Purpose: The Xerox Star computer... it was really a rock star of a computer back in its day (the 1980's), but it didn't get much respect due to its price tag and that it lacked a key piece of software... the spreadsheet. Last week we introduced you to the spreadsheet and Excel, and this week we're going to give you even more Excel fun.

By completing this lab, you should feel more comfortable using spreadsheets in your daily lives.

Specific concepts we're going to examine throughout this lab include these advanced Excel features:

- Conditional formatting
- COUNTIF function
- Sorting
- Scatterplot
- Linear regression (inserting a trend line)
- More experience using formulas, relative cell references, and absolute cell references

Part 1a: Using Excel to visualize class attendance data

Many students ask why we have an attendance policy, and one good answer is that attendance is important to succeeding in college. We've seen from experience that making attendance mandatory boosts the grades and contributes to the overall success in our class and labs. In this part of the lab, we will take a look at some real attendance data taken from previous semesters, and you’ll get the chance to calculate attendance grades for a large class. To start, go to the labs page, and download the attendance.xls spreadsheet to your I: drive.

Open up attendance.xls from your I: drive, and take a look at it. If you see a message at the top part of Excel with a button to Enable Editing, then go ahead and click Enable Editing.

The spreadsheet attendance.xls has two worksheets: attendance_checks and attendance_grades. Go ahead and click on the attendance_checks worksheet if it isn't already selected. The worksheet contains 14 columns of attendance checks (check1, check2, ..., check14) for 140 students that belong to various groups in the class. The attendance checks were taken at random dates throughout a previous semester, and a student was recorded as either being absent or present for the attendance check. There are two columns at the end of the worksheet that we
will fill in at the end of this part of the lab, and they are **Total Days Attended** and **Attendance Grade**.

At first glance, this data seems overwhelming, but we will use some cool features in excel to visualize the attendance data since looking at a bunch of “present” and “absent” values doesn’t really help us see the overall trend on who’s attending class regularly and who isn’t. The first tool we will use is conditional formatting to give us a great visualization to make the absent students stick out.

Click on the Home tab, and select all of the attendance check cells, cells C2 through P141. Once all of the attendance check cells have been selected, click on **Conditional Formatting**, choose **Highlight Cells Rules**, and then click on **Equal To**.

In the **Equal To** window that pops up, type in absent in the “**Format cells that are EQUAL TO**” box, and then select **Custom Format** from the drop down menu. A Format Cells window will pop up, click on the **Fill** tab, and then choose a **solid red color** to highlight those absent students so they will pop out with conditional formatting. Click ok in all of the opened windows, and take a look at the attendance values in the spreadsheet. Now with a simple glace, we easily see which students have an attendance problem, and which students don’t. This is the power of conditional formatting, it will allow you format a cell, in our case fill it with a red color, based on some rule such as equal to, less than, greater than, and a various other options. Conditional formatting is a great visualization tool for dealing with categorical data like attendance data.

Now that we have a good visualization of our data, go through the spreadsheet and find some students who have attendance problems. Also, note that two attendance checks seem have a high attendance rate (not many students were absent). Which two attendance checks have a high attendance rate based on just looking at the colored spreadsheet? Think for a moment on why two attendance checks have a high attendance rate. The reason for the high attendance rates for Check3 and Check10 are due to the fact that those attendance checks were during class exams, and attendance rates are quite high for class exams.

Let’s move on to calculating the **Total Days Attended**. All we have to do is count how many times a student was present to get this value, but we don’t want to do this by hand or we’ll be here all day. Luckily, Excel has a handy function that will count things that aren’t necessarily
numeric. Excel has a function called **COUNTIF** that we are going to use to find the total days attended. The COUNTIF function counts the number of cells within a range that meet the given criteria. Our criteria is that we want to count the cells that contain present.

Go ahead and click on cell Q2, and hit the insert function button [this can be found under the formulas tab. In the window that pops up, search for the COUNTIF function, select it, and click Ok. For the Function Arguments, input or select the range C2:P2 (the attendance values for that student), and the criteria is “present”. What this will do is count each cell that contains the word present (note that Excel will accept present with or without the double quotes surrounding it). Hit Ok, and you should see 13 in Q2. Go ahead and copy or use the fill handle to compute **Total Days Attended** for the rest of the students in the class.

![Countif Function](image)

We also need to calculate the **Attendance Grade**, which is easy since we now know each student’s total days attended. The attendance grade is calculated by taking the total days attended and dividing it by the number of attendance checks (which is 14). So go ahead put this formula (=q2/14) into R2, and you should get 0.92857. Copy or use the fill handle to compute the **Attendance Grades** for the rest of the students in the class. Since percentages are nicer to look at than decimal points, go ahead and change the attendance grade values to percentages that show 2 decimal digits of accuracy (hint look in the format dropdown). Check your results with the results pasted below for the first five rows. If they match, then congratulations, you’ve done the work of TA to compute attendance grades for a large class like CSCI 1100.

<table>
<thead>
<tr>
<th>Total Days Attended</th>
<th>Attendance Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>92.86%</td>
</tr>
<tr>
<td>12</td>
<td>85.71%</td>
</tr>
<tr>
<td>9</td>
<td>64.29%</td>
</tr>
<tr>
<td>10</td>
<td>71.43%</td>
</tr>
<tr>
<td>12</td>
<td>85.71%</td>
</tr>
</tbody>
</table>
There is one more thing to note, we can also sort our data to see if there are attendance trends in groups. Select the cells then Click on the Home tab, if you aren’t there already, then click on Sort & Filter, and then choose Custom Sort. In the Sort window that pops up, click Add Level if you don’t see a Sort by under the Column section, then choose Sort by Group, Sort On Values, Order Smallest to Largest. This will sort your data based on the group number instead of the student’s number.

You can sort by any column, and you can have multiple levels to your sort. Experiment with various sort options and Adding and Deleting sort levels, and take a look at your spreadsheet after each sort. Go ahead and sort by Attendance Grades to easily pick out the students who had perfect attendance, and which student had the lowest attendance. Keep in mind that by using Conditional Formatting, COUNTIF, and Sorting, professors and TAs can easily keep up trends in attendance pretty easily, and this not only applies to attendance data, but lots of real world data too. Save your work to your I: drive, and if a Compatibility Checker window pops up just uncheck the box and click Continue to save.
Part 1b: Using Excel to do some simple statistics

Earlier we stated that we’ve noticed that having a mandatory attendance policy leads to better grades. Now, let’s take a look at some data. In the attendance.xls file that you’ve been working on, click on the worksheet called attendance_grades. This worksheet contains students, their group, their attendance grades (what we calculated earlier), and their class grades. What we want to do is investigate whether or not attendance is correlated (or associated) with higher class grades. We will use two Excel tools to help us out: scatterplot and trend line (linear regression). Don’t worry if you haven’t used a scatterplot or trend line before, because we’ll discuss them in this lab.

A scatterplot (or scatter chart in Excel) is a useful way to plot an independent variable on the x-axis (horizontal axis) and a dependent variable on the y-axis. You’ve seen these types of plots in math, where you plot x against y in order to see a relationship. Here, we would like to plot the attendance grades and class grades with a scatterplot. Click on the Insert tab. Select the range C2 through D141, the attendance and class grades. Once they are selected, click on the Scatter chart symbol then click Scatter with only Markers. You should see a scatterplot appear somewhere in the spreadsheet that looks like the one below.

From the previous lab, we covered how to label horizontal and vertical axes, and how to add a title above the chart. So go ahead and label the horizontal axis as attendance grades, and the vertical axis as class grades, and the title of the chart should be something like Attendance and Class Grades. It is important to note that you should never have a plot that doesn’t have its axes labeled properly or doesn’t have a title. The legend that contains only Series1 doesn’t really provide us with useful information, so go ahead and delete it. Your chart should look like the one below. If you can’t see the legend then click on the sign that appears just beside the chart when you click on the chart area. Among the list given select the checkbox corresponding to legend. That’s it!
From your scatterplot, you can see an upward trend in the data. The higher an attendance grade generally corresponds to a higher class grade. Now, let’s add a trend line to the data. Click on your scatterplot, and click the sign beside the chart. This should give you a list of elements that you want to add to the chart. Amongst them, choose Trendline. This should add a linear trend line to your scatterplot.

Notice that trend line has a positive slope, which means that as attendance grades increase, the class grades also tend to increase. The trend line is known in statistics as linear best fit regression line. It is the line that comes closest to all of the data points. Right click the trend line in the scatterplot to bring up its context menu, and then click Format Trendline. Check the boxes for Display Equation on chart and display R-squared value on chart. Click Close, and you’ll see the equation for the linear regression line, and its R² value. You can move this information somewhere on your scatterplot so it doesn’t block your data points. Your scatterplot should resemble the one below.
Wow, by using Excel we just did some powerful statistics. We found the linear regression equation $y = 0.5565x + 0.3715$, and we found its corresponding $R^2$ value. Here $y$ stands for our class grades, and $x$ stands for our attendance grades. Now, we’ve all seen the equation for a line, and our linear regression equation is telling us that we have a positive slope of 0.5565, and an intercept of 0.3715. The intercept isn’t that interesting, but the slope of 0.5565 shows a moderate relationship between attendance grades and class grades: as attendance grades increase the class grades tend to increase, and as attendance grades decrease the class grades tend to decrease. Thus, we’ve shown some statistical justification that higher attendance grades tend to be associated with higher class grades. Therefore, one reason for a mandatory attendance policy is to boost class grades.

You may be wondering about that $R^2$ value of 0.6148. $R^2$ gives us a measurement on how well our linear regression equation fits our data. $R^2$ values close to 0 imply that your line doesn’t fit your data well, and $R^2$ values close to 1 imply that your line fits your data well. Our 0.6148 $R^2$ value is roughly saying that we can use our line to predict grades about 61.48% of the time. So attendance grades alone may not be that accurate in predicting class grades according to the $R^2$ value, but there are many factors that determine a class grade, and attendance is only one of them. So we would expect a moderate $R^2$ value like 0.6148 for our data. Save your work to your I: drive, and if a Compatibility Checker window pops up just click **Continue** to save.

Congratulations! You are now a statistician that can use some powerful tools from Excel.
Part 2: Should you buy a hybrid car? Let’s use Excel to explore this question. (Lab Exercise by Dr. Everett, edited by Doc)

Download the Excel spreadsheet Hybrid.xls from the labs page, and save Hybrid.xls to your I: drive. Open up Hybrid.xls from your I: drive. If you see a message at the top part of Excel with a button to Enable Editing, then go ahead and click Enable Editing. Remember to save your work frequently.

- Now move on to the worksheet, Hybrid Cost Comparison. In this scenario we will weigh the higher present cost of buying a hybrid vehicle against the future cost savings. This exercise will require lots of formulas that you will have to think about first, and then you must input these formulas into Excel using the appropriate relative and absolute cell references. Here’s a quick review on relative and absolute cell references:
  - Relative cell references like B11 refer to B11, and when a formula is copied or filled with the fill handle, Excel will update the columns and rows relative to where the formula was copied to. So relative cell references change based on which cell you copy a formula to.
  - Absolute cell references like $B$8 refer to cell B8 and only to cell B8. Note that the $ sign denotes an absolute reference. When a formula is copied or filled with the fill handle, Excel will not change the absolute references to cell $B$8. So absolute cell references do not change based on where you copy a formula to.
  - Cells and formulas can have a mix of relative and absolute cell references
    - $B11 means use an absolute (non-changing) reference to column B, but change the row according to the cell that the formula is copied to.
    - B$11 means use a relative (changing) reference to column B, but the absolute reference to row 11 means that it the formula will always refer to the 11th row.
    - A formula $B$5/D4 tells Excel to use an absolute cell reference to B5 and then divide it by the relative value found in D4. So when this formula is copied to another cell, it will always contain B5 in the numerator, but the denominator D4 will change relative to where the formula was copied to.

- Now go back to the Hybrid Cost Comparison worksheet. I have entered two pairs of vehicles: a pair of sedans and a pair of small SUVs. The Toyota Prius and Ford Escape are the hybrids (fuel efficient automobiles). The Toyota Camry and Subaru Forester are the non-hybrids. For this model to be valid, the two cars in each pair should be equivalent except for price and fuel mileage. I’m not a car expert, but I suspect that this assumption is not quite true. So this model may not be valid – it does not accurately represent reality. But I believe this model is useful – it illustrates something important that does exist in reality, in this case the cost tradeoff.

- Start by calculating the number of gallons of gas used per year by each vehicle in row 6. This will depend on the number of miles driven per year, in cell B5, and one other factor that depends on the vehicle. You should get the results 364, 245, 600, and 429 gallons in cells D6, F6, H6, and J6 respectively.
Now we will attempt to predict the price of gasoline in the future. It would be great if we could just say “$1 PER GALLON”, but unfortunately this may not be realistic. We based this exercise on the high gas prices of 2008, which made many people think about buying more fuel-efficient cars. We made the assumption that gasoline will cost $3.50 on average in 2009 and gas prices will go up ten cents a year on the average. “Your mileage may vary,” as the saying goes; as you can see, this model has so many assumptions that an accurate prediction is impossible. Predicting the future is like that, but we still can’t resist trying!

Enter $1 in cell C11, as the present value of $1 in 2009. (We are using 2009 as the “present” for this calculation, assuming you will buy the car in 2009.) Now we will use a clever algebraic shortcut to calculate the remaining present values. Enter this formula in cell C12:

\[ = \frac{C11}{1+B8} \]

Don’t worry about the algebraic derivation of this formula. For now we will just experiment with a few values:

1. Start with the initial value of ten percent for the discount rate. The result should be around 91 cents which you’ll see in cell C12.
2. Think about what should happen when the discount rate is zero. Plug in zero for the discount rate and see what happens.
3. A larger value for the discount rate means the future is less important. Think about whether the result should go up or down if you change the discount rate to 20%, then try it. Restore the original value of 10% when you are done.

Add absolute cell references as appropriate to the formula in C12, and copy the formula down. Format all the present values as currency. Using a 10% discount rate, the present value of $1 in 2020 should be only 35 cents.

If gas is $3.50 in 2009 and goes up at ten cents each year, what should it be in 2010? Enter a formula in cell B12 for the price in gas in 2010. Be a bit careful here, as the number we are using for the annual price increase happens to be the same as the discount rate. Be sure you get the right cell reference! You should find that gas will cost $4.60 per gallon in 2020.

Now how much will it cost to drive the car for a year? This depends on how many gallons of gas you use, and the price per gallon. Enter formulas in cells D11, F11, H11 and J11 for the annual fuel cost of each of the four vehicles. You should find that they cost $1272.73, $857.14, $2100, and $1500, respectively, to fuel. Add absolute cell references as appropriate, and copy the formulas down through the columns. You should find that the annual fuel cost goes up each year, tracking the presumably increasing cost of gas.

Yeah, but how much do we care about the future cost of gas? Not as much as we care about present expenses, because the dollars in the future are worth less than dollars today. Calculate the present value of each future gas bill in columns E, G, I, and K, using the cost of gas and the present value of $1. You should find that even as the cost of fuel goes up, the present value of that cost goes down.

Now we are ready to calculate the present value of the total cost of owning each vehicle over the 12-year period from 2009 through 2020. First, calculate the total present value of...
the fuel bills for each vehicle in cells E23, G23, I23, and K23. Then add this to the purchase price to get the total present value in cells E24, G24, I24, and K24.

- You should find that in each case the hybrid vehicle is a better bargain according to this model. Note that the advantage is greater for the hybrid SUV than for the sedan, although the fuel efficiency advantage is a mere 8 mpg. Because SUVs guzzle so much gas, those 8 mpg will save you more money than the extra 16 mpg the Prius gives you relative to the Camry.

- All this is based on assumptions, of course. Try the following scenarios. In each case, undo one scenario before starting the next:

  - The price of gas drops to $1.50 in 2009 and continues to go up 10 cents per year;
  - You only drive 6,000 miles per year, not 12,000.
  - The discount rate is really 30% rather than 10%. (The amount by which a person values the present more than the future can be taken as a subjective judgment rather than a cold business calculation)

In each case you should see that the extra cost of the hybrid now outweighs the future savings.

    Congrats! You've reached the end of the lab!