*Cross-tabulations, histograms, boxplots,*

*jitter plots, hypothesis testing, and regression.*

**Final**

**Project**

Fp

ALY6010 Probability Theory and Introductory Statistics

Milestone 3

**PREPERATION:**

By: John DiSessa

For: Professor Goulding

On: April 11th, 2021

Milestone 1

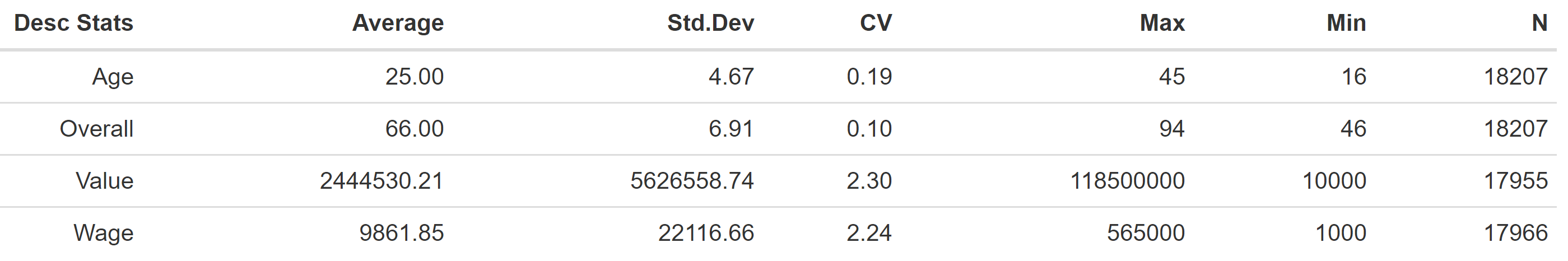
Introduction

For my final project, I decided to analyze the Fifa 19 dataset from the course files. After cleaning up the data, there were 18,207 rows of data and 62 variables. Many variables were discrete such as Age, Overall, and Potential. Some variables were continuous such as Value and Wage. There was even a mix of categorical variables such as the Player Nationality, Club, and Preferred Foot. Before I uploaded the “fifa19\_players.csv” dataset to R, I cleaned it up in several ways. First, I removed all of the columns that gave player overall ratings per position. I did not think this was useful since I would not need to analyze Lionel Messi’s defensive ratings, for example. Then I standardized the Value and Wage variables into their full euro amounts. Some values were “300M” for 300 million euros and some values were “300K” for 300 thousand euros. Now those variables read “300,000,000” and “300,000” respectively. I also made sure that any blank values were kept as blank and did not turn to zeros. Players with 0 for their value are not worthless, they just simply do not have a value listed from the game developers.

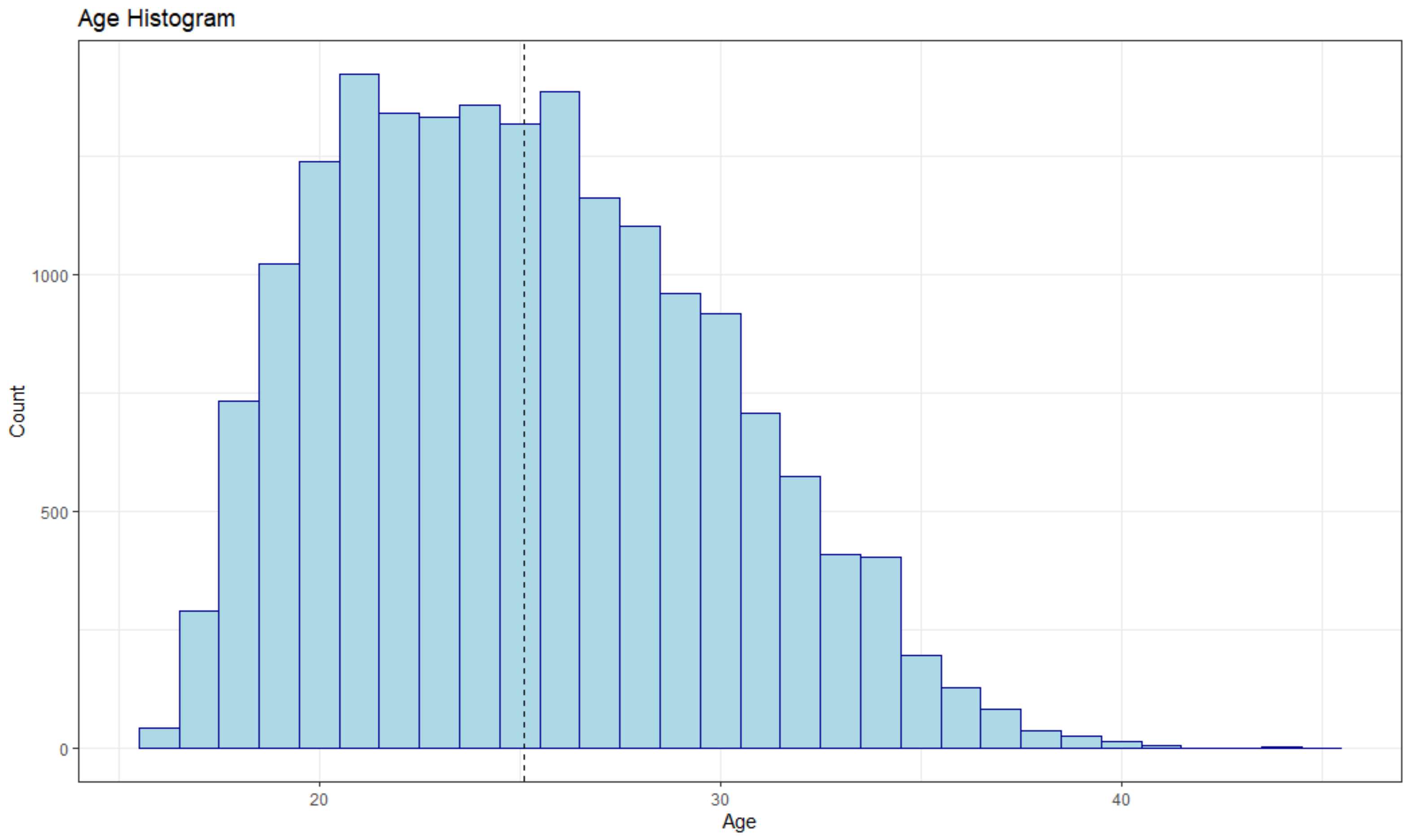
I chose this data because the Fifa series is my favorite sports franchise in video game history. I personally have spent hundreds of hours playing Fifa with most of them coming on Fifa 19 alone. I also love soccer so, in addition to analyzing my favorite game, I hope to draw conclusions that I can use in my real-world knowledge of soccer. Even though in theory it is incredible that EA was able to quantify every professional soccer on the planet, we have to understand that it is impossible to be completely accurate with their ratings due to personal bias in player evaluation and the constant improvement and changing of players talents.

Data Analysis

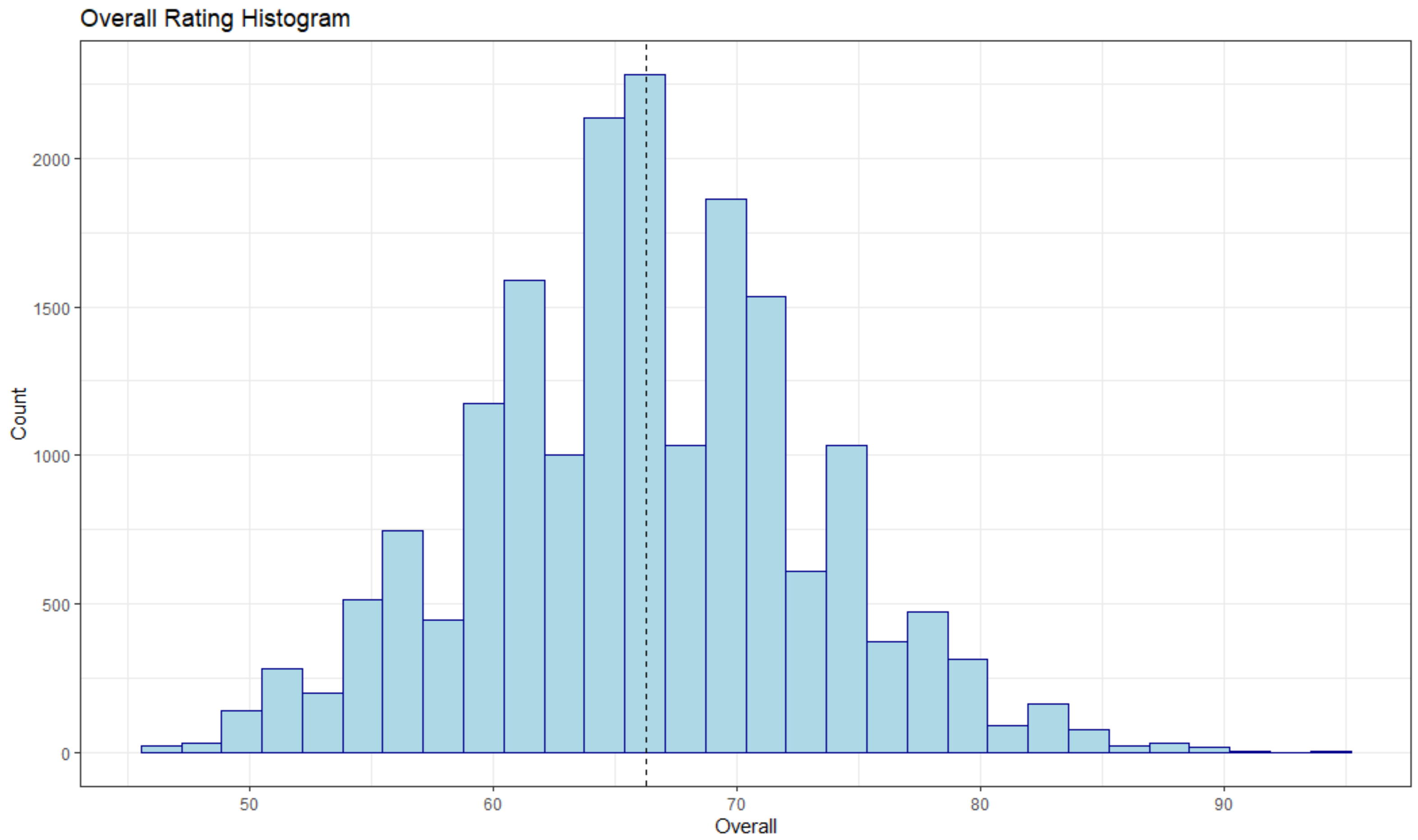
I first summarized the data by looking at four of the major variables (age, overall rating, value, and weekly wage).



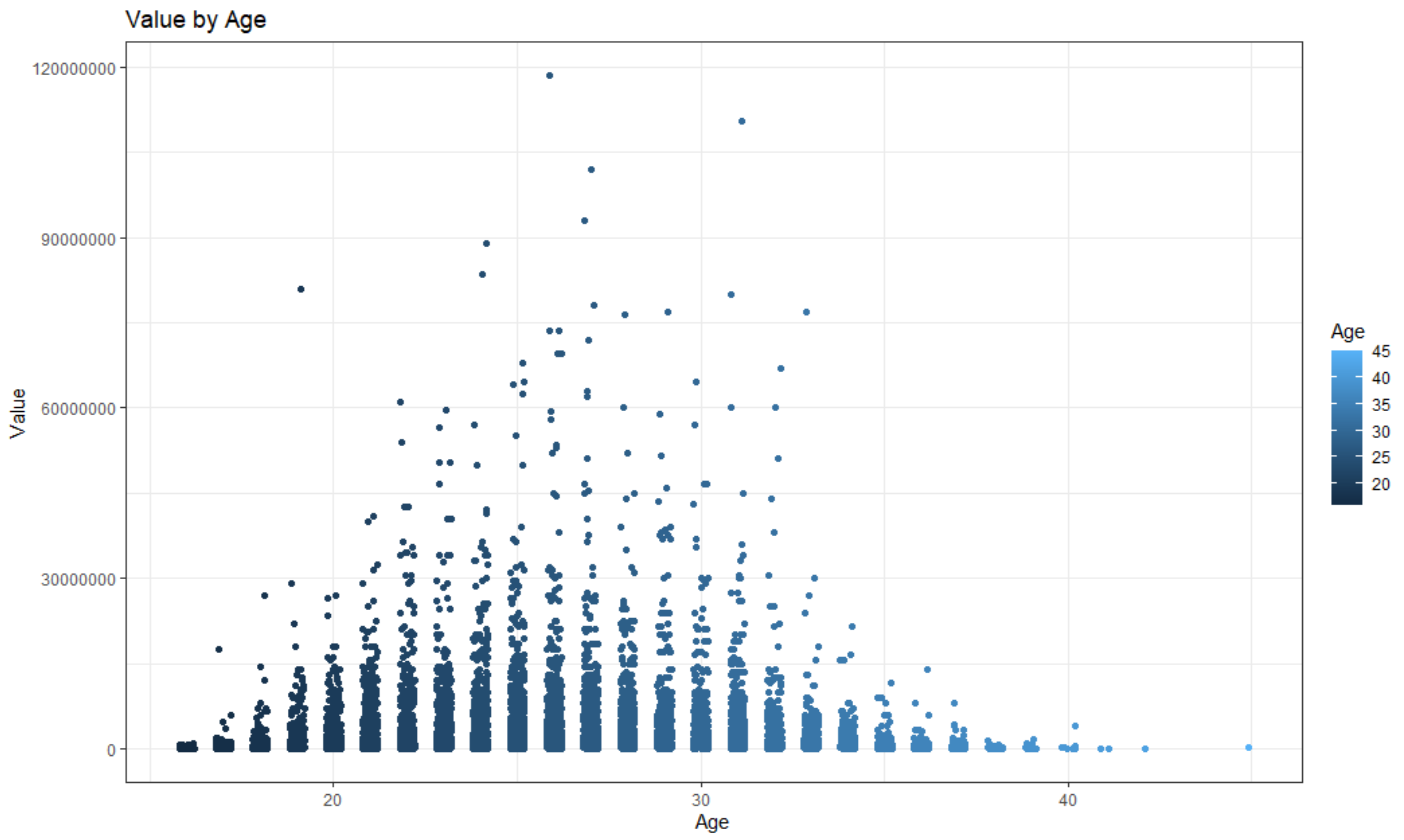
Fifa 19 players range from 16-45 years old, with the average being 25. I calculated the coefficient of variation in order to determine if my standard deviation of 4.67 years is high or low. Since the CV is only .19, this indicates that Age is not that variable. I created a histogram of ages in order to confirm this theory and Age looks about normally distributed, as seen below. Since the mean age of 25.12 is basically equal to the median age of 25, I can confidently say that age is not skewed. The dotted line on the graph below represents the mean.



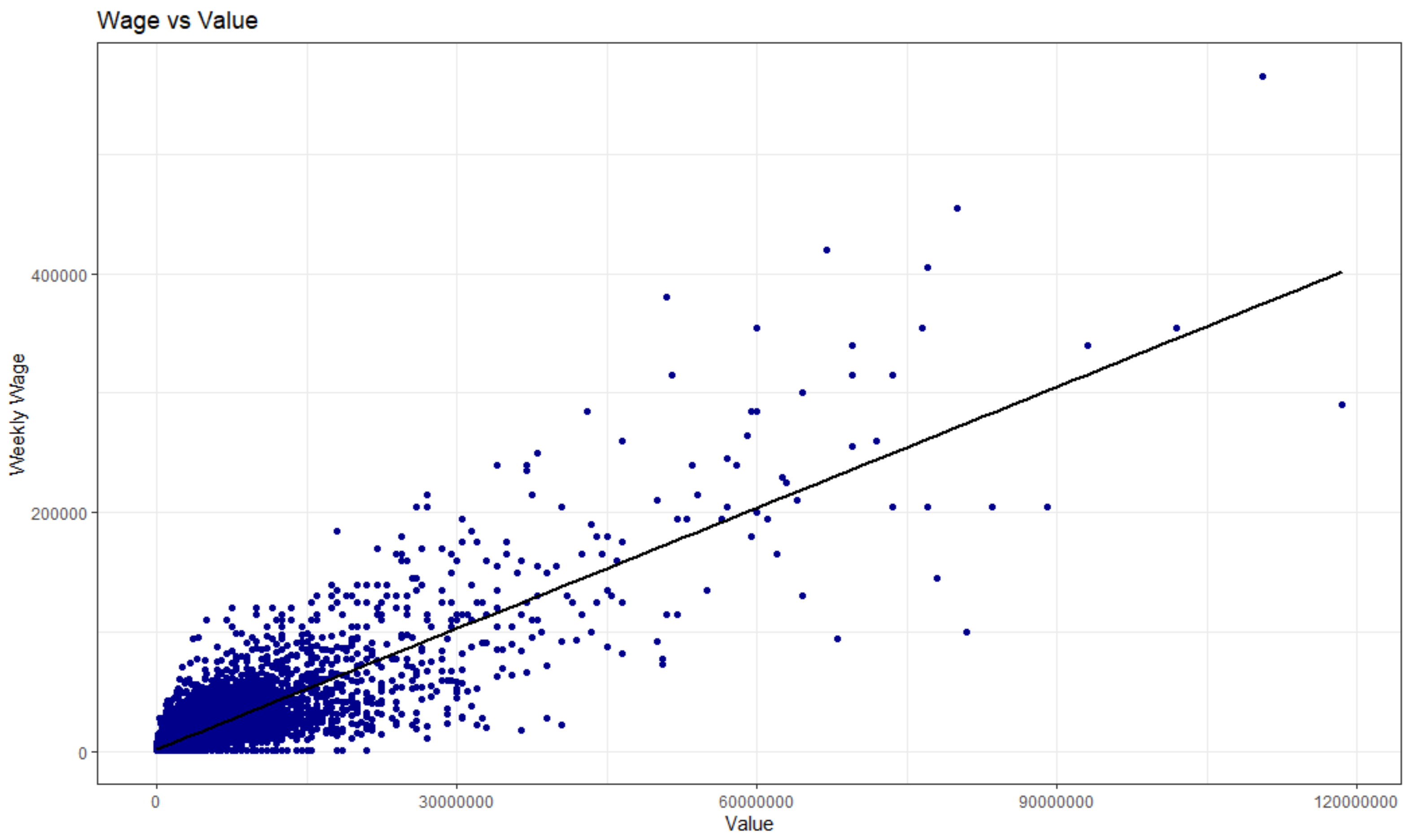
Next, I looked at the Overall variable which is the overall rating of each player, as determined by a formula that looks at each player’s skill rating in 34 categories. This is perhaps the most important variable since this is the first factor a video game player, like myself, looks at when determining if they want to use a certain player or team in the game. The mean, as represented by the dotted line in the histogram below, shows that the average player is rated 66.24 and the median is 66. Both of these values are virtually identical which means that Overall is normally distributed. Overall ratings are relatively arbitrary as assigned by the game developers, so it looks as though they did a great job making sure their ratings were balanced across the game.



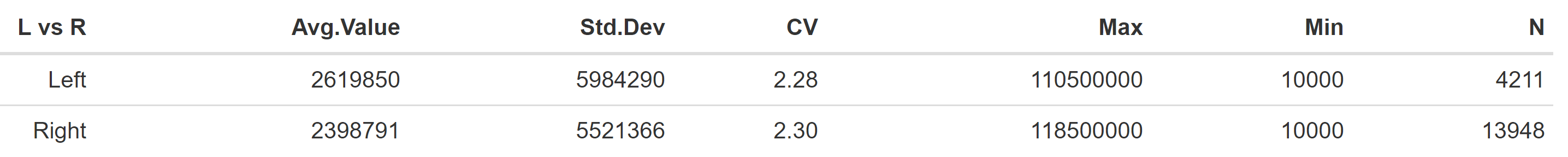
Now looking at player value, the coefficient of variation is 2.3 which is significantly greater than 1. My assumption was that it is more variable due to outliers like Lionel Messi and Cristiano Ronaldo. I created a boxplot of player values but it was unreadable due to the large number of outliers. This was not surprising because there are a few clubs that can afford to pay millions or billions of euros on players, but significantly many more clubs that are less funded. I then compared the mean value and median value to confirm this assumption. The mean player value is €2,444,530.21 but the median player value is only €700,000, indicating Value is heavily right skewed by the high value outliers. I also plotted value over age to see if my suspicion, that players have a ‘prime’ is true. Player value increases as they age until they hit their prime, then decrease with age as they lose some athleticism. While the following graph does seem to indicate that, there are too many outliers to come to concrete conclusions with the current dataset.



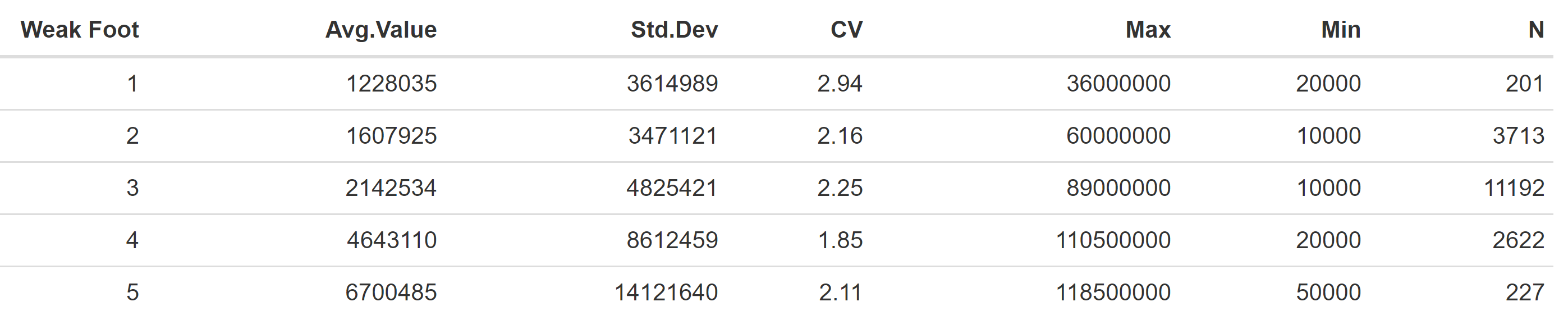
Next, I conducted similar analysis for a player’s weekly wage as I did for a player’s value. The mean wage is €9,861.85 but the median wage is €3,000 which again proves that wage is skewed heavily by outliers. Rather than plotting wage over time, I wanted to check the correlation between Wage and Value. The coefficient of correlation was .86 indicating a strong positive relationship where wage increases as value increases. The coefficient of determination was .74 so 74% of the variation in wage can be explained by value. The plot below also shows which players are overpaid and which players are underpaid. Players under the best fit line are underpaid, players on the best fit line are fairly paid, and players over the best fit line are overpaid. Interestingly, the highest value player in Fifa 19, Neymar Jr., is underpaid, the second highest value player, Lionel Messi, is overpaid, and the third highest value player, Kevin De Bruyne, is fairly paid.

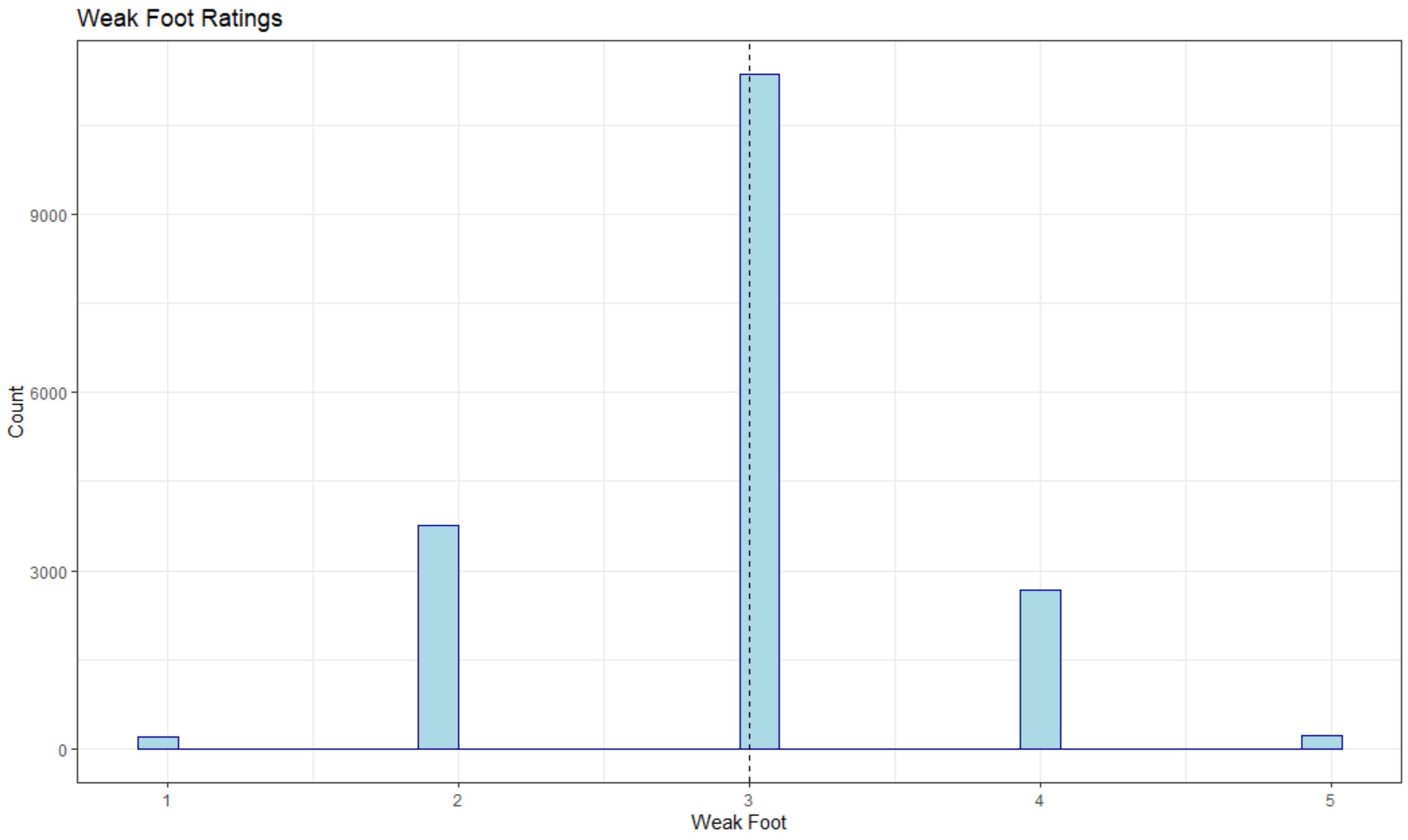


When analyzing subgroups, I first looked at left-footed vs right-footed players. I was interested to see if left-footed players are more valuable. Since the pitch can be divided in half with left-footed players playing primarily on the left side and vice versa for right-footed players, but there are three times as many right-footed players as left. This potentially would create some value for lefties. Even though the average value is slightly higher for lefties, the standard deviations and coefficient of variations are incredibly high so the results are insignificant.

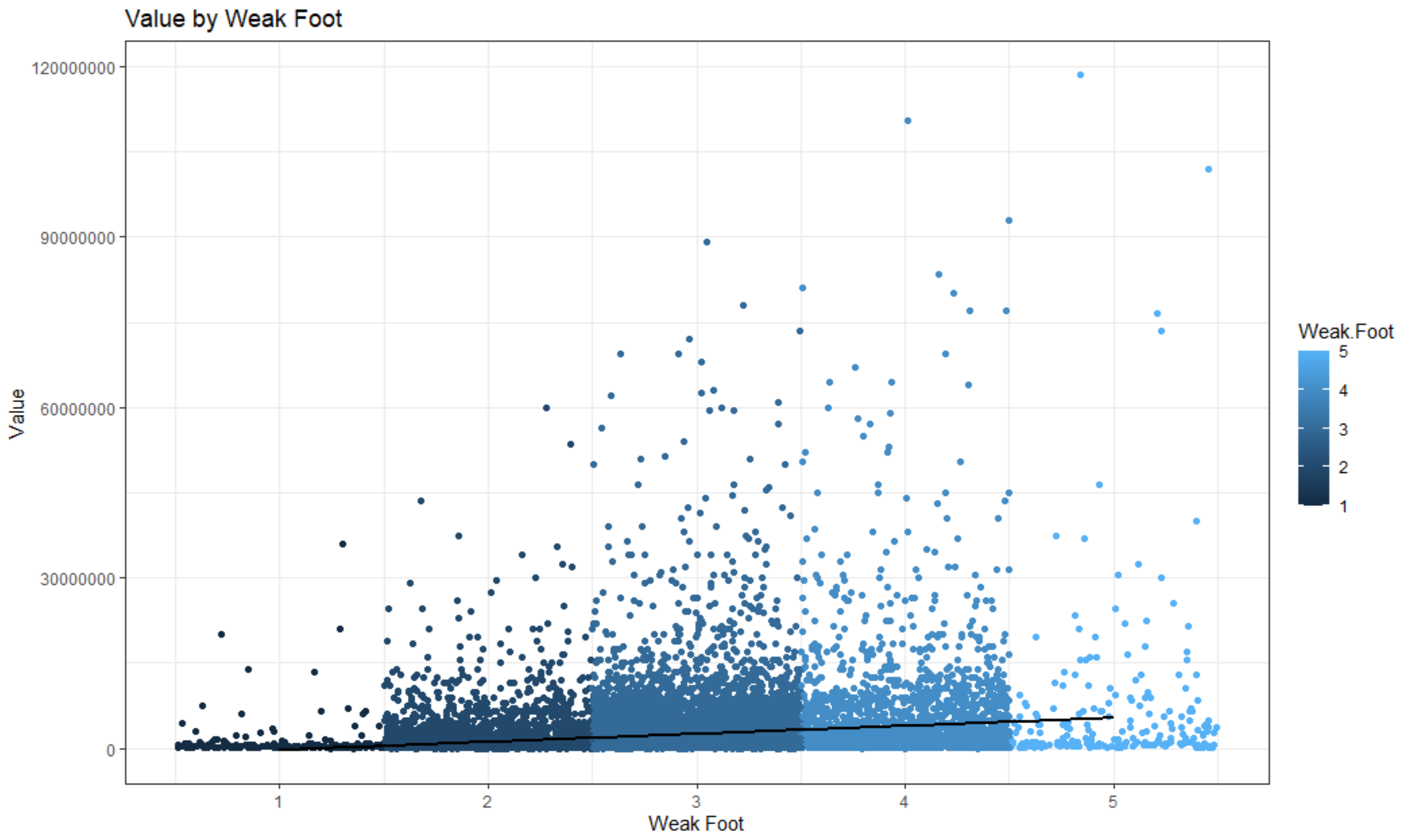


After the previous results were inconclusive, I thought maybe value could be found in players that are good with both feet instead of being one-foot dominant. The dataset categorizes players weak foot ratings on a scale from 1 to 5 with 5 being the best (players are equally good kicking with both feet).

As seen below, the developers gave players with average weak foot skills a 3. The developers made the game very balanced since weak foot skills are normally distributed with a median rating of 3 and a mean score of 2.94.



However, I noticed that the average player value increased as weak foot skills improved. I created the following jitter plot to see if there was significant correlation to match the increase in value. With a correlation coefficient of .17, there is barely a positive relationship between weak foot and value.



Summary

After analyzing many variables, the biggest takeaway is that the biggest superstars skew player value and wage. It is incredibly difficult and competitive to become a star player so it is not surprising to see only a handful of players have significantly high values. In the future, I think it would be helpful to remove some of the lowest rated players instead of the highest. Even though the highest rated players are outliers, they are the ones who get the most playing time and are worth analyzing. All of the lowest rated players barely play and could be removed to get a more realistic sense of value by age and value by weak foot skills. Another takeaway is that the game developers at EA not only looked at each player’s individual skills to assign them ratings, but they normalized the ratings to ensure even distribution of overall ratings and weak foot ratings. Future analysis could look into other variables the game developers put in effort to normalize, such as individual skill ratings or international reputation.

Milestone 2

Introduction

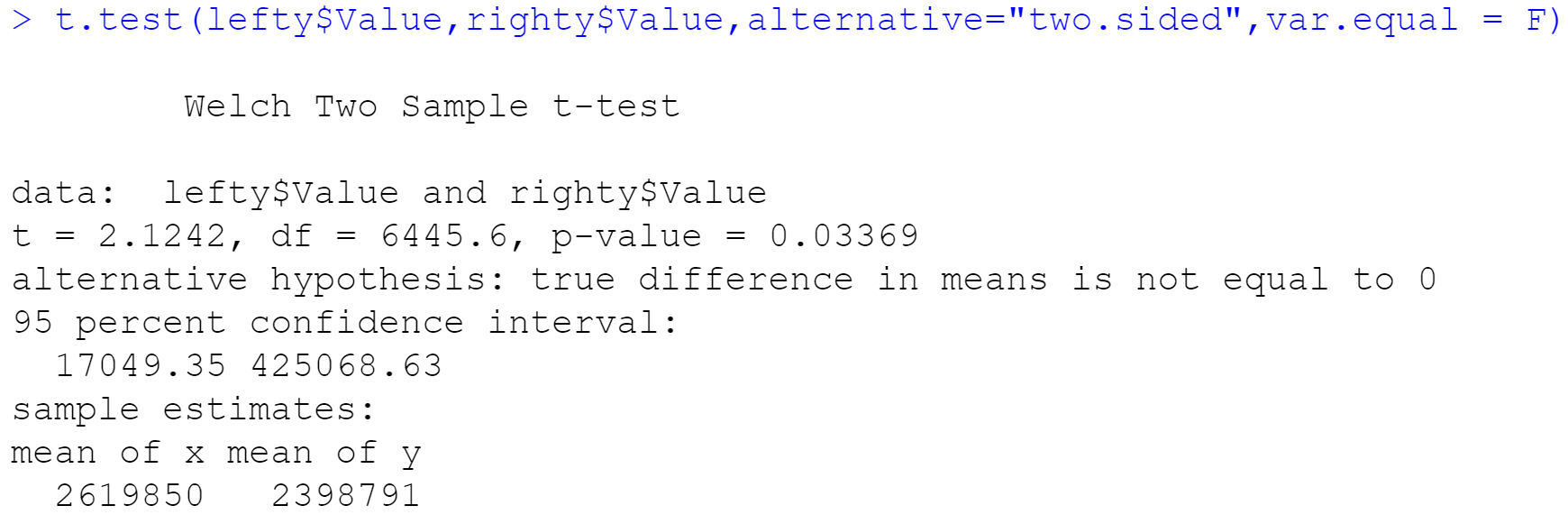
For milestone two, I used hypothesis testing to see if there were any statistically significant differences in player values based on certain variables. My analysis will answer the following questions: Are left footed players more valuable than right footed players? Are players with higher weak foot ratings more valuable than players with lower weak foot ratings? When clubs buy players in the transfer window, do they overpay based on Fifa 19’s values?

Analysis

Previous analysis in part one showed the mean value for left footed players was €2,619,850 and the mean value for right footed players was €2,398,791. However, the variations were very high compared to the difference in means so I did not follow up on my theory. For part two, I set out to get more conclusive answers by running an independent two sample t-test (at a 95% confidence level) to compare the difference in means. My null hypothesis was that the mean value of lefties was equal to the mean value of righties. My alternative hypothesis was that the mean values of lefties was not equal to the mean value of righties.

H0: µ1 = µ2

Ha: µ1 ≠ µ2

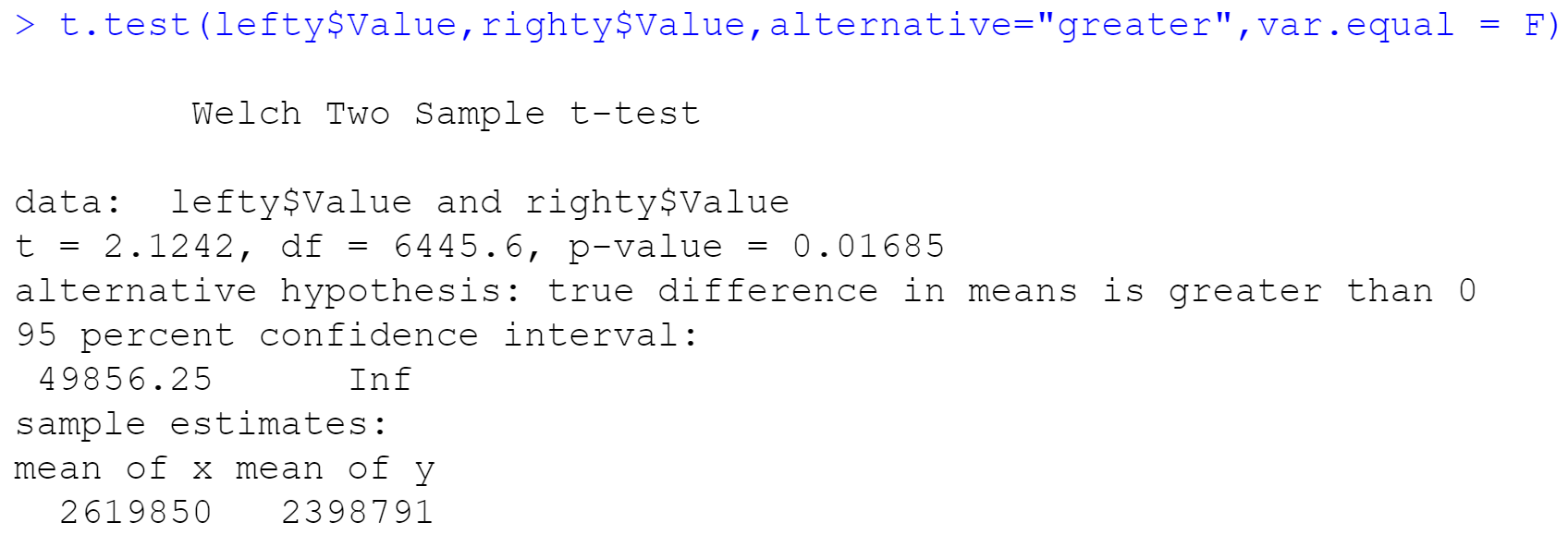


Since our p-value of .03 is less than our α of .05, I can confidently reject the null hypothesis in favor of the alternative and say that there is a statistically significant difference in mean player values. There is only a 3% chance that this conclusion could have happened due to chance. I then conducted a one-sided t-test to see if the mean value of lefties was greater than the mean value of righties.

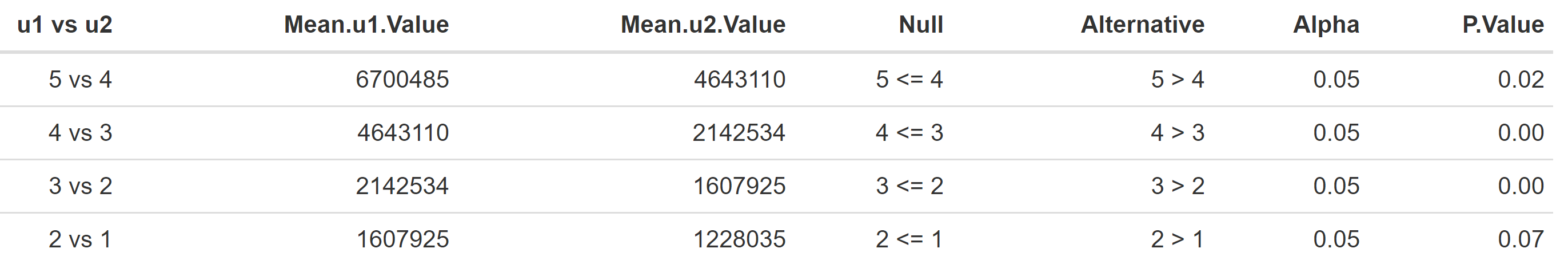
H0: µ1 ≤ µ2

Ha: µ1 > µ2

As you can see below, the p-value of .02 is less than .05 so lefties are in fact more valuable than righties.



Since part one’s results only showed a weak positive correlation between weak foot and value, I conducted one-sided independent two sample t-tests between different weak foot ratings at a 95% confidence level. The table below summarizes my hypothesis tests.

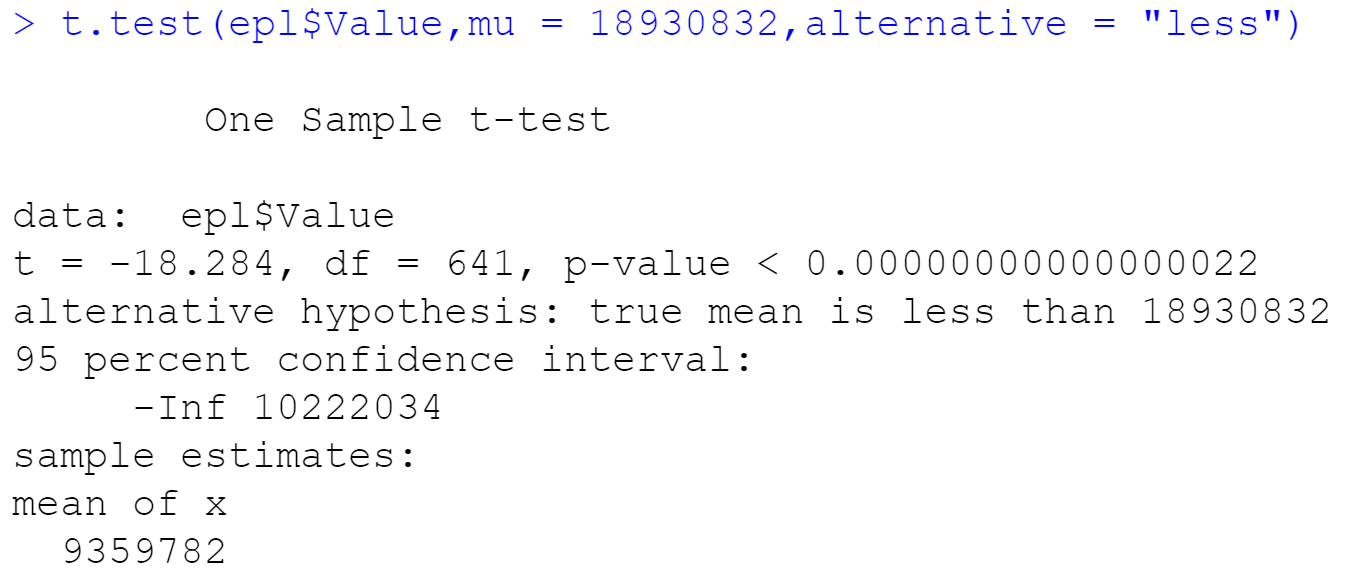


Our first three weak foot comparisons all had p-values less than our alpha value. We can reject the null hypotheses in favor of the alternatives and conclude that the mean value of 5 star weak foot players is greater than 4 star weak foot players. 4 star weak foot players are more valuable than 3 star weak foot players. 3 star weak foot players are more valuable than 2 star weak foot players. However, we cannot reject the null hypothesis for our fourth comparison and must conclude that 2 star weak foot players are just as valuable as 1 star weak foot players. Since in part one we determined the player values are skewed by outliers (players who play for the richest clubs), it is not surprising to see no difference in value between the lowest rated weak foot players. Those big clubs can afford to pay a premium for the better weak foot players, whereas the smaller clubs cannot afford to pay that premium and are stuck with the players the big clubs do not want. Conversely, since the major soccer leagues in Europe do not have salary caps, it is common practice for the big clubs to outbid each other. Any small competitive advantage a club can get they will try and get, so they will outbid the other clubs for better weak footed players.

The five major soccer leagues (Premier League in England, La Liga in Spain, Bundesliga in Germany, Serie A in Italy, and Ligue 1 in France) all have the same transfer windows where clubs can buy and sell players. Considering the massive financial competition among the big clubs, I was curious if clubs tend to overpay for players (price paid for a player exceeds the player’s value). I conducted a one sample t-test, at a 95% confidence level, to see if the average Premier League player value is less than the average transfer price from 2019 (€18,930,832) 1. I created a subset of all of the players from the 20 Premier League clubs and compared their average player value (€9,259,782) to that.

H0: µ1 ≥ €18,930,832

Ha: µ1 < €18,930,832



Since the p-value of 0 is less than our α of .05, we can reject the null in favor of the alternative and conclude that the mean Premier League player value is less than the average transfer price. Similarly, in part one, I would like to remove the lowest value players since they are the least desirable and less likely to be transferred. There is concern that the average transfer price is skewed because the better players more desirable players and more likely to be transferred, however I do not have a dataset specifically on Premier League transfers. Considering our p-value is so incredibly low and my calculations have been consistent, I am confident in concluding that Premier League clubs overpay for players. This is not necessarily a bad thing, however. A player’s market value can fluctuate and become different from their Fifa 19 value rather quickly depending on their recent play, teams’ depth at that player’s position, and the sheer competitive and threat of relegation of the Premier League that Fifa 19 player values to not take into consideration.

Summary

Since my analysis from part one looked promising, albeit inconclusive at times, I conducted hypothesis testing to finally answer these two questions: Are left footed players more valuable than right footed players? Do player values increase with an increase in their weak foot skills? My hypothesis tests were conclusive and we determined that lefties are more valuable than righties and player value increases as weak foot ratings increase from 2 to 5 (no difference in value between 1 and 2 rated players). I then decided to look at player value even further and give my analysis a more real-world reflection. I ran a one sample hypothesis test to compare the average Premier League player value to the average Premier League transfer price and concluded that Premier League teams tend to overpay.

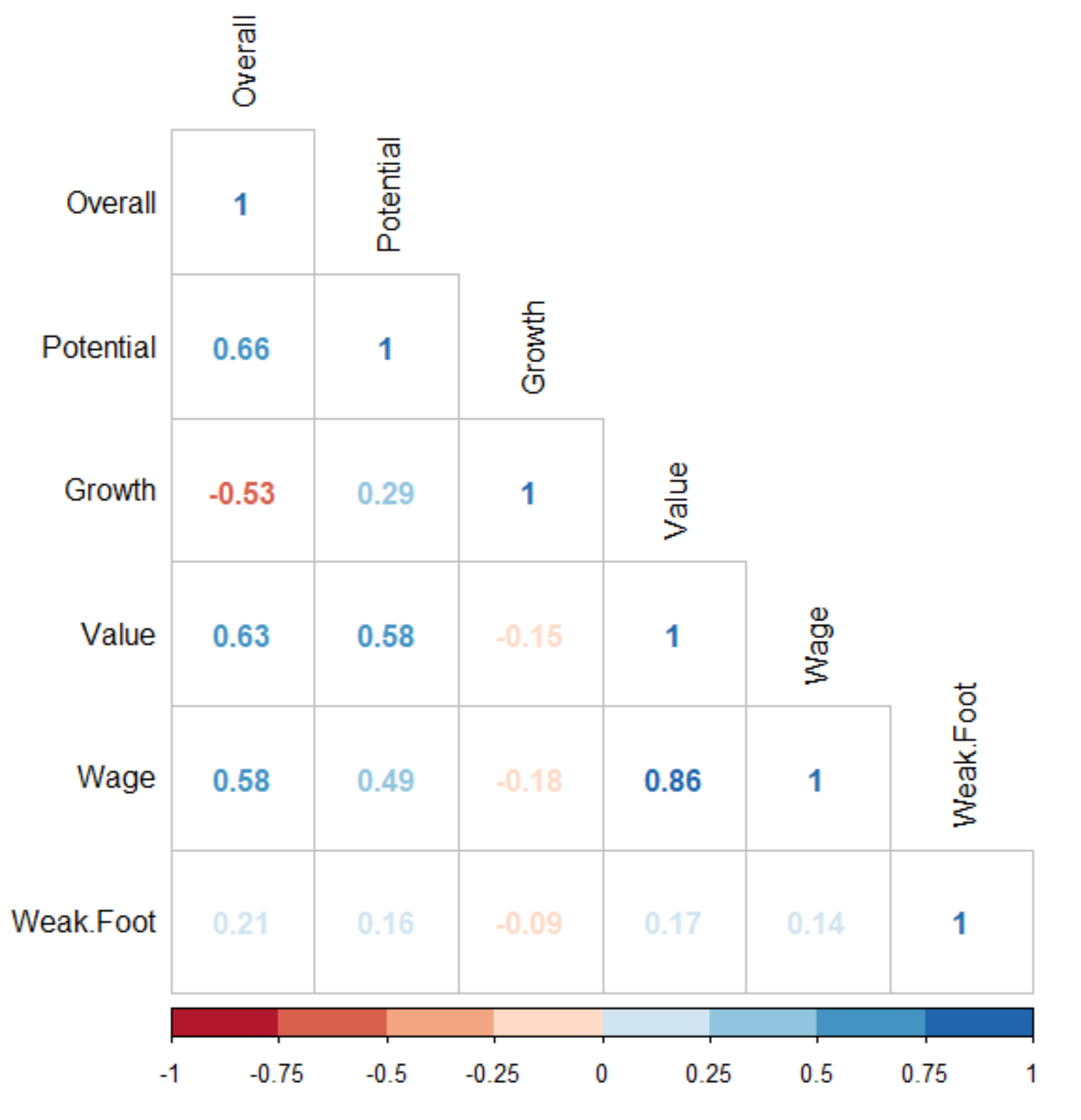
Milestone 3

Introduction

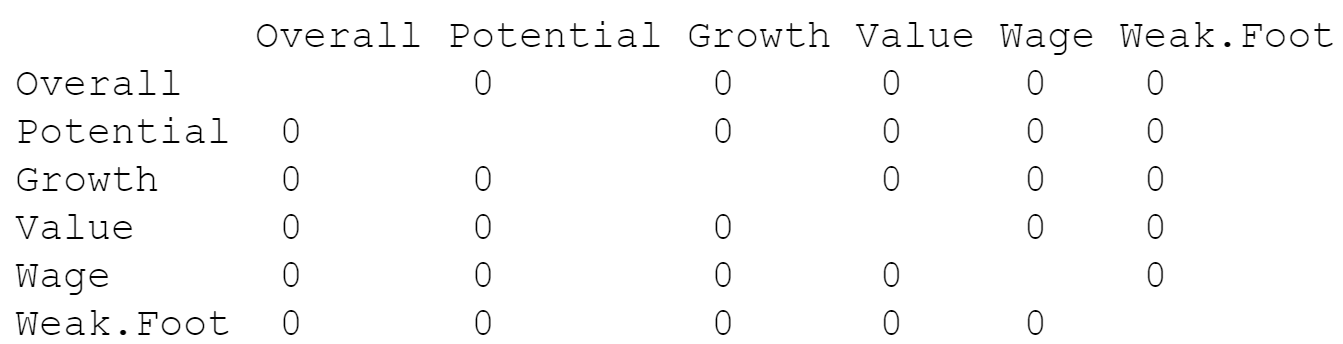
For the final milestone of this project, I used simple linear regression and multiple regression to build models that can predict certain attributes based on predictor variables. The first model I tried to build attempted to predict overall rating based on a player’s weekly wage. I then continued to explore the difference in values for left-footed and right-footed players based on the promising results from last milestone’s hypothesis tests. Lastly, I used multiple regression to build a full model predicting a player’s value.

Analysis

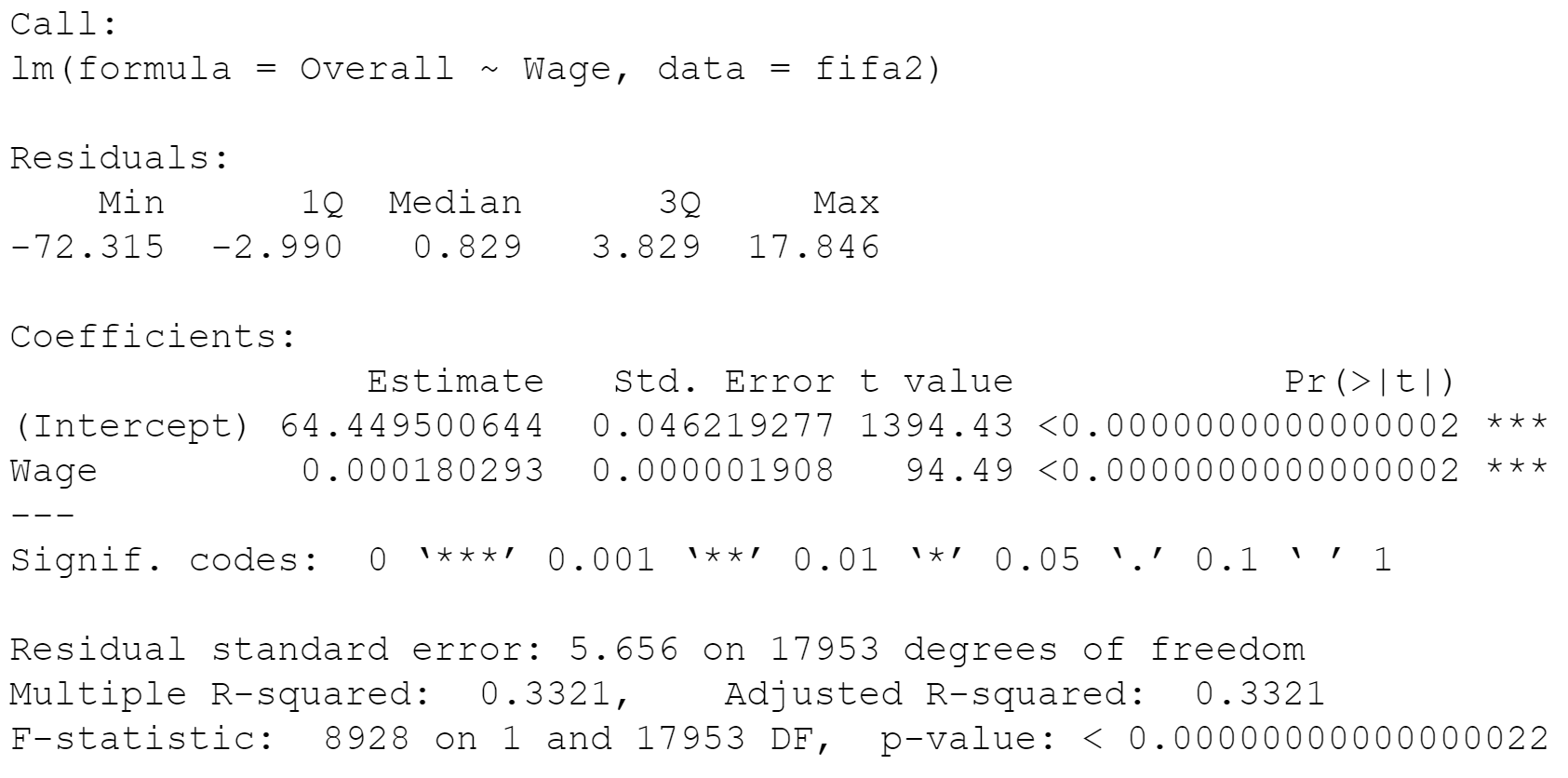
Before jumping directly into the regression, I created the following correlation table to see if my questions were worth exploring. Dark red r values indicate strong negative correlation (as one variable increased the other variable decreased) and dark blue r values indicate strong positive correlation (both variables increase or decrease at similar rates). Since we dealt with real-life data instead of manipulated data for statistical practice, we have to temper our expectations about seeing very strong positive or negative correlations since there can be more randomness involved. I considered r values near or above .5 or -.5 to be strong enough correlation coefficients to warrant further exploration. I also limited my correlation table to the 6 most important variables relevant to my analysis. My first correlation table had all r values of all variables but there were too many combinations to make the table simple enough to gather key takeaways. I also created a correlation table of p-values which showed some correlations were not statistically significant at a 95% confidence level, so I made sure to remove those as well before producing my final correlation table.



Based on the following p-value table, all of the correlations above are statistically significant at a 95% CL (α = .05).



As we already know from milestone 1, Wage and Value are strongly and positively correlated with an r value of .86 and Weak Foot ratings are not correlated with Value with an r value of .17. Since Wage and Overall have a correlation coefficient of .58, I wanted to see if Wage can predict Overall. Since the Wage variable is reported from a player’s real wage and Overall is merely a variable created from the Fifa 19 developers, I wanted to see just how much Wage influenced Overall in order to gauge how the developer’s Overall ratings were influenced from the only real-world variable in the dataset.

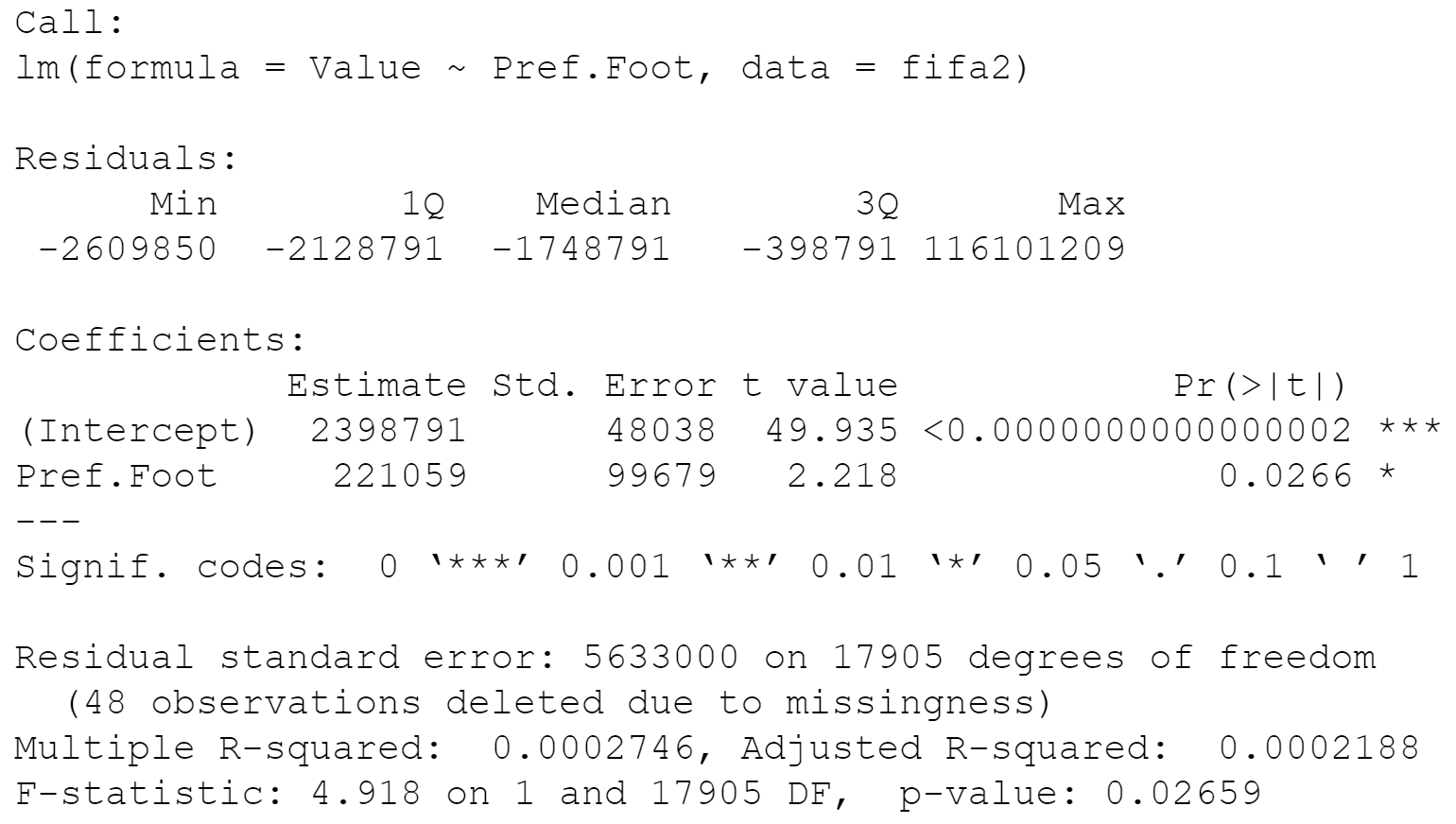


Since the model, Wage variable, and intercept all have p-values below .05, I used the regression results to create the following formula to predict Overall based on Wage.

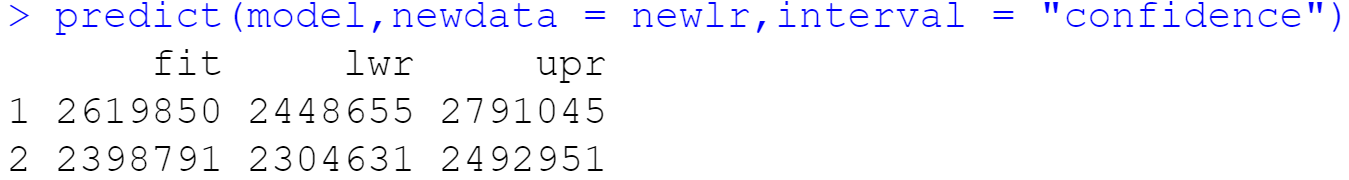
Overall = 64.45 + .0001803 (Wage) ± .047

If a player makes €88,000 per week, you can expect their Overall rating to be around 80. Liverpool, the defending Premier League champions, on average, pay that amount per player per week. Their average player’s Overall is 76, which is fairly accurate based on Liverpool’s overpayment of young players with high potential but currently low Overalls. However, the model’s adj R2 value of .33 means that only 33% of variation in Overall is explained by Wage (67% is unexplained). This is approximately what I expected considering the original dataset has 34 various attributes that contribute to Fifa 19’s Overall ratings and it would be foolish for Fifa 19 to completely base their Overalls on Wage.

In milestone 2 we learned that left-footed players are in fact more valuable than right footed players. I double-checked this conclusion by assigning dummy variables to Preferred Foot in order to run regression. I created a new column called Pref Foot and assigned a 1 for “Left” and 0 for “Right”. Pref.Foot is statistically significant so our baseline mean Value of €2,398,791 is the same mean Value for righties as we calculated earlier. We can calculate the mean Value for lefties by adding b0 + b1 (2398791 + 221059) to get a mean lefty Value of €2,619,850 which is also the exact same value we calculated in milestone 2. However, I included Pref Foot in my next multiple regression to find the best predictors of Value and it was the only variable not statistically significant, so I removed it from the model.



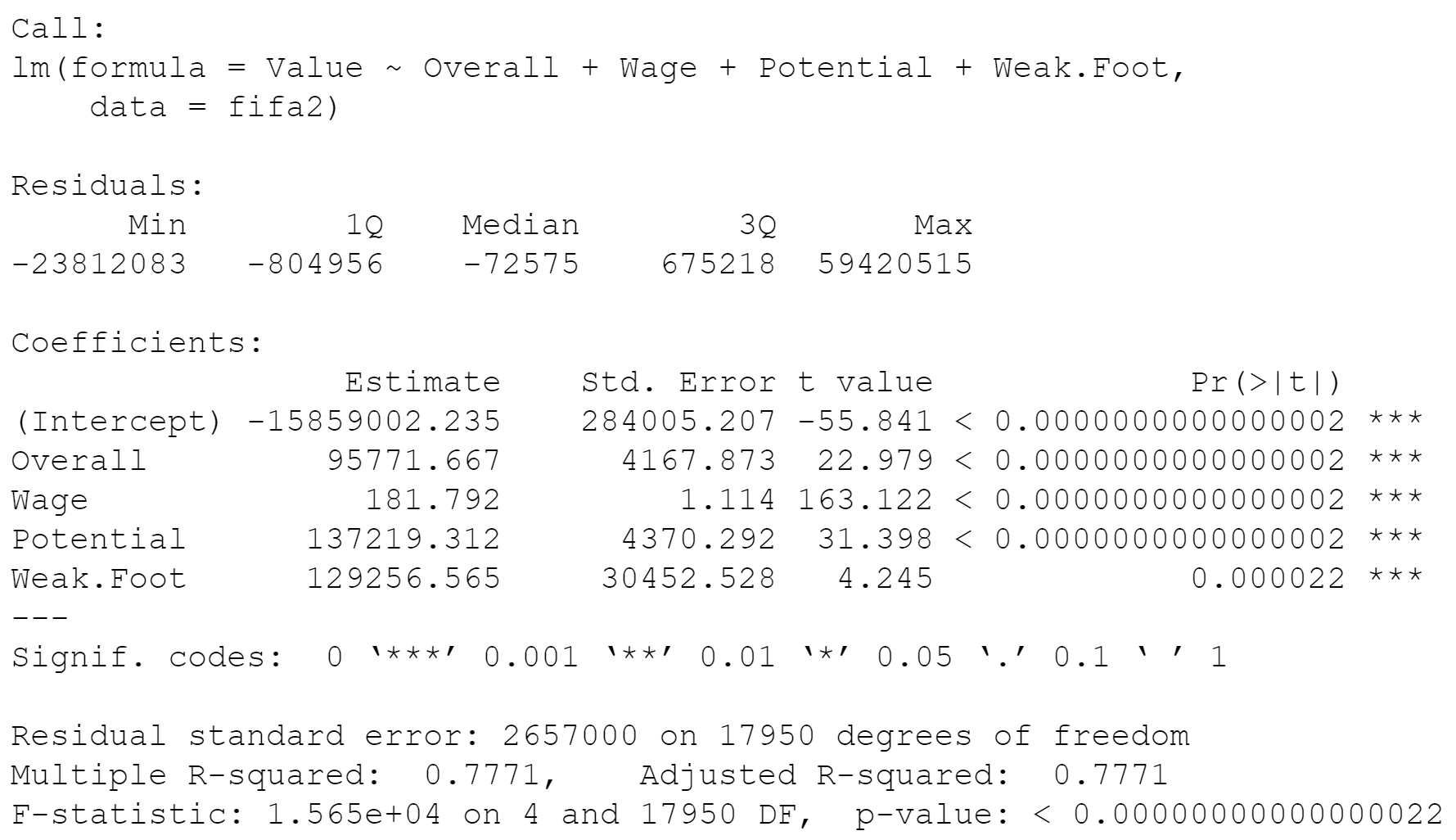
I created 95% Confidence Intervals for left-footed players and right-footed players to further illustrate the difference in Values. I am 95% confident that the mean lefty value is between €2,448,655 and €2,791,045. I am 95% confident that the mean righty value is between €2,304,631 and €2,492,951. There is very little overlap between these intervals so I am extremely confident that lefties are more valuable than righties.



After exploring Value in different ways for most of the report, I wanted to find the best predictors and the best model for Value. Earlier in this report we compared Wage and Weak Foot to Value, but for milestone 3 I ran a multiple regression to find the conclusive list of predictors. After trying all variables and removing the insignificant predictors based on an alpha of .05, I found that Overall, Wage, Potential, and Weak Foot are all significant predictors of Value. Based on the following regression output, I created this formula to estimate Value.

**Value = -15859002 + 95712\*(Overall) + 182\*(Wage) + 137219\*(Potential) + 129257\*(Weak Foot) ± 322997**

If a player is 75 Overall making €70,000 a week with a Potential of 90 and Weak Foot skills at a 4, we can expect that player to be worth about €16,926,136. This intuitively makes sense because an average Premier League player is rated around 75 and makes about €70,000. However, a 4 star weak foot rating is above average and a potential of 90 gives them the potential to be one of the best players in the league. Even though that player is not at that level yet, teams would pay more than what that player is currently worth in hopes that player reaches their potential. But their value is not yet that of a 90 rated player because there is risk that the player does not reach that level.



Since the model and each predictor have p-values less than .05, we determined that this model is a statistically significant predictor of Value. With an adjusted R2 value of .7771, 77.71% of the variation in Value is explained by these 4 predictors. That is a very high adj R2 value so we know this is a great model based on the Fifa 19 dataset. However, 22.29% of the variation is unexplained by these 4 variables. That seems accurate to me considering the whole dataset has over 30 variables. Even though Age was not a significant predictor of Value, if you take my recent example of a 75 rated player, a player who is 19 years old is more valuable than a player who is 23 because they would, in theory, have 4 extra years to reach their potential. There could also be certain individual attributes that are more valuable to teams, such as passing or shooting as compared to their penalty skills.

Summary

Throughout my entire project, I analyzed the Fifa 19 player ratings to see how fairly the developers created the rosters and how player values were influenced. Fifa 19 had normally distributed Overall and Weak Foot ratings as they should in order to keep the video game balanced and ensure there are significant differences between bad and good players. However, player wages and values were heavily skewed by outliers due to the abundance of low-level players and the rarity of highly talented players. I did not remove any outliers because the high value players get the most playing time at the popular clubs and are worth analyzing more than low-level players. In future analysis I would consider removing all of the low-level players and focus on the clubs in the top tier leagues around the world. I was also interested to see that left-footed players are more valuable than right-footed players. Even though I could not build a model to estimate a player’s value based on their preferred foot, lefties are more valuable due to the 3:1 ratio of righties to lefties but a desire to have balanced rosters based on popular soccer tactics. I was then able to determine the most important predictors of Value and explain 77% of the variation in Value by Overall, Wage, Potential, and Weak Foot rating as well as create a model that predicts Value given any set of those 4 predictors.

Citations

“ggplot2 Histogram Plot : Quick Start Guide - R Software and Data Visualization.” STHDA, www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization.

“How to Add Mean, and Mode to Ggplot Histogram?” Stack Overflow, 1 Sept. 1966, stackoverflow.com/questions/47000494/how-to-add-mean-and-mode-to-ggplot-histogram.

“ggplot2 Title : Main, Axis and Legend Titles.” STHDA, www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles.

“ggplot2 Scatter Plots : Quick Start Guide - R Software and Data Visualization.” STHDA, www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization.

Discrete and Continuous Data, www.mathsisfun.com/data/data-discrete-continuous.html.

“How to Deal with Missing Data Values in R.” Dummies, www.dummies.com/programming/r/how-to-deal-with-missing-data-values-in-r/.

OpenStax CNX, cnx.org/contents/xYwts2vl@1/Hypothesis-Testing-Two-Population-Means-Independent-Samples-Unknown-Population-Standard-Deviations.

“SPSS Tutorials: One Sample t Test.” LibGuides, libguides.library.kent.edu/SPSS/OneSampletTest.

1 Dubas-Fisher, David. “How Much Is the Average Premier League Transfer Fee - and How Does It Compare?” ChronicleLive, 8 July 2019, www.chroniclelive.co.uk/sport/football/football-news/how-much-average-premier-league-16552046.

“Unpaired Two-Samples T-Test in R.” STHDA, www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r.

“One-Sample T-Test in R.” STHDA, www.sthda.com/english/wiki/one-sample-t-test-in-r.

“How to Perform a One Sample t-Test by Hand and in R: Test on One Mean.” Stats and R, statsandr.com/blog/how-to-perform-a-one-sample-t-test-by-hand-and-in-r-test-on-one-mean/.

Kassambara, et al. “Predict in R: Model Predictions and Confidence Intervals.” STHDA, 10 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/.

Kassambara, et al. “Simple Linear Regression in R.” STHDA, 10 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/.

Kassambara, et al. “Regression with Categorical Variables: Dummy Coding Essentials in R.” STHDA, 11 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/.

Moon, Keon-Woong. GgPredict() - Visualize Multiple Regression Model, 6 Oct. 2020, cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html.

REUTERS “Every Premier League Club's Average Weekly Salary Listed for 2019-20 Campaign.” Mirror, 19 Nov. 2019, www.mirror.co.uk/sport/football/news/every-premier-league-clubs-average-20916779.

Appendix

See txt files:

“DiSessa – R Code – Final Project Milestone 1” submitted with assignment 1

“DiSessa – R Code – Final Project Milestone 2” submitted with assignment 2

“DiSessa – R Code – Final Project Milestone 3” submitted with assignment 3