

Data Intake Report

Name: G2M Insight for Cab Investment Firm

Report date: 12/12/2022

Internship Batch: LISUM16

Version: 1.0

Data intake by: Sam Embry

Data intake reviewer: N/A

Data storage location: https://github.com/skembr01/data_glacier_work/tree/main/Week_2

Tabular data details:

Cab_data

Total number of observations	359384
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

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	9 MB

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	759 bytes

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	1.1 MB

Proposed Approach:

- Join data into one df, so that all aspects of each of the four files can be represented
- Utilize D'Agonisto's K^2 and Kolmogrov-Smirnov tests for normality

- Chart variables for EDA, but also to visualize the hypotheses
- Utilize independent t-tests to calculate differences in means for hypotheses
 - Utilize Levene's test for variances related to use in these t-tests