Alexander Chang – Project 1 Milestone
Github Link: https://github.com/theamazingchang/BigDataProject
NRSG 741 Big Data

Working Title (informal): Booze, Happiness, Soccer…and Trump.

Working Title (formal): Is there a difference between countries that drink more in Happiness, Soccer, and Opinion?

Overview & Motivation: Ever since I started to really enjoy basketball, I knew that understanding the statistics and math behind the game was going to be important to my enjoying it in the years to come. It was not until my time as an undergraduate in college that I began taking statistics coursework which allowed me to enjoy the analytics that revealed the nuance underneath the surface of the sport. It was also about this time that Nate Silver, whose body of work has been primarily in election prediction and baseball, transitioned into writing extensively about basketball and other hot button topics in his blog, FiveThirtyEight. Inspired by his approach to writing and analytics, this project aims to leverage cleaned datasets from a variety of different sources to produce fun analysis between the amount of alcohol a country consumes, and other variables of which an association can be drawn.

The idea is to incorporate some of the stylistic elements behind the award winning news site that makes analytics comprehendible to a wide audience of people. In particular, using data visualization to guide the rhetoric. In doing this, I am preparing myself to produce deliverables that would better break down complex topics in public health or environmental health in a format that is accessible to others. This skill would be useful for consulting and academics, both being career paths that I aim to pursue following the time I have here at Rollins. Furthermore, this would serve as practice in utilizing the ggplot2 package in R, as most of my data visualization experience is grounded in both SAS and Excel.

Objective:

The focal research question is whether or not countries that drink more are happier, more interested in soccer, and see Trump in a more positive light.

What could be useful from this analysis is seeing the difference between the cultures of different countries by using alcohol sales as a proxy. More conservative regions drink less, and thereby have a vastly different life experience that states that drink more. This could possibly affect their happiness index, participation in the world's most popular sport, and their opinion over the man holding the highest office in America.

Data: There are 4 datasets that will be in use

Alcohol Consumption by Country – Total Liters of Alcohol Sold

https://github.com/fivethirtyeight/data/tree/master/alcohol-consumption

Fifa Population share – Country's share of both global population and global world cup TV Audience

https://github.com/fivethirtyeight/data/tree/master/fifa

World Happiness Index by County – Happiness Score, Happiness Rank

https://www.kaggle.com/unsdsn/world-happiness/data

Pew Research Data, World Trust of Trump – Trust in US Pres in World Affairs, Favorability of the US

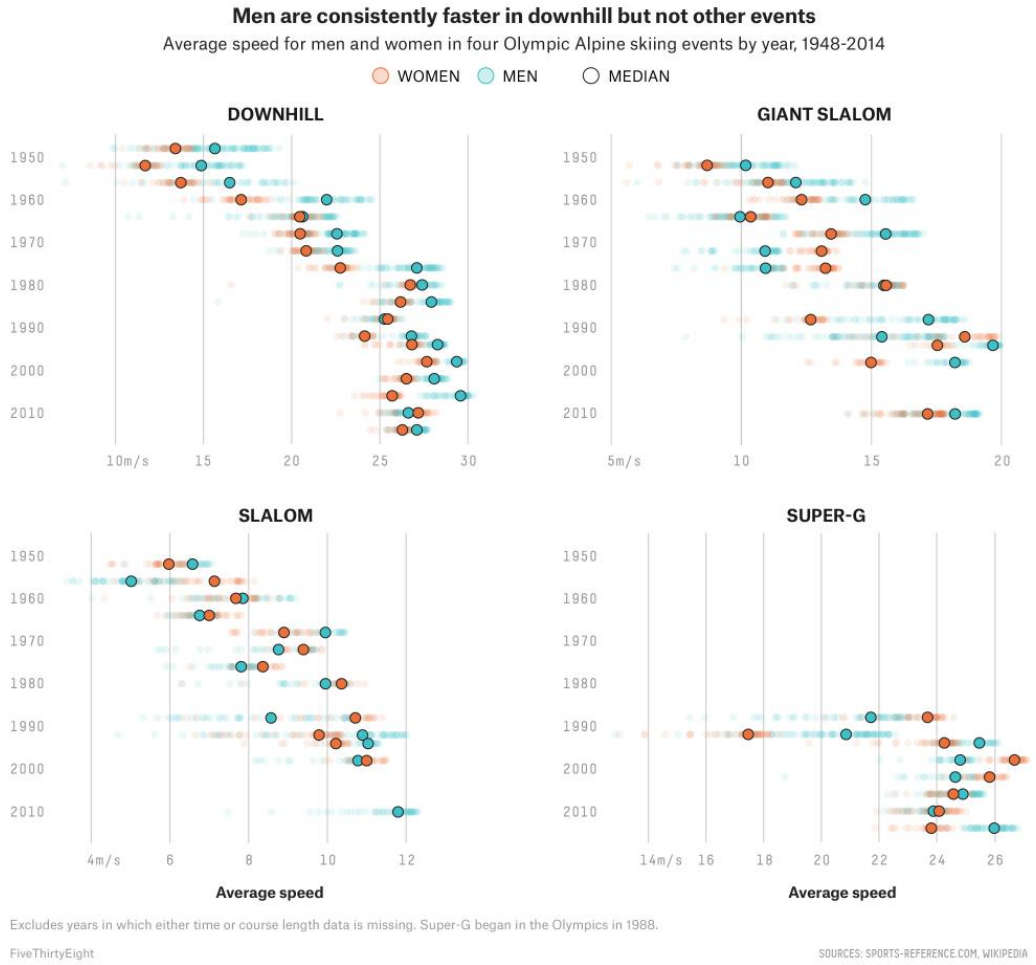
Alexander Chang – Project 1 Milestone
Github Link: https://github.com/theamazingchang/BigDataProject
NRSG 741 Big Data
https://www.kaggle.com/unsdsn/world-happiness/data

Data Wrangling:

The data was procured from formal datasets that have been previously cleaned, thus the only significant data wrangling is merging all the variables together by country. Afterwards it is necessary to comb through the data and find countries that are missing variables. Most notably, countries that do not have a FIFA team will most likely be removed from this analysis. Furthermore, there are only a select number of countries that was included in the Pew research Data, these countries will likely be set into a separate strata for their analysis.

Exploratory Analysis:

I would want to run the analysis and highlight countries of interest in a scatter plot and identify if there could be trends just through the scatter plot visualization. For example, having the USA and Belgium highlighted as separate colors and plotted along FIFA viewership, Trump Support, and Alcohol Consumption plots. Below is an example of what the visualization could look like:



Analysis: Linear Regression will be used to test the strength of association between individual continuous variables. However, I'm tempted to divide some of the countries into different strata (by Regions, *Middle East vs SEA*, by GDP, or Religion) and conduct a Chi Square test to see

different between these different categories and classes. This could tell us a bit more about the trends to look for and what exactly could divide these countries by these metrics.

Schedule:

Feb 14 – Feb -21: Exploratory Analysis and Data Wrangling

Feb 21 – Feb 28: Produce Visuals, Draft Results

Feb 28 – March 7: Draft Introduction

March 7 – March 14: Draft Discussion

Mach 14 – March 21: First rough draft of project

March 21 – March 28: Incorporate edits of draft into R Markdown doc