Final Project Description

The final project for the class is planned for Monday, December 6th. Working in groups of 3-4 people, you will make a 5-10 minute presentation on a question of your choosing. You get to pick your own question, but it must meet certain guidelines:

Requirements:

- You must use at least 50 points of data to answer your question, and you must have at least two populations. Each population must have at least 20 points of data, and you need to indicate the source of your data in your presentation.
- You need at least one graph in your presentation (more may be helpful, its ultimately up to you). In order to show the class your graph(s) you should be prepared to use the digital projector in the classroom. Microsoft Office is available on the classroom computer, so you should be able to display Office documents.
- All group members must speak at some point in your presentation.
- Groups must send me an email by Wednesday, December 1st, indicating who you are working with and what you want to investigate.

In the book we often make assumptions about data being normally distributed and observations being independent; now that you are using your knowledge to answer a real world question you should examine the assumptions of our models. Most projects will need some combination of tools (don't forget about or underestimate the usefulness of box plots, histograms, least square regression lines, in addition to the tests we learned in later chapters). In the real world you will often remove outliers before performing calculations, you should be able to justify why you were able to remove outliers. Make sure to explain not only any doubts you might have about the data or your assumptions, but also what conclusions you reached (no real world data fits our models perfectly, so there are always judgement calls involved in applying statistics).

Some example questions:

- Compare the times of people running in the Tucson marathon and the Boston marathon. Test the hypothesis that the runners in the Boston marathon are faster, but also look for other interesting trends (which race do you expect to have a larger IQR and what does that mean?). Or compare different years from the same marathon.
- Find some real data from a cholesterol drug study and compare it to their control group. How likely is it that the drug is effective? Or another idea, compare data from several different drugs to test whether there is a significant difference between them via ANOVA. You may be surprised at how fuzzy drug trial statistics are in the real world.
- Pick a police department and dig up arrest statistics comparing race and the category of arrest (you could divide it by violent versus nonviolent, or subdivide it into more categories). Use a χ^2-test to determine whether these factors are correlated. The statistics don't show any kind of causality, only correlation, but it might be interesting to compare arrest statistics in different kinds of neighborhoods.

These are just examples, pick something **that is interesting to you**.