Matt Dahle

Stat 479 Final Project Writeup

Analysis of the Countries Producing Players in the Champions League

Background

The purpose of this analysis is to look at certain factors in order to determine the countries that produce the best soccer players in the Champions League. The data I used is found at https://www.kaggle.com/karangadiya/fifa19, and then copied over to my GitHub account at https://raw.githubusercontent.com/mattdahle/STAT479DATA/master/FIFADATA.csv. The data set has a lot of variables since it comes from the popular video game FIFA 19, so I kept the variables I deemed necessary and only used 6 variables from the data set to prepare my analysis (Name, Age, Country, Rating, Club, and Wage).

Part 1: Reading in Data

The data set used was read in directly from the GitHub source using a filename mycsv command. I only read in the first few variables that I needed and named them accordingly. All variables were formatted with correct type and length. I also created the wage variable in this process. This was necessary because the wage value given in the data set was given in a string format that was displayed in Euros. I used an input(compress()) command to change the variable from string to numeric, and multiplied it by 1.12 (The conversion from Euros to U.S. Dollars) so the wage was formatted properly and easy to understand. For clarification, the wage variable is the *weekly wage in thousands of U.S. Dollars*. I then dropped the unnecessary variables and used a proc contents function to make sure all variable types were correct. Variables kept at this point are the 6 named in the background section of the paper

Part 2: Data Wrangling

There are over 18,000 players in the FIFA organization worldwide, and I wanted to focus on just the best players for my analysis. In order to do this, I created a new data set that only contained players on teams participating in the Champions League. The Champions League is an annual soccer tournament that showcases the best teams from each country's top leagues (Premier League in the UK, La Liga in Spain, etc). Furthermore, I only wanted to look at the top teams in the champions league, so I only kept the final 8 teams in the tournament from this year. In my data step, I used if/then/else logic to create a Boolean like variable "TopQualifier" to indicate if a player was on a top 8 team. I used subsetting to drop observations that were not in TopQualifer.

After this data step, I used a proc freq function to see which countries all the players on the top teams came from, and to make sure my data step worked correctly. I noticed there was still a lot of countries on this list, and a lot of countries were only producing a few players or even just 1 player. For my analysis of the data to not be skewed by these observations, I used a proc sql function to create another data set that only keeps players from countries that produce at least 10 players in the Champions League. After this, I used another proc freq function to see the list of

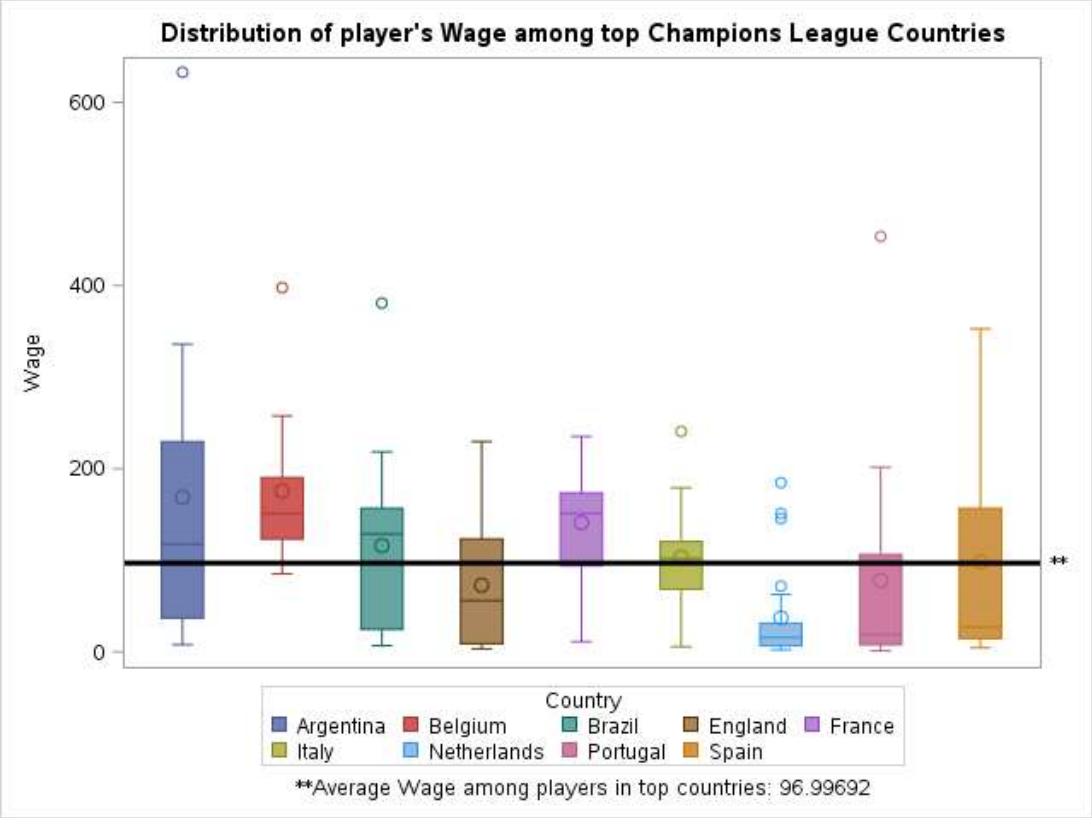
countries in the data set. After all the data cleansing, I had cut the number of observations from over 18,000 to 182, all from the top 9 countries in terms of player production.

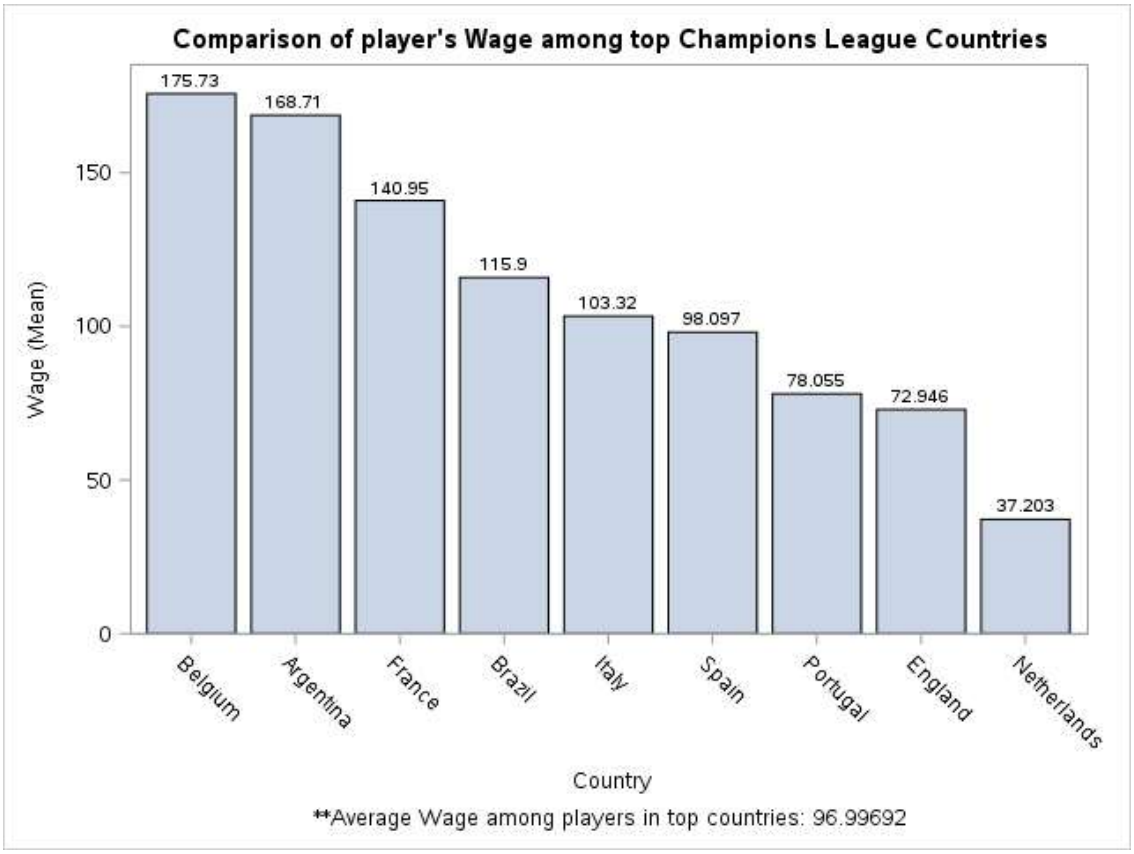
Part 3: Data Analysis

The rating variable from the FIFA 19 video game was interesting to me because it is essentially an aggregate score of the player's overall abilities and talent. To learn more about what factors effect this rating I performed a regression of rating on country, wage, and age. Looking at the p-values, all 3 of the variables were significant in estimating a players rating (p-value <0.01). The R^2 value of the regression was 0.76.

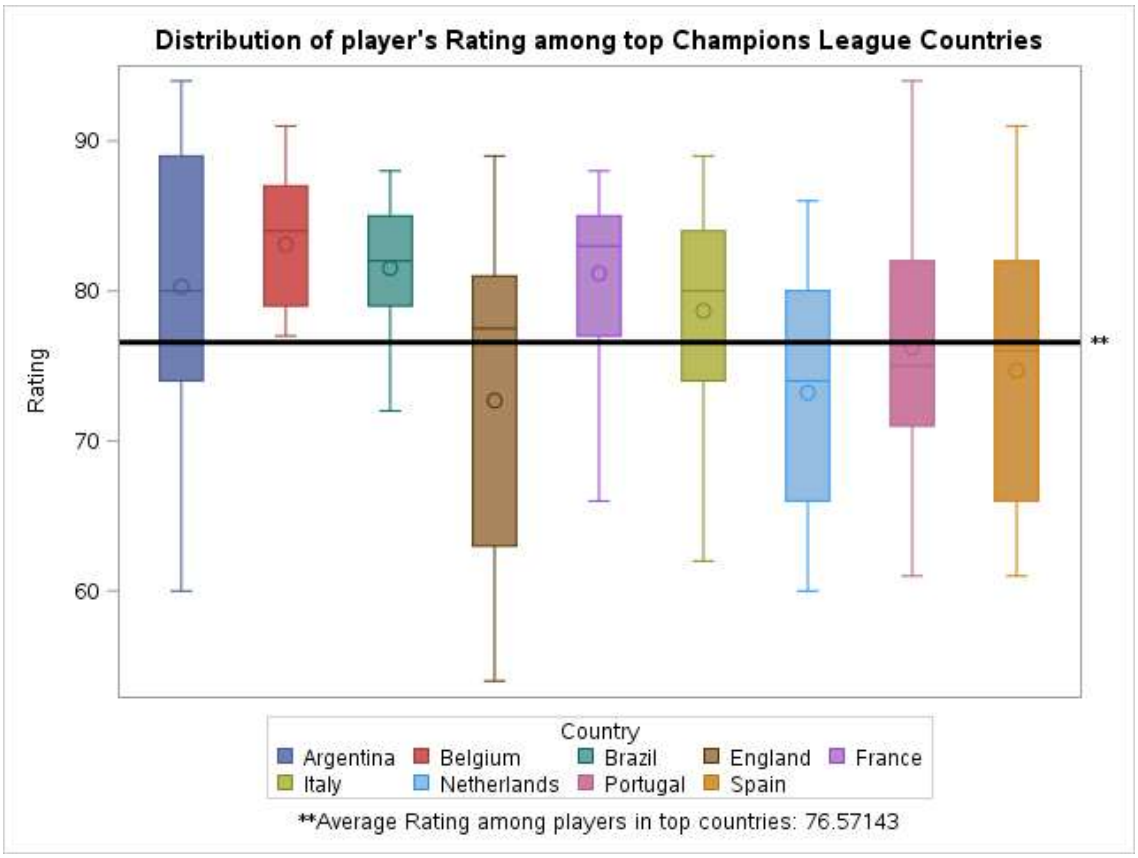
I now wanted to visualize the data and created a couple vbox and vbar plots using macros. The vbox plots show the distribution of wage and rating between all countries, and the vbar plots show the same in descending order with their average values displayed. We can see in both examples of rating and wage that Belgium seems to produce the highest earning and highest rated players in the Champions League. The average wage of Belgian players is about $175,000 per week, and the average rating of the players is an 83.1/100. All graphs are displayed below. I conclude from this analysis that in the Champions League, Belgium is the best country in producing top rated and top paid players.

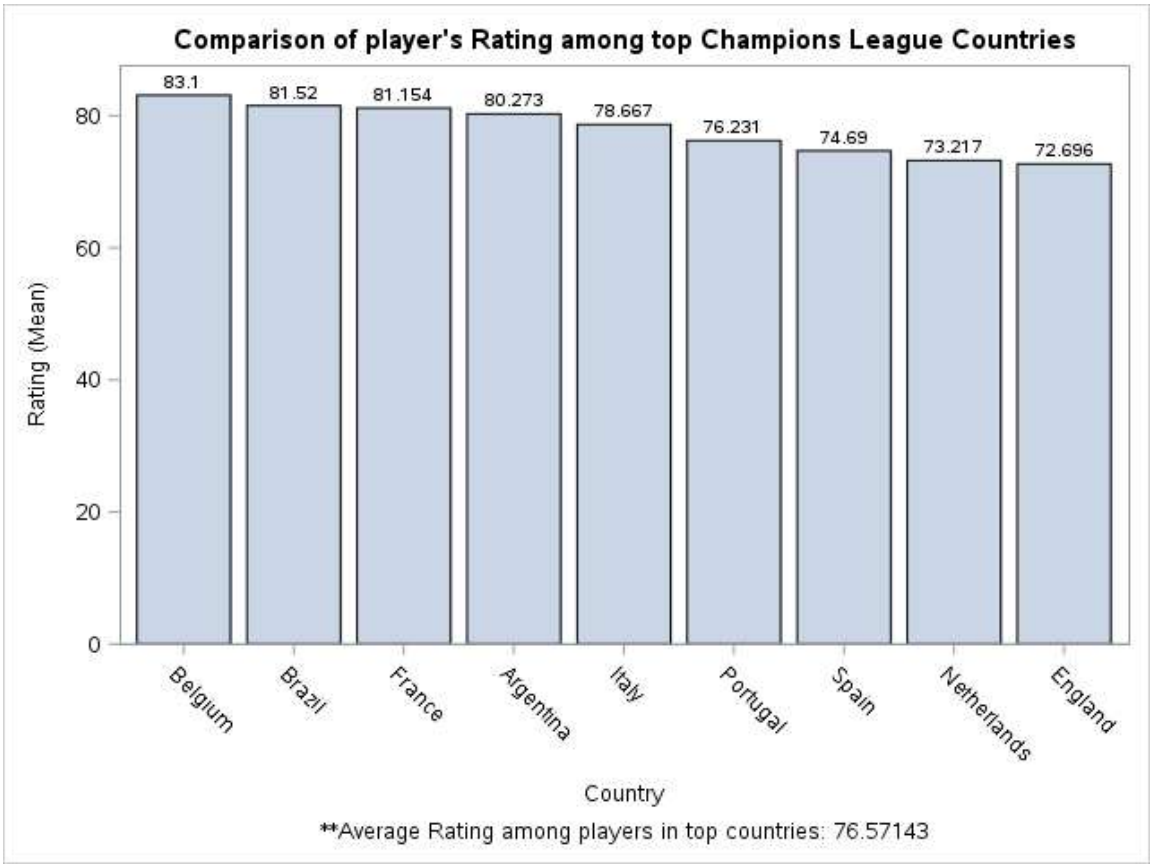
Relationships of Country and Wage:





Relationships of Country and Rating:




Comparison of player's Rating among top Champions League Countries
Rating (Mean)
0
20
40
60
80
83.1
81.52
81.154
80.273
78.667
76.231
74.69
73.217
72.696
Belgium
Brazil
France
Argentina
Italy
Portugal
Spain
Netherlands
England
Country
**Average Rating among players in top countries: 76.57143