Visualization Basketball Salaries

Brandon Onyejekwe, Conghan Liu, Adrian Monaghan, Sarah Costa, Mengting Tang
Northeastern University

Abstract

Visualizations depicting statistics from NBA players are necessary to help fans better understand the game and its players. For our project, we decided to take a look at statistical data we have collected from the National Basketball Association that will help us with our analysis of NBA players. We plan to visualize the salaries of various players in order to make comparisons and draw inferences based on their performance, and the ultimate goal is to determine whether a specified player's worth is reflected through their salaries.

https://github.com/DS4200-S22/final-project-visualization-basketball-salaries

1. Introduction

The topic of this project is to predict the salary of an NBA player according to their performance and background information, including age, position and team.

With the economic growth and the increase of the salary cap of the National Basketball Association, NBA players earn more money now than ever. At the same time, the differences in salaries and skills among players also continues to increase. As a result, our group wants to discover the relationship between NBA players' salary and their performance. At the same time, we also want to find whether or not background information (age, position and team) influences salary, and if so, how.

We want to use average points per game, the percentage of the minutes played in the game, rebounds per game, and other common basketball statistics to identify players' performance. We will also use age and position to identify background information and see how that plays a role in salary. Then, by comparing the dependent variables above with the corresponding NBA player's salary, we will try to discover any relationships that may exist between a player's salary and their performance.

Another goal of this project is to help the audience of the NBA have a better view of the performance differences between players. Where now, to compare players, a user would have to look at raw stats to have an idea of who is a better player. Our visualization will allow users to compare players' performance throughout a season visually.

2. Related Work

Our sources are all related, dealing with general sports data visualization and how it should be reflected to best show information relating to the sport.

One reference helps to highlight the difficult nature of sports data, stating that it "tends to be hypervariate, temporal, relational, hierarchical or a combination thereof."[5] This will be something that we will keep in mind when we create our visualizations which may be challenging for these reasons.

Another[1] aims to explore the usefulness of allowing users to control big picture ideas when it comes to sports visualizations, and then allows them to gain further detailed insights from there.

From the third reference, it gives us the ideas about how the salary of the NBA is built and some regression models of player's performance has been calculated. The statistic gives us some idea about the player's performance, and what we are going to face when we are working with our data.[3]

The fourth reference describes a novel visualization tool that was designed to be used by coaching staff during the game to allow for quicker human analysis of the data and make informed decisions. The teams used "glyph-based visual design … to visualize actions and events at a glance"[4]. We can use these insights gained by the researchers to allow our users to view our visualization and draw conclusions faster than with regular visualizations such as bar charts and scatter plots.

A final reference describes how to imply interactivity on a visualization in a website. The article talks about design cues that can be implemented that can allude to the presence of interactivity within the visualization[2].

3. Use Case

Using the visualization tool, D3, we aim to provide a visualization tool for the fans of the NBA to see how different positions perform in relation to their salary. The visualization tool will also allow fans to compare two players' performance and salaries throughout multiple seasons in the NBA.

4. Data

The data comes from two main websites. The first website contains the performance data in 23 categories

of all players in the NBA throughout three seasons: 2017-2018, 2018-2019, 2019-2020. The second data set contains the salary of each player in each season. We used Pandas, a Python library, to merge all the data into a single csv table, with each player's performance and their salary for each season.

The raw data in itself would not have any bias as it is reporting factual statistics about all players in the NBA. However, in order to simplify the visualization and data, we did not include players who did not play in all three of the seasons we were interested in. This could lead to some bias in the data as those players who may have only played for one or two seasons would not be represented in the dataset. Players who retired in one of these three seasons would also not be included in the dataset. For example, a player who might have had a good season, but had to be let go due to an injury would not be represented.

The links to the data are below:

- https://www.basketball-reference.com/leagues/NBA_2021.html
- https://www.nbastuffer.com/2017-2018-nba-player-stats/

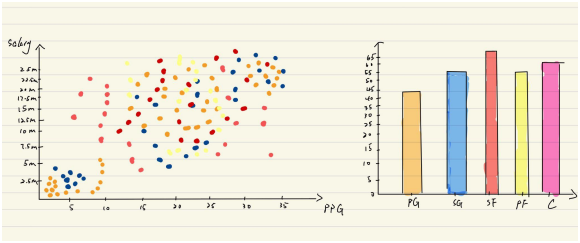
5. Task Analysis

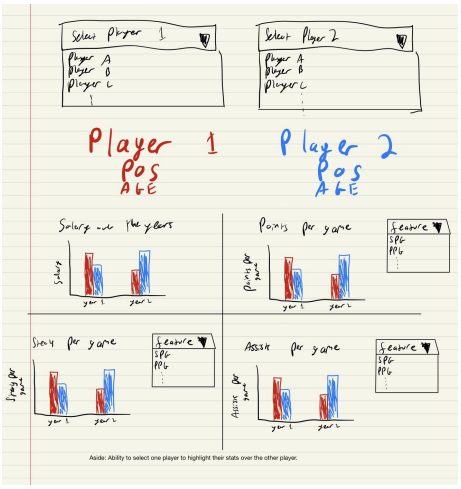
Task ID #	Domain Task	Analyze Task (high-level)	Search Task (mid-level)	Analytic Task (low-level, "query")
1	I want to show my friends why my favorite player is better than their favorite player	Consume -> present	Search -> lookup	Query -> compare
2	I want to make the best decision for my fantasy basketball team	Consume -> discover	Search -> Explore	Query -> Identify
3	I want to see trends of players over seasons	Consume -> discover	search-> lookup	Query -> summarize

The primary consumer of our visualization will be a basketball enthusiast who enjoys looking at data and wants to draw quick conclusions from it. We plan on primarily developing our visualization to easily discover information from our dataset. We think this is important because two out of the three tasks desired by the primary consumer is to discover information from the dataset. Comparison is the most important domain task desired by the user, however, we believe that a good discovery foundation will also allow for an easy comparison of players, even if we are unable to implement this in the final visualization.

6. Design Process

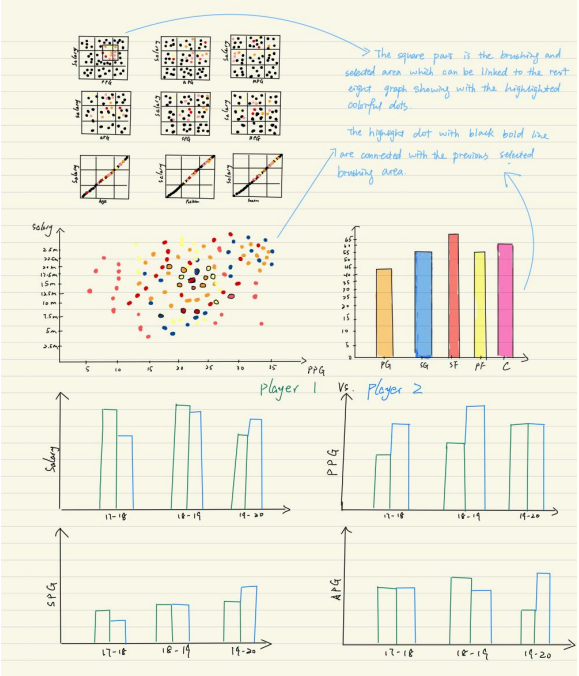
In designing our visualization, we started with multiple hand-drawn sketches that would help us visualize the performance and salaries of NBA players in different ways. We were most drawn to the ideas of having multiple scatterplots show the different statistics for each player, as well as one showing a comparison between different positions, and finally a direct comparison between two players. Our first implementation has a 3x3 matrix of scatterplots. These will allow a user to compare 9 different statistics for all players. Then, there are two large plots side by side: one of which is a bar chart showing the counts of all positions. The other is a scatter plot with all players showing a user specified stat. Finally, at the bottom, there are four sets of paired bar charts that will allow a userto compare four stats for two players.





These are some sketches from our initial design. The first of which is a scatter plot paired with a bar chart comparing the points per game to salary for each of the five positions. The second is a direct comparison over the years between the two players specified from the drop-down

menu, comparing for salary, points, steals, and assists per game.



Our final implementation utilizes 7 scatterplots (formatted as a 3 x 2 + 1) rather than a 3 x 3 scatterplot matrix as seen above. We felt that the remaining two categories did not give us enough information in their plots to justify keeping them as we wanted to focus on those that are most important. We also decided to add the ability to click on a bar in the bar chart and select those points in the adjacent scatterplot, as this would allow for a filter operation that lets the user only focus on one or a few positions at a time. We combined some of our initial sketches as we felt by combining the grid of scatter plots, the paired scatter and bar plot, and the grid of bar plots, we were best able to fulfill our users' tasks.

For our usability testing, we wanted to test the logic of our site, the extent to which our design was easy to use and intuitive, and the ability to compare two players' stats. We found that the tester found most of our site to be very intuitive and easy to use, but struggled with our third task of figuring out how best to compare players with each other. To resolve this, we decided to expand the drop-down menu options to make it easier to select players to compare. We also added the ability to click one of the seven scatter plots to focus in the lower section to allow the user to see correlations between different stats and salary in more detail. We also expanded the number of position categories in our dataset. Originally, some players played multiple positions which was reflected in the data. We decided to make it so those who played multiple positions were shown to play the position they played most often. For example, if a player played mostly guard and a little forward, their position was edited in our dataset to only be guard. Through our usability testing though, we found that this did not give enough granularity in the information presented. The user would have liked to have more positions represented such as point guard and power forward, for example. Therefore, we reverted to the original position data.

7. Visualization Design

The final visualization design is: at the top of the visualization is a grid of seven scatter plots. Each scatter plot shows the correlation between a different statistic and the player's salary. For example, one scatter plot will show the correlation between the average points per game over a season versus salary, and then in another scatter plot, the average assists per game over a season versus salary and so forth. This will allow the user to get a good idea of how different positions perform with regard to how much they get paid. The user will be able to select a scatter plot in the grid of nine scatter plots with the mouse, and that individual scatter plot will show separately below the other scatter plots. Next to the separate scatter plot will be a bar plot. The bar plot will show a distribution of player positions. The user will be able to brush over the larger and lower scatter plot. The points' positions selected in the scatter plot will highlight in the barchart of position distributions. For example, if the user brushes points that are played by guards, the guards bar in the bar plot will highlight. The goal of this part of the visualization with the larger scatter plot and bar plot is to allow the user to get a better idea of the distribution of positions that are selected. It also gives the user a more accurate correlation between the stat and the salary. Finally, at the very bottom of the page will be a place for two drop down menus where the user can select two players. The user will then be able to select four different stats, of a player, such as points per game, blocks per game or assists per game, and then a comparative bar plot will be shown that compares the statistics of the two different players over the three seasons of data we have. These two players selected will be highlighted in the corresponding scatter plots as well so the user is able to get an idea of the players performance in comparison to the rest of the league.

Since we decided to create scatter plots and bar charts, our marks are points for the scatter plots and bars for the bar charts. We chose to use scatter plots and bar charts because a lot of our data is quantitative and we typically only needed to show one qualitative feature at a time. Our channels for the scatter plots were horizontal and vertical position and color. We used vertical position to represent the salary because this is a continuous quantitative figure. This is the same for all seven of our scatter plots. We used horizontal position to represent the other features that change depending on the scatter plot

which include APG, PPG, age, MPG, BPG, SPG and RPG. These are also mostly continuous quantitative values which is why we chose to represent them with horizontal position. Lastly for the scatter plot, we used color to represent position because this is our only qualitative categorical feature that we are plotting in the scatter plots. SImilarly, our first bar chart uses vertical position and color to communicate our data. We use color to represent position once again because it is our only qualitative, categorical feature but for this chart we use vertical position to represent the number of players since this feature is measured quantitatively and this is the only other feature of the graph. Lastly for the bottom group of bar charts we once again use horizontal and vertical position as well as color to show our data. In this case, we used vertical position once again to show salary but also PPG, SPG, and APG making for a total of four bar charts in this last part of the visualization where we compare two players' statistics. Horizontal position here shows the three different seasons for which we have data which makes sense since this is a discrete number of ordinal categories. Finally, color here represents the player the statistics represent since this is a qualitative, categorical feature.

8. Discussion

The original database contained twenty three different statistics on each player. We decided against using all of them as we determined it would add unnecessary clutter to the visualization. We chose seven of what we deemed to be the most important out of the original twenty three. Then we created seven scatter plots by using those seven statistics and the salary to get an idea of the correlation between the salary and a given statistic. It is fairly clear to see that in general, the more points, steals, and rebounds you have per game, the more you get paid. Surprisingly, there does not seem to be much correlation between the assists per game and pay. It also seems that correlation between minutes played per game and salary seems almost exponential. Salary does not seem to really grow until players start playing a significant amount of the game, in general.

We created the bar chart showing the counts of players in different positions and allowed the user to filter which positions are shown in the corresponding scatter plot. In doing so, we found that players who were guards in general made significantly more than those of the other nine positions, most notably getting paid significantly more than centers and forwards.

During the design of visualization, at first we were going to use all the players from three seasons. However, while we were working on the grid of bar plots to compare two players, we found that if we wanted to compare two players, we needed to ensure that both players had played for three seasons. Therefore, we had to remove players from our dataset that did not play in all three seasons.

At the beginning of the design process, we sketched a significant number of visualizations. We chose what deemed necessary to complete the project and complete our users' tasks. However, while we were creating the visualization in D3 and performing some of our testing and use, we found that some of the original sketches were not intuitive to use or did not portray the information we originally set out to show. Therefore, we changed a lot from the original sketch while creating the visualization tool. It showed us, in practice, how iterative the design and creation of a visualization tool has to be in order to produce an effective tool.

Some limitations of our visualization tool is that we only have data for three seasons of players' performance. If we had more data over more seasons we could give the users more options to view their players or maybe how salary and different statistics changed over time.

9. Conclusion

In the end we created our own visualization that is able to be used to look at different groups of players across the NBA and how they perform on different statistics in regards to their salary. We were also able to create the visualization that allows users to be able to compare two players in the NBA.

We first started brainstorming the idea of wanting to visualize different basketball players and their salary as well as wanting to compare the performance of different

players. We started brainstorming and sketching different visualizations that we could feasibly create that would allow these tasks to be completed. We started the data collection through different websites and data manipulation through Python libraries such as Pandas. We were then able to start our visualization creation in d3. After the initial creation of the visualization, we performed a usability test to try to find issues and room for improvement. There were a few minor issues found with our visualization, however, nothing fundamental needed to be changed.

We did run into several issues when trying to implement our visualizations that forced us to sacrifice some functionality or form. However, we were able to still complete a majority of the tasks we originally set out to complete. This leaves us with plenty of room for future improvement. Ideally, when we click on a new scatter plot to show, we would like to animate this change instead of having it immediately pop up. We would also like to allow the user to zoom in to different parts of the main scatter plot. We found that it is difficult to see details in cluttered parts of the scatter plot. When the user is comparing two players on the lower portion of visualization we would like the players selected to highlight in the scatter plots to see how they compare to the rest of the NBA. Along with this, currently on the player comparison bar charts, the user is only able to compare players from a selection of about 30 players in the NBA. We would like to add the functionality to allow the user to select players from the entirety of the league. Along with this, we would also like the user to be able to compare more than two players at once.

10. Acknowledgements

We would like to acknowledge the work of the creators at the d3.js graph gallery. The code provided there for the visualizations was vital for the completion of our project. We would also like to acknowledge the exercises and assignments we had to complete for Professor Mosca's class. The assignments and in class activities allowed each of us to learn how to do basic visualizations in d3 which gave us a good foundation to start writing our own visualizations in d3.

References

1. *R. Vuillemot and C. Perin, "Sports Tournament Predictions Using Direct Manipulation," in IEEE Computer Graphics and Applications, vol. 36, no. 5, pp. 62-71, Sept.-Oct. 2016, doi: 10.1109/MCG.2016.90.*
2. J. Boy, L. Eveillard, F. Detienne and J. Fekete, "Suggested Interactivity: Seeking Perceived Affordances for Information Visualization," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 639-648, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2467201.
3. Sigler , K., & Compton, W. (2018, June 19). *NBA players' pay and performance: What counts?* The Sport Journal. Retrieved February 23, 2022, from https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/
4. Legg, P.A., Chung, D.H.S., Parry, M.L., Jones, M.W., Long, R., Griffiths, I.W. and Chen, M. (2012), *MatchPad*: Interactive Glyph-Based Visualization for Real-Time Sports Performance Analysis. Computer Graphics Forum, 31: 1255-1264. https://doi.org/10.1111/j.1467-8659.2012.03118.x
5. Basole, R. C., & Saupe, D. (2016, September). *CSDL: IEEE Computer Society*. CSDL | IEEE Computer Society. Retrieved February 23, 2022, from https://www.computer.org/csdl/magazine/cg/2016/05/mcg2016050024/13rRUxNmPIy

Appendix A: Group Charter

Group Purpose:

The reason for the group's formation is to complete the final project for DS4200. We all plan to make visualizations about basketball player salaries, and the purpose of the group is to complete the project well and on time, creating multiple distinct visualizations about basketball player statistics.

Group Goals:

Our group's project goals are to make the best visualization and report in the time given to us. We strive for a high quality of visualization. Everyone in the group is willing to put in enough effort to achieve an A- on this final project.

Group Roles and Responsibilities:

Responsibilities have been decided to be flexible and all members of the group will share different roles. These roles are preliminary and subject to change throughout the project.

Adrian Monaghan - Communicator
Brandon Onyejekwe - Documentation Coordinator
Conghan Liu - Information Manager
Mengting Tang - Meeting Facilitator
Sarah Costa - Group Leader

Ground Rules:

Our group will meet every Tuesday to work on the final project. We will agree to arrive at the place within 10 minutes of the meeting time without at least an hour of heads up to the rest of the group. We will make decisions by going with what the majority of the group members agree to.

We will handle dissenting views on a case by case basis. Those that are deemed not as important will be handled by the majority. For the dissenting views that are more important, we will make our best effort to come to a compromise on the decision, and if that still does not work we will seek advice and support from the teaching staff. We will hold each other accountable for this by airing our grievances with the team member that is causing the issue. If this still does not work we will seek advice and support from the teaching staff in how to further handle the member. We expect an equal level of commitment and participation from all team members.

Potential Barriers and Coping Strategies:

One possible barrier we may face is having all of us meet together at the same time to work on the project, as we all have different schedules that may not align together perfectly. To fix this, we plan to utilize a lot of communication and plan our work out effectively. In the past, members of our group have expressed the common feeling of having some people in a group not carry the same weight as others, and to combat this, we plan to be organized in terms of determining who is doing what, and making sure that everyone is on track with what needs to be accomplished. We anticipate that we will be able to work together effectively to do well on this project.

Review About Group Charter:

Our group members are all satisfied with the group roles we identified when we first created the group. We meet every Tuesday to work on our project milestones. In addition to meetings, we actively discuss the availability of meeting materials and meeting times in group chats. Additionally, all team members are well prepared for the team meeting. All of us have abided by what we have agreed in the beginning, and feel comfortable about our roles in the group.

Positive Thing About Group:

Each member is responsible and well-prepared for the group work. Every time group members are creative about our project idea.

Each group member does a great job of showing up to meetings on time and are responsive to messages that are sent out.

Each group member provides thoughtful insights to discussion when we have to make decisions for the project. Each group member comes to each meeting prepared.

People help each other to work on the pm to make sure we have everything done perfectly.

Everyone does a great job contributing to our conversation with their ideas. Each group member is prepared to help out with the PM and wants to help us do the best job possible.

Appendix B: Data Exploration

Initially we decided to remove certain features from the original Data set as they deemed not interesting or not relevant to the scope of this project.

Data Review

Full Name, Team, Position, Age, Points per Game (PPG), Assists per Game (APG), Minutes per Game (MPG), Rebounds per Game (RPG), Steals per Game (SPG), Blocks per Game (BPG), Season, Salary

The data Type for this database is Attribute

Full Name(Levels are Stephen Curry, CJ McCollum, DeMar DeRozan…), Team(Levels are Dal, Tor, Was…), and Position(Levels are PG, CF, SG, C…) are Categorical.

Season is Ordinal. The levels of the Season are 2017-2018, 2018-2019, and 2019-2020.

Age, Points per Game (PPG), Assists per Game (APG), Minutes per Game (MPG), Rebounds per Game (RPG), Steals per Game (SPG), Blocks per Game (BPG), Salary are Quantitative

There are 949 points in the data set. For age, the average is 26.9 years, the youngest player is 19.42 years old, the max is 43.55. The median age is 26.44 with a standard deviation of 4.19. The average points per game is 9.9 with a min of 0 and a max of 36.1. The median points per game is 8.2 and the standard deviation is 6.47. The average assists per game is 2.17, with a min of 0 and a max of 10.7. The median assists per game is 1.5 with a standard deviation of 1.86. The average minutes per game was 21.54 with a minimum of 1.0 and a maximum of 37.5. The median minutes per game is 21.7 with a standard deviation of 8.73. The average rebounds per game is 4.09 with a minimum of 0 and a maximum of 16. The median rebounds per game is 3.6 with a standard deviation of 2.62. The average steals per game is 0.68 with a minimum of 0 and a maximum of 2.35. The median steals per game is 0.6 with a standard deviation of 0.41. The average blocks per game is 0.45 with a minimum of 0 and a maximum of 2.93. The median blocks per game is 0.33 with a standard deviation of 0.44. The average salary is $8,080,000 with a maximum of $40,023,000 and a minimum of $4,737. The median salary is $3,628,000 with a standard deviation of $9,029,000. The only ordinal data feature in the data set is

the season, which ranges from the years 2017-2018 to 2019-2020, a range of three seasons. The distribution numbers and statistics are shown below in the screen shot. In summary, we can see that the data's distribution varies widely for all features.

	AGE	PPG	APG	MPG	RPG	SPG	BPG	Salar
count	949.000000	949.000000	949.000000	949.000000	949.000000	949.000000	949.000000	9.490000e+0
mean	26.908272	9.904426	2.167439	21.542044	4.089884	0.679947	0.449726	8.080541e+0
std	4.186631	6.467104	1.864244	8.732631	2.619084	0.413541	0.435216	9.029039e+0
min	19.420000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	4.737000e+0
25%	23.520000	5.000000	0.900000	14.500000	2.300000	0.380000	0.160000	1.599712e+0
50%	26.440000	8.200000	1.500000	21.700000	3.600000	0.600000	0.330000	3.627842e+0
75%	29.670000	13.500000	2.900000	28.900000	5.200000	0.900000	0.570000	1.219512e+0
max	43.550000	36.100000	10.700000	37.500000	16.000000	2.350000	2.930000	4.023176e+0

The distribution of data on a year by year basis is shown below. We can see that the average salary was greatest 2017-2018 and the lowest average salary being the 2018-2019 season.

season	AGE	PPG	APG	MPG	RPG	SPG	BPG	Salary
2017-2018	27.875000	17.957500	3.580000	31.390000	6.897500	1.032250	0.805000	2.493823e+07
2018-2019	26.833203	9.434423	2.099564	20.927015	3.933769	0.657495	0.431351	6.989084e+06
2019-2020	26.898911	9.668000	2.111111	21.294000	3.999556	0.671533	0.436889	7.695367e+06

Data Issues:

- Our data for both the salary and the season were each split between two separate columns. To solve this, we merged them into a single column for each feature.
- The salaries were also kept in an inconsistent manner. Some of them were stored as floating point numbers, while others were strings with the dollar sign symbol and were comma-delimited. We decided to convert everything into floating point numbers to help with analysis.
- Some of the columns, such as name and position, were repeated twice. These also needed to be merged into single feature columns.

Insights

With the rules and the guidelines of the NBA, players can score more easily than before. In that case, players need to get more and more points to prove themselves to get more salaries. There is a golden age for players in which they will earn the most salary during their career.

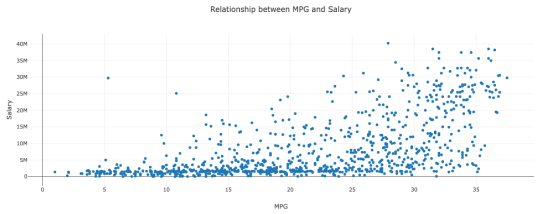
Also, another trend we found is salaries also vary from the positions of the players. Different positions may result in different salaries. Most NBA teams will pay the point guards the most salary than any position. So position is also an important factor that will influence the salary based on our insights.

To our surprise, there is one player who can get 29.6 points per game, but he only earns 6.273 M a year. It shows that if a player's performance is worth more than his salary, then there is also a player whose performance is not much to his salary.

The initial problem we still have may be in position. Since there are only five positions in the basketball rules, we got more than five positions in the Position. We need to fix some player's positions to make it a single position.

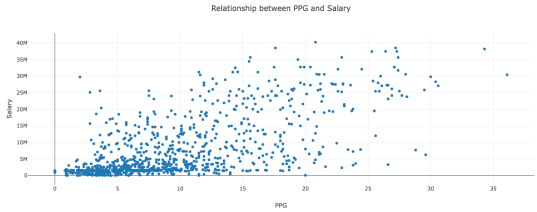
Screen Shots

1. MPG vs. Salary:



The first visualization we made is a scatter plot that shows the relationship between how many minutes per game a player plays and what their salary is. We did this using two features of this dataset: MPG and salary. We used a scatter plot because we wanted to show each individual players' MPG and salary yet also show this relationship as a whole to see if there was a positive or negative correlation between minutes played per game and salary. In general, it seemed like there was a positive correlation between MPG and salary: the more time a player spent playing per game, the more money they earned.

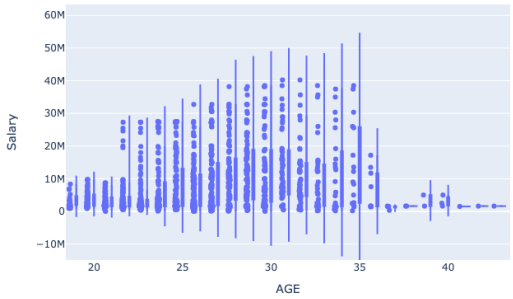
2. PPG vs. Salary:



We also made a scatter plot to show the relationship between points per game a player scores and their salary. We did this using two features of the dataset: PPG and salary. We used a scatter plot because we wanted to show each individual players' PPG and salary yet also show this relationship as a whole to see if there was a positive or negative correlation between the average points a player scored in a game and their salary. In general, it seemed like there was a positive correlation between PPG and salary: the

more a player scored on average per game, the more money they earned.

3. Age vs. Salary:



The third visualization we made is a violin plot that shows the relationship between the age of the player and how much they get paid. We are using the two features from the database called Age and Salary. The reason for using the violin to show the relationship is because the line of the violin can show the salary distribution for the different age periods. The higher line means there is a point with higher salary in the specific age scope. It potentially shows the golden age of the player starts from around 23, during this period of age, the player possibly gets a higher salary and more points. Also, there exists a great gap between the salary that different players will get paid. The highest salary that the player can get is around 80 million dollars, which is around three times of the lowest player which is less than 20 million dollars.

Data Snippet:

	FULL NAME	TEAM	POS	AGE	PPG	APG	MPG	RPG	SPG	BPG	season	Salary
0	Bradley Beal	Was	SG	24.00	22.6	4.5	36.3	4.4	1.18	0.44	2017-2018	25434263.0
1	Jrue Holiday	Nor	PG	27.00	19.0	6.0	36.1	4.5	1.52	0.79	2017-2018	25686667.0
2	CJ McCollum	Por	SG	26.00	21.4	3.4	36.1	4.0	0.95	0.43	2017-2018	23962573.0
3	Dwight Howard	Cha	C	32.00	16.6	1.3	30.4	12.5	0.59	1.62	2017-2018	23819725.0
4	DeMar DeRozan	Tor	SG	28.00	23.0	5.2	33.9	3.9	1.06	0.28	2017-2018	27739975.0
...	...	...	...	...	...	...	...	...	...	...	...	...
944	Luc Mbah a Moute	Hou	F	33.93	1.7	0.0	8.2	0.7	0.67	0.00	2019-2020	4320500.0
945	Isaiah Roby	Okc	F	22.53	0.0	0.0	3.8	0.7	0.00	0.00	2019-2020	1500000.0
946	Tyler Zeller	San	F-C	30.58	1.0	0.0	2.0	2.0	0.00	0.00	2019-2020	245686.0
947	Max Strus	Chi	G-F	24.38	2.5	0.0	3.1	0.5	0.00	0.00	2019-2020	898310.0
948	Ryan Anderson	Hou	F	32.27	2.5	1.0	7.0	3.5	0.50	0.00	2019-2020	1620564.0

Appendix C: Interview

End User Personas

As an end user I have not used a visualization before. One task that comes to mind is to compare players' performance across different seasons to have conversations with friends. Find the performance of players in specific areas such as best players for a given position, or best player for shooting three points for example. One last task is to compare the best three point shooters, defense to find who is the best three point shooter that plays defense. Or a group of defenders who play the best offense.

Currently to perform this task we look up raw data as numbers and rankings to compare players in the NBA.

Using this visualization tool we hope to save time in our day to day lives, not having to look up stats on different websites. Also we hope to use this visualization to help show others and ourselves, comparisons between players

Interview Script

What do you need a visualization for?

What is the purpose of your visualization? What do you hope to achieve with this visualization?

Follow Up: What do you want to learn from this visualization?

What would you want to use it for and what do you want it to show?

What measurements would you like us to use to show this?

Follow Up: Do you want to compare them?

Can you explain your workflow to me in five steps?

Follow Up: Ask for more information on a particular step

Please describe three positive aspects of your current process.

Please describe three pain points in your current process.

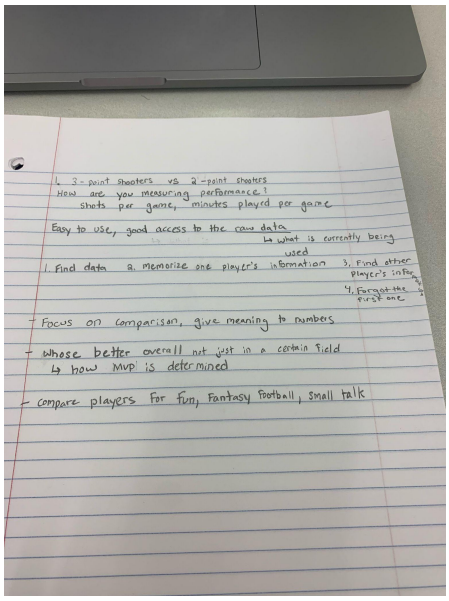
Follow Up: Is there anything that can be done to improve this?

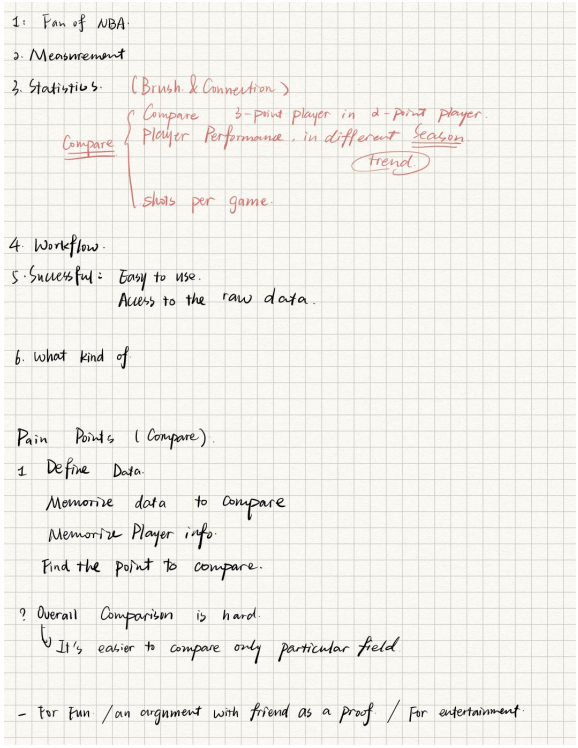
What tools do you need to be successful?

Who are the end users of your visualization?

Do you think your visualization is able to provide the end user with an easy tool to do things that they want to do so?

Interview Notes





Interview Results

<u>What do you need a visualization for?</u>

To compare the performance between two players for fun, fantasy basketball and small talk.

<u>How do you measure performance?</u>

Shots per game or minutes per game.

<u>Who is the end users of your visualization?</u>

Our end users are NBA fans who are interested in the data of the players and want to see how their favorite players perform versus their salaries.

<u>Do you think your visualization is able to provide the end user with na easy tool to do things that they want to do?</u>

Yes, we think that our visualization provides the end users with the ability to select the players that they are interested in and compare the attributes of them and come with the conclusion to the influence of their salaries. And easy brush to select the players.

<u>Are there any other factors about players that you'd like to compare?</u>

How many 3 point shots a player gets versus 2 point shots. To compare the player's performance and to see how they act in different seasons.

<u>What qualities should the visualization have in order to be successful?</u>

It should be easy to use and allow for easy access to the raw data

<u>Are you currently using this raw data to do these comparisons?</u>

Yes.

<u>Can you describe the process of how you do this and what in particular makes this challenging?</u>

First we find the data on a particular player and memorize it. Then we find the information of the player we want to compare this to, but by the time we do this we have forgotten the first player's data.

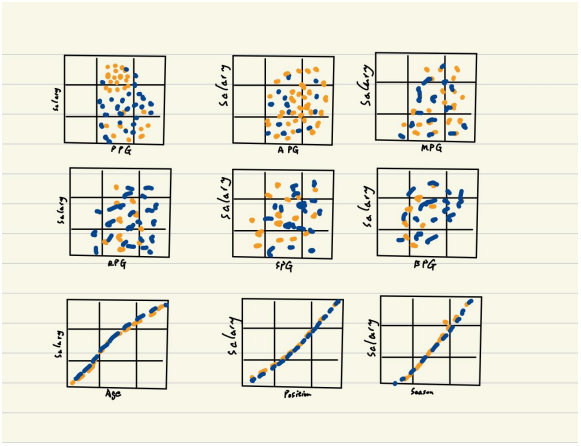
<u>What are you hoping to achieve with this visualization?</u>

We want to focus on comparisons in order to give meaning to the numbers in our data set. It's one thing to read numbers from a data set but having a visualization really helps to put them in perspective.

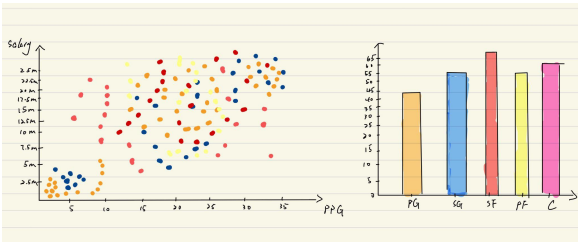
<u>What is the problem that you think this visualization might be hard to come up with?</u>

It is hard to create an overall comparison between the players to analyze whether or not the player is the best. Something we would be interested in is seeing how an MVP is determined by comparing all of these statistics to get an overall best player. However, it is easier to come up with to compare the players' performance in some particular field.

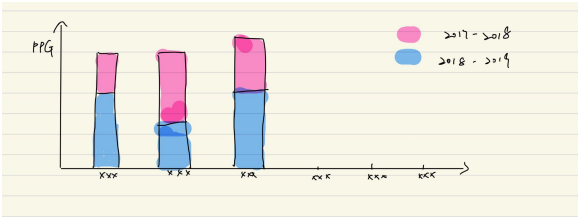
Appendix D: Design Sketch



Favorite 1.

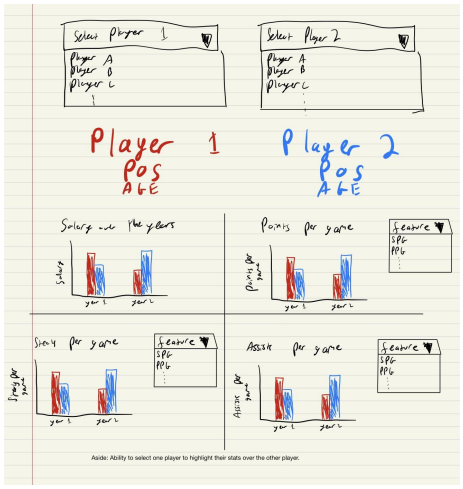


Favorite 2.

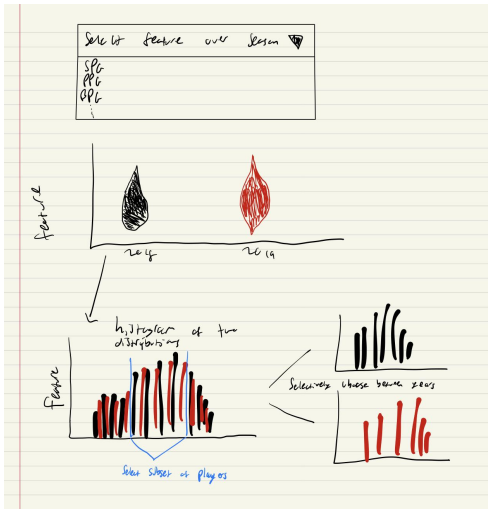


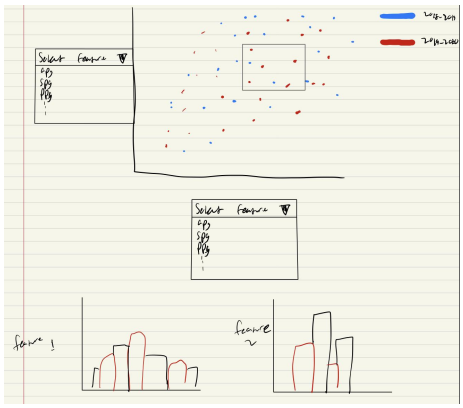
By Conghan Liu

1. For this chart, I use small multiples to compare each categories' relation with salary. It is linked with each other that when brushing a specific area of one chart, it will also be shown in the other eight charts. This is used to compare the difference between a pointed aspect of a category and compare the same player with their different aspect.
2. For this chart, I use scatter plot and bar chart to find the relationship between salar, point per game and position. Five different colors represent five different positions. The bar chart shows the number of players in each position, and in the scatter plot, use color to represent each position's player. We can use this chart to find the performance of each position's player and find the best position.
3. This is a bar chart that is used to compare the performance of a player in three years. It can be used to find whether the player has improved or fallen behind in the last three years.



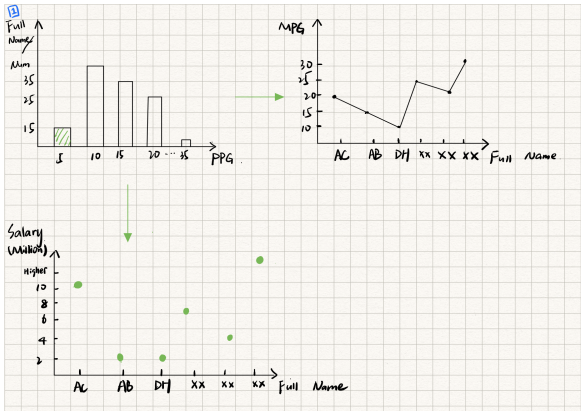
Favorite 3.

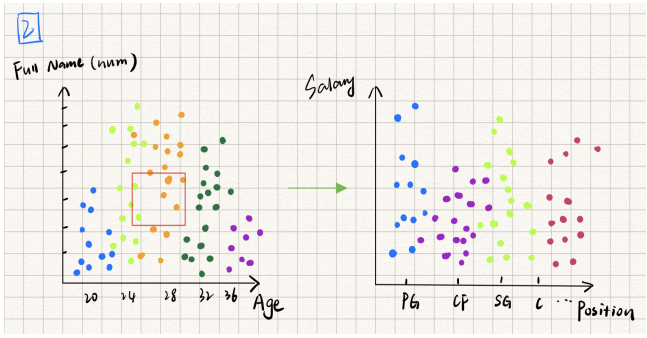


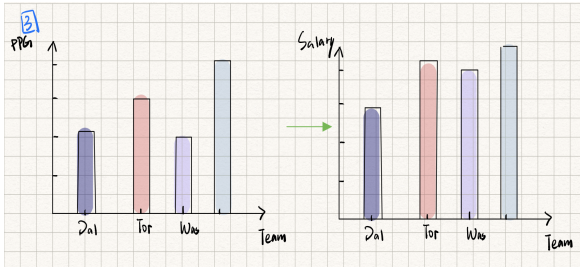


By Adrian Monaghan

1. For the first sketch I made a simple grid of barcharts. A user would be able to select two different players to compare along with four different features to compare them on. Each of the two players is given a certain color to easily distinguish between the two throughout the four barcharts. Barcharts were chosen to show the quantitative difference between the players as it is easy to determine the difference between the players as well as the magnitude of difference. This accomplishes the task of showing why my favorite player is better than a friend's favorite player.
2. For the second visualization I aimed to show the distribution of a given feature of several seasons. It performs this by using violin plots and bar charts. Violin plots can easily show the distribution of a given variable by the width of the plot at given values. A histogram of multiple seasons on one graph would allow for easier comparison in distributions across multiple seasons. Color coding the different seasons allows for easier comparison when shown on the histogram.
3. For the third visualization I aimed to allow the user to try and determine correlations between different feature variables of their choice. This would be done by allowing a choice of feature variables and then providing a scatter plot. Below would show two histograms of the distributions of the individual variables. The user would be able to brush the scatter plot to determine which data points they would like considered in the distribution of the histogram.

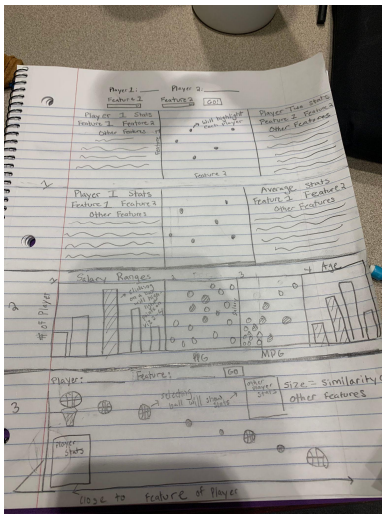






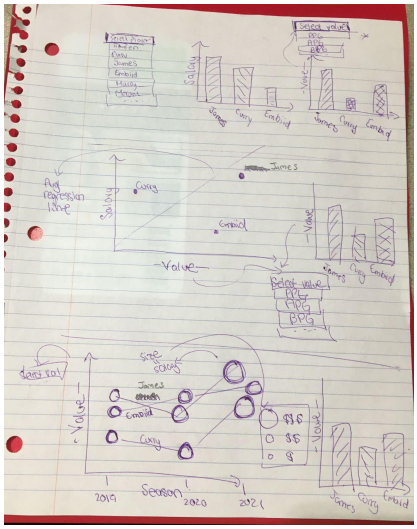
By Mengting Tang

1. For the first graph, I am trying to visualize the point per game and group by the number of how many players got this point and display it by the bar chart. Using the functionality of the selection, it will connect to the right and below graph which is showing the group that you select and display the group's minutes per game and the salary.
2. For the second visualization, I aimed to show the connection of the age to the player and how the position will influence the salary.
3. For the third visualization, I am going to focus on the team and to see how the team's performances vary and get paid averagely. Meanwhile, comparing the points per game for each team is also important. By the two charts, it can easily and directly see how the points they earned will influence their salary for each team.



By Sarah Costa

1. I chose my first dashboard to contain multiple scatterplots of points because I wanted to show how two quantitative features compare to each other for all players. Because I am using points on a scatter plot, both horizontal and vertical position is being used each to show one feature. Color is also being used to show which player or players have been selected. This dashboard helps us to achieve the goal of comparing two players by a particular trait or two by showing each player side by side, which helps achieve task #1.
2. I chose my second dashboard to contain multiple scatter plots and bar charts to convey how a few features each compare to a player's salary. I did this because a lot of our initial visualizations also focused heavily on salary. This visualization would use brushing and linking to compare each field together side by side. This visualization also uses horizontal position to show a player's MPG (average minutes played per game) and PPG (average points scored per game). It also uses vertical position to show a player's salary for the scatter plots and as a count of how many players fall into each range for the bar charts. It also uses color to show which points and bars are highlighted when brushing and linking. This also helps how trends in players over time by displaying players of different ages, which helps to achieve task #3.
3. My third visualization is an attempt to compare players as a whole based on a variety of different traits as was discussed during our interview. The viewer would choose a player and what feature they want to focus on (MPG, PPG etc.) the most and it would give them the top 10 players that have closest stats for that particular feature. The rest of the features are represented in the visualization by creating a similarity metric using the distances between each player and the chosen player for each feature and adding them together. This similarity metric is depicted in the visualization using size of the basketball. The players who are the most similar to the chosen player are larger than the ones who are less similar. I also used horizontal position to show how similar a player is to the designated player in the feature that was deemed most important by the user. This can help someone figure out what players are best for their particular needs and if they aren't available to pick for their fantasy team, then they can pick someone similar in statistics, thus this helps us achieve the task of task #2.



By Brandon Onyejekwe

1. For the first dashboard, I have two coordinated bar charts, where the first shows each player's salary, while the second shows the value of a specified feature for each player. Both the feature for the second chart and the set of players showing up for both charts are to be chosen with drop-down menus. Linking is done between the two charts such that when a player's bar is selected, it will be highlighted on the other chart. This dashboard is used to compare players with each other and show why one is better than another in terms of performance by directly comparing their stats.
2. For the second dashboard, I incorporated a scatter plot as the first visualization. This shows the same selected feature as the second bar chart, both of which are chosen through a single drop-down menu. This feature is plotted on the x axis for the scatter plot while salary is on the y

axis, and each point represents a specific player. There will be a line in the scatterplot showing the average salary for each feature value. Linking is coordinated in the same way; a player being selected in one plot is highlighted in the other. This dashboard is used to compare players with each other and show why one is better than another in terms of performance by showing them in relation to an average across all players..

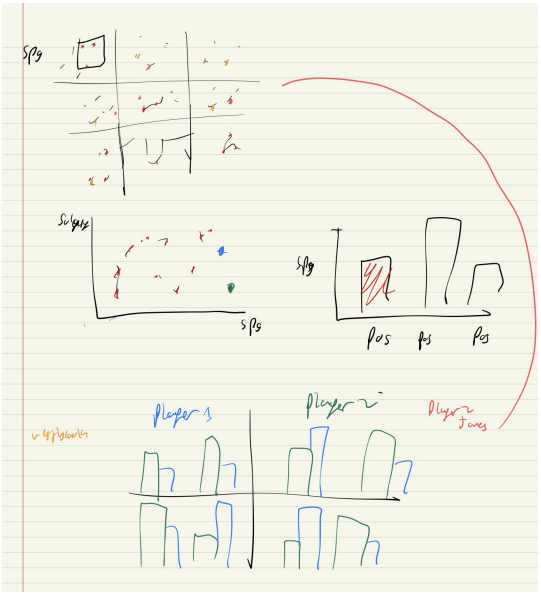
3. For the third dashboard, I used a scatter plot for the first visualization, where the x axis is the season played, and the y axis is the selected feature, coordinated with the second plot with the same drop-down menu. Each player has a line in the first visualization, connecting their value for the feature across different seasons. Size of the connecting points on each line is meant to show the relative salary size. Linking is coordinated in the same way; a player being selected in one plot is highlighted in the other. This is used to show the trends of players over seasons by plotting their performance over different years.

What data types and tasks do each support? Why are the depicted visual encodings effective?

In the first of the favorite visualizations we compare the distribution of the salaries of players to their different quantitative averages over seasons such as salary versus points per game, or salary versus assists per game. There is also an ordinal variable shown by the color of the dot on the map. This visualization is best for showing which players provide the best value for their skill. This visualization would easily allow a user to see which players provide the most value to their teams. The ability, as well, to select players in a certain subsection of salary versus shots per game, and see how they compare when considering salary versus blocks per game.

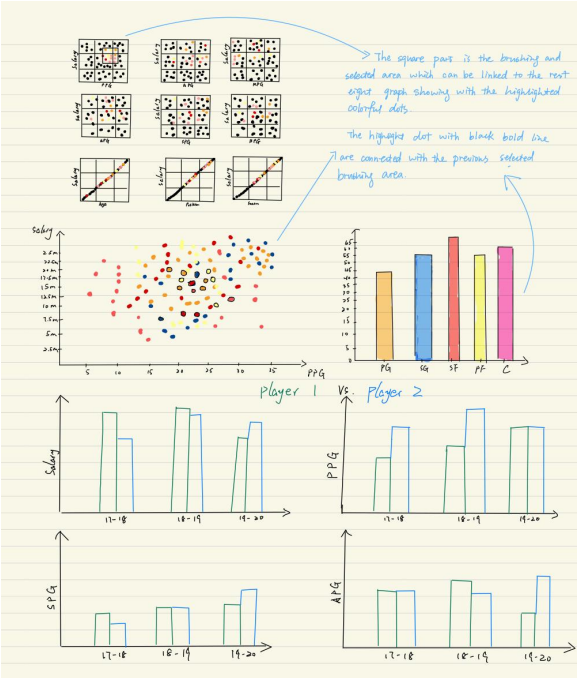
For the second visualization, we use the average points per game for each player, as well as the salary for each player, both of which are quantitative data. We use color to represent the categorical data of the 5 positions (PG, SG, SF, PF, C). The visualization is a scatter plot and a bar chart. The scatterplot has the specified feature on the x axis (ie. PPG) and the salary on the y axis, which each player corresponds to a single point. The bar chart has the positions on the x axis and the count on the y axis. These two visualizations will be linked together, so highlighting a point in one will highlight the corresponding position bar, and vice versa. This visualization will be effective in showing trends with respect to position.

For the third visualization, we can provide a direct and customizable comparison for two players. This visualization allows the user to easily select which categories they would like to compare players on as well as which players to compare. The use of bar charts as well allows for simple visual comparison and the use of colors to denote the difference between the two players allows for quick determination of who has the better stats between the two.



First a user would be able to select a subset from one of the nine scatter plots at the top of the page. For example, selecting a subset of the highest scoring and lowest paid athletes. This subset of players would then show up on a second scatter plot below allowing a user to get a better look at the distribution of players as well as the distribution of players in that position. The user would then be able to select two players from the subset of players to allow for a more direct comparison between the two players on four different, and user customizable, quantitative variables such as: points per game, minutes per game, assists per game, and steals per game. This would allow users to compare the individual players in a more direct and granular way.

Appendix E: Digital Sketch



We visualize a user using this visualization for two main tasks. First, we could see a user look at a player's value based on their salary and their performance on certain metrics. This would also allow a comparison between performance and position. From these graphs a user would then be able to select two players on the scatter plot of the selected statistic versus salary to compare their performance. For example they could compare the overall performance of a player that gets paid well but performs poorly and a player that gets paid poorly but performs well. The second task we could see a user using this visualization for is the direct comparison between two players of their choice. For example, they could compare the performance of Lebron James and Russel Westbrook in areas of a user's choice. And then the two players would be highlighted in the other graphs allowing further comparison between their performance and salary earned.

Appendix F: Usability Testing

Preparation

The first visualization is a 3 by 2 + 1 metrics scatter plot chat with salary on the y-axis and "PPG, APG, AGE, MPG, MPG, SPG, RPG " on the x-axis. The second is a scatter plot about PPG and the salary linked to a bar chart with the position and the number of each position. The third visualization is four bar charts to compare two players in the three seasons. The four charts use the same x-axis, season, to connect and compare the salary, PPG, SPG, and APG.

In order to test our visualization, firstly we want to see if the three visualizations accomplish the goal of what we are pursuing to show from this visualization and webpage. With this test, we want to see if the logic of the visualization works well. Secondly, we want to test if the interactive visualizations' functionalities perform in an easy and understandable way. The reason why we want to test it is because due to the requirements of the project, we want to see whether the visualization itself is easy to use for the intended audience. Last but not least, we want to compare the performance of two different players and find out which one is the better. By testing this, we can easily select the players that we want to compare to see their performance which fit our initial goal to help the intended audience to know the performance of the players they care about by visualizing it and with easy click.

Results

Some things our tester liked about our visualizations were the brushing and linking we implemented as well as the consistency of our y-axes across scatter plots. One issue that our tester had with our visualizations was that it was very tricky if not impossible to complete our third task because we had not implemented the necessary code to fulfill it.

In response to our first task, our usability tester said that judging by our visualizations, it seems like salary is correlated with everything besides age and BPG but that it was unclear if any position is earning more than the rest. In response to our second task, our usability tester said our visualizations are very intuitive and easy to understand and that having a consistent y-axis really helps to differentiate what is changing between graphs. In response to our third task, our usability tester said that it was hard to complete this task but this was due to our implementation not being fully complete at the time usability testing occurred. We also noticed that our tester didn't even notice our bar charts at the bottom of our page, which would have been used in this task.

We absorb the recommendations after the usability test. First, we make the all seven Scatter plots in the first part clickable. In that case, when the users want to see more clearly with any charts, it will have a clear view in the middle section. Then we change our number of positions from three to nine. In that case, it has more options for users to check whatever they want to find out. For the last part, we expand the button of the drop down list to make sure users have a clear view with it. Now, it is also possible to select players from a list to compare them.

Appendix G: Reflections

Brandon:

I believe that overall, we did a great job as a group. Personally, I think we did a good job distributing the work and making sure that everything was contributing well. I felt that I put a lot of effort into making some small improvements that made a difference, such as making the code for the scatterplots more functional and less repetitive. We did a good job scheduling meetings and making sure that we were productive during our meetings. In terms of things we could have done differently, I wish we were able to spend more time working on each project milestone each week. I also wish we had gone to the TAs and professor much more often when we had doubts, as that would have made the process go a lot smoother. Finally, I wish we had had a better sense of where we were in the project as time went along. I think it was hard to put into perspective how much of the project we had to accomplish each week. The most valuable and somewhat surprising thing I learned about visualization this year is that designing visualizations is a lot more than making subjective plots that look nice. When designing a visualization, there is so much you need to take into account and plan for before even making initial sketches, in order to make it as valuable and intuitive as possible.

Conghan:

I think we have a great group work on this project. I help to give the idea about the visualization and draw the digital sketch for our visualization. We discussed a lot during the design process of the visualization and made sure everything was a perfect match for our project. I also put effort into fixing the problems on the code to make it look better. If I have another chance to build this visualization, I will add more players to the comparison and more player's data into the visualization. It can make this visualization more professional. I wish I could fix the position sorted for the player. Right now we sorted players to nine different positions which is a little bit too much. People usually divided the player into five positions. Nine position kind of repetition in this. I also wish that I can add more analysis about player performance into this. Right now, we only have seven aspects of performance from a player, there are also other aspects that can affect player performance and their salary. It can also be used to find out the difference between a player's salary. The most important thing that I learned from this course is that what we designed at first is just what we want to do. When we were trying to code it, you will find out some are not fit for your design. In the meanwhile, designing a website page with the visualization is a long process that requires timely modification to make it perfect.

Adrian:

I felt I worked well in the group environment and provided value to the group to allow for a solid end product to be delivered. I felt the communication between my team was good. We never had any real misunderstandings and we were always able to find a time to meet that worked for everyone. I also felt that as a team, we were able to split up the work evenly and we were all able to provide equally to all parts of the project between the code, the presentation, and the report. One thing I would change about this project is having a better idea of what is due week to week as opposed to learning about what we have to work on for the visualization a week before it is due. I would have started work on the actual visualization much earlier in the semester as it felt that we had to cram it in the last few weeks. Finally, I would use TA and professor office hours sooner and more often as there were issues we ran into that could have been solved by speaking to a TA sooner in the project. I think the most surprising lesson I learned about visualization design through the course of this project was how complicated it is and how much work should go into designing even a simple visualization.

Sarah:

I think our group as a whole was really good at meeting consistently and all showing up to and committing to these meetings. I think personally I did a good job of facilitating our interview process and figuring out the things our group needed to work on each week. I would definitely go to office hours more often to learn more about coding in D3 to improve our visualizations but I would also try to explore more unique ways of coding visualizations in D3 outside of just bar charts and scatter plots. Lastly, I would also try to check over our documents more often. One thing I learned about visualization design that surprised me was how difficult it can be to create and implement visualizations that are easier to make in python with libraries like matplotlib and the importance of every single component of your visualization.

Mengting:

I think overall our group has a great environment of collaboration on the project. Personally, I think we all contributed equally and present actively during the meeting that we set to meet. I felt I provided value when designing the visualization and brainstorming the ideas. I also try my best to fix the problems in the coding part and try to fix the axis problem of our bar plots. If we have a second chance to do this project, I think we'd better start the coding parts more early than now. Since for now, we only started to do the coding part in the very last few weeks. And it makes the workload a little bit hard. If we start coding early, we will provide a better job with some minor details fixed. And

have more functionalities being added to our final version and make it more alike to what we imagined and planned to do at the beginning of our digital sketch. Besides this, I think we can also plan ahead to meet for the project milestone instead of meeting solid on Tuesday. Moreover, I think I will try to use the help of TAs and professor's office hours for the final project too to ask for the rubric of the project milestone and the coding part. Then we will not lose some minor points that stated in the project milestone and will have a better score. Through the course, one surprising thing that I learned is that I can handle the design of html by using the javascript. The functionalities like mouse event, tooltip, scatterplot and so on. And doing the project helps me with reinforcing the knowledge of designing the html by javascript.

Appendix H: Presentation Feedback

Project Name	**Specific aspect of the visualization tool or presentation that you liked the most**	**Follow-up Question that you have for the group**
Basketball Shot Percentages	The use of color was good to help determine which team is which as well as which teams made it to the playoffs. I like the way they pay attention to the playoffs, which is a great aspect.	Why didn't the basketball court visualization not end up working in the end? Do you think your compromise worked out well?
Wordle Visualization	The use of api's to get data for the visualization is very cool. It would be cool if they could make it so it updates as new data comes in.	What was the hardest decision you had to make in visualizing the wordle data?
Bechdel Test	The visualization is very clean and very intuitive. It looks really good. The amount of customization given to the user is very cool, however, it does not seem overwhelming which is very important.	Did the data come pre-cleaned and pre-computed for the bechdel test? Or did you have to go through and determine movies that passed and failed the test?

Baseball Salary	The bracket visualization is very cool and unique. It is good that you were able to use a visualization beyond the basic visualizations like bar charts and scatter plots.	How was the bracket made in d3? Is there a function in d3 or did you have to do it by hand?
LEARN	The map and interaction on the visualization looks very good and would give the user a good idea of the difference in education across MA?	How did you get the choropleth map in d3? Did you have to draw it? Did you find a map online? How did you get the data to coincide with a position on the visualization?
Plastic Pollution	We liked that even though a map looks better, you decided to use grouped bar charts that allows the users to have better/easier comparison between countries?	How late in the project did you decide to switch to barcharts from a map? Do you think in the future you would go back and add it back?
Visualizing Congress Stock Trading	We like the connection between the treemap and the bar chart. When you click on each company, the bar chart will change to the specific frequency of that stock. Moreover, it can also show the political party in the bar chart.	The tree map's color is just a little bit simple and hard to know the difference, do you ever think make it a little bit colorful so that people can differentiate it.
Covid-19 Dashboard	We like the idea of creating a covid-19 cases dashboard and using line charts to show how the cases rise in a period of time and having the functionalities of selecting the states.	What is the hardest problem that you encountered that made your visualization different that you first imagined it to be? Besides you mentioned just focus on one state, is there any other solution that can possibly solve the problem?
Expected Performance in the EPL	We liked the part that showed each team logo in the scatter plot instead of points. And the functionalities of drop down selection to show the season and the performance.	How do you make each point with a picture in d3, do you encounter any difficulties when you try to figure it out?
Once Upon a Viz	The way of combining the presentation about your ideas and your final visualization plots into the powerpoint is	Is the data cleaning process hard for you since you need to review the book reviews?

	great. The use of bubble charts and line charts is a great idea when you can link it together with the book point by brushing.	
MBTA Visualization	We liked the design of the MBTA rapid transit map. It provides the end users the functionality of zoom and pan around to select the station.	How did you make the map work in d3 with the dots showing the path of the train?
Visualizing the Recipe	We loved the idea of your visualization of the recipe. And your design of selecting the attribute of x axis and y axis and then to compare the two attributes are great. We also enjoyed the save functionality and the great use of word cloud.	What pushes you to not use the axis histograms through your design process? The save functionality is pretty useful, do you think about adding the functionality of save the lists to pdf?
Exoplanet Visualization	The animations to show changing radius and mass look very good. The use of tool tips to gain further information on the specific exoplanet is very smart too.	Do you think you could provide a visualization in the future that could show where these exoplanets are in relation to the stars they are for?
Which City Should I live in?	The idea of the project about showing the rates of the city to see whether it is a better city to live in is great. And the functionality of linking the three plots that we have and coming with the city's information showing in the tooltip is clear to see the three categories of rates in detail.	Only three categories of rates to show whether a city is good enough or not to be a choice to live is kind of simple, if there is a chance to make some future improvements, what are the other factors that you want to add to become the factors under consideration?
Crypto	The great use of different kinds of charts like bump chart, candlestick chart and the overview of the market is nice. The linking in the market overview and the highlight of the pie chart when mouse on is perfectly fit with the goal of educating the audience.	Do you think there is any other information you could include when you hover over information with your mouse?

Bike Allocation Optimization for Citi Bike	The count on both start station and end station is great that we can find different aspects of the bike in New York. The Scatter plot shows the each station and each point of the count.	Do you think there is the possibility of adding functionality or interactivity with a map of New York so users could have a better idea of what stations geographically would be more likely to have bikes?
CoOp'd	I liked the Apply and Reset button on the top of the visualization. The drop down list makes it easier to separate from college to college. The second bar chart shows the average pay, number of Co-op and the location of coop by dividing up to three times of Co-op which is great.	Would you implement functionality in the future to submit feedback on co-ops through your visualization so there is a running list of feedback?
Affordability of Diets Around the World	We like the topic of this project. The map of visualization shows how the diet costs in each country. There is also the comparison between income and the percentage of unable to afford nutrients with the demand information showing each scatter plot's country and information.	We are wondering how the income classification legend has been defined. And what is the color difference in the map?
Food Desert Visualization	We liked the use of the choropleth map and appreciated that you wanted to add the ability to zoom in state by state. We think that would be a huge improvement on the visualization.	What was the reason you weren't able to implement the functionality to zoom into a specific state?
Bar Hopper	We like the design of separate the potential users of the visualization into four parts and start the design of visualization. By easily clicking on the range of the bar, all plots are linked and provided with the opening time and direct hyperlink to yelp to see more comments on the selected bar is really convenient.	How did you think about using the user-surrounding scatterplot and how to implement it in d3? Why choose not showing the distances between the bar to the current position and instead showing the longitude and latitude?