NAME: OGWENO Washington Odhiambo
REG. NO.: 2024/HD05/21955K
STUDENT NO.: 2400721955

RETAIL BANKING DATASET

1. Introduction

The Exploratory Data Analysis have been performed on the Retail Banking Dataset to help in identifying any outlier data points and to understand the relationships between various attributes and structures within the dataset to draw logical findings and conclusions.

It further helps in framing questions and visualizing the results while paving the way to make an informed choice of the machine learning algorithm based on Client behaviour patterns within a Retail Banking Sector.

Questions:

Client Segmentations based on Age, Sex, Geographical locations, Professional and Transactional history.

CRM: Customers Support based the number of issues raised, priority assigned and Resolution.

Products: Highest and Lowest Product Consumed by Clients based on Segmentations.

Table of Contents

2. Dataset Description: Retail Banking

The dataset provides data from simulated environment of a retail banking, revolving around a original 1999 Czech banking dataset. The dataset consists of various files that have been stitched together to mimicked real-world data sources.
Any gaps identified in this process will lead to Dataset modifications and translation to suit the purpose of the assignment while maintaining its significant objectives.

3. Data uploads & Explorations

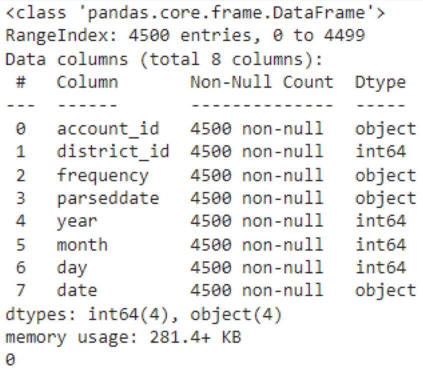
This stage involved loading raw Retail Banking CSV files and the use of Python and Panda libraries to perform basic explorations. e clean or transform to suit analysis.

Package Setups

```
!pip install missingno
import missingno as msno
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

1.completedacct.csv

```
# Loading & Exploring 1.completedacct.csv'dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/1.completedacct.csv')
df.info()
df.duplicated().sum()
```



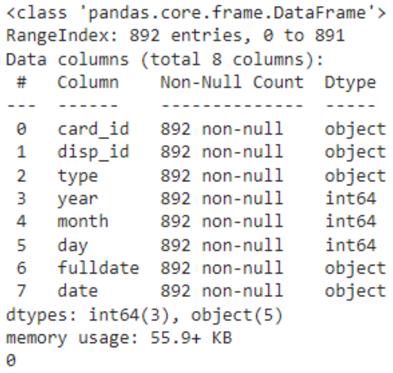
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4500 entries, 0 to 4499
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   account_id   4500 non-null   object
 1   district_id  4500 non-null   int64 
 2   frequency    4500 non-null   object
 3   parseddate   4500 non-null   object
 4   year         4500 non-null   int64 
 5   month        4500 non-null   int64 
 6   day          4500 non-null   int64 
 7   date         4500 non-null   object
dtypes: int64(4), object(4)
memory usage: 281.4+ KB
0
```

Observation 1: No missing values and duplicated rows.

Column Day, month & Year can be replaced by only column **date**.

2.completedcard.csv

```
# Loading & Exploring 2.completedcard.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/2.completedcard.csv')
df.info()
df.duplicated().sum()
```



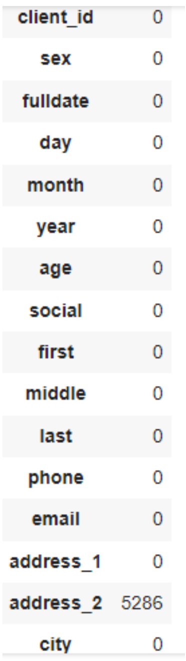
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 892 entries, 0 to 891
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   card_id   892 non-null    object
 1   disp_id   892 non-null    object
 2   type      892 non-null    object
 3   year      892 non-null    int64 
 4   month     892 non-null    int64 
 5   day       892 non-null    int64 
 6   fulldate  892 non-null    object
 7   date      892 non-null    object
dtypes: int64(3), object(5)
memory usage: 55.9+ KB
0
```

Observation 2: No missing values and duplicated rows.

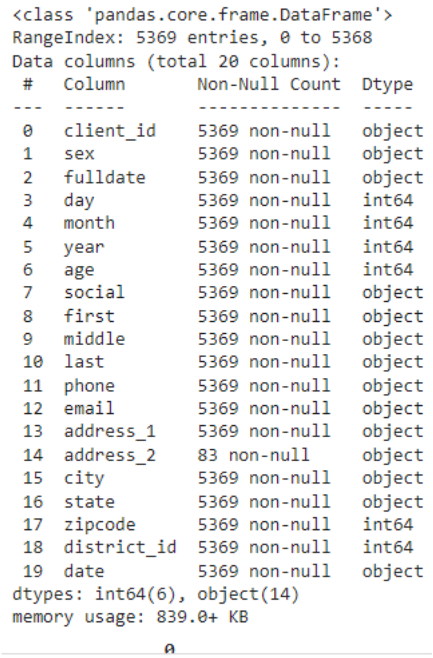
Column Day, month & fullyear can be replaced by only column **date**.

3.completedclient.csv

```
# Loading & Exploring 3.completedclient dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/3.completedclient.csv')
df.info()
df.duplicated().sum()
df.isnull().sum()
```



client_id	0
sex	0
fulldate	0
day	0
month	0
year	0
age	0
social	0
first	0
middle	0
last	0
phone	0
email	0
address_1	0
address_2	5286
city	0

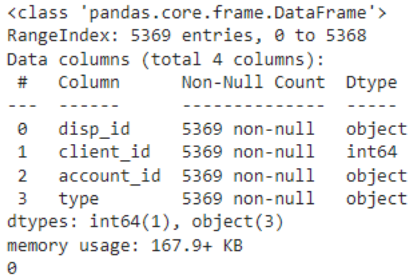


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5369 entries, 0 to 5368
Data columns (total 20 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   client_id    5369 non-null   object
 1   sex          5369 non-null   object
 2   fulldate     5369 non-null   object
 3   day          5369 non-null   int64
 4   month        5369 non-null   int64
 5   year         5369 non-null   int64
 6   age          5369 non-null   int64
 7   social       5369 non-null   object
 8   first        5369 non-null   object
 9   middle       5369 non-null   object
 10  last         5369 non-null   object
 11  phone        5369 non-null   object
 12  email        5369 non-null   object
 13  address_1    5369 non-null   object
 14  address_2    83 non-null     object
 15  city         5369 non-null   object
 16  state        5369 non-null   object
 17  zipcode      5369 non-null   int64
 18  district_id  5369 non-null   int64
 19  date         5369 non-null   object
dtypes: int64(6), object(14)
memory usage: 839.0+ KB
                0
```

Observation 3: address_2 column is missing 5286 records.
Column Day, month & year can be replaced by only column **fulldate**.

4.completeddisposition.csv

```
# Loading & Exploring 4.completeddisposition.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/4.completeddisposition.csv')
df.info()
df.duplicated().sum()
df.hist()
```

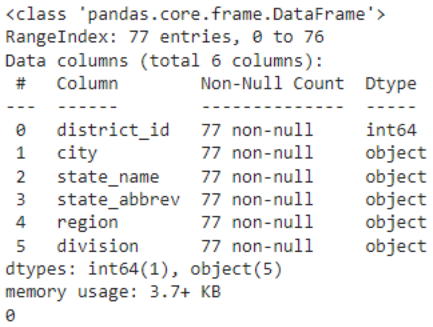


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5369 entries, 0 to 5368
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   disp_id     5369 non-null   object
 1   client_id   5369 non-null   int64
 2   account_id  5369 non-null   object
 3   type        5369 non-null   object
dtypes: int64(1), object(3)
memory usage: 167.9+ KB
0
```

Observation 4: No missing values and duplicated rows.

5.completeddistrict.csv

```
# Loading & Exploring 5.completeddistrict.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/5.completeddistrict.csv')
df.info()
df.duplicated().sum()
```

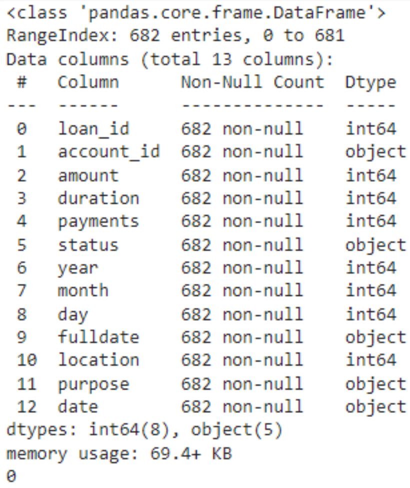


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77 entries, 0 to 76
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   district_id  77 non-null     int64 
 1   city         77 non-null     object
 2   state_name   77 non-null     object
 3   state_abbrev  77 non-null    object
 4   region       77 non-null     object
 5   division     77 non-null     object
dtypes: int64(1), object(5)
memory usage: 3.7+ KB
0
```

Observation 5: No missing values and duplicated rows.

6.completedloan.csv

```
# Loading & Exploring 6.completedloan.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-
7103/main/Retail%20Banking/data/raw/6.completedloan.csv')
df.info()
df.duplicated().sum()
```

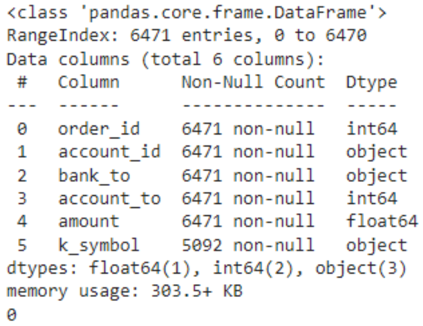


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 682 entries, 0 to 681
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   loan_id     682 non-null    int64 
 1   account_id  682 non-null    object
 2   amount      682 non-null    int64 
 3   duration    682 non-null    int64 
 4   payments    682 non-null    int64 
 5   status      682 non-null    object
 6   year        682 non-null    int64 
 7   month       682 non-null    int64 
 8   day         682 non-null    int64 
 9   fulldate    682 non-null    object
 10  location    682 non-null    int64 
 11  purpose     682 non-null    object
 12  date        682 non-null    object
dtypes: int64(8), object(5)
memory usage: 69.4+ KB
0
```

Observation 6: No missing values and duplicated rows.
Column Day, month & Year can be replaced by only column **fulldate**.

7.completedorder.csv

```
# Loading & Exploring 7.completedorder.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-
7103/main/Retail%20Banking/data/raw/7.completedorder.csv')
df.info()
df.duplicated().sum()
```

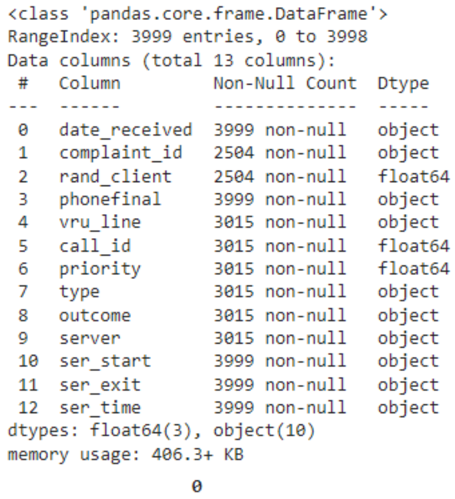


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6471 entries, 0 to 6470
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   order_id    6471 non-null   int64  
 1   account_id  6471 non-null   object 
 2   bank_to     6471 non-null   object 
 3   account_to  6471 non-null   int64  
 4   amount      6471 non-null   float64
 5   k_symbol    5092 non-null   object 
dtypes: float64(1), int64(2), object(3)
memory usage: 303.5+ KB
0
```

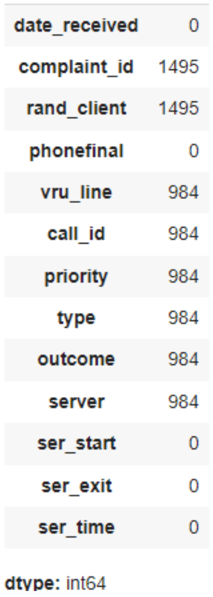
Observation 7: k_symbol column is missing 1379 records.

8.crm_call_center_logs.csv

```
# Loading & Exploring 8.crm_call_center_logs.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/8.crm_call_center_logs.csv')
df.info()
df.duplicated().sum()
df.isnull().sum()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3999 entries, 0 to 3998
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date_received  3999 non-null   object
 1   complaint_id   2504 non-null   object
 2   rand_client    2504 non-null   float64
 3   phonefinal     3999 non-null   object
 4   vru_line       3015 non-null   object
 5   call_id        3015 non-null   float64
 6   priority       3015 non-null   float64
 7   type           3015 non-null   object
 8   outcome        3015 non-null   object
 9   server         3015 non-null   object
 10  ser_start      3999 non-null   object
 11  ser_exit       3999 non-null   object
 12  ser_time       3999 non-null   object
dtypes: float64(3), object(10)
memory usage: 406.3+ KB
                   0
```



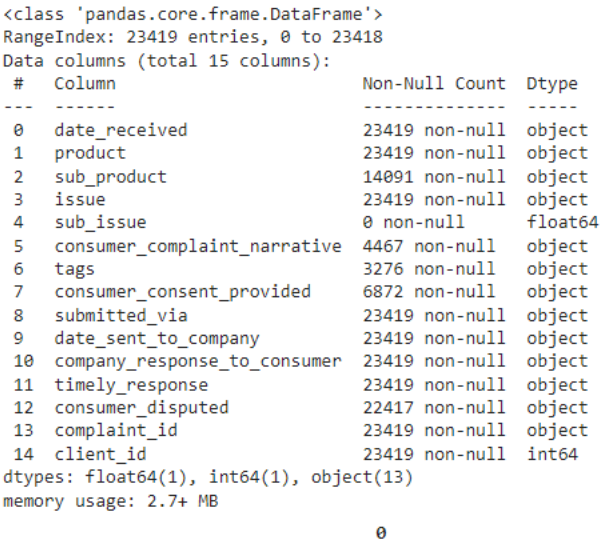
```
date_received     0
complaint_id   1495
rand_client    1495
phonefinal        0
vru_line        984
call_id         984
priority        984
type            984
outcome         984
server          984
ser_start         0
ser_exit          0
ser_time          0

dtype: int64
```

Observation 8: complaint_id column is missing 1495 records.
rand_client column is missing 1495 records.
vru_line column is missing 984 records.
call_id column is missing 984 records.
priority column is missing 984 records.
type column is missing 984 records.
outcome column is missing 984 records.
server column is missing 984 records.

9.crm_events.csv

```
# Loading & Exploring 9.crm_events.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/9.crm_events.csv')
df.info()
df.duplicated().sum()
df.isnull().sum()
```

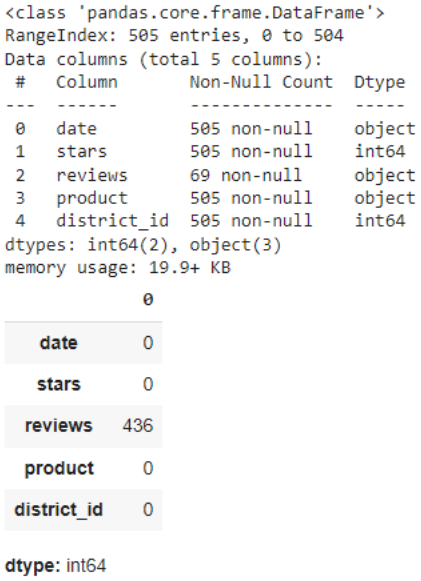


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23419 entries, 0 to 23418
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   date_received                 23419 non-null  object
 1   product                       23419 non-null  object
 2   sub_product                   14091 non-null  object
 3   issue                         23419 non-null  object
 4   sub_issue                     0 non-null      float64
 5   consumer_complaint_narrative  4467 non-null   object
 6   tags                          3276 non-null   object
 7   consumer_consent_provided     6872 non-null   object
 8   submitted_via                 23419 non-null  object
 9   date_sent_to_company          23419 non-null  object
 10  company_response_to_consumer  23419 non-null  object
 11  timely_response               23419 non-null  object
 12  consumer_disputed             22417 non-null  object
 13  complaint_id                  23419 non-null  object
 14  client_id                     23419 non-null  int64
dtypes: float64(1), int64(1), object(13)
memory usage: 2.7+ MB
                                     0
```

Observation 9: sub_product column is missing 1379 records.
sub_issue column has no entries.
consumer_complaint_narrative is missing 18952 records.
tags column is missing 20143 records.
consumer_consent_provided colun is missing 16547 records
consumer_disputed column is missing 1002 records.

10.crm_reviews.csv

```
# Loading & Exploring 10.crm_reviews.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-
7103/main/Retail%20Banking/data/raw/10.crm_reviews.csv')
df.info()
df.duplicated().sum()
df.isnull().sum()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 505 entries, 0 to 504
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   date         505 non-null    object
 1   stars        505 non-null    int64
 2   reviews      69 non-null     object
 3   product      505 non-null    object
 4   district_id  505 non-null    int64
dtypes: int64(2), object(3)
memory usage: 19.9+ KB
```

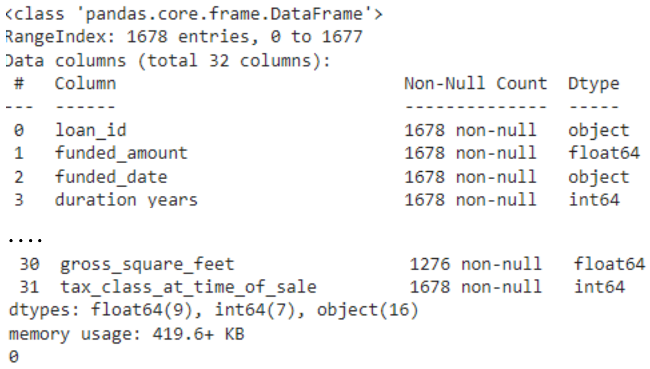
	0
date	0
stars	0
reviews	436
product	0
district_id	0

dtype: int64

Observation 10: reviews column is missing 436 records.

11.luxuryloanportfolio.csv

```
# Loading & Exploring 11.luxuryloanportfolio.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/11.luxuryloanportfolio.csv')
df.info()
df.duplicated().sum()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1678 entries, 0 to 1677
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   loan_id                         1678 non-null   object
 1   funded_amount                   1678 non-null   float64
 2   funded_date                     1678 non-null   object
 3   duration years                  1678 non-null   int64

....
 30  gross_square_feet               1276 non-null   float64
 31  tax_class_at_time_of_sale       1678 non-null   int64
dtypes: float64(9), int64(7), object(16)
memory usage: 419.6+ KB
0
```

Observation 11: No missing values and duplicated rows.

4. Data Cleaning

Observation 1: Column Day, month & Year can be replaced by only column date.

```
# Cleaning dataset 1.completedacct.csv
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/1.completedacct.csv')
df.info()
df.duplicated().sum()
df.drop('year', axis='columns', inplace=True)
df.drop('month', axis='columns', inplace=True)
df.drop('day', axis='columns', inplace=True)
df.drop('parseddate', axis='columns', inplace=True)
df.head()
```

Observation 2: Column Day, month & fullyear can be replaced by only column **date**.

```
# Cleaning 2.completedcard.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/2.completedcard.csv')
df.info()
df.drop('year', axis='columns', inplace=True)
df.drop('month', axis='columns', inplace=True)
df.drop('day', axis='columns', inplace=True)
df.drop('fulldate', axis='columns', inplace=True)
df.head()
```

Observation 3: address_2 column is missing 5286 records.
Column Day, month & year can be replaced by only column **fulldate**.

```
# Cleaning 3.completedclient dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/3.completedclient.csv')
print(round(df.isnull().sum()/len(df) * 100, 1))
df.drop('address_2', axis='columns', inplace=True)
df.info()
df.drop('day', axis='columns', inplace=True)
df.drop('month', axis='columns', inplace=True)
df.drop('year', axis='columns', inplace=True)
df.head()
```

Observation 6: Column Day, month & Year can be replaced by only column date.

```
# Cleaning 6.completedloan.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/6.completedloan.csv')
df.info()
df.drop('day', axis='columns', inplace=True)
df.drop('month', axis='columns', inplace=True)
df.drop('year', axis='columns', inplace=True)
df.drop('date', axis='columns', inplace=True)
df.head()
```

Observation 7: k_symbol column is missing 1379 records.

```
# Replace a missing column k_symbol data in 7.completedorder.csv with mode
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/7.completedorder.csv')
df['k_symbol'].fillna(df['k_symbol'].mode()[0], inplace=True)
df.info()
```

Observation 8: complaint_id column – Modified.
rand_client column – Modified.
vru_line column – Modified.
call_id column – Modified.
priority column – Modified.
type column – Modified.
outcome column – Modified.
server column – Modified.

```
# Cleaning 8.crm_call_center_logs.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-7103/main/Retail%20Banking/data/raw/8.crm_call_center_logs.csv')
print(round(df.isnull().sum()/len(df) * 100, 1))
df['complaint_id'].fillna(df['complaint_id'].mode()[0], inplace=True)
```

```
df['rand_client'].fillna(df['rand_client'].mode()[0], inplace=True)
df['vru_line'].fillna(df['vru_line'].mode()[0], inplace=True)
df['call_id'].fillna(df['call_id'].mode()[0], inplace=True)
df['priority'].fillna(df['priority'].mode()[0], inplace=True)
df['type'].fillna(df['type'].mode()[0], inplace=True)
df['outcome'].fillna(df['outcome'].mode()[0], inplace=True)
df['server'].fillna(df['server'].mode()[0], inplace=True)
df.info()
```

Observation 9: sub_product column - Modified
sub_issue column - Dropped
consumer_complaint_narrative column - Dropped
tags column - Dropped
consumer_consent_provided colum Dropped
consumer_disputed column - Modified.

```
# Cleaning 9.crm_events.csv dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-
7103/main/Retail%20Banking/data/raw/9.crm_events.csv')
print(round(df.isnull().sum()/len(df) * 100, 1))
df['sub_product'].fillna(df['sub_product'].mode()[0], inplace=True)
df.drop('sub_issue', axis='columns', inplace=True)
df.drop('consumer_complaint_narrative', axis='columns', inplace=True)
df.drop('tags', axis='columns', inplace=True)
df.drop('consumer_consent_provided', axis='columns', inplace=True)
df['consumer_disputed'].fillna(df['consumer_disputed'].mode()[0],
inplace=True)
df.info()
```

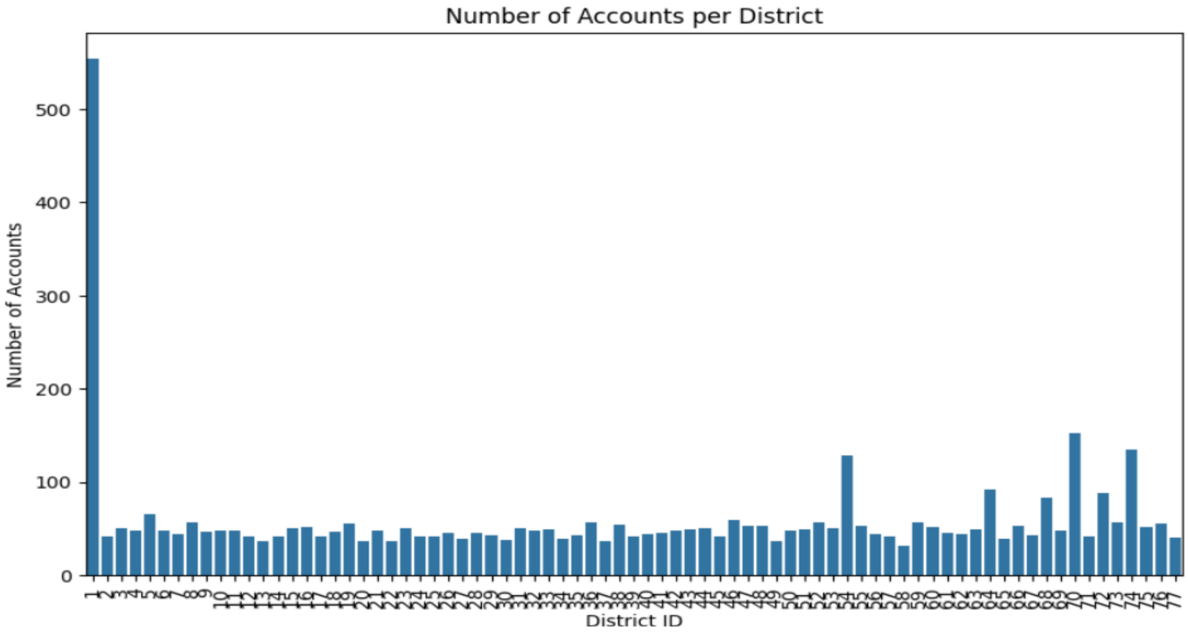
Observation 10: review column is missing 436 records.

```
# Droping address_2 from 3.completedclient dataset
df = pd.read_csv('https://raw.githubusercontent.com/wogweno/MCS-
7103/main/Retail%20Banking/data/raw/3.completedclient.csv')
print(round(df.isnull().sum()/len(df) * 100, 1))
df.drop('address_2', axis='columns', inplace=True)
df.info()
```

5. Examine Data Relationships & Outliers Identifications

.

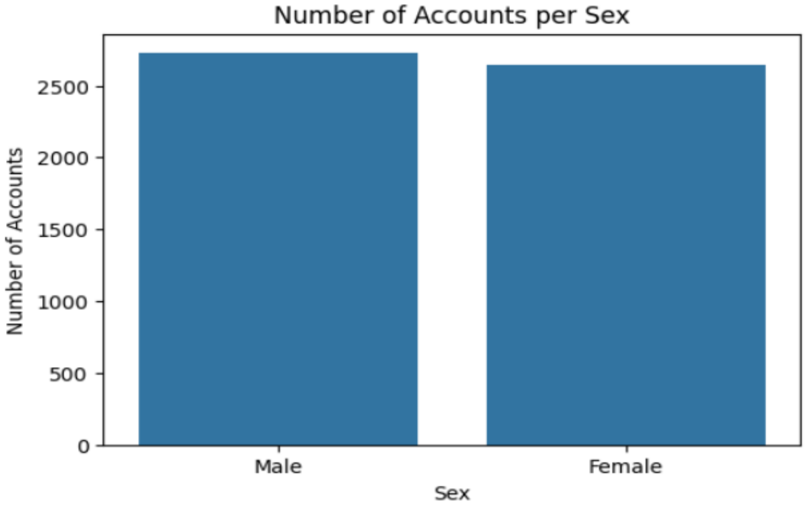
The following were Outliners Derived from the datasets.



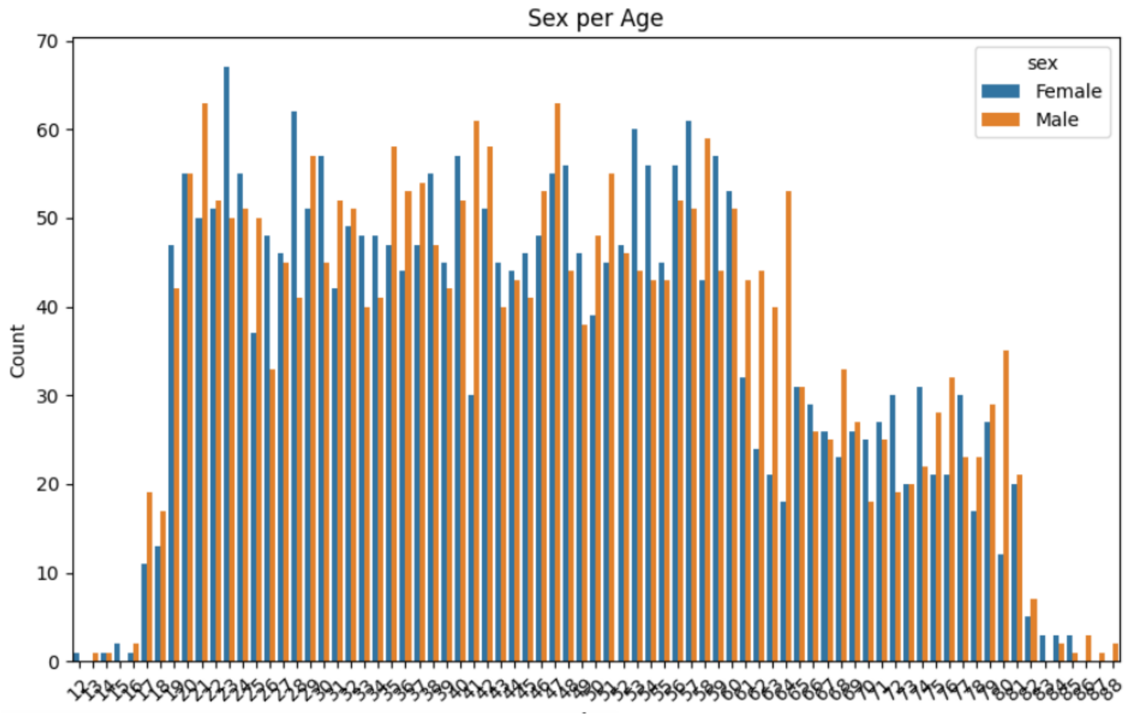
District 1 have most of the accounts.

District 70, 74 & 54 have over 100 account each.

The rest of the districts have less than 100 accounts.

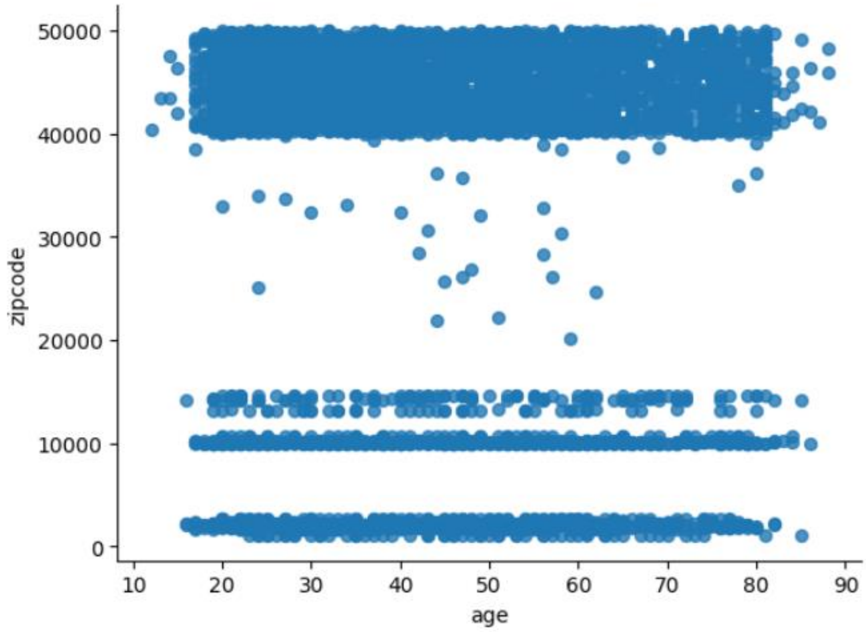


Accounts are fairly distributed among both Sex, however Males have more accounts that Females.

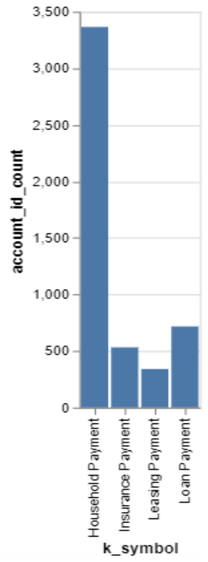


A female has the lowest age 12 with account while a male have the highest age of 88 among the account holders.

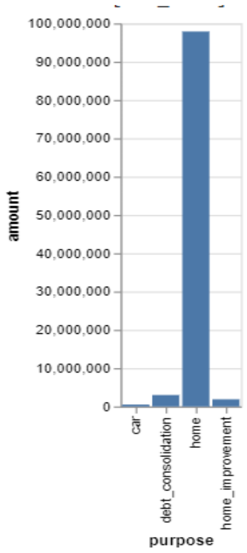
Female of age 23 are the highest account holders with 67 accounts.



Zipcode between 4000 to 5000 have the highest number of account holders.



Payments was dominated by Household payments.



There are more home loan request with car been the least loan made.

At this stage it was clear that the Retail Banking dataset chosen could still be optimized and normalized to produced further to emulate a real Banking Setup by revealing several components below:
Completed Files for Core Banking System: This section contains data related to the core banking system, where accounts are linked by identifiers...

CRM Datasets: Containing data related to customer relationship management, with a focus on customer interactions and complaints. The CRM events text can be parsed for sentiment analysis. Some phone calls from the call Center are matched to CRM event records. Additionally, some phone calls are made from known client numbers, allowing inference of the caller's identity. Certain clients have alternative phone numbers, providing backup contact information.

Loan Datasets: Containing data related to related to different loans products, presenting a good Product and Service template within a Financial Institutions.