UberDataset

Team G3

ICS474 Project

Abdulmajeed Alghamdi 202032860 Section02

Mohamad Alsakka 202044860 Section02

Abdullah Algada 202044680 Section01

1. **Dataset Overview**

The dataset is about uber trips data, and it focuses on trips purposes, distance of trip, start and end locations of trips, and the category of the trip.

1. **Feature Description**

The features of the dataset are as follows:

START\_DATE: The start date and time of the trip (Numeric, Interval).

END\_DATE: The end date and time of the trip (Numeric, Interval).

CATEGORY: Indicates if the trip was for "Business" or "Personal" purposes (Categorical, Nominal).

START: Starting location of the trip (Categorical, Nominal).

STOP: Destination or stopping location of the trip (Categorical, Nominal).

MILES: The distance traveled in miles (Numeric, Ratio), which could be used to categories trips (short or long trip).

PURPOSE: The reason for the trip (Categorical, Nominal).

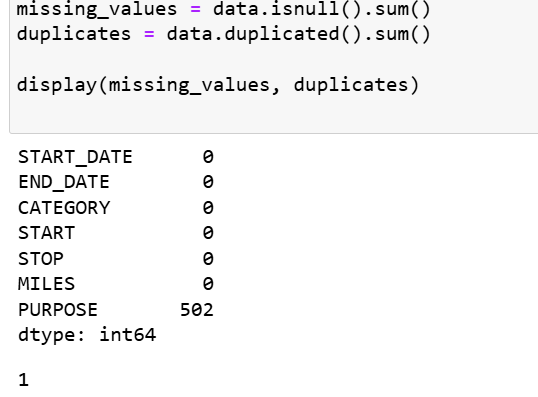
PURPOSE can be the Target variable.

1. **Dataset Structure**

The dataset contains 1155 rows and 7 columns.

1. **Missing Values and Duplicates**

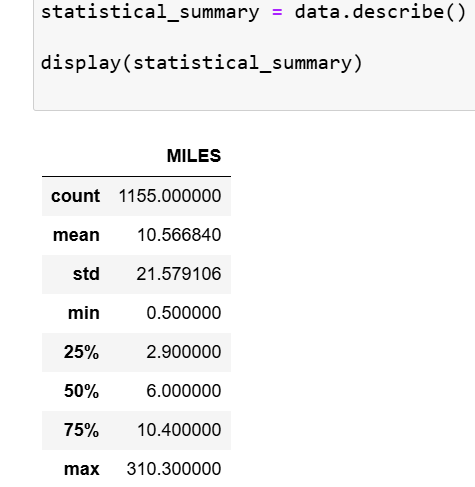
Using the following command will display the missing values and duplicates:



As we can see, the purpose column has many empty values which will limit the model accuracy, also it could lead to biased insights, especially if missing entries are related to certain trip types. So, we need to remove these rows because PURPOSE is the target and rows existence will reduce our model accuracy.

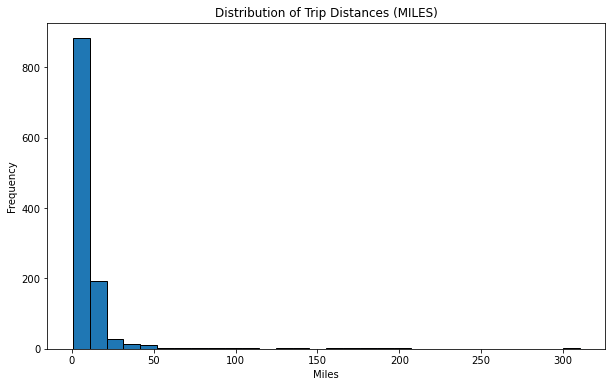
1. **Statistical Summary**

To get the statistical summary of numeric columns we can use the following commands:



1. **Data Distribution**

In the figure below it shows the distribution of trips miles



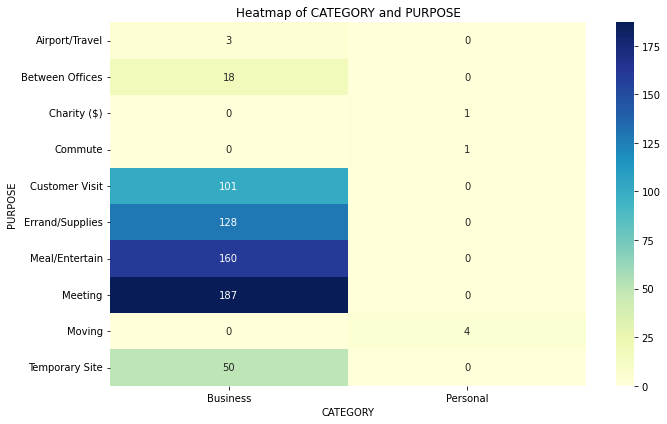
The figure below shows the distribution of purpose types

A graph of different colored bars

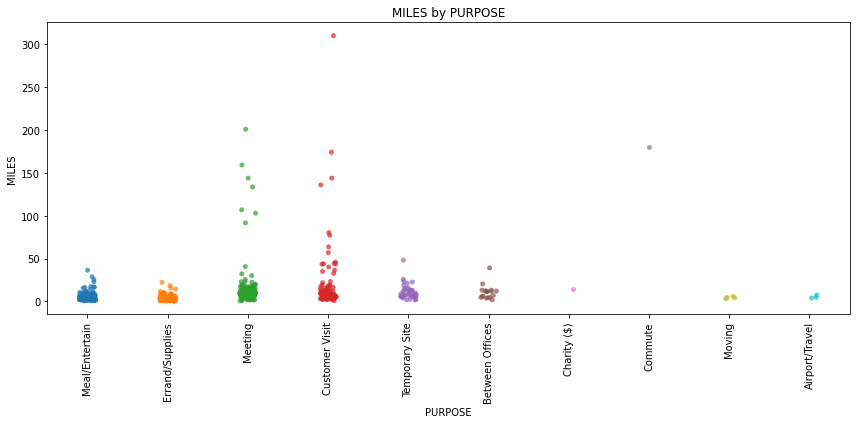
Description automatically generated

1. **Correlation Analysis**

The following figure is a heatmap to show the relationship between CATEGORY and PURPOSE



The following figure is a scatter plot to show the relationship between MILES and PURPOSE



1. **Outlier Detection**

The box plot of MILES reveals outliers in the right side, indicating that a few trips are much longer than average. These outliers could skew the analysis by inflating the mean distance, giving a misleading impression of average trip lengths.

