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).

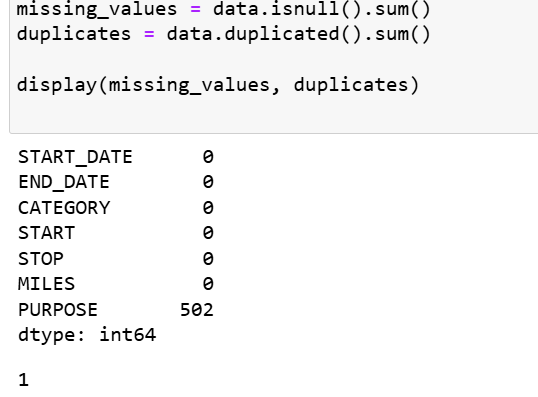
PURPOSE can be the Target variable.

1. **Dataset Structure**

The dataset contains 1155 rows and 7 columns.

1. **Missing Values and Duplicates**

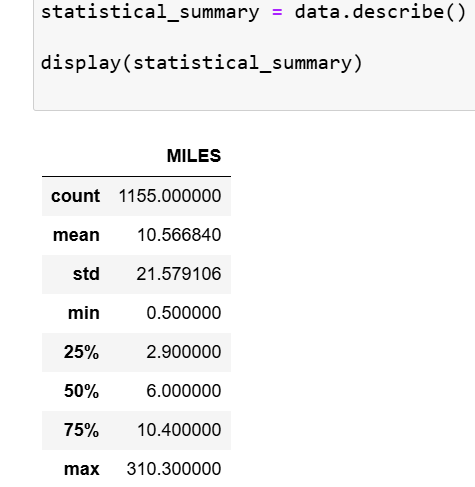
Using the following command will display the missing values and duplicates:



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.

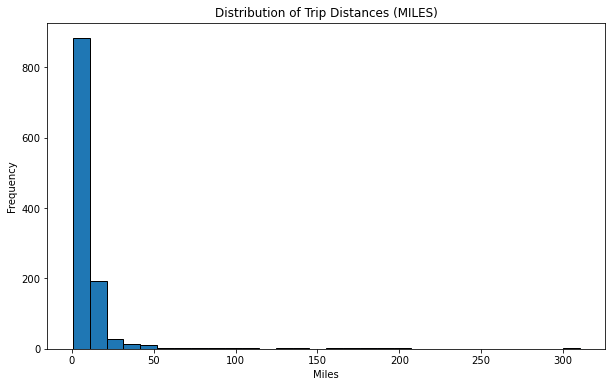
1. **Statistical Summary**

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



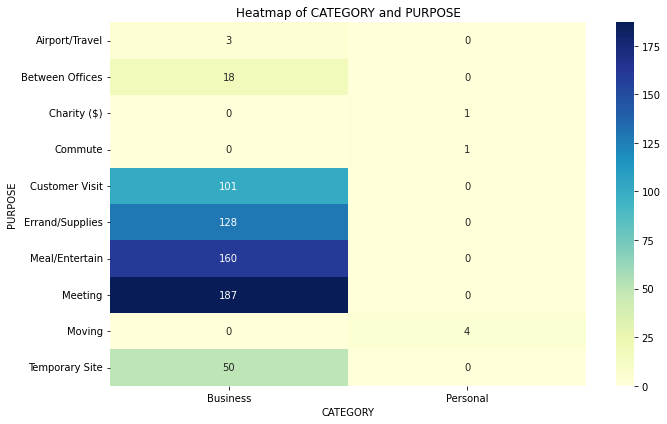
The figure below shows the distribution of purpose types

A graph of different colored bars

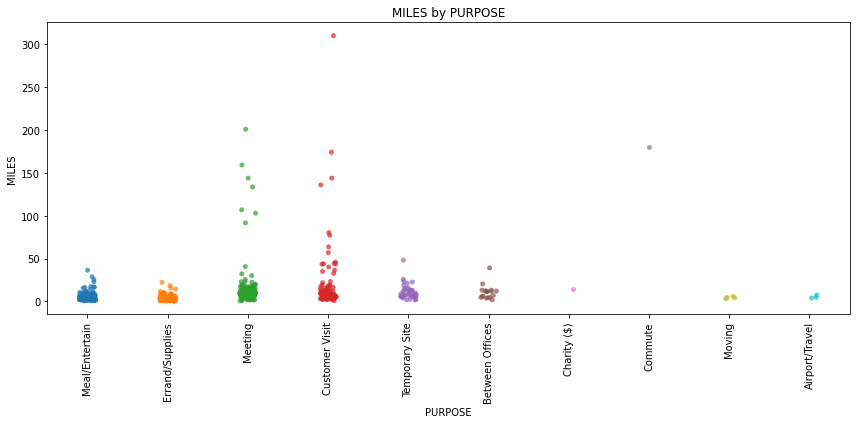
Description automatically generated

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



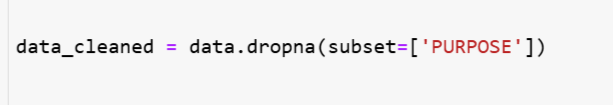
1. **Outlier Detection**

The box plot of MILES reveals outliers on 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.



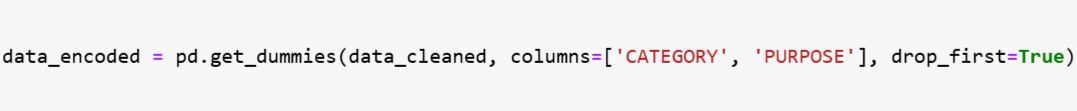
**9. Handling Missing Data**

Since the PURPOSE column has a significant portion of missing values (502 out of 1155 rows), there are two possible approaches to handle it. The first option is deletion, where we could drop the rows with missing values. This would help avoid introducing bias or noise through imputation. The second option is imputation, where we could fill in the missing values based on patterns in other fields, such as using the most common purpose for a given CATEGORY (e.g., imputing business trips with the most frequent purpose for that category). After evaluating both options, I’ve decided to proceed by dropping rows with missing PURPOSE values, as this will help retain original, unaltered data where purpose is known, ultimately enhancing the accuracy of models that might rely on the trip purpose.



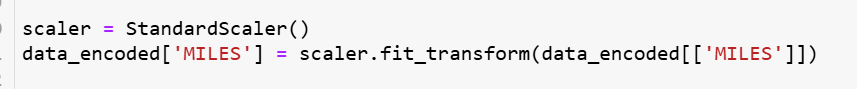
**10. Encoding Categorical Variables**

The dataset includes categorical variables such as CATEGORY (e.g., “Business” or “Personal”) and PURPOSE. To make these variables usable in machine learning models, the encoding technique chosen is One-Hot Encoding. This method will create binary columns for each unique value in the CATEGORY and PURPOSE variables, ensuring that there is no ordinal bias introduced and making the features compatible with algorithms that require numerical input.



**11. Feature Scaling**

MILES is the only continuous numerical feature in the dataset, and depending on the machine learning model used, scaling may be necessary. For models sensitive to feature scales, such as linear regression or SVMs, Standard Scaling (z-score normalization) will be applied. Alternatively, Min-Max Scaling can be used if there is a need to normalize MILES between 0 and 1, particularly for algorithms that are sensitive to feature range or when preparing data for visualization. Standard scaling is the chosen approach, as it ensures consistent scaling across features, especially for models that rely on distance calculations.

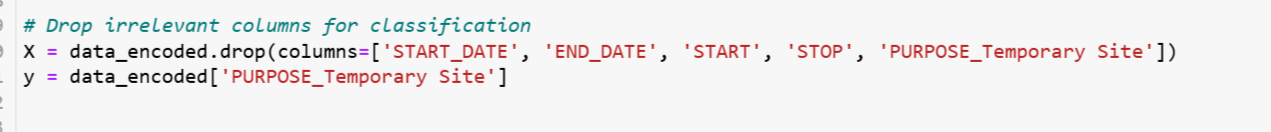


**12. Feature Selection**

The selected features for analysis are MILES, CATEGORY, and PURPOSE, as they are the most relevant for understanding travel behavior. Initially, START and STOP locations will be excluded unless the analysis requires route-specific models or geospatial analysis. MILES is central to the analysis, CATEGORY provides context for the purpose of trips, and PURPOSE itself reveals insights into the types of trips. Including START\_DATE or END\_DATE could be considered if temporal patterns are of interest, such as peak travel times, particularly if time series analysis is planned.

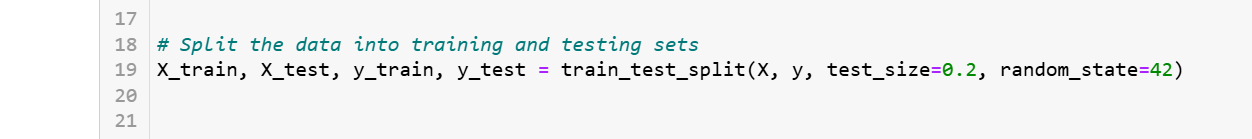
**13. Algorithm Selection**

Given the structure and features of the dataset, there are two possible tasks: classification (if predicting categories like trip purpose or business/personal trips) and regression (if predicting a continuous value like trip distance or MILES). For classification, suitable algorithms include Logistic Regression, Random Forest, or Support Vector Machines (SVM), all of which are effective for predicting categorical variables like PURPOSE. For regression, Linear Regression or Random Forest Regression would be appropriate for predicting trip distance. Since there is a predefined target, the chosen approach is to predict PURPOSE, a categorical variable, using a classification model.



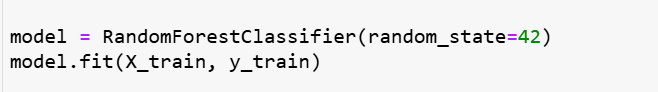
**14. Data Splitting**

The method chosen for data splitting is a hold-out approach, where 80% of the data is used for training and 20% for testing. This simple and effective method helps evaluate the model's performance while preventing data leakage. It works well for initial model validation, with the potential for cross-validation later to confirm model stability and reduce variance in performance.



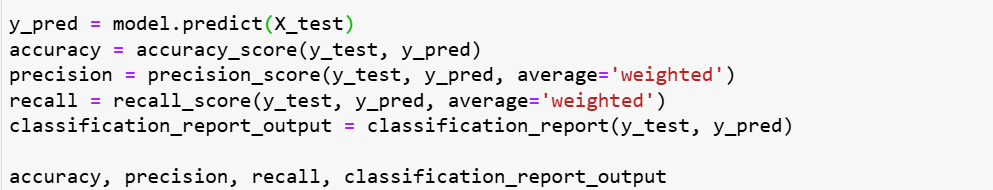
**15. Model Training**

For training, Random Forest will be used as the classification algorithm. It is well-suited for handling categorical data and is robust to overfitting, especially when tuned appropriately. Initially, the default parameters will be used for training, with hyperparameter tuning to be considered in the model improvement phase if needed.



**16. Model Evaluation**

To evaluate the model, accuracy will be the primary metric, providing a straightforward measure of classification performance. Additionally, Precision and Recall will be examined to gain insights into how well the model differentiates between specific classes, such as business versus personal trips, offering a more nuanced understanding of its performance.



**17. Performance Analysis**

* The model achieved an accuracy of 97.7% on the test set.
* Precision and Recall:
  + Weighted Precision: 97.8%, Showing how accurate the model is across categories.
  + Weighted Recall: 97.7%, Showing how accurate the model can identify the instances.
* Classification Report:
  + The main class (0) has both precision and recall close to 1.0, but class 1 has a lower recall of 0.73 because the model has difficulty with it.

**18. Model Improvement**

* Hyperparameter Tuning: Applying cross-validation and tuning parameters like the number of estimators, maximum depth, and minimum samples per leaf could improve performance for less frequent classes.
* Feature Engineering: Creating new features from START\_DATE (e.g., time of day, weekday/weekend) could enhance classification performance if certain purposes correlate with time.

**19. Validation**

Cross-validation can be used to ensure the model generalizes well across categories. K-Fold Cross-Validation (e.g., with 5 folds) can provide a more robust measure of model performance by averaging metrics across different data splits, helping us to detect any overfitting and ensuring stability across diverse samples.

We will perform cross-validation on the Random Forest model to validate its consistency.

**20. Final Model Selection**

To select the final model, we will compare models with each other and select our final model based on certain criteria

* Comparison: the cross-validated Random Forest model’s performance will be compared with another classifier, such as Logistic Regression. Random Forest is typically more accurate but also more complex, while Logistic Regression is simpler and may perform well if relationships are mostly linear.
* Selection Criteria: The final model will be chosen based on balanced performance metrics (accuracy, precision, recall) and practical considerations, like model complexity and interpretability.

Validation and Final Model Selection

* Cross-Validation Results:
  + Random Forest: Achieved a cross-validated accuracy of 98.9%, indicating strong generalization.
  + Logistic Regression: Achieved a cross-validated accuracy of 98.5%, which is close to Random Forest but slightly lower.

Final Model Selection

* Chosen Model: Random Forest model due to its slightly higher accuracy and robustness in handling non-linear relationships in the data.
* Justification: Random Forest might be more complex, but its superior performance and ability to handle the categorical and numerical mix make it more suitable for this dataset, so complexity can be justified here because it will give better results for this dataset.

**21. Data Distribution**

* Numerical Feature (MILES): Histogram and boxplot will be created to examine its distribution and detect any outliers.

A comparison of a graph

Description automatically generated

* Categorical Features (CATEGORY and PURPOSE): bar plots will be used to show the frequency of each category.

A graph of different colored bars

Description automatically generated with medium confidence

**22. Feature Importance**

* Since the final model is the Random Forest, feature importance scores can be visualized to show which features have the most impact on predictions.

A graph with different colored bars

Description automatically generated

**23. Model Performance Across Features**

* We will visualize how well the model performs across various categories in PURPOSE and CATEGORY to check for any imbalances or performance variations.
* Model performance will be measured using model accuracy by PURPOSE and by CATEGORY.
* The graphs below show a comparison between model accuracy, one for CATEGORY and one for PURPOSE.
* It can be seen from the graph that the model perform well across different categories, showing high and close accuracy rates, which indicates that the model does not exhibit performance imbalances across these categories.

A comparison of bar graph

Description automatically generated with medium confidence