Predicting a NBA Champion Using Historical Data

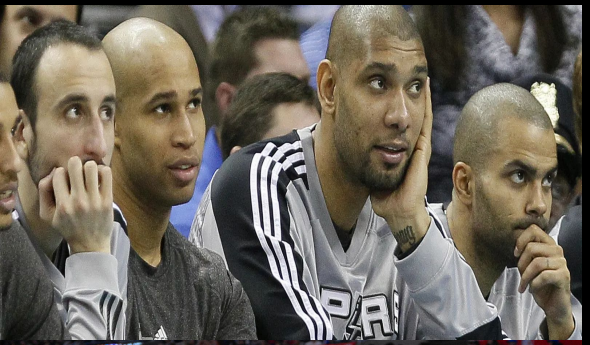
Ray LaForte
Rebecca Burnett
Nebe Samuel

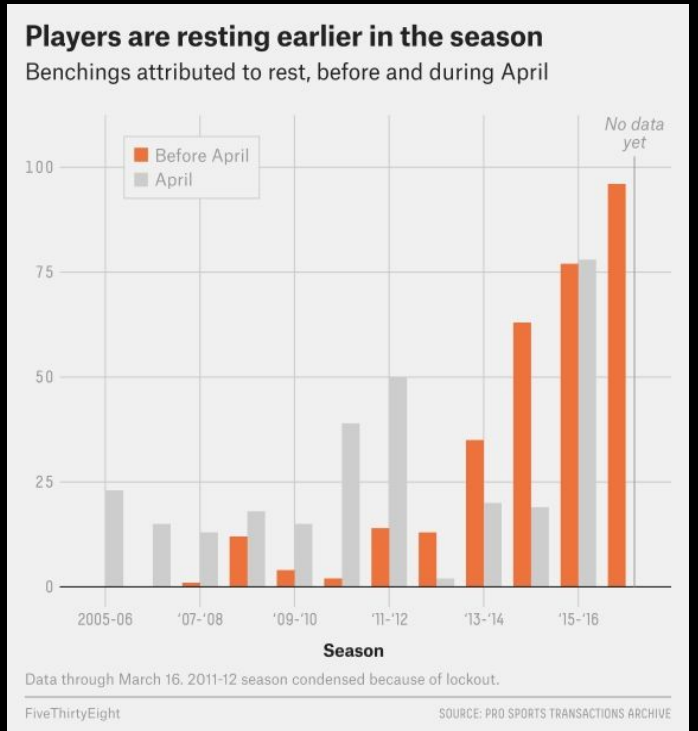
Project Pivot

- Our original goal
  - Use historical NBA team data to predict the outcome of the current NBA season
- Hurdles
  - Gaining access to team statistics for the current “bubble” season was expensive and ultimately unattainable
    - Covid + its implications
- Project Pivot
  - Utilize 4 seasons of historical NBA team data to create a model that can accurately predict the winner of the NBA championship

Analytics is TRANSFORMING the NBA

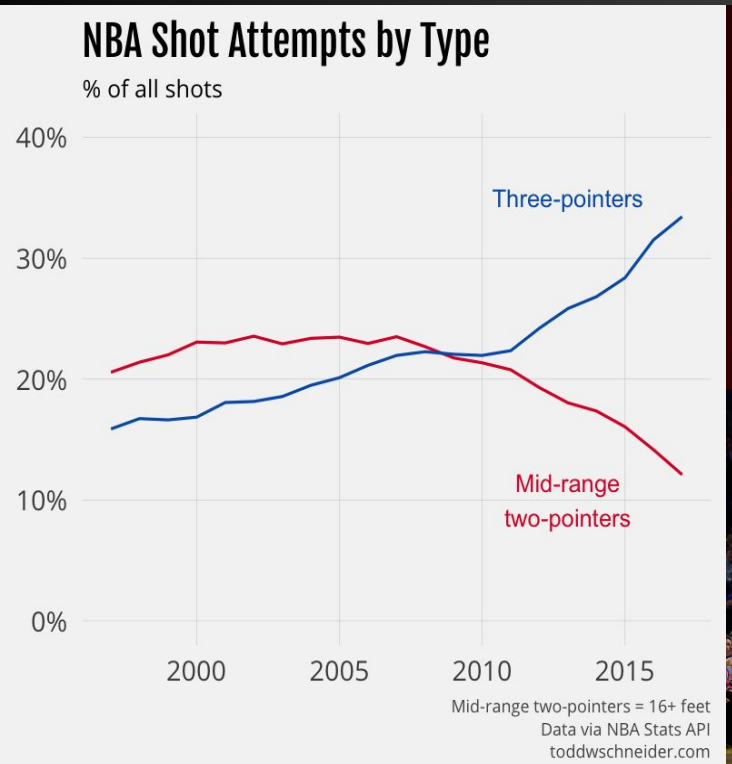
- Load Management

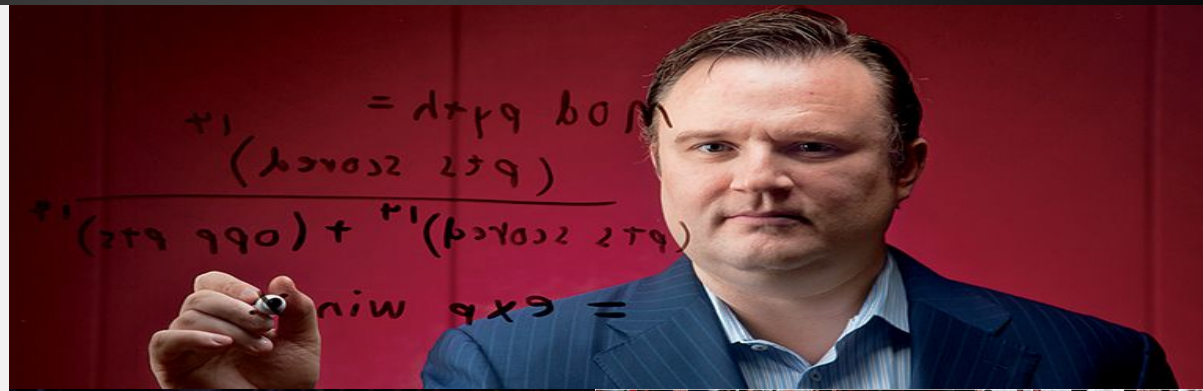




ANALYTICS, ANALYTICS, ANALYTICS

- Shot Selection + Team Makeup

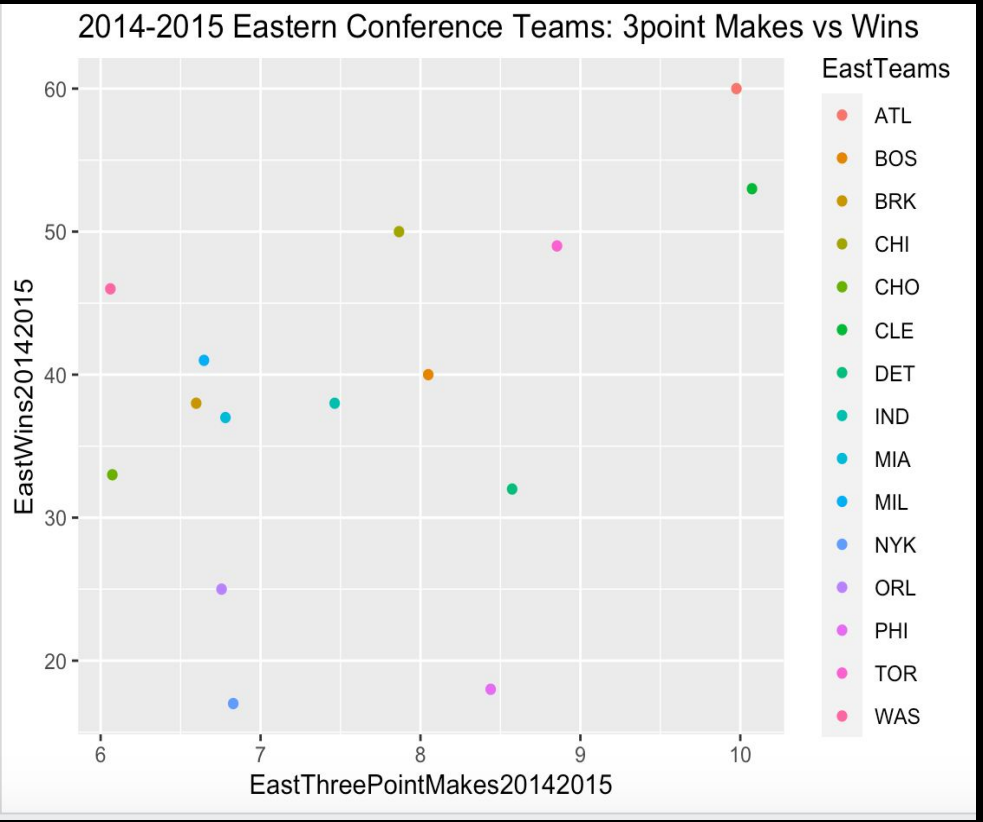


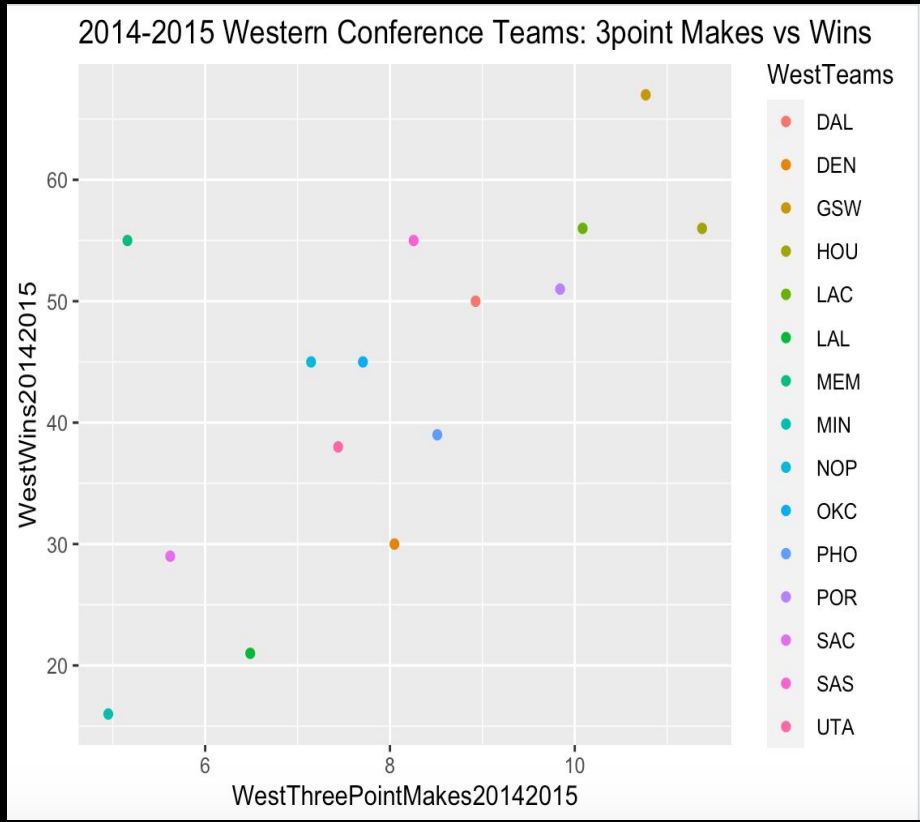


Dataset

- 2014-2018 Regular Season Data Set
  - 9840 observations with 41 variables
  - The dataset contains data on every team and every game played in the regular season from 2014-2018
- Granularity
  - One row in the dataset represents one team's stats against one opponent in an NBA Season
- Dimensions:
  - Team, Date, Home or Away, Opponent, Win or Loss, Team Points, Opponent Points, Field Goal %, 3 Point %, Amt of Free Throws, Free Throw Shooting %,Offensive Rebounds, Total Rebounds, Assists, Steals, Blocks, Turnovers, Total Fouls
  - The Dataset contains all dimension listed above for both the home team and the away team.

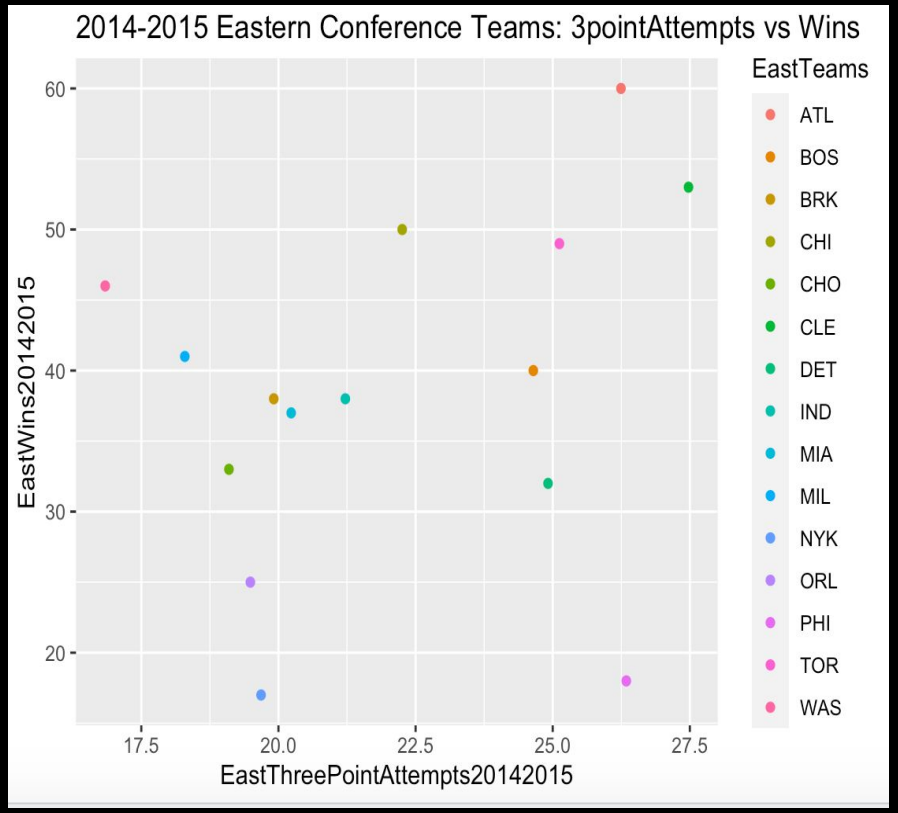
EDA: ScatterPlots for Each Conference



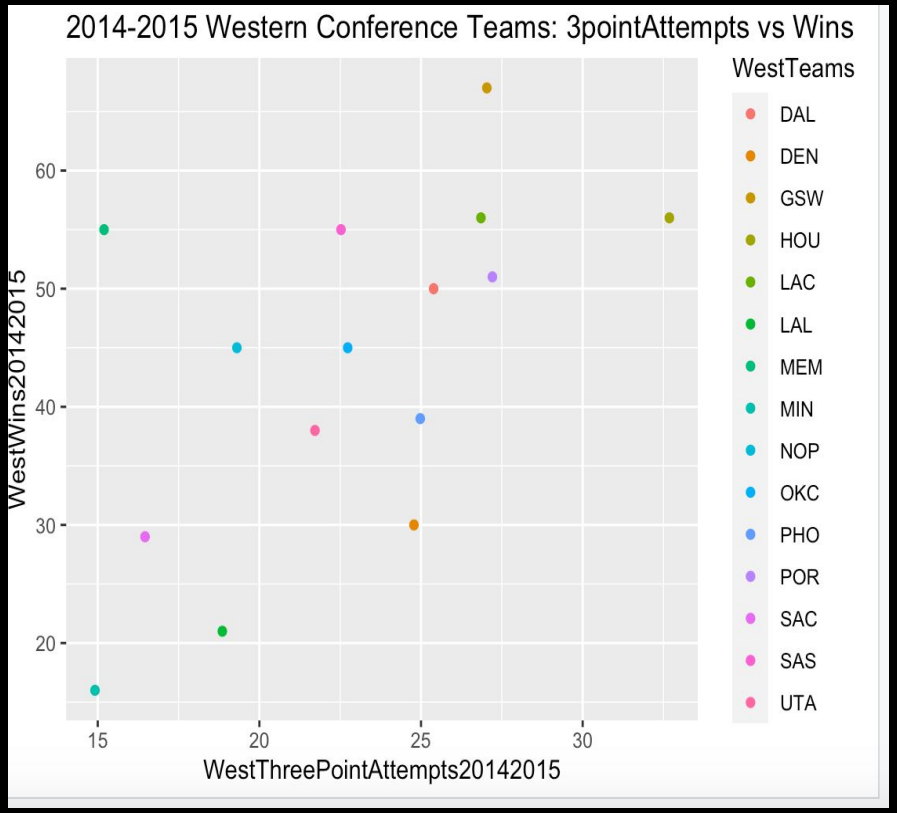



```
> mean(EastThreePointMakes20142015)
[1] 7.669106
```

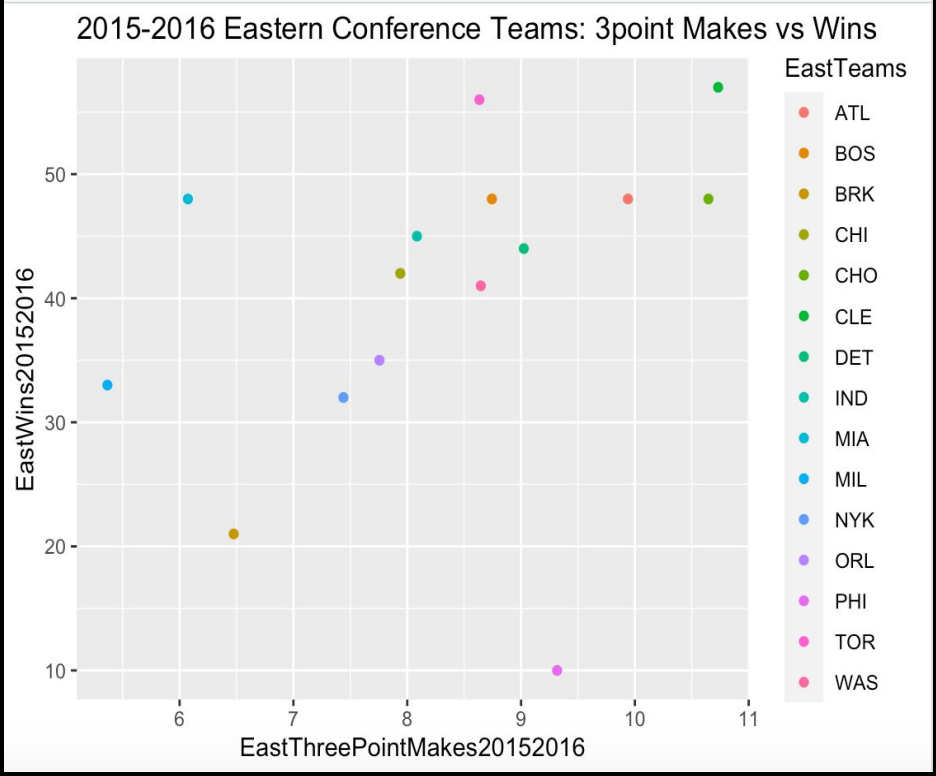
```
> mean(WestThreePointMakes20142015)
[1] 8.021951
```



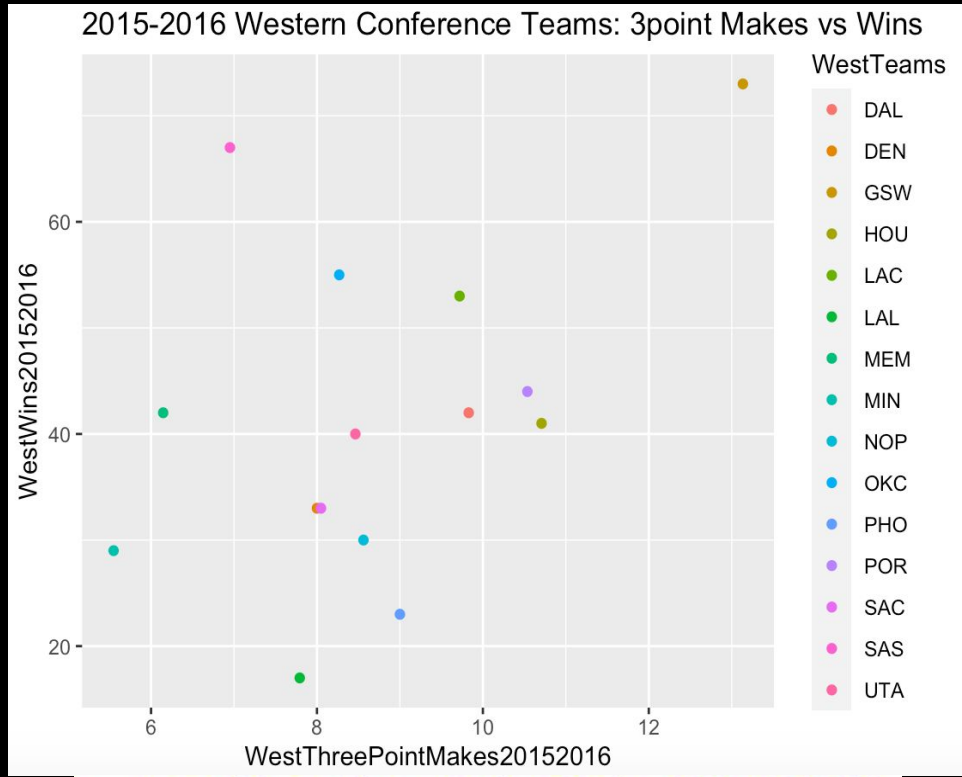

```
> mean(EastThreePointAttempts20142015)
[1] 22.11789
```



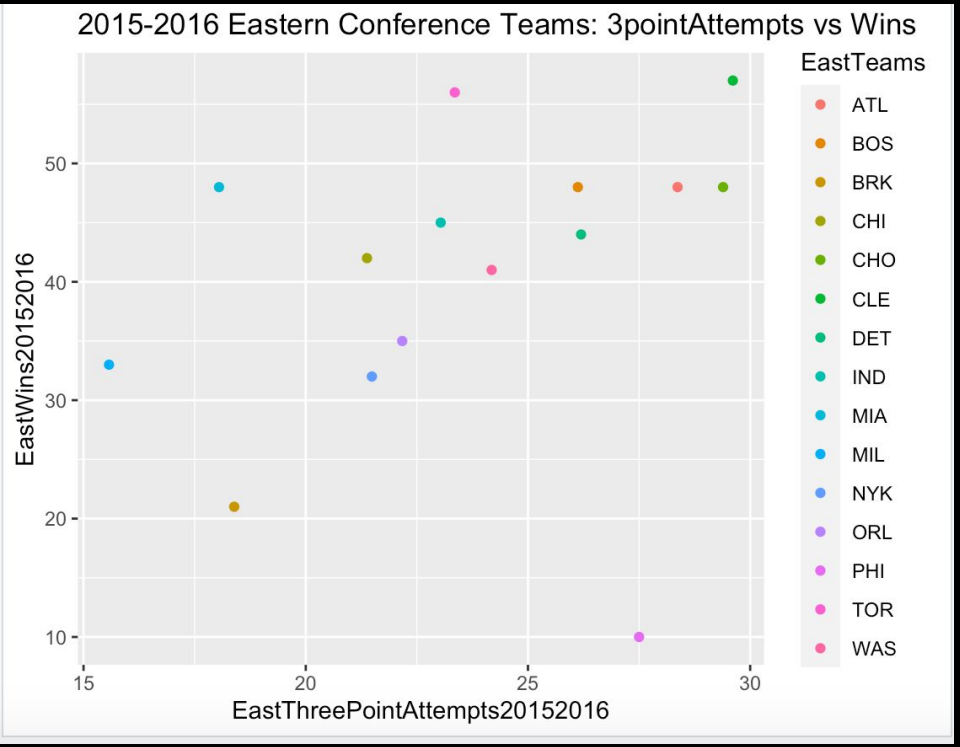

```
> mean(WestThreePointAttempts20142015)
[1] 22.70894
```

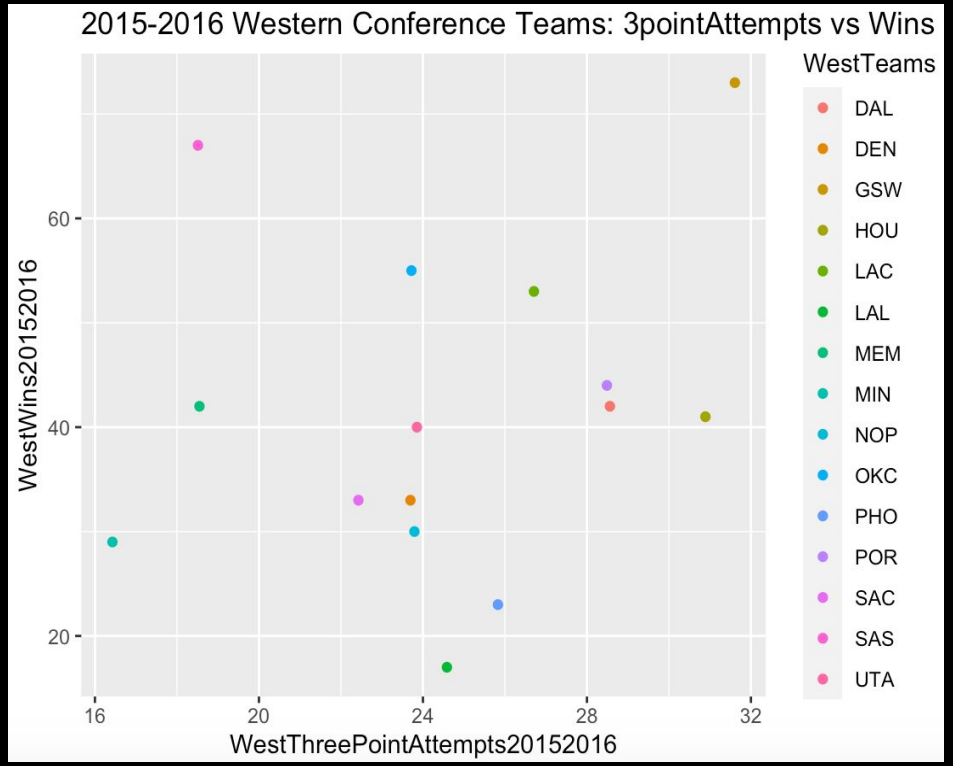



```
> mean(EastThreePointMakes20152016)
[1] 8.321138
```



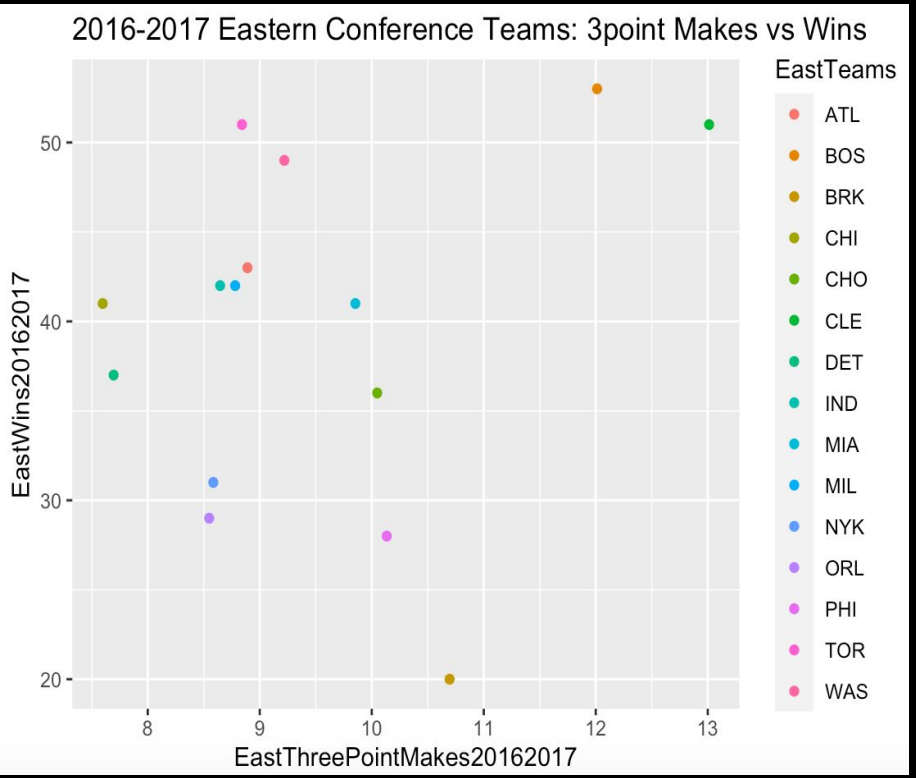

```
> mean(WestThreePointMakes20152016)
[1] 8.713821
```



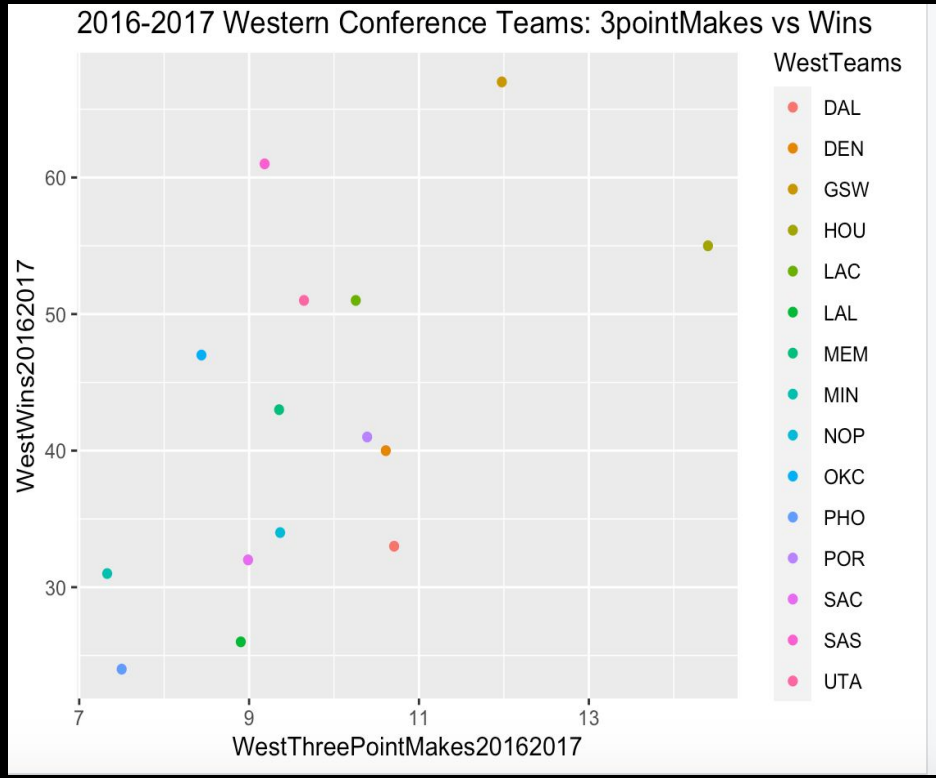



```
> mean(EastThreePointAttempts20152016)
[1] 23.65366
```

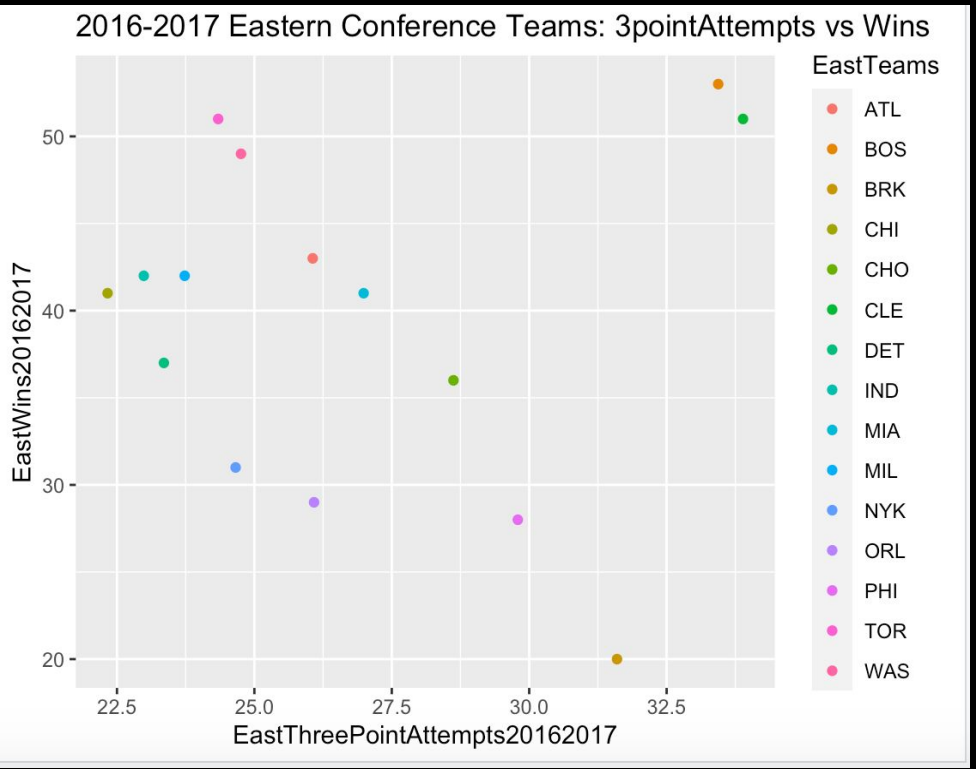
```
> mean(WestThreePointAttempts20152016)
[1] 24.50976
```

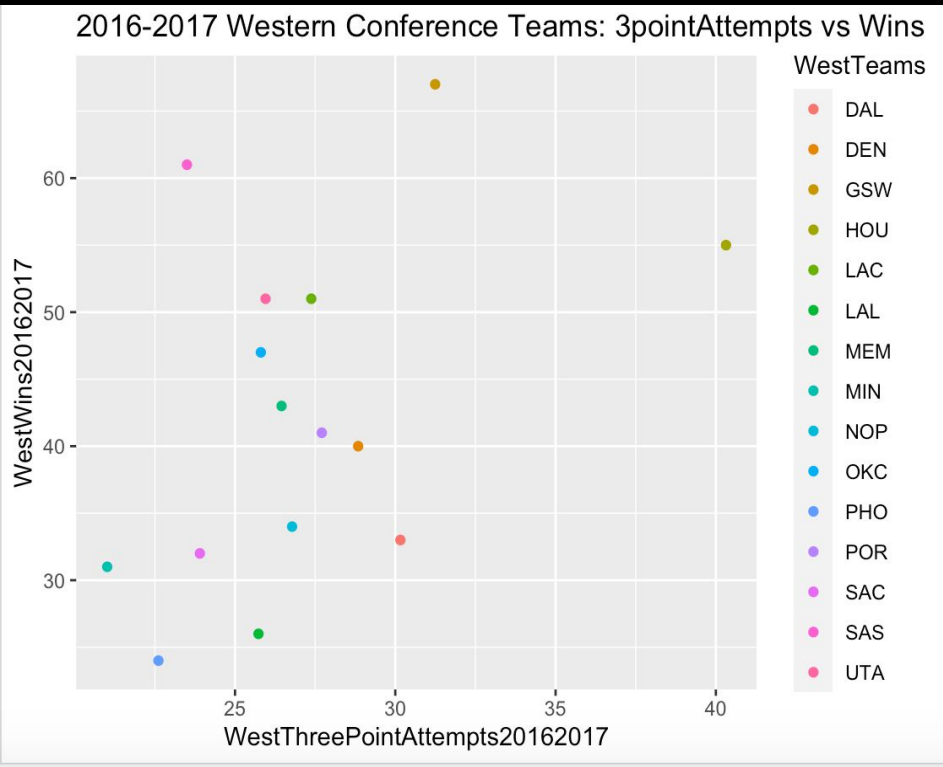



```
> mean(EastThreePointMakes20162017)
[1] 9.504065
```



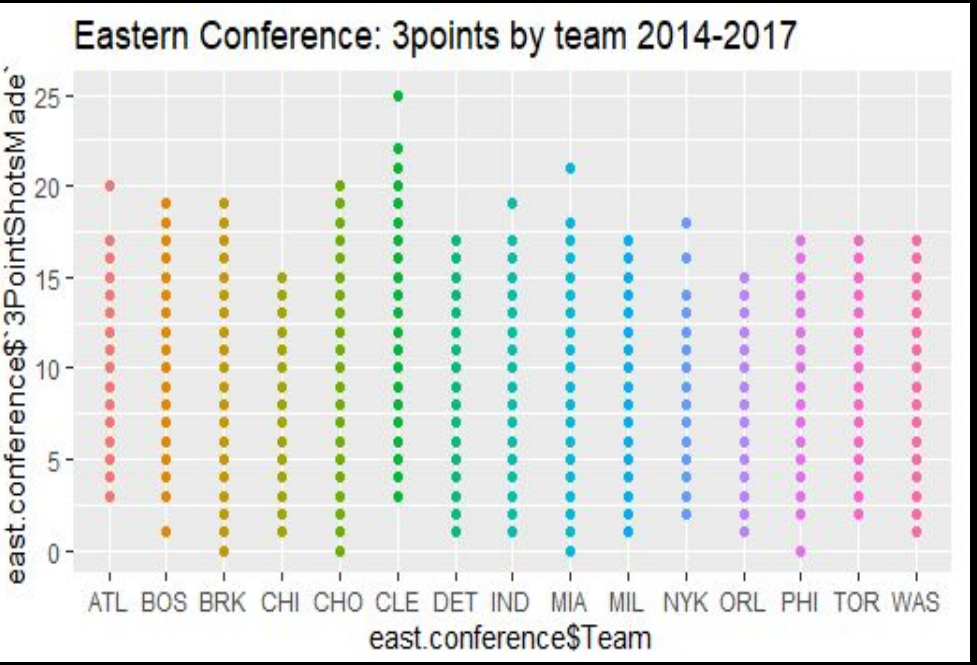

```
> mean(WestThreePointMakes20162017)
[1] 9.803252
```

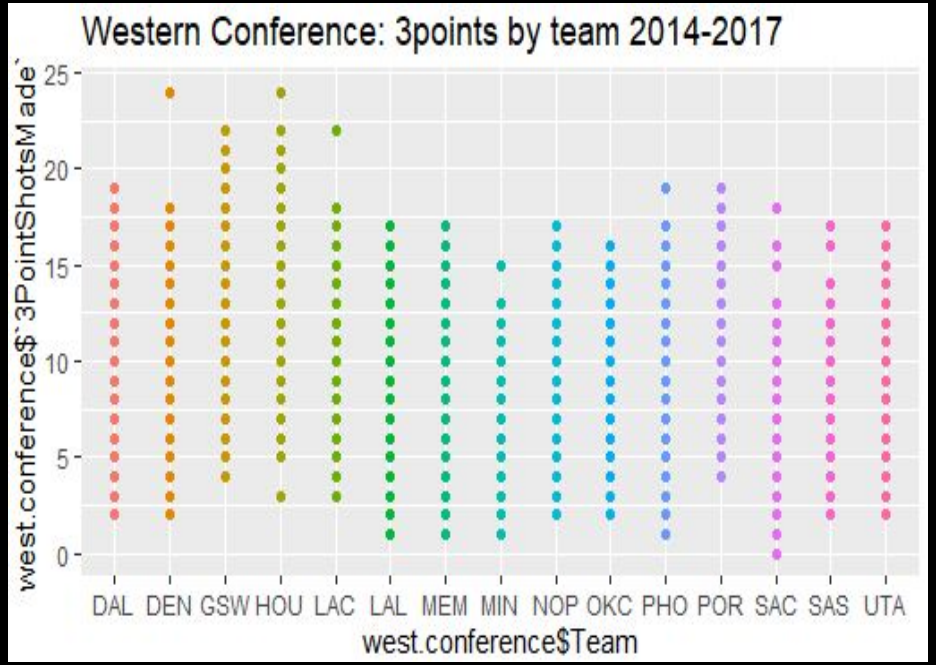


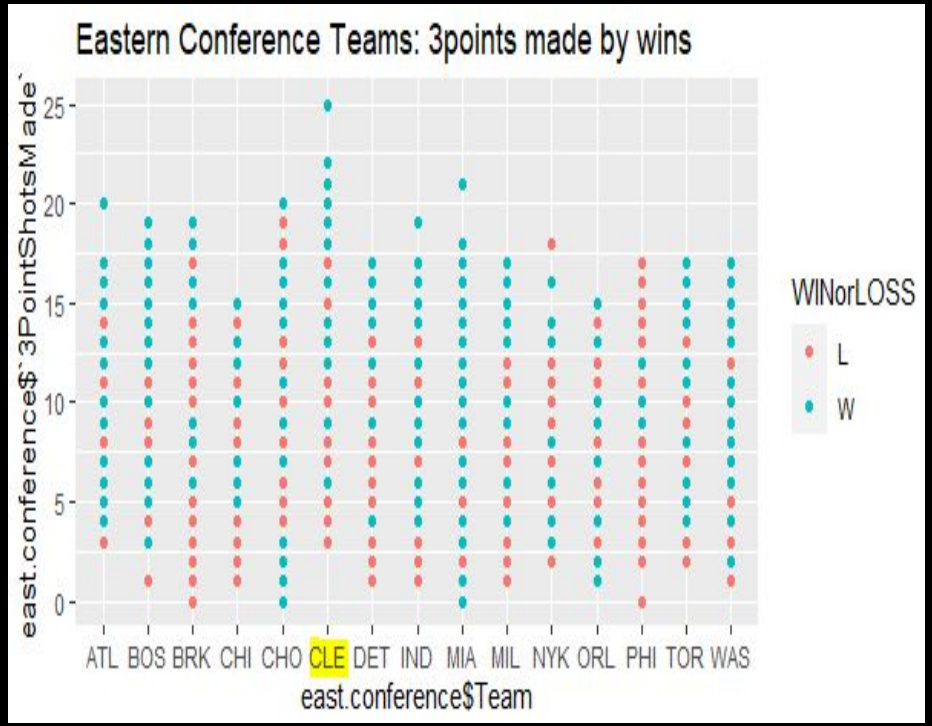


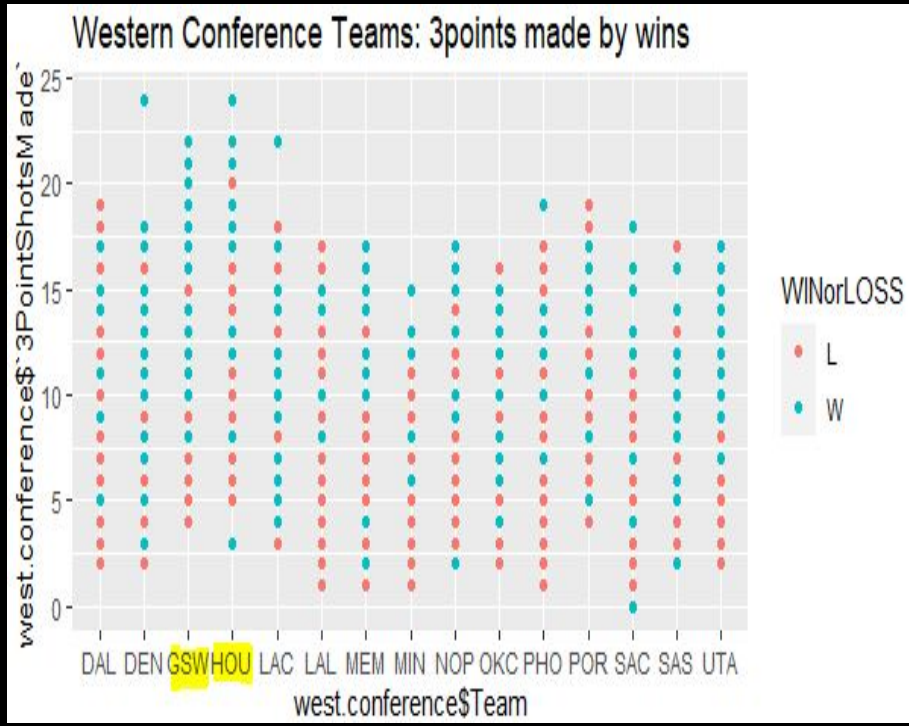

```
> mean(EastThreePointAttempts20162017)
[1] 26.84228
```

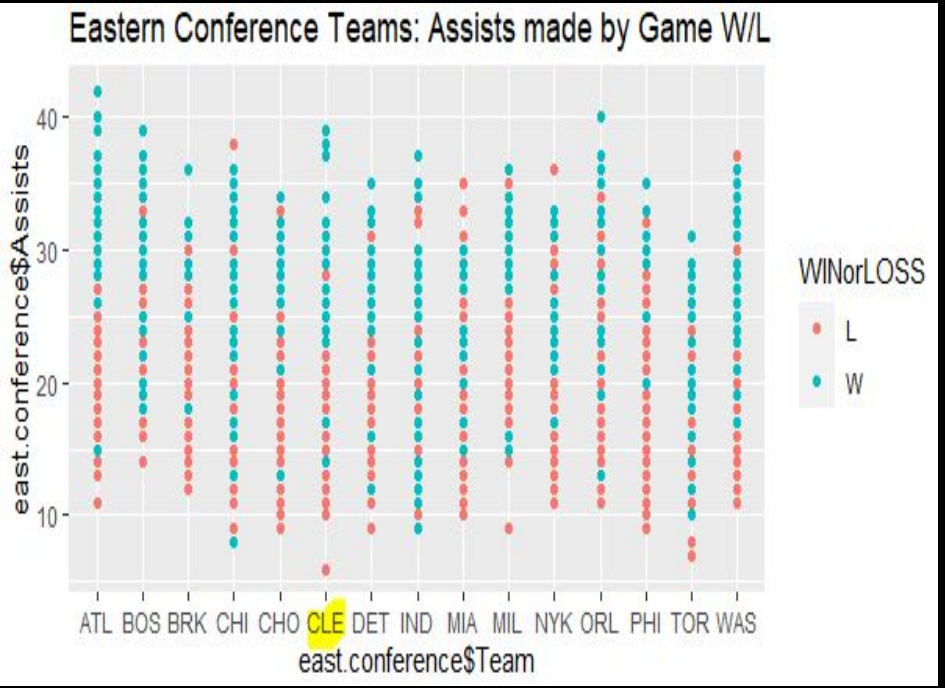
```
> mean(WestThreePointAttempts20162017)
[1] 27.15935
>
```

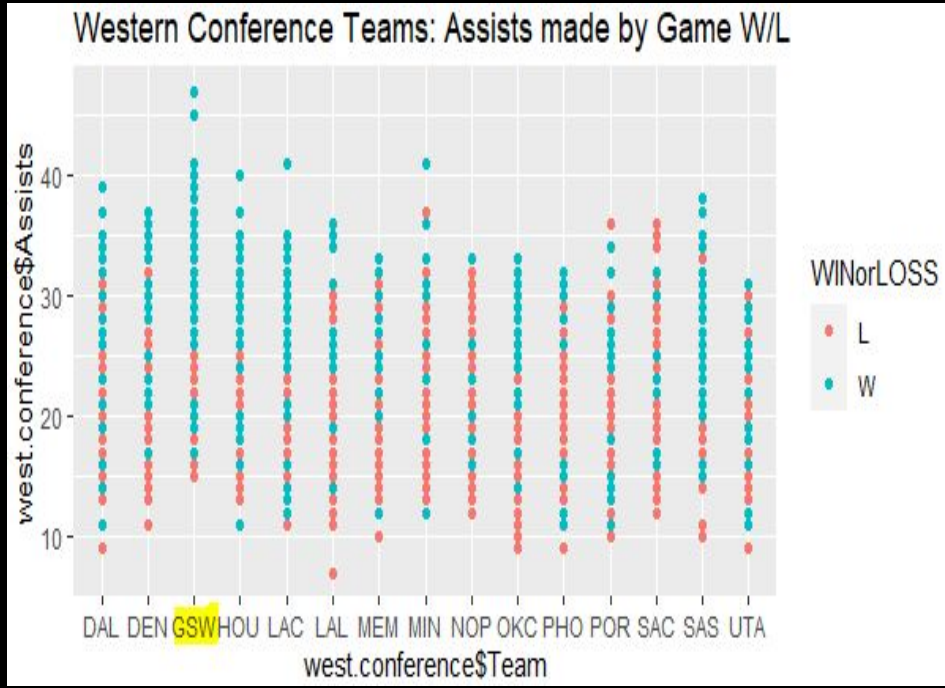

Eastern Conference: 3points by team 2014-2017
east.conference$`3PointShotsMade`
25
20
15
10
5
0
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS
east.conference$Team

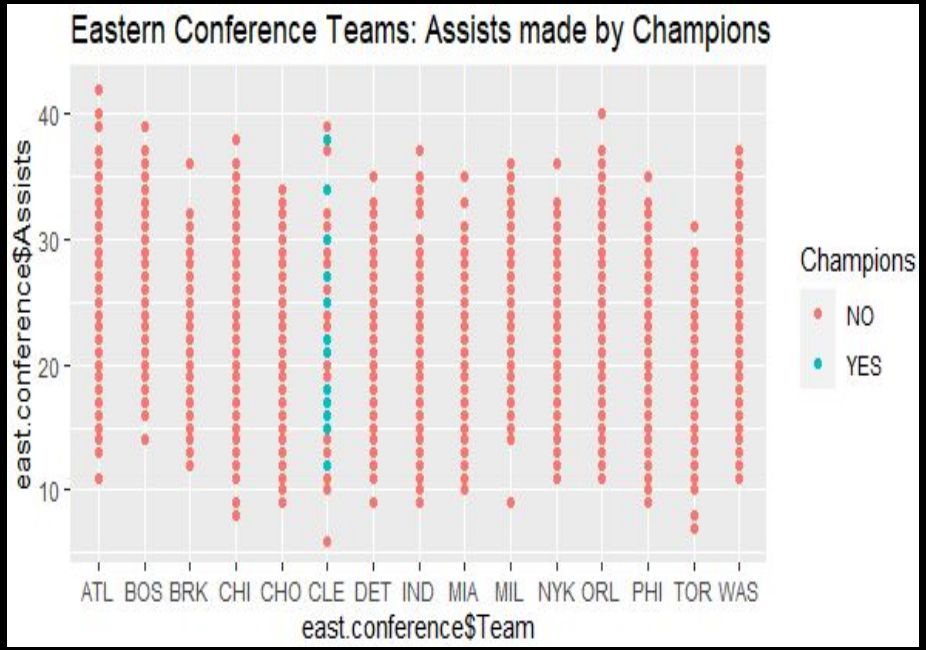

Western Conference: 3points by team 2014-2017
west.conference$`3PointShotsMade`
25
20
15
10
5
0
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA
west.conference$Team

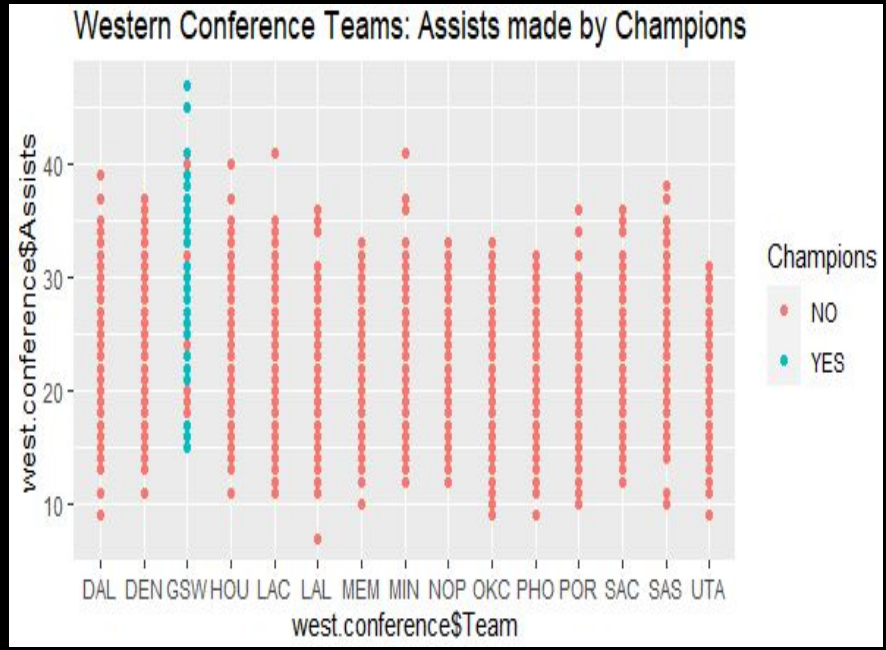

Eastern Conference Teams: 3points made by wins
east.conference$`3PointShotsMade`
25
20
15
10
5
0
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS
east.conference$Team
WINorLOSS
L
W


Western Conference Teams: 3points made by wins
west.conference$`3PointShotsMade`
25
20
15
10
5
0
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA
west.conference$Team
WINorLOSS
L
W

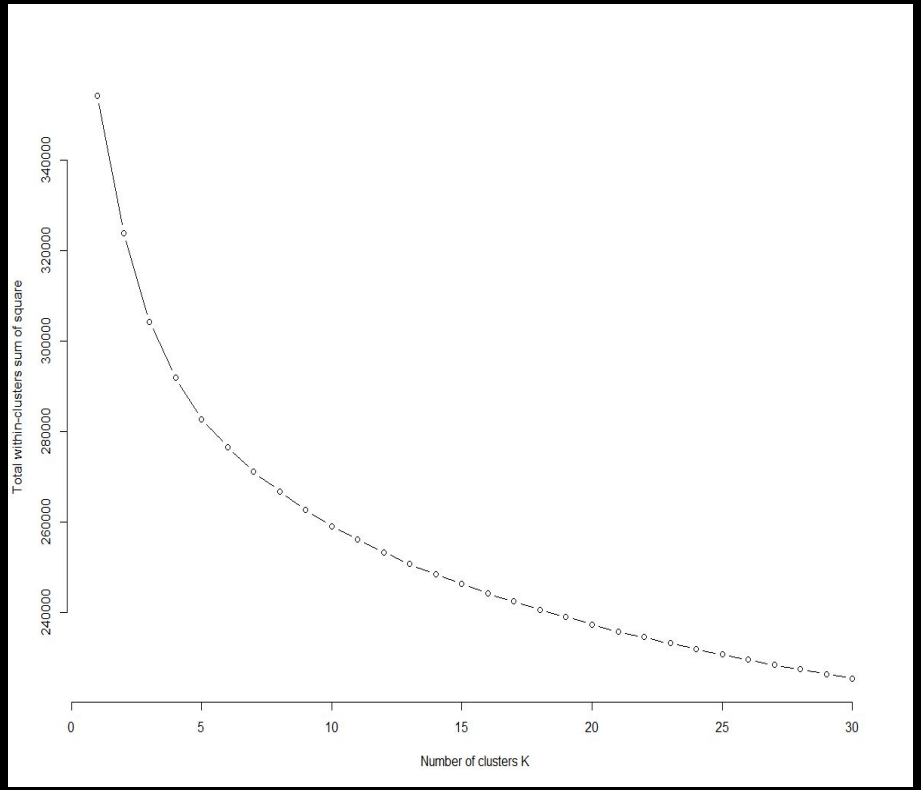

Eastern Conference Teams: Assists made by Game W/L
east.conference$Assists
40
30
20
10
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS
east.conference$Team
WINorLOSS
L
W

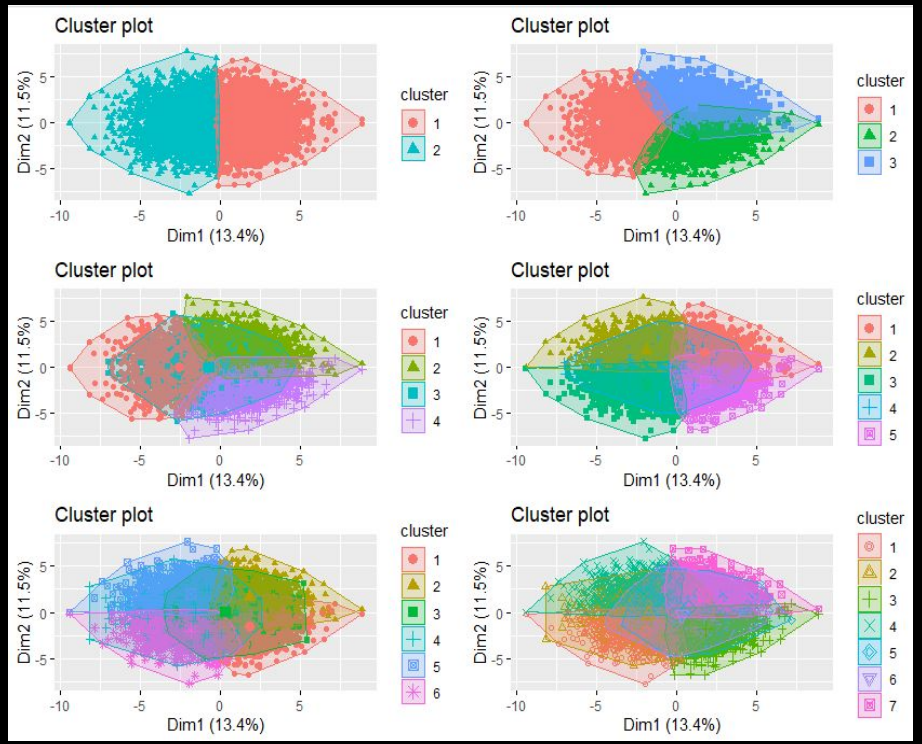

Western Conference Teams: Assists made by Game W/L
west.conference$Assists
40
30
20
10
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA
west.conference$Team
WINorLOSS
L
W


Eastern Conference Teams: Assists made by Champions
east.conference$Assists
40
30
20
10
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS
east.conference$Team
Champions
NO
YES


Western Conference Teams: Assists made by Champions
west.conference$Assists
40
30
20
10
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA
west.conference$Team
Champions
NO
YES

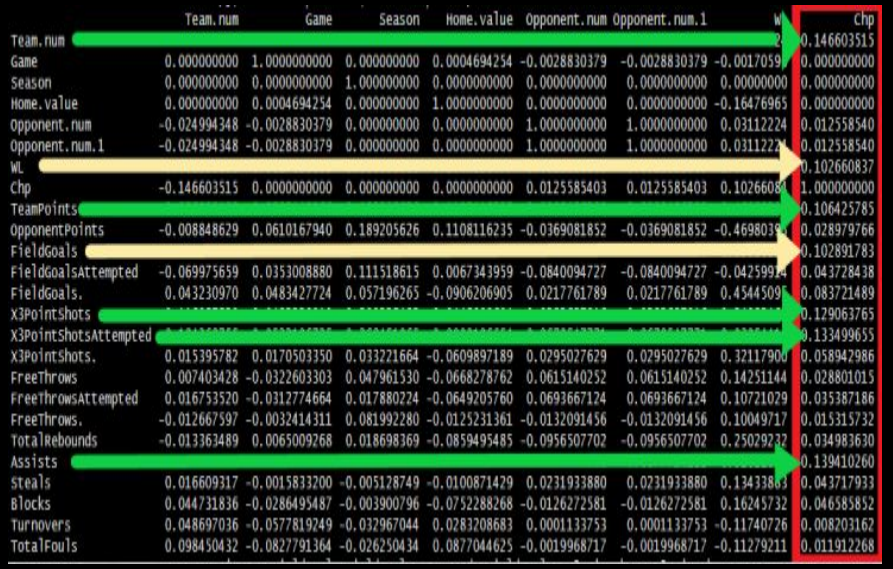
K-Means


Total within-clusters sum of square
240000
260000
280000
300000
320000
340000
0
5
10
15
20
25
30
Number of clusters K

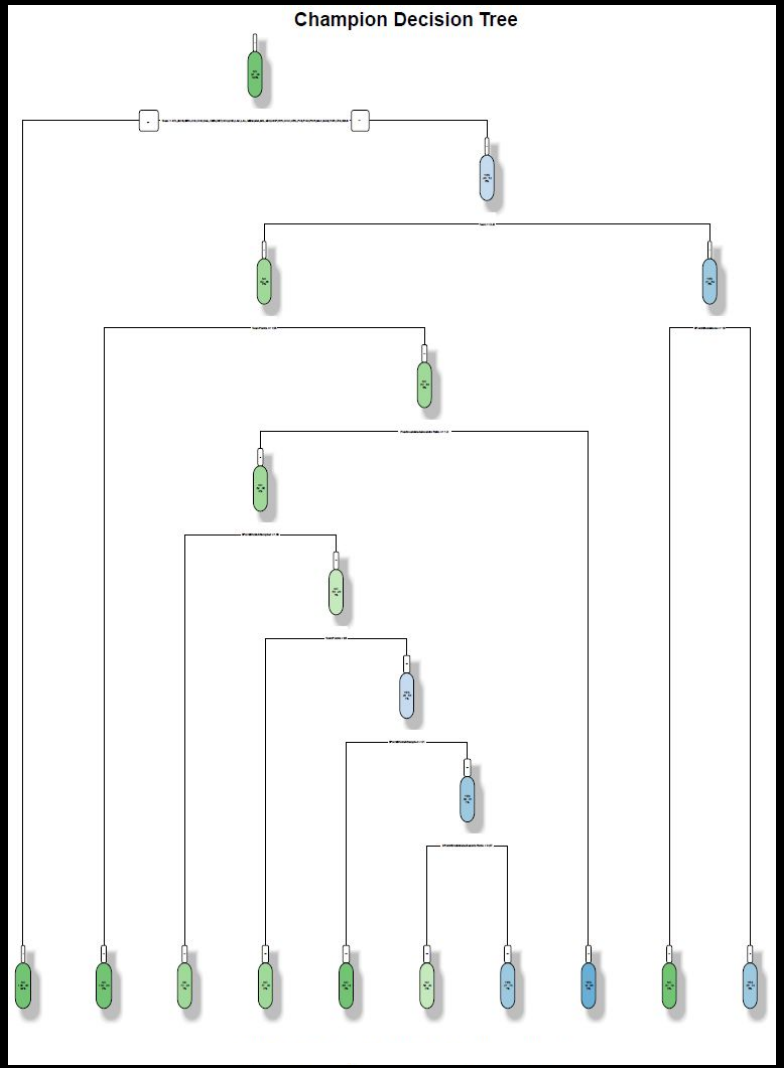

Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
3
Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
3
4
Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
3
4
5
Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
3
4
5
6
Cluster plot
Dim2 (11.5%)
Dim1 (13.4%)
cluster
1
2
3
4
5
6
7

Correlation Matrix

- Original data contained 47 variables and nearly 10,000 observations
- Data was initially trimmed into training data for 3 seasons and testing data of 1 season with the champion removed and set as TBD
- The training data was then divided again into an additional folds of training set of 80% (~40%) and 20% testing (~13% overall) to determine accuracy for which model should run the final unknown season to test the model.
- A correlation matrix was run to only consider the most significant variables in our datasets
- All non significant observations were dropped and only assists, 3 point shots made, 3 point shots attempted, field goals, and counts of wins and losses were retained



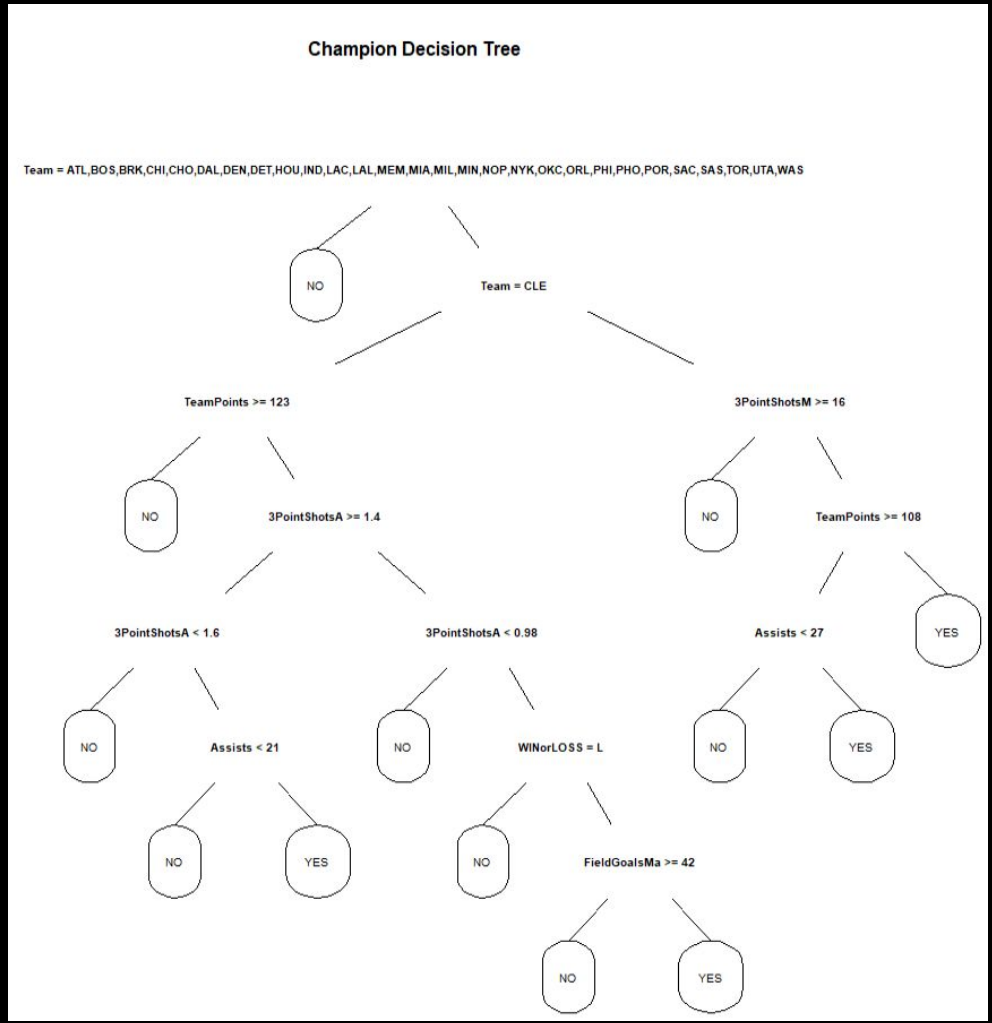
	Team.num	Game	Season	Home.value	Opponent.num	Opponent.num.1	WL	Chp
Team.num								0.146603515
Game	0.000000000	1.0000000000	0.000000000	0.0004694254	-0.0028830379	-0.0028830379	-0.0017059	0.000000000
Season	0.000000000	0.0000000000	1.000000000	0.0000000000	0.0000000000	0.0000000000	0.00000000	0.000000000
Home.value	0.000000000	0.0004694254	0.000000000	1.0000000000	0.0000000000	0.0000000000	-0.16476965	0.000000000
Opponent.num	-0.024994348	-0.0028830379	0.000000000	0.0000000000	1.0000000000	1.0000000000	0.03112224	0.012558540
Opponent.num.1	-0.024994348	-0.0028830379	0.000000000	0.0000000000	1.0000000000	1.0000000000	0.0311222	0.012558540
WL								0.102660837
Chp	-0.146603515	0.0000000000	0.000000000	0.0000000000	0.0125585403	0.0125585403	0.1026608	1.000000000
TeamPoints								0.106425785
OpponentPoints	-0.008848629	0.0610167940	0.189205626	0.1108116235	-0.0369081852	-0.0369081852	-0.4698039	0.028979766
FieldGoals								0.102891783
FieldGoalsAttempted	-0.069975659	0.0353008880	0.111518615	0.0067343959	-0.0840094727	-0.0840094727	-0.0425991	0.043728438
FieldGoals.	0.043230970	0.0483427724	0.057196265	-0.0906206905	0.0217761789	0.0217761789	0.4544509	0.083721489
X3PointShots								0.129063765
X3PointShotsAttempted								0.133499655
X3PointShots.	0.015395782	0.0170503350	0.033221664	-0.0609897189	0.0295027629	0.0295027629	0.3211790	0.058942986
FreeThrows	0.007403428	-0.0322603303	0.047961530	-0.0668278762	0.0615140252	0.0615140252	0.14251144	0.028801015
FreeThrowsAttempted	0.016753520	-0.0312774664	0.017880224	-0.0649205760	0.0693667124	0.0693667124	0.10721029	0.035387186
FreeThrows.	-0.012667597	-0.0032414311	0.081992280	-0.0125231361	-0.0132091456	-0.0132091456	0.10049717	0.015315732
TotalRebounds	-0.013363489	0.0065009268	0.018698369	-0.0859495485	-0.0956507702	-0.0956507702	0.25029232	0.034983630
Assists								0.139410260
Steals	0.016609317	-0.0015833200	-0.005128749	-0.0100871429	0.0231933880	0.0231933880	0.13433883	0.043717933
Blocks	0.044731836	-0.0286495487	-0.003900796	-0.0752288268	-0.0126272581	-0.0126272581	0.16245732	0.046585852
Turnovers	0.048697036	-0.0577819249	-0.032967044	0.0283208683	0.0001133753	0.0001133753	-0.11740726	0.008203162
TotalFouls	0.098450432	-0.0827791364	-0.026250434	0.0877044625	-0.0019968717	-0.0019968717	-0.11279211	0.011912268



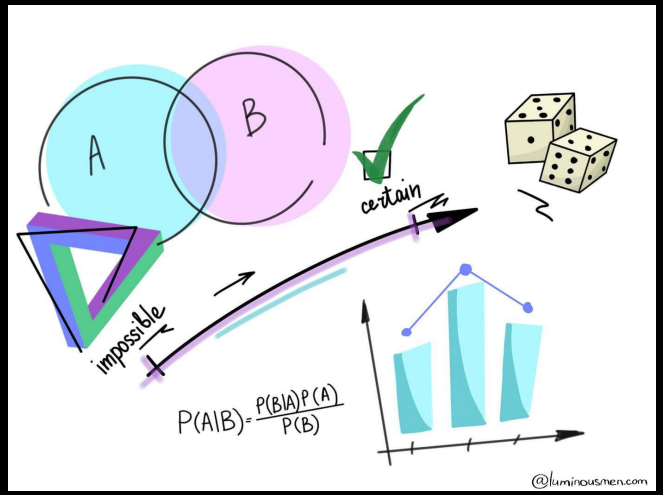
Decision Trees

Decision Trees continued

```
> confusionMatrix(t.test$Champions, decision_tree_predict)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  1395   20
       YES   31   30

               Accuracy : 0.9654
                 95% CI : (0.9548, 0.9742)
    No Information Rate : 0.9661
    P-Value [Acc > NIR] : 0.5937

                  Kappa : 0.5228

 Mcnemar's Test P-Value : 0.1614

            Sensitivity : 0.9783
            Specificity : 0.6000
         Pos Pred Value : 0.9859
         Neg Pred Value : 0.4918
             Prevalence : 0.9661
         Detection Rate : 0.9451
   Detection Prevalence : 0.9587
      Balanced Accuracy : 0.7891

       'Positive' Class : NO
```



Bayes




```
> confusionMatrix(t.test$Champions, nb_pred)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  1395   20
       YES   31   30

               Accuracy : 0.9654
                 95% CI : (0.9548, 0.9742)
    No Information Rate : 0.9661
    P-Value [Acc > NIR] : 0.5937

                  Kappa : 0.5228

 Mcnemar's Test P-Value : 0.1614

            Sensitivity : 0.9783
            Specificity : 0.6000
         Pos Pred Value : 0.9859
         Neg Pred Value : 0.4918
             Prevalence : 0.9661
         Detection Rate : 0.9451
   Detection Prevalence : 0.9587
      Balanced Accuracy : 0.7891

       'Positive' Class : NO
```

Random Forest

- Accuracy increased to 97%
- Process time was not cumbersome

```
> confusionMatrix(t.test$Champions, rf_pred)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  1400   15
       YES   29   32

               Accuracy : 0.9702
                 95% CI : (0.9602, 0.9783)
    No Information Rate : 0.9682
    P-Value [Acc > NIR] : 0.36300

                  Kappa : 0.5774

 Mcnemar's Test P-Value : 0.05002

            Sensitivity : 0.9797
            Specificity : 0.6809
         Pos Pred Value : 0.9894
         Neg Pred Value : 0.5246
             Prevalence : 0.9682
         Detection Rate : 0.9485
   Detection Prevalence : 0.9587
      Balanced Accuracy : 0.8303

       'Positive' Class : NO
```

SVM

- Since the SVM algorithm was the most efficient in both processing time and accuracy percentage that's what we will use for prediction

```
> confusionMatrix(t.test$Champions, svm_pred)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  1398   17
       YES   20   41

               Accuracy : 0.9749
                 95% CI : (0.9656, 0.9823)
    No Information Rate : 0.9607
    P-Value [Acc > NIR] : 0.001818

                  Kappa : 0.676

 Mcnemar's Test P-Value : 0.742308

            Sensitivity : 0.9859
            Specificity : 0.7069
         Pos Pred Value : 0.9880
         Neg Pred Value : 0.6721
             Prevalence : 0.9607
         Detection Rate : 0.9472
   Detection Prevalence : 0.9587
      Balanced Accuracy : 0.8464

       'Positive' Class : NO
```

SVM Prediction

The csv output was paired against the original test data set line by line and we were able to validate that the model built is correct and the SVM model was able to accurately predict who won the 2017-2018 NBA season

GameID	Champions	Team	Champions
10	YES	GSW	TBD
20	YES	GSW	TBD
37	YES	GSW	TBD
55	YES	GSW	TBD
75	YES	GSW	TBD
87	YES	GSW	TBD
97	YES	GSW	TBD
107	YES	GSW	TBD
117	YES	GSW	TBD
127	YES	GSW	TBD
144	YES	GSW	TBD
162	YES	GSW	TBD
182	YES	GSW	TBD
194	YES	GSW	TBD
204	YES	GSW	TBD
214	YES	GSW	TBD
224	YES	GSW	TBD
234	YES	GSW	TBD
251	YES	GSW	TBD
269	YES	GSW	TBD
289	YES	GSW	TBD
301	YES	GSW	TBD
311	YES	GSW	TBD
321	YES	GSW	TBD