NBA Analysis

By Shashwat Singh and Michelle Dry-Moran

Group 106 / May 11, 2020

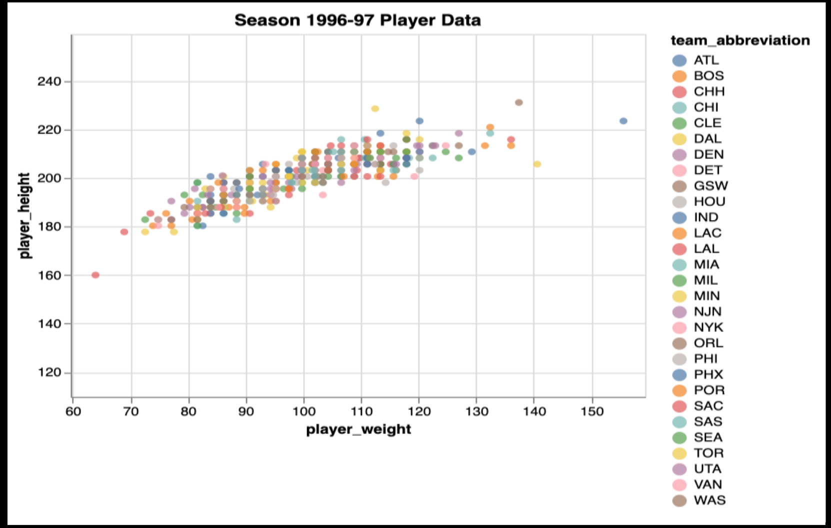
Proposal

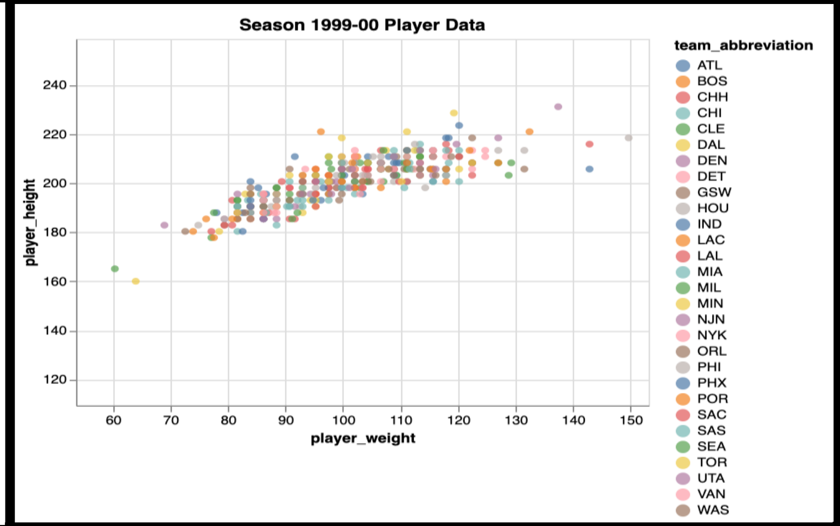
Prior to COVID19, Rutgers Men’s Basketball was ranked in the top 25 for the first time in 40 years, with the potential to play in the NCAA tournament.

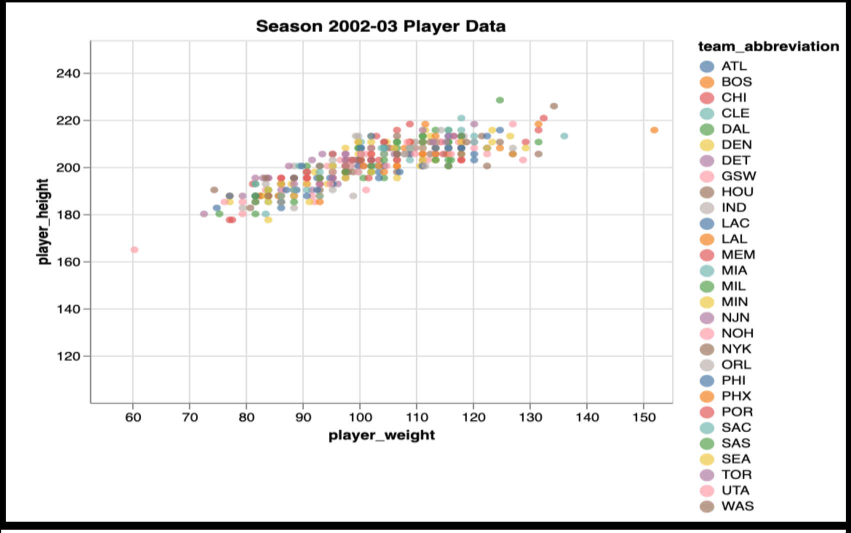
Therefore, it seemed fitting that our project be on BASKETBALL.

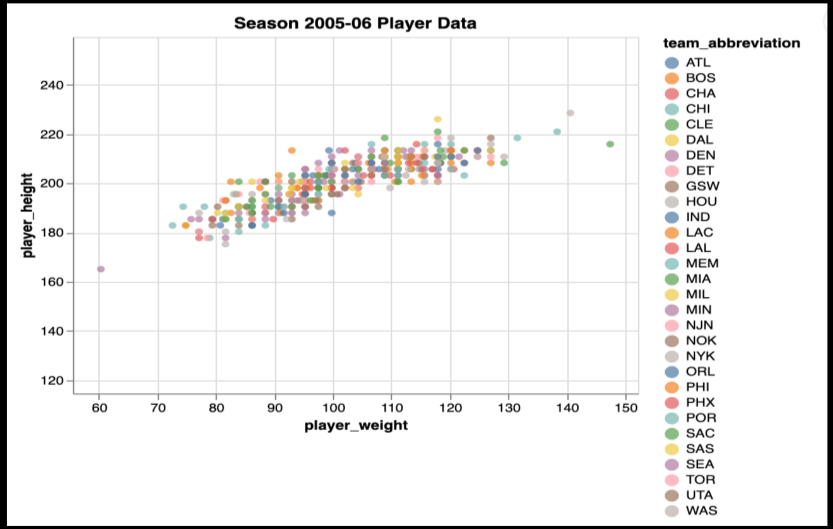
We chose to analyze NBA data to find out how basketball has changed over the years and if there really is a home court advantage. In addition, we tried to predict NBA playoff contenders.

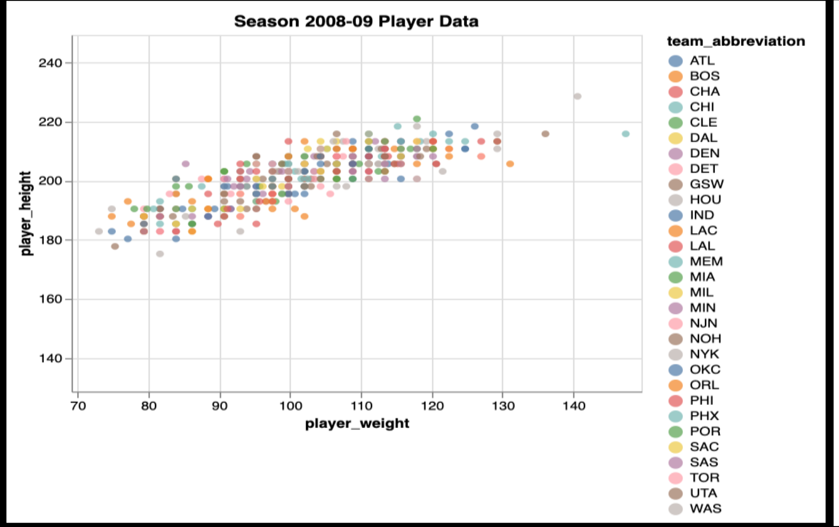
Player Attributes (height/weight) - Today vs Predecessors

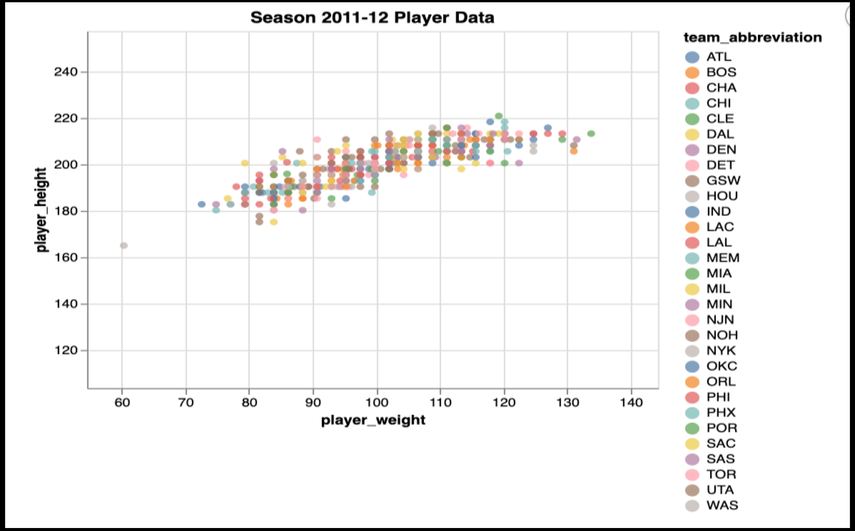




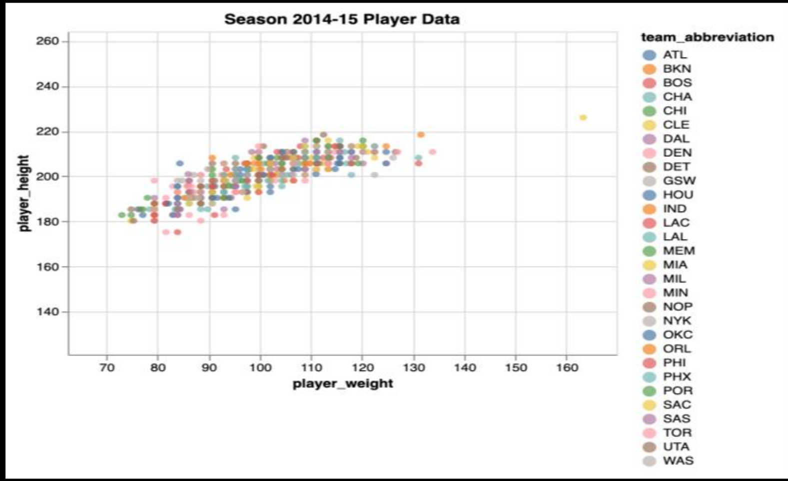


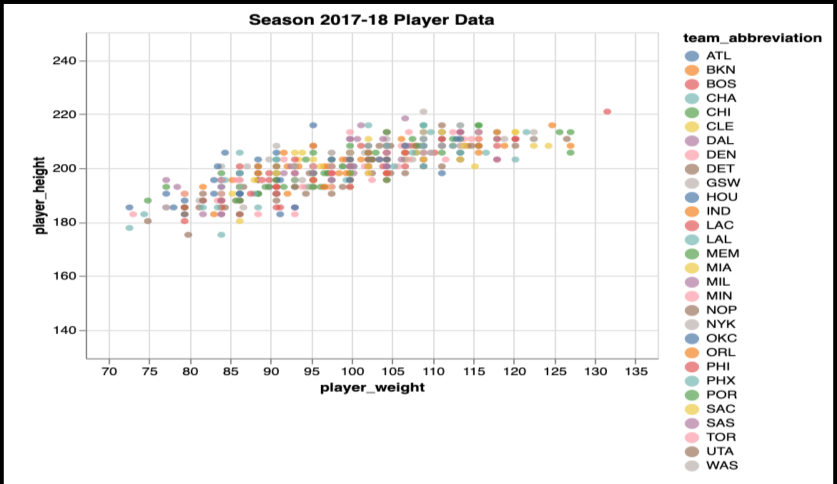


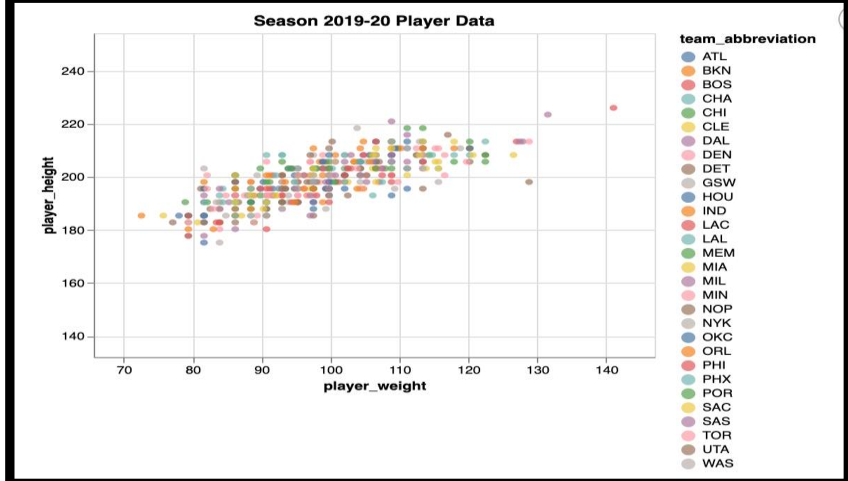




Player Attributes (height/weight) - Today vs Predecessors (continued)







<u>How are today's players attributes different from their predecessors?</u>

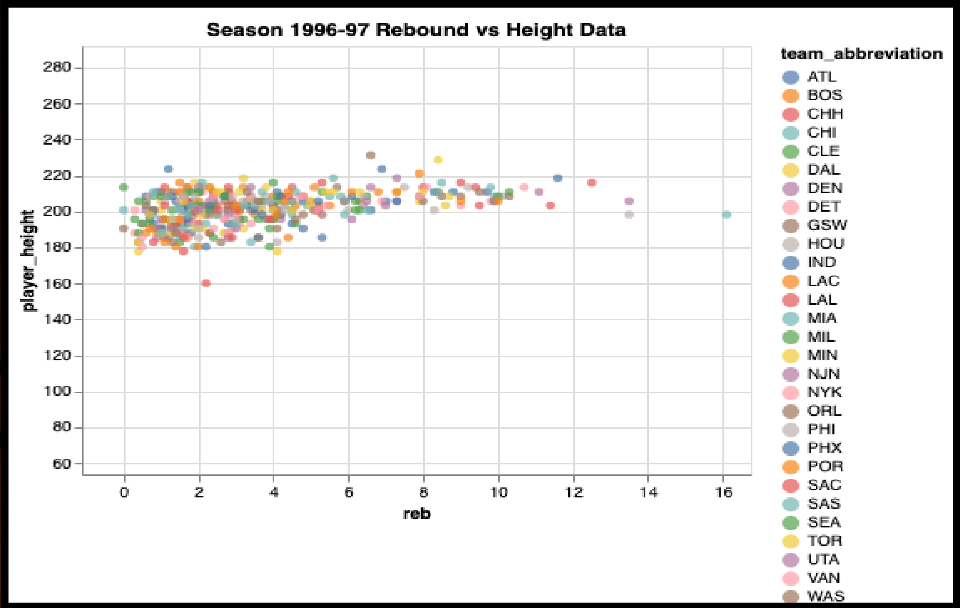
Overall, player's weights have shifted marginally upward as the seasons 2011-2014 period. From the 2017 to the 2019 season, we can see that there are fewer sub 80 kilogram athletes as well. Following the 2011-2014 period as well, there were no longer any players under 175 centimeters tall. In the earlier seasons, there were 1-4 players in the 160-175 height group. The 190-210 cm height group is the most populated height group through all of the seasons.

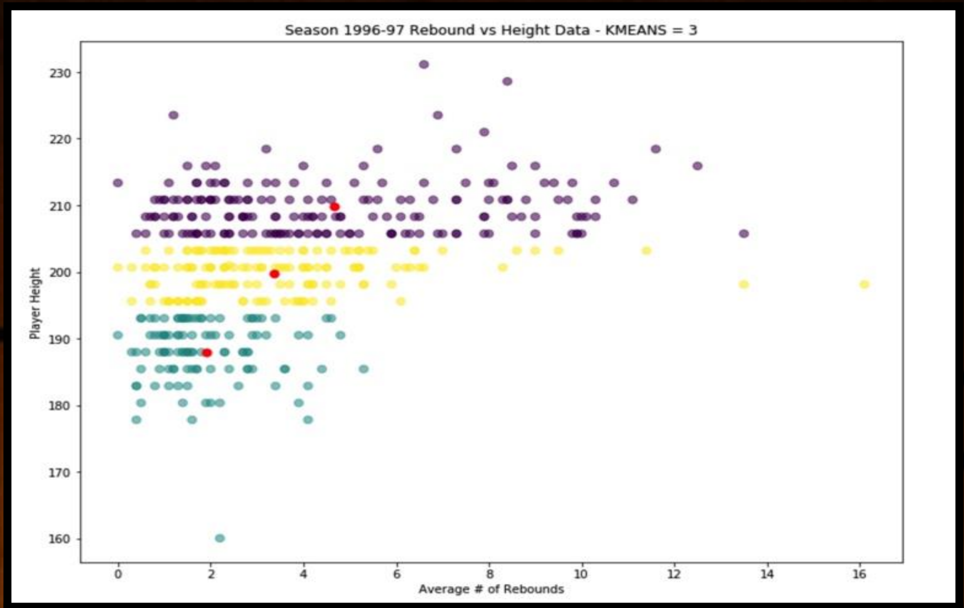
While doing this analysis, we found certain player's heights and weights did not change as their age did. A player's height should remain largely stable barring extraordinary circumstances. A player's weight on the other hand should change. In the earlier seasons, weight reporting wasn't checked as thoroughly and this could explain the lack of movement for this metric. Additionally, height is generally a more important measurement which has become more strictly monitored as the NBA moved forward.

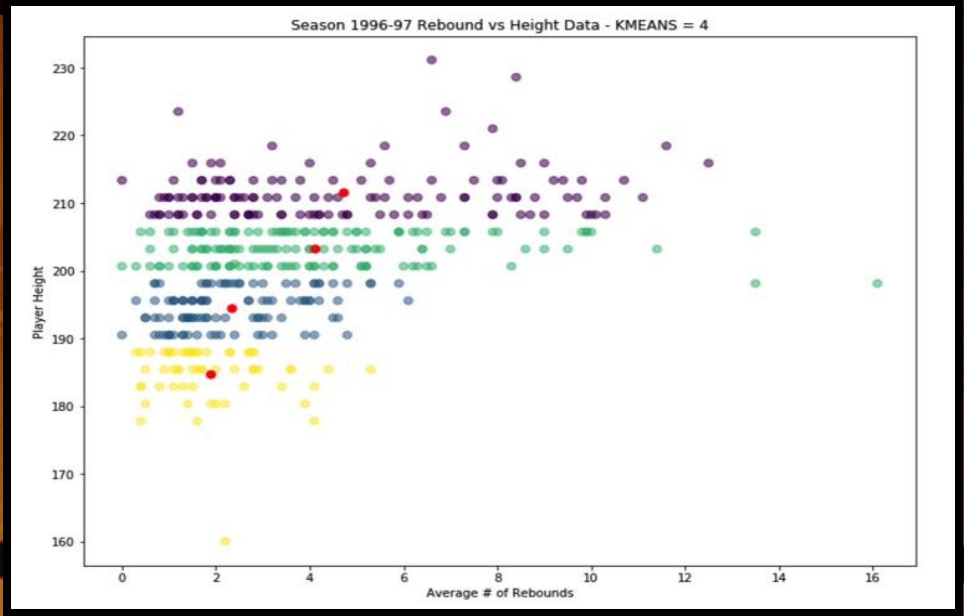
* Note: 2020 season was cancelled, and therefore not represented.

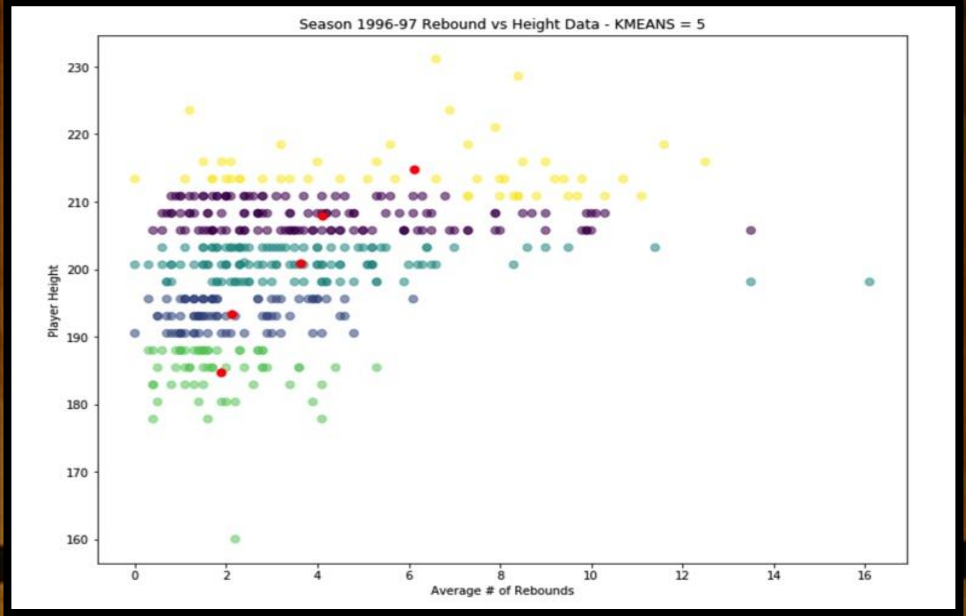
Do Taller Players Rebound More?

(1996 – 1997 Dataset)







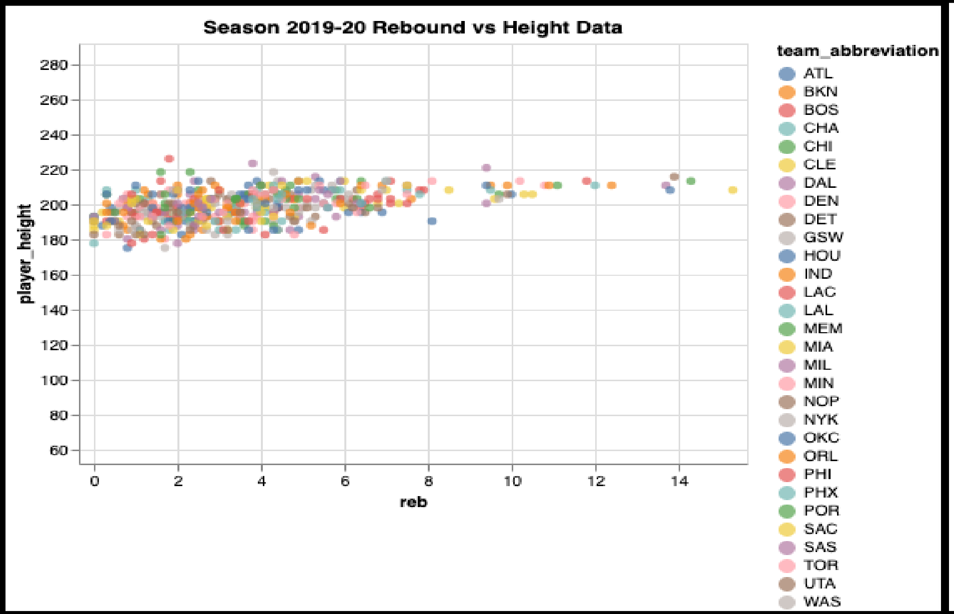


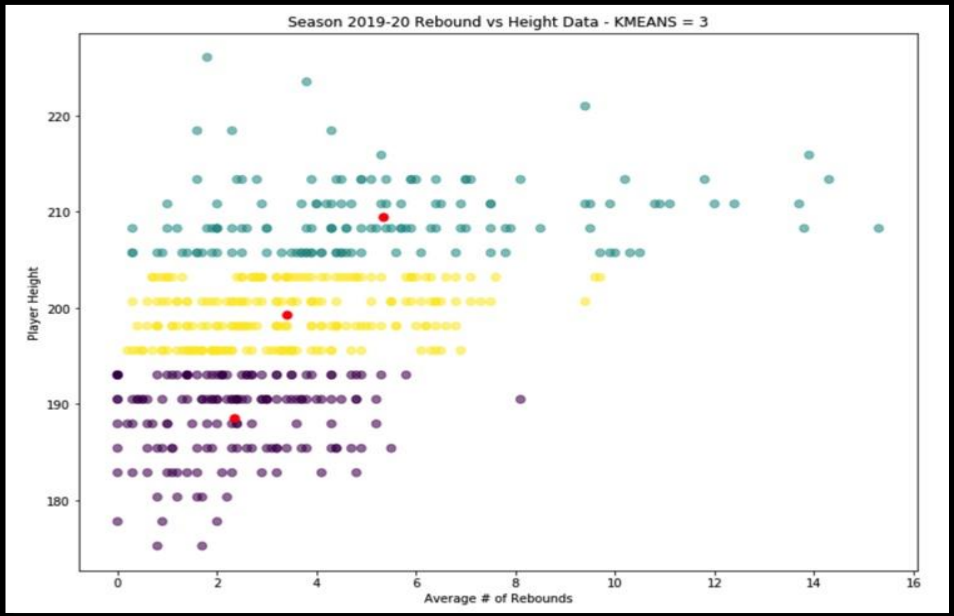
Looking at all of the data, you can clearly see that the players with the higher height, tend to have more rebounds.

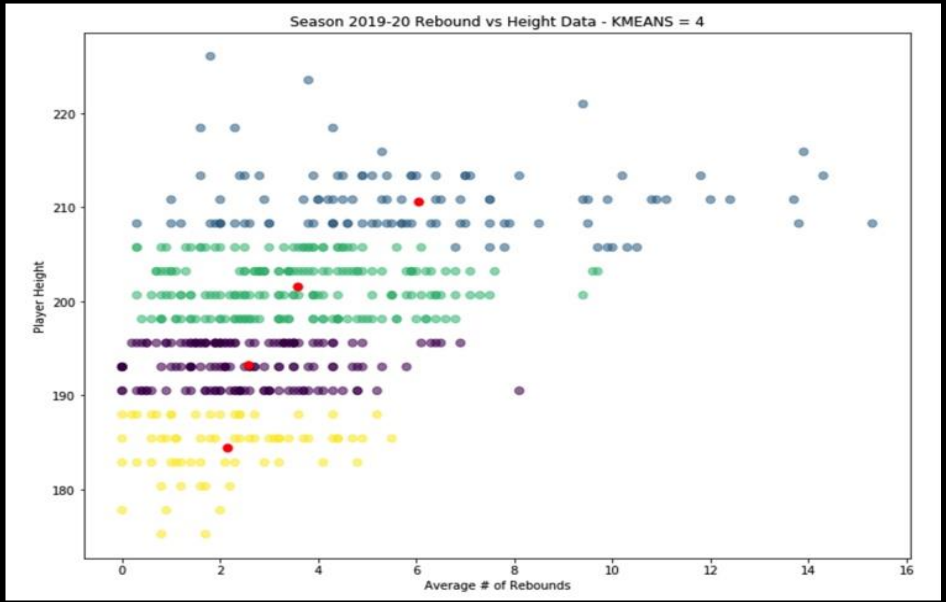
However height is not always a factor in the 1996-1997 dataset.

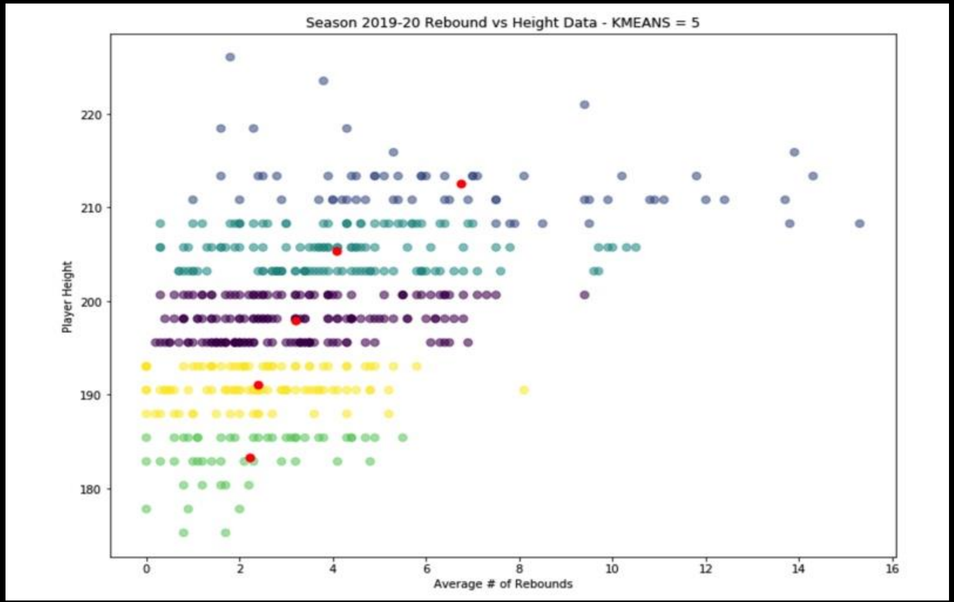
Do Taller Players Rebound More?

(2019 – 2020 Dataset)





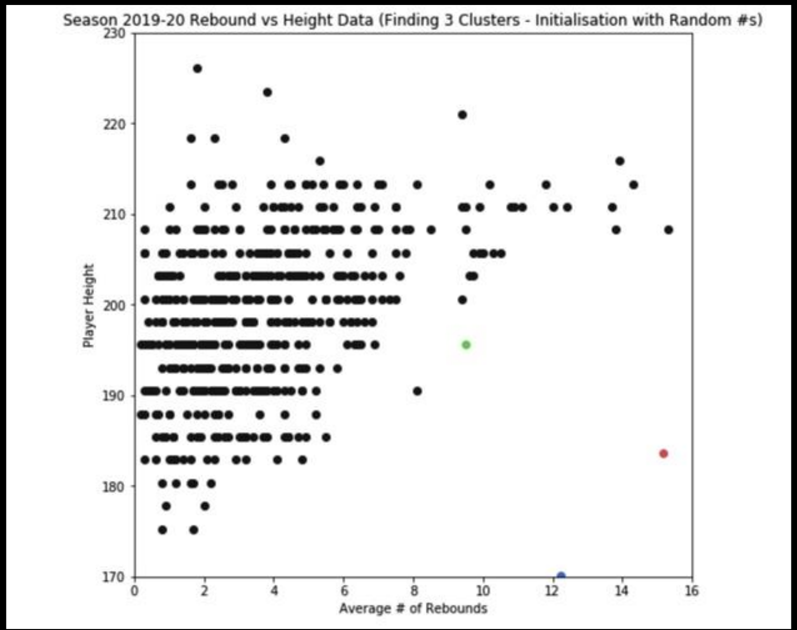


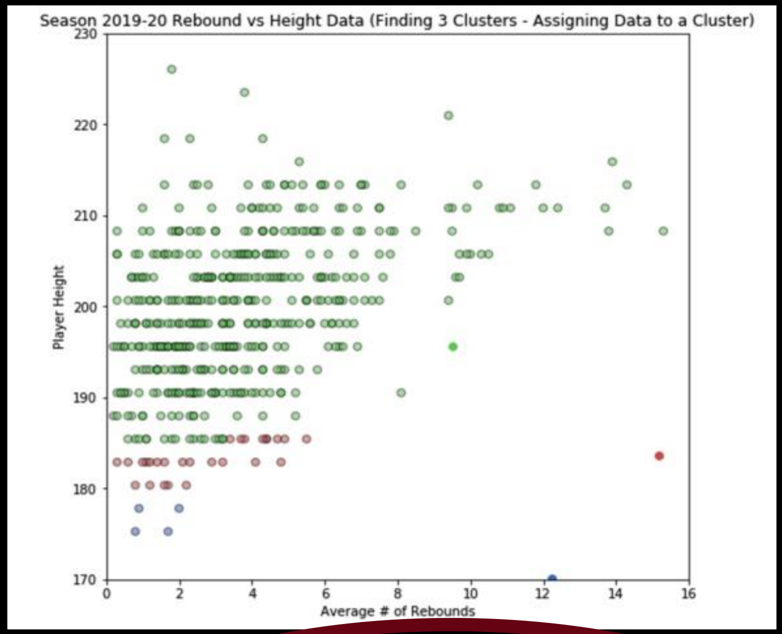


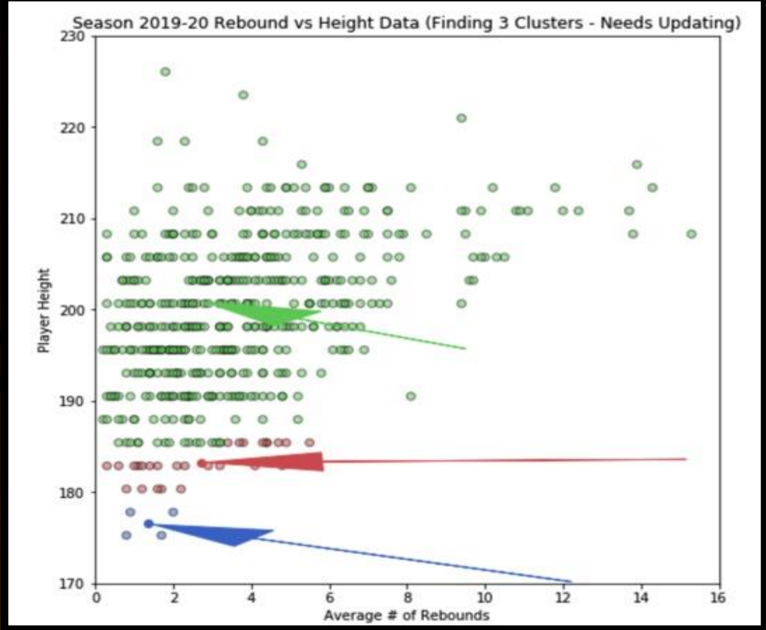
Again, there is a clear link to the number of rebounds as compared to height.

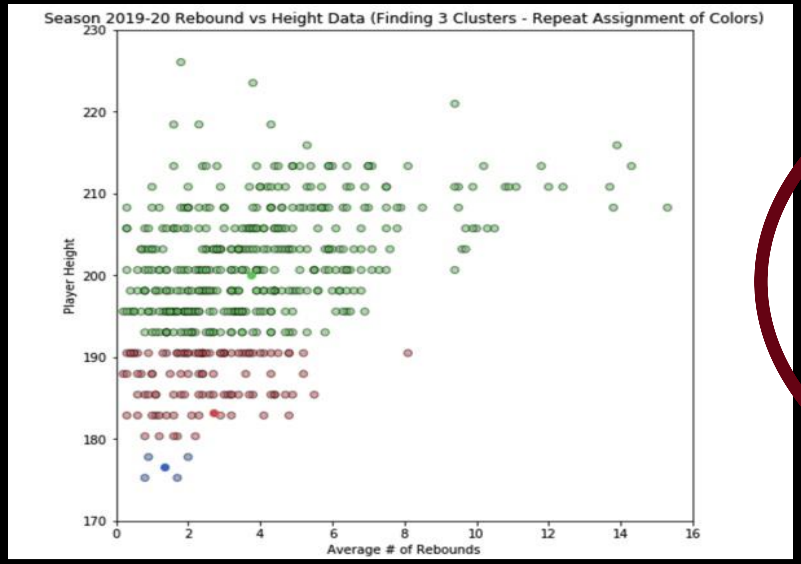
The data shows that 2019-2020 shows that this is even more prevalent today.

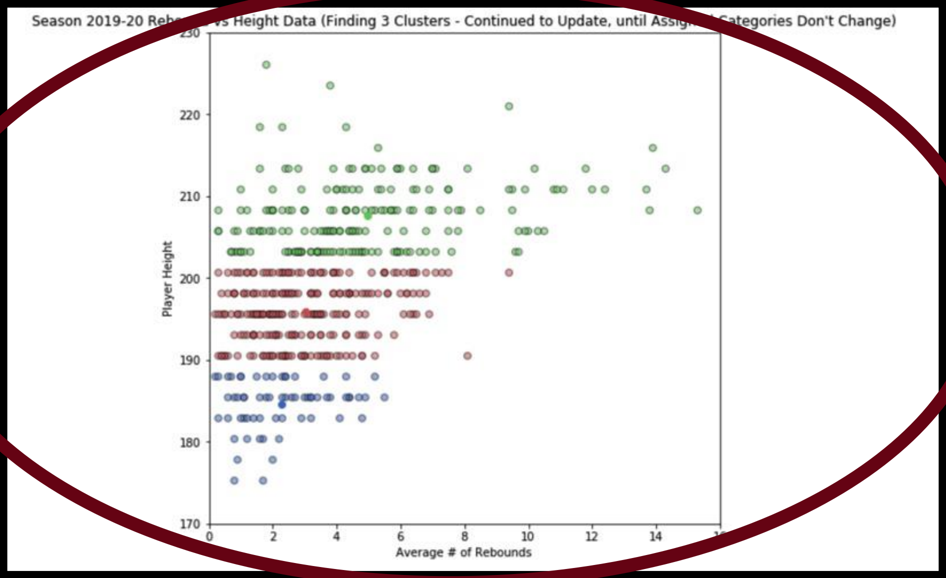
Manually Compute K=3 for:
2019-2020 Height vs Rebound Dataset









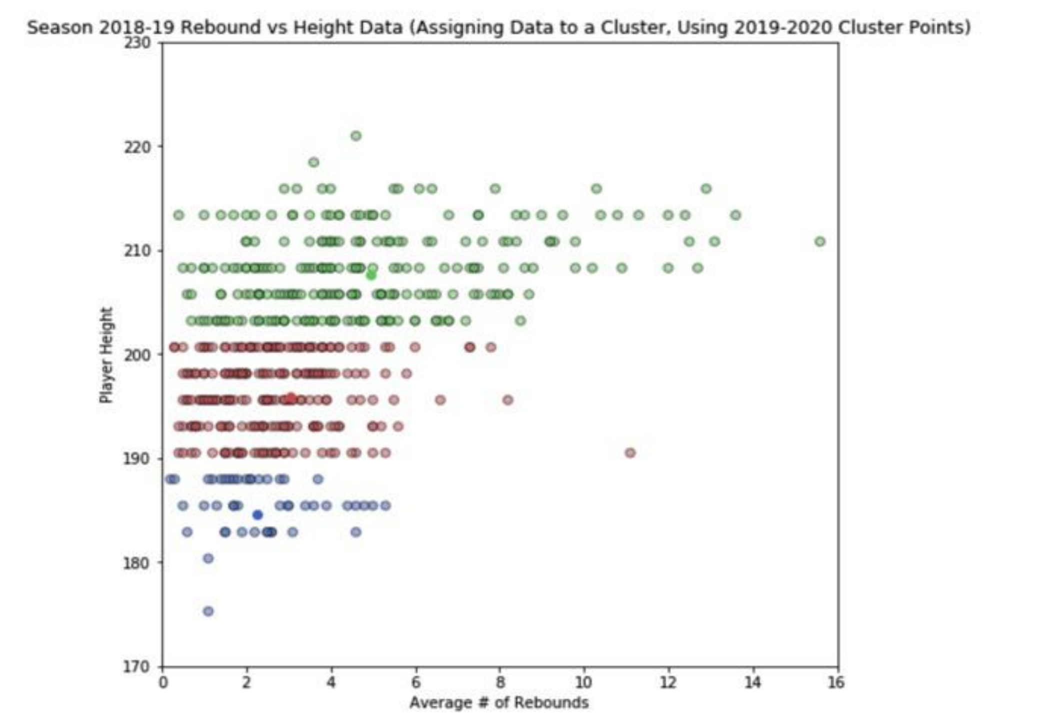


- ✓ The Manually Generated K = 3 graph matches the KMeans.fit function graph (from previous slide)

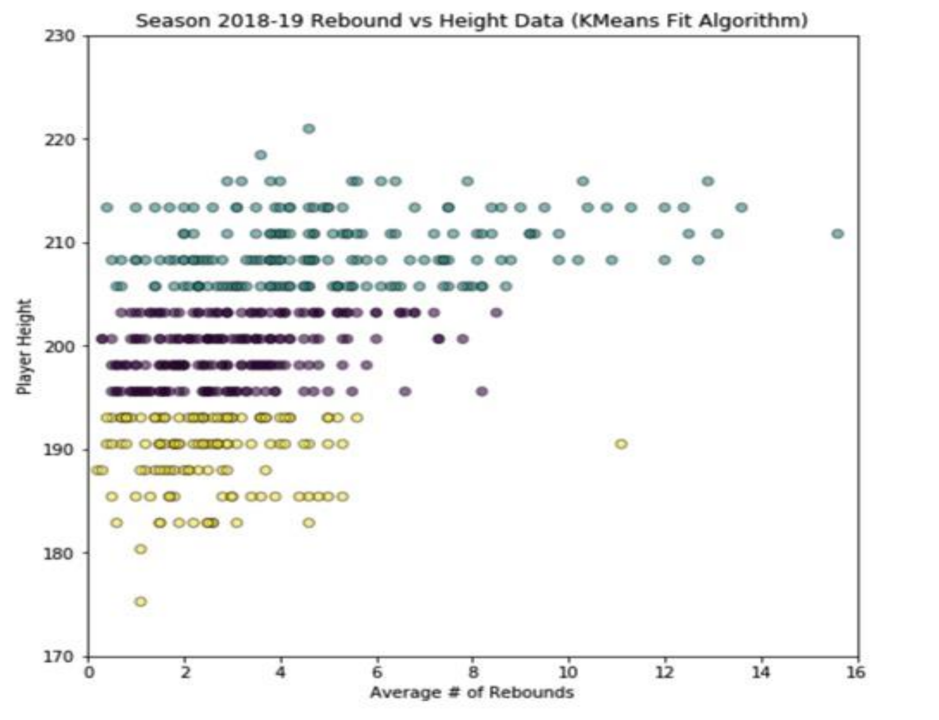
Will Centroids from 2019-20 data fit 2018-19 data?

Using the 2019-20 Centroid Points that were manually generated, see if this will fit the 2018-19 dataset.

2018-19 Manually Run K=3:



2018-19 Kmeans.Fit Generated:

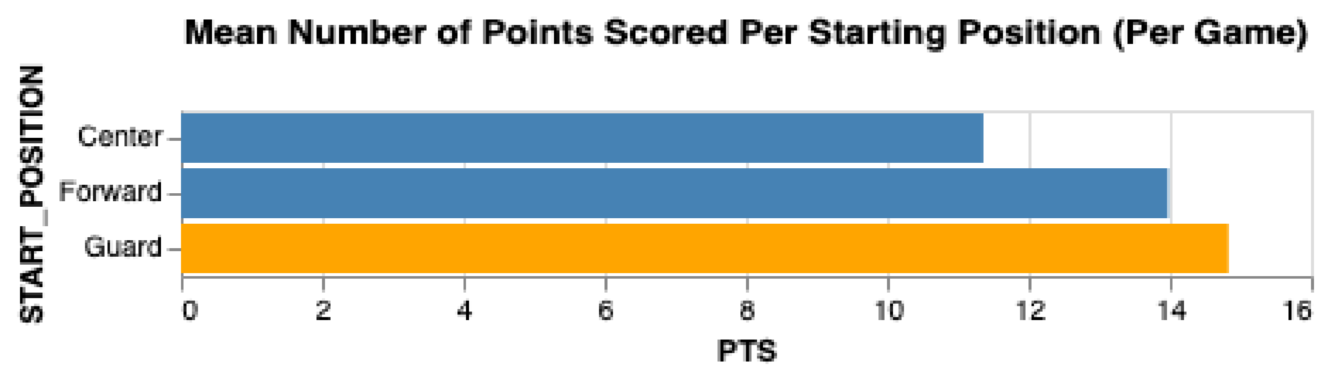


✓ These Two K-Means Graphs Similar

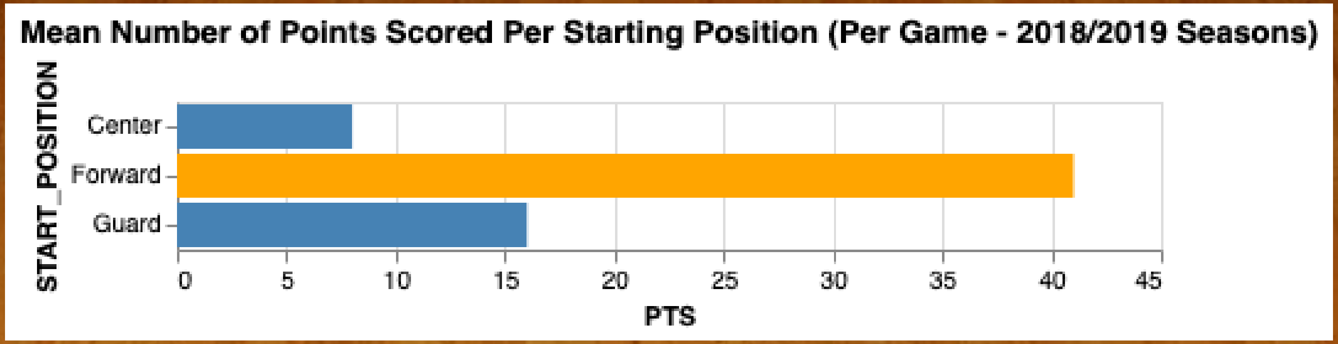
(However – it is not 100% Accurate)

Which Position Scores the Most Points?

ALL YEARS:



2018 & 2019:

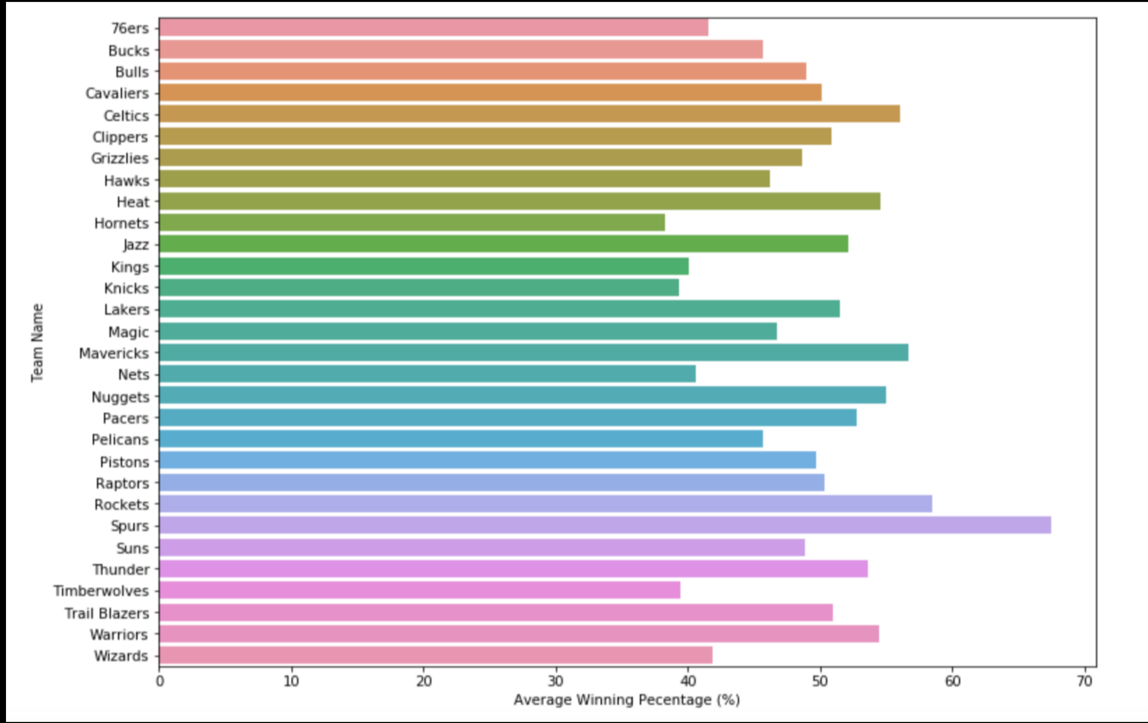


Looking at all of the data, it seems that, on average, the Guard scores the most points.

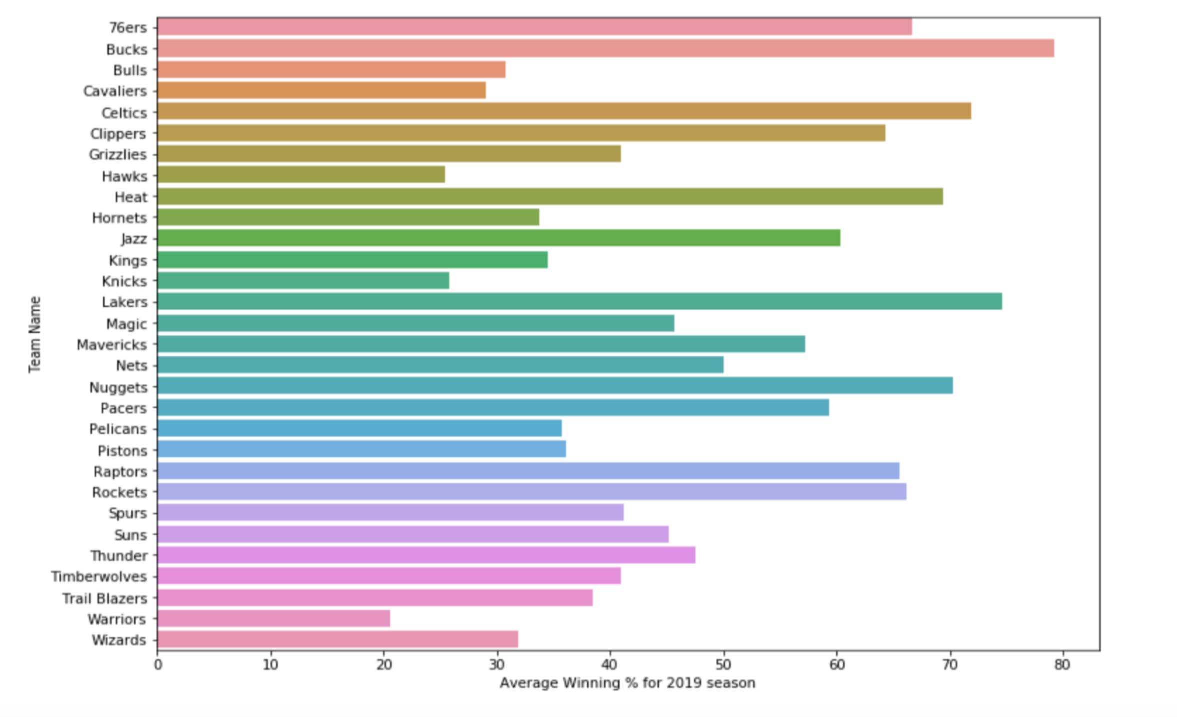
However, when we look at just the 2018 and 2019 seasons, we see that this has changed, and the Forward typically scores more. This is potentially likely linked to the height.

Analysis of Average Winning Percentage by Teams

ALL SEASONS



2019 SEASON

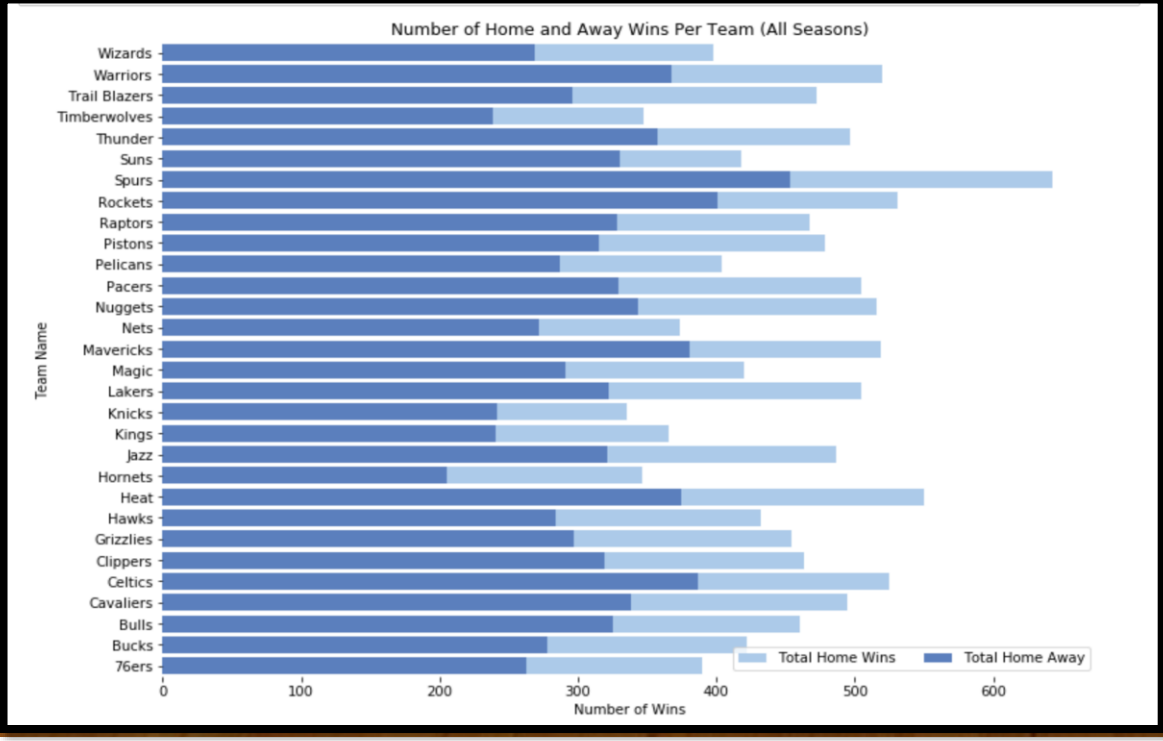


The tables above show that on average (spanning all seasons), the Spurs have a much higher winning percentage than any other teams, and the Heat had the lowest average winning percentage.

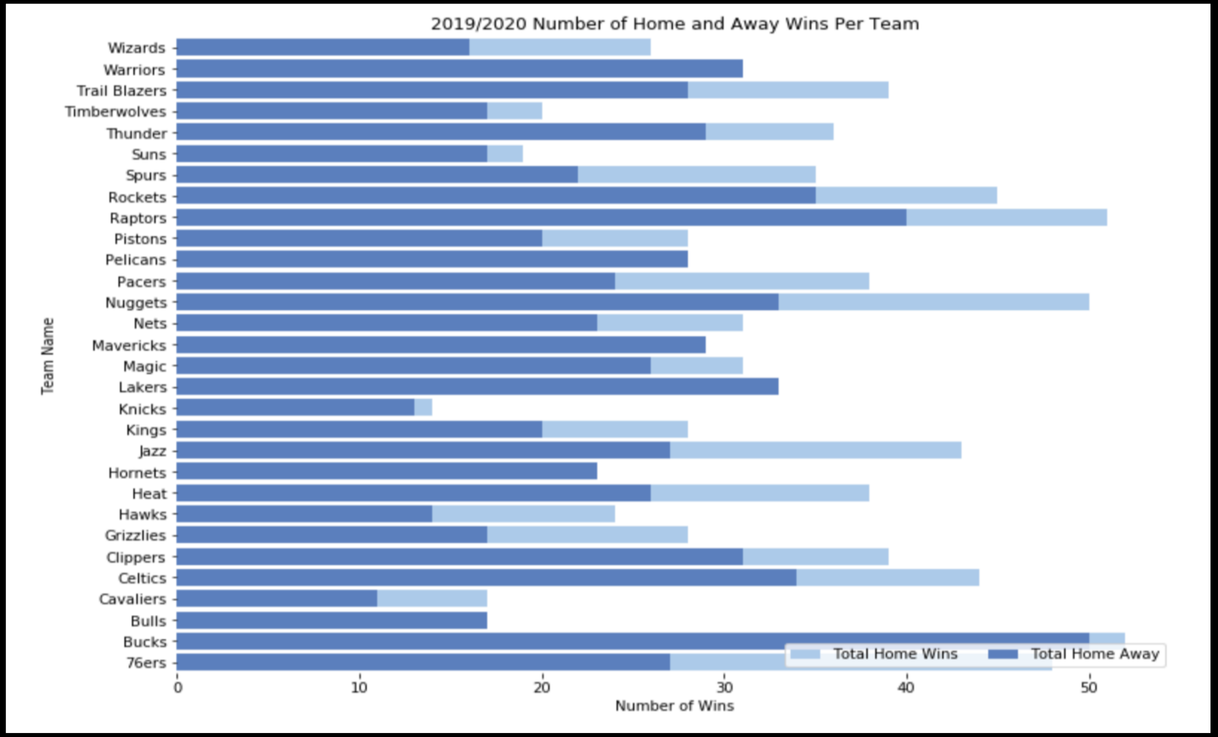
However, when you look at just the 2019 season, it seems that the Bucks and Lakers have the highest average winning percentages, while the Warriors have a low winning percentage. Additionally, while the Heat may have had many bad seasons, their current season is going strong.

Analysis of Home vs Away Games

ALL SEASONS



2019 SEASON



Overall, the tables show that the games have been won more at home, rather than being won on the road. Even when we look at the most current season (2019/2020), home games have more wins.

So there really is such a thing as a home field advantage!

Machine Learning - Decision Tree Classifier
Can We Predict Who Will Win (Home vs Away)?

Baseline and Assumptions:

Baseline for home teams winning = 54.72%

Assuming the home team ALWAYS wins – the F1 score = 38.70%

```
games_total = games2019['HOME_TEAM_WINS'].count()
games_won = games2019['HOME_TEAM_WINS'].sum()
winrate = games_won/games_total

print('Home Win percentage: {0:.2f}%'.format(100 * winrate))

#find baseline for home team winning
```

```
Home Win percentage: 54.72%
```

```
y_pred = [1] * len(y_true)

from sklearn.metrics import f1_score

print('F1: {0:.2f}%'.format(f1_score(y_true, y_pred, pos_label = None, average = 'weighted') * 100))

#assuming the home team always wins, calculate a baseline f1 score
```

```
F1: 38.70%
```

Machine Learning - Decision Tree Classifier
Can We Predict Who Will Win (Home vs Away)?

Features for our Decision Tree:

Does the home team have a better 3 point completion rate?
Does the home team have better free throw completion?
Does the home team have better rebounding?

```python
#create features to be used in decision tree

#whether the home team had a better 3 pt completion rate
games2019['home_3adv'] = games2019['FG3_PCT_home'] > games2019['FG3_PCT_away']
games2019['away_3adv'] = games2019['FG3_PCT_home'] < games2019['FG3_PCT_away']

#whether the home team had better free throw completion
games2019['home_ft'] = games2019['FT_PCT_home'] > games2019['FT_PCT_away']
games2019['away_ft'] = games2019['FT_PCT_home'] < games2019['FT_PCT_away']

#whether the home team had better rebounding
games2019['home_rb'] = games2019['REB_home'] > games2019['REB_away']
games2019['away_rb'] = games2019['REB_home'] < games2019['REB_away']

games2019.head()
```

Machine Learning - Decision Tree Classifier
Can We Predict Who Will Win (Home vs Away)?

Run The Decision Tree and Recalculate the F1 score

```
#make new dataframe with the chosen features
features = games2019[['home_3adv', 'away_3adv', 'home_ft', 'away_ft', 'home_rb', 'away_rb']].values

#take decision tree classifier from sklearn library
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier(random_state = 14)
```

```
#a scorer to determine decision tree effectiveness

from sklearn.metrics import f1_score, make_scorer

scorer = make_scorer(f1_score, pos_label = None, average = 'weighted')
```

```
from sklearn.model_selection import cross_val_score

#cross validation score will allow us to see if this algorithm produces better results

dtc = DecisionTreeClassifier(random_state = 14)
scores = cross_val_score(dtc, features, y_true, scoring = scorer)

print('F1: {0:.4f}%'.format(np.mean(scores) * 100))
```

```
F1: 72.0768%
```

The new F1 score of 72.0768% shows that the decision tree classifier with these features is able to predict which team will win and whether it will be a home team victory.

While the 3 point completion rate, free throw rate, and rebounds we used as features are from the games we are trying to predict, these can be replaced with an aggregation of the teams stats from previous games.

Any Questions?

Resources used:

https://github.com/clprice32/Predicting-NBA-Game-Winners

https://benalexkeen.com/k-means-clustering-in-python