Capstone Project - 3

Online Retail Customer Segmentation

ML Unsupervised Clustering

Individual Project:
Soumya Ranjan Mishra

Contents

- Problem Statement
- Data Summary
- Data Analysis
- Analysis Details
- RFM Table for Customer ID
- Data Preparation
- Implementing Various Clustering Algorithms
- RFM Table for Cluster ID
- Challenges
- Conclusions

Problem Statements

- Identify major customer segments on UK Based online retail dataset.
- Create RFM Table.

Data Summary

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with 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.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Basic Data Exploration

- The dataset has *541909* rows and 8 features(columns).

•Four categorical features 'InvoiceNo', 'StockCode', &

'Description', ' Country'.

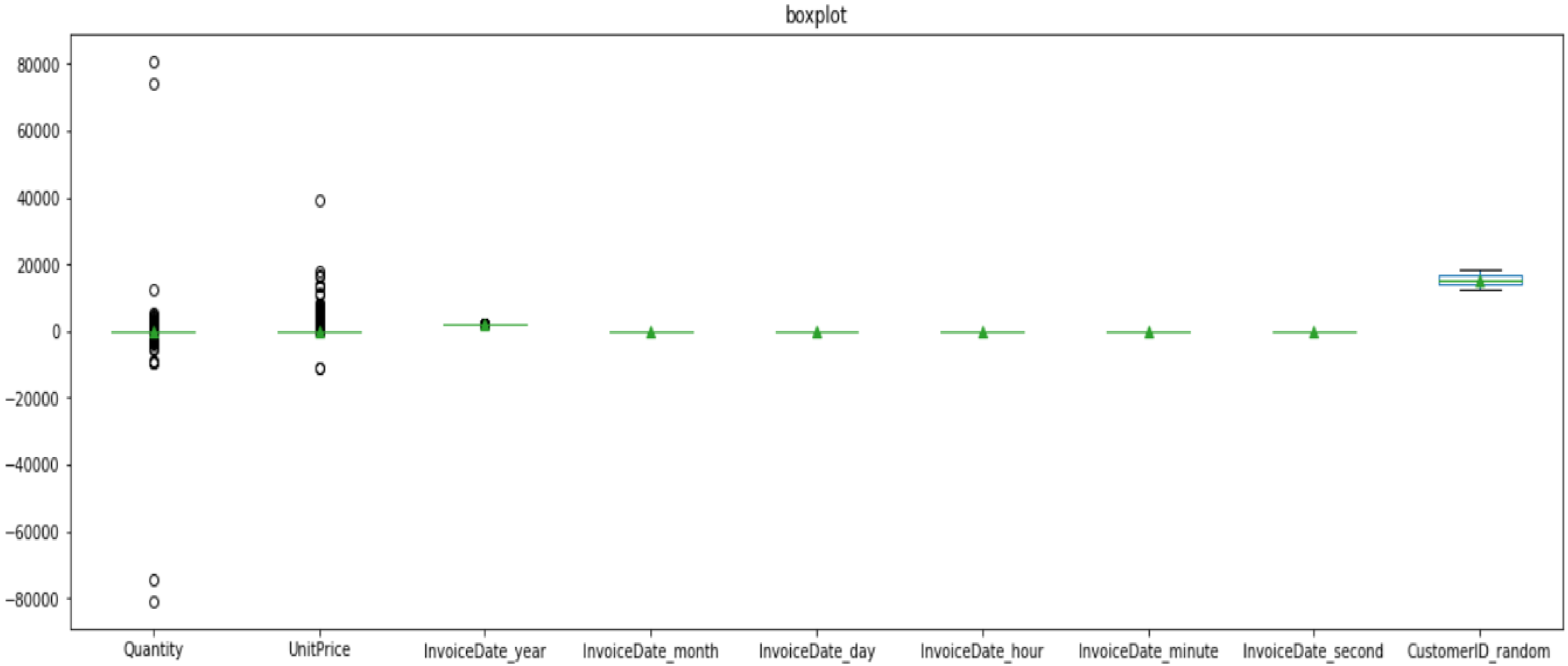
- One Datetime[ns] features 'InvoiceDate'.

- Outliers present only in "Quantity" & "UnitPrice"column.

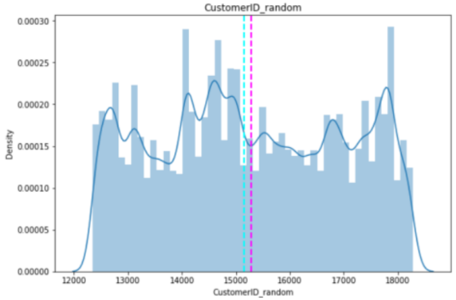
- Missing Values on *Description & CustomerID columns*.

- Duplicated values present.

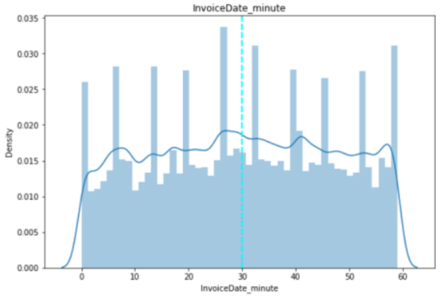
Outliers in the features



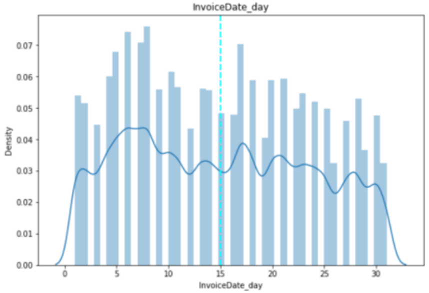
Mean Distribution of Features



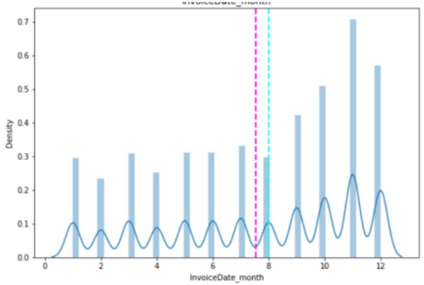
```
Skewness : 0.03468049946828383
Kurtosis : -1.179793510437348
```



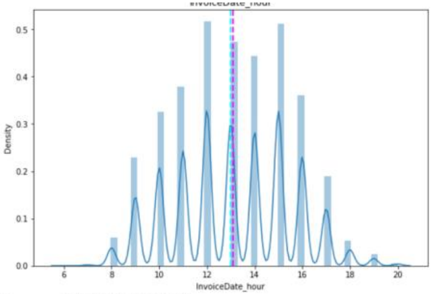
```
Skewness : -0.02534037213860687
Kurtosis : -1.136943576846362
```



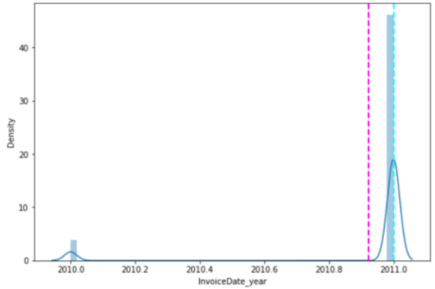
```
Skewness : 0.1425060392658862
Kurtosis : -1.1573667360751423
```



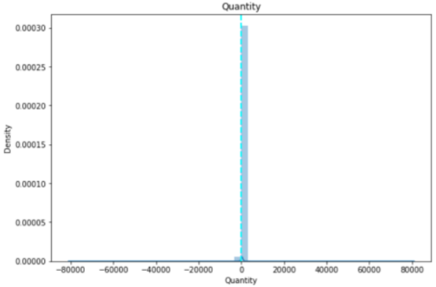
```
Skewness : -0.4112397637662981
Kurtosis : -1.121688905682585
```



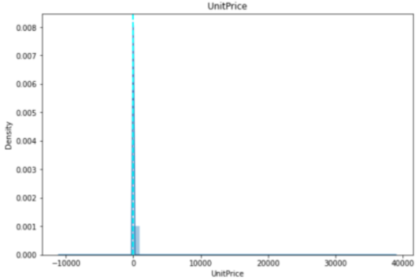
```
Skewness : 0.002850657774814659
Kurtosis : -0.6944981623468869
```



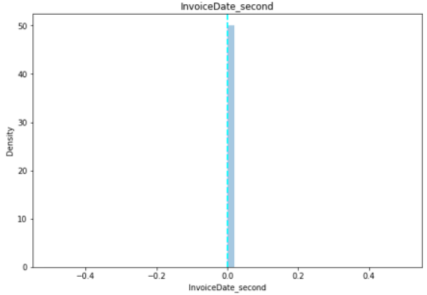
```
Skewness : -3.1413170786444153
Kurtosis : 7.867902311346793
```



```
Skewness : -0.2639207787329176
Kurtosis : 118645.54455234332
```

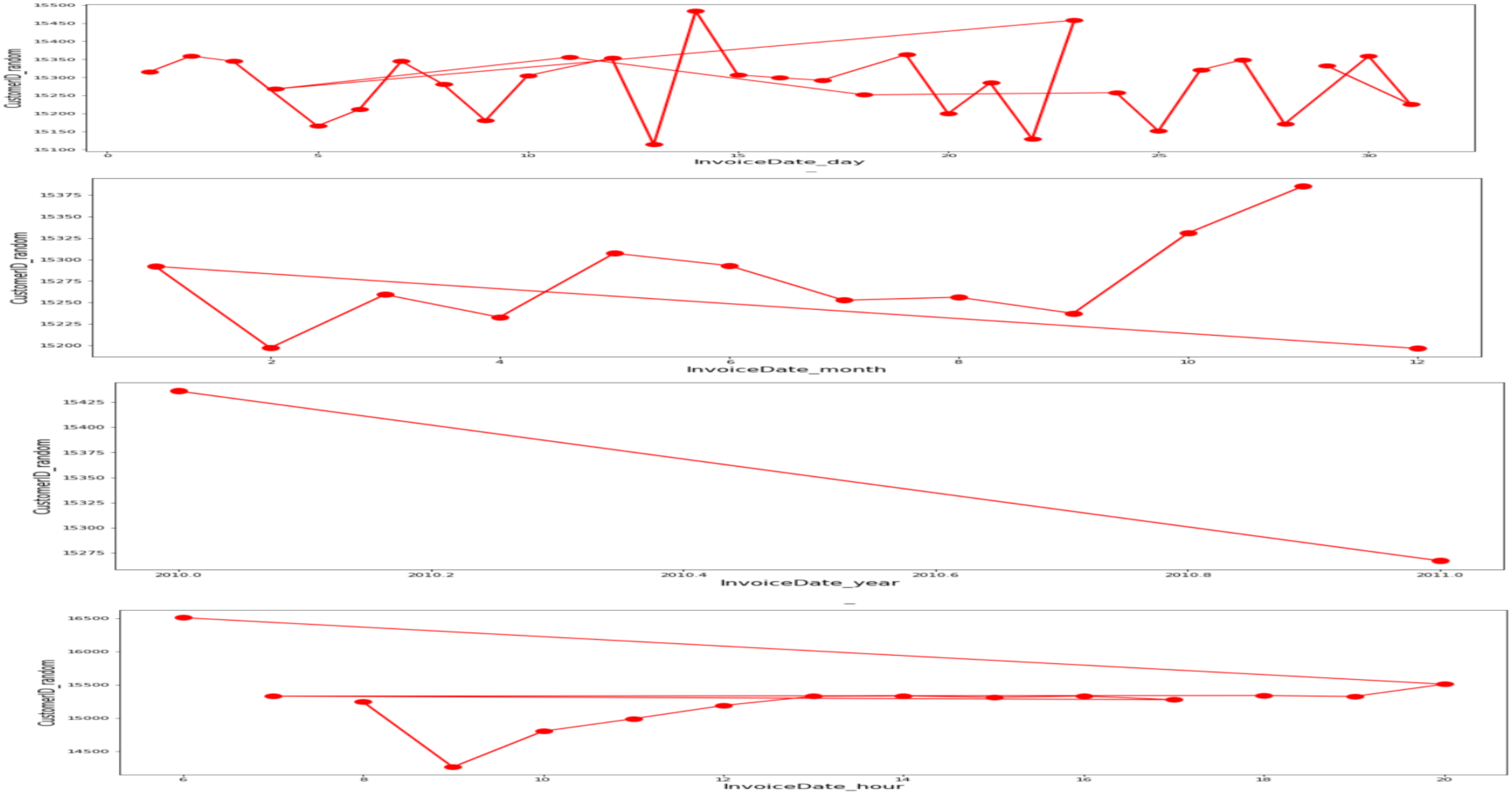


```
Skewness : 185.60044812920344
Kurtosis : 58433.135910585086
```

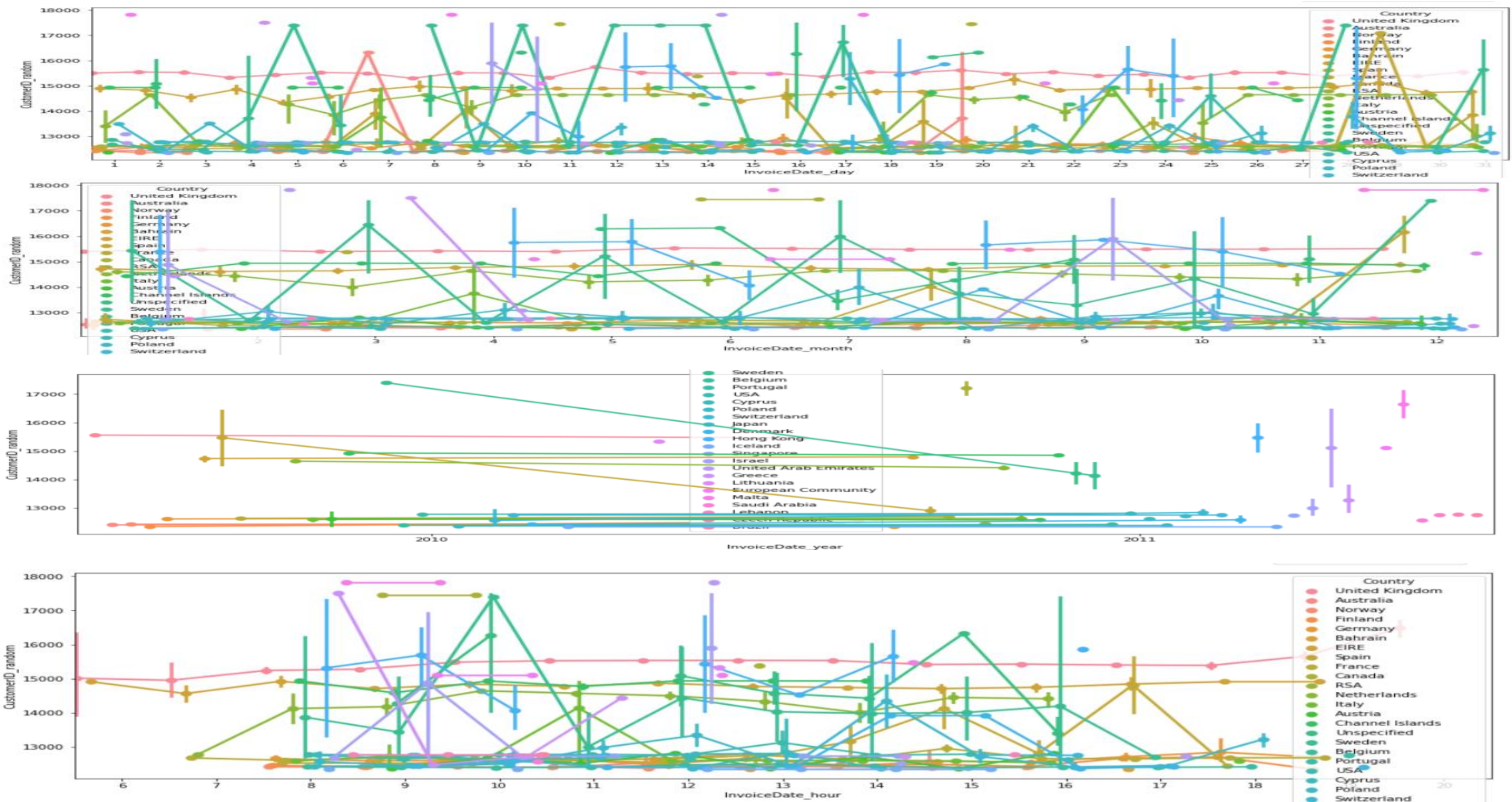


```
Skewness : 0
Kurtosis : 0
```

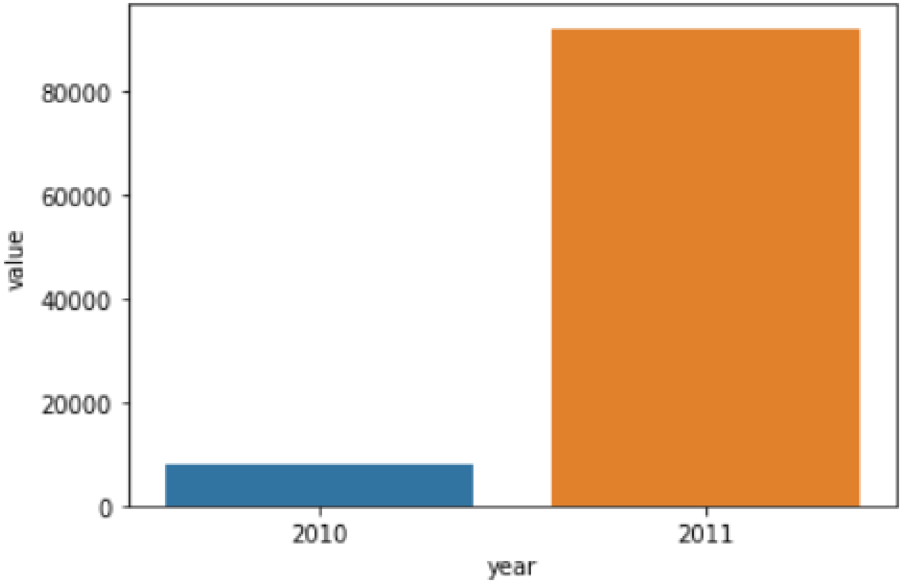
Spread over time



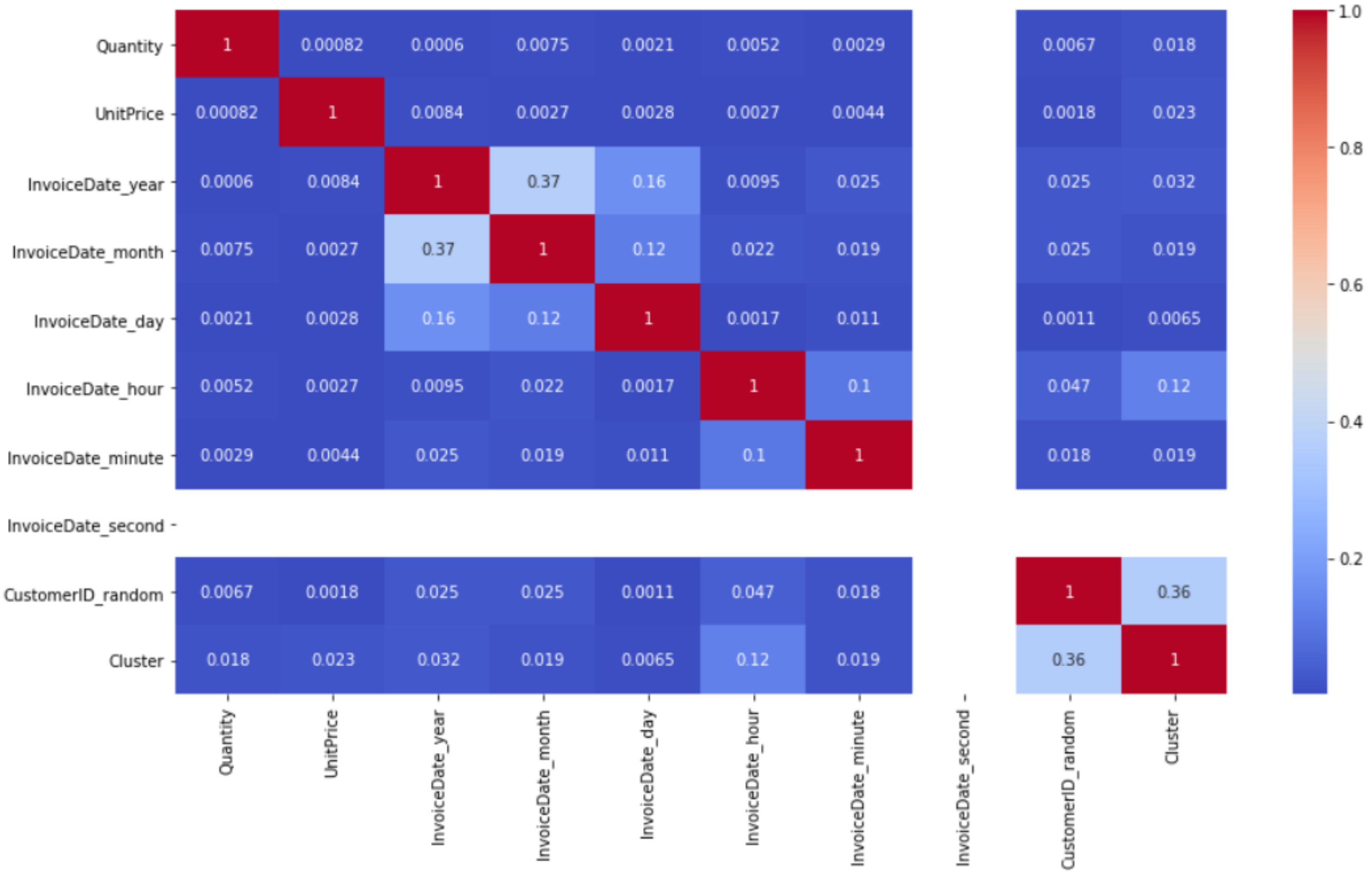
Spread over time and Country



Distribution of Number of reviews



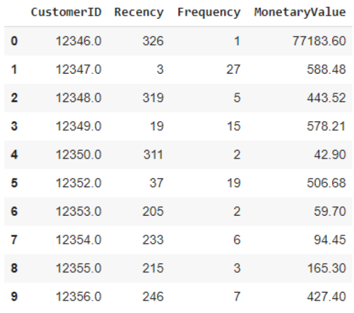
Correlation Matrix



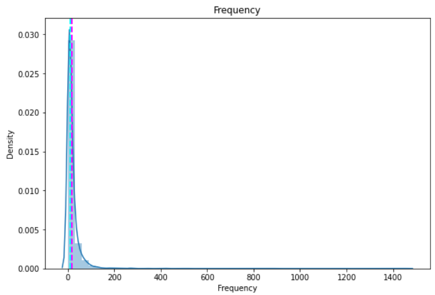
	Quantity	UnitPrice	InvoiceDate_year	InvoiceDate_month	InvoiceDate_day	InvoiceDate_hour	InvoiceDate_minute	InvoiceDate_second	CustomerID_random	Cluster
Quantity	1	0.00082	0.0006	0.0075	0.0021	0.0052	0.0029		0.0067	0.018
UnitPrice	0.00082	1	0.0084	0.0027	0.0028	0.0027	0.0044		0.0018	0.023
InvoiceDate_year	0.0006	0.0084	1	0.37	0.16	0.0095	0.025		0.025	0.032
InvoiceDate_month	0.0075	0.0027	0.37	1	0.12	0.022	0.019		0.025	0.019
InvoiceDate_day	0.0021	0.0028	0.16	0.12	1	0.0017	0.011		0.0011	0.0065
InvoiceDate_hour	0.0052	0.0027	0.0095	0.022	0.0017	1	0.1		0.047	0.12
InvoiceDate_minute	0.0029	0.0044	0.025	0.019	0.011	0.1	1		0.018	0.019
InvoiceDate_second										
CustomerID_random	0.0067	0.0018	0.025	0.025	0.0011	0.047	0.018		1	0.36
Cluster	0.018	0.023	0.032	0.019	0.0065	0.12	0.019		0.36	1

RFM Table for Customer ID

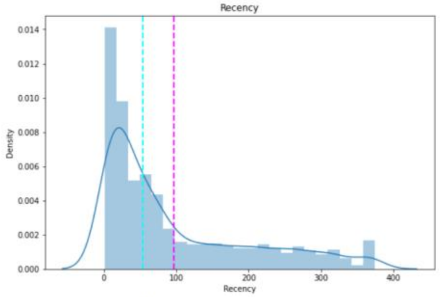
- R (Recency): Number of days since last purchase
- F (Frequency): Number of tracsactions
- M (Monetary): Total amount of transactions (revenue contributed)



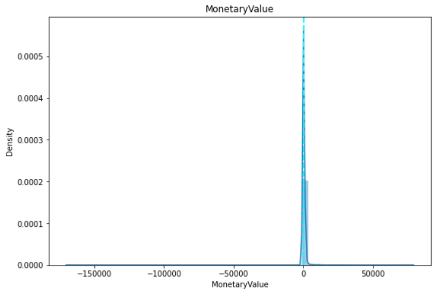
	CustomerID	Recency	Frequency	MonetaryValue
0	12346.0	326	1	77183.60
1	12347.0	3	27	588.48
2	12348.0	319	5	443.52
3	12349.0	19	15	578.21
4	12350.0	311	2	42.90
5	12352.0	37	19	506.68
6	12353.0	205	2	59.70
7	12354.0	233	6	94.45
8	12355.0	215	3	165.30
9	12356.0	246	7	427.40



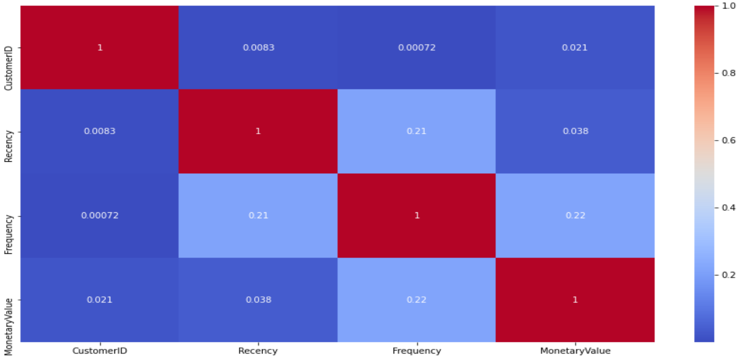
```
Skewness : 17.707469730773248
Kurtosis : 455.77098759018554
```

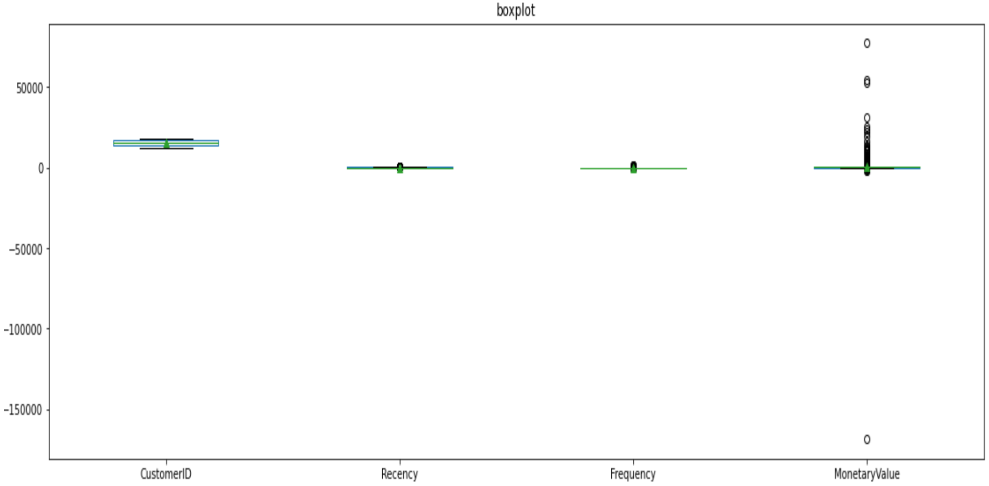


```
Skewness : 1.1894093874114786
Kurtosis : 0.2604433473318766
```



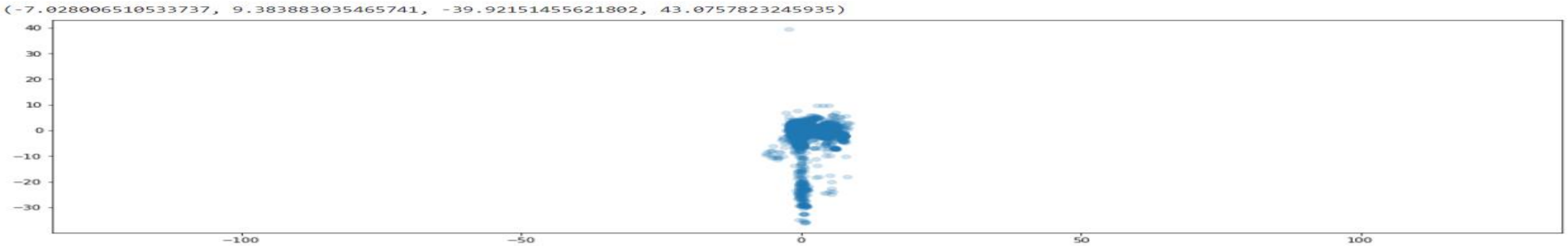
```
Skewness : -25.90170467892271
Kurtosis : 1691.90419354206
```





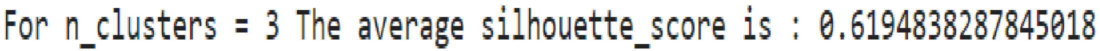
Data Preparation

- **One Hot Encoding**
- **Outlier Treatment**
- **Standard Scaler Scaling**
- **Principal Component Analysis (**n_components = 4**)**

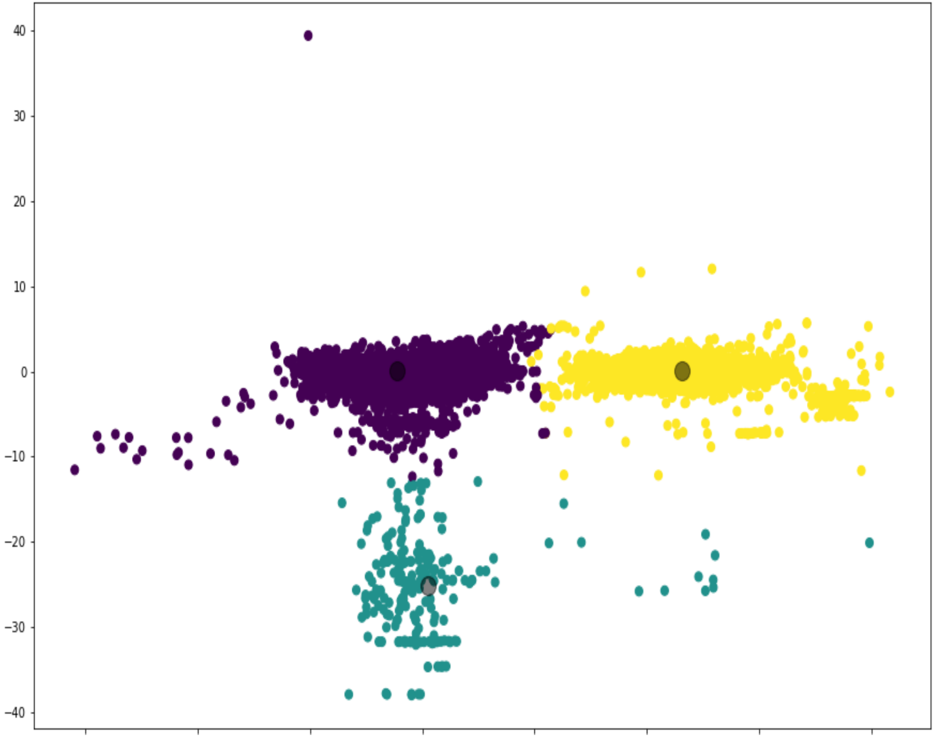


```
(-7.028006510533737, 9.383883035465741, -39.92151455621802, 43.0757823245935)
```

K Means Clustering

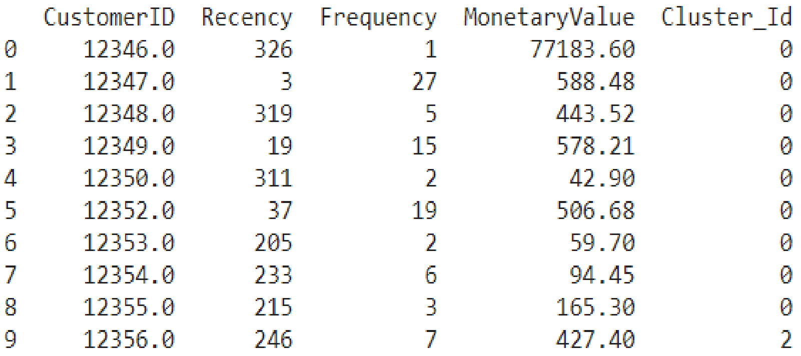


```
For n_clusters = 3 The average silhouette_score is : 0.6194838287845018
```

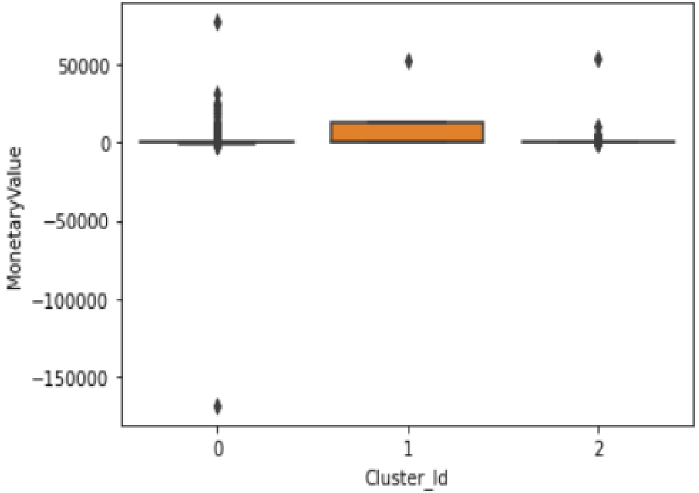


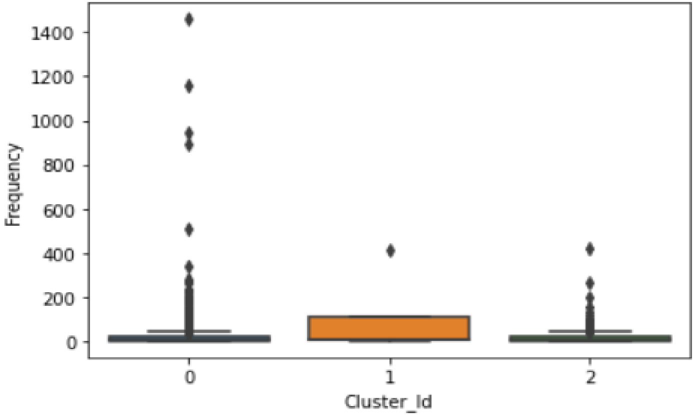
Hyper parameters
{n_clusters=3,
max_iter=1000,
random_state=10}

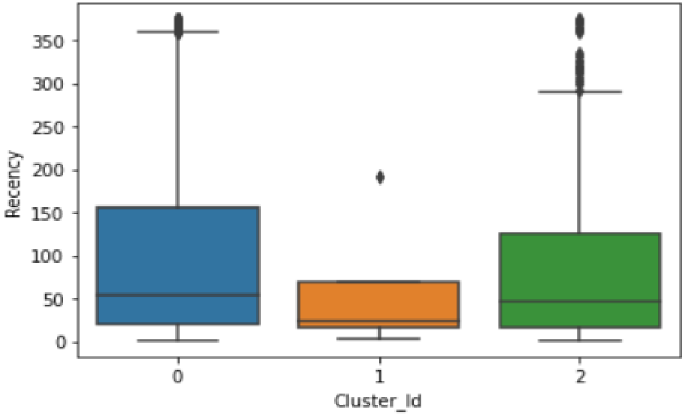
RFM For Cluster



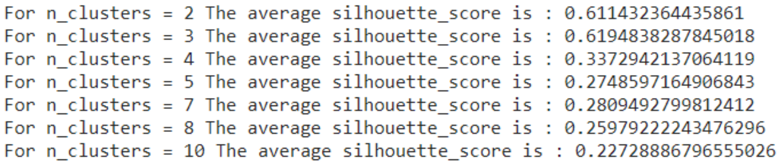
```
   CustomerID  Recency  Frequency  MonetaryValue  Cluster_Id
0     12346.0      326          1       77183.60           0
1     12347.0        3         27         588.48           0
2     12348.0      319          5         443.52           0
3     12349.0       19         15         578.21           0
4     12350.0      311          2          42.90           0
5     12352.0       37         19         506.68           0
6     12353.0      205          2          59.70           0
7     12354.0      233          6          94.45           0
8     12355.0      215          3         165.30           0
9     12356.0      246          7         427.40           2
```



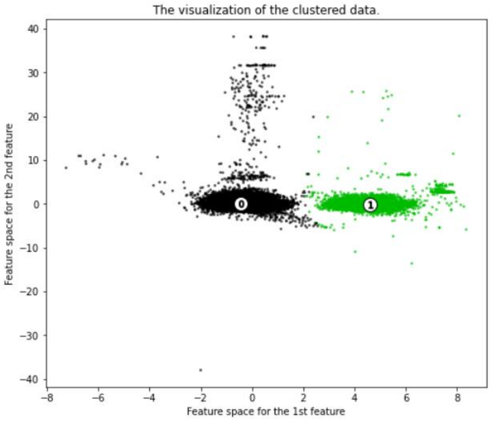


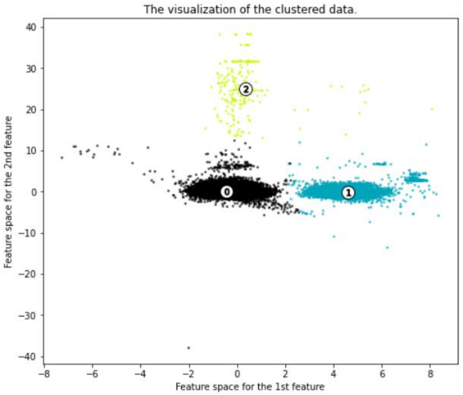


K-Means Clustering with Silhouette



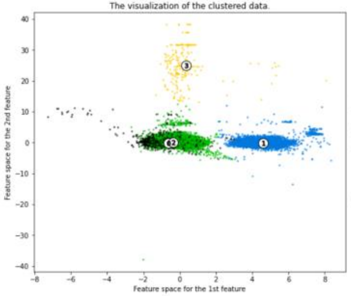
```
For n_clusters = 2 The average silhouette_score is : 0.611432364435861
For n_clusters = 3 The average silhouette_score is : 0.6194838287845018
For n_clusters = 4 The average silhouette_score is : 0.3372942137064119
For n_clusters = 5 The average silhouette_score is : 0.2748597164906843
For n_clusters = 7 The average silhouette_score is : 0.2809492799812412
For n_clusters = 8 The average silhouette_score is : 0.25979222243476296
For n_clusters = 10 The average silhouette_score is : 0.22728886796555026
```

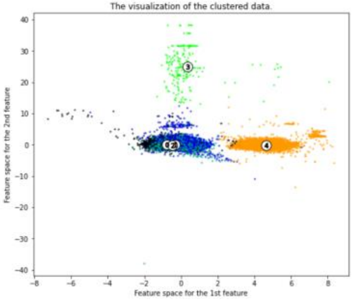


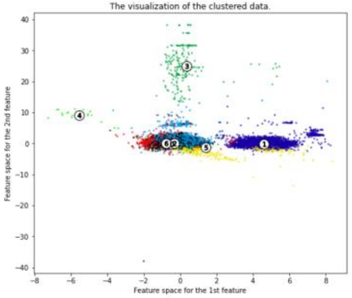


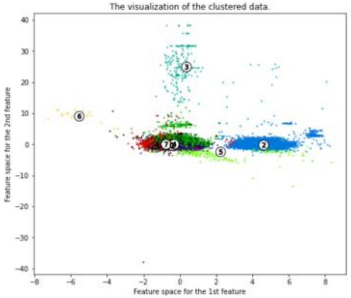
Hyper parameter

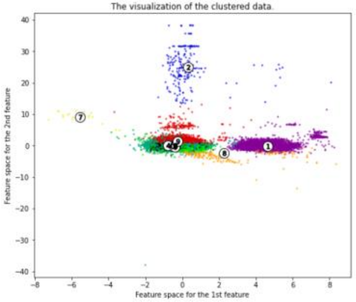
{ n_clusters=[2,3,4,5,7,8,10],
max_iter=1000,
random_state=10 }



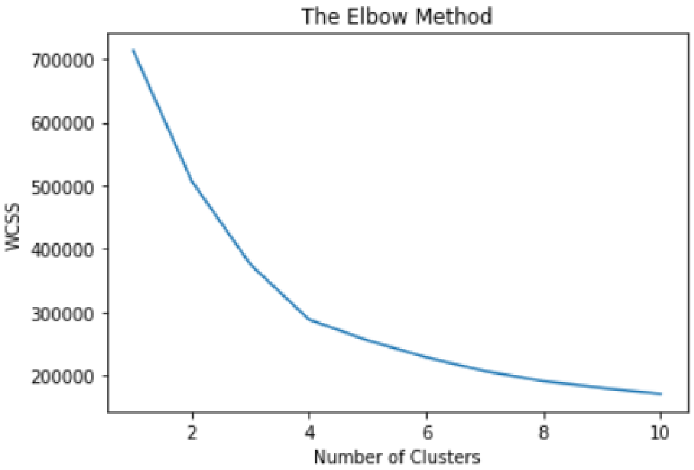




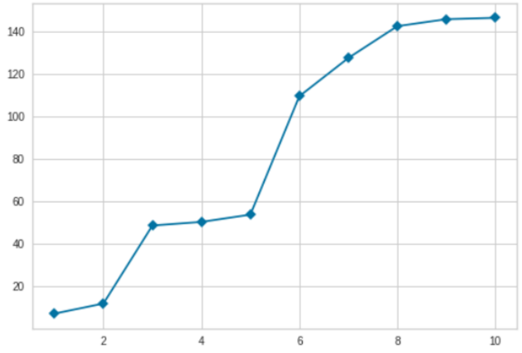




K-Means Clustering with Elbow method



```
KElbowVisualizer(ax=<matplotlib.axes._subplots.AxesSubplot object at 0x7f09590699d0>,
                 k=None, metric=None, model=None, timings=False)
```



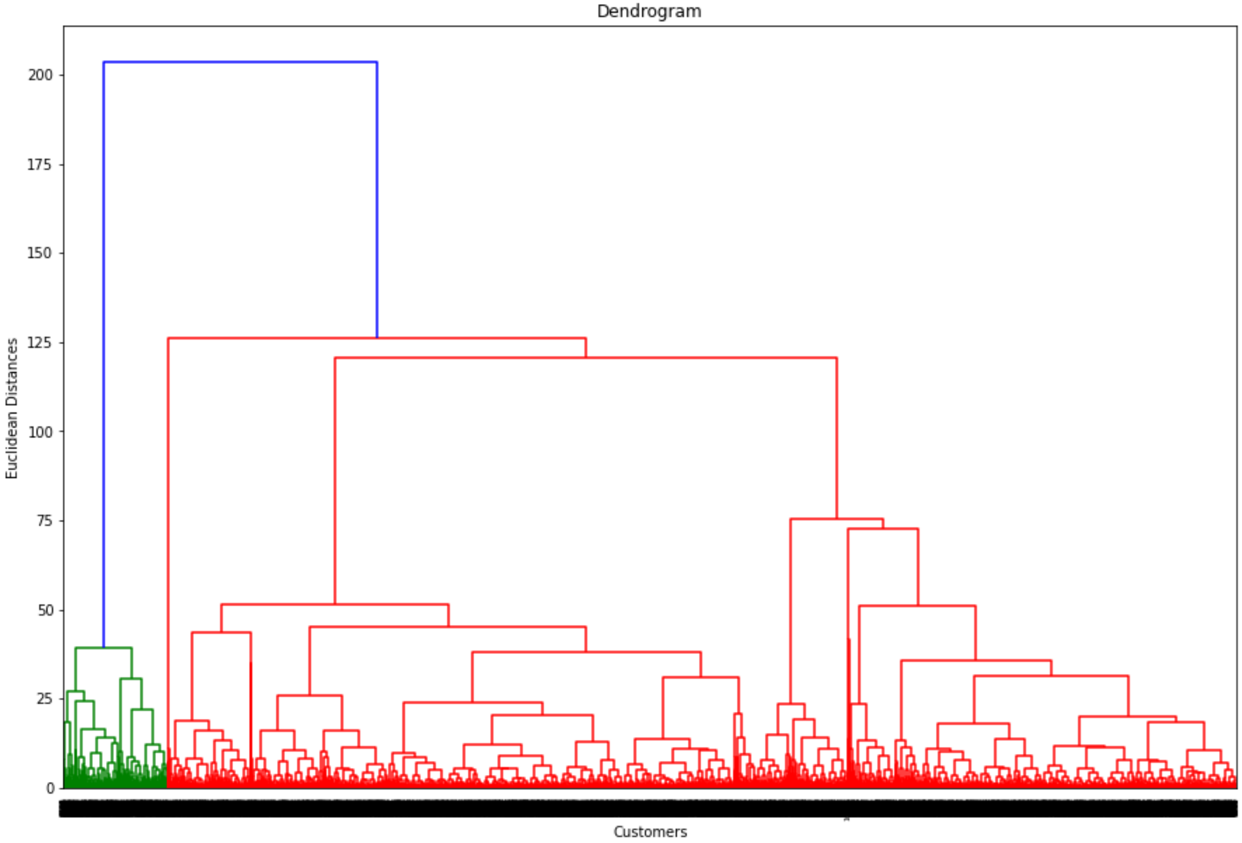
Hyper parameter

{ n_clusters=[1,10],
 init='k-means++',
random_state=0}

Hierarchical Clustering

Hyper parameteı

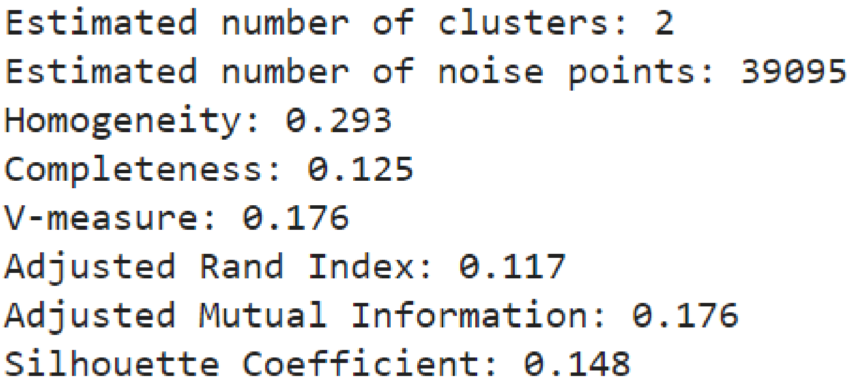
AgglomerativeClustering
{ n_clusters = 3,
affinity = 'euclidean',
linkage = 'ward‘}



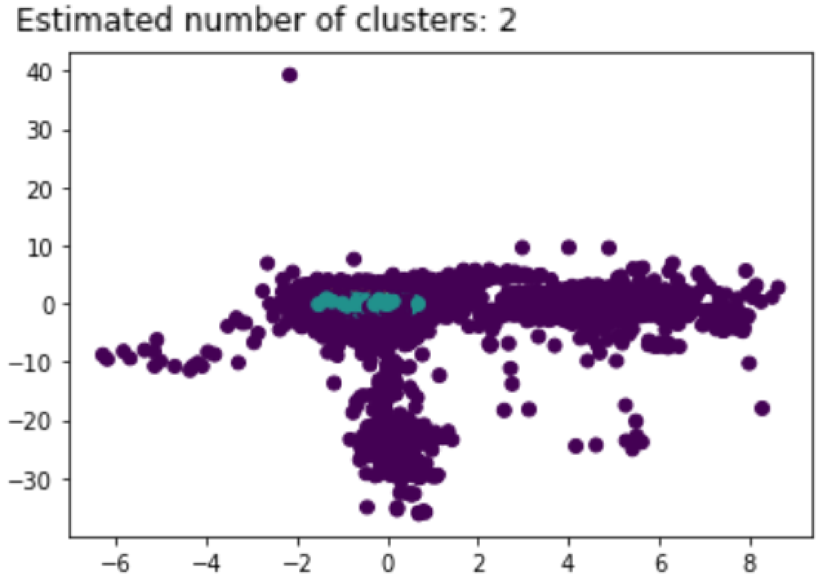
Density-Based Spatial Clustering Of Applications With Noise (DBSCAN)

Hyper parameter

{ eps=0.3,
min_samples=100 }



```
Estimated number of clusters: 2
Estimated number of noise points: 39095
Homogeneity: 0.293
Completeness: 0.125
V-measure: 0.176
Adjusted Rand Index: 0.117
Adjusted Mutual Information: 0.176
Silhouette Coefficient: 0.148
```



Challenges

- **Large Dataset to handle.**
- **Needs to plot lot of Graphs to analyse.**
- **Lot of NaN values.**
- **Continuous Runtime and RAM Crash due to large dataset.**
- **Carefully tuned Hyper parameters .**

Conclusion

- K-Means Clustering with Silhouette gives the highest score of 61.9% for number of clusters 3.

- Sales has been increased from 2010 to 2011.

- RFM for Cluster ID box plots tells well about Cluster detail.

- We can deploy this model.

THANK YOU

Q & A