Lab 2: Analyzing Fifa 2022 Data

w203: Sophie Yeh, Nathaniel Browning, Torrey Trahanovsky

April 12, 2022

1. An Introduction

With millions of players and billions of fans, football (also known as soccer in the United States) is the world's most popular sport both to play and to watch. Due to its popularity, professional football clubs have evolved into essentially companies with shareholders and managers. Fans have become customers that help clubs generate immense revenue. From a managerial perspective, one of the most important decisions that determine a club's chances for success is which players to employ and ensuring that they pay wages that correctly factor in the revenue the clubs will generate.

It's key for clubs to understand the market value of players on their team and ensure that as a club, like any company, has increased revenue, they focus on growing for shareholders and having the top talent in their club. Additionally, due to the nature of football having many variables and factors deciding a players ability on the field and value, it's easier to localize solely on players with the role goalkeepers to control their play expectations on the field. Thus, the main question we seek to answer is:

Research Question: In 2022, how does a goal keeper's performance, background, and physical attributes affect their market value in football?

To find the answer, the standard for data on how players play, are predicted to play, and their compensation is the Fifa dataset as this closely resembles the real sport variables fairly accurately and applies variables to skills. Fifa stands for the International Federation of Association Football (FIFA) and is an important body in the field of football.

Additionally, our measurement goal is to try to measure the combination of variables that are not accounted for in the overall ranking variable in Fifa or in the value variable. These variables being Preferred Foot,Nationality, Potential, and International Reputation. With the goal of seeing to measure how these variables might correlate with a players ranking and thus their value. We seek to answer whether a player with a preferred left-foot or perhaps from a certain country, might have a higher value than another player with other variables held constant. Another end goal is to hopefully provide the outcome of the research to clubs so that they can better factor into other perhaps unconsidered characteristics into their clubs and have a better overall club performance resulting in improved football performance in the professional leagues. This could also be phrased as another question of *"How can clubs benefit from improved understanding of value and other player attributes?"*.

The dataset Fifa 2022 data is quite well poised to answer the research question and provide measurements for the goals of the research. During the dataset exploratory data analysis, it's also expected that some variables and the way they interact may result in having an effect on player value or another skill that is unexpected yet important to note.

2. A description of the Data and Research Design

The data used to answer the main research question and measurement goals is composed of x and y variables. In which the X variables are: Age, Height, Weight, Preferred Foot, Overall Ranking (# overall), nationality,

potential, and international reputation. While the Y variable is Value, the market value of a player, which is precisely the contract transfer fee between clubs. The primary design of the research is causal and explanatory, attempting to analyze collinearity between various x variables and see how they might indicate whether the y variable or other variables are affected by variances. The secondary type of research is exploratory in which during exploratory data analysis, various graphs, tables, stargazer, tests, and other visualization techniques are used to understand the data. To complement the analysis, researching the sport of football and what such stats and variables mean are key as well to ensure the understanding of the data is correct. Also, researching what types of variables typically comprise value in a player in the sport among various journals is key as well. The analysis will further process our data in order to fit a linear regression model between our key variables and market value. The scope of our research is limited to three linear models that will provide insight on factors affecting market value. Finally, we validate our model through statistical tests and diagnostic plots and pinpoint model limitations.

Regarding the specific dataset, we will be using the FIFA22 dataset as previously introduced, specifically from Kaggle. This public dataset scrapes 2022 data from the official FIFA Index database and does not lose any data from the site while also maintaining a yearly view of stats. Player stats and scores are released by the official FIFA organization that governs international football.

Dataset (Public): https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database

Also, Journal sources where topic and Fifa learning came from:

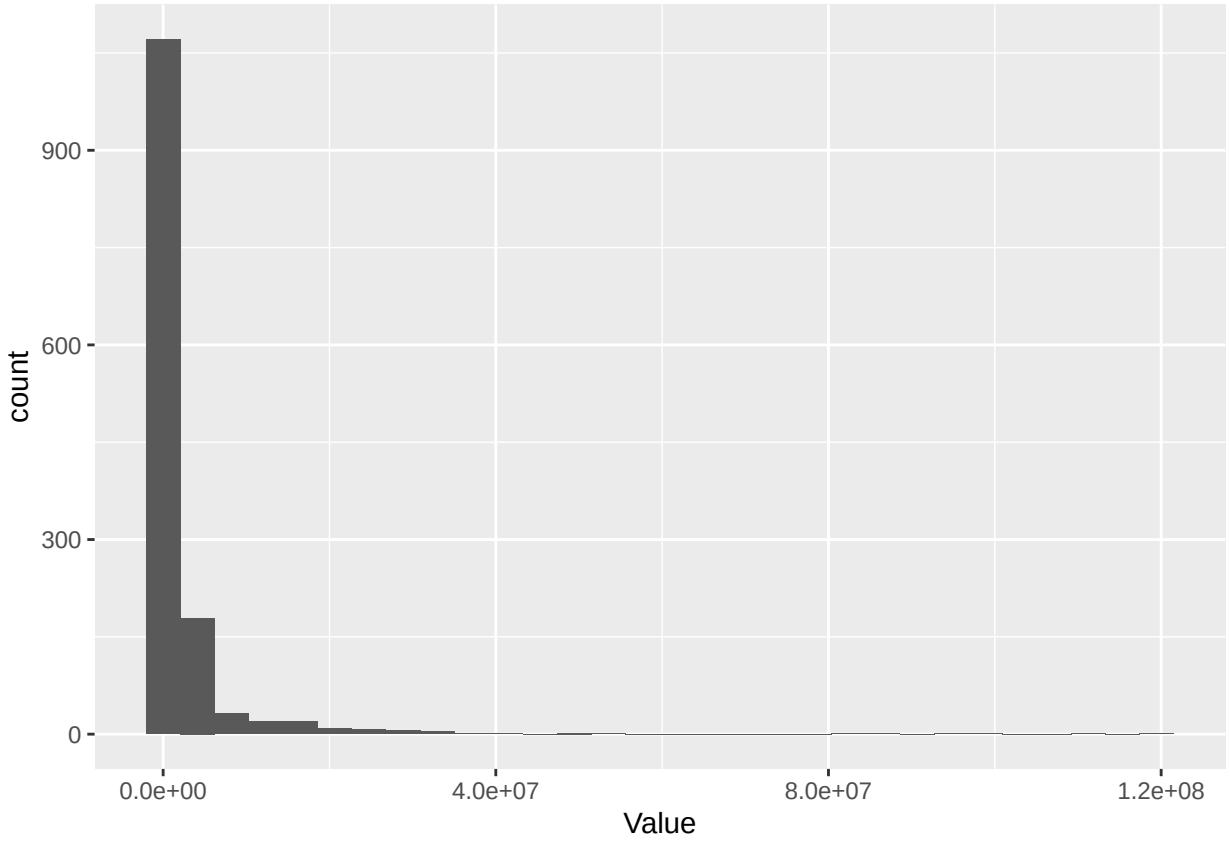
Journal One: https://home.kpmg/ch/en/blogs/home/posts/2018/08/how-much-do-you-value-your-favorite-football-star.html

Journal Two: https://medium.com/analytics-vidhya/fifa19-dataset-analysis-6837664cee89

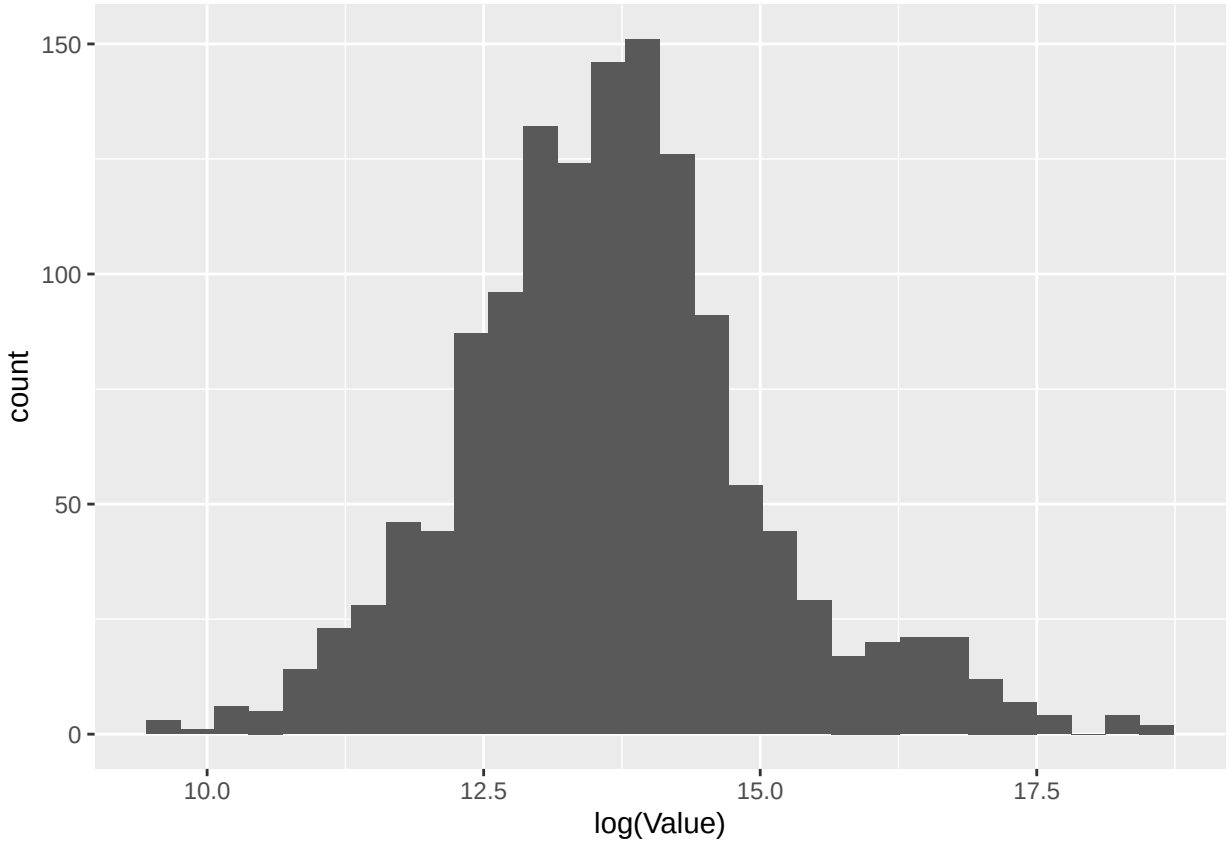
2a. A Model Building Process

The research question investigates factors influencing the market value of a player, and this is represented by the y variable, the 'Value' in the FIFA22 dataset.

The non-log histogram of "Value" shows a heavy skew in terms of actual pay, however, this is skewed by outliers.



To offset the outliers, a semi-log histogram of `Value` reveals that `log(Value)` is normally distributed. Thus in our linear model, we will be creating a linear model for `log(Value)` in order to fit the OLS IID assumption.



Several journal articles in our preliminary background research all identified similar key variables affecting player market value. Based on this preliminary research, we have identified key variables in the dataset in which we hypothesize will affect market value: `Age`, Height`,`Weight`and`Special‘. First, aging comes with physical limitations and increased chance of injury that can cause a player's value to decrease. Second, a player's height and weight itself, regardless of performance, may be more valued due to team managers' and scouts' bias. Performance is accounted for by the variable 'Special'. 'Special' has a high correlation with all the performance variables in the dataset. Although it is best practice to use causal variables that don't have other column variables affecting it, many performance variables overlap each other in different aspects, reverse causality and overfitting can become an issue. Instead, 'Special' is a composite representation of performance provided by experts based on professional considerations and priorities.

As for the actual aspects that need to be measured and the potential covariates for modeling goals, the collinearity and potential problem covariates need to be assessed. When we assess for collinearity, we run two main statistical tests to ensure that we don't have problematic covariates. The first test being a correlation matrix, and the second being a test for variance inflation factor. We can run the correlation matrix for values that are metric or on a scale. When running the matrix, we find that a few variables have correlation to each other. The first correlation we see is between the Height and Weight covariates. Due to these potentially playing an integral role in the decision behind Value, we want to ensure we look for the variable VIF values across our models later in the analysis. The second correlation we find is between the covariate "Special". This variable had extremely high correlation to factors regarding a players skill rating (Handling, Kicking, etc). We find that we might be able to replace these skill factors with the Special covariate that is likely to better fit the model. We will check for this later through ANOVA test analysis of our different models.

Finally, the Goalkeeper skill ratings were very highly correlated with each other and present reason to believe that these variables might contribute to violation of the collinearity assumption. Thus we conclude that it is safe to remove it from consideration for our models.

When assessing VIF, we have two major findings that we must factor in when considering our models. When

the covariate, Age, is added to the model, the VIF is found to exceed the threshold for meeting the colinearity assumption. Thus, we only include the quadratic form of the variable and remove the original covariate from our later models. Once the age variable has been removed from our models, when we reassess for VIF values, we find that no covariates exceed the threshold for concern. With this being said, the 3 models constructed include the following variables as shown below in the stargazer:

Stargazer regression table:

```
##
## ================================================
##                            Dependent variable:
##                        ---------------------------
##                                 log(Value)
##                           (1)       (2)       (3)
## ------------------------------------------------
## Age                     0.306***  0.447***  0.742***
##                          (0.043)   (0.039)   (0.043)
## I(Age2)                 -0.008*** -0.010*** -0.015***
##                          (0.001)   (0.001)   (0.001)
## Height                  0.030***  0.016***    0.012
##                          (0.007)   (0.006)   (0.007)
## Weight                    0.002    0.0003    -0.006
##                          (0.005)   (0.004)   (0.005)
## Special                 0.007***  0.006***
##                         (0.0002)  (0.0002)
## Contract.Years                    -0.193*** -0.263***
##                                    (0.019)   (0.021)
## Agility                                       0.003
##                                              (0.003)
## Strength                                     0.013***
##                                              (0.003)
## Jumping                                      0.014***
##                                              (0.003)
## Acceleration                                 0.009***
##                                              (0.003)
## Stamina                                       0.008*
##                                              (0.004)
## Weak.Foot                                    0.121***
##                                              (0.036)
## International.Reputation           0.847***  1.142***
##                                    (0.053)   (0.059)
## Constant                -2.395**   -0.867    -0.889
##                          (1.149)   (1.023)   (1.279)
## ------------------------------------------------
## Observations              1,358     1,358     1,358
## R2                        0.557     0.655     0.543
## Adjusted R2               0.555     0.653     0.539
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

The covariates that had transformations applied to them for our work in the model included Value, Wage, and Age. Value and Wage were covariates that are monetary values measured in euros that had a very heavy left skew. In order to adjust this for our model, we made sure to take the log of both covariates. This provided us with a more normal distribution of the data in order to better fit our model.

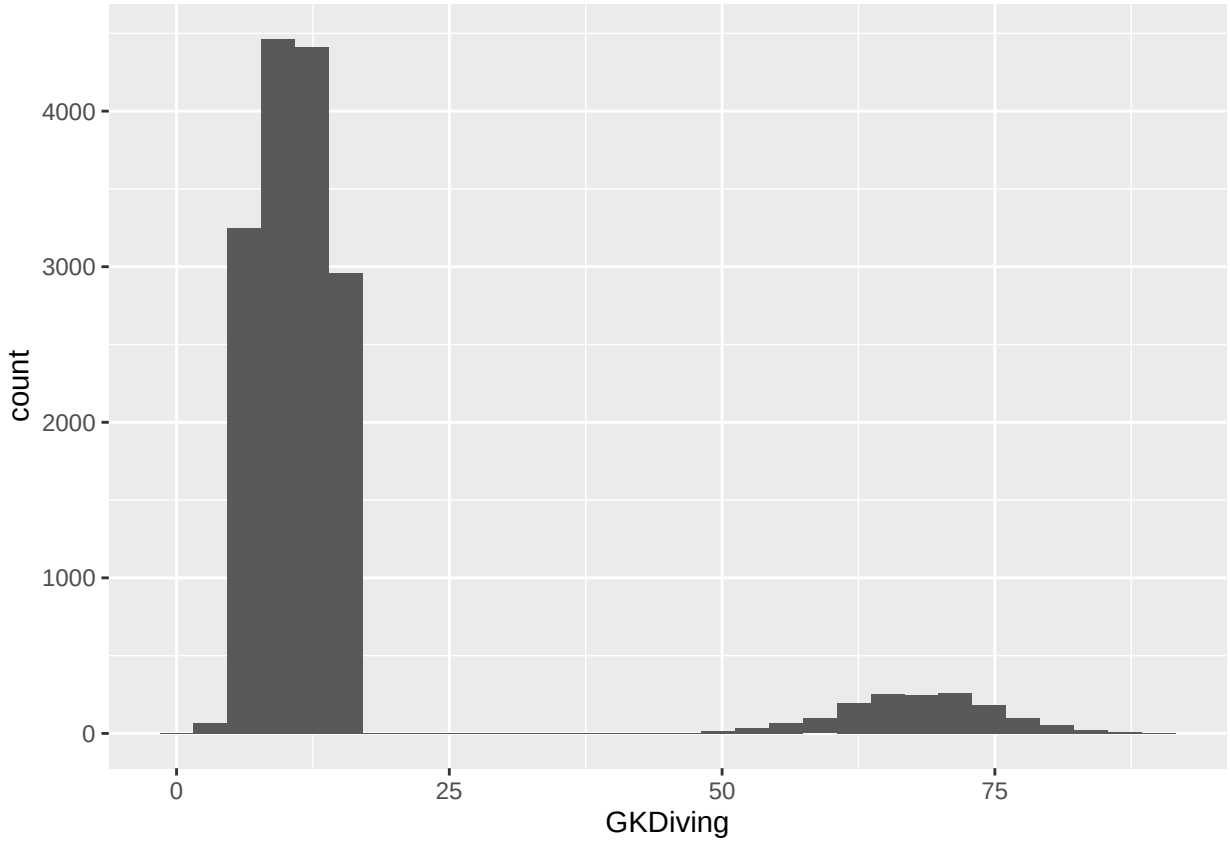
The original covariate for Age also proved to have a nonlinear relationship with the success metric of our study, Value. Thus, we needed to use the quadratic term for the variable which allowed for us to include it in the model without violating any necessary assumptions.

Also, it's clear that our choices are supported by EDA. We were able to detect that the overall rank and GK skill sets were not the only indicators used to determine the true overall rank of goalkeepers. Nonetheless, every rank would rarely be higher for someone with significantly less skill in a GK category than another GK player in the same category. Regarding Nans we found that of GK's, there were quite a lot of missing values in 2 columns: Loaned From and Marking. After these columns were removed, very few Nans existed, and when all rows with Nans were removed only about 100 players were removed. Additionally, more anomalies such as thirty-eight rows with players having no "Value" were also in the dataset that were removed, as players with no pay would likely be an error or skew the data. Also, some variables did have issues, such as weight, height, wage, and value. These had decimals or alphabetic characters used to represent numerical data. We had to alter this data to be fully numeric so it could be utilized and pad zeros to represent the numbers correctly.

Once we noticed and addressed these aspects, we proceeded to analyze the data by graphing with histograms and plots, this showed aspects such as skewed data, biased data, and areas where the best reflection of overall rank and skills could be seen. We were also able to understand what factors might indicate a well paid player or skill. Age was the least biased while other aspects such as wage were the highest biased among GKs. For statistical tests, t-tests were the easiest to perform and understand on the various variables. The p-values were however quite low due to the data not matching exactly for the tests being used. We proceeded to use our models to better understand the types of correlations and impacts of our data and how they relate. Also, by incorporating with stargazer we were better able to understand our data.

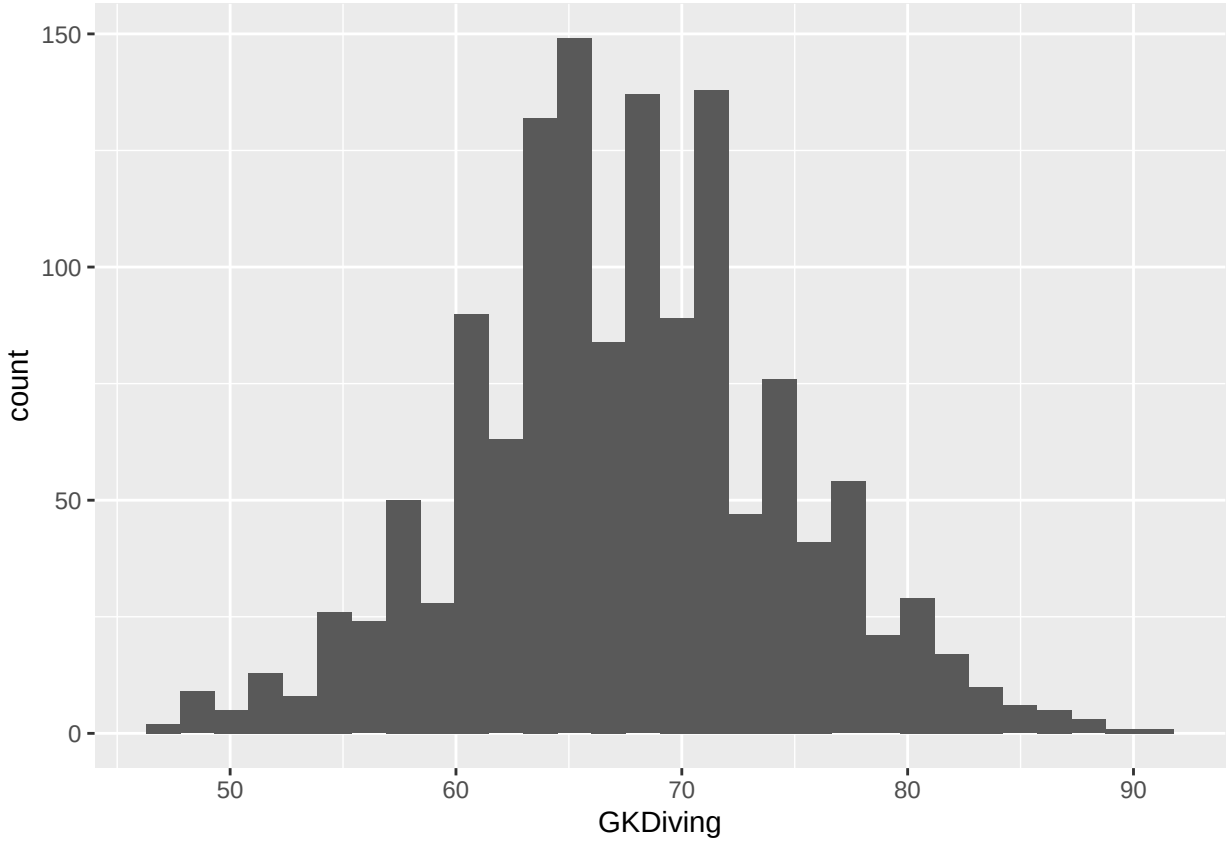
During our EDA there were some stand-out plots and analysis that are worth highlighting.

For example, the GKDiving histogram shows a heavily skewed distribution between GK normal distribution and the extremely low non-GK distribution when applied to the entire dataset of both non-GK and GK role

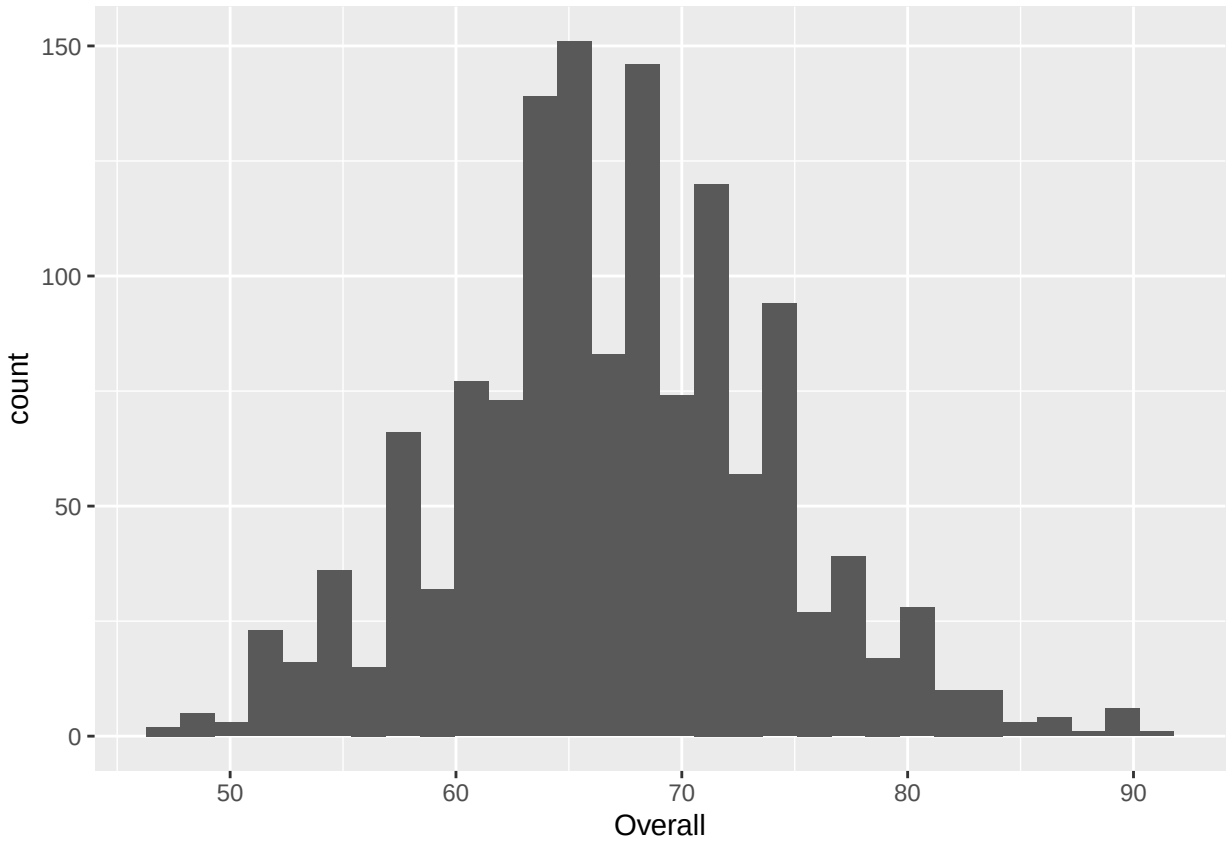


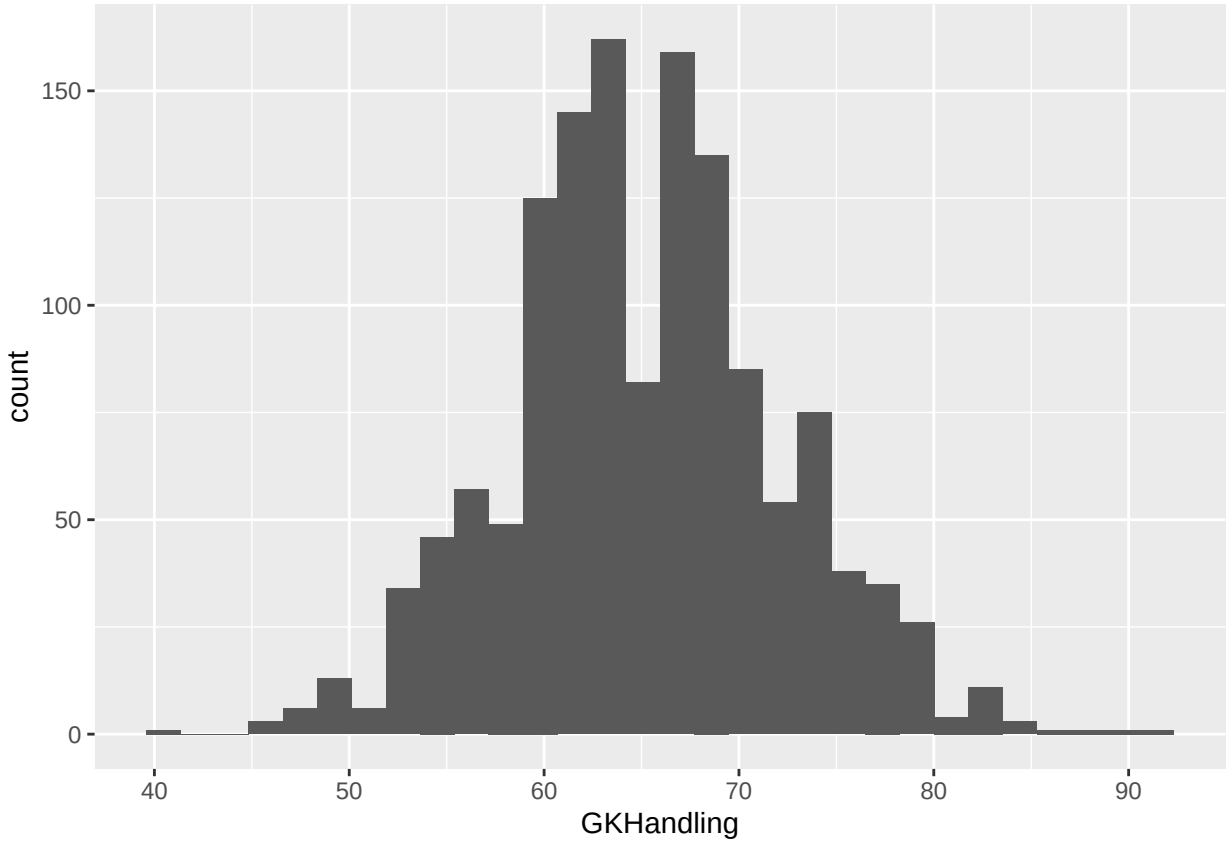
players.

However, zooming in on solely a GK oriented variable, such as GKDiving shows that the distribution of GK only players is normal, as opposed to the prior graph with all players skewed.



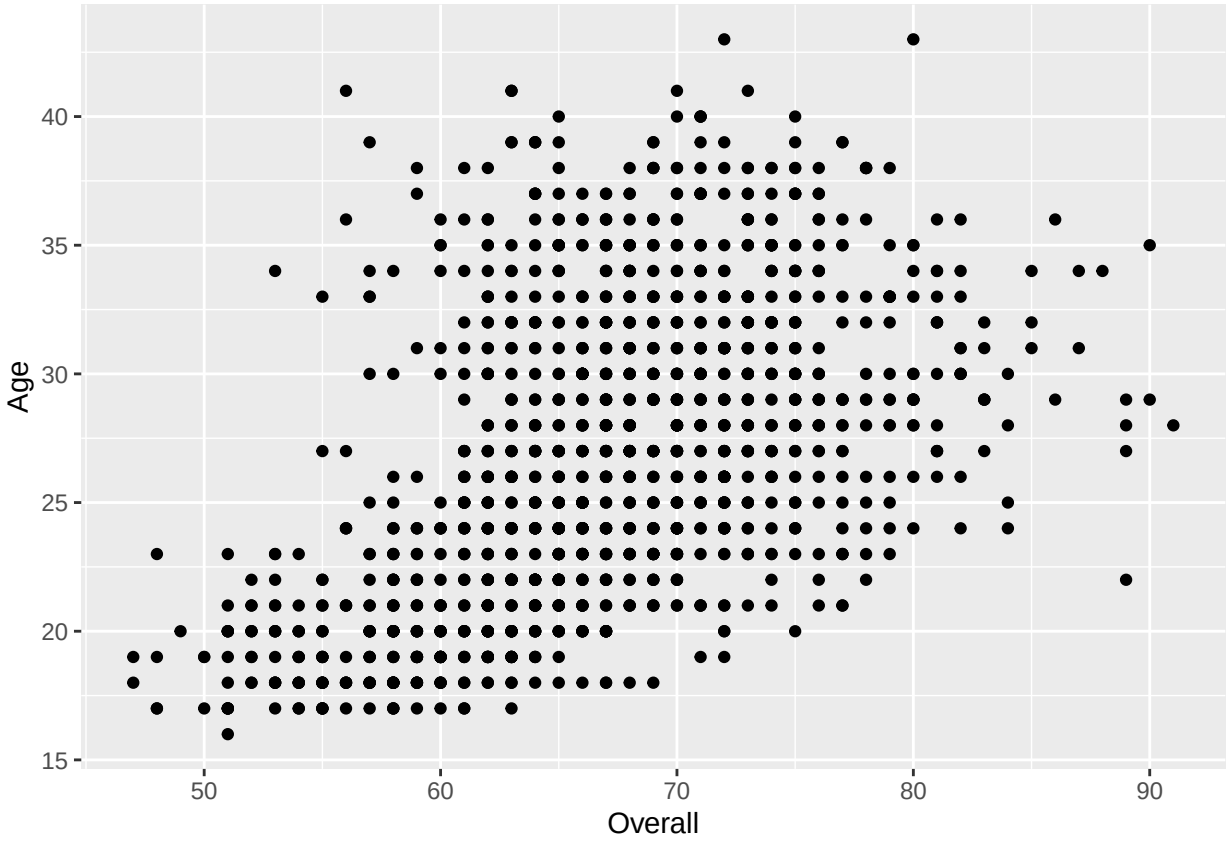
Other skills such as overall skill are surprisingly normally distributed among all professional players. As do the other graphs of GK skills when localized to solely GK role players, such as GKHandling, showing this is repeated in the data.


150
100
50
0
count
50
60
70
80
90
Overall

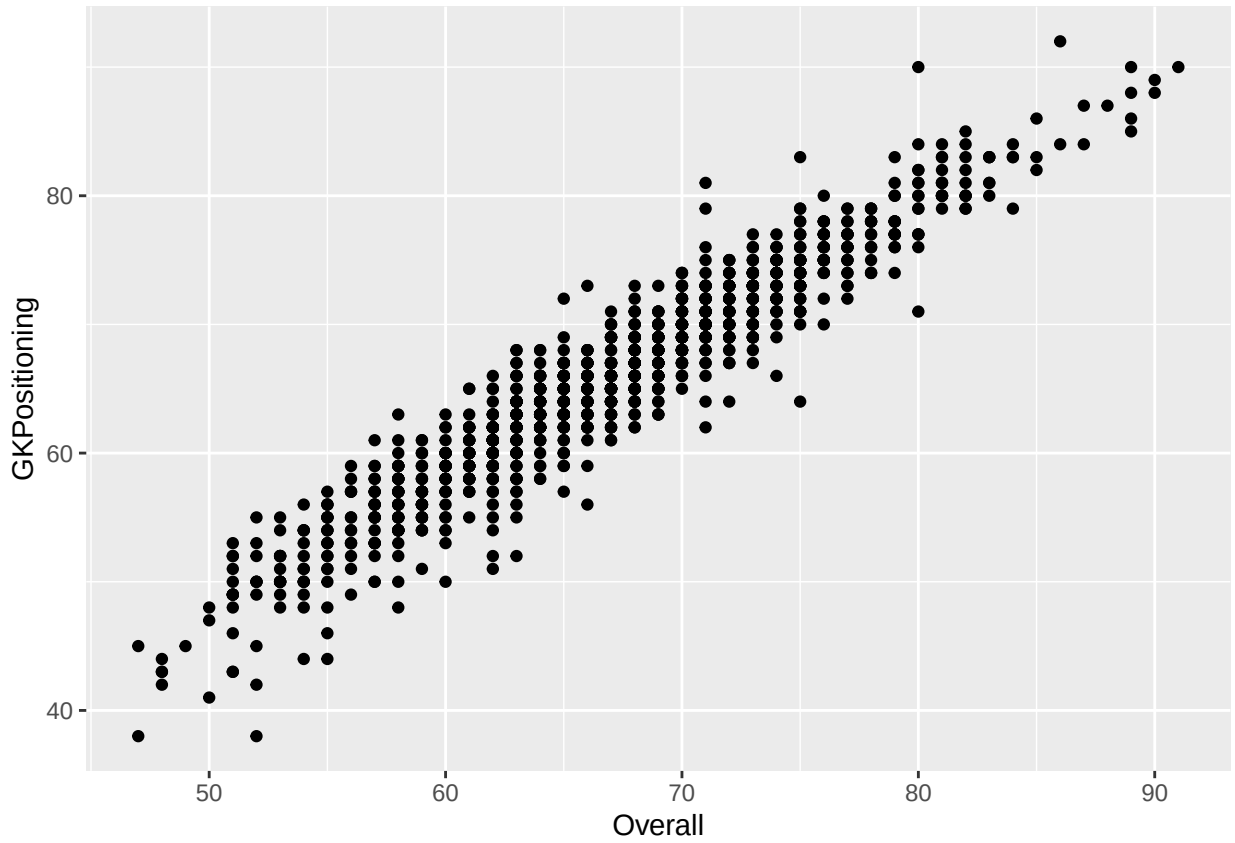


This normal distribution continues for GK players among their various GK stats, thus, this implies that GK are outliers among all players as well. Additionally, attempting to plot correlation, plotting age and overall ability shows a skewed nature towards higher age and more overall ranking. This is expected and seems to still match a normal distribution among the player group. While plotting the GK skill focused variables shows strong correlation, with GKKicking being the least indicative of a highly overall skilled GK while the plotting of GKPositioning shows that variables strongly have a high overall ranked GK. Also, as will be shown in the correlation plot. International reputation is very obviously separated by overall ability

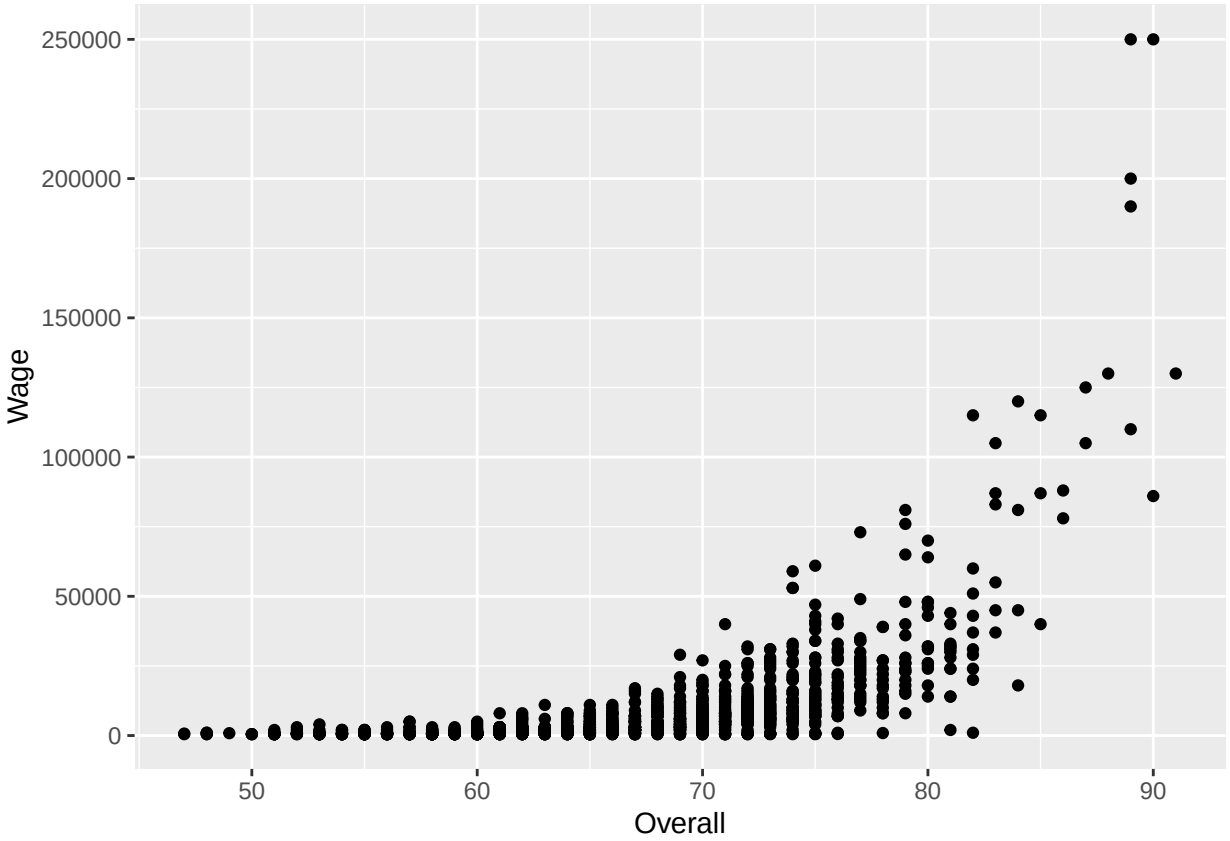
Plotting age and overall ability shows a skewed nature towards higher age and more overall skill. This seems to still match a normal distribution.



The plotting of GK positioning shows that falls the closest to a overall skilled GK, being a great predictor of a GK with the highest overall or perhaps indicating this is the most important GK skill in overall.



Overall skill and wage are quite skewed to the right with a long left tail, showing a large inequality.



Statistical tests were also applied during EDA and model building to better understand how variables may relate to one another. Running these tests on specifically wage, GKpositioning, and overall gave great detail into how these variables worked with one another and indicates that encompassing variables such as "Special" may have omitted variable bias. Which will be focused on in a later section of the paper.

```
##
##  Welch Two Sample t-test
##
## data:  dff$Overall and dff$Wage
## t = -17.073, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9715.550 -7713.036
## sample estimates:
##  mean of x  mean of y
##   67.03976 8781.33284
```

```
##
##  Welch Two Sample t-test
##
## data:  dff$Overall and dff$GKPositioning
## t = 3.9778, df = 2682.4, p-value = 7.138e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5921889 1.7435990
## sample estimates:
```

```
## mean of x mean of y
##  67.03976  65.87187


##
##  Welch Two Sample t-test
##
## data:  dff$GKPositioning and dff$Wage
## t = -17.076, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9716.718 -7714.204
## sample estimates:
##  mean of x  mean of y
##   65.87187 8781.33284


##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dff$Overall and dff$Wage
## W = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Thus, based on the EDA, research question, and measurement goals, the model specifications were narrowed down to as follows. For the first model the specification first was to only include the key variables desired to be measured as to keep covariates at the minimum number. Additionally to this, another specification of the first model was to accurately analyze the variable "Age" due to the potential large impact it could have on the model, thus two age variables were included to check which might impact the model. Thus for the first model, the X variables selected were those most likely to not duplicate or affect one another much from a covariate perspective. Height and weight were used due to omission from being already factored in overall and value. Special was added as well, as this has the ability to factor in the various variables that influence a players ability to become an outlier in a positive way unrelated to overall or potential. This has the ability to capture the omitted variables in a numeric fashion that value does not account for as special is not directly weighted in rank.

The second model adds the specification of including attributes of a player that likely contribute to a player's high value yet not already factored in the overall score or value for the player. Contract Years and International Reputation were selected as these variables meet the specification of improving the model yet also add another specification of seeking variables that should influence a players value yet are likely not currently accounted for due to abstraction from impact on the field. Special was also added and resulted in fantastic values overall and f-tests for the stargazer plot, that could be improved with more specifications or potentially not.

The third model expands on the investigation of success indication by adding all the core skills variables that are not directly used in Overall, these player ability traits and field skills are key as they will have a larger impact on value. Due to this, covariates are more likely, thus these were selected based on the collinearity plot to meet the specification of minimal covariates. These variables do seem to have an indication as to the value of a player based on how they impact the model based on the stargazer. The third specification is to ensure that there is limited crossover in using variables that already account for one another. The model seeks to address this however, the second model does the best in terms of actual results, which will be dived in more in the next section.

The end goal, which is perhaps another specification, is to prevent the collinearity from getting too high and close to one-hundred. This is to ensure that the model's indications can remain plausible instead of "too good to be true" with an abnormally high collinearity. This is done by checking the stargazer values and by reading the collinearity chart to ensure variables do not heavily influence one another.

4. A Results Section

Stargazer regression table:

```
##
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                  log(Value)
##                           (1)       (2)       (3)
## ------------------------------------------------
## Age                     0.306***  0.447***  0.742***
##                          (0.043)   (0.039)   (0.043)
## I(Age2)                 -0.008*** -0.010*** -0.015***
##                          (0.001)   (0.001)   (0.001)
## Height                  0.030***  0.016***    0.012
##                          (0.007)   (0.006)   (0.007)
## Weight                    0.002    0.0003    -0.006
##                          (0.005)   (0.004)   (0.005)
## Special                 0.007***  0.006***
##                         (0.0002)  (0.0002)
## Contract.Years                    -0.193*** -0.263***
##                                    (0.019)   (0.021)
## Agility                                       0.003
##                                              (0.003)
## Strength                                     0.013***
##                                              (0.003)
## Jumping                                      0.014***
##                                              (0.003)
## Acceleration                                 0.009***
##                                              (0.003)
## Stamina                                       0.008*
##                                              (0.004)
## Weak.Foot                                    0.121***
##                                              (0.036)
## International.Reputation           0.847***  1.142***
##                                    (0.053)   (0.059)
## Constant                -2.395**   -0.867    -0.889
##                          (1.149)   (1.023)   (1.279)
## ------------------------------------------------
## Observations              1,358     1,358     1,358
## R2                        0.557     0.655     0.543
## Adjusted R2               0.555     0.653     0.539
## ================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Statistical Significance:

```
## [1] "Model(1) VIF"
```

```
##        Age   I(Age^2)     Height     Weight    Special
## 102.765687  96.436092   1.487106   1.595312   1.749680
```

```
## [1] "Model(2) VIF"
```

```
##                      Age                 I(Age^2)                   Height
##                107.198673                100.997117                 1.517401
##                   Weight                  Special          Contract.Years
##                  1.601764                  2.084273                 1.147207
## International.Reputation
##                  1.285620
```

```
## [1] "Model(3) VIF"
```

```
##                      Age                 I(Age^2)                   Height
##                97.672622                95.391066                 1.693843
##                   Weight International.Reputation          Contract.Years
##                  1.796055                  1.201249                 1.116283
##                  Agility                  Strength                  Jumping
##                  1.760678                  1.561256                 1.736157
##             Acceleration                   Stamina                Weak.Foot
##                  2.044777                  1.513728                 1.021232
```

Our Model(1) includes only key variables based on our research question and preliminary EDA: `Age`, `Height`, `Weight`, and `Special`. Our EDA revealed that `Age` had a polynomial relationship with `log(Value)`, and so our linear model includes both `Age` and `Age^2`. Model(1) has a high VIF for `Age` (102.76) and `Age^2` (96.43). Although high VIFs are typically a concern, it makes sense that `Age` and `Age^2` have collinearity and the model does not aim to differentiate between `Age` and `Age^2`.

```
## model1 msr model2 msr model3 msr
##  0.8229954  0.6410832  0.8491938
```

All of the variables except `Weight` have statistical significance. Our initial Model(1) has an `R^2` of 0.557 and MSR of 0.823. Model(2) contains the key variables in addition to `Contract.Years` and `International.Reputation', which may also influence market value. Model(2) has an`R^2`of 0.655, MSR of 0.641 and stable VIFs. The stargazer regression model shows that`Contract.Years`and`Internationa` are both significant variables in addition to the significant variables in Model(1). When comparing Model(1) and Model(2) through the F-test, Model(2) has a significant p-value less than 2.2e-16 and thus improved the model's fit.

```
anova(model1, model2, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: log(Value) ~ Age + I(Age^2) + Height + Weight + Special
## Model 2: log(Value) ~ Age + I(Age^2) + Height + Weight + Special + Contract.Years +
##     International.Reputation
```

```
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1352 1117.63
## 2   1350  870.59  2    247.04 191.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After creating the linear model for our key variables, Model(3) investigates the impact of the omitted variable bias involved with using `Special` in place of independent performance variables. From the collinearity matrix in our exploratory data analysis, many performance metrics and scoring are collinear. To maintain independence as best as possible, `Agility`, `Strength`, `Jumping`, `Acceleration`, `Stamina`,and `Weak.Foot` were selected due to their weak correlation with each other. The stargazer regression table shows that all the variables except `Agility` and `Weight` are significant. Model(3)'s VIF's confirmed that these variables do not cause multicollinearity problems. Compared to Model(2), Model(3) had a higher MSR of 0.849 and lower `R^2` of 0.492. An F-test comparing Model(2) and Model(3) in the code below did not produce a significant p-value and so Model(3) did not improve Model(2). Although Model(3) may have reduced omitted variable bias, performance measurements are inherently related to each other because it is a measurement of a player's physical ability and can create causality problems. Additionally, our variable selection may not match FIFA's `Special` scoring process and contribute to Model(3)'s inaccuracy.

```
anova(model2, model3, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: log(Value) ~ Age + I(Age^2) + Height + Weight + Special + Contract.Years +
##     International.Reputation
## Model 2: log(Value) ~ Age + I(Age^2) + Height + Weight + International.Reputation +
##     Contract.Years + Agility + Strength + Jumping + Acceleration +
##     Stamina + Weak.Foot
##   Res.Df     RSS Df Sum of Sq F Pr(>F)
## 1   1350  870.59
## 2   1345 1153.21  5   -282.61
```

After evaluating the statistical significance of each model, Model(2) appears to be the best linear regression model among the three because of its higher `R^2`, lower MSR, and variable selection. When applying the t-test, all coefficients are significant except `Weight` and we can reject the null based on the null hypothesis that the p-value must be less than 0.1.

```
##
## t test of coefficients:
##
##                             Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)              -0.86699314  1.13153126  -0.7662  0.443684
## Age                       0.44657651  0.04398013  10.1541 < 2.2e-16 ***
## I(Age^2)                 -0.01036027  0.00078824 -13.1436 < 2.2e-16 ***
## Height                    0.01613469  0.00623590   2.5874  0.009774 **
## Weight                    0.00033100  0.00444658   0.0744  0.940672
## Special                   0.00567491  0.00024287  23.3661 < 2.2e-16 ***
## Contract.Years           -0.19314955  0.02077716  -9.2962 < 2.2e-16 ***
## International.Reputation  0.84687202  0.05826756  14.5342 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical Significance:

$Log(Value) = 0.447 * Age + -0.10 * Age^2 + 0.16 * Height + 0.003 * Weight + 0.006 * Special + 0.193 * Contract.Years + 0.847 * International.Reputation$

The Model(3) linear regression can be interpreted as how Log(Value) will change with increases in each variable. For example, a one-point increase in international reputation while keeping all else constant will lead to 0.847 increase in Log(Value) (or $2.33). The coefficients reveal how much each weight each variable carries in the determination of a player's market value. From this linear regression, international reputation has the largest impact, followed by Age, Contract.years, Height, Special, and Weight. Based on the selected linear model, it is surprising that `Special`, which measures a player's skill, does not play a large role in market value compared to other factors. International reputation and player skill can have a large influence on a club's revenue and performance while weight did not influence market value as much as we hypothesized. As a team manager or scout, this regression can support the determination of whether it is worth recruiting a high-market value player based on the team's priorities. A highly skilled player with low international reputation may have a lower market value compared to a player with high international reputation but is not as skilled. A team looking for a highly skilled player may find that recruiting the player with the lower market value is more beneficial.

5. Limitations of your Model

5a. Statistical limitations of your model

We know the data collected includes information on soccer players from the FIFA 2022 series and does not include any form of formal sampling, which is a potential violation. Rather each point in the data set is a unique player from the FIFA series, from a pool of all players in the league. With this knowledge, we can assume that the data is IID. This assumption of IID can help mitigate the issues caused from not having formal sampling.

In terms of a unique BLP, we run a test to ensure both that there is no perfect collinearity, and that E[XTX] is invertible. We know from previous work that there exists no perfect collinearity within the data that has been selected for our 3 models. We run the inversion of E[XTX] to find that there indeed exists a matrix that satisfies this condition. With that being said, we can say that the Large Sample Assumptions have been satisfied for our model.

Initially, we had planned to leverage some model variables that did not apply well to non-IID data, as we noticed some large non-log skewed histograms and plots of the player variables. However, when checking the large sample assumption violations and needing to mitigate them, we landed on ensuring that we used model variables that would fit to IID data well.

```
##                                Value           Age International.Reputation
## Value                   1.936189e-17  8.966355e-12            -1.572446e-10
## Age                     8.966355e-12  4.522965e-05            -9.757234e-05
## International.Reputation -1.572446e-10 -9.757234e-05             5.405479e-03
## Contract.Years           2.042353e-11 -4.026555e-05            -8.134411e-05
## Height                   2.466207e-12  3.951699e-06            -7.173825e-07
## Weight                  -1.117534e-12 -7.548542e-06            -1.348264e-05
## Special                 -5.678725e-13 -1.463929e-06            -1.948137e-06
## Weak.Foot               -5.408108e-13  2.316212e-05            -2.791212e-05
## Agility                 -1.526464e-13 -1.008096e-06             1.088026e-06
## Strength                -2.548573e-13  1.652363e-07             5.427484e-06
## Jumping                  1.369207e-12  1.031512e-06            -1.455716e-05
## Acceleration             1.786416e-12  4.589359e-06             8.835660e-06
## Stamina                  6.430750e-13  1.609595e-06             1.192106e-05
```

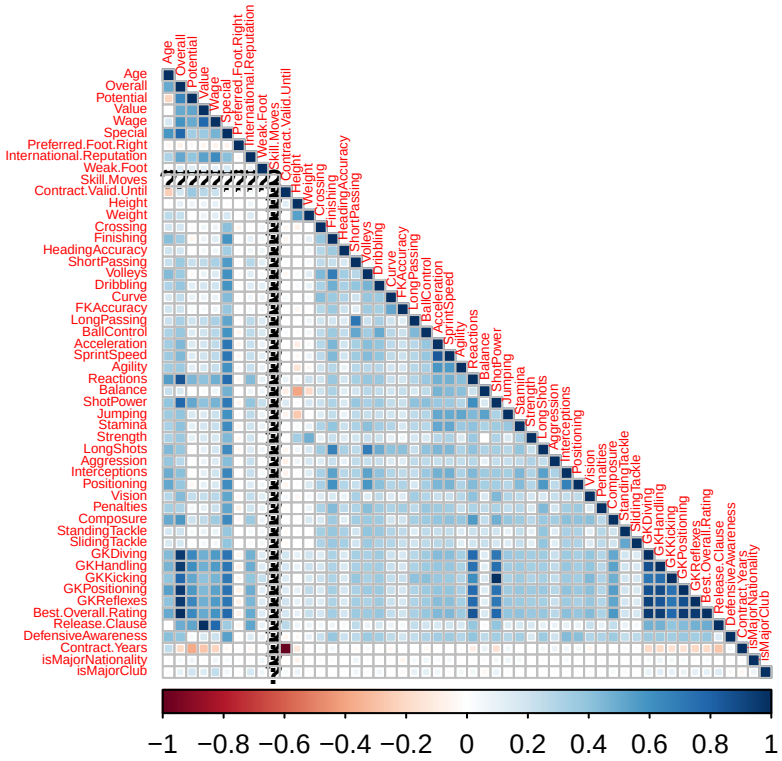
```
##                            Contract.Years        Height        Weight
## Value                       2.042353e-11  2.466207e-12 -1.117534e-12
## Age                        -4.026555e-05  3.951699e-06 -7.548542e-06
## International.Reputation   -8.134411e-05 -7.173825e-07 -1.348264e-05
## Contract.Years              5.768788e-04  4.410155e-06 -3.079784e-06
## Height                      4.410155e-06  6.341614e-06 -1.171131e-05
## Weight                     -3.079784e-06 -1.171131e-05  3.165123e-05
## Special                     1.465920e-06 -3.826358e-07  9.712513e-08
## Weak.Foot                  -1.950185e-05 -1.319097e-05  1.812295e-06
## Agility                    -1.923391e-06 -4.854470e-07  2.030005e-06
## Strength                   -8.117177e-07  1.038322e-06 -6.484298e-06
## Jumping                    -5.745417e-06 -1.714009e-07  4.357575e-08
## Acceleration               -2.748215e-06  1.795620e-06 -5.071582e-08
## Stamina                    -1.382219e-06  3.352939e-07  8.166150e-07
##                                 Special     Weak.Foot       Agility
## Value                     -5.678725e-13 -5.408108e-13 -1.526464e-13
## Age                       -1.463929e-06  2.316212e-05 -1.008096e-06
## International.Reputation  -1.948137e-06 -2.791212e-05  1.088026e-06
## Contract.Years             1.465920e-06 -1.950185e-05 -1.923391e-06
## Height                    -3.826358e-07 -1.319097e-05 -4.854470e-07
## Weight                     9.712513e-08  1.812295e-06  2.030005e-06
## Special                    2.062228e-07 -2.912201e-06 -2.766056e-07
## Weak.Foot                 -2.912201e-06  1.580790e-03  2.907927e-06
## Agility                   -2.766056e-07  2.907927e-06  1.009712e-05
## Strength                  -3.173245e-07  6.156373e-06  4.830901e-07
## Jumping                   -4.482648e-07 -1.094476e-06 -1.895367e-06
## Acceleration              -7.903201e-07  8.092737e-06 -2.769290e-06
## Stamina                   -6.929067e-07  4.002910e-06  5.161993e-07
##                                Strength       Jumping  Acceleration
## Value                     -2.548573e-13  1.369207e-12  1.786416e-12
## Age                        1.652363e-07  1.031512e-06  4.589359e-06
## International.Reputation   5.427484e-06 -1.455716e-05  8.835660e-06
## Contract.Years            -8.117177e-07 -5.745417e-06 -2.748215e-06
## Height                     1.038322e-06 -1.714009e-07  1.795620e-06
## Weight                    -6.484298e-06  4.357575e-08 -5.071582e-08
## Special                   -3.173245e-07 -4.482648e-07 -7.903201e-07
## Weak.Foot                  6.156373e-06 -1.094476e-06  8.092737e-06
## Agility                    4.830901e-07 -1.895367e-06 -2.769290e-06
## Strength                   1.007810e-05  9.240762e-07 -2.046880e-09
## Jumping                    9.240762e-07  9.173139e-06 -5.254134e-07
## Acceleration              -2.046880e-09 -5.254134e-07  1.464574e-05
## Stamina                   -1.153295e-06  1.680742e-07 -1.770330e-06
##                                Stamina
## Value                      6.430750e-13
## Age                        1.609595e-06
## International.Reputation   1.192106e-05
## Contract.Years            -1.382219e-06
## Height                     3.352939e-07
## Weight                     8.166150e-07
## Special                   -6.929067e-07
## Weak.Foot                  4.002910e-06
## Agility                    5.161993e-07
## Strength                  -1.153295e-06
## Jumping                    1.680742e-07
```

```
## Acceleration              -1.770330e-06
## Stamina                    2.168487e-05
```

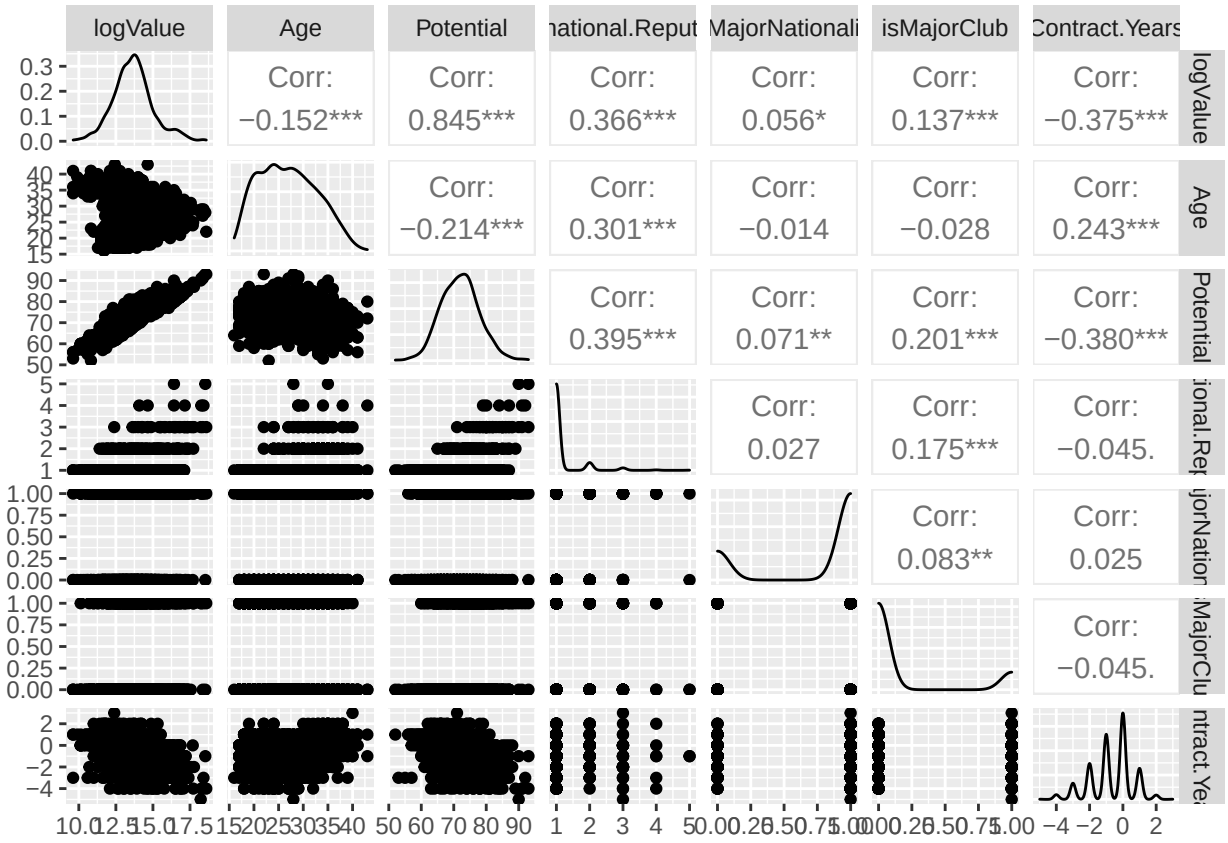
5b. Structural limitations of your model

The differences between model one and model two are essentially confined to expanding the variables to ensure that the assumed base assumptions are not skewing the model too much from being a reliable indicator. Thus, investigating model two and model three, it becomes apparent that the bias of "Special" in model two is being addressed in model three. "Special" is the most important omitted variable with another omitted variable that's not very biased being "Contract Years". These variables were unable to be analyzed as much as desired during EDA due to not knowing how they would influence the models until applied. The bias caused by "Contract Years" is quite minimal, having very little impact on the model and showing that it likely is normal. While for "Special" the bias is large, shown by a very skewed f-stat and a considerably lower residual error. Nonetheless, while the residual error is below the bounds that would likely cause the data to come into question (.90 and above) the influence of "Special" does not invalidate the results, but does call for further scrutiny. Hence, in model three "Special" is substituted by player abilities, the bias increase becomes apparent in the f-stat, constant, and individual dependent log values. This raises the residual error to near 1, the other tests and variables show that this is reducing omitted variable bias and resulting in a far more biased model.

Another way to visualize this reduction in omitted variable bias is by looking at the collinearity plot and seeing the ways the various variables interact. The variables replacing "Special" in model three have the most minimal collinearity compared to other variables with those in model three as shown in the collinearity plot. However, not the most optimal model like in model two.




```
## [1] 0.7533233
```



7. Conclusion

The above analysis focused on introducing the focus area of a "product" to improve, being clubs makeup and the criteria by which clubs should assess a skilled GK and determine the value of such GK. The FIFA dataset was used due to its reliability and accuracy for assessing players and giving a complex sport soccer quantitative and qualitative values for difficult to assess skills and field actions. The EDA process was especially important in which the data and potential model implications were explored, then the goal setting and model specifications were outlined. Finally, the models were built using the EDA process with variables cleaned or created to meet the research question and measurement goals outlined. The various potential biases and statistical limitations were also addressed after applying the models to the data via stargazer and checking various values such as collinearity, residual, f-stat, log values and deviations.

Using our analysis and addressing the research question with our models, we were able to assess the relationship between the typical player in FIFA at the goal-keeper position's market value and their physical attributes, performance, and background. We made use of our second model specifically such that it explains about 65.3% of the variance in the data. The statistically significant high-level findings from this assessment are as follows (while holding all other factors constant):

Citing below log transformations found here: https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

```
##          (Intercept)                  Age             I(Age^2)
##        -57.97868263          56.29522601          -1.03067869
##               Height               Weight                Special
##          1.62655595           0.03310543           0.56910410
```

```
##           Contract.Years International.Reputation
##             -17.56413178             133.23399142
```

- As the Age of a Goalkeeper is increased by 1 year, their market value is increased by 55.26%
- As the Height of a Goalkeeper is increased by 1 kilogram, their market value is increased by 1.63%
- As the "Special" ranking of a Goalkeeper is increased by 1 rating, their market value is increased by 0.56%
- As a Goalkeeper's contract years increase by 1 year, their market value decreases by 17.56%
- As a Goalkeeper's international reputation is increased by 1 rating, their market value increases by 133.23%

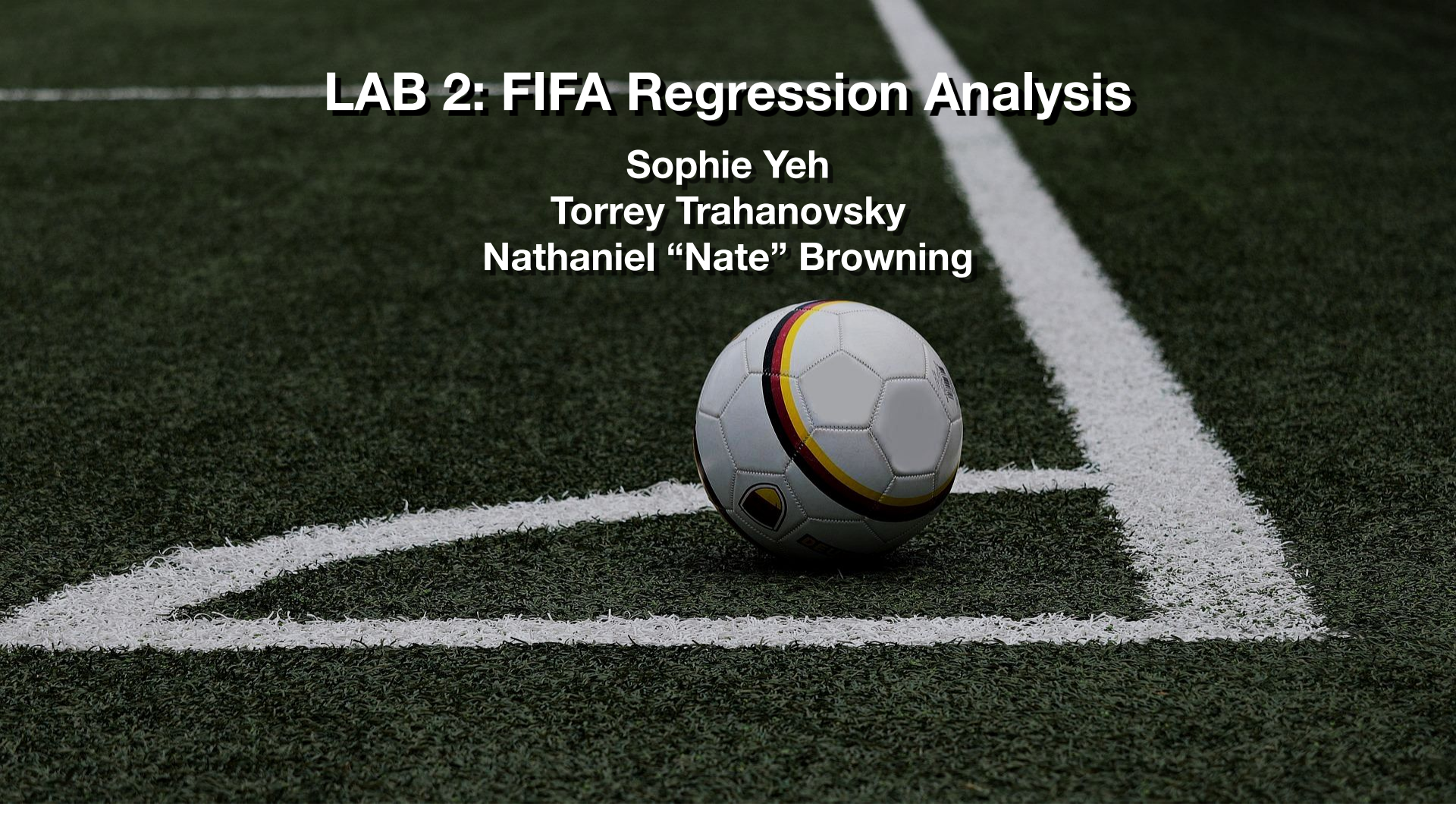
Thus regarding the research question:

Research Question: In 2022, how does a goal keeper's performance, background, and physical attributes affect their market value in football?

It becomes clear how Height, Special, Contract Years, and International reputation potentially influence market value. These variables should thus be used by Club's to improve their "product" of players in their Club and gain better value players. However, it could also be possible that these variables might not influence the value so much as instead be the reason behind value and already accounted for. Nonetheless, the insights provide enough in the models for Clubs to use the models on players to see who would be the best GK for their club and ensure they are getting the highest value or potential value players.

LAB 2: FIFA Regression Analysis

Sophie Yeh
Torrey Trahanovsky
Nathaniel “Nate” Browning



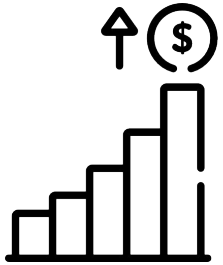
Intro

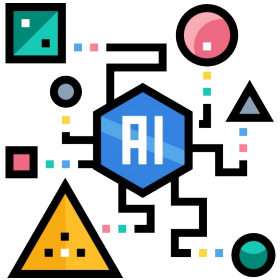
1

2

3

Research Question:
In 2022, how does a goal keeper's performance, background, and physical attributes affect their market value in football?





International Federation of Association Football (FIFA)

Research Question Created with measurement goals, model specs, and EDA deciding models

Use model to optimize "product" and provide tooling to FIFA

Research Question Addressing Process

1. Measurement Goals
2. Testing the research question via a null-hypothesis and various tests/stargazers
3. Design:
    a. **Primary design:** Causal and explanatory:
        i. Attempting to analyze collinearity between various x variables and see how they might indicate whether the y variable or other variables are affected by variances.
    b. **Secondary Design:** Exploratory:
        i. In which during exploratory data analysis, various graphs, tables, stargazer, tests, and other visualization techniques are used to understand the data.
4. Minimal bias models that increase in efficiency as developed to a final model.

Data

- **Limited Scope:** Goalkeepers
  - Nans are eliminated, data is cleaned, ready to EDA.

- **Related:** Outcome, Potential, Value, Skills
  - Outcome is made up of various other dataset values
  - Skills encompass other attributes
  - Potential is closely related to outcomes underlying values
  - Value already accounts for some pre-existing data

- **Peculiar yet Useful:** Special, Wage, Contract Years, Reputation
  - Special is a string of #’s yet indicates various attributes
  - Wage is monthly making conclusions on pay clearer
  - Contract_Years is custom and calculates years remaining
  - Reputation is noteworthy and potentially needs to contrast with nationality

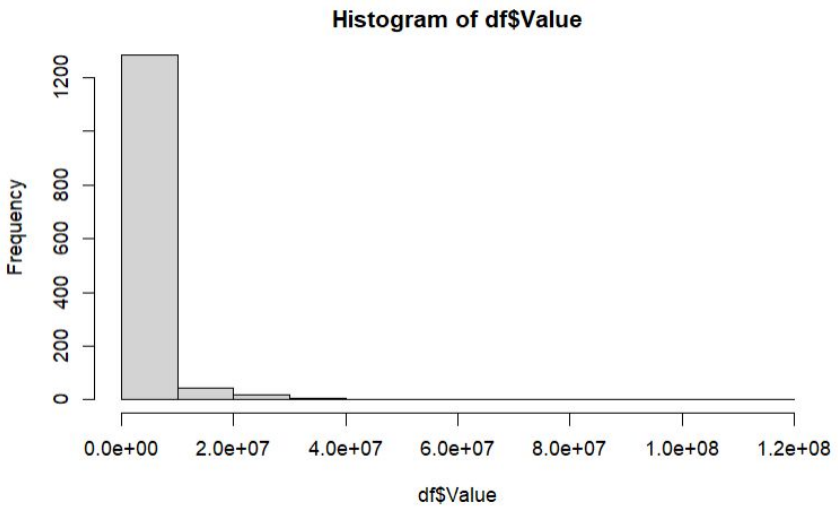
Key Covariates

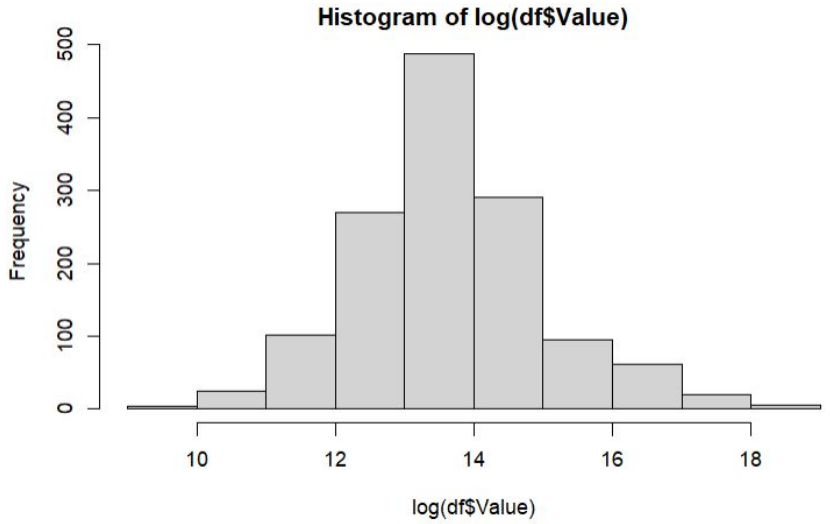
The dataset provided includes all players included in the FIFA 2022 video game including data on their physical description, position, and ratings of skill. The following covariates were considered key to our study and are included below:

- **Age:** The age of the player in the respective row
- **Height:** The height of the player in centimeters
- **Weight:** The weight of the player in kilograms
- **Special:** A rating metric to measure skills of a player
- **Value:** The market value for trading of a player measured in Euros

Covariate Transformations

- Age covariate has been made a quadratic to avoid a non-linear correlation to the success measure
- Value covariate has been logged to adjust for normality.





Large Sample Assumptions

- Two Large Sample Assumptions:
  - I.I.D. Data
  - A unique BLP exists

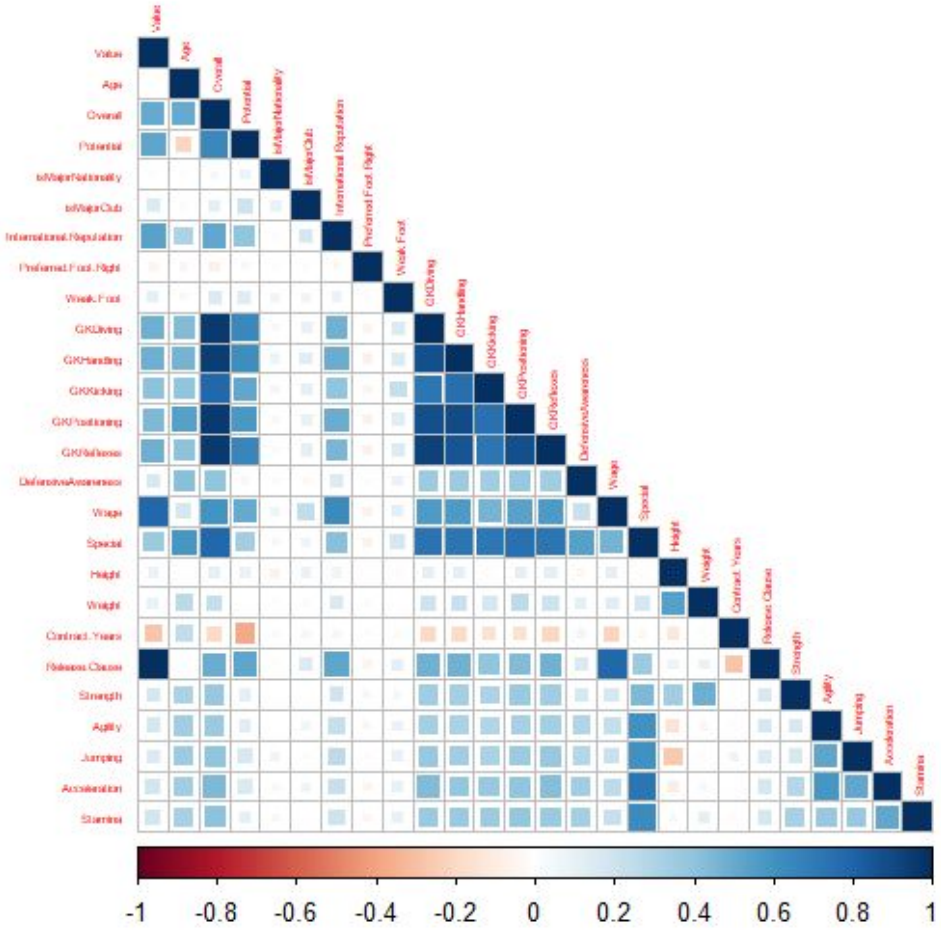
	Value	Age	International.Reputation	Contract.Years	Height	Weight
Value	1.936189e-17	8.966355e-12	-1.572446e-10	2.042353e-11	2.466207e-12	-1.117534e-12
Age	8.966355e-12	4.522965e-05	-9.757234e-05	-4.026555e-05	3.951699e-06	-7.548542e-06
International.Reputation	-1.572446e-10	-9.757234e-05	5.405479e-03	-8.134411e-05	-7.173825e-07	-1.348264e-05
Contract.Years	2.042353e-11	-4.026555e-05	-8.134411e-05	5.768788e-04	4.410155e-06	-3.079784e-06
Height	2.466207e-12	3.951699e-06	-7.173825e-07	4.410155e-06	6.341614e-06	-1.171131e-05
Weight	-1.117534e-12	-7.548542e-06	-1.348264e-05	-3.079784e-06	-1.171131e-05	3.165123e-05
Special	-5.678725e-13	-1.463929e-06	-1.948137e-06	1.465920e-06	-3.826358e-07	9.712513e-08
Weak.Foot	-5.408108e-13	2.316212e-05	-2.791212e-05	-1.950185e-05	-1.319097e-05	1.812295e-06
Agility	-1.526464e-13	-1.008096e-06	1.088026e-06	-1.923391e-06	-4.854470e-07	2.030005e-06
Strength	-2.548573e-13	1.652363e-07	5.427484e-06	-8.117177e-07	1.038322e-06	-6.484298e-06
Jumping	1.369207e-12	1.031512e-06	-1.455716e-05	-5.745417e-06	-1.714009e-07	4.357575e-08
Acceleration	1.786416e-12	4.589359e-06	8.835660e-06	-2.748215e-06	1.795620e-06	-5.071582e-08
Stamina	6.430750e-13	1.609595e-06	1.192106e-05	-1.382219e-06	3.352939e-07	8.166150e-07

	Special	Weak.Foot	Agility	Strength	Jumping	Acceleration	Stamina
Value	-5.678725e-13	-5.408108e-13	-1.526464e-13	-2.548573e-13	1.369207e-12	1.786416e-12	6.430750e-13
Age	-1.463929e-06	2.316212e-05	-1.008096e-06	1.652363e-07	1.031512e-06	4.589359e-06	1.609595e-06
International.Reputation	-1.948137e-06	-2.791212e-05	1.088026e-06	5.427484e-06	-1.455716e-05	8.835660e-06	1.192106e-05
Contract.Years	1.465920e-06	-1.950185e-05	-1.923391e-06	-8.117177e-07	-5.745417e-06	-2.748215e-06	-1.382219e-06
Height	-3.826358e-07	-1.319097e-05	-4.854470e-07	1.038322e-06	-1.714009e-07	1.795620e-06	3.352939e-07
Weight	9.712513e-08	1.812295e-06	2.030005e-06	-6.484298e-06	4.357575e-08	-5.071582e-08	8.166150e-07
Special	2.062228e-07	-2.912201e-06	-2.766056e-07	-3.173245e-07	-4.482648e-07	-7.903201e-07	-6.929067e-07
Weak.Foot	-2.912201e-06	1.580790e-03	2.907927e-06	6.156373e-06	-1.094476e-06	8.092737e-06	4.002910e-06
Agility	-2.766056e-07	2.907927e-06	1.009712e-05	4.830901e-07	-1.895367e-06	-2.769290e-06	5.161993e-07
Strength	-3.173245e-07	6.156373e-06	4.830901e-07	1.007810e-05	9.240762e-07	-2.046880e-09	-1.153295e-06
Jumping	-4.482648e-07	-1.094476e-06	-1.895367e-06	9.240762e-07	9.173139e-06	-5.254134e-07	1.680742e-07
Acceleration	-7.903201e-07	8.092737e-06	-2.769290e-06	-2.046880e-09	-5.254134e-07	1.464574e-05	-1.770330e-06
Stamina	-6.929067e-07	4.002910e-06	5.161993e-07	-1.153295e-06	1.680742e-07	-1.770330e-06	2.168487e-05

- IID Data exists with each player being a unique individual
- We are able to find the inverse of $E[X^TX]$
- No perfect collinearity exists within the data being used
- We can assume the data meets the large sample assumptions

Correlation Results

- Goalkeeper skills highly correlated with one another
- Player skills all highly correlated to the “Special” covariate
- “Special” VIF is close to 2, not higher than 4 so we will use in model



Linear Regression Model 1: Base Model

	Dependent variable:		
	log(Value)		
	(1)	(2)	(3)
Age	0.306***	0.447***	0.742***
	(0.043)	(0.039)	(0.043)
I(Age2)	-0.008***	-0.010***	-0.015***
	(0.001)	(0.001)	(0.001)
Height	0.030***	0.016***	0.012
	(0.007)	(0.006)	(0.007)
Weight	0.002	0.0003	-0.006
	(0.005)	(0.004)	(0.005)
Special	0.007***	0.006***	
	(0.0002)	(0.0002)	
Contract.Years		-0.193***	-0.263***
		(0.019)	(0.021)
Agility			0.003
			(0.003)
Strength			0.013***
			(0.003)
Jumping			0.014***
			(0.003)
Acceleration			0.009***
			(0.003)
Stamina			0.008*
			(0.004)
Weak.Foot			0.121***
			(0.036)
International.Reputation		0.847***	1.142***
		(0.053)	(0.059)
Constant	-2.395**	-0.867	-0.889
	(1.149)	(1.023)	(1.279)
Observations	1,358	1,358	1,358
R2	0.557	0.655	0.543
Adjusted R2	0.555	0.653	0.539
Residual Std. Error	0.909 (df = 1352)	0.803 (df = 1350)	0.926 (df = 1345)
F Statistic	339.813*** (df = 5; 1352)	465.862*** (df = 7; 1350)	133.052*** (df = 12; 1345)

model1 msr
0.8229954

Linear Regression Model 2

Anova Test:
Model 2 has a
p-value < 2.2e-16.

Differences between
Model 1 and 2 are
significant.

	Dependent variable:		
	log(Value)		
	(1)	(2)	(3)
Age	0.306***	0.447***	0.742***
	(0.043)	(0.039)	(0.043)
I(Age2)	-0.008***	-0.010***	-0.015***
	(0.001)	(0.001)	(0.001)
Height	0.030***	0.016***	0.012
	(0.007)	(0.006)	(0.007)
Weight	0.002	0.0003	-0.006
	(0.005)	(0.004)	(0.005)
Special	0.007***	0.006***	
	(0.0002)	(0.0002)	
Contract.Years		-0.193***	-0.263***
		(0.019)	(0.021)
Agility			0.003
			(0.003)
Strength			0.013***
			(0.003)
Jumping			0.014***
			(0.003)
Acceleration			0.009***
			(0.003)
Stamina			0.008*
			(0.004)
Weak.Foot			0.121***
			(0.036)
International.Reputation		0.847***	1.142***
		(0.053)	(0.059)
Constant	-2.395**	-0.867	-0.889
	(1.149)	(1.023)	(1.279)
Observations	1,358	1,358	1,358
R2	0.557	0.655	0.543
Adjusted R2	0.555	0.653	0.539
Residual Std. Error	0.909 (df = 1352)	0.803 (df = 1350)	0.926 (df = 1345)
F Statistic	339.813*** (df = 5; 1352)	365.862*** (df = 7; 1350)	133.0*** (df = 12; 1345)

model2 msr
0.6410832

Linear Regression Model 3

	Dependent variable		
	log(Value)		
	(1)	(2)	(3)
Age	0.306***	0.447***	0.742***
	(0.043)	(0.039)	(0.043)
I(Age2)	-0.008***	-0.010***	-0.015***
	(0.001)	(0.001)	(0.001)
Height	0.030***	0.016***	0.012
	(0.007)	(0.006)	(0.007)
Weight	0.002	0.0003	-0.006
	(0.005)	(0.004)	(0.005)
Special	0.007***	0.006***	
	(0.0002)	(0.0002)	
Contract.Years		-0.193***	-0.263***
		(0.019)	(0.021)
Agility			0.003
			(0.003)
Strength			0.013***
			(0.003)
Jumping			0.014***
			(0.003)
Acceleration			0.009***
			(0.003)
Stamina			0.008*
			(0.004)
Weak.Foot			0.121***
			(0.036)
International.Reputation		0.847***	1.142***
		(0.053)	(0.059)
Constant	-2.395**	-0.867	-0.889
	(1.149)	(1.023)	(1.279)
Observations	1,358	1,358	1,358
R2	0.557	0.655	0.543
Adjusted R2	0.555	0.653	0.539
Residual Std. Error	0.909 (df = 1352)	0.803 (df = 1350)	0.926 (df = 1345)
F Statistic	339.813*** (df = 5; 1352)	365.862*** (df = 7; 1350)	133.052*** (df = 12; 1345)

model3 msr
0.8491938

Final Linear Regression Model: Model 2

- Lowest MSR
- Highest R^2
- Added variables improved model

	Dependent variable:
	log(Value)
Age	0.447***
	(0.039)
I(Age2)	-0.010***
	(0.001)
Height	0.016***
	(0.006)
Weight	0.0003
	(0.004)
Special	0.006***
	(0.0002)
Contract.Years	-0.193***
	(0.019)
International.Reputation	0.847***
	(0.053)
Constant	-0.867
	(1.023)
Observations	1,358
R2	0.655
Adjusted R2	0.653
Residual Std. Error	0.803 (df = 1350)
F Statistic	365.862*** (df = 7; 1350)

Conclusion

Key Findings from Model:

- As the **Age** of a Goalkeeper is increased by 1 year, their market value is **increased** by **55.4%**
- As the **Height** of a Goalkeeper is increased by 1 kilogram, their market value is **increased** by **1.61%**
- As the "**Special**" ranking of a Goalkeeper is increased by 1 rating, their market value is **increased** by **0.57%**
- As a Goalkeeper's **contract years** increase by 1 year, their market value **decreases** by **17.56%**
- As a Goalkeeper's **international reputation** is increased by 1 rating, their market value **increases** by **133.23%**

Using the above, we are able to gain insight on how soccer clubs can maximize the market value that their Goalkeeper holds. There is reason to consider these values and how they might directly impact a team's defense on the market side as time passes. Regardless, the information is beneficial for both clubs and goalkeepers to ensure everyone is able to predict how certain factors will affect their market value.

Q&A