Appendix B. Spreadsheets 101

The modern spreadsheet is one of the best tools available for structuring information for use in visualizations. It can be successfully used as a method for simply storing the right item in its proper category, but a it also has the ability to transform and analyze the data in ways that can offer insight into underlying phenomenon. Spreadsheet do the latter by providing simple tools and functions for people to use to make the data dynamic and reactive to other pieces of information.

For example, a spreadsheet might be used to see if a person could afford purchasing a house by examining all the possible variable costs. The spreadsheet would automatically recalculate the total cost for the house if the price or interest changed. This makes it, easy for the user to test various combinations of rates to see what might be affordable. This dynamic ability for a spreadsheet to instantly react to item changes makes it such a powerful tool for the inquiry and structuring of data.

The most popular spreadsheet available is Microsoft's Excel, but other options include the freely available Open Office[32] and Google Docs[33], and the spreadsheet applications found in Apple's iWork and Microsoft's Works suites. All these tools work very much in the same general manner and are descendants from the groundbreaking *VisiCalc* program developed by Dan Bricklin and Bob Frankston in 1979, which inspired competitor *Lotus 1-2-3* to earn the title as the "killer application" for the nascent personal computer.

Modern spreadsheets include the capability to generate sophisticated charts and graphs that can facilitate a deeper understanding of the data through rudimentary visualization techniques. These spreadsheets can even do relatively sophisticated statistical analyses, such as correlations and regressions that can highlight trends between data sets, thereby eliminating the need for purchasing expensive and difficult-to-use statistical analysis applications such as SPSS.

"Cloud-based" spreadsheets such as Google Docs makes real-time collaboration on projects radically easier to manage. Because cloud-based applications store their files on a mutually accessible server in the Internet, any number of people can simultaneously edit data on a single file without engendering the inevitable confusion that develops with people working on multiple versions of the same project. Some tools, including In addition, VisualEyes, can treat access the online spreadsheet for use as a simple database of data for visualization.

[32] www.openoffice.org
[33] www.docs.google.com

Spreadsheet overview

A spreadsheet is like a simple table in a word processor. It keeps information organized in a grid-like collection of boxes, called *cells*. Each cell can be filled in with numbers or letters to organize information according to columns and rows. You can fill in a cell by clicking on it and typing directly into its box, or pasting text from your clipboard into the *formula bar* at the top of the page/screen.

	A	B
1	**name**	**age**
2	manny	40
3	moe	50
4	jack	60

At the top, the *columns* are labeled with letters from A to Z, and the *rows* labeled on the left with numbers. The combination of a column and row will uniquely point to a particular cell and forms its *address*, just like a real world street address. For example, the address **B2** refers to the 2rd row in the B column, which in this case has the number *40* in it.

This ability to refer to cells by their position in the grid (called *relative addressing*) is a little abstract to grasp at first, but makes the spreadsheet a very powerful tool for structuring information because we can easily apply operations to cells based on formulas.

Formulas

fx	=1+2		
	A	B	C
1	2		
2	20		

Formulas are similar to simple arithmetic statements, like 1+2=3 or 4*5=20. The formula itself appears in the *formula bar*, a box above the cell grid, and the results of a formula show up in the cell itself. The spreadsheet knows it's a formula, rather than text or a number because it starts with an equals sign. All the usual arithmetic operations can be used in formulas: add (+), subtract (-), multiply (*), and divide (/). Parenthesis can be used to group the order of these operations properly. In addition to these basic math operations, there are literally hundreds of specialized functions (described below) that can perform more sophisticated operations like averaging and statistical operations.

fx	=B1+C1		
	A	B	C
1	3	1	2
2	20	4	5

Formulas are more useful when they refer to the contents of other cells rather than specific numbers. This example returns the same results in the **A1** cell as the previous example, but gets the number 1 from the contents of cell **B1** and the number 2 from the contents of cell B2 by using the formula **=B1+C1** instead of having to specify the numbers explicitly.

Suppose we wanted to compare the weights of two groups of people, but the French group was in kilos and the English were in pounds. We would need to put them in one system or another. We could convert the English pounds into kilos by multiplying each entry by 2.2 using a calculator, but that would be time consuming. The ability of the spreadsheet to do this simple calculation simplifies our task by adding a new column that references the kilos cell and multiplies each weight by 2.2 using a simple formula.

Spreadsheet formulas begin with an equals sign, like this: **=A2*2.2**, which tells the spreadsheet to look at the number in the cell located at **A2** (*52* in this case), multiply it by 2.2 and write the result in the grid of 114 pounds. The formula itself shows up in the text box at the top of the grid, but the result appears in the cell. We only need to write the formula once, and then copy and paste it to the other cells we want to convert. The spreadsheet will automatically advance the address as appropriate (i.e. A3, A4, A5 etc.).

fx	=A2*2.2		
	A	B	C
1	kgs	lbs	
2	52	114	
3	60	132	
4	63	139	

Absolute cell addresses

The feature of automatically advancing the cell addresses is very convenient when the multiplying by a constant value, but there are times when we want to control all the calculations from single cell. Putting a dollar sign before the column and/or row indicator will prevent the spreadsheet app from advancing the column letter or the row number when copy and pasting cells.

fx	=A2*B$5		
	A	B	C
1	**price**	**cost**	
2	50,000	2,000	
3	60,000	2,400	
4	70,000	2,800	
5	**interest**	4%	

For example, when deciding whether to purchase house of a given price, the rate of interest will be an important deciding point. Isolating the interest rate in its own cell makes it easy to see the effect on the cost by simply changing the contents of cell **B5**, and then the cost column will be recalculated instantly to reflect the new interest rate's effect.

Naming cells

As an alternative to absolute addressing of cells, most spreadsheets let you name a cell with a unique name, such as "InterestRate." Instead referring to the cell by its address, you associate that name to the cell. This feature is found in the *Name* item in Excel's *Insert* menu and in the *Named Ranges* item from Google Docs' *Edit* menu. Once defined, you can use that name as a variable in fomulas: **=A2*B$5** becomes a more understandable **=A2*InterestRate**.

Cell Ranges

A cell range is similar to a regular cell address but refers to a block of cells rather than just one. They are specified by the top-left and bottom right corners separated by a colon. The range **B2:C3** refers to the cells located at **B2, B3, C2, and C3**. Ranges are used in functions such as those that average a group of cells.

	A	B	C	D
1	**name**	**age**	**sex**	**grade**
2	Bob	22	M	100
3	Ted	50	M	80
4	Carol	34	F	45
5	Alice	43	F	100

Lists

A spreadsheet can simply contain a series of items, arranged horizontally or vertically, but more commonly, they are arranged in a *list* format. This format is useful in its own right, as a clear way to organize information for later representation in a visualization, and most spreadsheet applications need the data in the list format for graphing and other advanced functions.

	A	B	C
1	**name**	**age**	**sex**
2	Bob	22	M
3	Ted	50	M
4	Carol	34	F
5	Alice	43	F

In a list, the vertical columns are used to group like things together (i.e. attributes), and each horizontal row is dedicated to enumerating those attributes for a particular entity. The top row, called the *header*, is usually dedicated to a list of names that describes the attributes, called *fields*. In this example, the rows contain individual people and the columns contain attributes about that person (their name, age and sex) under the appropriate field names.

Freezing of columns/rows

In larger data sets, it is useful to always know what field name you are currently viewing once the top header line that defines them has scrolled out of view. You can "freeze" the rows or columns at some point, so they do not scroll with the rest of the cell matrix to provide a constant reminder of their meanings. In Excel, this is done by the cell whose column and row you wish to retain the selecting the *Freeze Panes* item in the *Window* menu. In Google Docs, this is accomplished by selecting the *Freeze Rows* or *Freeze Columns* item in the *Tools* menu.

Paste special

When you copy a cell to the clipboard, for later inclusion into another cell, Excel has to make some hard choices: Does it need to copy the formula, or the end result of that formula. For example, if a cell contained **=A1+B1**, there may be times when we want to duplicate that formula for use in calculating different cells, and other times when we want to copy the number value generated when **A1** and **B1** were added.

By default, Excel copies the formula, but if we want the value, Excel has a very useful item called *Paste Special* in the *Edit* menu that offers some options to the way cells that have be copied into the clipboard can be pasted into new cells. This brings up a dialog box with some options under *Paste*, including one called *Values* which will only copy the results of the formula's calculation and not the formula itself.

Transposing columns and rows

Sometimes the cells are arranged in the right direction for the immediate task. We may need the cells to run horizontally instead of vertically, or vice-versa. The *Paste Special* dialog box has a checkbox that facilitates this.

Formatting cells

Most spreadsheet application offer a number of ways to change the appearance of the cells, in terms of changing the text font, size, and color, and the cell's internal and external color. From the data's point of view when used in a visualization, this graphical formatting is largely ignored as window dressing, but in more complex sets of data, the ability to graphically group items by color can shed insight into the nature of the data, and make it easier to navigate through larger datasets.

Cell data formats

The native value of a cell falls in three primary categories: *Text* which represents letters and numbers as a stream of characters; *Numbers* which represents whole and fractional numbers; and *Dates* which reflect the idea of time. Within these three basic categories, spreadsheets do offer some ways to format the data that makes it more understandable, such as adding comma to separate thousands in large numbers. This formatting does not change the underlying value of the cell, just how it is displayed.

For example, *.045* and *4.5%* represent the same number and mathematically, but the latter can more understandably express value to people without sacrificing accuracy. Likewise, consciously limiting the display of the number of decimal places (1.73346643 to 1.73) can help remove the appearance of complexity and increase understanding while again not sacrificing the accuracy of the base value. Likewise, the Number and Currency formats which add 1,000's commas and the $ symbols help make clarify the data.

Forcing a number as text

By default, Excel formats cells in a format called *General*, which is a hybrid of the text and number formats. If you type a number in a General cell, it will be treated as a number for calculation purposes, and if you type text, it will treat it as text.

There are some circumstances where you actually want to treat a number as text. If you were storing an archive record number such as *000623.0*, the spreadsheet would assume you wanted the value of that number and simplify it to *623.* You can tell the spreadsheet to treat the cell as a text value by setting its format to Text, or preceding the number with an apostrophe (i.e. '*000623.0*) to force it to be considered as text. One visual cue that spreadsheets offer as to the formatting is the alignment: Numbers are right-justified and text is left justified.

Dates

Spreadsheets handle dates in a more complicated manner, so some of the results from formulas and calculations may come out differently than expected. There are numerous ways to store dates (i.e. 1800, 5/1800, 5/2/1800, May 2, 1800, etc.) but they all reflect the same moment in time. Most spreadsheets convert that complex representation to a simpler native format when doing calculations on dates, by storing dates internally as the number of days (plus or minus) from January 1, 1900.

This allows for the proper sorting of dates, with *11/2/1800* showing up after *4/2/1800* instead of before if it were alphabetically sorted, and arithmetic operations such as finding the number of days between two dates possible. For example, the formula **=11/2/1800 - 4/2/1800** would result in 214 days (7 months) to be returned to the cell. There are a large number of built-in functions that make it easy to tease out the days, weeks, months, and years from dates.

Sorting

Sorting the data can be an excellent way to get some insight into the underlying phenomena it is representing. For example, if we wanted to get a quick idea of how many times the lunch menu had Tacos on Tuesday, we would sort the list by the *weekday* field, followed by a secondary sort by the *entree* field. This would cause all the rows in the list to first be group by the day, so all the Mondays would be grouped together, all the Tuesdays, etc. The second sort would then sort the entrees alphabetically within the day-wise groupings to make a clump of Taco entries stand out from the rest.

Clicking on the *Sort* item in the *Data* menu in Excel (in the *Tools* menu from Google Docs) will bring up a dialog box with the options to sort the list by. You can sort the data up or down by column, and add then further sort the information based on the order in other columns. The sort can be alphabetically ascending (i.e. A-Z) or descending. Because of the sort is alphabetical, dates and numbers can sometimes be incorrectly sorted. If your data has a header row, be sure to click on the button so that row does not get sorted with the rest of the data.

Formula functions

Using the hundreds of built-in *functions* in modern spreadsheet, not just simple arithmetic that can be performed between cells, but a wide variety of function ranging from simply adding a series of numbers together to sophisticated statistical and financial operations. Each function has a number of parameters within its parenthesis and returns the results of their calculation back to the cell, or as a parameter to another function.

The number and kind of parameters for each function is dependent upon that particular function. In Excel, all of the available functions are listed by their type by clicking on the ***fx*** button to the left of the formula bar, grouped by category. Choosing one will bring up its description and how to use it. Google Docs has many of the same functions available by clicking on the *Function* item in the *Insert* menu.

fx	=SUM(B2:B4)	
	A	B
1	**name**	**weight**
2	John	150
3	Paul	180
4	George	140
5	**Total**	470

For example, We could use the **SUM** function to add up a column of numbers by using the formula **=SUM(B2:B4)** and place the result in the cell. The parameter to the SUM function is the range of cells to add up. The range of cells to add was specified by the range B2:B4 to include B2, B3, and B4. The result from the **SUM** function could be just as well used in another operation, say to create an average: **=SUM(B2:B4)/3**, although there is an **AVERAGE** function built-in for this common operation.

Commonly used functions

The following functions are commonly used in visualizations and grouped by the general category they appear in. There are literally hundreds of functions available in most spreadsheets, so it is worthwhile to look at the help screens to see if a more appropriate one is available than the one listed below:

Math

- **ABS** (number) - Returns the absolute value (non-negative) of a number,
- **AVERAGE** (range) - Returns the average (mean) of a range of numbers.
- **CEILING** (number, 1) - Return the rounded-up integral part of a number.
- **FLOOR** (number, 1) - Return the lower integral part of a number.
- **MAX** (num1, num2, ...) - Return the largest of 2 or more numbers.
- **MIN** (num1, num2, ...) - Return the smallest of 2 or more numbers.
- **SUM** (range) - Returns the total of a range of numbers.

Date and Time

- **DATE** (year, month, day) - Returns the date based on year, month and date.
- **DAY** (date) - Returns the day number (1-31) from a date.
- **HOUR** (date) - Returns the hour number (0-23) from a date.
- **MINUTE** (date) - Returns the minute number (0-59) from a date.
- **MONTH** (date) - Returns the month number (1-12) from a date.
- **WEEKDAY** (date) - Returns the day text (Sunday-Saturday) from a date.
- **YEAR** (date) - Returns the year number (1900-9999) from a date.

Text

- **CLEAN** (text) - Returns the text cleaned of any non-printable characters.
- **CONCATONATE** (text1,text2, ...) - Returns the union of 2 or more strings.
- **LOWER** (text) - Returns the text in all lower-case letters.
- **PROPER** (text) - Returns the text in title case (1st letter of words in caps).
- **TRIM** (text) - Returns the with extraneous spaces removed.
- **UPPER** (text) - Returns the text in all upper-case letters.

Lookup

- **HLOOKUP** (value, range, index) - Returns the matching value for value found in a horizontal look up table (see section below on look-ups).
- **VLOOKUP** (value, range, index) - Returns the matching value for value found in a vertical look up table (see section below on look-ups).

Statistics

- **FTEST** (range1, range2) - Returns probably of 2 means being the same.
- **AVERAGE** (range) - Returns the mean of a range of numbers.
- **MEDIAN** (range) - Returns the median of a range of numbers.
- **MODE** (range) - Returns the mode of a range of numbers.
- **PEARSON** (range1, range2) - Returns the correlation between 2 ranges.
- **STDEV** (range) - Returns the standard deviation of a range of numbers.
- **TTEST** (r1, r2, tails, type) - Returns probably of 2 means being the same.
- **VAR** (range) - Returns the variance of a range of numbers.

Graphing

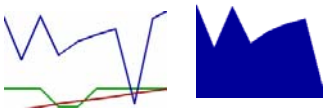
The ability to visually represent data is as valuable to the designers of interactive visualizations as it is to its consumers. Spreadsheets make it easy to quickly create charts and graphs to visually explore the data relationships. This can lead to insights into the underlying phenomenon that can direct the kind of visualizations that will be most effective.

Excel and Google Docs have excellent charting capabilities and work in a similar manner. You select the table you want to chart by highlighting the cells that define it and click on the *Chart* icon in the tool bar. A dialog box will offer a number of chart style and you can instantly see a preview of that style using your data.

Choosing the right kind of graph

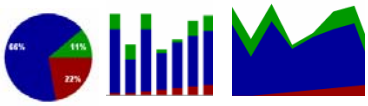
The various styles are useful in exploring different kinds of relationships between the data:

- **Time-Series** relationships, where the values of data are plotted vertically as time marches across are most fruitfully rendered by line and area styles.

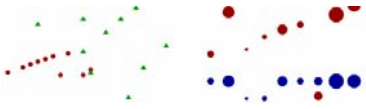


- **Quantitative** relationships between items in a data set are best drawn using bar, area and line charts

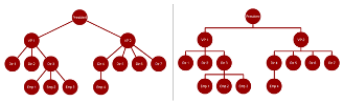
- **Part-to-whole** relationships, where the relative value of one item is compared with the group is well represented by pie, stacked bar/area charts.



- **Correlation** relationships, where a number data points are plotted using two variables are best drawn using a scatter, and bubble charts.



- **Hierarchical/Organizational** relationships between individual members can yield insight using organization maps, network diagrams, and trees.



Trend lines

A trend line is a line overlaid upon chart that "smoothes-out" the data an give you an overall sense of the trend the data is taking, and is useful in seeing whether in general, the values are increasing or decreasing over time. For more information, see the section on *Trends* in *Time series data* on page 37. In Excel, you click on the *Add trendline* option in the *Chart* menu.

Importing data

While it is certainly possible manually type in data into a spreadsheet, most people choose to import it from some existing source, which can be a text file, or a structured table from a database or website page.

Data from text files

The goal of putting data in a spreadsheet is to organize that data in some meaningful way using the fields have been defined. Unfortunately, many primary sources of data come from largely unstructured sources, such as letters, documents and other largely prosaic styles of organization.

It can be useful to use a word-processor, such as Microsoft Word or Google Docs to prepare the text prior to import into the more rigid constraints of a spreadsheet by creatively using the *Find and Replace* functionality to rearrange the data into the columns and rows required. The text can be easily reformatted so that it will slot into the proper columns and rows when pasted into the spreadsheet by making sure each row of data is listed on its own line, with tabs between each field.

To aid in this process, Word has some special character that allow you to search for and insert characters like new lines (^**p**) and tabs (^**t**). The *Special* button, exposed when the *More* options button is active will show all the available characters. For example, the following find and replace combination would look through the text for any double-line feeds followed by a Title: as the delimiter to a record and replace it with a simple line feed: Find what: *Title:^p^p* Replace with: *^p*

If the field values are separated by commas or some other separator, you will need to put tabs so the spreadsheet recognizes them as distinct fields with a following combination like this: Find what: *,* Replace with: *^t*

Finally, Excel has an option in the *Data* menu called *Text to Columns* that will parse some unformatted text in a cell or range of cells, and distribute the contents into columns based on one or more defined delineators, such as a tab, comma, etc.

Data from web pages

Existing websites are a great source of data for visualization projects, Many websites, like the US Census[34], OECD[35], and ManyEyes[36] make it easy to directly download data files in one or more format for editing within a spreadsheet. Some of these sites offer the data in Excel's native format, (.XLS) but comma-separated value (.CSV) and tab-delineated (.TXT) formats are more common. Most spreadsheets and Google Docs can load these formats easily.

Screen scraping table data

It is still possible to capture the data from web pages when no download option is thoughtfully provided by the page's author, through a process sometimes called *screen scraping*. Scraping is the process of pulling data directly from a fully formatted web page on the Internet and extracting the raw information for use in a spreadsheet

Modern web browsers make it easy to copy a table of data directly from a web page using the familiar copy and paste functionality. Once you have found a table of data you want, select the entire table and copy (CTRL-C key or the *Copy* item in the *Edit* menu) it into your computer's clipboard. Paste (CTRL-V) this data in an open Excel spreadsheet. You will probably need to clean up any extraneous bits and pieces, and most likely need to reset the formatting carried over from the web page to match your spreadsheet's formatting.

The Windows versions of Excel have a table scraping option built in that loads a web page in a dialog box, automatically identifies the data tables on the page and allows you to select one. A number of useful options are available that help import the data in the format desired and is found in the *Data* menu, in the *Get External Data / New Web Quer*y option.

[34] www.census.gov
[35] www.oecd.org/statsportal
[36] www-958.ibm.com/software/data/cognos/manyeyes

Look Up Tables

A lookup table is a method to use another list to provide one or more items of information for another. An example of this would be a price list's relationship to a catalog. We may want to keep all the prices in a single list, and look them up based on the item's description. To use it, our spreadsheet would contain both lists like the spreadsheet on the right. The first two columns are the catalog (tinted blue), and the last two (in green) are the price list. The formula for the **B** column uses the **VLOOKUP** function to look through the price list (located at **D2:E5**) for the item in the **A** column and place the match for from the **E** column.

fx	=VLOOKUP(A2,D$2:E$5,2)				
	A	B	C	D	E
1	**item**	**cost**		**item**	**price**
2	shoe	$40.99		pants	35.99
3	sock	$12.99		shirt	24.95
4	pants	$35.98		shoe	40.99
5	shirt	$24.95		sock	12.99
6					

The actual formula looks complicated but is simple once it is broken down by the three parameters required, the *value*, *range*, and *index*:

=VLOOKUP(A2, D$2:E$5, 2)

A2	The cell of the item to lookup (in this case "shoe")
D	Starting column of lookup list's range
$	Hold the row number constant when we copy and paste
2	Starting row of lookup list's range
:	Range separator
E	Ending column of lookup list's range
$	Hold the row number constant when we copy and paste
2	Ending row of lookup list range
2	Which column in lookup list to return (the 2nd with the price)

The **HLOOKUP** function works in very much the same way, but the lookup occurs horizontally instead of vertically and the index parameter refers to row rather than the column.

Filters

Excel has a feature where you can search through a data list and extract only the rows that meet certain criteria. This is useful from both the data structuring and data exploration perspectives. In the former, you can use filtering to create smaller subsets from larger ones, for example make a list that only contains males between the ages of 18 and 30. Secondly, filtering is useful in the exploratory phase to graph or view selected subsets within categories.

The implementation of filtering in Excel is as inelegant as it is powerful, with unfortunately very few supports to guide the process. The basic idea is that some criteria is applied against a list of data and any rows that meet that criteria are copied to a new list in the spreadsheet. To do this three cell ranges that need to be identified:

	A	B	C	D
1	**item**	**cost**	**size**	
2	shoe	$40.99	S	cost
3	sock	$12.99	M	>30
4	pants	$35.98	M	
5	shirt	$24.95	S	
6				
8	**item**	**cost**	**size**	
9	shoe	$40.99	S	
9	pants	$35.98	M	

1. The *list range* that will be filtered from. This is a typical Excel list, where the top row defines the fields and each row below it contains the data (shown in blue).

2. The *criteria range* specifies the what criteria needs to be met before being included in the rows that are selected (shown in green). In this case, any row whose cost is greater than $30 will be chosen.

3. The *extract range* is where the new rows that met the demands of the criteria will be copied to (shown in yellow).

Filtering in Excel

Clicking on the *Advanced Filter* option in the *Data* menu will bring up a dialog box where you tell Excel where on the spreadsheet's grid those three areas are located. The *List range* button will select the list to be filtered from, and the entire list, including the header row should be selected.

The *Criteria range* button is used to select the cells that define the criteria for inclusion in the new list. The criteria contains pairs of cells, with the top-most one specifying the field being examined (cost in this case) and the criteria below it (>30). Possible operators include less than, less than or equal, equal, not equal, greater than, and greater than or equal (<, <=, =, <>, >, >=).

While it is possible to filter the results in place, typically you want to create a second list to hold the results, so make sure the *Copy to another location* radio button is checked and a single cell selected using the *Copy to* button. Clicking the OK button will copy any rows that meet the criteria to that new range

Advanced filtering criteria

The example above is the simplest case of using a filter to select a subset of rows from a larger data set, but Excel's filtering can create complex queries using multiple sets of criteria, more complex formulas, and wildcard searches to artfully tease out intricate relationships from your data.

Multiple criteria cells

The example used only one pair of cells to define the criteria for inclusion, the rows where the cost field was over $30, but we define additional criteria to further refine or expand the rows selected. The criteria list can contain multiple columns, each one containing a field to search on.

cost	size
>30	M

If we wanted to search for items that cost over $40 *and* were size M, the criteria range to the left would search the data list and select only rows that met both criteria, in this case only *pants*. Because the criteria were on the same rows, both conditions needed to be met to be included.

cost	size
>30	
	M

If we wanted to search for items that cost over $40 *or* were size M, the criteria range to the right would search the data list and select only rows that met either one of the individual criteria, in this case *pants, shoes, and socks*. Because the criteria cells are on the different rows, either conditions needed to be met to be included.

Complex formulas

Almost any of the formulas and functions available to normal spreadsheet cells can be used in evaluating the criteria for a field, as long as the ultimate result of the cell returns a true/ false answer. For example, **=ISNUMBER(A2)** would return *true* and cause that row to be included in the selected rows if the value found in cell **A2** was a number. The formula **=A2>A7** would include only rows when the value of cell **A2** was greater than the value in the cell at **A7**.

Wildcards

Finally, you can use a number of *wildcards* when specifying text to search on. *Wildcards* are character that tell Excel that you want to broaden the acceptable value to find beyond what you have specified. For example, if we specified *shoes* as a search term, we would not get any results for *shoe*, since it is not an exact match. If we specify *shoe?*, any character that comes after shoe would be included (i.e. *shoe, shoes, shoe1*, but not *shoe12*). Specifying *show** makes the search broader, allowing any number of letters following shoe will be accepted (i.e. *shoe, shoes, shoe12, shoelace*). Wildcards can be placed in any spot of the search term (i.e. *shoe*s*, would return *shoes* and *shoelaces*, but not *shoetip*)

Statistics using spreadsheets

There are two basic kinds of statistics that can be applied to a data set; *descriptive statistics* and inferential statistics. Both Excel and Google Docs have a large number of these sophisticated techniques available as functions for inclusion into cell formulas. See the appendix section on statistics on page 94 for more information about using statistics.

Excel goes a step further than Google Docs (at least for now) by providing a menu-driven mechanism, very similar to statistical packages such as SPSS, that make it very easy to add and view the results of the most commonly used statistical tests, such as correlations, regressions, descriptives, and T-Tests. You may need to install the *Data Analysis add-in* (included on the Excel CD, but not typically installed by default).

Descriptive statistics

As the name implies, *descriptive statistics* provide simple summaries to begin to understand the nature of the data, through the average value (the *mean*), the middle-most value (the *median*), the common-most value (the *mode*), the largest and smallest values (the *range or min / max*), and the variance of from the mean (the *standard deviation*).

	A	B
1	**name**	**age**
2	Bob	34
3	Ted	50
4	Carol	34
5	Alice	43
6	Larry	53
7		
8	Mean	42.8
9	Median	43
10	Mode	34
11	Min	34
12	Max	53
13	SD	8.8

Using formulas for descriptive statistics

The following formulas appear in cells **B7** to **B12** to display the basic descriptive statistics for the ages in **B2** to **B5**:

- **=AVERAGE**(B2:B6) - The mean average age is 40.3
- **=MEDIAN**(B2:B6) - The middle-most person, age 43.
- **=MODE**(B2:B6) - Returns most common age of 34.
- **=MIN**(B2:B6) - Returns youngest person, at 34.
- **=MAX**(B2:B6) - Returns oldest person, Larry at age 53
- **=STDEV**(B2:B6) - Returns the standard deviation of ages of 8.81 years.

From this limited set of data we know and ages range from 34 to 53, the average age is about 40 and from the standard deviation, that 68% of the people in a set like this will be aged approximately between 31 and 40 (i.e. +/- 8.8 years from the average).

Descriptive statistics using Excel's Data Analysis

Clicking on the *Data Analysis* option in the *Tools* menu will being up a dialog box, listing *Descriptive Statistics* as one of its Tools. The *input range* button selects the cell range in the data to analyze. Other checkboxes add additional descriptive statistics to the output, such as confidence levels. The *output range* button specified the cells that the output will be written to. The table to the right contains just the description statistic basics for the age data set used by the formula method of generating them. See the appendix on statistics starting on page 94 for more information about using these measures.

Mean	42.8
Standard Error	3.942080669
Median	43
Mode	34
Standard Deviation	8.814760348
Sample Variance	77.7
Kurtosis	-2.71915048
Skewness	0.043509564
Range	19
Minimum	34
Maximum	53
Sum	214
Count	5

Inferential statistics

Inferential statistics build on the information provided in descriptive statistics and enable a series of tests that attempt to infer whether a two samples are from the same general population. There are a large number statistical functions available in both Excel and Google Docs, but only correlation and T-Test are listed below to show the basic procedure. Again, look the on statistics appendix section for more information about using these tests.

	A	B	C
1	**name**	**age**	**score**
2	Bob	34	80
3	Ted	50	70
4	Carol	34	90
5	Alice	43	56
6	Larry	53	76
7			
8	T-Test	.002	
9	Pearson	-.46	

Using formulas for descriptive statistics

- **=TTEST**(B2:B6,C2:C6,2,1) - The probability that the age and score are samples from the same data is .02 using a 2-tailed *T-Test*.
- **=PEARSON**(B2:B6,C2:C6) - The relationship between age and score is a negatively correlated by -.46 using a *Pearson Product Moment* test.

Descriptive statistics using Excel's Data Analysis

Clicking on the *Data Analysis* option in the *Tools* menu will being up a dialog box, listing the various tests that can be performed. Each test has its own options , but all will require defining the *input range* to selects the cell range(s) of the data to analyze, and the output range where the output will be written to.