1. **Data Exploration**
   1. **Understanding Data Structure**:
      * **Data Types**: Identify the types of data (numeric, categorical, datetime, text, etc.).
      * **Dimensions**: Check the shape of the dataset (number of rows and columns).
   2. **Descriptive Statistics**:
      * **Summary Statistics**: Calculate mean, median, standard deviation, min, max, and quartiles.
      * **Distribution Analysis**: Plot histograms, box plots, and KDE plots to understand the distribution of variables.
   3. **Data Visualization**:
      * **Univariate Analysis**: Visualize individual variables using histograms, bar plots, and pie charts.
      * **Bivariate Analysis**: Explore relationships between pairs of variables using scatter plots, correlation matrices, and heatmaps.
      * **Multivariate Analysis**: Use methods like pair plots and parallel coordinates plots to analyse relationships among multiple variables.
   4. **Correlations**:
      * **Correlation Matrix**: Compute and visualize the correlation between numeric variables.
      * **Heatmaps**: Use heatmaps to visualize and understand the strength of correlations.
   5. **Missing Values Analysis**:
      * **~~Missing Values Count~~**~~: Identify the number and percentage of missing values in each column.~~
      * **~~Patterns of Missingness~~**~~: Visualize missing data patterns using heatmaps or bar plots.~~
   6. **Outlier Detection**:
      * **Visual Methods**: Use box plots, scatter plots, and z-score for detecting outliers.
      * **Statistical Methods**: Apply statistical tests to identify anomalies.

**Data Preprocessing**

* 1. **Handling Missing Values**:
     + **Remove Missing Values**: Drop rows or columns with a high percentage of missing values if they are not critical.

1. [6:45 PM]
   1. **Imputation**: Fill missing values using mean, median, mode, or more sophisticated methods like K-Nearest Neighbors (KNN) imputation.
   2. **Outlier Treatment**:
      * **Cap and Floor**: Cap extreme values at a certain percentile.
      * **Transformation**: Apply transformations (e.g., log, square root) to reduce the effect of outliers.
      * **Removal**: Remove outliers based on z-scores or IQR (Interquartile Range).
   3. **Feature Engineering**:
      * **Creating New Features**: Derive new features from existing ones. For example, extract year, month, and day from a timestamp.
      * **Encoding Categorical Variables**: Convert categorical variables into numerical values using one-hot encoding, label encoding, or target encoding.
      * **Interaction Features**: Create interaction features by combining two or more features (e.g., product, ratio).
   4. **Scaling and Normalization**:
      * **Standardization**: Scale features to have zero mean and unit variance.
      * **Normalization**: Scale features to a range (e.g., 0 to 1) using Min-Max scaling or other normalization techniques.
      * **Robust Scaling**: Use RobustScaler for data with outliers, which scales data according to the median and the IQR.
   5. **Dimensionality Reduction**:
      * **Principal Component Analysis (PCA)**: Reduce dimensionality while retaining most of the variance.
      * **t-SNE and UMAP**: Techniques for visualization and reducing dimensionality, useful for exploratory data analysis.
   6. **Handling Imbalanced Data**:
      * **Resampling Techniques**: Use methods like oversampling (e.g., SMOTE) or undersampling to balance the classes.
      * **Algorithmic Techniques**: Use algorithms that handle class imbalance, such as XGBoost or weighted loss functions.
   7. **Feature Selection**:
      * **Univariate Selection**: Select features based on statistical tests and their relationship with the target variable.
2. [6:45 PM]
   1. **Recursive Feature Elimination (RFE)**: Iteratively select features by fitting models and removing the least important features.
      * **Feature Importance**: Use tree-based models to rank and select important features based on their contribution to model performance.