Final Project

Arnav, Casey, Harry, Maxx, Mohit

Introduction

Our data is from Kaggle: Football Player Data, the raw data contains 51 variables and 17954 observations. These variables contain many physical attributes as well as several skill based factors; both of these were pulled by the authors using data scraping tools. As a brief outline, our report will go over the following research questions:

- Which factors are the biggest overall contributors to overall FIFA rating?
- Do categorical ratings such as weak foot and skill moves affect overall FIFA rating?
- Does a player's overall rating contribute to their market value?
- What is the best model?

Loading and Cleaning the Data

After we load the data, we narrow down to 14 variables: overall_rating, weak_foot.1.5., skill_moves.1.5., sprint_speed, dribbling, strength, positions, height_cm, weight_kgs, stamina, value_euro, and wage_euro. We then look into NA values, and when we look at these specific 11 variables, there are only 255 NA's across only the wage_euro and value_euro columns. After we remove the NA's we still have 17699 ovbservations.

For the last research question we use training and testing data, additionally we use more variables. We take the full 51 variables. and split 20% of it for testing.

Removing Outliers:

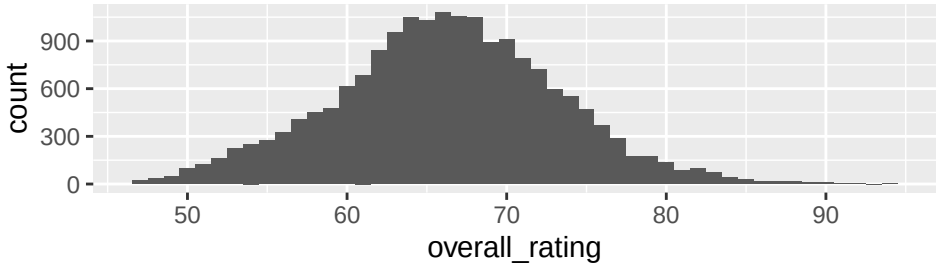
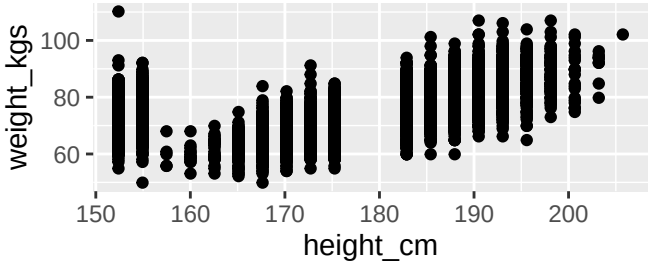


Fig. 1

As we can see, our main response variable, overall_rating, is approximately normally distributed and has no outliers.



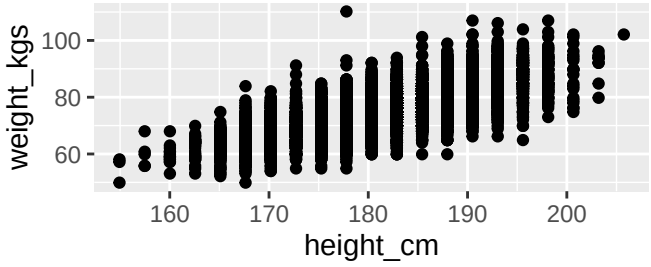


Fig. 2

We know from background information that height and weight should be positively correlated, and we expect to see that in this visualization. We mostly see this, but something is clearly off. We googled the heights to convert centimeters to inches and found that the outliers occur at heights 5'0" and 5'1". We also noticed that the data has an empty patch in the middle of the x axis, indicating that some heights are not included. We hypothesized that there was an error by the creators of the dataset in their data scraping process, in which 5'10" players were recorded as 5'0" and 5'11" was recorded as 5'1". The next code chunk shows us testing this hypothesis and seeing that 5'10" and 5'11" were indeed the missing heights, so we decided to rewrite the data set so that all 5'0" entries were corrected to 5'10" and all 5'1" entries were corrected to 5'11". However, the issue with this is that there could be a few players who really are 5'0" and 5'1" in the dataset and they need to be recorded with their accurate height. To solve this problem, we looked through the original data source (https://sofifa.com/players?col=hi&sort=asc&r=190043&set=true) and searched for 5'0" and 5'1" players. There were only 3, so we manually filtered by name to get their true height reflected in the dataset. After the following chunk the height column will have accurate data.

Now our goal is to detect, analyze, and assess whether outliers in our dataset should be removed or mitigated in another way. To do this, we examined overall rating, height, and weight for potential extreme values. We began by generating boxplots for key numerical variables to visualize any extreme values.

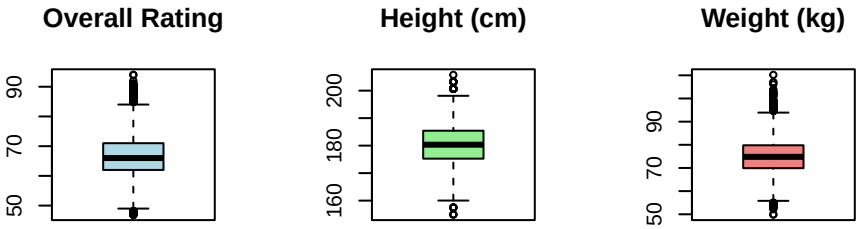


Fig. 3

From these boxplots we can see that height seems to be approximately distributed. However, there are outliers at both extremes for both Rating and Weight, with players rated above 85 and below 50 while in weight, some players are below 55kg or above 95kg. To anaylze these outliers further and see if they have an effect, we can identify the most extreme values and look at them:

The highest-rated outliers include Lionel Messi (94), Cristiano Ronaldo (94), and Neymar Jr. (92). Since these values accurately reflect the skill level of these players, they should not be removed. On the other end, the lowest-rated players, such as N. Fuentes and S. Squire (47 overall rating), are likely reserve or youth players. While they appear as outliers, they are realistic and do not indicate data errors.

The heaviest players include Adebayo Akinfenwa (110.2 kg) and multiple goalkeepers who exceed 100 kg. Given that goalkeepers and defenders tend to be heavier, these values are valid and should not be removed. For the lightest players, we found that Kazuki Yamaguchi and B. Al Mutairi weigh around 49.9 kg, which aligns with expectations for smaller midfielders and forwards. These values are not errors and should also be kept in the dataset.

While our analysis confirmed that the outliers in overall rating and weight are valid, extreme values can still influence the regression model. Instead of removing them, we can apply methods to reduce their impact while keeping the dataset intact. The first step was to log-transform weight to make the distribution more normal.

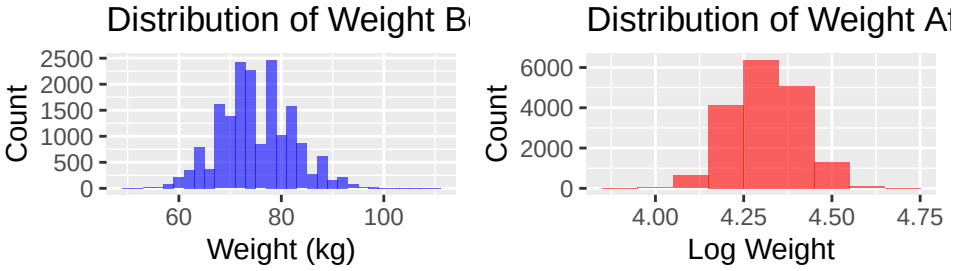


Fig. 4

Before applying the log transformation, the histogram of weight showed a clear right-skewed distribution, where a small number of very heavy players disproportionately influenced the overall weight distribution. After the transformation, the distribution became much more symmetrical, resembling a normal distribution. This adjustment ensures that extreme weight values do not overly influence the regression model while preserving all player observations.

Traditional ordinary least squares (OLS) regression is highly sensitive to extreme values. This means that a few very heavy or light players could disproportionately influence the regression coefficients. To prevent this, we implemented robust regression, which assigns less weight to extreme values while still considering them.

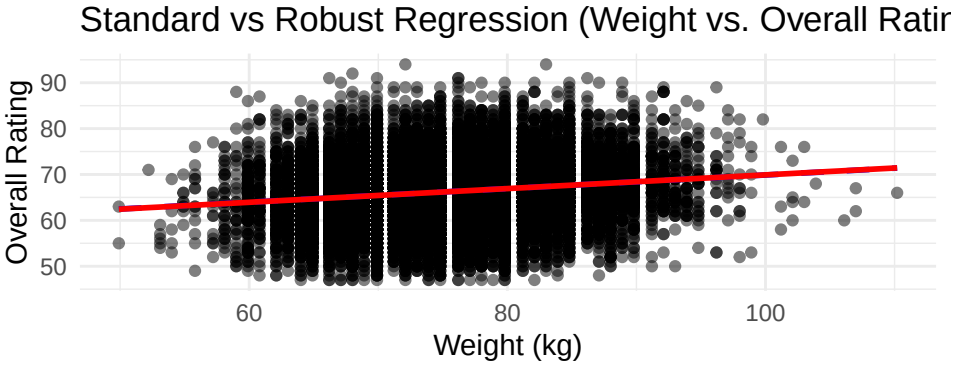


Fig. 5

The results of this comparison show a clear difference between standard regression and robust regression. In the scatter plot, the blue dashed line represents standard OLS regression, which is strongly influenced by extreme values. The red solid line represents robust regression, which follows the general trend of the data but is less affected by extreme values. By implementing robust regression, we ensure that our model is less sensitive to extreme weight values, making it more stable and interpretable.

To further validate whether keeping outliers affects model accuracy, we compared models with and without extreme weight values by examining the adjusted R^2 values and AIC information.

The Adjusted R^2 values for the full model (0.2985469) and the clean model (0.2991566) are nearly identical, suggesting that removing extreme weight values does not improve the model's explanatory power. The AIC values show a slight decrease when outliers are removed, with the clean model at 1.1213653×10^5 compared to 1.1271212×10^5 for the full model. While a lower AIC generally indicates a better model, the difference of about 580 points is relatively small considering the dataset size. This suggests that the improvement from removing outliers is minor and does not justify eliminating valid data points.

Conclusion: Since removing outliers does not significantly improve predictive accuracy, we retain all players while applying transformations and robust modeling to ensure a more stable regression analysis. These adjustments allow us to preserve real-world player data while preventing extreme values from skewing the results, ultimately leading to a more reliable model for predicting FIFA player ratings.

Response to Teacher Question 3:

"With such a large sample size, you will find a lot of significant predictors. Discuss why."

Our sample size is over 17000 (17898). This can result in many of our predictors being statistically significant because a large sample size leads to an even more precise discovery of relationships between variables, reducing variance. Thus, with a very large sample size, even if the predictor's impact is smaller, it could still be considered statistically significant. If we think in terms of confidence intervals, they become extremely narrow because the margin of error decreases. While a predictor could be statistically significant, it may not be an impactful predictor of the output in reality.

One way we can tackle this problem is by using corrections such as Bonferroni, Holm, or Benjamini Hochberg to lower false positive rates. Also, we used training and testing split in predictions, which can reduce the number of

samples during training and prevent overfitting to just the training data (and performing poorly on new data), and used AIC to choose models as well.

Research Question 1:

Which factors are the biggest contributors to overall FIFA ratings?

Here is how we set up the model

```
## lm(formula = overall_rating ~ weak_foot.1.5. + skill_moves.1.5. +
##     sprint_speed + dribbling + strength + stamina + log(value_euro) +
##     log(wage_euro), data = fifa)
```

The coefficients are:

```
##      (Intercept)   weak_foot.1.5. skill_moves.1.5.     sprint_speed
##      2.722617158     0.131623882    -0.028740319    -0.039284298
##        dribbling         strength          stamina  log(value_euro)
##     -0.004952933     0.043845834     0.015619906     4.139578778
##   log(wage_euro)
##      0.690921812
```

The R-squared for this model is: 0.9060477

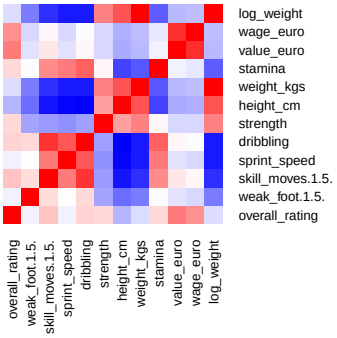


Fig. 6

So we started by looking at the relationship between each of our chosen predictors and the output variable of overall_rating

We saw roughly linear relationships for most cases. However, for value_euro and wage_euro, our scatter plots look very similar to a logarithmic curve.

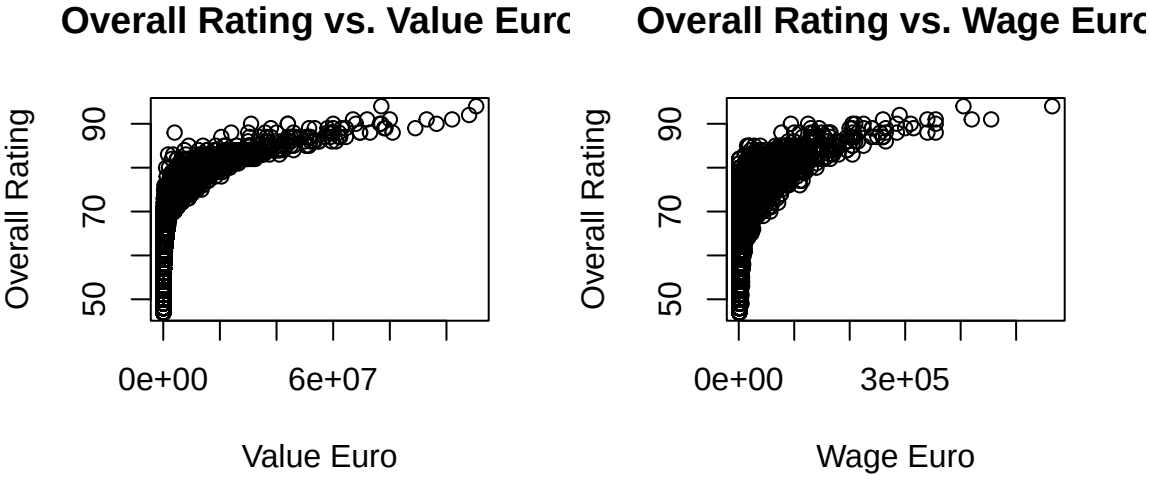


Fig. 7

Thus, we need to apply a transformation to our data and instead use ln(value_euro) and ln(wage_euro)

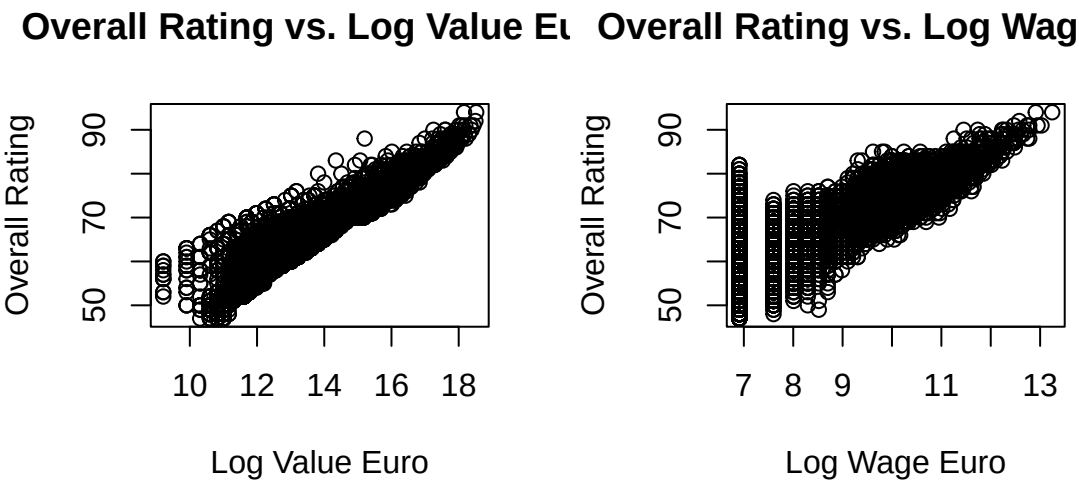


Fig. 8

This shows a much better linear relationship.

Now let's look at the R value or correlation coefficient of each variable to overall_rating:

```
##              Variable Correlation
## 1          Weak Foot   0.2161278
## 2        Skill Moves   0.4194500
## 3       Sprint Speed   0.2191865
## 4          Dribbling   0.3799389
## 5           Strength   0.3625752
## 6            Stamina   0.3704942
## 7 Log(Value in Euro)   0.9406430
## 8  Log(Wage in Euro)   0.8087885
```

Table 1

We can see we found particularly high values for log(value_euro), log(wage_euro).

Now, let us form our linear regression model and utilize an F test to see our results.

```
##                      Estimate      Pr(>|t|)
## (Intercept)       2.722617158  8.233291e-50
## weak_foot.1.5.    0.131623882  4.704007e-07
## skill_moves.1.5. -0.028740319  4.769179e-01
## sprint_speed     -0.039284298 1.069677e-113
## dribbling        -0.004952933  1.082617e-02
## strength          0.043845834 9.450453e-182
## stamina           0.015619906  1.065340e-21
## log(value_euro)   4.139578778  0.000000e+00
## log(wage_euro)    0.690921812 1.541023e-204
```

So, we found p-values < 0.05 (our alpha/significance level) for the following predictors: weak_foot, sprint_speed, dribbling, strength, stamina, log(value_euro), log(wage_euro) meaning that these predictors were found to have a statistically significant impact on overall_rating. Because skill_moves had a p-value > 0.05 (0.4769), it means it is not statistically. Thus, we will create a new model without this predictor and compare results. This likely means the other variables also account for the impact made by skill moves. Also, in this model we had a very high Adjusted R^2 of 0.9060477 which is much higher than our base model's R^2 of 0.5495 and means our model does is effective at explaining the variance in overall_rating.

```
##                     Estimate      Pr(>|t|)
## (Intercept)      2.733830750  1.445828e-50
## weak_foot.1.5.   0.130438707  5.646477e-07
## sprint_speed    -0.039334164 3.654369e-114
## dribbling       -0.005805689  1.477493e-04
## strength         0.043944476 4.729041e-184
## stamina          0.015644048  9.040184e-22
## log(value_euro)  4.137358847  0.000000e+00
## log(wage_euro)   0.690571626 1.932680e-204
```

Here we can see we only have significant predictors and the same R^2 of 0.906045

Running an ANOVA to compare our two models (model 2 has one less predictor):

```
## Analysis of Variance Table
##
## Model 1: overall_rating ~ weak_foot.1.5. + skill_moves.1.5. + sprint_speed +
##     dribbling + strength + stamina + log(value_euro) + log(wage_euro)
## Model 2: overall_rating ~ weak_foot.1.5. + sprint_speed + dribbling +
##     strength + stamina + log(value_euro) + log(wage_euro)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1  17690 80885
## 2  17691 80888 -1   -2.3133 0.5059 0.4769
```

As we can see from the ANOVA, we had a p-value of 0.2091. Remember, in the ANOVA test, our null hypothesis is that there is no difference between our models. Our p-value being > alpha 0.05 proves that we fail to reject the null and do not have significantly significant evidence there is a difference between the two models. We can interpret this as meaning that our smaller model is as effective as the larger one and is a better choice for us to use since it is more condensed.

Research Question 2:

Do categorical ratings such as weak foot and skill moves affect overall FIFA rating?

The columns weak foot and skill moves represent those respective ratings in FIFA. The ratings are a number 1-5 out of 5 stars. So this is a discrete numeric variable. First, we will do a visualization of the data to get an idea of what our data suggests. Because we only have 5 categories for rating, we can do a boxplot of overall rating for each category.

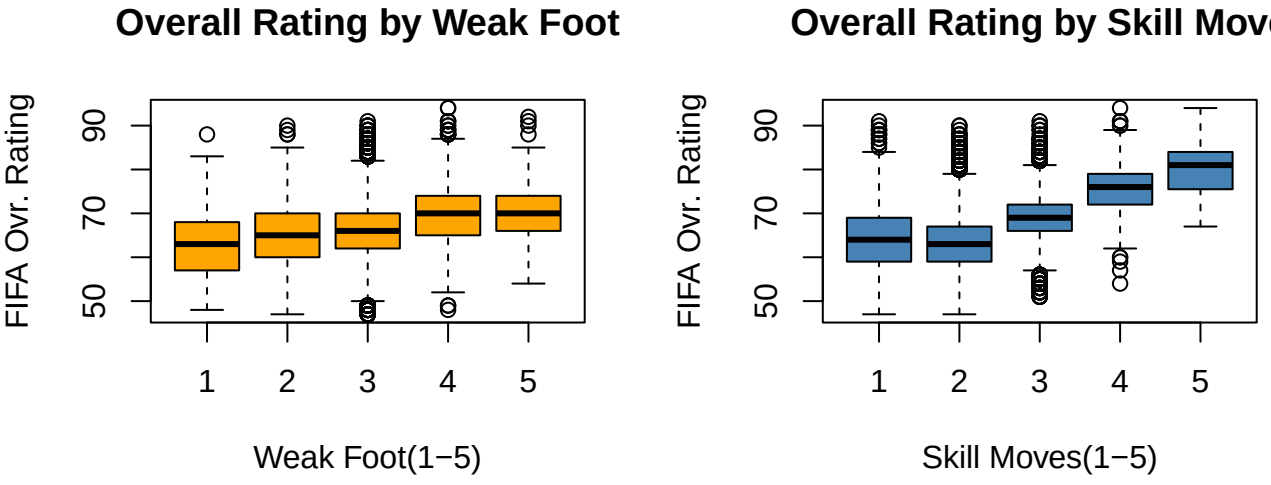


Fig. 9

Both these plots indicate that there is some correlation between these discrete predictors. For the most part, the quartiles and medians gradually increase as we move up by 1 star, and we can see a noticeable difference between 1 star and 5 stars, particularly in skill moves. However, the boxplots are labeling many points as outliers on

almost every boxplot. We already examined the overall rating column for outliers, so we know these points belong in our data. It makes me wonder if this is a case of correlation not causation, because these points seem unaffected by the respective discrete predictor. Additionally, it could be that these points have some other factor that is causing these discrete ratings to not matter.

We will run a linear regression on just these two predictors to expand on these comments:

```
##                     Estimate    Pr(>|t|)
## (Intercept)      55.2404094 0.00000e+00
## weak_foot.1.5.    0.8592129 1.92673e-29
## skill_moves.1.5.  3.5786017 0.00000e+00
```

This confirms what we found looking at the boxplots. This linear model found that both predictors are definitely statistically signficiant, as evidenced by the p-values. Based on the estimate for $\hat{\beta}_i$, it is indicated that skill moves has more correlation with overall rating, which is consistent with what we saw in the plots. Both are positively correlated. The model states that, if skill moves rating is held constant, for each star increase of weak foot rating, we can expect overall rating to increase by 0.8555. Also, if weak foot rating is held constant, for each star increase of skill moves rating, we can expect overall rating to increase by 3.5636, which is quite a significant increase.

However, we know just including these two variables is not going to be our best model. A more practical question is does including these variables in our base model making it significantly better vs without them. To test this, we will use an f-test using the anova test function.

```
## Analysis of Variance Table
##
## Model 1: overall_rating ~ sprint_speed + dribbling + strength + stamina +
##     log(value_euro) + log(wage_euro)
## Model 2: overall_rating ~ weak_foot.1.5. + skill_moves.1.5. + sprint_speed +
##     dribbling + strength + stamina + log(value_euro) + log(wage_euro)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  17692 81002
## 2  17690 80885  2    116.84 12.777 2.853e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The f-test indicated the base model was significantly better than the model removing discrete predictors, with the pvalue very close to 0. This indicates beyond a reasonable doubt that we should include these predictors in our model, they can be used to predict overall rating.

I'm still curious about the outliers we saw in the boxplots. I wonder if there are some positions for which these predictors don't matter.

Let's look at the same plot for only goalkeepers:

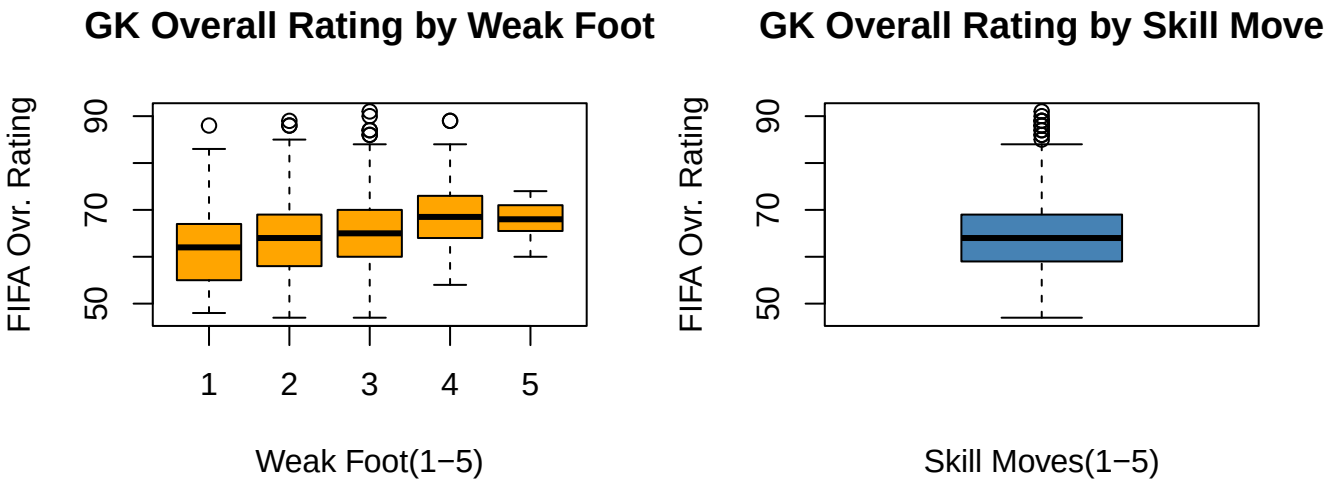


Fig. 10

It appears that the correlation for weak foot is weaker than for all positions. More significantly though, we found that every goalkeeper only has one star skill moves. This means that skill moves rating will have no predictive value for goalkeepers, and this is also effecting the signficance of this variable on our base model.

We also saw some similar issues in the data for centerbacks(CBs). Weak foot seems to have very little correlation with overall rating for CBs, and every CB in the game only have a skill move rating of 2 or 3.

*It is worth noting that the positions column lists all the positions a player can play in one string. So by filtering `positions == CB`, we are selecting players who play CB as their only position. Domagoj Vida, for example, has positions "CB,RB", so he is not included in this filter even though his primary position is CB. We played around with using the stringr library to include these players as well, but the results weren't as meaningful, so we chose to filter players who play exclusively CB, which is the majority of CBs. This wasn't an issue for goalkeepers as they all only play goalkeeper; outfield positions are more fluid creating this issue.

Repeating our analysis without GKs and CBs:

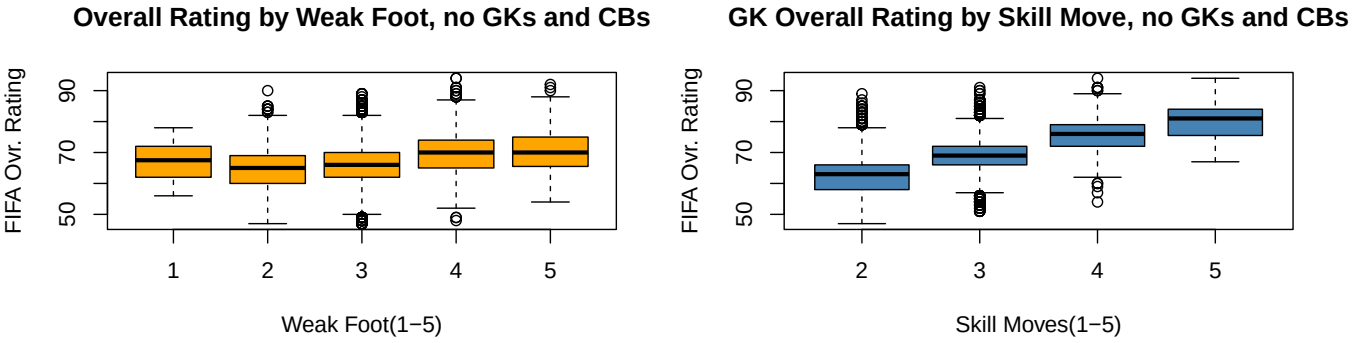


Fig. 11

```
##                    Estimate     Pr(>|t|)
## (Intercept)       47.041607 0.000000e+00
## weak_foot.1.5.     1.022446 8.582176e-41
## skill_moves.1.5.   6.185497 0.000000e+00
```

Based on the new boxplots, filtering by position didn't affect weak foot that much, but the skill moves plot looks very strongly correlated now. These observations are backed up by the updated linear model, which has a slightly higher $\hat{\beta}$ for weak foot but much higher for skill moves.

In conclusion, there is only a small amount of positive correlation between weak foot and overall rating. It is unclear if weak foot really has an effect on overall rating or if the two variables are just correlated. On the other hand, there is a strong positive correlation between skill moves and overall rating. Skill moves is a strong predictor of overall rating. It is either having a strong effect on overall rating, or it is very strongly correlated with a variable that is, such as dribbling or ball_control. However, all goalkeepers in FIFA have one star skill moves, and they are the only players in the game with this. Thus, skill moves is only a useful predictor for non goalkeepers, and using it in a dataset with goalkeepers will make it less effective of a predictor. A similar effect also applies to centerbacks, although less extreme.

Research Question 3:

Does a player's overall rating contribute to their market value?

Our R-squared value is 0.6809234. We have a moderately strong positive relationship between a player's overall rating and their market value from a correlation of 0.6309281, As we can see, this is horrendous. Let's look into a transformation!

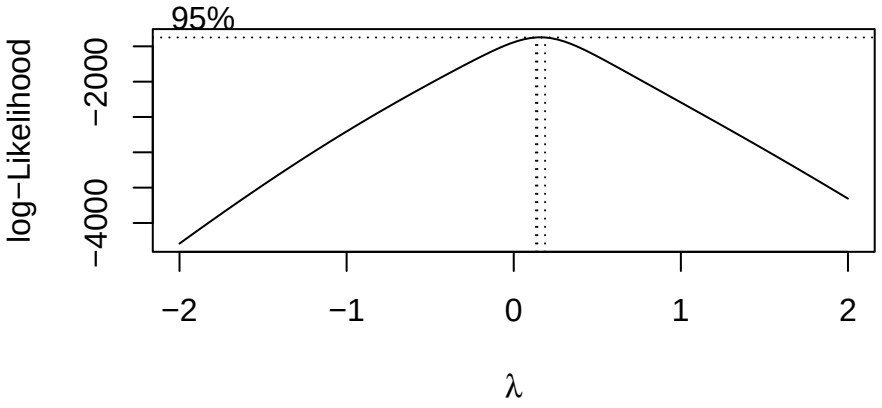


Fig. 12

We can see $\lambda = 0$ so let's try a log transformation!

Our log-transformation correlation is: 0.940643, and our log transformed R-squared is: 0.9067223. Our correlation suggests a strong relationship between player overall rating and market value after taking the log-transform of market value. Our r squared is much higher now too.

Research Question 4:

What is the best model?

We use adjusted R squared and AIC to find best model because a model that has more predictors obviously will have a higher R squared (proportional of variability explained by the model)

We make sure to be wary of our predictors that are significant, looking out for multicolinearity.

To answer the question of what is the best model, let us define our primary goal, that being prediction. So we want to fit our models to the training data and compare predicted results to the test actual responses.

We first make a guess of a good explanatory model based on our predictor knowledge after looking at the data:

```
##                      Estimate      Pr(>|t|)
## (Intercept)      1.055255e+01  2.828298e-18
## sprint_speed    -7.188757e-03  1.656826e-01
## dribbling       -4.521466e-02  1.829156e-13
## height_cm       -7.986574e-03  3.998945e-02
## log(value_euro)  3.219823e+00 6.153954e-220
## wage_euro        1.375896e-05  1.260399e-40
## national_rating  6.608623e-02  7.057406e-07
## crossing         1.513669e-02  1.140701e-03
## balance         -7.776278e-03  1.158688e-01
## jumping          1.427330e-02  2.175134e-04
## reactions        1.674184e-01  7.579432e-49
## penalties        9.156495e-03  2.910830e-02
```

Not a bad guess but let's now apply our best model tests and now do so on our generated training data!

```
##  [1] 0.9065747 0.9580267 0.9669220 0.9730219 0.9752599 0.9770353 0.9780083
##  [8] 0.9783318 0.9785972 0.9788521
```

From our output from above, we can get some really good models with adjusted R squared near 0.978. But we can do even better using the step function for AIC! First let's define our "complete" model on our training data, which uses nearly all predictors in the original data set besides a few we discussed that were not good and others we didn't need.

Then use step-wise selection for both directions

```
##                                   Estimate      Pr(>|t|)
## (Intercept)                   3.275989e+00  0.000000e+00
## age                           3.956606e-03  4.727365e-94
## log(value_euro)               5.209675e-02 1.581632e-254
## wage_euro                     5.904492e-08  9.189892e-07
## international_reputation.1.5. 3.888662e-03  1.409974e-05
## national_rating               7.805619e-04  2.775376e-08
## heading_accuracy             -1.014540e-04  3.716938e-03
## short_passing                -1.956050e-04  1.189831e-02
## sprint_speed                  1.751120e-04  2.831124e-04
## reactions                     8.340019e-04  6.729210e-11
## long_shots                   -1.125783e-04  2.621743e-02
## positioning                  -1.778064e-04  1.432103e-03
## composure                     2.158725e-04  6.870939e-03
## marking                       1.397934e-04  3.432288e-05
```

We got an exceptionally high adjusted R-squared value of 0.9803667, excellent!

Response to Teacher Question 4:

"Is your goal prediction or interpretation? I think prediction is suitable for your dataset. Compute prediction intervals and you could also use a training/test approach."

The best model for prediction isn't necessarily the best model found for analysis via stepwise regression or regsubsets(), but in this case, using our best analysis model works really well. Now that we have our best model, let's do some predicting:

```
##    Actual Predicted Lower_Bound Upper_Bound
## 4      88        87    85.12185    89.03771
## 8      89        89    87.29811    91.36712
## 11     89        89    87.13313    91.24975
## 16     89        88    85.83965    89.81464
## 21     87        87    84.58119    88.47715
## 24     87        88    85.73346    89.68389
```

```
## Root Mean Squared Error (RMSE): 0.8980265
```

Our MSE (and RMSE) is low, as we are able to consistently predict a player's rating or be really close to their actual. We decided to round our predicted values to help with interpretability, only increasing our MSE slightly by doing so.

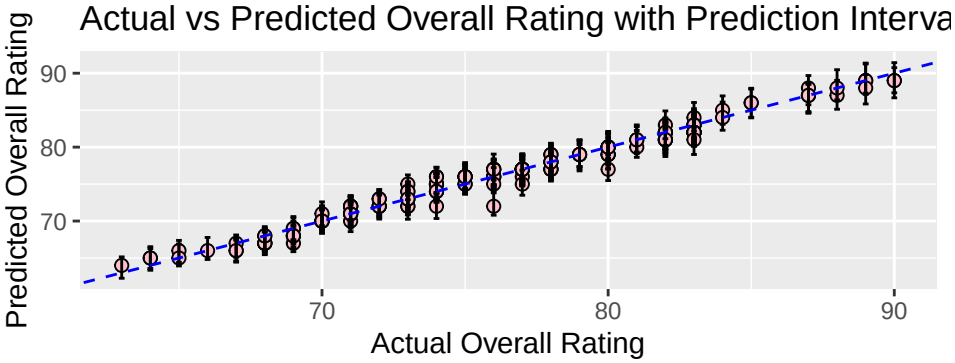


Fig. 13

Contributions

Itroduction: - Introduction: Maxx - Loading and Cleaning the Data: Maxx - Removing Outliers: Harry - Response to Teacher Question 3: Arnav

Research Question 1: Arnav

Research Question 2: Harry

Research Question 3: Casey

Research Question 4: Casey

Teacher Question 4: Casey and Mohit

Organization and formatting: Maxx