**NBA Player**

**Salary and Points Predictions**

By

**Samantha Beard**

DSC 550 | Term Project

Introduction

As a business, the National Basketball Association (NBA) team owners and coaches want to ensure they are employing players who perform well for them. They also need to know what a reasonable salary is for the players they are looking to add to their team. Often it can be difficult to know how a player will perform, especially if they are new to the league. Additionally, each team only has 15 positions on their team; they have limited resources, and they want to make wise decisions with those resources.

The predictive model that was created looks at the player statistics from the previous season, in this case we are reviewing the 2022 statistics. The following attributes were assessed: Position, Age, Team, Games Played, Games Started, FG Percent, 3P, 3P Percent, 2P Percent, Free Throws, Free Throw Percent, Offensive Rebound, Total Rebound, Assist, Steals, Block, Turnovers, Personal Fouls, Points, and Salary, with Salary and Points being the targets. This data was obtained from multiple Kaggle datasets, however, could be web scraped from ESPN or the NBA’s websites.

This model could be pitched to all the teams together or individually. To pitch it to individual teams, I would want to conduct further analysis on how players work together as that would allow them to not only identify players who will be individual contributors, but predict how they will work with the existing team.

Summary

The data required quite a bit of preparation to get it ready for modeling. First, dropping columns that had no significant data. This was determined by looking at what columns had little to no data or had columns of data that did not have a column name that made sense. I renamed columns to increase readability. This included changing things such as ‘Tm’ to ‘Team’ and ‘GS’ to ‘Games Started.’ I also dropped columns that had correlation scores over 90 percent, as that indicates that the data is redundant.

Since the salary information was based on 2022 data, I limited the player stats to the 2022 season as well. I joined the stats and salary data frames on the player name, and then dropped the player names to focus on the numerical components. I also dropped rows that only had salary information or only had statistics information. Additionally, there were four columns – field goal percent, three point percent, two point percent, and free throw percent – which had null values that I filled with the median percentiles in each of these categories. Lastly, I created dummy columns for position and team data. This cleaning took place across the first three milestones, and then was reorganized for the final code submission.

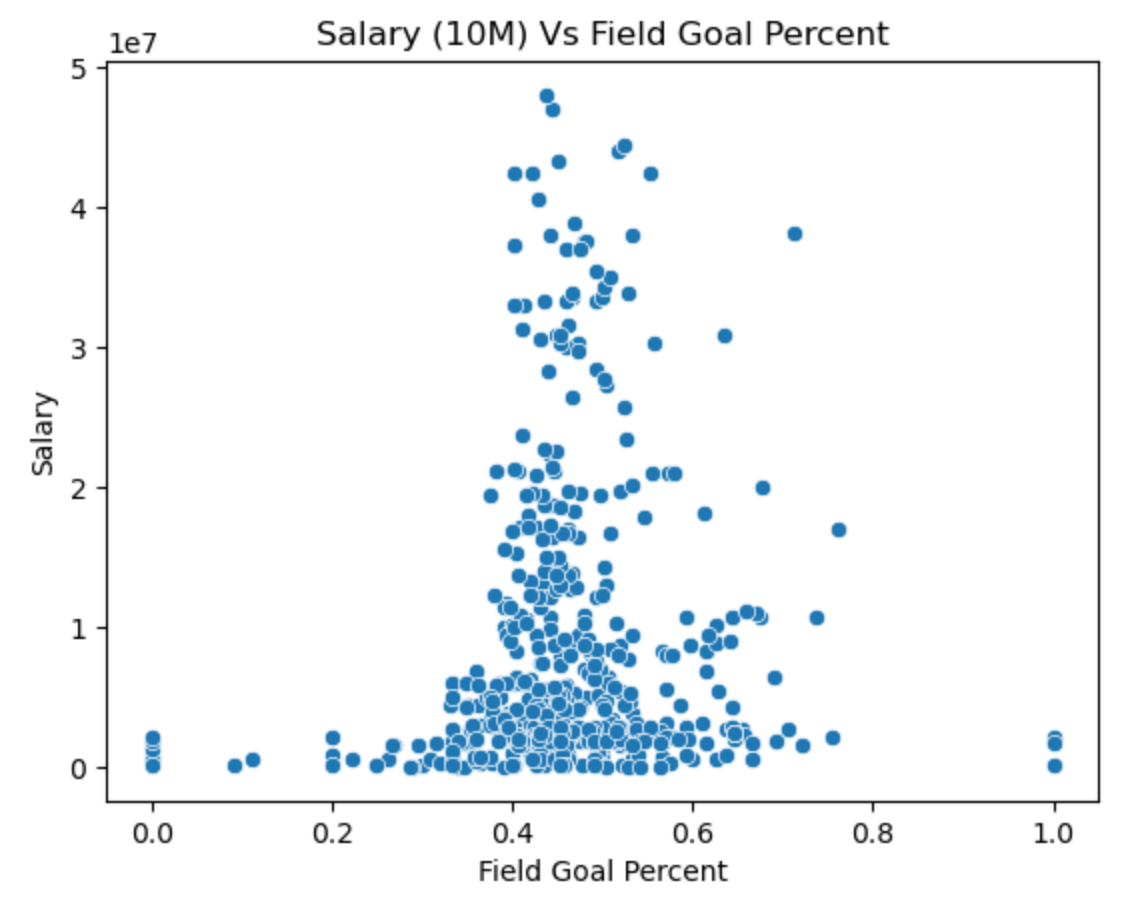
After getting all of the data clean, I was able to some exploratory data analysis (EDA) visualizations, comparing several columns to both points and salary.

When comparing salary and points, there’s a loose correlation.

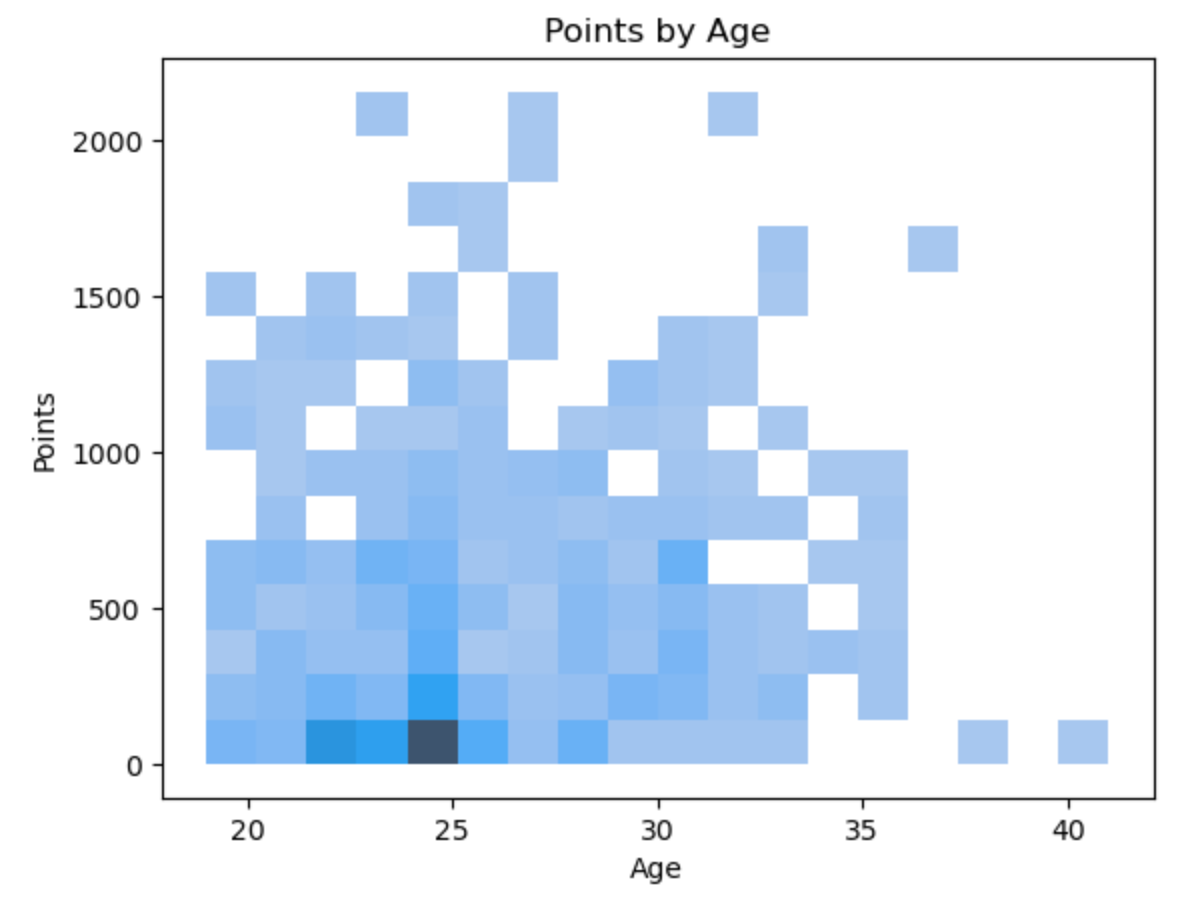
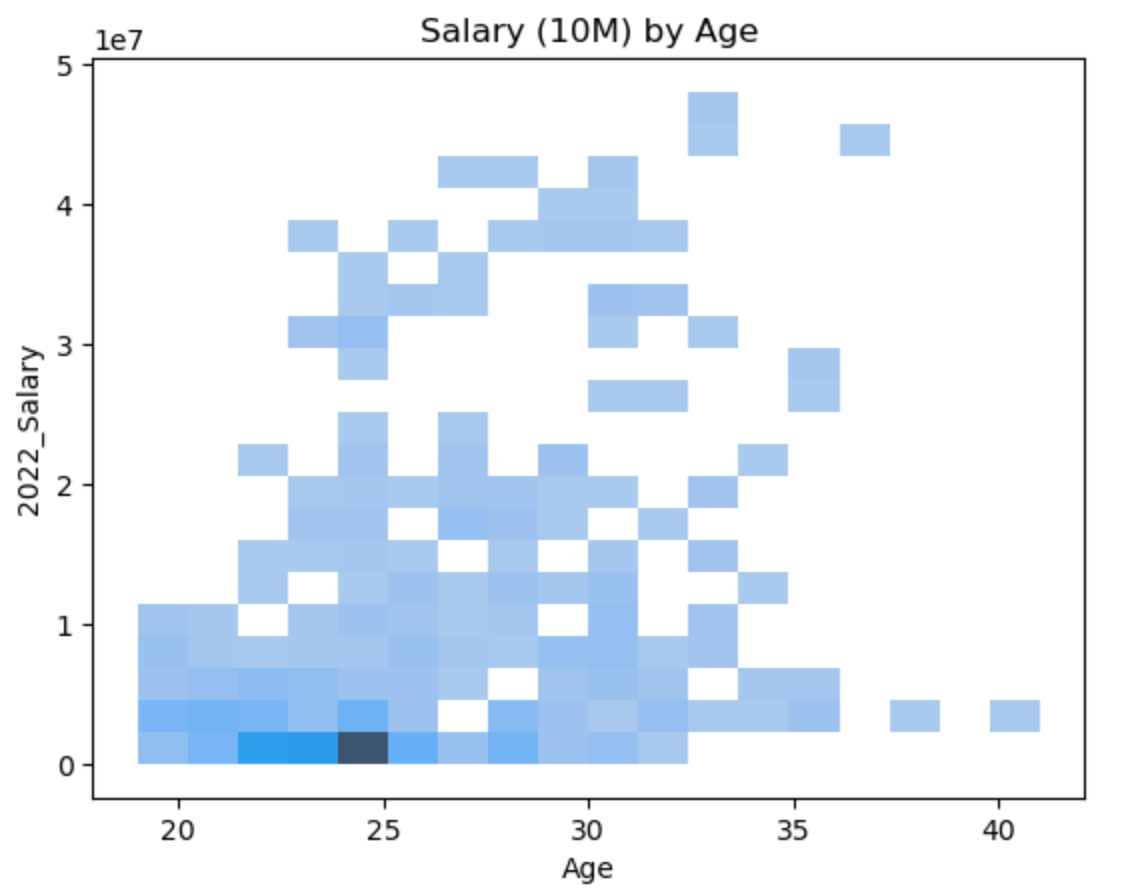
A graph of blue dots

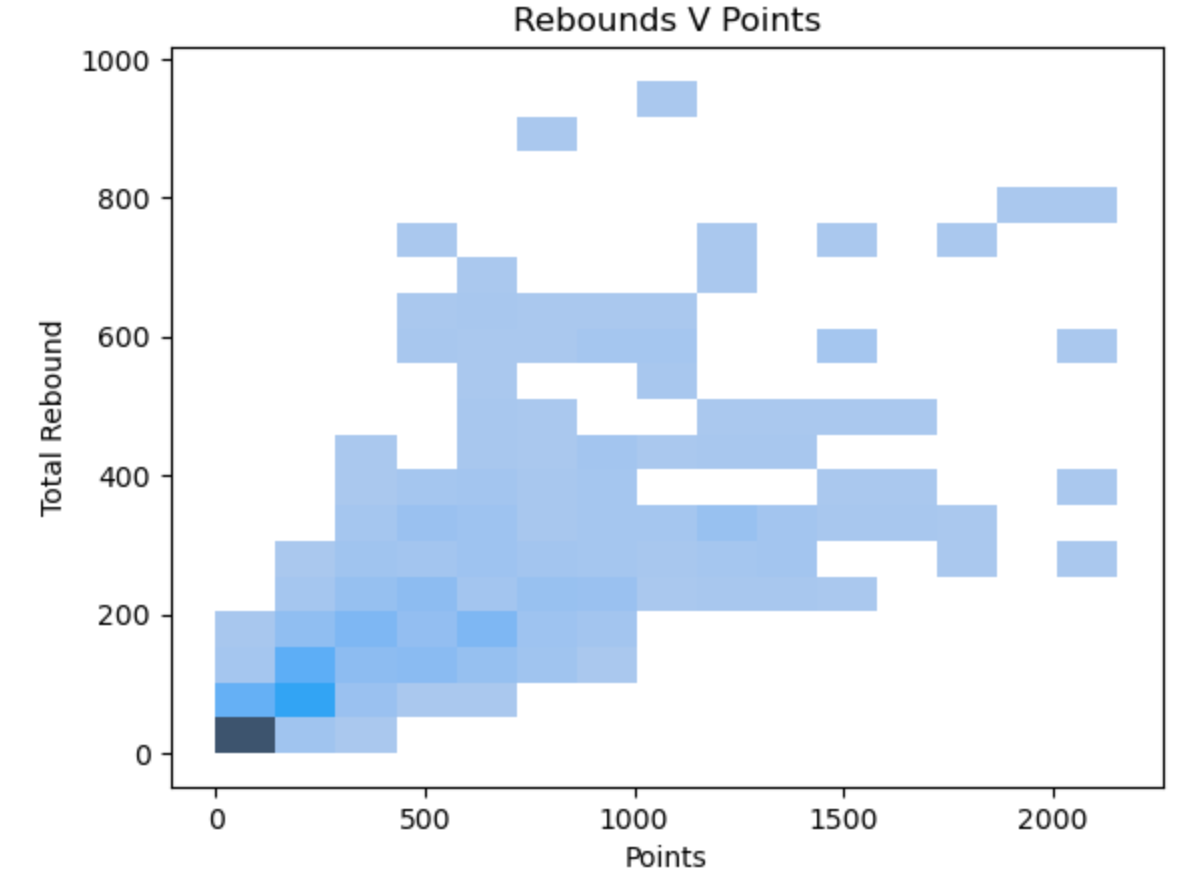
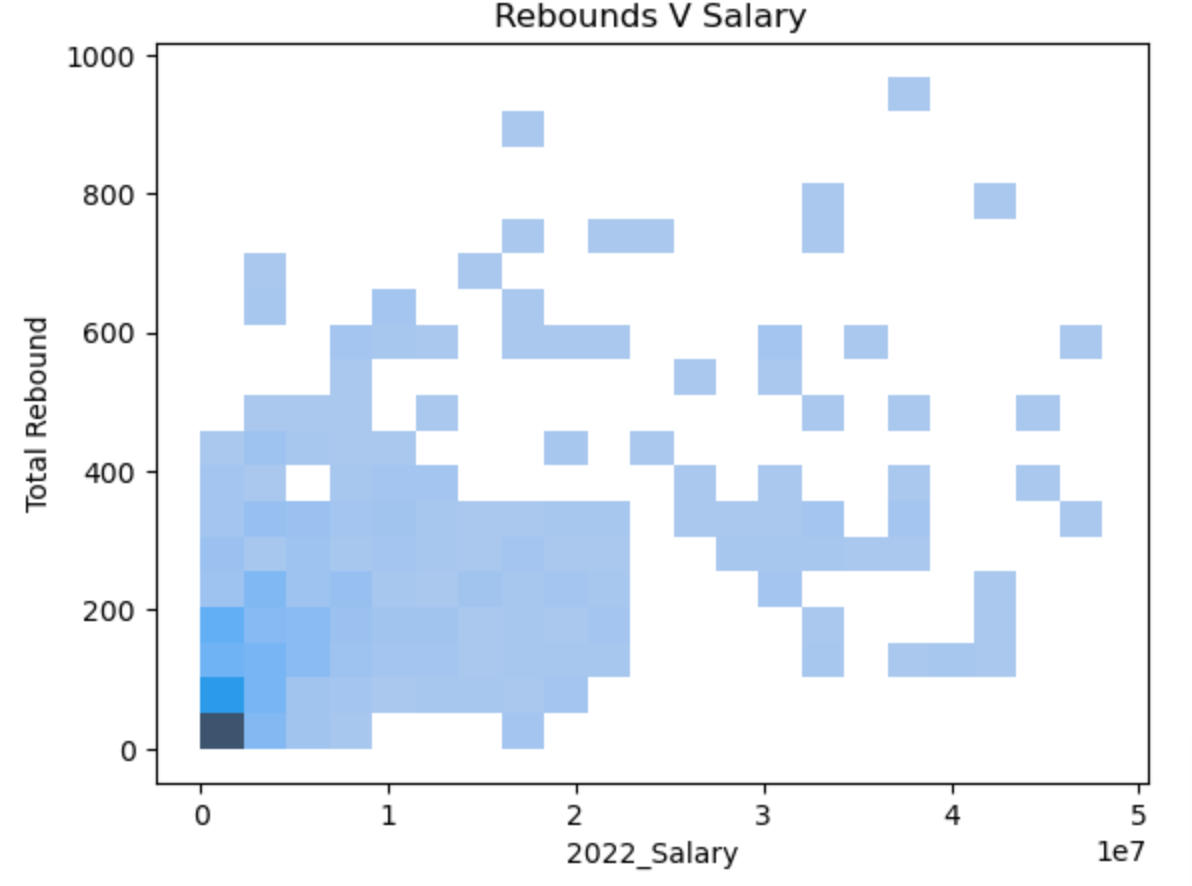
Description automatically generated

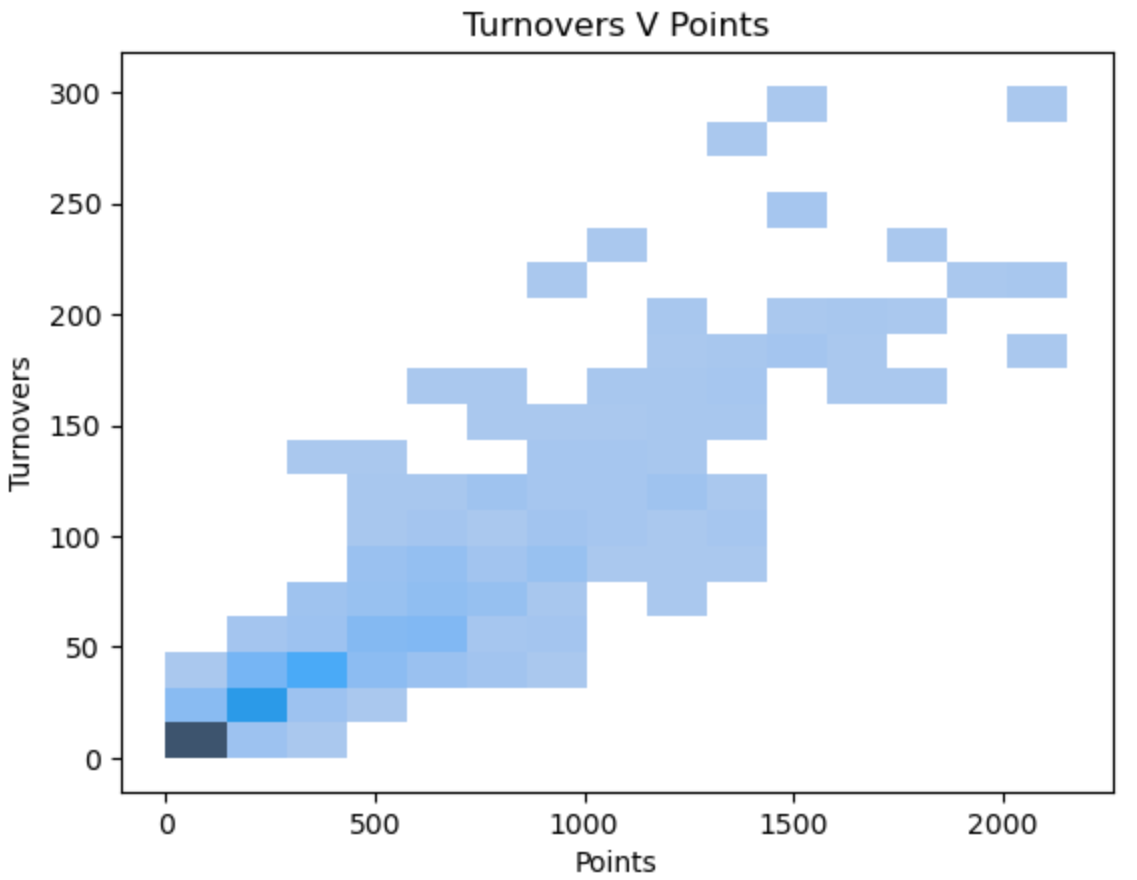
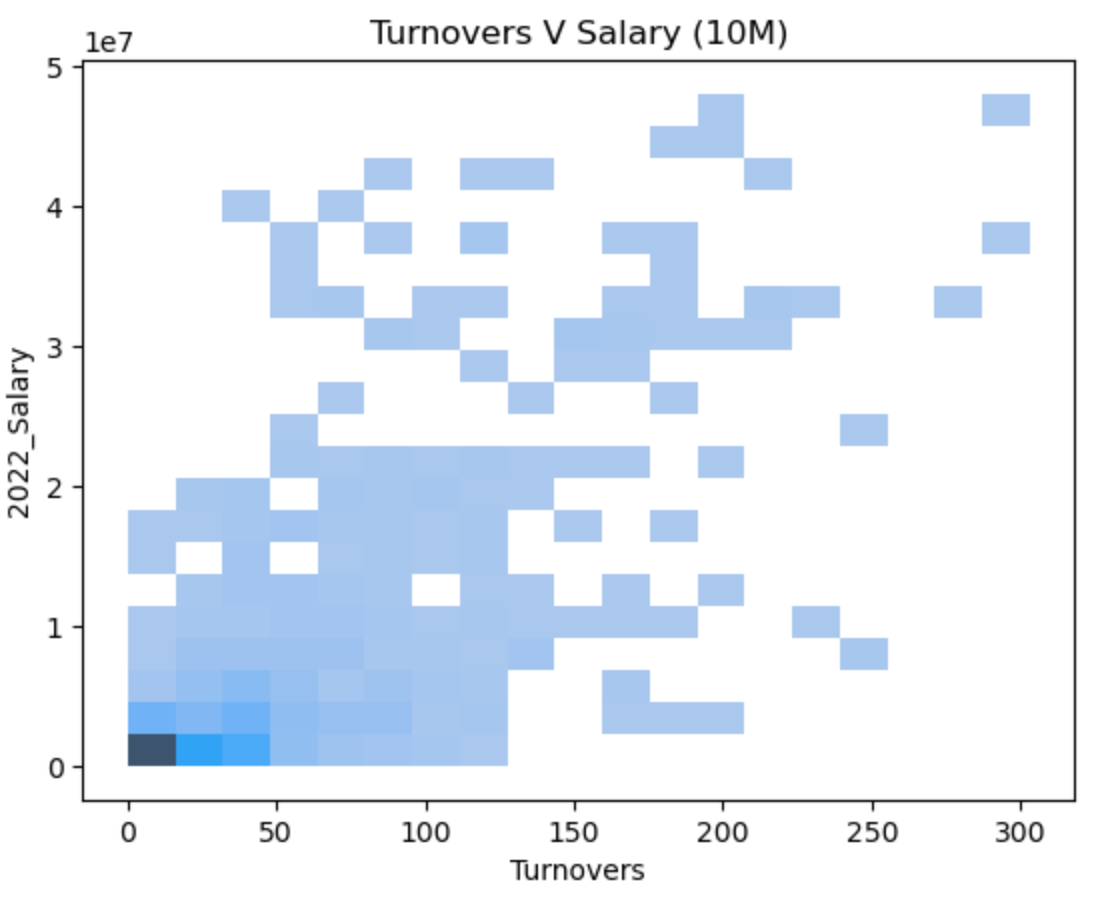
However, there is no correlation between salary and field goal percent, which is another target I considered. Players of all salary ranges appear to be in the 40-50 percent range.



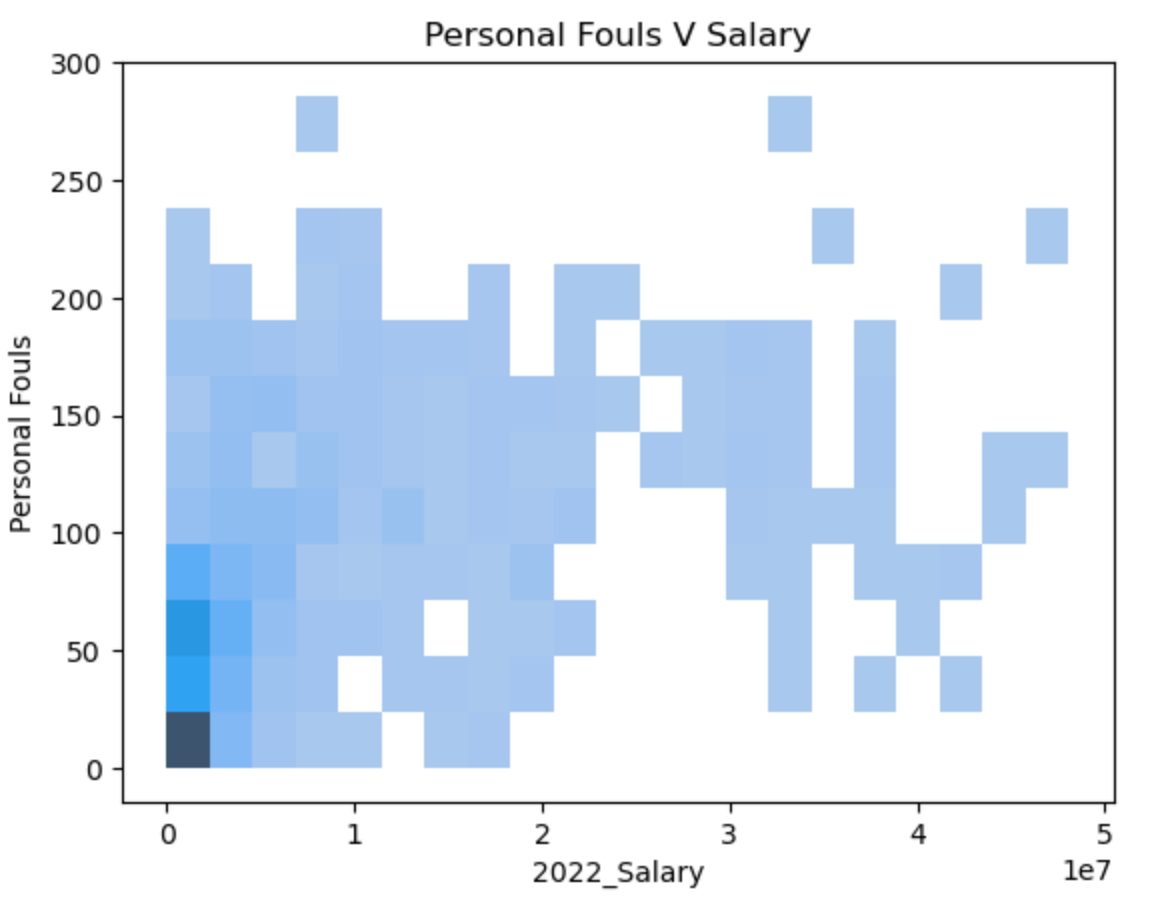
I then compared games played, age, rebounds, turnovers, and personal fouls to both points and salary. There were more trends when comparing these areas to points than there were when comparing to salary. As a business, I would want to get the best player for the least amount of money as possible.

A graph of a person's foul

Description automatically generated 

After this EDA, I was able to split the dataset into training and testing sets. I then fit a few different models. Decision tree regressor did not perform very well with an R2 score of .295 for salary and .773 for points. Random Forest performed significantly better with an R2 score of .705 for salary and .918 for points. I did have to play around with the parameters of the random forest to get these scores. The biggest factor was using a poisson criterion, meaning that the events occur randomly and independently of one another, but at continuous rates. I also looked at a multioutput regressor because I was wanting to predict salary and points of a player, and not just one target variable. Multioutput regression is best when variables are interconnected, so this helped minimally since this data performs best with poisson criterion. Multioutput regression also works with other regression models. I tried several, however, Random Forrest was still the best model, with the multioutput regression increasing the R2 score minimally to .716 for salary and .945 for points.

Conclusion

Given the EDA showed more trends across variables with points than with salaries, the r-squared score, which shows how much variability is accounted for, indicated that approximately 94% of the variability in points scored is accounted for and only 71% of the variability in salaries is accounted for. This would indicate that a player’s salary is less predictable than the number of points a player will score based on their other stats. Although not a perfect indicator of what the player should be paid, it would gives a ballpark to offer a player. Additionally, a team would have a good idea of how many points to expect from a player based on their other statistics.

Some areas that I wanted to explore, but due to the limited data and datasets not joining correctly I removed them from my project, include injuries, collegiate statistics, and demographic, such as height, weight, nationality, etc. Exploring injuries would show trends in how frequently a player may get injured, as well as how long they may be out of commission. Collegiate statistics would be great to explore when recruiting rookies and looking at how well they perform once they get to the NBA. Player demographics would be interesting to see how these things play into a players performance. Obviously a lot of basketball players are tall, but is a 6’ 8” player better than a 6’ 1” player?

I also think it would be interesting to see how players work together as a team and against specific players on opposing teams. This would allow coaches to have a more objective lineup, and be able to put players in who are going to have maximum performance for that particular game. Obviously there is human error and emotions, so I would not expect a statistical outcome greater than 80 percent.

Overall, I think the model is usable for high level analysis of players and what a team can expect from the players, but there is a lot more that could be done to provide additional insight into what makes a player perform and what a team can expect from players.