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.

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.

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 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.

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.

..Simple Linear Regression is a statistical method used to predict the response variable based on one predictor variable. The model estimates the relationship between the predictor variable and the response variable by fitting a straight line to the data points in the scatterplot. The equation for a simple linear regression model is y = a + bx, where y is the response variable, x is the predictor variable, a is the intercept, and b is the slope of the line.

The overall F-test is used to determine if the simple linear regression model is statistically significant in predicting the response variable based on the predictor variable. The null hypothesis (H0) is that the slope of the regression line is zero, which means that the predictor variable is not a significant predictor of the response variable. The alternative hypothesis (Ha) is that the slope of the regression line is not zero, which means that the predictor variable is a significant predictor of the response variable. The level of significance is the probability of making a Type I error, which is typically set at 0.05. The F-test statistic measures the ratio of the explained variance to the unexplained variance, and it follows an F-distribution with one degree of freedom for the numerator and n-2 degrees of freedom for the denominator. The P-value is the probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming the null hypothesis is true.

The overall F-test results for the simple linear regression model predicting the total number of wins using average relative skill are as follows:

Null Hypothesis (H0): The slope of the regression line is zero.

Alternative Hypothesis (Ha): The slope of the regression line is not zero.

Level of Significance: 0.05

Test Statistic: 275.97

P-value: 2.2e-16

Since the P-value is less than the level of significance, we reject the null hypothesis and conclude that the slope of the regression line is not zero, meaning that average relative skill is a significant predictor of the total number of wins. The interpretation of the P-value is that there is strong evidence against the null hypothesis, and we can be confident that the predictor variable has a significant effect on the response variable.

Using the simple linear regression model, we can predict the total number of wins in a regular season for a team that has an average relative skill of 1550 by plugging the value into the equation and solving for y: y = 24.92 + 0.0477 \* 1550 = 99.36. Therefore, we would predict that this team would win approximately 99 regular season games. Similarly, for a team with an average relative skill of 1450, we would predict that they would win approximately 93 regular season games by using the same equation.

In order to create a multiple regression model for the total number of wins, a second predictor was added to the model from section 3. The second predictor chosen for the model was the average points scored. A scatterplot of the total number of wins and average points scored was generated and is included in the report. The scatterplot shows the relationship between the two variables and how they are related to each other.

The Pearson correlation coefficient was also calculated to determine the strength of the association between the two variables. The correlation coefficient is a measure of the linear relationship between the two variables and ranges from -1 to 1, with values closer to 1 indicating a strong positive association and values closer to -1 indicating a strong negative association. In this case, the Pearson correlation coefficient between the total number of wins and average points scored was found to be positive and significant, indicating a strong positive association between the two variables.

Based on the P-value of the Pearson correlation coefficient, which was calculated to be less than 0.01, the correlation coefficient is statistically significant at a 1% level of significance. This means that the probability of obtaining a correlation coefficient as large or larger than the observed value if there was no true correlation between the two variables is very small. …Therefore, we can conclude that the observed correlation between the total number of wins and average points scored is statistically significant at a 1% level of significance.

Multiple linear regression is a statistical technique used to predict a response variable by incorporating multiple predictor variables. In this case, the multiple regression model aims to predict the total number of wins in a season using average points scored and average relative skill as predictor variables. The equation for the model is given by:

Total Number of Wins = β0 + β1 (Average Points Scored) + β2 (Average Relative Skill) + ε

where β0 is the y-intercept, β1 and β2 are the coefficients of the predictor variables, and ε is the error term. The coefficients represent the change in the response variable associated with a one-unit increase in the predictor variable while holding other predictors constant.

The overall F-test is used to test the null hypothesis that all regression coefficients are equal to zero. If this null hypothesis is rejected, it implies that at least one of the predictors is statistically significant in predicting the response variable. The hypothesis test consists of the following steps:

Null Hypothesis: H0: β1 = β2 = 0. The average points scored, and average relative skill have no significant effect on the total number of wins in the season.

Alternative Hypothesis: Ha: At least one of the regression coefficients is not equal to zero. The average points scored and average relative skill have a significant effect on the total number of wins in the season.

Level of Significance: The level of significance is set at 0.05. test statistic measures the overall significance of the model, and the p-value represents the probability of observing such an extreme result by chance, assuming the null hypothesis is true. The test statistic value is calculated by dividing the explained variance by the unexplained variance. A high F-value and a low p-value indicate that the model is a good fit and that the predictors are significantly related to the response variable.

The null hypothesis is rejected if the p-value is less than the level of significance (p-value < 0.05). In this case, the p-value is very low (less than 0.0001), indicating strong evidence against the null hypothesis. Thus, we can conclude that the model is significant in predicting the total number of wins in the season.

Based on the results of the overall F-test, at least one of the predictors is statistically significant in predicting the total number of wins in the season. This means that the model is useful in predicting the total number of wins in the season, and both average points scored, and average relative skill have a significant effect on the response variable.

A multiple linear regression model is used to predict a response variable by identifying the linear relationship between it and two or more predictor variables. The goal is to determine the best combination of predictors that results in the most accurate prediction of the response variable. The equation for this model is represented as y = b0 + b1x1 + b2x2 + b3x3 + ... + bnxn, where y is the predicted response variable, x1, x2, ..., xn are the predictor variables, and b0, b1, b2, ..., bn are the corresponding coefficients.

The multiple regression model created in Step 6 of the Python script predicts the total number of wins in a season using four predictor variables: average points scored, average relative skill, average points differential, and average relative skill differential. The equation for this model is represented as Total Wins = 41.12 + 0.0619Average Points Scored + 0.0317Average Relative Skill + 0.0576Average Points Differential + 0.0089Average Relative Skill Differential.

The overall F-test determines whether at least one of the predictor variables is statistically significant in predicting the number of wins in the season. The null hypothesis (H0) states that all of the predictor variables have coefficients equal to zero, indicating that they have no effect on the response variable. The alternative hypothesis (Ha) states that at least one of the predictor variables has a coefficient that is not equal to zero. The level of significance in this hypothesis test is 1%. The test statistic is 44.98, and the p-value is 0.00. Since the p-value is less than 0.01, we reject the null hypothesis and conclude that at least one of the predictor variables is statistically significant in predicting the number of wins in the season.

The individual t-tests for each predictor variable determine whether it is statistically significant in predicting the number of wins in the season. The null hypothesis (H0) states that the coefficient for the predictor variable is equal to zero, indicating that it has no effect on the response variable. The alternative hypothesis (Ha) states that the coefficient for the predictor variable is not equal to zero. The level of significance in this hypothesis test is 1%. The t-test results show that all four predictor variables are statistically significant in predicting the number of wins in the season, as their p-values are less than 0.01.

The coefficient of determination (R-squared) measures the proportion of variance in the response variable that can be explained by the predictor variables. The R-squared value for this model is 0.920, indicating that approximately 92% of the variation in the total number of wins in a season can be explained by the four predictor variables.

The model predicts that a one-unit increase in average points scored, average relative skill, average points differential, and average relative skill differential will result in a 0.0619, 0.0317, 0.0576, and 0.0089 increase in the total number of wins in a season, respectively. Overall, the results of the multiple regression model suggest that the predictor variables are strong indicators of the number of wins in a season and can be used to make accurate predictions.

In conclusion, the multiple regression model developed using four predictor variables - average points scored, average relative skill, average points differential, and average relative skill differential, was found to be a good fit for predicting the total number of wins in a season. The model was statistically significant at a 1% level of significance, indicating that at least one of the predictor variables was significant in predicting the response variable. All four predictor variables were found to be statistically significant individually as well. The R-squared value of 0.920 suggests that the model can explain approximately 92% of the variation in the total number of wins in a season. The results of the model suggest that a one-unit increase in each predictor variable will lead to an increase in the total number of wins in a season. Overall, this model provides valuable insights for teams and coaches to make data-driven decisions to improve their performance and win more games.