NBA GAME PREDICTION

TEAM MEMBERS:

AKSHTA
ARCHANA IYER
ASWINI SATISHKUMAR
LAAVANYA GANESH
PUJITHA PRASHANTH HEGDE
VAIBHAV RAVICHANDRAN

INTRODUCTION

The National Basketball Association (NBA) is the prominent men's professional basketball league in North America. This league was founded in New York on June 6th 1946. It was initially known as the Basketball Association of America (BAA) and it adopted the National Basketball Association name on August 3rd 1949 after merging with its rival National Basketball League (NBL). The NBA is one of the four major professional sports league in United States and Canada. NBA started with 8 teams and now there are 30 teams (29 in the United States and 1 in Canada).

OBJECTIVE OF THE PROJECT

The objective of this project is to predict the win percentage of the teams. In any sport a winning percentage is the fraction of games or matches a team or an individual has won. We can define winning percentage as wins divided by the total number of matches played i.e. wins plus losses. Another objective of this project would be to analyze and understand the player demographics and team performances.

We mainly consider the following independent variables to predict WIN PERCENTAGE which is the dependent or outcome variable.

INDEPENDENT VARIABLE	DEPENDENT VARIABLE
GOALS	WIN PERCENTAGE
FREE THROWS	
OFFENSIVE REBOUNDS	
DEFENSIVE REBOUNDS	
ASSISTS	
FOULS	
STEALS	
BLOCKS	
POINTS	

SOFTWARES USED:

We have used the following softwares:

- R - It is a programming language and software environment for statistical computing. R language is most commonly used for developing statistical software and data analysis. We have used R for regression and also exploratory data analysis.

- RapidMiner Studio - This is a powerful visual programming environment for rapidly building the complete predictive analytic workflow. This tool features pre-defined data preparation and also machine learning algorithms to efficiently support data analysis. We use RapidMiner Studio only for exploratory data analysis.

DATA MODEL

We have performed linear regression with feature selection using R programming language.

DATA COLLECTION & PROCESSING

We have collected the data from http://www.basketball-reference.com. The data we have chosen is the statistical data of all opponent teams with our home team. We then check for missing values and outliers. We select the variables and do principal component analysis.

DESCRIPTION OF DATASET

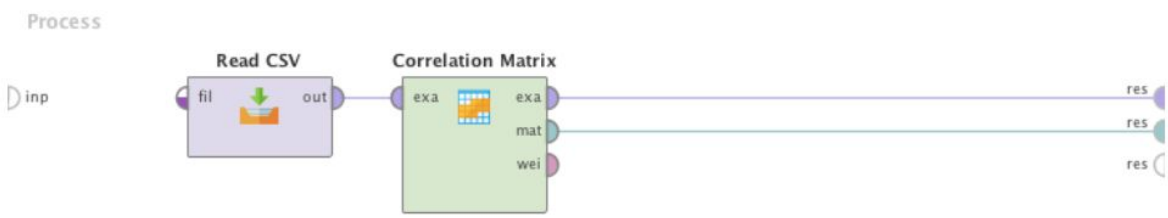
The dataset we obtained has 11,000 observations .

Variable Name	DataType	Sample Field Values
team (Team name)	Factor	“CHI” - Chicago,Bulls “CAR” - Carolina,Cougars
year (Year)	int	1946 to 2004
leag (League)	Factor	“A" and "N”
o_fgm (Opponent Field Goal Made)	int	1397 1879 1674 1437 1465 1510 ...
o_fga (Opponent Field Goal Attempts)	int	5133 6309 5699 5843 5255 ...
o_ftm (Opponent Free Throws Made)	int	811 939 903 923 951 1098 ...
o_fta (Opponent Free Throws Attempted)	int	1375 1550 1428 1494 1438 ...
o_oreb (Opponent Offensive Rebounds)	int	0 1 2 3 ...
o_dreb (Opponent Defensive Rebounds)	int	0 1 2 3 ...
o_reb (Opponent Rebounds)	int	0 1 2 3 ...
o_asts (Opponent Assists)	int	470 436 494 482 457 343 272 ...
o_pf (Opponent Personal Fouls)	int	1202 1473 1246 1351 1218 ...
o_stl (Opponent Steals)	int	0 1 2 3 ...
o_to (Opponent Turnovers)	int	0 1 2 3 ...
o_blk (Opponent Blocks)	int	0 1 2 3 ...
o_3pm (Opponent 3 Point Goals Made)	int	0 1 2 3 ...
o_3pa (Opponent 3 Point Goals Attempted)	int	0 1 2 3 ...
o_pts (Opponent Points)	int	3605 4697 4251 3797 3881 ...
d_fgm (Home Team Field Goals Made)	int	0 1 2 3 ...

d_fga (Home Team Field Goals Attempted)	int	0 1 2 3 ...
d_ftm (Home Team Free Throws Made)	int	0 1 2 3 ...
d_fta (Home Team Free Throws Attempted)	int	0 1 2 3 ...
d_oreb (Home Team Offensive Rebounds)	int	0 1 2 3 ...
d_dreb (Home Team Defensive Rebounds)	int	0 1 2 3 ...
d_reb (Home Team Rebounds)	int	0 1 2 3 ...
d_asts (Home Team Assists)	int	0 1 2 3 ...
d_pf (Home Team Personal Fouls)	int	0 1 2 3 ...
d_stl (Home Team Steals)	int	0 1 2 3 ...
d_to (Home Team Turnovers)	int	0 1 2 3 ...
d_blk (Home Team Blocks)	int	0 1 2 3 ...
d_3pm (Home Team 3 Point Goals Made)	int	0 1 2 3 ...
d_3pa (Home Team 3 Point Goals Attempted)	int	0 1 2 3 ...
d_pts (Home Team Points)	int	3900 4471 4308 3918 3840 ...
pace (Pace)	num	0 1 2 3 ...
won (Matches Won)	int	22 39 30 20 33 35 15 …
lost (Matches Lost)	int	38 22 30 40 27 25 45 ...

EXPLORATORY DATA ANALYSIS

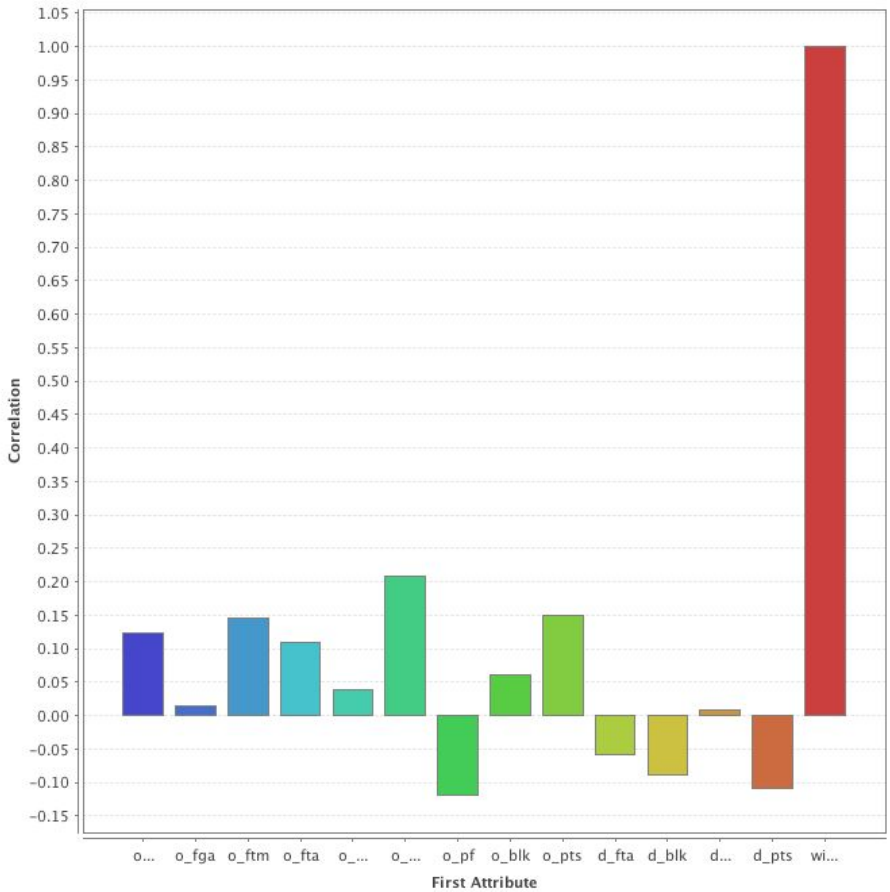
This step is performed in RapidMiner Studio. The process looks like:



We first import the data set into Rapidminer studio and then find the correlation coefficients. The correlation coefficients of few of the independent variables and our dependent variable can be viewed as :

Attribu...	o_fgm	o_fga	o_ftm	o_fta	o_dreb	o_asts	o_pf	o_blk	o_pts	d_fga	d_fta	d_blk	d_3pm	d_pts
o_fgm	The column Attributes	[illegible]	0.589	0.535	0.202	0.817	0.689	0.196	0.981	0.498	0.471	0.190	-0.160	0.951
o_fga	0.900	1	0.659	0.644	-0.087	0.604	0.717	-0.093	0.893	0.234	0.220	-0.083	-0.296	0.895
o_ftm	0.589	0.659	1	0.971	-0.264	0.424	0.667	-0.228	0.693	-0.055	-0.014	-0.226	-0.324	0.657
o_fta	0.535	0.644	0.971	1	-0.346	0.347	0.657	-0.295	0.635	-0.150	-0.096	-0.299	-0.355	0.610
o_dreb	0.202	-0.087	-0.264	-0.346	1	0.434	0.012	0.941	0.180	0.654	0.612	0.953	0.493	0.171
o_asts	0.817	0.604	0.424	0.347	0.434	1	0.562	0.428	0.806	0.526	0.482	0.413	0.011	0.753
o_pf	0.689	0.717	0.667	0.657	0.012	0.562	1	0.004	0.715	0.198	0.307	0.054	-0.239	0.750
o_blk	0.196	-0.093	-0.228	-0.295	0.941	0.428	0.004	1	0.179	0.621	0.584	0.922	0.475	0.165
o_pts	0.981	0.893	0.693	0.635	0.180	0.806	0.715	0.179	1	0.466	0.450	0.175	-0.084	0.964
d_fga	0.498	0.234	-0.055	-0.150	0.654	0.526	0.198	0.621	0.466	1	0.960	0.632	0.346	0.465
d_fta	0.471	0.220	-0.014	-0.096	0.612	0.482	0.307	0.584	0.450	0.960	1	0.620	0.344	0.466
d_blk	0.190	-0.083	-0.226	-0.299	0.953	0.413	0.054	0.922	0.175	0.632	0.620	1	0.494	0.200
d_3pm	-0.160	-0.296	-0.324	-0.355	0.493	0.011	-0.239	0.475	-0.084	0.346	0.344	0.494	1	-0.085
d_pts	0.951	0.895	0.657	0.610	0.171	0.753	0.750	0.165	0.964	0.465	0.466	0.200	-0.085	1
win_pct	0.125	0.014	0.145	0.109	0.039	0.208	-0.118	0.061	0.150	0.010	-0.058	-0.089	0.008	-0.109

This is a graph representing correlation coefficient of all attributes with win percentage.



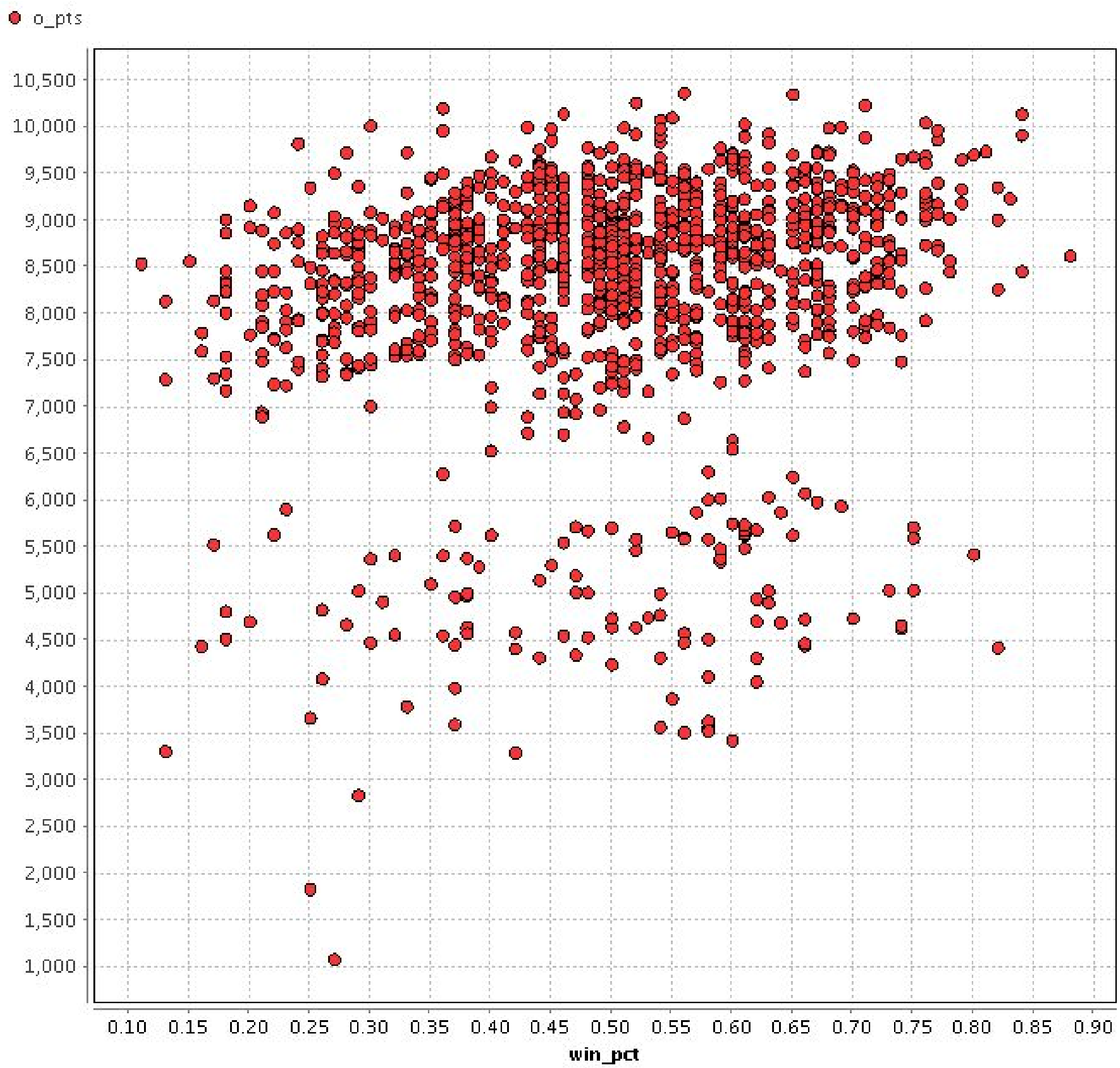
The following analysis can be made

- The correlation of the assists made by opponent has the highest correlation with the win percentage.
- The next highest correlation with win percentage is free throws made by the opponent.
- The third highest correlation is points scored by opponent with win percentage.
- Opponent personal fouls, Home team free throw attempt, Home team blocks and Home team points are negatively correlated with win percentage.

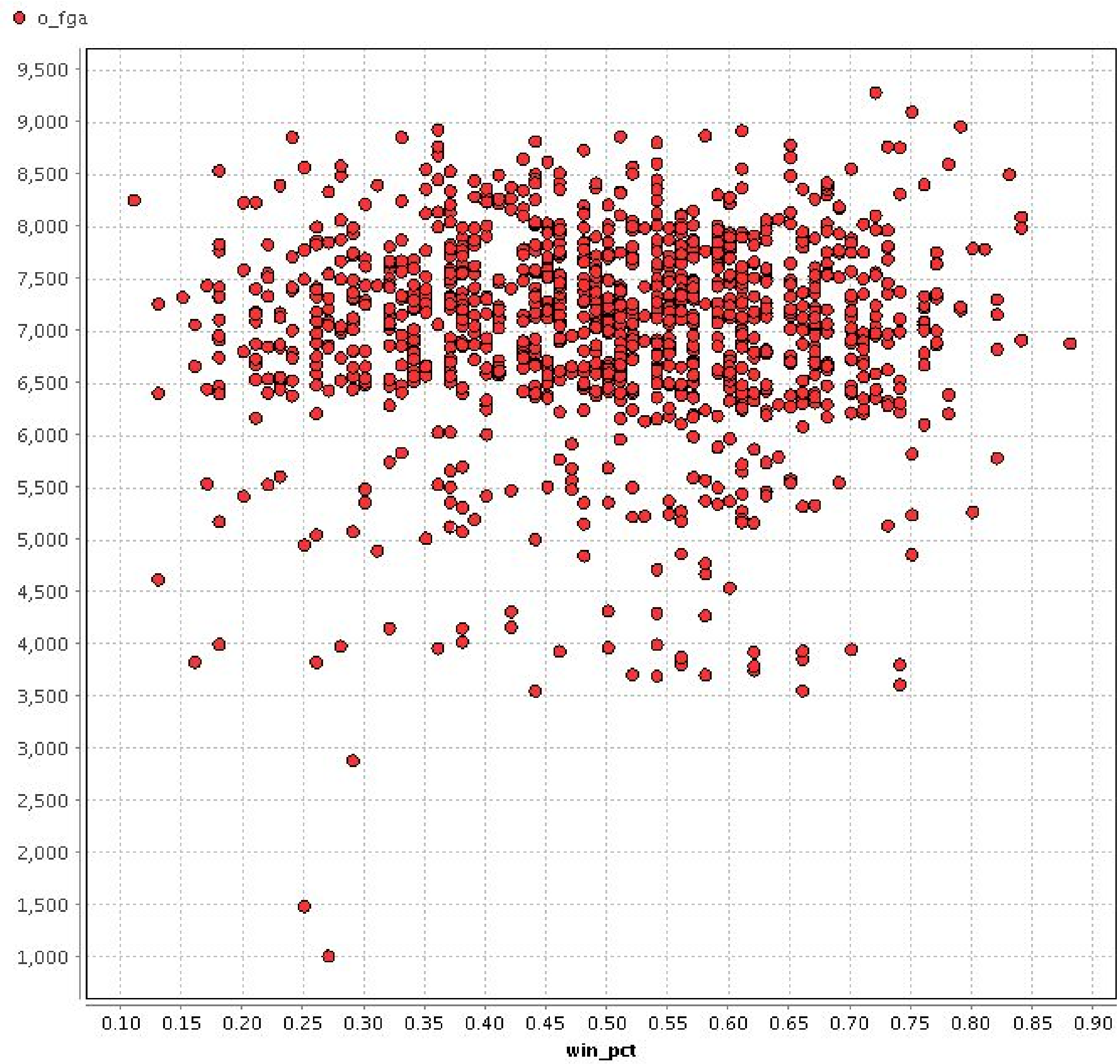
SCATTER PLOTS

We used the same RapidMiner Studio process to obtain the scatter plots.

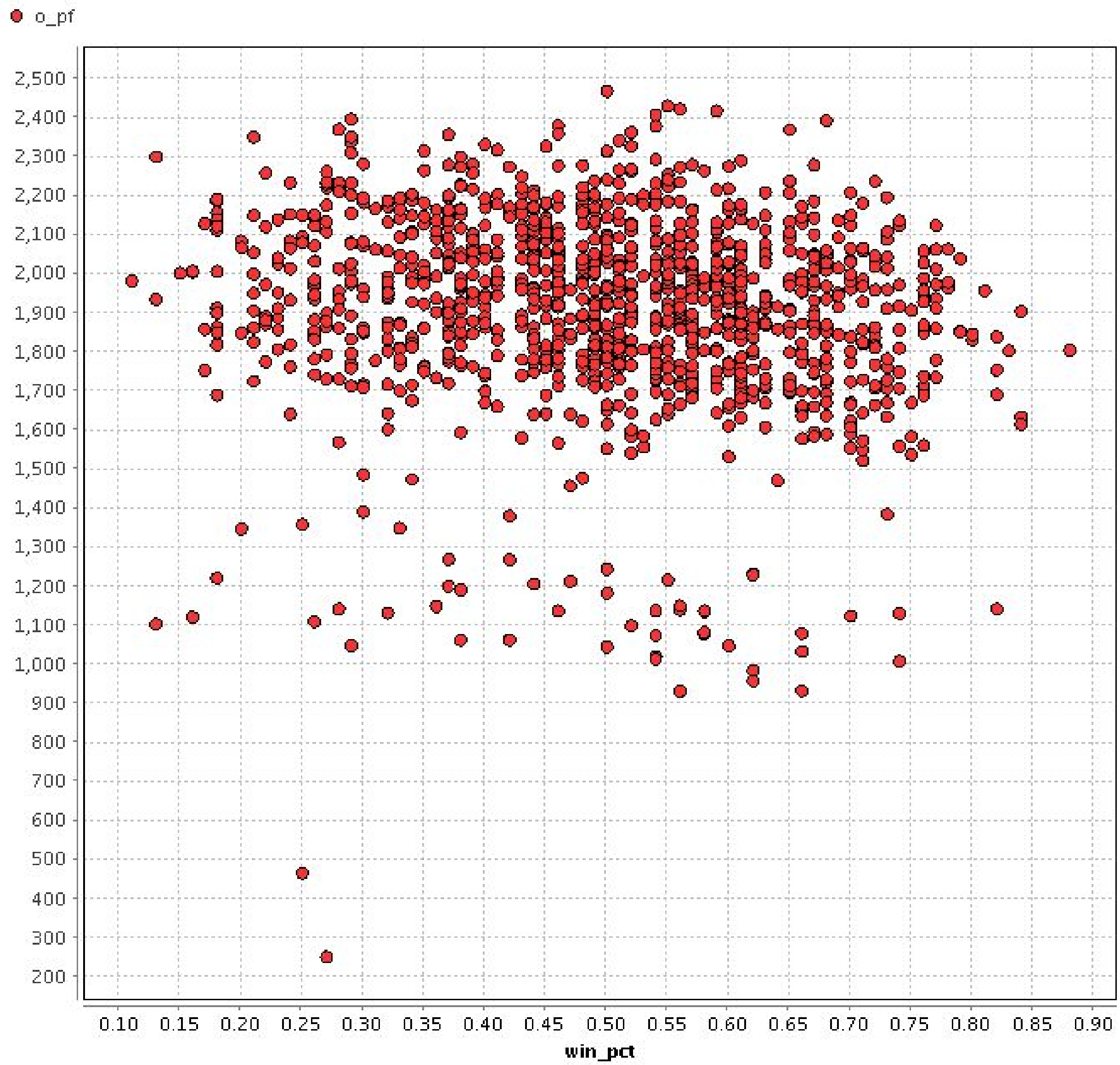
- Opponent Points



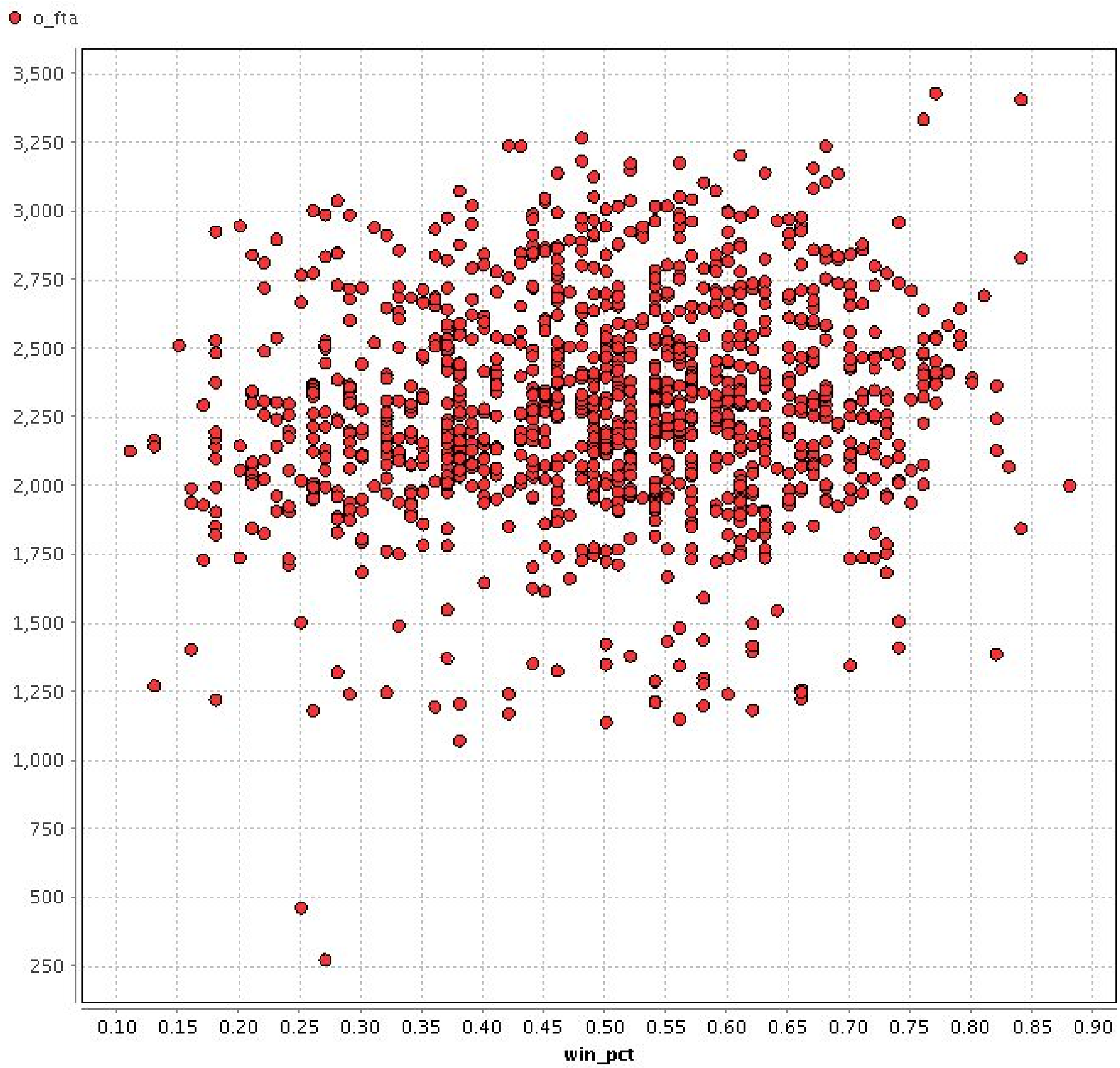
- Opponent Field Goals Made



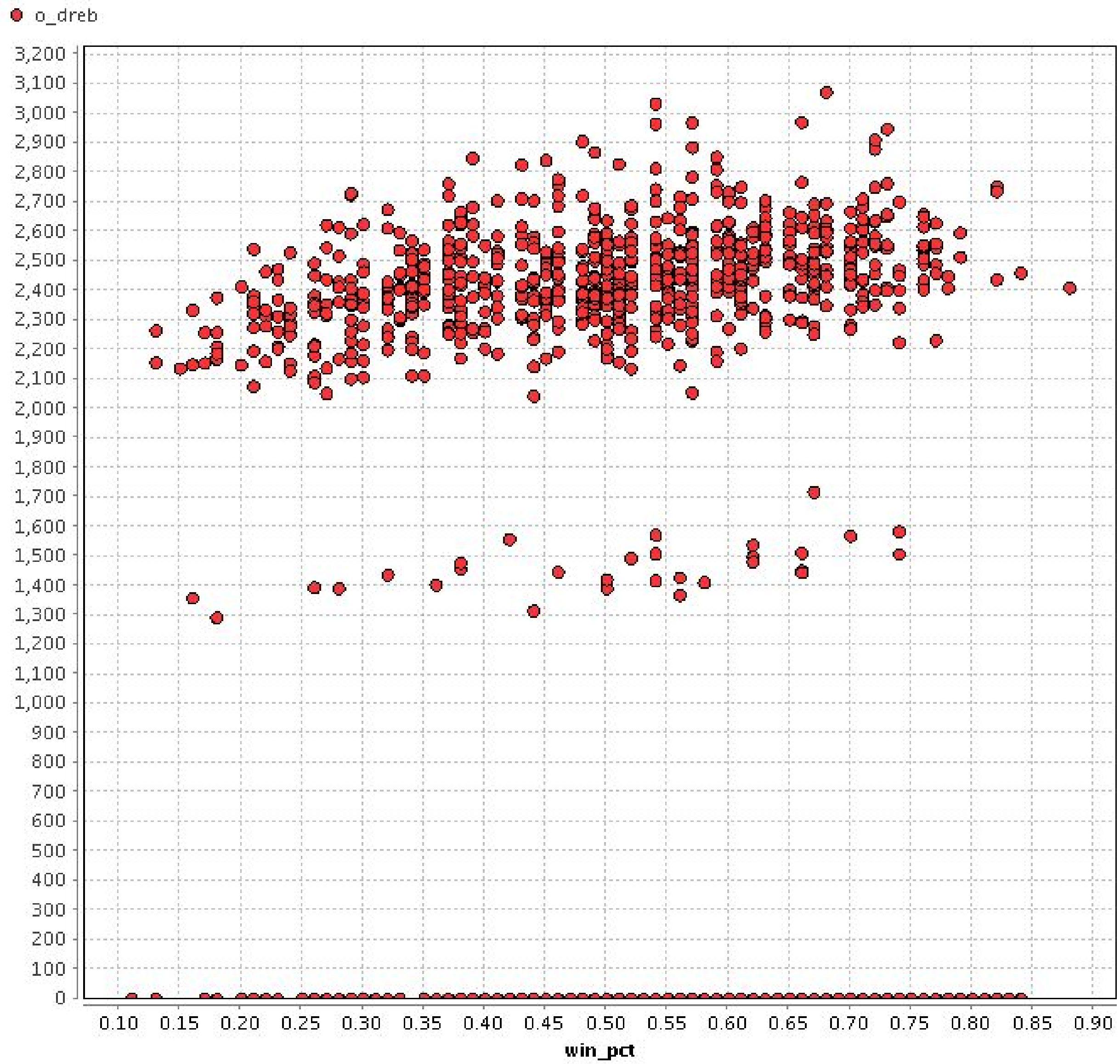
- Opponent Personal Fouls



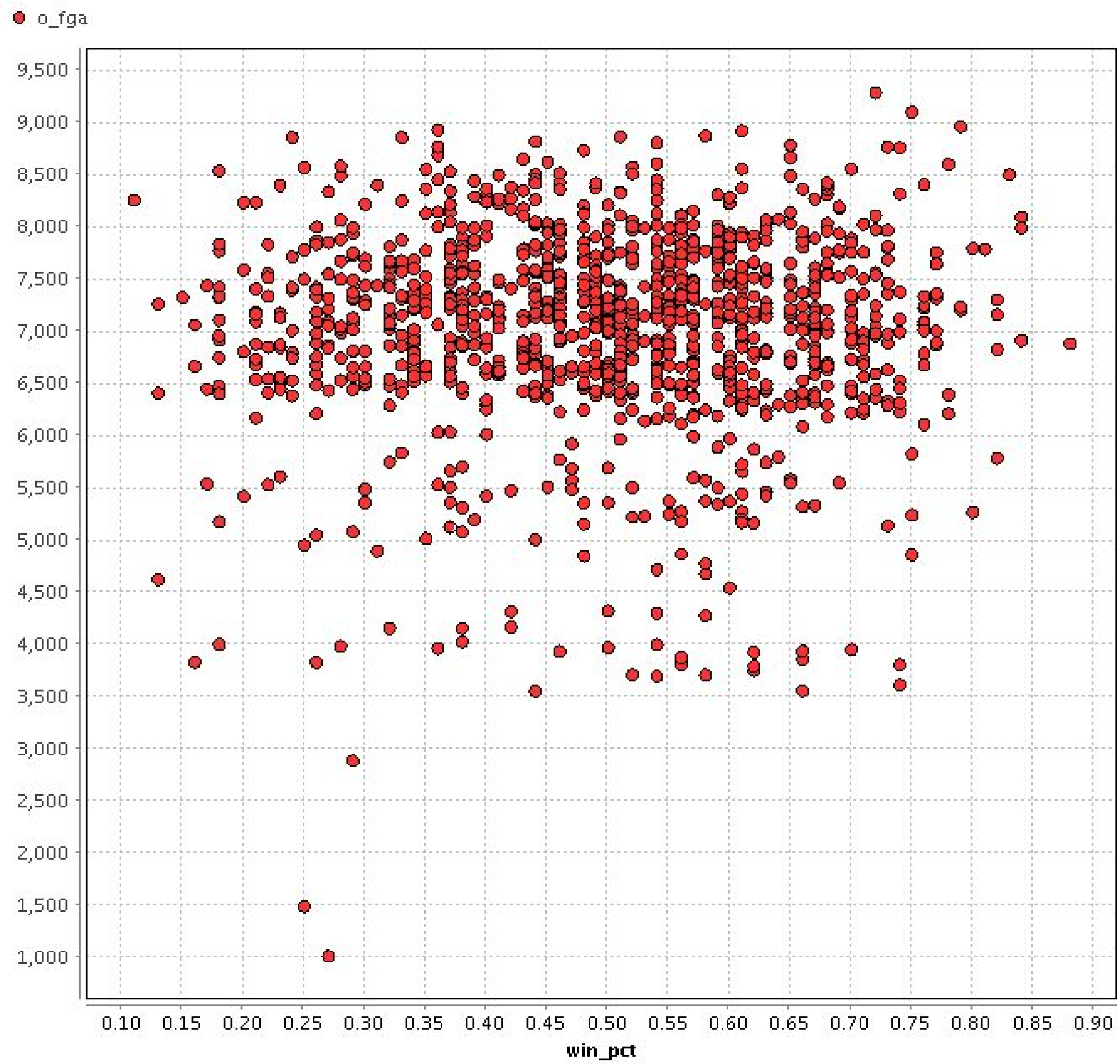
- Opponent Free Throws Attempted



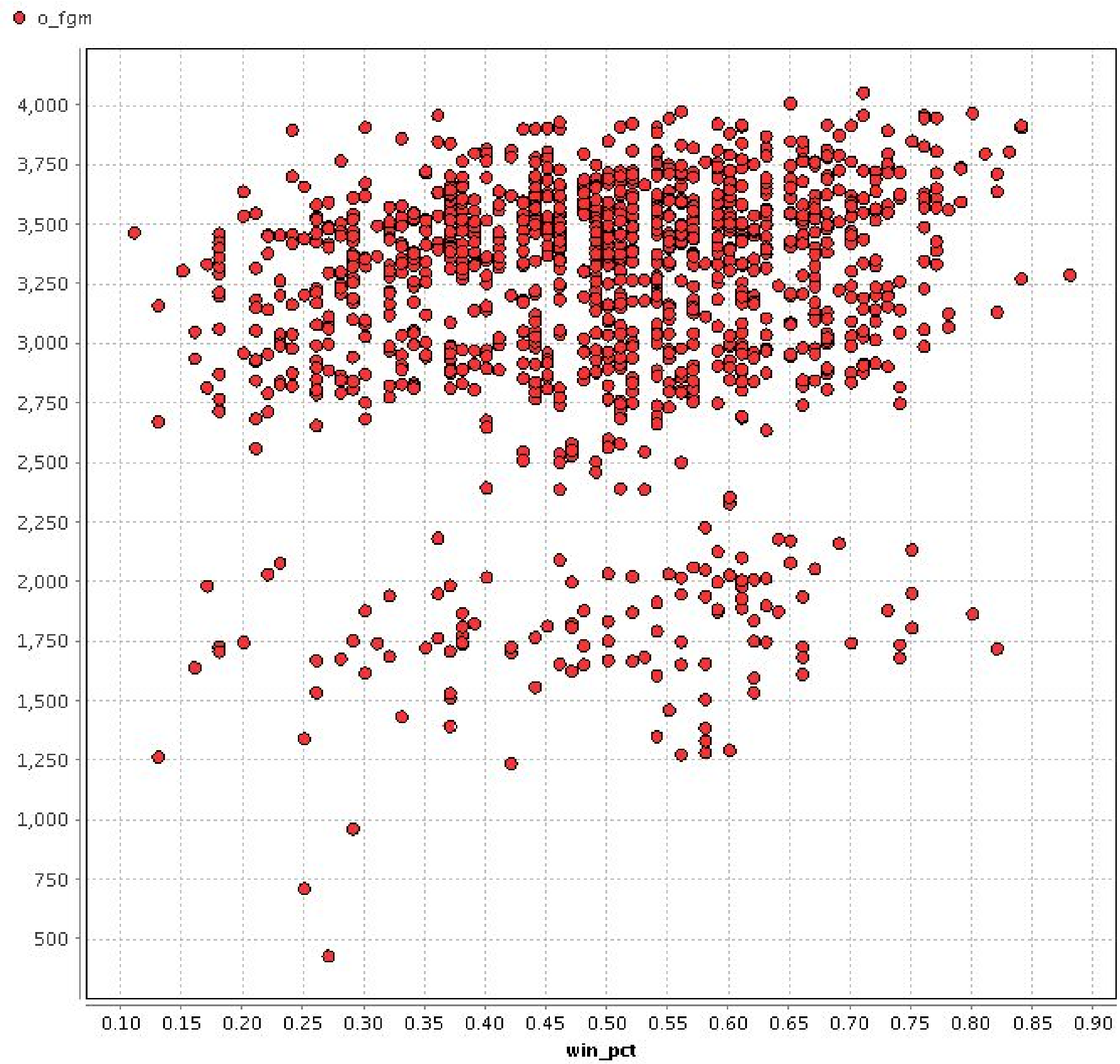
- Opponent Defensive Rebound



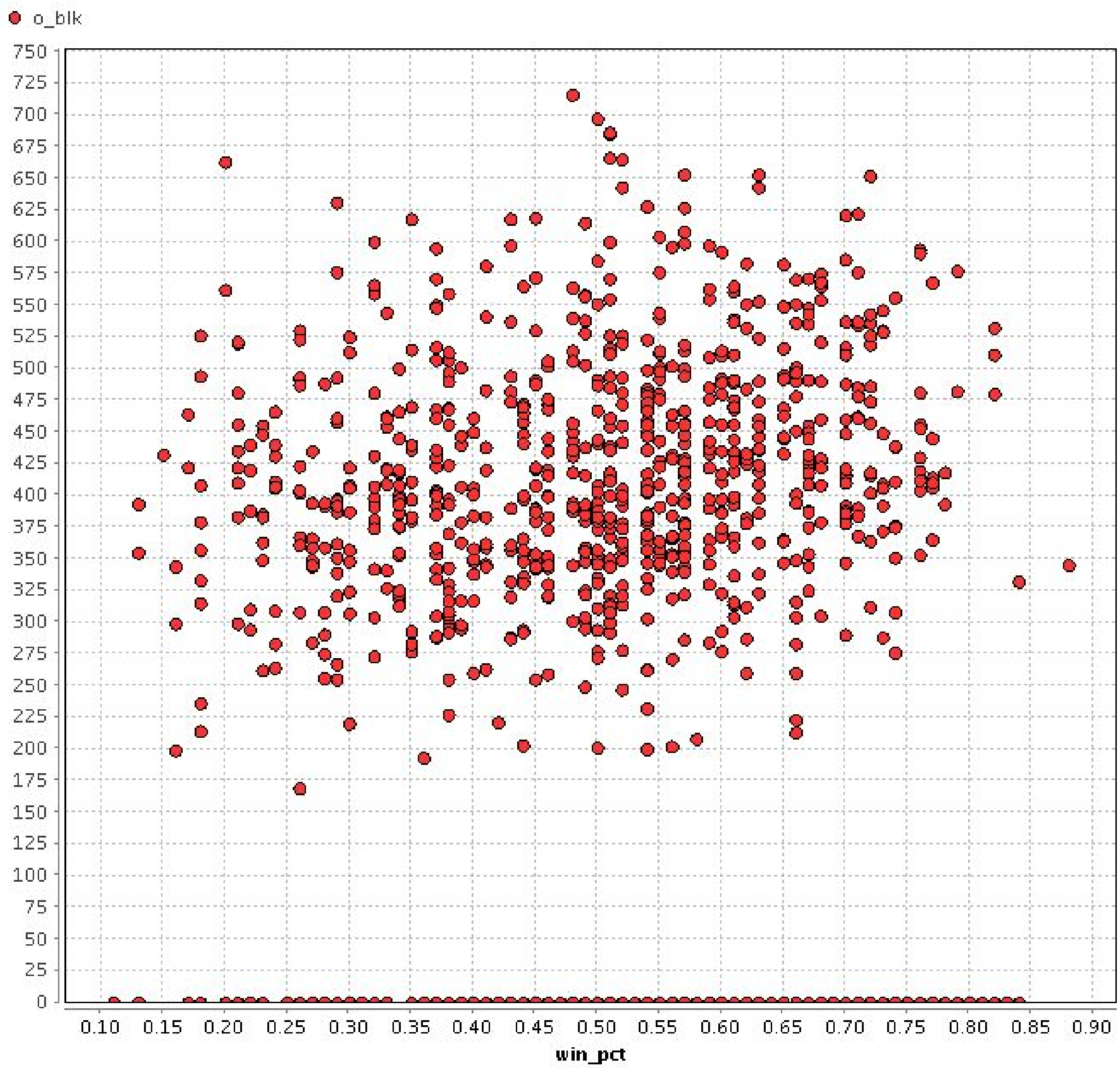
- Opponent Field Goal Attempts



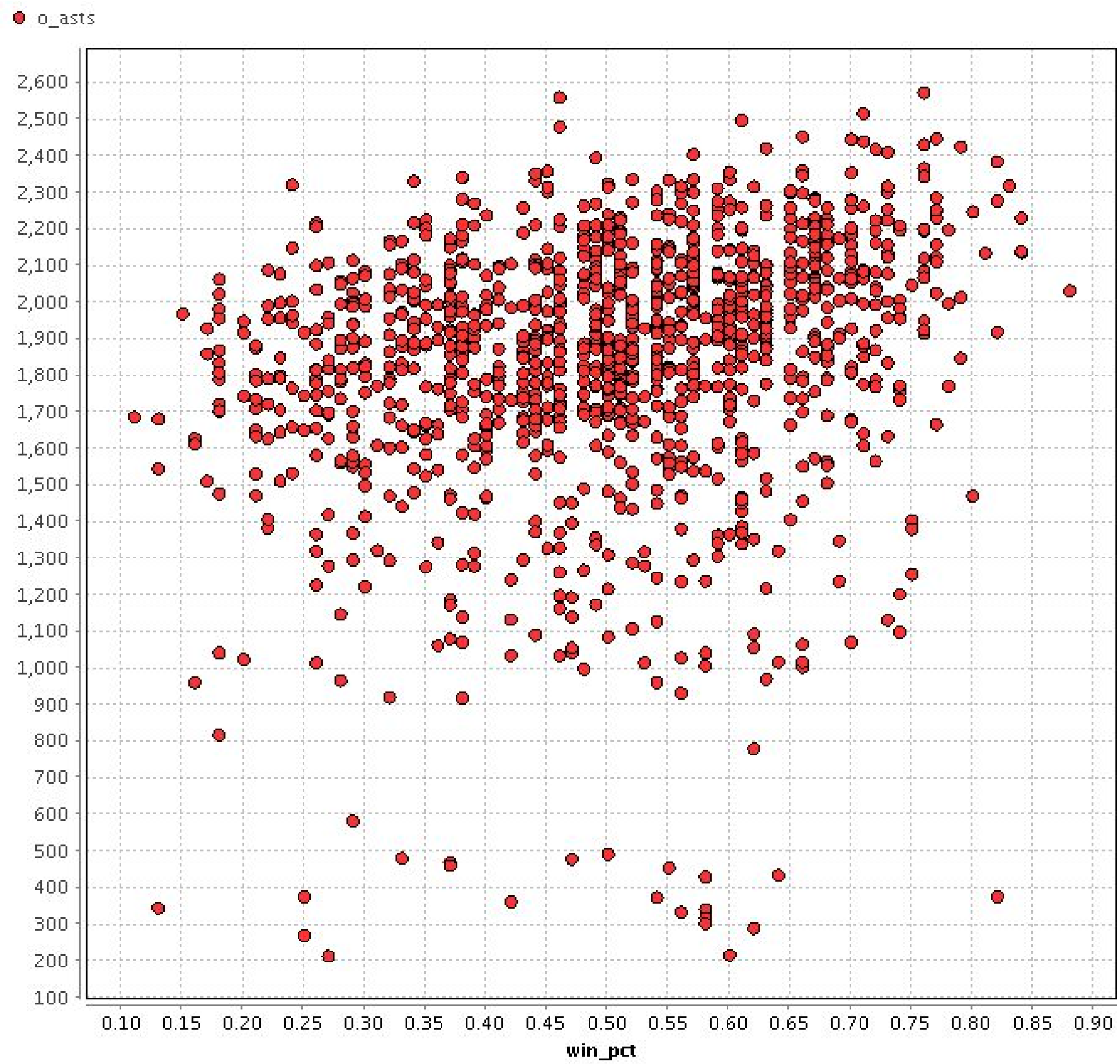
- Opponent Field Goal Made



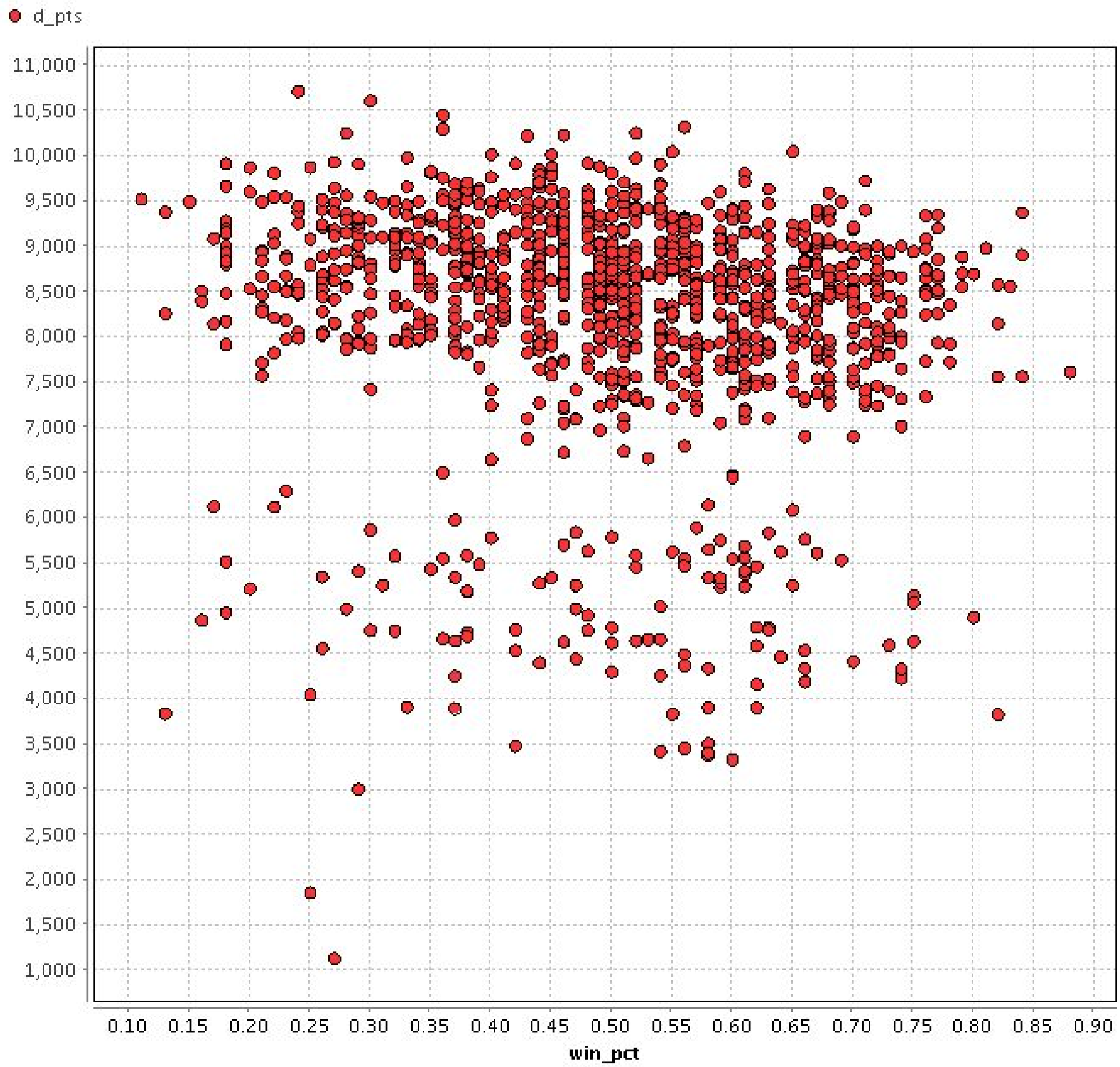
- Opponent Blocks



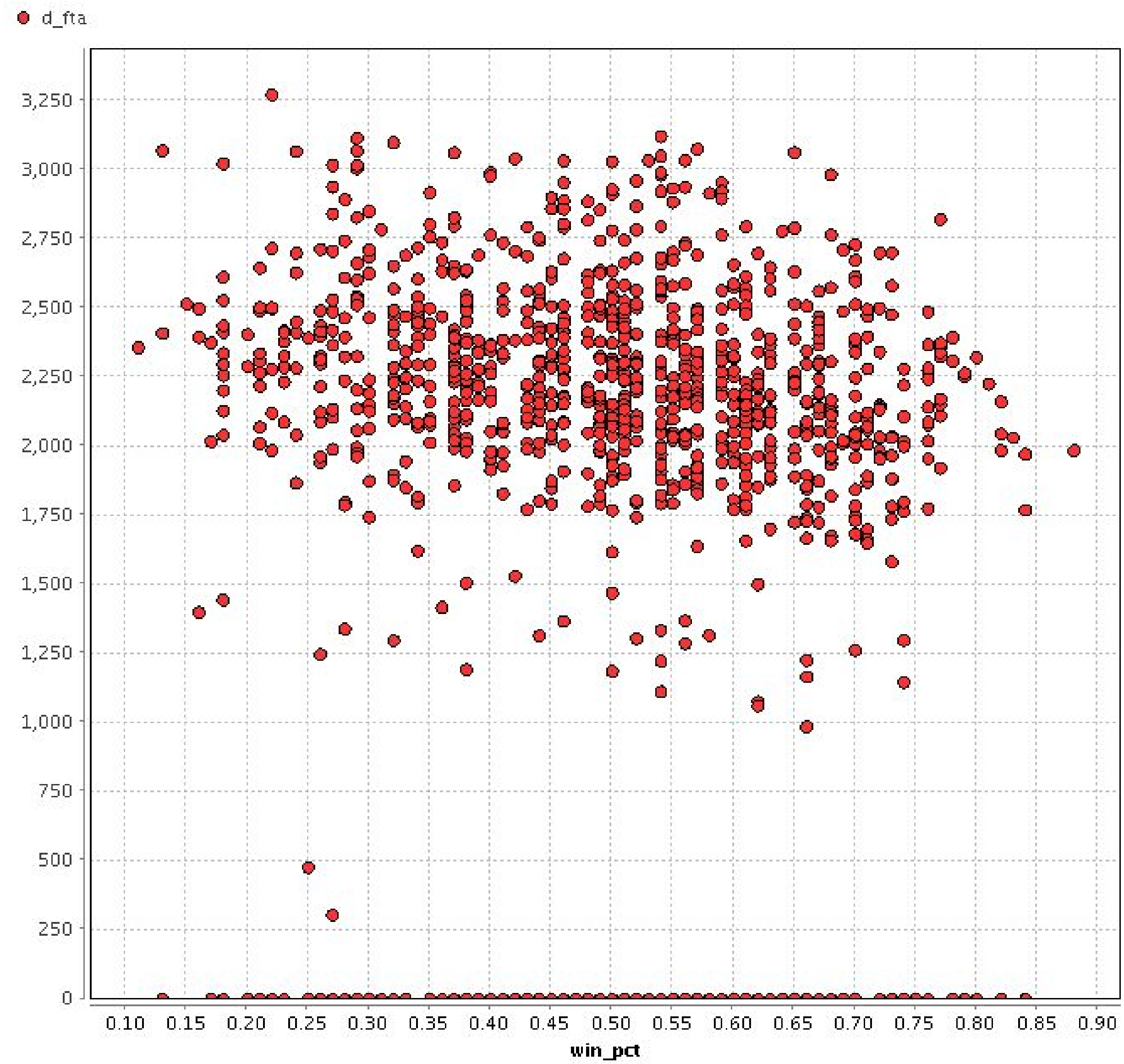
- Opponent Assists



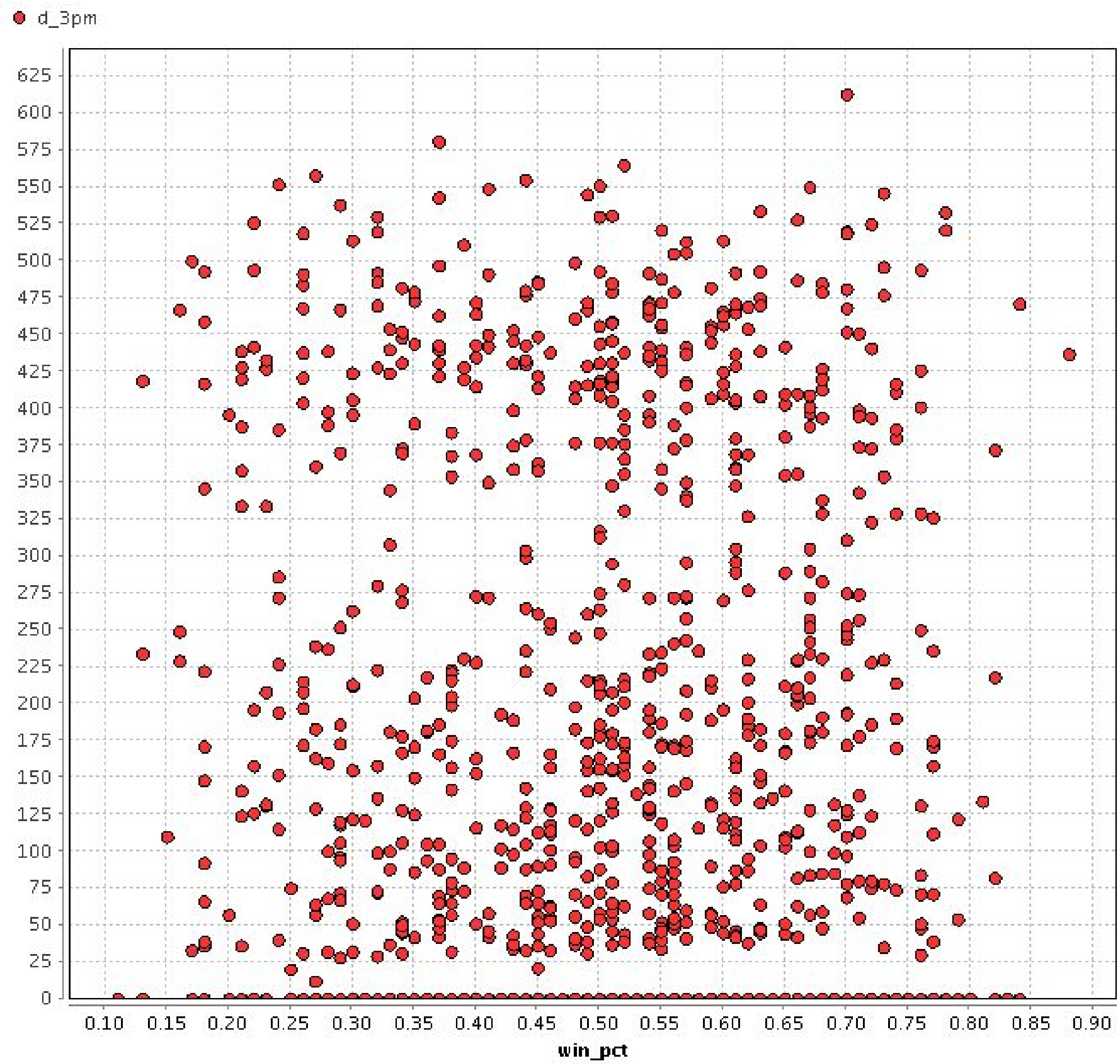
- Home Team Points



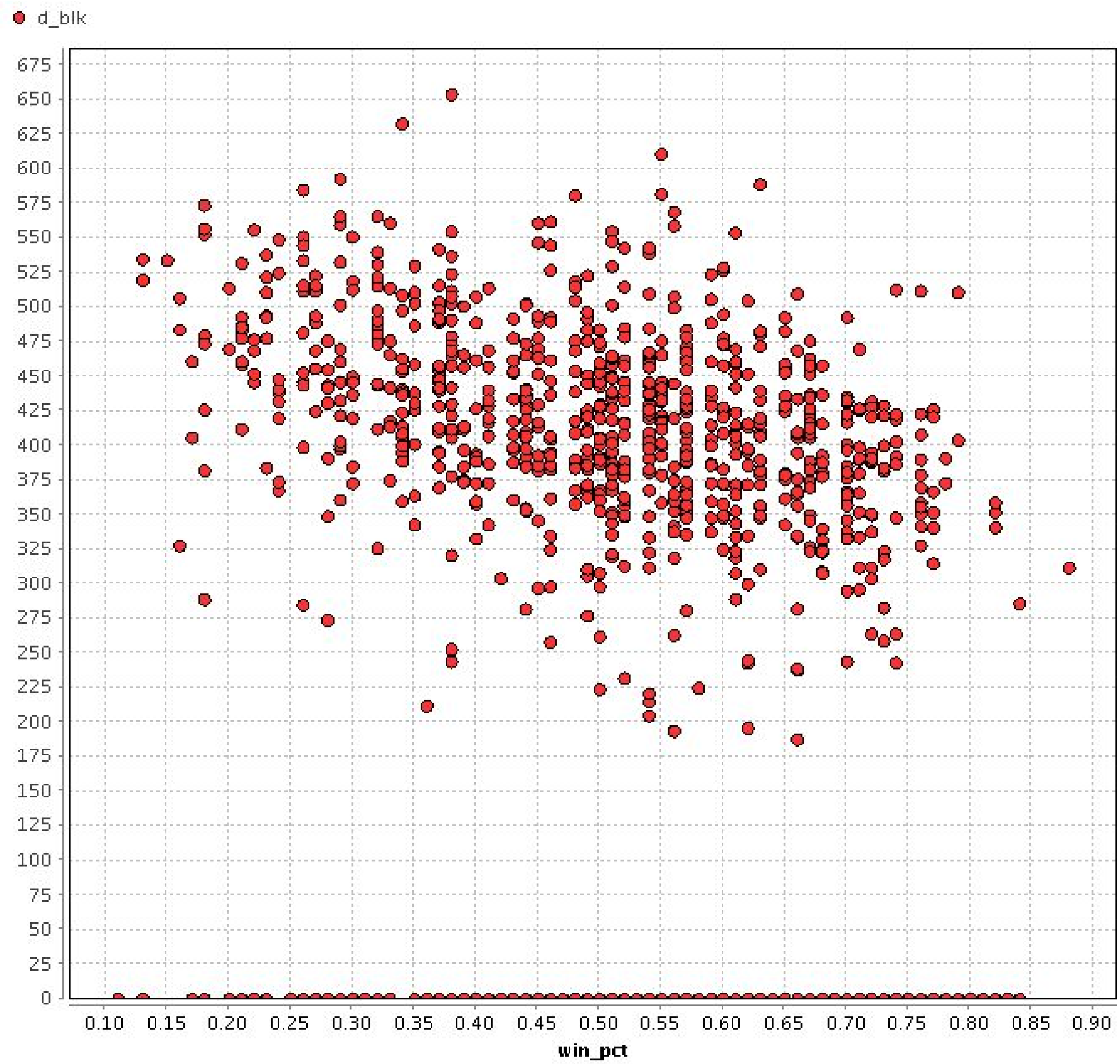
- Home Team Free Throws Attempted



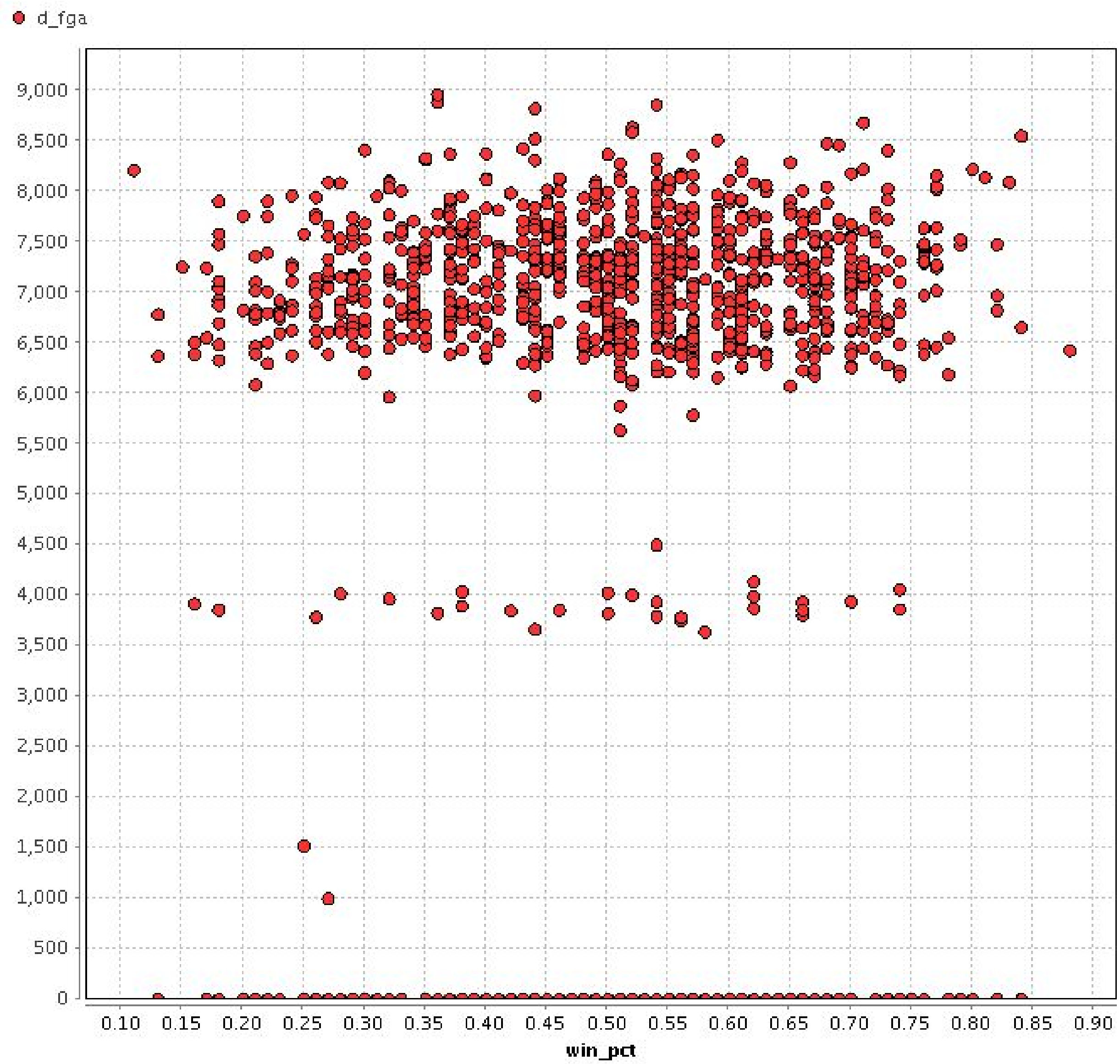
- Home Team 3 Points Goal Made



- Home Team Blocks



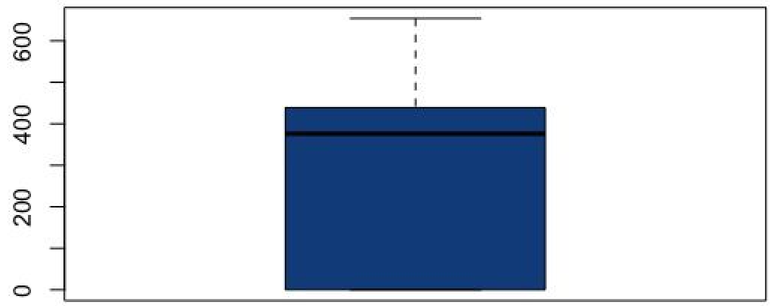
- Home Team Field Goals Attempted.



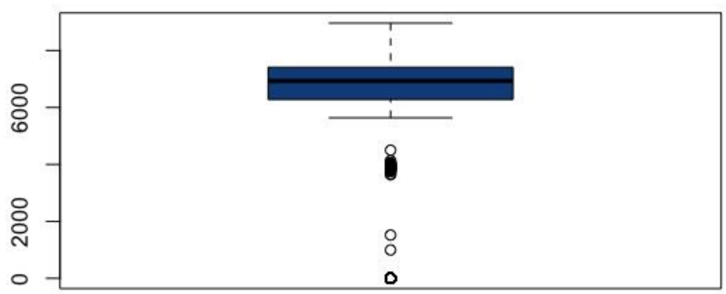
OUTLIER DETECTION

We have performed outlier detection in R programming language by box plots. Outliers are observations which are distant from other observations that is it does not follow the pattern which other observations in the attribute follow. In a boxplot, an outlier is defined as data point located outside the fences of the boxplot. The following are the outlier checks using boxplot for Home Team Blocks,Home Team Field Goals Attempted, Home Team Free Throws Attempted and Home Team 3 Point Goals Made.

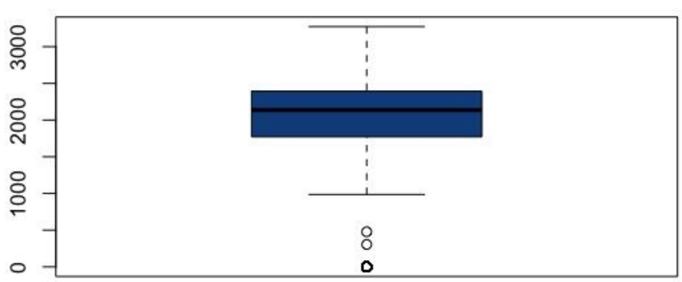
Outlier Check for Home Team Blocks

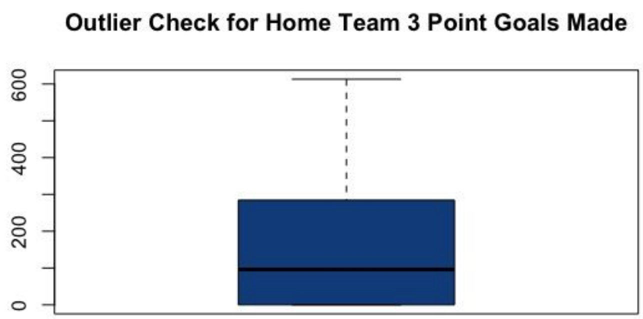


Outlier Check for Home Team Field Goals Attempted



Outlier Check for Home Team Free Throws Attempted





LINEAR REGRESSION MODEL

Linear regression is the basic model used in predictive analysis which is quite commonly used. Its attempts to model the relationship between two variables by fitting a linear equation to the observations. Regression estimates are used to describe the data and explain the dependent or outcome variable with one or more independent or explanatory variables. In regression we try to fit a single line through the scatter plots.The most simplest form is

$$y = c + b*x$$

Where y= estimated dependent variable
c= constant
b= regression coefficient
x= independent variable.

In our project we use linear regression to predict the win percentage. This is done in R programming language. WIN PERCENTAGE can be calculated using the following formula,
WIN Percent= Total Games Won / (Total Number of Games Played)

Total Number of Games Played can also be calculated as

Total Number of Games Played = Total Games Won + Total Games Lost.

This Win Percent indicates the percent of wins of a team for a season. The following steps were performed in prediction.

1) **Data Normalization**

We normalize the data as different attributes can have different ranges. In order to bring all the attributes to a same range we perform data normalization.

2) <u>Variable Selection</u>

a) Stepwise Selection

In stepwise regression procedure we build the regression model from a set of selected variables by entering and removing predictor variables in a stepwise manner into the model until there is no reason for us to enter or remove any other variables. The list of predictor variables must include all variables which actually predict the response or outcome variable.

The variables which we consider are

- FG
- FGA
- THREEP
- FT
- FTA
- ORB
- AST
- STL
- TOV
- OFGM
- OFGA
- OFTM
- O3PM
- OppORB
- OppDRB
- OppBlk

```
          Df Sum of Sq    RSS    AIC
<none>                 1899.6 526.19
+ THREEPA  1      15.9 1883.7 526.21
+ OppTOV   1      12.5 1887.0 526.63
+ PF       1      11.1 1888.5 526.81
- AST      1      22.0 1921.6 526.91
- OppORB   1      23.7 1923.3 527.12
- OppBlk   1      27.5 1927.1 527.58
+ OFTA     1       4.9 1894.7 527.58
+ O3PA     1       4.8 1894.8 527.59
+ BLK      1       4.4 1895.1 527.64
+ OppPF    1       2.1 1897.4 527.92
+ OppAsst  1       0.8 1898.8 528.09
+ DRB      1       0.3 1899.3 528.15
+ OppSTL   1       0.2 1899.4 528.17
- STL      1      52.7 1952.3 530.65
- FTA      1      95.7 1995.2 535.79
- TOV      1     100.9 2000.5 536.41
- OFGA     1     120.3 2019.9 538.69
- OppDRB   1     135.2 2034.7 540.41
- ORB      1     349.6 2249.2 564.06
- FGA      1     413.2 2312.8 570.64
- O3PM     1     422.7 2322.3 571.61
- OFTM     1     786.9 2686.4 605.99
- FT       1     802.0 2701.5 607.31
- THREEP   1    2037.7 3937.3 696.20
- FG       1    2554.9 4454.5 725.33
- OFGM     1    5477.7 7377.3 844.39

Call:
lm(formula = W ~ FG + FGA + THREEP + FT + FTA + ORB + AST + STL +
    TOV + OFGM + OFGA + OFTM + O3PM + OppORB + OppDRB + OppBlk,
    data = nbaData)

Coefficients:
(Intercept)           FG          FGA       THREEP           FT          FTA          ORB
  59.873187     0.074412    -0.026224     0.032381     0.038810    -0.010941     0.031601
        AST          STL          TOV         OFGM         OFGA         OFTM         O3PM
   0.002961     0.011411    -0.012117    -0.065153     0.010574    -0.023913    -0.026497
     OppORB       OppDRB       OppBlk
  -0.007519     0.014700    -0.008985
```

b) PCA

Principal Component Analysis is a statistical method which uses orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables known as principal components. The number of principal components is less than or equal to the number of original variables.

In our project we performed PCA on 15 variables of home and opponent teams like Free Throws made, 3 point Goals made, etc. We considered the first 3 principal components which described 98% of all the characteristics of the home team. This is true eigenvector based multivariate analysis and is very closely related to factor analysis. This was performed in RapidMiner Studio

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	4083.249	0.851	0.851
PC 2	1245.613	0.079	0.930
PC 3	996.531	0.051	0.980
PC 4	432.569	0.010	0.990
PC 5	316.456	0.005	0.995
PC 6	175.662	0.002	0.997
PC 7	144.051	0.001	0.998
PC 8	126.026	0.001	0.998
PC 9	103.702	0.001	0.999
PC 10	80.613	0.000	0.999
PC 11	63.526	0.000	1.000
PC 12	61.230	0.000	1.000
PC 13	49.476	0.000	1.000
PC 14	40.518	0.000	1.000
PC 15	22.643	0.000	1.000

ExampleSet (1187 examples, 2 special attributes, 3 regular attributes)

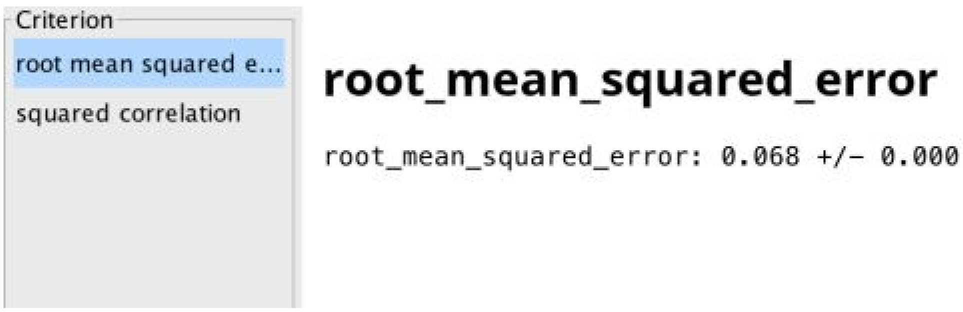
Row No.	team	winpercent	pc_1	pc_2	pc_3
1	BOS	0.367	-8418.875	2409.059	1767.616
2	CH1	0.639	-8326.453	1947.585	1449.924
3	CL1	0.500	-8352.836	2079.319	1540.614
4	DE1	0.333	-8415.961	2394.511	1757.601
5	NYK	0.550	-8428.586	2457.550	1800.999
6	PH1	0.583	-8416.933	2399.361	1760.939
7	PIT	0.250	-8393.625	2282.982	1680.821
8	PRO	0.467	-8329.528	1962.941	1460.495
9	ST1	0.623	-8417.256	2400.977	1762.052
10	TO1	0.367	-8360.605	2118.112	1567.320
11	WSC	0.817	-8429.558	2462.399	1804.337
12	BA1	0.583	-8502.394	2826.082	2054.707
13	BOS	0.417	-8485.237	2740.414	1995.731
14	CH1	0.583	-8481.353	2721.018	1982.378
15	NYK	0.542	-8495.434	2791.330	2030.783
16	PH1	0.562	-8489.931	2763.852	2011.866
17	PRO	0.125	-8427.777	2453.509	1798.217
18	ST1	0.604	-8510.164	2864.875	2081.413
19	WSC	0.583	-8497.700	2802.645	2038.572
20	BA1	0.483	-8251.836	1575.012	1193.434
21	BOS	0.417	-8278.057	1705.938	1283.567
22	CH1	0.633	-8273.201	1681.693	1266.876
23	FTW	0.367	-8297.480	1802.920	1350.333
24	INJ	0.300	-8279.028	1710.787	1286.906

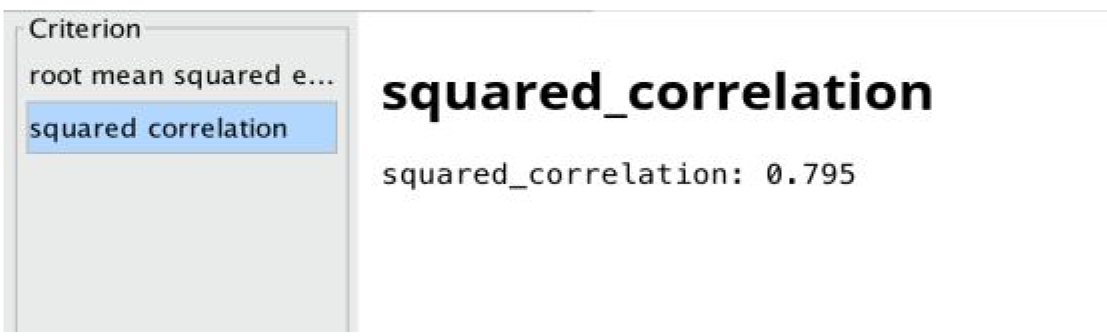
3) Segmenting the data into training and validation set data.

We segment the data into training and validation set in order to train a number of models on the training data and then test the models on the validation set data and choose the best model.

MODEL RESULTS

The RMSE & Adjusted R- Squared values when PCA Attributes were used:





The RMSE & Adjusted R- Squared values when stepwise variable selection was used:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.2208 -2.0304  0.0067  1.9612  8.1020 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 59.873187   8.550778   7.002 3.05e-11 ***
FG           0.074412   0.004336  17.162  < 2e-16 ***
FGA         -0.026224   0.003799  -6.902 5.46e-11 ***
THREEP       0.032381   0.002113  15.327  < 2e-16 ***
FT           0.038810   0.004036   9.615  < 2e-16 ***
FTA         -0.010941   0.003294  -3.321 0.001050 ** 
ORB          0.031601   0.004978   6.349 1.23e-09 ***
AST          0.002961   0.001860   1.592 0.112730    
STL          0.011411   0.004627   2.466 0.014436 *  
TOV         -0.012117   0.003553  -3.411 0.000771 ***
OFGM        -0.065153   0.002593 -25.130  < 2e-16 ***
OFGA         0.010574   0.002839   3.725 0.000249 ***
OFTM        -0.023913   0.002511  -9.525  < 2e-16 ***
O3PM        -0.026497   0.003796  -6.981 3.45e-11 ***
OppORB      -0.007519   0.004546  -1.654 0.099588 .  
OppDRB       0.014700   0.003724   3.948 0.000106 ***
OppBlk      -0.008985   0.005046  -1.780 0.076391 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.945 on 219 degrees of freedom
Multiple R-squared:  0.9438,	Adjusted R-squared:  0.9397
```

LINEAR REGRESSION RESULT

The following variables determine the result of the game better than other variables:

Game Metric	Wins per 100
My Shooting Ability (FGs, 3Ps etc)	**21.7**
Their Shooting ability (OFGs, O3Ps etc)	**16.3**
Free throws (FTs)	**9.6**
Rebounding (ORBs)	**6.5**
Their Free throws (OFTs)	**3.1**
Netplay (TOVs, ASTs)	**2.8**
Taking The Ball (STLs, BLKs)	**1.5**

R CODE

```
nbaData_team <- team_season

nbaData_team$win_per <- nbaData_team$won/(nbaData_team$won+nbaData_team$lost)

normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }

nbaData_team$o_fgm <- normalize(nbaData_team$o_fgm)
nbaData_team$o_fga <- normalize(nbaData_team$o_fga)
nbaData_team$o_ftm <- normalize(nbaData_team$o_ftm)
nbaData_team$o_fta <- normalize(nbaData_team$o_fta)
nbaData_team$o_oreb <- normalize(nbaData_team$o_oreb)
nbaData_team$o_dreb <- normalize(nbaData_team$o_dreb)
nbaData_team$o_reb <- normalize(nbaData_team$o_reb)
nbaData_team$o_asts <- normalize(nbaData_team$o_asts)
nbaData_team$o_pf <- normalize(nbaData_team$o_pf)
```

```
nbaData_team$o_stl <- normalize(nbaData_team$o_stl)
nbaData_team$o_to <- normalize(nbaData_team$o_to)
nbaData_team$o_blk <- normalize(nbaData_team$o_blk)
nbaData_team$o_3pm <- normalize(nbaData_team$o_3pm)
nbaData_team$o_3pa <- normalize(nbaData_team$o_3pa)
nbaData_team$o_pts <- normalize(nbaData_team$o_pts)

nbaData_team$d_fgm <- normalize(nbaData_team$d_fgm)
nbaData_team$d_fga <- normalize(nbaData_team$d_fga)
nbaData_team$d_ftm <- normalize(nbaData_team$d_ftm)
nbaData_team$d_fta <- normalize(nbaData_team$d_fta)
nbaData_team$d_oreb <- normalize(nbaData_team$d_oreb)
nbaData_team$d_dreb <- normalize(nbaData_team$d_dreb)
nbaData_team$d_reb <- normalize(nbaData_team$d_reb)
nbaData_team$d_asts <- normalize(nbaData_team$d_asts)
nbaData_team$d_pf <- normalize(nbaData_team$d_pf)
nbaData_team$d_stl <- normalize(nbaData_team$d_stl)
nbaData_team$d_to <- normalize(nbaData_team$d_to)
nbaData_team$d_blk <- normalize(nbaData_team$d_blk)
nbaData_team$d_3pm <- normalize(nbaData_team$d_3pm)
nbaData_team$d_3pa <- normalize(nbaData_team$d_3pa)
nbaData_team$d_pts <- normalize(nbaData_team$d_pts)

nbaData_team$win_per <- normalize(nbaData_team$win_per)

set.seed(214)
ind <- sample(2, nrow(nbaData_team), replace=TRUE, prob=c(0.75, 0.25))
nba_TrainingData <- nbaData_team[ind==1,]
nba_TestData <- nbaData_team[ind==2,]


model_team_null = lm(win_per ~ 1, data = nba_TrainingData) # Includes only the intercept
model_team_full = lm(win_per ~ d_fgm + d_fga + d_fta + d_oreb + d_dreb + d_asts + d_stl +
d_blk + d_to + d_pf + d_3pm + d_3pa + d_pts + o_fgm + o_fga + o_ftm + o_fta + o_oreb +
o_dreb + o_asts + o_stl + o_blk + o_to + o_pf + o_3pm + o_3pa + o_pts, data =
nba_TrainingData)
step(model_team_null, scope=list(lower=model_team_null, upper=model_team_full),
direction="both")
```

```
model_team_stepvars <- lm(formula = win_per ~ d_pts + o_pts + o_3pm + d_3pm, data =
nba_TrainingData)
summary(model_team_stepvars)

test_data_pred <- predict(model_team_stepvars, newdata = nba_TestData)
summary(test_data_pred)

test_data_pred_nba <- table(test_data_pred, nba_TestData$win_per)
test_data_pred_nba

test_data <- cbind(nba_TestData, as.data.frame(test_data_pred))
write.csv(test_data, "D:/Desktop/StatsProject/test_data.csv")
```

CONCLUSION

We can also do further analysis on the dataset by obtaining answers to the following questions

- Which team will win a season or championship?
- Is there a trend or seasonal variation in the pattern of teams with higher winning percentage?
- Which are the major opponents of each team?
- Career of major or popular NBA players and how they are influential to their teams?
- Which are the factors that are influential or affecting the way of play of teams?

Also our model predicted the win percentage of Philadephia 76ers as 90% for the 1967 season. Their actual win-loss ratio was 68-13 which gives their win percentage as 83.95% which is fairly close.

REFERENCES

http://www.basketball-reference.com
https://en.wikipedia.org/wiki/R_(programming_language)
http://docs.rapidminer.com/
https://www.r-statistics.com/tag/boxplot-outlier/
http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
https://onlinecourses.science.psu.edu/stat501/node/329