**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 5 May 2024 |
| Team ID | 737906 |
| Project Title | Walmart Sales Analysis For Retail Industry With Machine Learning. |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Template**

For Walmart recruiting stores sales forecasting, thorough data exploration and preprocessing are essential. This involves understanding the dataset's structure, identifying missing values, and visualizing key distributions. Categorical variables need encoding, while feature engineering and scaling enhance predictive accuracy. Time series-specific preprocessing, such as stationarity checks and lag feature incorporation, is crucial for accurate forecasting. Outlier detection and feature selection refine the dataset, ensuring model reliability. Finally, data splitting facilitates model validation, enabling informed decision-making based on reliable predictions.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | To understanding the basic statistics, dimensions, and structure of the data is essential for effective analysis and modeling.  1. Basic Statistics:  Mean: Calculate the average sales amount, as well as the mean of other relevant variables such as price, quantity, and store performance metrics.  Median: Determine the median sales amount to understand the central tendency of the data and assess its robustness to outliers.  Standard Deviation: Compute the standard deviation of sales and other variables to measure the variability or dispersion around the mean.  Minimum and Maximum: Identify the minimum and maximum values of sales and other variables to understand the range of the data.  Percentiles: Calculate percentiles (e.g., 25th, 50th, 75th) to understand the distribution of sales amounts and other variables.  2. Dimensions of Data:  Number of Observations: Count the total number of observations or records in the dataset to understand its size.  Number of Variables: Determine the total number of variables or features available in the dataset, including both independent and dependent variables.  Variable Types: Identify the types of variables present in the dataset, such as numerical, categorical, or ordinal.  Time Dimension: If applicable, consider the temporal dimension of the data by examining timestamps or date-related variables.  3. Structure of Data:  Data Format: Determine the format of the data (e.g., tabular, relational database, JSON) and ensure compatibility with analysis tools and libraries.  Missing Values: Check for missing values in the dataset and decide on strategies for handling them (e.g., imputation, removal).  Variable Names: Review variable names to ensure clarity and consistency, and consider renaming variables for better interpretability.  Data Integrity: Assess the overall integrity of the data by checking for duplicates, inconsistencies, or errors that may require cleaning or preprocessing. |
| Univariate Analysis | Univariate involves examining individual variables one at a time to understand their distribution, central tendency, variability, and other statistical properties.  performing univariate analysis involves calculating and interpreting measures of central tendency such as mean, median, and mode for relevant variables. Let's focus on analyzing the sales data:  Mean:  Mean Sales: Calculate the average sales amount across all transactions. This provides an overall view of the typical sales value.  Mean Sales by Time Period: Compute the average sales for different time periods (e.g., daily, weekly, monthly) to identify any trends or seasonality.  Median:  Median Sales: Determine the median sales amount. Unlike the mean, the median is less affected by extreme values and provides a measure of central tendency that is robust to outliers.  Median Sales by Store Type: Calculate the median sales for different types of stores (e.g., supercenter, neighborhood market) to understand variations in sales performance.  Mode:  Mode of Sales: Identify the sales amount that occurs most frequently. This can indicate common transaction sizes or popular price points.  Mode of Sales by Product Category: Determine the most common sales amounts for different product categories, helping to understand consumer preferences. |
| Bivariate Analysis | bivariate analysis involves examining the relationships between pairs of variables, particularly focusing on the correlation between sales and other relevant factors including correlation analysis and scatterplots:  Correlation Analysis:  Pearson Correlation Coefficient: Calculate the correlation coefficient between sales and other numerical variables to measure the strength and direction of their linear relationship.  Spearman Rank Correlation: Assess the correlation between sales and ordinal variables or non-linear relationships.  Scatterplots:  Sales vs. Time: Plot sales against time-related variables (e.g., date, hour) to visualize trends and seasonality.  Sales vs. Store Attributes: Scatterplot sales against store attributes such as size or performance metrics to identify any patterns or correlations.  Sales vs. Product Attributes: Explore the relationship between sales and product characteristics like price or category using scatterplots.  Sales vs. External Factors: Examine how sales correlate with external factors like weather conditions or holidays. |
| Multivariate Analysis | Exploring relationships and patterns involving multiple variables is crucial for understanding the dynamics of sales and identifying key factors driving performance. Here's how you can investigate these relationships and patterns:  1. Exploratory Data Analysis (EDA):  Correlation Matrix: Compute the correlation between all pairs of variables to identify significant relationships. Visualize the correlation matrix to understand the strength and direction of these relationships.  Pairwise Scatterplots: Plot pairwise scatterplots between sales and other relevant variables to visually examine their relationships. Use different colors or markers to represent different categories or clusters if applicable.  2. Feature Engineering:  Create Interaction Terms: Generate new features by combining existing variables to capture potential synergistic effects. For example, create a "Promotion × Store Type" feature to analyze how different types of stores respond to promotions.  Temporal Features: Extract temporal features such as day of the week, month, or season from the date variable to explore sales patterns over time.  3. Regression Analysis:  Multiple Regression: Build a multiple regression model with sales as the dependent variable and multiple predictors such as store attributes, product features, promotional activities, and external factors. Analyze the coefficients to understand the relative importance of each predictor.  Regularization Techniques: Apply regularization techniques like Lasso or Ridge regression to handle multicollinearity and select the most important predictors.  4. Clustering Analysis:  Customer Segmentation: Use clustering algorithms such as K-means to segment customers based on demographic variables, purchasing behavior, and geographical location. Analyze the characteristics of each segment and their corresponding sales patterns.  Store Clustering: Cluster stores based on attributes like size, location, and performance metrics. Compare the sales performance of different clusters to identify high-performing and low-performing store groups.  5. Time Series Analysis:  Seasonal Decomposition: Decompose the sales time series into trend, seasonal, and residual components to identify seasonal patterns and long-term trends.  Lagged Variables: Incorporate lagged variables (e.g., sales from previous periods) into the analysis to capture temporal dependencies and autocorrelation. |
| Outliers and Anomalies | Identification and Treatment of Outliers:  1. Identify Outliers:  Statistical Methods: Use statistical techniques like Z-score or interquartile range (IQR) to identify data points that deviate significantly from the rest of the distribution.  Visualization: Plot histograms, box plots, or scatterplots to visually inspect the data for any extreme values or unusual patterns.  2. Understand the Nature of Outliers:  Domain Knowledge: Utilize domain expertise to understand whether outliers represent genuine anomalies or errors in the data.  Contextual Analysis: Investigate the circumstances surrounding outlier data points to determine their potential causes.  3. Choose Outlier Treatment Methods:  Imputation:  Replace outliers with a measure of central tendency (e.g., mean, median) to maintain the overall distribution of the data.  Use interpolation techniques to estimate outlier values based on neighboring data points.  Transformation:  Apply transformations like log transformation to stabilize variance and reduce the impact of outliers.  Removal:  Remove outliers from the dataset entirely if they are deemed to be erroneous or unrepresentative of the underlying distribution.  Consider trimming the dataset by removing a fixed percentage of data points from both tails of the distribution.  4. Implement Outlier Treatment:  Apply the chosen outlier treatment method to the dataset, ensuring that it is performed consistently across all relevant variables.  Document the rationale behind the outlier treatment decisions for transparency and reproducibility.  5. Evaluate the Impact:  Assess the impact of outlier treatment on the distribution and statistical properties of the data.  Determine whether outlier treatment has improved the quality and reliability of the dataset for subsequent analysis and modeling tasks.  6. Sensitivity Analysis:  Conduct sensitivity analysis to explore the effects of different outlier treatment strategies on downstream analysis and model performance.  Evaluate the robustness of conclusions and recommendations under various outlier handling scenarios.  7. Continuous Monitoring:  Establish mechanisms for ongoing monitoring of data quality to detect and address outliers as new data becomes available.  Incorporate outlier detection and treatment as part of regular data maintenance processes. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Checking For Null Values |  |
| Handling Negative Values |  |
| Exploratory Data Analysis |  |
| Handling Categorical Values |  |

|  |  |
| --- | --- |
| Splitting Data Into Train And Test |  |