# MAT 243 Project One Summary Report

Jared Baker

Jared.baker@snhu.edu

Southern New Hampshire University

## Introduction: Problem Statement

I have been assigned the task of using data analytics to to provide descriptive statistics and data visualization in comparing NBA teams. I am using the data set of all NBA teams, their scores, wins & losses, and whether they were home or away. I will be using python to visualize the data, assess points of central tendency, and provide confidence intervals for average relative skill of teams (elo).

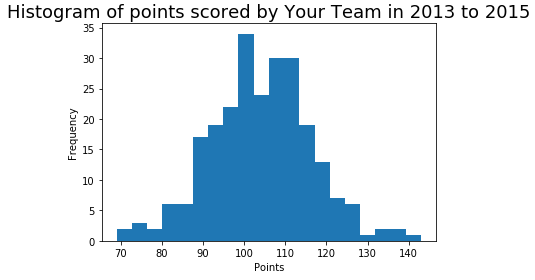
## Introduction: Your Team and the Assigned Team

Table 1. Information on the Teams

|  | **Name of Team** | **Assigned Years** |
| --- | --- | --- |
| 1. Yours | Nuggets | 2013 - 2015 |
| 2. Assigned | Bulls | 1996 – 1998 |

## Data Visualization: Points Scored by Your Team

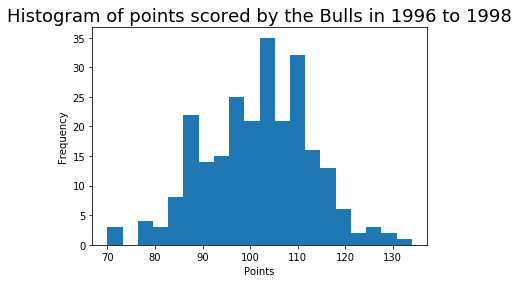
Data visualization is used to study data distribution and trends by reformatting a long list of figures into one clear picture that can easily show trends, outliers, and relative amounts. I chose the histogram representation of points scored by the Nuggets because it clearly shows central tendency. The scatterplot does not give much information as the x-axis is too compressed.



You cannot tell how many games are scoring in the central tendency because all of the dots are overlapping. The distribution of points scored appears to be very similar to a normal distribution, with a slight tendency to score under the median rather than over. This signifies that the Nuggets are very likely to score between 90-120 points.

## Data Visualization: Points Scored by the Assigned Team

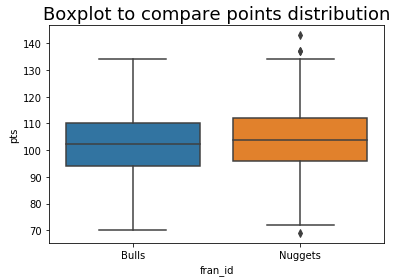
I chose the histogram for the Bulls via the same reasoning as the Nuggets. While the scatterplot shows the trend line through the years clearly, the histogram shows the point distribution, which is the metric we are currently interested in.



The distribution of the Bulls appears to have more inconsistency in points, shown by the spikes in the graph. Overall, the distribution is relatively normal and very similar to the Nuggets.

## Data Visualization: Comparing the Two Teams

Data visualization is used to compare two different data distributions by clearly showing one metric next to or on top of another metric. This paints a very clear picture of how the data differs as well as how the data is similar.



I chose the boxplot comparison because it clearly shows that the Nuggets score a little higher on average than the Bulls. The Nuggets have a higher median and mean as well as a smaller range. It also shows the Nuggets have 3 outlying games (two of which well above what is expected of them, vs 1 lower), while the Bulls do not have any. The histogram comparison shows the two previous visualizaitons on top of one another. While the plot looks interesting, there are no useful conclusions to draw without having to do your own math based on the graph.

## Descriptive Statistics: Points Scored By Your Team in Home Games

Table 2. Descriptive Statistics for Points Scored by Your Team in Home Games

| **Statistic** | **Value** |
| --- | --- |
| Mean | 106.2 |
| Median | 106.0 |
| Variance | 147.8 |
| Standard Deviation | 12.16 |

Measures of central tendency and variability are used to describe where a certain metric is expected to fall, and how likely it is for that metric to vary from that central point. The mean of 106.2 points means that the average game played by the Nuggets at a home game is 106.2 points. The median of 106.0 represents that the middle game in a list of all games ordered by points is 106.0. The mean being higher than the median shows that there are more games above 106.0 than below. The variance of 147.8 is a measure of how likely it is for the total points to stray from the mean. This is more clearly demonstrated by the standard deviation of 12.6 points. This means that 1 unit of standard deviation on the distribution of points is 12.6. Therefore about 34% of the data falls 12.6 points above the mean. This also indicates that points 25.2 points above or below the mean are considered outliers (2 standard deviations above or below the mean). The distribution is bell shaped, with a mean and median that are almost identical. Either value is a fair measure of central tendency to base statistics upon.

## Descriptive Statistics: Points Scored By Your Team in Away Games

Table 3. Descriptive Statistics for Points Scored by Your Team in Away Games

| **Statistic** | **Value** |
| --- | --- |
| Mean | 101.8 |
| Median | 103.0 |
| Variance | 148.33 |
| Standard Deviation | 12.18 |

* The mean of 101.8 points shows that the average game played by the Nuggets at an away game is 101.8 points. The median of 103.0 points shows that the middle game in an ordered list of points is 103.0 points. The median being higher than the mean shows that there are more games below the median points than above. The variance of 148.33 shows the spread of the data in the set from the measures of central tendency. The standard deviation of 12.18 shows that about 34% of the data is 12.18 points above the mean. This also indicates that outlying data is considered anything farther than 24.36 points from the mean (2 standard deviations away). The distribution of points is slightly left leaning, meaning that more of the games fall below the median.
* After analyzing the data, we can see there is a tendency for the Nuggets to score less at away games than home games. The mean of games at home are 106.2 versus 101.8 when away. The games played away are also more consistent in their slightly lower scores as shown by the lower standard deviation. From this data we can conclude that the Nuggets are likely to score a few points higher when playing at home rather than away.

## Confidence Intervals for the Average Relative Skill of All Teams in Your Team’s Years

Table 4. Confidence Interval for Average Relative Skill of Teams in Your Team’s Years

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| 95% | (1502.02, 1507.18) |

* Confidence intervals are used in estimating the measures of central tendency by showing that, in this case, if a random team was pulled from the dataset, there is a 95% chance that the team’s elo would be between 1502.02 and 1507.18. Confidence intervals are used because some populations are too large or impracticle to calculate the true values of central tendency.
* If the chosen confidence level was higher, than the interval would be larger. This is because in order to fit more teams in the range, you must increase the range. If the confidence level was lower than the interval would be smaller, closing in on the mean.
* Choice 2 (st.norm.cdf is the correct code for finding the liklihood that a team has a relative skill level lower than that of the Nuggets.

## Confidence Intervals for the Average Relative Skill of All Teams in the Assigned Team’s Years

Table 5. Confidence Interval for Average Relative Skill of Teams in Assigned Team’s Years

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| 95% | (1487.66, 1493.65) |

* 17 years prior, the confidence interval is about ~15 points lower. This shows that the average team in 2013-2015 was scoring more points than the average team in 1996-1998. The competition has become more fierce as time has gone on.
* If the chosen confidence level was higher, than the interval would be larger. This is because in order to fit more teams in the range, you must increase the range. If the confidence level was lower than the interval would be smaller, closing in on the mean.

## Conclusion

The practical importance of the analyses that were performed in this paper is that it demonstrates how the Bulls from 1996-1998 compare to the Nuggets from 2013-2015. It also compares the entire NBA’s relative skill levels from the Bull’s years to the Nugget’s years.

We have shown through measures of central tendency that the Nuggets tend to perform higher point games in 2013-2015 than the Bulls did in 1996-1998. The Nuggets are also more likely to have outlying performances.

Through calculating confidence intervals for the entire NBA in both the 1996-1998 time frame and the 2013-2015 time frame, we have shown that the average team is scoring higher as time has gone on. This is also supported tangentially by the nuggets having better statistics specifically than the Bulls in the 90’s.

Another observation we can draw from our data analytics is that the Nuggets perform better at home games than away games over the 3 year span of the data frame.