# MITS6005

# Big Data

# Session 3

# Analyze & Present Data

# *Analyze data*

# *Pivot tables*

- A pivot table is an aggregation of a source table
- Supports summary statistics:
  - Count
  - Sum
  - Min/Max
  - Average
- This summary is presented in a table format which can be formatted and filtered
- A table with categories down the rows and across the columns is a ***cross table***

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Rec | Salespers | Code | Item | Region | Month | Year | Units | Sales |
| 2 | 1 | Fred | A | Diary | East | Jan | 1991 | 12 | 3 |
| 3 | 2 | Fred | D | Dictionary | East | Jan | 1991 | 34 | 40 |
| 4 | 3 | Fred | E | Encyclope | East | Jan | 1991 | 52 | 416 |
| 5 | 4 | Fred | N | Novel | East | Jan | 1991 | 34 | 17 |
| 6 | 5 | Fred | A | Diary | East | Feb | 1991 | 6 | 1 |
| 7 | 6 | Fred | D | Dictionary | East | Feb | 1991 | 5 | 6 |
| 8 | 7 | Fred | E | Encyclope | East | Feb | 1991 | 55 | 440 |
| 9 | 8 | Fred | N | Novel | East | Feb | 1991 | 33 | 16 |
| 10 | 9 | Fred | A | Diary | East | Mar | 1991 | 65 | 19 |
| 11 | 10 | Fred | D | Dictionary | East | Mar | 1991 | 34 | 40 |
| 12 | 11 | Fred | E | Encyclope | East | Mar | 1991 | 87 | 696 |
| 13 | 12 | Fred | N | Novel | East | Mar | 1991 | 23 | 11 |
| 14 | 13 | Bert | A | Diary | South | Jan | 1991 | 98 | 29 |
| 15 | 14 | Bert | D | Dictionary | South | Jan | 1991 | 55 | 66 |
| 16 | 15 | Bert | E | Encyclope | South | Jan | 1991 | 21 | 168 |
| 17 | 16 | Bert | N | Novel | South | Jan | 1991 | 11 | 6 |

| **Sum of Sales** | **Month** | | | |
|---|---|---|---|---|
| Salesperson | Jan | Feb | Mar | Grand Total |
| Bert | 7508 | 10360 | 6281 | 24149 |
| Bill | 5113 | 8916 | 7642 | 21671 |
| Fred | 7561 | 9735 | 11221 | 28517 |
| Harry | 10513 | 3583 | 9452 | 23548 |
| Grand Total | 30695 | 32594 | 34596 | 97885 |

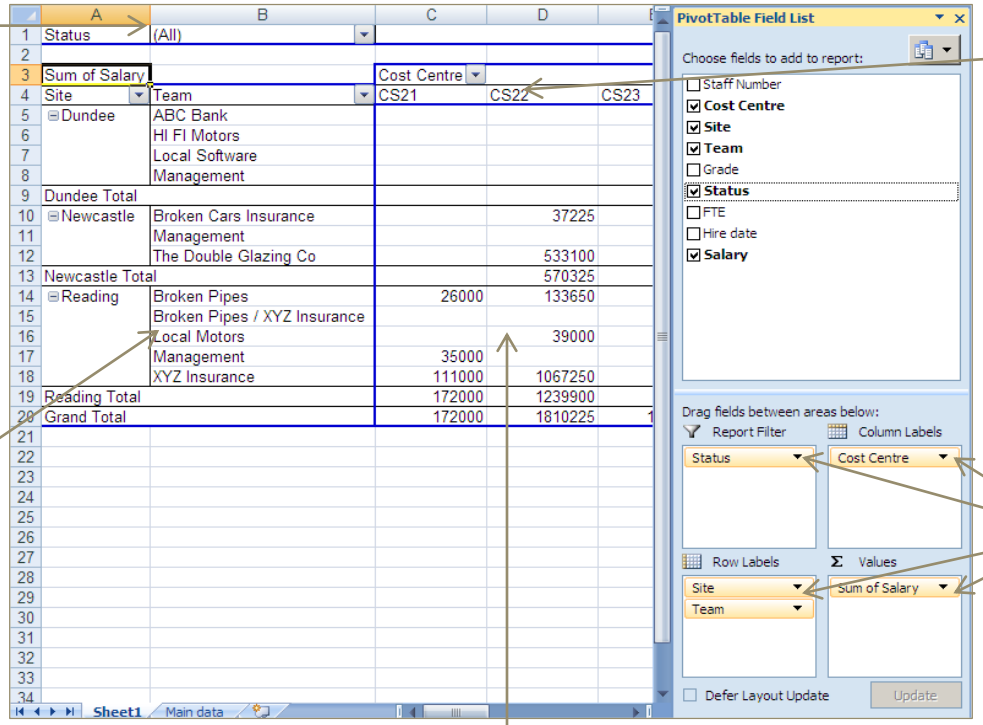
# *Customizing pivot tables*

**Filters**

Change total sample

**Rows**

Can stack multiple fields, e.g. site broken down by team




**Columns**

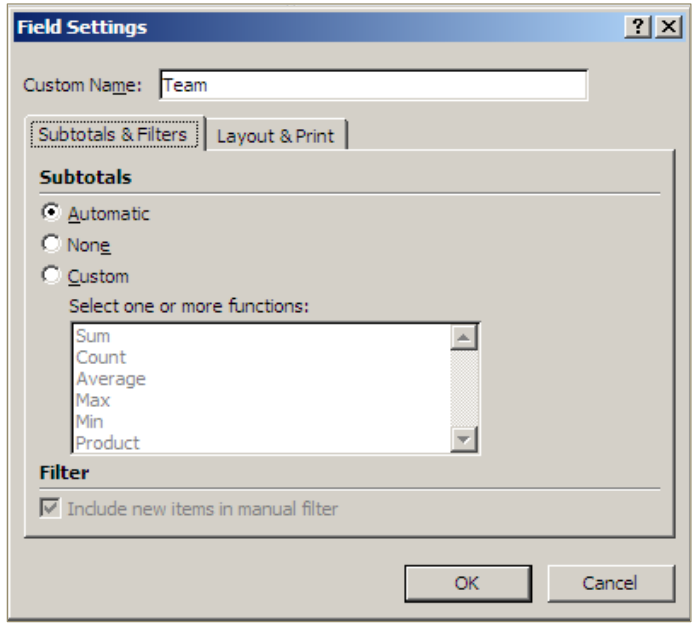
**Options**

Details of the way fields are displayed in the table are changed here

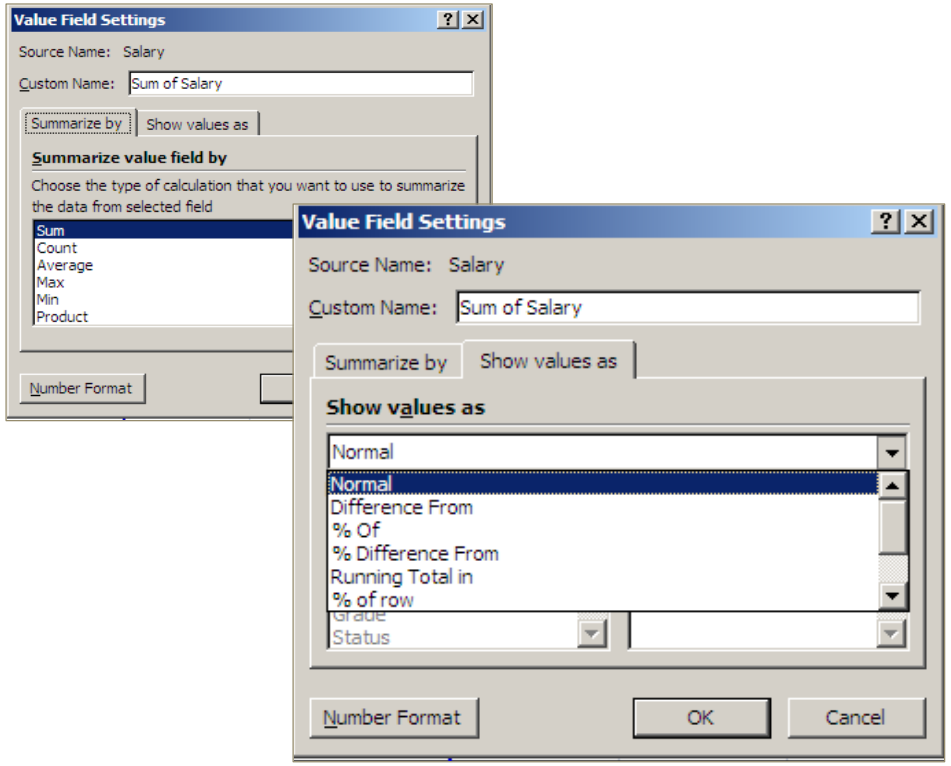
**Cells:** A variable summarised by rows/columns/filters

# *Customizing pivot tables*

**Rows and Columns** can have subtotals added to them, names changed, and additional options applied to the layout.



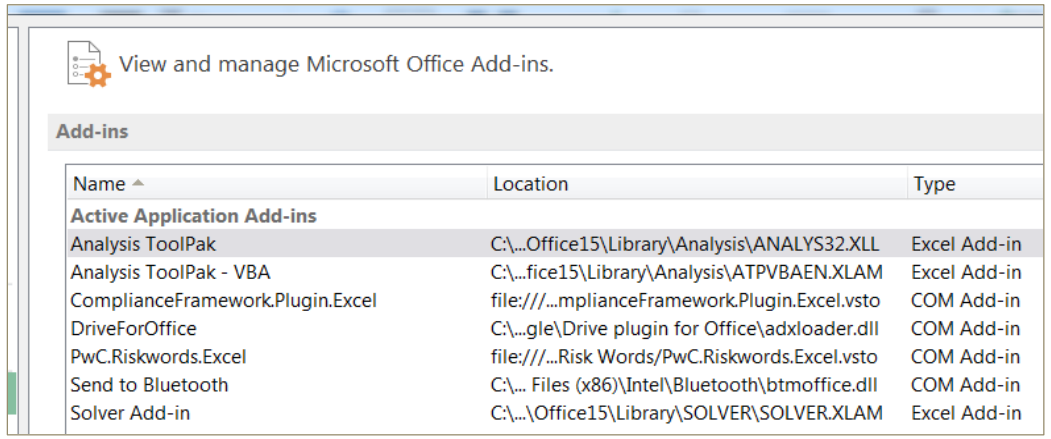

**Value Fields/Cells** can be summarised by different functions and shown as various values

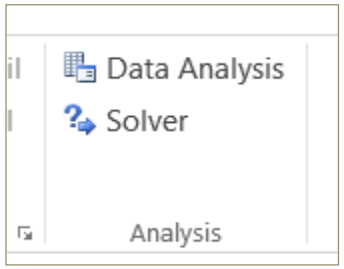


# *Analysis ToolPak*

Excel can be used for more advanced analysis than summary statistics and charts

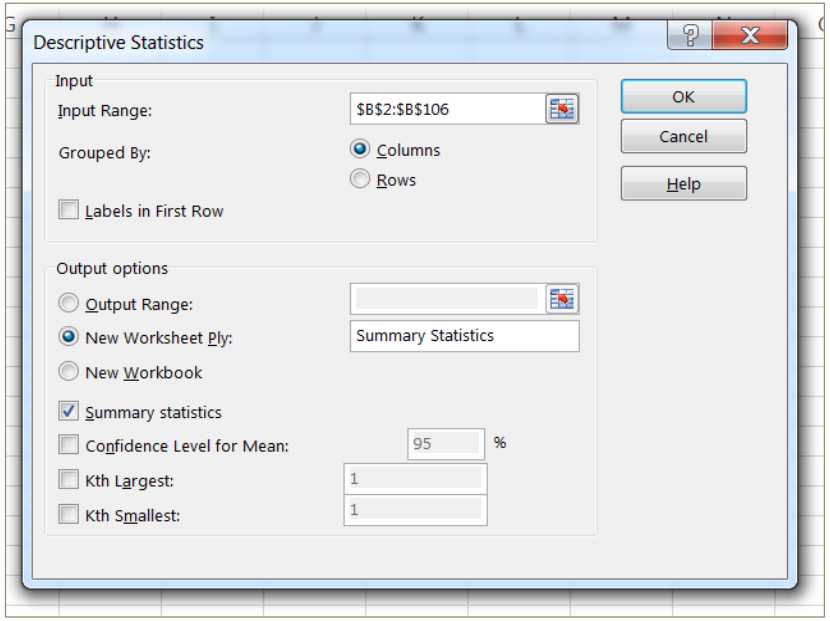
1. Click the File tab, click Options, and then click the Add-Ins category
2. In the Manage box, select Excel Add-ins and then click ‘Go’
3. In the Add-Ins box, check the ‘Analysis ToolPak’ and ‘Solver’ check boxes, and then click ‘OK’
4. Go to the ‘Data’ tab and look for the Analysis section



# *Descriptive statistics*

- Instead of using individual functions, the 'Data analysis' tab enables the user to calculate various statistics for a column at the same time




| *Number of Participants* | |
|---|---|
| Mean | 39.69524 |
| Standard Error | 5.648011 |
| Median | 16 |
| Mode | 2 |
| Standard Deviation | 57.87489 |
| Sample Variance | 3349.502 |
| Kurtosis | 6.352246 |
| Skewness | 2.515343 |
| Range | 271 |
| Minimum | 1 |
| Maximum | 272 |
| Sum | 4168 |
| Count | 105 |
| Largest(2) | 252 |
| Smallest(2) | 1 |
| Confidence Level(95.0%) | 11.20022 |

# *Histograms*

A histogram depicts the frequency or probability distribution of a numeric variable across bins of equal width

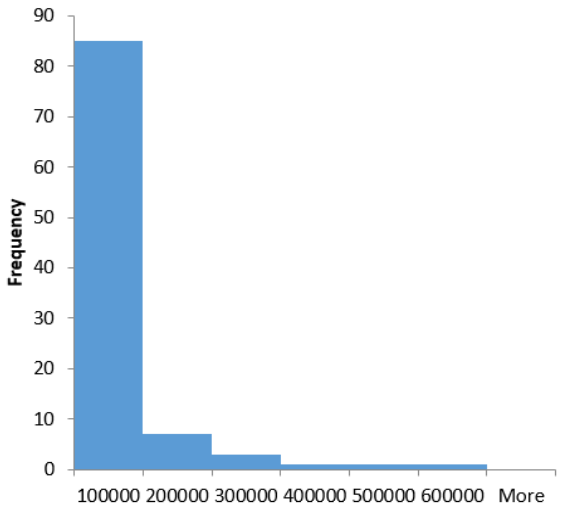
Variable distributions may appear statistically “normal” or display non-normal characteristics such as skewness or kurtosis

Create a histogram through ‘Data analysis’ by specifying the bins in a separate range

| Values |
|---|
| 3973 |
| 25190 |
| 197343 |
| 1434 |
| 25757 |
| 59694 |
| 47147 |
| 74572 |
| 16296 |
| 70653 |
| 11837 |
| 311773 |
| 1813 |
| 8771 |
| 47863 |
| 564760 |
| 17982 |
| 214227 |
| 277423 |

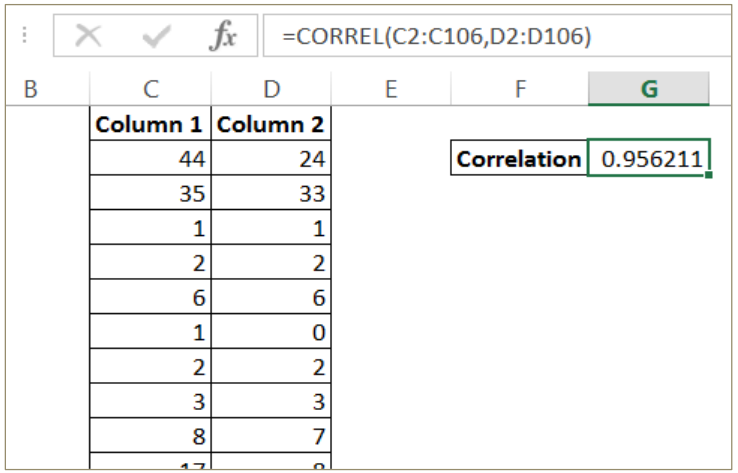
| Bins |
|---|
| 0 |
| 100000 |
| 200000 |
| 300000 |
| 400000 |
| 500000 |
| 600000 |

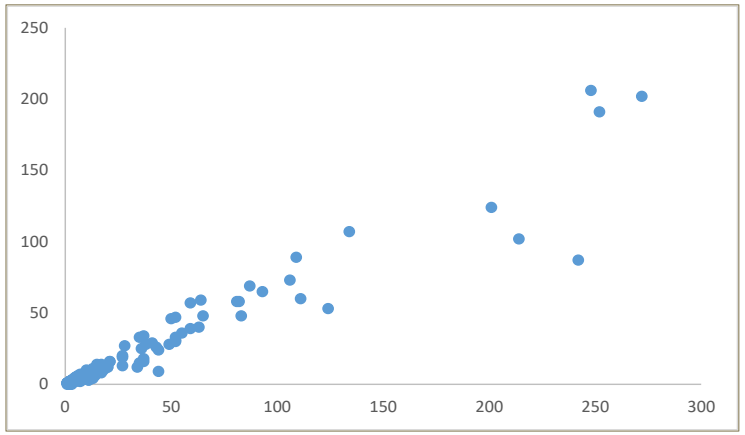
| *0* | *Frequency* |
|---|---|
| 100000 | 85 |
| 200000 | 7 |
| 300000 | 3 |
| 400000 | 1 |
| 500000 | 1 |
| 600000 | 1 |
| More | 0 |



# *Correlation*

- The CORREL() function (and Data Analysis correlation option) quantify the direction and degree of the relationship between two variables
- Correlation coefficient is between -1 and 1:
  - ***Negative values*** indicate the variables are inversely related
  - ***Positive values*** indicate the variables are directly related
  - ***Values close to zero*** indicate a weak correlation
  - ***Values close to 1*** indicate a strong correlation
- Correlation is often best visualized with a scatter plot

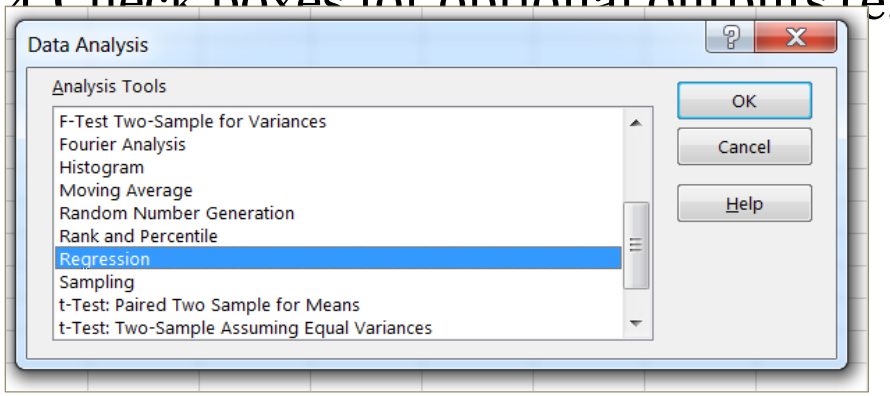


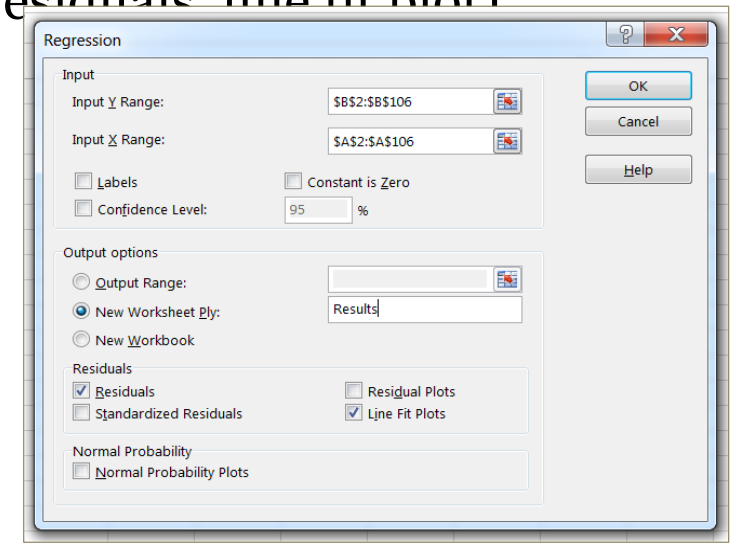



# *Linear regression (univariate)*

In addition to descriptive statistics, Excel can build and evaluate simple predictive models:

1. Select ‘Regression’ from the Data Analysis section
2. Select your dependent (Y) variable (to be predicted)
3. Select your independent (X) variable (used to predict the Y variable)
4. Check boxes for optional outputs (e.g. residuals, line fit plot)



# *Linear regression output*

Actual values

Goodness of fit

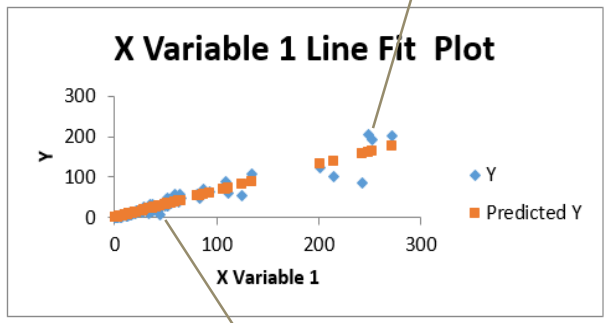
SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.956211442 |
| R Square | 0.914340321 |
| Adjusted R Square | 0.913508674 |
| Standard Error | 11.62475051 |
| Observations | 105 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *ignificance F* |
|---|---|---|---|---|---|
| Regression | 1 | 148571.6 | 148571.6 | 1099.433 | 8.94E-57 |
| Residual | 103 | 13918.89 | 135.1348 | | |
| Total | 104 | 162490.5 | | | |

| | *Coefficients* | *andard Err* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *ower 95.0%* | *pper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.342812827 | 1.377775 | 0.248816 | 0.803998 | -2.38968 | 3.075305 | -2.38968 | 3.075305 |
| X Variable 1 | 0.653072134 | 0.019696 | 33.15769 | 8.94E-57 | 0.61401 | 0.692134 | 0.61401 | 0.692134 |




Predicted values

Coefficient and intercept

# *Exercise #3*

In ‘Payroll Data.xlsx’:

1. What is the average percentage of total compensation that is overtime pay, not including staff with no overtime pay?
2. Create a pivot table with the following statistics by Operating Unit:
   - Number of employees
   - Average total compensation
   - Average tenure (in years)
3. Add Job Title to the pivot table across the columns
4. What is the correlation between total compensation and tenure?
5. Create a histogram for total compensation… how would you characterize the distribution? What if you transform the data to reduce any skewness?
6. Create a linear regression model using tenure to predict total compensation… what is the adjusted R-squared value?

# *Present findings*

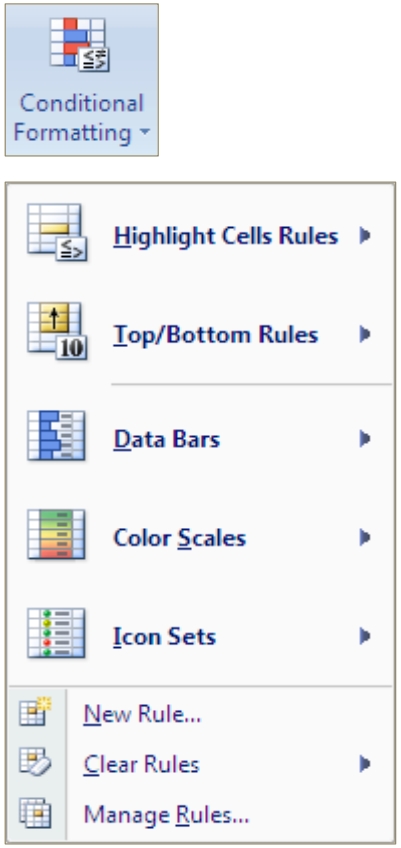
# *Visualisation in excel*

- Visualization is often essential for gaining an understanding of the data and presenting findings to a new audience
- Excel can easily produce a variety of basic charts
- For an effective visualization, always consider:
  - What question do I want to answer?
  - What message do I want my audience to take away?
  - How can I keep it simple?
- We'll focus more on visualization later in the course

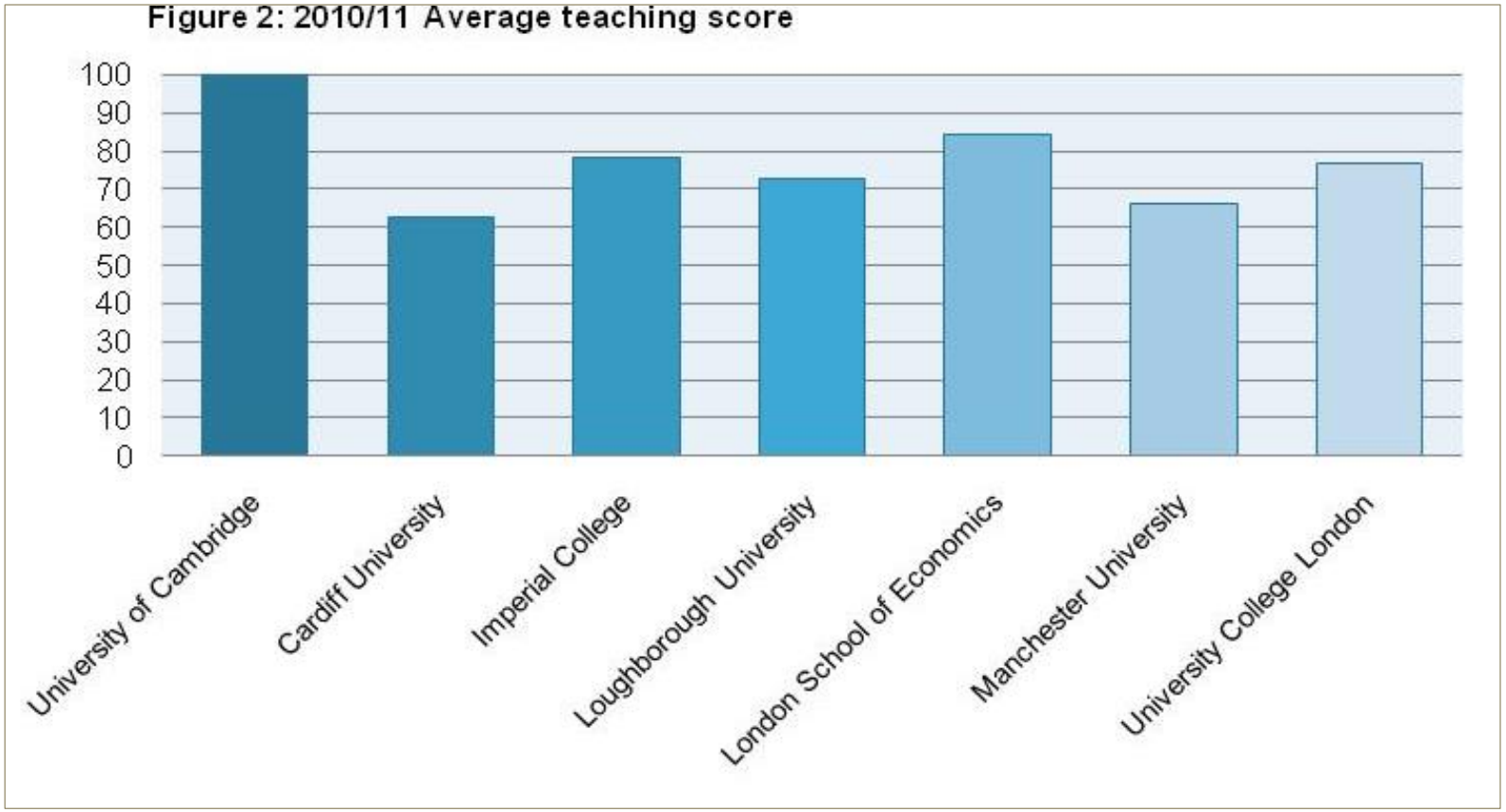
# *Conditional formatting*

- Conditional formatting adjusts the color of a cell according to the relative magnitude of the values or established rules
- It is commonly used for RAG (Red/Amber/Green) reports, in

| Average of Salary | Column Labels | | | | |
|---|---|---|---|---|---|
| Row Labels | CT | NJ | NT | NY | WA |
| Communication & Strategic Sales | $ 138,128.00 | $223,768.50 | | $ 27,605.00 | |
| Finance | | $ 94,055.25 | | | |
| Human Resources | | $ 80,350.00 | | | |
| IT Services | $ 168,449.00 | $ 81,555.33 | $76,752.00 | $ 75,240.60 | |
| Luxury Treats | $ 81,241.93 | $ 67,250.20 | | $152,621.25 | $ 88,688.14 |
| Marketing | | $ 62,922.05 | | $137,842.00 | |
| Natural Health Foods | $ 119,963.13 | $ 84,642.50 | | $165,881.67 | |
| Plant Support Services | | $ 83,708.20 | | $130,655.32 | |
| Tasty Biscuits | $ 80,390.89 | $ 58,220.67 | | $115,482.36 | $ 35,814.25 |
| Tasty Bizkits | $ 94,501.00 | | | $ 72,242.00 | |
| Tasty Snacks Food | $ 105,900.00 | $184,037.00 | | $ 64,903.33 | $ 93,503.00 |
| Tasty Soups | $ 112,294.00 | $ 45,017.67 | | $ 52,969.63 | $120,435.50 |

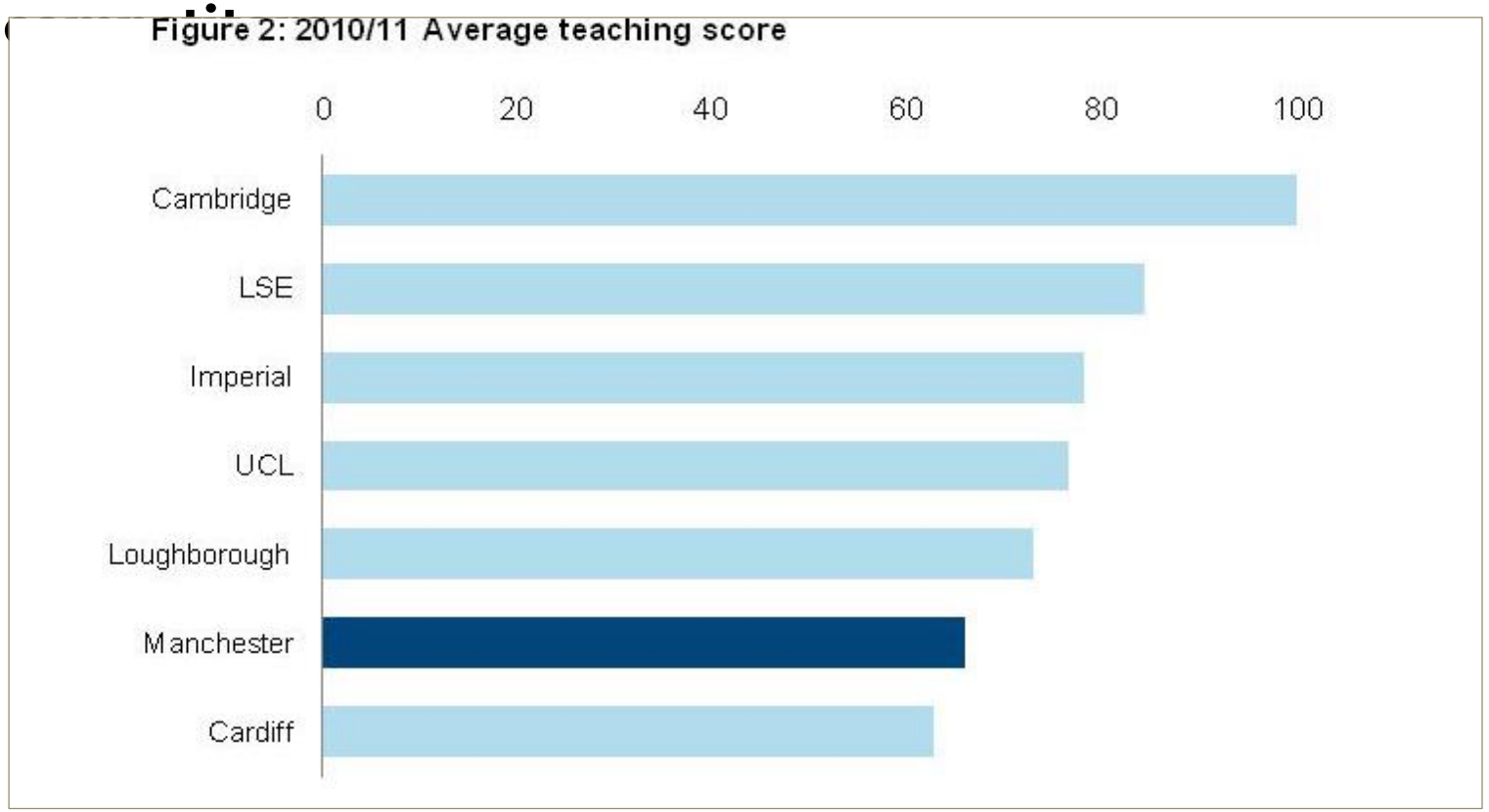


# *What do you think of this visualization?*



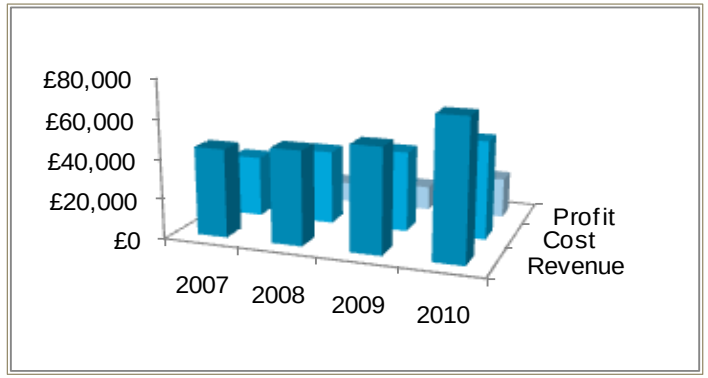
# *What about this one?*

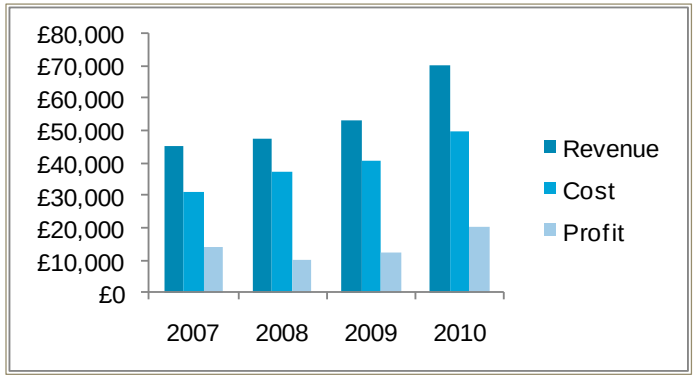
**Manchester University has low teaching score compared to its**



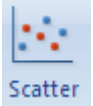
# *Effective visualisation*

- More information, less ink
- Clear out the junk!
- Use color, shape, placement, etc. to draw attention
- Use 2D rather than 3D charts



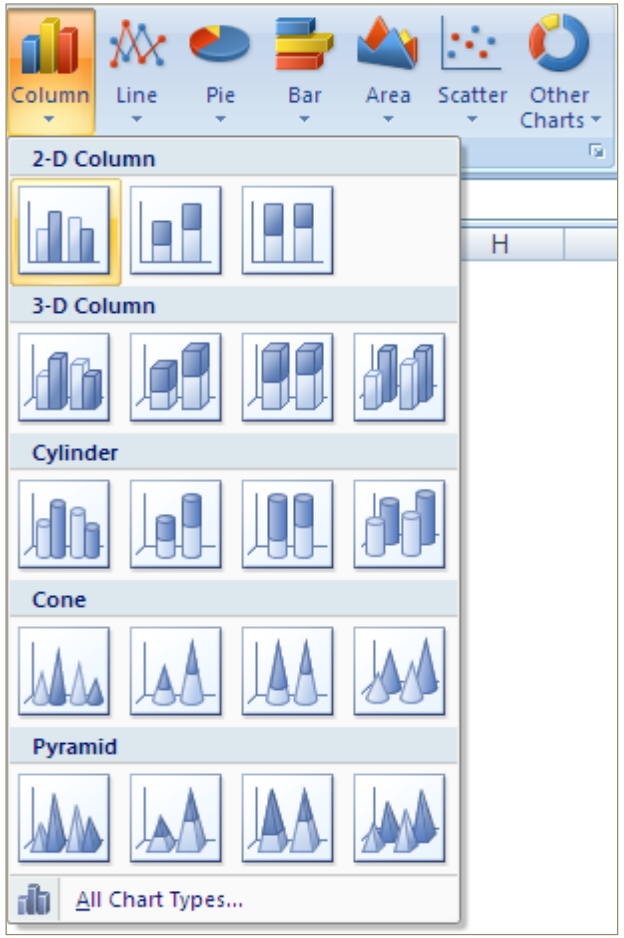
# *Basic charts*

| Type of chart | When to use them | Examples |
|---|---|---|
| Bar/Column     | Discrete values | Number of employees at different offices |
| Line   | Continuous values, over time | Annual revenue |
| Scatter   | Continuous values, two variables | Quantity and price for different products |

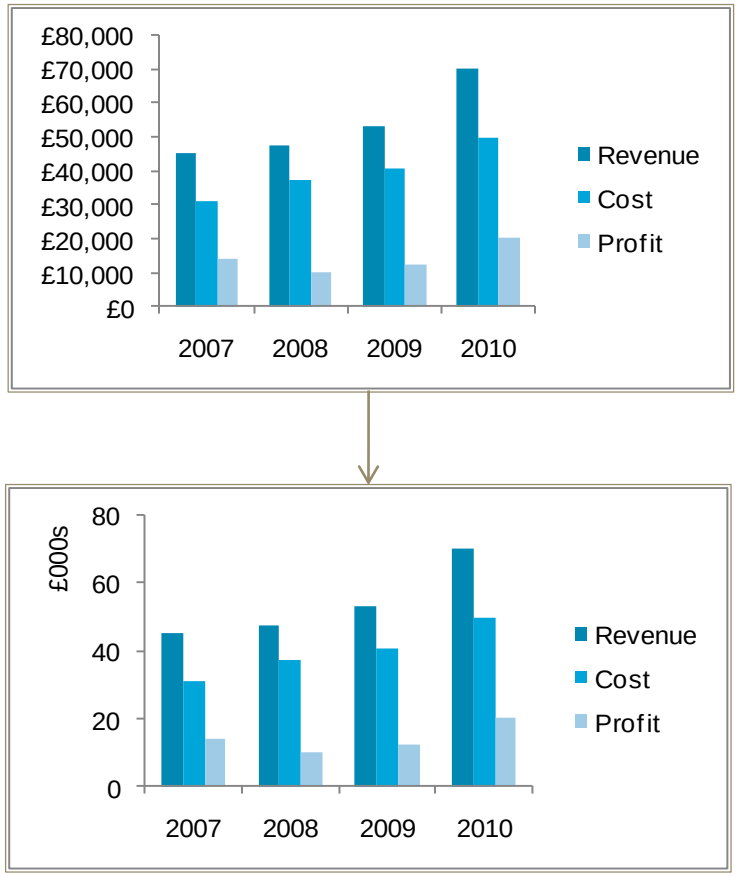
# *Insert chart*

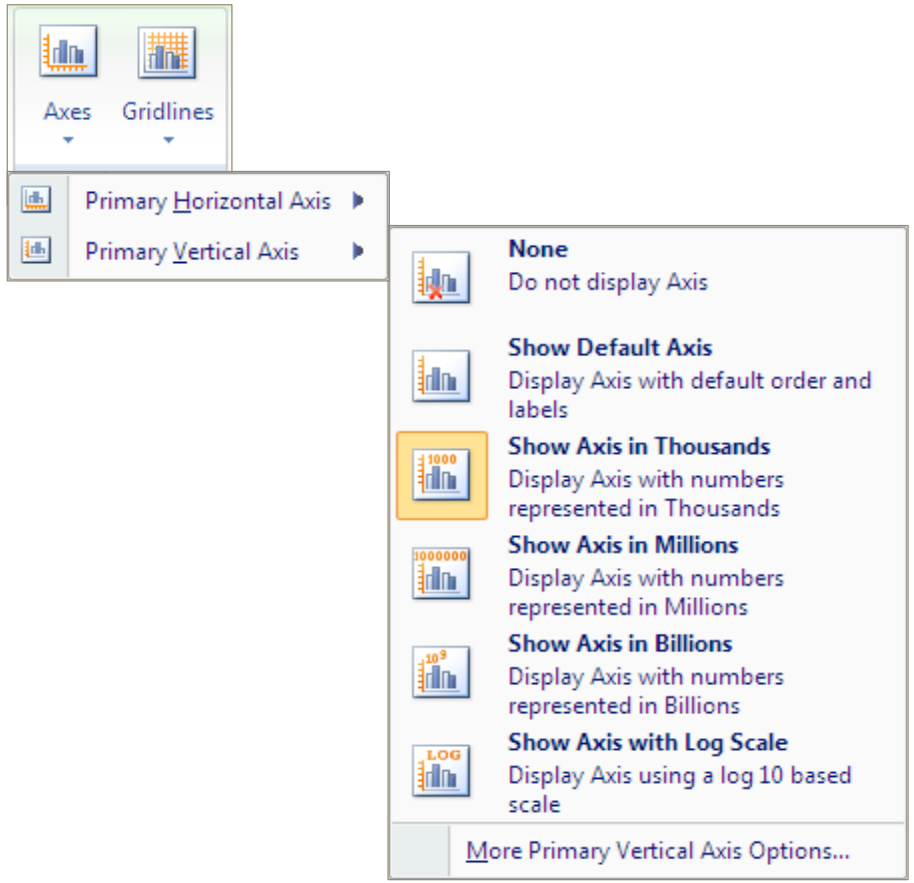
- The first step to setting up your chart is to arrange the data into a table
- Then select the table and select the type of graph you are looking

| | Revenue | Cost | Profit |
|---|---|---|---|
| 2006 | £ 44,300 | £ 28,500 | £ 15,700 |
| 2007 | £ 46,000 | £ 31,100 | £ 13,900 |
| 2008 | £ 47,000 | £ 37,600 | £ 10,200 |
| 2009 | £ 53,000 | £ 40,800 | £ 12,200 |
| 2010 | £ 70,200 | £ 49,800 | £ 20,400 |

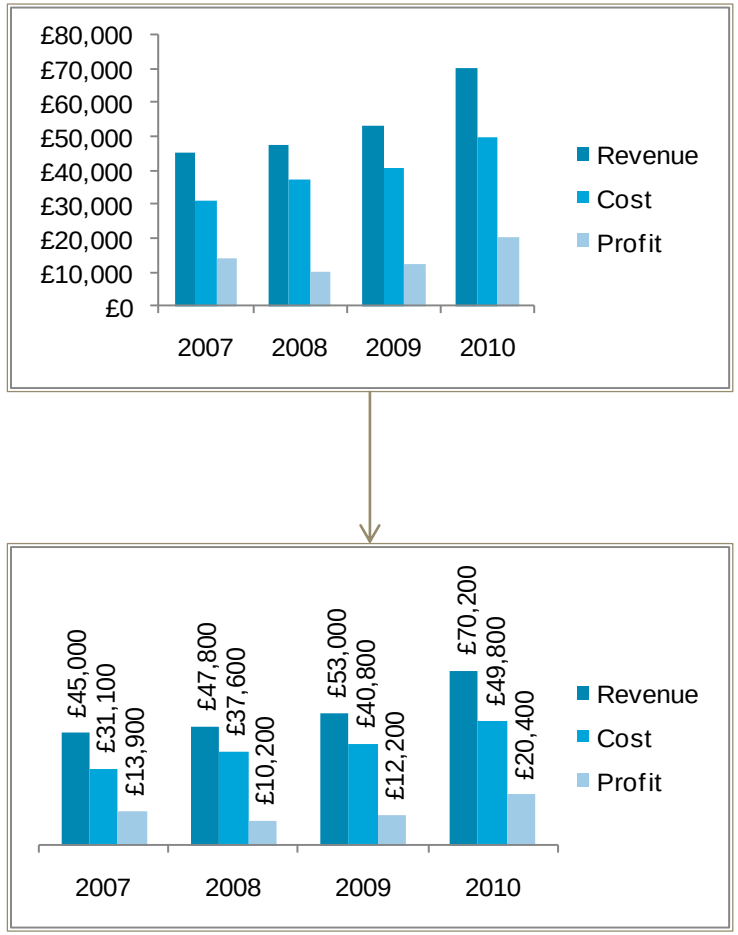


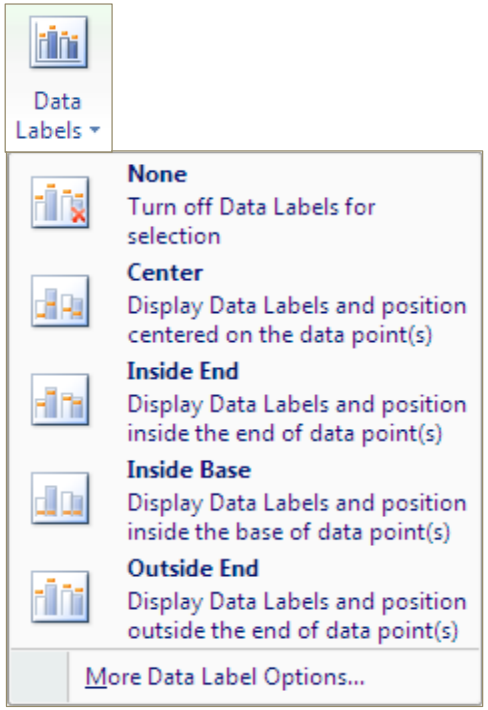
# *Axis settings*



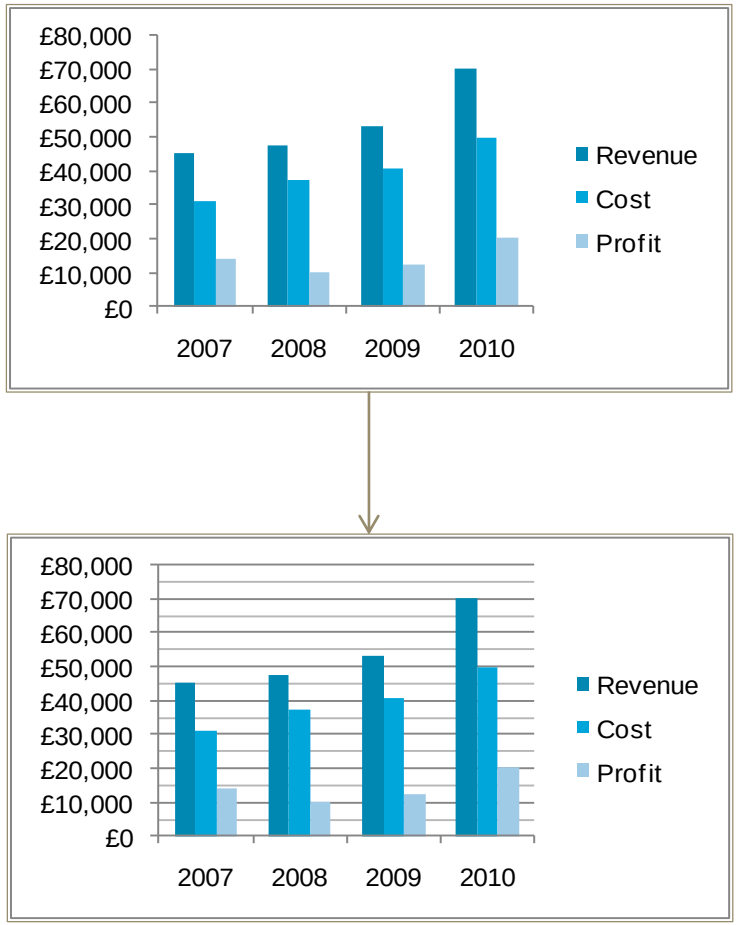



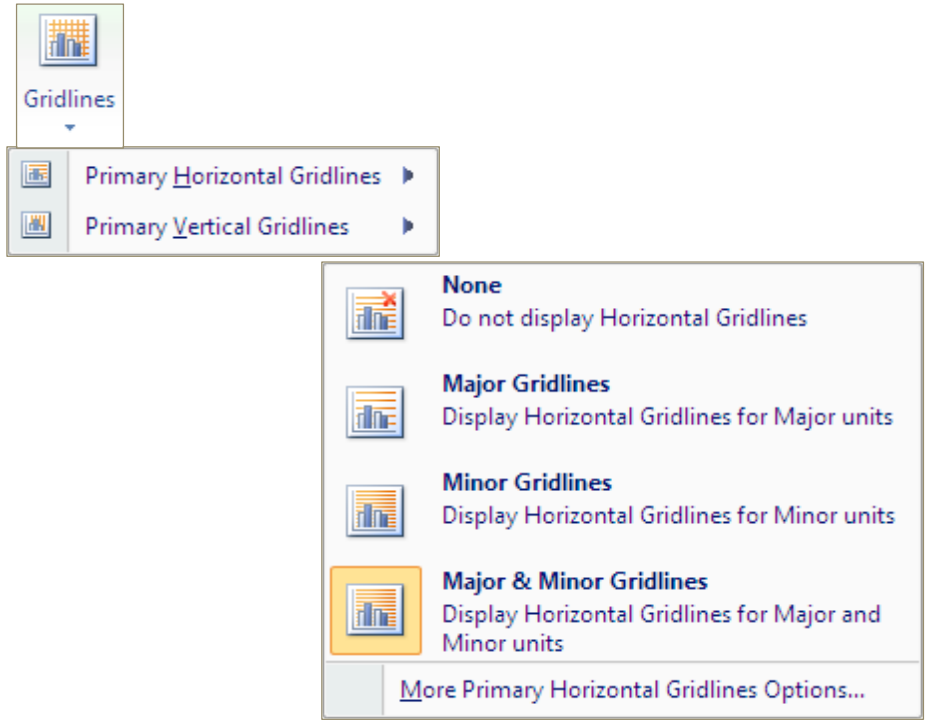
# *Data labels*


£80,000
£70,000
£60,000
£50,000
£40,000
£30,000
£20,000
£10,000
£0
2007
2008
2009
2010
Revenue
Cost
Profit
£45,000
£31,100
£13,900
£47,800
£37,600
£10,200
£53,000
£40,800
£12,200
£70,200
£49,800
£20,400
Revenue
Cost
Profit
2007
2008
2009
2010

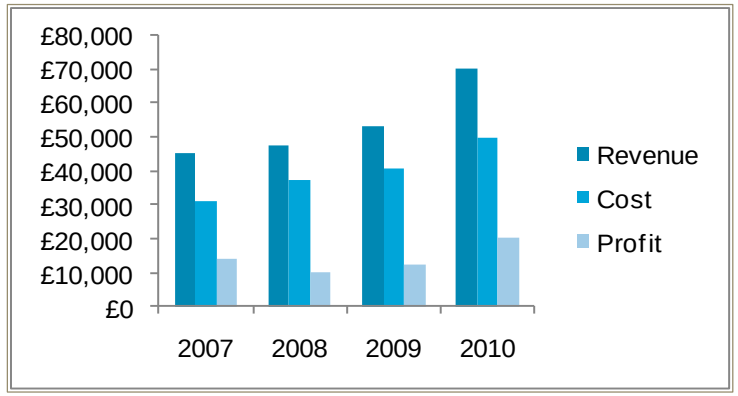

Data Labels
None
Turn off Data Labels for selection
Center
Display Data Labels and position centered on the data point(s)
Inside End
Display Data Labels and position inside the end of data point(s)
Inside Base
Display Data Labels and position inside the base of data point(s)
Outside End
Display Data Labels and position outside the end of data point(s)
More Data Label Options...

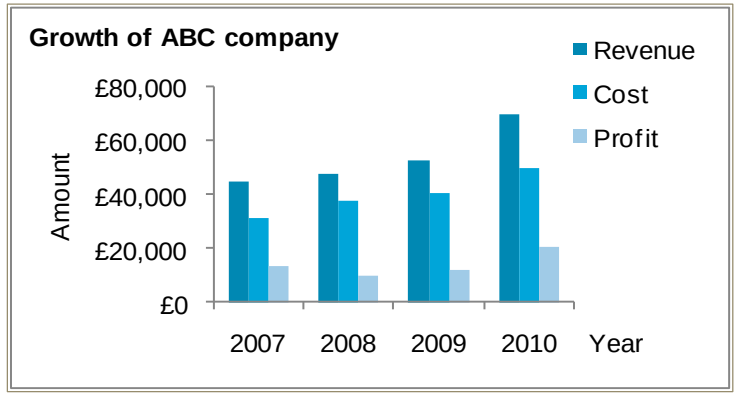
# *Gridlines*


£80,000
£70,000
£60,000
£50,000
£40,000
£30,000
£20,000
£10,000
£0
2007
2008
2009
2010
Revenue
Cost
Profit


Gridlines
Primary Horizontal Gridlines
Primary Vertical Gridlines
None
Do not display Horizontal Gridlines
Major Gridlines
Display Horizontal Gridlines for Major units
Minor Gridlines
Display Horizontal Gridlines for Minor units
Major & Minor Gridlines
Display Horizontal Gridlines for Major and Minor units
More Primary Horizontal Gridlines Options...

# Titles


£80,000
£70,000
£60,000
£50,000
£40,000
£30,000
£20,000
£10,000
£0
2007
2008
2009
2010
Revenue
Cost
Profit


Growth of ABC company
Amount
£80,000
£60,000
£40,000
£20,000
£0
2007
2008
2009
2010
Year
Revenue
Cost
Profit

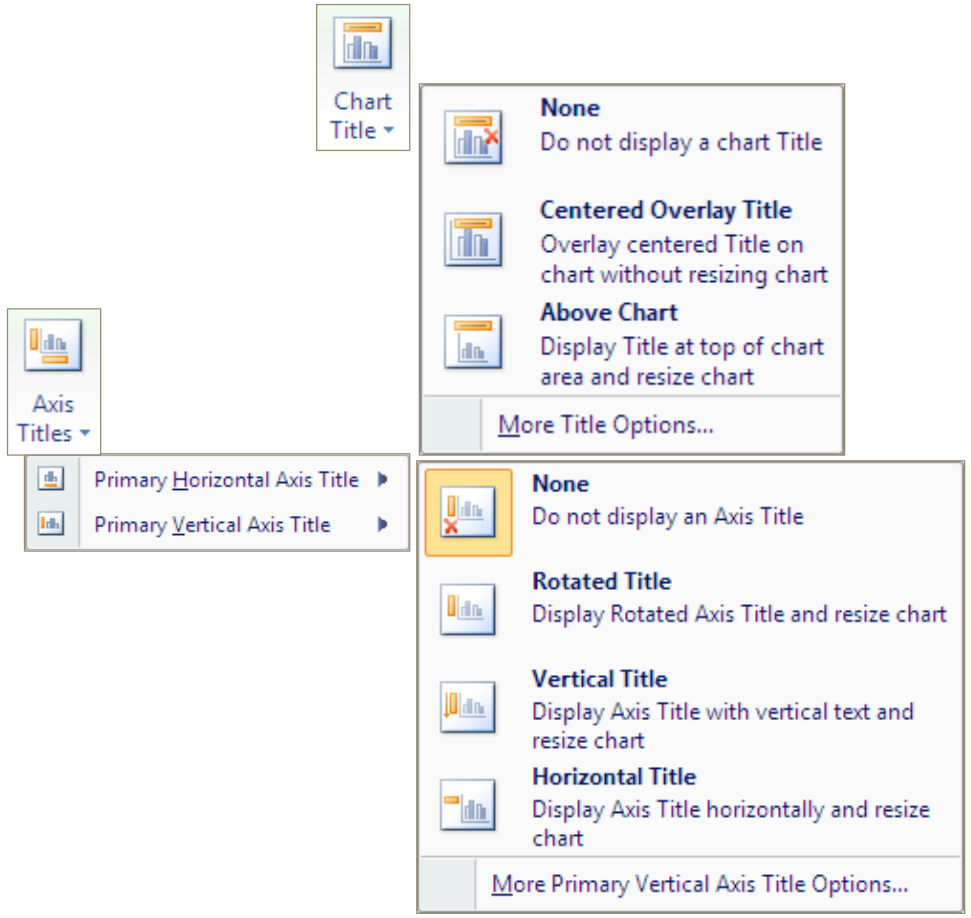
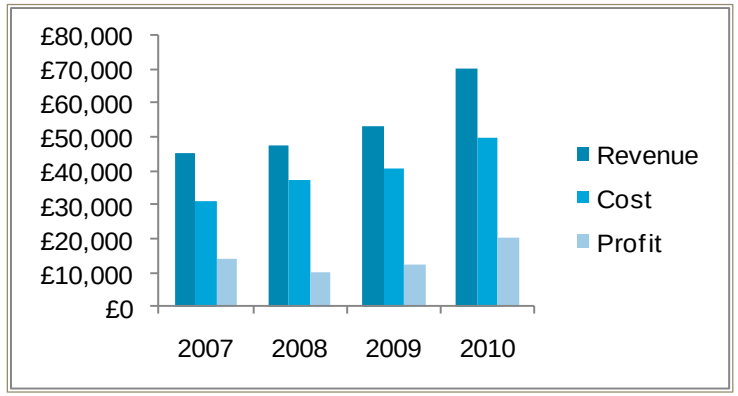


Chart Title
None
Do not display a chart Title
Centered Overlay Title
Overlay centered Title on chart without resizing chart
Above Chart
Display Title at top of chart area and resize chart
More Title Options...
Axis Titles
Primary Horizontal Axis Title
Primary Vertical Axis Title
None
Do not display an Axis Title
Rotated Title
Display Rotated Axis Title and resize chart
Vertical Title
Display Axis Title with vertical text and resize chart
Horizontal Title
Display Axis Title horizontally and resize chart
More Primary Vertical Axis Title Options...

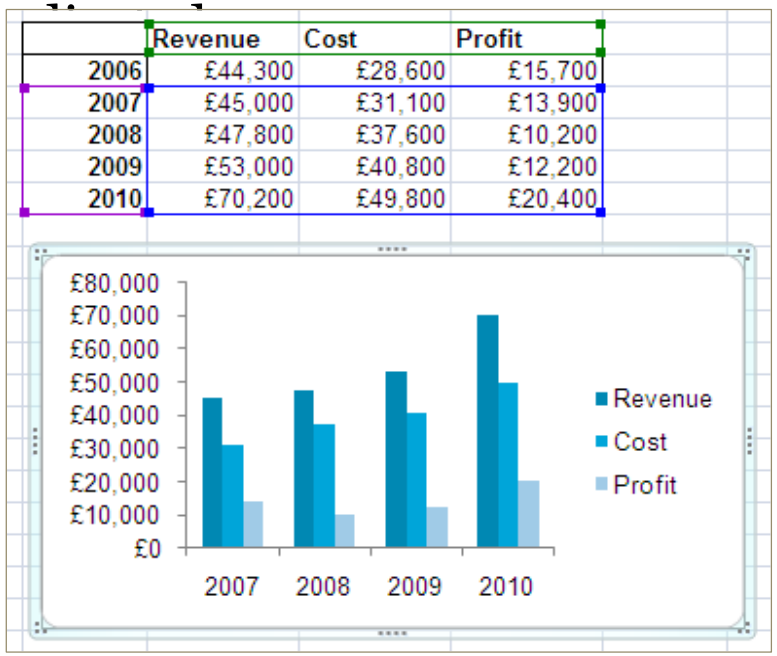
# *Data in rows or columns*

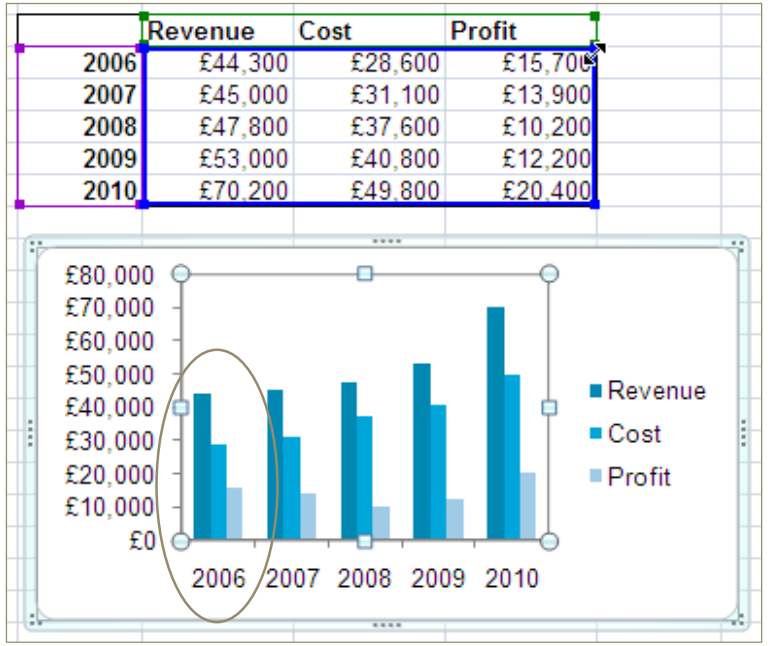



# *Changing data range*

Sometimes we want to change the data range covered by a chart or add an extra series

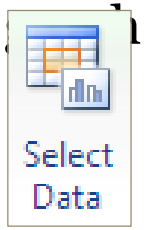
When you select a data series on a chart, the source data range can be



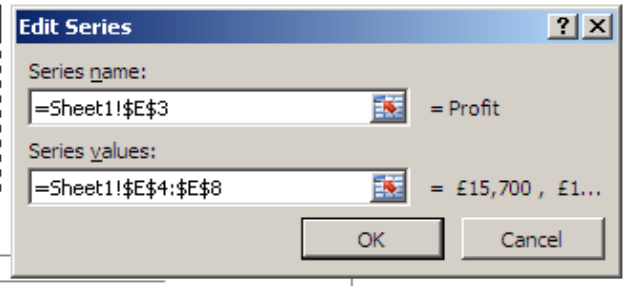



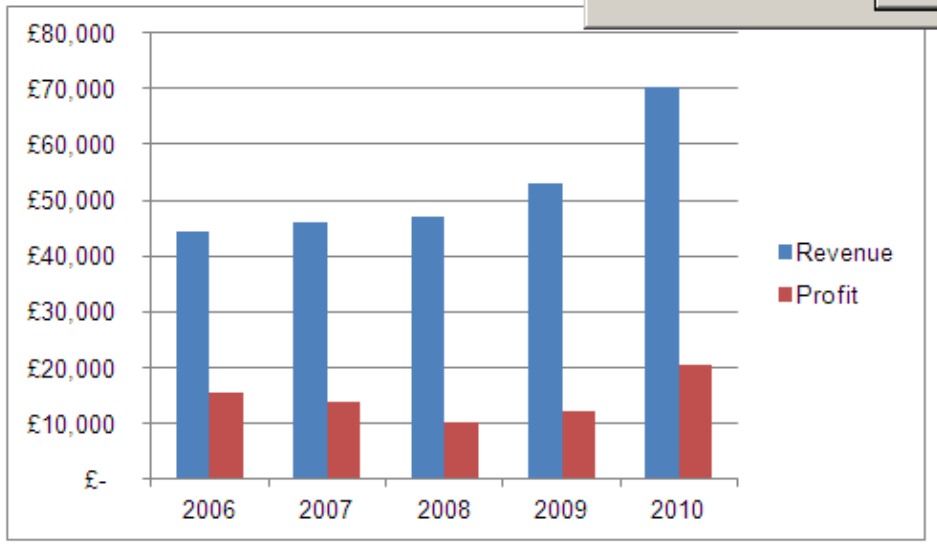
# *Select data source*

You can also use the Select Data Source dialog box from the Design tab in the ribbon or by right-clicking the chart to add a new series to a




| | Revenue | Cost | Profit |
|---|---|---|---|
| 2006 | £ 44,300 | £ 28,500 | £ 15,700 |
| 2007 | £ 46,000 | £ 31,100 | £ 13,900 |
| 2008 | £ 47,000 | £ 37,600 | £ 10,200 |
| 2009 | £ 53,000 | £ 40,800 | £ 12,200 |
| 2010 | £ 70,200 | £ 49,800 | £ 20,400 |



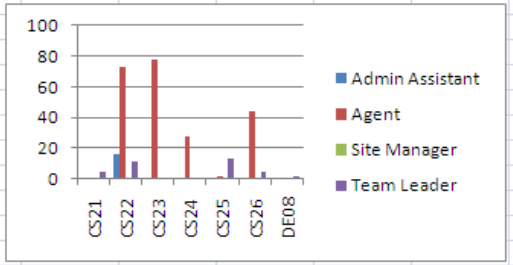



# *Pivot charts*

In addition to creating a Pivot table on the data, you can also create a Pivot Chart which is based on the Pivot table itself

Changes made to the chart are replicated on the table and vice versa

| Sum of FTE | Grade | | | | |
|---|---|---|---|---|---|
| Cost Centre | Admin Assistant | Agent | Site Manager | Team Leader | Grand Total |
| CS21 | | | 1 | 5 | 6 |
| CS22 | 16 | 72.8 | | 11 | 99.8 |
| CS23 | | 77.6 | | | 77.6 |
| CS24 | | 28 | | | 28 |
| CS25 | | 1.85 | 1 | 13 | 15.85 |
| CS26 | | 43.7 | 1 | 5 | 49.7 |
| DE08 | | | | 2 | 2 |
| Grand Total | 16 | 223.95 | 3 | 36 | 278.95 |



# *Exercise #4*

In ‘Payroll Data.xlsx’:

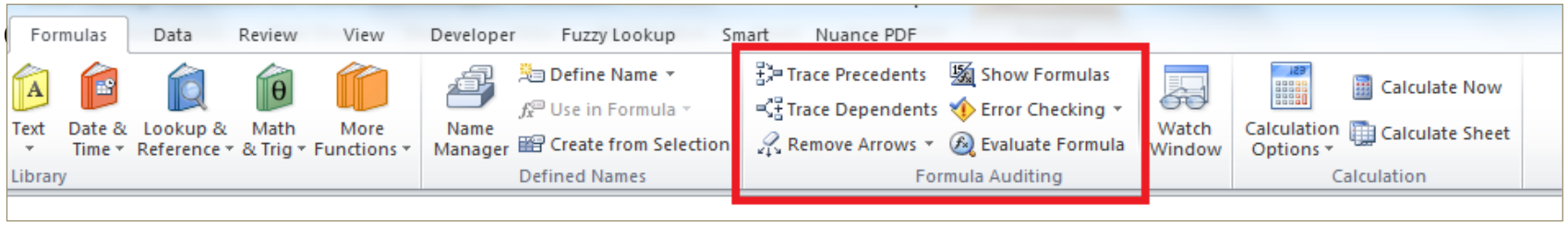
1. Create a pivot table with average salary by Operating Unit and State and add conditional formatting.
2. Create a bar plot with the number of employees by State
3. Create a scatter plot of salary and tenure
4. Create a line chart to show the cumulative number of employees hired by year.

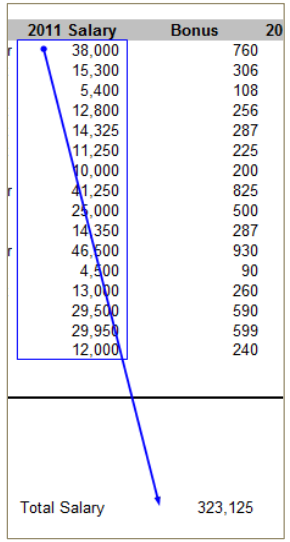
# *Preparing a final workbook*

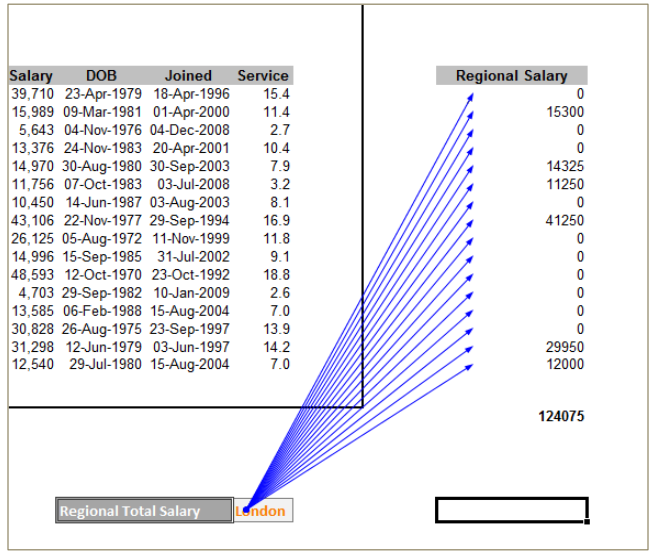
- Getting an Excel file into a presentable state can be a challenge
- Storyboard your workbook and consider how someone would “read” through it
- Keep it simple!
- The following approaches can help with the finishing touches:
  - Formula auditing
  - Page layout
  - Freeze panes
  - Validation
  - Removing gridlines

# *Formula auditing*

To check that formulas have the right cell references, use Trace Precedents or Trace Dependents to display an arrow between linked

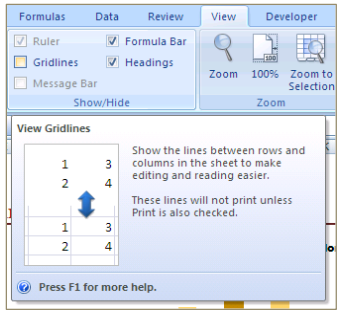


# *Removing gridlines*

One quick tip for making a workbook look instantly less cluttered is to remove the gridlines




**Examples:**

**Approx. one third of the world's population is in two countries**

Fig.1 List of countries with largest population

| Rank | Country | Population | % of World Population |
|---|---|---|---|
| 1 | China | 1,340 | 19% |
| 2 | India | 1,210 | 17% |
| 3 | US | 313 | 4% |
| 4 | Indonesia | 238 | 3% |
| 5 | Brazil | 192 | 3% |
| 6 | Pakistan | 178 | 3% |
| 7 | Nigeria | 162 | 2% |
| 8 | Russia | 143 | 2% |
| 9 | Bangladesh | 142 | 2% |
| 10 | Japan | 128 | 2% |
| 11 | Mexico | 112 | 2% |
| 12 | Philippines | 94 | 1% |
| 13 | Vietnam | 87 | 1% |
| 14 | Ethiopia | 82 | 1% |
| 15 | Germany | 82 | 1% |
| 16 | Egypt | 81 | 1% |
| 17 | Iran | 76 | 1% |
| 18 | Turkey | 74 | 1% |
| 19 | Thailand | 70 | 1% |
| 20 | The Congo | 68 | 1% |
| 21 | France | 65 | 1% |
| 22 | UK | 62 | 1% |
| 23 | Italy | 61 | 1% |
| 24 | South Africa | 51 | 1% |
| 25 | South Korea | 49 | 1% |

**Fig. 2 US Market index**

Source: Bloomberg Stocks and Markets - 23rd Jan

| Country | Index |
|---|---|
| United States | DOW JONES INDUS. AVG |
| United States | DOW JONES TRANS. AVG |
| United States | DOW JONES UTILITIES AVG |
| United States | DOW JONES COMP. AVG |
| United States | DJTSM US |
| United States | S&P 100 INDEX |
| United States | S&P 500 INDEX |
| United States | S&P 400 MIDCAP INDEX |
| United States | S&P 600 SMALLCAP INDEX |
| United States | S&P 1500 Composite Index |
| United States | NYSE COMPOSITE INDEX |
| United States | NYSE U.S. 100 INDEX |
| United States | NYSE INTERNATIONAL 100 |
| United States | NYSE TMT INDEX |
| United States | NYSE WORLD LEADERS INDX |
| United States | VALUE LINE ARITHMETIC |
| United States | NASDAQ COMPOSITE INDEX |
| United States | NASDAQ OTHER FINANCIAL |
| United States | NYSE HEALTHCARE INDEX |
| United States | NYSE FINANCIAL INDEX |
| United States | NASDAQ INDUSTRIAL INDEX |
| United States | NASDAQ TRANSPORTATION IX |
| United States | NASDAQ TELECOMM INDEX |
| United States | NASDAQ INSURANCE INDEX |
| United States | NASDAQ BANK INDEX |

**Examples:**

**Approx. one third of the world's population is in two countries**

Fig.1 List of countries with largest population

| Rank | Country | Population | % of World Population |
|---|---|---|---|
| 1 | China | 1,340 | 19% |
| 2 | India | 1,210 | 17% |
| 3 | US | 313 | 4% |
| 4 | Indonesia | 238 | 3% |
| 5 | Brazil | 192 | 3% |
| 6 | Pakistan | 178 | 3% |
| 7 | Nigeria | 162 | 2% |
| 8 | Russia | 143 | 2% |
| 9 | Bangladesh | 142 | 2% |
| 10 | Japan | 128 | 2% |
| 11 | Mexico | 112 | 2% |
| 12 | Philippines | 94 | 1% |
| 13 | Vietnam | 87 | 1% |
| 14 | Ethiopia | 82 | 1% |
| 15 | Germany | 82 | 1% |
| 16 | Egypt | 81 | 1% |
| 17 | Iran | 76 | 1% |
| 18 | Turkey | 74 | 1% |
| 19 | Thailand | 70 | 1% |
| 20 | The Congo | 68 | 1% |
| 21 | France | 65 | 1% |
| 22 | UK | 62 | 1% |
| 23 | Italy | 61 | 1% |
| 24 | South Africa | 51 | 1% |
| 25 | South Korea | 49 | 1% |

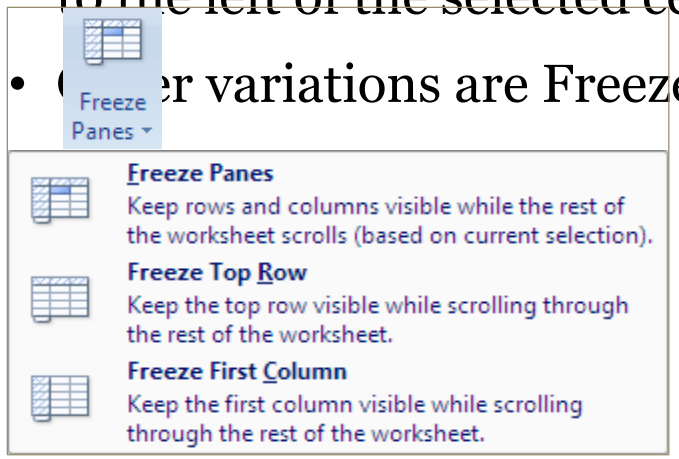
**Fig. 2 US Market index**

Source: Bloomberg Stocks and Markets - 23rd Jan

| Country | Index |
|---|---|
| United States | DOW JONES INDUS. AVG |
| United States | DOW JONES TRANS. AVG |
| United States | DOW JONES UTILITIES AVG |
| United States | DOW JONES COMP. AVG |
| United States | DJTSM US |
| United States | S&P 100 INDEX |
| United States | S&P 500 INDEX |
| United States | S&P 400 MIDCAP INDEX |
| United States | S&P 600 SMALLCAP INDEX |
| United States | S&P 1500 Composite Index |
| United States | NYSE COMPOSITE INDEX |
| United States | NYSE U.S. 100 INDEX |
| United States | NYSE INTERNATIONAL 100 |
| United States | NYSE TMT INDEX |
| United States | NYSE WORLD LEADERS INDX |
| United States | VALUE LINE ARITHMETIC |
| United States | NASDAQ COMPOSITE INDEX |
| United States | NASDAQ OTHER FINANCIAL |
| United States | NYSE HEALTHCARE INDEX |
| United States | NYSE FINANCIAL INDEX |
| United States | NASDAQ INDUSTRIAL INDEX |
| United States | NASDAQ TRANSPORTATION IX |
| United States | NASDAQ TELECOMM INDEX |
| United States | NASDAQ INSURANCE INDEX |
| United States | NASDAQ BANK INDEX |

# *Freeze panes*

- As worksheets can get very large, it is important to ensure that the data being viewed on screen at all points has titles and comments representing the appropriate columns
- Freeze panes allows for parts of the Excel document to be frozen, useful to preserve titles and headings
- Freeze Panes freezes all rows above the selected cell and all columns to the left of the selected cell
- Other variations are Freeze Top Row and Freeze First Column



# *Page layout*

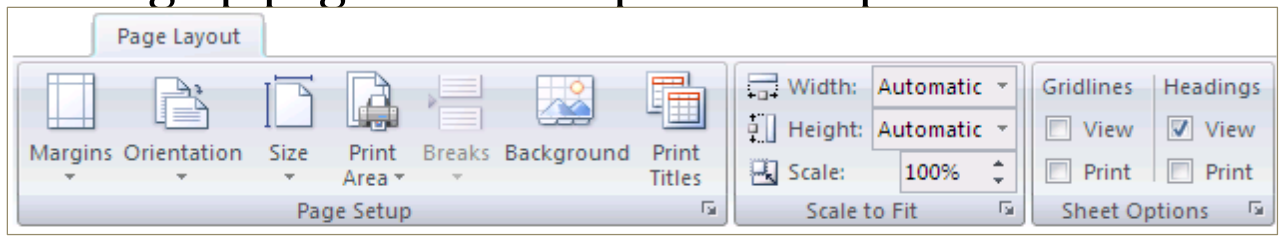
To ensure that the worksheet prints neatly, you may need to change settings on the Page Layout tab:

Changing the page orientation (Portrait vs Landscape)

Adjusting the margins

Specifying a print area (the part of the worksheet that will be printed)

Setting up page breaks at particular points

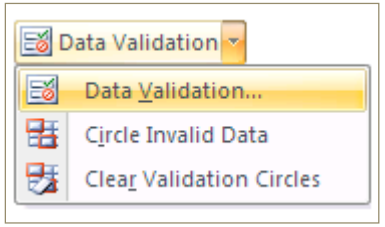


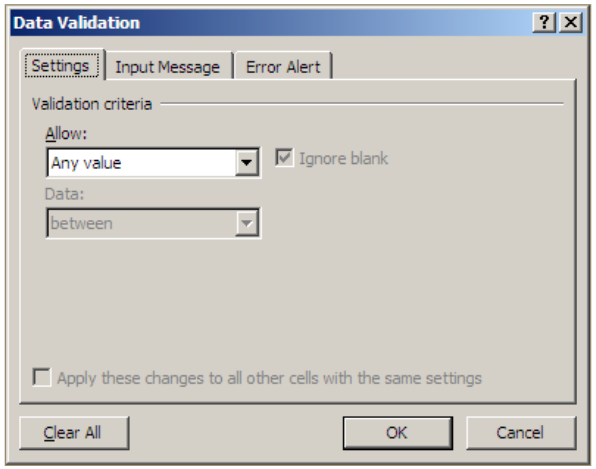

Use Print Preview to check how your worksheet will look when printed

# *Data validation*

Data validation controls what type of data a user can input into a cell, for example:

- Only a number within a certain range
- Only a time/date within a certain period
- Only an item from a predefined list



# *More leading practices*

- Save your workbook with A1 as the active cell
- Put a title and description for the workbook in the upper left corner of the first sheet
- Name your spreadsheet tabs so users can easily navigate throughout your workbook
- Don’t hide columns or rows; instead, group them
- Consider if hardcoded parameters make sense or should be avoided
- Be careful with merged cells
- Keep source data in the workbook
- Break down complicated formulas

# *Summary*

# *Acquire data*

| Task | Description | Excel |
|---|---|---|
| Data access | Connect to a data source | • File > Open<br>• Open in text editor and copy/paste |
| Importing data | Read the data into an analytical environment | • Text Import Wizard<br>• Data > Text to Columns |
| Data profiling | Review data dimensions and summary statistics | • COUNT()<br>• MIN(), MAX(), etc. |
| Data quality assessment | Identify aspects of the data that pose challenges for subsequent analysis | • Sort<br>• Filter<br>• COUNTBLANK() |
| Data simulation | Generate data based on analytical requirements | • RAND()<br>• RANDBETWEEN()<br>• CHOOSE() |

# *Transform data*

| Task | Description | Excel |
|---|---|---|
| Cleaning data | Address data quality issues to facilitate analysis | • Find/Replace |
| Changing data types | Convert a value to the appropriate format for analysis | • Format |
| Filtering data | Create subsets of records and features based on specified conditions | • Filter<br>• IF() |
| Deriving data | Create new features from original features | • MID()<br>• FIND()<br>• LEN()<br>• ROUND()<br>• WEEKDAY()<br>• … |
| Scaling data | Put features with different ranges of values on the same scale while preserving relative values | • SUM()<br>• AVERAGE()<br>• EXP() |

# *Transform data*

| Task | Description | Excel |
|---|---|---|
| Sampling data | Create subsets of records based on a probability distribution | • RAND()<br>• RANDBETWEEN() |
| Aggregating data | Return a statistic or value for one feature according to different values of another feature | • Pivot Table |
| Reshaping data | Change whether values are represented in different records or different features | • Pivot Table |
| Concatenating data | Combine data sets through juxtaposition | • Cut and paste |
| Merging data | Combine data sets by matching records on a common identifier | • VLOOKUP()<br>• HLOOKUP()<br>• INDEX()/MATCH() |

# *Analyze data*

| Task | Description | Excel |
|---|---|---|
| Summary analysis | Calculate representative statistics for features of interest | • AVERAGE()<br>• MEDIAN()<br>• PERCENTILE.INC() |
| Perform statistical tests | Estimate the probability that the data supports a specific claim | • Data Analysis Toolpak |
| Clustering | Identify similar groups of records | |
| Predictive modeling | Use one set of features to predict the value of another feature | • Data Analysis Toolpak |
| Network analysis | | |

# *Present findings*

| Task | Description | Excel |
|---|---|---|
| Data visualization | Display data using lines, shapes, colors, and other abstract representations | • Charts |
| Dashboarding | Create a collection of dynamic visualizations | • Charts |
| Exporting data | Produce output from an analytical environment for future use | • File > Save As |
| Make recommendations | Use results of data analysis to guide decision-making | |