Data Intake Report

Name: <G2M insight for Cab Investment firm >
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/Cab_Data.csv>

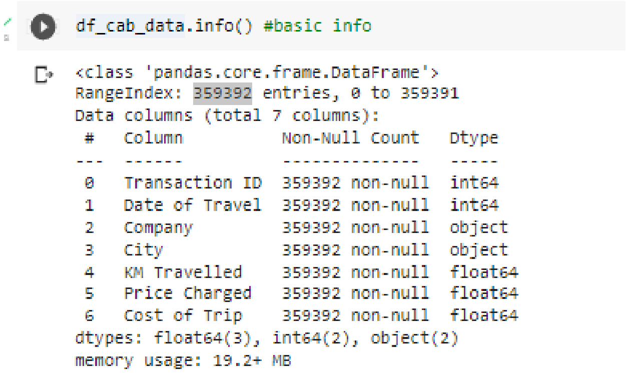
Tabular data details:

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	<59.9 MB>

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

Proposed Approach:

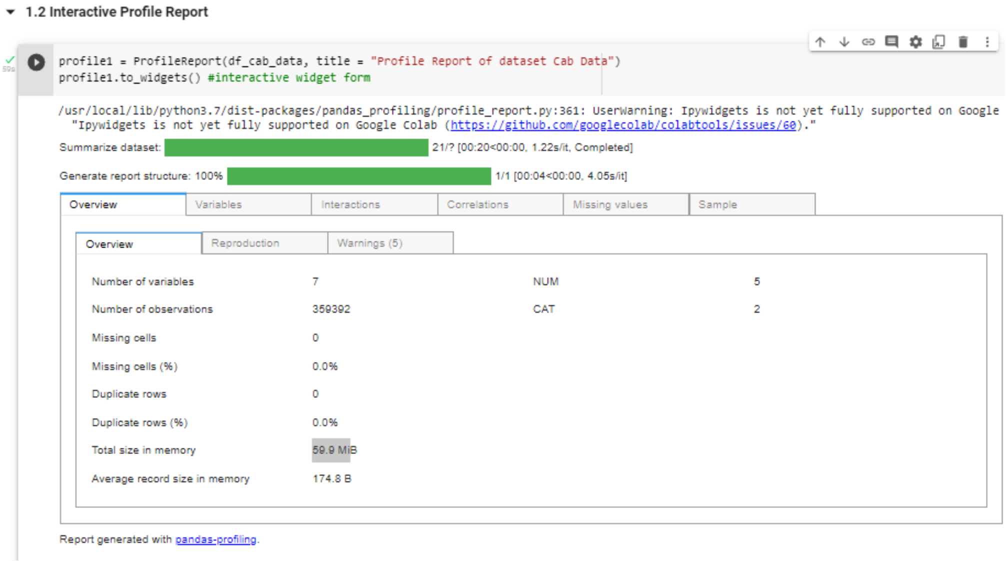
- Two methods were used to explore initial characteristics (univariate analysis)
  1. **Pandas code : df. Info**
     Sample screeenshot



```
df_cab_data.info() #basic info
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 359392 entries, 0 to 359391
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   Transaction ID  359392 non-null  int64  
 1   Date of Travel  359392 non-null  int64  
 2   Company         359392 non-null  object 
 3   City            359392 non-null  object 
 4   KM Travelled    359392 non-null  float64
 5   Price Charged   359392 non-null  float64
 6   Cost of Trip    359392 non-null  float64
dtypes: float64(3), int64(2), object(2)
memory usage: 19.2+ MB
```

  2. **Pandas Profiling: widget command of pandas profiling module**
     Sample screenshot



1.2 Interactive Profile Report

```
profile1 = ProfileReport(df_cab_data, title = "Profile Report of dataset Cab Data")
profile1.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
  "Ipywidgets is not yet fully supported on Google Colab (https://github.com/googlecolab/colabtools/issues/60)."
Summarize dataset: 21/? [00:20<00:00, 1.22s/it, Completed]
Generate report structure: 100% 1/1 [00:04<00:00, 4.05s/it]
```

Number of variables	7	NUM	5
Number of observations	359392	CAT	2
Missing cells	0		
Missing cells (%)	0.0%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	59.9 MiB		
Average record size in memory	174.8 B		

Report generated with pandas-profiling.