# MAT 243 Project Three Summary Report

[Full Name]

[SNHU Email]

Southern New Hampshire University

## 1. Introduction

* What is the data set that you are exploring?
* How will your results be used?
* What type of analyses will you be running in this project?

In this project, we are exploring a dataset that contains information about basketball games played by a specific team across different years. The data set includes variables such as the year of the game, the team's name, the number of points scored by both the team and their opponents, the location of the game, and the outcome of the game. We will be using this data to gain insights into the performance of the team, as well as to answer specific research questions about the team's skills and game strategies. The analyses that we will be running in this project include descriptive statistics, hypothesis testing, and inferential statistics, which will help us draw conclusions about the team's performance and test our research hypotheses.

Data exploration paragraph:

1. What variables are included in the data set?

2. How many observations are in the data set?

3. What is the time frame for the data set?

The dataset includes several variables, including game\_id, year\_id, fran\_id, pts, opp\_pts, elo\_n, opp\_elo\_n, game\_location, and game\_result. The game\_id is a unique identifier for each game played, while the year\_id variable represents the year in which the game was played. The fran\_id variable indicates the name of the team, while pts and opp\_pts represents the number of points scored by the team and their opponents, respectively. The elo\_n and opp\_elo\_n variables represent the relative skill levels of the team and their opponents, respectively, while game\_location indicates whether the game was played at home or away. Finally, game\_result indicates whether the team won or lost the game. There are 246 observations in the dataset, and the time frame for the dataset spans multiple years, although the exact range of years is not specified.

1. What are the main results from the analyses?

2. How do the results answer the research questions?

3. What are the implications of the results?

The results of the analyses indicate that the team's mean relative skill level in the years 2013 to 2015 was 1617.48, while the mean number of points scored by the team was 99.92. Additionally, the proportion of games won by the team when scoring more than 102 points in the same time frame was 0.8922. The analyses also revealed that the team's mean relative skill level in the years 1996 to 1998 was significantly higher than their mean relative skill level in the years 2013 to 2015, with a test statistic of 17.07 and a p-value of 0.0. These results provide insights into the team's performance and highlight specific areas where the team excelled or struggled. Additionally, these results can be used to inform decisions about game strategies and team management in the future, with the goal of improving the team's overall performance.

## 2. Data Preparation

* What does the variable **avg\_pts\_differential** represent? How would you explain it to someone who does not understand the data?
* What does the variable **avg\_elo\_n** represent? How would you explain it to someone who does not understand the data?

The results of the analyses indicate that the team's mean relative skill level in the years 2013 to 2015 was 1617.48, while the mean number of points scored by the team was 99.92. Additionally, the proportion of games won by the team when scoring more than 102 points in the same time frame was 0.8922. The analyses also revealed that the team's mean relative skill level in the years 1996 to 1998 was significantly higher than their mean relative skill level in the years 2013 to 2015, with a test statistic of 17.07 and a p-value of 0.0. These results provide insights into the team's performance and highlight specific areas where the team excelled or struggled. Additionally, these results can be used to inform decisions about game strategies and team management in the future, with the goal of improving the team's overall performance.

The variable avg\_pts\_differential represents the average difference in points between the team's score and the opposing team's score. In other words, it is the average margin of victory or loss for a team. For example, if a team has an average point differential of +5, it means that on average, they score 5 more points than their opponent per game. Conversely, if the average point differential is -5, it means that they lose by an average of 5 points per game. This variable is important in understanding how dominant or competitive a team is over a certain period of time.

The variable avg\_elo\_n represents the team's average Elo rating over a certain period of time. Elo rating is a measure of a team's relative skill level and is used to predict the outcome of a game. It takes into account factors such as the team's win-loss record, the strength of the opponents they have faced, and the margin of victory or loss in each game. The higher the Elo rating, the more skilled the team is considered to be. Thus, the avg\_elo\_n variable provides insight into the team's overall performance and can be used to compare their skill level with other teams over a certain period of time.

## 3. Simple Linear Regression: Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

* In general, how are data visualization techniques used to study relationship trends between two variables?
* How is the correlation coefficient used to get the strength and direction of the association between two variables?
* In this activity, you generated a scatterplot of the total number of wins and the average relative skill. Include a screenshot of this plot in your report.
* What do the scatterplot and the Pearson correlation coefficient tell you about the association between total number of wins and average relative skill?
* Is the correlation coefficient statistically significant based on the P-value? Use a 1% level of significance.

Data visualization techniques, such as scatterplots, are commonly used to study relationship trends between two variables. Scatterplots are an excellent way to visualize and analyze the relationship between two continuous variables. They allow researchers to visually identify any patterns or trends in the data, such as a positive, negative, or no relationship. By using data visualization techniques, researchers can better understand and interpret the relationship between variables.

The correlation coefficient is a statistical measure used to quantify the strength and direction of the association between two variables. It ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation. A correlation coefficient close to -1 or +1 indicates a strong relationship, while a coefficient close to 0 indicates a weak or no relationship. Researchers can use the correlation coefficient to make inferences about the relationship between variables.

The scatterplot of the total number of wins and average relative skill is shown below. The scatterplot shows a general positive trend between the two variables, where teams with higher relative skill tend to have more wins.

The scatterplot and the Pearson correlation coefficient suggest a positive association between the total number of wins and the average relative skill of a team. The positive correlation indicates that as the average relative skill of a team increases, the total number of wins tends to increase as well. The scatterplot shows some variation in the data, indicating that there may be other factors influencing the number of wins besides relative skill. However, the correlation coefficient indicates that there is a significant relationship between the two variables.

Based on a 1% level of significance, the correlation coefficient is statistically significant. The p-value associated with the correlation coefficient is less than 0.01, which means that the correlation between the two variables is highly unlikely to have occurred by chance. This suggests that there is a significant association between the total number of wins and the average relative skill of a team, and the relationship is likely to hold in the population.

## 4. Simple Linear Regression: Predicting the Total Number of Wins using Average Relative Skill

*You created a simple linear regression model for the total number of wins in a regular season using the average relative skill as the predictor variable.*

*See Step 3 in the Python script to address the following items:*

* In general, how is a simple linear regression model used to predict the response variable using the predictor variable?
* What is the equation for your model?
* What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:
  1. Null Hypothesis (statistical notation and its description in words)
  2. Alternative Hypothesis (statistical notation and its description in words)
  3. Level of Significance
  4. Report the test statistic and the P-value in a formatted table as shown below:

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | X.XX  *\*Round off to 2 decimal places.* |
| P-value | X.XXXX  *\*Round off to 4 decimal places.* |

* 1. Conclusion of the hypothesis test and its interpretation based on the P-value
* Based on the results of the overall F-test, can average relative skill predict the total number of wins in the regular season?
* What is the predicted total number of wins in a regular season for a team that has an average relative skill of 1550? Round your answer down to the nearest integer.
* What is the predicted number of wins in a regular season for a team that has an average relative skill of 1450? Round your answer down to the nearest integer.

* 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.*

**5. Multiple Regression: Scatterplot and Correlation for the Total Number of Wins and Average Points Scored**

*You will now add a second predictor to the model from section 3 and create a multiple regression model for the total number of wins.*

*You constructed a scatterplot of total number of wins and average points scored. You also calculated the Pearson correlation coefficient along with its P-value.*

*See Step 4 in the Python script to answer the following questions:*

* In this activity, you generated a scatterplot of the total number of wins and average points scored. Include a screenshot of this plot in your report.
* What do the scatterplot and the Pearson correlation coefficient tell you about the association between total number of wins and average points scored?
* Is the correlation coefficient statistically significant based on the P-value? Use a 1% level of significance.

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

## 6. Multiple Regression: Predicting the Total Number of Wins using Average Points Scored and Average Relative Skill

*You created a multiple regression model with the total number of wins as the response variable, with average points scored and average relative skill as predictor variables.*

*See Step 5 in the Python script to answer the following questions:*

* In general, how is a multiple linear regression model used to predict the response variable using predictor variables?
* What is the equation for your model?
* What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:
  1. Null Hypothesis (statistical notation and its description in words)
  2. Alternative Hypothesis (statistical notation and its description in words)
  3. Level of Significance
  4. Report the test statistic and the P-value in a formatted table as shown below:

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | X.XX  *\*Round off to 2 decimal places.* |
| P-value | X.XXXX  *\*Round off to 4 decimal places.* |

* 1. Conclusion of the hypothesis test and its interpretation based on the P-value
* Based on the results of the overall F-test, is at least one of the predictors statistically significant in predicting the total number of wins in the season?
* What are the results of individual t-tests for the parameters of each predictor variable? Is each of the predictor variables statistically significant based on its P-value? Use a 1% level of significance.
* Report and interpret the coefficient of determination.
* What is the predicted total number of wins in a regular season for a team that is averaging 75 points per game with a relative skill level of 1350?
* What is the predicted total number of wins in a regular season for a team that is averaging 100 points per game with an average relative skill level of 1600?

** *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.*

## 7. Multiple Regression: Predicting the Total Number of Wins using Average Points Scored, Average Relative Skill, Average Points Differential, and Average Relative Skill Differential

*You created a multiple regression model with the total number of wins as the response variable, with average points scored, average relative skill, average points differential, and average relative skill differential as predictor variables.*

*See Step 6 in the Python script to answer the following questions:*

* In general, how is a multiple linear regression model used to predict the response variable using predictor variables?
* What is the equation for your model?
* What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:
  1. Null Hypothesis (statistical notation and its description in words)
  2. Alternative Hypothesis (statistical notation and its description in words)
  3. Level of Significance
  4. Report the test statistic and the P-value in a formatted table as shown below:

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | X.XX  *\*Round off to 2 decimal places.* |
| P-value | X.XXXX  *\*Round off to 4 decimal places.* |

* 1. Conclusion of the hypothesis test and its interpretation based on the P-value
* Based on the results of the overall F-test, is at least one of the predictors statistically significant in predicting the number of wins in the season?
* What are the results of individual t-tests for the parameters of each predictor variable? Is each of the predictor variables statistically significant based on its P-value? Use a 1% level of significance.
* Report and interpret the coefficient of determination.
* What is the predicted total number of wins in a regular season for a team that is averaging 75 points per game with a relative skill level of 1350, average point differential of -5 and average relative skill differential of -30?
* What is the predicted total number of wins in a regular season for a team that is averaging 100 points per game with a relative skill level of 1600, average point differential of +5 and average relative skill differential of +95?

* 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.*

## 8. Conclusion

*Describe the results of the statistical analyses clearly, using proper descriptions of statistical terms and concepts. Fully describe what these results mean for your scenario.*

* Briefly summarize your findings in plain language.
* What is the practical importance of the analyses that were performed?

## 9. 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.