1.

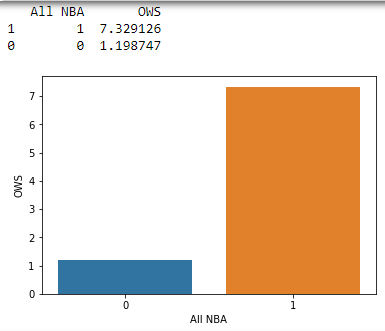
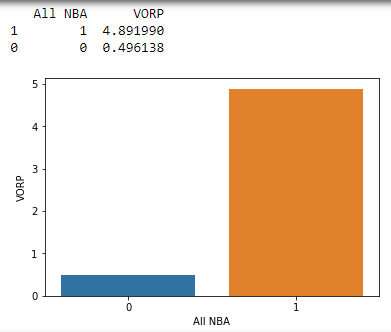
I approached this project by contemplating the programming language and initial dataset about past NBA players’ season stats publicly available. I decided to use Python and the Jupyter Notebook web application along the lines with the ‘NBA Players stats since 1950’ dataset from Kaggle (<https://www.kaggle.com/drgilermo/nba-players-stats>) to begin the data pre-processing step (seasons until 2017). After shrinking the dataset’s size to seasons from 1990 to 2017 and utilizing NBA’s official website to add three more numerical features essentially called ‘All NBA’, ‘All NBA Accumulated’, and ‘All NBA Total Career’, I transferred entries of Irving, Towns, and Curry to another small dataset that I manually completed with data from seasons 2018 to 2020 as of 11/28/2019. Additionally, I extended some features to be displayed as per-game stats. Using Python, I filled in empty cells with the features’ medians, and deleted confusing rows explained by mid-season transfers. My main feature selection is based on an algorithm filtering out attributes with a higher overall correlation than 0.95. Besides, I deleted ‘Player’, ‘Year’, ‘All NBA’ and ‘All NBA Accumulated’ due to their insignificance for the model, and ‘PER’ due to its excessive number of outliers.

Next, I split my dataset into train and test data to train a linear regressor to predict the career total number of All NBA selections for any given player at any given point in their career. I used this model to complete task 3 of the project.

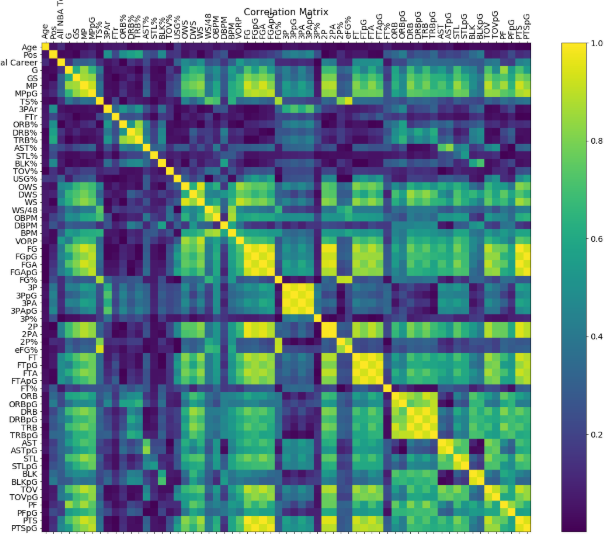
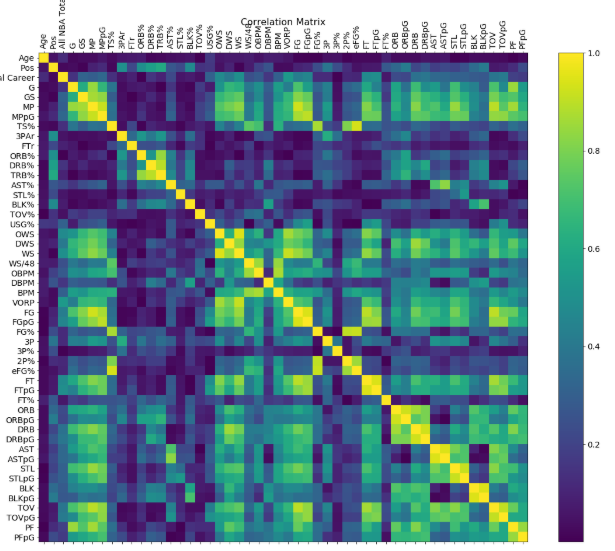
I used another approach that took the player’s average career data, concluding in a single row per player within the dataset. My resulting model had better scores but predicted the number of All NBA Selections for the four players worse than my initial model.

2.

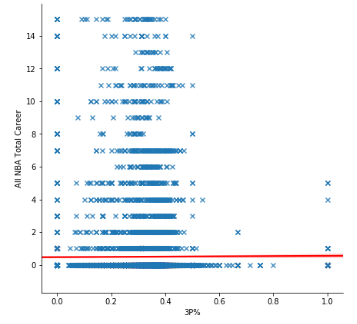
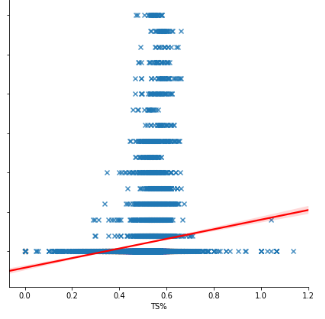
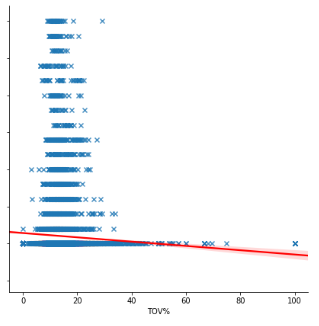
During the pre-processing step in developing my model, I wanted to achieve visual information about possible key features in determining an All NBA Selection in general for a better understanding of the data. Using matplotlib, I plotted every single feature of the dataset in contrast to the attribute All NBA. The two most significant features ended up being OWS and VORP illustrated by the driagrams below.



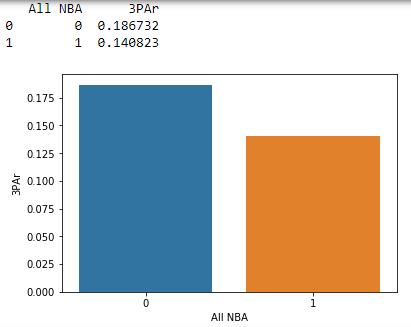
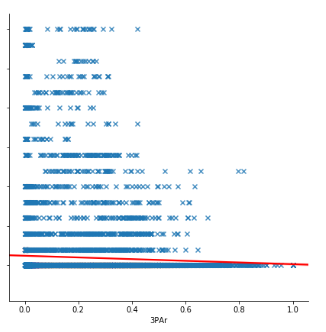
The next visualization represents the underlying correlation matrix of all features of the dataset excluding ‘Year’ and ‘Player’ before and after removing highly correlated features. The denser the yellow color, the higher the correlation of that feature in the bigger scope. The goal of my feature selection was to mitigate data redundancy and model confusion by using a Python algorithm that returns a list of all features with a higher correlation than 0.95. The red circles illustrate a few instances in which high correlation was removed from the dataset.

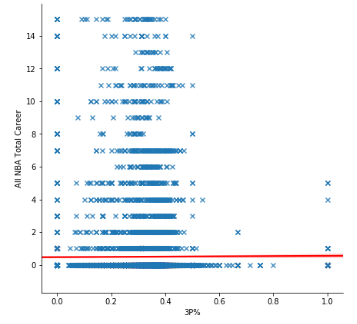


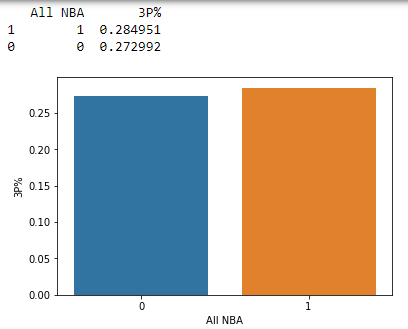
The following diagrams are visualizations of fitted regression lines calculated by my model. With close to 50 different fitted regression lines, I chose features that represent a positive, negative, and close to no correlation respectively. 3P% shows close to no correlation to All NBA Total Career which is illustrated by a relatively horizontal line. TS% shows a positive correlation with the dependent variable All NBA Total Career tendentially increasing as the independent variable TS% increases. Vice versa, TOV% shows a negative correlation which means that as TOV% increases, All NBA Total Career decreases. The trends are substantiated by the correlation matrix from above. I decided to incorporate these visualizations to understand better how my model works.



The negative correlation of 3PAr, and the close to no correlation of 3P% are also visually illustrated by the first type of visualization I provided.







3i)

Using my model, I predicted the number of remaining All NBA Selections for each of the four player’s careers:

Luka Doncic: **4**

Karl-Anthony Towns: **3**

Kyrie Irving: **2**

Stephen Curry: **0**

The number of remaining All NBA selections is calculated by subtracting the All NBA Accumulated number from the predicted number of total All NBA Selections. With their current statistics of season 2020, the remaining All NBA Selections are predicted to be six, three, two, and zero.

3ii)

Since the linear regressor already implies probability, the model predicts Luka Doncic to have the highest probability to receive the greatest number of All NBA Selections remaining among this group followed by Towns, Irving, and Curry in both, the previous, and current season.