'15 Celtics v '98 Bulls ELO Summary Report

James Michael
www.lucid-solutions.org

1. Introduction: Problem Statement

For the outset of this report, I am tasked as a data analyst for an NBA basketball team. Equipped with a large set of historical data, I will be working to analyze and find patterns of player behavior and performance, to aid management in making decisions to further improve our team's performance. I will be drawing data from the FiveThirtyEight NBA Elo dataset, provided by Kaggle. During this presentation I will be engaging descriptive statistics and providing data visualizations to support this statistical analysis.

I hope you find this information enlightening. Let us proceed.

2. Introduction: Your Team and the Assigned Team

For the purpose of this report I have selected the Boston Celtics (herein referred to as 'your team'), and draw data from their games set between 2013 – 2015. Likewise, I have been assigned the Chicago Bulls as an opposing team, and will review their games played between 1996 – 1998. Both teams played a total of 246 games across the course of 3 years.

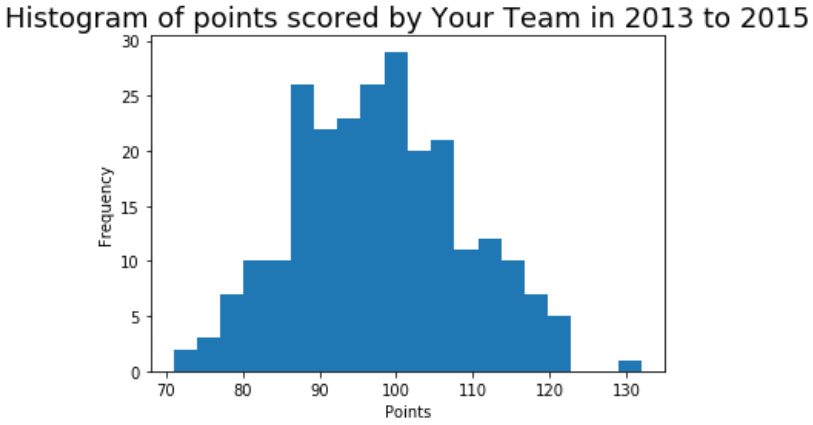
Table 1. Information on the Teams

	Name of Team	Assigned Years
1. Yours	Celtics	2013 - 2015
2. Assigned	Bulls	1996 - 1998

3. Data Visualization: Points Scored by Your Team

Data visualization is a time machine. It enables us to jump ahead, letting us understand hours of statistical analysis in mere moments. By using graphical representations to statistical data, we can quickly observe useful patterns, spot outliers, and determine central tendencies.

Histograms (as attached), are great for studying the frequency distribution of a variable. This allows us to visually assess the concentration and spread of data, revealing frequency of data

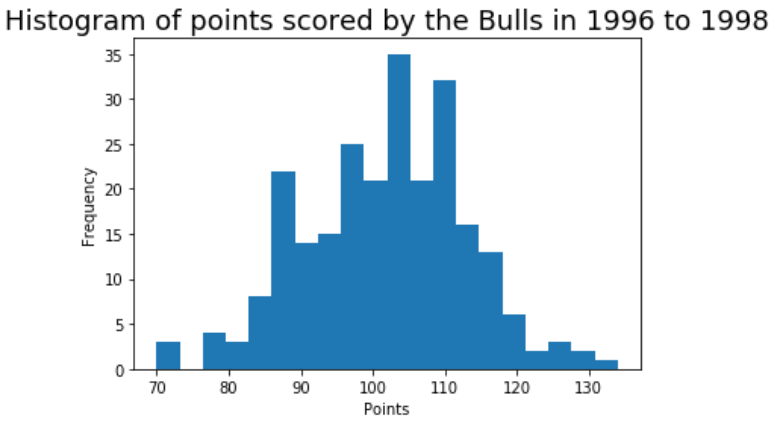


points across various ranges and highlighting the concentration of values, a skew towards one side or another, etcetera.

In this particular, the shape of our distribution suggests a normal distribution, often called a 'bell-curve'. This signifies the points scored but the Celtics averages out to approximately 100 per game, rarely above 130, and never below 70.

4. **Data Visualization: Points Scored by the Assigned Team**

We also analyzed similar data from the Chicago Bulls, over the course of 1996-1998. As our intention with this report is to compare and contrast, we will use another Histogram to represent the assigned team's stats as well. This graph format allows us to visualize the data quickly and accurately, and will set us up for a later comparison to the previous observations.



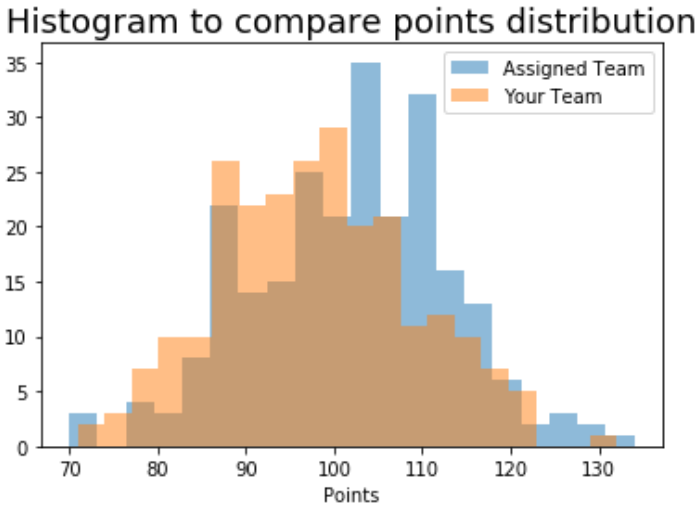
At first glance, we can observe many similarities between the Celtics and the Bulls; a middling bell-curve suggests a normal distribution, with outliers sitting near 70 and 130 points respectively. The Bulls average somewhere between 100 and 110 points per game played, and we see a high frequency of total points. A looser middle grouping suggests greater consistency at

specific point ranges; the Bulls are as likely to sink ~87 points in any given game as they are ~102 or ~108.

5. Data Visualization: Comparing the Two Teams

Data visualization is used to compare two different data distributions by overlaying or placing side-by-side graphical representations, such as our histograms. These visual tools allow for a direct comparison of the central tendencies, spread, and shape of the distributions. By examining the position of the peaks, the width of the spread, and the presence of skewness or outliers, we can quickly assess differences in the distributions, such as whether one dataset has higher or lower values, greater variability, or a different distribution shape. This visual comparison helps to identify trends, differences in data characteristics, and potential relationships between the two datasets.

Below are the histograms for the Celtics and the Bulls, overlaid:



This particular plot was chosen for its ease of readability; while other options may offer a more precise approximation of our datasets, with this we can directly compare and contrast the score potential of both teams at a single glance.

As we see the majority of overlap (brown) suggests a very similar point average between these two teams, however the Chicago Bulls (blue) peaked out more often, and with higher points than the Boston Celtics (orange).

6. Descriptive Statistics: Points Scored By Your Team in Home Games

Measures of central tendency (mean, median) and measures of variability (variance, standard deviation) are essential in analyzing a data distribution because they provide insights into both the "typical" value and the spread of the data. The mean and median help to identify the center or average value of the dataset, while the variance and standard deviation give an understanding of how much the values deviate from the center, indicating the consistency or unpredictability of the data.

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

Statistic	Value
Mean	98.82
Median	99.5
Variance	107.67
Standard Deviation	10.38

The mean (98.82) represents the average points scored by your team in home games, summarizing the overall performance. The median (99.5), which is slightly higher than the mean, suggests a right-skewed distribution. This means there are a few games where your team scored significantly higher points, pulling the mean down while leaving the median relatively unaffected.

The variance (107.67) measures the spread of scores, indicating how much the points deviate from the mean on average. The standard deviation (10.38), the square root of the variance, shows that most of the games' scores deviate by about 10.38 points from the mean, reflecting some variability in performance.

Given that the mean is slightly lower than the median, the distribution is right-skewed, meaning most scores are clustered near the lower end, with a few high-scoring games stretching the tail. In skewed distributions like this, the median is the preferred measure of central tendency, as it is less sensitive to outliers and more accurately represents the typical score.

In this case, the median is a more reliable measure of the "typical" value since it better reflects the center of the distribution without being distorted by extreme high scores. The mean, while useful, can be misleading in the presence of skewness.

7. **Descriptive Statistics: Points Scored By Your Team in Away Games**

Comparatively, let us examine the data of points scored by the Celtics in away games.

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

Statistic Name	**Value**
Mean	97.28
Median	96.0
Variance	121.43
Standard Deviation	11.02

The mean (97.28) represents the average points scored by your team in away games, providing a general sense of the team's performance. However, the mean can be sensitive to outliers, which is evident here since the median (96.0) is slightly lower than the mean. This suggests a right-skewed distribution, where a few high-scoring games pull the mean upwards, while most games are clustered around the median.

The variance (121.43) indicates that the scores deviate significantly from the mean, showing greater variability in away games compared to home games. The standard deviation (11.02)

further highlights this spread, indicating that the scores in away games are more spread out, with some scores being much higher or lower than the mean. This large deviation suggests that away games feature more inconsistency in performance.

Given the right-skewed nature of the distribution, the median is the better measure of central tendency for away games. The median is less sensitive to outliers and provides a more accurate representation of the center of the distribution in this case.

When comparing home games to away games, your team scores slightly more points at home (98.82 vs. 97.28). Additionally, the standard deviation for away games is higher, indicating more variability in away performance. This suggests that the team has a more consistent performance at home, while away games show greater fluctuation in scoring.

8. **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 statistical analysis to estimate the range within which a population parameter, such as the mean, is likely to fall. For example, when estimating the

average relative skill (ELO) of teams over a set of years, a confidence interval provides a range of values within which the true mean is expected to lie, with a certain level of confidence. A 95% confidence interval means we are 95% confident that the true population mean is within the specified range, accounting for variability in the sample. This helps to provide a more reliable estimate than using the sample mean alone, especially when dealing with large datasets.

In this case, the 95% confidence interval for the average relative skill of teams in the selected years is (1502.02, 1507.18), which suggests that the true mean ELO of teams in those years is likely to fall within this range. The interval indicates that, based on the data collected, the average skill level of teams falls between 1502.02 and 1507.18, providing a reasonable estimate of their overall performance. If a different confidence level were used, such as 99%, the interval would likely be wider, reflecting a greater level of certainty that the true mean lies within the range. Conversely, a lower confidence level (e.g., 90%) would result in a narrower interval, offering less certainty but a more precise estimate.

Incidentally, we also calculated the probability any given team in the league would have a lower ELO than the Celtics; this number rounds out to 50%, meaning any given team has a fifty-fifty shot at being in a lower ELO bracket. Which admittedly is not great.

9. 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)

The 95% confidence interval for the average relative skill (ELO) of all teams in the assigned years is (1487.66, 1493.65). This means that we are 95% confident that the true mean ELO of all teams in those years lies within this range. The interval gives us a reliable estimate of the overall skill level of teams, suggesting that their average ELO falls between 1487.66 and 1493.65. This range represents the average performance across the entire league, rather than focusing on a single team, providing insight into the competitive landscape during the assigned years.

If a different confidence level had been used, the interval would change in width. A 99% confidence interval would be wider, indicating a higher level of certainty that the true mean lies within the interval, but with less precision. Conversely, a 90% confidence interval would be narrower, offering more precise estimates but with less certainty. Comparing this confidence interval with the previous one for your team, the average relative skill of all teams (1487.66 to 1493.65) is slightly lower than your team's average (1502.02 to 1507.18), which indicates that your team performed above the average skill level of the league during the selected years. This suggests that, on average, your team’s relative skill was higher compared to the broader league performance during those years.

10. Conclusion

The statistical analyses performed provide valuable insights into the relative skill levels (ELO) of teams over a specific time period. By calculating the confidence intervals for both the average relative skill of all teams and the average skill of a particular team, we can estimate the range within which the true mean ELO is likely to fall. The 95% confidence interval for all teams (1487.66, 1493.65) and for the selected team (1502.02, 1507.18) suggests that the selected team performed above average compared to other teams in the league. The narrower confidence interval for the specific team's ELO highlights greater precision in estimating the team's performance, while the broader interval for all teams reflects greater variability in the league's performance.

The practical importance of these analyses lies in providing a quantitative measure of how a team's performance compares to the league as a whole, allowing coaches, analysts, and decision-makers to assess the team's standing and make data-driven decisions. For example, knowing that your team's average skill is above the league's mean could influence strategy and expectations. Additionally, the confidence interval helps quantify the level of certainty in these estimates, which is crucial when making predictions about future performance. By understanding these statistical concepts, teams can better allocate resources, tailor their strategies, and set realistic goals based on both individual and collective performance trends.