Final Project Data Analysis and Linear Model

By Muhammad Areeb and Elizabeth Grace Hiebert

Data Analysis:

The dataset contains basketball NBA statistics for players that have a a higher EFF value. The dataset contains information such as their in-game stats, points, rebounds, assists, blocks, field goal percentage, etc. There are two categorical variables, one of which is a response/target variable for our model, `CHANGED_FRANCHISE` and the other one is `SEASON_ID`.

Trends: Basic Statistical Overview:

```
df <- read.csv('./train.csv')%>% mutate(CHANGED_FRANCHISE = factor(CHANGED_FRANCHISE))
head(df, 10)
```

```
##    X PLAYER_ID RANK            PLAYER    TEAM_ID TEAM GP  MIN FGM  FGA FG_PCT
## 1  0       764    1    David Robinson 1610612759  SAN 82 3019 711 1378  0.516
## 2  1       893    2    Michael Jordan 1610612741  CHI 82 3090 916 1850  0.495
## 3  2       252    3       Karl Malone 1610612762  UTH 82 3113 789 1520  0.519
## 4  3       165    4   Hakeem Olajuwon 1610612745  HOU 72 2797 768 1494  0.514
## 5  4       255    5        Grant Hill 1610612765  DET 80 3260 564 1221  0.462
## 6  5       787    6   Charles Barkley 1610612756  PHX 71 2632 580 1160  0.500
## 7  6       358    7 Anfernee Hardaway 1610612753  ORL 82 3015 623 1215  0.513
## 8  7       431    8        Shawn Kemp 1610612760  SEA 79 2631 526  937  0.561
## 9  8       913    9     Larry Johnson 1610612766  CHH 81 3274 583 1225  0.476
## 10 9       304   10     John Stockton 1610612762  UTH 82 2915 440  818  0.538
##    FG3M FG3A FG3_PCT FTM FTA FT_PCT OREB DREB  REB AST STL BLK TOV  PF  PTS
## 1     3    9   0.333 626 823  0.761  319  681 1000 247 111 271 190 262 2051
## 2   111  260   0.427 548 657  0.834  148  395  543 352 180  42 197 195 2491
## 3    16   40   0.400 512 708  0.723  175  629  804 345 138  56 199 245 2106
## 4     3   14   0.214 397 548  0.724  176  608  784 257 113 207 247 242 1936
## 5     5   26   0.192 485 646  0.751  127  656  783 548 100  48 263 242 1618
## 6    49  175   0.280 440 566  0.777  243  578  821 262 114  56 218 208 1649
## 7    89  283   0.314 445 580  0.767  129  225  354 582 166  41 229 160 1780
## 8     5   12   0.417 493 664  0.742  276  628  904 173  93 127 315 299 1550
## 9    67  183   0.366 427 564  0.757  249  434  683 355  55  43 182 173 1660
## 10   95  225   0.422 234 282  0.830   54  172  226 916 140  15 246 207 1209
##     EFF AST_TOV STL_TOV CHANGED_FRANCHISE SEASON_ID
## 1  2626    1.30    0.58             False   1995-96
## 2  2368    1.79    0.91             False   1995-96
## 3  2323    1.73    0.69             False   1995-96
## 4  2173    1.04    0.46             False   1995-96
## 5  2016    2.08    0.38             False   1995-96
## 6  1978    1.20    0.52             False   1995-96
## 7  1967    2.54    0.72             False   1995-96
```

```
## 8  1950    0.55    0.29             False   1995-96
## 9  1835    1.95    0.30             False   1995-96
## 10 1834    3.72    0.57             False   1995-96
```

```
print(summary(df))
```

```
##        X           PLAYER_ID            RANK        PLAYER
##  Min.   :    0   Min.   :      2   Min.   :  1   Length:13829
##  1st Qu.: 3457   1st Qu.:   1594   1st Qu.:119   Class :character
##  Median : 6914   Median : 101129   Median :239   Mode  :character
##  Mean   : 6914   Mean   : 378186   Mean   :241
##  3rd Qu.:10371   3rd Qu.: 203473   3rd Qu.:358
##  Max.   :13828   Max.   :1642013   Max.   :603
##     TEAM_ID              TEAM                GP              MIN
##  Min.   :1.611e+09   Length:13829       Min.   : 1.00   Min.   :   1
##  1st Qu.:1.611e+09   Class :character   1st Qu.:31.00   1st Qu.: 372
##  Median :1.611e+09   Mode  :character   Median :57.00   Median :1102
##  Mean   :1.611e+09                      Mean   :51.12   Mean   :1198
##  3rd Qu.:1.611e+09                      3rd Qu.:73.00   3rd Qu.:1905
##  Max.   :1.611e+09                      Max.   :85.00   Max.   :3485
##       FGM              FGA             FG_PCT            FG3M
##  Min.   :  0.0   Min.   :   0.0   Min.   :0.0000   Min.   :  0.00
##  1st Qu.: 43.0   1st Qu.: 101.0   1st Qu.:0.3990   1st Qu.:  0.00
##  Median :143.0   Median : 320.0   Median :0.4390   Median : 13.00
##  Mean   :187.2   Mean   : 411.1   Mean   :0.4367   Mean   : 37.97
##  3rd Qu.:286.0   3rd Qu.: 628.0   3rd Qu.:0.4830   3rd Qu.: 62.00
##  Max.   :978.0   Max.   :2173.0   Max.   :1.0000   Max.   :402.00
##       FG3A           FG3_PCT            FTM              FTA
##  Min.   :   0.0   Min.   :0.0000   Min.   :  0.00   Min.   :  0.0
##  1st Qu.:   3.0   1st Qu.:0.0000   1st Qu.: 16.00   1st Qu.: 24.0
##  Median :  43.0   Median :0.3090   Median : 55.00   Median : 76.0
##  Mean   : 106.2   Mean   :0.2487   Mean   : 90.29   Mean   :119.2
##  3rd Qu.: 176.0   3rd Qu.:0.3680   3rd Qu.:127.00   3rd Qu.:170.0
##  Max.   :1028.0   Max.   :1.0000   Max.   :756.00   Max.   :972.0
##      FT_PCT            OREB             DREB            REB
##  Min.   :0.0000   Min.   :  0.00   Min.   :  0.0   Min.   :   0
##  1st Qu.:0.6470   1st Qu.: 12.00   1st Qu.: 42.0   1st Qu.:  57
##  Median :0.7500   Median : 35.00   Median :123.0   Median : 163
##  Mean   :0.6999   Mean   : 55.66   Mean   :155.4   Mean   : 211
##  3rd Qu.:0.8180   3rd Qu.: 77.00   3rd Qu.:226.0   3rd Qu.: 305
##  Max.   :1.0000   Max.   :443.00   Max.   :894.0   Max.   :1247
##       AST              STL              BLK              TOV
##  Min.   :  0.0   Min.   :  0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 19.0   1st Qu.: 10   1st Qu.:  4.00   1st Qu.: 18.00
##  Median : 66.0   Median : 30   Median : 13.00   Median : 53.00
##  Mean   :111.3   Mean   : 38   Mean   : 24.35   Mean   : 69.01
##  3rd Qu.:152.0   3rd Qu.: 57   3rd Qu.: 31.00   3rd Qu.:103.00
##  Max.   :935.0   Max.   :231   Max.   :332.00   Max.   :464.00
##        PF              PTS              EFF             AST_TOV
##  Min.   :  0.0   Min.   :   0.0   Min.   :  -8.0   Min.   : 0.000
##  1st Qu.: 39.0   1st Qu.: 113.0   1st Qu.: 135.0   1st Qu.: 0.820
##  Median :100.0   Median : 376.0   Median : 453.0   Median : 1.330
##  Mean   :104.3   Mean   : 502.6   Mean   : 565.5   Mean   : 1.476
##  3rd Qu.:159.0   3rd Qu.: 768.0   3rd Qu.: 869.0   3rd Qu.: 2.000
```

```
##  Max.   :371.0   Max.   :2832.0   Max.   :3039.0   Max.   :21.000
##     STL_TOV       CHANGED_FRANCHISE  SEASON_ID
##  Min.   :0.0000   False:9539        Length:13829
##  1st Qu.:0.3700   True :4290        Class :character
##  Median :0.5300                     Mode  :character
##  Mean   :0.6088
##  3rd Qu.:0.7500
##  Max.   :7.0000
```

It can be observed from the code above, that there are two categorical variables: `CHANGED_FRANCHISE` and `SEASON_ID`

Correlation Analysis:

```r
cor_matrix <- df %>%
  select(where(is.numeric)) %>%
  cor(use = "complete.obs")
print(cor_matrix)
```

```
##                        X   PLAYER_ID         RANK       TEAM_ID           GP
## X          1.0000000000  0.69758068  0.188089646  0.0008072137 -0.127965017
## PLAYER_ID  0.6975806789  1.00000000  0.239181730  0.0215986827 -0.181465233
## RANK       0.1880896464  0.23918173  1.000000000  0.0045755525 -0.851062299
## TEAM_ID    0.0008072137  0.02159868  0.004575553  1.0000000000  0.006424455
## GP        -0.1279650173 -0.18146523 -0.851062299  0.0064244546  1.000000000
## MIN       -0.1262424516 -0.17695208 -0.939386069 -0.0015658663  0.860512759
## FGM       -0.0539410040 -0.11554195 -0.891443223 -0.0049534671  0.725021118
## FGA       -0.0684890272 -0.12612815 -0.879182688 -0.0046799720  0.731005091
## FG_PCT     0.0505689123  0.01092808 -0.367647969  0.0001148484  0.305932995
## FG3M       0.1717371509  0.06023054 -0.518874879 -0.0258053187  0.483023061
## FG3A       0.1783718551  0.06735812 -0.531988171 -0.0256319845  0.498144872
## FG3_PCT    0.1678313503  0.11388656 -0.209747391  0.0031076195  0.240353001
## FTM       -0.1186178116 -0.15276703 -0.783217597  0.0031229577  0.595410070
## FTA       -0.1358204296 -0.16256860 -0.796019205  0.0045738777  0.609977864
## FT_PCT     0.0074393483 -0.05256895 -0.393150309  0.0047170228  0.396849677
## OREB      -0.1553248198 -0.15187326 -0.683920147 -0.0010021762  0.582553869
## DREB      -0.0469579716 -0.12134592 -0.848612461 -0.0088076964  0.705640652
## REB       -0.0821521636 -0.13514393 -0.828171865 -0.0067105810  0.692928971
## AST       -0.0432926079 -0.09827119 -0.688746979 -0.0047070247  0.560885680
## STL       -0.1257618042 -0.15142420 -0.804528172 -0.0016970160  0.717438097
## BLK       -0.0803930385 -0.09781998 -0.568029861  0.0049296410  0.468990682
## TOV       -0.1441344572 -0.17442657 -0.853739053 -0.0046089680  0.708803020
## PF        -0.1978707507 -0.20652752 -0.868152599 -0.0008470234  0.857025992
## PTS       -0.0470588046 -0.11173605 -0.881312483 -0.0057282847  0.714221741
## EFF       -0.0402785069 -0.11369899 -0.925318112 -0.0072267092  0.750343608
## AST_TOV    0.1598303283  0.10045472 -0.155599597 -0.0011475117  0.165300616
## STL_TOV    0.0859808763  0.08254657  0.015601172  0.0080966682  0.060231191
##                     MIN          FGM          FGA        FG_PCT         FG3M
## X          -0.126242452 -0.053941004 -0.068489027  0.0505689123  0.171737151
## PLAYER_ID -0.176952077 -0.115541952 -0.126128145  0.0109280829  0.060230536
## RANK       -0.939386069 -0.891443223 -0.879182688 -0.3676479686 -0.518874879
## TEAM_ID    -0.001565866 -0.004953467 -0.004679972  0.0001148484 -0.025805319
## GP          0.860512759  0.725021118  0.731005091  0.3059329947  0.483023061
## MIN         1.000000000  0.922759354  0.928241051  0.2619212620  0.600240538
```

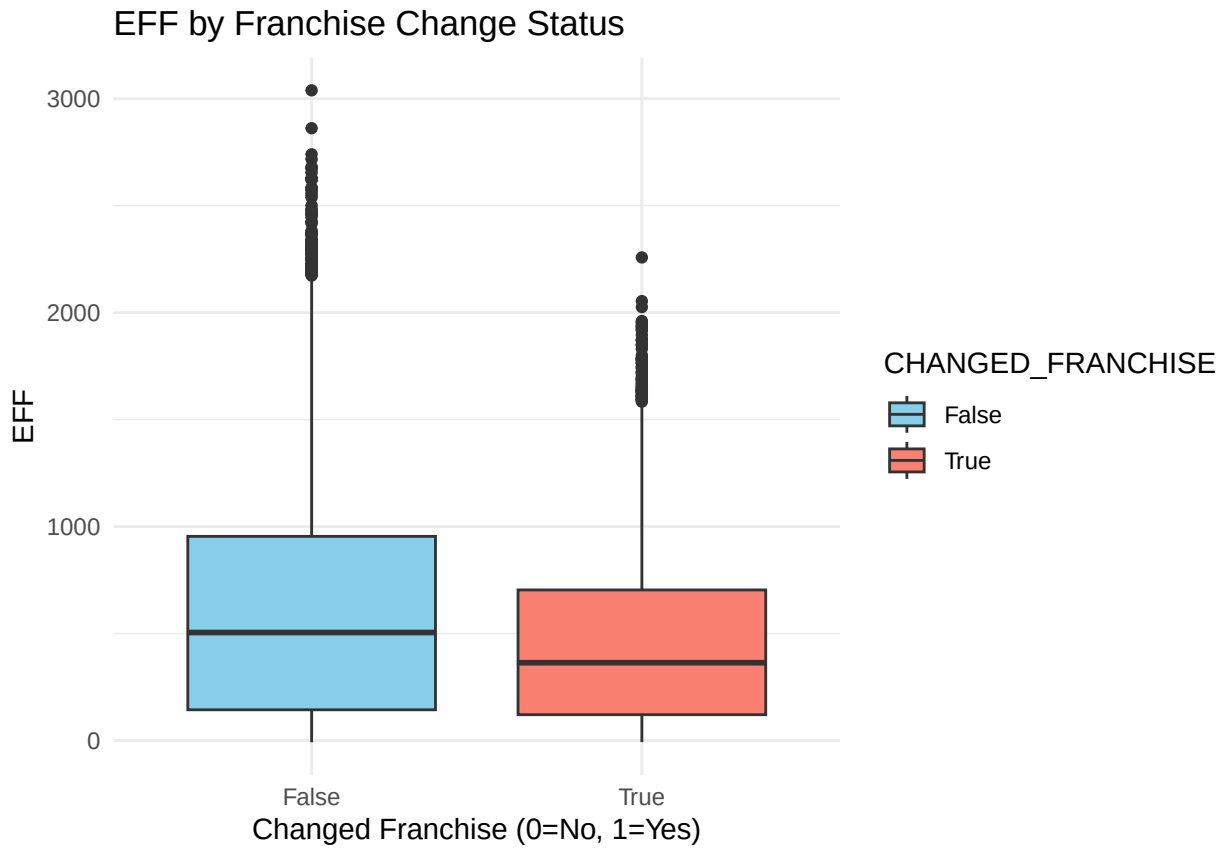
```
## FGM        0.922759354  1.000000000  0.988779093  0.2699827082  0.599261029
## FGA        0.928241051  0.988779093  1.000000000  0.2052189500  0.657262296
## FG_PCT     0.261921262  0.269982708  0.205218950  1.0000000000 -0.001889547
## FG3M       0.600240538  0.599261029  0.657262296 -0.0018895471  1.000000000
## FG3A       0.617178100  0.615853495  0.679630690 -0.0156888286  0.991505798
## FG3_PCT    0.262199215  0.253334333  0.290558782  0.0080153964  0.501796708
## FTM        0.803485176  0.891036455  0.882012097  0.2187249815  0.450741405
## FTA        0.809292355  0.889169721  0.871354138  0.2494614152  0.395463983
## FT_PCT     0.362515557  0.336525489  0.359294122  0.1631717565  0.354374099
## OREB       0.623451868  0.578558183  0.514681499  0.3905539224 -0.026553784
## DREB       0.825575909  0.803725350  0.762075394  0.3454793971  0.317306555
## REB        0.792733795  0.762973328  0.712954685  0.3718265110  0.222381606
## AST        0.736393909  0.723587381  0.749041082  0.1032383305  0.561862173
## STL        0.859038891  0.789423512  0.809356490  0.1750122064  0.554529677
## BLK        0.511563730  0.485658516  0.427424225  0.3318046704  0.007902937
## TOV        0.885247948  0.901467292  0.907195699  0.2156481422  0.509633432
## PF         0.872434179  0.765794698  0.752559111  0.3284779732  0.365642013
## PTS        0.917497434  0.993230238  0.989289990  0.2456726999  0.646111954
## EFF        0.929454217  0.956197643  0.927348452  0.3295870244  0.525293243
## AST_TOV    0.191082496  0.138082348  0.167186121 -0.0566261003  0.290810327
## STL_TOV    0.002043001 -0.075229466 -0.064331401  0.0086239699  0.065332789
##                  FG3A      FG3_PCT          FTM          FTA       FT_PCT
## X          0.17837186  0.167831350 -0.118617812 -0.135820430  0.007439348
## PLAYER_ID  0.06735812  0.113886558 -0.152767027 -0.162568599 -0.052568952
## RANK      -0.53198817 -0.209747391 -0.783217597 -0.796019205 -0.393150309
## TEAM_ID   -0.02563198  0.003107619  0.003122958  0.004573878  0.004717023
## GP         0.49814487  0.240353001  0.595410070  0.609977864  0.396849677
## MIN        0.61717810  0.262199215  0.803485176  0.809292355  0.362515557
## FGM        0.61585350  0.253334333  0.891036455  0.889169721  0.336525489
## FGA        0.67963069  0.290558782  0.882012097  0.871354138  0.359294122
## FG_PCT    -0.01568883  0.008015396  0.218724982  0.249461415  0.163171757
## FG3M       0.99150580  0.501796708  0.450741405  0.395463983  0.354374099
## FG3A       1.00000000  0.494462570  0.473366659  0.419747759  0.359809840
## FG3_PCT    0.49446257  1.000000000  0.170371192  0.127050848  0.331884430
## FTM        0.47336666  0.170371192  1.000000000  0.987178184  0.307386073
## FTA        0.41974776  0.127050848  0.987178184  1.000000000  0.260056969
## FT_PCT     0.35980984  0.331884430  0.307386073  0.260056969  1.000000000
## OREB      -0.01817250 -0.156603805  0.524809210  0.597194474  0.075350690
## DREB       0.33125217  0.064091149  0.713887753  0.756147860  0.217162033
## REB        0.23510765 -0.001846324  0.681067423  0.734160356  0.181114092
## AST        0.58802015  0.303921305  0.672048334  0.647442988  0.309236852
## STL        0.58107882  0.277805680  0.706390421  0.705370275  0.311077115
## BLK        0.01367898 -0.128494008  0.442706604  0.508509829  0.047366992
## TOV        0.53802630  0.214324558  0.867384229  0.872096527  0.312314610
## PF         0.37921808  0.101792692  0.670362709  0.702517917  0.285673567
## PTS        0.66232579  0.277239057  0.921348021  0.911189845  0.352396411
## EFF        0.53848776  0.197593676  0.877270338  0.886745226  0.313339217
## AST_TOV    0.29949463  0.345670529  0.080102557  0.048552018  0.265486079
## STL_TOV    0.06623744  0.140395599 -0.116499164 -0.120730281  0.101045393
##                 OREB         DREB          REB          AST          STL
## X          -0.155324820 -0.046957972 -0.082152164 -0.043292608 -0.125761804
## PLAYER_ID  -0.151873260 -0.121345921 -0.135143932 -0.098271193 -0.151424200
## RANK       -0.683920147 -0.848612461 -0.828171865 -0.688746979 -0.804528172
## TEAM_ID    -0.001002176 -0.008807696 -0.006710581 -0.004707025 -0.001697016
```

```
## GP          0.582553869  0.705640652  0.692928971  0.560885680  0.717438097
## MIN         0.623451868  0.825575909  0.792733795  0.736393909  0.859038891
## FGM         0.578558183  0.803725350  0.762973328  0.723587381  0.789423512
## FGA         0.514681499  0.762075394  0.712954685  0.749041082  0.809356490
## FG_PCT      0.390553922  0.345479397  0.371826511  0.103238331  0.175012206
## FG3M       -0.026553784  0.317306555  0.222381606  0.561862173  0.554529677
## FG3A       -0.018172498  0.331252174  0.235107651  0.588020146  0.581078823
## FG3_PCT    -0.156603805  0.064091149 -0.001846324  0.303921305  0.277805680
## FTM         0.524809210  0.713887753  0.681067423  0.672048334  0.706390421
## FTA         0.597194474  0.756147860  0.734160356  0.647442988  0.705370275
## FT_PCT      0.075350690  0.217162033  0.181114092  0.309236852  0.311077115
## OREB        1.000000000  0.837270141  0.917662243  0.206039272  0.440744270
## DREB        0.837270141  1.000000000  0.985604169  0.484378476  0.645257012
## REB         0.917662243  0.985604169  1.000000000  0.415714262  0.605199099
## AST         0.206039272  0.484378476  0.415714262  1.000000000  0.777677146
## STL         0.440744270  0.645257012  0.605199099  0.777677146  1.000000000
## BLK         0.753435042  0.723554590  0.758784417  0.145495244  0.345128376
## TOV         0.530761481  0.744082454  0.704850896  0.837485519  0.815423224
## PF          0.742686765  0.813925928  0.821135417  0.518242220  0.718796069
## PTS         0.536373181  0.780038568  0.732715980  0.738546655  0.793533740
## EFF         0.697376053  0.903039586  0.871885608  0.739809715  0.807619912
## AST_TOV    -0.174557811 -0.001908639 -0.055360949  0.456518981  0.272933602
## STL_TOV    -0.061375992 -0.056532965 -0.060061158 -0.043124695  0.179048249
##                     BLK          TOV           PF          PTS          EFF
## X          -0.080393038 -0.144134457 -0.1978707507 -0.047058805 -0.040278507
## PLAYER_ID  -0.097819981 -0.174426567 -0.2065275244 -0.111736046 -0.113698992
## RANK       -0.568029861 -0.853739053 -0.8681525994 -0.881312483 -0.925318112
## TEAM_ID     0.004929641 -0.004608968 -0.0008470234 -0.005728285 -0.007226709
## GP          0.468990682  0.708803020  0.8570259917  0.714221741  0.750343608
## MIN         0.511563730  0.885247948  0.8724341791  0.917497434  0.929454217
## FGM         0.485658516  0.901467292  0.7657946981  0.993230238  0.956197643
## FGA         0.427424225  0.907195699  0.7525591108  0.989289990  0.927348452
## FG_PCT      0.331804670  0.215648142  0.3284779732  0.245672700  0.329587024
## FG3M        0.007902937  0.509633432  0.3656420127  0.646111954  0.525293243
## FG3A        0.013678980  0.538026302  0.3792180812  0.662325790  0.538487764
## FG3_PCT    -0.128494008  0.214324558  0.1017926918  0.277239057  0.197593676
## FTM         0.442706604  0.867384229  0.6703627085  0.921348021  0.877270338
## FTA         0.508509829  0.872096527  0.7025179167  0.911189845  0.886745226
## FT_PCT      0.047366992  0.312314610  0.2856735670  0.352396411  0.313339217
## OREB        0.753435042  0.530761481  0.7426867649  0.536373181  0.697376053
## DREB        0.723554590  0.744082454  0.8139259283  0.780038568  0.903039586
## REB         0.758784417  0.704850896  0.8211354166  0.732715980  0.871885608
## AST         0.145495244  0.837485519  0.5182422200  0.738546655  0.739809715
## STL         0.345128376  0.815423224  0.7187960692  0.793533740  0.807619912
## BLK         1.000000000  0.435903416  0.6248471940  0.453984718  0.602875331
## TOV         0.435903416  1.000000000  0.7694513881  0.906173450  0.893995836
## PF          0.624847194  0.769451388  1.0000000000  0.747899868  0.812549620
## PTS         0.453984718  0.906173450  0.7478998682  1.000000000  0.950128653
## EFF         0.602875331  0.893995836  0.8125496198  0.950128653  1.000000000
## AST_TOV    -0.161081192  0.152325629  0.0160242690  0.150173222  0.156083892
## STL_TOV    -0.056113583 -0.149420392 -0.0259122088 -0.073677927 -0.047509393
##                 AST_TOV      STL_TOV
## X           0.159830328  0.085980876
## PLAYER_ID   0.100454722  0.082546572
```

```
## RANK      -0.155599597  0.015601172
## TEAM_ID   -0.001147512  0.008096668
## GP         0.165300616  0.060231191
## MIN        0.191082496  0.002043001
## FGM        0.138082348 -0.075229466
## FGA        0.167186121 -0.064331401
## FG_PCT    -0.056626100  0.008623970
## FG3M       0.290810327  0.065332789
## FG3A       0.299494634  0.066237441
## FG3_PCT    0.345670529  0.140395599
## FTM        0.080102557 -0.116499164
## FTA        0.048552018 -0.120730281
## FT_PCT     0.265486079  0.101045393
## OREB      -0.174557811 -0.061375992
## DREB      -0.001908639 -0.056532965
## REB       -0.055360949 -0.060061158
## AST        0.456518981 -0.043124695
## STL        0.272933602  0.179048249
## BLK       -0.161081192 -0.056113583
## TOV        0.152325629 -0.149420392
## PF         0.016024269 -0.025912209
## PTS        0.150173222 -0.073677927
## EFF        0.156083892 -0.047509393
## AST_TOV    1.000000000  0.416675373
## STL_TOV    0.416675373  1.000000000
```

Looking at the large blob above, there seems to be a very high correlation (often times ≈ 0.9) among variables that represent similar or related stats such as `c('FGM', 'FG_PCT', 'FG3M')` that represent different 'Field Goal' metrics.

```
box_plot <- ggplot(df, aes(x = CHANGED_FRANCHISE, y = EFF, fill = CHANGED_FRANCHISE)) +
  geom_boxplot() +
  scale_fill_manual(values = c("skyblue", "salmon")) +
  labs(title = "EFF by Franchise Change Status",
       x = "Changed Franchise (0=No, 1=Yes)",
       y = "EFF") +
  theme_minimal()
print(box_plot)
```



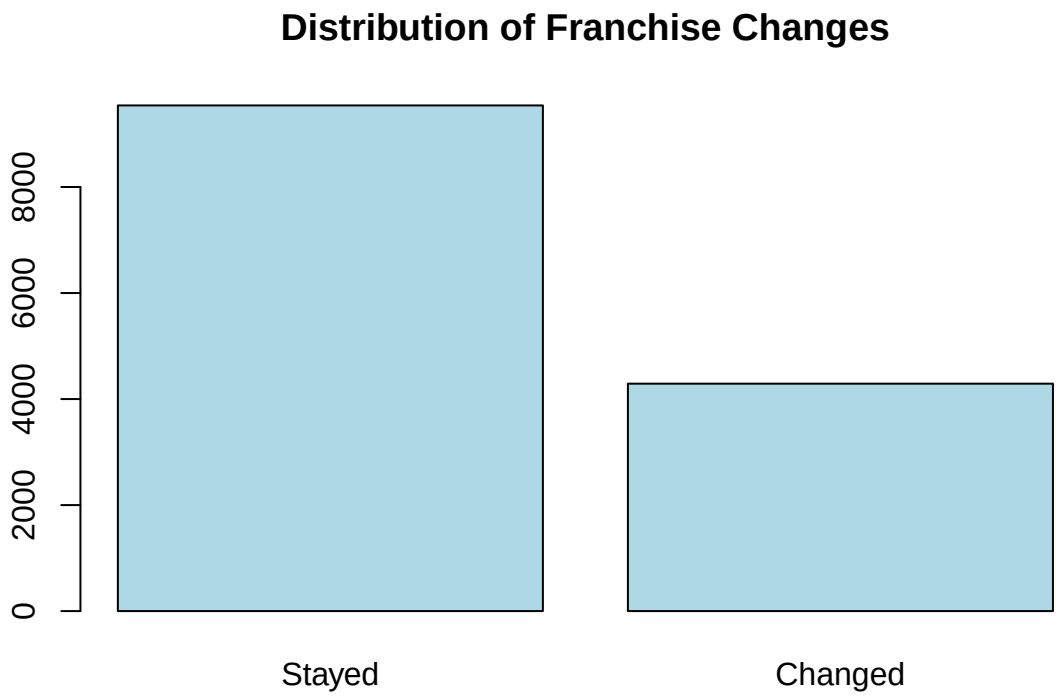
Visualization:

There are statistically significant outliers as the EFF increases for both Categories (`CHANGED_FRANCHISE`: True or False)

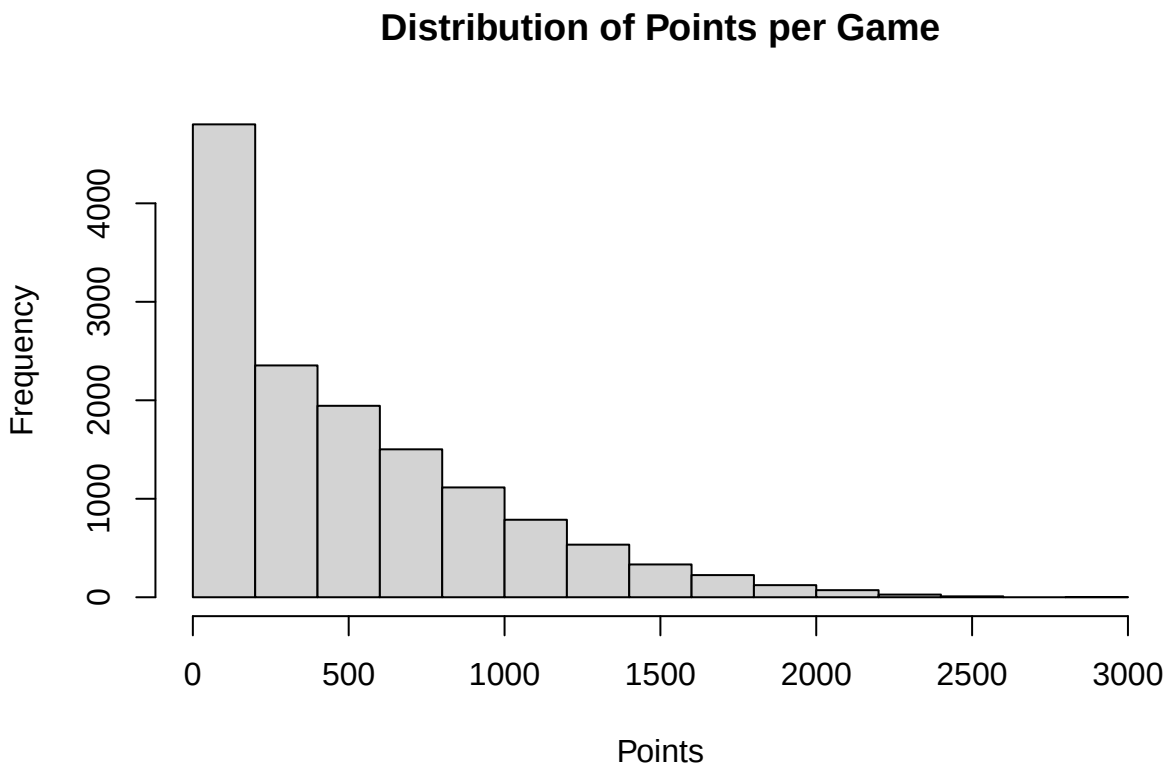
```
table(df$CHANGED_FRANCHISE)
```

```
##
## False  True
##  9539  4290
```

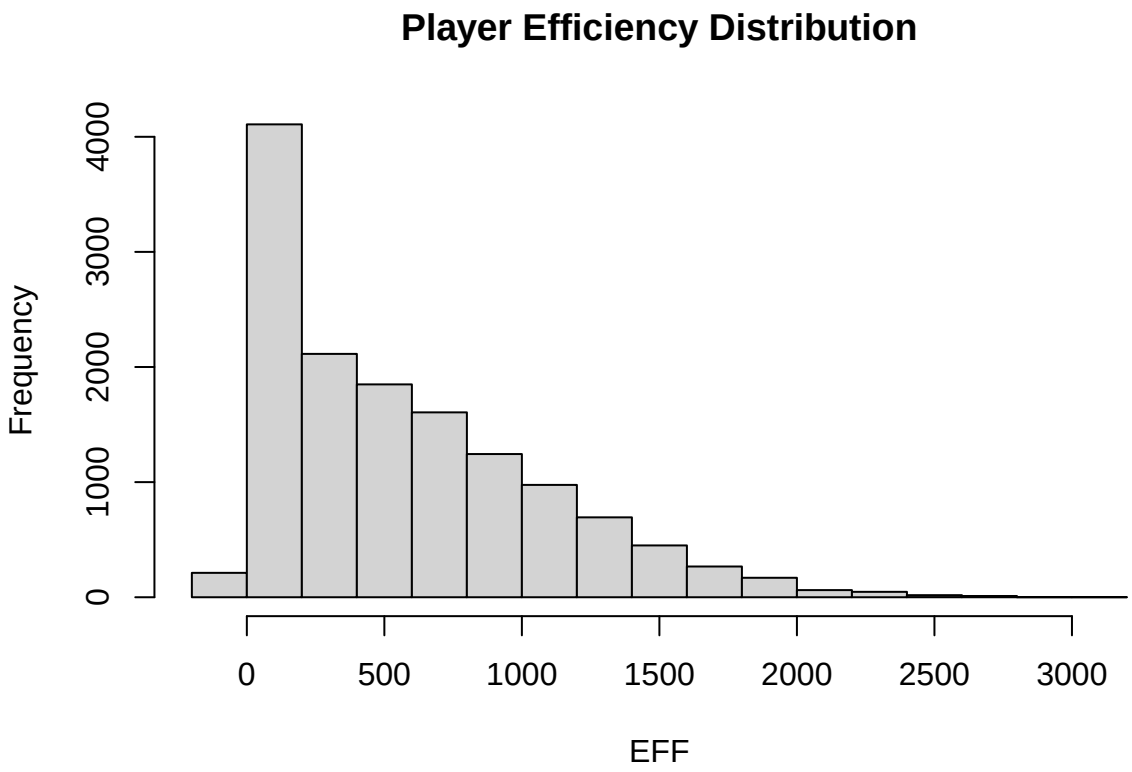
```
barplot(table(df$CHANGED_FRANCHISE),
        names.arg = c("Stayed", "Changed"), col = "lightblue",
        main = "Distribution of Franchise Changes")
```




```r
hist(df$PTS, main = "Distribution of Points per Game", xlab = "Points")
```




```r
hist(df$EFF, main = "Player Efficiency Distribution", xlab = "EFF")
```




```
numeric_cols <- c("EFF")
z_scores <- df %>%
  select(all_of(numeric_cols)) %>%
  scale() %>%
  abs()

outliers <- z_scores > 3
df_outliers <- df %>%
  filter(rowSums(outliers) > 0) %>%
  select(PLAYER, all_of(numeric_cols))

print(df_outliers)
```

Understanding Outliers:

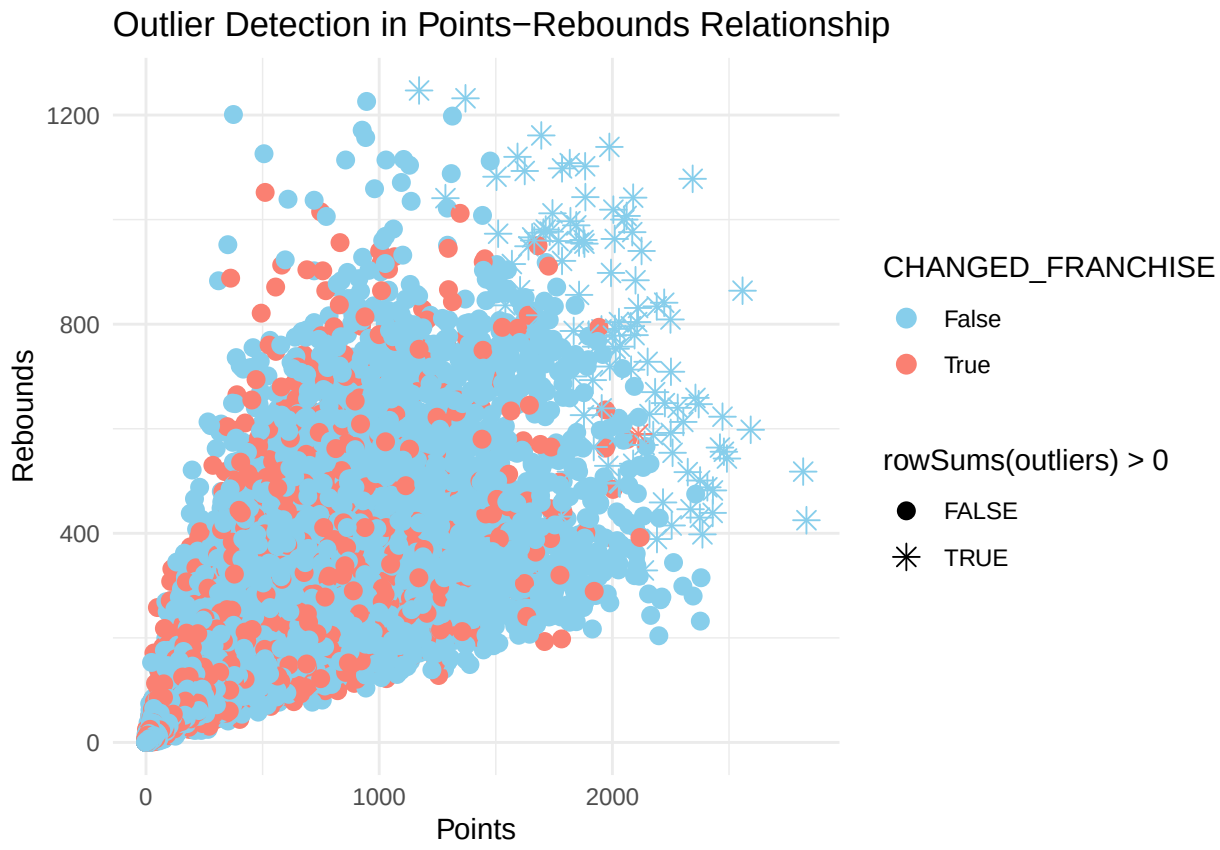
```
##                    PLAYER  EFF
## 1           David Robinson 2626
## 2           Michael Jordan 2368
## 3              Karl Malone 2323
## 4         Hakeem Olajuwon 2173
## 5              Karl Malone 2478
## 6           Michael Jordan 2215
## 7               Grant Hill 2130
```

```
## 8              Karl Malone 2370
## 9               Tim Duncan 2170
## 10       Shaquille O'Neal 2672
## 11          Kevin Garnett 2330
## 12             Karl Malone 2223
## 13             Gary Payton 2159
## 14            Chris Webber 2093
## 15       Shaquille O'Neal 2293
## 16          Kevin Garnett 2251
## 17             Tim Duncan 2125
## 18             Tim Duncan 2558
## 19          Kevin Garnett 2279
## 20          Kevin Garnett 2630
## 21             Tim Duncan 2425
## 22            Kobe Bryant 2298
## 23          Dirk Nowitzki 2218
## 24          Tracy McGrady 2160
## 25          Kevin Garnett 2717
## 26          Kevin Garnett 2621
## 27           LeBron James 2259
## 28          Dirk Nowitzki 2194
## 29       Amar'e Stoudemire 2141
## 30           Shawn Marion 2073
## 31           Shawn Marion 2337
## 32           LeBron James 2323
## 33          Kevin Garnett 2303
## 34            Elton Brand 2253
## 35            Kobe Bryant 2226
## 36          Dirk Nowitzki 2218
## 37          Kevin Garnett 2217
## 38            Kobe Bryant 2129
## 39          Dirk Nowitzki 2098
## 40           LeBron James 2275
## 41       Amar'e Stoudemire 2247
## 42              Chris Paul 2231
## 43          Dwight Howard 2194
## 44            Kobe Bryant 2181
## 45           LeBron James 2501
## 46              Chris Paul 2376
## 47            Dwyane Wade 2314
## 48          Dwight Howard 2133
## 49           LeBron James 2464
## 50           Kevin Durant 2293
## 51              David Lee 2186
## 52          Dwight Howard 2101
## 53           LeBron James 2258
## 54          Dwight Howard 2208
## 55          Blake Griffin 2102
## 56              Pau Gasol 2083
## 57           Kevin Durant 2462
## 58           LeBron James 2446
## 59           Kevin Durant 2572
## 60             Kevin Love 2328
## 61           LeBron James 2255
```

```
## 62            Blake Griffin 2082
## 63             James Harden 2202
## 64               Chris Paul 2125
## 65            Stephen Curry 2073
## 66            Stephen Curry 2424
## 67        Russell Westbrook 2283
## 68             James Harden 2281
## 69             Kevin Durant 2149
## 70             LeBron James 2092
## 71        Russell Westbrook 2740
## 72             James Harden 2623
## 73       Karl-Anthony Towns 2485
## 74            Anthony Davis 2336
## 75             LeBron James 2291
## 76    Giannis Antetokounmpo 2270
## 77             LeBron James 2681
## 78            Anthony Davis 2476
## 79       Karl-Anthony Towns 2384
## 80        Russell Westbrook 2359
## 81    Giannis Antetokounmpo 2306
## 82             James Harden 2170
## 83           Andre Drummond 2125
## 84             James Harden 2581
## 85    Giannis Antetokounmpo 2538
## 86       Karl-Anthony Towns 2338
## 87             Nikola Jokić 2311
## 88           Nikola Vučević 2242
## 89              Rudy Gobert 2183
## 90             Kevin Durant 2178
## 91           Andre Drummond 2153
## 92        Russell Westbrook 2118
## 93              Paul George 2110
## 94             James Harden 2220
## 95    Giannis Antetokounmpo 2180
## 96             Nikola Jokić 2585
## 97             Nikola Jokić 2862
## 98    Giannis Antetokounmpo 2343
## 99              Joel Embiid 2304
## 100      Karl-Anthony Towns 2088
## 101            Nikola Jokić 2622
## 102       Domantas Sabonis 2456
## 103             Joel Embiid 2369
## 104            Luka Dončić 2214
## 105           Jayson Tatum 2209
## 106 Shai Gilgeous-Alexander 2073
## 107            Nikola Jokić 3039
## 108       Domantas Sabonis 2679
## 109   Giannis Antetokounmpo 2655
## 110            Luka Dončić 2580
## 111           Anthony Davis 2548
## 112 Shai Gilgeous-Alexander 2416
## 113            LeBron James 2126
## 114            Kevin Durant 2075
```

Looking at the list of outliers based on **EFF** values, it can be said that the list boils down to some of the more popular household names in basketball.

```
pts_iqr <- df %>%
  summarise(
    Q1 = quantile(PTS, 0.25, na.rm = TRUE),
    Q3 = quantile(PTS, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1
  )

pts_outliers <- df %>%
  filter(PTS < (pts_iqr$Q1 - 1.5 * pts_iqr$IQR) |
           PTS > (pts_iqr$Q3 + 1.5 * pts_iqr$IQR))


outlier_plot <- ggplot(df, aes(x = PTS, y = REB)) +
  geom_point(aes(color = CHANGED_FRANCHISE, shape = rowSums(outliers) > 0), size = 3) +
  scale_color_manual(values = c("skyblue", "salmon")) +
  scale_shape_manual(values = c(16, 8)) +
  labs(title = "Outlier Detection in Points-Rebounds Relationship",
       x = "Points",
       y = "Rebounds") +
  theme_minimal()
print(outlier_plot)
```



Model Creation:

Logistic Regression Model:

The model aims to predict `CHANGED_FRANCHISE`, a categorical variable of boolean type based on all of the other relavent parameters.

We decided to use a logistic regression because our response variable is binomial, a player changes franchise or does not change franchise.

```
# Selecting desired columns.
cols_list <- c('RANK', 'TEAM_ID', 'GP', 'MIN', 'FGM', 'FG_PCT', 'FG3M', 'FG3A',
               'FG3_PCT', 'FTM', 'FTA', 'FT_PCT', 'OREB', 'DREB', 'REB', 'AST',
               'STL', 'BLK', 'TOV', 'PF', 'PTS', 'EFF', 'AST_TOV', 'STL_TOV',
               'CHANGED_FRANCHISE', 'SEASON_ID')
df <- df[, cols_list]
#head(train_df, 10)

# Convert to factor.
df$CHANGED_FRANCHISE <- as.factor(df$CHANGED_FRANCHISE)
df$SEASON_ID <- as.factor(df$SEASON_ID)

#Logistic regression model
logit_mod <- glm(CHANGED_FRANCHISE ~ ., family=binomial(link='logit'), data = df)
```

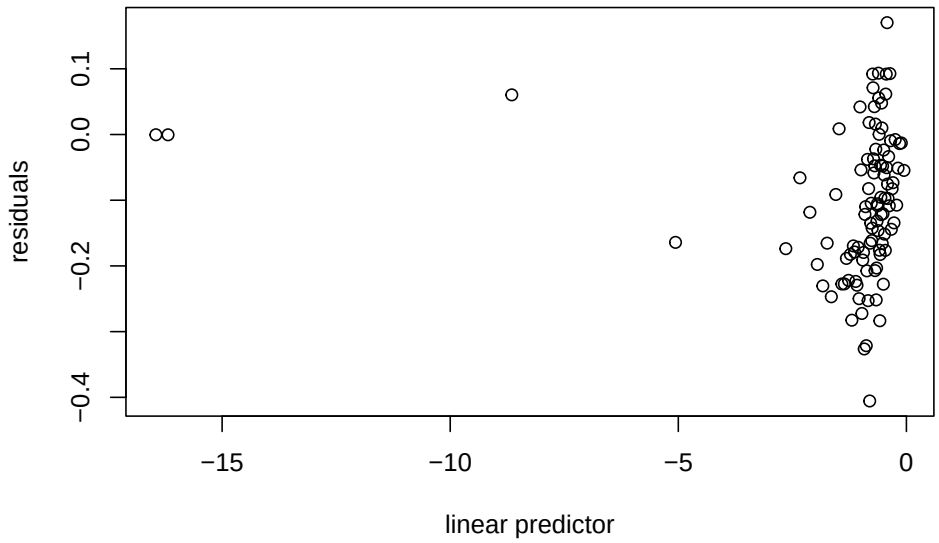
```
summary(logit_mod)
```

```
##
## Call:
## glm(formula = CHANGED_FRANCHISE ~ ., family = binomial(link = "logit"),
##     data = df)
##
## Coefficients: (2 not defined because of singularities)
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       9.040e+06  3.534e+06   2.558 0.010521 *
## RANK             -4.685e-03  7.051e-04  -6.645 3.04e-11 ***
## TEAM_ID          -5.612e-03  2.194e-03  -2.558 0.010521 *
## GP                6.780e-04  1.996e-03   0.340 0.734053
## MIN               3.643e-05  1.229e-04   0.296 0.767000
## FGM              -1.261e-03  9.497e-04  -1.328 0.184273
## FG_PCT           -1.762e-01  2.360e-01  -0.747 0.455131
## FG3M              8.186e-03  4.025e-03   2.034 0.041980 *
## FG3A             -3.087e-03  1.504e-03  -2.053 0.040073 *
## FG3_PCT           2.397e-02  1.349e-01   0.178 0.858962
## FTM               8.610e-04  2.246e-03   0.383 0.701448
## FTA              -8.039e-04  1.661e-03  -0.484 0.628459
## FT_PCT            1.225e-01  1.160e-01   1.055 0.291288
## OREB             -8.225e-04  1.222e-03  -0.673 0.500836
## DREB              3.064e-03  8.854e-04   3.460 0.000539 ***
## REB                      NA         NA      NA       NA
## AST               9.928e-05  9.098e-04   0.109 0.913103
## STL              -1.036e-03  1.694e-03  -0.612 0.540649
## BLK              -4.722e-03  1.400e-03  -3.372 0.000745 ***
## TOV               2.853e-03  1.707e-03   1.671 0.094683 .
```

```
## PF               -7.020e-04  8.128e-04  -0.864 0.387774
## PTS                      NA         NA      NA       NA
## EFF              -2.424e-03  7.161e-04  -3.384 0.000713 ***
## AST_TOV           5.576e-02  2.823e-02   1.975 0.048223 *
## STL_TOV          -6.334e-03  5.777e-02  -0.110 0.912684
## SEASON_ID1996-97  1.593e+01  1.122e+02   0.142 0.887104
## SEASON_ID1997-98  1.575e+01  1.122e+02   0.140 0.888314
## SEASON_ID1998-99  1.523e+01  1.122e+02   0.136 0.892002
## SEASON_ID1999-00  1.559e+01  1.122e+02   0.139 0.889497
## SEASON_ID2000-01  1.570e+01  1.122e+02   0.140 0.888664
## SEASON_ID2001-02  1.565e+01  1.122e+02   0.140 0.889008
## SEASON_ID2002-03  1.553e+01  1.122e+02   0.138 0.889853
## SEASON_ID2003-04  1.600e+01  1.122e+02   0.143 0.886544
## SEASON_ID2004-05  1.620e+01  1.122e+02   0.144 0.885170
## SEASON_ID2005-06  1.565e+01  1.122e+02   0.139 0.889067
## SEASON_ID2006-07  1.537e+01  1.122e+02   0.137 0.891015
## SEASON_ID2007-08  1.562e+01  1.122e+02   0.139 0.889263
## SEASON_ID2008-09  1.579e+01  1.122e+02   0.141 0.888038
## SEASON_ID2009-10  1.581e+01  1.122e+02   0.141 0.887891
## SEASON_ID2010-11  1.621e+01  1.122e+02   0.144 0.885125
## SEASON_ID2011-12  1.540e+01  1.122e+02   0.137 0.890792
## SEASON_ID2012-13  1.598e+01  1.122e+02   0.142 0.886707
## SEASON_ID2013-14  1.598e+01  1.122e+02   0.142 0.886714
## SEASON_ID2014-15  1.587e+01  1.122e+02   0.142 0.887471
## SEASON_ID2015-16  1.578e+01  1.122e+02   0.141 0.888146
## SEASON_ID2016-17  1.589e+01  1.122e+02   0.142 0.887371
## SEASON_ID2017-18  1.578e+01  1.122e+02   0.141 0.888114
## SEASON_ID2018-19  1.598e+01  1.122e+02   0.142 0.886690
## SEASON_ID2019-20  1.592e+01  1.122e+02   0.142 0.887129
## SEASON_ID2020-21  1.591e+01  1.122e+02   0.142 0.887213
## SEASON_ID2021-22  1.600e+01  1.122e+02   0.143 0.886567
## SEASON_ID2022-23  1.591e+01  1.122e+02   0.142 0.887192
## SEASON_ID2023-24  1.597e+01  1.122e+02   0.142 0.886758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17128  on 13828  degrees of freedom
## Residual deviance: 16225  on 13778  degrees of freedom
## AIC: 16327
##
## Number of Fisher Scoring iterations: 15
```

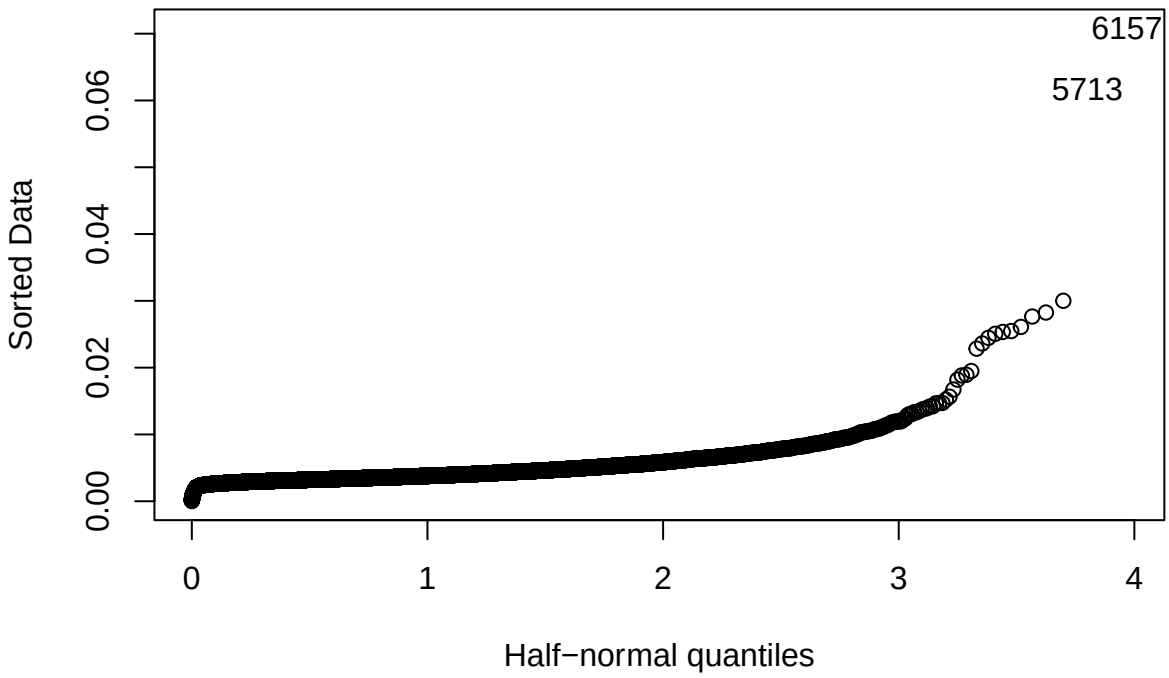
```
df <- mutate(df,residuals = residuals(logit_mod),linpred = predict(logit_mod))

gdf <- group_by(df,bin = cut(linpred, breaks = unique(quantile(linpred, (1:100)/101))))

diagdf <- summarise(gdf, residuals = mean(residuals), linpred = mean(linpred))
plot(residuals ~ linpred, diagdf, xlab = "linear predictor")
```

Diagnostics For Logistic Regression:



The values seem to cluster up close to zero, and this breaks our constant variance assumption. Because there is not a random scatter in our fitted versus residual plot, we can say that this model does not meet our assumption of linearity.

```
#Half norm plot
library(faraway)
halfnorm(hatvalues(logit_mod))
```



From the half norm plot, we can see that there are only two outliers. The plot seems to look similar to a sigmoid function, which is expected since the model is logistic regression.

```
#Smaller model from ommitting non significant predictors
columns <- c('RANK', 'TEAM_ID', 'FG3M', 'FG3A', 'DREB', 'BLK', 'EFF', 'AST_TOV', 'CHANGED_FRANCHISE')
new_train <- df[, columns]

new_train$CHANGED_FRANCHISE <- as.factor(new_train$CHANGED_FRANCHISE)

logit_mod_reduced <- glm(CHANGED_FRANCHISE ~ ., family=binomial(link='logit'), data = new_train)
#Summary of new log mod
summary(logit_mod_reduced)
```

Model Comparison:

```
##
## Call:
## glm(formula = CHANGED_FRANCHISE ~ ., family = binomial(link = "logit"),
##     data = new_train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.938e+06  3.484e+06   2.566   0.0103 *
```

```
## RANK        -2.638e-03  3.886e-04  -6.788 1.14e-11 ***
## TEAM_ID     -5.550e-03  2.163e-03  -2.566   0.0103 *
## FG3M         4.091e-03  2.997e-03   1.365   0.1722
## FG3A        -9.430e-04  1.150e-03  -0.820   0.4121
## DREB         3.328e-03  4.265e-04   7.802 6.08e-15 ***
## BLK         -4.883e-03  1.046e-03  -4.668 3.05e-06 ***
## EFF         -2.203e-03  1.629e-04 -13.524  < 2e-16 ***
## AST_TOV      1.033e-01  2.029e-02   5.093 3.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17128  on 13828  degrees of freedom
## Residual deviance: 16681  on 13820  degrees of freedom
## AIC: 16699
##
## Number of Fisher Scoring iterations: 4
```

```
#Compare the reduced vs full models
anova(logit_mod_reduced, logit_mod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHANGED_FRANCHISE ~ RANK + TEAM_ID + FG3M + FG3A + DREB + BLK +
##     EFF + AST_TOV
## Model 2: CHANGED_FRANCHISE ~ RANK + TEAM_ID + GP + MIN + FGM + FG_PCT +
##     FG3M + FG3A + FG3_PCT + FTM + FTA + FT_PCT + OREB + DREB +
##     REB + AST + STL + BLK + TOV + PF + PTS + EFF + AST_TOV +
##     STL_TOV + SEASON_ID
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     13820      16681
## 2     13778      16225 42   456.43 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 2.2e-16 which is extremely small. This means that the additional predictors in Model 2(the full model) significantly improve the fit of the model and help explain the response variable(changed franchise).

Hypothesis Test Null Hypothesis(H_0): The reduced model(Model 1) fits just as well as the full model(Model 2) and the additional predictors in the full model do not improve the model.

Alternative Hypothesis(H_1): The full model(Model 2) is better fit than the reduced model(Model 1). alpha = 0.05

D = $Deviance_{Model1}$ - $Deviance_{Model2}$ = 16681 - 16225 = 456.43

D = 456.43

df = 42

p-value = 2.2e-16

Conclusion:

Since the p-value is extremely small, we reject the null hypothesis. This indicates that the full model offers a significantly improved fit over the reduced model, and the additional predictors included in the full model are meaningful in explaining the response variable.

```r
ci <- confint.default(logit_mod) # default 95% CI based on standard error.
print(ci)
```

Confidence Interval:

```
##                           2.5 %        97.5 %
## (Intercept)       2.113917e+06  1.596509e+07
## RANK             -6.066968e-03 -3.303133e-03
## TEAM_ID          -9.912439e-03 -1.312501e-03
## GP               -3.233260e-03  4.589204e-03
## MIN              -2.045272e-04  2.773803e-04
## FGM              -3.122370e-03  6.004597e-04
## FG_PCT           -6.387294e-01  2.862465e-01
## FG3M              2.968644e-04  1.607413e-02
## FG3A             -6.033614e-03 -1.398778e-04
## FG3_PCT          -2.404015e-01  2.883386e-01
## FTM              -3.540946e-03  5.262984e-03
## FTA              -4.060248e-03  2.452347e-03
## FT_PCT           -1.049855e-01  3.499189e-01
## OREB             -3.217301e-03  1.572263e-03
## DREB              1.328452e-03  4.799117e-03
## REB                         NA            NA
## AST              -1.683820e-03  1.882375e-03
## STL              -4.355693e-03  2.283248e-03
## BLK              -7.466725e-03 -1.977690e-03
## TOV              -4.929675e-04  6.198874e-03
## PF               -2.294969e-03  8.910467e-04
## PTS                         NA            NA
## EFF              -3.827187e-03 -1.020099e-03
## AST_TOV           4.357724e-04  1.110850e-01
## STL_TOV          -1.195542e-01  1.068855e-01
## SEASON_ID1996-97 -2.039316e+02  2.357823e+02
## SEASON_ID1997-98 -2.041034e+02  2.356105e+02
## SEASON_ID1998-99 -2.046269e+02  2.350870e+02
## SEASON_ID1999-00 -2.042714e+02  2.354425e+02
## SEASON_ID2000-01 -2.041531e+02  2.355608e+02
## SEASON_ID2001-02 -2.042020e+02  2.355119e+02
## SEASON_ID2002-03 -2.043220e+02  2.353919e+02
## SEASON_ID2003-04 -2.038521e+02  2.358617e+02
## SEASON_ID2004-05 -2.036569e+02  2.360570e+02
## SEASON_ID2005-06 -2.042104e+02  2.355035e+02
## SEASON_ID2006-07 -2.044869e+02  2.352270e+02
## SEASON_ID2007-08 -2.042382e+02  2.354757e+02
## SEASON_ID2008-09 -2.040643e+02  2.356496e+02
## SEASON_ID2009-10 -2.040434e+02  2.356705e+02
## SEASON_ID2010-11 -2.036505e+02  2.360634e+02
## SEASON_ID2011-12 -2.044552e+02  2.352587e+02
## SEASON_ID2012-13 -2.038752e+02  2.358387e+02
## SEASON_ID2013-14 -2.038763e+02  2.358376e+02
## SEASON_ID2014-15 -2.039838e+02  2.357301e+02
## SEASON_ID2015-16 -2.040797e+02  2.356342e+02
## SEASON_ID2016-17 -2.039696e+02  2.357443e+02
```

```
## SEASON_ID2017-18 -2.040750e+02  2.356388e+02
## SEASON_ID2018-19 -2.038728e+02  2.358410e+02
## SEASON_ID2019-20 -2.039352e+02  2.357787e+02
## SEASON_ID2020-21 -2.039471e+02  2.357668e+02
## SEASON_ID2021-22 -2.038553e+02  2.358586e+02
## SEASON_ID2022-23 -2.039442e+02  2.357697e+02
## SEASON_ID2023-24 -2.038826e+02  2.358313e+02
```

A coefficient is statistically significant if its confidence interval doesn't include 0, and if it does than it is not statistically significant.

This happens because including 0 means that there is a plausible chance that the true effect is zero i.e., no effect.

List of Significant Parameters:

```
has_zero <- ci[, 1] <0 & ci[, 2] >0
significant_params <- rownames(ci)[!has_zero]
significant_params
```

```
##  [1] "(Intercept)" "RANK"        "TEAM_ID"     "FG3M"        "FG3A"
##  [6] "DREB"        NA            "BLK"         NA            "EFF"
## [11] "AST_TOV"
```

List of Insignificant Parameters:

```
insignificant_params <- rownames(ci)[has_zero]
insignificant_params
```

```
##  [1] "GP"               "MIN"              "FGM"              "FG_PCT"
##  [5] "FG3_PCT"          "FTM"              "FTA"              "FT_PCT"
##  [9] "OREB"             NA                 "AST"              "STL"
## [13] "TOV"              "PF"               NA                 "STL_TOV"
## [17] "SEASON_ID1996-97" "SEASON_ID1997-98" "SEASON_ID1998-99" "SEASON_ID1999-00"
## [21] "SEASON_ID2000-01" "SEASON_ID2001-02" "SEASON_ID2002-03" "SEASON_ID2003-04"
## [25] "SEASON_ID2004-05" "SEASON_ID2005-06" "SEASON_ID2006-07" "SEASON_ID2007-08"
## [29] "SEASON_ID2008-09" "SEASON_ID2009-10" "SEASON_ID2010-11" "SEASON_ID2011-12"
## [33] "SEASON_ID2012-13" "SEASON_ID2013-14" "SEASON_ID2014-15" "SEASON_ID2015-16"
## [37] "SEASON_ID2016-17" "SEASON_ID2017-18" "SEASON_ID2018-19" "SEASON_ID2019-20"
## [41] "SEASON_ID2020-21" "SEASON_ID2021-22" "SEASON_ID2022-23" "SEASON_ID2023-24"
```