Investigation to NBA Players and Their Salary

Summary of Questions and Results

1.What's the relationship between salary and average points gained per game?

In different periods, the comparison of salary is meaningless. The "absolute" salary of NBA players is unrelated to their scoring or defensive abilities due to variations over time. However, the "relative" salary of players (compared to other players in the same season) is **positively linearly correlated** with their scoring abilities.

2.What attributes (from height, scoring ability and 3 point goal ability) of players make a difference to their salary?

We have observed from the scatter plot that a player's salary increases significantly with an improvement in their three-point shooting abilities. However, it is important to consider that NBA player salaries are higher now than in the past, and present-day NBA players place more emphasis on their three-point shooting abilities. Therefore, it is possible that the high salaries of players with exceptional three-point shooting abilities are driving this trend in the plot. Certainly, we can still recommend future NBA players to **practice their three-point shooting abilities**, as the current trend in the league shows an increased demand for three-point shooters. Improving one's three-point shooting can potentially lead to a **higher salary and increased opportunities in the league.**

3.How does height influence those factors?

Although height has an objective effect on a player's performance on the court, height is **not** a **determining factor** in a player's performance on the court, and it is **not enough** to affect a player's salary.

4.Given the college (non-NBA) player's information, can we predict their salary if they want to get into the NBA?

The answer is **YES**. The R-Squared value of our prediction model is about 0.4. Although this is not a very high value, it is enough to make a reasonable prediction.

Motivation

As one of the most popular sports, basketball affects people deeply. Many people started playing basketball when they were little kids. They don't play just for fun but also for dreams. By watching the NBA, they can learn about players who have superb skills and legendary deeds. They dream to be one of them. Our program can analyze the relationship between salary of an NBA player and his other conditions comprehensively and users can also use this program to predict their salary if they were NBA players.

Dataset

https://github.com/KiriSchrieffer/NBA-salaries
This dataset contains two csv files. One is about players' personal information and career record. The other one is about the salary.
Data sourced from basketball-reference.com, which has Basketball Stats and History Statistics, scores for the NBA players.
Salary Cap data sourced from https://basketball.realgm.com/nba/info/salary_cap which showed the salary cap from 1985 - 2035 (future salary cap)

Use python crawler to get test dataset from http://www.espn.com/nba/salaries and https://www.basketball-reference.com/leagues/NBA_2023_totals.html

Method

Research Question 1:

We use the Altair Library in Python to build an interactive graph.

We begin to calculate the player's career average salary in Salary data using groupby and calculate the mean of each player's average career salary. Store the calculated dataset into a new DataFrame name average_salary.

Next step, filter the Player Info data with three remaining columns 'player_id' and 'salary' and 'start_season', we calculate each player's career average points per game in Player Info data using groupby function to calculate the mean of each player's average points per game.

After that, we will merge the two new DataFrame together by '_id' and 'player_id' (which is unique for each player) and use Altair.Chart to plot the scatter point graph of the relationship between salary and points gain per game.

Then, we preprocessed the salary cap dataset with Excel built in function to transform the 'season' (i.e., 2021 - 2022) to 'start_season' (i.e., 2021) in order to merge with the Salary dataset.

The merged dataset group by ‘player_id’ and calculate the mean of career salary and average salary cap.

Merge the processed dataset with Player Info dataset and use **Altair.Chart** to plot the scatter graph, and this plot will help us determine the relationship between salary and scoring ability.

Research Question 2:

We filtered the Player Info dataset, with ‘_id', 'career_PTS', 'height', 'name', and 'career_FG3%' remaining, and we stored the new DataFrame.

Merge the new DataFrame with the Player Info Dataset by ‘player_id’ and ‘_id’.

We use **Altair.Chart** to plot the scatter point graph, with repeat function set attribute columns as 'height', 'career_PTS' and, 'career_FG3%', attribute row is ‘salary’.

Finally, we save the chart as .html form.

Research Question 3:

Before all

Each color represents a player. Since there are too many players participating in the survey, the chart will look very crowded and messy if all of them are labeled out. As a result, we decide to hide the legend. If looking closely, it is confusing that the same color of points will appear several times. This is because each row only represents the data of this player in this year, and our data has a certain time span, which means that if the same player has been in the service for a long time, then he will appear several times.

The first analysis is how height affects a player's performance on the field. A player's performance is overwhelmingly measured by the player's on-court stats, and scoring is one stat that is highly valued. Therefore, the first comparison we designed is the comparison of players' height and field goal percentage. It is a common perception that most people believe that smaller players are better at making shots. Especially in modern basketball, small players who don't have consistent shooting ability are generally seen as the alternative. But is this really the case?

We first plotted using a new library, called **plotly**, that was learned outside of class. Since we want to know the effect of height on field goal percentage, we set height as the independent variable and field goal percentage as the dependent variable. Although there are outliers, the outliers are not very obvious, and the overall distribution of points in the image is very random and there is no clear trend. In order to enhance the rigor of the conclusion we choose to verify by r-square value. **R-squared (R2) value** is a statistical metric that evaluates how well the regression line aligns with the data in a scatter plot. Its scale ranges from 0 to 1, where a higher value indicates that the regression line is more closely aligned with the data points. Then using sklearn's linear_model: LinearRegression we calculated the r-square value.

Then we thought that a very common and efficient way for tall players to score is to create fouls in the opponent's three-second zone and score on free throws. As a result, many tall players spend a lot of time and effort practicing their free throw consistency and shooting percentage. Considering this factor, we devised the next comparison: height and three-point shooting percentage.

Following the same process as we did in the first comparison, we set height as the independent variable and three-point shooting percentage as the dependent variable and plotted the scatter plot using plotly. Then, we calculated the r-square value. However, because of the explicit existence of outliers, we decided to improve our plots using a statistical method: processing the data by dividing them into quantiles. The above process was then repeated to obtain a new plot and r-square value.

We also all know that players' salaries are closely tied to their performance on the court, so we want to see if there is a correlation between a player's height and their salary. This process is basically the same as the second one: first plotted with height as the independent variable and salary as the dependent variable. Then, because of the existence of large number of outliers, we improved by dividing our data into quantiles, and then plotted the scatter plot and calculated the r-square value using plotly and LinearRegression model.

Research Question 4:

In this part, we analyzed the relationship between NBA players' performance and their salary by linear regression. The ultimate goal of this model is to provide accurate predictions of the salary of players, if they attend the NBA, based on certain factors.

Firstly, imports the necessary libraries for data manipulation, machine learning, and data visualization. Among them, LinearRegression is a new class.

And then, creates two new data frames by selecting certain columns from the player data and salary data, merges the two data frames based on the season start year, groups the merged data by player ID and calculates the mean salary and salary cap for each player, and finally calculates the ratio of a player's salary to the league's salary cap.

To develop our model, we used three datasets: players.csv, salaries_1985to2018.csv, and salary_cap.vsc. The data included various performance metrics. In this model, we use Field Goal Percentage, 3 Point Field Goal Percentage, and assists in the players.csv. Salary information for each player in one year came from salary_1985to2018.csv. The dataset salary_cap.csv contains the salary cap information in the NBA from 1985 to 2018.

We used machine learning algorithms to analyze the data and find patterns between performance and salary. The Field Goal Percentage, 3 Point Field Goal Percentage and assists are considered as performance of a player. And the result of salary divided by salary cap is used to train the model. The reason why we use the salary cap is that our data cover through several decades. The players in the 1980s got much less salary than players in the 2010s get. To eliminate the error caused by inflation, we use the percentage of an individual player’s salary in the salary cap.

Explanation of salary cap: The salary cap in the NBA is a limit on the total amount of money that each team is allowed to spend on player salaries in a given season. The NBA's salary cap is determined by the league's revenue, with a percentage of the revenue allocated to player salaries. This percentage is negotiated between the league and the players' union during collective bargaining.Each team's salary cap is adjusted each season based on the league's revenue for the previous season. This means that if the league's revenue increases, the salary cap will also increase, allowing teams to spend more money on player salaries.

The use of a salary cap increases our **R-squared value** from about 0.4 to about 0.6

While filtering the data, we found that many players got 0 or 100 on their 3 point field goal percentage.

This kind of extreme data also influences the model effectively, so we ignore them.

Results

Research Question 1:

For question 1, we use Altair Library to present the interactive graph and plot each NBA Player's Career Average Points and their Corresponding Salary in the graph. According to the graph, we found the Career Average Points are positively correlated to their salary. However, we found some special points (NBA player) with fairly high career average scores and a comparatively low salary.

Therefore, we decided to have depth research on the special points (NBA players).

Depended on the tooltips function in Altair Library, we found the players' draft year is much earlier than others who has similar career average score (i.e., Michael Jordan[1] joined NBA earlier than Kobe Bryant[2]) due to the NBA Salary Cap at a specific period.

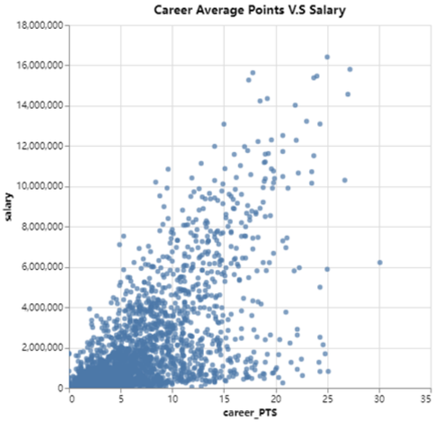


Fig 1: Career Average Point V.S Carrer Average Salary

Depending on the increasing salary cap in the past 40 years, we redesign the method and plot a graph based on the percent of single player's salary in the salary cap in that period.

We found that the better the scoring ability, the higher salary in their period, which also showed that NBA focused and offered more salary to offensive players.

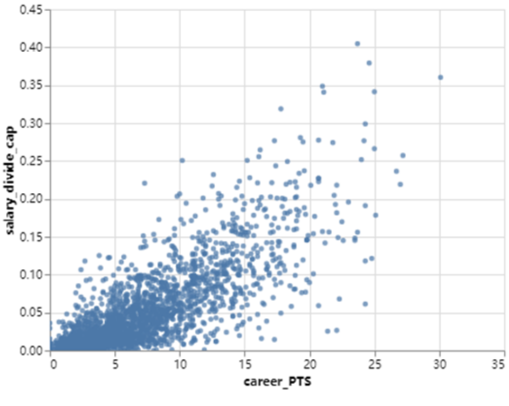


Fig 2: Percent of Salary in Salary Cap V.S Career Average Score

[1] American businessman and former professional basketball player, and his draft year: 1984
[2] American professional basketball player, and his draft year: 1996

Research Question 2:

For question 2, we want to find the relationship between some attributes with salary. So we consider three considerable attributes for a basketball player: height, scoring ability, and 3-point shooting ability. We found the height in the NBA will not affect the salary significantly unless the player is much lower (less than 175cm) than the majority of players. Besides, combining with question 1, the scoring ability will affect the salary seriously. Most of the time, the stronger the scoring ability, the higher salary in their period.

What is surprising is that the average salary increases significantly when 3-point shooting ability is strong. So, we will recommend potential NBA players practice their 3-point shooting ability to pursue higher salary.

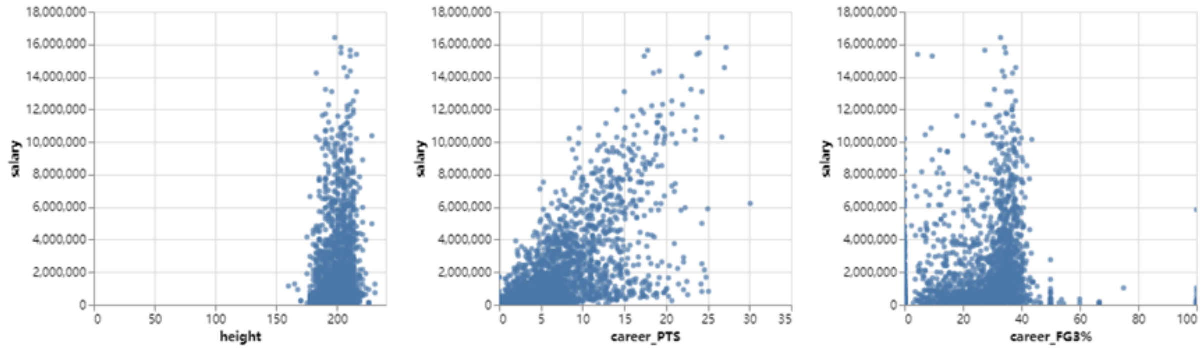
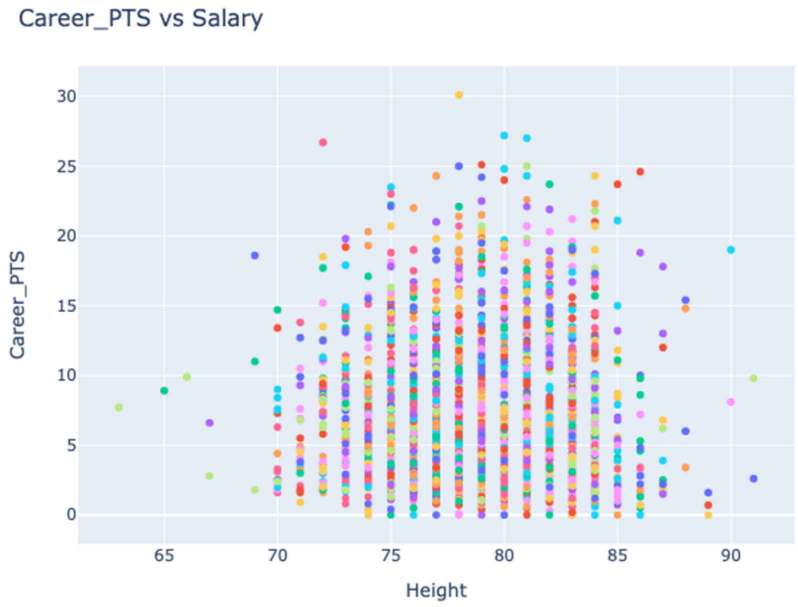


Fig 3: Height/ Career Average Points/ 3-points Ability V.S Salary

We still consider the percentage of a player's salary in the salary cap of the team in that year. However, the percentage is scattered broadly over the graph which is meaningless to check the trend.

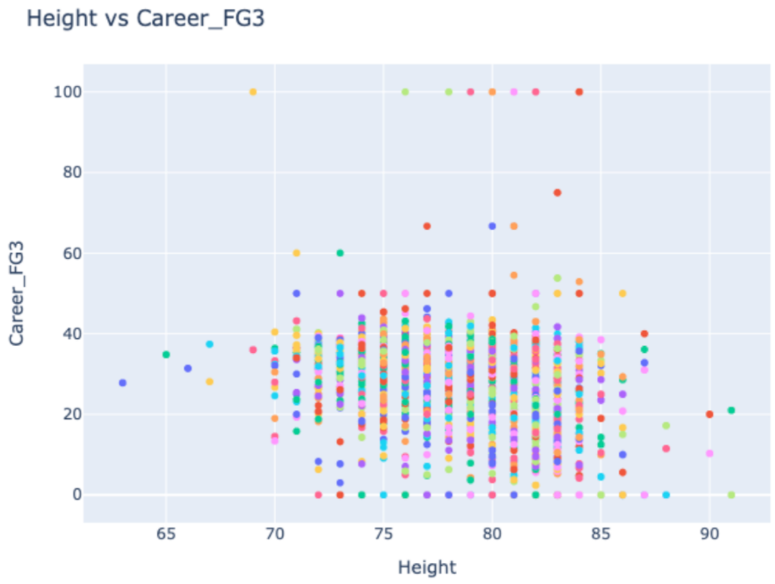
Research Question 3:

Below is the scatter plot we got with height as the independent variable and field goal percentage as the dependent variable. The overall direction of this plot is random, which means there is no overall direction for location of points.

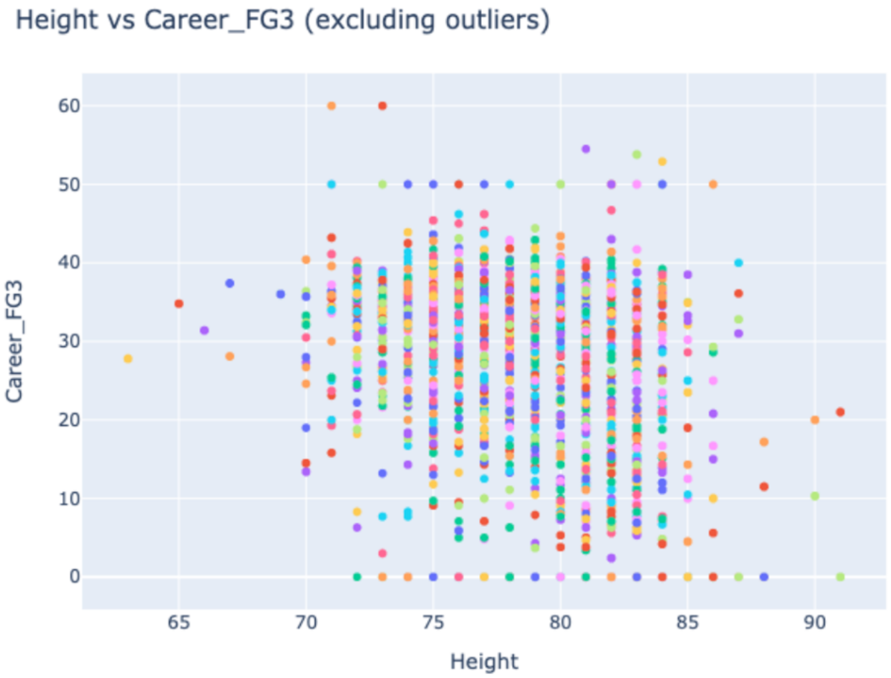


We can therefore draw the preliminary conclusion that there is no strong correlation between these two variables. To further verify our conjecture, we need to go further and calculate the r square value. We calculated the r-square value between the two to be 0.0059. This is a value that converges infinitely to 0. More specifically, the r-square value obtained of 0.0059 suggests that merely 0.59% of the variability in career_PTS can be clarified by height variable when using linear regression model. This result suggests that height by itself is not a reliable predictor of career_PTS. Therefore, we can conclude that height does not have a significant impact on shooting percentage and there is little correlation between the two.

Below is the scatter plot we got with height as the independent variable and three-point shooting percentage as the dependent variable.

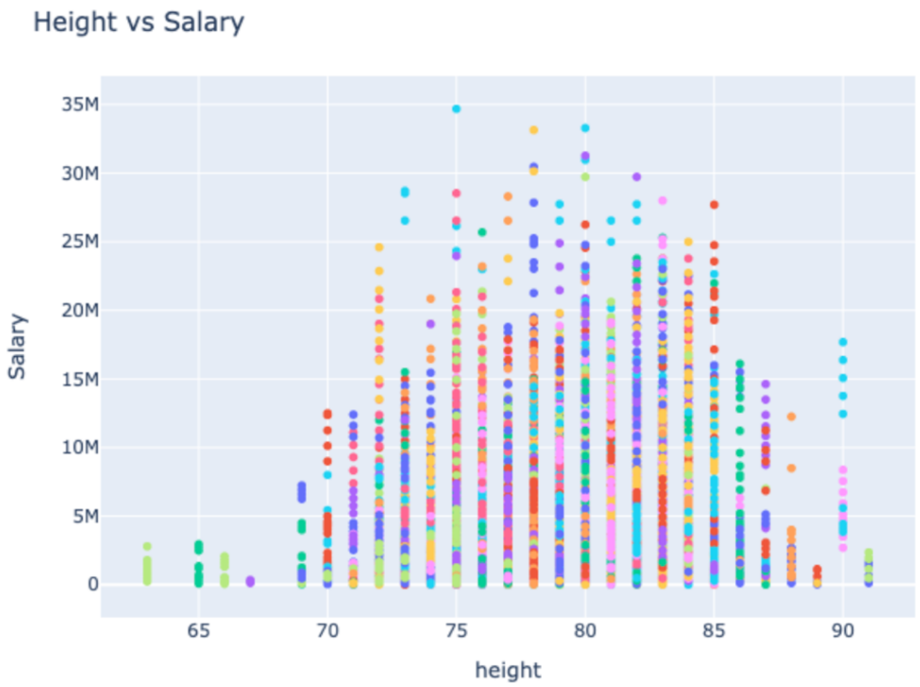


Following the same process as we did in the first comparison, we calculated the r-square value between the two to be 0.18. This is much higher compared to the r-square value for the comparison between height and field goal percentage, which means the correlation between three-point shooting and height is greater. However, if looking closely at the above diagram, we can see the presence of very obvious outliers. The presence of these things may affect our judgment about the correlation. We therefore decided to eliminate the potential influence of outliers on our results by dividing the quantile. The following plot is the new one we get by dividing the quantile.

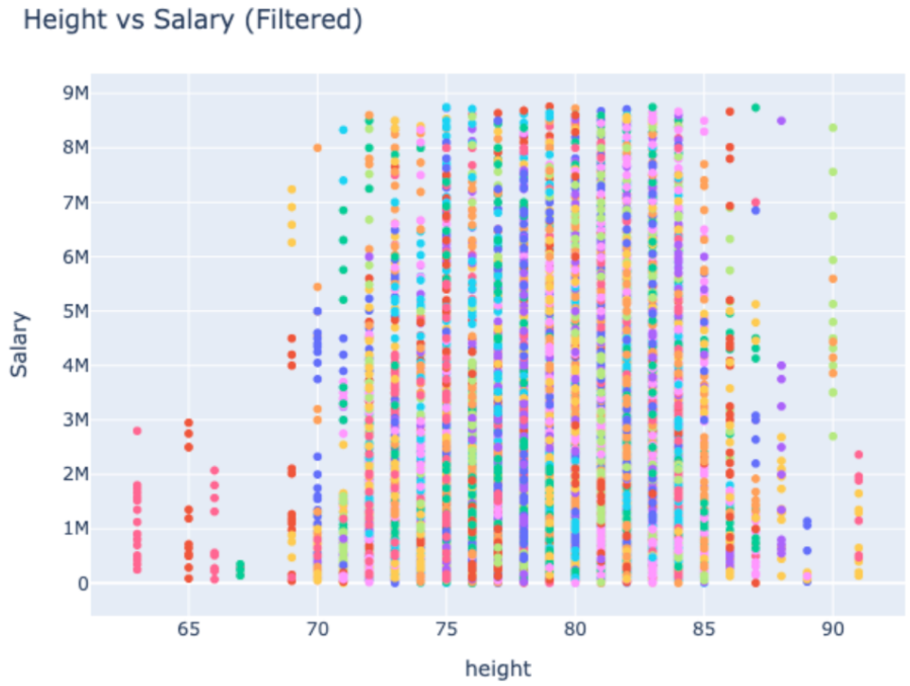


We recalculate the r-square value and get 0.20. This indicates that approximately 20% of the career three-point shooting percentage can be explained by players' heights in the regression model. It is indeed a very big improvement compared to the first one. This goes to show that height has an effect on a player's performance. However this data is not sufficient for the r-square value to indicate that there is a very good fit between these two variables. In other words, the regression model is not a very good fit for the data as there is a lot of unexplained variation. Therefore, we can conclude that height has an objective effect on player performance, but it is not a decisive factor in determining a player's performance.

Below is the scatter plot we got with height as the independent variable and salary as the dependent variable.



It’s also clear that the overall direction of this plot is random, which means there is no overall direction for location of points. In addition, we get the r-square value equal to 0.0084, which is a very small value for r-square value and means only 0.84% of the variance in salary can be explained by the height variable using a linear regression model. However, outliers still exist in the plot, which may lead to some confusion and disagreement. In order to eliminate the bias for this conclusion, we decide to reduce and filter out outliers. Below is the new one we get by dividing the quantile.



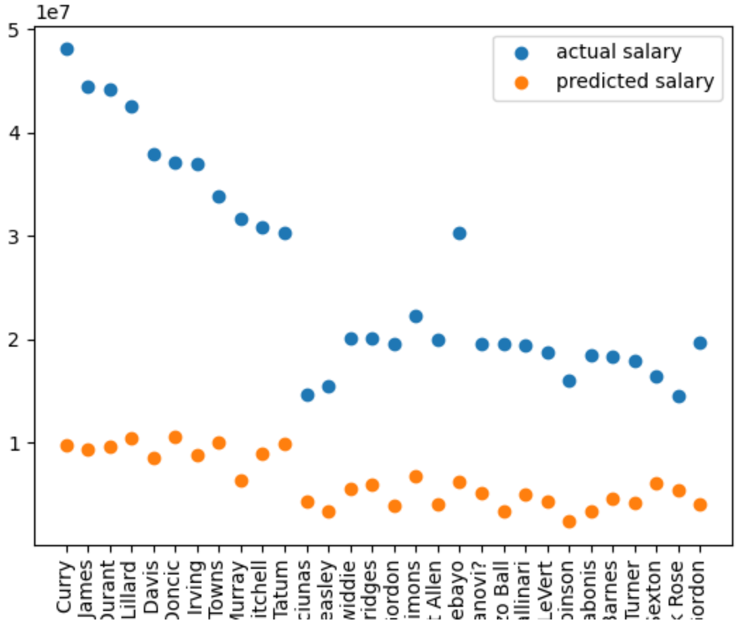
For this one, we get a lower r-square value: 0.006, which reinforces our conclusion: height alone is not a good predictor of salary. In other words, there exists other significant factors that contribute to a player's salaries. For example, a player who is shorter than average but has exceptional skills and a track record of strong performance may command a higher salary than a taller player who is less skilled or less experienced. Additionally, factors such as team budget and market demand may also play a role in determining player salaries. In addition, NBA player salaries are typically determined through intricate contract negotiations involving players, agents, and team owners. These negotiations are influenced by various factors, not solely limited to the player's height.

In conclusion, although height has an objective effect on a player's performance on the court, height is not a determining factor in a player's performance on the court, and it is not enough to affect a player's salary.

Research Question 4:

Before using salary cap:

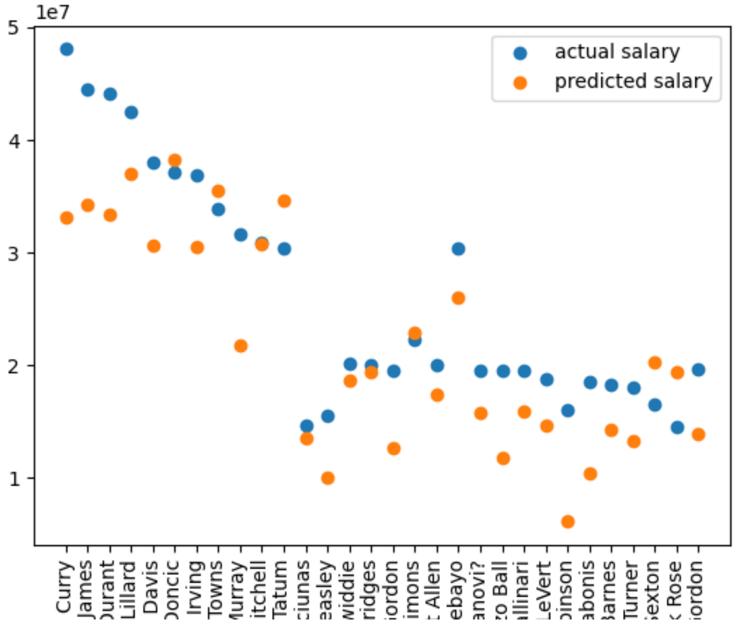
Our predict salaries for the thirty NBA players are:



And then, to check the validity of our result, we compute the relative error and the percentage of relative error less than 50%.
The percentage is 0.
The R-squared value is about 0.4.

After using salary cap:

Our predict salaries for the thirty NBA players are:



And then, to check the validity of our result,
we compute the relative error and the percentage of relative error less than 50%.
The percentage is 83.34%.
The R-squared value is about 0.6.

An R-squared value of 0.6 indicates that 60% of the variance in the dependent variable can be explained by the independent variables included in the model. While this may seem low, it's important to keep in mind that the significance of an R-squared value depends on the context of the problem and the domain knowledge. In summary, while an R-squared value of 0.6 is not extremely high, it's still a reasonable result and further investigation is needed to determine whether it is acceptable for the specific problem at hand.

Impact and Limitations

The potential impact is that our study may lead to future basketball players focusing more on long-range shooting ability and neglecting the development of assists, steals, and other abilities.

NBA scouts may benefit from our research, as it will give scouts a reference for player ability and reduce the workload when scouting future stars.

Players who are more basketball-oriented (Center) may be hurt by the lack of scoring ability or three-point shooting ability and be offered a lower salary.

The data of player’s salary is from 1985 to 2018, nowadays, the average salary is much higher than 30 years ago. Therefore, the potential bias in the dataset is that the older generation basketball has lower salary since they have poor ability in basketball compared with young generation.

However, we redesigned the algorithms which evaluate the basketball players’ ability (percent of their salary in the salary cap), and then predict the potential salary depended on the salary cap in the future.

The limitation of this analysis is that the dataset is old for nowadays NBA players, and consider too much hard ability (scoring ability) and neglect other ‘soft’ ability (i.e., collaboration ability) and defending ability. However, I still proposed that the analysis should be used, since this conclusion could help the NBA find more potential stars in the future.

Challenge Goals

1. New Library

We choose to use two new visualization libraries: altair and plotly in our visualization part.

Altair is a declarative visualization library in Python that allows you to create interactive visualizations with just a few lines of code. Altair allows you to create a wide range of visualizations, including scatter plots, line charts, area charts, bar charts, and more. It also supports interactive features such as zooming, panning, and tooltip information. The property of interactive helps us find the special points in question 1 and then build a more reasonable algorithm for machine learning by solving the special points problem.

Plotly is a Python library for creating interactive, publication-quality visualizations in web browsers. Plotly allows you to create interactive visualizations with hover effects, zooming, and panning. You can also add annotations, text, and images to your plots to provide additional context and information. In question 3, we used Plotly to act as the visualization library to plot and visualize relationships between different variables.

2. Result Validity

This challenge goal we have divided into two parts. The first part is that we add the component of statistical analysis to the method, for example, we use the method in Q3 in combination with plotly about using the division of different quantile to reduce the potential influence of outliers on the results and by calculating the r-square of the regression value of the regression to make our conclusion more convincing. In Q4, we use relative error and R-squared value to verify the result validity.

Work Plan and Evaluations:

Data Management (5 hours)

Randy and Kiri will deal with our chosen datasets (filter out our needed columns, delete content or elements involving “bad input” like NaN or None, etc) and clean and organize the data to ensure it is accurate and suitable for analysis.

John will do some extra research to better understand the background of our research. More specifically, John will read through some past research or even materials like magazines or newspapers to see if there exists some other potential factors that influence the players’ salaries. We will consider these potential influential factors when dealing with “Result Validity" later.

Evaluation:

In general, the team's strategy for conducting data cleaning and research is comprehensive and worked well.

1. **Data Analysis** (10 hours)

Randy, Kiri, and John will work together and analyze the filtered data. In detail, Randy, Kiri, and John will dig into the filtered datasets and use techniques taught in class (pandas, machine learning, etc) to analyze relationships between factors identified in the research questions.

Kiri and John will test the accuracy of the data and the validity of the analysis through setting extra tests.

Evaluation:

The team met some problems when filtering data, but finally worked it out successfully. However, this process took the team more time than expected: about 12 hours.

2. **Data Interpretation** (10 hours)

Randy, Kiri, and John will use the analysis done in the second part to analyze and get conclusions about the relationship between NBA players' salaries and each of these factors identified in the research questions.

Evaluation:

The team did surprisingly fast and smoothly in this process: it only took the team about 5 hours to finish and wrapped up all the tasks.

3. **Result Interpretation** (5 hours)

Randy will combine the conclusions we got in the third part, while Kiri and John will test the result validity through use of our conclusion (model) on other datasets (also NBA players' salaries) to see the accuracy of our conclusion (model).

Evaluation:

As with the previous task, the team completed the task more efficiently and in less time: about 3 hours.

4. **Report and Presentation** (6 hours)

Randy, Kiri, and John will work together for our final report (we will divide our work through dividing this report into parts like introduction and conclusion with the same workload) and our final presentation (online, so we need to record for each part).

Evaluation:

The team did a great job in this part and basically followed the expected time.

General Evaluation:

The team did a great job in this project!

Testing

Q4:

Crawl data from other two websites for new data in 2023.

Use the model to predict salaries of players in new datasets and compare the predicted results with actual salaries.

We computed relative error for each predicted result and gave a limit of 50% relative error, which means those results have relative error less than 50% is considered valid. And then finally got the accuracy of 51.7%.

Collaboration

Q1 and Q2:

https://www.basketball-reference.com/

Salary Cap for Different Year:
https://basketball.realgm.com/nba/info/salary_cap

Q3:

https://github.com/plotly/plotly.py
https://plotly.com/python/plotly-express/
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.query.html
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rename.html
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.quantile.html
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
https://pandas.pydata.org/docs/reference/api/pandas.to_numeric.html

Q4:

https://www.w3schools.com/python/python_ml_scatterplot.asp
http://www.espn.com/nba/salaries
https://www.basketball-reference.com/leagues/NBA_2023_totals.html