

Data Intake Report

Name: G2M Insight for Cab Investment Firm

Report date: 10/14/2022

Internship Batch: LISUM14

Version:<1.0>

Data intake by: Ammar Sidhu

Data intake reviewer: <intern who reviewed the report>

Data storage location: https://github.com/ammarsidhu18/G2M_Insight_Cab_Investment

Tabular data details:

Name – Cab_data.csv

Total number of observations	359392
Total number of files	1
Total number of features	8
Base format of the file	.csv
Size of the data	22.08 MB

Tabular data details:

Name – Customer_ID.csv

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	1.027 MB

Tabular data details:

Name – Transaction_ID.csv

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	8.778 MB

Tabular data details:

Name – Transaction_ID.csv

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	0.001 MB
------------------	----------

Proposed Approach:

- Approach:
 - 1) Merge Datasets on Common Columns
 - 2) Check for Missing Values, and Duplicates
 - 3) Create Target Variable – the Margins
 - 4) Split Data into Pink and Yellow Cab Company
 - 5) Explore Outliers for Margin Variable
 - 6) Create Univariate and Bivariate Data Visualizations for Features and Target
 - 7) Propose and Answer Questions About Cab Companies
 - 8) Correlation Analysis and Hypothesis Testing
 - 9) Conclude Best Cab Company for Investment
 - 10) Develop Machine Learning Models to Predict Margins given Features
- Assumptions:
 - 1) No Duplicate Values
 - 2) No Missing Values
 - 3) The 4 Datasets Can be Mapped and Merged to Each Other