Introduction to Microsoft Excel: 10/19
Intermediate Excel: 10/20
Instructor: Abigail Haddad

Microsoft Excel is a spreadsheet program that offers a host of features for cleaning and analyzing data, creating visuals, and constructing interactive forms and tools for others to use. With a small amount of training, you can do all of these on a level that will useful in a variety of roles.

This syllabus contains the materials for two, one-day workshops: the introductory class and the intermediate class. You can take either or both. Both of these are geared at people who will be using Excel for analyst-type tasks: cleaning, analyzing, and presenting data.

You need a computer running Excel (2010 or later recommended).

On the first day, we'll learn the following:

- Basic formulas and functions
- Pivot tables
- Making a graph or chart
- Dashboards

On the second day we'll learn:

- More advanced functions
- Data management tools (merging, grouping by, summarizing, string functions)
- Macros

What can you do in Excel?

- Clean data
- Analyze data
- Make a table
- Make a graph
- Make a form for someone else to fill out
- A bunch of the same stuff you can do in SQL (merge tables, select, drop, summarize)
- Write code (or modify someone else's code) to automate all of these things

What should you do in Excel? This depends on:

- What other programs you and/or your collaborators know how to use
- What other programs you have access to
- Your timeline
- Whether what you're doing needs to be easily checkable and/or repeatable
- How much data you're working with
- Whether there are privacy concerns which mean you can't put your stuff on the web

Alternatives to Excel

- For cleaning/analyzing data: SQL, SAS, Stata, Python, R, SPSS, Minitab – any command-line framework will be able to handle the basic data cleaning/analysis that most Excel users use it for
- For data visualization, you can also use one of those frameworks, or you can use a data visualization/business intelligence software like Tableau or Qlik.
- For forms: google forms, Adobe Reader, Word

Introduction to Excel
10/19

Download the following files:

http://www.education.pa.gov/Documents/Data%20and%20Statistics/Cohort%20Graduation%20Rates/2015-2016%20Pennsylvania%204-Year%20Cohort%20Grad%20Rates.xlsx

http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/SAT-ACT/2016%20SAT%20Scores%20Public%20Schools.xlsx

Basic terms

- A cell is one box in Excel
- A column is a vertical set of cells
- A row is a horizontal set of cells
- A spreadsheet is made up of rows and columns,
- A tab contains one spreadsheet; a file can have multiple tabs
- A pivot table is a data structure that reads in the data from a spreadsheet and summarizes it

Manipulating Data

Exercise 1

- Replicate the graduation rate percentages in the "Statewide Graduation Rate" tab. Use both a formula and the quotient function
- Make a table with these numbers and format it – borders, fonts, percent, number of digits - the way you would for putting in a report.

Exercise 2

- Do the same thing with the first five rows of data in the "Grad Rate by LEA" tab:
  - Copy the LEA, Total Grad, and Total Cohort cells into a new tab
  - Generate percentages using either a formula or the quotient function
  - Format it for publication

Making Graphs

Exercise 1

- Make a bar chart of the "Statewide Graduation Rate" results
- Give it a title, try different design templates, customize axis formatting, data labels via "add chart element"
- Make a pie chart for female graduation rate, using formulas to generate the numbers you need. Copy the chart and swap out data to replace it with male graduation rate.
- Make a scatterplot showing the relationship between male and female graduation rates. Play with different trend lines.

Exercise 2

- Do the same thing for your "Grad Rate by LEA" table, including creating pie charts and swapping out data and creating a scatterplot comparing two sets of percentages.

Using Functions

Exercise 1

- We're interested in whether big LEAs have on average a higher or lower graduation rate relative to smaller LEAs
- Use the "median" function to get the median cohort size by LEA
- Insert a column for Cohort Size; use the "if" function in combination with your median result to declare whether an LEA is big or small
- Use "countif" to verify that half of your LEAs are characterized as big and half as small
- Use "averageif" to find the means of each group
- Format the result in a table

Exercise 2

- Do the same thing but at the school level

Using Pivot Tables

Exercise 1

- We're going to do the same thing we just did with LEAs via functions, but using a pivot table instead.

Exercise 2

- Do the same thing, but at the school level.

Creating Dashboards

We're going to create an interactive dashboard using some school data, and then you're going to create your own.

Additional Exercises using SAT data

- What are the top-scoring schools? Use filter to get the top 10 scoring schools in terms of composite average score. Copy LEA, school name, and score into a separate tab, and format it for a report.
- Create a scatterplot showing the relationship between number of students tested and composite average score.
- Create a scatterplot showing the relationship between average math score and average reading score.
- Use a pivot table to group by LEA and get average scores. Get both average score of a school within the LEA and average score of a student within the LEA.
- Do charter schools have higher or lower composite SAT scores relative to non-charter schools? Charter schools have "CS" in their names. Use the search function in a new variable for Charter School: SEARCH("CS", D2) to search a cell for whether it finds "CS". If it does, it'll return a number, if it doesn't, it'll return an error. Use the filter to order results, and replace the errors with "Not Charter School" and the non-errors with "Charter School." Use the averageif function to determine which set of schools have the higher scores.

Intermediate Excel

Download the following files:

http://www.education.pa.gov/Documents/Data%20and%20Statistics/Cohort%20Graduation%20Rates/2015-2016%20Pennsylvania%204-Year%20Cohort%20Grad%20Rates.xlsx

http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/SAT-ACT/2016%20SAT%20Scores%20Public%20Schools.xlsx

Merging Data and Sorting

Exercise 1

- We're interested in looking at which schools have graduation rates which deviate significantly from the graduation rates of their LEAs.
- In order to do this, we need to join or merge data from the LEA tab with the school tab.
- Create a column in the school tab for LEA graduation rate.
- Use "match" and "index." The result will be the same as a left inner join.
- Use the absolute value function to get overall difference, and use "sort" to see which schools differ the most from their LEAs.

Exercise 2

- Do the same thing, but get male graduation rate.

Grouping By and Summarizing

Exercise 1

- We're going to use the school data to duplicate the overall LEA graduation rate using a pivot table.
- Verify if the results match by merging the data via match and index, and then countif to count the true and false results.

Exercise 2

- Do the same thing using a different column/demographic group's graduation rate.

Grouping By II and Summarizing

Exercise 1

- Now we're going to group by multiple variables at once.
- Create a column that indicates whether the school is in the top half in terms of cohort size or the bottom half, using the median function.
- Concatenate that result with the LEA. You now have a variable which indicates both which LEA a school is in and whether they're a big or a small school, relative to the schools overall.
- Use a pivot table to group by this new variable to get graduate rates.
- Copy the results to get them out of the pivot table, and then manipulate this so you have an LEA column, a grad-rate-for-small-schools column, and a grad-rate-for-big-schools column.
- Create a new column for the difference between the two. Make some observations/use Excel tools (functions, graps) to describe this column.

Exercise 2

- Do the same thing, but with LEA type and cohort size quartile (use the quartile function to generate quartiles and if statements to connect those to each school's cohort size).

Additional Data Management Exercises

Finding Graduation Rate Information for Philadelphia Schools

- Using "2015-2016 Pennsylvania 4-Year Cohort Grad Rates.xlsx", determine the name of the Philadelphia LEA (Local Educational Agency) by sorting the "Total Cohort" column in the "Grad Rate by LEA" tab. (We expect the Philadelphia schools to have the highest number of students in the state, and this is the case.)
- When you've determined the name, go into the next tab, "Grad rate by school" and filter the data so you're only looking at Philadelphia schools.
- Let's figure out some information about Philadelphia schools. How many are there? Use "count".
- What's the average graduation rate by school? Use "AVERAGE".
- How does this differ from the overall graduation rate for the school system? (That is, are lower-graduating schools typically bigger or smaller?) Take the "SUM" of total cohort and the "SUM" of total grads and divide them.
- Duplicate this via a pivot table to check results.

Putting Graduation Rates in Bins and Making a Histogram/Bar Chart

- Binning, or grouping the graduation rates by ranges like 70%-75% and 75%-80% will let you make a bar chart to better see your data.
- We'll first go through how you could create a bin using "COUNTIF".
- We'll then do this more simply using the histogram functionality.
- Modify the formatting to make it more readable; we'll play with "add chart element" and "format cells".

Merging Graduation Rate Data with SAT Data

- Filter the "2016 SAT Scores Public Schools" data so that you're just looking at the Philadelphia data; copy it into another tab in the Grad Rate file.
- Use "match" with the School Number/School Nmbr. Field. Correct an error if you get one. Use absolute referencing on the "lookup array" argument, and get an exact match.
- Use "index" to get composite SAT data. Your "row number" term will be the result of your "match" function; again, use absolute referencing on your array.
- Use copy and "paste special" to get the values into that row rather than the formulas.

Relationship between Graduation Rate and SAT Scores

- Now that we've merged graduation rate data with SAT data, we can make a scatterplot showing these.
- Add a trendline and an r^2 value; play with the different model options and see if anything seems to describe the relationship well.
- Play with formatting – try adding a title, changing font, font size, axes titles, etc. Try a template and then modify that.

Using Macros

We're going to work through some of the following:

Meeting VBA macros in Excel:
https://msdn.microsoft.com/en-us/library/office/ee814735(v=office.14).aspx

Appending data from multiple workbooks:
https://msdn.microsoft.com/en-us/library/office/gg549168(v=office.14).aspx

Using Excel functions via VBA:
https://msdn.microsoft.com/en-us/library/office/hh211481(v=office.14).aspx

Additional Exercises

Simulate Coin Flips and Dice Rolls

Typing =rand() to a cell will generate a uniformly distributed random variable between 0 and 1. You can translate this into coin flip by using it as an argument in an IF statement that says the following: if the outcome of rand() is greater than .5, output “heads”, otherwise output “tails.” Try implementing this.

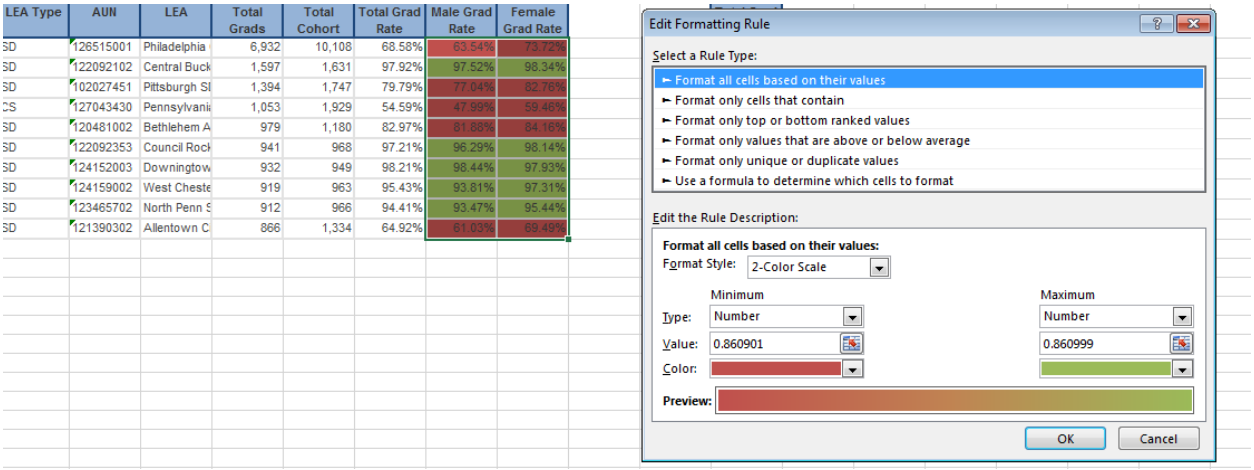
Now simulate 10 coin flips 10 different times to create a 10 x 10 table. Summarize your results using COUNTIF to get a 10 x 1 table showing the number of heads you got in each set of 10 coin flips.

Create bins for each integer between 0 and 10, and then either use the data analysis add-in to bin your data in a histogram, or use countif. Create a bar graph from the results showing the frequency of each result.

Simulate the outcome of a six-sided die. Play with the formatting and wording as well as with review>protect sheet so that the user a) only sees the instructions and the outcome and b) can’t change your formatting. (The easiest way to do this is to put your computations on a separate tab, right-click on that tab, and select “hide.”)

Use Conditional Formatting for Colors

Find the statewide overall graduation rate. Find the top 10 LEAs by cohort size and copy columns A-H into a new tab. Select the male and female graduation rate numbers and select home>conditional formatting>new rule. By setting it as follows, the values below the overall graduation rate should be green and those above it red:



LEA Type	AUN	LEA	Total Grads	Total Cohort	Total Grad Rate	Male Grad Rate	Female Grad Rate
SD	126515001	Philadelphia	6,932	10,108	68.58%	63.54%	73.72%
SD	122092102	Central Buck	1,597	1,631	97.92%	97.52%	98.34%
SD	102027451	Pittsburgh SI	1,394	1,747	79.79%	77.04%	82.76%
CS	127043430	Pennsylvania	1,053	1,929	54.59%	47.99%	59.46%
SD	120481002	Bethlehem A	979	1,180	82.97%	81.88%	84.16%
SD	122092353	Council Rock	941	968	97.21%	96.29%	98.14%
SD	124152003	Downingtow	932	949	98.21%	98.44%	97.93%
SD	124159002	West Cheste	919	963	95.43%	93.81%	97.31%
SD	123465702	North Penn S	912	966	94.41%	93.47%	95.44%
SD	121390302	Allentown C	866	1,334	64.92%	61.03%	69.49%

Now implement a three-color scheme. Take the top 30, determine what the total range of graduation rates in that group is, and divide that range into thirds. Play with the formatting rules so that the top third are green, middle are yellow, and bottom red. (Your Format Style will be 3-Color Scale.)

Description of Excel Tools

Formulas

You tell Excel you want to compute something in a cell by typing = . For instance =1+1 would output 2, = and B2/D2 would divide the content of cell B2 by the content in cell D2.

You can think of functions as a subset of these.

Functions
Functions are the building block of cleaning data in Excel. You can also do a significant amount of other data management and analysis via functions, although there are often easier ways to do it using the data analysis add-on or pivot tables. A function expects data to be in a particular format, or contain a particular type/number of arguments.

- **Logical functions**
  - "AND" lets you string together multiple statements; the output is "true" only if all of the statements are true.
  - "OR" lets you string together multiple statements; the output is "true" if any of the statements are true
  - "IF" takes a statement (which can be an AND or OR function) and outputs a value if it's true and another value if it's false.
- **Statistical functions**
  - "MEAN", "MEDIAN", "MIN", "MAX", "RANGE", "VAR", "COUNT", "SUM", "AVERAGE" give you that information about a set of numbers
- **Data management functions**
  - "CONCATENATE" pastes together fields
  - "INDEX" and "MATCH" let you compare two different columns to find matches, and then return information in columns which are adjacent to the match. This is how we will merge data sets.

Standard Data Management Tools
These are the kinds of data management tools you'd cover if you were handling data sets in SQL or another programming language/framework.

- **Filtering data**
  - *Filtering data means getting only the data with certain values; for instance, in PA school data, you might only want data from the Philadelphia school system.*
  - Data>filter is the easiest way to filter data.
  - You could also use logical operators in new columns or worksheets, in combination with "remove duplicates"
- **Sorting data**
  - *Sorting"data means putting it in the order you want, either numerically or alphabetically, ascending or descending*
  - Data>filter and data>sort both let you sort data
- **Summarizing data**
  - *Summarizing data means getting information about the values in your data. For instance, you might want to see the mean and variance of high school graduation rates.*
  - You can use statistical functions like mean, median, min, and max
  - You can build a pivot table
  - You can use the "data analysis" add-on
- **Merging data sets**
  - *Merging connects data sets which have a field in common; for instance, you might have two data sets with different information about the same set of schools, and you can join them based on the school name or ID.*
  - The "INDEX" and "MATCH" functions will let you do a "left join", where you keep all of the data in your initial (left) data set and add any matches in your second data set
- **Group by**
  - *Grouping by lets you aggregate data by the values in one or more columns. For instance, if you have a data set where each row represents a school, you wish to see the number of graduates by school system.*
  - You can group data by using pivot tables. If you want to group by more than one variable, you can concatenate (or paste together) the variables you are grouping by prior to creating the pivot table.
- **Split**
  - *Splitting lets you divide a cell or column into multiple cells/columns, defined either by the width of the cell/column or by a particular character. For instance, whenever it sees a comma, Excel can recognize it should start a new column.*
  - You can do this with data>text to columns
- **Drop duplicates**
  - *If you don't want duplicates – maybe you have a data set with every student in it but you just want a list of all the schools they go to, without repeats – you can drop duplicates.*
  - Drop duplicates via data>drop duplicates.

Miscellaneous Other Tools

- **Absolute and relative references**
  - When you refer to another cell in a formula/function, and then wish to copy this formatting to another cell, Excel will default to relative referencing, where it changes the cell references. You can use $s in your formula/function to preserve the formatting across cells.
- **Generating random numbers**
  - You can use RAND and RANDBETWEEN to generate random numbers
- **Search and replace**
  - You can use “Find & select” or the “SEARCH” and “REPLACE” functions
- **Format cells**
- Right-click on a cell, column or row, or worksheet to bring up “format cells”. From here you can change the formatting (font, size, etc.), merge cells, or change the data type (number of digits, text (string), number, percent, date.)