View All View Keyframes

## What Is Driving Analytics?

1. Data availability
2. Inexpensive data storage
3. Faster computer processing

4 of 70

## Data: The Four Vs

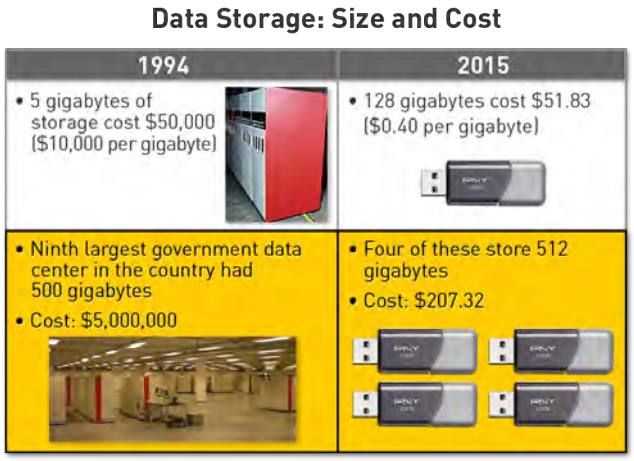
- Volume: amount of data
- Velocity: how fast you get new data
- Variety: different types of data
- Veracity: how accurate is the data

13 of 70

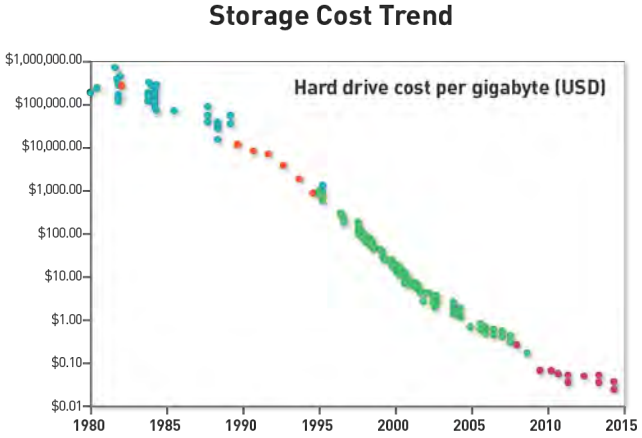
## Herb Simon, Nobel Laureate, Carnegie Mellon

- Data overload
  - "In an informaon-rich w orld, the wealth of informaon means a dearth of some thing else: a scarcity of whatever it is that informaon c onsumes."
  - "What informaon c onsumes is rather obvious: it consumes the a enon of its r ecipients."
  - "Hence a wealth of informaon cr eates a poverty of a enon and a need t o allocate that a enon e fficiently among the overabundance of informaon sources that might consume it."

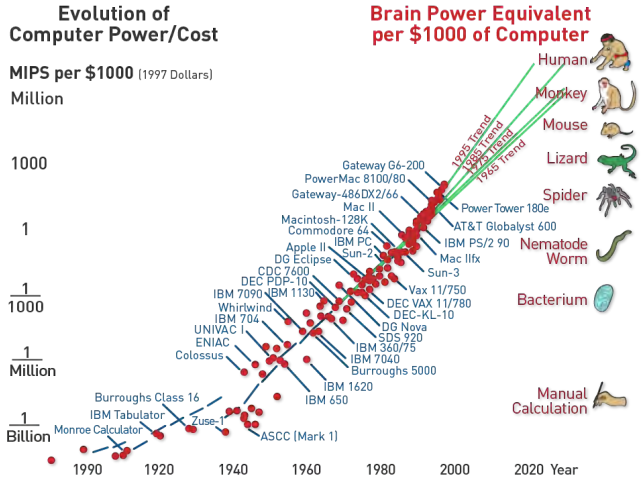
16 of 70




29 of 70




30 of 70




37 of 70

## Computer Trends

- 1989: *Deep Thought* chess program developed by Carnegie Mellon researchers
  - Defeated by Garry Kasparov
- 1996: IBM refines program, renames it *Deep Blue*
  - Defeated again by Kasparov 4-2
- 1997: IBM's *Deep Blue* upgraded further
  - Defeats Kasparov 3.5-2.5
- 2011: IBM *Watson* defeats Jeopardy champions
- 2021–2074: Projected that computers will be superintelligent, i.e., smarter than humans

55 of 70

## How Do These Trends Apply to Business?

- 1990s: retail outlets tailored customer prices to the individual
- 2000: Amazon.com tailored customer web pages based on browsing and purchasing behavior in 350 milliseconds (0.35 seconds)
- 2010: highly customized content delivery to customers

60 of 70

## Business Analytics Applications

- Retail sales
- Financial services
- Risk and credit
- Markeng and pricing
- Fraud
- Inventory control and supply chain

67 of 70

## Questions

- In your industry, how has analycs a  ffected your business?
- How do you expect it to affect your company in the next decade?

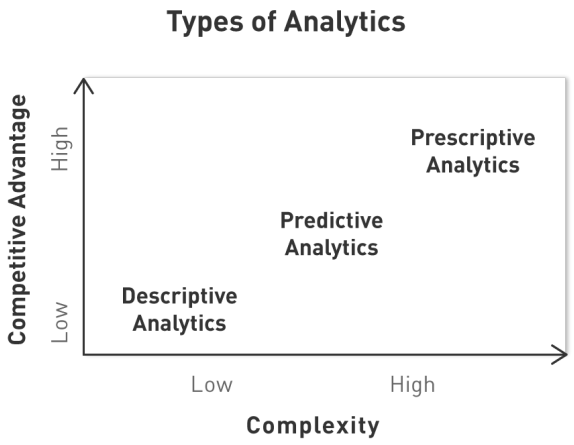
70 of 70

View All View Keyframes

## Business Analytics: What Is It?

- Goal:
  - Analyzing data to facilitate achievement of business objecv es
- How?
  - Reporng of da ta to analyze trends
  - Creang pr edicv e models for forecasng
  - Opmizing business pr ocesses for enhanced performance

7 of 56

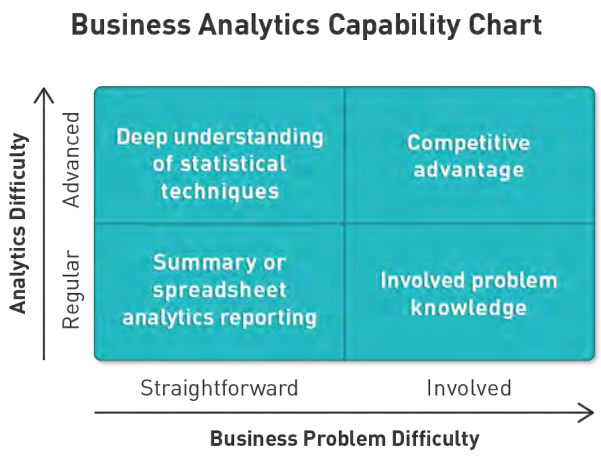



8 of 56

## Summary: Types of Analytics

- Descripv e analycs
  - Describe the data to be er understand facts and make be er decisions
  - Start with visualizaon
- Predicv e analycs
  - Forecasng and modeling
  - Cluster analysis, regression, neural networks
- Prescripv e analycs
  - Opmiz aon

23 of 56




24 of 56

## Why Is Business Analytics Difficult?

- Data is power.
  - Gartner Group reports that analycs will fuel disrup on in or ganizaons and business pr ocesses
- Industry lacks the skills to quickly move into analycs.
  - McKinsey reports a shortage of 140,000–190,000 staff with analycs e xperse
  - And a shortage of 1.5 million managers and analysts to understand and make decisions

41 of 56

## Business Analytics Pitfalls

1. Informaon and disin formaon o verload
2. Overanalysis
3. Oversimplificaon
4. Determinisc thinking (be c areful with causality)
5. Overdependence on industry knowledge
6. Tunnel thinking
7. Overconfidence
8. Unpiloted big bang launches

50 of 56

## Cultural Myths

- *Data* in my industry are different, beyond outsider comprehension.
- *Stas c al tools* in my industry are different, beyond outsider comprehension.

54 of 56

## Question

- What are the barriers to using analycs in y our organizaon?

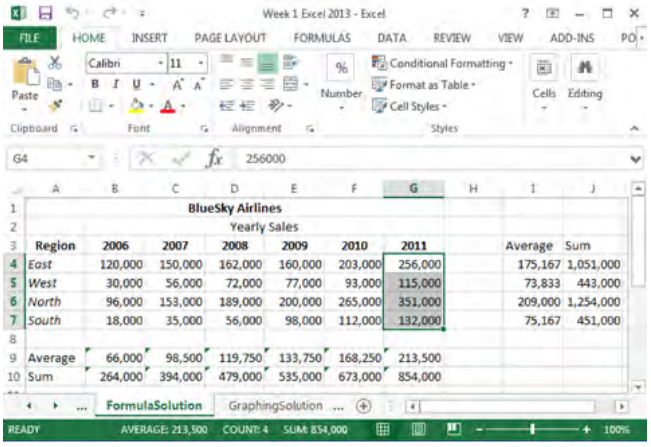
56 of 56

View All View Keyframes

## Excel Overview

- Calculaons and f ormulas
- Graphing and visualizaon
- Sorng and filt ers
- Pivot tables and charts
- Power View

6 of 25

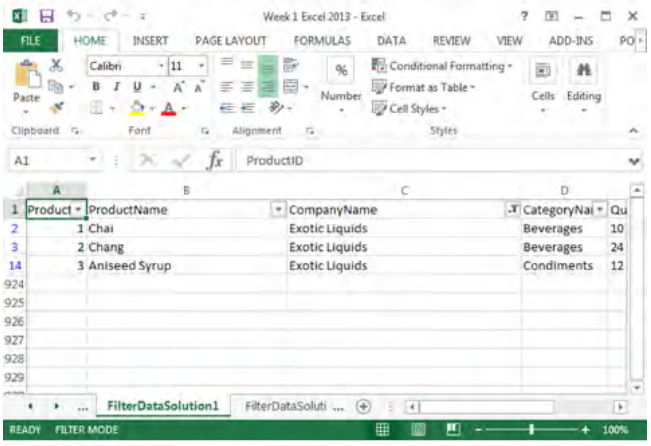



7 of 25

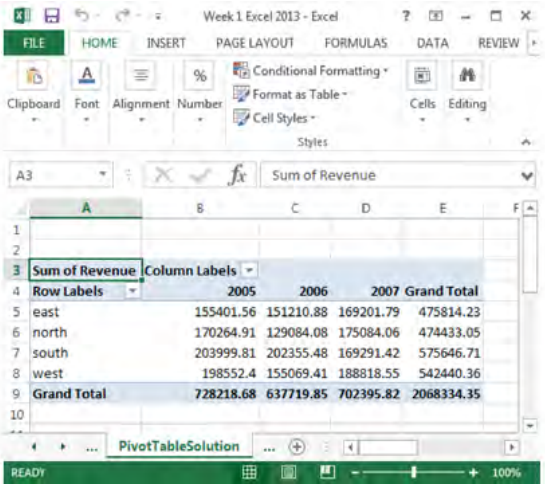
## Graphing and Visualization

- Ability to generate:
  - Bar charts
  - Sca er diagrams
  - Pie charts
  - 3–D representaons of da ta

16 of 25

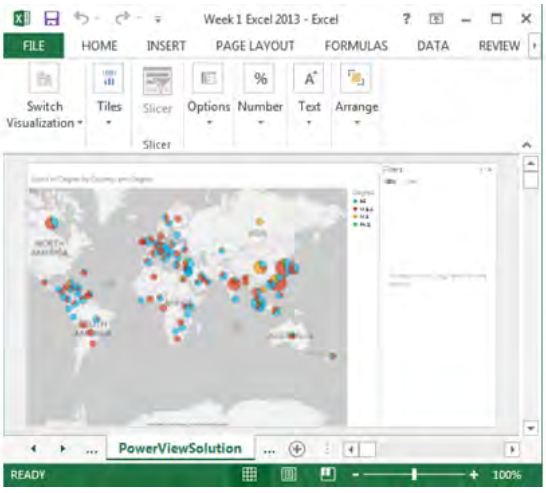



17 of 25




| Sum of Revenue | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | 2005 | 2006 | 2007 | Grand Total |
| east | 155401.56 | 151210.88 | 169201.79 | 475814.23 |
| north | 170264.91 | 129084.08 | 175084.06 | 474433.05 |
| south | 203999.81 | 202355.48 | 169291.42 | 575646.71 |
| west | 198552.4 | 155069.41 | 188818.55 | 542440.36 |
| Grand Total | 728218.68 | 637719.85 | 702395.82 | 2068334.35 |

20 of 25




24 of 25

# SCM 651: Business Analytics

1

Week 1

# Agenda



- Introductions
- Course Overview
- How to be successful in this course
- Why Business Analytics?
- Review of Hands-on Exercises
- Group Discussion of Articles from Industry
- Examples from Industry
- Team Formation (deadline next class)

# Overview



- Introductions
  - Name, role, why did you take Business Analytics
- Course overview
  - Designed in fall 2013
  - Met with executives to identify technology and techniques
    - McKinsey, IRI, Proctor & Gamble, Unilever, National Grid, Constellation Brands, VWR International, Publicis Kaplan Thayer, Transaver
  - Objective: include technology & business interpretation of results
  - Reviewed ~50 books – no book purchase required
  - Reviewed numerous articles – free download from library
  - Will use software recommended by executives – Excel, Access, Google Analytics, R, Tableau (free access)
  - Basis for analytics courses in Accounting, Finance. Marketing, and Supply Chain Management; also good for Lean Six Sigma

# Course content



- Syllabus review
  - Four group homework assignments
  - Individual participation and peer review score
  - Individual final exam
- Grading curve
  - May be adjusted at end of semester
- Weekly articles
  - Download articles from library (free), read and be prepared to discuss in the live session
- Hands-on exercises
  - Run all exercises in instructions on website

# How to be successful in this course



- Review all asynchronous videos before attending the live session
- Use the instructions and datasets which accompany the asynchronous sessions to work through all examples before the live sessions
- Read the articles each week and participate in the discussion (prepare answers to questions before the live session)
- Attend all live sessions on time
- Work with your teams to learn from each other
- Attend office hours if you don't understand something

# Why Business Analytics?



- Demand for expertise
  - Accenture Analytics now has 20,000 employees in the management consulting and technology area[1]
  - Procter & Gamble it is in the process of quadrupling its analytics expertise[1]
  - 140,000 to 190,000 shortfall of talent with deep analytical skills (McKinsey)[2]
  - 4.4 million demand for analytical skills (Gartner Group) [2]

(1) http://www.informationweek.com/software/information-management/big-data-talent-war-10-analytics-job-trends/d/d-id/1103566?
(2) http://fortune.com/2013/09/04/the-big-data-employment-boom/

# Remote Access to Hands-on Exercises



- Using any browser, go to:

  remote.whitman.syr.edu

- All software is on the remote site
- All course data files are available in the G: drive at:

  g:/WHIT/SCM651/Online/

# Hands-on Exercises



- Calculations and formulas
- Graphing and visualization
- Sorting and filters
- Pivot tables and charts
- PowerView (2013 or add-in in 2016)

# Article #1: Business Analytics



- Business Analytics Insight: Hype or Here to Stay?
  - How are companies using analytics?
  - Which analytics techniques does your company use?

# Article #2: GE



- GE and the Culture of Analytics
  - How is GE using analytics for process improvement?
  - What did they do to accelerate analysis and results?
  - Why is culture important?

# Article #3: Location Analytics



- Location Analytics: Bringing Geography Back
  - What are some applications of geographic data in businesses?
  - What are the advantages of consolidating individual accounts by location? What are the risks of consolidation?
  - How is your company using geographic data?

# Example of Analytics in Industry



- L.L. Bean pricing
- Amazon.com pricing

# Google Analytics & Team Formation



- Google Analytics
  - A Google Analytics account will be set up by your instructor
- Teams
  - Form teams of 4-5 students each
  - Send email to instructor with your teams no later than Friday
  - Teams to be finalized in next class
- Office Hours

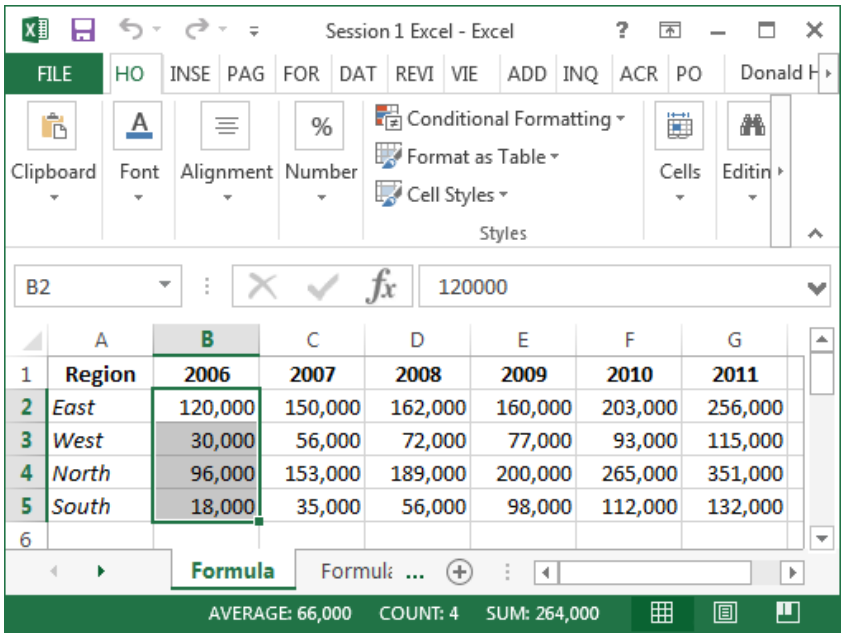
# Microsoft Excel: Basics

## Microsoft Excel

- Go to the Course Website
- Download and print “Business Analytics Week 1 Instructions.doc”
- Download “Business Analytics - Week 1 Excel.xlsx”

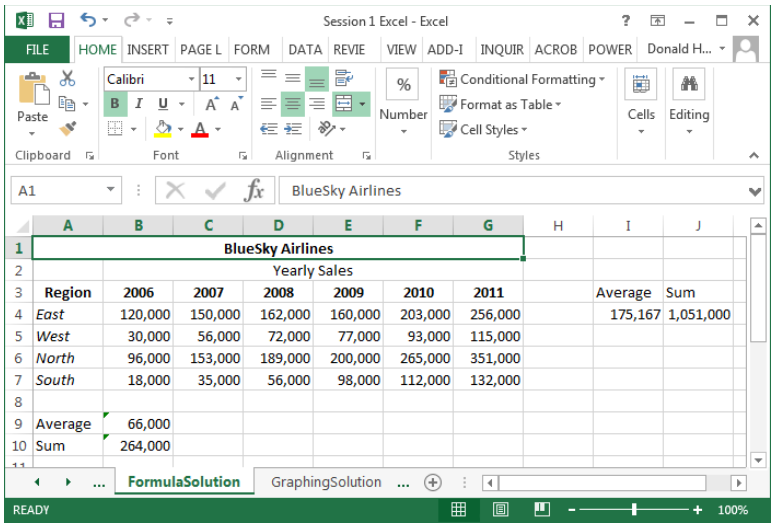
## Session 1.7: Calculations and Formulas

- Highlight the four sales values for 2006. Notice at the bottom of the page, it calculates the average, count and sum of the values.



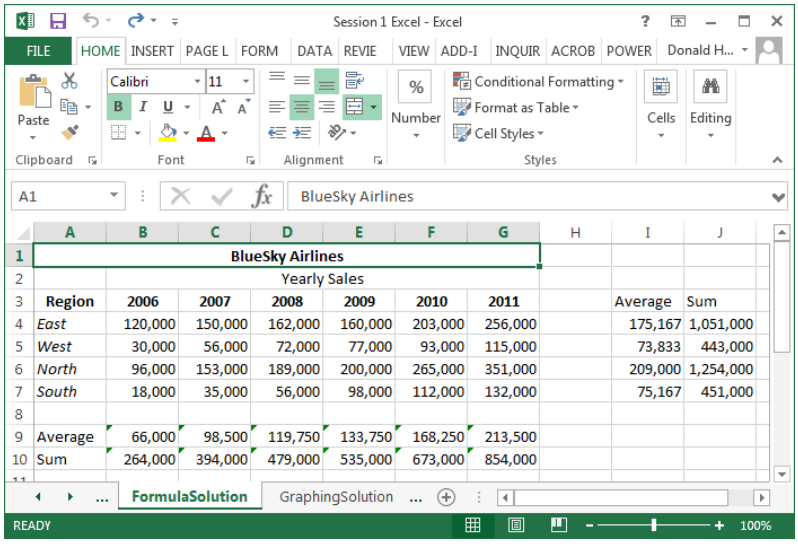

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Region | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
| 2 | East | 120,000 | 150,000 | 162,000 | 160,000 | 203,000 | 256,000 |
| 3 | West | 30,000 | 56,000 | 72,000 | 77,000 | 93,000 | 115,000 |
| 4 | North | 96,000 | 153,000 | 189,000 | 200,000 | 265,000 | 351,000 |
| 5 | South | 18,000 | 35,000 | 56,000 | 98,000 | 112,000 | 132,000 |
| 6 | | | | | | | |

- Next, we’ll enter a formula in the spreadsheet to automatically calculate averages and sums
- In cell A9, type the word Average; also enter Average in I3
- In cell A10, type the word Sum; also enter Sum in J3
- To calculate a formula, use the equal sign, the name of the formula, and the data range
    - In cell B9, enter =average(b4:b7)
    - In cell B10, enter =sum(b4:b7)
    - In cell I4, enter =average(b4:g4)
    - In cell J4, enter =sum(b4:g4)
- Notice that after you type the equals sign and part of the formula name, it gives you several options that start with the same spelling; you can click on the one you want, then highlight the range of data



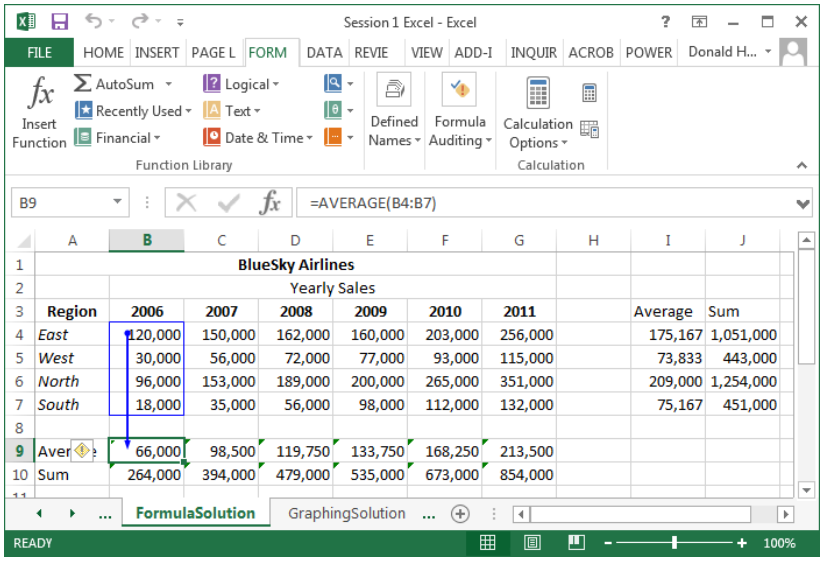

## Copy and Paste

- Copy and paste works the same in Excel as in Microsoft Word, but you can also copy and paste formulas and Excel will automatically update formulas
- Copy the average and sum formulas across all columns and rows
    - Click on cell B9, click on the copy button in the Clipboard section (or use control-C), highlight cells C9 to G9, and click the paste button in Clipboard
    - Copy cell B10 across for the columns in the same way
    - You can copy several formulas at once. Highlight cells I4 and J4. Click on copy, then highlight the range I5 to J7, and click paste.
    - Look at the formulas in B9 and C9. How are they different?



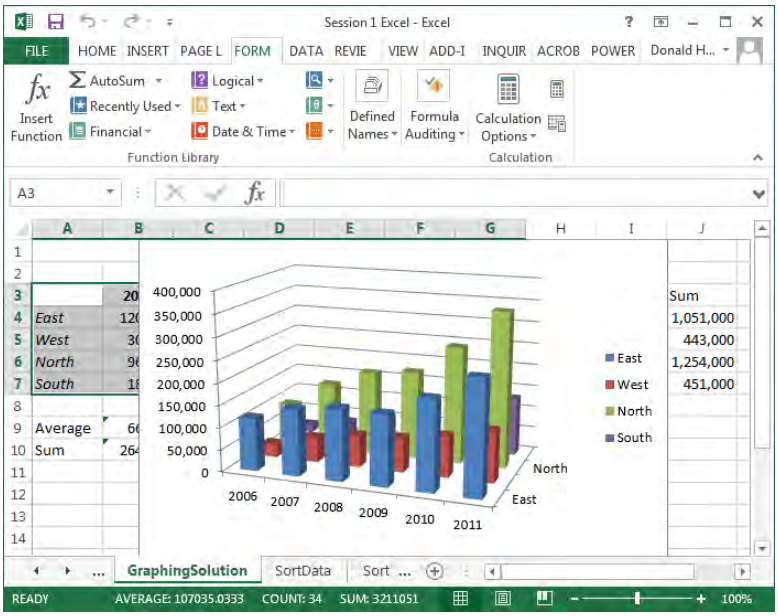
## Checking Formulas

- Excel has an auditing feature to check formulas; click on the tab labeled Formulas
- Click on B9, then click on Trace Precedents in Formula Auditing
- To clear the arrows, click on Remove Arrows



## Session 1.8: Graphing Data

- Sometimes you might want to graph your data. Excel can graph columns and rows of data, but has difficulty if anything is in the upper left corner. First, delete the word Region by clicking on cell A3 and pressing the delete key.
- Highlight the data by clicking on the cell labeled A3 through G7.
- Next click on the Insert tab at the top of the screen.
- In the section labeled Charts, click on the tiny icon in the lower right corner of charts to bring up the possible chart options.
- Move the cursor over each options and select the 3-D column chart, click OK




- Let's move the chart to another sheet. Click on the upper right corner of Excel, select Move Chart Location, click on New Sheet, then OK.
- At the top of the screen click on Chart Tools: Design: Type, Change Chart Type to change the type of chart
- In the Chart Tools: Design: Data, click on Switch Row/Column. What happens?
- In the Chart Tools: Design: Chart Layouts, Quick Layout, there are several options (scroll down). How are they different?
- Next, try Chart Tools: Layout. Experiment with Labels and Axes. Under 3-D Rotation, change the X and Y rotation.

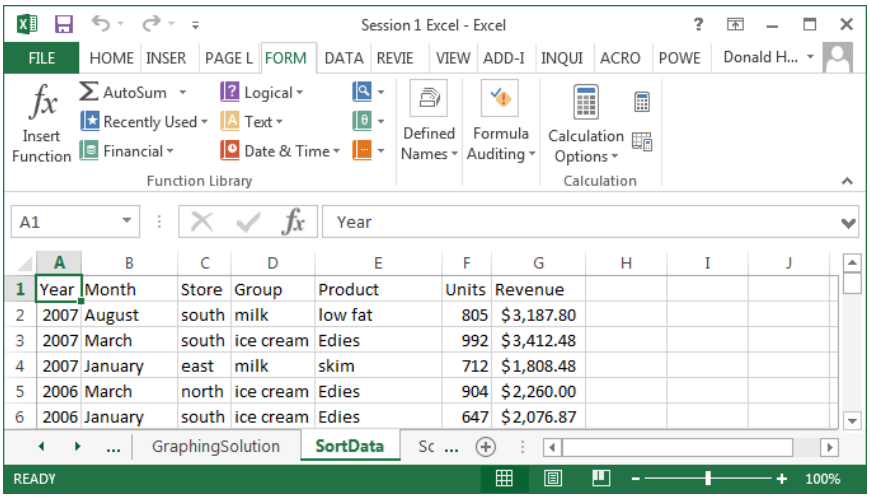
## Saving your Spreadsheet

To save a spreadsheet, click on the Microsoft Office icon in the upper left corner of the screen, move the cursor over Save As. You'll notice that you can save your Spreadsheet in a variety of formats.

# Microsoft Excel: Sorts, Filters, Pivot Tables

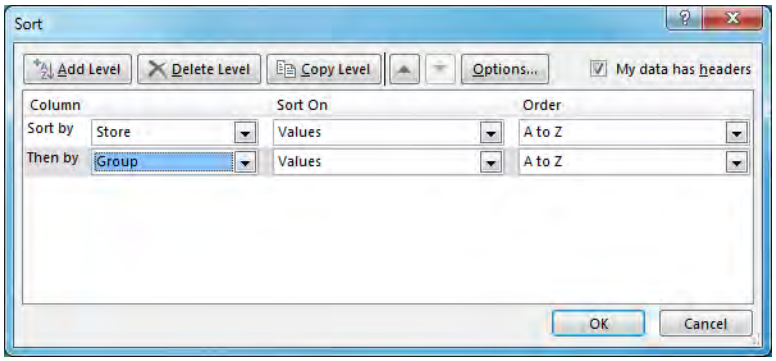
## Session 1.9: Sorting Data

- Use the SortData spreadsheet for this exercise.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Month | Store | Group | Product | Units | Revenue |
| 2 | 2007 | August | south | milk | low fat | 805 | $3,187.80 |
| 3 | 2007 | March | south | ice cream | Edies | 992 | $3,412.48 |
| 4 | 2007 | January | east | milk | skim | 712 | $1,808.48 |
| 5 | 2006 | March | north | ice cream | Edies | 904 | $2,260.00 |
| 6 | 2006 | January | south | ice cream | Edies | 647 | $2,076.87 |

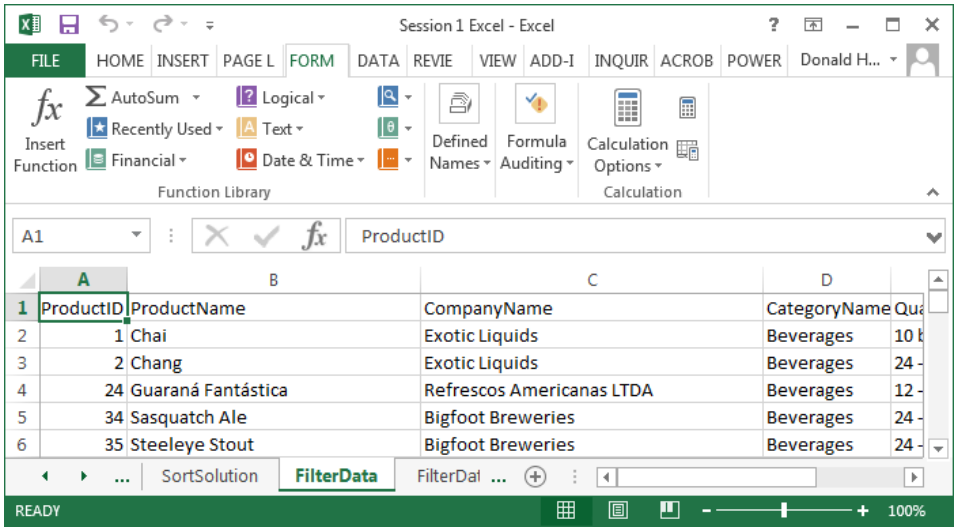
- To sort data, highlight the top of the columns (A through G). The entire columns should be highlighted.
- Click on the Data tab at the top of the spreadsheet.
- Click on Sort; note that “My data has headers” is checked. Why is it checked?
- Select the first sort criteria: Store
- Click on Add Level, then add the second sort criteria: Group



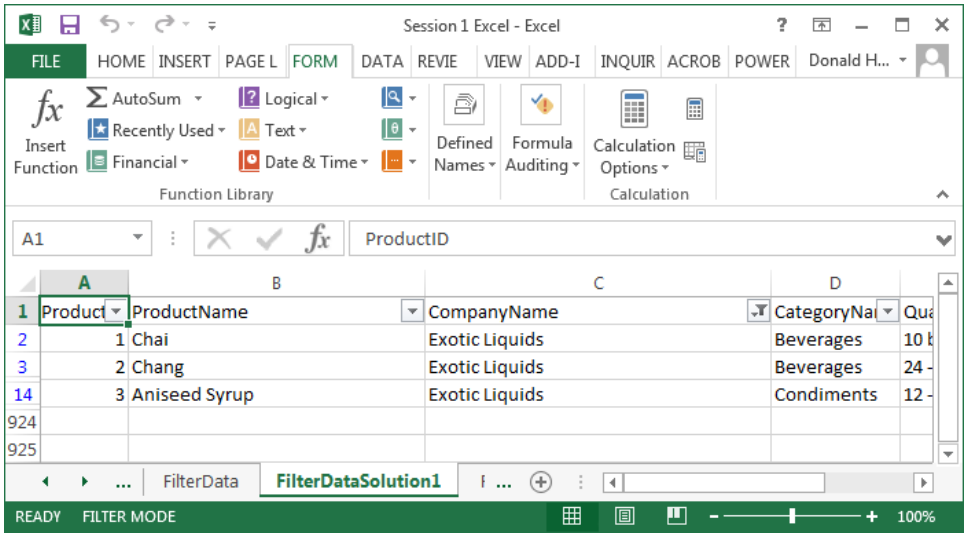

- Click OK, and the data is sorted
- How would you sort from Z to A (reverse alphabetical order)? Why would you want to sort in a different order?

## Filters

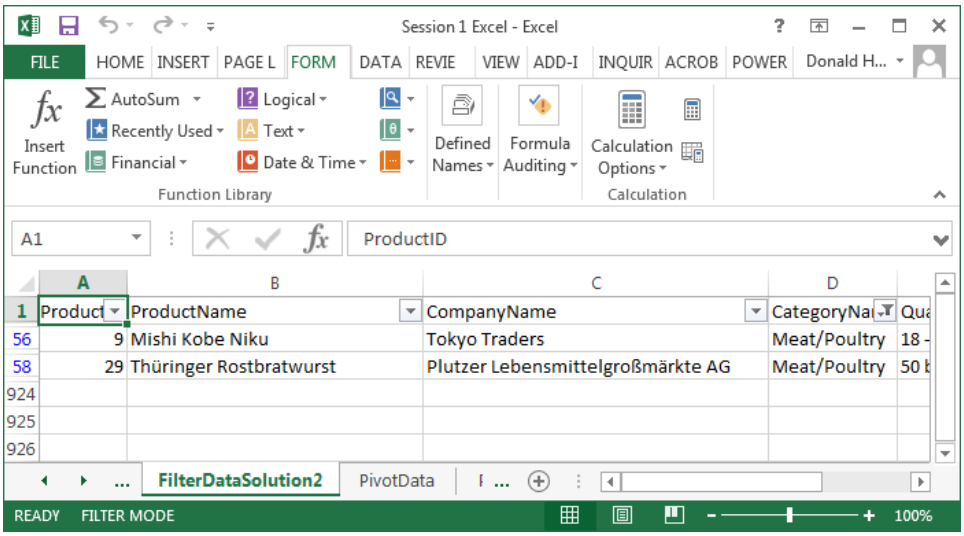
- Use the FilterData spreadsheet for this exercise
- A filter allows you to select specific data out of a long list
- First, click on cell A1, then click on the Data tab, then Filter (looks like a funnel)




- Notice that there are drop down arrows for each column. Let's find all products which Exotic Liquids produces.
- Click on the drop down arrow next to company name; uncheck "(Select All)", then check Exotic Liquids and OK. What happens?



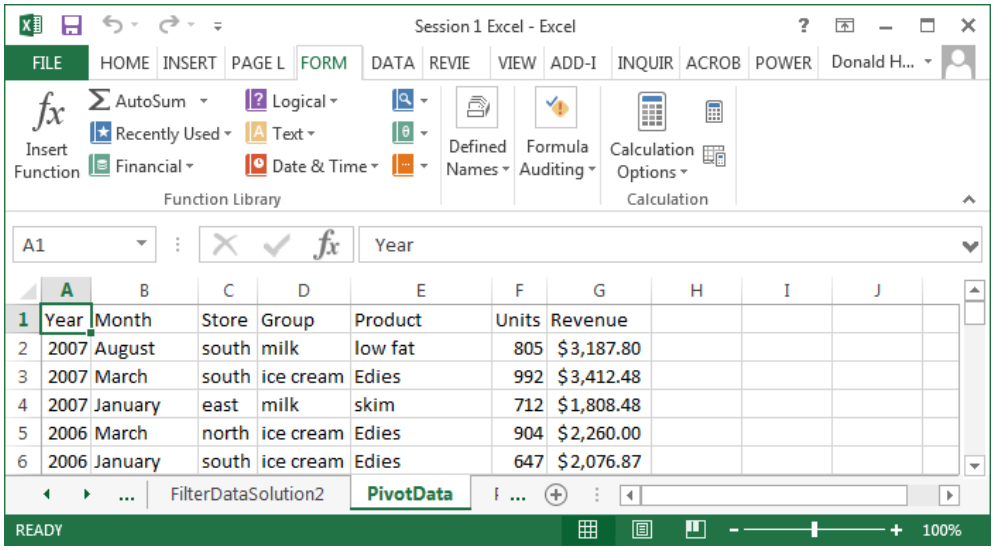

- Go back and click Select All for Company Name
- To find which products have a price greater than 50, click on the down arrow next to unit price, Number Filters, Greater Than, enter 50 in the field, and click OK
- You can turn on multiple filters. Add a filter for Category Name equal to Meat/Poultry to see which Meat/Poultry products cost more than 50




- To turn off the filter, click on the Filter icon and the drop down arrows should disappear.

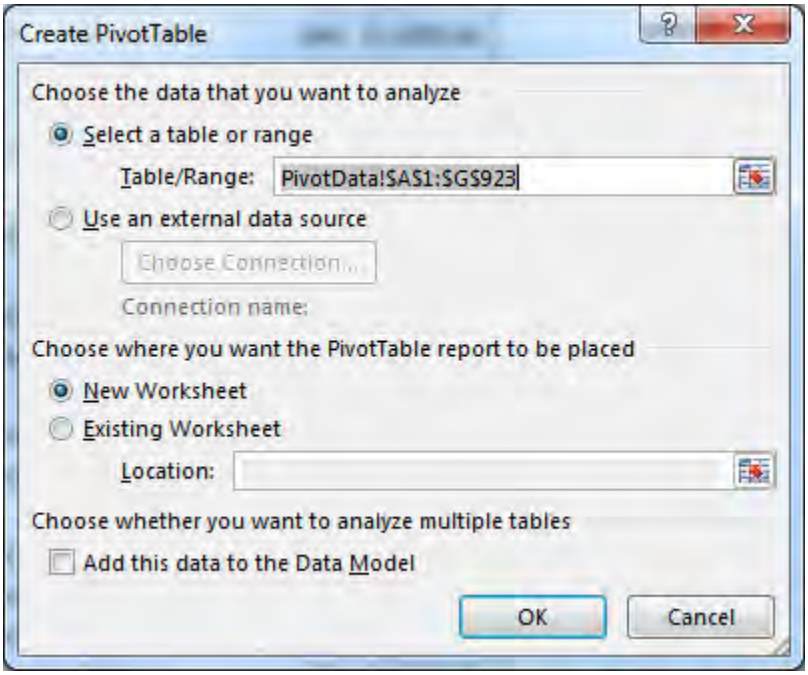
## Session 1.10: Pivot Tables and Charts

Pivot tables and charts are powerful techniques to quickly summarize and display large amounts of data. For this example, use the PivotData spreadsheet.

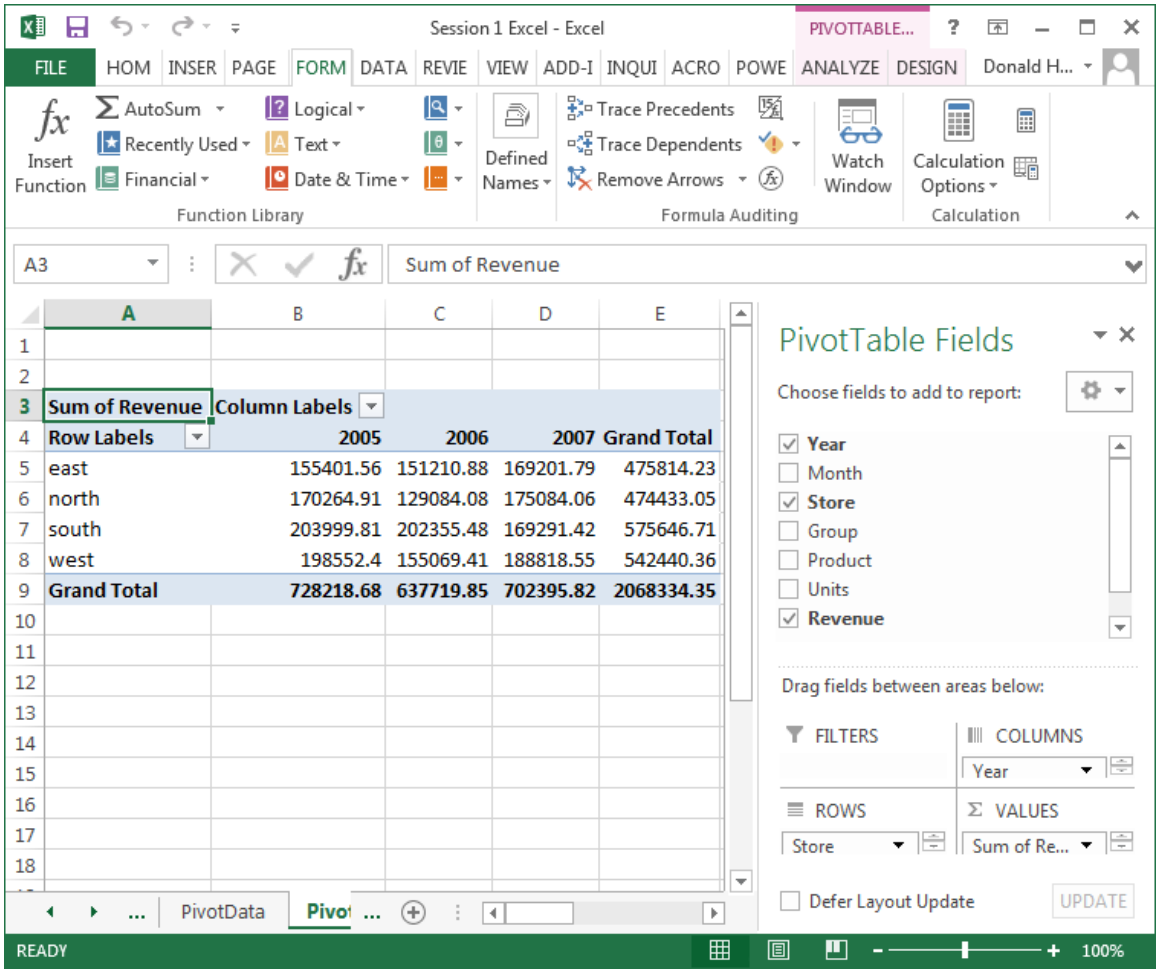



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Month | Store | Group | Product | Units | Revenue |
| 2 | 2007 | August | south | milk | low fat | 805 | $3,187.80 |
| 3 | 2007 | March | south | ice cream | Edies | 992 | $3,412.48 |
| 4 | 2007 | January | east | milk | skim | 712 | $1,808.48 |
| 5 | 2006 | March | north | ice cream | Edies | 904 | $2,260.00 |
| 6 | 2006 | January | south | ice cream | Edies | 647 | $2,076.87 |

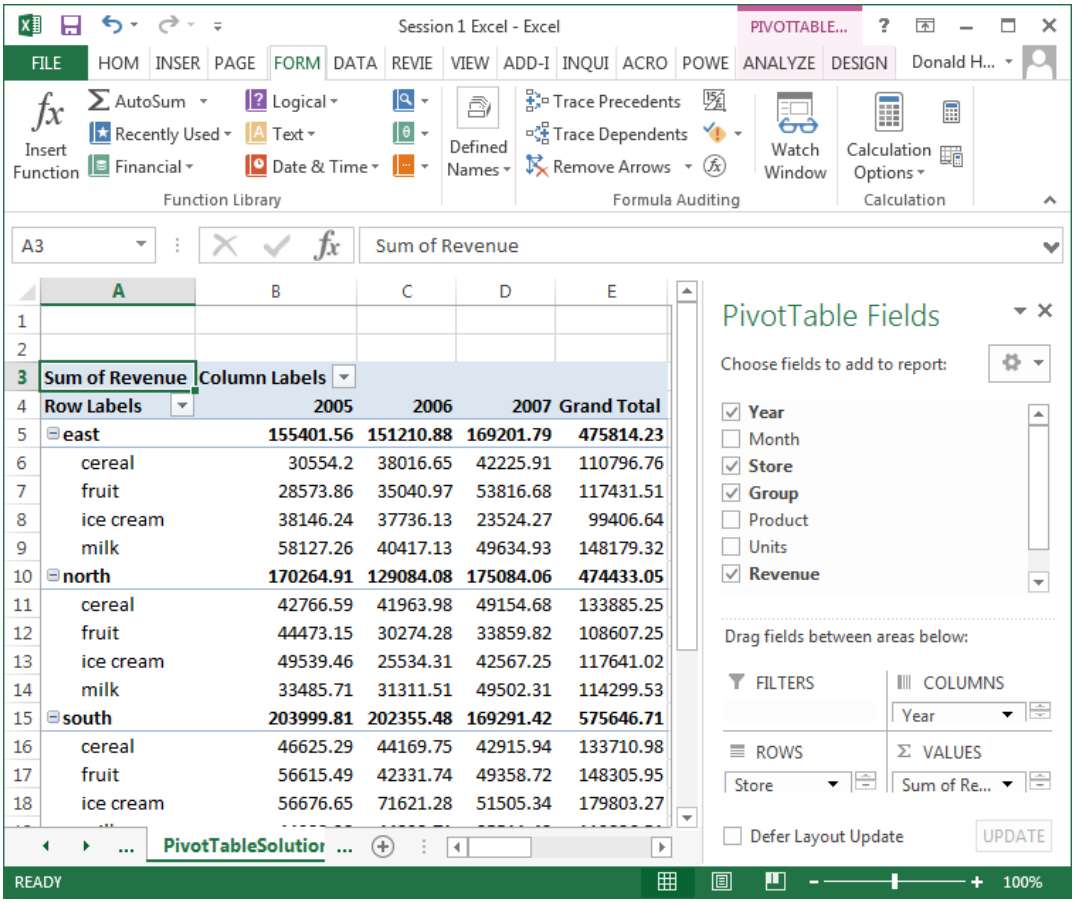
1. The headings must be in the first row of your data.
2. Click anywhere in the data table.
3. Click on the Insert tab.
4. Click on PivotTable; it should automatically highlight the entire table.
5. Make sure New Worksheet is checked, then click OK.



6. In the pivot table dialog box, we want to add row labels, column labels, and identify what goes into the table.
7. For this example, let's build a table with store revenue by year. First click on store and drag it to the row label. These will become the row labels on the left side of your table.
8. Next, click on year and drag it to the column label. These will become the column labels at the top of your table.
9. Finally, click on revenue and drag it to the values cell. These values will feed the interior of the table. The default function is Sum. If you want to change this function, click on the down arrow in the values box, then select the function (average, min, max, etc.)

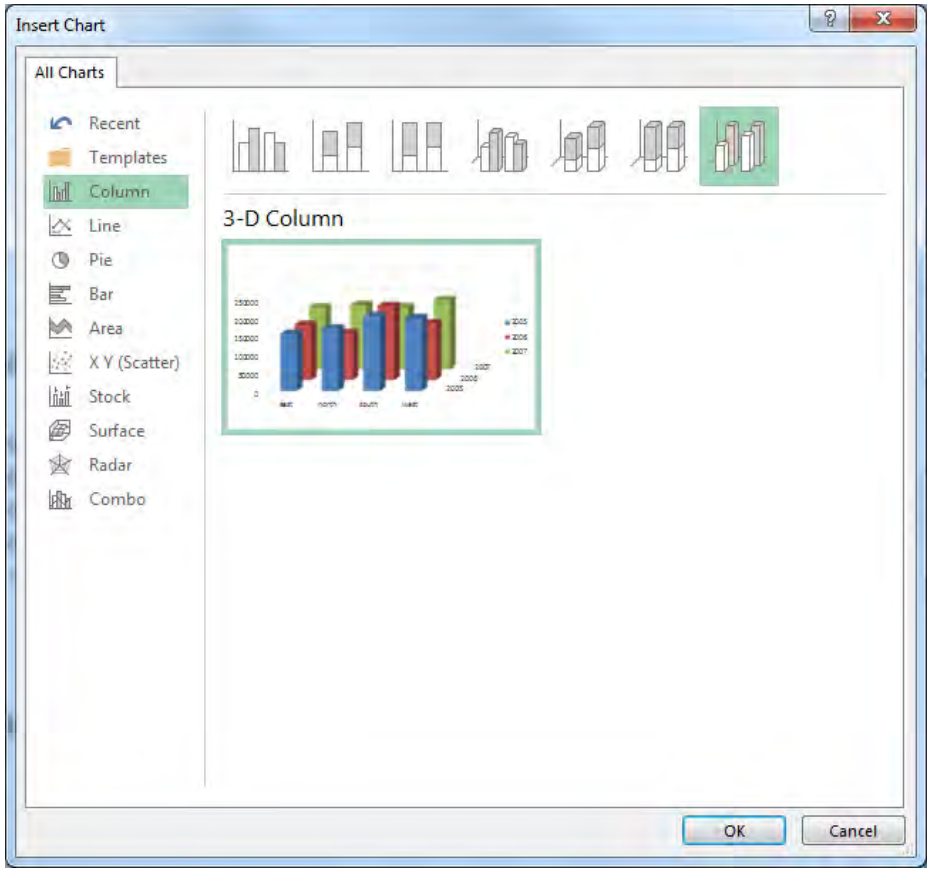


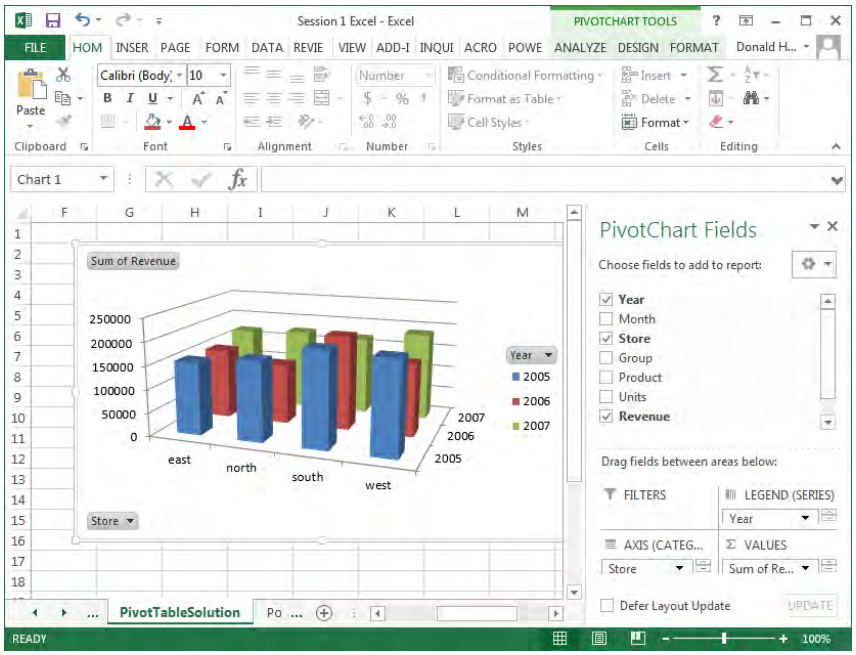
Add group to the row labels to see what happens.




Let's now create a Pivot Chart. The steps are:

1. Uncheck Group to simplify the data.
2. Click anywhere inside your Pivot Table. An Options tab will appear. Click on Pivot Chart.
3. A chart dialog box will appear. Click on your preferred type of chart.




To move the chart to its own page:

1. Right click on the picture
2. Click on Move Chart
3. Select New Sheet and name the new sheet

# Microsoft Excel: PowerView

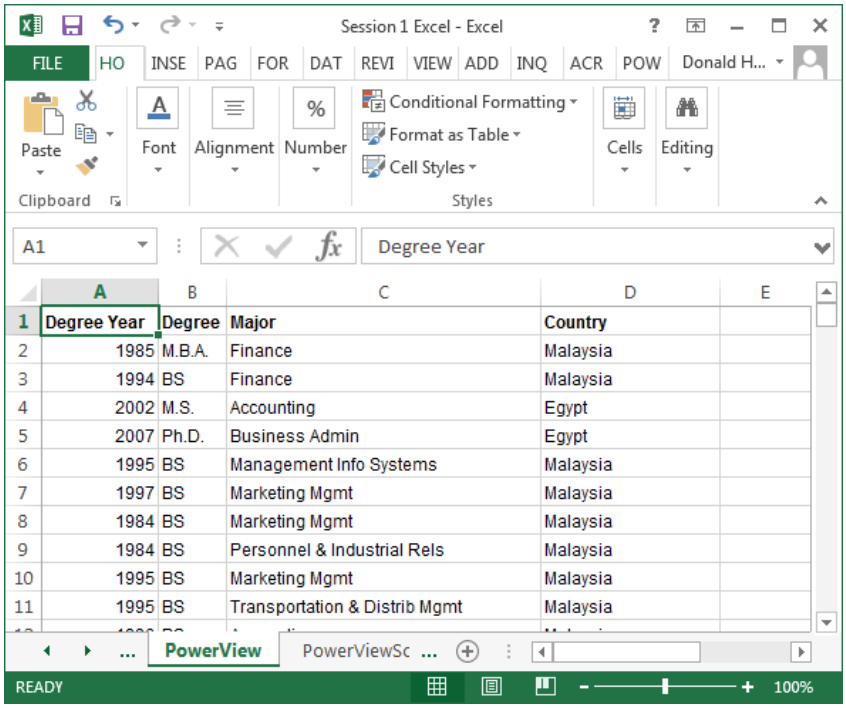
**Session 1.11: Powerview**

**PowerView Add-in**

The PowerView options are available as an add-in to Excel. The steps to add it are:

1. In Excel, click on the File tab, then Options
2. Click on Add-Ins
3. In the Manage box at the bottom of the screen, select COM Add-ins, then Go
4. Check the box for PowerView, then OK

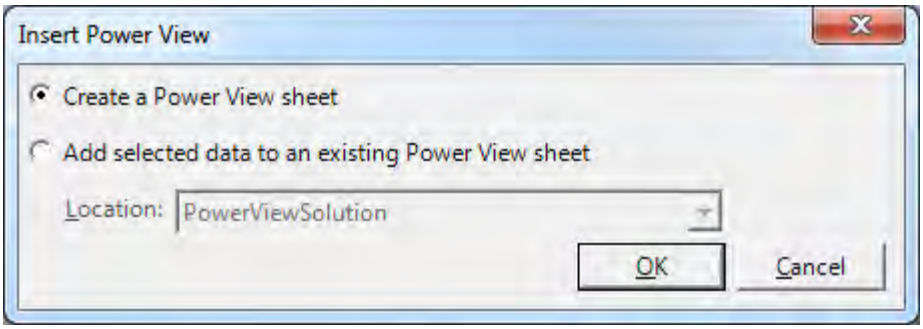
PowerView allows you to quickly generate business analytics graphs of complex data. PowerView is only available in Microsoft Excel 2013. Click on the PowerView spreadsheet for this exercise. This data is a list of international alumni from the Whitman School.

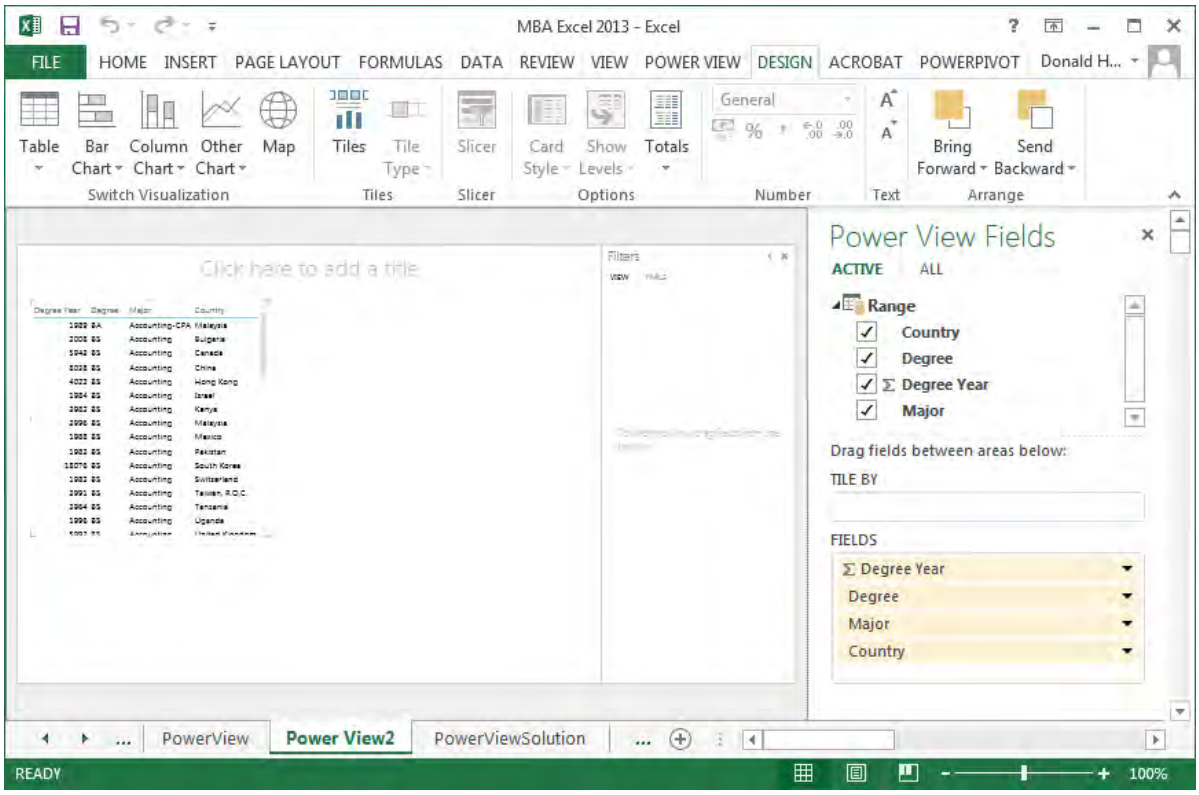



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Degree Year | Degree | Major | Country | |
| 2 | 1985 | M.B.A. | Finance | Malaysia | |
| 3 | 1994 | BS | Finance | Malaysia | |
| 4 | 2002 | M.S. | Accounting | Egypt | |
| 5 | 2007 | Ph.D. | Business Admin | Egypt | |
| 6 | 1995 | BS | Management Info Systems | Malaysia | |
| 7 | 1997 | BS | Marketing Mgmt | Malaysia | |
| 8 | 1984 | BS | Marketing Mgmt | Malaysia | |
| 9 | 1984 | BS | Personnel & Industrial Rels | Malaysia | |
| 10 | 1995 | BS | Marketing Mgmt | Malaysia | |
| 11 | 1995 | BS | Transportation & Distrib Mgmt | Malaysia | |

To generate a PowerView report:

1. Click on the Insert tab at the top of the screen
2. In the Reports section, click on Power View
3. In the Insert Power View screen, click OK.

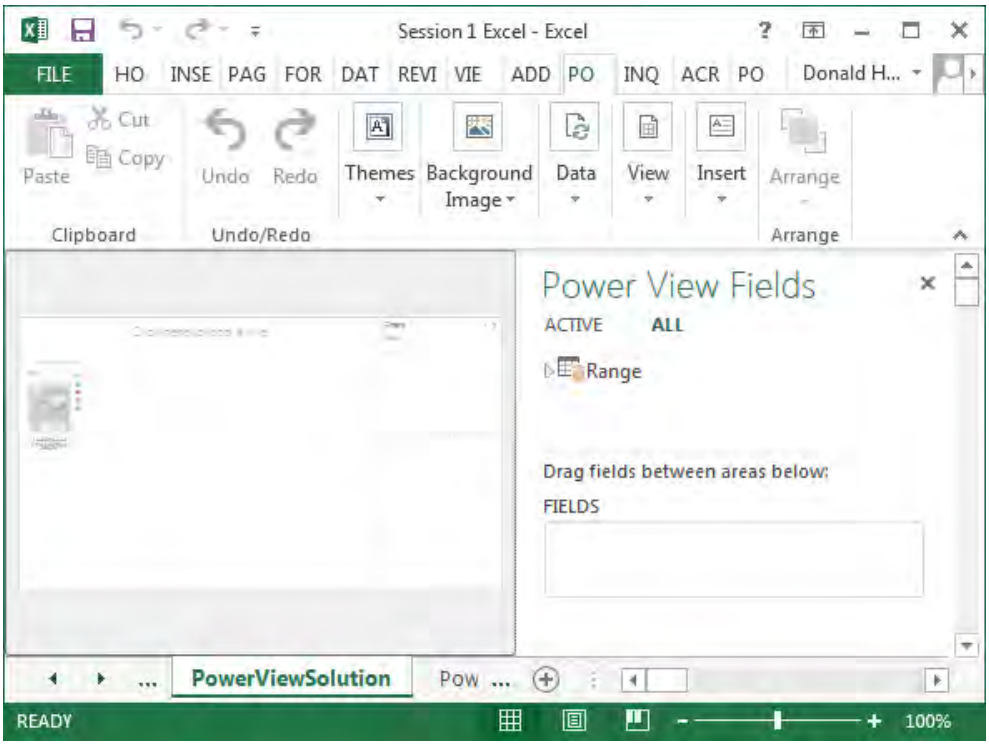



On the left side of the screen is your data. On the right side are the data fields or column labels (country, degree, degree year, major).

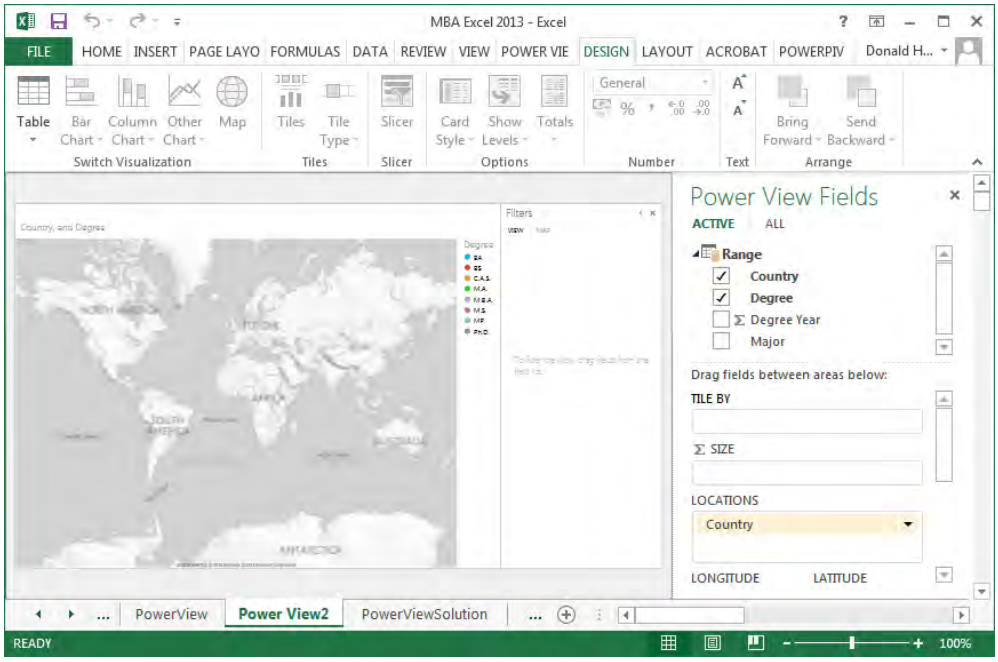
To create a map of degrees by country:

1. Uncheck Degree Year; uncheck Major
2. Click on the Map icon in the upper left portion of your screen

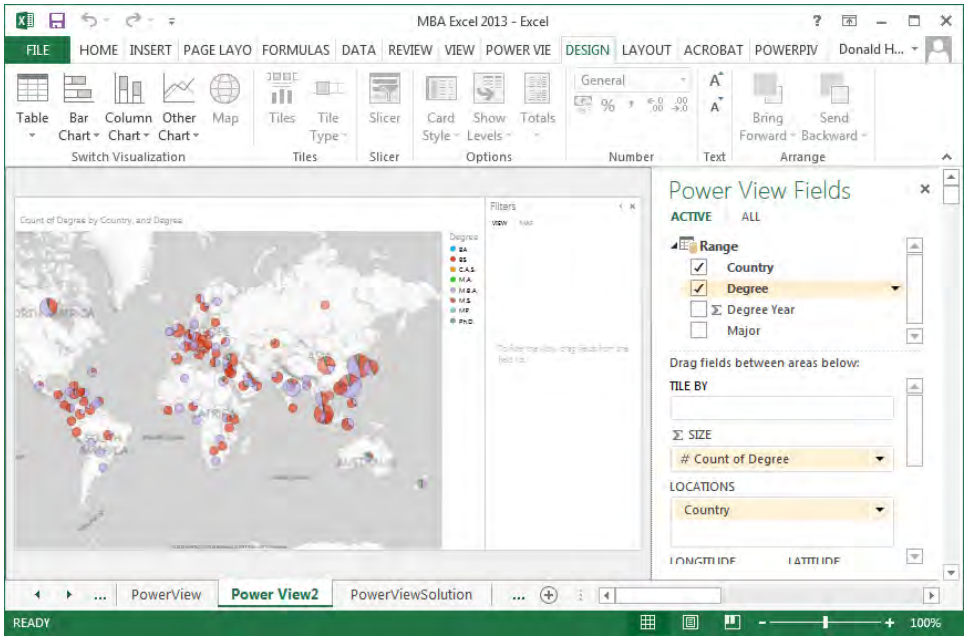



The map on the left is not very clear.

1. Move your cursor over the map.
2. An arrow in the upper right of the picture should appear.
3. Move the cursor over the arrow; it should say Pop out.
4. Click on the arrow.



Now we have a picture, but no data. Notice that in the field labeled ∑ SIZE has no field entered. This field determines the data to be displayed. Drag Degree from Power View Fields to ∑ SIZE.



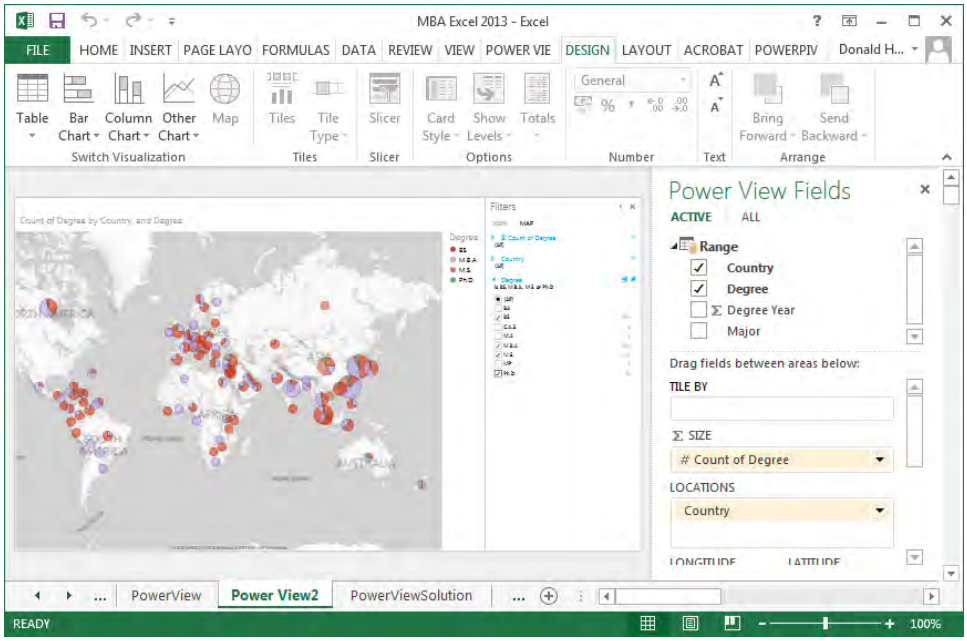

Be sure to have the following on the right side of the screen:

1. Drag Degree to ∑ SIZE; it should show # Count of Degree
2. LOCATIONS should have Country
3. COLOR should have Degree

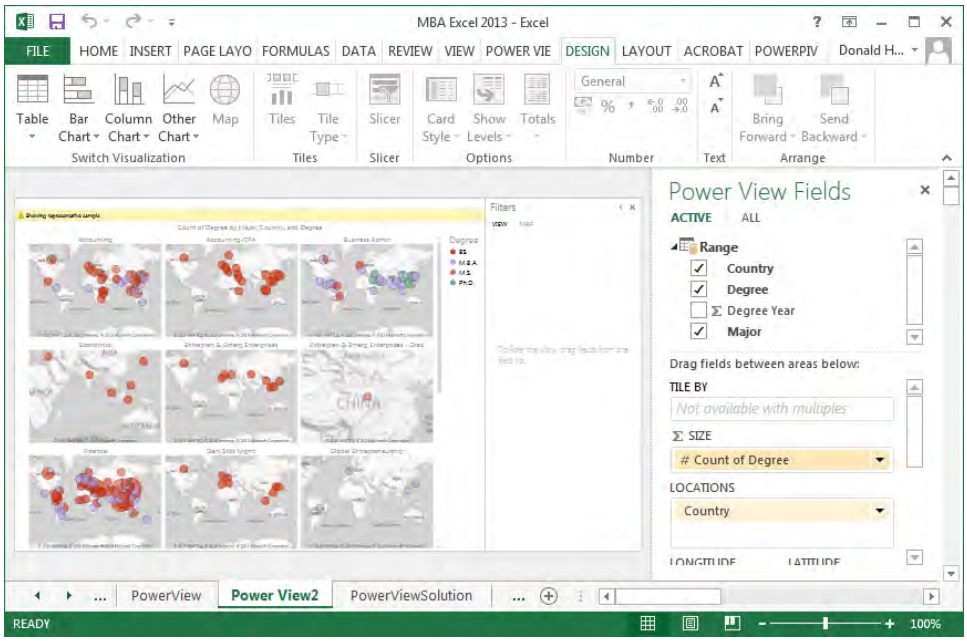
Put your cursor over a country's pie chart. By moving the cursor over different parts of the pie chart, it will give you the data associated with each pie segment.

You can also filter data to limit the data displayed on the screen. To activate a filter:

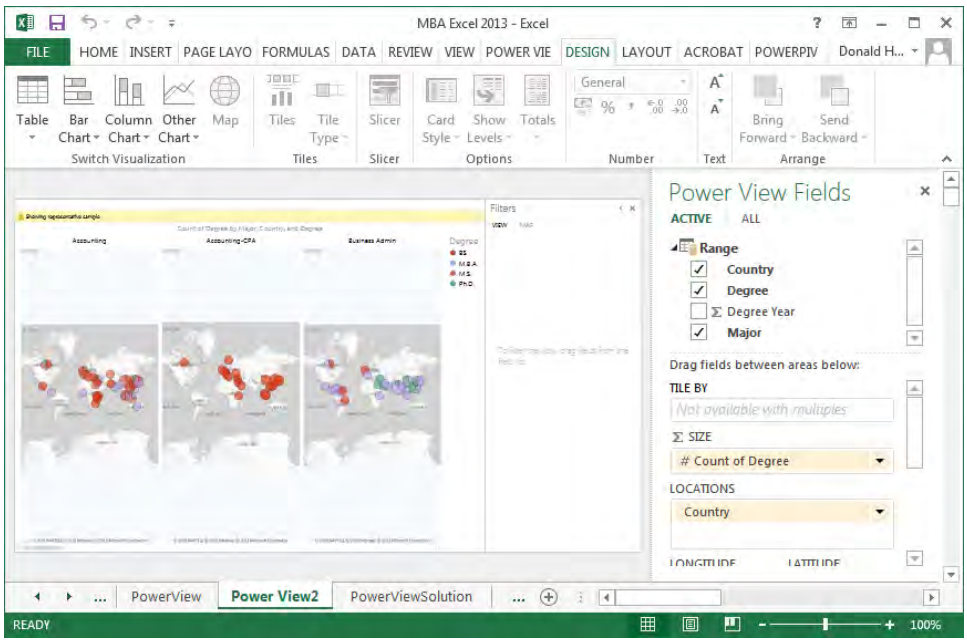
1. Click on the funnel to the upper right of the map.
2. In the Filters arrow to the right of the map, click on the arrow next to Degree
3. A list of degrees should appear. Check the degrees BS, MBA, MS and PhD
4. To reset to all degrees, click on All


MBA Excel 2013 - Excel
FILE
HOME
INSERT
PAGE LAYO
FORMULAS
DATA
REVIEW
VIEW
POWER VIE
DESIGN
LAYOUT
ACROBAT
POWERPIV
Donald H...
Table
Bar Chart
Column Chart
Other Chart
Map
Switch Visualization
Tiles
Tile Type
Slicer
Card Style
Show Levels
Totals
Options
General
Number
Text
Bring Forward
Send Backward
Arrange
Count of Degree by Country, and Degree
Filters
Power View Fields
ACTIVE
ALL
Range
Country
Degree
Degree Year
Major
Drag fields between areas below:
TILE BY
SIZE
# Count of Degree
LOCATIONS
Country
PowerView
Power View2
PowerViewSolution
READY
100%

To add an additional layer of analysis, drag Major to Vertical Multiples.




Drag it to Horizontal Multiples.



View All View Keyframes

## Time Value of Money

- Would you prefer $1,000 today or $1,000 one year from now?
  - Today
  - Opportunity to invest
- Would you prefer $1,000 today or $1,200 one year from now?
  - Dependent upon interest rate or opportunity to invest
- **Time value of money:** value of money decreases over mea

7 of 21

## Net Present Value (NPV)

Converts future cash flow stream into today's dollars

- Cash flow by date
  - Investments: negav e numbers
  - Returns and savings: posiv e numbers
- Interest rate
- **Net present value:** what a given cash flow is worth to you today

14 of 21

## NPV Interpretation

- **Positive NPV:** make more money than bank interest rate
  - Compare alternav es for highest NPV
- **Negative NPV:** lose money compared to bank interest rate

20 of 21

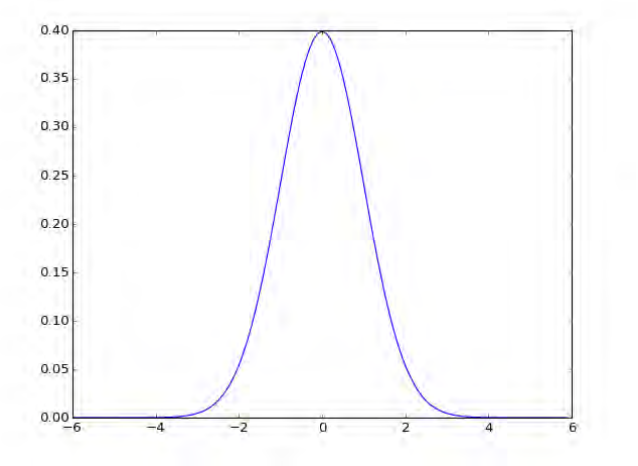
View All View Keyframes

## Internal Rate of Return (IRR)

- Measures the effecv e rate of return for your cash flow stream
- Calculates exact interest rate you're earning
- For example:
  - Posiv e NPV: goal of 7% interest rate converts to 7.3% IRR
  - Negav e NPV: goal of 7% interest rate converts to 3% IRR
- Calculates exact interest rate you would achieve if NPV is set to zero

6 of 7

View All View Keyframes




8 of 14

## Descriptive Statistics

- Simple measures of data
- Examples:
  - Mean: arithmec a  verage
  - Median: middle point in distribuon
  - Mode: most common value (highest frequency of occurrence)
  - Kurtosis: height of data peak relav  e to normal distribuon
  - Skewness: left or right posion of da  ta relav  e to normal distribuon
  - Standard deviaon: measur  e of spread
  - Range: highest value minus lowest value

14 of 14

View All View Keyframes

## Correlations Overview

- The measure of how two variables are related
- Examples:
  - Educaon and earning pot enal (posiv e correlaon)
  - Product price and product demand (negav e correlaon)
- Posiv e correlaon: v ariables moving in same direcon
- Negav e correlaon: v ariables moving in opposite direcons

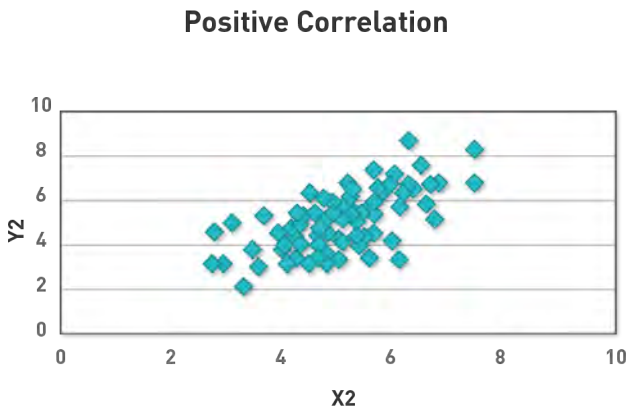
8 of 24

## Types of Correlations

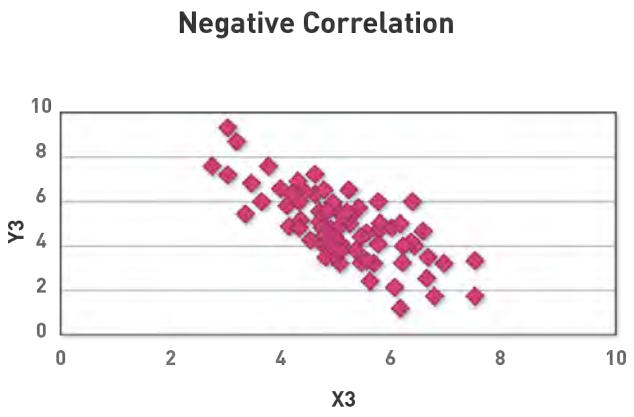
- Negav e
- Posiv e
- Zero

12 of 24

13 of 24




14 of 24




15 of 24

## Correlation Measurements

- Measured on a scale from −1 to +1.
  - Perfectly negav e correlaon: −1
  - Perfectly posiv e correlaon: +1
  - No relaonship be tween variables: 0
- Number reflects how strong the correlaon is.
- Correlaon does not mean c ausaon.
  - A much stronger relaonship is r equired to show causaon.

23 of 24

View All View Keyframes

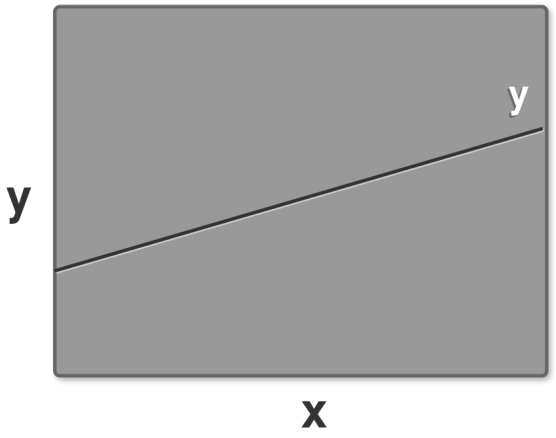
Correlation: whether variables move together or in different directions

1 of 39

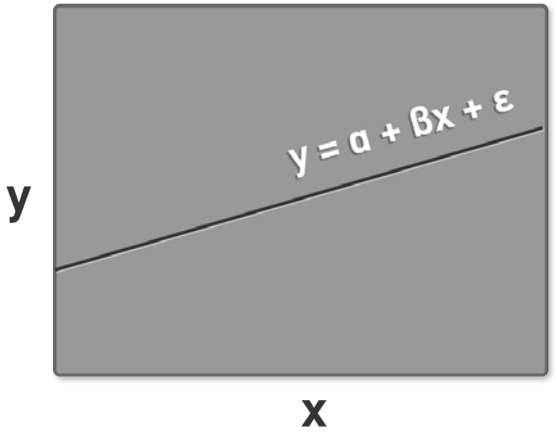
## Regression

- Measure of change associated with one variable and its effect on another
- E.g., how much interest rates affect mortgage applicaons

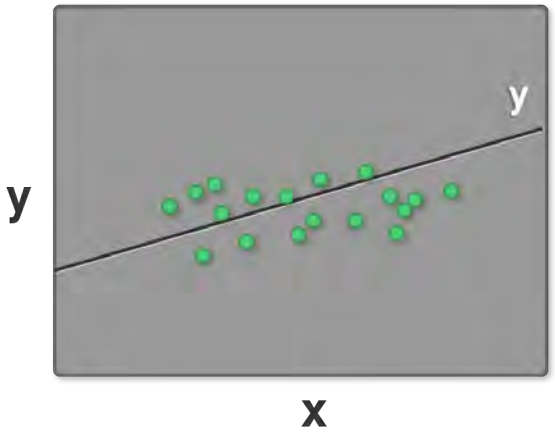
3 of 39



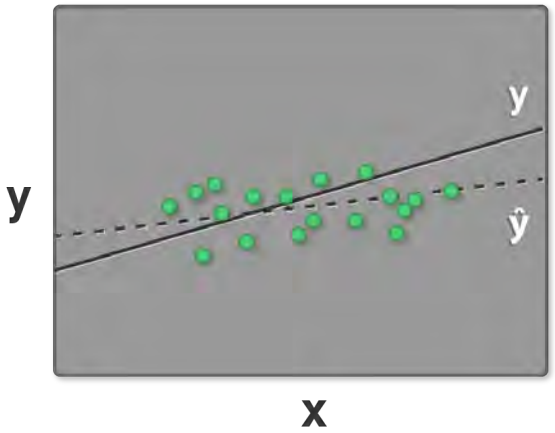

4 of 39



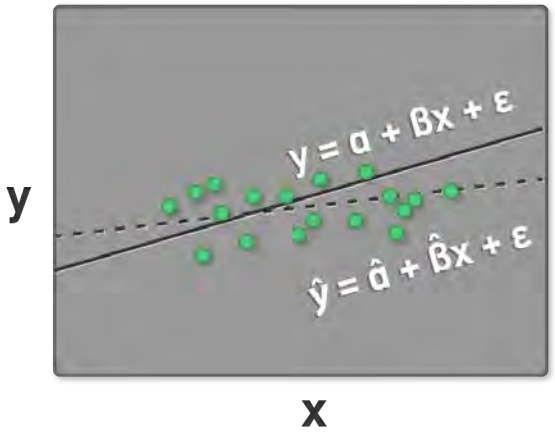

5 of 39




10 of 39




11 of 39




13 of 39

## Linear Regression Assumptions

1. Relaonship be  tween X and Y is linear.
2. If there is more than one X variable, they are not correlated.
3. The error terms (distance from data point to predicted line):
    - Have zero mean and constant variance (no heteroscedascity)
    - Are independent (no serial correlaon)
    - Are normally distributed

23 of 39

## Regression Measurements

- F-statistic: measures if we have confidence in the equaon
  - Should have p-value (likelihood of an error) of less than 0.05
- R-squared ($R^2$): measures percent of variaon in Y v  ariable explained by X variables
- Coefficient: measures relaonship be  tween X and Y, or measures intercept
- T-statistic: measures if we have confidence in the coefficient of a variable
  - Should have p-value of less than 0.05 (95% confidence)

34 of 39

## p-Values

- 90% confidence = 0.10 p-value
- 99% confidence = 0.01 p-value
- 95% confidence = 0.05 p-value

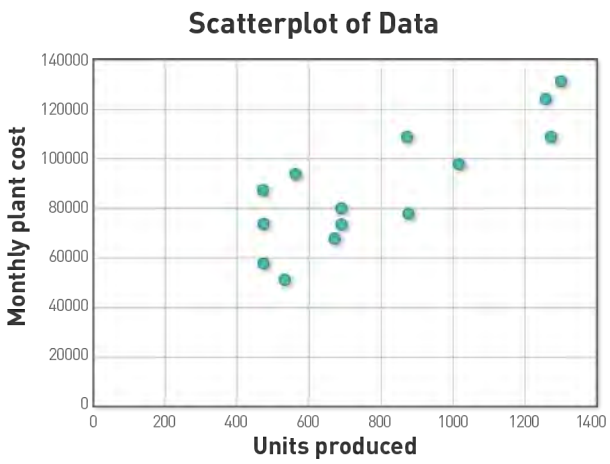
38 of 39

View All View Keyframes

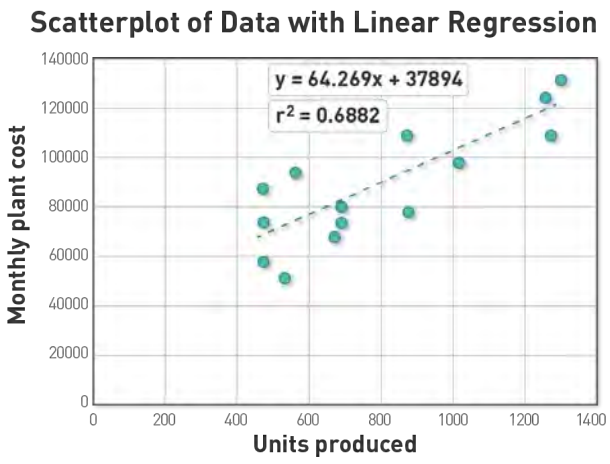
## Univariate Linear Regression

- This regression has only one X variable.
- Excel can quickly generate these results from the graphing package.
- Trend lines, equaons, and $R^2$ values can be added to sca erplots.
- This does not generate stas c al confidence levels.

6 of 21




7 of 21




11 of 21

## Fixed Costs and Variable Costs

- **Fixed costs:** expenses paid by factory independent of business acvity
- **Variable costs:** costs that vary depending on factory's producon v olume

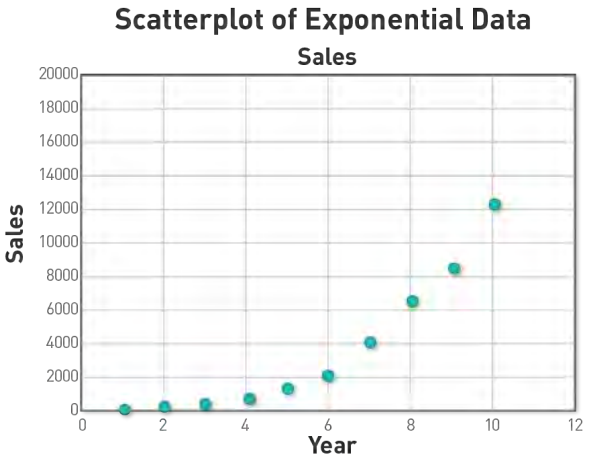
15 of 21

View All View Keyframes

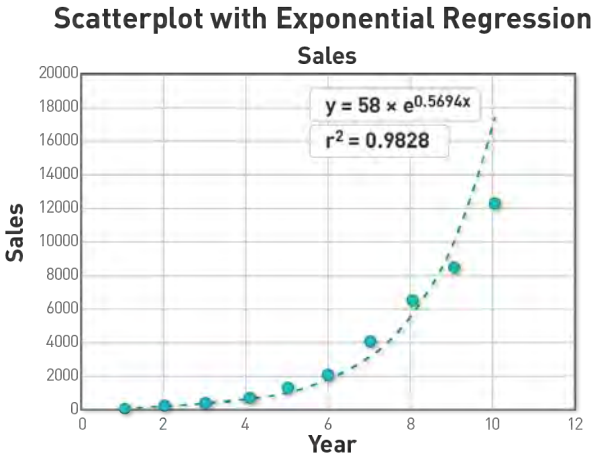
## Exponential Regression

- Not all data is linear.
- Nonlinear pa erns can be made evident by generang a sc a erplot.
- Exponenal gr owth is a form on nonlinear behavior.
  - E.g., Annual sales data for an organizaon

5 of 9




6 of 9




7 of 9

View All View Keyframes

## Multivariate Regression

- Mulple X v ariables
- No sca erplots
- Form of a mulv ariate equaon:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon$$

5 of 35

## Results of a Multivariate Regression

Regression Statistics

| | |
|---|---|
| Mulple R | 0.803398744 |
| $R^2$ | 0.645449542 |
| Adjusted $R^2$ | 0.57453945 |
| Standard Error | 1252.763898 |
| Observaons | 19 |

ANOVA

| | df | SS | MS | F | Significance F |
|---|---|---|---|---|---|
| Regression | 3 | 42856229.89 | 14285409.96 | 9.102365067 | 0.001126532 |
| Residual | 15 | 23541260.74 | 1569417.383 | | |
| Total | 18 | 66397490.63 | | | |

| | Coefficients | Standard Error | t Stat | P-value |
|---|---|---|---|---|
| Intercept | 35102.90045 | 1837.226911 | 19.10645889 | 6.11198E−12 |
| A Made | 2.065953296 | 1.664981779 | 1.240826369 | 0.23372682 |
| B Made | 4.176355531 | 1.681252566 | 2.484073849 | 0.025287785 |
| C Made | 4.790641037 | 1.789316107 | 2.677358695 | 0.017222643 |

11 of 35

## General Conclusions

- We have a significant equaon.
- It explains 64–65% of change in producon c osts.
- We have confidence in our intercept (fixed cost).
- We have confidence in variable costs for B and C.

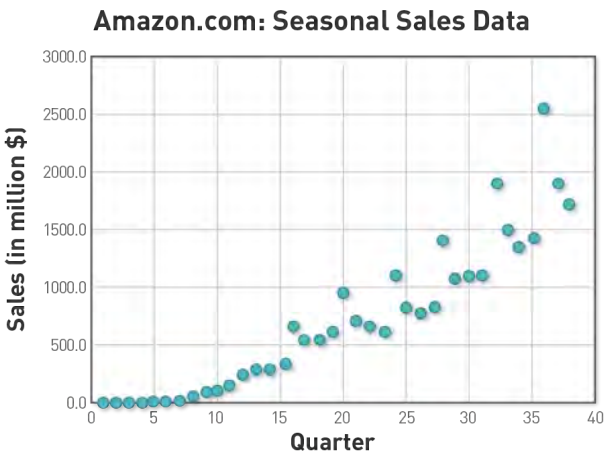
35 of 35

View All View Keyframes

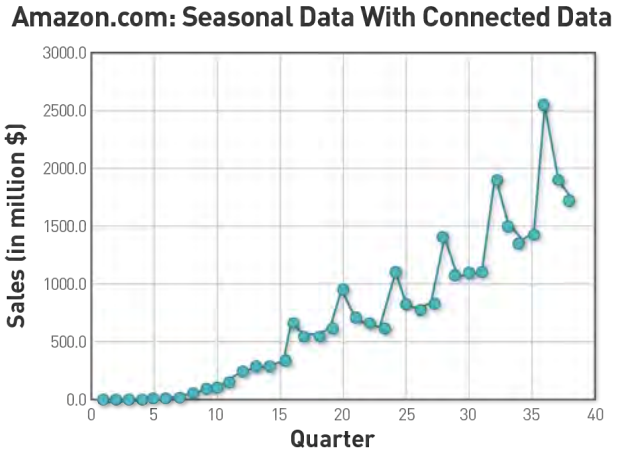
## Simple Time Series: Moving Average

- **Seasonality:** when data experiences regular changes that occur every year
- Examples:
  - Holiday sales in the retail industry
  - Home sales during summer
  - Gasoline sales during summer vacaons
- **Moving average:** reduces effect of seasonality so true pa erns can be seen

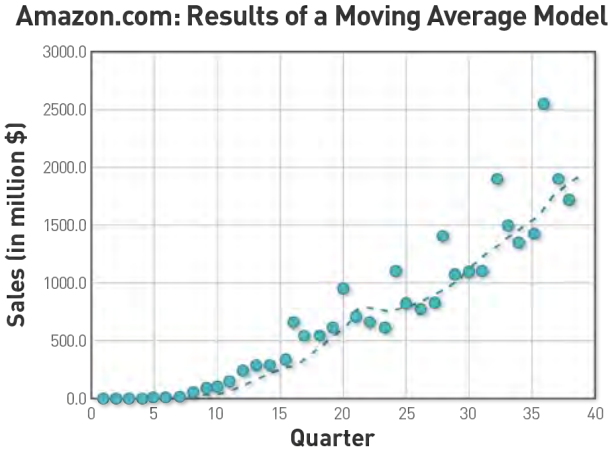
7 of 13




8 of 13




10 of 13



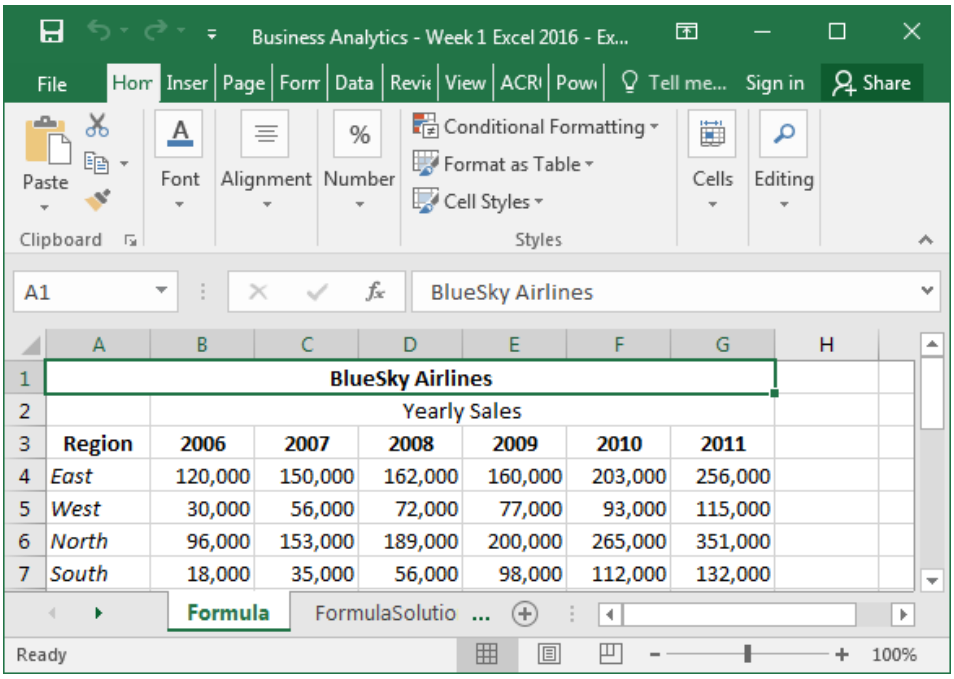
# Microsoft Excel 2016: Basics

## Microsoft Excel

1. Go to the Course Website
2. Download and print “Business Analytics Week 1 Excel 2016 Instructions.doc”
3. Download “Business Analytics - Week 1 Excel 2016.xlsx”

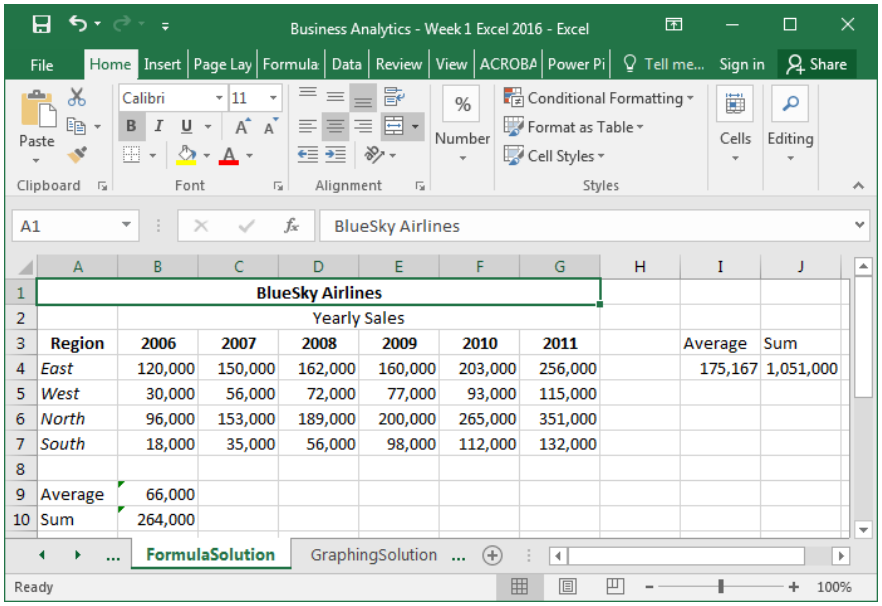
## Session 1.7: Calculations and Formulas

1. Highlight the four sales values for 2006. Notice at the bottom of the page, it calculates the average, count and sum of the values.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | BlueSky Airlines | | | | | | |
| 2 | | Yearly Sales | | | | | |
| 3 | **Region** | **2006** | **2007** | **2008** | **2009** | **2010** | **2011** |
| 4 | *East* | 120,000 | 150,000 | 162,000 | 160,000 | 203,000 | 256,000 |
| 5 | *West* | 30,000 | 56,000 | 72,000 | 77,000 | 93,000 | 115,000 |
| 6 | *North* | 96,000 | 153,000 | 189,000 | 200,000 | 265,000 | 351,000 |
| 7 | *South* | 18,000 | 35,000 | 56,000 | 98,000 | 112,000 | 132,000 |

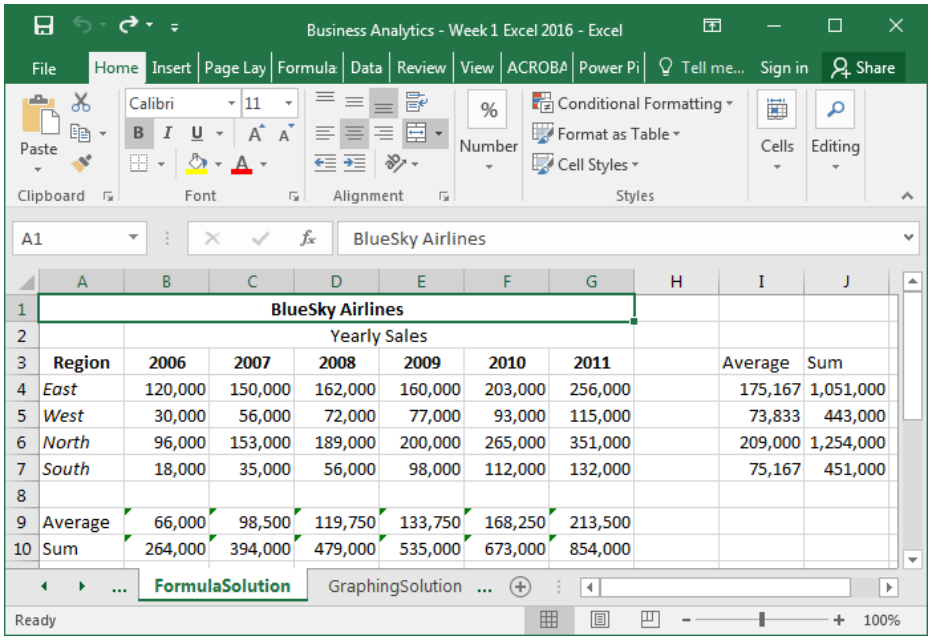
2. Next, we’ll enter a formula in the spreadsheet to automatically calculate averages and sums
3. In cell A9, type the word Average; also enter Average in I3
4. In cell A10, type the word Sum; also enter Sum in J3
5. To calculate a formula, use the equal sign, the name of the formula, and the data range
    a. In cell B9, enter =average(b4:b7)
    b. In cell B10, enter =sum(b4:b7)
    c. In cell I4, enter =average(b4:g4)
    d. In cell J4, enter =sum(b4:g4)
6. Notice that after you type the equals sign and part of the formula name, it gives you several options that start with the same spelling; you can click on the one you want, then highlight the range of data




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BlueSky Airlines | | | | | | | | | |
| 2 | | Yearly Sales | | | | | | | | |
| 3 | **Region** | **2006** | **2007** | **2008** | **2009** | **2010** | **2011** | | Average | Sum |
| 4 | *East* | 120,000 | 150,000 | 162,000 | 160,000 | 203,000 | 256,000 | | 175,167 | 1,051,000 |
| 5 | *West* | 30,000 | 56,000 | 72,000 | 77,000 | 93,000 | 115,000 | | | |
| 6 | *North* | 96,000 | 153,000 | 189,000 | 200,000 | 265,000 | 351,000 | | | |
| 7 | *South* | 18,000 | 35,000 | 56,000 | 98,000 | 112,000 | 132,000 | | | |
| 8 | | | | | | | | | | |
| 9 | Average | 66,000 | | | | | | | | |
| 10 | Sum | 264,000 | | | | | | | | |

## Copy and Paste

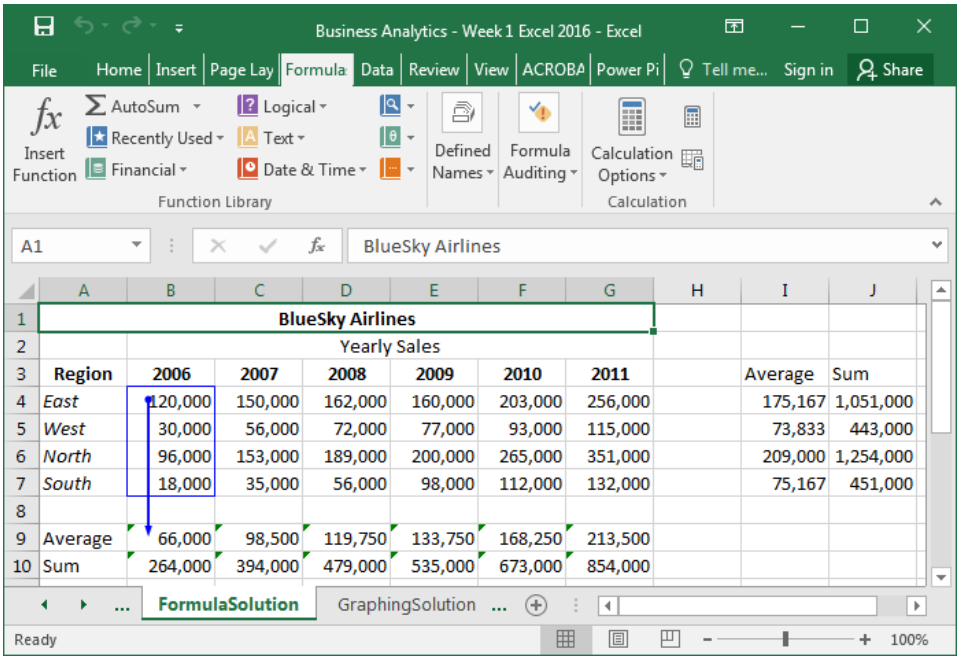
1. Copy and paste works the same in Excel as in Microsoft Word, but you can also copy and paste formulas and Excel will automatically update formulas
2. Copy the average and sum formulas across all columns and rows
    a. Click on cell B9, click on the copy button in the Clipboard section (or use control-C), highlight cells C9 to G9, and click the paste button in Clipboard
    b. Copy cell B10 across for the columns in the same way
    c. You can copy several formulas at once. Highlight cells I4 and J4. Click on copy, then highlight the range I5 to J7, and click paste.
    d. Look at the formulas in B9 and C9. How are they different?




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BlueSky Airlines | | | | | | | | | |
| 2 | | Yearly Sales | | | | | | | | |
| 3 | **Region** | **2006** | **2007** | **2008** | **2009** | **2010** | **2011** | | Average | Sum |
| 4 | *East* | 120,000 | 150,000 | 162,000 | 160,000 | 203,000 | 256,000 | | 175,167 | 1,051,000 |
| 5 | *West* | 30,000 | 56,000 | 72,000 | 77,000 | 93,000 | 115,000 | | 73,833 | 443,000 |
| 6 | *North* | 96,000 | 153,000 | 189,000 | 200,000 | 265,000 | 351,000 | | 209,000 | 1,254,000 |
| 7 | *South* | 18,000 | 35,000 | 56,000 | 98,000 | 112,000 | 132,000 | | 75,167 | 451,000 |
| 8 | | | | | | | | | | |
| 9 | Average | 66,000 | 98,500 | 119,750 | 133,750 | 168,250 | 213,500 | | | |
| 10 | Sum | 264,000 | 394,000 | 479,000 | 535,000 | 673,000 | 854,000 | | | |

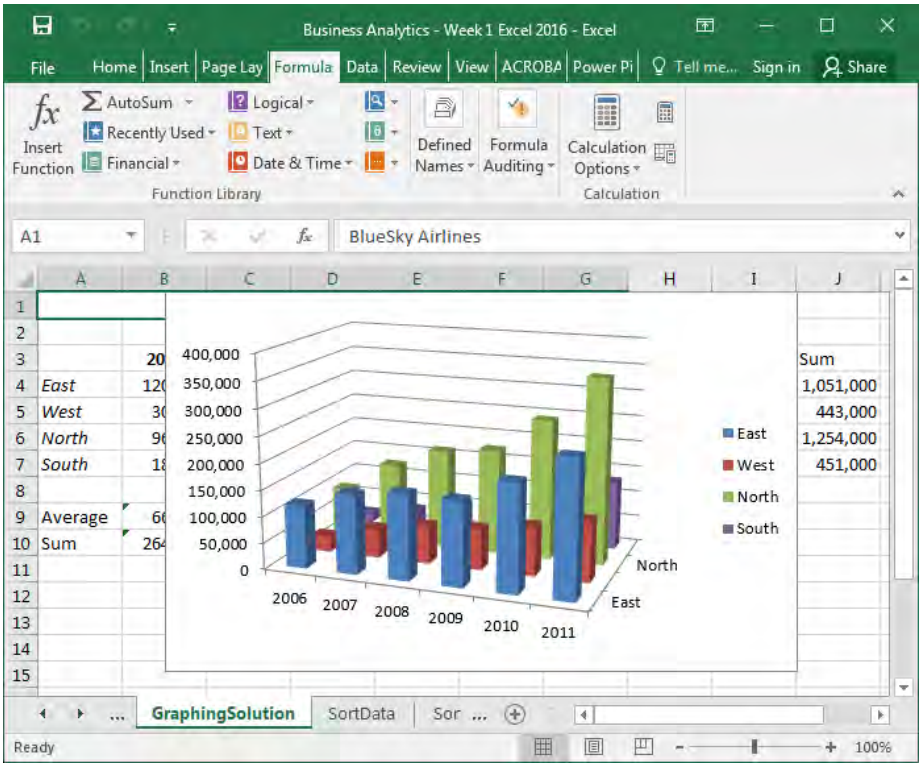
## Checking Formulas

1. Excel has an auditing feature to check formulas; click on the tab labeled Formulas
2. Click on B9, then click on Trace Precedents in Formula Auditing
3. To clear the arrows, click on Remove Arrows



## Session 1.8: Graphing Data

1. Sometimes you might want to graph your data. Excel can graph columns and rows of data, but has difficulty if anything is in the upper left corner. <u>First, delete the word Region by clicking on cell A3 and pressing the delete key.</u>
2. Highlight the data by clicking on the cell labeled A3 through G7.
3. Next click on the Insert tab at the top of the screen.
4. In the section labeled Charts, click on the tiny icon in the lower right corner of charts to bring up the possible chart options.
5. Move the cursor over each options and select the 3-D column chart, click OK




6. Let's move the chart to another sheet. Click on the upper right corner of Excel, select Move Chart Location, click on New Sheet, then OK.
7. At the top of the screen click on Chart Tools: Design: Type, Change Chart Type to change the type of chart
8. In the Chart Tools: Design: Data, click on Switch Row/Column. What happens?
9. In the Chart Tools: Design: Chart Layouts, Quick Layout, there are several options (scroll down). How are they different?
10. Next, try Chart Tools: Layout. Experiment with Labels and Axes. Under 3-D Rotation, change the X and Y rotation.

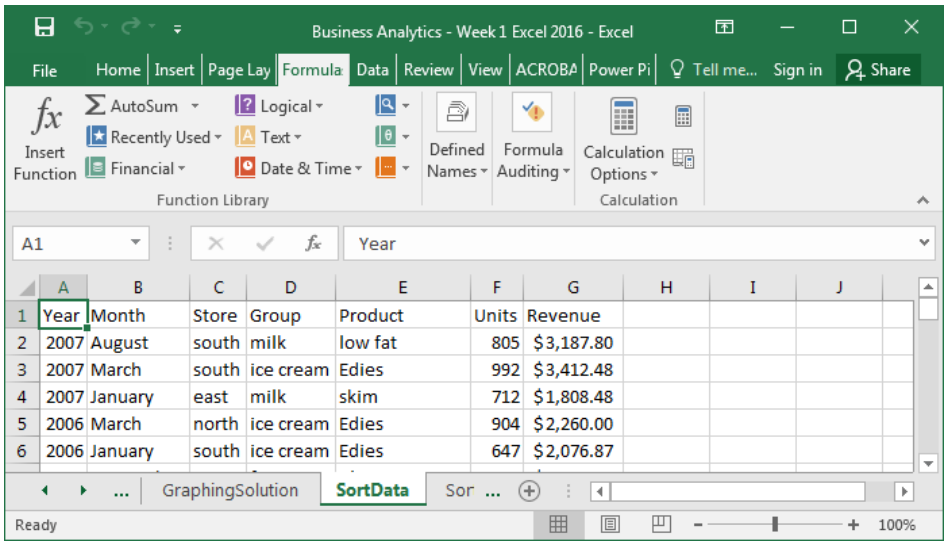
## Saving your Spreadsheet

To save a spreadsheet, click on the File tab in the upper left corner of the screen, click on Save As. You'll notice that you can save your Spreadsheet in a variety of formats.

# Microsoft Excel: Sorts, Filters, Pivot Tables

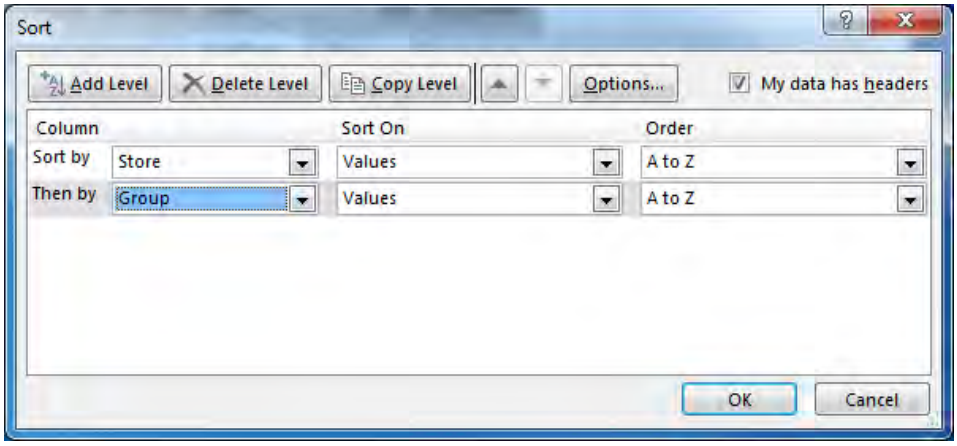
## Session 1.9: Sorting Data

1. Use the SortData spreadsheet for this exercise.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Month | Store | Group | Product | Units | Revenue |
| 2 | 2007 | August | south | milk | low fat | 805 | $3,187.80 |
| 3 | 2007 | March | south | ice cream | Edies | 992 | $3,412.48 |
| 4 | 2007 | January | east | milk | skim | 712 | $1,808.48 |
| 5 | 2006 | March | north | ice cream | Edies | 904 | $2,260.00 |
| 6 | 2006 | January | south | ice cream | Edies | 647 | $2,076.87 |

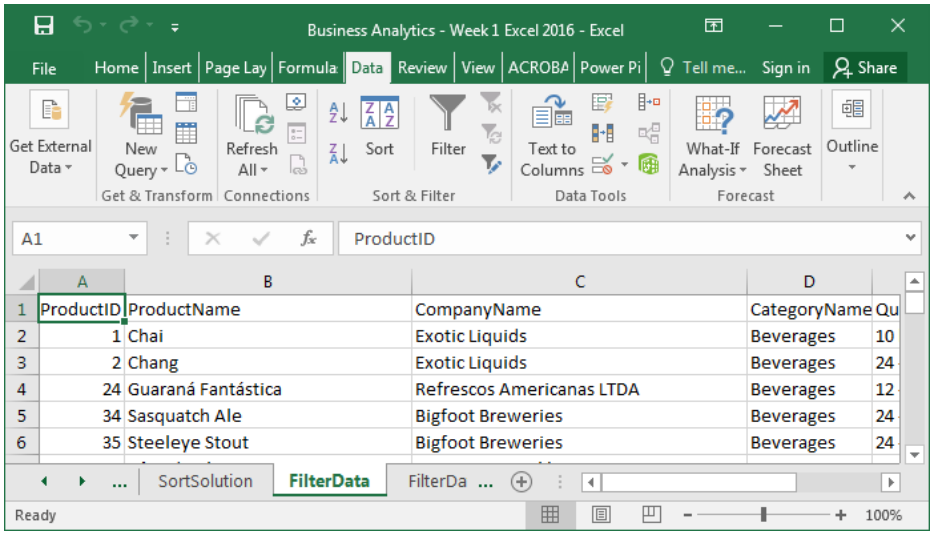
2. To sort data, highlight the top of the columns (A through G). The entire columns should be highlighted.
3. Click on the Data tab at the top of the spreadsheet.
4. Click on Sort; note that “My data has headers” is checked. Why is it checked?
5. Select the first sort criteria: Store
6. Click on Add Level, then add the second sort criteria: Group



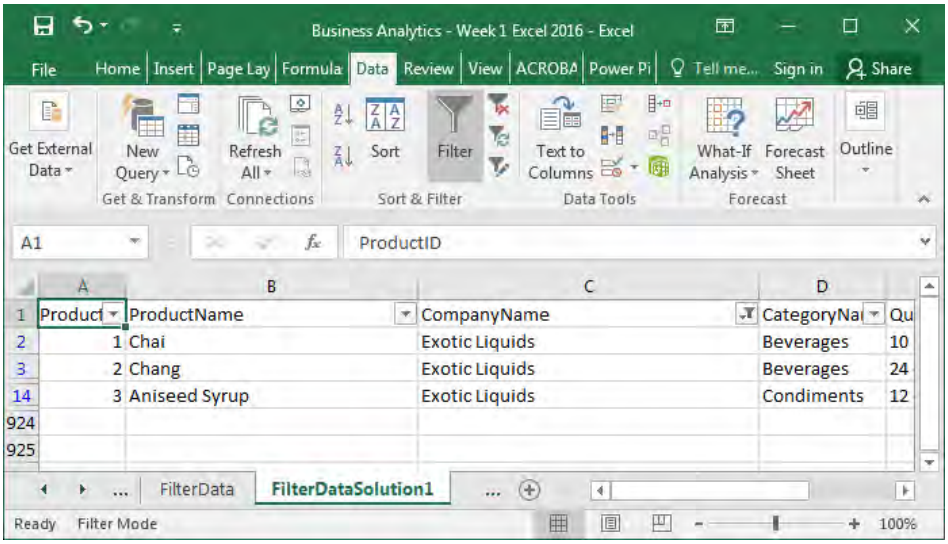

7. Click OK, and the data is sorted
8. How would you sort from Z to A (reverse alphabetical order)? Why would you want to sort in a different order?

## Filters

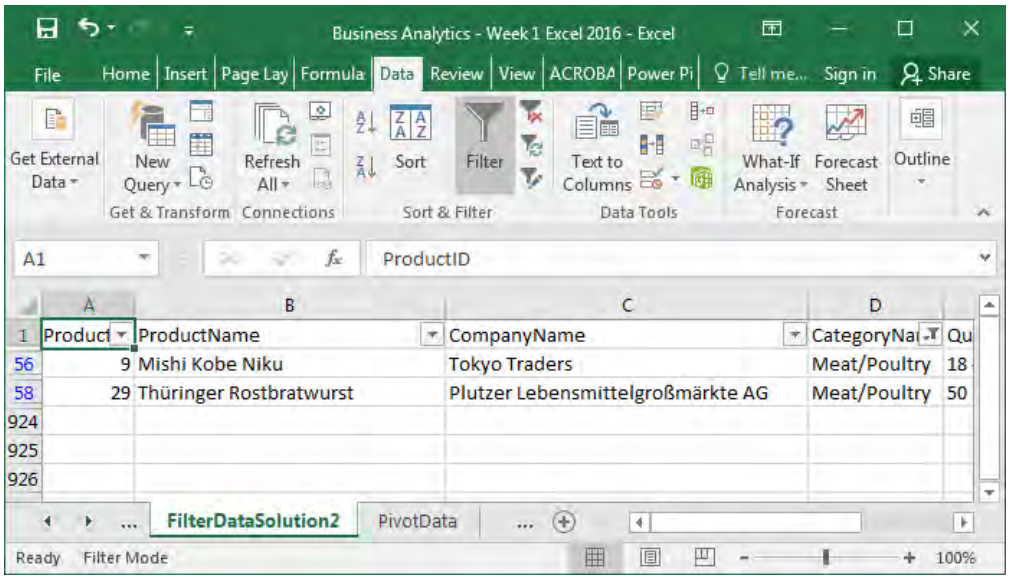
1. Use the FilterData spreadsheet for this exercise
2. A filter allows you to select specific data out of a long list
3. First, click on cell A1, then click on the Data tab, then Filter (looks like a funnel)




4. Notice that there are drop down arrows for each column. Let's find all products which Exotic Liquids produces.
5. Click on the drop down arrow next to company name; uncheck "(Select All)", then check Exotic Liquids and OK. What happens?



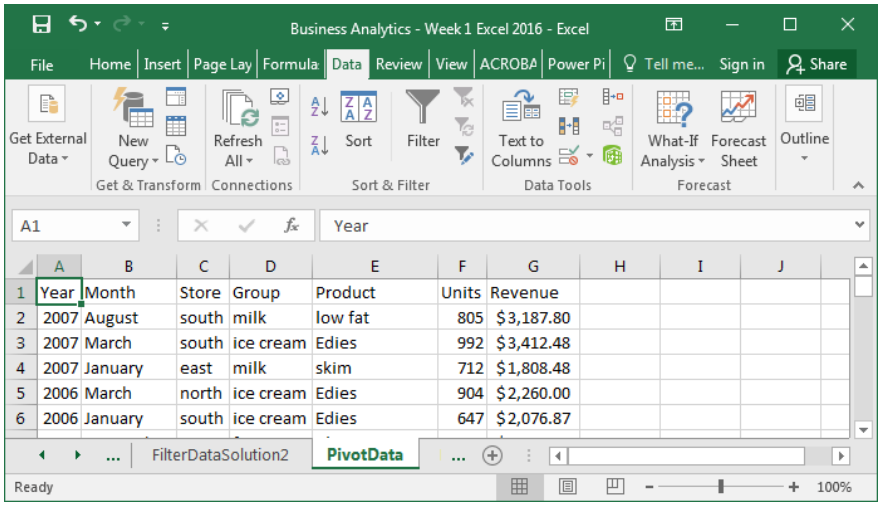

6. Go back and click Select All for Company Name
7. To find which products have a price greater than 50, click on the down arrow next to unit price, Number Filters, Greater Than, enter 50 in the field, and click OK
8. You can turn on multiple filters. Add a filter for Category Name equal to Meat/Poultry to see which Meat/Poultry products cost more than 50



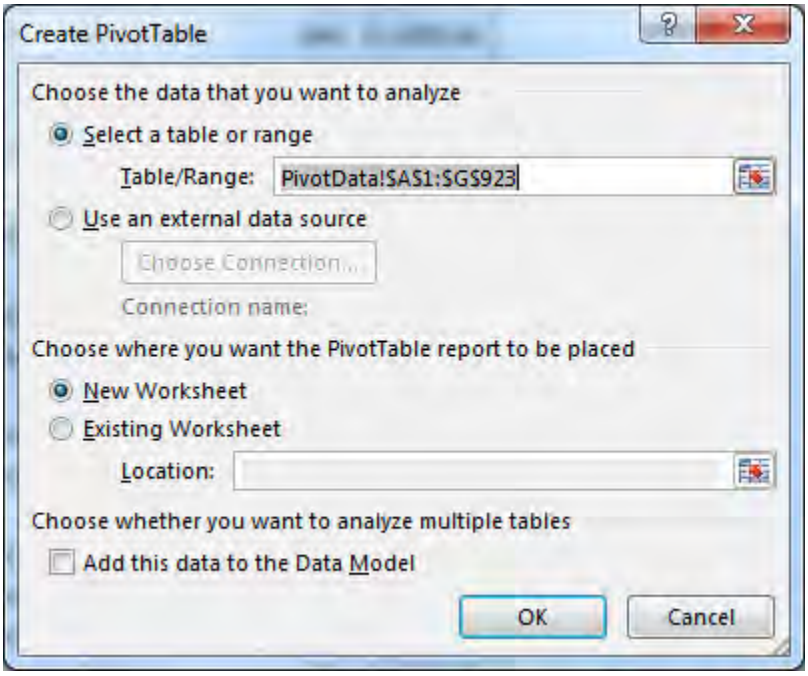

9. To turn off the filter, click on the Filter icon and the drop down arrows should disappear.

## Session 1.10: Pivot Tables and Charts

Pivot tables and charts are powerful techniques to quickly summarize and display large amounts of data. For this example, use the PivotData spreadsheet.

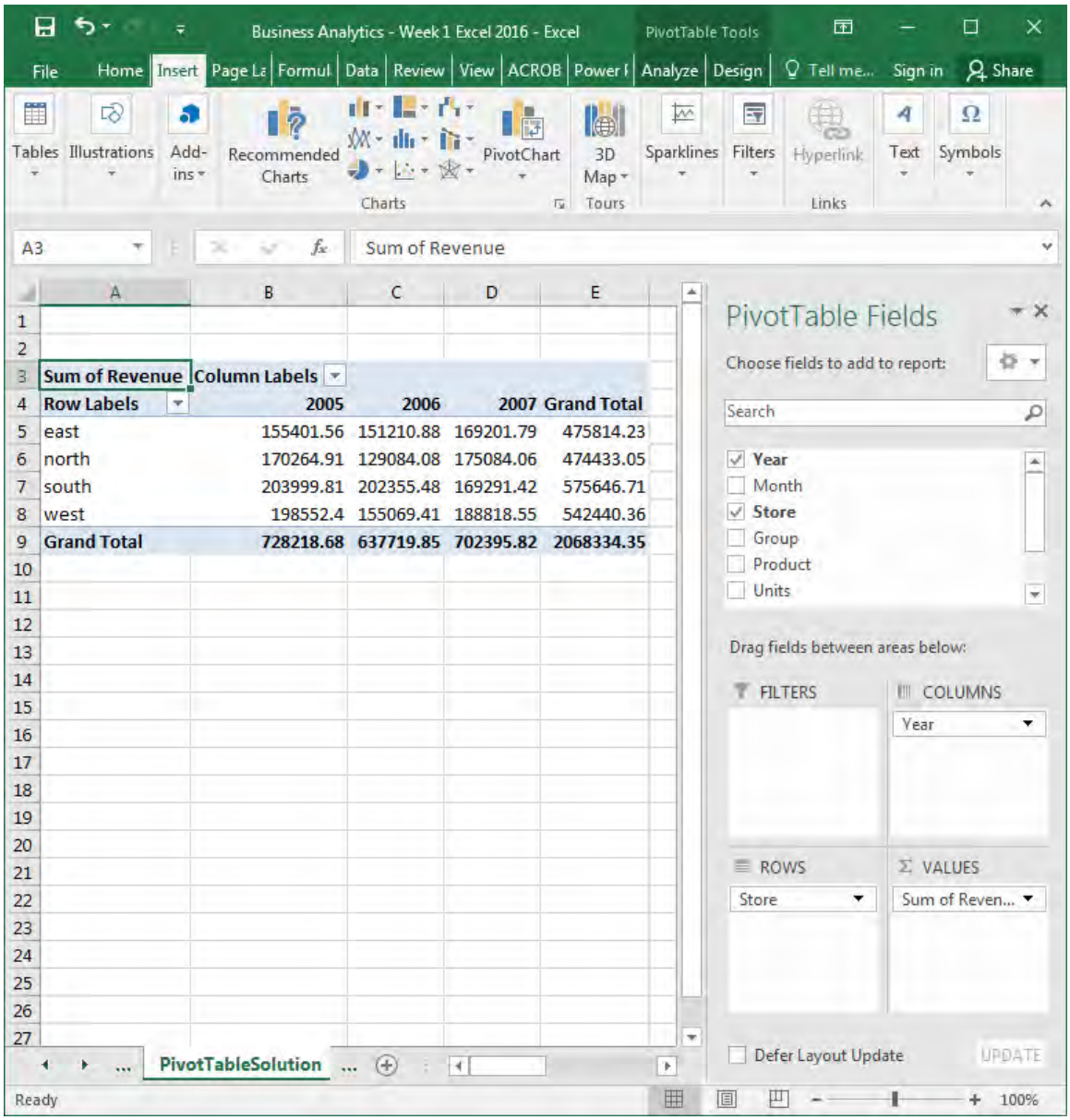



1. The headings must be in the first row of your data.
2. Click anywhere in the data table.
3. Click on the Insert tab.
4. Click on PivotTable; it should automatically highlight the entire table.
5. Make sure New Worksheet is checked, then click OK.

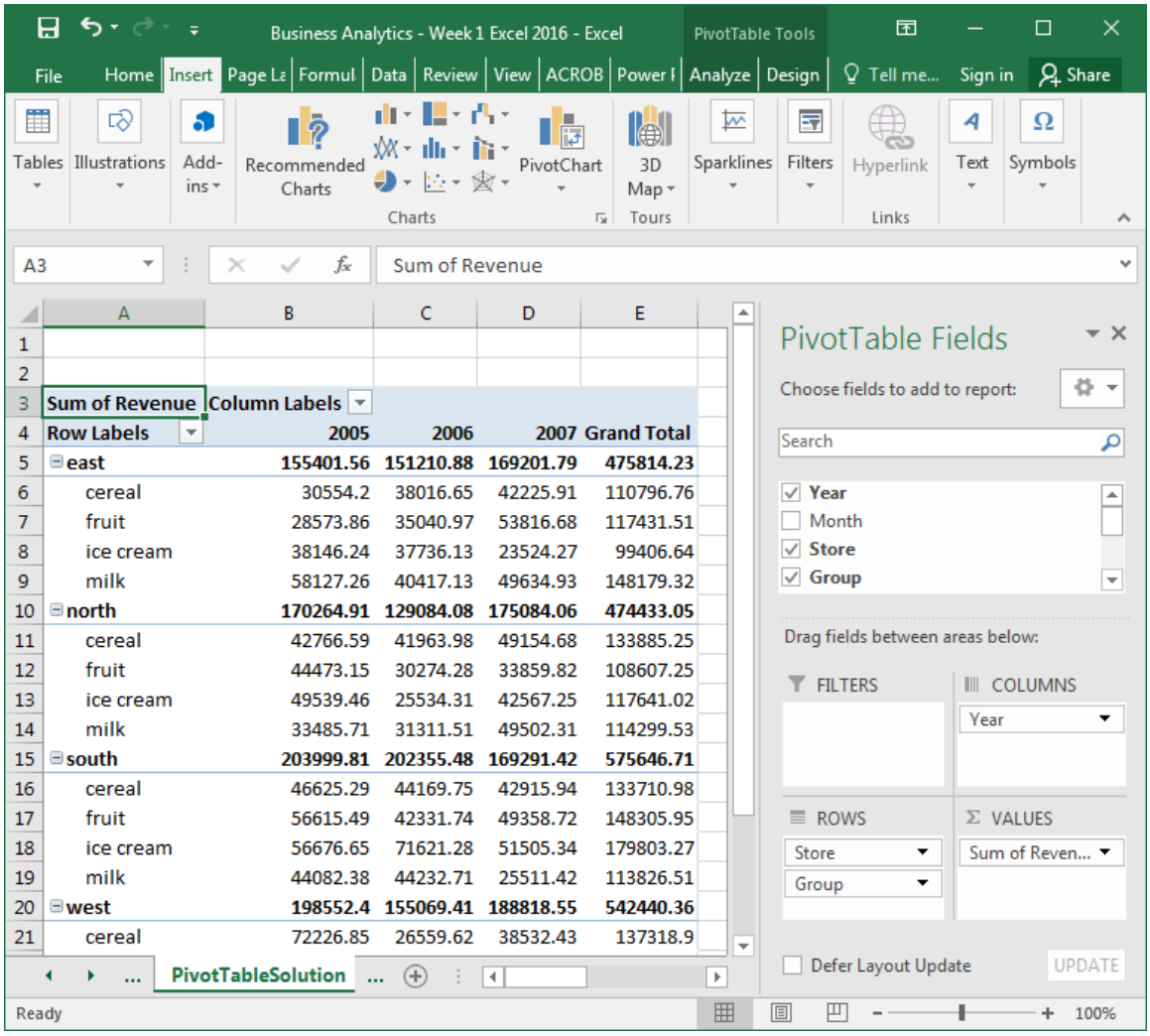



6. In the pivot table dialog box, we want to add row labels, column labels, and identify what goes into the table.

7. For this example, let's build a table with store revenue by year. First click on store and drag it to the row label. These will become the row labels on the left side of your table.
8. Next, click on year and drag it to the column label. These will become the column labels at the top of your table.
9. Finally, click on revenue and drag it to the values cell. These values will feed the interior of the table. The default function is Sum. If you want to change this function, click on the down arrow in the values box, then select the function (average, min, max, etc.)



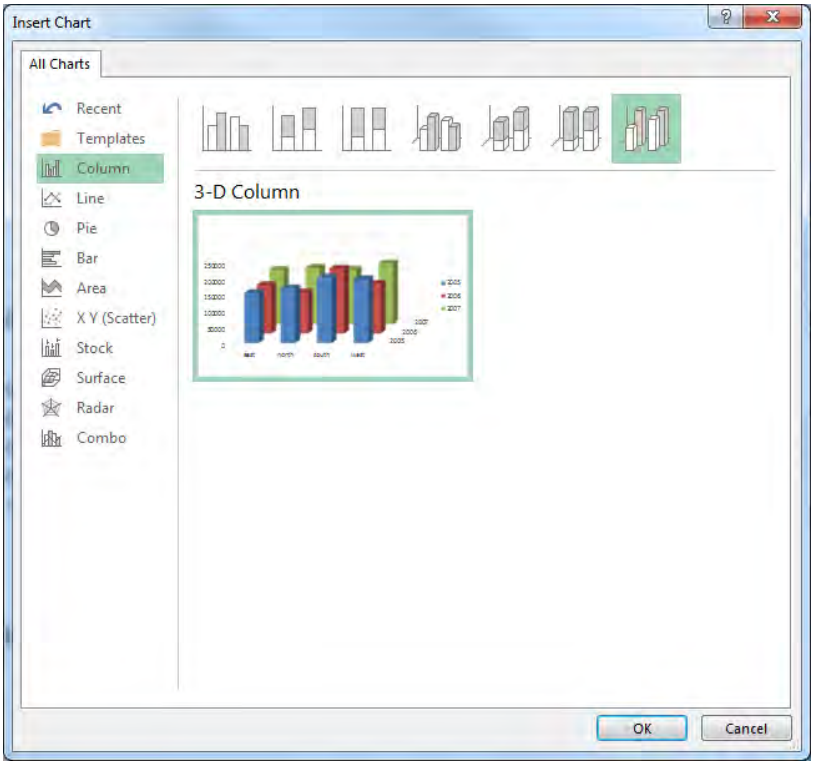
Add group to the row labels to see what happens.

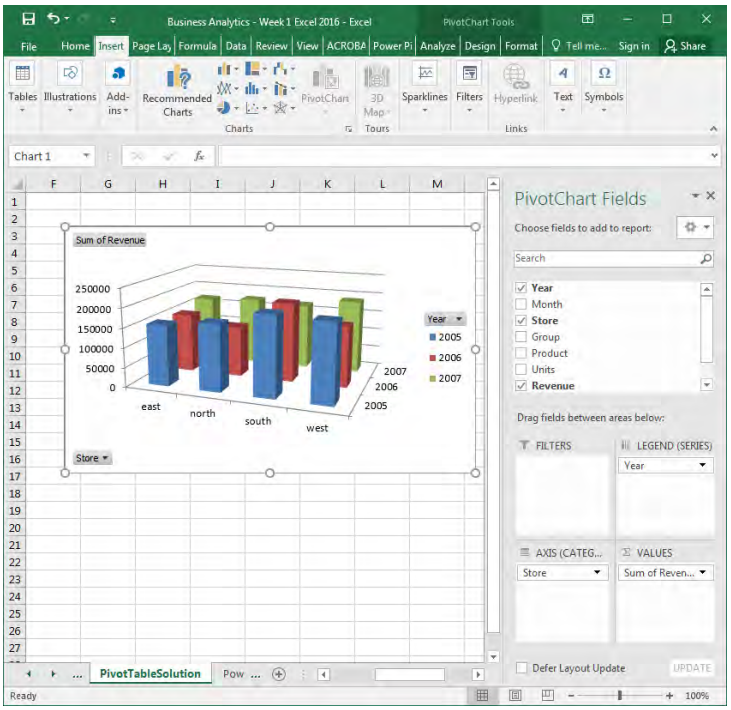



| Sum of Revenue | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | 2005 | 2006 | 2007 | Grand Total |
| east | 155401.56 | 151210.88 | 169201.79 | 475814.23 |
| cereal | 30554.2 | 38016.65 | 42225.91 | 110796.76 |
| fruit | 28573.86 | 35040.97 | 53816.68 | 117431.51 |
| ice cream | 38146.24 | 37736.13 | 23524.27 | 99406.64 |
| milk | 58127.26 | 40417.13 | 49634.93 | 148179.32 |
| north | 170264.91 | 129084.08 | 175084.06 | 474433.05 |
| cereal | 42766.59 | 41963.98 | 49154.68 | 133885.25 |
| fruit | 44473.15 | 30274.28 | 33859.82 | 108607.25 |
| ice cream | 49539.46 | 25534.31 | 42567.25 | 117641.02 |
| milk | 33485.71 | 31311.51 | 49502.31 | 114299.53 |
| south | 203999.81 | 202355.48 | 169291.42 | 575646.71 |
| cereal | 46625.29 | 44169.75 | 42915.94 | 133710.98 |
| fruit | 56615.49 | 42331.74 | 49358.72 | 148305.95 |
| ice cream | 56676.65 | 71621.28 | 51505.34 | 179803.27 |
| milk | 44082.38 | 44232.71 | 25511.42 | 113826.51 |
| west | 198552.4 | 155069.41 | 188818.55 | 542440.36 |
| cereal | 72226.85 | 26559.62 | 38532.43 | 137318.9 |

Let's now create a Pivot Chart. The steps are:

1. Uncheck Group to simplify the data.
2. Click anywhere inside your Pivot Table. An Options tab will appear. Click on Pivot Chart.
3. A chart dialog box will appear. Click on your preferred type of chart.




To move the chart to its own page:

1. Right click on the picture
2. Click on Move Chart
3. Select New Sheet and name the new sheet

# Microsoft Excel: PowerView

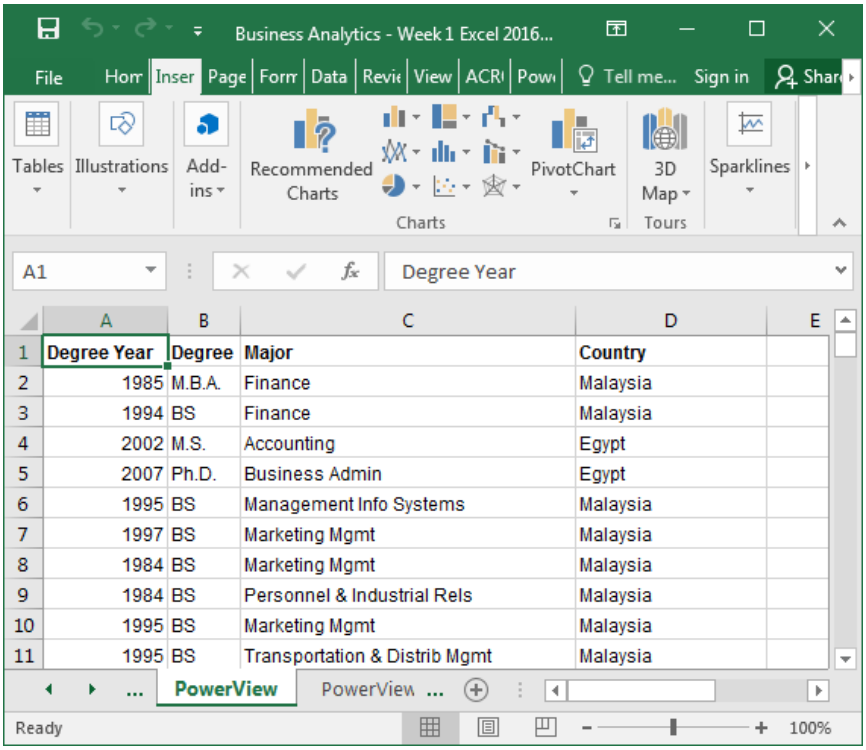
**Session 1.11: Powerview**

**PowerView Add-in**

The PowerView options are available as an add-in to Excel 2016. The steps to add it are:

1. In Excel 2016, click on the File tab, then Options, then Customize Ribbon
2. Under Main Tabs, click on Insert, then click on New Group
3. Under Choose commands from, click Commands Not in the Ribbon, then select Insert a Power View Report
4. With both Insert a Power View Report and New Group (custom) selected, click Add, then move the New Group (custom) to where you want it on the ribbon
5. Click New Group (custom), then Rename, then in the Display Name box, type Reports or the name you want
6. Click OK twice
7. The first time you insert a Power View sheet, you might be prompted to enable the Power View add-in. Click Enable.

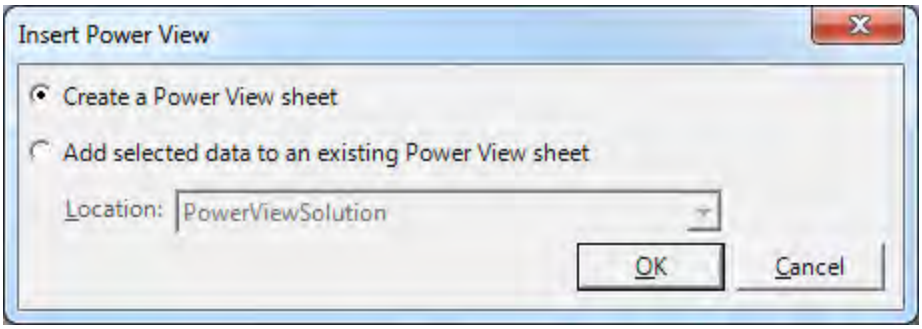
PowerView allows you to quickly generate business analytics graphs of complex data. PowerView is only available in Microsoft Excel 2013. Click on the PowerView spreadsheet for this exercise. This data is a list of international alumni from the Whitman School.

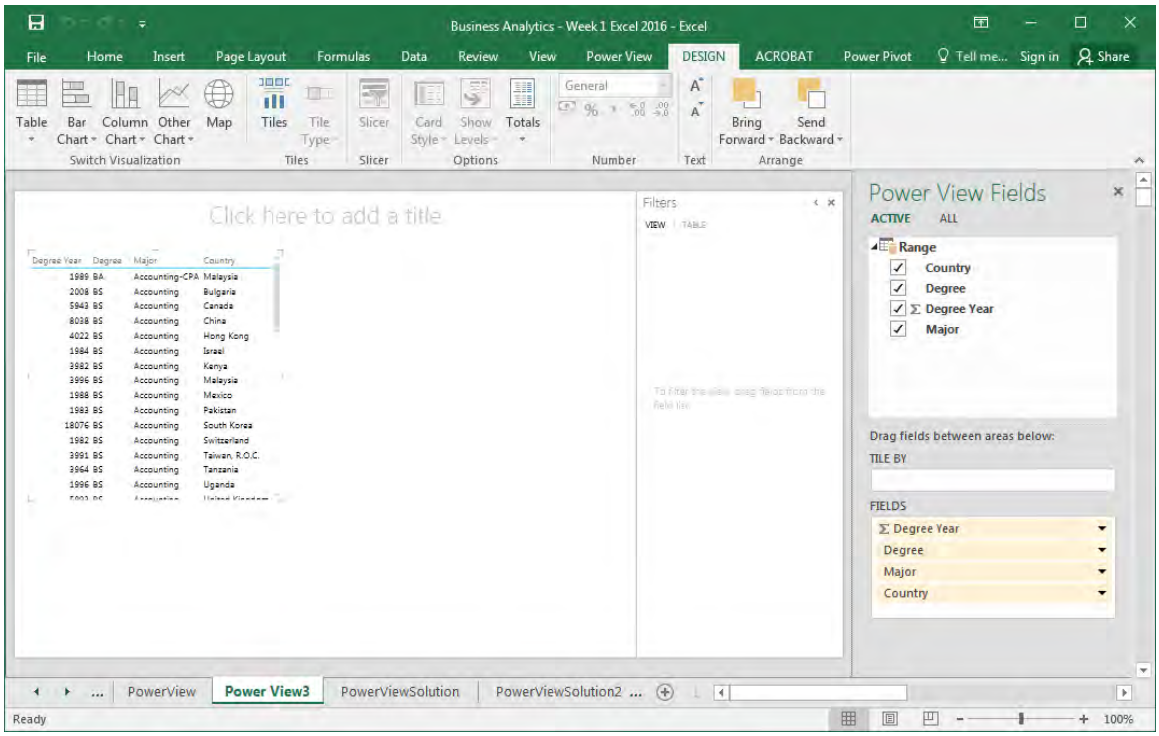



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Degree Year | Degree | Major | Country |
| 2 | 1985 | M.B.A. | Finance | Malaysia |
| 3 | 1994 | BS | Finance | Malaysia |
| 4 | 2002 | M.S. | Accounting | Egypt |
| 5 | 2007 | Ph.D. | Business Admin | Egypt |
| 6 | 1995 | BS | Management Info Systems | Malaysia |
| 7 | 1997 | BS | Marketing Mgmt | Malaysia |
| 8 | 1984 | BS | Marketing Mgmt | Malaysia |
| 9 | 1984 | BS | Personnel & Industrial Rels | Malaysia |
| 10 | 1995 | BS | Marketing Mgmt | Malaysia |
| 11 | 1995 | BS | Transportation & Distrib Mgmt | Malaysia |

To generate a PowerView report:

1. Click on the Insert tab at the top of the screen
2. In the Reports section, click on Power View
3. In the Insert Power View screen, click OK.

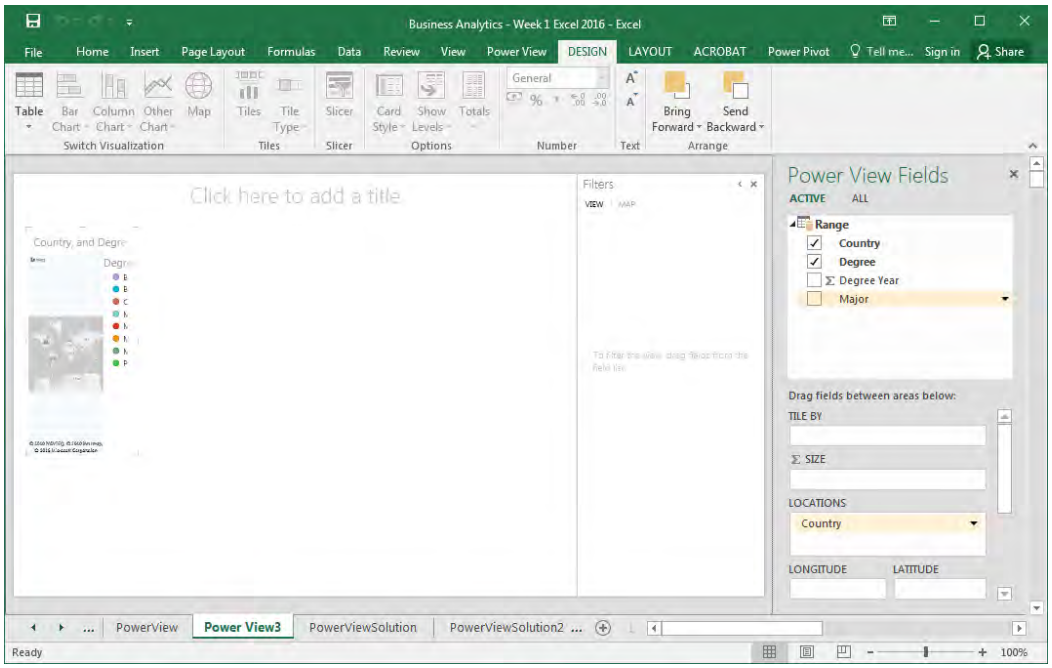



On the left side of the screen is your data. On the right side are the data fields or column labels (country, degree, degree year, major).

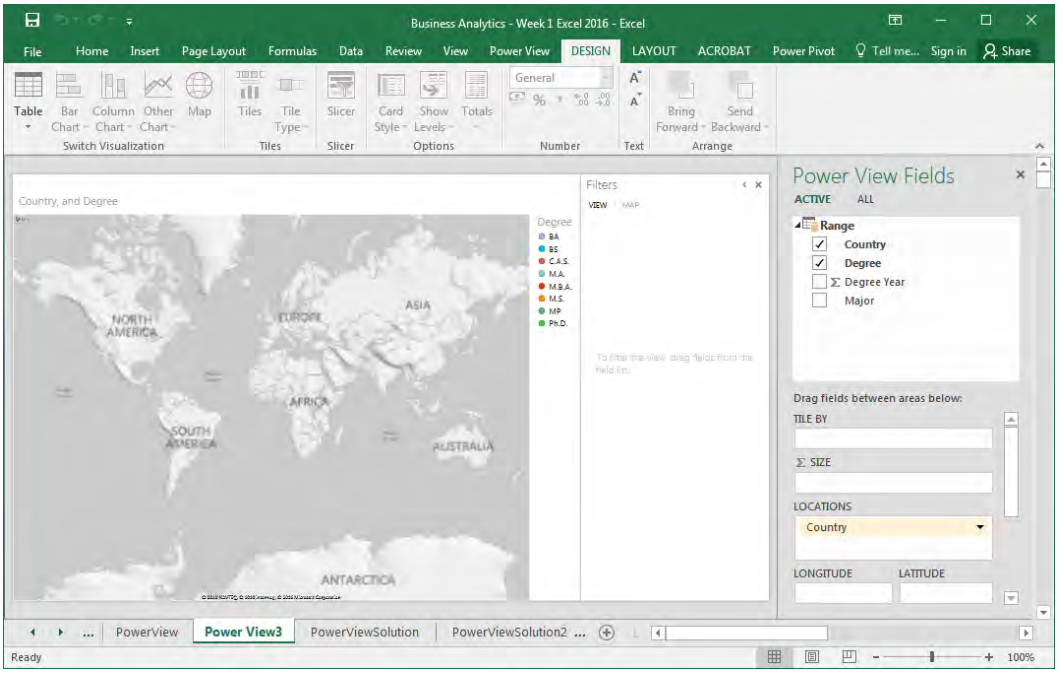
To create a map of degrees by country:

1. Uncheck Degree Year; uncheck Major
2. Click on the Map icon in the upper left portion of your screen



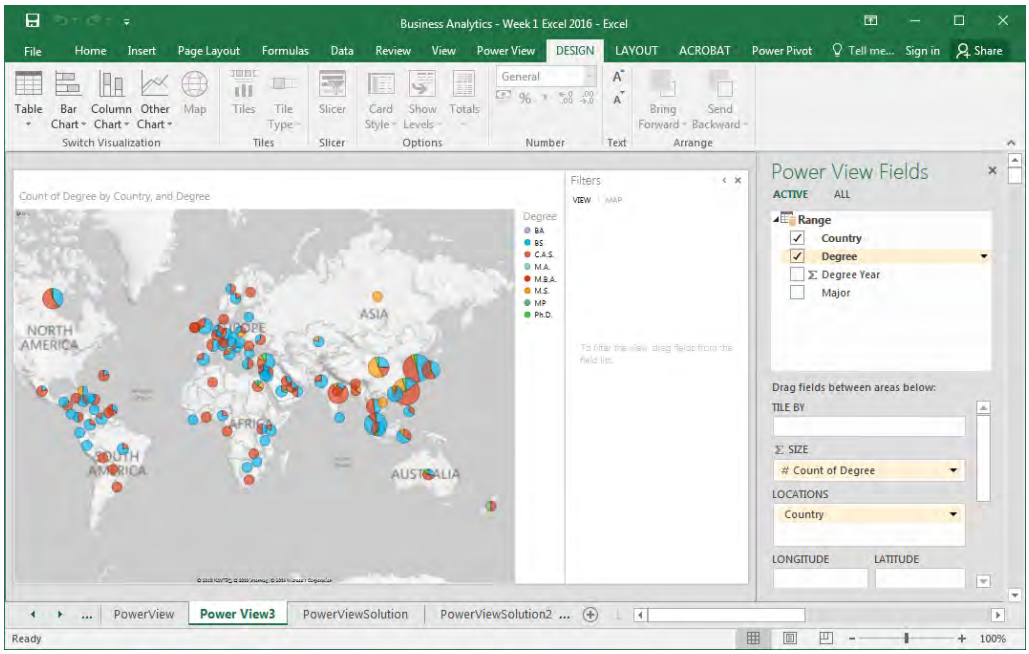

The map on the left is not very clear.

1. Move your cursor over the map.
2. An arrow in the upper right of the picture should appear.
3. Move the cursor over the arrow; it should say Pop out.
4. Click on the arrow.



Now we have a picture, but no data. Notice that in the field labeled ∑ SIZE has no field entered. This field determines the data to be displayed.

1. Drag Degree from Power View Fields to ∑ SIZE.



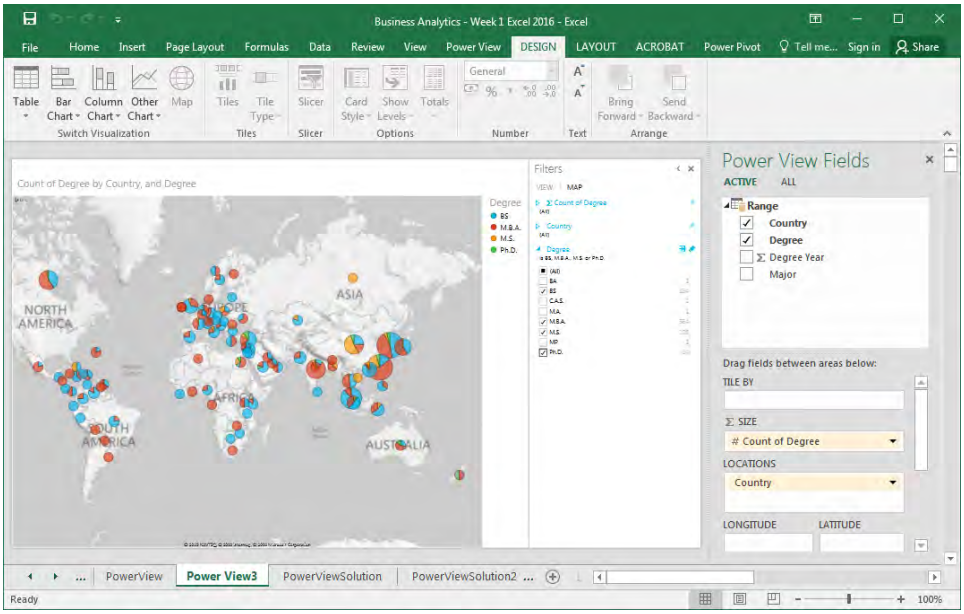

Be sure to have the following on the right side of the screen:

1. Drag Degree to ∑ SIZE; it should show # Count of Degree
2. LOCATIONS should have Country
3. COLOR should have Degree

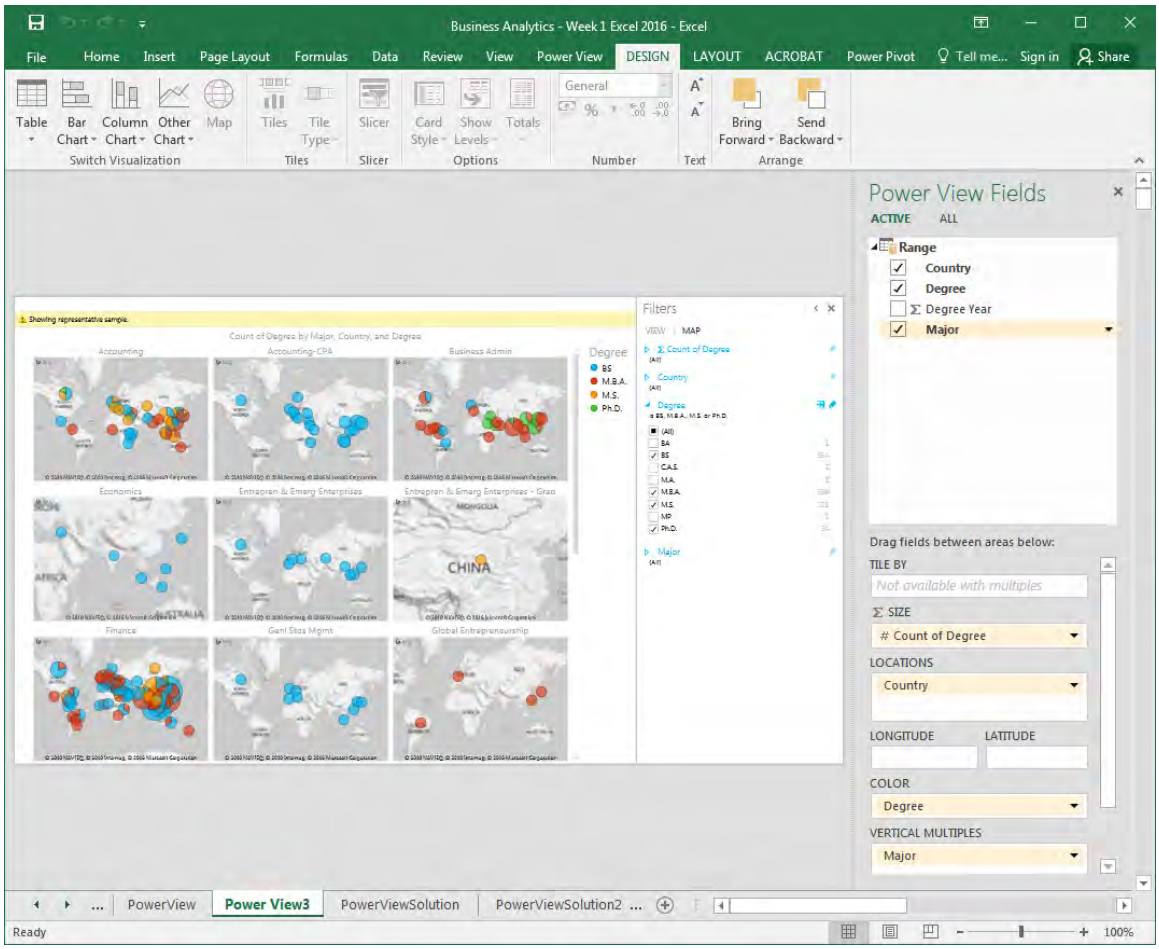
Put your cursor over a country's pie chart. By moving the cursor over different parts of the pie chart, it will give you the data associated with each pie segment.

You can also filter data to limit the data displayed on the screen. To activate a filter:

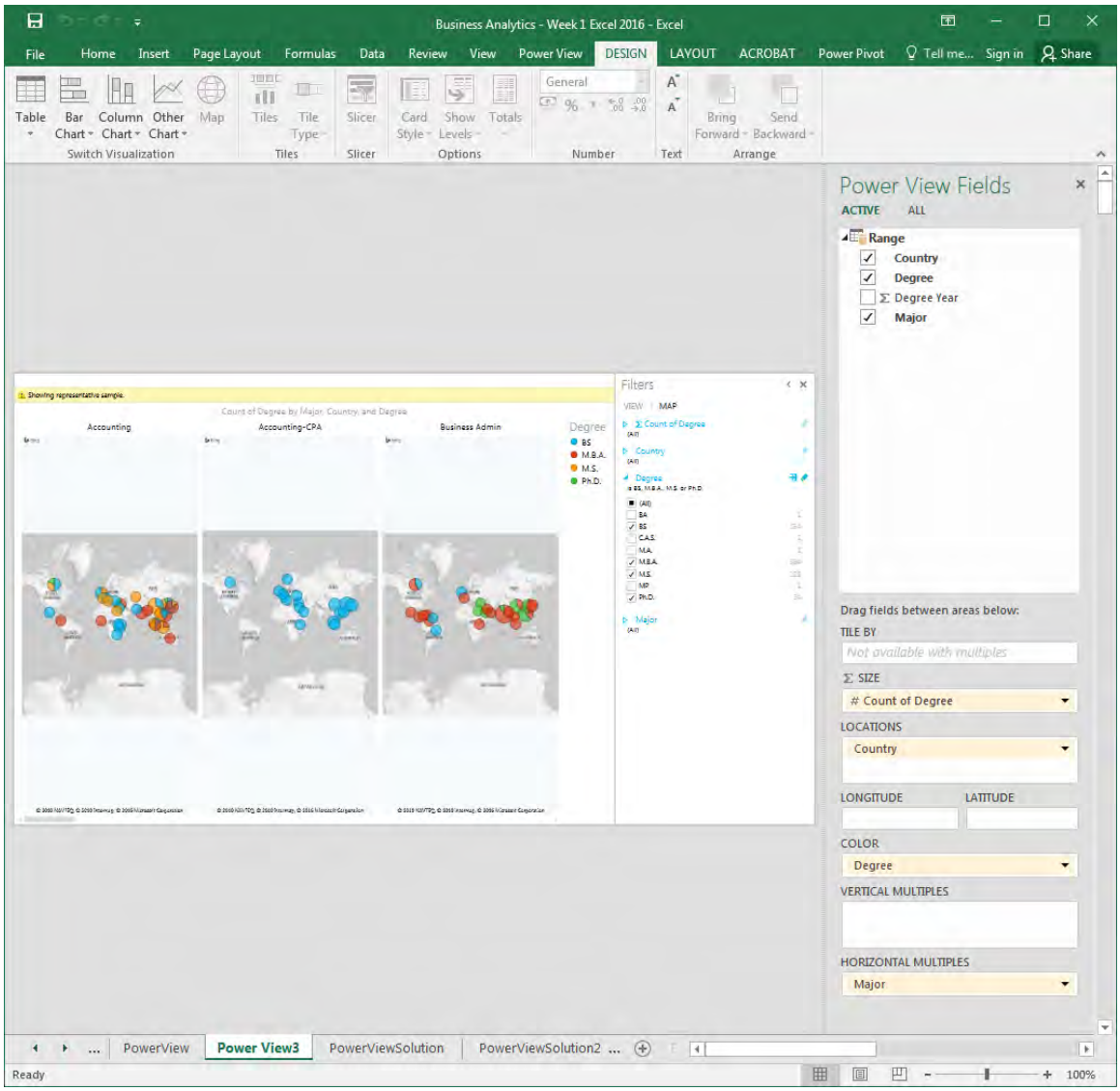
1. Click on the funnel to the upper right of the map.
2. In the Filters arrow to the right of the map, click on the arrow next to Degree
3. A list of degrees should appear. Check the degrees BS, MBA, MS and PhD
4. To reset to all degrees, click on All




To add an additional layer of analysis, drag Major to Vertical Multiples.



Drag it to Horizontal Multiples.



# SCM 651: Business Analytics

1

Week 2

# Agenda



- Google Analytics Access
- Teams
- Homework #1
- Review of concepts
- Group discussion of articles
  - Sustaining an Analytics Advantage
  - Creating Business Values with Analytics
  - Raising the Bar with Analytics

# Google Analytics, Team Formation & Homework #1



- Google Analytics
  - An account will be set up with your access
  - You will receive an email when your access is authorized
- Teams
  - Form teams of 4-5 students each
- Homework #1 – Pivot Tables & Charts, Correlation, Regression
  - due before class in Week 4 live session

# Week 2 - Review



- NPV
  - Calculates today's value of a cash flow stream
  - Investments are entered as negative numbers
- IRR
  - IRR > interest rate → positive NPV
  - IRR < interest rate → negative NPV
- Statistics
  - Skewness: curve is shifted left or right
  - Kurtosis: curve is peaked higher or lower

# Week 2 - Review



- Correlation
  - Strongest (highest positive or negative correlation): INTC & MSFT = 0.39
  - Weakest (closest to zero): CAT & MSFT = 0.08
- Correlation versus regression
  - If one variable changes, does the other variable go up or down? (correlation)
  - If one variable changes, how much does the other change? (regression)

# Week 2 - Review



- Regression Example
  - Fixed Costs
    - Measured by intercept
  - Variable costs
    - Measured by coefficient of variable
- Exponential regression
  - Compounded growth
- Power regression
  - Learning curve or volume efficiencies

# Week 2 - Review



- Multivariate regression (multiple X variables)
  - T-statistics measures the significance of one coefficient
  - F-statistic measures the significance of the entire equation
  - $R^2$ measures the goodness of fit of the equation, i.e., how much of the change in Y is explained by changes in X
- Seasonality (check answer)
  - Periodicity of 4 – quarterly
  - Periodicity of 12 – monthly
  - Periodicity of 52 - weekly

# Article #1: Sustaining an Analytics Advantage



- Sustaining an Analytics Advantage
  - What are some examples of creating competitive advantage with analytics (companies and their techniques)?

# Article #1: Sustaining an Analytics Advantage



- Sustaining an Analytics Advantage
  - What are some examples of creating competitive advantage with analytics (companies and their techniques)?
    - Wal-Mart: keep analytics techniques secret (consumer choice and human resources)
    - ABB: implement analytics fast (customer choice)
    - Procter & Gamble: apply to the right problem (reengineer the supply chain)
    - American Airlines (Sabre): data is more important (schedules)
    - Amazon: become data driven (algorithms)

# Article #1: Sustaining an Analytics Advantage



- Sustaining an Analytics Advantage
  - Analytics does not provide a sustainable competitive advantage
  - Analytics capability to change and innovate does provide a sustainable competitive advantage
  - Analytics is becoming a competitive necessity; ATM machines were initially a competitive advantage, now are a competitive necessity
    - ATM: Barclays Bank, London, 1967
    - ATM: Chemical Bank, Rockville Centre, New York, 1969

# Article #2: Creating Business Value with Analytics



- Creating Business Value with Analytics
  - What are the differences between competencies in information management and analytics expertise?
  - What are the advantages of starting with each?

# Article #2: Creating Business Value with Analytics



- Creating Business Value with Analytics
  - What are the differences between competencies in information management and analytics expertise?
    - Information management: develop enterprise wide data systems
    - Analytics: developing functional expertise
  - What are the advantages of starting with each?
    - Information management: break down cultural barriers, leverage customer focused data (expand sales)
    - Analytics: leverage algorithms to optimize activities (order placement, fulfillment, shipping, delivery)

# Article #3: Raising the Bar with Analytics



- Raising the Bar with Analytics
  - What new opportunities did StyleSeek and Entravision encounter when they used analytics?
  - What opportunity allowed MillerCoors to create efficiencies with analytics?

# Article #3: Raising the Bar with Analytics



- Raising the Bar with Analytics
  - What new opportunities did StyleSeek and Entravision encounter when they used analytics?
    - StyleSeek: sold their technology to partners
    - Entravision: expanded beyond media spots to information services for the Latino market
  - What opportunity allowed MillerCoors to create efficiencies with analytics?
    - MillerCoors: applied analytics to identify efficiencies with the joint venture

# Microsoft Excel: Financial Tools

## Microsoft Excel

- Go to the Course Website
- Download and print “Business Analytics – Week 2 Instructions.doc”
- Download “Business Analytics – Week 2 Excel 2013”

### Session 2.3: Net present value

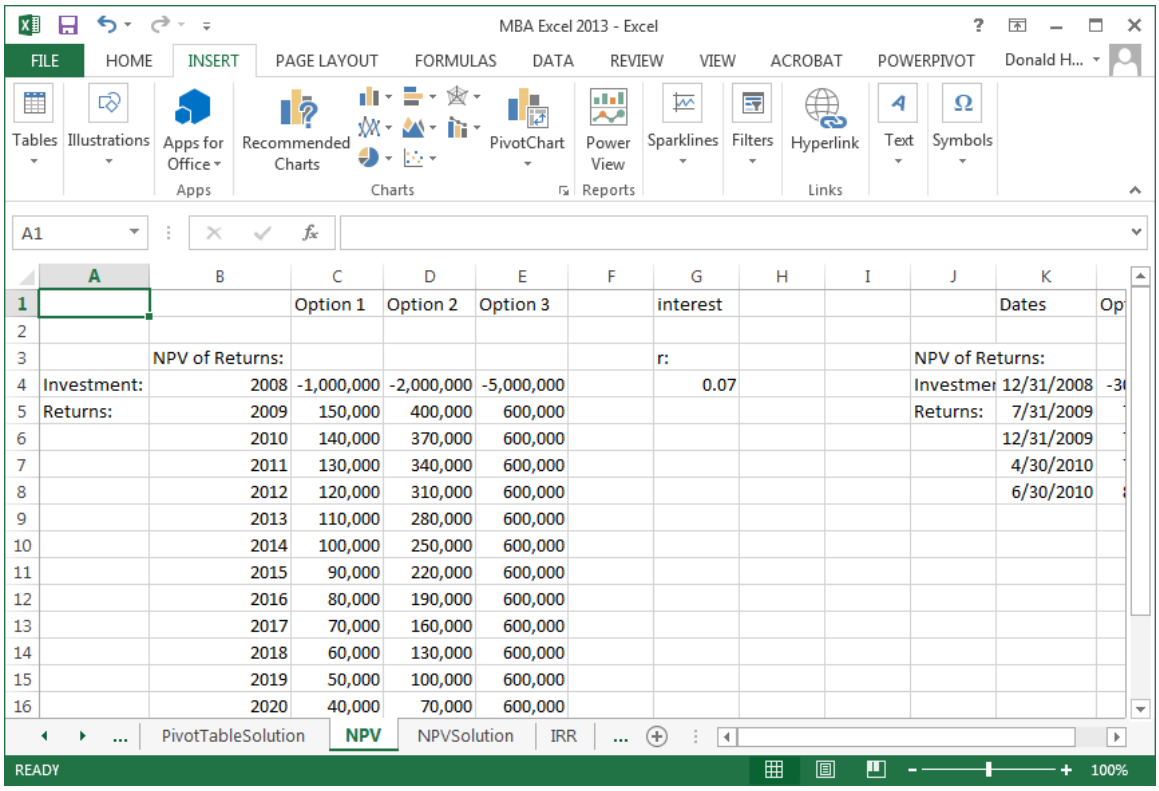
Net present value is used in finance to determine a value of a future stream of money in today’s dollars. Net present value determines what a revenue stream would be worth in today’s dollars. There are two NPV functions: NPV calculates the net present value of a stream of money at regular intervals; XNPV calculates the NPV of a stream of money at irregular intervals.

NPV(rate, range of cells)
XNPV(rate, values, dates)

The NPV calculation assumes that all cash flow is at the end of the period.

In the example below (NPV spreadsheet), investments are entered as negative numbers, returns as positive. Calculate the NPV for Options 1, 2 and 3 and enter the formulas in cells C3, D3, E3.

1. To calculate the NPV of Option 1, in cell C3, enter the formula for NPV.
2. For rate, refer to cell G4
3. For range of cells, highlight the investment and returns (cells C4:C20)
4. Similarly, calculate the NPV for Options 2 and 3




| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | Option 1 | Option 2 | Option 3 | | interest | | | | Dates |
| 2 | | | | | | | | | | | |
| 3 | | NPV of Returns: | | | | | r: | | | NPV of Returns: | |
| 4 | Investment: | 2008 | -1,000,000 | -2,000,000 | -5,000,000 | | 0.07 | | | Investmer | 12/31/2008 |
| 5 | Returns: | 2009 | 150,000 | 400,000 | 600,000 | | | | | Returns: | 7/31/2009 |
| 6 | | 2010 | 140,000 | 370,000 | 600,000 | | | | | | 12/31/2009 |
| 7 | | 2011 | 130,000 | 340,000 | 600,000 | | | | | | 4/30/2010 |
| 8 | | 2012 | 120,000 | 310,000 | 600,000 | | | | | | 6/30/2010 |
| 9 | | 2013 | 110,000 | 280,000 | 600,000 | | | | | | |
| 10 | | 2014 | 100,000 | 250,000 | 600,000 | | | | | | |
| 11 | | 2015 | 90,000 | 220,000 | 600,000 | | | | | | |
| 12 | | 2016 | 80,000 | 190,000 | 600,000 | | | | | | |
| 13 | | 2017 | 70,000 | 160,000 | 600,000 | | | | | | |
| 14 | | 2018 | 60,000 | 130,000 | 600,000 | | | | | | |
| 15 | | 2019 | 50,000 | 100,000 | 600,000 | | | | | | |
| 16 | | 2020 | 40,000 | 70,000 | 600,000 | | | | | | |

Now calculate the NPV of Option 4 using the XNPV formula and enter it in cell L3.

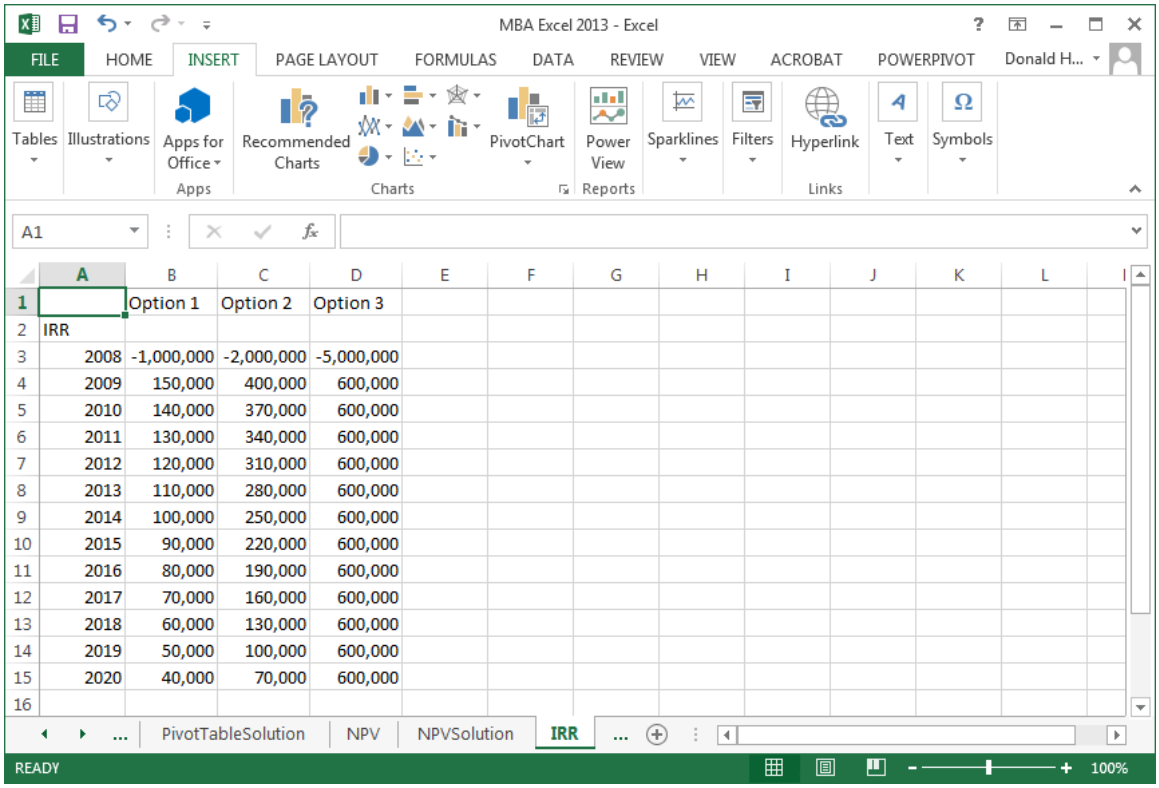
**Session 2.4: Internal rate of return**

The internal rate of return (IRR) calculates what the interest rate would have to be so that the NPV is zero. In some situations, there is no IRR. The spreadsheet is set up similar to the NPV problem, but investments are listed as negative returns. The function for IRR is:

IRR(data range)

The IRR spreadsheet gives an example of three investment options, revenue streams, and the calculated IRR. A corporation would compare the IRR to the possible returns available elsewhere to determine if a project was worthwhile.

In cell B2, enter the formula for IRR of Option 1. Similarly, calculate the IRR for Options 2 and 3 in cells C2 and D2.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Option 1 | Option 2 | Option 3 |
| 2 | IRR | | | |
| 3 | 2008 | -1,000,000 | -2,000,000 | -5,000,000 |
| 4 | 2009 | 150,000 | 400,000 | 600,000 |
| 5 | 2010 | 140,000 | 370,000 | 600,000 |
| 6 | 2011 | 130,000 | 340,000 | 600,000 |
| 7 | 2012 | 120,000 | 310,000 | 600,000 |
| 8 | 2013 | 110,000 | 280,000 | 600,000 |
| 9 | 2014 | 100,000 | 250,000 | 600,000 |
| 10 | 2015 | 90,000 | 220,000 | 600,000 |
| 11 | 2016 | 80,000 | 190,000 | 600,000 |
| 12 | 2017 | 70,000 | 160,000 | 600,000 |
| 13 | 2018 | 60,000 | 130,000 | 600,000 |
| 14 | 2019 | 50,000 | 100,000 | 600,000 |
| 15 | 2020 | 40,000 | 70,000 | 600,000 |
| 16 | | | | |

# Microsoft Excel: Statistics

**Session 2.5: Data Analysis Add-in**

The statistics options are available as an add-in to Excel. The steps to add it are:

1. In Excel, click on the File tab, then Options
2. Click on Add-Ins
3. Click Analysis TookPak Add-in, then Go
4. Check the box for Analysis ToolPak, then OK

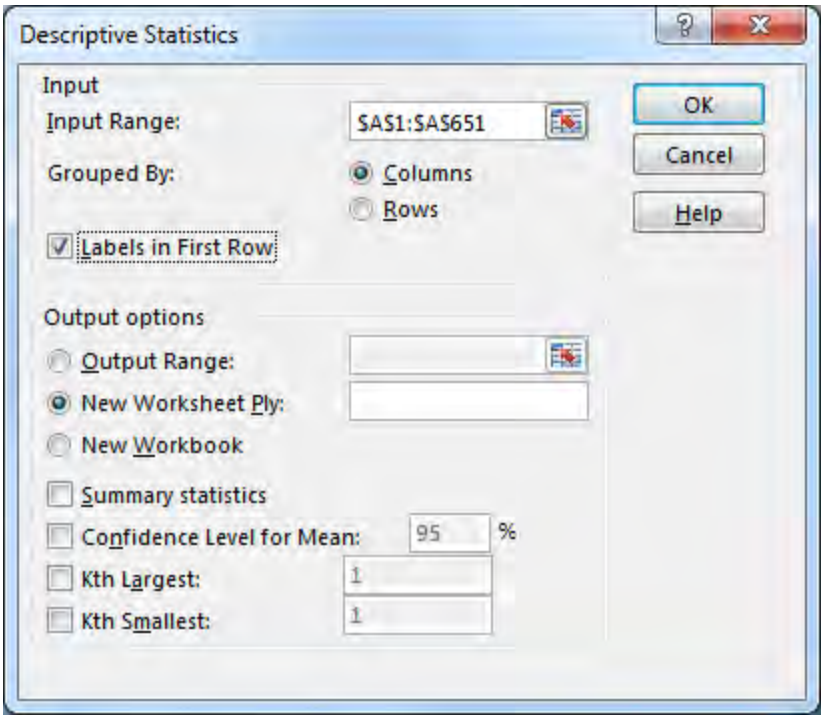
**Session 2.6: Data Analysis: Descriptive statistics**

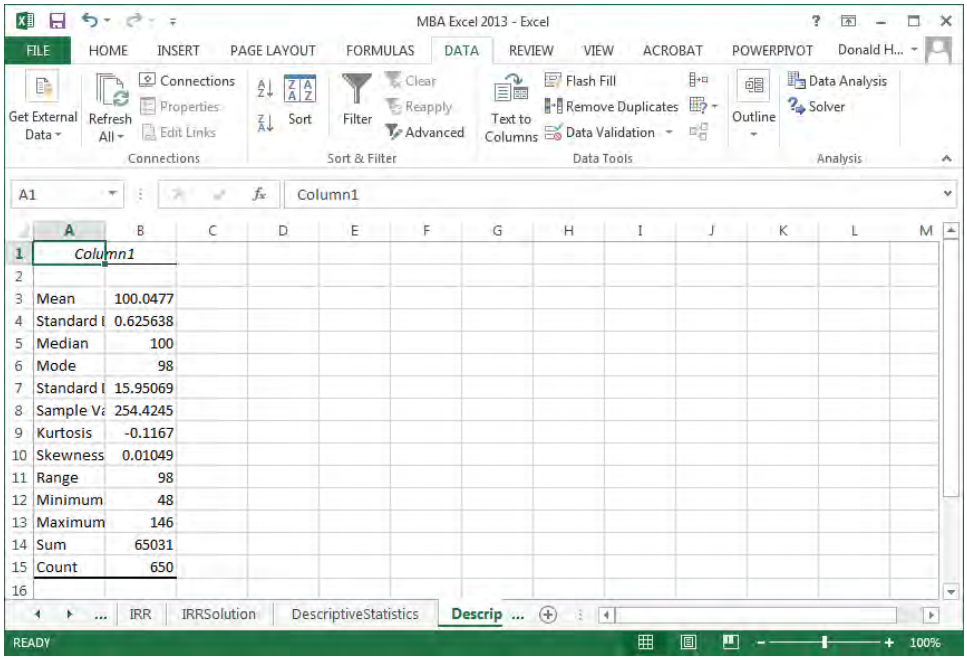
Use the DescriptiveStatistics spreadsheet tab for this exercise.
There are a number of descriptive statistics that can be automatically generated, including:

- Mean: arithmetic average
- Median: middle point in distribution
- Mode: most common value (highest frequency of occurrence)
- Kurtosis: is the data peaked higher or lower than normal?
- Skewness; is the peak shifted left or right?
- Standard deviation: measure of spread
- Range: highest value minus lowest value

To calculate the descriptive statistics:

1. Click on the data tab, then data analysis, descriptive statistics, and OK.
2. Enter the input range for the IQ data; if you include the header, click on Labels in first row.
3. Check Summary Statistics, then OK

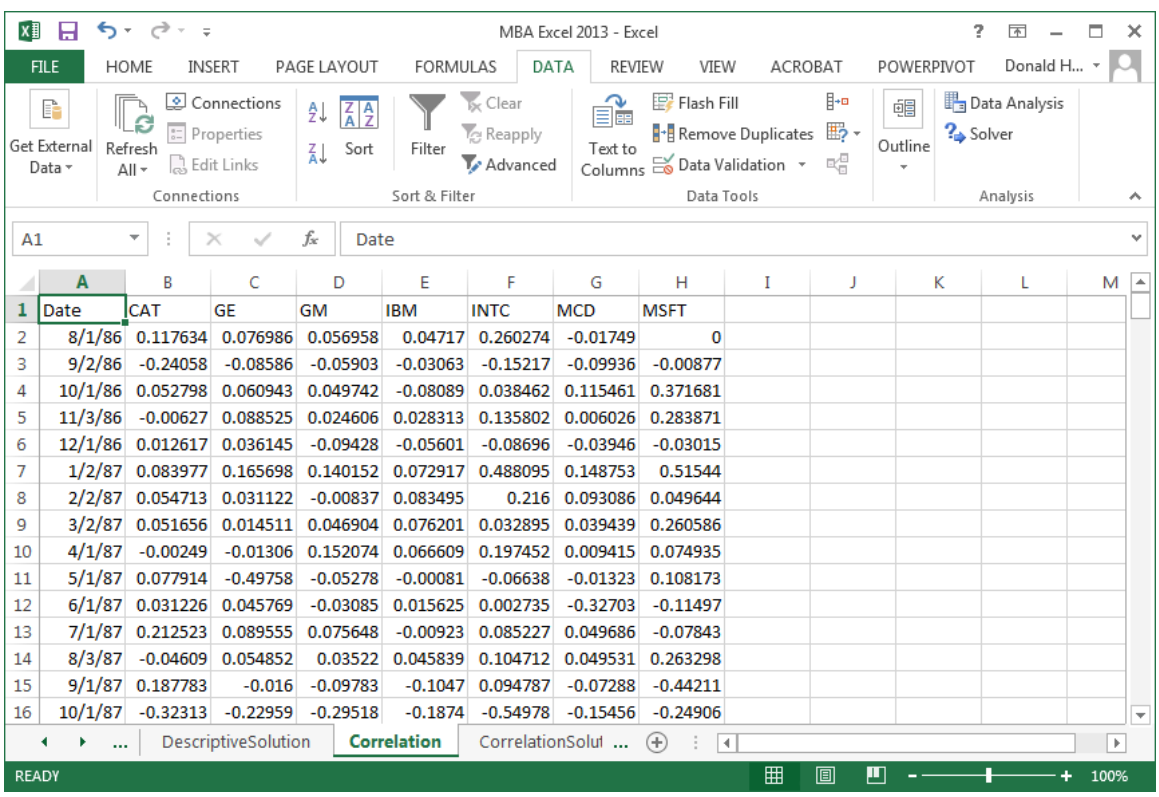




| | Column1 |
|---|---|
| Mean | 100.0477 |
| Standard E | 0.625638 |
| Median | 100 |
| Mode | 98 |
| Standard D | 15.95069 |
| Sample Va | 254.4245 |
| Kurtosis | -0.1167 |
| Skewness | 0.01049 |
| Range | 98 |
| Minimum | 48 |
| Maximum | 146 |
| Sum | 65031 |
| Count | 650 |

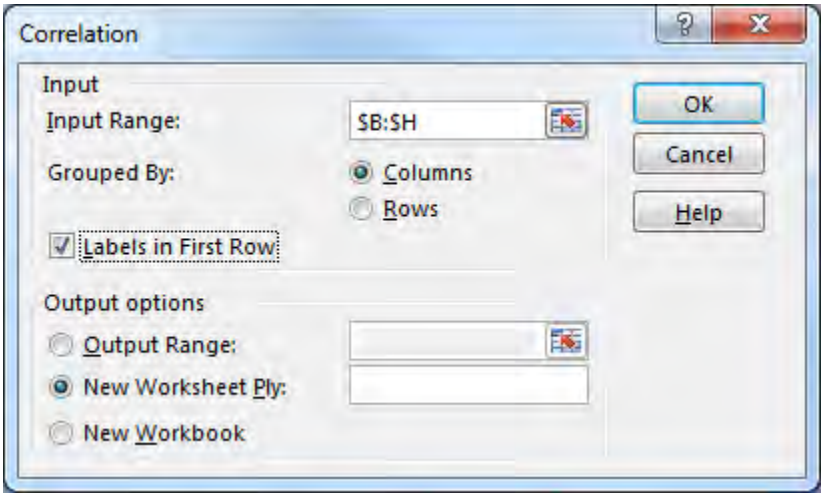
**Session 2.7: Correlations**

Correlation analysis identifies how two or more variables are related. For this exercise, use the Correlation spreadsheet. This spreadsheet records the upward or downward movement of stock by month.

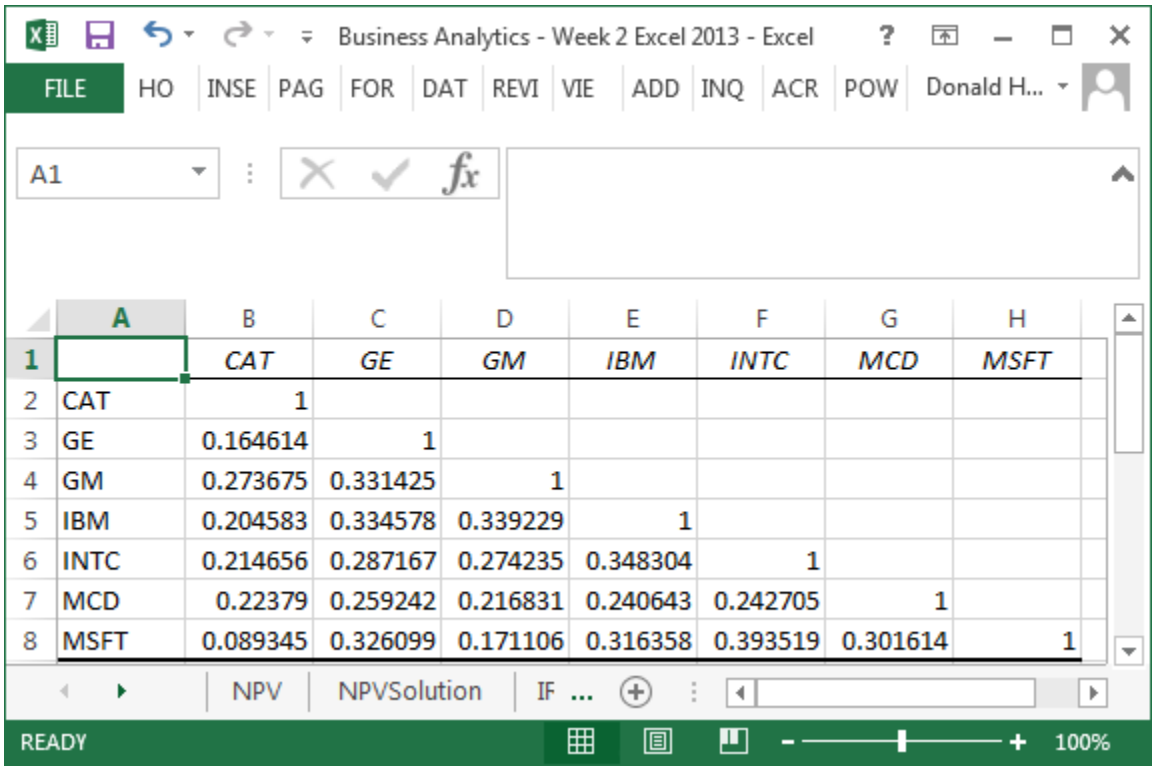


| Date | CAT | GE | GM | IBM | INTC | MCD | MSFT |
|---|---|---|---|---|---|---|---|
| 8/1/86 | 0.117634 | 0.076986 | 0.056958 | 0.04717 | 0.260274 | -0.01749 | 0 |
| 9/2/86 | -0.24058 | -0.08586 | -0.05903 | -0.03063 | -0.15217 | -0.09936 | -0.00877 |
| 10/1/86 | 0.052798 | 0.060943 | 0.049742 | -0.08089 | 0.038462 | 0.115461 | 0.371681 |
| 11/3/86 | -0.00627 | 0.088525 | 0.024606 | 0.028313 | 0.135802 | 0.006026 | 0.283871 |
| 12/1/86 | 0.012617 | 0.036145 | -0.09428 | -0.05601 | -0.08696 | -0.03946 | -0.03015 |
| 1/2/87 | 0.083977 | 0.165698 | 0.140152 | 0.072917 | 0.488095 | 0.148753 | 0.51544 |
| 2/2/87 | 0.054713 | 0.031122 | -0.00837 | 0.083495 | 0.216 | 0.093086 | 0.049644 |
| 3/2/87 | 0.051656 | 0.014511 | 0.046904 | 0.076201 | 0.032895 | 0.039439 | 0.260586 |
| 4/1/87 | -0.00249 | -0.01306 | 0.152074 | 0.066609 | 0.197452 | 0.009415 | 0.074935 |
| 5/1/87 | 0.077914 | -0.49758 | -0.05278 | -0.00081 | -0.06638 | -0.01323 | 0.108173 |
| 6/1/87 | 0.031226 | 0.045769 | -0.03085 | 0.015625 | 0.002735 | -0.32703 | -0.11497 |
| 7/1/87 | 0.212523 | 0.089555 | 0.075648 | -0.00923 | 0.085227 | 0.049686 | -0.07843 |
| 8/3/87 | -0.04609 | 0.054852 | 0.03522 | 0.045839 | 0.104712 | 0.049531 | 0.263298 |
| 9/1/87 | 0.187783 | -0.016 | -0.09783 | -0.1047 | 0.094787 | -0.07288 | -0.44211 |
| 10/1/87 | -0.32313 | -0.22959 | -0.29518 | -0.1874 | -0.54978 | -0.15456 | -0.24906 |

The stocks listed are Caterpillar, General Electric, General Motors, IBM, Intel, McDonalds and Microsoft. Click on the data tab, data analysis, correlation, then OK. Highlight columns B through H, group by columns, check Labels in First Row, then OK.




The result is shown below. A positive correlation means that when one variable increases, the other increases. A negative correlation means that when one increases, the other decreases.




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | *CAT* | *GE* | *GM* | *IBM* | *INTC* | *MCD* | *MSFT* |
| 2 | CAT | 1 | | | | | | |
| 3 | GE | 0.164614 | 1 | | | | | |
| 4 | GM | 0.273675 | 0.331425 | 1 | | | | |
| 5 | IBM | 0.204583 | 0.334578 | 0.339229 | 1 | | | |
| 6 | INTC | 0.214656 | 0.287167 | 0.274235 | 0.348304 | 1 | | |
| 7 | MCD | 0.22379 | 0.259242 | 0.216831 | 0.240643 | 0.242705 | 1 | |
| 8 | MSFT | 0.089345 | 0.326099 | 0.171106 | 0.316358 | 0.393519 | 0.301614 | 1 |

## Session 2.9: Univariate Linear Regression

### Regression Assumptions

Regression is a technique that attempts to measure the relationship between and outcome variable (dependent) and explanatory variables (independent). To use linear regression, there are three key assumptions

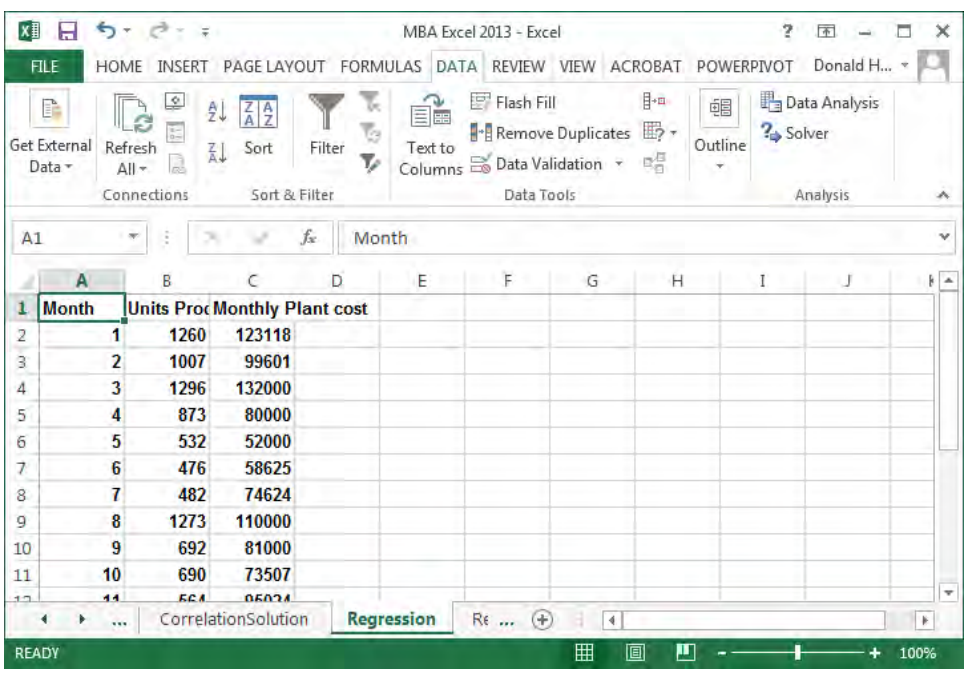
1. relationship between x and y is linear
2. the x's are fixed numbers, not random variables (non-stochastic), not related to each other, i.e., independent: $Corr(X_i,X_j)=0$
3. the error terms:
    a. have zero mean and constant variance: $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$
    b. the error terms are independent: $Cov(\varepsilon_i,\varepsilon_j) = 0$
    c. the error terms are normally distributed $\sim N(0,\sigma^2)$

Violation of these assumptions requires the use of more sophisticated techniques.

### Straight line relationships

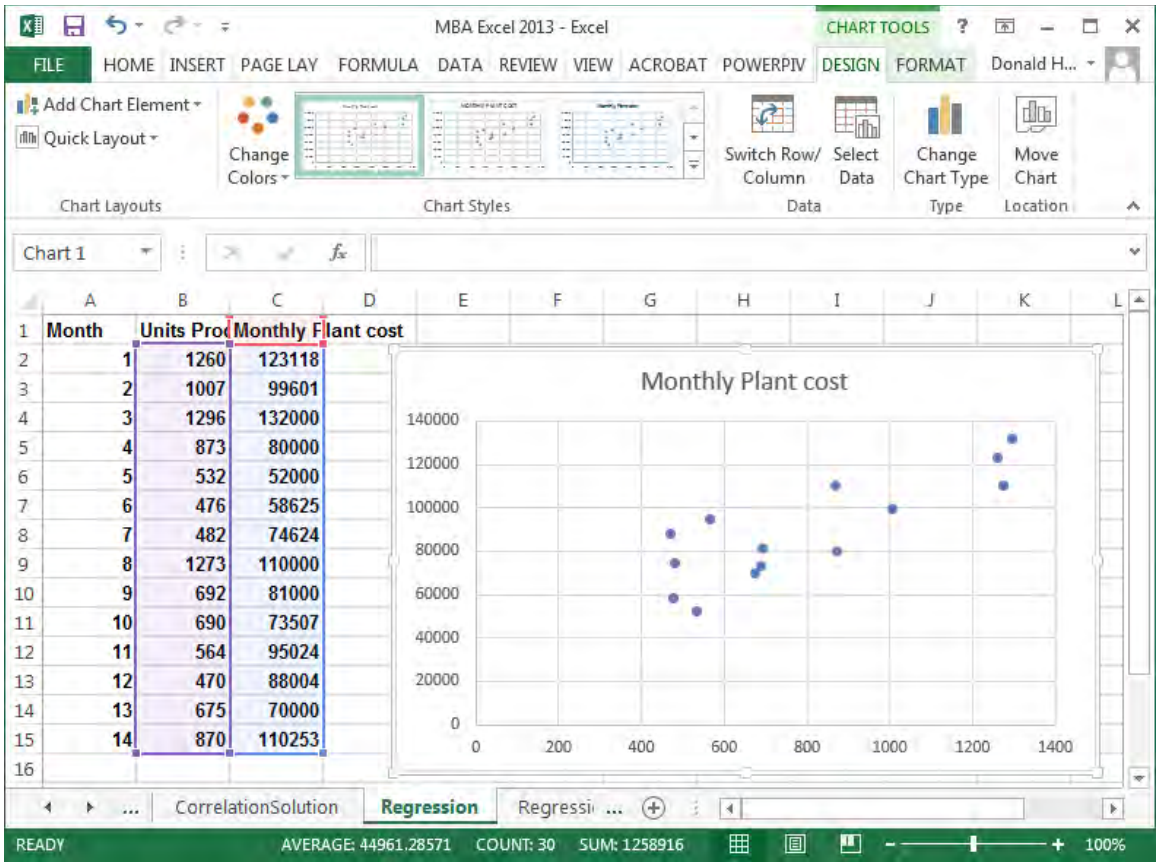
When you want to determine if there is a straight line relationship in statistics, you can run a regression. Excel has the ability to perform regression analysis. For example, if you wanted to model the relationship between items produced and factory costs, you could estimate the linear relationship. Units produced would be called the independent variable; production costs would be the dependent variable. The output, costs, depends on the input, number of units produced.

For this example, use the Regression spreadsheet.



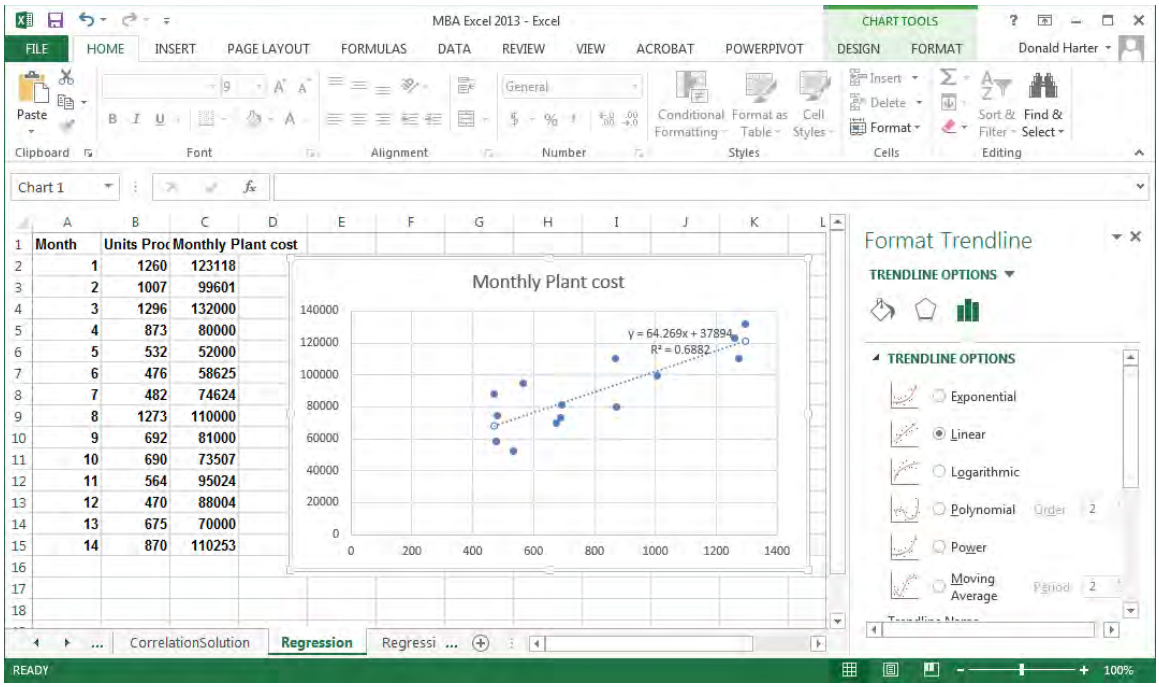
Let's first draw a scatterplot to see what the data looks like.

1. Click on the Insert tab
2. Highlight the the cells b1:c15
3. Click on Scatter in the charts group.




Reviewing the chart, it appears that there is a linear relationship. We will therefore perform a linear regression. Click on any data point, right click, then add trendline.

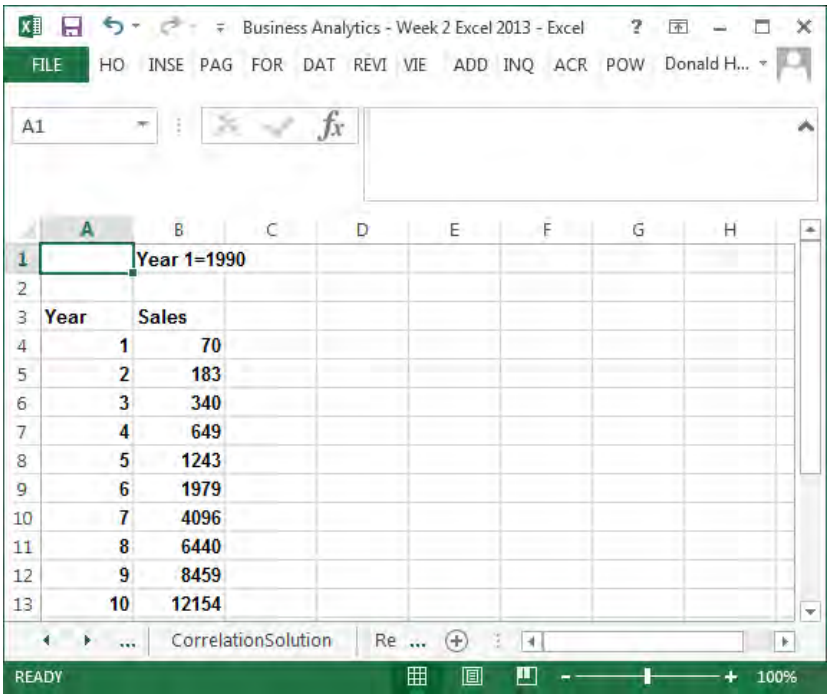
In the Format Trendline, Trendline Options, select Linear, then check the boxes for Display Equation and Display R-squared value.




In the picture above, the coefficient on x is approximately 64. This means that as unit production increases by one, costs increase by $64. What does the number 37,894 represent? What does the $R^2 = 0.6882$ mean?

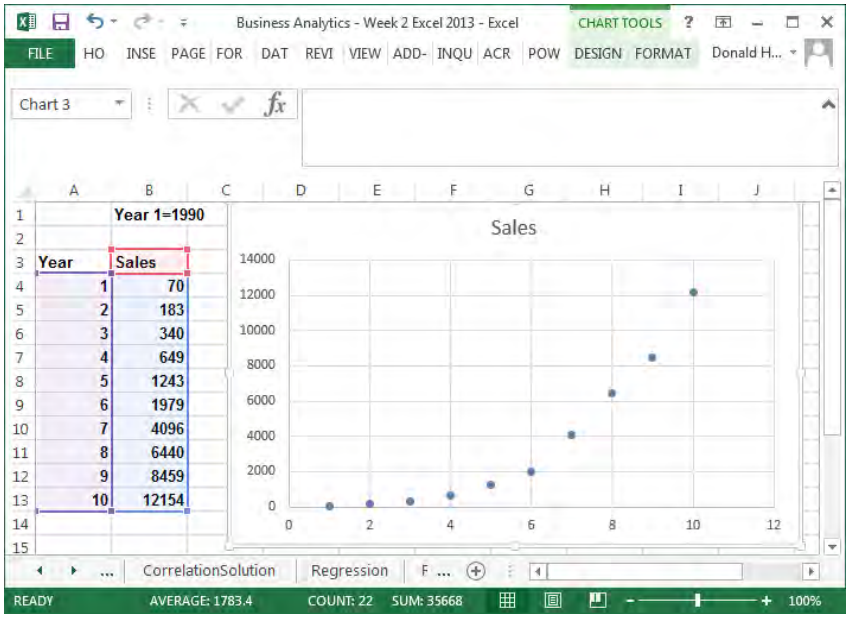
**Session 2.10: Exponential regression**

Some data relationships are not linear, but grow at an increasing rate. These curves often follow the exponential growth curve. An exponential growth curve will have the same percentage growth per period compounded over time. Use the Exponential spreadsheet.

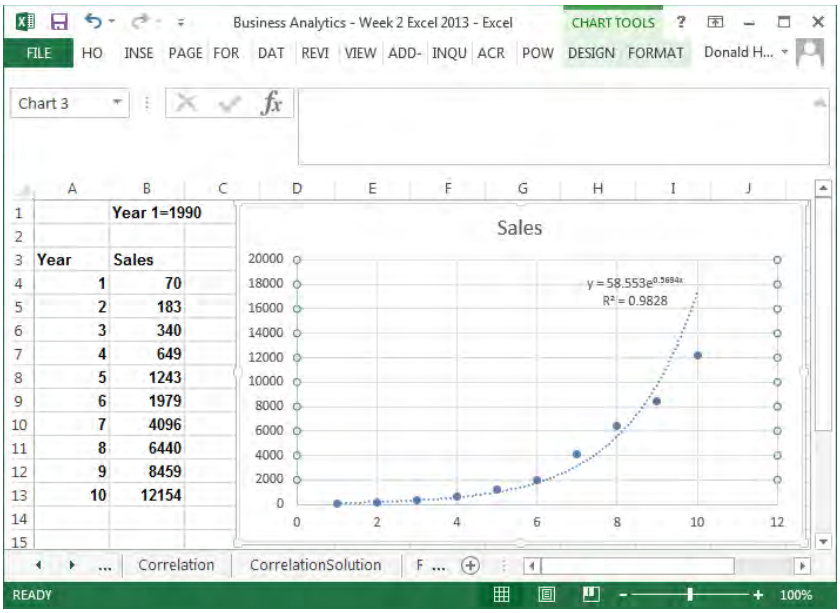



Let's first draw a scatterplot to see what the data looks like.

1. Click on the Insert tab
2. Highlight the the cells a3:b13
3. Click on Scatter in the charts group.

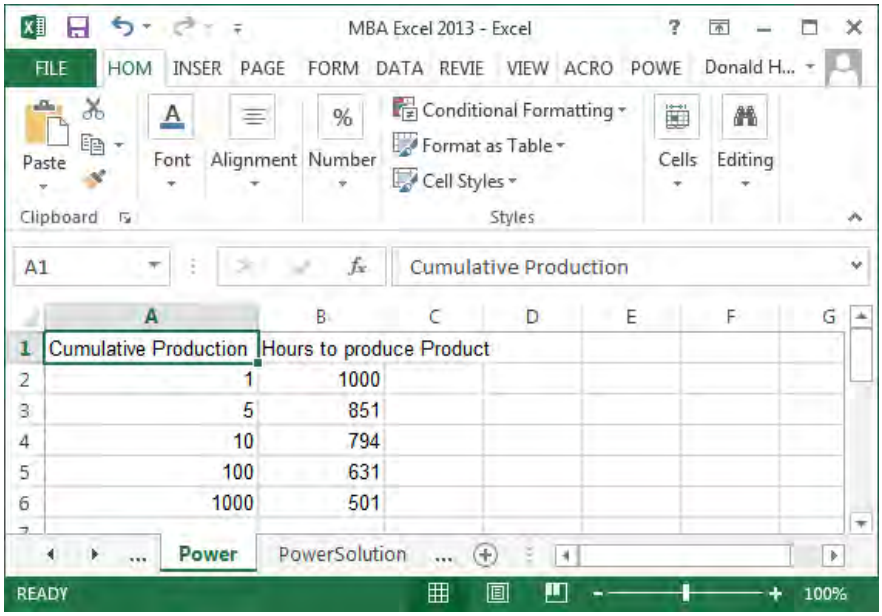


This data definitely does not look linear. So let's use the exponential curve. Click on any data point, right click, then add trendline. Select exponential, display equation and display R-squared, then Close.

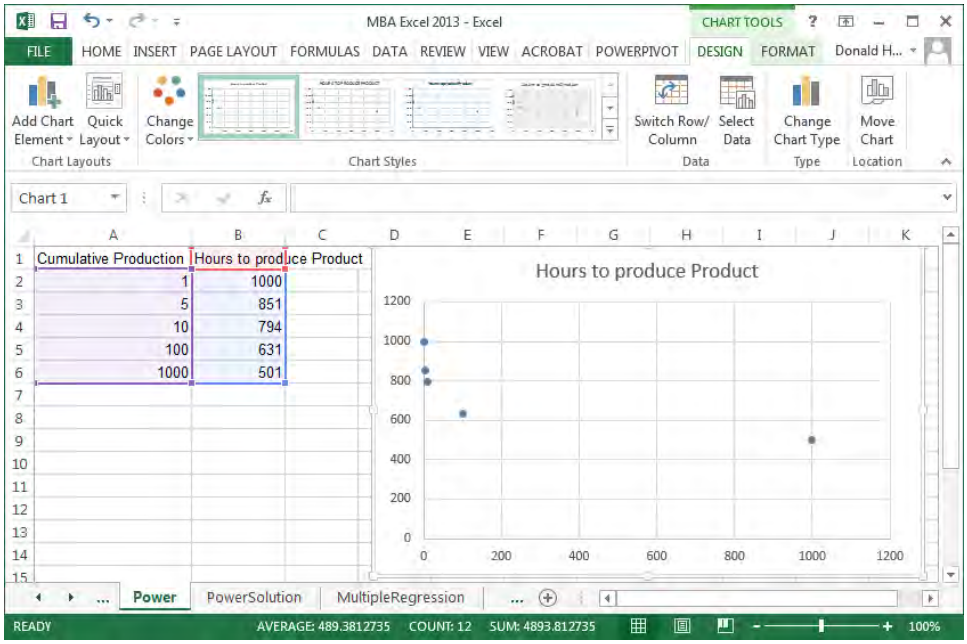



**Session 2.11: Power regression**

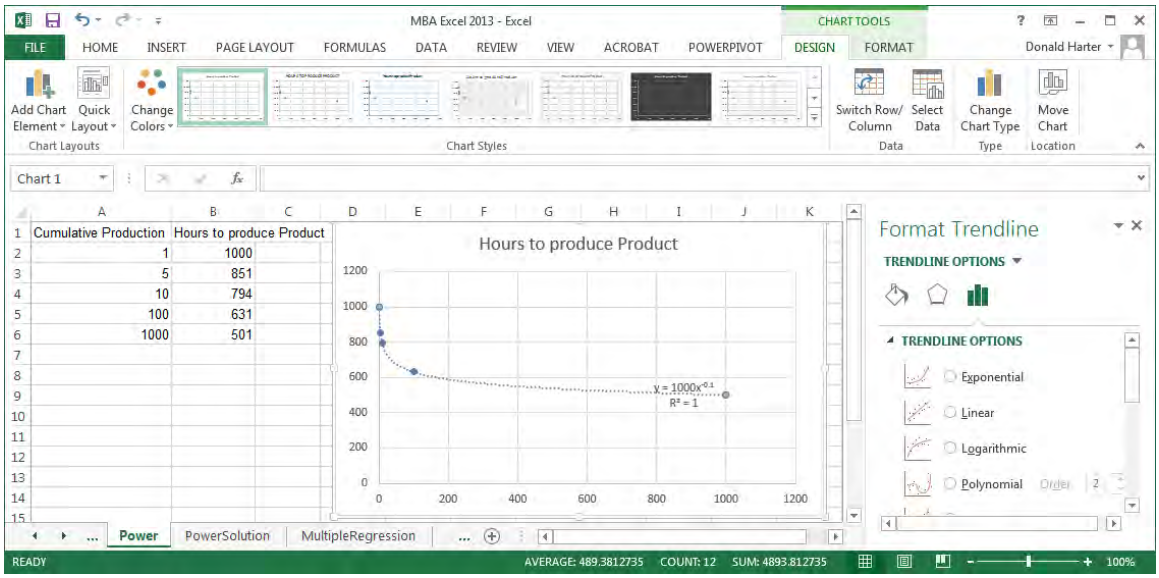
The power curve allows you to examine economies of scale and diseconomies of scale. Economies of scale means that you become more efficient as volume increases. Diseconomies of scale means that you become less efficient as volume increases. Use the Power spreadsheet.



Let's graph as before. Click on Insert, Scatter.

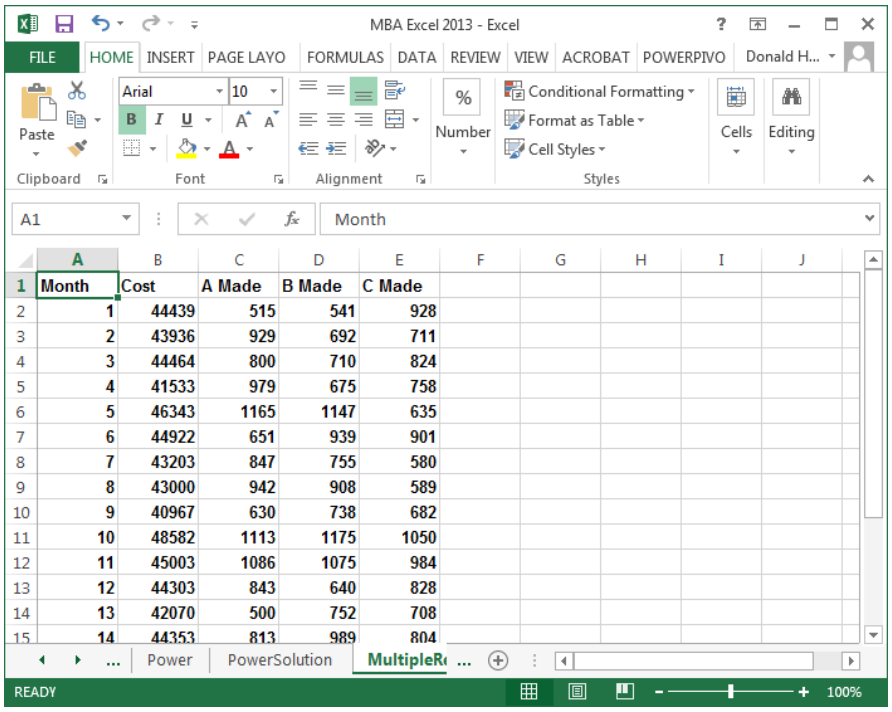



Now click on any data point in the graph, right click, add trendline. Click on Power, display equation, display R-squared.



**Session 2.12: Multivariate regression**

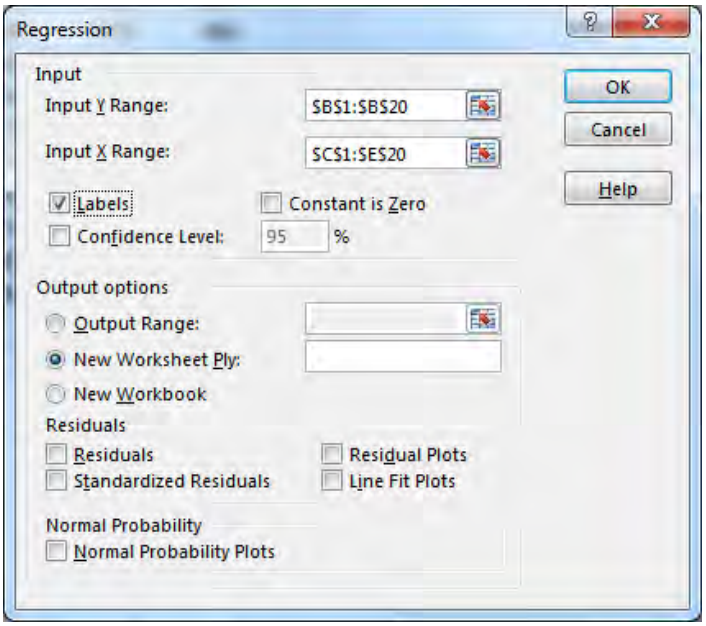
When we reviewed linear regression earlier, we only had one independent variable. Multiple regression includes several independent variables. Use the Multiple Regression spreadsheet.

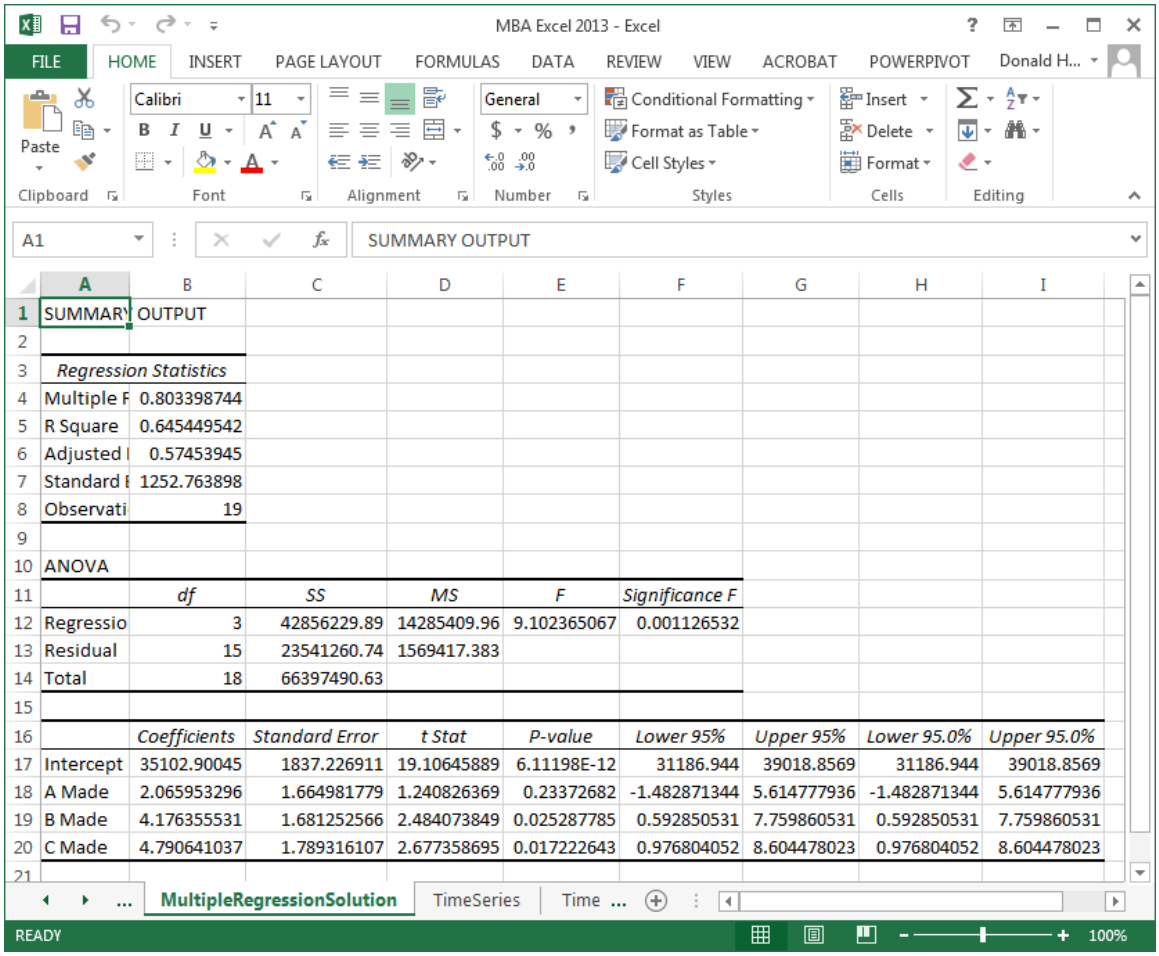



| Month | Cost | A Made | B Made | C Made |
|---|---|---|---|---|
| 1 | 44439 | 515 | 541 | 928 |
| 2 | 43936 | 929 | 692 | 711 |
| 3 | 44464 | 800 | 710 | 824 |
| 4 | 41533 | 979 | 675 | 758 |
| 5 | 46343 | 1165 | 1147 | 635 |
| 6 | 44922 | 651 | 939 | 901 |
| 7 | 43203 | 847 | 755 | 580 |
| 8 | 43000 | 942 | 908 | 589 |
| 9 | 40967 | 630 | 738 | 682 |
| 10 | 48582 | 1113 | 1175 | 1050 |
| 11 | 45003 | 1086 | 1075 | 984 |
| 12 | 44303 | 843 | 640 | 828 |
| 13 | 42070 | 500 | 752 | 708 |
| 14 | 44353 | 813 | 989 | 804 |

To run a multiple regression:

1. Click on the data tab, data analysis, regression, then OK.
2. For the Y-range, highlight the values in the B column for cost
3. For the X-range, highlight the values in the C, D, and E columns.
4. If you included the headings at the top of the columns, click labels.



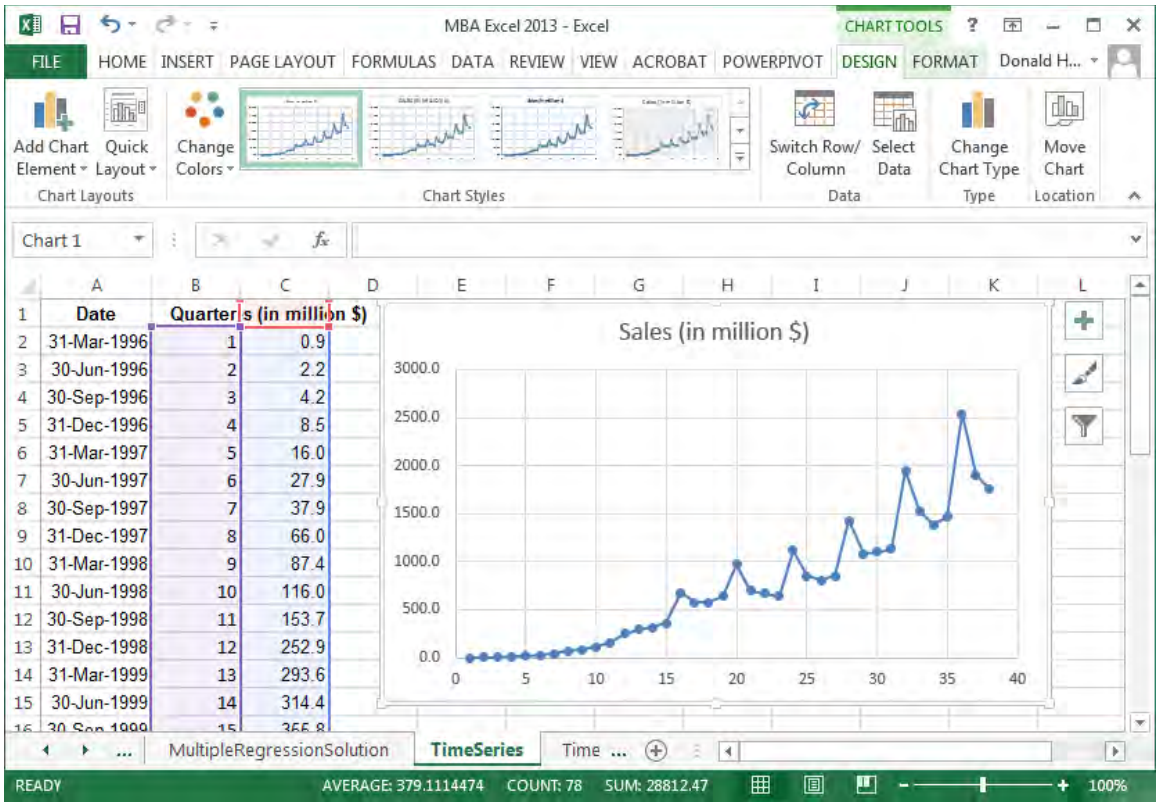


| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY | OUTPUT | | | | | | | |
| 2 | | | | | | | | | |
| 3 | *Regression Statistics* | | | | | | | | |
| 4 | Multiple F | 0.803398744 | | | | | | | |
| 5 | R Square | 0.645449542 | | | | | | | |
| 6 | Adjusted I | 0.57453945 | | | | | | | |
| 7 | Standard E | 1252.763898 | | | | | | | |
| 8 | Observati | 19 | | | | | | | |
| 9 | | | | | | | | | |
| 10 | ANOVA | | | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 12 | Regressio | 3 | 42856229.89 | 14285409.96 | 9.102365067 | 0.001126532 | | | |
| 13 | Residual | 15 | 23541260.74 | 1569417.383 | | | | | |
| 14 | Total | 18 | 66397490.63 | | | | | | |
| 15 | | | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| 17 | Intercept | 35102.90045 | 1837.226911 | 19.10645889 | 6.11198E-12 | 31186.944 | 39018.8569 | 31186.944 | 39018.8569 |
| 18 | A Made | 2.065953296 | 1.664981779 | 1.240826369 | 0.23372682 | -1.482871344 | 5.614777936 | -1.482871344 | 5.614777936 |
| 19 | B Made | 4.176355531 | 1.681252566 | 2.484073849 | 0.025287785 | 0.592850531 | 7.759860531 | 0.592850531 | 7.759860531 |
| 20 | C Made | 4.790641037 | 1.789316107 | 2.677358695 | 0.017222643 | 0.976804052 | 8.604478023 | 0.976804052 | 8.604478023 |
| 21 | | | | | | | | | |

**Session 2.13: Time series moving average regression**

Time series problems have data where one data point is dependent on the previous data point. For example, the closing price of Microsoft stock can be tracked day by day. Today's price is dependent on yesterday's price. This dependency from one day to the next, or one time period to the next, is a characteristic of time series data.

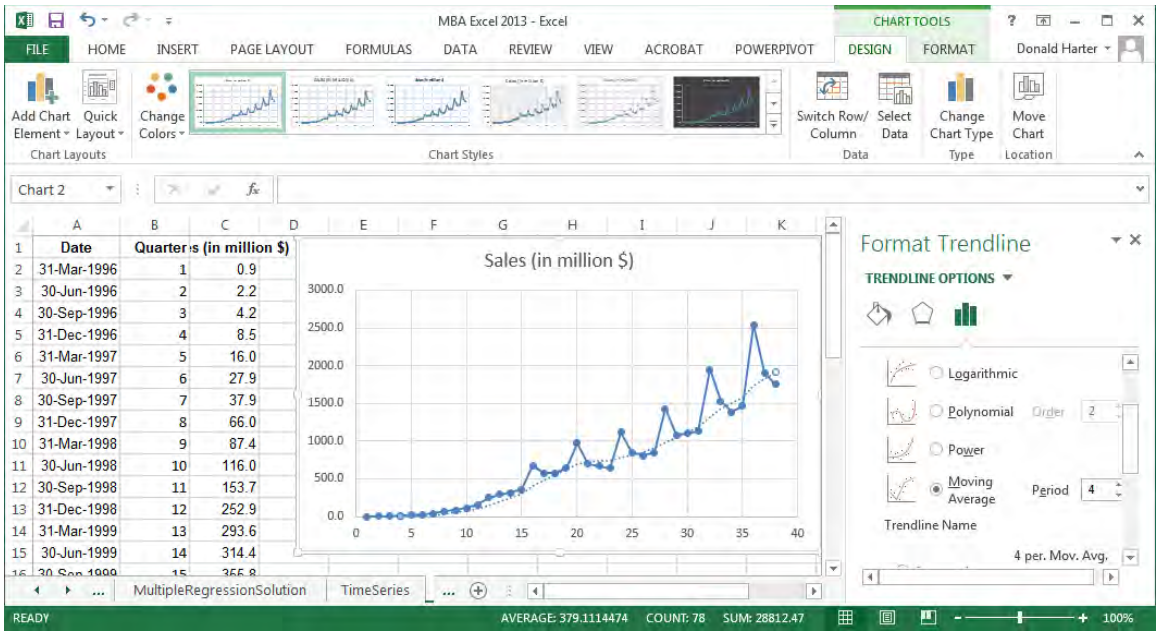
Often there is so much variation in time series data that it's hard to see trends. Seasonality also masks a trend. Seasonality is variations in data due to high or low points that occur at regular intervals. Create a scatter plot for the Amazon data using the Time Series spreadsheet. For this scatter plot, select the option to connect the dots.




Notice that there is seasonality in the data. Amazon sales tend to peak during the fourth quarter of each year due to holiday sales. However, this seasonality masks the true trend. A moving average helps to see the trend.

To add a moving average line, follow these steps:

1. Right click on a data point.
2. Click on Add Trendline
3. Click the checkbox for moving average. Since we have quarterly data, let's identify the number of periods as four.




| | Date | Quarters (in million $) | |
|---|---|---|---|
| 2 | 31-Mar-1996 | 1 | 0.9 |
| 3 | 30-Jun-1996 | 2 | 2.2 |
| 4 | 30-Sep-1996 | 3 | 4.2 |
| 5 | 31-Dec-1996 | 4 | 8.5 |
| 6 | 31-Mar-1997 | 5 | 16.0 |
| 7 | 30-Jun-1997 | 6 | 27.9 |
| 8 | 30-Sep-1997 | 7 | 37.9 |
| 9 | 31-Dec-1997 | 8 | 66.0 |
| 10 | 31-Mar-1998 | 9 | 87.4 |
| 11 | 30-Jun-1998 | 10 | 116.0 |
| 12 | 30-Sep-1998 | 11 | 153.7 |
| 13 | 31-Dec-1998 | 12 | 252.9 |
| 14 | 31-Mar-1999 | 13 | 293.6 |
| 15 | 30-Jun-1999 | 14 | 314.4 |

The moving average line is superimposed on the graph. It's now very clear what the trend looks like when a moving average accounts for seasonality.

View All View Keyframes

## Assumptions about Values

- In the real world, you don't know.
- **Sensitivity analysis:** Explore what happens if your assumpons ar en't right.
  - Determine how answers change based on variables.

5 of 18

## One-Way Sensitivity Analysis

- Change one variable to see how it affects results.
- Excel can look at a range of values.

8 of 18

## One-Way Sensitivity Analysis Example

| Price | Profit |
|---|---|
| | $57,950.00 |
| $1.00 | $(14,200.00) |
| $1.25 | $(2,000.00) |
| $1.50 | $9,075.00 |
| $1.75 | $19,025.00 |
| $2.00 | $27,850.00 |
| $2.25 | $35,550.00 |
| $2.50 | $42,125.00 |
| $2.75 | $47,575.00 |
| $3.00 | $51,900.00 |
| $3.25 | $55,100.00 |
| $3.50 | $57,175.00 |
| $3.75 | $58,125.00 |
| $4.00 | $57,950.00 |

17 of 18

Up next: Perform a one-way sensitivity analysis over a range of values in Excel.18 of 18

View All View Keyframes

## Two-Way Sensitivity Analysis

- Change two variables simultaneously to see their interacons.
- See how assumpons a ffect results.

3 of 17

## Two-Way Sensitivity Analysis Example

| | | Unit Cost | | |
|---|---|---|---|---|
| | $57,950.00 | $0.30 | $0.35 | $0.40 |
| | $1.50 | $16,800.00 | $14,225.00 | $11,650.00 |
| | $1.75 | $26,412.50 | $23,950.00 | $21,487.50 |
| | $2.00 | $34,900.00 | $32,550.00 | $30,200.00 |
| | $2.25 | $42,262.50 | $40,025.00 | $37,787.50 |
| | $2.50 | $48,500.00 | $46,375.00 | $44,250.00 |
| | $2.75 | $53,612.50 | $51,600.00 | $49,587.50 |
| Price | $3.00 | $57,600.00 | $55,700.00 | $53,800.00 |
| | $3.25 | $60,462.50 | $58,675.00 | $56,887.50 |
| | $3.50 | $62,200.00 | $60,525.00 | $58,850.00 |
| | $3.75 | $62,812.50 | $61,250.00 | $59,687.50 |
| | $4.00 | $62,300.00 | $60,850.00 | $ 59,400.00 |
| | $4.25 | $60,662.50 | $59,325.00 | $57,987.50 |
| | $4.50 | $57,900.00 | $56,675.00 | $55,450.00 |
| | $4.75 | $54,012.50 | $52,900.00 | $51,787.50 |
| | $5.00 | $49,000.00 | $48,000.00 | $47,000.00 |

4 of 17

## Two-Way Sensitivity Analysis Exercise

- Perform two-way sensivity analy sis in Excel.
- Change unit cost and price.
- Find opmal pr ofit.

17 of 17

## Google Analytics

- Facilitates collecon and analy sis of data streams
- Helps determine pa erns and effecv eness of website and ads



## Types of Data

- Audience
  - Demographics
  - Geography
  - Technology
- Acquision
  - How to drive traffic to site
- Behavior
  - How engaged users are
- Visitor flow
  - How users navigate site



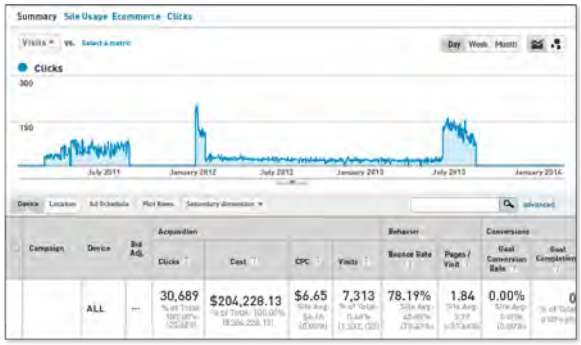
## Audience Data

## Audience Data: Geography

| Demographics | Country/Territory | Visits | % Visits |
|---|---|---|---|
| Language | 1. United States | 29,968 | 73.82% |
| Country/Territory | 2. China | 3,142 | 7.74% |
| City | 3. India | 2,786 | 6.86% |
| System | 4. Canada | 379 | 0.93% |
| Browser | 5. United Kingdom | 321 | 0.79% |
| Operating System | 6. Philippines | 246 | 0.61% |
| Service Provider | 7. Germany | 204 | 0.50% |
| Mobile | 8. Taiwan | 200 | 0.49% |
| Operating System | 9. Pakistan | 188 | 0.46% |
| Service Provider | 10. South Korea | 175 | 0.43% |
| Screen Resolution | | | view full report |

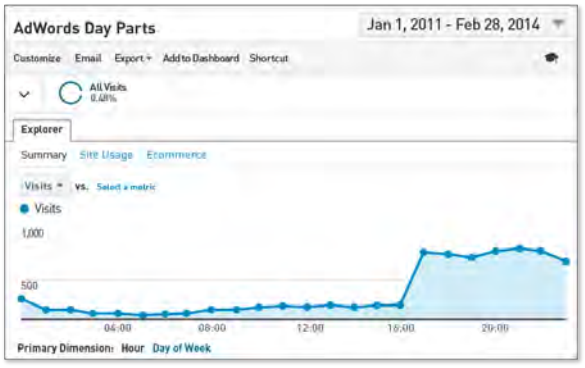
33 of 52

## Google Adwords




40 of 52

## Google Adwords: Day Parts




46 of 52

## Behavior

## Visitor Flow

# SCM 651: Business Analytics

1

Week 3

# Agenda



- Google Analytics Access
- Teams
- Homework #1: due before next class
- Review of concepts
- Group discussion of articles
  - Web Analytics: Enhancing Customer Relationship Management
  - How eBay Uses Data and Analytics to Get Closer to Its (Massive) Customer Base

# Google Analytics, Team Formation & Homework #1



- Google Analytics
  - Required for Homework #2
- Teams
  - Form teams of 4-5 students each
- Homework #1 – Regression
  - due before class in Week 4 live session

# Week 3 - Review



- One-way sensitivity analysis
  - Only one variable changes, allowing you to see the effect on the outcome
- Two-way sensitivity analysis
  - Two variables changes simultaneously, allowing you to see the joint effect

# Week 3 - Review



- Conditional formatting
  - Colors, bars, icons to facilitate interpretation
  - When are some better than others?
- Dashboards
  - Combine table and graphical representation

# Week 3 - Review



- Google Analytics: general measures
  - Visits, new visits, % new visits
  - Bounce rate, average visit duration
- Google Analytics: locations
  - Countries, states, cities
  - Visits, new visits, % new visits
- Google analytics: behavior
  - Frequency and recency
  - Engagement (length of time on site)

# Week 3 - Review



- Google Analytics: technology
  - Operating system & browser
  - Network provider
  - Mobile device
- Google Analytics: visitor flow
  - Entry and exit point

# Week 3 - Review



- Google Analytics: Acquisition
  - Channels (organic search, referral, direct, social, email)
  - Source/medium (google/organic, direct/none, syr.edu/referral, bing/organic, va.gov/referral)

# Week 3 - Review



- Google Analytics: advertising
  - Advertising (key words, adwords)
  - Dayparting (time of day)
- Google Analytics: social
  - Facebook, LinkedIn, Twitter, Google+

# Week 3 - Review



- Google Analytics: Behavior
  - Site content (ability to drill down to each web page for page views, average time on page, bounce rate, exit %)
  - Landing pages (first page viewed by user)
  - Exit pages (last page viewed by user)
  - Site speed, page timings, speed suggestions

# Article #1: Web Analytics



- Web Analytics: Enhancing Customer Relationship Management
  - Describe the four main categories of metrics and relate to the Google analytics lessons
    - Website usability; Traffic sources; Visitor profiles; Conversion statistics
  - Describe the common techniques for Web analytics
    - Clustering/classification; Association rules; Path analysis; Sequential patterns
  - What are some business applications of web analytics?

# Article #2: How eBay Uses Data



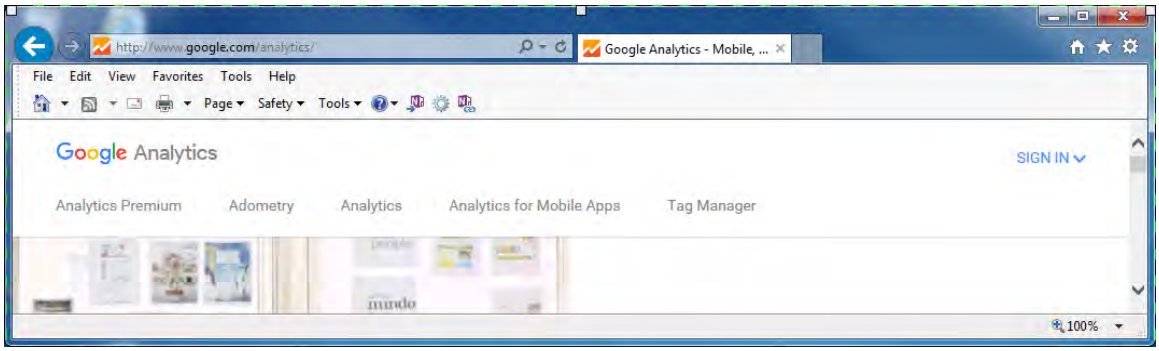
- How eBay Uses Data and Analytics to Get Closer to Its (Massive) Customer Base
  - What is an A/B test and what is its purpose?
  - Describe the three biggest challenges of web data
    - Data at a large scale
    - Collecting the right data
    - New kinds of data
  - How can Power Sellers use data better?
  - Why are web analytics better than surveys?

# Google Analytics

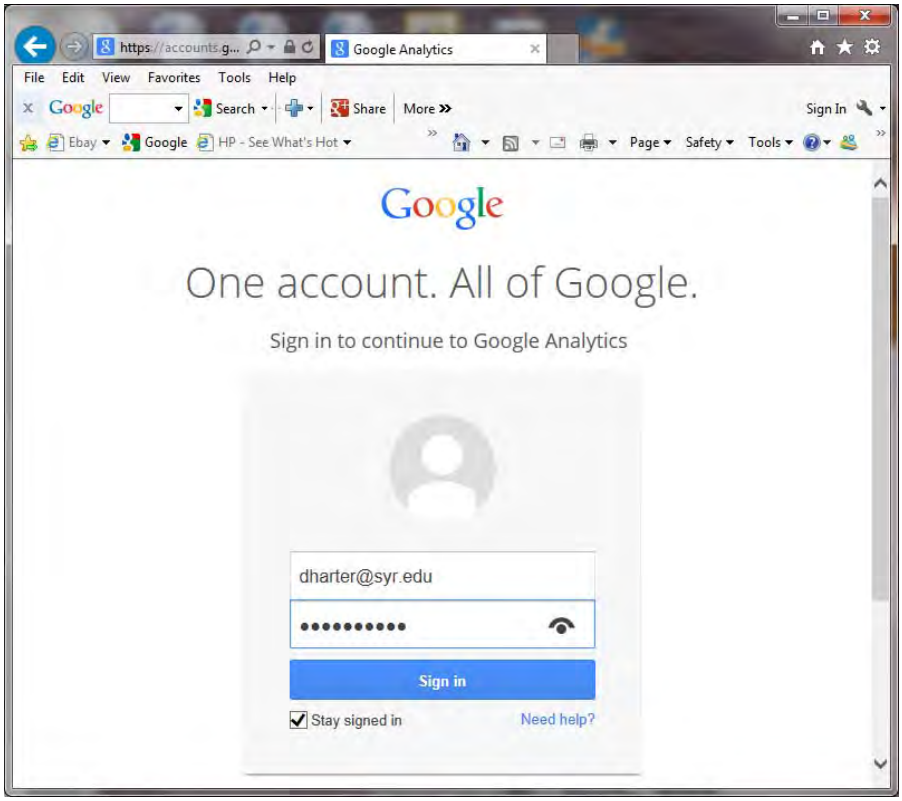
**Session 3.8: Google Analytics**

To begin Google Analytics, go to the web page:

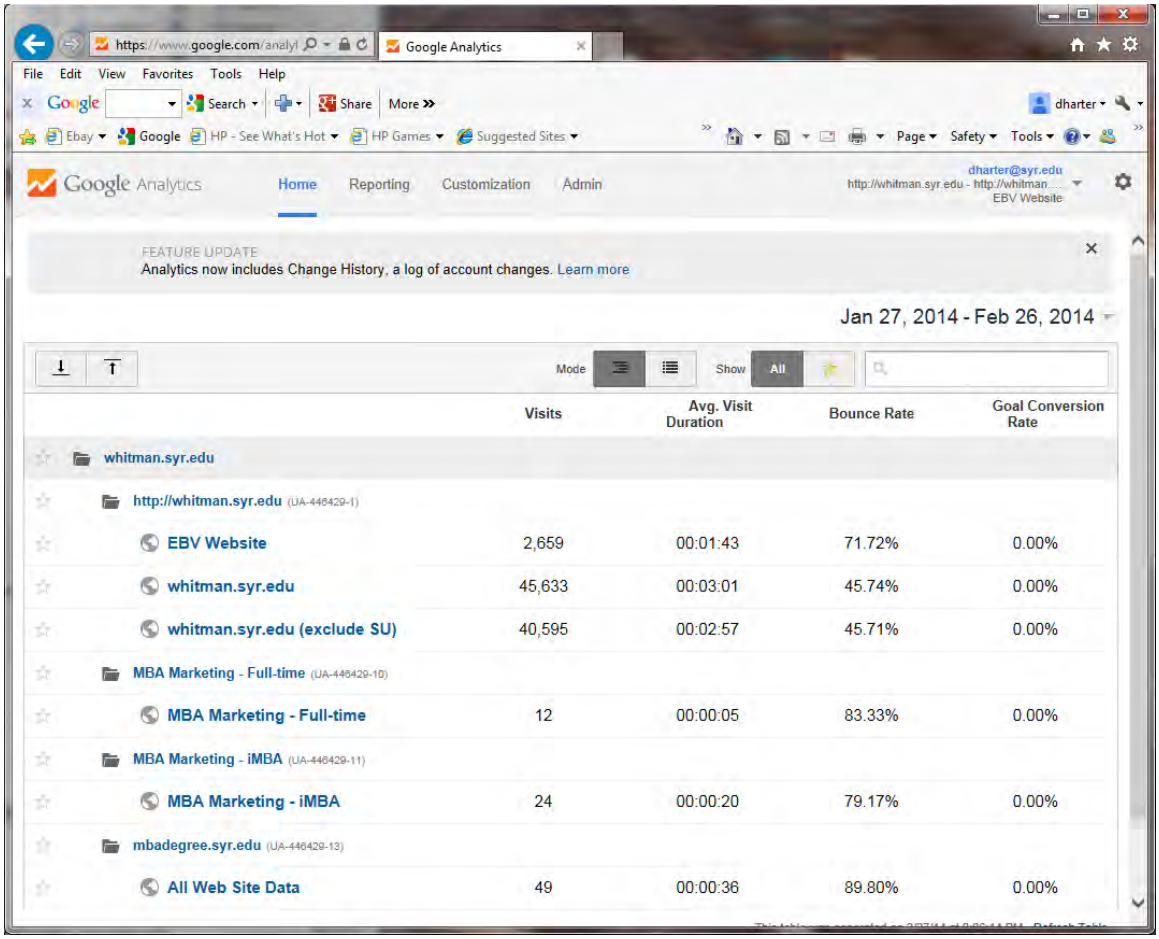
www.google.com/analytics




Click on Sign in, select Google Analytics, then enter your Syracuse email address and the password that you set up.



You should see a page similar to the one below. Note: The data is updated daily; your page will have a similar format but different data.




**Website traffic (Home)**

Note that the Home in the upper left should be highlighted in blue. If you do not see the page above, click on Home in the upper left corner of the screen.

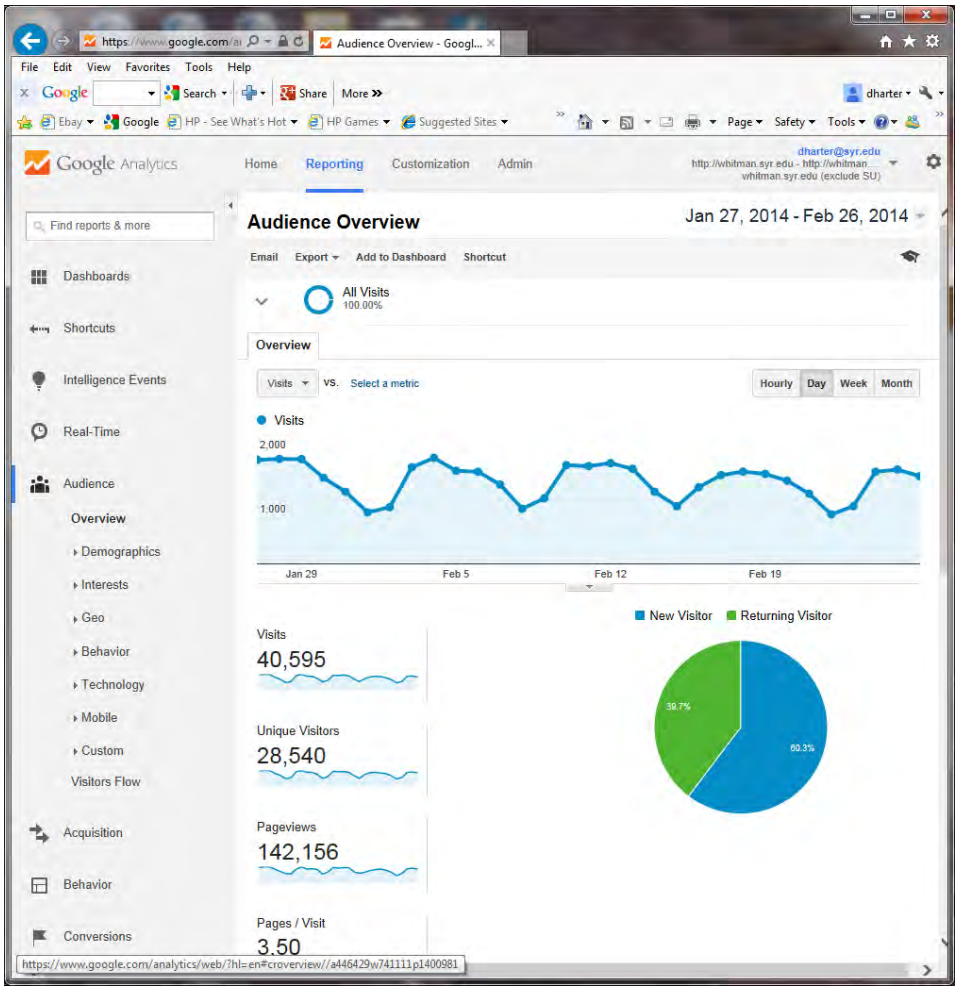
On the left are all the websites where you have Google Analytics access. Notice that there are multiple variations for the whitman.syr.edu website. One shows whitman.syr.edu and the next shows whitman.syr.edu (exclude SU). The first is all traffic to the Whitman site; the second excludes traffic from Syracuse University computers (lab computers, office computers, computers connected via the wireless system, etc.). Why is it important to distinguish between inside and outside traffic?

**Definitions**

- Sessions: number of times someone visited the site (does not count pages)
- Avg. Session Duration: time spent on website before leaving
- Bounce Rate: percentage of visitors who do not visit a second page
- Goal Conversion Rate: percentage of visitors who completed your goal

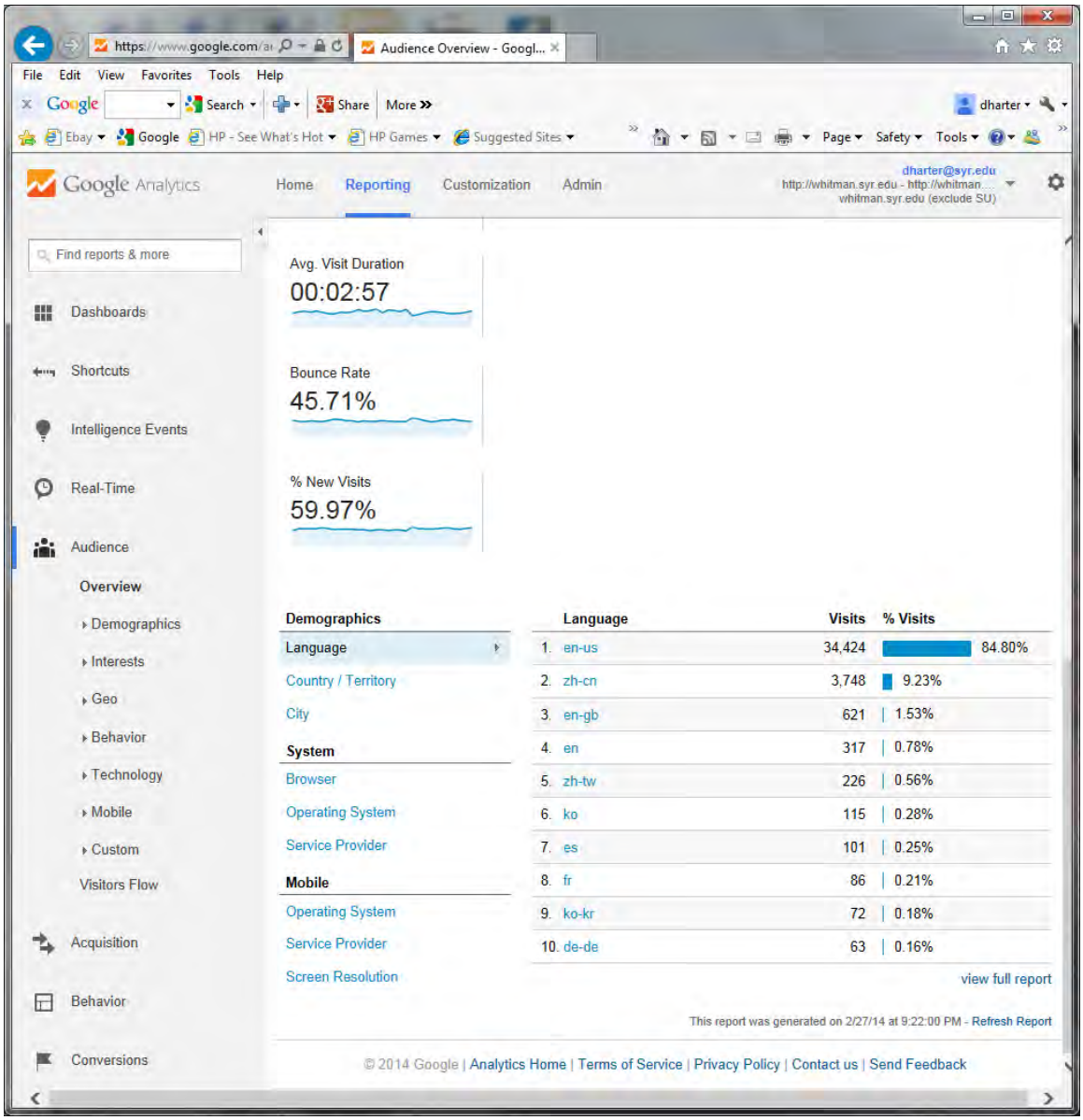
## Reporting

Let's examine traffic for whitman.syr.edu, excluding SU internal traffic. Click on "whitman.syr.edu (exclude SU)".




The default report is the Audience Overview. It includes:

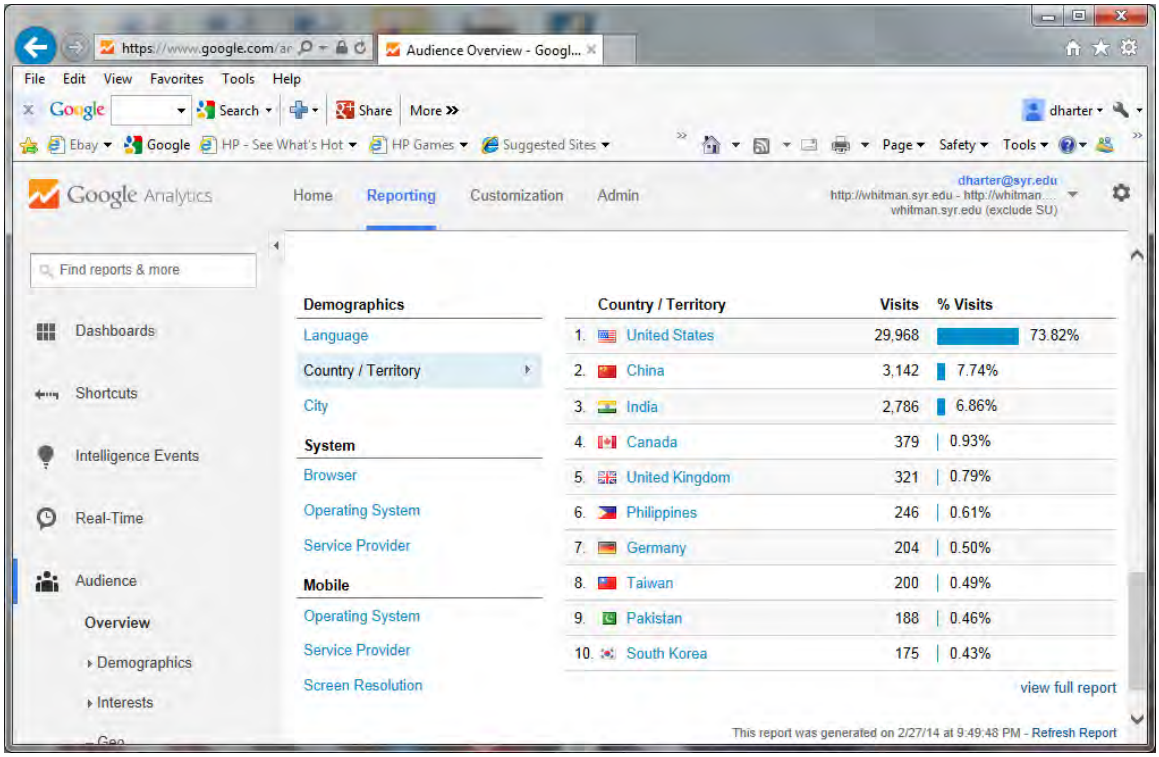
1. Date range in the upper right corner. We will change this later.
2. Option to view different categories of data. Click the drop-down button below Overview
   a. Default is Sessions
   b. Change to % New Sessions, Avg. Session Duration, Bounce Rate, Pages/Session, Pageviews, Unique Visitors, Sessions
3. Option to change to hourly, daily, weekly, and monthly graphs (to the right and above the graph). Change to weekly to observe the smoothing of data by week to see trends?
4. Graph of Sessions
5. Pie chart of new visitor versus returning visitor percentage
6. Number of Sessions
7. Unique users (tracked by your IP address on your computer)
8. Page views (how many individual pages were viewed
9. Pages per session (how many pages, on average, each person viewed)
10. Average visit duration (2 minutes, 57 seconds in this example [next page])
11. Bounce rate
12. New visits




Below the data is additional demographic, system, and mobile information. The default is language. In this example, 84.80% of the traffic to Whitman from outside of SU came from the United States, 9.23% from China. In this example, can you identify Taiwan, Korea, Spain, France, and Germany?

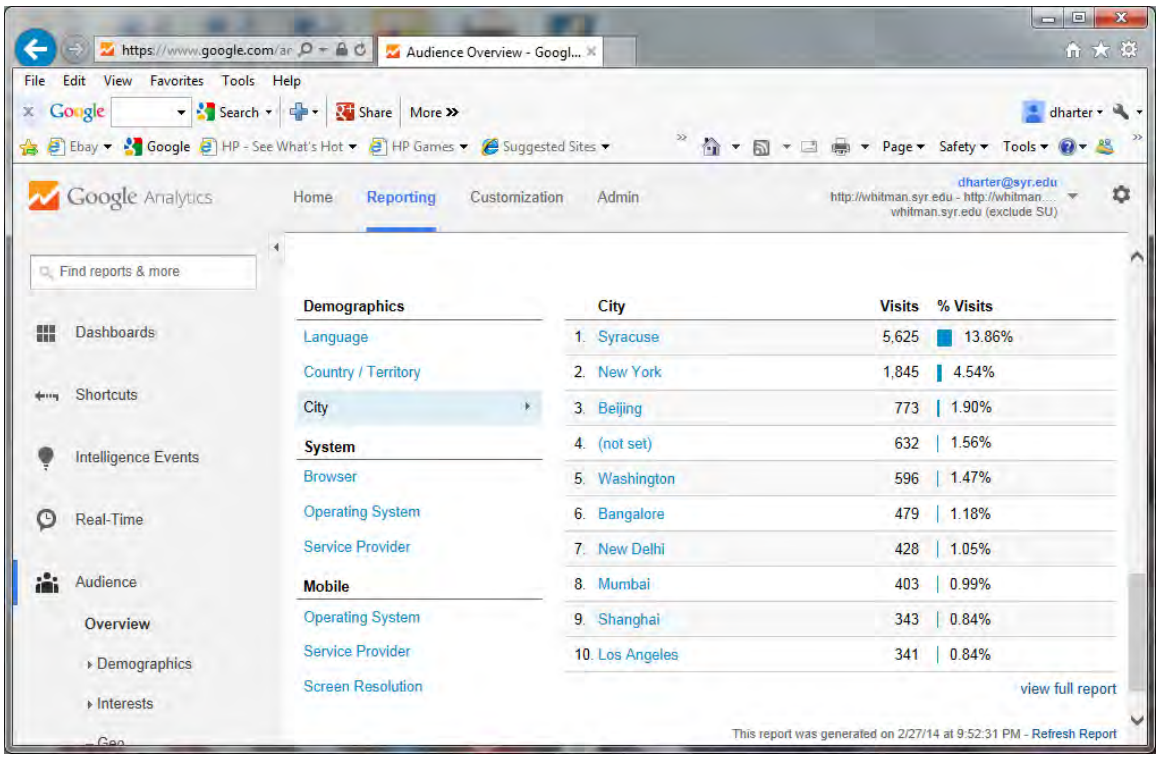
1. To view the entire report of all languages, click on “view full report” in the lower right corner.
2. Use the left and right arrows in the lower right corner to scroll through all the countries.

Click on Country/Territory. This presents the data in a format that is easier to review than languages.

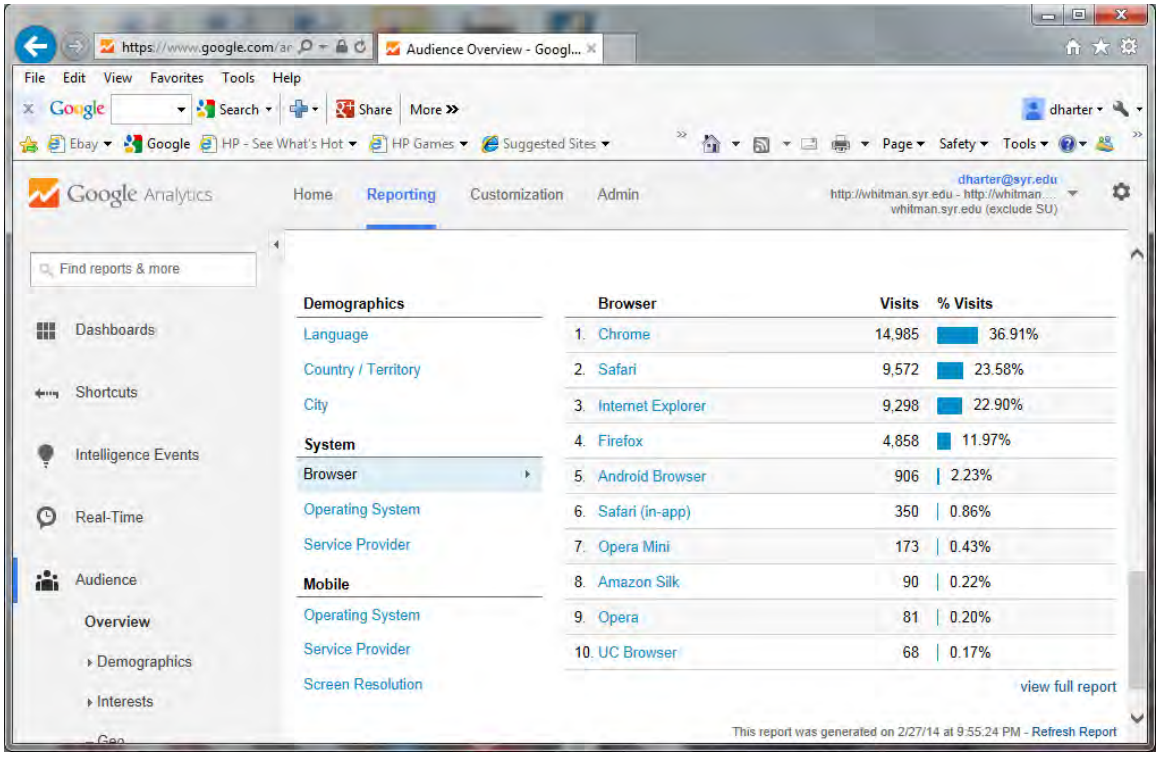



Why would you sometimes want to see Languages, other times Countries?

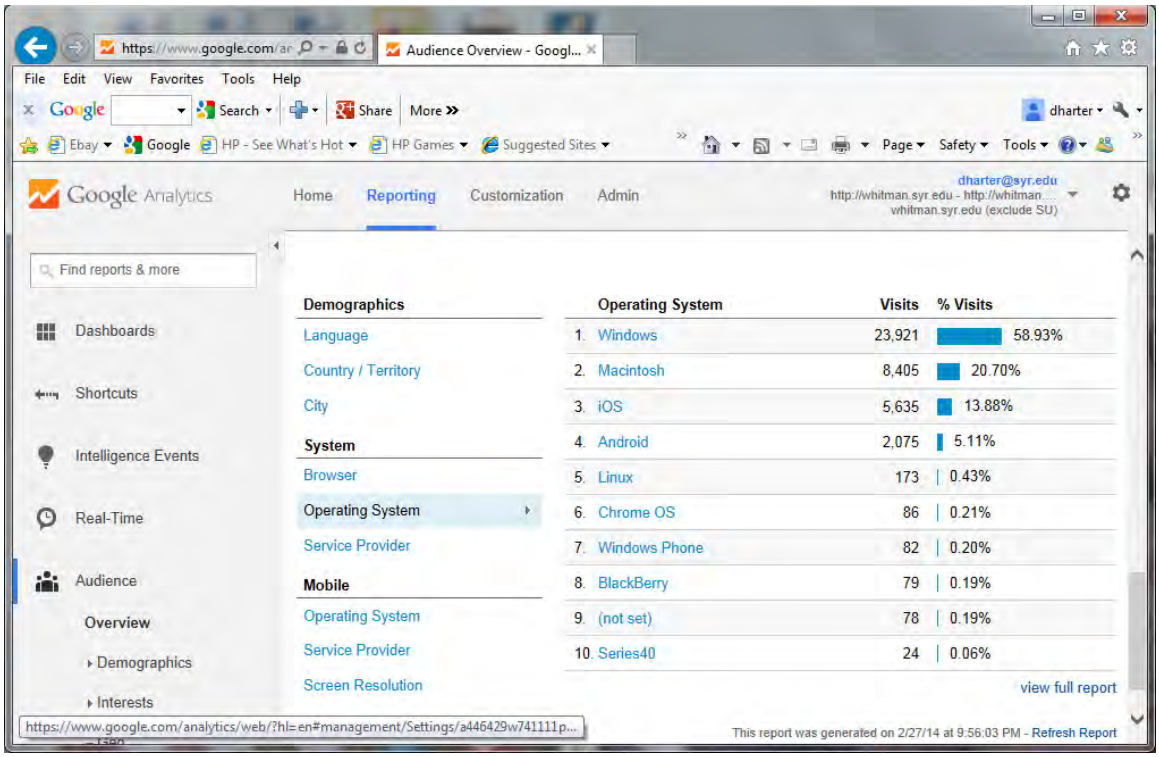
Next try City. Would this information help you in developing your recruiting strategy? Why is the Syracuse traffic so high, since we already excluded university and student computers on campus?



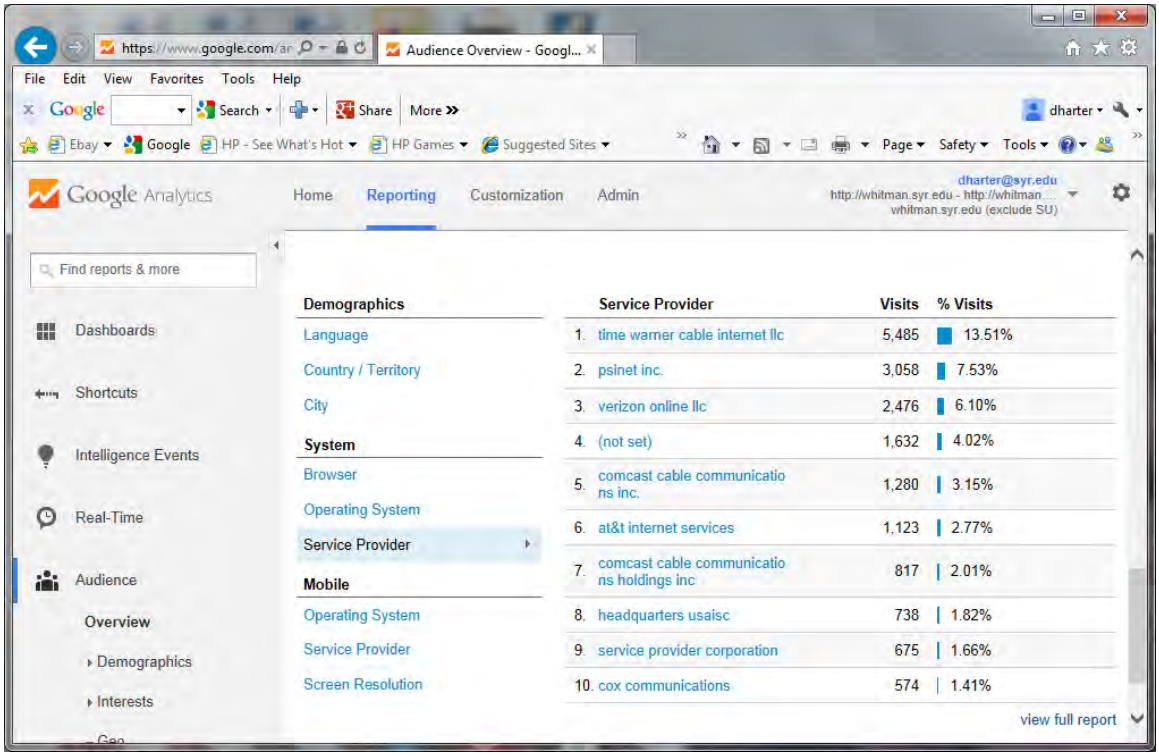
You can also view what computer system characteristics for visitors to the website. Browser identifies the web browser being used. Why is this important?



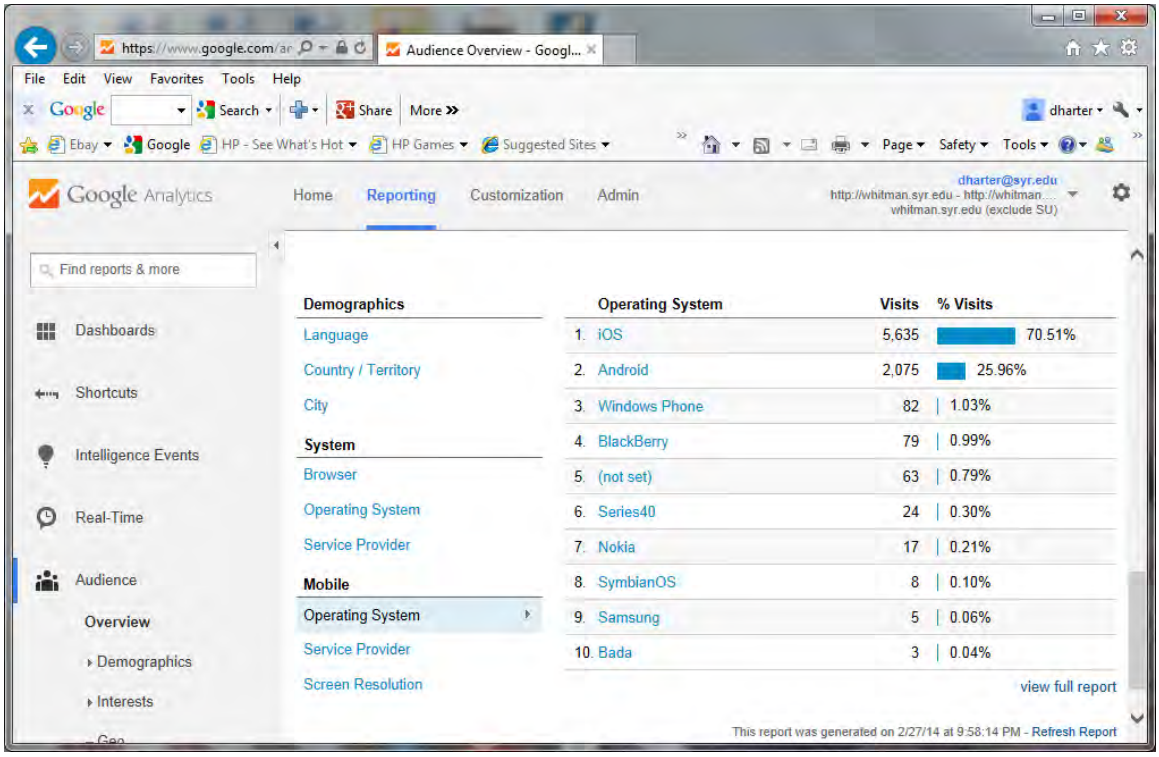

Operating system is next:



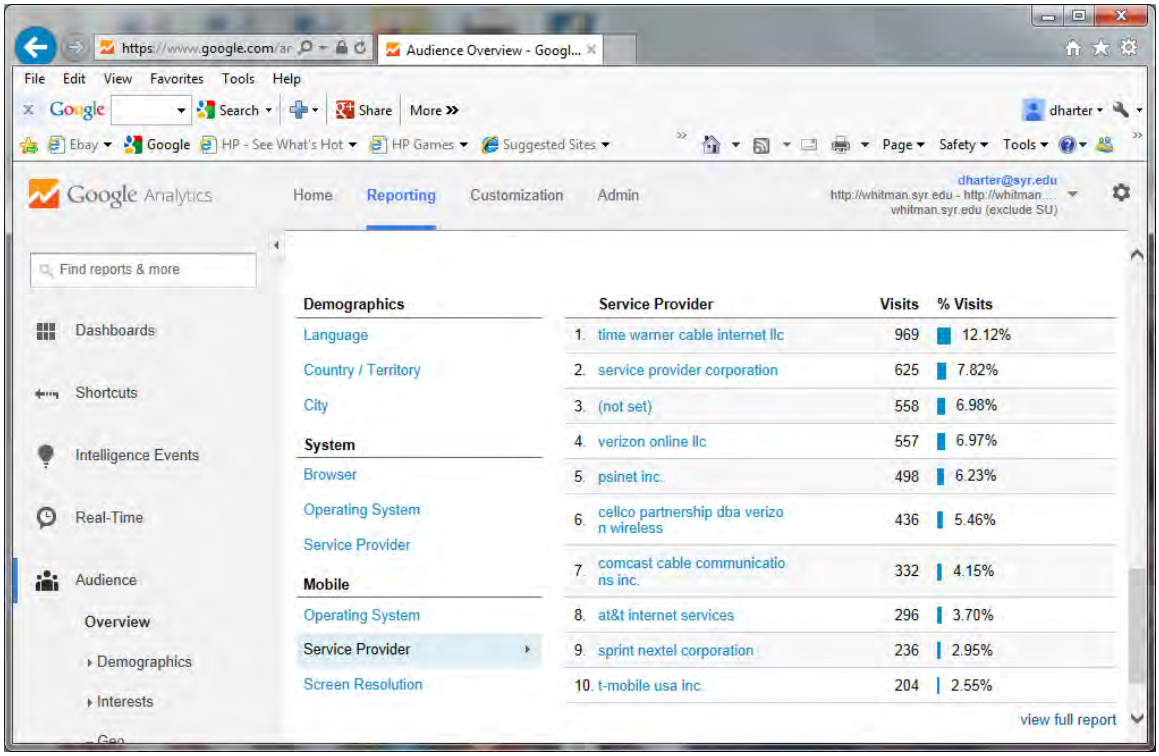
Next is Service Provider. Why is Time Warner Cable at the top?



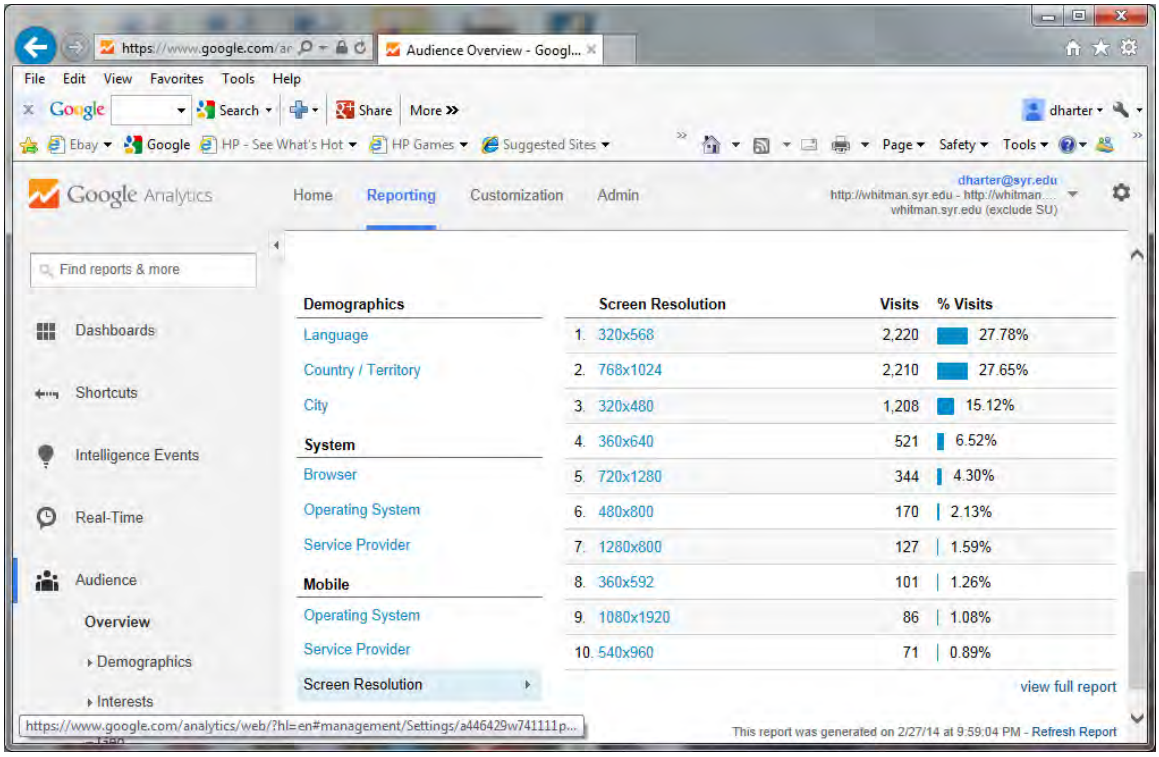

View the Mobile characteristics by first clicking on Operating System.



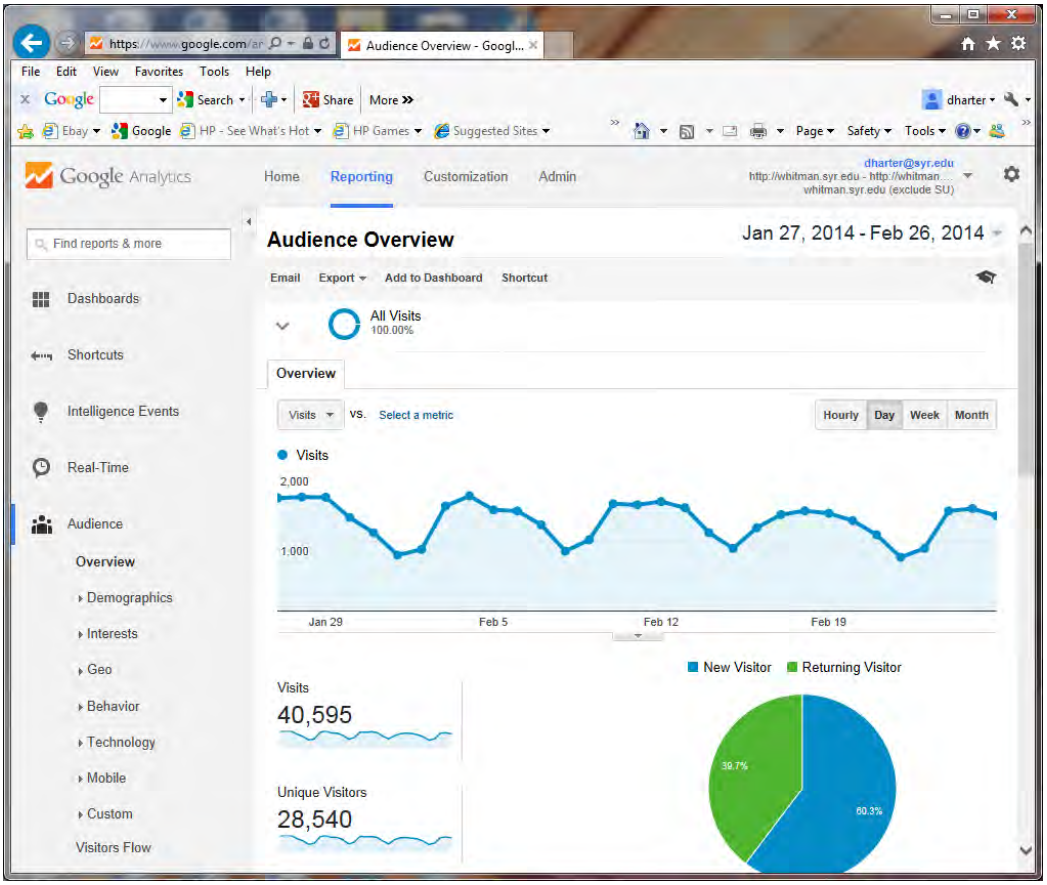
Next try Mobile Service Provider.




Finally, click on Screen Resolution. Why is this important? Does this change over time?



Scroll to the top of the screen. If you don't see the picture below, click on Home.



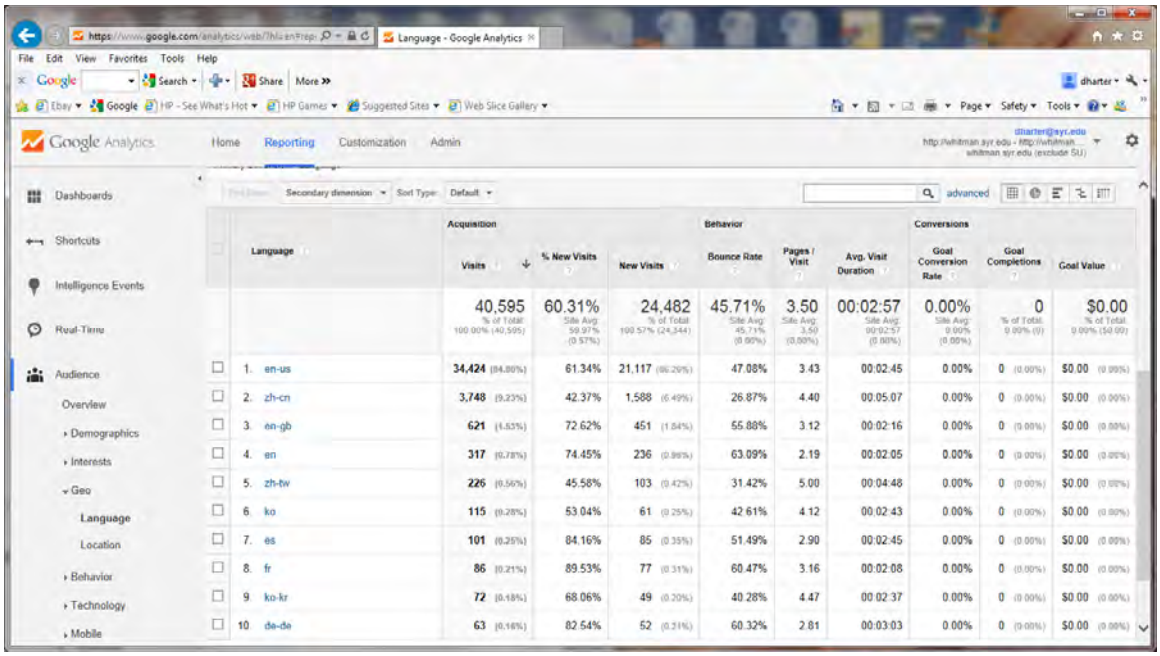

On the left side of the screen are different options under Audience. Click on each in the steps below.

1. Demographics - only activated by website administrator; recommend that privacy policy indicate that you are collecting this data
    a. Overview
    b. Age
    c. Gender
2. Interests - activated by website administrator and should be in privacy policy
    a. Overview
    b. Affinity Categories
    c. In-Market Segments
    d. Other Categories
3. Geo (geographic data)
    a. Language – more detail on language (example below)
    b. Location – more detail on language (example below)
4. Behavior
    a. New vs. Returning
    b. Frequency & Recency
    c. Engagement
5. Technology
    a. Browser & OS
    b. Network
6. Mobile

    a. Overview
    b. Devices
7. Custom
    a. Custom Variables
    b. User Defined
8. Visitors Flow

**Geographic - Language**

Within Audience, click on Geo, then Language.

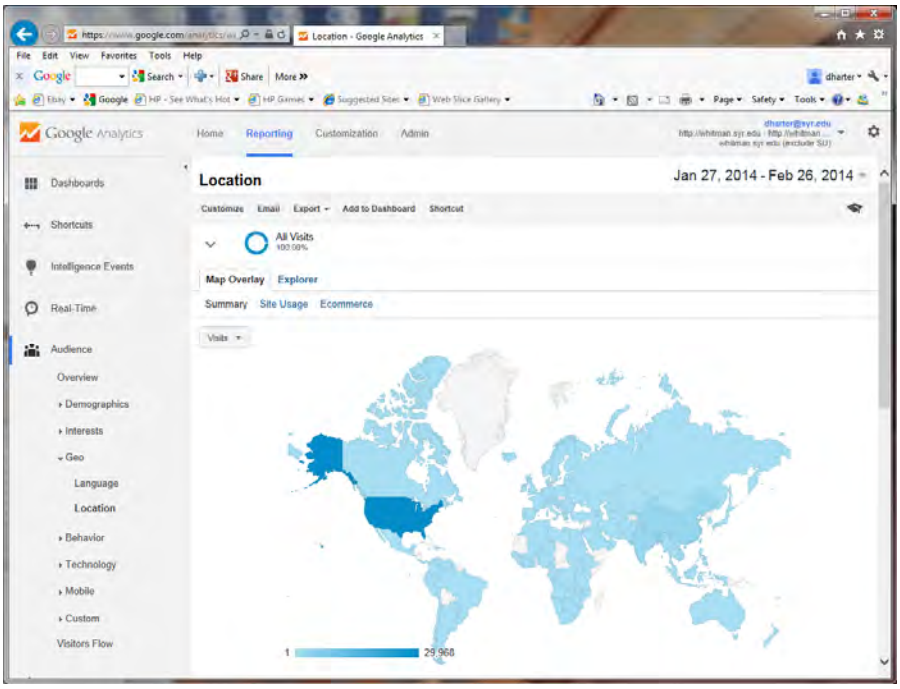



This displays more detailed data by language, including:

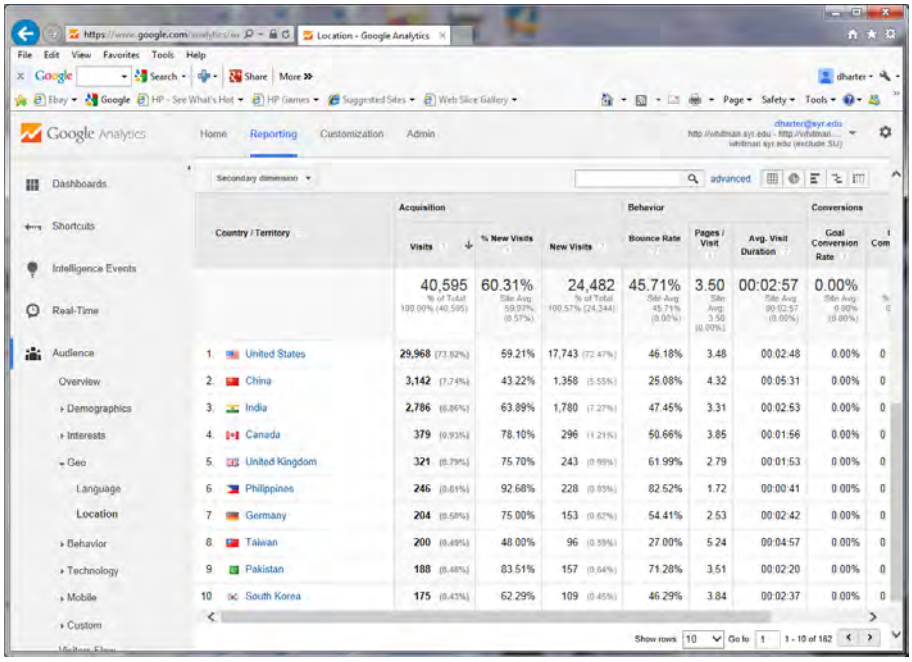
1. Acquisition
    a. Visits
    b. New visits and % new visits
2. Behavior
    a. Bounce rate
    b. Pages/visit
    c. Avg. visit Duration
3. Conversion
    a. Goal conversion rate
    b. Goal completions
    c. Goal value

**Geographic – Location**

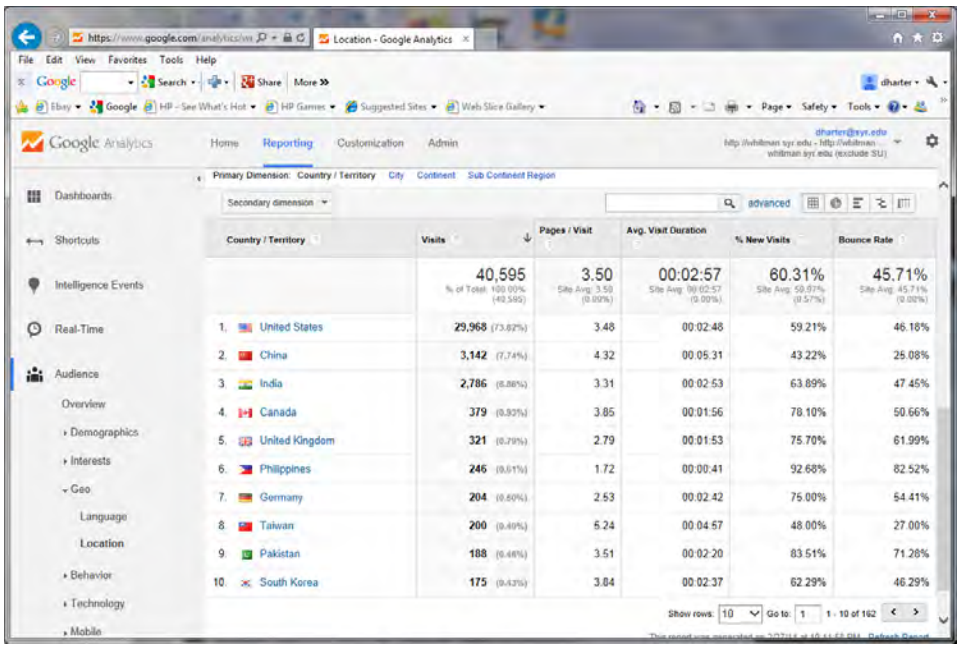
Under Audience, click on Geo, Location. This displays traffic on a Map Overlay (notice the tabs Map Overlay and Explorer above and to the left of the picture).



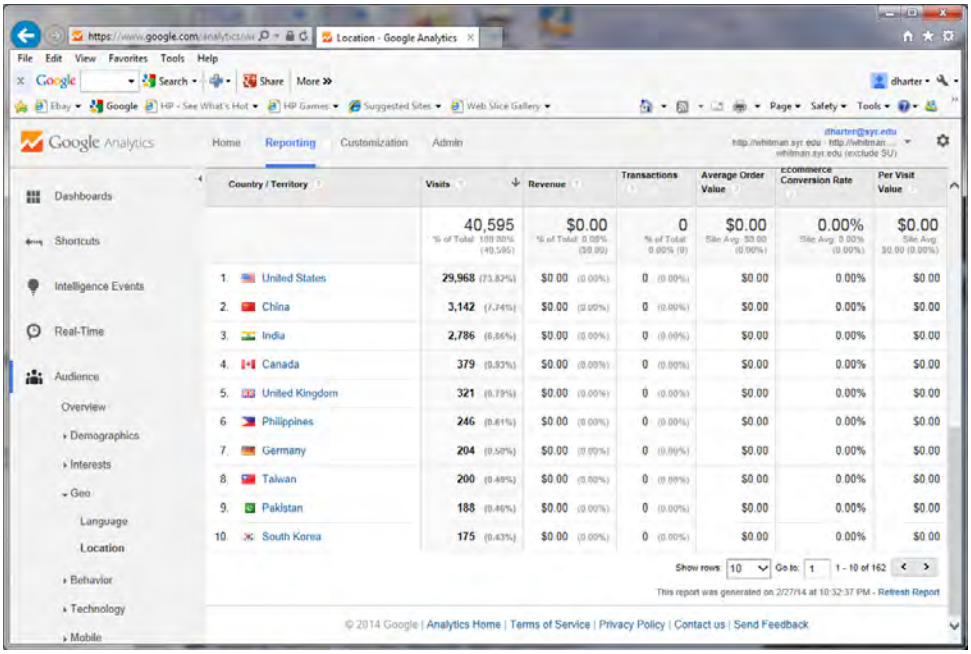

Move your cursor over Brazil. What happens? Next, scroll down to see the detailed data. It includes Acquisition, Behavior, and Conversion data.



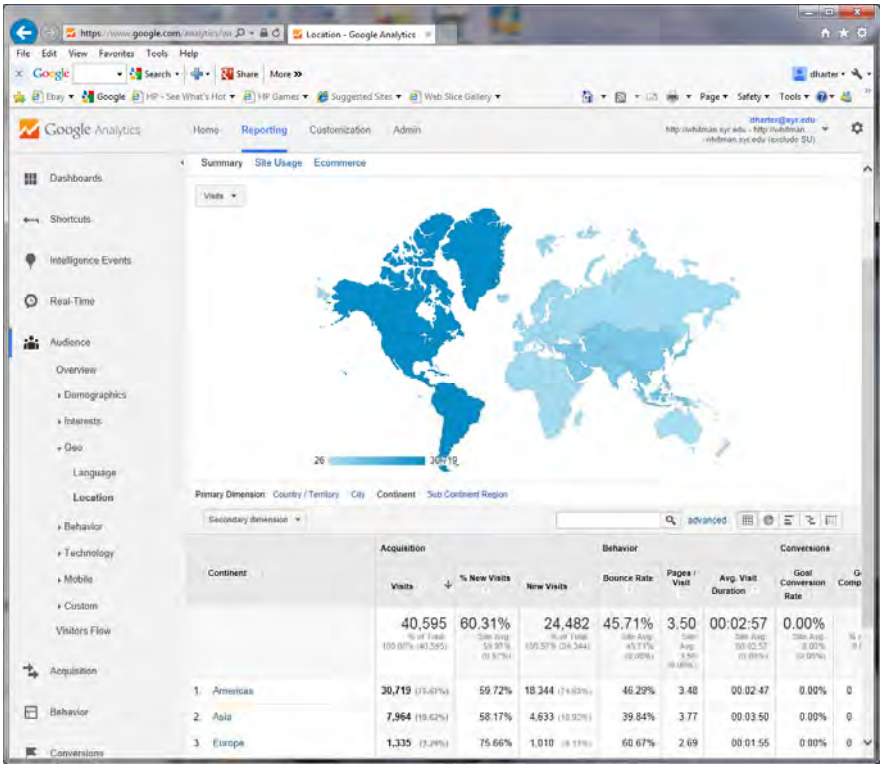
Scroll back to the top of the page. Under Map Overlay, click on Site Usage, then scroll to the bottom. The view is now Visits, Pages/Visit, Avg. Visit Duration, % New Visits, and Bounce Rate.




Scroll back to the top of the page. Under Map Overlay, click on Ecommerce. Now we can see Visits, Revenue, Transactions, Average Order Value, Ecommerce Conversion Rate, and Per Visit Value. Why is everything zero?

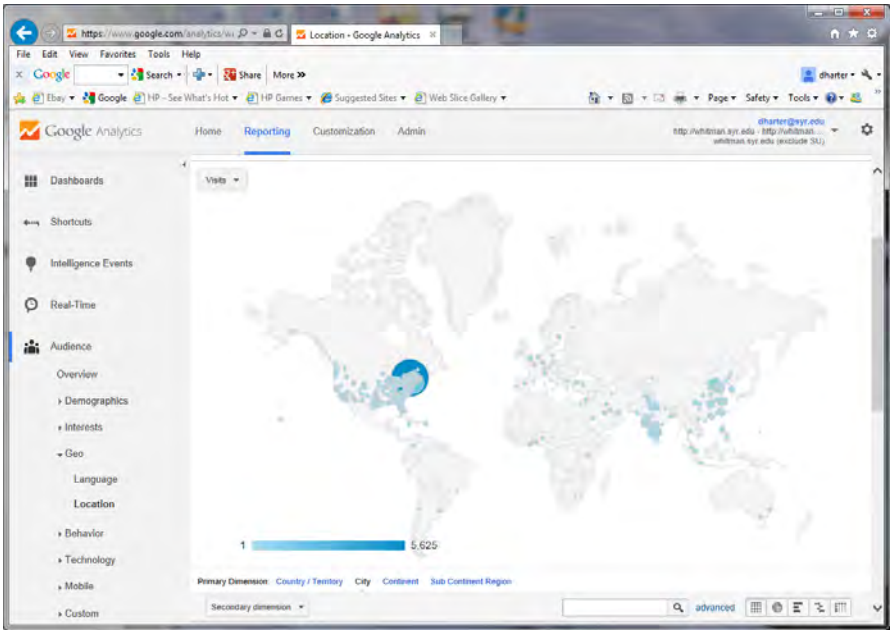


Scroll back to the top. Under Map Overlay, click on Summary. Scroll down to just below the map and click on Continent.

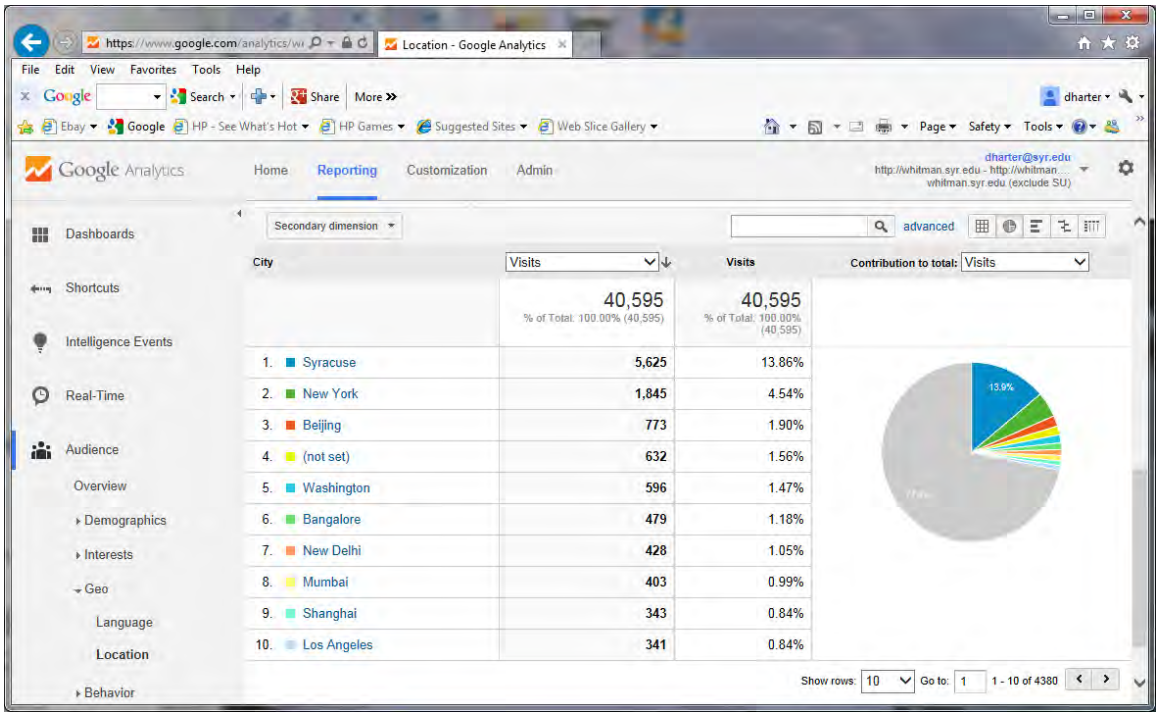



Click on Sub Continent Region. How is it different?

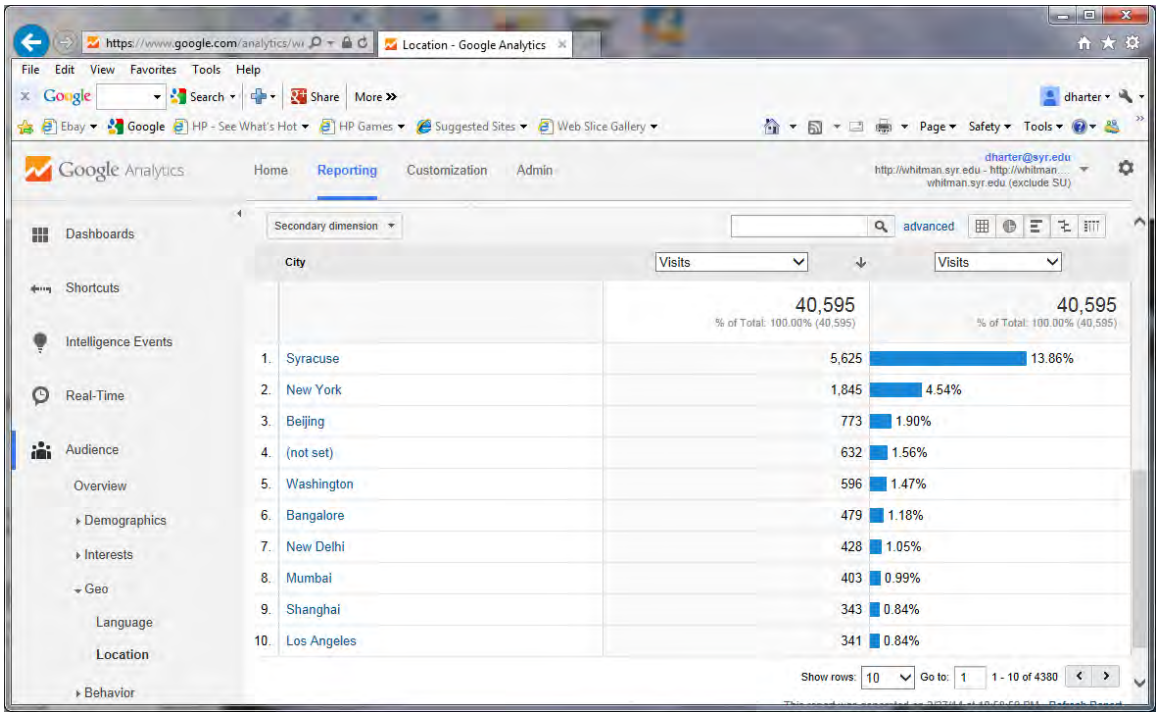
Now click on City. Scroll down to see the traffic by city. Put your cursor over a city.



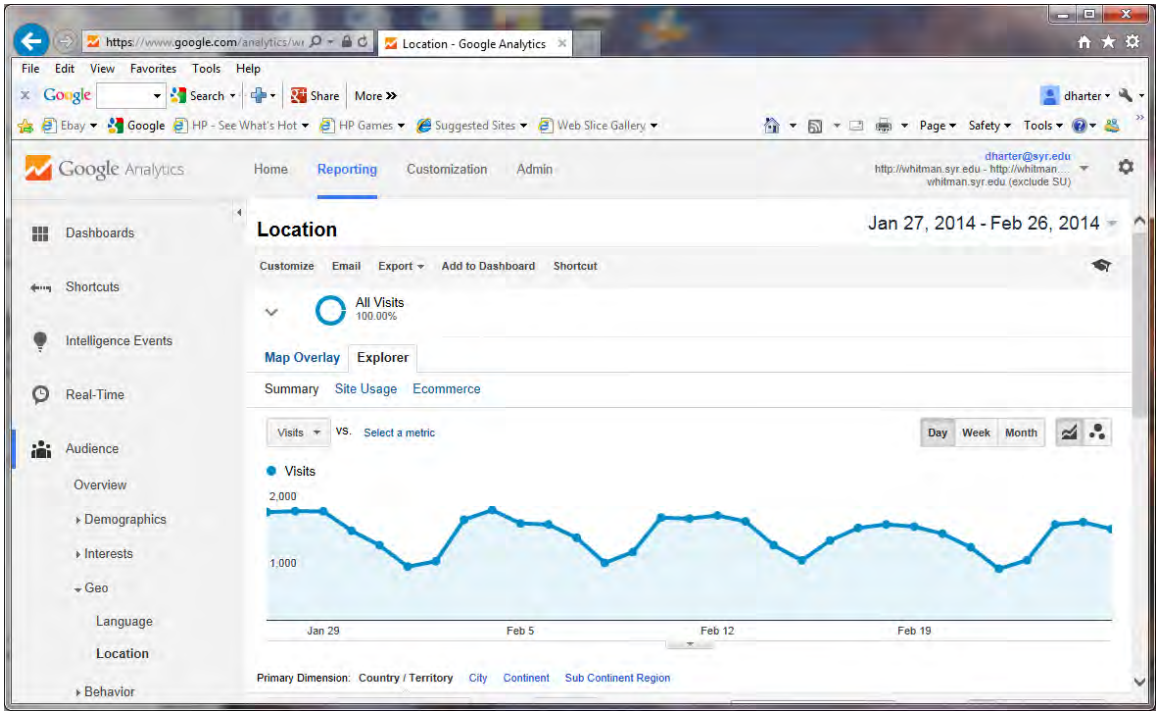
To the lower right of the picture, and to the right and above the data, are five options. The five options are Data, Percentage, Performance, Comparison, Pivot. We have been viewing data; now click on Percentage.



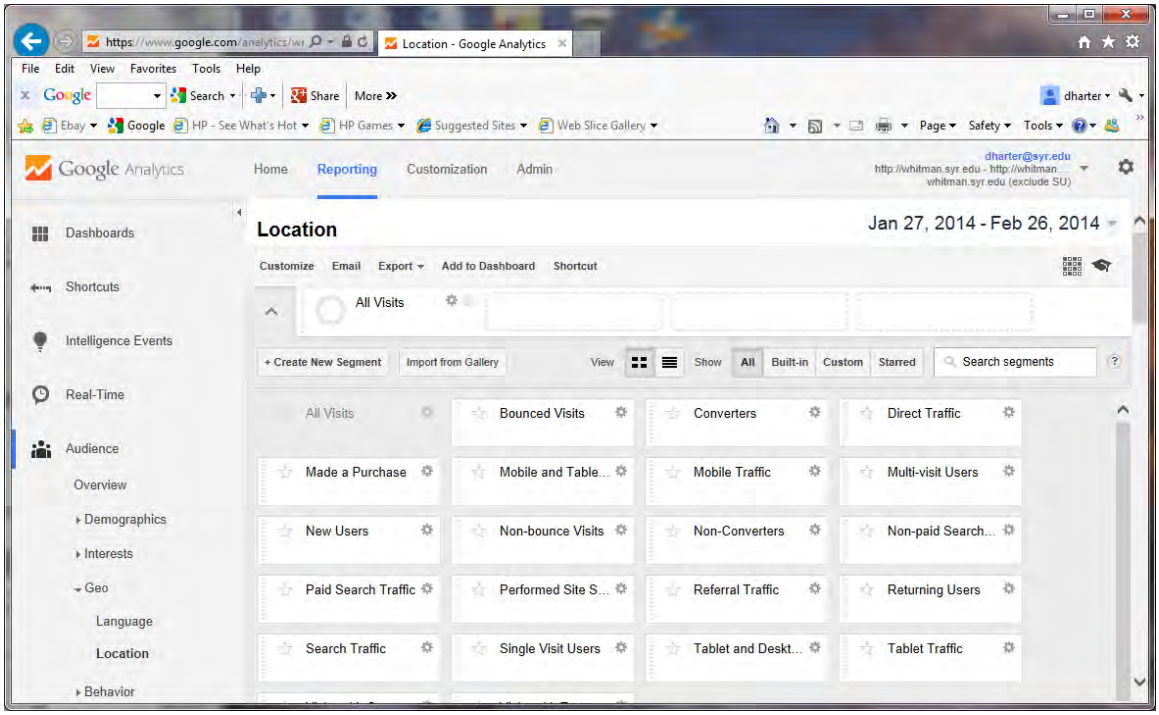

Next, click on Performance.



Scroll back to the top of the page. Instead of Map Overlay, click on Explorer (to the right of Map Overlay).

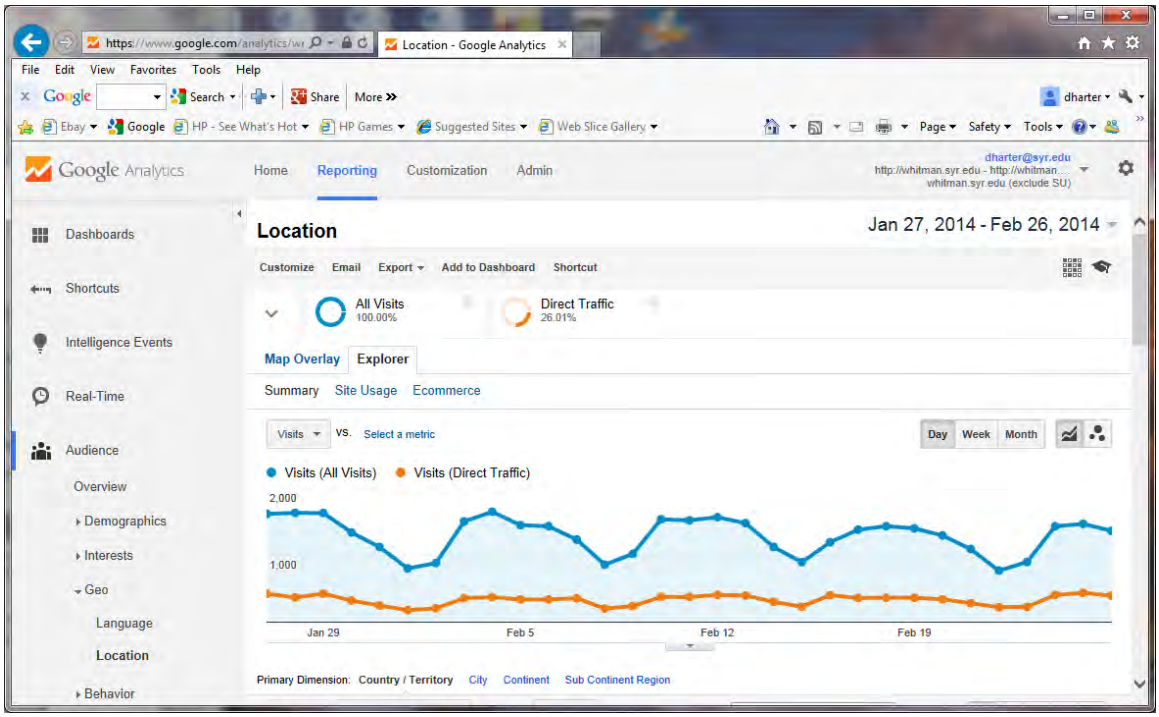



Notice that there is a small down arrow above the words Map Overlay. Click on the down arrow. It changes to an up arrow and displays more options.

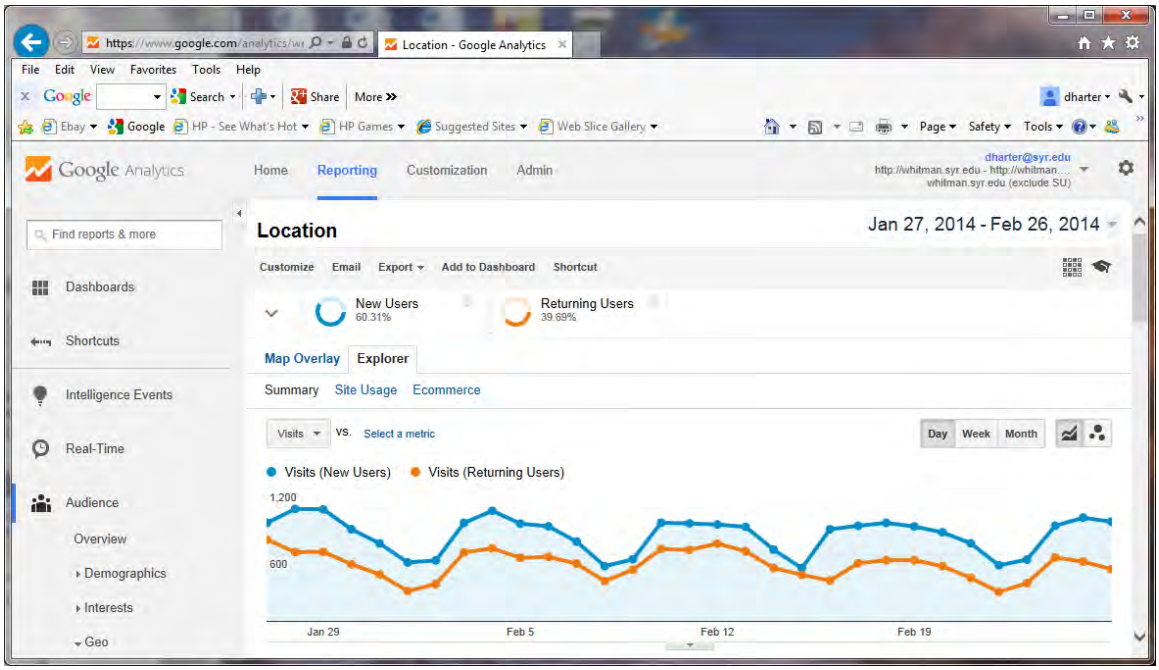



Only “All Visits” is shaded. Click on Direct Traffic, then click on the up arrow.

We can now see direct traffic versus all visits. Direct traffic is from individuals who type in your web address or who click on a bookmark. All traffic includes visits directed from advertisements and other websites. If someone first visited syr.edu, then clicked through to whitman.syr.edu, would that be in direct traffic or all traffic?

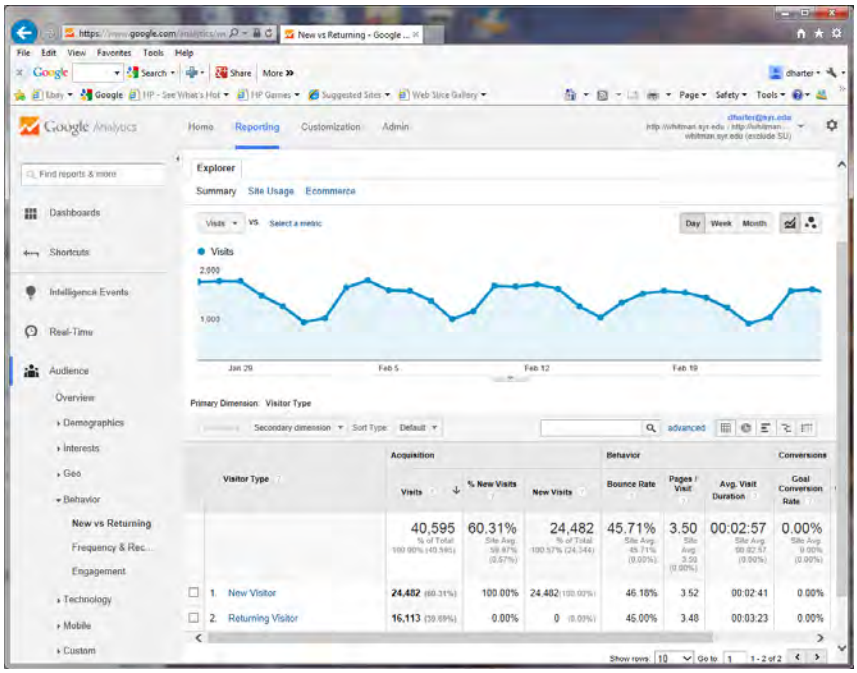



You could also look at New Users versus Returning Users. Click on the down arrow, click on All Visits to turn it off, click on New Users and Returning Users, then click on the up arrow.

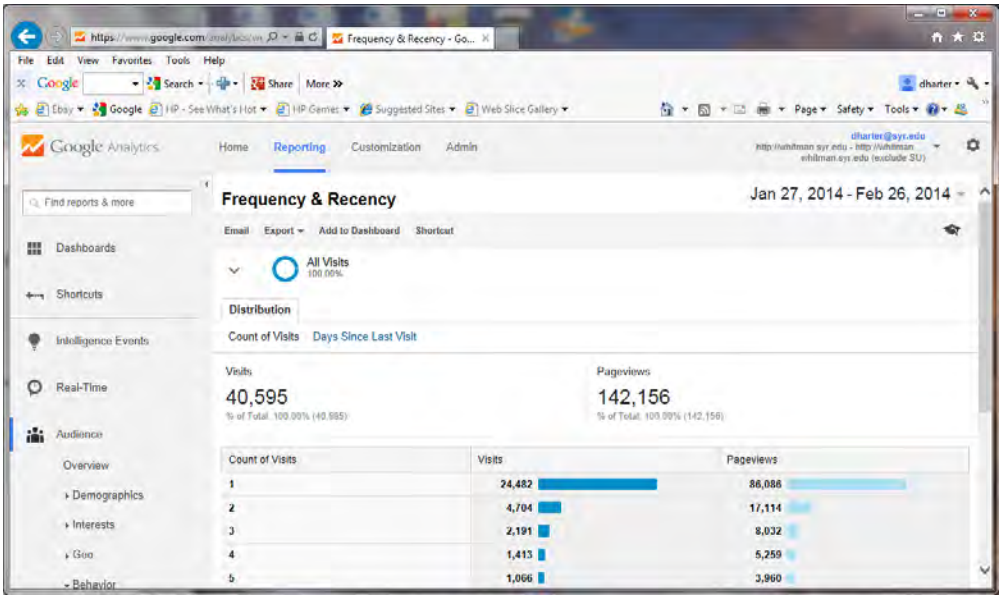


**Behavior**

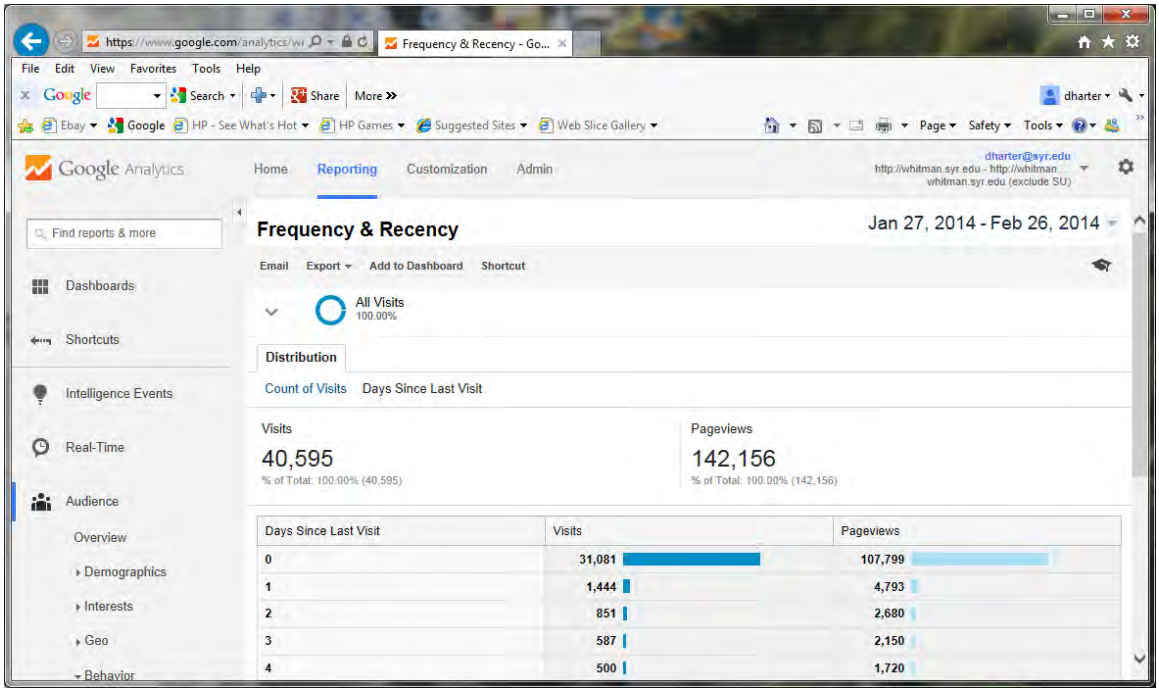
On the left side of the screen, now click on Behavior. Reset options to All Visits. Click on New versus Returning.



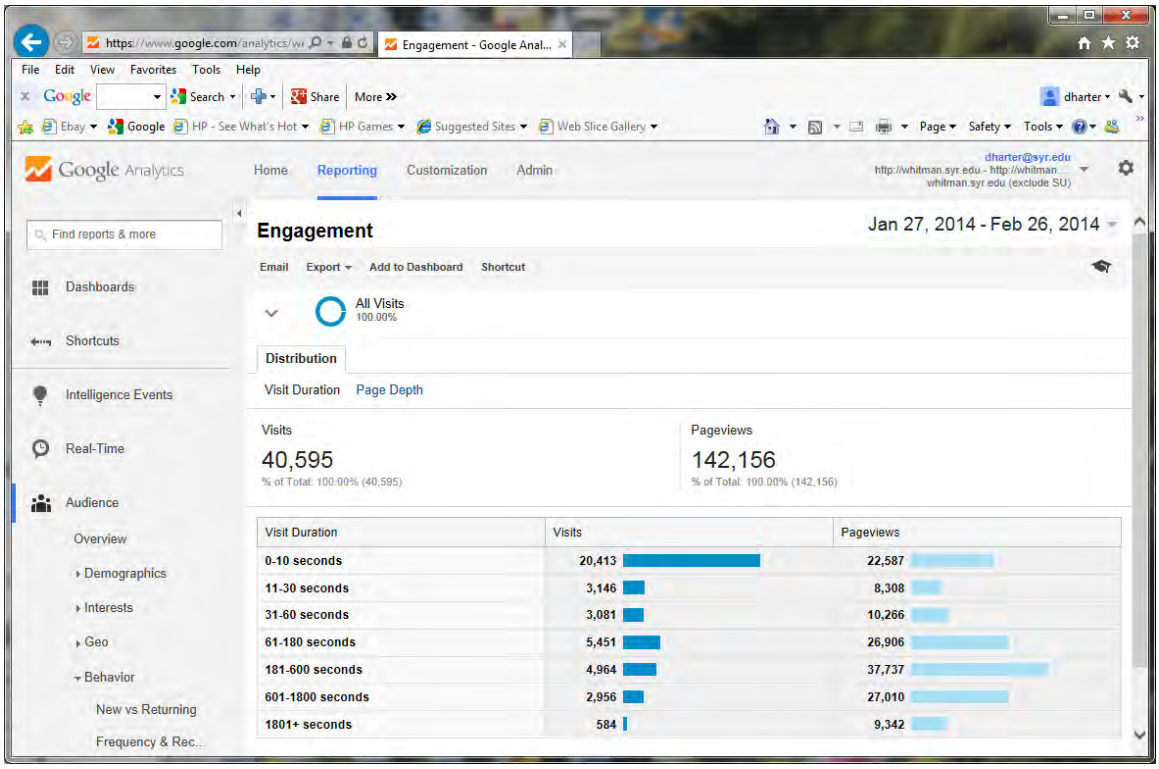

Now try Frequency and Recency. Frequency is how often someone visits. Recency measures the time since their last visit. In the example below, there were 24,482 people who visited the site once. How many people visited the site five times during this period?



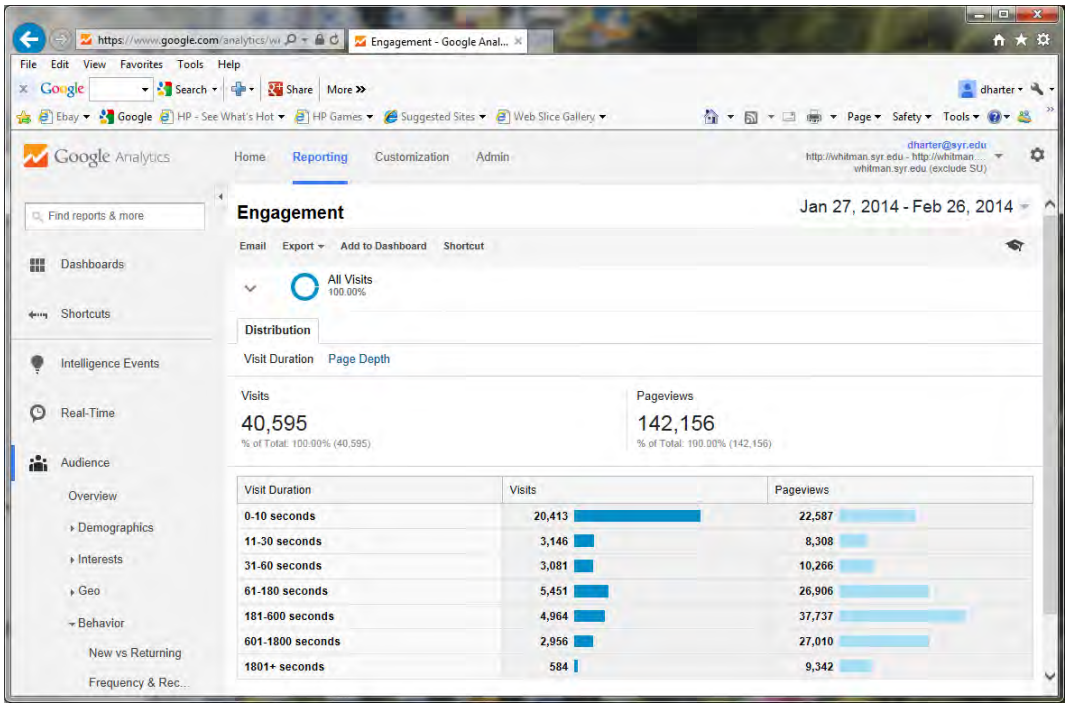
Scroll to the top of the page. Under the tab Distribution, you have been viewing Count of Visits. Click on Days Since Last Visit.




Go back to the left of the screen, and click on Behavior, then Engagement. This displays the length of time a visitor is on your website. What is the difference between Visits and Pageviews?

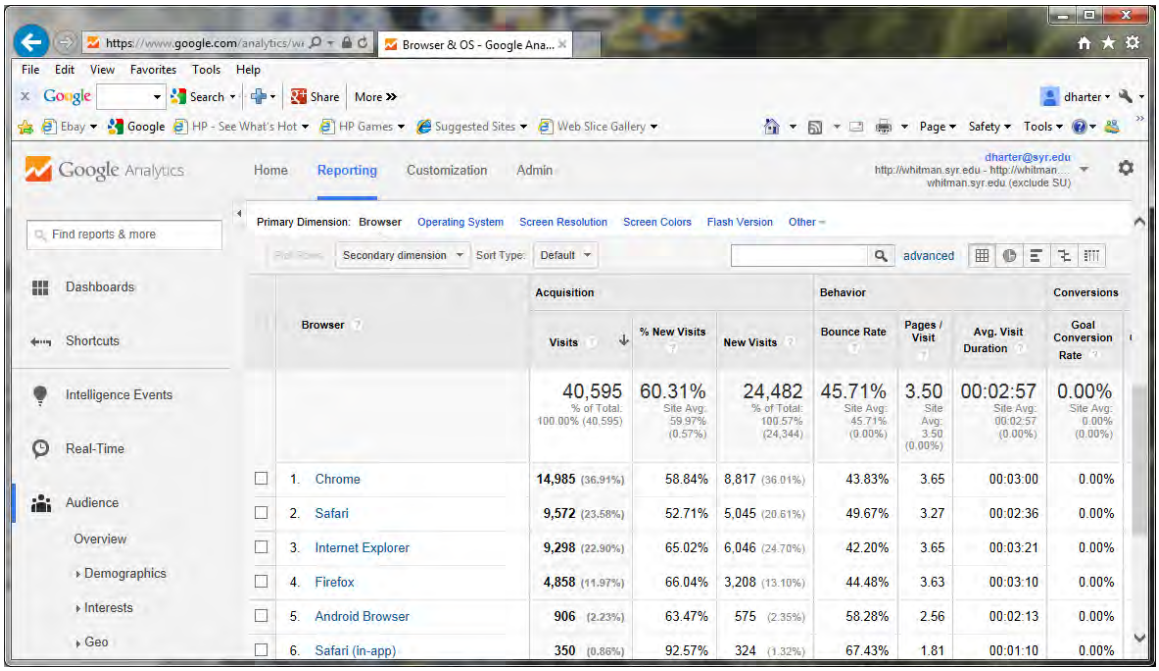


Scroll to the top of the page. Under Distribution, click on Page Depth.



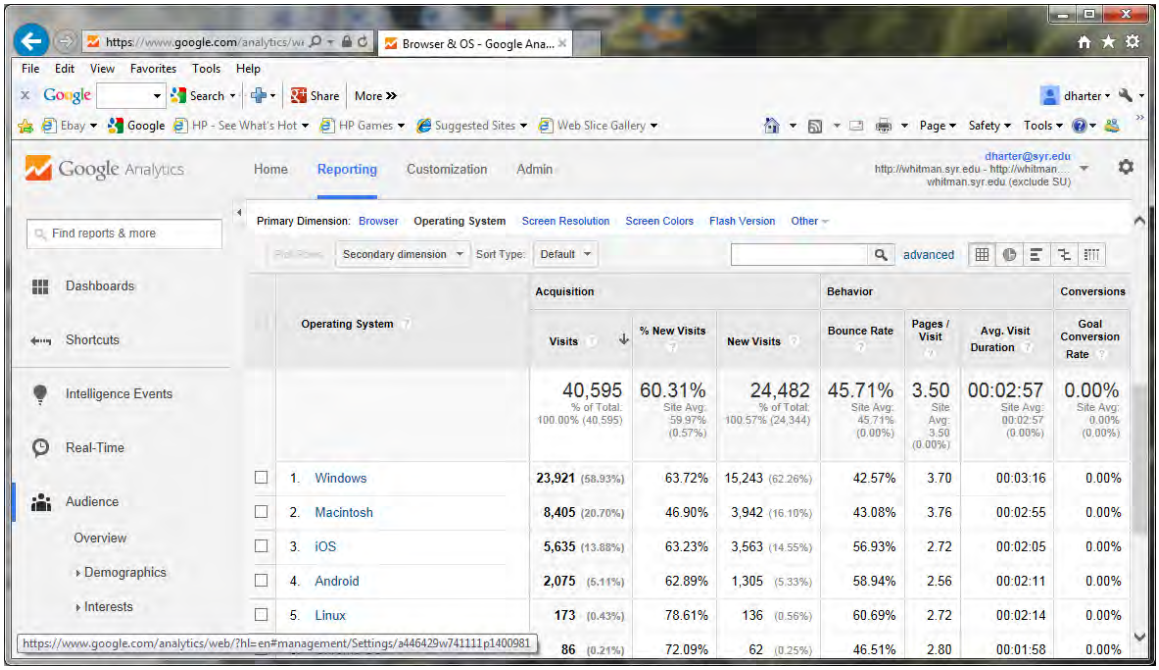

## Technology

Click on Technology on the left of the screen, then Browser.

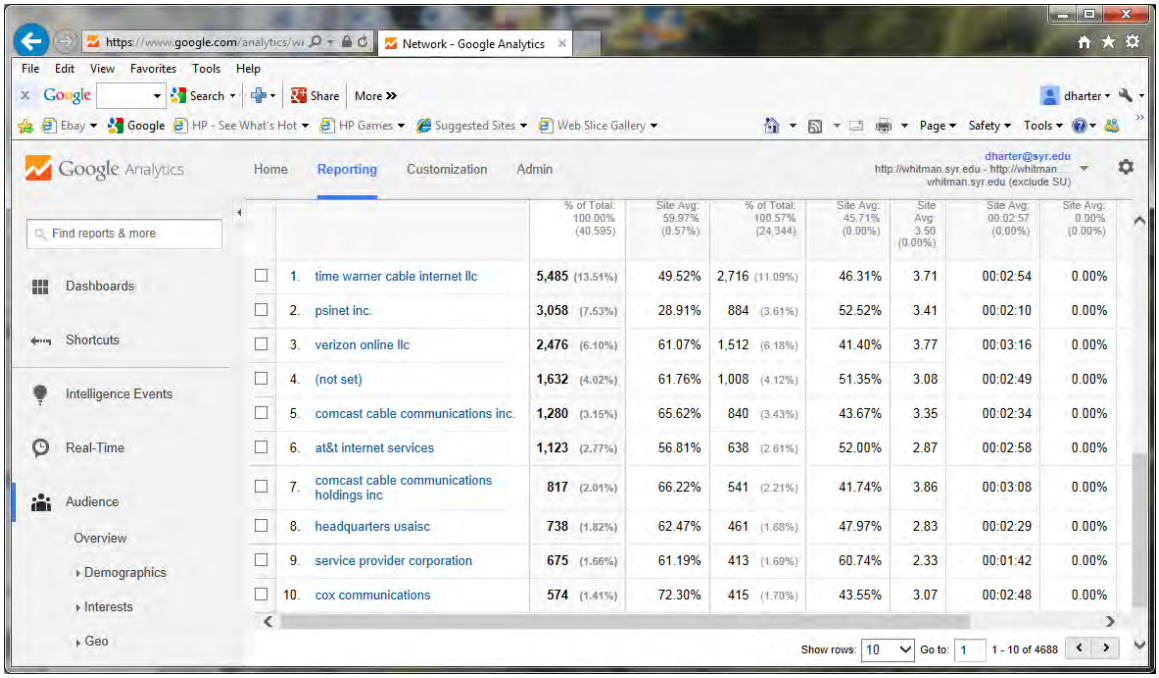



Why is it important to know your customer's browser?

Now find Primary Dimension on your screen (under the graph). Change to Operating System, Screen Resolution, Screen Colors, Flash Version, and Other.

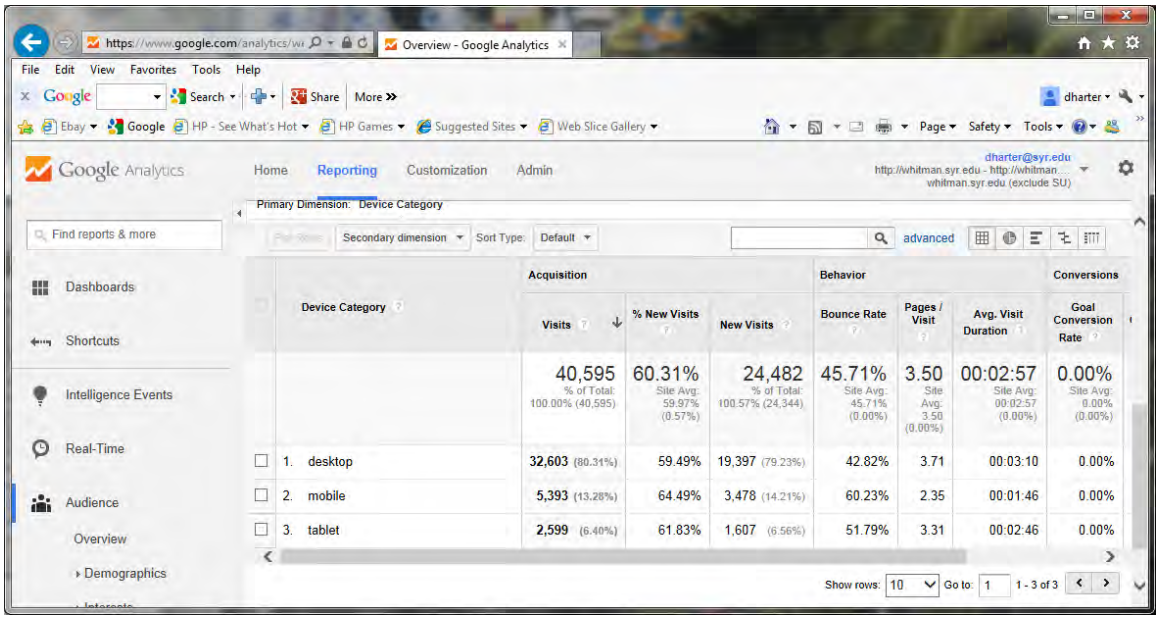



Click on Technology, Network.

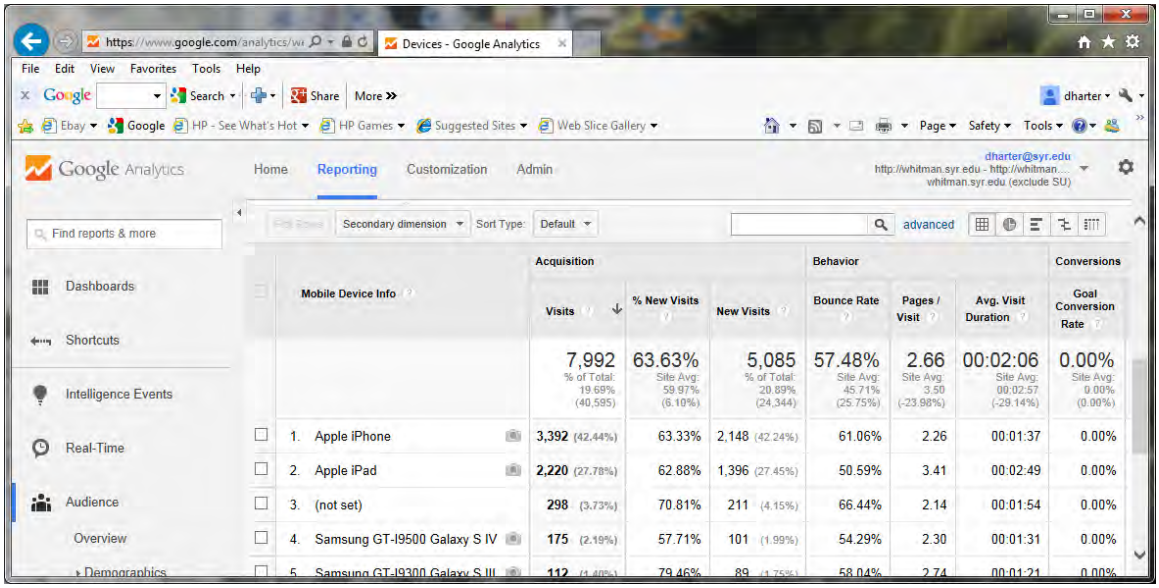


## Mobile

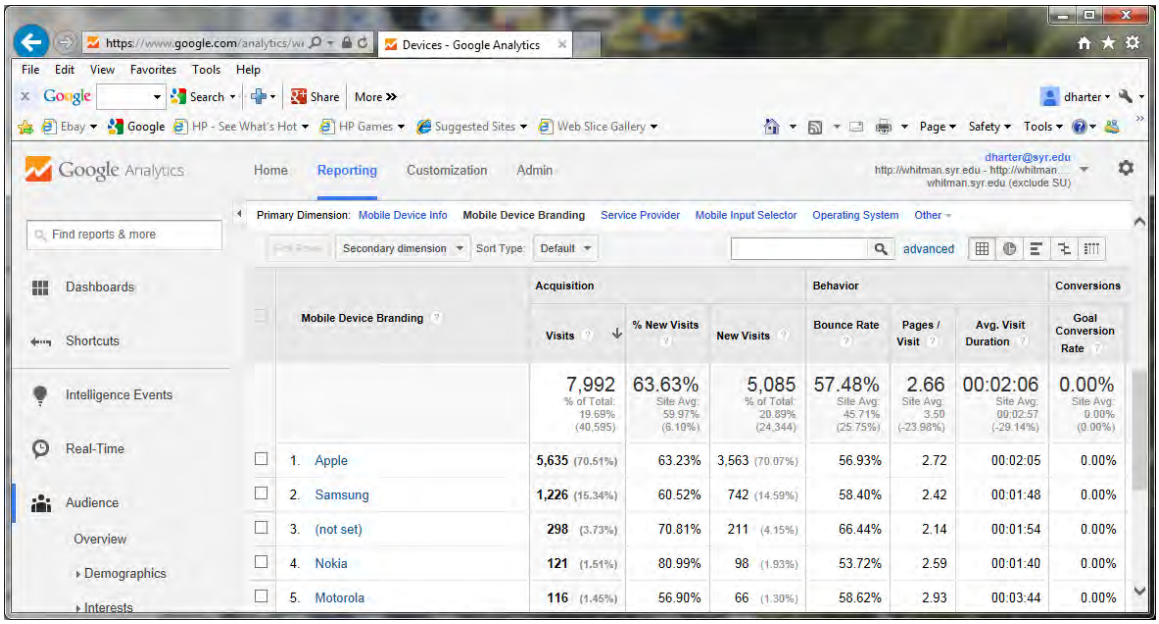
The Mobile option on the left of the screen allows you to distinguish between desktop, mobile, and tablet traffic. Click on Mobile, Overview.



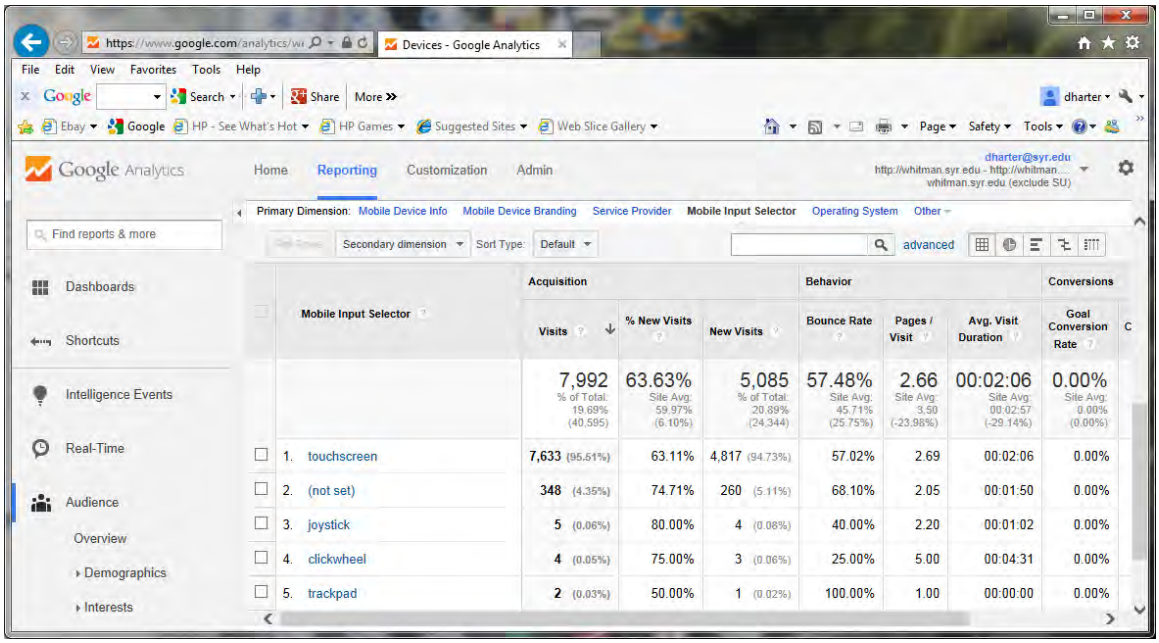

Mobile, Devices distinguish between different mobile devices used to access your website.



In Primary Dimension, change to Mobile Device Branding to roll up to the brand level.

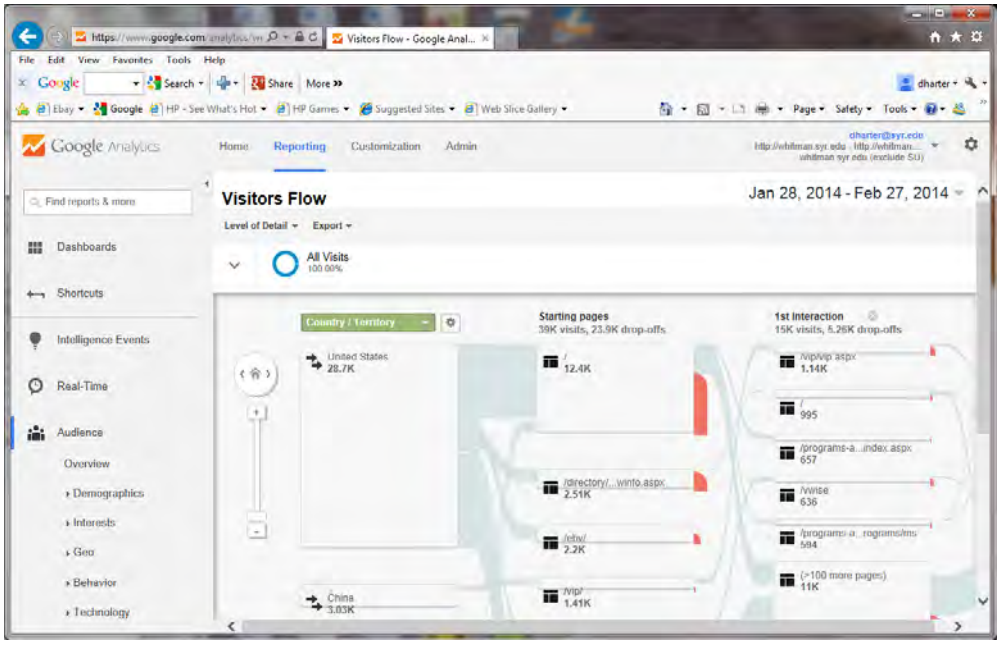



Next try Mobil Input Selector.



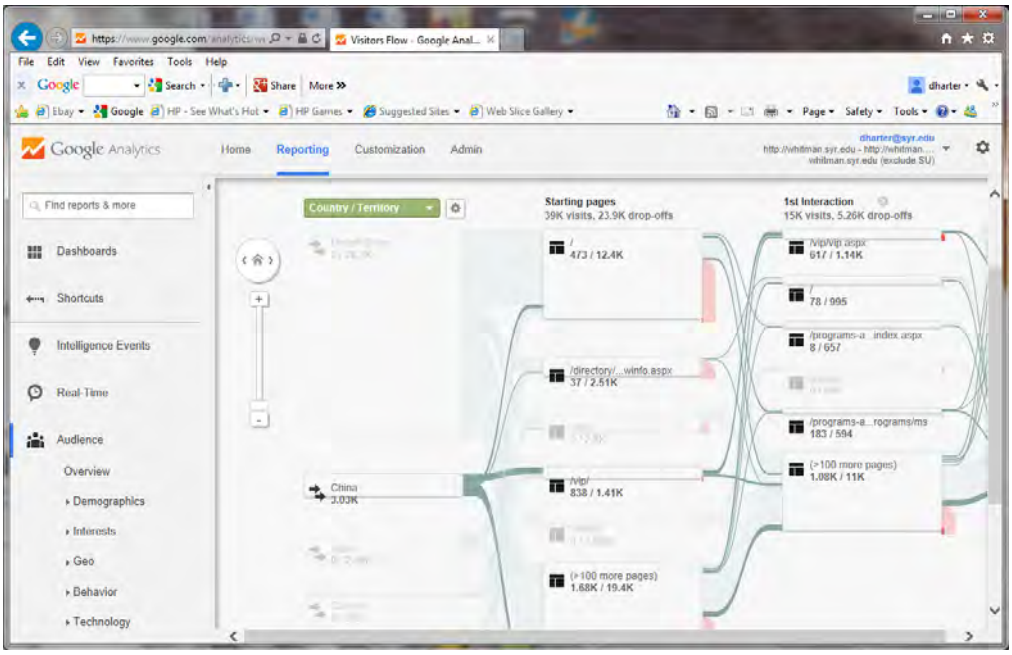
### Visitors Flow

The final section within Audience is Visitors Flow. Click on Visitors Flow on the left of the screen.

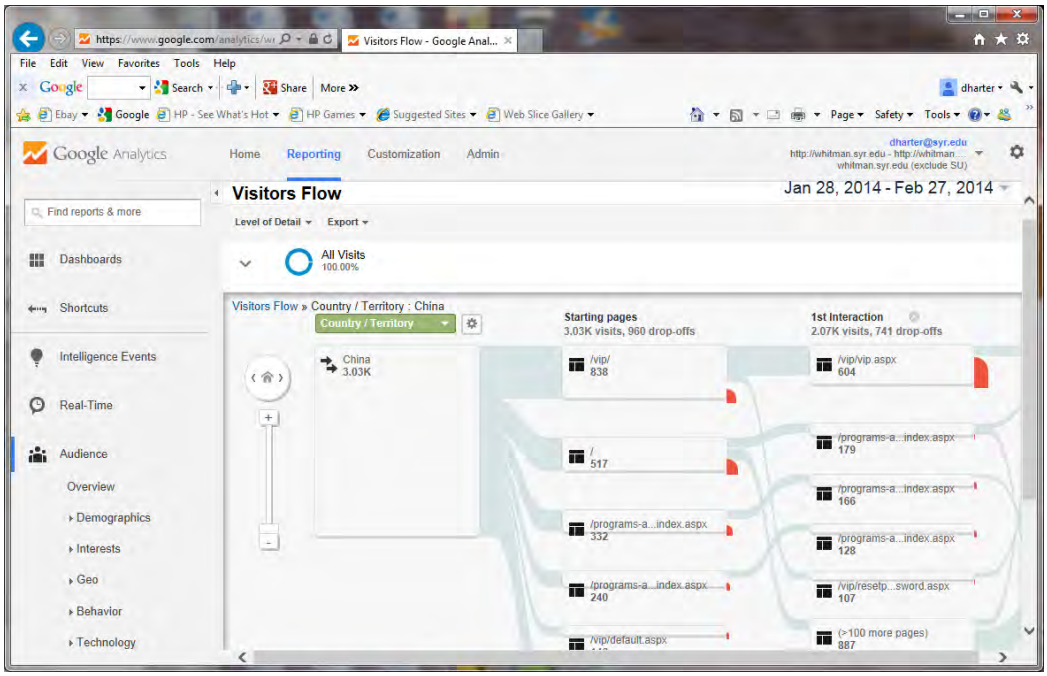



Traffic flow is highlighted in blue. Drop-offs are highlighted in red. Move your cursor over each area to identify additional information.

Click on China, then click on “Highlight traffic through here”. Which web page has the highest traffic from China?

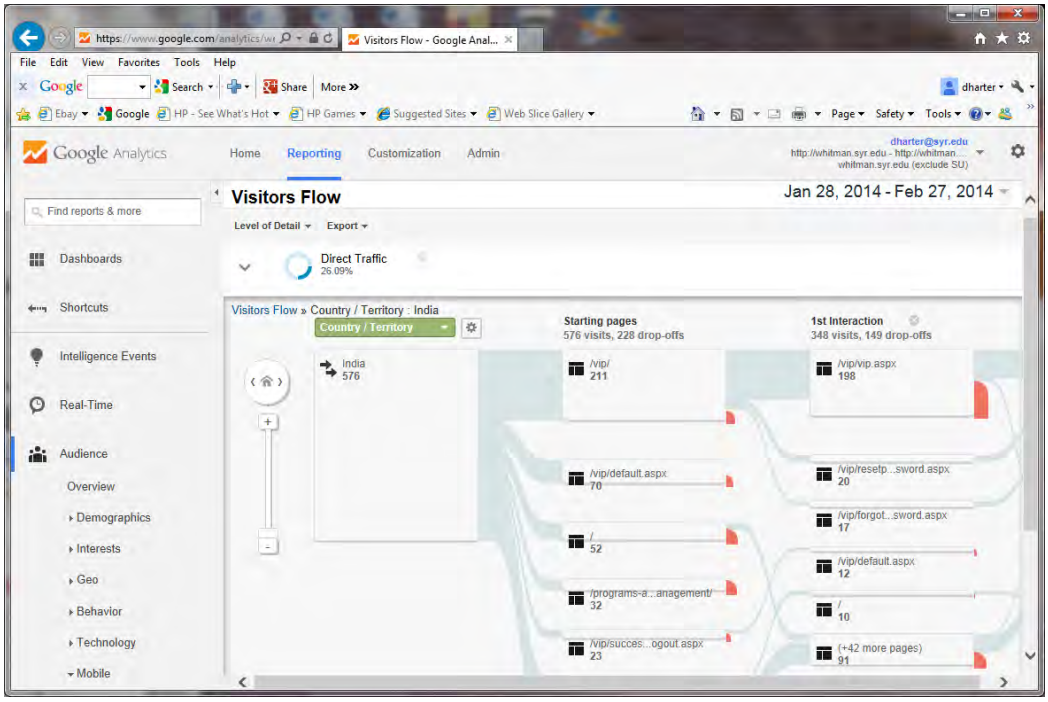


Now, click on China, then View only this segment. Is it easier to see the traffic flows for China?




To return to all Visitors Flow, look for “Visitors Flow >> Country/Territory: China”, then click on the first part, Visitors Flow.

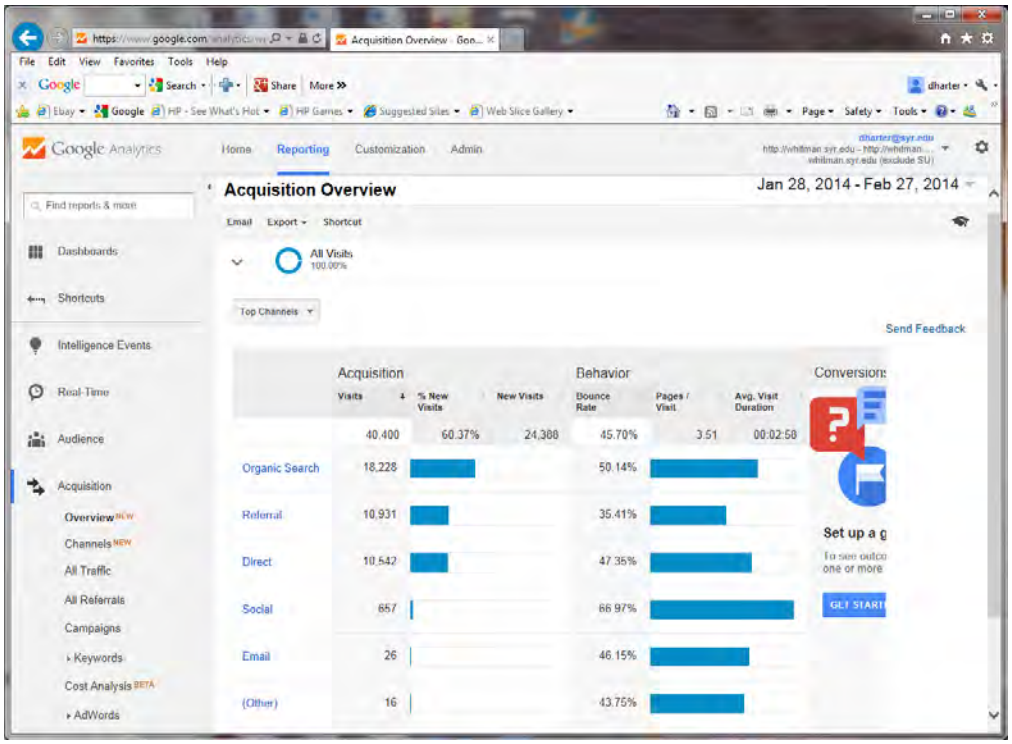
Change to India, then use the drop down arrow to switch from All Visits to Direct Traffic.



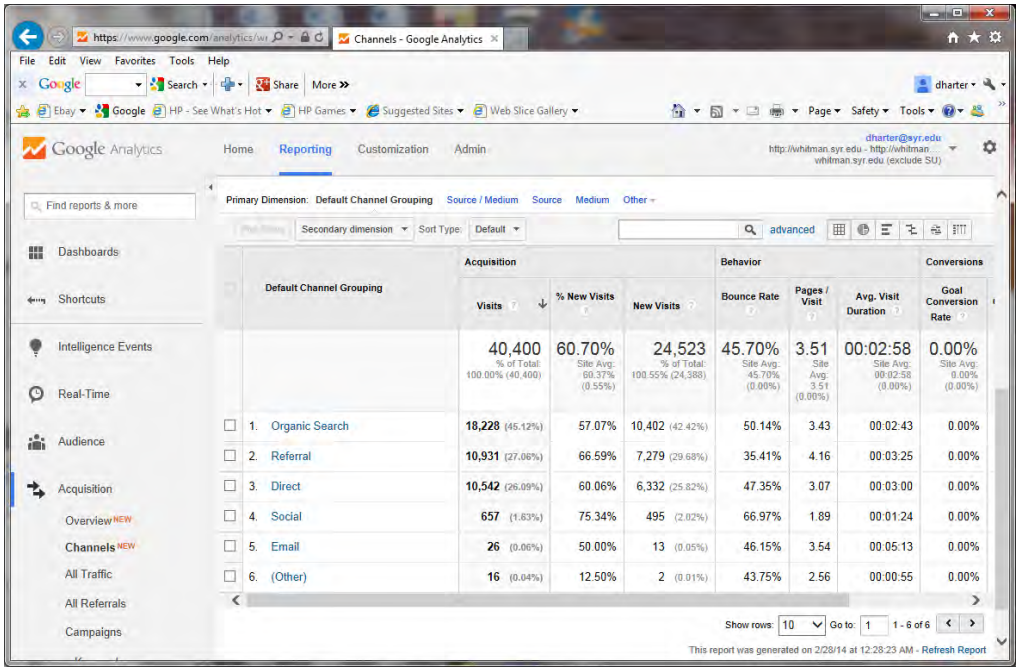

Switch back to All Visits.

**Session 3.9: Acquisition**

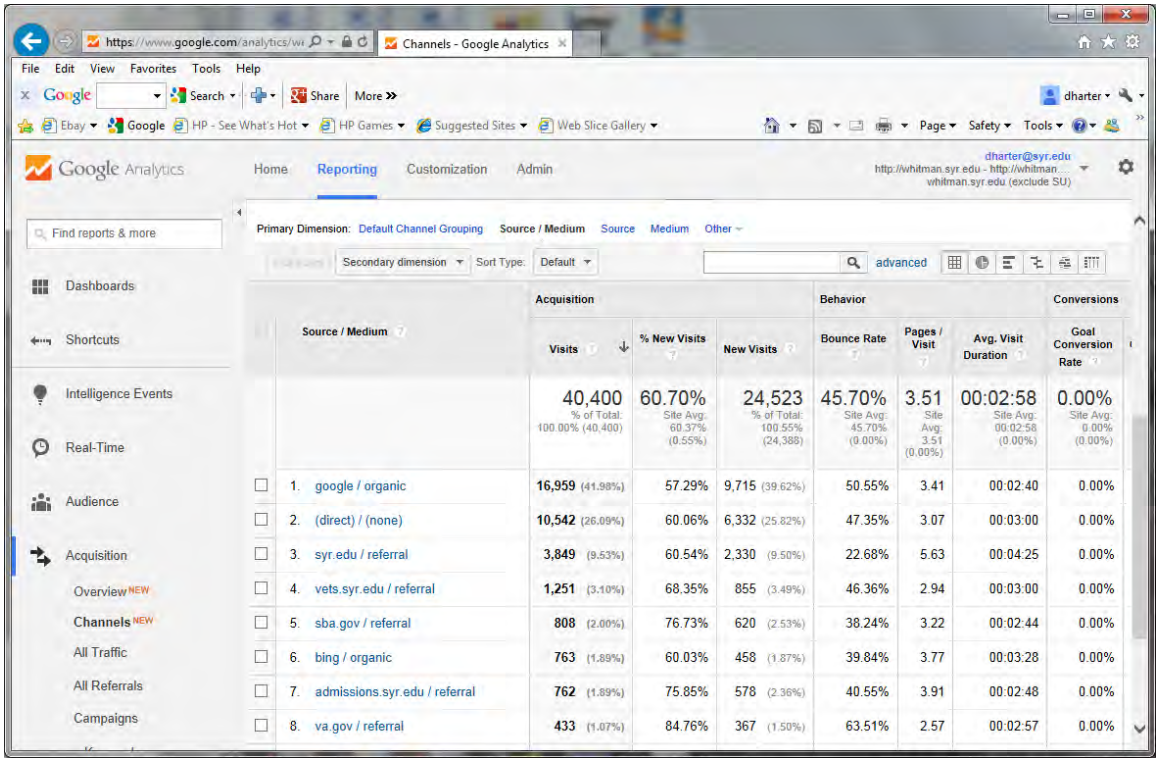
Acquisition shows how traffic arrives at your website. On the left, click on Acquisition and Overview.



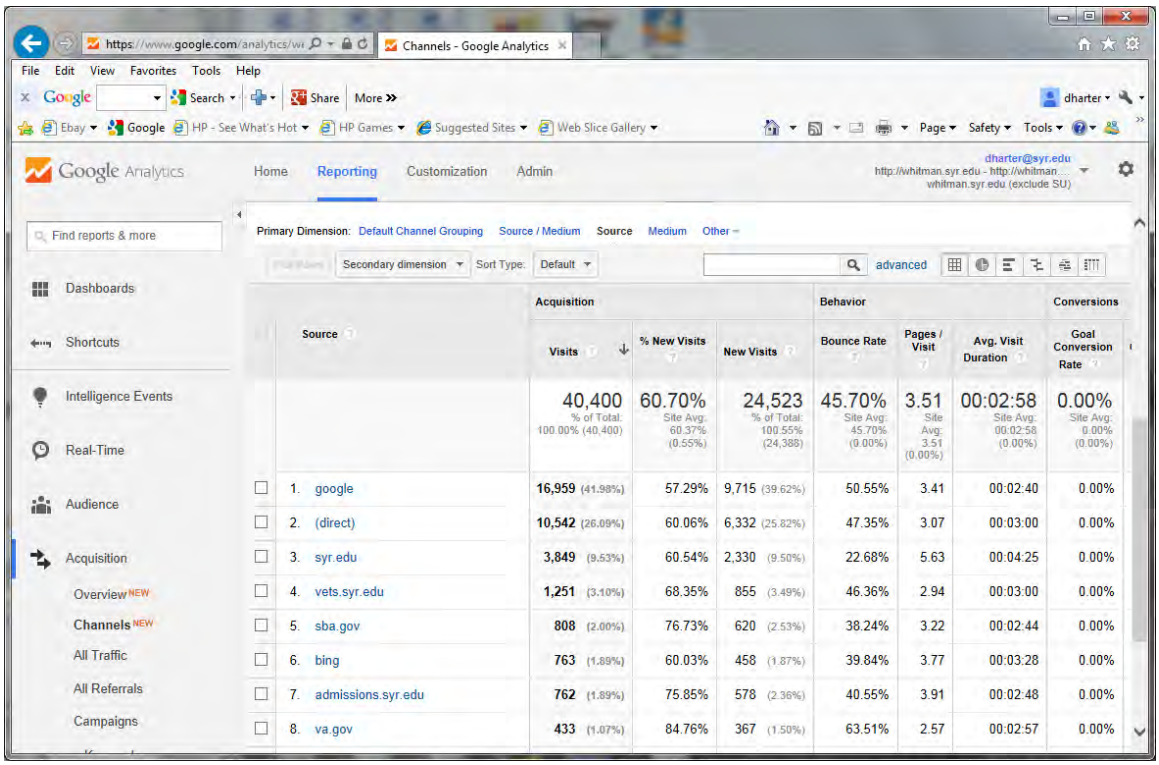

Next, click on Acquisition, Channels. What's the difference in the categories?



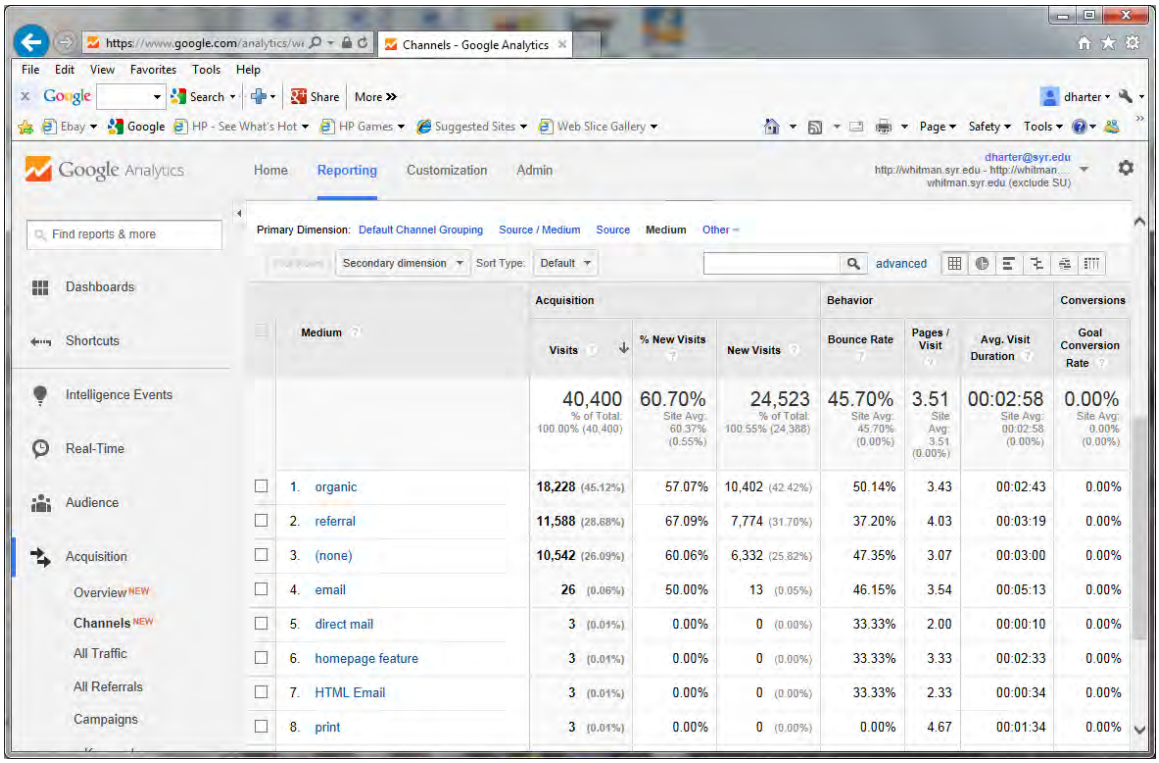
Change the Primary Dimension to Source/Medium. Can you interpret what these are? What is the difference between organic and referral?



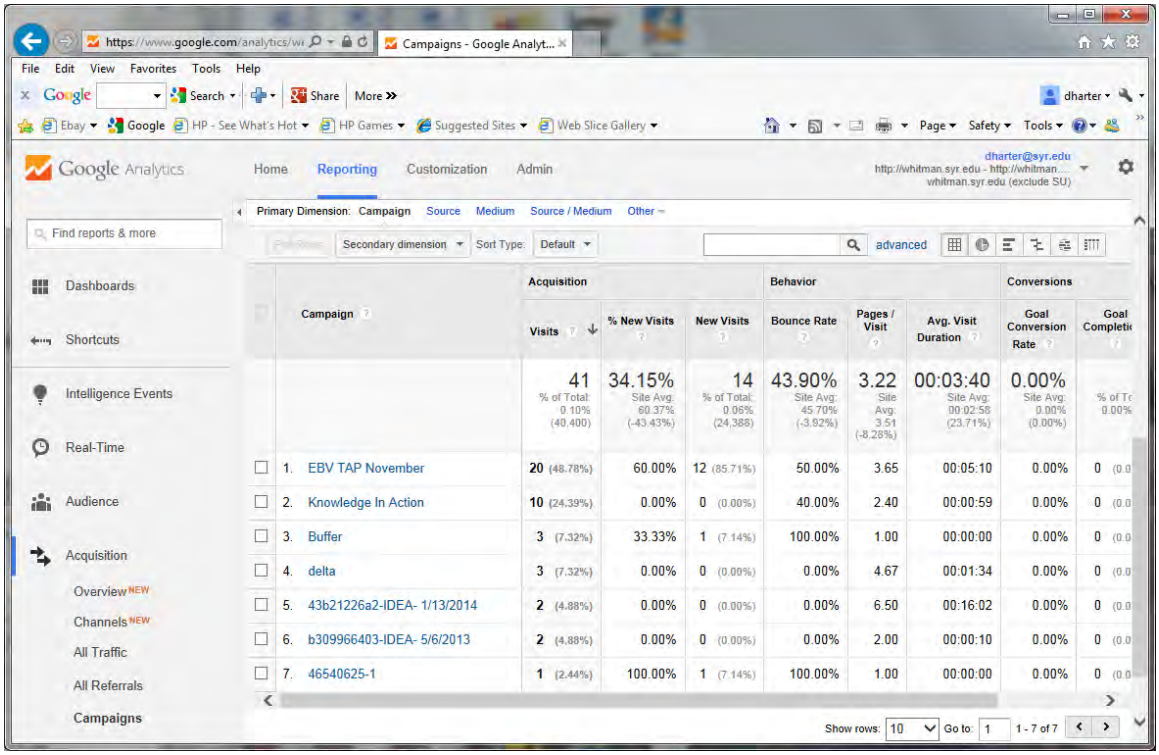

Change Primary Dimension to Source.



Finally, change Primary Dimension to Medium.

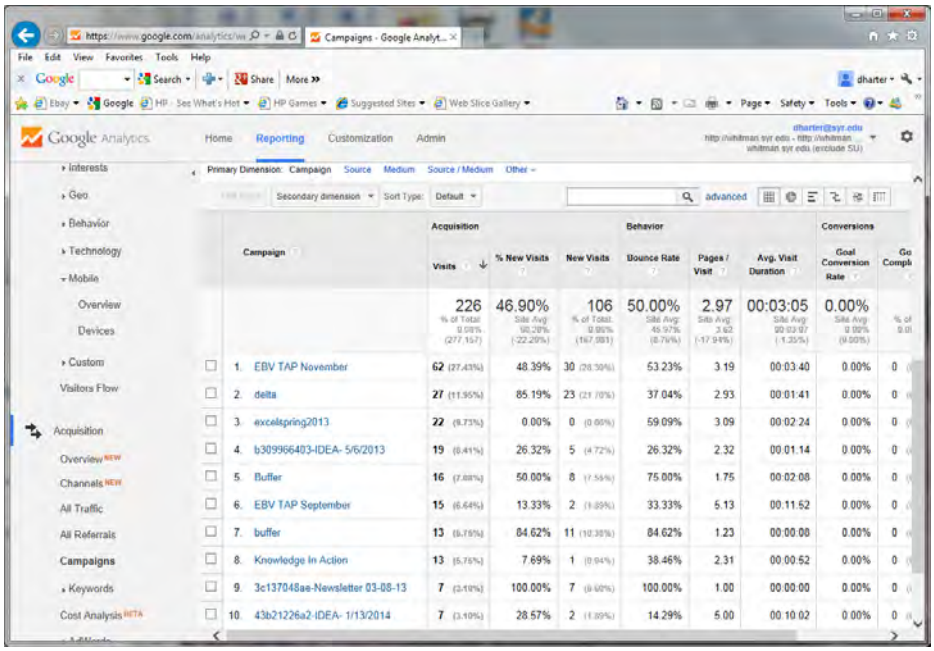



Now examine Acquisition, Campaigns.



Let's expand the time period of the data. In the upper right corner of the screen is the date range.

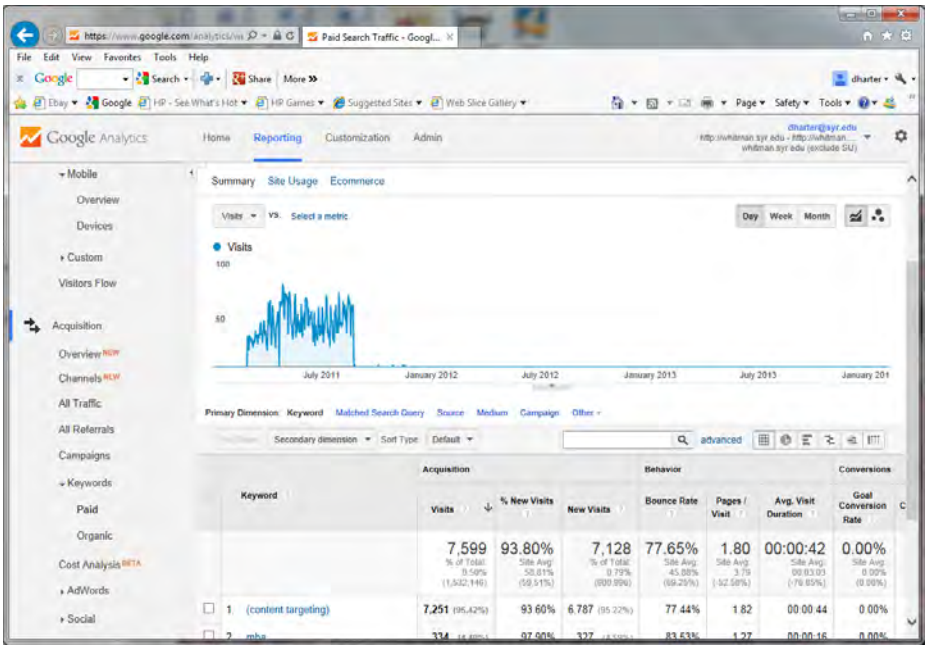
1. Click on the date
2. A calendar will appear
3. Click on the "from" date, then scroll to the left in the calendar and select August 1, 2013.
4. Click on the "to" date, then scroll to the left until you get to today.
5. Click Apply



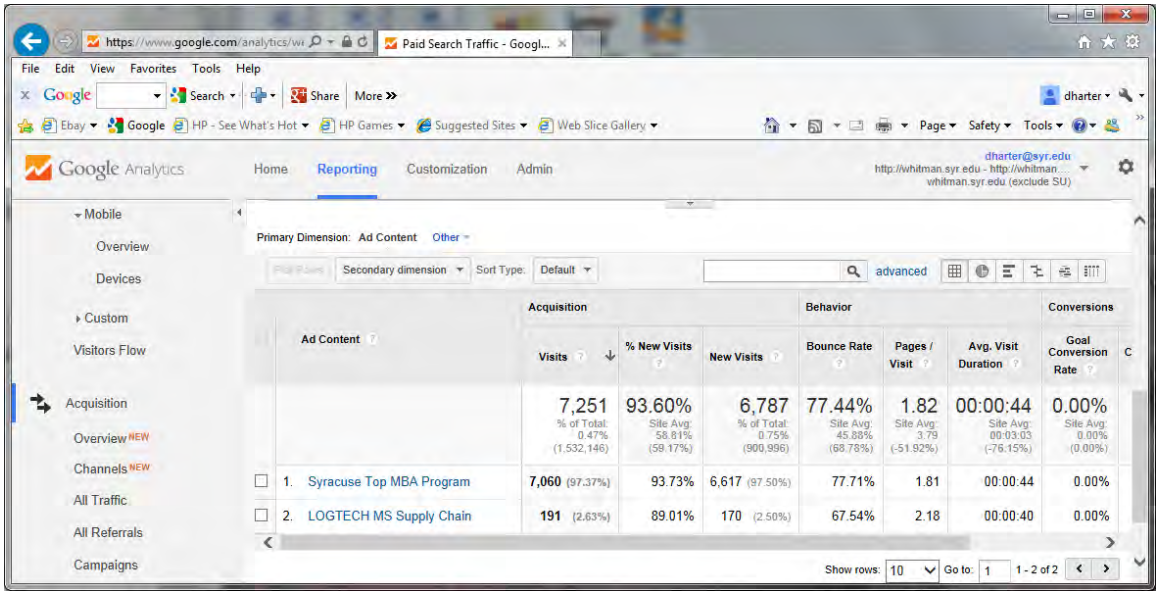

The "delta" campaign was an advertisement placed in Delta Airline magazine. How does its bounce rate compare to other campaigns?

**Acquisition – Key Words**

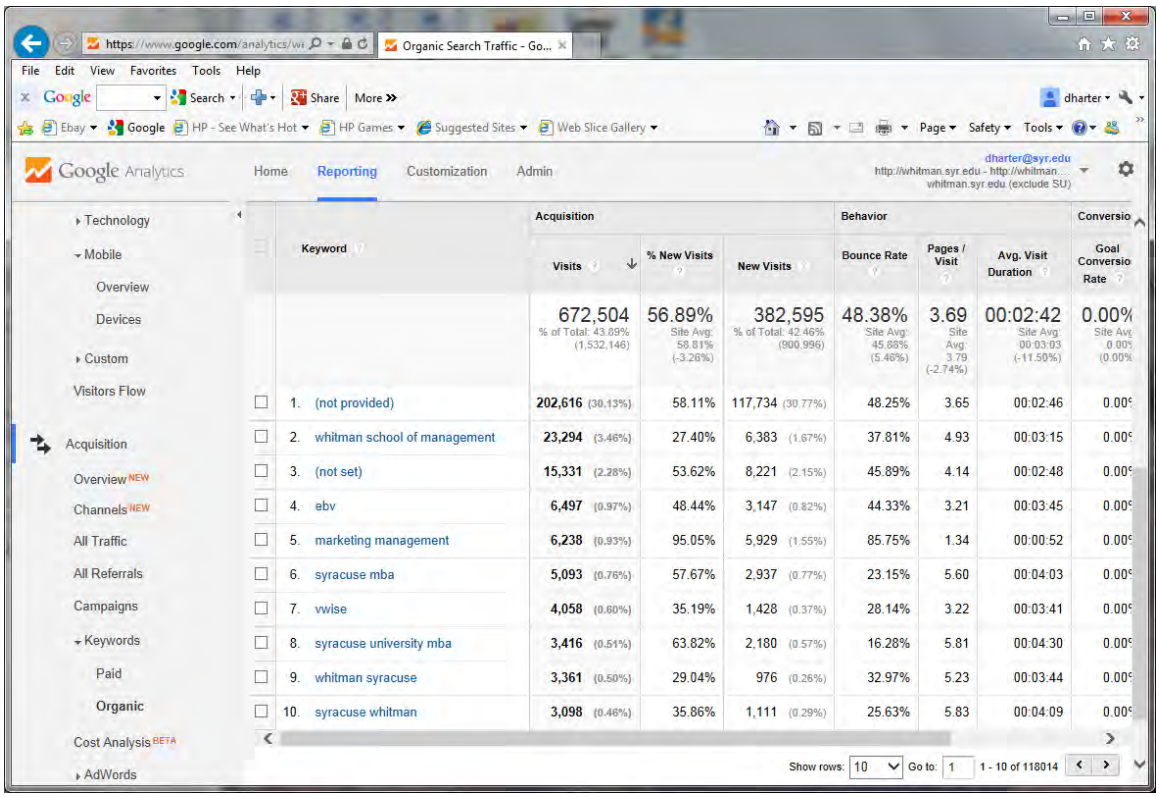
Change the date range to the range January 1, 2011 through today. Click on Keywords, Paid.




Now click on “(content targeting)”.

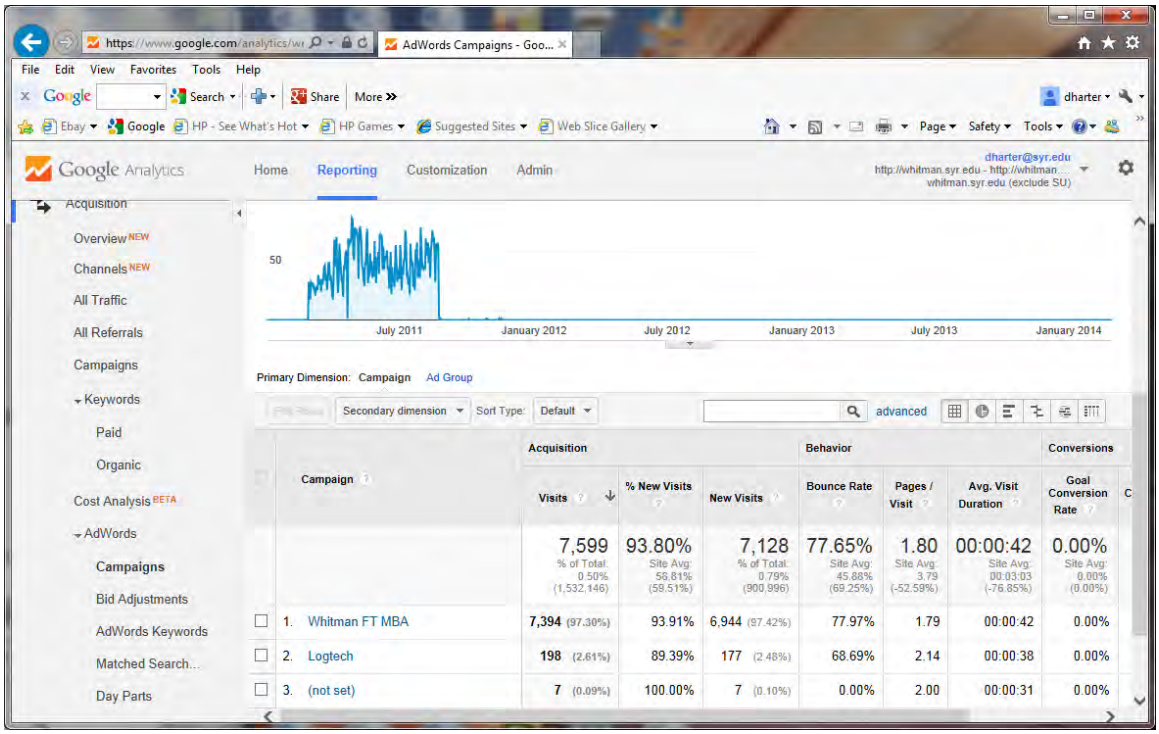


Next, on the left of the screen, click on Keywords, Organic. Organic means keyword searches that a visitor entered into a search engine, that were not associated with an advertisement. How many different keyword combinations were used in this period?

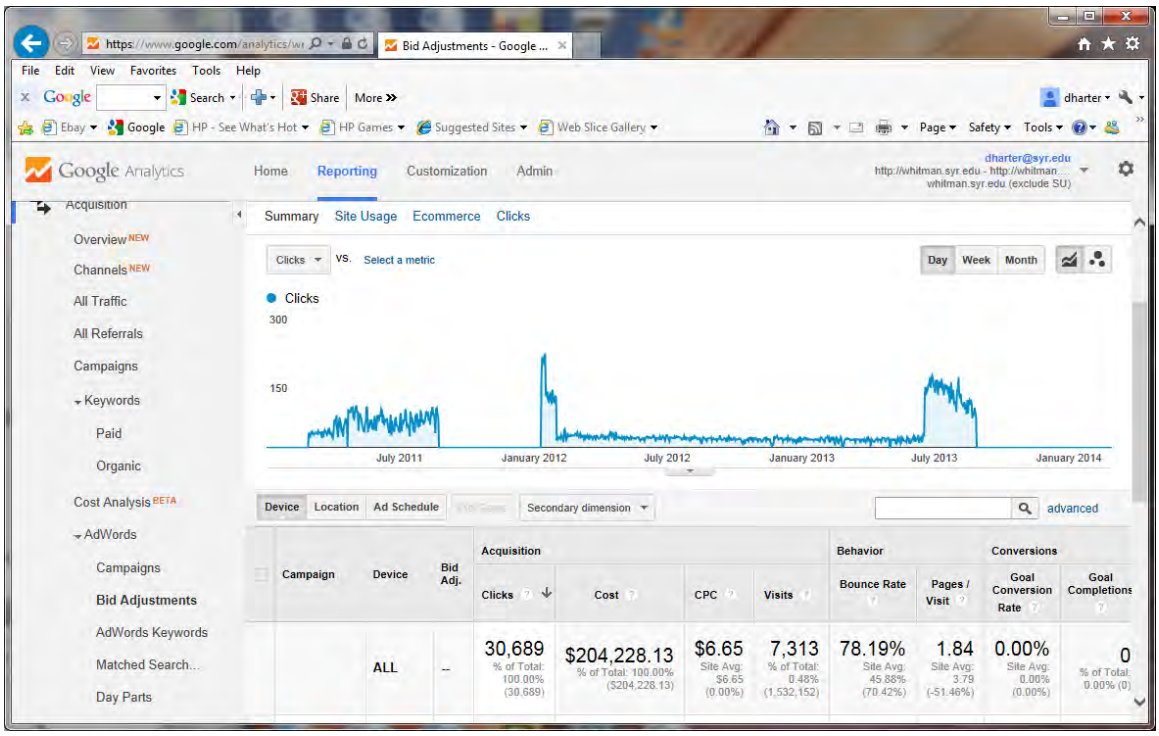


## Adwords

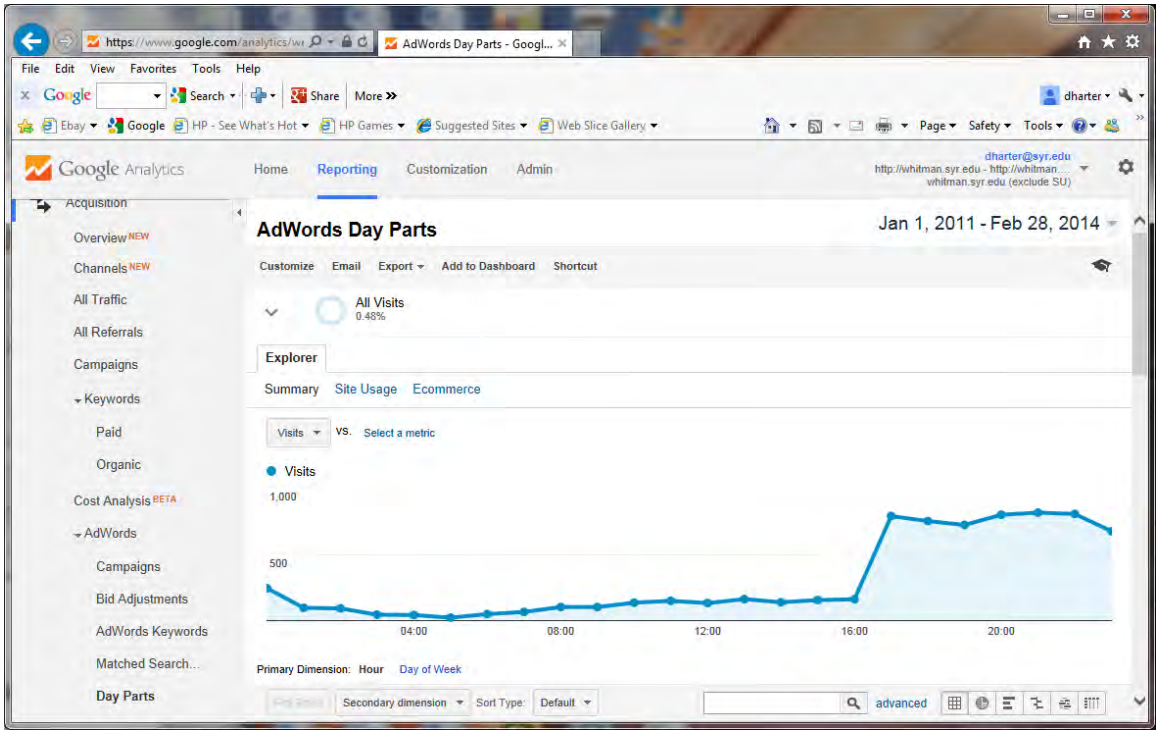
Advertising on Google is through Google Adwords. Click on Adwords, Campaigns on the left side of the screen.



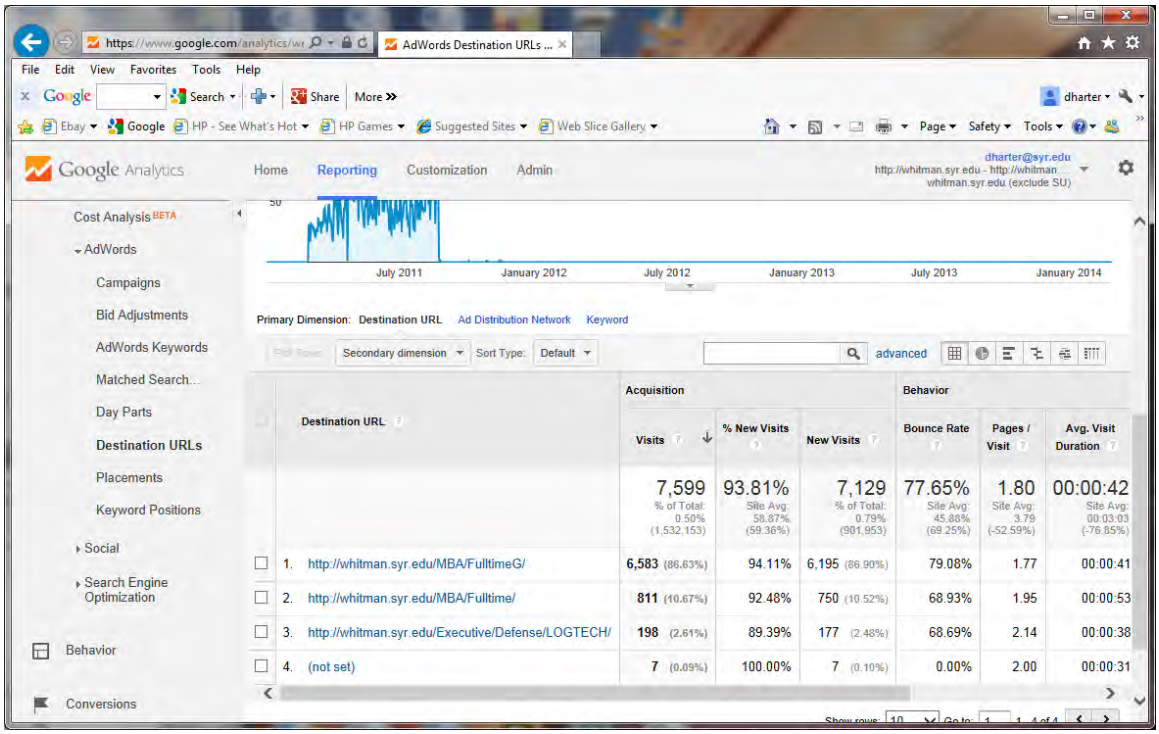

Now switch to Adwords, Bid Adjustment.



Change to Adwords, Day Parts. This is time of day advertising. When is the highest level of traffic?

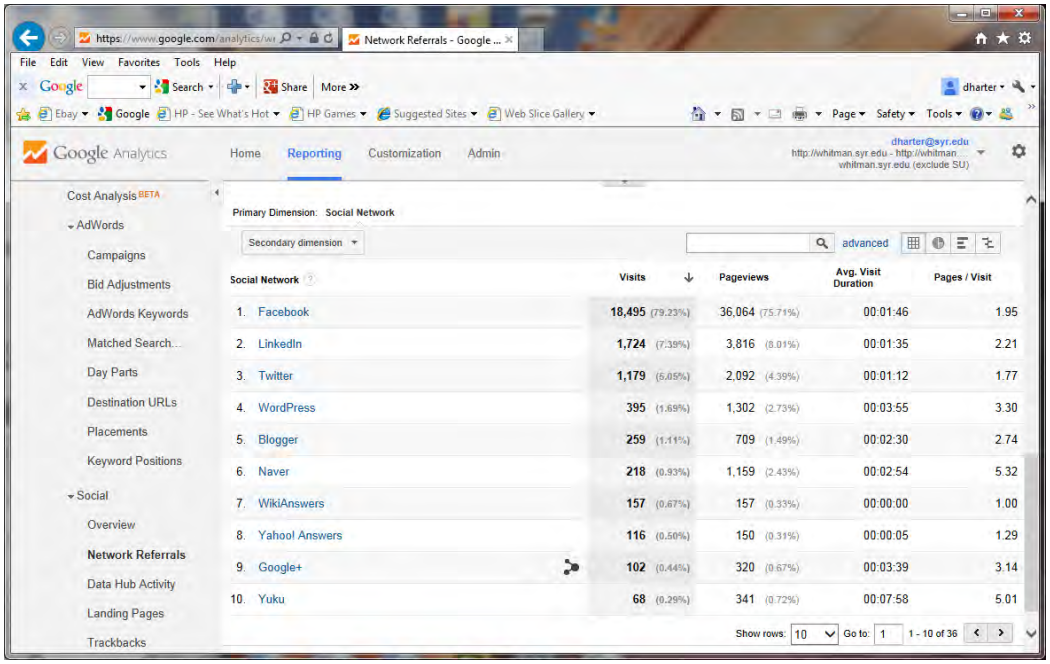



Click on Adwords, Destination URL. Where is the traffic going?

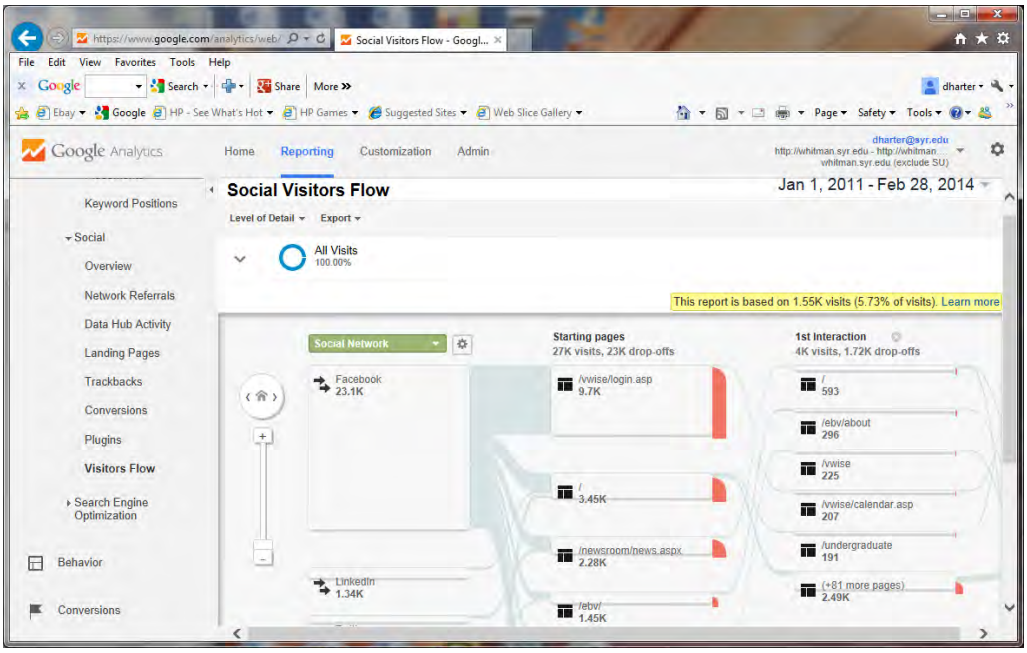


## Social

Click on Social, Network Referrals. What social media site sends the most traffic to the website?

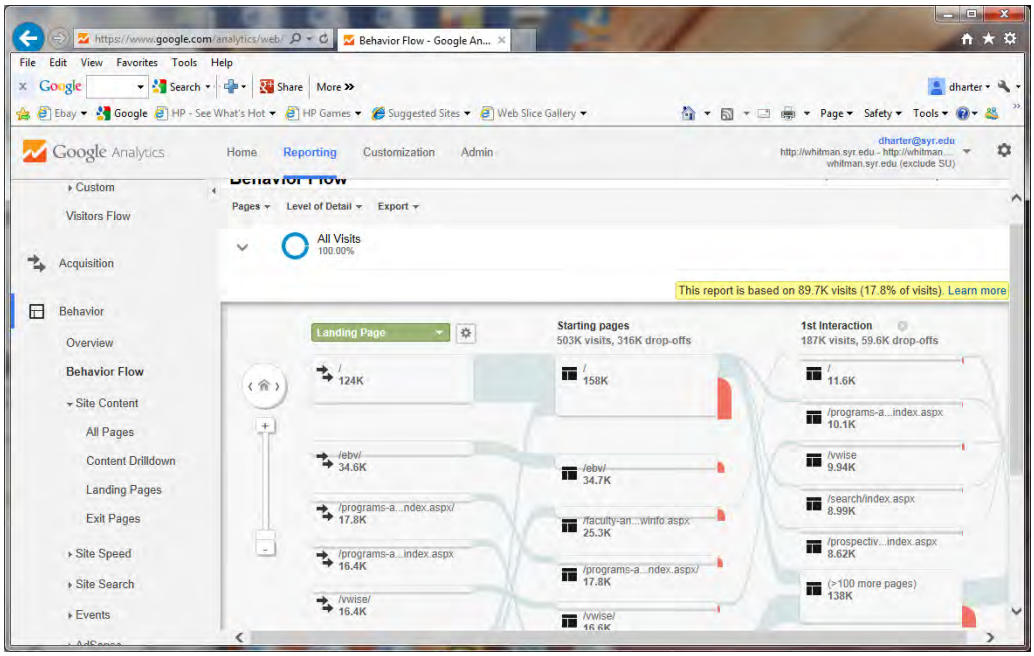



Click on Social, Visitors Flow to see a graphic representation of the traffic flow from social websites.

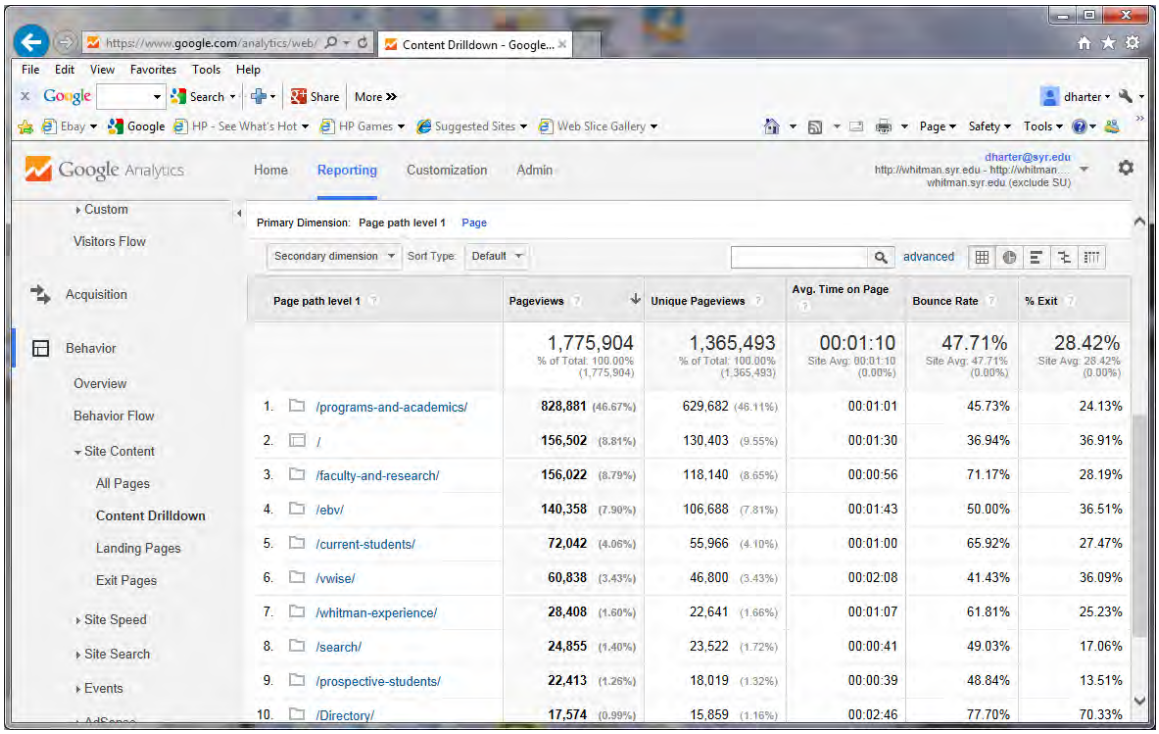


**Session 3.10: Behavior – Site Content**

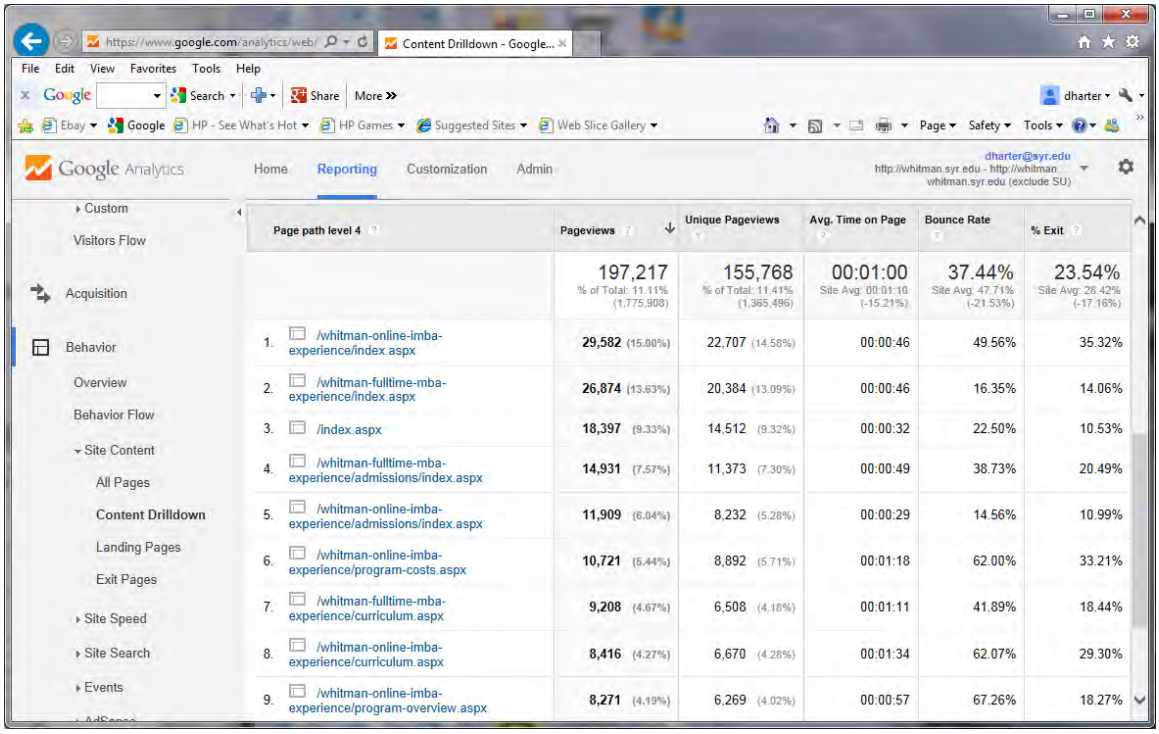
The new Whitman website was created in January 2013. Change your date range to start with February 1, 2013 and end with today. Then click on Behavior, Behavior Flow on the left side.



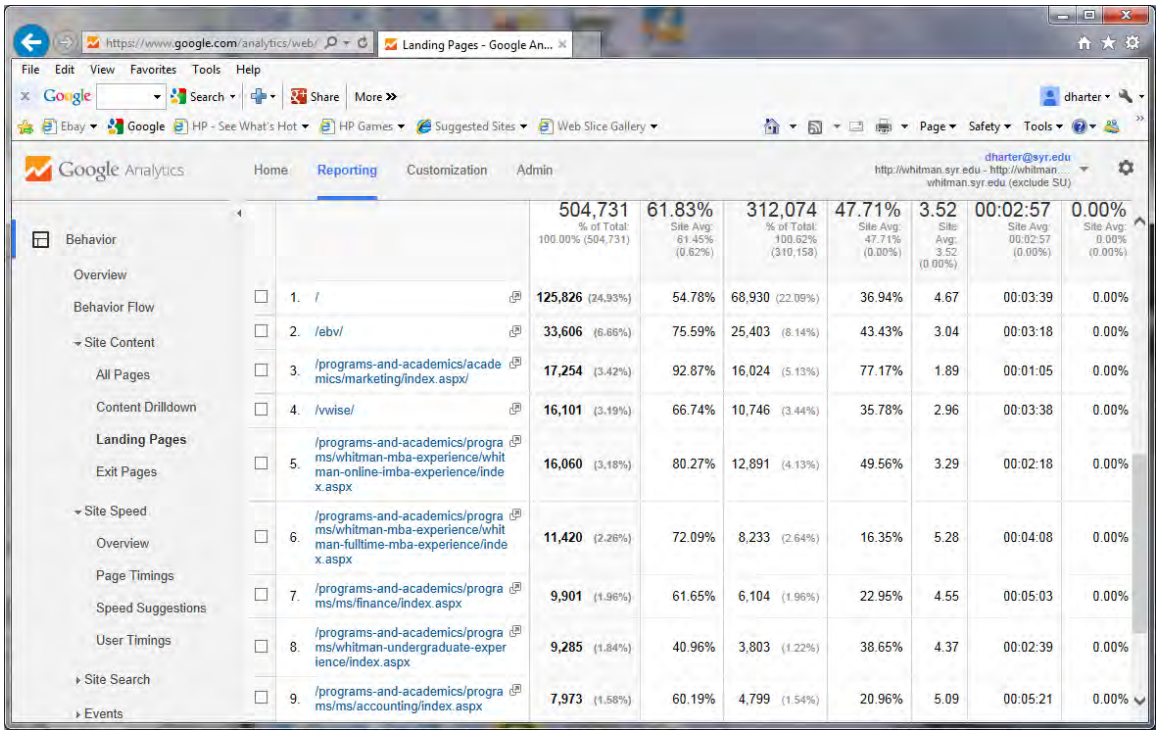
Next, click on Site Content, Content Drilldown.



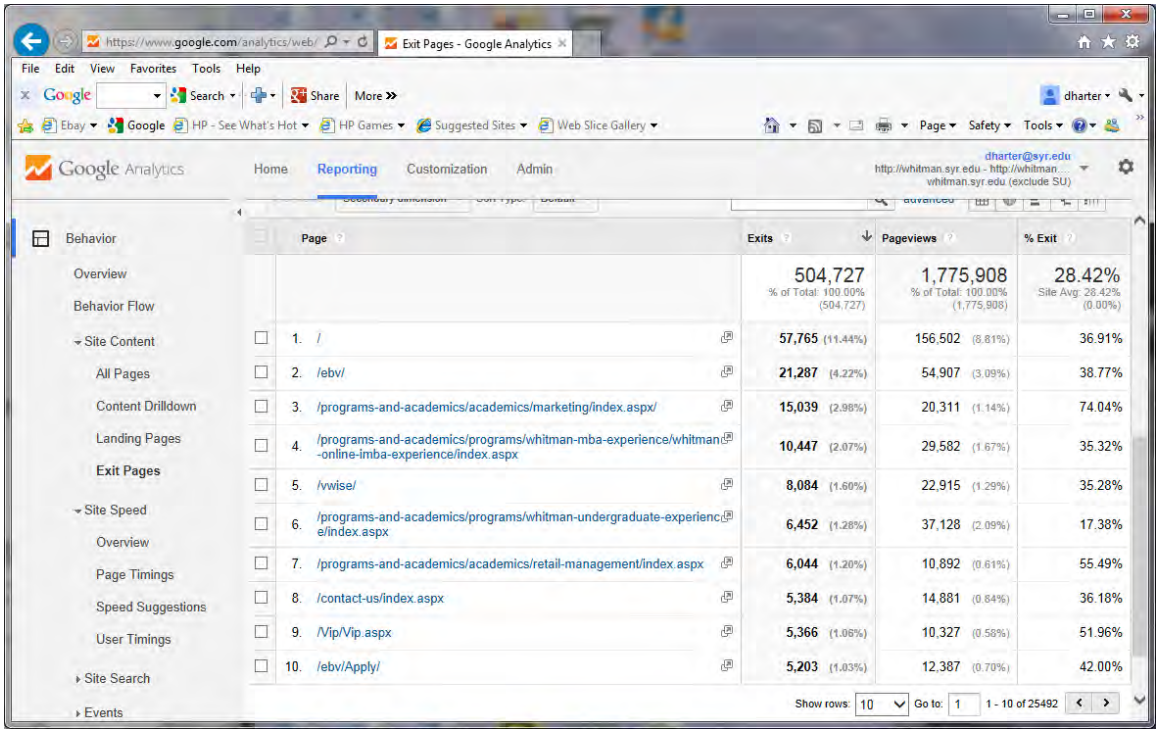

Click on: programs-and-academics, then programs, then whitman-mba-experience.



Click on Site Content, Landing Pages.

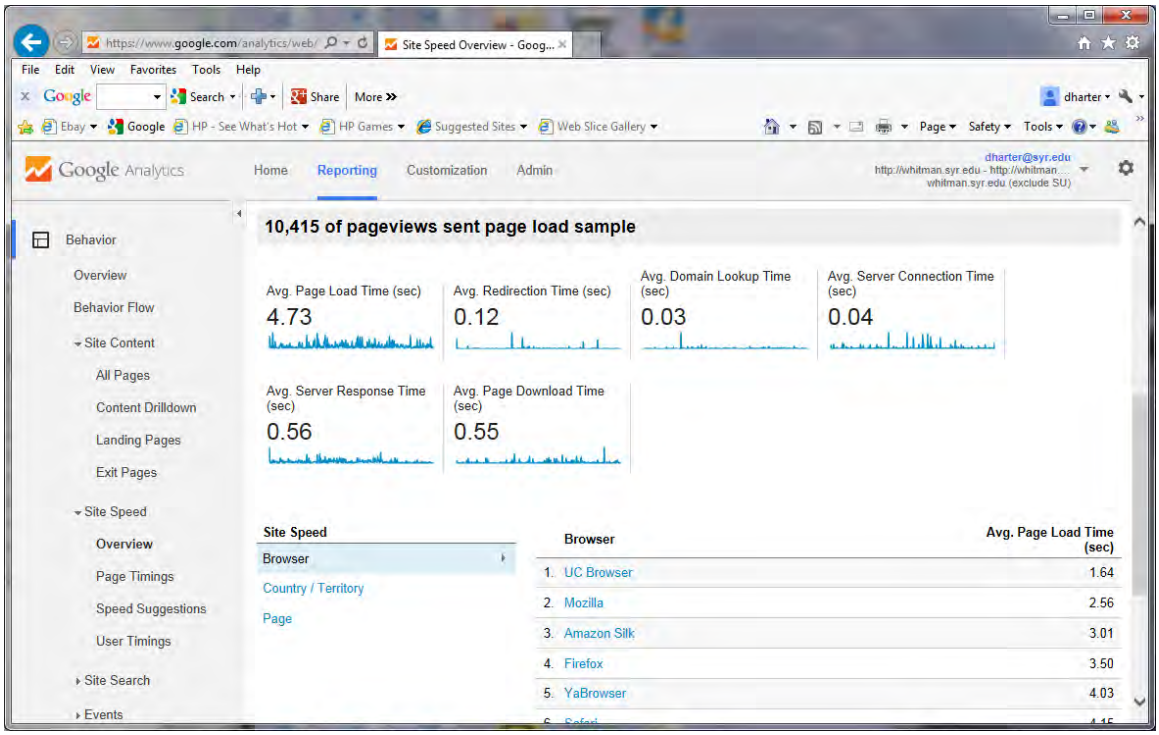



Click on Site Content, Exit Pages.

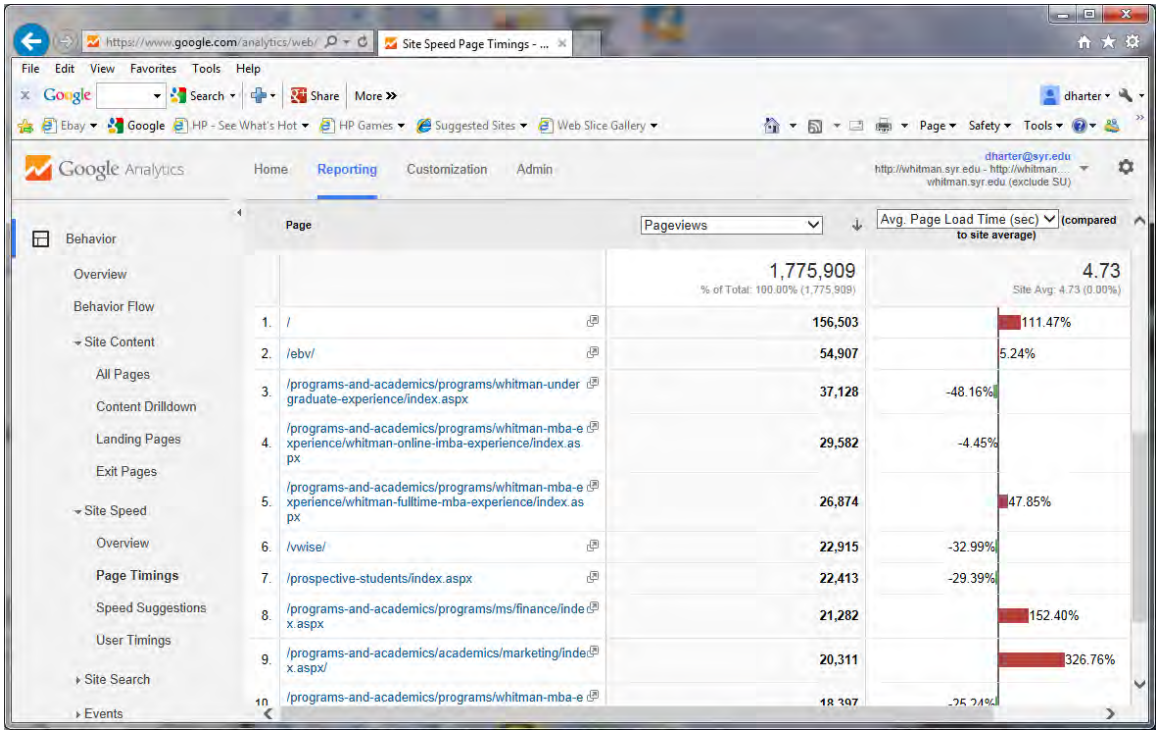


**Behavior – Site Speed**

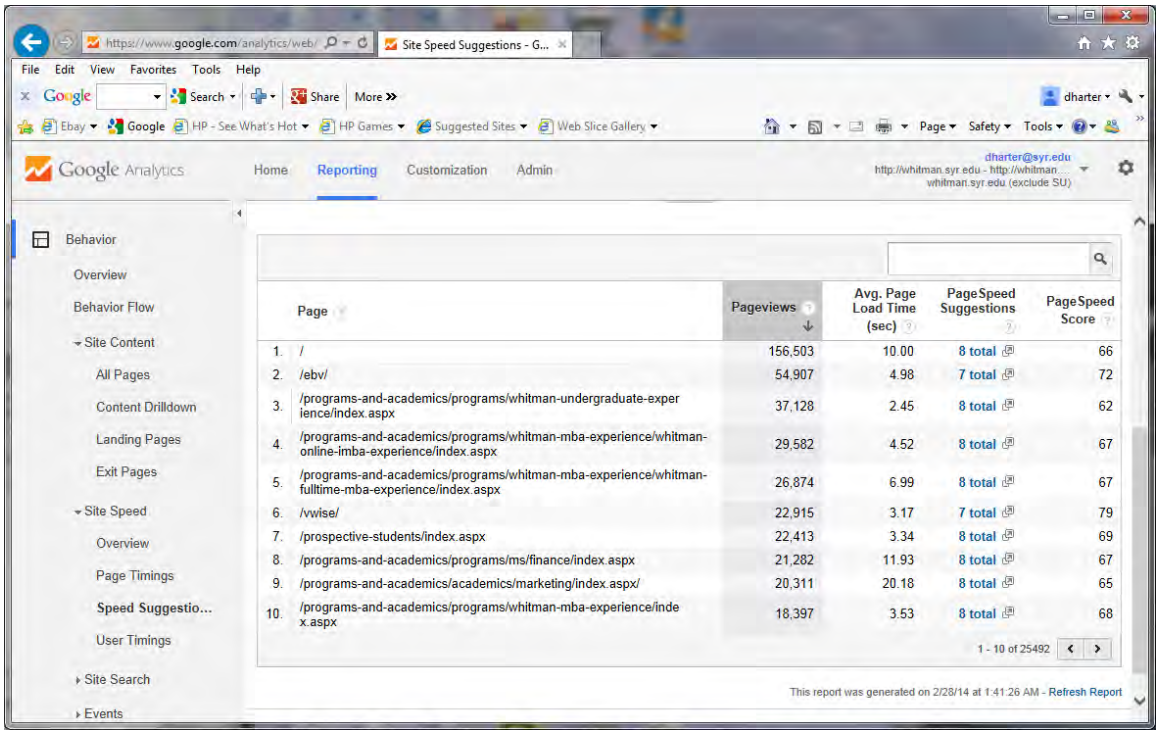
Click on Behavior, Site Speed, Overview.




Click on Site Speed, Page Timings. What do you notice about the main web page?



Click on Site Speed, Speed Suggestions.



# Microsoft Excel:
# Sensitivity Analysis & Conditional Formatting

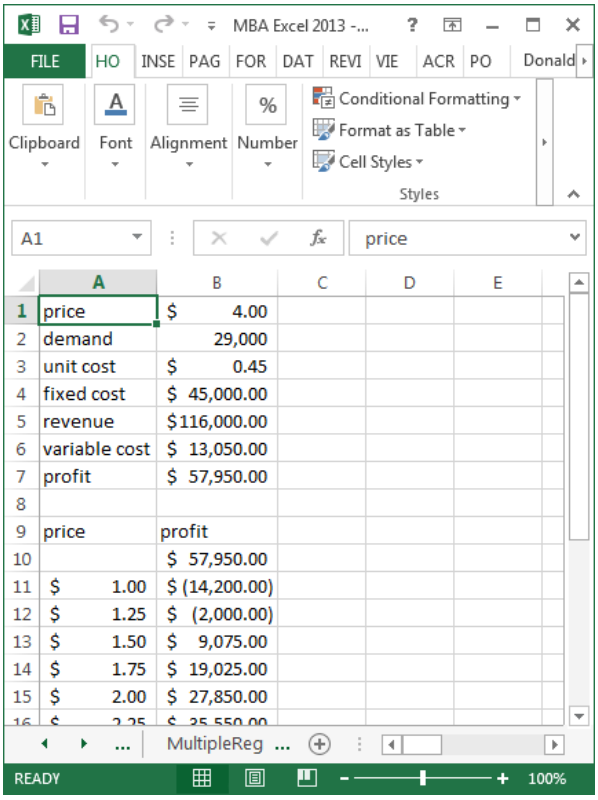
## Microsoft Excel

- Go to the Course Website
- Download and print “Business Analytics – Week 3 Instructions.doc”
- Download “Business Analytics – Week 3 Excel.xlsx”

### Session 3.3: One-way Sensitivity Analysis

Sensitivity analysis allows you to see the effect of a changing variable on an outcome variable. In the example below, use the spreadsheet tab Sensitivity.

The spreadsheet provides the demand curve, costs, revenue and profit for a widget. By changing the price, you can observe the effect on sales and profitability. But this approach only allows you to look at one price at a time. Using the Sensitivity spreadsheet, we will build the spreadsheet shown in the diagram below.




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | price | $ 4.00 | | | |
| 2 | demand | 29,000 | | | |
| 3 | unit cost | $ 0.45 | | | |
| 4 | fixed cost | $ 45,000.00 | | | |
| 5 | revenue | $116,000.00 | | | |
| 6 | variable cost | $ 13,050.00 | | | |
| 7 | profit | $ 57,950.00 | | | |
| 8 | | | | | |
| 9 | price | profit | | | |
| 10 | | $ 57,950.00 | | | |
| 11 | $ 1.00 | $ (14,200.00) | | | |
| 12 | $ 1.25 | $ (2,000.00) | | | |
| 13 | $ 1.50 | $ 9,075.00 | | | |
| 14 | $ 1.75 | $ 19,025.00 | | | |
| 15 | $ 2.00 | $ 27,850.00 | | | |
| 16 | $ 2.25 | $ 35,550.00 | | | |

The following steps build a table to examine the effect price has on profits, revenue and costs.

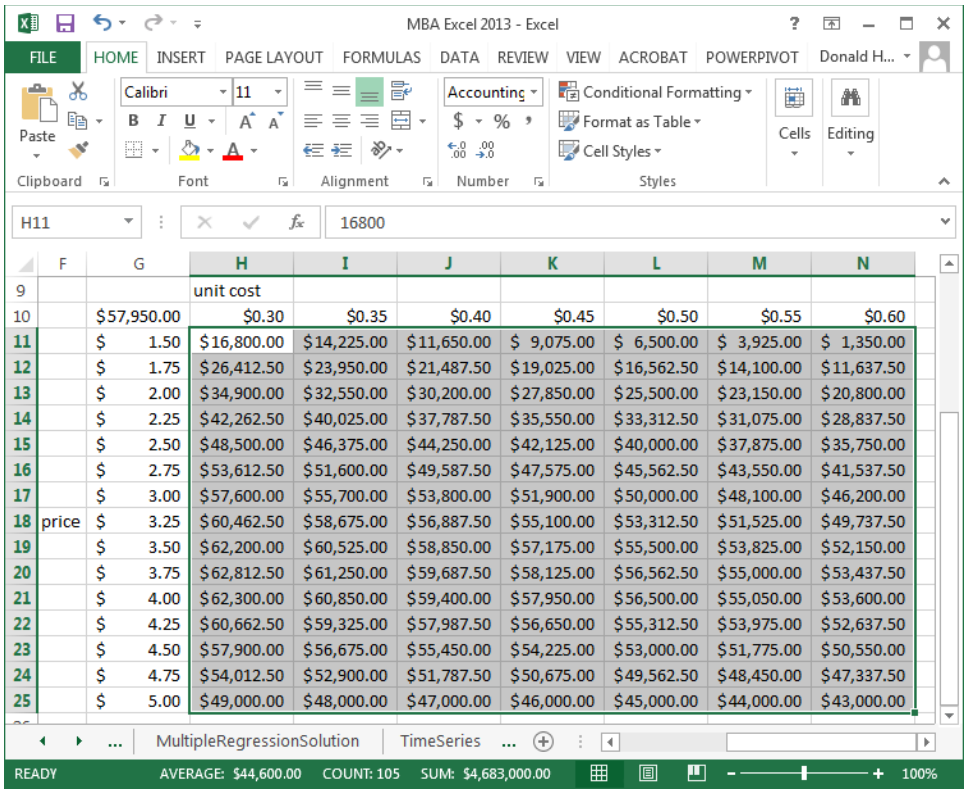
1. To set up a one-way table, enter the labels of your one input variable and all of the output variables in a row. In this example, enter price in A9 and profit in B9.
2. Next, enter a column of prices under price **but skip the first cell below the word price**. In this example, enter a column for price with price ranging from $1 to $4 by $0.25 increments.
3. Immediately under profit in cell B10, enter =B7, so the profit calculation is filled in.

4. Finally, we will create a data table. Highlight the fields immediately under the labels and continue until all of the prices are highlighted. In this case, we highlight A10:B23. Click on the Data tab at the top of the page, click on What-If Analysis, Data Table, then enter in the column input cell the cell reference for variable that you want to vary (i.e., $B$1, which is price), then click OK. This should populate the data table.

What you really did: The price column represents on which prices you would like to perform a what-if analysis, and the three formulas are the outcomes that you want to analyze when price changes. The first step was to set up the price points and the formulas. When you created the data table, you identified what variable changes, but used an absolute address.

## Session 3.4: Two-way Sensitivity Analysis

If you want to change two variables at once, we create a two-way table. In this example, let's change both price and unit cost for our widgets. The result that we want will look like the picture below.



| | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|
| 9 | | | unit cost | | | | | | |
| 10 | | $57,950.00 | $0.30 | $0.35 | $0.40 | $0.45 | $0.50 | $0.55 | $0.60 |
| 11 | | $ 1.50 | $16,800.00 | $14,225.00 | $11,650.00 | $ 9,075.00 | $ 6,500.00 | $ 3,925.00 | $ 1,350.00 |
| 12 | | $ 1.75 | $26,412.50 | $23,950.00 | $21,487.50 | $19,025.00 | $16,562.50 | $14,100.00 | $11,637.50 |
| 13 | | $ 2.00 | $34,900.00 | $32,550.00 | $30,200.00 | $27,850.00 | $25,500.00 | $23,150.00 | $20,800.00 |
| 14 | | $ 2.25 | $42,262.50 | $40,025.00 | $37,787.50 | $35,550.00 | $33,312.50 | $31,075.00 | $28,837.50 |
| 15 | | $ 2.50 | $48,500.00 | $46,375.00 | $44,250.00 | $42,125.00 | $40,000.00 | $37,875.00 | $35,750.00 |
| 16 | | $ 2.75 | $53,612.50 | $51,600.00 | $49,587.50 | $47,575.00 | $45,562.50 | $43,550.00 | $41,537.50 |
| 17 | | $ 3.00 | $57,600.00 | $55,700.00 | $53,800.00 | $51,900.00 | $50,000.00 | $48,100.00 | $46,200.00 |
| 18 | price | $ 3.25 | $60,462.50 | $58,675.00 | $56,887.50 | $55,100.00 | $53,312.50 | $51,525.00 | $49,737.50 |
| 19 | | $ 3.50 | $62,200.00 | $60,525.00 | $58,850.00 | $57,175.00 | $55,500.00 | $53,825.00 | $52,150.00 |
| 20 | | $ 3.75 | $62,812.50 | $61,250.00 | $59,687.50 | $58,125.00 | $56,562.50 | $55,000.00 | $53,437.50 |
| 21 | | $ 4.00 | $62,300.00 | $60,850.00 | $59,400.00 | $57,950.00 | $56,500.00 | $55,050.00 | $53,600.00 |
| 22 | | $ 4.25 | $60,662.50 | $59,325.00 | $57,987.50 | $56,650.00 | $55,312.50 | $53,975.00 | $52,637.50 |
| 23 | | $ 4.50 | $57,900.00 | $56,675.00 | $55,450.00 | $54,225.00 | $53,000.00 | $51,775.00 | $50,550.00 |
| 24 | | $ 4.75 | $54,012.50 | $52,900.00 | $51,787.50 | $50,675.00 | $49,562.50 | $48,450.00 | $47,337.50 |
| 25 | | $ 5.00 | $49,000.00 | $48,000.00 | $47,000.00 | $46,000.00 | $45,000.00 | $44,000.00 | $43,000.00 |

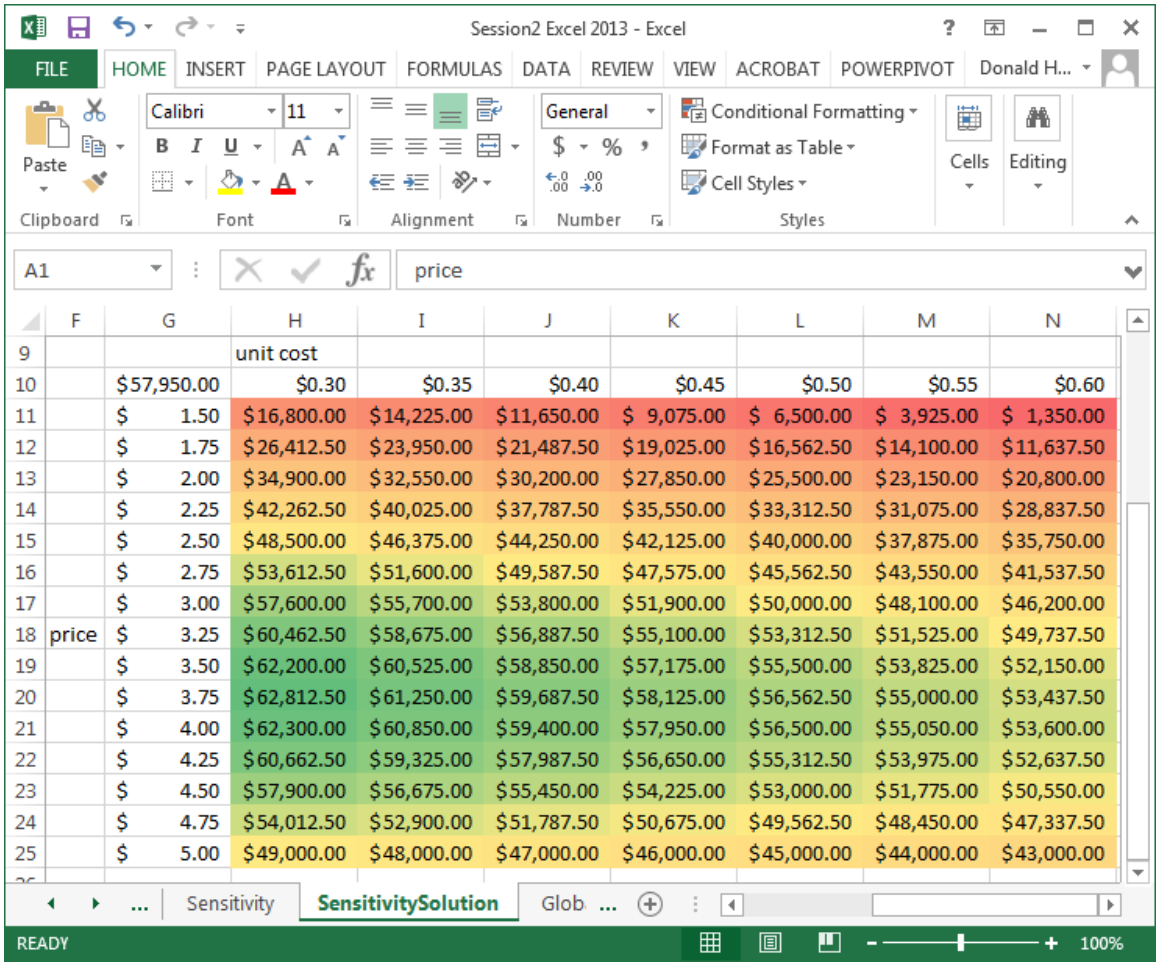
Note that in a two-way table, you can only have one outcome variable. In this case, let's use profit as the outcome variable. The following steps create a two-way what-if analysis:

1. As before, enter a column for prices. In the picture, the prices are in cells G11:G23. You can type a label of price to the left of any cell.
2. Enter variable costs in H10:N10 from $.30 to $.60 in increments of $.05. You can enter a label for variable costs above the first entry (cell H9).
3. The price column and variable cost row will form the boundary of the table.
4. Now enter the formula that you want to calculate where the column and row intersect, in this case, cell G10. The formula we want to use is for profit; enter =B7.
5. Now highlight the entire table of prices and variable costs G10:N25. Click on the Data tab at the top of the spreadsheet, then What-if Analysis and Data Table. Since the variable that we are varying across the row is unit cost, and unit cost is a named variable, enter unit cost for the row variable. Likewise, since price varies down the column and is named variable, enter price as the column variable. Click OK.

**Session 3.5: Conditional formatting**

Conditional formatting allows you to display data differently based on the values of the data. For example, you can change the font color, background color, or other characteristics to make the data jump out at you when it satisfies certain conditions. Let's convert the data in the Sensitivity spreadsheet to something more visual.

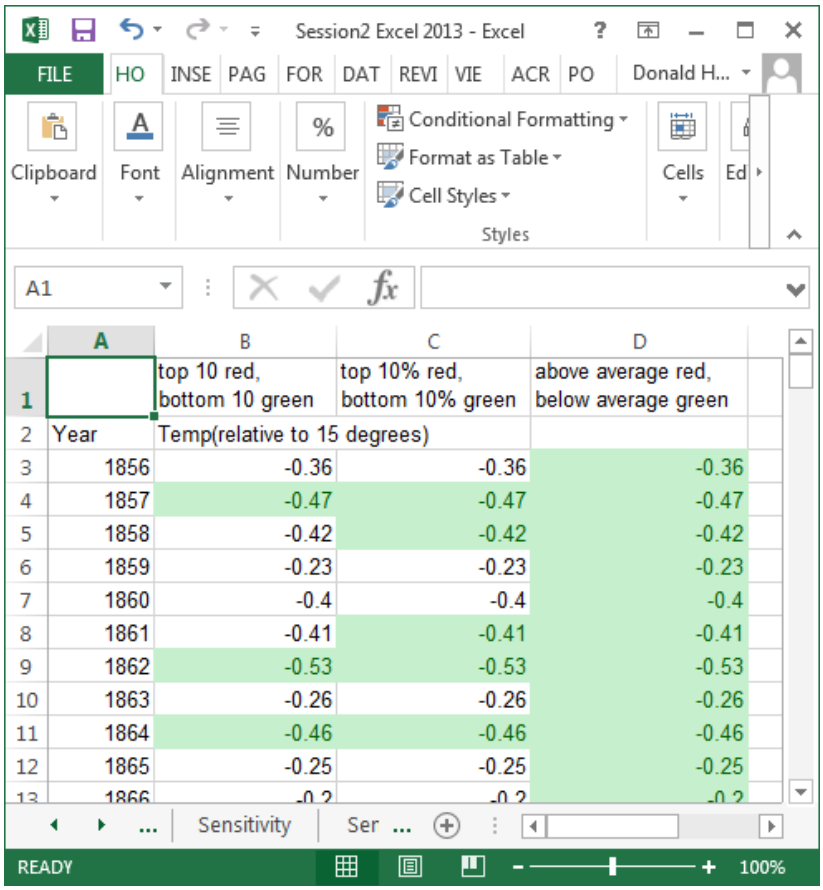
1. Continuing to use the Sensitivity data, highlight cells H11:N25.
2. Click on the Home tab, Conditional Formatting, Color Scales, then select a color scheme from the right.




| | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|
| 9 | | | unit cost | | | | | | |
| 10 | | $57,950.00 | $0.30 | $0.35 | $0.40 | $0.45 | $0.50 | $0.55 | $0.60 |
| 11 | | $ 1.50 | $16,800.00 | $14,225.00 | $11,650.00 | $ 9,075.00 | $ 6,500.00 | $ 3,925.00 | $ 1,350.00 |
| 12 | | $ 1.75 | $26,412.50 | $23,950.00 | $21,487.50 | $19,025.00 | $16,562.50 | $14,100.00 | $11,637.50 |
| 13 | | $ 2.00 | $34,900.00 | $32,550.00 | $30,200.00 | $27,850.00 | $25,500.00 | $23,150.00 | $20,800.00 |
| 14 | | $ 2.25 | $42,262.50 | $40,025.00 | $37,787.50 | $35,550.00 | $33,312.50 | $31,075.00 | $28,837.50 |
| 15 | | $ 2.50 | $48,500.00 | $46,375.00 | $44,250.00 | $42,125.00 | $40,000.00 | $37,875.00 | $35,750.00 |
| 16 | | $ 2.75 | $53,612.50 | $51,600.00 | $49,587.50 | $47,575.00 | $45,562.50 | $43,550.00 | $41,537.50 |
| 17 | | $ 3.00 | $57,600.00 | $55,700.00 | $53,800.00 | $51,900.00 | $50,000.00 | $48,100.00 | $46,200.00 |
| 18 | price | $ 3.25 | $60,462.50 | $58,675.00 | $56,887.50 | $55,100.00 | $53,312.50 | $51,525.00 | $49,737.50 |
| 19 | | $ 3.50 | $62,200.00 | $60,525.00 | $58,850.00 | $57,175.00 | $55,500.00 | $53,825.00 | $52,150.00 |
| 20 | | $ 3.75 | $62,812.50 | $61,250.00 | $59,687.50 | $58,125.00 | $56,562.50 | $55,000.00 | $53,437.50 |
| 21 | | $ 4.00 | $62,300.00 | $60,850.00 | $59,400.00 | $57,950.00 | $56,500.00 | $55,050.00 | $53,600.00 |
| 22 | | $ 4.25 | $60,662.50 | $59,325.00 | $57,987.50 | $56,650.00 | $55,312.50 | $53,975.00 | $52,637.50 |
| 23 | | $ 4.50 | $57,900.00 | $56,675.00 | $55,450.00 | $54,225.00 | $53,000.00 | $51,775.00 | $50,550.00 |
| 24 | | $ 4.75 | $54,012.50 | $52,900.00 | $51,787.50 | $50,675.00 | $49,562.50 | $48,450.00 | $47,337.50 |
| 25 | | $ 5.00 | $49,000.00 | $48,000.00 | $47,000.00 | $46,000.00 | $45,000.00 | $44,000.00 | $43,000.00 |

For the next example, use the Global Warming spreadsheet. In the column labeled “top ten red…”, we will use conditional formatting to change the background based on the values in the column. To do this, use the following steps:

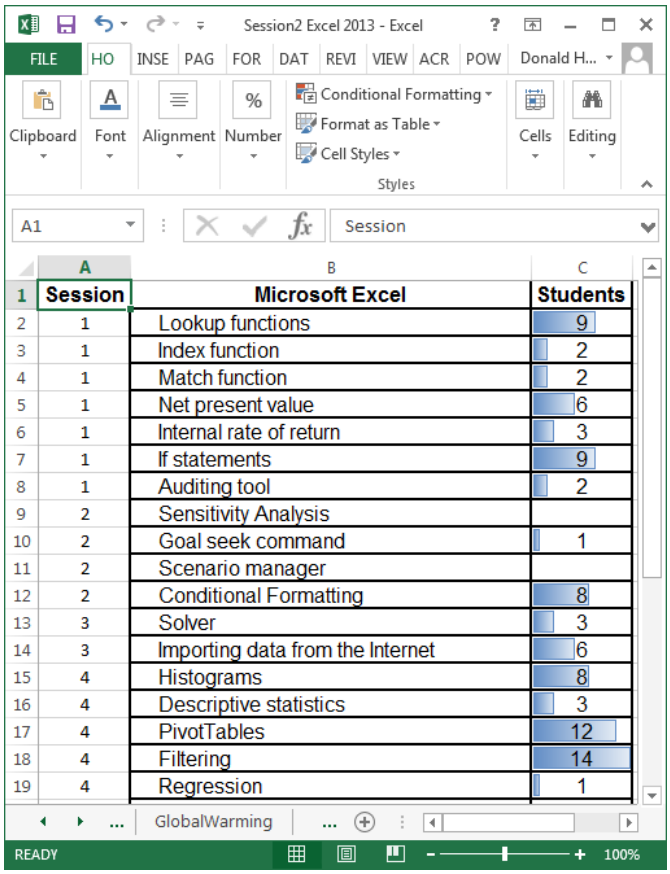
1. Highlight the values B3:B152
2. Click on the Home tab at the top of your spreadsheet.
3. Click on Conditional Formatting, Top/Bottom Rules, Top 10. Note that in the next screen, you can change top 10 to top anything. Select red as the color, then OK.
4. Click on Conditional Formatting a second time, Top/Bottom rules, Bottom 10. Change the font and background color to match the column heading.
5. For the next column, use the same approach but select the top and bottom 10%.
6. For the last column, use the same approach but select above and below average.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | | top 10 red, bottom 10 green | top 10% red, bottom 10% green | above average red, below average green |
| 2 | Year | Temp(relative to 15 degrees) | | |
| 3 | 1856 | -0.36 | -0.36 | -0.36 |
| 4 | 1857 | -0.47 | -0.47 | -0.47 |
| 5 | 1858 | -0.42 | -0.42 | -0.42 |
| 6 | 1859 | -0.23 | -0.23 | -0.23 |
| 7 | 1860 | -0.4 | -0.4 | -0.4 |
| 8 | 1861 | -0.41 | -0.41 | -0.41 |
| 9 | 1862 | -0.53 | -0.53 | -0.53 |
| 10 | 1863 | -0.26 | -0.26 | -0.26 |
| 11 | 1864 | -0.46 | -0.46 | -0.46 |
| 12 | 1865 | -0.25 | -0.25 | -0.25 |
| 13 | 1866 | -0.2 | -0.2 | -0.2 |

A feature of Excel is data bars to enhance data display. Data bars, color scales and icon sets can be used to enhance tabular data. Using the spreadsheet Experience, the steps to implement these new features are listed below.

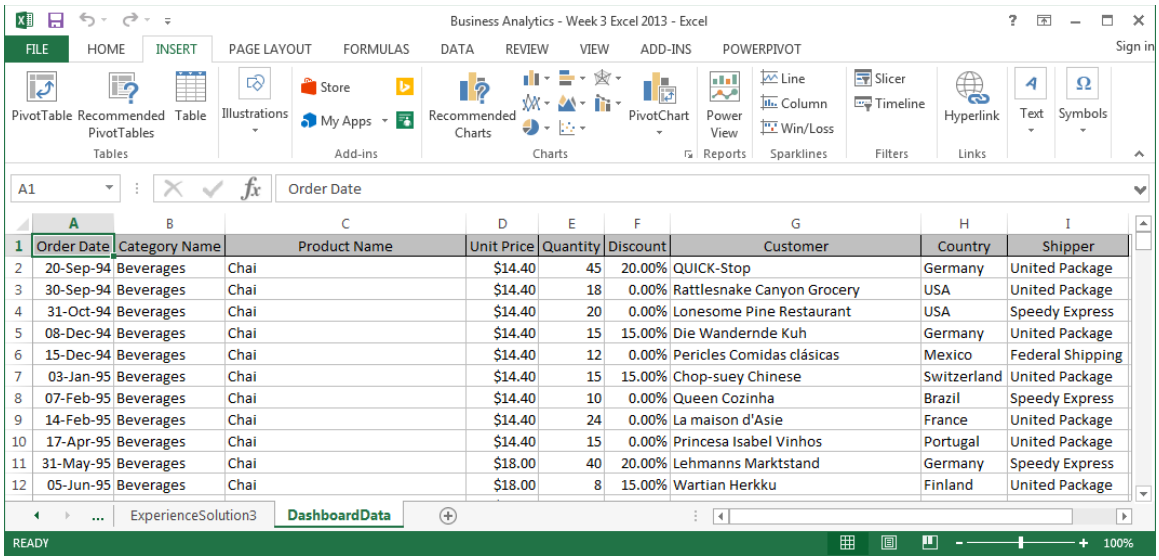
1. Select the range of values under the title students.
2. Click on Conditional Formatting.
3. First try Data Bars, and place your cursor over the options that appear on the right.
4. Next try Color Scales, placing your cursor over the options on the right.
5. Finally try Icon Sets, again placing your cursor over the options listed.
6. For each of these features, you can click on more rules and specify additional combinations.




| | A | B | C |
|---|---|---|---|
| 1 | Session | Microsoft Excel | Students |
| 2 | 1 | Lookup functions | 9 |
| 3 | 1 | Index function | 2 |
| 4 | 1 | Match function | 2 |
| 5 | 1 | Net present value | 6 |
| 6 | 1 | Internal rate of return | 3 |
| 7 | 1 | If statements | 9 |
| 8 | 1 | Auditing tool | 2 |
| 9 | 2 | Sensitivity Analysis | |
| 10 | 2 | Goal seek command | 1 |
| 11 | 2 | Scenario manager | |
| 12 | 2 | Conditional Formatting | 8 |
| 13 | 3 | Solver | 3 |
| 14 | 3 | Importing data from the Internet | 6 |
| 15 | 4 | Histograms | 8 |
| 16 | 4 | Descriptive statistics | 3 |
| 17 | 4 | PivotTables | 12 |
| 18 | 4 | Filtering | 14 |
| 19 | 4 | Regression | 1 |

## Session 3.6: Dashboards in Excel

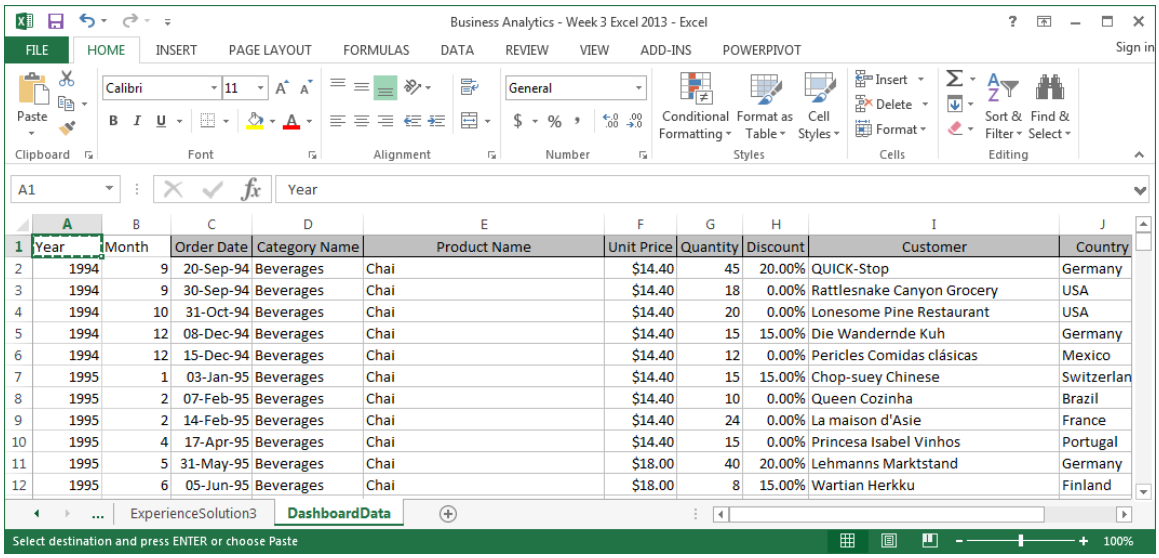
Dashboards allow multiple views of the data to be displayed simultaneously. For this exercise, use the tab labeled DashboardData. The data reflects orders of products by various customers and shipped by three shipping companies.



| Order Date | Category Name | Product Name | Unit Price | Quantity | Discount | Customer | Country | Shipper |
|---|---|---|---|---|---|---|---|---|
| 20-Sep-94 | Beverages | Chai | $14.40 | 45 | 20.00% | QUICK-Stop | Germany | United Package |
| 30-Sep-94 | Beverages | Chai | $14.40 | 18 | 0.00% | Rattlesnake Canyon Grocery | USA | United Package |
| 31-Oct-94 | Beverages | Chai | $14.40 | 20 | 0.00% | Lonesome Pine Restaurant | USA | Speedy Express |
| 08-Dec-94 | Beverages | Chai | $14.40 | 15 | 15.00% | Die Wandernde Kuh | Germany | United Package |
| 15-Dec-94 | Beverages | Chai | $14.40 | 12 | 0.00% | Pericles Comidas clásicas | Mexico | Federal Shipping |
| 03-Jan-95 | Beverages | Chai | $14.40 | 15 | 15.00% | Chop-suey Chinese | Switzerland | United Package |
| 07-Feb-95 | Beverages | Chai | $14.40 | 10 | 0.00% | Queen Cozinha | Brazil | Speedy Express |
| 14-Feb-95 | Beverages | Chai | $14.40 | 24 | 0.00% | La maison d'Asie | France | United Package |
| 17-Apr-95 | Beverages | Chai | $14.40 | 15 | 0.00% | Princesa Isabel Vinhos | Portugal | United Package |
| 31-May-95 | Beverages | Chai | $18.00 | 40 | 20.00% | Lehmanns Marktstand | Germany | Speedy Express |
| 05-Jun-95 | Beverages | Chai | $18.00 | 8 | 15.00% | Wartian Herkku | Finland | United Package |

The Order Date field identifies each order by specific day. Let's create two new columns and calculate the year and month for each order.

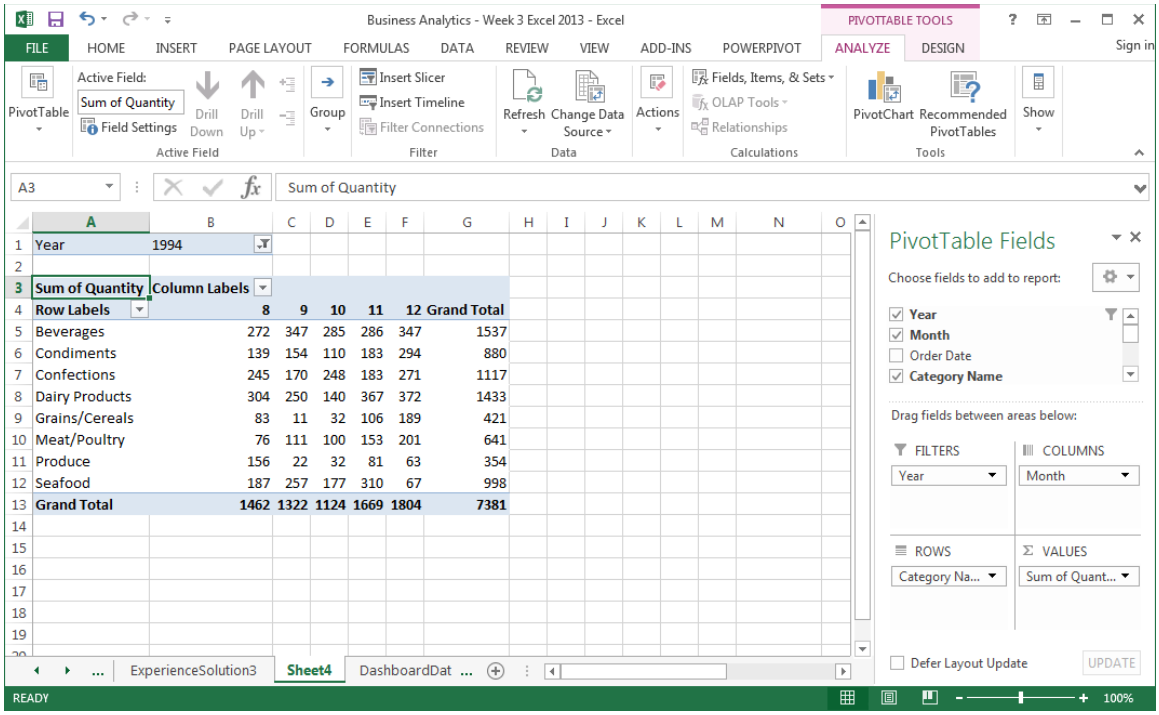
1. On the home tab, click on A above the first column, then click Insert, Insert Sheet Columns. Do this twice to insert two new columns to the left of your data.
2. Enter Year in cell A1; Month in cell B1
3. In cell A2, enter the formula =year(C2)
4. In cell B2, enter the formula =month(C2)
5. Copy the formulas down for all data



| Year | Month | Order Date | Category Name | Product Name | Unit Price | Quantity | Discount | Customer | Country |
|---|---|---|---|---|---|---|---|---|---|
| 1994 | 9 | 20-Sep-94 | Beverages | Chai | $14.40 | 45 | 20.00% | QUICK-Stop | Germany |
| 1994 | 9 | 30-Sep-94 | Beverages | Chai | $14.40 | 18 | 0.00% | Rattlesnake Canyon Grocery | USA |
| 1994 | 10 | 31-Oct-94 | Beverages | Chai | $14.40 | 20 | 0.00% | Lonesome Pine Restaurant | USA |
| 1994 | 12 | 08-Dec-94 | Beverages | Chai | $14.40 | 15 | 15.00% | Die Wandernde Kuh | Germany |
| 1994 | 12 | 15-Dec-94 | Beverages | Chai | $14.40 | 12 | 0.00% | Pericles Comidas clásicas | Mexico |
| 1995 | 1 | 03-Jan-95 | Beverages | Chai | $14.40 | 15 | 15.00% | Chop-suey Chinese | Switzerlan |
| 1995 | 2 | 07-Feb-95 | Beverages | Chai | $14.40 | 10 | 0.00% | Queen Cozinha | Brazil |
| 1995 | 2 | 14-Feb-95 | Beverages | Chai | $14.40 | 24 | 0.00% | La maison d'Asie | France |
| 1995 | 4 | 17-Apr-95 | Beverages | Chai | $14.40 | 15 | 0.00% | Princesa Isabel Vinhos | Portugal |
| 1995 | 5 | 31-May-95 | Beverages | Chai | $18.00 | 40 | 20.00% | Lehmanns Marktstand | Germany |
| 1995 | 6 | 05-Jun-95 | Beverages | Chai | $18.00 | 8 | 15.00% | Wartian Herkku | Finland |

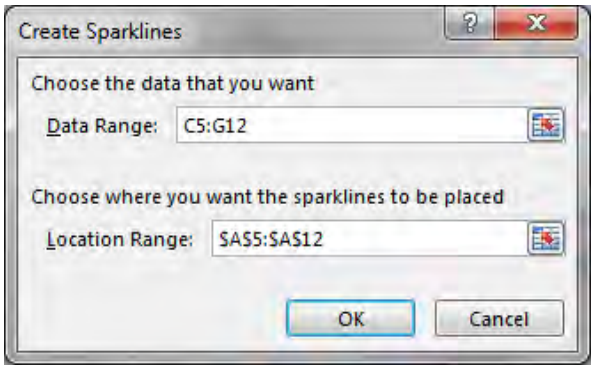
Next, insert a Pivot Table.

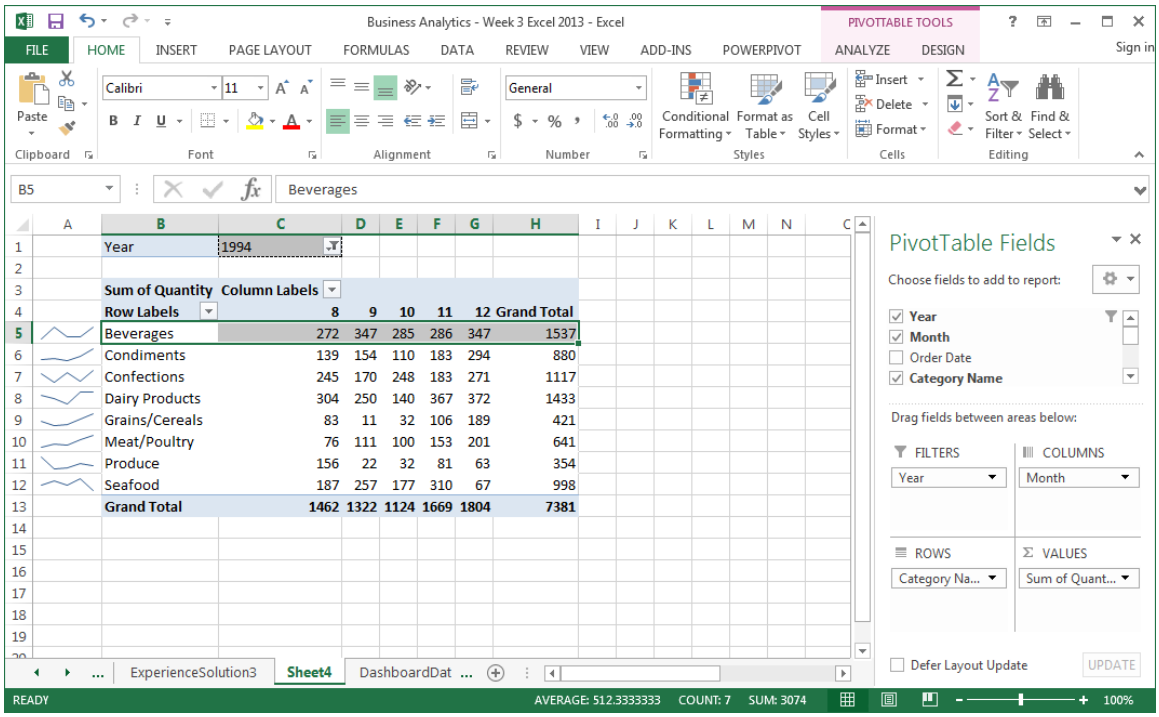
1. Click on the Insert tab, then Pivot Table
2. Drag Year to Filters (we will only use 1994 data)
3. Drag Month to Columns
4. Drag Category Name to Rows
5. Drag Quantity to Values
6. In cell B2 of the Pivot Table, use the drop down arrow to change All to 1994.




Now, let's insert a sparkline graph to the left of the data by first creating a column to the left of the Pivot Table.

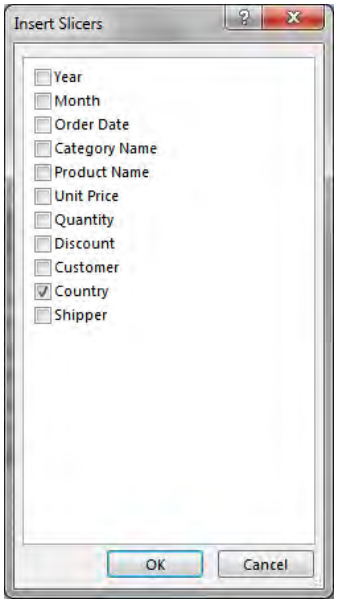
1. Click on the letter A at the top of the first column.
2. On the home tab, click on Insert, Insert Sheet Columns
3. Highlight the quantity data in cells C5 to G12
4. Click on the Insert tab at the top of the screen
5. Click on Sparkline: Line
6. In Location Range, enter A5:A12, the location to put the sparkline graphs
7. Click OK

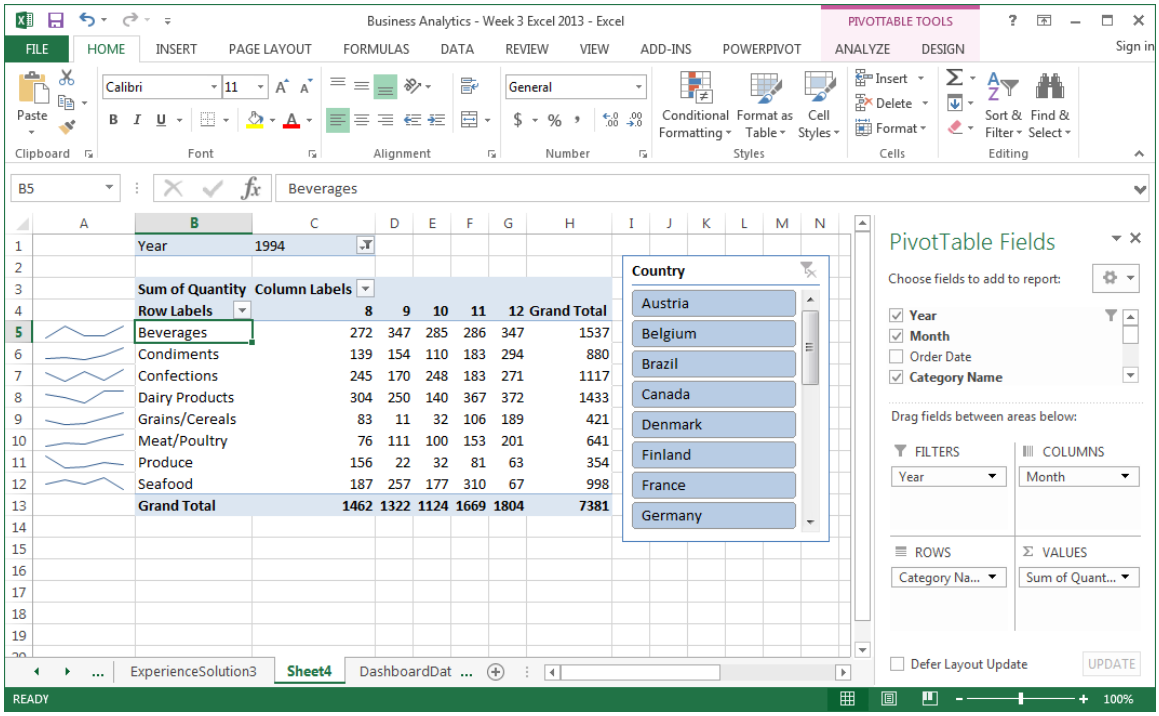




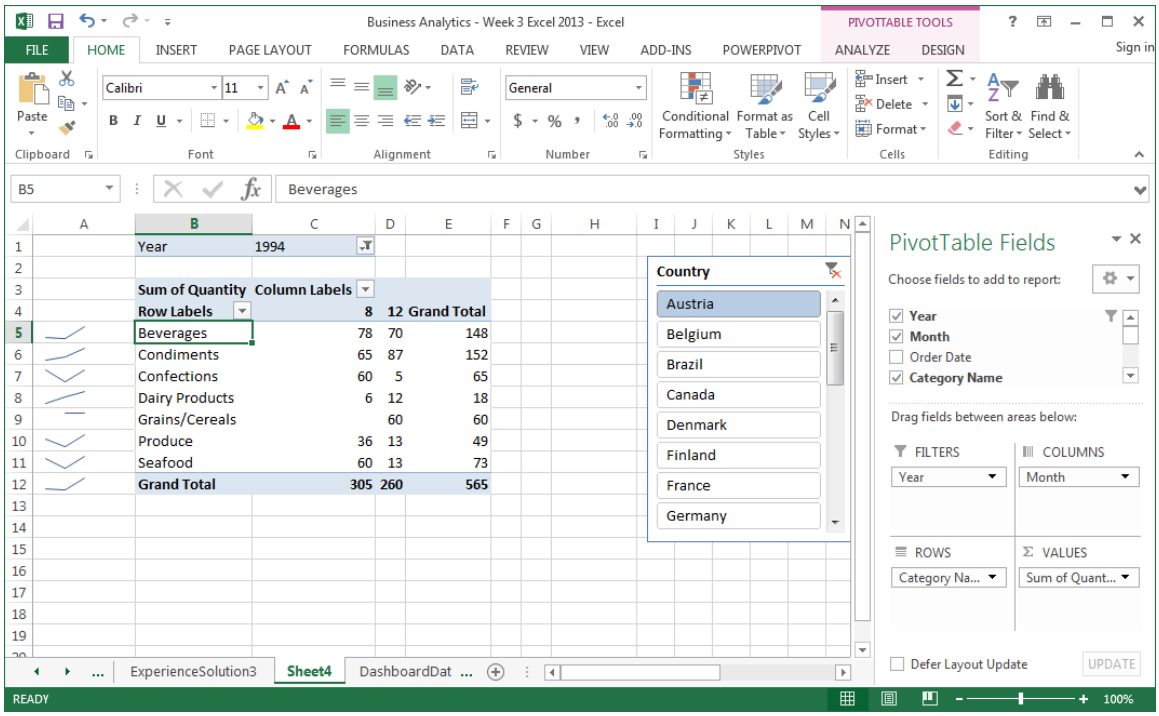

Finally, let's add a slicer, the ability to filter data on the fly.

1. Click on the Insert tab
2. Click on Slicer
3. Select Country, then OK






The Pivot Table and Sparkline initially shows all countries (note that countries are all highlighted in blue). By clicking on Austria, we can view only the data for Austria. By holding down the control key, you can select multiple countries. Click on the picture of a filter to the right of Country to restore all countries.



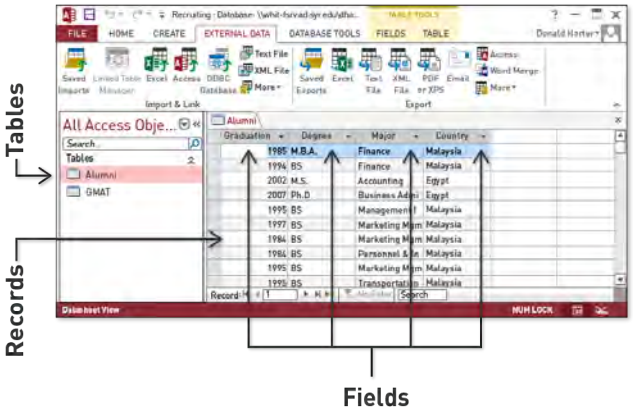
## What Is a Database?

- Collecon of gr  oups of data similar to an Excel workbook
- Different terminology because databases were created decades before spreadsheets

| Excel | Database |
|---|---|
| Spreadsheet | Table |
| Column | Field |
| Row | Record |



## Database Example



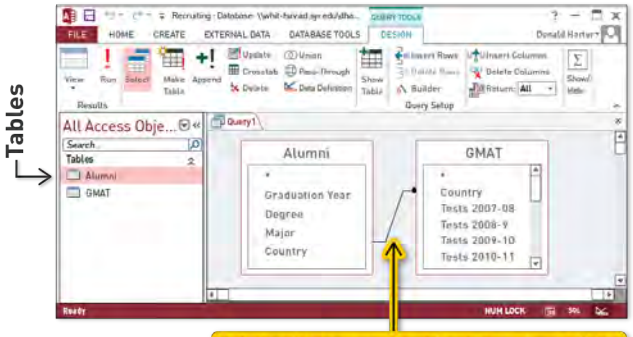



## Differences between Databases and Spreadsheets

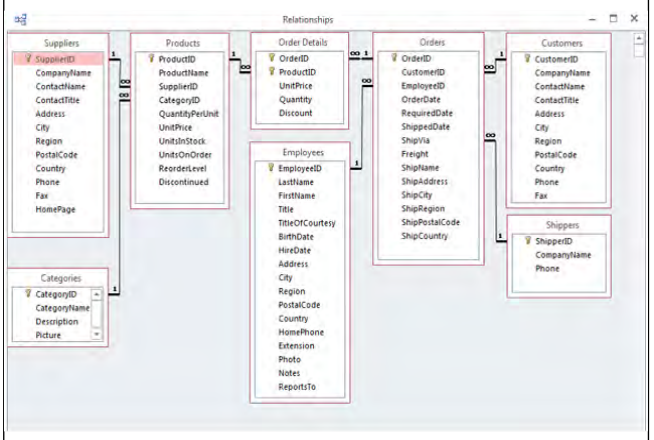
- Spreadsheet workbooks allow you to view one sheet at a me.
  - Formulas can cross spreadsheets, but difficult to combine data from mulple sheeąts
- Databases allow you to create relaonship  s between tables and view "joined" tables.



## Joined Tables



33 of 49

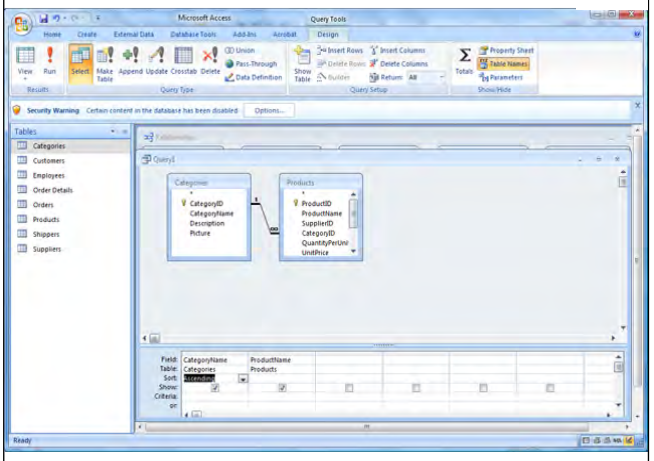



38 of 49

## Database Relationships

- 1-to-1: Each row (record) in one table matches one record (row) in the second.
  - Example: Table of high school students and table of high school football players
- 1-to-many: Each record in the first table links to many records in the second.
  - Example: Couple and dependents on income tax form

39 of 49




45 of 49

## Queries

- Retrieve data from database
- Can span mulple t ables, using relaonship t o join them
- Can retrieve raw data, group and sort data, perform calculaons, and filt er data
- Look like a table
- Can be saved and used in other queries (subqueries)

## Hands-On Activity

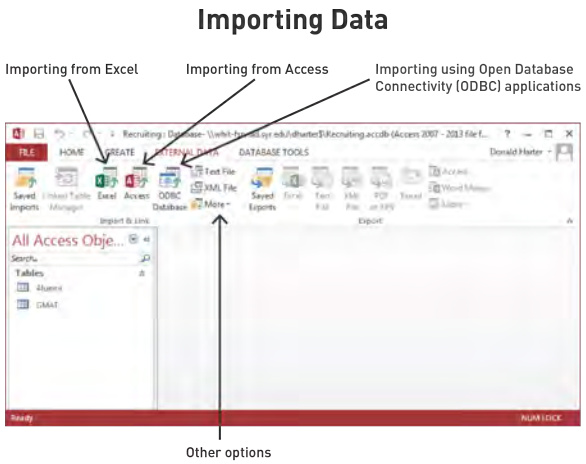
- Enter data.
- Connect data with relaonship s.
- Perform queries.

View All View Keyframes

## Accessing and Importing Data

- Import from other files or systems.
  - Excel, Microsoft Access, text files, XML data, etc.
- Databases can read and import data stored in other systems.

4 of 13




5 of 13

## Hands-On Activity

- Import data from various sources.
- Begin building first database.

13 of 13

View All View Keyframes

## Relationships

- Determine how data is used when retrieving data from mulple t ables
- Do not typically exist when data tables are imported into database
- Must be manually idenfied/ created
- Created by finding common field(s) between two tables

5 of 13

## Relationships

11 of 13

## Hands-On Activity

Manually create joins and relaonship s in Access.

13 of 13

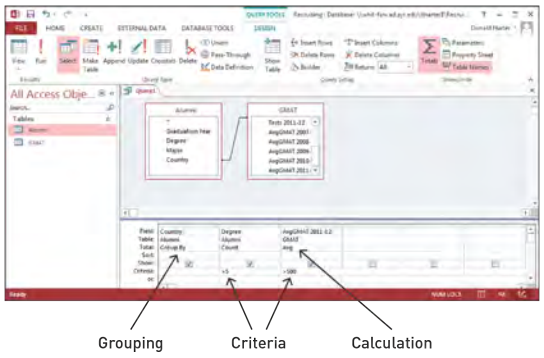
View All View Keyframes

## Simple Queries

- Simple queries allow retrieval of data.
  - By idenf  ying fields (columns) in tables to be retrieved
- More complex queries group data and perform calculaons.
  - E.g., number of alumni from a country or average GMAT for country
- Criteria allow filtering of data.
  - E.g., limit retrievals to countries with more than five alumni or average GMAT score more than 500

7 of 15

## Simple Query Example




8 of 15

## Function of a Simple Query

- Build on simple data by grouping, calculang  , and using criteria to filter.
- Pracce in hands-on acvity  .

15 of 15

View All View Keyframes

## Dirty Data

- May be due to misspelled or mistyped informaon in da tabase
  - Difficult to correct, but new matching technology can make recommendaons
- May result when two or more parts of database use different idenfier s for same item
  - E.g., "New York City" and "NYC"
  - Can be idenfied using an op on in r elaonship joins

6 of 39

## Types of Joins

- **Inner join:** only includes records from the two tables when the joining fields from each table match exactly
- **Left join:** includes all records from left table and only records from right table where there is an exact match
  - Can result in voids
- **Right join:** includes all records from right table and only records from left table where there is an exact match
- Left and right joins give broader spectrum of data than inner join.

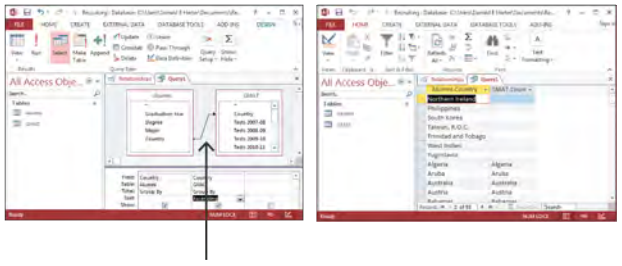
15 of 39

## Identifying Relationships with Dirty Data

- Run query using left join.
  - Look for records with missing data that should have been retrieved from right side.
- Run query using right join.
  - Look for records with missing data that should have been retrieved from left side.
- Correct mismatches due to dirty data.

21 of 39

## Dirty Data Example: Left Join



Left join: Direction of arrow

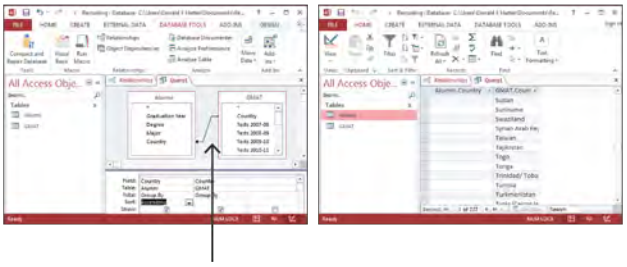
22 of 39

## Using Left Join

- Selects all data from le t able and only exact matches from right table
- GMAT database uses "Taiwan," not "Taiwan, R.O.C.," so no exact match in alumni database
- Picks up data from le t able and finds what is missing from right table

31 of 39

## Dirty Data Example: Right Join



Right join: Direction of arrow

32 of 39

### Fixing Dirty Data

- Relaonship s can idenf y inconsistencies in data.
- Correcng da ta in tables gives a more accurate representaon in queries.

39 of 39

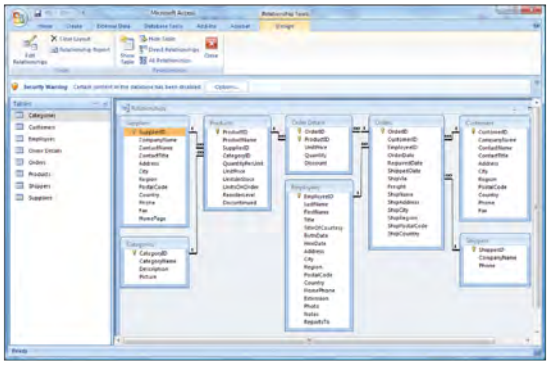
View All View Keyframes

## Complex Queries

- Can retrieve data from larger numbers of tables and enr e databases
- Must have correctly defined relaonship s and dirty data corrected
- Incorporate filters to focus retrieval

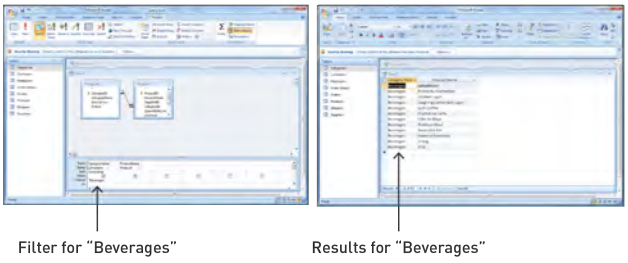
4 of 14

## Complex Database



5 of 14

## Complex Query with Filters




7 of 14

## Using Complex Queries

- Can expand to mulple t ables, even thousands
- Need team to keep track of data and relaonship s
- Not much more difficult than simple queries

14 of 14

# SCM 651: Business Analytics

1

Week 4

# Agenda



- Homework #1: review
- Homework #2: discussion
  - Google analytics hints
- Review of concepts
- Group discussion of articles
  - Minding the Analytics Gap
  - Innovating with Analytics

# Homework #1



- Pivot table
- Pivot charts
- Correlation – any negative correlations?
- Regression – any non-intuitive results?
- Prediction model
- Non-intuitive results and additional data

# Homework #2



- Campaign time frames, costs, effectiveness
- Future campaign: geographic regions, key words, day of week and time of day
- Allocation of costs by program and region
- Performance measures
- Other data that would be helpful

- Example of how to find information in Google Analytics

# Week 4 - Review



- Importing data into Access
  - Excel spreadsheets, text files, XML, et al.
- Relationships can be created
  - When retrieving data
  - Between common fields in two tables

# Week 4 - Review



- Grouping
  - Collapses together rows of data according to the field grouped
  - It does not form calculations
- Criteria
  - Identifies a subset of data
- Calculations
  - Can be used to calculate min, max, sums, average, standard deviations, variance, counts, etc.

# Week 4 - Review



- Dirty data
  - To identify inconsistent key fields used to match two tables, use both a left and right join
- Complex queries
  - Important to clean out dirty data and ensure that you have correct relationships before performing a complex query

# Article #1: Minding the Analytics Gap



- Minding the Analytics Gap
  - What is the barrier to using analytics?
  - What can be done on the production and consumption side of analytics to overcome this barrier?
  - Describe the three levels of analytics maturity

# Article #1: Minding the Analytics Gap



- Minding the Analytics Gap
  - What is the barrier to using analytics?
    - Translating analytics into business actions
  - What can be done on the production and consumption side of analytics to overcome this barrier?
    - Production: data analysts can learn more about business
    - Consumption: managers can takes steps to become savvier at understanding analytical results
  - Describe the three levels of analytics maturity
    - Analytically Challenged: lack data management and analytical skills
    - Analytical Practitioners: use analytics for operational purposes; "just good enough data"
    - Analytical Innovators: more strategic in analytics applications; place high value on data, higher levels of data management and analytical skills

# Article #2: Innovating with Analytics



- Innovating with Analytics
  - Describe the three characteristics of analytics innovators

# Article #2: Innovating with Analytics



- Innovating with Analytics
  - Describe the three characteristics of analytics innovators
    - Tend to use more data: strong correlation between analytics to create competitive advantage and how much data they use
    - Manage information more effectively: strong correlation between competitive advantage and how company manages information transformation (capturing data, analyzing, aggregating, integrating, insights to strategy)
    - Speed: stronger need for speed, process to analyze data more quickly

# Microsoft Access

## Session 4.4: Overview

This document is a quick reference guide for reviewing the techniques in the laboratory.

For these exercises, download the files:
“Business Analytics – Week 4 Instructions.doc”
“Business Analytics – Week 4 Excel 2013.xls”
“Business Analytics – Week 4 Orders NoRelationships Database.acc”
“Business Analytics – Week 4 Orders Relationships Database.acc”

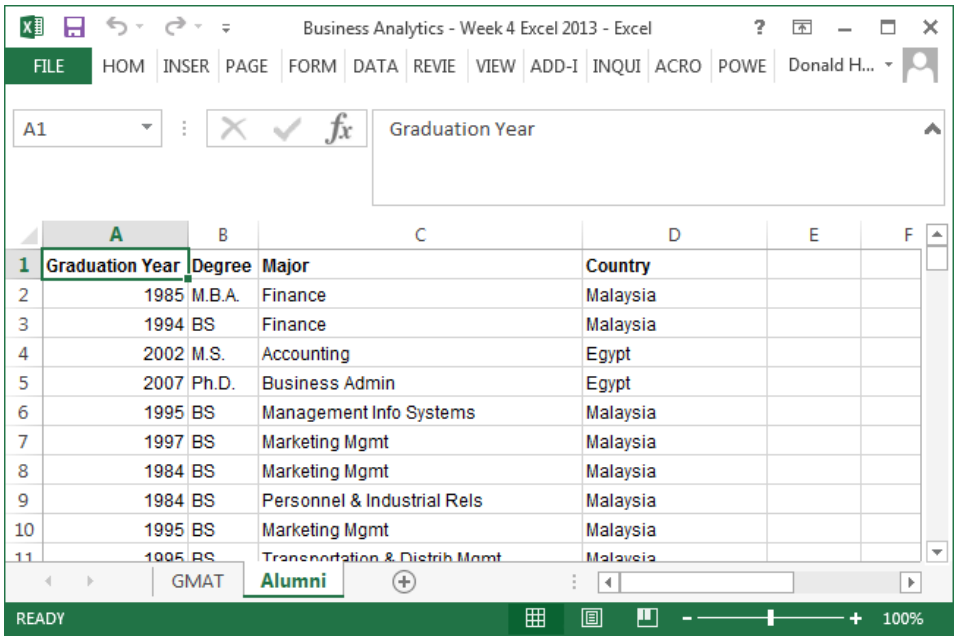
## Background

In Excel, we could have multiple spreadsheets within a workbook. However, these spreadsheets did not necessarily relate to each other. In Microsoft Access, we will use a similar concept, but it’s called tables. A table is like a spreadsheet. The key difference is that multiple tables can relate to each other and be combined to form calculations, summaries, queries and reports. When tables are combined, we say they are joined. Tables can be joined based on relationships. This tutorial explores tables, relationships, and their uses in retrieving and adding data to a database.

## Importing Data

Generally, if you are starting from scratch, it’s easiest to build data into an Excel spreadsheet and import it into Access. Excel allows data to easily be imported from the web or entered directly. Editing and formulas can manipulate the data before you convert it to Access format.

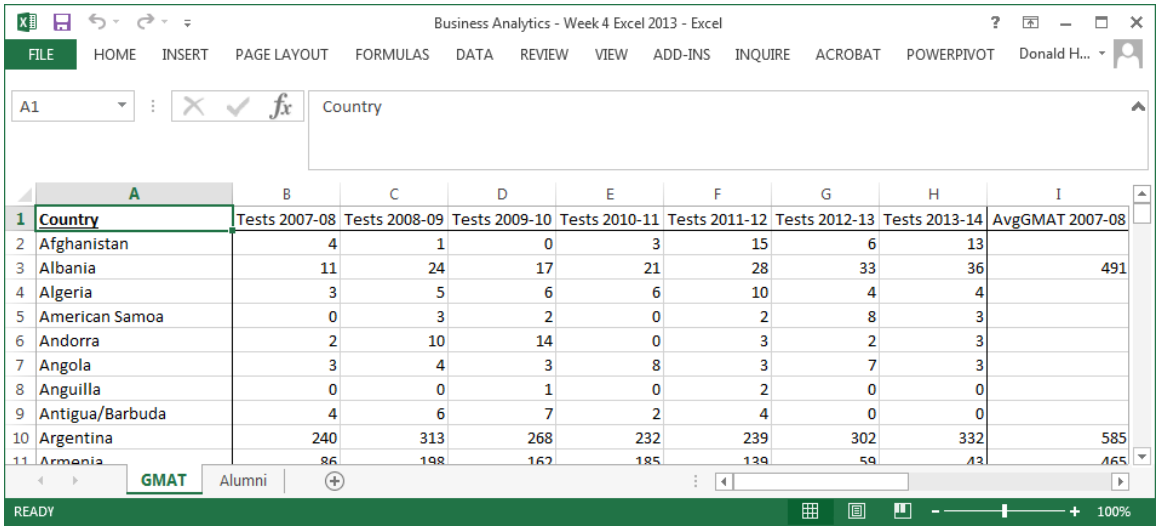
Let’s first examine data in the Alumni tab of your spreadsheet.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Graduation Year | Degree | Major | Country | | |
| 2 | 1985 | M.B.A. | Finance | Malaysia | | |
| 3 | 1994 | BS | Finance | Malaysia | | |
| 4 | 2002 | M.S. | Accounting | Egypt | | |
| 5 | 2007 | Ph.D. | Business Admin | Egypt | | |
| 6 | 1995 | BS | Management Info Systems | Malaysia | | |
| 7 | 1997 | BS | Marketing Mgmt | Malaysia | | |
| 8 | 1984 | BS | Marketing Mgmt | Malaysia | | |
| 9 | 1984 | BS | Personnel & Industrial Rels | Malaysia | | |
| 10 | 1995 | BS | Marketing Mgmt | Malaysia | | |
| 11 | 1995 | BS | Transportation & Distrib Mgmt | Malaysia | | |

Data includes graduation year, degree, major and country for each student.

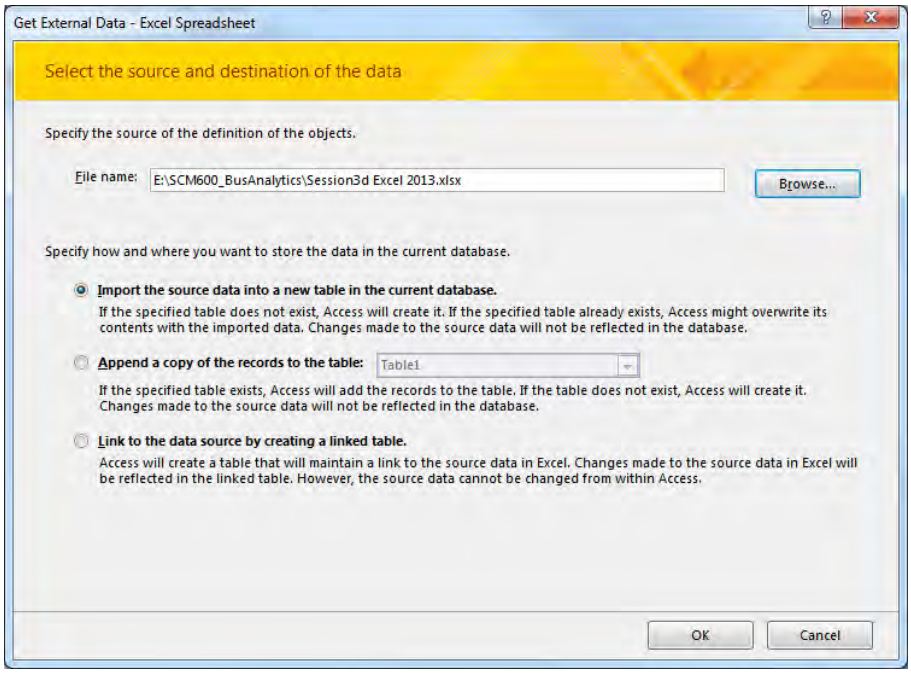
Data for GMAT test takers is shown in the GMAT tab.



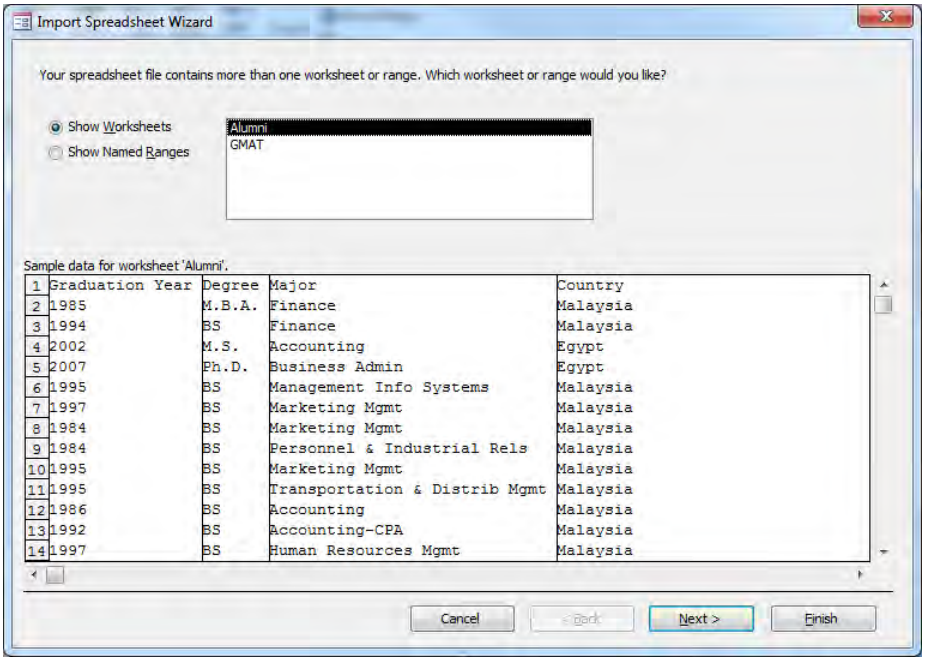

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Country | Tests 2007-08 | Tests 2008-09 | Tests 2009-10 | Tests 2010-11 | Tests 2011-12 | Tests 2012-13 | Tests 2013-14 | AvgGMAT 2007-08 |
| 2 | Afghanistan | 4 | 1 | 0 | 3 | 15 | 6 | 13 | |
| 3 | Albania | 11 | 24 | 17 | 21 | 28 | 33 | 36 | 491 |
| 4 | Algeria | 3 | 5 | 6 | 6 | 10 | 4 | 4 | |
| 5 | American Samoa | 0 | 3 | 2 | 0 | 2 | 8 | 3 | |
| 6 | Andorra | 2 | 10 | 14 | 0 | 3 | 2 | 3 | |
| 7 | Angola | 3 | 4 | 3 | 8 | 3 | 7 | 3 | |
| 8 | Anguilla | 0 | 0 | 1 | 0 | 2 | 0 | 0 | |
| 9 | Antigua/Barbuda | 4 | 6 | 7 | 2 | 4 | 0 | 0 | |
| 10 | Argentina | 240 | 313 | 268 | 232 | 239 | 302 | 332 | 585 |
| 11 | Armenia | 86 | 198 | 162 | 185 | 139 | 59 | 43 | 465 |

Let's import the data from Excel to Access by using the Access Import capability.

1. Open Microsoft Access by clicking on the Start button, all Programs, Microsoft Office, Access.
2. Click on Blank Database.
3. In the File Name field, enter Recruiting, then click Create.
4. Click on the tab External Data, then Excel in the Import section.
5. Use the Browse button to find your spreadsheet, then click OK.

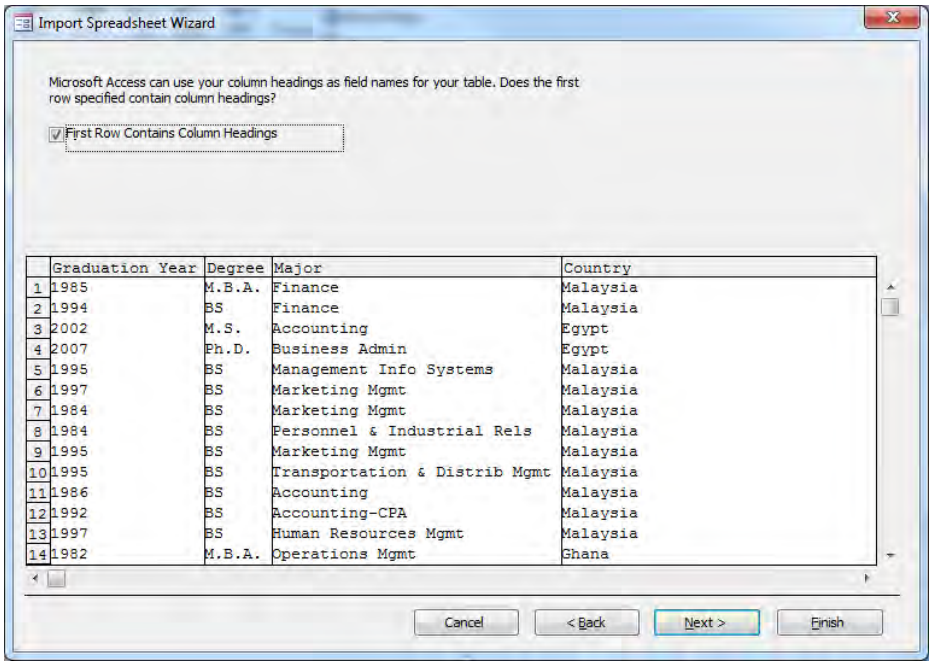



6. Select the appropriate worksheet from the workbook. In this case, use Alumni, then Next.



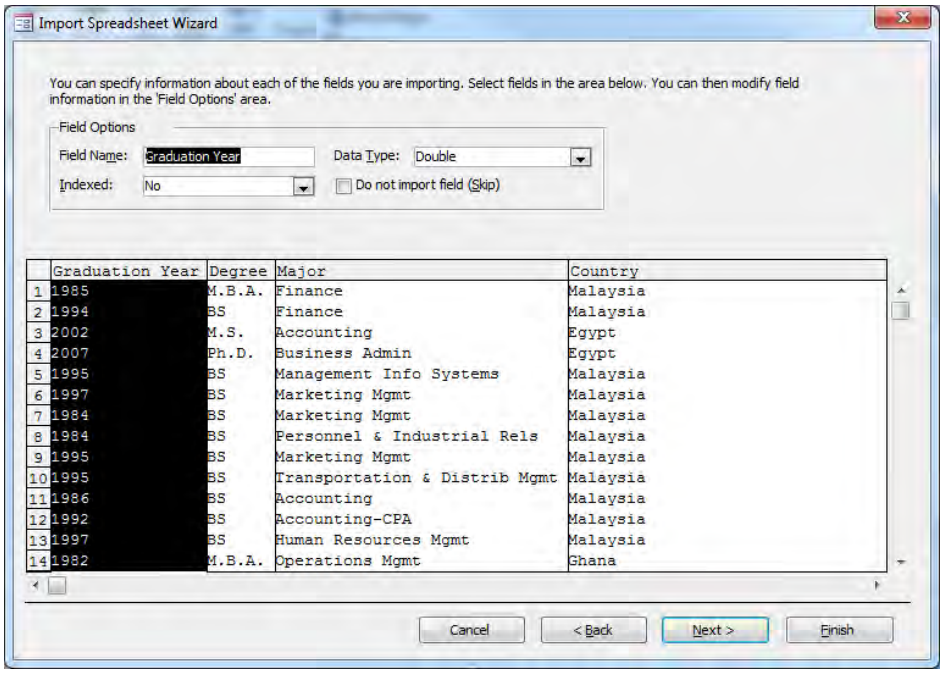

| | Graduation Year | Degree | Major | Country |
|---|---|---|---|---|
| 1 | 1985 | M.B.A. | Finance | Malaysia |
| 2 | 1994 | BS | Finance | Malaysia |
| 3 | 2002 | M.S. | Accounting | Egypt |
| 4 | 2007 | Ph.D. | Business Admin | Egypt |
| 5 | 1995 | BS | Management Info Systems | Malaysia |
| 6 | 1997 | BS | Marketing Mgmt | Malaysia |
| 7 | 1984 | BS | Marketing Mgmt | Malaysia |
| 8 | 1984 | BS | Personnel & Industrial Rels | Malaysia |
| 9 | 1995 | BS | Marketing Mgmt | Malaysia |
| 10 | 1995 | BS | Transportation & Distrib Mgmt | Malaysia |
| 11 | 1986 | BS | Accounting | Malaysia |
| 12 | 1992 | BS | Accounting-CPA | Malaysia |
| 13 | 1997 | BS | Human Resources Mgmt | Malaysia |

7. Access will try to detect if you have labels in the first row. Confirm and click Next.

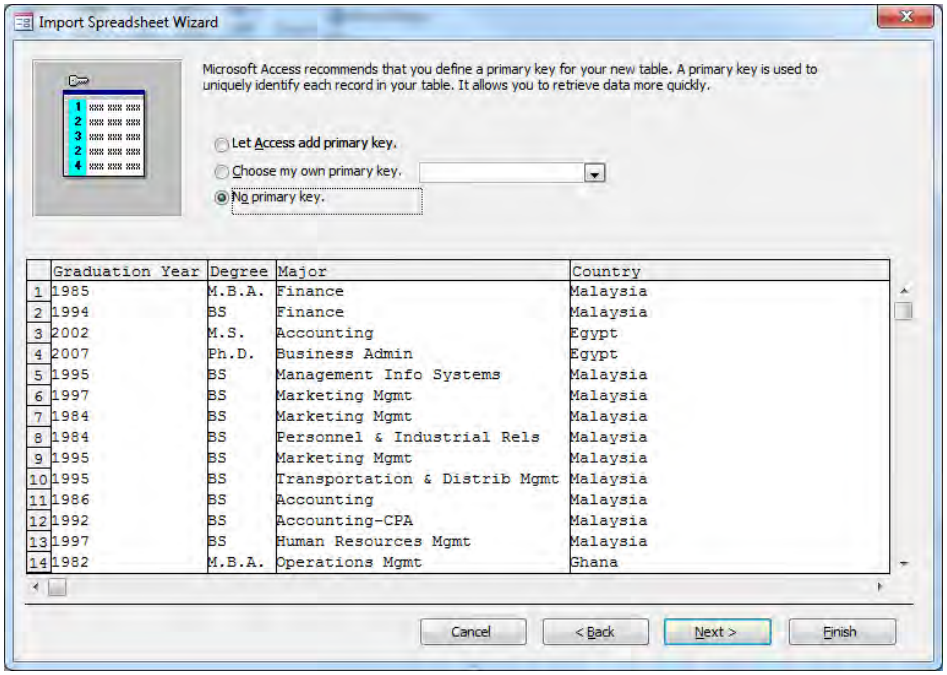



| | Graduation Year | Degree | Major | Country |
|---|---|---|---|---|
| 1 | 1985 | M.B.A. | Finance | Malaysia |
| 2 | 1994 | BS | Finance | Malaysia |
| 3 | 2002 | M.S. | Accounting | Egypt |
| 4 | 2007 | Ph.D. | Business Admin | Egypt |
| 5 | 1995 | BS | Management Info Systems | Malaysia |
| 6 | 1997 | BS | Marketing Mgmt | Malaysia |
| 7 | 1984 | BS | Marketing Mgmt | Malaysia |
| 8 | 1984 | BS | Personnel & Industrial Rels | Malaysia |
| 9 | 1995 | BS | Marketing Mgmt | Malaysia |
| 10 | 1995 | BS | Transportation & Distrib Mgmt | Malaysia |
| 11 | 1986 | BS | Accounting | Malaysia |
| 12 | 1992 | BS | Accounting-CPA | Malaysia |
| 13 | 1997 | BS | Human Resources Mgmt | Malaysia |
| 14 | 1982 | M.B.A. | Operations Mgmt | Ghana |

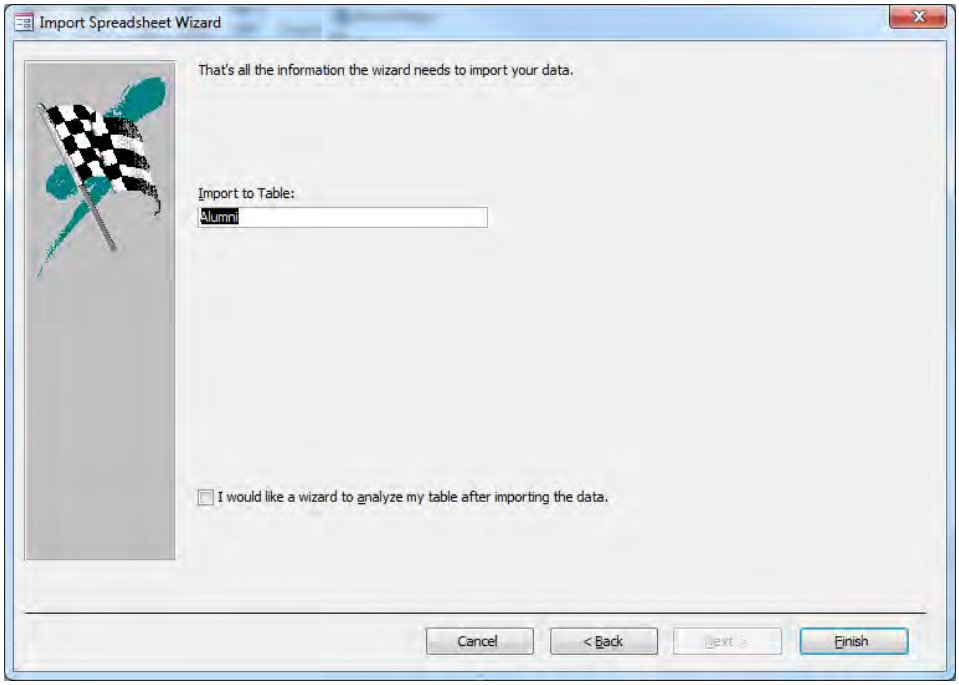
8. The next screen allows you to determine whether each column is text, numeric, etc., and set the characteristics. Access is pretty good at setting these automatically. Click Next.



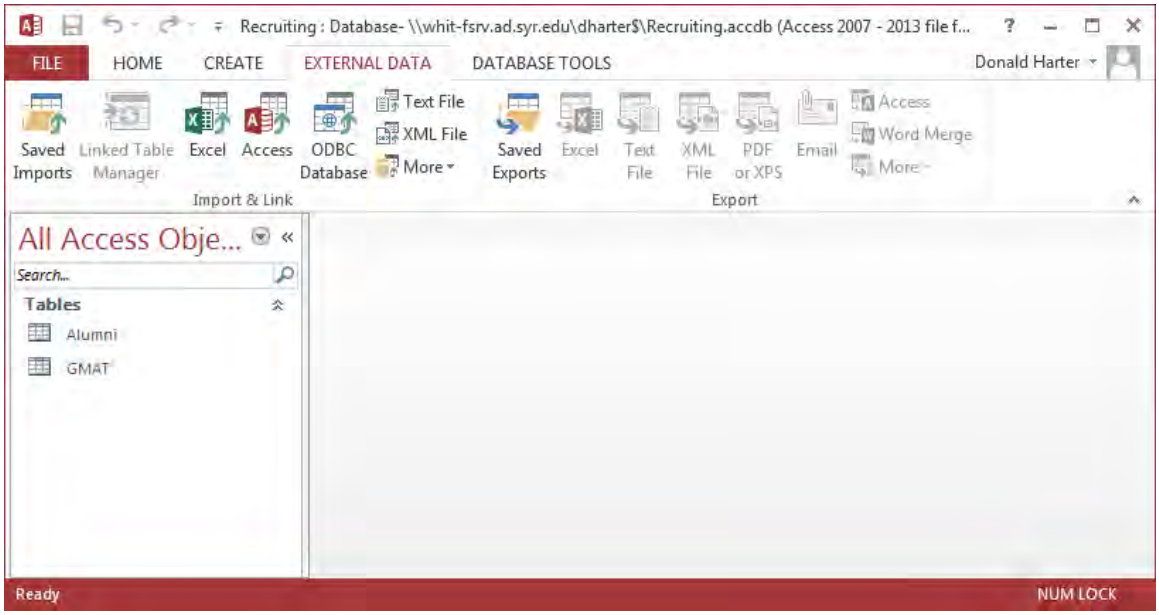

9. Next, Access allows you to add a primary key. Primary keys allow you to uniquely identify each row in a table. However, if you let Access add one, it always adds an arbitrary key that means nothing. Click on No primary key, then Next.



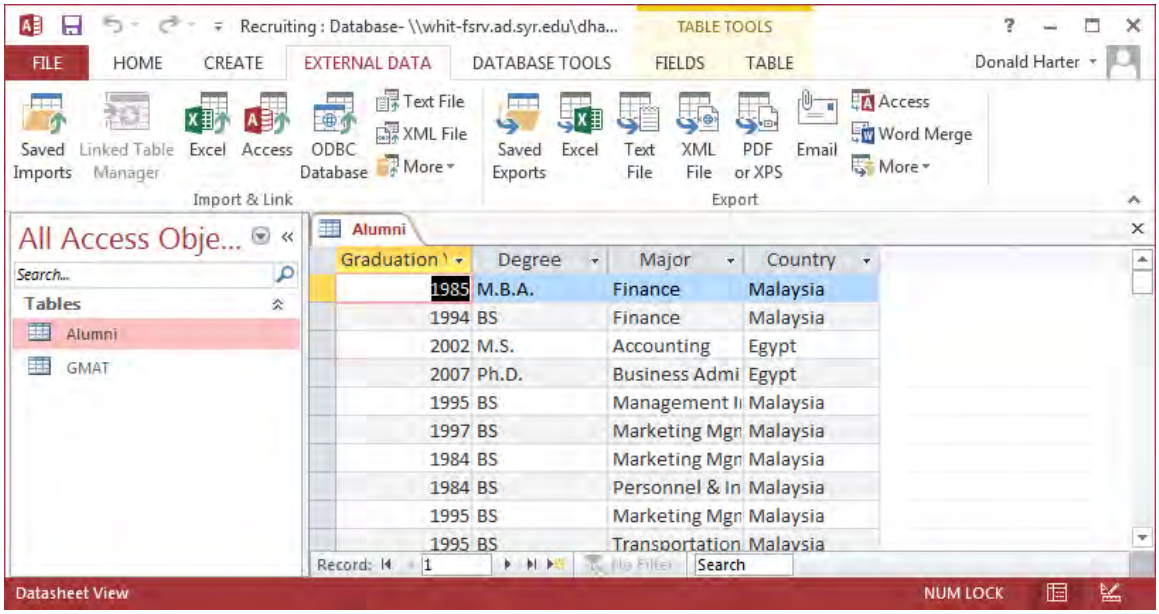

10. The last screen allows you to name the table. Let's call it Aumni. Click Finish.




Now follow the same set of steps to add the GMAT data from your spreadsheet. Both tables should appear in your Access database.



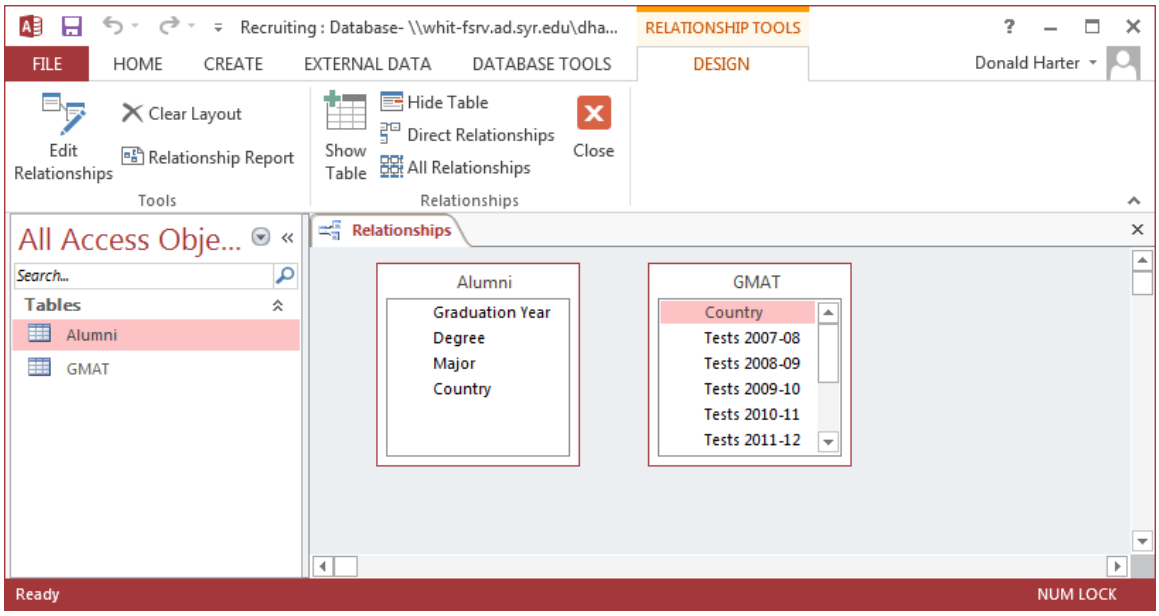
Double click on the table name Alumni to view the data.



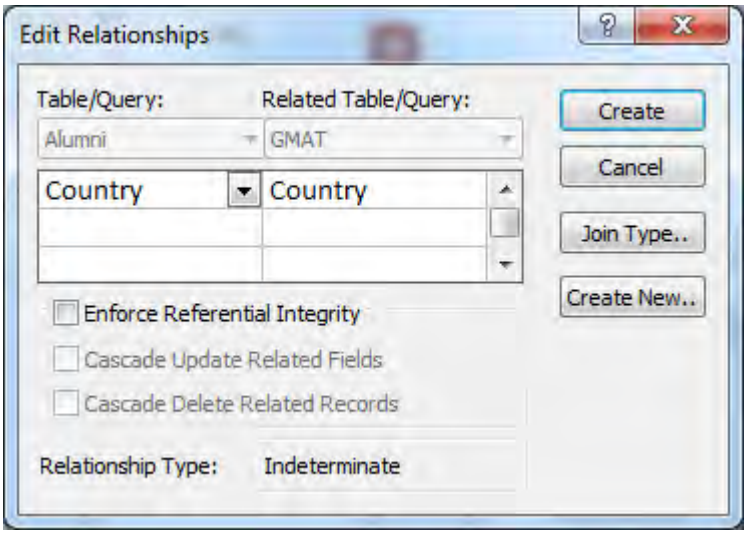

**Session 4.5: Relationships**

Relationships describe how tables relate to each other. Click on the Database Tools tab, then relationships. You should see the following pop-up.

Click on Alumni, then Add. Click on GMAT, then Add. Click Close.




There is no relationship (yet) between these table. Click on Country in Alumni, then drag to Country in GMAT.



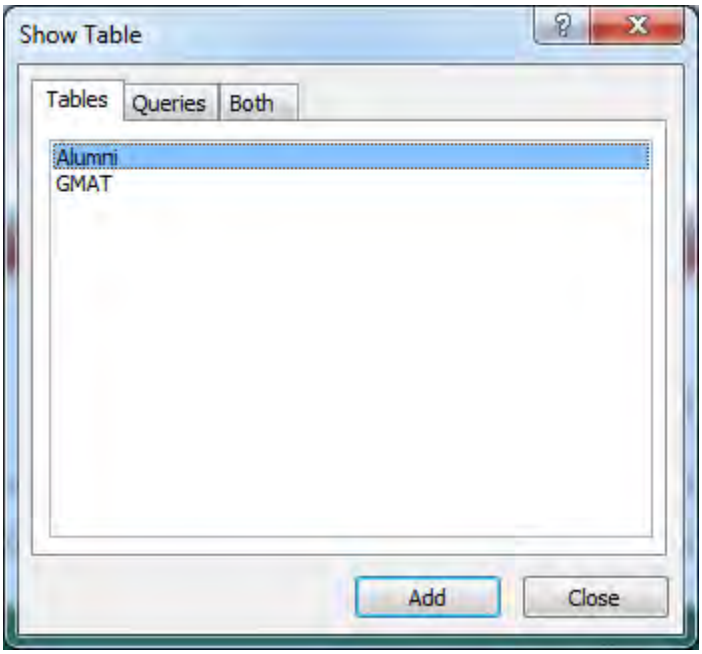

Now click Create. This “joins” the two tables using the field Country in each table.

Click on the X to the right of Relationships. When it asks if you want to save Relationships, click Yes.

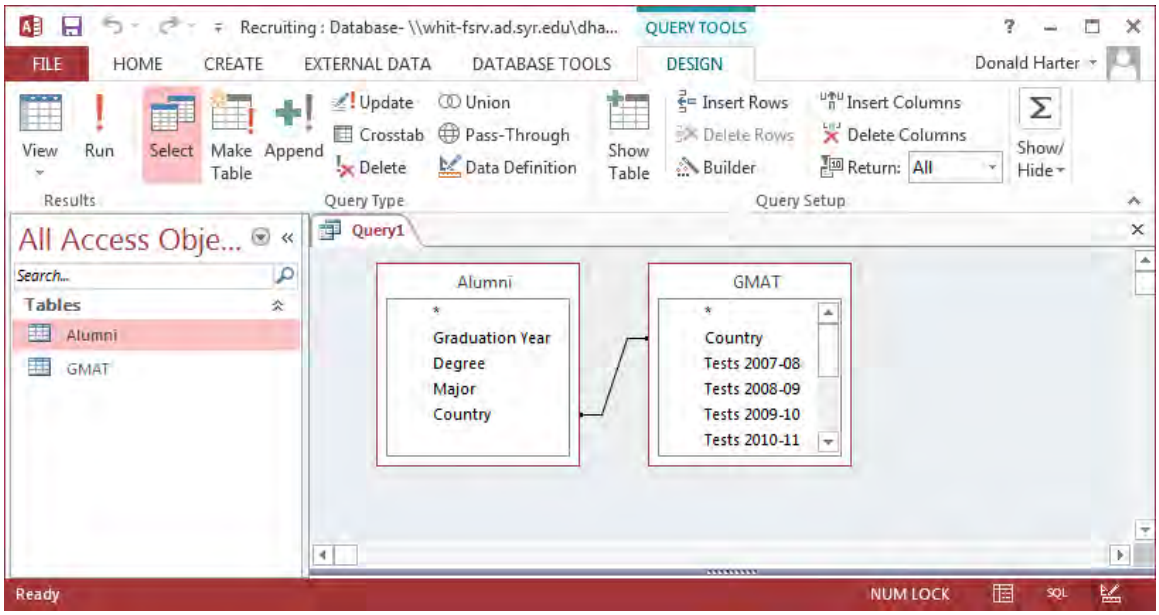
**Session 4.6: Simple Queries**

When you want to retrieve specific data from a database, you can use a query. A query is like a filter in Excel, but much more powerful. It can allow you to restrict what data is displayed, sort the data, and perform calculations. However, dirty data can corrupt your queries. Let's see how.

To create a query, click on CREATE, then Query Design.

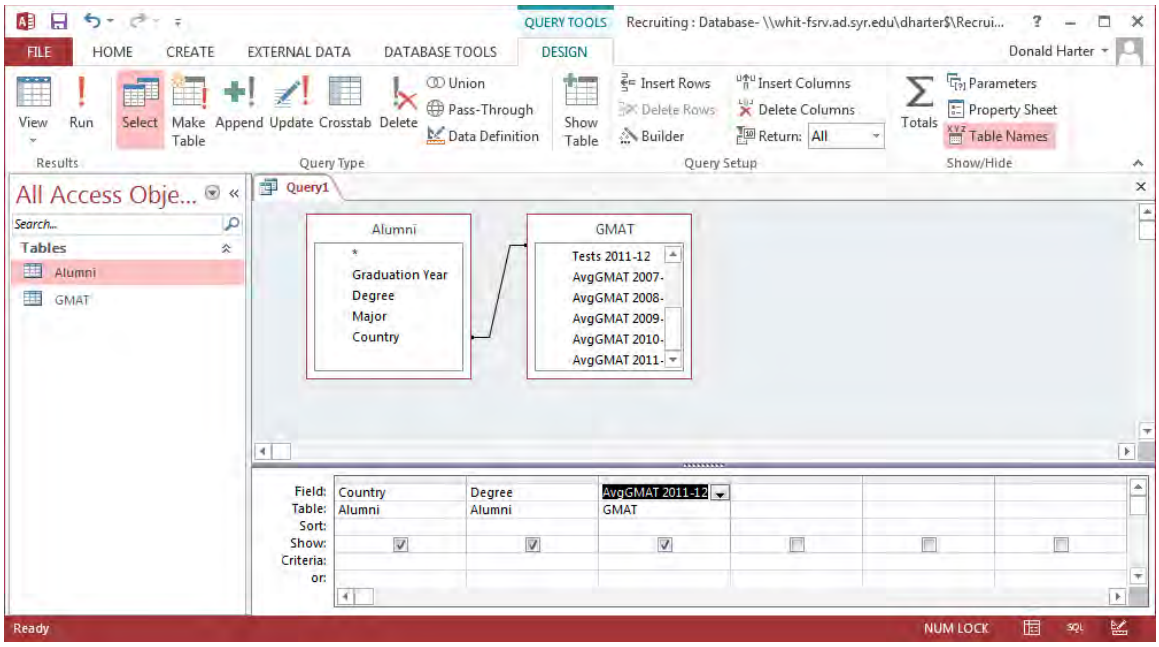



Click on Alumni, then Add; GMAT, then Add.

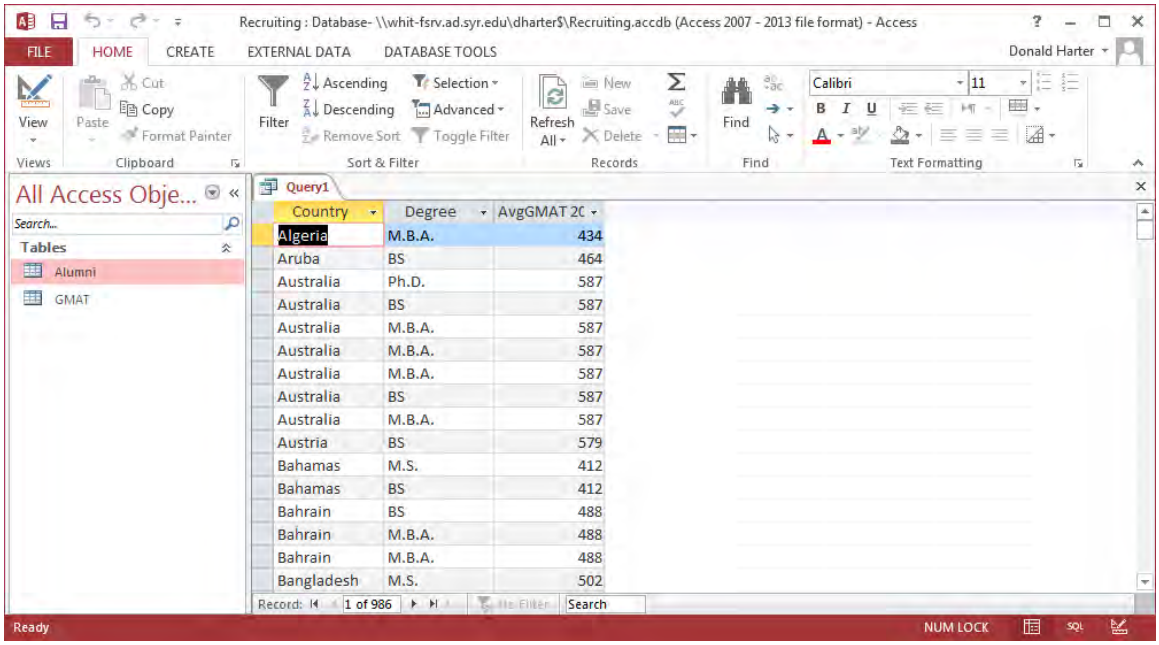



Let's compare Countries in the Alumni table to Countries in the GMAT table.

1. In the table Alumni, double click Country.
2. In the table Alumni, double click Degree
3. In the table GMAT, double click AvgGMAT 2011-12.
4. Click Run !



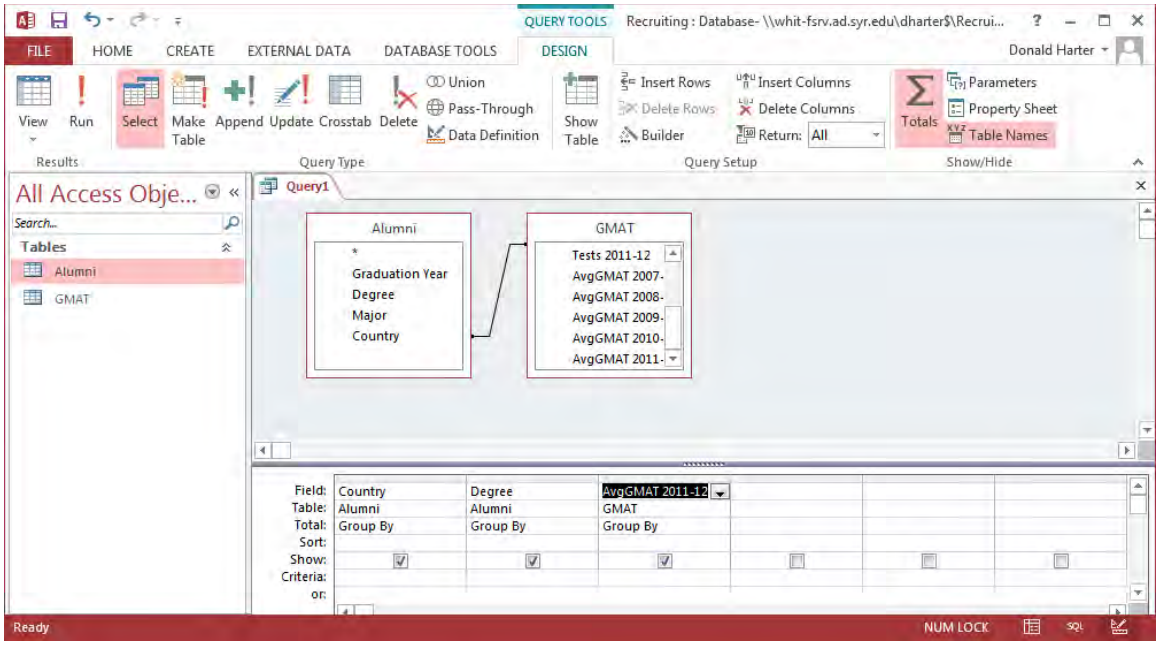

Now click Run ! in the upper right corner.




This lists each student, rather than a summary.

Click on View, Design View, to go back to the query.

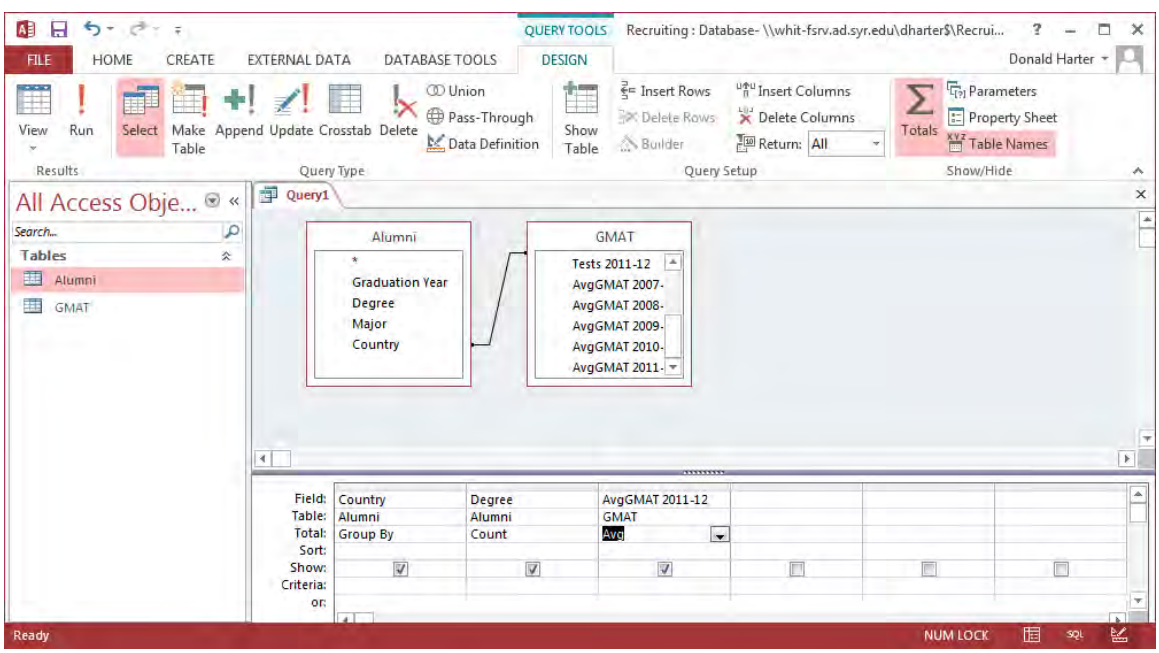
Click on the Sigma (summation sign).



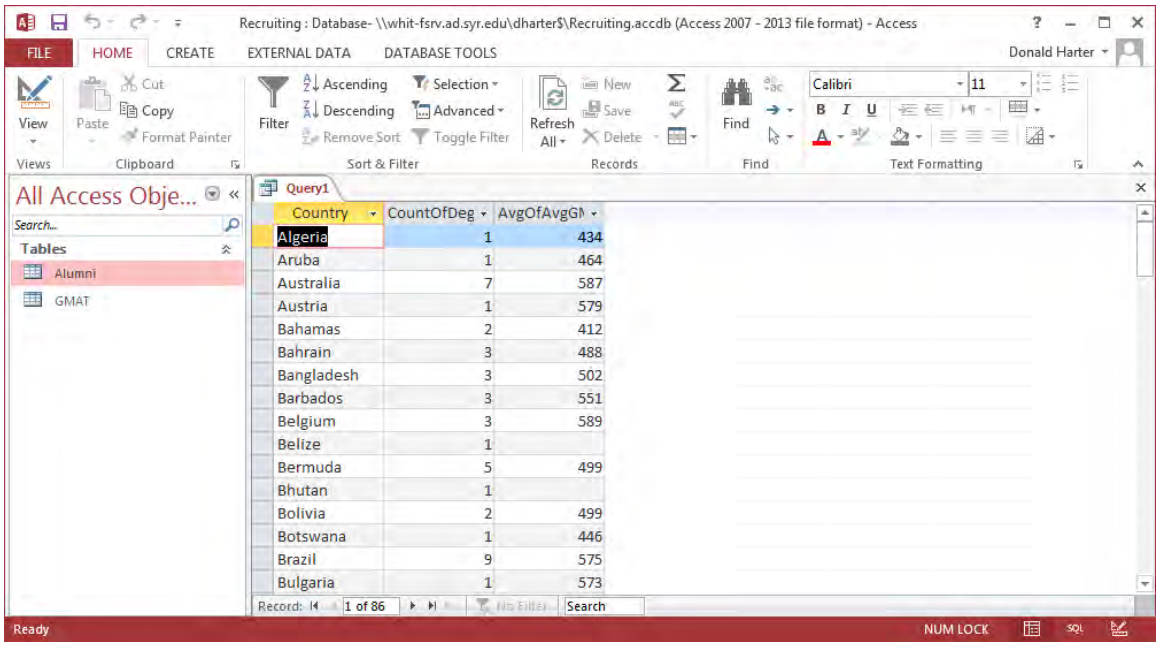

It now adds the Total line Group By. We want to Group By country, but we really want to count degrees and average the GMAT scores.

Where it says Group By under Degree, use the drop down menu to change to Count.

Where it says Group By under AvgGMAT 2011-2012, use the drop down menu to change to Average.

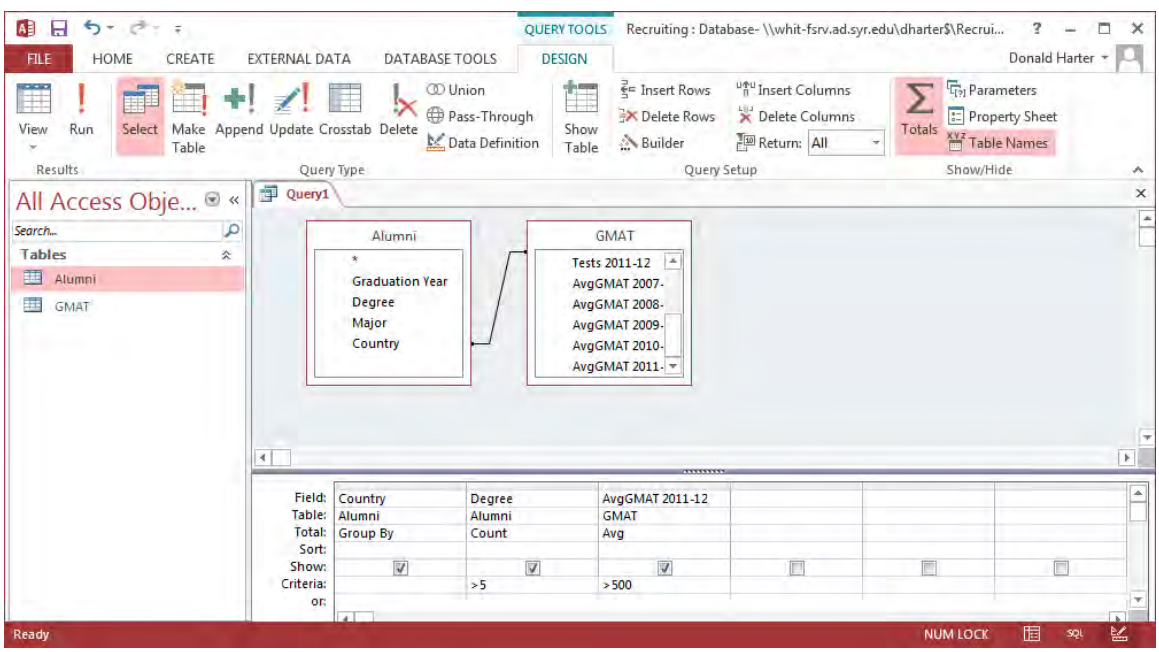



Now click Run !.

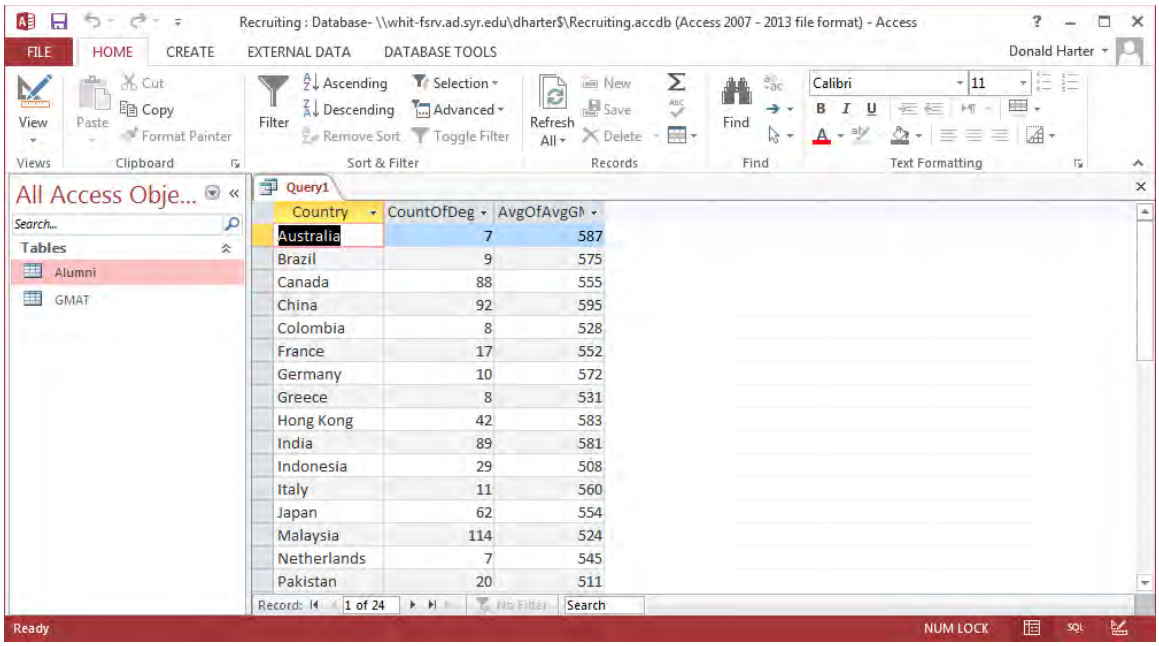



Go back to the Design and use the Sort field to change how data is sorted.

The criteria field allows you to filter results. Find the countries that have more than five alumni and average GMAT more than 500 by using >5 and >500.



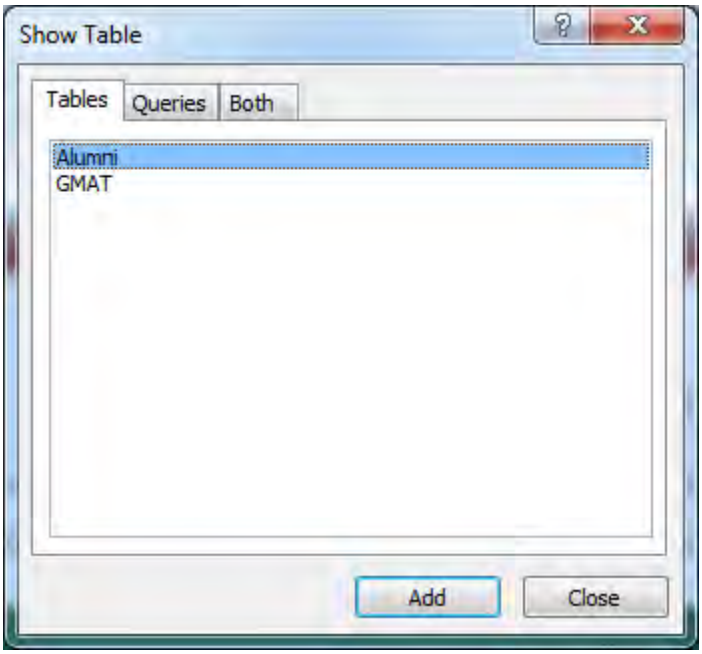

Click Run !.


Recruiting : Database- \\whit-fsrv.ad.syr.edu\dharter$\Recruiting.accdb (Access 2007 - 2013 file format) - Access
FILE
HOME
CREATE
EXTERNAL DATA
DATABASE TOOLS
Donald Harter
View
Paste
Cut
Copy
Format Painter
Filter
Ascending
Descending
Remove Sort
Selection
Advanced
Toggle Filter
Refresh All
New
Save
Delete
Find
Calibri
11
Views
Clipboard
Sort & Filter
Records
Find
Text Formatting
All Access Obje...
Search...
Tables
Alumni
GMAT
Query1
Country
CountOfDeg
AvgOfAvgG
Australia 7 587
Brazil 9 575
Canada 88 555
China 92 595
Colombia 8 528
France 17 552
Germany 10 572
Greece 8 531
Hong Kong 42 583
India 89 581
Indonesia 29 508
Italy 11 560
Japan 62 554
Malaysia 114 524
Netherlands 7 545
Pakistan 20 511
Record: 1 of 24
No Filter
Search
Ready
NUM LOCK

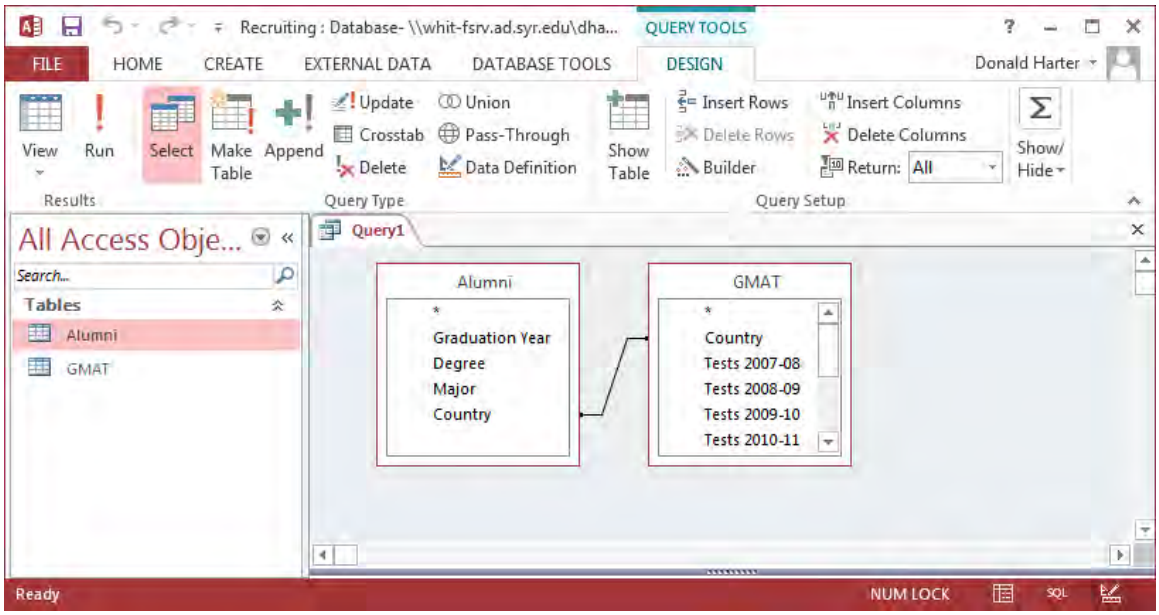
**Session 4.7: Queries & Fixing Dirty Data**

When you want to retrieve specific data from a database, you can use a query. A query is like a filter in Excel, but much more powerful. It can allow you to restrict what data is displayed, sort the data, and perform calculations. However, dirty data can corrupt your queries. Let's see how.

To create a query, click on CREATE, then Query Design.

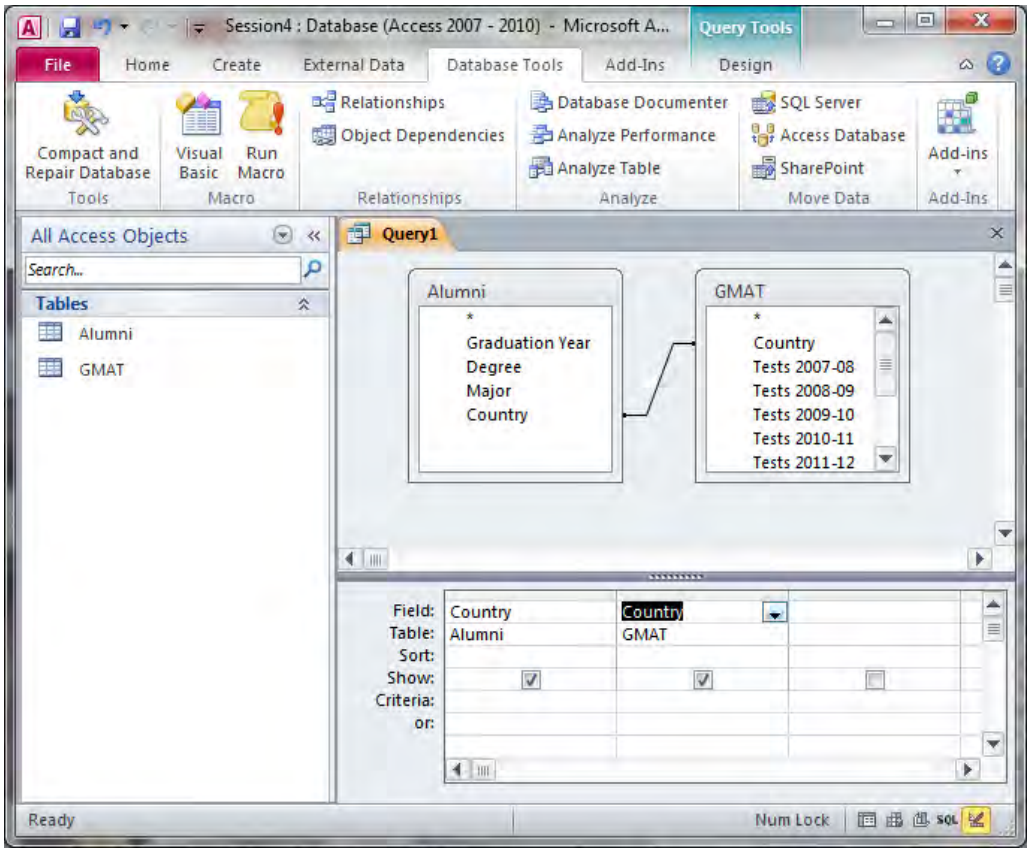



Click on Alumni, then Add; GMAT, then Add. Click Close.

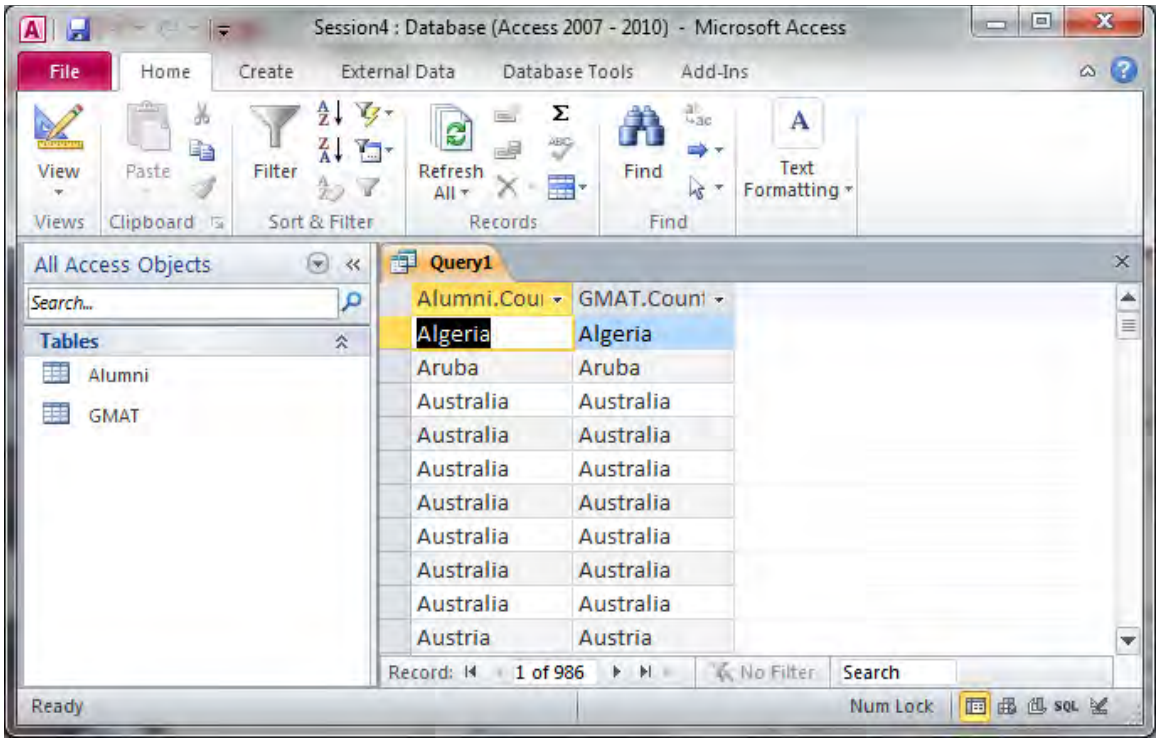



Let's compare Countries in the Alumni table to Countries in the GMAT table.

1. In the table Alumni, double click Country.
2. In the table GMAT, double click Country.
3. Click Run !

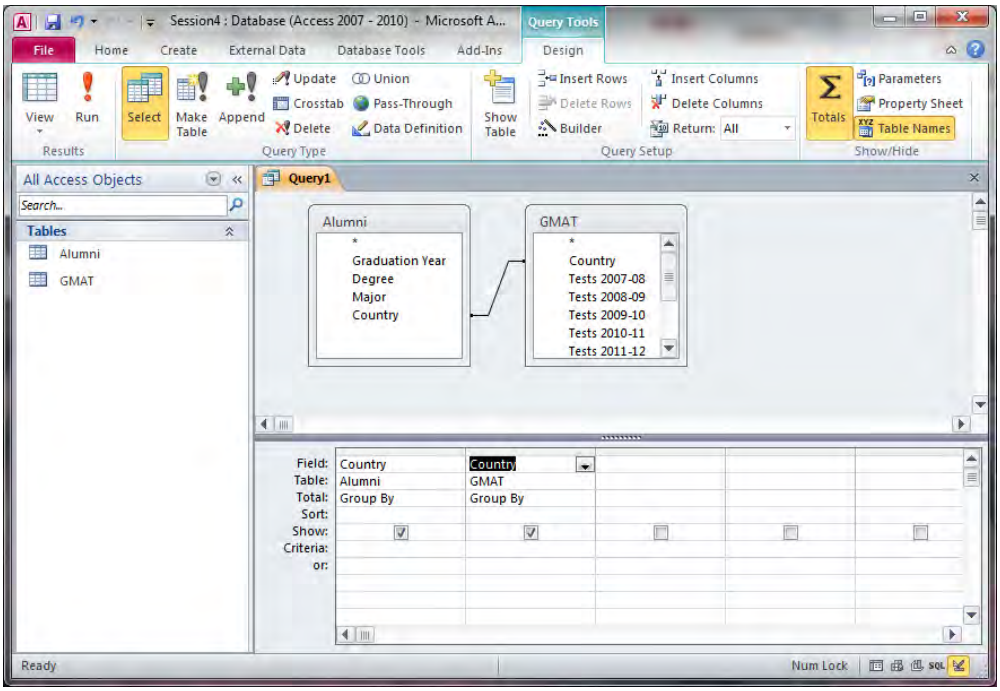



Now click Run ! in the upper left corner.

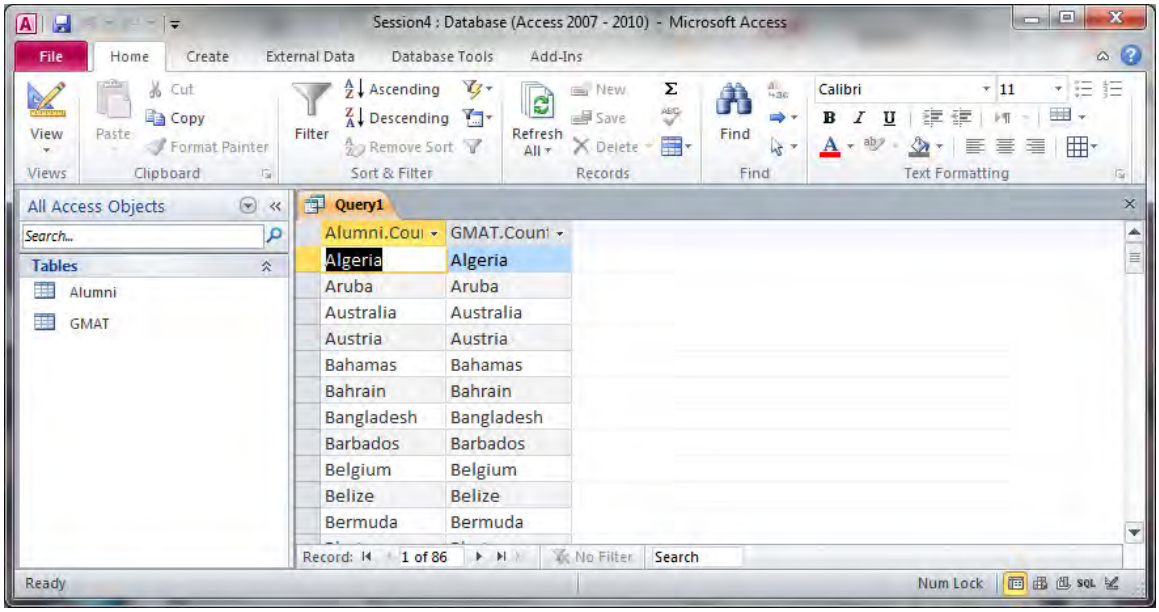



Notice that this lists duplicates of each country. To eliminate duplicates:

1. Click on View, Design View.
2. Click on the summation sign (Sigma) for Totals.
3. A Group By appears.
4. Click Run !



Now duplicates are eliminated.




Look for Taiwan in the data. Does it appear?

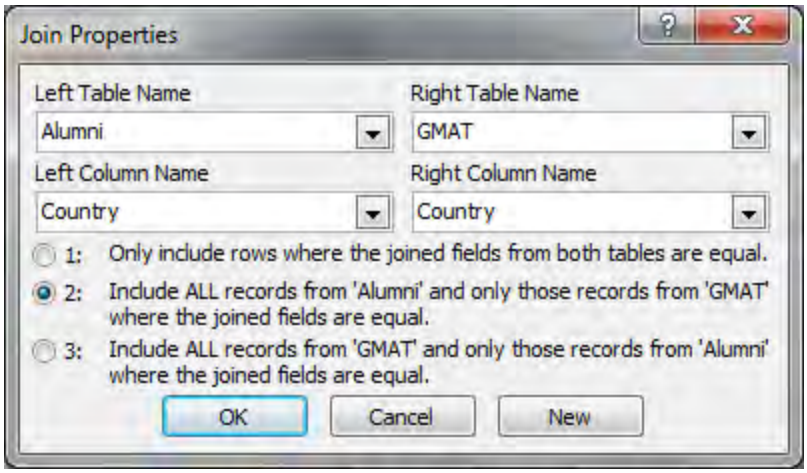
Open the table for Alumni. Does Taiwan appear?

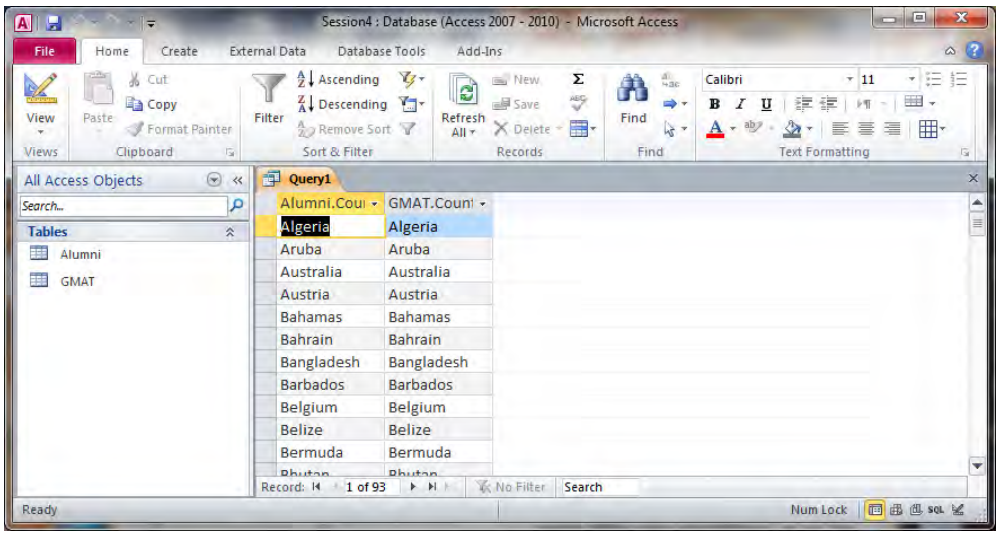
Open the table for GMAT. Does Taiwan appear?

What's the difference?

Now, let's allow Access to identify data discrepancies.

1. In your query, click on View, Design View.
2. Right click on the line connecting the tables.
3. Click on Join Properties
4. Change the selection to 2: Include All records from "Alumni" and only those records from "GMAT" where the joined fields are equal, then click OK.
5. Click Run !
6. Are there any entries in column 1 (Alumni.Country) that have blank entries in column 2 (GMAT.Country)? Why?

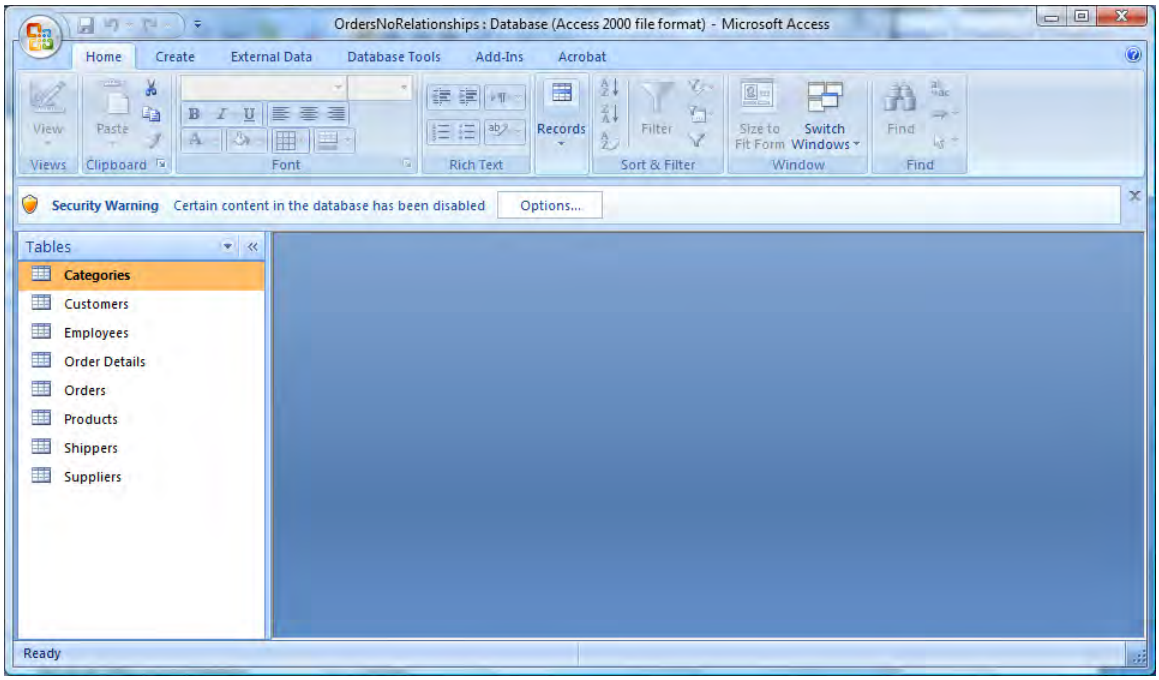


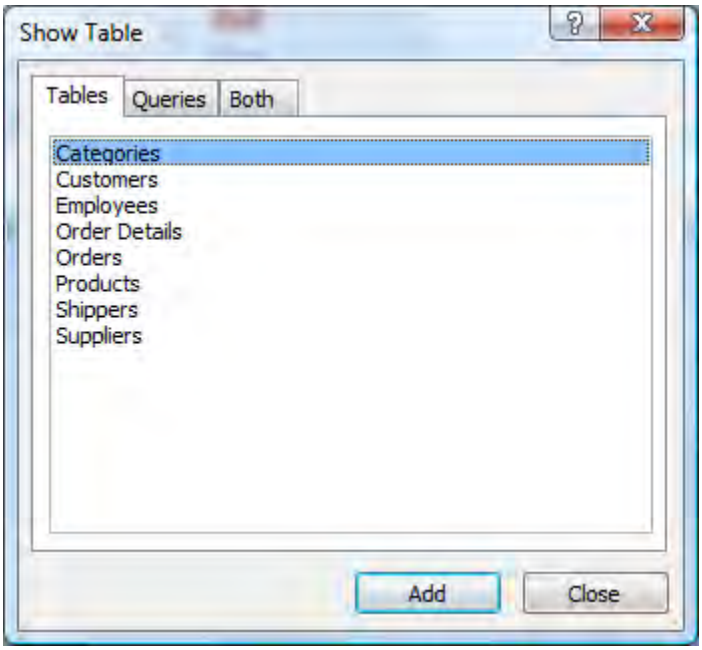

Change the relationship to selection 3. What happens?

**Session 4.8: Complex Queries**

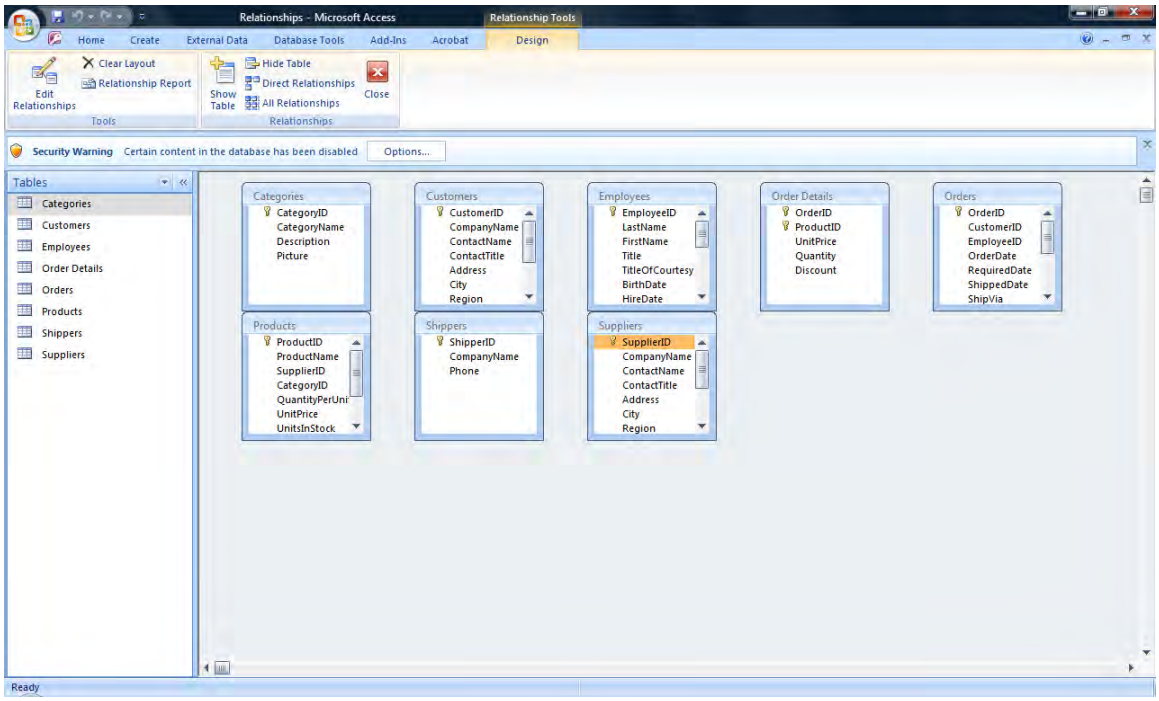
Open the database titled Business Analytics – Week 4 Orders NoRelationships Database. We are going to create relationships.



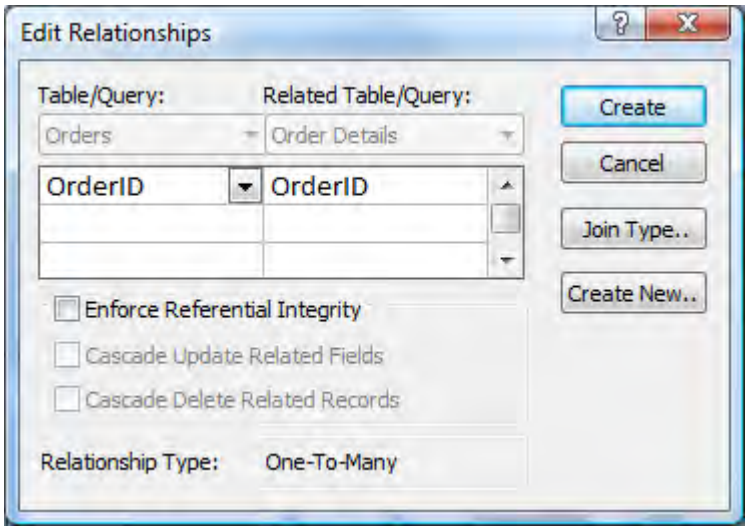

1. Click on Database Tools, Show Relationships, Show Table.



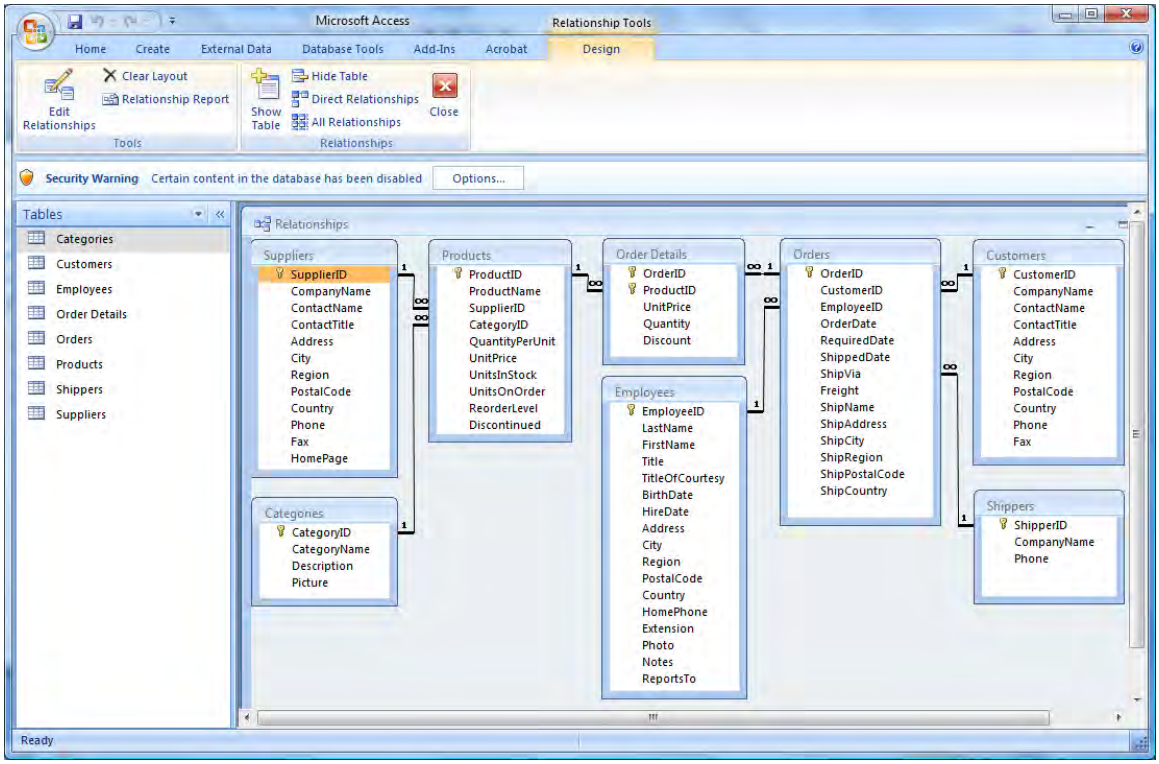

2. Click on each table, then add. When you have added all tables, click close.




3. The key fields are identified with a picture of a key. To create relationships, drag a field to the similarly named field in another table. Drag Order ID in Orders to Order ID in Order Details.



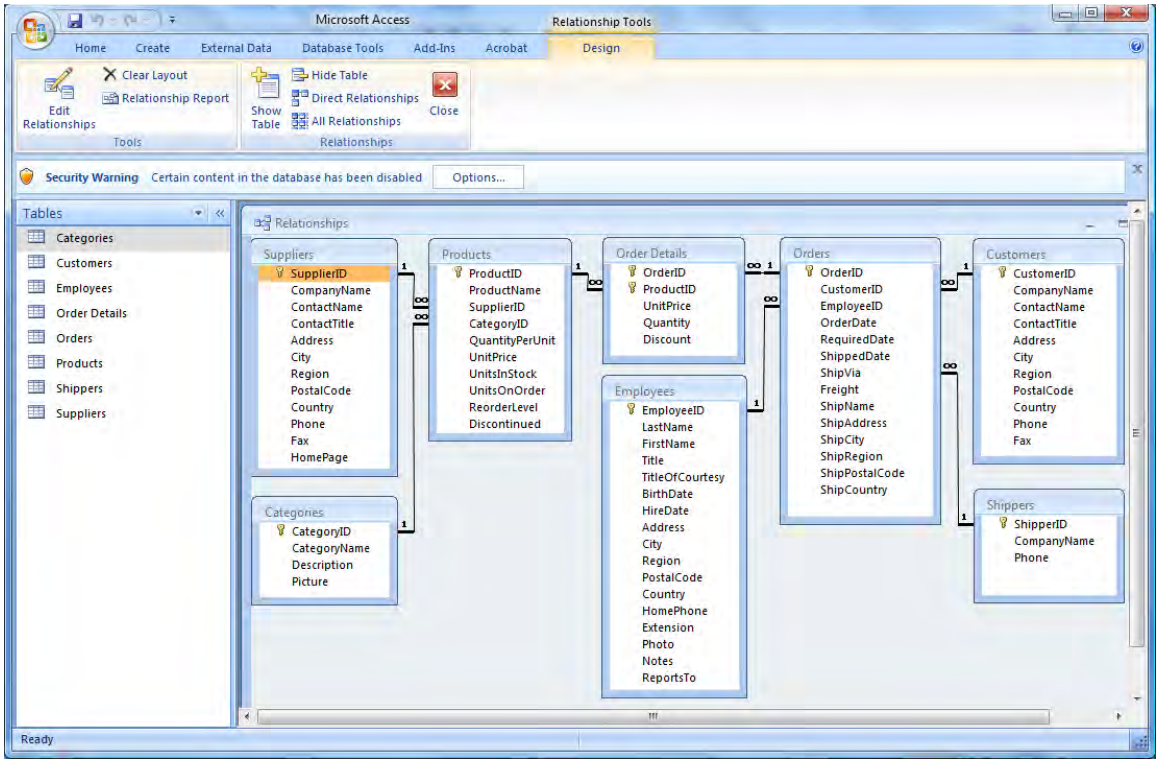

4. You should see the screen above. If you want to ensure that data in one table matches data in the other (called referential integrity), check the box. Notice that Access has determined that this relationship will be one-to-many. Click Create.
5. Similarly link the other tables. When done, you should have a picture that looks like this. Notice that I've rearranged the tables to clean up the lines.
6. Close the relationship box and save the new relationships.


Microsoft Access
Relationship Tools
Home
Create
External Data
Database Tools
Add-Ins
Acrobat
Design
Edit Relationships
Clear Layout
Relationship Report
Tools
Show Table
Hide Table
Direct Relationships
All Relationships
Close
Relationships
Security Warning
Certain content in the database has been disabled
Options...
Tables
Categories
Customers
Employees
Order Details
Orders
Products
Shippers
Suppliers
Relationships
Suppliers
SupplierID
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
Country
Phone
Fax
HomePage
Products
ProductID
ProductName
SupplierID
CategoryID
QuantityPerUnit
UnitPrice
UnitsInStock
UnitsOnOrder
ReorderLevel
Discontinued
Order Details
OrderID
ProductID
UnitPrice
Quantity
Discount
Orders
OrderID
CustomerID
EmployeeID
OrderDate
RequiredDate
ShippedDate
ShipVia
Freight
ShipName
ShipAddress
ShipCity
ShipRegion
ShipPostalCode
ShipCountry
Customers
CustomerID
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
Country
Phone
Fax
Categories
CategoryID
CategoryName
Description
Picture
Employees
EmployeeID
LastName
FirstName
Title
TitleOfCourtesy
BirthDate
HireDate
Address
City
Region
PostalCode
Country
HomePhone
Extension
Photo
Notes
ReportsTo
Shippers
ShipperID
CompanyName
Phone
Ready

## Queries

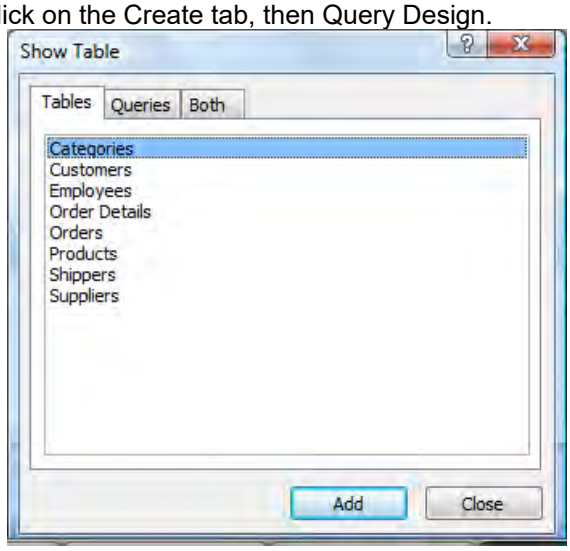
When you want to retrieve specific data from a database, you can use a query. A query is like a filter in Excel, but much more powerful. It can allow you to restrict what data is displayed, sort the data, and perform calculations.

For this section, use the Business Analytics – Week 4 Orders Relationships Database. Let's recall how the tables relate by clicking on the Database Tools tab, then Relationships.

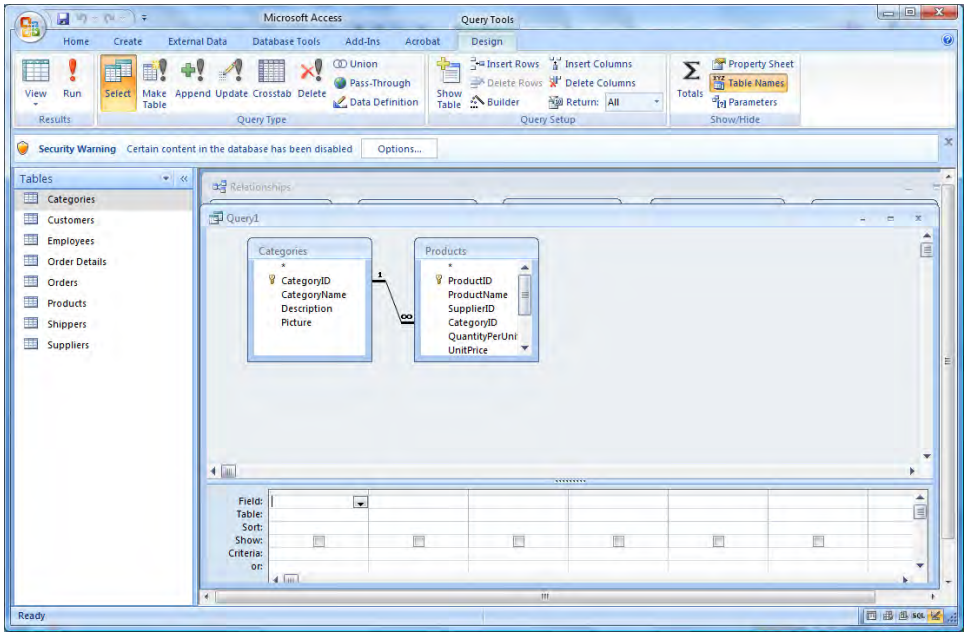



Let's find all ProductNames that have a CategoryName of beverages. Looking at the tables in the relationship diagram, can you find ProductNames and CategoryName? What tables are they in?

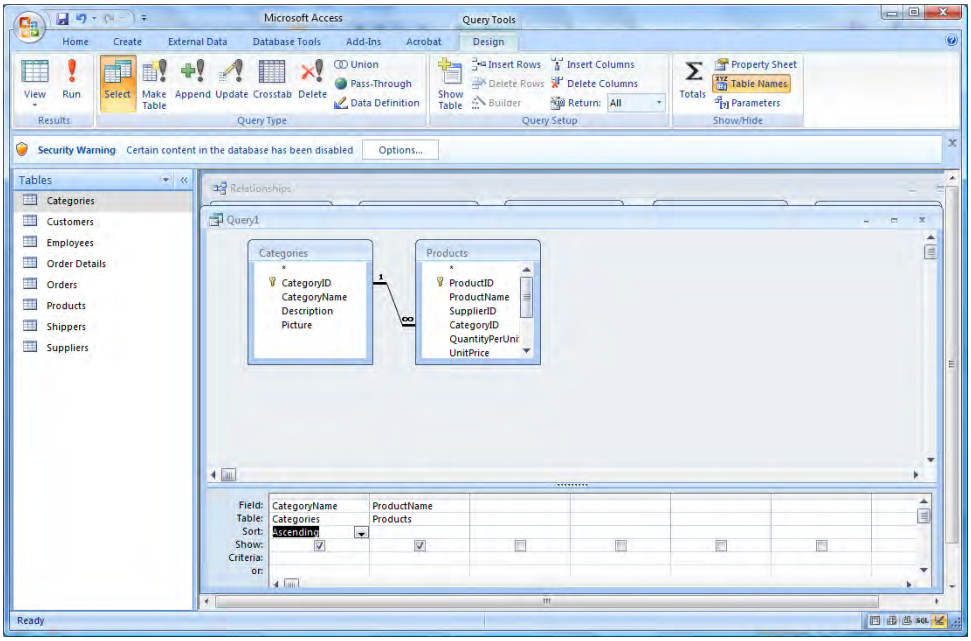
1. To create a query, click on the Create tab, then Query Design.



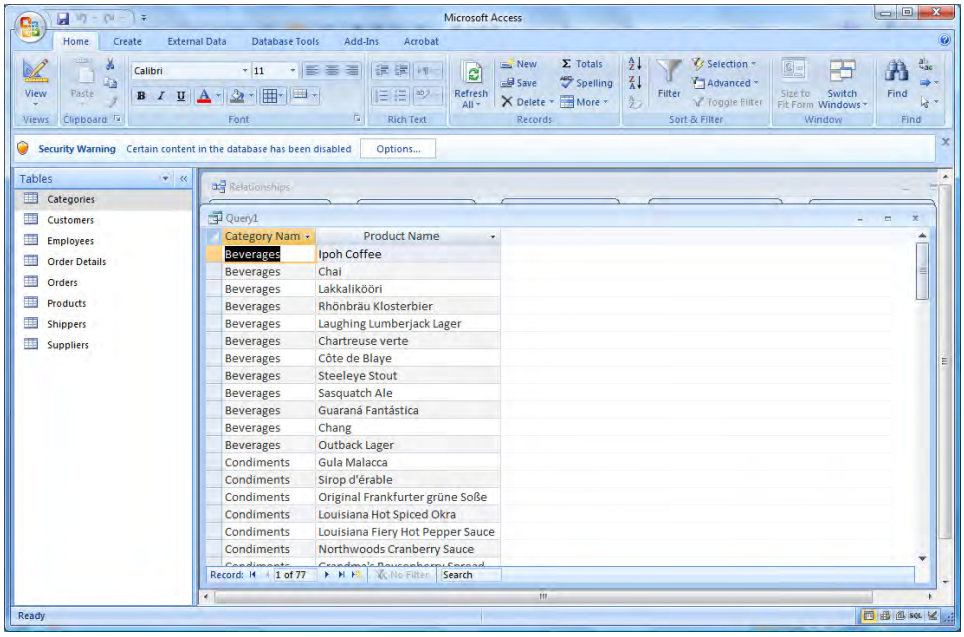
2. Since we want to use the tables Products and Categories, click on each one, then Add, then Close




3. At the top of the screen are the two tables. At the bottom is a set of options to set up the query.
4. We want to find all the Product Names for each Category Name. First, double click on Category Name. It should appear at the bottom of the page
5. Next, double click on Product Name. It should appear next to Category Name.
6. If you want to sort the data, click on sort under Category Name, then ascending.



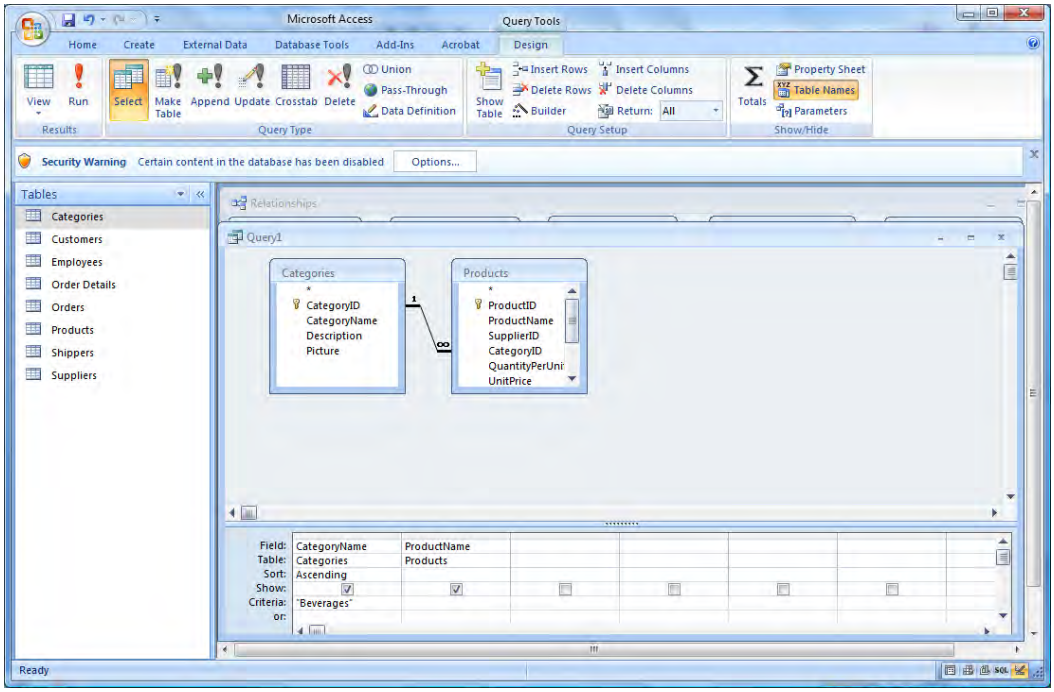

7. To run the query, click Run (see red exclamation point). You should see the following result.



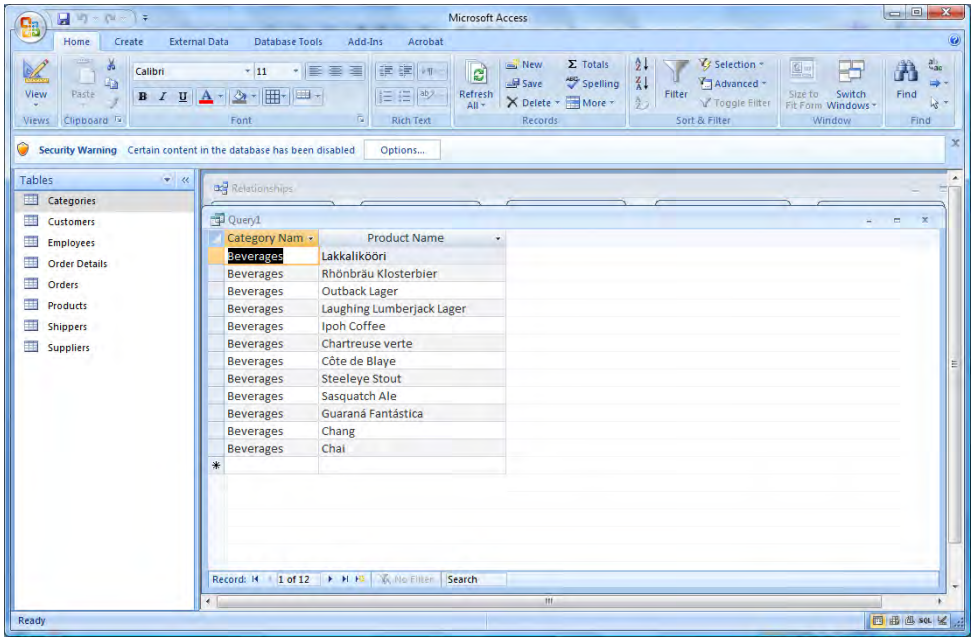

## Filtering or Using Criteria to Restrict the Output

Just as you filtered data in Excel, you can filter data in Access. If we only want Beverages listed, use the following steps.

1. On the Home tab, click on View, Design View. This takes you back to the design of the query.
2. In the field called Criteria under Category name, enter the word Beverages. Notice that it puts quotation marks around it automatically.



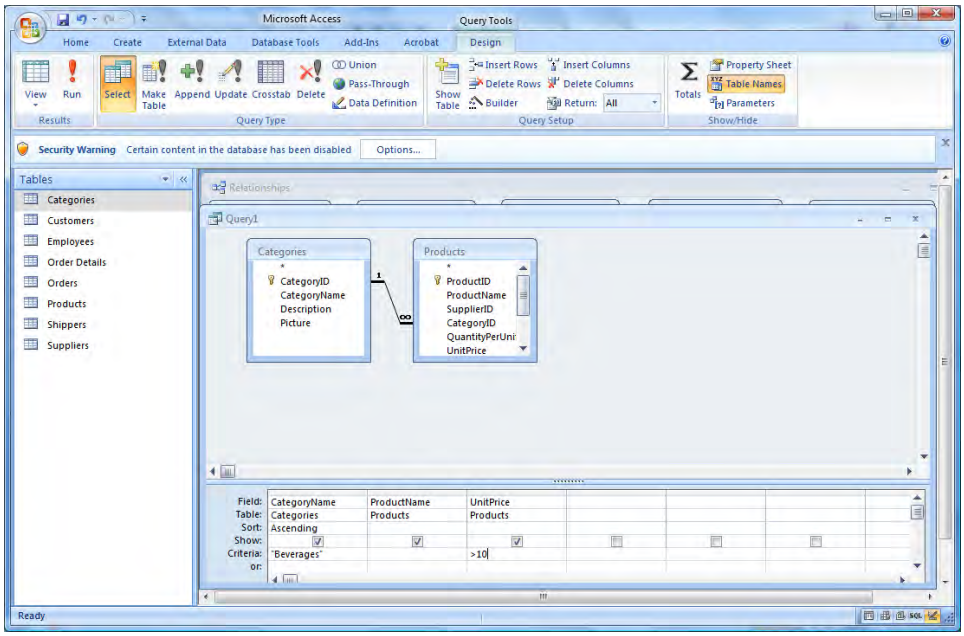
3. Run the query. Only Beverages are displayed.



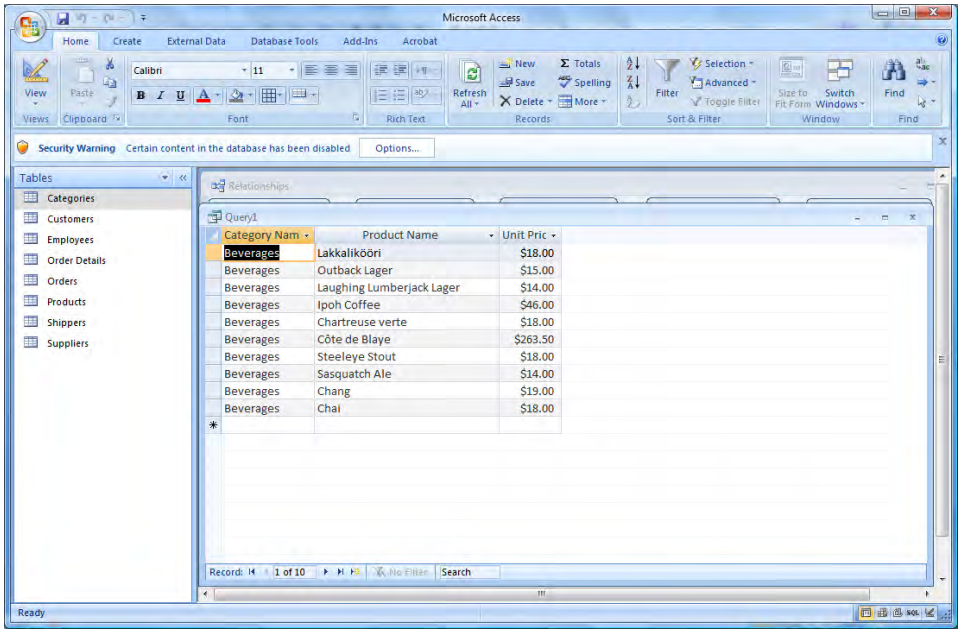

## Queries with Two or More Criteria Using "And"

You can also run queries where multiple conditions must be true. Let's find beverages with Unit Price greater than 10.

1. Click on the Home tab, View, Design View.
2. Since we don't have Unit Price in the query screen, add it by double clicking on Unit Price in Products.
3. In the criteria for Unit Price, enter >10

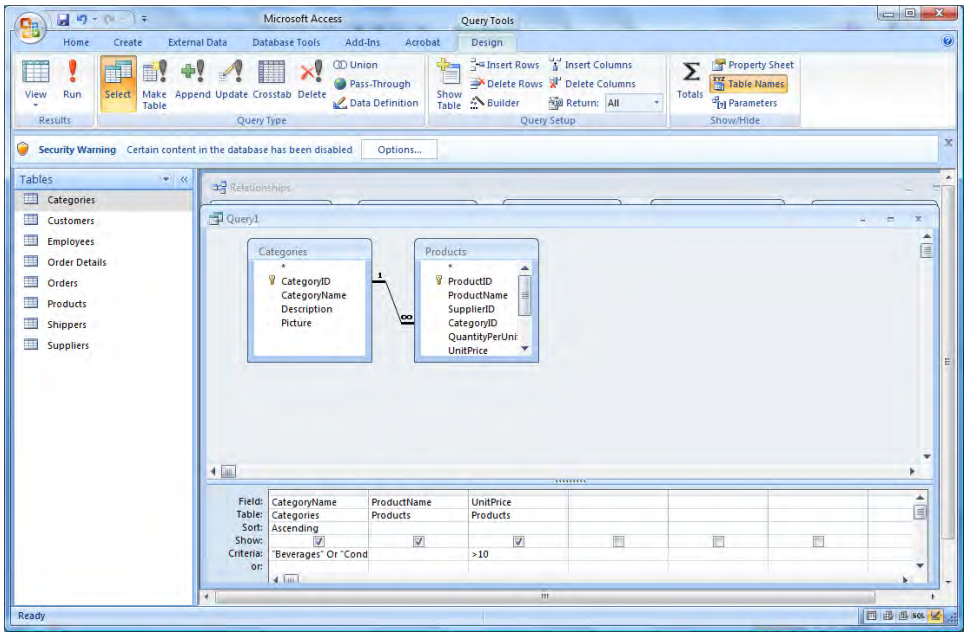


4. Run the query

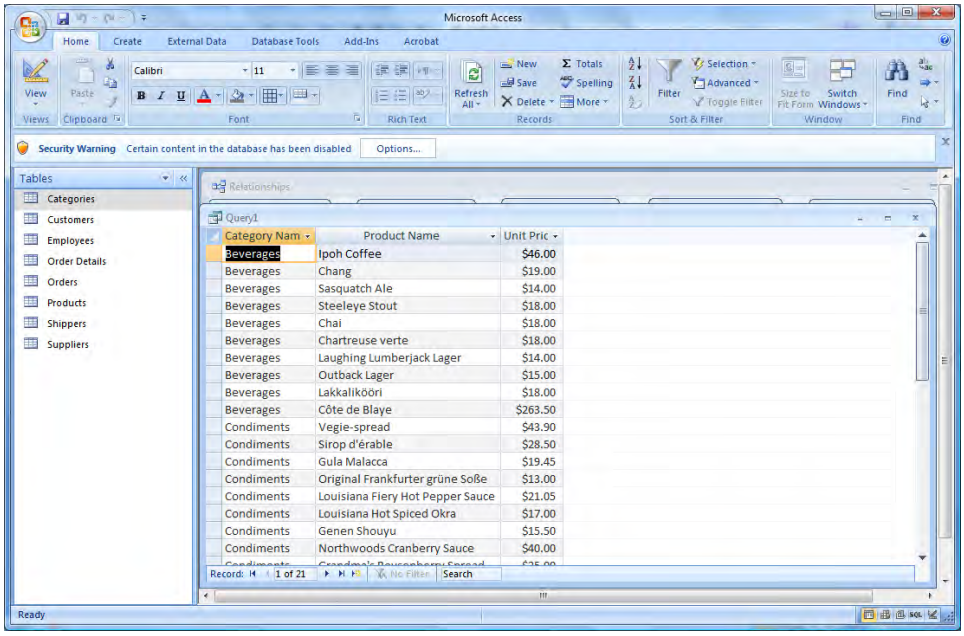



## Query with an "Or" Condition – within one field

You can also have an OR condition. For example, if we want Beverages or Condiments, we can put "Beverages" OR "Condiments" in the criteria field for Category Name.

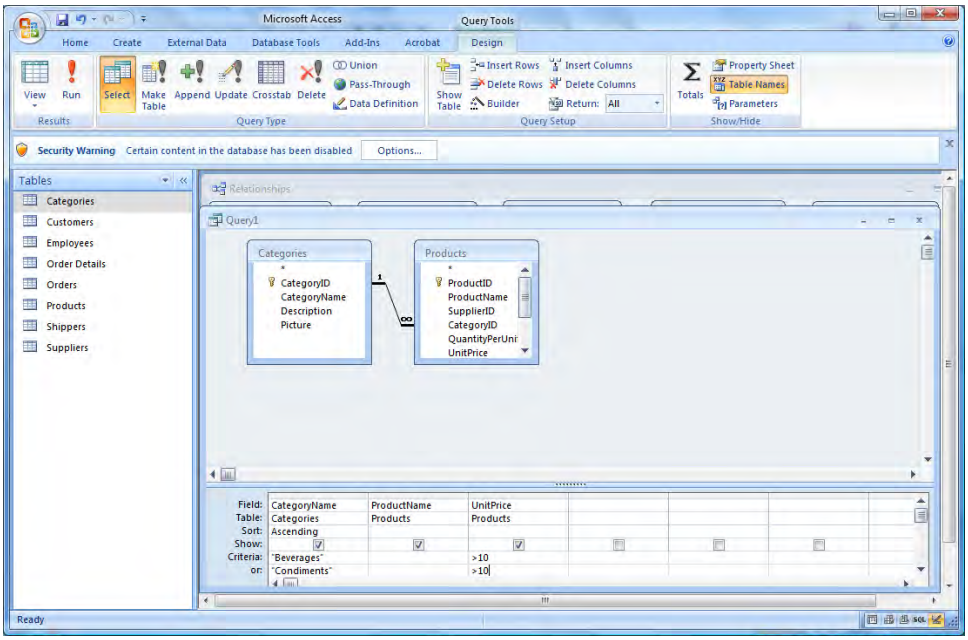


The result of running the query is shown below.




## Query with an "Or" Condition – between fields

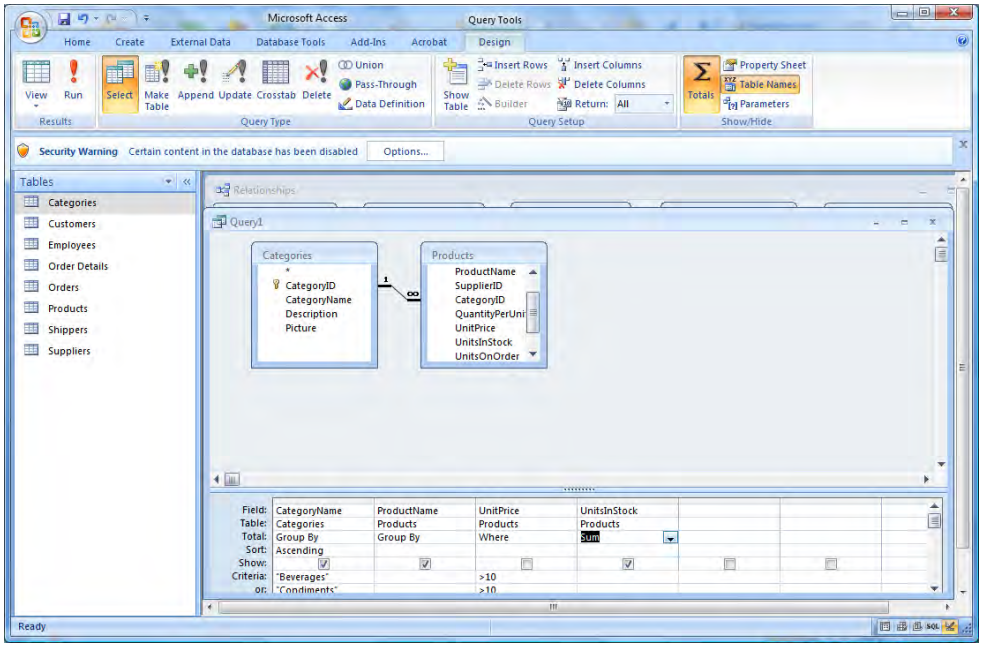
You can also use multiple criteria rows for an OR condition instead of cramming everything into one cell. Let's use multiple lines for the previous query. Note that if we also want >10, that >10 has to be on both lines.



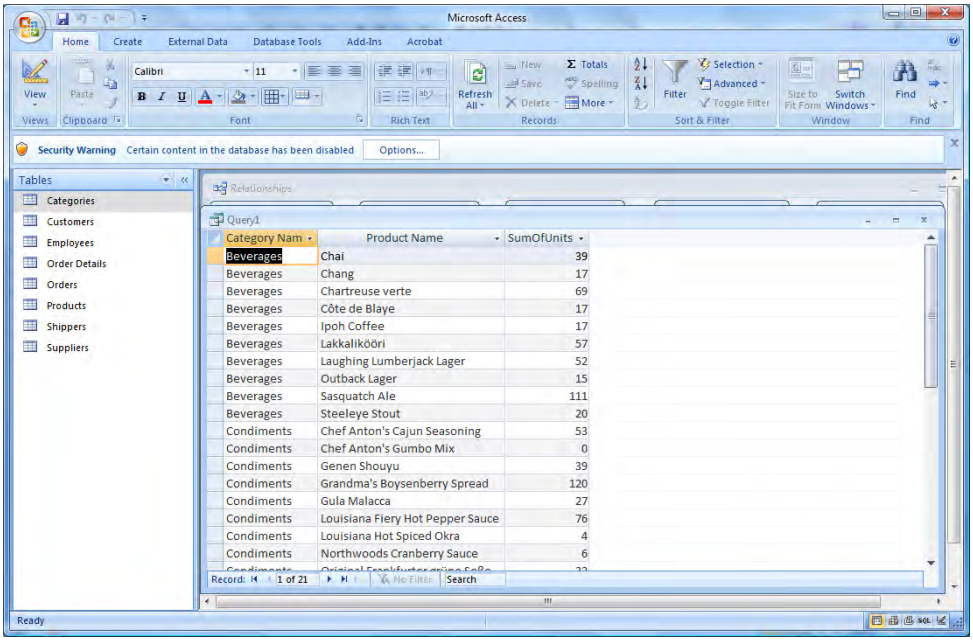

## Calculations

The calculation of sums, averages, counts, minimum, maximum can also be performed in a query. Let's sum the total number of units in stock using our current criteria.

1. First, add UnitsInStock from the Products table
2. Click on the Design tab, then click on the summation symbol ∑
3. A total line should appear, where it says group by for each variable.
4. We want to have groupings of Categories and Products, but not UnitPrice. For UnitPrice, use the down arrow on GroupBy and change it to Where. A Where clause is used to filter data; you never group on a where.
5. We don't want to group on UnitsInStock, but we want to sum the total instead. Change GroupBy for UnitsInStock to Sum. You should have the following.

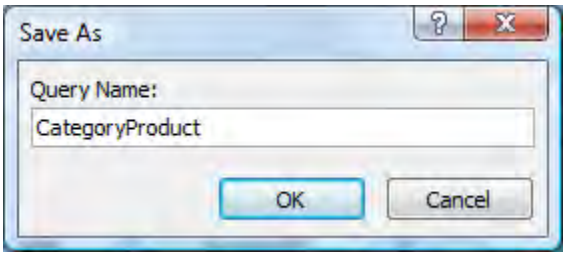



6. Now run the query.

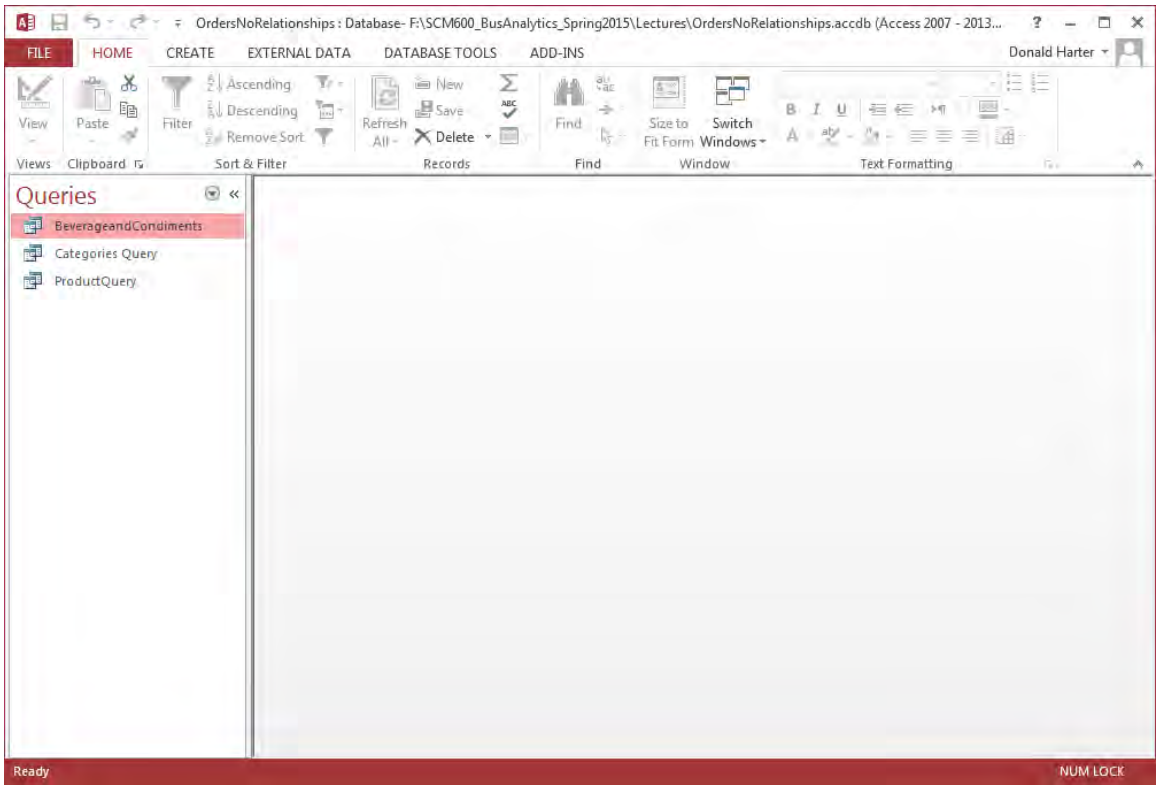


**Saving a Query**

To save a query, click on the Microsoft button in the upper left of your screen, click Save, then enter the name of your query. Click OK.




You can retrieve old queries by clicking on the down arrow next to the word Tables, then selecting Queries.



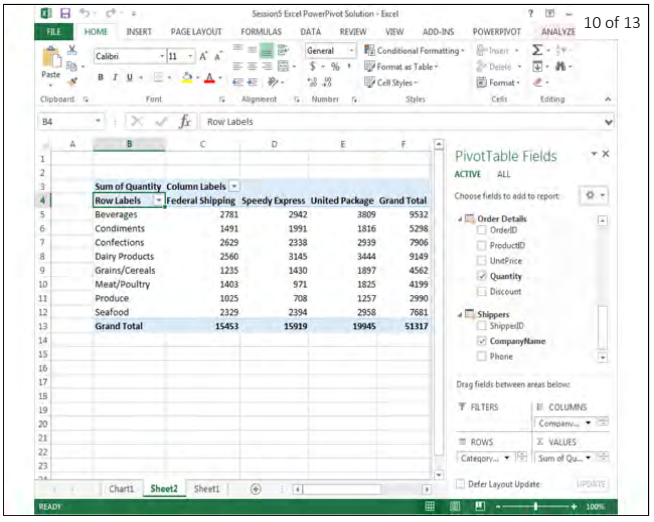
## PowerPivot

- Delivered with Excel 2013 and later, but might require installaon of the add-in
- Similar to pivot tables and charts, large amounts of data can quickly be categorized and summarized
- PowerPivot requires you to:

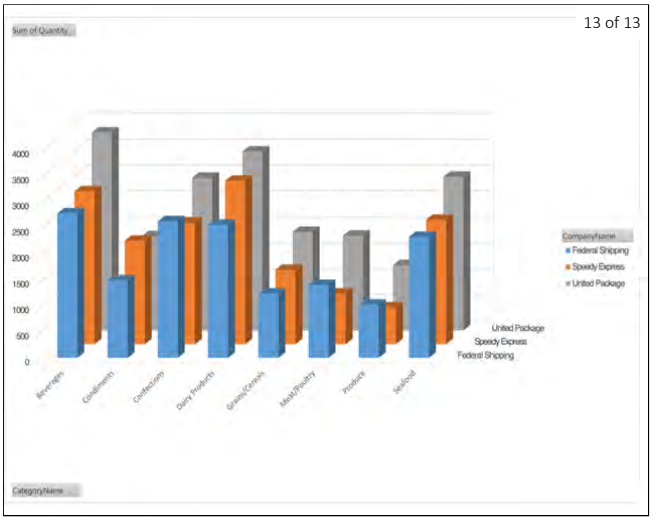
1. Connect to your data sources
2. Establish relaonship s
3. Generate tables and charts

- PowerPivot connects to the external data and has the ability to manually refresh data as the data changes







View All View Keyframes

## Importing Data to PowerPivot

- One of the most useful features of spreadsheets is the ability to quickly calculate tabular summaries of data.
- A weakness of pivot tables and charts: They rely on one worksheet.
- PowerPivot can pull data from mulple daàtabase tables or spreadsheets, or retrieve through database queries.
- Similar to databases, PowerPivot requires that the user join together data files.

5 of 19

## Getting Data into PowerPivot

- Step 1: Connecng t  o data from other files or systems
  - Database tables and SQL servers
  - Excel spreadsheets
  - Reports from other systems
  - Text files

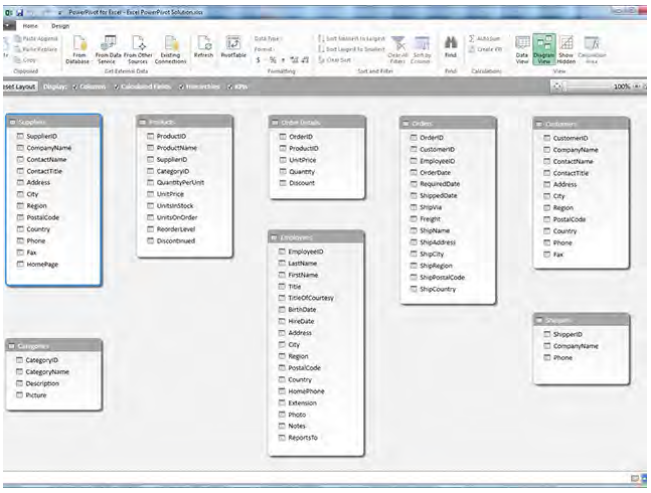
11 of 19

View All View Keyframes

## Relationships

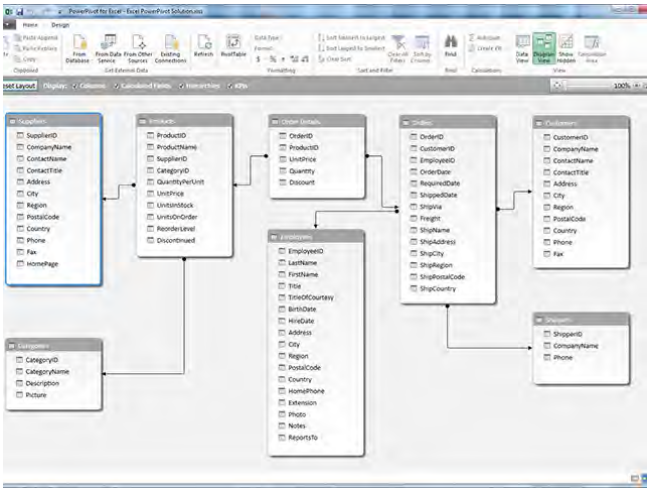
- Since PowerPivot draws from diverse data sources, it's important to integrate the data.
- PowerPivot establishes relaonship s between and among spreadsheets, database tables, queries, etc.
  - Databases only create relaonship s between and among tables.
- Similar process—find the field(s) in common between two data sources.

5 of 7



6 of 7

Placeholder for a. er



7 of 7

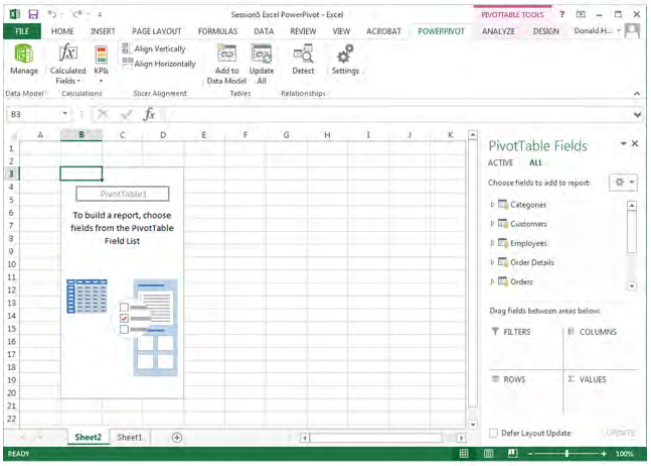
View All View Keyframes

## PowerPivot Properties and Filters

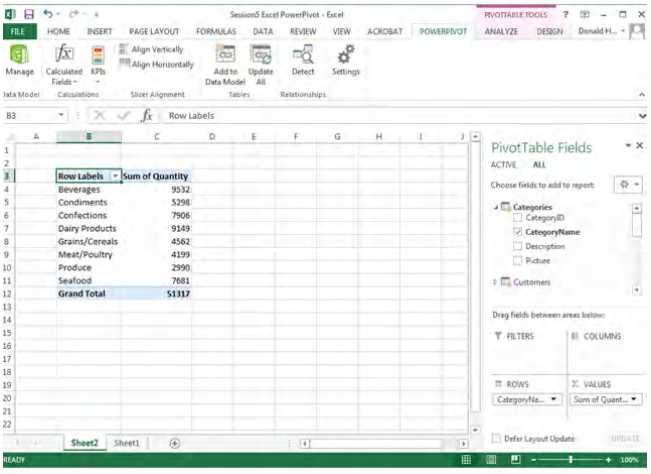
- **Properties:** Used to maintain connecons t o remote systems
  - How it connects to other data sources
  - Last me tha t the data was refreshed
- **Filters:** Used to select and summarize data meeng specific crit eria

5 of 10

View All View Keyframes




4 of 16




9 of 16

## PowerPivot Tables

- Created once data has been imported and relaonship s established
- Pivot vs. PowerPivot? PowerPivot:
  - Allows data to be retrieved from mulple sour ces
  - Enables ability to refresh data
  - Permits data to be organized with PivotTable

16 of 16

View All View Keyframes

## Slicer

- Menu-based filters
- Allows you to drill deeper into the data by selecng v arious “slices”

8 of 8

View All View Keyframes

## Timeline

- Special type of slicer that uliz es dates
- Generates a meline menu—clicking on specific da tes in the menu slices or filters the data for that date range
- Opon t o select mulple mon ths, years, or any combinaon

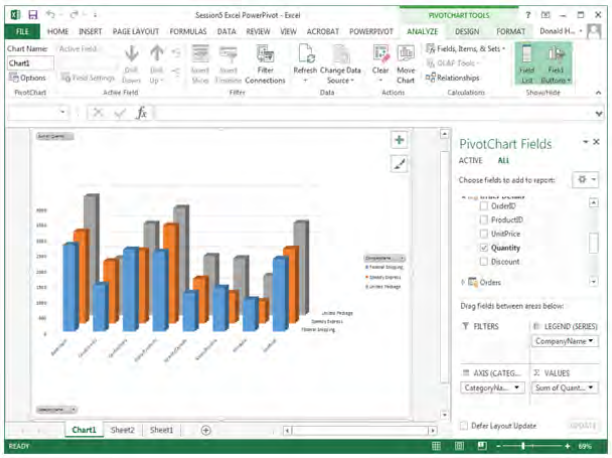
9 of 9

View All View Keyframes

## Charts and Graphs

- PowerPivot can create charts similar to PivotCharts.
- However, PowerPivot can draw data from mulple da  ta sources.

3 of 7




4 of 7

# SCM 651: Business Analytics

1

Week 5

# Agenda



- Homework #2: discussion
  - Google analytics hints
- Review of concepts
- Group discussion of articles
  - Innovating with Airborne Analytics
  - A New, Analytics-Based Era of Banking Dawns at State Street

# Homework #2



- Campaign time frames, costs, effectiveness
- Future campaign: geographic regions, key words, day of week and time of day
- Allocation of costs by program and region
- Performance measures
- Other data that would be helpful

- Example of how to find information in Google Analytics

# Week 5 - Review



- Power Pivot
  - Data can be imported from database tables, Excel spreadsheets, SQL servers, other types of files
  - Relationships can be established among database tables, excel spreadsheets and other files
- Power Pivot properties
  - Shows connection to data sources and last time data was refreshed

# Week 5 - Review



- Slicers
  - Creates a menu structure to slice the data by a specified characteristic
- Timelines
  - Similar to a slicer, but using a date timeline rather than a categorical classification
- Power Pivot charts
  - Same capability as Pivot Charts
  - Can be created using database tables, Excel spreadsheets and other files

# Article #1: Innovating with Airborne Analytics



- Innovating with Airborne Analytics
  - What are the three areas of focus for their strategy for the next 3-4 years? Give examples of what they are doing
  - What are the challenges with hiring data scientists?
  - What are the risks trying to use technology in the organization? Give an example of each

# Article #1: Innovating with Airborne Analytics



- Innovating with Airborne Analytics
  - What are the three areas of focus for their strategy for the next 3-4 years? Give examples of what they are doing
    - Operational efficiency:
      - engine performance, reliability, fuel burn
      - Optimize crew and shift deployment (match to plane types, destinations, breaks for staff)
    - Customer intimacy:
      - Frequent flyer preferences, flying preferences
      - Click stream data
      - Customer complaint data
    - Innovation:
      - Site monitoring: download Financial Times & WSJ
      - Lost-bag retrieval (LED tags & iPhone), customer follow-up

# Article #1: Innovating with Airborne Analytics



- Innovating with Airborne Analytics
  - What are the challenges with hiring data scientists?
    - Not needed all the time
    - Sometimes they blindly follow the data, not incorporating experience
  - What are the risks trying to use technology in the organization?
    - Interoperability: merging data from different sources
    - Data security: privacy
    - Taxonomies: architecture, naming conventions, context
    - Data ownership: internal versus external, and internal rivalries

# Article #2: A New, Analytics-Based Era of Banking Dawns at State Street



- A New, Analytics-Based Era of Banking Dawns at State Street
  - What were the key questions that they asked which analytics could help?
  - What was the challenge of launching analytics at State Street?
  - How do they continue to push analytics forward?

# Article #2: A New, Analytics-Based Era of Banking Dawns at State Street



- A New, Analytics-Based Era of Banking Dawns at State Street
  - What were the key questions that they asked?
    - How do we better manage, measure, visualize risk?
    - How do they deal with regulatory change?
    - How do they find new sources of return?
    - What new types of investments do they need to increase yields?
    - How do they keep costs down?

# Article #2: A New, Analytics-Based Era of Banking Dawns at State Street



- A New, Analytics-Based Era of Banking Dawns at State Street
  - What was the challenge of launching analytics at State Street? Answer: Culture
    - Remaining agile in decision making
    - Inspiring staff toward new organization
    - Be more nimble and market savvy
  - How do they continue to push analytics forward?
    - Quarterly pulse check survey of employees
    - Working list of goals every quarter

# Microsoft Excel: PowerPivot

## Session 5.4: Microsoft Excel

This document is a quick reference guide for reviewing the techniques in the laboratory.

For these exercises, download the files:
“Business Analytics – Week 5 Instructions.doc”
“Business Analytics – Week 5 Orders NoRelationships Database.acc”

## PowerPivot Add-in

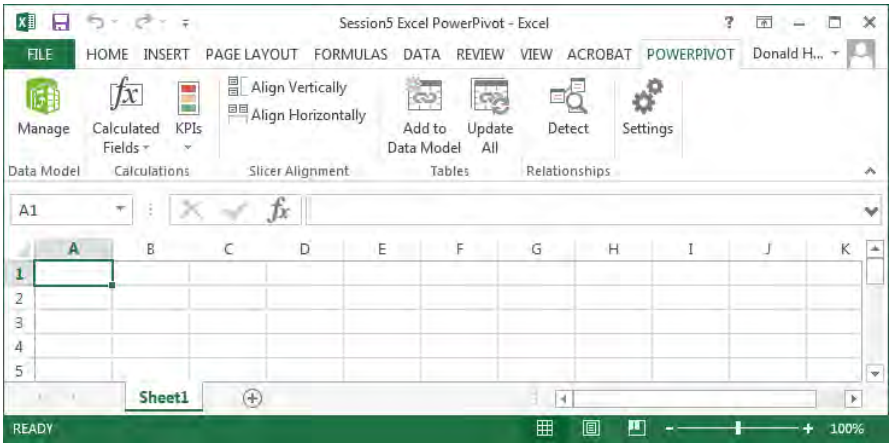
PowerPivot is an add-in to Excel. To add the capability:

1. Open Excel
2. Go to **File** > **Options** > **Add-Ins**
3. In the **Manage** box, click **COM Add-ins**, then **Go**.
4. Check the **Microsoft Office Power Pivot in Microsoft Excel 2013** box, then click **OK**

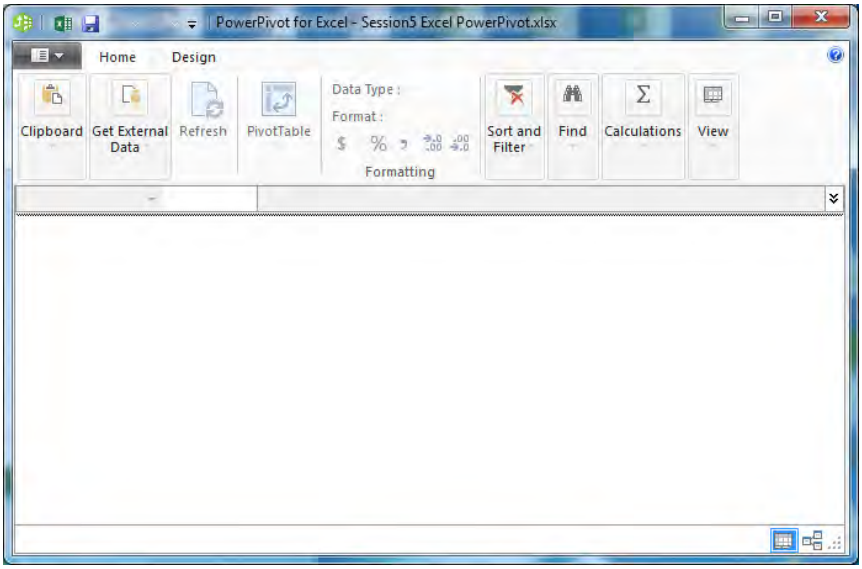
## Session 5.5: Getting Started & Importing Data

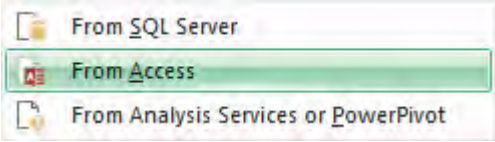
We want to launch PowerPivot and import data from other sources.

1. Click on the PowerPivot tab at the top of the screen
2. Click on Manage in the upper left corner

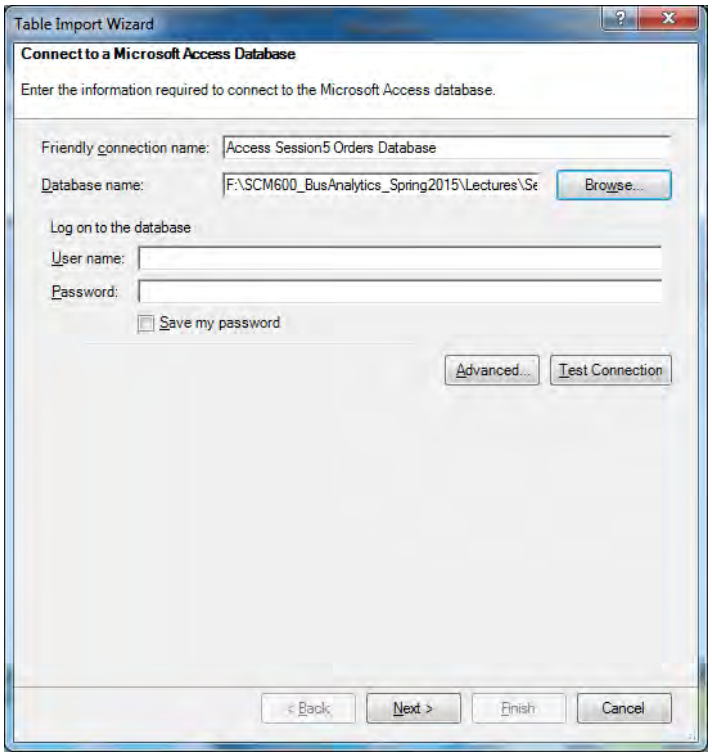


3. Click on Get External Data, then Database, then From Access

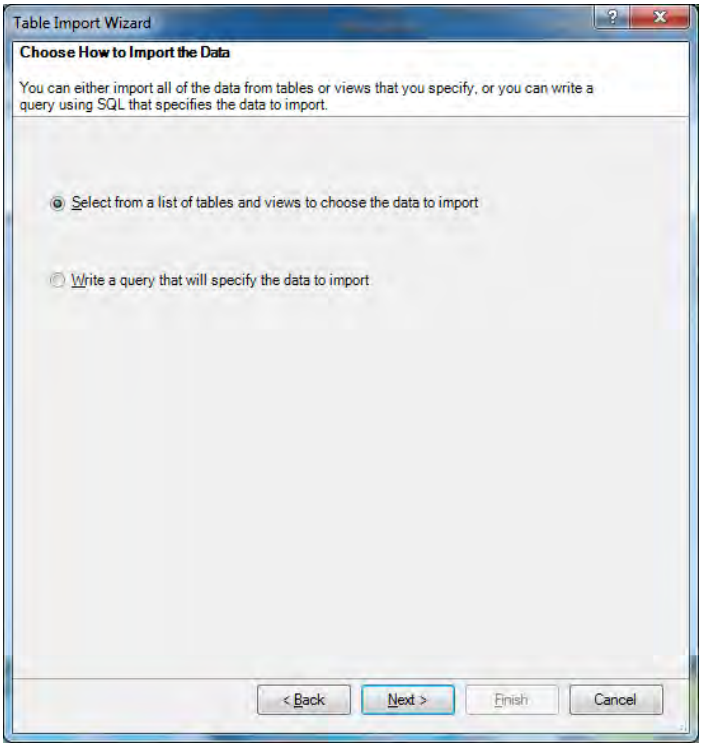


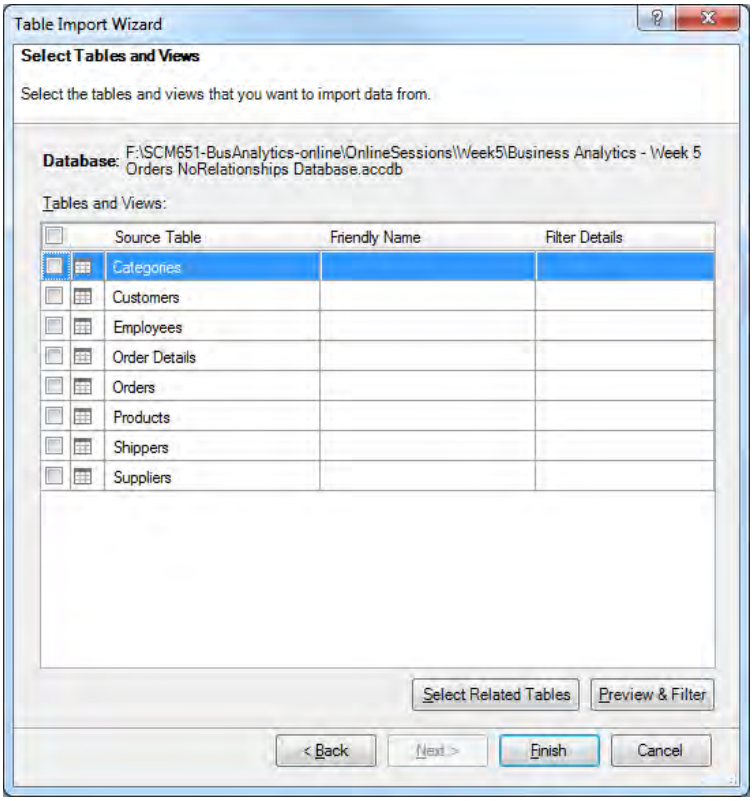

4. Click on Browse, and find the Business Analytics – Week 5 Orders NoRelationships Database downloaded from the website; click Next



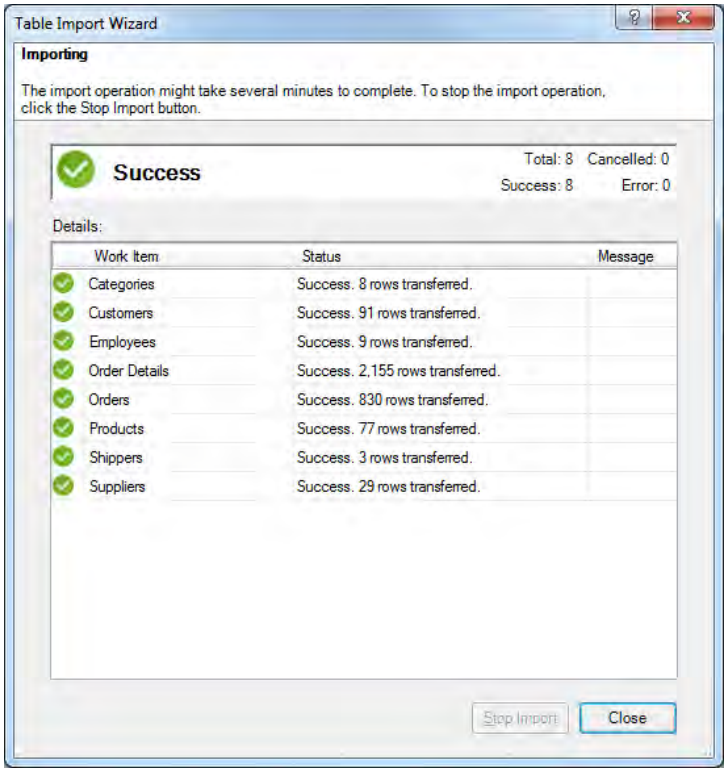
5. Select the default “Select from a list of tables”, then click Next



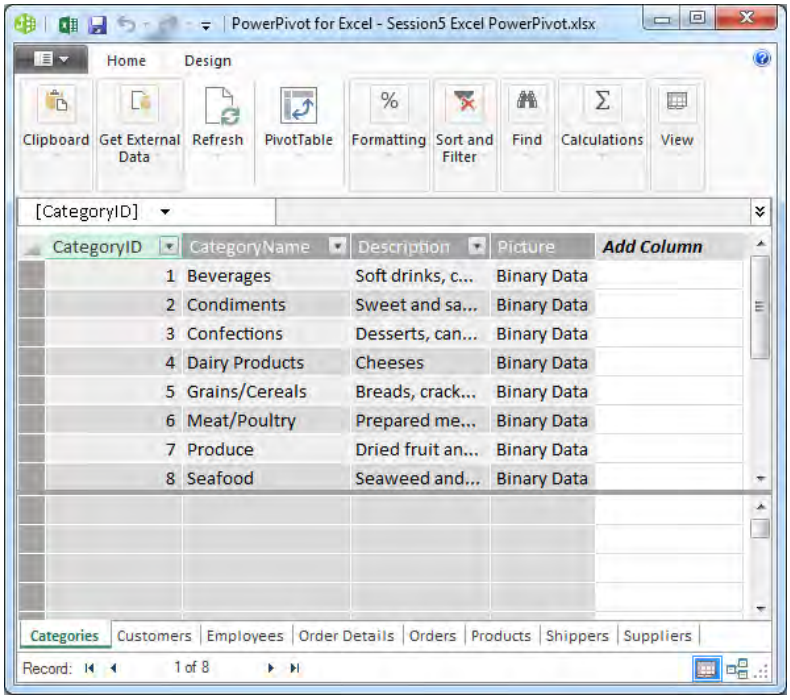

6. Check the box for tables Categories through Suppliers, then Finish



7. The completed import will list the tables and number of rows transferred; click Close.



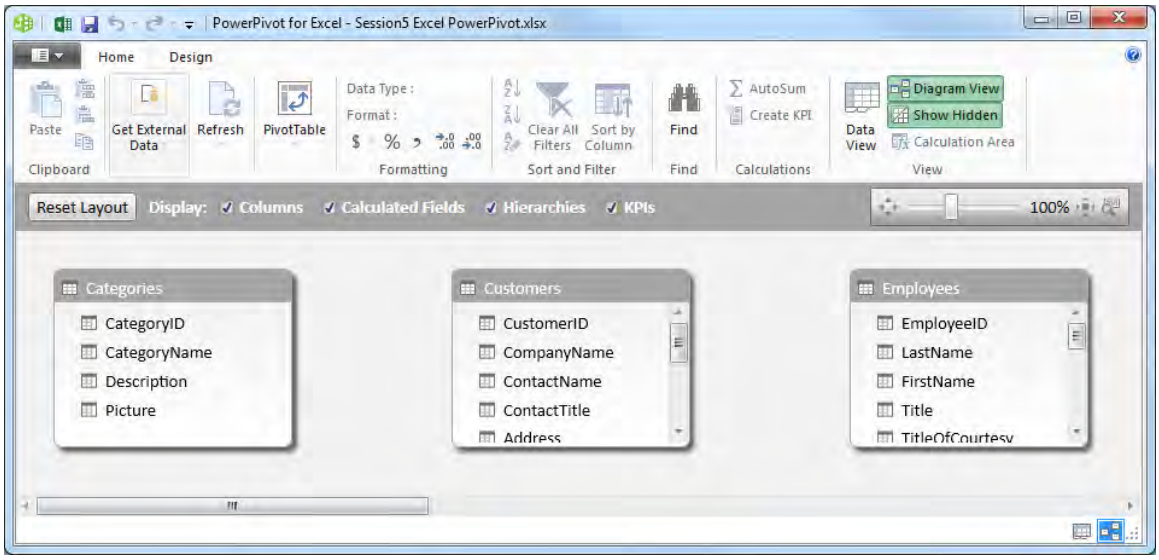

8. The data grid should appear. The tabs at the bottom include all of the tables.



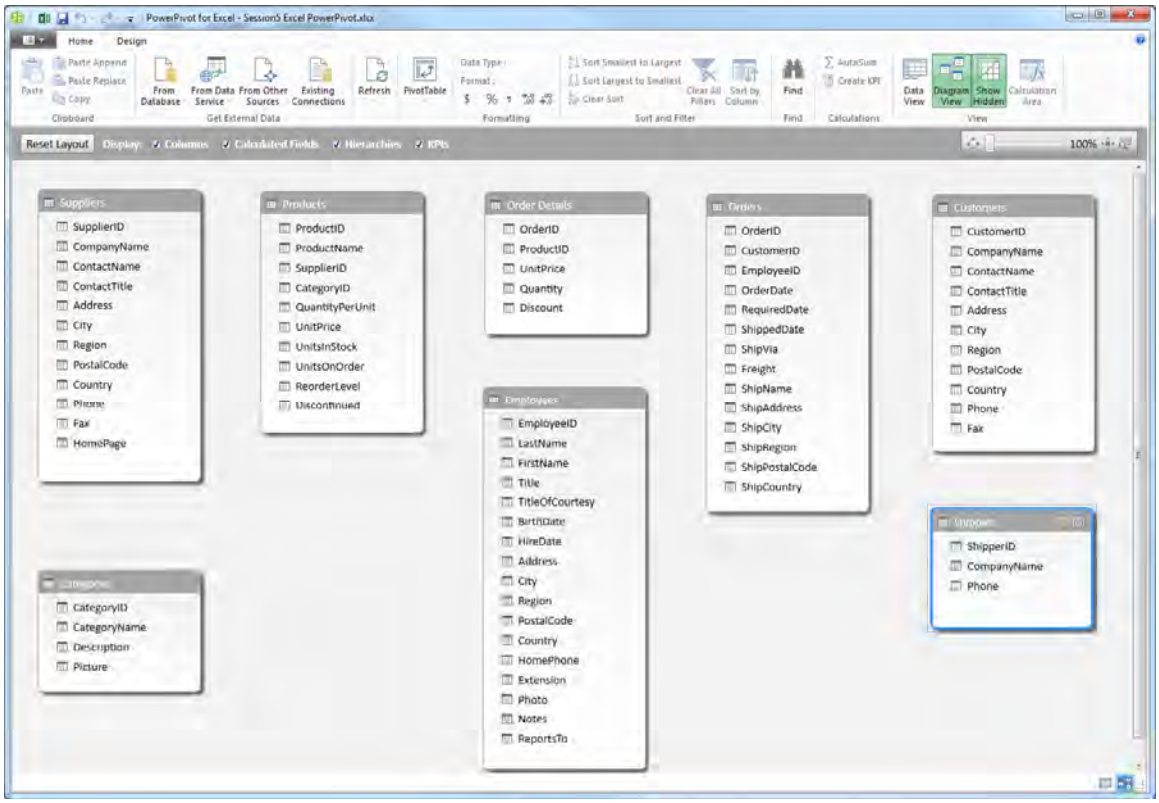
## Session 5.6: Relationships

In Microsoft Access, we were able to establish relationships between tables. We can create similar relationships in Excel PowerPivot.

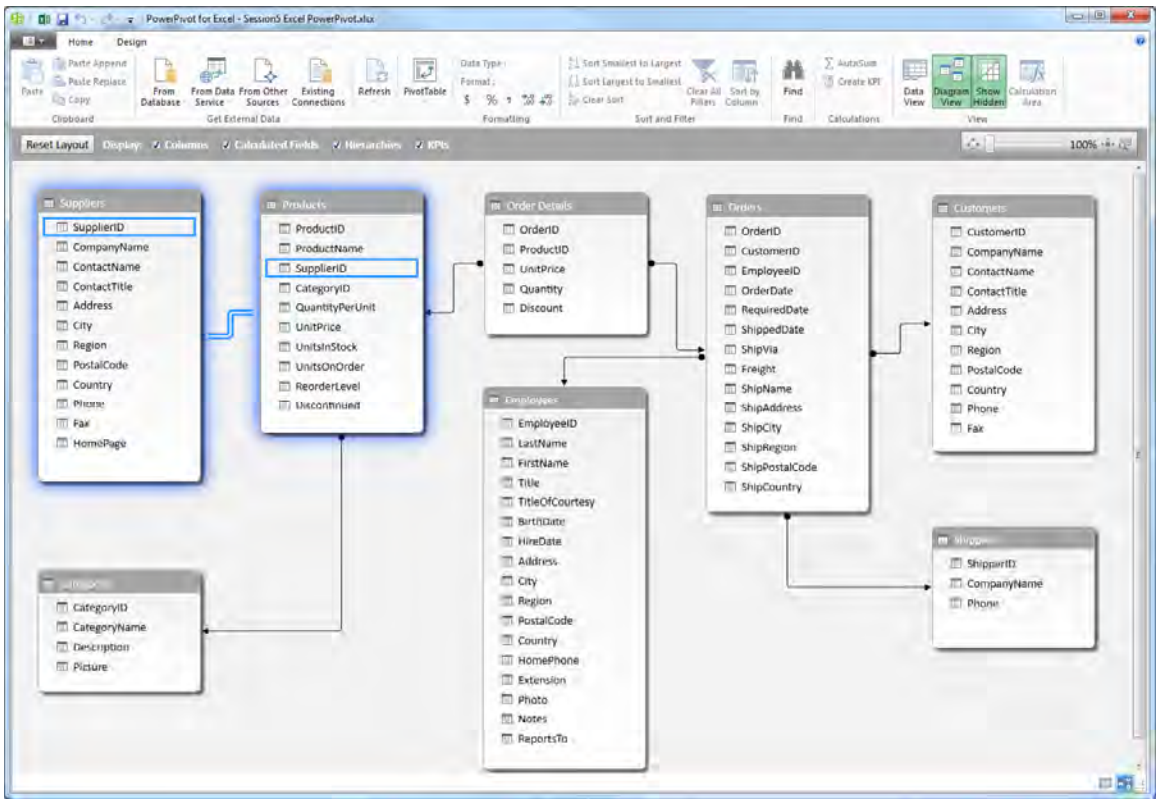
1. In the lower right corner of the Data Grid, click on the diagram view. Notice that there is a scroll bar; you can scroll to see all the tables.




2. I've reorganized the tables so it's easier to see the data



3. Now, create relationships by dragging and dropping the unique field in one table that is in common with another.
    a. Suppliers table: SupplierID with Products table: SupplierID
    b. Categories table: CategoryID with Products table: CategoryID
    c. Products table: ProductID with OrderDetails table: ProductID
    d. OrderDetails table: OrderID with Orders table: OrderID
    e. Orders table: EmployeeID with Employee table: EmployeeID
    f. Orders table: CustomerID with Customers table: CustomerID
    g. Orders table: ShipVia with Shippers table: ShipperID (Note: the fields used in establishing the relationship can have different names, as long as they have the same data)
4. When you are done, click on any relationship. It will show the two tables connected and the field names used for the relationship.

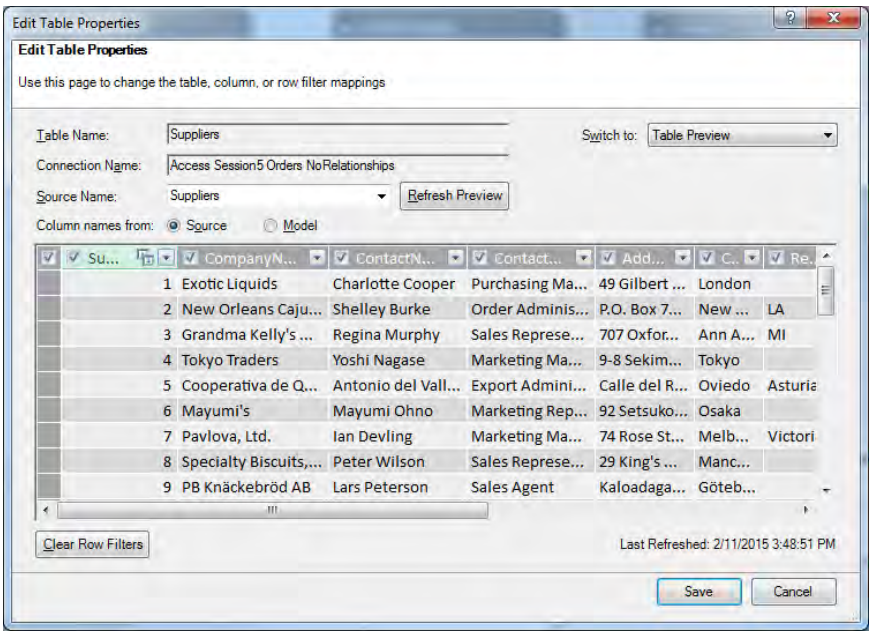



5. Save your work by clicking on the icon in the upper left, and saving.
6. In the upper right of the screen, switch back to the Data View.
7. If you close this view, you can always come back by clicking Manage.

## Session 5.7: Table Properties

Table properties presents additional information on the source of your data, data display, and filters.

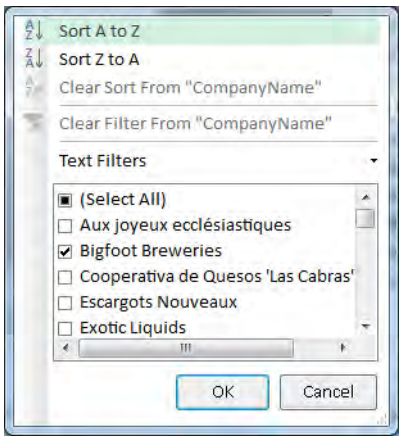
1. In PowerPivot, click on the Design tab in the upper left corner
2. Click on the title of the Suppliers table (it should highlight the table)
3. Click on Table Properties at the top of the screen
4. Notice that the Connection Name identifies where the data originated



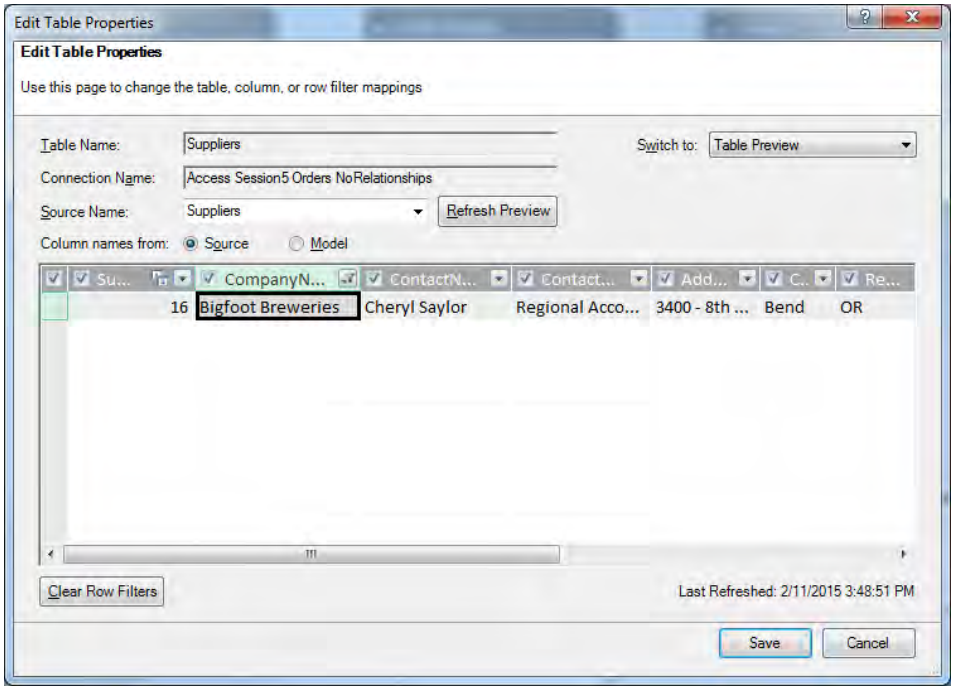

## Table Filters

Filters are also possible in this view.

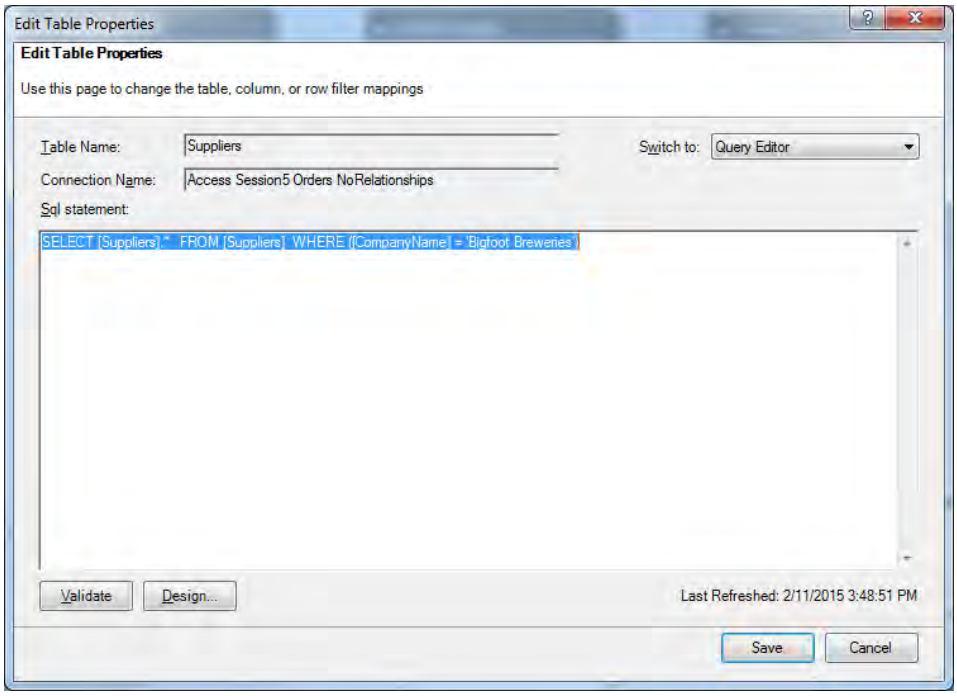
1. Click on the down arrow for Company Name
2. Click on the box before (Select All) to uncheck the box
3. Check the box for Bigfoot Breweries, then OK



You should see the following filtered data.




In the upper right of the view, click on the “Switch to:” drop down menu and change from Table Preview to Query Editor. It now displays the SQL to implement this filter.

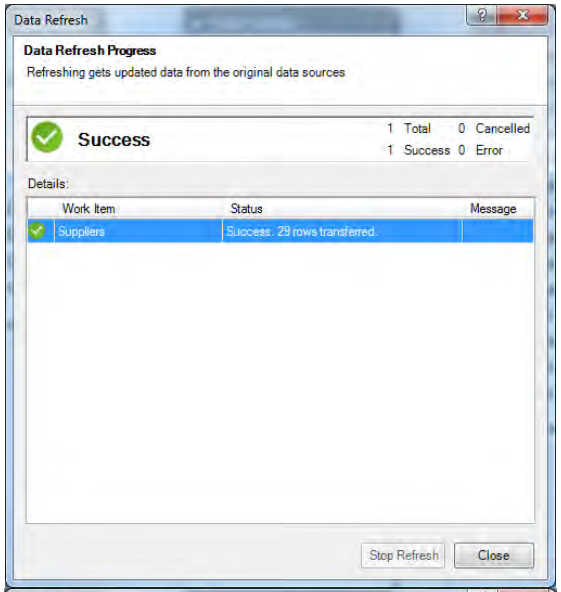


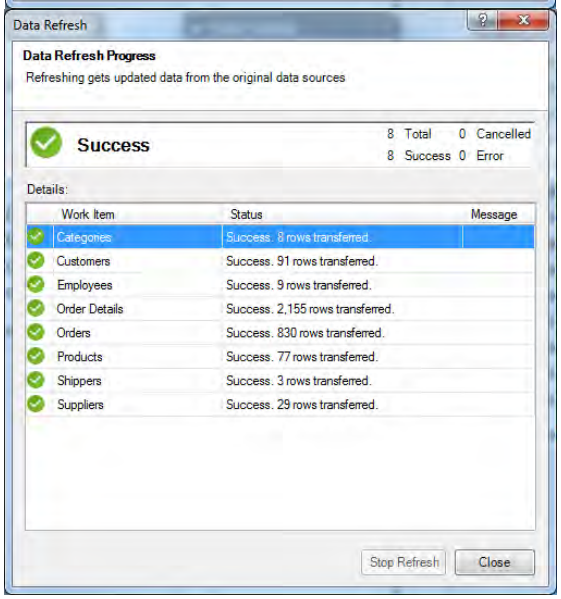

Change back Switch to: Table Preview, then change the filter to Select All.

## Refreshing Data

In the earlier description of Table Properties, we saw that there was an active connection to our original database. What happens if your database is updated? You don't need to re-import the data, just use the refresh option.

1. In the PowerPivot screen, click on the Home tab in the upper left corner
2. At the top of the screen is Refresh button. Hover over the button and it says: "Refresh the data that was imported from external data sources"
3. Click on the Suppliers table to highlight it.
4. Click on the down arrow below the refresh button.
5. Click on Refresh. PowerPivot goes to the Access database, retrieves the Suppliers data, and updates your table.
6. Similarly, try Refresh All.

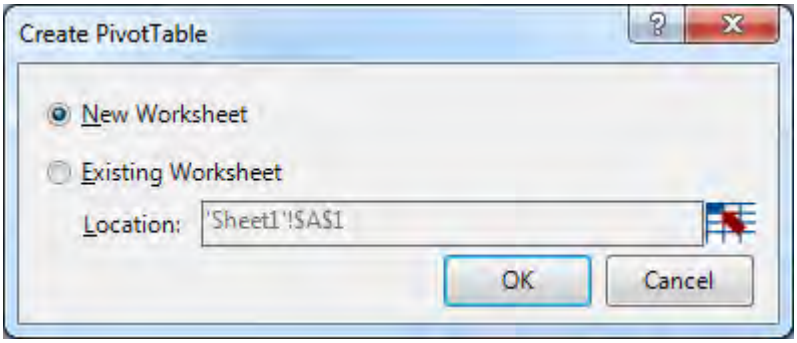


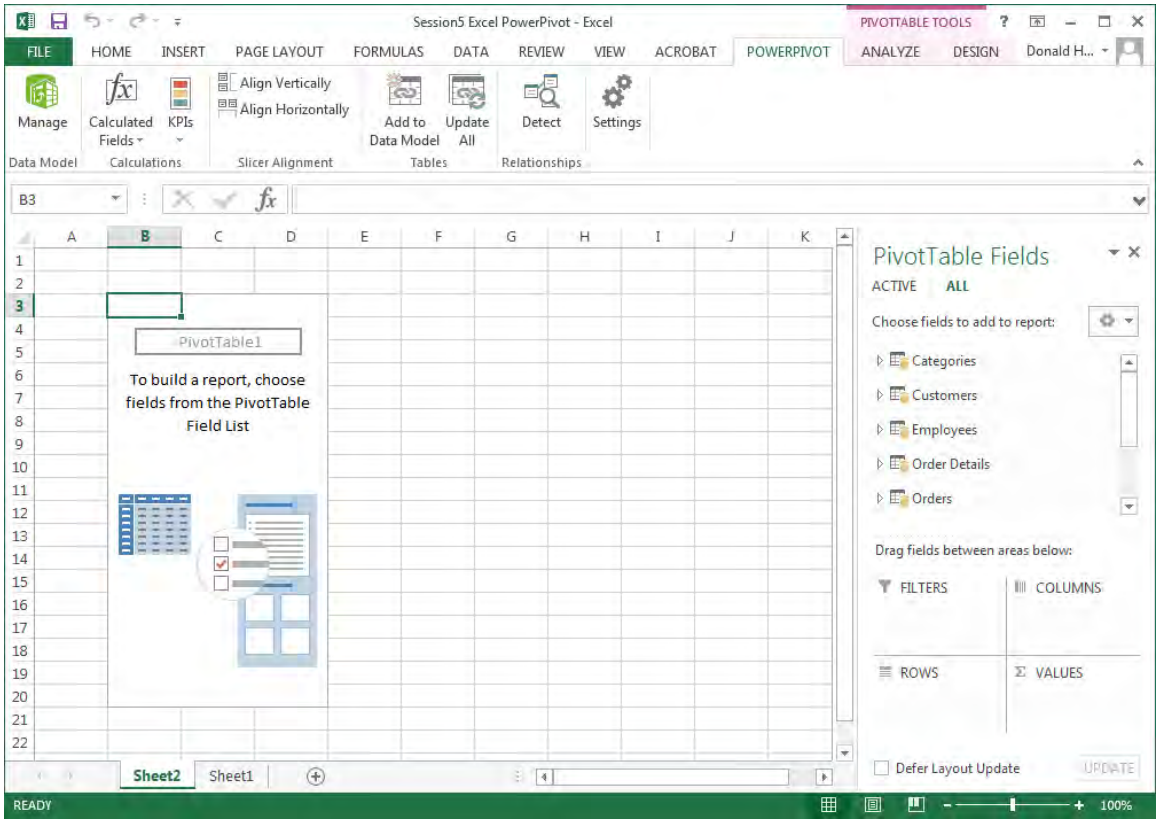



## Session 5.8: Creating Pivot Tables with PowerPivot

Let's now create Pivot Tables, incorporating data across multiple tables.

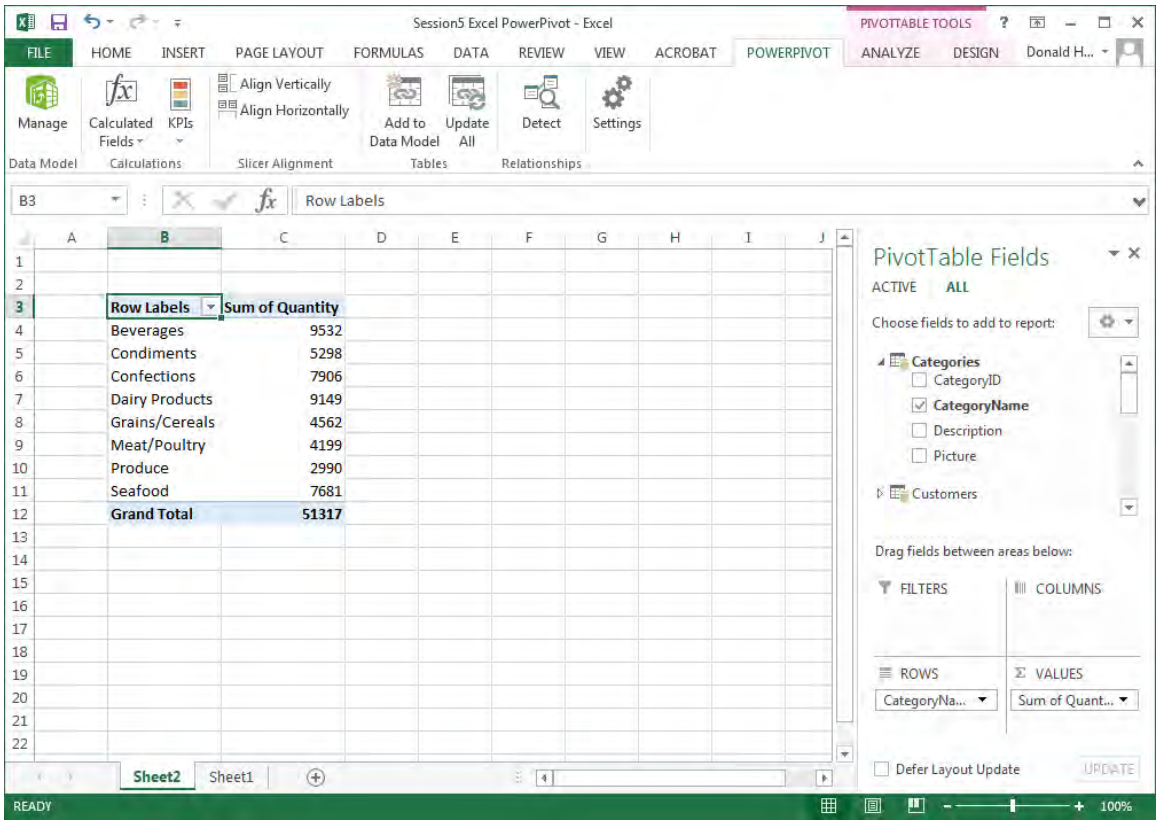
1. In PowerPivot, click on the Home tab in the upper left corner
2. At the top of the screen, click on the drop down arrow below PivotTable and select Pivot Table
3. Select New Worksheet, then OK




4. Notice that in the upper right of the pivot table, it lists PivotTable Fields. In this case, we now have all of the tables, not just one table.
5. Let's build a Pivot Table with the total quantity ordered for each category of products. Which table has CategoryName? Which table has Quantity (ordered)?

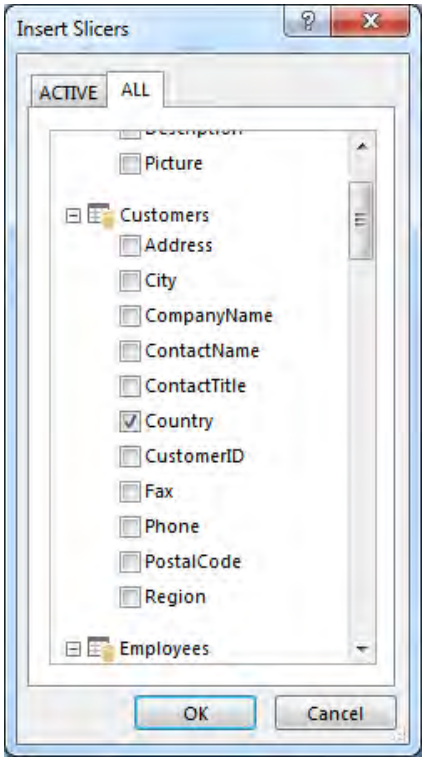
6. Click on the arrow next to the table Categories. Click on CategoryName and drag to Rows.
7. Scroll down in the PivotTable Fields until you see Order Details, then click on the arrow next to the table Order Details. Click on Quantity and drag it to Values.
8. You should see the view below.

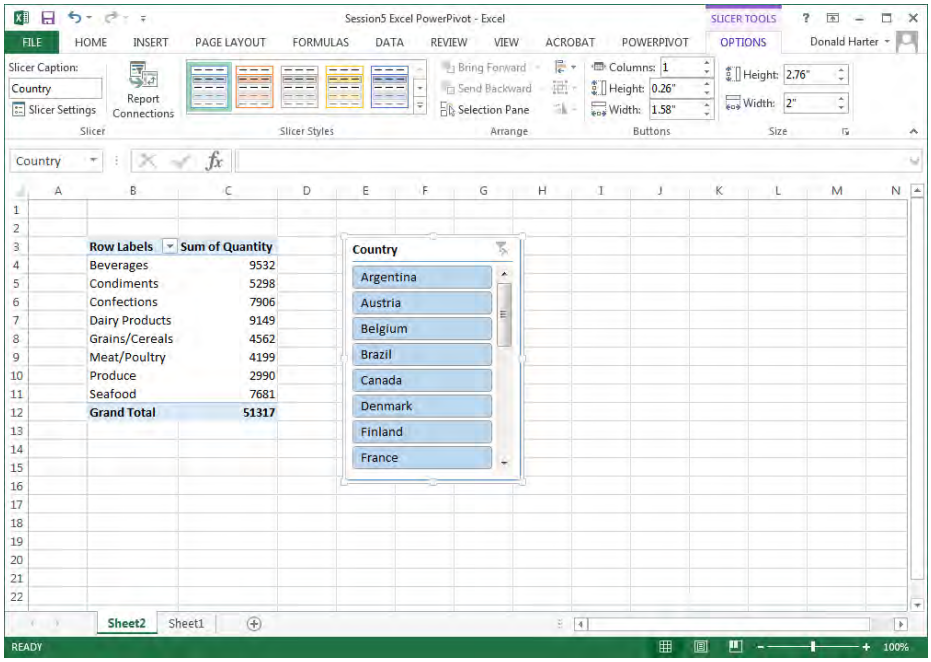


**Session 5.9: Slicers**

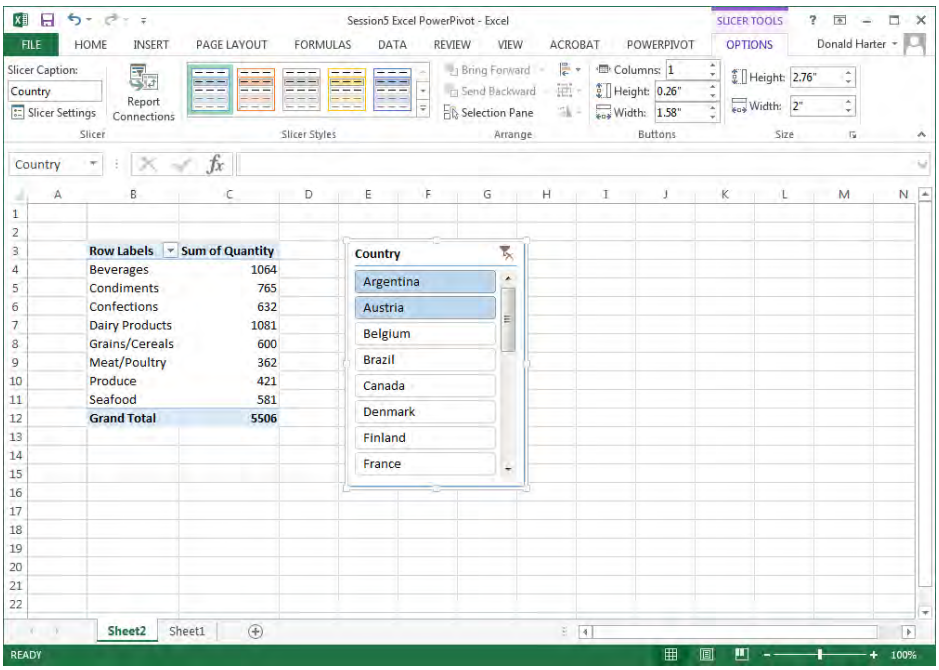
Let's dive deeper and examine quantity ordered by category and by a customer's country. A Slicer allows you to select data in varying "slices".

1. First, find where a customer's country is stored. It's in the table Customers, in the field Country.
2. In the PivotTable screen, look for PivotTable Tools in the upper right corner of the screen. Click on Analyze.
3. In the top middle of the screen, click on Insert Slicer
4. The Active tables are the ones we have used so far. But Country is stored in the Customers table. Click on ALL to see all tables, then check the box for Country in the Customers table. Click OK.






5. Notice that all countries are highlighted. Click on Argentina and watch how the numbers change.
6. To select more than one country, hold down the Control (Ctrl) key and select multiple countries. Select Argentina and Austria.
7. To return to the entire list, click on the funnel picture (filter picture) in the upper right corner of the slicer.

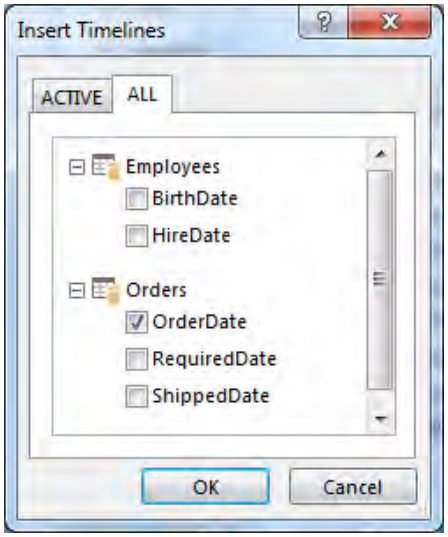


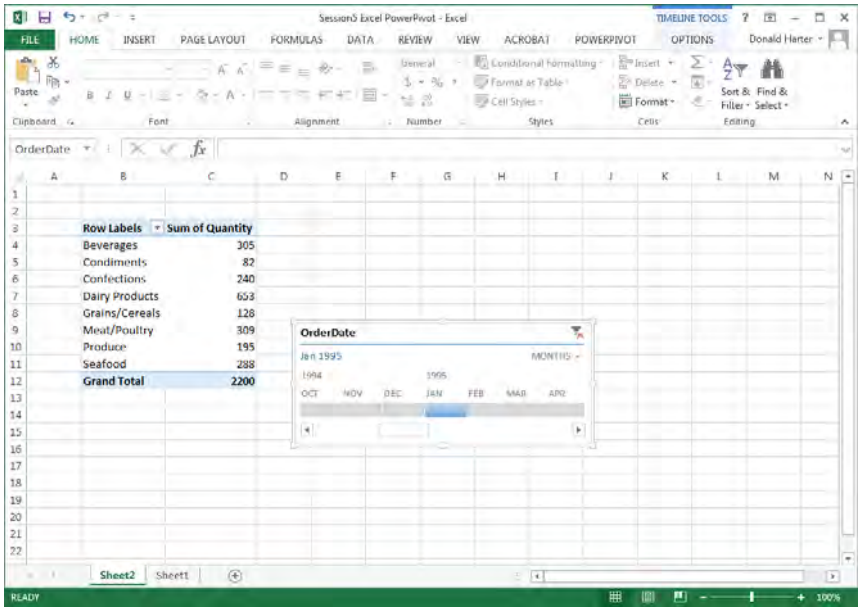

8. To remove the slicer, right click on the slicer and click on Remove “Country”

**Session 5.10: Timelines**

A useful technique for filtering data by dates is the Timeline option.

1. In the PivotTable Tools, Analyze tab, click on Insert Timeline
2. Click on the All tab and check the box for OrderDate in the Orders table, then OK
3. Notice that all dates are highlighted in blue. Click on January 1995 to highlight only that month
4. You can select multiple months by holding down the control key can clicking on multiple months.
5. Click on the filter in the upper right corner of the Timeline screen to bring back all data.
6. Right click on the Timeline and click Remove Timeline

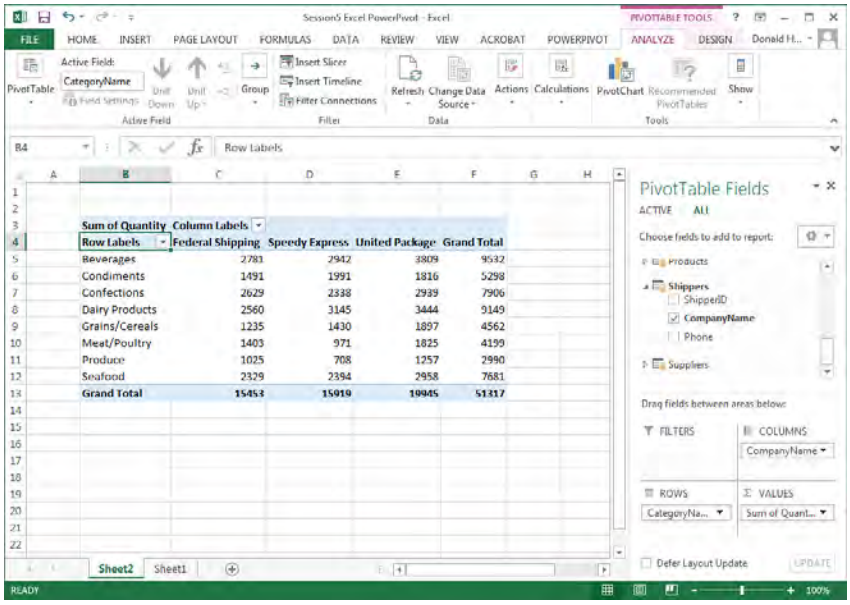


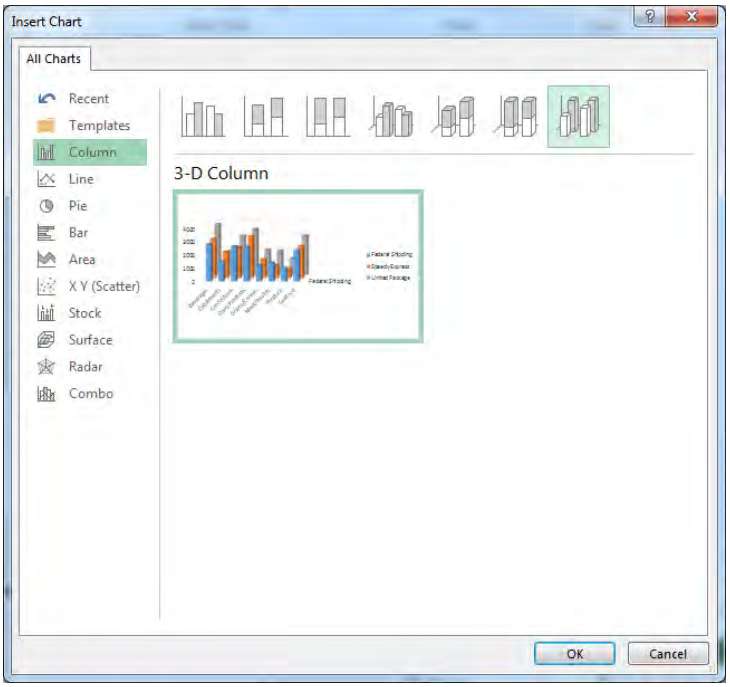
## Session 5.11: Pivot Charts

Pivot charts can also be created in Power Pivot.

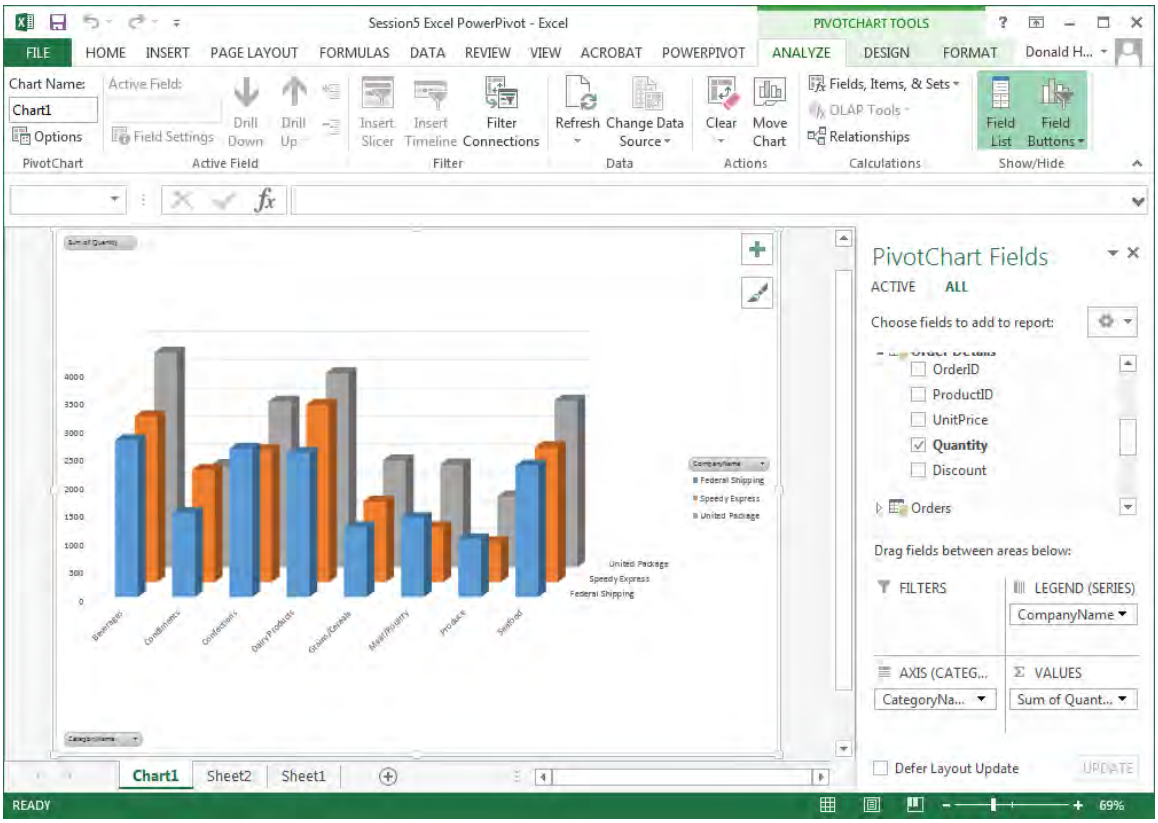
1. Click on PivotTable Tools, Analyze
2. Click on Field List to pop up the field list on the right side of the screen
3. Click on the arrow next to Shippers, check the CompanyName box, and drag it to Columns




4. Next, click on PivotChart in the upper right corner of the screen
5. Select my favorite, 3-D Column, then OK



6. Right click on the chart, select Move Chart, then check the button for New Sheet



## Optimization

- Set of techniques to find the maximum or minimum soluon for a problem
- **Goal Seek:** alter one variable to determine which value achieves a specific objecv e (e.g., product price)
- **Solver (unconstrained):** alter mulple v ariables to determine which combinaon r esults in a maximum or minimum result
- **Solver (constrained):** solve objecv e by altering mulple variables, subject to constraints



## Examples

- Find the demand for a product that results in a breakeven point.
- Find the product price that maximizes profit.
- Determine the opmal mix of pr oducts to be produced to maximize profit (subject to raw material and labor constraints).
- Schedule a workforce to minimize number of workers that sll c overs all shifts.
- Determine product distribuon plan fr om warehouses that sas fies demand and minimizes transportaon c osts.
- Select best projects in capital budgeng tha t opmiz e NPV (subject to budget constraints).



## History

- 1827—Joseph Fourier proposed inial op miz aon techniques.
- 1939—Kantorovich developed more modern approaches during WWII to opmiz e military planning (won Nobel Prize).
- 1947—Dantzig developed the simplex method, which is used in most soluons of linear pr oblems today.



## Terminology

- **Objective (objective function):** the variable that you are trying to opmiz e
  - E.g., maximize profit, minimize travel me
- **Constraints:** condions tha t limit the search for a soluon
  - E.g., labor, raw material, inventory

## Techniques

1. Linear programming assumes that the goal is linear and the changing variables are connuous.
2. Nonlinear programming assumes that the goal is nonlinear and the changing variables are connuous.
3. Integer programming assumes the changing variable must be a whole number.

All of these are automated within Excel.

View All View Keyframes

## Goal Seek

- Excel technique that searches for values of a variable that will result in a precise outcome
- Only allows one variable to be changed
- Measures the effect of these changes on the objecv e unl the g oal is achieved

4 of 8

## Sample Problem

How many products must be sold to reach the breakeven point (profit = zero) for that product?

6 of 8

## Sample Problem

How many products must be sold to reach the breakeven point (profit = zero) for that product?

- Goal Seek allows you to set profit at zero.
- Vary demand unl g oal is reached.

8 of 8

View All View Keyframes

## Unconstrained Optimization with Solver

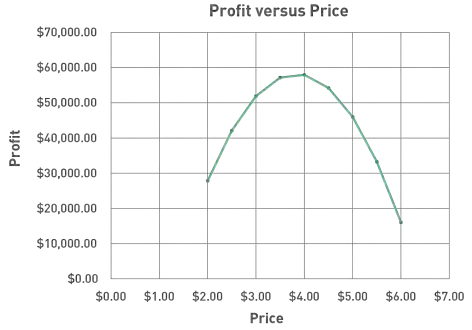
- Opmiz aon ma ximizes or minimizes objecv e.
- Unconstrained opmiz aon has no limits on the possibilies of the v ariables.
- Relies on linear, nonlinear, or integer programming techniques to find the soluon, depending on the pr oblem.
- **Sample problem:** Find the price of a product that maximizes profit, recognizing that as price increases, demand will decrease.

6 of 14

## Demand and Profit

7 of 14

## Demand and Profit



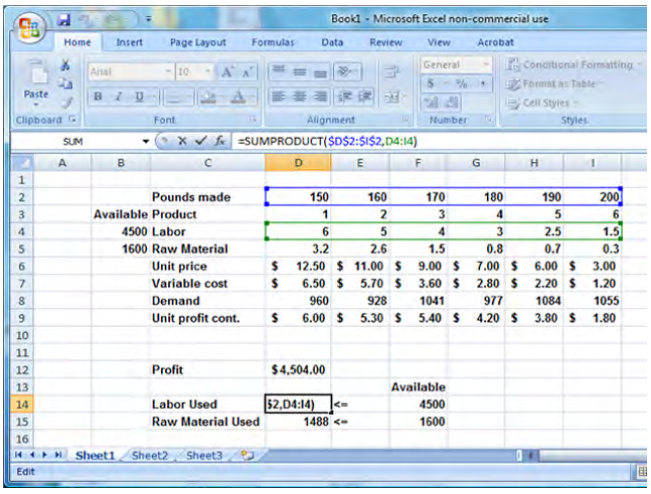

9 of 14

View All View Keyframes

## Optimization and Useful Solvers

- Opmiz ation often requires seng up c omplex equaons, which r equires you to mulply mulple r ows of data.
- SUMPRODUCT takes the product of two separate rows/columns, mulplies them t ogether, and calculates the sum.

3 of 12




4 of 12

## Example

- Labor used = SUMPRODUCT (D2:I2, D4:I4)
- Mulplies D2*D4, E2*E4, … , then adds sums t ogether

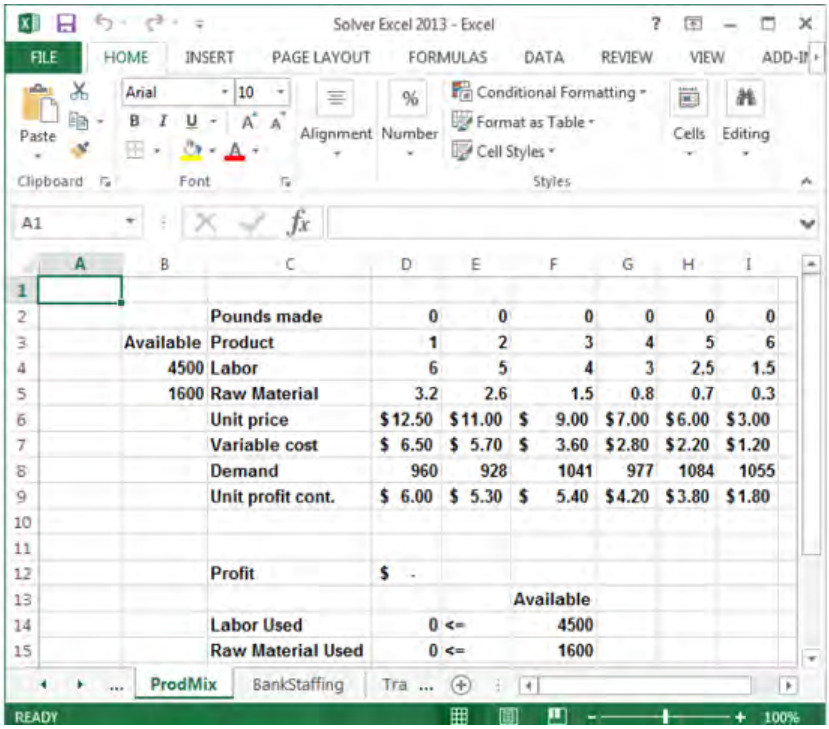
12 of 12

View All View Keyframes

## Product Mix

- Company generates a variety of products.
  - Each product generates profit.
  - Each product requires raw material and labor.
- **Product mix** determines how much of each product to manufacture subject to constraints of raw material and labor.

5 of 9




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | Pounds made | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | Available | Product | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | 4500 | Labor | 6 | 5 | 4 | 3 | 2.5 | 1.5 |
| 5 | 1600 | Raw Material | 3.2 | 2.6 | 1.5 | 0.8 | 0.7 | 0.3 |
| 6 | | Unit price | $12.50 | $11.00 | $ 9.00 | $7.00 | $6.00 | $3.00 |
| 7 | | Variable cost | $ 6.50 | $ 5.70 | $ 3.60 | $2.80 | $2.20 | $1.20 |
| 8 | | Demand | 960 | 928 | 1041 | 977 | 1084 | 1055 |
| 9 | | Unit profit cont. | $ 6.00 | $ 5.30 | $ 5.40 | $4.20 | $3.80 | $1.80 |
| 10 | | | | | | | | |
| 11 | | | | | | | | |
| 12 | | Profit | $ - | | | | | |
| 13 | | | | | Available | | | |
| 14 | | Labor Used | 0 <= | | 4500 | | | |
| 15 | | Raw Material Used | 0 <= | | 1600 | | | |

6 of 9

View All View Keyframes

## Workforce Scheduling

- Problems are common where shift work requires minimal but different levels of staffing.
- Must determine how many staff are needed by shift rotation to cover all staffing needs.
- Overall goal is to minimize worker expense.

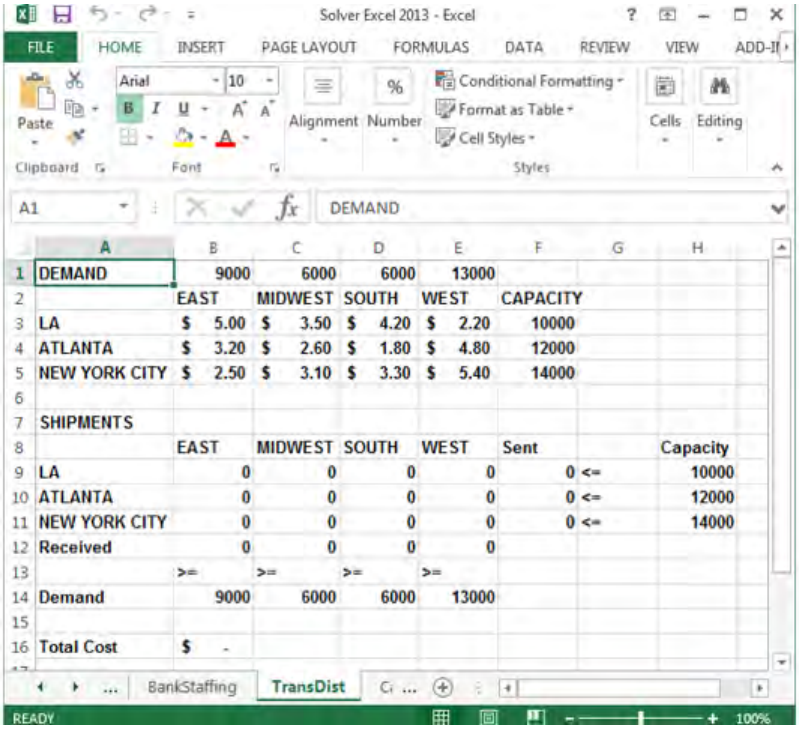
7 of 17

View All View Keyframes

## Transportation and Distribution

- Corporations try to find the cheapest way to transport products from warehouses to customers.
- The challenge is that warehouses and customers are located all over the country.
- The goal is to satisfy demand with the lowest transportation cost.

4 of 13



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | DEMAND | 9000 | 6000 | 6000 | 13000 | | | |
| 2 | | EAST | MIDWEST | SOUTH | WEST | CAPACITY | | |
| 3 | LA | $ 5.00 | $ 3.50 | $ 4.20 | $ 2.20 | 10000 | | |
| 4 | ATLANTA | $ 3.20 | $ 2.60 | $ 1.80 | $ 4.80 | 12000 | | |
| 5 | NEW YORK CITY | $ 2.50 | $ 3.10 | $ 3.30 | $ 5.40 | 14000 | | |
| 6 | | | | | | | | |
| 7 | SHIPMENTS | | | | | | | |
| 8 | | EAST | MIDWEST | SOUTH | WEST | Sent | | Capacity |
| 9 | LA | 0 | 0 | 0 | 0 | 0 | <= | 10000 |
| 10 | ATLANTA | 0 | 0 | 0 | 0 | 0 | <= | 12000 |
| 11 | NEW YORK CITY | 0 | 0 | 0 | 0 | 0 | <= | 14000 |
| 12 | Received | 0 | 0 | 0 | 0 | | | |
| 13 | | >= | >= | >= | >= | | | |
| 14 | Demand | 9000 | 6000 | 6000 | 13000 | | | |
| 15 | | | | | | | | |
| 16 | Total Cost | $ - | | | | | | |



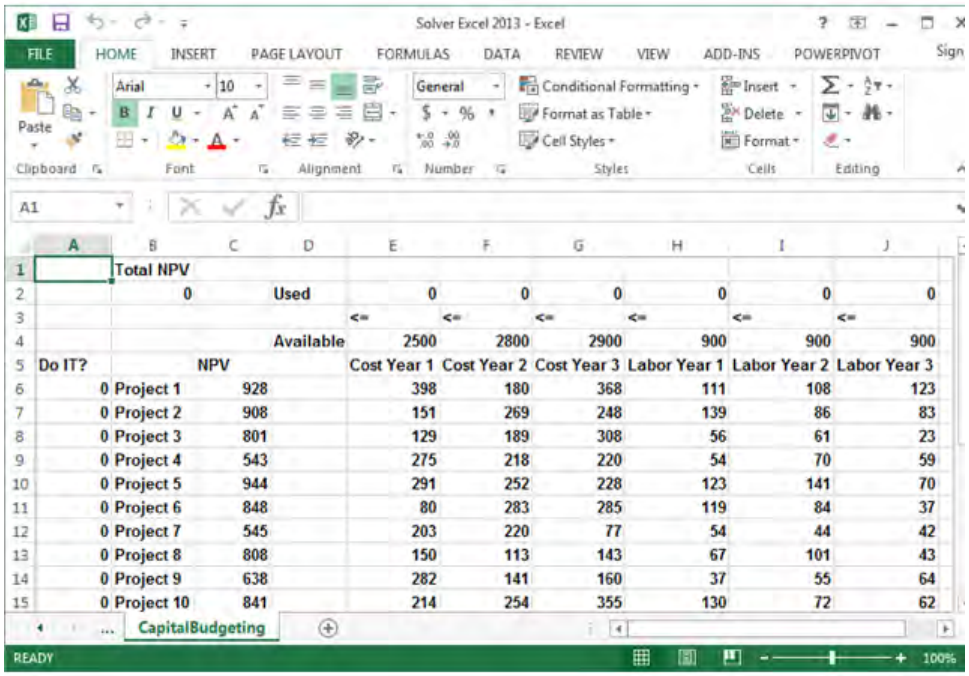
5 of 13

View All View Keyframes

## Capital Budgeting

- Organizations have many projects they want to take on yet are constrained by their budget.
- Capital budgeting researches evaluates to determine which projects should be chosen.
- Goal: Maximize NPV subject to investment cost and labor constraints.

4 of 10



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Total NPV | | | | | | | | |
| 2 | | 0 | | Used | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | | | | | <= | <= | <= | <= | <= | <= |
| 4 | | | | Available | 2500 | 2800 | 2900 | 900 | 900 | 900 |
| 5 | Do IT? | | NPV | | Cost Year 1 | Cost Year 2 | Cost Year 3 | Labor Year 1 | Labor Year 2 | Labor Year 3 |
| 6 | 0 | Project 1 | 928 | | 398 | 180 | 368 | 111 | 108 | 123 |
| 7 | 0 | Project 2 | 908 | | 151 | 269 | 248 | 139 | 86 | 83 |
| 8 | 0 | Project 3 | 801 | | 129 | 189 | 308 | 56 | 61 | 23 |
| 9 | 0 | Project 4 | 543 | | 275 | 218 | 220 | 54 | 70 | 59 |
| 10 | 0 | Project 5 | 944 | | 291 | 252 | 228 | 123 | 141 | 70 |
| 11 | 0 | Project 6 | 848 | | 80 | 283 | 285 | 119 | 84 | 37 |
| 12 | 0 | Project 7 | 545 | | 203 | 220 | 77 | 54 | 44 | 42 |
| 13 | 0 | Project 8 | 808 | | 150 | 113 | 143 | 67 | 101 | 43 |
| 14 | 0 | Project 9 | 638 | | 282 | 141 | 160 | 37 | 55 | 64 |
| 15 | 0 | Project 10 | 841 | | 214 | 254 | 355 | 130 | 72 | 62 |

5 of 10

# SCM 651: Business Analytics

1

Week 6

# Agenda



- Homework #2: discussion
- Homework #3: overview
- Review of concepts
- Group discussion of articles
  - Modern Analytics and the Future of Quality Performance Excellence
  - A Process of Continuous Innovation: Centralizing Analytics at Caesars

# Homework #2



- Campaign time frames, costs, effectiveness
- Future campaign: geographic regions, key words, day of week and time of day
- Allocation of costs by program and region
- Performance measures
- Other data that would be helpful

# Homework #3



1. Graph, regression, calculated sales, revenue, profit
2. Constrained optimization
3. Discussion of risks, other data which would be valuable

# Week 6 - Review



- Goal Seek
  - Searches for one goal such as break even point (profit = 0)
  - Allows one variable to be changed in search of the goal
  - Does not allow constraints on the search
- Unconstrained optimization
  - Does not constrain any variables in the search
  - Can search for maximum, minimum, or specific values
  - Can use linear programming (straight line functions) or non-linear programming (curved functions)

# Week 6 - Review



- Valuable functions: Sumproduct
  - Multiplies rows or columns together, than adds result
- Constrained optimization
  - Can set variables to less than or greater than some constraint
  - Can set variables to be integer or binary

# Article #1: Modern Analytics and the Future of Quality Performance Excellence



- Modern Analytics and the Future of Quality Performance Excellence
  - Define analytics (page 6)
  - How are companies using analytics in (page 7):
    - Banks
    - Manufacturing
    - Retail
    - Pharmaceuticals
    - Sports
  - Modern analytics integrates which three fields (page 8)?
  - What are some examples of data sources (page 9)?
  - What are examples of data visualization (page 11)?

# Article #1: Modern Analytics and the Future of Quality Performance Excellence



- Modern Analytics and the Future of Quality Performance Excellence
  - Define analytics (page 6)
    - "a process of transforming data into actions through analysis and insights in the context of organizational decision making and problem solving"
    - "the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions"

# Article #1: Modern Analytics and the Future of Quality Performance Excellence



- Modern Analytics and the Future of Quality Performance Excellence
  - How are companies using analytics in (page 7):
    - Banks: prevent fraud
    - Manufacturing: production planning, purchasing, inventory mgt
    - Retail: recommend products and optimize marketing promotions
    - Pharmaceuticals: get drugs to market more quickly
    - Sports: determine game strategy and optimal ticket prices

# Article #1: Modern Analytics and the Future of Quality Performance Excellence



- Modern Analytics and the Future of Quality Performance Excellence
  - Modern analytics integrates which three fields (page 8)?
    - Business intelligence/information systems
    - Statistics
    - Quantitative methods/operations research
  - What are some examples of data sources (page 9)?
    - Supermarket scanners
    - Click streams from the web
    - Customer transactions
    - Email, tweets, social media
  - What are examples of data visualization (page 11)?
    - Dashboards and scorecards

# Article #2: A Process of Continuous Innovation: Centralizing Analytics at Caesars



- A Process of Continuous Innovation: Centralizing Analytics at Caesars
  - Why does Caesars use analytics (pages 1 & 2)?
  - What are four lessons learned from their experience (page 3)?

# Article #2: A Process of Continuous Innovation: Centralizing Analytics at Caesars



- A Process of Continuous Innovation: Centralizing Analytics at Caesars
  - Why does Caesars use analytics (pages 1 & 2)?
    - Create a rich customer experience
    - Marketing based not only on their preferences but on their actions
    - Gaming analytics, revenue management, finance, marketing analytics, hotel operations and labor
  - What are four lessons learned from their experience (page 3)?
    - Sense of scale
    - Adequate infrastructure
    - Communications with stakeholders
    - Visible and meaningful wins

# Microsoft Excel: Goal Seek and Solver

**Session 6.4: Goal seek**

This document is a quick reference guide for reviewing the techniques in the laboratory.

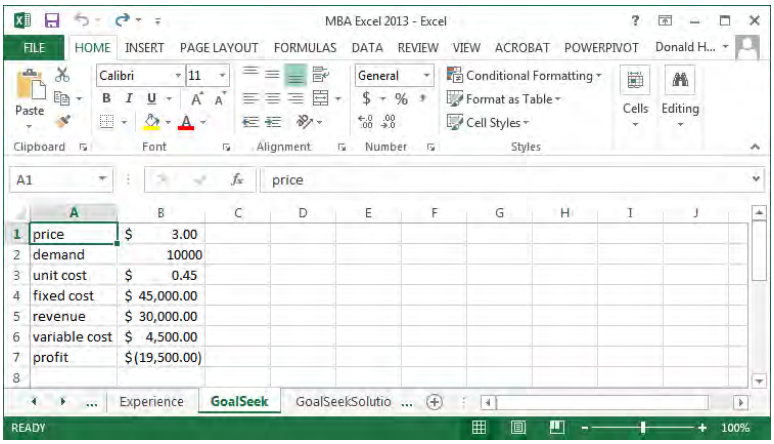
For these exercises, download the files:
“Business Analytics – Week 6 Instructions.doc”
“Business Analytics – Week 6 Excel 2013.xls”

The goal seek command searches for a solution by varying a parameter until a solution is found.

Definitions:

Set cell: formula for the answer that you are seeking
To value: value that you want to achieve
By changing cell: the cell that changes until the Set cell matches the To value.

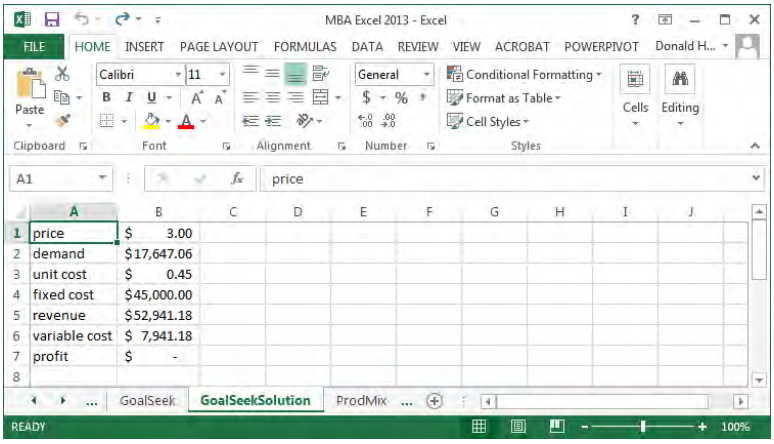
Let’s find the number of widgets that needs to be sold to break even for a given price. In other words, we will vary price, which will drive demand, until profit equals zero. Use the spreadsheet GoalSeek.




The following commands run GoalSeek.

1. Click on the Data tab at the top of the spreadsheet.
2. Click on What-if Analysis, Goal Seek.
3. For Set Cell, enter the cell that is the outcome variable that you want to match. Since we are measuring profit as the outcome, enter $B$7.
4. For the To Value, enter zero, since we are trying to find the break-even point.
5. For the By Changing Cell, enter the cell which will change, in this case price $B$1.
6. Click OK to find the solution.

You should get the solution shown below.




| | A | B |
|---|---|---|
| 1 | price | $ 3.00 |
| 2 | demand | $17,647.06 |
| 3 | unit cost | $ 0.45 |
| 4 | fixed cost | $45,000.00 |
| 5 | revenue | $52,941.18 |
| 6 | variable cost | $ 7,941.18 |
| 7 | profit | $ - |
| 8 | | |

**Session 6.5: Solver Add-in**

The Solver option is available as an add-in to Excel. The steps to add it are:

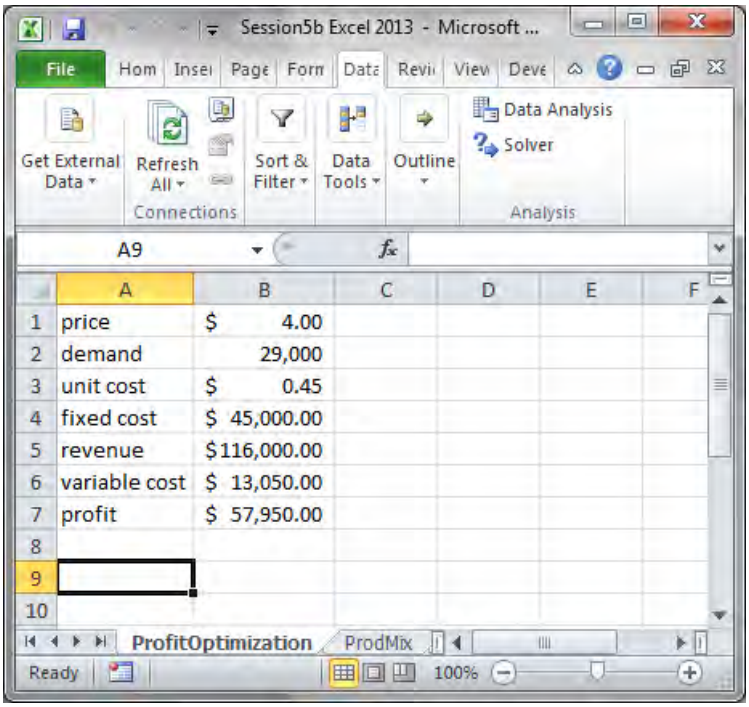
1. In Excel, click on the File tab, then Options
2. Click on Add-Ins
3. Click Solver Add-in, then Go
4. Check the box for Solver Add-in, then OK

**Introduction to solver**

Solver has the ability to search for an optimal solution subject to constraint. In contrast, goal seek would search for a solution that would match a specific value, such as demand-supply=0. Solver will find the maximum or minimum of a function, called the optimal feasible solution, if one exists. In some cases, particularly overly constrained problems, there will be no solution.

**Session 6.6: Simple Optimization**

For this example, use the ProfitOptimization spreadsheet. Recall the price-demand function from earlier.

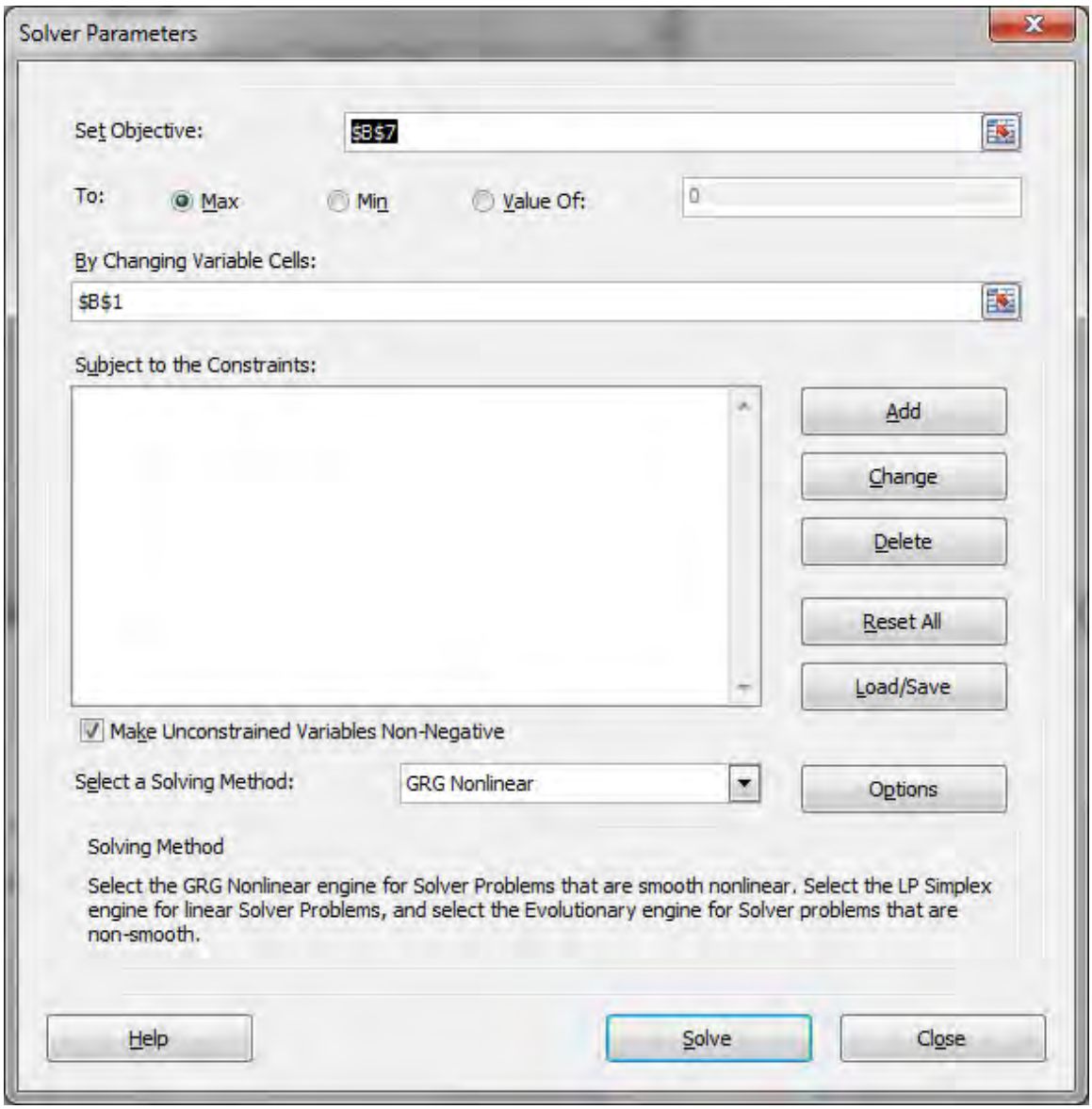



| | A | B |
|---|---|---|
| 1 | price | $ 4.00 |
| 2 | demand | 29,000 |
| 3 | unit cost | $ 0.45 |
| 4 | fixed cost | $ 45,000.00 |
| 5 | revenue | $116,000.00 |
| 6 | variable cost | $ 13,050.00 |
| 7 | profit | $ 57,950.00 |
| 8 | | |
| 9 | | |
| 10 | | |

As price increases, demand decreases. Attempt to find the price that maximizes profit by changing the price.

Let's use Solver to quickly calculate the optimal price.

1. Click on the Data tab, then click on Solver.
2. The Objective is the cell that you want to maximize or minimize. We want to maximize profit, so enter B7 in the objective, and click on Max
3. Since we want to find the price that results in maximum profit, we will change price until we find a solution. In the box labelled "By Changing Variable Cells", enter B1, which is where price is stored.
4. Now click Solve.

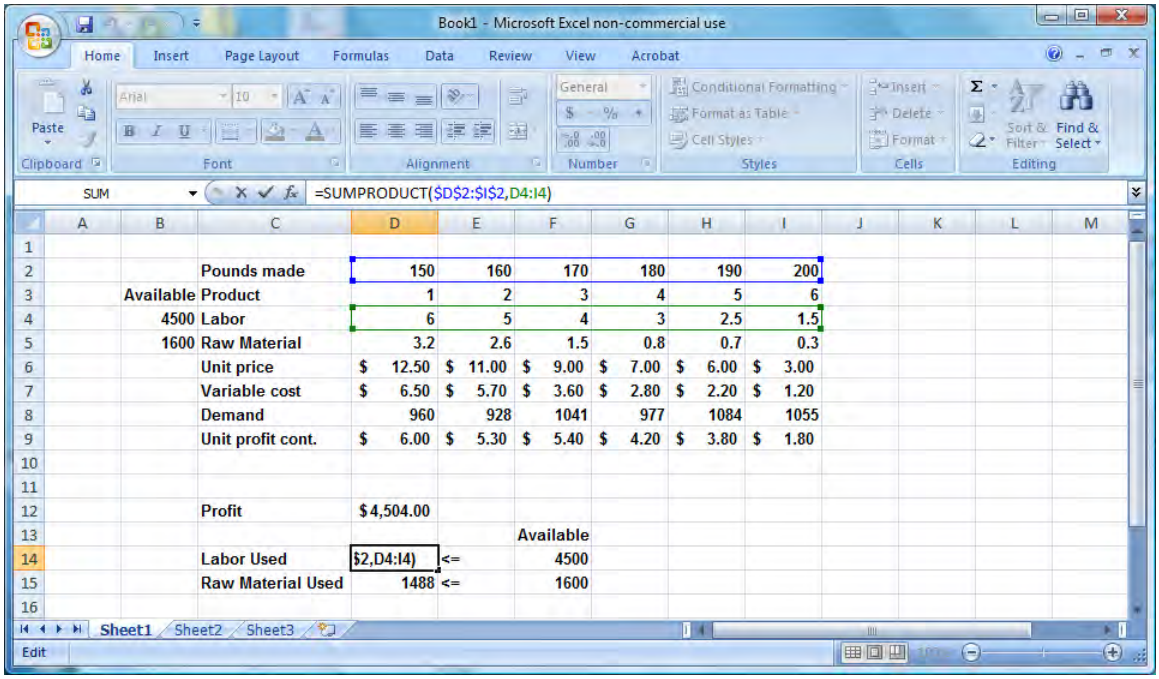



You can keep the solver solutions, or restore original values.

**Session 6.7: Useful Functions for Solver**

One function that is helpful in setting up optimization problems is the SUMPRODUCT function. SUMPRODUCT first multiplies two columns or rows together (product), then adds the resulting values (sum). In the example below, =SUMPRODUCT($D$2:$I$2,D4:I4) will multiply the values in D2:I2 against the paired value in D4:I4, i.e., D2*D4, E2*E4, etc., then add together the results. This concept will be used in each of the following chapters.

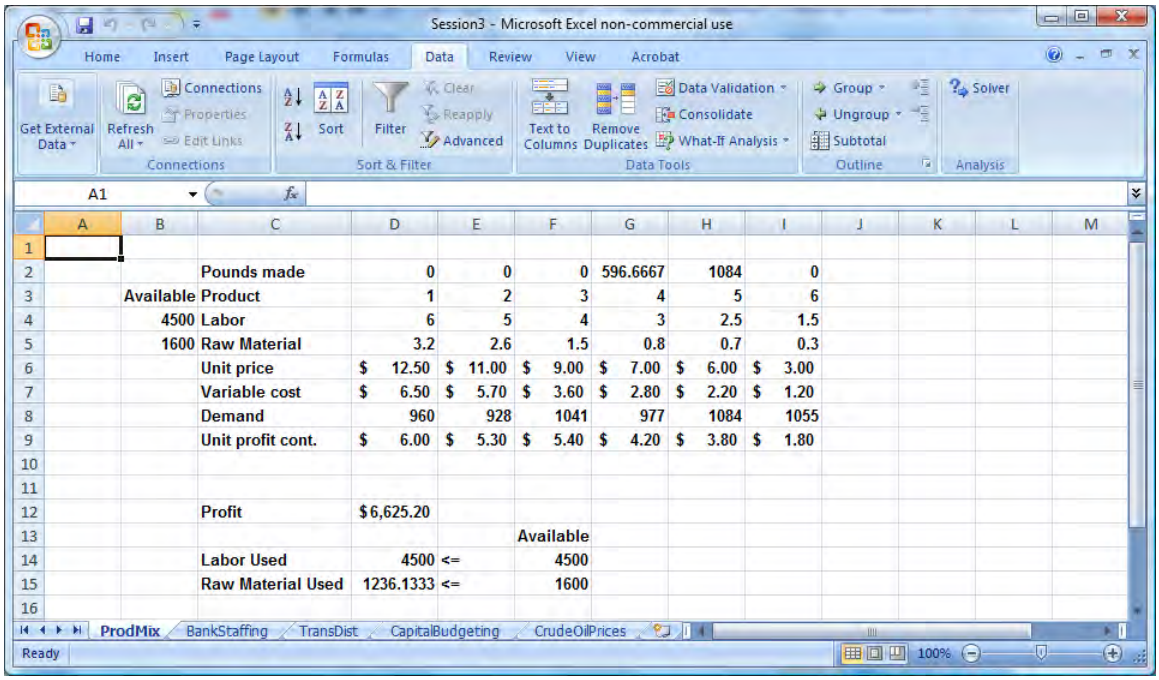
Use the ProdMix spreadsheet to see the example below.




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | Pounds made | 150 | 160 | 170 | 180 | 190 | 200 |
| 3 | | Available | Product | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | | 4500 | Labor | 6 | 5 | 4 | 3 | 2.5 | 1.5 |
| 5 | | 1600 | Raw Material | 3.2 | 2.6 | 1.5 | 0.8 | 0.7 | 0.3 |
| 6 | | | Unit price | $ 12.50 | $ 11.00 | $ 9.00 | $ 7.00 | $ 6.00 | $ 3.00 |
| 7 | | | Variable cost | $ 6.50 | $ 5.70 | $ 3.60 | $ 2.80 | $ 2.20 | $ 1.20 |
| 8 | | | Demand | 960 | 928 | 1041 | 977 | 1084 | 1055 |
| 9 | | | Unit profit cont. | $ 6.00 | $ 5.30 | $ 5.40 | $ 4.20 | $ 3.80 | $ 1.80 |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | Profit | $4,504.00 | | | | | |
| 13 | | | | | | Available | | | |
| 14 | | | Labor Used | $2,D4:I4) | <= | 4500 | | | |
| 15 | | | Raw Material Used | 1488 | <= | 1600 | | | |
| 16 | | | | | | | | | |

**Session 6.8: Optimal product mix**

For this problem, use the ProdMix spreadsheet.
In the optimal product mix problem, we are trying to produce six different products, where the labor and raw materials are constrained. The goal is to find the production volumes for each product that maximizes profitability.

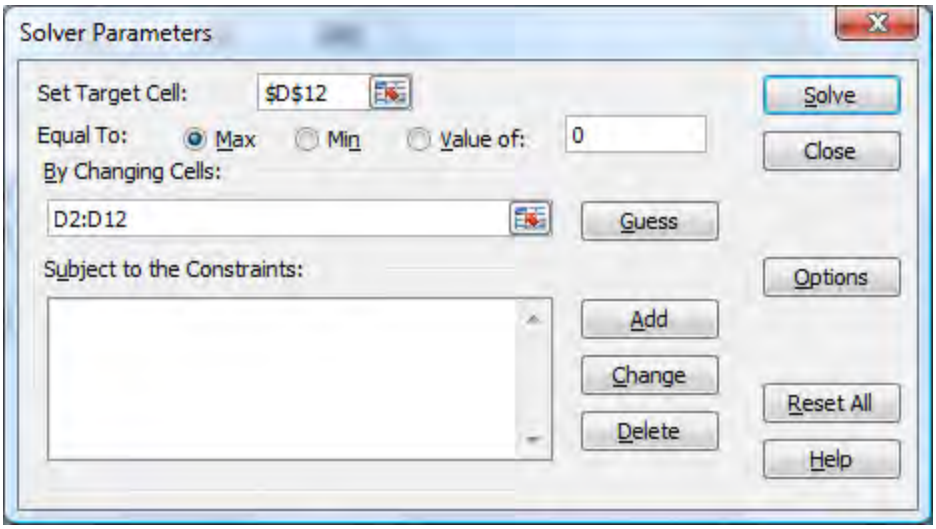


| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | Pounds made | 0 | 0 | 0 | 596.6667 | 1084 | 0 |
| 3 | | Available | Product | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | | 4500 | Labor | 6 | 5 | 4 | 3 | 2.5 | 1.5 |
| 5 | | 1600 | Raw Material | 3.2 | 2.6 | 1.5 | 0.8 | 0.7 | 0.3 |
| 6 | | | Unit price | $ 12.50 | $ 11.00 | $ 9.00 | $ 7.00 | $ 6.00 | $ 3.00 |
| 7 | | | Variable cost | $ 6.50 | $ 5.70 | $ 3.60 | $ 2.80 | $ 2.20 | $ 1.20 |
| 8 | | | Demand | 960 | 928 | 1041 | 977 | 1084 | 1055 |
| 9 | | | Unit profit cont. | $ 6.00 | $ 5.30 | $ 5.40 | $ 4.20 | $ 3.80 | $ 1.80 |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | Profit | $6,625.20 | | | | | |
| 13 | | | | | | Available | | | |
| 14 | | | Labor Used | 4500 | <= | 4500 | | | |
| 15 | | | Raw Material Used | 1236.1333 | <= | 1600 | | | |
| 16 | | | | | | | | | |

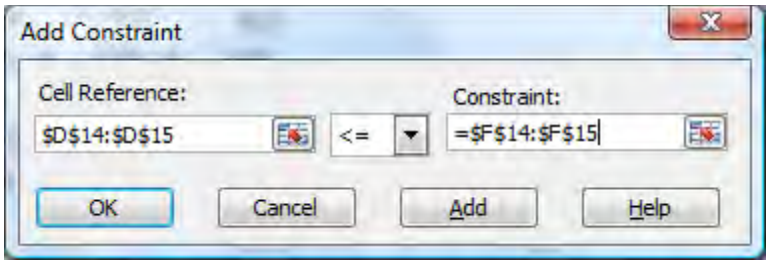
The product numbers are listed in row 3, with the labor hours and raw material required to produce one pound of that product in rows 4 and 5. The maximum demand is in row 8. The maximum labor and raw materials available are also identified. Demand, labor and raw materials are constraints.

To use solver, follow these steps:

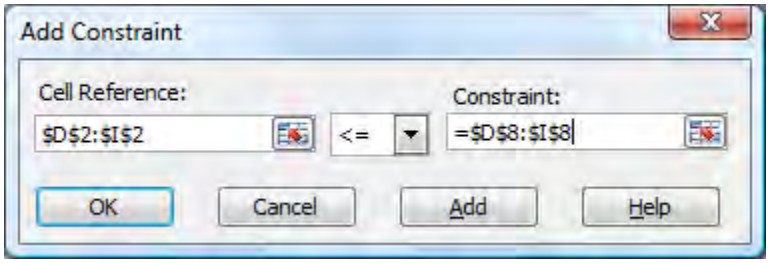
1. Click on the Data tab
2. In the Analysis group, click Solver.
3. Set Target Cell is the cell that you want to maximize or minimize. In this case, set it to profit ($D$12)
4. Since we want to maximize profit, make sure the Max button is checked.
5. By Changing Cell refers to what you want to vary. In this case, we want to vary product quantities to find the optimal production quantities. Click in the box and set this to D2:D12. The screen should look like the picture below.



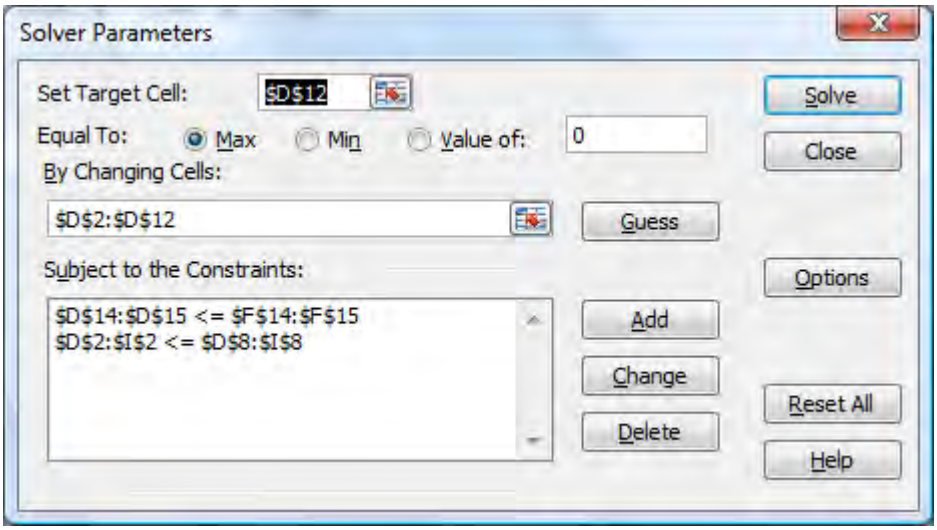

6. Next, add the constraints. First, click the Add button.
7. We want to impose a labor and resource constraint. We can do both simultaneously by entering D14:D15 and F14:F15 as below. Set the constraint direction to <=.
8. Click Add to add another constraint.



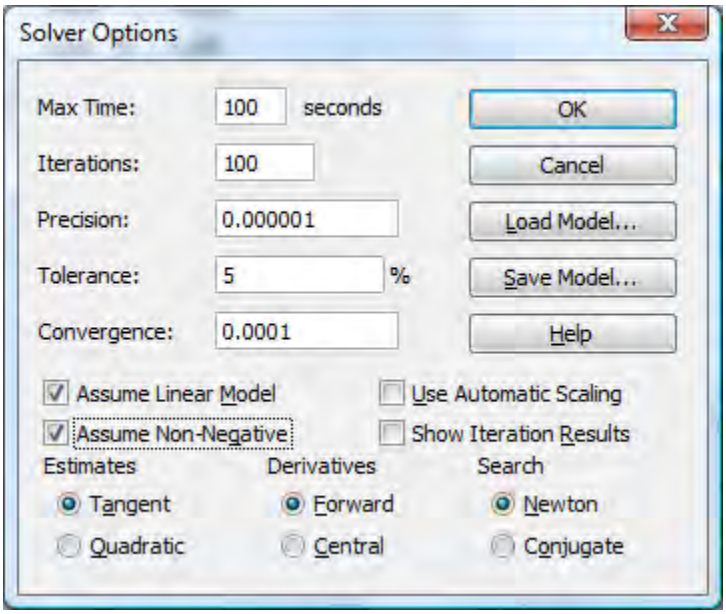

9. We also want to make sure that we don't produce more than is demanded. After clicking add in the earlier step, Add Constraint let's you add more constraints. In this case, production (row 2) should not exceed demand (row 8). Set this demand constraint by entering D2:I2 and D8:I8. Set the constraint direction to <=.



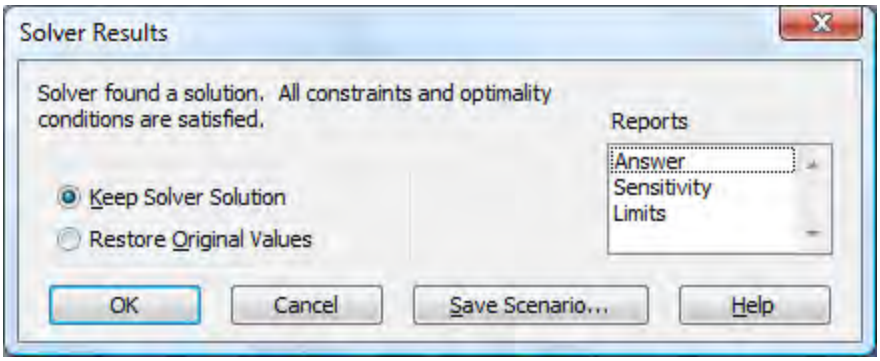

10. Since this is the last constraint, click OK. You should be returned to the following screen.



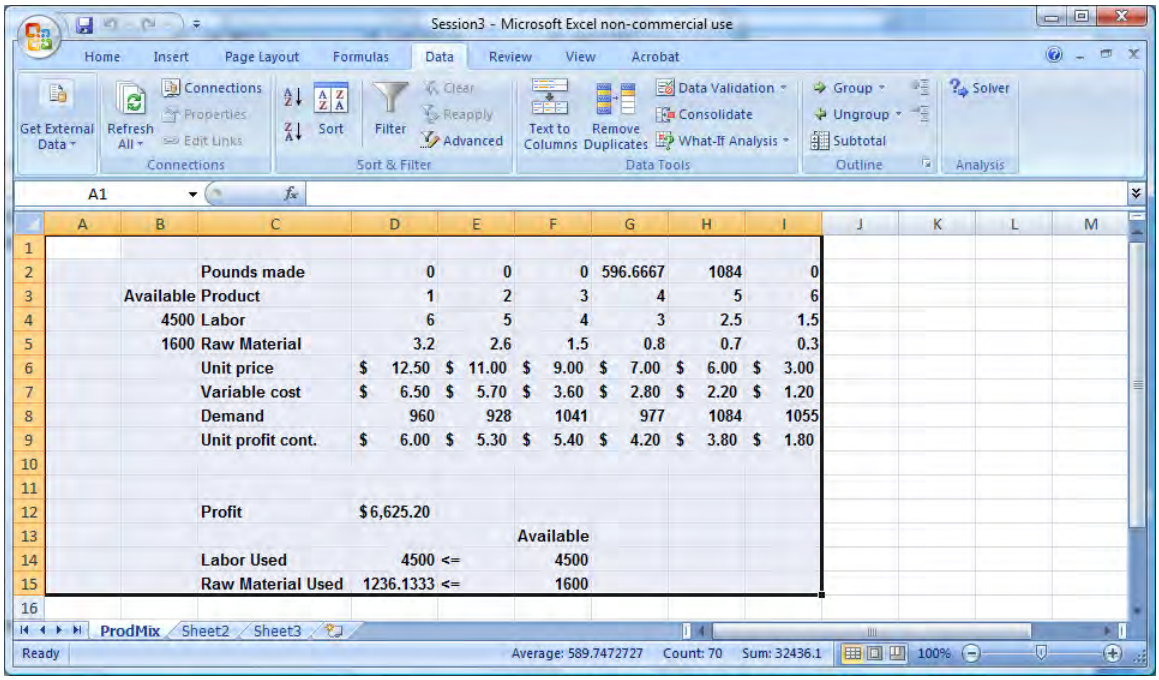

11. Next, we want to enter a constraint to make sure that the changing cells are not negative. Click the Options button, check Assume Linear Model and check Assume Non-Negative, then click OK.




12. Linear models are guaranteed to converge to a solution. Non-linear models might not converge to the optimal solution.
13. Finally, click on Solve. Select Keep Solver Solution if you want to save your answer. You can select three reports (Answer, Sensitivity, Limits) by clicking on them.



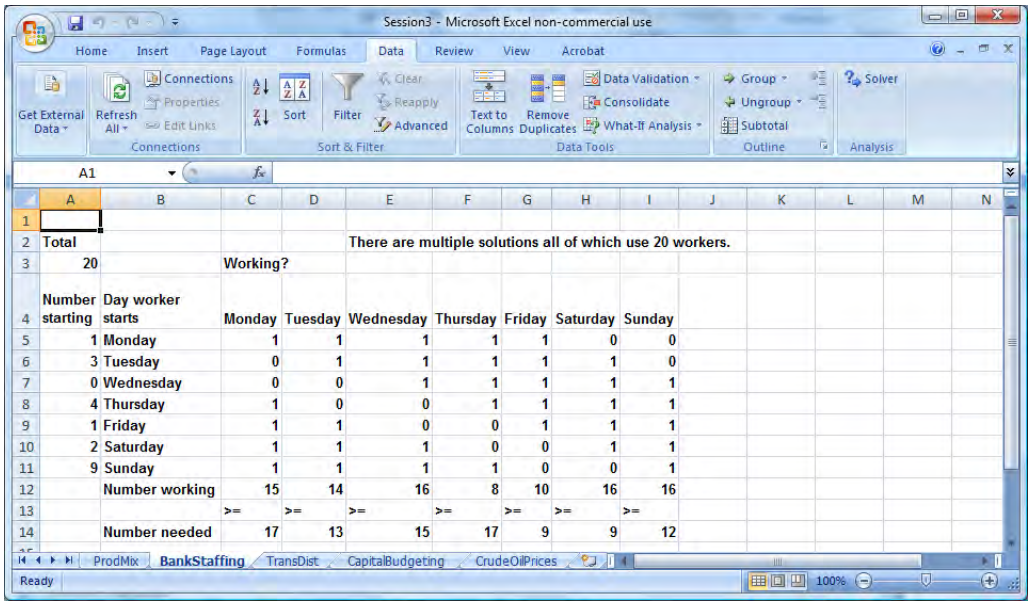

14. When you've made your selections, click OK. You should see the solution shown below. If you clicked on the additional reports, they will be listed as separate tabs.




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Pounds made | 0 | 0 | 0 | 596.6667 | 1084 | 0 |
| | Available | Product | 1 | 2 | 3 | 4 | 5 | 6 |
| | 4500 | Labor | 6 | 5 | 4 | 3 | 2.5 | 1.5 |
| | 1600 | Raw Material | 3.2 | 2.6 | 1.5 | 0.8 | 0.7 | 0.3 |
| | | Unit price | $ 12.50 | $ 11.00 | $ 9.00 | $ 7.00 | $ 6.00 | $ 3.00 |
| | | Variable cost | $ 6.50 | $ 5.70 | $ 3.60 | $ 2.80 | $ 2.20 | $ 1.20 |
| | | Demand | 960 | 928 | 1041 | 977 | 1084 | 1055 |
| | | Unit profit cont. | $ 6.00 | $ 5.30 | $ 5.40 | $ 4.20 | $ 3.80 | $ 1.80 |
| | | Profit | $6,625.20 | | | | | |
| | | | | | Available | | | |
| | | Labor Used | 4500 | <= | 4500 | | | |
| | | Raw Material Used | 1236.1333 | <= | 1600 | | | |

**Session 6.9: Schedule workforce**

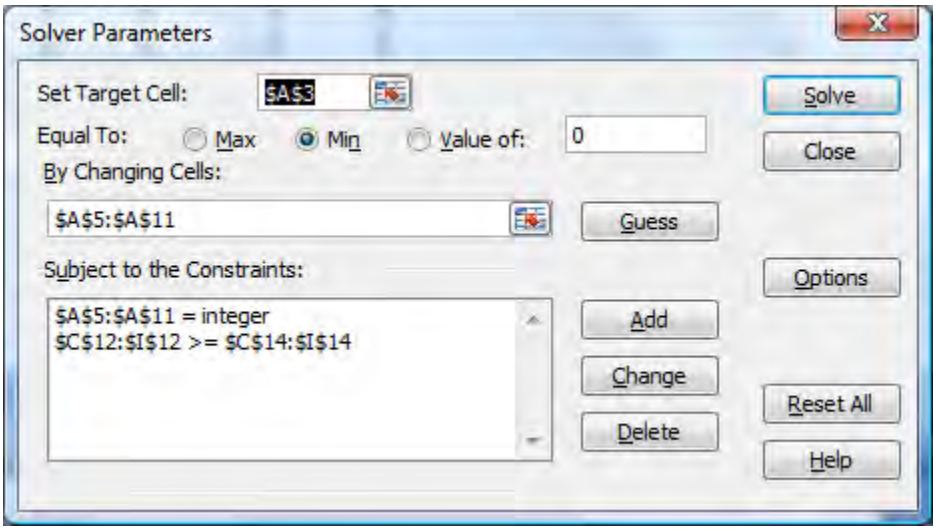
In this example, we are trying to staff a bank with a minimum number of employees by day. The employees can work five consecutive days, starting on any one day. How many employees are needed to work in each shift? Use the Bank Staffing spreadsheet.



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | Total | | | | There are multiple solutions all of which use 20 workers. | | | | |
| 3 | 20 | | Working? | | | | | | |
| 4 | Number starting | Day worker starts | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
| 5 | 1 | Monday | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 6 | 3 | Tuesday | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 7 | 0 | Wednesday | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 8 | 4 | Thursday | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 9 | 1 | Friday | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 10 | 2 | Saturday | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 11 | 9 | Sunday | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 12 | | Number working | 15 | 14 | 16 | 8 | 10 | 16 | 16 |
| 13 | | | >= | >= | >= | >= | >= | >= | >= |
| 14 | | Number needed | 17 | 13 | 15 | 17 | 9 | 9 | 12 |

Cells A5:A11 have an initial assignment of people to shifts. Cell A5=1 means that one person will start working on a Monday. The cells C5:I5 show the days worked for a “Monday” shift. To calculate how many people are working each day, the Number working is the sumproduct. Management has determined the minimum number of employees that are required per day, listed as number needed. Now, let’s find a solution. The steps are:

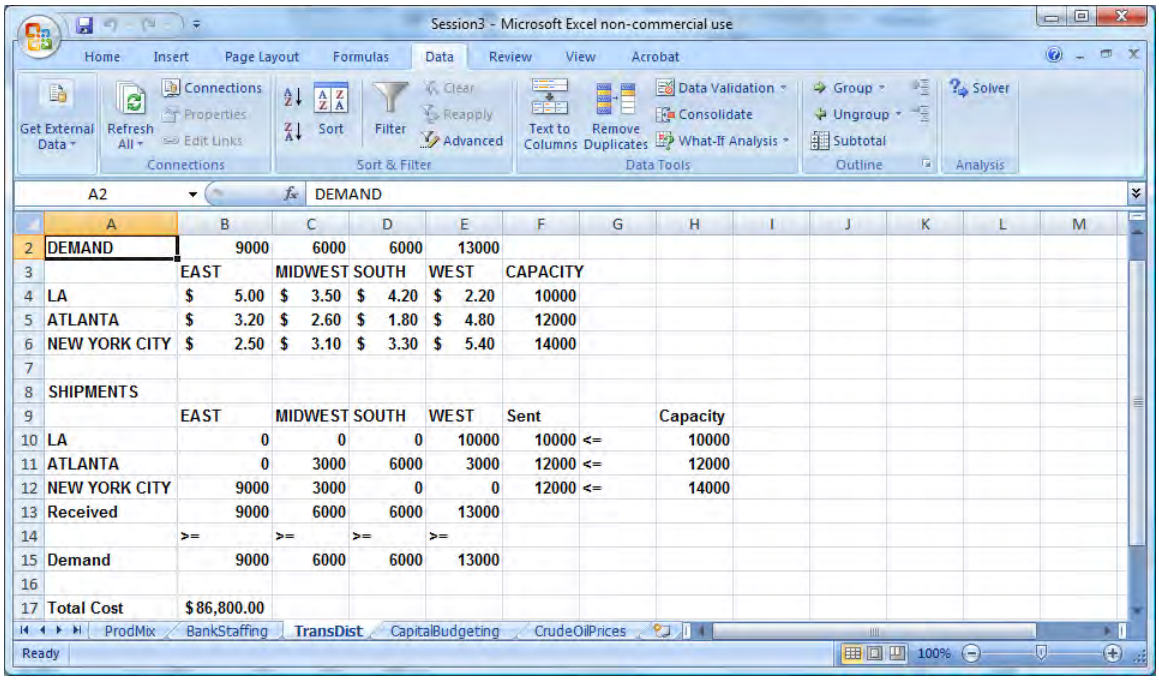
1. Target cell: minimize the number of employees
2. Changing cells: number of employees who start work for each shift
3. Constraints: number of employees working must be greater than or equal to the number of employees required.
4. Also, set a constraint that you should have whole people, i.e., people are integer.




5. Finally select the options Assume Linear Model and Assume Non-Negative for the changing cells.
6. Run the solver as before.

**Session 6.10: Transportation and distribution**

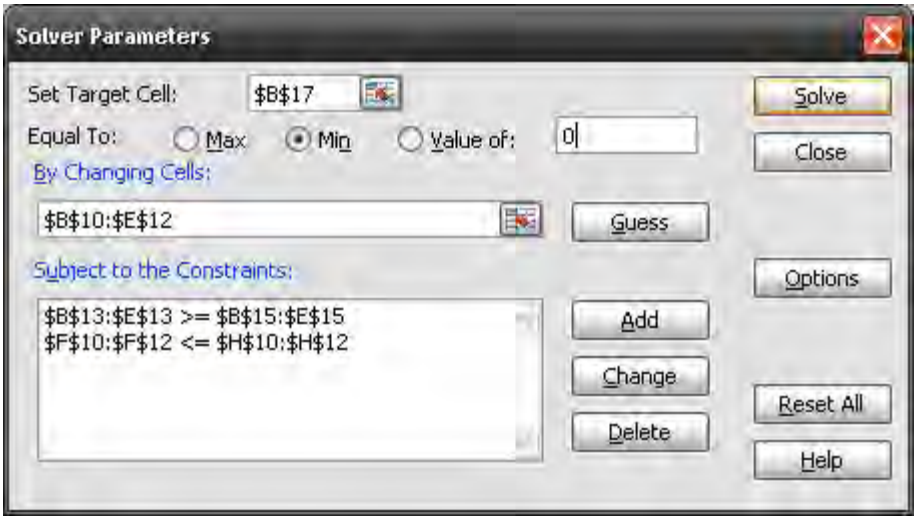
The transportation and distribution problem describes an opportunity to minimize transportation costs. There are three distribution centers with inventory to be shipped to four destinations. The goal is to minimize transportation costs but satisfy demand for the products at all locations. Use the TransDist spreadsheet. The initial spreadsheet looks like the picture below:



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 2 | DEMAND | 9000 | 6000 | 6000 | 13000 | | | |
| 3 | | EAST | MIDWEST | SOUTH | WEST | CAPACITY | | |
| 4 | LA | $ 5.00 | $ 3.50 | $ 4.20 | $ 2.20 | 10000 | | |
| 5 | ATLANTA | $ 3.20 | $ 2.60 | $ 1.80 | $ 4.80 | 12000 | | |
| 6 | NEW YORK CITY | $ 2.50 | $ 3.10 | $ 3.30 | $ 5.40 | 14000 | | |
| 7 | | | | | | | | |
| 8 | SHIPMENTS | | | | | | | |
| 9 | | EAST | MIDWEST | SOUTH | WEST | Sent | | Capacity |
| 10 | LA | 0 | 0 | 0 | 10000 | 10000 | <= | 10000 |
| 11 | ATLANTA | 0 | 3000 | 6000 | 3000 | 12000 | <= | 12000 |
| 12 | NEW YORK CITY | 9000 | 3000 | 0 | 0 | 12000 | <= | 14000 |
| 13 | Received | 9000 | 6000 | 6000 | 13000 | | | |
| 14 | | >= | >= | >= | >= | | | |
| 15 | Demand | 9000 | 6000 | 6000 | 13000 | | | |
| 16 | | | | | | | | |
| 17 | Total Cost | $86,800.00 | | | | | | |

The steps to minimize cost are:

1. Set the target cell to the total cost. See the formula in the cell. Why does this work?
2. Set the changing cells to amount produced by site which is shipped to each destination.
3. Set the constraints to reflect the capacity of the distribution center. Also, the values of the changing cell should be non-negative.

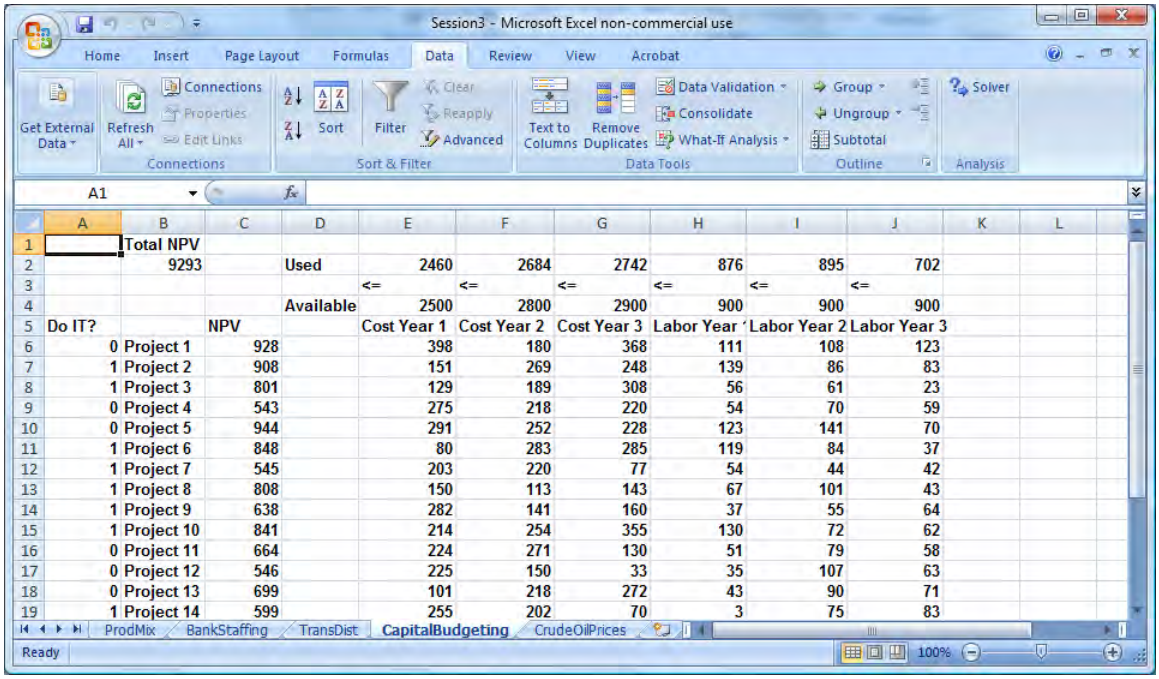



Work through this example, create the reports, and interpret the results.

**Session 6.11: Capital budgeting**

Using NPV (net present value) and IRR (internal rate of return), we are able to compare projects. Solver can be used for Capital Budgeting to select the optimal projects from a budgeting perspective.

In this case, we will use a trick called binary changing cells. A binary changing cell has a value of one or zero, reflecting that a project is selected or not. We will use a constraint of “bin” to designate a binary changing cell.



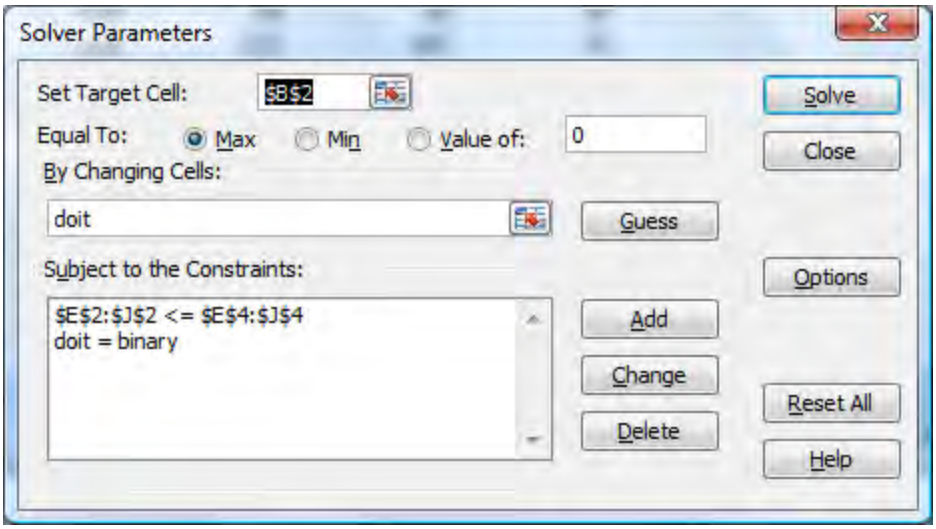
| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Total NPV | | | | | | | | |
| 2 | | 9293 | | Used | 2460 | 2684 | 2742 | 876 | 895 | 702 |
| 3 | | | | | <= | <= | <= | <= | <= | <= |
| 4 | | | | Available | 2500 | 2800 | 2900 | 900 | 900 | 900 |
| 5 | Do IT? | | NPV | | Cost Year 1 | Cost Year 2 | Cost Year 3 | Labor Year ´ | Labor Year 2 | Labor Year 3 |
| 6 | 0 | Project 1 | 928 | | 398 | 180 | 368 | 111 | 108 | 123 |
| 7 | 1 | Project 2 | 908 | | 151 | 269 | 248 | 139 | 86 | 83 |
| 8 | 1 | Project 3 | 801 | | 129 | 189 | 308 | 56 | 61 | 23 |
| 9 | 0 | Project 4 | 543 | | 275 | 218 | 220 | 54 | 70 | 59 |
| 10 | 0 | Project 5 | 944 | | 291 | 252 | 228 | 123 | 141 | 70 |
| 11 | 1 | Project 6 | 848 | | 80 | 283 | 285 | 119 | 84 | 37 |
| 12 | 1 | Project 7 | 545 | | 203 | 220 | 77 | 54 | 44 | 42 |
| 13 | 1 | Project 8 | 808 | | 150 | 113 | 143 | 67 | 101 | 43 |
| 14 | 1 | Project 9 | 638 | | 282 | 141 | 160 | 37 | 55 | 64 |
| 15 | 1 | Project 10 | 841 | | 214 | 254 | 355 | 130 | 72 | 62 |
| 16 | 0 | Project 11 | 664 | | 224 | 271 | 130 | 51 | 79 | 58 |
| 17 | 0 | Project 12 | 546 | | 225 | 150 | 33 | 35 | 107 | 63 |
| 18 | 0 | Project 13 | 699 | | 101 | 218 | 272 | 43 | 90 | 71 |
| 19 | 1 | Project 14 | 599 | | 255 | 202 | 70 | 3 | 75 | 83 |

In the spreadsheet above, there are 20 projects, each with an NPV, yearly cost and yearly labor requirement. However, there also is a limit to available funds and personnel, listed in the Available row. Your objective is to select the projects within the funding and personnel constraints which maximize total NPV.

The following steps maximize NPV:

1. Set target cell to Total NPV value.
2. Set changing cells to the range of “Do it” zeroes and ones.
3. Set constraints to reflect that funds and employees used cannot exceed available.

The set up picture follows. Now run solver to see the result.




How would you handle other constraints, such as, if you select project 4, then project 3 must also be selected? If you only have four project managers, how would you limit the selection to four projects?

View All View Keyframes

## Overview of R

- Over the past decade, companies have been using a variety of statistical packages, including SPSS, SAS, Minitab, Stata, and others.
- Recently, companies have begun to migrate to a powerful statistical and data mining package called **R**.
- R is an open source software.
    - Free
    - Cost of other packages has driven new users to R.
- R is a language, but we will mostly operate it using a graphical user interface.

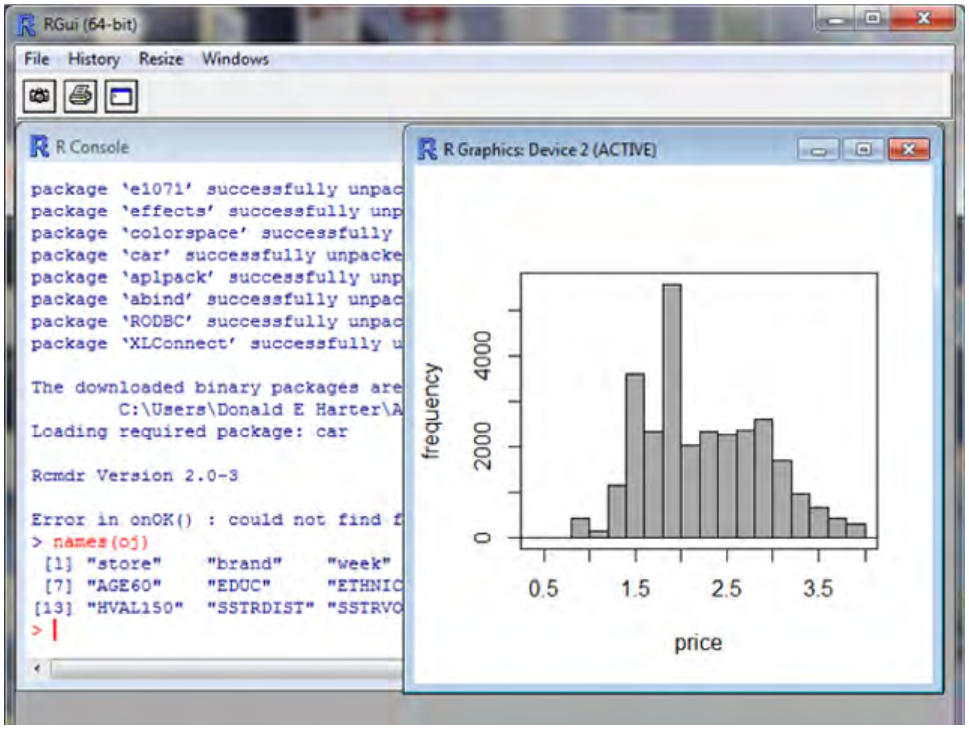
7 of 7

View All View Keyframes

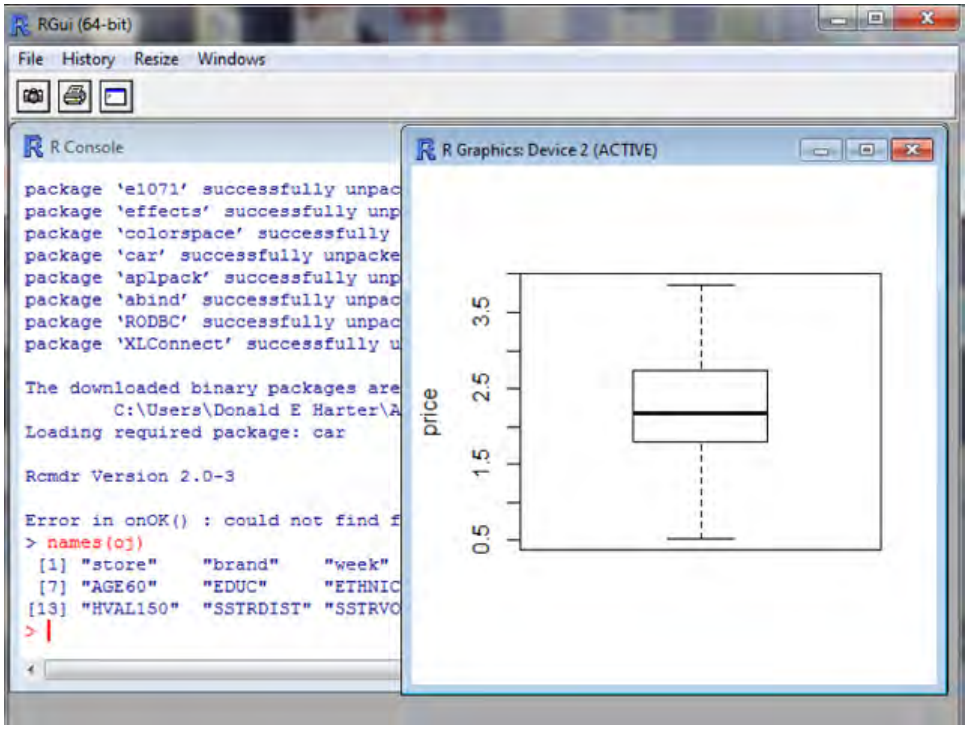
## Visualization

- R has the ability to generate histograms, boxplots, scatterplots, and XY plots similar to Excel.
- Graphs from R can differentiate by brand, product category, or other variable.
  - Gives greater insight into patterns

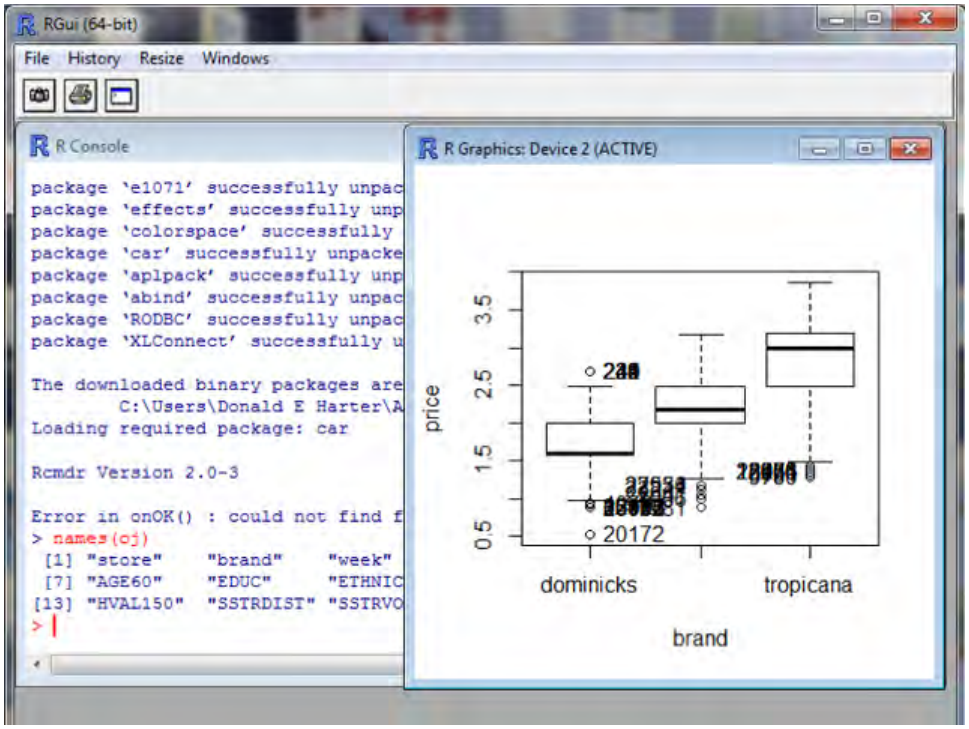
4 of 39




5 of 39



10 of 39

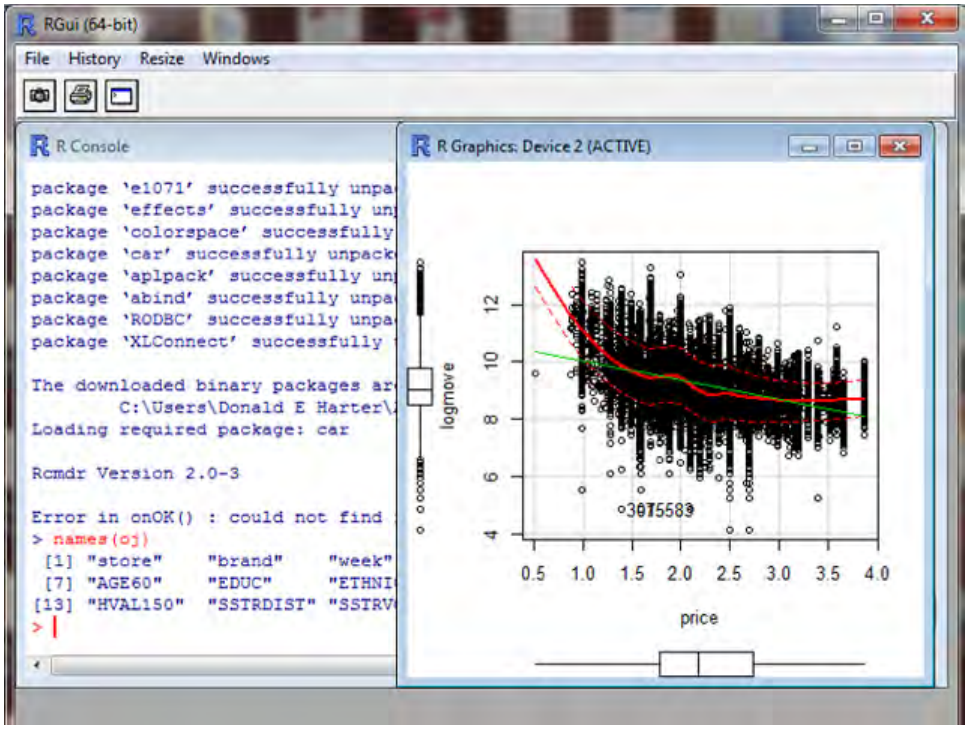



19 of 39

## Histograms and Boxplots

- **Histograms** show frequency of data within intervals or bins
  - Easy to develop in Excel or R.
- **Boxplots** identify:
  - Maximum and minimum—whiskers
  - 25%-ile and 75%-ile—box
  - Median—center bar
- Boxplots in R can also differentiate by brands.

23 of 39

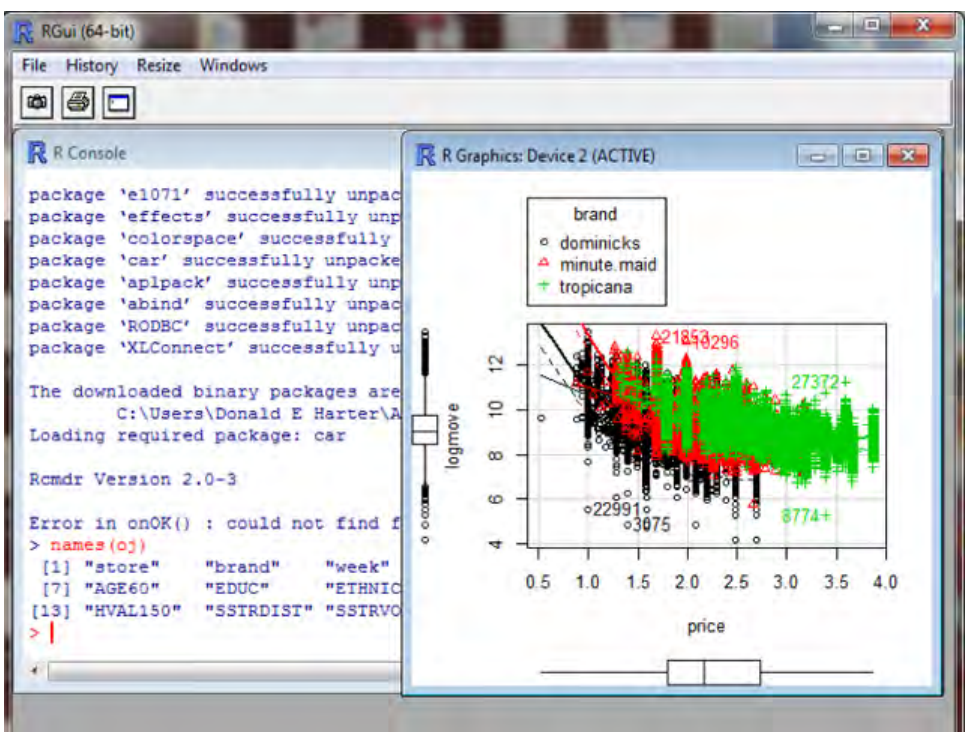



24 of 39

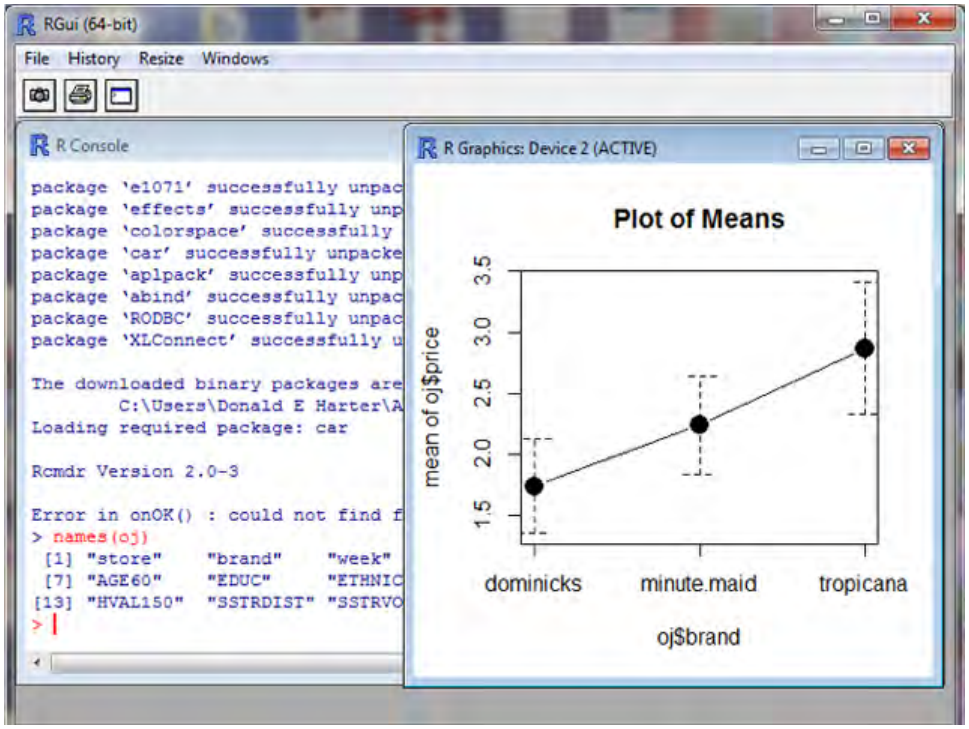
## Scatterplots

- **Scatterplots** offer not only trend line, but price by price.
- From example:
  - Solid red line—average sales by price
  - Green line—trend line
  - Dotted red line—standard deviation
- Scatterplots of log sales by price can also be categorized by brand.

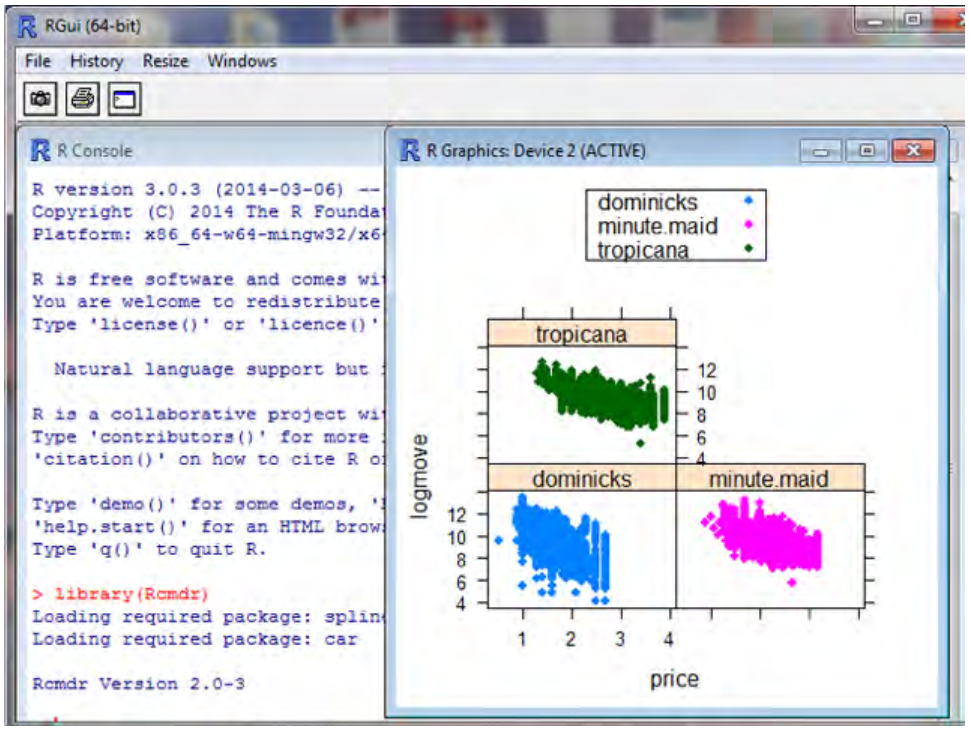
30 of 39



31 of 39




36 of 39



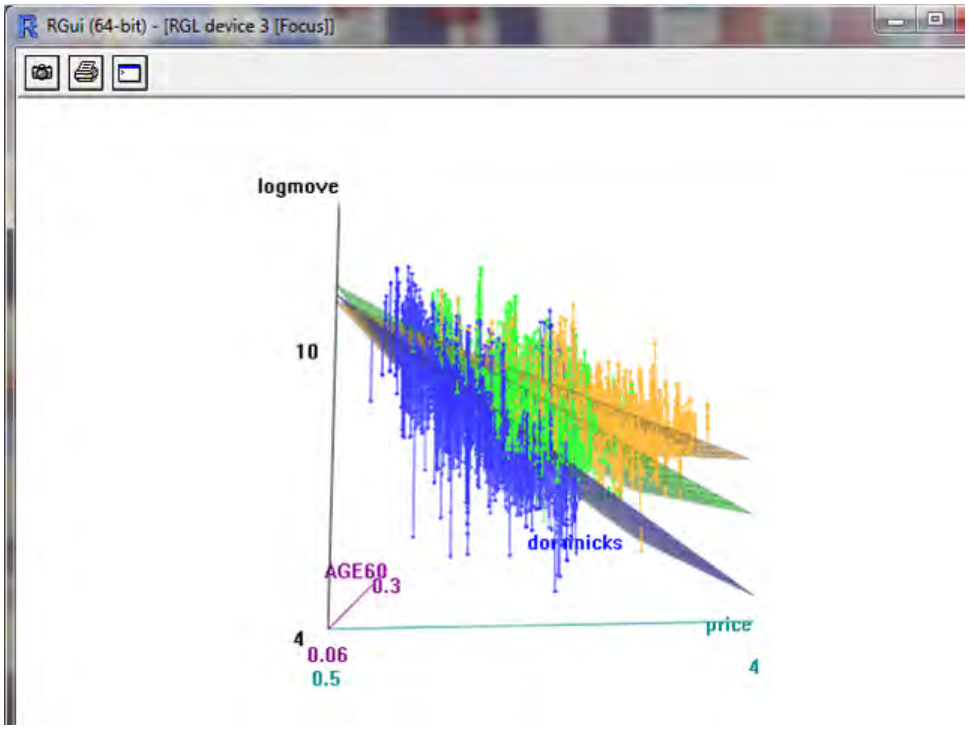

38 of 39

## Plot of Means and XY Plots

- **Plot of means** is where the average price by brand is plotted.
- **XY plots** display the effect of price on sales per brand by separating and placing plots next to each other.

39 of 39

View All View Keyframes



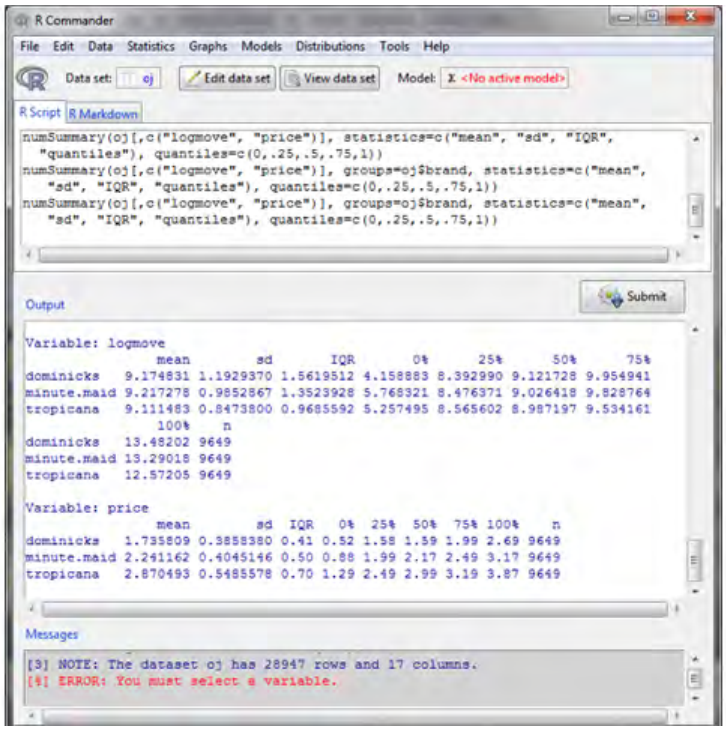

5 of 13

## 3-D Visualization

- R can perform more sophisticated 3-D graphing than Excel.
  - Plot the planes—helps find patterns
  - Rotate the graphs—offers different senses of parameters.
- 3-D allows you to see regression lines and planes to identify patterns.

13 of 13

View All View Keyframes




4 of 13

## Statistical Summaries

- Similar to Excel, R can calculate descriptive statistics.
- R has the additional ability to calculate descriptive statistics by brand.
- Stratify and characterize data while performing analysis.

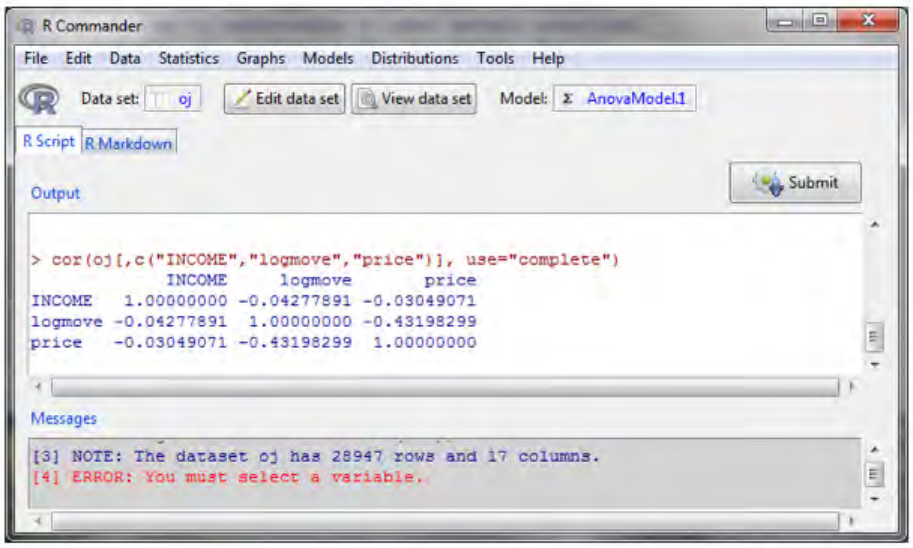
13 of 13

View All View Keyframes

## Correlations

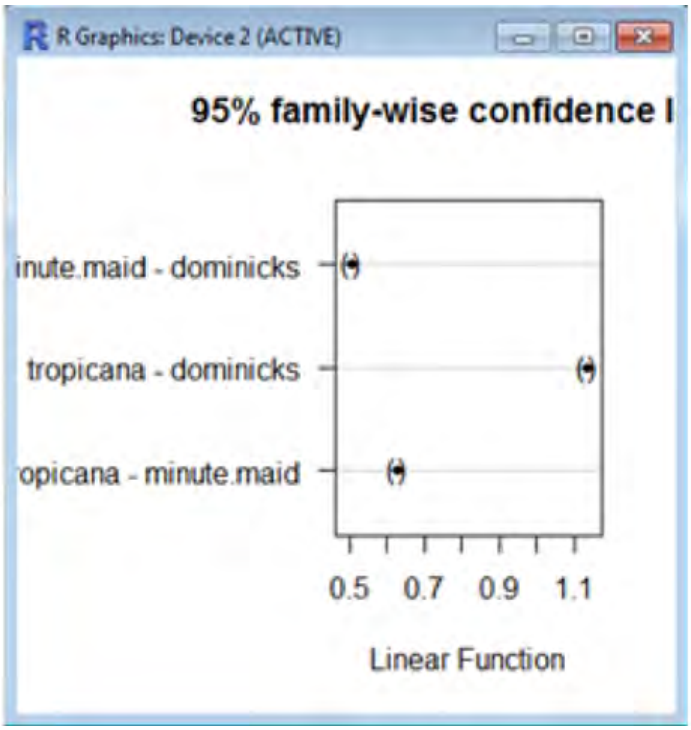
- Measure how two variables are related
- Positively correlated: one variable increases, the other one also increases
- Negatively correlated: one variable increases, the other one decreases

4 of 11




5 of 11

View All View Keyframes



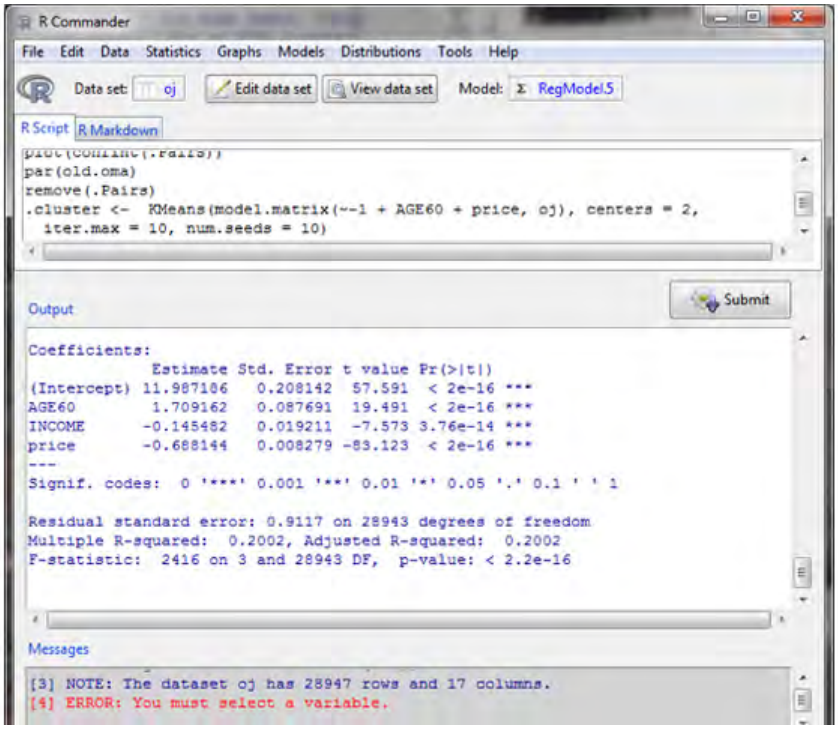

8 of 11

## Analysis of Variance

- Store managers want to compare product placement positions, sales, and prices of different products or brands.
- Analysis of variance (ANOVA) is used to compare the averages for different categories or brands.
- ANOVA calculates the average of each item, combines with standard deviation/variance, and determines if they are statistically different from one another.
- ANOVA is a valuable technique for product placement.

11 of 11

View All View Keyframes



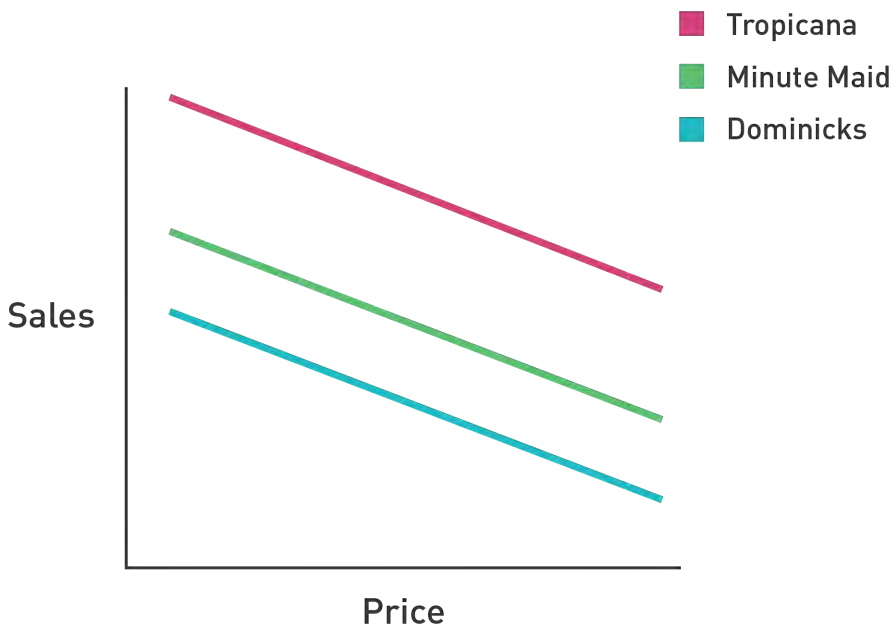

4 of 9

## Regression

- R can perform linear regression.
  - Output is similar to Excel's multivariate regression output
- Analysis produces coefficients, statistical significance, and R-squared values.
  - Sets foundation to view more complex regressions

9 of 9

View All View Keyframes



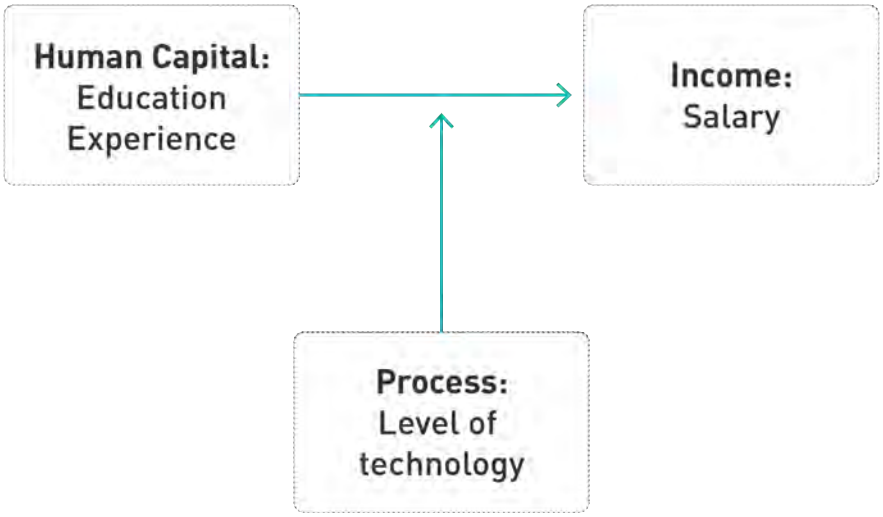

5 of 10

## Regression with Dummy Variables

- Dummy variables change the intercept of some items to be different from others.
  - E.g., one brand will always be higher in sales than another
  - Reflected in different intercepts or heights of the regression line
- This allows you to look at price differentiation in different product markets.

10 of 10

View All View Keyframes

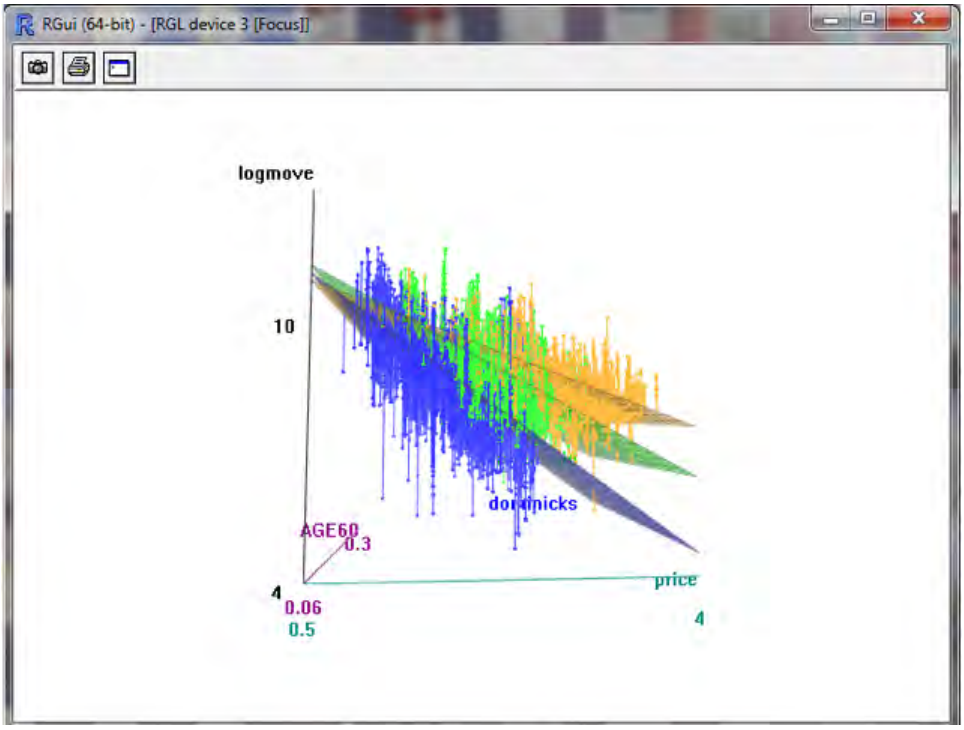



10 of 19

## Regression with Moderating Effects

- Dummy variables can only change the intercepts.
- Moderating effects allow the slope of the line to change.
  - Recognizes that two variables might interact, magnifying their effects
- Education and experience both have a positive effect on income.
  - When an experienced, educated professional learns technology, the technology is the moderating effect.
- **Moderating effect:** An interaction of variables that leverage each other
- The moderating effect acts as a catalyst to accelerate the effect of certain variables.

18 of 19


RGui (64-bit) - [RGL device 3 [Focus]]
logmove
10
dominicks
AGE60
0.3
4
0.06
0.5
price
4

19 of 19

# R: Basics

**Session 7.3: Installation of R**

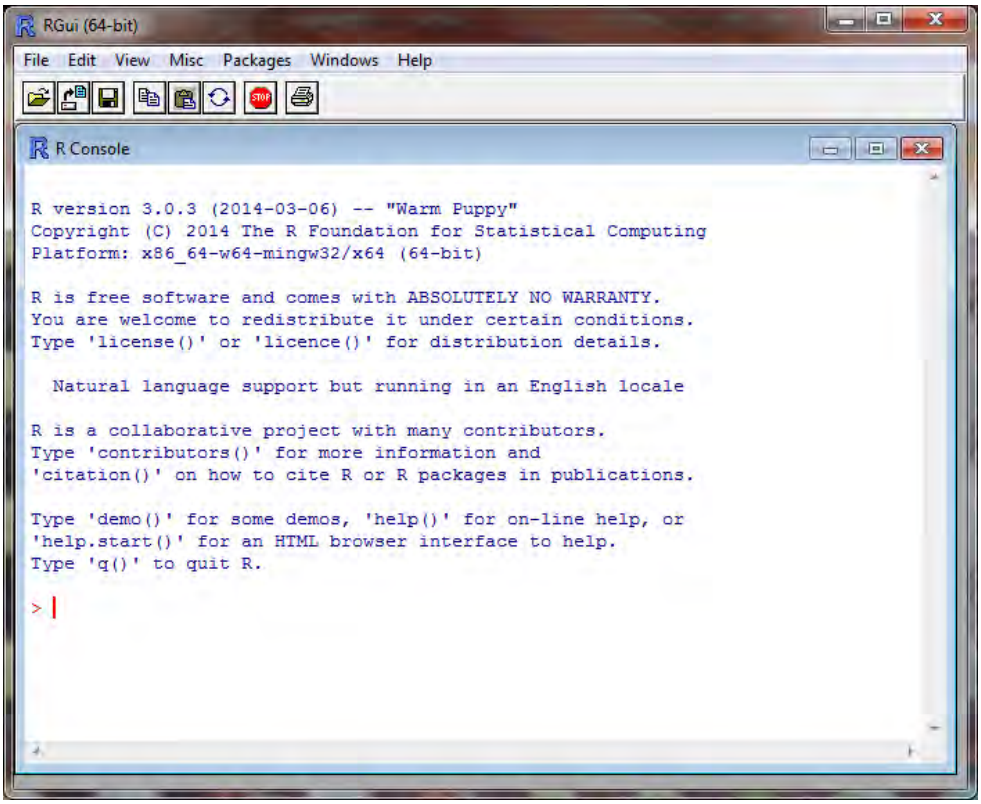
For these exercises, download the files:
"Business Analytics – Week 7 Instructions.doc"
"Business Analytics – Week 7 oj.xls"

R is a free downloadable package capable of performing sophisticated statistical analysis and data mining. The software is already installed on the classroom laptops. To install on your own personal computer:

1. go to the website: https://cran.r-project.org/
    a. Windows: http://cran.r-project.org/bin/windows/base/
    b. Mac: https://cran.r-project.org/bin/macosx/
2. Click on Download R 3.2.2 (note: the version changes frequently)
3. Click on Run, and follow the install instructions

**Starting R**

1. Click on the Start button in the lower left corner of Windows
2. Click on All Programs, then click on the R folder, then R

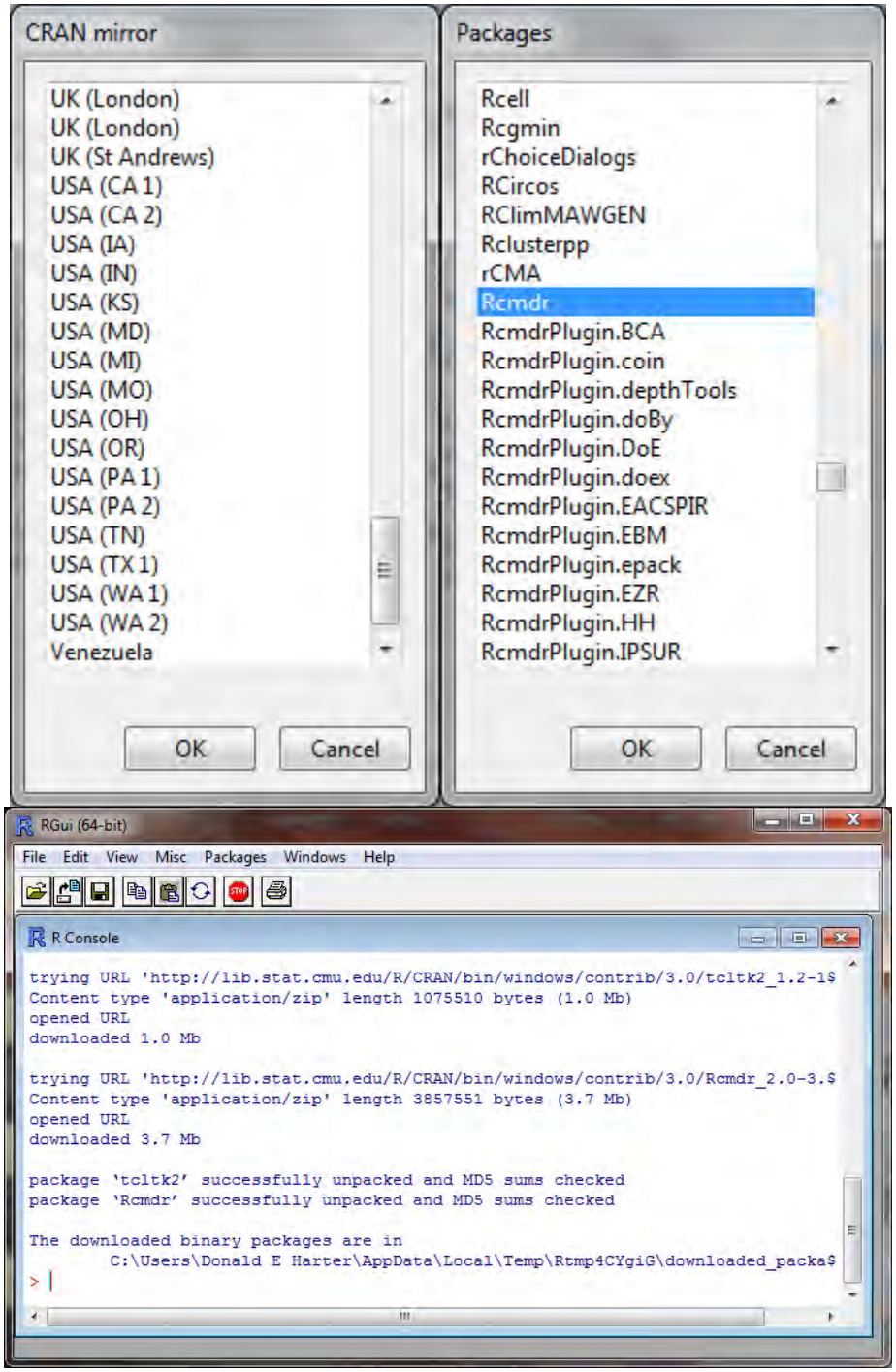



This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

**Installing R Commander**

Follow these steps only if you don't already have Rcmdr installed.

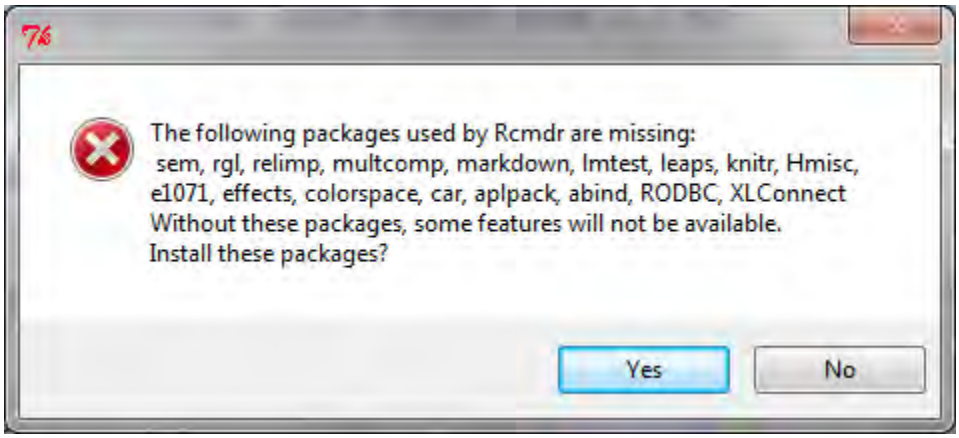
1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; then click OK
4. In the Packages screen, click on Rcmdr, then OK
5. When prompted to create a personal library, click Yes

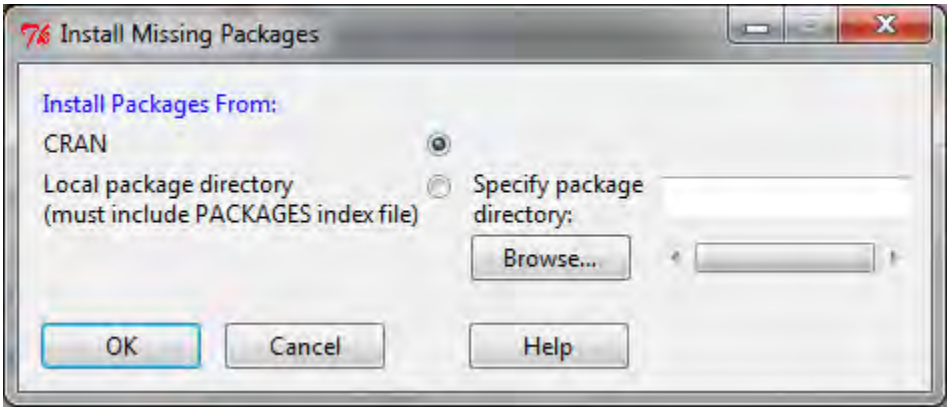


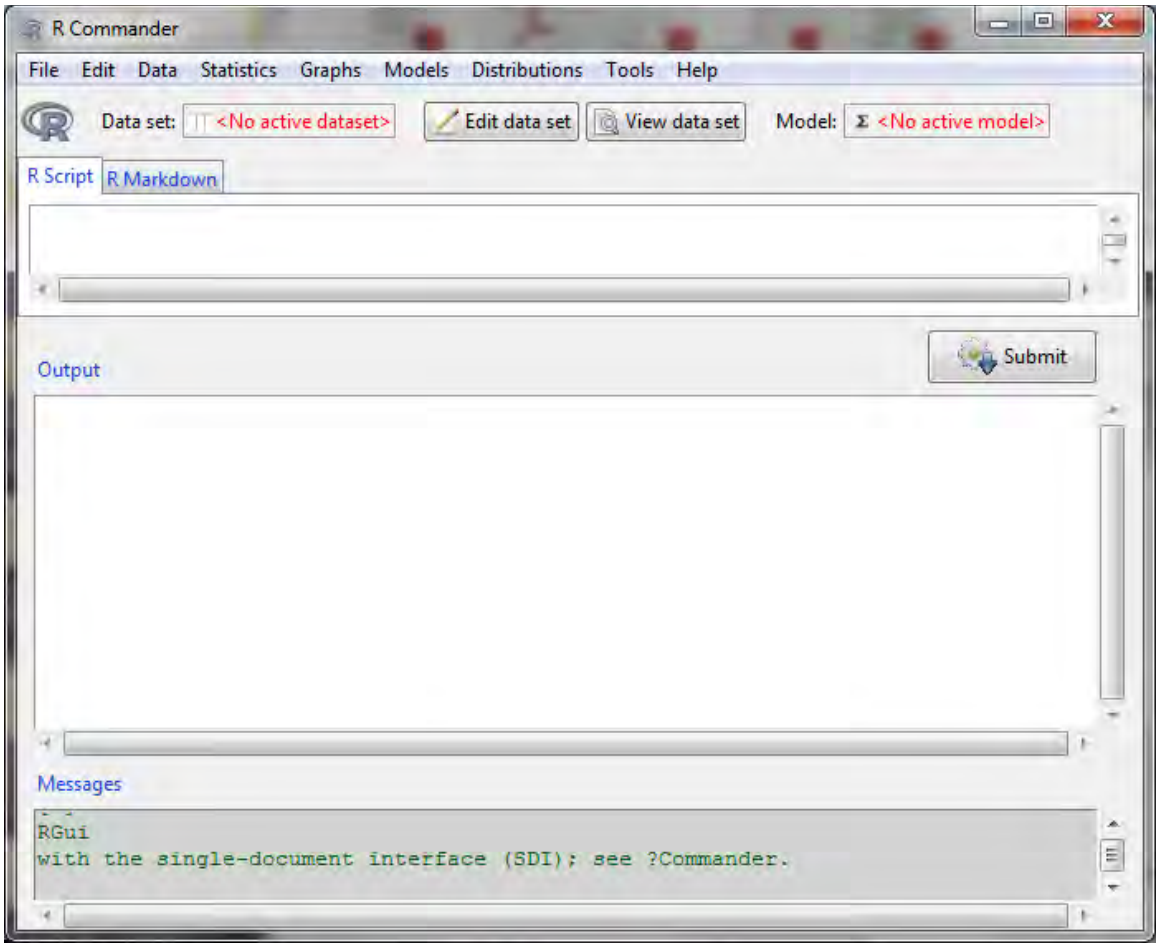
**Launch Rcmdr (R Commander)**

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type library(Rcmdr)
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK
4. R will install the necessary software
5. The R Commander screen will appear






**Session 7.4: Download Datasets**

To access some excellent data sets used in the book "Data Mining and Business Analytics with R," by Johannes Ledolter:

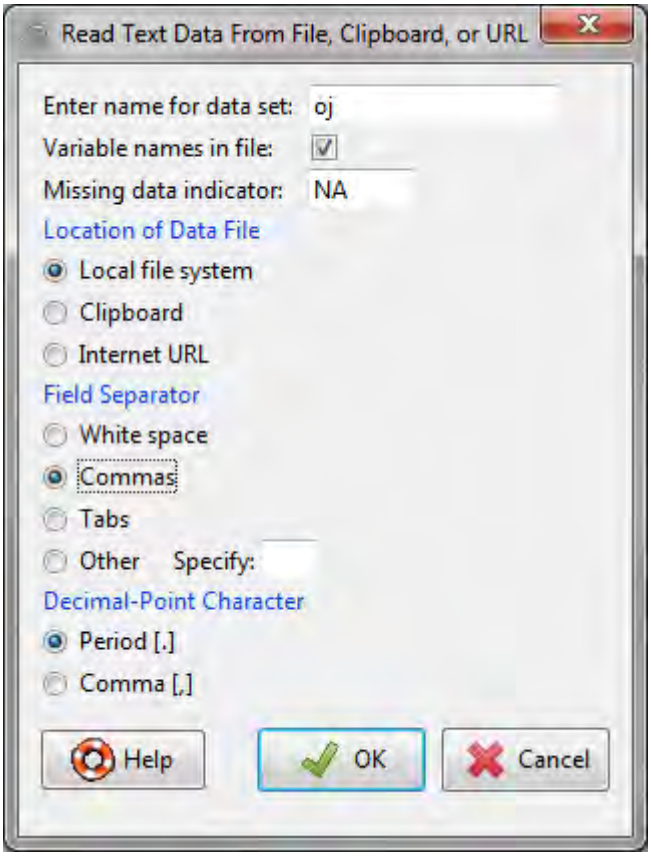
1. Go to the website:
   http://www.biz.uiowa.edu/faculty/jledolter/DataMining
2. Click on Data Text
3. Right click on oj.csv, then save on your computer
4. Remember where you saved the file

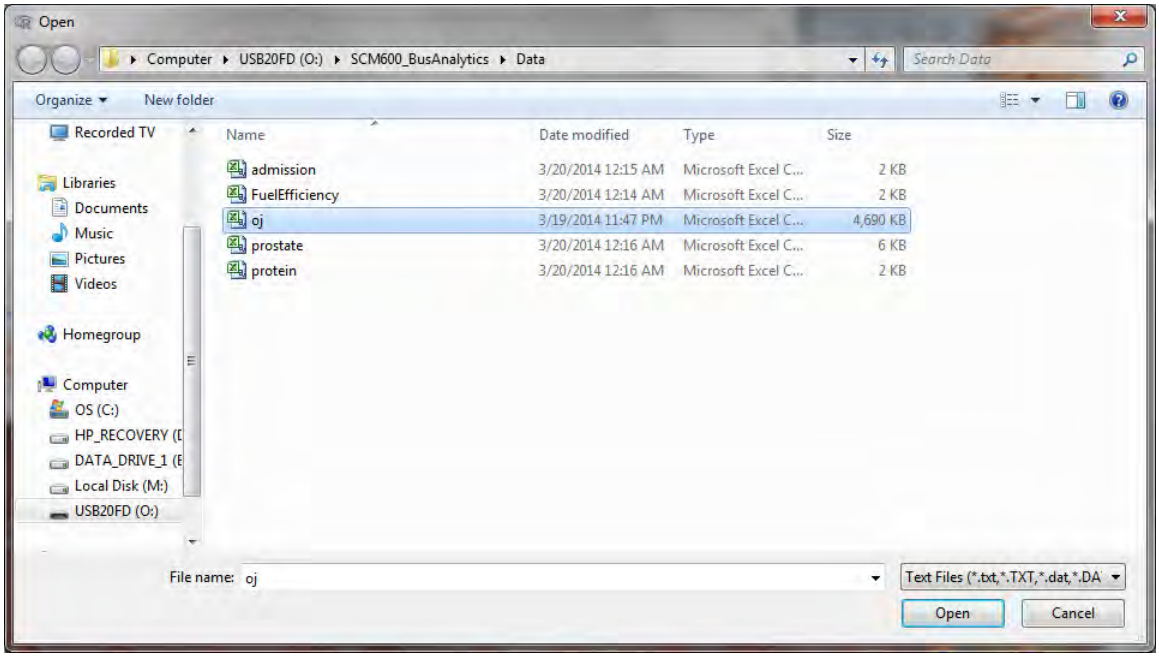
The Business Analytics - Week 7 oj.csv (orange juice) file can be downloaded from the course website.

**Loading Data**

To load data into R:

1. Click on Data at the top of the screen
2. Click on Import Data > From text file …
3. Enter the name that you would like to use for this data set; type in oj
4. Change Field Separator to Commas, then OK
5. Click on the oj file, then Open

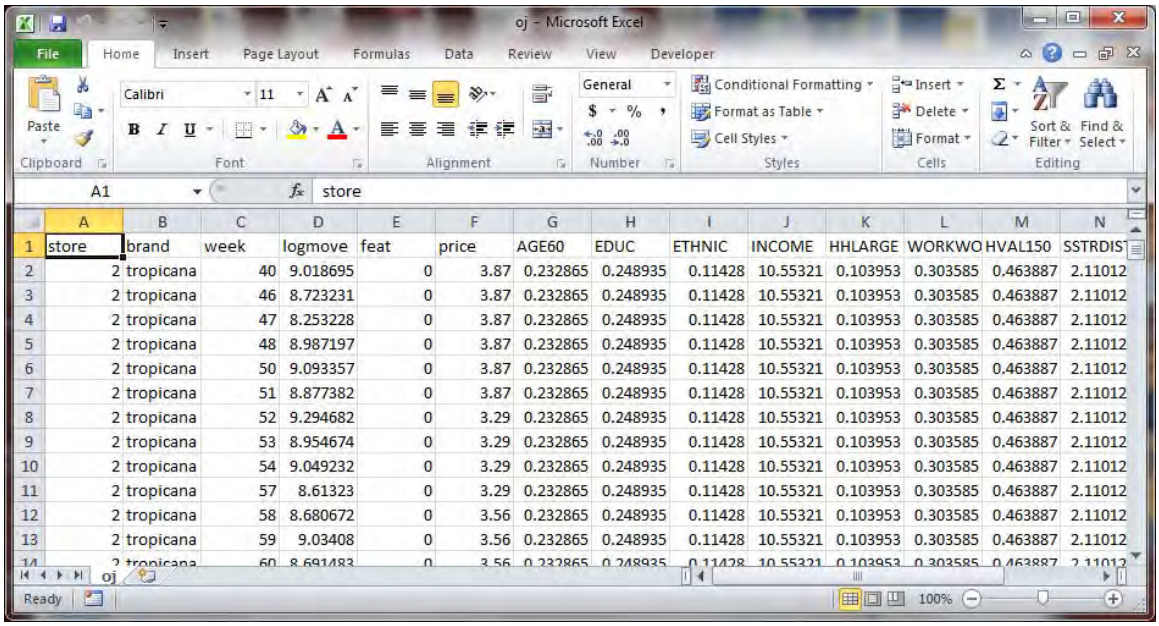



Note that the dataset oj has 28,947 rows and 17 columns.

### Viewing data fields

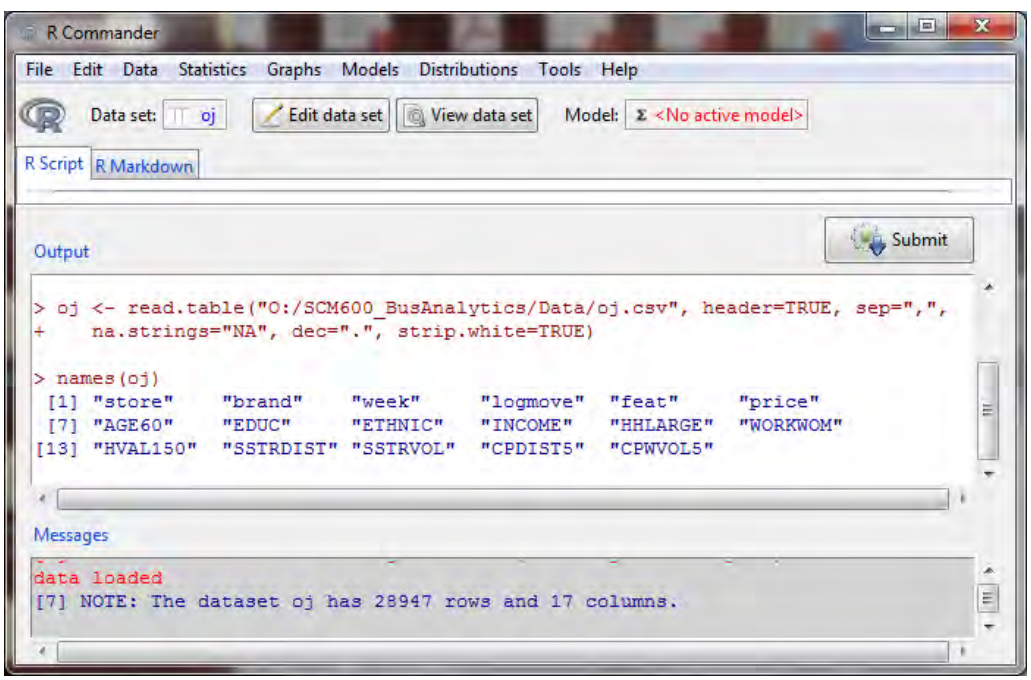
This data set lists weekly sales over 83 stores for three brands of products. Let's view the data. The easiest way to view is simply by opening the original Excel spreadsheet. Find the spreadsheet oj.csv that you downloaded and double click on it. The variable logmove is the logarithm of sales (how much product moved in a week).



| store | brand | week | logmove | feat | price | AGE60 | EDUC | ETHNIC | INCOME | HHLARGE | WORKWO | HVAL150 | SSTRDIS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | tropicana | 40 | 9.018695 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 46 | 8.723231 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 47 | 8.253228 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 48 | 8.987197 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 50 | 9.093357 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 51 | 8.877382 | 0 | 3.87 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 52 | 9.294682 | 0 | 3.29 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 53 | 8.954674 | 0 | 3.29 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 54 | 9.049232 | 0 | 3.29 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 57 | 8.61323 | 0 | 3.29 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 58 | 8.680672 | 0 | 3.56 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |
| 2 | tropicana | 59 | 9.03408 | 0 | 3.56 | 0.232865 | 0.248935 | 0.11428 | 10.55321 | 0.103953 | 0.303585 | 0.463887 | 2.11012 |

Now return to R. To view the variables in R,

1. Click on Data, Active Data Set, Variables in Active Data Set

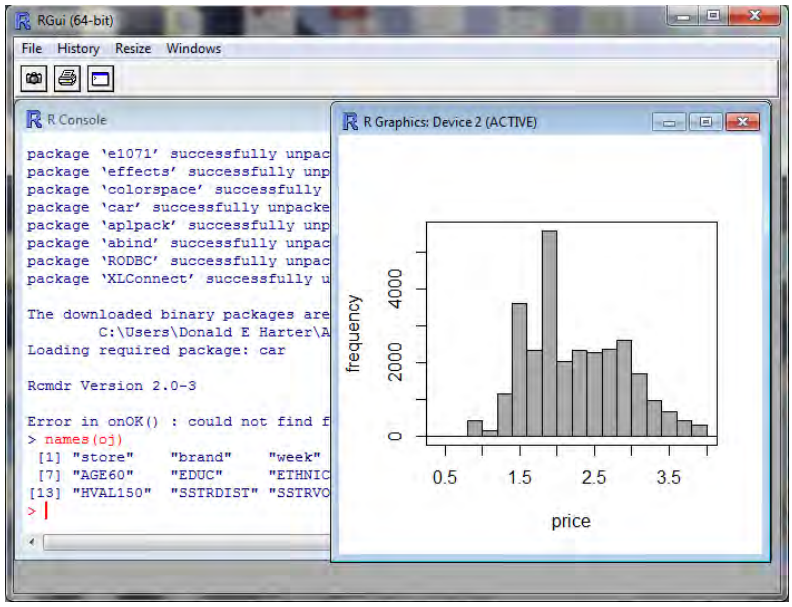



Notice that R generates the command names(oj). This is the command line version.

**Session 7.5: Histograms**

To create a histogram,

1. Click on Graphs, Histogram, price, OK

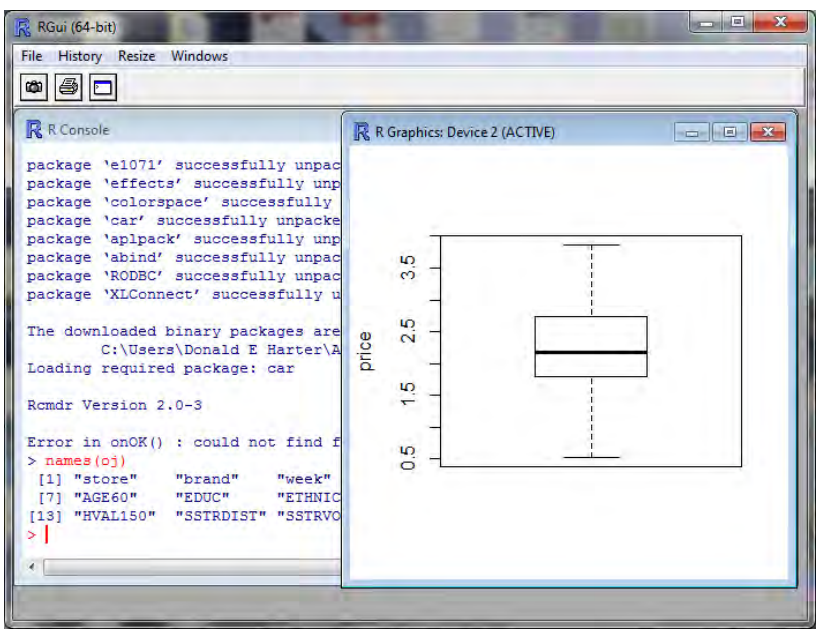



Try it again, but in the Options tab, change to Percentages (click Apply), Densities, and change number of bins to 5.

**Boxplots**

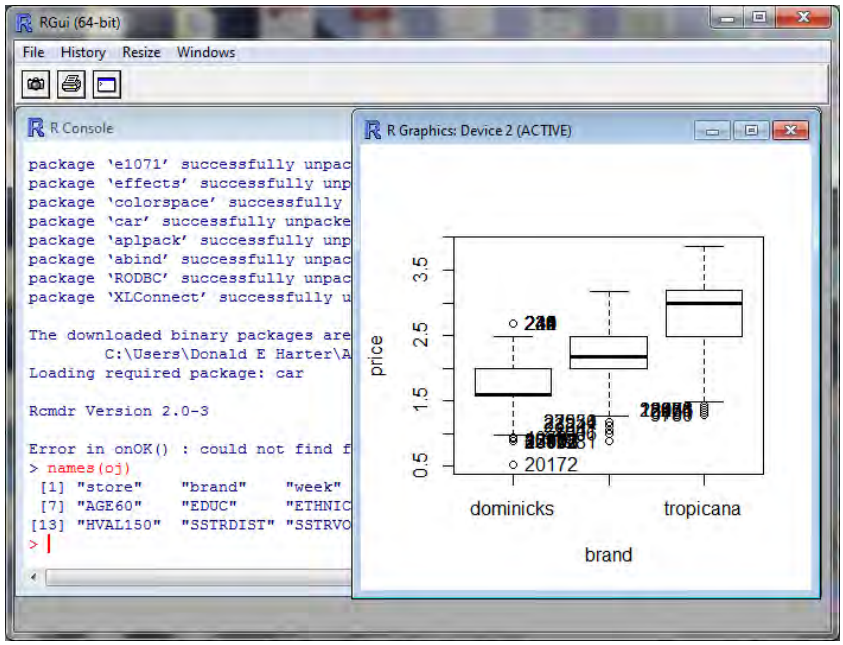
To create a boxplot,

1. Click on Graphs, Boxplot, price, OK



To create a boxplot by brand,

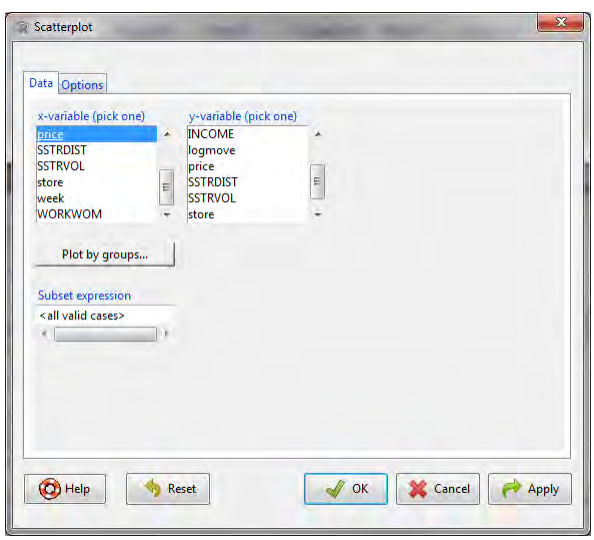
1. Click on Graphs, Boxplot, price
2. Click on Plot by groups, select brand, OK
3. Click OK




**Scatterplots**

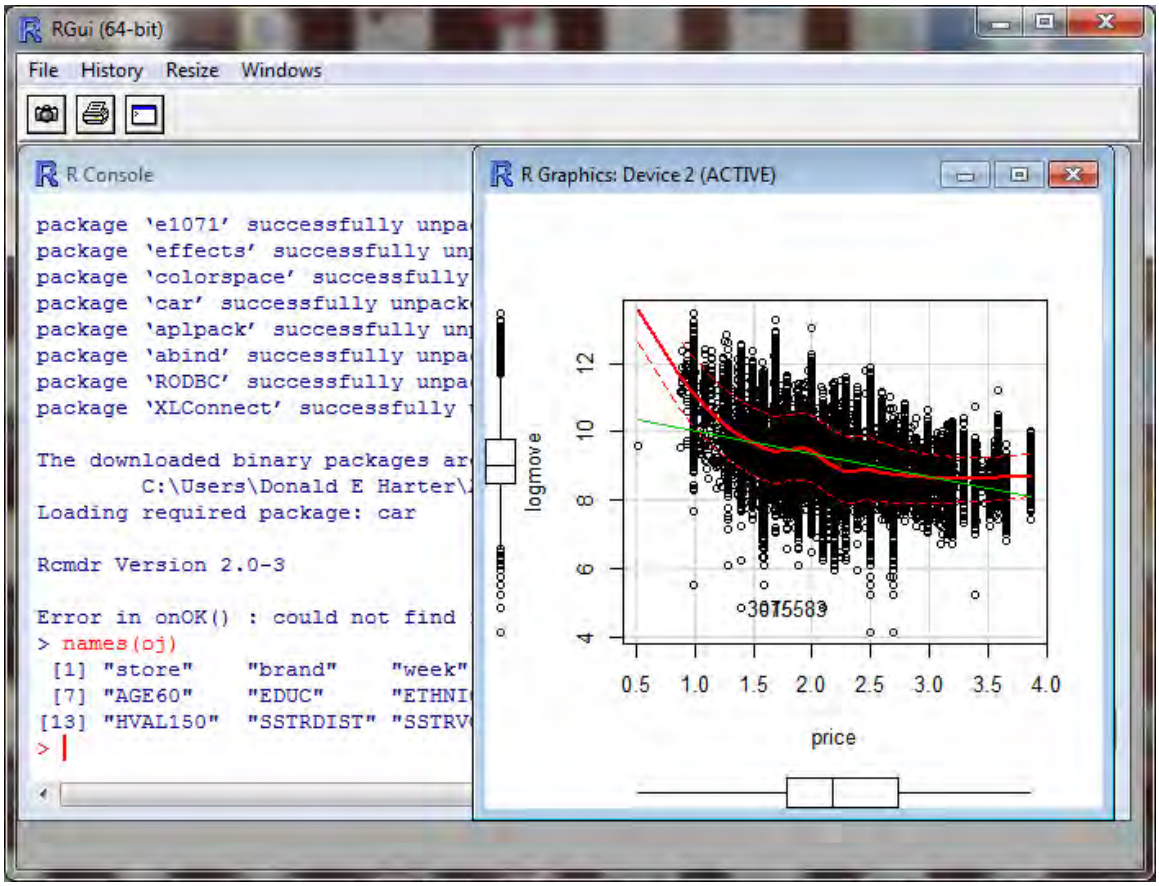
To generate a scatter plot,

1. Click on Graphs, Scatterplot
2. Select price as the x-variable
3. Select logmove as the y-variable
4. Click on OK



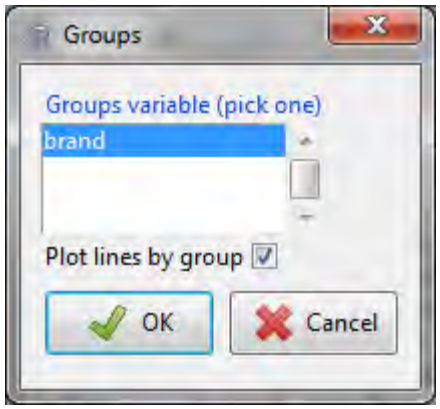
To interpret the chart:

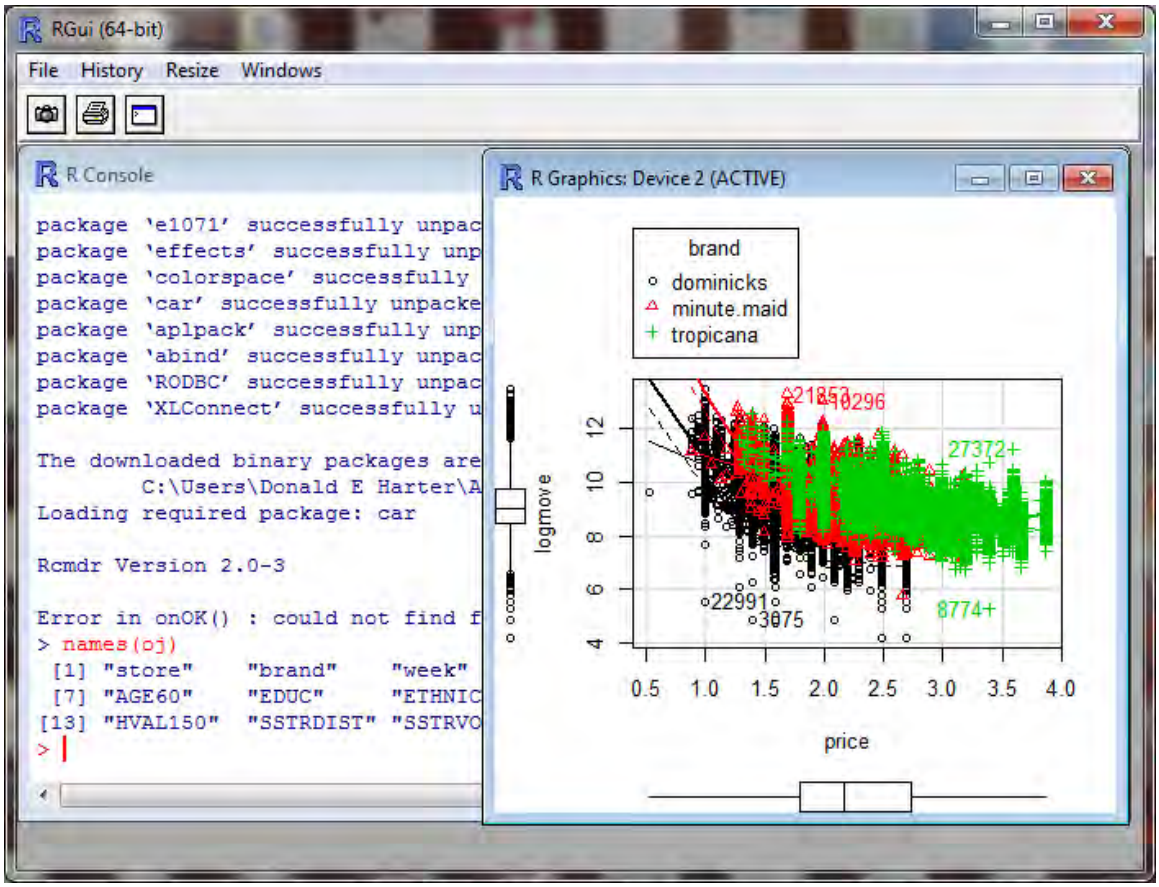
1. The black dots are the price versus log(sales) for each time period, store and brand.
2. The green line is the linear regression line through the data
3. The red lines are the averages and plus or minus one standard deviation
4. Below and to the left of the chart are box and whisker diagrams
    a. The center line is the average
    b. The box is the 25%-ile to 75%-ile range
    c. The whiskers show the range




Now generate a scatter plot by brand,

1. Click on Graphs, Scatterplot
2. Select price as the x-variable
3. Select logmove as the y-variable
4. Click on Plot by Groups, select brand, then OK.
5. Click on OK




Interpretation

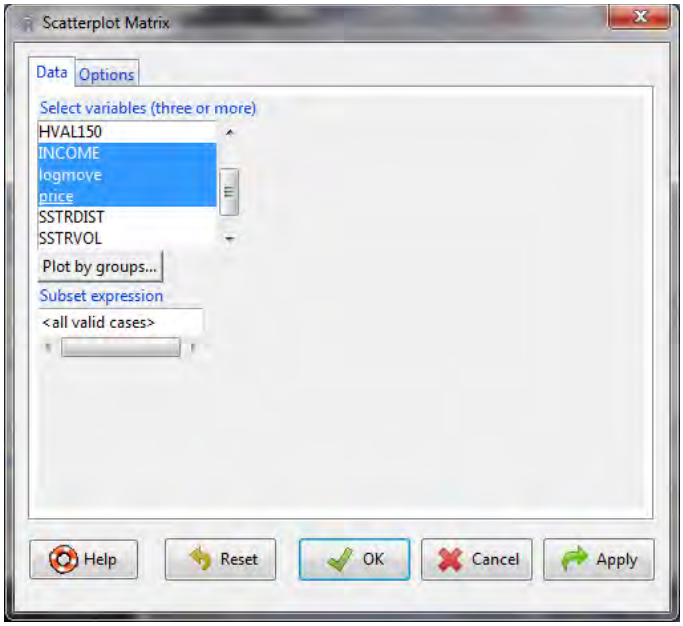
1. Brand dominicks is in black
2. Brand minute maid is in red
3. Brand Tropicana is in green

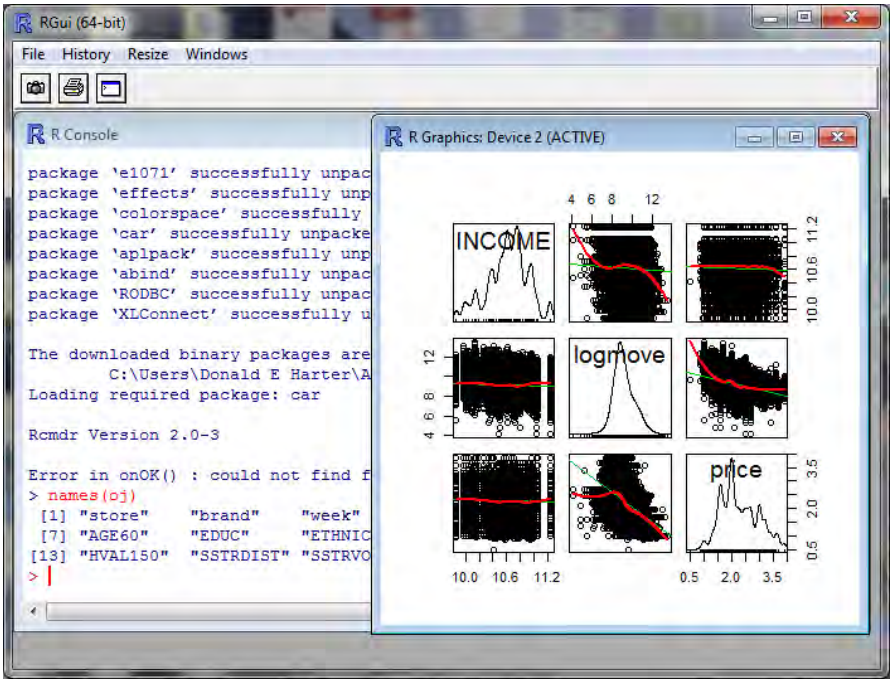
Which is the premium brand?

**Plotting pairwise scatterplots with more than two variables**

To create a matrix of scatterplots with more than two variables,

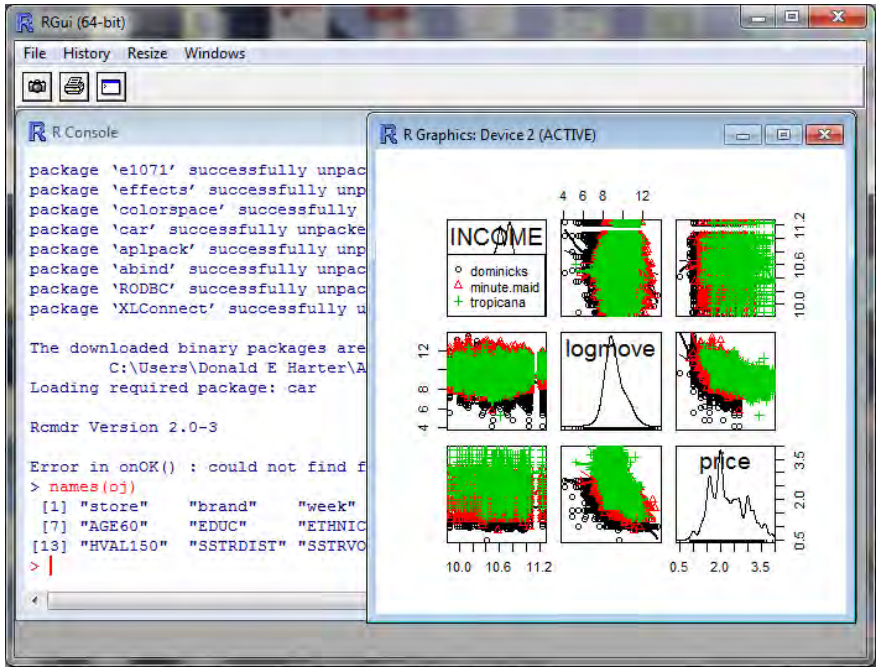
1. Click on Graphs
2. Click on Scatterplot Matrix
3. To select multiple variables, hold down the control key, then select INCOME, logmove, and price
4. Click OK




The diagonal is the distribution of data points (density function). Off-diagonal are the scatterplots for the pair of variables listed to the side and above/below the scatterplot.

Now perform a Scatterplot Matrix by Groups (brand)

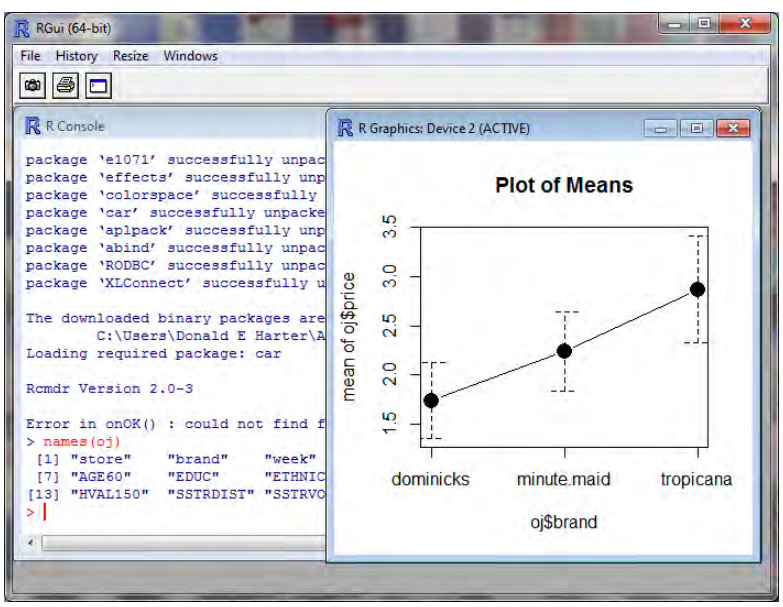



Once again, the brands are color coded.

**Plot of Means**

To determine if the different brands have different prices, on average, plot the means:

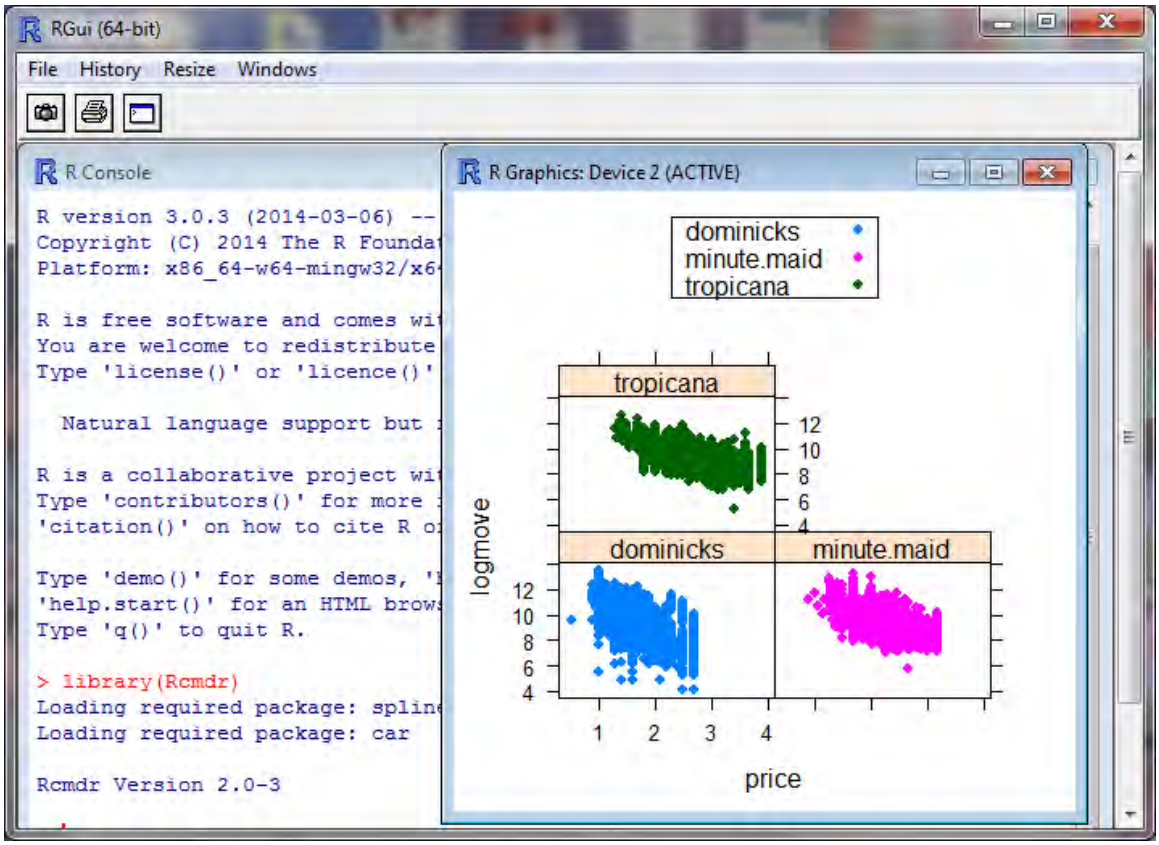
1. Click on Graphs, Plot of Means
2. Select price
3. In the Options tab, click on standard deviations, then OK



**XY Plots**

Now let's generate XY plots by brand.

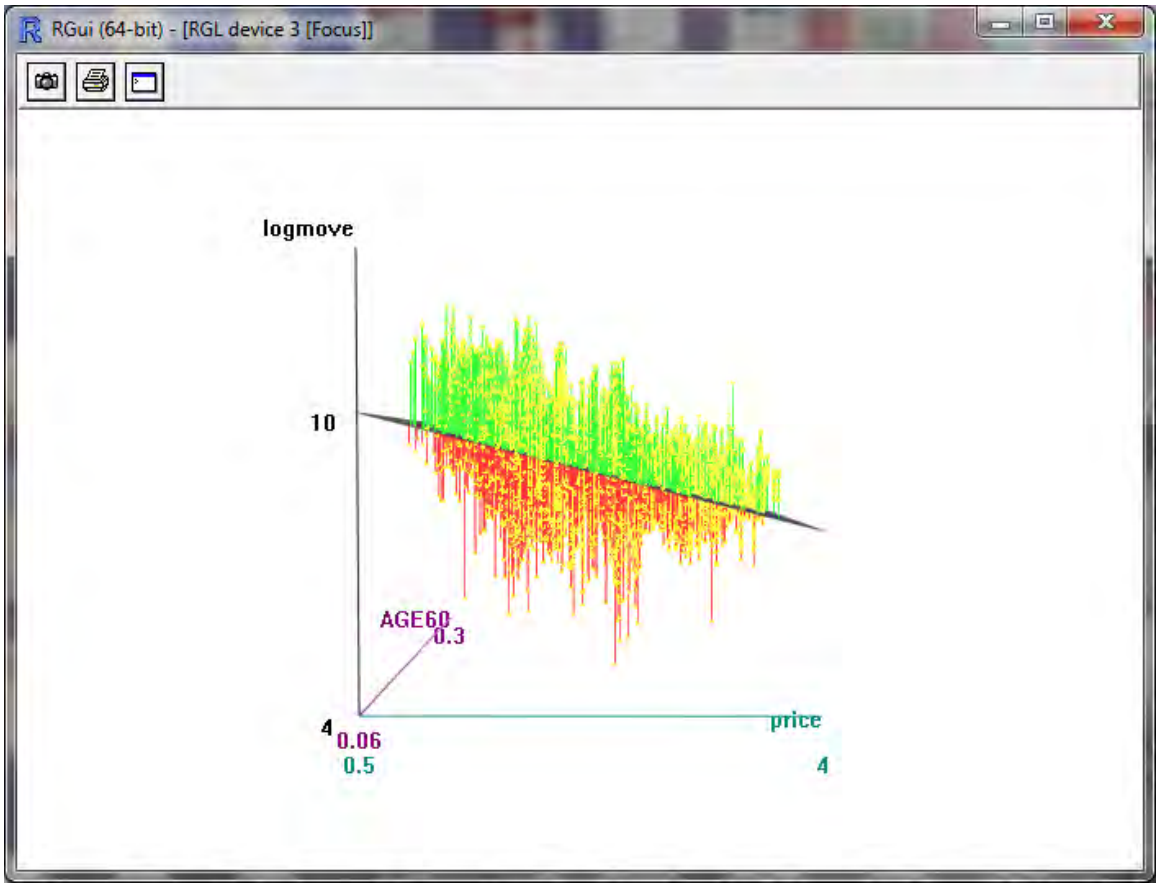
1. Click on Graphs, XY Conditioning Plot
2. Select price for the explanatory variable
3. Select logmove for the response variable
4. Click on brand for each
5. Click OK



**Session 7.6: 3D Graphs**

To generate 3D graphs,

1. Graphs, 3D Graph, 3D Scatterplot
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price
3. Select logmove (log of sales) as the response variable
4. Click OK
5. Expand the window by clicking on the box in the upper right corner of your graph
6. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move

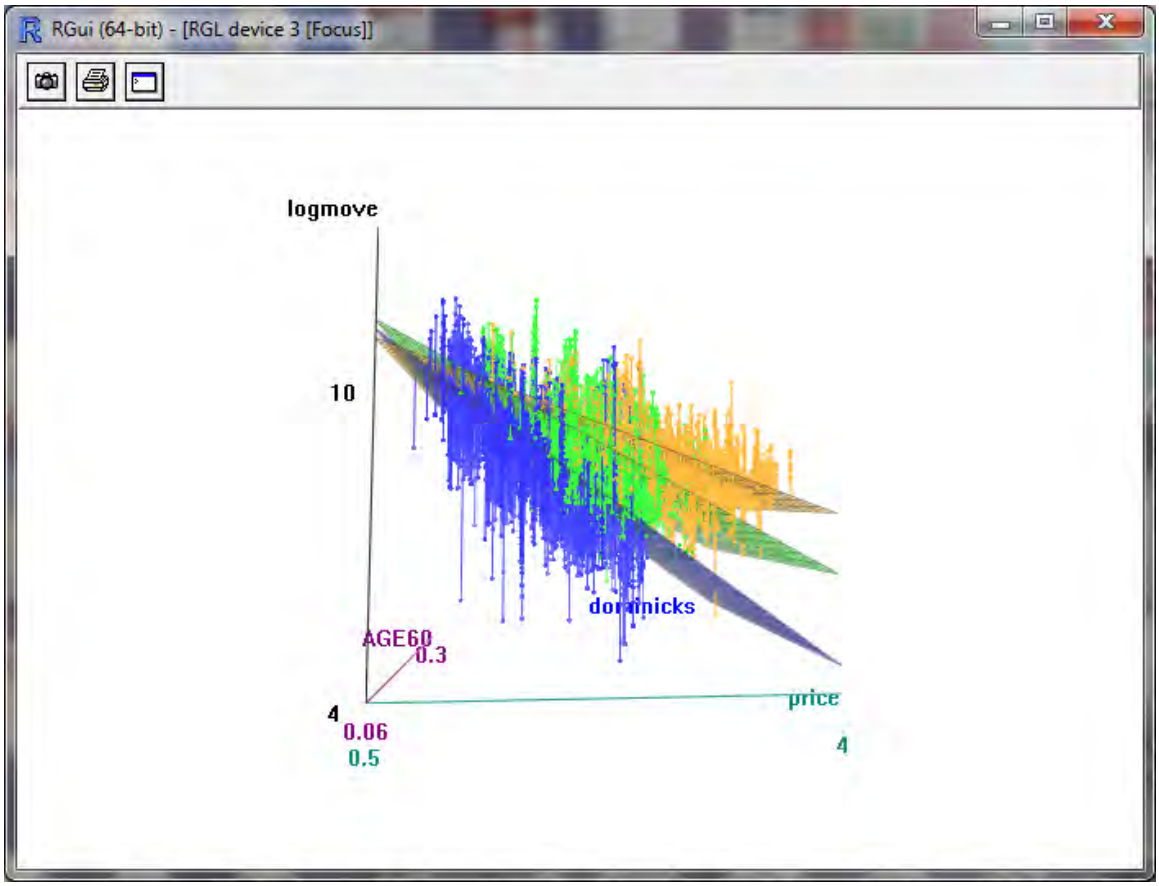



Does price affect sales?

Does age affect sales?

To generate 3D graphs by brand,

1. Graphs, 3D Graph, 3D Scatterplot
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price
3. Select logmove (log of sales) as the response variable
4. Click Plot by groups, select brand, then OK
5. Click OK
6. Expand the window by clicking on the box in the upper right corner of your graph
7. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move




Which brand is more sensitive to price?

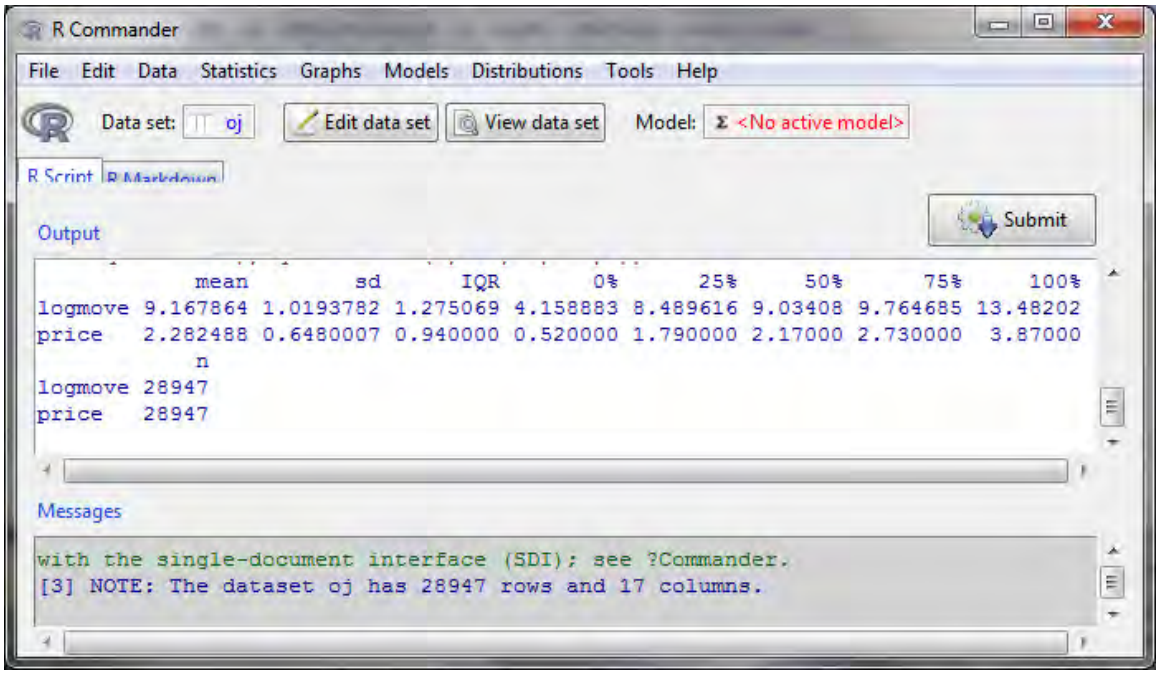
**Saving Graphs**

You can save graphs by clicking Graphs, Save Graph to File, then select the type of file.

**Session 7.7: Statistical Summaries**

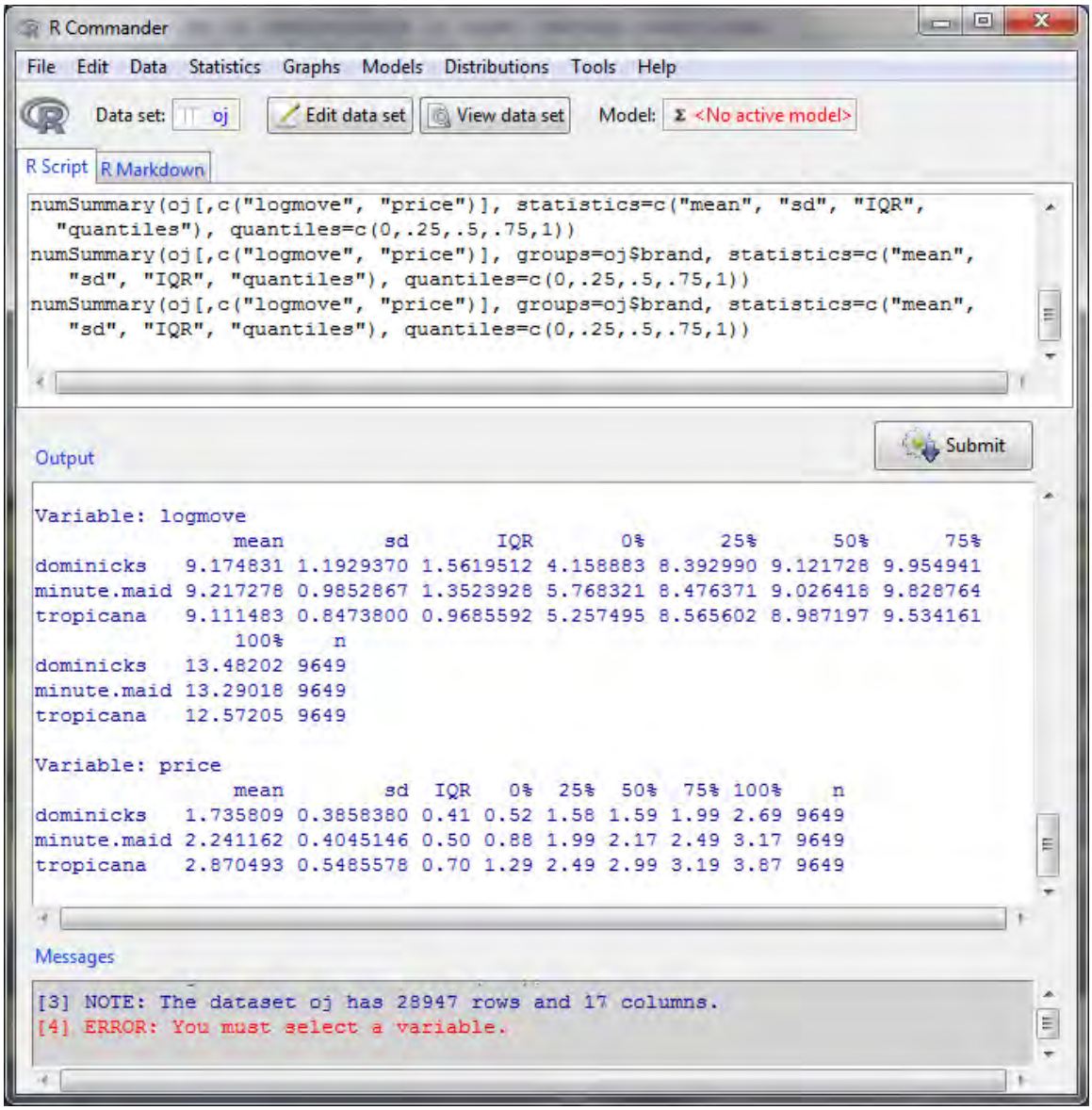
The mean, standard deviation and quartiles can be found by:

1. Click on Statistics, Summaries, Numerical Summaries
2. Select logmove, price by holding down the control key
3. Click OK




To categorize by brand, the mean, standard deviation and quartiles can be found by:

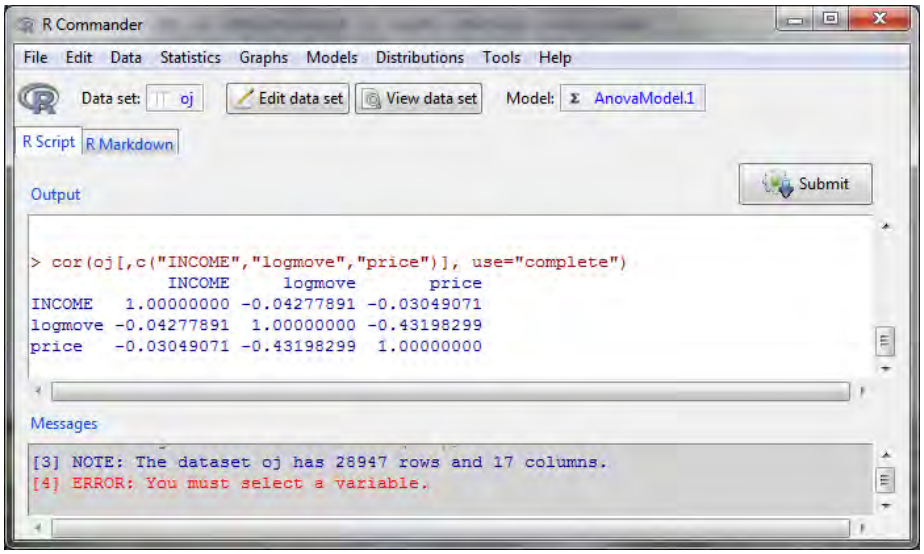
1. Click on Statistics, Summaries, Numerical Summaries
2. Select logmove, price by holding down the control key
3. Click Summarize by Groups, then OK
4. Click OK


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: oj
Edit data set
View data set
Model: <No active model>
R Script
R Markdown
numSummary(oj[,c("logmove", "price")], statistics=c("mean", "sd", "IQR",
"quantiles"), quantiles=c(0,.25,.5,.75,1))
numSummary(oj[,c("logmove", "price")], groups=oj$brand, statistics=c("mean",
"sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
numSummary(oj[,c("logmove", "price")], groups=oj$brand, statistics=c("mean",
"sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
Submit
Output
Variable: logmove
mean sd IQR 0% 25% 50% 75%
dominicks 9.174831 1.1929370 1.5619512 4.158883 8.392990 9.121728 9.954941
minute.maid 9.217278 0.9852867 1.3523928 5.768321 8.476371 9.026418 9.828764
tropicana 9.111483 0.8473800 0.9685592 5.257495 8.565602 8.987197 9.534161
100% n
dominicks 13.48202 9649
minute.maid 13.29018 9649
tropicana 12.57205 9649
Variable: price
mean sd IQR 0% 25% 50% 75% 100% n
dominicks 1.735809 0.3858380 0.41 0.52 1.58 1.59 1.99 2.69 9649
minute.maid 2.241162 0.4045146 0.50 0.88 1.99 2.17 2.49 3.17 9649
tropicana 2.870493 0.5485578 0.70 1.29 2.49 2.99 3.19 3.87 9649
Messages
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.

**Session 7.8: Correlation**

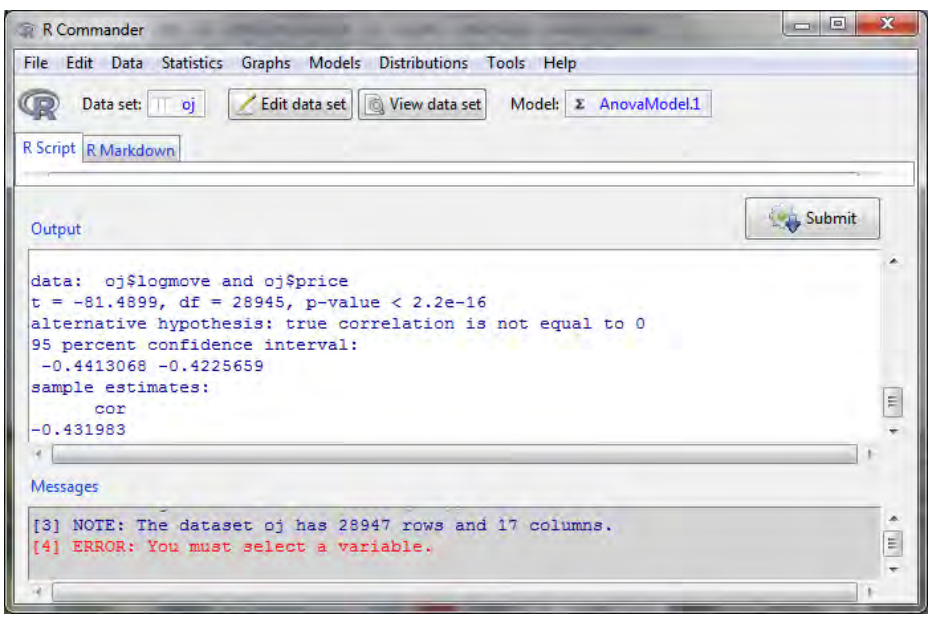
To generate a correlation matrix,

1. Click on Statistics, Summaries, Correlation Matrix
2. Hold down the control key and select INCOME, logmove, price
3. Click OK




The matrix shows the correlation, but not the statistical significance. To calculate significance,

1. Click on Statistics, Summaries, Correlation Test
2. Select both logmove and price
3. Click OK

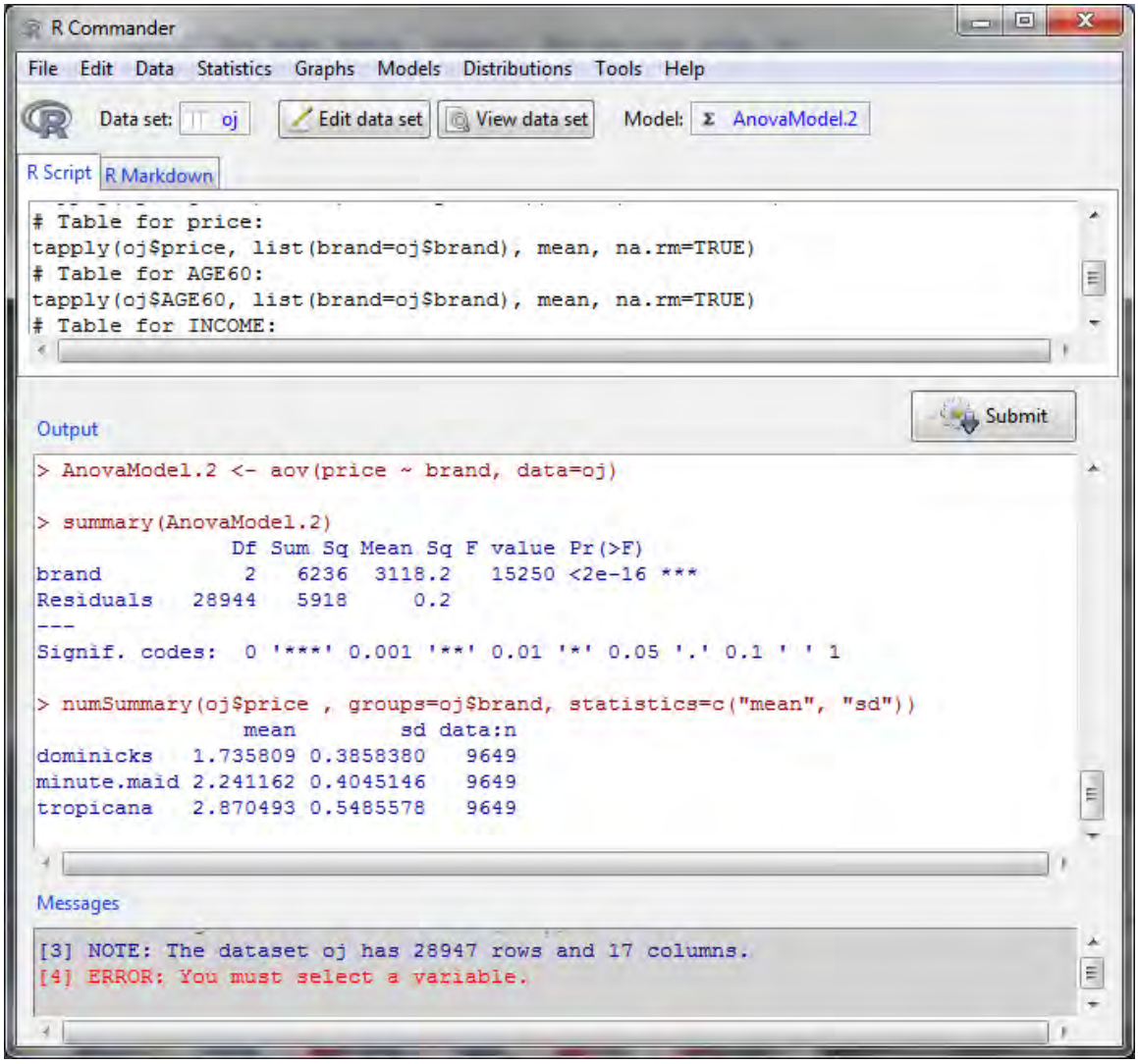



The p-value is less than 0.05, so the correlation is statistically significant.

**Session 7.9: ANOVA**

ANOVA stands for Analysis of Variance. It compares the means of several groups to determine if the groups are different. Let's see if prices are different across brands.

1. Click on Statistics, Means, One-way ANOVA
2. For the response variable, click on price
3. Click OK

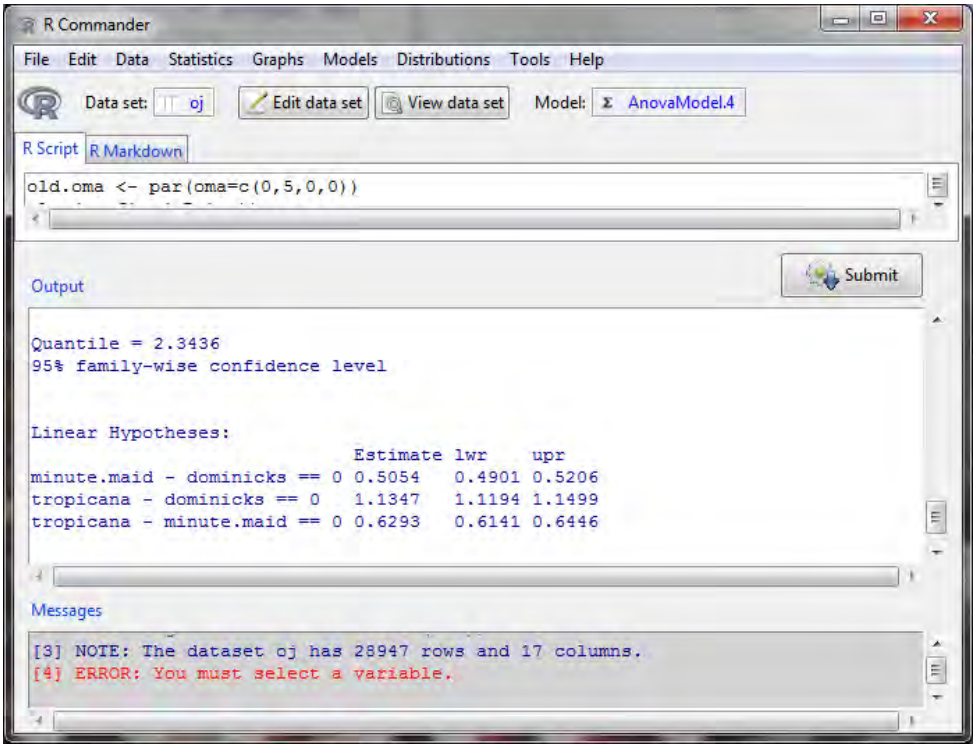



The F-statistic p-value [Pr(>F)] is less than 0.05. That means that one of the brands has a price that is statistically different from the others.

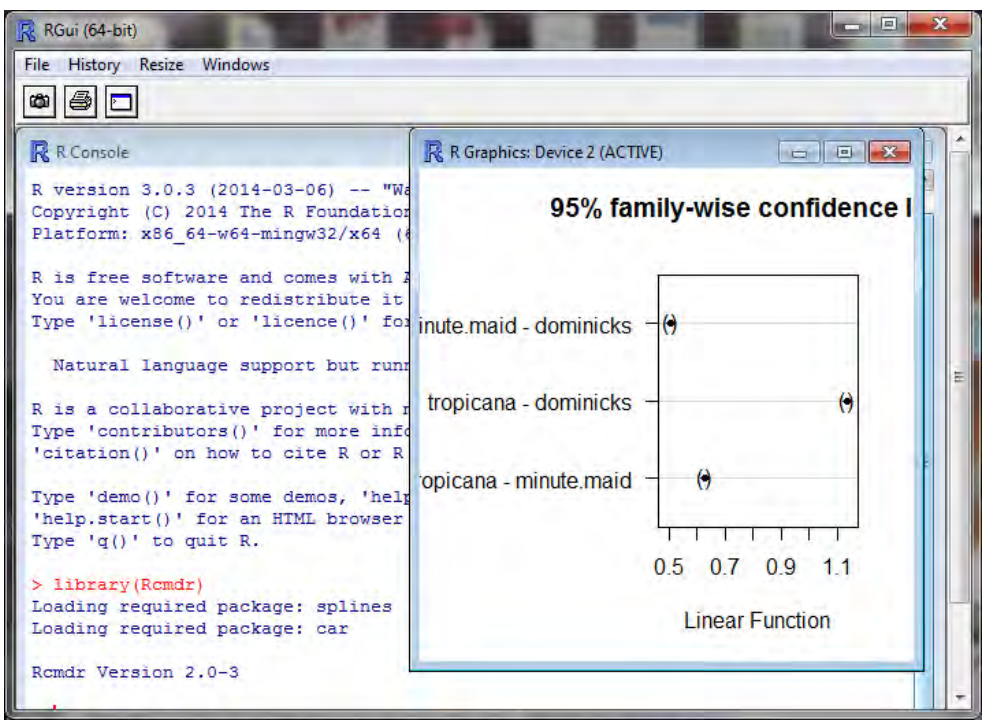
Let's now perform a pairwise comparison.

1. Click on Statistics, Means, One-way ANOVA
2. For the response variable, click on price
3. Check the box Pairwise comparison of means
4. Click OK

The pairwise comparison estimates the price and confidence interval for each brand (lower and upper interval values). Do they overlap?



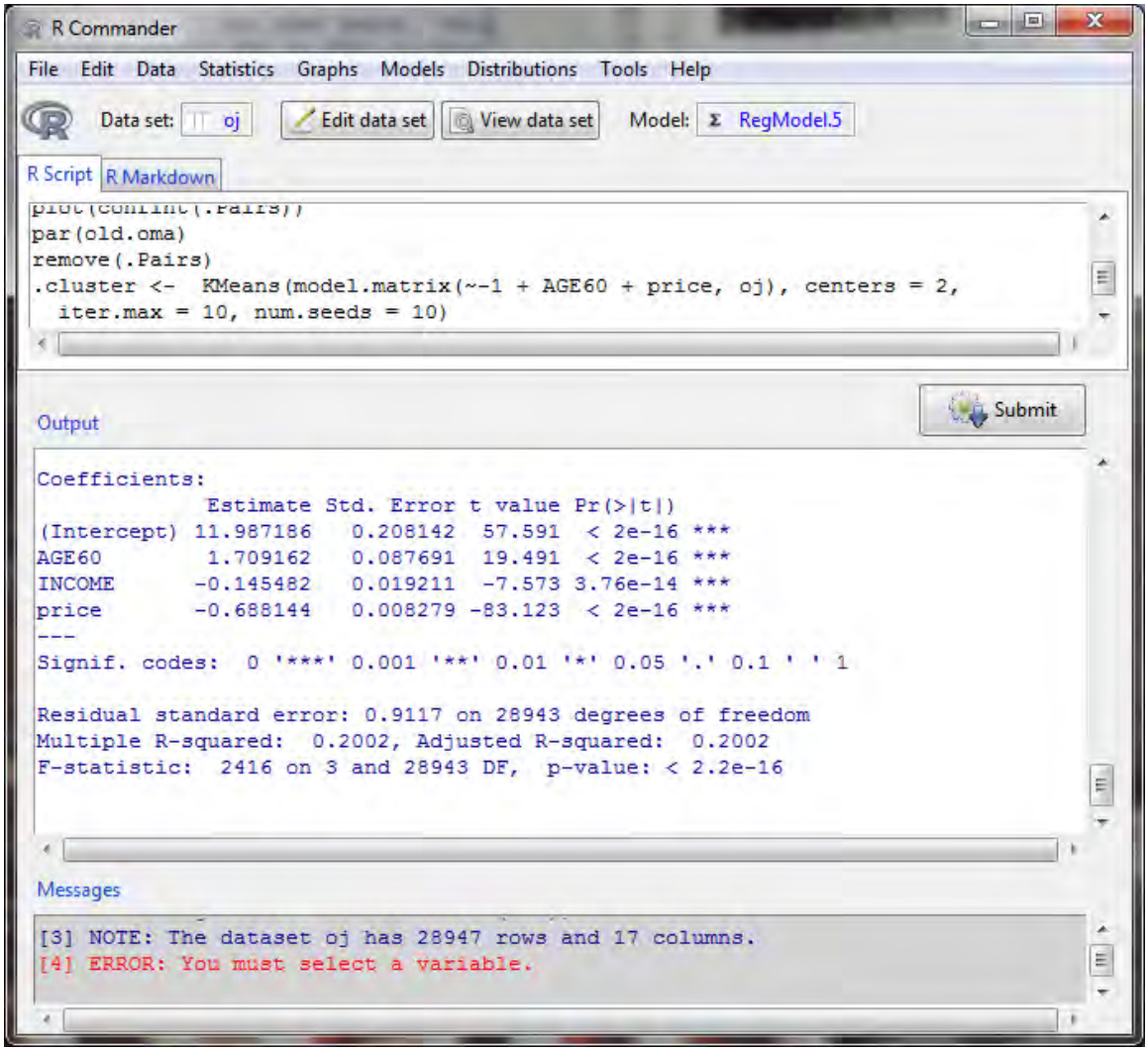

The graph portrays the estimate of price for each brand and the confidence interval.



**Session 7.10: Regression**

Linear regression of the log of sales against age, income and price can be performed by:

1. Click on Statistics, Fit Models, Linear Regression
2. For response variable, click on logmove
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
4. Click OK




Is the equation statistically significant?

How much of the variability in the log of sales is explained by the explanatory variables?

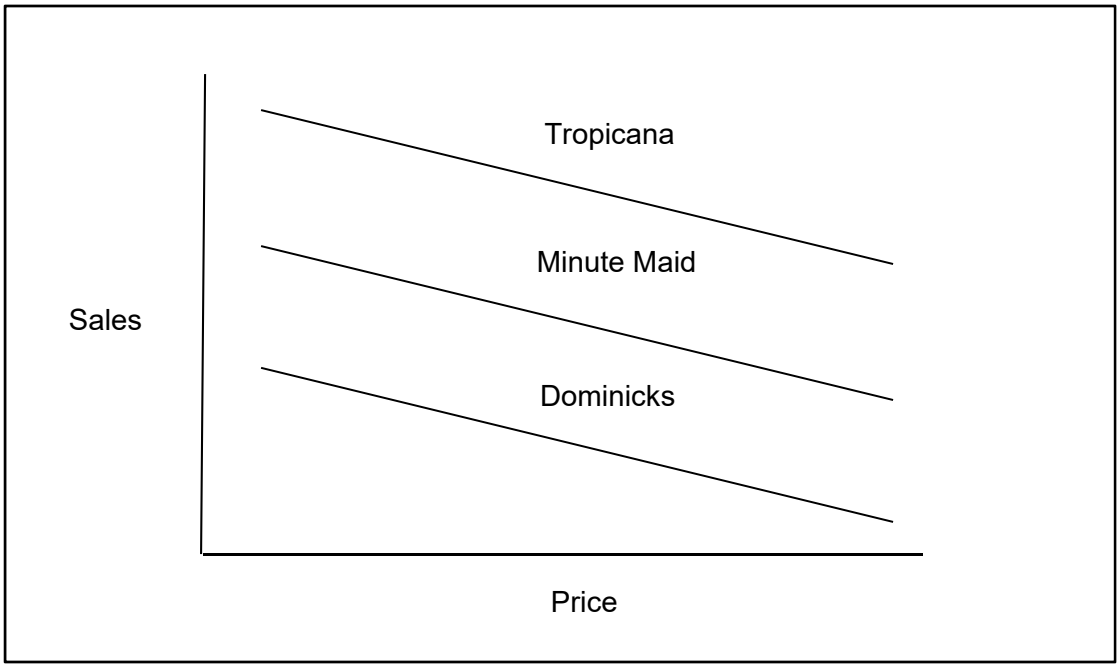
Which explanatory variables are statistically significant?

How does each explanatory variable affect sales? Which affect it positively and which negatively? How do you interpret this (what does it really mean)?

**Session 7.11: Regression with Dummy Variables**

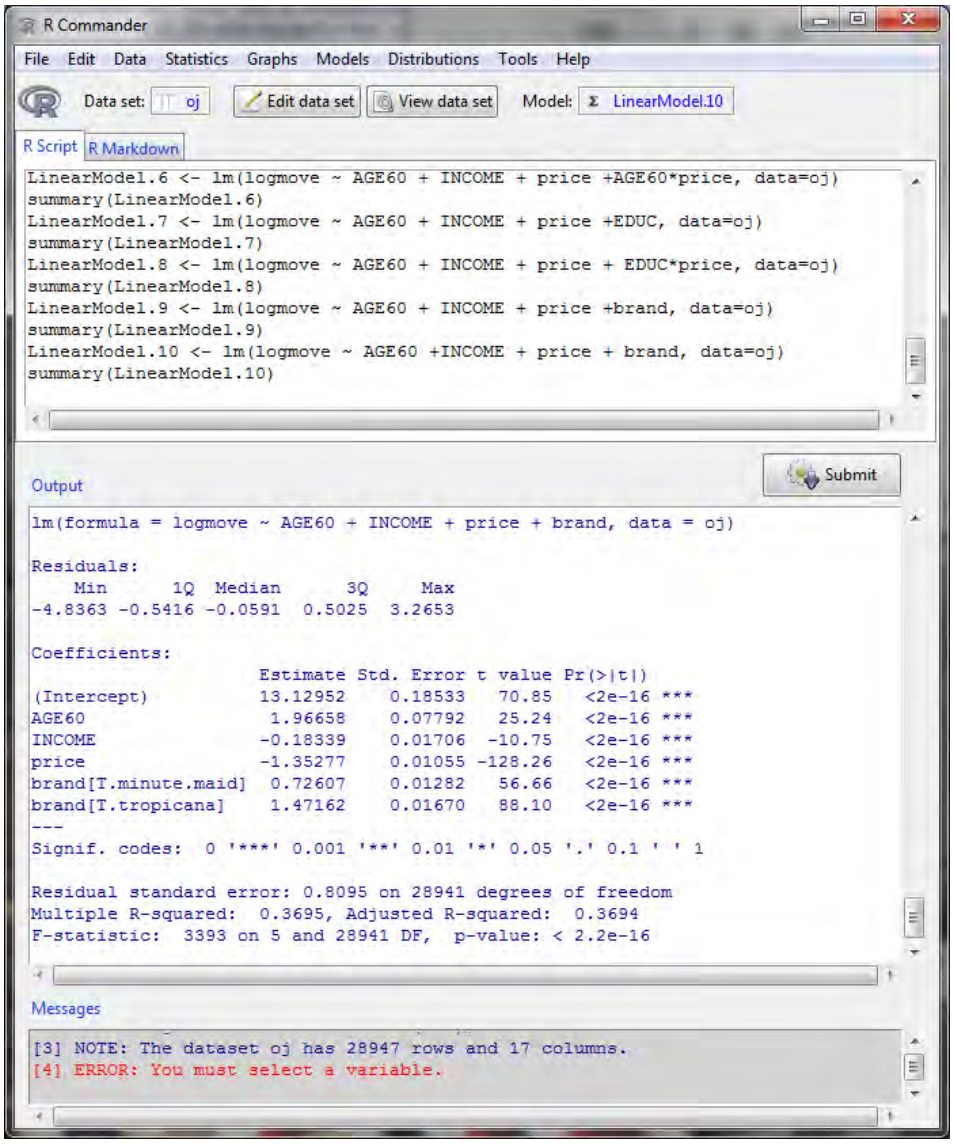
A dummy variable in a regression can assist in determining if the constant changes when the brand changes.

In the example below, the slopes of the lines are the same, which means the coefficients in the regression are the same. However, the intercept terms are different. This difference in intercept term is represented by a dummy variable (zero or one) where the coefficient in the dummy variable is the change in the intercept. One brand is identified as the base case, e.g., in this case it's Dominicks. The coefficient of the dummy variable for Tropicana is the shift in the intercept for Tropicana; similarly the coefficient of the dummy variable for Minute Maid is the shift in the intercept for Minute Maid.




To perform this more sophisticated regression,

1. Click on Statistics, Fit Models, Linear Model
2. Click Reset
3. Double click on logmove
4. Double click on AGE60
5. Double click on INCOME
6. Double click on price
7. Double click on brand (notice it says it's a factor)
8. Click on OK

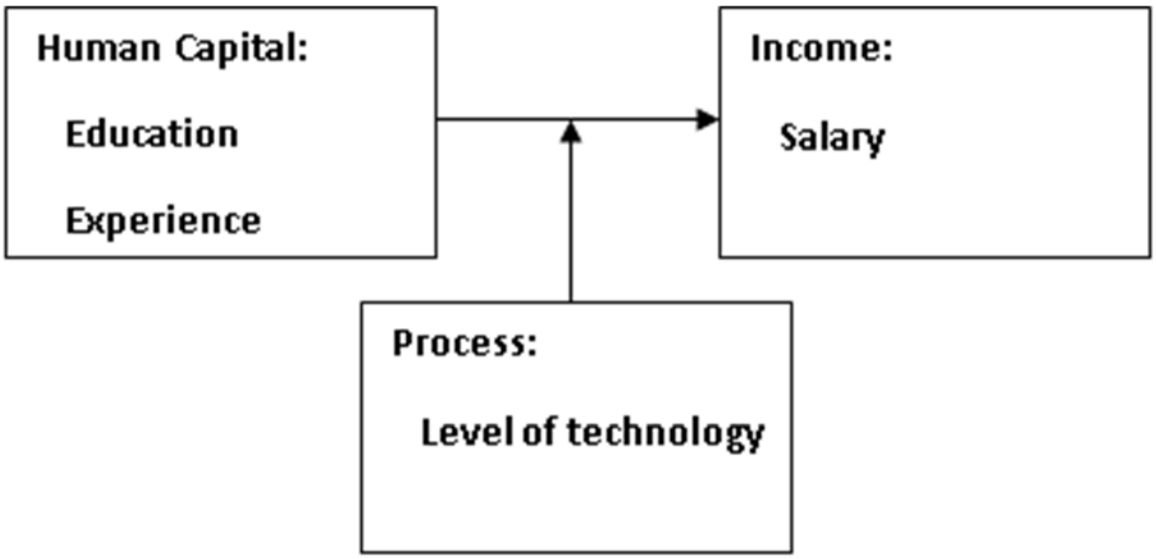



In the Output section, the intercept is 13.12952. This is the intercept for dominicks. The intercept for minute maid is 13.12952 + 0.72607. The intercept for Tropicana is 13.12952 + 1.47162.

This means, all else being equal, the log of sales is highest for Tropicana.

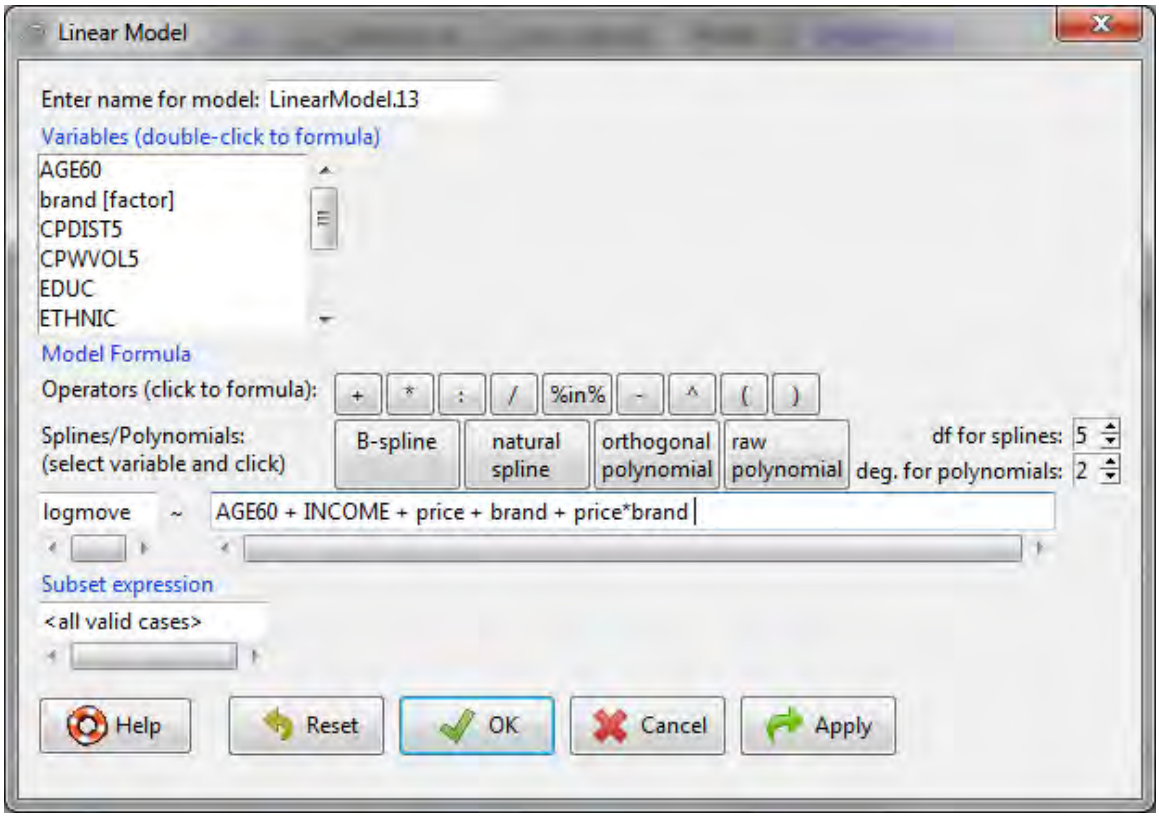
**Session 7.12: Moderating effects (interactions of price and brand)**

The interaction of two variables can be modeled and is called a moderating effect. A moderating effect magnifies the effect of one variable by changing another variable. In the following example, salary is dependent on education and experience of an individual. The technology that a person has available can enhance the effect of education and experience on their salary. Think of this as a catalyst which can increase or decrease the effect of a variable. This moderation is represented by an interaction, or multiplying two variables together.




In the previous example, we examined if the intercept is different for each brand. It's possible that the slope of the relationship between price and sales varies by brand. To test this, we create what is called an interaction term. An interaction is two variables multiplied together.

1. Click on Statistics, Fit Models, Linear Model
2. Click Reset
3. Double click on logmove
4. Double click on AGE60
5. Double click on INCOME
6. Double click on price
7. Double click on brand (notice it says it's a factor)
8. Double click on price (again)
9. Click on the multiplication sign (*)
10. Click on brand
11. Click on OK




On the next page, the coefficient on price is -1.94480. That means as price increases, the log of sales declines. But since we included an interaction term, this only applies to dominicks. We need to include the price*brand effect for minute maid and Tropicana. For minute maid, the coefficient is -1.94480 + 0.47545 = -1.46935. For Tropicana, the coefficient is -1.94480 + 0.94817 = -0.99663.

Which brand is more sensitive to price?

R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: oj | Edit data set | View data set | Model: LinearModel.13

R Script | R Markdown

```
summary(LinearModel.7)
LinearModel.8 <- lm(logmove ~ AGE60 + INCOME + price + EDUC*price, data=oj)
summary(LinearModel.8)
LinearModel.9 <- lm(logmove ~ AGE60 + INCOME + price +brand, data=oj)
summary(LinearModel.9)
LinearModel.10 <- lm(logmove ~ AGE60 +INCOME + price + brand, data=oj)
summary(LinearModel.10)
LinearModel.12 <- lm(logmove ~ AGE60 + INCOME + price + brand + price*brand,
   data=oj)
summary(LinearModel.12)
LinearModel.13 <- lm(logmove ~ AGE60 + INCOME + price + brand + price*brand,
   data=oj)
summary(LinearModel.13)
```

Submit

Output

```
Call:
lm(formula = logmove ~ AGE60 + INCOME + price + brand + price *
    brand, data = oj)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1482 -0.5318 -0.0519  0.4937  3.5030

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               14.20227    0.18385  77.248   <2e-16 ***
AGE60                      1.96066    0.07606  25.776   <2e-16 ***
INCOME                    -0.18754    0.01666 -11.259   <2e-16 ***
price                     -1.94480    0.02086 -93.218   <2e-16 ***
brand[T.minute.maid]      -0.04030    0.05854  -0.689    0.491
brand[T.tropicana]        -0.57834    0.05668 -10.204   <2e-16 ***
price:brand[T.minute.maid] 0.47545    0.02882  16.500   <2e-16 ***
price:brand[T.tropicana]   0.94817    0.02549  37.194   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7902 on 28939 degrees of freedom
Multiple R-squared:  0.3992, Adjusted R-squared:  0.3991
F-statistic:  2747 on 7 and 28939 DF,  p-value: < 2.2e-16
```

Messages

```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```

# SCM 651: Business Analytics



Week 7

# Agenda



- Overview of homework #3 (Regression and Optimization)
- Overview of homework #4 (Logit, Probit, Neural networks: info in week 9 videos)
- Review of hands-on exercises
- Group discussion of articles
  - Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients
  - A Review of Analytics and Clinical Informatics in Health Care

# Homework #3



1. Graph, regression, calculated sales, revenue, profit
2. Constrained optimization
3. Discussion of risks, other data which would be valuable

# Homework #4



1. Logit and probit analysis (see week 9)
2. Moderating effects (week 7)
3. Final logit & probit models with interaction effects (moderating effects), prediction of outcome, sensitivity analysis
4. Neural network analysis
5. Neural network prediction model and sensitivity analysis (new material in handout in week 9)

# Week 7 - Review



- Histograms, Boxplots, and other charts
  - Boxplots show minimum, 25%-ile, median, 75%-ile, maximum
  - Histograms show frequency of data within intervals
  - Scatterplots add trend line, average, standard deviation
- Statistical summaries
  - Replicates capability seen in Excel: minimum, average, median, maximum
  - Also can calculate by brand or other category

# Week 7 - Review



- Correlation
  - Positive correlation means when one variable increases, the other increases
  - Negative correlation means when one variable increases, the other decreases
  - Correlation does not measure the magnitude of the change
- ANOVA (Analysis of Variance)
  - Compares the means of two or more populations
- Regression
  - Linear regression in Excel is similar to multivariate regression in R

# Week 7 - Review



- Dummy variables
  - Dummy variables are used to measure the differences in intercepts between two groups, for example, different brands
- Moderating effects (interaction term $x_1*x_2$)
  - Moderating effects (interaction effects) measures the difference in slope between two groups

# Article #1: Big Data in Health Care



- Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients
  - What are the six opportunities to reduce costs through analytics?
  - How can cost be reduced in each

# Article #1: Big Data in Health Care



- Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients
  - What are the six opportunities to reduce costs through analytics?
    - High cost patients
    - Readmissions
    - Triage
    - Decompensation
    - Adverse events
    - Diseases affecting multiple organs
  - How can cost be reduced in each

# Article #2: A Review of Analytics in Clinical Informatics in Health Care



- A Review of Analytics and Clinical Informatics in Health Care
  - What are some methods for improvement in health care using analytics (page 2)
  - What are some challenges for analytics in health care (page 4)

# Article #2: A Review of Analytics in Clinical Informatics in Health Care



- A Review of Analytics and Clinical Informatics in Health Care
  - What are some methods for improvement in health care using analytics (page 2)
    - Identify patients of greatest risk: early detection in neo-natal care
    - Wearable monitors: disease prevention though monitoring
    - Costs savings and resolution of billing anomalies: revenue leakage
    - Better manage resource allocation: patient's length of stay
  - What are some challenges for analytics in health care (page 4)
    - Garbage in, garbage out
    - Increased demand for professionals well versed in analytics and medicine

## Regression Assumptions and Diagnostics

- Linear regression can only be performed when a set of assumptions are satisfied.
- Excel can perform regressions but has limited ability to test the assumptions.
- R has diagnostics to determine if the assumptions are met.



## Linear Regression Assumptions

- The relationship between X and Y is linear.
- If there are multiple X variables, they are not correlated (no multicollinearity).
- The error terms (distance from the data point to the predicted line):
    - Have zero mean and constant variance (no heteroscedasticity)
    - Are independent (no serial correlation)
    - Are normally distributed (includes no outliers)

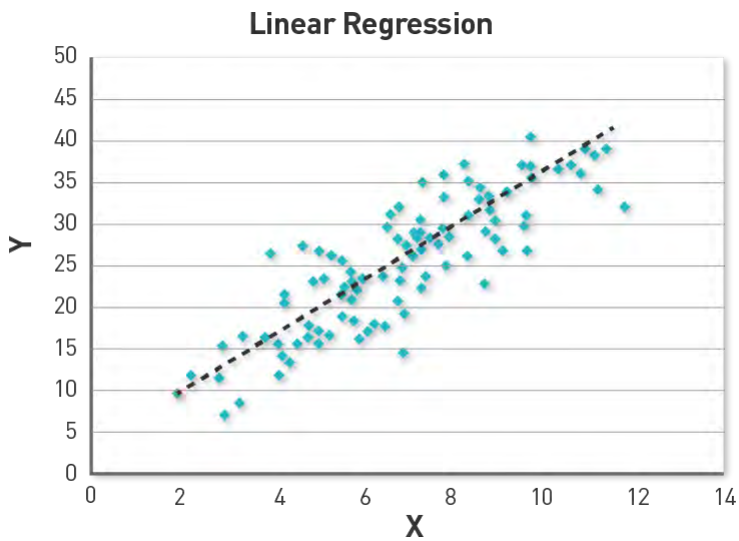
Each of these will be explained in the video for each assumption and diagnostic test.

## Diagnostics and Solutions

| Assumption | Diagnostic | Solution |
|---|---|---|
| Linearity | Ramsey | Transformation |
| No multicollinearity | Variance inflation factor | Combine variables or drop one |
| Heteroscedasticity | Breusch-Pagan | Transformation |
| No serial correlation | Durbin-Watson | Time series analysis |
| Normality/Outliers | Bonferroni outlier test | Drop outliers |

View All View Keyframes

## Regression Assumption: Linearity

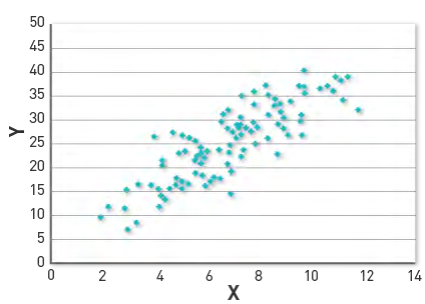



For linear regression, the data must be from a linear relationship.

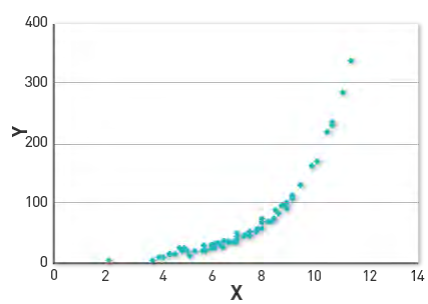
2 of 10

## Linear vs. Nonlinear Relationship Illustration

Linear relationship




Nonlinear relationship




5 of 10

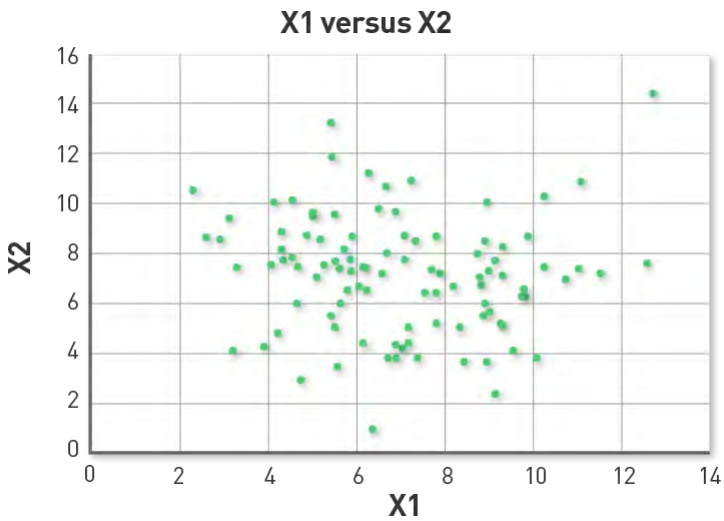
## Diagnostics and Solutions

- Diagnostic test:
  - Ramsey Regression Equation Specification Error Test (RESET) (1969)
- Solution:
  - Take a transformation of the data (e.g., logarithm, square, square root, or inverse).
    - Advanced techniques (i.e., Box-Cox for Y variables, Box-Tidwell for X variables) help determine the appropriate transformation.

10 of 10

View All View Keyframes

## Regression Assumption: Collinearity

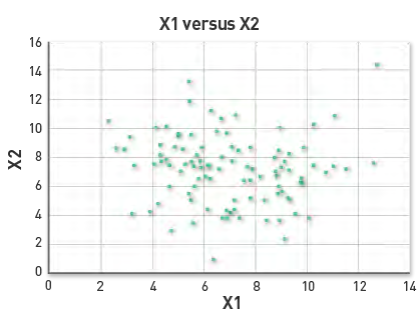



Each pair of X variables must not be correlated.

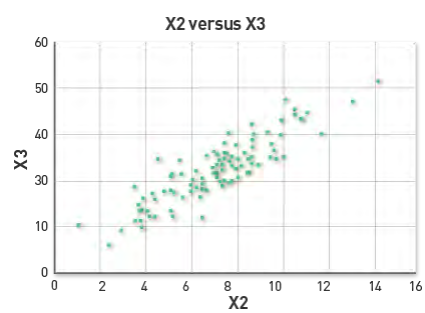
2 of 10

## Not Correlated vs. Correlated: Illustration

Not correlated




Correlated




5 of 10

## Diagnostics and Solution

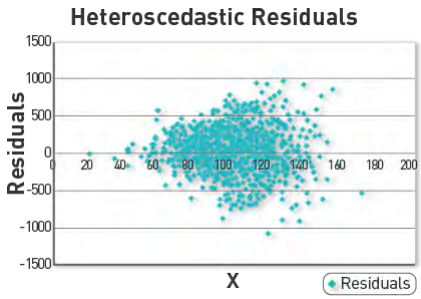
- Diagnostic test:
  - Variance inflation factor (VIF)
- Solution:
  - Drop one of the variables.
  - Use advanced techniques (e.g., factor analysis) to combine variables into one factor.

10 of 10

View All View Keyframes

## Regression Assumption: Heteroscedasticity

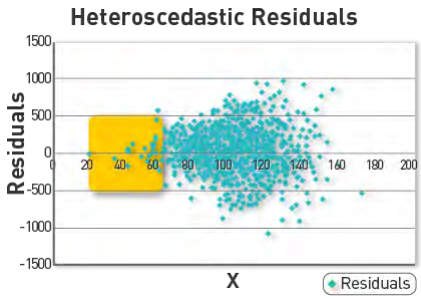
- Residuals (error terms) are the difference between the regression line and the data point.
- The error terms must have zero mean and constant variance.




3 of 12

## Heteroscedasticity, Illustrated

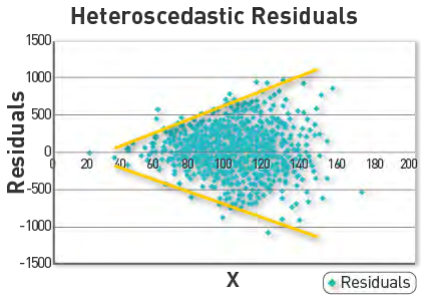
- Is there a pattern?




6 of 12

## Heteroscedasticity, Illustrated (cont.)

- Is there a pattern?
- Variance increases as X increases (i.e., variance is not constant).




7 of 12

## Diagnostics and Solutions

- Diagnostic test:
  - Breusch-Pagan test of heteroscedasticity

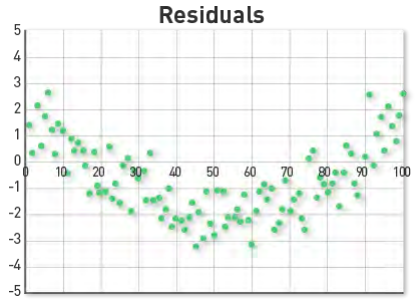
- Solution:
  - A transformation usually fixes this problem (e.g., logarithm, square, square root, or inverse).
  - If the problem still occurs after a transformation, Huber regression corrects for the error.

12 of 12

View All View Keyframes

## Regression Assumption: Serial Correlation

- Error terms must not be related to each other.
- Serial correlation often occurs with time-series data.




3 of 7

## Diagnostics and Solutions

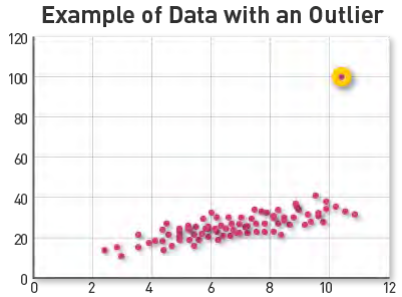
- Diagnostic test:
  - Durbin-Watson test of serial correlation
- Solution:
  - Advanced techniques include rho differencing and ARCH.

7 of 7

View All View Keyframes

## Regression Assumption: Serial Correlation

- Outliers are data points that are significantly different from the other points.
- An influential outlier can twist the regression line away from its true position.




5 of 10

## Diagnostics and Solutions

- Diagnostic test:
  - Bonferroni outlier test
- Solution:
  - If the data point is clearly an outlier, drop the bad data point.
  - Always report that outliers are dropped from the analysis.

10 of 10

View All View Keyframes

## Data Mining

- Data mining is a data exploration process that searches for patterns or relationships within data.
- Techniques combine machine learning, statistics, and database systems.
- Some approaches include decision trees, logistic regression, perceptrons, and neural networks.
- Data mining requires new features in R, including installation of the Rattle package.

5 of 5

View All View Keyframes

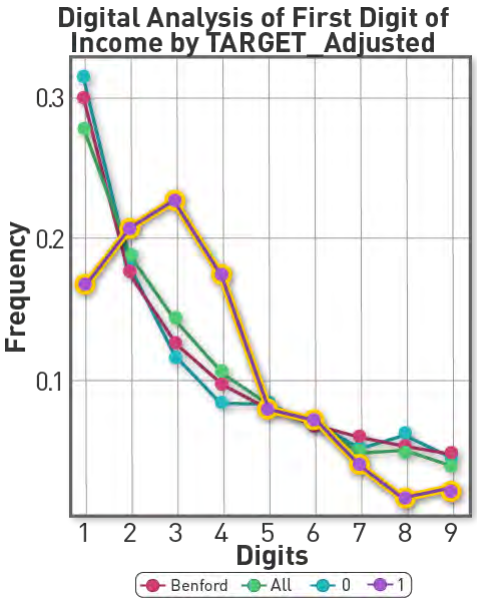
## Benford's Law

- Accountants and auditors use data mining techniques to detect fraud.
  - One technique relies on Benford's law.
- Benford's law states that in financial data, small digits appear more often than larger digits at the beginning of a number.
  - E.g., 30% of all financial numbers begin with the number 1; fewer than 5% begin with the number 9.

4 of 7

## Benford's Law Illustration

- If the initial digits in the data do not match the frequency specified by Benford's law, the numbers are likely fraudulent.




7 of 7

View All View Keyframes

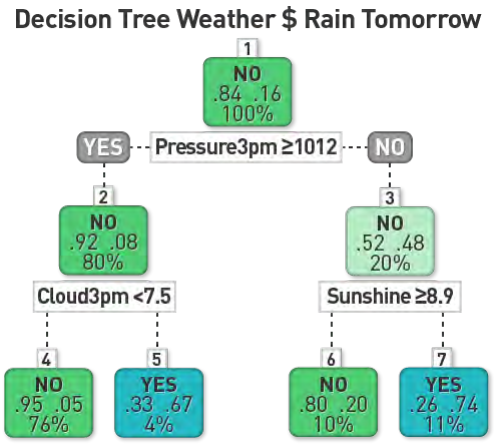
## Decision Trees

- Decision trees identify which variables are most important for making decisions or predictions.
- The technique removes noise (entropy) and identifies which variables contribute the most information.
- The result is a set of rules that can assist in decision making.

4 of 8

## Decision Tree Illustration: Weather Forecasting

- A simple decision tree can help predict if it will rain tomorrow.




8 of 8

# R: Intermediate

**Datafiles**

For these exercises, download the files:
"Business Analytics – Week 8 Instructions.doc"
"Business Analytics – Week 8 oj.xls"

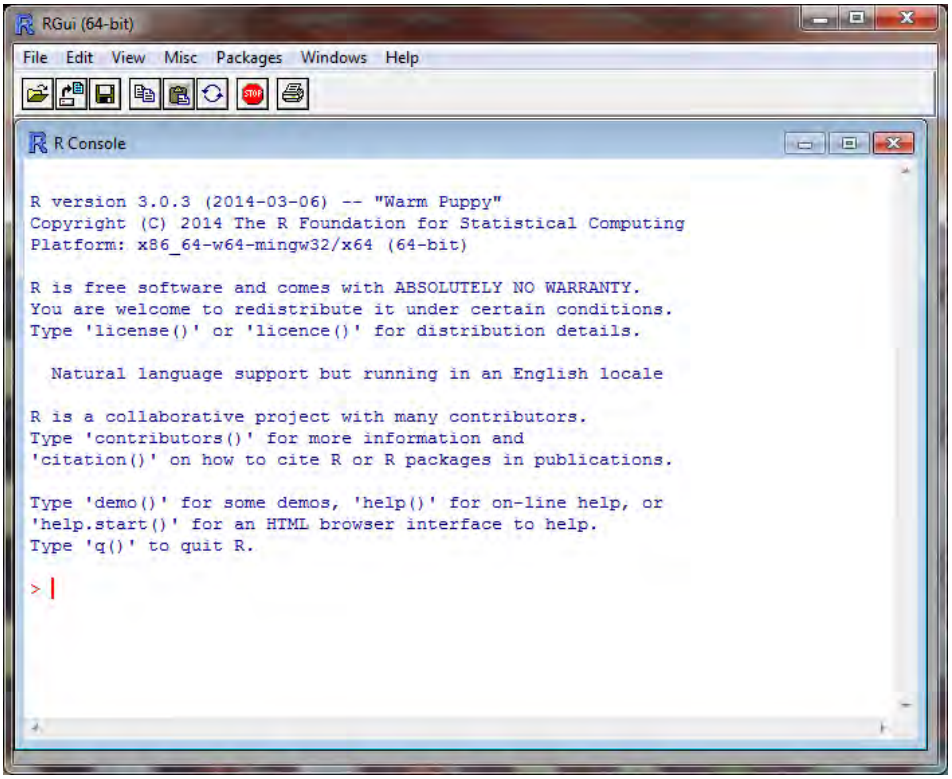
**Review: Installation of R**

R is a free downloadable package capable of performing sophisticated statistical analysis and data mining. The software is already installed on the classroom laptops. To install on your own personal computer:

1. Go to the website: http://cran.r-project.org/bin/windows/base/
2. For a Mac, go to http://cran.r-project.org/bin/macosx/
3. Click on Download R 3.0.3 for Windows
4. Click on Run, and follow the install instructions

**Starting R**

1. Click on the Start button in the lower left corner of Windows
2. Click on All Programs, then click on the R folder, then R

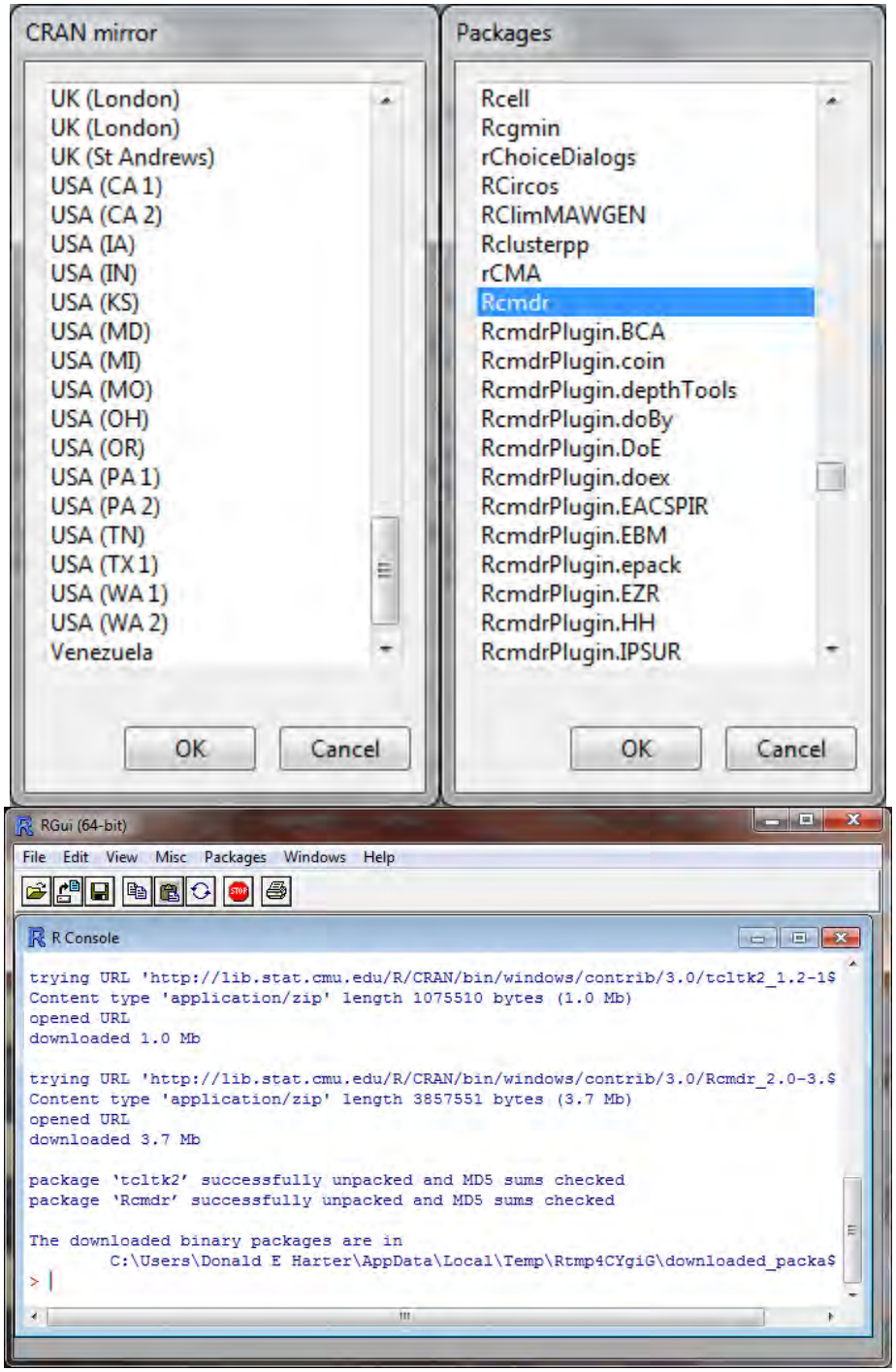



This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

**Installing R Commander**

Follow these steps only if you don't already have Rcmdr installed.

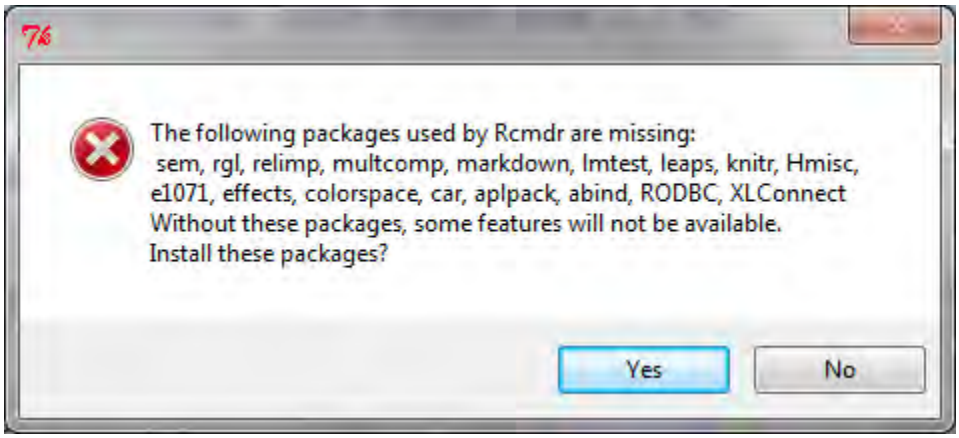
1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on Rcmdr, then OK
5. When prompted to create a personal library, click Yes

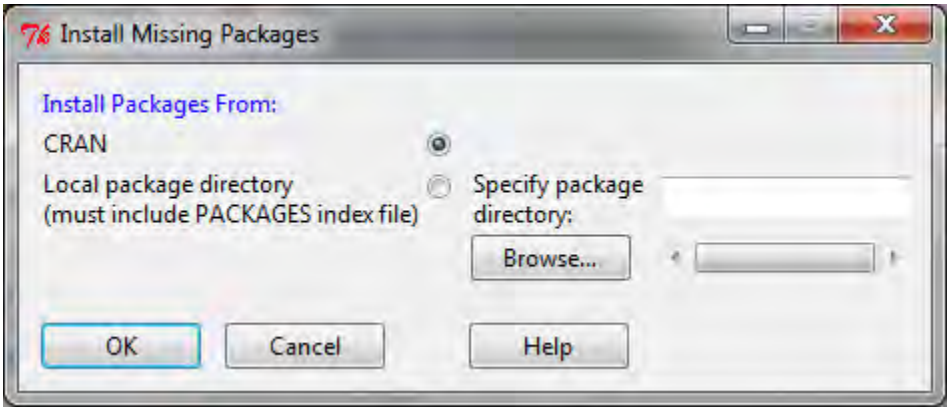


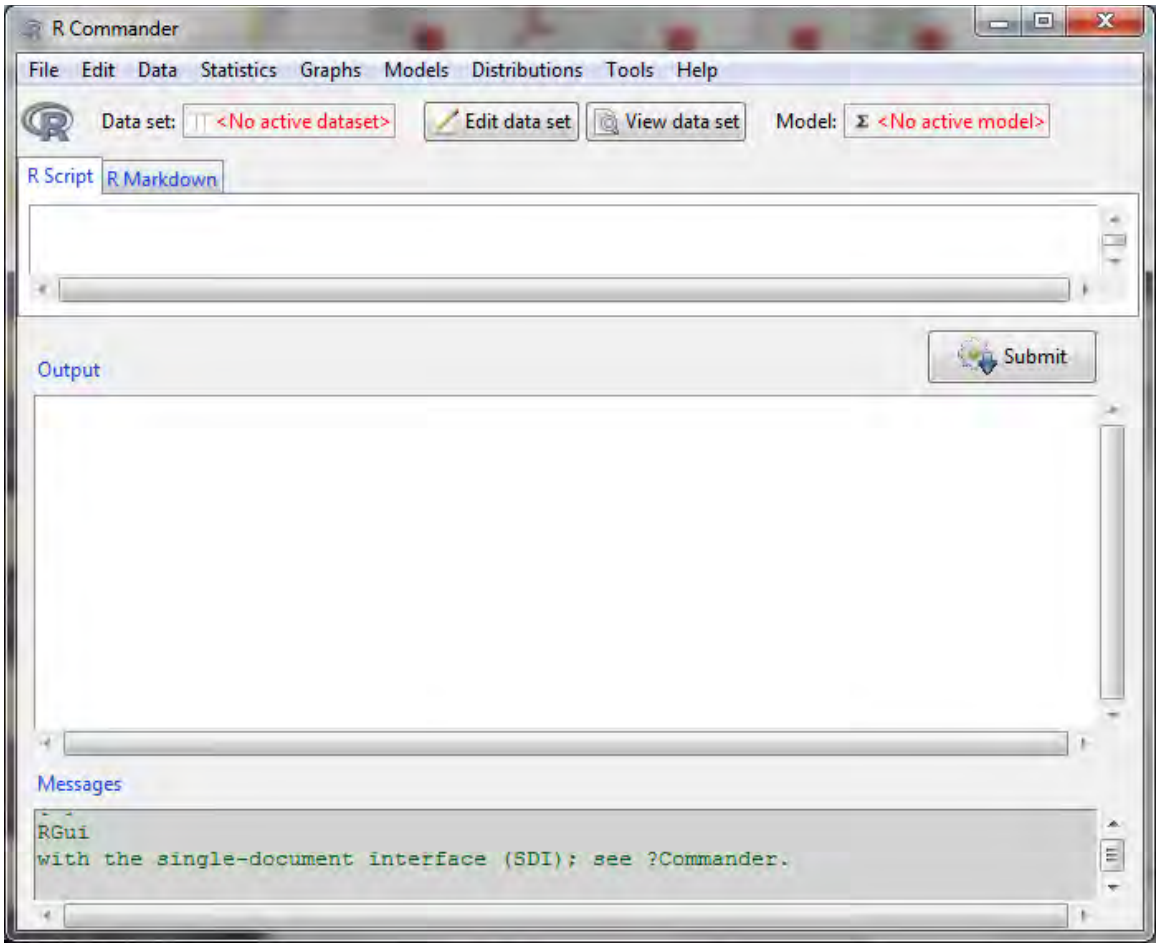
**Launch Rcmdr (R Commander)**

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type library(Rcmdr)
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK
4. R will install the necessary software
5. The R Commander screen will appear



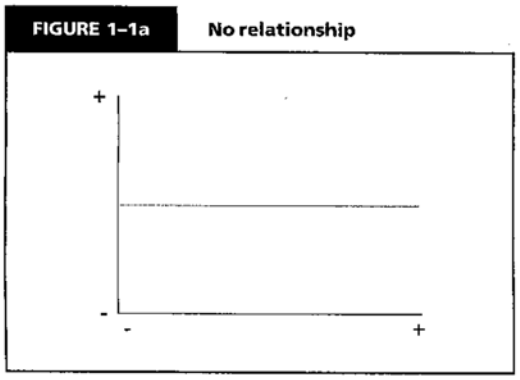

R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: <No active dataset>
Edit data set
View data set
Model: <No active model>
R Script
R Markdown
Output
Submit
Messages
RGui
with the single-document interface (SDI); see ?Commander.

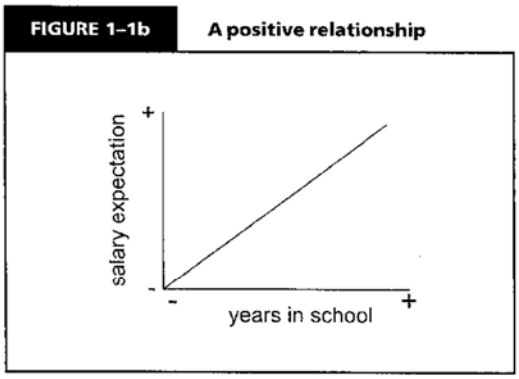
**Modeling Background (Correlation & Regression)**

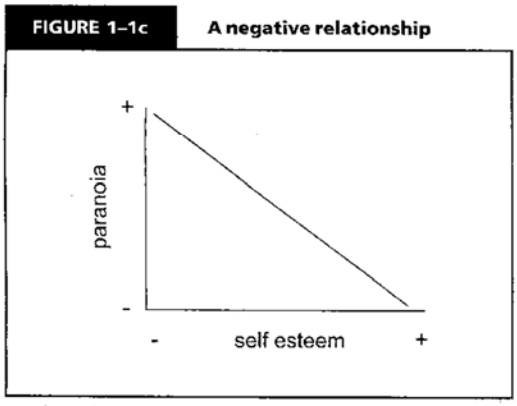
Identifying data relationships is key to modeling behavior of customer, student, and corporate data. First, let's consider two variables and the relationships between them. When comparing two data variables, you can have:

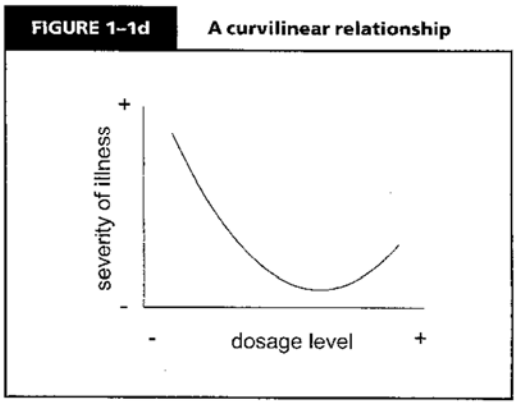
1. No relationship between the variables
2. A positive relationship (when one variable goes up, the other goes up)
3. A negative relationship (when one variable goes up, the other goes down)
4. A curvilinear relationship (a non-linear relationship

Examples of these, from The Research Methods Knowledge Base by Trochim & Donnelly (2007):



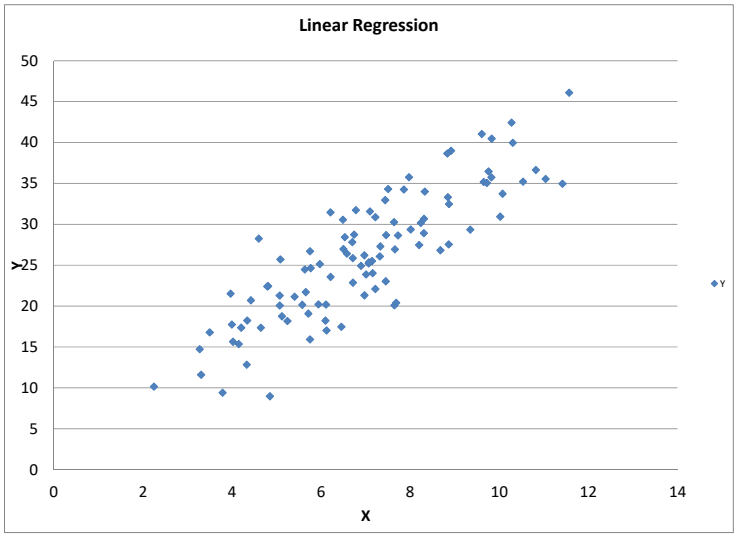



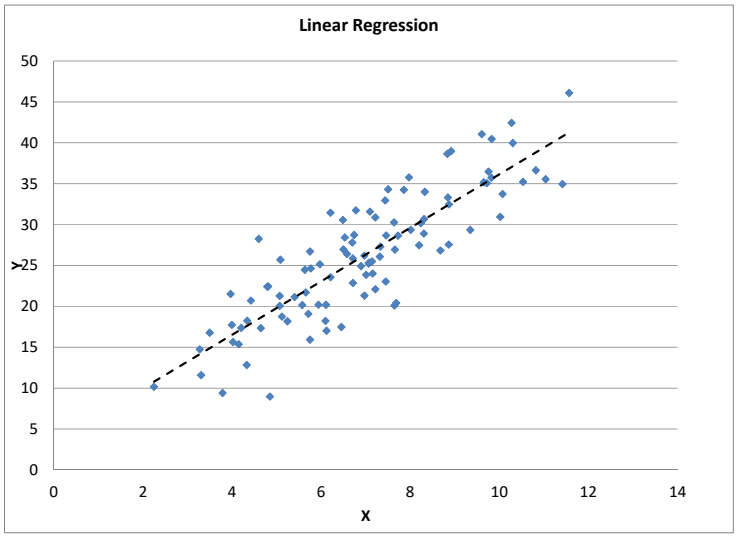



## Regression

Linear regression is a technique that calculates the relationship between a dependent variable Y and one or more independent variables, or X's. Assume that you have data similar to the picture below.




You can calculate a regression trend line based on the data. This dashed line represents $\hat{Y}$ which is the estimate of the Y equation.




The vertical distance between the line and the data point is called the residual or error term.

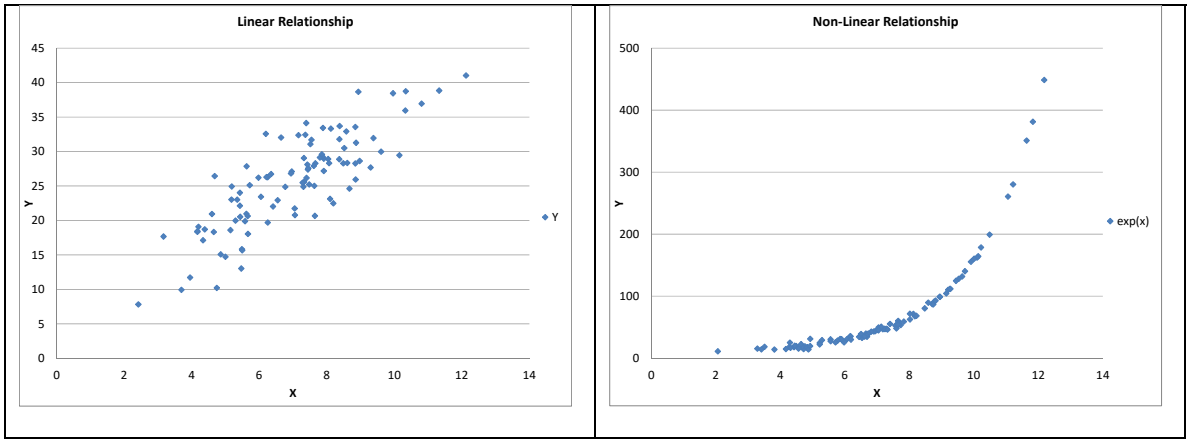
### Regression Diagnostics

There are several assumptions of linear regression:

1. The relationships are linear
2. The X variables (explanatory variables) are not correlated
3. Distribution of residuals
    a. The error terms have constant variance
    b. The errors terms are not correlated
    c. There are no outliers

### Assumption #1: the relationship is linear

Let's examine each of these assumptions. In the pictures below, the left picture has data with a linear relationship, the right picture had non-linear data. Linear regression can only be used on data with a linear relationship. Transformations can be used to transform non-linear data into linear data. For example, exponential data like the data on the right can be converted into a linear relationship by taking the logarithm of both the Y and X variables.




### Test for Linearity

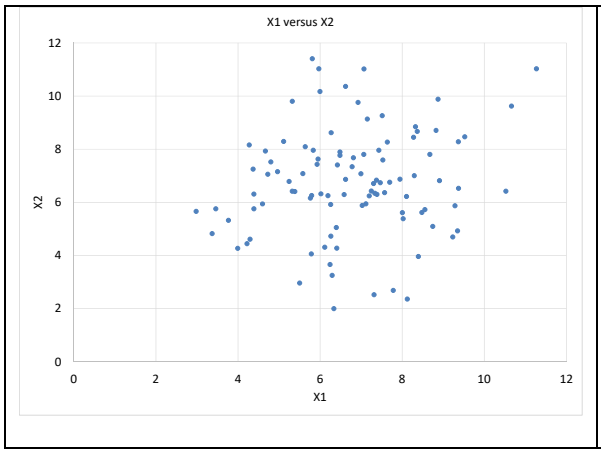
The Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity

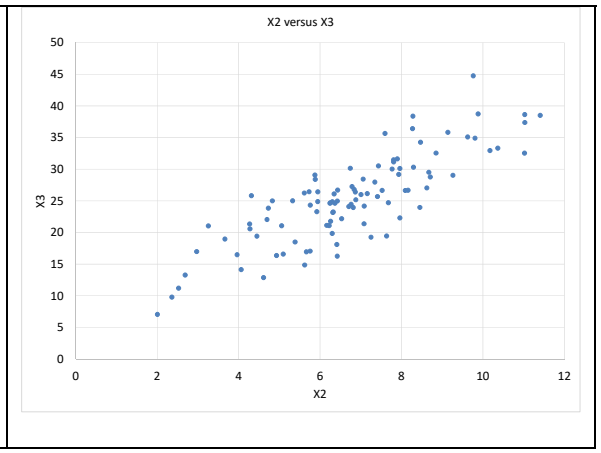
### Solution to non-linearity

The best solution for non-linear data is to transform the data using logarithms, squares, square roots, or inverses (1/variable). There are more advanced techniques which can assist in determining the correct transformation (Box-Cox for the Y variable; Box-Tidwell for the X variables).

**Assumption #2: The X variables are not correlated (no multi-collinearity)**

When including more than on explanatory or independent variable (i.e., X variable) in an analysis, you must ensure that they are not related to each other. If you plot the X variables, you should see no pattern, such as the picture on the left between variables X1 and X2. If you see a relationship, such as on the right between X2 and X3, then multi-collinearity exists.




**Test for Multi-collinearity**

The Variance Inflation Factor test of correlated explanatory variables

**Solution to Multi-collinearity**

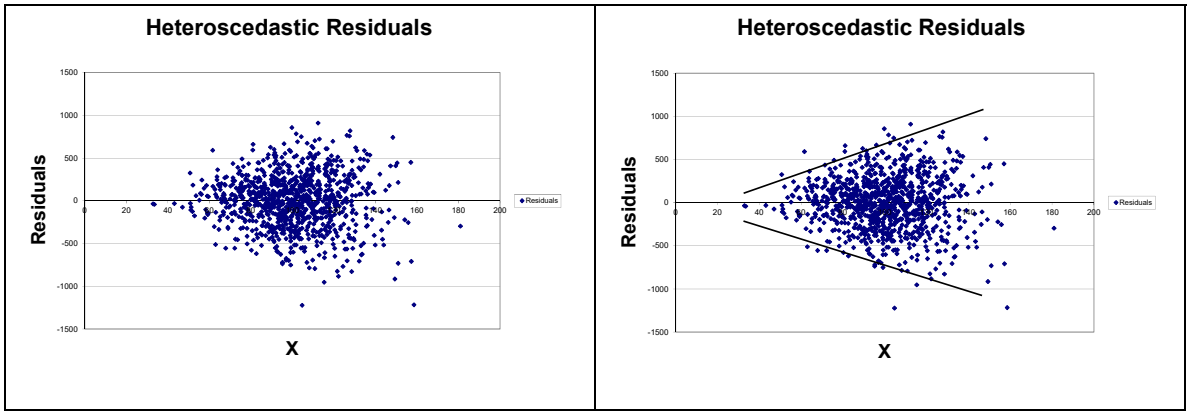
If two or more variables are collinear (highly correlated), there are three solutions:

1. Combine the variables, for example, take an average of the variables
2. Drop one of the variables
3. Use factor analysis to combine variables

**Assumption #3a: The error terms do not have constant variance (Heteroscedasticity)**

The residuals (error terms) of a regression must have constant variance over a range of X values. If the size of the error terms depends on an X value, this is called heteroscedasticity. Heteroscedasticity is often caused by performing a linear regression on non-linear data. In the charts below, there is no relationship between the X variable and the error term. On the right, the residuals or errors are heteroscedastic; the size of the error is dependent on the X value.

The picture below shows heteroscedastic residuals. Notice that the variability of the errors or residuals tends to grow larger for larger values of X. The picture on the right has lines added indicating the general growth in variability.




**Test for Heteroscedastiticy**

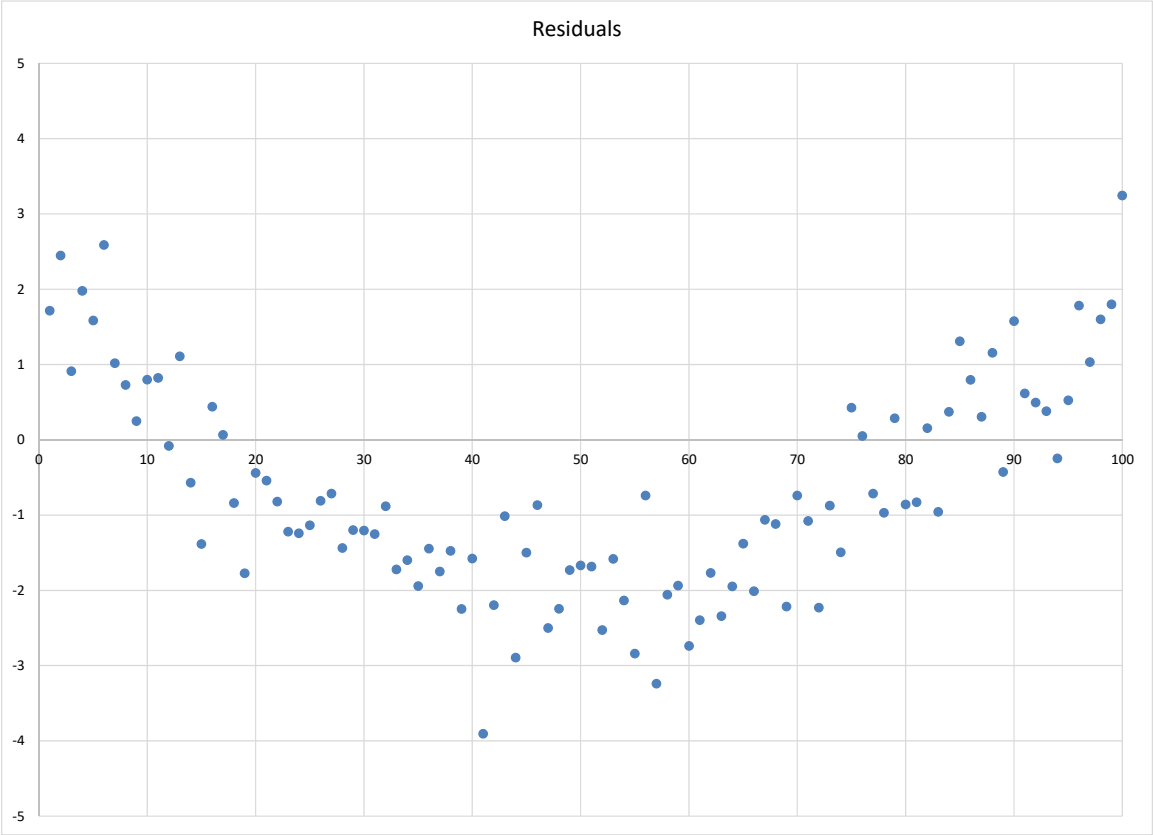
Breusch-Pagan test of heteroscedasticity

**Solution to Heteroscedasticity**

Heteroscedasticity is often caused by performing linear regression on non-linear data. Generally, solving non-linearity problems with transformations reduces or eliminates heteroscedasticity. If the problem is not completely resolved with a transformation, additional advanced techniques including Huber regression can correct lingering issues.

**Assumption #3b: The error terms are not correlated (Serial Correlation)**

When dealing with data over time, it's possible for the error terms from one time period to be highly correlated with the previous time period. This is called serial correlation. The error terms or residuals will have a pattern that is not random, such as in the picture below.




**Test for Serial Correlation**

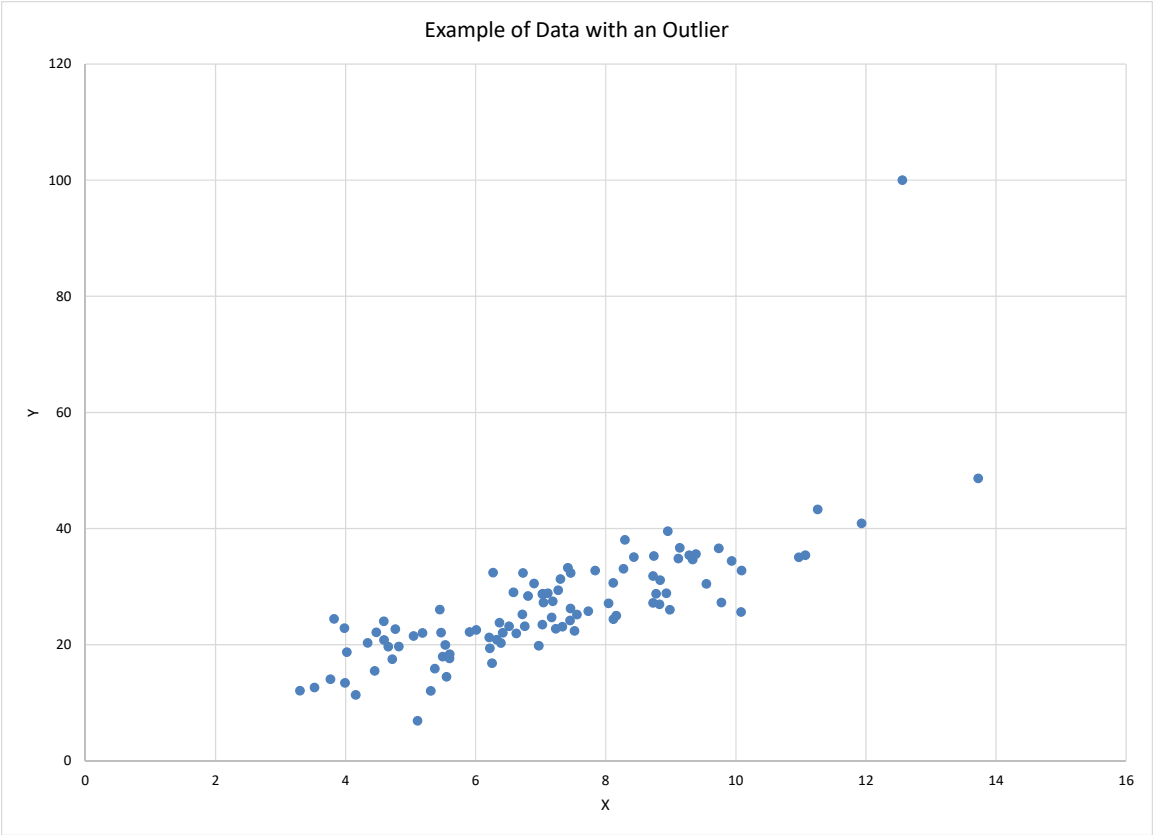
Durbin-Watson test of serial correlation

**Solution to Serial Correlation**

To correct for serial correlation there are a number of techniques in time series, including rho differencing and ARCH.

**Assumption #3c: There are no outliers**

An outlier is a data point that is significantly different from other data points. Outliers are often the result of unusual circumstances or data entry errors. The data below has an outlier.




**Test for Outliers**

Bonferroni outlier test

**Solution to Outliers**

If the data point is clearly an outlier, you can drop the bad data point, but mention in your analysis that you dropped outliers.

**Download Datasets**

To access some excellent data sets used in the book "Data Mining and Business Analytics with R," by Johannes Ledolter:

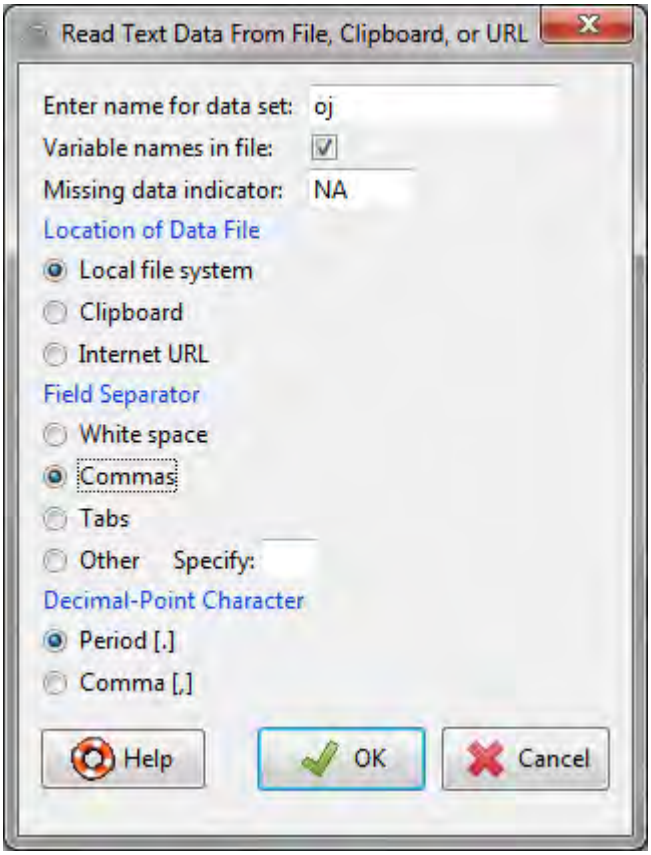
1. Go to the website:
   http://www.biz.uiowa.edu/faculty/jledolter/DataMining
2. Click on Data Text
3. Right click on oj.csv, then save on your computer
4. Remember where you saved the file

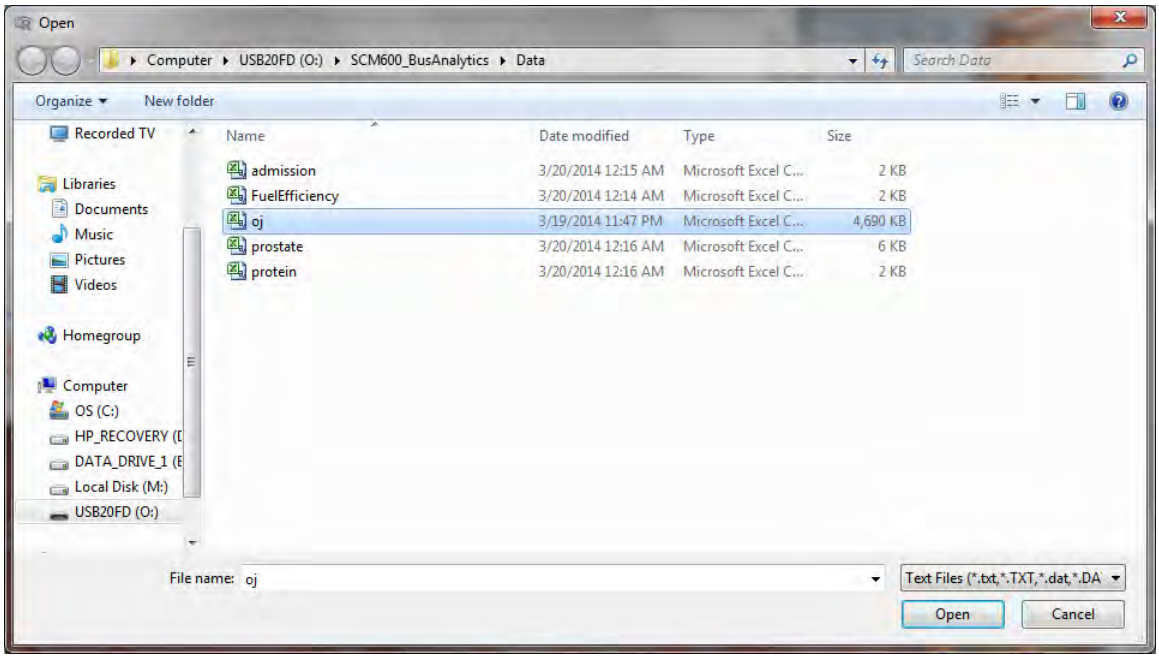
The Business Analytics - Week 8 oj.csv (orange juice) file can be downloaded from the course website.

**Loading Data**

To load data into R:

1. Click on Data at the top of the screen
2. Click on Import Data > From text file …
3. Enter the name that you would like to use for this data set; type in oj
4. Change Field Separator to Commas, then OK
5. Click on the oj file, then Open

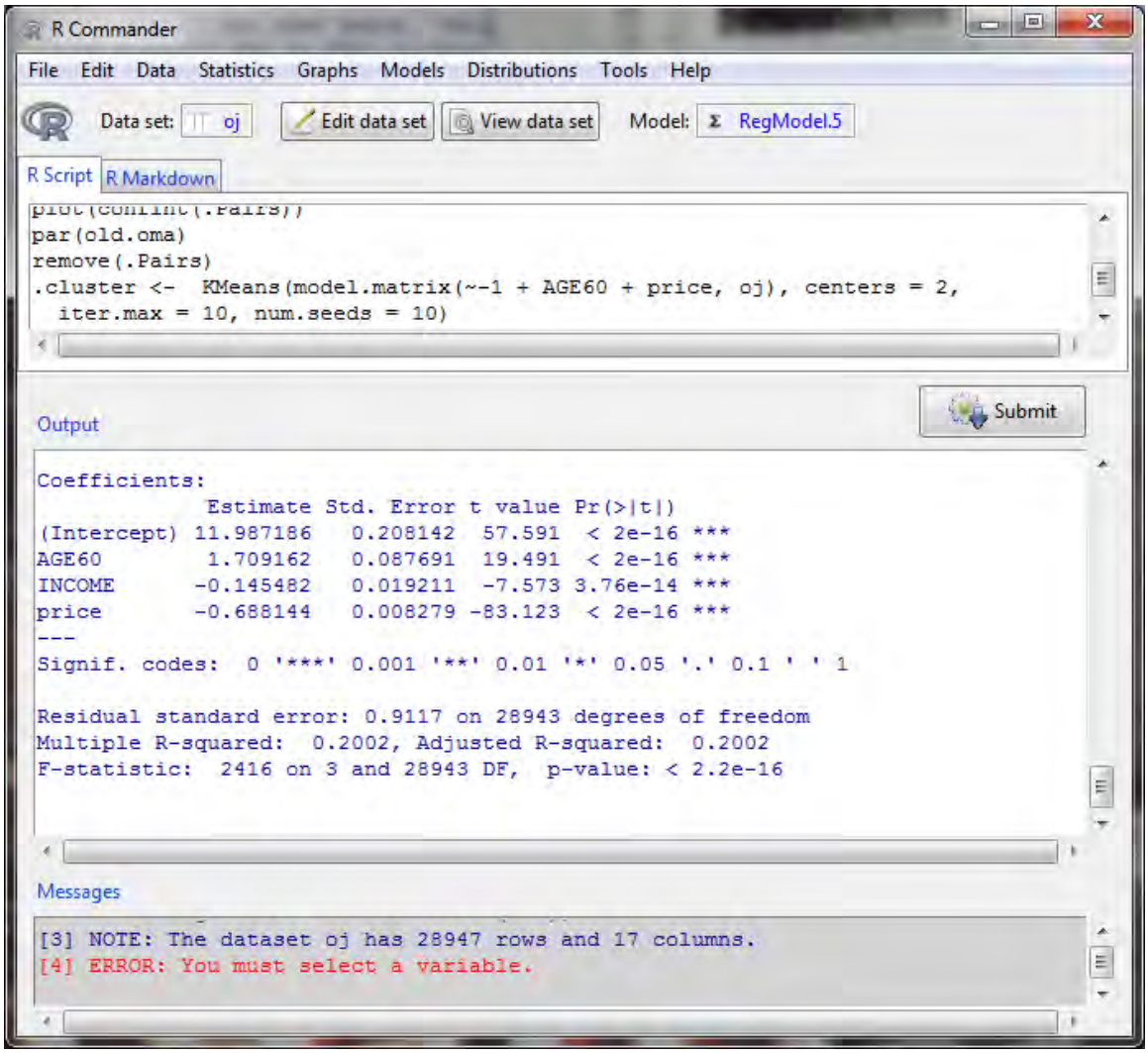



Open
Computer ▸ USB20FD (O:) ▸ SCM600_BusAnalytics ▸ Data
Search Data
Organize
New folder
Recorded TV
Libraries
Documents
Music
Pictures
Videos
Homegroup
Computer
OS (C:)
HP_RECOVERY (D
DATA_DRIVE_1 (E
Local Disk (M:)
USB20FD (O:)
Name
Date modified
Type
Size
admission 3/20/2014 12:15 AM Microsoft Excel C... 2 KB
FuelEfficiency 3/20/2014 12:14 AM Microsoft Excel C... 2 KB
oj 3/19/2014 11:47 PM Microsoft Excel C... 4,690 KB
prostate 3/20/2014 12:16 AM Microsoft Excel C... 6 KB
protein 3/20/2014 12:16 AM Microsoft Excel C... 2 KB
File name: oj
Text Files (*.txt,*.TXT,*.dat,*.DA
Open
Cancel

**Session 8.4: Regression and the RESET test for linearity**

Linear regression of the log of sales against age, income and price can be performed by:

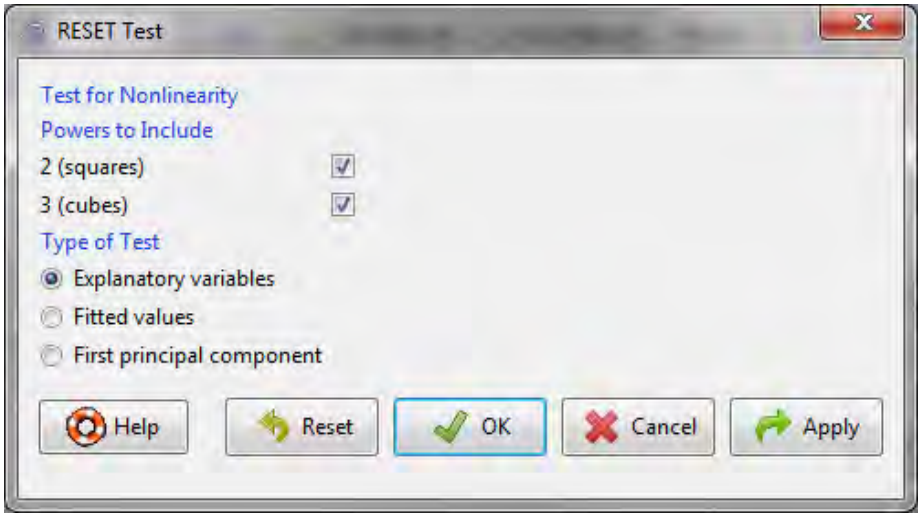
1. Click on Statistics, Fit Models, Linear Regression
2. For response variable, click on logmove
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
4. Click OK

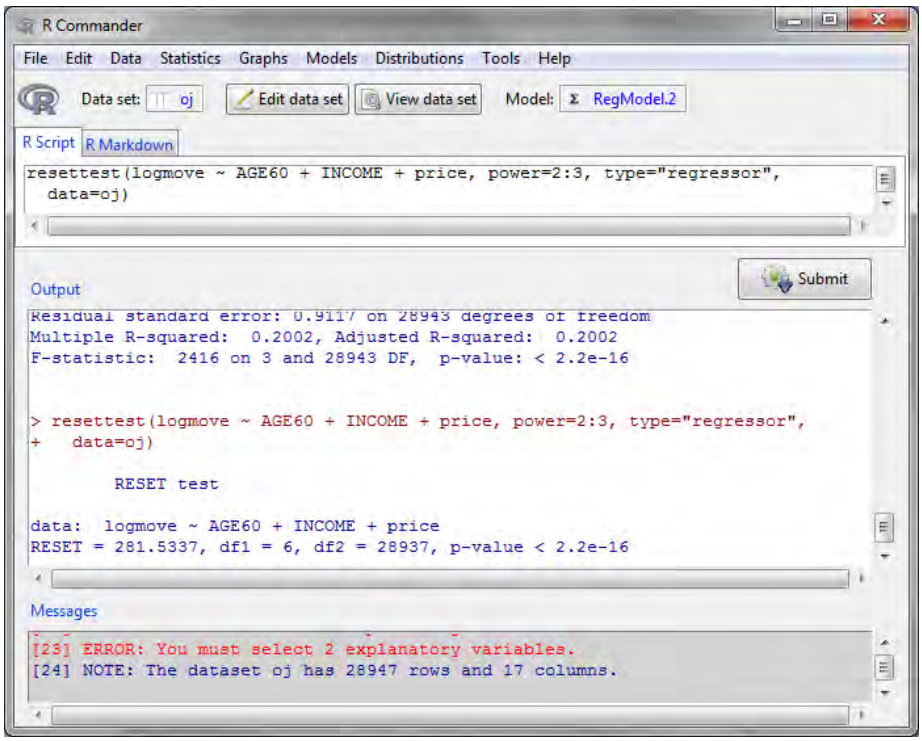


**Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity**

To test if your equation is linear:

1. Click on Models, Numerical Diagnostics, RESET test for Non-linearity
2. Click OK



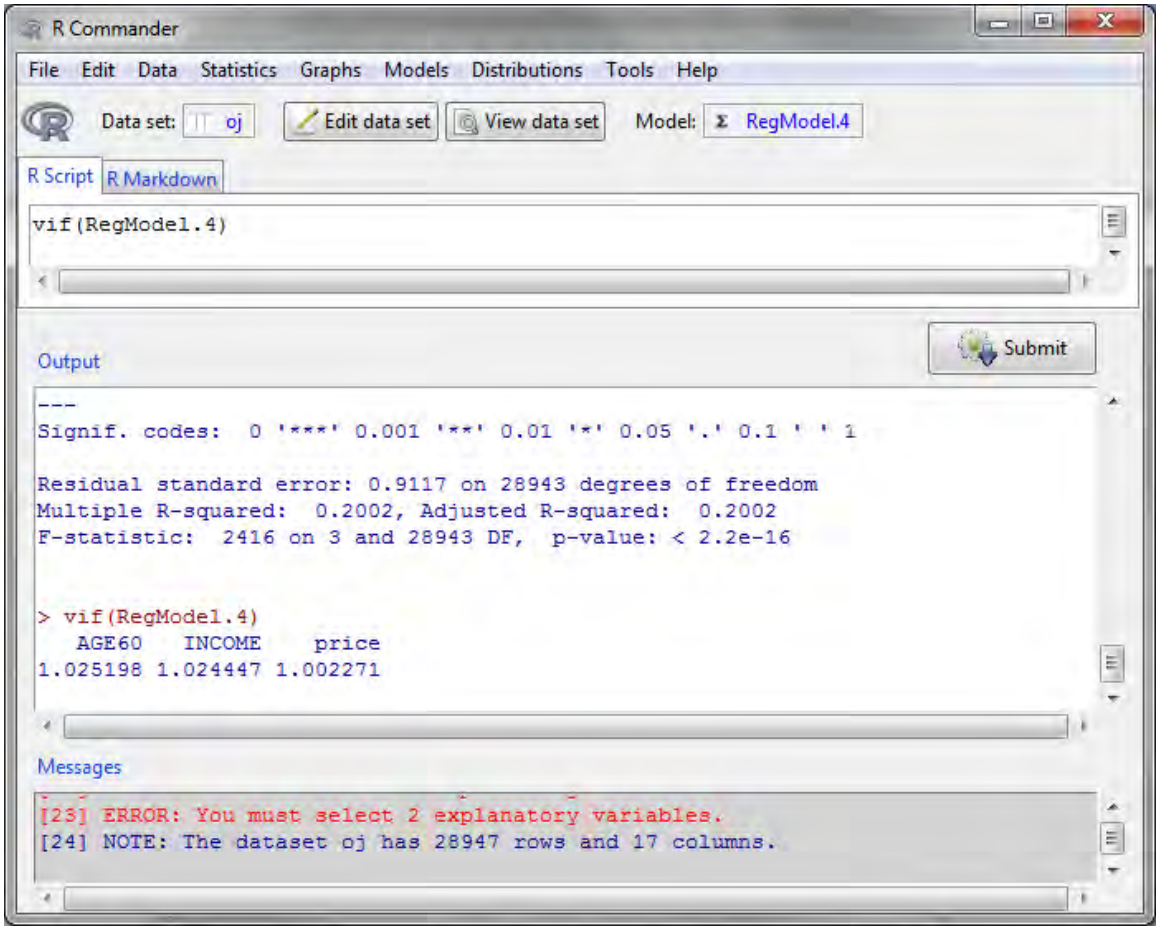




If the p-value is less than 0.05, then there is a non-linearity problem.

**Session 8.5: Variance Inflation Factor test of correlated explanatory variables**

To calculate the Variance Inflation Factor:

1. Click on Models, Numerical Diagnostics, Variance Inflation Factor

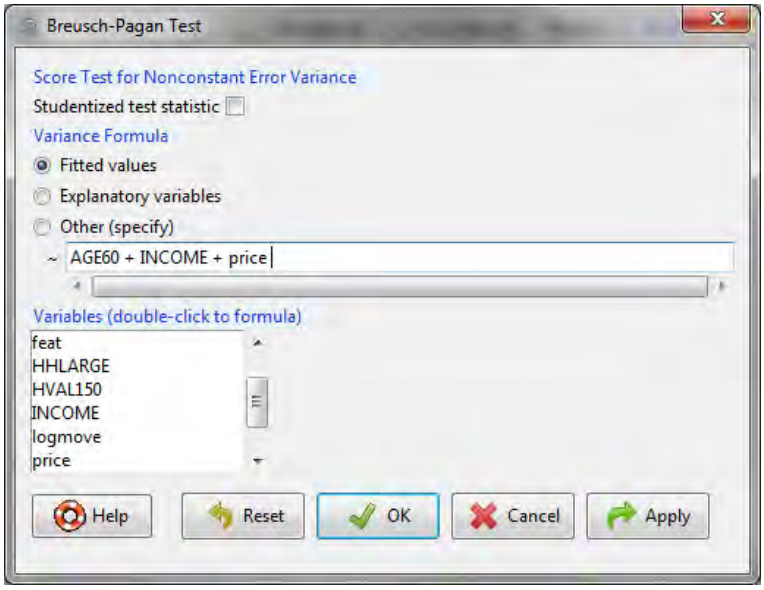


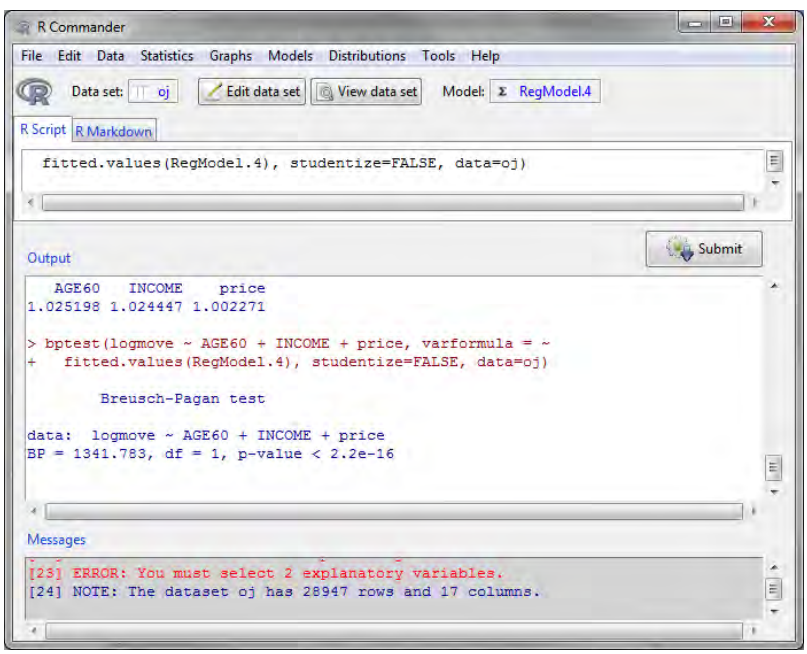

If the variance inflation factors are less than 10, then there is no multi-collinearity. If multi-collinearity exists, then drop variables or combine variables. Factor analysis is one technique for combining variables.

**Session 8.6: Breusch-Pagan test of heteroscedasticity**

Heteroscedasticity means that the error terms are vary depending on values of the explanatory variables. To test for heteroscedasticity:

1. Click on Models, Numerical Diagnostics, Breusch-Pagan test for heteroscedasticity
2. Double click on AGE60, INCOME, price
3. Click on OK



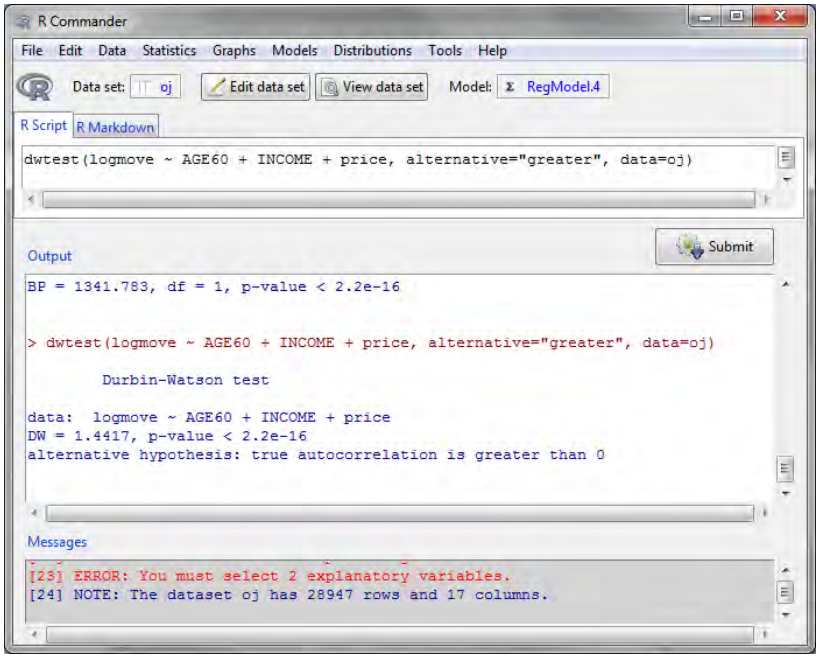




If the p-value is less than 0.05, then there is a problem with heteroscedasticity. Generally, this is a sign that the equation is non-linear.

**Session 8.7: Durbin-Watson test of serial correlation**

Serial correlation occurs when the errors terms are correlated. To test this,

1. Click on Models, Numerical Diagnostics, Durbin-Watson test for autocorrelation

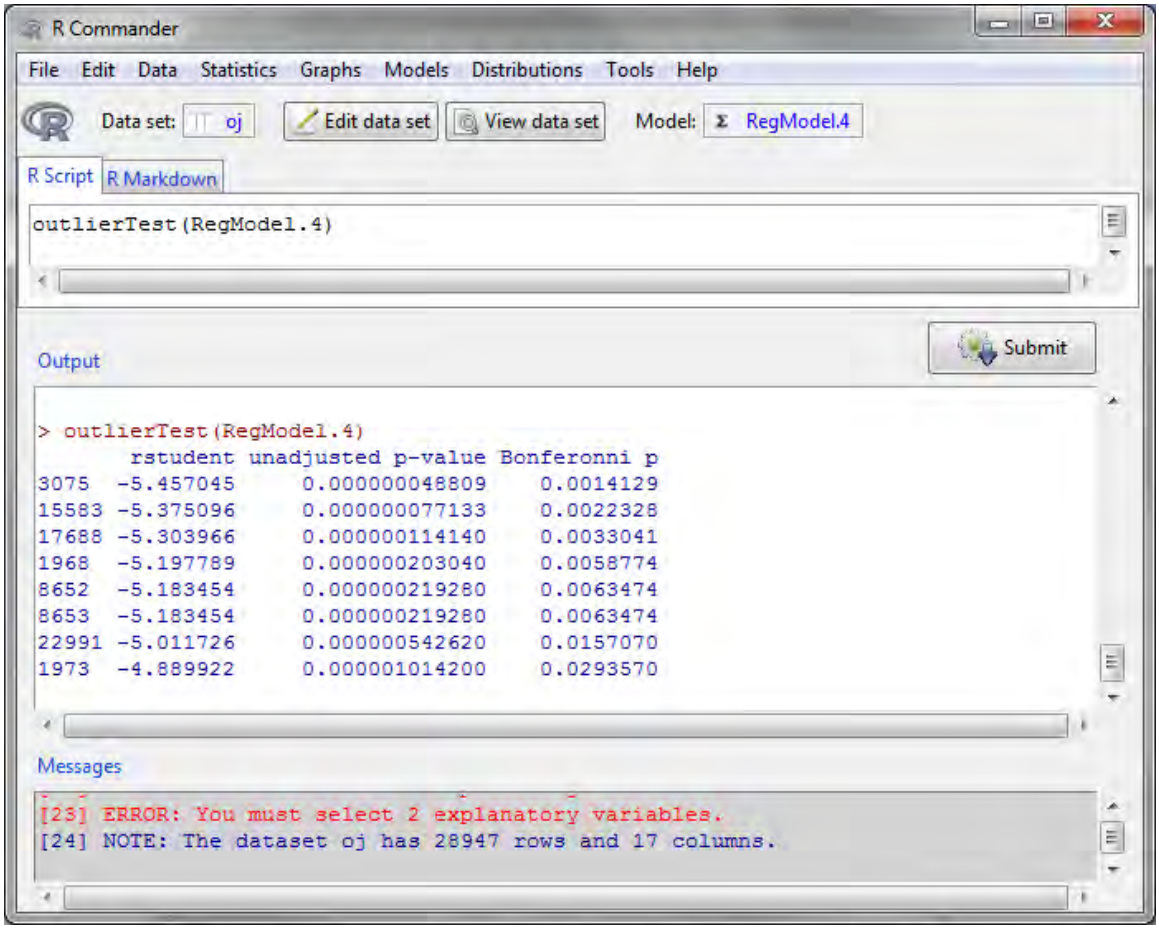



If the p-value is less than 0.05, there is a problem with serial correlation.

**Session 8.8: Bonferroni outlier test**

Outliers are extreme data points that can influence the results and lead to incorrect coefficients. To identify outliers,

1. Click on Models, Numerical Diagnostics, Bonferroni outlier test




In this example, data points numbered 3075, 15583, 17688, etc., are outliers. It's usually best to remove these data points from your data.

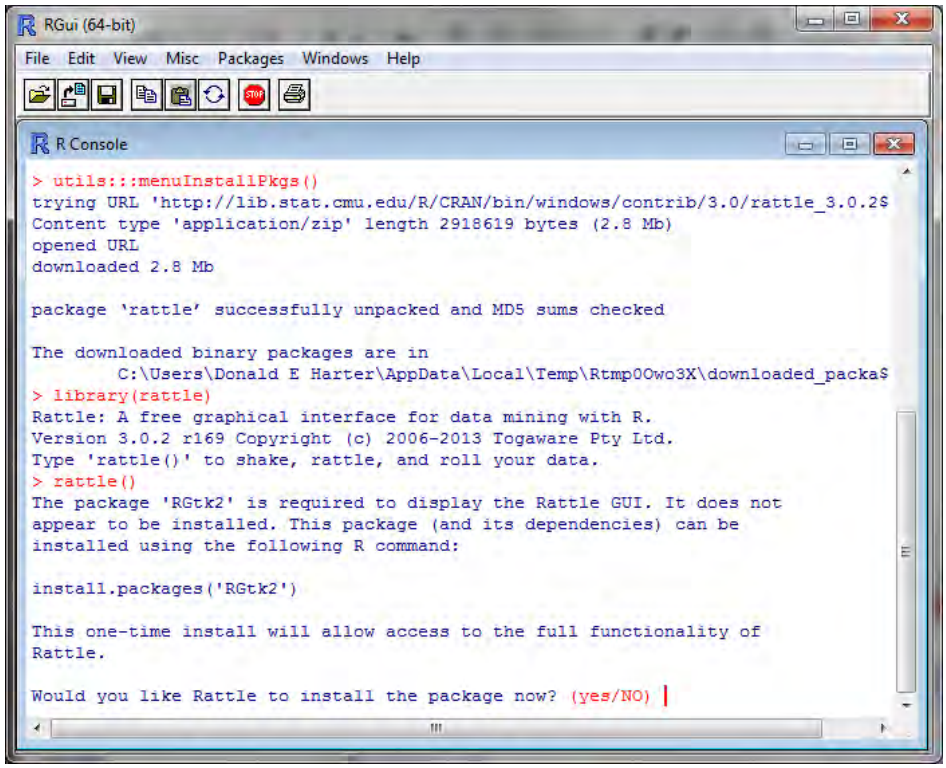
**Session 8.9: Data Mining**

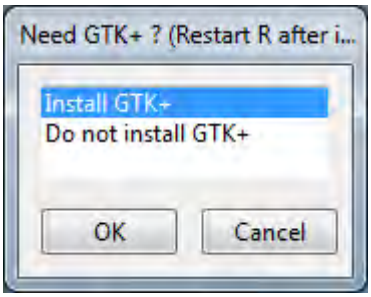
Data mining tools allow you to explore in more detail groupings of data and more sophisticated analysis. Rattle is an add-in to R that facilitates data mining.

**Installing Rattle**

Follow these steps only if you don't already have Rattle installed.

1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on rattle, then OK
5. Type library(rattle)
6. Type rattle()
7. When it asks "Would you like Rattle to install …", type yes
8. If you receive an error message about GTK+, then install GTK+ by clicking OK
9. If you receive an error message about XML, click Yes to install
10. Similarly for cairoDevice



### Launch Rattle

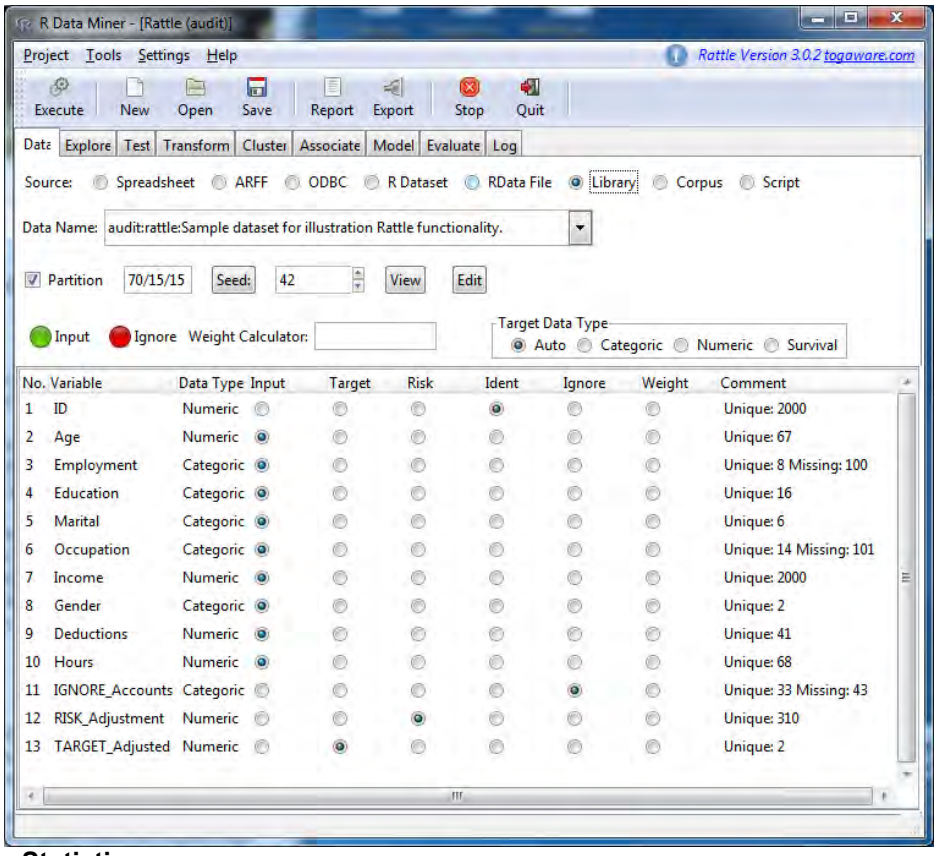
After Rattle has already been installed, you can always launch Rattle by:

1. Type library(rattle)
2. Type rattle(), press enter, then type yes, press enter

### Loading Data

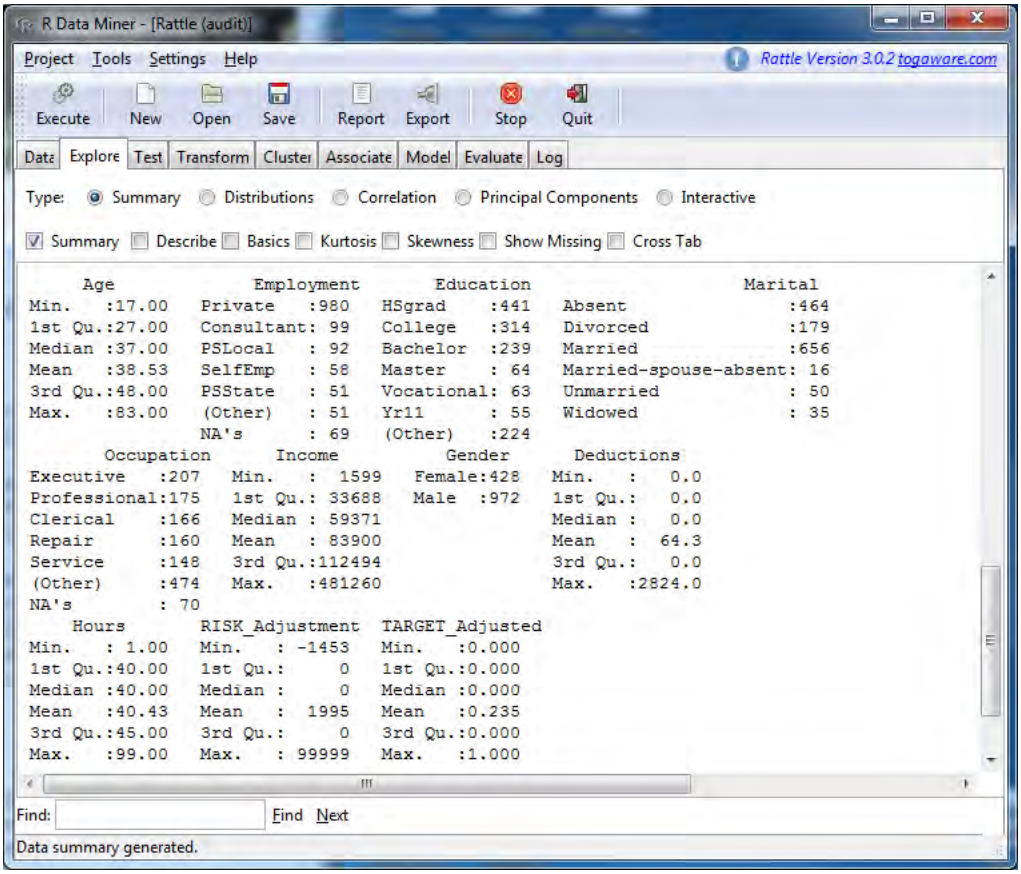
The package R has some built in data sets. To load data into R:

1. In the R Data Miner [Rattle] window, click on the Data tab
2. Check that Source indicates Library
3. Next to Data Name, use the drop down menu to select audit: rattle: Sample dataset
4. Click on Execute
5. This data set represents income tax audit data




### Summary Statistics

1. Click on the tab Explore
2. Check the radio button Summary
3. Click on Execute


R Data Miner - [Rattle (audit)]
Project Tools Settings Help
Rattle Version 3.0.2 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Summary Distributions Correlation Principal Components Interactive
Summary Describe Basics Kurtosis Skewness Show Missing Cross Tab
Age Employment Education Marital
Min. :17.00 Private :980 HSgrad :441 Absent :464
1st Qu.:27.00 Consultant: 99 College :314 Divorced :179
Median :37.00 PSLocal : 92 Bachelor :239 Married :656
Mean :38.53 SelfEmp : 58 Master : 64 Married-spouse-absent: 16
3rd Qu.:48.00 PSState : 51 Vocational: 63 Unmarried : 50
Max. :83.00 (Other) : 51 Yr11 : 55 Widowed : 35
NA's : 69 (Other) :224
Occupation Income Gender Deductions
Executive :207 Min. : 1599 Female:428 Min. : 0.0
Professional:175 1st Qu.: 33688 Male :972 1st Qu.: 0.0
Clerical :166 Median : 59371 Median : 0.0
Repair :160 Mean : 83900 Mean : 64.3
Service :148 3rd Qu.:112494 3rd Qu.: 0.0
(Other) :474 Max. :481260 Max. :2824.0
NA's : 70
Hours RISK_Adjustment TARGET_Adjusted
Min. : 1.00 Min. : -1453 Min. :0.000
1st Qu.:40.00 1st Qu.: 0 1st Qu.:0.000
Median :40.00 Median : 0 Median :0.000
Mean :40.43 Mean : 1995 Mean :0.235
3rd Qu.:45.00 3rd Qu.: 0 3rd Qu.:0.000
Max. :99.00 Max. : 99999 Max. :1.000
Find: Find Next
Data summary generated.

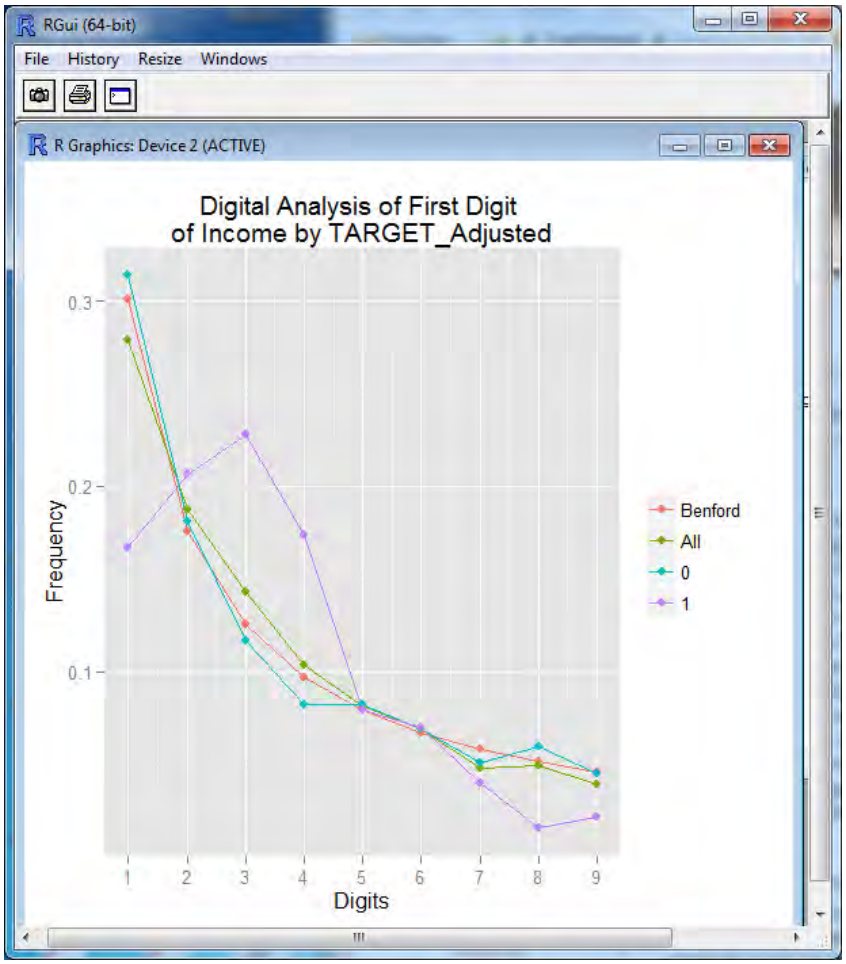
**Session 8.10: Benford's Law – Detecting Fraud with Data Mining**

In auditing (accounting, financial audits, tax audits), there is a rule called Benford's Law that specifies the frequency of the first digit in almost any financial number. For example, approximately 30% of financial numbers start with the digit 1.

The data set that was loaded describes 2000 income tax audits. To compare the result of income tax audits to Benford's Law:

1. Click on the tab Explore
2. Check the radio button Distributions
3. In the line for Income, check the box under Benford
4. Click Execute

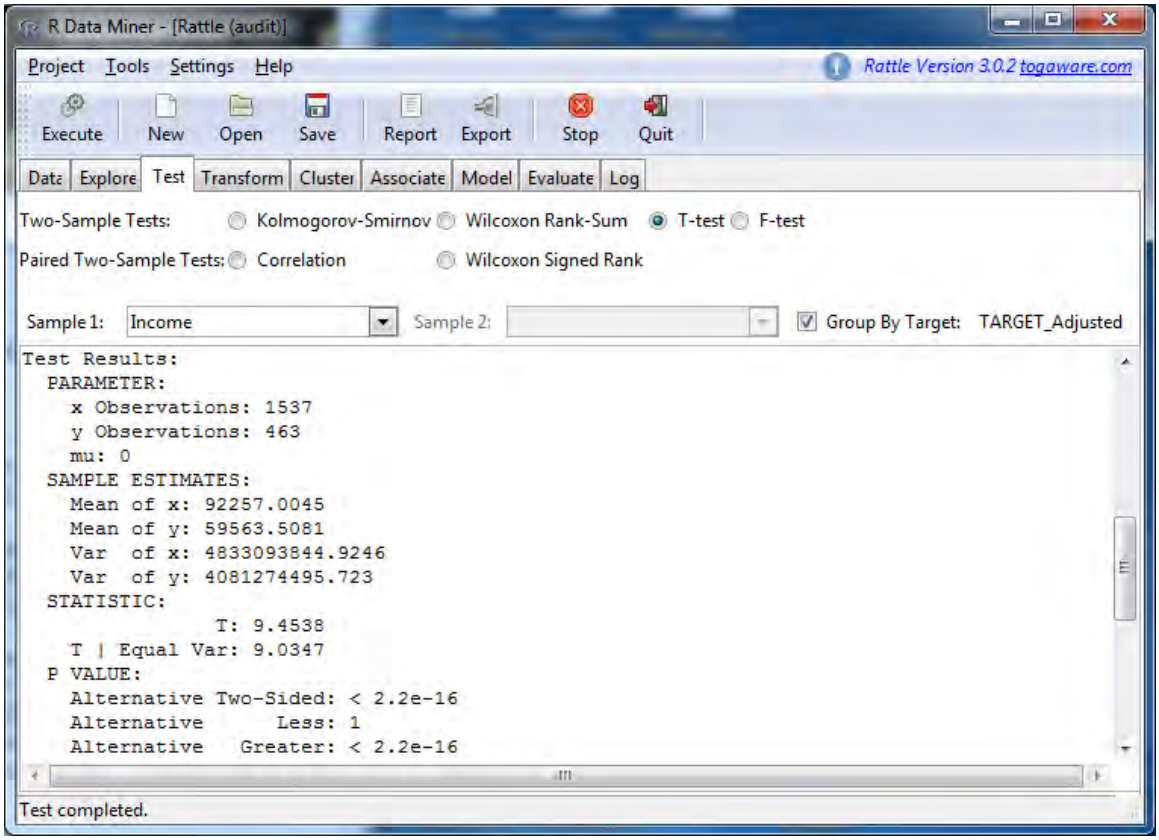
The Benford line is the expected frequency of the first digit of tax payers' income. The All line is the frequency of the first digit of all tax returns filed for 2000 people. The 1 line is the frequency of first digits of income for tax payers who were asked to fix their tax returns. The 0 line is those tax payers were not asked to fix tax returns. The 1 line departs significantly from the Benford line.



**Statistical Tests**

It appeared from the Benford curve that the distribution of tax violators (coded as 1) were different from non-violators. Let's test the means of the two distributions to see if they are different.

1. Click on the tab Test
2. For Two-Sample Tests, click on the radio button T-test
3. For Sample 1, use the drop down arrow to select Income
4. Click Execute



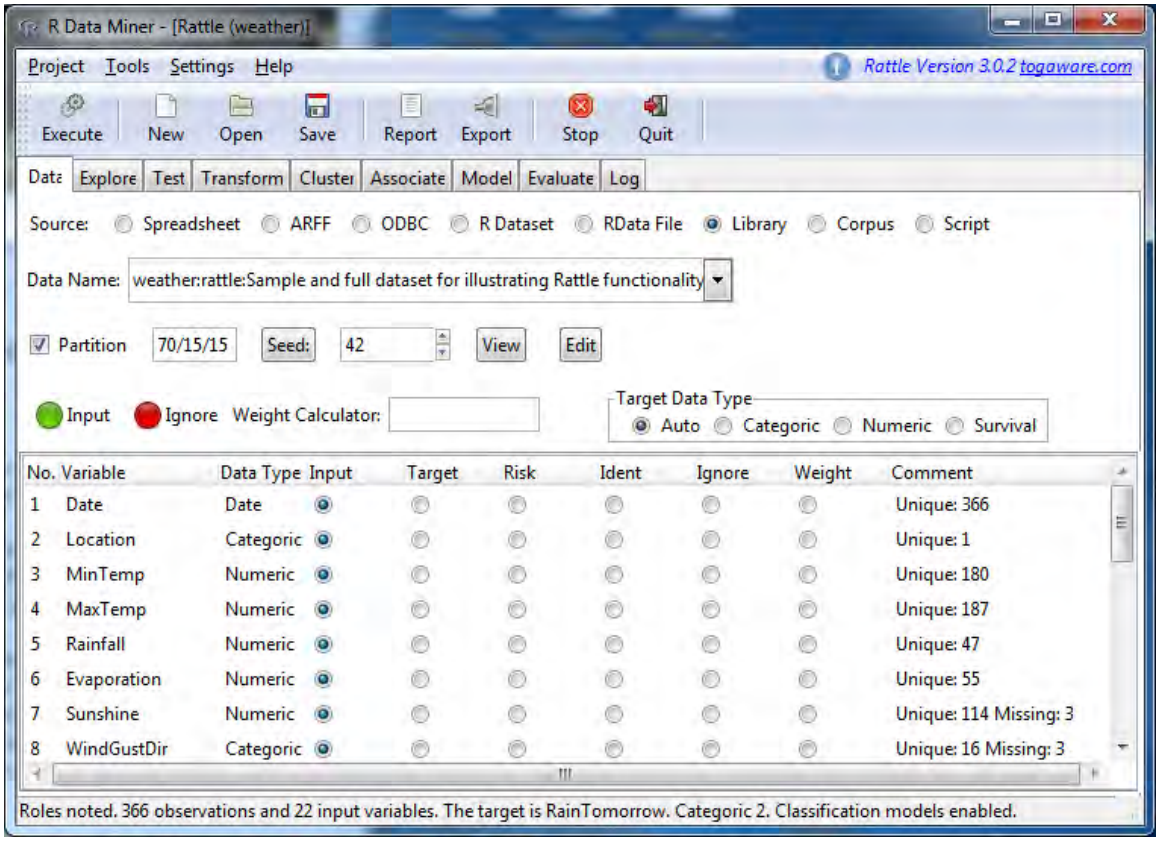

The X observation is for those coded as 0; Y observation for those coded 1. Look at the p-value for the test. Are the two groups different? What is the average income for each group? Which group appears to be misreporting their income more frequently? The higher or lower income group?

**Session 8:11: Decision Trees**

**Loading Data**

To load data into R:

1. In the R Data Miner [Rattle] window, click on the Data tab
2. Check that Source indicates Library
3. Next to Data Name, use the drop down menu to select weather: rattle: Sample dataset
4. Click Execute
5. This data set represents historical weather data

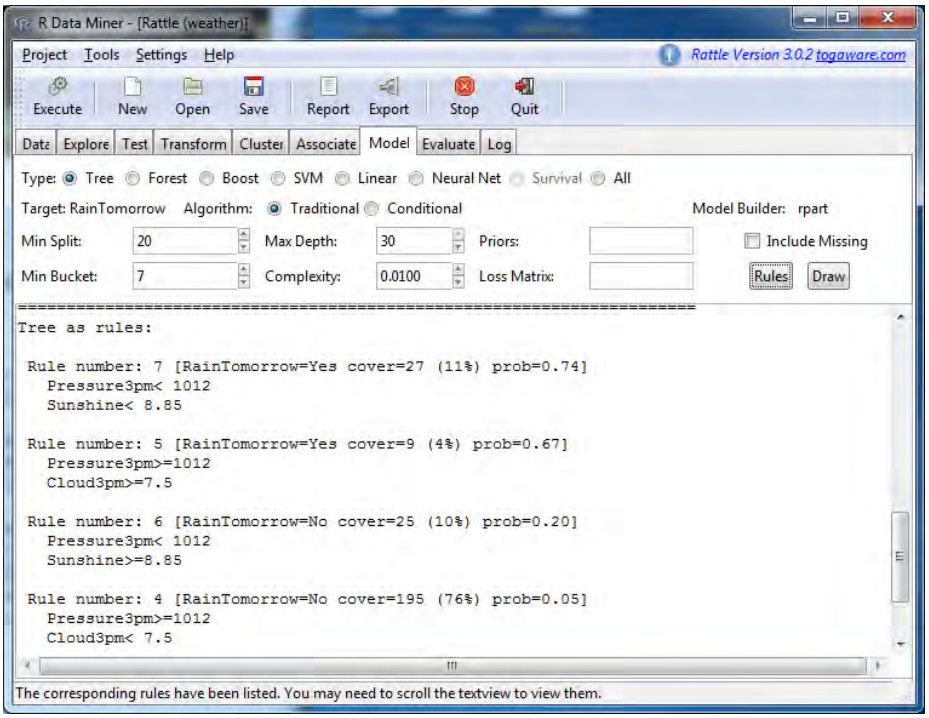



**Generating Decision Tree**

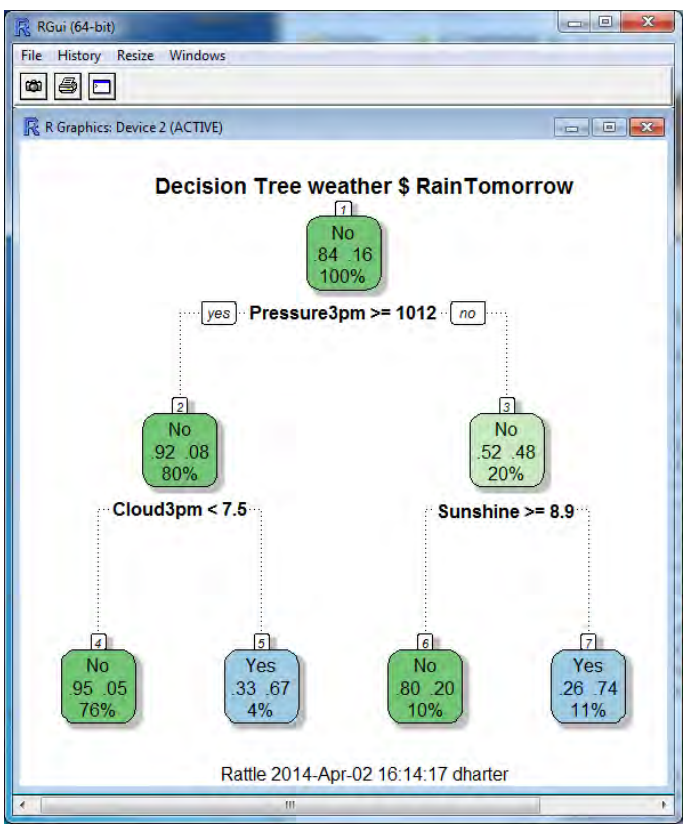
Create a decision tree to determine how to predict tomorrow's weather.

1. Click on the Model tab
2. Check the radio button Tree
3. Click Execute
4. Click on the button Draw

The resulting decision rules to predict rain tomorrow are:




The decision tree is:



# SCM 651: Business Analytics

1

Week 8

# Agenda



- Review of homework #3 (Regression and Optimization)
- Overview of homework #4 (Logit, Probit, Neural networks: info in week 9 videos)
- Review of hands-on exercises
- Group discussion of articles
  - What Businesses Can Learn from Sports Analytics?
  - Team GB: Using Analytics (and Intuition) to Improve Performance

# Homework #3



1. Graph, regression, calculated sales, revenue, profit
2. Constrained optimization
3. Discussion of risks, other data which would be valuable

# Homework #4



1. Logit and probit analysis (see week 9)
2. Moderating effects (week 7)
3. Final logit & probit models with interaction effects (moderating effects), prediction of outcome, sensitivity analysis
4. Neural network analysis
5. Neural network prediction model and sensitivity analysis (new material in handout in week 9)

# Week 8 - Review



- Regression Assumption #1: linearity
  - Violation: non-linear data
  - Solution: logarithm, square, inverse, other
- Regression Assumption #2: X variables are not correlated
  - Violation: multicollinearity
  - Solution: drop or combine variables

# Week 8 - Review



- Regression Assumption #3a: errors are random with constant variance
  - Violation: heteroscedasticity, or wedge shape to error terms in scatterplot
  - Solution: logarithm, square, inverse, or Huber regression
- Regression Assumption #3b: error terms are correlated
  - Violation: serial correlation
  - Solution: rho differencing
- Regression Assumption #3c: outliers
  - Violation: outlier influences slope of line
  - Solution: drop outlier data points

# Week 8 - Review



- Benford's law
  - Financially reported numbers tend to start with smaller digits
- Decision trees
  - Use entropy reduction to reduce the amount of error in the data to make a decision
  - Identify the most important variables in making a decision
  - Create a series of rules to make a decision

# Article #1: What Businesses Can Learn from Sports Analytics



- What Businesses Can Learn from Sports Analytics
  - Describe the five key lessons of analytics in sports (give an example of each)

# Article #1: What Businesses Can Learn from Sports Analytics



- What Businesses Can Learn from Sports Analytics
  - Describe the five key lessons of analytics in sports (give an example of each)
    - Align leadership at multiple levels
      - Player acquisition, player payment, strategies for performance
    - Focus on human dimension
      - Individual-level game performance
      - Performance in context (plus/minus analysis)
    - Exploit locational data
      - NYY player acquisition based on homerun measurement
    - Broader ecosystem (partnerships)
      - Business operations, dynamic ticket pricing, digital strategy
    - Support "analytic amateurs"
      - Players becoming analytics specialists

# Article #2: Team GB: Using Analytics (and Intuition) to Improve Performance



- Team GB: Using Analytics (and Intuition) to Improve Performance
  - What is the value of predicting team performance? (page 2)
  - What is the biggest challenge? (page 2)
  - What are some of the barriers? (page 3)
  - Where is the power of the data? (page 5)

# Article #2: Team GB: Using Analytics (and Intuition) to Improve Performance



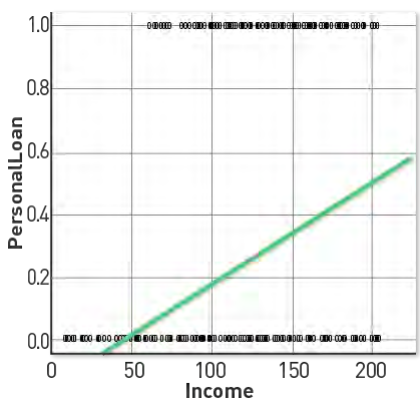
- Team GB: Using Analytics (and Intuition) to Improve Performance
  - What is the value of predicting team performance? (page 2)
    - Priorities: GB only funds sports which are likely to produce medals
  - What is the biggest challenge? (page 2)
    - Difficulty in collecting data – some sports are hard to collect
  - What are some of the barriers? (page 3)
    - Elite coaches rely on experience, rather than data
  - Where is the power of the data? (page 5)
    - Longitudinal data rather than snapshots

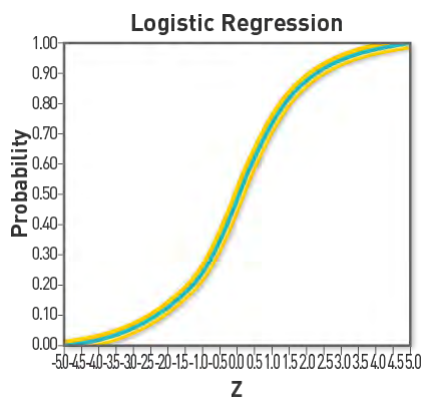
## Choice Models

- Choice models examine binary behavior or choices.
  - E.g., a customer walks into a store and either buys a TV or doesn't.
  - The Y variable is not a continuous number but is binary (i.e., zero or one).
- Choice models apply to:
  - Purchases
  - Elections of government candidates
  - Loan and credit approvals
  - Anything where there is a choice of yes or no, zero or one.



## A Problem with Choice Models




- Linear models don't fit.
- Logit (logistic) models do fit.

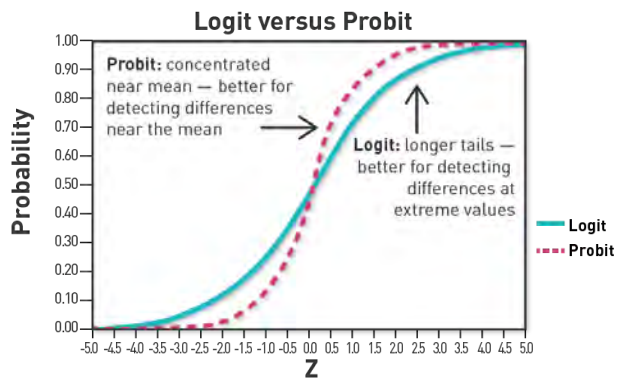
## Logistic Regression and Logit

- Logistic regression assumes that the Y variable is a probability, ranging from zero to one.
- The regression determines the likelihood that someone will make a purchase, vote for a candidate, etc.
- It relies on the logistic function:

$$P(Y = 1) = \exp(\sum \beta_i X_i) / [1 + \exp(\sum \beta_i X_i)]$$



## Logit vs. Probit



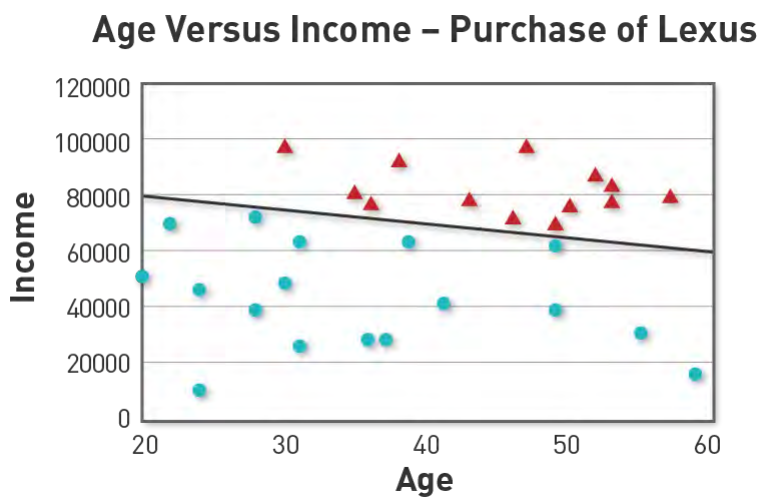

| Logit | Probit |
|---|---|
| • Relies on the logistic distribution.<br>• Is more sensitive to extreme values. | • Relies on the normal distribution.<br>• Is more sensitive to values near the mean. |

- Logit and probit usually produce similar results.

View All View Keyframes

## Perceptrons

- Perceptrons are an advanced decision-making tool, first developed in the 1960s.
- They focus on choices, where a straight line drawn through data can separate patterns.




- Perceptrons work well for some problems but not for others.

8 of 35

## Perceptron Example 1: 'And' Condition

- Consider two Lexus buyer variables: rich (or not rich) and young (or not young).

| | Rich | Not Rich |
|---|---|---|
| Young | X | O |
| Not Young | O | O |

| | Rich | Not Rich |
|---|---|---|
| Young | X | O |
| Not Young | O | O |

- This example uses an "And" condition: a buyer must be both rich and young.
- With the "And" condition, a straight line can divide the buyers from the nonbuyers.

18 of 35

## Perceptron Example 2: 'Or' Condition

| | Rich | Not Rich |
|---|---|---|
| Young | X | X |
| Not Young | X | O |

| | Rich | Not Rich |
|---|---|---|
| Young | X | X |
| Not Young | X | O |

- This example uses an "Or" condition: a buyer must be either rich or young.
- With the "Or" condition, a straight line can divide the buyers from the nonbuyers.

24 of 35

## Perceptron Example 3: 'Exclusive Or' Condition

| | Rich | Not Rich |
|---|---|---|
| Young | O | X |
| Not Young | X | O |

| | Rich | Not Rich |
|---|---|---|
| Young | O | X |
| Not Young | X | O |

- This example uses an "Exclusive Or" condition: a buyer must be either rich or young, but not both rich and young.
- With the "Exclusive Or" condition, a straight line cannot divide the buyers from the nonbuyers.

30 of 35

## Perceptrons to Neural Networks

- Although innovative, perceptrons could not solve the "Exclusive Or" problem (A or B but not both).
- It's failure was reliance on the straight line or linear function.
- Breaking the linearity constraint (i.e., allowing curved lines) solved the problem.
- Applying the logistic function curve led to the development of neural networks.

35 of 35

View All View Keyframes

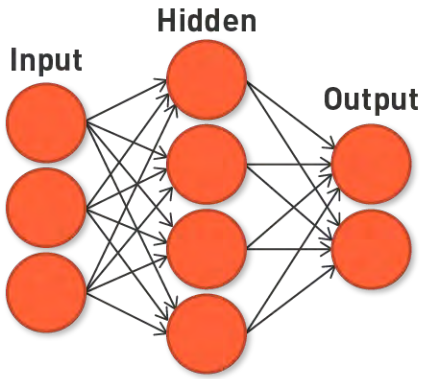
## Neural Networks

- Neural networks are a powerful technique to identify trends and patterns; they expand on the concept of perceptrons.
- Neural networks work in a way that is similar to the human brain.

3 of 53

## Neural Network Characteristics

- Neural networks use the logistic function to represent nonlinear behavior.
- A "hidden" layer of variables allows modeling of any mathematical representation.
- Neural networks are slow to learn but very effective in decision making.




9 of 53

## Neural Network Variables

- Inputs:
  - These are the X variables that would be included in a regression.
- Hidden nodes:
  - These are calculated by the neural network.
  - The user specifies how many hidden nodes to use.
  - More hidden nodes increase the ability of the neural network to find a pattern.
  - More hidden nodes also increase the time required to find a solution.

17 of 53

## Neural Network Variables (cont.)

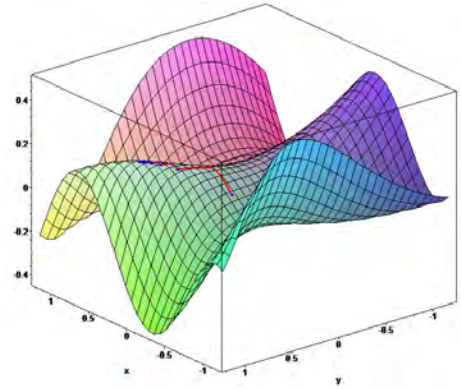
- Outputs:
  - These are the Y variables.
  - There can be more than one Y variable (e.g., "buy a TV" or "buy a cell phone").

20 of 53

## Neural Networks: Hill Climbing

- The neural network process is like

climbing a hill.
- From a given starting point: find the direction of maximum slope, take a step in that direction, then repeat.




25 of 53

## Neural Networks: Gradient Search

- Gradient search is the mathematical process of hill climbing.
- First, the neural network examines where it is in the landscape.
- Second, it calculates the slope in all directions, by taking the derivative of the logistic function, *f*(*x*):

$$f'(x) = f(x) * (1 - f(x))$$

- Third, it takes a step uphill, then repeats the process until it reaches the top of the hill.

32 of 53

## Problems With Gradient Search

- Local optima challenge:
  - The problem with gradient search, or hill climbing, is that you might be climbing a small hill, not the tallest hill in the area.
  - That means the process may find a solution that is suboptimal.
- Solution:
  - Run the neural network multiple times with different starting points in the landscape.
  - The neural network package in R randomly selects a new starting point each time.

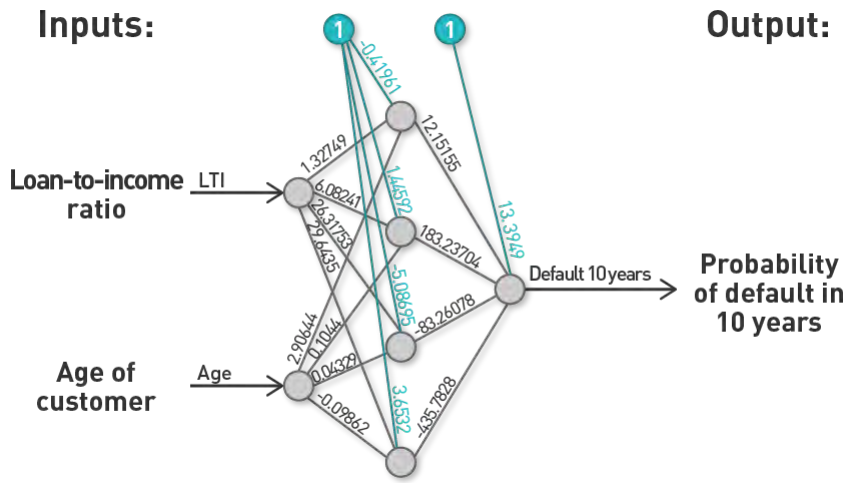
39 of 53

## Neural Network Implementation

| | |
|---|---|
| **Step 1:** | Collect data. |
| **Step 2:** | Identify input variables (i.e., the X variables). |
| **Step 3:** | Identify output variables (i.e., the Y variables). |
| **Step 4:** | Determine number of hidden nodes. |
| **Step 5:** | Run neural network. |

45 of 53

## Neural Network Solution Example

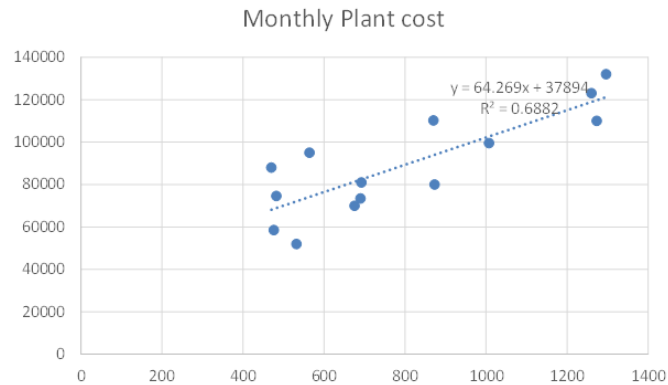

Inputs:
Output:
Loan-to-income ratio
LTI
Age of customer
Age
Default 10 years
Probability of default in 10 years
1
1
-0.41961
12.15155
1.32749
1.44592
13.3949
6.08241
26.31753
29.6435
183.23704
-5.08695
-83.26078
2.90644
0.1044
0.04329
-0.09862
3.6532
-435.7828

53 of 53

View All View Keyframes

## Regression: A Review

Our earlier regression estimated plant cost based on units:

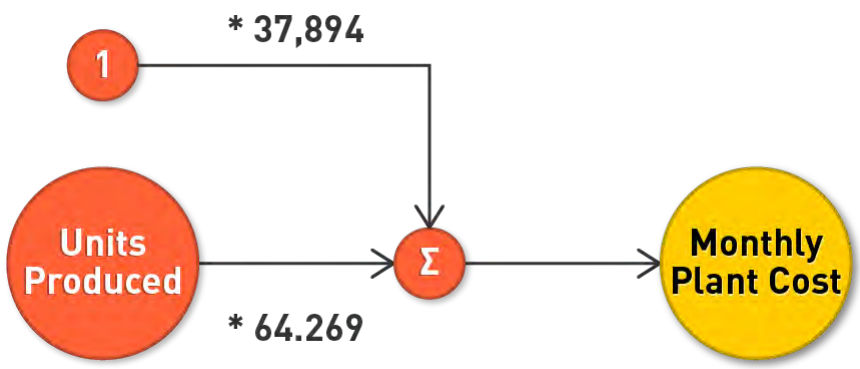



$$Monthly\ Plant\ Cost = 37{,}894 + 64.269 * Units\ Produced$$

4 of 33

## Another Way to View Regression: Example 1

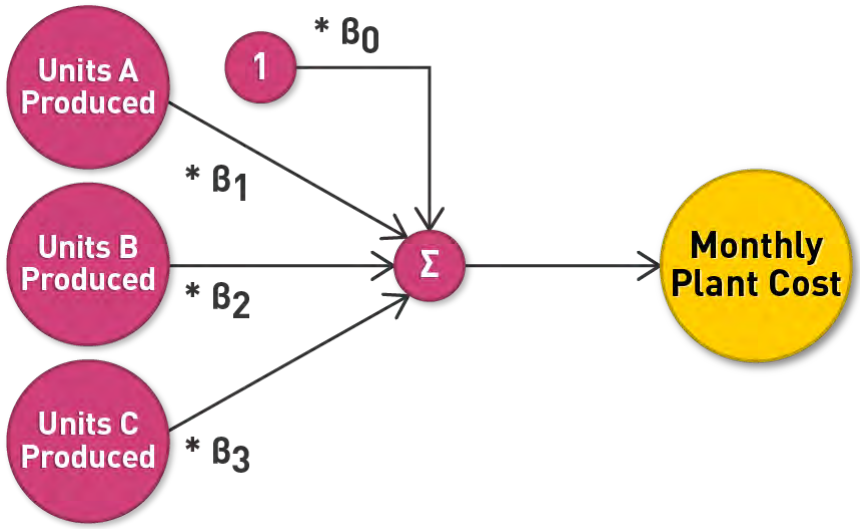
$$Monthly\ Plant\ Cost = 37{,}894 + 64.269 * Units\ Produced$$




11 of 33

## Another Way to View Regression: Example 2

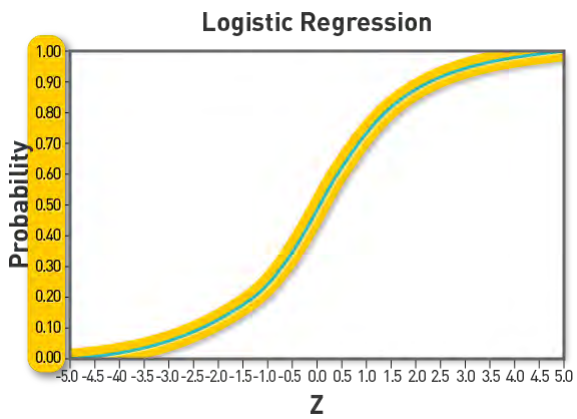
$$Monthly\ Plant\ Cost = \beta_0 + \beta_1 * A\ produced + \beta_2 * B\ produced + \beta_3 * C\ produced$$




18 of 33

## Logit: A Different Perspective

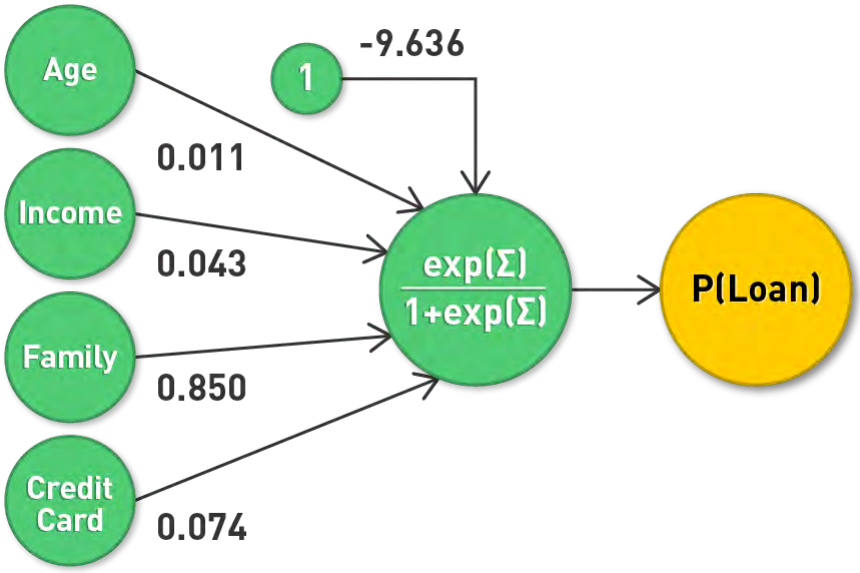
- Let's predict a loan approval based on:
  - Age
  - Income
  - Family size
  - Credit card average balance
- This is a choice model, so we'll use logit.




27 of 33

## Another Way to View Logit

$$P(Y=1) = \frac{\exp(-9.636 + 0.011 * Age + 0.043 * Income + 0.850 * Family + 0.074 * CCAvg)}{[1 + \exp(-9.636 + 0.011 * Age + 0.043 * Income + 0.850 * Family + 0.074 * CCAvg)]}$$


Age
1
-9.636
0.011
Income
0.043
Family
0.850
Credit Card
0.074
exp(Σ)
1+exp(Σ)
P(Loan)

33 of 33

# SCM 651: Business Analytics

1

Week 9

# Agenda



- Overview of homework #4 (Logit, Probit, Neural networks: info in week 9 videos)
- Review of hands-on exercises
- Group discussion of articles
  - An introduction to data mining and other techniques for advanced analytics

# Homework #4



1. Logit and probit analysis (see week 9)
2. Moderating effects (week 7)
3. Final logit & probit models with interaction effects (moderating effects), prediction of outcome, sensitivity analysis
4. Neural network analysis
5. Neural network prediction model and sensitivity analysis (new material in handout in week 9)

# Week 9 - Review



- Logit & Probit
  - Predict probabilities
- Logit
  - Logistic distribution
  - More sensitive in detecting differences at extreme values of your variables
- Probit
  - Normal distribution
  - More sensitive in detecting differences at values near the mean of your variables

# Week 9 - Review



- Neural networks
  - Uses the logistic function to build relationships
  - Also has at least three levels, the X input variables, one or more hidden layers of variables, and the Y output variables
  - To predict the neural network outcome:
    - First, predict the hidden variables from the inputs, just like a logit prediction
    - Second, predict the Y output variables from the hidden variables, again like a logit prediction

# Week 9 – Neural Network Prediction and Sensitivity Analysis



- In-class example
  - Logit analysis of Titanic survivor data
  - Creation of prediction model of logit results
  - Sensitivity analysis of logit results
  - Neural network analysis of Titanic survivor data
  - Creation of a prediction model of neural network results
  - Sensitivity analysis of neural network results

# Article #1: An Introduction to Data Mining and Other Techniques for Advanced Analytics



- An introduction to data mining and other techniques for advanced analytics
  - What are the key differences between statistical analysis and data mining? (page 140)
  - Describe tools for advanced analytics (page 149-151)
    - Data visualization
    - Text mining
    - Social network analysis
    - Contact optimization
  - How do you mitigate the risks of data mining? (page 152)

# Article #1: An Introduction to Data Mining and Other Techniques for Advanced Analytics



- An introduction to data mining and other techniques for advanced analytics
  - What are the key differences between statistical analysis and data mining? (page 140)
    - Creation of a hold-out sample
    - Use the hold-out sample to test the model
  - Describe tools for advanced analytics (page 149-151)
    - Data visualization: scatter plots and heat maps, geographic data
    - Text mining: extract structure from unstructured text files
    - Social network analysis: identify networks of calling circles, influencers
    - Contact optimization: best solution for customers calling in and marketers calling out

# Article #1: An Introduction to Data Mining and Other Techniques for Advanced Analytics



- An introduction to data mining and other techniques for advanced analytics
  - How do you mitigate the risks of data mining? (page 152)
    - Focus on good data quality, strong business focus, sound user training

# R: Advanced

**Datafiles**

For these exercises, download the files:
“Business Analytics – Week 9 Instructions.doc”
“Business Analytics – Week 9 Universal Bank.xls”
“Business Analytics – Week 9 creditset.xls”

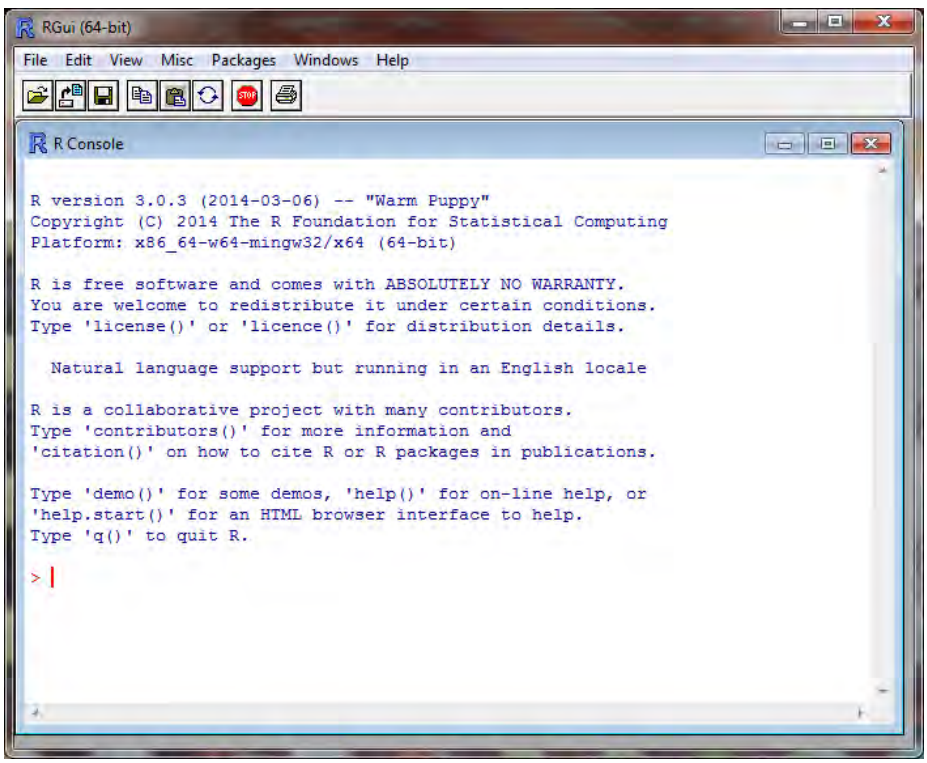
**Installation of R**

R is a free downloadable package capable of performing sophisticated statistical analysis and data mining. The software is already installed on the classroom laptops. To install on your own personal computer:

1. Go to the website: http://cran.r-project.org/bin/windows/base/
2. For a Mac, go to http://cran.r-project.org/bin/macosx/
3. Click on Download R 3.0.3 for Windows
4. Click on Run, and follow the install instructions

**Starting R**

1. Click on the Start button in the lower left corner of Windows
2. Click on All Programs, then click on the R folder, then R

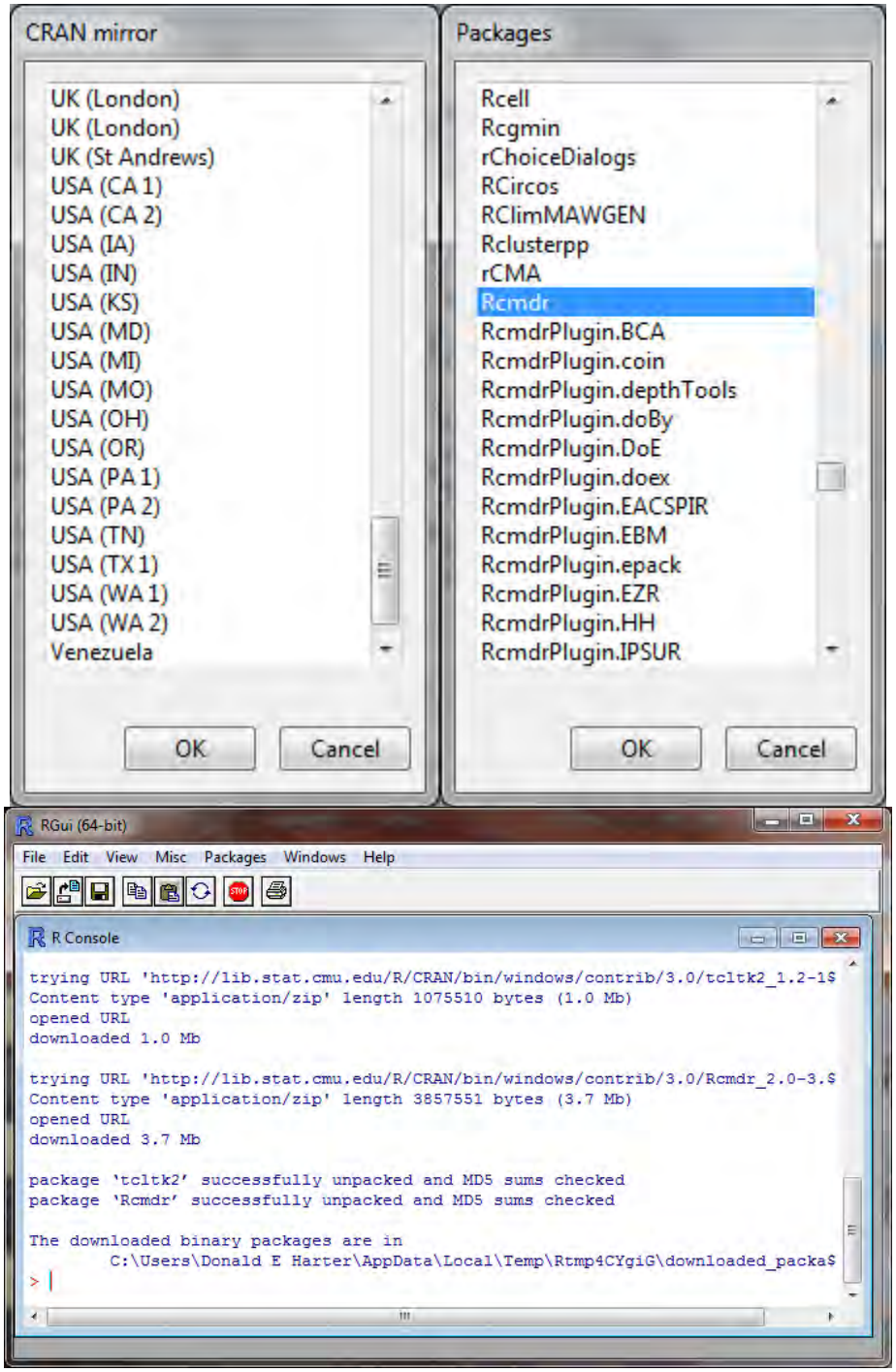



This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

**Installing R Commander**

Follow these steps only if you don't already have Rcmdr installed.

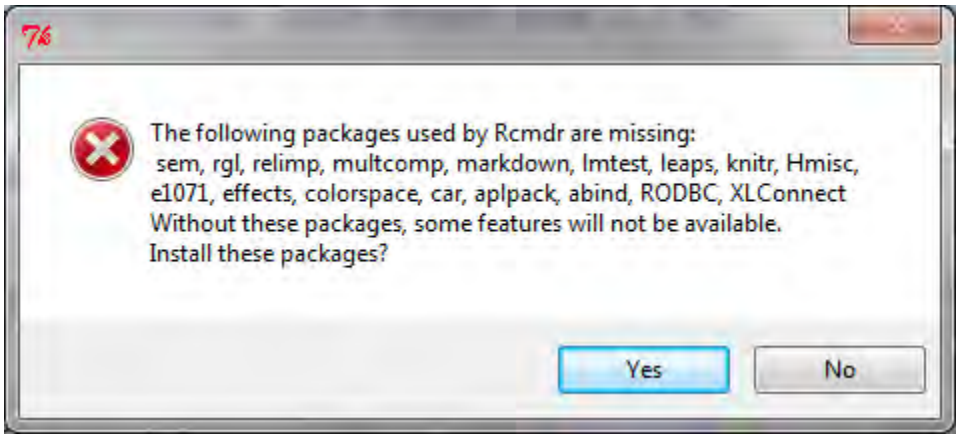
1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on Rcmdr, then OK
5. When prompted to create a personal library, click Yes

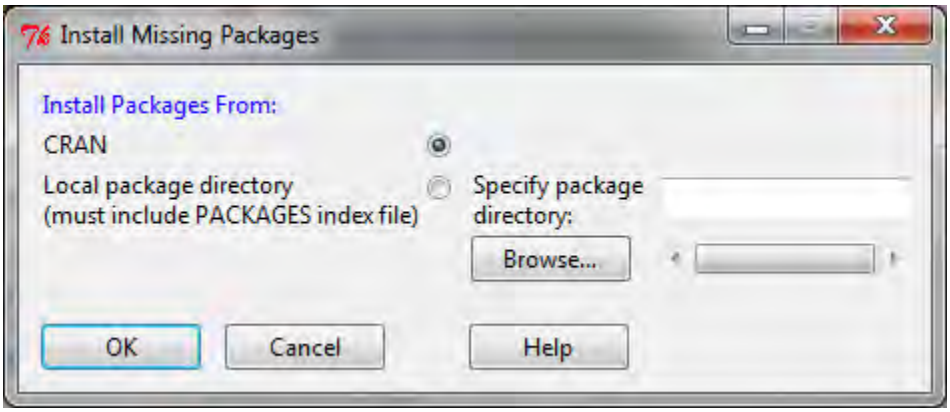


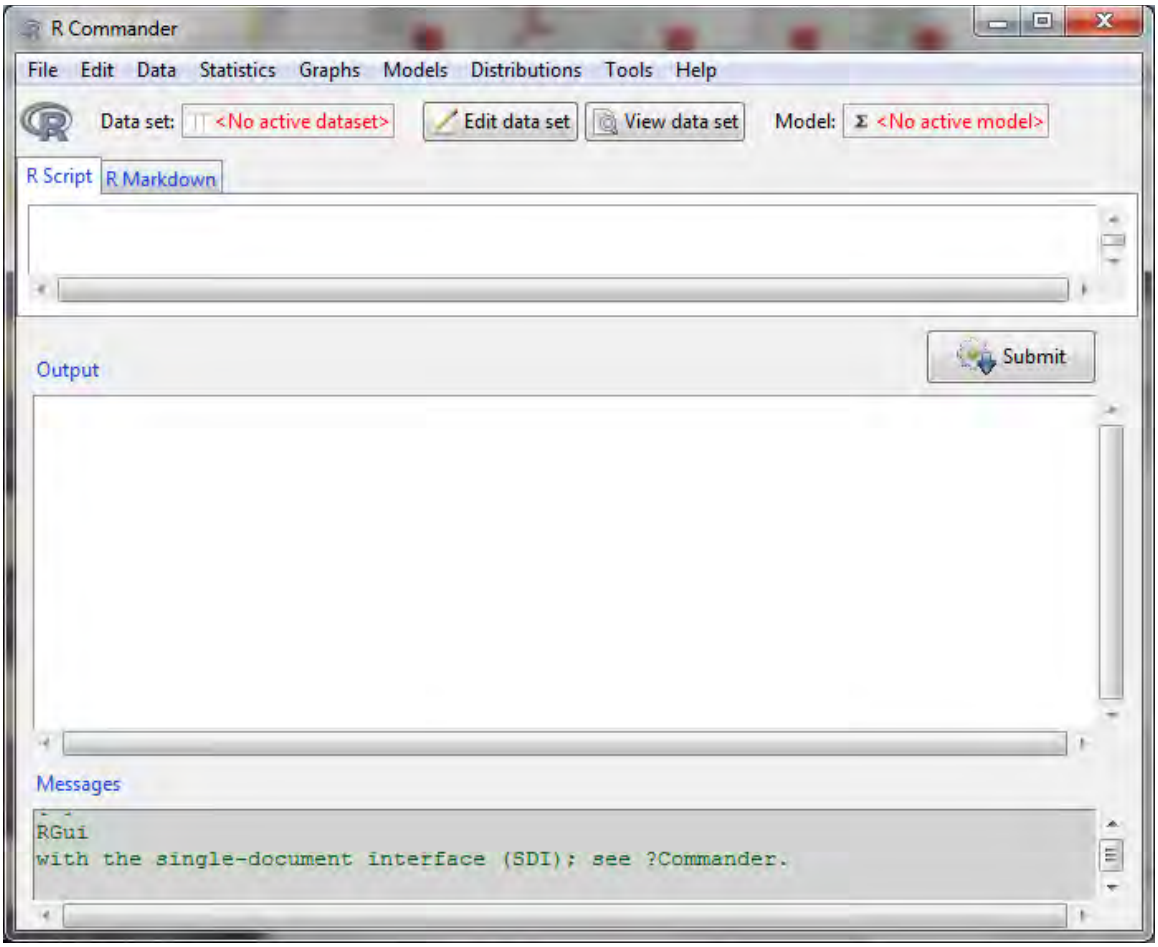
**Launch Rcmdr (R Commander)**

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type library(Rcmdr)
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK
4. R will install the necessary software
5. The R Commander screen will appear




R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: <No active dataset>
Edit data set
View data set
Model: <No active model>
R Script
R Markdown
Output
Submit
Messages
RGui
with the single-document interface (SDI); see ?Commander.

**Session 9.4: Logit Analysis - Download Datasets**

To access some excellent data sets used in the book "Data Mining and Business Analytics with R," by Johannes Ledolter:

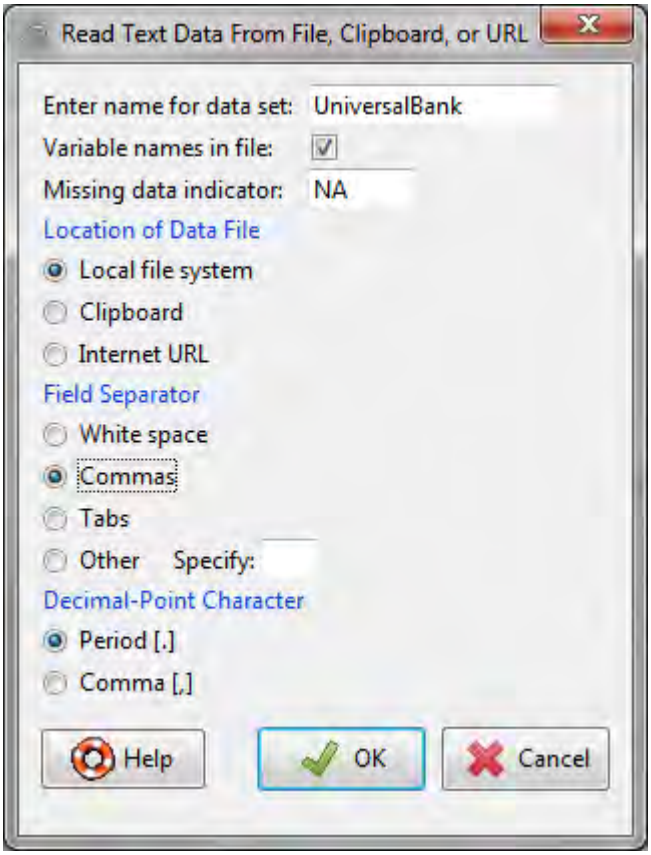
1. Go to the website:
   http://www.biz.uiowa.edu/faculty/jledolter/DataMining
2. Click on Data Text
3. Right click on UniversalBank.csv, then save on your computer
4. Remember where you saved the file

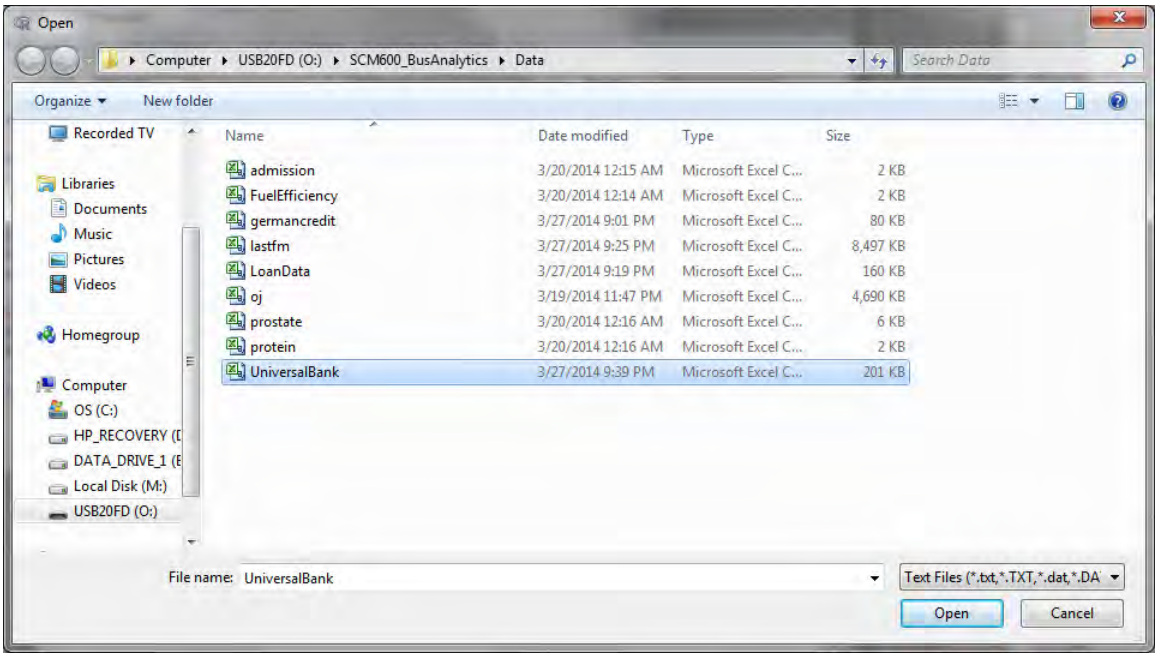
The Business Analytics - Week 9 Universal Bank.csv file can be downloaded from the course website.

**Loading Data**

To load data into R:

1. Click on Data at the top of the screen
2. Click on Import Data > From text file …
3. Enter the name that you would like to use for this data set; type in UniversalBank
4. Change Field Separator to Commas, then OK
5. Click on the UniversalBank file, then Open



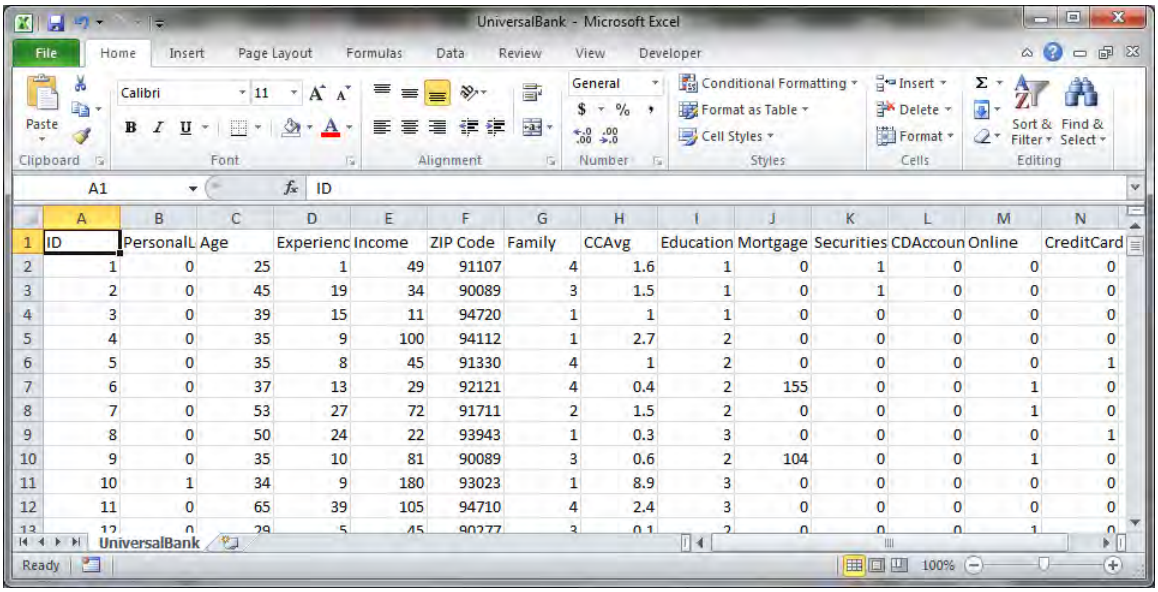



Note that the dataset UniversalBank has 5000 rows and 14 columns.

**Viewing data fields**

This data set lists loan characteristics for 5000 loan applications. Let's view the data. The easiest way to view is simply by opening the original Excel spreadsheet. Find the spreadsheet UniversalBank.csv that you downloaded and double click on it. The variables are defined below.

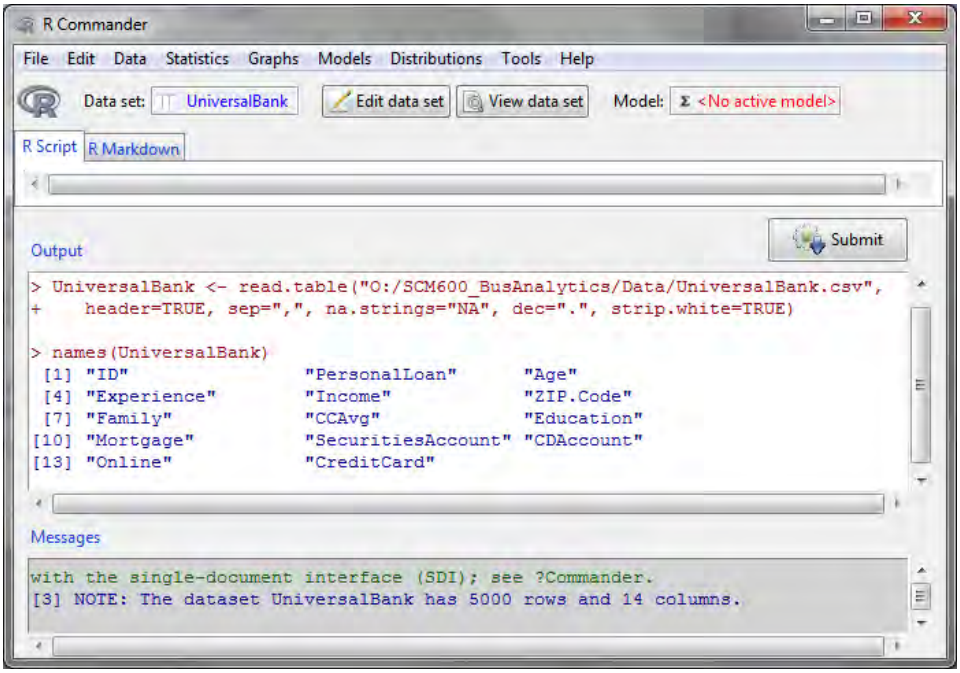
PersonalLoan: 0 for did not take loan, 1 if took loan
Age: age of customer
Experience: professional experience of customer
Income: income of customer
Family: family size of customer
CCAvg: average monthly credit card spending
Education: three categories (undergraduate, graduate, professional)
Mortgage: size of mortgage
SecuritiesAccount: No/Yes (0,1)
CDAccount: No/Yes (0,1)
Online: No/Yes (0,1)
CreditCard: No/Yes (0,1)



| ID | PersonalL | Age | Experienc | Income | ZIP Code | Family | CCAvg | Education | Mortgage | Securities | CDAccoun | Online | CreditCard |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 25 | 1 | 49 | 91107 | 4 | 1.6 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 45 | 19 | 34 | 90089 | 3 | 1.5 | 1 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 39 | 15 | 11 | 94720 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 35 | 9 | 100 | 94112 | 1 | 2.7 | 2 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 35 | 8 | 45 | 91330 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 37 | 13 | 29 | 92121 | 4 | 0.4 | 2 | 155 | 0 | 0 | 1 | 0 |
| 7 | 0 | 53 | 27 | 72 | 91711 | 2 | 1.5 | 2 | 0 | 0 | 0 | 1 | 0 |
| 8 | 0 | 50 | 24 | 22 | 93943 | 1 | 0.3 | 3 | 0 | 0 | 0 | 0 | 1 |
| 9 | 0 | 35 | 10 | 81 | 90089 | 3 | 0.6 | 2 | 104 | 0 | 0 | 1 | 0 |
| 10 | 1 | 34 | 9 | 180 | 93023 | 1 | 8.9 | 3 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 65 | 39 | 105 | 94710 | 4 | 2.4 | 3 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 29 | 5 | 45 | 90277 | 3 | 0.1 | 2 | 0 | 0 | 0 | 1 | 0 |

Now return to R. To view the variables in R,

1. Click on Data, Active Data Set, Variables in Active Data Set

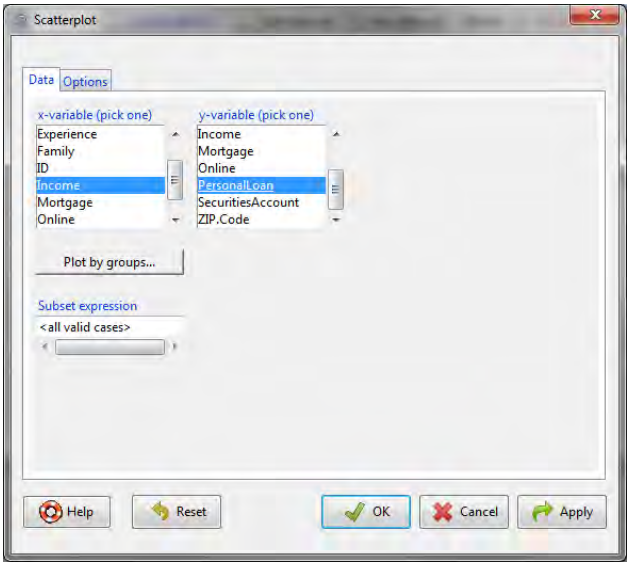



Notice that R generates the command names(UniversalBank). This is the command line version.

**Scatterplots**

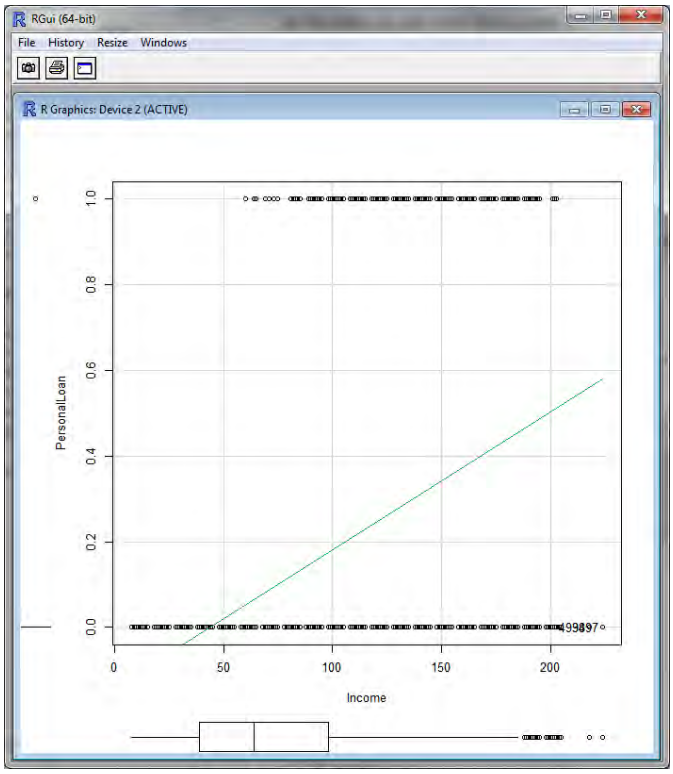
To generate a scatter plot,

1. Click on Graphs, Scatterplot
2. Select Income as the x-variable
3. Select PersonalLoan as the y-variable
4. Click on OK



To interpret the chart:

1. The black dots are the Income versus Loan (1) or No Loan (0)
2. The green line is the linear regression line through the data
3. Does the linear regression line make sense?




Linear regression assumes that there is a linear relationship between the X and Y variables. In this case, that doesn't make sense. A better solution looks like this:

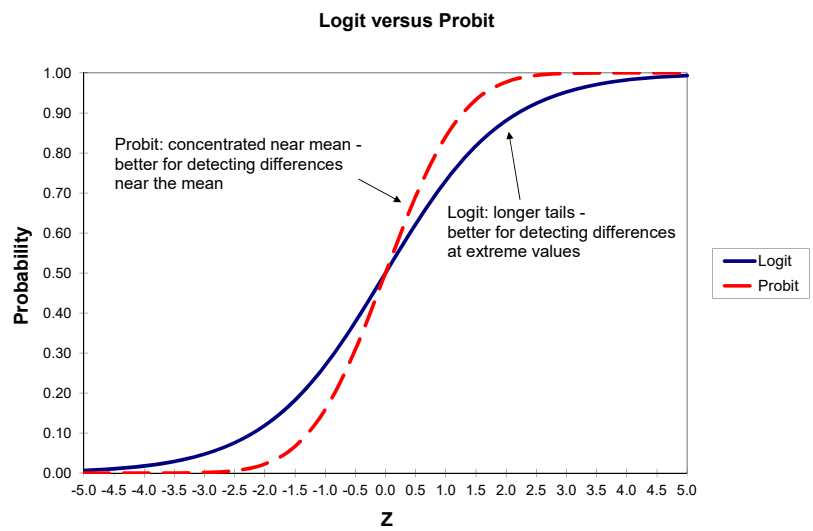
Logit and probit are techniques that assume the dependent variable (Y) is zero or one, and finds the relationship between the explanatory variables (X) and the dependent variable (Y). Logistic regression and logit are based on the logistic distribution. Probit is based on the normal distribution. Logit is more sensitive to extreme values of the X variable. Probit is more sensitive to values near the mean.

The Logit regression uses the logistic function to calculate the probability:

$$P(Y=1) = exp(\Sigma\beta_i X_i)/[1+ exp(\Sigma\beta_i X_i)]$$

The Probit regression uses the normal distribution to calculate the probability:

$$P(Y=1) = \Phi(\Sigma\beta_i X_i)$$ where $\Phi$ is the normal distribution



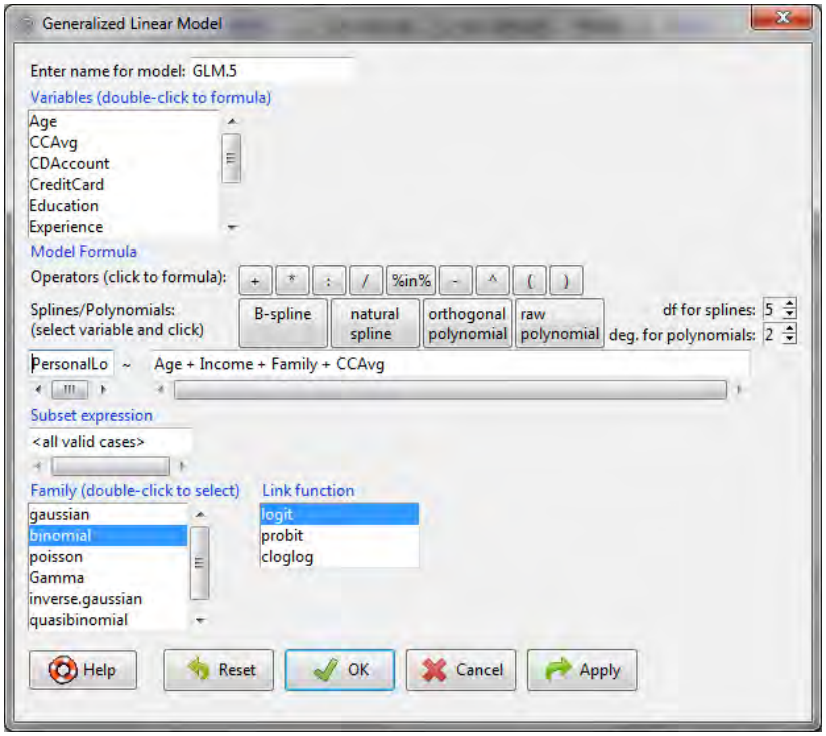

**Logit Analysis**

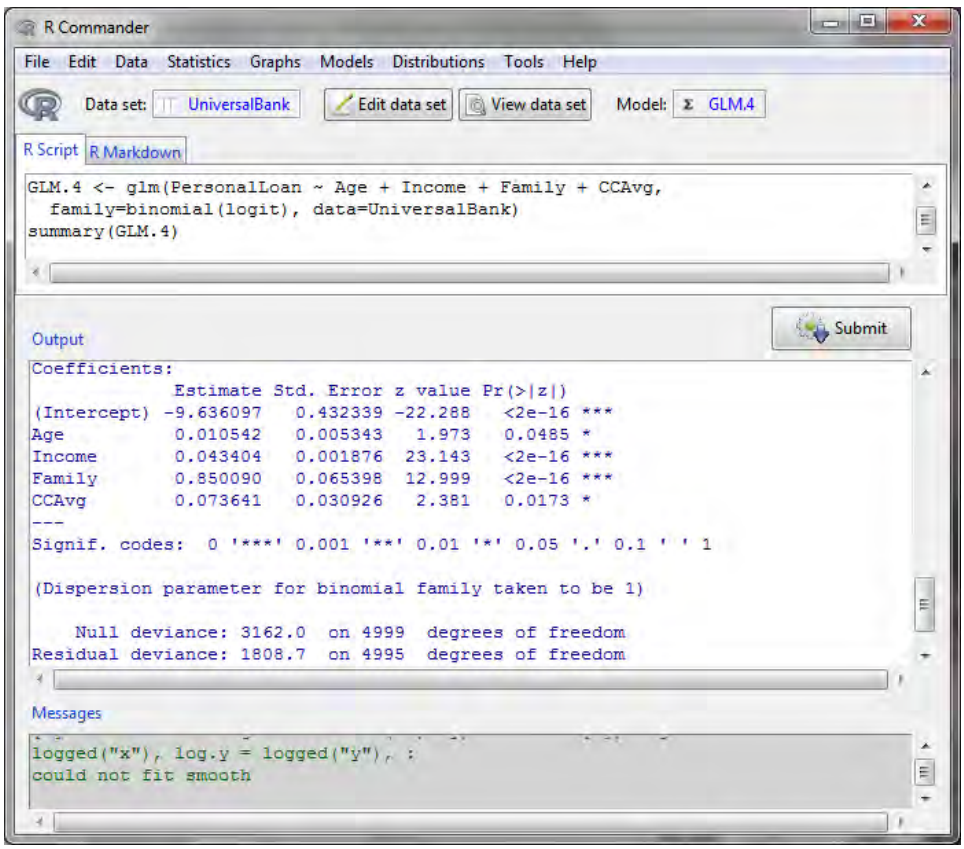
To perform a logit analysis on our data, where the Y variable is PersonalLoan and the explanatory variables are age, income, family size, and credit card average balance:

1. Click on Statistics, Fit models, Generalized linear model
2. Double click on PersonalLoan for the dependent variable
3. Double click on Age, Income, Family and CCAvg for the explanatory variables
4. Select binomial family
5. Select link function as logit
6. Click OK

Are the coefficients positive or negative?

Are the coefficients statistically significant?


Generalized Linear Model
Enter name for model: GLM.5
Variables (double-click to formula)
Age
CCAvg
CDAccount
CreditCard
Education
Experience
Model Formula
Operators (click to formula): + * : / %in% - ^ ( )
Splines/Polynomials:
(select variable and click)
B-spline
natural spline
orthogonal polynomial
raw polynomial
df for splines: 5
deg. for polynomials: 2
PersonalLo ~ Age + Income + Family + CCAvg
Subset expression
<all valid cases>
Family (double-click to select)
gaussian
binomial
poisson
Gamma
inverse.gaussian
quasibinomial
Link function
logit
probit
cloglog
Help
Reset
OK
Cancel
Apply

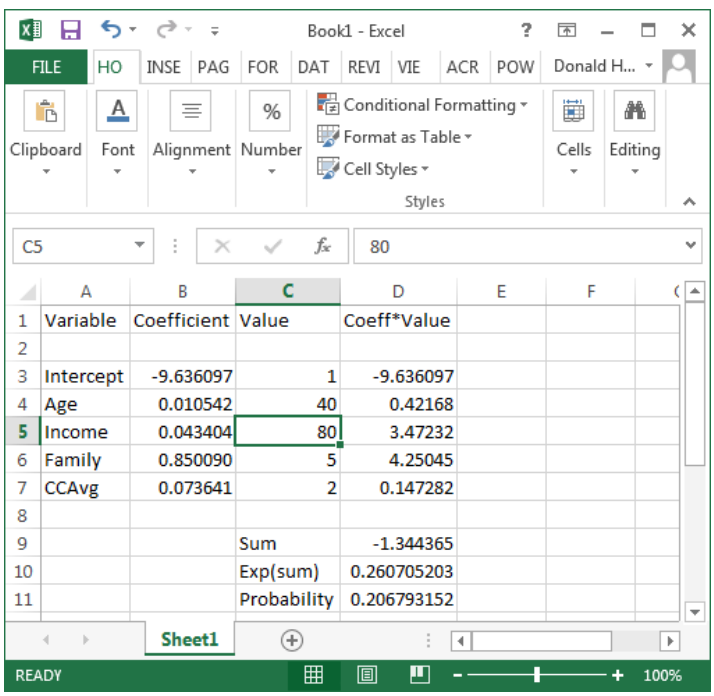

R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: UniversalBank
Edit data set
View data set
Model: GLM.4
R Script
R Markdown
GLM.4 <- glm(PersonalLoan ~ Age + Income + Family + CCAvg,
family=binomial(logit), data=UniversalBank)
summary(GLM.4)
Submit
Output
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.636097 0.432339 -22.288 <2e-16 ***
Age 0.010542 0.005343 1.973 0.0485 *
Income 0.043404 0.001876 23.143 <2e-16 ***
Family 0.850090 0.065398 12.999 <2e-16 ***
CCAvg 0.073641 0.030926 2.381 0.0173 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 3162.0 on 4999 degrees of freedom
Residual deviance: 1808.7 on 4995 degrees of freedom
Messages
logged("x"), log.y = logged("y"), :
could not fit smooth

**Session 9.5: Logit Predictions**

You can use a spreadsheet to calculate logit probabilities for different combinations of the explanatory variable values. Let's calculate the probability of taking out a loan for someone who is 40 years old, with $80,000 in income, family of 5, and credit card average balance of $2,000. Use the data from the previous page.

1. Create the labels in row 1 for Variable, Coefficient, Value, and Coeff*Value
2. In Column A, below Variable, list Intercept and each of the explanatory variables
3. In Column B, below Coefficient, enter the coefficients from your Logit regression
4. In Column C, below Value, enter 1 for intercept and the values that you want to evaluate
    a. Enter 40 for age
    b. Since income is in the dataset in thousands, enter 80 for $80,000
    c. Enter 5 for family
    d. Since credit card average balance is in the dataset in thousands, enter 2 for $2,000
5. In Column D, below Coeff*Value, enter the formula to multiply the coefficient and value; e.g., for cell D3, enter =B3*C3
6. In cell D9, enter the formula for the sum of the column D calculations; i.e., =sum(D3:D7)
7. In cell D10, calculate the exponential of the sum; i.e., =exp(D9)
8. In cell D11, calculate the probability; i.e., =D10/(1+D10)

Vary the age, income, family size, and credit card balance to determine the effect on the probability of taking out a loan.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Variable | Coefficient | Value | Coeff*Value |
| 2 | | | | |
| 3 | Intercept | -9.636097 | 1 | -9.636097 |
| 4 | Age | 0.010542 | 40 | 0.42168 |
| 5 | Income | 0.043404 | 80 | 3.47232 |
| 6 | Family | 0.850090 | 5 | 4.25045 |
| 7 | CCAvg | 0.073641 | 2 | 0.147282 |
| 8 | | | | |
| 9 | | | Sum | -1.344365 |
| 10 | | | Exp(sum) | 0.260705203 |
| 11 | | | Probability | 0.206793152 |

**Session 9.6: Probit Analysis**

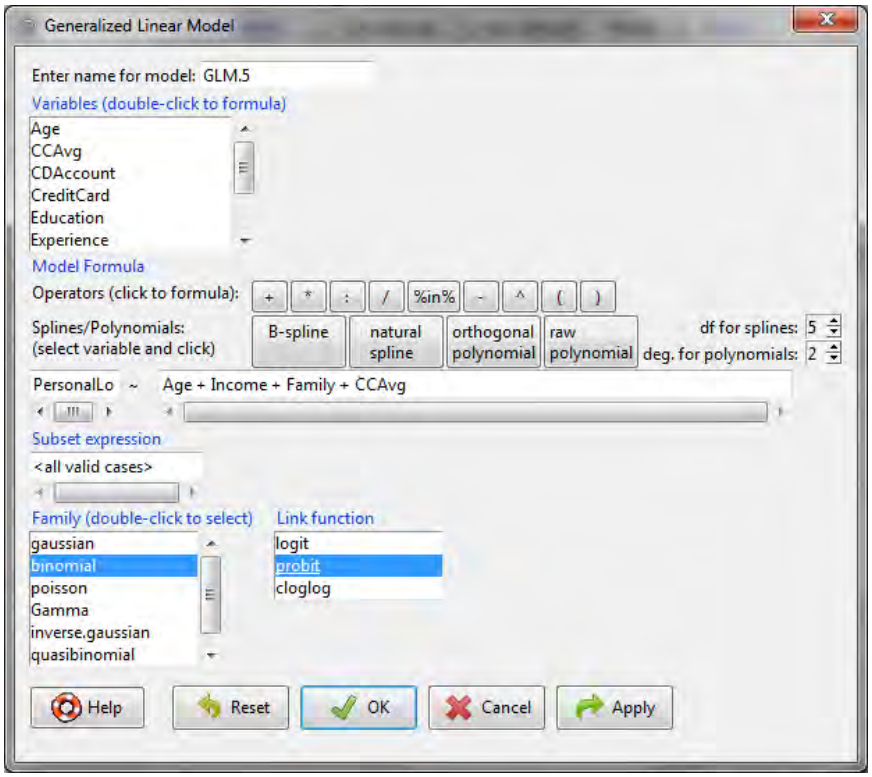
To perform a probit analysis on our data, where the Y variable is PersonalLoan and the explanatory variables are age, income, family size, and credit card average balance:

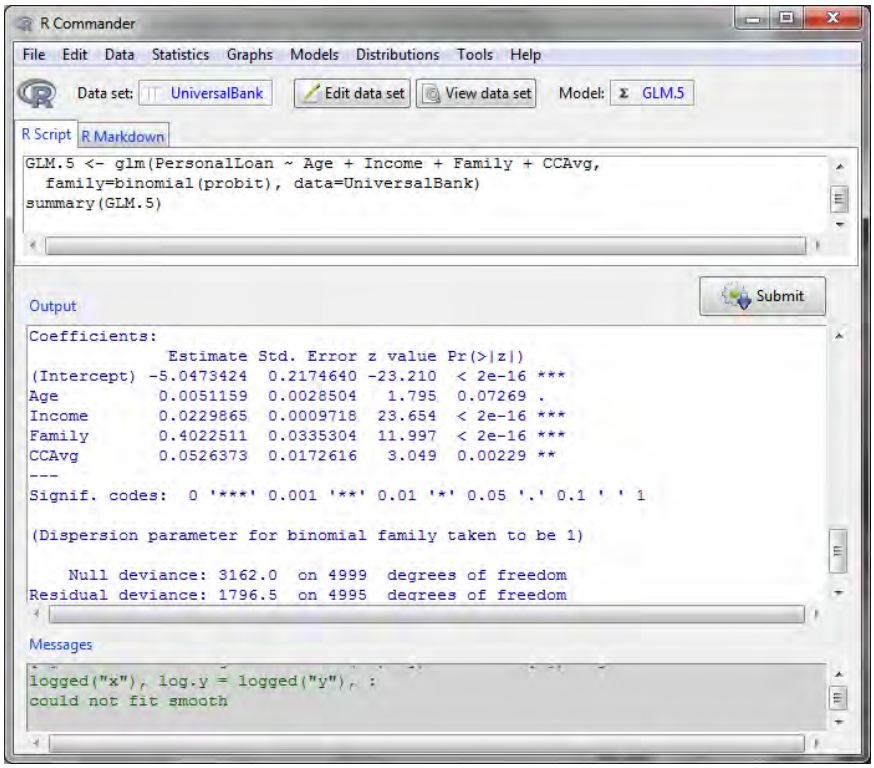
1. Click on Statistics, Fit models, Generalized linear model
2. Double click on PersonalLoan for the dependent variable
3. Double click on Age, Income, Family and CCAvg for the explanatory variables
4. Select binomial family
5. Select link function as probit
6. Click OK

Are the coefficients positive or negative?

Are the coefficients statistically significant?

Is there a difference between logit and probit?





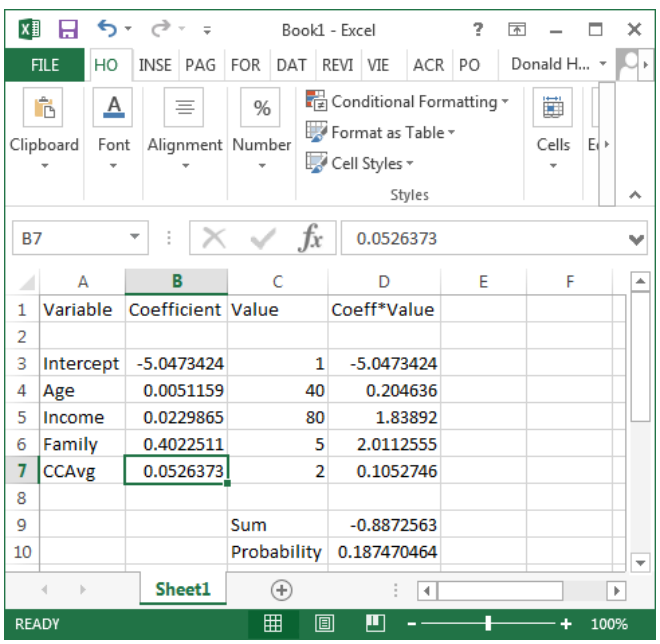

As an in class exercise, try adding Online as an explanatory variable. Is it significant in either logit or probit?

**Session 9.7: Probit Predictions**

You can use a spreadsheet to calculate probit probabilities for different combinations of the explanatory variable values. Let's calculate the probability of taking out a loan for someone who is 40 years old, with $80,000 in income, family of 5, and credit card average balance of $2,000. Use the data from the previous page.

1. Create the labels in row 1 for Variable, Coefficient, Value, and Coeff*Value
2. In Column A, below Variable, list Intercept and each of the explanatory variables
3. In Column B, below Coefficient, enter the coefficients from your Logit regression
4. In Column C, below Value, enter 1 for intercept and the values that you want to evaluate
    a. Enter 40 for age
    b. Since income is in the dataset in thousands, enter 80 for $80,000
    c. Enter 5 for family
    d. Since credit card average balance is in the dataset in thousands, enter 2 for $2,000
5. In Column D, below Coeff*Value, enter the formula to multiply the coefficient and value; e.g., for cell D3, enter =B3*C3
6. In cell D9, enter the formula for the sum of the column D calculations; i.e., =sum(D3:D7)
7. In cell D10, calculate the probability for the standard normal distribution using =NORM.S.DIST(D9,TRUE)

Vary the age, income, family size, and credit card balance to determine the effect on the probability of taking out a loan.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Variable | Coefficient | Value | Coeff*Value |
| 2 | | | | |
| 3 | Intercept | -5.0473424 | 1 | -5.0473424 |
| 4 | Age | 0.0051159 | 40 | 0.204636 |
| 5 | Income | 0.0229865 | 80 | 1.83892 |
| 6 | Family | 0.4022511 | 5 | 2.0112555 |
| 7 | CCAvg | 0.0526373 | 2 | 0.1052746 |
| 8 | | | | |
| 9 | | | Sum | -0.8872563 |
| 10 | | | Probability | 0.187470464 |

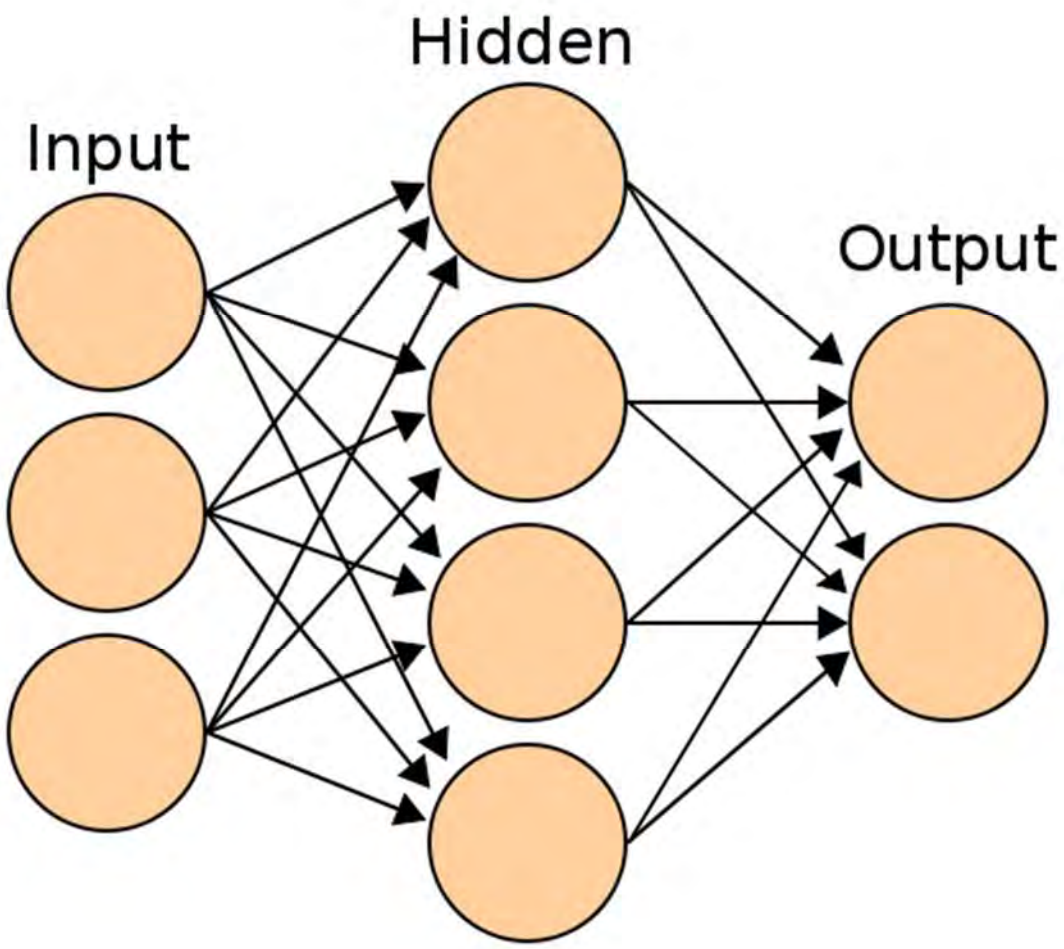
**Session 9.8: Neural Networks**

Reference:

Rumelhart, D.E; James McClelland (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press.

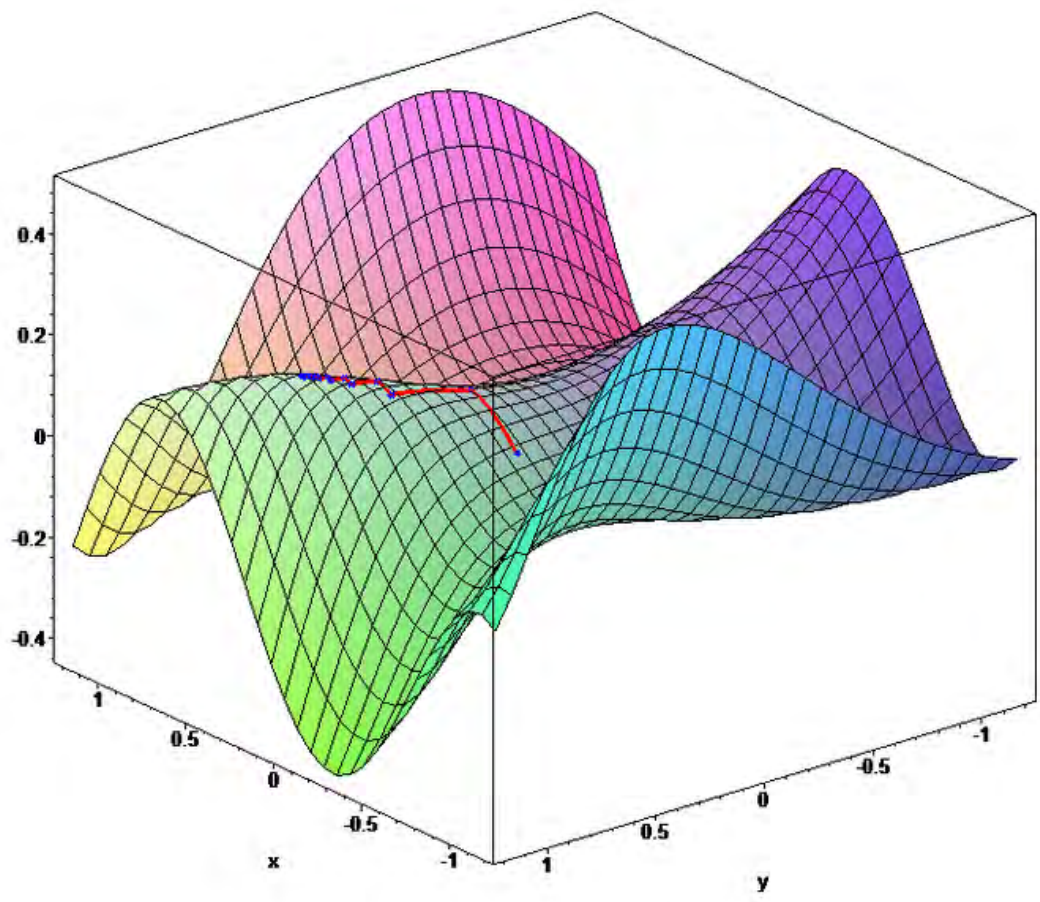
David Rumelhart and Jay McClelland recognized the limitation of a linear perceptron and proposed two innovations in 1986.

1. Use the logistic function to represent non-linear behavior
2. Add another layer (called the hidden layer) to produce more complex functions and represent more complex relationships



Why use the logistic function (used in Logit) rather than the normal distribution (used in Probit)? The logistic function has a very simple derivative. Why is this important?

Neural network searches use gradient search. Imagine that you are climbing a hill. To reach the peak in the shortest amount of time, look at where you are standing and find the direction with the steepest slope. Head in that direction, then decide in a new direction.




The risk is that there might be multiple high points, where some are local optima. Gradient search uses multiple starting points to find the global optimum.

So, why is a simply derivative for the logistic function important? The derivative gives you the slope so you can determine search direction. Other functions could be used, but the logistic function is the most popular because the derivative is easy to calculate.

If f(x) is the logistic function, then the derivative is f(x) * (1 – f(x)).

**Neural Networks with R**

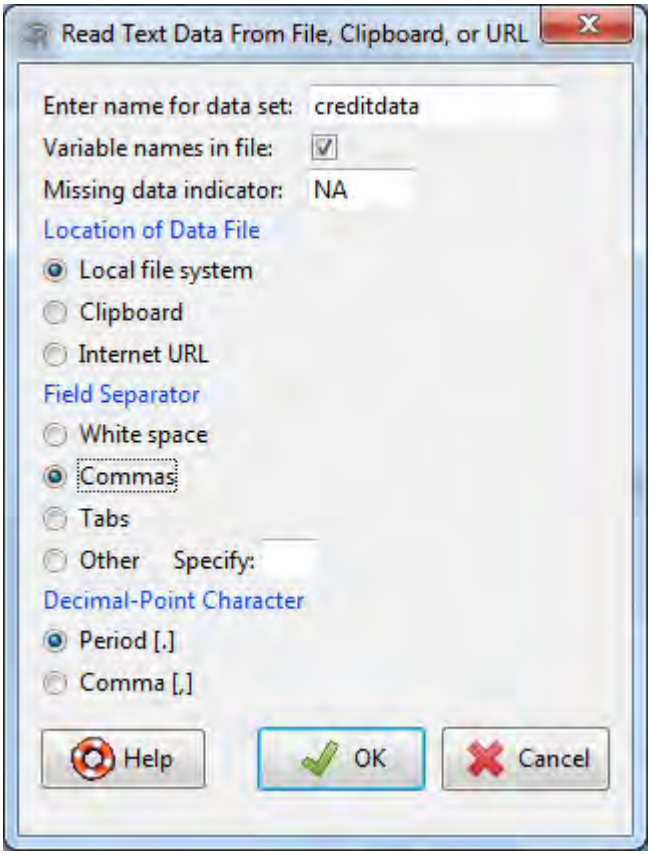
**Download Datasets**

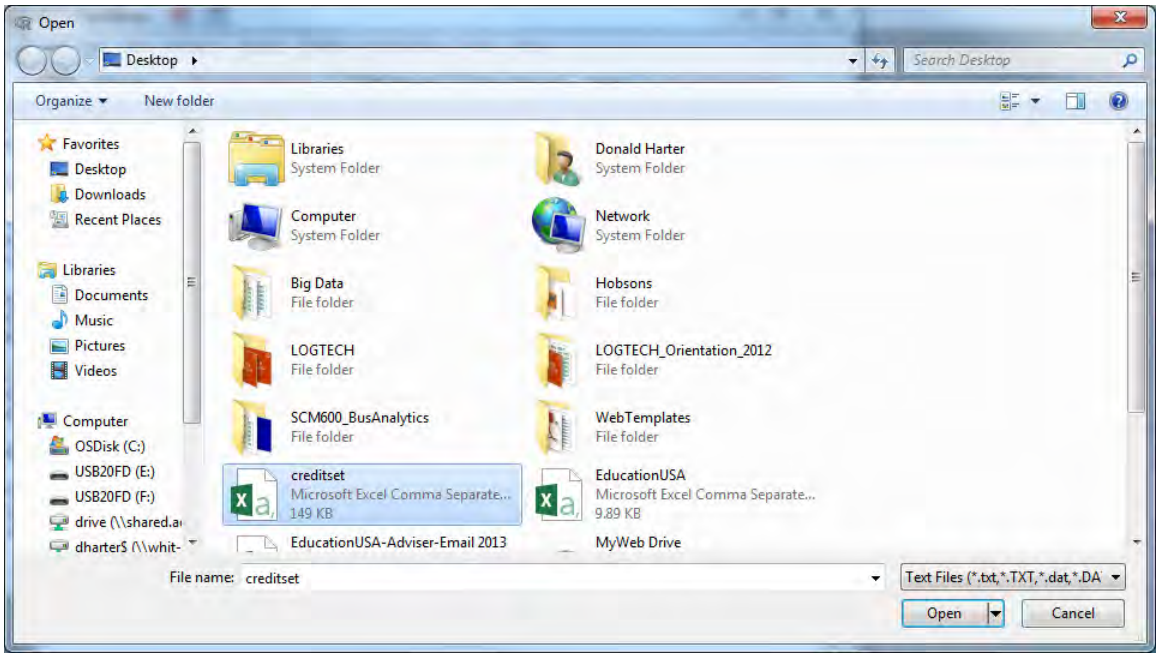
The Business Analytics - Week 9 creditset.csv file will be used for this exercise.

**Loading Data**

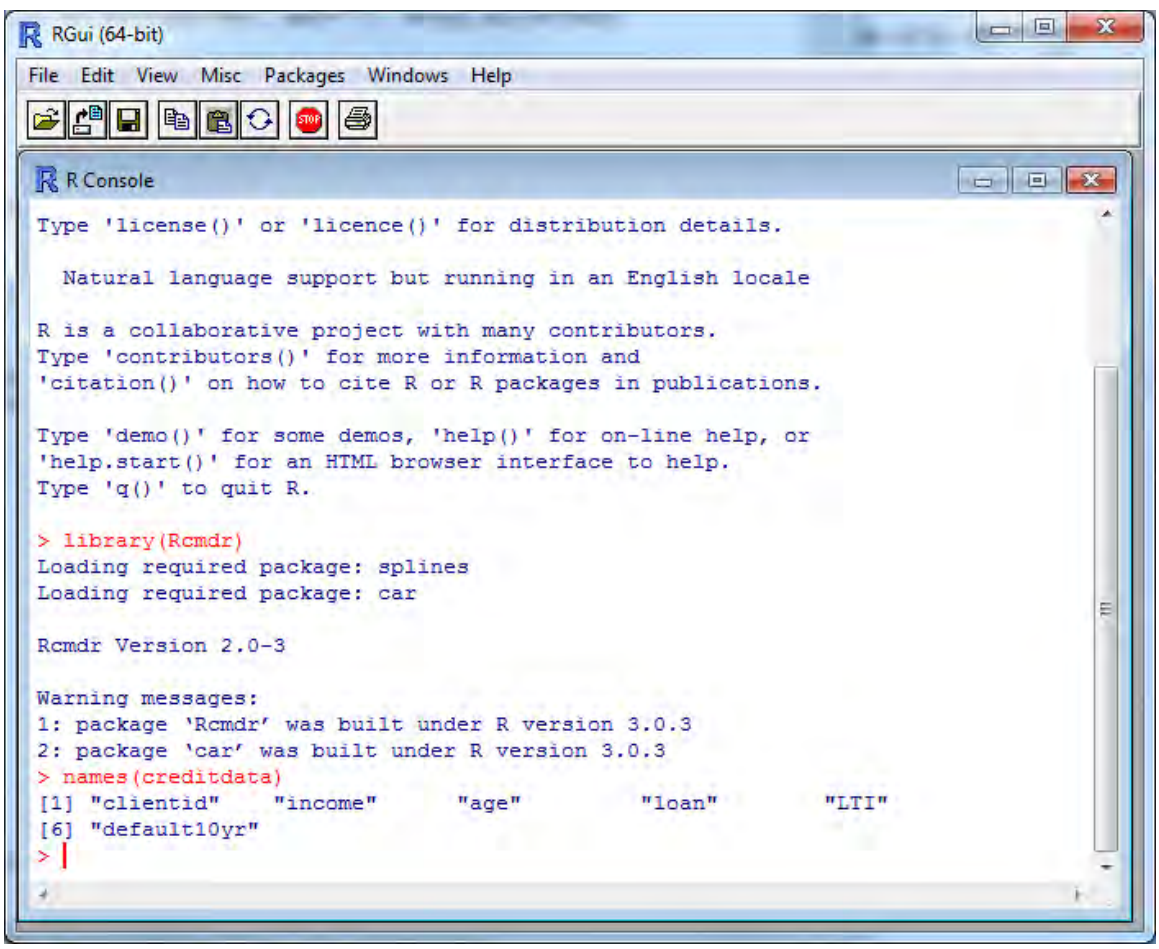
To load data into R:

1. Click on Data at the top of the Rcmdr screen
2. Click on Import Data > From text file …
3. Enter the name that you would like to use for this data set; type in creditdata
4. Change Field Separator to Commas, then OK
5. Click on the creditset.csv file, then Open

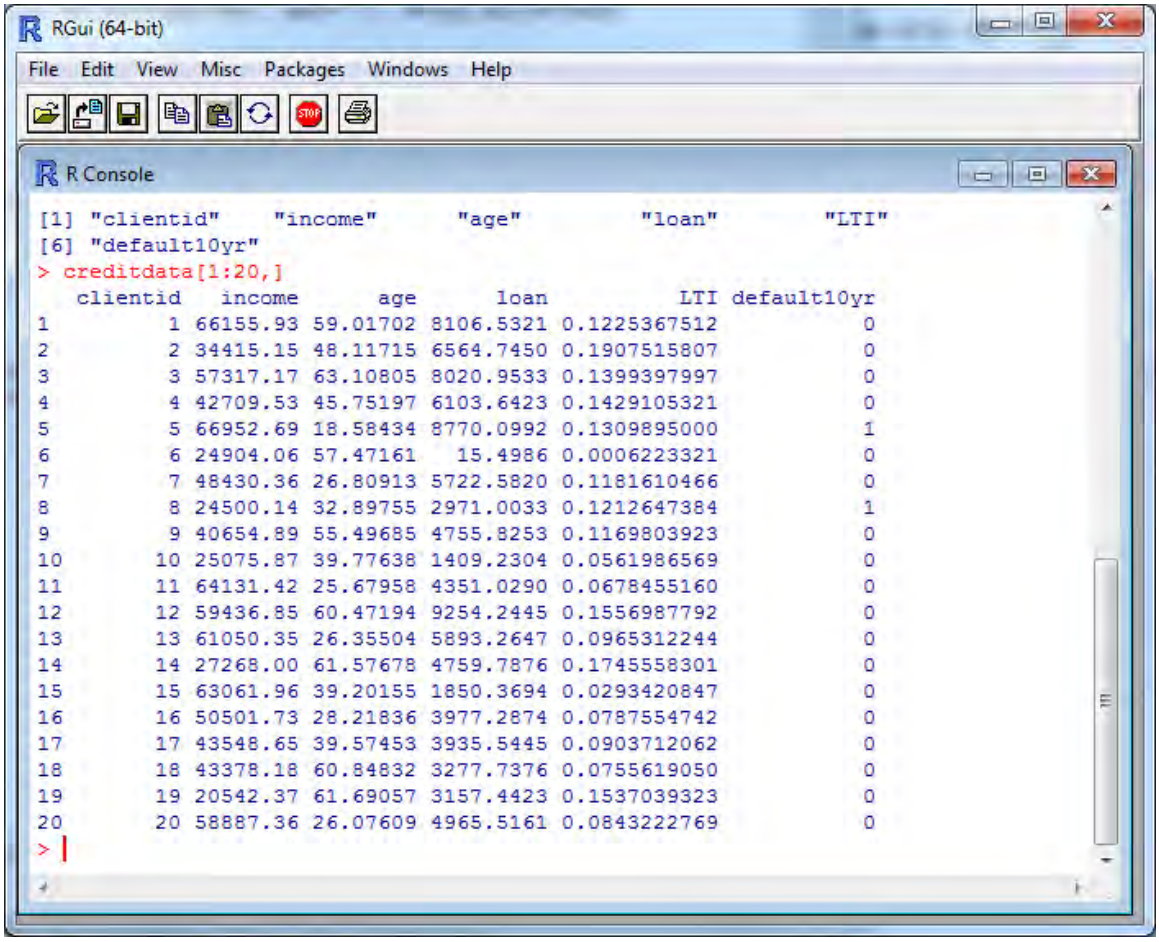





In the R Console (RGui), type names(creditdata).



In the R Console (RGui), type creditdata[1:20,]



```
[1] "clientid"    "income"      "age"         "loan"        "LTI"
[6] "default10yr"
> creditdata[1:20,]
   clientid   income      age      loan          LTI default10yr
1         1 66155.93 59.01702 8106.5321 0.1225367512           0
2         2 34415.15 48.11715 6564.7450 0.1907515807           0
3         3 57317.17 63.10805 8020.9533 0.1399397997           0
4         4 42709.53 45.75197 6103.6423 0.1429105321           0
5         5 66952.69 18.58434 8770.0992 0.1309895000           1
6         6 24904.06 57.47161   15.4986 0.0006223321           0
7         7 48430.36 26.80913 5722.5820 0.1181610466           0
8         8 24500.14 32.89755 2971.0033 0.1212647384           1
9         9 40654.89 55.49685 4755.8253 0.1169803923           0
10       10 25075.87 39.77638 1409.2304 0.0561986569           0
11       11 64131.42 25.67958 4351.0290 0.0678455160           0
12       12 59436.85 60.47194 9254.2445 0.1556987792           0
13       13 61050.35 26.35504 5893.2647 0.0965312244           0
14       14 27268.00 61.57678 4759.7876 0.1745558301           0
15       15 63061.96 39.20155 1850.3694 0.0293420847           0
16       16 50501.73 28.21836 3977.2874 0.0787554742           0
17       17 43548.65 39.57453 3935.5445 0.0903712062           0
18       18 43378.18 60.84832 3277.7376 0.0755619050           0
19       19 20542.37 61.69057 3157.4423 0.1537039323           0
20       20 58887.36 26.07609 4965.5161 0.0843222769           0
>
```

The data includes:

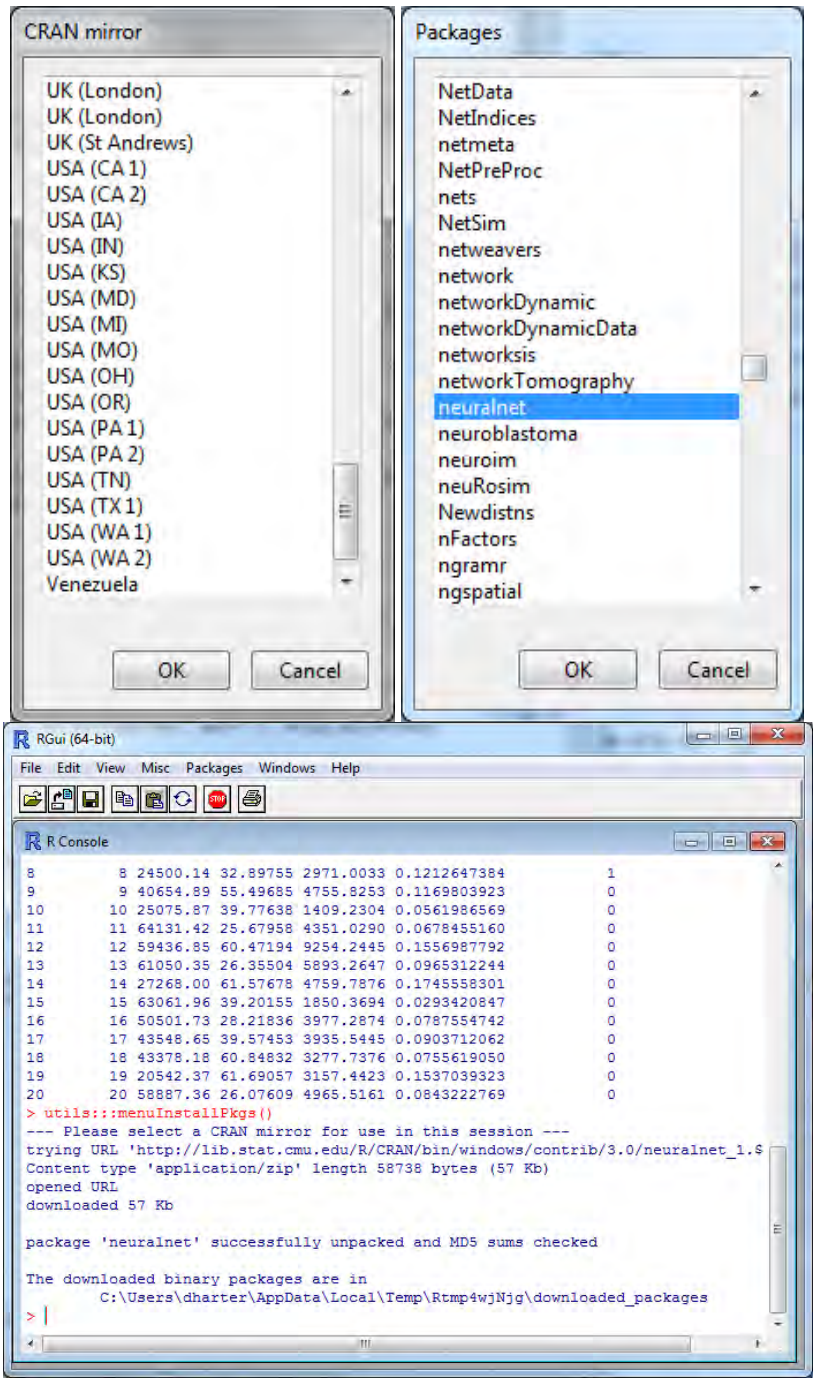
| | |
|---|---|
| Client id | unique identifier for each loan client |
| Income | annual income in Euros |
| Age | age of customer |
| Loan | loan size in Euros |
| LTI | loan to yearly income ratio |
| Default10yr | 1 if a default occurred in 10 years; 0 if no default occurred in 10 years |

Which variables should affect the probability of a loan default?

**Installing NeuralNet**

Follow these to install neuralnet.

1. In the R Console (RGui), at the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on neuralnet, then OK
5. If prompted to create a personal library, click Yes



**Launch neuralnet**

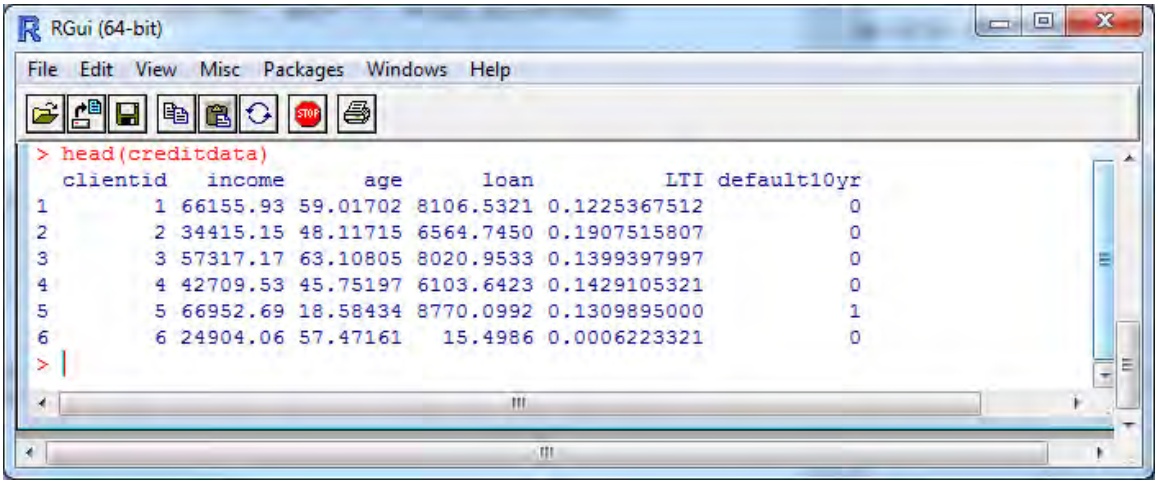
neuralnet is the R software that performs neural network calculations:

1. Type library(neuralnet)
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK
4. R will install the necessary software

**Viewing a sample of data**

Another way to view the data headers and sample data is with the head command.

1. In the R console (RGui), type head(creditdata)




**Neural network training and testing data**

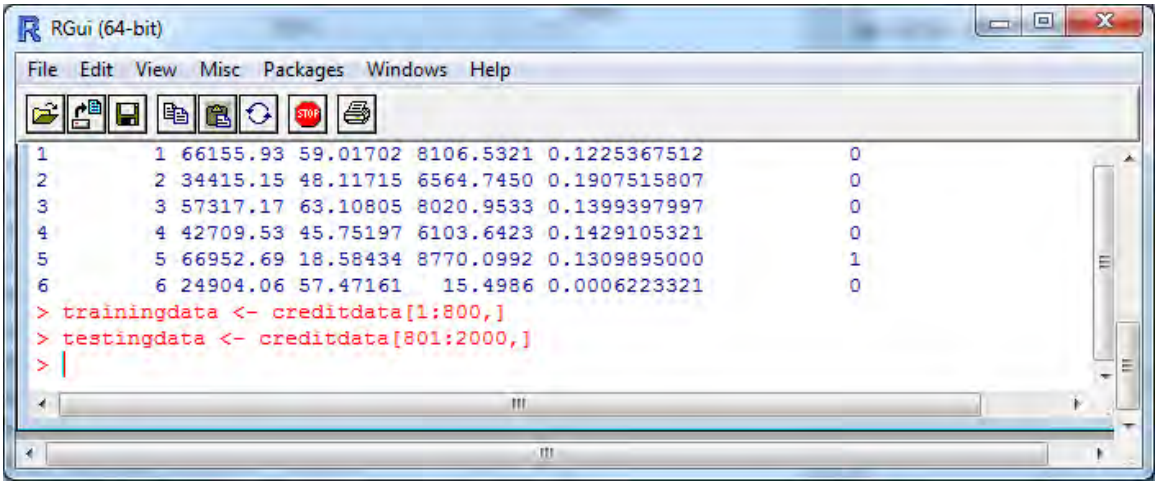
Whenever you builds analytics models, it's a good idea to use some data to build the model and other data to test the model. R allows you to split the dataset into training and testing subsets.

There are 2000 observations in our creditdata data set. Since the data is randomly distributed (not sorted), we will select the first 800 data rows for the training data, and the remainder of the data for testing purposes.

To create the training data and testing data:

1. Type the following to create the training data
   trainingdata <- creditdata[1:800,]
2. Type the following to create the testing data
   testingdata <- creditdata[801:2000,]

The training data that the neural network will use to learn is stored in trainingdata. The testing data that we will use to test the neural network is stored in testingdata.

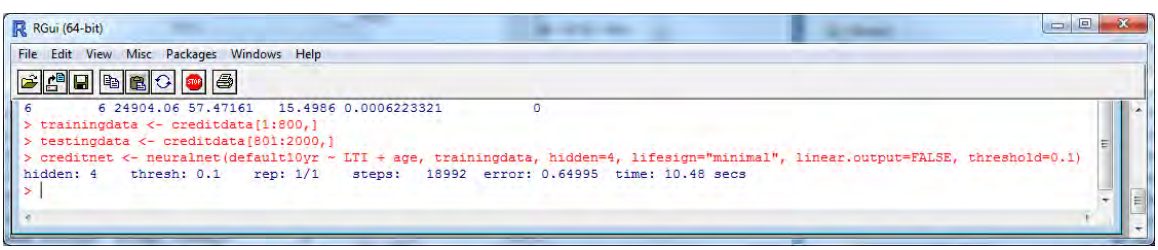



**Neural network analysis**

To run the neural network on loan defaults with inputs of loan to income ratio (LTI) and age, enter the command into the R console:

creditnet <- neuralnet(default10yr ~ LTI + age, trainingdata, hidden=4, lifesign="minimal", linear.output=FALSE, threshold=0.1)

| | |
|---|---|
| creditnet | stores the results |
| neuralnet | program which runs the neural network analysis |
| default10yr | dependent variable |
| LTI & age | independent variables |
| traningdata | data to be used for training the network |
| hidden | number of hidden nodes |
| lifesign | amount of output |
| linear.output | whether you want linear or non-linear model |
| threshold | error term threshold |

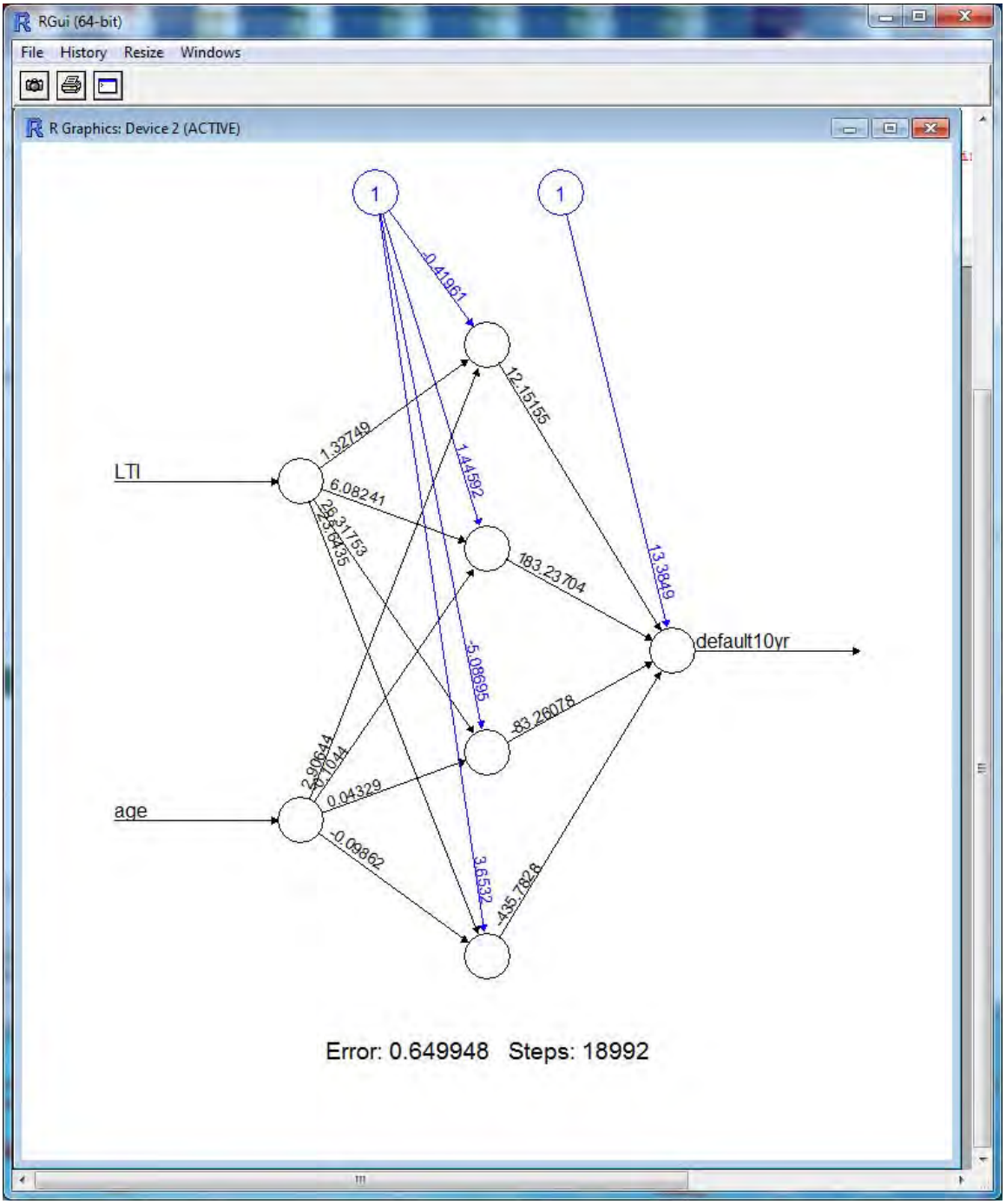



The neural network algorithm will perform a gradient search to find a solution that minimizes the error of making a mistake. In this case, the algorithm took 18,992 iterations or steps.

**Neural Network Model**

The result of the model can be displayed by plotting the model

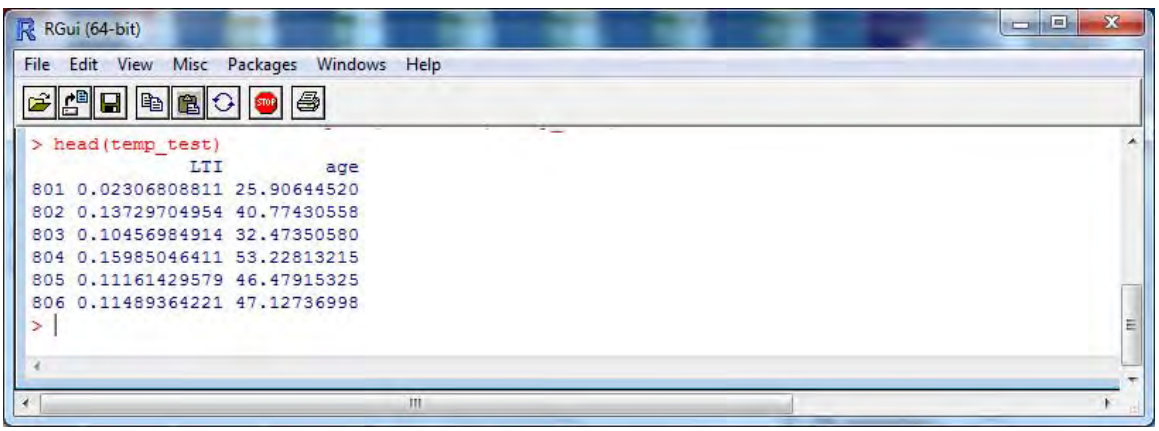
1. In the R console (RGui), type the command
   plot(creditnet,rep="best")

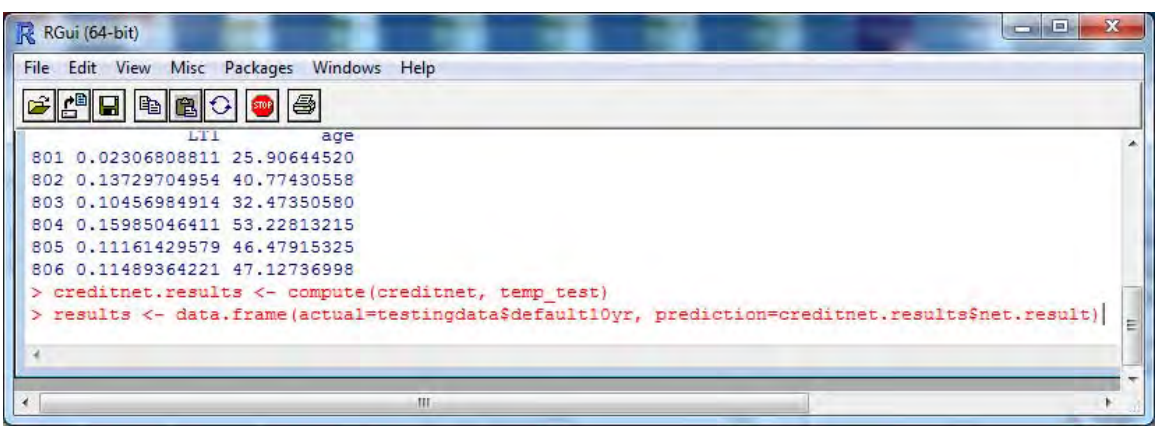


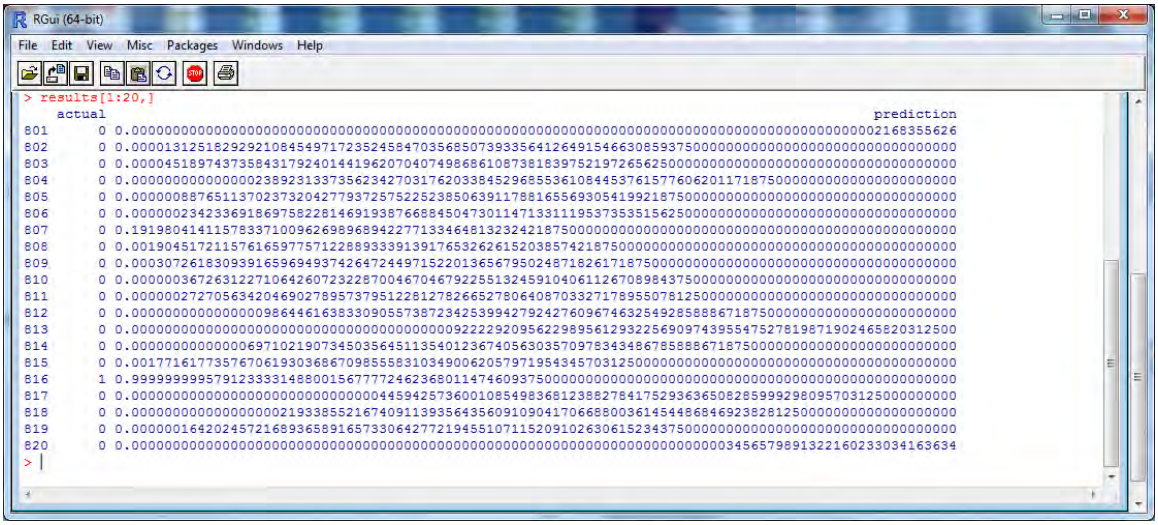
**Testing the Neural Network Prediction Ability**

To test our neural network with the testing data, we first need to reduce the test data to only the variables needed for the model.

1. In the R Console (RGui), type
   temp_test <- subset(testingdata, select = c("LTI", "age"))
2. To view this subset of data, type
   head(temp_test)
3. To predict loan defaults for the test data, type
   creditnet.results <- compute(creditnet, temp_test)
4. To create a view of the predictions, type the two lines below onto one line
   results <- data.frame(actual=testingdata$default10yr, prediction=creditnet.results$net.result)
5. Finally, to view the results, type
   results[1:20,]
6. There are too many decimal places. To round off the number, type
   results$prediction <- round(results$prediction)
   results [1:20,]



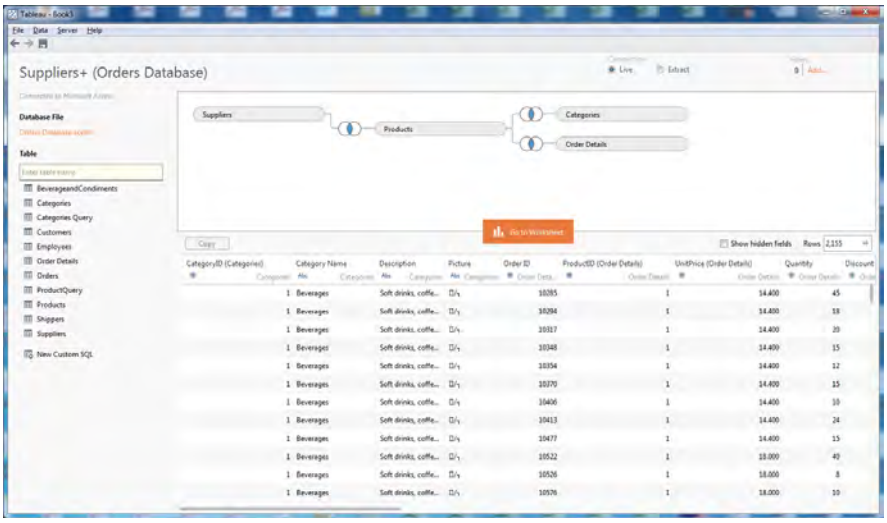

RGui (64-bit)
File Edit View Misc Packages Windows Help
> results[1:20,]
actual prediction
801 0 0.0000000000

## Tableau

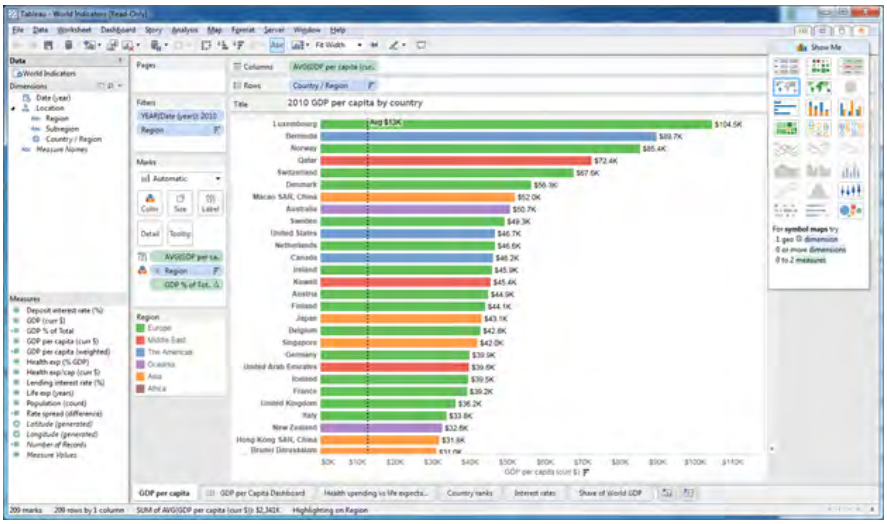
- Tableau is a commercial reporting package that can connect to various data sources.
    - Similar to PowerPivot, but also able to manage multiple tables and charts on a dashboard
- Tableau can use geolocations to display data, perform calculations, and filter data.



## Tableau Data and Relationships



## Sample Workbook



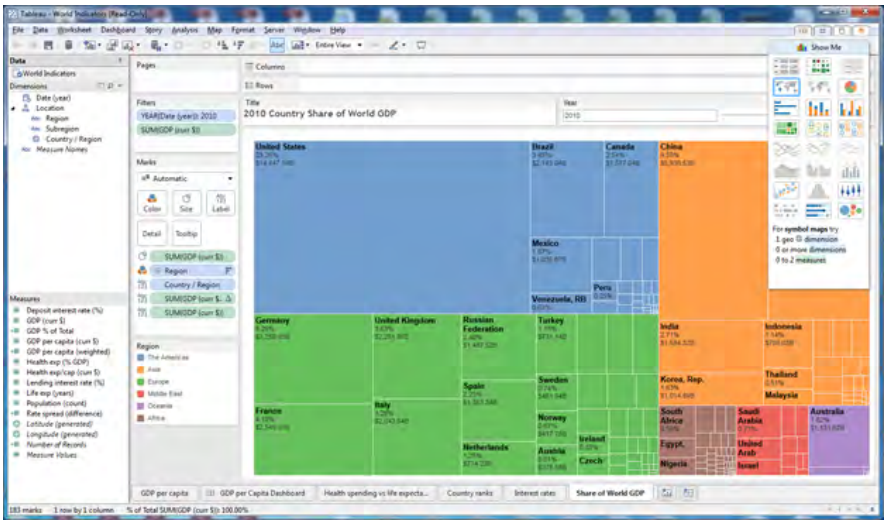

- Workbooks can represent data graphically.



## Multiple Views

- Multiple views of the data can be presented simultaneously.
- In this example, trends appear at the top, and status appears at the bottom.

## Variety of Representations




- Box sizes, colors, groupings, and other representation methods can also convey information.

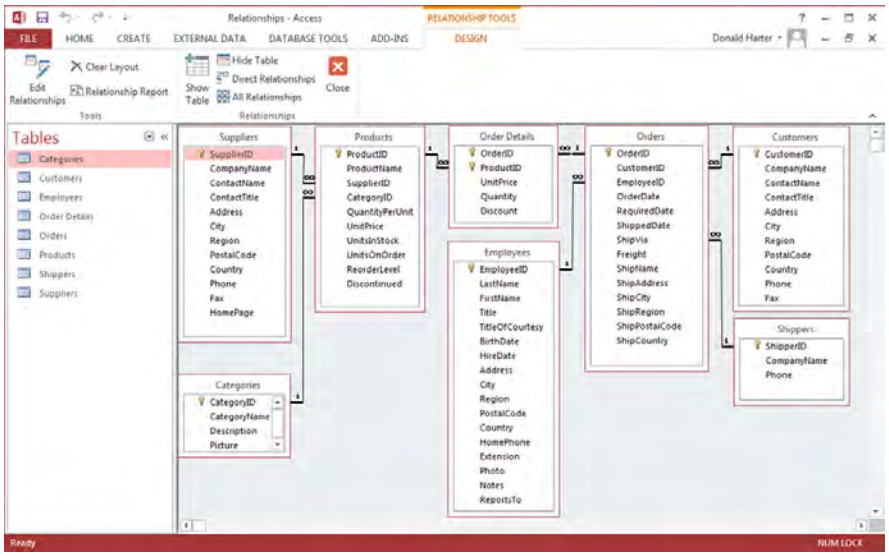
View All View Keyframes

## Tableau Data Connections

- Like PowerPivot and Access, Tableau can connect to remote databases.
- Tableau can take data from database tables, spreadsheets, text files, and other media and bring it into its own dashboards.
    - Tableau can integrate data that would otherwise be difficult to pull together.
- Data is automatically refreshed from sources, as needed.

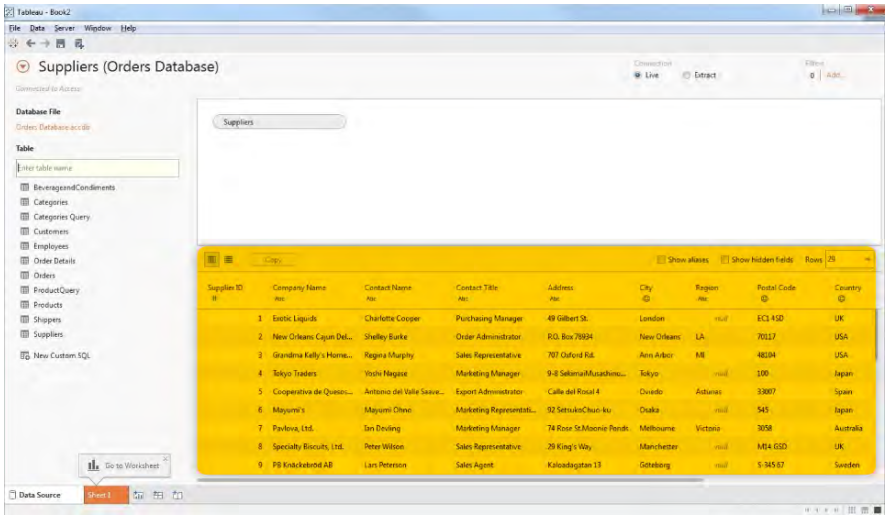
5 of 10

## Tableau Data Connections:
## Access Database




6 of 10

## Convert Data to Tableau




10 of 10

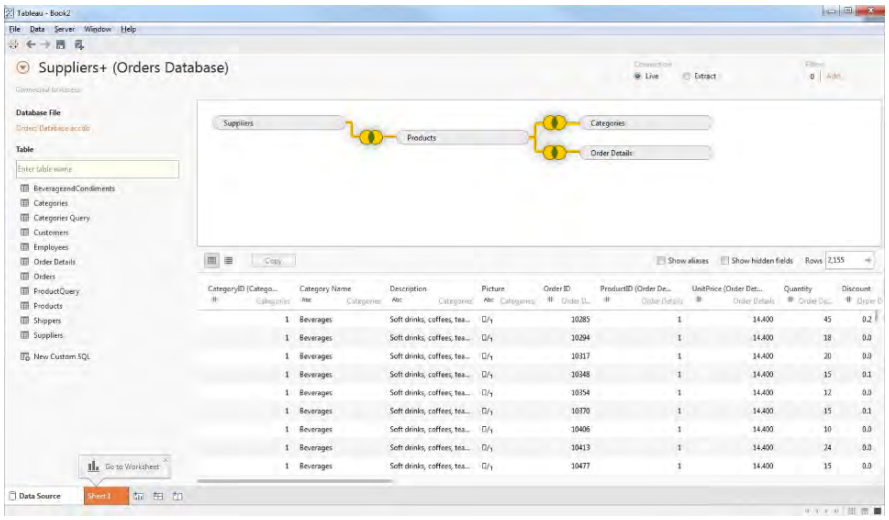
View All View Keyframes

## Tableau Relationships

- Relationships can be automatically built from preexisting relationships in the source databases.
    - Relationships can also be constructed manually.
- Common fields between tables join the tables into a relationship.

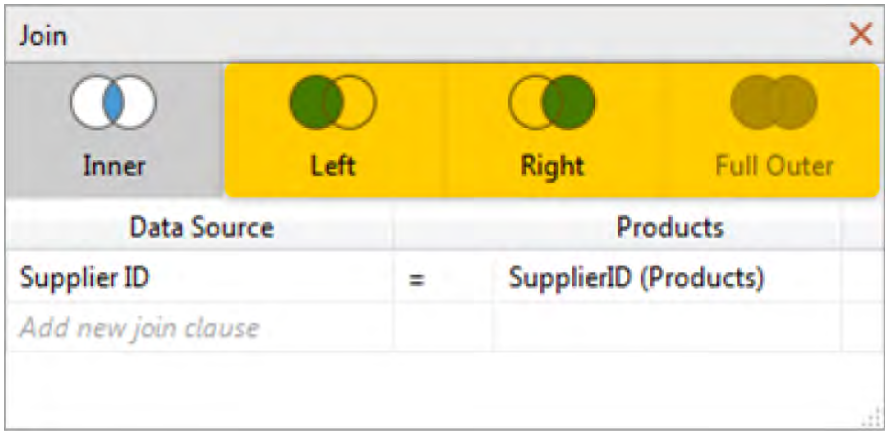
4 of 11

## Tableau Relationships (cont.)




8 of 11

## Option to Select Type of Join




- The inner join is the type we have used most frequently.
- The left join, right join, and outer join are other options.

11 of 11

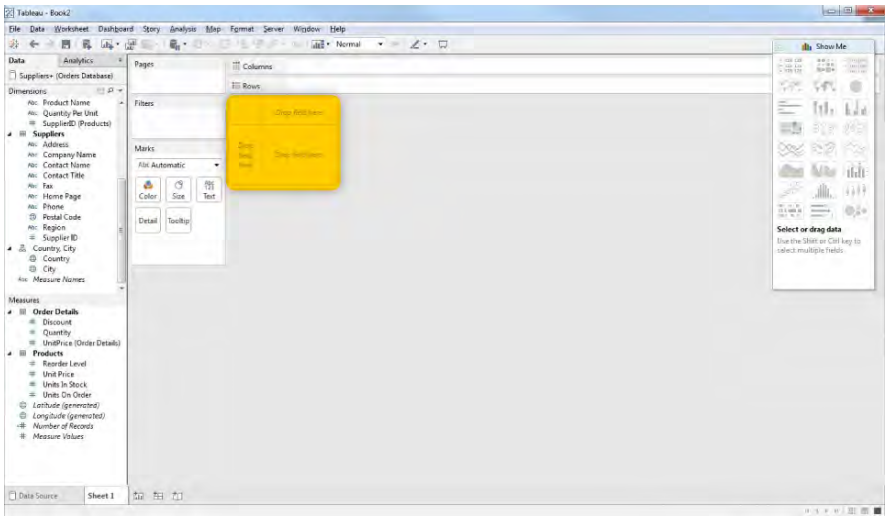
View All View Keyframes

## Building Tableau Worksheets

- Worksheets are the basic unit of Tableau.
- Worksheets can include tabular data or graphical representations of data.
- Data can be drawn from any data source connected to Tableau.

4 of 25

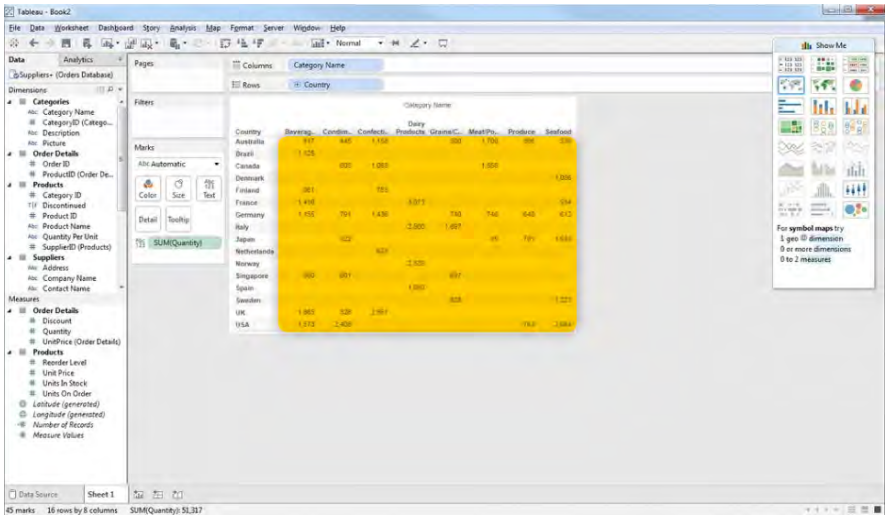
## Tableau Worksheet



- Data sources are shown on the left, including in this example: the product category, order detail, product, and supplier databases.
- From the data sources on the left, we drag and drop items into the middle area to create the worksheet.

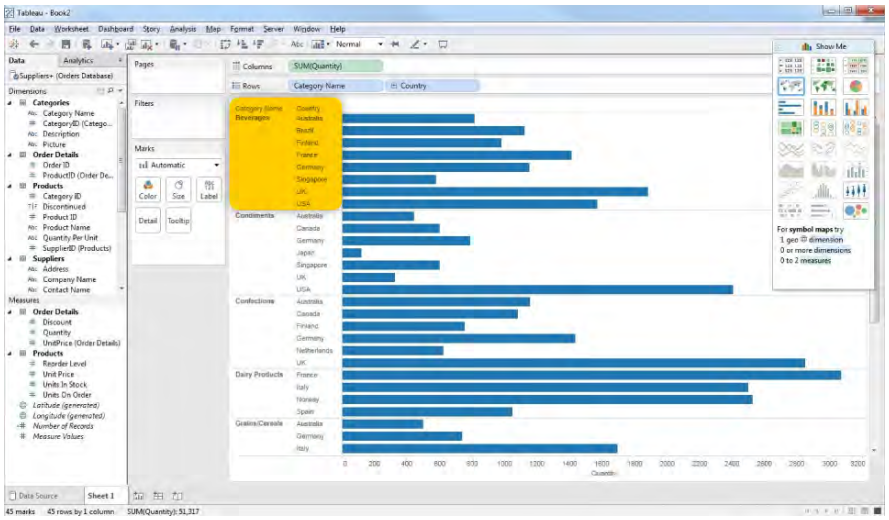
8 of 25

## Tabular Data Example



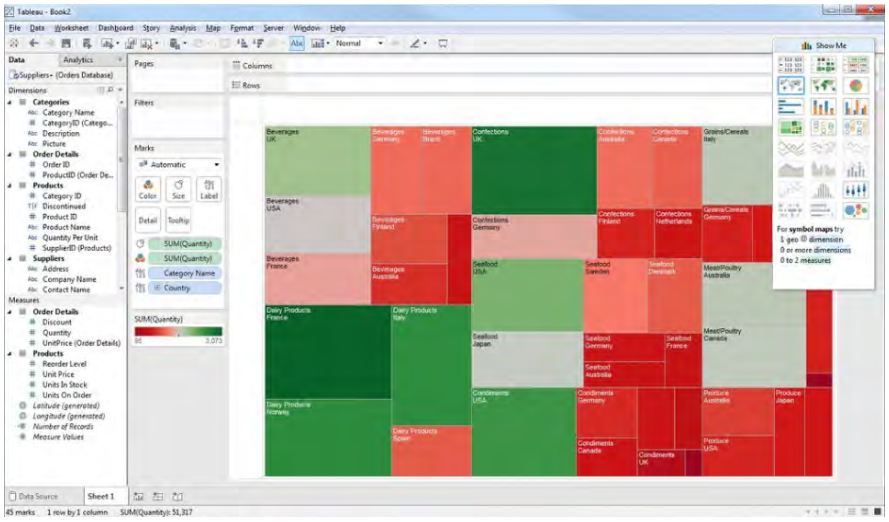
12 of 25

## Horizontal Bar Example



15 of 25

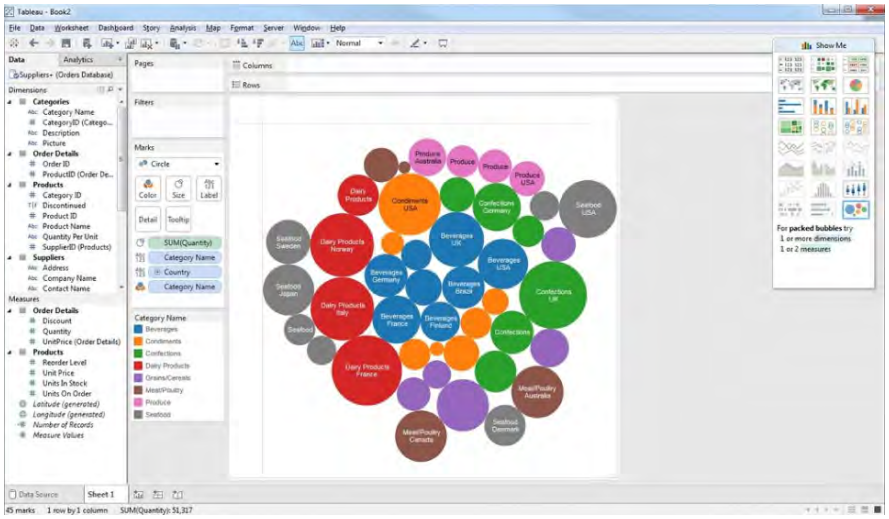
## Treemap Example



- Size, color, and location represent different features.
- In this example, box size and color indicate quantity, and location indicates product category.

21 of 25

## Packed Bubble Example




- Circle sizes represent sales.
- Colors represent product categories.

25 of 25

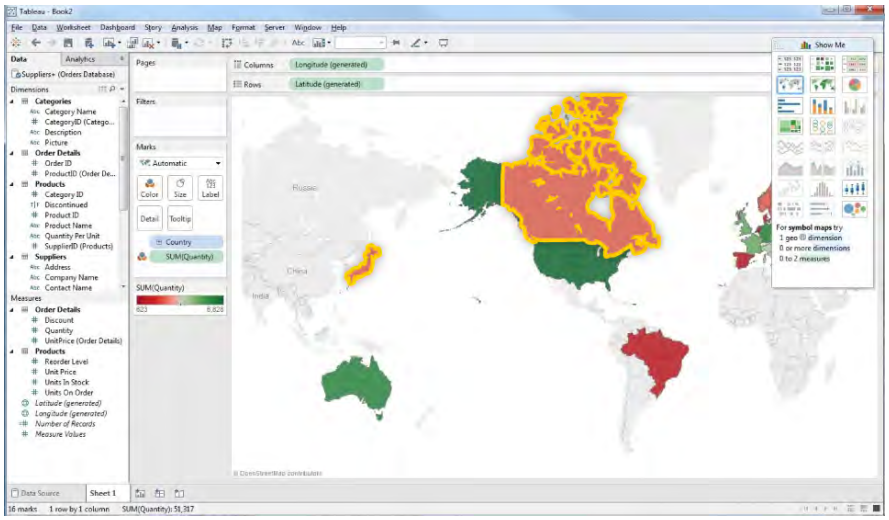
View All View Keyframes

## Geolocations

- Tableau uses geolocation information to merge data with a mapping representation.
- Tableau will automatically scan data to find geographic information such as postal codes, cities, countries, etc.

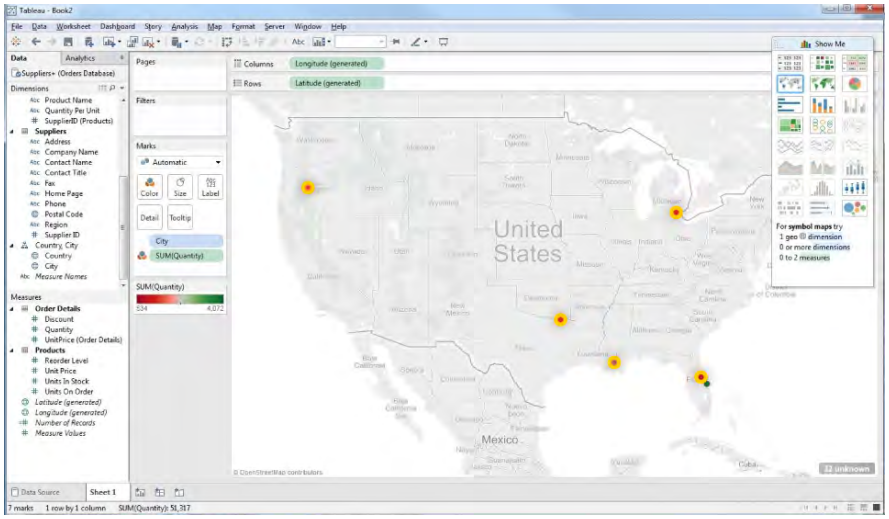
3 of 10

## World Map Example:
## Sales Data by Country



7 of 10

## Domestic Map Example:
## Sales by City




10 of 10

View All View Keyframes

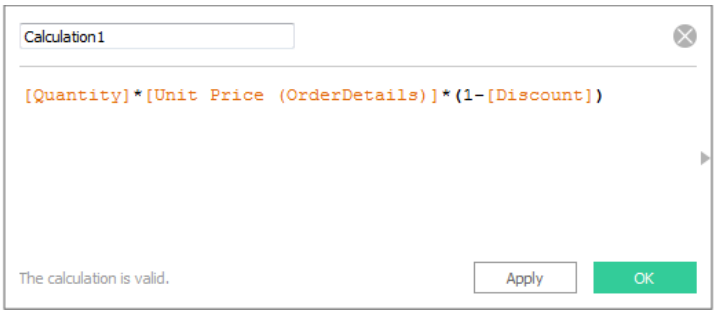
## Calculations

- Calculations can be added when generating tabular or graphic data.
- All mathematical functions are available and can be combined for tailored calculations, such as sales revenue after accounting for discounts:

$$Quantity * Price * (1 - Discount\ Rate)$$

4 of 10

## Example Calculation




- Calculations are built via point-and-click.
- The results can then drive tabular and graphic data.

10 of 10

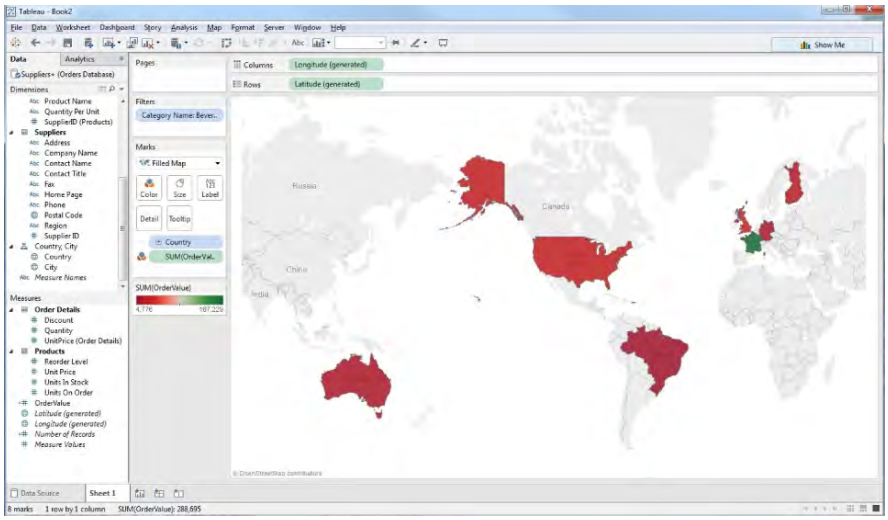
View All View Keyframes

## Filters

- Tableau filters work similarly to filters in Excel and Access.
- Filters allow you to select data based on specific characteristics.
- Filters can be used in both tables and graphs.

4 of 9

## Example Filter




9 of 9

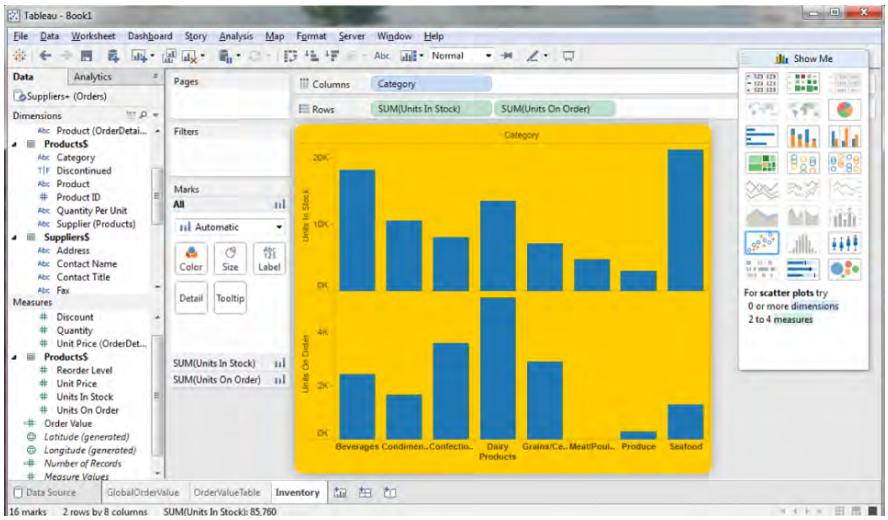
View All View Keyframes

## Building a Dashboard

- A dashboard is a collection of worksheets.
- The first step in building a dashboard is to build all the worksheets to be used in the dashboard.
    - These can include tables, charts, geographical representations, etc.
- The second step is to drag and drop the worksheets into the desktop (after activating Dashboard mode).

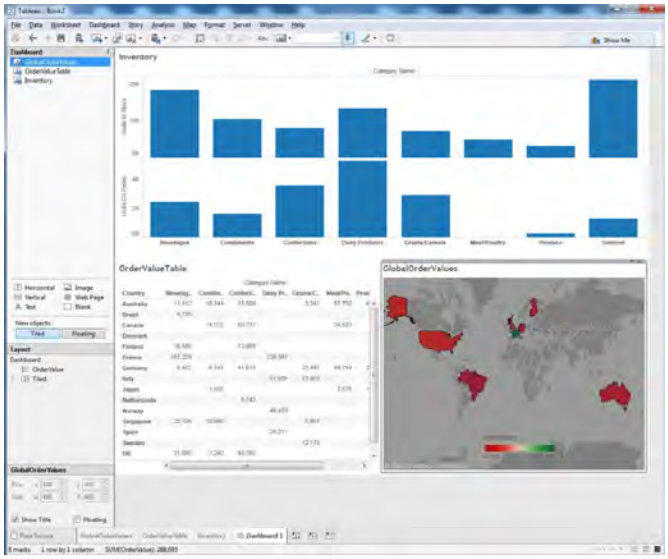
5 of 13

## Drag and Drop First Worksheet




7 of 13

## Drag and Drop Additional Worksheets




- As you position each item, Tableau will automatically reallocate space on the screen.
- The chart, table, and geographical representation will automatically update whenever the data are refreshed.

13 of 13

# SCM 651: Business Analytics

1

Final Exam Review

# Final Exam



- Academic integrity
  - Do your own work, no collaboration
- My philosophy
  - More questions, each worth fewer points
    - Advantage: if you don't know the answer on a question, it's only worth a few points
  - Test breadth of knowledge (multiple choice)
  - Test depth of knowledge (short answer)
- Test Taking Strategy
  - Strive for full credit on a question – there is no extra credit for elaborate answers, so don't spend too much time on any question
  - Use your time wisely

# Final Exam



- Content
  - Part 1: Concepts – Short Answer
  - Part 2: Tools – Multiple Choice
  - Part 3: Techniques – Multiple Choice
  - Part 4: Regression Assumptions – Multiple Choice
  - Part 5: Interpretation – Short Answer
  - Part 6: Business Issues from Articles – Short Answer
- Summary
  - 15-20 multiple choice questions
  - 15-20 short answer questions

- Final exam will be sent to you via email at your syr.edu address
- You will need MS Word
- You will not run any other software – just answer the questions

# Final Exam



- Content
  - Part 1: Concepts – Short Answer
    - Define or describe a concept or business application
  - Part 2: Tools – Multiple Choice
    - Identify which tool was used in a given example (e.g., Excel, Access, Google Analytics, R, Tableau)
  - Part 3: Techniques – Multiple Choice
    - Identify which technique is presented in example (e.g., correlation, linear regression, exponential regression, power regression, moving average, logit, probit, neural network)
  - Part 4: Regression Assumptions – Multiple Choice
    - Identify assumption violations, corrections (linearity, multi-collinearity, heteroscedasticity, serial correlation, outliers)
  - Part 5: Interpretation – Short Answer
    - Interpret output results of a technique
  - Part 6: Business Issues from Articles – Short Answer
    - Provide a short answer to questions from the articles

# Final Exam Topics



- Week 1
  - Background
    - What drives analytics?
    - Why is analytics difficult?
    - What are business examples where analytics is important?
  - Tools
    - Formulas
    - Sorting
    - Filters
    - Pivot tables and charts
    - Powerview

# Final Exam Topics



- Week 2
  - NPV
  - IRR
  - Correlation
  - Linear regression
  - Exponential regression
  - Power regression
  - Time series

# Final Exam Topics



- Week 3
  - Sensitivity analysis
  - Conditional formatting
  - Dashboards in Excel
  - Google analytics

# Final Exam Topics



- Week 4
  - Importing data
  - Access tables
  - Access relationships
  - Access queries
    - Grouping
    - Criteria
    - Calculations

# Final Exam Topics



- Week 5
  - PowerPivot importing
  - PowerPivot relationships
  - PowerPivot tables
  - PowerPivot charts

# Final Exam Topics



- Week 6
  - Goal seek
  - Solver (unconstrained)
  - Solver (constrained)

# Final Exam Topics



- Week 7
  - R: 3D visualization
  - ANOVA
  - Dummy variables
  - Moderating effects

# Final Exam Topics



- Week 8
  - Regression Assumptions
    - Know what each looks like
    - Know what a violation looks like
    - Know the solutions to the assumption violation
  - Solutions
    - Linearity
      - Solution: transformation
    - Multi-collinearity
      - Solution: Combine variables or drop one
    - Heteroscedasticity
      - Solution: transformation
    - Serial correlation
      - Solution: Time series analysis
    - Outliers
      - Solution: drop outliers

# Final Exam Topics



- Week 8 (continued)
  - Benford's Law
  - Decision trees

# Final Exam Topics



- Week 9
  - Logit
    - Logistic distribution
    - More sensitive at extreme values of X variables
  - Probit
    - Normal distribution
    - More sensitive at values of variables near their means
  - Perceptrons
    - Early linear attempt at machine learning

# Final Exam Topics



- Week 9
  - Neural networks
    - Uses logistic function
    - Has at least three levels: inputs (X), hidden (H), and outputs (Y)
      - Can have multiple hidden layers (deep neural networks)
    - Subject to local optima

# Final Exam Topics



- Week 10
  - Tableau
    - Importing data
    - Creating relationships
    - Tables and charts
    - Dashboards

# Tableau

**Session 10.1: Overview**

Tableau is a data visualization and data analysis tool. It has the capability to connect to external data sources, join various data sets, summarize and display data in dashboard format.

For these exercises, download the files:
“Business Analytics – Week 10 Instructions.doc”
“Business Analytics – Week 10 Orders.xls”

**Installing Tableau**

1. Go to http://www.tableausoftware.com/tft/activation
2. Desktop Key: **TDLF-067F-C3E0-1561-FBD8**
3. Click on the link above and select Get Started. Enter your university email address for “Business email”; and under "Organization", please input the name of your school

**Starting Tableau**

On pre-windows 7 computers:

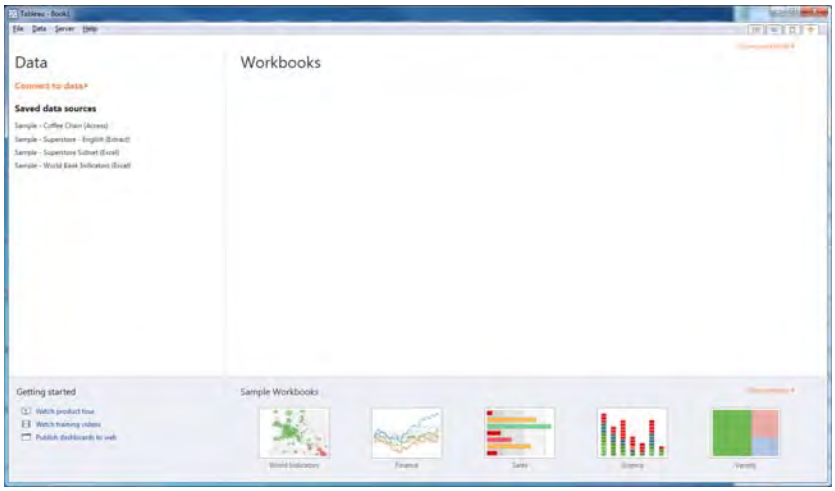
1. Click on the Start button in the lower left corner of Windows
2. Click on All Programs, then click on Tableau

On Windows 7 or later computers:

1. Move the cursor to the upper right corner of your screen
2. Click on search
3. Search for Tableau

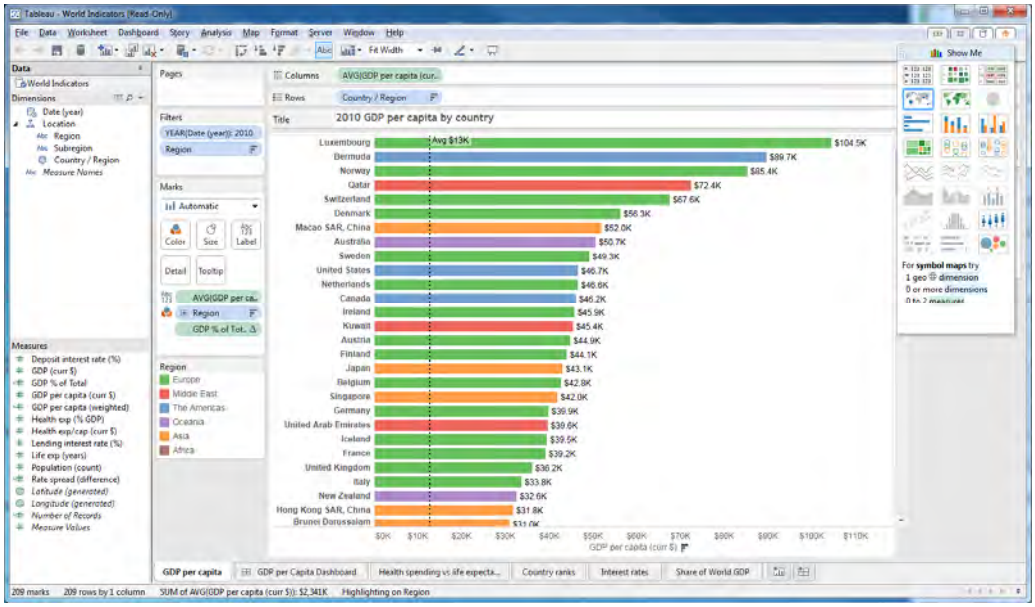
**Guided Tour of Tableau Capabilities**

After opening Tableau, you should see a screen similar to the following:

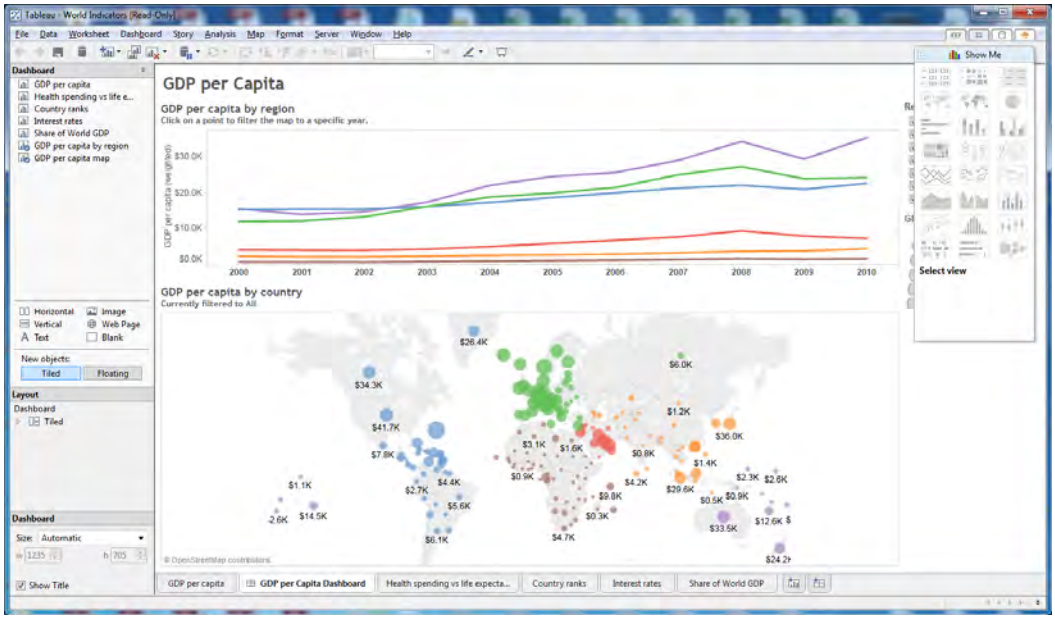



The area labeled Workbooks will display previously used workbooks (workbook history). If this is the first time you’ve opened Tableau, this will be blank. At the bottom of the screen are sample workbooks.

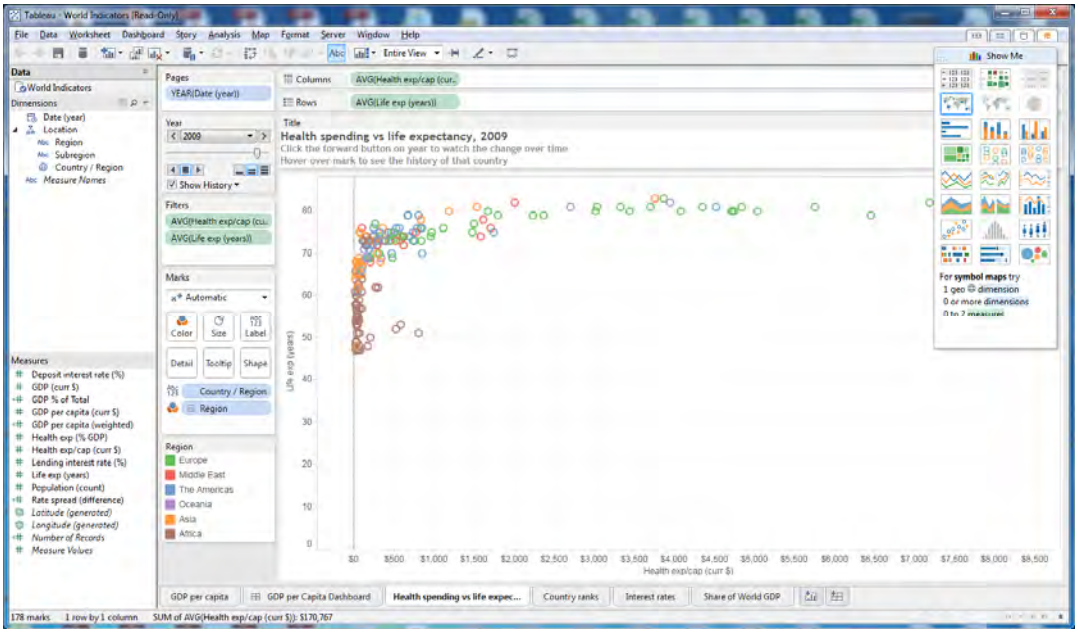
1. Click on the World Indicators sample workbook.



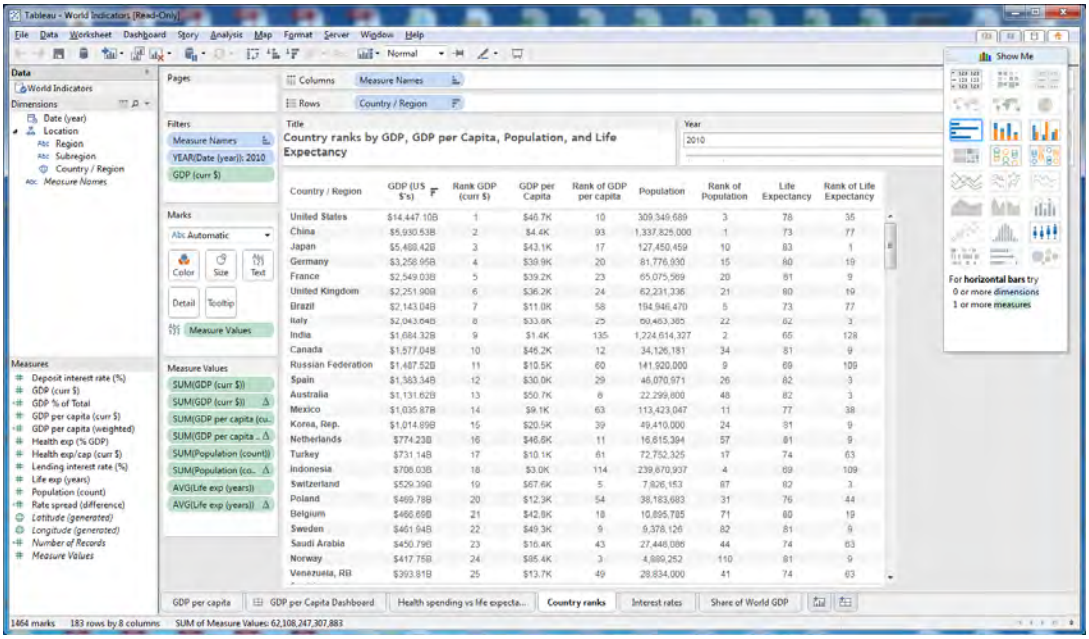

2. In the upper left corner is a section labeled Data. The data source is listed below the word Data, in this case, World Indicators
3. In the lower left corner are the data measures (columns in Excel).
    a. # means raw data
    b. =# is a calculated field
    c. ⊕ is a generated geographical field
4. In the main part of the screen is the graphic display, in this case, a bar chart where the columns are average GDP and the rows are Country
5. At the bottom are the Tableau tabs. Click on GDP per Capita Dashboard. A dashboard can have multiple views of the data simultaneously.



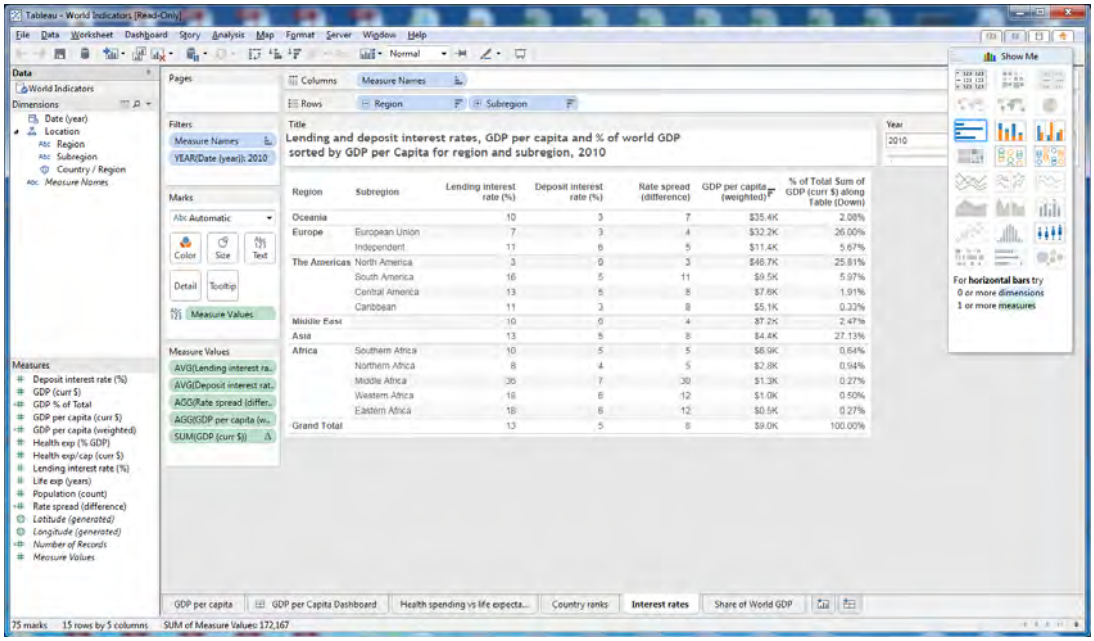
6. In this example, there is time trend data of GDP per capita over time at the top of the dashboard and a geographic representation of the data in the bottom of the dashboard.
7. Click on the tab for Health spending vs life expectancy. The X-axis is health care expenditures per capita. The Y-axis is life expectations. The countries are color-coded by continent with the legend on the left side of the screen. Which continent has the highest health care spending?



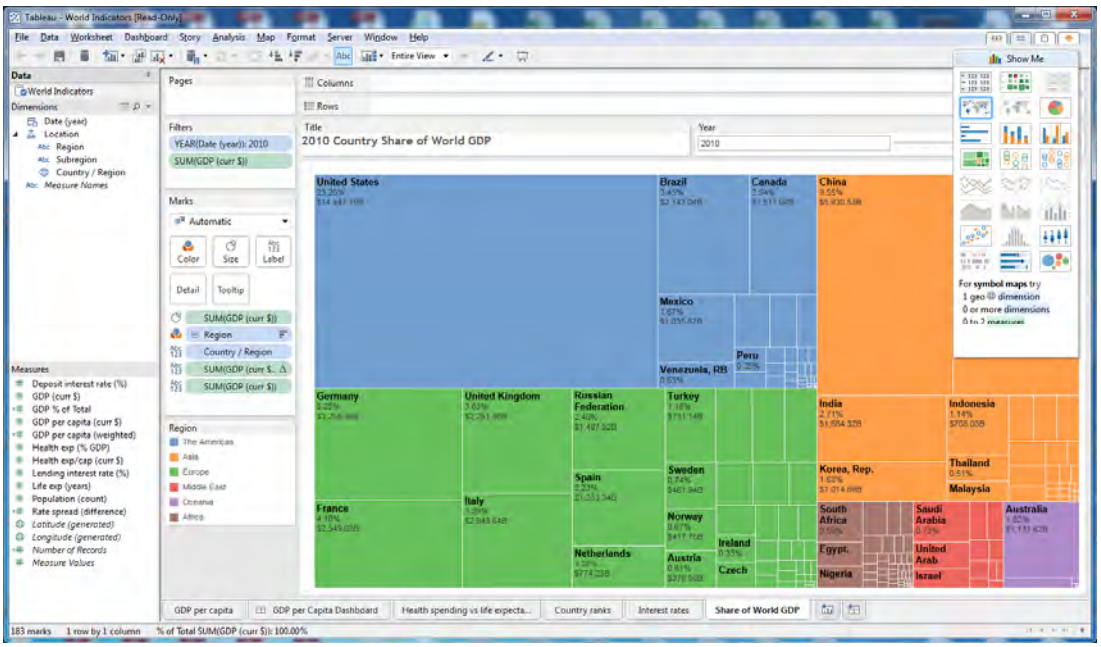

8. Click on the Country ranks tab. This is an example of tabulated data by country, similar to a Pivot Table in Excel.



9. Click on the Interest rates tab. This is similar to the previous example, but with Region and Subregion summaries.




10. Finally, click on Share of World GDP. This representation uses blocks to show the relative size of GDP by country. The color coding categorizes countries by Region.




11. Move your cursor over each box. The detailed data will pop up for each box.
12. The Show Me menu in the upper right corner displays the graphic options and describes how many dimensions and measures can be used.

## Session 10.5: Download Data Using Excel

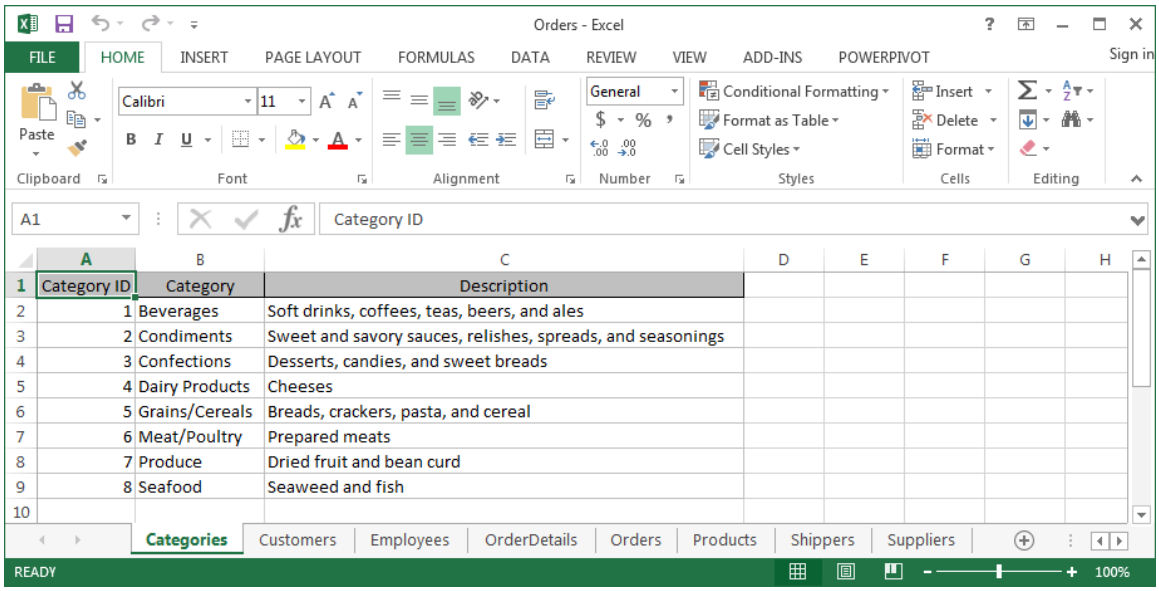
Use the Business Analytics – Week 10 Orders.xls file for this exercise.

### Opening a new Tableau Workbook

1. To open a new Tableau Workbook, click on File, New

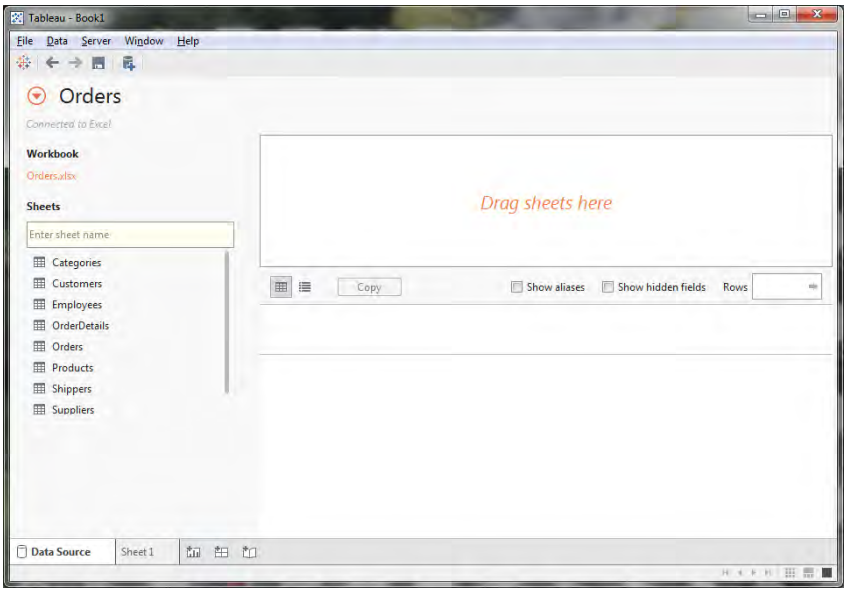
### Connecting to Data in Tableau

We will use Orders data extracted from an Excel spreadsheet to construct our dashboards. The data is shown below.



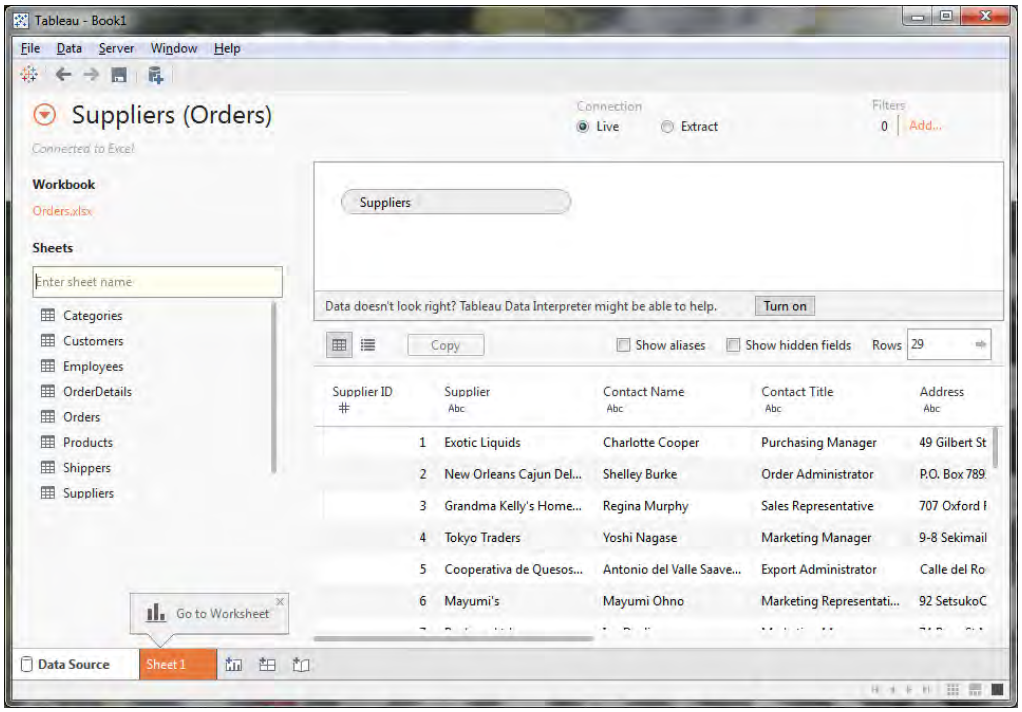

| Category ID | Category | Description |
|---|---|---|
| 1 | Beverages | Soft drinks, coffees, teas, beers, and ales |
| 2 | Condiments | Sweet and savory sauces, relishes, spreads, and seasonings |
| 3 | Confections | Desserts, candies, and sweet breads |
| 4 | Dairy Products | Cheeses |
| 5 | Grains/Cereals | Breads, crackers, pasta, and cereal |
| 6 | Meat/Poultry | Prepared meats |
| 7 | Produce | Dried fruit and bean curd |
| 8 | Seafood | Seaweed and fish |

1. To connect to the orders data, click on Data, Connect to Data
2. Under the word Connect, In a File, click on Excel
3. Locate and click on your Orders Excel spreadsheet, then click Open

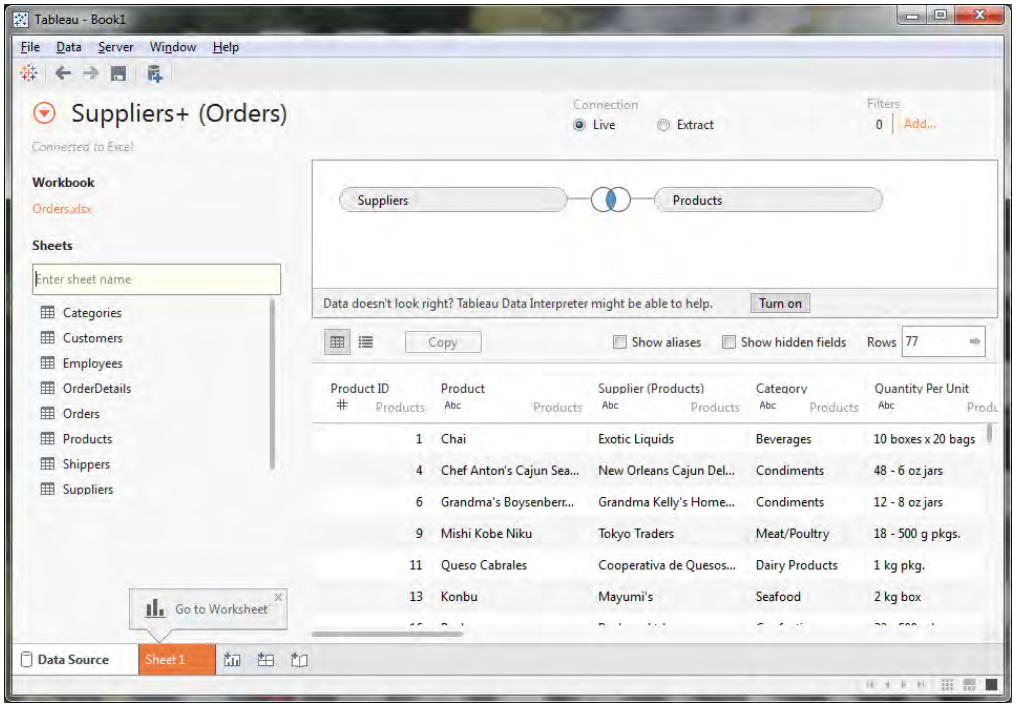


**Session 10.6: Relationships**

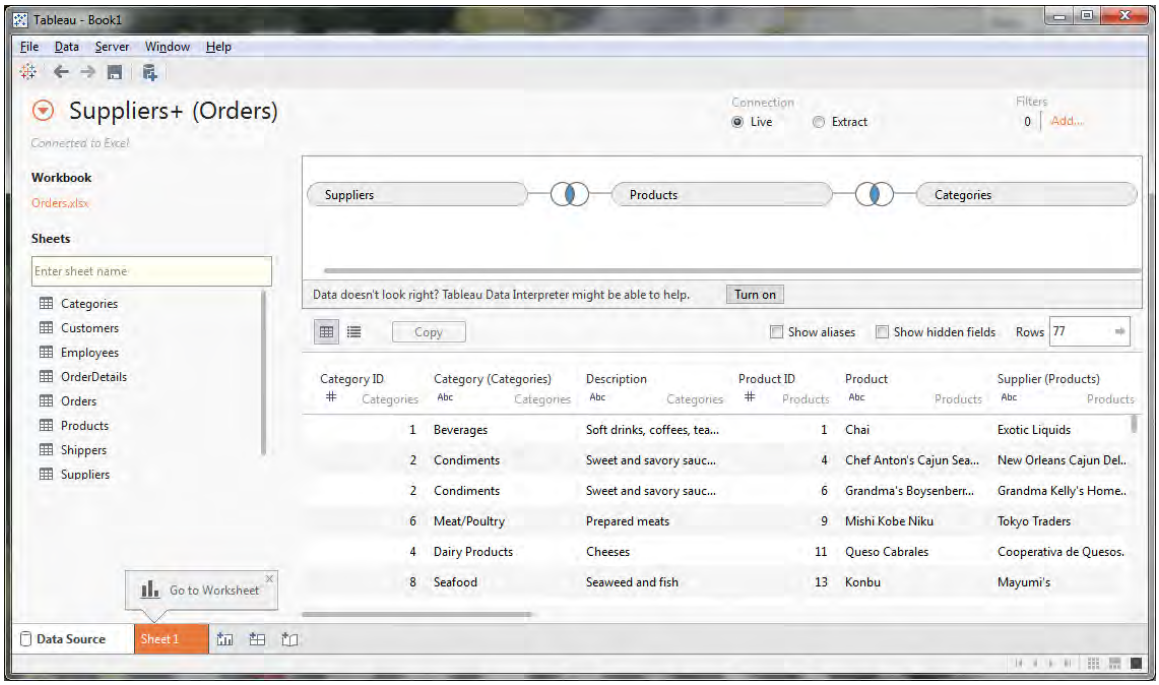
1. The worksheets from the Excel workbook are listed on the left. Double click on Suppliers. The table appears it the top box.



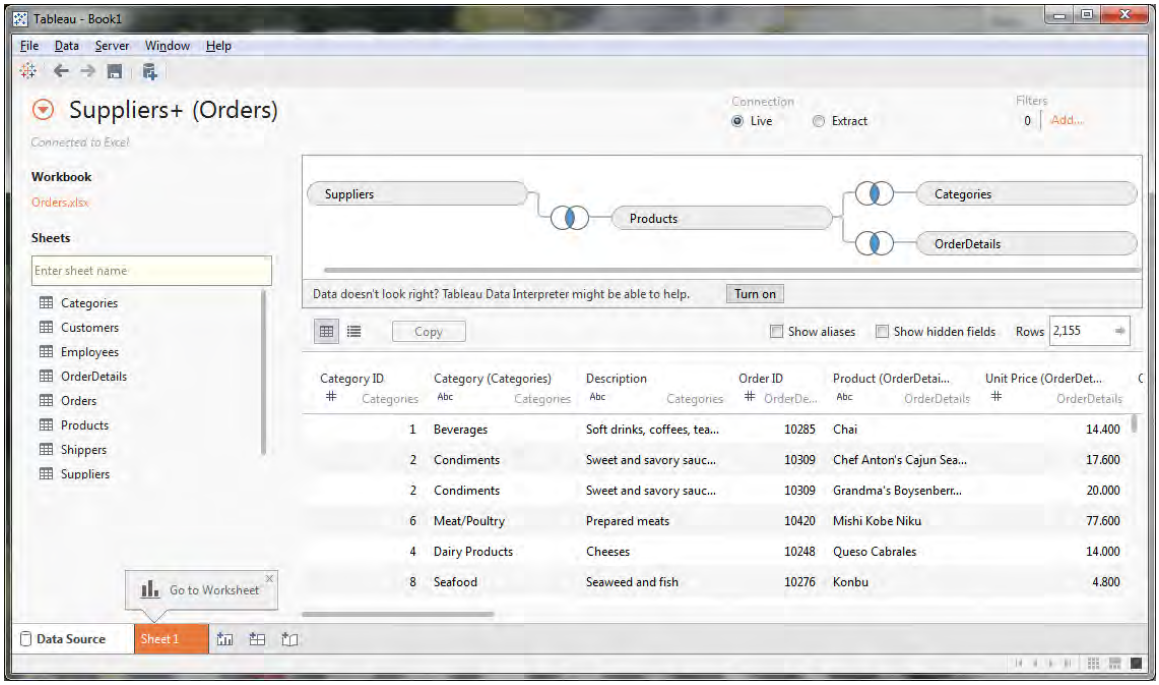

2. Next, double click on Products. Note that Products has been added to the list of tables at the top. Also, a relationship has been created. Click on the relationship. Which fields are used to create the join? What type of join is used?



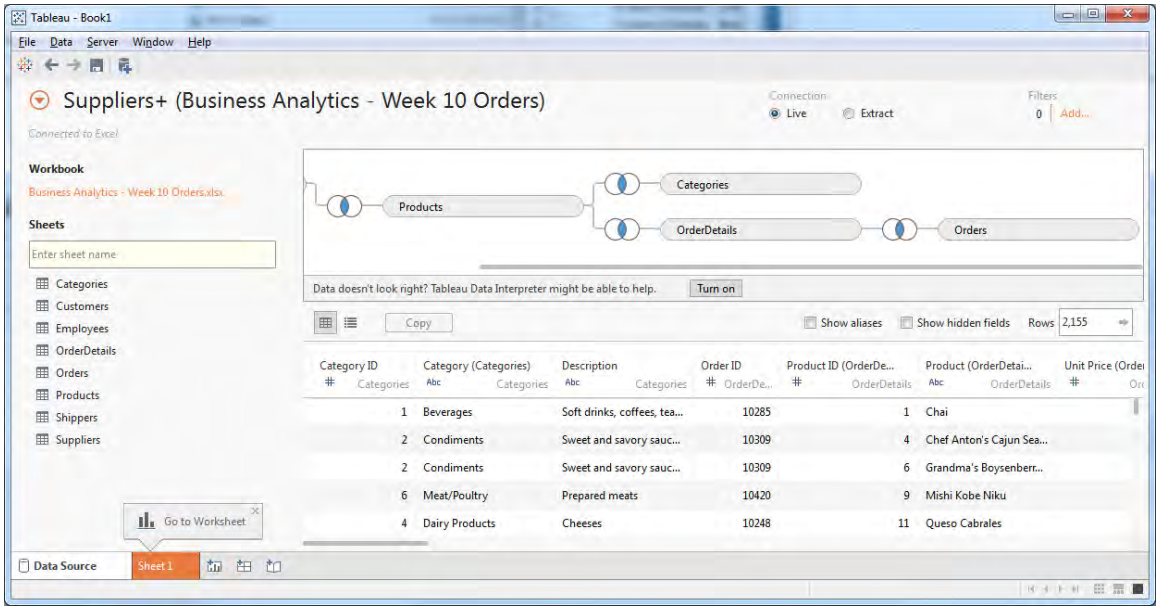
3. Next, add Categories. Again the relationship between Products and Categories is created.



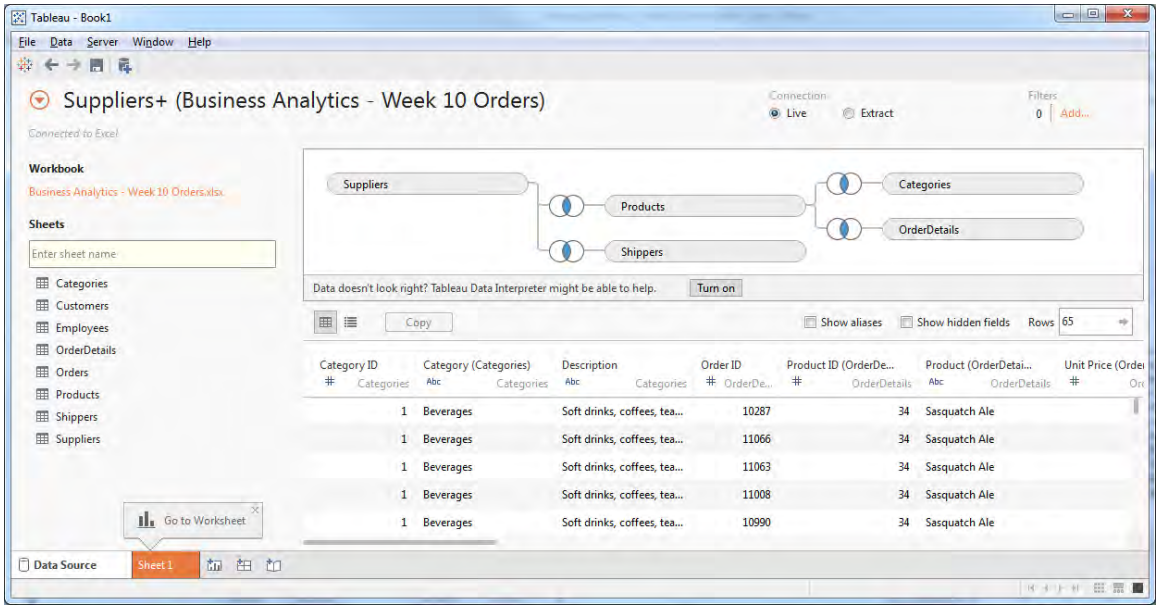

4. Next, add Order Details. The relationship links Order Details to other tables through the Products table.



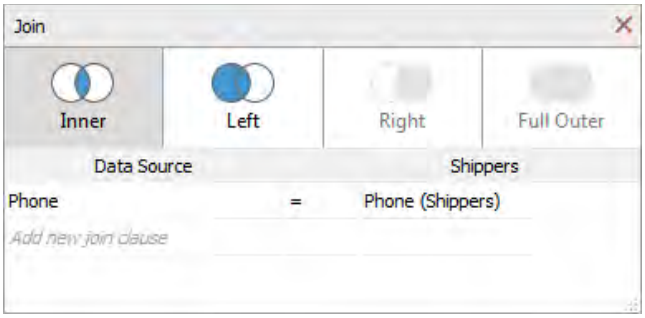
5. Next add Orders. The relationship links Orders with Orders Details



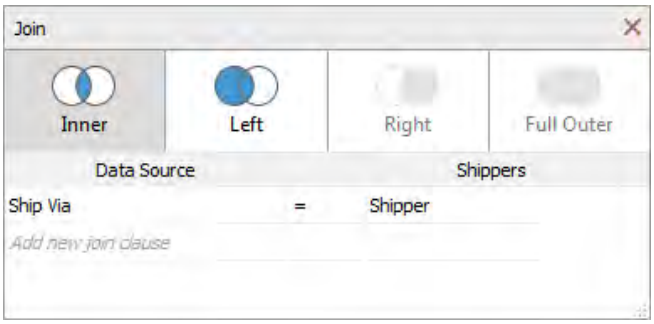

6. Finally, let's add the Shippers table.



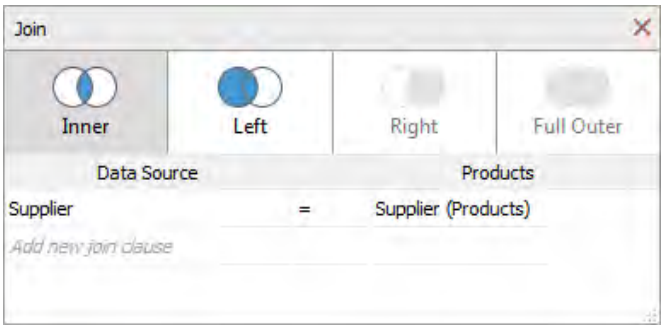
7. The Shippers table is incorrectly connected to the Suppliers table. It's linking phone numbers, although it should join using the Shipper ID.




To correct this, click on the left side, down arrow, and change Phone to Ship Via in the Orders table. Click on the right side, down arrow, and change Phone (Shippers) to Shipper in the Shippers table. You will need to manually create joins when the fields have different names.



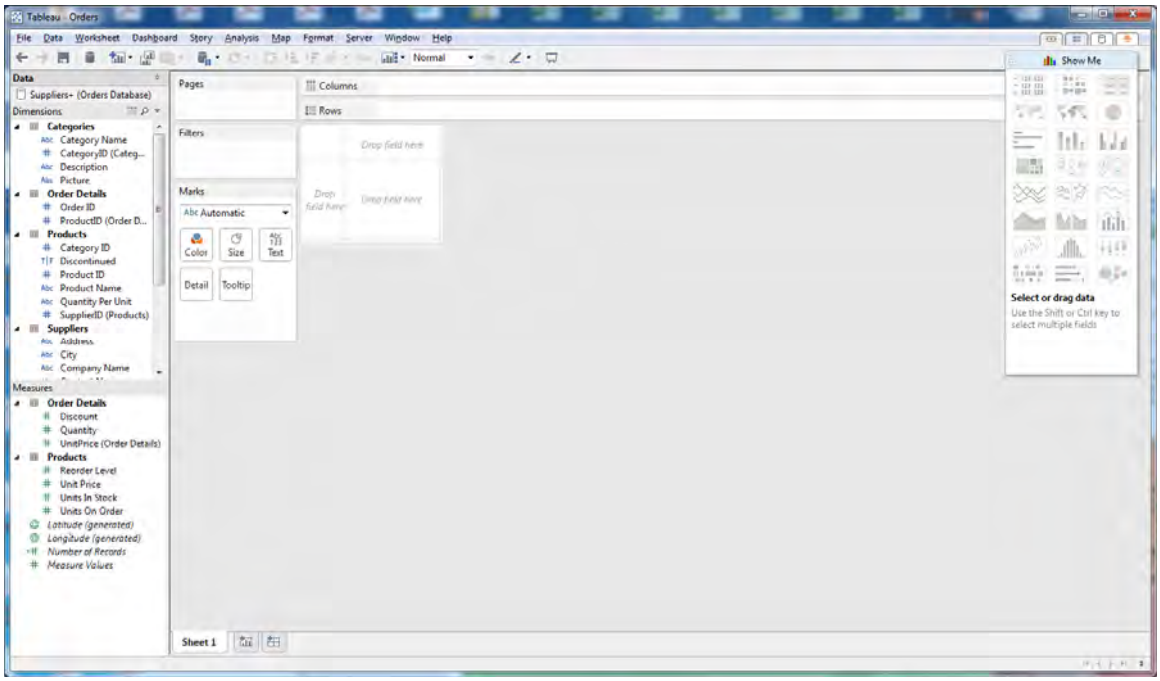

8. Click on each relationship. There are four different types of joins.
    a. Inner join: the data rows selected require that the fields (in this case, Supplier) must match from both tables; this is a simple intersection
    b. Left Join: all rows are selected from the table on the left; those rows with matching fields (Supplier) on the right table are also selected; this includes all rows from the left table and those rows from the right table that intersect
    c. Right Join: all rows are selected from the table on the right; those rows with matching fields (Supplier) on the left table are also selected; this includes all rows from the right table and those rows from the left table that intersect
    d. Full Outer Join: includes all rows from both the left and right table, inserting blanks when a value for the key field (Supplier) appears in one table but not the other



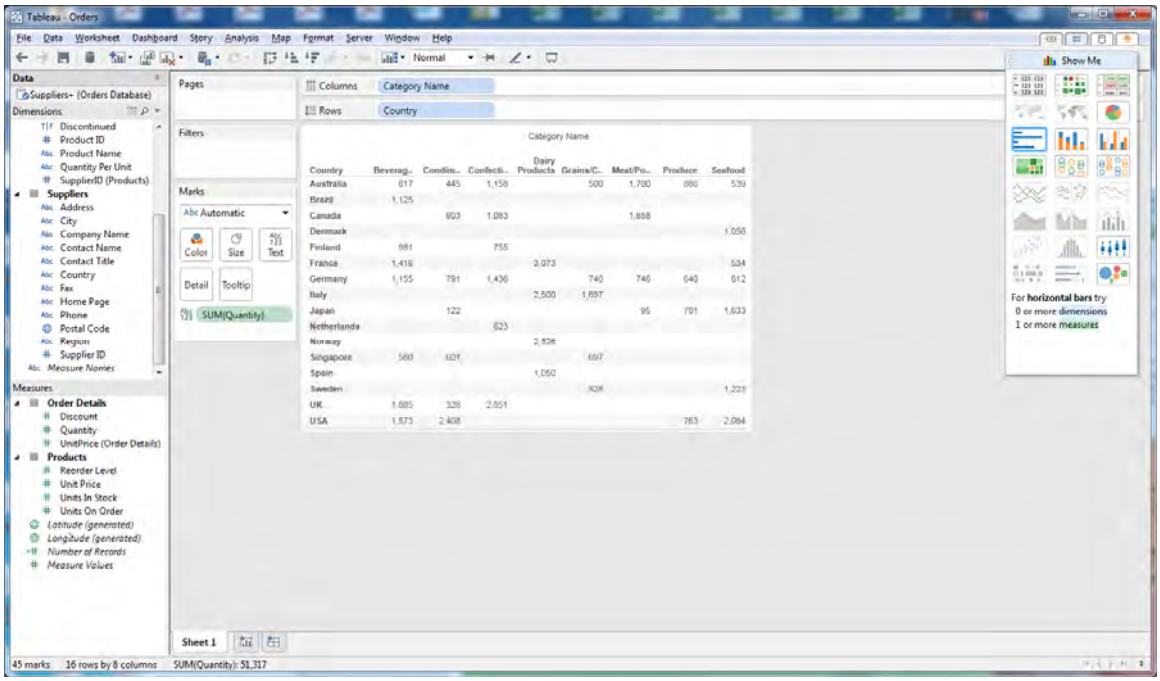

**9.** Save your workbook by clicking on File, Save As, then enter File name: Orders and click Save.

**Session 10.7: Building Worksheets**

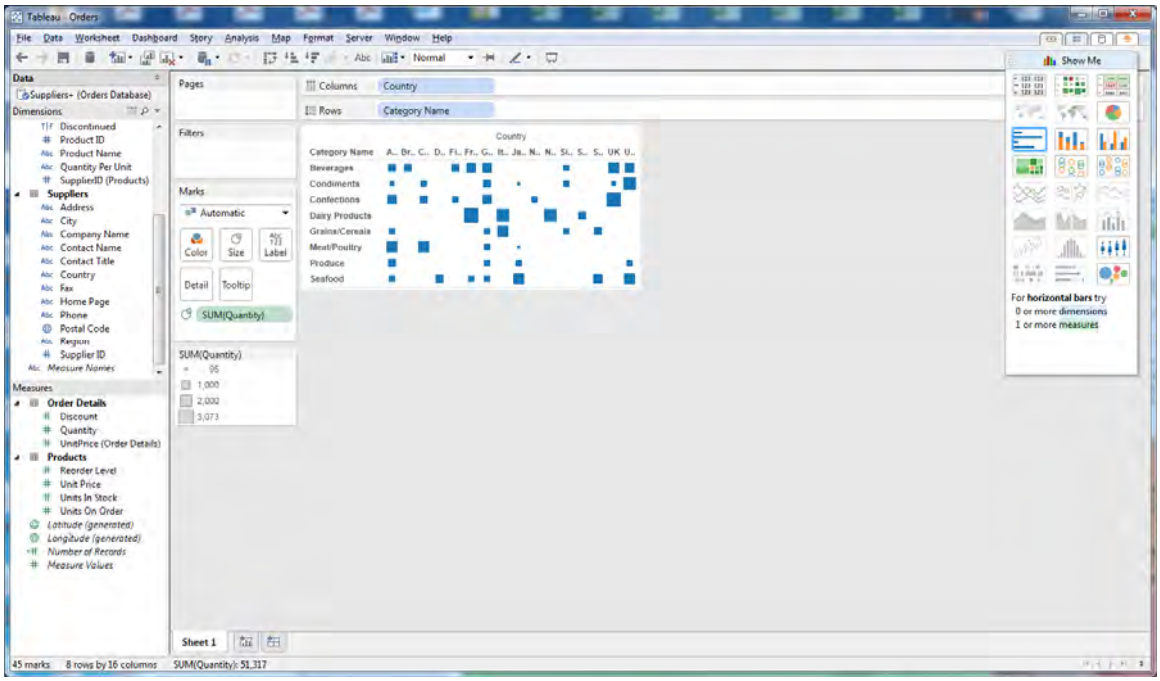
1. To return to Worksheet mode, click on Go to Worksheet (version 8.3) or Go to Worksheet – Sheet 1 (version 9.0)



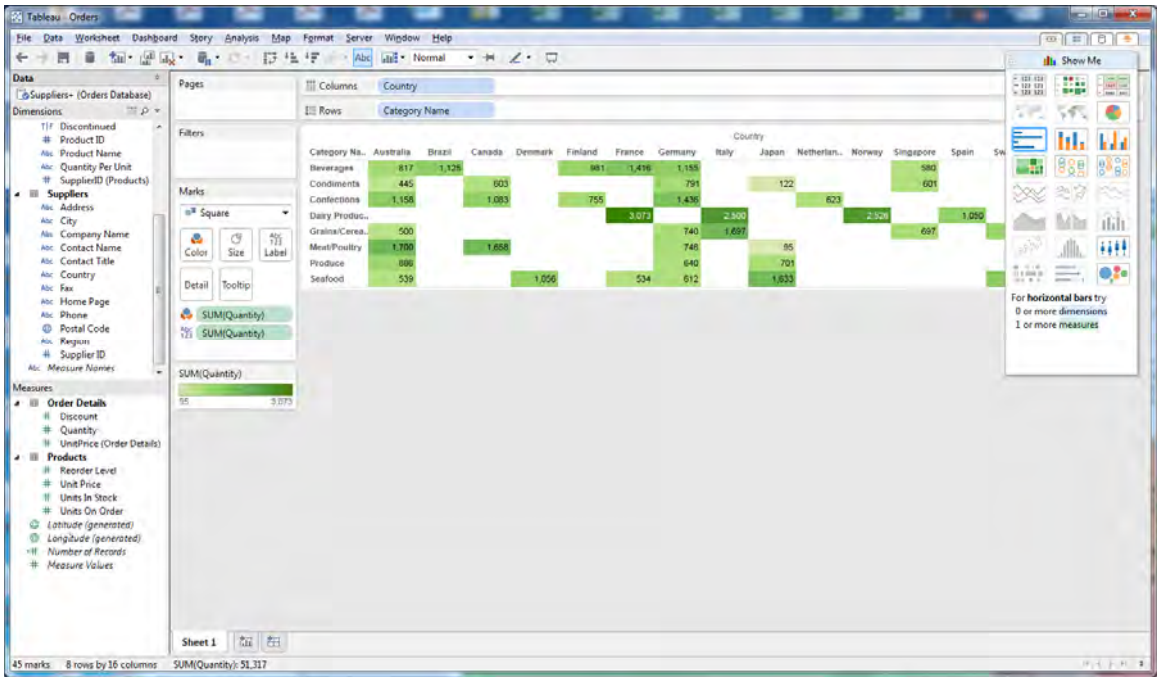

2. In the upper left, under Data, it shows that we are connected to the Orders database
3. Also in the upper left, the dimensions are listed. Dimensions are alphanumeric fields that can be used to categorize data.
4. In the lower left are the Measures. Measures are quantitative data which can be used for calculations. Only the Order Details table and the Products table have measurable data. Notice that Latitude and Longitude are generated from our raw data.
5. Near the top center of the page are sections labeled columns and rows. These will be filled in soon
6. Below the Columns and Rows sections is a Drop field here area. This is where we build our table of data (like a Pivot Table)
7. Let's build a table with quantity ordered by category and supplier country.
8. First, from Dimensions drag Categories: Category Name to Columns
9. Next, from Dimensions drag Supplier: Country to Rows
10. Finally, from Measures drag Order Details: Quantity to the main part of the table
11. You can easily swap columns and rows by clicking on the icon at the top which shows columns and rows with an arrow between the two.



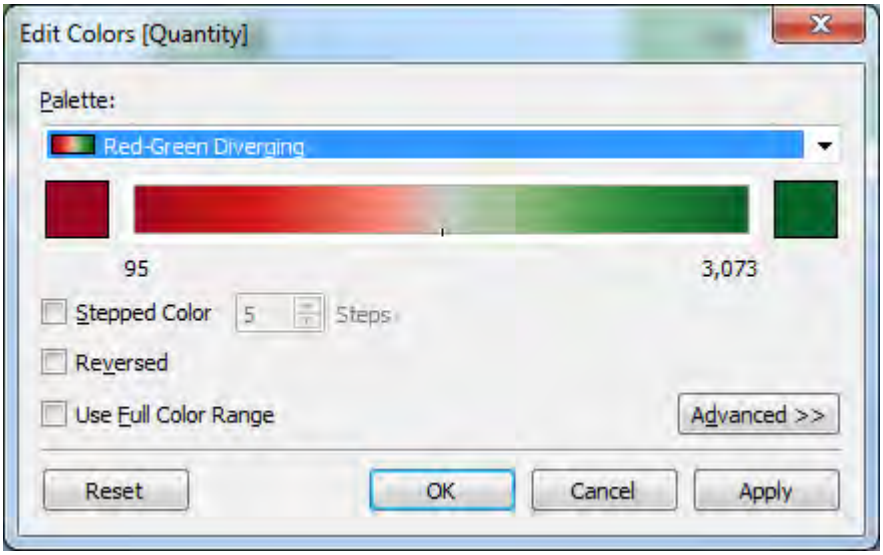

12. Let's now experiment with different formats and representations. Go to the upper right of the screen labeled Show Me. Move the cursor over each of the options in the Show Me box. The pop-up identifies how many dimensions and measures are supported by each option.
13. Click on the top middle icon in the Show Me box. This displays size graphic instead of numbers, where the legend is on the left.

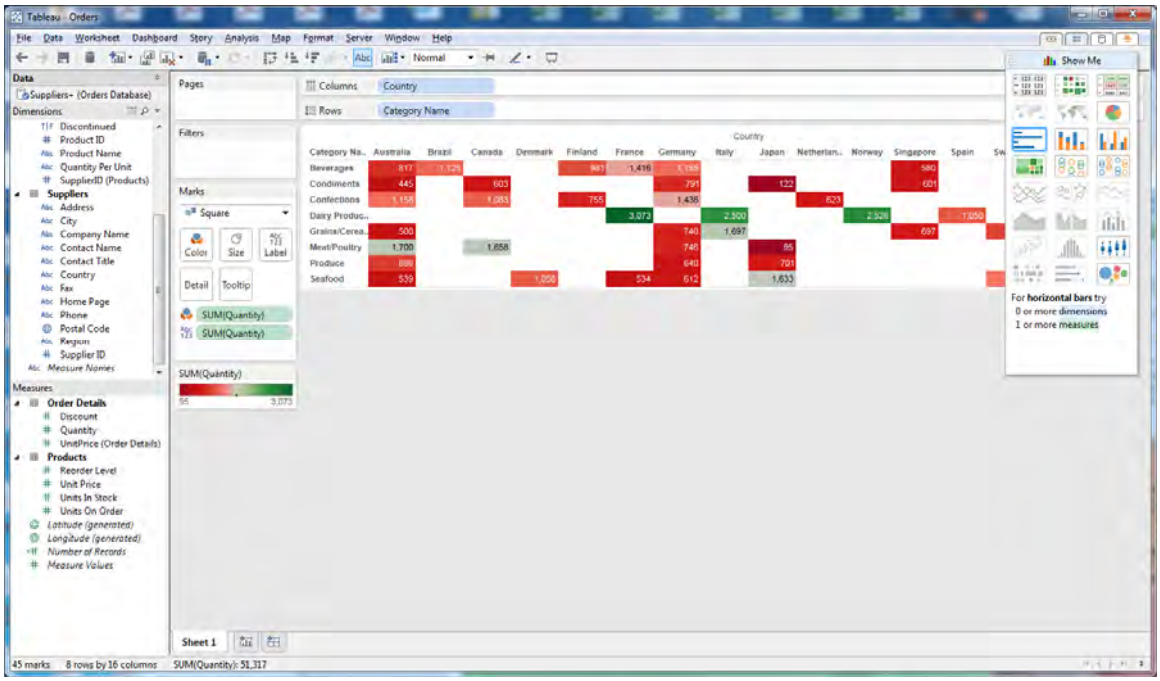



14. Next, click on the icon in the upper right of the Show Me box. This is our data table with conditional formatting applied. The legend for the color scale is on the left.

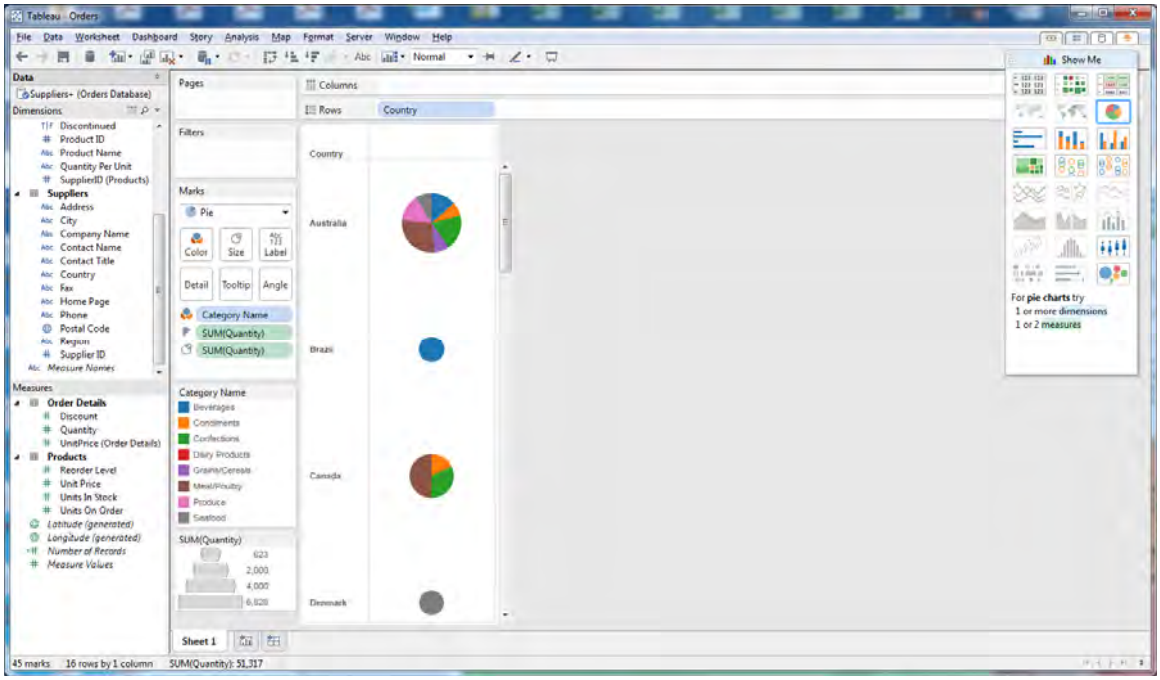



15. The colors for conditional formatting can be edited. Move the cursor over SUM(Quantity) which displays the legend, click the down arrow, then click Edit Colors.
16. Change the Palette from Automatic to Red-Green Diverging

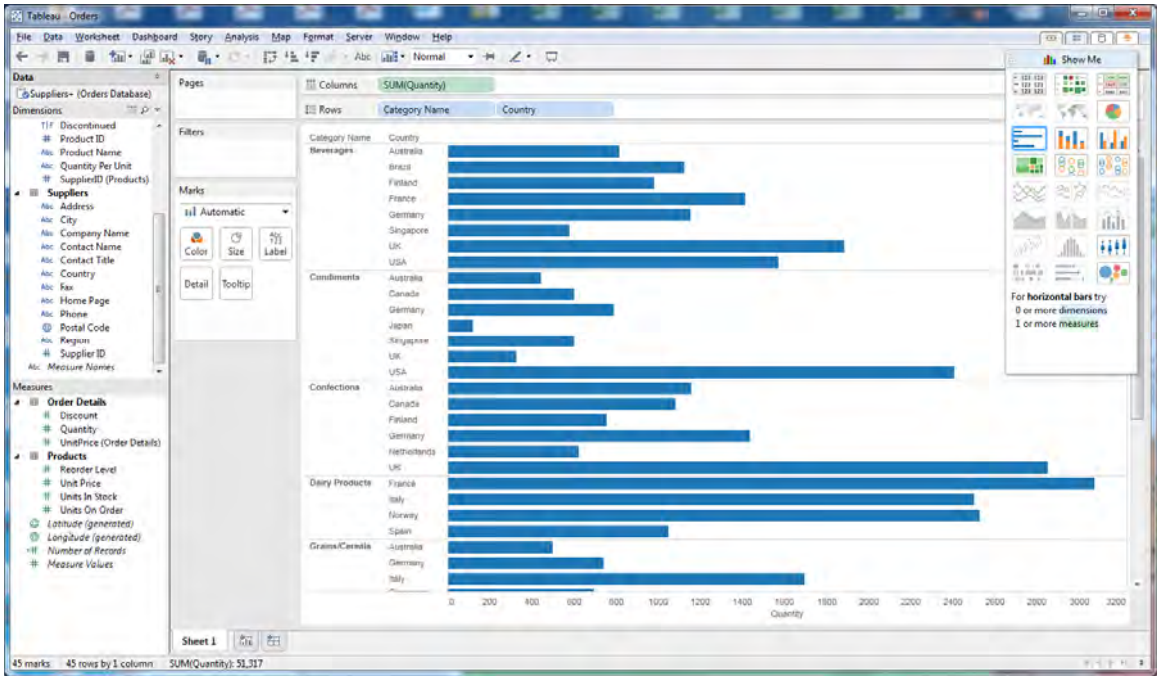




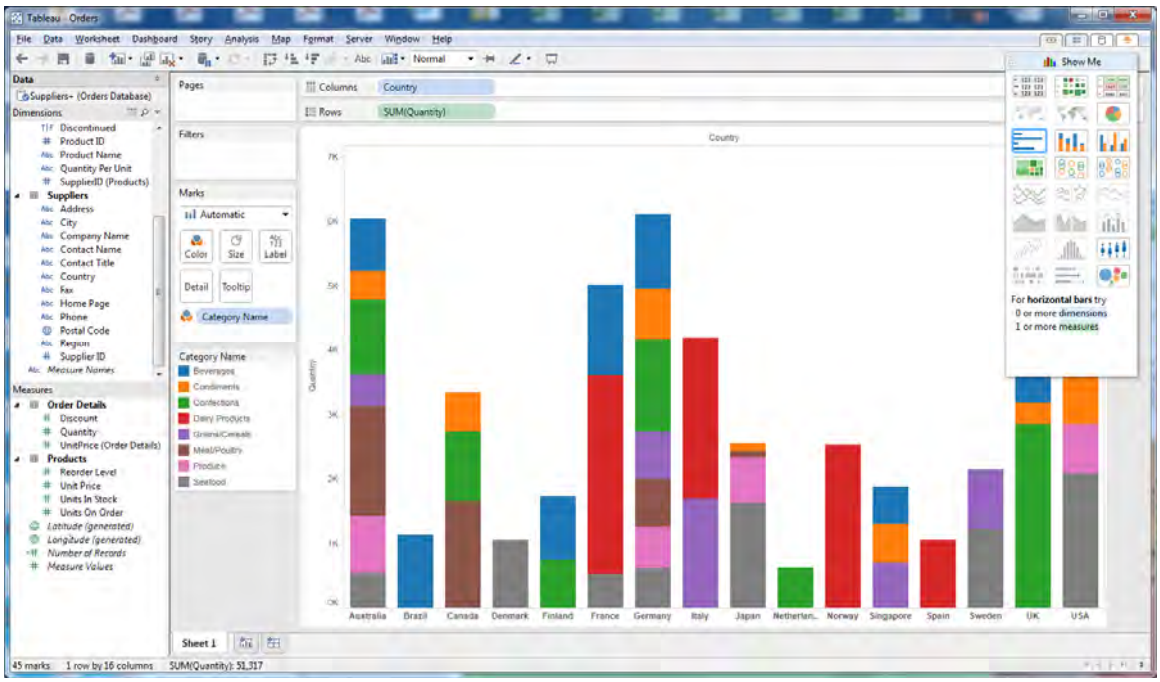

17. Next, try the Pie Charts in Show Me. Note that all data is collapsed to a row with a pie chart for each row



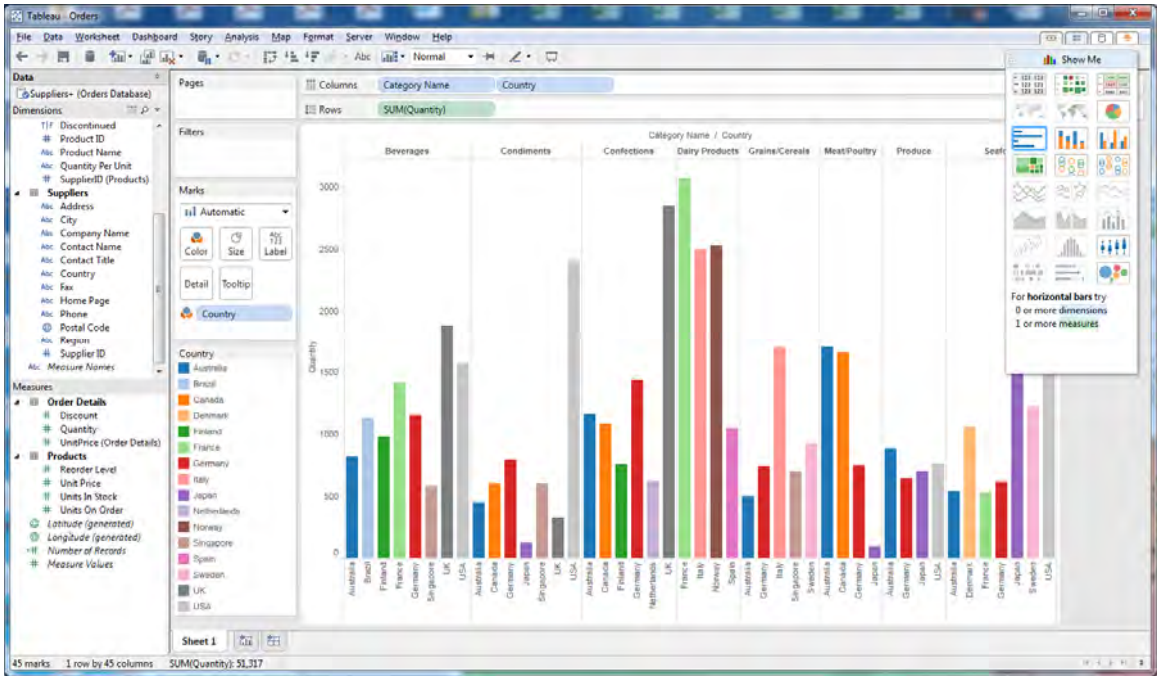
18. Horizontal bar charts: Categories are countries are grouped. Drag Category Name after Country; what happens?



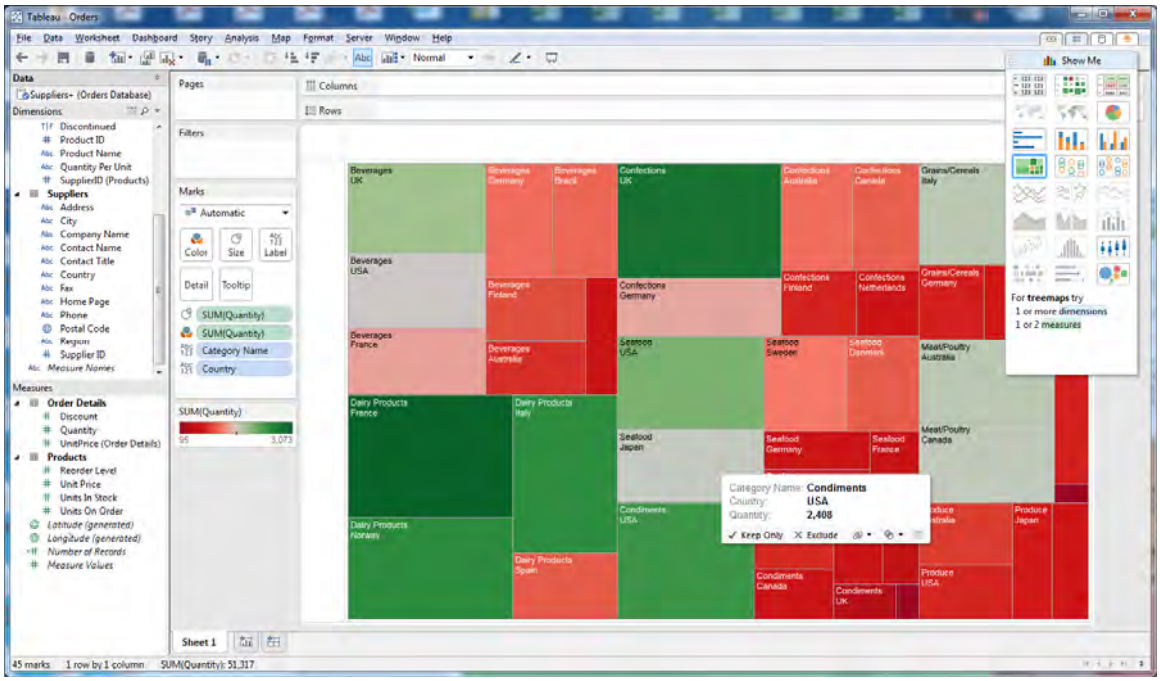

19. Stacked bar charts: Click on Confections in the Legend. What happens? Hold down the control key and click on Beverages, Confections, and Dairy Products. What happens?



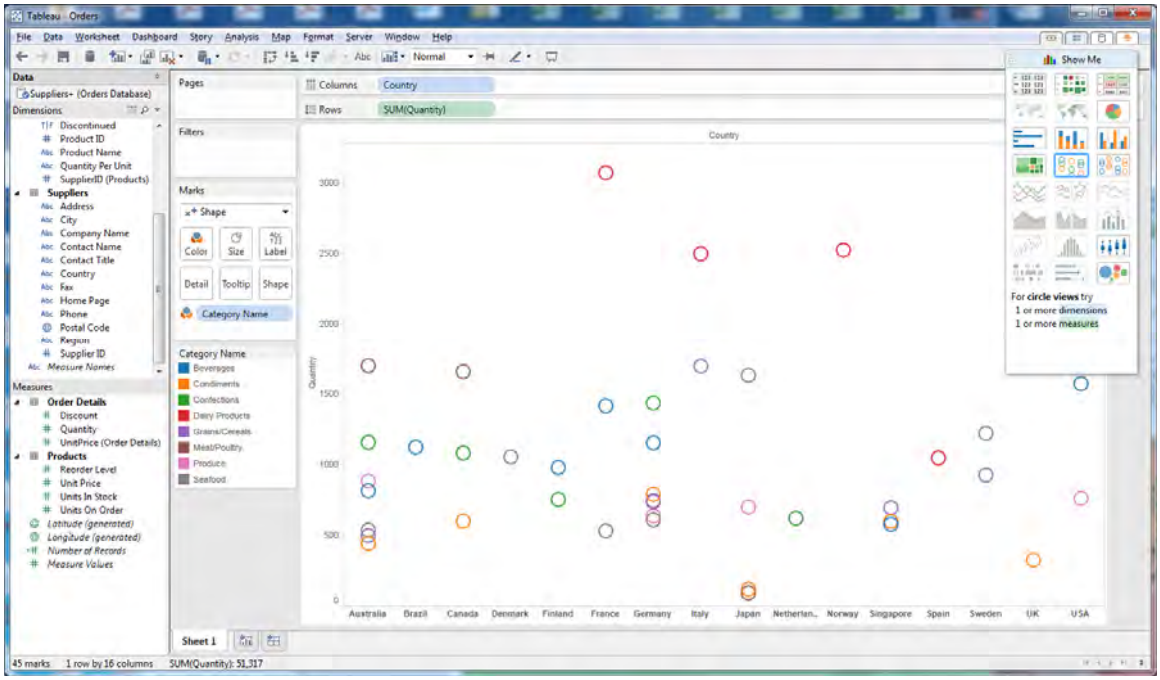
20. Side-by-side bar charts: Click on Beverages at the top of the graph.



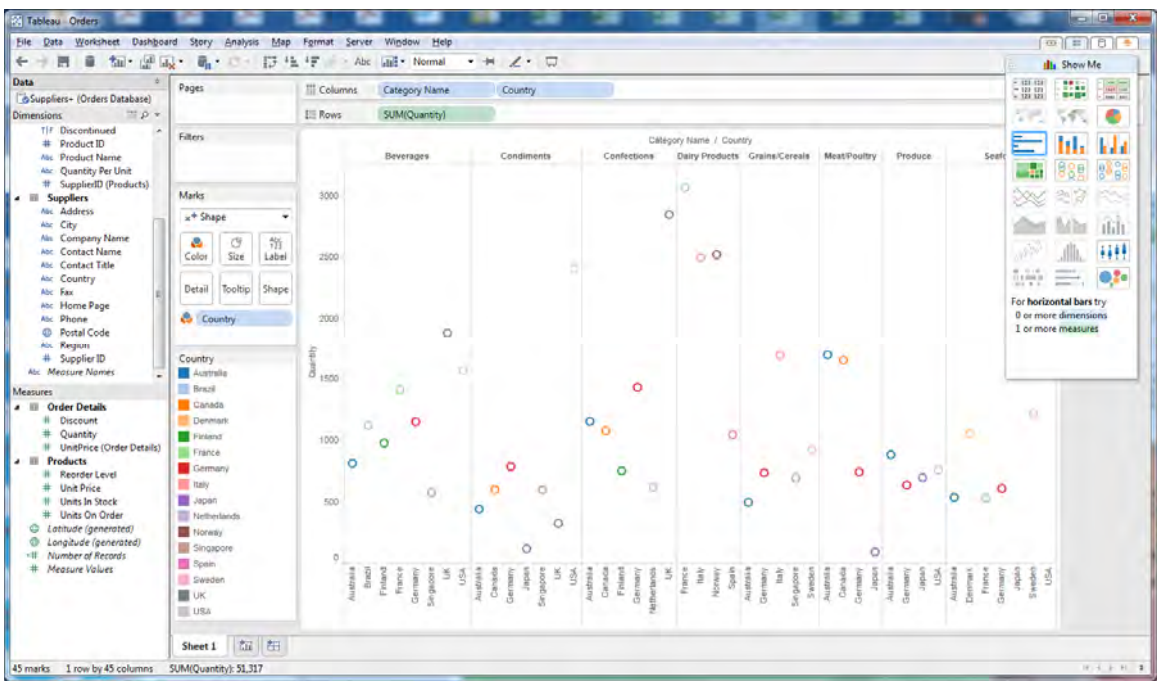

21. Treemaps display data by translating the numeric values into an area.



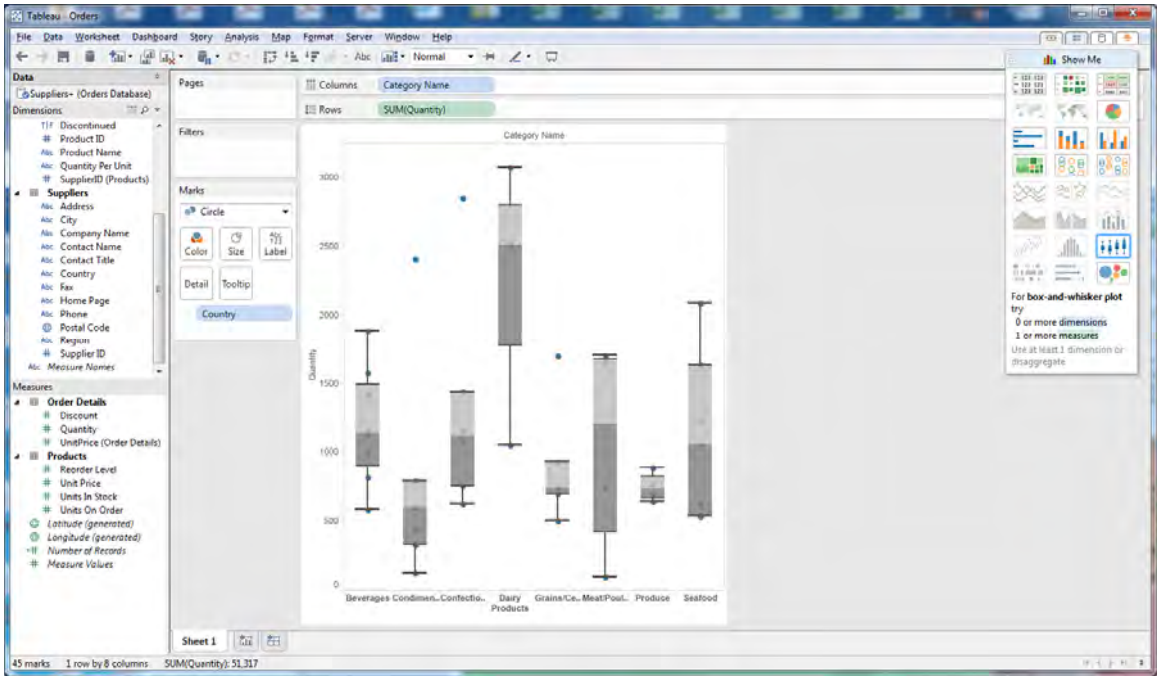
22. Circle views display the quantity of product by country, with the color of circle reflecting the category name.



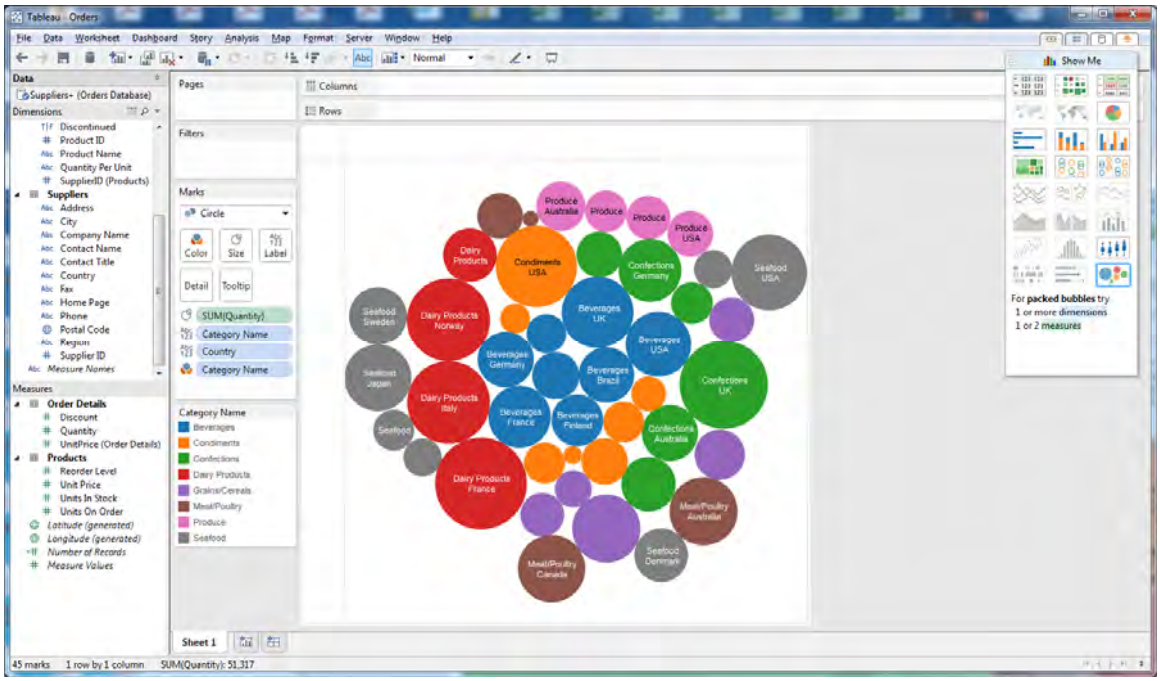

23. Side-by-side circle view swaps the dimensions



24. Box and whisker portray data ranges from minimum, 25%-ile, median, 75%-ile, and maximum.




25. Packed bubbles use the area of a circle (bubble) to reflect the magnitude of the numbers

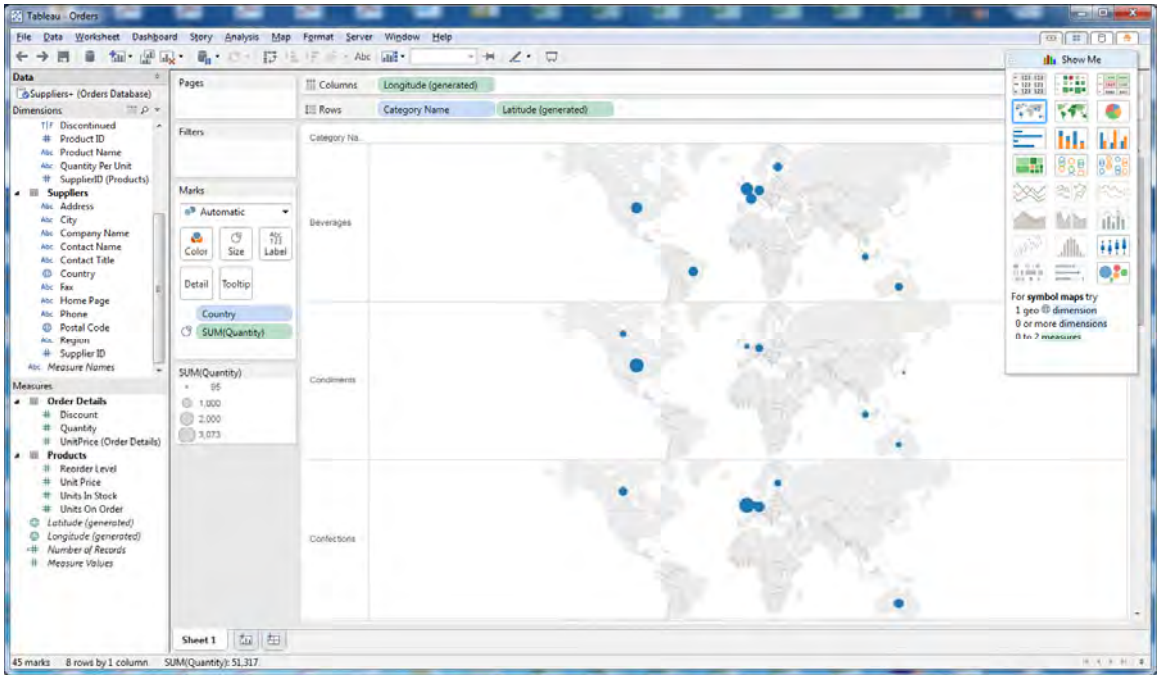


**Session 10.8: Adding New Geolocations**

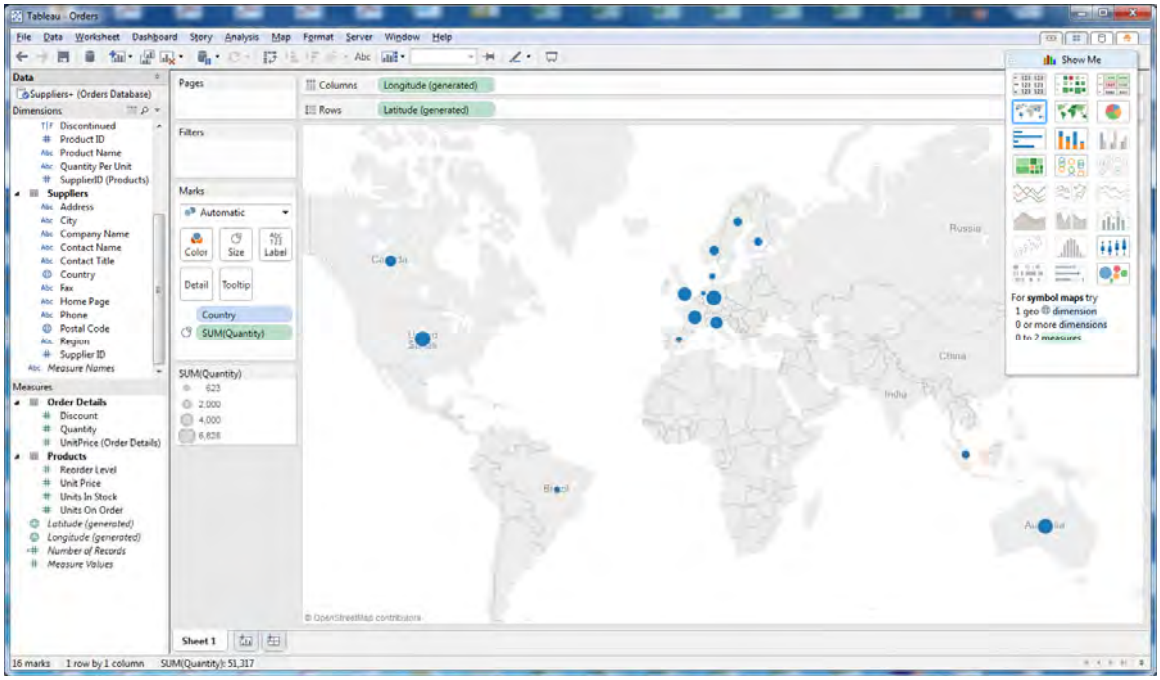
**In version 8.3, you will need to set up Geolocations. Version 9.0 automatically detects these. Steps 1, 2, and 3 will not be necessary if you have version 9.0.**

Several of the graphs in Show Me were grayed out. This is because data was not available to plot specific graphs. Move your cursor over the pictures of world maps. Each requires a geo dimension. In reviewing the dimensions on the left side of the screen, only Postal Code has the symbol for geo dimension. Tableau will automatically scan data and try to identify geo dimensions, but it's not perfect. Other geographic locations include City and Country, but they are listed as Abc, which means alphanumeric data not associated with a location. Let's convert Suppliers: Country to a geo location.

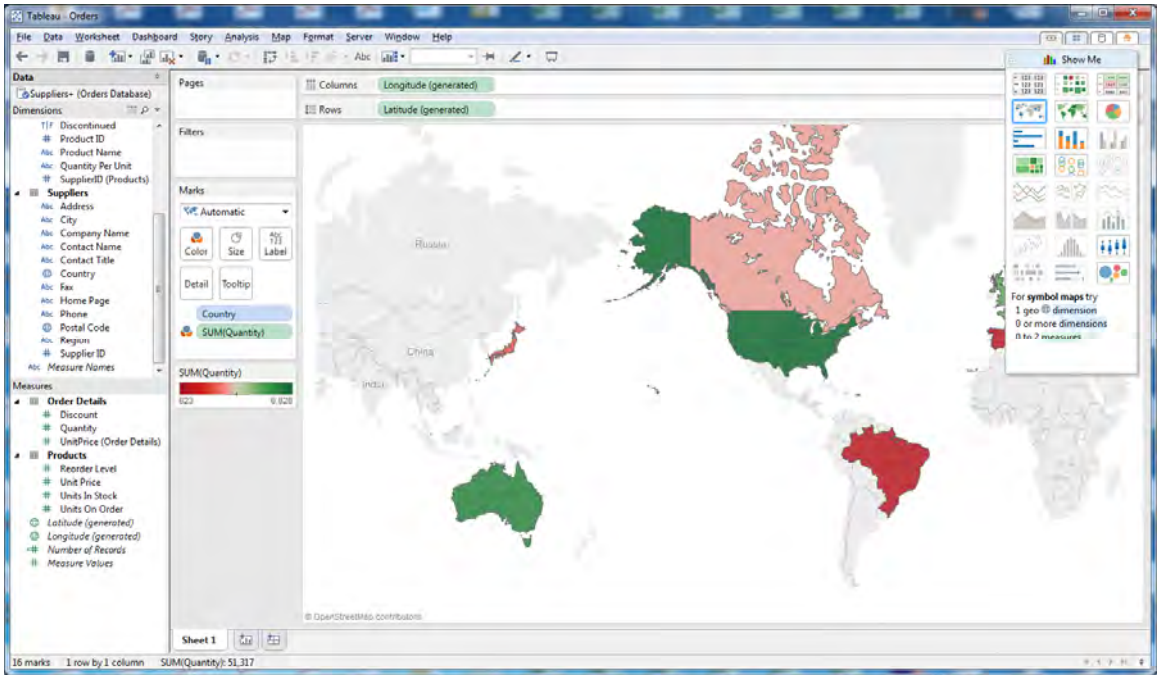
1. Put your cursor over Country in the Suppliers table. A drop down arrow should appear.
2. Click on the drop down arrow and select Geographic Role, then Country/Region
3. Note that the symbol in front of Country has changed to a geo location. Also note that the map icon under Show Me are now available.
4. Click on the map icon on the left of Show Me. The charts are for each Category.



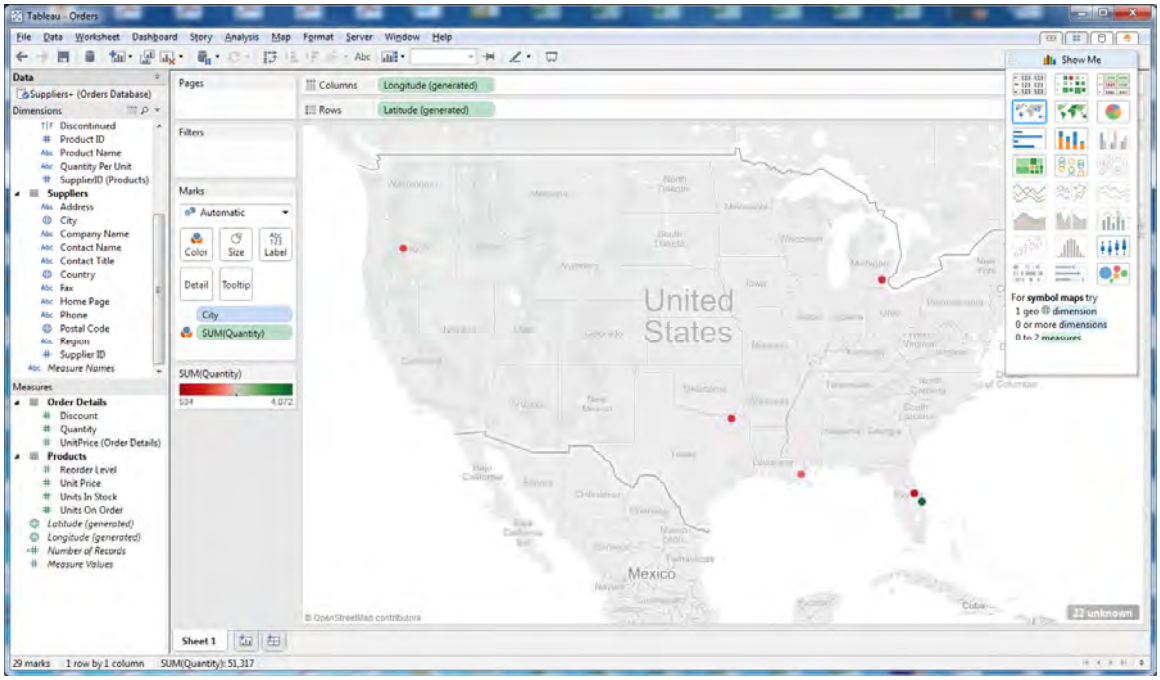

5. Click on Category Name in Rows, then delete. How did the graph change?




6. Now click on the second map icon in Show Me.




7. If you wanted to use City instead of Country, what would you do differently?
    a. Click on City, Geographic Role, City
    b. Drag City from Suppliers and drop over Country in Marks

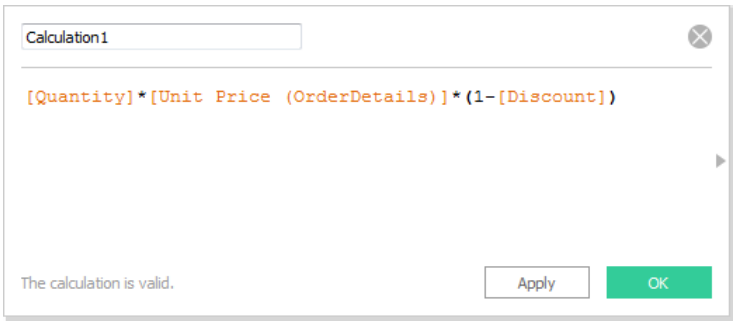



8. What limitation does automatic translation of cities reveal? Hint: are there any international cities?
9. Change back to Country.

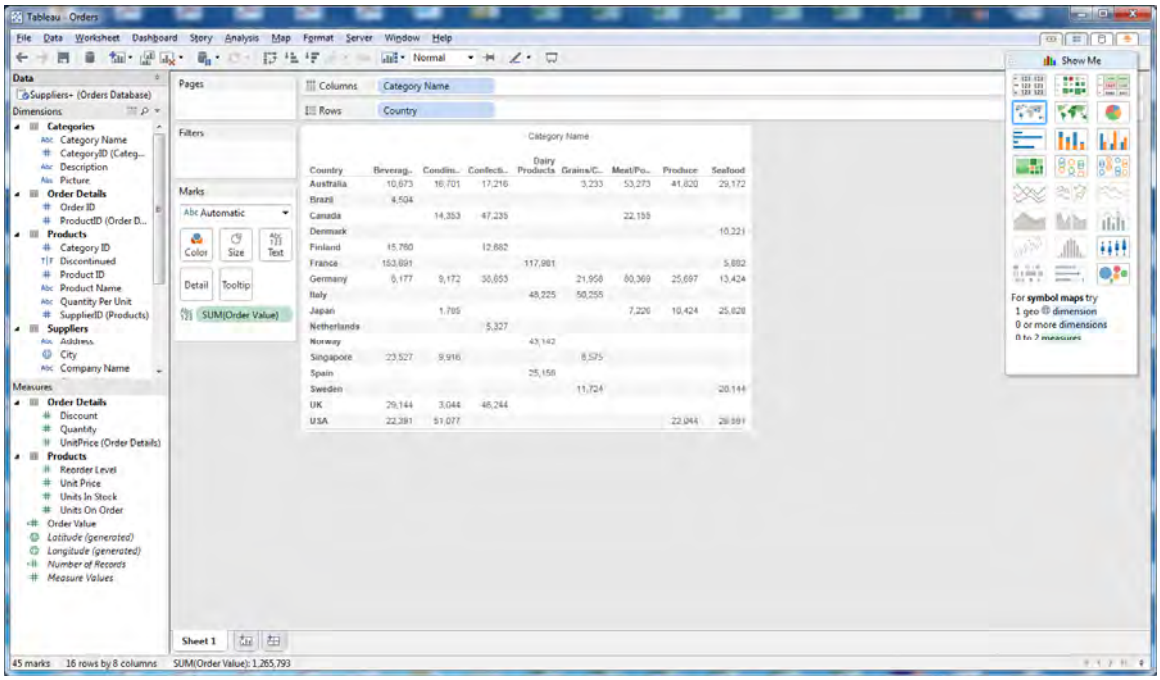
**Session 10.9: Adding Calculations**

Sometimes you will want to create new values which are calculated from data in the worksheet. For example, let's calculate the value of an order, which is Quantity * UnitPrice * (1-Discount).

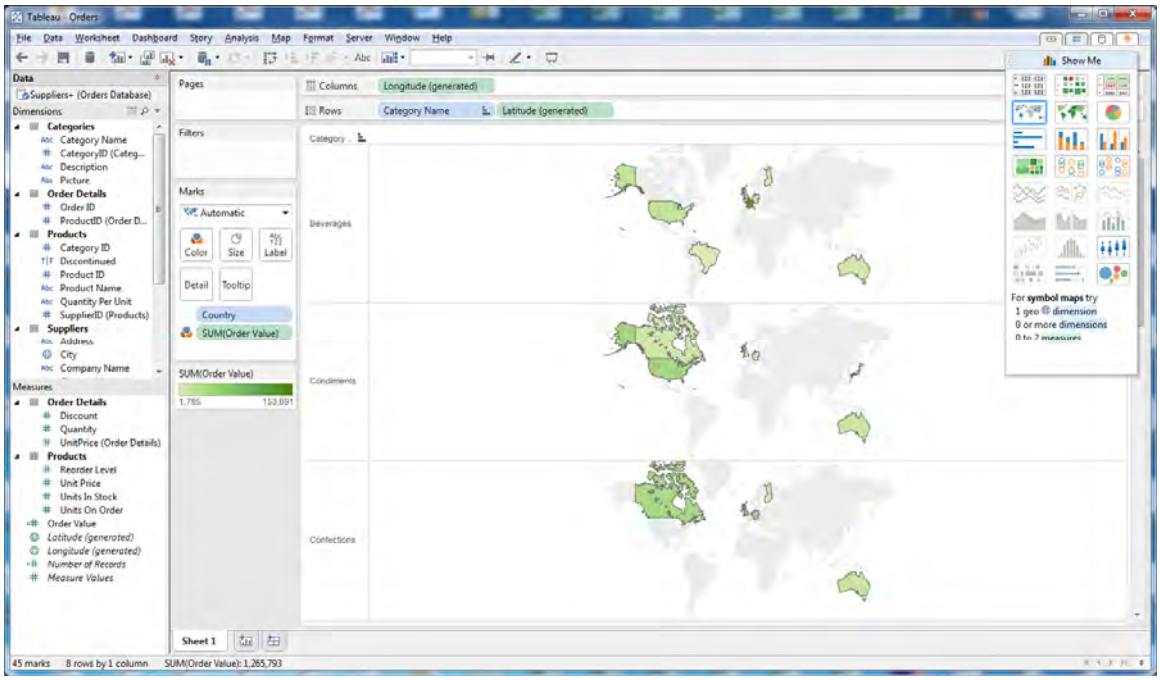
1. At the top of the worksheet, click on Analysis, Create Calculated Field.
2. Replace the name Calculation1 with Order Value
3. Drag Quantity in the blank area below the title Order Value
4. Enter *
5. Drag UnitPrice [Order Details] into the box
6. Enter *, then (1-
7. Drag Discount into the box
8. Enter ), then click OK.
9. Order Value now appears in Measures



1. Return to the table format in Show Me (upper left corner). Make sure that Category is in Columns, Country is in Rows, and Order Value is in the table.

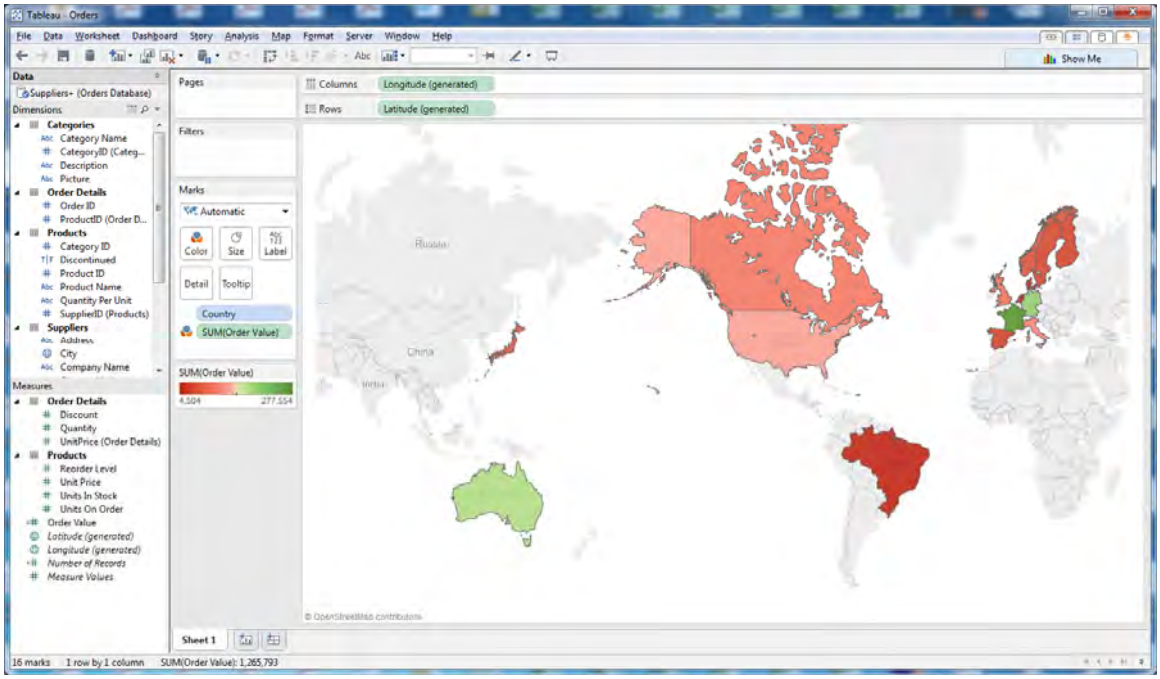



2. Change to map format (use map icon in Show Me)

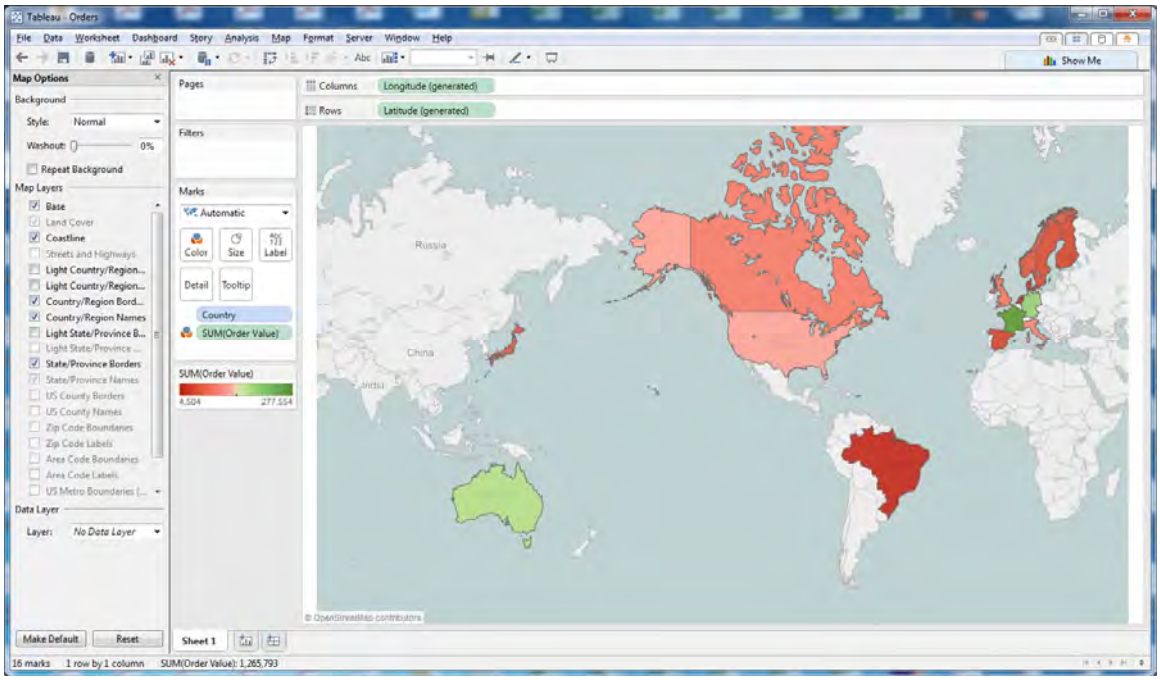



3. Drop Category Name by clicking on it and pressing the delete key to consolidate for all categories. To hide the Show Me box, click on the words Show Me.

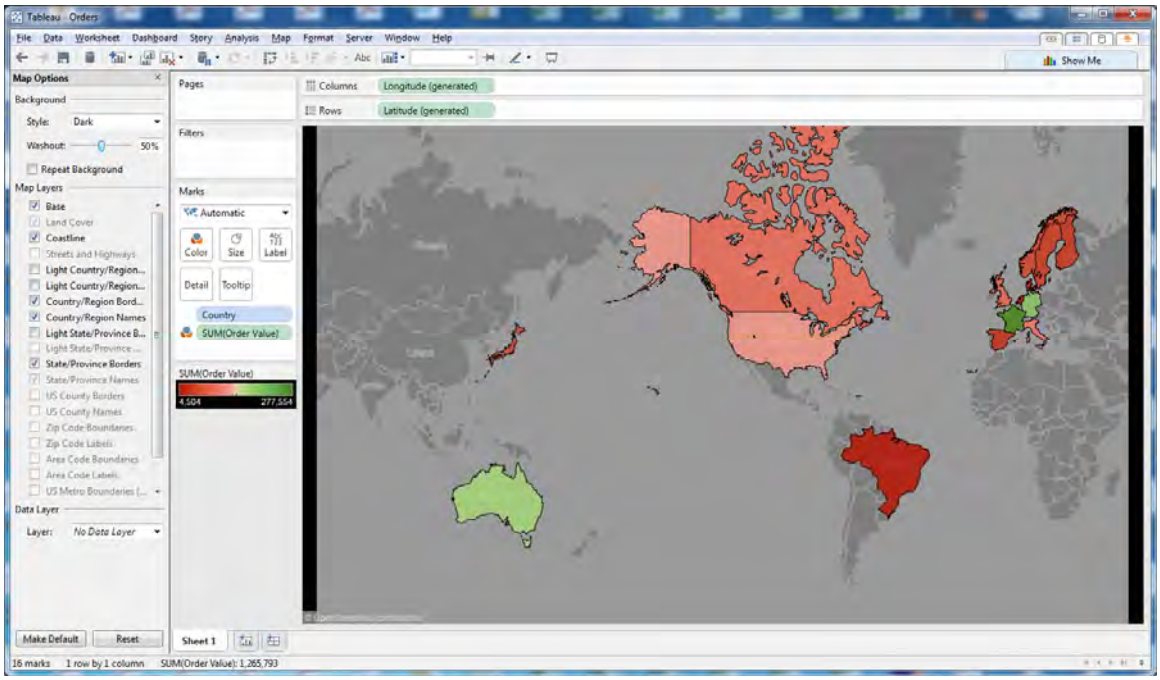
4. Change the colors by clicking on the drop down arrow next to SUM(Order Value), Edit Colors, then Area Red-Green Diverging.




5. There are a variety of format options for maps. Click on Map at the top of the screen, then Map Options. Change Background Style to Normal and check the Coastline box.



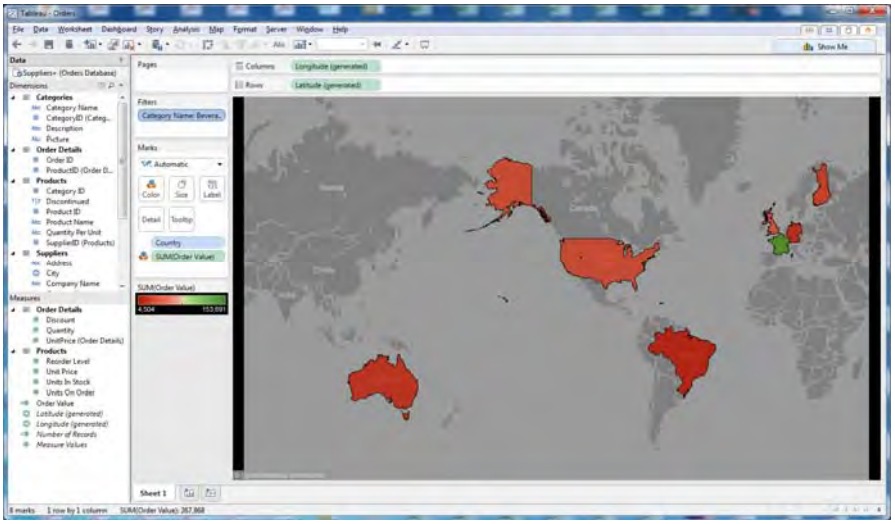
6. Change Background Style to Dark, Washout to 50%.



**Session 10.10: Filters**

Just like Excel, you can add filters to your data. Using the previous map of Order Value, let's add a filter for Category.

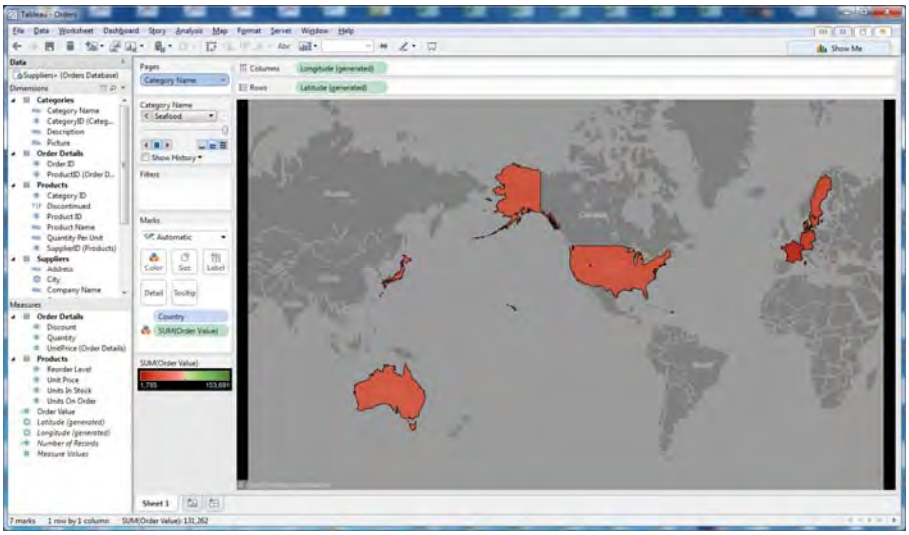
1. Drag Category Name to the Filters box. Click on All, then OK. The map should not change, but Category Name will remain in the Filters box.
2. In the Filters box, place your cursor over Category Name, click the down arrow, then Filter
3. Uncheck everything except Beverages. (You can click None, then Beverages). Click OK.
4. The new map only reflect Beverages.



**Pages**

The Pages feature allows you to perform a filter, with each Category on a different page.

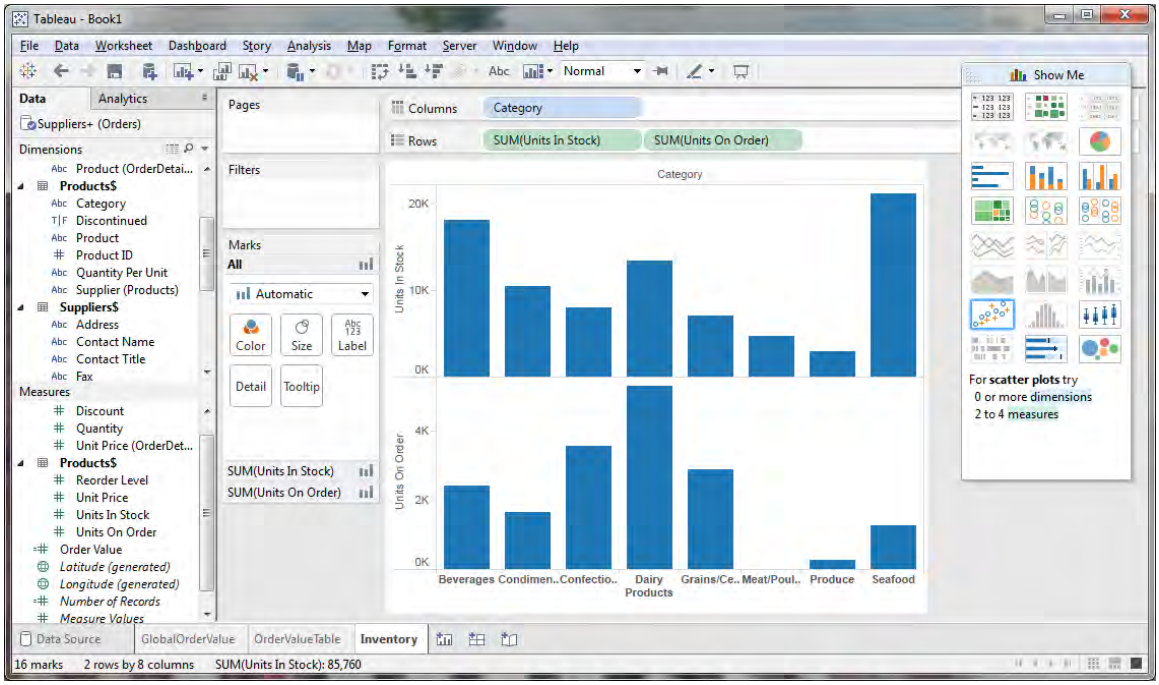
1. Drag Category Name to the Pages box.
2. The arrows below are play buttons. Click on the right arrow to play through the pages.
3. Click on the picture of a screen for full screen presentation) mode; escape to return.



**Session 10.11: Building Dashboards**

In this section, we will move from building worksheets to building dashboards. A dashboard is simply a collection of worksheets.

1. At the bottom of the screen is the label Sheet 1. Rename this by double clicking on Sheet 1, the renaming it GlobalOrderValue
2. Next to GlobalOrderValue is an icon for New Worksheet. Click on the icon. Sheet 2 should appear.
3. Drag Category to Columns, Country to Rows, and Order Value to the center of the table
4. Rename Sheet 2 OrderValueTable
5. Click on New Worksheet again. Sheet 3 should appear.
6. Drag Category to Columns
7. Drag Units in Stock to Rows, then also drag Units on Order to Rows
8. Rename Sheet 3 Inventory




9. To create a dashboard, click on the icon for New Dashboard
10. On the left side of the screen, each of your Worksheets should be listed.
11. Drag the Inventory worksheet to “Drop sheets here”
12. Drag OrderValueTable to the bottom of the screen
13. Drag GlobalOrderValue to the lower right of the screen
14. Notice that OrderValue legend is taking up the entire right side of the screen. Right click on the Legend, then click Floating
15. Click on the floating legend and drag it into the map.

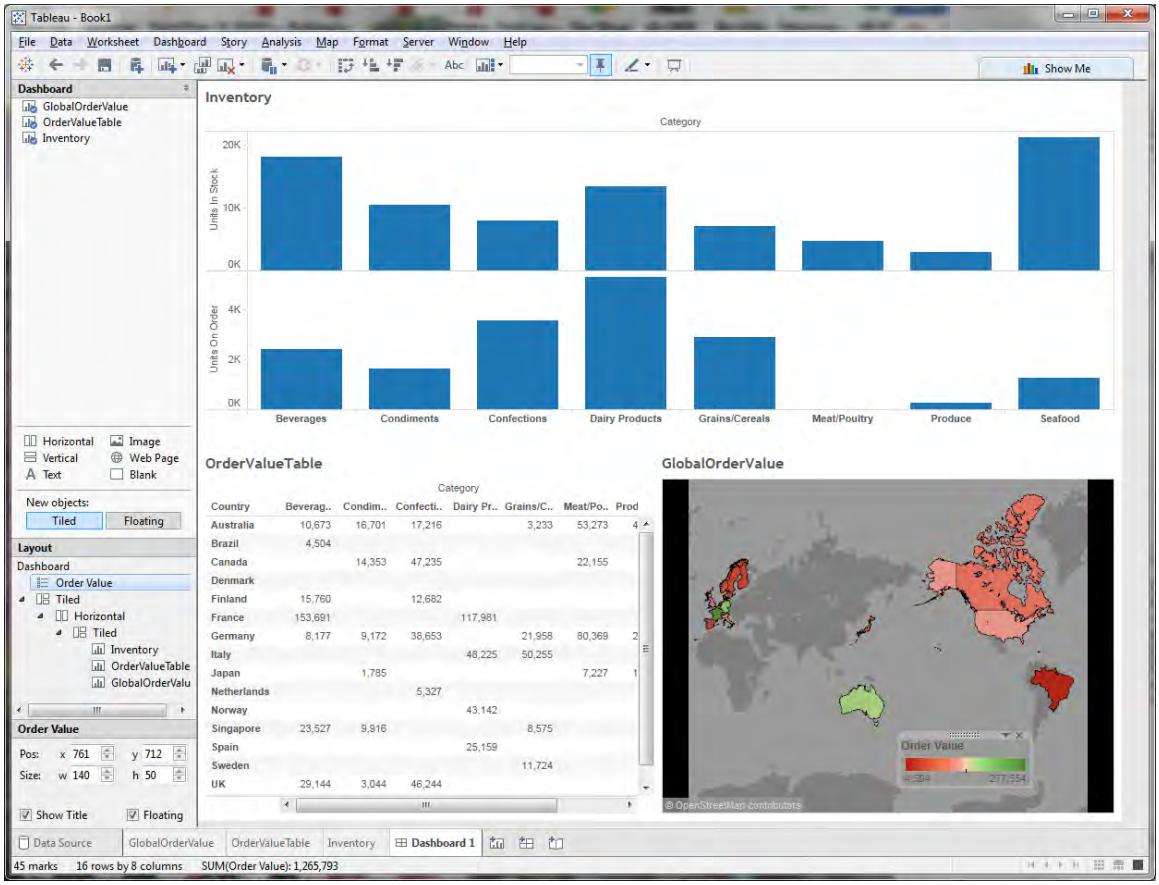


Tableau - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Show Me
Dashboard
GlobalOrderValue
OrderValueTable
Inventory
Horizontal
Image
Vertical
Web Page
Text
Blank
New objects:
Tiled
Floating
Layout
Dashboard
Order Value
Tiled
Horizontal
Tiled
Inventory
OrderValueTable
GlobalOrderValu
Order Value
Pos: x 761 y 712
Size: w 140 h 50
Show Title
Floating
Inventory
Category
Units In Stock
20K
10K
0K
Units On Order
4K
2K
0K
Beverages
Condiments
Confections
Dairy Products
Grains/Cereals
Meat/Poultry
Produce
Seafood
OrderValueTable
Category
Country Beverag.. Condim.. Confecti.. Dairy Pr.. Grains/C.. Meat/Po.. Prod
Australia 10,673 16,701 17,216 3,233 53,273
Brazil 4,504
Canada 14,353 47,235 22,155
Denmark
Finland 15,760 12,682
France 153,691 117,981
Germany 8,177 9,172 38,653 21,958 80,369
Italy 48,225 50,255
Japan 1,785 7,227
Netherlands 5,327
Norway 43,142
Singapore 23,527 9,916 8,575
Spain 25,159
Sweden 11,724
UK 29,144 3,044 46,244
GlobalOrderValue
Order Value
© OpenStreetMap contributors
Data Source
GlobalOrderValue
OrderValueTable
Inventory
Dashboard 1
45 marks 16 rows by 8 columns SUM(Order Value): 1,265,793

**Each week we will review the assigned articles in the live session. Questions are listed on the following pages.**

**Week 1**

**Article #1: Business Analytics Insight**

How are companies using analytics?

- Analytics are used to analyze current business processes and to design better ones, and analytics are integrated into work flows to monitor and increase their efficiency and effectiveness.
- Predictive analytics can be used to identify the attributes that are best for predicting job performance and thus score applicants for positions. With this approach, every application is considered, but only the "best" applicants are interviewed.
- If BI directors fail to plan for and integrate big data into their BI strategy, governance, architecture, technologies, processes, and activities, they risk facing a vacuum filled by the business units, resulting in a new generation of analytic silos

Which analytics techniques does your company use?

*Kathi Fox - Time Series, Regression, Risk adjusted O/E ratios are very important for my work*
*Michael - Predictive and Descriptive*
Eric - We use descriptive, predictive, and optimization techniques for everything from # of people to hire, to marketing, energy production, AI, and space travel.
Ryan - mainly descriptive, maybe more predictive for executive management (i hope)
Jaycee- Intel uses, predictive, descriptive, and Optimization analytics. From a finance perspective we use data analytics to forecast revenue, predict budgetary needs, set commission goals, and make investment trade off decisions.
John - primarily descriptive for internal and starting to develop predictive products to sell to customers
Kelly - we use predictive and descriptive- one visual we use is stoplights to tell us what area we are over/under- staffed in. We find the executive staff really likes this :)

Amber

- Optimization analytics
  - Mathematical programming (e.g., linear, integer), simulation
- Predictive analytics
  - Decision trees, CART, genetic algorithms, neural networks
- Descriptive analytics

- Data visualization
- Dashboards/scorecards
- Published reports

- Predictive analytics

Bria- we use a lot of predictive analytics within our department (sales) with some descriptive analytics as well

**Article #2: GE and the Culture of Analytics**

How is GE using analytics for process improvement?
They break up analytics into two categories; “big machine data” applications to analyze technical issues with customer assets and how to utilize the incoming data to improve commercial objectives to continuously improve performance for the long term.

What did they do to accelerate analysis and results?
GE borrowed the concept “agile methodology” which scopes work load and encourages “sprint planning”. They use data methodology like Lean to drive change in behavior and organize the team through daily calls, identifying the benefit up front, and asking the team “how they know they will be successful or unsuccessful. They utilize the Measurement and Control group to look at ROI and analyze impact quantitatively.

Why is culture important?
“Innovation is contingent on that culture” without a supportive environment, there would be a lack of striving for improvement and innovation. Employees could quantify success criteria and measure activities.

**Article #3: Location Analytics**

What are some applications of geographic data in businesses?
Political maps showing electoral votes, GPS applications using geospatial data, city identification geospatially using data for addresses, trees, landmarks, etc., product specific marketing campaigns dependent on geographical location

What are the advantages of consolidating individual accounts by location? What are the risks of consolidation?
Risks may include multiple business (strip malls) that may contaminate other nearby businesses, multiple device login congestion notification, fradualant activity, marketing analytics to better understand household profiles (netflix, utilities, amazon/shopping).

How is your company using geographic data?

*Kathi Fox - The VA medical system is broken in regional districts and sub-districts. Demographics are affected by geographical reach so occasionally, the regions are adjusted based on access for the veterans.*

Parsing of inbound customer calls to regionally specific call centers to provide better customer service

**Week 2**

**Article #1: Sustaining an Analytics Advantage**

What are some examples of creating competitive advantage with analytics (companies and their techniques)?

1. *Keep your analytics secret.: Walmart kept their supply chain analytics secret and it has maintained the competitive advantage.*
2. *Implement the analytics fast and defeat your competitors before they can react: analytics can be effective in a time of intense competition, when the analytics first mover advantage becomes magnified. ABB targeted transformer market. Others could not afford to keep up.*
3. *Apply your analytics to the right problems: Many groups prioritize the low hanging fruit where they can have easy wins and save money or drive incremental revenue, however it is easy to replicate and does not provide a sustained advantage. Instead prioritize the meaty projects to develop a competitive advantage. Procter and Gamble and the ICBC (Chinese Bank) and branch network design*
4. *Recognize that sometimes control of the data is more important than control of the analytics. American Airlines developed a crew algorithm to schedule crews. It then sold the algorithm to other airlines so that the other airlines would not build our their own data set and algorithm and find something that could potentially create a competitive advantage for them. Less here is to own and keep tight control of data driving the strategic analytics even though the analytics themselves can be sold to others.*
5. *Become a truly data driven corporation. Need tops down management support to have affective analytics teams. Best teams are rewarded and supported by senior leaders. Amazon is a clear example of this, analytics are mentioned in earnings report, they are run by an engineer, etc. IBM also mentioned*

**Article #2: Creating Business Values with Analytics**

What are the differences between competencies in information management and analytics expertise?

Information management and analytic expertise complement each other. Information management (data management) is based on structuring the data environment to keep tight control on the quality and timeliness of relevant data delivery. Information management specializes in collaborative organization since it is already focused on data integration. Analytic expertise is built off like the article says, off of “talent, tools and technology” and understanding the data and making quality decisions on the data. Analytic expertise supports specialized organization and focuses more on probability rather than delivery.

What are the advantages of focusing first on information management versus analytics expertise?

One major advantage is that information management already has the data infrastructure necessary for building an analytics platform and has already gone through the growth pains for developing it. Data structure needs to be in place to be able to access data in a meaningful manner. Focusing on analytics first is more cumbersome due to probability of data not being as organized as information management.

**Article #3: Raising the Bar with Analytics**

What new opportunities did StyleSeek and Entravision encounter when they used analytics?

**Clifford:**

Both StyleSeek and Entravision discovered opportunities outside of their initial business models to leverage their analytics platform to create other business opportunities. In the case of Entravision, the company created a separate entity to sell the insights it extrapolates from its customer's behavior to advertisers and marketing companies. Similarly, StyleSeek used its data insight to create a different business model that focuses on reselling the customer's intelligence it gains from the company's analytics platform.

They have an opportunity to enhance their business models viewing their analytics platforms as assets with strategic and commercial value that can help other businesses use analytics in more effective ways. In essence, this was an opportunity to develop a new revenue stream through the selling of analytics based on expertise in their specific industries.

The ability to increase sales at partner sites (by getting users "StyleDNA") and StyleSeek is working to become a better producer of data to increase data effectiveness for other businesses Entravision used data from its stores to set up a its own analytics division (Luminar) to sell the insights they'd been gathering to media spot buyers

What opportunity allowed MillerCoors to create efficiencies with analytics?

**Clifford:**

Analytics allowed MillerCoors to streamline its operation by using data to increase efficiencies and lower its operating cost. I can imagine that bringing two companies together would create a lot of redundancy and duplicity in the operation of the business. By collecting data and using analytics to extrapolate useful data to drive the decision on the optimal use of the processes, human resources, financial, production, logistics, and market segmentation, etc..

The merging of brewing companies to create MillerCoors highlighted the need for efficiency increases, waste reduction, and behavioral change at the senior leadership levels
Data analytics provided the path to this end through providing key components of information that allowed decision makers to better invest resources and fully leverage the data (this impacted decisions for the better and the company is more competitive as a result)

**Week 3**

**Article #1: Web Analytics: Enhancing Customer Relationship Management**

1. Describe the four main categories of metrics and relate to the Google analytics lessons

· Website usability - the tracking of how long a person stays on each page was demonstrated in Google Analytics and provides feedback on issues in a website and how engaging the content is on each page.

· Traffic sources - Google Analytics tracks the origination point for traffic and has a visual flow which shows how a person navigates through a website and at what point they exit

· Visitor profiles - It is important to understand the location, language, technology and keywords from users to tailor the website for these needs. GA allows for this type of tracking.

· Conversion statistics - Tracking visitors through IP, "cookies", or site login allows for the tracking of the status of visitors (new, returning or defectors) but also has privacy implications that are now regulated through GDPR (Global Data Protection Regulation). GA allows this type of tracking, but they have also been criticized for not being more open about data protection and privacy.

2. Describe the common techniques for Web analytics

Clustering/classification - Clustering items with similar characteristics. Enhances the ability to discover relationships that are otherwise not obvious.

Association rules - Utilized to predict the correlation of items. Affiliates one item with a set of items.

Path analysis- Generation of a graph that represents relations defined on web pages

Sequential patterns - discovering patterns that indicate user visit patterns over a certain period.

3. What are some business applications of web analytics?

· Science, health, marketing and finance all use Big Data and web analytics to derive meaning and business advantages (gaining more customers, learning about competitors, identifying critical industry trends)

· CRM (customer relationship management) and CEM (customer experience management) can be better managed using web analytics

· User behavior can be predicted and taken advantage of

· Ads and marketing campaigns can be more accurately and powerfully tailored

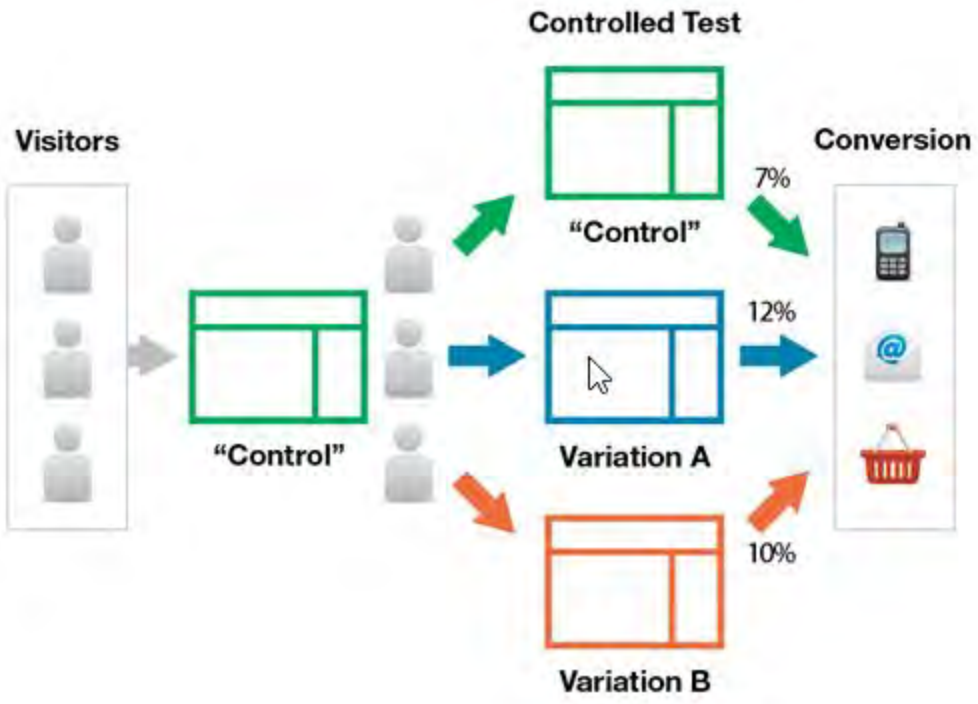
- Evaluate Promotional campaigns
- Provide users dynamic information based on their interests.
- Target electronic ads based on user patterns.
- Design more effective marketing strategies.

- Predict user behavior.
- Provide better customer care and customer intimacy.
- Maintain and manage customer base.

**Article #2: How eBay Uses Data and Analytics to Get Closer to Its (Massive) Customer Base**

1. What is an A/B test and what is its purpose?

AB **testing** is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.




2. Describe the three biggest challenges of web data

- Data at a large scale - managing scale - how can we ensure we can handle the amount of data coming in from a hardware and software aspect? The platform needs to handle the data in terms of storage but also in accessibility

- Collecting the right data - the data has to be relevant, Ensuring new data is adding value and not just creating noise - in addition to this notion, cutting the data the right way is also key
- New kinds of data - Being able to deal with/handle new types of data and understand what they mean such as audio input

3. How can Power Sellers use data better?
Power sellers are using data to run their own experiments on how their policy changes affect demand. Example what is the effect if I offer free shipping vs. if I charge $5 for shipping. Power sellers are testing consumer behavior and collecting the data to make business decisions.

4. Why are web analytics better than surveys? Web analytics run off of data rather than opinions. Ebay found the those who use surveys tend to be those that complain. The complainers were also found to complain a lot. Paired with the opposite approach to dissatisfaction, leave the site and never come back, quality data is hard to achieve or derive quality information from. Some behaviors that cannot be captured in surveys are better seen in data. It helps us understand the friction points when they use our system.

Overall, surveys don't get an accurate picture of the whole, rather, they get a lot of "squeaky wheels" that complain a lot and this can alter the actual truth.

- There is a discrepancy between what people say and do

**Week 4**

**Article #1: Minding the Analytics Gap**

1. What is a barrier to using analytics?

The number one barrier by far was translating analytics into business actions — in other words, making business decisions based on the results, not producing the results themselves. Another gap identified in the article is the gap in management's ability to understand the analytics being run and thus they have a hard time translating the analytics into business actions.

Translating analytics into business action or otherwise stated making business decisions based on the results. Middle management skills necessary to interpret the data in meaningful ways/understanding the data are lacking.

2. What can be done on the production and consumption side of analytics to overcome this barrier?
From the consumption side, a couple of key things can be done. First, data analysts can learn more about the business. Many organizations actually utilize time and resources to train their analysts about the business. Managers could also receive additional training in order for them to have a better understanding of the data and how to apply it.

From the production side, organizations can systematically improve infrastructure and processes; improved data quality, for example, can make it easier to turn data into competitive advantage.

a. Production Side: better industry specific training for the data analysists could lead to better displays of data that are easier for management to interpret. Organizations can also make it easier to collect quality data through process and infrastructure improvement.
b. Consumption Side: Management training so the data is easier to understand and interpret actionable items based on the analytical results. Also, “identifying trustworthy analytic professionals within the organization, requiring straightforward explanations and asking detailed questions” also assists with better managerial interpretation and decision making.

3. Describe the three levels of analytics maturity
Analytically Challenged – organizations generally rely more on management experience than data analytics and tend to lack data management and analytical skills
Analytical Practitioners – organizations tend to use analytics for operational purposes, have “just good enough data” and are working to become more data driven.

Analytical Innovators – organizations are more strategic in their application of analytics, place a high value on data, and have higher levels of data management and analytical skills.
1. Analytically Challenged: organizations rely more heavily on management experience than data analysis. Generally, data management and analytical skills are lacking.
2. Analytical Practitioners: Organizations utilize the analytics for operational decisions and continue to work to become more data driven.
3. Analytical Innovators: Organizations are generally more strategic in it's application of analytics, consider data highly valuable and employees/managers are skilled in higher levels of data management and data analytics.

**Article #2: Innovating with Analytics**

1. Describe the three characteristics of analytics innovators

Innovators tend to use more data- generally most to all of the data their organization generates or collects is used.

Innovators manage the information more effectively throughout the entire information transformation cycle (capturing data, analyzing information, aggregating and integrating data, using insights to guide future strategy and disseminating information and insights, through which, provides better decision making and a competitive advantage over others.

Innovators also understand the need and are capable of processing and analyzing data more quickly. Innovators are focused on utilizing speed for customer service/experience, pricing and innovative advancements.

**Week 5**

**Article #1: Innovating with Airborne Analytics**

What are the three areas of focus for their strategy for the next 3-4 years? Give examples of what they are doing

Operational efficiency/improved performance – they are fine tuning their engines to improve engine performance which will ultimately reduce fuel costs, which account for 40% of cost. They also use analytics to optimize crew and shift deployment that needs to take into account plane types, destinations and special breaks that are needed.

Customer intimacy – they are reviewing click-stream data to see customer stage, as well in trends in customer complaints

Innovation – understanding customer preferences and catering to those, such as understanding the types of things customers were downloading before getting on a flight allowed them to tailor and streamline their cache to allow customers to download *The Financial Times* and *The Wall Street Journal* onto their iPads in the lounge. Using RFID tags to track luggage and get to root cause analysis and customer satisfaction.

What are the challenges with hiring data scientists?
Data Scientists: specialized data scientists are not budget friendly; they're expensive. Hire as consultants as opposed to full time employment. They also need to be paired with an industry expert to assist with what if hypothesis and analysis/prediction models.

What are the risks trying to use technology in the organization? Give an example of each
Interoperability - data from source system can't be imported in analytics system, common understanding of how all the data comes together (Ford example from prof Harter)
Security - users who need access cannot get to it due to security concerns
Proper Taxonomies - inconsistent classification can cause conflict/information mismatch
Data Ownership - wrong group trying to change/edit when they don't own
Lack of Judgement - trusting any data, especially from the internet, w/o validating

**Article #2: A New, Analytics-Based Era of Banking Dawns at State Street**

What were the key questions that they asked which analytics could help?

a. How do I better manage, measure, visualize and report on risk?
b. How do I deal with all of this regulatory change?
c. How do I continue to find new sources of return?

d. What new types of investments, asset classes and geographies do I have to expand into to achieve better yields?

e. How do I keep my costs down, and how do I made sure I'm getting a fair shake in my trading activity?

What was the challenge of launching analytics at State Street?
-Getting manager buyoff/executive support to drive the analytics initiative by providing air cover while the program was working up to launch and continues to improve.
Massive culture shift, teaching an old dog new tricks.

Larger companies require many discussions prior to acting and tend to move slower than smaller companies. Moving fast enough to compete is the goal. Inspiring everyone to engage and break away from the cultural safe haven employees were used to by integrating individuals from different parts of the organization together.

How do they continue to push analytics forward?
Lead by example
Be dedicated to disruption
Being smart about how, when, and where to enact change - sometimes push, sometimes pull

**Week 6**

**Article #1: Modern Analytics and the Future of Quality Performance Excellence**

Define analytics (page 6):is the use of data, information technology, statistical analysis, quantitative methods and mathematical or computer based models to help managers gain improved insight about their business operations and make better, fact based decisions.

A process of transforming data into actions through analysis and insights in the context of organizational decision making and problem solving.
Analytics is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions

*"a process of transforming data into actions through analysis and insights in the context of organizational decision making and problem solving."* "the use of data, information technology and statistical analysis, quantitative methods, and math or computer based models to help managers gain improved insight about their business operations and make better fact based decisions."

How are companies using analytics in (page 7)?
· Banks – use analytics to predict and prevent credit fraud
· Manufacturing – use analytics
for production planning, purchasing and inventory management
· Retail – use analytics
to recommend products to customers and optimize marketing promotions
· Pharmaceuticals – use analytics
to get life-saving drugs to market more quickly
· Sports – use analytics to
determine both game strategy and optimal ticket prices

· Banks --- predict and prevent credit fraud
· Manufacturing—production planning, purchasing and inventory management
· Retail – recommend products and optimize marketing promotions
· Pharmaceuticals—drug research
· Sports—tickets sales and game strategy

Modern analytics integrates which three fields (page 8)?

- Business intelligence/information systems (BI/IS), statistics, and quantitative methods/operations research.

Descriptive Analytics – understand past and current performance and make informed decisions

Predictive Analytics – analyzing past performance in an effort to predict the future by examining historical data, detecting patterns or relationships in these data and then extrapolating these relationships forward in time.

Prescriptive Analytics – using optimization to identify the best alternatives to minimize or maximize some objective.

Business intelligence (BI) - collection, management, analysis and reporting of data
Statistics- summarizes data in a coherent manner and can find unknown or interesting relationships
Quantitative Methods/Operations - determine the relationship between one thing [an independent variable] and another [a dependent or outcome variable] within a population.

What are some examples of data sources (page 9)?
Click streams from websites, customer transactions, social media info

a. Numerical
b. Textual
c. Audio
d. Video

- Big data is captured using sensors such as scanners, click streams, transactions, emails, tweets, etc.
- . Sensor data (consumer and industrial)
-

Big data – unstructured and messy.
Other sources include numerical (customer transactions, clicks), textual (social media, tweets, emails), audio/video (smart tv activity, youtube)

What are examples of data visualization (page 11)?
Space

a. Time
b. Multivariate
c. Text
d. Graph
e. Network
f. Effective Data visualization:
    i. Both axes and units of measure are clearly labeled
    ii. Levels and trends are reported for a key performance measure- defects per million opportunities

iii. Results are presented for several years
iv. An arrow indicated that a downward trend is good for this measure
v. Appropriate comparisons are shown clearly
vi. Segmentation
vii. Improved performance.
viii. Known as dashboards/scorecards

**Article #2: A Process of Continuous Innovation: Centralizing Analytics at Caesars**

Why does Caesars use analytics (pages 1 & 2)?
Ultimate goal is to garner a better understanding of customer & operations. Gary Loveman (prior Chief Analytics Officer) changed the culture in Caesars to be forward thinking and adapt Analytics into a more centralized organization – reorganization turned the silos into a more cohesive all encompassing environment – increasing efficiency, communications and utilization of it's own resources.
To create a rich customer experience, marketing to customer based not only on their preferences, but their actions while at the Caesars property, including whether they've had a winning day or a losing one.

Caesars uses analytics:
- To guide business strategy
- Operations
- Innovation
- Creating a rich customer experience
- Gaming
- Revenue Management
- Finance
- Marketing
- Labor

What are four lessons learned from their experience (page 3)?
1-consolidating and centralizing different analytics departments (labor, marketing, casino, etc) required major reorganization which in turn required effective leadership for the new formed sub-organizations. Because of the previous isolations of the various sites and departments, consideration and selection had to be very deliberate.
2- Infrastructure needed to be severely overhauled. Degradation from acquisitions, communications and connections had to be addressed while at the same time, the ideology of 'one version of the truth' needed to be ingrained into the entire organization. So not only had physical and abstract obstacles but also emotional and cognizant challenges. Major challenge – developing a common language, buy in, costs and benefits to each department.

3-identifying and maintain communications with the major stakeholders through transparency and soliciting and acting on their feedback.
4- Visibility of the ‘big wins’ were communicated upfront and maintained high visibility particularly whenever a big win was approaching so the entire organization was aware of what this meant to themselves.

**Week 7**

**Article #1: Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients**

What are the six opportunities to reduce costs through analytics?
**1. High cost patients**
**2. Readmissions**
**3. Triage**
**4. Decompensation (when a patient's condition worsens)**
**5. Adverse events**
**6. Treatment optimization for diseases affecting multiple organs**

How can cost be reduced in each?

1. Use analytics to identify potentially high-cost patients to determine the patient's' specific needs and gaps in care. This is especially important for behavioral health issues (eg depression). Tailored interventions and risk analysis for follow-ups.
2. Use of algorithms to predict who is likely to be re-admitted
    a. Tailoring the interventions
    b. Ensuring patients receive the intended interventions
    c. Ensuring a low ratio of patients flagged to patients who experience a readmission
    d. Monitoring with IoT device (phone, monitoring device)
3. Estimating the risk of complications when a patient first presents to a hospital. This information can be used to manage staffing and bed resources, anticipating transfers and informing the overall strategy for managing the patient. Examples of studies - Apgar score for newborns and adult emergency room visits
4. Monitor patients in general care and use indices such as Rothman Index to prevent decompensation. IoT monitors can capture the breathing/heart/movement of bed-ridden patients and alert care-givers.
5. Early detection and on-going monitoring for Renal Failure, Infection and Adverse Drug Effects
6. Use analytics to project the trajectory of a patient's disease to target high-cost treatments to those who will benefit the most

**Article #2: A Review of Analytics and Clinical Informatics in Health Care**

What are some methods for improvement in health care using analytics? (page 2)

1. utilizing risk assessment analytics to process EHR data to identify the patients at greatest risk for utilizing more resources than their peers with the goal of improving patient outcomes and managing costs

2. (Ryan) value of aggregating data enhanced with real-time analytics to provide point-of-care information to oncologists that was tailored to individual patients

3. Unlabeled and free-text databases such as mammography data can be transformed into searchable and accessible collections that are usable for large scale health analytics

4. Analytics can supplement real-time analysis of physiological data streams in the neonatal intensive care unit for earlier detection of worsening medical conditions

5. Enhance less sophisticated, rule based systems already in use.  reduce errors and improve clinical outcomes in certain settings, such as pediatric intensive care units  and CDS can result in performance improvement on perioperative quality and process measures

6. Wearable monitors that patients use at home. Wearable health monitoring systems consist of a variety of sensors, actuators, and multimedia devices, and enable low-cost, noninvasive options for continuous monitoring of health, activity, mobility, and mental status, both indoors and outdoors reduce health-care costs by disease prevention and enhance the quality of life with disease management, and can be tailored to specific uses such as intelligent health monitoring of the elderly in nursing homes and for individuals with dementia or Parkinson's disease

7 (Eric).  cost savings based on improved patient care and outcomes, but also the identification of simple billing anomalies (i.e., revenue leakage). Most health care organizations find billing anomalies via a combination of rules-based approaches and manual audits; however, this approach is time-consuming and error-prone. Advanced analytics approaches (e.g., machine learning and predictive modeling) can be used to find patterns in billing records that are most likely associated with missing or erroneous charges.

What are some challenges for analytics in health care? (page 4)

Data is both quantitative and qualitative making it more difficult to pull and assimilate in data files to analyze. This can also make it difficult to produce accurate results

**Week 8**

**Article #1: What Businesses Can Learn from Sports Analytics?**

Describe the five key lessons of analytics in sports (give an example of each)

1. Align Leadership at multiple levels: Need to align decision making on which players to acquire, what strategy to adopt, and how much to pay players.
2. Focus on the human dimension: Teams assess individual performance through typical stat keeping, they also assess performance in context (what combinations of players are on the court at the same time).
3. Exploit Video and Location Data: track players movements on the field to predict how they will play in different arenas ( NY Yankees Baseball). Other companies are using this to optimize routes for delivery and reduce customer wait times.
4. Work with a broader ecosystem: draw as much from the partner while maintaining key internal capabilities. Orlando Magic established a close relationship with a software vendor and Disney, but maintain expertise of their business model and operations. Does allow other teams to catch up and can hinder a long lasting competitive advantage.
5. Support Analytic Amateurs: players are now studying their specific data. Example- Analyzing Fielding Independent Pitching metric to understand that you throw too many foul balls due to your pitching style.
    a. Performance metrics, sales CRM

**Article #2: Team GB: Using Analytics (and Intuition) to Improve Performance**

What is the value of predicting team performance? (page 2)
Predicting team performance helps to determine how a team stacks up against its competition, and also helps to determine which sports to investment in, which in the U.K. is based on medal potential at the Olympic Games.

What is the biggest challenge? (page 2)
The biggest challenge is difficulty in collecting the data. Analytics are very sport specific, and some sports don't produce data that is as quantifiable and/or data collection methods don't yet exist for them. Or, sensors do exist but are simply too expensive.

What are some of the barriers? (page 3)

Resistance to change by coaches who have used intuition and experience in the past plus they may not understand the technology (lack of computer skills among coaches)

Intellectual Property – if I write programs, I don't want them used by everyone

Where is the power of the data? (page 5)

- more data about themselves, can help prevent injury and train smarter instead of harder
- glean more information about the competition, more likely to beat them
- With these two pieces of information, models can be created to to predict performance

The power resides in good longitudinal information rather than in snapshots.

**Week 9**

**Article #1: An introduction to data mining and other techniques for advanced analytics**

What are the key differences between statistical analysis and data mining? (page 140)

- Statistical analysis deals with small samples, while DM deals with entire populations - requires different approach, else you'll have misleading results
- DM does not provide confidence intervals
- Data mining is more detailed by including diagnostic results to indicate likely business benefits from the model

Describe tools for advanced analytics (page 149-151)

· Data visualization a. makes it easier to interpret large amounts of data and transform it into multidimensional pictures and animations while highlighting areas of interest. This can be completed prior to engaging in DM or after the model has been created (black box- neural networks/genetic algorithms). Most popular techniques include scatter plots, heat maps and geographical maps in addition to animations.

· Text mining: a. converting valuable information from unstructured format within text files so it becomes a structured variable to be included in DM. It uses linguistic analysis to extract facts from unstructured text (identify certain words with high association of fraudulent claims).

· Social network analysis : a. Identifies groups connected together in some way by applying network theory concepts such as nodes and links (nodes- individuals within the networks; links- relationships between the individuals). SNA metrics are correlated with customer loyalty. SNA extracts useful new variables such as size, strength and composition of each customers calling circle for use in DM projects (churn prediction)

· Contact optimization: a. Optimizes the communications budget by identifying which channels to use to connect with customers. It takes all the analytical model predictions as inputs and searches for the optimal allocation of products and channels to customers over time. The allocation itself must meet budget constraints, contact rules and min/max volume limitations. Inbound vs. outbound:

- Inbound- concerned with delivering the best solution for each individual customer who contacts the company's call center or logs onto their website. Extension of CRM, supply the next best offer for each customer based on a set of predicted propensities for the available products.
- Outbound- aims to find the best solution at an individual level and at the same time meet overall outbound marketing business targets and constraints.

How do you mitigate the risks of data mining? (page 152)

- Data quality - spend time before modeling and deployment to insure the data is accurate and structured appropriately
- Train workers so they don't make misleading or nonsensical results from highly automated modeling tools
- Workers need the experience to identify meaningful patterns in the data
- Training to identify model efficiency, understand descriptive statistics and not modeling outside the domain of the data

**Other articles of interest**

Deep Neural Networks:

DeepMind AlphaGo defeats world champion in game of Go (March 2016; Jan 2017 update):
https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/
http://fortune.com/2017/01/07/google-alphago-ai/

Google Brain's neural network develop AI encryption (November 2016):
https://www.scmagazine.com/google-brains-neural-networks-develops-ai-encryption/article/570049/
http://www.wired.co.uk/article/google-artificial-intelligence-encryption

DeepStack Defeats 10 out of 11 poker champions (March 2017), develops intuition:
http://www.cnn.com/2017/03/02/health/artificial-intelligence-poker-intuition-study/index.html
https://www.scientificamerican.com/article/time-to-fold-humans-poker-playing-ai-beats-pros-at-texas-hold-rsquo-em/

Not Neural Networks:

DeepBlue defeats Gary Kasporov in chess (1996):
https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)

DeepThought defeats chess grand master Brent Larsen (1988), but loses to Gary Kasparov in chess (1989):
https://en.wikipedia.org/wiki/Deep_Thought_(chess_computer)

**Week 10**

**Article #1: Business Analytics: Transforming the Role of Management Accountants**

What are some external and internal data sources for accountants? (page 3)

Internal: Files (Spreadsheets, CSVs, MS Access) and Databases (SQL, ERP, Data Warehouses)
External: Google Analytics, SEC XBRL databases, Salesforce, ZenDesk
Other: Email, Sensors, Videos, Tweets….

What are four challenges for accountants using analytics? (page 4)
Awareness – smaller companies are often constrained because their resources are much more limited. Access is limited by cost constraints, inadequate IT structure and technical knowledge. As Helen Brand, CEO of ACCA (Association of Chartered Certified Accountants), said, "The key issue now is how the finance function adapts to the fast-evolving technology to better serve businesses in the future."3
Interoperability – internet data is unstructured so there is a challenge to mesh existing structured data with internet data. World Wide Web Consortium (W3C) is working on developing standardizations of web data sources into usable formats.
Security – data security becomes even more of an issue as web data is pulled in. Integrity of data is paramount since breaches can have traumatic repercussions for the company, so the accountants also need to monitor and protect the data and ensure privacy controls are practiced.
Analysis Quality – companies need to focus on structured design processes for filtering in data. (Garbage in – garbage out) becomes amplified if the wrong data is focused on.

What are five areas for leveraging analytics in accounting? (page 4)

1. franchise sales analysis- analyze franchisor sales metrics to determine point of sale promotions
2. Accounts receivable and credit analysis- days sales out tracking, accounts receivable collection history. Companies can develop credit and collection policies that minimize bad debt risk.
3. accounts payable analysis and payment monitoring- payments are constantly monitored and evaluated against payment decision rules that flag fraudulent transactions. Can find fraudulent payments more efficiently and reduce the risk of overpayment.
4. mergers and acquisitions (M&A) due diligence - accurate business valuations
5. forensic accounting- fraud detection analytics to identify fraud, bribery, and corruption in companies. Textual analysis on payment description fields

**Article #2: Elevating Data, Analytics to the C-Suite**

What are the steps to elevate a department using analytics? (page 5)

Data first (not necessarily perfect data)
Then reporting
Then analytics
Then quantitative and predictive modeling
People have to trust your information and analysis

How should you address non-perfect data? (page 5)

Data cannot be perfect only reasonable. The time and effort to obtain perfect data is not worth the investment of resources and time. By the time data is perfected to 100%, there is the potential it may be obsolete also. Data that is 80% good provides adequate support for making decisions. It's very important to communicate the deficits in the data to help make more informed decisions.

Acknowledge that data has to be good and clean as well knowing that data quality is a journey. Understand that in the absence of data, a decision will be made anyway. Also acknowledge that there may be blind areas that could impact, but may not necessarily stop a decision if acknowledged and managed

Should analytics teams be centralized or decentralized? (page 6)

- analytics teams need to be connected to their lines of business, and should support and the decision makers. When organizations are large, it's very difficult not wise — to centralize as there is too much distance between the organization and the decision makers and the people developing the information for them.
- For smaller organizations, it's easier to centralize; however, it is a difficult and constant balancing act for an organization. Centralizing has all the benefits of better training, skill growth, and career paths, but has the down side of losing connectivity to the business