IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

January 19,21 2016

Week 2 Presentation

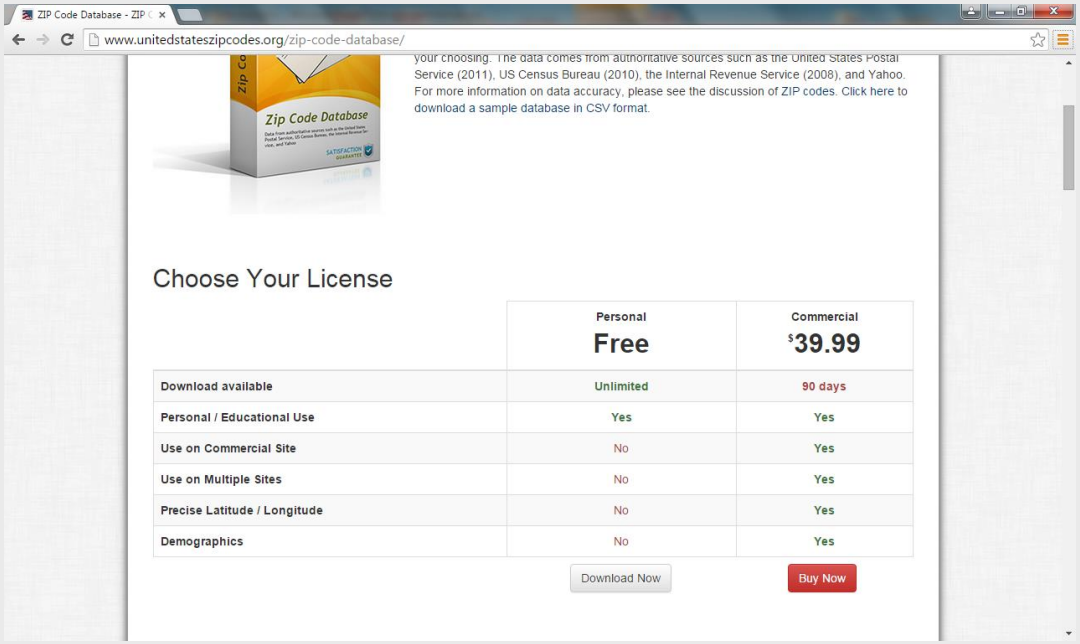
Week 2 Topic: Expectation & Agenda

- Read Chapter 1 of Data Smart
- Reviewed course materials from last week
- Reviewed Excel Basics and Installed Solver Add-in
- Review assignment on zip code
- Optimization Modeling Overview and Readings

Week 2 Topic:
Zip Code: Analysis using Pivot Table

We will demonstrate using Pivot Tables by downloading a zip code data set from:
http://www.unitedstateszipcodes.org/zip-code-database/

This will be useful in the future as a "look up" to do further analysis. To begin, download the *free* version into Excel:

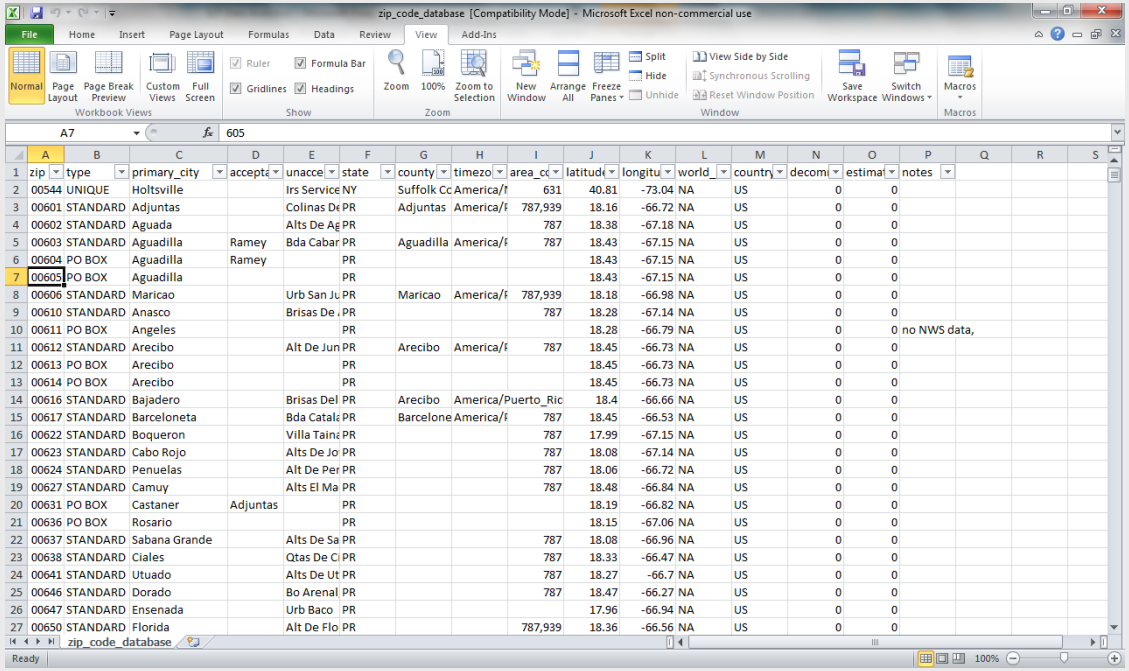


Week 2 Topic:
Zip Code: Raw Data

Investigate the data:

- What are the different types of zip codes e.g., Unique, Standard?
- What data should I keep versus not?
- Do I care for Military zip codes?

To answer some of these questions and to analyze the data further, we create a Pivot Table in the next slide.



zip	type	primary_city	accepta	unacce	state	county	timezo	area_co	latitud	longitu	world_	country	decomi	estimat	notes
00544	UNIQUE	Holtsville		Irs Service	NY	Suffolk Co	America/I	631	40.81	-73.04	NA	US	0	0	
00601	STANDARD	Adjuntas		Colinas De	PR	Adjuntas	America/I	787,939	18.16	-66.72	NA	US	0	0	
00602	STANDARD	Aguada		Alts De Ag	PR			787	18.38	-67.18	NA	US	0	0	
00603	STANDARD	Aguadilla	Ramey	Bda Cabar	PR	Aguadilla	America/I	787	18.43	-67.15	NA	US	0	0	
00604	PO BOX	Aguadilla	Ramey		PR				18.43	-67.15	NA	US	0	0	
00605	PO BOX	Aguadilla			PR				18.43	-67.15	NA	US	0	0	
00606	STANDARD	Maricao		Urb San Ju	PR	Maricao	America/I	787,939	18.18	-66.98	NA	US	0	0	
00610	STANDARD	Anasco		Brisas De	PR			787	18.28	-67.14	NA	US	0	0	
00611	PO BOX	Angeles			PR				18.28	-66.79	NA	US	0	0	no NWS data,
00612	STANDARD	Arecibo		Alt De Jun	PR	Arecibo	America/I	787	18.45	-66.73	NA	US	0	0	
00613	PO BOX	Arecibo			PR				18.45	-66.73	NA	US	0	0	
00614	PO BOX	Arecibo			PR				18.45	-66.73	NA	US	0	0	
00616	STANDARD	Bajadero		Brisas Del	PR	Arecibo	America/Puerto_Ric		18.4	-66.66	NA	US	0	0	
00617	STANDARD	Barceloneta		Bda Catala	PR	Barcelone	America/I	787	18.45	-66.53	NA	US	0	0	
00622	STANDARD	Boqueron		Villa Taina	PR			787	17.99	-67.15	NA	US	0	0	
00623	STANDARD	Cabo Rojo		Alts De Jo	PR			787	18.08	-67.14	NA	US	0	0	
00624	STANDARD	Penuelas		Alt De Per	PR			787	18.06	-66.72	NA	US	0	0	
00627	STANDARD	Camuy		Alts El Ma	PR			787	18.48	-66.84	NA	US	0	0	
00631	PO BOX	Castaner	Adjuntas		PR				18.19	-66.82	NA	US	0	0	
00636	PO BOX	Rosario			PR				18.15	-67.06	NA	US	0	0	
00637	STANDARD	Sabana Grande		Alts De Sa	PR			787	18.08	-66.96	NA	US	0	0	
00638	STANDARD	Ciales		Qtas De Ci	PR			787	18.33	-66.47	NA	US	0	0	
00641	STANDARD	Utuado		Alts De Ut	PR			787	18.27	-66.7	NA	US	0	0	
00646	STANDARD	Dorado		Bo Arenal	PR			787	18.47	-66.27	NA	US	0	0	
00647	STANDARD	Ensenada		Urb Baco	PR				17.96	-66.94	NA	US	0	0	
00650	STANDARD	Florida		Alt De Flo	PR			787,939	18.36	-66.56	NA	US	0	0	

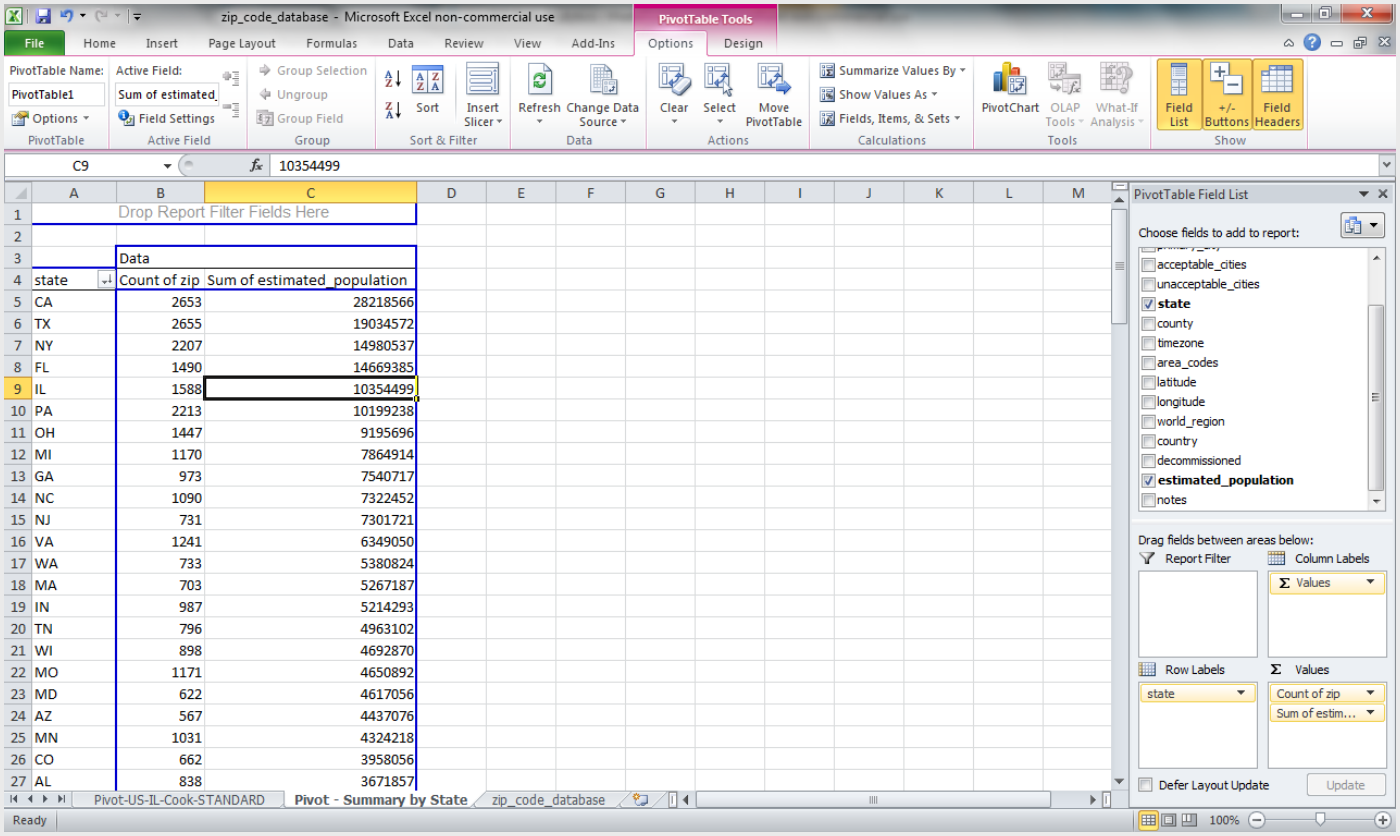
Week 2 Topic:
Zip Code: Fun Facts

Going back to the website with some questions from looking at the data, we learn some interesting facts on zip codes:

- The term ZIP stands for Zone Improvement Plan. The basic 5-digit format was first introduced in 1963 and later extended to add an additional 4 digits after a dash to form a ZIP+4 code. The additional 4 digits help USPS more precisely group mail for delivery. Though ZIP codes were originally developed for USPS, many other shipping companies such as United Parcel Service (UPS), Federal Express (FedEx), DHL, and others make use of ZIP codes for sorting packages and calculating the time and cost of shipping a package (the shipping rate).
- Types of Zip Codes:
  - Unique/single high volume address (ex. 20505 for the CIA in Washington, DC)
  - PO Box only (ex. 22313 for the PO Boxes of Alexandria, VA)
  - Military
  - Standard (all other ZIP codes)
- There are some interesting reads on geographic boundary issues in assigning ZIP codes. ZIP codes were made to make mail delivery easier. They weren't made to correspond to existing boundaries such as cities, counties, or even states. If it is more efficient for a mail carrier to drive across a state line to deliver mail, the ZIP code "boundary" will cross the state lines. ZIP codes don't usually cross state lines, but some do (65733 is a good example).
- Knowing what we know about ZIP codes, we may need to pay special attention on using ZIP codes for as criteria for cluster analysis, as an example. It may not always serve up customers that the company wants to target or may include states that the company is not licensed to operate.

Week 2 Topic:
Zip Code: Summary by State

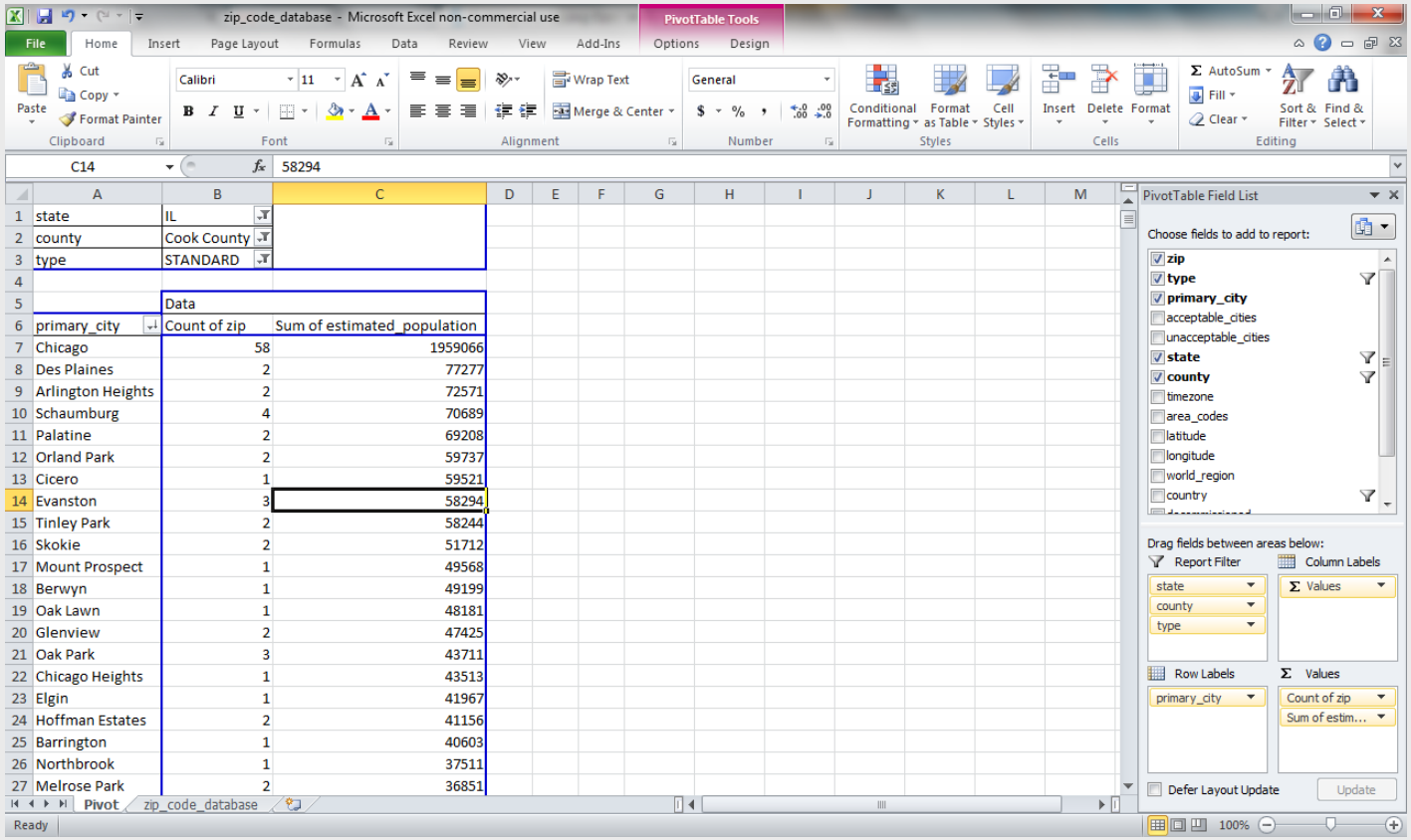
Count of zip (all categories) and estimated population by state:



state	Count of zip	Sum of estimated_population
CA	2653	28218566
TX	2655	19034572
NY	2207	14980537
FL	1490	14669385
IL	1588	10354499
PA	2213	10199238
OH	1447	9195696
MI	1170	7864914
GA	973	7540717
NC	1090	7322452
NJ	731	7301721
VA	1241	6349050
WA	733	5380824
MA	703	5267187
IN	987	5214293
TN	796	4963102
WI	898	4692870
MO	1171	4650892
MD	622	4617056
AZ	567	4437076
MN	1031	4324218
CO	662	3958056
AL	838	3671857

Week 2 Topic:
Zip Code: US-IL-Cook View

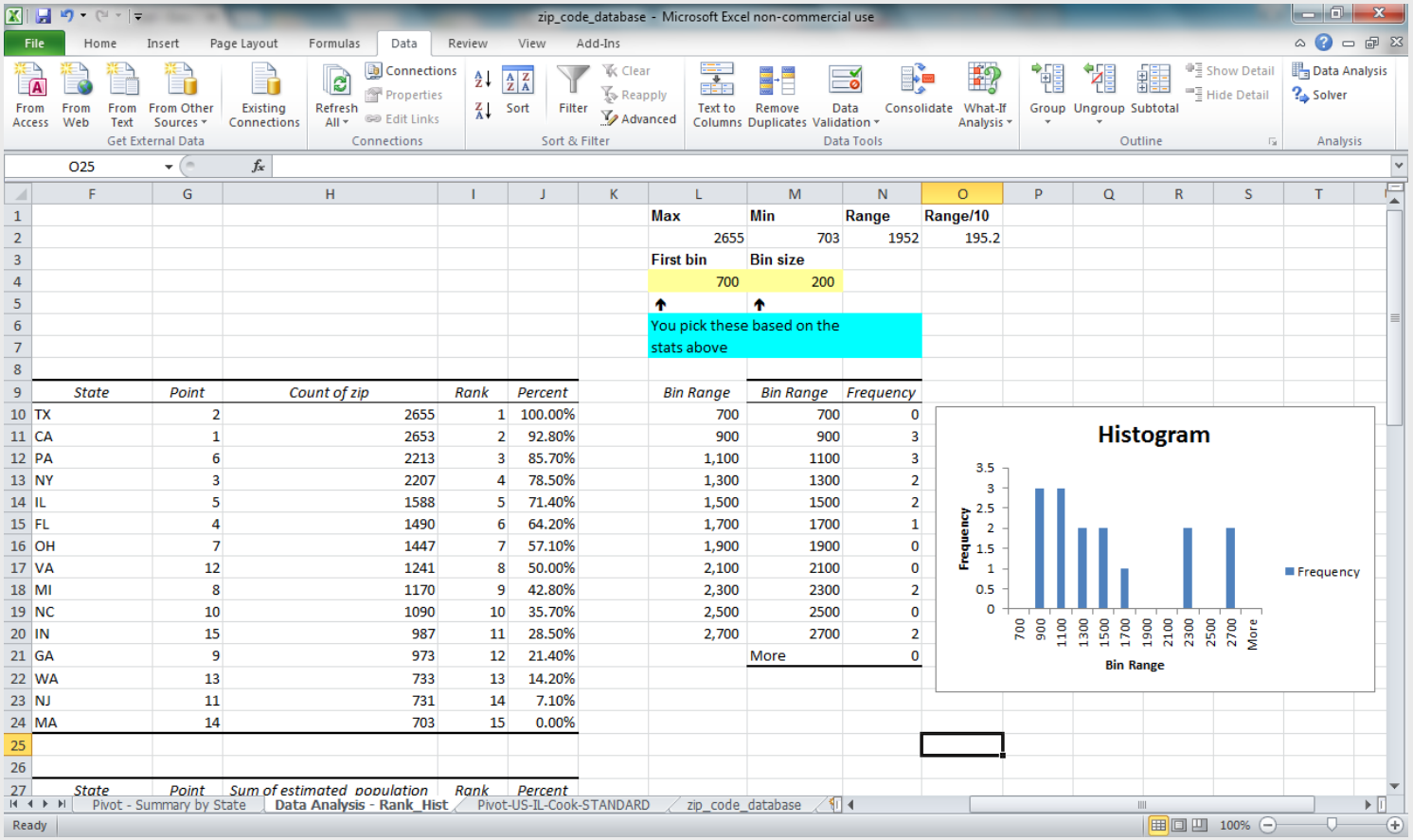
Filtering for US-IL-Cook County then sort high to low on count of zip code:



	A	B	C
1	state	IL	
2	county	Cook County	
3	type	STANDARD	
4			
5		Data	
6	primary_city	Count of zip	Sum of estimated_population
7	Chicago	58	1959066
8	Des Plaines	2	77277
9	Arlington Heights	2	72571
10	Schaumburg	4	70689
11	Palatine	2	69208
12	Orland Park	2	59737
13	Cicero	1	59521
14	Evanston	3	58294
15	Tinley Park	2	58244
16	Skokie	2	51712
17	Mount Prospect	1	49568
18	Berwyn	1	49199
19	Oak Lawn	1	48181
20	Glenview	2	47425
21	Oak Park	3	43711
22	Chicago Heights	1	43513
23	Elgin	1	41967
24	Hoffman Estates	2	41156
25	Barrington	1	40603
26	Northbrook	1	37511
27	Melrose Park	2	36851

Week 2 Topic: Zip Code: Data Analysis

We demonstrate using performing data analysis in Excel:



Week 2 Topic: Analysis Objectives

- **Actionable** – analyst does the analysis, not the reader. Information, not data, needs to be presented in a way that is easy to understand. Decisions should be made from presentation of data.
  - Listed data – is a presentation of just data
  - Sorted data – is information that can be acted upon

- **Anticipate** – what the reader will ask, think, and need. Don't just stop with what you have, think of what might be asked and do the next ten steps of analysis.
  - Derive data if further analysis needs it
  - Get additional data if the next question to be asked needs it
  - Be creative

- **Purposeful** – any graph, chart, dataset that you present needs to have purpose. Don't waste the reader's time if there is no decision to be made from the analysis. Ask the "so what" in the analysis.

Week 2 Topic: Analysis Objectives (cont.)

- **Simple** – clean and thoughtful presentation of information is always simple. Analysis can be complicated but this does not mean the presentation has to be. Spend the time to simplify the presentation. Don't mask the decisions in complexity
  - One chart, graph, pivot, or table per screen view
  - Organize the information so that there is a *flow*

- **Annotate** – always explain what the reader is reading. The reader shouldn't have to guess what's going on.
  - Add an intro/summary sentence to a chart, graph, or table
  - Titles, labels, legends, etc. all matter
  - Freeze columns and rows as needed to make sure the reader knows when scrolling

Week 2 Topic:
Zip Code analysis – summary section

What does the reader know about the postal code system in the country that you chose?

- Assume nothing
- Provide just enough information in the beginning so that the reader understands the purpose of the analysis
- Include metadata information so that the reader understands the data content before stepping into the analysis
- Further details on the data and analysis can be referenced in an appendix or back of reports

Week 2 Topic:
Zip Code analysis – data section

A reader should have been provided the following information:

- What's the hierarchy?
- What information do I need to know versus not for the analysis?
- What columns of information was included in the analysis?
- It would have been nice to see some graphical representation of the codes as it is based on some kind of geographic breakout.
- Was the data complete?
- Were there any data issues?
- Include metrics on data files e.g., file row counts, header information, version, date published, etc.

Week 2 Topic:
Zip Code analysis – analysis section

- Statics:
  - How many postal codes are there for the country? Does this match the row count for the file? If not, why?
  - State why you chose to look at a certain level breakout and counts e.g., by geographic zone, state, etc.
  - If you chose to focus on a specific state or area, state why you did that e.g., your home state, state with most postal codes, etc.
- Presentation:
  - Always SORT data
  - Do not include grand totals in charts & graphs
  - Default pivot calculation is SUM, don't leave it be unless it makes sense e.g., sum of zip code means nothing

Week 2 Topic: Presentation - *best practices*

- Any inserts from Excel should be copy/pasted as a picture. There should be just one source of data. If that source of data is Excel, then keep it there. Imbedding objects can get tricky.
- If you insert charts or graphs from Excel, just insert the chart and/or graph. Exclude the menu bar and other surroundings.
- If you insert data from Excel, make sure to insert just the relevant cells and data, without background gridlines.
- With accompanying summary description – describe what the reader needs to know about the content that was presented, don't just copy/paste without a statement
- With formatting – uniform gridlines, wraps in cells, size the cells appropriately, size the chart/graph
- With readable sizing – no magnifying glasses but not 80 years old
- Without background gridlines – just turn it off for printing altogether
- Without dropdowns and filters that are Excel specific – if you include everything in Excel, might as well just share the Excel file

Week 2 Topic:
Data in Excel - *best practices*

- Summary to the front – start with a "main or summary" sheet that explains content of the workbook e.g., table of content, versions, dates, etc. Also, consider including:
  - Summary of objectives or an introduction to the workbook
  - Key metadata/information needed to understand the analysis/data
  - Summary of findings
  - Updates to the workbook (if versioned)
- Data to the back – raw data sets, look ups, lists, etc. are references that should be in the back for referencing.
- Links to navigate – don't make the reader spend half their time searching for references. Create links to datasets, if in another worksheet.
- Present the information first with a sentence or two before presenting the data – a reader should know what he/she is looking at before wondering what he/she is looking at
- Always sort the information. It's about pattern. If you present chaos without pattern, it's no information at all.
- Highlight derivation of data – always account for where the data came from and why you are using it for the analysis.

Week 2 Topic:
Class Exercises – due 1/28th

Create pivot table using US zip code data:

a) What characteristics are there e.g., exceptions, boundary definitions, history
b) How many postal designations are there for the country?
c) What geographic area has the most postal codes?
e) Submit both Excel Workbook and present results in a presentation (similar to in-class example on zip code)

Week 2 Topic: Overview of Optimization Modeling

Optimization is the process of finding the best values of the variables for a particular criterion or the best decisions for a particular measure of performance.

- Components of an Optimization Model:
  - Objective – mathematical function to minimize or maximize some measure of performance
  - Parameters – Numerical inputs for calculations that may correspond to raw data, estimates, forecasts, or predictions
  - Constraints – logical conditions or calculations representing real world limits
  - Decision Variables – An unknown quantity or variables to be determined via model run
- References for Optimization Modeling:
  - Data Smart
  - www.solver.com
  - http://opensolver.org
  - Optimization Modeling with Spreadsheets (What's in the PDF handed out. Additional content will be discussed/distributed as needed)
  - Step-By-Step_Optimization_S.pdf (in Week 3 Readings)

Week 2 Topic:
Basic Structure of Optimization Modeling

- The basic structure of a typical mathematical optimization problem formulation is shown here:

  min|max objective function

  subject to constraints
  variable bounds

- The formulation is easier to understand if it is followed by a description of the decision variables, sets, and parameters.

Week 2 Topic: Linear Programming Problem

- Each linear constraint can be either an inequality or an equation
- Bounds can be ±∞, so xj can be restricted to be nonnegative (lj = 0 and uj = +∞) or free (lj = -∞ and uj = +∞)
- Exhibits proportionality (contribution from any given decision variable to the objective grows in proportion to its value), additivity (contribution from one decision is added to contributions of other decisions), and divisibility (fractional decision variable is meaningful).

$$\begin{aligned} &\min|\max && f_1x_1 + \ldots + f_nx_n \\ &\text{subject to} && \mathbf{Ax}\ \{\leq, =, \geq\}\ \mathbf{b} \\ & && l_j \leq x_j \leq u_j \quad (j = 1,2,\ldots,n) \end{aligned}$$

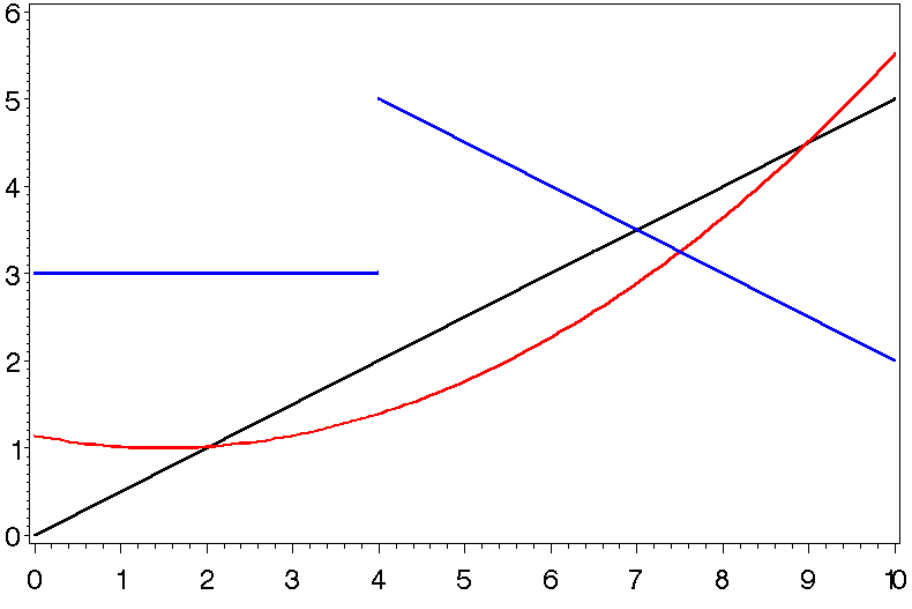
Week 2 Topic: Nonlinear Programming Problem

- $f(\mathbf{x})$ and $c_i(\mathbf{x})$ are continuous functions
- With the following constraints:
  - **unconstrained**: There are no constraints or bounds.
  - **bound constrained**: There are no constraints other than bounds.
  - **linearly constrained**: All functions $c_i(\mathbf{x})$ are linear.
  - **nonlinearly constrained**: At least one of the functions $c_i(\mathbf{x})$ is nonlinear.

$$\begin{aligned} &\min|\max \quad f(\mathbf{x}) \\ &\text{subject to} \quad c_i(\mathbf{x}) \ \{\leq,=,\geq\} \ b_i \quad (i=1,2,\ldots,m) \\ &\qquad\qquad l_j \leq x_j \leq u_j \quad (j=1,2,\ldots,n) \end{aligned}$$

Week 2 Topic:
What's expected using Optimization

- Able to set up Optimization problems using add in function of Excel
- Understand the three modeling available:
  - GRG Nonlinear: For nonlinear, smooth (red)
  - Simplex LP: For linear, smooth (black)
  - Evolutionary: For nonlinear, non-smooth (blue)

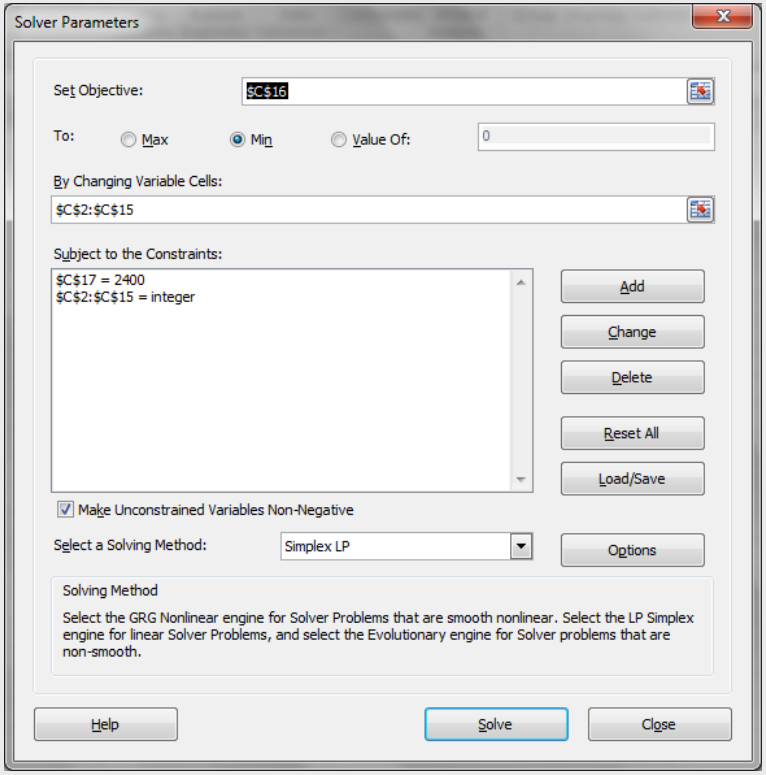


Week 2 Topic: Usage in business

- **Production Planning:** Determine which of several possible mixes of products should be produced to achieve the highest profit.
- **Facility Location:** Find the "best" site for, say, a new factory, in relation to the location of materials suppliers, distribution centers, and so on.
- **Portfolio Selection:** Maximize ROI, balancing return versus risk.
- **Personnel Assignment:** Match personnel to work requirements in order to meet current needs and anticipated changes, subject to budget and HR requirements.
- **Supply Chain Planning:** Find the lowest-cost way to move product from factories to distribution centers to stores, and plan for possible disruptions or expansion.
- **Promotional Marketing:** Determine the best combination of promotional offers, delivery channels, and customers to maximize the overall return on marketing investment.
- **Supplier Selection and Evaluation:** Choose which suppliers to deal with in order to satisfy requirements and maximize leverage, rating suppliers using a variety of criteria simultaneously.
- **Inventory Replenishment:** Set inventory policies (reorder levels and maximum stock levels) to meet customer service goals and minimize costs.
- **Pricing Decisions:** Establish and maintain optimal everyday prices based on costs, regional demand patterns, and competitive price information.

Week 2 Topic:
Back to concession stand example

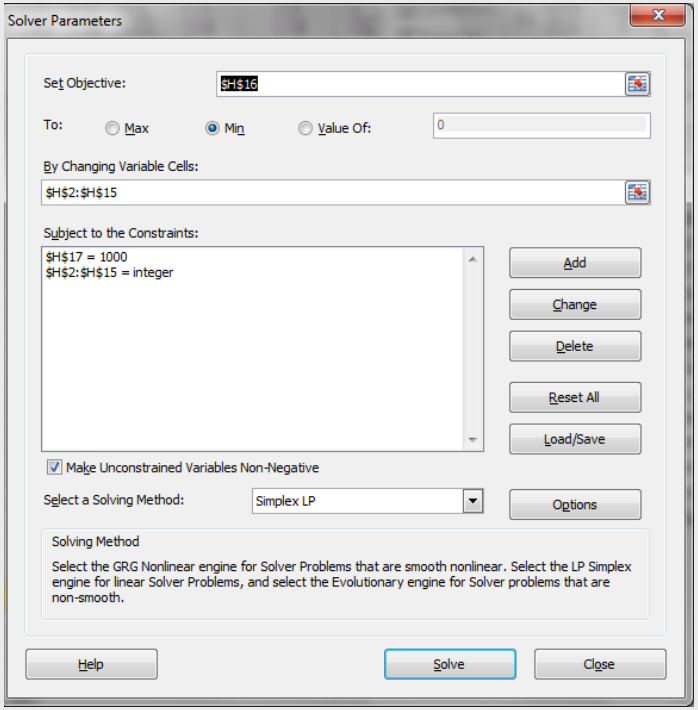
Using Simplex LP: The model run reflects the book:



Item	Calories	How many?
Beer	200	0
Bottled Water	0	0
Chocolate Bar	255	0
Chocolate Dipped Cone	300	0
Gummy Bears	300	0
Hamburger	320	0
Hot Dog	265	0
Ice Cream Sandwich	240	0
Licorice Rope	280	1
Nachos	560	2
Pizza	480	0
Popcorn	500	2
Popsicle	150	0
Soda	120	0
	Total Items:	5
	Total Calories:	2400

Week 3 Topic: Back to concession stand example

Using Simplex LP: The model says to just sell popcorn.

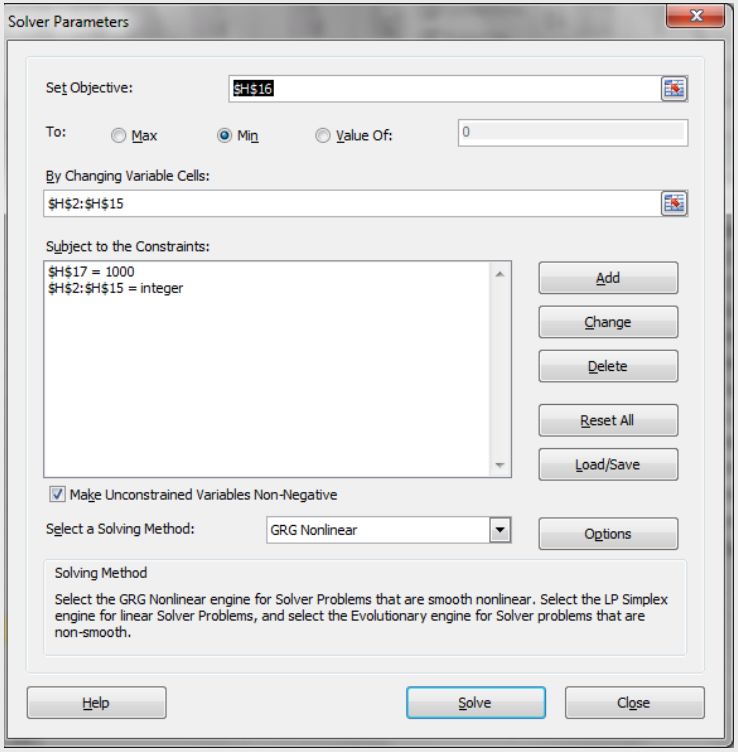


Item	Unit Price	How many?
Popcorn	5	200
Beer	4	0
Bottled Water	3	0
Chocolate Dipped Cone	3	0
Hamburger	3	0
Ice Cream Sandwich	3	0
Nachos	3	0
Popsicle	3	0
Soda	2.5	0
Chocolate Bar	2	0
Gummy Bears	2	0
Licorice Rope	2	0
Pizza	2	0
Hot Dog	1.5	0
	Total # of Items:	200
	Revenue:	1000

ITM - 527

Week 2 Topic:
Back to concession stand example (cont.)

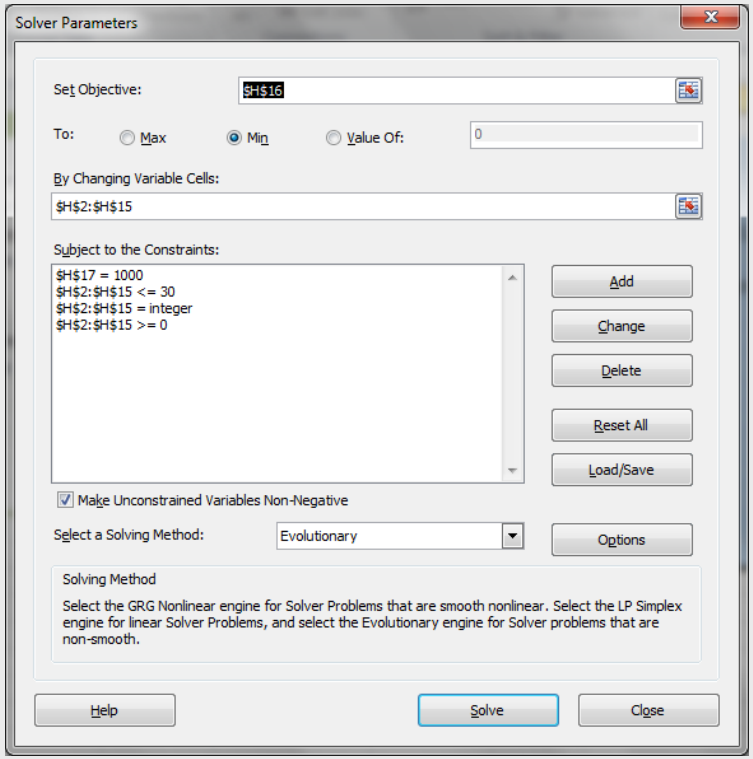
Using GRG Nonlinear: The model says to just sell popcorn. Linear is a subset of nonlinear models. Hence you can solve linear problems using nonlinear models. However, we use linear models on linear problems as it's more dependable.



Item	Unit Price	How many?
Popcorn	5	200
Beer	4	0
Bottled Water	3	0
Chocolate Dipped Cone	3	0
Hamburger	3	0
Ice Cream Sandwich	3	0
Nachos	3	0
Popsicle	3	0
Soda	2.5	0
Chocolate Bar	2	0
Gummy Bears	2	0
Licorice Rope	2	0
Pizza	2	0
Hot Dog	1.5	0
	Total # of Items:	200
	Revenue:	1000

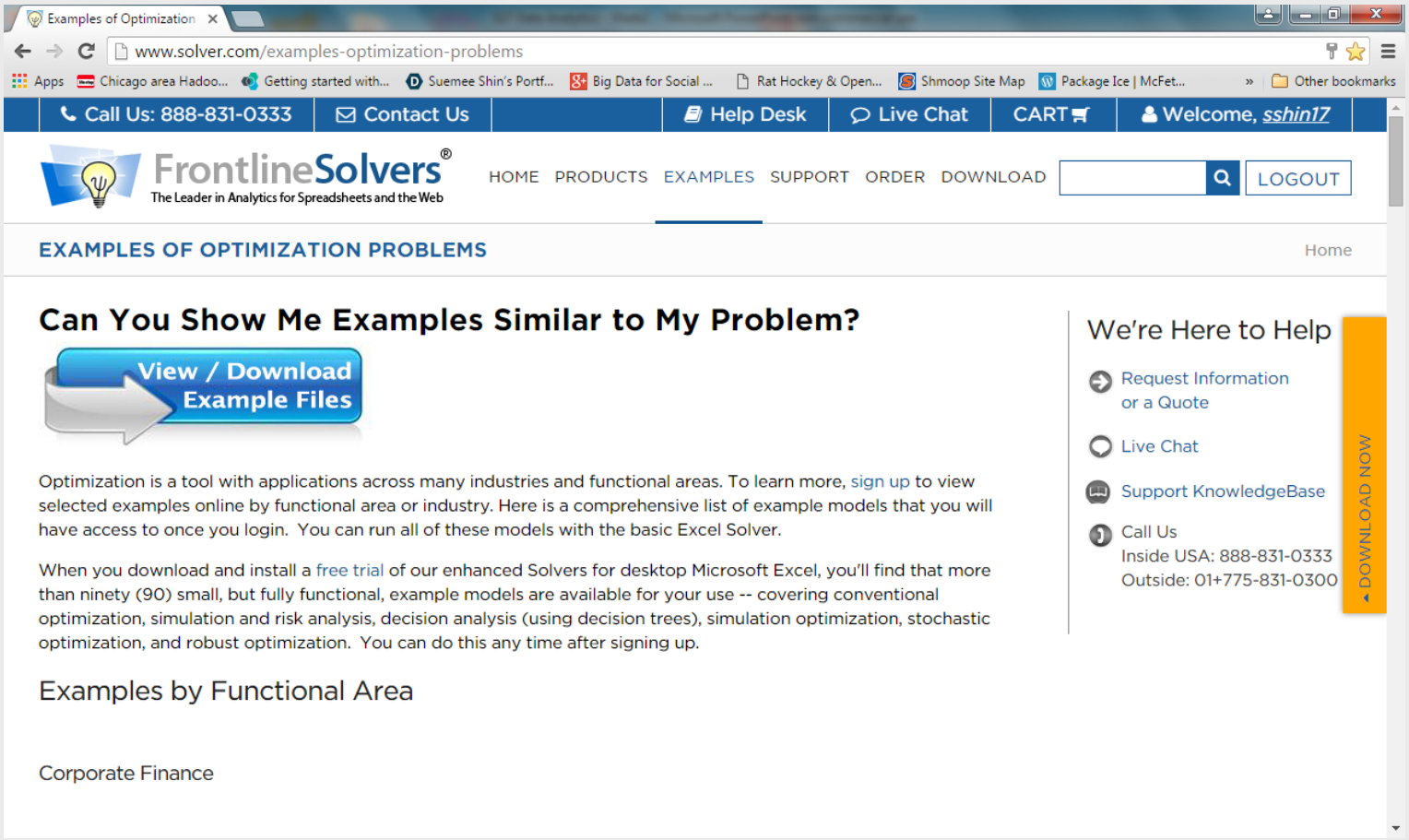
Week 2 Topic:
Back to concession stand example (cont.)

Using Evolutionary and adding bounds to the variables:



Item	Item Price	Count of item
Popcorn	5	30
Beer	4	30
Bottled Water	3	30
Chocolate Dipped Cone	3	30
Hamburger	3	30
Ice Cream Sandwich	3	30
Nachos	3	30
Popsicle	3	30
Soda	2.5	30
Chocolate Bar	2	27
Gummy Bears	2	0
Licorice Rope	2	0
Pizza	2	30
Hot Dog	1.5	0
	Total # of items:	327
	Revenue target:	999

Week 2 Topic: Other examples



Week 2 Topic: Product Mix Example

- Your company manufactures TVs, stereos and speakers, using a common parts inventory of power supplies, speaker cones, etc. Parts are in limited supply and you must determine the most profitable mix of products to build.

		TV set	Stereo	Speaker	
	Number to Build->		0	0	0
Part Name	Inventory	No. Used			
Chassis	450	0	1	1	0
Picture Tube	250	0	1	0	0
Speaker Cone	800	0	2	2	1
Power Supply	450	0	1	1	0
Electronics	600	0	2	1	1
		Profits:			
		By Product	$0	$0	$0
		Total	**$0**		

Week 2 Topic:
Product Mix Example – Set up

Problem

Your company builds TVs, stereos and speakers, using a common parts inventory of power supplies, speaker cones, etc. Parts are in limited supply. What is the best combination of products to build that maximizes profit?

Solution

1) The variables are clearly the number of TVs, stereos and speakers to build. In this worksheet, they are given the name Number_to_build.

2) The constraints specify that the number of parts used cannot exceed the supply. This leads to

Number_used <= Number_available

There is also the logical constraint

Number_to_build >= 0 via the Assume Non-Negative option

3) The objective is to maximize profit. In the ProductMix worksheet this is defined as Total_profit.

Week 2 Topic:
Product Mix Example - Remarks

Remarks

Although this is a good example of a product mix problem, bear in mind the limitations of the model. For example, market demand and price elasticity is not included in the model -- we assume that it doesn't matter how many TVs we build, we will always be able to sell them. Nor are there any pre-specified minimum or maximum of products that need to be made. The effect of introducing these restrictions can be studied by examining a Sensitivity Report, which you can select from the

dialog appears with the message 'Solver found a solution.'