Capstone Project 1: NBA_Salary_Prediction

Statistical Data Analysis Report

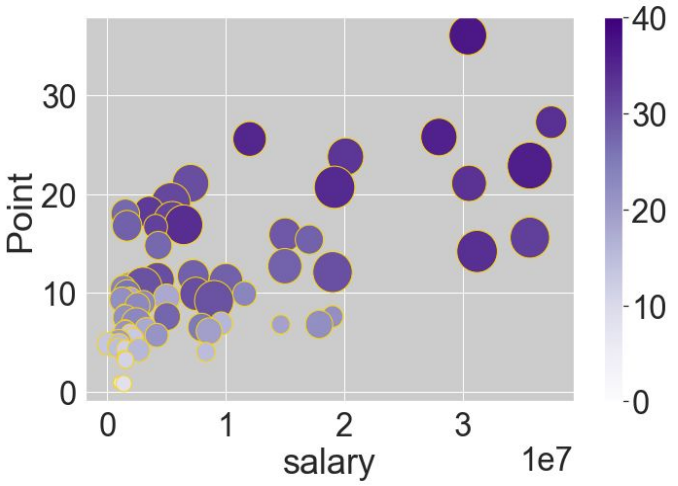
To analyze my variables, I, first, looked at the scatter plot shown below, to see if positions of players affect their points, salary and minutes they play. I mainly focused on two main positions.

- **C = Centers** who are usually the tallest players in the team and defending rim from short range shots and collecting the rebounds.

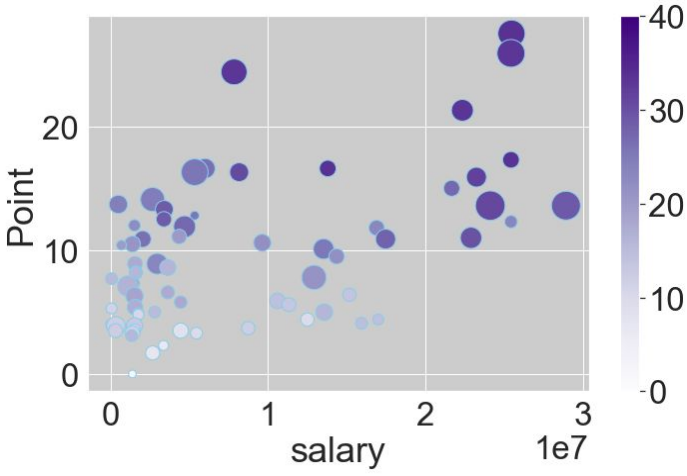
- **PG = Point-guard** who are the brains of the team sets the game, holds the ball mostly and directs other players.

Some information about plots that will be shown below:

Color of circle = Minutes played, the darker color the more minutes per game
Area of circle = Assists made
X-axis = salary earned in 10 Million $
Y-axis = Points made per game



Point-guard position plot



Center position plot

If you observe two plots, we can easily see the abundance of **larger and darker color** circles in the first plot(**Point-guard position)**, which shows that **PG** players are able to assist more(directing the game as said before) and stay in the game longer than center players. And also, they are earning more than Center players, because we cannot see any C player earning more than 30 Million $. PG players are also scoring leaders as seen in the plots. Y-axis shows that number of PG players scoring over 20 points per game is much more than number of C players scoring over 20 points per game.

T-tests were performed to prove these observations.

H0 = Null Hypothesis =>Center players are staying in the game as long as point guard players.

H01 = Null Hypothesis =>Center players are scoring per game as much as point guard players.

HA = Alternative Hypothesis => There is significant difference between minutes of Center players and minutes of PG players.

HA1 = Alternative Hypothesis => There is significant difference between points of Center players and points of PG players.

p-value is 0.002 and t-value is 3.2143507802050766 for minutes
Since p-value is 0.002 less than 0.05 for minutes, we reject the null hypothesis, and there is a statistically significant difference between minutes of PG and C players.

p-value is 0.038 and t-value is 2.102429387246607 for point
Since p-value is 0.038 less than 0.05, we reject the null hypothesis, and there is a statistically significant difference between points of PG and C players.

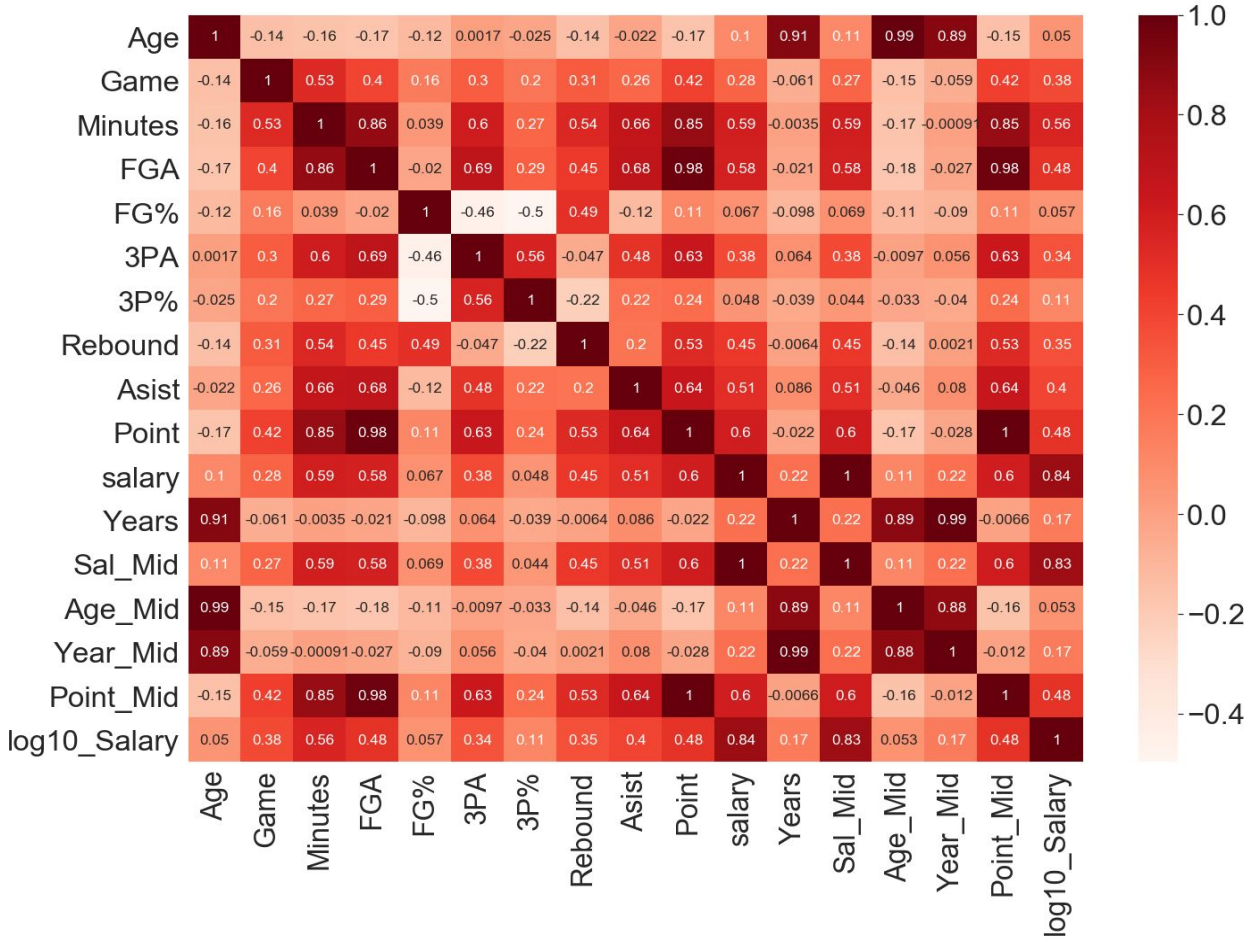
Another T-test was performed to see if they are actually earning different amounts;

H0 = Null Hypothesis => Salaries of center players are not significantly different than salaries of point guard players.

HA = Alternative Hypothesis => There is significant difference between salaries of different positions.

p-value is 0.344 and t-value is 0.9515878705510601
Since p-value is 0.344 more than 0.05, we failed to reject the null hypothesis, and salaries of center players ARE NOT significantly different than point guard players.

Seaborn heatmap is produced to see the correlations between dependent and independent variables.
Dependent variable = salary
Independent variables = all other stats



When we just take a look at the heat map and the correlations between salary(dependent variable) and other stats(independent variables), These independent variable are highly correlated with our dependent variable salary ->

'Point' :0.6, 'Asist':0.51, 'Rebound': 0.45, 'FGA'(Field Goal attempt): 0.58, 'Minutes':0.59, and '3PA'(3 point attempt):0.38.

The reason why age and years are not correlated is because we drop the rookie players who have 3 or less years experience, and mostly younger.
Pearson correlation coefficient between point and between salary is **0.599**
Pearson correlation coefficient between minutes and between salary is **0.595**
Pearson correlation coefficient between assist and between salary is **0.508**
Pearson correlation coefficient between field-goal and between salary is **0.581**
Pearson correlation coefficient between rebound and between salary is **0.450**
Pearson correlation coefficient between three-points and between salary is **0.378**

By looking at the heatmap we were able to see some correlations between independent variables;

When we just take a deeper look at the heat map and the correlations between independent variables, These independent variables are highly correlated with each other ->

'Point' : 'FGA' = **0.98**, The more shoot trials the more points.

'Point' : 'Asist' = **0.64**, The more assists to teammates the more shooting trials.

'Point' : 'Minutes' = **0.85**, The longer staying in the game the more chance to score.

There is also negative correlation between rebound and 3P%, we can easily say that Center players are not good at shooting 3 points.

'Rebound' : '3P%'(3 point percentage) = **-0.22.**