Capstone Project – 4

Machine Learning – Unsupervised

Team **Champions** : Online Retail Customer Segmentation

Team Members

Varsha Rani
Vivek Chandrakant Pawar
Rabista Parween
Tushar Gaikwad

POINTS FOR DISCUSSION

- Problem Description
- Data Summary
- Importing Libraries & Data Inspection
- Data Cleaning
- Exploratory Data Analysis
- RFM(Recency Frequency Monetary) Analysis
- Model Preparation
- Data Modeling
- Cluster 0 Analysis
- Cluster 1 Analysis
- Conclusion

Problem Description

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

We are given the following dataset:

Online Retail.xlsx

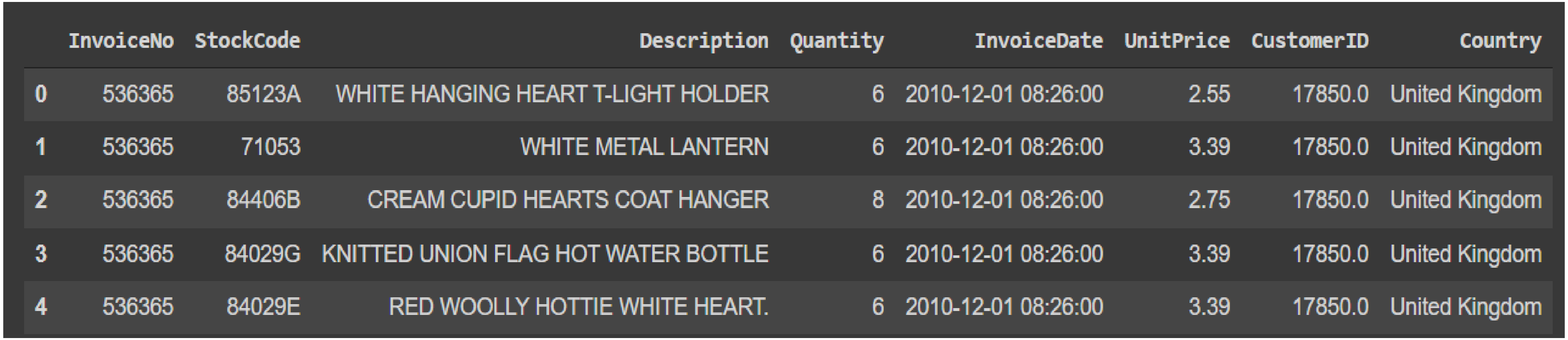
Data Summary

Attribute Information:

- 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.

Importing Libraries & Data Inspection

- Pandas – Manipulation of tabular data in Dataframes
- Numpy – Mathematical operations on arrays
- Matplotlib – Visualization
- Seaborn – Visualization
- Sklearn – Data Modeling



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Original Dataset contains 8 columns and 541909 rows.

Data Cleaning

After dropping the duplicate values, now we have 8 columns and 536641 rows.

After dropping the null values, now we have 8 columns and 401604 rows.

We added 7 more columns to our dataset:

Amount_spent = df['Quantity'] * df['UnitPrice’]

Year = df['InvoiceDate'].dt.year

Month = df['InvoiceDate'].dt.month

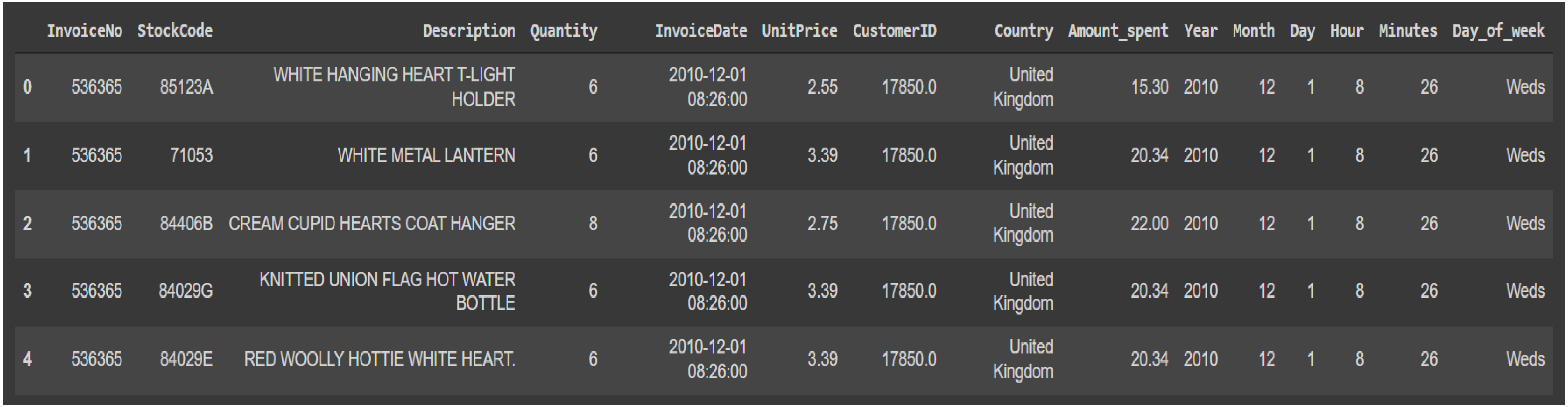
Day = df['InvoiceDate'].dt.day

Hour = df['InvoiceDate'].dt.hour

Minutes = df['InvoiceDate'].dt.minute

Day_of_week = df['InvoiceDate'].dt.dayofweek

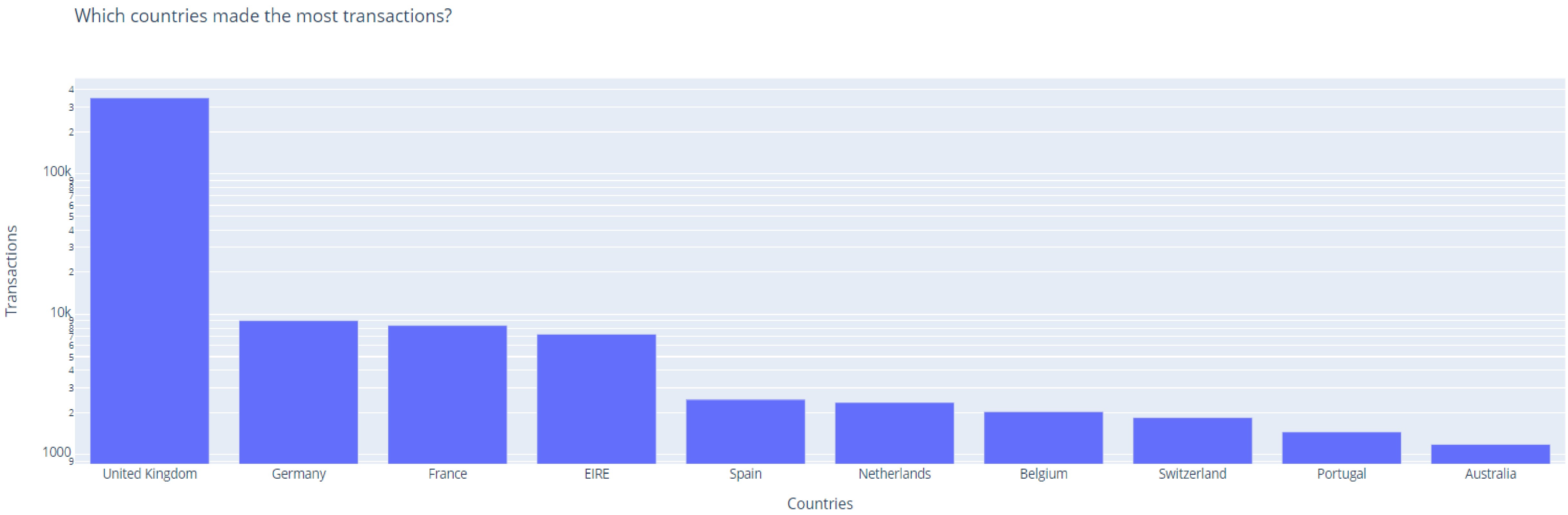
After cleaning and adding more column, we have the following data:



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Amount_spent	Year	Month	Day	Hour	Minutes	Day_of_week
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	2010	12	1	8	26	Weds
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010	12	1	8	26	Weds
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00	2010	12	1	8	26	Weds
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010	12	1	8	26	Weds
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010	12	1	8	26	Weds

Exploratory Data Analysis

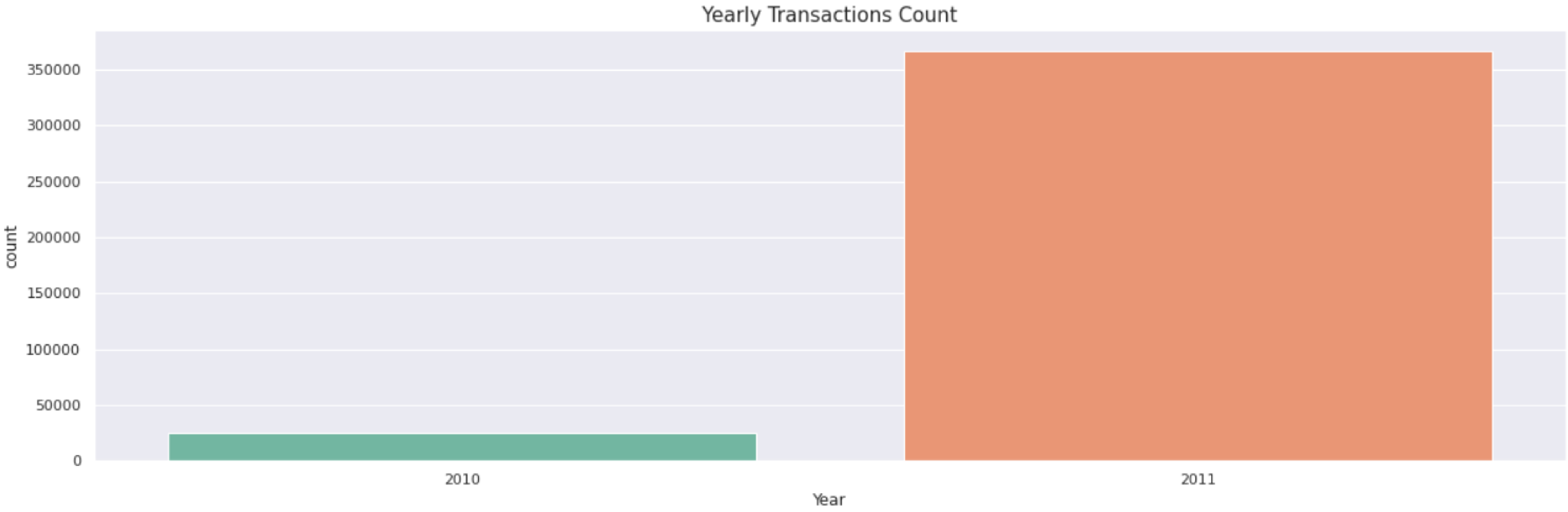
Which countries made the most transactions?



The above graph shows that which country have made the most transactions. United Kingdom have made most transactions followed by Germany.

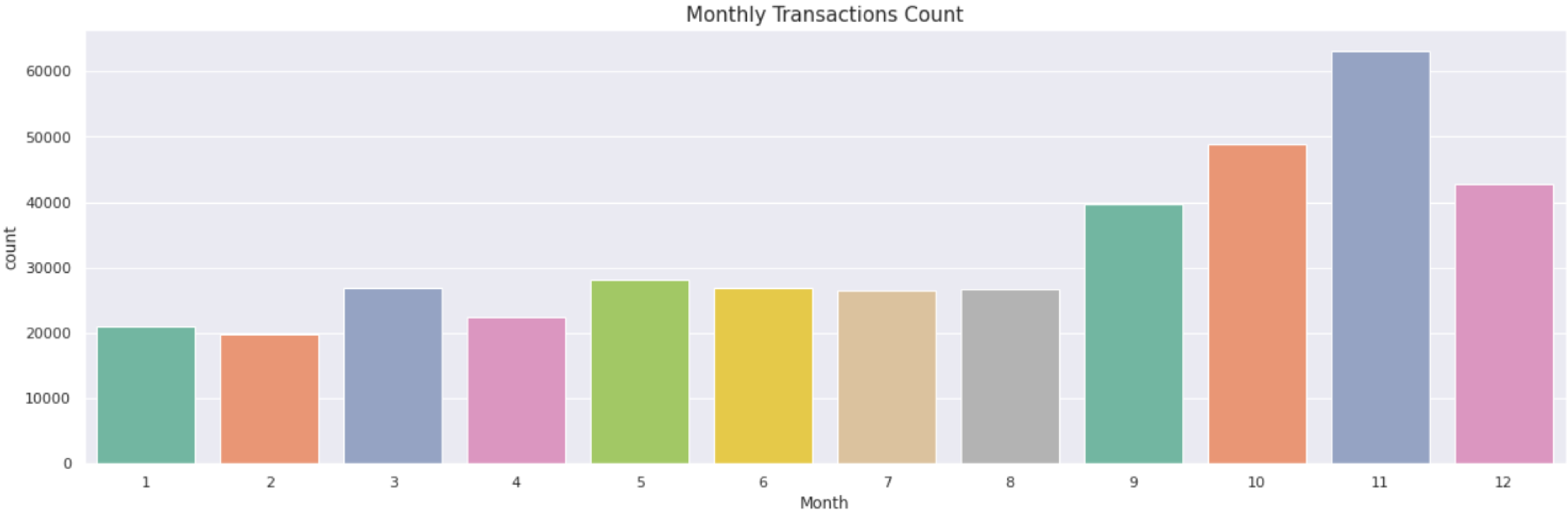
EDA Continued…

Yearly transactions Count



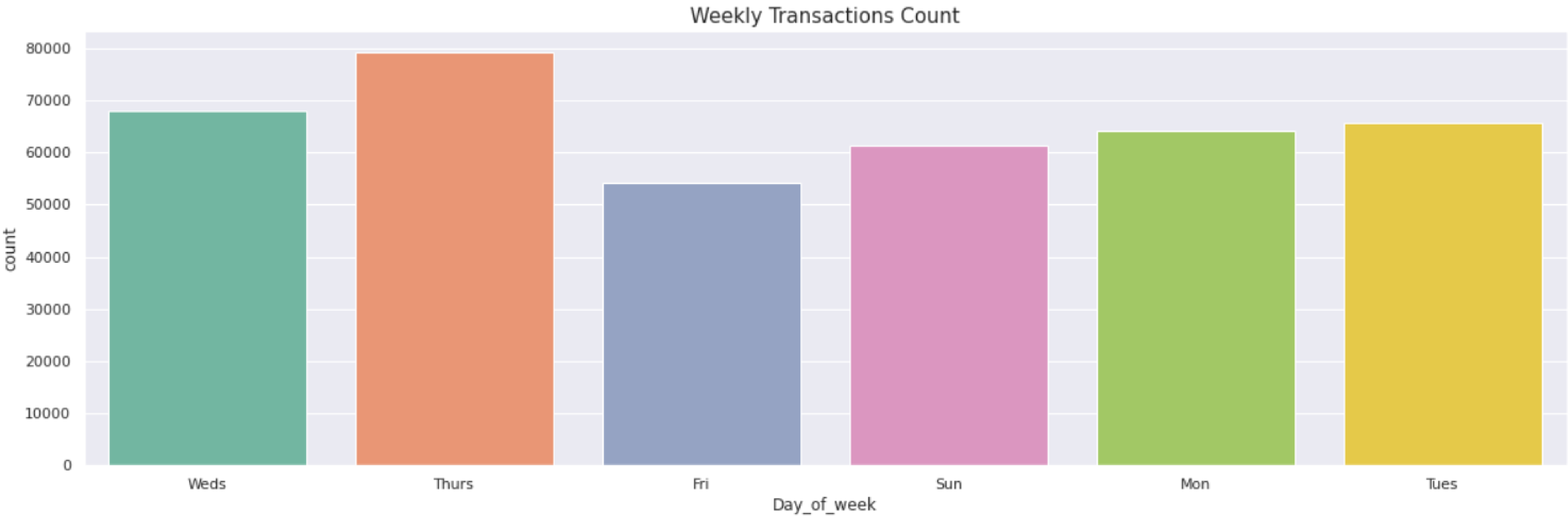
EDA Continued…

Monthly Transactions Count



EDA Continued…

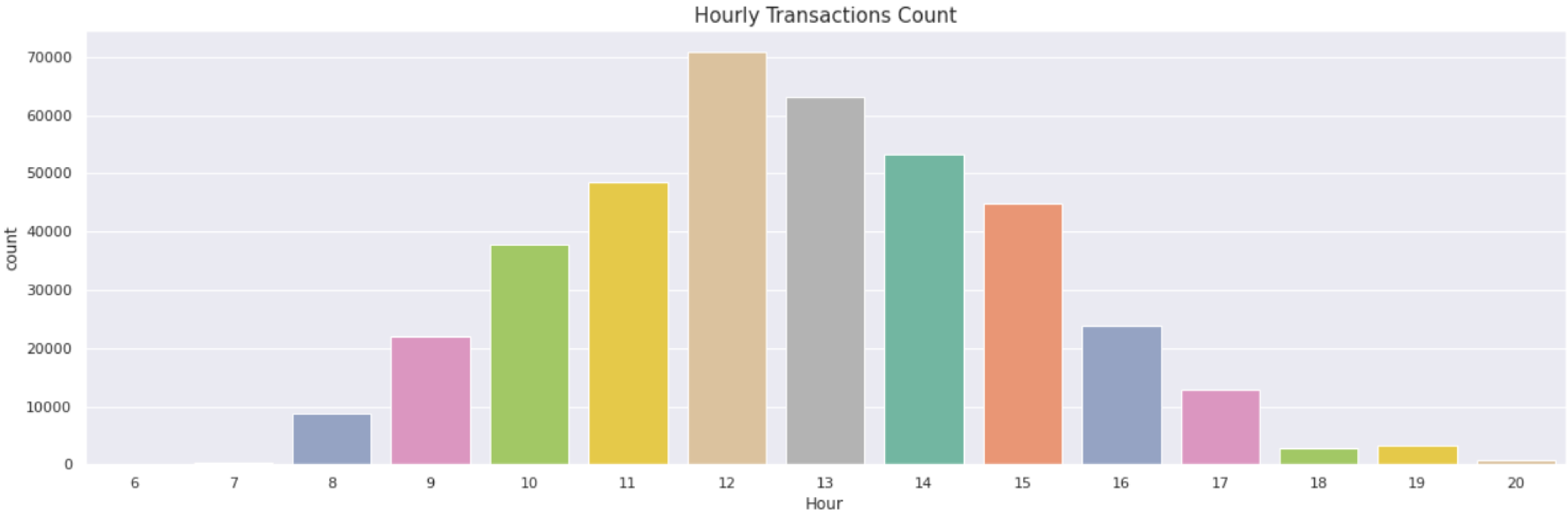
Weekly Transaction Count



-

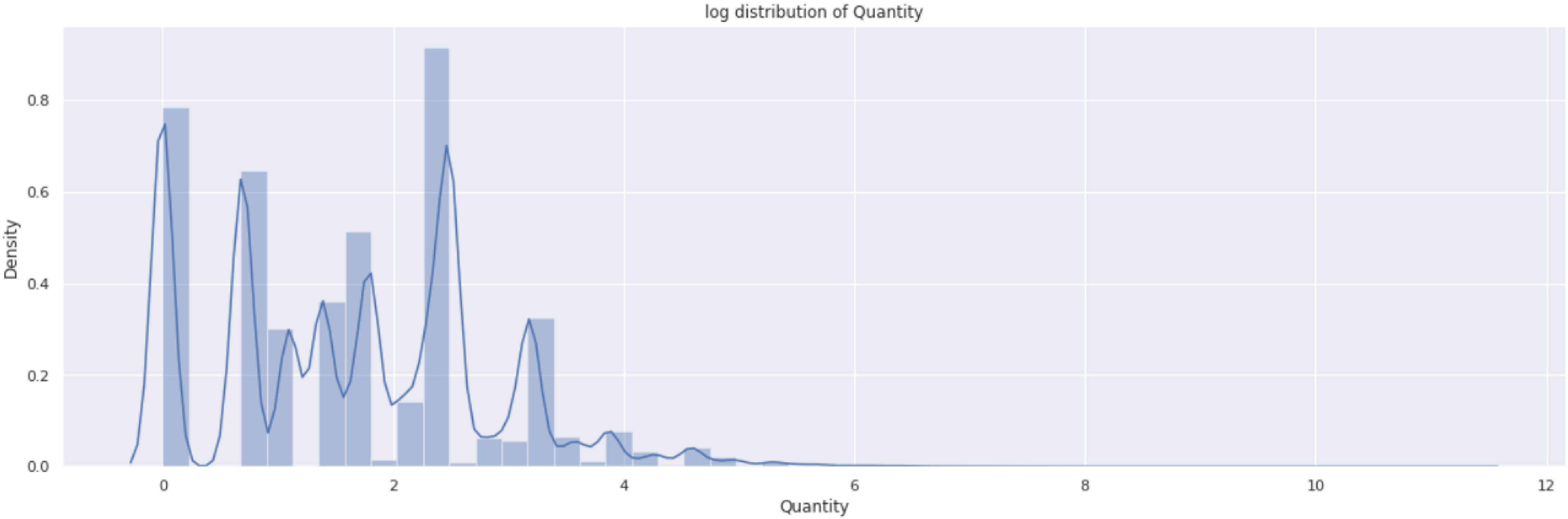
EDA Continued…

Hourly Transactions Count



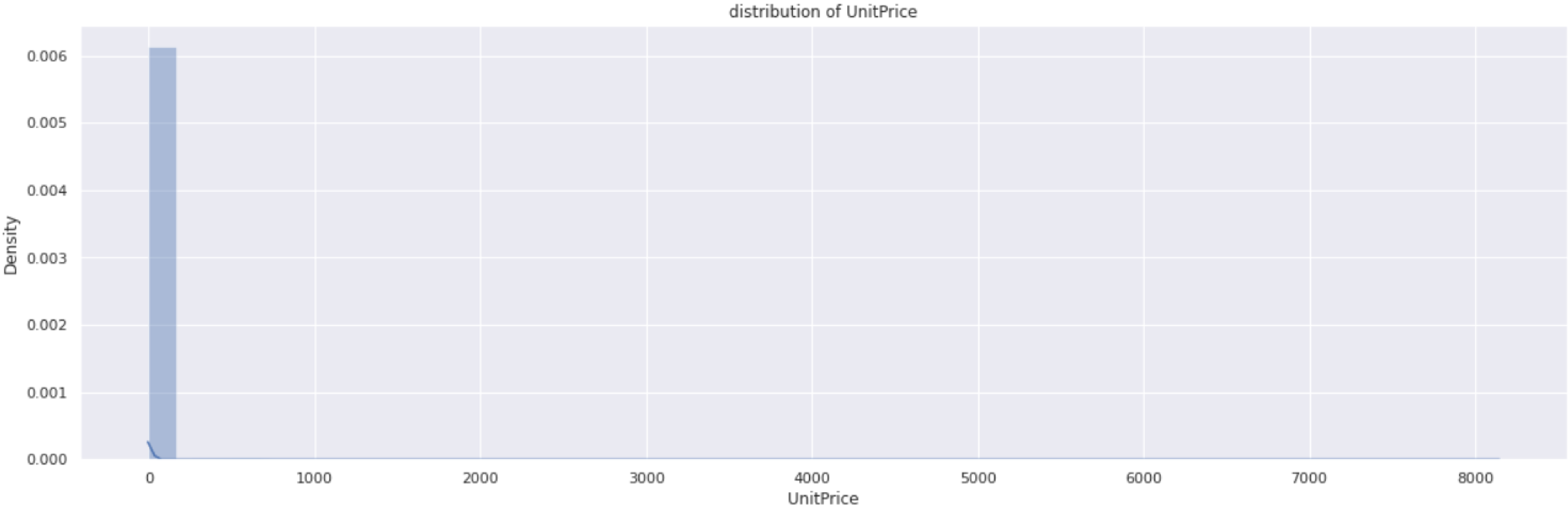
EDA Continued…

Log Distribution of Quantity



EDA Continued…

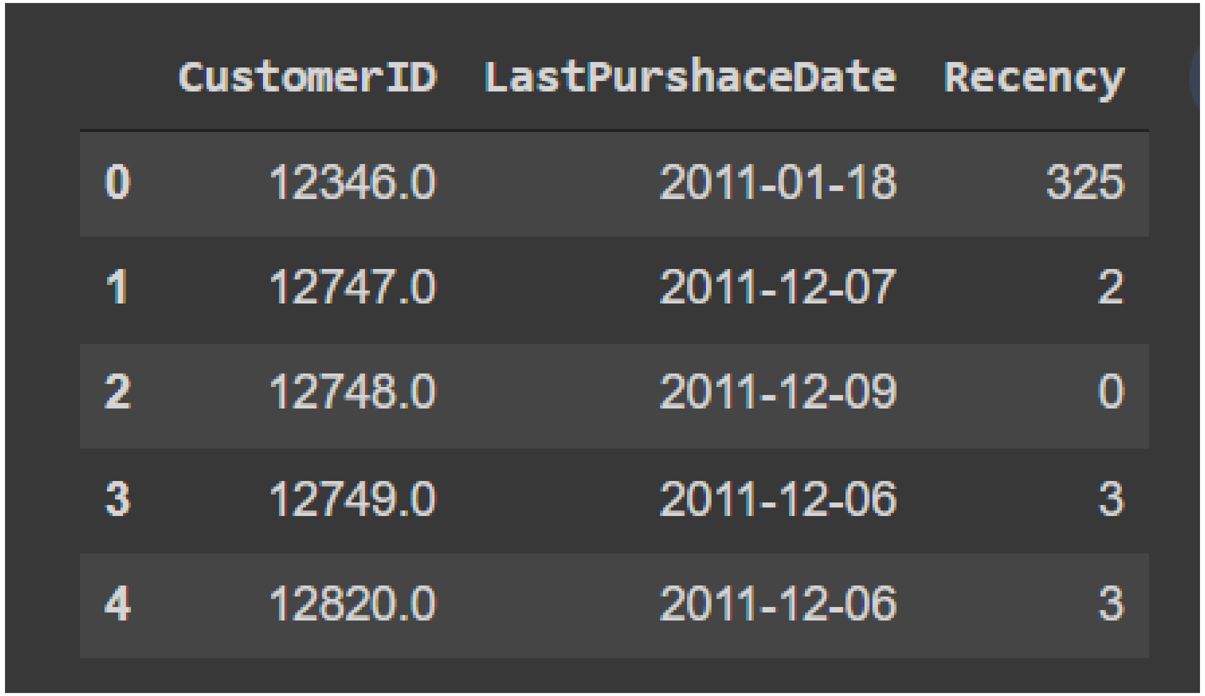
Distribution of UnitPrice



RFM(Recency Frequency Monetary) Analysis

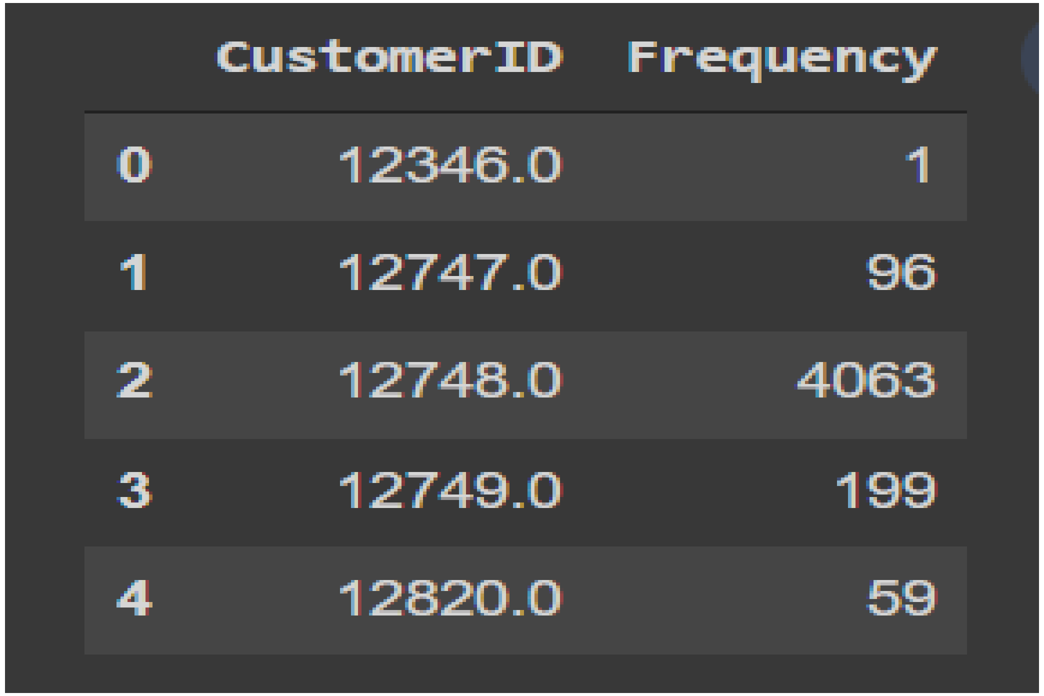
- RECENCY (R): Days since last purchase
- FREQUENCY (F): Total number of purchases
- MONETARY VALUE (M): Total money this customer spent.

Recency



	CustomerID	LastPurshaceDate	Recency
0	12346.0	2011-01-18	325
1	12747.0	2011-12-07	2
2	12748.0	2011-12-09	0
3	12749.0	2011-12-06	3
4	12820.0	2011-12-06	3

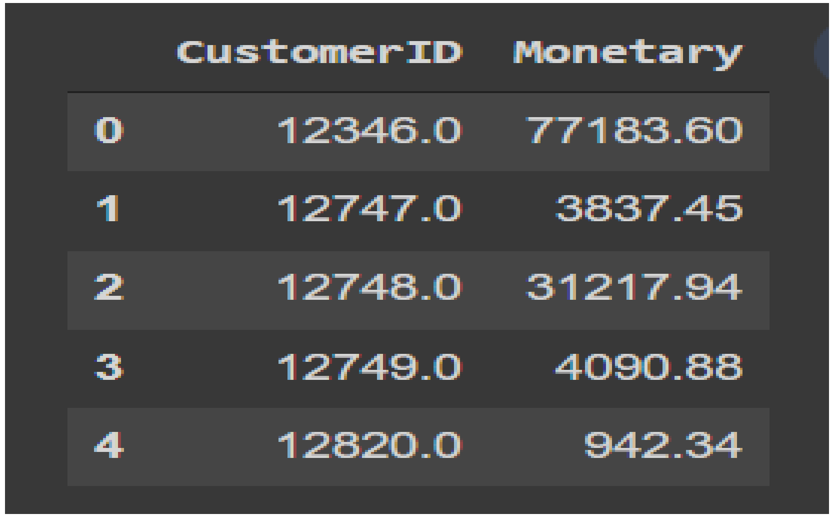
Frequency



	CustomerID	Frequency
0	12346.0	1
1	12747.0	96
2	12748.0	4063
3	12749.0	199
4	12820.0	59

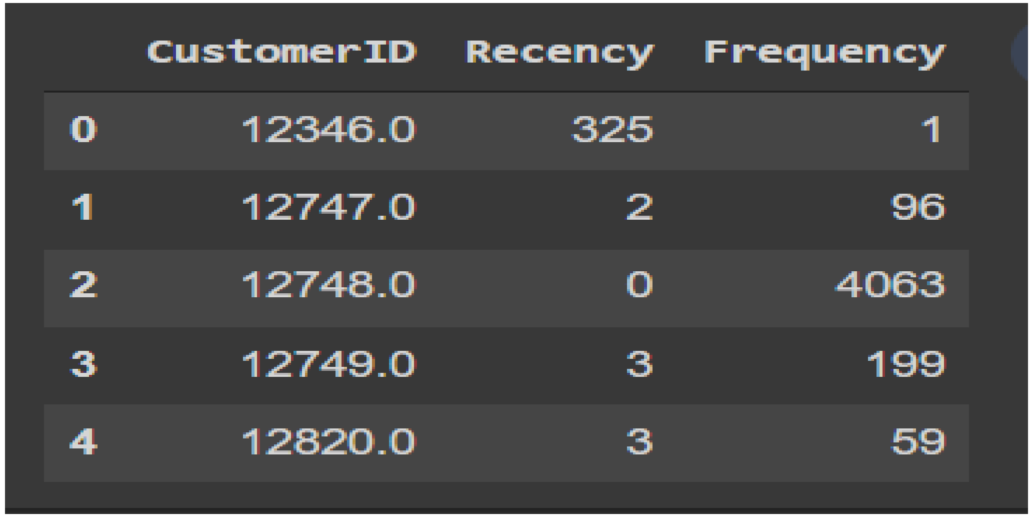
RFM Analysis Continued…

Monetary



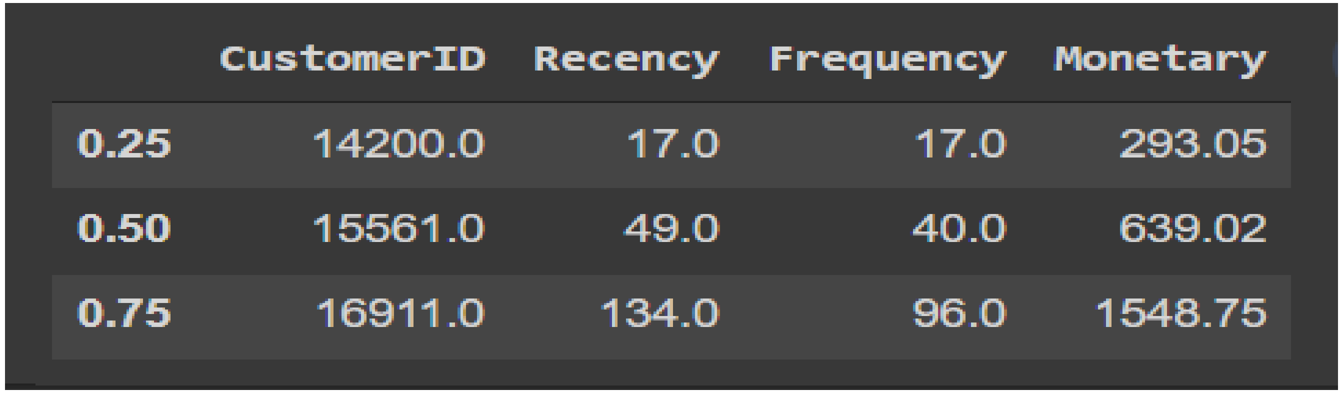
	CustomerID	Monetary
0	12346.0	77183.60
1	12747.0	3837.45
2	12748.0	31217.94
3	12749.0	4090.88
4	12820.0	942.34

Recency with Frequency



	CustomerID	Recency	Frequency
0	12346.0	325	1
1	12747.0	2	96
2	12748.0	0	4063
3	12749.0	3	199
4	12820.0	3	59

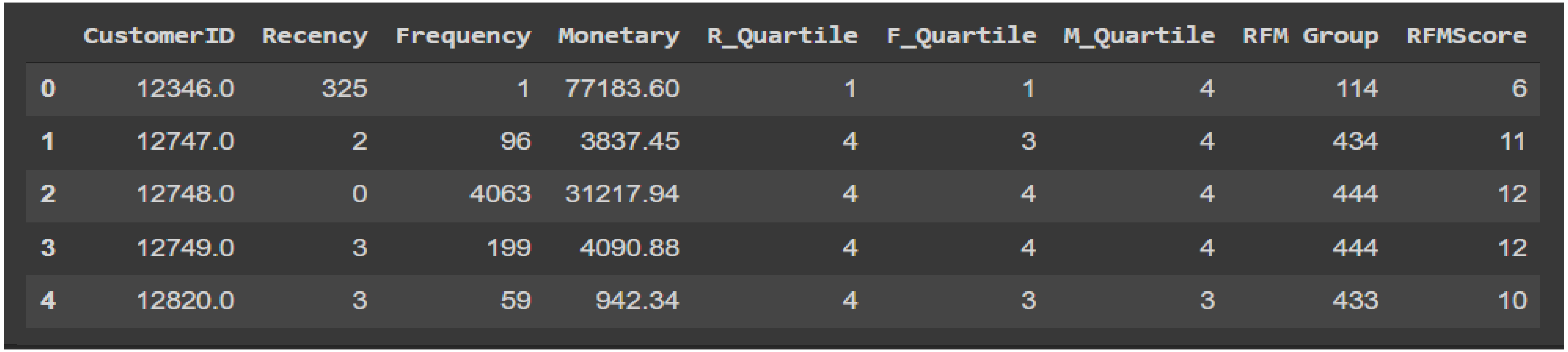
RFM Quantiles(Recency, Frequency and Monetary)



	CustomerID	Recency	Frequency	Monetary
0.25	14200.0	17.0	17.0	293.05
0.50	15561.0	49.0	40.0	639.02
0.75	16911.0	134.0	96.0	1548.75

RFM Analysis Continued…

RFM Score



	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFM Group	RFMScore
0	12346.0	325	1	77183.60	1	1	4	114	6
1	12747.0	2	96	3837.45	4	3	4	434	11
2	12748.0	0	4063	31217.94	4	4	4	444	12
3	12749.0	3	199	4090.88	4	4	4	444	12
4	12820.0	3	59	942.34	4	3	3	433	10

Best Recency Score = 4 (most recently purchase)

Best Frequency Score = 4 (most quantity purchase)

Best Monetary Score = 4 (spent the most)

Number of different types of customers:

Best Customers = 404

Loyal Customers = 961

Big Spenders = 966

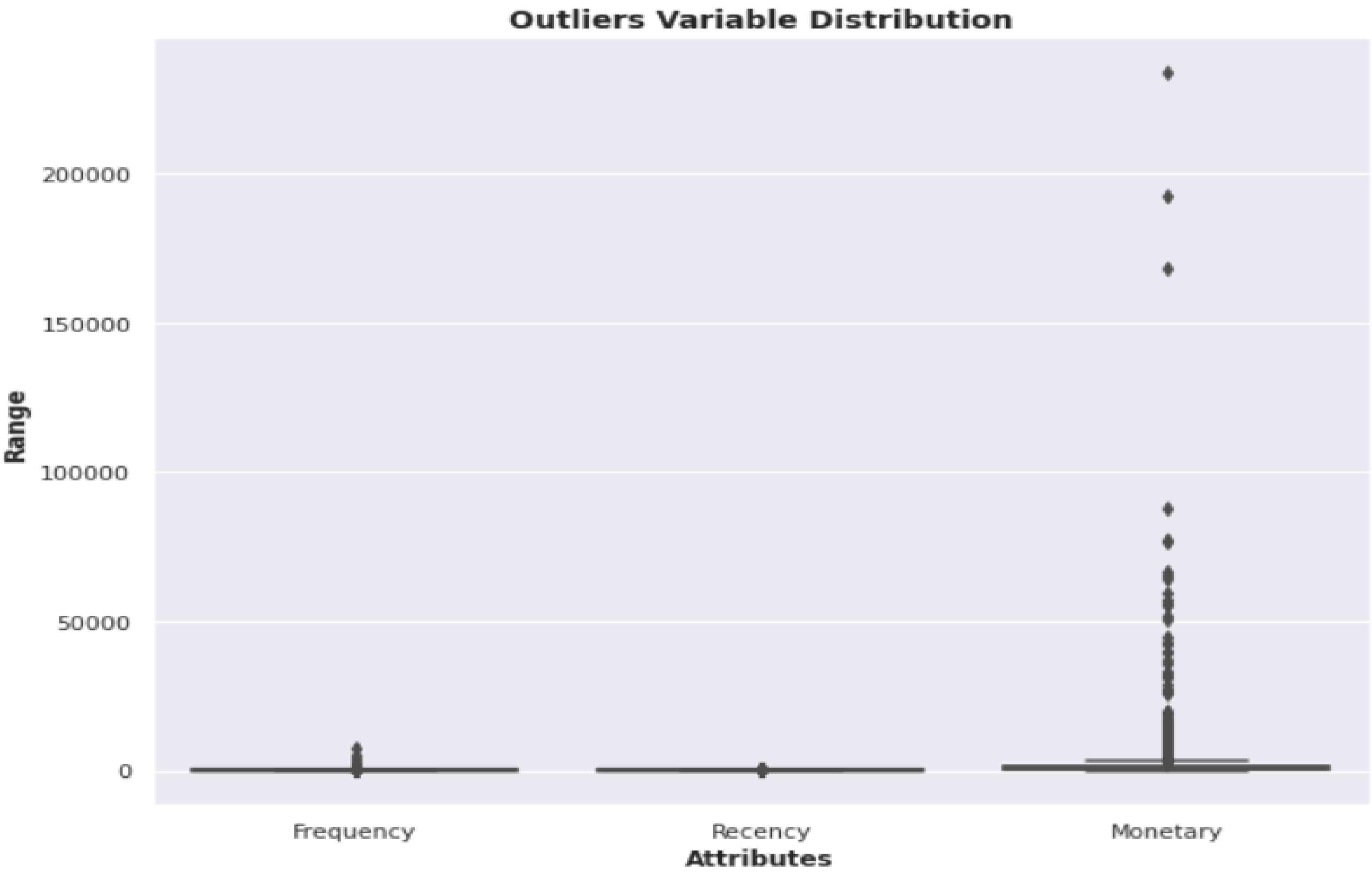
Almost Lost = 96

Lost Customers = 18

Lost Cheap Customers = 337

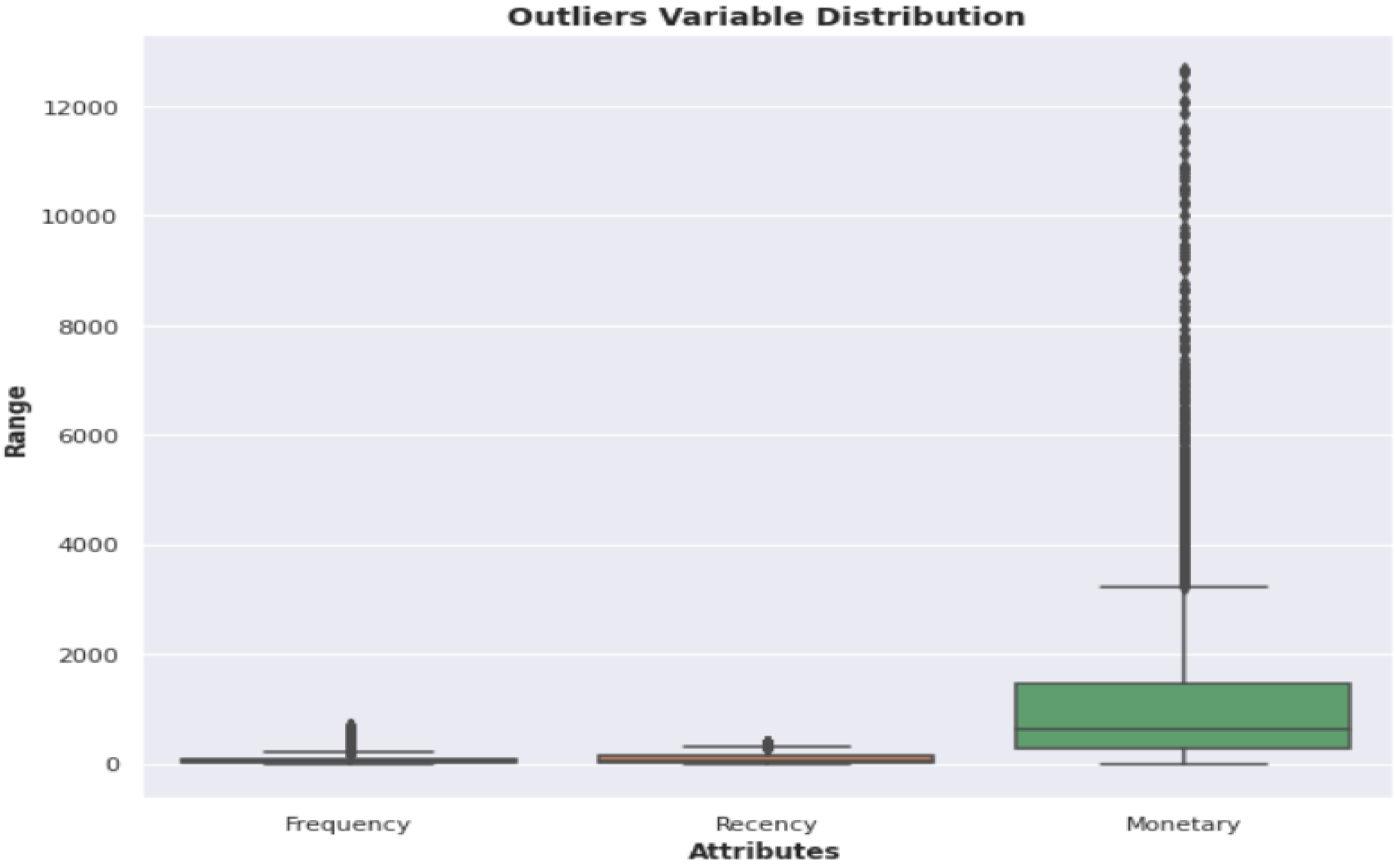
Model Preparation

Outliers Variable Distribution



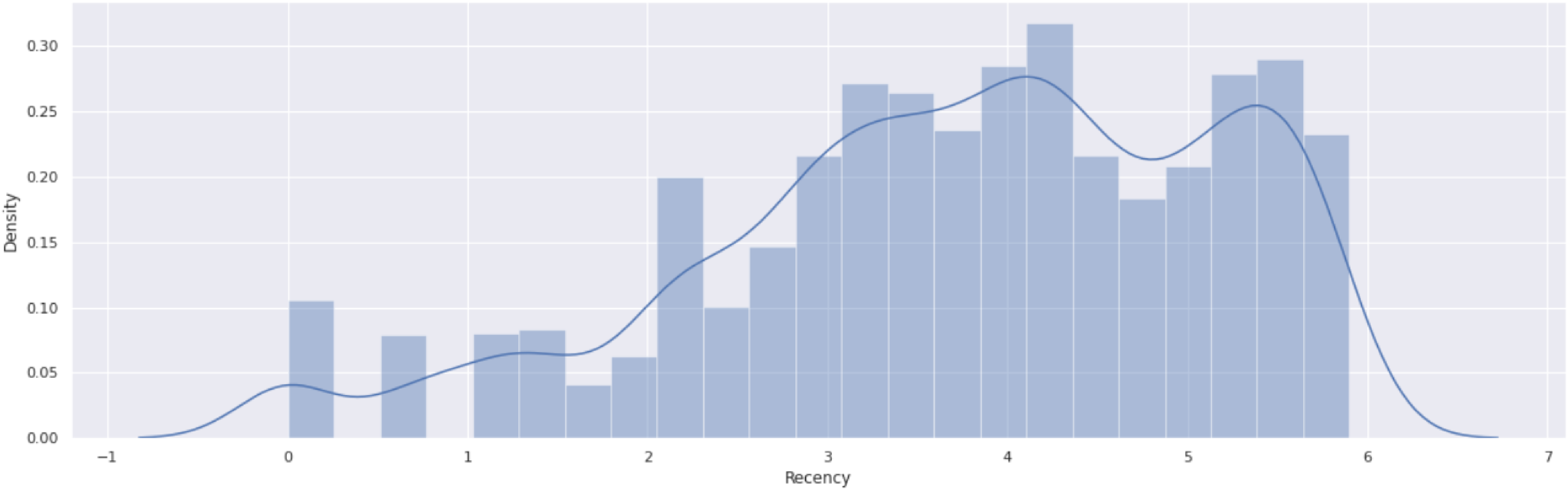
Model Preparation

Outliers Variable Distribution (After removing outliers for Amount)



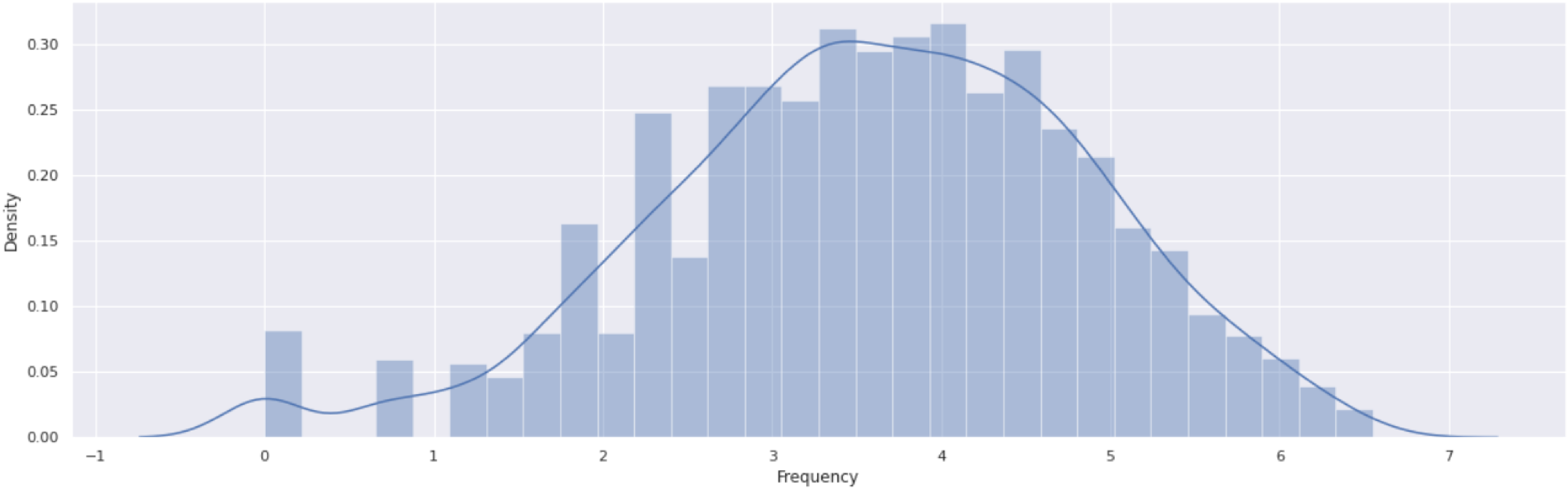
Model Preparation

Data distribution after data normalization for Recency



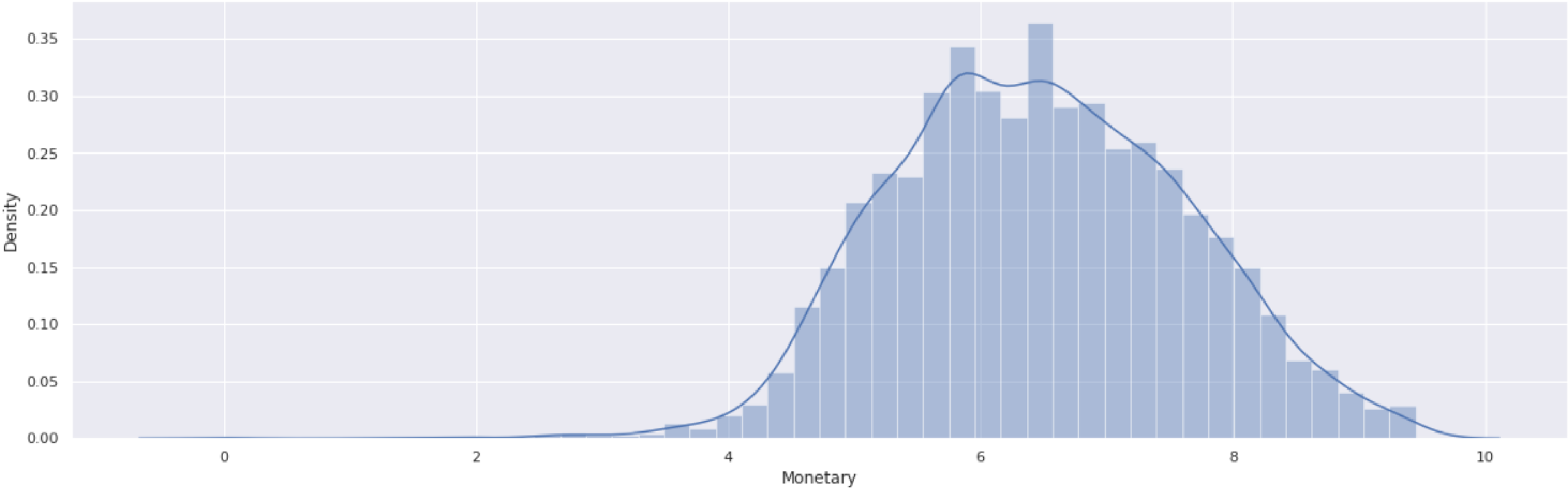
Model Preparation

Data distribution after data normalization for Frequency



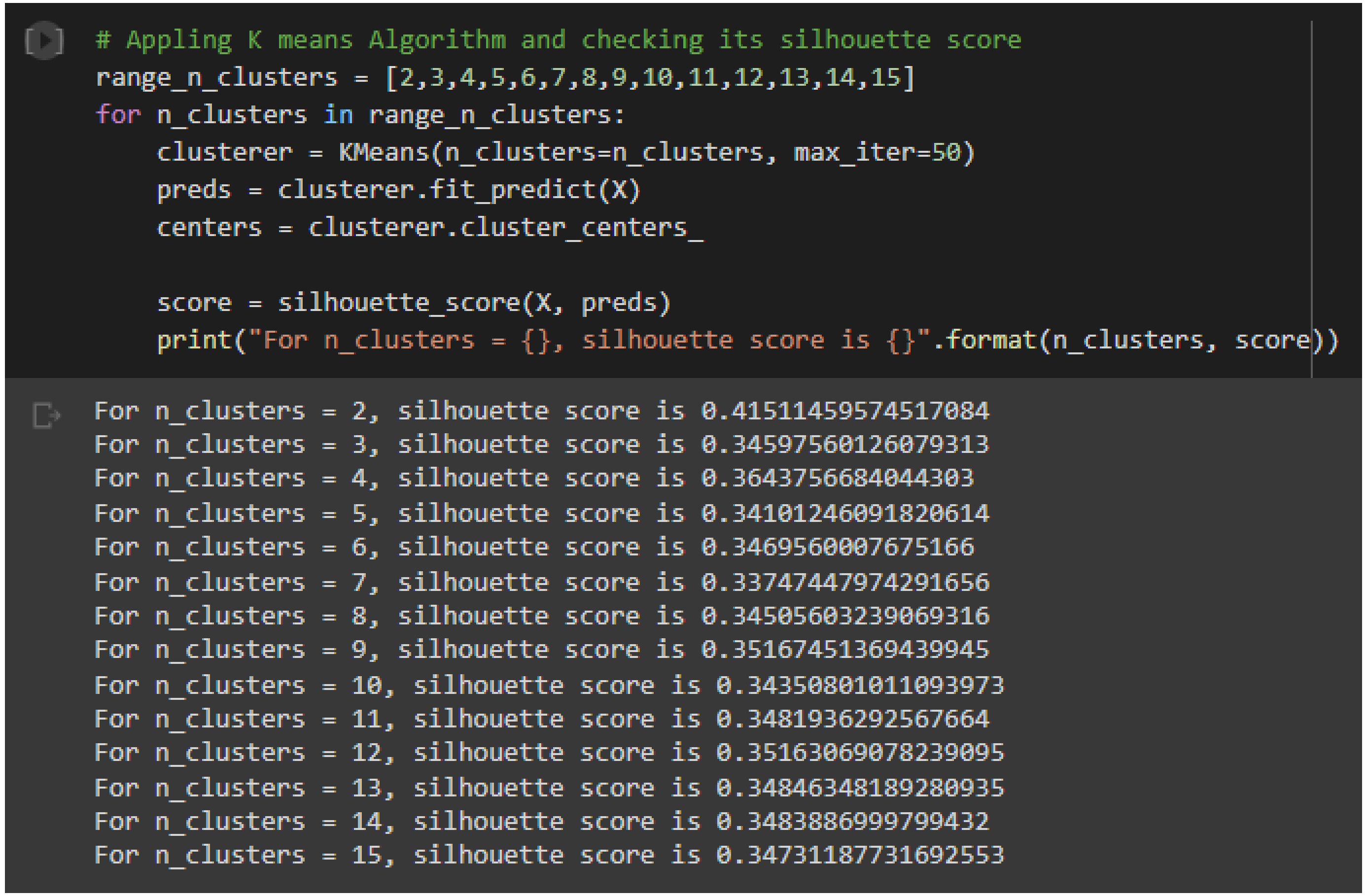
Model Preparation

Data distribution after data normalization for Monetary



Data Modeling

Apply Silhouette Score Method on Recency and Monetary



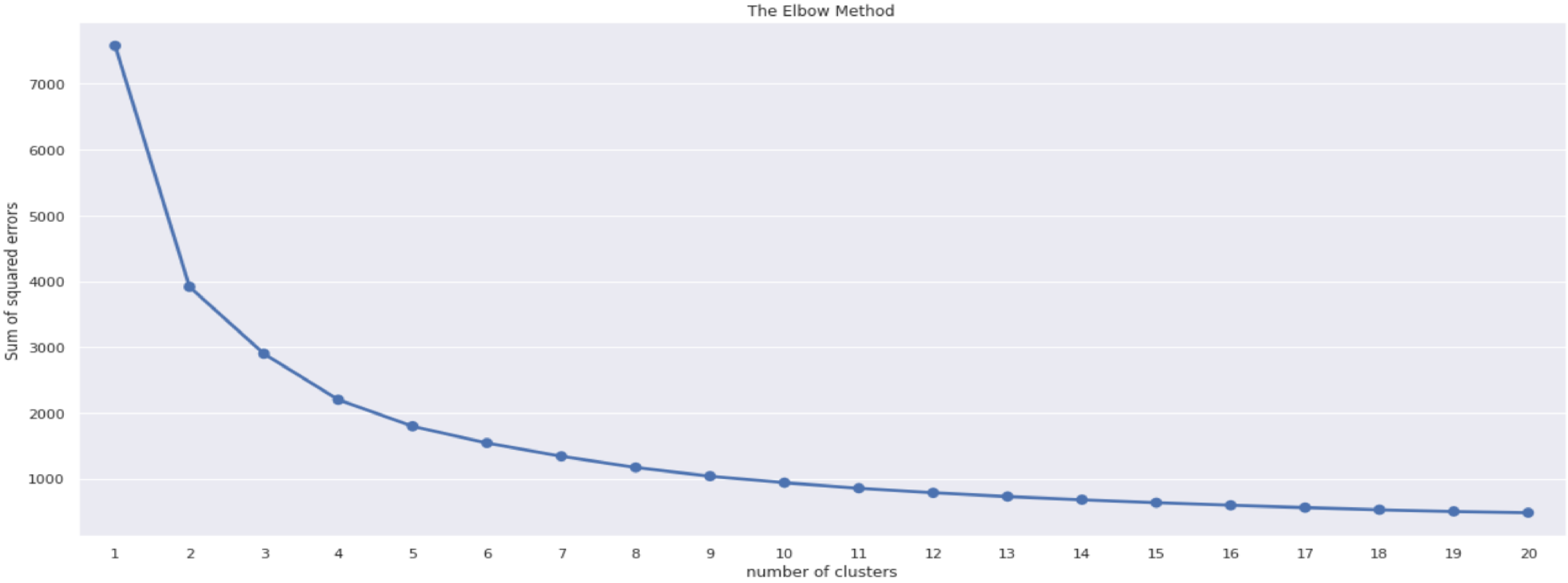
```
# Appling K means Algorithm and checking its silhouette score
range_n_clusters = [2,3,4,5,6,7,8,9,10,11,12,13,14,15]
for n_clusters in range_n_clusters:
    clusterer = KMeans(n_clusters=n_clusters, max_iter=50)
    preds = clusterer.fit_predict(X)
    centers = clusterer.cluster_centers_

    score = silhouette_score(X, preds)
    print("For n_clusters = {}, silhouette score is {}".format(n_clusters, score))
```

```
For n_clusters = 2, silhouette score is 0.41511459574517084
For n_clusters = 3, silhouette score is 0.34597560126079313
For n_clusters = 4, silhouette score is 0.3643756684044303
For n_clusters = 5, silhouette score is 0.34101246091820614
For n_clusters = 6, silhouette score is 0.3469560007675166
For n_clusters = 7, silhouette score is 0.33747447974291656
For n_clusters = 8, silhouette score is 0.34505603239069316
For n_clusters = 9, silhouette score is 0.35167451369439945
For n_clusters = 10, silhouette score is 0.34350801011093973
For n_clusters = 11, silhouette score is 0.3481936292567664
For n_clusters = 12, silhouette score is 0.35163069078239095
For n_clusters = 13, silhouette score is 0.34846348189280935
For n_clusters = 14, silhouette score is 0.3483886999799432
For n_clusters = 15, silhouette score is 0.34731187731692553
```

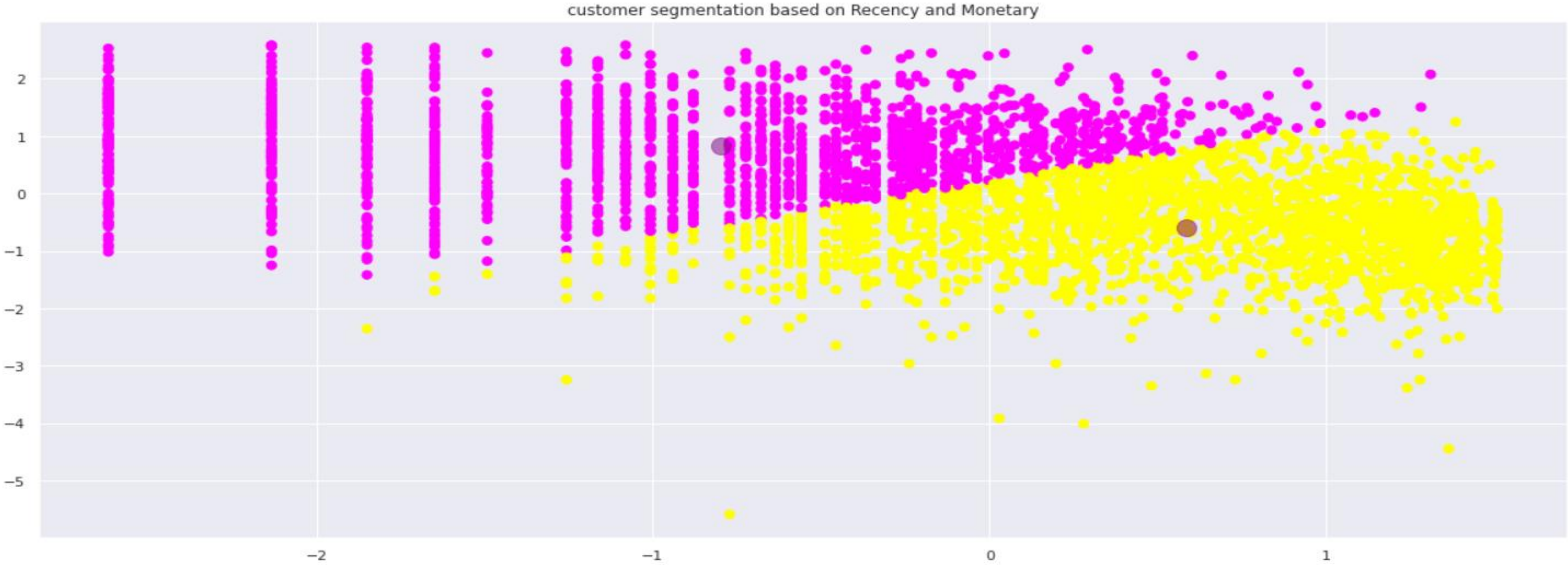
Data Modeling

Elbow Method on Recency and Monetary



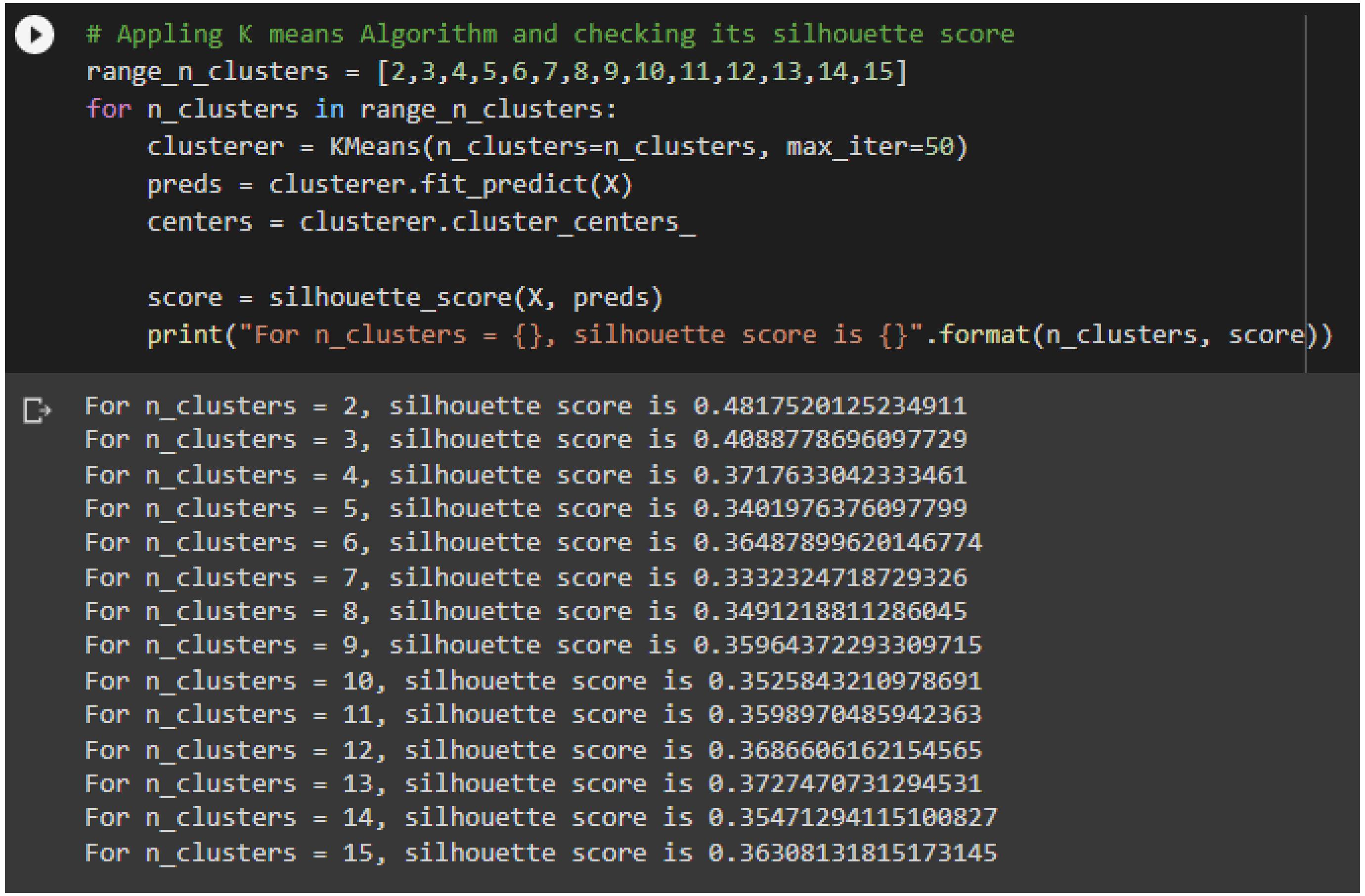
Data Modeling

Customer segmentation based on Recency and Monetary



Data Modeling

Apply Silhouette Score Method on Frequency and Monetary



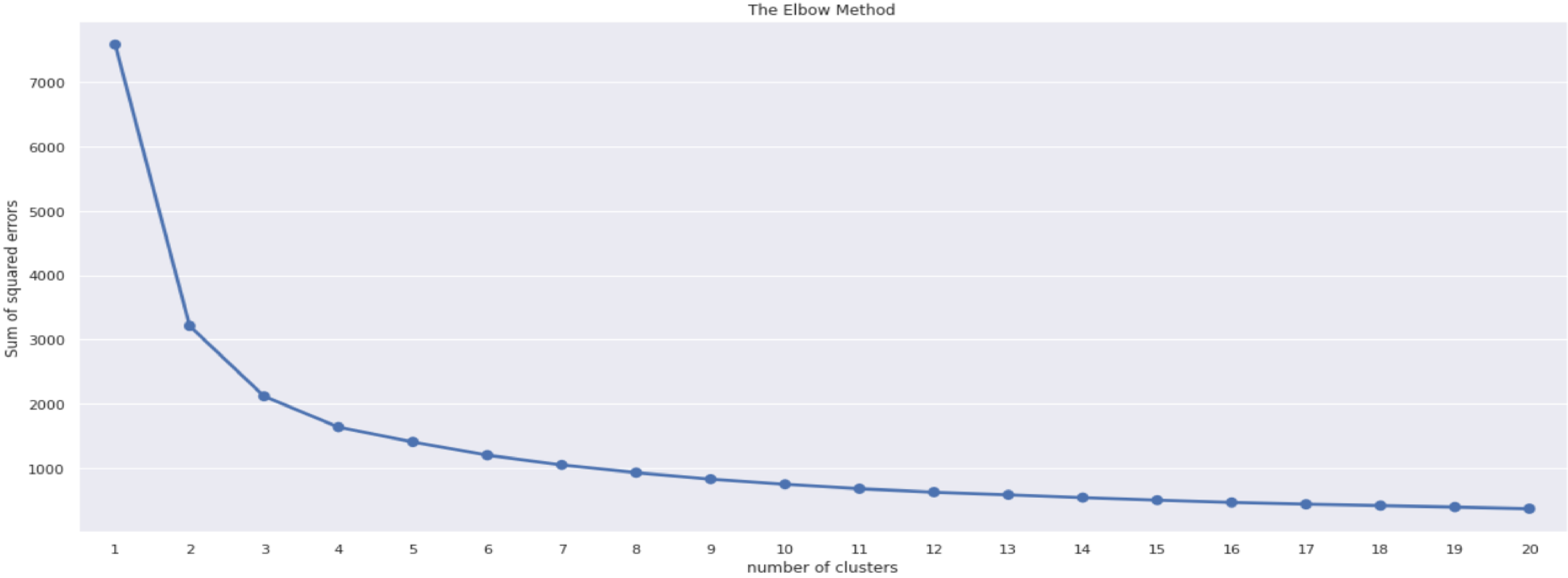
```
# Appling K means Algorithm and checking its silhouette score
range_n_clusters = [2,3,4,5,6,7,8,9,10,11,12,13,14,15]
for n_clusters in range_n_clusters:
    clusterer = KMeans(n_clusters=n_clusters, max_iter=50)
    preds = clusterer.fit_predict(X)
    centers = clusterer.cluster_centers_

    score = silhouette_score(X, preds)
    print("For n_clusters = {}, silhouette score is {}".format(n_clusters, score))
```

```
For n_clusters = 2, silhouette score is 0.4817520125234911
For n_clusters = 3, silhouette score is 0.4088778696097729
For n_clusters = 4, silhouette score is 0.3717633042333461
For n_clusters = 5, silhouette score is 0.3401976376097799
For n_clusters = 6, silhouette score is 0.36487899620146774
For n_clusters = 7, silhouette score is 0.3332324718729326
For n_clusters = 8, silhouette score is 0.3491218811286045
For n_clusters = 9, silhouette score is 0.35964372293309715
For n_clusters = 10, silhouette score is 0.3525843210978691
For n_clusters = 11, silhouette score is 0.3598970485942363
For n_clusters = 12, silhouette score is 0.3686606162154565
For n_clusters = 13, silhouette score is 0.3727470731294531
For n_clusters = 14, silhouette score is 0.35471294115100827
For n_clusters = 15, silhouette score is 0.36308131815173145
```

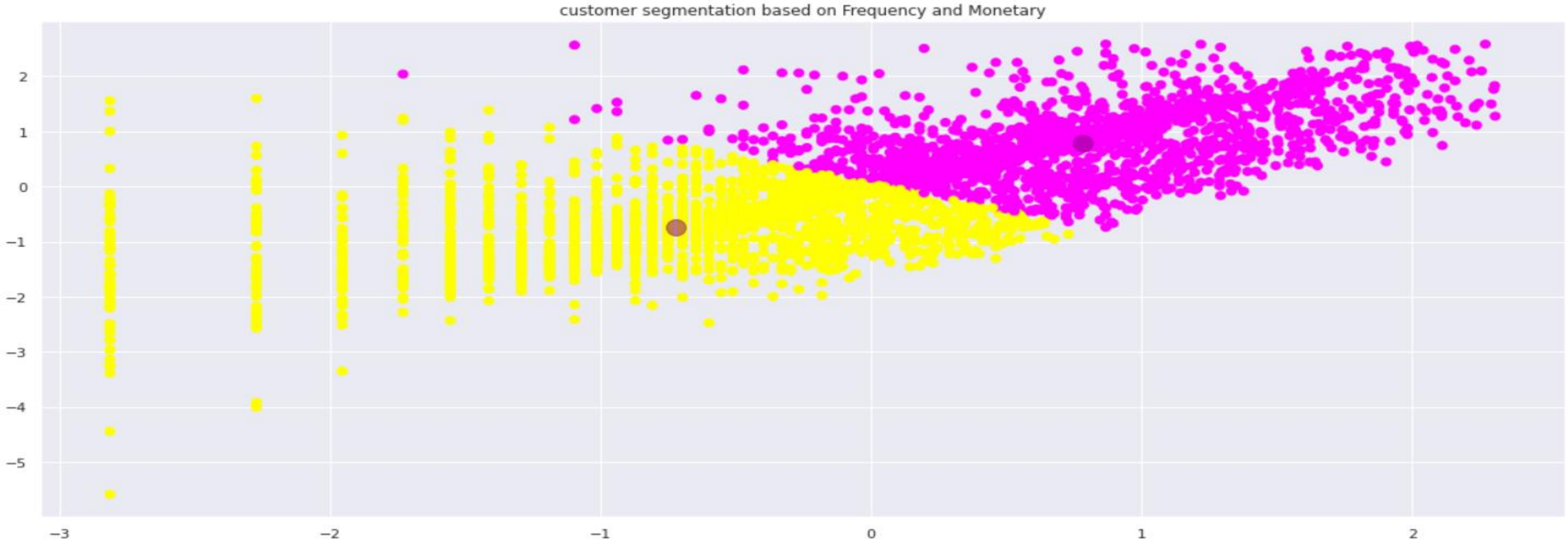
Data Modeling

Elbow method on Frequency and Monetary



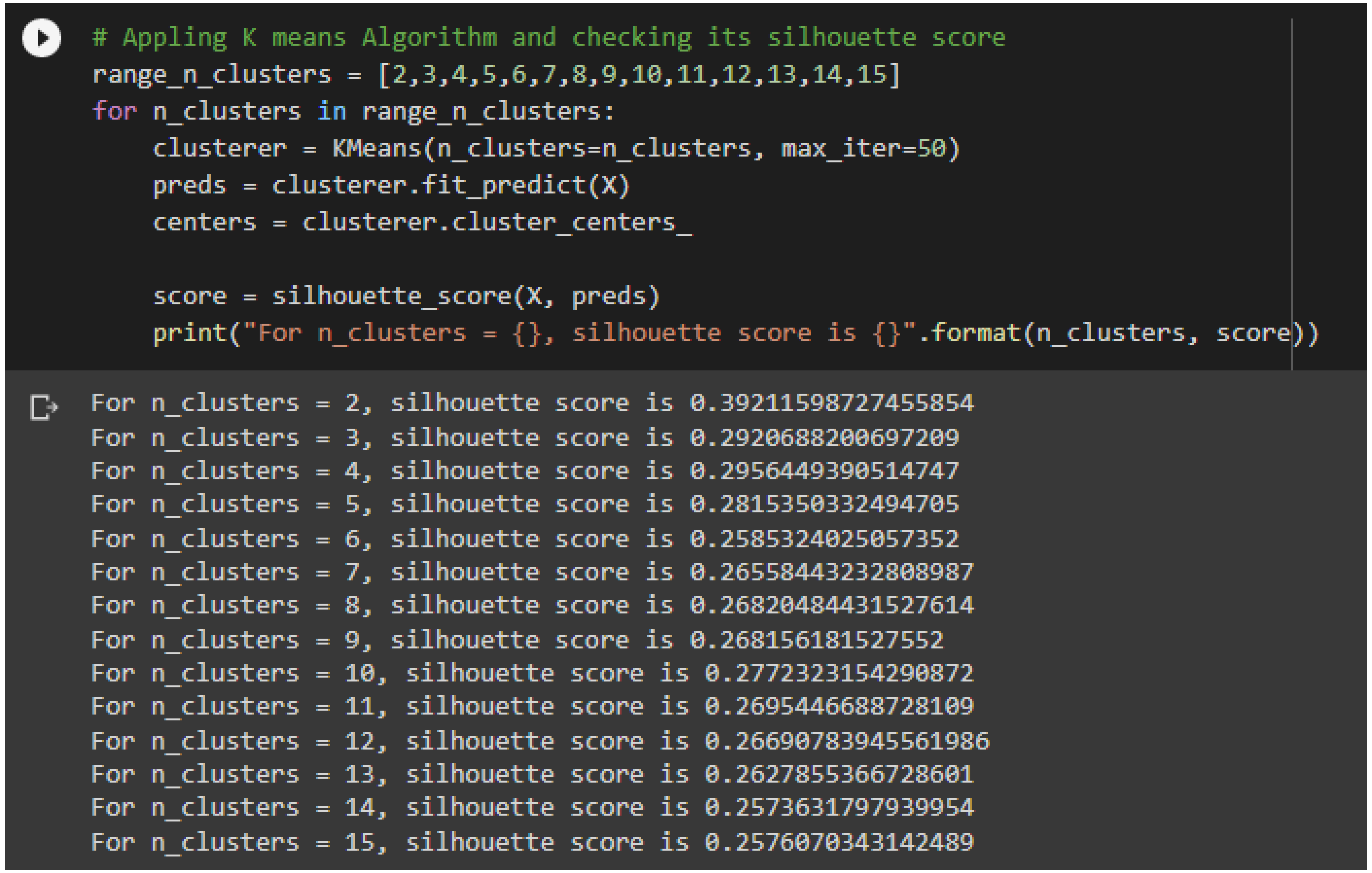
Data Modeling

Customer segmentation based on Frequency and Monetary



Data Modeling

Applying silhouette score method on Recency, Frequency and Monetary



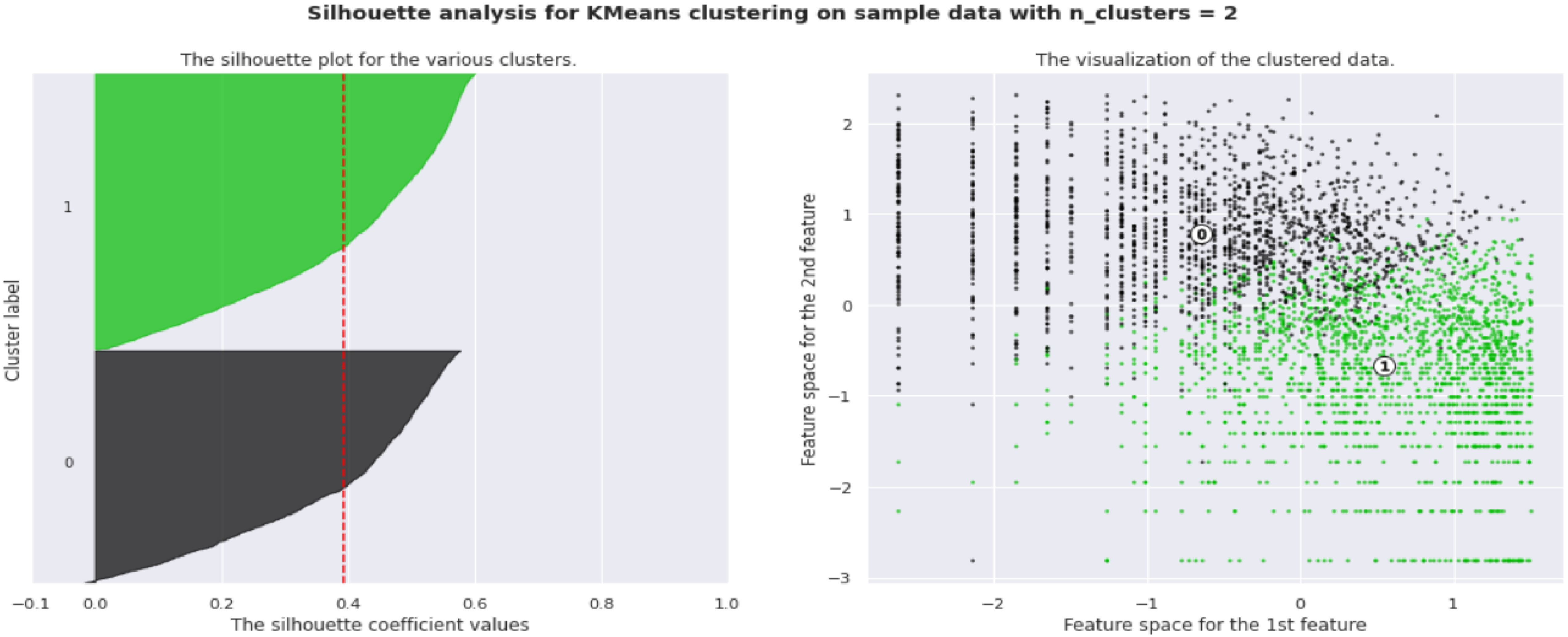
```
# Appling K means Algorithm and checking its silhouette score
range_n_clusters = [2,3,4,5,6,7,8,9,10,11,12,13,14,15]
for n_clusters in range_n_clusters:
    clusterer = KMeans(n_clusters=n_clusters, max_iter=50)
    preds = clusterer.fit_predict(X)
    centers = clusterer.cluster_centers_

    score = silhouette_score(X, preds)
    print("For n_clusters = {}, silhouette score is {}".format(n_clusters, score))
```

```
For n_clusters = 2, silhouette score is 0.39211598727455854
For n_clusters = 3, silhouette score is 0.2920688200697209
For n_clusters = 4, silhouette score is 0.2956449390514747
For n_clusters = 5, silhouette score is 0.2815350332494705
For n_clusters = 6, silhouette score is 0.2585324025057352
For n_clusters = 7, silhouette score is 0.26558443232808987
For n_clusters = 8, silhouette score is 0.26820484431527614
For n_clusters = 9, silhouette score is 0.268156181527552
For n_clusters = 10, silhouette score is 0.2772323154290872
For n_clusters = 11, silhouette score is 0.2695446688728109
For n_clusters = 12, silhouette score is 0.26690783945561986
For n_clusters = 13, silhouette score is 0.2627855366728601
For n_clusters = 14, silhouette score is 0.2573631797939954
For n_clusters = 15, silhouette score is 0.2576070343142489
```

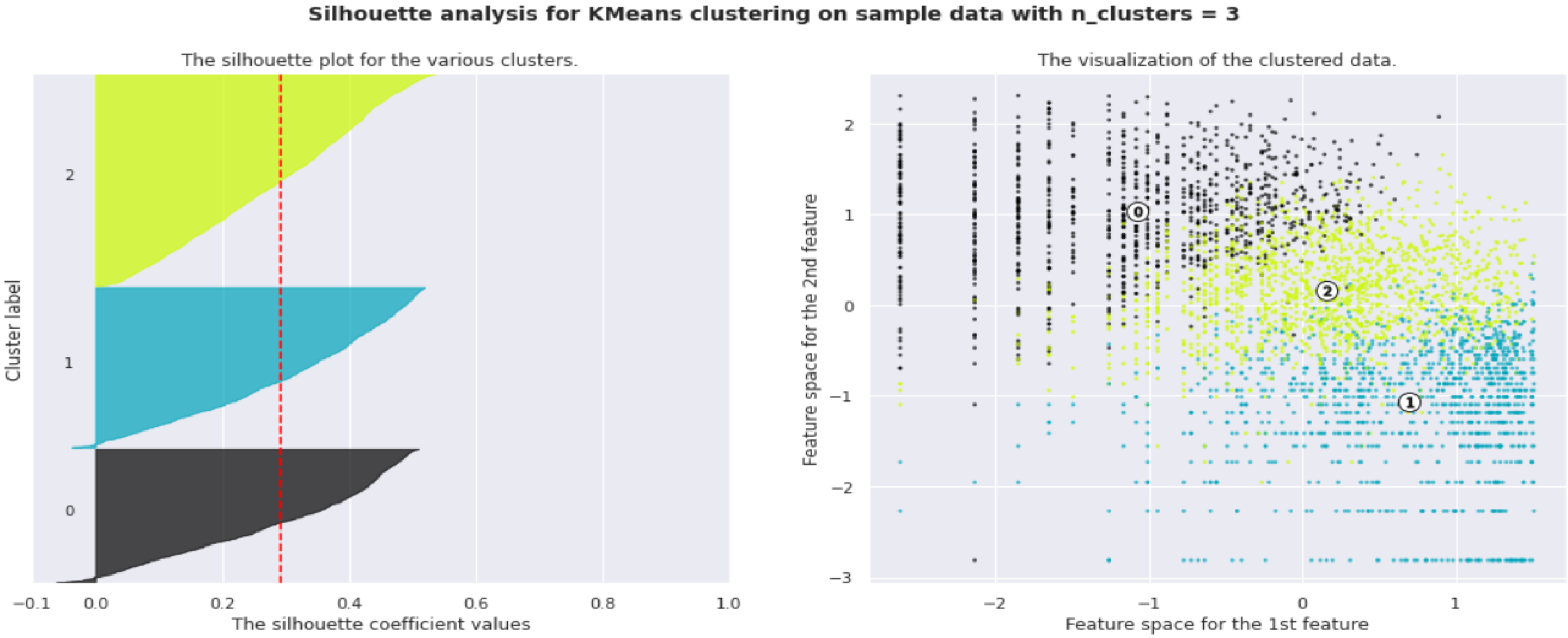
Data Modeling

Applying silhouette score method on Recency, Frequency and Monetary



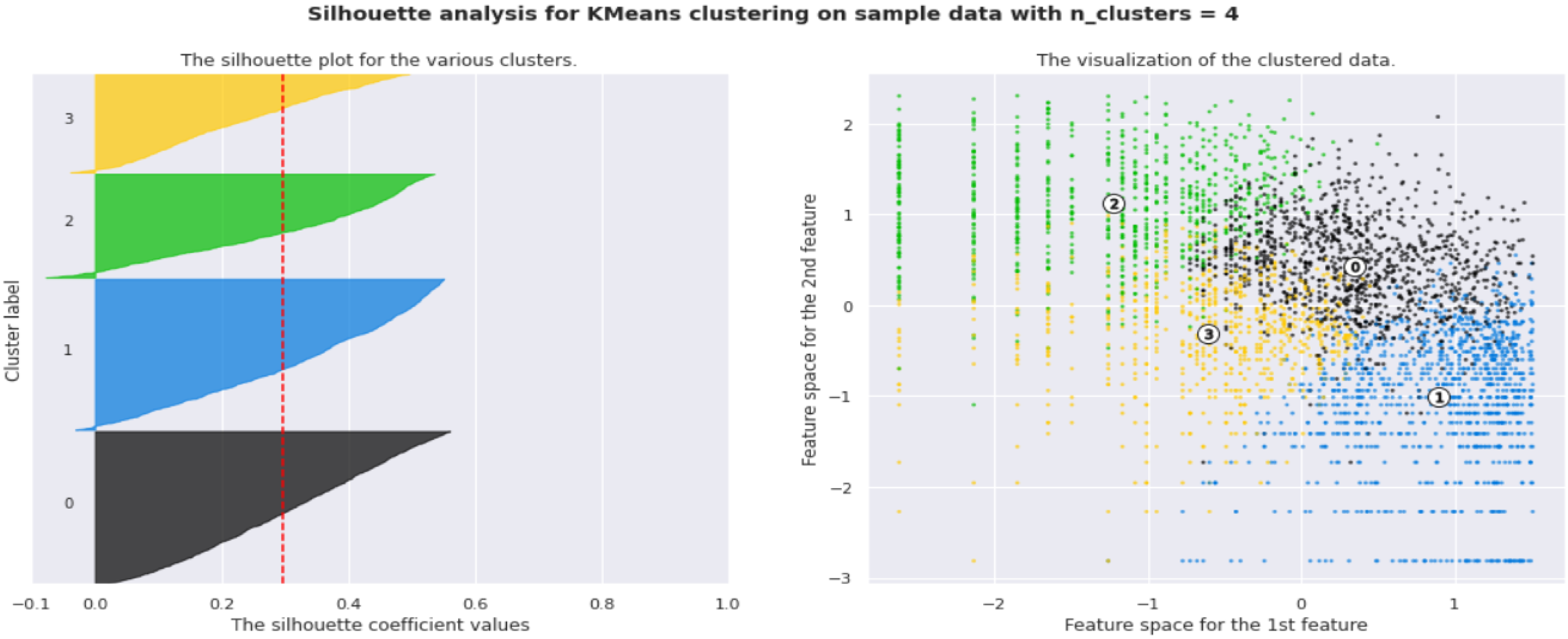
Data Modeling

Applying silhouette score method on Recency, Frequency and Monetary



Data Modeling

Applying silhouette score method on Recency, Frequency and Monetary



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

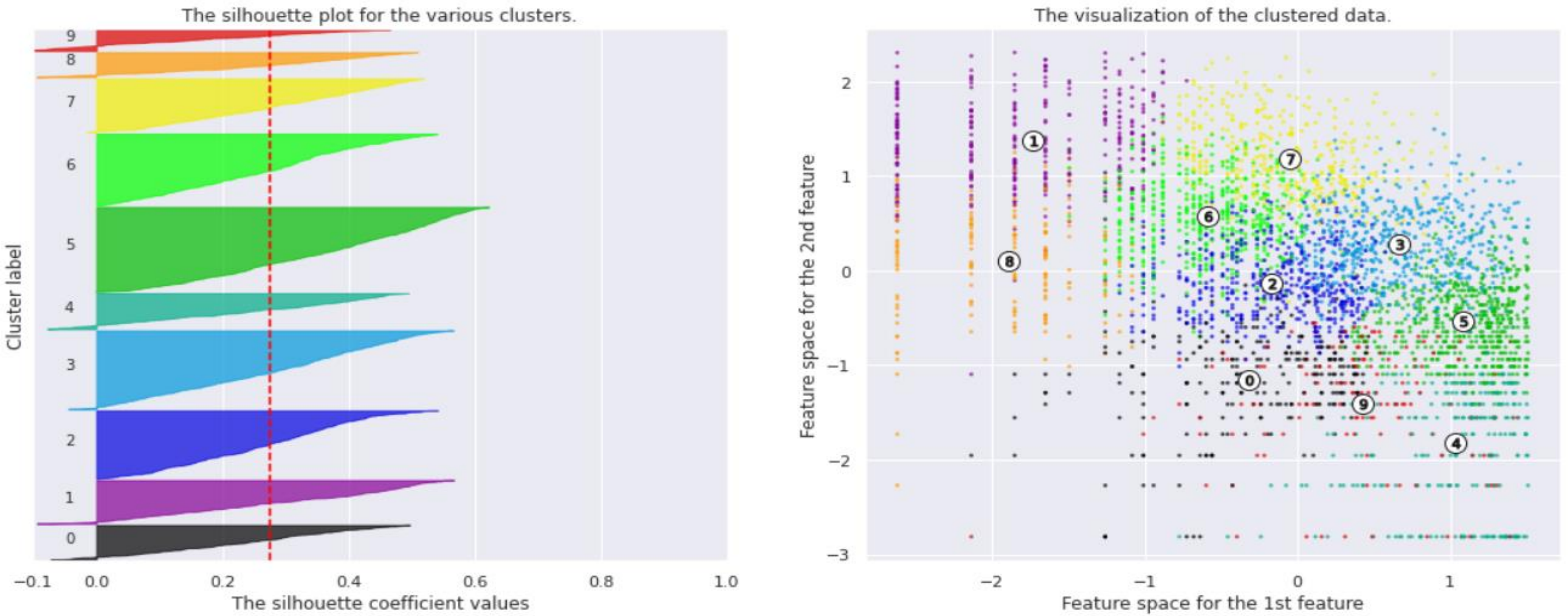
The silhouette plot for the various clusters.

The visualization of the clustered data.

Data Modeling

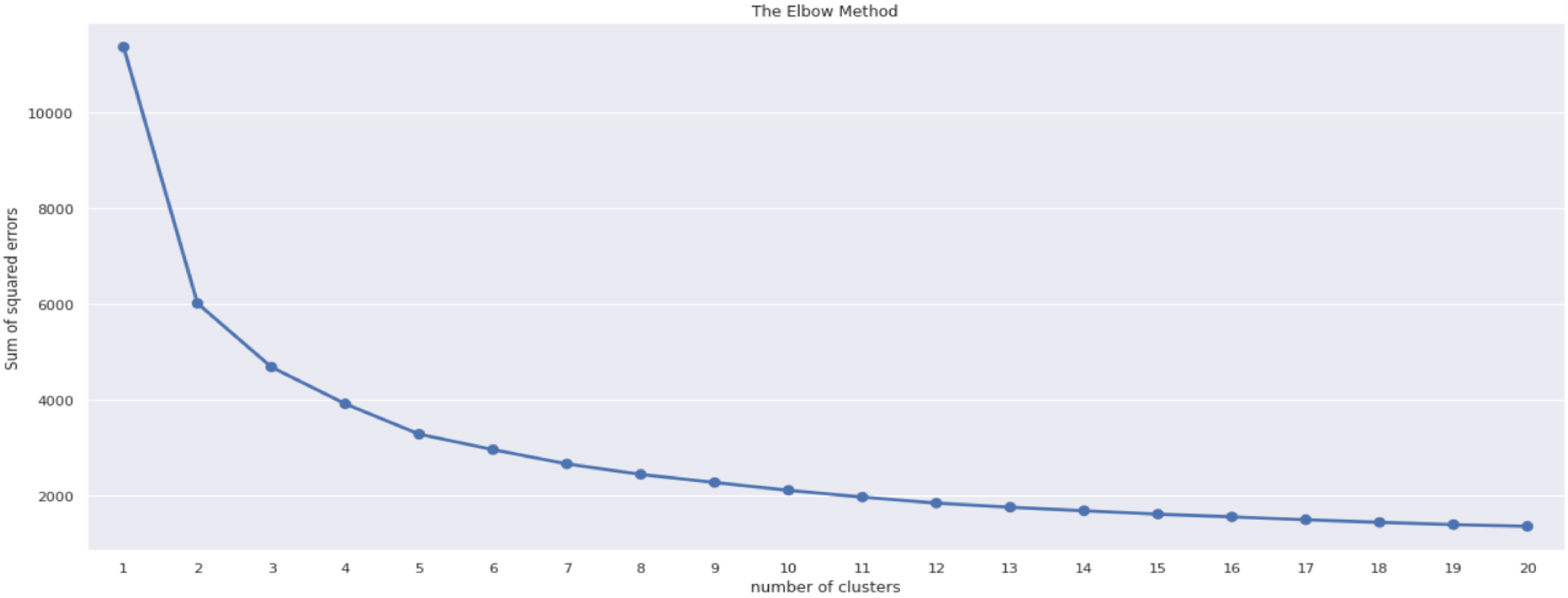
Applying silhouette score method on Recency, Frequency and Monetary

Silhouette analysis for KMeans clustering on sample data with n_clusters = 10



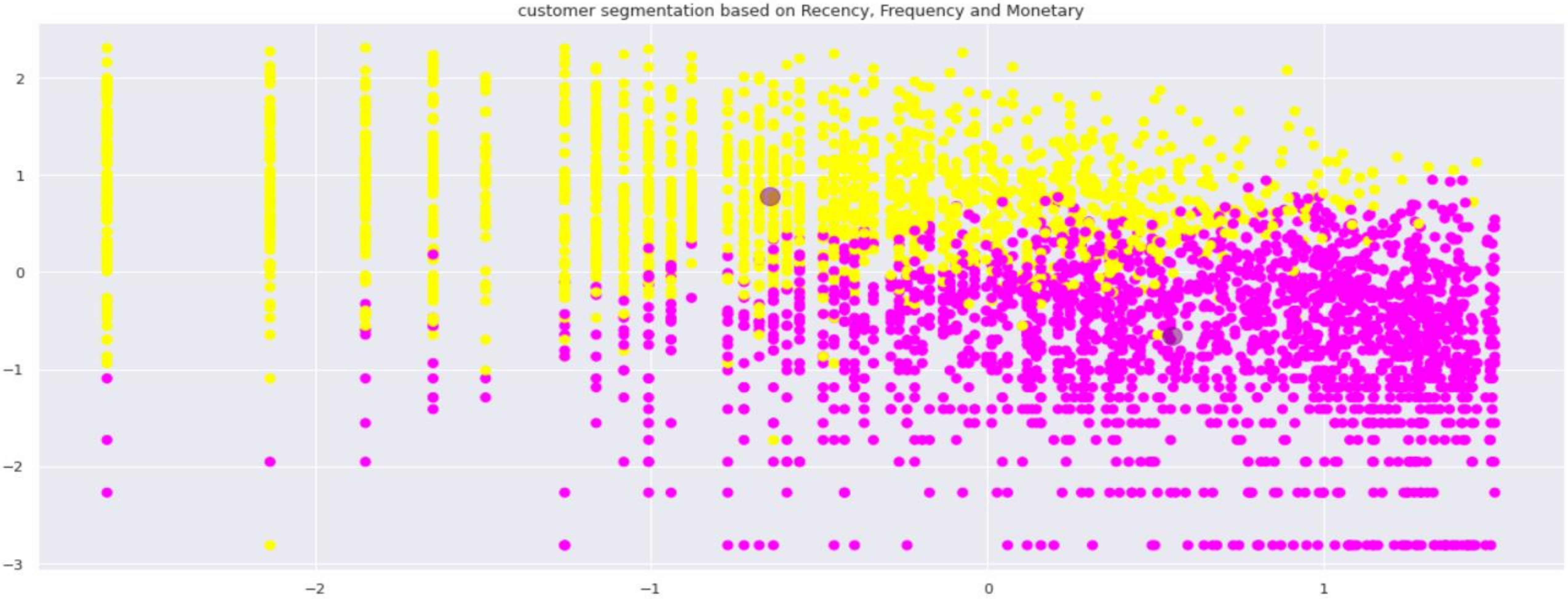
Data Modeling

Applying Elbow method on Recency, Frequency and Monetary



Data Modeling

Customer segmentation based on Recency, Frequency and Monetary



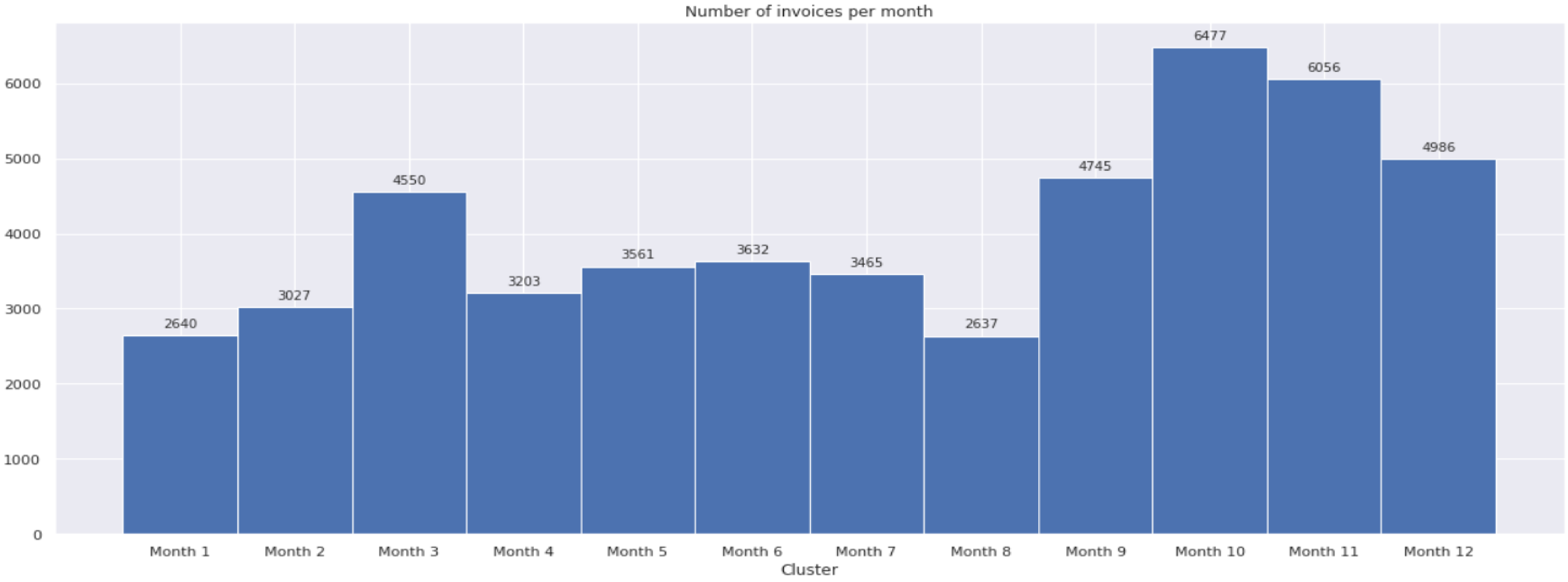
COUNT OF NUMBER OF CUSTOMERS IN EACH CLUSTER

CLUSTER 0 = 1735

CLUSTER 1 = 2055

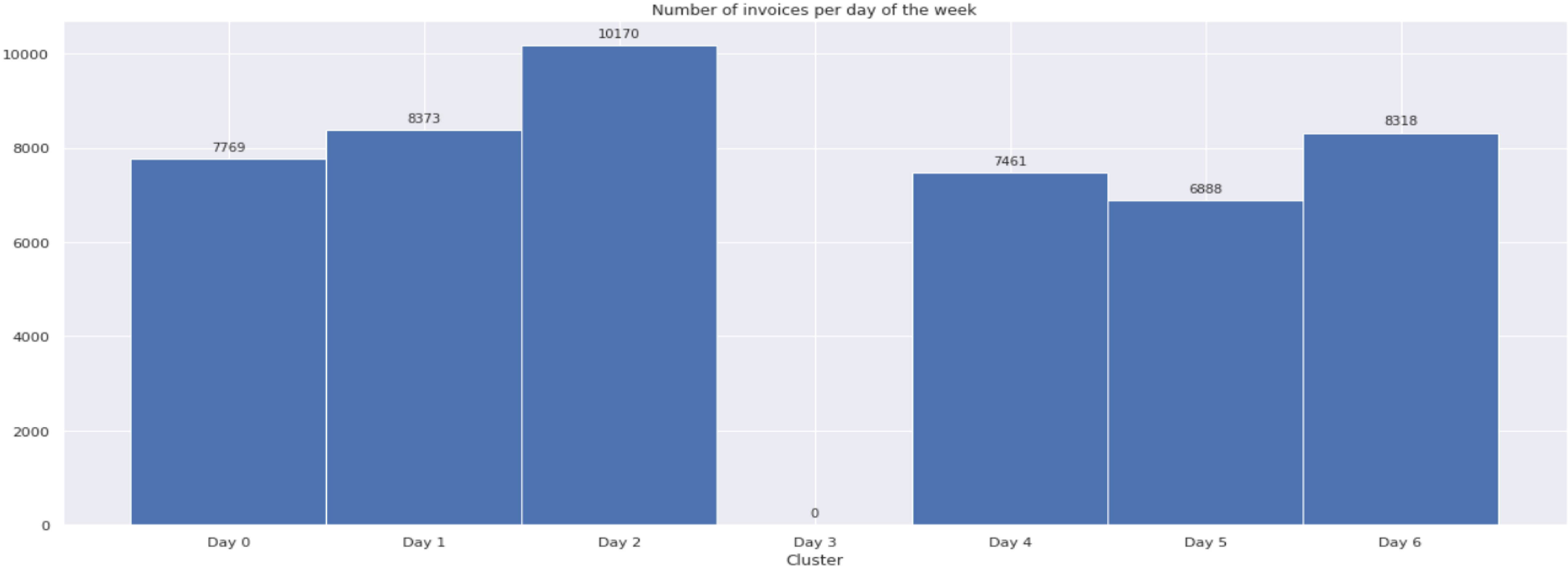
Cluster 0 Analysis

Number of invoices per month



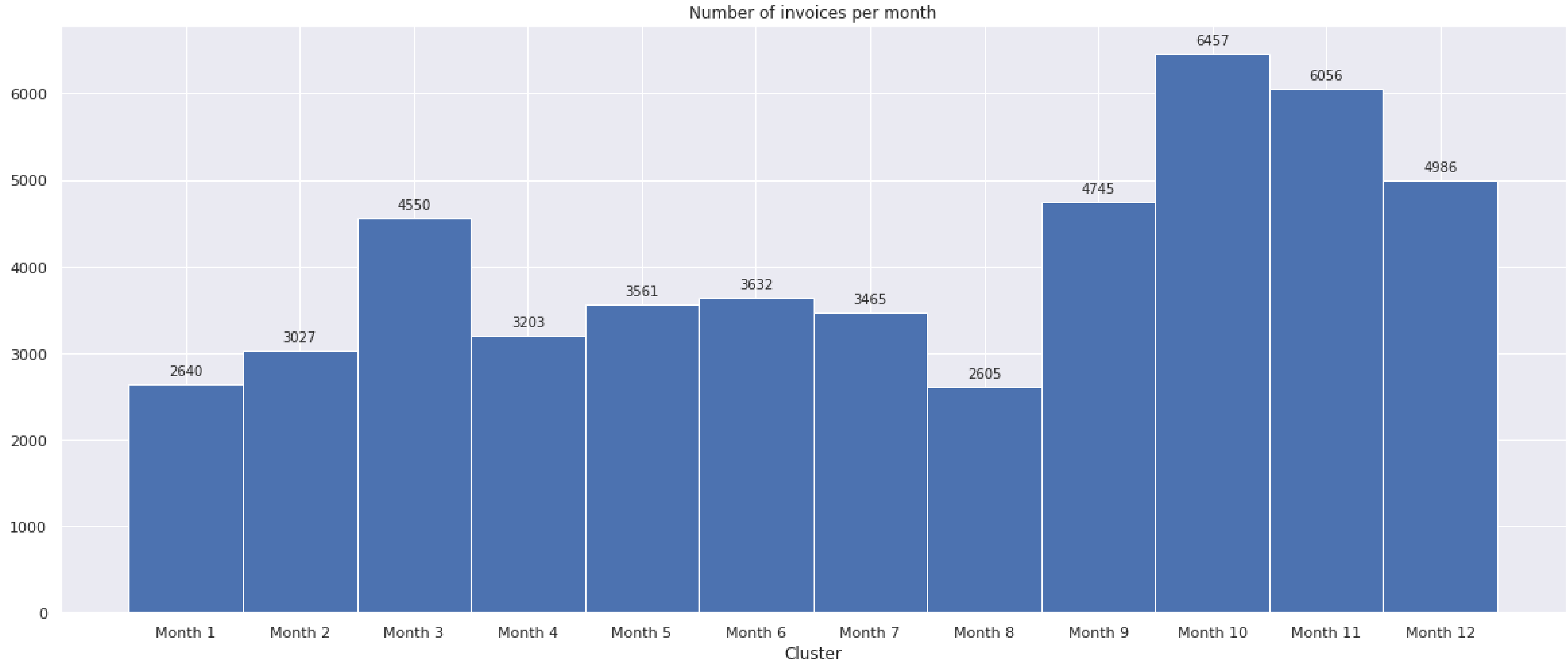
Cluster 0 Analysis

Number of invoices per day of the week



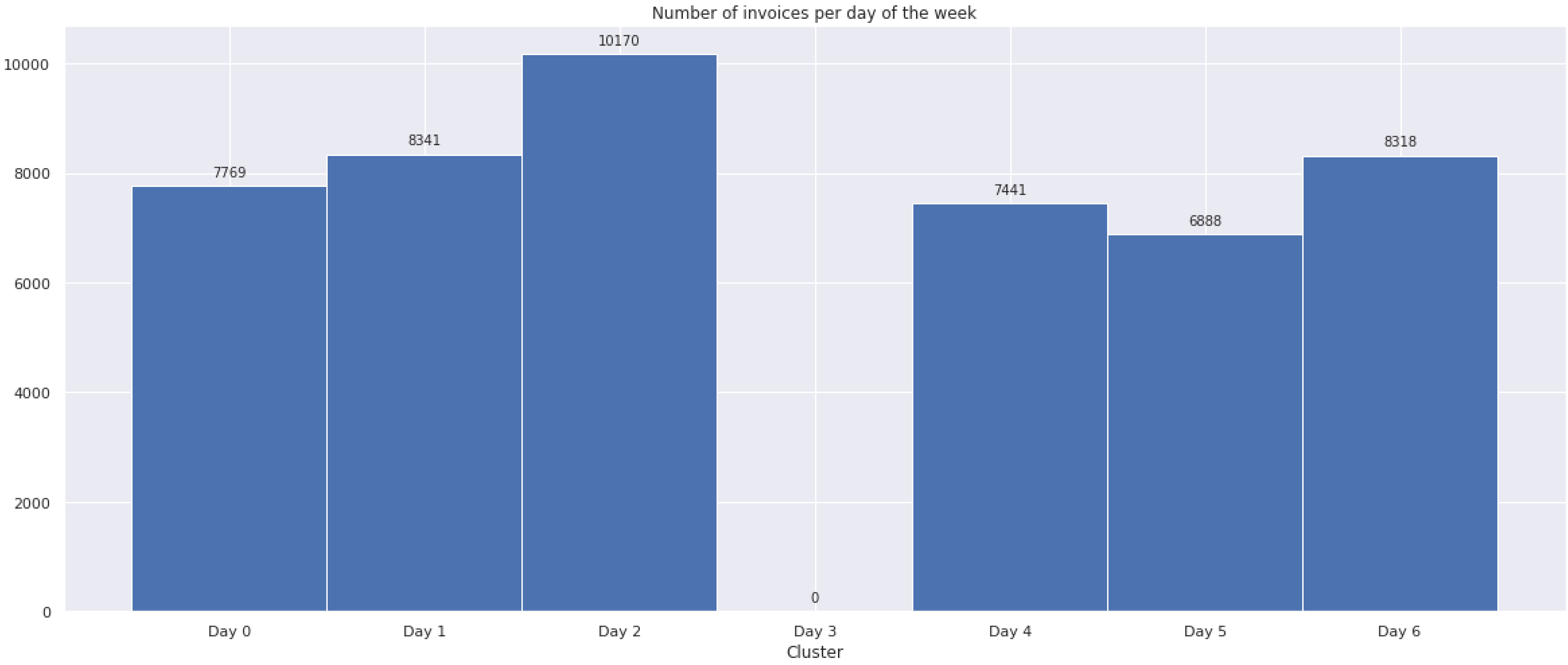
Cluster 1 Analysis

Number of invoices per month



Cluster 1 Analysis

Number of invoices per day of the week



Conclusion

We have got 2 clusters by applying k means algorithm.
So the customers got segmented into 2 clusters.
Online Retail Customer marketing team can now use different approaches to acquire the customers.

Cluster 0

Key Figures

- Frequency : 28.68
- Recency : 230
- Monetary : 3070
- RFM Score : 10.71

Top 5 Products

- WHITE HANGING HEART T-LIGHT HOLDER : 339
- REGENCY CAKESTAND 3 TIER : 268
- ASSORTED COLOUR BIRD ORNAMENT : 235
- PARTY BUNTING : 229
- REX CASH+CARRY JUMBO SHOPPER : 202

Conclusion

Cluster 1

Key Figures

- Frequency : 37.67
- Recency : 134.64
- Monetary : 447.40
- RFM Score : 5.90

Top 5 Products

WHITE HANGING HEART T-LIGHT HOLDER	344
REGENCY CAKESTAND 3 TIER	271
ASSORTED COLOUR BIRD ORNAMENT	239
PARTY BUNTING	232
REX CASH+CARRY JUMBO SHOPPER	204

Thank You!!