Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 20th September 2022
Internship Batch: LISUM13: 30
Version: 1.0
Data intake by: Rania Tarek Fleifel
Data intake reviewer:
Data storage location: G2M Cab datasets

Tabular data details:

1) **Cab_data.csv**

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	20.1 MB
Unique identifier feature	Transaction ID
Dupe validation	No duplicates

2) **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 MB
Unique identifier feature	Customer ID
Dupe validation	No duplicates

3) **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	8.58 MB
Unique identifier feature	Transaction ID
Dupe validation	No duplicates

4) **City.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	759 bytes

Unique identifier feature	City
Dupe validation	No duplicates

Variables after data cleaning and new features are: (Total:19)

- 'trans_id'--> ride's unique identifier
- 'company'--> yellow_cab or pink_cab
- 'city'--> ride's location
- 'kms'--> ride's distance covered in kms
- 'c_paytype'--> ride's payement method cash or credit
- 'costumer_id' --> rider's unique identifier
- 'gender'--> rider's gender
- 'c_age'--> age of rider
- 'c_income'--> income USD/month of rider
- 'population'--> city's population
- 'users'--> city's cab users
- 'traveldate'--> ride's date %d-%M-%Y
- 'cost_km'--> ride's 'cost/kms'
- 'charged_km'--> ride's 'charged/kms'
- 'weekendnot'--> whether date of ride is a weekend (Sat,Sun) or not
- 'holidaynot'--> whether date of ride is a US holiday (**https://www.timeanddate.com/holidays/us/?hol=25) or not
- 'agegrp'--> rider's age group (youth or adult or older)
- 'year'--> extracted from traveldate
- 'month'--> extracted from traveldate

Deduplication Approach:

- For each file, find the number of unique values for each feature, the feature with the maximum number of unique features is the "Unique identifier" of the file.
- **If** the unique identifier feature has number of unique values equal the total number of observations, there's no duplicated observations (entirely repeated rows).
- **Else,** keep only one of the duplicated rows & drop the rest.

Assumptions:

- If a feature doesn't have neither zeros or empty values, there's no missing data
- If a feature has values of different types, there's faulty data
- Merge transaction_ID.csv and costumer_ID.csv on "Customer_ID" feature.
- Data where the company choice is not provided is not helpful in making final decision(e.g. costumer IDs that don't have any transaction IDs within our desired time frame)
- Data provided is for a time interval more than the interval I am looking into for decision making
- The market un-utilized by both companies is up-for-grabs (i.e. there are no other competitors)
- The cost of a ride is based on kms only, the rest of the charged amount is the profit per ride
- XYZ is interested in the company with biggest growth potential, not current market share.