PREDICTING BASKETBALL GAMES

Brian Chirn

BASKETBALL 101

Game played 5 on 5

Team scores points that getting the basketball into the hop

Team with the most points wins

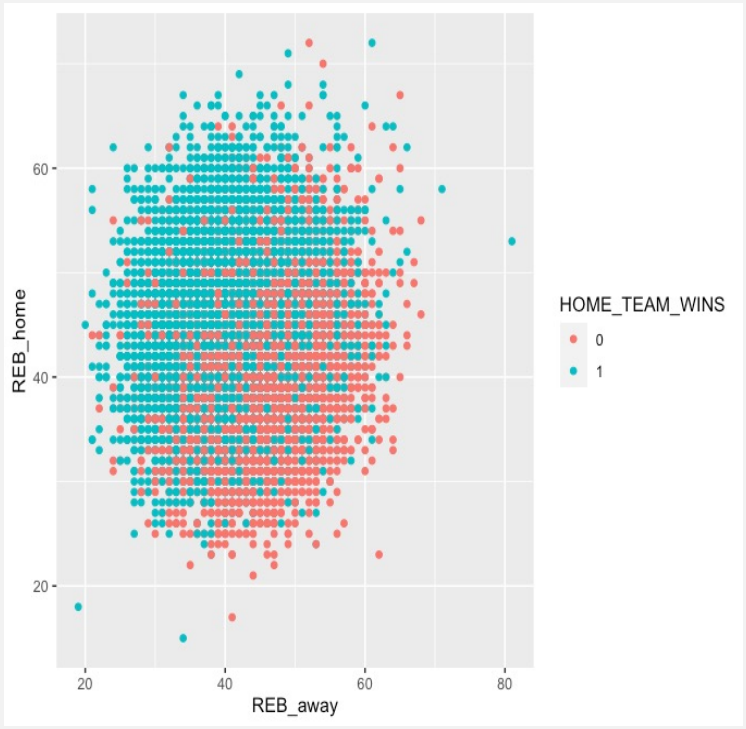
DESCRIBE THE DATA

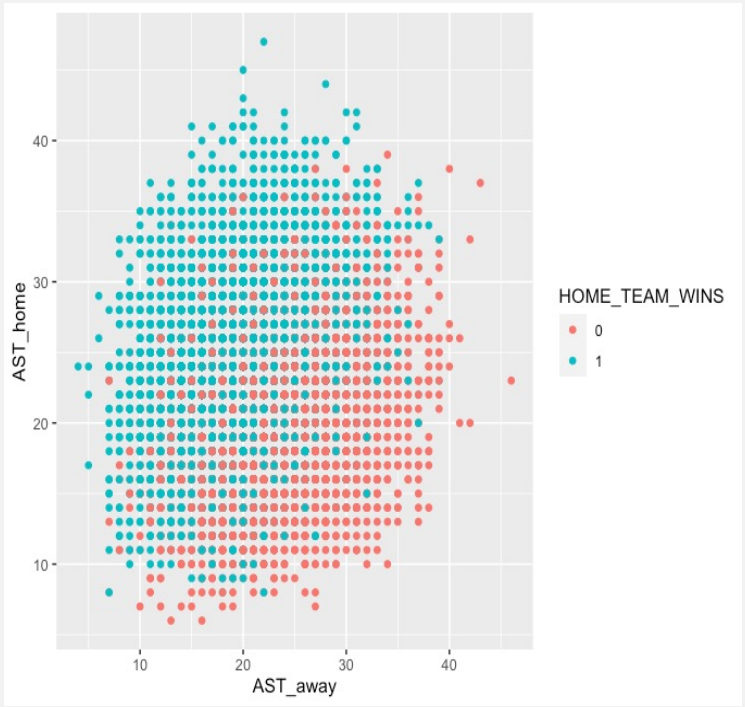
Data collected has every game between 2004-2020 taken from nba website

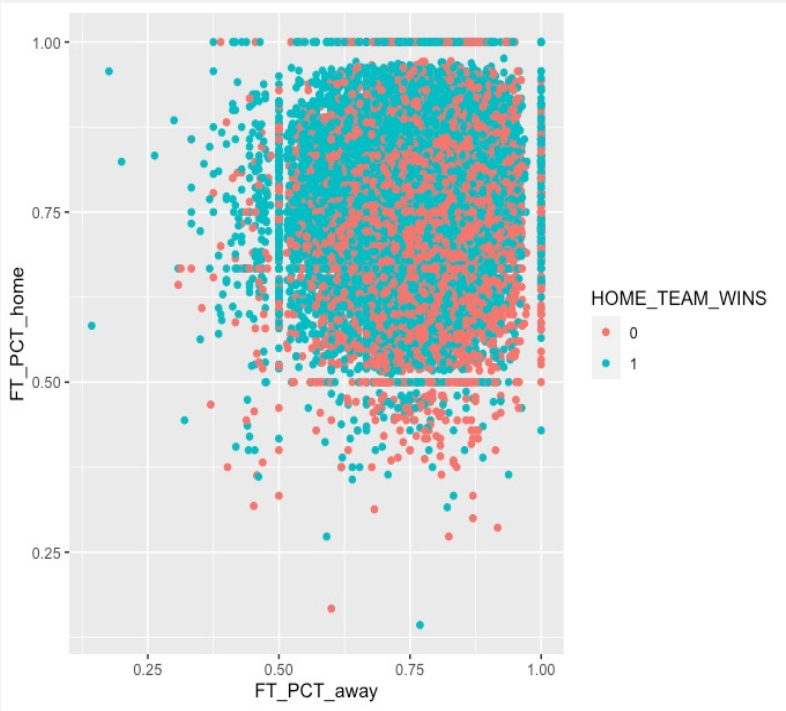
Contains stats like Points, Assist, Rebounds, Field Goal Percentage etc.. For Both the Home Team and Visiting team

```
> head(dat)
  GAME_DATE_EST  GAME_ID GAME_STATUS_TEXT HOME_TEAM_ID VISITOR_TEAM_ID SEASON TEAM_ID_home PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home
1    2020-03-01 21900895            Final   1610612766      1610612749   2019   1610612766        85       0.354       0.900        0.229       22
2    2020-03-01 21900896            Final   1610612750      1610612742   2019   1610612750        91       0.364       0.400        0.310       19
3    2020-03-01 21900897            Final   1610612746      1610612755   2019   1610612746       136       0.592       0.805        0.542       25
4    2020-03-01 21900898            Final   1610612743      1610612761   2019   1610612743       133       0.566       0.700        0.500       38
5    2020-03-01 21900899            Final   1610612758      1610612765   2019   1610612758       106       0.407       0.885        0.257       18
6    2020-03-01 21900900            Final   1610612740      1610612747   2019   1610612740       114       0.421       0.818        0.219       24
  REB_home TEAM_ID_away PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away HOME_TEAM_WINS
1       47   1610612749       93       0.402       0.762        0.226       20       61              0
2       57   1610612742      111       0.468       0.632        0.275       28       56              0
3       37   1610612755      130       0.505       0.650        0.488       27       37              1
4       41   1610612761      118       0.461       0.897        0.263       24       36              1
5       51   1610612765      100       0.413       0.667        0.429       23       42              1
6       52   1610612747      122       0.515       0.900        0.371       23       36              0
>
```

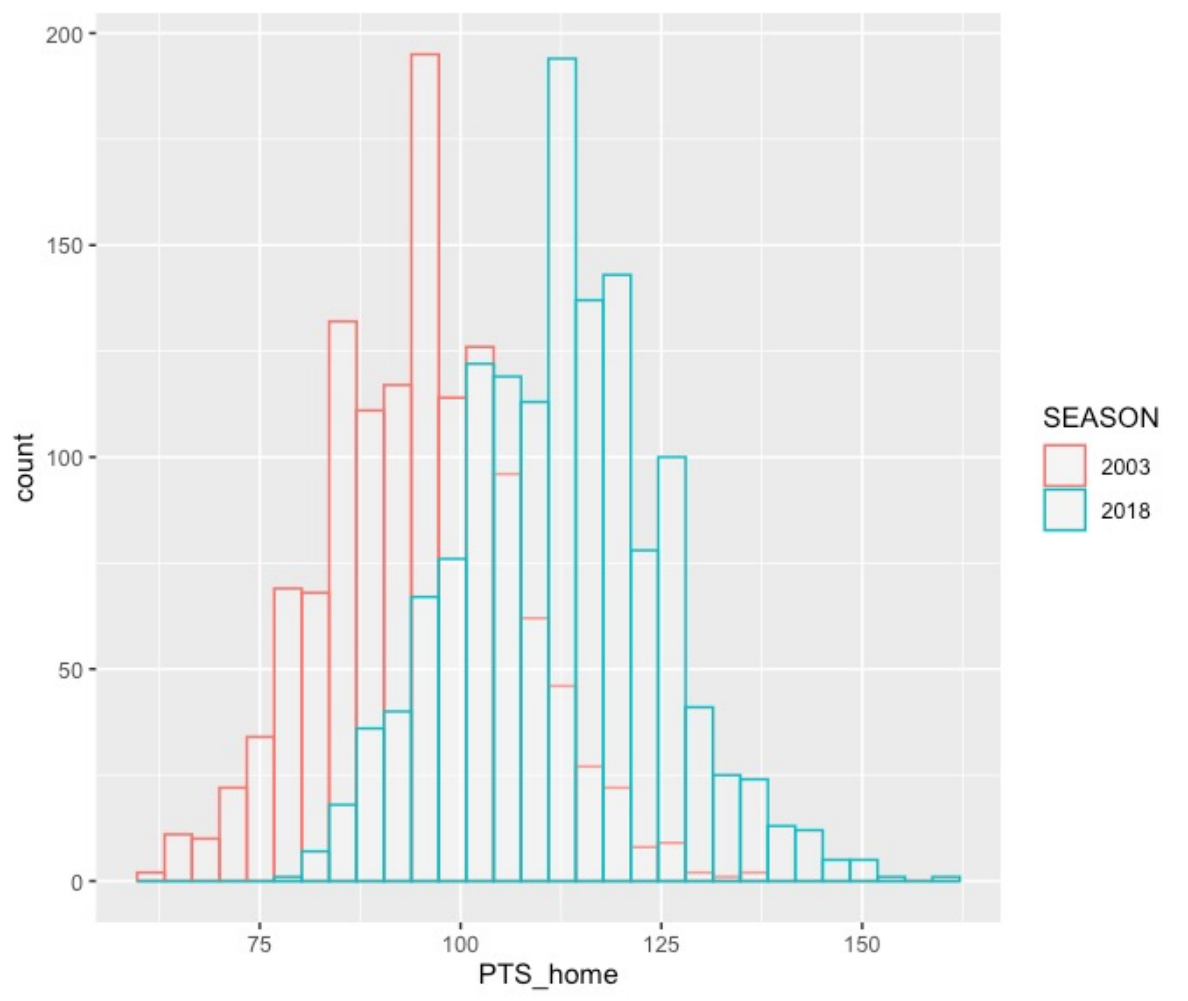
EDA: REBOUNDS AND ASSISTS CORRELATE WITH WINNING; FREE THROW PRECENTAGE UNCLEAR



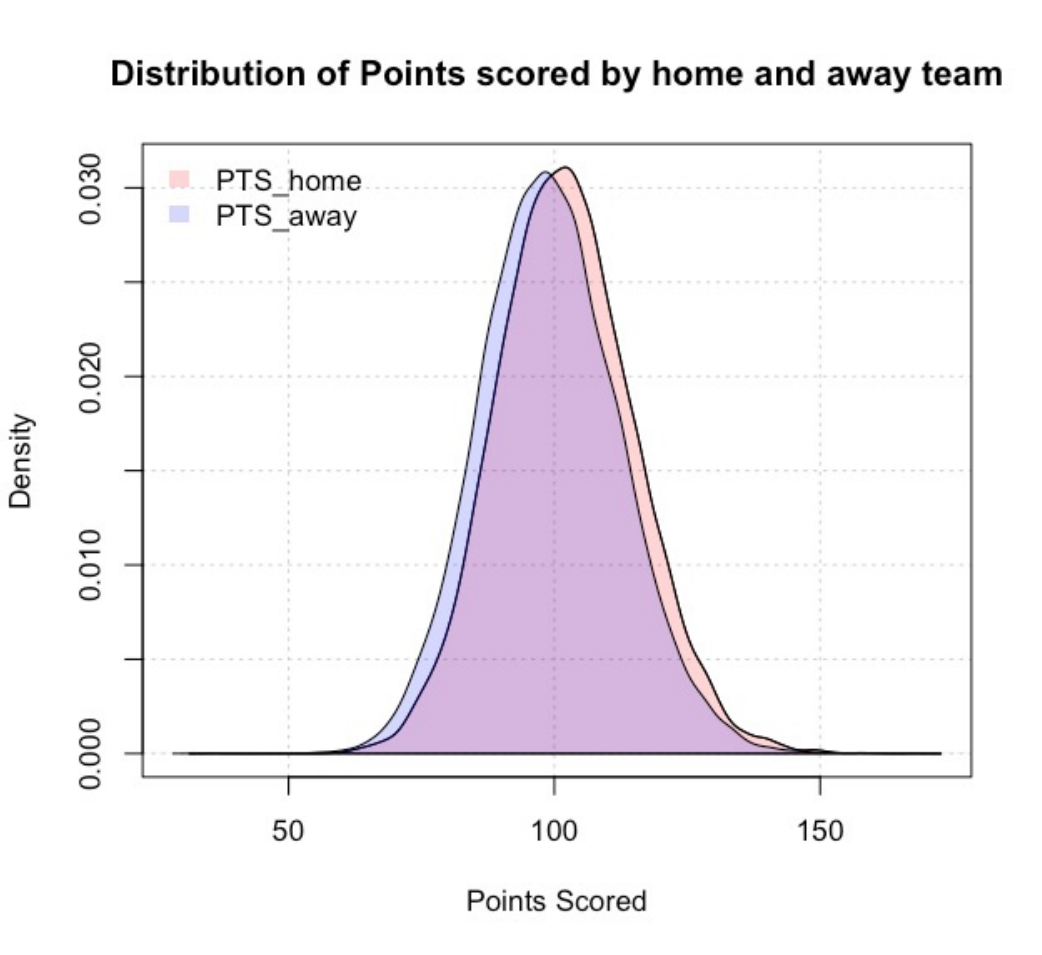


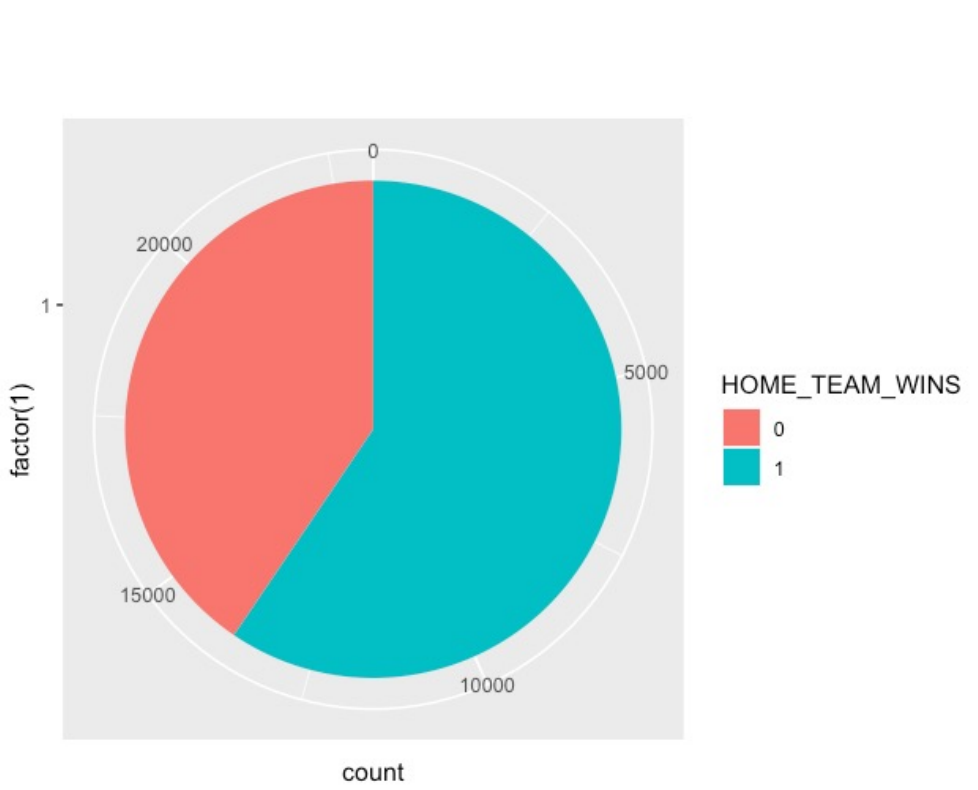


EDA: THE MODERN NBA IS SCORING MORE POINTS PER GAME THAN 15 YEARS AGO

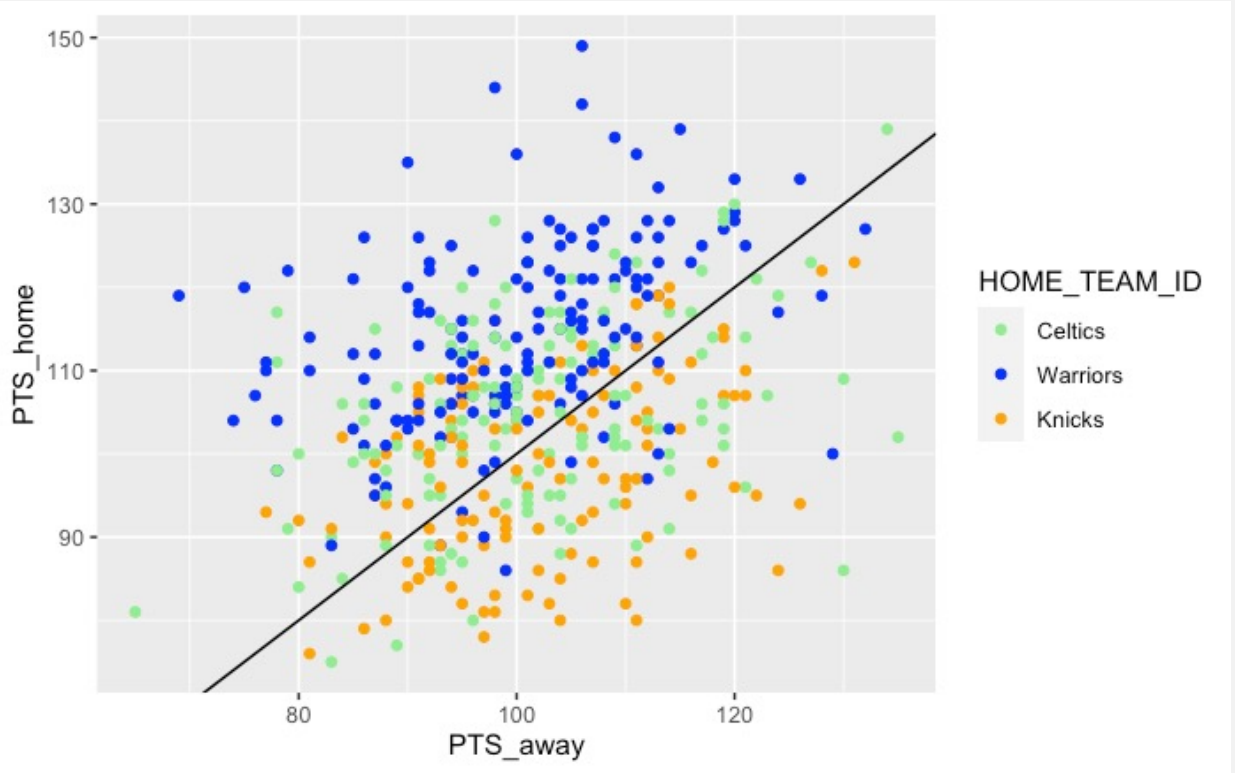


EDA: HOME COURT ADVANTAGE VISUALIZED. HOME TEAM SCORES MORE POINTS ON AVERAGE AND WINS 59% OF THE GAMES





EDA: WARRIORS DOMINANCE (2014-2016).



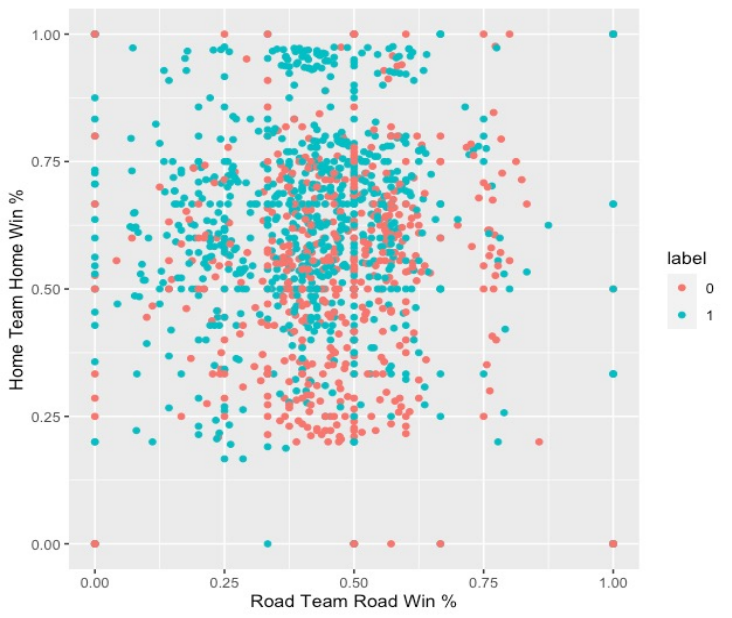
CREATING A NEW DATASET

- We want our model to predict the wins BEFORE a game is played.

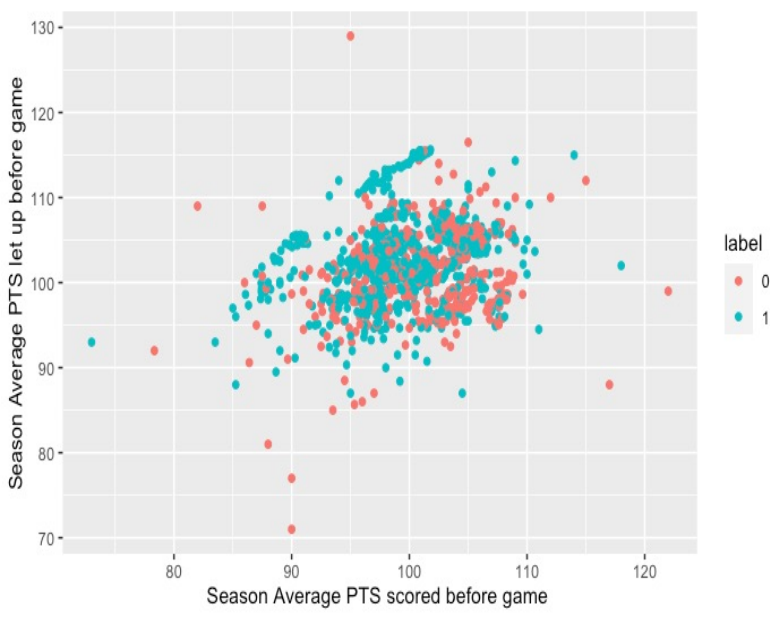
- **Calculate the average stats for home team and visiting team for each game played**

```
> colnames(df_before)
 [1] "date"                    "HOME_TEAM_ID"            "PTS_home"                "FG_PCT_home"             "FT_PCT_home"
 [6] "FG3_PCT_home"            "AST_home"                "REB_home"                "PTS_away"                "FG_PCT_away"
[11] "FT_PCT_away"             "FG3_PCT_away"            "AST_away"                "REB_away"                "HOME_WIN_PCT"
[16] "Visitor_VISITOR_TEAM_ID" "Visitor_PTS_away"        "Visitor_FG_PCT_away"     "Visitor_FT_PCT_away"     "Visitor_FG3_PCT_away"
[21] "Visitor_AST_away"        "Visitor_REB_away"        "Visitor_PTS_home"        "Visitor_FG_PCT_home"     "Visitor_FT_PCT_home"
[26] "Visitor_FG3_PCT_home"    "Visitor_AST_home"        "Visitor_REB_home"        "ROAD_WIN_PCT"            "label"
>
```

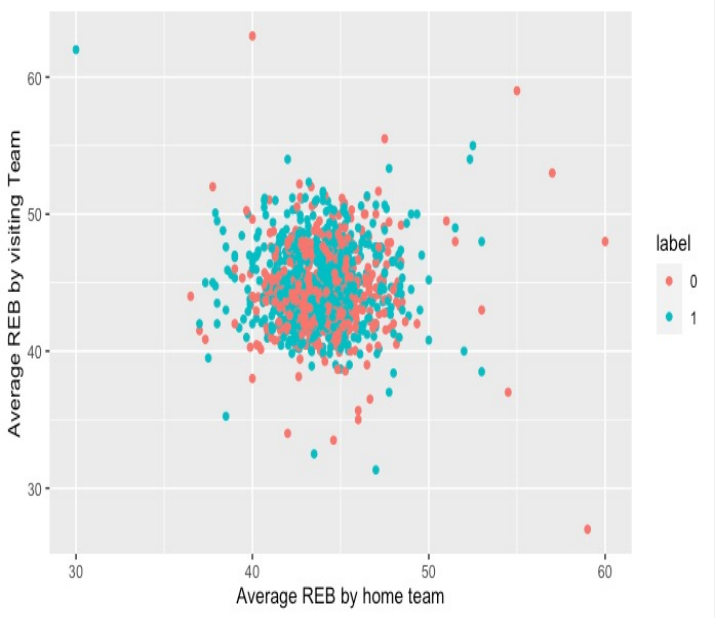
BEFORE EACH GAME, SEASON STATS
HOME AND ROAD WIN%



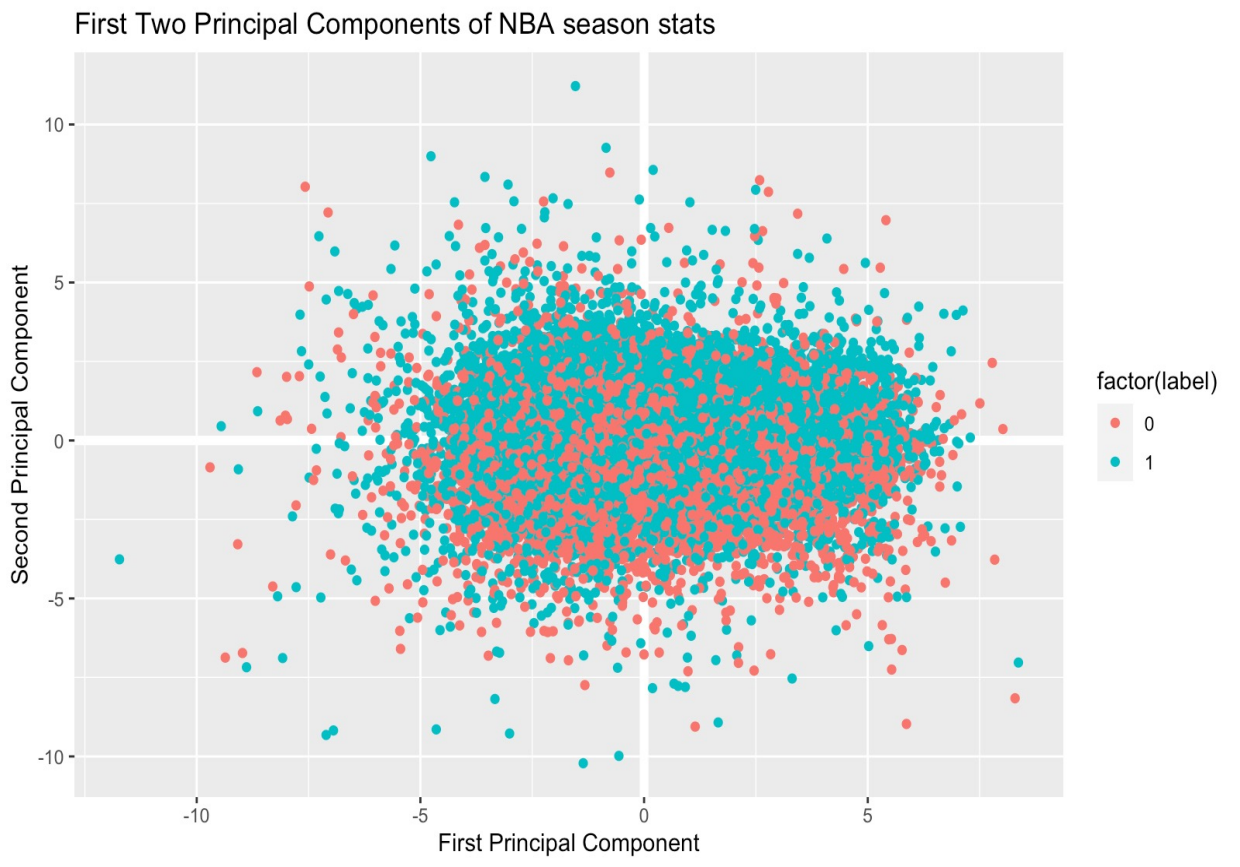
BEFORE EACH GAME, SEASON STATS
PTS SCORED VS PTS LET UP



BEFORE GAME AVERAGE REBOUNDS BY HOME TEAM AND AWAY TEAM

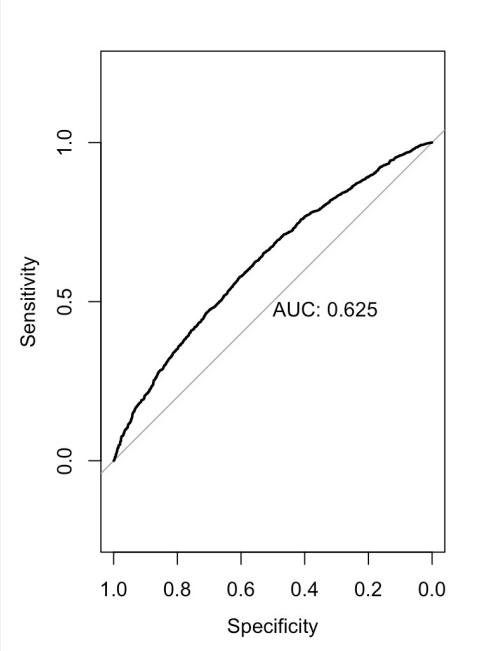


PCA

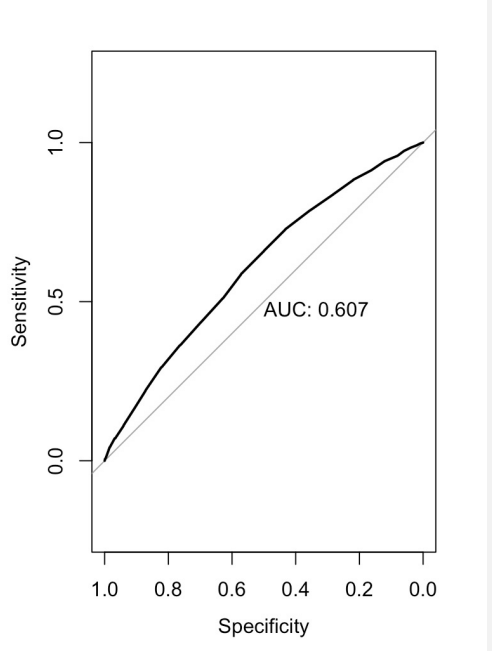

First Two Principal Components of NBA season stats
Second Principal Component
First Principal Component
10
5
0
-5
-10
-10
-5
0
5
factor(label)
0
1

LOGISTIC REGRESSION AND KNN (15 DAY AVERAGE)

	Logistic Regression	KNN
AUC	.625	.607
Accuracy	61.64%	61.62%



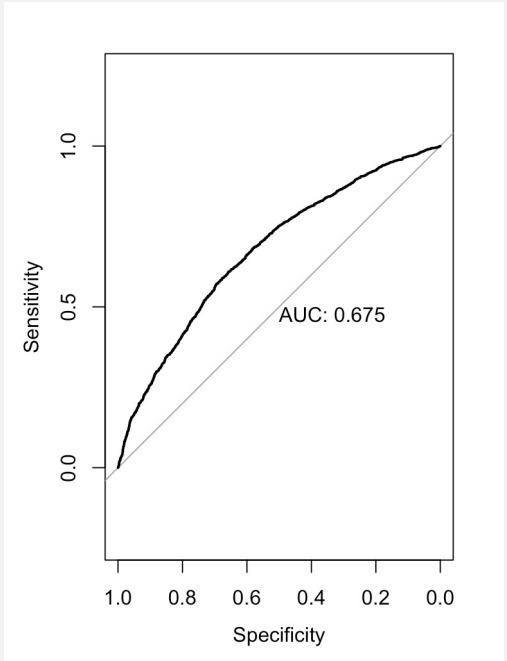
Logistic Regression



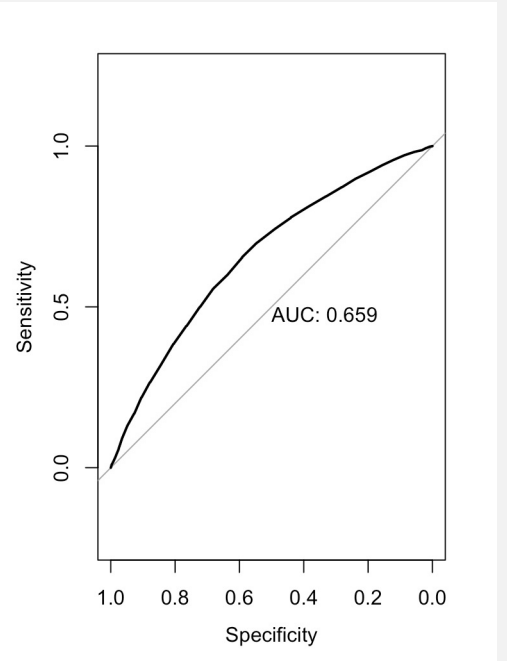
KNN

LOGISTIC REGRESSION AND KNN (SEASON AVERAGE)

	Logistic Regression	KNN
AUC	0.675	0.659
Accuracy	64.46%	64.31%



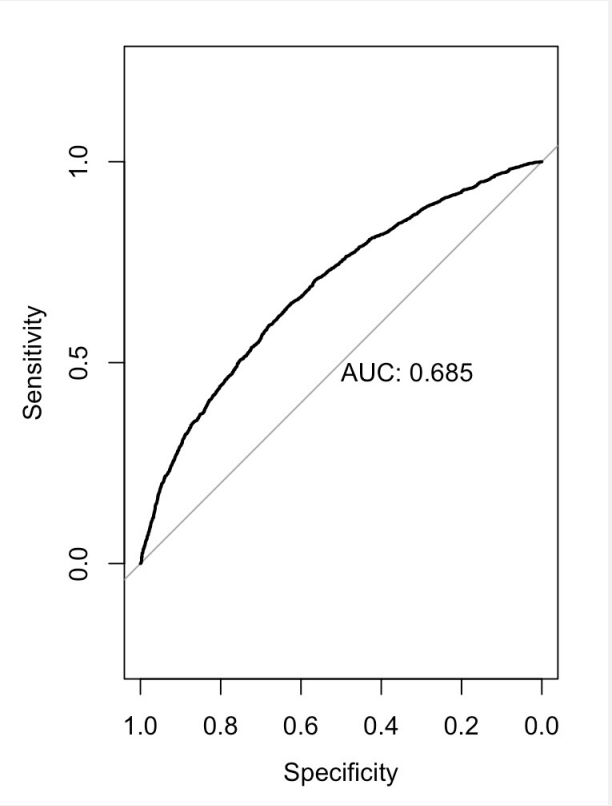
Logistic Regression

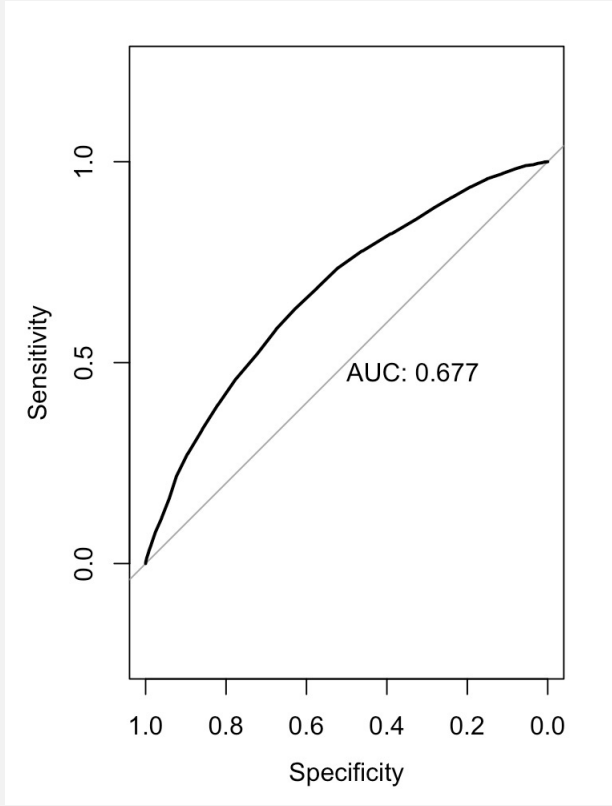


KNN

LOGISTIC REGRESSION AND KNN WITH PREVIOUS SEASON INCLUDED

	Logistic Regression	KNN
AUC	0.685	0.677
Accuracy	64.43%	64.30%





SUMMARY

	LR 15	**LR CurrentSeason**	**LR Previous Season + Current Season**	**KNN 15**	**KNN Current Season**	KNN Previous Season + Current Season
AUC	.625	0.675	**0.685**	.607	0.659	0.677
Accuracy	61.64%	**64.46%**	64.43%	61.62%	64.31%	64.30%

COMPARISONS

Another student who had a similar idea to me had accuracy results (62.07% with SVM, 63.75% LR, 64.95%) [1]

A group was able to achieve ~70% accuracy by incorporating individual player statistics rather than looking at team data as well as looking at previous season statistics. [2]

NBA experts predict games correctly ~70% of the time as well. [2]

1) http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf

2)https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf

CONCLUSION

Using the previous and current team statistics we developed a Logistic regression model with ~65% accuracy.

In the future, I would like to incorporate player data into the model, specifically examining the strength of the starting players and bench players.