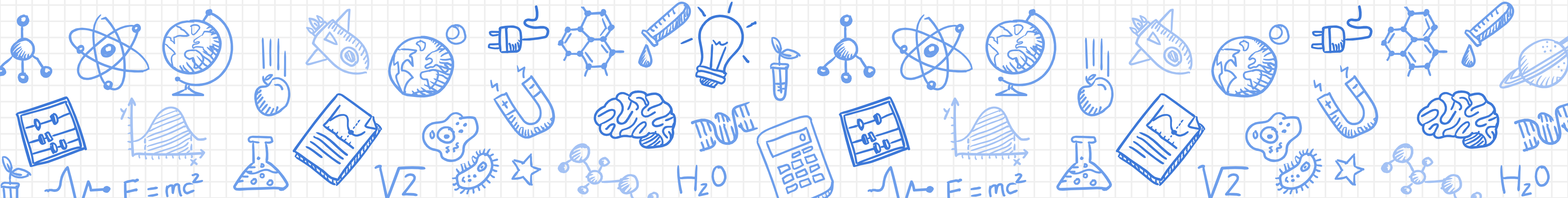
SUCI AULYA PUTRI

CUSTOMER SEGMENTATION on Online Retail Dataset

Using Python Programming



Use Case Summary

Objective Statement:

1. Get business insight about how many product sold every month.
2. Get business insight about how much customer spend their money every month.
3. Get business insight about how many customers make transactions each month.
4. Get business insight about how much is the frequency of transactions in months, days, and hours.
5. Get business insight about the most popular products.
6. Get business insight about the most consumers by country.
7. To reduce risk in deciding where, when, how, and to whom a product, service, or brand will be marketed.
8. To increase marketing efficiency by directing effort specifically toward the designated segment in a manner consistent with that segment's characteristics.

Challenges:

1. Large size of data, can not maintain by excel spreadsheet.
2. Demography data have a lot missing values.

Business Benefit:

1. Helping Business Development Team to create product differentiation based on the characteristic for each customer.
2. Know how to treat customer with specific criteria.

Expected Outcome:

1. Know how many product sold every month.
2. Know how much customer spend their money every month.
3. Know how many customers make transactions each month.
4. Know how much is the frequency of transactions in months, days, and hours.
5. Know the most popular products.
6. Know the most customer by the country.
7. Customer segmentation analysis.
8. Recommendation based on customer segmentation.

Business Understanding

Retail is the process of selling consumer goods or services to customers through multiple channels of distribution to earn a profit.

This case has some **business question** using the data:

- How many product sold every month?
- How much customer spend their money every month?
- How many customers make transactions each month?
- How much is the frequency of transactions in months, days, and hours?
- What products are the most popular?
- Most consumers by country?
- How about Customer segmentation analysis?
- How about recommendation based on customer segmentation?

Data Understanding

The data **consists** of 2 datasets.

Dataset I

- Online Retail Dataset between 01/12/2009 until 09/12/2010.
- The dataset consists of 525461 rows and 8 columns..

Dataset II

- Online Retail Dataset between 01/12/2010 until 09/12/2011.
- The dataset consists of 541910 rows and 8 columns..

Dataset

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 until 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

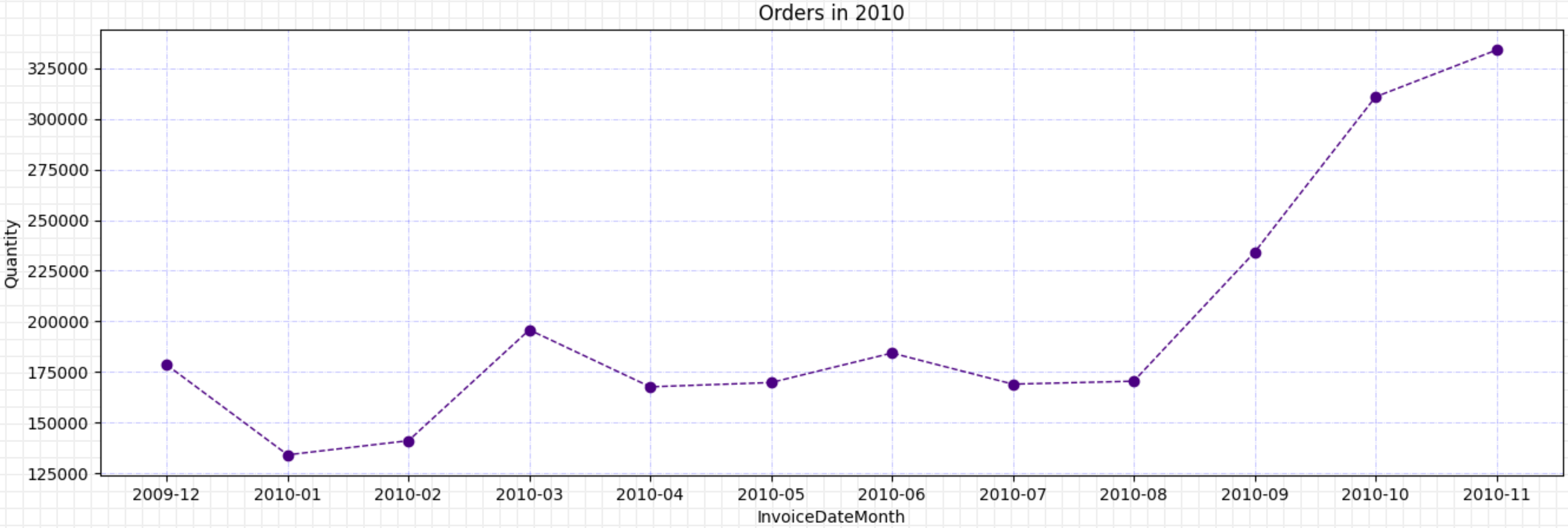
Data Dictionary

- Invoice : Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- StockCode : Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
- Description : Product (item) name. Nominal.
- Quantity : The quantities of each product (item) per transaction. Numeric.
- Invoice Date : Invice date and time. Numeric. The day and time when a transaction was generated.
- Price : Unit price. Numeric. Product price per unit in sterling (Â£).
- Customer ID : Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
- Country : Country name. Nominal. The name of the country where a customer resides.

Source Data

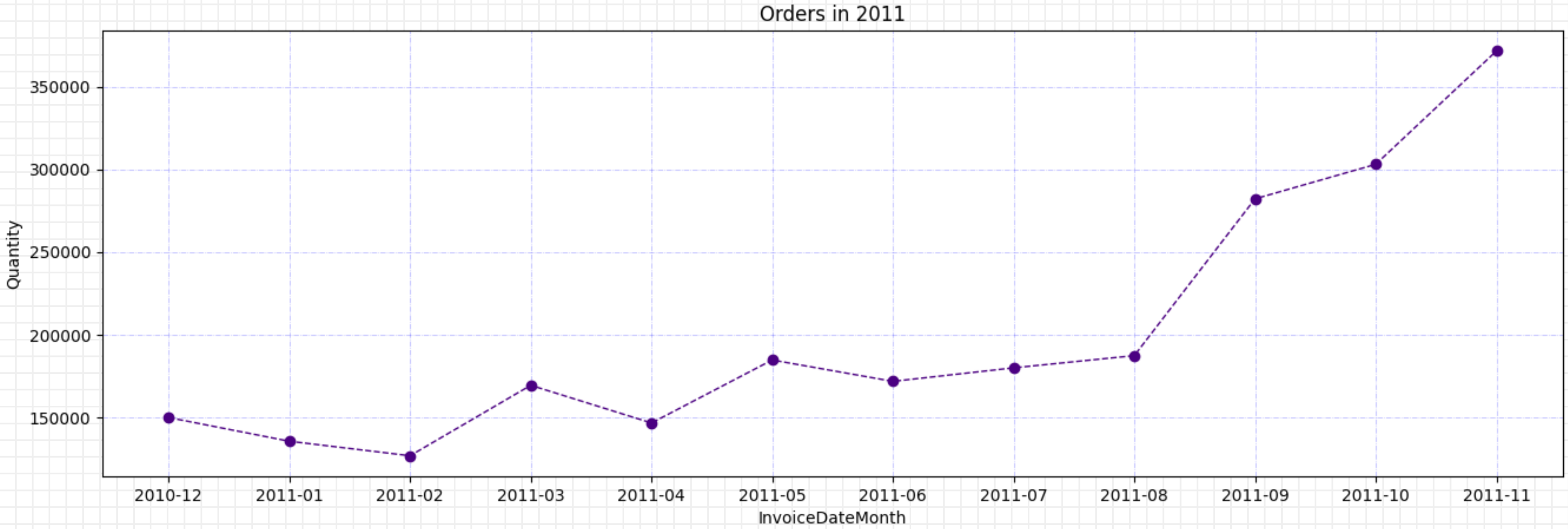
https://www.kaggle.com/mathchi/online-retail-ii-data-set-from-ml-repository

NUMBER OF PRODUCTS SOLD EACH MONTH IN 2009 - 2010



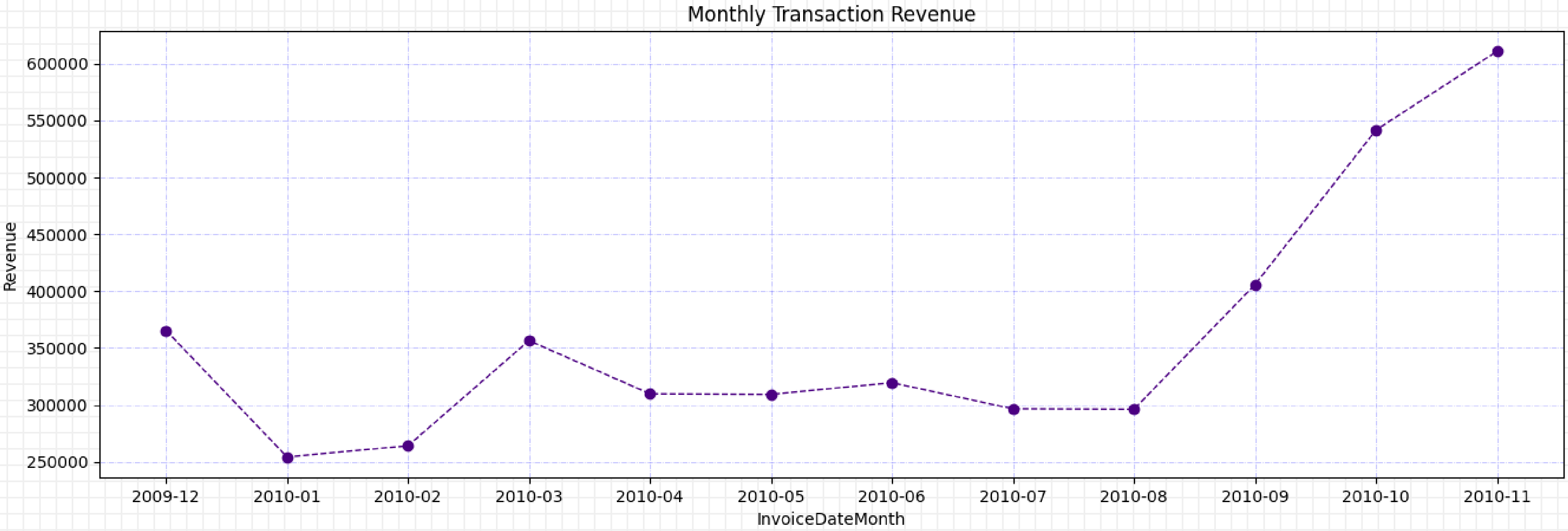
Product sold in November has the highest quantity that has around 13,97% product sold from all transaction along 1 year. Therefore the business team can increase sales in this month such as promoting new products to customers in this month.

NUMBER OF PRODUCTS SOLD EACH MONTH IN 2010 - 2011



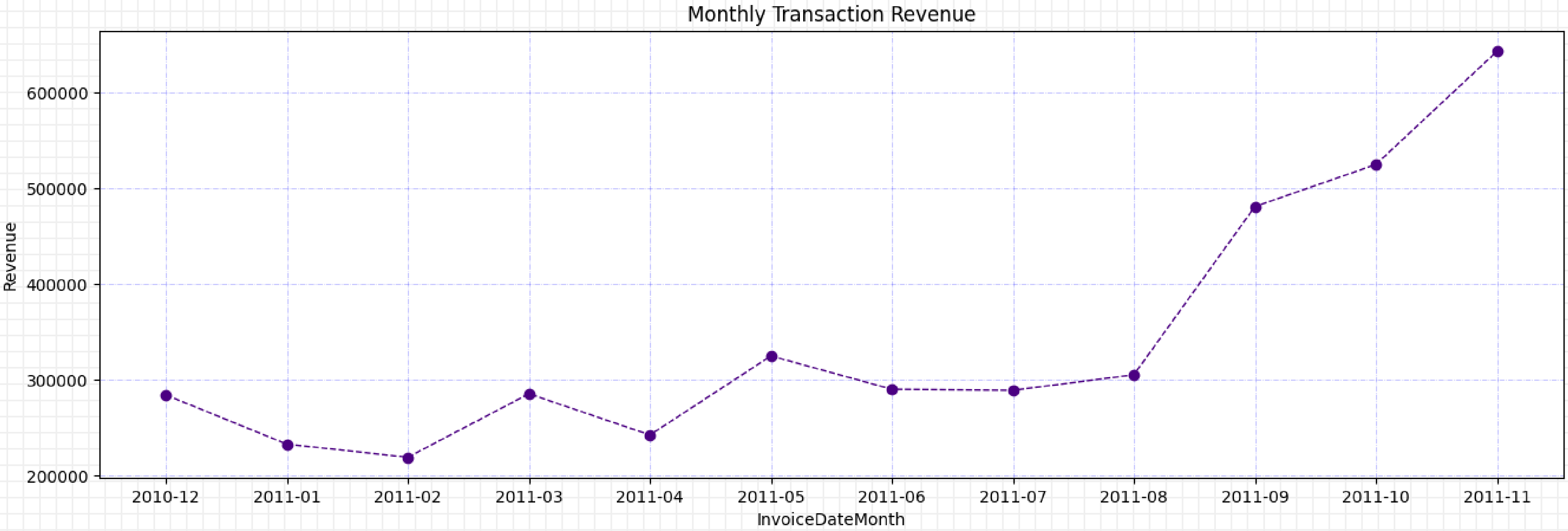
Same as in 2010, product sold in November has the highest quantity that has around 15,42% product sold from all transaction along 1 year. Therefore the business team can increase sales in this month such as promoting new products to customers in this month.

THE AMOUNT OF MONEY THAT CUSTOMERS SPEND ON EACH MONTH IN 2009 - 2010



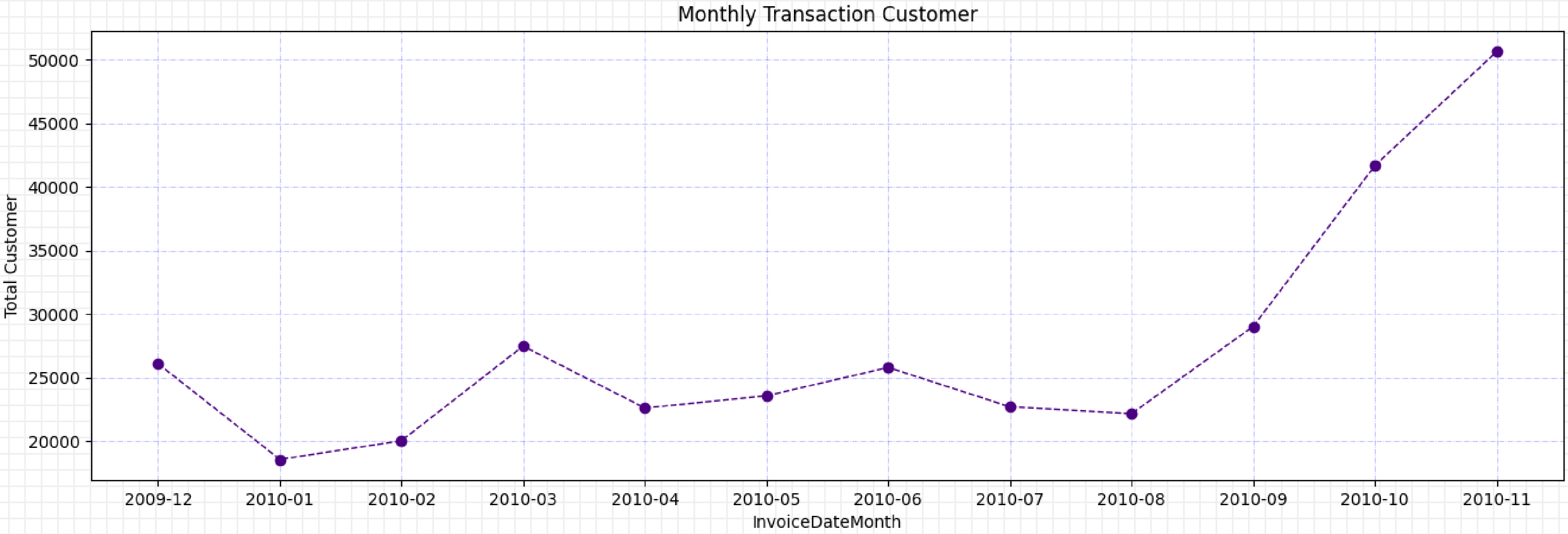
Revenue in November has the highest amount that has around 14,11% revenue from total revenue along 1 year. Therefore the business team can replicate the success of sales strategies in November to be implemented in other months.

THE AMOUNT OF MONEY THAT CUSTOMERS SPEND ON EACH MONTH IN 2010 - 2011



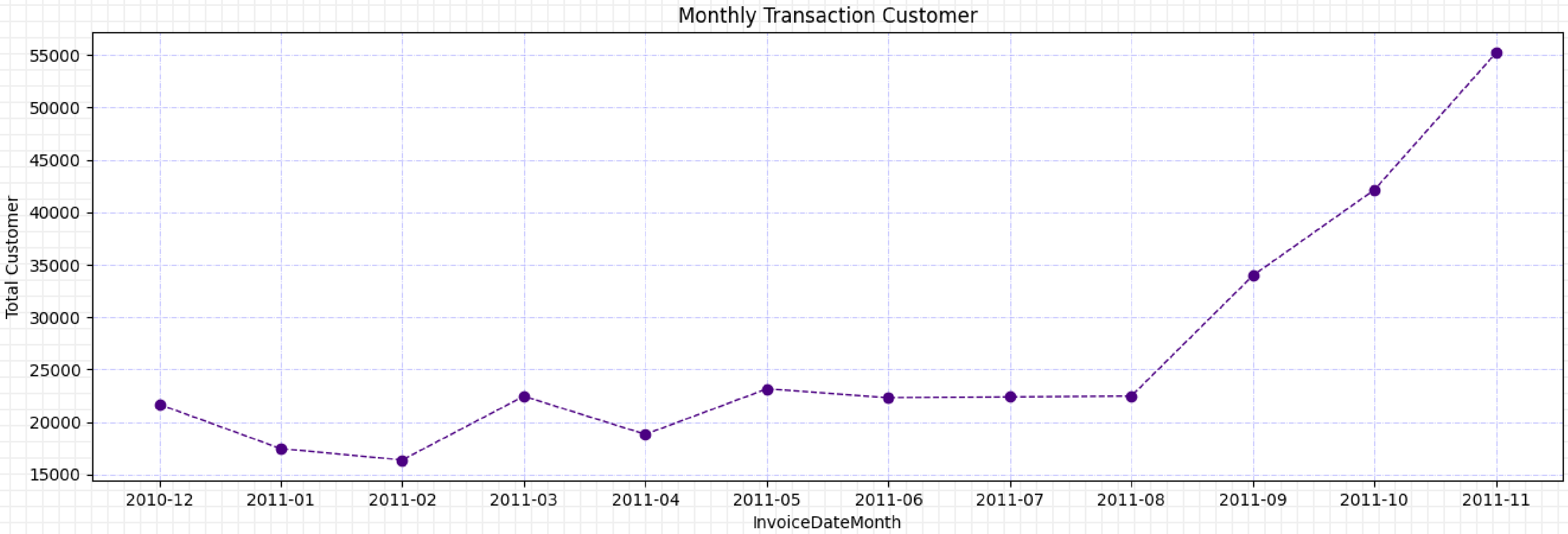
Same as in 2010, ,revenue in November has the highest amount that has around 15,6% revenue from total revenue along 1 year. Therefore the business team can replicate the success of sales strategies in November to be implemented in other months.

NUMBER OF CUSTOMERS WHO MAKE TRANSACTIONS EVERY MONTH IN 2009 - 2010



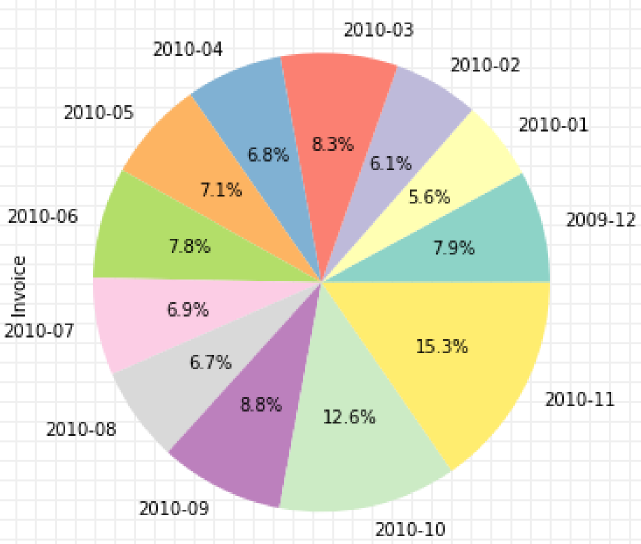
The number of customers from December 2009 to November 2010 was fluctuating. However, in general, the number of customers almost every month tends to show an increase, only in January, April, July, and August do the number of customers show a decrease.The business team can provide special discounts in January, April, July, and August to increase the number of customers and sales in this month.

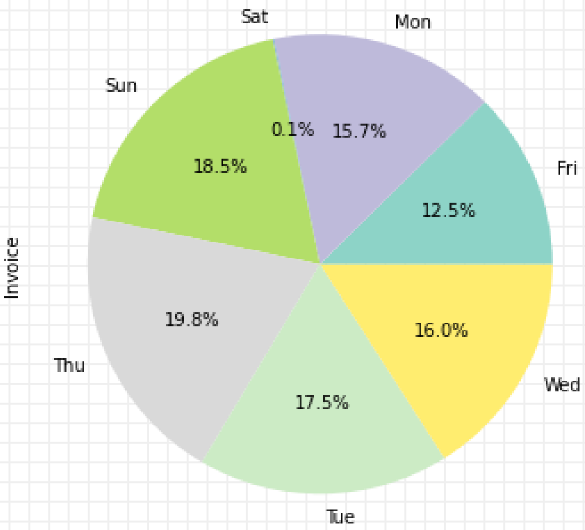
NUMBER OF CUSTOMERS WHO MAKE TRANSACTIONS EVERY MONTH IN 2010 - 2011

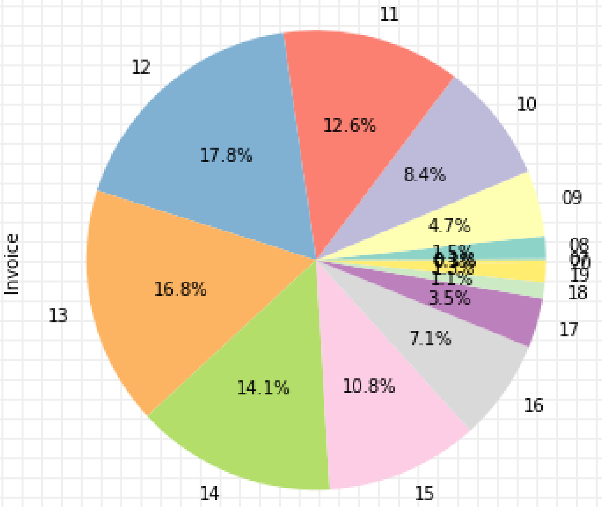


The number of customers from December 2010 to November 2011 was fluctuating. However, in general, the number of customers almost every month tends to show an increase, only in January, February,and April do the number of customers show a decrease.The business team can provide special discounts in January, February,and April to increase the number of customers and sales in this month.

TRANSACTION FREQUENCY EVERY MONTH, DAY, AND HOUR IN 2009 - 2010


Transaction Frequency Every Month in 2010


Transaction Frequency Every Day in 2010

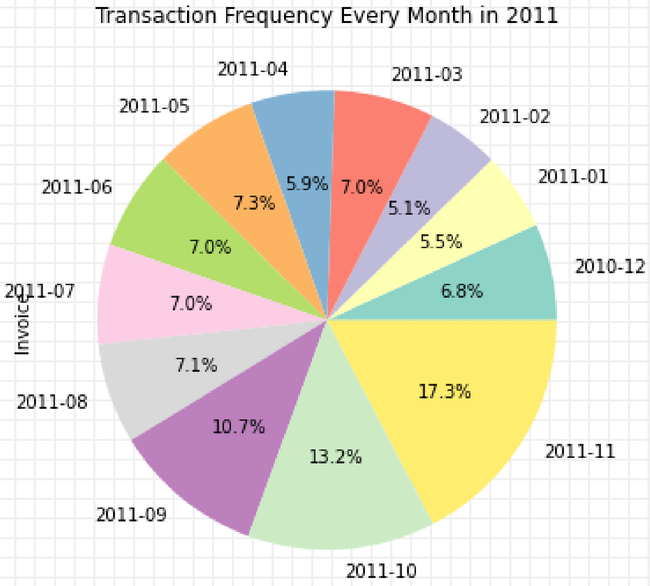

Transaction Frequency Every Hour in 2010

The number of customers in November is the highest number of customers that has around 15,3% of the total customers along 1 year. Business teams can increase sales by promoting new products to customers on November.

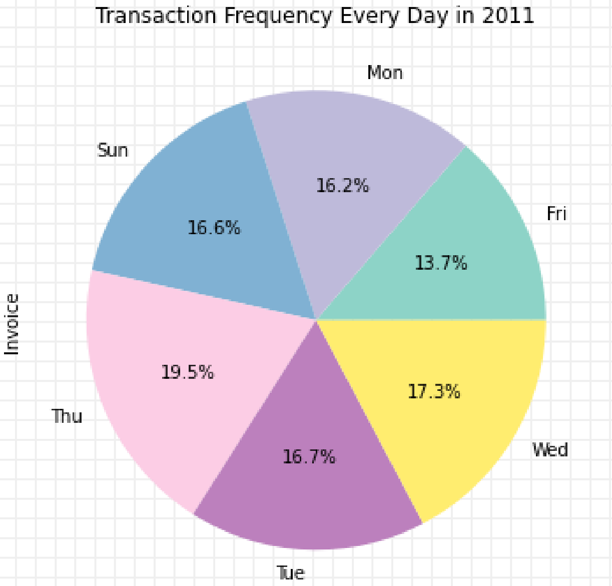
Most consumers make transactions on Thursday, which is around 19,8% of the total daily transactions. Business teams can increase sales by promoting new products to customers on Thursday.

Most consumers order the products at 12 AM with a transaction amount of 17.8% of the total daily transactions. Business teams can increase sales by promoting new products to customers at 12 A.M

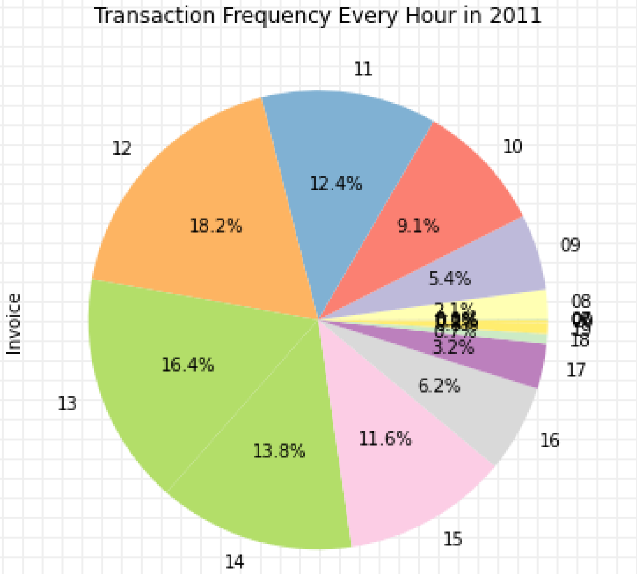
TRANSACTION FREQUENCY EVERY MONTH, DAY, AND HOUR IN 2010 - 2011



The number of customers in November is the highest number of customers that has around 17,3% of the total customers along 1 year. Business teams can increase sales by promoting new products to customers on November.

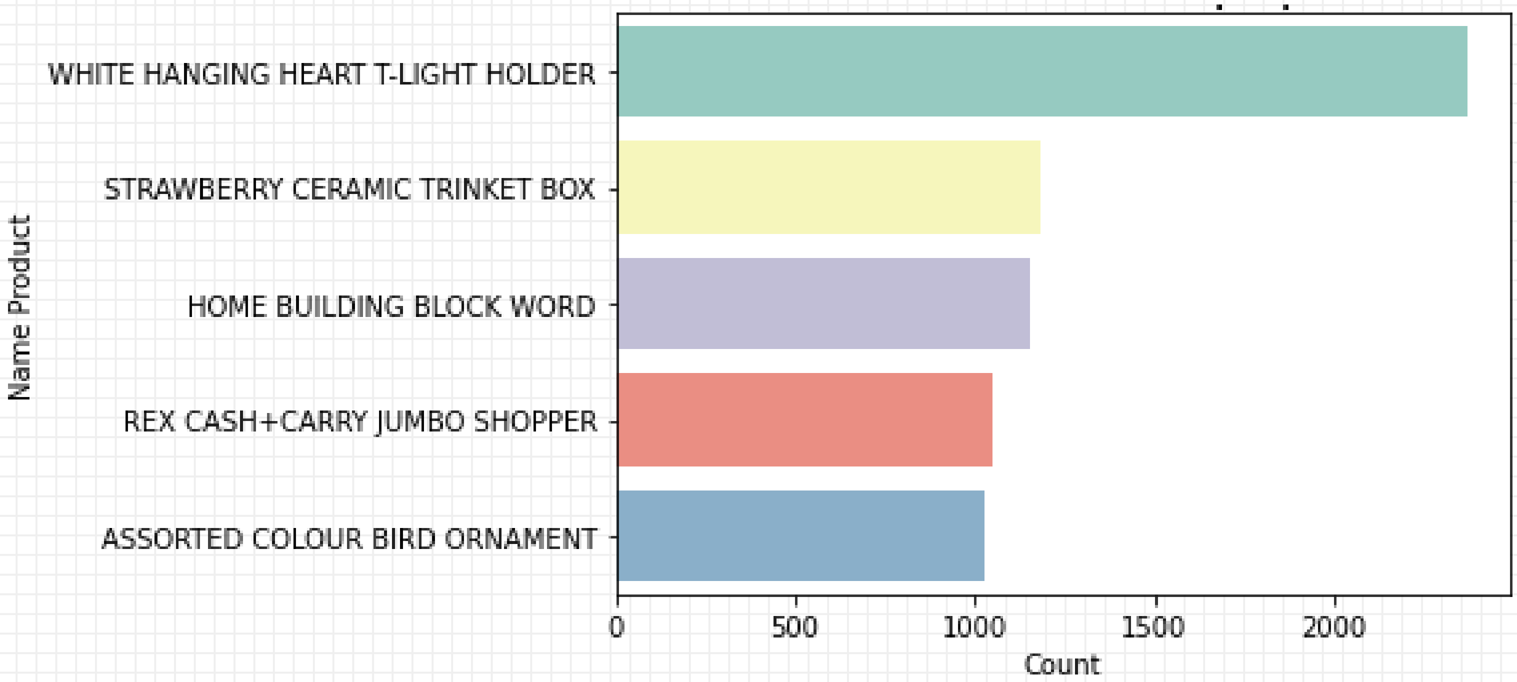


Most consumers make transactions on Thursday, which is around 19,5% of the total daily transactions. Business teams can increase sales by promoting new products to customers on Thursday.



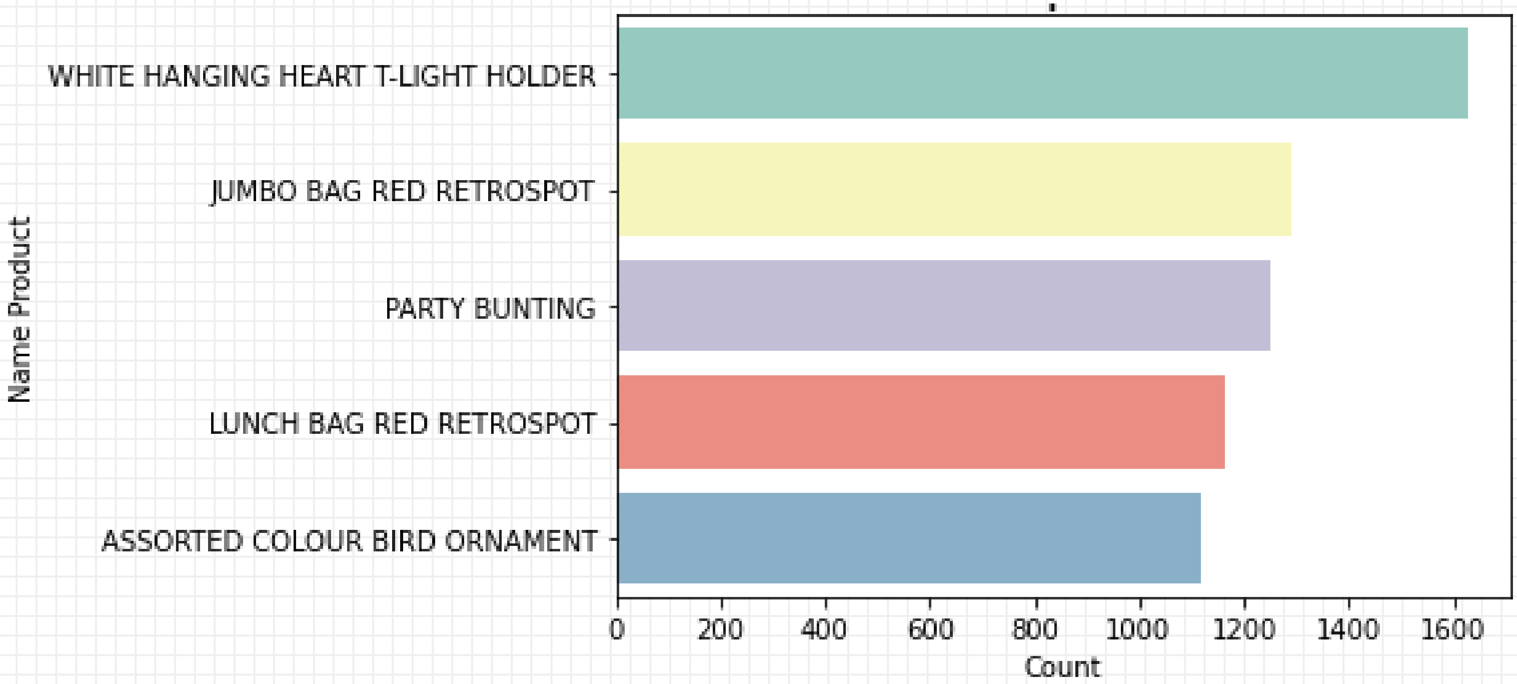
Most consumers order the products at 12 AM with a transaction amount of 18,2% of the total daily transactions. Business teams can increase sales by promoting new products to customers at 12 A.M

THE MOST POPULAR PRODUCT IN 2009 - 2010



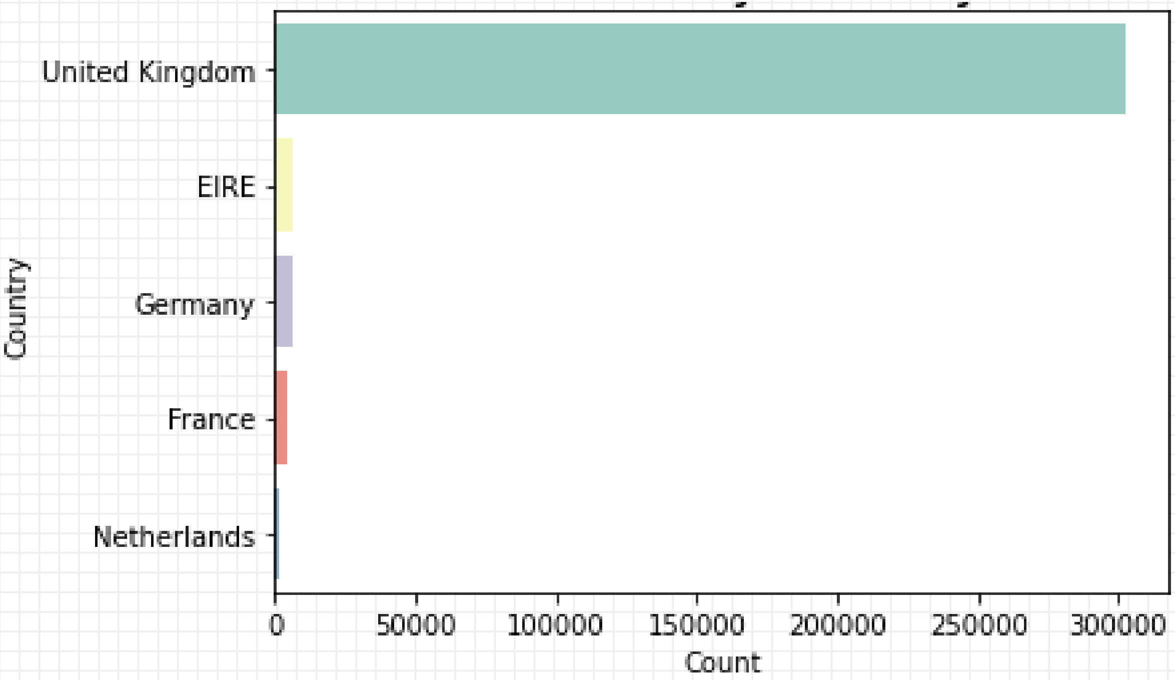
White Hanging Heart T-Light Holder became the product that was most in-demand by consumers in 2010. The number of purchases of White Hanging Heart T-Light Holder reached 2369 units in 2010.The business team can provide special discounts from this product to attract more users.

THE MOST POPULAR PRODUCT IN 2010 - 2011



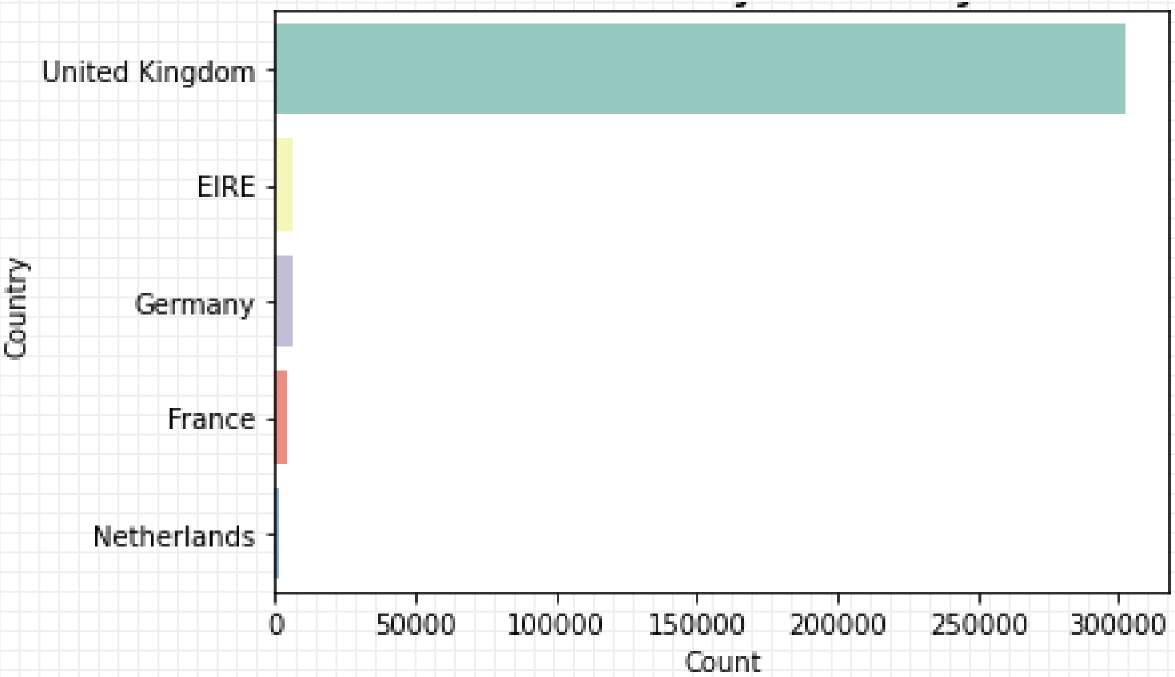
Same as in 2010, White Hanging Heart T-Light Holder became the product that was most in-demand by consumers in 2011. The number of purchases of White Hanging Heart T-Light reached 1625 units in 2011.The business team can provide special discounts from this product to attract more users.

THE MOST CUSTOMERS BY COUNTRY IN 2009 - 2010



The United Kingdom became the city with the highest number of customers in 2010. The total number of customers in United Kingdom reached 302776 (91.71%) customers in 2010. The business team can focus on promotions in the United Kingdom to increase sales.

THE MOST CUSTOMERS BY COUNTRY IN 2010 - 2011



The United Kingdom became the city with the highest number of customers in 2011. The total number of customers in United Kingdom reached 286683 (90%) customers in 2011. The business team can focus on promotions in the United Kingdom to increase sales.

Recency, Frequency, Monetary Value (RFM) Analysis

Recency, Frequency, Monetary Value (RFM) analysis method is a method of customer analysis and segmentation based on customer habits.

The variables used to perform RFM analysis are:

- Recency : How recently the customer made a transaction.
- Frequency : How often customers make transactions
- Monetary : How many transactions the customer has made

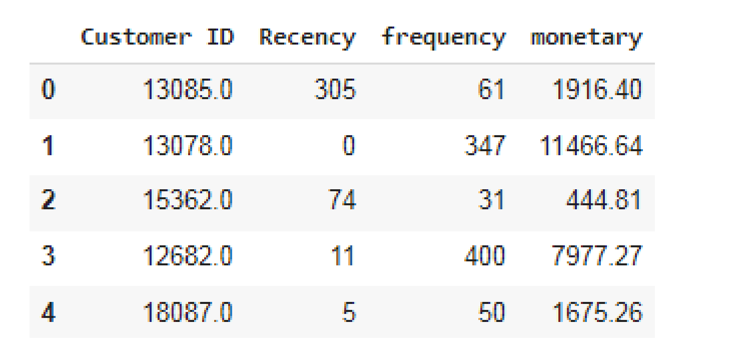
In this case, the dataset contains transaction data from 01/12/2009 to 01/12/2011, so the RFM Value is treated as follows:

- Recency : The difference between the last day the customer made a transaction and the day he did the analysis. In this case, the day of analysis uses the data of the last day of the transaction.
- Frequency : The number of transactions made by customers from 01/12/2009 to 01/12/2011.
- Monetary : Total order amount issued by customers from 01/12/2009 to 01/12/2011.

Recency, Frequency, Monetary Value (RFM) Analysis

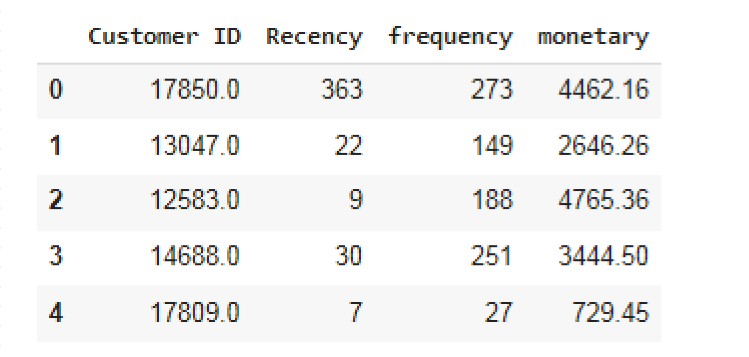
Here are the **steps** in RFM analysis:

1. Calculate RFM Value



	Customer ID	Recency	frequency	monetary
0	13085.0	305	61	1916.40
1	13078.0	0	347	11466.64
2	15362.0	74	31	444.81
3	12682.0	11	400	7977.27
4	18087.0	5	50	1675.26

RFM Value in 2009 - 2010



	Customer ID	Recency	frequency	monetary
0	17850.0	363	273	4462.16
1	13047.0	22	149	2646.26
2	12583.0	9	188	4765.36
3	14688.0	30	251	3444.50
4	17809.0	7	27	729.45

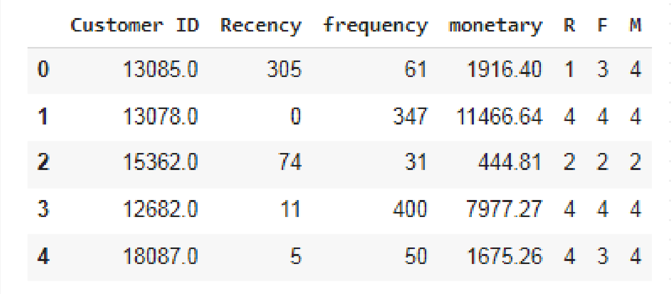
RFM Value in 2010 - 2011

2. Calculate RFM Score

The calculation of the individual RFM Score can be done using the Quartile statistical method. The steps is:

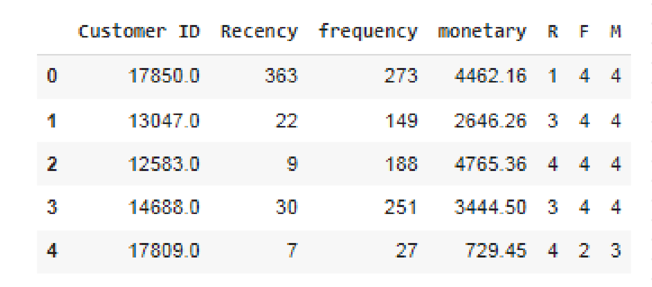
1. Split the metrics into segments using quantiles.
2. Assign a score from 1 to 4 to Recency, Frequency and Monetary.
3. Four is the best/highest value, and one is the lowest/worst value.

Recency, Frequency, Monetary Value (RFM) Analysis



	Customer ID	Recency	frequency	monetary	R	F	M
0	13085.0	305	61	1916.40	1	3	4
1	13078.0	0	347	11466.64	4	4	4
2	15362.0	74	31	444.81	2	2	2
3	12682.0	11	400	7977.27	4	4	4
4	18087.0	5	50	1675.26	4	3	4

RFM Score in 2009 - 2010

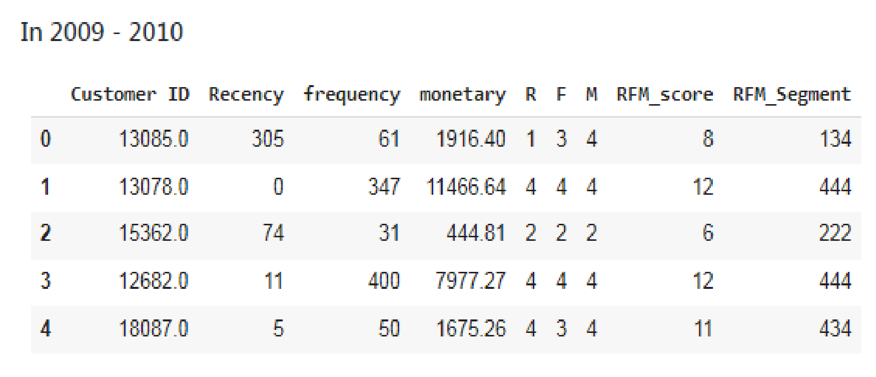


	Customer ID	Recency	frequency	monetary	R	F	M
0	17850.0	363	273	4462.16	1	4	4
1	13047.0	22	149	2646.26	3	4	4
2	12583.0	9	188	4765.36	4	4	4
3	14688.0	30	251	3444.50	3	4	4
4	17809.0	7	27	729.45	4	2	3

RFM Score in 2010 - 2011

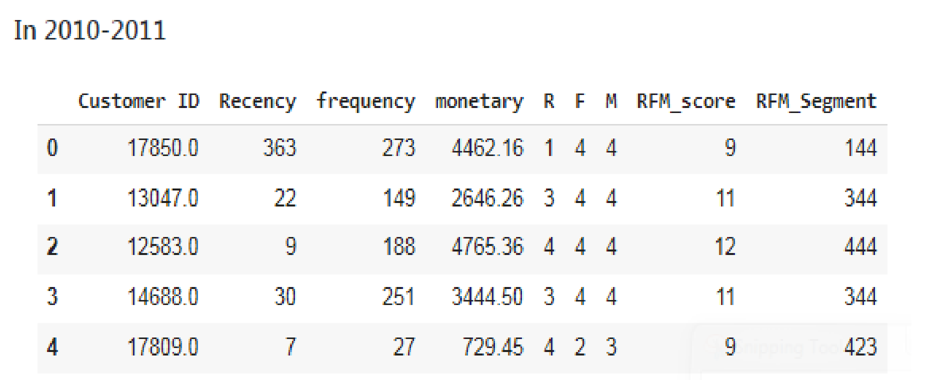
3. Calculate the total RFM score

A total RFM score is calculated simply by combining individual RFM score numbers.



In 2009 - 2010

	Customer ID	Recency	frequency	monetary	R	F	M	RFM_score	RFM_Segment
0	13085.0	305	61	1916.40	1	3	4	8	134
1	13078.0	0	347	11466.64	4	4	4	12	444
2	15362.0	74	31	444.81	2	2	2	6	222
3	12682.0	11	400	7977.27	4	4	4	12	444
4	18087.0	5	50	1675.26	4	3	4	11	434



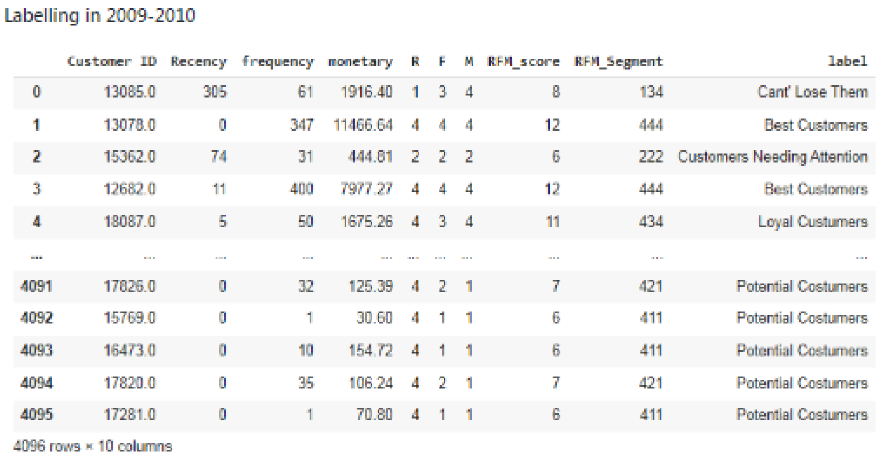
In 2010-2011

	Customer ID	Recency	frequency	monetary	R	F	M	RFM_score	RFM_Segment
0	17850.0	363	273	4462.16	1	4	4	9	144
1	13047.0	22	149	2646.26	3	4	4	11	344
2	12583.0	9	188	4765.36	4	4	4	12	444
3	14688.0	30	251	3444.50	3	4	4	11	344
4	17809.0	7	27	729.45	4	2	3	9	423

Recency, Frequency, Monetary Value (RFM) Analysis

4. Labelling

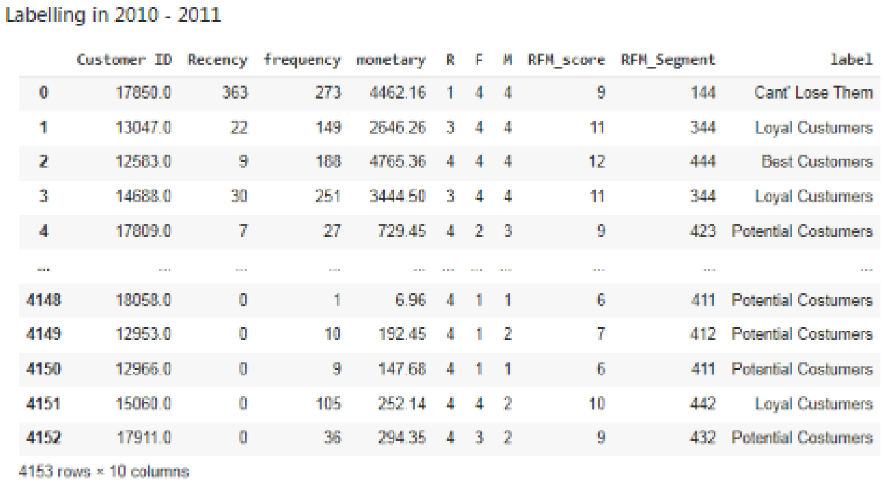
Segment	RFM Score Range		
	R	F	M
Best Customers	4	4	4
Loyal Customers	3-4	3-4	3-4
Potential Customers	3-4	1-3	1-3
Customers Needing Attention	1-2	1-3	1-3
Cant' Lose Them	1-2	4	4
At Risk Customers	1-2	2-4	2-4
Lost Customers	1	1	1



Labelling in 2009-2010

	Customer ID	Recency	frequency	monetary	R	F	M	RFM_score	RFM_Segment	label
0	13085.0	305	61	1916.40	1	3	4	8	134	Cant' Lose Them
1	13078.0	0	347	11466.64	4	4	4	12	444	Best Customers
2	15362.0	74	31	444.81	2	2	2	6	222	Customers Needing Attention
3	12682.0	11	400	7977.27	4	4	4	12	444	Best Customers
4	18087.0	5	50	1675.26	4	3	4	11	434	Loyal Custumers
...	...	...	...	...	...	...	...	...	...	...
4091	17826.0	0	32	125.39	4	2	1	7	421	Potential Costumers
4092	15769.0	0	1	30.60	4	1	1	6	411	Potential Costumers
4093	16473.0	0	10	154.72	4	1	1	6	411	Potential Costumers
4094	17820.0	0	35	106.24	4	2	1	7	421	Potential Costumers
4095	17281.0	0	1	70.80	4	1	1	6	411	Potential Costumers

4096 rows × 10 columns

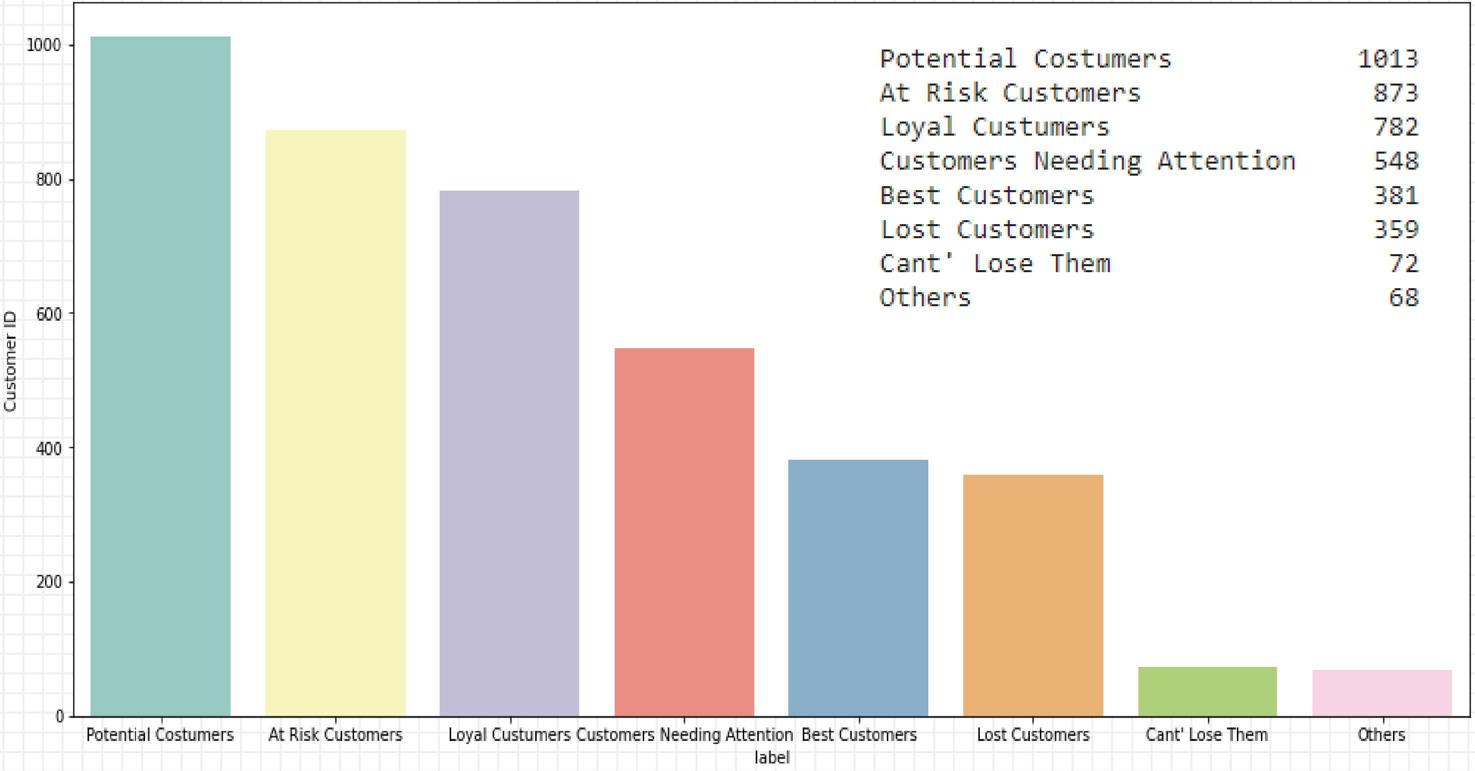


Labelling in 2010 - 2011

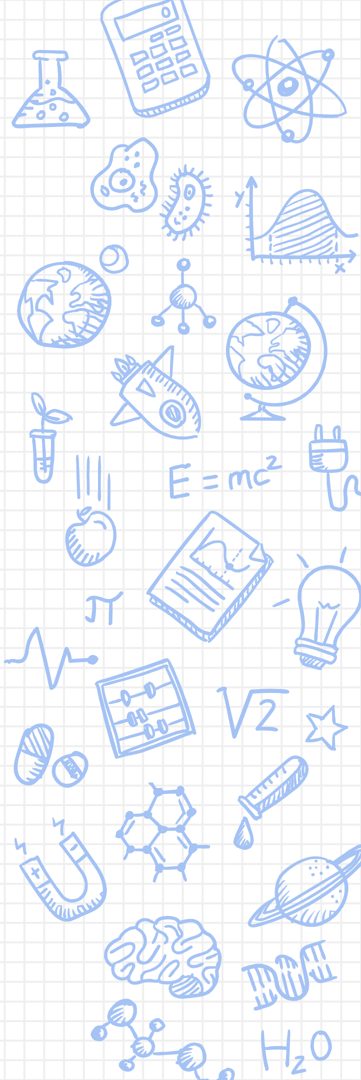
	Customer ID	Recency	frequency	monetary	R	F	M	RFM_score	RFM_Segment	label
0	17850.0	363	273	4462.16	1	4	4	9	144	Cant' Lose Them
1	13047.0	22	149	2646.26	3	4	4	11	344	Loyal Custumers
2	12583.0	9	188	4765.36	4	4	4	12	444	Best Customers
3	14688.0	30	251	3444.50	3	4	4	11	344	Loyal Custumers
4	17809.0	7	27	729.45	4	2	3	9	423	Potential Costumers
...	...	...	...	...	...	...	...	...	...	...
4148	18058.0	0	1	6.96	4	1	1	6	411	Potential Costumers
4149	12953.0	0	10	192.45	4	1	2	7	412	Potential Costumers
4150	12966.0	0	9	147.68	4	1	1	6	411	Potential Costumers
4151	15060.0	0	105	252.14	4	4	2	10	442	Loyal Custumers
4152	17911.0	0	36	294.35	4	3	2	9	432	Potential Costumers

4153 rows × 10 columns

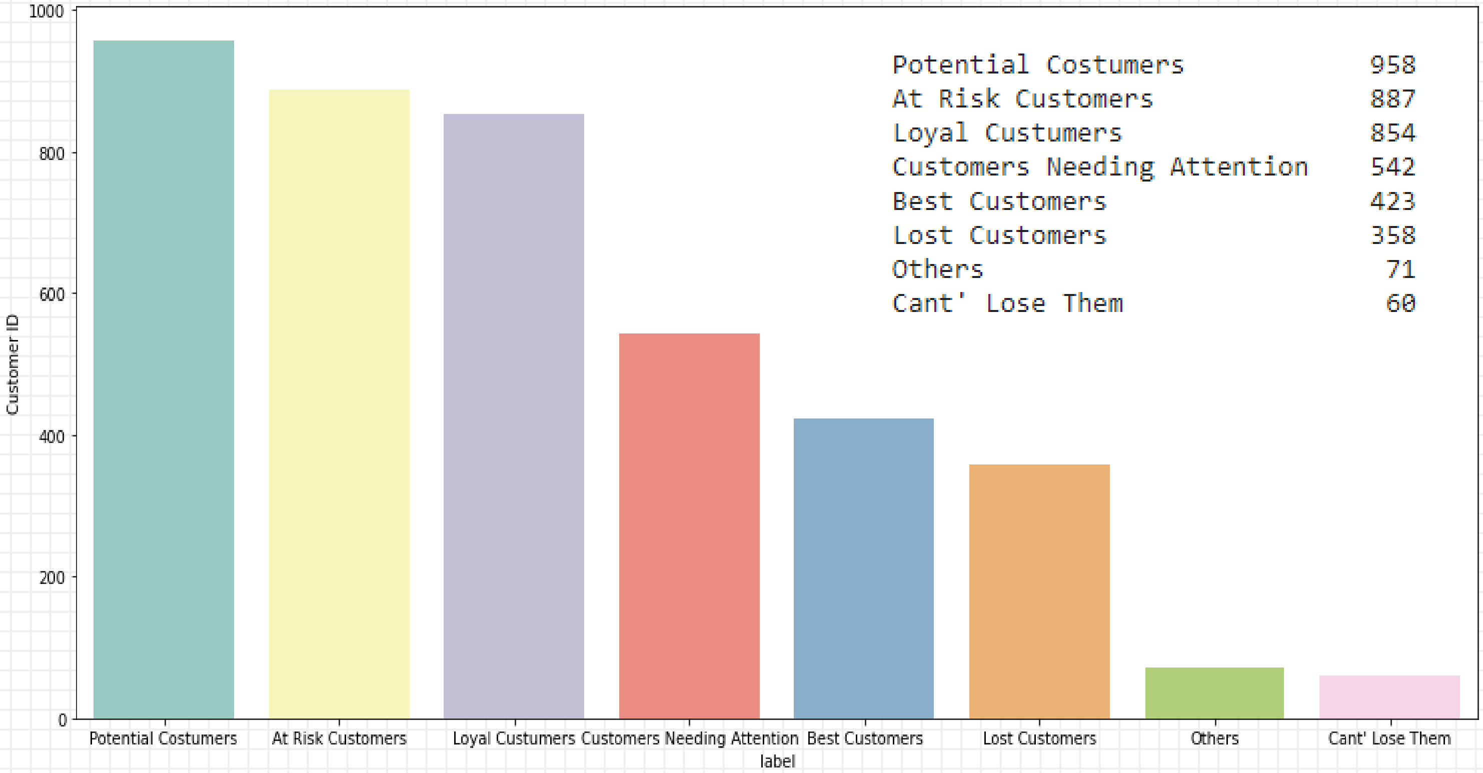
RFM Analysis in 2009 - 2010



```
Potential Costumers                 1013
At Risk Customers                    873
Loyal Custumers                      782
Customers Needing Attention          548
Best Customers                       381
Lost Customers                       359
Cant' Lose Them                       72
Others                                68
```



RFM Analysis in 2010 - 2011



```
Potential Costumers             958
At Risk Customers               887
Loyal Custumers                 854
Customers Needing Attention     542
Best Customers                  423
Lost Customers                  358
Others                           71
Cant' Lose Them                  60
```

K-Means Clustering

K-Means clustering algorithm is an unsupervised machine learning algorithm that uses multiple iterations to segment the unlabeled data points into K different clusters in a way such that each data point belongs to only a single group that has similar properties.

K-means gives the best result under the following conditions:

1. Data’s distribution is not skewed.

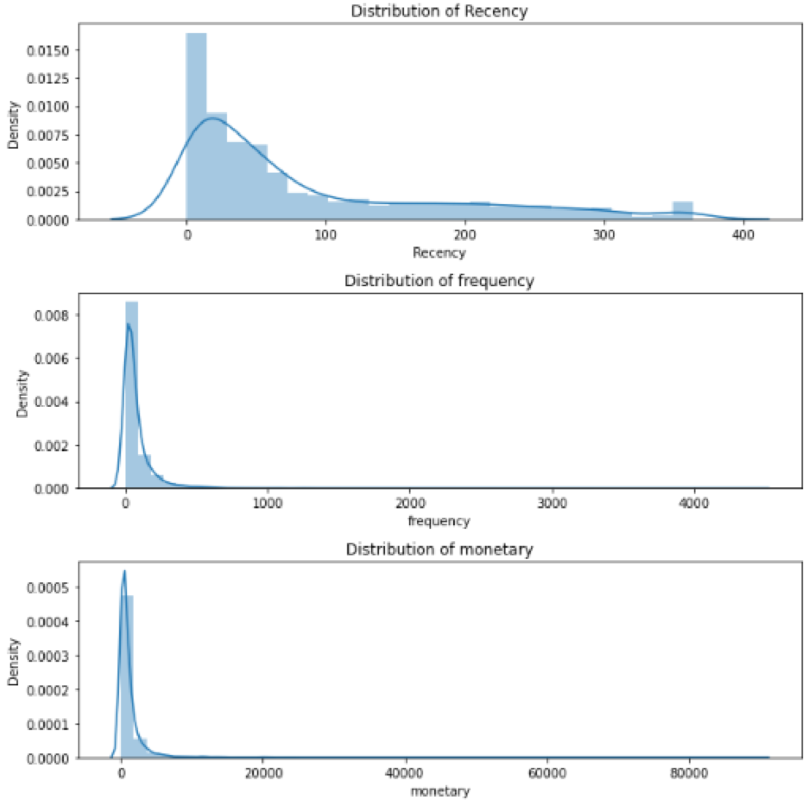
2. Data is standardised

   (i.e. mean of 0 and standard deviation of 1).

K-Means Clustering

K-means gives the best result under the following conditions:

1. Data’s distribution is not skewed.

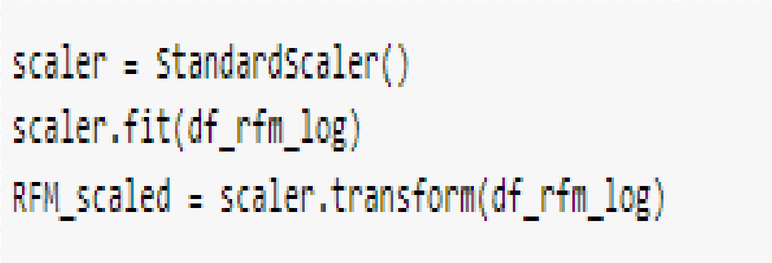

Distribution of Recency

Distribution of frequency

Distribution of monetary

The data is highly skewed,therefore we will perform log transformations to reduce the skewness of each variable.I add a small constant as log transformation demands all the values to be positive.

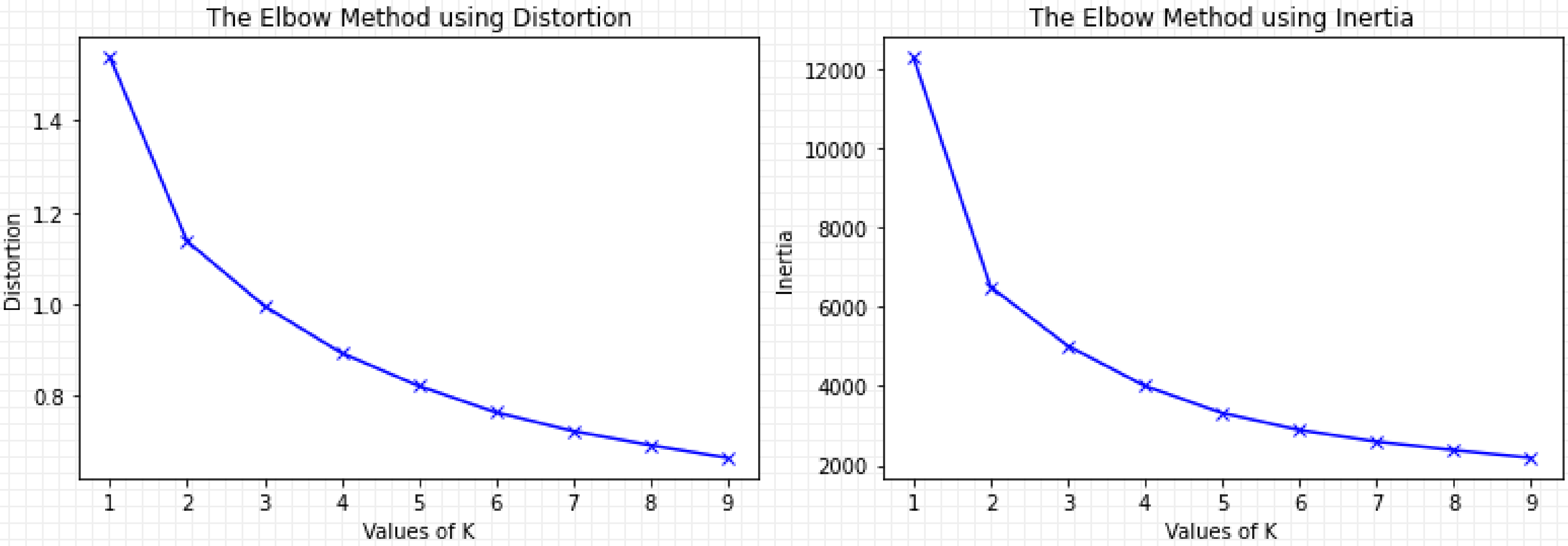
2. Data is standardised
   (i.e. mean of 0 and standard deviation of 1).



```
scaler = StandardScaler()
scaler.fit(df_rfm_log)
RFM_scaled = scaler.transform(df_rfm_log)
```

DATASET 1 (2009 – 2010)

FINDING THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD

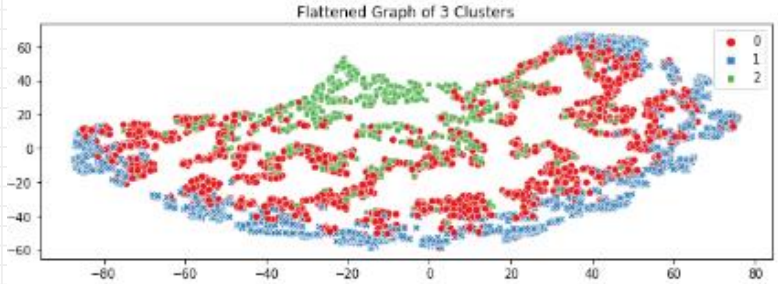

The Elbow Method using Distortion

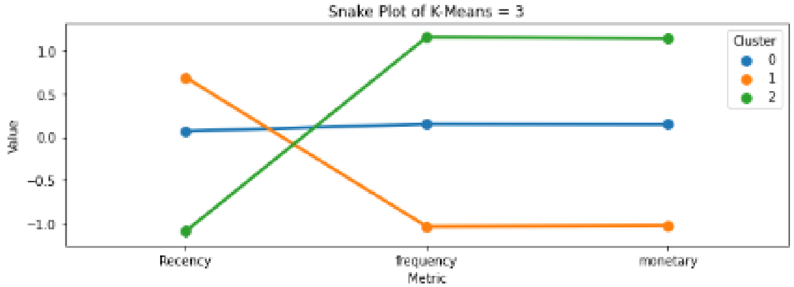
The Elbow Method using Inertia

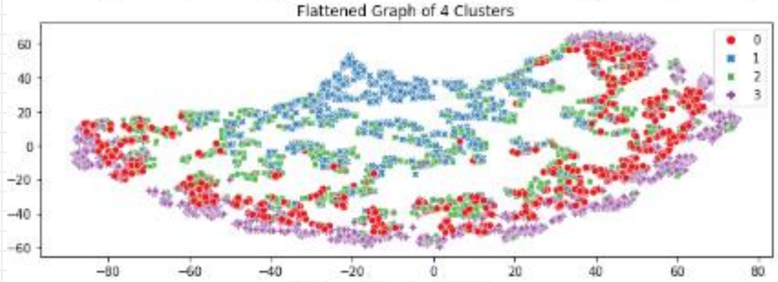
The cluster value where this decrease in distortion value and inertia value becomes constant can be chosen as the right cluster value for our data. Looking at the above elbow curve, we can choose any number of clusters between 3 to 5.

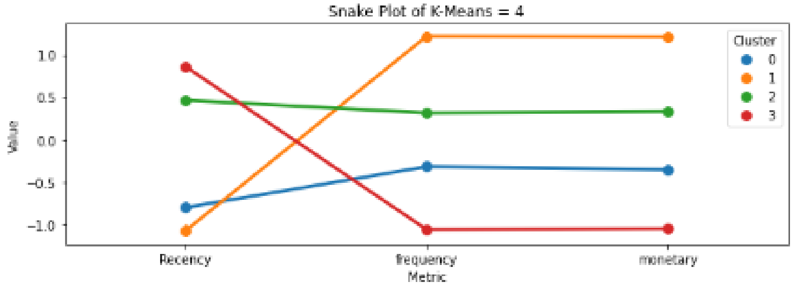
DATASET 1 (2009 – 2010)

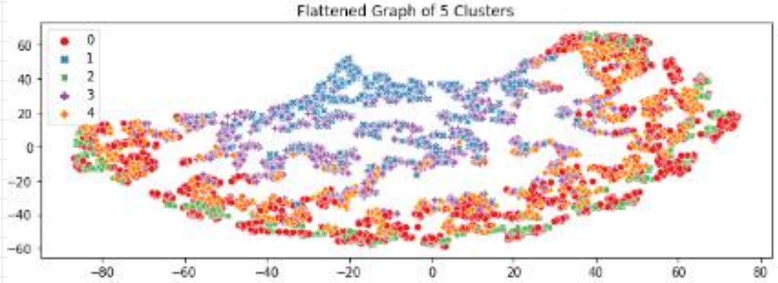
FINDING THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD

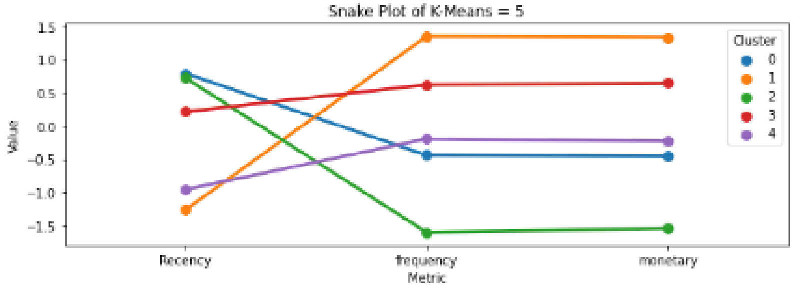

Flattened Graph of 3 Clusters


Snake Plot of K-Means = 3


Flattened Graph of 4 Clusters


Snake Plot of K-Means = 4

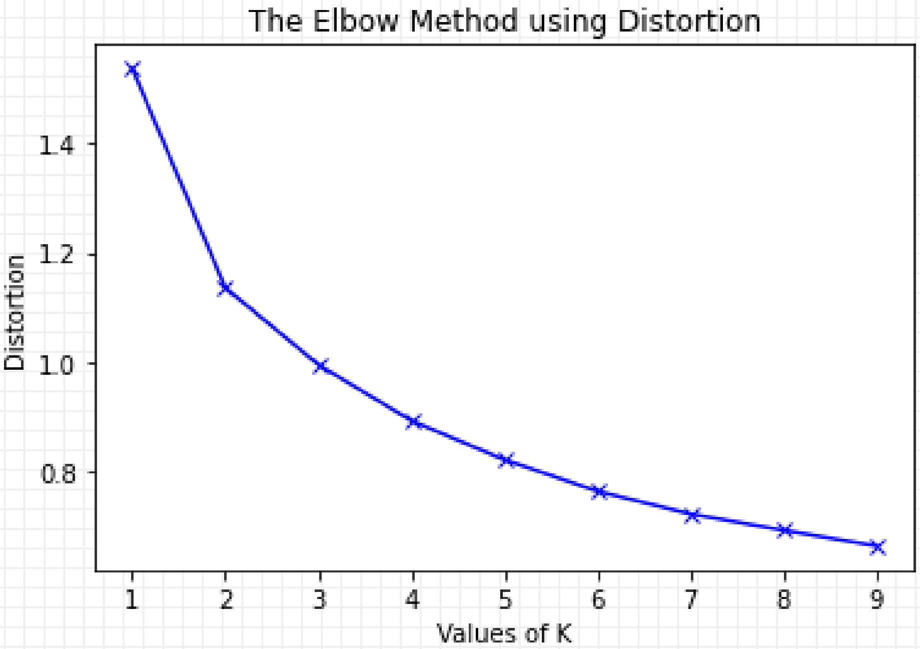

Flattened Graph of 5 Clusters

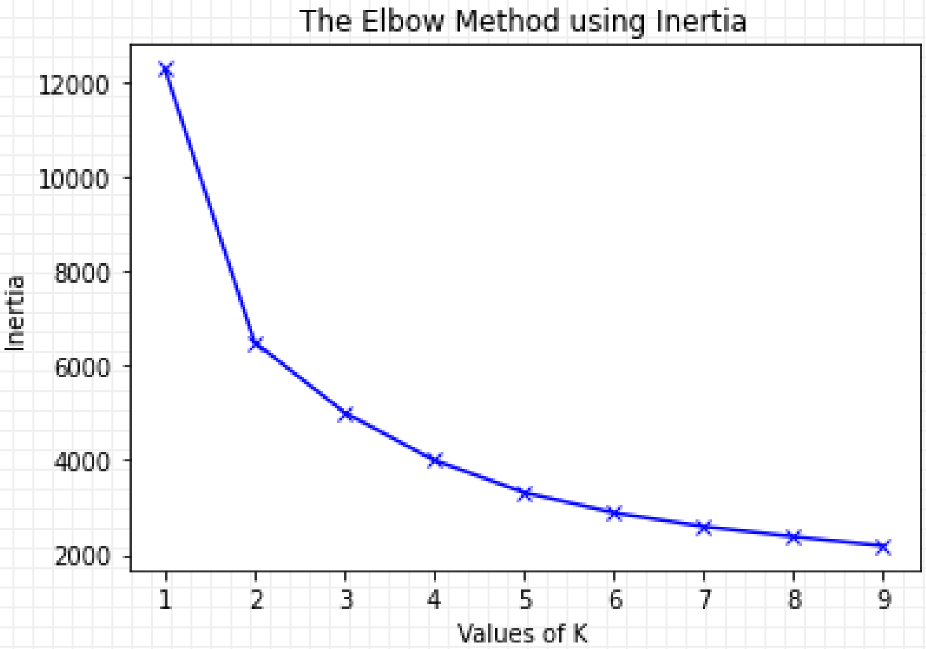

Snake Plot of K-Means = 5

From the flattened graphs and the snake plots it is evident that having a cluster value of 4, segments our customers well. We could also go for higher number of clusters, it completely depends on how the company wants to segment their customers.

DATASET 2 (2010 – 2011)

FINDING THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD

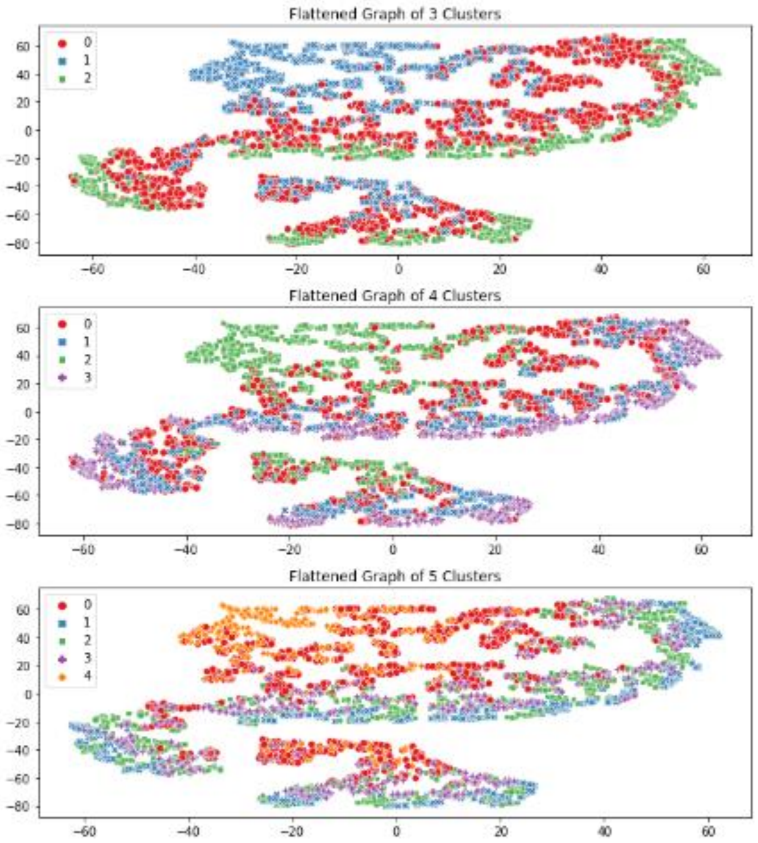




The cluster value where this decrease in distortion value and inertia value becomes constant can be chosen as the right cluster value for our data. Looking at the above elbow curve, we can choose any number of clusters between 3 to 5.

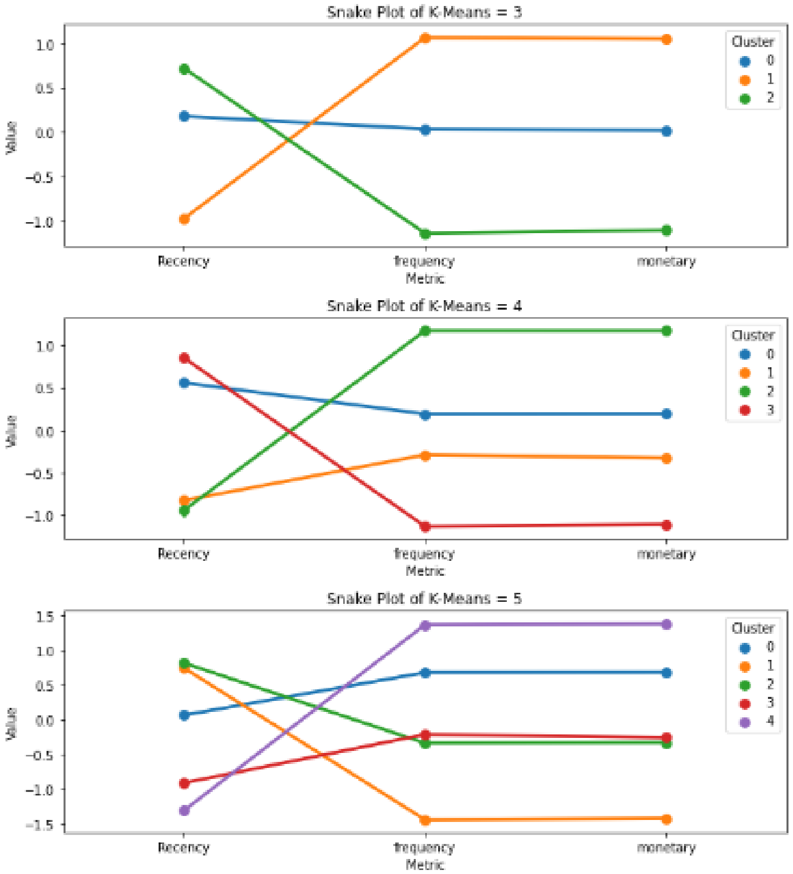
DATASET 2 (2010 – 2011)

FINDING THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD


Flattened Graph of 3 Clusters

Flattened Graph of 4 Clusters

Flattened Graph of 5 Clusters


Snake Plot of K-Means = 3

Snake Plot of K-Means = 4

Snake Plot of K-Means = 5

From the flattened graphs and the snake plots it is evident that having a cluster value of 4, segments our customers well. We could also go for higher number of clusters, it completely depends on how the company wants to segment their customers.

DATASET 1 (2009 – 2010)

EVALUATING MODEL

1. Davies Bouldin Score

Davies Bouldin Score is a metric for evaluating clustering algorithms. The smaller Davies Bouldin Score is the more optimal the cluster.

K-Means Cluster	Davies Bouldin Score
3	1.1010195514551004
4	0.9915658231949979 (smaller)
5	0.9947174153325206

K-Means with 4 clusters has lowest davies bouldin score than other cluster. Therefore the optimum cluster is 4.

2. Silhouetter Score

Silhoutter Score is a metric for evaluating clustering algorithms. The higher Silhouter Score is the more optimal the cluster.

K-Means Cluster	Silhouetter Score
3	0.302
4	0.3146 (higher)
5	0.30197

K-Means with 4 clusters has higher silhouette score than other cluster. Therefore the optimum cluster is 4.

DATASET 2 (2010 – 2011)

EVALUATING MODEL

1. Davies Bouldin Score

Davies Bouldin Score is a metric for evaluating clustering algorithms. The smaller Davies Bouldin Score is the more optimal the cluster.

K-Means Cluster	Davies Bouldin Score
3	1.088402821719592
4	0.9887352361190681 (smaller)
5	1.013345257005478

K-Means with 4 clusters has lowest davies bouldin score than other cluster. Therefore the optimum cluster is 4.

2. Silhouetter Score

Silhoutter Score is a metric for evaluating clustering algorithms. The higher Silhouter Score is the more optimal the cluster.

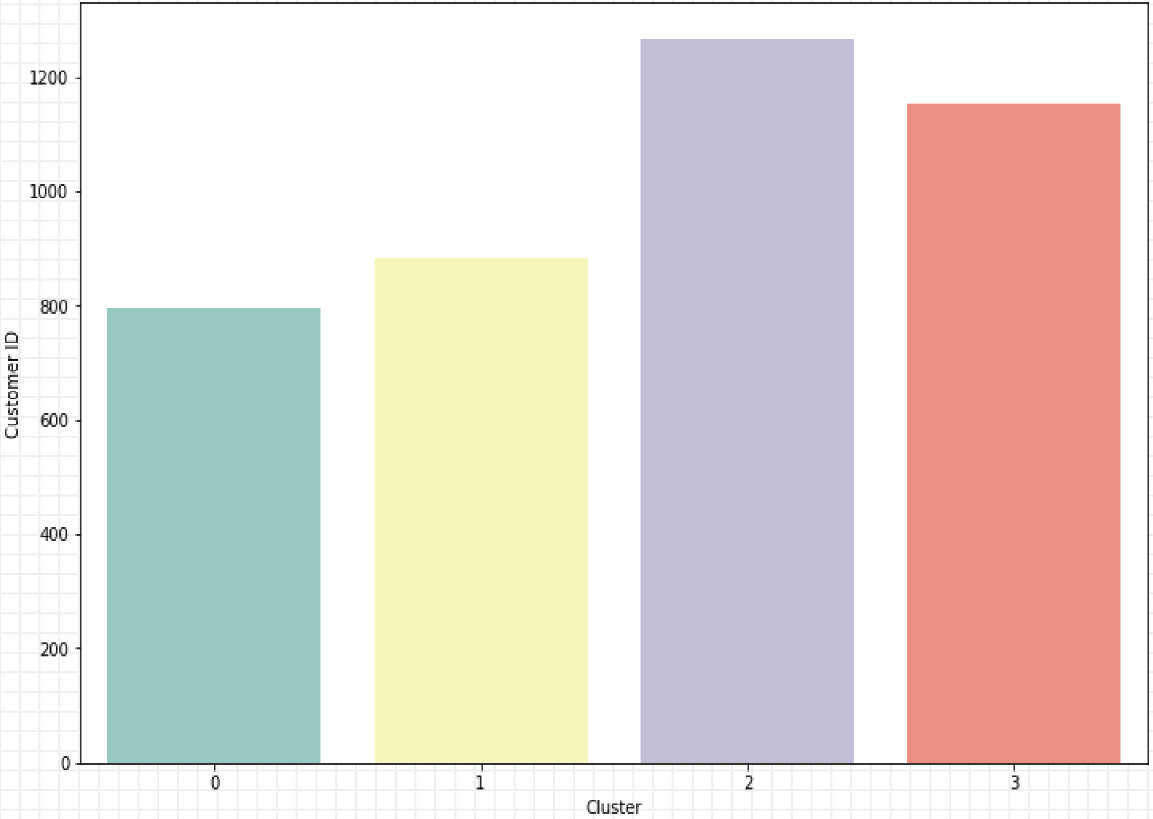
K-Means Cluster	Silhouetter Score
3	0.306
4	0.3187 (higher)
5	0.29514

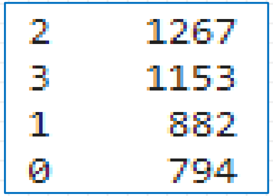
K-Means with 4 clusters has higher silhouette score than other cluster. Therefore the optimum cluster is 4.

DATASET 1 (2009 – 2010)

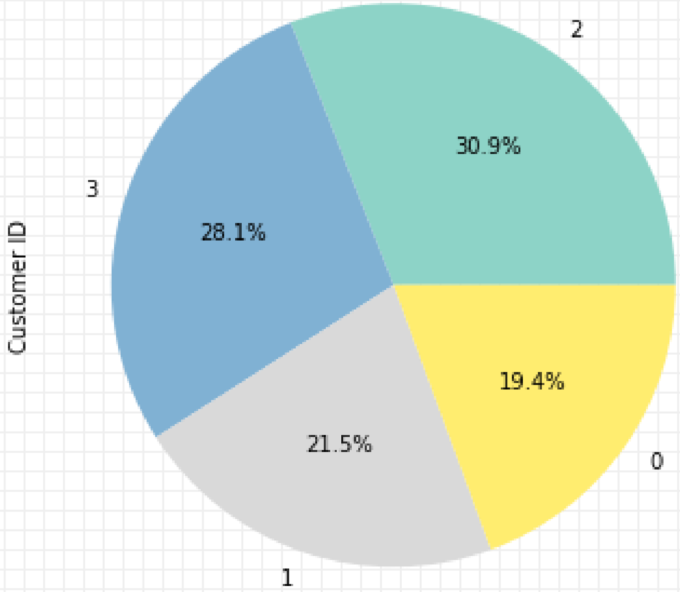
INTERPRETATION OF THE CLUSTERS FORMED USING K-MEANS

K – Means 4 Clusters





```
2      1267
3      1153
1       882
0       794
```



DATASET 1 (2009 – 2010)

INTERPRETATION OF THE CLUSTERS FORMED USING K-MEANS

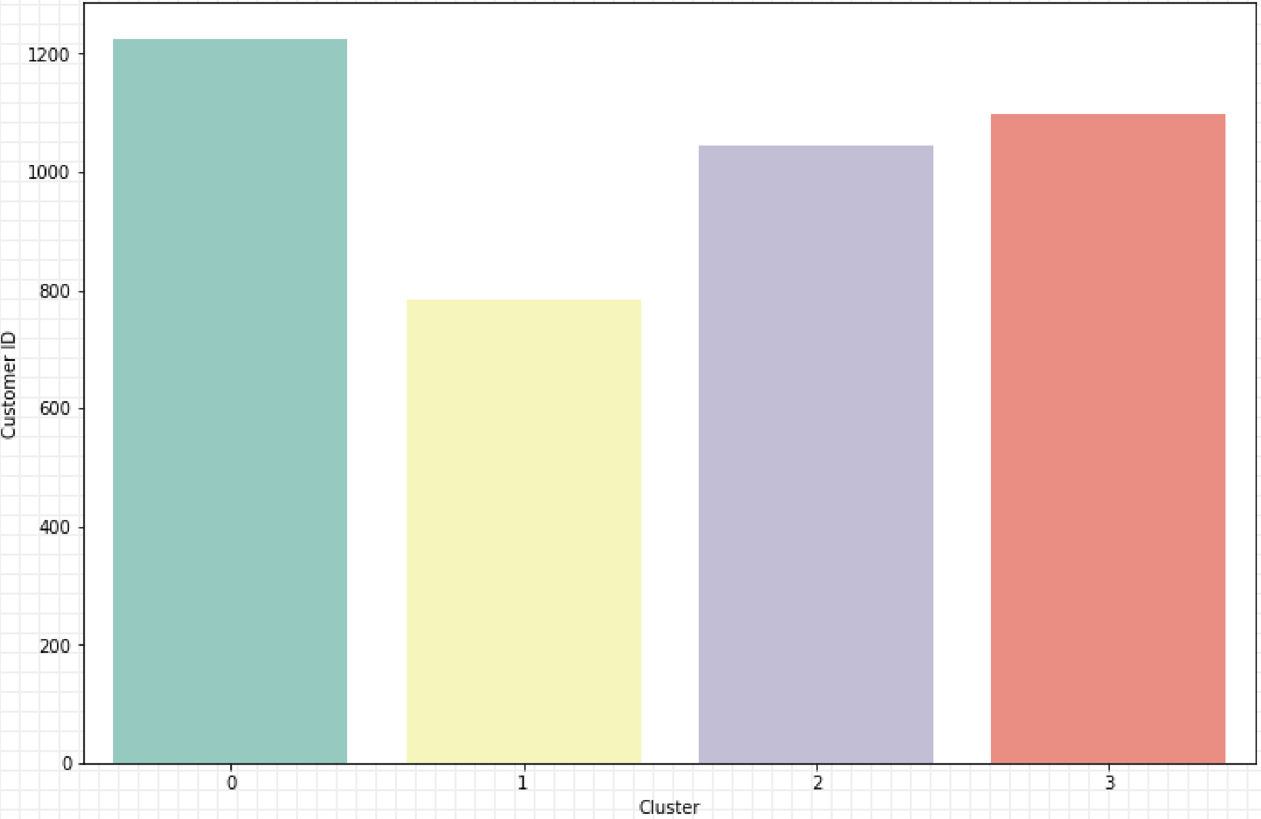
K – Means 4 Clusters

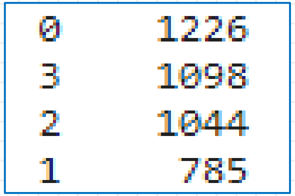
Cluster	RFM Score			Segment	Explanation Segmentation	% Customers
	R	F	M			
0	3	2	2	Potential Customers	Recent customers, but spent a good amount and bought more than once.	19.4%
1	4	4	4	Best Customers	Bought recently, buy often and spend the most.	21.5%
2	2	3	3	Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently	30.9%
3	1	1	1	Lost Customers	Last purchase was long back, low spenders and low number of orders.	28.1%

DATASET 2 (2010 – 2011)

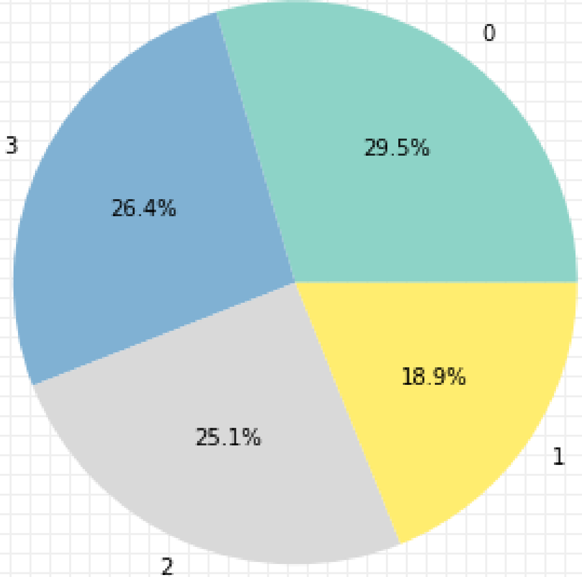
INTERPRETATION OF THE CLUSTERS FORMED USING K-MEANS

K – Means 4 Clusters





```
0     1226
3     1098
2     1044
1      785
```



DATASET 2 (2010 – 2011)

INTERPRETATION OF THE CLUSTERS FORMED USING K-MEANS

K – Means 4 Clusters

Cluster	RFM Score			Segment	Explanation Segmentation	% Customers
	R	F	M			
0	3	4	4	Loyal Customers	Spend good money with us often and responsive to promotions.	29.5%
1	2	3	3	Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently	18.9%
2	4	2	2	Potential Customers	Recent customers, but spent a good amount and bought more than once.	25.1%
3	1	1	1	Lost Customers	Last purchase was long back, low spenders and low number of orders.	26.4%

DATASET 1 (2009 – 2010)

RECOMMENDATION

Based on the 4 clusters, we could formulate marketing strategies relevant to each cluster:

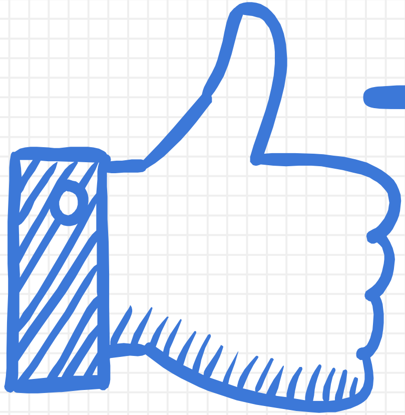
Cluster	Segment	Recommendation
0	Potential Customers	Offer membership / loyalty program and recommend other products.
1	Best Customers	Reward them, can be early adopters for new products. Will promote your brand, and ask for reviews.
2	Customers Needing Attention	Make limited time offers, recommend based on past purchases, share valuable resources, recommend popular products / renewals at discount, and reconnect with them.
3	Lost Customers	Revive interest with reach out campaign, ignore otherwise, offer other relevant products and special discounts, and recreate brand value.

DATASET 2 (2010 – 2011)

RECOMMENDATION

Based on the 4 clusters, we could formulate marketing strategies relevant to each cluster:

Cluster	Segment	Recommendation
0	Loyal Customers	Upsell higher value products, ask for reviews, and engage them.
1	Customers Needing Attention	Make limited time offers, recommend based on past purchases, share valuable resources, recommend popular products / renewals at discount, and reconnect with them.
2	Potential Customers	Offer membership / loyalty program and recommend other products.
3	Lost Customers	Revive interest with reach out campaign, ignore otherwise, offer other relevant products and special discounts, and recreate brand value.



THANKS!

https://www.linkedin.com/in/suciaulyaputri/

https://github.com/suciaulyaputri