Sample-1

For your data project,

1. Introduction: Tell me what problem you are working on? Why is this problem interesting and important. State specific research questions your group will work on. Introduce recent research done in area related to your problem. You can pack all this together to motivate us. Do keep it short, to the point, and interesting.

The National Basketball Association (NBA) is the top basketball league in the world, and currently has 30 teams, with about 450 players in the league at a time. Each year, there are awards handed out to a number of top players in the league. There are five awards specifically that I am interested in looking into. The first three of these are given to one player each season. The Most Valuable Player (MVP) award is given to the best performing player in the regular season, and has been given out since 1956. The Defensive Player of the Year (DPOY) award is given the most outstanding defensive player of the regular season, and has been given out since 1983. The Sixth Man of the Year (6MOTY) award is given to the best performing player in the regular season who came off of his team's bench, or did not start. 6MOTY has been given out since 1983. The last two awards, All-NBA and All-Defensive, are given out to a numbers of player each season. All-Defensive honors have been given out to 10 players (except in cases of a tie) since its inaugural year in 1969. All-NBA was given to 10 players from 1950 to 1988, and has been given to 15 players since. Each award is broken into "teams" (1st and 2nd for All-Defensive and 1st, 2nd, and 3rd for All-NBA), but for this analysis I will only be looking at if the player made any of the teams, as generally there is not considered to much distinction in the quality of players between the 1st, 2nd, or 3rd All-NBA/All-Defensive teams. For each of these awards, I want to find out what sort of statistical performance a player needs to achieve the honor. NBA statistics are generally broken into two category: counting stats and advanced stats. Counting stats, generally communicated on a per game basis, are stats that can be recorded by simply watching the game. For example, one can watch a game and see a player score 11 points. If you see this player score 15 points in the next game, you know they have a 13 points per game average for those two games. Advanced stats are created by weighing basic counting stats in various formulas or models, and except for a few of them, are not available for the earliest NBA seasons. For example, the Value Above Replacement Player (VORP) advanced statistic only has values from the 1974 season onwards, and is not available for seasons before then.. The awards are voted on by a cast of about 75 rotating media members, of which all certainly have their own biases which impact how they vote. However, I do not believe there is a suitable way to quantify this bias, and thus it will be ignored in this analysis.

Data: Tell me about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.

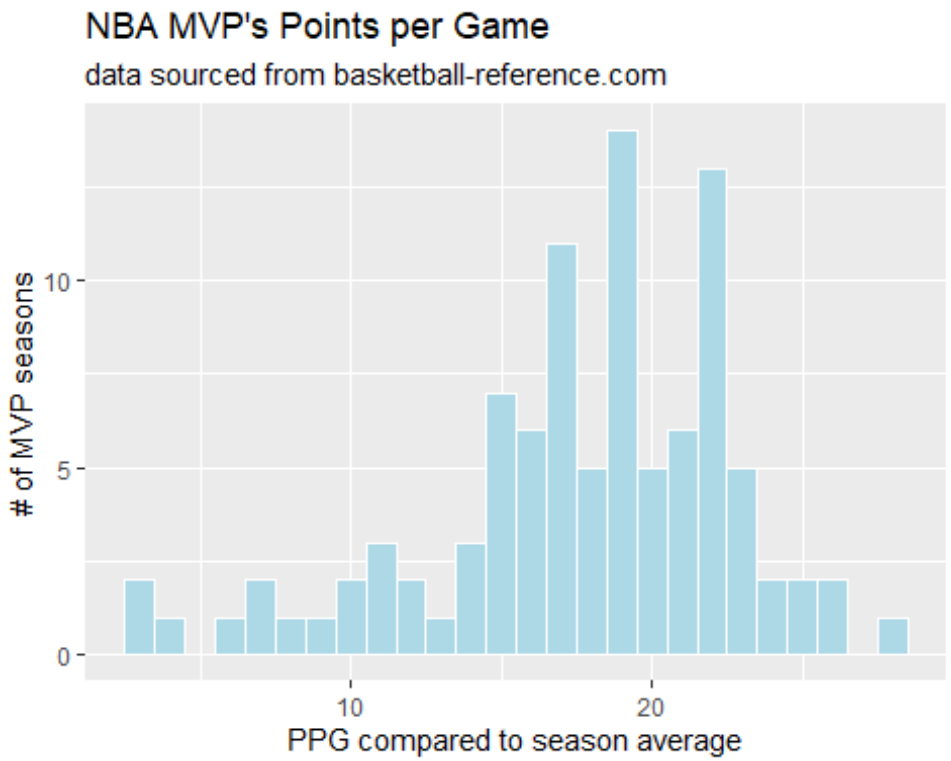
I used 4 data frames from the NBA Stats (1947-present) dataset on Kaggle. The author on Kaggle used web-scraping techniques to source all the data from basketball-reference.com, the best free resource available for accessing NBA statistics. The first of the 4 data frames I

used are Player.Per.Game, which gives per game numbers for basic counting stats (such as points, assists, and rebounds) for every NBA player's every season they played in the league. The next dataframe is Advanced, which gives advanced statistics for every NBA player's every season the played. The third dataframe is Player_Award_Shares, which gives every player who finished top 5 in voting for the 5 major individual awards, including variables such as how many first place votes the player received, the percentage of total possible voting points they received, and whether the player ended up winning the award. The final dataframe I will use is End_of_Season_Teams, which gives all players in each season who were selected to the All-NBA and All-Defense teams. The resource I used to supplement these data sets is basketball-reference.com itself, in any case where I was unsure what a variable meant and/or what it was measuring.

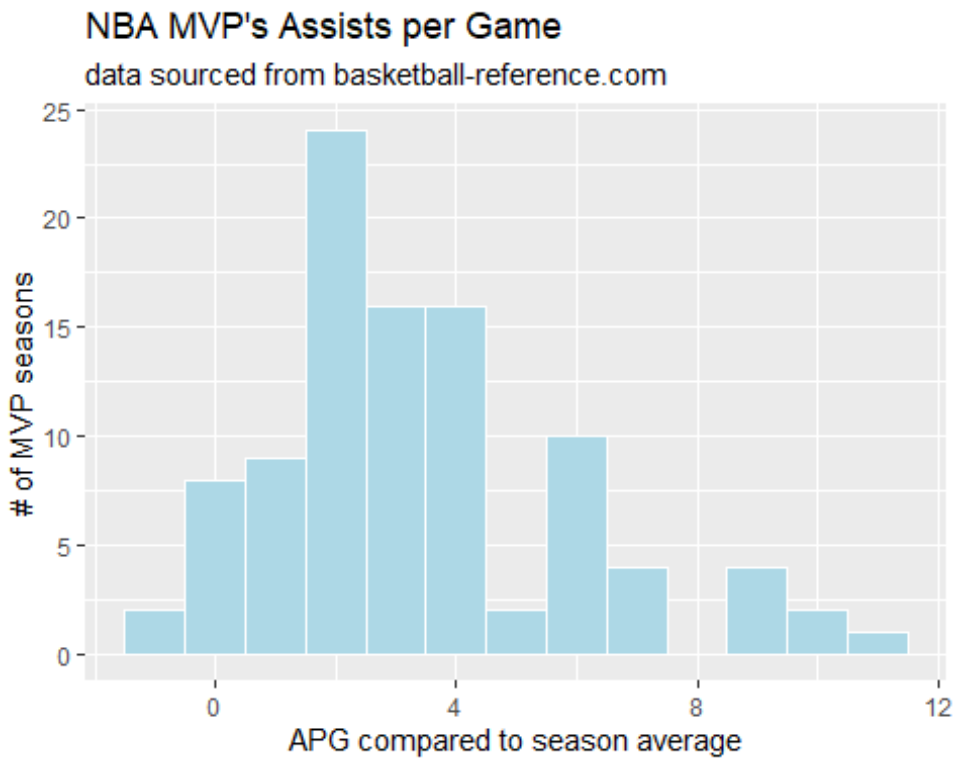
EDA: Use your dataset to make data visualizations that explain the variables of interest and how information through the graphics provides easy solution for your research questions. Explain your steps on how these visualizations help with your project.

The first research goal I will use EDA for is to get an idea of what sort of statistical performance a player needs to win MVP.

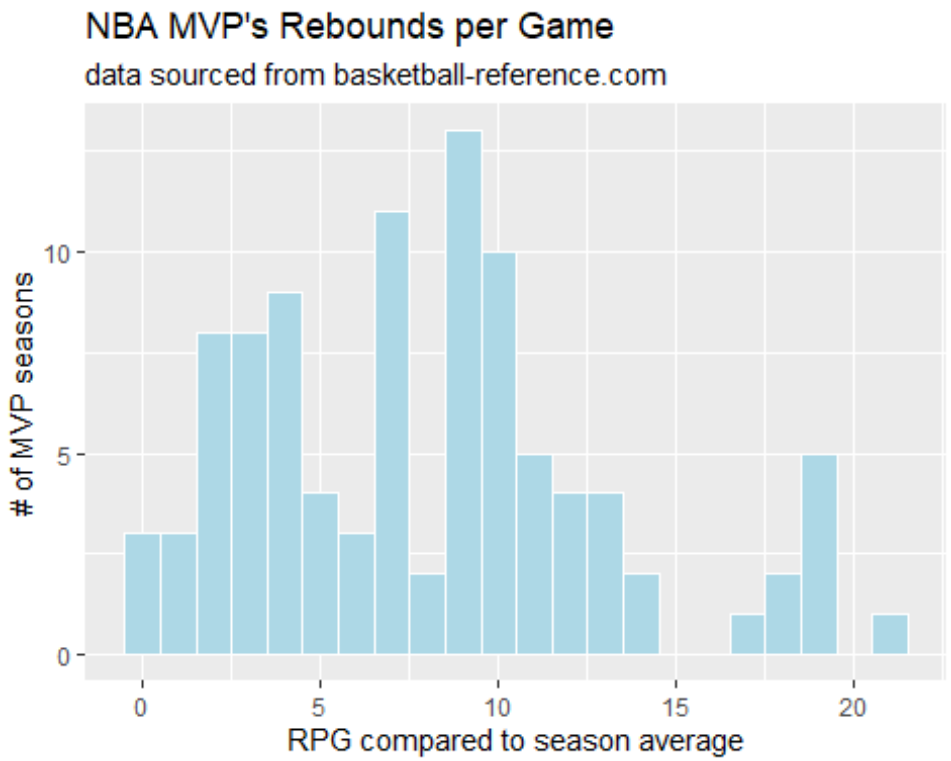
```
NBA_mvpwinners = NBAPlayers_FullStats %>% subset(winner==TRUE & award=="nba m
vp")

mvp_ppg = NBA_mvpwinners %>% ggplot(aes(x = pts_per_game.y)) + geom_histogram
(binwidth=1,

color="white", fill="lightblue")
mvp_ppg + labs(x = "PPG compared to season average", y="# of MVP seasons",
               title = "NBA MVP's Points per Game", subtitle="data sourced fr
om basketball-reference.com")
```



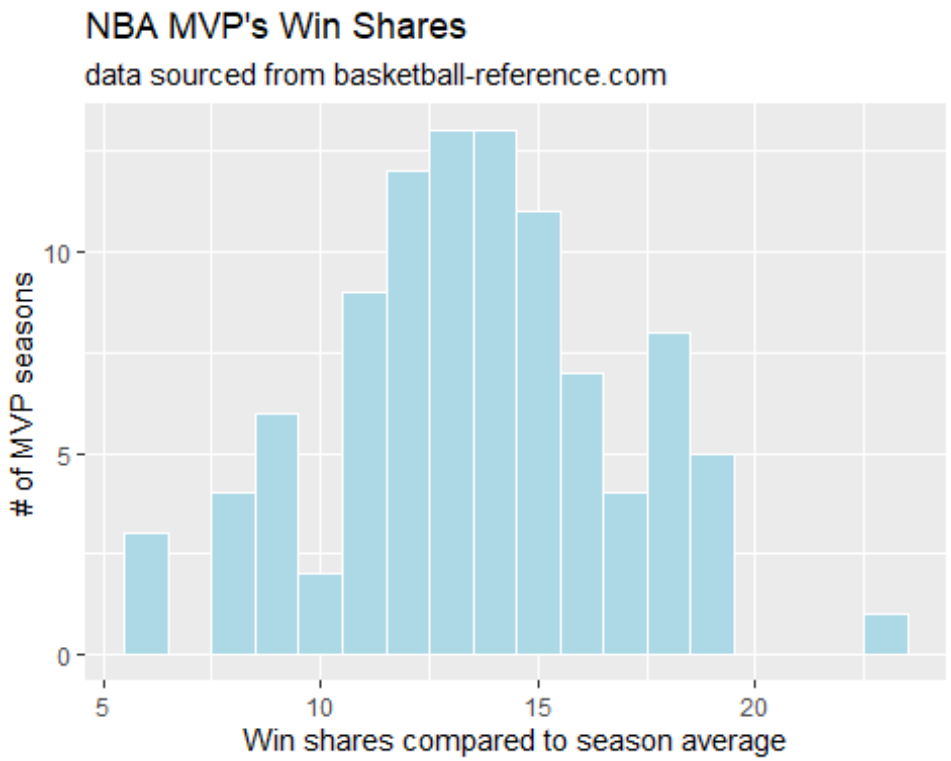

```
mvp_astpg =  NBA_mvpwinners %>% ggplot(aes(x = ast_per_game.y)) + geom_histog
ram(binwidth=1,

color="white", fill="lightblue")
mvp_astpg + labs(x = "APG compared to season average", y="# of MVP seasons",
                            title = "NBA MVP's Assists per Game", subtitle="da
ta sourced from basketball-reference.com")
```



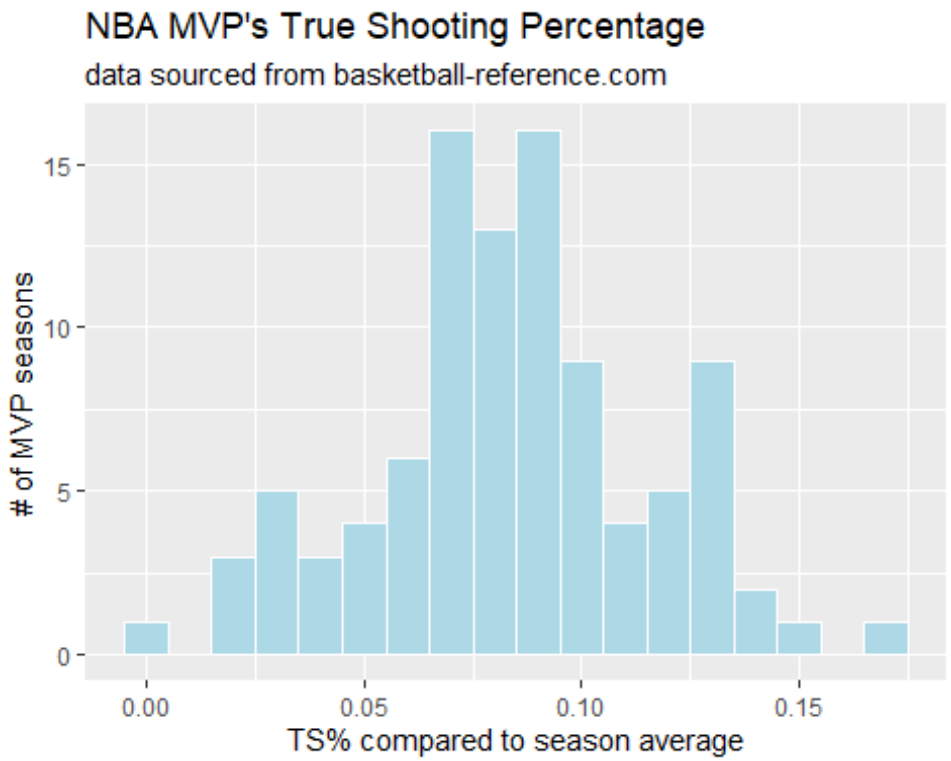

```
mvp_rebpg = NBA_mvpwinners %>% ggplot(aes(x = trb_per_game.y)) + geom_histogr
am(binwidth=1,

color="white",fill="lightblue")
mvp_rebpg + labs(x = "RPG compared to season average", y="# of MVP seasons",
                 title = "NBA MVP's Rebounds per Game", subtitle="data source
d from basketball-reference.com")
```



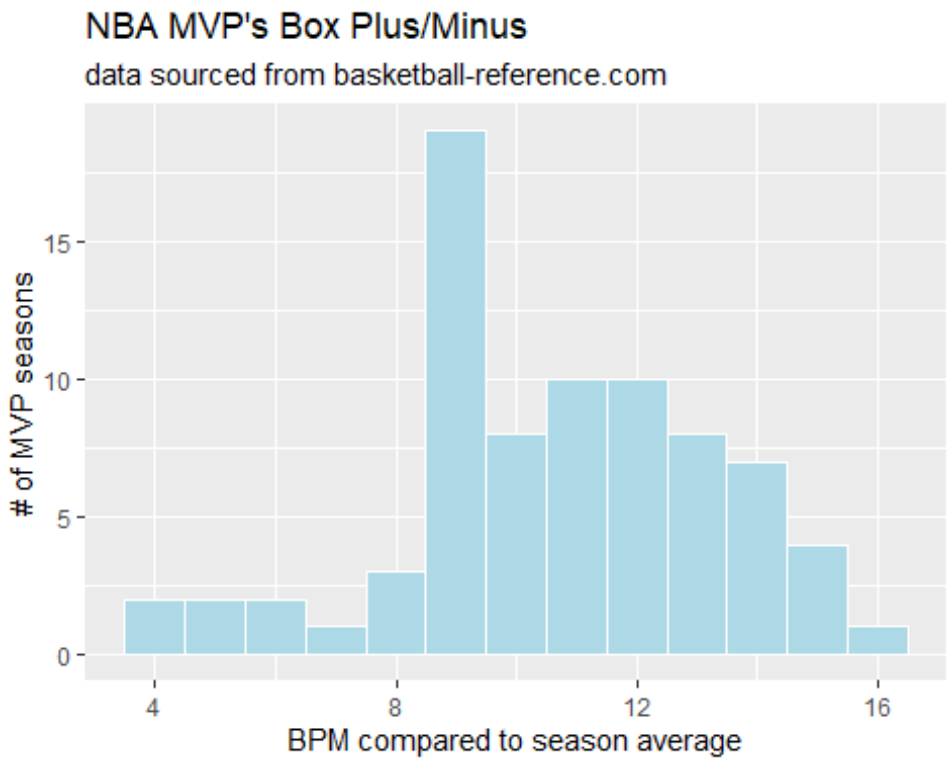

```
mvp_ws = NBA_mvpwinners %>% ggplot(aes(x = ws.y)) + geom_histogram(binwidth=1
,

color="white",fill="lightblue")
mvp_ws + labs(x = "Win shares compared to season average", y="# of MVP season
s",
                  title = "NBA MVP's Win Shares", subtitle="data sourced from
basketball-reference.com")
```




```
mvp_ts = NBA_mvpwinners %>% ggplot(aes(x = ts_percent.y)) + geom_histogram(bi
nwidth=0.01,
                                                                    color="whi
te",fill="lightblue")
mvp_ts + labs(x = "TS% compared to season average", y="# of MVP seasons",
              title = "NBA MVP's True Shooting Percentage", subtitle="data sour
ced from basketball-reference.com")
```




```
mvp_bpm = NBA_mvpwinners %>% ggplot(aes(x = bpm.y)) + geom_histogram(binwidth
=1,
                                                                     color="w
hite",fill="lightblue")
mvp_bpm + labs(x = "BPM compared to season average", y="# of MVP seasons",
               title = "NBA MVP's Box Plus/Minus", subtitle="data sourced from
basketball-reference.com")

## Warning: Removed 21 rows containing non-finite values (`stat_bin()`).
```



Initial findings:

When looking at MVP winners, I decided to look at three counting stats and three advanced stats. The counting stats I used were points per game (when a player scores), assists per game (when a player gives the ball to a teammate who then scores), and rebounds per game (when a player collects the ball after a missed shot). I chose these not only because there are available for every MVP winner in history, but also because they are the most commonly cited counting stats when discussing a player's performance. MVP winners have a much lesser assists per game advantage compared to league average than they do in rebounds per game and points per game. This makes sense, as there are much fewer assists in a single NBA game on average compared to points and rebounds. In comparison to rebounds per game, MVP winners have a much greater points per game average. This makes sense, but the difference is not as great as I would expect, considering there are many more points per NBA game than rebounds. A reason I think this might be because a plurality of MVP winners have played the center position. Centers are generally the tallest player on the court and play closer to the basket, which naturally lends itself to grabbing more rebounds.