RevoU Full-Stack Data Analytics Course

2024

Customer Segmentation

Presented by
Raffael Nathaniel Siregar

Next Slide

Problem Statement

How is the sales performance for each customer in the last 3 years?

Analysis Goals

- To explore sales performance in the last 3 years
- To create customer segmentation

Next Slide

Cleaning Process

Match Data
Dictionary

Correcting
Data Type

Remove
Duplicates

Handle
Null Values

Customer's
Age Filter

Match Data Dictionary

Using .value_counts()

We need to check the given data dictionary with given dataset. In the dataset, we use .value_counts() for categorical column to make sure the defined categories in data dictionary are the same with the dataset

Result

Fortunately, the dataset categories are the same with data dictionary, so we don't need to impute or manipulate anything.

```
df['MOB'] = df['MOB'].astype(int)
df['birth_date'] = pd.to_datetime(df['birth_date'])
df['account_id'] = df['account_id'].astype(str)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12559 entries, 0 to 12558
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   account_id                12559 non-null  object
 1   account_activity_level    12559 non-null  object
 2   customer_value_level      12559 non-null  object
 3   MOB                       12559 non-null  int64
 4   flag_female               12559 non-null  int64
 5   avg_sales_L36M            11820 non-null  float64
 6   cnt_sales_L36M            12559 non-null  int64
 7   last_sales                12559 non-null  float64
 8   month_since_last_sales    12559 non-null  int64
 9   count_direct_promo_L12M   12559 non-null  int64
 10  birth_date                12559 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(5), object(3)
memory usage: 1.1+ MB
```

Correcting Data Type

'birth_date', 'MOB' , and 'account_id' columns are not at the right data type. In order to fix this, we need to change the columns datatype using the functions beside. Because, 'birth_date' column has datetime value and 'MOB' is month differences. This also match with data dictionary. For the 'account_id', we need to change it to string since it's a unique identifier

Remove Duplicates

Found 72 duplicated values

By using df[df.duplicated()] we found 72 duplicated values in this dataset. So, we remove the duplicated values using df = df.drop_duplicates()

No duplicates in ‘account_id’

After handling duplicated values in all columns. we need to ensure that no duplicated ‘account_id’. By using df[’account_id’].duplicated(), no duplicated values are found

Handle Null Values

736 Null Values in ‘avg_sales_L36M’

Using for loop, we iterate through all columns by summing all null values.

Correlate with other column

‘avg_sales_L36M’ correlate with ‘cnt_sales_L36M’. In ‘cnt_sales_L36M’ column, it has 0 values, so it will be more logial if we impute with 0

Customer's Age Filter

Create 'age' column

We need to find the difference between toda's date and customer's birthdate and assign in in age column

Filter 'age' < 21

We filter out the dataframe using
df = df[~(df['age'] < 21)]
Because, people who are less than 21 usually don't make money themselves

End of Milestone 1

Total Sales

Assumptions :

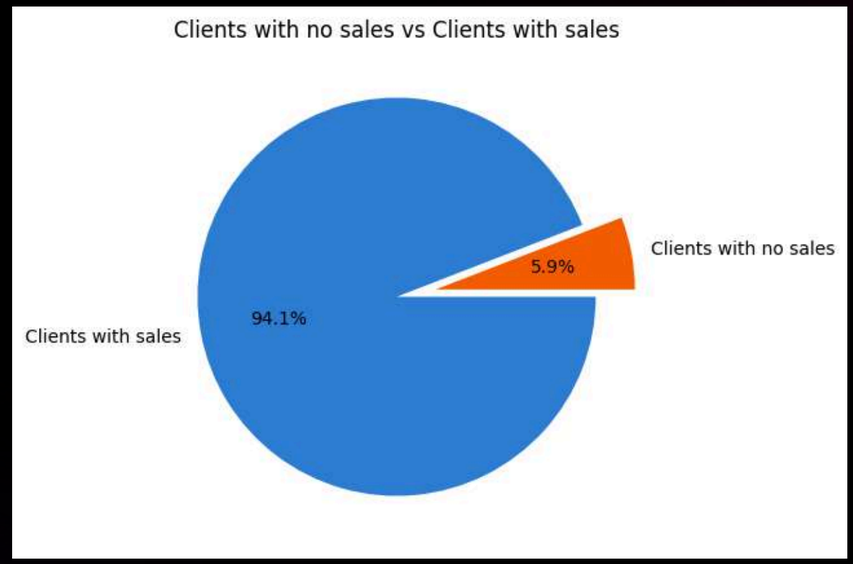
- Total Sales in Euro
- RevoBank is one of the banks located in UK that offers credit card service
- UK has 350 Banks that offer credit card service

Total Sales

The total sales of Credit Card in UK in 1 year (Aug 2023-Jul 2024) is **257.15 billions Euro**. Meanwhile, RevoBank's credit card total sales is **402.58 millions Euro** in 36 months (3 years). It means, RevoBank has an average of **134.2 millions Euro** in 1 year. Based on the assumptions, RevoBank has about **0.05% credit card Revenue Share** (compared to all over UK) and placing RevoBank much below the average of credit card sales each bank in 1 year of **734.71 millions Euro** each bank.

Clients Overview

Clients with Sales
11,631 clients

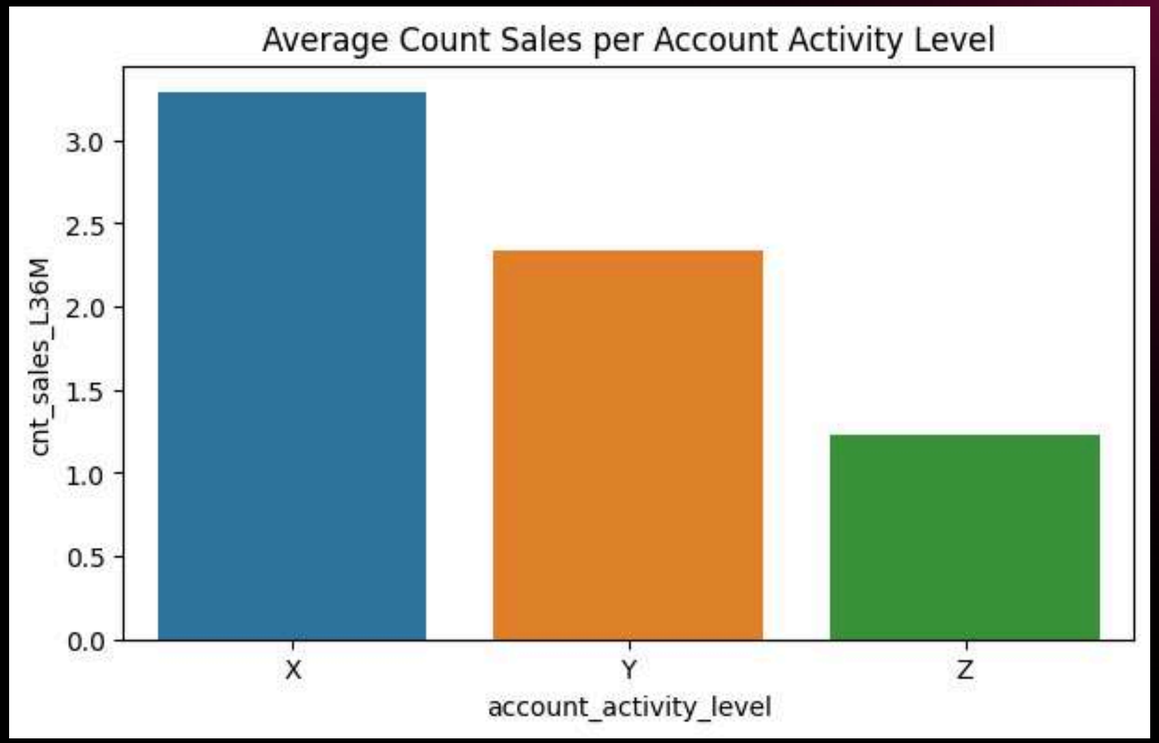


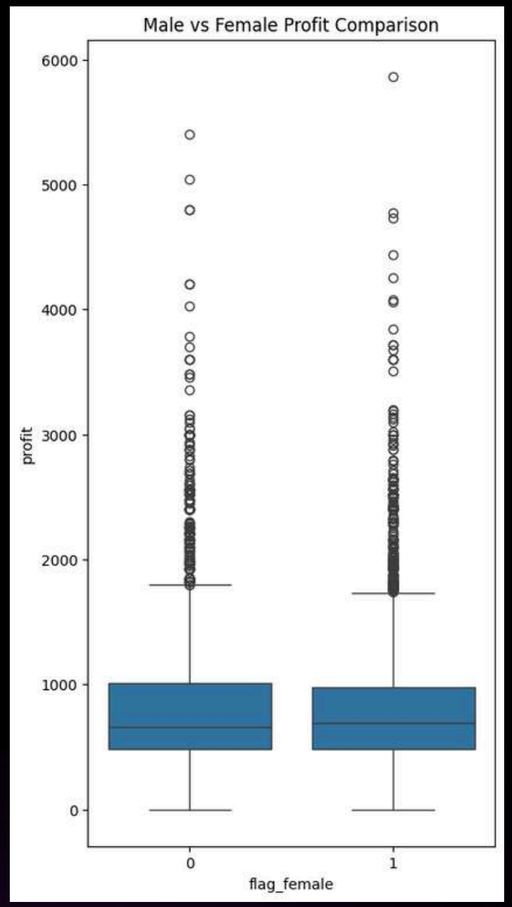
Clients with no Sales
726 clients

From the chart before, the clients with no sales are only 726 clients. It's only 5.9% from total clients. This number is quite low but it still needs further segmentation for furhter insights and actions

Account Activity

The account activity level is divided into 3 categories. They are X, Y, and Z. By aggregating the average of count sales (cnt_sales_L36M), we found that the average of count sales for X is above 3, y is 2-3 counts, and Z is 1-2 counts

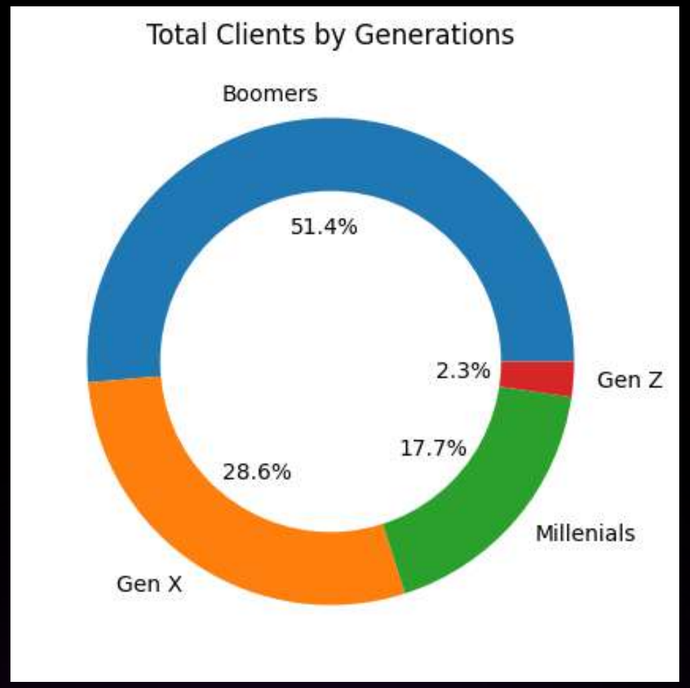


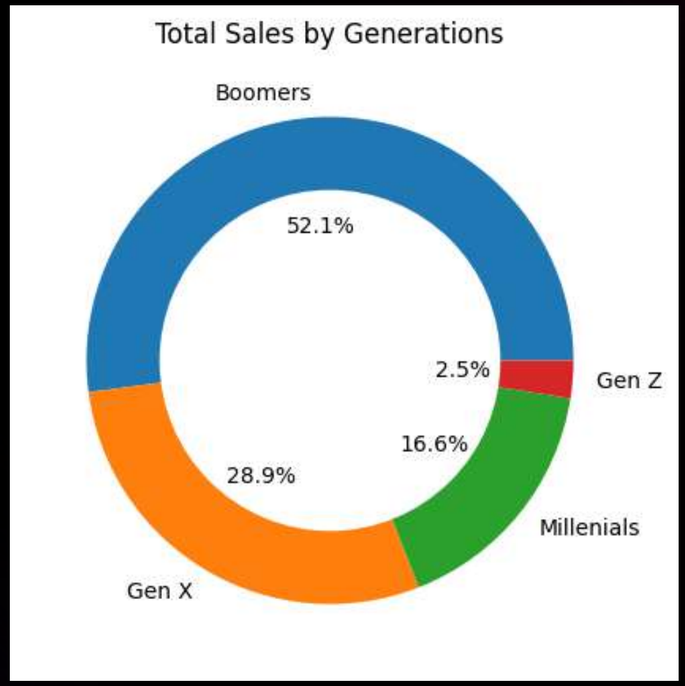


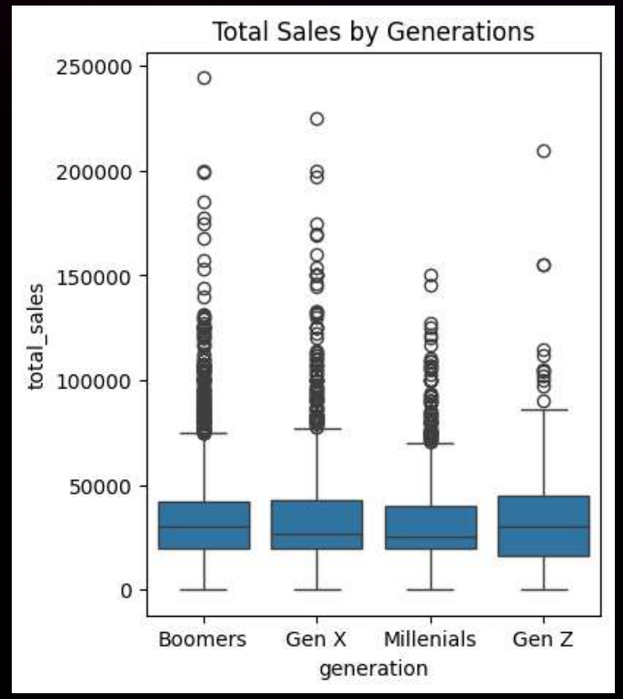
Profit Comparison

From the boxplot beside, it seems there is no significant difference of the profit generated by Male or Female. The male average profit is slightly higher with about 788 Euro compared to Female with 776 Euro.

Sales Proportions





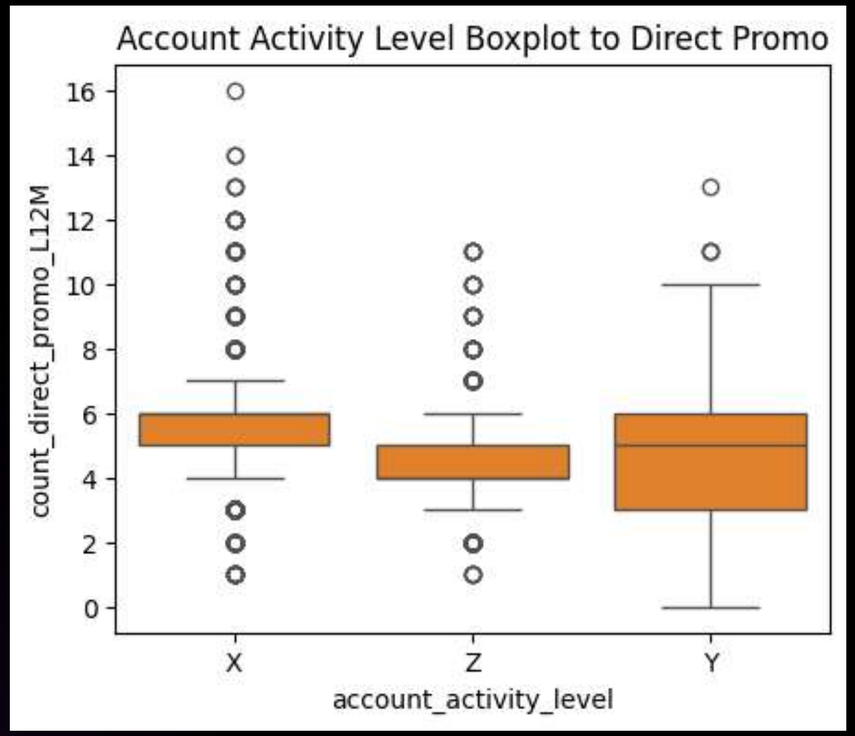


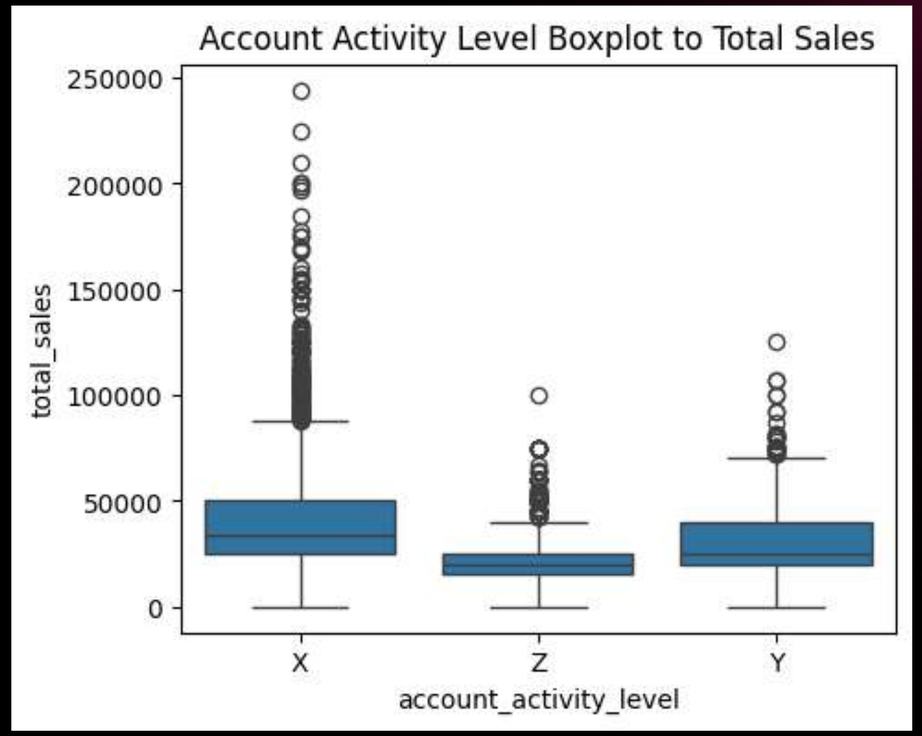
Next Slide

Sales Proportions

The chart before proves us the sales performance by each generation. The overall performance is quite good and has similar average between generations. It tells us each person in each generation performs similarly well. But, because the RevoBank customers are mostly come from Boomers Generation, so the Boomers generated much higher Total Sales compared to all generations. This made Boomers dominate the total sales by 52.1%.

Sales & Promo Relationship





Sales & Promo Relationship

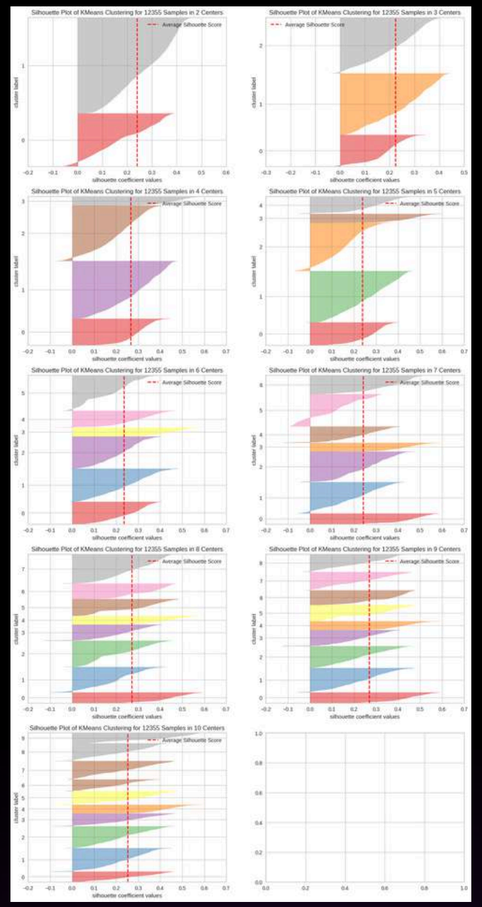
From the account activity definition and our findings, we knew that the account activity level indicates the average of clients using their credit cards. The Boxplot of Account activity level to Direct Promo indicates that the “X” level receive more direct promo while “Y” level receive more vary direct promo. As a result, In the Account activity level to total sales boxplot, “X” and “Y” level has similar average of total sales. The correlation value between Direct Promo and Total Sales is 0.55, it indicates a relatively strong correlations. If the clients receive more direct promo, the total sales more likely to increase.

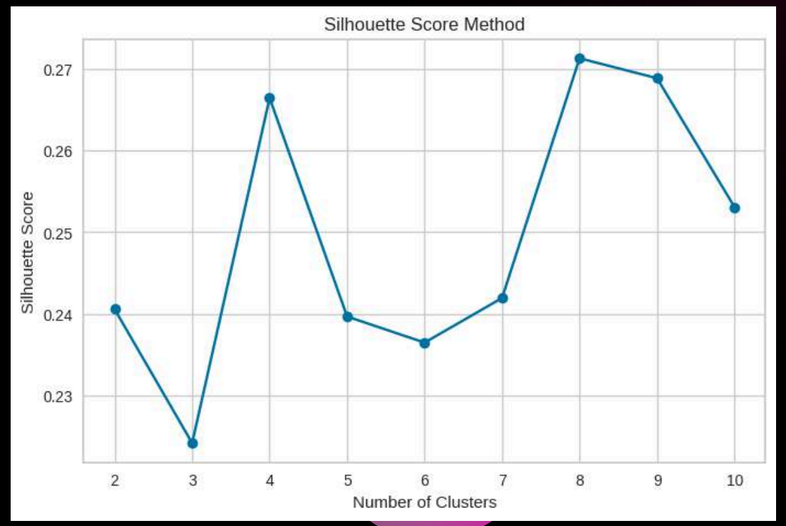
End of Milestone 2

KMeans Clustering

For this RevoBank case, KMeans clustering is more reliable. These are the reasons why Kmeans Clustering is choosen :

- RevoBank is one of banks in Europe with small market capitalization. So, more segmented offerings are more effective.
- The first point become possible because KMeans allow multiple features beyond purhcase behavior
- It provides objective segmentation based on the inherent patterns in the data.
- KMeans can reveal patterns in customer activity, value, and engagement that may not be evident from simpler segmentation methods like RFM.



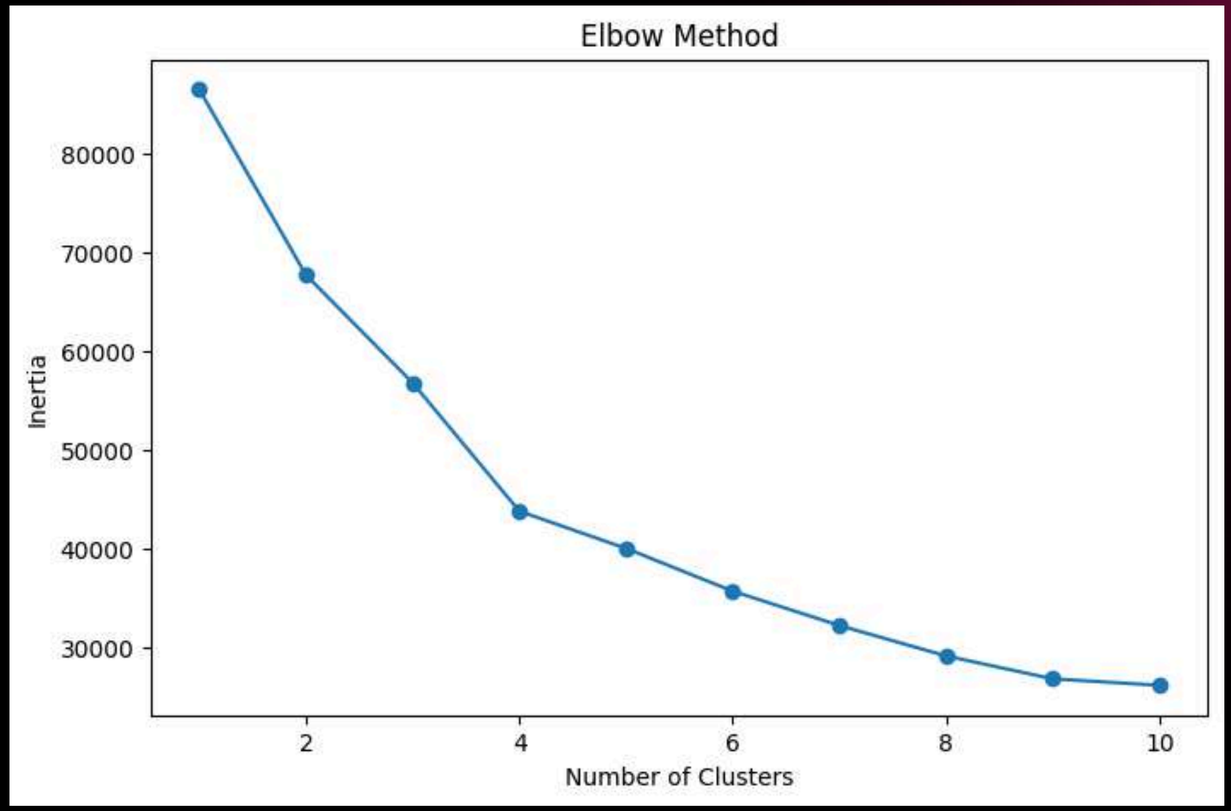


Silhouette Method

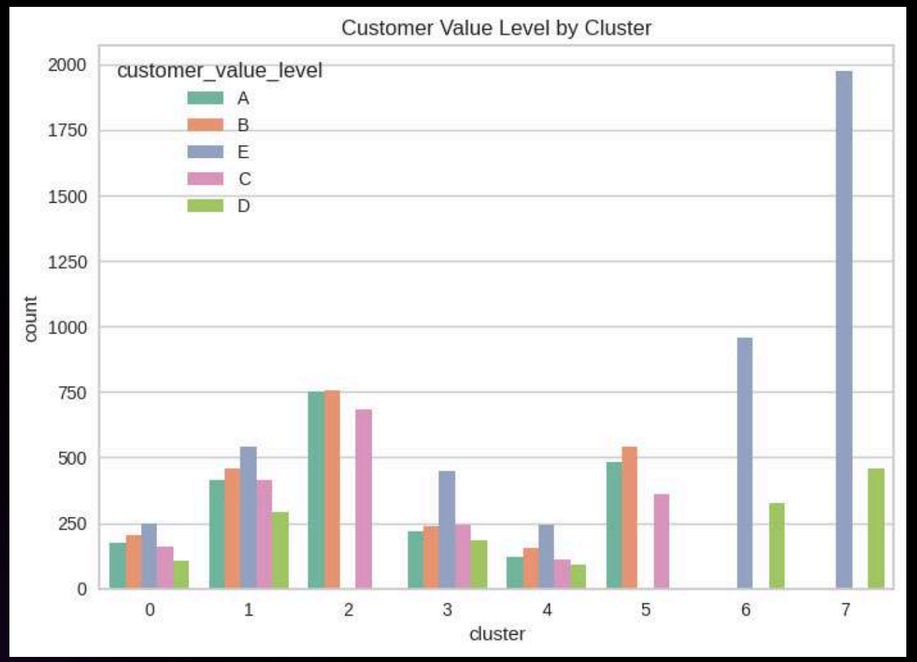
According to the Silhouette plot, number of 8 clusters show the highest score and fullfil the silhouette plot terms. The ‘saw’ graph all above the red line, minimal “minus”, and the area of the “saw” looks similiar.

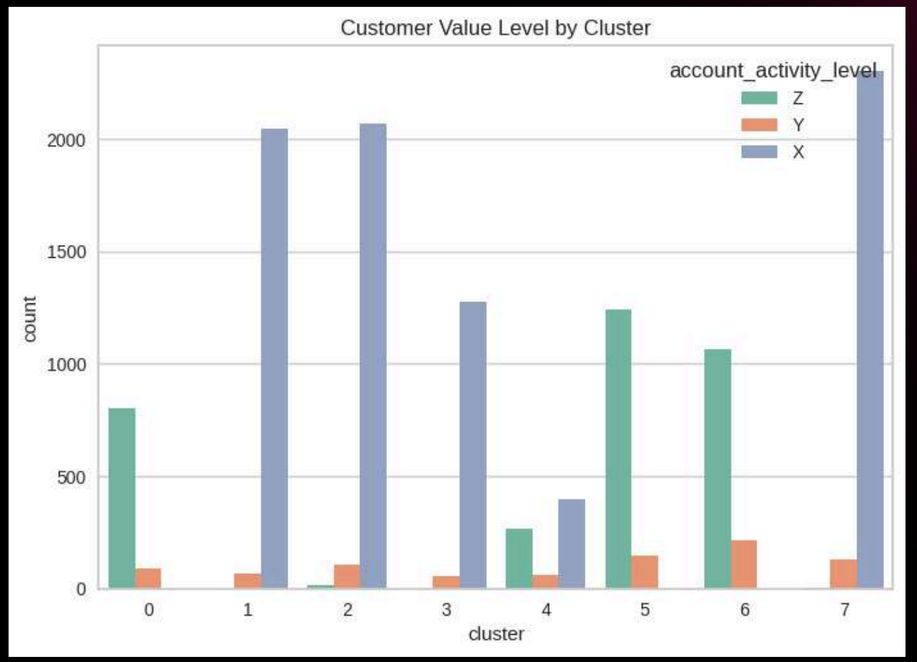
Elbow Method

Since the elbow is about 4 clusters, but with 8 clusters got much lower inertia. 8 clusters are still considerable.



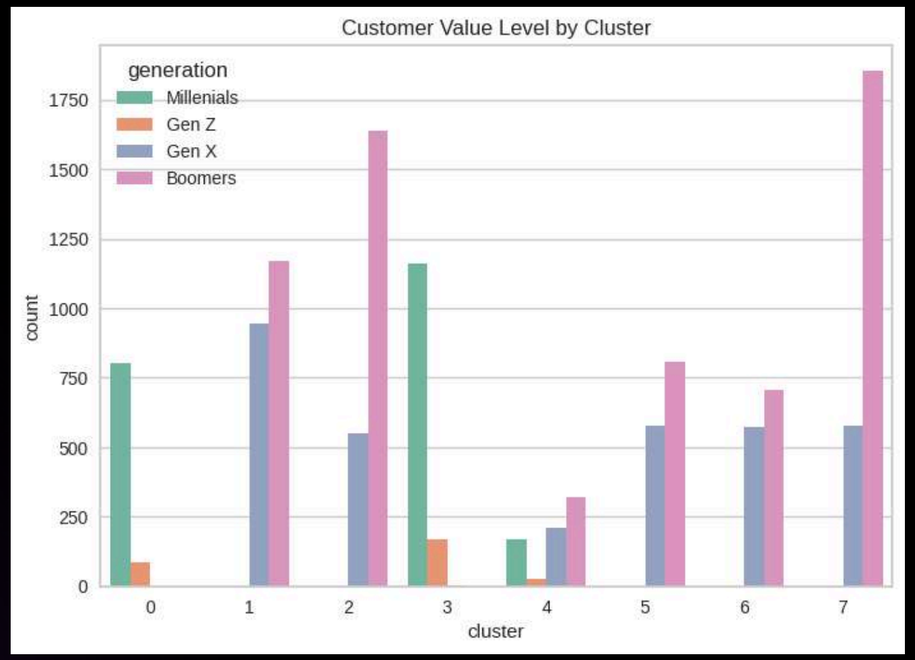
Clusters Visualization

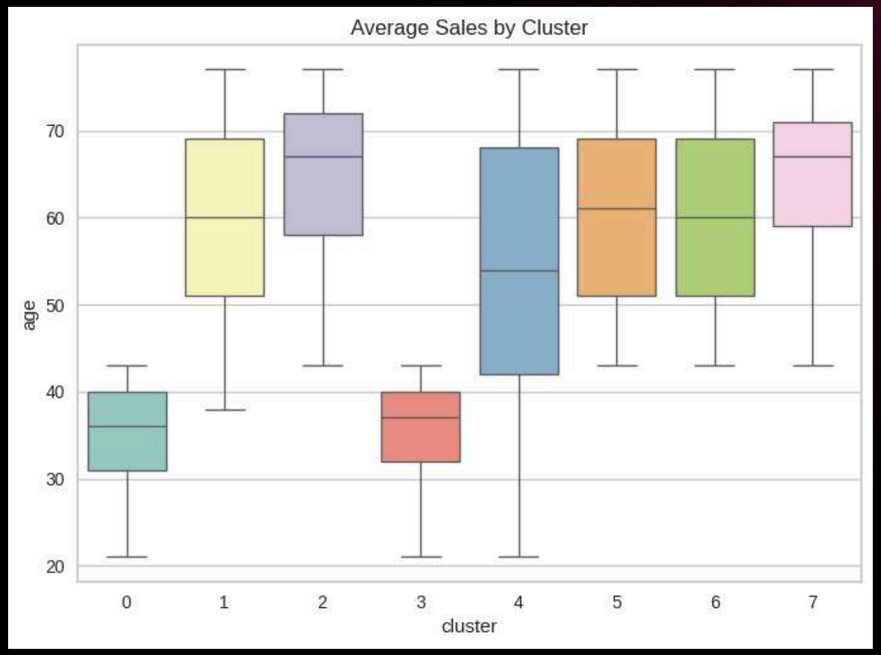




Next Slide

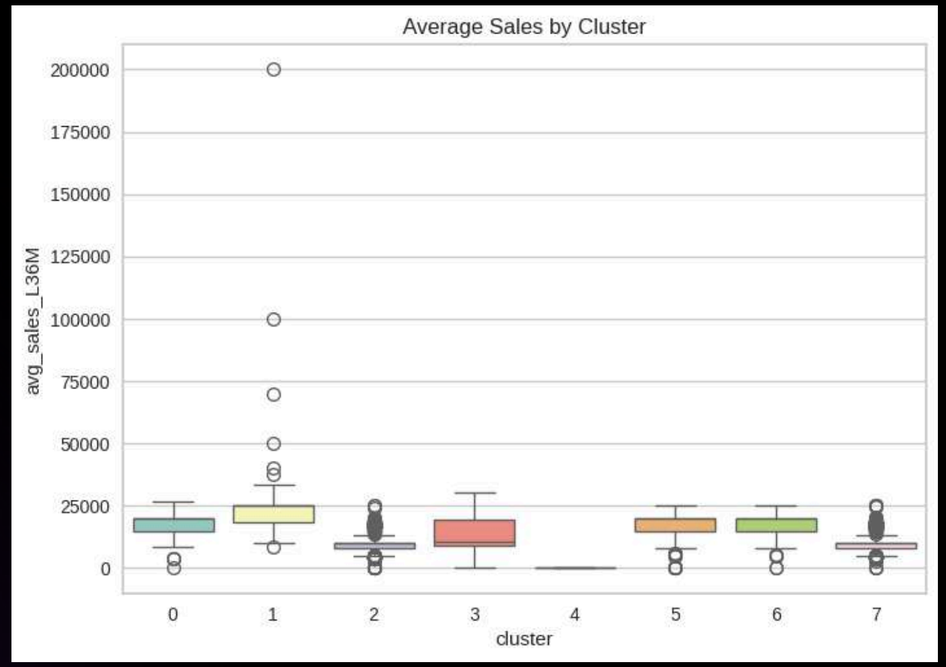
Clusters Visualization

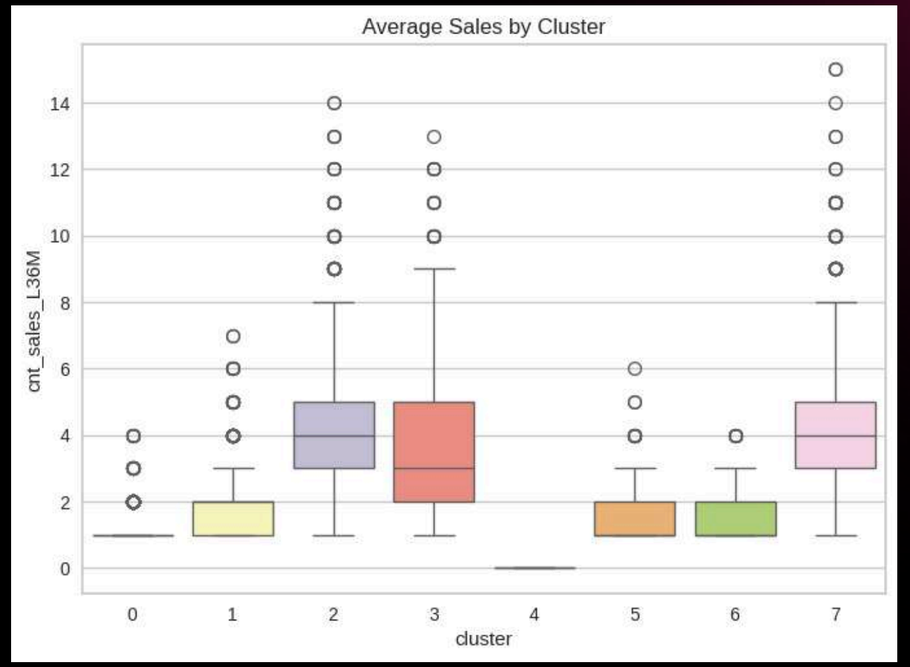




Next Slide

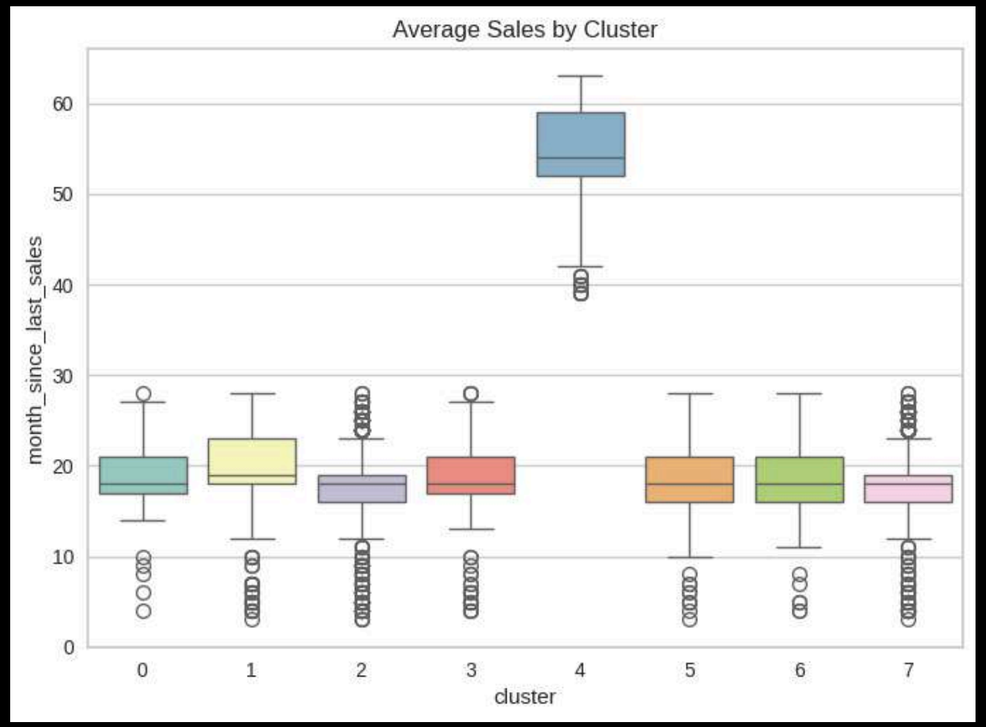
Clusters Visualization





Next Slide

Clusters Visualization



Clusters Definition

Cluster 0 : Low traffic transactions (shown from low `cnt_sales_L36M`), young clients (low `age`), mostly Milenials `generation`

Cluster 1 : Mostly labeled X `account_activity_level`, moderate transaction traffics (shown from quite low `cnt_sales_L36M`), highest `avg_sales_L36M`, quite old with high `age`, dominated by gen x and boomers `generation`

Cluster 2 : quite high `cnt_sales_L36M`, dominated by boomers `generation`, low `avg_sales_L36M`

Cluster 3 : quite high `avg_sales_L36M`, young `age`, dominated by Milenials `generation`

Cluster 4 : very high `month_since_last_sales`, very low `cnt_sales_L36M`, very low `avg_sales_L36M`, from vary `age`

Cluster 5 : Moderate transaction traffices (`cnt_sales_L36M`), quite high `avg_sales_L36M`, old `age` with mostly Z `account_activity_level`

Cluster 6 : Moderate transaction traffices (`cnt_sales_L36M`), quite high `avg_sales_L36M`, Mostly E `customer_value_level`

Cluster 7 : High `cnt_sales_L36M`, moderate `avg_sales_L36M`, mostly boomers `generation` with X `account_activity_level` and E `customer_value_level`

Clusters Definition

1st Cluster :

- The milenials with low traffic transactions but generate quite high average sales

2nd Cluster :

- Gen X & Boomers with high activity and generate highest average sales

3rd Cluster :

- Vary clients value boomers with high activity but generate low average sales

4th Cluster :

- “X” labeled (high activity) Milenials with quite high average sales

5th Cluster :

- They are that “lost”, long time no transaction (inactive for a long time)

6th Cluster :

- The old age clients with moderate traffic that generate quite high average sales

7th Cluster :

- “E” Customer Value Level with moderate traffic that generate quite high average sales

8th Cluster :

- “E” Value Boomers with quite high activity but generate moderate average sales

Clusters Insights

Since the RevoBank wants to evaluate their 3 years performance, the sales/revenue is the most important parameter. According to the cluster, 3rd and 8th cluster may become highlights. They are Boomers (probably have more money than other gen) with moderate to low average sales. Their activity is not that bad, even the most recent. Some promotions/offerings needs to be evaluated or any other factors that causing their low average sales.

The 5th Cluster needs to be investigated why they generated very low average sales since they don't come from a specific generation/ages. They are vary.

The rest, need to be maintained according to their specific criterion, such as 1st cluster, Milenials with low activity but generate high average sales must be treated differently with 4th cluster, Milenials with quite high average sales but high activity level, etc.

Thank You!