Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 10/14/2023
Internship Batch: LISUM26
Version: <1.0>
Data intake by: Matthew Iversen
Data intake reviewer: <intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets

Cab_Data

Total number of observations	359,392
Total number of features	7
Base format of the file	.csv
Size of the data	62.83 MB

City_Data

Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	738 Bytes

Customer_Data

Total number of observations	49,171
Total number of features	4
Base format of the file	.csv
Size of the data	4.22 MB

Transaction_Data

Total number of observations	440,098
Total number of features	3
Base format of the file	.csv
Size of the data	33.89 MB

Proposed Approach:

- Import the datasets to understand the structure, dimensions, and data types.
- Check the first few records of each dataset to get a sense of the data.
- The deduplication process was approached by checking for exact matches across all columns in the datasets.
- Check for any missing values in the datasets.
- Assess the nature and impact of the missing data to determine if any imputation or removal was necessary.
- Visualize data distributions using boxplots to inspect for potential outliers visually.
- Apply the IQR (Interquartile Range) method to identify outliers statistically.

- Identify common columns (keys) across datasets to perform merges.
- Create a comprehensive merged dataset for a holistic analysis, ensuring data consistency and integrity.
- Formulate hypotheses based on data attributes and business questions.
- Conduct both univariate and bivariate analyses to derive insights.
- Use visualizations to support findings and insights.

Assumptions:

- Consistent Data: It was assumed that the data provided, especially where IDs are concerned (like Customer ID, Transaction ID), was consistent across datasets.
- Data Completeness: While missing values were checked for, it was assumed that the datasets provided represented a complete picture for the given time frame. Any data not present was considered as not collected or not relevant.
- Outliers: We assumed that extreme values in terms of payment could be legitimate, especially in the context of cab rides, which can vary greatly in distance and price.
- Currency: All monetary values were assumed to be in the same currency, likely USD, given the location of the cities.
- Time Frame Consistency: The data was assumed to be consistently recorded across the mentioned time frame without significant interruptions or changes in data collection methods.