Airline Customer Segmentation

Presented by: Julius Hermanto

Workflow

- Objective
- Data Understanding
- Descriptive Statistics
- Exploratory Data Analysis
- Data Preparation
- Feature Selection & Engineering
- Outlier Analysis
- Modeling
- Characteristic Analysis
- Recommendations
- References

Objectives

- Identify the customer segmentation in an aviation company.
- Analyze the characteristics of different clusters.
- Suggest business strategies to customer categories of different values.

Data Understanding

Data Understanding

- The dataset has 62988 rows and 23 columns:
  - 15 columns contain numerical data
  - 8 columns contain non numeric data

Code	Description
MEMBER_NO	ID Member
FFP_DATE	Frequent Flyer Program Join Date
FFP_TIER	Frequent Flyer Program Tier
LOAD_TIME	Observed time
FLIGHT_COUNT	Customer flight count
BP_SUM	Flight Plan
SUM_YR_1	Fare Revenue
SUM_YR_2	Votes Prices
SEG_KM_SUM	Total flight kilometers in observation window
LAST_FLIGHT_DATE	Last Flight Date
LAST_TO_END	The last flight time to the end of the observation window
AVG_INTERVAL	Average flight interval
MAX_INTERVAL	Maximum flight interval
EXCHANGE_COUNT	Exchange count
avg_discount	Average discount rate
Points_Sum	The number of points earned by the customer
Point_NotFlight	Point not used by customers

Descriptive Statistics

Descriptive Statistics

Based on non-numerical data, we can see that most customers are **Male.** The top city of origin is **Guangzhou**, in the province of **Guangdong**. The top country is **China**, so we can assume that we are analyzing a Chinese airline company.

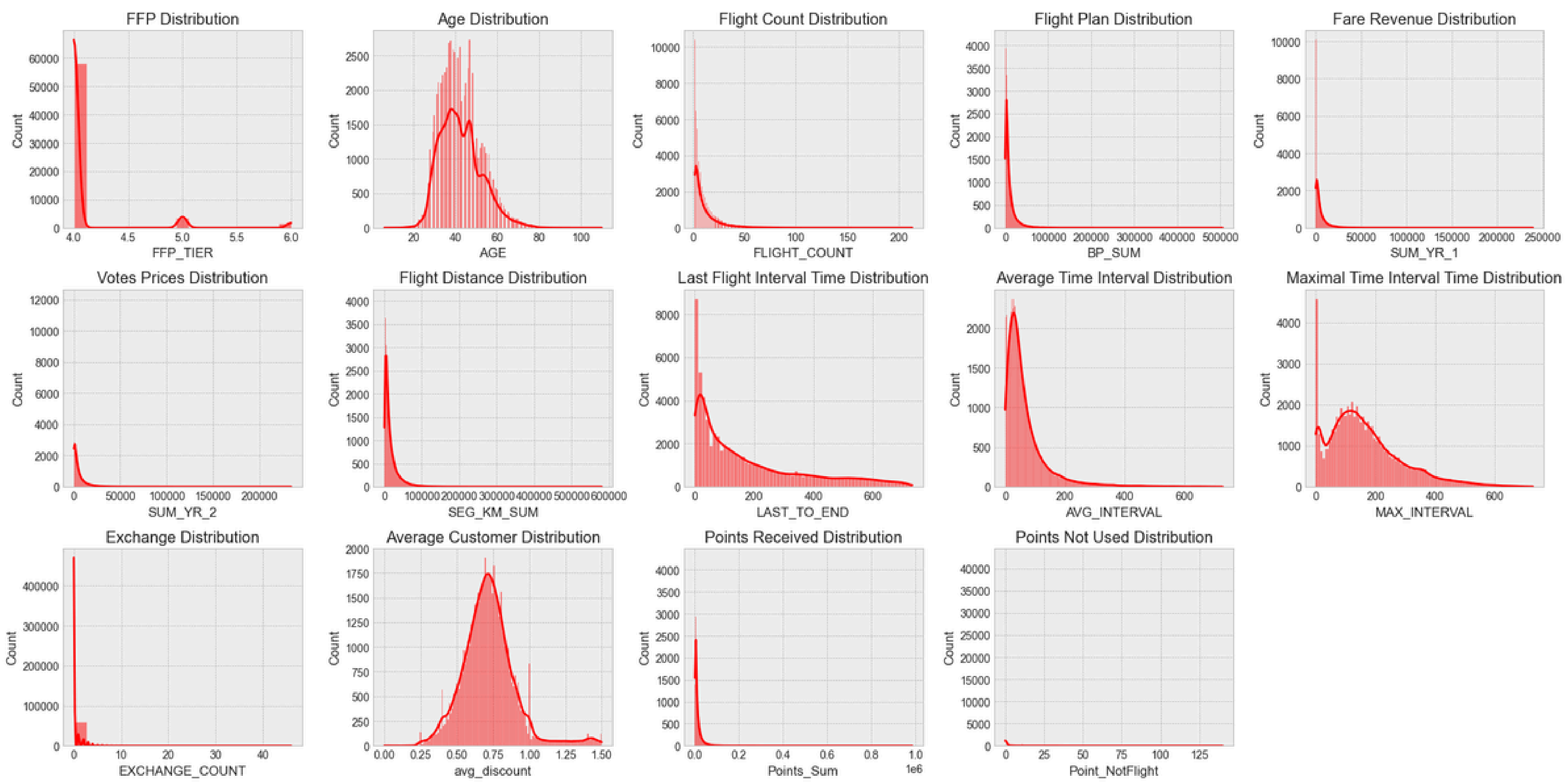
Based on numerical data, here are as follows:

1. Average age: **42 years old**
2. Average flight frequency: **11 flights**
3. Average cumulative distance: **17,123 km**
4. Average time interval: **67 minutes**
5. Average maximal time interval: **166 minutes**
6. Average discount rate: **70%**
7. Average points accumulated: **12,545 points**

Exploratory Data Analysis

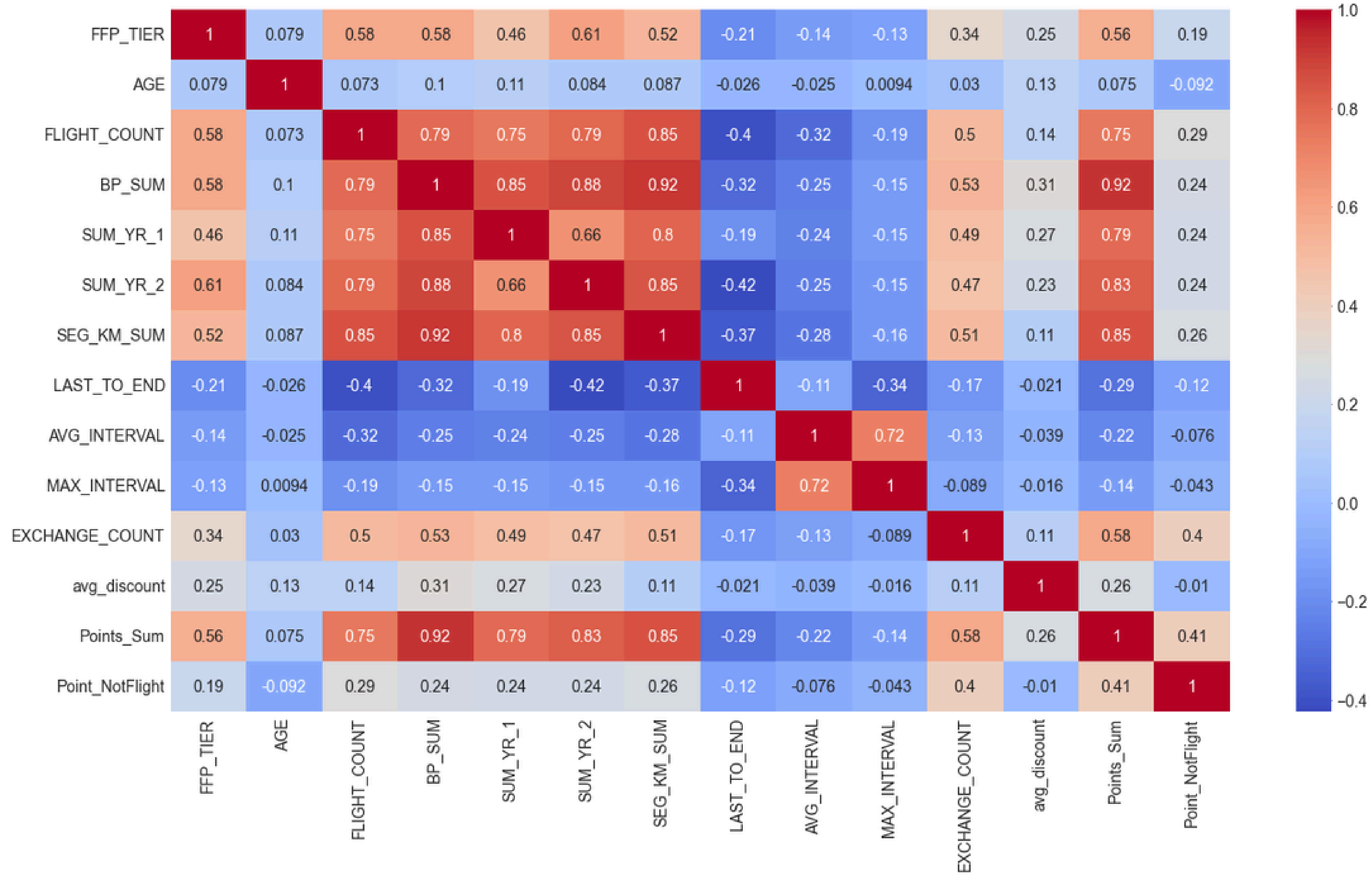
Univariate Analysis

- **FPP_TIER** is a categorical value since it has discrete values (4 to 6).
- We did not analyze **MEMBER_NO** since it has too many values and it may not give insights respectedly.
- Most data shown tend to be positively skewed with some data having extreme value differences.
- Features chosen to be used for modeling must be normalized later on to even out distribution.



Multivariate Analysis

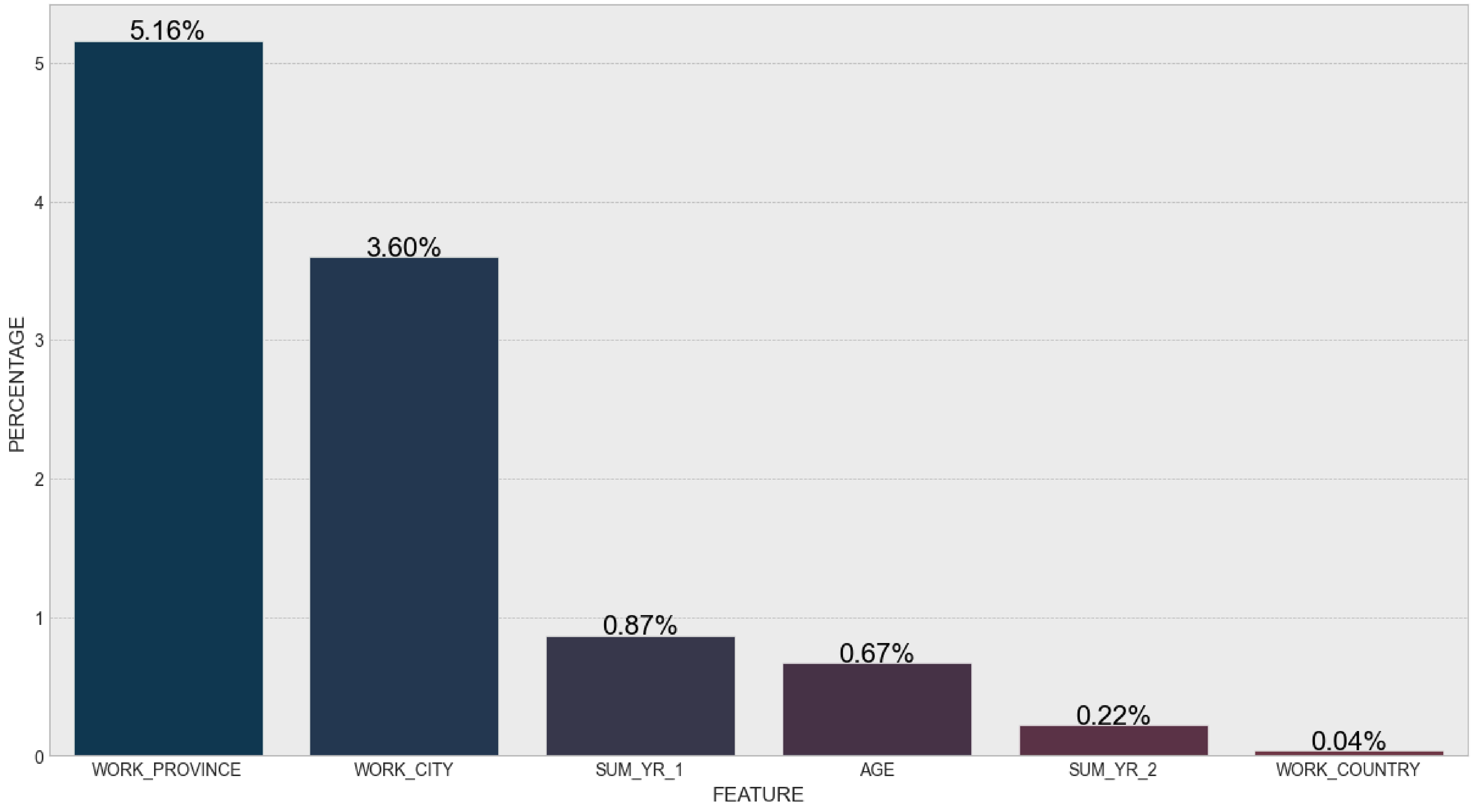
- **FLIGHT_COUNT**, **BP_SUM**, **SUM_YR_1**, **SUM_YR_2**, **REG_KM_SUM**, and Point_Sum have high correlation between each other.
- These features may be put into consideration for modelling later on.



Data Preparation

Missing Values

- We can see that the missing values totaled **less than 10%** of all the data. We drop missing values since it may affect our model.
- There also has 1 duplicate values, we also drop the duplicate value.



Feature Selection

Feature Selection

Here we use an extended version of **RFM** called the **LRFMC model**. The selection is as follows:

- **Length: LOAD_TIME - FPP_DATE**
  - The number of days from the passenger's first day of registration to the time of observation. A higher number indicates a longer membership period.
- **Recency: LAST_TO_END**
  - The time interval between a passenger's most recent consumption and the observation window. The lower the number, the more recent the flight.
- **Frequency: FLIGHT_COUNT**
  - The frequency of the passenger's consumption over a predetermined duration. Greater numbers indicate more frequent flights.
- **Monetary: SEG_KM_SUM**
  - The mean expenditure over a given time frame. A higher number indicates that they made larger purchases.
- **Discount (C): avg_discount**
  - The typical space-discount percentage for travelers within a specific time frame A higher figure indicates greater usage of discounts.

Outlier Handling

Feature Engineering

Here we create a new feature to incorporate to **Length** variable of the model, where **LOAD_TIME - FFP_DATE**

```
dff['FFP_DATE']= pd.to_datetime(dff['FFP_DATE'])
dff['LOAD_TIME']= pd.to_datetime(dff['LOAD_TIME'])
```

✓ 0.1s

```
data['L'] = data['LOAD_TIME'] - data['FFP_DATE']
data['L'] = (data['L'].astype(str).str.split().str[0]).astype(int)
```

✓ 0.5s

Duplicate Values

After engineering features, we check for duplicates and drop them if available.

```
data.duplicated().sum()
```

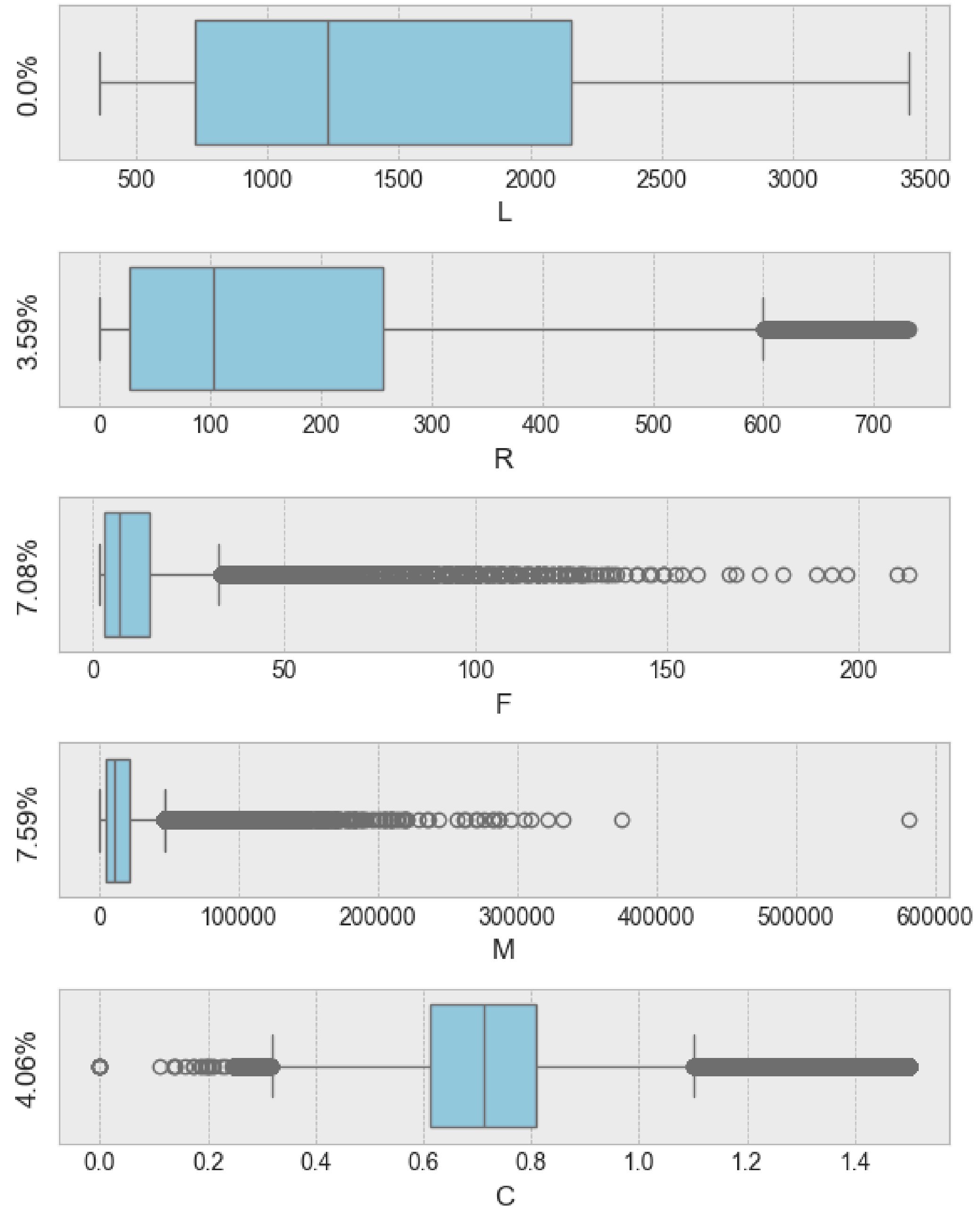
✓ 0.0s

```
81
```

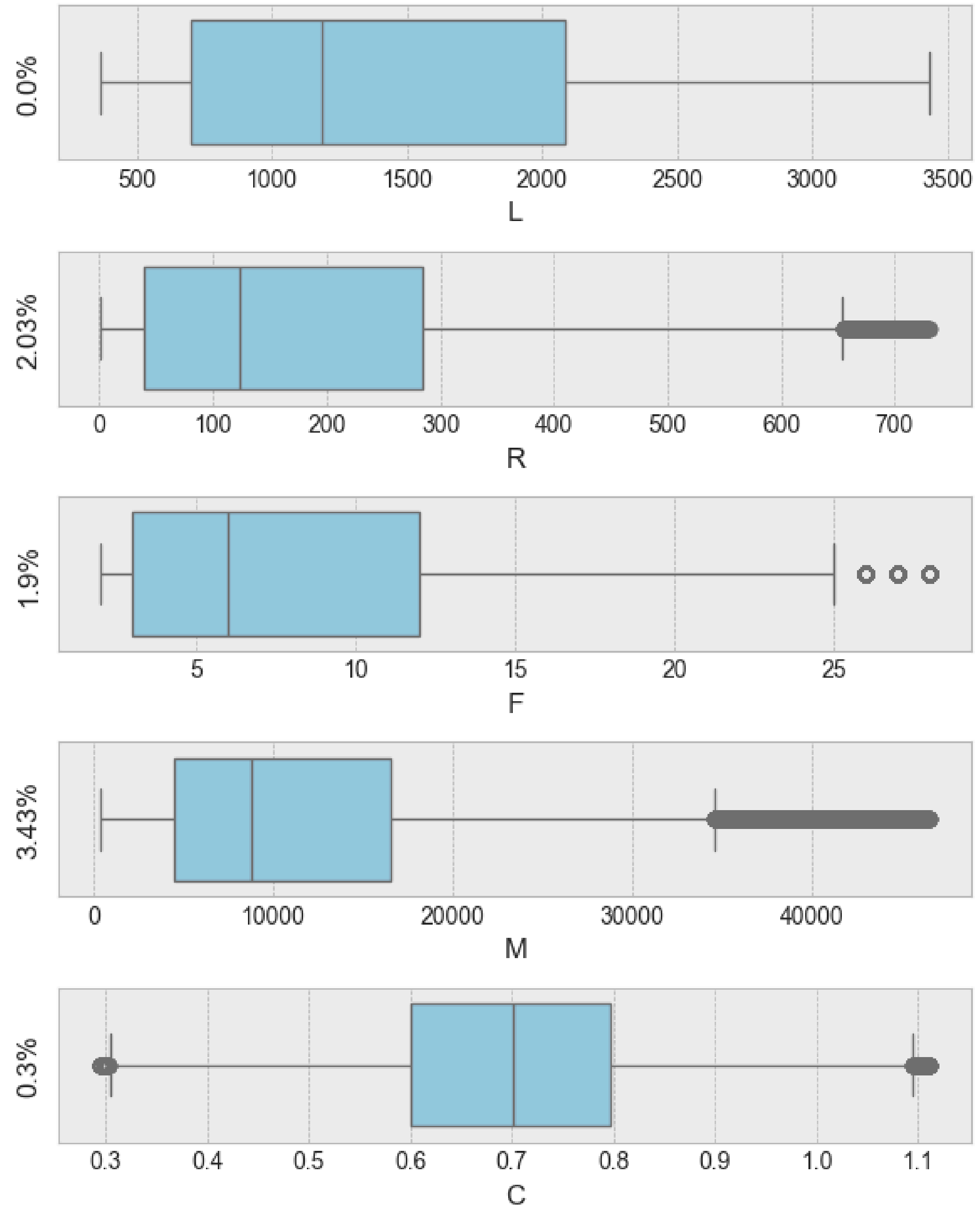
```
data = data.drop_duplicates()
```

✓ 0.0s

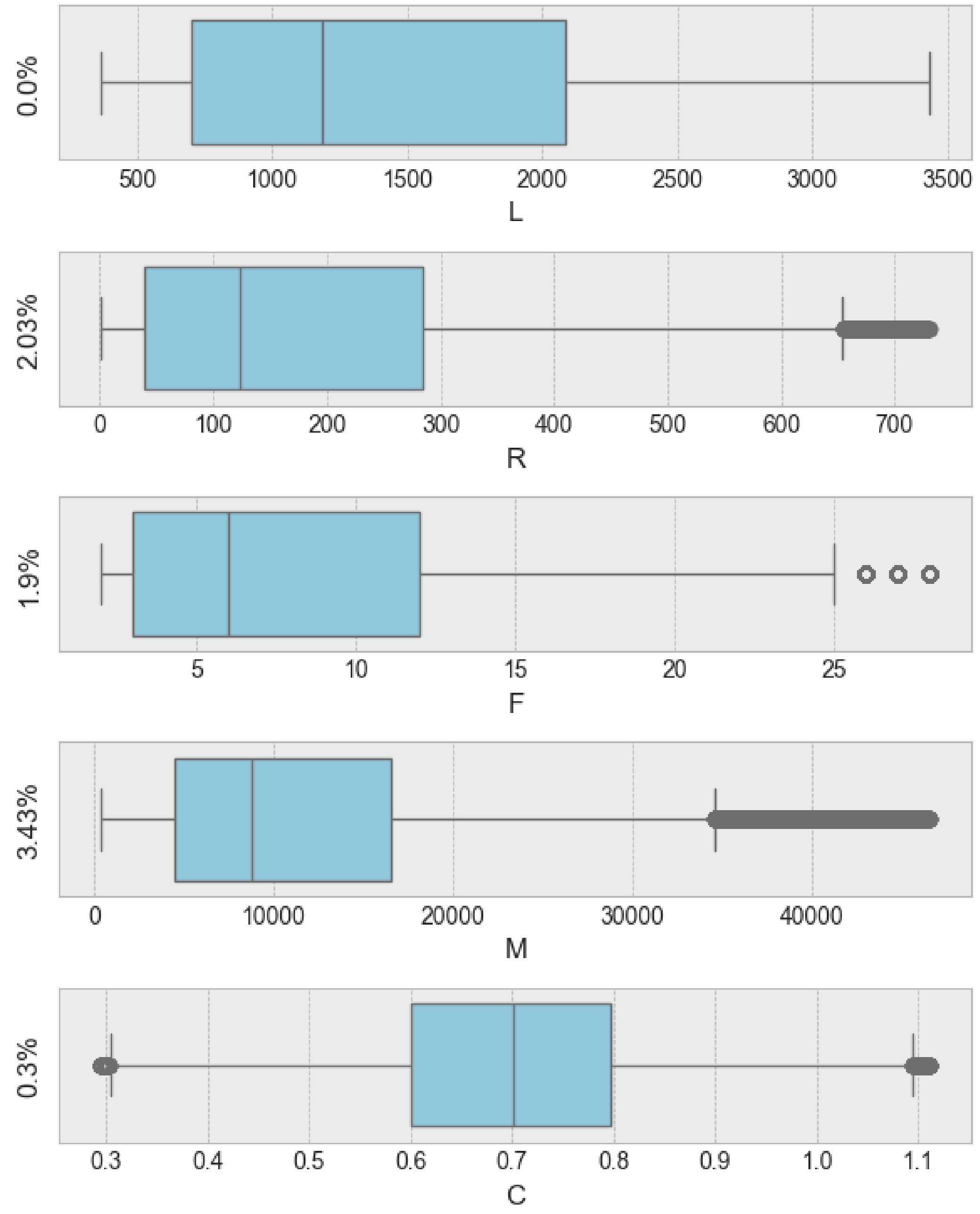
Outlier Handling



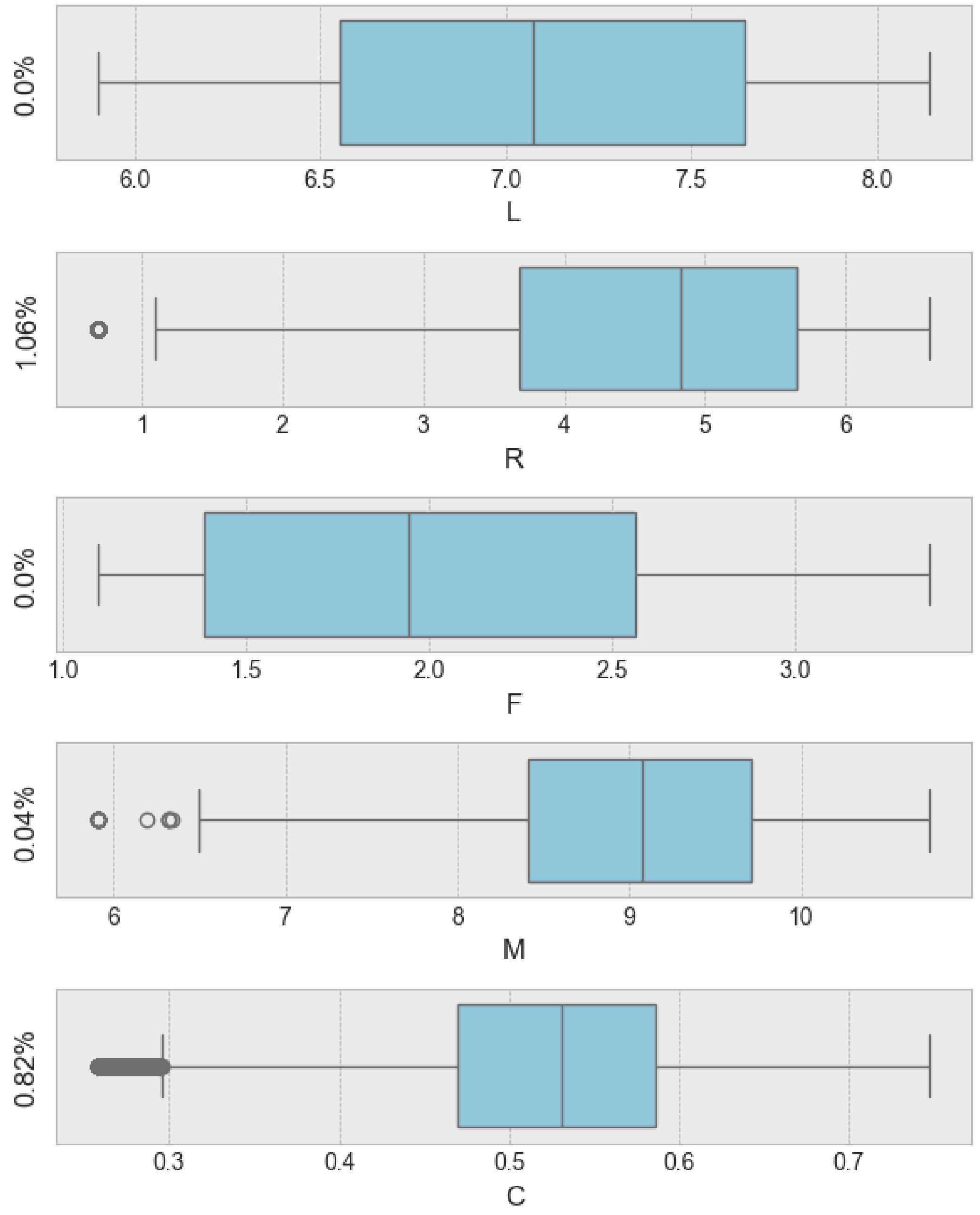
- On the **left**, we can see **F, M,** and **C** has tight distribution and **F** & **M** has extreme values.
- We drop the outliers, resulting to the outlier distribution on the **right**



Scaling



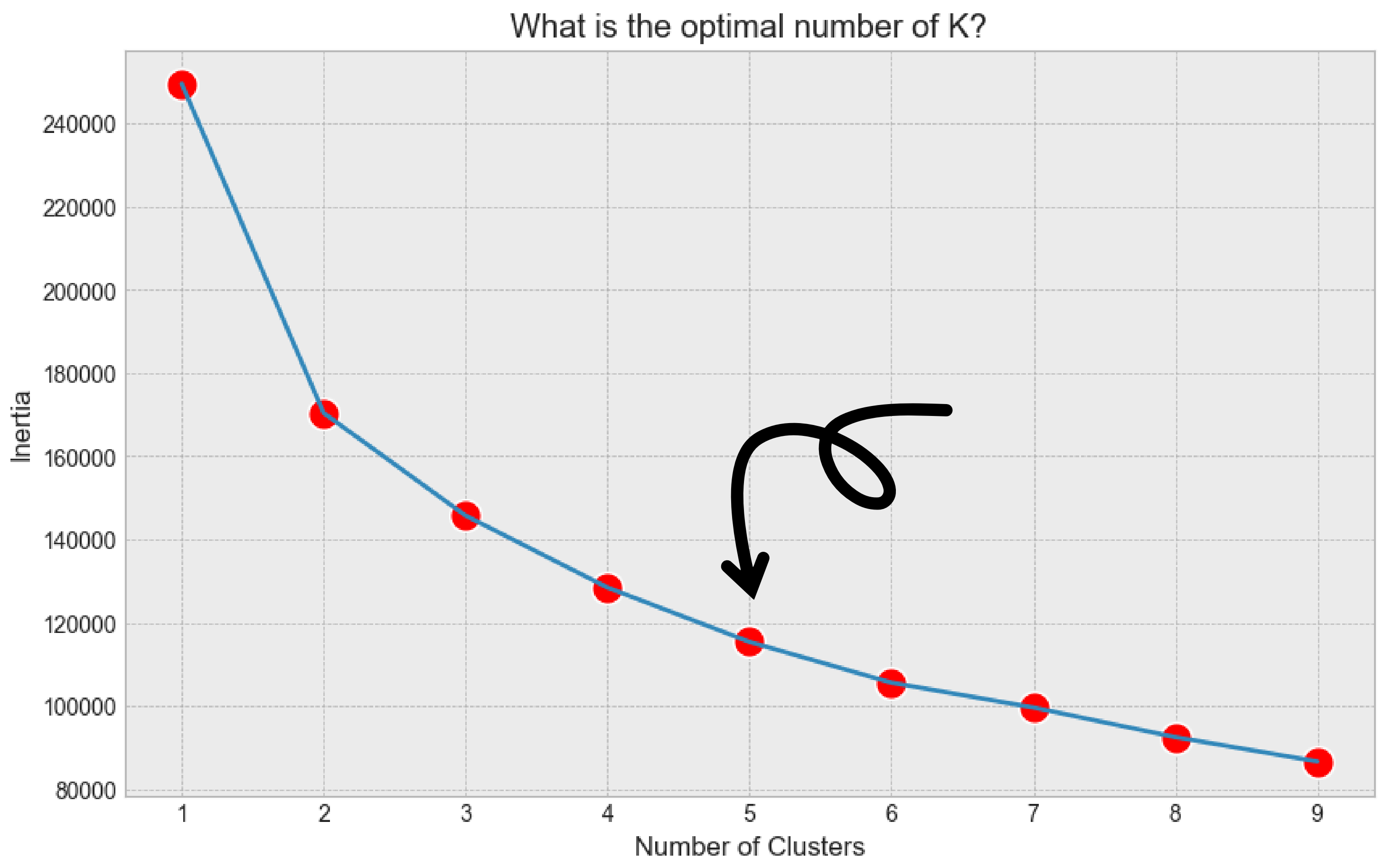
- On this modeling we use **StandardScaler** to normalize distribution reducing skewness (right).

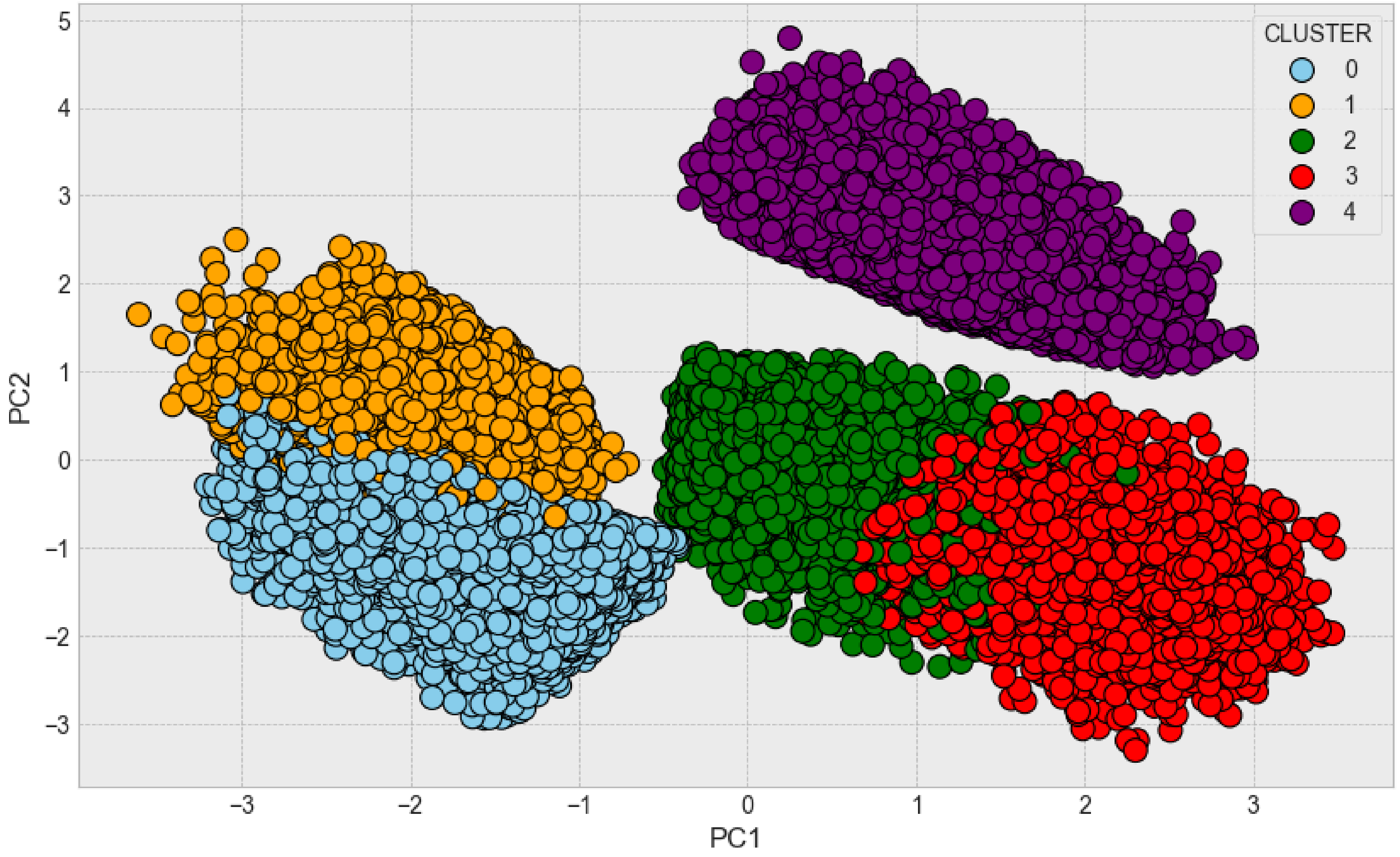


Modeling

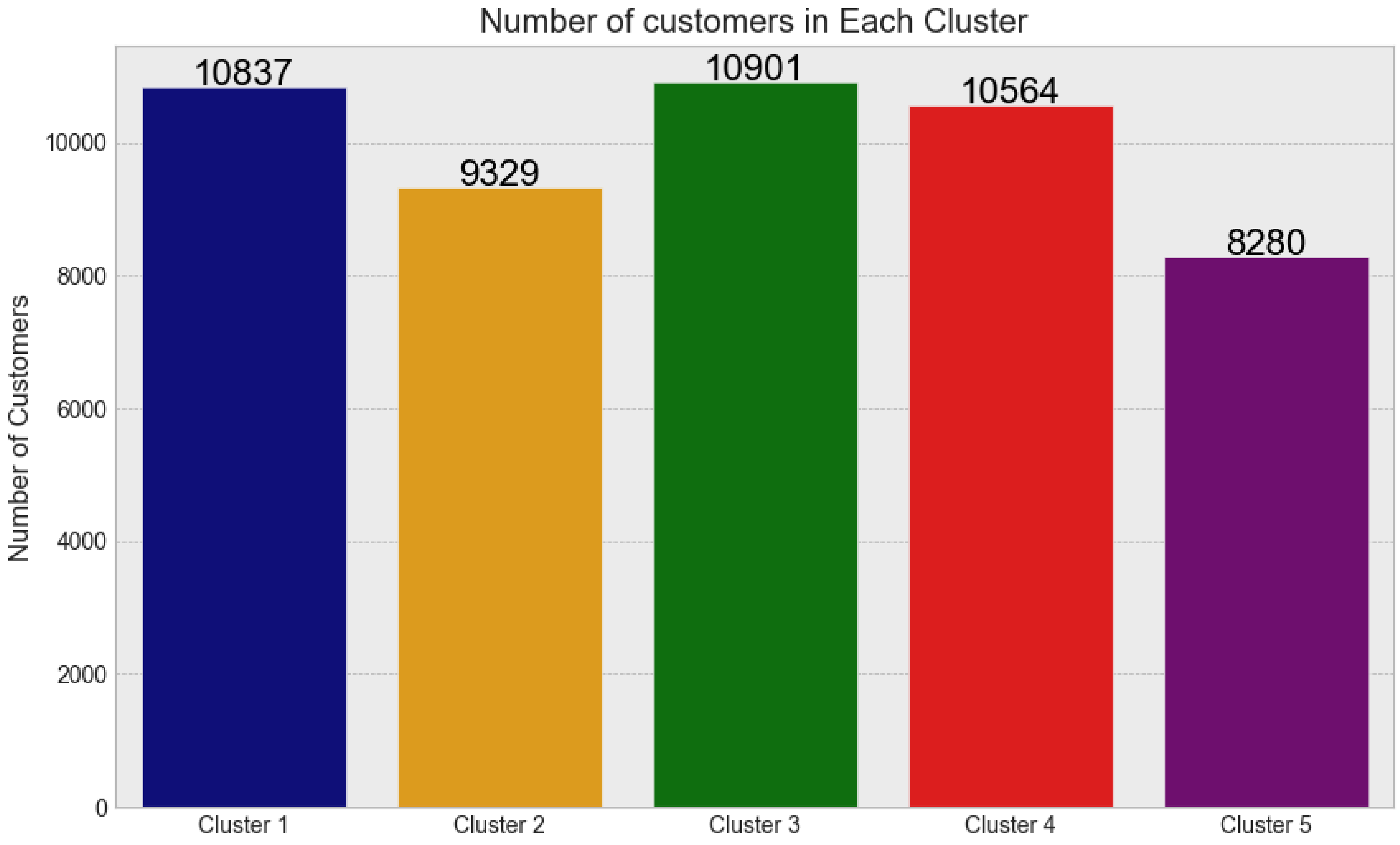
Clustering

- Optimal number of cluter based on elbow plot method is **5**.
- Visualizing the plot based on number of cluster 5 can be seen it is evenly clustered.





Clustering with k=5

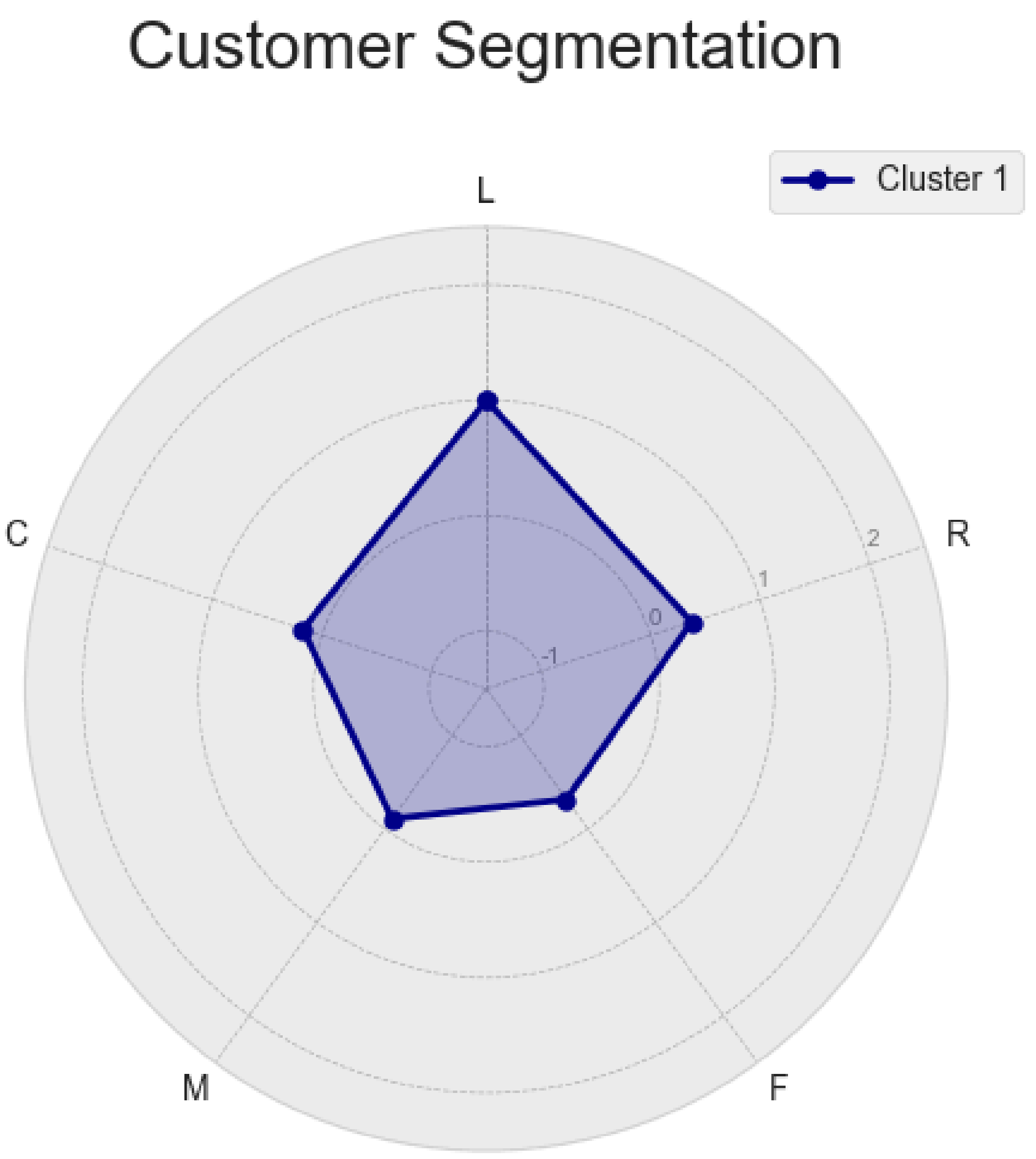


The clusters are as followed:

- Cluster 1 - **Promising Customers**
- Cluster 2 - **Price Sensitive Customers**
- Cluster 3 - **Customers that Need Attention**
- Cluster 4 - **Champion Customers**
- Cluster 5 - **Hibernating Customers**

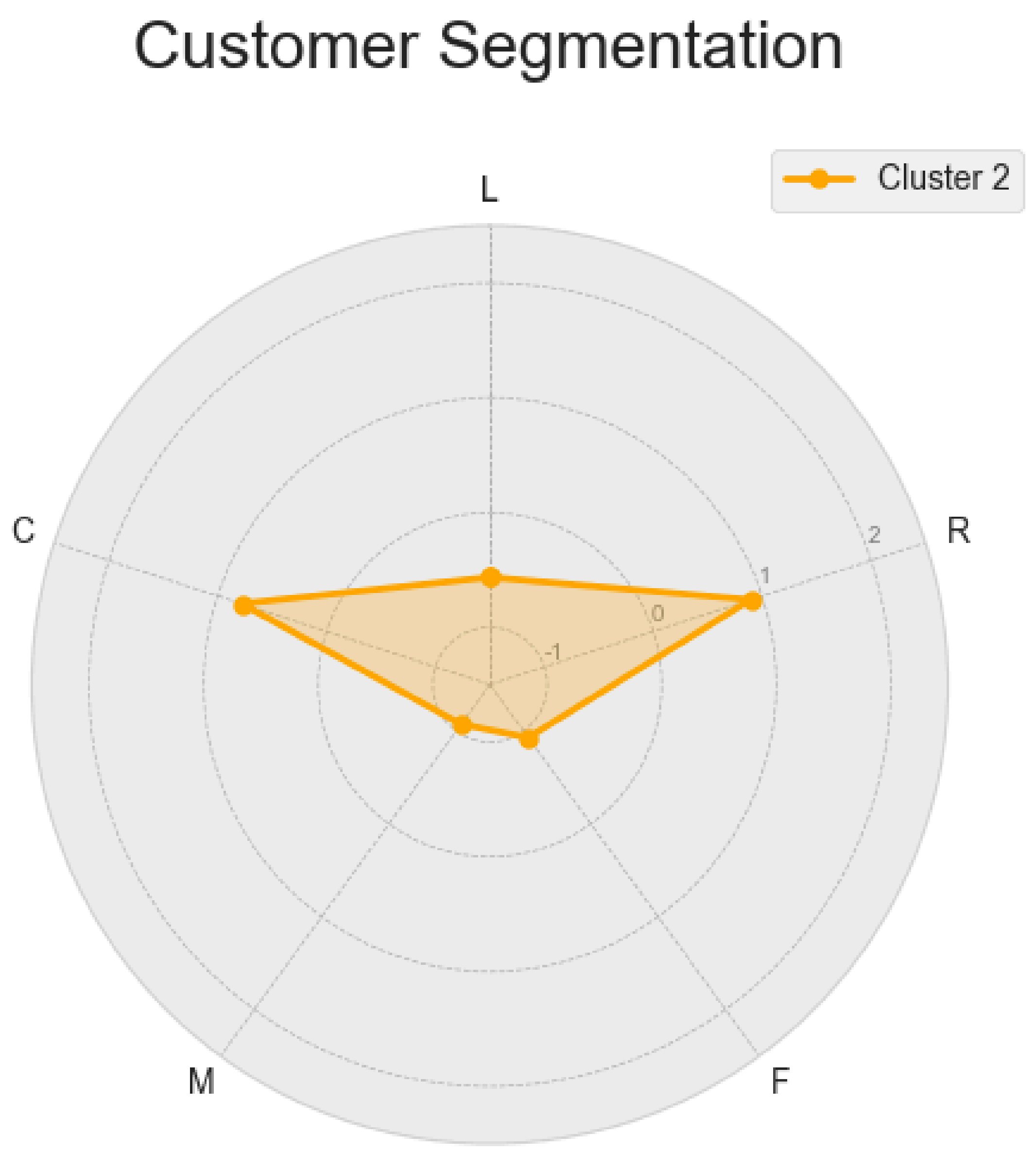
Characteristic Analysis

(1) Promising Customers



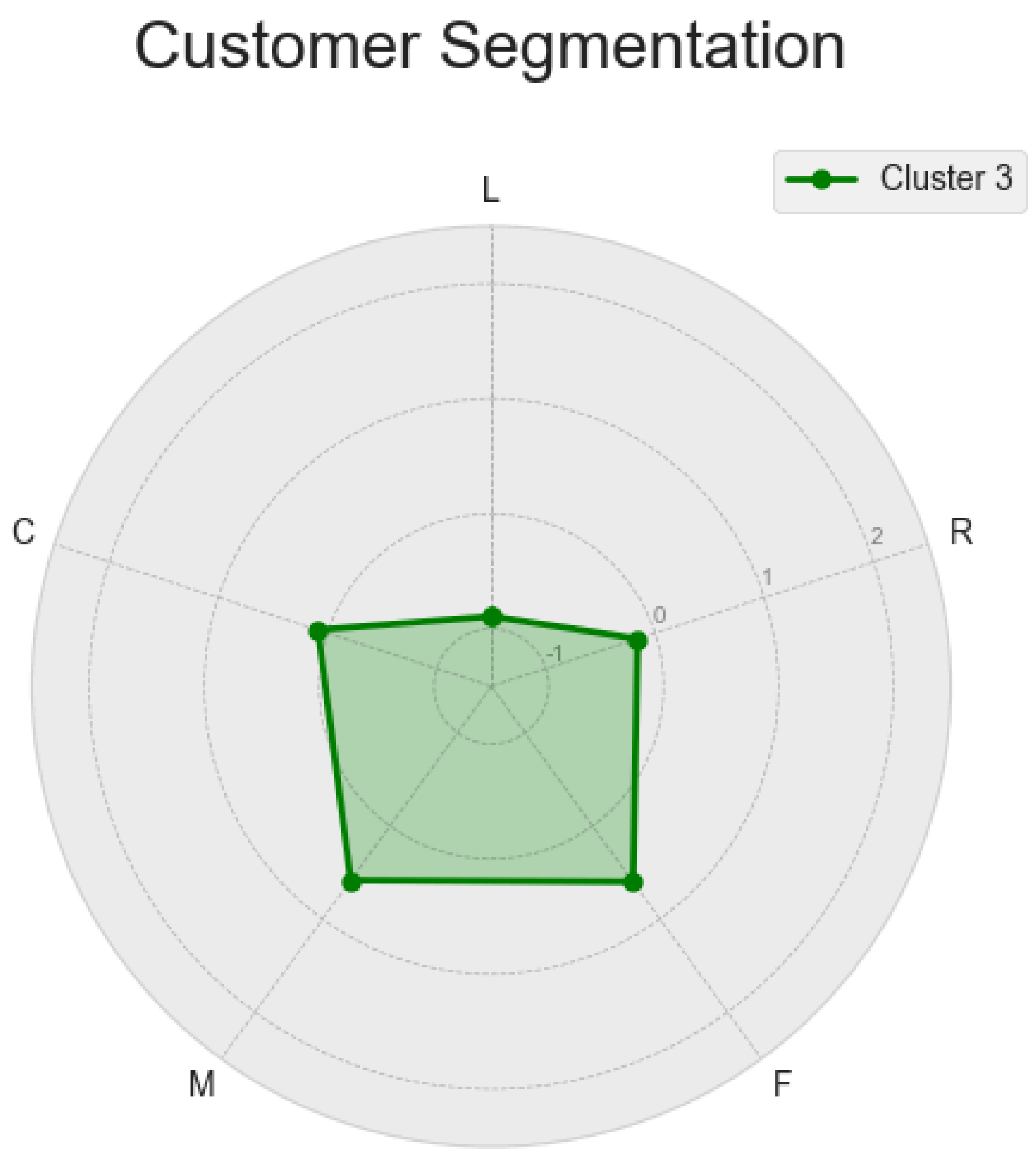
- Has average L, R, and C values
- Has not so fairly low F and M values
- These customers are relatively new but show the potential for becoming loyal
- They have recently started engaging with the airline but have not yet reached the highest level of activity or spending

(2) Price Sensitive Customers



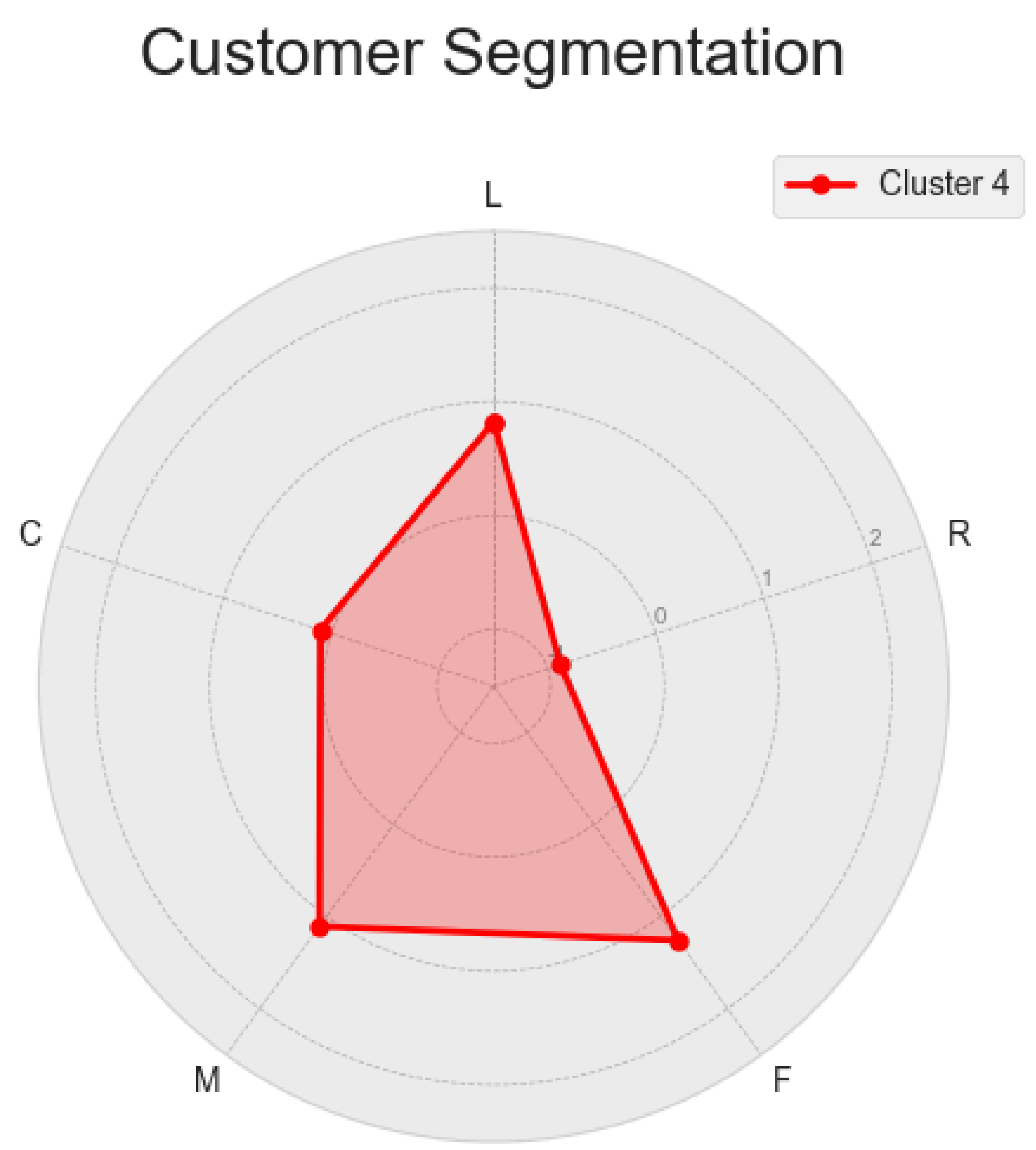
- Has high R, and C values while low on every other aspect
- These customers are highly motivated by price and may switch airlines for better deals
- They are likely to book flights based on cost rather than loyalty

(3) Customers Needing Attention



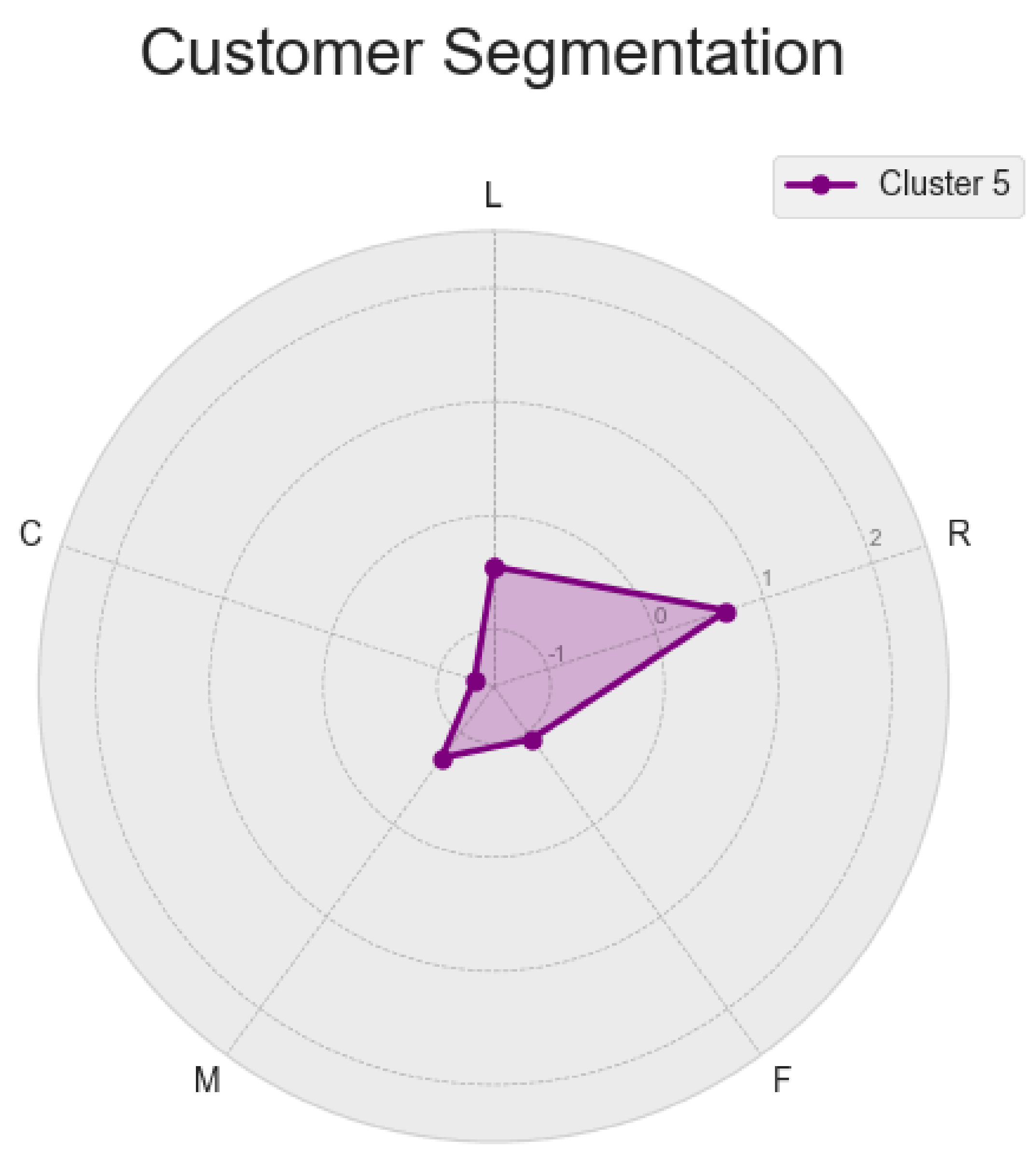
- Has fairly average F & M values
- Has average C & R values
- Has low L value
- These customers may have previously been active but have reduced their engagement
- They might be dissatisfied or have found alternatives

(4) Champion Customers



- Extremely high L, F & M value
- Has average C value
- Has low R value
- These are the airline's most loyal and valuable customers
- They frequently book flights and contribute significantly to revenue

(5) Hibernating Customers



- Has average L, R & M values
- Has low F and C values
- These customers have not engaged with the airline for a long time
- They are inactive but have the potential to return

Business Recommendations

Business Recommendations

1. Promising Customers

- Provide special welcome discounts or miles to encourage repeat purchases.
- Introduce them to the benefits of the airline's loyalty program.
- Send personalized messages highlighting the benefits and services of the airline.

2. Price Sensitive Customers

- Regularly offer discounts, deals, and special promotions to keep them engaged.
- Provide options like fare classes that allow for cost-saving.
- Highlight cost-saving benefits like free baggage, meals, or flexible change policies.
- Create loyalty program tiers that cater to budget-conscious travelers.

3. Customers that Need Attention

- Send targeted campaigns to re-engage them, such as special offers or reminders of upcoming trips.
- Offer incentives to try new or improved services.

Business Recommendations (2)

4. Champion Customers

- Offer exclusive perks such as priority boarding, lounge access, and upgrades.
- Provide personalized travel experiences and dedicated account managers.
- Recognize their loyalty through special events, gifts, or acknowledgment in communication.
- Encourage them to refer others by offering referral bonuses.

5. Hibernating Customers

- Launch win-back campaigns with strong incentives to reactivate their engagement.
- Send periodic check-ins or updates about new routes, services, or offers.
- Send targeted seasonal offers that align with common travel times or holidays.

Thank You!

References

- LRFMC Model Paper