Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 06-14-2024, Time: 21.30PM EST
Internship Batch: LISUM34,Version:1.0
Data intake by: Shradhanjali Pradhan
Data intake reviewer: -
Data storage location: https://github.com/shradhanjalipradhan/Week_2

Tabular data details: Cab_Data

Total number of observations	**359392**
Total number of files	1
Total number of features	7
Base format of the file	CSV
Size of the data	20.1 MB

Tabular data details: Customer_ID

Total number of observations	**49171**
Total number of files	1
Total number of features	4
Base format of the file	CSV
Size of the data	1.3 MB

Tabular data details: Transaction_ID

Total number of observations	**440098**
Total number of files	1
Total number of features	3
Base format of the file	CSV
Size of the data	8.7 MB

Tabular data details: City

Total number of observations	**20**
Total number of files	1
Total number of features	3
Base format of the file	CSV
Size of the data	1 KB

Data Preprocessing and Merging Report

Introduction

This report outlines the steps and methods used to preprocess and merge multiple datasets. The datasets include cab data, customer data, transaction data, and city data. The objective was to clean and merge these datasets for further analysis.

Steps and Methods

Step 1: Reading the CSV files

The initial step involves reading the CSV files into pandas DataFrames using the `pd.read_csv()` method.

Step 2: Ensuring column names are consistent

Column names in each DataFrame were stripped of any leading or trailing spaces to ensure consistency. This was done using the `str.strip()` method.

Step 3: Renaming Columns

The `Customer ID` column in the customer data was renamed to `Customer_ID` to ensure consistency during merging. This was achieved using the `rename()` method.

Step 4: Converting relevant columns to string

Relevant columns such as `Transaction_ID` and `Customer_ID` in various DataFrames were converted to string type using the `astype(str)` method. This step was crucial for accurate merging of the datasets.

Step 5: Inspecting Columns

Column names in each DataFrame were printed to ensure correctness before proceeding with the merge operations.

Step 6: Merging the DataFrames

The datasets were merged step-by-step using the `pd.merge()` method:

1. `cab_df` and `transaction_df` were merged on `Transaction_ID`.
2. The resulting DataFrame was merged with `customer_df` on `Customer_ID`.
3. Finally, the merged DataFrame was combined with `city_df` on `City`.

Step 7: Displaying the final DataFrame

The head of the final merged DataFrame was printed to verify the merge operations. Additionally, the shape of the final merged DataFrame was printed to understand the dimensions of the resulting dataset.

Step 8: Displaying the final DataFrame

Exploratory data analysis using box plots, pair plots, heatmap, pie charts, bar graphs to analyze the relation of data.

Step 9: Performed the hypothesis test

There are 4 hypothesis test method I used.

Conclusion

The preprocessing and merging steps ensured that the datasets were cleaned and combined accurately. These steps are fundamental for any data analysis or machine learning tasks that will be performed on the dataset.