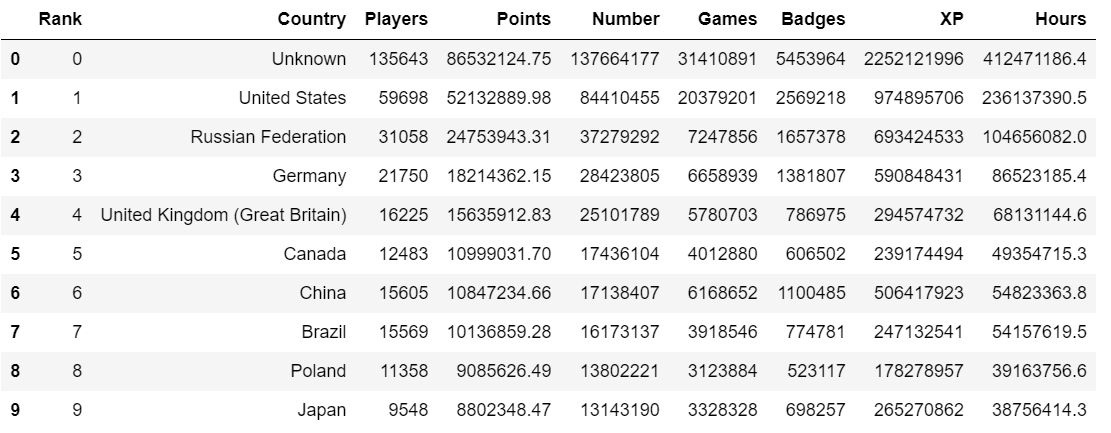
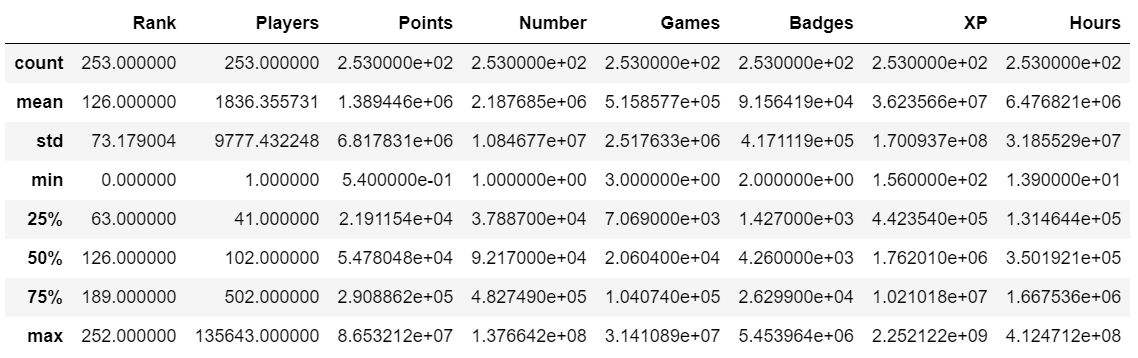
**Steam Players Country-wise Comparison**

This project focuses on taking in the Steam Country Dataset, that include the Countries, its no. of Players, and the Points or XP gained overall by that country till date. Then, the data would be cleaned and three regression models would be applied to predict the number of points if a player plays for ‘x’ no. of hours. The basic results show the Linear, L1 and L2 regression outputs, and also show the best data frame to use after several changes to the main dataset.



Looking at the original dataset, it was observed that the data is very large, and it may or may not affect the overall model or algorithm.



The describe function shows how the large the data is clearly. If we look at the mean of the values in any column, say ‘XP’, it is in the format ‘number’ \* e07. This necessarily doesn’t mean that it would ruin the performance of the model, but creating another dataset with cleaner values would show the dominance of the correct dataset, and clear us of the same doubt.

Over the next few steps, the data would be compressed and would be refined to the values at which the Linear model can be implemented.

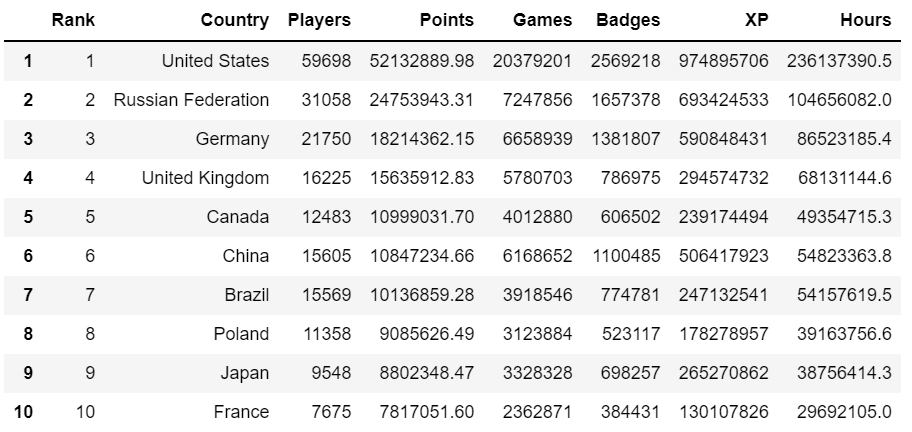


We have no null values in the dataset. But we have a row called as that has the value of ‘Unknown’ in the Country column (the first row of the dataset). It is unnecessary, and removing it would be beneficial. Also, the name United Kingdom (Great Britain) is a big name, that isn't required (row no. 5). Simply converting it to United Kingdom makes the value short and precise. The column – ‘Number’, is also unnecessary and is of no importance. So, that would also be removed.

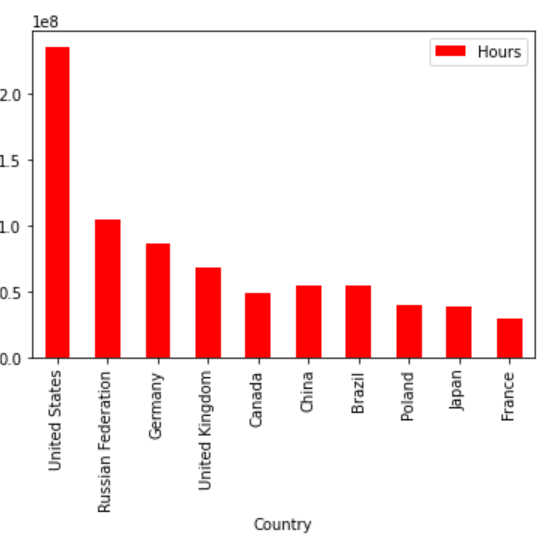
This is how the dataset looks after performing all the above steps.



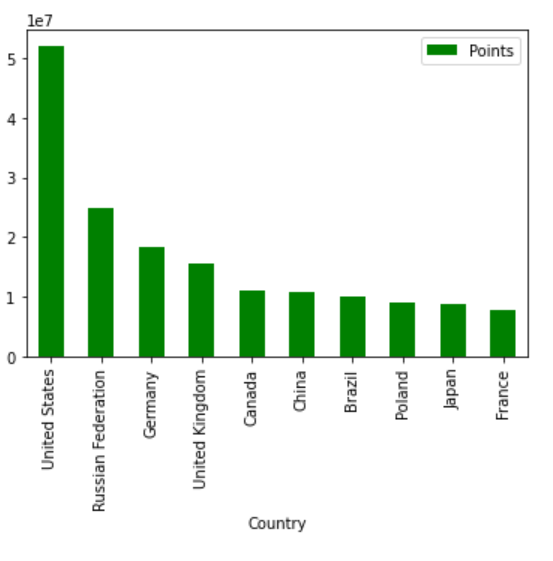
Now, due to a lot of large and different values, it is difficult to interpret how they have ranked the countries and on the basis of what column they say this country is at this position. So, a top 10 dataframe will be created, and it will be used to create different comparative bar graphs. If we find a constant decrease after each country and the value for which we are checking, it would confirm the same as the main factor for ranking the countries.



This is the top 10 dataframe. Now, we will create a bar graph comparison between the countries and the other columns.



Here, for example, when we used ‘Hours’ in the comparison, we saw that it decreases constantly till Canada, but increases again after that. So, it is not the main factor of ranking.



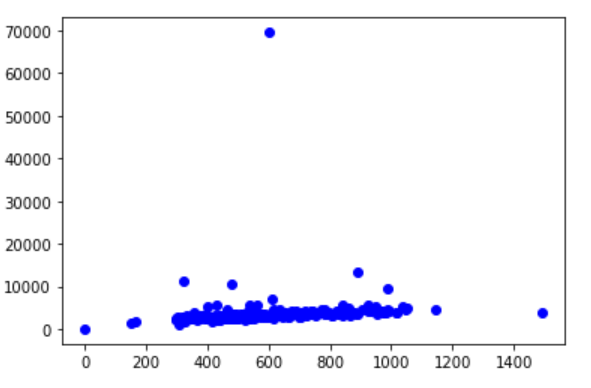
After comparing with all the columns, only the points column shows a constant decrease. So, we can use the ‘Points’ column in the model as decided.

As discussed previously, we will be creating a dataset with reduced values and better interpretability. To do this, we will create an average dataframe, that takes the values average per player in a country. This can be achieved by dividing each column with the ‘Players’ column.

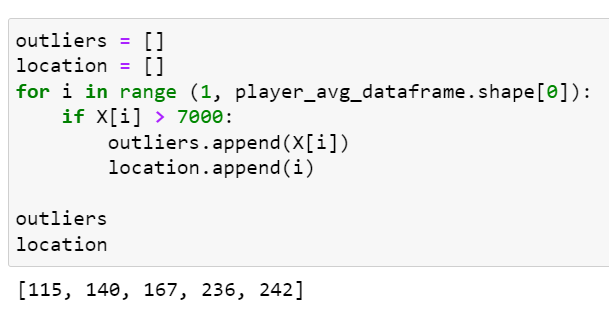


Now, for the final step in this dataframe, we would remove the Rank, Country and Players column as they are of no use after.





After looking at the graph, it appears to be very compressed due to some outliers. So, we will remove the outliers to make the dataframe usable and efficient for a model.



We took the values above 7000, and used the append() function to add them to a list named ‘outliers’, and simultaneously append the locations of the outliers in the dataframe.

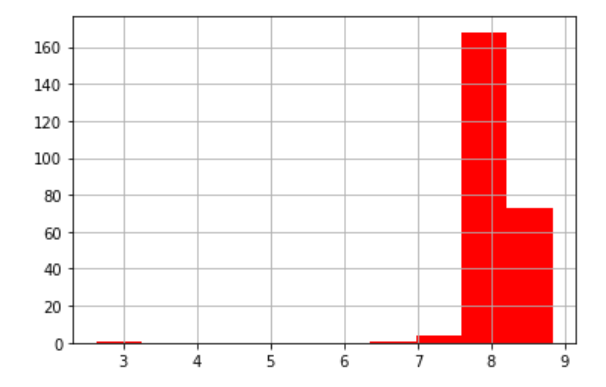
Here, we can conclude that the elements at location 115, 140, 167, 236 and 242 would be considered as outliers. So, we will remove them from the dataframe.

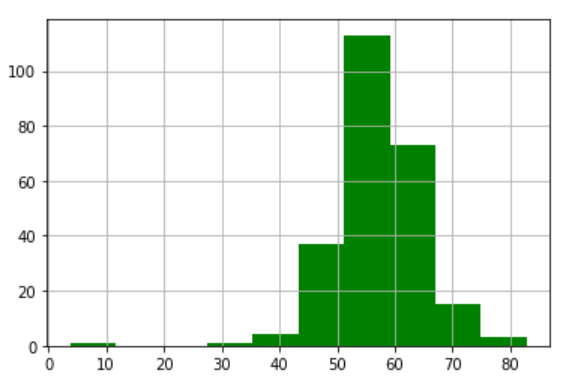
Now, we need to normalize the data with any of the three methods, which are

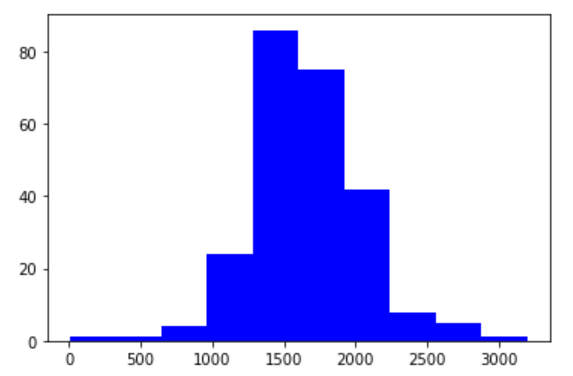
1. Square Root Transformation

2. Log Transformation

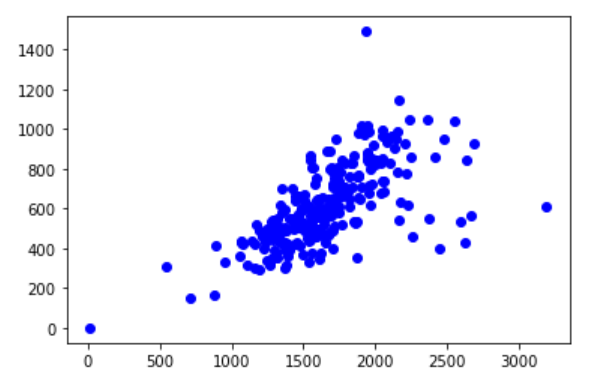
3. Boxcox Method





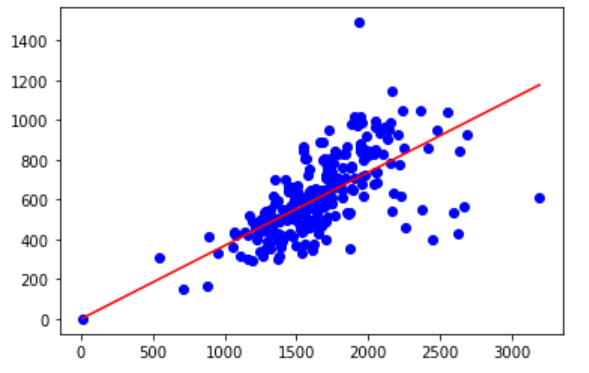


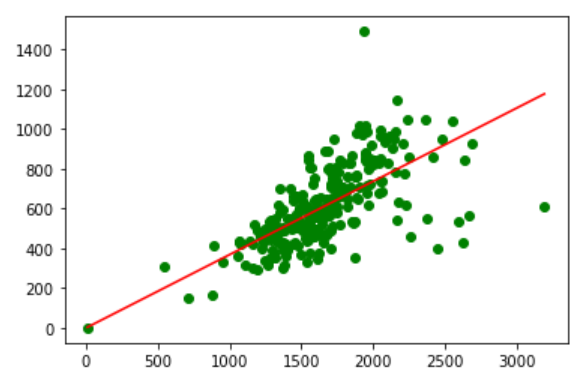
The normaltest for Boxcox method proved out to be having the closest p-value to 0.05. We will be using the boxcox normalized values for the regression models.

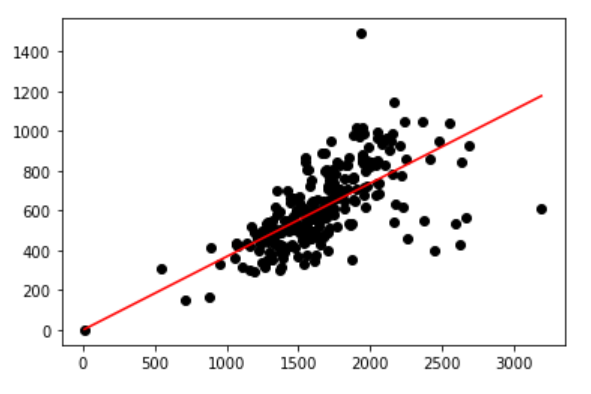


This is the dataset in a scatterplot after doing all the changes.

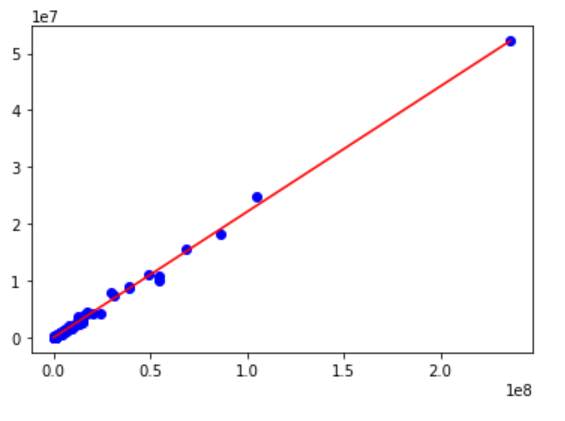
The regression models (Linear regression, L1 and L2) are applied to the dataset.

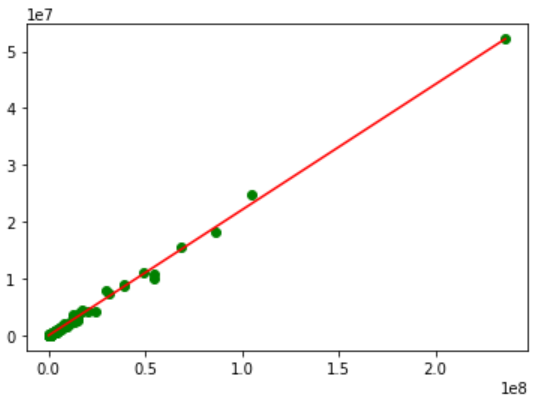


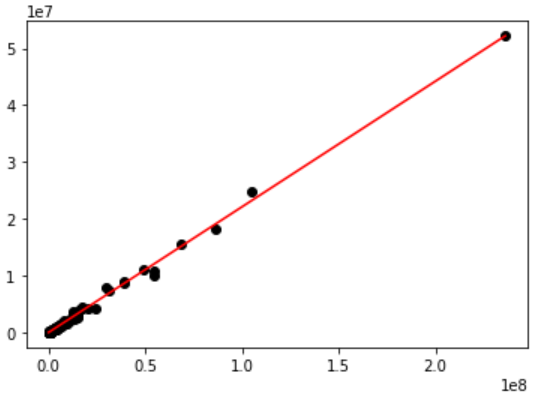




All three Regression techniques give the same values and plots. The r-squared value came out to be around 0.44. Now, we will take the original dataframe and repeat the same steps for regression.







Still the same in all three regressions, but a better visualization of the linear data. This dataset seems to be highly linear in nature, and could be used to predict any value easily. The r2 score of 0.99> shows how less the errors are in the overall dataset, which were not that visible after taking the average of the dataframe values.

Hence, the regression models were successfully applied and visualized.