Purdue University

Exploratory Data Analysis for Business Insights using Python

MGMT586: Python Programming

Group: Haden Foster, Ashvin Raj

Date: 10/12/2023

MGMT586 Final Project

Exploratory Data Analysis for Business Insights using Python   
Haden Foster, Ashvin Raj

10/12/2023

Introduction:  
  
With the NBA market's rapid growth, there has been a surge of controversy related to players demanding higher salaries through contracts and extensions. One prominent example is the "Super Max" contract, granting players 35% of the team's salary cap, resulting in substantial financial commitments. Consequently, NBA owners and general managers have come under scrutiny for offering exorbitant salaries to players whose skills may not necessarily justify such compensation, placing immense pressure on teams to validate these high salary investments with commensurate performance (Kraft, 2011).

Our objective is to comprehensively explore the relationship between player salaries and a team's success. This analysis takes into account various factors, such as win percentages, individual player statistics, and overall team performance. Through this examination, we aim to uncover the fundamental drivers behind a team's triumph in the NBA.

The insights derived from this project hold significant value for NBA team managers and decision-makers. They provide a deeper understanding of how the allocation of financial resources within a team, particularly in terms of player salaries, can impact a team's win-loss performance. This analysis aims to address critical questions, such as whether teams that invest more in player salaries tend to achieve superior results on the basketball court.

Data Acquisition:

The first dataset we have is from HoopsHype (https://hoopshype.com/salaries/players/2022-2023/), a basketball news and rumors website. It provides up-to-date information, news, and coverage related to the National Basketball Association (NBA) and basketball in general. HoopsHype is a valuable resource for basketball enthusiasts, fans, and those interested in staying informed about the latest happenings in the world of professional basketball. This data set has information about each player's salary for the 2022-2023 season. HoopsHype itself does not collect or create the data; rather, it compiles and reports on information already available in the public domain. It is an independent sports news and information platform that generates revenue through advertising, sponsorships, and other media-related sources.

The second dataset, obtained from Kaggle, provides a wealth of player attributes, including information such as their age, position, points scored, team affiliation, fouls committed, and field goals made. This dataset complements and enriches the analysis conducted using the previous dataset on salary.

When linking these two datasets, it becomes evident that the Kaggle dataset contributes additional player-specific attributes to our analysis. These attributes offer more comprehensive insights into player performance and characteristics, which can be correlated with the salary statistics available in the first dataset. By combining the two datasets, we gain a more holistic understanding of the factors influencing player and team success in the NBA, allowing for a more robust and nuanced analysis.

(https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular?select=2022-2023+NBA+Player+Stats+-+Regular.csv)

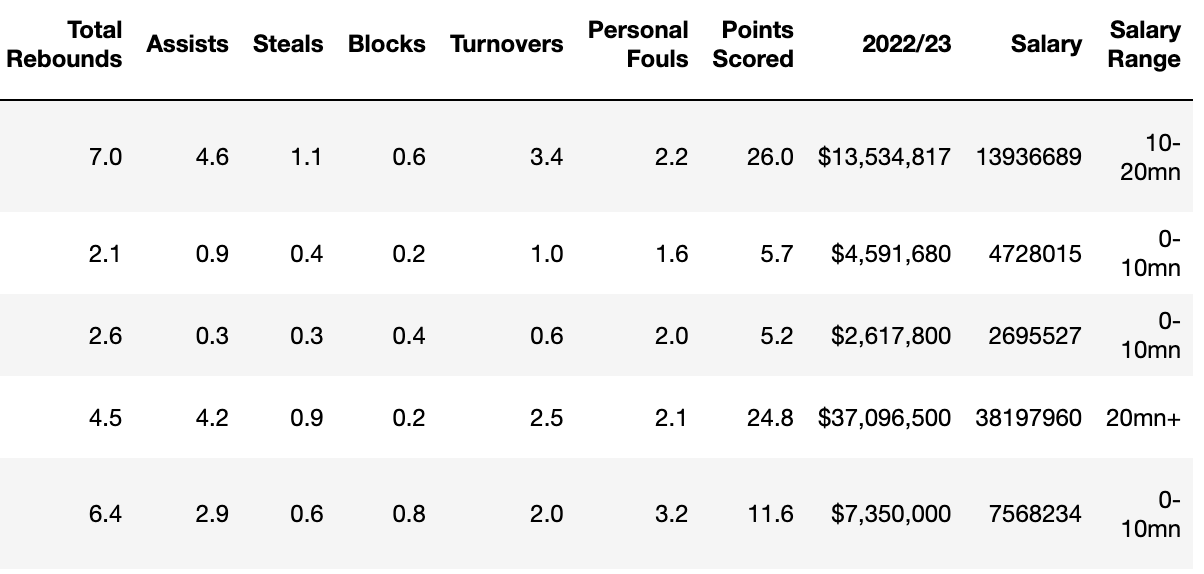
Data Cleaning

Because we incorporated data from two distinct sources, we encountered a few data management challenges. Approximately 6-7% of the players in our dataset had missing salary information. To address this gap, we employed a strategy of imputing these missing values by using the mean salary of all players, thus ensuring a more comprehensive dataset for our analysis.

The Kaggle dataset also contained duplicate records for some players. This duplication arose when a player changed teams during the season, leading to the presence of two rows for the same player. In order to rectify this issue, we adopted a criterion for selecting a single, representative row for each player. Our selection was based on the player's most recent team affiliation, as this reflects their current status within the league.

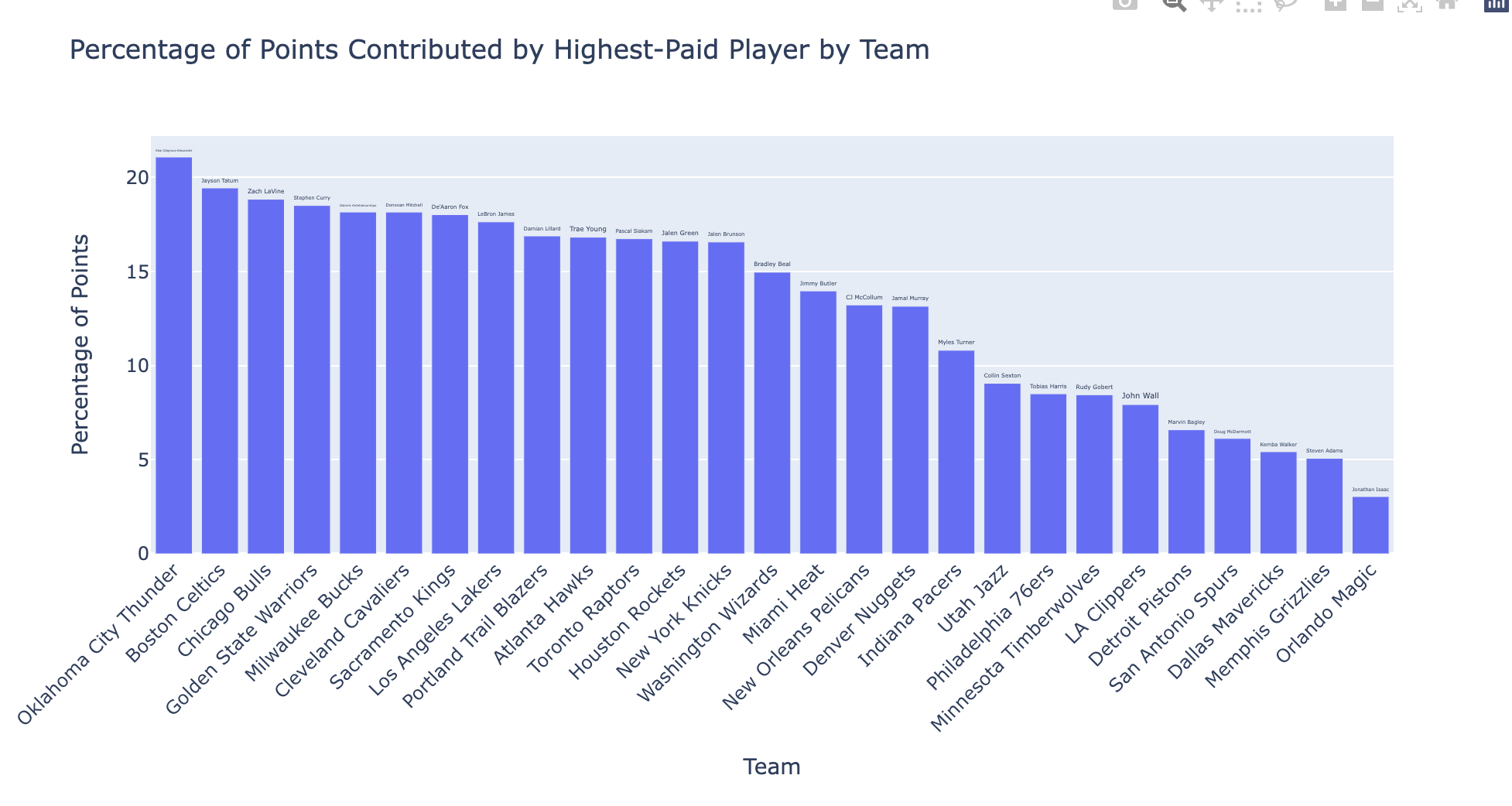
In our analysis, we deliberately narrowed down our focus to specific columns. This decision was made because the dataset contained an extensive array of columns, and not all of them were directly pertinent to the objectives of our analysis. By selecting only the relevant columns, we ensured that our analysis was concise and concentrated on the aspects of player and team performance that were within the scope of our investigation.

Exploratory Data Analysis (EDA):

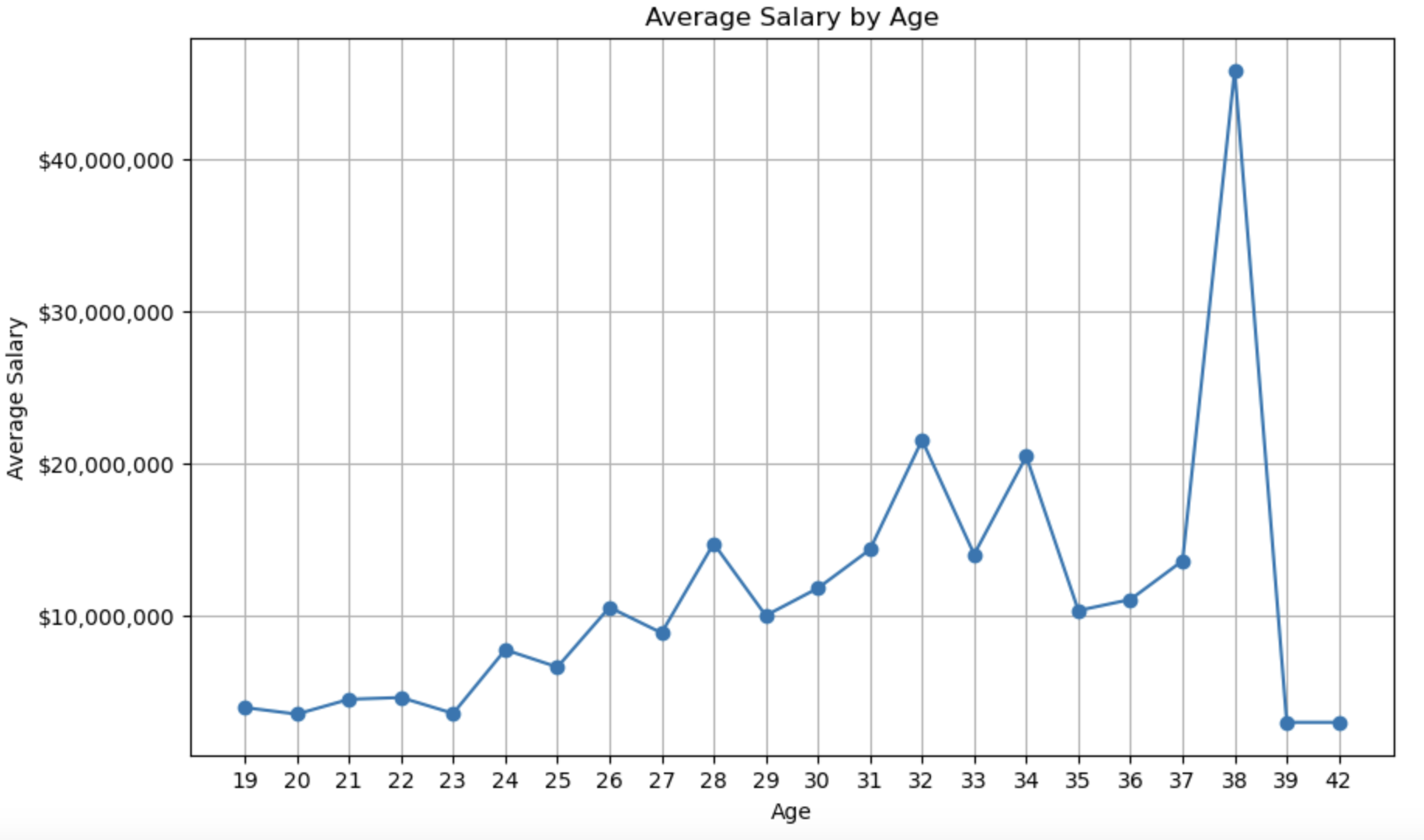


This is our finalized dataframe. We ended up combining our player stats table with the player salary table. Along with joining the two tables, We also had to rename a couple of different columns.

**Salary Analysis**

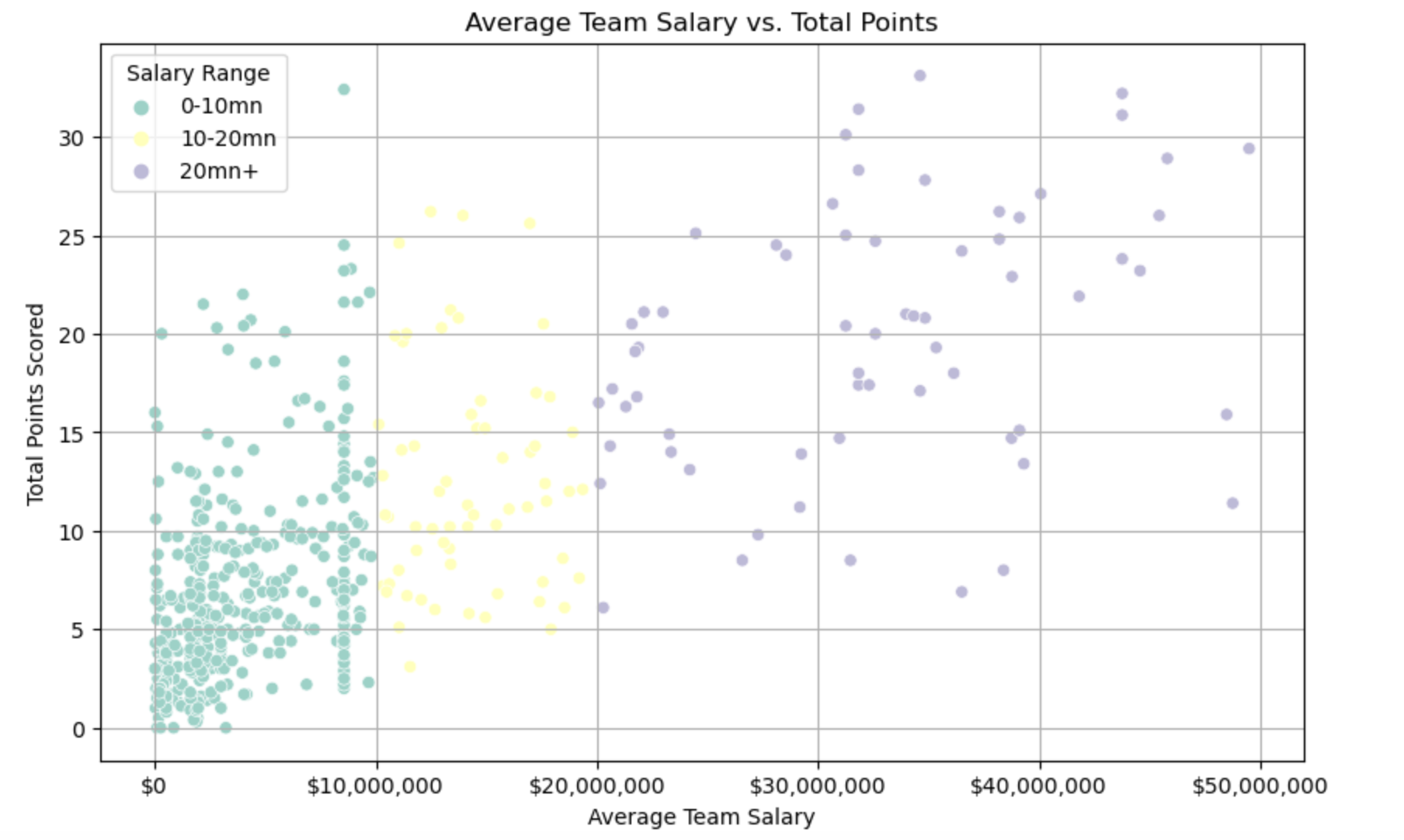


As the above graph shows for each team the highest paid player contributes quite significantly to the percentage of points scored by that team so it makes sense to have high salary players if they are contributing more.



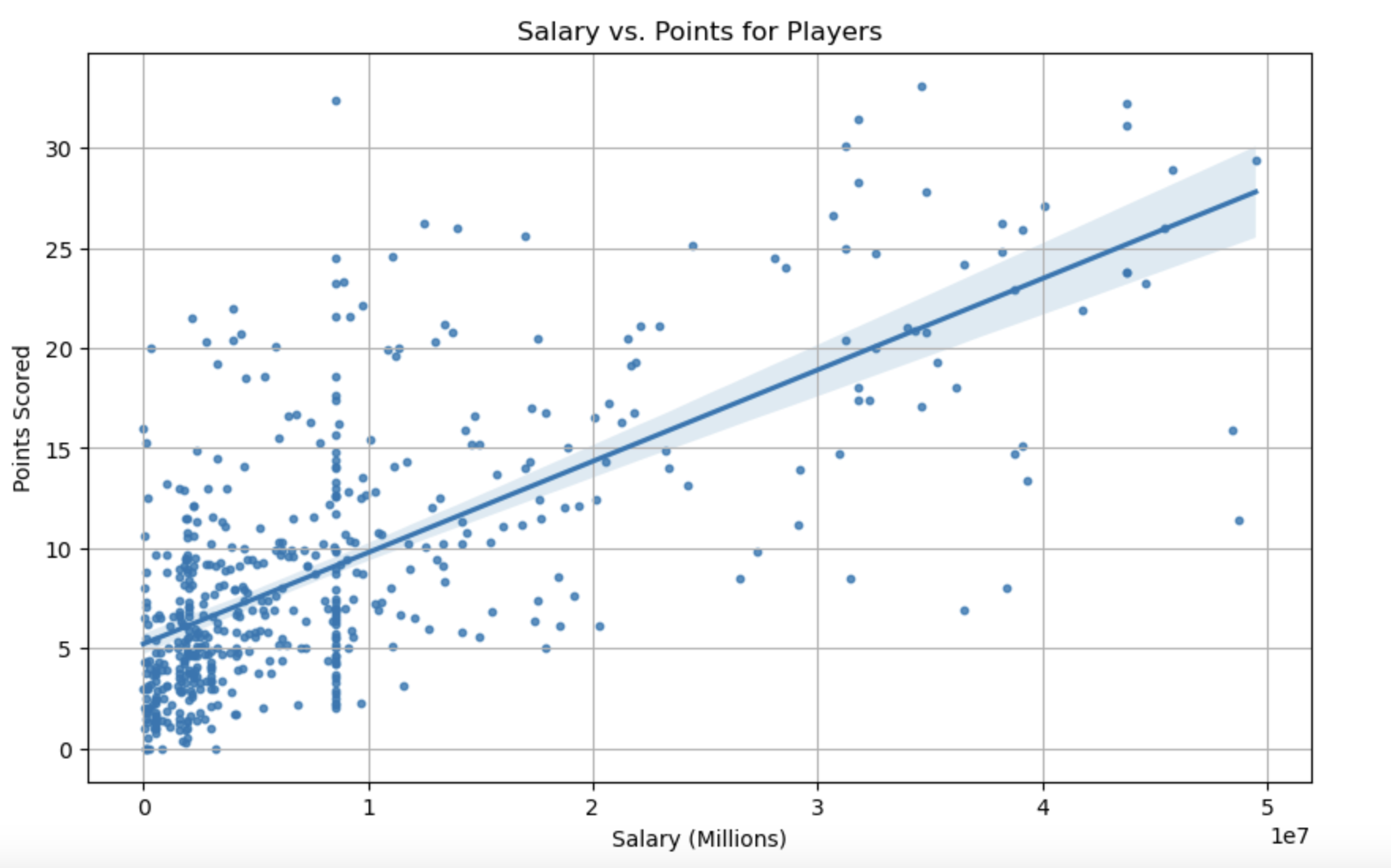
- The graph displays the relationship between basketball player age and their average salaries, showing consistent growth in salaries up to around 35 years.

- After this age, most players see a reduction in their earnings, with notable outliers, such as LeBron James at age 38, influencing spikes in average salary.

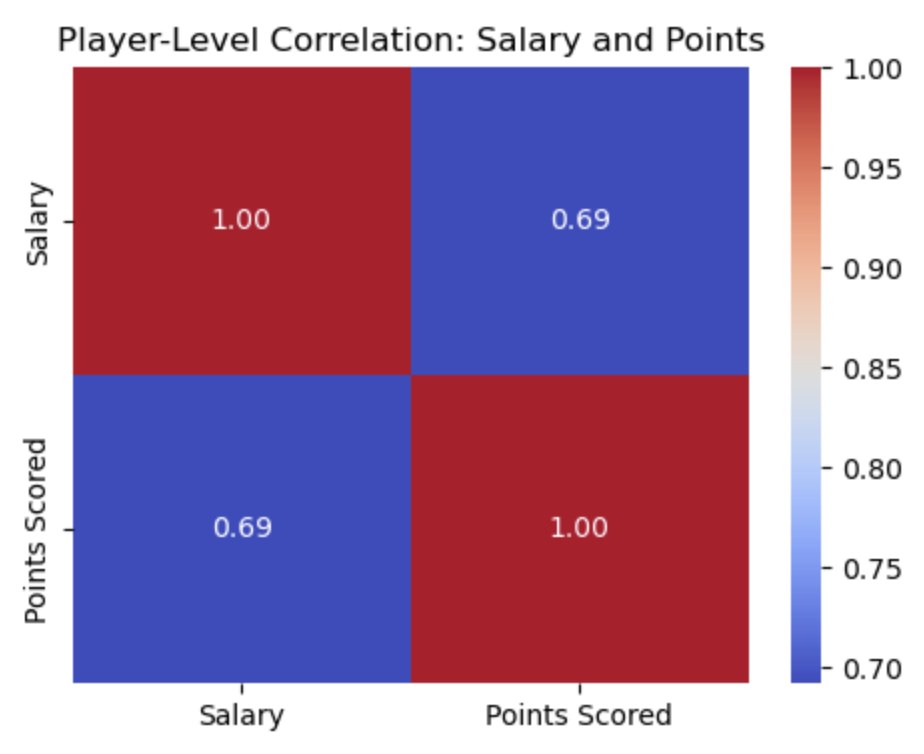


This is a graph that shows the average break down of team salary relative to the total amount of points scored. Please reference the key in the top left to help gain an understanding of where these points fall on the graph and why.

As we can see, those teams in purple which have a high team salary have high points as well.

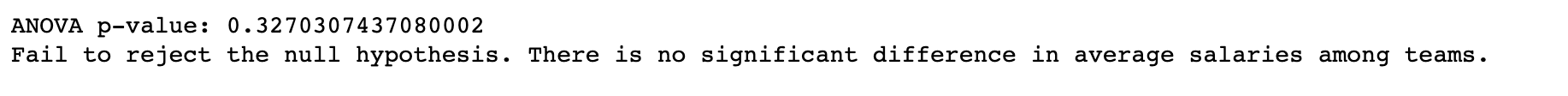


The above graph titled “Salary vs. Points for Players” illustrates the trend of increasing salary for players who score more points. The data indicates that most players that score more than 20 points on average are making upwards of 2 million dollars. The outliers in the data are indicated by those players who are scoring 20 or more points and are not making at least 2 million dollars in salary. This can be explained via incoming draft picks who are not fully established in the new league that they are playing in. With time and exposure these players will grow to make more money as a direct result of their exceptional performances on the court.

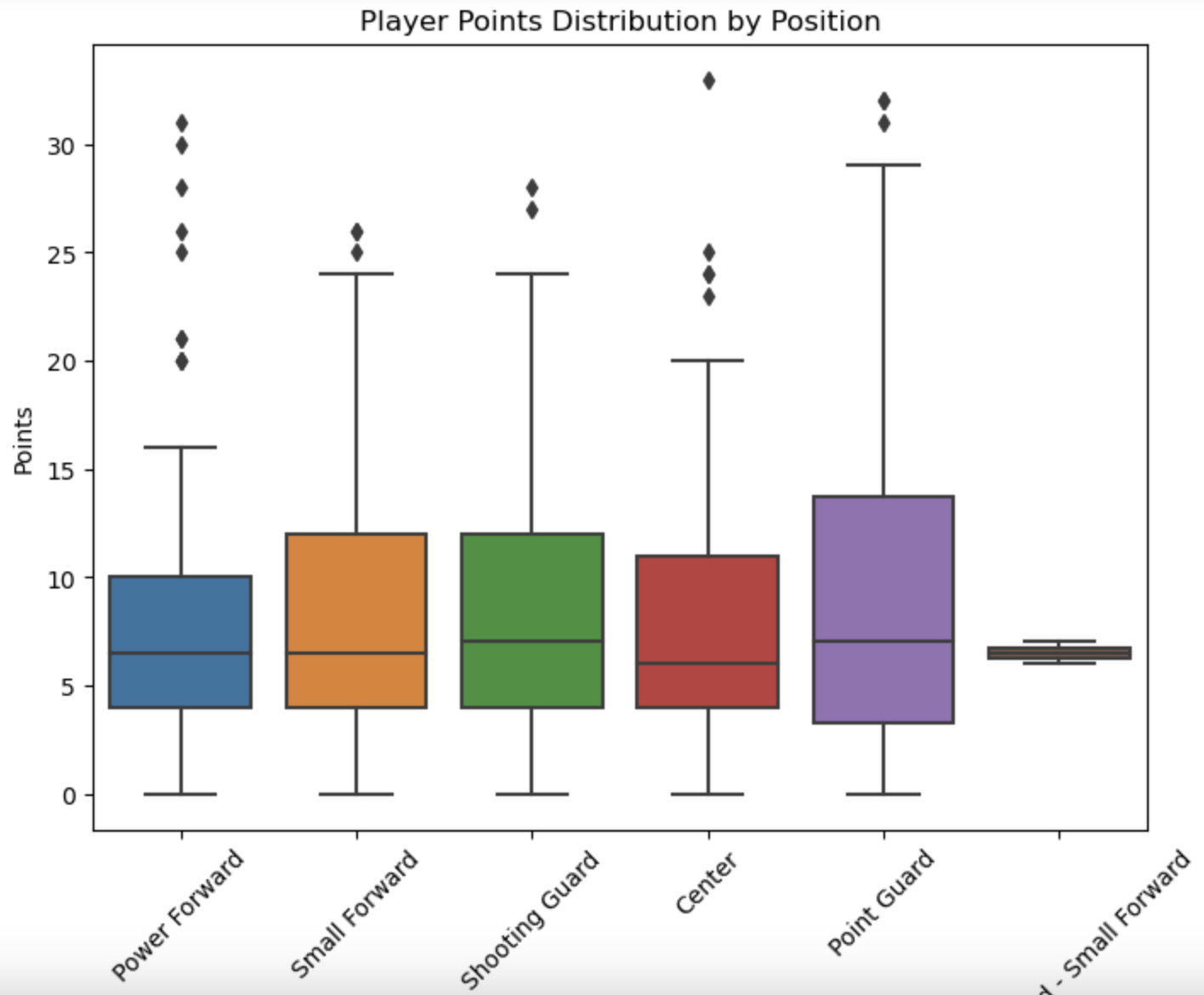


This correlation plot also shows that there is a strong correlation between percentage points and salary, coming out to be at 0.69.

To test the hypothesis on the team level we conducted a statistical test to see whether there is a significant difference in average salaries among teams. We failed to reject null-hypothesis since the p-value came out to be very large.



**Points Analysis**



This box plot is used to show the relationship between each position on the court and the points that they score. As you can see from the box-plot above, Point Guards typically score more points than the other positions on a basketball team.This is assuming that we are talking about the winning team.

To help give some further insights into this data, we should go over what each of these positions are responsible for on the basketball court.

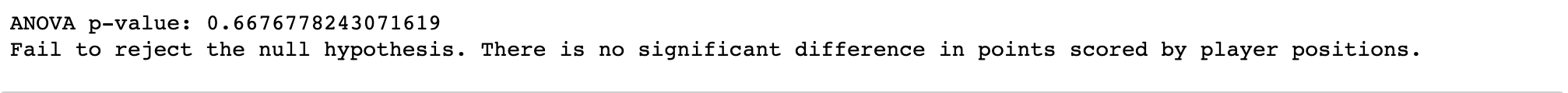
**Center:**The center is the tallest player on each team, playing near the basket. On offense, the center tries to score on close shots and rebound. But on defense, the center tries to block opponents’ shots and rebound their misses.

**Power Forward:**The power forward does many of the things a center does, playing near the basket while rebounding and defending taller players. But power forwards also take longer shots than centers.

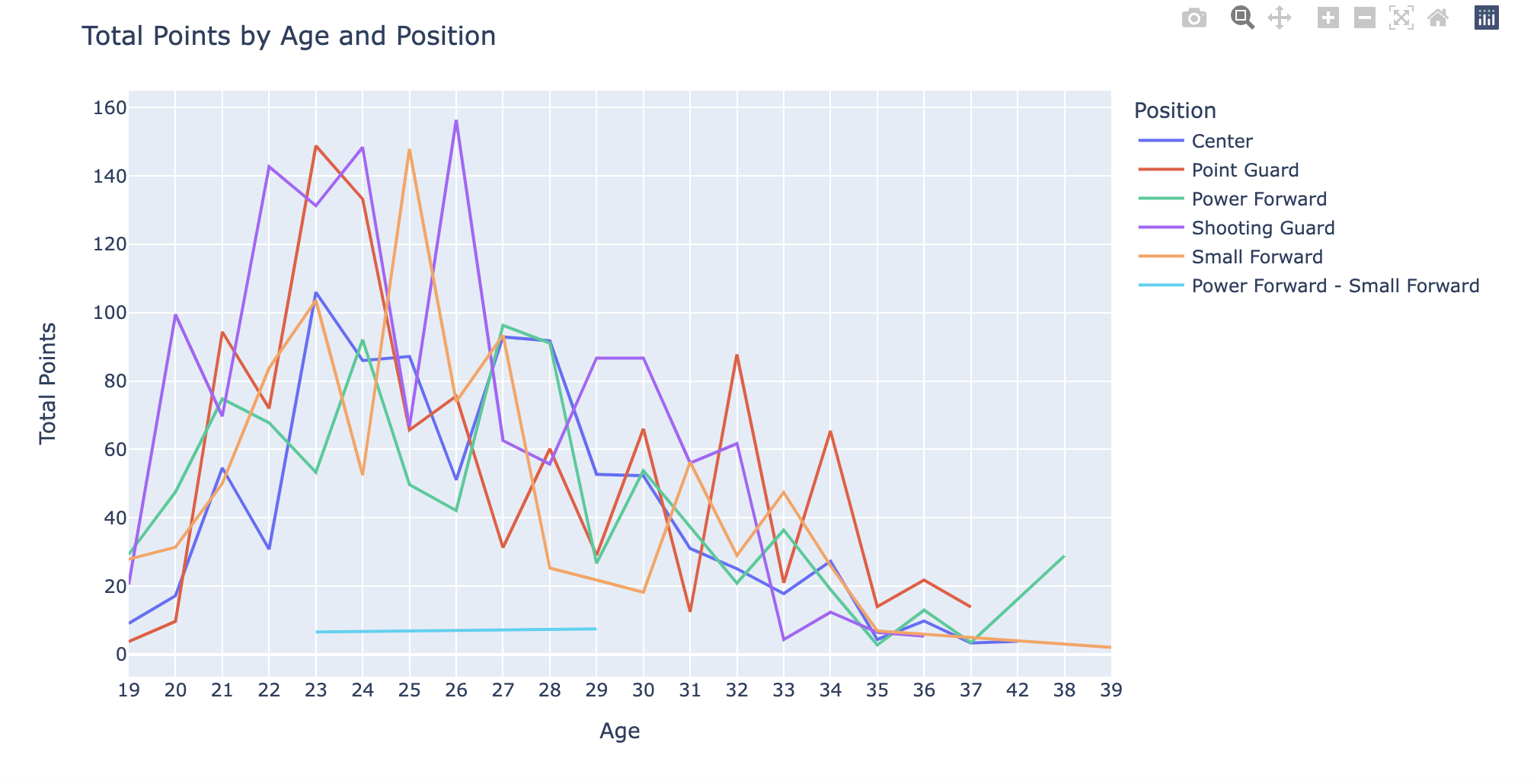
**Small Forward:**The small forward plays against small and large players. They roam all over on the court. Small forwards can score from long shots and close ones.

**Point Guard:**The point guard runs the offense and usually is the team’s best dribbler and passer. The point guard defends the opponent’s point guard and tries to steal the ball.

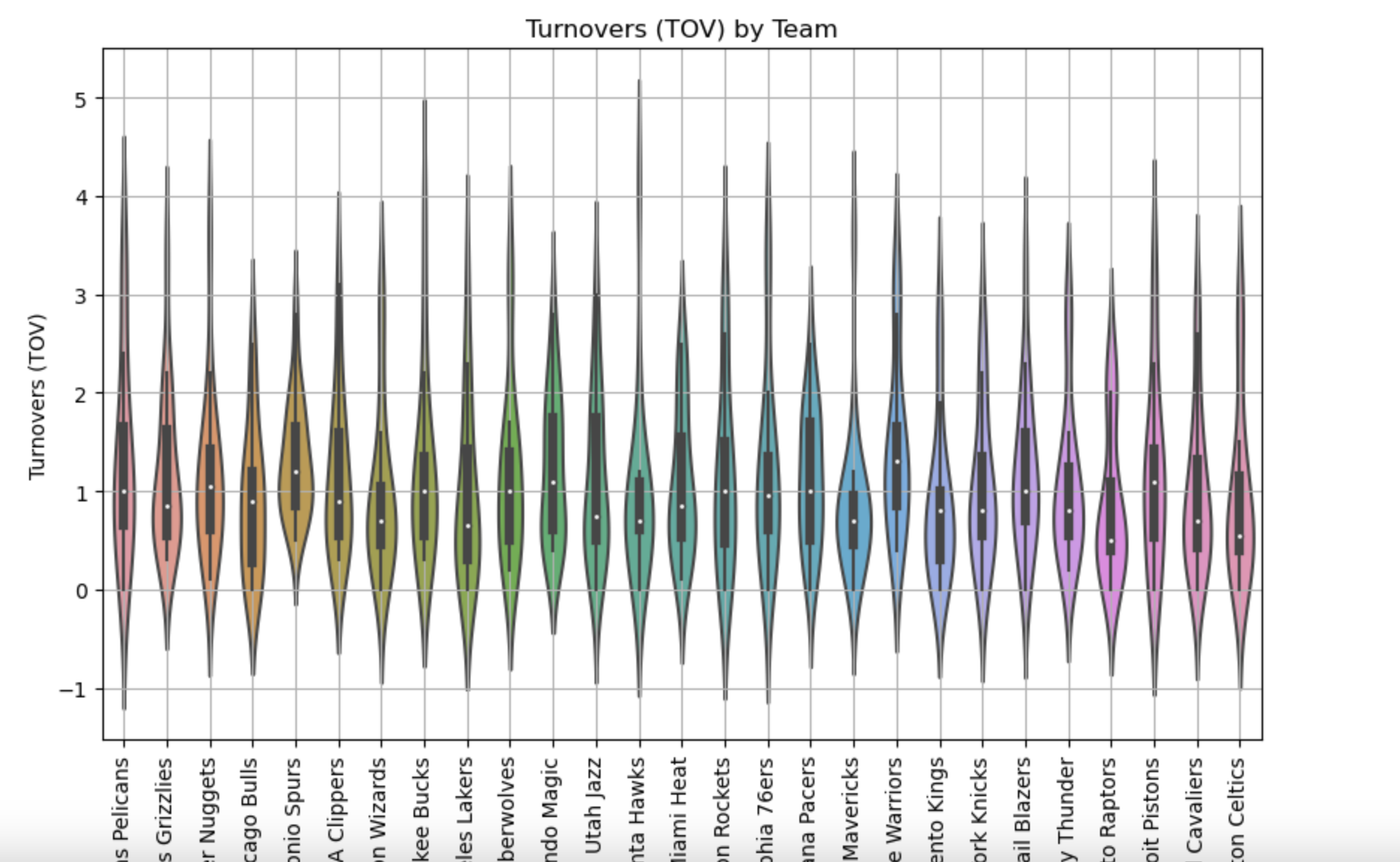
**Shooting Guard:**The shooting guard is usually the team’s best shooter. The shooting guard can make shots from long distances and also is a good dribbler.  
  
After further analyzing the positions of a basketball team, it is apparent that a point guard should always be the one scoring most of the points due to the fact that they are responsible for the offense within a basketball team.  
  
Now that we have analyzed how many points these positions score on average, lets see how much of a factor age plays on an athletes performance with scoring points.



Even though the box plot shows data that supports basketball position playing a significant role on the amount of points a player scores, we decided to conduct a hypothesis test to further this analysis. Given the ANOVA analysis, we have identified that there is no significant difference between the points scored by the player and their assigned position.



After analyzing the scoring averages of all players relative to position, we can also see what the prime age is for the best scoring athletes in the NBA. The common trend amongst all of the highest scoring Athletes via position is an age of roughly 23-26 years old. Across all positions, the highest scorers fall within this age range. After 26 years old, we begin to see a trend in the data where the older the athletes, the less they score.

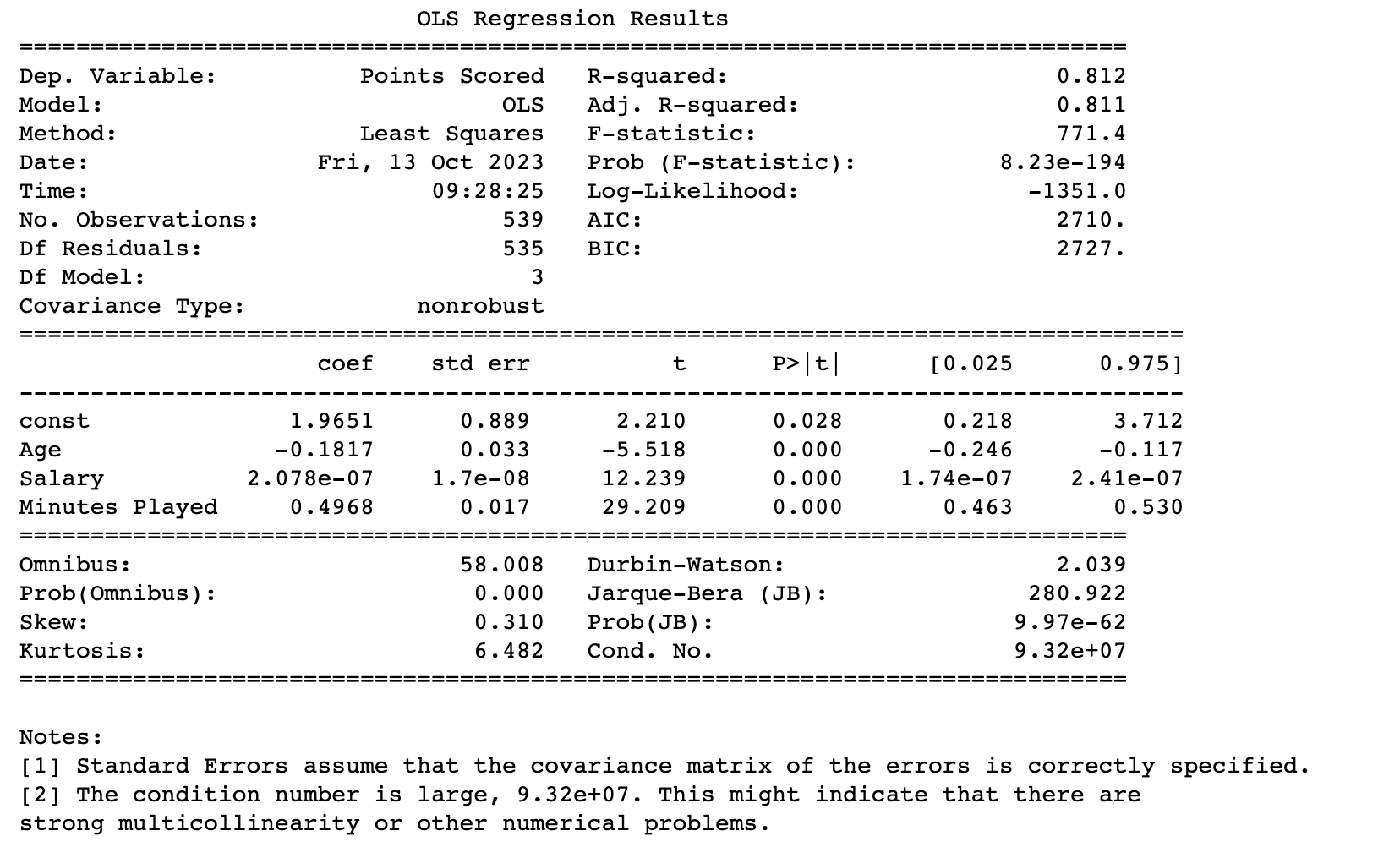


The following graph shows the amount of turnovers relative to each team within the NBA. For some context, A turn over is when one team loses the ball to the opposing team before they get a chance to score any points. This graph helps to illustrate the amount of times each team has turned over the ball on average per game. Teams that yield more turnovers will have fewer players who are scoring an above average amount of points, regardless of position.

**Point Analysis Conclusion:**

After observing our analysis in the above graphs and charts, we can recognize that the amount of points scored is relevant to the age of the player as well as the position. A player who is a point guard is more likely to score more points than the other positions. A point guard within the age of 23 to 26 is going to be the most likely to be at their optimal athletic performance level which will as a result yield a greater average amount of points scored. The athlete's position on the basketball court can help hedge the amount of points they are capable of scoring, however there is no direct relation to the position of the player compared to the amount of points that they score. There are still several players who may even play a position that focuses more on defense than offense and manage to score a significant amount of points.

**Multiple Linear Regression Model Summary: Predicting Points Scored**



We conducted a regression analysis to see how Points Scored are related to age, salary and minutes played.

This model provides valuable insights into the relationship between Age, Salary, Minutes Played, and Points Scored. The variables 'Age' and 'Salary' are statistically significant predictors of 'Points Scored,' while 'Minutes Played' also has a significant positive impact.

**R-squared (R²):** 0.812

* Indicates that approximately 81.2% of the variation in Points Scored is explained by the independent variables.

**Coefficients:**

* **Age:** -0.1817 (p < 0.001)
  + For each year increase in Age, Points Scored is estimated to decrease by 0.1817 units.
* **Salary:** 2.078e-07 (p < 0.001)
  + For each unit increase in Salary, Points Scored is estimated to increase by 2.078e-07 units.
* **Minutes Played:** 0.4968 (p < 0.001)
  + For each additional minute played, Points Scored is estimated to increase by 0.4968 units.