Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: Mar 11, 2024
Internship Batch: LISUM31
Version: 1.0
Data intake by: Shuju Sun
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/jeff-suen/VC_Week2/tree/main/Datasets

Tabular data details:

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	21.2MB

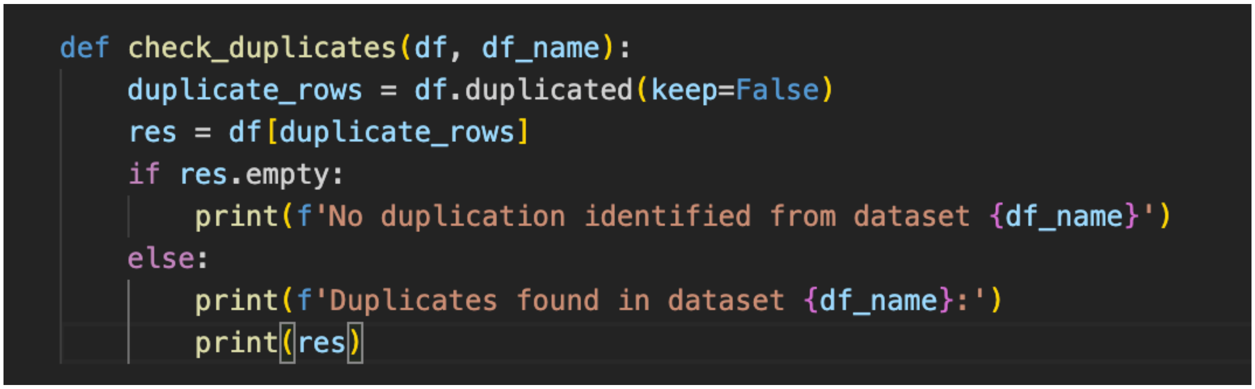
Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

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	1.1MB

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	9MB

Proposed Approach:

- Approach of duplication validation (identification)
  I would approach this question by writing a duplication identification function shown below:



```python
def check_duplicates(df, df_name):
    duplicate_rows = df.duplicated(keep=False)
    res = df[duplicate_rows]
    if res.empty:
        print(f'No duplication identified from dataset {df_name}')
    else:
        print(f'Duplicates found in dataset {df_name}:')
        print(res)
```

- Mention your assumptions (if you assume any other thing for data quality analysis)

1. I assume that all other aspects of the data collected meet the requirement of the project requirements.
2. The dataset is assumed to be consistent within itself and across time and data sources. There are no contradictions or discrepancies in the data.
3. Data is assumed to be correctly recorded and to reflect the true values of the intended attributes. This means that there are no errors or deviations from the true values due to mishandling or misreporting.