Project 1 Part 1 - NBA Salary Statistics

Nithin Vijayakumar

Dataset Information

Background Information

The dataset used for analysis in this report is about players in the National Basketball Association (NBA). In particular it focuses on the 2017-2018 regular season. This means data from the summer league, preseason, all-star break, and playoffs is not included.

The dataset used in this project was merged from three separate datasets. The first came from Kaggle and it contained annual salary information for every player in the NBA for the 2017-2018 season. It included trades/signs/waives, and the amount that the new team paid the player. Originally, this data was published by the NBA. The second dataset contained every player who has played for the NBA from 1951 to 2018. This dataset had height, weight, college, and starting/ending year for each player and came from Kaggle as well. The third dataset contained game statistics for the 2017-2018 season and was downloaded as a `.csv` from `basketball-reference.com` who originally gathered the data from the NBA. In addition to the player, it included fields such as position, points per game, and field goal percentage. These datasets were all merged on the character vector containing player names, and cleaned further by removing duplicates and removing team data for each player to simplify how trades were dealt with in each dataset.

After merging, each observation in the final, merged dataset represents a player in the NBA that played during the 2017-2018 season. Because the data is exhaustive, meaning every player is accounted for, it represents a population of NBA players. All of data originally came from the NBA, with game statistics being recorded during each game, and attributes, such as height/position/weight are self-reported by the teams each player plays on. A description for each of the columns is included in the section below, *Data Dictionary*.

Data Dictionary

Variable in Data	Description	Data Type
`Player`	Player name	Categorical
`Pos`	Position (PG, SG, SF, PF, C)	Categorical

Variable in Data	Description	Data Type
Age	Age of player	Numeric
G	Games played	Numeric
GS	Games started	Numeric
MP	Minutes played per game	Numeric/Percentage
FG	Field goal percentage	Numeric
FGA	Field goals attempted per game	Numeric
FG.	Field goals made per game	Numeric
X3P	Three point percentage	Numeric/Percentage
X3PA	Three pointers attempted per game	Numeric
X3P.	Three pointers made per game	Numeric
X2P	Two point percentage	Numeric/Percentage
X2PA	Two pointers attempted per game	Numeric
X2P.	Two pointers made per game	Numeric
eFG.	Effective field goal percentage	Numeric/Percentage
FT	Free throw percentage	Numeric/Percentage
FTA	Free throws attempted per game	Numeric
FT.	Free throws made per game	Numeric
ORB	Offensive rebounds per game	Numeric
DRB	Defensive rebounds per game	Numeric
TRB	Total rebounds per game	Numeric
AST	Assists per game	Numeric
STL	Steals per game	Numeric
BLK	Blocks per game	Numeric
TOV	Turnovers per game	Numeric
PF	Personal fouls per game	Numeric
PTS	Points per game	Numeric
year_start	First year in the NBA	Numeric
height	Height of player in ft/in	Categorical
weight	Weight of player in lbs.	Numeric
birth_date	Birth date of player in MM-DD-YYYY	Categorical
college	Name of college attend, blank if none	Categorical
Salaries	Salary in the 2017-2018 season in dollars	Numeric

Potential Issues

As mentioned before, each of the datasets had a unique way of dealing with trades. For the dataset from 1951-2018, since team was not included as an attribute, trades were not recorded at all. The dataset with game statistics recorded trades thoroughly with players having separate rows for each different team. Another row was given for these players detailing their overall game statistics for the year, regardless of team. For salaries, not every team a player played for during the season was listed. Only the teams that paid the player had entries. This means if a player was waived, i.e. a player is dropped from a team and is free to be signed by another team, but is still being paid by the original team, the new team is not listed. This causes complications when merging as salaries statistics would be duplicated for each team traded.

Another potential issue comes with the attribute `POS` in the data representing position. The values of this factor were PG, SG, SF, PF, C, PG-SG, SF-SG. Because some of the factors are combinations of other positions, when doing analysis with this attribute, assumptions will have to be made for converting the combined positions into a single position.

Analysis

Numerical Analysis

Correlation

```
library(knitr)
final_df <- read.csv('final_merged_nba.csv', stringsAsFactors = FALSE)
column_class <- sapply(final_df, class)
cols_numeric <- which(column_class == "numeric" | column_class == "integer")
cols_with_null <- which(apply(is.na(final_df), 2, sum) != 0)
numeric_df <-select(final_df, setdiff(cols_numeric, cols_with_null))
correlations <- cor(numeric_df)[, "Salaries"]
kable(cbind(Salaries = correlations[abs(correlations) > .5]),
      "markdown", caption = "Title of the table", booktabs = T)
```

	Salaries
GS	0.5615022
MP	0.5529134
FG	0.5912200
FGA	0.5650776

	Salaries
X2P	0.5472367
X2PA	0.5335000
FT	0.5659135
FTA	0.5687985
DRB	0.5071293
TOV	0.5133355
PTS	0.5990800
Salaries	1.0000000

This numerical summary represents the correlation coefficient found between each numerical attribute and salary. Only the values where the correlation was above 0.5 were displayed. These values can be seen as potentially important in a modeling application.

Position and Salary

```
to_graph <- final_df %>%
  mutate(Pos = substr(final_df$Pos, 1, 2)) %>% group_by(Pos) %>%
  summarise(min_sal = min(Salaries),  f_qrtr_sal = quantile(Salaries, .25),
            med_sal = median(Salaries), t_qrtr_sal = quantile(Salaries, .75),
            max_sal = max(Salaries))
kable(to_graph, "markdown", format.args = list(scientific = F, big.mark = ","),
      bookends = T, col.names = c("Position", "Minimum Salary", "25th %ile Salary",
                          "Median Salary", "75th %ile Salary", "Maximum Salary"))
```

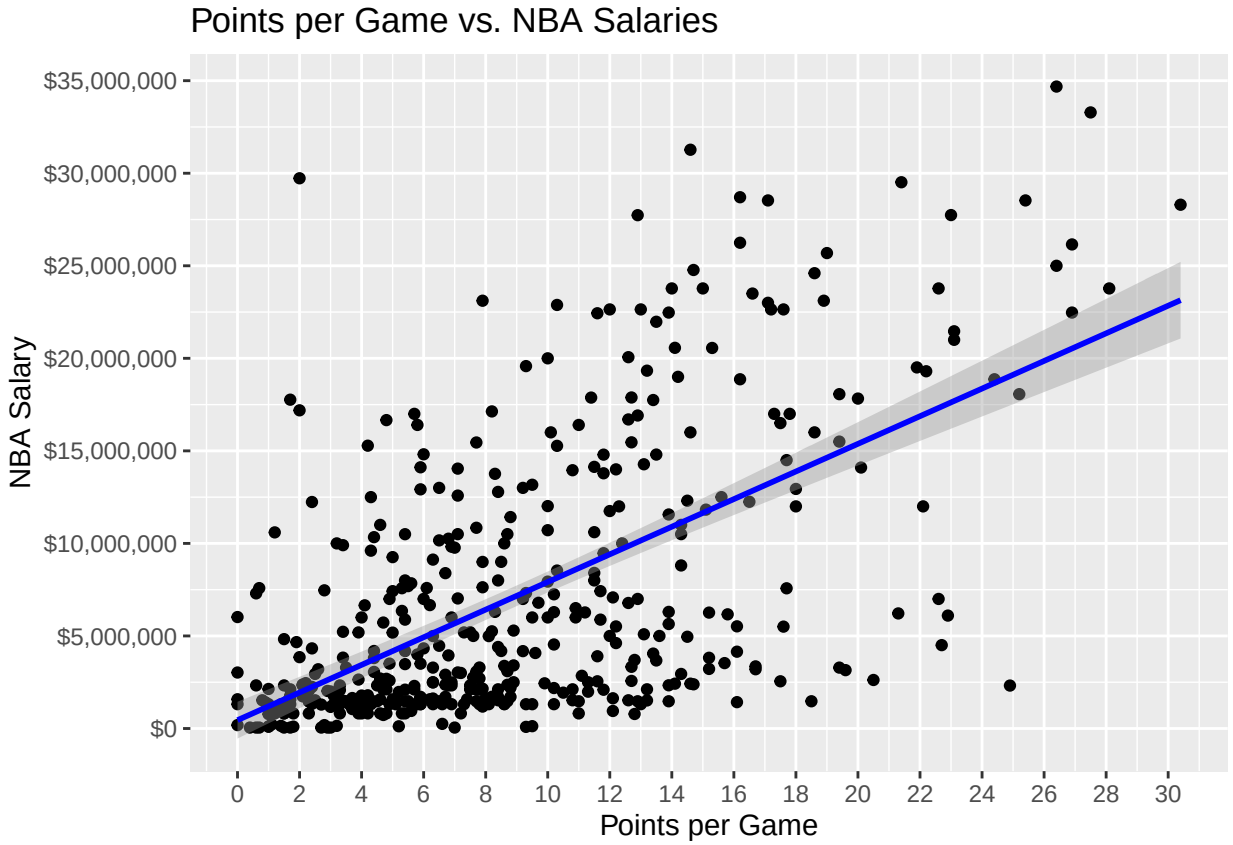
Position	Minimum Salary	25th %ile Salary	Median Salary	75th %ile Salary	Maximum Salary
C	100,000	1,709,538	4,995,120	13,000,000	27,734,405
PF	50,000	1,522,309	3,885,820	9,343,750	33,285,709
PG	50,000	1,471,382	2,645,888	7,645,577	34,682,550
SF	50,000	1,635,660	3,223,962	10,438,726	29,727,900
SG	50,000	1,509,496	3,817,899	9,955,357	28,299,399

This is a five-number-summary for each of the positions. As stated before, an assumption was made that the first two characters in a string was the position used, e.g. `"PG-SG"` was converted to `"PG"`

Graphical Analysis

Points per Game and Salary

```
pretty.print.sal <- function(x) {paste0("$", prettyNum(x,big.mark=",",scientific=F))}
ggplot(final_df, aes(x = PTS, y = Salaries)) + geom_point() +
geom_smooth(method = "lm", color = 'blue') + scale_x_continuous(breaks = seq(0,30,2)) +
scale_y_continuous(breaks = seq(0, 35000000, 5000000), labels = pretty.print.sal) +
labs(title="Points per Game vs. NBA Salaries", x = "Points per Game", y = "NBA Salary")
```



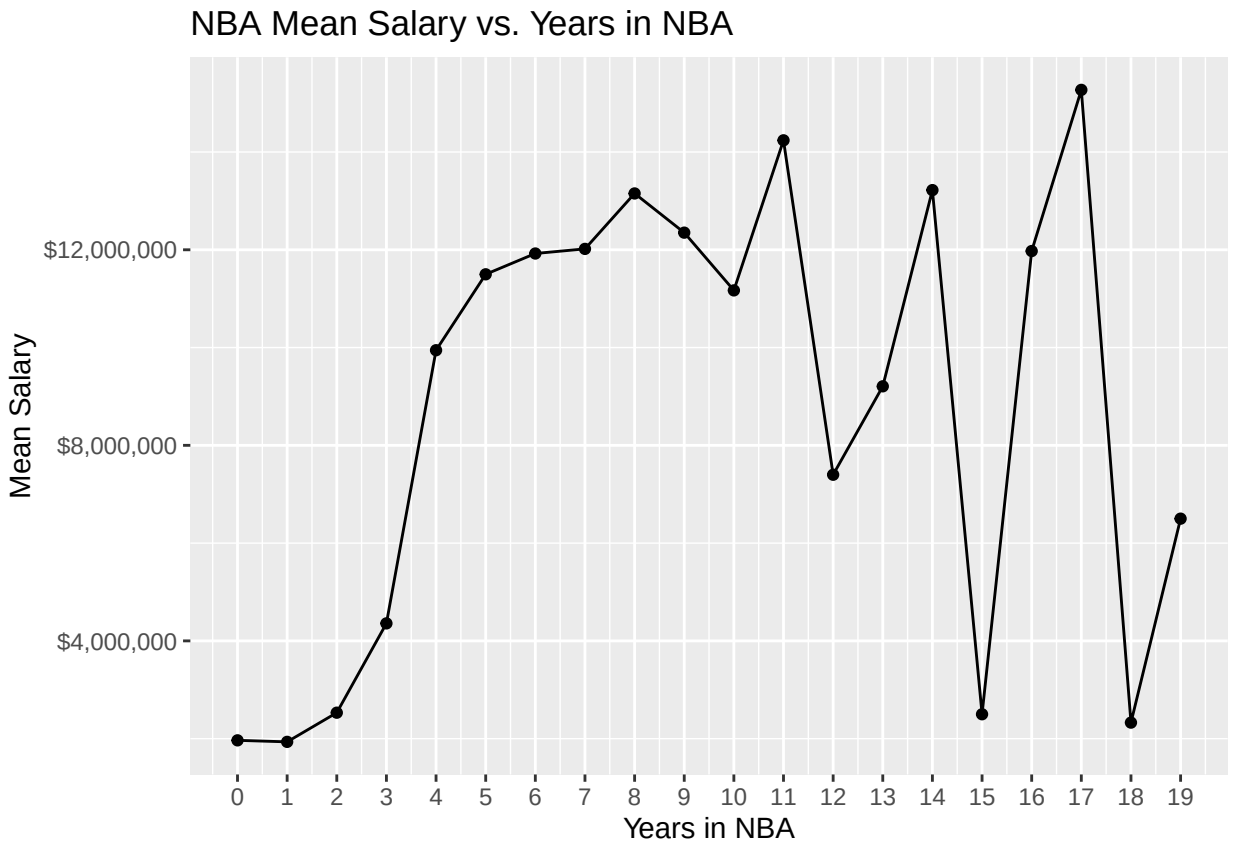
Points per game was chosen as a metric for the scatterplot because it was the attribute with the highest correlation coefficient at ~0.599.

Years Pro and Salary

```
to_graph <- final_df %>% mutate(years_pro=2018-year_start) %>%
  group_by(years_pro) %>% summarise(mean_sal = mean(Salaries))

ggplot(to_graph, aes(x=years_pro, y=mean_sal)) + geom_line() + geom_point() +
  scale_x_continuous(breaks = seq(min(to_graph), max(to_graph))) +
  scale_y_continuous(labels = pretty.print.sal) +
```

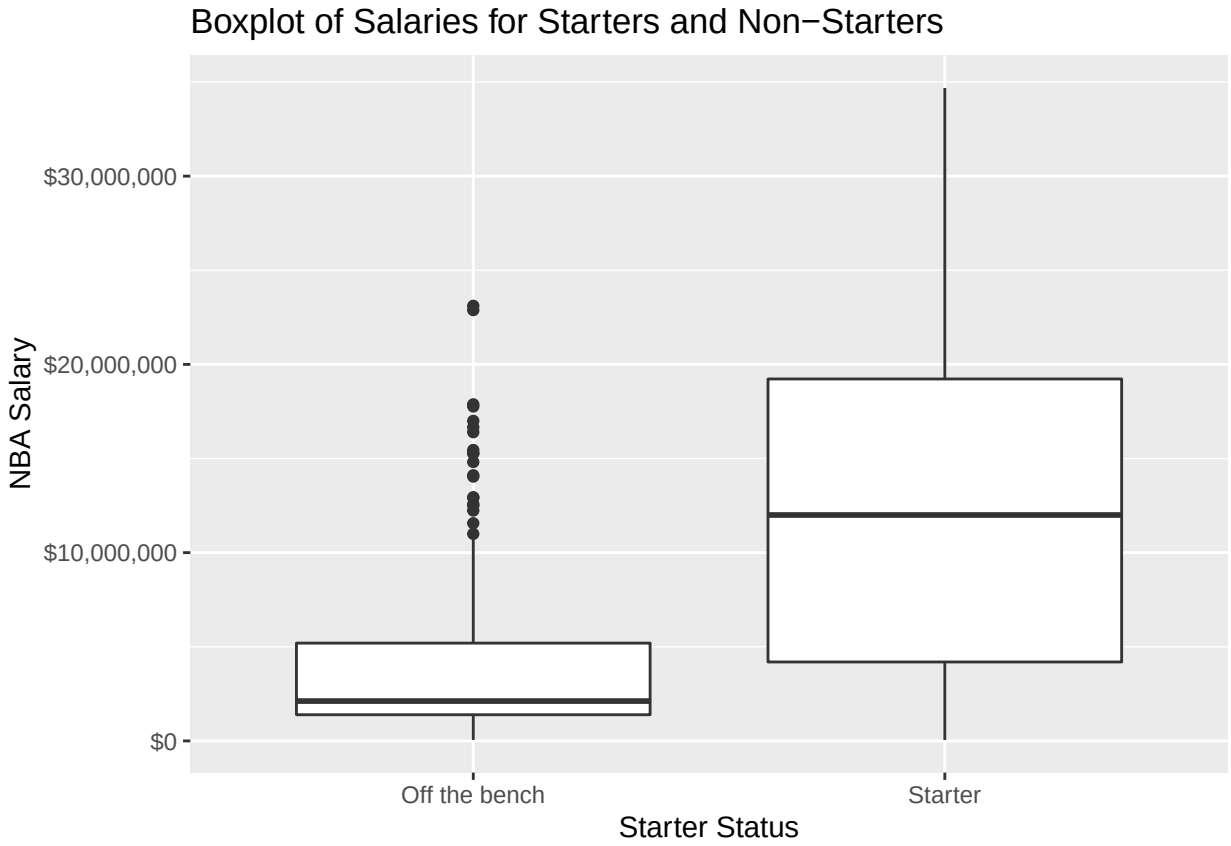
```
labs(title="NBA Mean Salary vs. Years in NBA",x="Years in NBA",y="Mean Salary")
```



This plot shows the average salary for how long a player is in the NBA. Generally, after the 3rd year, mean salaries increase and remain fairly consistent. It is important to note that salaries are generally multi-year contracts, so older players retire whenever the contract ends, e.g. a player can have the same salary for years 8-11 if they sign a 4 year contract.

Starter Status and Salary

```
final_eng_df <- final_df %>% mutate(starter = GS/G >= .5)
final_eng_df %>% ggplot(aes(x = starter, y = Salaries)) + geom_boxplot() +
  scale_y_continuous(labels = pretty.print.sal) +
  scale_x_discrete(labels = c("Off the bench", "Starter")) +
  labs(title="Boxplot of Salaries for Starters and Non-Starters",
       x = "Starter Status", y = "NBA Salary")
```



This plot shows the distribution of starter status on salary. Starter status was determined by if at least 50% of games were started by the player. There seems to be a drastic difference in salaries between starter and coming off the bench.

Conclusion

With these representations of salary, the correlation and trends can be seen between various attributes and NBA salary. The correlation coefficients seen in the report demonstrated the linear relationship that each variable had with salary, but it can be seen that correlation coefficient is not a conclusive indicator. Derived interactions, such as the boxplots seen above, indicate vast differences when the data is looked at deeper. In addition, the number of years in the NBA seems to have some relationship, just perhaps not linear in nature. With these visualizations, it is safe to assume that much of the variation in salary can be attributed to the variables in the dataset. Future work visualizing quadratic and interaction terms using more advanced statistical methods needs to be done to be able to properly make any conclusions about relationships between these variables and salary. Only then, can effective attempts at modeling occur.

References

1. Dataset - Kaggle https://www.kaggle.com/drgilermo/nba-players-stats
2. Dataset - Basketball Reference https://www.basketball-reference.com/leagues/NBA_2018_per_game.html
3. `setdiff()` https://stat.ethz.ch/R-manual/R-devel/library/base/html/sets.html
4. `sapply(., class)` https://stackoverflow.com/questions/21125222/determine-the-data-types-of-a-dat
5. dplyr Cheat Sheet https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
6. kable https://haozhu233.github.io/kableExtra/awesome_table_in_pdf.pdf
7. prettyNum https://stat.ethz.ch/R-manual/R-devel/library/base/html/formatc.html
8. For Data Cleaning
    1. join https://dplyr.tidyverse.org/reference/join.html
    2. distinct https://dplyr.tidyverse.org/reference/distinct.html
    3. gsub https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/grep
    4. iconv http://ugrad.stat.ubc.ca/R/library/base/html/iconv.html