Stats and ml techniques combined with python libraries to draw insight from NBA data

NBA Shot Analysis Report

Sanseerat Virk

Sanseerat Virk

Table of Contents

Sanseerat Virk

Intro

This report aims to find insights reveled by analyzing NBA shooting data, examining shooting focusing on the 2022-2023 NBA season. Scenario ex (“We are commissioned by an NBA team seeking insights into the impact of three-point shots on game outcomes”). An additional personal request was made by the NBA team, we've conducted a detailed analysis focusing on the Denver Nuggets, the reigning basketball champions. Our objective is to provide a visual representation of where the team's starting players pose the greatest scoring threats on the court. This report incorporates references to both machine learning (ML) techniques and statistical methodologies.

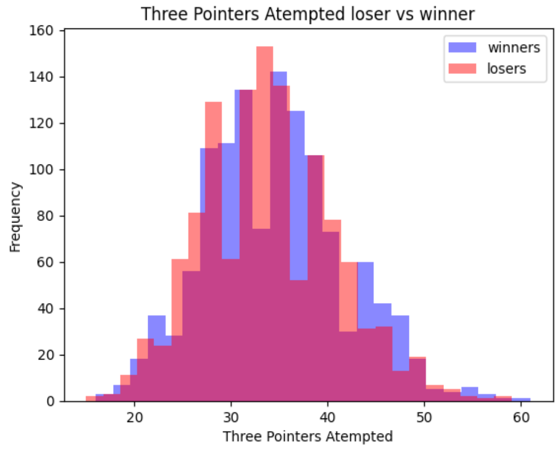
Three Point shots for the 2022-2023 season

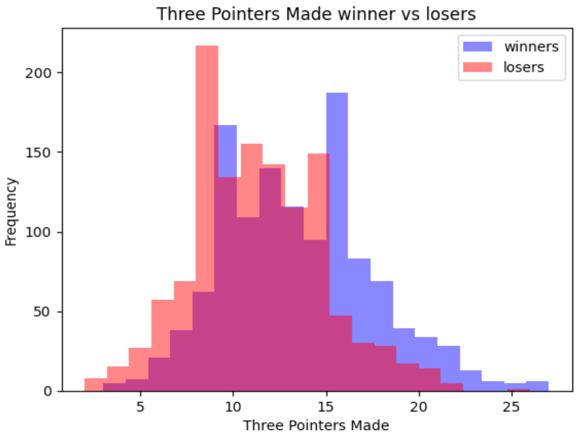
Gathering Data

Using the NBA API which takes care of making the requests to the NBA database. We first requested a Schedule for the 2022-2023 season, which allowed us to obtain the game id. Using the game id for each we could extract the data for every game and see how the winner were for the game. Along with players stats and various other types of information. Getting this data between each run was not plausible as it took very long to make the request to the NBA API. So, a custom python script to gather the data and filter to our needs was made. This script was written by referencing information from https://github.com/swar/nba_api.

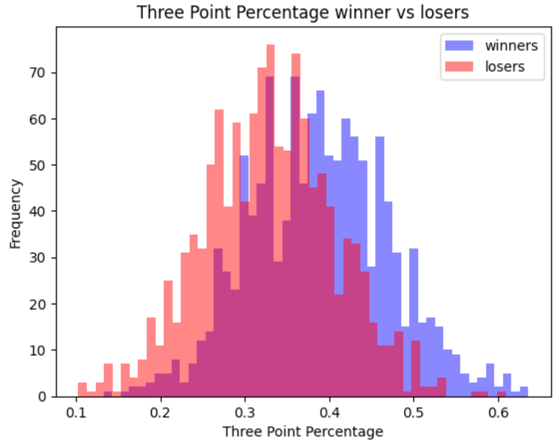
Analyzing Data

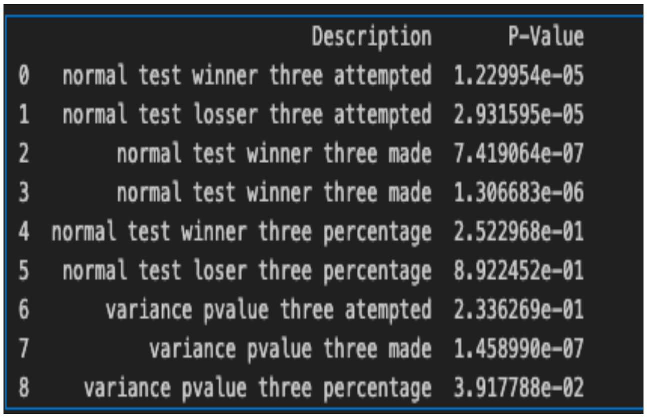
Since each game had to have a winner and loser, we could separate the data into data frames of equal length. Then use a histogram to visualize the three-point field goal attempts, makes and percentage. Data was tested for normality, equal variance. The test concluded data was not normal and did not have equal variance as the p value was very small for three attempts, three made, percentage. This directed us to use the Mann W test to compare the winner and loser for the three categories.





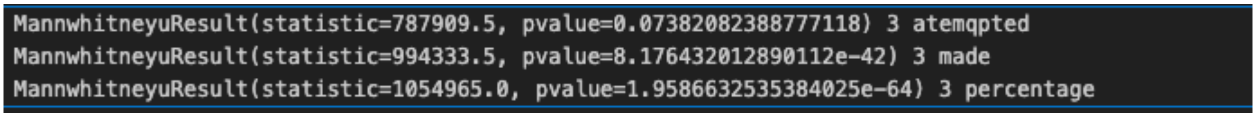
Sanseerat Virk





```
                          Description       P-Value
0   normal test winner three attempted  1.229954e-05
1   normal test losser three attempted  2.931595e-05
2          normal test winner three made  7.419064e-07
3          normal test winner three made  1.306683e-06
4 normal test winner three percentage  2.522968e-01
5  normal test loser three percentage  8.922452e-01
6        variance pvalue three atempted  2.336269e-01
7            variance pvalue three made  1.458990e-07
8      variance pvalue three percentage  3.917788e-02
```

Result Mann W test

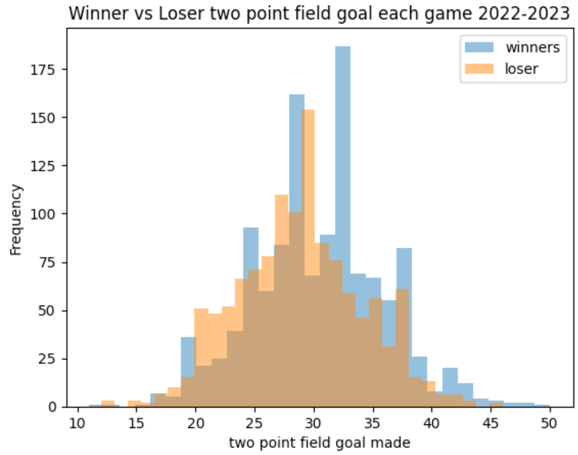


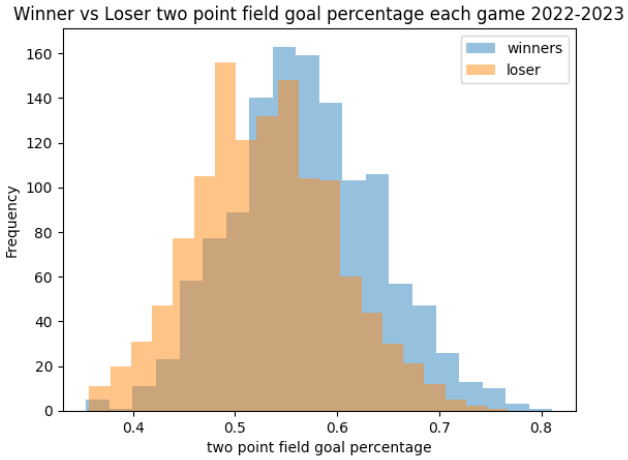
```
MannwhitneyuResult(statistic=787909.5, pvalue=0.07382082388777118) 3 atemqpted
MannwhitneyuResult(statistic=994333.5, pvalue=8.176432012890112e-42) 3 made
MannwhitneyuResult(statistic=1054965.0, pvalue=1.9586632535384025e-64) 3 percentage
```

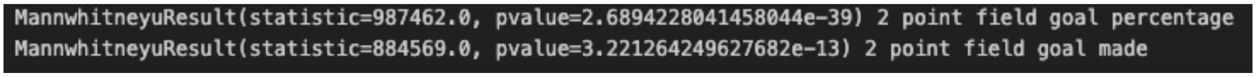
The Mann W test concluded that there was no statistical difference between the three attempted between losers and winner. However, for three made and percentage for the winner and loser it indicated one group was different from the other. Since we visualized the data, by overlapping the histograms for the three categories. It was evident that winners had a higher three-point percentage along with more threes made then their opponent.

Diving deeper Mann W test

we know from the previous part the winner of the game is likely the team which has a higher 3-point percentage/3 point made, does this relationship also hold for 2-point shots, (we count anything not a three pointer a 2-point field goal).







```
MannwhitneyuResult(statistic=987462.0, pvalue=2.6894228041458044e-39) 2 point field goal percentage
MannwhitneyuResult(statistic=884569.0, pvalue=3.221264249627682e-13) 2 point field goal made
```

Preforming the Mann W test for the two-point percent and two made we got similar results as well. Winners generally had higher percentage and shots made then their opponent who were the losers.

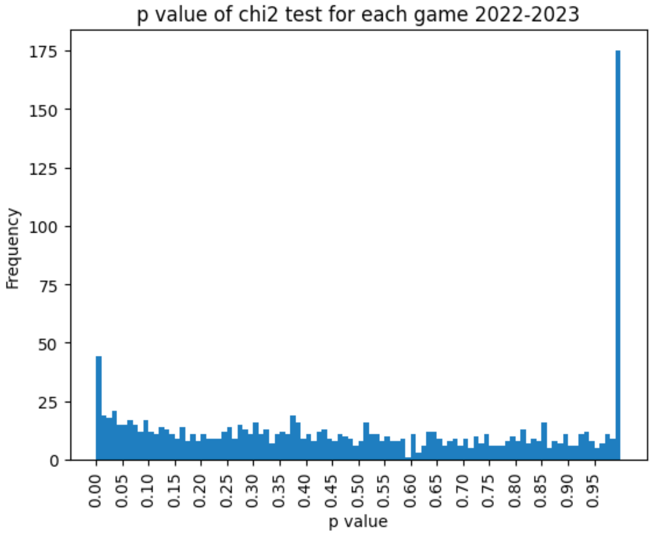
Sanseerat Virk

Diving deeper Chi-Square Test

To further test this point we preformed the ci-squared test for each game, the two categories being the winner and loser, and the rows begin the three and the two-point shot. We came up with an initially hypothesis. Weather you shoot the three or two it doesn't matter the difference between the winner and loser for shots made wither three or two will be the same. After preforming this test for every game and plotting the p-values we could visually see that majority of the p-values were greater then 0.05. Therefore, we failed to reject the null hypothesis.

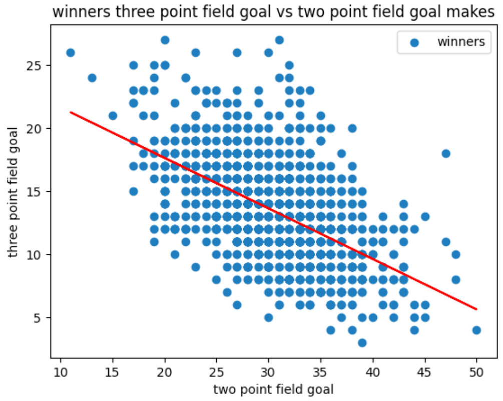
A table may look like this for one game, we compute this table and calculate the p value for each game then plot a histogram of the p values.

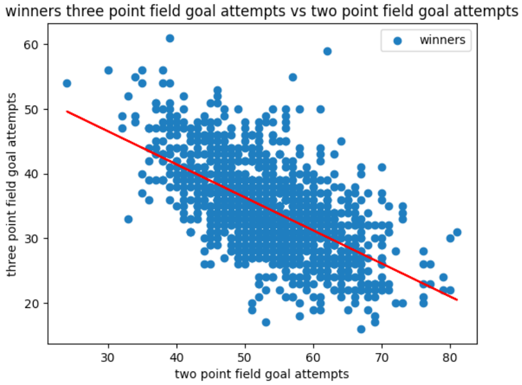
	Winner	Loser
Two made	A	B
Three made	C	D

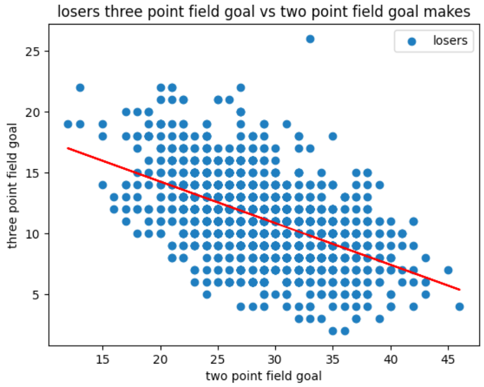


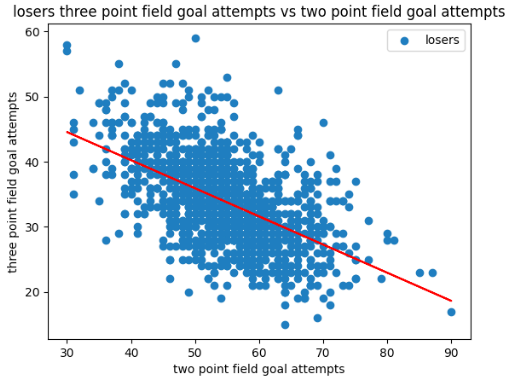
Correlation between three-point vs two point shot for both winners and losers

Is there a deeper relation here, if the team is shooting more three then are they likely to shoot two point shots









r value winner three point shot vs two points makes= -0.5448768629824203
r value winner three-point attempts vs two-point attempts= -0.6092580192930822
r value loser three-point vs two point makes = -0.5127821579278091
r value loser three-point attempts vs two-point attempts = -0.5494549200362506

visually and referencing the r vale we can conclude the team that will shoot more threes is likely to shoot less two pointers. This makes sense since on each possession a team must decide if they want to shoot a two pointer or a three pointer so if a team shoots more threes, then two pointers will be less.

No signification correlation between the percentage was observed if the team shoot higher percentages threes, then it does not mean they will shoot lower percentage two-point shot

ML classification, Denver Nuggets 2023 championship

Explore

The NBA team Denver nuggets are the current champions, having won the championship this year during the 2023 playoffs. It is important insight to know where their players were scoring shots on the court. Team may use this information for their defensive strategies. A visual representation would make it easy to interpret this information.

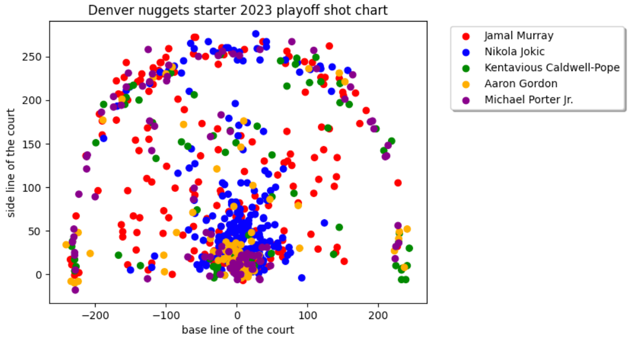
Gathering Data

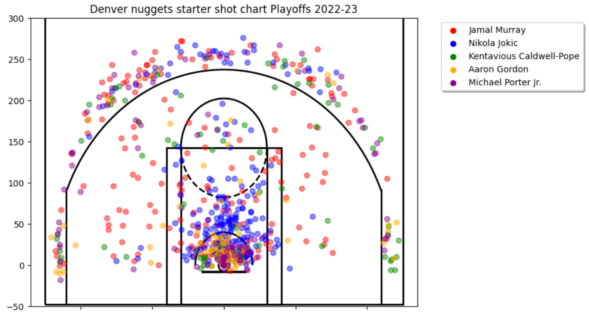
To formulate this information, we look at shot distribution for the Regular season and the playoffs. Since regular season has more games than playoffs it has more samples so it would make a good visual comparison with the playoff data.

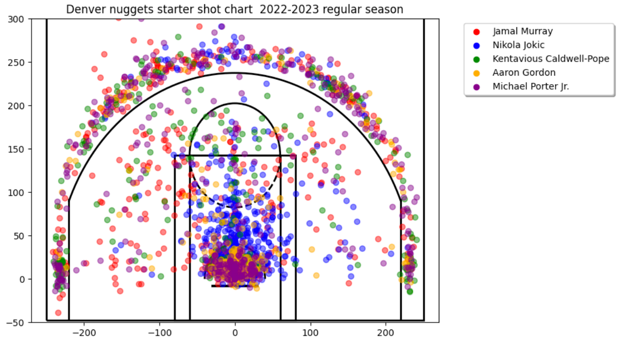
The shot data will comprise of only the starter since these individuals spend the most time together.
The unit with the most starts (spent most time together on court) for the 2022-2023
- K. Caldwell-Pope, A. Gordon, N. Jokic, J. Murray, M. Porter

Sanseerat Virk

By using the NBA API we obtain the shots for the regular season and plot the (x,y) of where the player took the shot on the court, assigning each player their own unique colour. To make this a more readable plot we draw a basketball in the background with a helpful function.







Cleaning Data

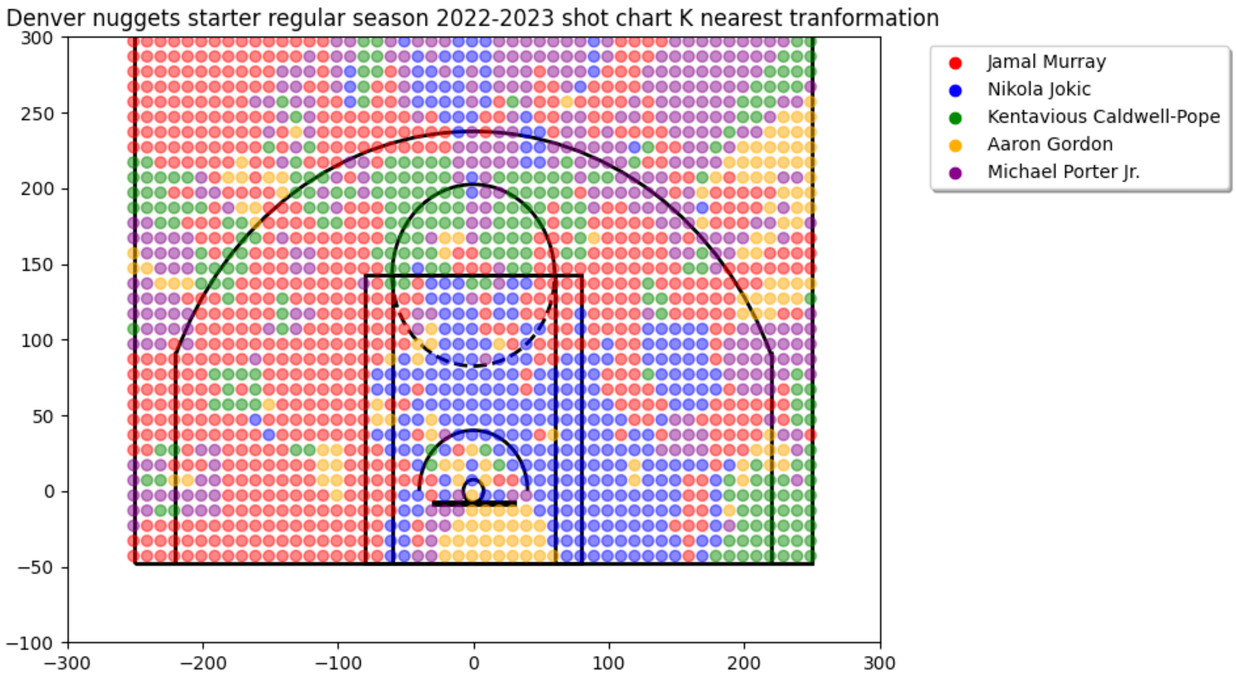
After we plotted the made shot location for the player, we saw many overlapping points and it was not clear visually which player is more likely to shoot and score from the spot. we reasoned a player that is likely to score from a particular spot has already scored the from near the spot. This made the k nearest neighbour classification a perfect candidate.

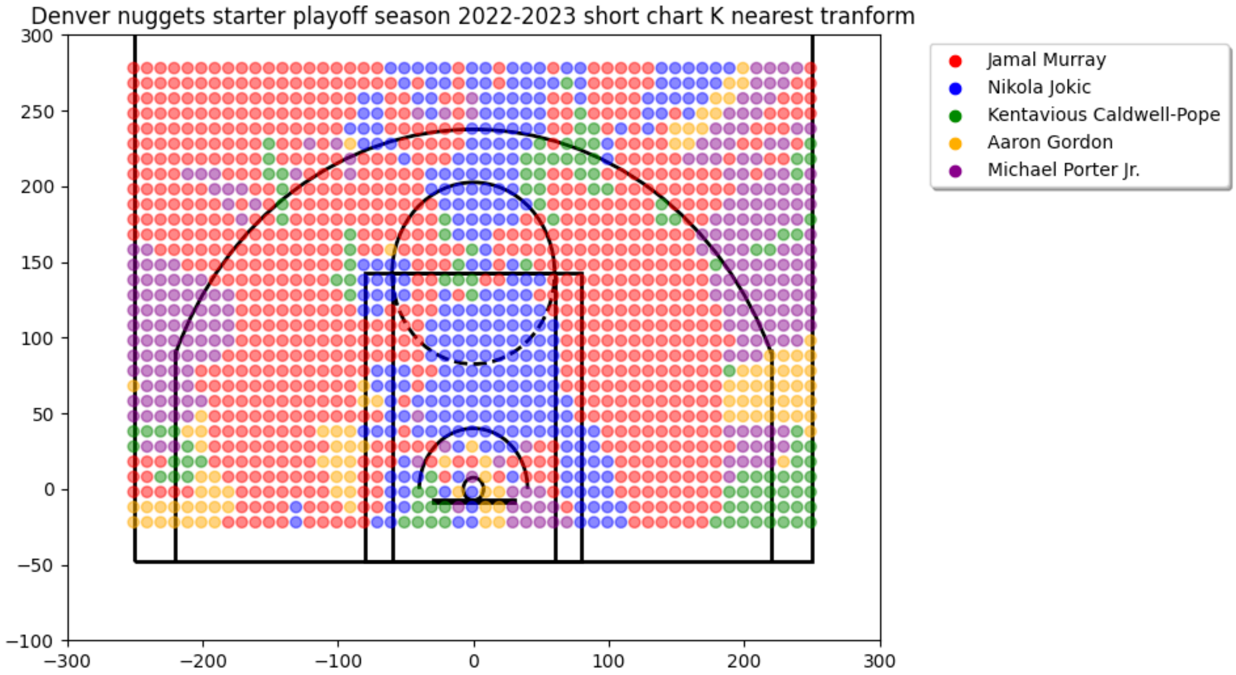
Using the k nearest model classification the goal was to clean the overlapping data point, so it was visually clear which player was the greatest threat form the spot. The score of the model was not as important. 5 percent of the data was spilt to generate score.

K=5
K neighbour model Playoff score = 0.5545454545454546
K neighbour model Regular season score 0.5362182502351834

Considering no modification were applied to the data these are not bad scores.

After building the model we simply generated all the points that could fit on the court and predicated what player was to shoot using the model. This gave us a nice visual allowing us to differentiate what player and where on the court were they the greatest threat to score which we could not when looking at the original data. This information can now be shared with the team.


Denver nuggets starter regular season 2022-2023 shot chart K nearest tranformation


Denver nuggets starter playoff season 2022-2023 short chart K nearest tranform

Example key observation we see Murray being more dominate in the playoffs then regular season, redder in playoffs. Pope being less dominant, less green than regular season.

Result

This visualization may not be accurate representation, but it does come close to show how out of the 5 started of the Denver nuggets where they are the most dominant based on how many shots they made near a spot. Also, it makes for a neat visual which crunches a large dataset making it interpretable without too much thought.

Limitations:

- As we started to the statistical tests with our data, it reveled more relationship we could test to offer more insights. Doing such we would be deviating from the goals of the report, so we had to set an upper limit and called it a day doing such we may just touched the surface of things leaving some insights off the table.
- The ml model made for visualization could factor in more parameters currently it consumes only (x, y) coordinates of made shots along with the name of the player. Deeper domain knowledge could factor in more prams for a more accurate representation of where the championship team player has the highest probability of scoring the basket.

Project Experience Summary:

- Successfully researched a way of obtaining NBA data. Developed a pipeline with custom script to get relevant data filtering out data not required. Custom script used the NBA API to obtain the NBA schedules. The game ids inside these schedules were then used to game details. Final output was the data of winner and losers for every NBA game for the 2022-2023 season in CSV format.
- Employed statistical techniques to find differences between winner and losers in three-point shots. Conducted Mann-Whitney test, T-tests. Normal-test, equal variance, and Chi-test. Analyzing their p values to draw insights and take correct order of actions to further find the truth.
- Constructed a visual representation of the shot data of the Denver nuggets team based the x,y coordinate for the championship and regular season 2022-2023. Cleaned up this data using K-nearest neighbor to construct a more readable and meaningful visual.