What Influence Sucess in NBA

2024-12-01

```r
options(max.print = 10000)
```

Introduction

Our project focuses on understanding what defines success in the NBA. Success in this context can be measured in various ways, such as a team's budget, player salaries, or win percentages. Our goal is to explore how these factors influence success and uncover key drivers behind achieving excellence in the league. In this analysis we will explore:

- Player's Salaries vs Player demographics. We want to see if certain demographics like height, age, weight influence performance and salaries
- Team Budget vs Team performance : With this we'll investigate relationship between teams budgets and overall performance

Datasets

The datasets used in the analysis are Players Salary, NBA demographics,Performance and Coaches dataset(We included more details of the datasets in our report) From the dataset mentioned these, we created two merged datasets, which serve as the foundation for our analysis.

Analysis

Our analysis focuses on the 2020-2021 to 2022-2023 NBA seasons.We will aim to uncover insights on how player demographics influence their salaries and how it correlate to overall teams success. By applying multiple linear regression, the expected outcome is to identify significant predictors for each season. These predictors will demonstrate their relevance to the model and provide a clearer understanding of how they relate to teams success in the NBA.

Combined dataset

The final dataset used for analysis consists of 22 columns and 1,468 rows. The columns include:

Player Information: player_name: The name of the player. team_abbreviation: The abbreviation of the team the player belongs to (e.g., "TOR" for Toronto Raptors). age: The player's age (in years). player_height: The player's height (in centimeters). player_weight: The player's weight (in kilograms). college: The city where the college or university the player attended is located. country: The country the player is from. draft_year: The year the player was drafted into the NBA or "Undrafted" if not selected. draft_round: The round in which the player was drafted (e.g., "1" for the first round, "Undrafted" if not selected). draft_number: The overall pick number in the draft (e.g., "8" for the 8th pick, "Undrafted" if not selected). season: The NBA season (e.g., "2020-21").

Performance Stats: gp: Games played during the season. pts: Average points scored per game (points per game). reb: Average rebounds per game (includes offensive and defensive rebounds). ast: Average assists per game. **Advanced Metrics:** net_rating: The difference between the team's offensive rating and defensive rating when the player is on the court. oreb_pct: Offensive rebound percentage — the percentage of available offensive rebounds grabbed by the player. dreb_pct: Defensive rebound percentage — the percentage of available defensive rebounds grabbed by the player. usg_pct: Usage percentage — an estimate of the percentage of team plays used by the player while on the court. ts_pct: True shooting percentage — a measure of shooting efficiency that accounts for field goals, 3-point field goals, and free throws. ast_pct: Assist percentage — an estimate of the percentage of teammates' field goals assisted by the player while on the court.

Response Variable salary: The player's salary for the corresponding season (in dollars).

Step one : Prepare the data:

```r
player_salaries <- read.csv("players_salaries_2020_2023.csv")

head(player_salaries)
```

```
##      player_name team_abbreviation age player_height player_weight
## 1 Gary Trent Jr.               TOR  22        195.58      94.80073
## 2    Gary Harris               ORL  26        193.04      95.25432
## 3     Gary Clark               PHI  26        198.12     102.05820
## 4   Gabriel Deck               OKC  26        198.12     104.77975
## 5 Garrett Temple               CHI  35        195.58      88.45044
## 6   Gabe Vincent               MIA  25        190.50      90.71840
##                   college   country draft_number gp  pts reb ast net_rating
## 1                    Duke       USA            3 58 15.3 2.6 1.4       -1.8
## 2          Michigan State       USA            2 39  9.9 2.0 2.0       -4.2
## 3              Cincinnati       USA            5 39  3.1 2.9 0.8       -7.7
## 4                         Argentina            5 10  8.4 4.0 2.4      -12.7
## 5         Louisiana State       USA            5 56  7.6 2.9 2.2        1.5
## 6 California-Santa Barbara       USA            5 50  4.8 1.1 1.3       -5.9
##   oreb_pct dreb_pct usg_pct ts_pct ast_pct  season    salary
## 1    0.014    0.069   0.204  0.534   0.067 2020-21  1.663861
## 2    0.019    0.054   0.164  0.511   0.102 2020-21 19.610714
## 3    0.044    0.125   0.097  0.436   0.064 2020-21  2.018458
## 4    0.067    0.118   0.159  0.548   0.160 2020-21  3.870370
## 5    0.019    0.082   0.126  0.525   0.104 2020-21  4.767000
## 6    0.016    0.068   0.184  0.498   0.161 2020-21  0.660750
```

Clean the data:

```r
player_salaries <- na.omit(player_salaries)
```

Assigning variables to the seasons we'll be working with below:

```r
player_salaries_2020_21 <- player_salaries[player_salaries$season == "2020-21", ]
player_salaries_2021_22 <- player_salaries[player_salaries$season == "2021-22", ]
player_salaries_2022_23 <- player_salaries[player_salaries$season == "2022-23", ]
```

```r
# looking at 2020-2021 season
tail(player_salaries_2020_21)
```

```
##             player_name team_abbreviation age player_height player_weight
## 439           Max Strus               MIA  25        195.58      97.52228
## 440          Maxi Kleber               DAL  29        208.28     108.86208
## 441 Matthew Dellavedova               CLE  30        190.50      90.71840
## 442          Matt Thomas               UTA  26        193.04      86.18248
## 443     Matisse Thybulle               PHI  24        195.58      91.17199
## 444        Mason Plumlee               DET  31        210.82     115.21237
##                           college   country draft_number gp  pts reb ast
## 439                        DePaul       USA            5 39  6.1 1.1 0.6
## 440                                 Germany            5 50  7.1 5.2 1.4
## 441 St.Mary's College of California Australia            5 13  2.8 1.8 4.5
## 442                    Iowa State       USA            5 45  3.1 1.0 0.4
## 443                    Washington       USA            2 65  3.9 1.9 1.0
## 444                          Duke       USA            2 56 10.4 9.3 3.6
##     net_rating oreb_pct dreb_pct usg_pct ts_pct ast_pct  season   salary
## 439       -4.2    0.011    0.073   0.179  0.597   0.074 2020-21 0.647098
## 440        4.6    0.035    0.151   0.103  0.606   0.066 2020-21 8.475000
## 441       -3.1    0.029    0.085   0.125  0.312   0.337 2020-21 2.174318
## 442       -9.3    0.020    0.112   0.187  0.522   0.096 2020-21 1.517981
## 443        4.6    0.023    0.070   0.090  0.508   0.064 2020-21 2.711280
## 444       -4.9    0.095    0.264   0.163  0.638   0.205 2020-21 8.000000
```

```r
# looking at 2021-2022 season
tail(player_salaries_2021_22)
```

```
##          player_name team_abbreviation age player_height player_weight
## 875 Tim Hardaway Jr.               DAL  30        195.58      92.98636
## 876    Tobias Harris               PHI  29        200.66     102.51179
## 877 Tomas Satoransky               WAS  30        200.66      95.25432
## 878     Tony Bradley               CHI  24        208.28     112.49082
## 879       Tony Snell               NOP  30        198.12      96.61510
## 880     Terry Rozier               CHA  28        185.42      86.18248
##            college        country draft_number gp  pts reb ast net_rating
## 875       Michigan            USA            2 42 14.2 3.7 2.2        1.3
## 876      Tennessee            USA            2 73 17.2 6.8 3.5        3.2
## 877                Czech Republic            3 55  3.6 2.3 3.3       -8.0
## 878 North Carolina            USA            2 55  3.0 3.4 0.5        5.2
## 879     New Mexico            USA            2 53  3.5 1.9 0.5       -7.8
## 880     Louisville            USA            2 73 19.3 4.3 4.5        1.4
##     oreb_pct dreb_pct usg_pct ts_pct ast_pct  season    salary
## 875    0.010    0.114   0.214  0.520   0.115 2021-22 21.306816
## 876    0.032    0.164   0.214  0.566   0.158 2021-22 35.995950
## 877    0.029    0.110   0.124  0.461   0.285 2021-22 10.000000
## 878    0.123    0.196   0.128  0.600   0.065 2021-22  1.789256
## 879    0.018    0.107   0.099  0.541   0.047 2021-22  2.389641
## 880    0.021    0.101   0.227  0.566   0.197 2021-22 17.905263
```

```r
# looking at 2022-2023 season
tail(player_salaries_2022_23)
```

```
##         player_name team_abbreviation age player_height player_weight
## 1324     Joe Ingles               MIL  35        205.74      99.79024
## 1325   Joe Wieskamp               TOR  23        198.12      92.98636
## 1326    Joel Embiid               PHI  29        213.36     127.00576
## 1327   John Collins               ATL  25        205.74     102.51179
## 1328   Jericho Sims               NYK  24        208.28     113.39800
## 1329 JaMychal Green               GSW  33        205.74     102.96538
##          college   country draft_number gp  pts  reb ast net_rating oreb_pct
## 1324             Australia            5 46  6.9  2.8 3.3        2.5    0.012
## 1325        Iowa       USA            3  9  1.0  0.4 0.3        1.0    0.000
## 1326      Kansas  Cameroon            1 66 33.1 10.2 4.2        8.8    0.057
## 1327 Wake Forest       USA            2 71 13.1  6.5 1.2       -0.2    0.035
## 1328       Texas       USA            5 52  3.4  4.7 0.5       -6.7    0.117
## 1329     Alabama       USA            5 57  6.4  3.6 0.9       -8.2    0.087
##      dreb_pct usg_pct ts_pct ast_pct  season    salary
## 1324    0.102   0.122  0.616   0.181 2022-23  6.479000
## 1325    0.068   0.115  0.321   0.083 2022-23  2.909261
## 1326    0.243   0.370  0.655   0.233 2022-23 33.616770
## 1327    0.180   0.168  0.593   0.052 2022-23 23.500000
## 1328    0.175   0.074  0.780   0.044 2022-23  1.639842
## 1329    0.164   0.169  0.650   0.094 2022-23  8.200000
```

Create Multiple linear regression model

In this analysis, we have performed a Multiple linear regression where the dependent variable is salary, and the independent variables include age, player height, player weight, games played (GP), points (PTS), rebounds (REB), assists (AST), net rating, offensive rebound percentage (OREB_PCT), defensive rebound percentage (DREB_PCT), usage percentage (USG_PCT), true shooting percentage (TS_PCT), assist percentage (AST_PCT), team abbreviation, college, country, and draft number.

Step 1: Create a full additive model:

The first step in this analysis is to create a comprehensive model that includes all possible variables and predictors. To streamline the process of model creation, a function has been developed to automate this step:

```
create_model_by_season <- function(players_salaries_Variable) {
  # Ensure salary is numeric
  players_salaries_Variable$salary <- as.numeric(players_salaries_Variable$salary)

  factors <- c("team_abbreviation", "college", "country", "draft_number")

for (factor in factors) {
  if (length(unique(players_salaries_Variable[[factor]])) > 1) {
    players_salaries_Variable[[factor]] <- as.factor(players_salaries_Variable[[factor]])
  } else {
    players_salaries_Variable[[factor]] <- NULL # Remove columns with only one unique value
 }
}

 # Full Model
```

```r
model <- lm(salary ~ age + player_height + player_weight + gp + pts + reb + ast +net_rating + oreb_pct +
              team_abbreviation + college + country  + draft_number
              , data = players_salaries_Variable)

  # Return the model summary
  return((model))
}
```

To create the model, it is necessary to call the function with the corresponding dataset. Note that the dataset has been filtered by season, so we need three separate variables, each holding the data for the respective seasons.

```r
# Create and summarize models for each season
full_model_2020_21 <- create_model_by_season(player_salaries_2020_21)
full_model_2021_22 <- create_model_by_season(player_salaries_2021_22)
full_model_2022_23 <- create_model_by_season(player_salaries_2022_23)

# Print summaries
#cat("Model Summary for 2020-21 Season:\n")
print(summary(full_model_2020_21))
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + player_weight + gp +
##     pts + reb + ast + net_rating + oreb_pct + dreb_pct + usg_pct +
##     ts_pct + ast_pct + team_abbreviation + college + country +
##     draft_number, data = players_salaries_Variable)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.795  -2.284   0.000   1.907  17.773
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -23.49724   17.35208  -1.354 0.177019
## age                                       0.65038    0.09766   6.660    2e-10
## player_height                             0.06097    0.08916   0.684 0.494809
## player_weight                             0.02036    0.05601   0.364 0.716530
## gp                                       -0.07343    0.02557  -2.871 0.004467
## pts                                       0.55365    0.19938   2.777 0.005940
## reb                                       0.47416    0.46092   1.029 0.304694
## ast                                       1.47370    0.65781   2.240 0.026027
## net_rating                                0.02105    0.05369   0.392 0.695318
## oreb_pct                                -27.70085   17.80603  -1.556 0.121154
## dreb_pct                                 -1.92189   15.98297  -0.120 0.904393
## usg_pct                                   2.94030   16.34229   0.180 0.857374
## ts_pct                                   -4.95387    5.38158  -0.921 0.358263
## ast_pct                                  -4.50412   12.69171  -0.355 0.723000
## team_abbreviationBKN                      6.70001    2.94519   2.275 0.023833
## team_abbreviationBOS                      3.59270    2.81426   1.277 0.203028
## team_abbreviationCHA                      4.74796    3.21556   1.477 0.141162
## team_abbreviationCHI                      3.65748    2.96366   1.234 0.218421
## team_abbreviationCLE                      3.26816    2.94117   1.111 0.267652
```

```
## team_abbreviationDAL                    3.54490    2.72260   1.302 0.194210
## team_abbreviationDEN                    5.95661    2.82002   2.112 0.035742
## team_abbreviationDET                    2.85662    2.80352   1.019 0.309301
## team_abbreviationGSW                    4.91174    2.98008   1.648 0.100678
## team_abbreviationHOU                    3.40250    3.10814   1.095 0.274790
## team_abbreviationIND                    5.31114    2.85391   1.861 0.064020
## team_abbreviationLAC                    3.66832    3.02665   1.212 0.226753
## team_abbreviationLAL                    5.61513    2.99187   1.877 0.061812
## team_abbreviationMEM                    3.46040    2.83177   1.222 0.222962
## team_abbreviationMIA                    1.84886    2.83684   0.652 0.515225
## team_abbreviationMIL                    3.28144    3.00981   1.090 0.276744
## team_abbreviationMIN                    5.56088    2.73299   2.035 0.043026
## team_abbreviationNOP                    4.41342    2.96126   1.490 0.137492
## team_abbreviationNYK                    1.53111    2.91157   0.526 0.599485
## team_abbreviationOKC                    0.52648    2.95274   0.178 0.858642
## team_abbreviationORL                    1.67472    3.02070   0.554 0.579833
## team_abbreviationPHI                    6.28746    2.80102   2.245 0.025739
## team_abbreviationPHX                    5.65946    3.03731   1.863 0.063693
## team_abbreviationPOR                    3.14236    3.26276   0.963 0.336510
## team_abbreviationSAC                    3.09705    2.98984   1.036 0.301356
## team_abbreviationSAS                    3.16953    2.93690   1.079 0.281624
## team_abbreviationTOR                    4.80148    3.22434   1.489 0.137821
## team_abbreviationUTA                    6.16419    2.94765   2.091 0.037606
## team_abbreviationWAS                    5.03359    2.98962   1.684 0.093599
## collegeAlabama                         -2.73884    4.83176  -0.567 0.571375
## collegeArizona                          2.98976    3.03982   0.984 0.326379
## collegeArizona State                    3.34171    4.98979   0.670 0.503715
## collegeArkansas                         2.62543    3.72285   0.705 0.481386
## collegeArkansas-Little Rock             3.76099    6.57588   0.572 0.567923
## collegeAuburn                           5.32904    4.17374   1.277 0.202960
## collegeBaylor                           4.23807    4.79072   0.885 0.377274
## collegeBelmont                          1.09259    6.45348   0.169 0.865708
## collegeBoise State                     -1.58372    6.63586  -0.239 0.811581
## collegeBoston College                  -5.00163    4.17355  -1.198 0.231990
## collegeBowling Green                   -2.65853    6.55645  -0.405 0.685499
## collegeBucknell                        -1.60764    6.53619  -0.246 0.805933
## collegeButler                           5.08151    4.80835   1.057 0.291707
## collegeCal Poly                        -0.59656    6.52199  -0.091 0.927199
## collegeCalifornia                       8.75626    6.47148   1.353 0.177367
## collegeCalifornia-Santa Barbara         6.01582    6.47976   0.928 0.354173
## collegeCentral Florida                 -9.00502    9.63994  -0.934 0.351213
## collegeCincinnati                       2.04079    6.37453   0.320 0.749147
## collegeCollege of Charleston            1.13981    6.53083   0.175 0.861605
## collegeColorado                        -4.57578    4.20413  -1.088 0.277557
## collegeConnecticut                      8.61512    3.72539   2.313 0.021632
## collegeCreighton                       -7.74152    4.14329  -1.868 0.062971
## collegeDavidson                        13.49222    6.80301   1.983 0.048527
## collegeDayton                           8.60053    6.52526   1.318 0.188803
## collegeDePaul                           5.22384    4.85031   1.077 0.282604
## collegeDrexel                           0.92340    6.57967   0.140 0.888513
## collegeDuke                             1.08389    2.40276   0.451 0.652341
## collegeFlorida                          0.39150    4.42867   0.088 0.929635
## collegeFlorida Gulf Coast               8.20271    6.62670   1.238 0.217042
## collegeFlorida State                    3.10529    3.09887   1.002 0.317363
```

```
## collegeFresno State                      4.24569    4.21692   1.007 0.315078
## collegeGeorge Washington                 2.90486    9.79054   0.297 0.766963
## collegeGeorgetown                      -12.07245    6.56118  -1.840 0.067059
## collegeGeorgia                           0.37490    4.11178   0.091 0.927432
## collegeGeorgia Tech                     -0.97221    3.98075  -0.244 0.807273
## collegeGonzaga                          -3.65149    4.21285  -0.867 0.386981
## collegeHouston                          -1.19392    6.46174  -0.185 0.853574
## collegeIllinois                          3.57671    6.63956   0.539 0.590617
## collegeIndiana                           3.02044    3.22603   0.936 0.350115
## collegeIndiana-Purdue Fort Wayne         6.24031    6.53260   0.955 0.340451
## collegeIndiana-Purdue Indianapolis      -4.85543    6.41389  -0.757 0.449813
## collegeIona                              5.37011    6.69905   0.802 0.423598
## collegeIowa                              5.90745    6.55180   0.902 0.368184
## collegeIowa State                       -1.78885    3.50621  -0.510 0.610405
## collegeKansas                            0.09374    3.57481   0.026 0.979103
## collegeKansas State                      5.17544    4.85889   1.065 0.287926
## collegeKentucky                          1.04162    2.31836   0.449 0.653644
## collegeLehigh                            9.85667    6.55629   1.503 0.134110
## collegeLipscomb                          5.15754    6.59448   0.782 0.434960
## collegeLouisana-Lafayette               -0.63520    6.45683  -0.098 0.921719
## collegeLouisiana State                   3.81629    3.42917   1.113 0.266918
## collegeLouisiana Tech                   -2.36445    6.60328  -0.358 0.720618
## collegeLouisville                       -1.71244    3.76156  -0.455 0.649361
## collegeMarquette                         3.23331    3.48301   0.928 0.354221
## collegeMarshall                         -4.61364    6.61621  -0.697 0.486304
## collegeMaryland                          2.21541    4.16561   0.532 0.595356
## collegeMemphis                           0.29411    3.83759   0.077 0.938976
## collegeMiami                            -1.95852    6.61889  -0.296 0.767574
## collegeMichigan                          0.01754    2.94845   0.006 0.995258
## collegeMichigan State                    3.11489    3.01399   1.033 0.302466
## collegeMinnesota                         6.36796    5.09769   1.249 0.212867
## collegeMississippi                       4.06954    6.57147   0.619 0.536349
## collegeMississippi State                 4.21602    5.01378   0.841 0.401285
## collegeMissouri                          3.18098    4.84810   0.656 0.512397
## collegeMissouri State                   -1.84697    7.09218  -0.260 0.794770
## collegeMontana State                     5.66366    6.69103   0.846 0.398179
## collegeMurray State                     -5.55053    4.84639  -1.145 0.253279
## collegeNebraska-Lincoln                  3.27348    6.53896   0.501 0.617123
## collegeNevada                            0.70686    6.75654   0.105 0.916770
## collegeNevada-Reno                      -0.56335    4.21086  -0.134 0.893690
## collegeNew Mexico                       12.26903    6.57884   1.865 0.063466
## collegeNew Mexico State                  6.91807    9.59768   0.721 0.471760
## collegeNorth Carolina                    1.25393    2.68170   0.468 0.640521
## collegeNotre Dame                        2.79249    6.60639   0.423 0.672912
## collegeOakland                          -0.44659    6.53057  -0.068 0.945539
## collegeOhio State                        6.47947    3.67655   1.762 0.079334
## collegeOklahoma                          4.83424    4.53348   1.066 0.287388
## collegeOklahoma State                   -1.40208    6.56721  -0.213 0.831128
## collegeOld Dominion                     -0.87711    6.57612  -0.133 0.894011
## collegeOregon                           -0.17485    4.07775  -0.043 0.965836
## collegeOregon State                      6.42458    6.56836   0.978 0.329048
## collegePenn State                        1.11347    4.87761   0.228 0.819631
## collegePittsburgh                       -1.07838    6.62878  -0.163 0.870912
## collegeProvidence                        3.88706    6.79205   0.572 0.567680
```

```
## collegePurdue                            -0.55133    4.99549  -0.110 0.912215
## collegeRadford                            5.84360    6.58505   0.887 0.375789
## collegeSan Diego State                    3.51379    4.11154   0.855 0.393653
## collegeSouth Carolina                     1.21044    4.90966   0.247 0.805482
## collegeSouth Carolina Upstate            -0.25977    6.64516  -0.039 0.968851
## collegeSouthern California                3.96004    3.19426   1.240 0.216335
## collegeSouthern Methodist                -1.73786    4.12268  -0.422 0.673757
## collegeSt. John's                         1.97317    6.48055   0.304 0.761041
## collegeSt. Joseph's (PA)                  1.51735    6.74133   0.225 0.822116
## collegeSt.Mary's College of California  -6.53379    7.01106  -0.932 0.352352
## collegeStanford                           4.06155    3.52705   1.152 0.250704
## collegeSyracuse                          -3.46298    3.39125  -1.021 0.308256
## collegeTCU                                3.46999    4.80670   0.722 0.471085
## collegeTennessee                          8.25422    4.05027   2.038 0.042700
## collegeTennessee State                    8.74388    6.68022   1.309 0.191867
## collegeTexas A&M                          5.37257    4.14760   1.295 0.196500
## collegeTexas Tech                         4.55044    4.87168   0.934 0.351253
## collegeTexas-Austin                       3.46542    3.09510   1.120 0.264031
## collegeTulsa                              3.57257    4.87241   0.733 0.464168
## collegeUCLA                               3.69583    2.75857   1.340 0.181645
## collegeUniversity of Texas at Austin      0.73235    4.86685   0.150 0.880519
## collegeUNLV                               3.20175    4.61763   0.693 0.488774
## collegeUtah                              -2.72280    4.82389  -0.564 0.573005
## collegeUtah State                         5.92052    6.58504   0.899 0.369546
## collegeVanderbilt                         1.43194    3.45390   0.415 0.678831
## collegeVillanova                         -0.51296    3.07302  -0.167 0.867576
## collegeVirginia                           1.18403    3.06353   0.386 0.699489
## collegeVirginia Tech                     -4.91984    6.79118  -0.724 0.469529
## collegeWake Forest                        3.77747    3.48411   1.084 0.279411
## collegeWashington                         0.95554    2.89467   0.330 0.741623
## collegeWashington State                   2.53061    4.99010   0.507 0.612552
## collegeWeber State                        6.14240    6.63921   0.925 0.355848
## collegeWest Virginia                      4.05067    6.64080   0.610 0.542486
## collegeWestern Kentucky                   1.67115    6.75568   0.247 0.804842
## collegeWichita State                      1.47076    4.97698   0.296 0.767870
## collegeWilliam & Mary                     9.54040    6.71370   1.421 0.156661
## collegeWisconsin                         -7.02692    6.50715  -1.080 0.281328
## collegeWisconsin-Green Bay                3.88532    6.58731   0.590 0.555891
## collegeWyoming                            2.55582    4.80184   0.532 0.595061
## collegeXavier                            -0.98159    4.86179  -0.202 0.840175
## collegeYale                               6.79646    6.48721   1.048 0.295890
## countryArgentina                         -5.74195    8.69168  -0.661 0.509513
## countryAustralia                         -3.74960    7.75498  -0.484 0.629195
## countryAustria                           -0.26833   10.57280  -0.025 0.979775
## countryBahamas                           -3.77177    9.06077  -0.416 0.677598
## countryBosnia and Herzegovina            -5.18831    9.72447  -0.534 0.594181
## countryBrazil                            -6.27233    7.91251  -0.793 0.428764
## countryCameroon                          -2.46824    9.97014  -0.248 0.804693
## countryCanada                            -4.42785    7.39670  -0.599 0.550012
## countryCroatia                           -3.03129    8.17406  -0.371 0.711096
## countryCzech Republic                    -5.02491    9.77632  -0.514 0.607754
## countryDominican Republic                -0.09827   10.32962  -0.010 0.992418
## countryDRC                               -6.92565    9.71933  -0.713 0.476838
## countryEgypt                             -6.35348   10.08018  -0.630 0.529128
```

```
## countryFinland                        -11.01988    9.88596  -1.115 0.266142
## countryFrance                          -1.54066    7.78831  -0.198 0.843363
## countryGabon                           -0.09457   10.41886  -0.009 0.992765
## countryGeorgia                         -3.62004    9.67913  -0.374 0.708746
## countryGermany                         -3.24549    7.88876  -0.411 0.681157
## countryGreece                          -2.90762    8.76797  -0.332 0.740480
## countryGuinea                          -4.79953    9.79199  -0.490 0.624496
## countryIsrael                          -4.81663    9.66021  -0.499 0.618534
## countryItaly                            0.05186    8.41464   0.006 0.995088
## countryJamaica                         -4.97964    9.68199  -0.514 0.607522
## countryJapan                           -7.27339   10.22996  -0.711 0.477812
## countryLatvia                           1.65603    8.26206   0.200 0.841315
## countryLithuania                       -4.74145    8.17458  -0.580 0.562467
## countryMontenegro                      -5.76353   10.07582  -0.572 0.567870
## countryNew Zealand                     16.94520   11.72826   1.445 0.149870
## countryNigeria                         -1.38562    8.40852  -0.165 0.869256
## countryRepublic of the Congo           -8.72458    9.72307  -0.897 0.370493
## countrySaint Lucia                     -5.21086   10.44346  -0.499 0.618285
## countrySenegal                          6.47666   10.06715   0.643 0.520641
## countrySerbia                          -3.16016    7.80207  -0.405 0.685824
## countrySlovenia                        -9.73124    8.30309  -1.172 0.242409
## countrySouth Sudan                     -4.54870    8.77435  -0.518 0.604671
## countrySpain                           -9.70324    8.29103  -1.170 0.243079
## countrySudan                           -3.51016   10.40852  -0.337 0.736244
## countrySwitzerland                      5.49249    9.75331   0.563 0.573887
## countryTurkey                          -9.56386    8.05190  -1.188 0.236147
## countryUkraine                         -9.81183    8.90949  -1.101 0.271927
## countryUnited Kingdom                 -16.31586    9.99722  -1.632 0.104039
## countryUSA                             -6.86293    7.17211  -0.957 0.339627
## draft_number2                          -0.90051    1.13475  -0.794 0.428257
## draft_number3                          -1.53634    1.22073  -1.259 0.209473
## draft_number4                          -2.06003    1.46813  -1.403 0.161915
## draft_number5                          -4.63868    1.31851  -3.518 0.000524
##
## (Intercept)
## age                                   ***
## player_height
## player_weight
## gp                                    **
## pts                                   **
## reb
## ast                                   *
## net_rating
## oreb_pct
## dreb_pct
## usg_pct
## ts_pct
## ast_pct
## team_abbreviationBKN                  *
## team_abbreviationBOS
## team_abbreviationCHA
## team_abbreviationCHI
## team_abbreviationCLE
## team_abbreviationDAL
```

```
## team_abbreviationDEN                   *
## team_abbreviationDET
## team_abbreviationGSW
## team_abbreviationHOU
## team_abbreviationIND                   .
## team_abbreviationLAC
## team_abbreviationLAL                   .
## team_abbreviationMEM
## team_abbreviationMIA
## team_abbreviationMIL
## team_abbreviationMIN                   *
## team_abbreviationNOP
## team_abbreviationNYK
## team_abbreviationOKC
## team_abbreviationORL
## team_abbreviationPHI                   *
## team_abbreviationPHX                   .
## team_abbreviationPOR
## team_abbreviationSAC
## team_abbreviationSAS
## team_abbreviationTOR
## team_abbreviationUTA                   *
## team_abbreviationWAS                   .
## collegeAlabama
## collegeArizona
## collegeArizona State
## collegeArkansas
## collegeArkansas-Little Rock
## collegeAuburn
## collegeBaylor
## collegeBelmont
## collegeBoise State
## collegeBoston College
## collegeBowling Green
## collegeBucknell
## collegeButler
## collegeCal Poly
## collegeCalifornia
## collegeCalifornia-Santa Barbara
## collegeCentral Florida
## collegeCincinnati
## collegeCollege of Charleston
## collegeColorado
## collegeConnecticut                     *
## collegeCreighton                       .
## collegeDavidson                        *
## collegeDayton
## collegeDePaul
## collegeDrexel
## collegeDuke
## collegeFlorida
## collegeFlorida Gulf Coast
## collegeFlorida State
## collegeFresno State
```

```
## collegeGeorge Washington
## collegeGeorgetown                     .
## collegeGeorgia
## collegeGeorgia Tech
## collegeGonzaga
## collegeHouston
## collegeIllinois
## collegeIndiana
## collegeIndiana-Purdue Fort Wayne
## collegeIndiana-Purdue Indianapolis
## collegeIona
## collegeIowa
## collegeIowa State
## collegeKansas
## collegeKansas State
## collegeKentucky
## collegeLehigh
## collegeLipscomb
## collegeLouisana-Lafayette
## collegeLouisiana State
## collegeLouisiana Tech
## collegeLouisville
## collegeMarquette
## collegeMarshall
## collegeMaryland
## collegeMemphis
## collegeMiami
## collegeMichigan
## collegeMichigan State
## collegeMinnesota
## collegeMississippi
## collegeMississippi State
## collegeMissouri
## collegeMissouri State
## collegeMontana State
## collegeMurray State
## collegeNebraska-Lincoln
## collegeNevada
## collegeNevada-Reno
## collegeNew Mexico                     .
## collegeNew Mexico State
## collegeNorth Carolina
## collegeNotre Dame
## collegeOakland
## collegeOhio State                     .
## collegeOklahoma
## collegeOklahoma State
## collegeOld Dominion
## collegeOregon
## collegeOregon State
## collegePenn State
## collegePittsburgh
## collegeProvidence
## collegePurdue
```

```
## collegeRadford
## collegeSan Diego State
## collegeSouth Carolina
## collegeSouth Carolina Upstate
## collegeSouthern California
## collegeSouthern Methodist
## collegeSt. John's
## collegeSt. Joseph's (PA)
## collegeSt.Mary's College of California
## collegeStanford
## collegeSyracuse
## collegeTCU
## collegeTennessee                        *
## collegeTennessee State
## collegeTexas A&M
## collegeTexas Tech
## collegeTexas-Austin
## collegeTulsa
## collegeUCLA
## collegeUniversity of Texas at Austin
## collegeUNLV
## collegeUtah
## collegeUtah State
## collegeVanderbilt
## collegeVillanova
## collegeVirginia
## collegeVirginia Tech
## collegeWake Forest
## collegeWashington
## collegeWashington State
## collegeWeber State
## collegeWest Virginia
## collegeWestern Kentucky
## collegeWichita State
## collegeWilliam & Mary
## collegeWisconsin
## collegeWisconsin-Green Bay
## collegeWyoming
## collegeXavier
## collegeYale
## countryArgentina
## countryAustralia
## countryAustria
## countryBahamas
## countryBosnia and Herzegovina
## countryBrazil
## countryCameroon
## countryCanada
## countryCroatia
## countryCzech Republic
## countryDominican Republic
## countryDRC
## countryEgypt
## countryFinland
```

```
## countryFrance
## countryGabon
## countryGeorgia
## countryGermany
## countryGreece
## countryGuinea
## countryIsrael
## countryItaly
## countryJamaica
## countryJapan
## countryLatvia
## countryLithuania
## countryMontenegro
## countryNew Zealand
## countryNigeria
## countryRepublic of the Congo
## countrySaint Lucia
## countrySenegal
## countrySerbia
## countrySlovenia
## countrySouth Sudan
## countrySpain
## countrySudan
## countrySwitzerland
## countryTurkey
## countryUkraine
## countryUnited Kingdom
## countryUSA
## draft_number2
## draft_number3
## draft_number4
## draft_number5                          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.753 on 230 degrees of freedom
## Multiple R-squared:  0.8078, Adjusted R-squared:  0.6297
## F-statistic: 4.537 on 213 and 230 DF,  p-value: < 2.2e-16
```

```
#cat("\nModel Summary for 2021-22 Season:\n")
#print(summary(full_model_2021_22))

#cat("\nModel Summary for 2022-23 Season:\n")
#print(summary(full_model_2022_23))
```

The results obtained from this full model:

Adjusted R-squared: 0.6297 **p-value:** < 2.2e-16

Adjusted R-squared (0.6297): Indicate that approximately 62.97% of the variation in the dependent variable (player salaries) can be explained by the model, accounting for the number of predictors.

p-value (< 2.2e-16)is less than 0.05 at significance level , suggesting that the model as a whole is statistically significant.

Step 1.1

Selecting the best additive model

Once the Models had been created. It is possible to use the selection methods to create and chose the best additive model for this analysis. In this case, the subset predictor method is the one being used to select the best additive method:

Note: Our criteria for selecting the best additive model the criteria is base on:

- High R^2
- Low RSE
- Low AIC
- Low BIC

In this case, **Step Backward Procedure** is the one being used to select the best additive method:

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
backward_model_2020_2021=ols_step_backward_p(full_model_2020_21, p_val = 0.05)
```

In order to make this reduce method compatible with the annova command in needs to be saved:

```
reduce_model_2020_21<- backward_model_2020_2021$model
```

To display the information from the model we use:

```
summary(backward_model_2020_2021$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(c(include, cterms), collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8621  -3.2901  -0.0413   2.8946  20.4295
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -38.24184    6.98503  -5.475 7.42e-08 ***
## age             0.63705    0.06332  10.061  < 2e-16 ***
## player_height   0.12131    0.03393   3.575  0.00039 ***
## gp             -0.05204    0.01571  -3.312  0.00100 **
## pts             0.65533    0.06547  10.010  < 2e-16 ***
```

```
## ast             1.40450    0.22205   6.325 6.30e-10 ***
## draft_number2  -1.16745    0.77091  -1.514  0.13066
## draft_number3  -2.06649    0.82939  -2.492  0.01309 *
## draft_number4  -2.66862    0.97641  -2.733  0.00653 **
## draft_number5  -3.52393    0.81123  -4.344 1.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.464 on 434 degrees of freedom
## Multiple R-squared:  0.6729, Adjusted R-squared:  0.6661
## F-statistic: 99.18 on 9 and 434 DF,  p-value: < 2.2e-16
```

The results obtained form the above model are:

Adjusted R-squared: 0.6661 **p-value:** < 2.2e-16

Adjusted R-squared (0.6661): This means that approximately 66.61% of the variation in the dependent variable (player salaries) is explained by the model, after adjusting for the number of predictors.

p-value (< 2.2e-16): small p_value less than 0.05 at significant level suggest that the model is statistically significant. This means there is strong evidence to suggest that the predictors in the model are having a meaningful effect on the dependent variable.

Step 1.2

Create Anonova table to test the new additive model

From test the new additive model it is necessary to compare with the previous full model and analyze the p value to make sure that the dropped values were not significant and that the model pass the hypothesis test.

$$H_0 : B_1 = B_{2...} = B_{droppedPredictors} = 0 \quad VS \quad H_a : B_i \neq 0$$

- The null hypothesis test that all the dropped predictors are insignificant (equal to zero).
- The alternative hypothesis test that at least one predictor is significant (different from zero) for this model.

The anova table to compare the models:

```
anova(reduce_model_2020_21, full_model_2020_21)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ age + player_height + gp + pts + ast + draft_number
## Model 2: salary ~ age + player_height + player_weight + gp + pts + reb +
##     ast + net_rating + oreb_pct + dreb_pct + usg_pct + ts_pct +
##     ast_pct + team_abbreviation + college + country + draft_number
##   Res.Df     RSS  Df Sum of Sq      F Pr(>F)
## 1    434 12955.0
## 2    230  7612.4 204    5342.6 0.7913 0.9564
```

From the results from the anova table we can see that the **p-value** is **0.9564** these being bigger than α at **0.05**. Then we fail to reject the null hypothesis and conclude that the dropped predictors were insignificant and accept our reduce model.

the reduce model then would look like this:

$$\widehat{Salary} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{age} + \widehat{\beta}_2 X_{playerHeight} + \widehat{\beta}_3 X_{gp} + \widehat{\beta}_4 X_{pts} + \widehat{\beta}_5 X_{ast} + \widehat{\beta}_6 X_{draftNumber2} + \widehat{\beta}_7 X_{draftNumber3} + \widehat{\beta}_8 X_{draftNumber4} + \widehat{\beta}_9$$

Step 2: Use of the interaction model

After having a reduce additive model, the next step is to create an interactive model to verify its behavior.

```
interactive_reduce_model <- function(players_salaries_Variable) {

intereactive_model <- lm(salary ~ (age + player_height + gp + pts + ast + draft_number)^2, data=players_

}
```

```
# Call the funtion to create the interactive model, here we are assuming that the predictor are same...
interactive_reduce_model_2020_2021 <- interactive_reduce_model(player_salaries_2020_21)
interactive_reduce_model_2021_2022 <- interactive_reduce_model(player_salaries_2021_22)
interactive_reduce_model_2022_2023 <- interactive_reduce_model(player_salaries_2022_23)


# Print summaries
#cat("Interactive Model Summary for 2020-21 Season:\n")
print(summary(interactive_reduce_model_2020_2021))
```

```
##
## Call:
## lm(formula = salary ~ (age + player_height + gp + pts + ast +
##     draft_number)^2, data = players_salaries_Variable)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5373  -2.2926  -0.1398   1.5112  21.3932
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.325e+01  4.338e+01  -0.306  0.76013
## age                        6.183e-01  1.472e+00   0.420  0.67471
## player_height              1.176e-01  2.112e-01   0.557  0.57788
## gp                        -1.251e-01  3.713e-01  -0.337  0.73635
## pts                       -3.605e+00  1.372e+00  -2.627  0.00894 **
## ast                       -3.165e+00  4.303e+00  -0.736  0.46237
## draft_number              -1.911e-01  4.269e+00  -0.045  0.96431
## age:player_height         -4.361e-03  7.123e-03  -0.612  0.54073
## age:gp                    -5.209e-03  3.034e-03  -1.717  0.08670 .
## age:pts                    8.162e-02  1.432e-02   5.700 2.25e-08 ***
## age:ast                    1.400e-01  4.244e-02   3.298  0.00106 **
## age:draft_number           3.315e-02  4.082e-02   0.812  0.41720
## player_height:gp           7.628e-04  1.834e-03   0.416  0.67763
## player_height:pts          1.127e-02  6.299e-03   1.789  0.07435 .
## player_height:ast          3.006e-03  2.027e-02   0.148  0.88219
## player_height:draft_number -7.182e-03  2.069e-02  -0.347  0.72870
```

```
## gp:pts                       2.471e-03  3.277e-03   0.754  0.45116
## gp:ast                       1.013e-02  1.240e-02   0.817  0.41457
## gp:draft_number              1.455e-02  9.347e-03   1.557  0.12033
## pts:ast                     -2.331e-02  2.241e-02  -1.040  0.29884
## pts:draft_number            -7.959e-02  3.969e-02  -2.005  0.04557 *
## ast:draft_number            -1.094e-01  1.436e-01  -0.762  0.44668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.536 on 422 degrees of freedom
## Multiple R-squared:  0.7807, Adjusted R-squared:  0.7698
## F-statistic: 71.54 on 21 and 422 DF,  p-value: < 2.2e-16
```

```
#cat("\nInteractive Model Summary for 2021-22 Season:\n")
#print(summary(interactive_reduce_model_2021_2022))

#cat("\nInteractive Model Summary for 2022-23 Season:\n")
#print(summary(interactive_reduce_model_2022_2023))
```

The results obtained from the above model are :

Adjusted R-squared: 0.7698 **p-value:** < 2.2e-16

Adjusted R-squared (0.7698): This value indicates that approximately 76.98% of the variation in the dependent variable is explained by the interactive model, after adjusting for the number of predictors.

p-value (< 2.2e-16): The small p-value indicates that the model as a whole is statistically significant. l

Test the hypothesis test

$$H_0 : B_1 = B_{2...} = B_n = 0 \quad VS \quad H_a : B_i \neq 0$$

- The null hypothesis test that all of the predictors are insignificant (equal to zero).
- The alternative hypothesis test that at least one predictor is significant (different from zero) for this model.

The summary shows that the global p value for the interaction model is 2.2e-16, this value is less that α, therefore we reject the null hypothesis in favor to the alternative and conclude that at least one of the predictors are significant for the model.

Step 2.1

Reduced Interaction Model

From the previous summary, it is possible to identify p values that are bigger than $\alpha = 0.05$, therefore they can be dropped to simplify the model.

The new reduce interaction model should hold only the following predictors.

$\widehat{Salary} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{age} + \widehat{\beta}_2 X_{playerHeight} + \widehat{\beta}_3 X_{gp} + \widehat{\beta}_4 X_{pts} + \widehat{\beta}_5 X_{ast} + \widehat{\beta}_6 X_{draftNumber} + \widehat{\beta}_7 X_{age} X_{pts} + \widehat{\beta}_8 X_{age} X_{ast} + \widehat{\beta}_9 X_{draft_numbe}$

To create the reduction reduce interaction model:

```r
reduce_interactive_model <- function(players_salaries_Variable) {

  reduce_intereactive_model <- lm(salary ~ age + player_height + gp + pts + ast + factor(draft_number) +

}
```

```r
# Call the funtion to create the interactive model, here we are assuming that the predictor are same...
reduce_interactive_model_2020_2021 <- reduce_interactive_model(player_salaries_2020_21)
reduce_interactive_model_2021_2022 <- reduce_interactive_model(player_salaries_2021_22)
reduce_interactive_model_2022_2023 <- reduce_interactive_model(player_salaries_2022_23)


# Print summaries
#cat("Interactive Model Summary for 2020-21 Season:\n")
print(summary(reduce_interactive_model_2020_2021))
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + ast +
##     factor(draft_number) + age * pts + age * ast + pts * draft_number,
##     data = players_salaries_Variable)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1476  -2.3439  -0.2762   1.7150  20.8481
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -15.43471    6.22048  -2.481 0.013472 *
## age                   -0.31159    0.09453  -3.296 0.001061 **
## player_height          0.12433    0.02855   4.355 1.67e-05 ***
## gp                    -0.03249    0.01369  -2.373 0.018080 *
## pts                   -1.00223    0.34928  -2.869 0.004315 **
## ast                   -2.79398    1.05548  -2.647 0.008416 **
## factor(draft_number)2  -0.75379    0.75278  -1.001 0.317228
## factor(draft_number)3  -1.17089    0.95572  -1.225 0.221192
## factor(draft_number)4  -1.52825    1.14235  -1.338 0.181664
## factor(draft_number)5  -1.51372    1.13631  -1.332 0.183519
## draft_number                NA         NA      NA       NA
## age:pts                0.06935    0.01255   5.528 5.62e-08 ***
## age:ast                0.14495    0.03745   3.870 0.000126 ***
## pts:draft_number      -0.06270    0.02931  -2.139 0.032999 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.594 on 431 degrees of freedom
## Multiple R-squared:  0.7703, Adjusted R-squared:  0.7639
## F-statistic: 120.5 on 12 and 431 DF,  p-value: < 2.2e-16
```

```r
#cat("\nInteractive Model Summary for 2021-22 Season:\n")
#print(summary(interactive_reduce_model_2021_2022))
```

```r
#cat("\nInteractive Model Summary for 2022-23 Season:\n")
#print(summary(interactive_reduce_model_2022_2023))
```

Step 2.2

Create Anonova table to test the new reduced interaction model

From test the new interaction model it is necessary to compare with the previous full interaction model and analyze the p value to make sure that the dropped values were not significant and that the model pass the hypothesis test.

$$H_0 : B_1 = B_{2...} = B_{droppedPredictors} = 0 \quad VS \quad H_a : B_i \neq 0$$

- The null hypothesis test that all the dropped predictors are insignificant (equal to zero).
- The alternative hypothesis test that at least one predictor is significant (different from zero) for this model.

The anova table to compare the models:

```r
anova(reduce_interactive_model_2020_2021,interactive_reduce_model_2020_2021)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ age + player_height + gp + pts + ast + factor(draft_number) +
##     age * pts + age * ast + pts * draft_number
## Model 2: salary ~ (age + player_height + gp + pts + ast + draft_number)^2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    431 9094.6
## 2    422 8684.6  9    410.02 2.2137 0.02035 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results from the annova table we can see that the **p-value** is **0.02035** these smaller than α at **0.05**. Then we reject the null hypothesis and conclude that the dropped predictors were insignificant and accept the reduce interaction model.

the reduce interaction model then would look like this:

$$\widehat{Salary} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{age} + \widehat{\beta}_2 X_{playerHeight} + \widehat{\beta}_3 X_{gp} + \widehat{\beta}_4 X_{pts} + \widehat{\beta}_5 X_{ast} + \widehat{\beta}_6 X_{draftNumber} + \widehat{\beta}_7 X_{age} X_{pts} + \widehat{\beta}_8 X_{age} X_{ast} + \widehat{\beta}_9 X_{draft_numbe}$$

Step 3: Use of Higher order

The next step is to verify if the model can have any of the terms a higher order. To do this the GGally package is used:

The first thing to used this is to reduce the data sets to hold only the variables that we are analyzing: those are: * Salary * Age * playerHeight * gp * pts * draftNumber

The colunms that were removed for the analysis are:

```r
removed_columns <- c("dreb_pct", "usg_pct", "ast_pct", "player_weight", "net_rating", "college", "ts_pc
```

The new data sets would look like that:

```r
suppressPackageStartupMessages(library(dplyr))
```

```r
reduce_player_salaries_2020_21 <- select(player_salaries_2020_21,  -all_of(removed_columns))

reduce_player_salaries_2021_22 <- select(player_salaries_2021_22, -all_of(removed_columns))

reduce_player_salaries_2022_23 <- select(player_salaries_2022_23,  -all_of(removed_columns))
```

Display the new dataset:

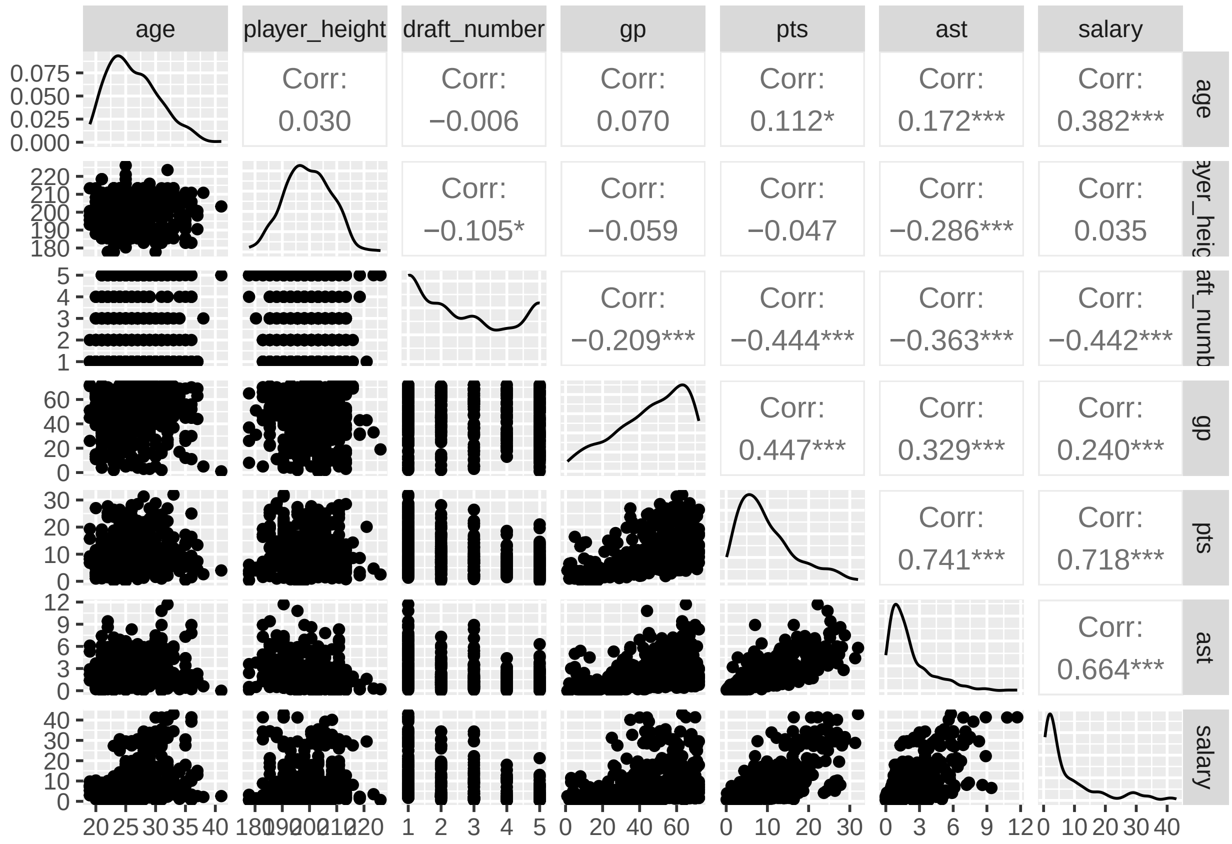
```r
head(reduce_player_salaries_2022_23)
```

```
##     age player_height draft_number gp  pts ast    salary
## 881  23        195.58            2 71 11.3 2.1  2.277000
## 882  30        195.58            3 61  7.6 1.3  5.728393
## 883  23        198.12            1 73 19.6 2.8 10.900634
## 884  23        203.20            2 55  9.2 0.9  2.840160
## 885  20        200.66            3 23  3.3 0.5  2.193960
## 886  34        187.96            3 67  6.2 2.9 13.801614
```

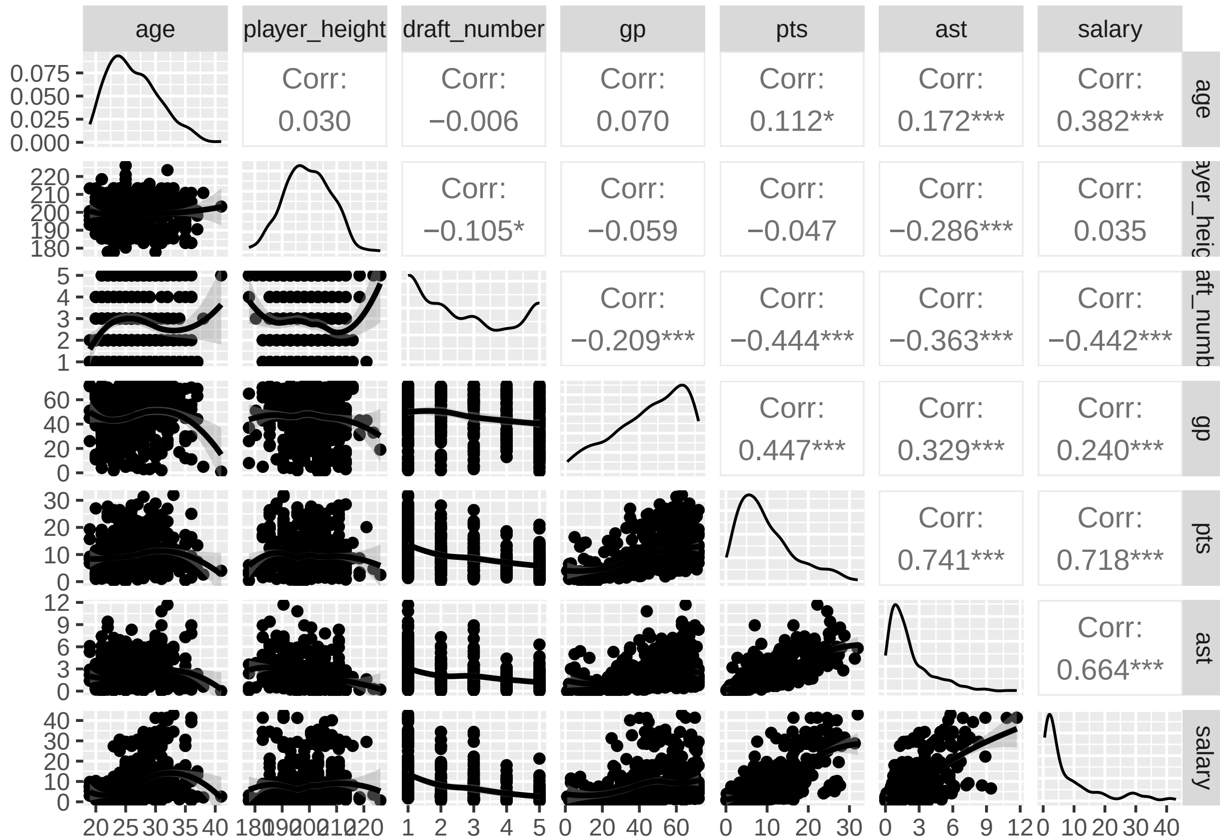
To display the charts to see how the response looks with respect to each independent variable use the command ggpairs():

```r
suppressPackageStartupMessages(library(GGally))
```

```r
ggpairs(reduce_player_salaries_2020_21, progress = FALSE)
```




```r
#ggpairs(reduce_player_salaries_2020_21)
ggpairs(reduce_player_salaries_2020_21, progress = FALSE,lower = list(continuous = "smooth_loess", comb
```



From the graph above, it is possible that the gp, pts and ast variables might have a higher order relationship.

Step 3.1

Add higher Order Relationships

The next step after identifying the possible higher order relationship variables, is to add those relationships and add them to the model.

Lets start with the gp (games play):

```
gp_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + I(gp^2) + pts + ast + draft_number +

summary(gp_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + I(gp^2) + pts +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
##     data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8949  -2.3821  -0.2561   1.7545  21.0113
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.629e+01  6.278e+00  -2.595 0.009787 **
## age              -2.950e-01  9.437e-02  -3.126 0.001890 **
## player_height     1.246e-01  2.836e-02   4.395 1.40e-05 ***
## gp               -5.996e-03  5.324e-02  -0.113 0.910371
## I(gp^2)          -3.256e-04  6.304e-04  -0.517 0.605751
## pts              -9.596e-01  3.461e-01  -2.773 0.005795 **
## ast              -2.770e+00  1.051e+00  -2.635 0.008712 **
## draft_number     -3.281e-01  2.665e-01  -1.231 0.219009
## age:pts           6.839e-02  1.252e-02   5.464 7.86e-08 ***
## age:ast           1.440e-01  3.733e-02   3.859 0.000131 ***
## pts:draft_number -6.988e-02  2.750e-02  -2.541 0.011404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.585 on 433 degrees of freedom
## Multiple R-squared:  0.7701, Adjusted R-squared:  0.7648
## F-statistic: 145.1 on 10 and 433 DF,  p-value: < 2.2e-16
```

From the summary it is evident that the quadratic level is not applicable for the model.

Moving on to the next variable pts (Average points scored per game):

```r
pts_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + pts + I(pts^2) + ast + draft_number

summary(pts_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(pts^2) +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
##     data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6754  -2.2752  -0.3772   1.6150  21.0179
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -15.304665   6.199696  -2.469 0.013949 *
## age               -0.278302   0.094838  -2.935 0.003518 **
## player_height      0.125115   0.028298   4.421 1.24e-05 ***
## gp                -0.026966   0.014211  -1.898 0.058425 .
## pts               -1.154379   0.370427  -3.116 0.001953 **
## I(pts^2)           0.007182   0.005223   1.375 0.169836
## ast               -2.798814   1.048939  -2.668 0.007911 **
## draft_number      -0.513266   0.301062  -1.705 0.088939 .
## age:pts            0.066731   0.012555   5.315 1.71e-07 ***
## age:ast            0.145223   0.037252   3.898 0.000112 ***
## pts:draft_number  -0.049041   0.031655  -1.549 0.122053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.577 on 433 degrees of freedom
```

```
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.7657
## F-statistic: 145.8 on 10 and 433 DF,  p-value: < 2.2e-16
```

From the summary it is evident that the quadratic level is applicable for the model, and the age variable.

Moving on to the next variable ast (Average assists per game.):

```r
ast_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + pts + I(ast^2) + ast + draft_number

summary(ast_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(ast^2) +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
##     data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8483  -2.4616  -0.2514   1.8182  20.9034
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -17.01777    6.35217  -2.679 0.007664 **
## age               -0.31744    0.09510  -3.338 0.000917 ***
## player_height      0.13135    0.02926   4.490 9.16e-06 ***
## gp                -0.03368    0.01369  -2.461 0.014253 *
## pts               -0.95384    0.34583  -2.758 0.006059 **
## I(ast^2)          -0.04028    0.04550  -0.885 0.376434
## ast               -2.66622    1.05830  -2.519 0.012116 *
## draft_number      -0.26525    0.27199  -0.975 0.330001
## age:pts            0.06818    0.01250   5.454 8.28e-08 ***
## age:ast            0.15308    0.03855   3.971 8.38e-05 ***
## pts:draft_number  -0.07607    0.02804  -2.713 0.006930 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.582 on 433 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7651
## F-statistic: 145.3 on 10 and 433 DF,  p-value: < 2.2e-16
```

From the summary it is evident that the quadratic level is not applicable for the model for this variable.

Moving on to the next (demographic) variable age (players age):

```r
age_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + pts + I(age^2) + ast + draft_number

summary(age_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
```

```
##     data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9316  -2.3541  -0.2491   1.7631  20.3936
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -36.27358    9.88424  -3.670 0.000273 ***
## age                 1.22954    0.58548   2.100 0.036301 *
## player_height       0.12559    0.02813   4.464 1.03e-05 ***
## gp                 -0.03452    0.01355  -2.547 0.011211 *
## pts                -0.90155    0.34390  -2.622 0.009061 **
## I(age^2)           -0.02756    0.01040  -2.649 0.008373 **
## ast                -2.92549    1.04415  -2.802 0.005310 **
## draft_number       -0.36985    0.26441  -1.399 0.162592
## age:pts             0.06590    0.01244   5.298 1.87e-07 ***
## age:ast             0.15027    0.03710   4.051 6.05e-05 ***
## pts:draft_number   -0.07344    0.02724  -2.696 0.007287 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.55 on 433 degrees of freedom
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.7684
## F-statistic:   148 on 10 and 433 DF,  p-value: < 2.2e-16
```

From the summary it is evident that the quadratic level is applicable for the model, and the age variable. Therefore the next step is to move on to the cubic level.

```r
age_higer_model_2020_2021_cubic <- lm(salary ~ age + player_height + gp + pts + I(age^2) + I(age^3) + as

summary(age_higer_model_2020_2021_cubic)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(age^2) +
##     I(age^3) + ast + draft_number + age * pts + age * ast + pts *
##     draft_number, data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5694  -2.4356  -0.2683   1.7685  20.4840
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -10.141071  40.360597  -0.251  0.80173
## age                -1.663789   4.371806  -0.381  0.70371
## player_height       0.124895   0.028171   4.433 1.18e-05 ***
## gp                 -0.035816   0.013700  -2.614  0.00925 **
## pts                -0.869797   0.347385  -2.504  0.01265 *
## I(age^2)            0.077329   0.157401   0.491  0.62347
## I(age^3)           -0.001238   0.001854  -0.668  0.50459
## ast                -2.928103   1.044825  -2.802  0.00530 **
```

```
## draft_number      -0.351357   0.266021  -1.321  0.18727
## age:pts            0.064870   0.012540   5.173 3.53e-07 ***
## age:ast            0.149921   0.037123   4.039 6.36e-05 ***
## pts:draft_number  -0.073994   0.027270  -2.713  0.00693 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.553 on 432 degrees of freedom
## Multiple R-squared:  0.7739, Adjusted R-squared:  0.7681
## F-statistic: 134.4 on 11 and 432 DF,  p-value: < 2.2e-16
```

From the summary it is visible that the cubic level does not apply for the **age** variable.

Moving on to the next variable player_height (players height):

```
player_height_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + pts + I(age^2) + I(player_

summary(player_height_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(age^2) +
##     I(player_height^2) + ast + draft_number + age * pts + age *
##     ast + pts * draft_number, data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9022  -2.3666  -0.2597   1.7531  20.3574
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.306e+01  9.355e+01  -0.140  0.88907
## age                  1.227e+00  5.862e-01   2.092  0.03698 *
## player_height       -1.070e-01  9.325e-01  -0.115  0.90868
## gp                  -3.444e-02  1.357e-02  -2.538  0.01150 *
## pts                 -9.031e-01  3.443e-01  -2.623  0.00903 **
## I(age^2)            -2.750e-02  1.042e-02  -2.639  0.00860 **
## I(player_height^2)   5.828e-04  2.335e-03   0.250  0.80304
## ast                 -2.918e+00  1.046e+00  -2.790  0.00550 **
## draft_number        -3.768e-01  2.662e-01  -1.416  0.15757
## age:pts              6.592e-02  1.245e-02   5.294 1.91e-07 ***
## age:ast              1.499e-01  3.717e-02   4.033 6.50e-05 ***
## pts:draft_number    -7.272e-02  2.742e-02  -2.652  0.00829 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.555 on 432 degrees of freedom
## Multiple R-squared:  0.7737, Adjusted R-squared:  0.7679
## F-statistic: 134.2 on 11 and 432 DF,  p-value: < 2.2e-16
```

From the summary it is evident that the quadratic level is not applicable for the model for this variable.

In conclusion, there is only one evident higher level for the variables in the existent model. The predictive model gets updated adding the new predictor $I(age^2)$.

$\widehat{Salary} = \widehat{\beta}_0+\widehat{\beta}_1 X_{age}+\widehat{\beta}_2 X_{playerHeight}+\widehat{\beta}_3 X_{gp}+\widehat{\beta}_4 X_{pts}+\widehat{\beta}_5 X_{ast}+\widehat{\beta}_6 X_{draftNumber}+\widehat{\beta}_7 X_{age} X_{pts}+\widehat{\beta}_8 X_{age} X_{ast}+\widehat{\beta}_9 X_{draft_numbe}$

```r
age_higer_model_2020_2021 <- lm(salary ~ age + player_height + gp + pts + I(age^2) + ast + draft_number

summary(age_higer_model_2020_2021)
```

```
##
## Call:
## lm(formula = salary ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
##     data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9316  -2.3541  -0.2491   1.7631  20.3936
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -36.27358    9.88424  -3.670 0.000273 ***
## age                1.22954    0.58548   2.100 0.036301 *
## player_height      0.12559    0.02813   4.464 1.03e-05 ***
## gp                -0.03452    0.01355  -2.547 0.011211 *
## pts               -0.90155    0.34390  -2.622 0.009061 **
## I(age^2)          -0.02756    0.01040  -2.649 0.008373 **
## ast               -2.92549    1.04415  -2.802 0.005310 **
## draft_number      -0.36985    0.26441  -1.399 0.162592
## age:pts            0.06590    0.01244   5.298 1.87e-07 ***
## age:ast            0.15027    0.03710   4.051 6.05e-05 ***
## pts:draft_number  -0.07344    0.02724  -2.696 0.007287 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.55 on 433 degrees of freedom
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.7684
## F-statistic:   148 on 10 and 433 DF,  p-value: < 2.2e-16
```

Step 4: Regression Diagnostics

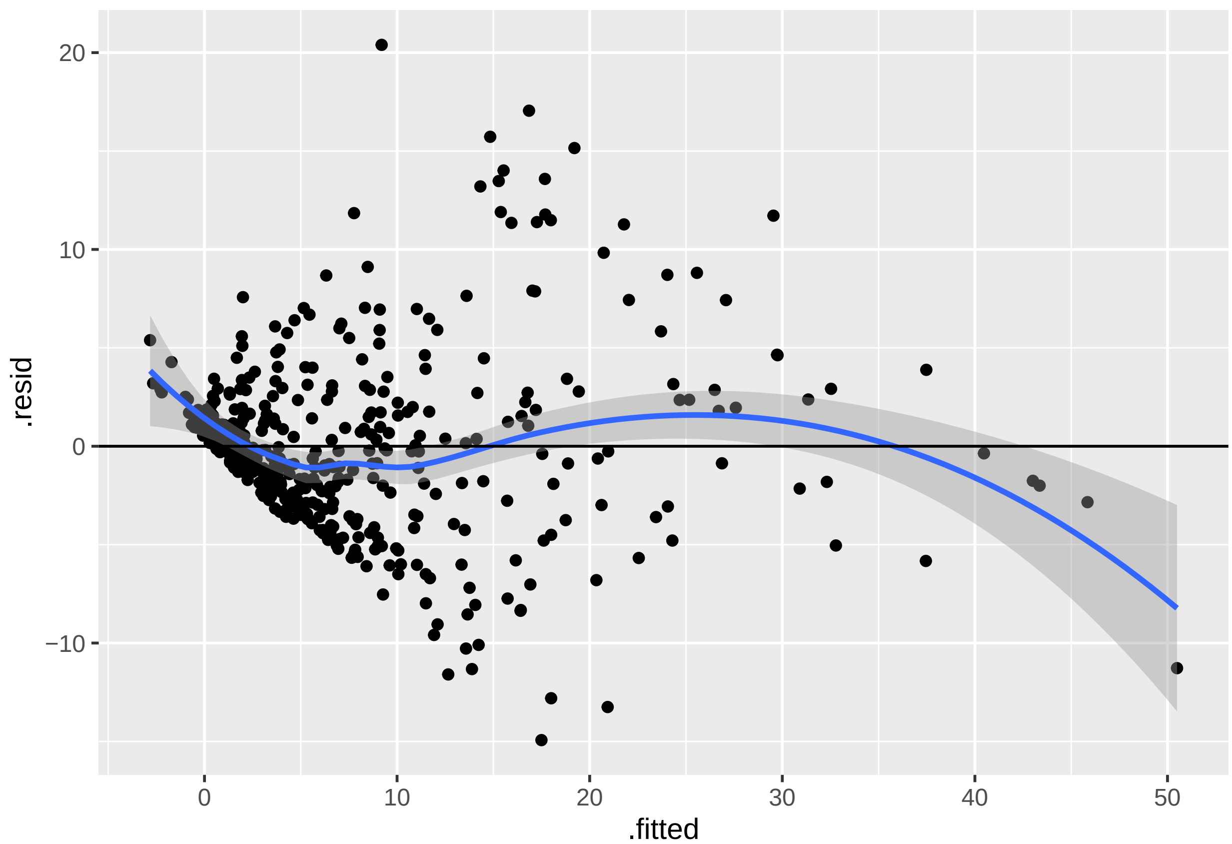
In this section assumptions will be investigate, to confirm that the selected model meets all the assumption and is well justify.

Step 4.1

Linear Assumption The linear regression model assumes that there is a straight-line (linear) relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect and the prediction accuracy of the model can be significantly reduced

```r
ggplot(age_higer_model_2020_2021, aes(x=.fitted, y=.resid)) +
geom_point() +geom_smooth()+
geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The adjuster r squared for the quadratic model for age is 0.7684 indicates the variation in salary that can be explained by this model is 76% with RMSE 4.55.

Step 4.2

Independence Assumption

```
head(reduce_player_salaries_2020_21)
```

```
##   age player_height draft_number gp  pts ast    salary
## 1  22        195.58            3 58 15.3 1.4  1.663861
## 2  26        193.04            2 39  9.9 2.0 19.610714
## 3  26        198.12            5 39  3.1 0.8  2.018458
## 4  26        198.12            5 10  8.4 2.4  3.870370
## 5  35        195.58            5 56  7.6 2.2  4.767000
## 6  25        190.50            5 50  4.8 1.3  0.660750
```

```
#Create the group as cathegorical:

data <- data.frame(draft_number = reduce_player_salaries_2020_21$draft_number)
# Define the cut points and labels
cut_points <- c(-Inf, 1, 2, 3, 4, Inf)
labels <- c("Level 1", "Level 2", "Level 3", "Level 4", "Level 5")

# Convert numeric variable to categorical
```

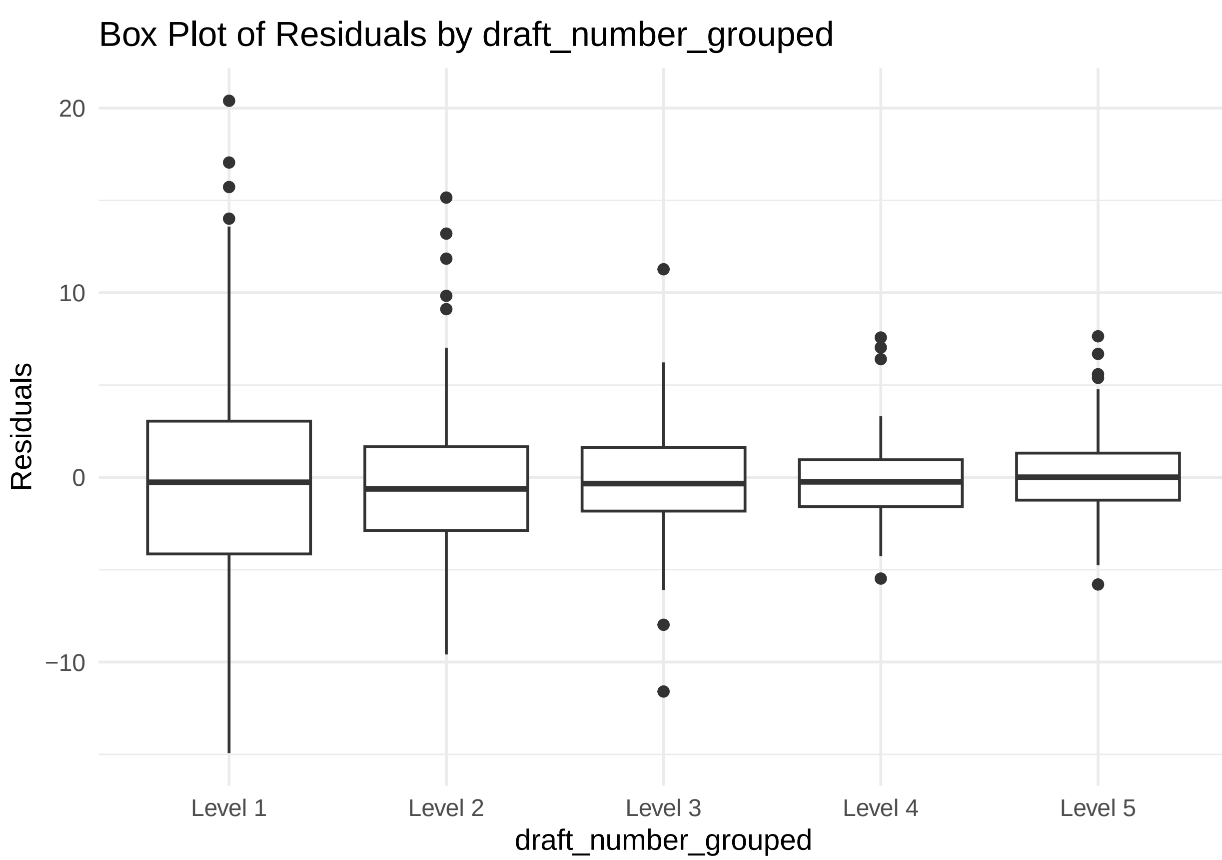
```r
data$categorical_var <- cut(data$draft_number, breaks = cut_points, labels = labels)

# Assign the new categorical variable back to the original data frame
reduce_player_salaries_2020_21$draft_number_grouped <- data$categorical_var
# Print the updated original data frame to verify
print(head(reduce_player_salaries_2020_21))
```

```
##   age player_height draft_number gp  pts ast    salary draft_number_grouped
## 1  22        195.58            3 58 15.3 1.4  1.663861              Level 3
## 2  26        193.04            2 39  9.9 2.0 19.610714              Level 2
## 3  26        198.12            5 39  3.1 0.8  2.018458              Level 5
## 4  26        198.12            5 10  8.4 2.4  3.870370              Level 5
## 5  35        195.58            5 56  7.6 2.2  4.767000              Level 5
## 6  25        190.50            5 50  4.8 1.3  0.660750              Level 5
```

```r
reduce_player_salaries_2020_21$residuals <- residuals(age_higer_model_2020_2021)

# Create box plots of residuals by a categorical variable
ggplot(reduce_player_salaries_2020_21, aes(x = draft_number_grouped, y = residuals)) + geom_boxplot() +
```



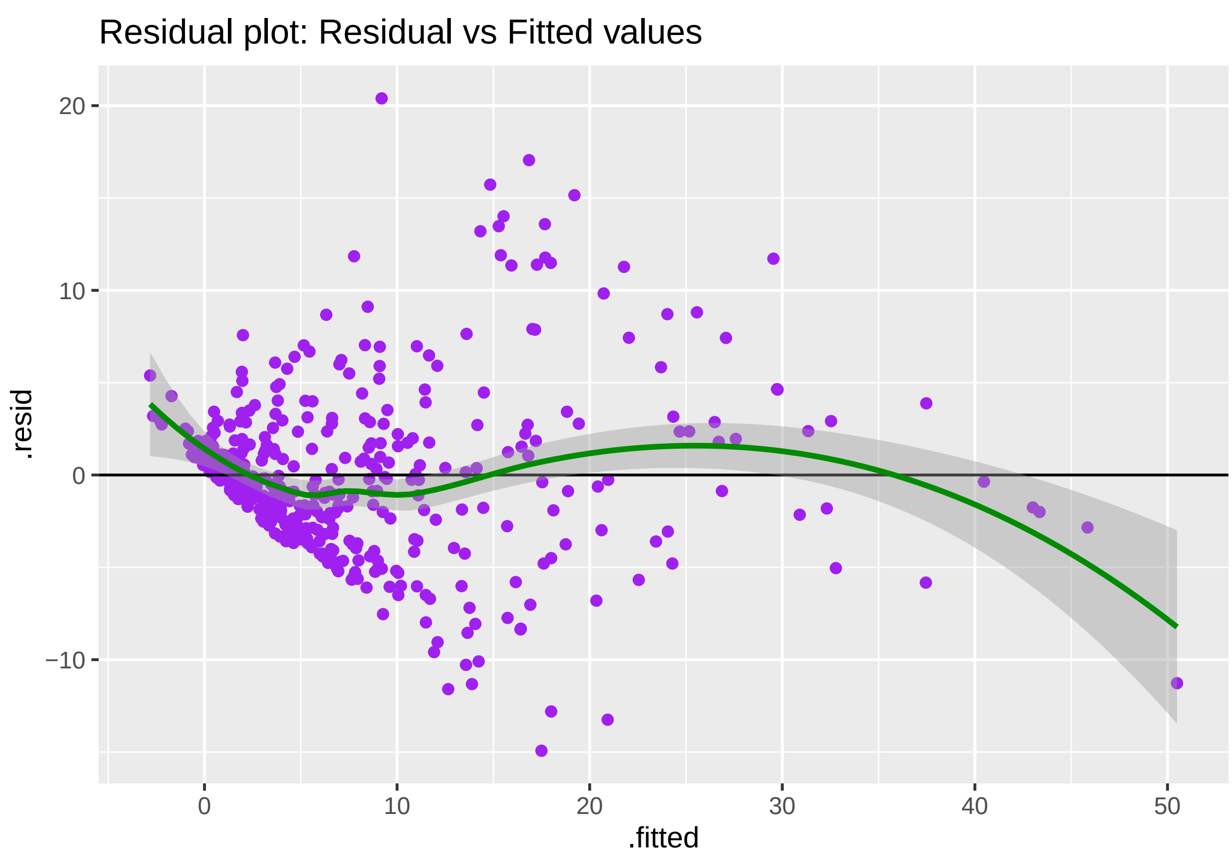
From the picture above we can assume that the Independence Assumption is met.

Step 4.3

Equal Variance Assumption

```r
ggplot(age_higer_model_2020_2021, aes(x=.fitted, y=.resid)) +
geom_point(colour = "purple") +
geom_hline(yintercept = 0) +
geom_smooth(colour = "green4")+
ggtitle("Residual plot: Residual vs Fitted values")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```r
bcmodel_log=lm(log(salary) ~ age + player_height + gp + pts + I(age^2) + ast + draft_number + age*pts +
summary(bcmodel_log)
```

```
##
## Call:
## lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + age * pts + age * ast + pts * draft_number,
##     data = player_salaries_2020_21)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8399 -0.4430  0.0326  0.4398  1.7487
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.445399   1.323980  -4.868 1.58e-06 ***
```

```
## age                0.255741   0.078424   3.261 0.001198 **
## player_height      0.018198   0.003769   4.829 1.91e-06 ***
## gp                 0.001793   0.001815   0.988 0.323756
## pts                0.001589   0.046064   0.034 0.972505
## I(age^2)          -0.003589   0.001394  -2.575 0.010342 *
## ast               -0.078918   0.139862  -0.564 0.572871
## draft_number      -0.339235   0.035417  -9.578  < 2e-16 ***
## age:pts            0.001541   0.001666   0.925 0.355660
## age:ast            0.006860   0.004969   1.381 0.168126
## pts:draft_number   0.012521   0.003649   3.431 0.000658 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6095 on 433 degrees of freedom
## Multiple R-squared:  0.7306, Adjusted R-squared:  0.7244
## F-statistic: 117.4 on 10 and 433 DF,  p-value: < 2.2e-16
```

```
#bptest(bcmodel1)
#summary(bcmodel1)


#+ age*pts + age*ast  este es del normal
#+ + age*pts + age*ast  este es del hemoscedasticity
```

From the previous model, we can drop some of the interactions:

```
bcmodel_log_reduce=lm(log(salary) ~ age + player_height + gp + pts + I(age^2) + ast + draft_number + pts
summary(bcmodel_log_reduce)
```

```
##
## Call:
## lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + pts * draft_number, data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93447 -0.43556  0.06826  0.43396  1.68722
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.305019   1.296418  -5.635 3.15e-08 ***
## age                0.293966   0.077452   3.795 0.000168 ***
## player_height      0.018163   0.003804   4.775 2.45e-06 ***
## gp                 0.001477   0.001829   0.808 0.419613
## pts                0.040078   0.010256   3.908 0.000108 ***
## I(age^2)          -0.003757   0.001402  -2.680 0.007633 **
## ast                0.115392   0.025017   4.613 5.24e-06 ***
## draft_number      -0.339102   0.035768  -9.481  < 2e-16 ***
## pts:draft_number   0.012472   0.003682   3.387 0.000770 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6155 on 435 degrees of freedom
```

```
## Multiple R-squared:  0.724,  Adjusted R-squared:  0.7189
## F-statistic: 142.6 on 8 and 435 DF,  p-value: < 2.2e-16
```

```r
suppressPackageStartupMessages(library(lmtest))
```

```r
bptest(bcmodel_log_reduce)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bcmodel_log_reduce
## BP = 17.133, df = 8, p-value = 0.02876
```

This number is smaller that α at **0.05**

Step 4.4

Normality Assumption

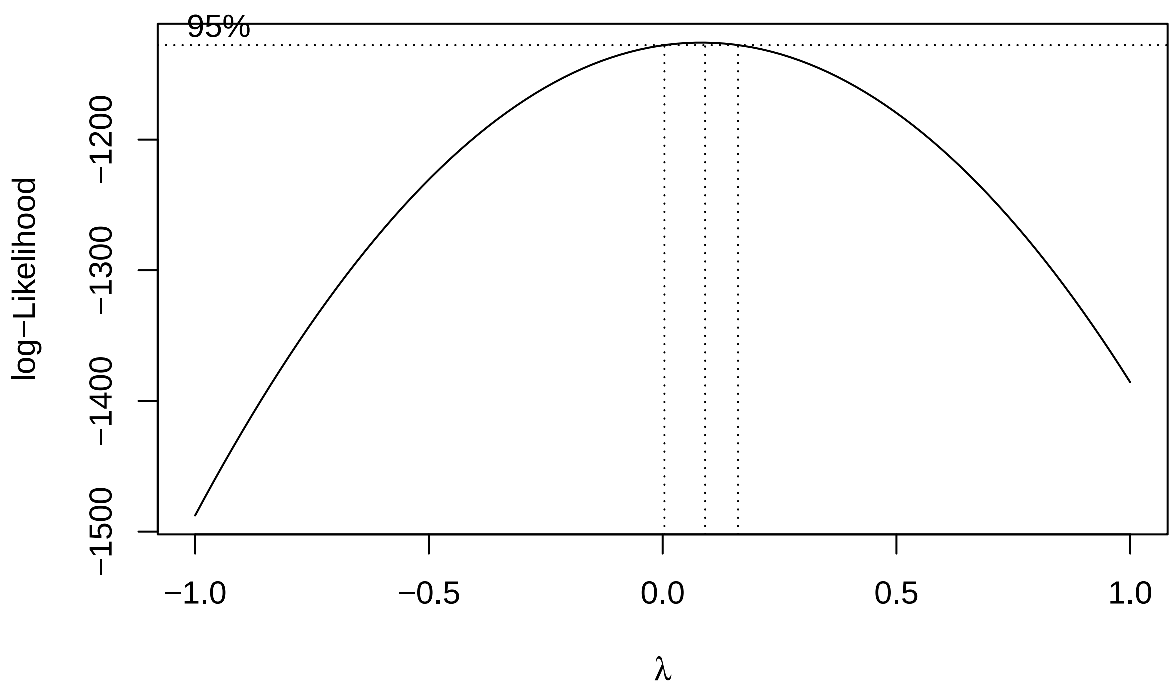
```r
library(MASS) #for the boxcox()function
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:olsrr':
##
##     cement
```

```r
bc=boxcox(age_higer_model_2020_2021,lambda=seq(-1,1))
```




```
#extract best lambda
bestlambda=bc$x[which(bc$y==max(bc$y))]
bestlambda
```

```
## [1] 0.09090909
```

From the command above we got that the best lamba is in the range of -0.1 to 0.1

Then choosing the values 0 and 0.09

```
bcmodel1=lm(log(salary) ~ age + player_height + gp + pts + I(age^2) + ast + draft_number  + pts*draft_nu
summary(bcmodel1)
```

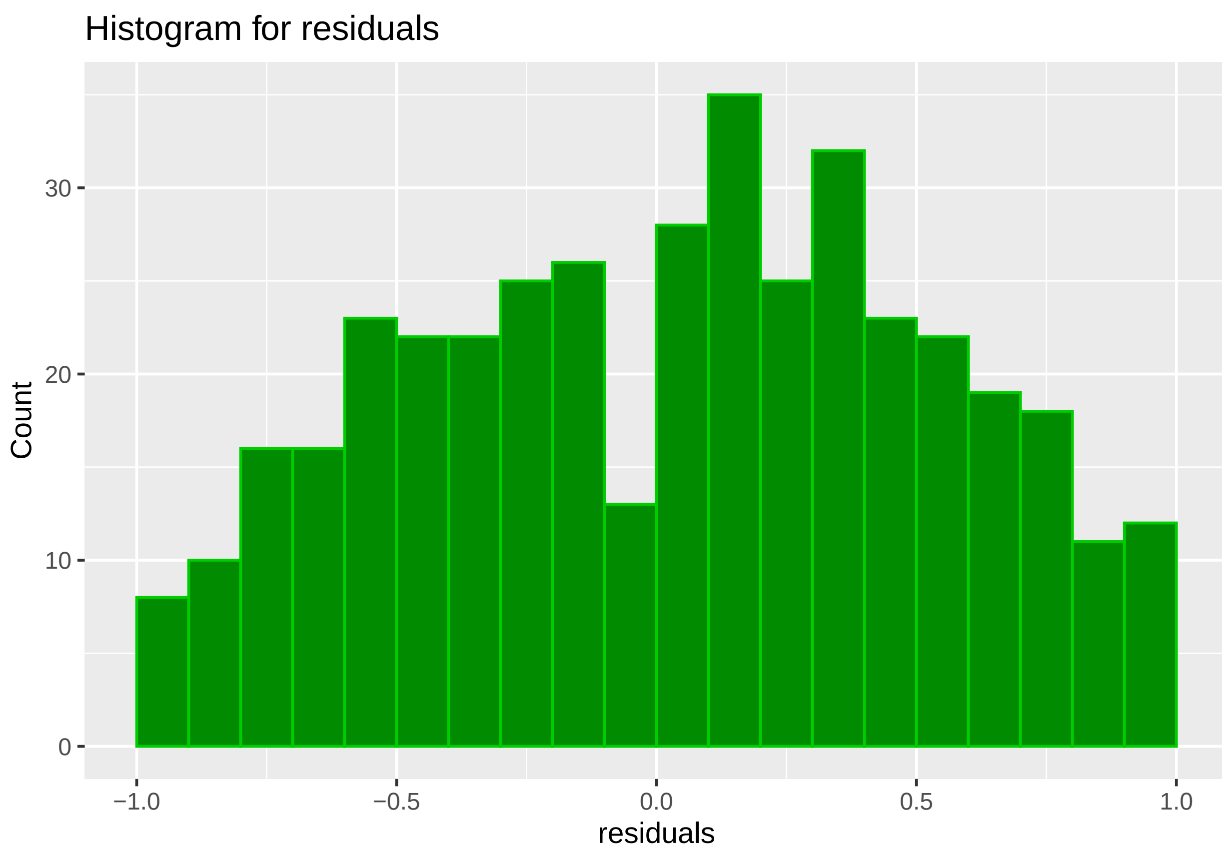
```
##
## Call:
## lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + pts * draft_number, data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93447 -0.43556  0.06826  0.43396  1.68722
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.305019   1.296418  -5.635 3.15e-08 ***
```

```
## age                 0.293966   0.077452   3.795 0.000168 ***
## player_height       0.018163   0.003804   4.775 2.45e-06 ***
## gp                  0.001477   0.001829   0.808 0.419613
## pts                 0.040078   0.010256   3.908 0.000108 ***
## I(age^2)           -0.003757   0.001402  -2.680 0.007633 **
## ast                 0.115392   0.025017   4.613 5.24e-06 ***
## draft_number       -0.339102   0.035768  -9.481  < 2e-16 ***
## pts:draft_number    0.012472   0.003682   3.387 0.000770 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6155 on 435 degrees of freedom
## Multiple R-squared:  0.724,  Adjusted R-squared:  0.7189
## F-statistic: 142.6 on 8 and 435 DF,  p-value: < 2.2e-16
```

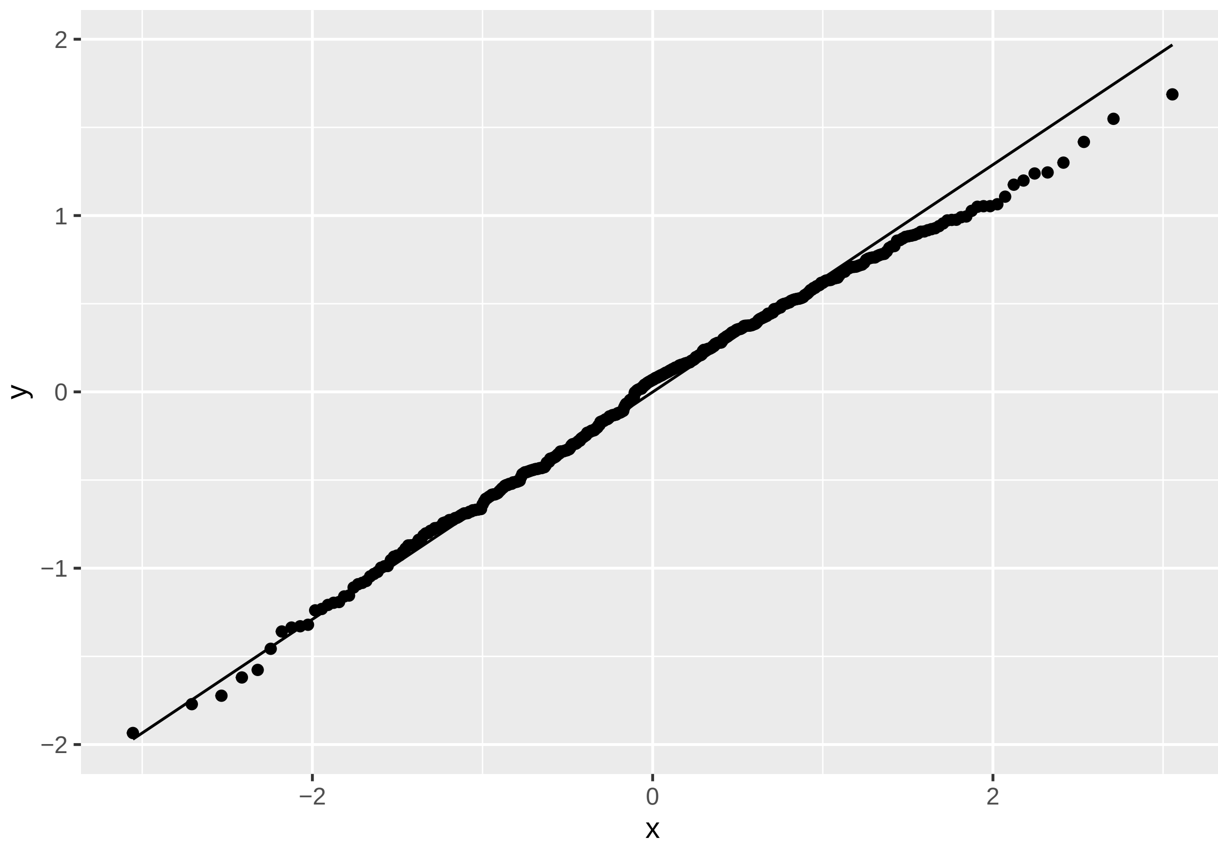
```
shapiro.test(residuals((bcmodel1)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((bcmodel1))
## W = 0.994, p-value = 0.07742
```

```
ggplot(data=player_salaries_2020_21, aes(residuals(bcmodel1))) +
geom_histogram(breaks = seq(-1,1,by=0.1), col="green3", fill="green4") +
labs(title="Histogram for residuals") +
labs(x="residuals", y="Count")
```




```r
#normal QQ plot
ggplot(data=player_salaries_2020_21, aes(sample=bcmodel_log_reduce$residuals)) +
stat_qq() +
stat_qq_line()
```




```r
#optional histogram
par(mfrow=c(1,2))
```

The outputs show that the residual data have normal distribution (from histogram and Q-Q plot). Shapiro-Wilk normality test also confirms that the residuals are normally distributed as the p- value=0.07742 >0.05.We fail to reject the null hypothesis that we have normality.

Step 4.5

Multicollinearity Assumption

During this analysis, it is expected to have a Multicollinearity, due to the interactions between variables as $I(age^2)$ that would be strong related to the age predictor. However this interaction is being ignored if is shown in the analysis.

```r
#install.packages("mctest")
library(mctest)
imcdiag(bcmodel_log_reduce, method="VIF")
```

```
##
## Call:
## imcdiag(mod = bcmodel_log_reduce, method = "VIF")
##
##
```

```
##  VIF Multicollinearity Diagnostics
##
##                        VIF detection
## age              124.2594         1
## player_height      1.2187         0
## gp                 1.3576         0
## pts                5.4319         0
## I(age^2)         123.8949         1
## ast                2.7178         0
## draft_number       3.5577         0
## pts:draft_number   3.8737         0
##
## Multicollinearity may be due to age I(age^2) regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

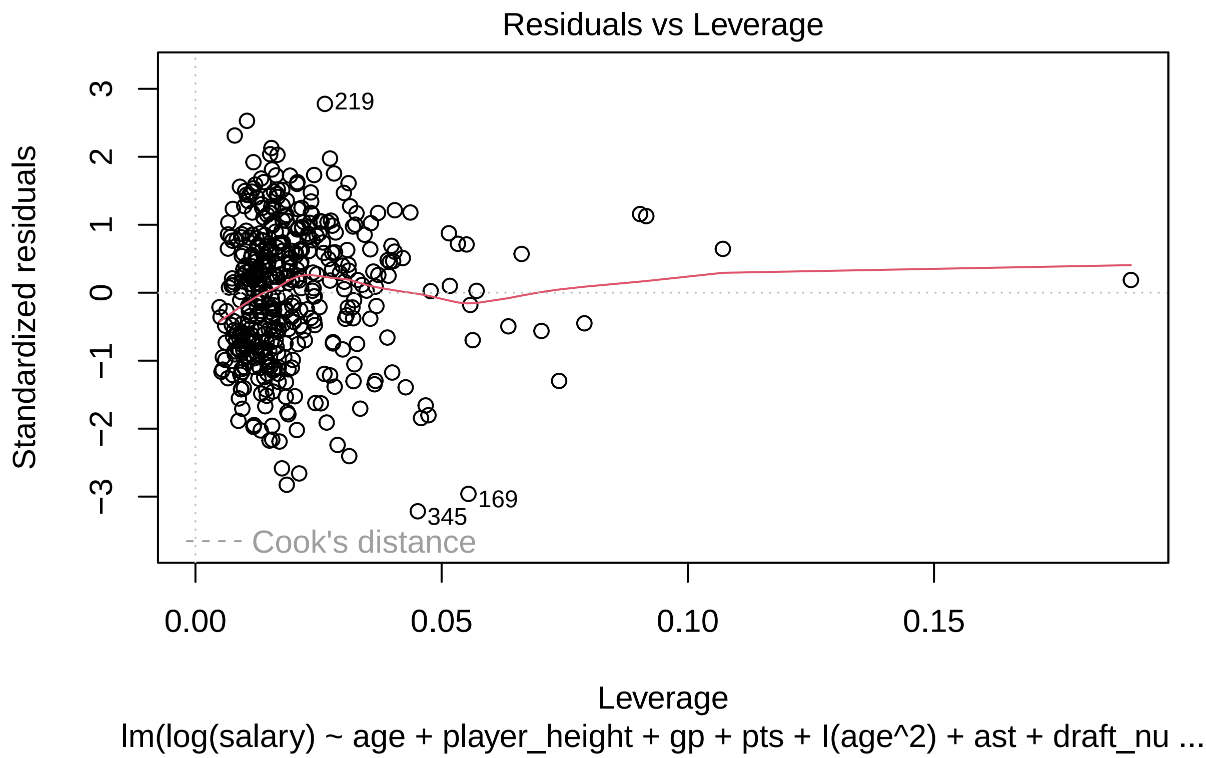
As assumed the Multicollinearity between the variables age and $I(age^2)$ is present. This is to be ignore, That being said the model meets the Multicollinearity Assumption.

Step 4.6

Outliers

In this section we are analyzing the outliers if present and if they are influential in the behaviour of the data and the model:

```r
plot(bcmodel_log_reduce,which=5)
```

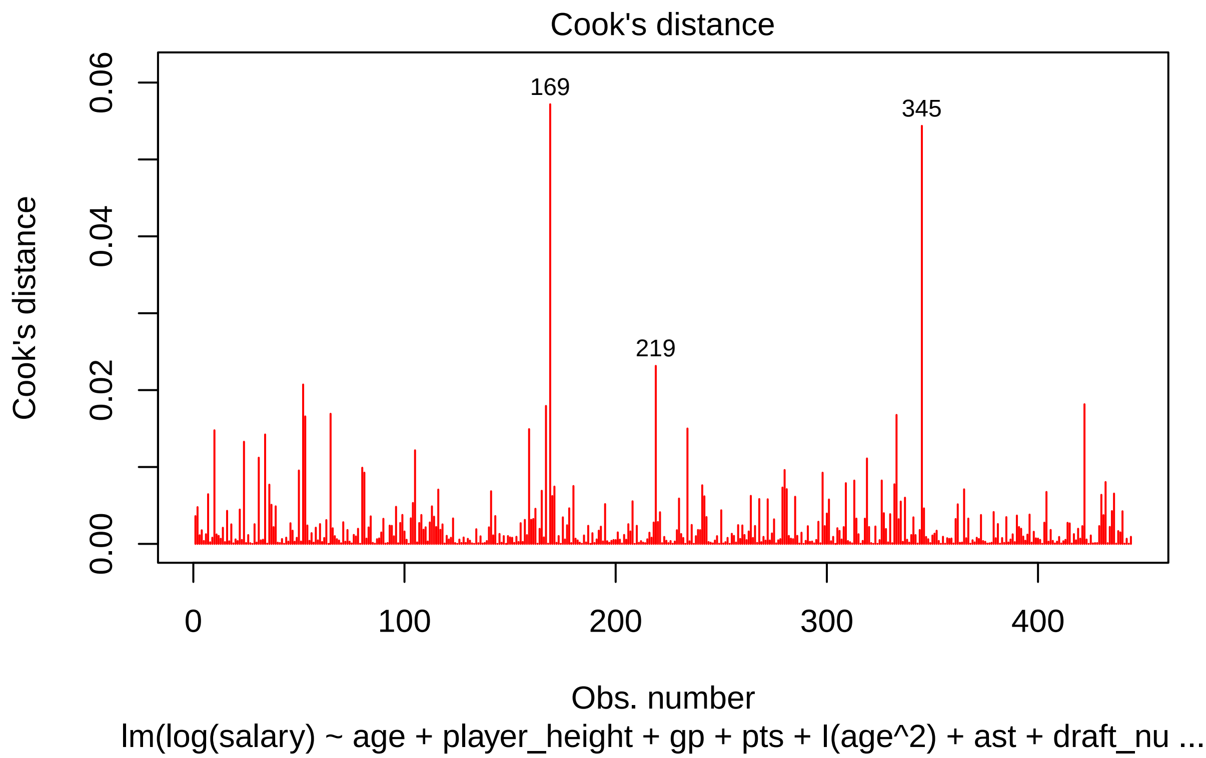


From the graphic, it is not evident any bad behavior, therefore a deeper analysis is required:

```
player_salaries_2020_21[cooks.distance(bcmodel_log_reduce)>0.5,]
```

```
##  [1] player_name        team_abbreviation age                player_height
##  [5] player_weight      college            country            draft_number
##  [9] gp                 pts                reb                ast
## [13] net_rating         oreb_pct           dreb_pct           usg_pct
## [17] ts_pct             ast_pct            season             salary
## <0 rows> (or 0-length row.names)
```

```
plot(bcmodel_log_reduce,pch=18,col="red",which=c(4))
```



From this analysis it is noticeable that there are no strange behavior from outliers they are under the cook distance 0.06 and that is lower to our set 0.05 value. Form this analysis we can conclude that there is no necessary to study the leverage points.

Conclusion:

The final model for the Regression analysis is:

```
summary(bcmodel_log_reduce)
```

```
##
## Call:
## lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
##     ast + draft_number + pts * draft_number, data = player_salaries_2020_21)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93447 -0.43556  0.06826  0.43396  1.68722
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.305019   1.296418  -5.635 3.15e-08 ***
## age               0.293966   0.077452   3.795 0.000168 ***
## player_height     0.018163   0.003804   4.775 2.45e-06 ***
```

```
## gp                 0.001477   0.001829   0.808 0.419613
## pts                0.040078   0.010256   3.908 0.000108 ***
## I(age^2)          -0.003757   0.001402  -2.680 0.007633 **
## ast                0.115392   0.025017   4.613 5.24e-06 ***
## draft_number      -0.339102   0.035768  -9.481  < 2e-16 ***
## pts:draft_number  0.012472   0.003682   3.387 0.000770 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6155 on 435 degrees of freedom
## Multiple R-squared:  0.724,  Adjusted R-squared:  0.7189
## F-statistic: 142.6 on 8 and 435 DF,  p-value: < 2.2e-16
```

Construct our final selected model:

$$\widehat{Salary} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{age} + \widehat{\beta}_2 X_{playerHeight} + \widehat{\beta}_3 X_{gp} + \widehat{\beta}_4 X_{pts} + \widehat{\beta}_5 X_{ast} + \widehat{\beta}_6 X_{I(age^2)} + \widehat{\beta}_7 X_{draftNumber} + \widehat{\beta}_8 X_{draftNumber4} * X_{pts}$$

This variables explain 71.89% of the data and the model .

Intercept: -7.305

Estimate: This is the expected salary when all predictor variables are zero

age: 0.294

Estimate: For each one-year increase in age, the salary increases by approximately 0.294 units, holding other factors constant.

player_height: 0.018

Estimate: For each centimeter increase in height, the salary increases by 0.018 units.

gp (games played): 0.001

Estimate: For each additional game played, the salary increases by 0.001 units.

pts (points per game): 0.040

Estimate: For each additional point scored per game, the salary increases by 0.040 units.

I(age^2): -0.004

Estimate: The squared age term is included to capture any non-linear relationship between age and salary. Here, a negative coefficient suggests diminishing returns of age on salary at higher ages.

ast (assists per game): 0.115

Estimate: For each additional assist per game, the salary increases by 0.115 units.

draft_number: -0.339

Estimate: For each increase in draft number (i.e., a worse draft position), the salary decreases by 0.339 units.

pts:draft_number (interaction term): 0.012

Estimate: This indicates the combined effect of points per game and draft number on salary. For each unit increase in the product of points per game and draft number, the salary increases by 0.012 units.