Task 6.1: Sourcing Open Data

Sruthy Sreekanth

Bank Customer Segmentation:

- **Data Sourcing:** This data is an open-source data and downloaded from Kaggle.com. The dataset includes Customer demographics and transactions data from an Indian Bank.

- **Data Collection:** The data was collected as a part of research project in collaboration with a bank. Since it was shared by a bank as part of a research project in 2016 and not uptodate, the analysis will not reflect the recent trends.

- **Data Contents:** This dataset includes more than 800,000 clients' transactions totalling over a million from an Indian bank throughout the months of August and October of 2016. It contains information such as Transaction ID, Customer ID, Customer age (DOB), Location, Gender, Account balance at the time of the transaction, Transaction date & time and Transaction amount in INR.

- **Data Limitations:** Since this data was collected from bank in 2016, the information is only available for the year 2016 and not for recent years. The dataset has many missing values in columns- CustomerDOB, CustGender, and CustAccount Balance. Additionally, the incorrect DOB of 1/1/1800 was entered in column CustomerDOB, 36367 data.

- **Data Ethics:** The data does not include any private information about bank customers, which is ethically acceptable.
- **Data Relevance:** This is a reliable source of information for my analysis. It is helpful to perform customer segmentation analysis to identify popular customer groups based on specific features, perform location analysis & transaction related analysis. The bank marketing team will utilize the findings to develop a new marketing strategy.

Data Profiling:

- The dataset has 1048567 rows and 9 columns. During EDA, found that the transaction time code is in HHMMSS format. So converted the same to readable format using date time library.
- I have created a subset of data, because I am analyzing the data with locations having number of transactions >5000. So the new dataset contains 660304 rows and 9 colum ns.
- There are no duplicate records in dataset.

		Data Types			
Variables	Description	time -variant/-invariant	structured/unstructured	qualitative/quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
TransactionID	Unique identifier for Transaction	Time -invariant	Structured	Qualitative	ordinal
CustomerID	Unique identifier for Customer	Time -invariant	Structured	Qualitative	ordinal
CustomerDOB	Customer's age at the time of the transaction	Time-Variant	Structured	Quantitative	Continuous
CustGender	Gender of the customer	Time-invariant	Structured	Qualitative	Nominal
CustLocation	Location where the transaction took place	Time-invariant	Structured	Qualitative	Nominal
CustAccountBalance	Account balance of customer at the time of transaction	Time-invariant	Structured	Quantitative	Discrete/Continuous
TransactionDate	Date of transaction	Time-invariant	Structured	Quantitative	Discrete/Continuous
TransactionTime	Time of Transaction	Time-invariant	Structured	Quantitative	Discrete/Continuous
TransactionAmount	Amount of Transaction	Time-invariant	Structured	Quantitative	Discrete/Continuous

Challenges faced during Data Cleaning:

1. The author of this dataset mentioned that the 'TransactionTime' in dataset is in Unix timestamp. So I tried to convert the TransactionTime from Unix Timestamp to readable format using below command: But the output shows a timestamp with year 1970, which is wrong.

```
# TransactionTime is in Unix timestamp.So changing to readable time format
df_bank['TransactionTime'] =pd.to_datetime(df_bank['TransactionTime'], unit='s')
```

   Then I understood that the time code is in HHMMSS format, not a Unix timestamp. So I converted the same to readable format using datetime library.

```
# TransactionTime is in HHMMSS format.So changing to readable time format
df_bank['TransactionTime'] = df_bank['TransactionTime'].apply(lambda x : datetime.utcfromtimestamp(int(x)).strftime('%H:%M:%S'))
```

2. Two Locations in column 'CustLocation' were modified to follow uniform naming prac tices. So replaced DELHI to NEW DELHI & NAVI MUMBAI to MUMBAI.

3. The missing values in the columns (CustomerDOB, CustGender) may be the result of cu stomer not providing the information, while the column (CustAccountBalance) may be due to zero balance. So, I am imputing missing values in DOB and Gender with 'NA' an d in CustAccountBalance with '0'.

Questions to explore:

- How many customers in bank based on age, Gender and locations?
- How many transactions on daily basis or per month ?
- How many transactions distributed over locations?
- What is the total amount of transaction per month/daily/location?
- which customer gender has been doing the most transactions?
- Monthly comparison of spending habits of male & female/ spending habits across location?