Milestone Report - NBA_Salary_Prediction

Abdurrahman Akkurek

1. Introduction

1.1 Background

The NBA has moved past Major League Baseball as the second-most popular sport in the United States and has millions of fans around the world. Moreover, the NBA is still growing and attracting more people as cited in Wikipedia link.

Moreover, across last year's season, the NBA generated about $8 billion in revenue. And the 30 teams making up the NBA have an average valuation of $1.9 billion each. And also with this growth and popularity, the average salary cap for each team has also skyrocketed now it is around 120 Million $ per year. How players on a team perform is the most important factor that determines which team wins the championship. Players' pays are largely based on their past performances. However, player performance changes from season to season. Each year there are a number of players who improve dramatically over last year. Those players bring a lot of value, both competitively and economically, to the teams they belong to.

That's why salary negotiation is vital for players along with their performances.

1.2 Problem

Season data that might contribute to determining player salary includes his performance last season, his age, his experience, his position, and other stats that describe what kind of player he is. This project aims to predict whether and how much salary a player will be worth for the next season based on these data.

Beside this,this prediction might help players during salary negotiation and provide them with insight to foresee what they are worth before signing any contract.

1.3 Interest

Obviously, NBA teams and players would be very interested in accurate prediction of player salary, it is useful for players as a competitive advantage and for teams as business values. Others who are interested in the NBA such as fans and fantasy basketball players may also be interested.

1.4 Questions

1- Which statistic or statistics are the best indicators(predictors) for the player's salaries?

2- How reliable is the machine learning model I will build to predict NBA salaries?

3- Measuring the error of the model and analyzing top 20 players with their predicted salaries and actual salaries to see if it can be predicted correctly.

2. Data acquisition and cleaning

2.1 Data sources

NBA player stats, position, age, and draft position data can be found on here. However, this data has only stats not years of experience and salary information. I scraped basketball-reference.com for players drafted beteen 1998-2019 and player salary for the 2018 season. Datasets can be also downloaded as csv files from the same website. Initial dataset has 578 players stats and 18 attributes..

2.2 Data cleaning

Data downloaded or scraped from the sources given above were loaded and combined into one Pandas dataframe. While reading the csv file, I encountered an encoding error, that is why files were loaded with (encoding = "ISO-8859-1) parameter.

There are several problems with the datasets. First, there were a lot of missing values for FG%, 2P% and 3P% because of not having any FGA and 3PA for that season.It is double checked and seen that those values are zeros. I decided to fill those values with 0 because there is no field goal attempt at all.

Second, there are players with duplicated names, which is because those players changed their team during mid season. That is why, I decided to drop duplicates and leave the one that player played the most minutes per game. Players were sorted in descending order by minutes and then dropped duplicates by keeping the first duplicate. Now our dataset has 402 rows, 34 columns.

Third, the salary column was categorical and had $ sign. $ sign was removed and the column was converted to numeric. Beside that position column had ‘C’, ‘PG’, ‘SG’, ‘SF’, ‘PF’ as 5 different positions. I wrote a script and assigned numbers for each position, which is also used in the NBA. ‘PG’ is number 1, ‘C’ is number 5 position.

After fixing these problems, I dropped these columns because they will not be used throughout analysis. ['Rk', 'username_y', 'username_x', 'eFG%', 'PF', 'ORB', 'DRB', 'TOV', 'GS', 'FG', '3P', '2P', 'FT', 'BLK', 'FTA', 'FT%', 'STL', '2PA', '2P%']. The reason why I dropped these because I already have another stat represents most of these columns. For example, TRB(total rebound) is equal to addition of ORB(offensive rebound) and DRB(defensive rebound). We don’t need these two columns separately.

Moreover, these column names were changed to more meaningful ones to be more clear about what they are. ('Salary 2018-19': 'salary', 'Yrs': 'Years', 'Pos': 'Position', 'Tm': 'Team', 'G':'Game', 'MP':'Minutes', 'TRB':'Rebound', 'AST':'Asist', 'PTS': 'Point'). And then, for aesthetics of the data and visualizations, I wrote scripts that classify Point, Assist, Salary, Age, Years, and Minutes, Games, and Three point attempts and most of them were reduced to 10-15 different groups instead of more. Years column was displaying years of experience as of 2019, that is why, I wrote another script that all years were reduced by 1 to have 2018 season experiences data.

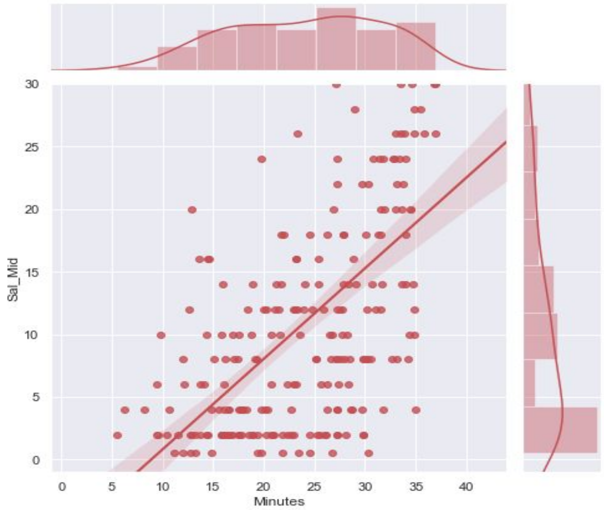
2.3 Feature selection

After data cleaning, there were 402 samples and 34 features in the data. After examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of the number of rebounds a player collected, and another feature of defensive and offensive rebounds he collected. These two features contained very similar information (a player's ability to rebound). Such total vs. rate relationships also existed between other features. These features are problematic for two reasons: (1) A player's certain abilities were duplicated in two features. (2) A player's playing time was duplicated in multiple features. In order to fix this, I decided to keep all features that were total in nature, and drop their cumulative counterparts (Table 1). There were also other redundancies, such as that total rebounds are the sum of offensive rebounds and defensive rebounds. For features that can be calculated by sum of other features, I decided to drop them (Table 1). 2 After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated (Pearson correlation coefficient > 0.9). For example, shots attempted, shots made, and points scored were highly correlated. This makes sense, after all, you score points by making shots. From these highly correlated features, only one was kept, others were dropped from the dataset.For example Years and Age is also highly correlated. Age was dropped. After all, 15 features were selected

Table 1. Simple feature selection during data cleaning.

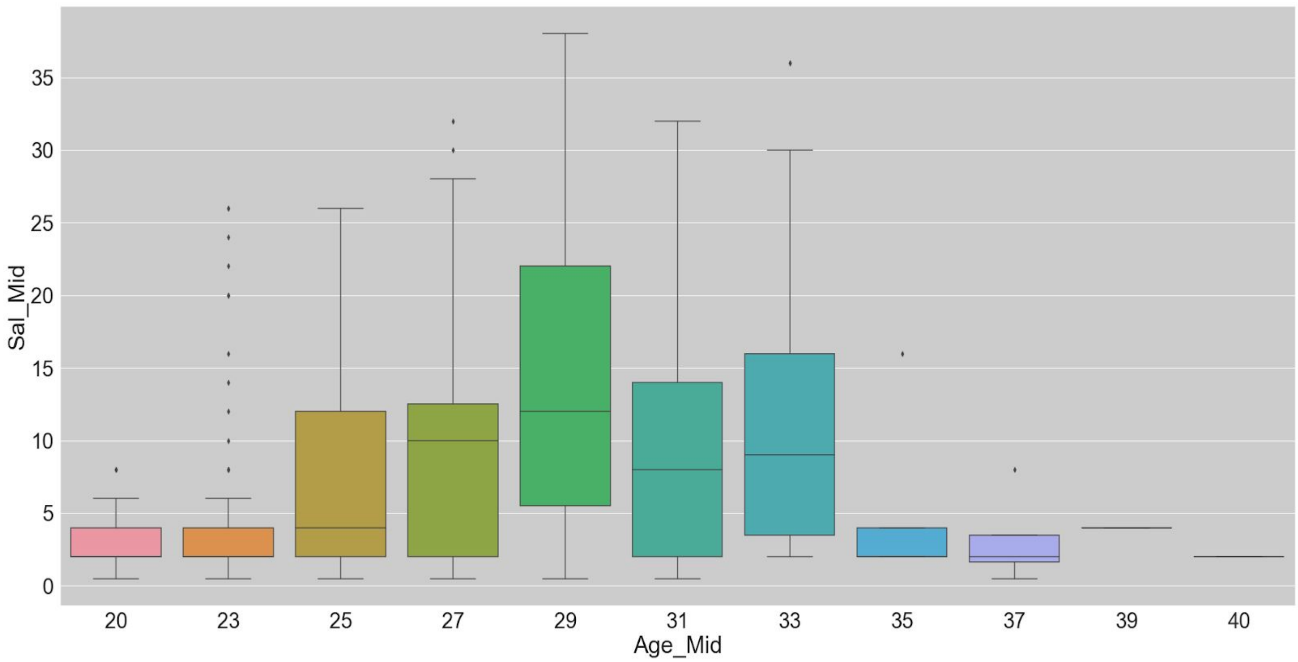
Kept features	Dropped features	Reason for dropping features
TRB	DRB, ORB	Total = offense + defense. Dropped defense.
FGA, FG, FG%, 3PA	2PA, 2P, 2P%, 3P, 3P%	Field goal = 2-point shots + 3-point shots. Dropped 2-point and 3-point shots. Just kept 3PA
Asist, Point, Games played, Minutes, Years	FTA, FT%, FT, PF(fouls), BLK(blocks), STL(steals), GS(games started)	FT% are very similar, no need to add in our prediction, personal fouls and steals are the same. We just kept the main attributes.

2. Exploratory Data Analysis (EDA)

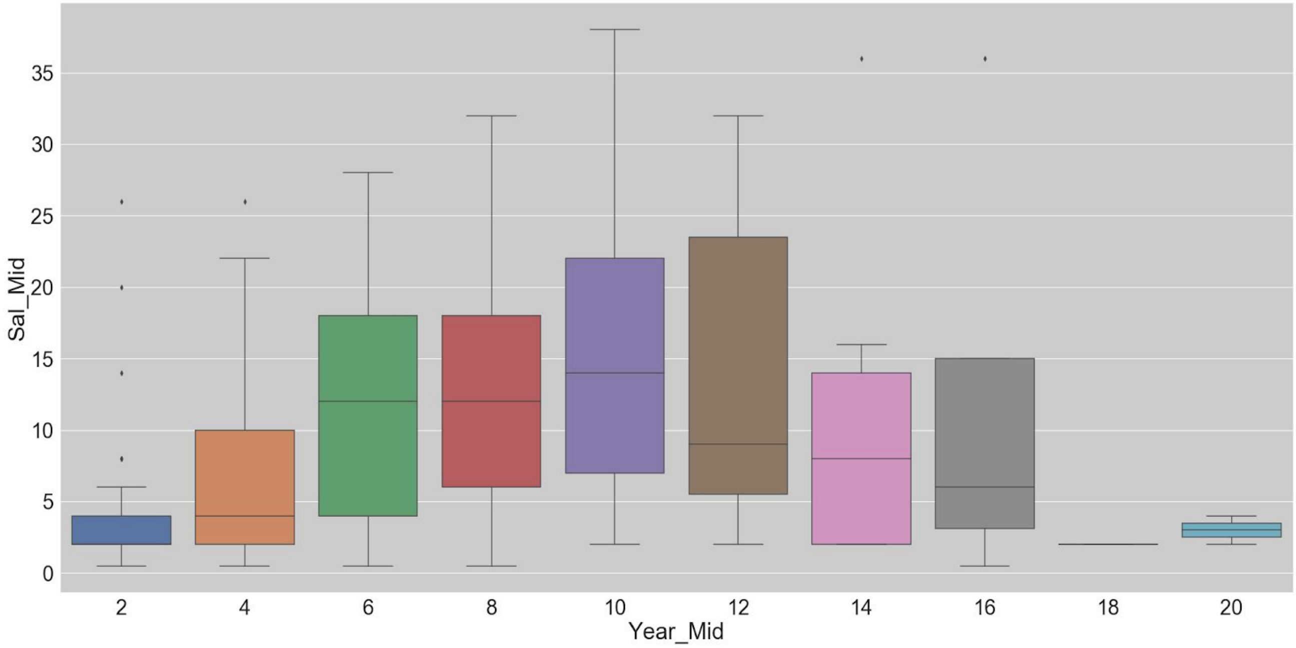


There were so many outliers in our data, many players who played less than 10 minutes and obviously they did score almost any point and they didn't earn a good salary. That's why players who played less than 10 minutes will be dropped.

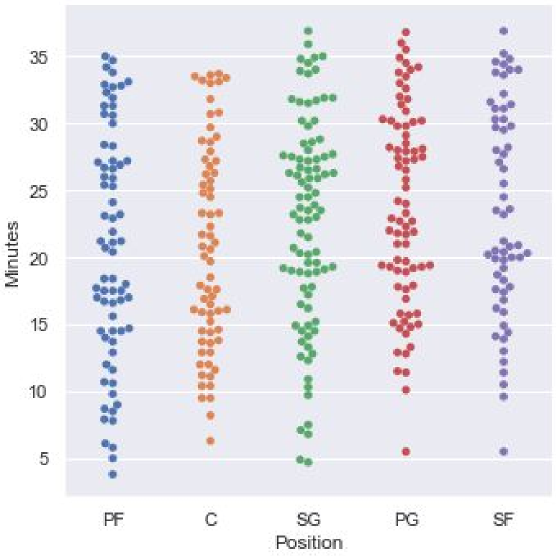
And also, players who did not play more than 20 games were also dropped. A season has 82 games and players should be eligible at least quarter of these games for the sake of analysis.

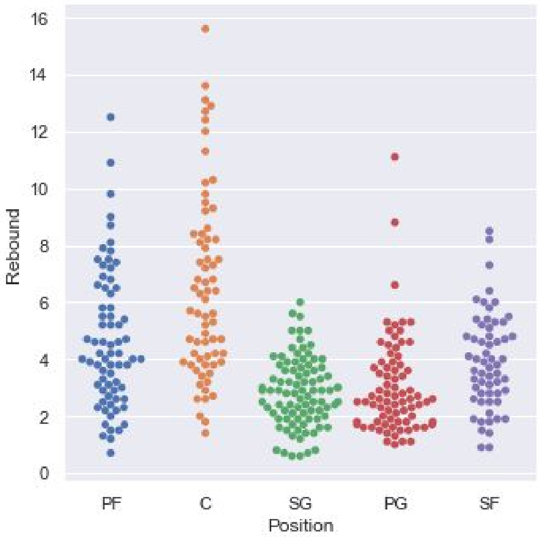


Based on this box plot, we see that ages between 27-33 are the most productive ages for players. And also this plot shows that retirement age is around 35 for NBA players. The reason why 20 years old players are not earning well is because rookie players have a 3-year rookie contract which is relatively small, and they cannot change their salary until it is over.

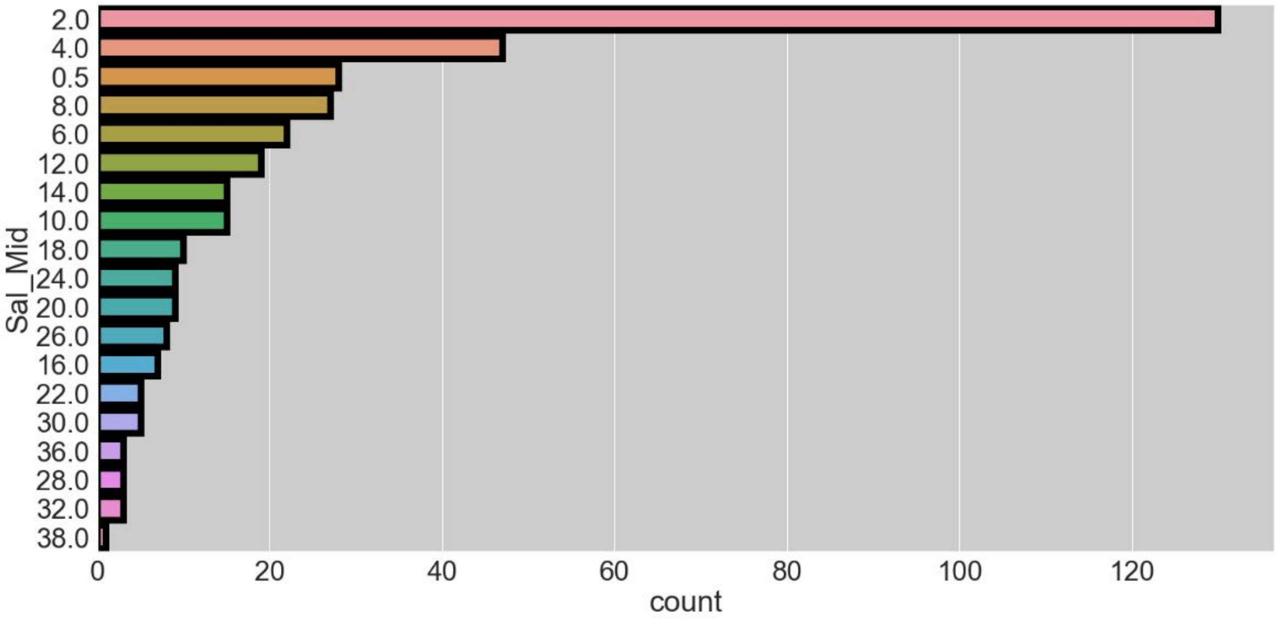


This plot also shows that experience between 8-12 years are most productive years.

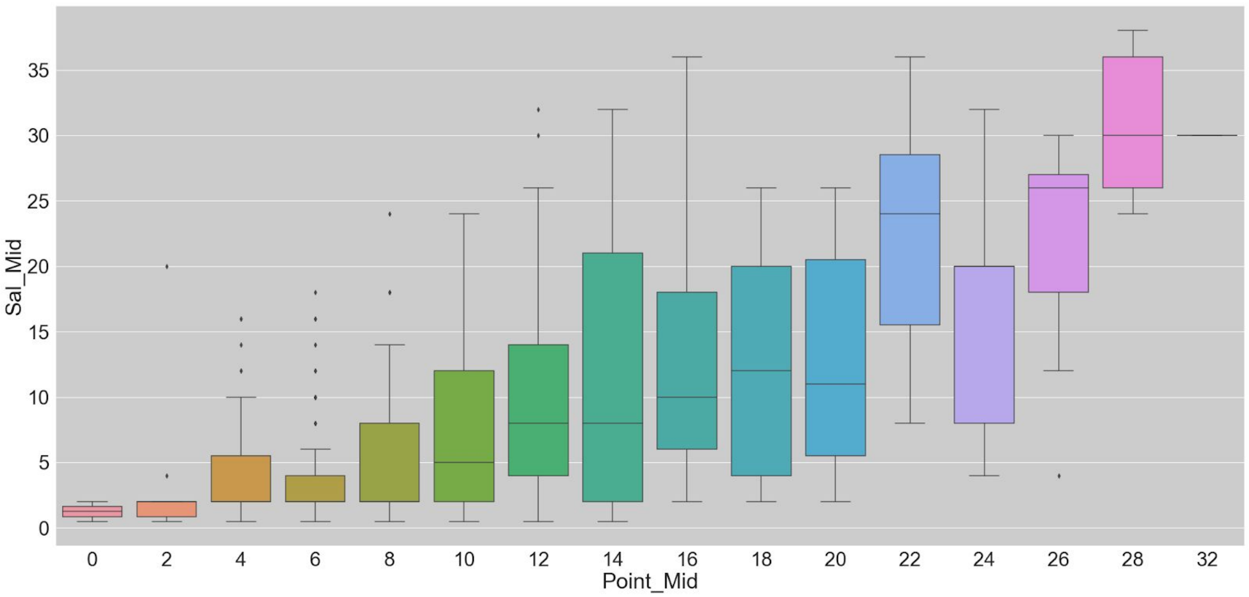




On the scatter plot above, what we notice is that the importance of position differs based on what statistics we are looking for. Center players are able to collect most rebounds along with Power-forward players, however, they are having slightly less minutes than other positions.



In this count plot, we see that most players are earning between 0-14 Million $. It is not easy to earn huge amount of money in NBA unless you show outstanding performances.



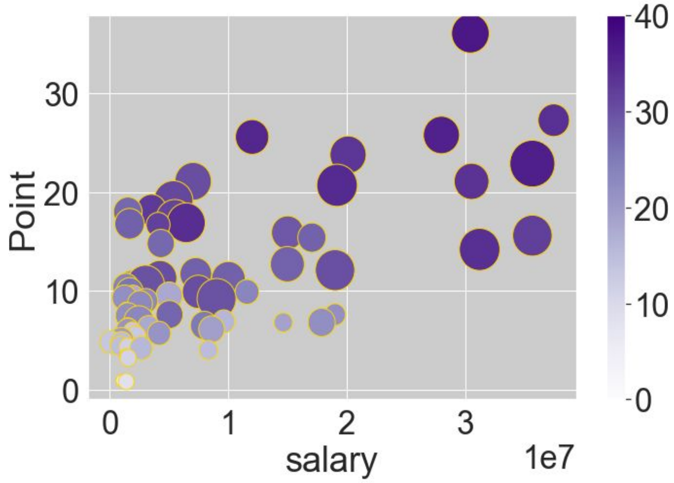
On this box plot, the correlation between points and salary is obviously noticed.If you can score around 28 points, you kind of guarantee over 25 Million $ up to 40 Million $.

3. Statistical Data Analysis

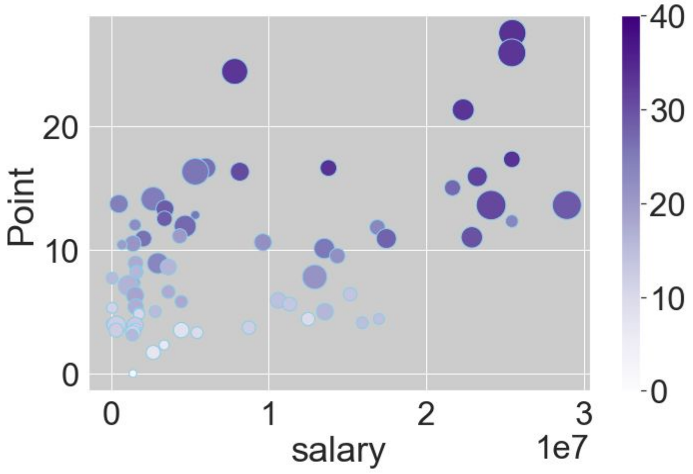
To analyze my variables, I, first, looked at the scatter plot shown below, to see if positions of players affect their points, salary and minutes they play. I mainly focused on two main positions.

- **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 the number of PG players scoring over 20 points per game is much more than the 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.

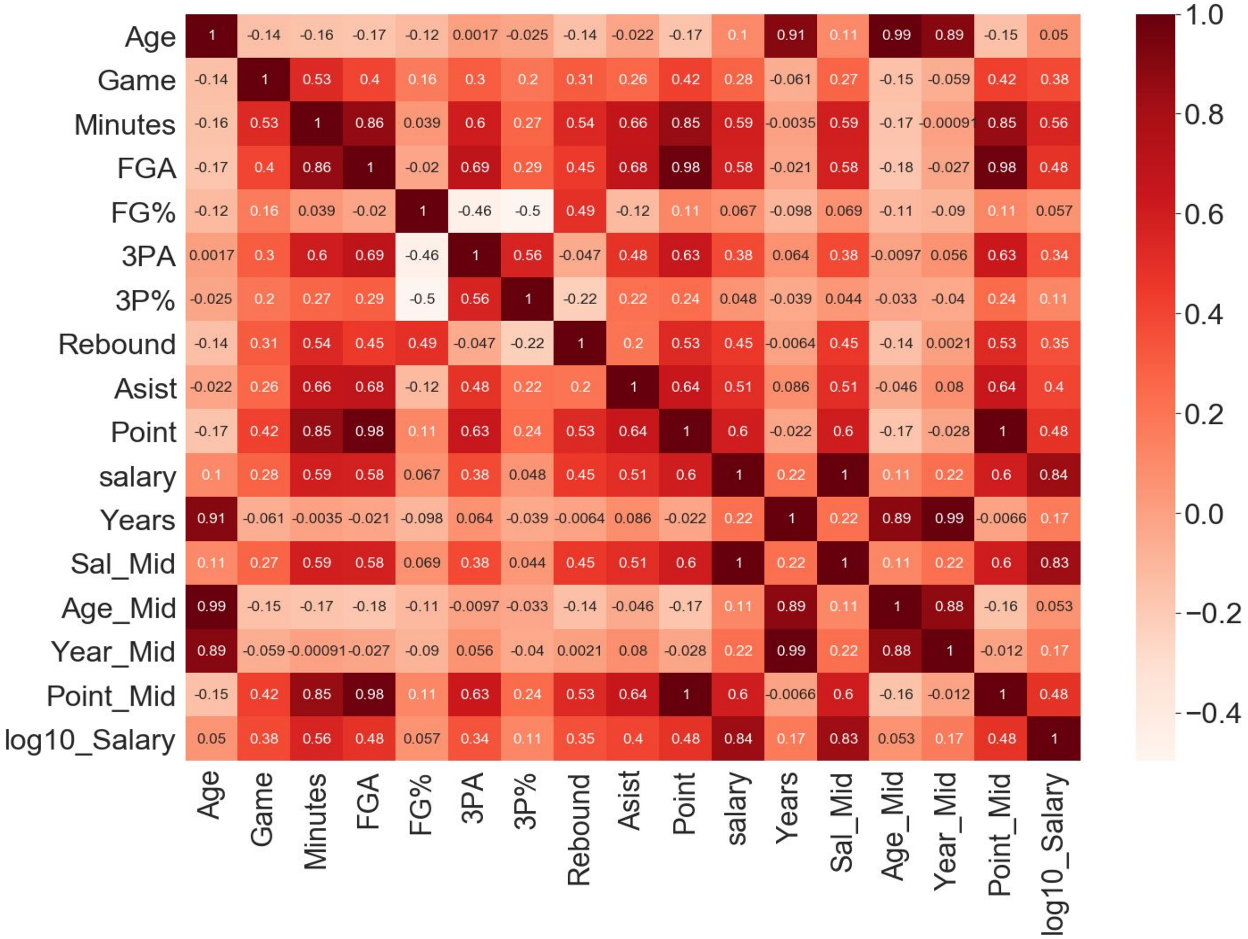
Another T-test was performed to see if they are actually earning different amounts;

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



	Age	Game	Minutes	FGA	FG%	3PA	3P%	Rebound	Asist	Point	salary	Years	Sal_Mid	Age_Mid	Year_Mid	Point_Mid	log10_Salary
Age	1	-0.14	-0.16	-0.17	-0.12	0.0017	-0.025	-0.14	-0.022	-0.17	0.1	0.91	0.11	0.99	0.89	-0.15	0.05
Game	-0.14	1	0.53	0.4	0.16	0.3	0.2	0.31	0.26	0.42	0.28	-0.061	0.27	-0.15	-0.059	0.42	0.38
Minutes	-0.16	0.53	1	0.86	0.039	0.6	0.27	0.54	0.66	0.85	0.59	-0.0035	0.59	-0.17	-0.00091	0.85	0.56
FGA	-0.17	0.4	0.86	1	-0.02	0.69	0.29	0.45	0.68	0.98	0.58	-0.021	0.58	-0.18	-0.027	0.98	0.48
FG%	-0.12	0.16	0.039	-0.02	1	-0.46	-0.5	0.49	-0.12	0.11	0.067	-0.098	0.069	-0.11	-0.09	0.11	0.057
3PA	0.0017	0.3	0.6	0.69	-0.46	1	0.56	-0.047	0.48	0.63	0.38	0.064	0.38	-0.0097	0.056	0.63	0.34
3P%	-0.025	0.2	0.27	0.29	-0.5	0.56	1	-0.22	0.22	0.24	0.048	-0.039	0.044	-0.033	-0.04	0.24	0.11
Rebound	-0.14	0.31	0.54	0.45	0.49	-0.047	-0.22	1	0.2	0.53	0.45	-0.0064	0.45	-0.14	0.0021	0.53	0.35
Asist	-0.022	0.26	0.66	0.68	-0.12	0.48	0.22	0.2	1	0.64	0.51	0.086	0.51	-0.046	0.08	0.64	0.4
Point	-0.17	0.42	0.85	0.98	0.11	0.63	0.24	0.53	0.64	1	0.6	-0.022	0.6	-0.17	-0.028	1	0.48
salary	0.1	0.28	0.59	0.58	0.067	0.38	0.048	0.45	0.51	0.6	1	0.22	1	0.11	0.22	0.6	0.84
Years	0.91	-0.061	-0.0035	-0.021	-0.098	0.064	-0.039	-0.0064	0.086	-0.022	0.22	1	0.22	0.89	0.99	-0.0066	0.17
Sal_Mid	0.11	0.27	0.59	0.58	0.069	0.38	0.044	0.45	0.51	0.6	1	0.22	1	0.11	0.22	0.6	0.83
Age_Mid	0.99	-0.15	-0.17	-0.18	-0.11	-0.0097	-0.033	-0.14	-0.046	-0.17	0.11	0.89	0.11	1	0.88	-0.16	0.053
Year_Mid	0.89	-0.059	-0.00091	-0.027	-0.09	0.056	-0.04	0.0021	0.08	-0.028	0.22	0.99	0.22	0.88	1	-0.012	0.17
Point_Mid	-0.15	0.42	0.85	0.98	0.11	0.63	0.24	0.53	0.64	1	0.6	-0.0066	0.6	-0.16	-0.012	1	0.48
log10_Salary	0.05	0.38	0.56	0.48	0.057	0.34	0.11	0.35	0.4	0.48	0.84	0.17	0.83	0.053	0.17	0.48	1

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.**