Effect of Fifa Player Ratings on Real Life Wages

Alan Qin

03/04/2020

Contents

Project Introduction

My project is about players in the game FIFA 20 and how wages are affected by rating. This is interesting to me because I want to see how the best soccer players' wages are affected by their FIFA rating. The source of the data is from Kaggle.com, specifically (https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset). The Y (outcome) in this data are the wages of the players while the X's are the ratings of the player. There are well over 50 X's in this dataset but the most important in my opinion are, overall rating, potential rating, shooting rating, and international reputation. In this project, I am using different techniques to find out if professional wages are affected by FIFA rating.

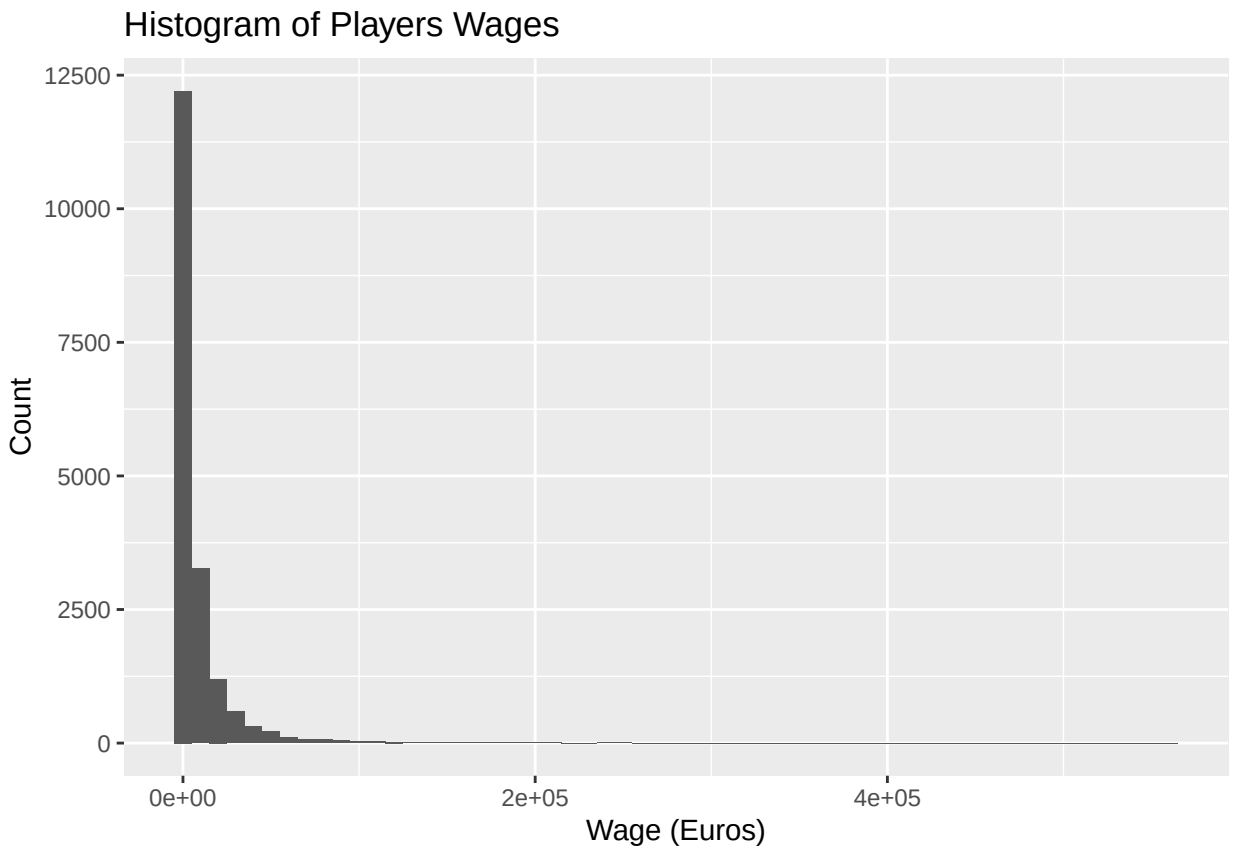
Some of the variables are

- **overall**: overall attribute of the player
- **potential**: potential attribute of the player
- **value_eur**: value in EUR of the player
- **wage_eur**: wage in EUR of the player
- **pace**: player pace rating
- **shooting**: player shooting rating

Report 1

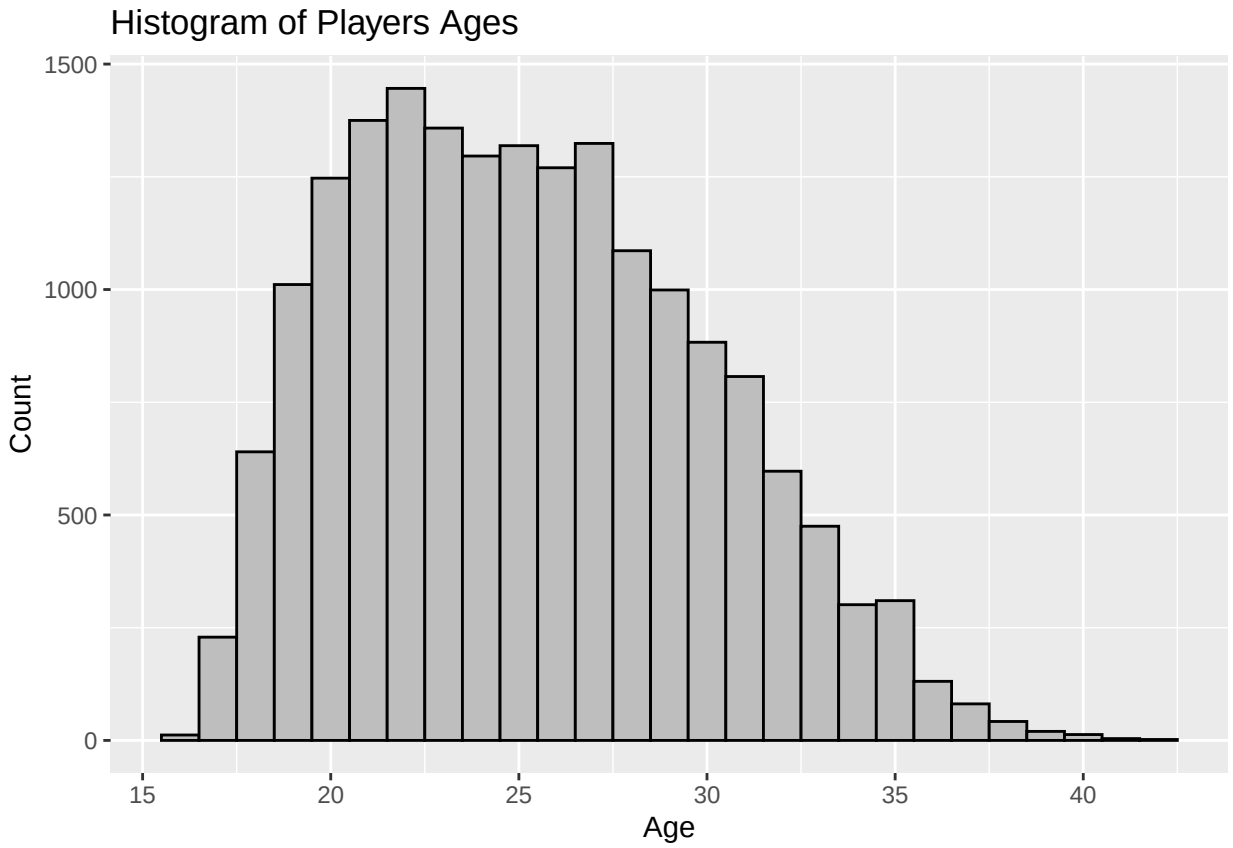
Histogram for my X's and Y's

```
ggplot(data = data, aes(x = wage_eur)) +
  geom_histogram(binwidth = 10000) +
  ggtitle('Histogram of Players Wages') +
  labs(x = 'Wage (Euros)', y = "Count")
```



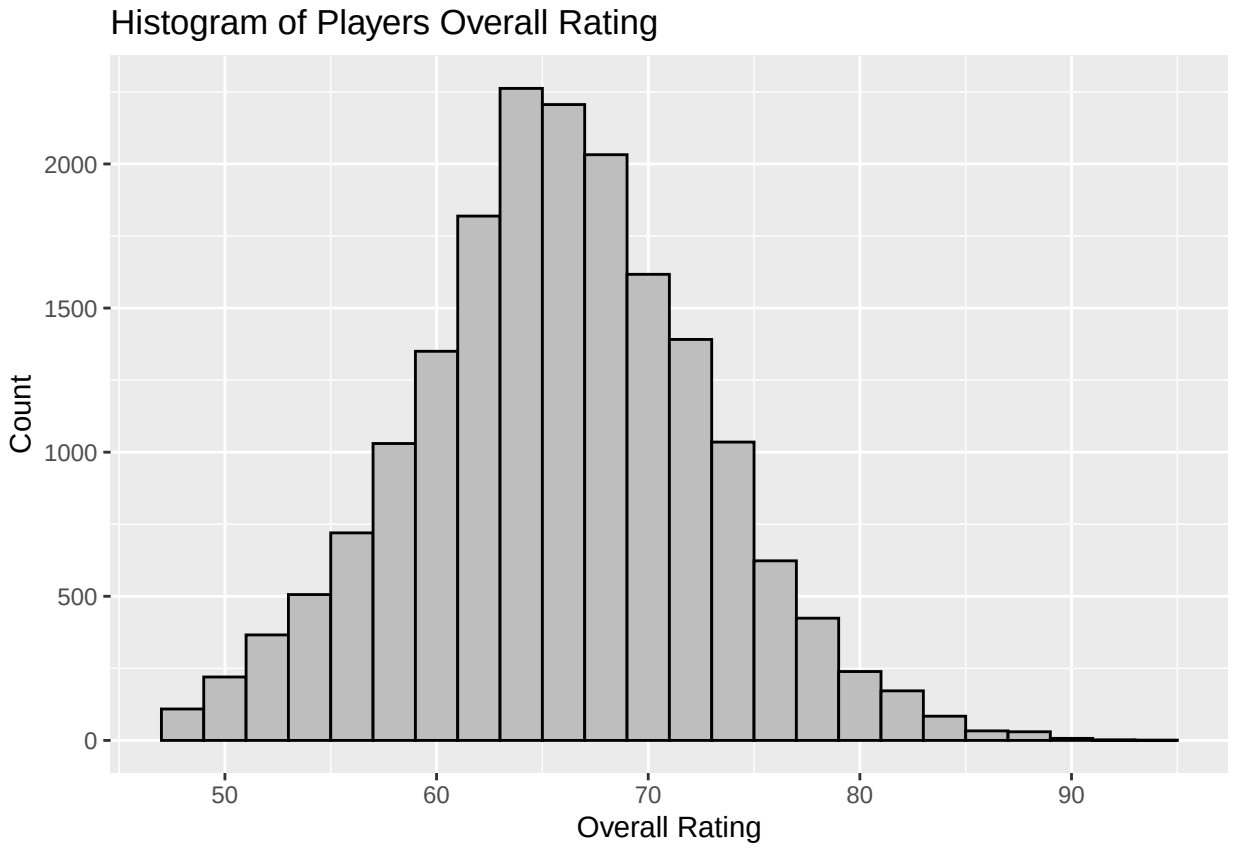
This histogram shows the wages of all the players in fifa. As you can see, there are over ten thousand players with wages of around 0. As we go up in wage, there are less and less players with one player (Lionel Messi) making around 500000 euros a week.

```r
ggplot(data = data, aes(x = age)) +
  geom_histogram(binwidth = 1, color = 'black', fill = 'gray') +
  ggtitle('Histogram of Players Ages') +
  labs(x = 'Age', y = "Count")
```



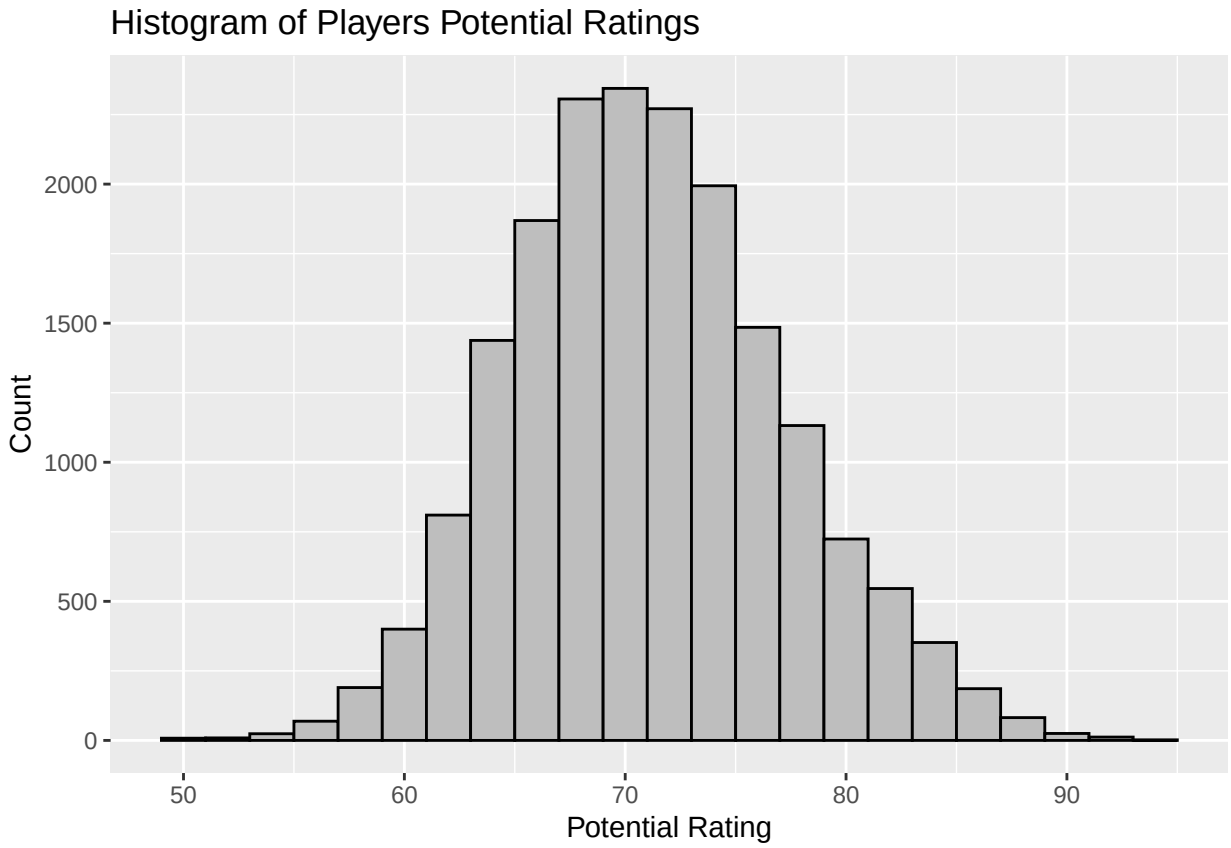
As we can see from this histogram, there are many players within the age range of 20-29 while there are few players younger than 18 and older than 35. There are the most players within the age range of 20-29 because that is their physical peak.

```r
ggplot(data = data, aes(x = overall)) +
  geom_histogram(binwidth = 2, color = 'black', fill = 'gray') +
  ggtitle('Histogram of Players Overall Rating') +
  labs(x = 'Overall Rating', y = "Count")
```



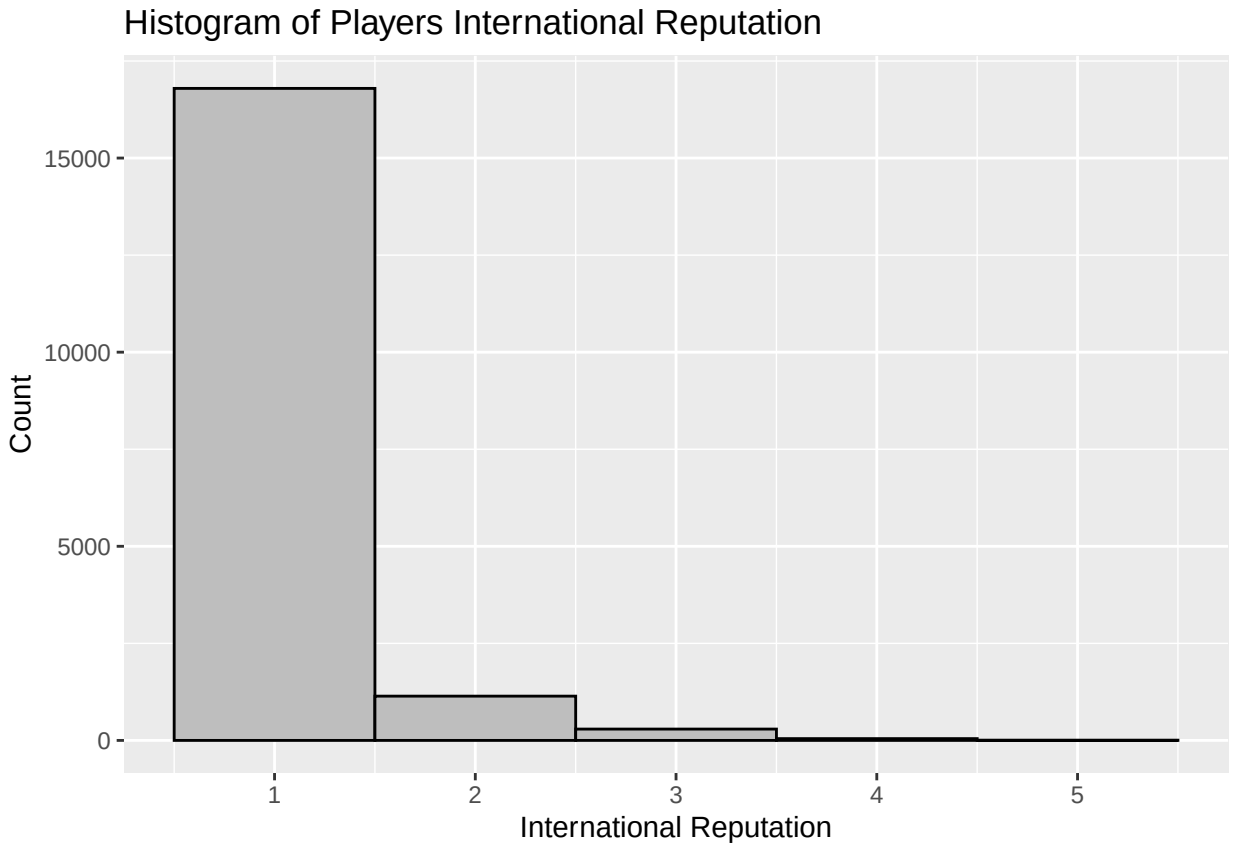
As we can see from this histogram, the mode of the overall rating of a player in FIFA 20 is around 65 points. The histogram also looks like it is normally distributed.

```r
ggplot(data = data, aes(x = potential)) +
  geom_histogram(binwidth = 2, color = 'black', fill = 'gray') +
  ggtitle('Histogram of Players Potential Ratings') +
  labs(x = 'Potential Rating', y = "Count")
```



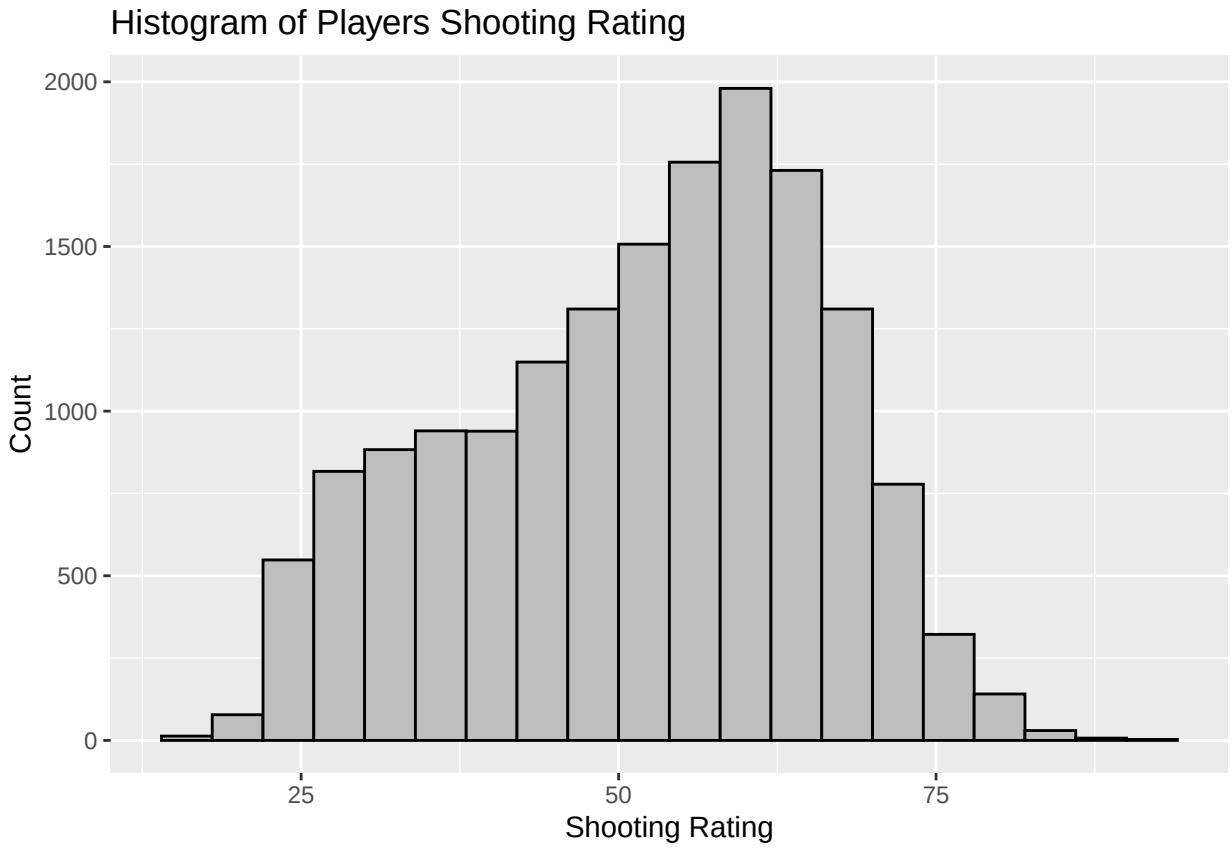
Simliarly to the Overall Rating Histogram, the potential ratings of players are also normally distributed. However, the average potential rating of FIFA players increased to around 70 points instead of 65.

```r
ggplot(data = data, aes(x = international_reputation)) +
  geom_histogram(binwidth = 1, color = 'black', fill = 'gray') +
  ggtitle('Histogram of Players International Reputation') +
  labs(x = 'International Reputation', y = "Count")
```



This histogram shows that there are many players with an international reputation of 1 with less than 2000 players with a rating above. I think this is the case becuase international repuation is based off of how well the player plays for their national team. Since there are only so many spots on a national team it makes sense that the vast majority of players have an international reputation of 1.

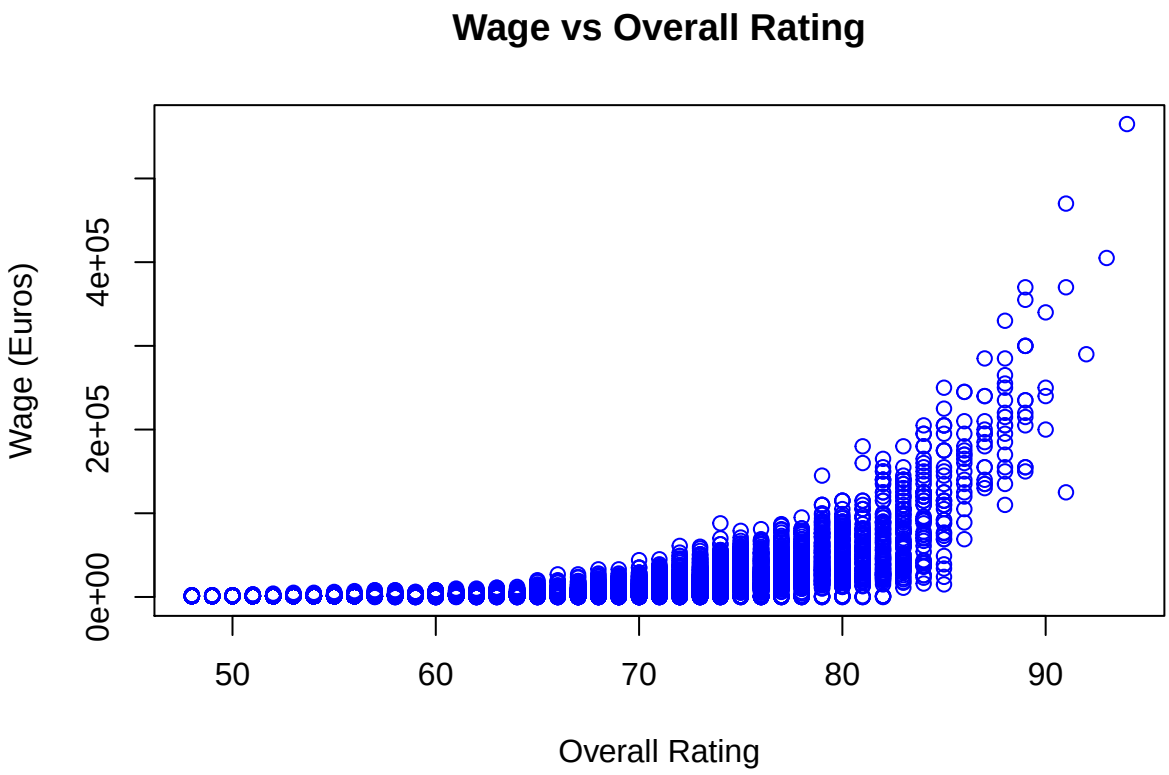
```r
ggplot(data = data, aes(x = shooting)) +
  geom_histogram(binwidth = 4, color = 'black', fill = 'gray') +
  ggtitle('Histogram of Players Shooting Rating') +
  labs(x = 'Shooting Rating', y = "Count")
```



As we can see from this histogram of players' shooting rating,

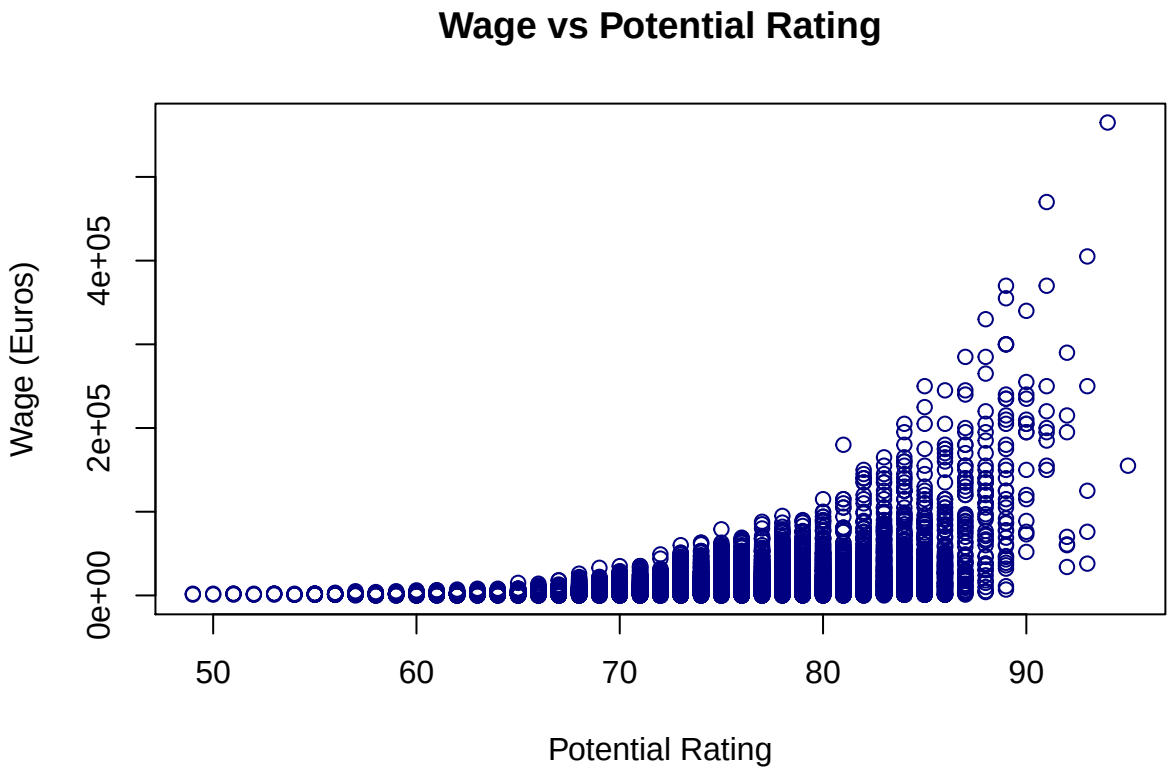
Plots of X's vs Y's

```
plot(x = data$overall,
     y = data$wage_eur,
     ylab = "Wage (Euros)",
     xlab = "Overall Rating",
     main = 'Wage vs Overall Rating',
     col = 'blue')
```



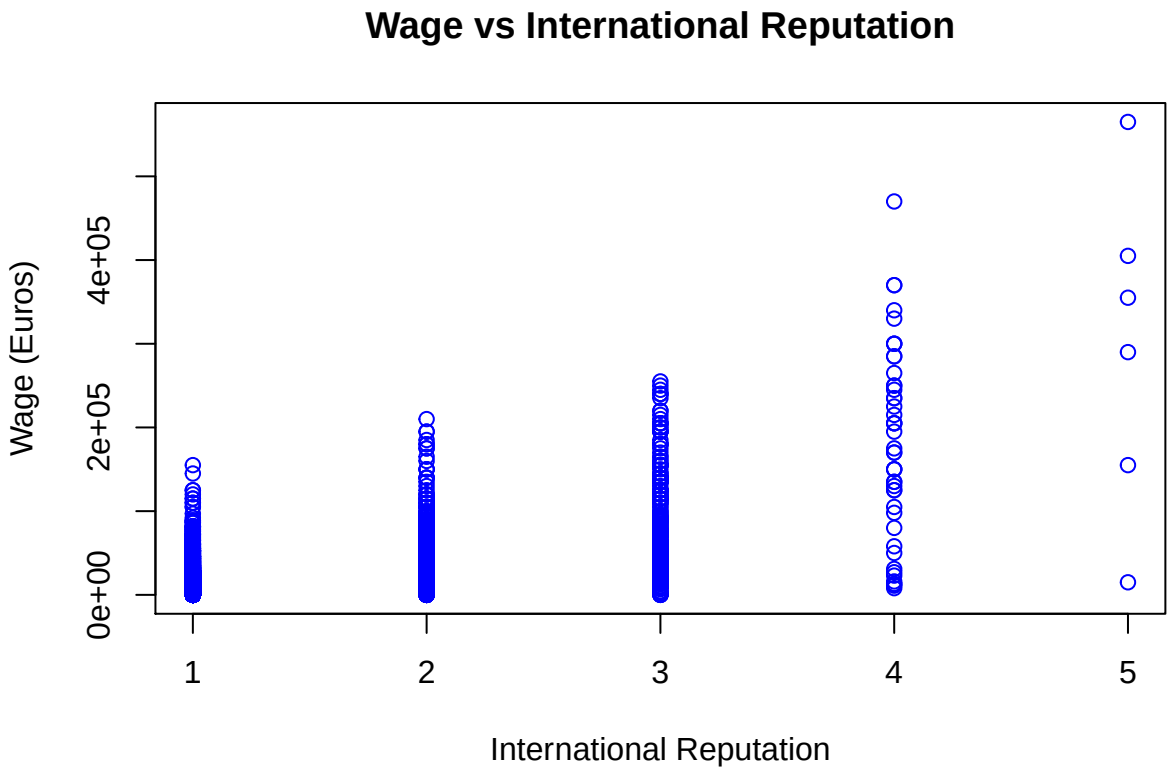
In this plot, you can see that overall rating has a clear correlation with wage. As overall rating goes up, wage also goes up. This plot also looks like an exponential function.

```r
plot(x = data$potential,
     y = data$wage_eur,
     ylab = "Wage (Euros)",
     xlab = "Potential Rating",
     main = 'Wage vs Potential Rating',
     col = 'navy')
```



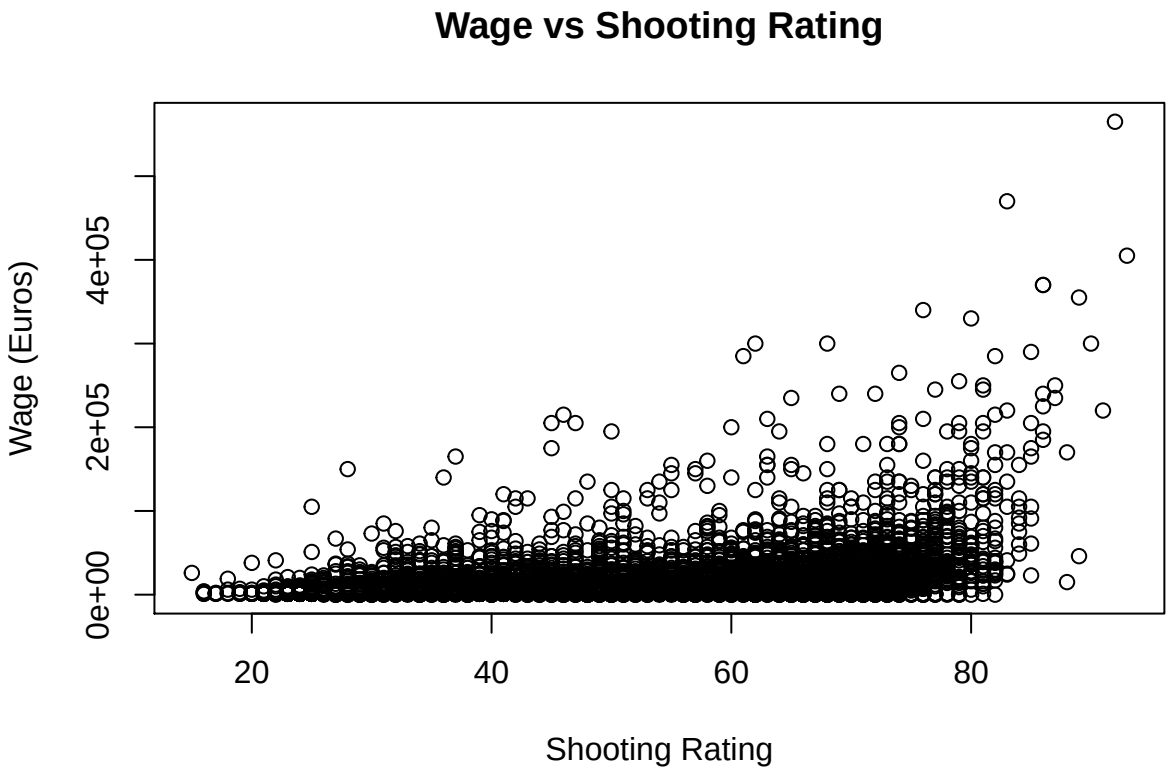
In this plot, you can also see that it is very similar to the results of the overall rating. This is because I think that overall rating is almost the same as potential rating on the players with the highest wages.

```r
plot(x = data$international_reputation,
     y = data$wage_eur,ylab = "Wage (Euros)",
     xlab = "International Reputation",
     main = 'Wage vs International Reputation',
     col = 'blue')
```



Again, this box plot shows that there are many players with an international reputation of 3 or less but, they do not make that much money. Most of the players with an international repuation of 4 or higher earn a much larger salary than those with a rating of 3 or less.

```r
plot(x = data$shooting, y = data$wage_eur,
     ylab = "Wage (Euros)",
     xlab = "Shooting Rating",
     main = 'Wage vs Shooting Rating')
```



In this plot, shooting rating does not tell us much about the wage of the players unless their shooting stat is very high compared to other players. As we can see, the players with 90 shooting or above has high salaries compared to 85 and below. I think this is because attackers in soccer are paid more than defenders.

```r
plot(x = data$age,
     y = data$wage_eur,
     ylab = "Wage (Euros)",
     xlab = "Age",
     main = 'Wage vs Age')
```

As we can see from this plot, most of the fifa players are between 23 and 33 years old. This is also where the players get paid the most because of their combination of physical form and experience.These are peak years for athletes and they have the wages to back that up.

Report 2

Finding the Best Regression

Fixing Release Clause Data

```
data$release_clause_eur[is.na(data$release_clause_eur)]= 0
```

Since there are some release clauses that are NA, I set them to 0 to make sure that the dimensions of my regressions and residuals are the same.

Regression for different X's on wage

```
fit_release = lm(data = data, wage_eur ~ release_clause_eur)
summary(fit_release)$adj.r.squared
```

```
## [1] 0.7102299
```

The variable I chose were players' release clause in Euros. This variable would make sense because the release clause should be in a similar range to wages.

```
fit_value = lm(data = data, wage_eur ~ value_eur)
summary(fit_value)$adj.r.squared
```

```
## [1] 0.736238
```

The variable I chose were players' value in Euros. This variable would make sense because player value should be in a similar range to wages.

```
fit_pass = lm(data = data, wage_eur ~ passing)
summary(fit_pass)$adj.r.squared
```

```
## [1] 0.1630528
```

The variable I chose were the players' passing rating. This variable would make sense because as a player gets better at passing the more a player should get paid.

```
fit_drib = lm(data = data, wage_eur ~ dribbling)
summary(fit_drib)$adj.r.squared
```

```
## [1] 0.1376096
```

The variable I chose were the players' dribbling rating. This variable would make sense because as a player gets better at dribbling the more a player should get paid.

```r
fit_overall = lm(data = data, wage_eur ~ overall)
summary(fit_overall)$adj.r.squared
```

```
## [1] 0.3289551
```

The variable I chose were the players' overall rating. This relationahip would make sense because as a player gets better as a player the more they should get paid.

```r
fit_potential = lm(data = data, wage_eur ~ potential)
summary(fit_potential)$adj.r.squared
```

```
## [1] 0.2260481
```

The variable I chose was the potential rating of the players. The relationship between wage and potential rating makes sense because soccer clubs pay for player's potential.

```r
fit_best = lm(wage_eur ~ release_clause_eur +
                value_eur + overall +
                potential, data = data)
summary(fit_best)$adj.r.squared
```

```
## [1] 0.7402041
```

The variables for this regression were release clause in euros, value in euros, overall rating, and potential rating. The relationship between these variables make sense because all these variables should affect the wage of a player.

Finding the best fit

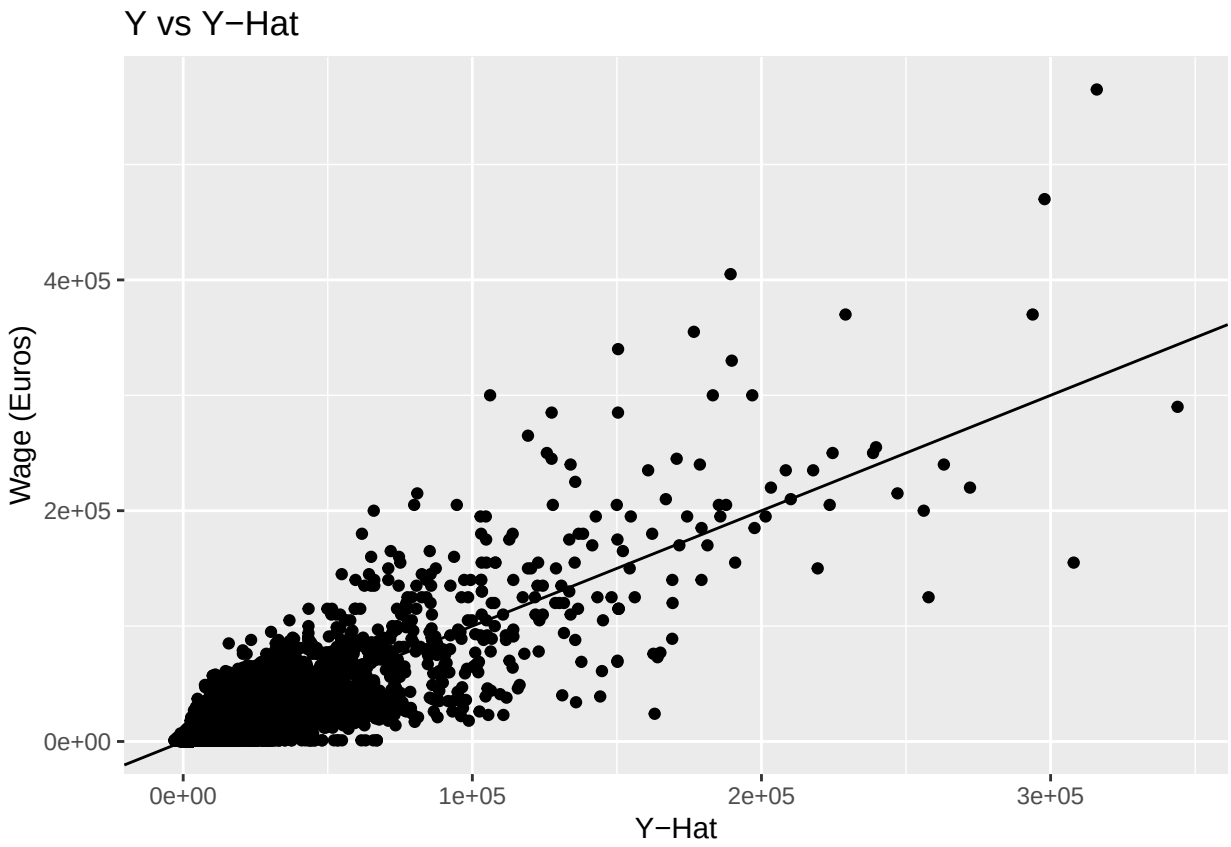
```
summary(fit_best)
```

```
##
## Call:
## lm(formula = wage_eur ~ release_clause_eur + value_eur + overall +
##     potential, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152988   -2117    -491    1434  249032
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.105e+03  1.248e+03   1.686   0.0918 .
## release_clause_eur  2.281e-04  3.575e-05   6.382 1.79e-10 ***
## value_eur           2.810e-03  7.080e-05  39.688  < 2e-16 ***
## overall             2.390e+02  1.684e+01  14.191  < 2e-16 ***
## potential          -2.302e+02  1.794e+01 -12.826  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10880 on 18273 degrees of freedom
## Multiple R-squared:  0.7403, Adjusted R-squared:  0.7402
## F-statistic: 1.302e+04 on 4 and 18273 DF,  p-value: < 2.2e-16
```

From my testing, I believe that fit_best is the best regression. The regression includes the X variables release clause (euros), value (euros), overall rating, and potential rating. The Adjusted R-Squared value is 0.7402 which is relatively high. In statistics, the Adjusted R-Squared value explains the percentage of the variation around the mean. In other words, the regression can explain 74.02 % of the variation around the mean. The p-value for all the X variables are significant. This means that the variable's p-value is less than .05. As we can see from the summary, the p-values of the variables are far less than .05 resulting in significant variables. The mean response for release clause was 2.281e-04, mean response for value was 2.810e-03, the mean response for overall rating was 2.390e+02, and the mean response for potential rating was -2.302e+02. That means holding everything else equal, an increase in overall rating will cause an increase in wage by 239 Euros. This makes sense as the better the player, the better the wages. The same thing goes with release clause and value of a player. The more valuable the player, the more the player will get paid. The one weird thing I did notice in the regression was the mean response for potential rating because it was negative. One would think that an increase in potential rating would increase wages but that is simply not the case. One explanation could be that since overall rating and potential rating are very simliar they would cancel each other out in our regression line. Our F statistic is 1.302e+04 on 4 and 18273 Degrees of Freedom with a p-value of 2.2e-16. This menas that our statistics are significant and that our Adjusted R-Squared value is also significant.

Y vs Yhat

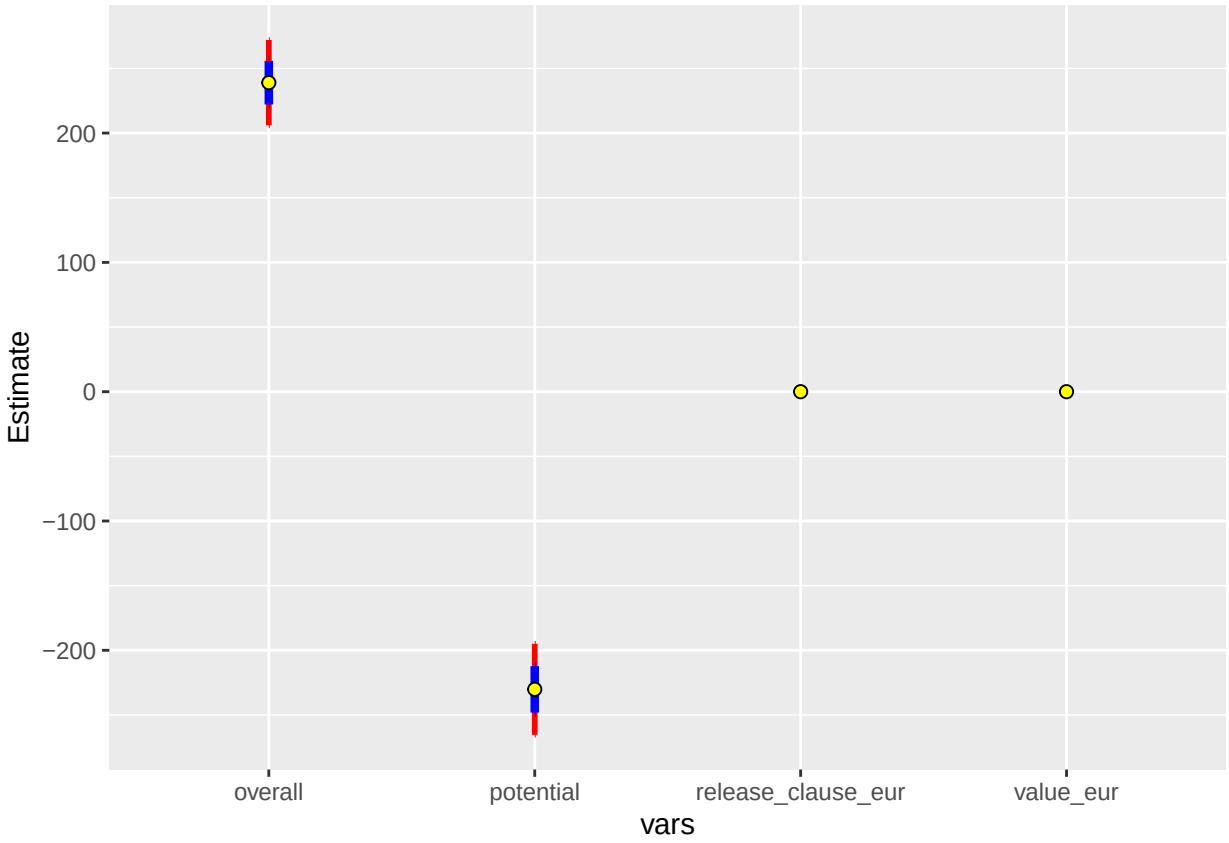
```
pred = fit_best$fitted.values
ggplot() + geom_point(aes(x = pred, y = data$wage_eur)) + geom_abline() + ggtitle('Y vs Y-Hat') +
  labs(x = 'Y-Hat', y = 'Wage (Euros)')
```



When plotting Y-hat onto Y, the desired result is a one-to-one relationship. As we can see from the plot, Y is almost a one to one function of the prediction but, the variance of the residuals increases for high values. So we have heteroscedasticity in the wage dimnesion.

Estimated Coefficients and Standard Error

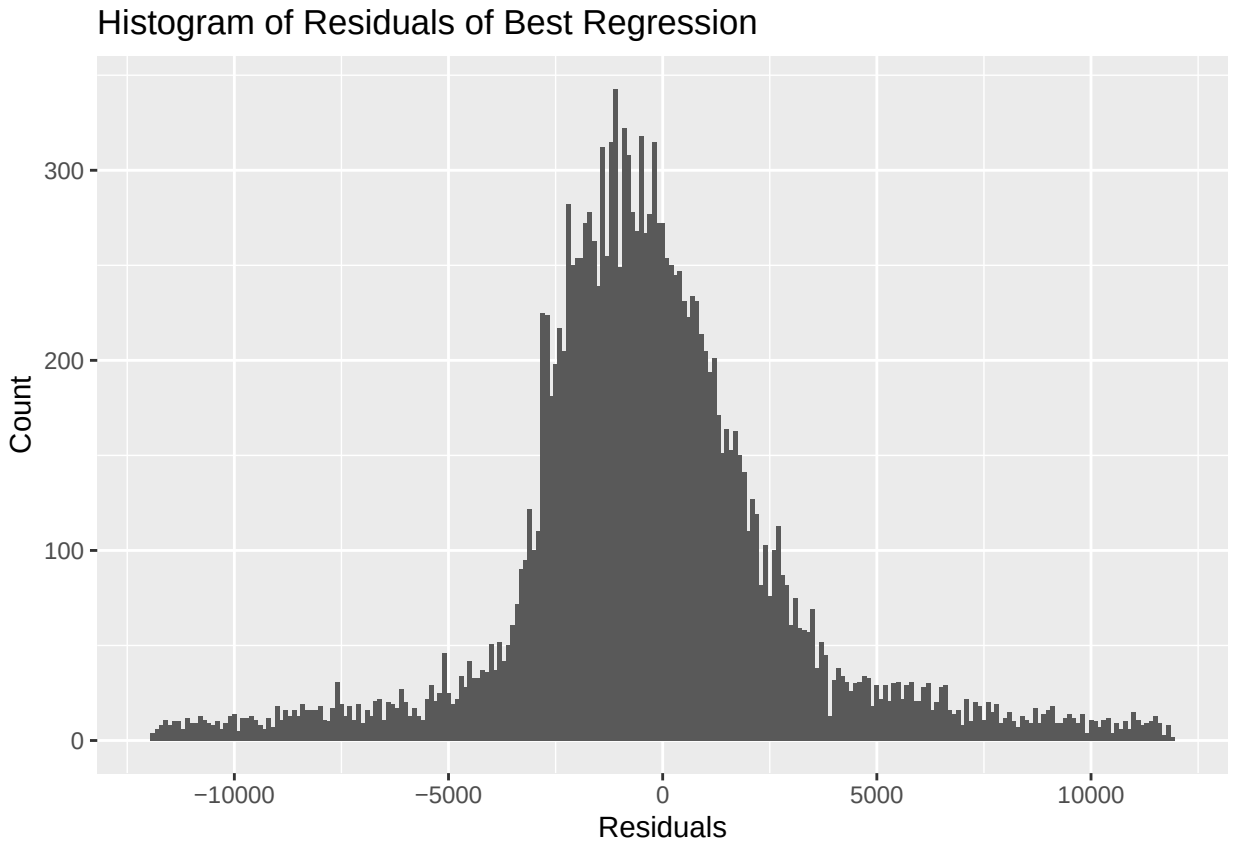
```
fifa_data = summary(fit_best)
coefs = as.data.frame(fifa_data$coefficients[-1, 1:2])
names(coefs)[2] = "se"
coefs$vars = rownames(coefs)
ggplot(coefs, aes(vars, Estimate)) +
  geom_errorbar(aes(ymin=Estimate - 1.96*se, ymax=Estimate + 1.96*se), lwd=1,
                colour="red", width=0) +
  geom_errorbar(aes(ymin=Estimate - se, ymax=Estimate + se),
                lwd=1.5, colour="blue", width=0) +
  geom_point(size=2, pch=21, fill="yellow")
```



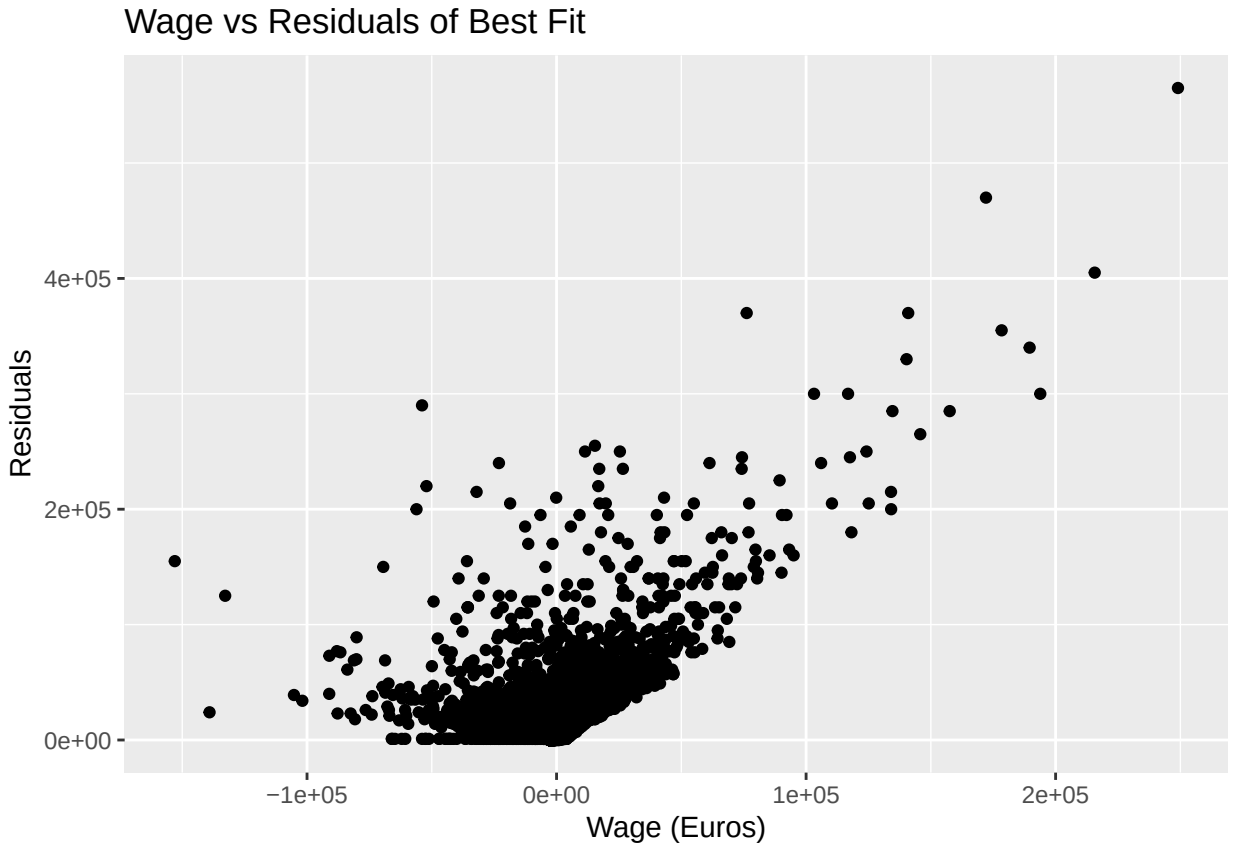
From the standard error plot, we can see the 95% confidence interval of the 4 different X variables in the FIFA 20 dataset. Clearly, overall and potential rating's confidence interval do not include 0. This results in those variables being significant. On the other hand, release clause and value include 0 in their confidence interval, resulting in those variables being insignificant. This is different than fit_best in which I found that all 4 X variables are significant.

Histograms of the Residuals

```
ggplot(data = data, aes(x = residuals(fit_best))) +
  geom_histogram(binwidth = 100) +
  xlim(-12000,12000) +
  ggtitle('Histogram of Residuals of Best Regression') +
  labs(x = 'Residuals', y = 'Count')
```




```
ggplot() + geom_point(aes(y = data$wage_eur, x = residuals(fit_best))) +
  ggtitle('Wage vs Residuals of Best Fit') +
  labs(x = 'Wage (Euros)', y = 'Residuals')
```



The histogram of the residuals are normally distributed which is the desired result. For the plot of Wage vs Residuals, the desired results are a homosketastic plot and to see if the residuals are dependent on wage. As we can see from the plot, the plot is not homosketastic but not super heterosketastic which is good. The most important part though, is that the residuals are not dependent on wage.

Conclusion

From the Histograms and Plots of wage against the different X variables I have chosen, we can see that there is definitely a correlation between wage and the X's I have chosen. For example in Wage vs Overall, Shooting, and Potential Rating, there is an exponential correlation between the X's and wage. From the regression analysis, I found that the best regression for finding wage included release clause, value, overall rating, and potential rating. From this regression I found that all those variables were significant resulting in a good model. Furthermore, the residual histograms and residual plots against wage support the fact that those variables are significant.