Data Intake Report

Name: G2M Case Study
Report date: 6/13/2022
Internship Batch: LISUM10: 30
Version: 1.0
Data intake by: Jefferson Pelera
Data storage location: https://github.com/JeffersonTPelera/DataGlacier/tree/main/week2

Tabular data details:

"Cab_Data.csv

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	csv
Size of the data	22418 KB

city.csv

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

Customer_Id.csv

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	csv
Size of the data	1027 KB

Transaction_Id.csv

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	csv
Size of the data	8788 KB

master.csv

Total number of observations	359392
Total number of files	
Total number of features	14
Base format of the file	csv
Size of the data	42293 KB

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

To find the number of rows and columns of a data file you use .shape.

To merge all the files into one
master = pd.DataFrame(cab_data)
master = pd.merge(master, transaction_id, on=['Transaction ID'])
master = pd.merge(master, city, on=['City'])
master = pd.merge(master, customer_id, on=['Customer ID'])

All the merges are inner by default.
I then used .shape to make sure it has the same number of rows as the Cab_Data.csv, which it does.

To remove duplicates you would just need to write it in the format below with df being the name of your dataframe in my case it would just be master
df = df.drop_duplicates(subset=[columns…..], keep=False)

I wrote:

```
unique_customer = master.groupby(['Company'])['Customer ID'].nunique()
```

to find the number of unique customers in the master file.

Questions Asked:
Which company has more users?
Where do those users mainly reside?
How much money(profit) does each company earn?
Does each company have the same rate for expenses?
What is the demographic of the customers?