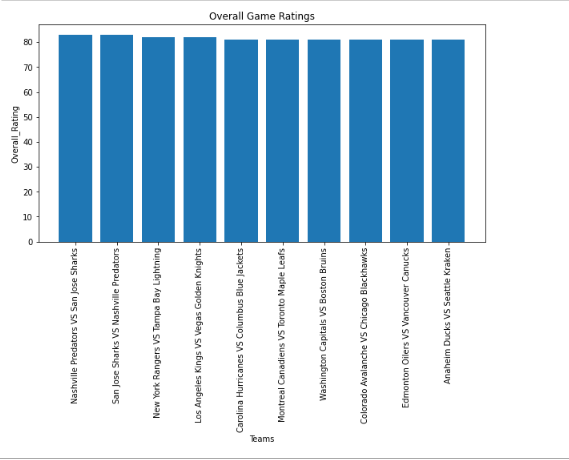
I am glad to state that I made the decision to complete the project on my own. As the first stage, I conducted some research and went over the list of sources provided for data mining. I concentrated on data sources that were both intriguing to me and straightforward to get information from. Finally, I chose to base my judgment on the data I gathered from the websites Kaggle and FiveThirtyEight.com. Let's look at each of these websites separately because the methods they employ to collect data differ. As previously stated, data access and modification will be handled through the use of a Google Collab notebook. In addition, the code is available on my Github page.

**Five Thirty Eight website**

The data obtained from this website was the National Hockey League (NHL) data. It contains data for the team ratings, match ratings and scores between different teams among others. The website provides a link the csv file; we therefore read the file directly using the read\_csv method of Pandas. Our main interest was to find out the highest rated games in NHL. We did this by ranking sorting the dataframe based on the 'game\_overall\_rating' column and taking the first 10 samples of the sorted data using the ‘head’ method of a Pandas dataframe.

We then plotted a bar graph to show the overall game ratings using matplotlib library. Below is the output.



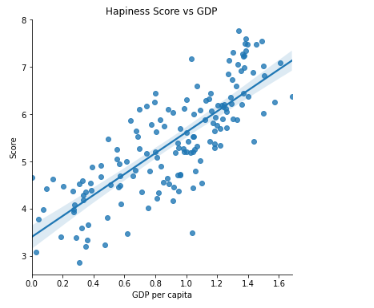
Further, we were able to note from the data, that the higher the game quality and the game importance ratings, the higher the game overall rating.

**Kaggle**

The data obtained from Kaggle is the World happiness data. It contains data such as the Overall rank, Country or region, Score, GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity and Perceptions of corruption.

Extraction of data from Kaggle is quite different when compared to the Four Thirty Eight website. We make use of the opendatasets library to download the dataset from Kaggle. This prompts us to provide the details like the Kaggle username and the key, we are obtained in a json file downloaded from Kaggle under the ‘Account’ option. After provideng these details, the dataset is downloaded and exracted to our working directory. From here, we can then read the csv file using the read\_csv method of Pandas.

Our main interest here was to find the correlation between the happiness score and the GDP per capita. To investigate this, we made a scatterplot using seaborn and this included a regression line. The output graph showed that the two are directly proportional in most parts of the world. Below is the output graph:



We also noted that, in the countries that the freedom to make life choices was high, the happiness score was also high.

**Challenges Faced**

The main challenge we faced was in the extraction of data from Kaggle since we did not know that a json file from Kaggle was required so as to download the data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |

**Conclusion**

From the project, I was able to import different sets of data and also extract meaningful insights from them.

For the first case, which is the data from FiveThirtyEight website, I am able to first read data from a link using Pandas. The data contained different data types which included 11 which were float type, 6 integers and 7 which were object type. It is noted that there is a high correlation between the columns for the game ratings and the win probability columns. In addition to that, when the game ratings are plotted, I noted that the rankings of the teams are almost in groups; that is, highly ranked, averagely ranked and lowly ranked games. Matplotlib library was utilized in visualizing the game rankings from high to low.

From this part, I was able to learn how to properly use pandas in importing the data from a website link, know the data types in the data, and compute the measures of central tendency in the data. Filtering out columns of a data frame and ranking them is another concept that I learnt too. In addition to that, I understood how to find the correlation between the data variables. I also knew how to visualize the correlation of the variables using Seaborn.

In the second case where the data was going to be obtained from Kaggle. The data was first downloaded using the Opendatasets library and then read using pandas. The data types it contained were 7 float variables, one integer and also one object type variable. The object type variable contained the list of countries. The measures of central tendency were also checked. The observation was that there was a high correlation between Happiness score and GDP per capita, social support, Healthy life expectancy & Freedom to make life choices. These columns could therefore be used to predict score which is the target variable.

In this part, I understood how to make use of the Opendatasets library to download data from Kaggle. To download, a json file from Kaggle, which contained the username and the key was required; we therefore downloaded it. The other concept reinforced here was checking the measures of central tendency like the mean, median and percentiles with Pandas, checking the correlation between the variables and visualizing this using the correlation heatmap with seaborn. Seaborn library was also utilized in visualizing the data and drawing a regression line, especially for those columns that had a positive correlation with Happiness score.

The project was therefore a success coupled with great learning experience and activities that sharpened my programming skills.