# MAT 243 Project One Summary Report

Fabian Rodriguez

Fabian.rodriguez@snhu.edu

Southern New Hampshire University

## Introduction: Problem Statement

The following analysis is based on an NBA data starting on the 1950 until 2020. The data from this dataset will be analyzed using descriptive statistics and data visualization techniques to study distributions of key variables associated with the performance of different teams. Statistical methods such as mean, median, variance, standard deviation and more will be used to better understand the performance of the selected teams.

## Introduction: Your Team and the Assigned Team

Table 1. Information on the Teams

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

The Lakers is the team I chose for this comparison. The assigned years for the Lakers ranges from 2013 – 2015. In contrast, we have the Bulls ranging from 1996 – 1998 as the assigned team.

## Data Visualization: Points Scored by Your Team

Data visualization is a method of representing data in a graphical format to help people see and understand the information. There are many ways that data visualization can be used to study data distributions and trends. One way is by using graphs, charts, or tables to help people visualize the information. Another way is by using interactive graphics to allow people to explore the data in more detail. Data visualization helps us comprehend the information that we are looking at more easily because it makes it easier for us to see patterns and trends in the data. It also helps make it clear when there are inconsistencies in the data or when something seems off about it.

For the team I selected, I chose the histogram as the plot. Histograms are a graphical representation of the distribution of data. They are used to visualize the distribution of data and how it changes over time. Histograms are useful for understanding how data is distributed and can be used for many purposes. They can be used to find outliers, measure variability, or compare distributions. Histograms are easier to read than bar charts because they use two dimensions instead of three dimensions. This means that we can see more information at once in one glance.

The histogram allows us to easily see that the highest point frequency is at 100 points, we can easily deduce this by looking at the center of the histogram. Although there are a few outliers.

Chart, histogram

Description automatically generated

## Data Visualization: Points Scored by the Assigned Team

To keep our data visualization uniform, we also used a histogram. Thus, we can easily identify that the highest frequency is around 104 points.

Chart, histogram

Description automatically generated

## Data Visualization: Comparing the Two Teams

Data visualization can be used to compare two different data distributions by representing the data in a visual format. Data visualization helps us understand how one distribution compares with another distribution by showing us how the data is distributed. Data visualization can also show us an overview of the entire dataset at once instead of showing individual values for each piece of information. This makes it easier for viewers to see patterns or trends in the dataset without having to look at every single piece of information separately.

The plot selected for this section is the Box Plot. A box plot is a data visualization that uses a box and several lines to depict the distribution of data in a dataset. A box spans the middle 50% of the data, with Q1 as the lower boundary of the box and Q3 as the upper boundary of the box. The median is shown as a line inside the box. Two lines, known as whiskers, extend from the lower boundary of the box to the minimum and from the upper boundary of the box to the maximum. The whiskers represent the lower and upper 25% of the data.

Using the Box plot we can see that the median is really close at 102, although the Q3 for the Bulls is better than the Lakers by a few points. Something else we can clearly see is that the Bulls have more outliers than the Bulls.

Chart, box and whisker chart

Description automatically generated

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

| **Type** | **Value** |
| --- | --- |
| Statistic  (for example, Mean) | X.XX  \*Round off to 2 decimal places. |
| Mean | 101.7 |
| Median | 102 |
| Variance | 149.18 |
| Standard Deviation | 12.21 |

There are four main measures of central tendency that are used to analyze a data distribution. These include the mean, median, mode, and weighted mean. The most used measure of central tendency is the mean. This is because it is easy to calculate and does not require any special calculations or knowledge about the data set.

The skew for this data is bell shaped. The shape in this case is due to the mean and the median beign very close.

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

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

| **Statistic Name** | **Value** |
| --- | --- |
| Statistic  (for example, Mean) | X.XX  \*Round off to 2 decimal places. |
| Mean | 100.71 |
| Median | 101.0 |
| Variance | 88.16 |
| Standard Deviation | 9.39 |

Mean and average can be used interchangeably. Quoted from Investopedia, “**The** **median** **is** **the** middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average”. The variance and the. Standard deviation allows us to see how far some data points are from the mean, thus giving us insight on the outliers.

Like the team we selected the skew for this data is also bell shaped, because the mean and the median are quite close. However, the variance and the standard deviation are lower than the team we selected. This information highlights the consistency of the Bulls performance.

## 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 generally used in estimating the measures of central tendency for a population. The interval is usually calculated by using the sample statistic and the population parameter. Confidence intervals are a measure of the degree of uncertainty in an estimate. They are used to help us make decisions about whether we can be confident that a particular value is close to the true population parameter. The probability that a given team in the league has a relative skill level less than that of the team is .2853. It is not unusual that a team has a skill level less than your team.

## 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) |

Considering all the data reviewed previously it’s clear that the Bulls performed better than the Lakers. The Bulls had a probability of being better than any other team 97 out of 100 times. The previous statement indicates that percentage wise the chance of a team being better than the Bulls is 3 %. If we compare it to the Lakers, which is coming in at a 28% there is quite the difference, showcasing the Bulls as the more performant team.

## Conclusion

The importance of the analyses we just performed can easily help us understand the performance of multiple different teams based on a large dataset. We used massive amounts of data and convert it to an easy to understand data visualization. The results of this scenario demonstrated that the Bulls was the suprior team in our comparison.

## Citations

Ganti, A. (2022, September 11). Median: What it is and how to calculate it, with examples. Investopedia. Retrieved September 25, 2022, from https://www.investopedia.com/terms/m/median.asp