MAT 243 Project One Summary Report

Traver Yates

traver.yates@snhu.edu

Southern New Hampshire University

What is the problem you are going to solve?

What data set are you using?

What statistical methods will you be using to do the analysis for this project?

The problem that is being solved is analyzing a dataset of basketball games to understand the factors that contribute to a team winning or losing. The data set being used is a collection of 246 games from the 1996 NBA season, which includes information such as the teams playing, the scores, the location of the game, and the outcome of the game. To perform the analysis, statistical methods such as linear regression and logistic regression will be used to examine the relationship between different variables and the game's outcome. Additionally, data visualization techniques will be used to explore and understand the data.

|  | **game\_id** | **year\_id** | **fran\_id** | **pts** | **opp\_pts** | **elo\_n** | **opp\_elo\_n** | **game\_location** | **game\_result** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 199511030CHI | 1996 | Bulls | 105 | 91 | 1598.2924 | 1531.7449 | H | W |
| **1** | 199511040CHI | 1996 | Bulls | 107 | 85 | 1604.3940 | 1458.6415 | H | W |
| **2** | 199511070CHI | 1996 | Bulls | 117 | 108 | 1605.7983 | 1310.9349 | H | W |
| **3** | 199511090CLE | 1996 | Bulls | 106 | 88 | 1618.8701 | 1452.8268 | A | W |
| **4** | 199511110CHI | 1996 | Bulls | 110 | 106 | 1621.1591 | 1490.2861 | H | W |

printed only the first five observations...

Number of rows in the data set = 246

*In this project, you picked a team and you were assigned a team to do comparative analysis.*

What team did you pick and what years were picked to do the analysis?

What team and range of years were you assigned for the comparative study? (Hint: This is called the assigned team in the Python script.) Present this information in a formatted table as shown below.

The table shows that the team that was picked for the analysis is the Heat and the years picked for the analysis are 2013 - 2015, while the team assigned for the comparative study is the Bulls and the range of years is 1996 - 1998. The data provided in the table corresponds to the years and teams that were mentioned in the table 1.

This information can be used to conduct a comparative analysis between the performance of the Heat during the years 2013-2015 and that of the Bulls during the years 1996-1998. This can be done by comparing various statistics such as points scored, points allowed, win-loss ratio, etc. Additionally, Data visualization techniques can also be used to explore and understand the data.

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

|  |  | **game\_id** | **year\_id** | **fran\_id** | **pts** | **opp\_pts** | **elo\_n** | **opp\_elo\_n** | **game\_location** | **game\_result** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** |  | 199511030CHI | 1996 | Bulls | 105 | 91 | 1598.2924 | 1531.7449 | H | W |
| **1** |  | 199511040CHI | 1996 | Bulls | 107 | 85 | 1604.3940 | 1458.6415 | H | W |
| **2** |  | 199511070CHI | 1996 | Bulls | 117 | 108 | 1605.7983 | 1310.9349 | H | W |
| **3** |  | 199511090CLE | 1996 | Bulls | 106 | 88 | 1618.8701 | 1452.8268 | A | W |
| **4** |  | 199511110CHI | 1996 | Bulls | 110 | 106 | 1621.1591 | 1490.2861 | H | W |

printed only the first five observations...

Number of rows in the data set = 246

Data Visualization: Points Scored by Your Team

In the Python script, you created a visualization for the distribution of points scored by your team.

See Step 3 in the Python script to address the following items in a paragraph response:

In general, how is data visualization used to study data distributions and trends?

In this activity, you were asked to pick one of the two plots that best describes the data distribution of the variable for your team. Include a screenshot of this plot in your report.

Why did you pick this plot? Explain.

What can you say about the distribution of the variable by visually inspecting this plot? What does this signify?

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include Python code in your report.*

Data visualization is a powerful tool for studying data distributions and trends. By creating visual representations of data, we can quickly identify patterns, outliers, and other important features of the data. In this activity, I picked a histogram as the best plot to describe the data distribution of the variable. I chose this plot because it is a simple and effective way to show the distribution of a variable and how often certain values occur. By visually inspecting the histogram, we can see that the distribution of the variable is roughly symmetric, with most of the data points concentrated in the middle of the range. This signifies that there is no clear bias in the data, and that it is likely to be a representative sample. Additionally, the histogram allows us to see any outliers, skewness and kurtosis in the data.

| **game\_id** | **year\_id** | **fran\_id** | **pts** | **opp\_pts** | **elo\_n** | **opp\_elo\_n** | **game\_location** | **game\_result** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 201210300MIA | 2013 | Heat | 120 | 107 | 1666.3193 | 1586.1121 | H | W |
| **1** | 201211020NYK | 2013 | Heat | 84 | 104 | 1647.6675 | 1548.2699 | A | L |
| **2** | 201211030MIA | 2013 | Heat | 119 | 116 | 1650.0934 | 1554.4674 | H | W |
| **3** | 201211050MIA | 2013 | Heat | 124 | 99 | 1656.5652 | 1504.0280 | H | W |
| **4** | 201211070MIA | 2013 | Heat | 103 | 73 | 1659.7239 | 1361.5804 | H | W |

printed only the first five observations...

Number of rows in the data set = 246

Data Visualization: Points Scored by the Assigned Team

*In the Python script, you created a visualization for the distribution of points scored by the assigned team.*

*See Step 4 in the Python script to address the following items in a paragraph response:*

In this activity, you were asked to pick one of the two plots that best describes the data distribution of the variable for the assigned team. Include this plot in your report.

Why did you pick this plot? Explain.

What can you say about the distribution of the variable by visually inspecting this plot? What does this signify?

*Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include Python code in your report.*

I

n this activity, I picked a boxplot as the best plot to describe the data distribution of the variable for the assigned team. I chose this plot because it is a compact and informative way to show the distribution of a variable and its basic statistical properties, such as the median, quartiles, and outliers.

By visually inspecting the boxplot, we can see that the distribution of the variable is roughly symmetric and has no outliers. The median, represented by the line inside the box, is near the center of the distribution and the quartiles, represented by the top and bottom of the box, show that the data is tightly distributed. This signifies that the variable is likely to be a representative sample with no extreme values. Additionally, we can see that the distribution of the variable is not skewed to the left or right, indicating a normal distribution.

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Data Visualization: Comparing the Two Teams

*In the Python script, you created a visualization for the difference in the distributions of points scored by your team and the assigned team.*

*See Step 5 in the Python script to address the following items in a paragraph response:*

In general, how is data visualization used to compare two different data distributions?

In this activity, you were asked to pick one of the two plots that best compares the data distributions of your team with the assigned team. Include a screenshot of this plot in your report.

Why did you pick this plot? Explain.

How do the two distributions compare to each other?

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include Python code in your report.*

In general, data visualization is used to compare two different data distributions by creating visual representations of the data, such as bar charts, line plots, and scatter plots. These types of plots allow us to quickly identify patterns, outliers, and other important features of the data, and to compare them with other data sets.

In this activity, I picked a side-by-side boxplot as the best plot to compare the data distributions of my team with the assigned team. I chose this plot because it is a compact and informative way to show the distribution of a variable and its basic statistical properties, such as the median, quartiles, and outliers, for two different data sets on a single graph.

By visually inspecting the side-by-side boxplots, we can see that the two distributions are similar in shape and spread, but the assigned team has a slightly higher median and more outliers. This signifies that the two teams are similar in terms of performance, but the assigned team has a slightly better performance with some extreme values.

*Chart, box and whisker chart

Description automatically generated*

*Chart, histogram

Description automatically generated*

Descriptive Statistics: Points Scored By Your Team in Home Games

*In the Python script, you calculated descriptive statistics on the points scored by your team in games played at home venue. These included the mean, median, variance, and standard deviation for the relative skill of your team.*

*See Step 6 in the Python script to address the following items:*

Summarize **all** statistics in a formatted table as shown below. Use one row for each statistic. You will need to add rows to the table in order to include all of your statistics.

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

| **-** | **Value** |
| --- | --- |
| Statistic  *(for example, Mean)* | X.XX  *\*Round off to 2 decimal places.* |

In general, how are the measures of central tendency and variability used to analyze a data distribution?

Interpret each statistic in detail and explain what it represents in this scenario.

Use the mean and the median to describe the distribution of points scored by your team in home games.

Describe the skew: Is it left, right, or bell-shaped?

Explain which measure of central tendency is best to use to represent the center of the distribution based on its skew.

* Answer the questions in a paragraph response. Remove all questions and this note (but not the table) before submitting! Do not include Python code in your report.*

In general, measures of central tendency and variability are used to analyze a data distribution. Measures of central tendency, such as mean, median and mode, provide a single value that represents the center of the distribution. Measures of variability, such as range, variance, and standard deviation, provide information about the spread of the data.

The Mean is the average of the data set and in this scenario, it is 102.56. It represents the center of the distribution by considering all the values of the data set.

The Median is the middle value of the data set and in this scenario, it is 103.0. It represents the center of the distribution by dividing the data set into two equal parts.

The Variance is a measure of the spread of the data set, representing how far each number in the set is from the mean, and is calculated as 126.05.

The Standard deviation is the square root of the variance, a measure of the spread of the data set, and it is calculated as 11.23.

The mean and the median are both measures of central tendency and are used to describe the distribution of points scored by your team in home games. The mean of 102.56 and median of 103.0 are close to each other, showing that the data is symmetric and bell-shaped. This indicates that the data is evenly distributed around the center.

The best measure of central tendency to use to represent the center of the distribution based on its skew is the median. Since the data is symmetric and bell-shaped, the median is the best measure of central tendency as it is not affected by outliers or extreme values, it is more robust and provide a better representation of the central tendency of the data.

Statistic Value

Mean 102.56

Median 103.0

Variance 126.05

Standard Deviation 11.23

Points Scored by Your Team in Home Games (2013 to 2015)

-------------------------------------------------------

Mean = 102.56

Median = 103.0

Variance = 126.05

Standard Deviation = 11.23

Descriptive Statistics: Points Scored By Your Team in Away Games

*In the Python script, you calculated descriptive statistics on the points scored by your team in games played at opponent’s venue (away). These included the mean, median, variance, and standard deviation for the relative skill of the assigned team.*

*See Step 7 in the Python script to address the following items:*

Summarize **all** statistics in a formatted table as shown below. Use one row for each statistic. You will need to add rows to the table in order to include all of your statistics.

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.* |

Interpret each statistic in detail and explain what it represents in this scenario.

Use the mean and the median to describe the distribution of points scored by your team in away games.

Describe the skew: Is it left, right, or bell-shaped?

Explain which measure of central tendency is best to use to represent the center of the distribution based on its skew.

Is your team performing better in games played at home than those played away? Use the mean and the standard deviation to answer this question. What can be deduced by comparing the standard deviation of points scored in home games and points scored in away games?

*Answer the questions in a paragraph response. Remove all questions and this note (but not the table) before submitting! Do not include Python code in your report.*

Points Scored by Your Team in Away Games (2013 to 2015)

-------------------------------------------------------

Mean = 97.28

Median = 97.0

Variance = 101.4

Standard Deviation = 10.07

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

*In the Python script, you calculated a 95% confidence interval for the average relative skill of all teams in the league during the years of your team. Additionally, you calculated the probability that a given team in the league has a relative skill level less than that of the team that you picked.*

*See Step 8 in the Python script to address the following items:*

Report the confidence interval in a formatted table as shown below.

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

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| XX% (for example, 95%) | (X.XX, X.XX)  *\*Round off to 2 decimal places.* |

Describe how confidence intervals are generally used in estimating the measures of central tendency for a population.

Provide a detailed interpretation of the confidence interval in terms of the average relative skill of teams in the range of years that you picked.

How would your interval be different if you had used a different confidence level?

What is the probability that a given team in the league has a relative skill level less than that of the team that you picked? Is it unusual that a team has a skill level less than your team?

*Answer the questions in a paragraph response. Remove all questions and this note (but not the table) before submitting! Do not include Python code in your report.*

Confidence intervals are generally used in estimating the measures of central tendency for a population. They provide a range of values that are likely to contain the true population mean or median, with a certain level of confidence. The level of confidence is usually set at 95% or 99%, meaning that if the process of sampling and estimation is repeated many times, the true population mean or median would be included in the calculated interval in 95% or 99% of the cases.

In this scenario, the 95% confidence interval for the average relative skill (ELO) of teams in the years 2013 to 2015 is (1502.02, 1507.18). This means that we are 95% confident that the true average relative skill of teams in the league lies between 1502.02 and 1507.18 in the years 2013 to 2015.

If a different confidence level had been used, the interval would have been different. For example, if a 99% confidence level was used, the interval would be wider, resulting in more uncertainty around the true population mean.

The probability that a given team in the league has a relative skill level less than that of the team that you picked is 0.841. It is not unusual that a team has a skill level less than your team as the probability is relatively high.

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

Confidence Interval for Average Relative Skill in the years 2013 to 2015

------------------------------------------------------------------------------------------------------------

95% confidence interval (unrounded) for Average Relative Skill (ELO) in the years 2013 to 2015 = (1502.0236894390478, 1507.1824625533618)

95% confidence interval (rounded) for Average Relative Skill (ELO) in the years 2013 to 2015 = ( 1502.02 , 1507.18 )

Probability a team has Average Relative Skill LESS than the Average Relative Skill (ELO) of your team in the years 2013 to 2015

----------------------------------------------------------------------------------------------------------------------------------------------------------

Which of the two choices is correct?

Choice 1 = 0.159

Choice 2 = 0.841

Choice #2 is correct

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

*In the Python script, you calculated a 95% confidence interval for the average relative skill of all teams in the league during the years of the assigned team. Additionally, you calculated the probability that a given team in the league has a relative skill level less than that of the assigned team.*

*See Step 9 in the Python script to address the following items:*

Report the confidence interval in a formatted table as shown below.

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

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| XX% (for example, 95%) | (X.XX, X.XX)  *\*Round off to 2 decimal places.* |

Provide a detailed interpretation of the confidence interval in terms of the average relative skill of teams in the assigned team’s range of years.

Discuss how your interval would be different if you had used a different confidence level.

How does this confidence interval compare with the previous one? What does this signify in terms of the average relative skill of teams in the range of years that you picked versus the average relative skill of teams in the assigned team’s range of years?

* Answer the questions in a paragraph response. Remove all questions and this note (but not the table) before submitting! Do not include Python code in your report.*

The 95% confidence interval for the average relative skill (ELO) of teams in the years 1996 to 1998 is (1487.71, 1493.6). This means that we are 95% confident that the true average relative skill of teams in the league lies between 1487.71 and 1493.6 in the years 1996 to 1998.

If a different confidence level had been used, the interval would have been different. For example, if a 99% confidence level was used, the interval would be wider, resulting in more uncertainty around the true population mean.

When compared to the previous confidence interval, this interval is narrower. This signifies that the relative skill of teams in the range of years 1996 to 1998 is more consistent than the average relative skill of teams in the range of years 2013 to 2015. This could be due to a number of factors, such as differences in the overall competitiveness of the league, changes in the rules or coaching strategies, or other external factors.

|  |  |
| --- | --- |
| Confidence Level (%) | Confidence Interval |
| 95% | (1487.71, 1493.6) |

Confidence Interval for Average Relative Skill in the years 1996 to 1998

------------------------------------------------------------------------------------------------------------

95% confidence interval (unrounded) for Average Relative Skill (ELO) in the years 2013 to 2015 = (1487.7065262875703, 1493.596609849239)

95% confidence interval (rounded) for Average Relative Skill (ELO) in the years 2013 to 2015 = ( 1487.71 , 1493.6 )

Probability a team has Average Relative Skill LESS than the Average Relative Skill (ELO) of your team in the years 2013 to 2015

----------------------------------------------------------------------------------------------------------------------------------------------------------

Which of the two choices is correct?

Choice 1 = 0.1629

Choice 2 = 0.8371

Number 2 is correct

Conclusion

*Describe the results of your statistical analyses clearly, using proper descriptions of statistical terms and concepts.*

What is the practical importance of the analyses that were performed?

Describe what these results mean for the scenario.

*Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include Python code in your report.*

The statistical analyses that were performed in this scenario provided information about the distribution of points scored by a team in home games, the comparison of two different data distributions, and the estimation of the measures of central tendency for a population.

The practical importance of these analyses is that they allow us to better understand the performance of a team and to identify patterns, outliers, and other important features of the data. By understanding the distribution of points scored by a team in home games, we can identify trends and make predictions about future performance. By comparing two different data distributions, we can identify similarities and differences between teams and make inferences about their relative performance. By estimating the measures of central tendency for a population, we can make inferences about the overall performance of teams in a league over a certain period.

In terms of the scenario, these results mean that we can identify patterns in the performance of the team in home games and make predictions about future performance. They also indicate that the team is similar in terms of performance to the assigned team, but the assigned team has a slightly better performance with some extreme values. Furthermore, the results indicate that the relative skill of teams in the league in the range of years 1996 to 1998 is more consistent than the average relative skill of teams in the range of years 2013 to 2015. This could be due to several factors, such as differences in the overall competitiveness of the league, changes in the rules or coaching strategies, or other external factors. Overall, these results provide a better understanding of the performance of teams in the league and allow for more informed decision making and predictions about future performance.

Citations

*You were* ***not*** *required to use external resources for this report. If you did not use any resources, you should remove this entire section. However, if you did use any resources to help you with your interpretation, you* ***must*** *cite them. Use proper APA format for citations.*

Insert references here in the following format:

Author's Last Name, First Initial. Middle Initial. (Year of Publication). Title of book: Subtitle of book, edition. Place of Publication: Publisher.