**Task Plan**

Question 1:

Calculation of Proportions:

* Calculate the proportion of wins by higher-ranked players in same-handed matches.
* Calculate the proportion of wins by higher-ranked players in opposite-handed matches.

Statistical Analysis:

* Conduct a statistical test to compare the proportions of higher-ranked wins between same-handed matches and opposite-handed matches.

Interpretation of Results:

* Analyze the results of the statistical test to determine whether there is a significant difference in the proportion of higher-ranked wins between same-handed matches and opposite-handed matches.

Visualization

* Create visualizations, such as bar plots or pie charts, to illustrate the proportions of higher-ranked wins in same-handed and opposite-handed matches.

Discussion and Conclusion:

* Discuss the implications of the findings and potential reasons for any observed differences in the proportion of higher-ranked wins between same-handed matches and opposite-handed matches.
* Conclude by summarizing the main findings and their relevance to the initial research question.

Question 2:

Feature Selection:

* Select relevant match statistics that are likely to be influenced by court surface type, such as winners, unforced errors, aces, and double faults.

Clustering Analysis:

* Perform clustering analysis on the dataset to group matches based on winner and loser statistics.
* Explore different numbers of clusters to identify meaningful patterns in the data.

Association with Court Surface:

* Analyse the clusters to determine if there are any associations between match statistics and court surface type.
* Evaluate whether certain types of matches (clusters) are more common on specific court surfaces (e.g., more aggressive play on grass courts, longer rallies on clay courts).
* Use visualization techniques such as scatter plots or heatmaps to visualize the distribution of clusters across different court surfaces

Comparison with Grand Slam Events:

* Investigate how well the identified clusters align with specific Grand Slam events.
* Examine whether certain types of matches are more prevalent at certain tournaments (e.g., faster-paced matches at Wimbledon due to grass courts, slower-paced matches at the French Open due to clay courts).
* Assess the consistency of match statistics across different Grand Slam events.

Question 3:

Feature Selection:

* Select the difference in ranks as the predictor variable (independent variable).
* Define the outcome variable (dependent variable) as whether the higher-ranked player wins the match (1 for win, 0 for loss).

Model Training and Testing:

* Split the dataset into training and testing sets to evaluate the performance of the logistic regression model.

Model Evaluation:

* Evaluate the performance of the logistic regression model in predicting match outcomes based on the difference in ranks.
* Interpret the coefficients of the logistic regression model to understand the relationship between the predictor variable (difference in ranks) and the probability of the higher-ranked player winning the match.

Interpretation of Results:

* Interpret the results of the logistic regression model in the context of the research question.
* Assess the significance of the predictor variable (difference in ranks) and its impact on the likelihood of the higher-ranked player winning the match.