Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Data Glacier

Your Deep Learning Partner

G2M insight for Cab Investment firm

Joseph Nnodim

13th August 2022

Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Agenda

1. Introduction
2. Data understanding
3. Data Exploration and Data Quality Check
4. Data Cleaning
5. Statistical Analysis
6. Interpretation

Introduction

Business Analysis

Problem definition

XYZ, a private firm in US needs to gain insights into the cab industry and leverage opportunities in the growing market, as per their Go-to-Market(G2M) strategy. This will inform its decision on the right investment.

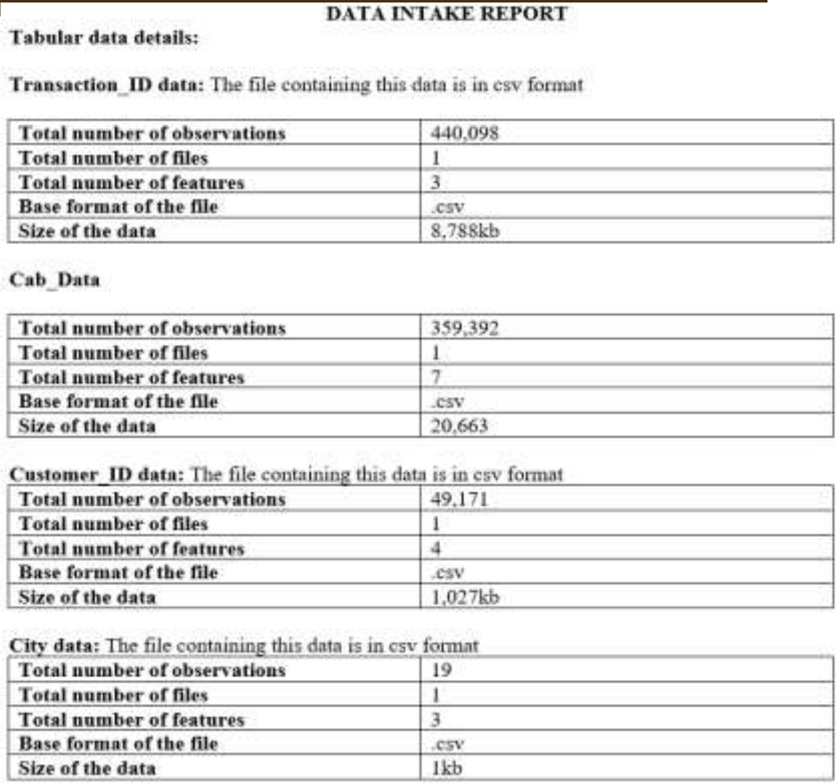
Objectives

- To investigate and understand the dataset in terms of schema, structure and quality
- To handle existing data quality issues
- To carry out exploratory through visualization and analytical approaches for the two companies

-To recommend the better company for XYZ’s investment that will drive the most value and attain the highest profit.

Solution Requirements

Explore, transform, analysis and generate insight from data using statistical techniques.



DATA INTAKE REPORT

Tabular data details:

Transaction_ID data: The file containing this data is in csv format

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788kb

Cab_Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663

Customer_ID data: The file containing this data is in csv format

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027kb

City data: The file containing this data is in csv format

Total number of observations	19
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1kb

Data Exploration- Understanding the data

Schema

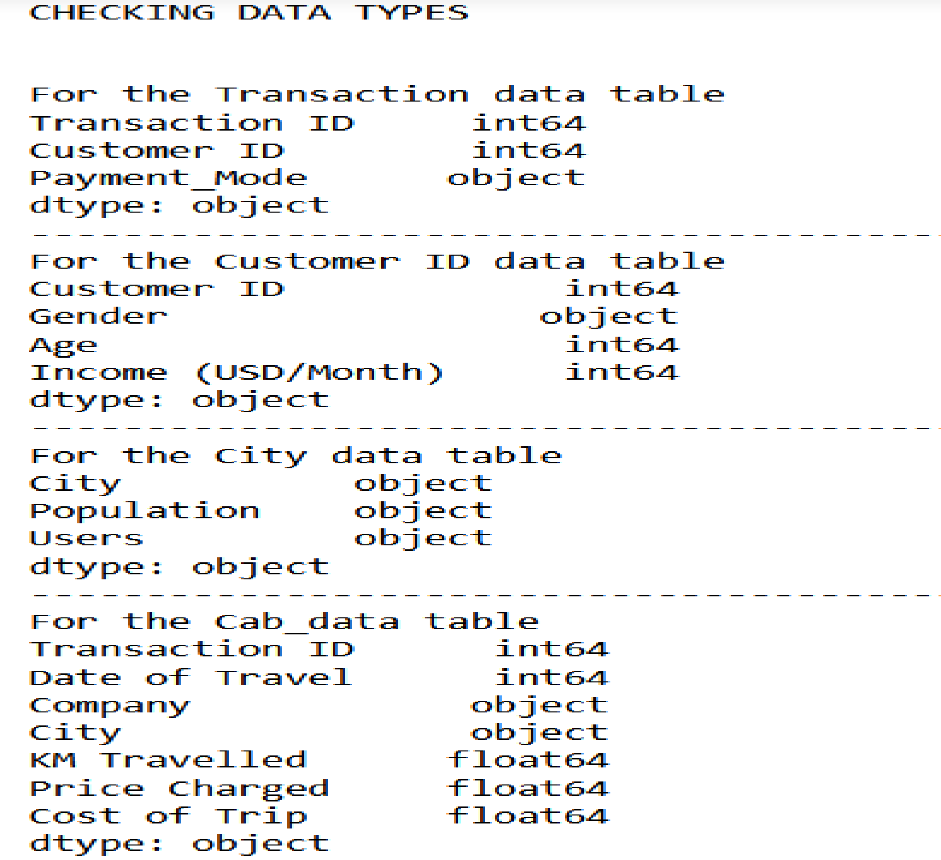
The schema was checked to know the **data types** for each of the tables, which will inform the type of analysis that will be carried out on each column. The schema analysis shows that:

In the Customer demographic table,

-Name, gender, job title, job industry category, wealth segment, deseaced indicator, owns_car, address, state and country are text data types

-Customer id, past_3_years_bike_related_purchases and tenure are integer data types

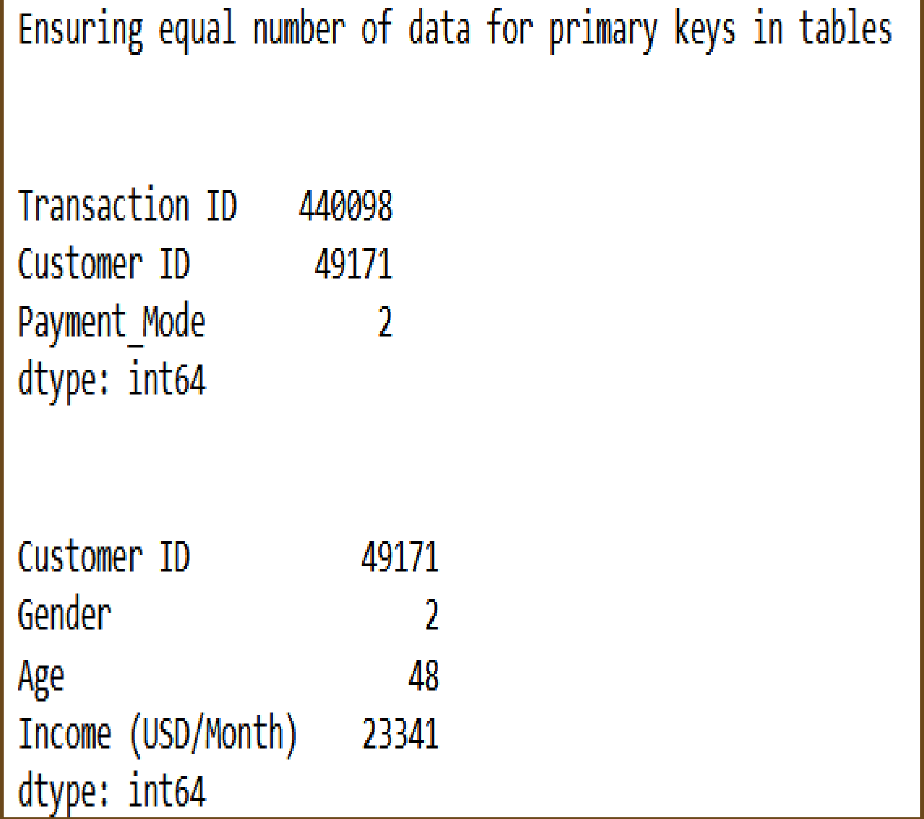
-DOB is date data types



```
CHECKING DATA TYPES


For the Transaction data table
Transaction ID      int64
Customer ID         int64
Payment_Mode       object
dtype: object
--------------------------------------------------
For the Customer ID data table
Customer ID                 int64
Gender                     object
Age                         int64
Income (USD/Month)          int64
dtype: object
--------------------------------------------------
For the City data table
City          object
Population    object
Users         object
dtype: object
--------------------------------------------------
For the Cab_data table
Transaction ID      int64
Date of Travel      int64
Company            object
City               object
KM Travelled      float64
Price Charged     float64
Cost of Trip      float64
dtype: object
```

Data Exploration- Checking for consistency amongst tables

Customer ID occur in both the Transaction data table and the customer id table, so some checks was done to ensure the number of values for the customer IDs column in both tables



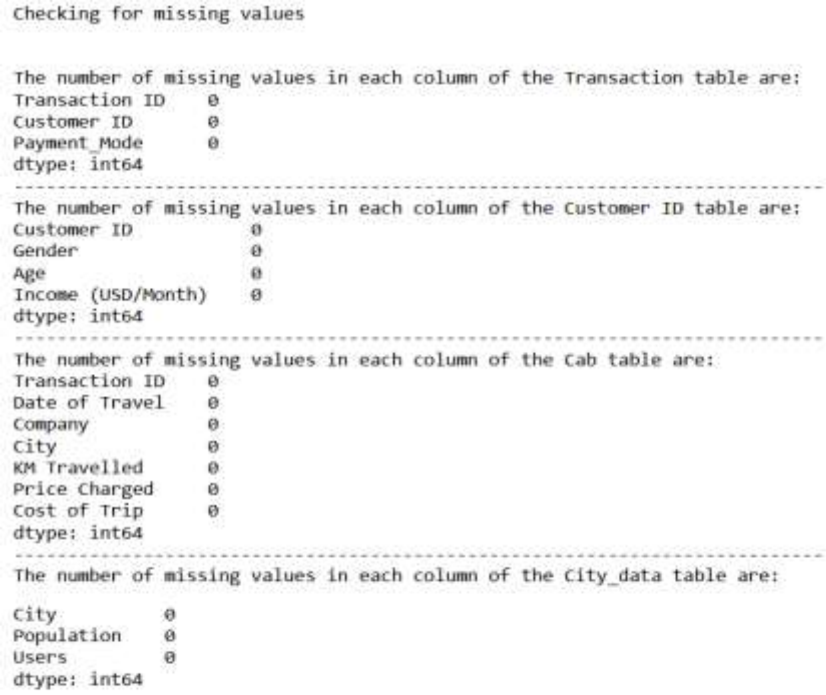
```
Ensuring equal number of data for primary keys in tables



Transaction ID    440098
Customer ID        49171
Payment_Mode           2
dtype: int64



Customer ID            49171
Gender                     2
Age                       48
Income (USD/Month)     23341
dtype: int64
```

Data Exploration- Data Quality Check continues…

Missing Values

A comprehensive exploration of the dataset was done to check for quality issues and gain a deep understanding of the properties, qualities and relationship between features in the data.

The data was checked for missing values

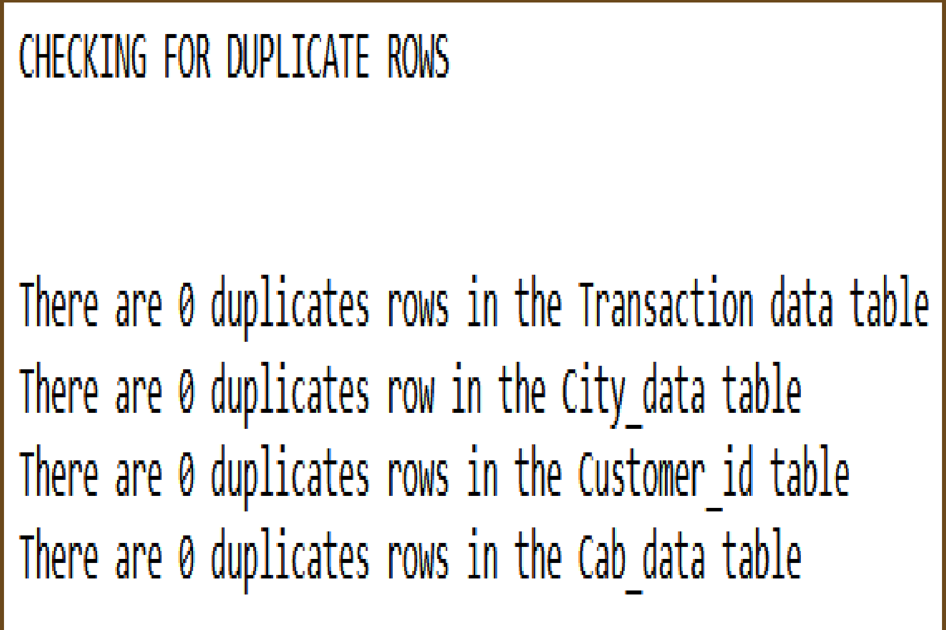


```
Checking for missing values


The number of missing values in each column of the Transaction table are:
Transaction ID    0
Customer ID       0
Payment_Mode      0
dtype: int64
------------------------------------------------------------------------------------------
The number of missing values in each column of the Customer ID table are:
Customer ID           0
Gender                0
Age                   0
Income (USD/Month)    0
dtype: int64
------------------------------------------------------------------------------------------
The number of missing values in each column of the Cab table are:
Transaction ID    0
Date of Travel    0
Company           0
City              0
KM Travelled      0
Price Charged     0
Cost of Trip      0
dtype: int64
------------------------------------------------------------------------------------------
The number of missing values in each column of the City_data table are:

City          0
Population    0
Users         0
dtype: int64
```

Data Exploration- Data Quality Check continues…

Duplicate rows

The data was checked for duplicate rows



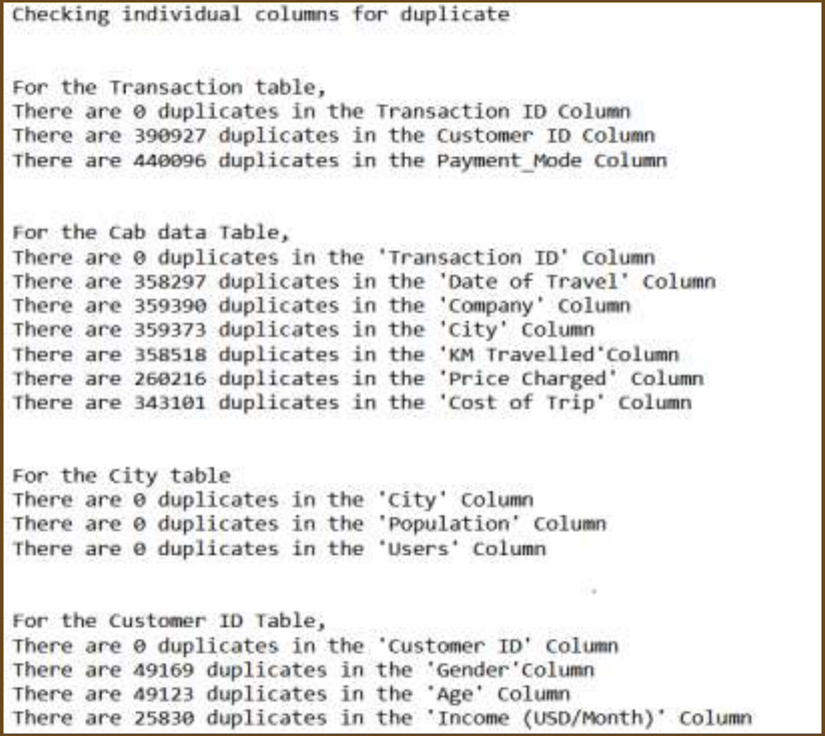
```
CHECKING FOR DUPLICATE ROWS



There are 0 duplicates rows in the Transaction data table
There are 0 duplicates row in the City_data table
There are 0 duplicates rows in the Customer_id table
There are 0 duplicates rows in the Cab_data table
```

Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Data Exploration- Data Quality Check continues…

Duplicate in individual columns

Individual columns were checked for duplicates



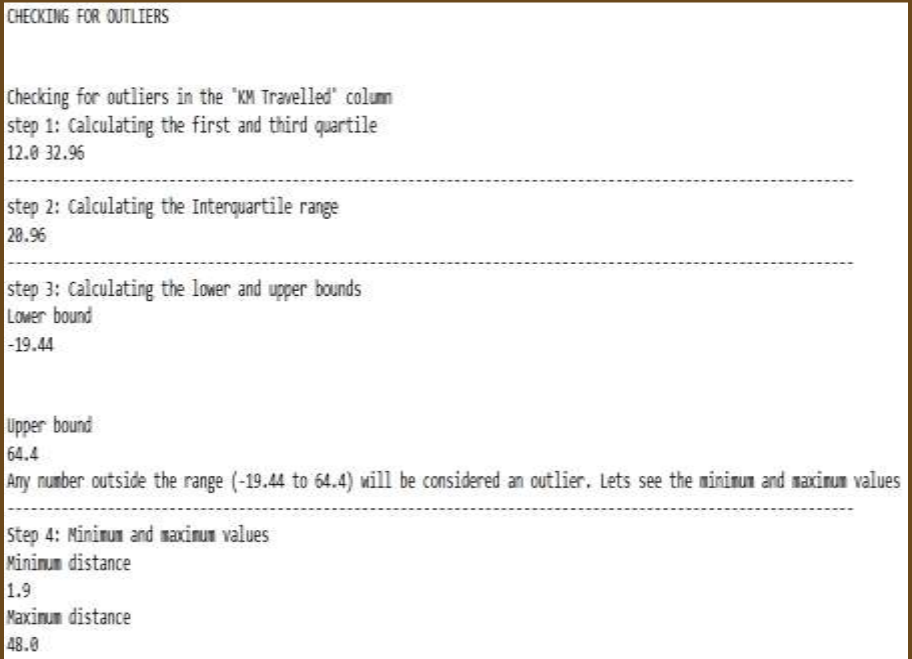
```
Checking individual columns for duplicate


For the Transaction table,
There are 0 duplicates in the Transaction ID Column
There are 390927 duplicates in the Customer ID Column
There are 440096 duplicates in the Payment_Mode Column


For the Cab data Table,
There are 0 duplicates in the 'Transaction ID' Column
There are 358297 duplicates in the 'Date of Travel' Column
There are 359390 duplicates in the 'Company' Column
There are 359373 duplicates in the 'City' Column
There are 358518 duplicates in the 'KM Travelled'Column
There are 260216 duplicates in the 'Price Charged' Column
There are 343101 duplicates in the 'Cost of Trip' Column


For the City table
There are 0 duplicates in the 'City' Column
There are 0 duplicates in the 'Population' Column
There are 0 duplicates in the 'Users' Column


For the Customer ID Table,
There are 0 duplicates in the 'Customer ID' Column
There are 49169 duplicates in the 'Gender'Column
There are 49123 duplicates in the 'Age' Column
There are 25830 duplicates in the 'Income (USD/Month)' Column
```

Data Exploration- Data Quality Check continues…

Outliers

The data was checked for the presence of outliers.

No Outlier detected



```
CHECKING FOR OUTLIERS


Checking for outliers in the 'KM Travelled' column
step 1: Calculating the first and third quartile
12.0 32.96
----------------------------------------------------------------------------------------------------
step 2: Calculating the Interquartile range
20.96
----------------------------------------------------------------------------------------------------
step 3: Calculating the lower and upper bounds
Lower bound
-19.44


Upper bound
64.4
Any number outside the range (-19.44 to 64.4) will be considered an outlier. Lets see the minimum and maximum values
----------------------------------------------------------------------------------------------------
Step 4: Minimum and maximum values
Minimum distance
1.9
Maximum distance
48.0
```

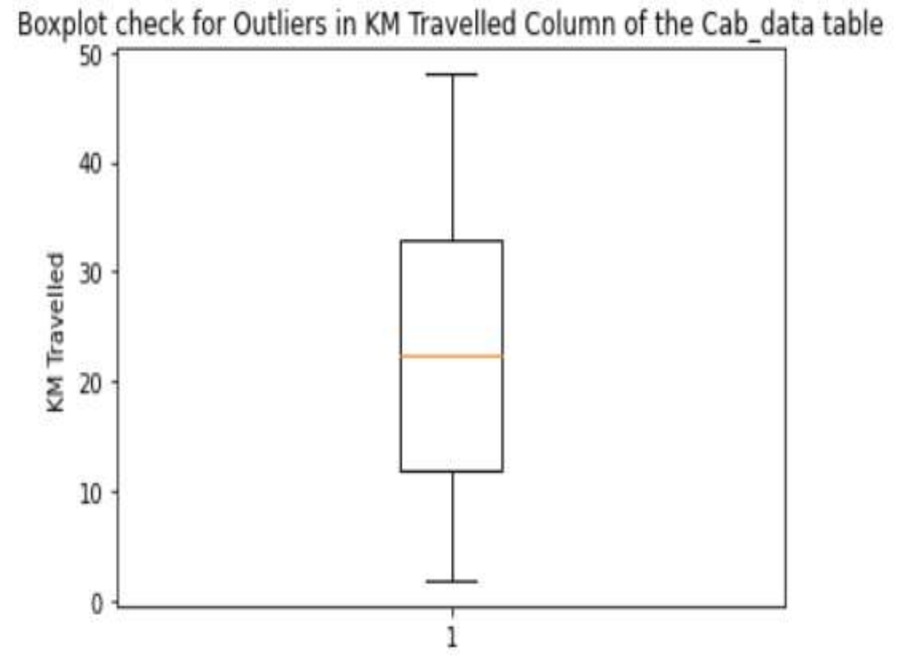
Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Data Exploration- Data Quality Check continues…

Outliers …

The data was visualised for the presence of outliers.

No Observed extreme identified for the KM Travelled column



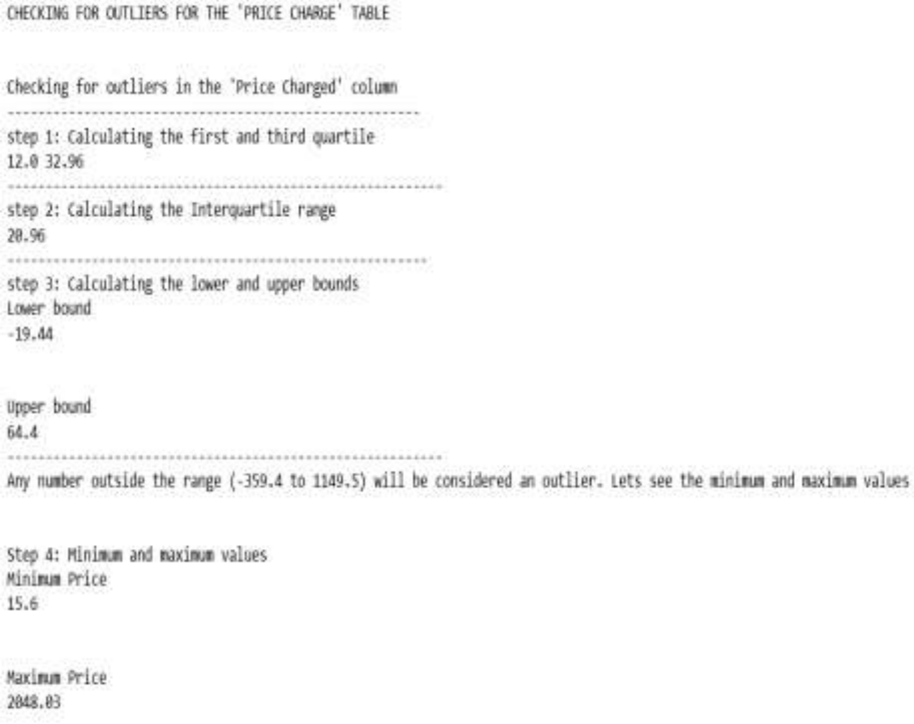
Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Data Exploration- Data Quality Check continues…

Outliers …

The ‘Price Charged’ was visualized for the presence of outliers.

Extreme identified for the Price Charged



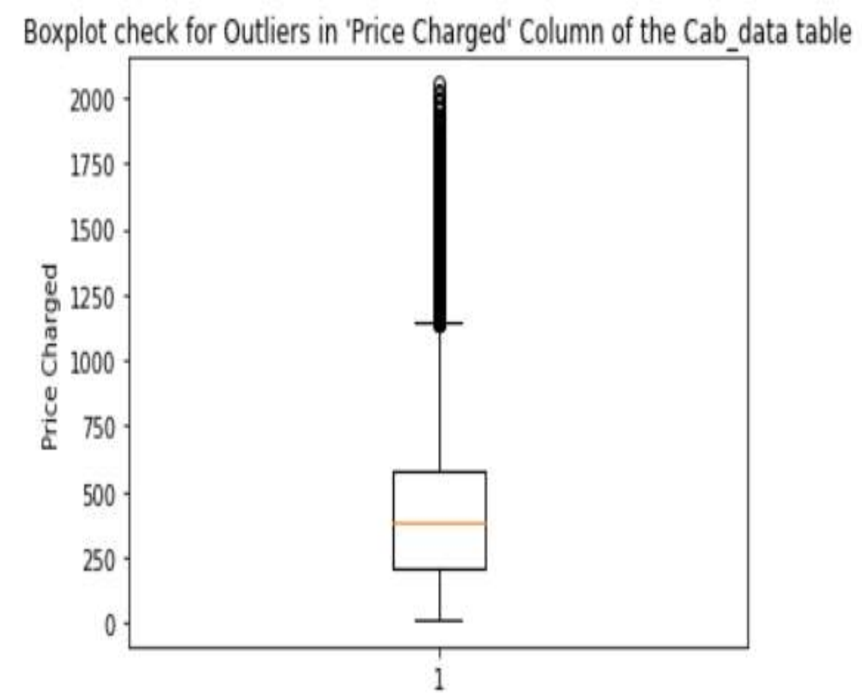
```
CHECKING FOR OUTLIERS FOR THE 'PRICE CHARGE' TABLE


Checking for outliers in the 'Price Charged' column
------------------------------------------------------------
step 1: Calculating the first and third quartile
12.0 32.96
----------------------------------------------------------------
step 2: Calculating the Interquartile range
20.96
--------------------------------------------------------------
step 3: Calculating the lower and upper bounds
Lower bound
-19.44


Upper bound
64.4
----------------------------------------------------------------
Any number outside the range (-359.4 to 1149.5) will be considered an outlier. Lets see the minimum and maximum values


Step 4: Minimum and maximum values
Minimum Price
15.6


Maximum Price
2048.03
```

Data Exploration- Data Quality Check continues…

Outliers …

The ‘Price Charged’ was visualized for the presence of outliers.

Extreme identified for the Price Charged

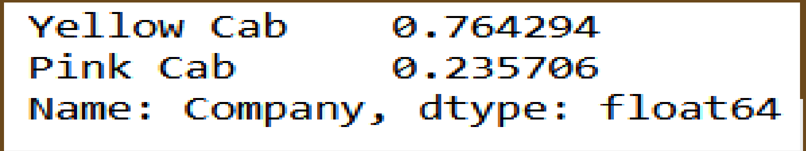


Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

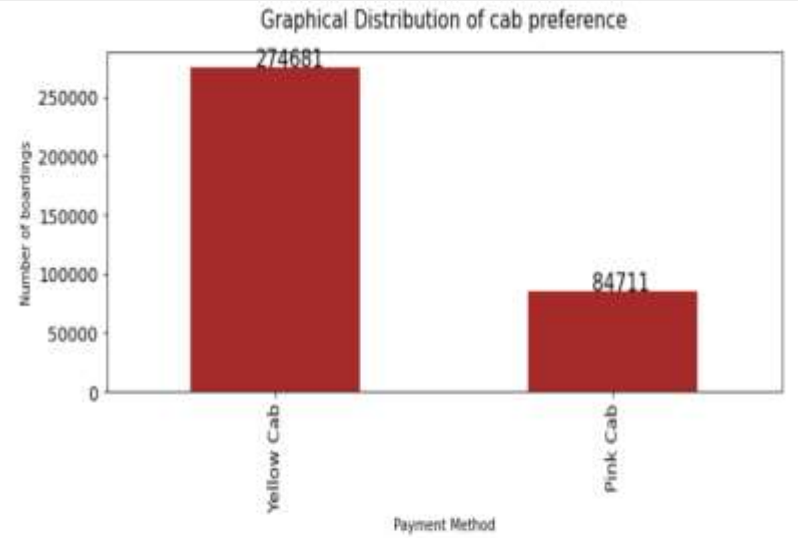
Data Exploration- Statistical Analysis

Duplicate in individual columns

The two companies were compared based on patronage



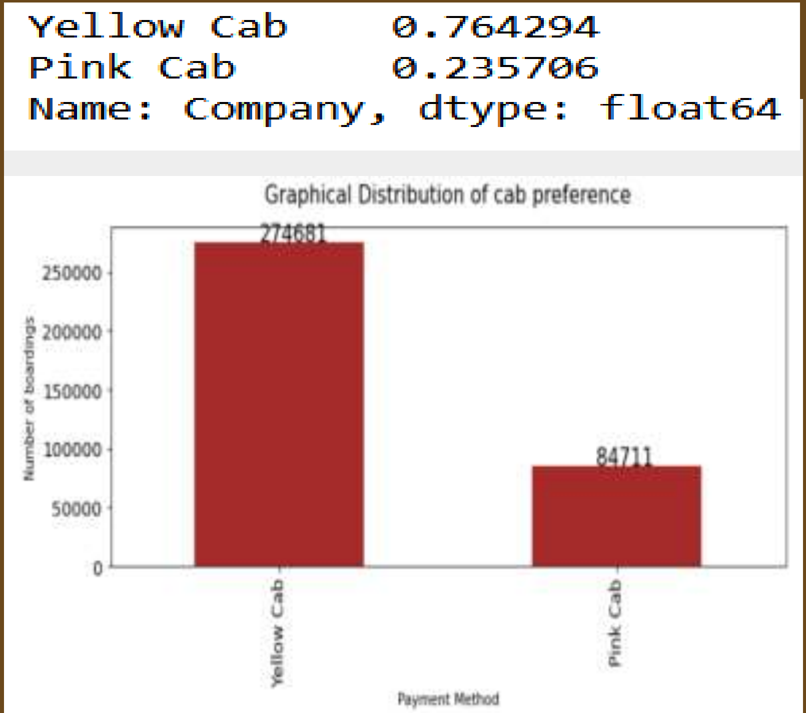
```
Yellow Cab    0.764294
Pink Cab      0.235706
Name: Company, dtype: float64
```



Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

DInterpretation

Based on the current analysis and the number of people who boarded the cab, the yellow cab is recommended for investment



```
Yellow Cab     0.764294
Pink Cab       0.235706
Name: Company, dtype: float64
```

Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 13th August 2022
Internship Batch: LISUM12: 30 July - 30 October 2022
Version: 1.0
Data intake by: Joseph Nnodim
Data intake reviewer:
Data storage location:

DATA INTAKE REPORT

Tabular data details:

Transaction_ID data: The file containing this data is in csv format

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788kb

Cab_Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663

Customer_ID data: The file containing this data is in csv format

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027kb

City data: The file containing this data is in csv format

Total number of observations	19
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1kb

Proposed Approach:

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