

Data Intake Report

Name: <G2M insight for Cab Investment firm Transaction_id>

Report date: <14/07/2022>

Internship Batch:<LISUM11>

Version:<1.0>

Data intake by:<Md Khalid Siddiqui>

Data intake reviewer:<intern who reviewed the report>

Data storage location: < https://raw.githubusercontent.com/DataGlacier/DataSets/main/Transaction_ID.csv

>

Tabular data details:

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	<32.3 MB>

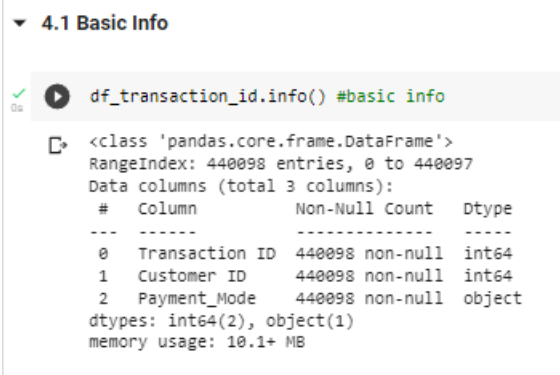
Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Two methods were used to explore initial characteristics (univariate analysis)

1. Pandas code : df_dataframe. Info

Sample screenshot



```
4.1 Basic Info

df_transaction_id.info() #basic info

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440098 entries, 0 to 440097
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Transaction ID  440098 non-null  int64
1   Customer ID    440098 non-null  int64
2   Payment_Mode   440098 non-null  object
dtypes: int64(2), object(1)
memory usage: 10.1+ MB
```

2. Pandas Profiling: widget command of pandas profiling module

Sample screenshot

▼ 4.2 Interactive Profile Report

215

▶

```
profile4 = ProfileReport(df_transaction_id, title = "Profile Report of dataset customer_id")
profile4.to_widgets() #interactive widget form
```

↑ ↓ ↻ ⌂ ⚙ 📄 🗑 ⋮

/usr/local/lib/python3.7/dist-packages/pandas_profiling/profile_report.py:361: UserWarning: Ipywidgets is not yet fully supported on Google Colab (https://github.com/googlecolab/colabtools/issues/60).
"Ipywidgets is not yet fully supported on Google Colab (https://github.com/googlecolab/colabtools/issues/60)."

Summarize dataset: ██████████ 17/? [00:07<00:00, 2.47it/s, Completed]

Generate report structure: 100% ██████████ 1/1 [00:01<00:00, 1.95s/it]

Overview	Variables	Interactions	Correlations	Missing values	Sample
----------	-----------	--------------	--------------	----------------	--------

Overview	Reproduction	Warnings (2)
Number of variables	3	NUM 2
Number of observations	440098	CAT 1
Missing cells	0	
Missing cells (%)	0.0%	
Duplicate rows	0	
Duplicate rows (%)	0.0%	
Total size in memory	32.3 MiB	
Average record size in memory	77.0 B	

Report generated with [pandas-profiling](#).

◀

▶