Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 05/03/2021
Internship Batch: LISP01
Version:1.0
Data intake by Beril Chepkorir
Data intake reviewer:
Data storage location: https://github.com/dangerous2

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	17
Base format of the file	.csv
Size of the data	56.6+ MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

Data Validation (Identification)

I merged various datasets using common columns into a master data and analyzed it using various python techniques and visualization

Assumptions

Users feature of city dataset is treated as number of cab users in the city.
we have assumed that this can be other cab users as well (including Yellow and Pink cab)
We have outliers in our dataset
We assume that the profit is calculated by Price Charged -Cost of trip

Datasets Used

Time period of data is from 31/01/2016 to **31/12/2018.**

The list of datasets which are provided for the analysis:

Cab_Data.csv – this file includes details of transaction for 2 cab companies

Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details

Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode

City.csv – this file contains list of US cities, their population and number of cab users

Cab Dataset

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	01/08/16	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	01/06/16	Pink Cab	ATLANTA GA	28.62	358.52	334.854
2	10000013	01/02/16	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	01/07/16	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	01/03/16	Pink Cab	ATLANTA GA	8.73	114.62	97.776

City dataset

	City	Population	Users
0	NEW YORK NY	8405837	302149
1	CHICAGO IL	1955130	164468
2	LOS ANGELES CA	1595037	144132
3	MIAMI FL	1339155	17675
4	SILICON VALLEY	1177609	27247

Customer Dataset

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536

Transaction Dataset

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card

The final dataset after combining all the files -master_data

Date of ravel	Company	City	KM Travelled	Price Charged	Cost of Trip	Payment_Mode	Gender	Age	Income (USD/Month)	Population	Users	Profits	Month	Year	Day	agerange
016-1-08	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	Card	Male	28	10813	814885	24701	57.3150	1	2016	8	18-25
018-7-21	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	Cash	Male	28	10813	814885	24701	281.2772	7	2018	21	18-25
018-1-23	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	Card	Male	28	10813	814885	24701	194.6480	11	2018	23	18-25
016-1-06	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	Card	Male	27	9237	814885	24701	23.6660	1	2016	6	18-25
018-4-21	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	Card	Male	27	9237	814885	24701	253.9808	4	2018	21	18-25