Data Wrangling and Descriptive Statistics Report

1 Introduction

This report presents the exploration, cleaning, feature engineering, and analysis of a dataset obtained from Kaggle by handling missing values, identifying outliers, engineering features, and visualizing key trends.

2 Data Exploration and Cleaning

2.1 Dataset Overview

The dataset includes both numerical and categorical variables. The structure of the dataset was examined using .info(), .describe(), and .head() functions. The dataset contains missing values and potential outliers that required preprocessing.

2.2 Handling Missing Values

- Numerical columns: Missing values were interpolated.
- Categorical columns: Missing values were replaced with the most frequent category (mode).
- Columns with more than 50% missing values were dropped.

2.3 Outlier Detection and Removal

Outliers were identified using the Interquartile Range (IQR) method. Extreme values lying outside 1.5 times the IQR were removed to improve data consistency.

3 Feature Engineering and Descriptive Statistics

3.1 Feature Engineering

- Binning Continuous Variables The `Distance_km` column was transformed into a categorical variable by binning its values into three distinct categories: low, medium, and high distances. This approach simplifies the interpretation of continuous data and helps capture non-linear relationships.
- Encoding Categorical Variables The categorical column `Weather` was processed using one-hot encoding to convert it into a binary matrix format. This transformation ensures that the model can interpret categorical data effectively without assuming any ordinal relationship between the categories.

3.2 Descriptive Statistics

- Mean, median, and standard deviation were calculated for numerical variables.
- Frequency counts were determined for categorical variables.

4 Data Visualization and Analysis

4.1 Visualization of Distributions

- Histograms and box plots were used to visualize numerical distributions and detect outliers.
- Bar plots were used to display the frequency distribution of categorical variables.

4.2 Pairwise Relationships

- A pairplot was generated to explore relationships between numerical variables.
- Scatter plots were used to identify correlations and trends.

5 Key Insights

- Higher distances generally lead to longer delivery times.
- Traffic levels significantly impact delivery time.
- Weather conditions influence vehicle type selection and overall delivery efficiency.

6 Conclusion

The data wrangling process improved the dataset's quality by handling missing values, removing outliers, and engineering relevant features. The descriptive statistics and visualizations revealed important trends and relationships in the data, which can be further used for predictive modeling.