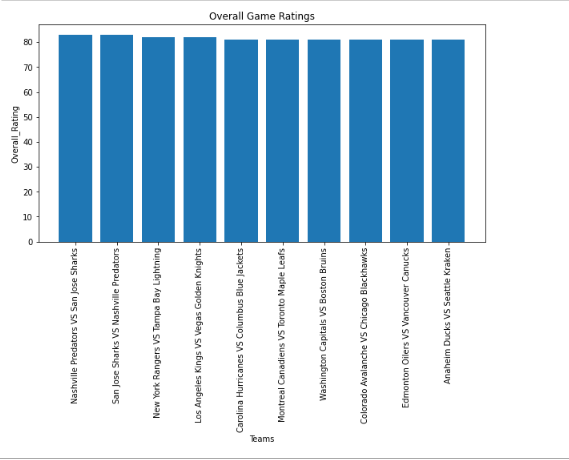
In the project, we decided to collaborate with a friend in order to get the work done. We first met in a zoom and went through the list of websites provided so that we could extract the data from them. We focused on the sources that provided us with interesting data as well as those that would easily be extracted. We finally chose to settle with data from Kaggle and the Five Thirty Eight websites. The extraction methods for the data from these sites is different so let us look at them one at a time. We also unanimously decided to use Jupyter notebook for data reading and manipulation. On top of that, we agreed to upload the code to my Github account.

**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.



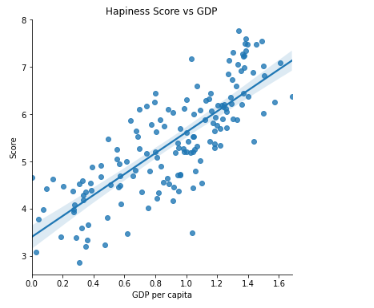
Further, we were able to note from the data, that the higher the game quality and the game importance ratings, the higher the game overall rating.

**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.

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