deliverable1

Craig Le

10/5/2020

##Introduction

For this project I am analyzing statistics from the 2018-19 NBA season. I chose to analyze last season's game data because this season took place during the Covid pandemic, and as a result there is incomplete team data because not every team played all 82 regular season games. Basketball is one of the sports I watch the most, and I wanted to do a deeper dive on team statistics.

Also some specifics I would also like to study: home team win percentage, home team average margin of victory compared to road margin of victory, home team fourth quarter scoring statistics, # of home team victories when the game is within five points.

##Datasets

All of my data was compiled from basketball-reference.com. Basketball-reference is part of the Sports Reference collection of sites. Basketball-reference receives their data from Sportradar which is the official stat partner of the NBA. Some limitations to this source is that there are few unnecessary columns, but those can easily be omitted. Also the column name for home and away points is the exact same, so that needed to distinguished.

```
NBA_stats <- read_csv("2018-2019_NBA_game_logs - Sheet1.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_character(),
##   `Start (ET)` = col_character(),
##   `Visitor/Neutral` = col_character(),
##   PTS = col_double(),
##   `Home/Neutral` = col_character(),
##   PTS_1 = col_double(),
##   Attendance = col_double(),
##   Notes = col_character()
## )
```

```
summary(NBA_stats)
```

```
##      Date            Start (ET)        Visitor/Neutral         PTS
##  Length:1230        Length:1230        Length:1230        Min.   : 68.0
##  Class :character   Class :character   Class :character   1st Qu.:101.0
##  Mode  :character   Mode  :character   Mode  :character   Median :110.0
##                                                           Mean   :109.8
##                                                           3rd Qu.:118.0
##                                                           Max.   :168.0
##  Home/Neutral            PTS_1          Attendance        Notes
##  Length:1230         Min.   : 77.0   Min.   :10079   Length:1230
##  Class :character    1st Qu.:104.0   1st Qu.:16682   Class :character
##  Mode  :character    Median :112.0   Median :18256   Mode  :character
##                      Mean   :112.6   Mean   :17857
##                      3rd Qu.:121.0   3rd Qu.:19461
##                      Max.   :161.0   Max.   :21852
```

From the summary it is easily seen that home teams in general score more points thn the away team. The home team has a higher minimum, mean, and median compared to away teams. However, the max value in the away team column is higher, but these points are outliers considering how far off they are away from the mean and median. In the attendance data there is a large gap between the minimum and maximum fan attendance. This is not much of a concern because there are more popular and less popular and less popular teams in the NBA, so in general fan attendance should have quite a lot of variance to it.

##Variables

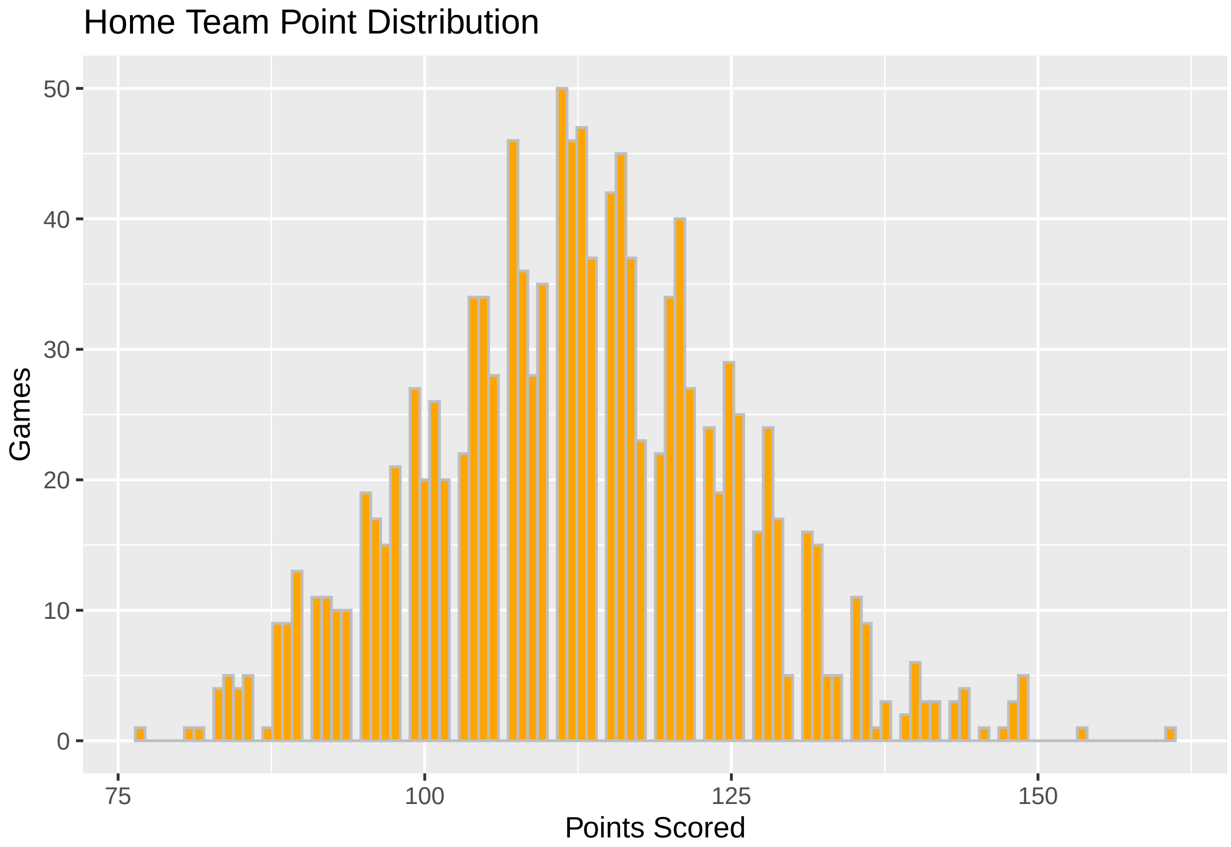
Variables I would like to analyze: *`Date` - when game took place (char)* `Visitor/Neutral` - the away team (char) *`Home/Neutral` - the home team (char)* `PTS` - points scored by away team (double) *`PTS_1` - points scored by home team (double)* `Attend` - total fan attendance (double) *`FT%` - away team free throw percentage* `FT_1%` - home team free throw percentage *`TOV` - away team number of turnovers* `TOV_1` - home team number of turnovers *`FG%` - away team field goal percentage* `FG_1%` - home team field goal percentage *`3P%` - away team 3 point fg percentage* `3P_1%` - home team 3 point fg percentage

##Analysis

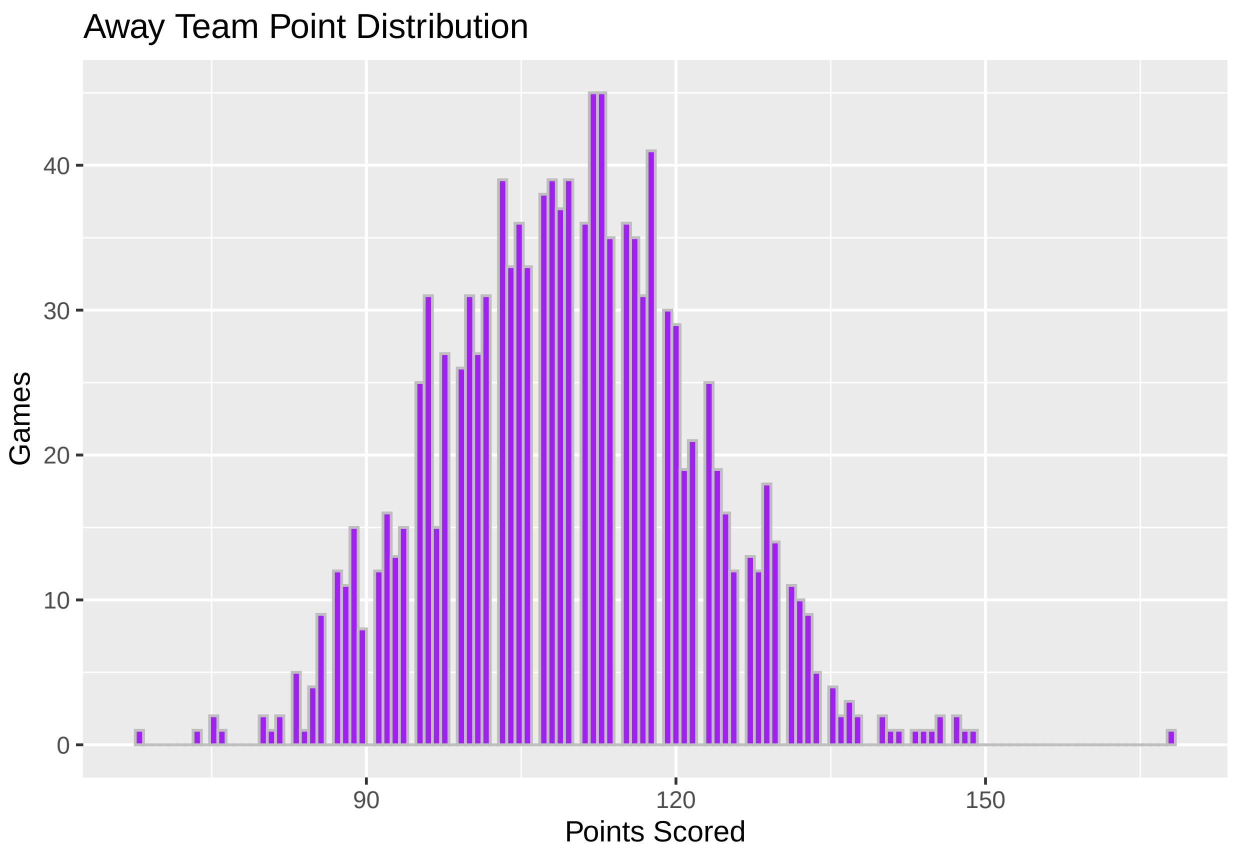
```r
NBA_stats <- read_csv("2018-2019_NBA_game_logs - Sheet1.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_character(),
##   ‘Start (ET)‘ = col_character(),
##   ‘Visitor/Neutral‘ = col_character(),
##   PTS = col_double(),
##   ‘Home/Neutral‘ = col_character(),
##   PTS_1 = col_double(),
##   Attendance = col_double(),
##   Notes = col_character()
## )
```

```r
##Ttl_Home_points <- sum(NBA_stats$PTS_1)
##Ttl_Away_points <- sum(NBA_stats$PTS)
ggplot(data = NBA_stats) +
  geom_histogram(mapping = aes(x = PTS_1), binwidth = .8, fill = 'orange', color = 'gray') +
  labs(title = "Home Team Point Distribution", x = "Points Scored", y = "Games")
```



```
ggplot(data = NBA_stats) +
  geom_histogram(mapping = aes(x = PTS), binwidth = .8, fill = 'purple', color = 'gray') +
  labs(title = "Away Team Point Distribution", x = "Points Scored", y = "Games")
```

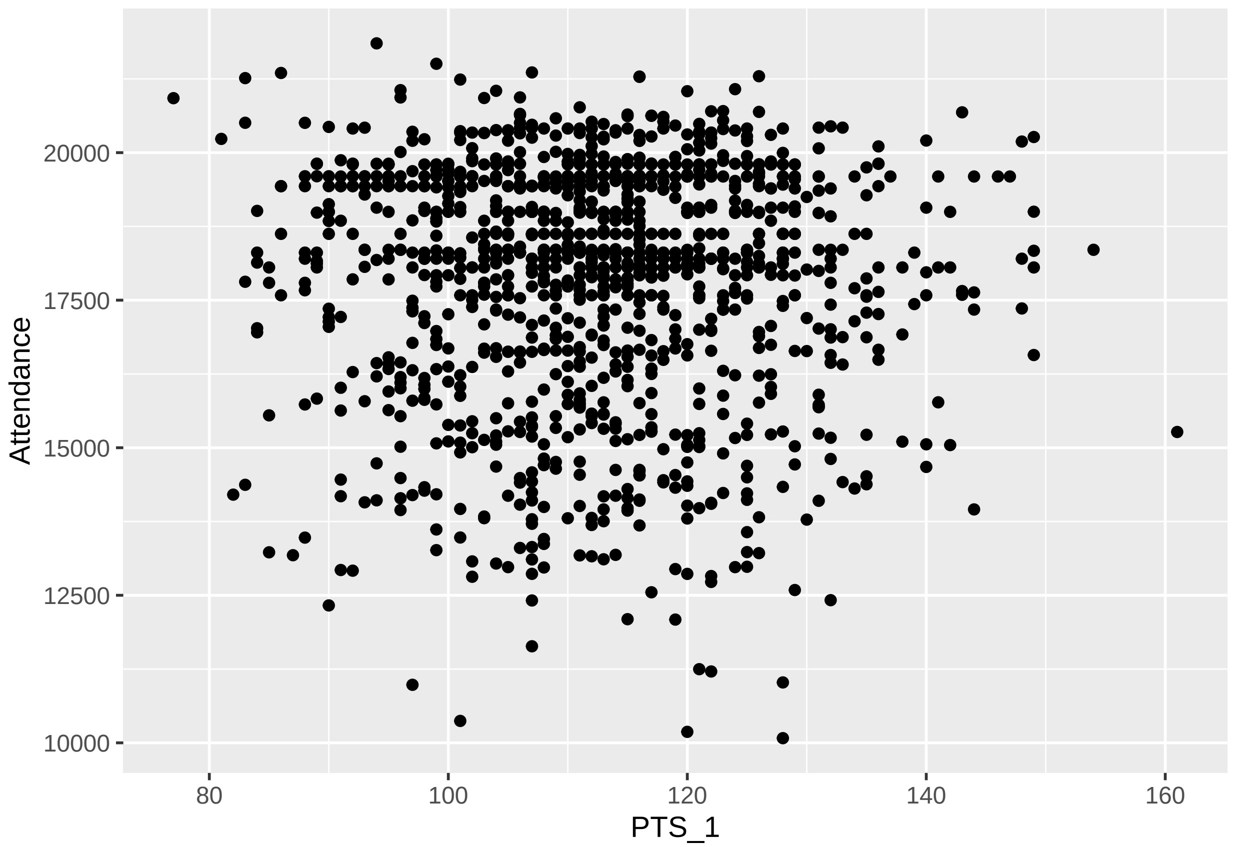


Each histogram shows the total point distribution for every game played in the season. The first graph is home teams and the bottom is away teams. When comparing both of the distributions they look very similar. However, in a more detailed look seem to generally have a larger count of games that are higher scoring. Also the away teams have more games with lower scoring totals. It is easily seen in these graphs where the average is, and it is easy to pick out some of the outlier scores.

```
NBA_stats <- read_csv("2018-2019_NBA_game_logs - Sheet1.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_character(),
##   ‘Start (ET)‘ = col_character(),
##   ‘Visitor/Neutral‘ = col_character(),
##   PTS = col_double(),
##   ‘Home/Neutral‘ = col_character(),
##   PTS_1 = col_double(),
##   Attendance = col_double(),
##   Notes = col_character()
## )
```

```
ggplot(data = NBA_stats) +
  geom_point(aes(x = PTS_1, y = Attendance))
```



This scatter plot is aimed at trying to visualize a potential correlation between fan attendance and points scored by the home team. The graph does not really depict any sort of clear correlation between these two variables. There is however a larger concentration of higher points scored and higher attendance numbers, but there are also many points that are lower scoring with the similar attendance numbers.
A better analysis might be to take average fan attendance for each team and average points scored. Also a limitation to this is that not every NBA stadium holds the same number of people, so one stadium might not even have a max capacity as other stadiums.

##Further Questions

Some of the initial questions I am asking are as follows: Could we develop a model that predicts how much better an NBA team plays at home court vs away? Will they have a better field goal, free throw, 3pt percentage? Will the home team be more "clutch" in close games? (home team fourth quarter scoring statistics, # of home team victories when the game is within five points) I would also like to use some data from the games that were in the covid bubble. In the bubble there were no fans in attendance, so the atmosphere would be completely different.