Zachary Blitz
SI 330
Professor Chris Teplovs
Final Project

MOTIVATION:

Throughout my life, I have always enjoyed playing video games. To this day, there is no game I enjoy playing more than FIFA. I decided to analyze the game of FIFA using the FIFA 18 player dataset and the REST countries API to try find out if there is a correlation between player ratings and country populations. From my statistical analysis, I will be able to see if a player is likely to be higher rated in FIFA if he is from a country with a higher population.

DATA SOURCES:

FIFA DATASET URL: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

REST COUNTRIES API URL**:** https://restcountries.eu/rest/v2/name/

The first data source that I used was the FIFA 18 full player dataset, which is a csv file. I found the dataset on kaggle.com, one of the websites suggested to use for my project in the instructions. The file contains 75 columns of all player statistics (name, age, country, overall rating, acceleration, aggression, etc.) and 17,982 rows of players. The two important columns I needed to grab information from was the country column and the overall rating column.

The second data source that I used was the REST countries API. The rest countries API requires a user to input a country and then it returns information about the country such as its capital, region, time zone, population, etc. When I called the API, I got my results returned in JSON format. I retrieved information of 155 different countries. The most important variables contained in all the information retrieved from this API was the population of the country.

DATA MANIPULATION METHODS:

To begin, I used csv.DictReader to collect data from the FIFA 18 full player dataset. I created a unique list called countries that appended each country in the dataset to the list if it were not already in there. Then, I created a dictionary called country_rating_dict to keep track of each country and all of the overall rating values of each country with the country as the key and all the ratings as the value (see below).

country_rating_dict = {'Spain': [80,80,85,90,90], 'France': [88, 90, 92]} *

I then created a function called get_rating_average that takes a list and returns the average of the values in it (the sum of the list divided by the length of the list). I then iterated through the country_rating_dict dictionary and ran the function against the all of the values of it. I created a dictionary called average_rating_by_country, and as I iterated through the previous dictionary, I added to this new dictionary with each country as the key and average rating as the value (see below).

average_rating_by_country = {'Spain': 85, 'France': 90} *

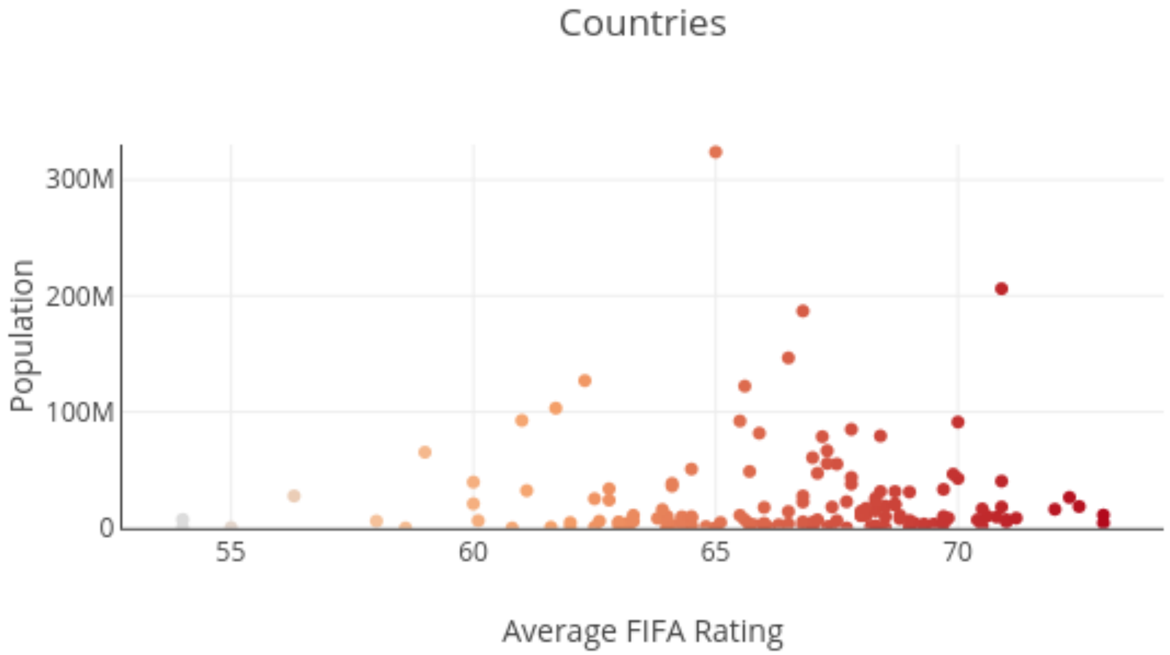
Next, I needed to find the population of each country. I created a dictionary called country_population_dict with the intention of making each key a different country and each corresponding value the country's population. Iterating through the unique countries list, I ran each country in the API and retrieved each population. As the iteration was happening, I added to the dictionary making each country the key and its population as the value (see below).

country_population_dict = {‘Spain’: 8,939,843, ‘France’: 10,483,234} *

I made a JSON file caching all the data from the API so that the program would run faster and not have to grab from the internet every time I ran it. As I was making my project, I had to deal with two problems. First, some of the countries in the dataset were not recognized countries by the API. To combat this, if the name of the country was not recognized, I had the program continue in the iteration and move on to the next row. Second, some of the countries in the dataset had a slightly different required input name for the API. For example, in the FIFA dataset the country was called Bosnia Herzegovina. However, for the API the country needed to be called Bosnia and Herzegovina. To combat this, as I iterated through the countries, I changed the name of each country so that the name was able to be run through the API.

ANALYSIS AND VISUALIZATION:

After aggregating and manipulating my data, I decided to output the data to a new excel file. This excel file is called fifa_final_output. There is a total of 155 rows and 3 columns containing each country, the average FIFA rating for that country, and the population of that country. I used plot.ly to make a scatterplot of all the data (see below).



After making the scatterplot, I decided to go back into my code and find the r-correlation coefficient value to see if there was any correlation between the average FIFA rating of a country and the population of the country. After running the numbers, I found a r value of -0.0000054510, indicating there is no correlation between the variables. Now I know that a player is not likely to be higher rated in FIFA if he is from a country with a larger population relative to other countries.

***dictionary values in this write-up are fictional – they do not represent actual values**