Data Intake Report

Name: G2M Insight for Cab Investment Firm XYZ
Report date: 26-08-2023
Internship Batch: LISUM 24
Version: 1.0
Data intake by: Sai Vishal Arram
Data intake reviewer:
Data storage location:

Tabular data details:

Cab Dataset:

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	21.8 MB

Transaction ID Dataset:

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	8.58 MB

Customer ID Dataset:

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 MB

City Dataset:

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

US Airport Dataset:

Total number of observations	20
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	1 KB

US Holiday Dataset:

Total number of observations	342
Total number of files	1
Total number of features	6
Base format of the file	.csv
Size of the data	16 KB

Proposed Approach:

Data Preparation and Exploration:

Joined all the tables (excluding holiday and airport data) and prepared a final data frame. Checked for any unique values, but none were found. No missing values were found. Removed any duplicate values. Explored all categorical and numeric variables.

Exploration Relationship Between Features and Hypothesis Testing:

Correlation matrix to identify which features are correlated with profit. Assumed various hypothesis and tested their significance.

Analyzing Derived Features:

Analyzed the relationship between number of airports in the city and their relation with rides and profit. Same with the impact of US holidays

Recommendations:

Based on the findings, the best company to invest in.

Assumptions:

- Profit has been calculated by subtracting the cost of the trip from the cost charged to the customer
- The outliers have been identified in the price charged and profit, however, they are being considered as outliers due to the fact that charge could depend on various other factors which are not included (peak times, seasonality etc.,)

- The data represents the population
- Data is not imbalanced