SC1015 - Mini Project

NBA Datasets

By: SC13 - Group 1 - Ananthan Srinath Adhvait, Wan Kai Jie and Anapana Dinesh Kumar

Table of Contents

Practical Motivation

- The National Basketball Association (NBA) is a league composed of 30 teams and is one of the major sports league in the US and Canada.
- We as a group, are very interested in the NBA and enjoy playing basketball as well. As a result, we closely follow the NBA season and have always had a few areas that intrigued us.

Sample Collection

Problem 1:

- NBA Stats 2016-2019 - Used for forming the Model
- NBA Stats 2020-20021 - Used for testing the model and predicting All-Stars

Problem 2:

- Games_updated_csv. - Used to determine home court advantage
- Basketball Reference NBA Summary - Used in prediction

2 Problems Explored

1. What variables are important to predict NBA All-Stars?

2 . How effective n is home court advantage?

Problem 1: What are the statistics important for predicting the NBA All-Stars

- Aim is to find out what statistics are important for a player to be considered an All-Star.

- Main Variables used:
    - Total Rebounds
    - Total Points
    - Win Share
    - Total Assists

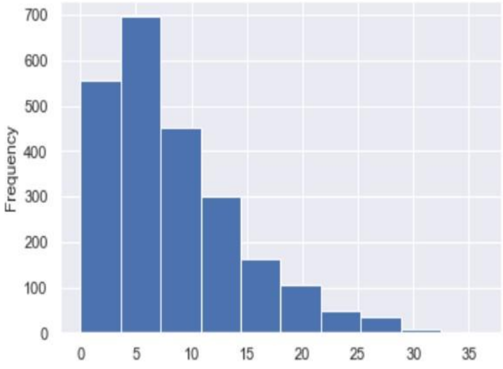
Exploratory Data Analysis

Finding the Average statistics of NBA players in the NBA and comparing them to the statistics of NBA All-Stars

Exploratory Data Analysis: Exploring the average number of Assists, Rebounds and Points.

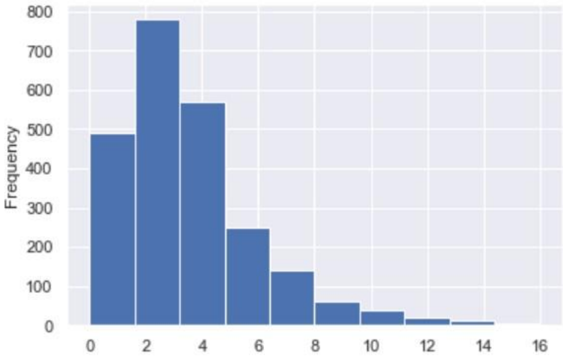
Total Points

```
count    2360.000000
mean        8.313644
std         6.042272
min         0.000000
25%         3.900000
50%         6.800000
75%        11.500000
max        36.100000
Name: PTS, dtype: float64
```



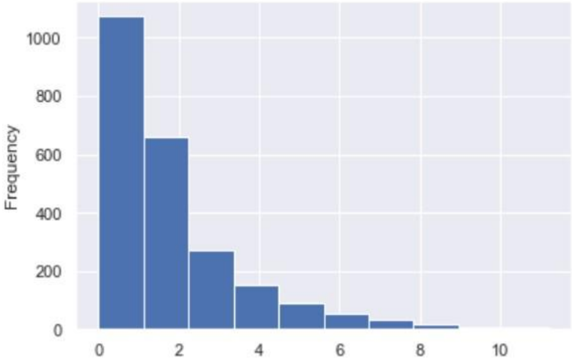
Total Rebounds

```
count    2360.000000
mean        3.489068
std         2.459760
min         0.000000
25%         1.800000
50%         3.000000
75%         4.600000
max        16.000000
Name: TRB, dtype: float64
```



Total Assists

```
count    2360.000000
mean        1.822331
std         1.716730
min         0.000000
25%         0.700000
50%         1.300000
75%         2.400000
max        11.200000
Name: AST, dtype: float64
```



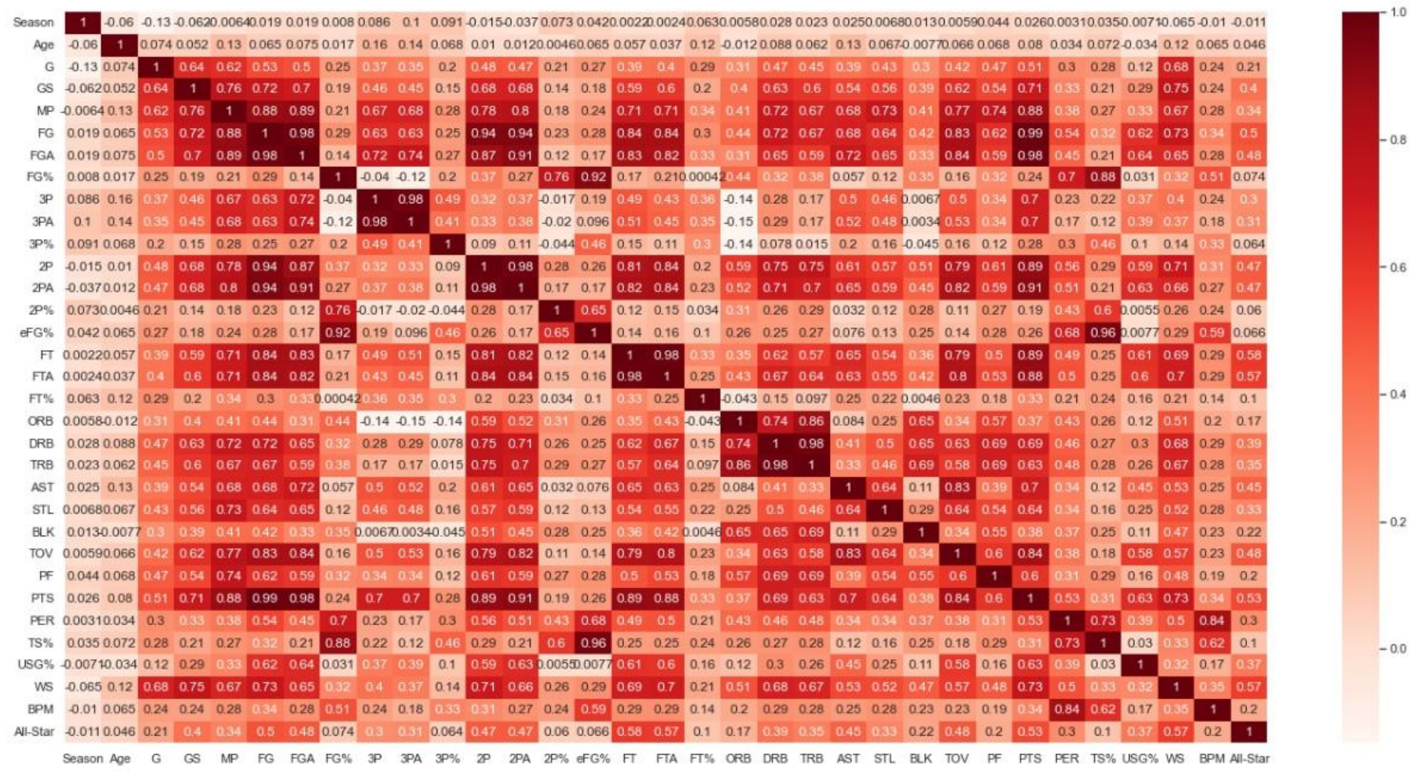
Exploratory Data Analysis: Comparison of statistics with NBA All-Stars

Analytic Visualization and Pattern Recognition

To find out the effectiveness of each variable in determining if a player becomes an All-Star

Visualization of the effectiveness of the variables using a heatmap

Exploratory Data Analysis: Most Effective Statistics to Determine an All-Star



Exploratory Data Analysis: Most Effective Variables to Determine an All-Star

```
<bound method NDFrame.head of GS          0.397897
MP          0.344354
FG          0.499682
FGA         0.482143
3P          0.302392
3PA         0.306670
2P          0.473187
2PA         0.467463
FT          0.582824
FTA         0.570977
DRB         0.391594
TRB         0.347066
AST         0.449640
STL         0.328841
TOV         0.483782
PTS         0.527794
USG%        0.372814
WS          0.568998
All-Star    1.000000
Name: All-Star, dtype: float64>

<bound method NDFrame.head of FT          0.582824
FTA         0.570977
PTS         0.527794
WS          0.568998
All-Star    1.000000
Name: All-Star, dtype: float64>
```

- Variables that have a correlation with All-Stars of 0.30 and above

- Variables that have a correlation with All-Stars of 0.5 and above.

Machine Learning

Developing models that can be used to predict All-Stars

1. Cleaning dataset and forming a desirable dataframe
2. Forming the Models
3. Implementing the model in the new dataset

Cleaning Dataset

- Dropping all columns that have a correlation of less than 0.3 with All-Star.
- Cleaning dataset by replacing ‘0’ with True.
- Creating a new dataframe that can be used to create the model

```python
#Cleaning Data
playerdata.fillna(0, inplace = True)

#Dropping irrelevant columns to help us predict All-Star
newplayerdata = playerdata.drop(['Player', 'Pos', 'Tm','Season', 'Age', 'G', 'GS', 'All-Star'], axis = 1)
print(newplayerdata.head())
```

```
     MP   FG  FGA    FG%   3P  3PA    3P%   2P  2PA    2P%  ...  STL  BLK  \
0  15.5  2.0  5.0  0.393  1.4  3.6  0.381  0.6  1.4  0.426  ...  0.5  0.1
1   8.0  0.8  2.8  0.294  0.2  1.2  0.143  0.7  1.7  0.400  ...  0.0  0.0
2  15.9  2.0  4.8  0.425  1.1  2.6  0.434  0.9  2.2  0.414  ...  0.4  0.5
3  29.9  4.7  8.2  0.571  0.0  0.0  0.000  4.7  8.2  0.572  ...  1.1  1.0
4  25.9  3.0  6.9  0.440  1.0  2.5  0.411  2.0  4.4  0.457  ...  0.3  0.1

   TOV   PF   PTS   PER    TS%  USG%   WS   BPM
0  0.5  1.7   6.0  10.1  0.560  15.9  2.1  -1.6
1  0.3  1.5   2.2  -1.4  0.355  20.0 -0.1 -14.3
2  0.6  1.8   6.5  13.1  0.587  16.5  1.1  -0.9
3  1.8  2.4  11.3  16.5  0.589  16.2  6.5  -0.2
4  0.7  1.7   8.4   8.9  0.559  14.4  1.4  -3.6

[5 rows x 28 columns]
```

ML Technique Used: Random Forest Classifier

- Random Forest is a machine learning technique that can be used for regression and classification problems.

- In this case, Random Forest Classifier is necessary because we need produce good predictions and, it can be used to handle large datasets such as the one used for this problem.

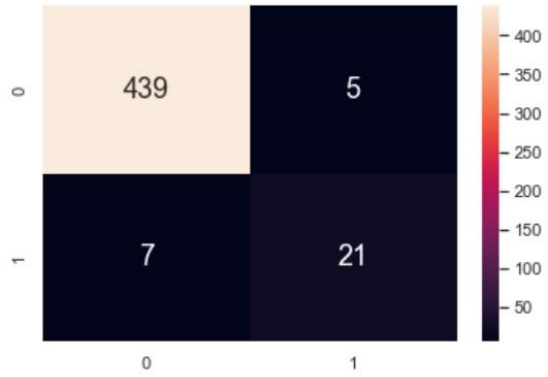
- Moreover, it has a higher accuracy than the Decision Tree Classifier.

Model 1

- With the new dataset, we split the dataframe into Train and Test data to find out the accuracy of the model.
- The accuracy of the Train Model was 1.0
- The accuracy of the Test Model was 0.9745.
- But is it possible to get a better model?

Train Data

Test Data



Model 2

- Similar to the first model, we split the dataframe into Train and Test data to find out the accuracy of the model.

- The accuracy of the Train model was 1.0 and the accuracy of the Test model was 0.9766, which is actually lower than the first model!!

Train Data

Test Data

Predicting All-Stars

- After loading in the new dataset, we train it to fit the chosen model.
    - This can be done by dropping the unnecessary columns.
- Next, after using the model to predict, replace all 1's with 'All-Star' to determine the names of the predicted players.

```
pred = rf.predict(mydata[finalstats])

print(pred)
```

```
[0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
```

```
mydata['All-Star'] = pred.tolist()

print(mydata)
```

```
      Rk             Player Pos  Age   Tm   G  GS    MP   FG   FGA  ...  BLK  \
0      1   Precious Achiuwa  PF   21  MIA  61   4  12.1  2.0   3.7  ...  0.5
1      2       Jaylen Adams  PG   24  MIL   7   0   2.6  0.1   1.1  ...  0.0
2      3       Steven Adams   C   27  NOP  58  58  27.7  3.3   5.3  ...  0.7
3      4        Bam Adebayo   C   23  MIA  64  64  33.5  7.1  12.5  ...  1.0
4      5  LaMarcus Aldridge   C   35  SAS  21  18  25.9  5.5  11.8  ...  0.9
..   ...                ...  ..  ...  ...  ..  ..   ...  ...   ...  ...  ...
621  536       Delon Wright  PG   28  SAC  27   8  25.8  3.9   8.3  ...  0.4
622  537     Thaddeus Young  PF   32  CHI  68  23  24.3  5.4   9.7  ...  0.6
623  538         Trae Young  PG   22  ATL  63  63  33.7  7.7  17.7  ...  0.2
624  539        Cody Zeller   C   28  CHO  48  21  20.9  3.8   6.8  ...  0.4
625  540        Ivica Zubac   C   23  LAC  72  33  22.3  3.6   5.5  ...  0.9

     TOV   PF   PTS   PER    TS%  USG%   WS    BPM  All-Star
0    0.7  1.5   5.0  14.2  0.550  19.5  1.3   -4.5       0.0
1    0.0  0.1   0.3  -6.5  0.125  18.6 -0.1  -19.8       0.0
2    1.3  1.9   7.6  15.1  0.596  11.7  4.0   -0.8       0.0
3    2.6  2.3  18.7  22.7  0.626  23.7  8.8    4.7       1.0
4    1.0  1.7  13.7  15.1  0.545  22.7  0.8   -1.2       0.0
..   ...  ...   ...   ...    ...   ...  ...    ...       ...
621  1.3  1.1  10.0  15.9  0.563  17.1  1.3    0.8       0.0
622  2.0  2.2  12.1  20.3  0.578  22.3  5.1    3.2       0.0
623  4.1  1.8  25.3  23.0  0.589  33.0  7.2    3.6       1.0
624  1.1  2.5   9.4  18.2  0.599  18.3  3.3   -0.3       0.0
625  1.1  2.6   9.0  19.1  0.693  15.1  6.9    0.7       0.0

[626 rows x 36 columns]
```

Statistical Inference

Checking whether the predicted players are accurate

Are the Predicted Players Accurate?

Players Predicted by the Model:

	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	\
3	4	Bam Adebayo	C	23	MIA	64	64	33.5	7.1	12.5	...	
14	12	Giannis Antetokounmpo	PF	26	MIL	61	61	33.0	10.3	18.0	...	
41	38	Bradley Beal	SG	27	WAS	60	60	35.8	11.2	23.0	...	
95	84	Jimmy Butler	SF	31	MIA	52	52	33.6	7.0	14.2	...	
134	116	Stephen Curry	PG	32	GSW	63	63	34.2	10.4	21.7	...	
136	118	Anthony Davis	PF	27	LAL	36	36	32.3	8.4	17.0	...	
151	130	Luka Dončić	PG	21	DAL	66	66	34.3	9.8	20.5	...	
162	140	Kevin Durant	PF	32	BRK	35	32	33.1	9.3	17.2	...	
168	146	Joel Embiid	C	26	PHI	51	51	31.1	9.0	17.6	...	
198	173	Paul George	SF	30	LAC	54	54	33.7	8.2	17.6	...	
229	200	James Harden	SG	31	HOU	8	8	36.3	7.5	16.9	...	
230	200	James Harden	PG	31	BRK	36	35	36.6	7.8	16.6	...	
279	242	Brandon Ingram	SF	23	NOP	61	61	34.3	8.4	18.0	...	
280	243	Kyrie Irving	PG	28	BRK	54	54	34.9	10.2	20.1	...	
290	251	LeBron James	PG	36	LAL	45	45	33.4	9.4	18.3	...	
303	262	Nikola Jokić	C	25	DEN	72	72	34.6	10.2	18.0	...	
335	286	Zach LaVine	SG	25	CHI	58	58	35.1	9.8	19.4	...	
343	293	Kawhi Leonard	SF	29	LAC	52	52	34.1	8.9	17.5	...	
348	297	Damian Lillard	PG	30	POR	67	67	35.8	9.0	19.9	...	
402	348	Donovan Mitchell	PG	24	UTA	53	53	33.4	9.0	20.6	...	
485	421	Julius Randle	PF	26	NYK	71	71	37.6	8.5	18.6	...	
513	444	Domantas Sabonis	PF	24	IND	62	62	36.0	7.8	14.6	...	
544	473	Jayson Tatum	SF	22	BOS	64	64	35.8	9.5	20.6	...	
567	491	Karl-Anthony Towns	C	25	MIN	50	50	33.8	8.5	17.5	...	
582	504	Nikola Vučević	C	30	ORL	44	44	34.1	9.9	20.6	...	
599	517	Russell Westbrook	PG	32	WAS	65	65	36.4	8.4	19.0	...	
611	528	Zion Williamson	PF	20	NOP	61	61	33.2	10.4	17.0	...	
623	538	Trae Young	PG	22	ATL	63	63	33.7	7.7	17.7	...	

The Actual All-Stars of the 2020-2021 Season are:

1. Kevin Durant
2. Lebron James
3. Giannis Antetokounmpo
4. Stephen Curry
5. Luka Doncic
6. Nikola Jokic
7. Jaylen Brown
8. Paul George
9. Rudy Gobert
10. Damian Lillard
11. Chris Paul
12. Domantas Sabonis
13. Ben Simmons
14. Bradley Beal
15. Joel Embiid
16. Kyrie Irving
17. Kawhi Leonard
18. Jayson Tatum
19. Devin Booker
20. Anthony Davis
21. James Harden
22. Zach Lavine
23. Donovan Mitchell
24. Julius Randle
25. Nikola Vucevic
26. Zion Williamson
27. Mike Conley

Problem 2: How effective is home court advantage?

- Home court advantage refers to the benefit that the home team is said to gain over the visiting (away) team
- To find out if home court advantage exists and if it does, how effective is it?
- Variables include:
    - Win-loss percentage of all teams in general
    - Points of home and away teams respectively
    - Win-loss margin from the perspective of home teams

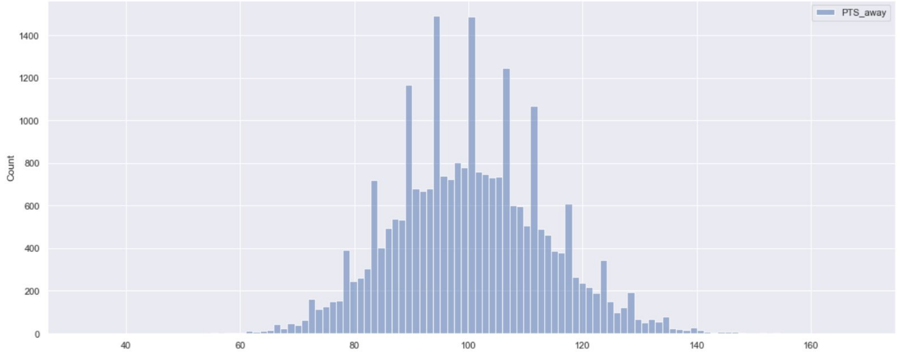
Exploratory Data Analysis

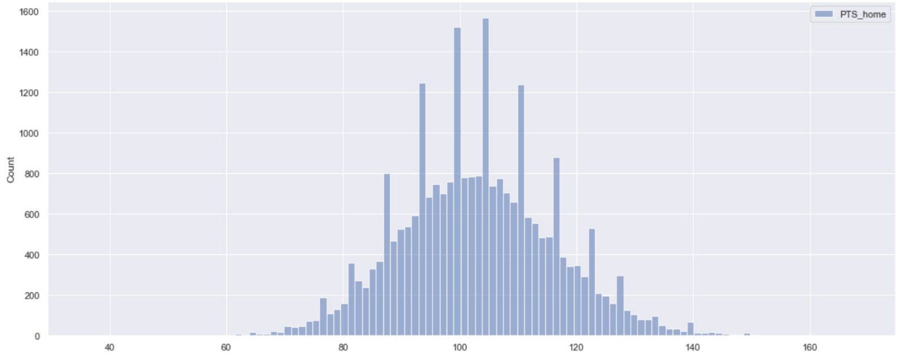
Refined the data file by changing the team IDs to their corresponding actual team names

1. **Compare points of home team (PTS_home) and points of away teams (PTS_away)**

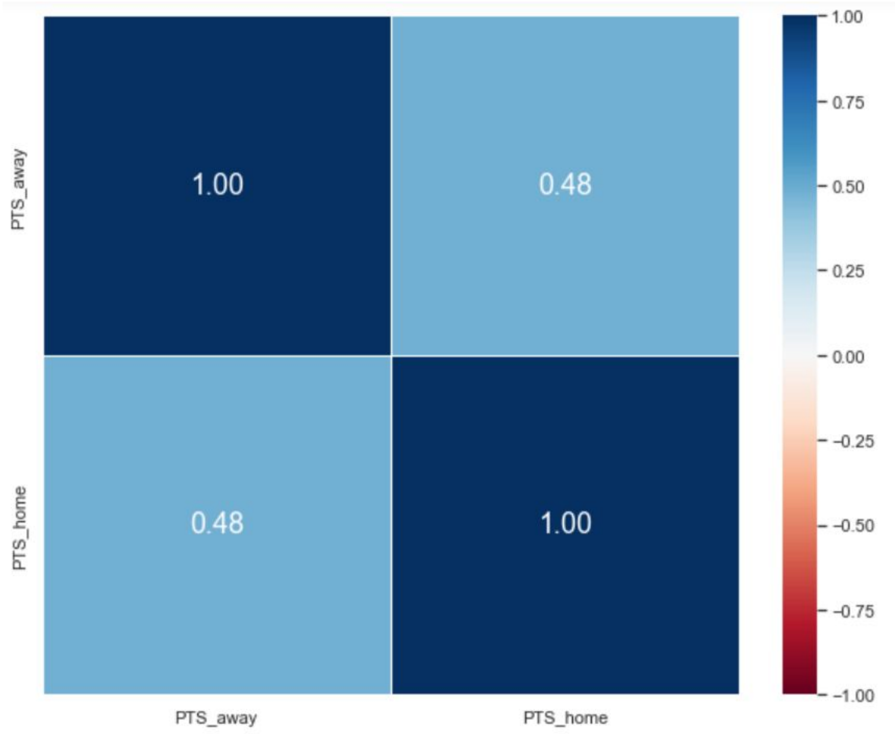
Exploratory Data Analysis: Comparison of points between home and away teams

- Similar distribution of points scored by both away and home teams, hence not a good indicator of home court advantage and its effectiveness





Exploratory Data Analysis: Comparison of points between home and away teams



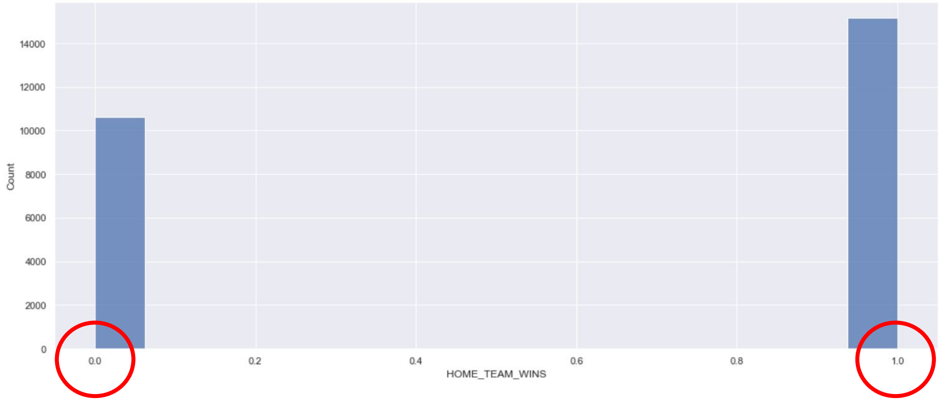
- Correlation between points scored by home and away teams quite low
- Not a good approach to determining the existence of home court advantage and its effectiveness

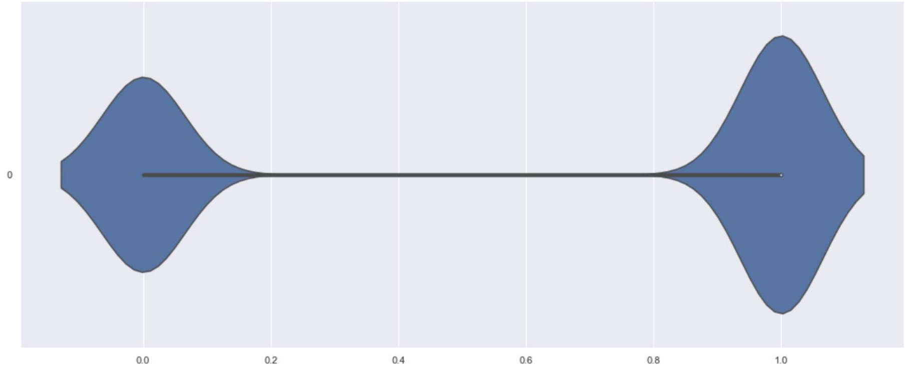
Exploratory Data Analysis

Directly looking at the wins obtained by the home teams

2. Exploring home team wins

Exploratory Data Analysis: Exploring home team wins





- Clear distinction between the number of losses and wins attained by the home teams
- Home teams definitely have an advantage during games

How effective is home court advantage?

Description of home team wins when grouped by home team names:

26/30 of home teams have higher frequency of wins as compared to losses

Probability of having a home team win is approximately 86.7% !

Exploratory Data Analysis

Begin by cleaning and refining the dataset for analysis later on

- Replace null values of points scored by 0
- Replace ‘1’ and ‘0’ under home team wins (HOME_TEAM_WINS) by ‘win’ and ‘loss’ respectively

3. Exploring effectiveness of home court advantage

Exploratory Data Analysis: Exploring effectiveness of home court advantage

- Obtain points margin from the perspective of home teams (win_loss)
- Median: all margins are positive; at least 50% of the home teams managed to score an edge over the away teams across seasons

SEASON	count	mean	std	min	25%	50%	75%	max
2003	1385.0	3.542238	11.620684	-32.0	-4.0	4.0	11.0	47.0
2004	1362.0	3.223201	12.340925	-35.0	-6.0	4.0	11.0	40.0
2005	1432.0	3.233939	12.338138	-33.0	-6.0	4.0	11.0	45.0
2006	1419.0	2.904863	12.730973	-50.0	-6.0	4.0	11.0	45.0
2007	1411.0	3.724309	13.640671	-42.0	-6.0	5.0	13.0	52.0
2008	1425.0	3.018947	13.289466	-58.0	-6.0	4.0	12.0	48.0
2009	1424.0	2.908708	13.226237	-50.0	-7.0	4.0	11.0	43.0
2010	1422.0	3.372011	12.790748	-51.0	-6.0	4.0	11.0	55.0
2011	1104.0	2.962862	13.297335	-39.0	-6.0	4.0	12.0	44.0
2012	1420.0	3.267606	13.149606	-45.0	-6.0	4.0	12.0	45.0
2013	1427.0	2.473721	13.279173	-48.0	-7.0	3.0	11.0	45.0
2014	1418.0	2.279972	13.428549	-54.0	-7.0	4.0	11.0	53.0
2015	1416.0	2.947740	13.597654	-51.0	-6.0	4.0	12.0	50.0
2016	1405.0	3.110320	13.898439	-44.0	-6.0	4.0	12.0	49.0
2017	1382.0	2.294501	13.837663	-48.0	-7.0	3.0	11.0	61.0
2018	1378.0	2.736575	14.403496	-56.0	-7.0	4.0	12.0	50.0
2019	1241.0	1.808219	14.287157	-41.0	-8.0	2.0	11.0	49.0
2020	1249.0	1.137710	15.198112	-57.0	-9.0	2.0	11.0	53.0
2021	1076.0	1.756506	15.256511	-68.0	-8.0	3.0	12.0	73.0

Exploratory Data Analysis: Exploring effectiveness of home court advantage

HOME_TEAM_ID	count	mean	std	min	25%	50%	75%	max
Atlanta Hawks	855.0	1.467836	12.928051	-41.0	-8.00	3.0	10.00	46.0
Boston Celtics	914.0	4.386214	12.749274	-44.0	-5.00	5.0	12.00	53.0
Brooklyn Nets	848.0	-0.306604	13.405328	-39.0	-9.25	1.0	9.00	44.0
Charlotte Hornets	772.0	-0.031088	13.017959	-68.0	-8.00	-1.0	9.00	61.0
Chicago Bulls	874.0	2.061785	13.751046	-56.0	-7.00	3.0	10.00	47.0
Cleveland Cavaliers	890.0	2.492135	14.123131	-41.0	-7.00	4.0	11.00	45.0
Dallas Mavericks	875.0	4.654857	13.039623	-34.0	-4.00	5.0	12.00	53.0
Denver Nuggets	848.0	5.558962	13.022143	-38.0	-4.00	6.0	14.00	52.0
Detroit Pistons	877.0	2.579247	12.913912	-44.0	-7.00	4.0	11.00	45.0
Golden State Warriors	871.0	5.819747	14.333976	-37.0	-4.00	7.0	15.00	50.0
Houston Rockets	874.0	4.577803	13.412702	-49.0	-5.00	5.0	13.00	49.0
Indiana Pacers	867.0	4.058824	12.617210	-37.0	-5.00	5.0	12.00	42.0
Los Angeles Clippers	874.0	2.979405	14.331600	-51.0	-6.00	4.0	12.00	49.0
Los Angeles Lakers	945.0	2.810582	13.330746	-48.0	-6.00	4.0	11.00	55.0
Memphis Grizzlies	852.0	2.314554	12.943323	-37.0	-7.00	4.0	11.00	73.0
Miami Heat	922.0	4.584599	13.324233	-47.0	-4.00	6.0	12.75	42.0
Milwaukee Bucks	852.0	2.427230	13.402671	-54.0	-7.00	3.0	10.00	47.0
Minnesota Timberwolves	818.0	-0.086797	13.311853	-33.0	-9.00	-2.0	9.00	43.0
New Orleans Pelicans	821.0	1.222899	12.868292	-58.0	-8.00	2.0	9.00	48.0
New York Knicks	812.0	-0.514778	13.213453	-50.0	-9.00	-2.0	8.00	43.0
Oklahoma City Thunder	868.0	3.407834	13.506444	-57.0	-5.00	5.0	12.00	45.0
Orlando Magic	847.0	1.146399	14.565408	-40.0	-8.00	2.0	11.00	54.0
Philadelphia 76ers	844.0	1.209716	13.731361	-51.0	-8.00	2.0	10.00	47.0
Phoenix Suns	859.0	2.951106	14.018218	-51.0	-6.50	4.0	11.00	48.0
Portland Trail Blaze	851.0	2.924794	12.749403	-40.0	-6.00	4.0	11.00	48.0
Sacramento Kings	818.0	-0.019560	12.934904	-49.0	-9.00	-1.0	9.00	44.0
San Antonio Spurs	915.0	7.732240	12.952083	-37.0	-1.00	9.0	16.00	40.0
Toronto Raptors	860.0	3.501163	12.514625	-30.0	-6.00	4.0	11.00	53.0
Utah Jazz	845.0	5.355030	13.231331	-45.0	-5.00	6.0	14.00	48.0
Washington Wizards	828.0	1.218599	12.733024	-38.0	-7.25	2.0	10.00	43.0

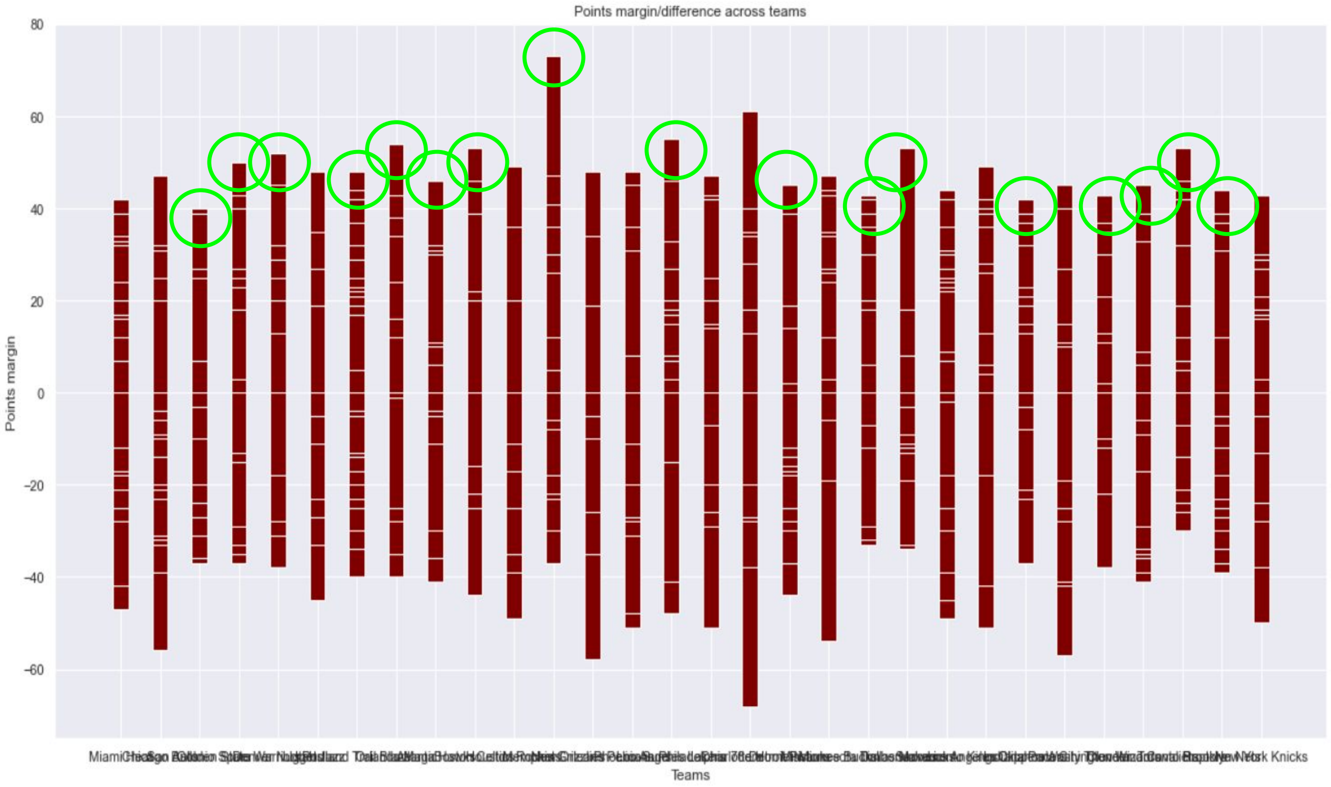
19/30 teams have a larger margin of win points than loss points when we compare the maximum and minimum points margin; 28/30 teams have a larger margin of win points as compared to loss points when we compare the 1st and 3rd quartile

Analytic Visualization and Pattern Recognition

To confirm that home court advantage is effective and show how effective it is

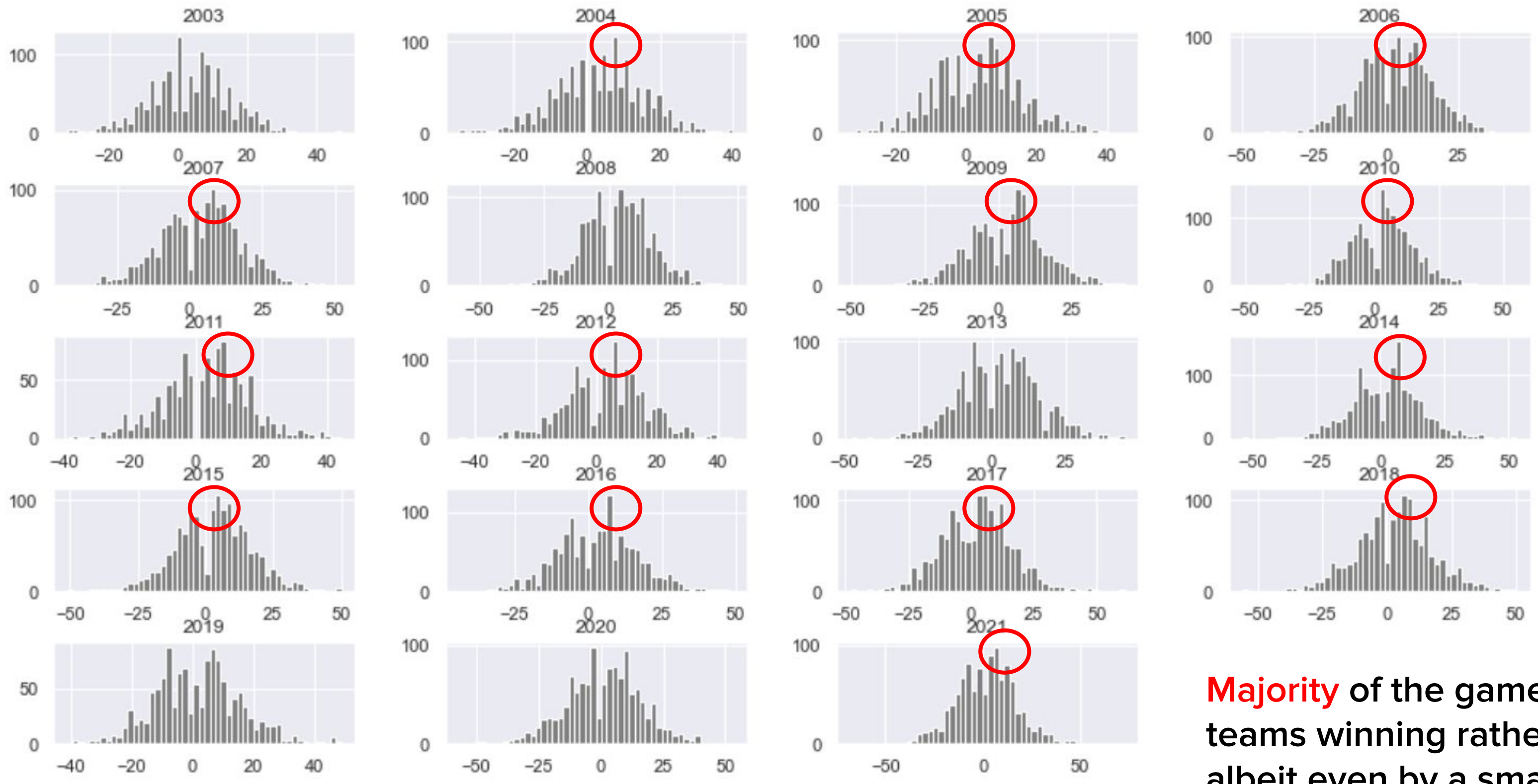
Visualization of the effectiveness of home court advantage using bar plots

Analytic Visualization and Pattern Recognition



Across the teams, approximately and confidently 16-17/30 teams have a larger positive points margin (win) than negative points margin (loss)

Analytic Visualization and Pattern Recognition



Across game seasons 2003 to 2021, most of the plots have peaks appearing on the positive (right) side of the win_loss difference axis

Majority of the games have home teams winning rather than losing, albeit even by a small margin

All these shows that home court advantage may indeed be quite effective in games

Values used to represent the rate of home court effectiveness for each team

```
HOME_TEAM_ID
Atlanta Hawks              1.467836
Boston Celtics             4.386214
Brooklyn Nets             -0.306604
Charlotte Hornets         -0.031088
Chicago Bulls              2.061785
Cleveland Cavaliers        2.492135
Dallas Mavericks           4.654857
Denver Nuggets             5.558962
Detroit Pistons            2.579247
Golden State Warriors      5.819747
Houston Rockets            4.577803
Indiana Pacers             4.058824
Los Angeles Clippers       2.979405
Los Angeles Lakers         2.810582
Memphis Grizzlies          2.314554
Miami Heat                 4.584599
Milwaukee Bucks            2.427230
Minnesota Timberwolves    -0.086797
New Orleans Pelicans       1.222899
New York Knicks           -0.514778
Oklahoma City Thunder      3.407834
Orlando Magic              1.146399
Philadelphia 76ers         1.209716
Phoenix Suns               2.951106
Portland Trail Blaze       2.924794
Sacramento Kings          -0.019560
San Antonio Spurs          7.732240
Toronto Raptors            3.501163
Utah Jazz                  5.355030
Washington Wizards         1.218599
Name: win_loss, dtype: float64
```

Machine Learning

Forming equations for best fit regression lines

Prediction of winning rate of teams based on models including and excluding home court advantage

Equations :

1. Log(win) = a + Log(points) + log(home_court) + Log(ast) + log(rebounds) + log(turnovers) + Error

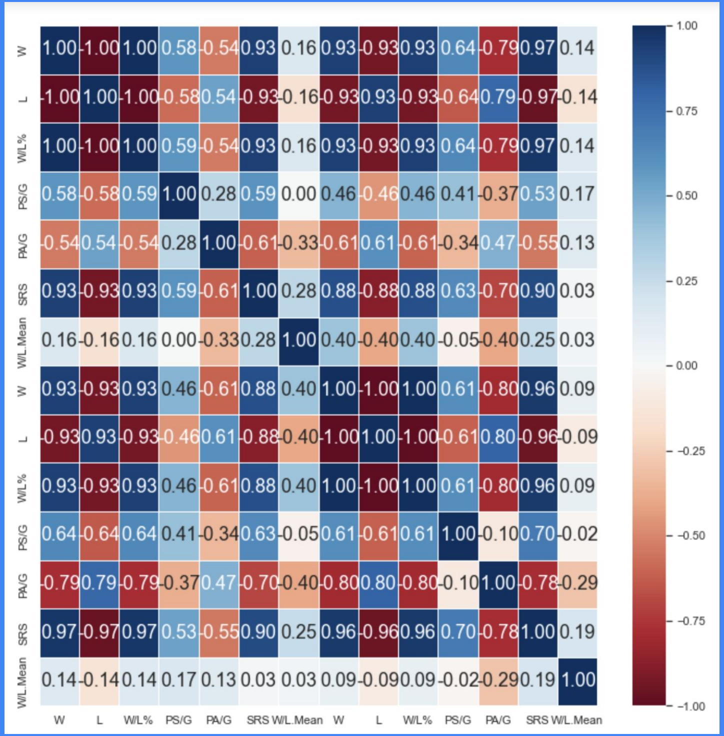
2. Win = a + b(points) + c(assists) + d(rebounds) + e(turnovers) + f(home_court advantage) + Error

	Rk	Team	G	MP	FG	FGA	FG%	3P	3PA	\
0	1.0	Minnesota Timberwolves*	82	241.2	41.6	91.0	0.457	14.8	41.3	
1	2.0	Memphis Grizzlies*	82	241.2	43.5	94.4	0.461	11.5	32.7	
2	3.0	Milwaukee Bucks*	82	240.9	41.8	89.4	0.468	14.1	38.4	
3	4.0	Charlotte Hornets	82	242.4	42.8	91.4	0.468	13.9	38.2	
4	5.0	Phoenix Suns*	82	240.6	43.7	90.1	0.485	11.6	31.9	
5	6.0	Atlanta Hawks*	82	240.3	41.5	88.3	0.470	12.9	34.4	
6	7.0	Utah Jazz*	82	240.6	40.6	86.2	0.471	14.5	40.3	
7	8.0	San Antonio Spurs	82	241.5	43.2	92.7	0.467	11.3	32.0	
8	9.0	Brooklyn Nets*	82	240.9	42.0	88.4	0.475	11.5	31.7	
9	10.0	Denver Nuggets*	82	241.5	41.7	86.3	0.483	12.7	35.9	
10	11.0	Los Angeles Lakers	82	243.7	41.6	88.8	0.469	12.0	34.5	
11	12.0	Boston Celtics*	82	242.7	40.7	87.4	0.466	13.2	37.1	
12	13.0	Chicago Bulls*	82	240.6	41.7	86.9	0.480	10.6	28.8	
13	14.0	Indiana Pacers	82	242.4	41.4	89.5	0.463	12.2	35.4	
14	15.0	Golden State Warriors*	82	240.6	40.5	86.4	0.469	14.3	39.4	
15	16.0	Sacramento Kings	82	241.5	40.5	88.1	0.460	11.4	33.2	
16	17.0	Miami Heat*	82	242.1	39.6	84.8	0.467	13.6	35.8	
17	18.0	Philadelphia 76ers*	82	241.5	39.4	84.5	0.466	11.6	31.8	
18	19.0	Houston Rockets	82	240.9	39.4	86.4	0.456	13.5	38.7	
19	20.0	Toronto Raptors*	82	242.1	40.6	91.3	0.445	11.9	34.2	
20	21.0	New Orleans Pelicans*	82	240.9	40.2	88.0	0.457	10.6	32.1	
21	22.0	Washington Wizards	82	241.8	40.6	86.0	0.472	10.5	30.6	
22	23.0	Los Angeles Clippers	82	241.2	40.1	87.4	0.458	12.8	34.2	
23	24.0	Dallas Mavericks*	82	240.9	39.3	85.1	0.461	13.1	37.4	
24	25.0	Cleveland Cavaliers	82	240.6	39.7	84.6	0.469	11.6	32.8	
25	26.0	New York Knicks	82	241.2	37.7	86.2	0.437	13.2	36.9	
26	27.0	Portland Trail Blazers	82	240.6	38.5	87.1	0.442	12.7	36.8	
27	28.0	Detroit Pistons	82	241.2	38.2	88.6	0.431	11.3	34.6	
28	29.0	Orlando Magic	82	241.2	38.3	88.3	0.434	12.2	36.9	
29	30.0	Oklahoma City Thunder	82	241.5	38.3	89.1	0.430	12.1	37.4	
30	NaN	League Average	82	241.4	40.6	88.1	0.461	12.4	35.2	

	3P%	...	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	W/L.Mean	W/L%
0	0.358	...	32.9	44.2	25.7	8.8	5.6	14.3	21.8	115.9	-0.086797	0.561
1	0.353	...	35.0	49.2	26.0	9.8	6.5	13.2	19.8	115.6	2.314554	0.683
2	0.366	...	36.5	46.7	23.9	7.6	4.0	13.4	18.2	115.5	2.427230	0.622
3	0.365	...	33.7	44.6	28.1	8.6	4.9	13.3	19.9	115.3	-0.031088	0.524
4	0.364	...	35.5	45.3	27.4	8.6	4.4	12.9	19.9	114.8	2.951106	0.780
5	0.374	...	33.9	44.0	24.6	7.2	4.2	11.9	18.7	113.9	1.467836	0.524
6	0.360	...	35.6	46.3	22.4	7.2	4.9	14.0	18.9	113.6	5.355030	0.598
7	0.352	...	34.3	45.3	27.9	7.6	4.9	12.7	18.1	113.2	7.732240	0.415
8	0.361	...	34.1	44.4	25.3	7.1	5.5	14.1	20.4	112.9	-0.306604	0.537
9	0.353	...	34.9	44.1	27.8	7.2	3.7	14.5	20.0	112.7	5.558962	0.585
10	0.347	...	34.5	44.0	24.0	7.6	5.2	14.5	20.2	112.1	2.810582	0.402
11	0.356	...	35.5	46.1	24.8	7.2	5.8	13.6	18.5	111.8	4.386214	0.622
12	0.369	...	33.7	42.3	23.9	7.1	4.1	12.8	18.8	111.6	2.061785	0.561
13	0.344	...	32.6	43.9	25.4	7.1	5.6	14.4	20.4	111.5	4.058824	0.305
14	0.364	...	35.7	45.5	27.1	8.8	4.5	14.9	21.0	111.0	5.819747	0.646
15	0.344	...	33.4	42.9	23.7	7.2	4.5	14.1	18.9	110.3	-0.019560	0.366
16	0.379	...	33.9	43.7	25.5	7.4	3.2	14.6	20.5	110.0	4.584599	0.646
17	0.364	...	33.8	42.3	23.7	7.7	5.3	12.5	19.4	109.9	1.209716	0.622
18	0.349	...	32.4	42.0	23.6	7.3	4.7	16.5	20.6	109.7	4.577803	0.244
19	0.349	...	32.0	45.3	22.1	9.0	4.6	12.5	19.6	109.4	3.501163	0.585
20	0.332	...	33.2	45.2	25.0	8.3	4.0	14.1	19.7	109.3	1.222899	0.439
21	0.342	...	34.1	43.1	25.0	6.4	5.0	13.1	18.8	108.6	1.218599	0.427
22	0.374	...	34.9	44.0	24.0	7.4	5.0	13.7	18.6	108.4	2.979405	0.512
23	0.350	...	33.8	43.0	23.4	6.7	4.0	12.5	19.7	108.0	4.654857	0.634
24	0.355	...	34.0	44.2	25.2	7.1	4.2	14.4	17.5	107.8	2.492135	0.537
25	0.357	...	34.6	46.1	21.9	7.0	4.9	13.3	20.4	106.5	-0.514778	0.451
26	0.346	...	32.5	42.9	22.9	8.0	4.5	14.5	21.1	106.2	2.924794	0.329
27	0.326	...	32.0	43.0	23.5	7.7	4.8	14.2	21.9	104.8	2.579247	0.280
28	0.331	...	35.2	44.3	23.7	6.8	4.5	14.5	19.7	104.2	1.146399	0.268
29	0.323	...	35.2	45.6	22.2	7.6	4.6	14.0	18.3	103.7	3.407834	0.293
30	0.354	...	34.1	44.5	24.6	7.6	4.7	13.8	19.6	110.6	2.749491	0.500

We combined the data achieved from home court advantage with the general statistics of all the teams.

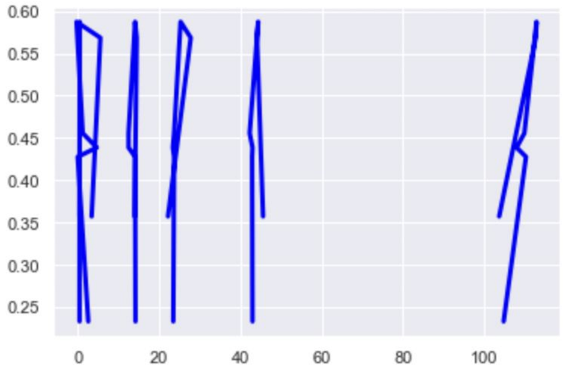
Home Court Effectiveness with Respect to other Factors



Results

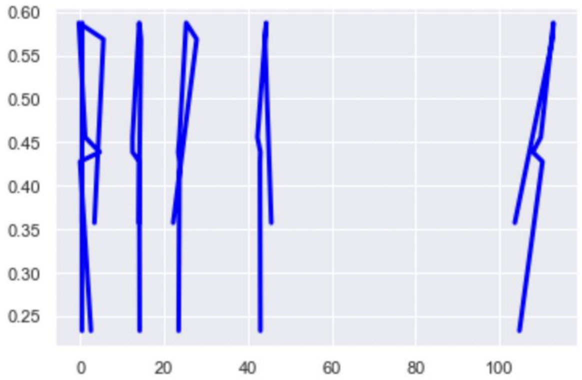
Regression Model including home court advantage:

```
[-1.07318390e-03  6.25294403e+00  3.40939420e-02 -6.58137709e-03
 -4.11076809e-02  6.92865561e-04]
r2 socre is 0.4597921375704427
mean_sqrd_error is== 0.01125599151458733
root_mean_squared error of is== 0.10609425768903484

[<matplotlib.lines.Line2D at 0x7fd2c8bdfac0>,
 <matplotlib.lines.Line2D at 0x7fd2c8bdf5e0>,
 <matplotlib.lines.Line2D at 0x7fd2c8c2e460>,
 <matplotlib.lines.Line2D at 0x7fd2c8bdf940>,
 <matplotlib.lines.Line2D at 0x7fd2c8bdffa0>,
 <matplotlib.lines.Line2D at 0x7fd2c8bdfdf0>]
```



Regression Model Excluding home court advantage:

```
[-1.25450413e-03  6.29720793e+00  3.43483838e-02 -6.46728487e-03
 -4.08242690e-02]
r2 score is 0.5228450664469511
mean_sqrd_error is== 0.009942194952627064
root_mean_squared error of is== 0.09971055587362385

[<matplotlib.lines.Line2D at 0x7fd2fbc67cd0>,
 <matplotlib.lines.Line2D at 0x7fd2fbc67f40>,
 <matplotlib.lines.Line2D at 0x7fd2fbc673a0>,
 <matplotlib.lines.Line2D at 0x7fd2fbc67d60>,
 <matplotlib.lines.Line2D at 0x7fd2fbc67e20>,
 <matplotlib.lines.Line2D at 0x7fd2fbc676a0>]
```



Statistical Inference

Home court advantage on its own does have a significant impact

However, when placed alongside other crucial factors, it does not become as effective

Concluding on whether home court advantage is effective

What we learnt from this project and Outcome of it

What we learned from our Project

- Random Forest Classifier
- A different way to use Multivariate Regression

Outcome of our Project

- The Random Forest Model used in Problem 1 can be used in the future to predict NBA All-Stars for upcoming seasons.
- The Multivariate regression line can be used to determine the effectiveness of various factors that can affect the Win-Loss% for a team.

Data Driven Insights and Recommendations

Problem 1

- Out of all the variables, Win Share proved to be the most effective as it had the highest correlation with All-Star.
- Our approach, however, could be improved if it took into consideration the votes of fans, players, and managers. This is due to the fact that voting is a factor in selecting whether or not a player is an All-Star.

Problem 2

- Home court advantage does play a part in affecting the team's ability to win. However, there are many statistical factors that can affect it which are hard to track for example, how often does a player does something out of the ordinary or how long it is required for a player to perform at his optimum?

References

- NBA Stats 2016-2019 - https://www.kaggle.com/datasets/toniabiru/nba-stats-20162019-seasons
- NBA Stats 2020-2021 - https://www.kaggle.com/datasets/toniabiru/20202021-season-stats
- Games.csv - https://www.kaggle.com/datasets/nathanlauga/nba-games?select=teams.csv
- Basketball Reference NBA Summary - https://www.basketball-reference.com/leagues/NBA_2022.html#all_confs_standings_E