ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Proposal (or Final Report)
By: Nathan Stefanick

Step 1:
"Identify a problem," well I would have to say that the biggest problem in my life right now is that I am a poor college student, and I need a solid source of revenue to fund my addiction to Jimmy Johns, the local sandwich delivery service. So I decided the best way to increase my funds is to invest, but what do I invest in? I want to make a lot of money, so I decided that my best investment would be in some sort of global commodity. At first I wanted to invest the beverage trade: tea, coffee, hot chocolate, etc. However doing some research into I found a much more interesting commodity, oils. There is a huge demand for oil that I never really thought about, of course there is oil in the food production industry, almost everything we eat is cooked in oil or uses some sort of oil as a non-stick agent. However natural oils also go into shampoos, conditioners and many more beauty products. Oils are also used in the health industry to manufacture sunscreen, anti-acne cream, deodorant, and toothpaste. So how do I make money off all these uses for oil? The foundation of my project is to use the most basic principal of investment: "buy low, sell high."
Using the data that I find, I will compare the historical trends in prices of two oils, soybean and coconut, using a complete linear regression analysis. I will use the models I obtain to make inferences on the trends of oil prices, I will determine if there is a positive or negative correlation on oil prices for Soybean vs. Coconut oil. I will use measures of significance to determine if the model estimate is actually a good estimate or if there is too much variance to tell. Additionally I will look at the trends in oils from year to year to determine if there is any seasonal correlations in oil prices. What this could indicate is an increased production in oil during certain months due to limited availability and decreased price due to much more supply.

Step 2:
In my search for data I came across a spreadsheet for the United Nations Conference on Trade and Development. This spreadsheet contains a very large amount of historical information. It has the prices of eight different kinds of oil produced in various places around the world. These are the free market commodity prices as determined by the UN. This data goes back to January 1960 and goes until August 2015; the data is recorded on a monthly basis. It luckily can be downloaded in either Excel or comma delimited form, which makes it easy to translate into different programs. I plan on using SPSS, a statistical analysis program that I used in a stat class previously. Using this program I have many statistical tools at my disposal. Much of what I can and will do is with linear regression and analysis of variance.

Step 3:
I did Regression analysis for the following hypothesis test:
H_o: Soybean = $B_o + B_1*0 + \varepsilon_o$
H_1: Soybean = $B_o + B_1*$Coconut $+ \varepsilon_o$
Where B_o is the average price of soybean oil, and B_1 is the slope of the least squares regression line. The values of these are $93.70 and .73 respectively. This is the positive slope that we were looking for and it is pretty close to 1. If it were a perfect relationship the value of B_1 would be 1, this would indicate that for every dollar that the price of soybean oil increased, the price of coconut oil also increase 1 dollar. In this case when the price of coconut oil increases by 1 dollar, the price of soybean oil increases by 73 cents. A practical use for this stat in planning for buying stocks would be: if you see the price of soybean oil starting to drop, it would be a good idea to

sell your stock in coconut oil because chances are that the price for that will drop. Complementary, if you were to see the prices of Soybean oil increase, you would want to purchase stock in coconut oil, because your stocks are likely to increase in value, making you money. There are however some regression anomalies that are present in the model, most noteworthy is the heteroscedacisity. Most of the values, for both soybean and coconut oil are focused on the $300-600 area and then spread out upwards to $1500-2000. There is also some evidence of nonlinearity, although that only happens towards the upper values on the scatter plot. And based on all the other values on the scatter plot I would count out the outliers, there might have been a sort of political embargo or natural disaster effecting one or both of the oils. The last thing I took into account was the significance value of the test, this gives us a good overall idea of if the model is a good fit or not. The significance value is 0.000 so at and alpha of .01 we can reject the null hypothesis and go with the alternative hypothesis as a good fit for the data.

The next thing I did was look to see if there was any seasonal correlation in oil prices. So I averaged the prices of oil for each month for 52 years, and then made a histogram for each type of oil. I did this so I could get a good visual representation of how the average price of oil went up and down during each season. The result of this is very clear. For soybean oil the price significantly drops in the month of January, and then again rises until about May, where it stays the same for the rest of the year. In the month of January, it is Winter in the northern hemisphere, and Summer in the southern hemisphere. After doing some research, the biggest soybean production comes from both Brazil, and the United States. This brings up a strange predicament. Due to the fact that the United States and Brazil have opposite seasons, one would think that the price fluctuation would cancel each other out yearly based on seasonal price fall and rise. However that is not the case as evidently shown by the graph. So there seems to be another factor at play. Perhaps there is a global summit that attempts to bring down the price of the oil every January, or the data that I have only takes into account production from a certain area of the world. Either way, for the goal that I am trying to accomplish, it doesn't really matter. I want to strategize my profits by buying a lot of stock in soybean oil in January when the price is low, and then selling when the price goes back up in July.
For coconut oil there is a similar trend, except in a different month. From January to June the price stays pretty much the same. From July to September the price steadily decreases, and then for the rest of the year until December the price consistently rises until it reaches the same price that it started at in January. Using this information, I would buy stock in September and then sell in January. Like the soybean, the coconut exists pretty evenly to the North and the South of the equator. This would indicate that there is some other factor at play for seasonal price correlation.
We can tell that this find is significant for both Oils, because if the season was truly an unimportant, random factor, we would see these graphs to be random, or straight. The fact that the pattern exists is indication that there is a seasonal effect on price, what it is though, I don't know.

In conclusion, I you want to make money of the natural oil trade, then follow these simple steps. First, when you see there to be pricing movement in either coconut oil or soybean oil, make the same purchasing or selling with each of the oils to maximize profits. Also, as a general seasonal rule, buy soybean stock in January and sell in July, also buy coconut oil stock in September, and sell in January. If I were to go more in depth with this project there would be two things that I would analyze. The firs would be checking to see if over the long run the prices of Soybean and Coconut Oil rise or fall, or remain the same. This would be determinant if I should buy a lot of stock early on and just let it sit until I retire. The next thing I would do is check the variance of seasonal graphs, to determine if they are actually good estimates for predictive seasonal prices.