BUS 212a – Analyzing Big Data II, Spring 2019
Project

Assignment 2: Multiple Regression

Team:

Data Incubator

Part 0. Data Wrangling process for the newly chosen FIFA dataset

1. Dataset includes latest edition FIFA 2019 players attributes like Age, Nationality, Overall, Potential, Club, Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Height, Weight, score for each position, Crossing, Finishing, Heading, Accuracy, and so on.

2. Problems solved during data wrangling: ununified units for currency-related columns; missing values; mismatched data in need of format converting; detecting zero value in a variable, redundant variable deleting, etc.

Part 1. Descriptive Statistics

1. Show descriptive statistics for relevant and important variables.

- **the minimum, maximum, and average (mean, median, mode) and standard deviation / variance of important variables.**

```
##                          vars     n       mean          sd  min       max
## Age                         1 14743      25.11        4.59   16        39
## Overall                     2 14743      66.38        6.89   46        94
## Potential                   3 14743      71.33        6.10   48        95
## Wage                        4 14743    9990.64    22834.38 1000    565000
## International.Reputation    5 14743       1.12        0.40    1         5
## Release.Clause              6 14743 4554829.34 11258078.99 13000 228000000
## Nationality                 7 14743        NaN          NA  Inf      -Inf
## Club                        8 14743        NaN          NA  Inf      -Inf
## Work.Rate                   9 14743        NaN          NA  Inf      -Inf
## Position                   10 14743        NaN          NA  Inf      -Inf
##                              range       se
## Age                             23     0.04
## Overall                         48     0.06
## Potential                       47     0.05
## Wage                        564000   188.06
## International.Reputation         4     0.00
## Release.Clause           227987000 92719.56
## Nationality                   -Inf       NA
## Club                          -Inf       NA
## Work.Rate                     -Inf       NA
## Position                      -Inf       NA
```

```r
#Descriptive statistics for the top 2-4 clubs on FIFA 2019 ranking
ClubStat["Real Madrid"]
##                          vars  n        mean          sd   median
## Age                         1 29       23.62        4.38       22
## Overall                     2 29       78.10        9.78       80
## Potential                   3 29       85.03        5.29       86
## Wage                        4 29   154068.97   130716.41   120000
## International.Reputation    5 29        2.03        1.27        1
## Release.Clause              6 29 57034482.76 51946045.33 46000000
## Nationality*                7 29      107.41       50.45      138
## Club*                       8 29      474.00        0.00      474
## Work.Rate*                  9 29        5.24        3.08        3
## Position*                  10 29       13.41        8.81       14
##                              trimmed         mad     min       max
## Age                            23.44        4.45      17        32
## Overall                        78.24       11.86      63        91
## Potential                      85.24        7.41      75        92
## Wage                       146000.00   142329.60    9000    420000
## International.Reputation        1.96        0.00       1         4
## Release.Clause           53560000.00 65234400.00 1000000 156000000
## Nationality*                  110.96        0.00       7       158
## Club*                         474.00        0.00     474       474
## Work.Rate*                      5.28        2.97       1         9
## Position*                      13.32       13.34       2        26
##                              range  skew kurtosis         se
## Age                             15  0.40    -1.14       0.81
## Overall                         28 -0.34    -1.49       1.82
## Potential                       17 -0.36    -1.26       0.98
## Wage                        411000  0.48    -1.16   24273.43
## International.Reputation         3  0.55    -1.50       0.24
## Release.Clause           155000000  0.49    -1.16 9646138.45
## Nationality*                   151 -0.87    -1.08       9.37
## Club*                            0   NaN      NaN       0.00
## Work.Rate*                       8  0.15    -1.75       0.57
## Position*                       24  0.13    -1.51       1.64

ClubStat["FC Bayern MÃ¼nchen"]

##                          vars  n        mean          sd     median
## Age                         1 24       24.21        5.44       23.0
## Overall                     2 24       77.29       10.60       83.0
## Potential                   3 24       83.67        5.14       84.5
## Wage                        4 24    73958.33    55949.88    85000.0
## International.Reputation    5 24        2.50        1.22        2.5
## Release.Clause              6 24 39306416.67 34691843.33 41000000.0
## Nationality*                7 24       69.17       32.15       60.0
## Club*                       8 24      219.00        0.00      219.0
## Work.Rate*                  9 24        6.12        3.44        9.0
```

```
## Position*                   10 24        9.75        7.84        7.0
##                                trimmed         mad    min       max
## Age                              23.90        6.67     17        35
## Overall                          77.85        4.45     59        90
## Potential                        84.05        5.19     72        90
## Wage                          70100.00    56338.80   3000    205000
## International.Reputation          2.50        2.22      1         4
## Release.Clause             37000000.00 45219300.00 660000 127000000
## Nationality*                     67.65        0.00     10       138
## Club*                           219.00        0.00    219       219
## Work.Rate*                        6.35        0.00      1         9
## Position*                         9.00        7.41      1        26
##                                range  skew kurtosis         se
## Age                               18  0.39    -1.12       1.11
## Overall                           31 -0.66    -1.37       2.16
## Potential                         18 -0.81    -0.57       1.05
## Wage                          202000  0.27    -0.70   11420.72
## International.Reputation           3  0.00    -1.64       0.25
## Release.Clause             126340000  0.45    -0.62 7081442.87
## Nationality*                     128  0.80     0.07       6.56
## Club*                              0   NaN      NaN       0.00
## Work.Rate*                         8 -0.39    -1.81       0.70
## Position*                         25  0.74    -0.75       1.60
```

```
ClubStat["FC Barcelona"]
##                            vars  n        mean          sd   median
## Age                           1 29       23.86        4.81       23
## Overall                       2 29       78.62        9.02       82
## Potential                     3 29       85.72        4.38       86
## Wage                          4 29   153379.31   137559.96   125000
## International.Reputation      5 29        2.28        1.33        2
## Release.Clause                6 29 56448275.86 56271399.87 53000000
## Nationality*                  7 29       97.00       54.09      138
## Club*                         8 29      217.00        0.00      217
## Work.Rate*                    9 29        5.76        3.28        7
## Position*                    10 29       11.55        9.05        8
##                                trimmed         mad     min       max
## Age                              23.72        5.93      18        32
## Overall                          78.64        7.41      64        94
## Potential                        85.80        4.45      77        94
## Wage                         136240.00   149742.60   11000    565000
## International.Reputation          2.16        1.48       1         5
## Release.Clause             49720000.00 63751800.00 2000000 226000000
## Nationality*                     99.96        0.00       7       155
## Club*                           217.00        0.00     217       217
## Work.Rate*                        5.88        2.97       1         9
## Position*                        11.24        8.90       1        26
##                                range  skew kurtosis          se
## Age                               14  0.35    -1.39        0.89
## Overall                           30 -0.24    -1.34        1.67
```

```
## Potential                         17 -0.17    -0.91        0.81
## Wage                          554000  1.17     1.06    25544.24
## International.Reputation           4  0.56    -1.05        0.25
## Release.Clause             224000000  1.13     0.89 10449336.63
## Nationality*                     148 -0.55    -1.61       10.04
## Club*                              0   NaN      NaN        0.00
## Work.Rate*                         8 -0.22    -1.79        0.61
## Position*                         25  0.30    -1.60        1.68
```

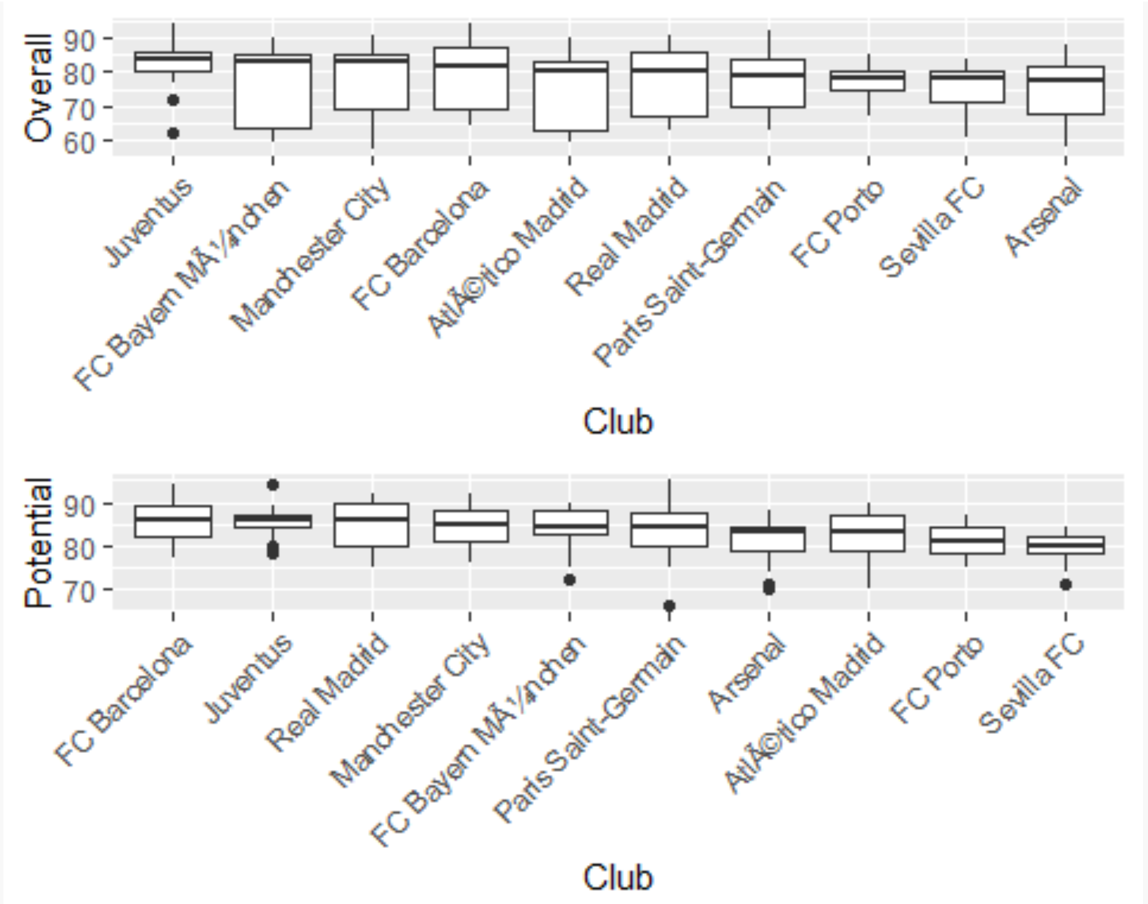
1. We used both the built-in summary() function and the describe() function in the Psych package. The describe function can show all the descriptive statistics asked.

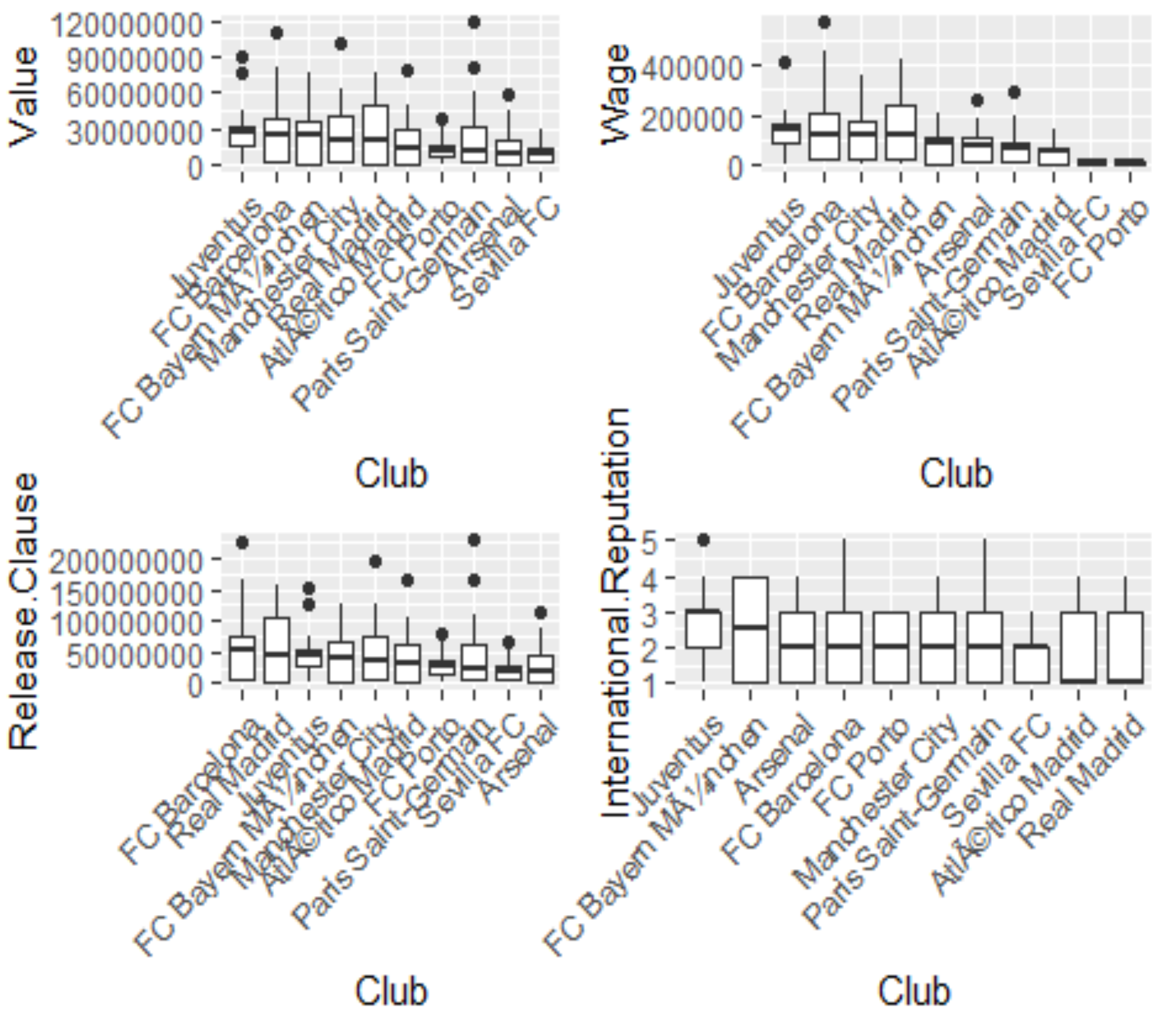
2. Besides describing the overall statistics of target variables, we also compared the descriptive statistics of the FIFA top 3 clubs on March 8 ("Real Madrid","FC Bayern MÃ¼nchen","FC Barcelona"). We can tell that all three clubs have lower-than-average age, and higher Overall and Potential scores, average wage and International Reputation than the average. Therefore, we assume that there will be some linear and curvilinear relationships in the data. As there are a lot of columns showing scores of a player for different positions and for different physical and mental capabilities. As there are overall score, potential score and detailed scores, we doubt that there may be redundancy and many variables will be take out while building the predictive model, so we first looked at several variables that we think are relevant and important and show how these variables can contribute to the market value of a player (visualized by clubs).

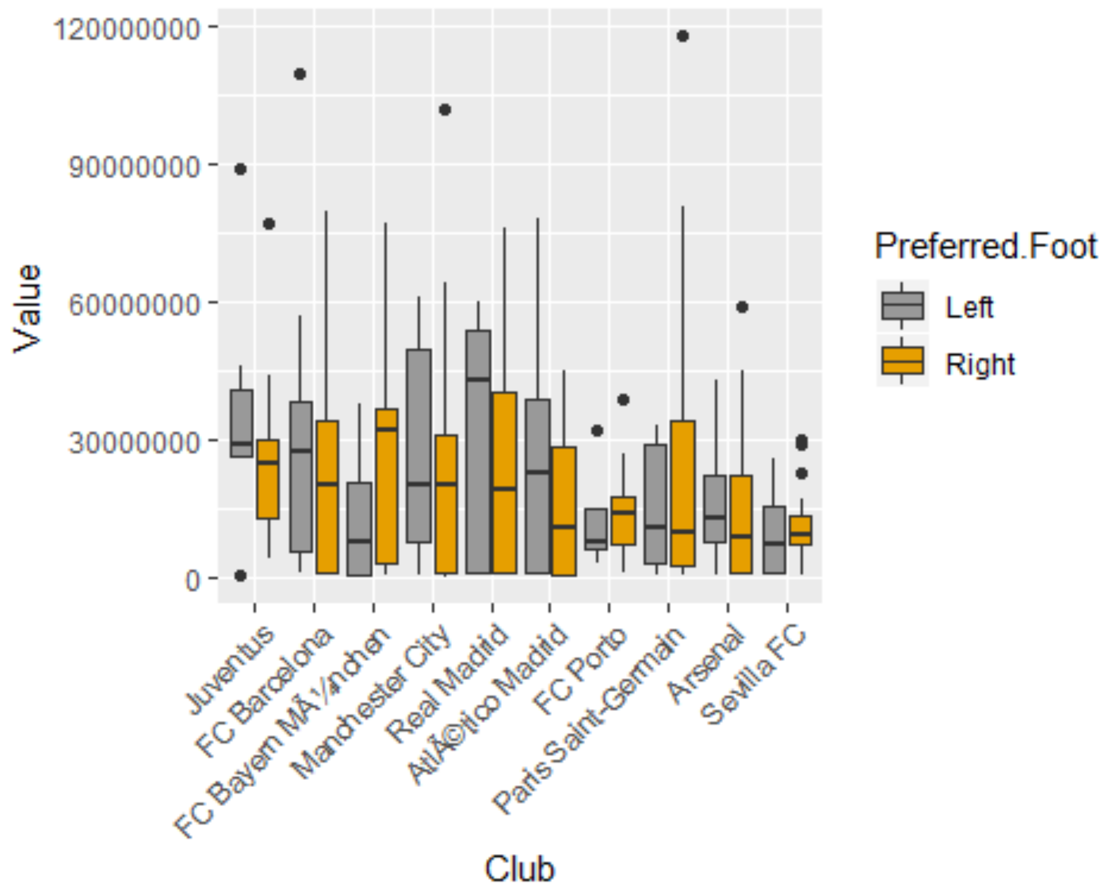
- **box-and-whisker plots for relevant and important variables.**

Going further by grouping observations into different clubs, we made box plots for the top 10 clubs on some important variables and reordered the plots by medians to show rankings of the top 10 clubs in different aspects. For example, Juventus has the highest overall score while FC Barcelona has the highest potential score. Juventus is the best in value, wage and international reputation while FC Barcelona has the highest in released clause. The difference of these rankings with the Mar 8 club rankings may come from the fact that the FIFA dataset we use in this project has not been updated to include data till Mar 8 2019.

```
# Top 10 clubs on Mar 8. source: https://www.footballseeding.com/club-ranking/a2018-2019/
```

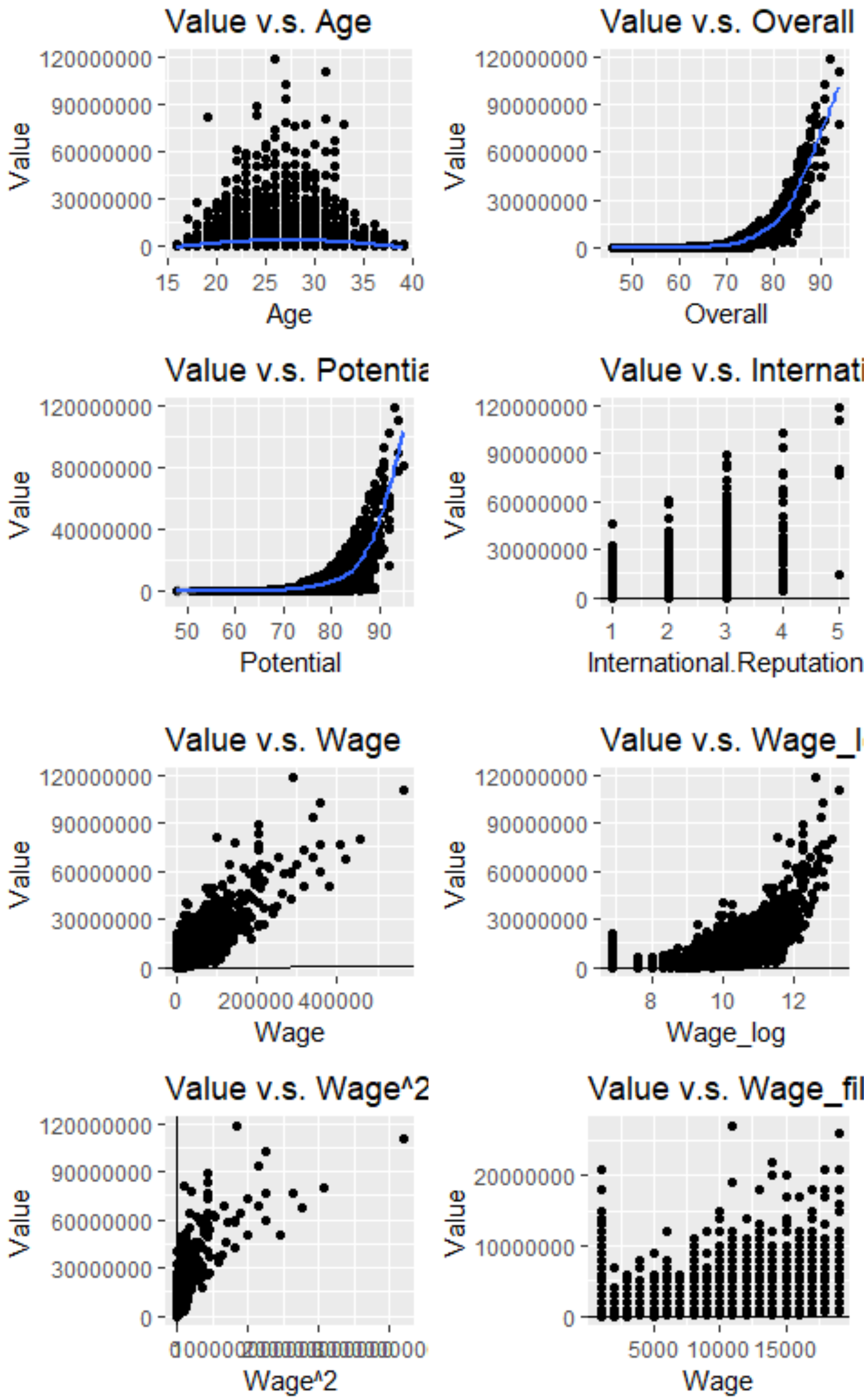

Overall
90
80
70
60
Juventus
FC Bayern MÃ¼nchen
Manchester City
FC Barcelona
AtlÃ©tico Madrid
Real Madrid
Paris Saint-Germain
FC Porto
Sevilla FC
Arsenal
Club
Potential
90
80
70
FC Barcelona
Juventus
Real Madrid
Manchester City
FC Bayern MÃ¼nchen
Paris Saint-Germain
Arsenal
AtlÃ©tico Madrid
FC Porto
Sevilla FC
Club

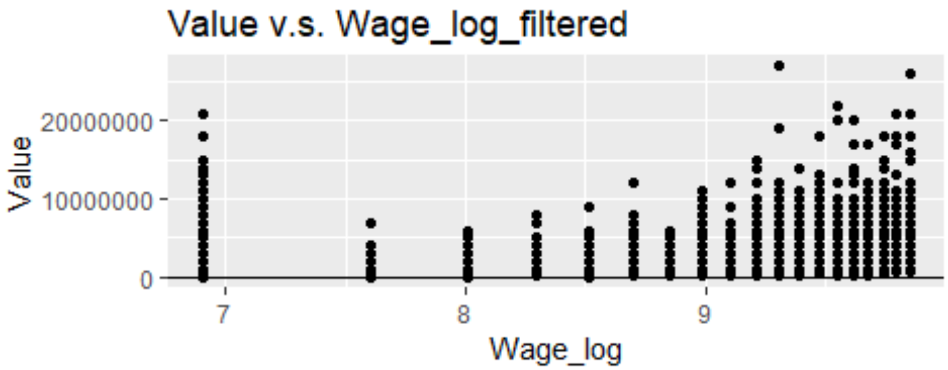

Value
120000000
90000000
60000000
30000000
0
Juventus
FC Barcelona
FC Bayern MÃ¼nchen
Manchester City
Real Madrid
AtlÃ©tico Madrid
FC Porto
Paris Saint-Germain
Arsenal
Sevilla FC
Club
Wage
400000
200000
0
Juventus
FC Barcelona
Manchester City
Real Madrid
FC Bayern MÃ¼nchen
Arsenal
Paris Saint-Germain
AtlÃ©tico Madrid
Sevilla FC
FC Porto
Club
Release.Clause
200000000
150000000
100000000
50000000
0
FC Barcelona
Real Madrid
Juventus
FC Bayern MÃ¼nchen
Manchester City
AtlÃ©tico Madrid
FC Porto
Paris Saint-Germain
Sevilla FC
Arsenal
Club
International.Reputation
5
4
3
2
1
Juventus
FC Bayern MÃ¼nchen
Arsenal
FC Barcelona
FC Porto
Manchester City
Paris Saint-Germain
Sevilla FC
AtlÃ©tico Madrid
Real Madrid
Club



2. Create a scatterplot among the variables to find potentially linear or curvilinear relationships. That should help you identify both a target variable and candidate predictor variables.

We tested the relationship between value (market value) and potential candidate predictor variables including age, Overall score, Potential score, International.Reputation and wage. The result shows some trends like value peaks at the age of 27-28, increases before this age and decreases afterwards; higher international reputation contributes to higher value. And there’s a clear curvilinear relationship between value and overall or potential score. However, it’s hard to find a clear relationship between wage and market value, so more attention could be paid on this when doing multiple regression.


Value v.s. Age
Value
Age
Value v.s. Overall
Value
Overall
Value v.s. Potentia
Value
Potential
Value v.s. Internati
Value
International.Reputation
Value v.s. Wage
Value
Wage
Value v.s. Wage_l
Value
Wage_log
Value v.s. Wage^2
Value
Wage^2
Value v.s. Wage_fil
Value
Wage



Choose a target variable and justify that choice.

Therefore, given the descriptive statistics analysis above, we decided to use Value as our target variable. This makes sense because considering the content and aim of FIFA data, we can probably tell that a popular use of the FIFA data is to predict the market value of players and clubs. And the descriptive analysis also shows that it's possible to predict the value using multiple candidate variables for multiple regression.

Part 2. Predictive Modeling: Multiple Regression

1. Perform Multiple Regression

We take the "*Market Value*" of players (in thousands of dollars) as our target value, which is a continuous numeric variable. And select *"Age", "Overall", "Potential", "Value", "Wage", "Preferred.Foot", "International.Reputation", "Weak.Foot", "Work.Rate", "Body.Type", "Height" and "Weight"* as our independent variables to train the model.

And we choose to use *forward selection* to select the best model with the lowest AIC.

Below is the diagnostics plots of the *training model*:

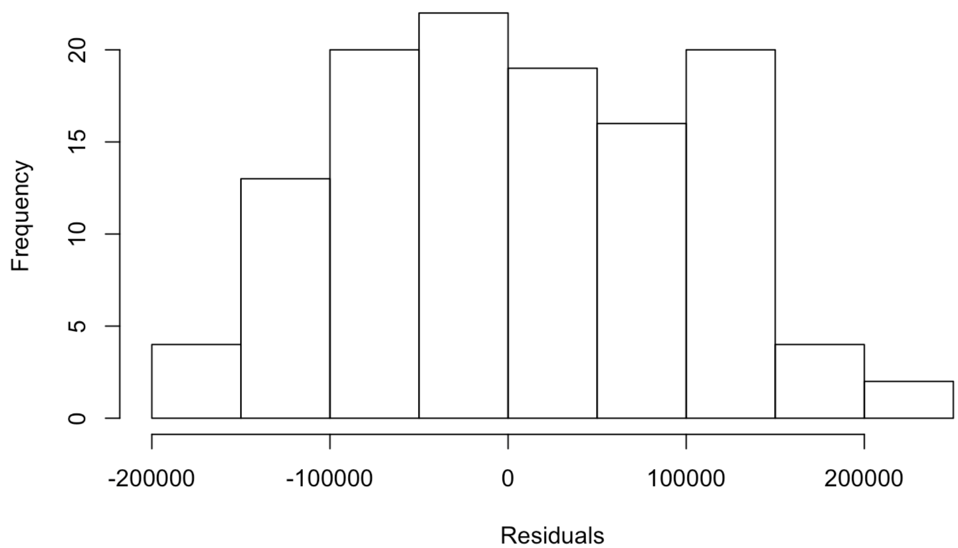


Diagram 1. After Partitioning the dataset and Removing Outliers

After removing the outliers, the histogram of the residuals from the training model (*Diagram 1*) is slightly right-skewed and there is a little bump between 100000 and 150000 but overall is Gaussian distributed.

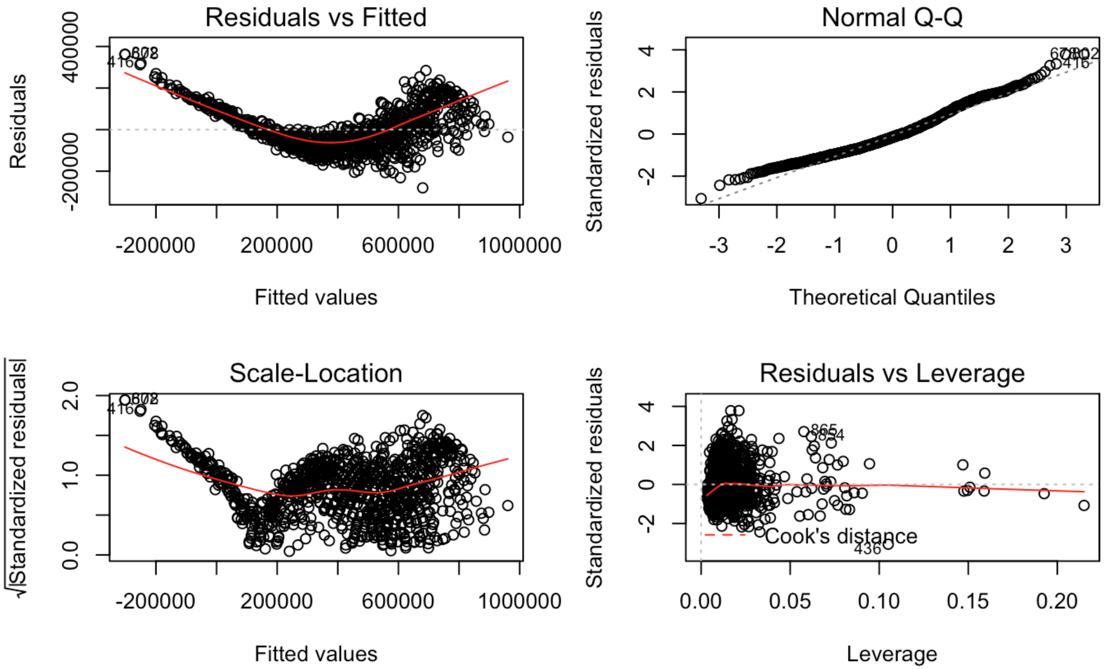


Diagram 2. Residuals vs. Fitted Values & Normal Probability Plot of Residual

Since this dataset includes several positions for different soccer players, and different position has different attributes. Besides, a group of famous players has a significantly different distribution to their peers. Thus, to better fit the model, we only choose the data for "*Striker*" position and remove the players who earn more than 1,000,000 thousand dollars per year. And *Diagram 2* is the final result.

4. **What is the adjusted R-Squared value of your best model? What is the RMSE? Include some diagnostic residual plots with your final, best model, to show that you have minimized outliers.**

After the model selection process by using *'forward selection'* approach, the software gives us the best model which only includes *Overall, Age, Wage, Body.Type, and Potential*. The reported adjusted R-squared is 0.891, indicating that the model has a quite good in-sample fit. The reported RMSE is 87650. Compared with the RMSE in the training model (95189), it also indicates that this selected model has a good out-of-sample fit.

The following Diagram 3 shows the *Residuals vs. Fitted Values* and *Normal Probability Plot of Residual* plot for the best model selected. As we can see from the *Residuals vs. Fitted Values* plot, data points are more randomly distributed compared with the one in Diagram 2. And by comparing the *Normal Probability Plot of Residual* Plot in Diagram 2 and 3, it is obvious that the number of outliers has largely decreased.

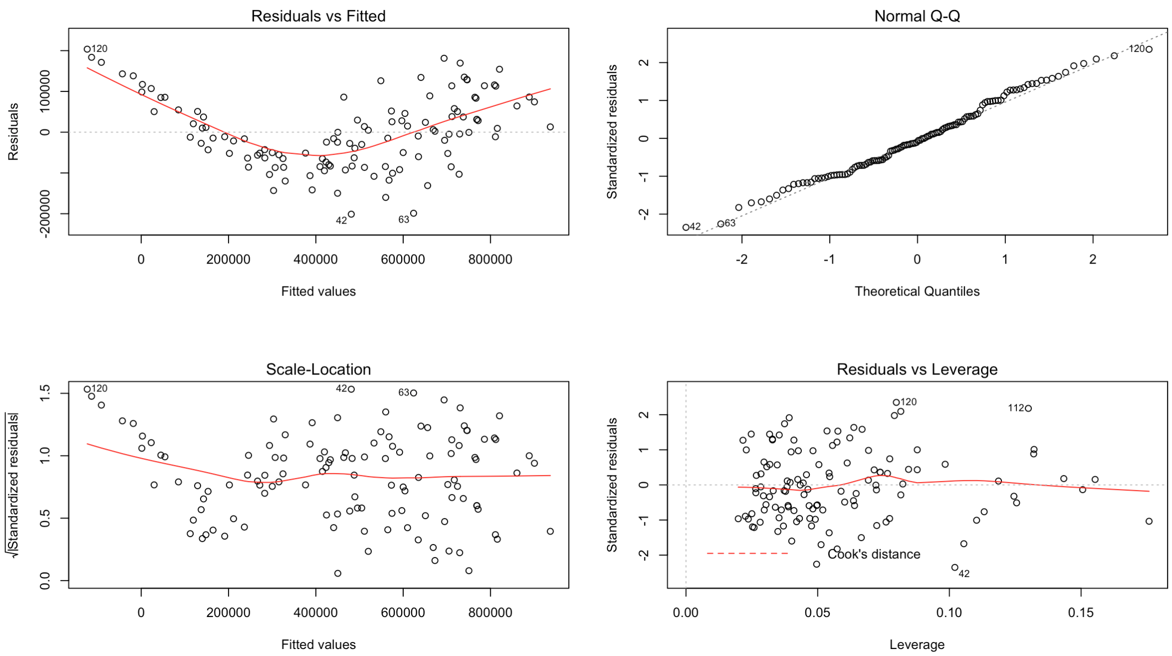


Diagram 3. Residuals vs. Fitted Values & Normal Probability Plot of Residual (Best Model)

The following Diagram 4 shows the *Histogram of Residuals* for the best model. Even though it is a little bit right-skewed, but still it is Gaussian distributed.

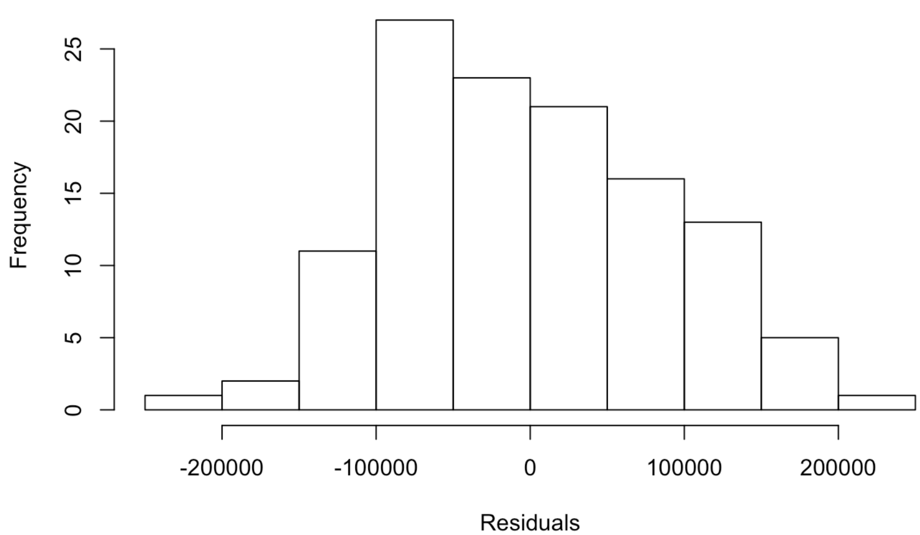


Diagram 4. Histogram of Residual (Best Model)

Create and try to include at least one interaction or polynomial term, i.e., higher-order term.

To further explore a potential better model, we add 2 second-order variables and 4 interaction term to our best model: the square of age(*Age_sq)*, the square of wage(*Wage_sq*), *Body.Type*Age, Body.Type*Potential, Body.Type*Wage* and *Body.Type*Overall*.

5. What is your final adjusted R-Squared after trying to include higher-order/interaction terms? What is the RMSE?

(1)Result from regression after including the square of Age:

```
Call:
lm(formula = Value ~ Overall + Age + Wage + Body.Type + Potential +
    Age_sq, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-242945  -72701  -19176   50355  377221

Coefficients:
                  Estimate  Std. Error t value             Pr(>|t|)
(Intercept)     -3910614.67   143431.08  -27.26 < 0.0000000000000002 ***
Overall            50393.34     1629.98   30.92 < 0.0000000000000002 ***
Age                67406.86     9683.83    6.96      0.0000000000059 ***
Wage                  12.43        1.39    8.91 < 0.0000000000000002 ***
Body.TypeNormal    -1489.02     6210.44   -0.24                 0.81
Body.TypeStocky     4570.95    12239.12    0.37                 0.71
Potential           8695.60     1404.62    6.19      0.0000000008521 ***
Age_sq             -1639.75      164.87   -9.95 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 93600 on 1070 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.868
F-statistic: 1.02e+03 on 7 and 1070 DF,  p-value: <0.0000000000000002

            ME  RMSE   MAE MPE MAPE
Test set 10520 89424 76279   3   26
```

The adjusted R-squared after including the square of Age is 0.868 and the RMSE is 89424.

(2)Result from regression after including the square of Wage:

```
Call:
lm(formula = Value ~ Overall + Age + Body.Type + Potential +
    Wage + Wage_sq, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-286472  -71666  -16268   59887  383978

Coefficients:
                       Estimate      Std. Error t value            Pr(>|t|)
(Intercept)     -2605117.768221    62563.879093  -41.64 < 0.0000000000000002 ***
Overall            61839.323984     1210.682589   51.08 < 0.0000000000000002 ***
Age               -28153.477348     1282.300112  -21.96 < 0.0000000000000002 ***
Body.TypeNormal      173.548732     6488.243472    0.03                0.979
Body.TypeStocky     3069.446729    12790.505986    0.24                0.810
Potential          -1658.690860      989.885542   -1.68                0.094 .
Wage                  14.048328        3.434406    4.09             0.000046 ***
Wage_sq               -0.000117        0.000239   -0.49                0.626
---
Signif. codes:  0 ‘***’ 0 ‘**’ 0 ‘*’ 0 ‘.’ 0 ‘ ’ 1

Residual standard error: 97800 on 1070 degrees of freedom
Multiple R-squared:  0.857,     Adjusted R-squared:  0.856
F-statistic:  917 on 7 and 1070 DF,  p-value: <0.0000000000000002

           ME  RMSE   MAE MPE MAPE
Test set 7385 96332 81265   1   27
```

The adjusted R-squared after including the square of Wage is 0.856 and the RMSE is 96332.

(3)Result from regression after including *Body.Type*Age*:

```
Call:
lm(formula = Value ~ Overall + Age + Body.Type + Body.Type *
    Age + Potential + Wage, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-293455  -71078  -17062   60688  381874

Coefficients:
                       Estimate  Std. Error t value            Pr(>|t|)
(Intercept)         -2631679.73    64593.92  -40.74 <0.0000000000000002 ***
Overall                61604.10     1214.80   50.71 <0.0000000000000002 ***
Age                   -26959.64     1647.19  -16.37 <0.0000000000000002 ***
Body.TypeNormal        21239.70    31502.95    0.67               0.500
Body.TypeStocky       109301.44    64922.44    1.68               0.093 .
Potential              -1420.58      996.49   -1.43               0.154
Wage                      12.49        1.46    8.56 <0.0000000000000002 ***
Age:Body.TypeNormal      -925.84     1328.22   -0.70               0.486
Age:Body.TypeStocky     -4195.90     2506.77   -1.67               0.094 .
---
Signif. codes:  0 ‘***’ 0 ‘**’ 0 ‘*’ 0 ‘.’ 0 ‘ ’ 1

Residual standard error: 97700 on 1069 degrees of freedom
Multiple R-squared:  0.858,     Adjusted R-squared:  0.856
F-statistic:  804 on 8 and 1069 DF,  p-value: <0.0000000000000002

           ME  RMSE   MAE MPE MAPE
Test set 7928 96372 81829   1   27
```

The adjusted R-squared after including *Body.Type*Ag*e is 0.856 and the RMSE is 96372.

(4)Result from regression after including *Body.Type*Potential*:

```
Call:
lm(formula = Value ~ Overall + Age + Body.Type + Body.Type *
    Potential + Potential + Wage, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-294505  -70832  -15137   59739  390159

Coefficients:
                             Estimate  Std. Error t value            Pr(>|t|)
(Intercept)               -2450362.54    80773.34  -30.34 <0.0000000000000002 ***
Overall                     61718.82     1195.60   51.62 <0.0000000000000002 ***
Age                        -27821.42     1282.87  -21.69 <0.0000000000000002 ***
Body.TypeNormal           -260788.35    87951.67   -2.97              0.0031 **
Body.TypeStocky           -301504.02   198896.44   -1.52              0.1298
Potential                   -3867.47     1229.33   -3.15              0.0017 **
Wage                           12.43        1.45    8.56 <0.0000000000000002 ***
Body.TypeNormal:Potential    3810.11     1281.41    2.97              0.0030 **
Body.TypeStocky:Potential    4465.02     2939.82    1.52              0.1291
---
Signif. codes:  0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 97400 on 1069 degrees of freedom
Multiple R-squared:  0.858,     Adjusted R-squared:  0.857
F-statistic:  810 on 8 and 1069 DF,  p-value: <0.0000000000000002

           ME  RMSE   MAE MPE MAPE
Test set 6053 95022 80527  -0   26
```

The adjusted R-squared after including *Body.Type*Potential* is 0.857 and the RMSE is 95022.

(5)Result from regression after including *Body.Type*Wage*:

```
Call:
lm(formula = Value ~ Overall + Age + Body.Type + Body.Type *
    Wage + Potential + Wage, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-298718  -71287  -16607   60597  385334

Coefficients:
                        Estimate  Std. Error t value            Pr(>|t|)
(Intercept)          -2607526.60    61930.47  -42.10 < 0.0000000000000002 ***
Overall                 61961.34     1200.57   51.61 < 0.0000000000000002 ***
Age                    -28163.35     1286.06  -21.90 < 0.0000000000000002 ***
Body.TypeNormal         -3604.96     9515.96   -0.38                0.705
Body.TypeStocky         -3656.30    20955.19   -0.17                0.862
Wage                       11.51        2.29    5.04           0.00000056 ***
Potential               -1654.98      991.22   -1.67                0.095 .
Body.TypeNormal:Wage        1.47        2.71    0.54                0.588
Body.TypeStocky:Wage        2.59        6.21    0.42                0.677
---
Signif. codes:  0 ‘***’ 0 ‘**’ 0 ‘*’ 0 ‘.’ 0 ‘ ’ 1

Residual standard error: 97900 on 1069 degrees of freedom
Multiple R-squared:  0.857,     Adjusted R-squared:  0.856
F-statistic:  802 on 8 and 1069 DF,  p-value: <0.0000000000000002

           ME  RMSE   MAE MPE MAPE
Test set 7385 96418 81341   1   27
```

The adjusted R-squared after including *Body.Type*Wage* is 0.856 and the RMSE is 96418.

(6)Result from regression after including *Body.Type*Overall*:

```
Call:
lm(formula = Value ~ Overall + Age + Body.Type + Body.Type *
    Overall + Potential + Wage, data = train.t_more)

Residuals:
    Min      1Q  Median      3Q     Max
-297561  -70435  -14044   61272  402225

Coefficients:
                           Estimate  Std. Error t value            Pr(>|t|)
(Intercept)             -2500781.78    77226.19  -32.38 <0.0000000000000002 ***
Overall                   60369.58     1365.53   44.21 <0.0000000000000002 ***
Age                      -28362.12     1282.69  -22.11 <0.0000000000000002 ***
Body.TypeNormal         -177760.07    80043.81   -2.22               0.027 *
Body.TypeStocky         -255112.93   195006.39   -1.31               0.191
Potential                 -1781.73      988.60   -1.80               0.072 .
Wage                         12.37        1.46    8.49 <0.0000000000000002 ***
Overall:Body.TypeNormal    2929.88     1312.45    2.23               0.026 *
Overall:Body.TypeStocky    4176.05     3098.80    1.35               0.178
---
Signif. codes:  0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 97600 on 1069 degrees of freedom
Multiple R-squared:  0.858,     Adjusted R-squared:  0.857
F-statistic:  807 on 8 and 1069 DF,  p-value: <0.0000000000000002

           ME  RMSE   MAE MPE MAPE
Test set 6364 95209 80088   1   26
```
The adjusted R-squared after including *Body.Type*Overall* is 0.857 and the RMSE is 95209.

6. How many of your observations were removed outliers? What percentage of your observations is that? Does that seem acceptable?

We implement two steps to remove the outliers. First, after we filter the dataset based on whether the players are strikers and partition the dataset on a 90% and 10% base. We plot the training dataset and figure out the there are basically two clusters of residuals, which have different distributions. Therefore, we decide to set a cut-off point at 1000000 based on the *Value* variable of players. This will allow us seperate the "super star" players from the "average" players. As a result, 1198 observations("average" players) are selected out of 1924 observation. Next, We further removed 12 outliers from the training dataset to ensure that the residuals are Gaussian distributed.

Overall, 738 observations are removed outliers, which is around 38%. Even if the percentage seems to be fairly large, most of the outliers are removed because they followed a non-Gaussian distribution. Therefore, the percentage seems acceptable because it allows us to generate a model with unbiased estimation of coefficients.

7. If your final model is different, because of higher-order terms, what is it?

Interpret the (beta) coefficients?

***Bolded variables are additional ones based on original best model.**

Model Variables	Adjusted R-squared	RMSE
Age, Wage, Body.Type, and Potential	0.891	87650
Age, **Age_sq**, Wage, Body.Type, and Potential	0.868	89424
Age, Wage, **Wage_sq** Body.Type, and Potential	0.856	96332
Age, Wage, Body.Type, **Body.Type*Age** and Potential	0.856	96372
Age, Wage, Body.Type, **Body.Type*Potential** and Potential	0.857	95022
Age, Wage, Body.Type, **Body.Type*Wage** and Potential	0.856	96418
Age, Wage, Body.Type, **Body.Type*Overall** and Potential	0.857	95209

After comparing the adjusted R-squared and RMSE, our final model will still be our previous best model generated by forward selection with no addition of interaction and polynomial terms.

The following is the result from the final best model:

```
Coefficients:
                  Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)     -2708283.86   192961.36  -14.04 < 0.0000000000000002 ***
Overall           69010.35     3272.14   21.09 < 0.0000000000000002 ***
Age              -30966.38     3358.41   -9.22   0.0000000000000019 ***
Wage                 15.85        4.57    3.47              0.00074 ***
Body.TypeNormal    5883.53    20030.75    0.29              0.76951    
Body.TypeStocky   73277.93    29719.63    2.47              0.01518 *  
Potential         -5786.23     3014.67   -1.92              0.05746 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90300 on 113 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.891 
F-statistic:  164 on 6 and 113 DF,  p-value: <0.0000000000000002

         ME  RMSE   MAE MPE MAPE
Test set -0 87650 71626   4   28
```

Based on the above coefficients, we can reach the following conclusions:

(1) As players' overall rating increments by 1 point, their market values will increase around €69010 on average.
(2) The market values for players will drop around €30966 on average as players get one year elder, so young players definitely have an advantage.
(3) As players wage increase by €1, their market values will increase around €15.85.
(4) After we filter the original dataset based on players' position and market value, only players with three categories of body types are left : lean, normal and stocky. Compared with the market values of players with lean body type, the market values of players with normal type increases by €5885 on average. However, this effect is not significant. The market values of players with stocky body type increases by €73278 on average compared to those of players with lean type.
(5) To our surprise, the market values for players will decrease around €5786 on average as players potential rating increase by 1 point. However, the effect for the potential rating is barely significant.

Overall, incremental rating, wage and stock body type will boost players' market values, whereas incremental age and potential rating will reduce the market values of players.

Model Interpretation and Reflections

What conclusions do you draw from your model, your interpretation of the coefficients, and your process in terms of insights for manufacturing, marketing, financing, or other business functions?

- ## Model Interpretation

We build this regression model in order to help club managers to detect underrated football strikers and to make an informed decision.

Based on our regression model and FIFA 2019 player dataset, we can conclude that the market values of strikers are determined by the players' overall rating, age, wage, body type and potential. The higher the overall rating, the higher the market value. Age is also a significant determinant for player's market value. As the age goes up, the market value decreases. The player's current wage affects the market value in the same direction. When wage grows, market value also increases. Stocky body type is an advantage for football players in striker position. Stocky strikers have higher average market value. Surprisingly, the market value decreases a little when the player's potential rating goes up. Further discussion is required to explain the reason behind it.

If we look at the significance and magnitude of the coefficients, we will have some more interesting discoveries.

1. Overall rating includes the player's performance on different positions. When players' overall rating increases by 1 point, the market value increases by €69010.35. This coefficient is both statistically and economically significant.

2. Aging is a big problem for football players. Their market value falls dramatically as their age increases. On average, when the player grows 1 year older, his market value decreases by €30966.38. This coefficient is also statistically and economically significant, which is a cruel news for a market that the players' average age is around 25 years old.

3. Although players' current wage is a statistically significant variable in terms of predicting the market value, its effect on market value is relatively small. €1 increase in wage only results in €15.85 increase in market value. Compare to the average wage and the average market value, this coefficient is not economically significant.

4. People tend to have this stereotype that stocky football players generally have better performance. We cannot make concrete conclusion that stocky players have better performance during the game, but our research does reveal some correlation between body type and the players' market value. Stocky players' average market value is €73278 more than lean players'.

- ## Reflection

We confronted with lots of obstacles during our model building process.

We started the model building process in the hope of finding a magic regression model that can detect all the factors that determine a player's market value, so that coaches can make training plan for the players accordingly.

However, when we plotted out the residual vs. fitted values & normal probability plot of residual diagram, we found out that the outliers were more than we could get rid of manually. We tried several ways to deal with the outliers, including deleting the first and last 10% observations after the data was sorted by residuals, deleting all the guard keepers and using k means clustering to divide the players into groups that had different average market value. None of the above methods worked. Eventually, after consulting our professor and having tons of group discussion, we removed all the "superstars" with high market value and focused our attention on strikers.

Although this process was painful at first, we learned that taking opinions from different sources and learning from failure can help us approaching our final goal. Also, this process enhanced our problem-solving skills and team-playing skills.

Although our final model tells us some interesting stories about the market value, it cannot help coaches make training program because most of the variables that determine the market value like age and body type are not controllable. However, we found out that our model can help club managers to detect underrated football strikers and to make an informed decision when they want to purchase a valuable player.