Through this project, I have learned some key steps in event data analysis. First, during the problem formulation phase, it's important to think deeply and identify the urgent problems that need to be solved. This involves understanding what insights or answers we want to derive from the data, so we have a clear direction for our analysis.

Next, during the data collection phase, it is crucial to find reliable and valid data sources that are relevant to the event we are studying. This means seeking accurate and trustworthy data, and avoiding unreliable or misleading information that could lead to inaccurate analysis conclusions.

Furthermore, the type and structure of the data are essential for subsequent modeling and analysis. We need to ensure that the collected data can be properly used as numerical types for effective calculations and statistical analysis. Additionally, providing accurate and distinct column names for the features in the data is important for correctly classifying and identifying the data.

Lastly, during the data preprocessing stage, it is necessary to filter and clean the data to ensure the model functions properly during its construction. This may involve handling missing values, addressing outliers, and removing duplicate data. By following these steps, we can ensure that the model is not affected by invalid or disruptive data during training and testing, thereby improving the accuracy and reliability of the analysis results.

In our project, initially, we encountered a large amount of character-based data that couldn't be directly used for mathematical modeling. However, we did have a significant portion of data consisting of Yes or No responses. So, I decided to replace these data points with 1 or 0 to enable the use of suitable mathematical models for data analysis.

Nevertheless, I realized that during the data collection phase, it would be more beneficial to gather more appropriate data that could facilitate more accurate data modeling. This is because in our existing datasets, there was very little data that could be used directly, necessitating transformations in order to conduct the analysis effectively.