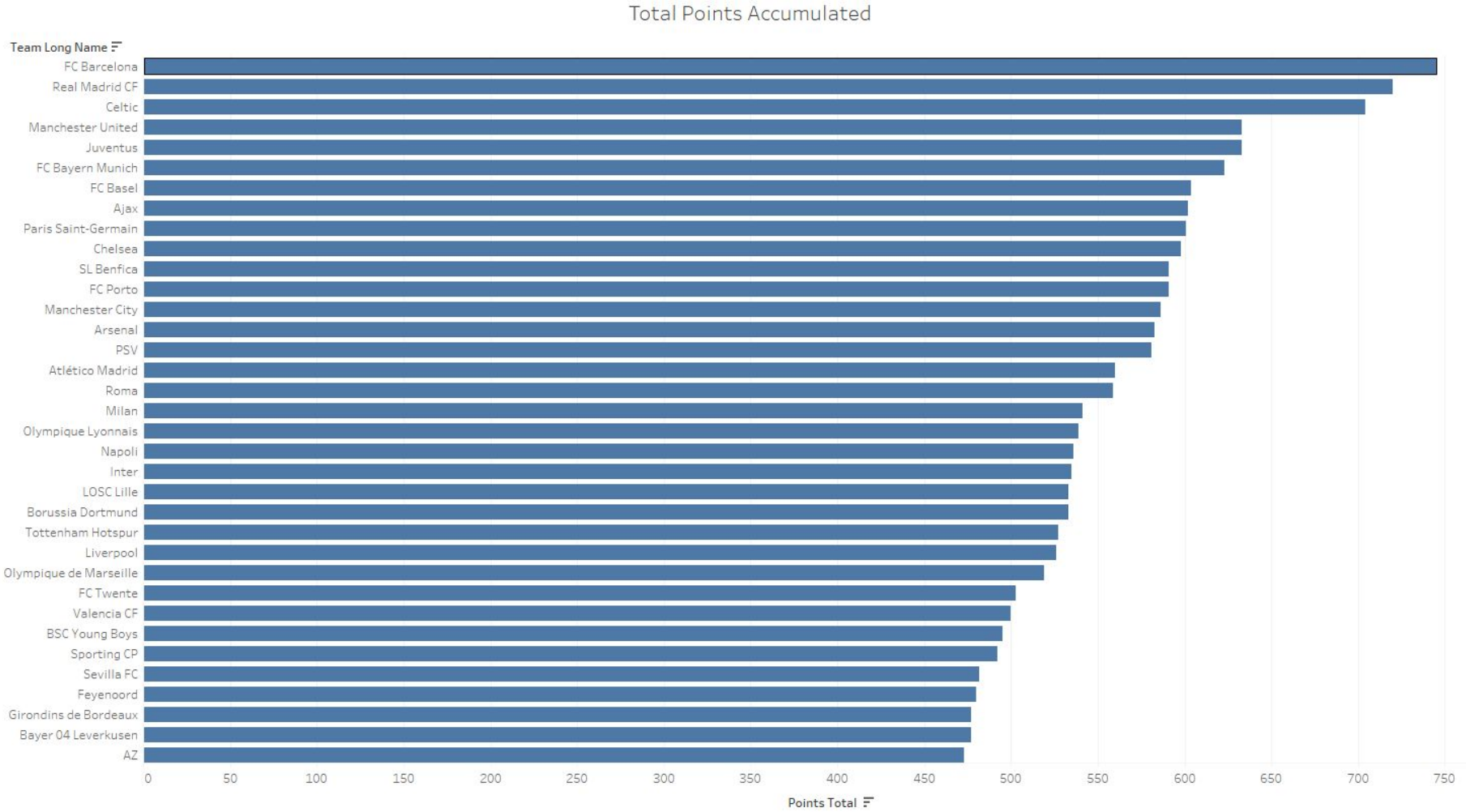
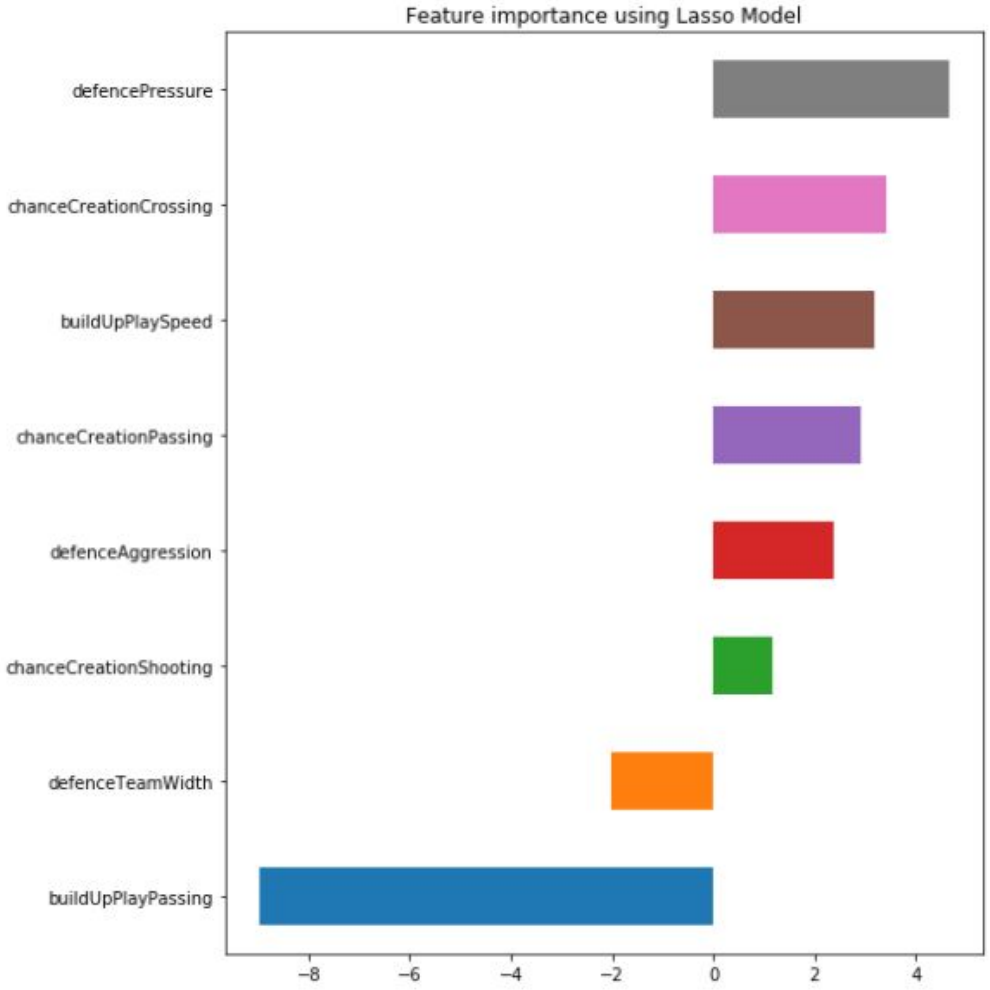
European Soccer Database Project

Ben Green

What I did in this project

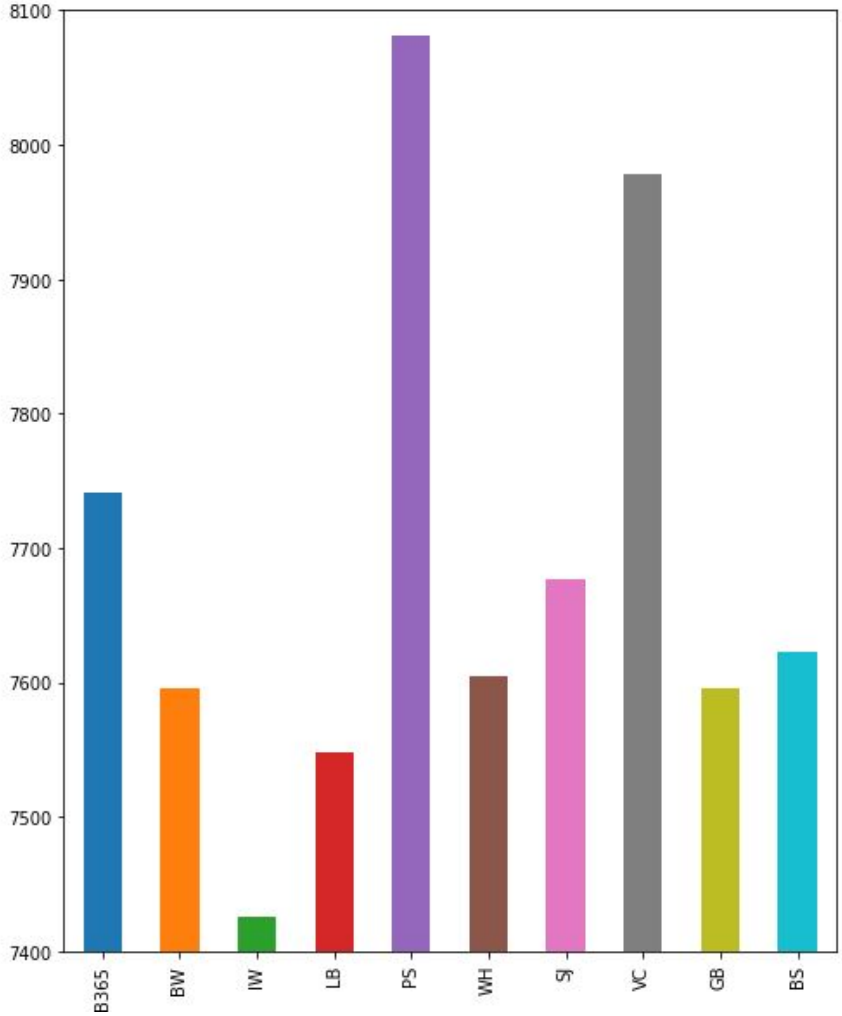
- First, I explored the data to see what I was working with and what analysis I could perform.
- Then for my first analysis, I got the total points accumulated for each team over all of their matches. Then, I compared that to their Fifa Team statistics.
- To see if there was any correlation between the Fifa Team statistics and total points accumulated, I ran some regressions.
- Next, I took a look at the betting odds given for each game and wanted to determine who gave the worst odds and therefore would be the best to bet with
- After, I looked at the Fifa player attributes and grabbed who was the most improved player over the time the dataset has info for
- Still using the Fifa player attributes I tried to classify the players by position


Total Points Accumulated
Team Long Name
FC Barcelona
Real Madrid CF
Celtic
Manchester United
Juventus
FC Bayern Munich
FC Basel
Ajax
Paris Saint-Germain
Chelsea
SL Benfica
FC Porto
Manchester City
Arsenal
PSV
Atlético Madrid
Roma
Milan
Olympique Lyonnais
Napoli
Inter
LOSC Lille
Borussia Dortmund
Tottenham Hotspur
Liverpool
Olympique de Marseille
FC Twente
Valencia CF
BSC Young Boys
Sporting CP
Sevilla FC
Feyenoord
Girondins de Bordeaux
Bayer 04 Leverkusen
AZ
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
Points Total

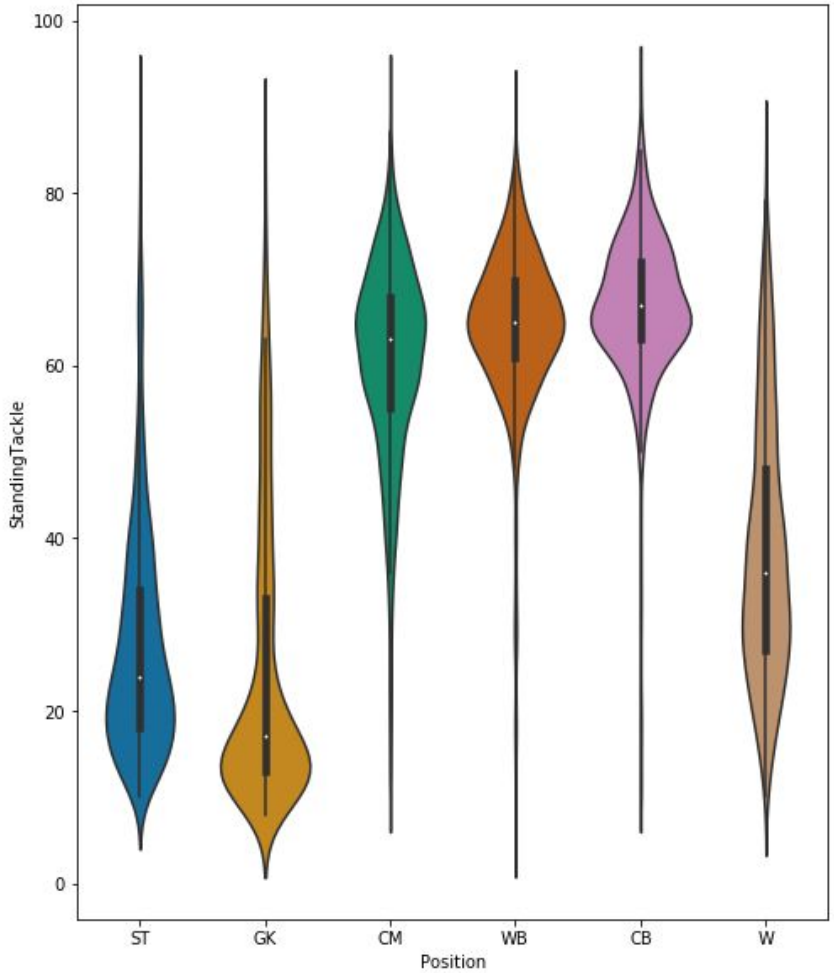


Best Alpha 0.350
Best Score 0.210

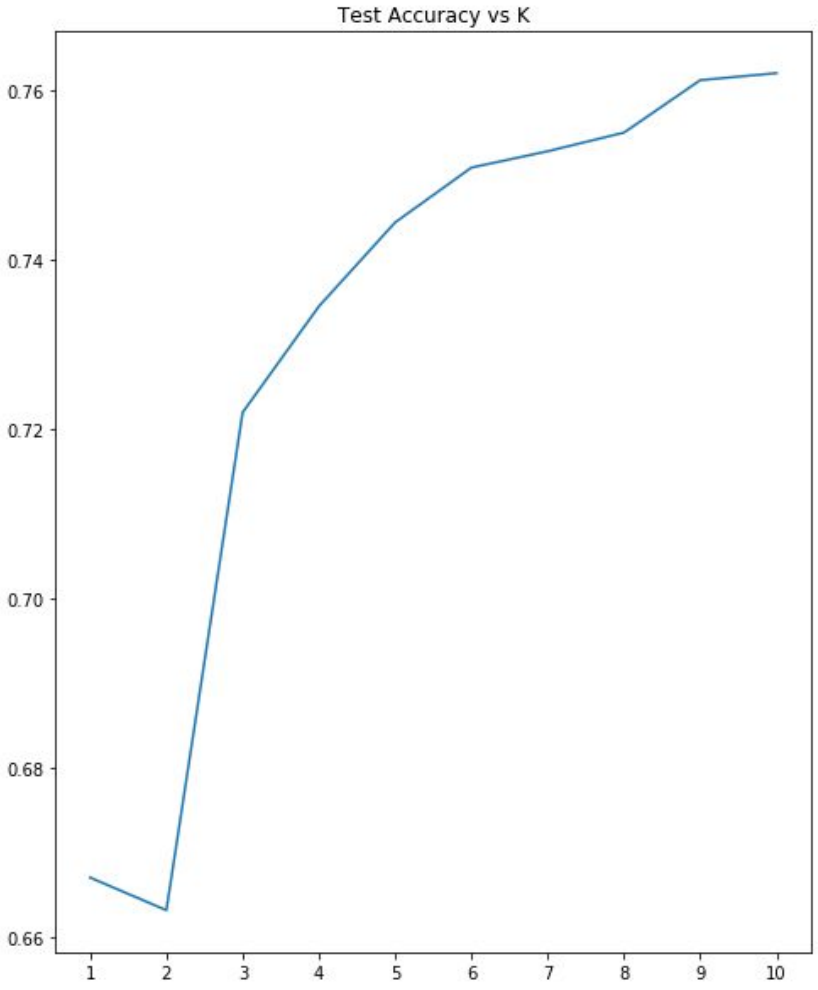
Based on these results, there is no real Correlation between total points and Any of the team attributes. However, the most important features are defencePressure and buildUpPlayPassing



Based on these results, placing a bet with PS
Would yield the highest payoff



This graph shows the players grouped by position And their ‘StandingTackle’ ratings. As you can see, There are some distinct differences with just this One rating. Based on this we suspect we could Predict their position off all the ratings.



This graph shows the accuracy of the prediction based on different K values in the KNN model. As we increase K, we get more accurate with diminishing returns up to 76%

The most accurate model I ran was the logistic regression with an accuracy of 79%