## Data

1. Information on sales in 4 different stores are given – Andheri (Mumbai), Dwarka (Delhi), Kachrapara (WB) and Moulali (WB). There are 5,80,781 datapoints (about 5.8 Lakhs).

2. For each transaction, the corresponding customer ID, ornament type, ornament category code, and several other feature values are provided. There is a total of 29 feature vectors.

3. For amount of sales against each transaction, we use the feature LINEAMOUNT.

4. We extract those points for which LINEAMOUNT > 0. (For some data points, this features takes negative values in the data). For clustering, we take LINEAMOUNT >= 0, to account for recency and frequency of customers.

5. The extracted data has 5,28,585 datapoints.

# GRAPHICAL REPRESENTATION

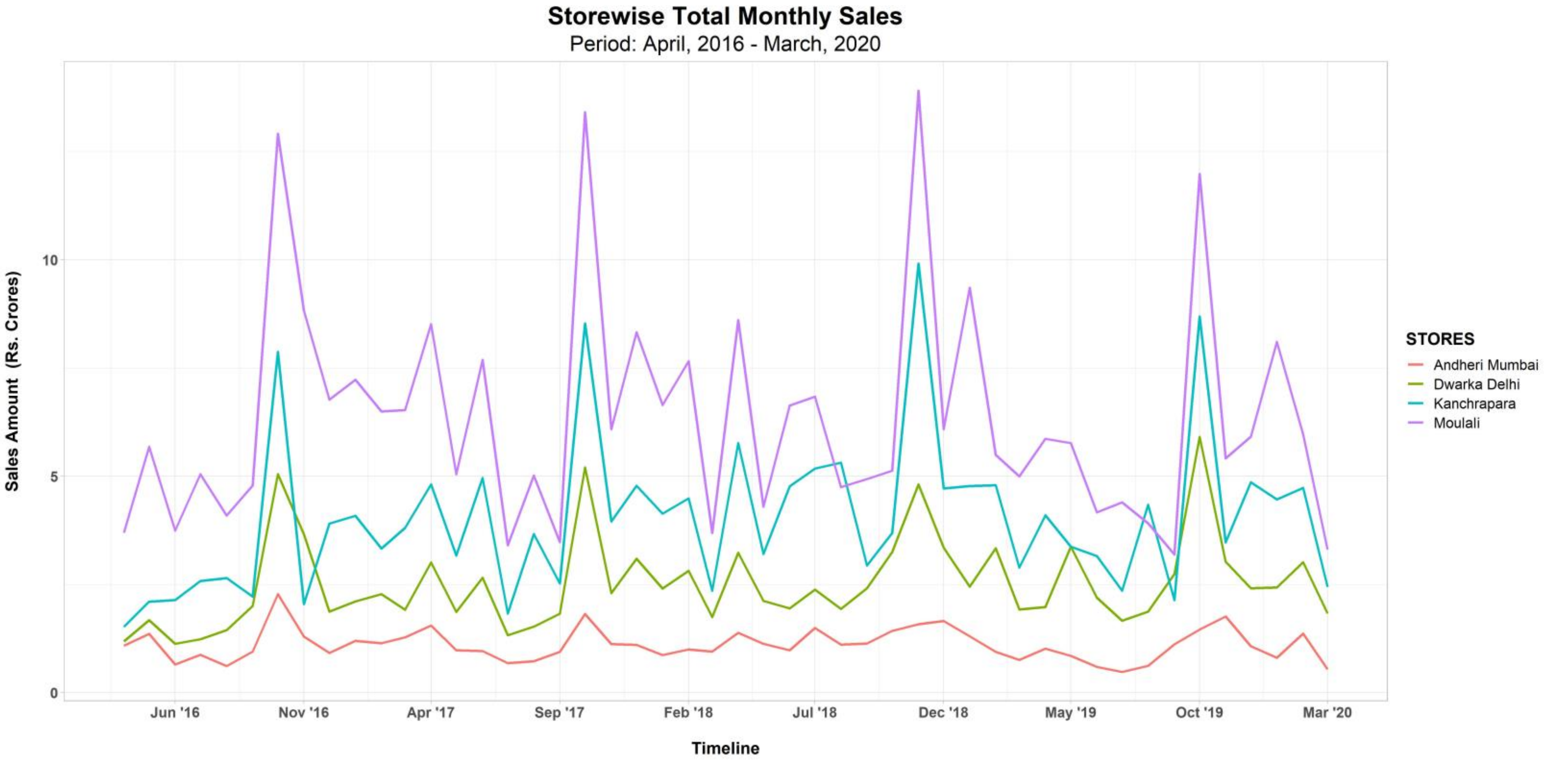
## 1. Total Monthly Sales for each store

For a given store, we plot the total sales amount for each month over the given period in the data.

Observations :

- Moulali has the highest total monthly sales for every month except August 2018 and August 2019. At both of these time points, Moulali is surpassed by Kanchrapara.

- Andheri Mumbai has the lowest total monthly sales for every month.

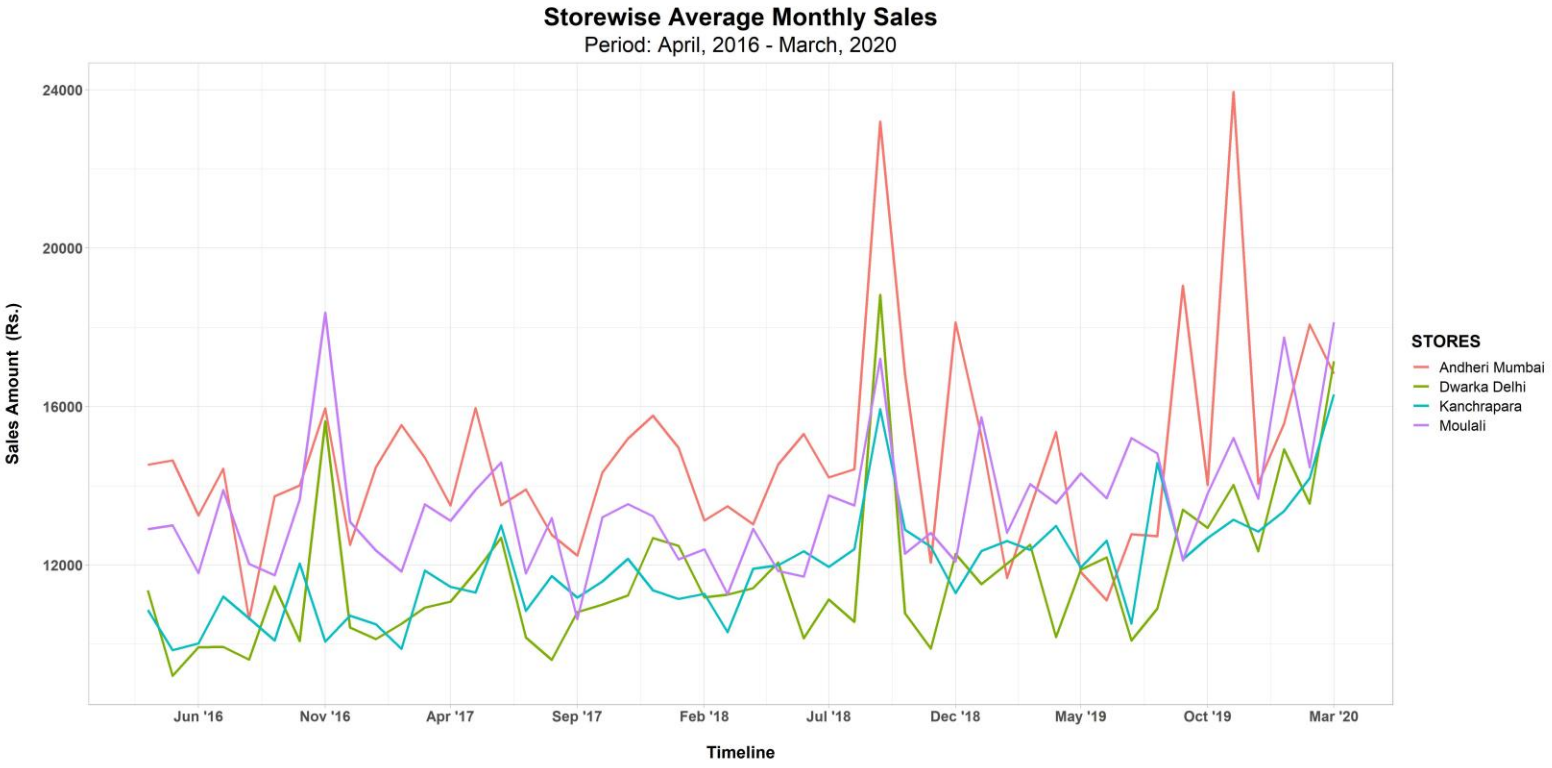


## 2. Average Monthly Sales (per transaction) for each store

For a given store, we plot the total monthly sales amount divided by the number of transactions of that month, for each month over the given period in the data.

<u>Observations :</u>

- Andheri Mumbai now becomes the store wit highest sales amount for 36 months out of 4 months.
- This shows that the conclusion from the previous graph does not hold for this graph, even though both the graphs give us a comparison of sales amount for each store.



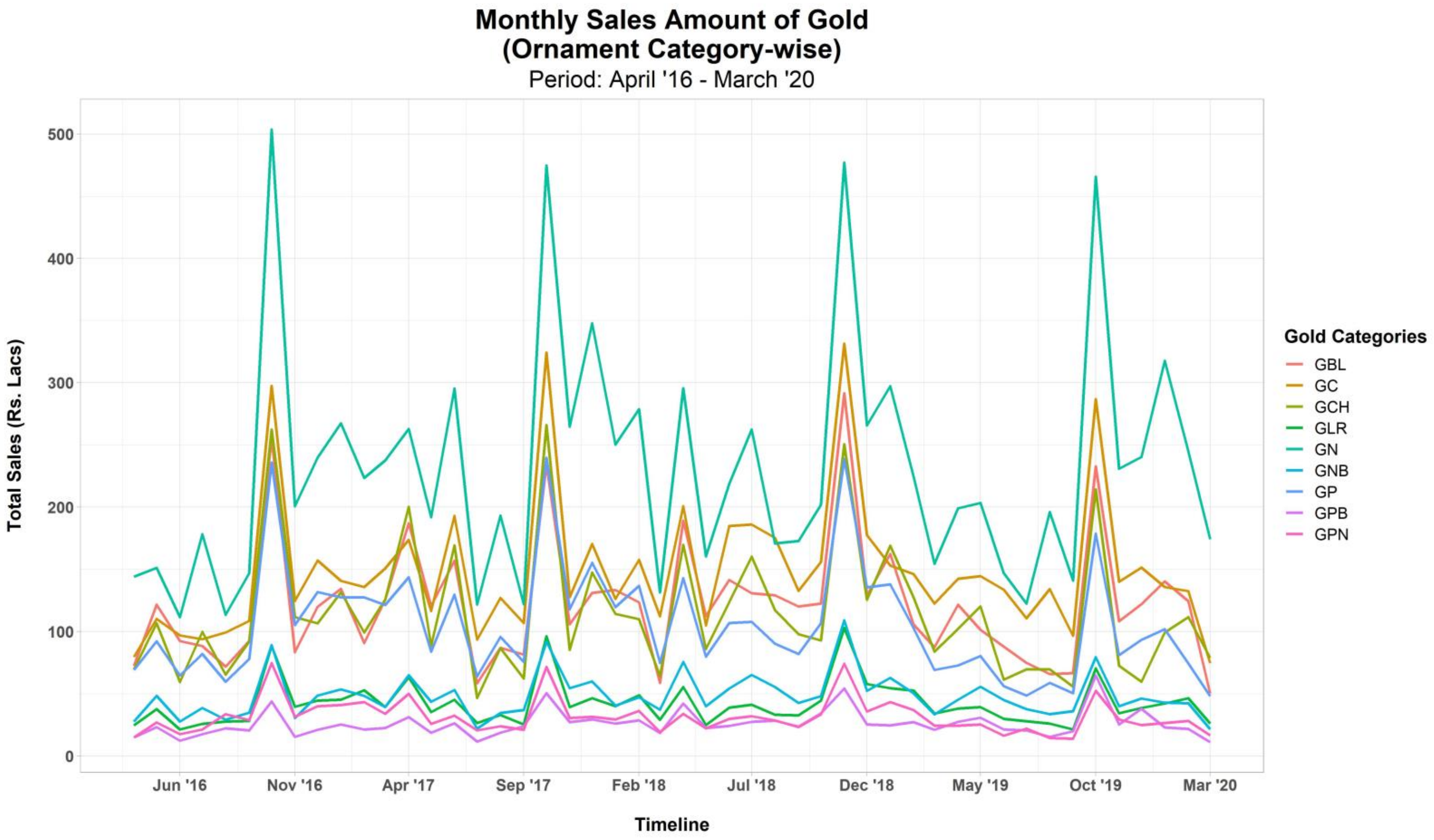
Even though **total sales amount is more in Moulali**, the **value per transaction is much higher in Andheri**.

This means,
**Moulali sells low value items to a very large number of customers**
**Andheri sells very high value items but have much smaller number of customers**

## Further scope of Graphical representation

1. Draw similar graphs for each Ornament Type

2. For each ornament type, extract ornament categories whose frequency is above a chosen threshold (for example, the 90th percentile). For these categories, draw similar graphs. For example, we take Ornament Type **GOLD** and select those categories whose frequency is above the 90th percentile. Then, the graph is as follows :



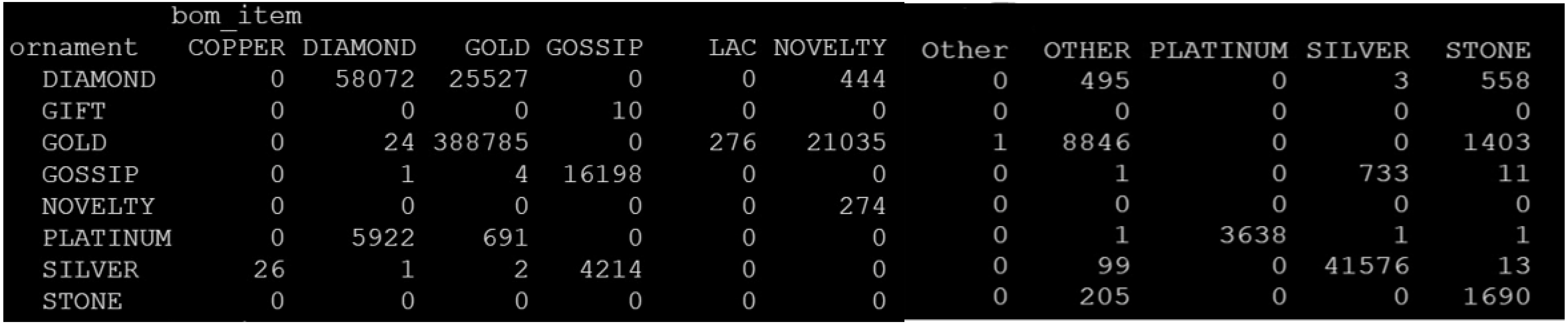
**Comments :**

1. This graph is taken over all stores. It shows that GLR is the highest selling category, combined over all stores.

2. A similar representation could be performed for each store, to obtain relevant insights.

# EXPLORATORY DATA ANALYSIS

**1. We plot a 2-way table of OrnamentType and BomItemType and obtain the following result:**

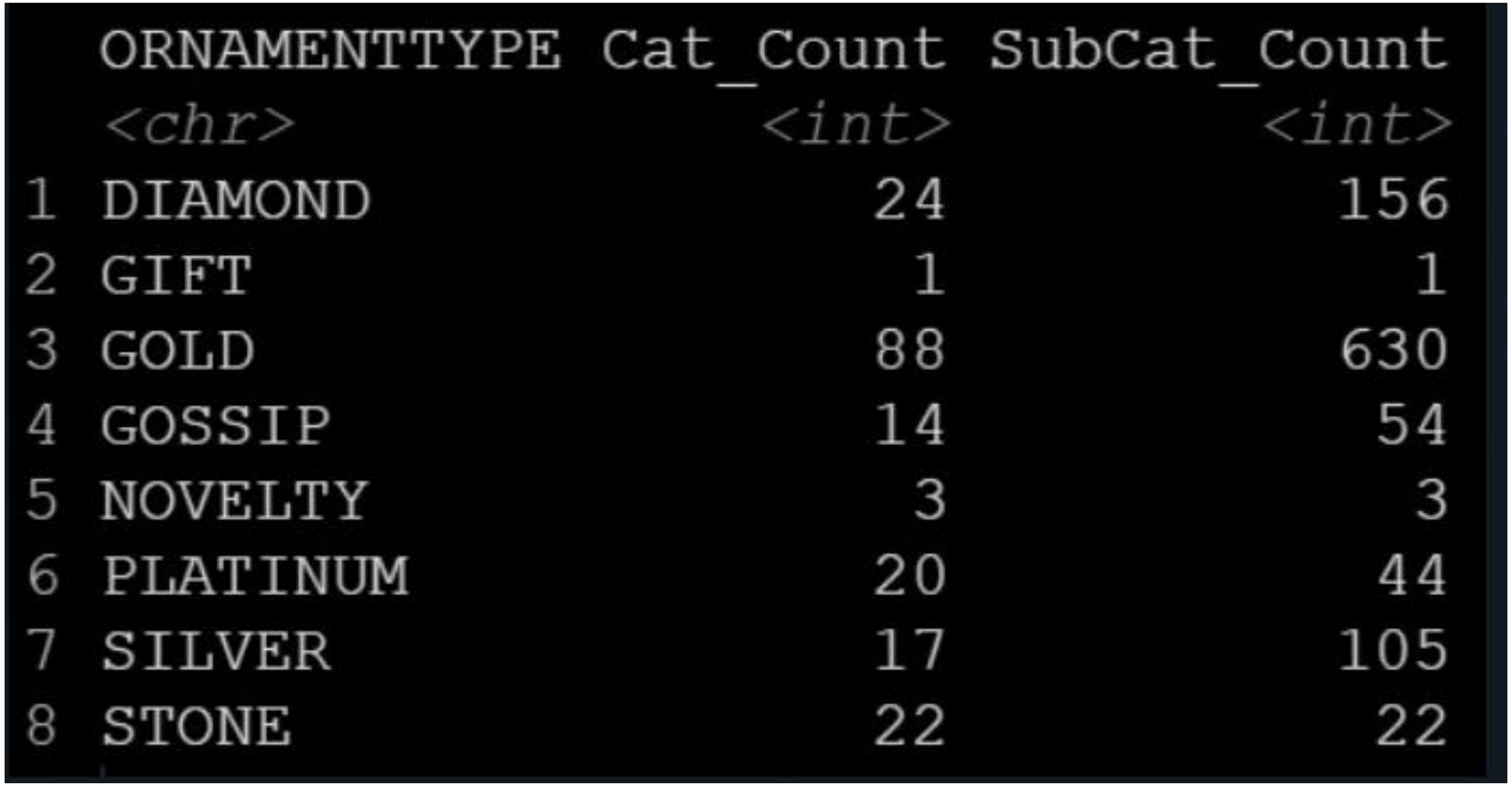


```
          bom_item
ornament    COPPER DIAMOND   GOLD GOSSIP    LAC NOVELTY  Other  OTHER PLATINUM SILVER  STONE
  DIAMOND        0   58072  25527      0      0     444      0    495        0      3    558
  GIFT           0       0      0     10      0       0      0      0        0      0      0
  GOLD           0      24 388785      0    276   21035      1   8846        0      0   1403
  GOSSIP         0       1      4  16198      0       0      0      1        0    733     11
  NOVELTY        0       0      0      0      0     274      0      0        0      0      0
  PLATINUM       0    5922    691      0      0       0      0      1     3638      1      1
  SILVER        26       1      2   4214      0       0      0     99        0  41576     13
  STONE          0       0      0      0      0       0      0    205        0      0   1690
```

Some basic observations :

1. Every point with BomItemType **LAC** corresponds to OrnamentType **GOLD**
2. Every point with BomItemType **Copper** corresponds to OrnamentType **Silver**
3. We merge ‘Other’ and ‘OTHER’, since there is only one observation for the former. This suggests that the difference in upper and lower case is an error in data entry.

## 2. Counts of OrnamentCategory and OrnamentSubCategory for each OrnamentType

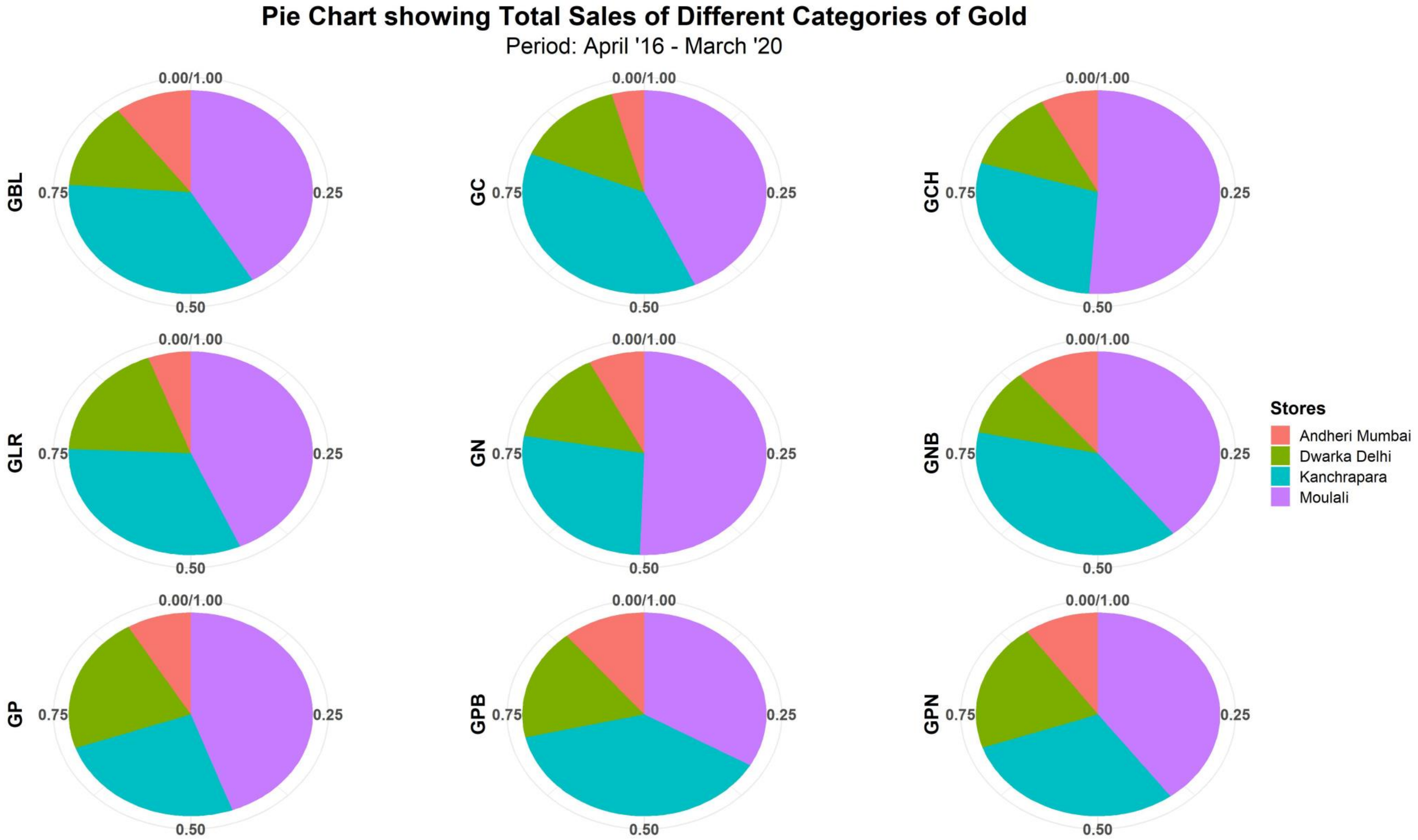


```
  ORNAMENTTYPE Cat_Count SubCat_Count
  <chr>            <int>        <int>
1 DIAMOND             24          156
2 GIFT                 1            1
3 GOLD                88          630
4 GOSSIP              14           54
5 NOVELTY              3            3
6 PLATINUM            20           44
7 SILVER              17          105
8 STONE               22           22
```

Some basic observations :

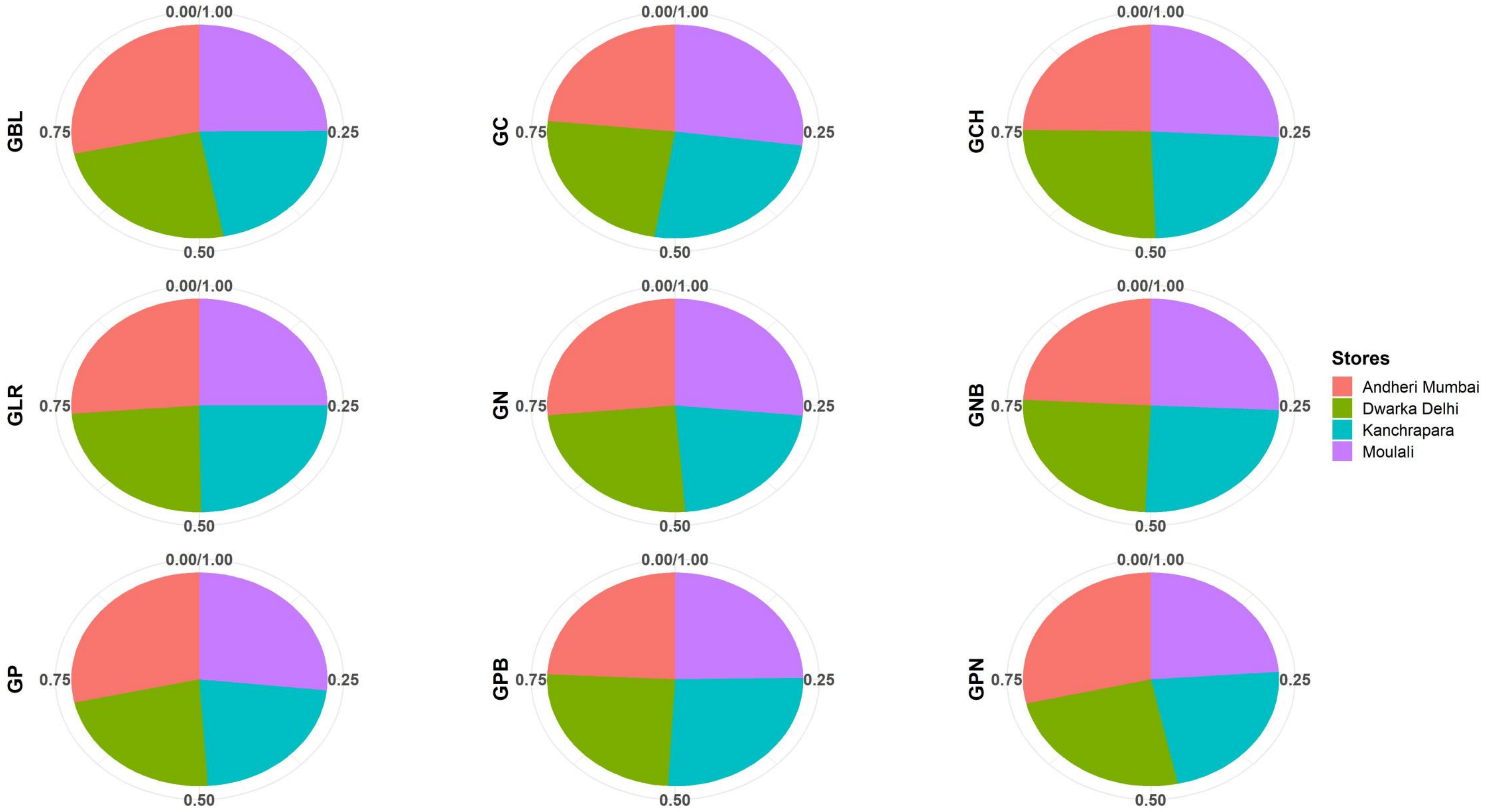
1. Only one category and one subcategory for **GIFT**
2. Only 3 categories and 3 subcategories for **NOVELTY**
3. Equal number of categories and subcategories (=22) for **STONE**

We now study the sale of the relevant categories of gold, as a percentage for each store. In other words, for each category (among the most frequent categories), we express their percentage of sales from each store.



# Pie Chart showing Average Sales of Different Categories of Gold

Period: April '16 - March '20

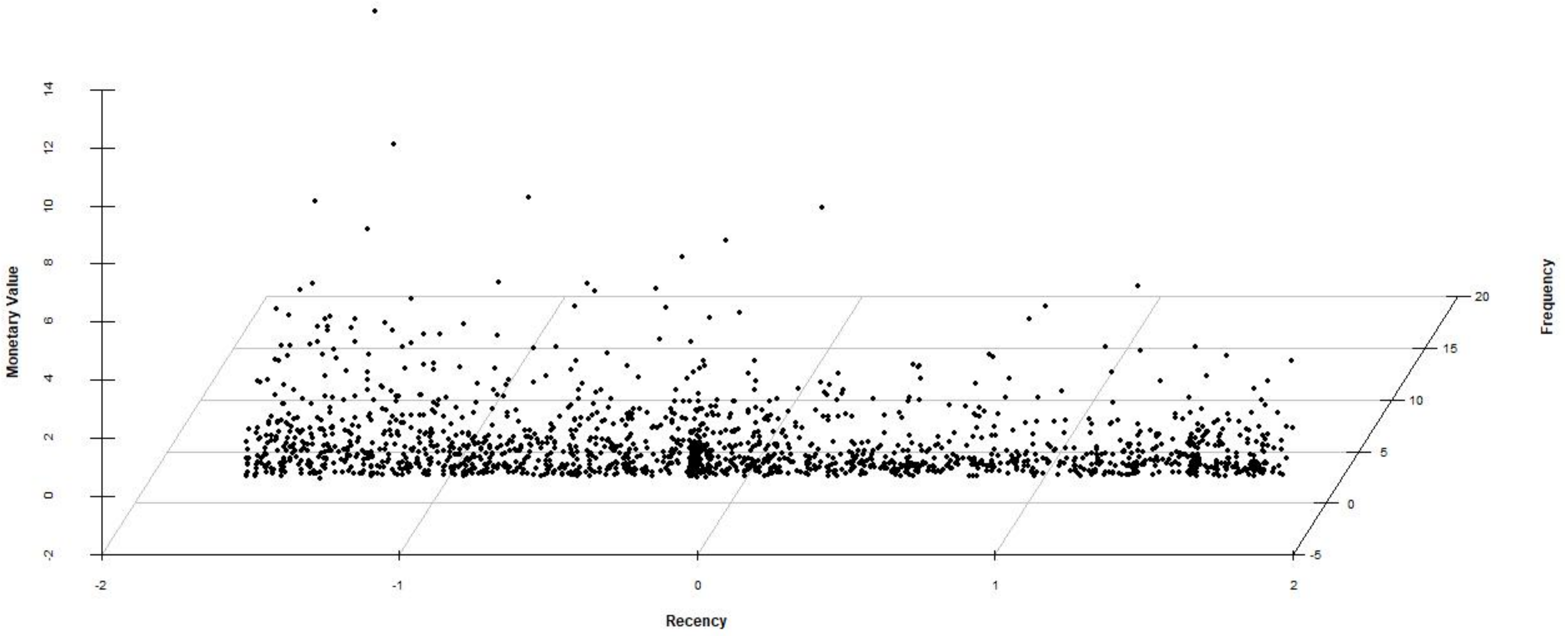


# CUSTOMER SEGMENTATION

We perform customer segmentation using RFM Model (Recency-Frequency-Monetary value) using k-means++ algorithm. This procedure will enable us to cluster customers based on their value to the company. As a result, targeted strategies could be devised to retain high-value customers and convert mid-value customers to high value customers.

We perform this segmentation for each store over 4 years of the data, to study how customer behaviour has changes over the years (in each store). This includes a study of the customer base, the gradual change in behaviour of long-term customers and the spending pattern of customers at each store.

## 1. ANDHERI MUMBAI

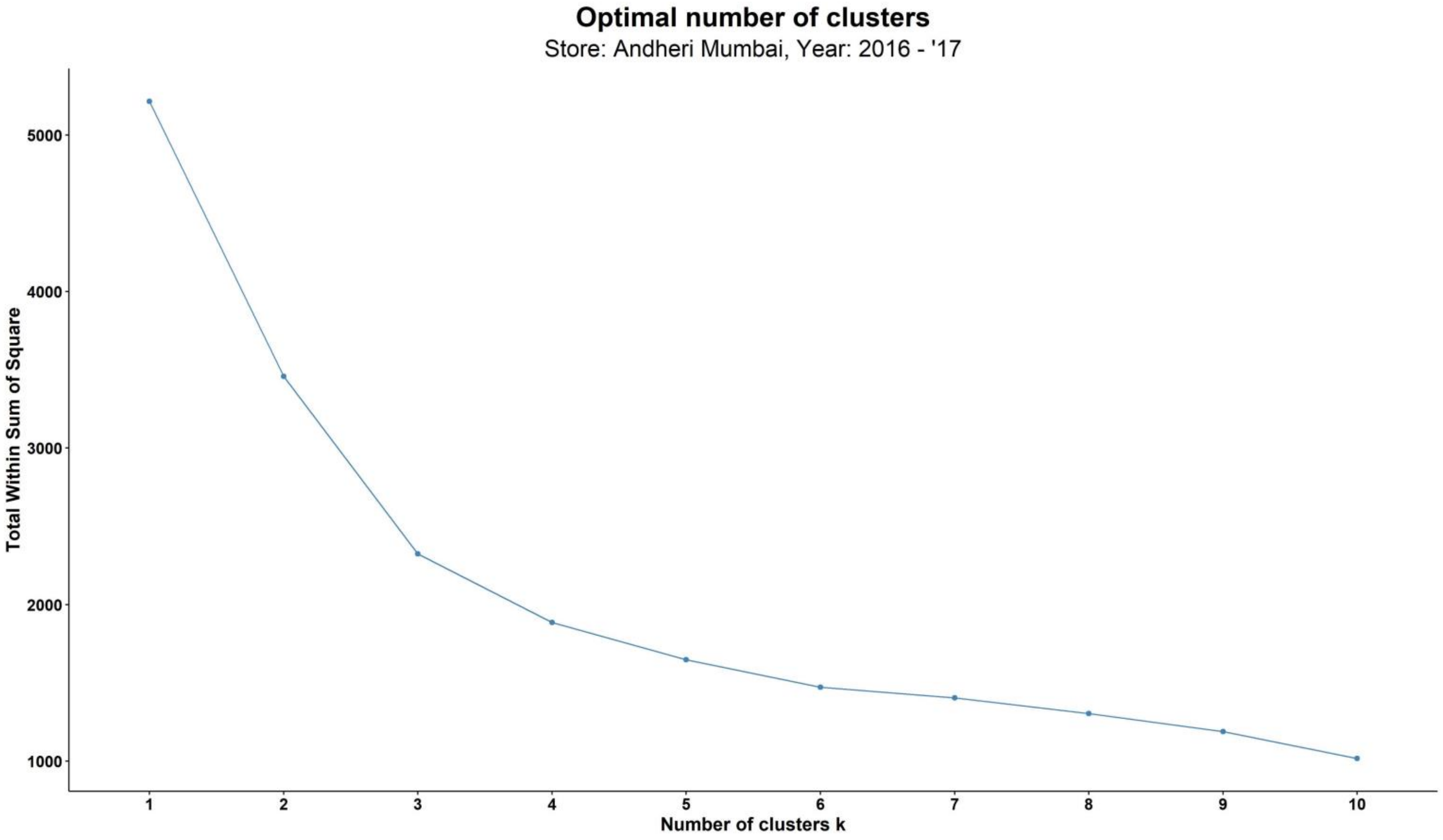


The adjacent figure is the RFM Model for Andheri Mumbai in the year **2016-2017**. Each point uniquely corresponds to a customer account. The position of a point is dictated by that customer's Recency, Frequency and Monetary value.

**Recency** of a customer is defined by the number of days that have elapsed between the last transaction date of that customer and the last available transaction date in the data.

**Frequency** of a customer is defined by the number of transactions made by the customer throughout the time period under study.

**Monetary value** of a customer is defined by the total amount of money spent by that customer during the time period under study.

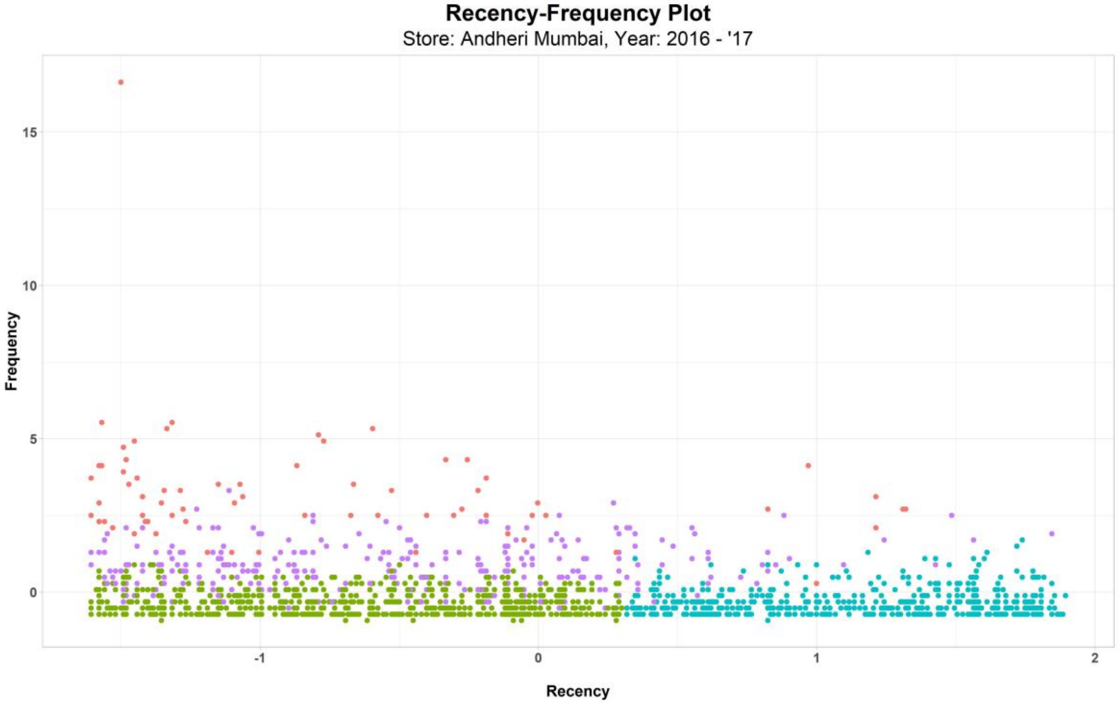

Optimal number of clusters

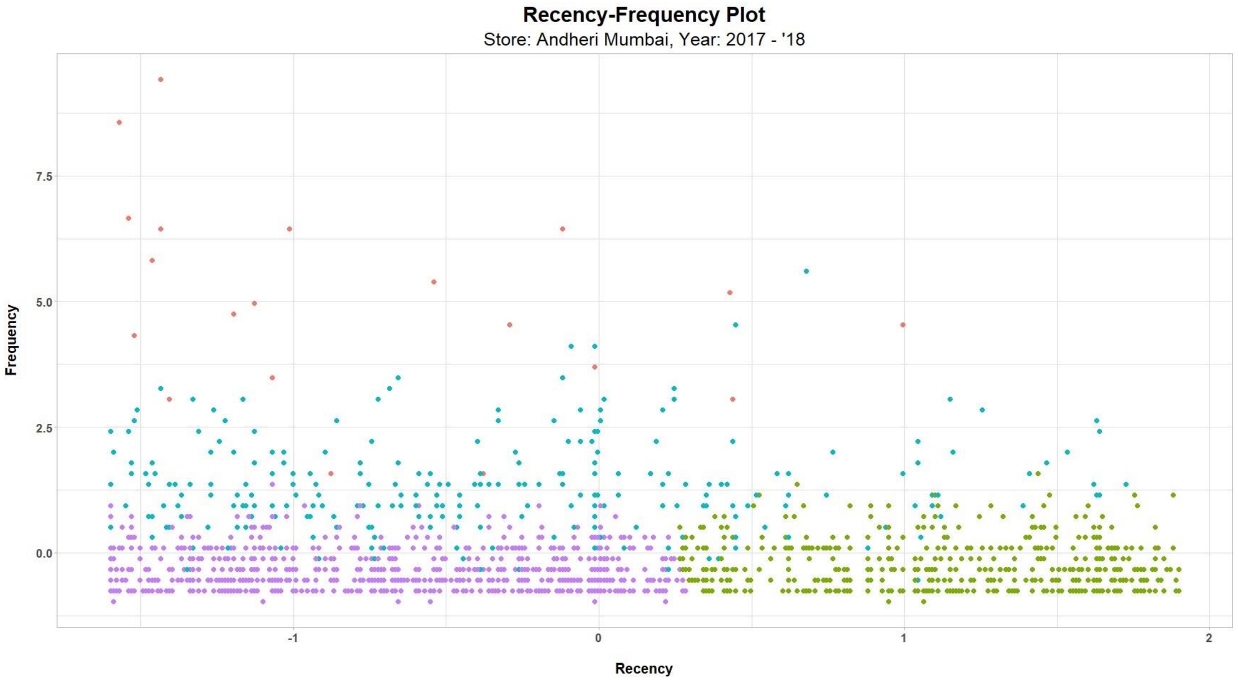
Store: Andheri Mumbai, Year: 2016 - '17

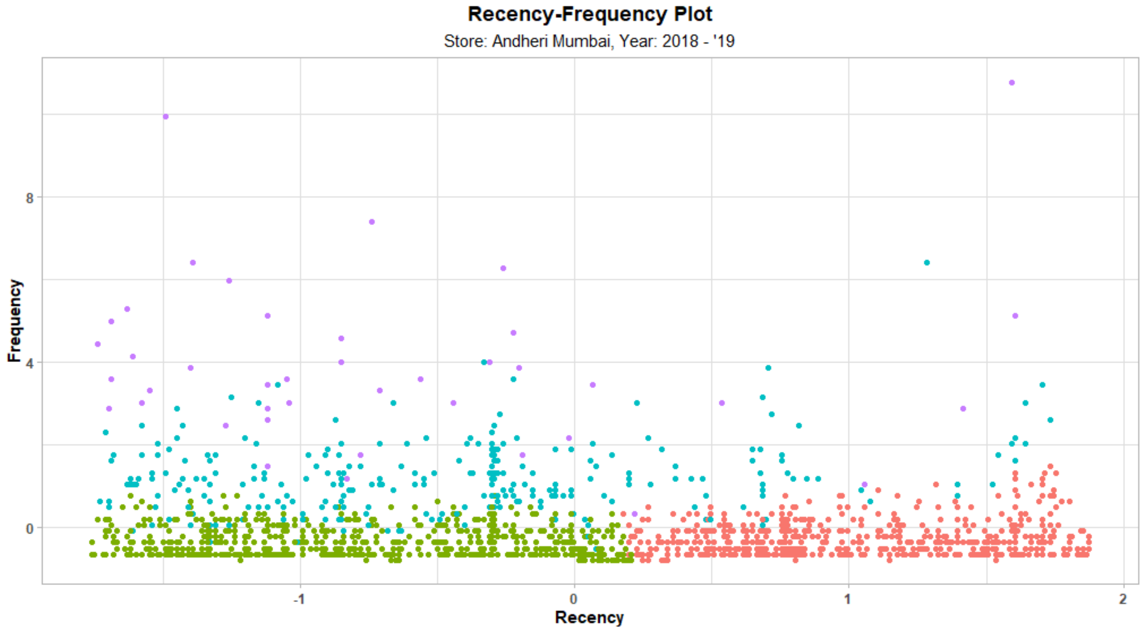
The adjacent figure is called an 'elbow plot'. We plot the final value of the objective function against each value of 'k', where k denotes the number of clusters.

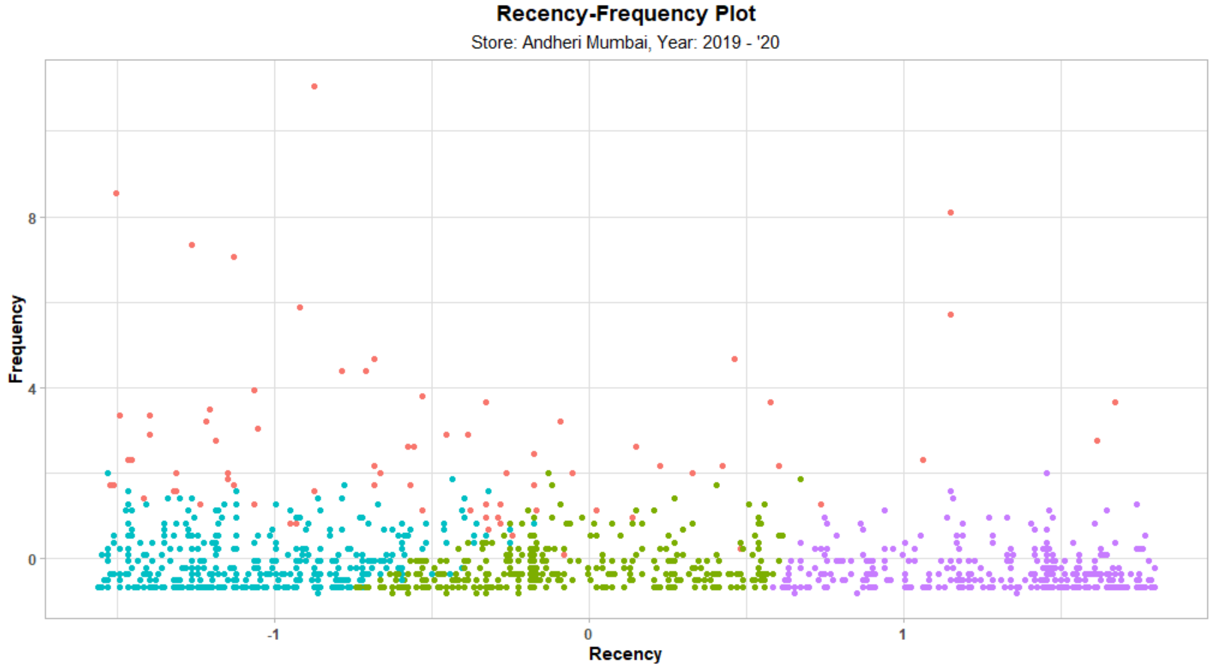
We perform elbow plot for the RFM Model for **Andheri Mumbai** for the year **2016-2017**. From the figure given, we choose k=4.

Elbow plots for the remaining 3 fiscal years are plotted similarly.


Recency-Frequency Plot
Store: Andheri Mumbai, Year: 2016 - '17
Frequency
15
10
5
0
Recency
-1
0
1
2


Recency-Frequency Plot
Store: Andheri Mumbai, Year: 2017 - '18
Frequency
7.5
5.0
2.5
0.0
Recency
-1
0
1
2


Recency-Frequency Plot
Store: Andheri Mumbai, Year: 2018 - '19
Frequency
8
4
0
Recency
-1
0
1
2


Recency-Frequency Plot
Store: Andheri Mumbai, Year: 2019 - '20
Frequency
8
4
0
Recency
-1
0
1

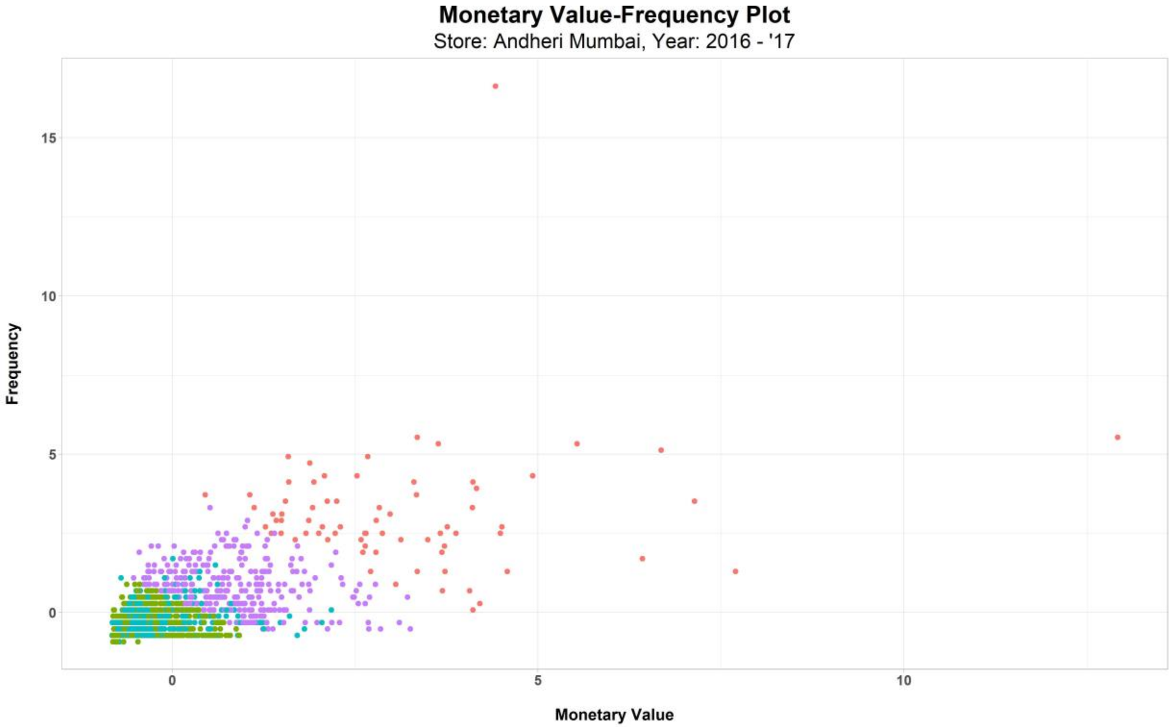
In Page 11, we plot Frequency against Recency for each customer.
Let us use the following notations to denote the clusters in Page 11 :

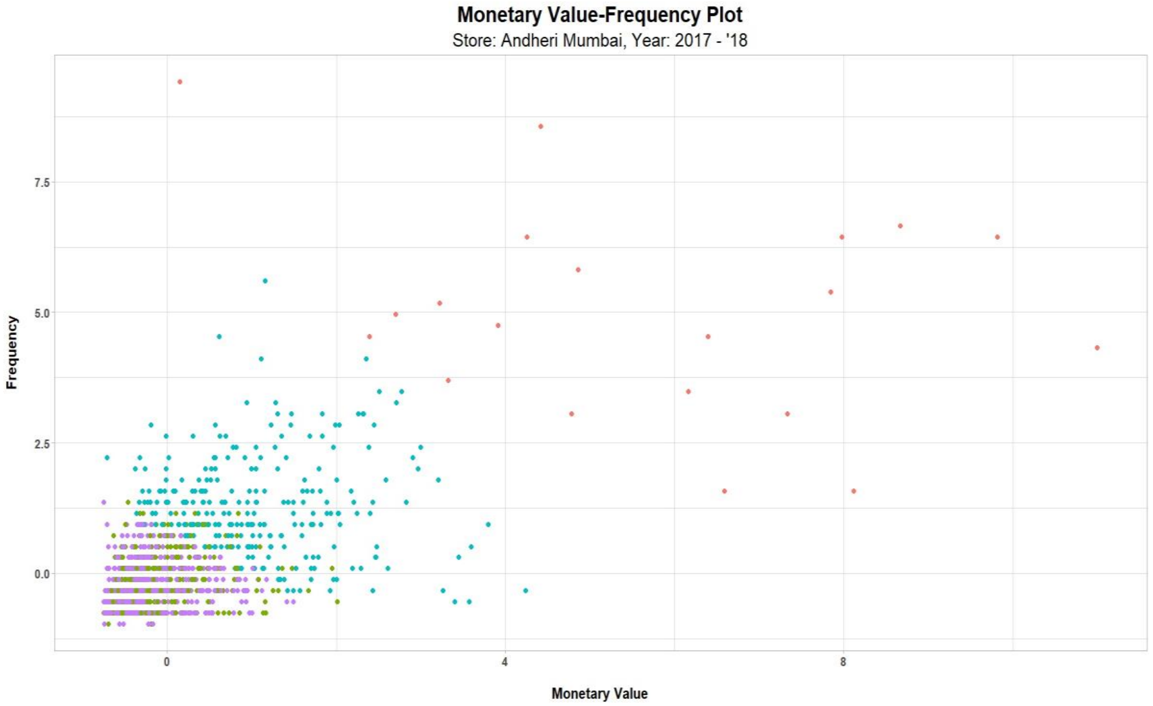
| **2016-2017** | **2017-2018** | **2018-2019** | **2019-2020** |
|---|---|---|---|
| A1 : Blue | A1 : Green | A1 : Red | A1 : Purple |
| A2 : Green | A2 : Purple | A2 : Green | A2 : Green |
| A3 : Purple | A3 : Blue | A3 : Blue | A3 : Blue |
| A4 : Red | A4 : Red | A4 : Purple | A4 : Red |

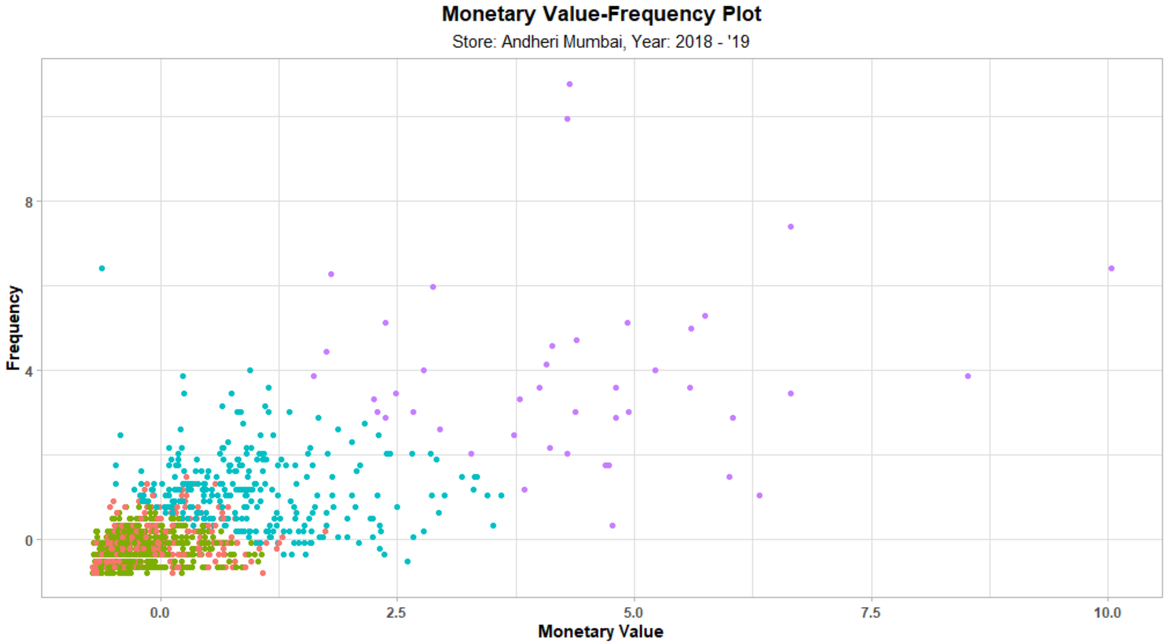
1. Frequency of customers seem to have increased over the years, especially for cluster A4. One exception is in cluster A4 in the year 2016-2017, where a customer has a frequency score of above 15.

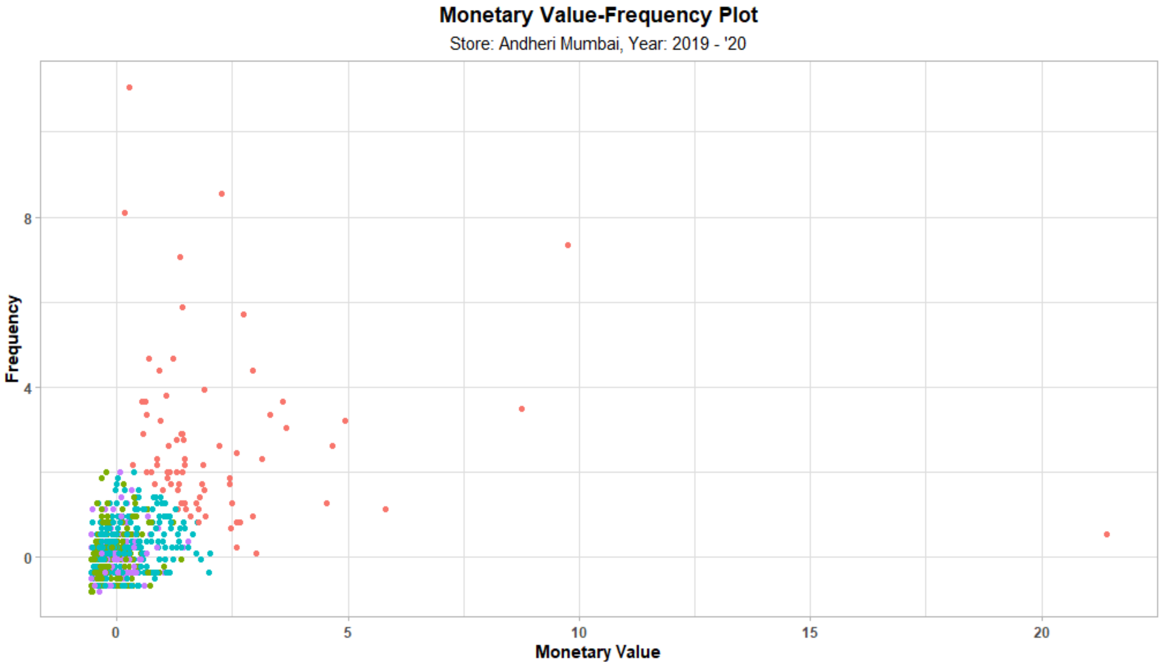
2. Clusters A1 and A2 are only distinguished among themselves through Recency. On the other hand, A1 and A2 differ from A3 and A4 in terms of frequency.

3. In all 4 years, the clusters can be ordered in increasing order of customer value as A1<A2<A3<A4. Hence, customers in cluster A4 are the highest value customers in the company.


Monetary Value-Frequency Plot
Store: Andheri Mumbai, Year: 2016 - '17
Frequency
15
10
5
0
0
5
10
Monetary Value


Monetary Value-Frequency Plot
Store: Andheri Mumbai, Year: 2017 - '18
Frequency
7.5
5.0
2.5
0.0
0
4
8
Monetary Value


Monetary Value-Frequency Plot
Store: Andheri Mumbai, Year: 2018 - '19
Frequency
8
4
0
0.0
2.5
5.0
7.5
10.0
Monetary Value


Monetary Value-Frequency Plot
Store: Andheri Mumbai, Year: 2019 - '20
Frequency
8
4
0
0
5
10
15
20
Monetary Value

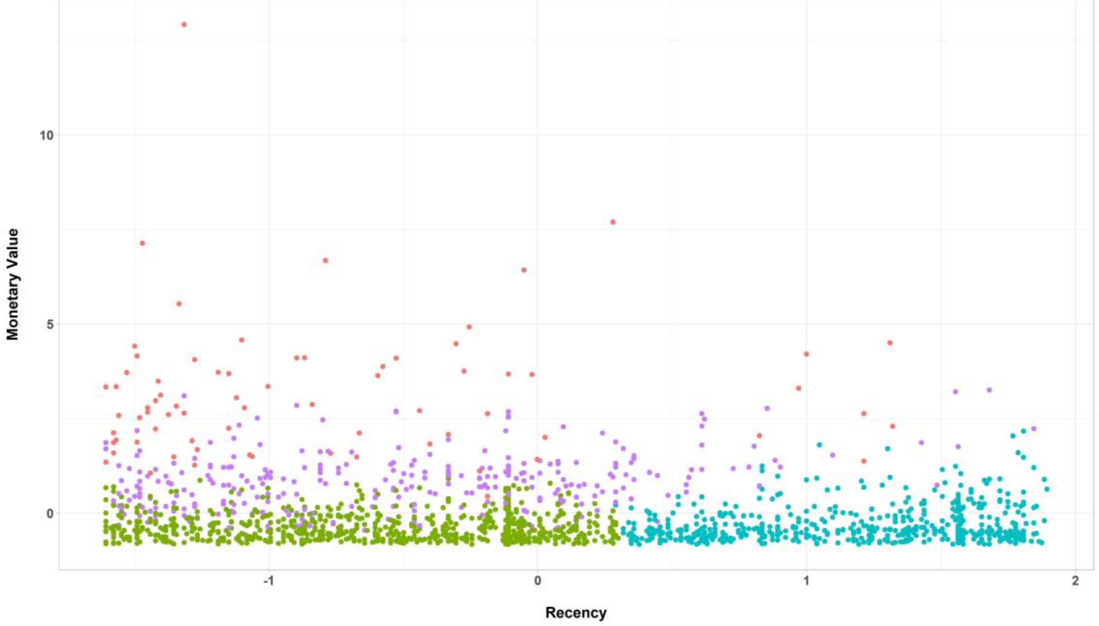
In Page 13, we plot Frequency against Monetary value for each customer.
Let us use the following notations to denote the clusters in Page 13 :

| 2016-2017 | 2017-2018 | 2018-2019 | 2019-2020 |
|---|---|---|---|
| A1 : Blue | A1 : Green | A1 : Red | A1 : Purple |
| A2 : Green | A2 : Purple | A2 : Green | A2 : Green |
| A3 : Purple | A3 : Blue | A3 : Blue | A3 : Blue |
| A4 : Red | A4 : Red | A4 : Purple | A4 : Red |

1. Frequency and monetary value of customers play no role in separating clusters A1 and A2. In 2019-2020 A1, A2 and A3 cannot be distinguished in terms of frequency and monetary value. These clusters only differ in terms of recency.

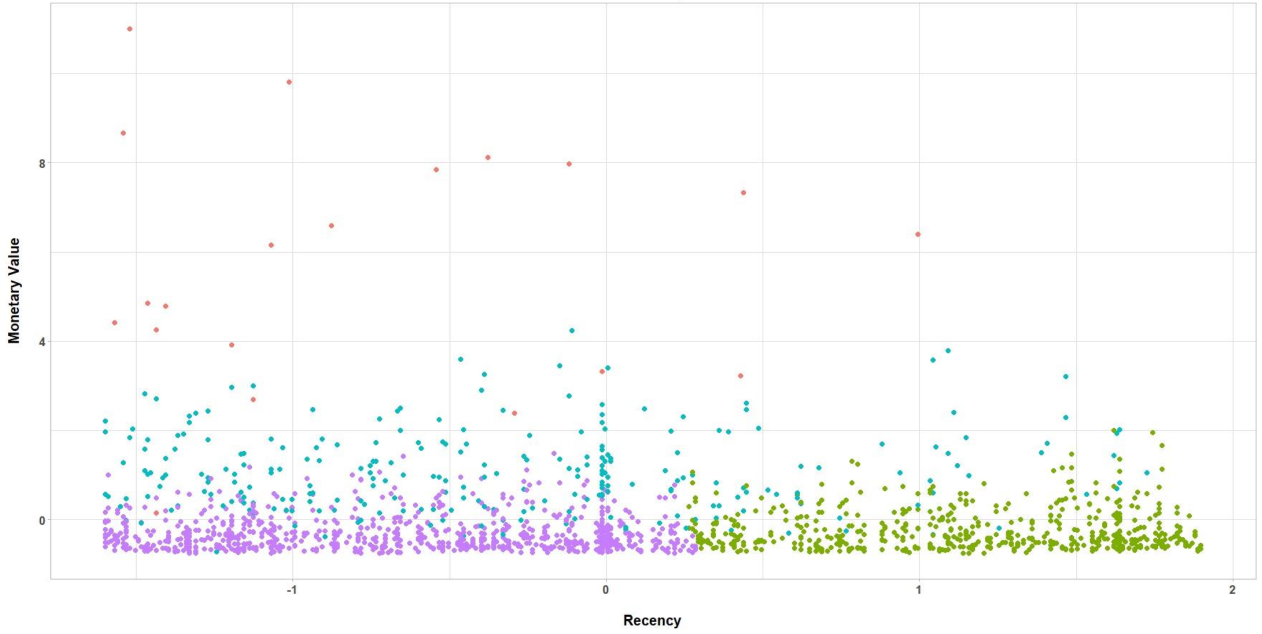
2. Cluster A4 has the highest value customers.

3. Points in Cluster A4 which occur in the lower right corner of the cluster spend high amounts of money per transaction. Points in Cluster A4 which occur in the upper left corner of the cluster are highly frequent, and hence have a high cumulative monetary value.



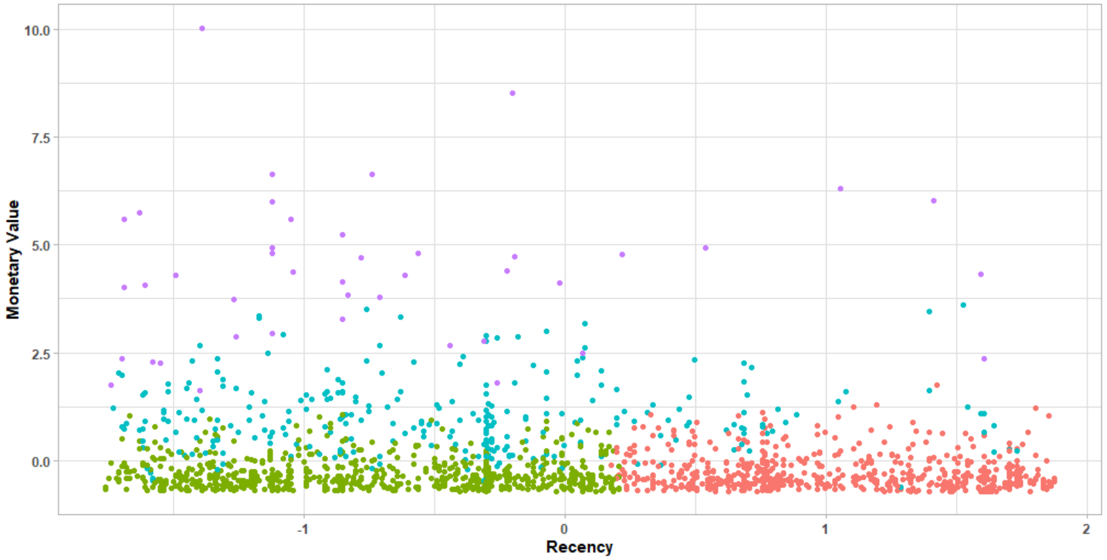
Recency-Monetary Value Plot

Store: Andheri Mumbai, Year: 2016 - '17



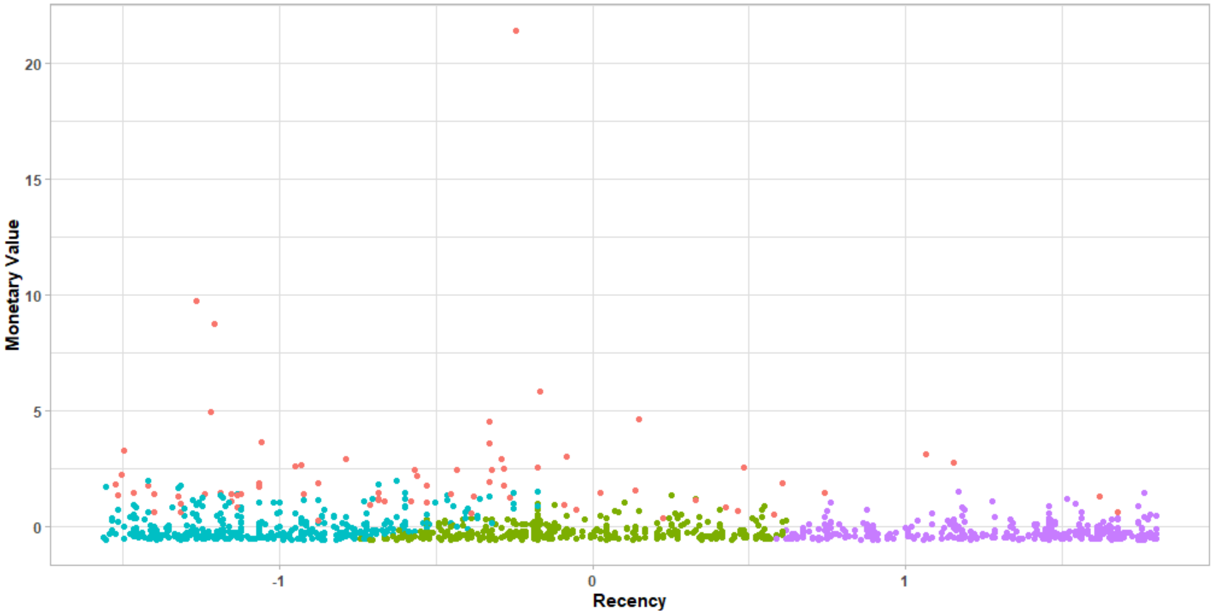
Recency-Monetary Value Plot

Store: Andheri Mumbai, Year: 2017 - '18



Recency-Monetary Value Plot

Store: Andheri Mumbai, Year: 2018 - '19



Recency-Monetary Value Plot

Store: Andheri Mumbai, Year: 2019 - '20

In Page 15, we plot Monetary value against Recency for each customer.
Let us use the following notations to denote the clusters in Page 15 :

| 2016-2017 | 2017-2018 | 2018-2019 | 2019-2020 |
|---|---|---|---|
| A1 : Blue | A1 : Green | A1 : Red | A1 : Purple |
| A2 : Green | A2 : Purple | A2 : Green | A2 : Green |
| A3 : Purple | A3 : Blue | A3 : Blue | A3 : Blue |
| A4 : Red | A4 : Red | A4 : Purple | A4 : Red |

1. Cluster A4 has the highest monetary value customers.

2. Monetary value usually has an upper limit in 10. There is one exception in 2019-2020, with a monetary value above 20.

3. Order of customer value remains same as that in Page-10.

**Note** :
**On approval of the relevance of this study, further observations such as gradual change in each cluster, following behavioural pattern of a given customer, etc. could be performed.**