Soccer Data Analysis

Tufei Cai, William Fitzhugh

Author contributions

1. William worked on tidying the dataset, prepared the data description, create some data visualization, and wrote the background information.

2. Tufei worked on tidying the dataset, created some data visualization, finish exploratory analysis.

Abstract

Soccer is an incredibly popular and complex sport. The motivation of this analysis is to gain insight on the top soccer leagues in Europe. This analysis has two aims. The first is to create metrics which summarize aspects of a soccer league and compare the relationships between these metrics among European soccer leagues. The second aim is to search for a relationship between players' physical attributes and players' overal ability through the use of Fifa stats. The results of our analysis show that there is little association between any physucal characteristic and overall ability. There is, however, an negative associtionn between league goals per game and league competitiveness, and a positive association between league goals per game and the signifcance of home advantage.

Background

Soccer is the most popular game in the world. Played by billions of people, almost everyone has some knowledge of the game and their country's league. However, soccer is played differently throughout the world, with certain countries known for their flair and others known for their discipline. The goal of our project is to explore how these differences manifest themselves in data. The main data we will be working on is a collection of match results from eleven different European soccer leagues over a number of years (roughly 2010-2016). We will combine this data with information on player abilities, team play styles, and betting odds, to create insights into how these top leagues differ from one another. The infomation we create in this project could be used to aid the accuracy of a model used to predict soccer results.

Aims

The aim of the first approach was to find correlstions between metrics across Europe's top soccer leagues. The metrics created were league competitiveness, scoring rate, and significance of home advantage. Competitiveness was evaluated with two measurements: the number of unique teams finishing in top 4 positions, and the average spread of points per game over all season. Significance of home advantage was evaluated with two measurements, as well. The first being the average difference between goal difference at home and goal difference away from home. The second being the avgerage difference between points per game at home and points per game away from home. The third metric is the scoring rate in each league, which is an average of goals per game over all seasons.

The second approach is the player and team attributes correlated with success according to FIFA soccer player data. By focusing on the physical characteristics of players who had a 90 or higher overall rating we have found out that the physical characteristic is not straight related to the win rate or player's overall rating. But it still has little correlation that can lead to the win rate or overall rating a little bit higher. By using linear regression as the tool to predict results we have found out that a player's height and weight play an important role in the soccer game. In other words, a player's physical characteristic is going to increase the win rate which means players with better physical data had a higher chance to win.

Materials and methods

Datasets

The match data is contains the home and away teams involved in the match, the goals scored by each team, and the result information. This data comes from a kaggle project (https://www.kaggle.com/datasets/hugomathien/soccer) which aimed to collect data on European club soccer. Data was collected via web scraping, as all of this infomation is readily availbale online, but needed to be collected and stored. The relevant population is the eleven leagues for which matches are recorded, all of which are in Europe. We will focus on the biggest leagues, being England, Spain and Italy. The scope of inference is small, if anything, because this data is years old. If this were up to date data, then our finding would have some bearing on how these leagues will perform going

forward, but given that this data is at least five years old, the leagues would have had enough time to change that our finding will likely have little to no bearing on European soccer today. Data that we will pair with the match infomation is the player and team attributes, which are taken from Fifa, the most popular sports video game in the world. Fifa is a soccer video game which maticulously tracks player abilities and team styles throughout the year. While these stats are obviously not a direct representaion of real life, they provide a good idea of how teams should play on paper. This data was also web scraped, and was found on the same kaggle page.

Data Structure

Dataset 1 :

The following data demonstrate the question of winning and scoring patterns the same across different European soccer leagues?(home advantage, goals scored, variability of top teams each year).

> **Table 1**: Variable descriptions and units for each variable in the dataset

Variable	Description	Units
country	Country of league	Nan
season	Year of season	Year
team	Team name	NaN
home_GF	Goals for per game while playing at home	goals per game
home_GA	Goals conceded per game while playing at home	goals per game
home_PF	Points earned per game while playing at home	points per game
home_PA	Points earned per game by opponent while playing at home	points per game
home_GD	Goal difference per game while playing at home	goals per game
away_GF	Goals for per game while playing away from home	goals per game
away_GA	Goals conceded per game while playing away from home	goals per game
away_PF	Points earned per game while playing away from home	points per game
away_PA	Points earned per game by opponent while playing away from home	points per game
away_GD	Goal difference per game while playing away from home	goals per game

> **Example Data:** This raw data took a signifcant amount of manipulation to create the tables used in the analysis of aim #1. An example of the data is below:

country	season	team	home_GF	home_GA	home_PF	home_PA	home_GD	away_GA	away_GF	away_PA	away_PF	away_GD
Belgium	2008/2009	Beerschot AC	0.941176	0.558824	0.882353	0.529412	0.382353	0.676471	0.352941	0.970588	0.352941	-0.323529
		Club Brugge KV	1.088235	0.676471	1.029412	0.411765	0.411765	0.794118	0.647059	0.705882	0.705882	-0.147059
		FCV Dender EH	0.617647	0.676471	0.558824	0.823529	-0.058824	1.029412	0.676471	0.911765	0.470588	-0.352941
		KAA Gent	0.911765	0.441176	0.882353	0.529412	0.470588	0.794118	1.058824	0.500000	0.852941	0.264706
		KRC Genk	0.676471	0.617647	0.764706	0.588235	0.058824	0.882353	0.735294	0.794118	0.705882	-0.147059

Dataset 2:

The following dataset focus on real player's physical characteristic and FIFA player data

> **Table 2**: Variable descriptions and units for each variable in the dataset

Variable	Description	Units
ID	ID of the Player	NaN
Name	Name of the Player	NaN
Height	Height of the Player	Centimeters (cm)
Weight	Weight of the Player	Pound(lb)
date	Date of the data update	NaN
overall_rating	Overall Rating of the Player	Score from 1 - 100
preferred_foot	Player's preferred foot	NaN
attacking_work_rate	Attacking Rating	Low, Medium, High
defensive_work_rate	Defensive Rating	Low, Medium, High

> **Table 3**: Example of relative abundance data.

ID	Name	Height	Weight	date	overall_rating	preferred_foot	attacking_work_rate	defensive_work_rate
30723	Alessandro Nesta	187.96	174	2007-08-30 00:00:00	91.0	right	medium	high
30723	Alessandro Nesta	187.96	174	2007-02-22 00:00:00	91.0	right	medium	high
30955	Andres Iniesta	170.18	150	2013-06-07 00:00:00	90.0	right	high	medium
30955	Andres Iniesta	170.18	150	2013-05-24 00:00:00	90.0	right	high	medium
30955	Andres Iniesta	170.18	150	2013-05-17 00:00:00	90.0	right	high	medium

Methods

The exploratory analysis aimed at soccer players' physical characteristics and FIFA soccer player data. By applying data manipulation the dataset transferred to focus on the best player in FIFA against a real player's physical characteristics. Subsequently, multiple linear regression was performed, and learning the key features from the data to predict the overall rating by player's height and weight. Also, by applying principal components analysis to simplify datasets and determine the data features to demonstrate the result. In addition, data visualization was made to understand the data better. (Data visualization package Altair was used).

The analysis on league metrics began with a lot of exploratory analysis and data manipulation. The exploratory analysis lead me to chose the metrics that would be involved in the correlation matrix. Again, the metrics are competativeness, scoring rate, and signifcance of home advantage. These metrics were calculated using aggreation methods in Pandas and were visualized using Altair. Once metrics were calculated for each league, the correlation between each measuremnt was found, and visualized with a heatmap to highlight positive an dnegative relationships.

Results

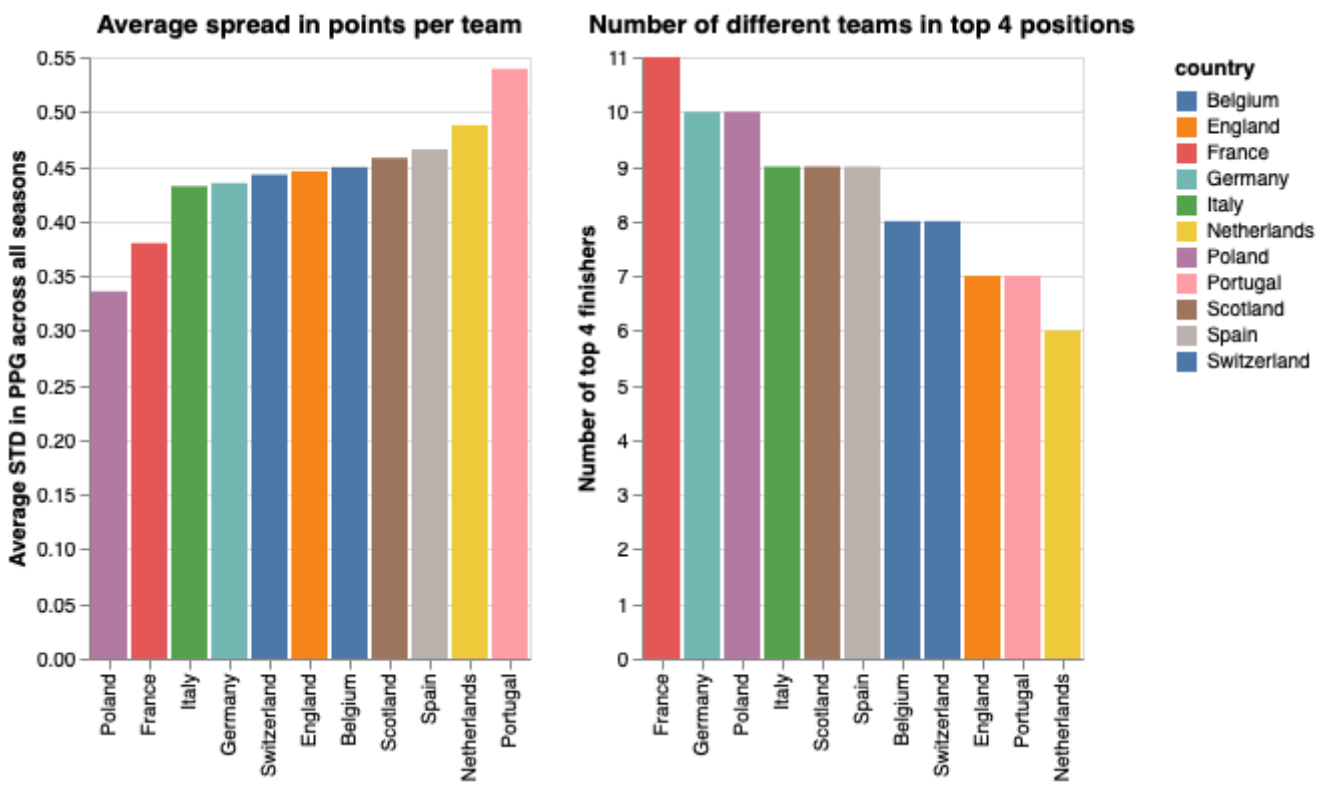
League metrics

Assessing League Competitivenss

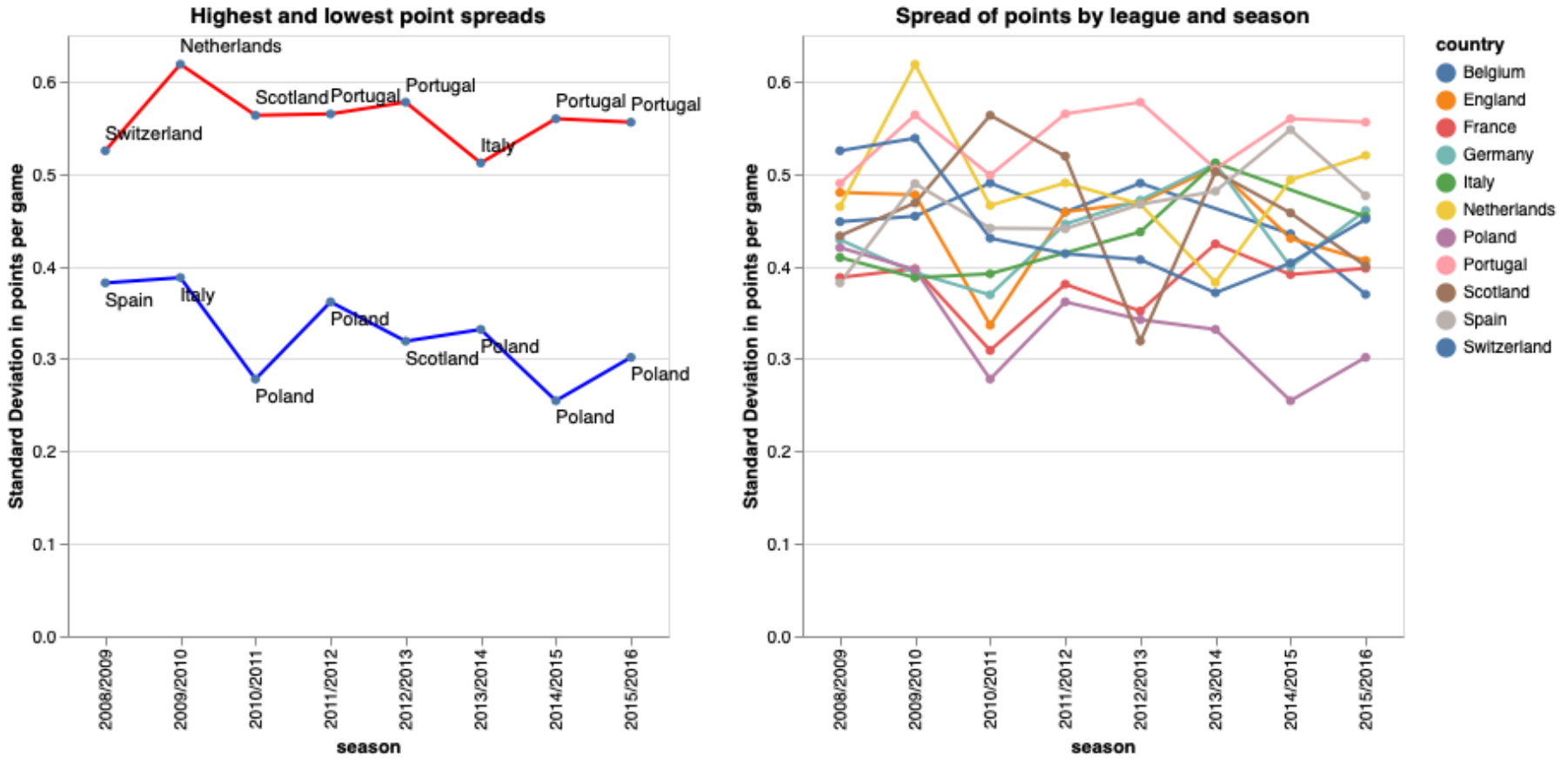
This section aims to visualize the comparison of competition levels and the distribution of talent amogst teams in Europe's top soccer leagues.

Competitivenss is first measured by the number of different teams which have finished in the top 4 positions in each league during this time period. A greater number of unique top 4 finishers indicates a league in which success (ie. talent) is widely distributed across a greater number of teams, and hence indicates a more competitive league. The second metric used to assess league competitiveness is the standard deviation of points in each leagues final standing for each season. In this way, a league with closely grouped teams suggest a competitive league, whereas a league that is dominated by the top teams (a two or three horse race) would be indicative of a league with low competition levels.

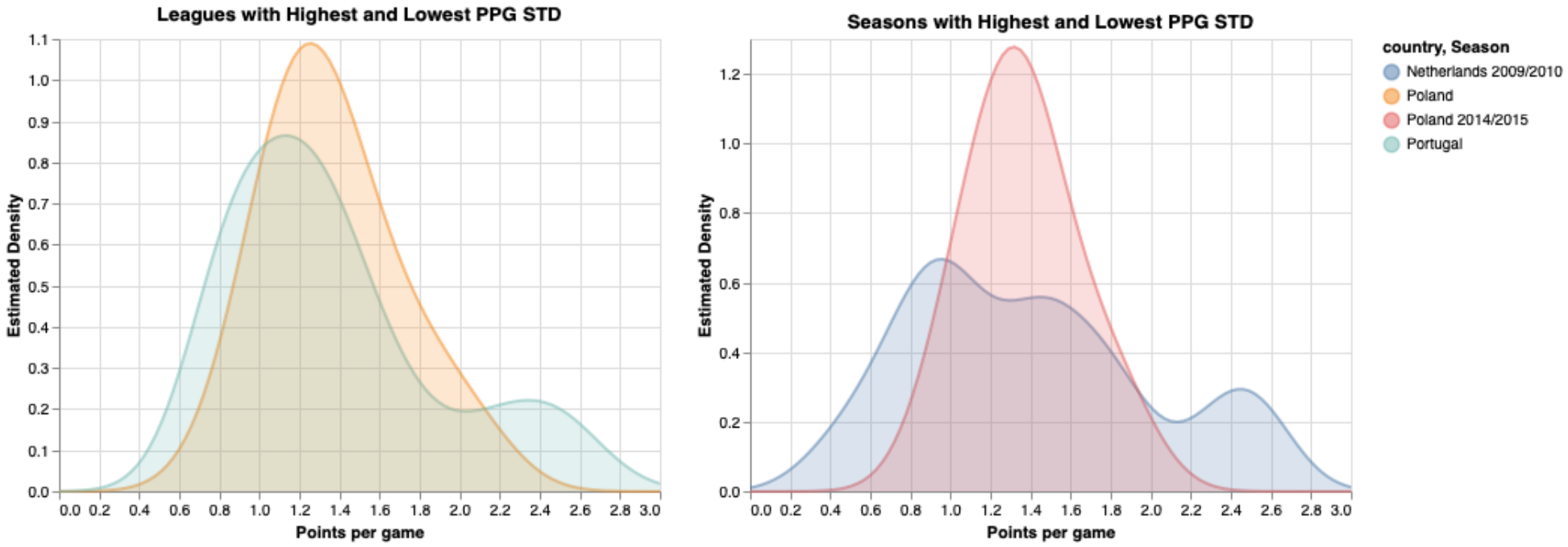
> **Figure 1**: Comparison of competitiveness among European soccer leagues:



The following figures show how the spread in points per game is a measurement of league competitiveness: > **Figure 2**: The spread in points per game varies across leagues and seasons:



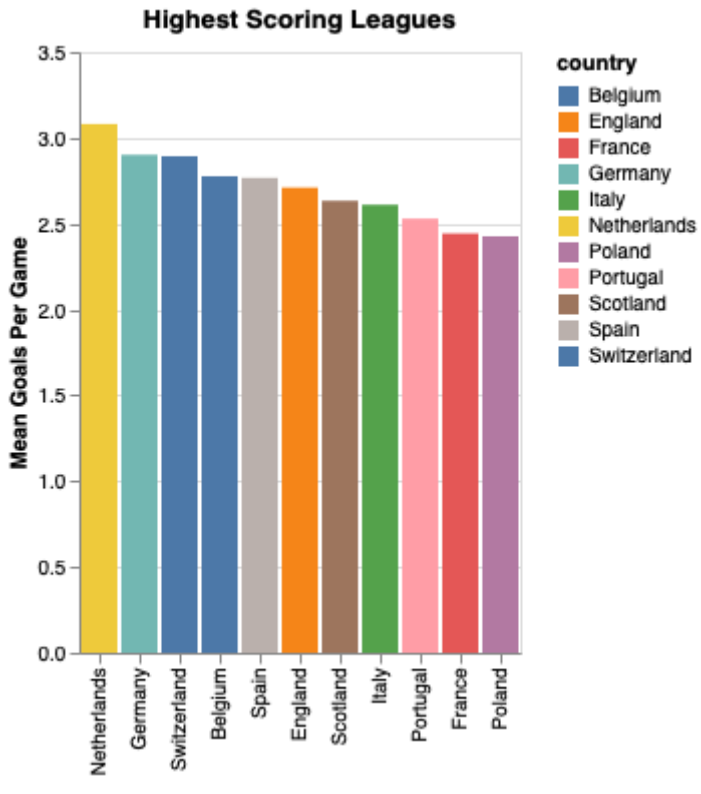
> **Figure 3**:The differences between a highly competitive league and a league which is dominated by just a few teams:



Highest Scoring Leagues

The second metric used to compare leagues is the rate of goals per game in each league. This is simply a measurement of the average of goals per game in each league over each season recorded.

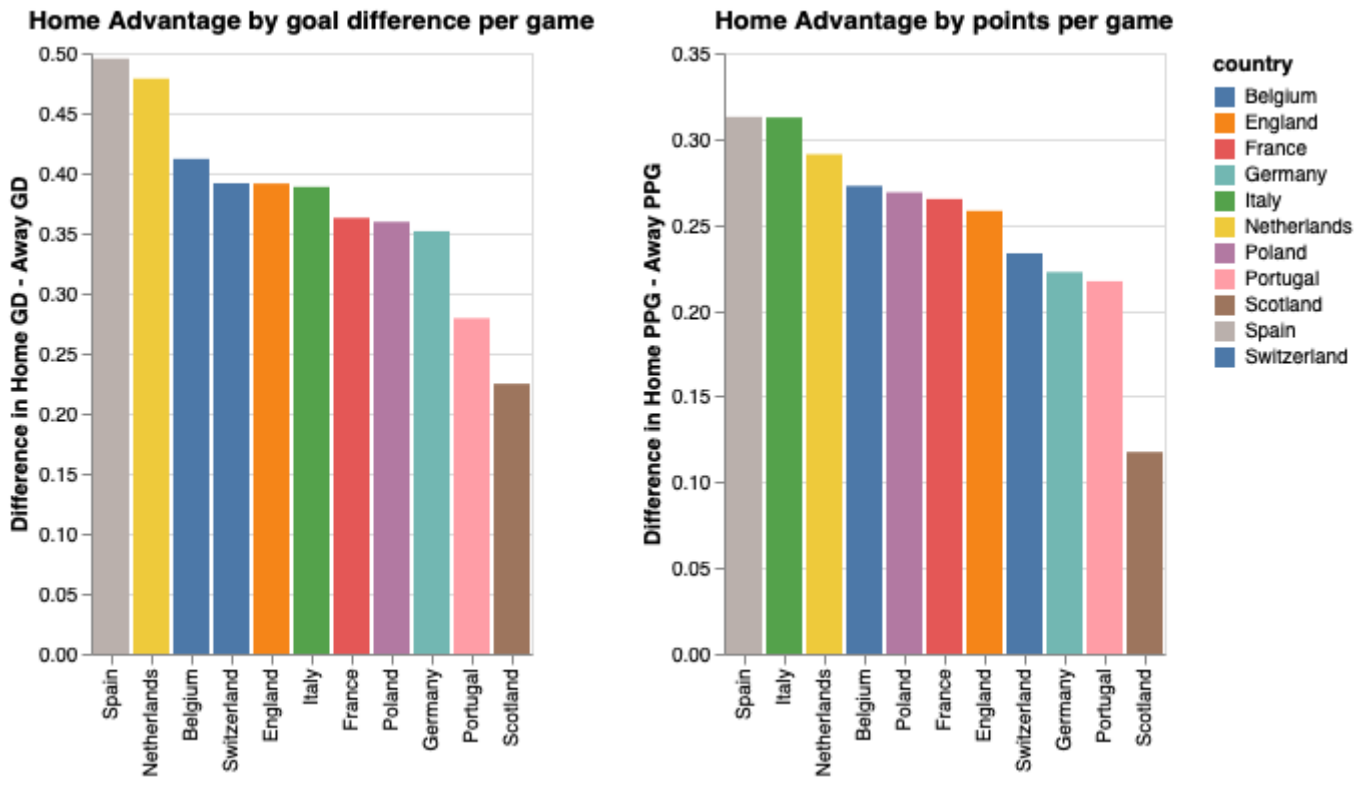
> **Figure 4**: Rate of scoring in Europe's top soccer leagues:



Significance of Home Advantage

The third metric used to compare leagues is the signifcance of home advantage. Home advantage was assessed in two ways: comparing the goal difference (goals scored - goals conceded) for home teams and for away teams; comparing the points per game for home teams and away teams.

> **Figure 5**: Comparison of home advantage across Europe's top soccer leagues:



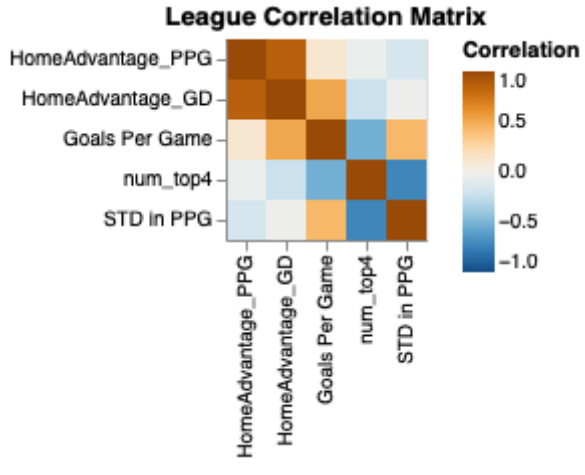
The first table can be interpretted in the following manner: in Spain, the home team scores an average of ~0.5 goals more than the away team every game in this time period. The second table shows the comparitive success of home teams against away teams in each league.

It is worth noting that there is a clear home advantage in every league, and that the focus of this analysis is on which leagues have the biggest home advantage.

Correlation of League Metrics

The crux of this analysis is the correlation between these metrics. The correlation matrix heatmap below visualizes the relationships between each metric:

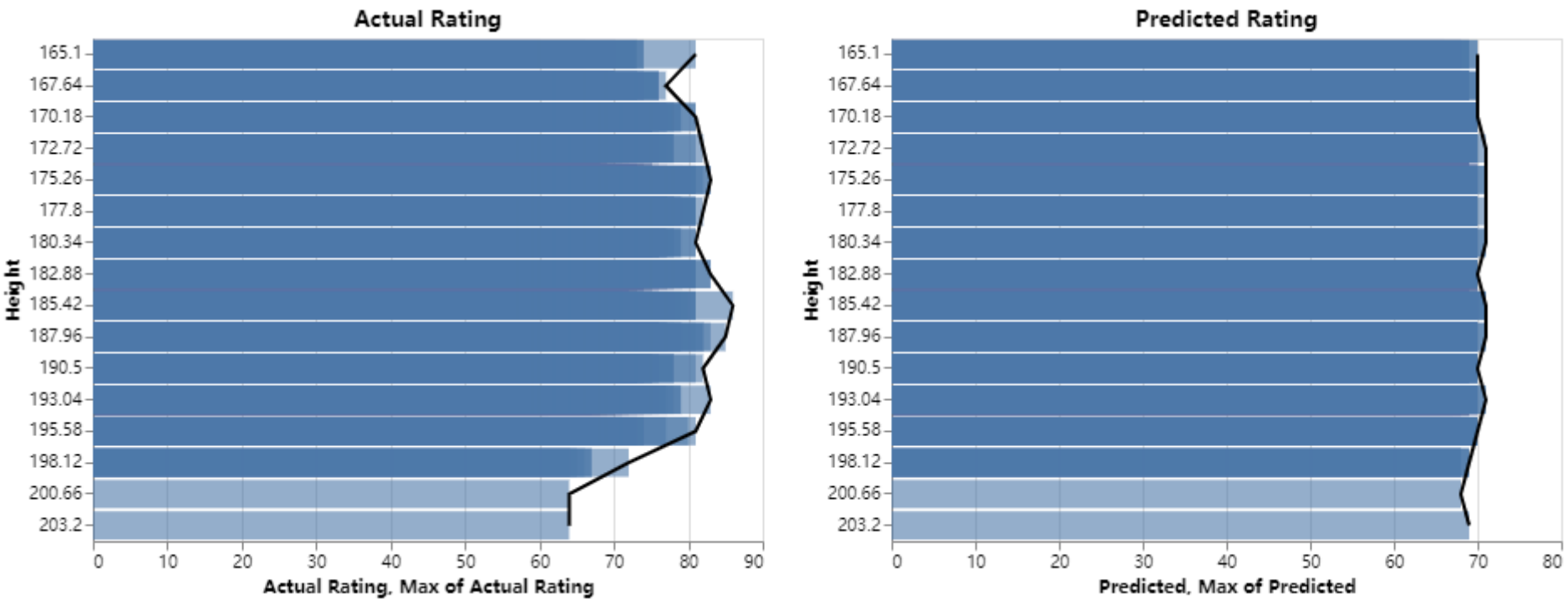
> **Figure 6**: Correlation heatmap visualizing the relationships between league metrics:



Relationship between height and weight against overall rating of soccer player

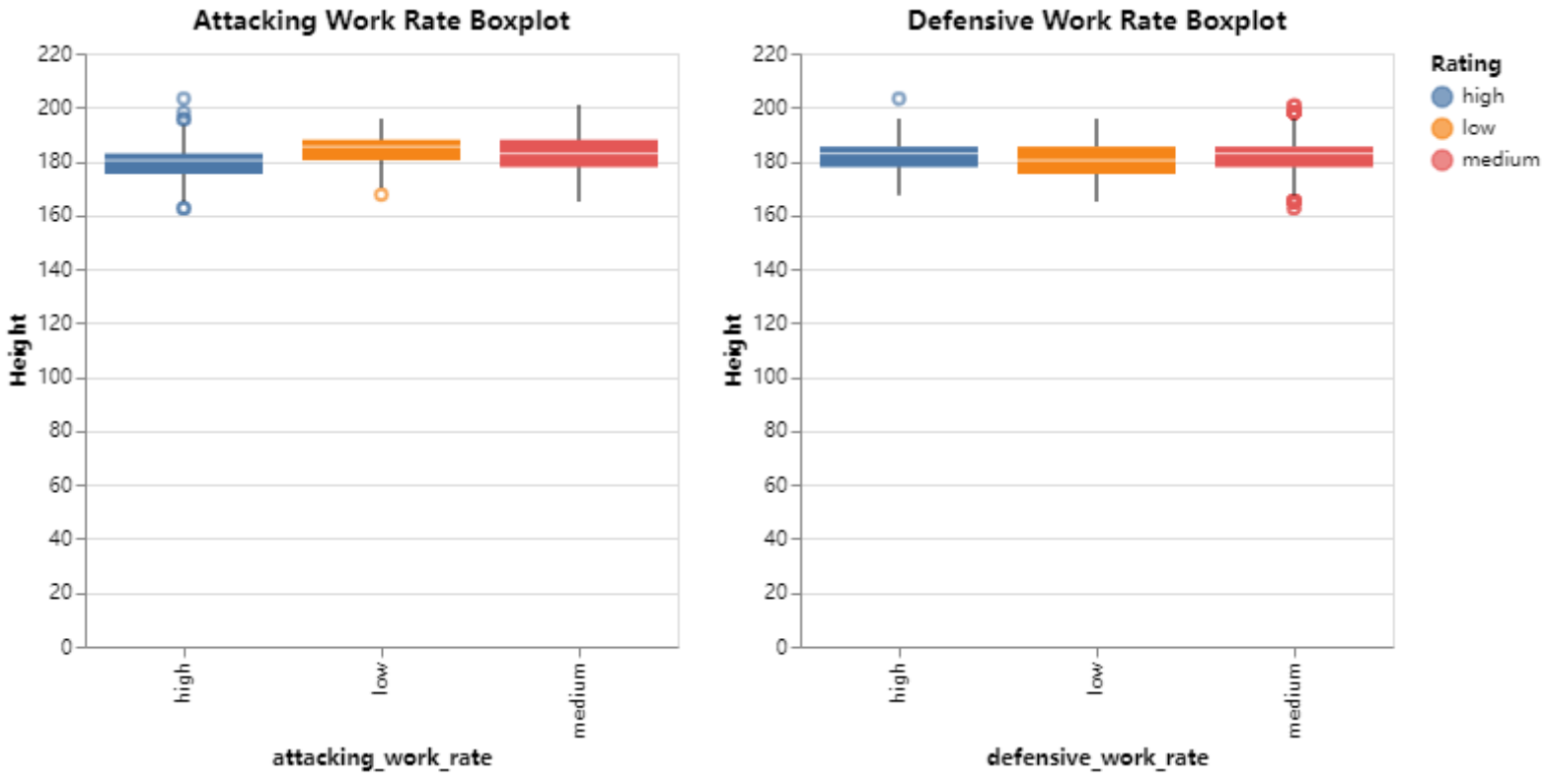
By applying multiple linear regression has demonstrated that the height is one of the factor that will influence the overall rating for soccer players. After using linear regression model to train with data features we found that certain height are having larger amount of high overall rating than others. The actual rating system had a major curve between 182cm to 187cm and the linear regression model had a small curve with the same height. In other words, it demonstrates that player's height will influence their rating. The purpose of doing this is because we understand that there are many other elements that will affect the rating. However, when using linear regression we only focus on the data and data features which means by only looking at player's height we found out that certain height of people always perform better than others.

> **Figure 7** : Figure 2 demonstrates that whether the real data or predicted data both show that players arround 182cm - 187cm had a better overall rating than others.



Further analysis of how player physical characteristics affect attacking and defensive work rate during the game. The data focus on the player's height and their attacking/defensive work rating. By using boxplot, it demonstrates the minimum, maximum, and median height of players from different stage.

> **Figure 8**: Figure 3 shows the different level of attacking / defensive work rate of players. The graph shows the median of different stages and it demonstrates the average height for different rating. According to Figure 3, the median height of high attacking rating is arround 180cm and the high defensive rating is arrount 182cm. In other words, this data visualization has proved that player's physical characteristic will be correlated with overall rating.



Discussion

The correaltion matrix highlights some interesting relationships in Europe's top soccer leagues.

One signifcant finding is the relationship between league competitiveness and league goals per game. Goals per game is apparently negatively associated with league competitiveness. Meaning, an increase in goals per game leads to a decrease in the number of unique top 4 teams, and an increase in the spread of points per game. Therefore, this analysis shows that higher scoring leagues will tended to have a more unequal distribution of talent and success among teams in this period.

Another interesting finding is that, while the association between home advantage in terms of points per game and goals per game is not particularly strong, the association between home advatage in terms of goal difference and goals per game is significant. This suggests that goals per game is positively associated with a significant home advantage. In other words, the leagues in which home advantage made a big difference tended to be the higher scoring leagues. This may be because home advantage manifests itself more when teams are scoring 2 or 3 goals per game, rather than fighting out 0-0 draws.

There is not a particularly strong correlation between home advantage and league competitiveness, but this analysis suggests that home advantage will be slightly more significant among competitive leagues. Which makes sense, becuase it doesn't matter where the match is played if one team is far more talented than the other. One could expect home advantage to have a more significant effect if the involved teams are otherwise evenly matched.

According to our dataset, different heights had different advantages which means some of the heights are good at dribbling and some others will be good at free-kick, or passing. Because a soccer game is based on the whole team that means not only the player who got the goal performs well. Good defense and good passing also play an important role in the soccer game. Besides, if we focus on the player's height in certain positions we can figure out a better physical characteristic that has a better overall rating. In other words, if we go deeper to compare all the data on the same position we can find out what is the best physical characteristic for each position. In conclusion, we will be able to predict the overall rating by looking at the player's physical characteristics in each position.