



G4-Final

Customer Segmentation Project

G4 Workflow



Nebi

- EDA



İsmail

- RFM



Nursal

- K-Mean



Serdar

- Cohort Analysis

About Data

- An online store data in UK (one year sales information)
- The company mainly sells unique all-occasion gifts
- Many customers of the company are wholesalers
- We have «InvoiceNo» «StockCode» «Description» «Quantity» «InvoiceDate» «UnitPrice» «CustomerID» «Country» information

Dedicates

- ▶ Feature Information
 - ▶ **InvoiceNo**: Invoice number.
 - ▶ **StockCode**: Product (item) code.
 - ▶ **Description**: Product (item) name
 - ▶ **Quantity**: The quantities of each product (item) per transaction
 - ▶ **InvoiceDate**: Invoice Date and time.
 - ▶ **UnitPrice**: Unit price.
 - ▶ **CustomerID**: Customer number.
 - ▶ **Country**: Country name

Project Structures



- ▶ Data Cleaning & Exploratory Data Analysis (EDA)
- ▶ RFM Analysis
- ▶ Customer Segmentation
- ▶ Applying K-Means Clustering
- ▶ Create Cohort and Conduct Cohort Analysis

• Data Cleaning & EDA

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.550	17850.000	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.750	17850.000	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.650	17850.000	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.250	17850.000	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.850	17850.000	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.850	17850.000	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.690	13047.000	United Kingdom

In this table, we see in which structure the data established

Relationships between InvoiceNo, Quantity and UnitPrice

```
df[df["Quantity"]<0][~(df[df["Quantity"]<0]["InvoiceNo"].str.startswith("C")== True)]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2406	536589	21777	NaN	-10	2010-12-01 16:50:00	0.000	NaN	United Kingdom
4347	536764	84952C	NaN	-38	2010-12-02 14:42:00	0.000	NaN	United Kingdom
7188	536996	22712	NaN	-20	2010-12-03 15:30:00	0.000	NaN	United Kingdom
7189	536997	22028	NaN	-20	2010-12-03 15:30:00	0.000	NaN	United Kingdom
7190	536998	85067	NaN	-6	2010-12-03 15:30:00	0.000	NaN	United Kingdom
...
535333	581210	23395	check	-26	2011-12-07 18:36:00	0.000	NaN	United Kingdom
535335	581212	22578	lost	-1050	2011-12-07 18:38:00	0.000	NaN	United Kingdom
535336	581213	22576	check	-30	2011-12-07 18:38:00	0.000	NaN	United Kingdom
536908	581226	23090	missing	-338	2011-12-08 09:56:00	0.000	NaN	United Kingdom
538919	581422	23169	smashed	-235	2011-12-08 15:24:00	0.000	NaN	United Kingdom

1336 rows x 8 columns

We know that invoice numbers starting with C are canceled orders, but there are also negative transactions on some lines even though this expression is not present.

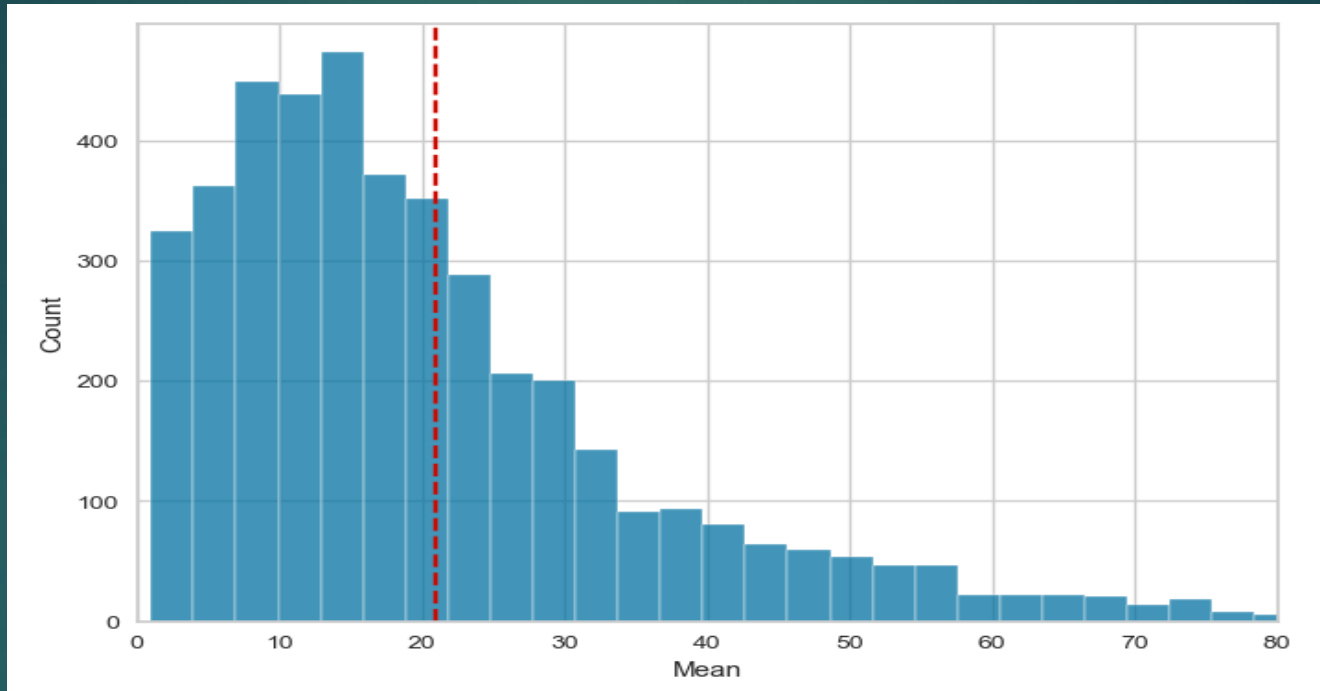
•Handling Missing Values

```
for i in df.columns:  
    print(f"Null values in {i} column is = %", round(df[i].isnull().sum()/df.shape[0]*100,2))
```

```
Null values in InvoiceNo column is = % 0.0  
Null values in StockCode column is = % 0.0  
Null values in Description column is = % 0.0  
Null values in Quantity column is = % 0.0  
Null values in InvoiceDate column is = % 0.0  
Null values in UnitPrice column is = % 0.0  
Null values in CustomerID column is = % 24.82  
Null values in Country column is = % 0.0
```

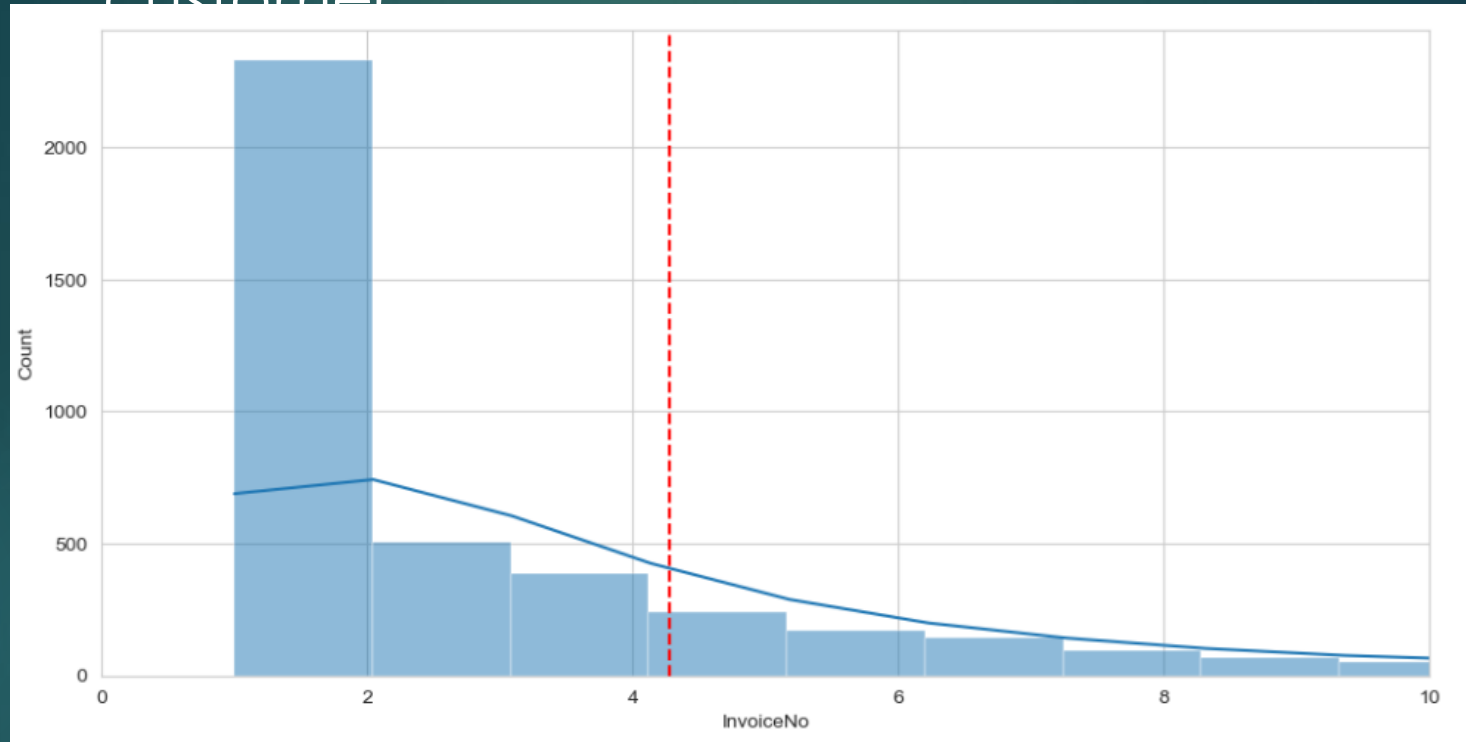
When we examine the null values, we see a significant ratio in the CustomerID column, these values indicate that there are purchases made without membering. We update our data by dropping these records.

- Explore the Orders
 - The number of unique products per customer



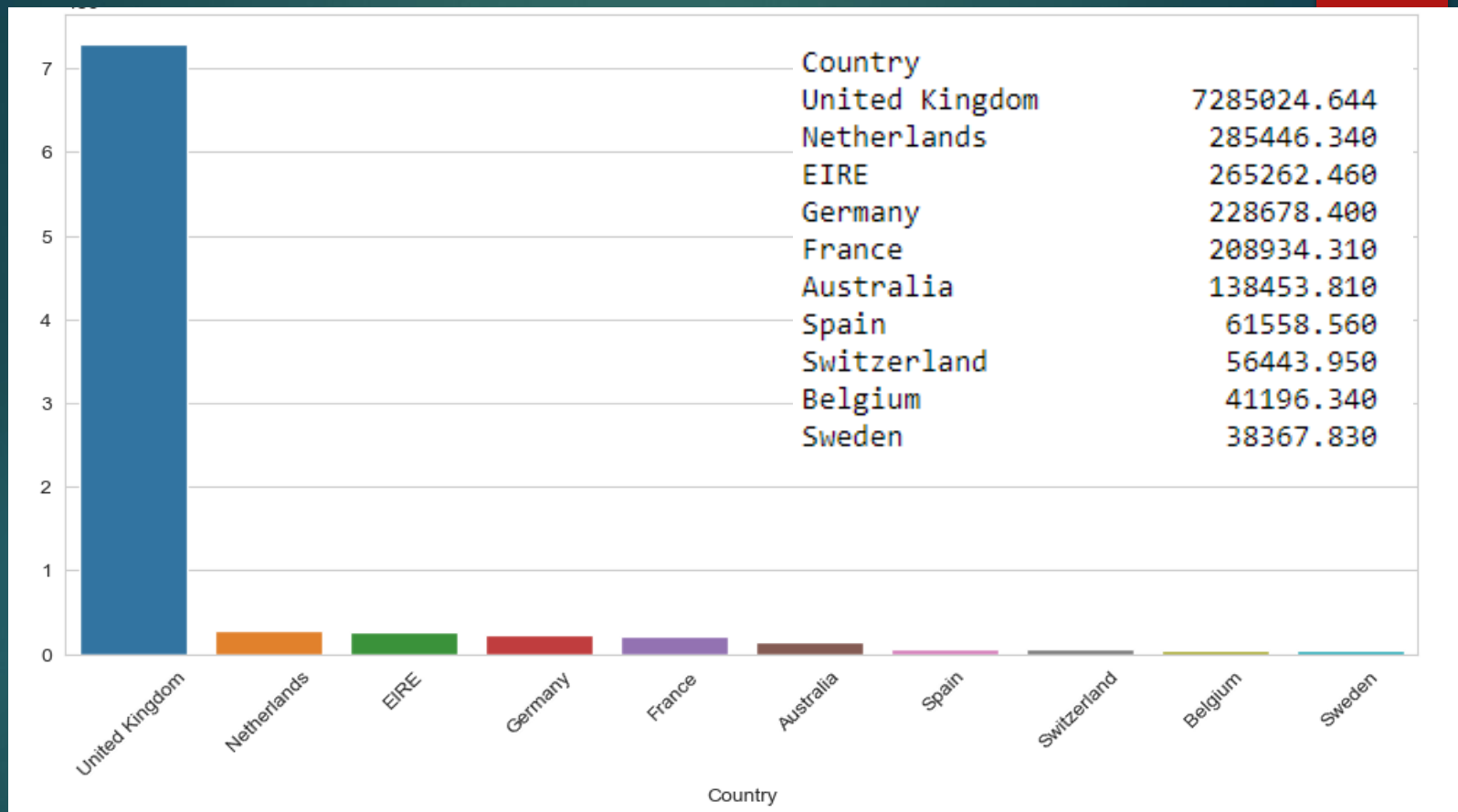
When examining how many types of products the customers buy, it is observed that 21 types of products are bought on average and the density is between 8-15.

- Average number of unique InvoiceNo per customer

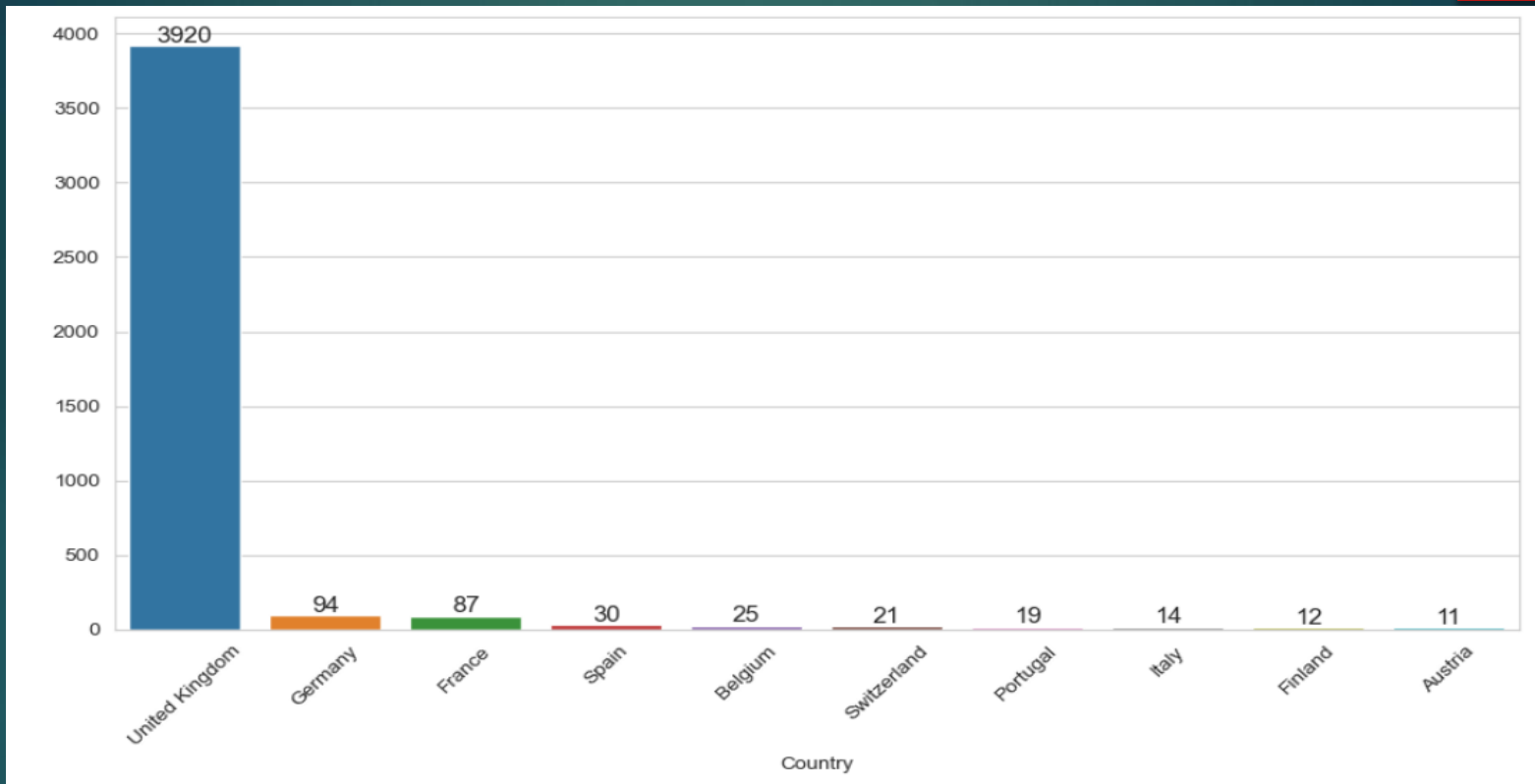


In this table, it is seen that there are too many customers who shop only once, on average, customers shop 4 times.

•Explore Customers by Country

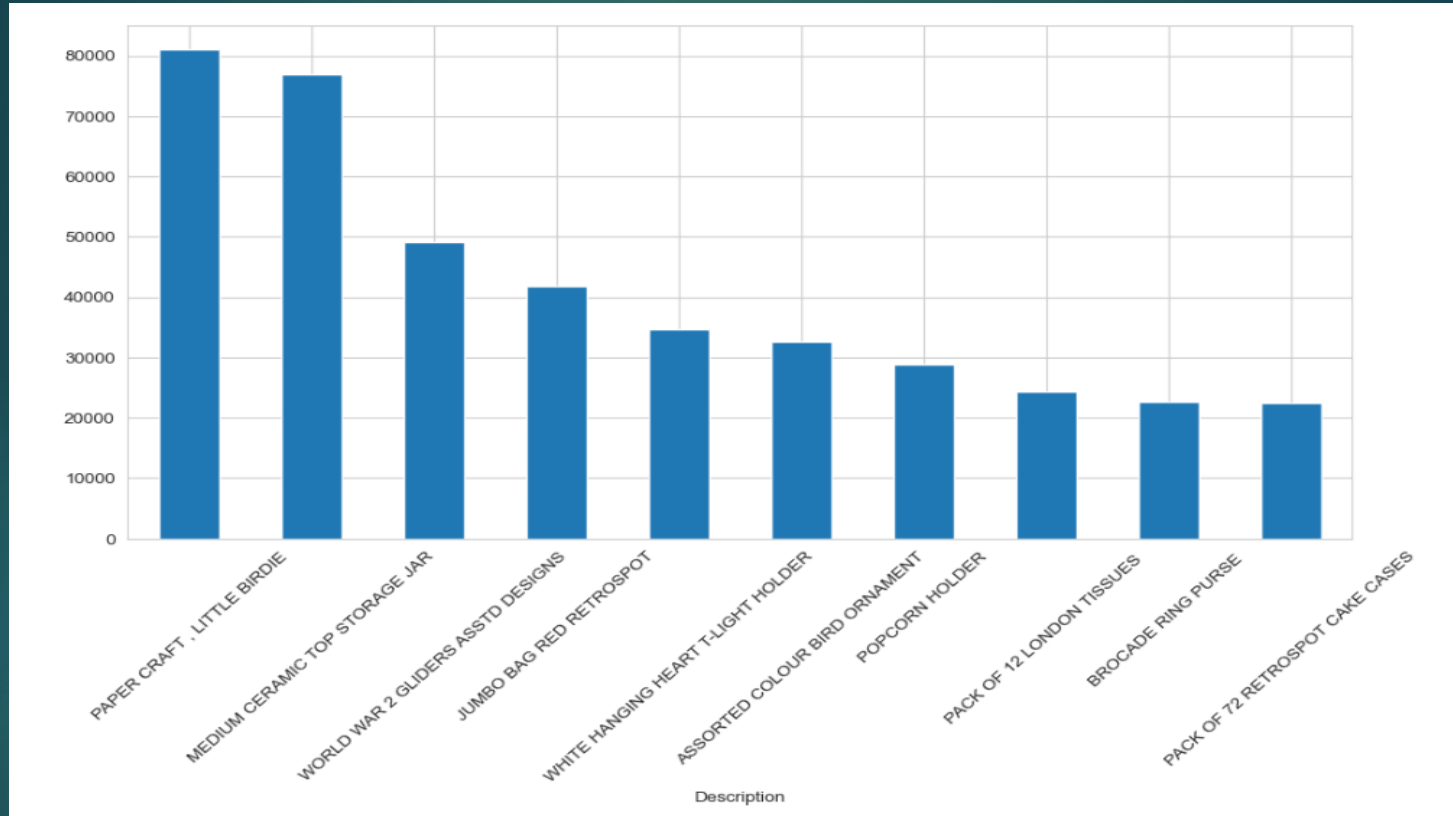


•Explore Customers by Country



We see that the most of the customers are UK based according to total revenue and customer count so we will drop other country purchases.

- Explore the UK Market



Top products of the company that sells souvenirs in general

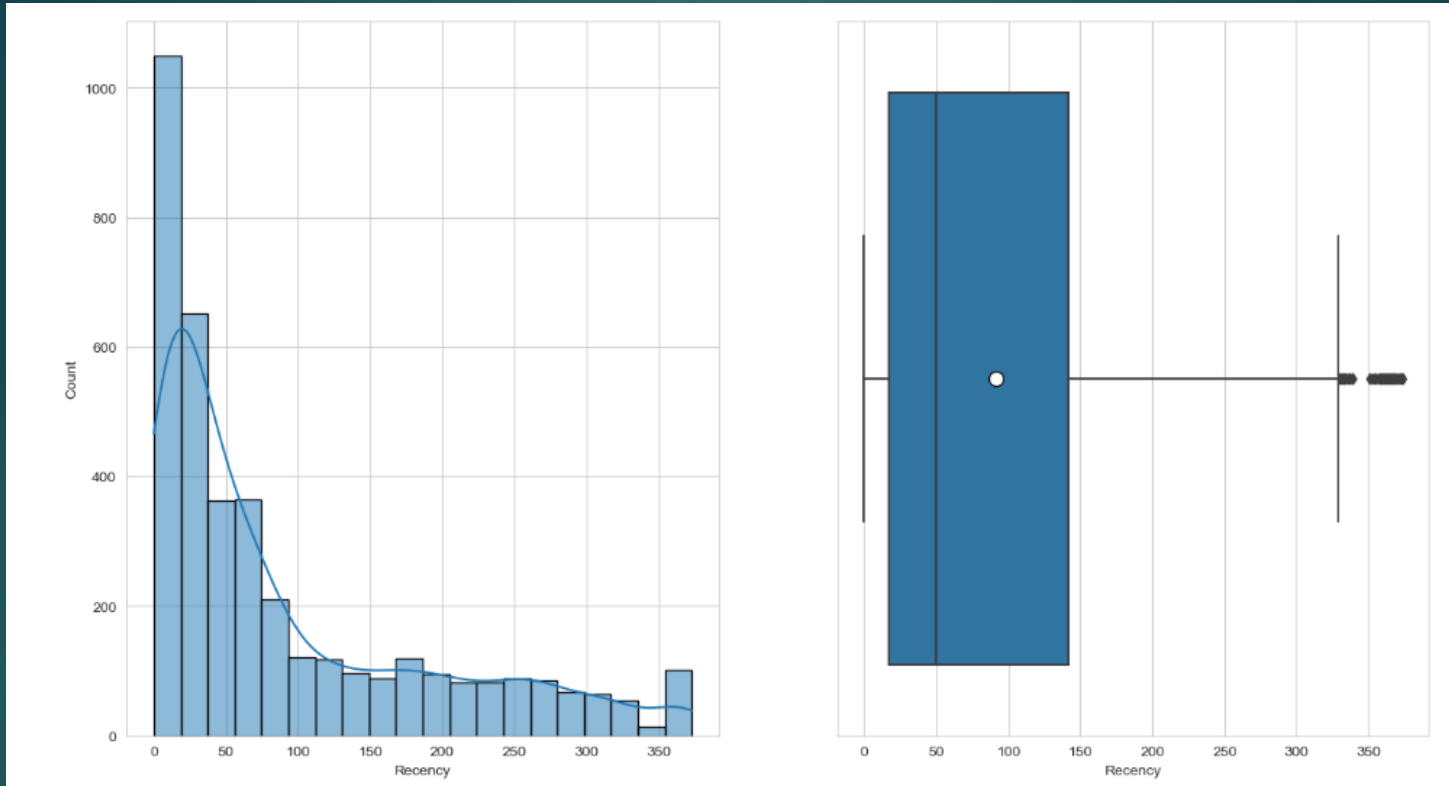
•RFM (Recency, Frequency, Monetary) Analysis

- RECENCY (R): Time since last purchase
- FREQUENCY (F): Total number of purchases
- MONETARY VALUE (M): Total monetary value

•We will try to find these questions' answers

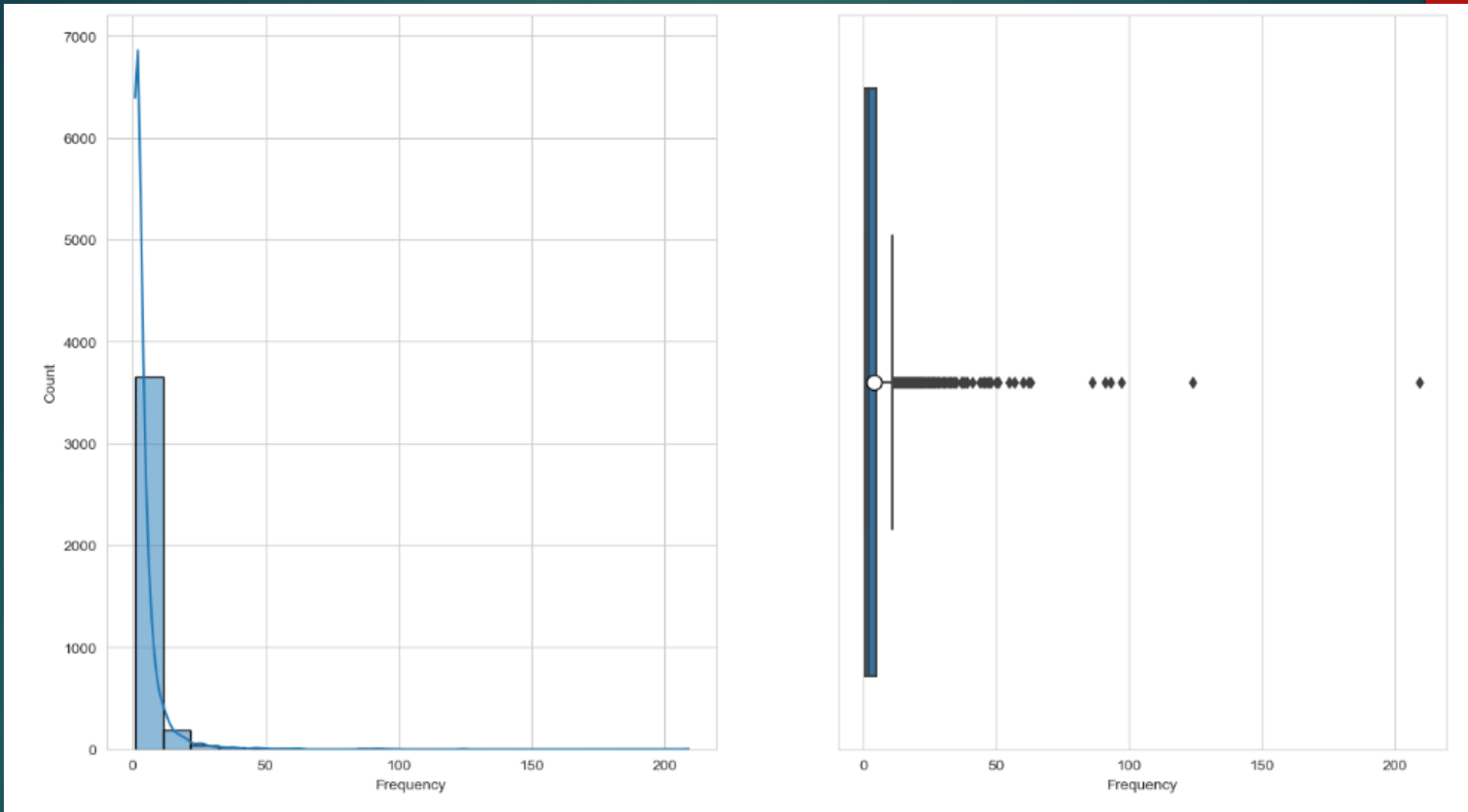
- Who are our best customers?
- Who has the potential to be converted into more profitable customers?
- Which customers do we need to retain?
- Which group of customers is most likely to respond to our marketing campaign?

- Recency: Days since last purchase



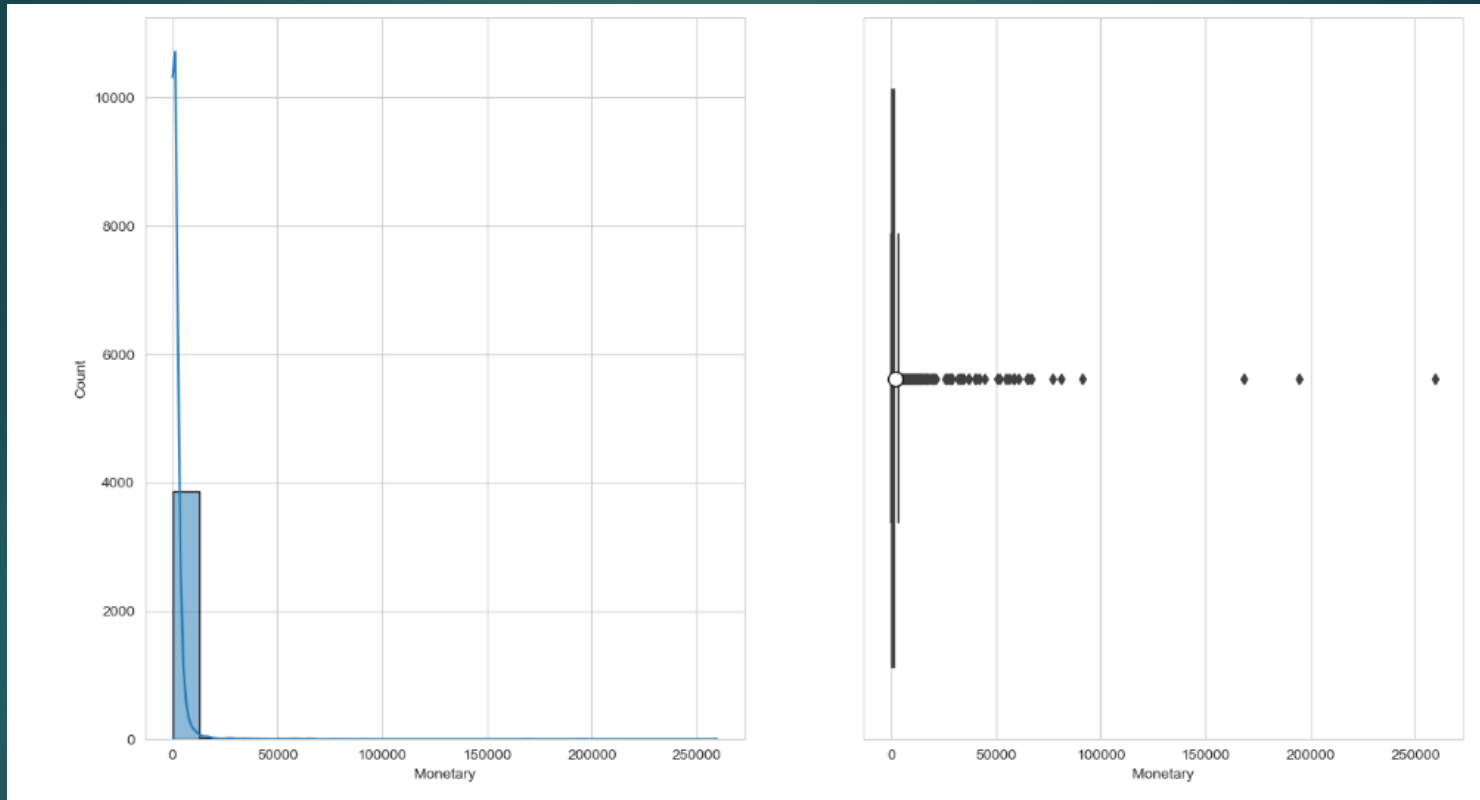
Top products of the company that sells souvenirs in general

- Frequency: Number of purchases



When the frequency of visits of the customers is examined, it is seen that the weight is concentrated in low numbers.

- Monetary: Total amount of money spent



When we calculate the total spending of customers, we see a concentration of under £1000.

•Create RFM Table

- Merge the recency, frequency and monetary dataframes

We combined the calculated columns and created a new dataframe.

	CustomerID	Recency	Frequency	Monetary
0	12346	325	1	77183.600
1	12747	2	11	4196.010
2	12748	0	209	33053.190
3	12749	3	5	4090.880
4	12820	3	4	942.340

We divide our new futures into quartiles and classify them

	Recency	Frequency	Monetary
0.250	17.000	1.000	298.185
0.500	50.000	2.000	644.975
0.750	142.000	5.000	1571.285

•Creating the RFM Segmentation Table

	CustomerID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score
0	12346	325	1	77183.600	1	1	4
1	12747	2	11	4196.010	4	4	4
2	12748	0	209	33053.190	4	4	4
3	12749	3	5	4090.880	4	3	4
4	12820	3	4	942.340	4	3	3
...
3915	18280	277	1	180.600	1	1	1
3916	18281	180	1	80.820	1	1	1
3917	18282	7	2	178.050	4	2	1
3918	18283	3	16	2045.530	4	4	4
3919	18287	42	3	1837.280	3	3	4

3920 rows x 7 columns

We created a new table by assigning points to the classes we created.

•Creating the RFM Segmentation Table

```
#segment4
# Best_Customers, who bought most recently, most often, and are heavy spenders.
# Loyal_Customers customers with high frequency, high recency and average monetary.
# New Customers are your customers high recency, no frequency average monetary
# At Risk Customers are your customers spent average amounts, but haven't purchased recently and not frequently.

def rfm_level4(RFM) :
    if (RFM[0] in ["3","4"]) and (RFM[1] == "4") and (RFM[2] in ["3", "4"]):
        return "Best_Customers"
    elif (RFM[0] in ["3","4"]) and (RFM[1] in ["3", "4"]) and (RFM[2] in ["1","2","3", "4"]):
        return "Loyal_Customers"
    elif (RFM[0] in ["3","4"]) and (RFM[1] in ["1", "2"]) and (RFM[2] in ["1","2","3", "4"]):
        return "New_Customers"
    else:
        return "At_Risk_Customers"
```

Using our business knowledge, we created class names that would correspond to customers' scores.

•RFM Segmentation Table

	Segment4	At_Risk_Customers	Best_Customers	Loyal_Customers	New_Customers
Recency	max	373.000	50.000	50.000	50.000
	min	51.000	0.000	0.000	0.000
	mean	164.940	13.751	20.439	24.293
Frequency	max	34.000	209.000	7.000	2.000
	min	1.000	6.000	3.000	1.000
	mean	2.151	13.301	3.865	1.507
Monetary	max	77183.600	259657.300	12393.700	168472.500
	min	3.750	716.000	36.560	6.900
	mean	791.860	6555.720	1302.944	717.264
	size	1948.000	678.000	631.000	663.000

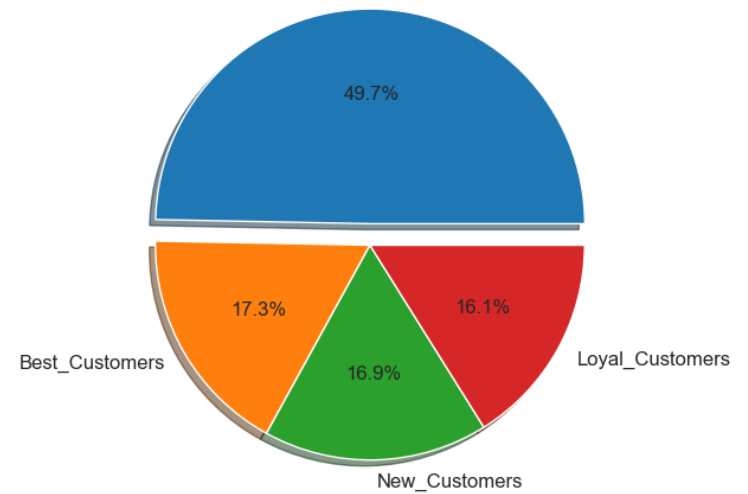
When the min, max, mean values of these classes are examined, it is seen that the classes are generally assigned correctly.

•Customer distribution by segmentation

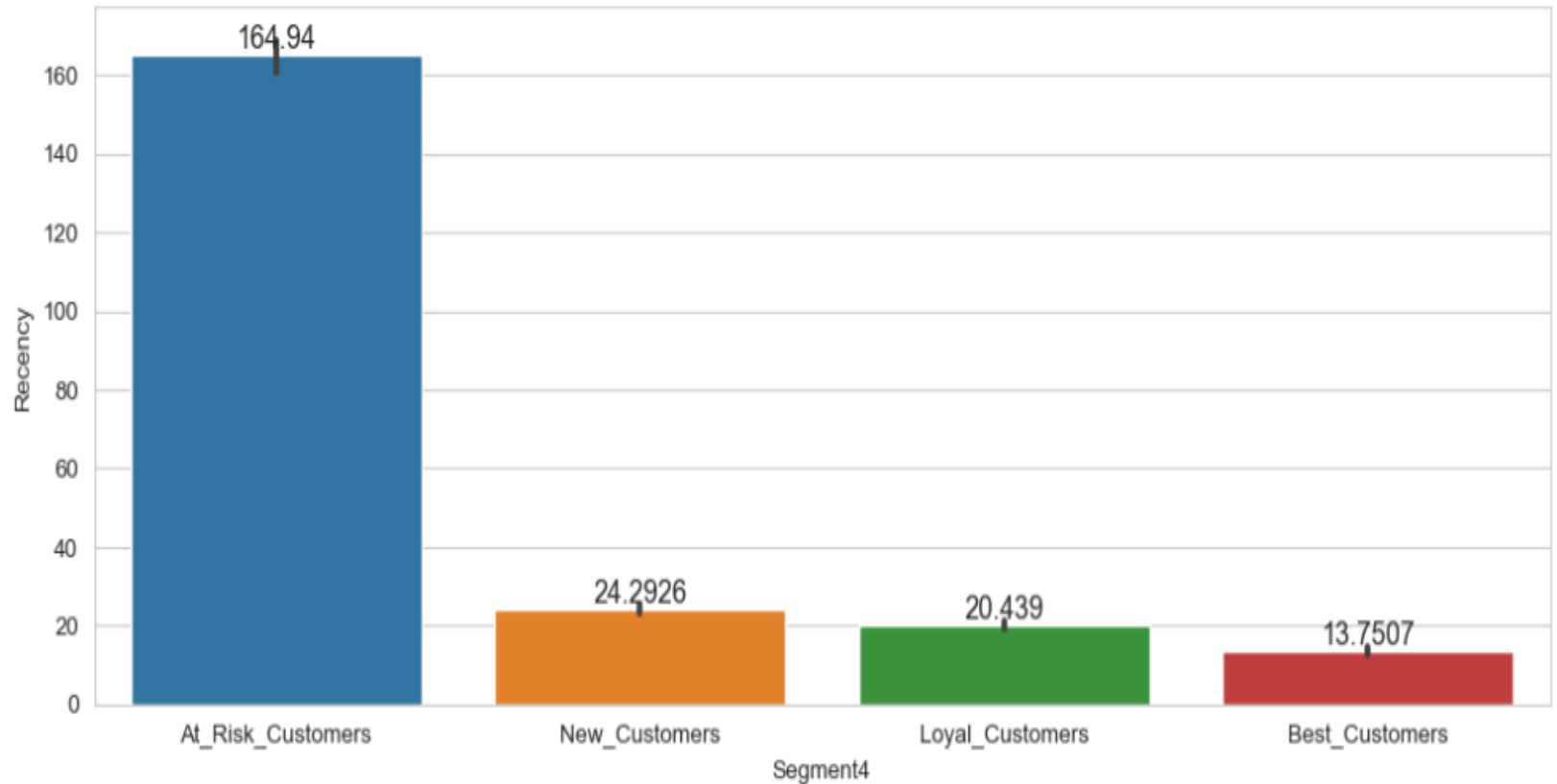
RFM Segment4



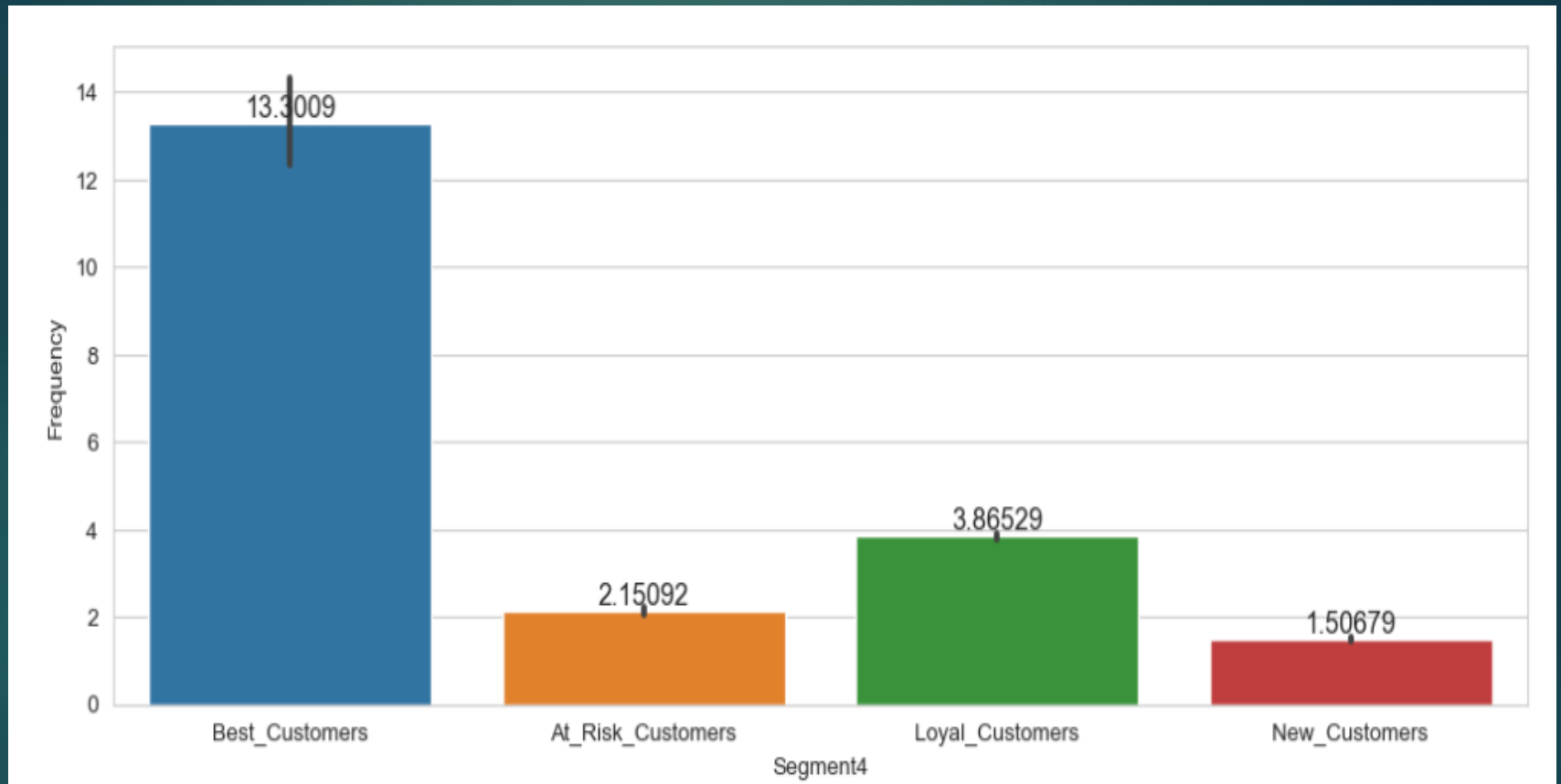
RFM 4 Customer Segments
At_Risk_Customers



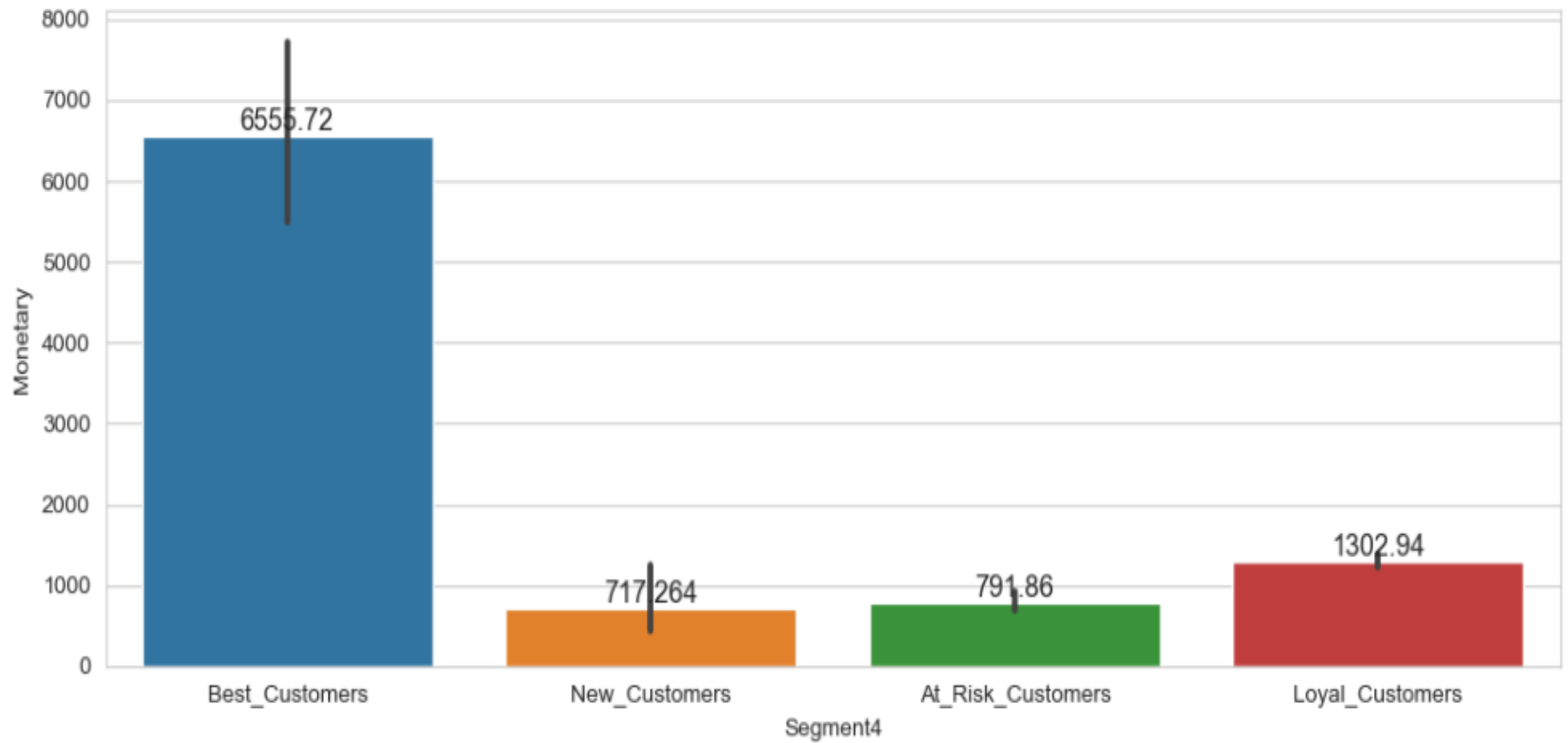
•Plot RFM Segments-Recency



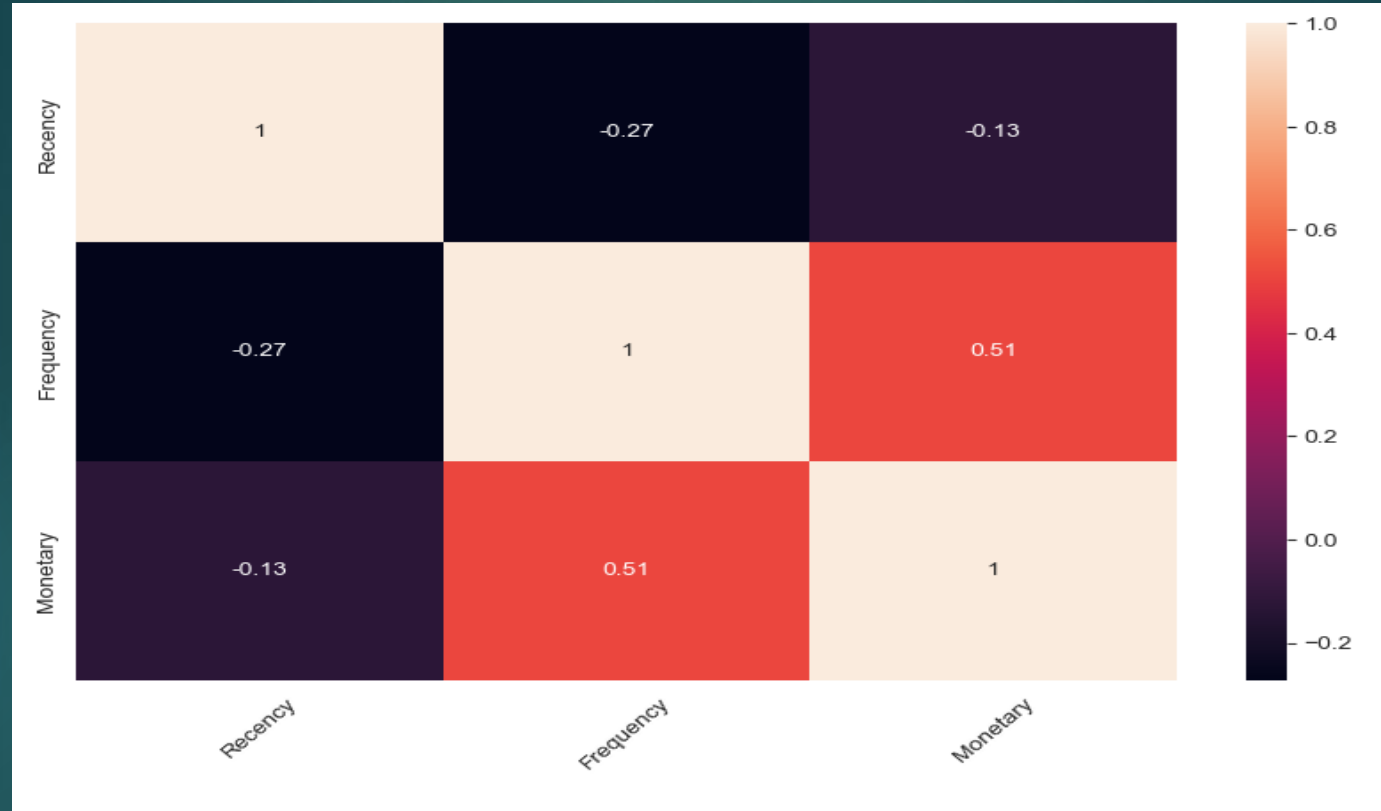
•Plot RFM Segments-Frequency



•Plot RFM Segments-Monetary

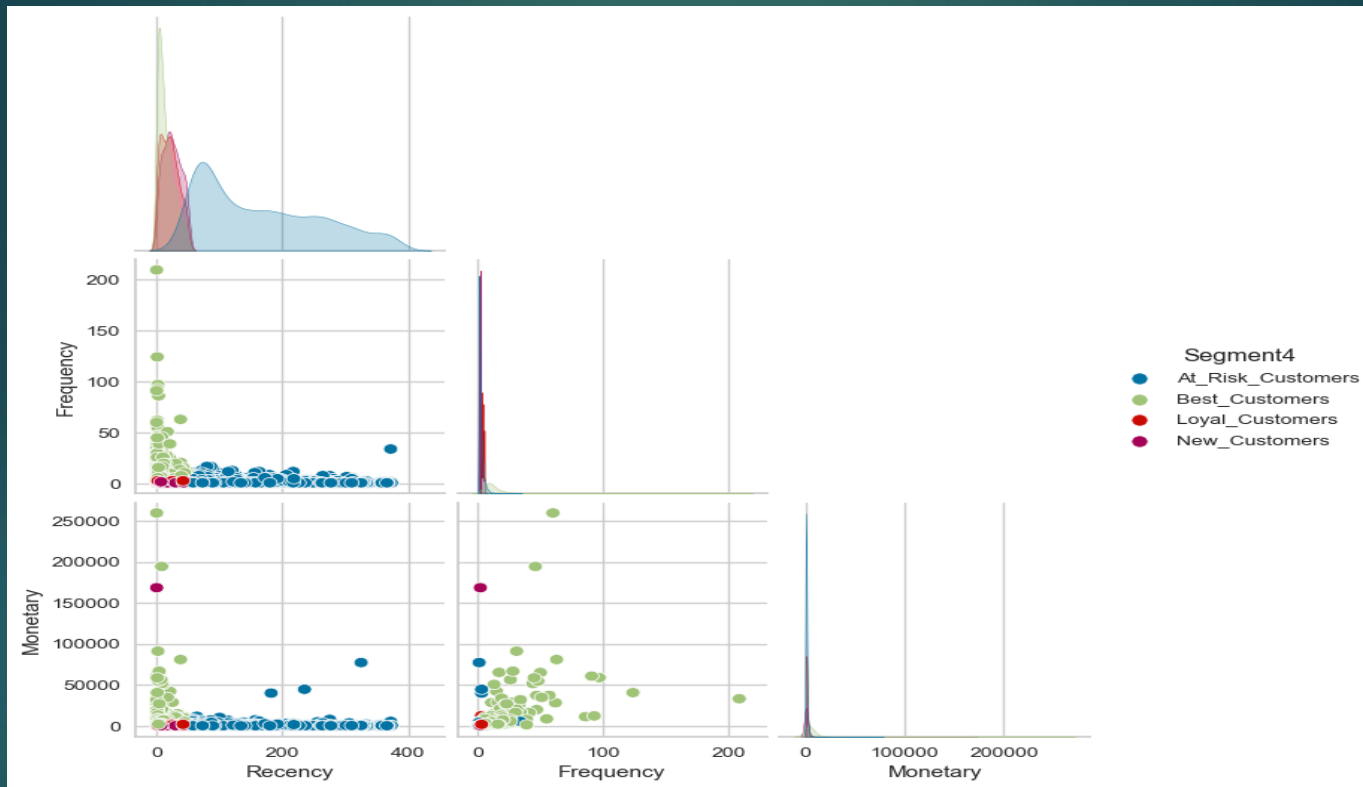


- Applying K-Means Clustering
 - Define and Plot Feature Correlations



When the correlation between futures is examined, there is a strong relationship between frequency and turnover. This is an expected situation.

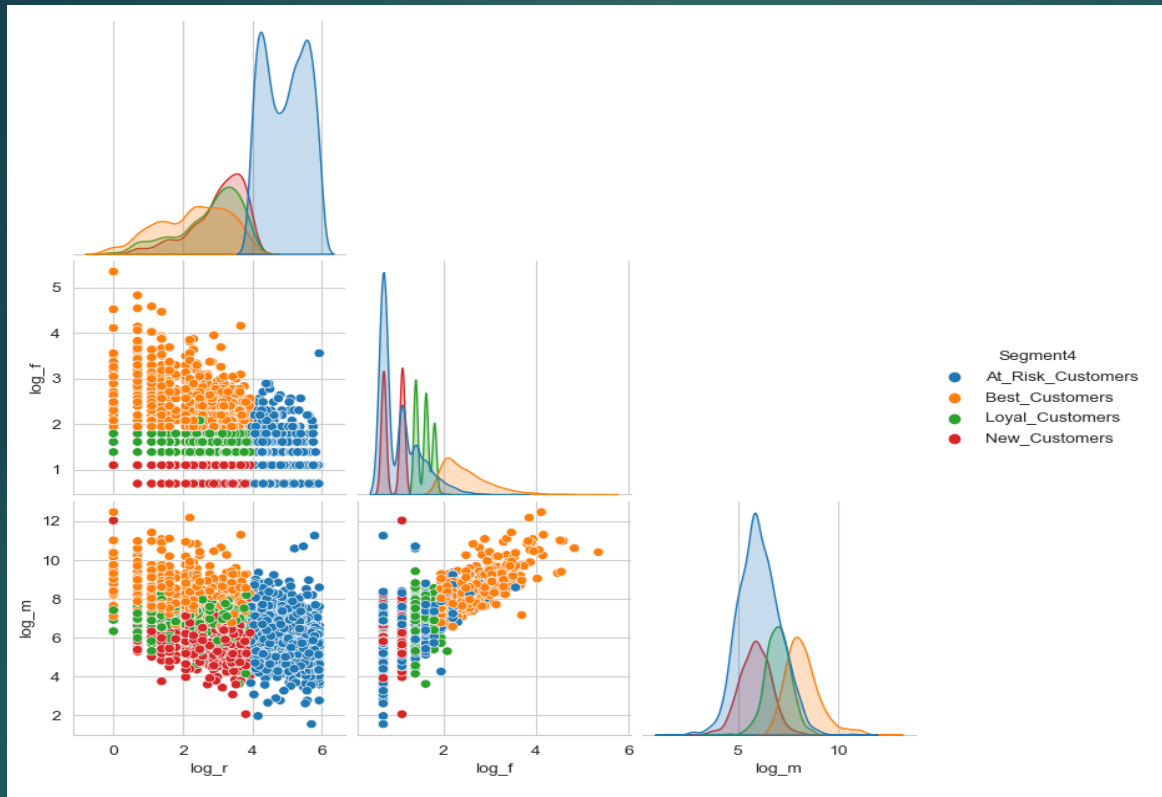
•Visualize Feature Distributions



While examining the table, we cannot clearly see that the classes can be differentiated, but we can say that the loyal and best customers have high frequency, high purchasing and low recency values.

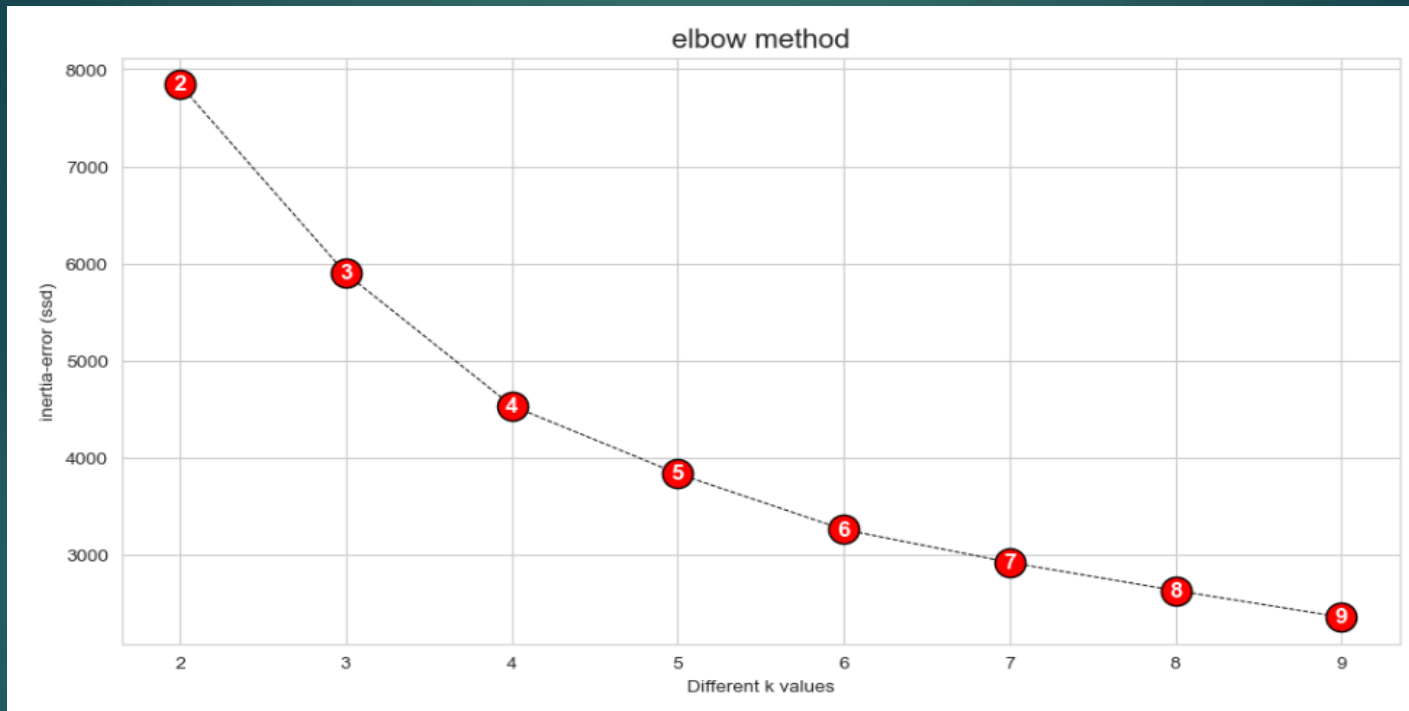
•Data Normalization

While examining customer segment scores, we applied log-normalization in order to see the details more clearly.



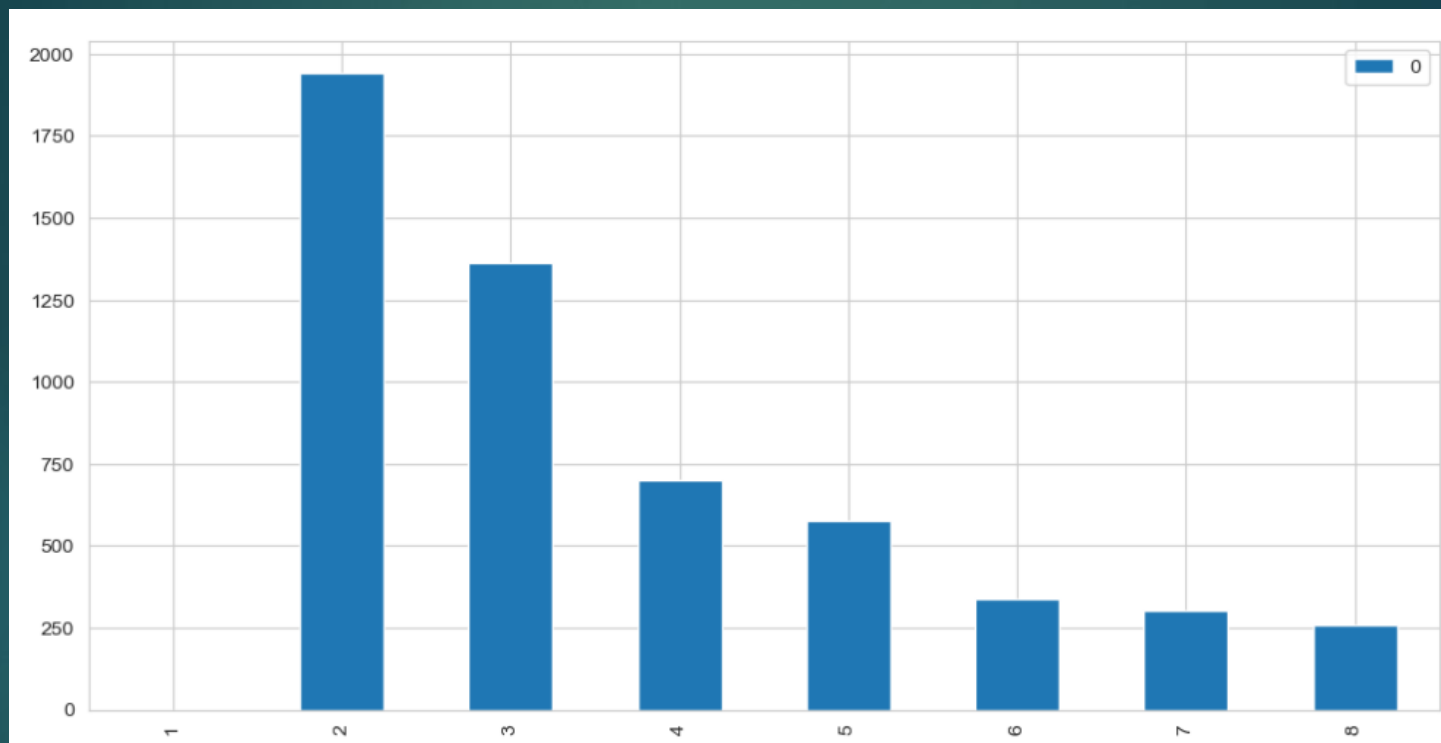
In this table, the separation of customers can be seen more clearly; risky customers with low scores and loyal & best customers with high scores.

- K-Means Implementation
 - Define the Optimal Number of Clusters



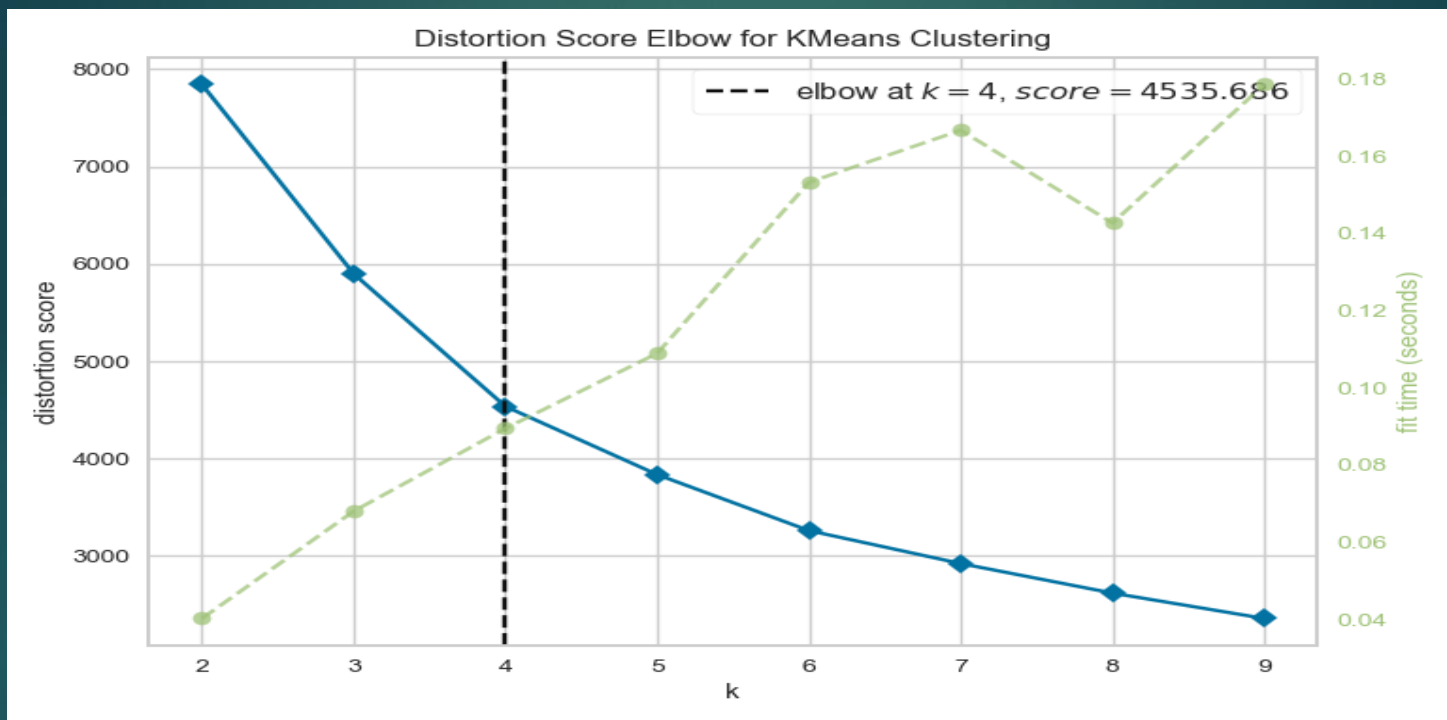
We tried to find the optimum number of segments with the Elbow method, we can say that the net breaks decrease after 4 or 5 segments.

- K-Means Implementation
 - Define the Optimal Number of Clusters



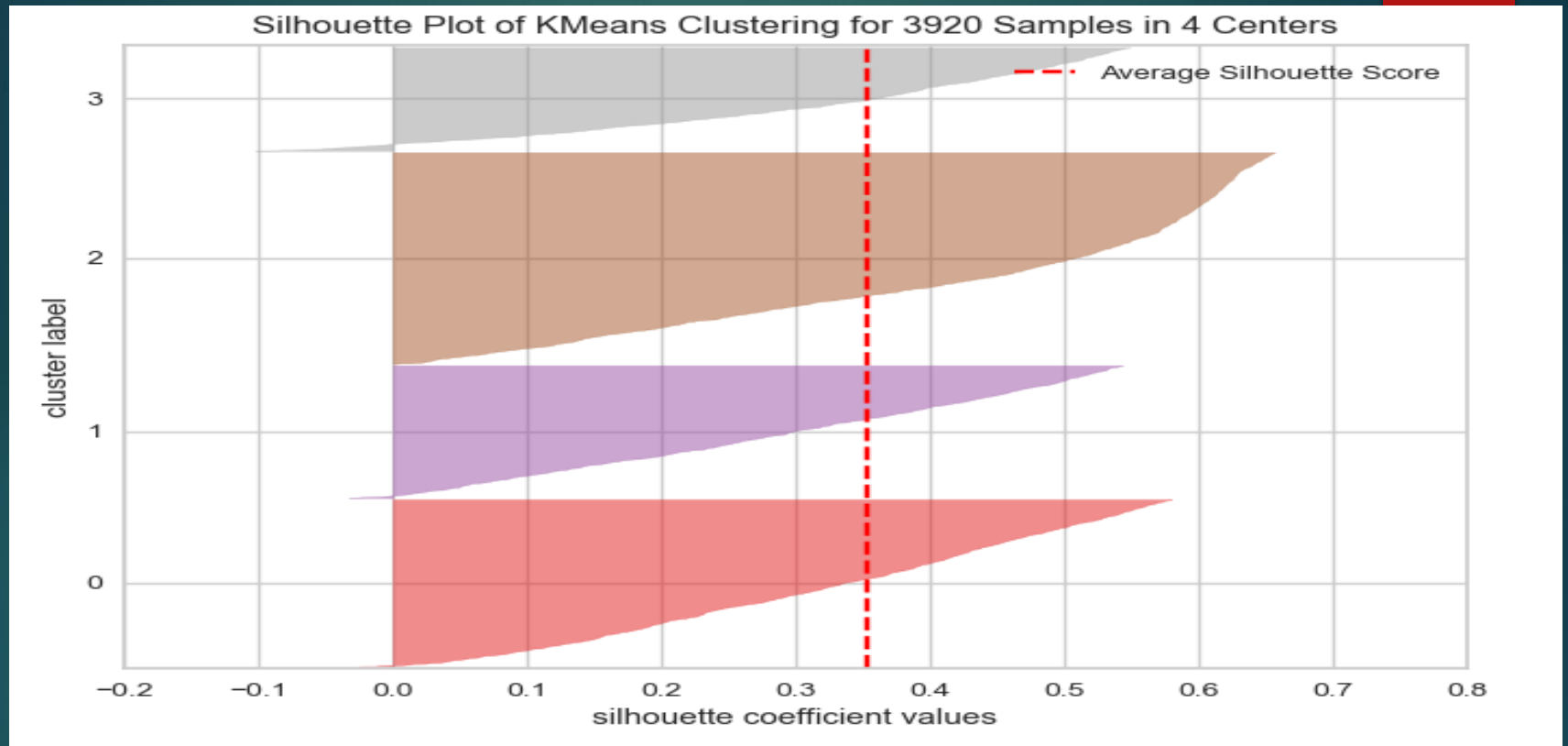
When we examine this table with differences as a bar plot, we see that the breakouts after 2 and 3 decrease dramatically, but less so after 4.

- K-Means Implementation
 - Define the Optimal Number of Clusters



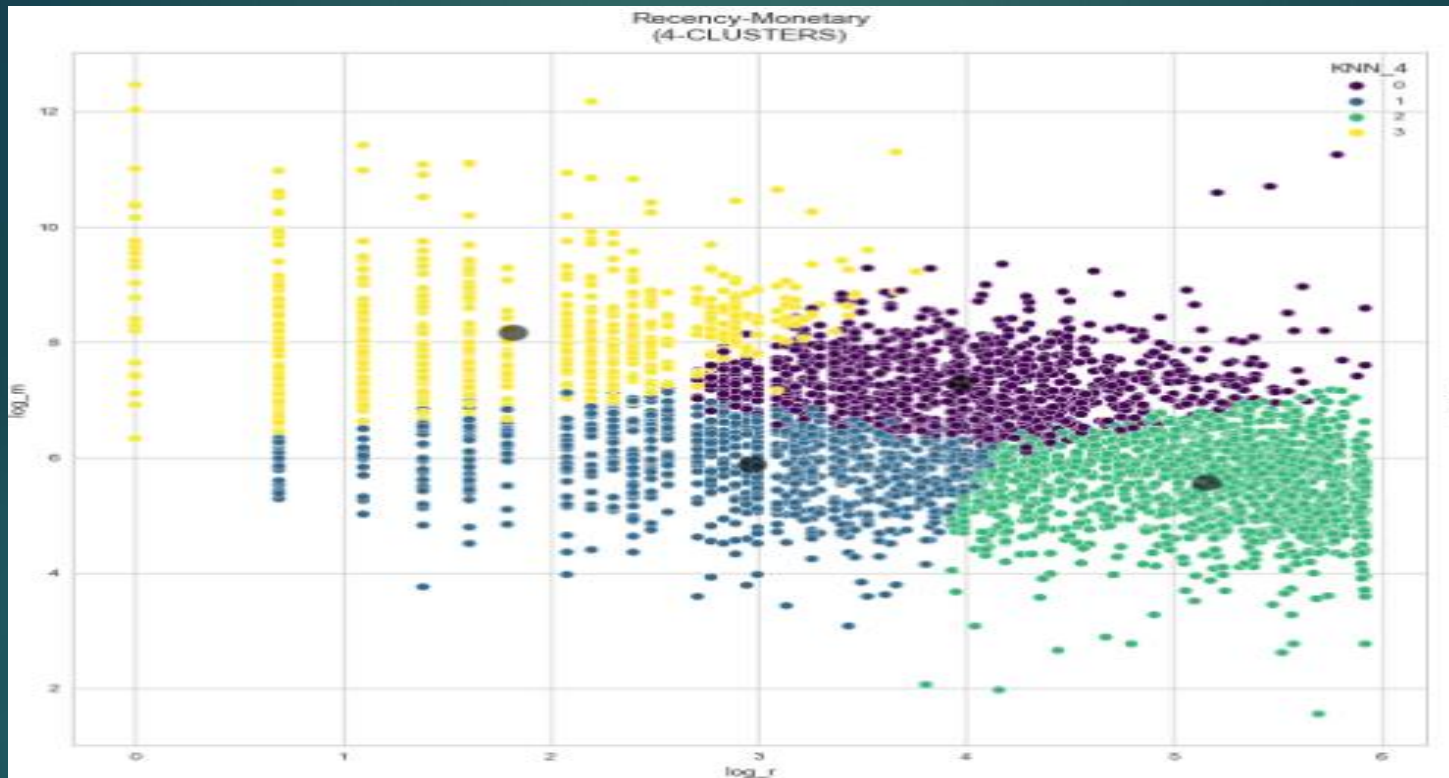
When the timeline is checked in the calculation of the segmentation, we see that the cost increases significantly when the number of clusters exceeds 4. This gives us one more logical reason to divide them into 4 classes.

•Silhouette Score



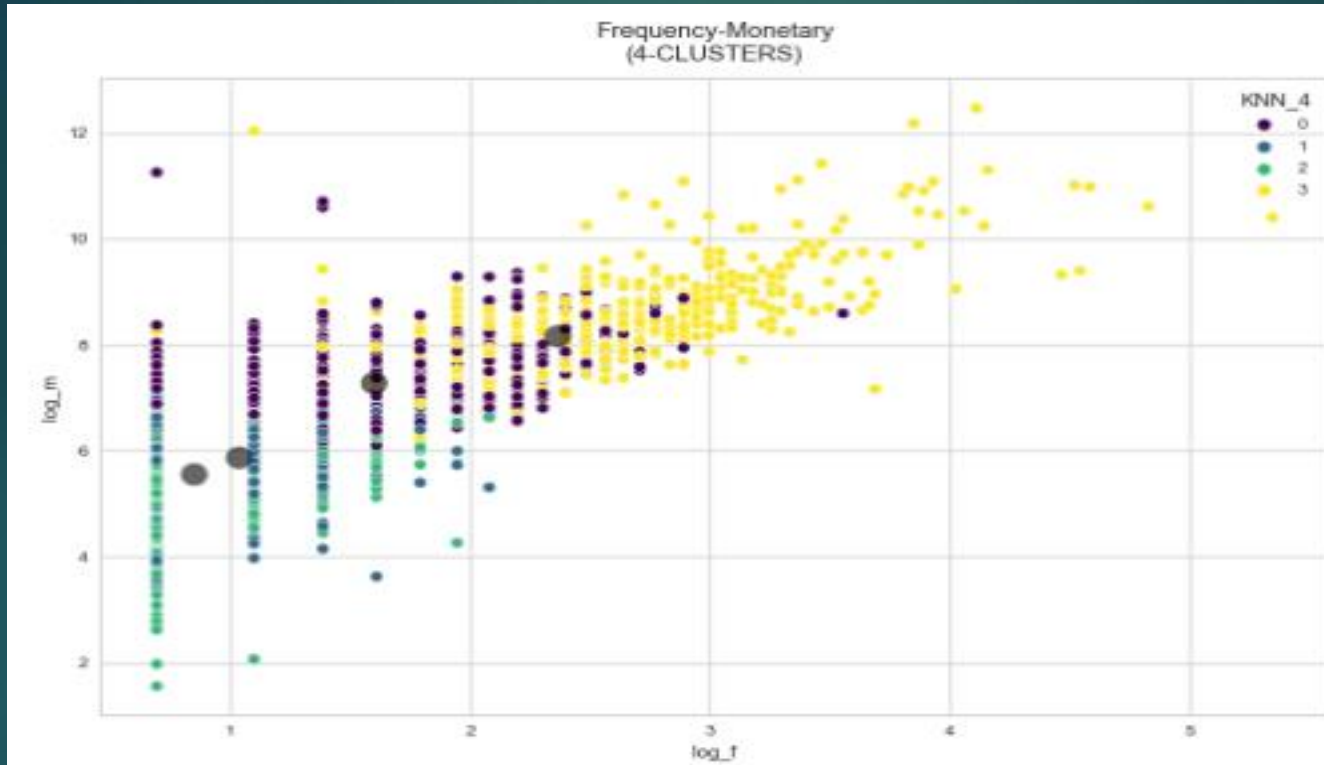
When the silhouette scores are examined, we understand that the negative values are very few and the positive values are close to 1, so the clusters are well segregated.

- Visualize the Clusters (Recency-Monetary)



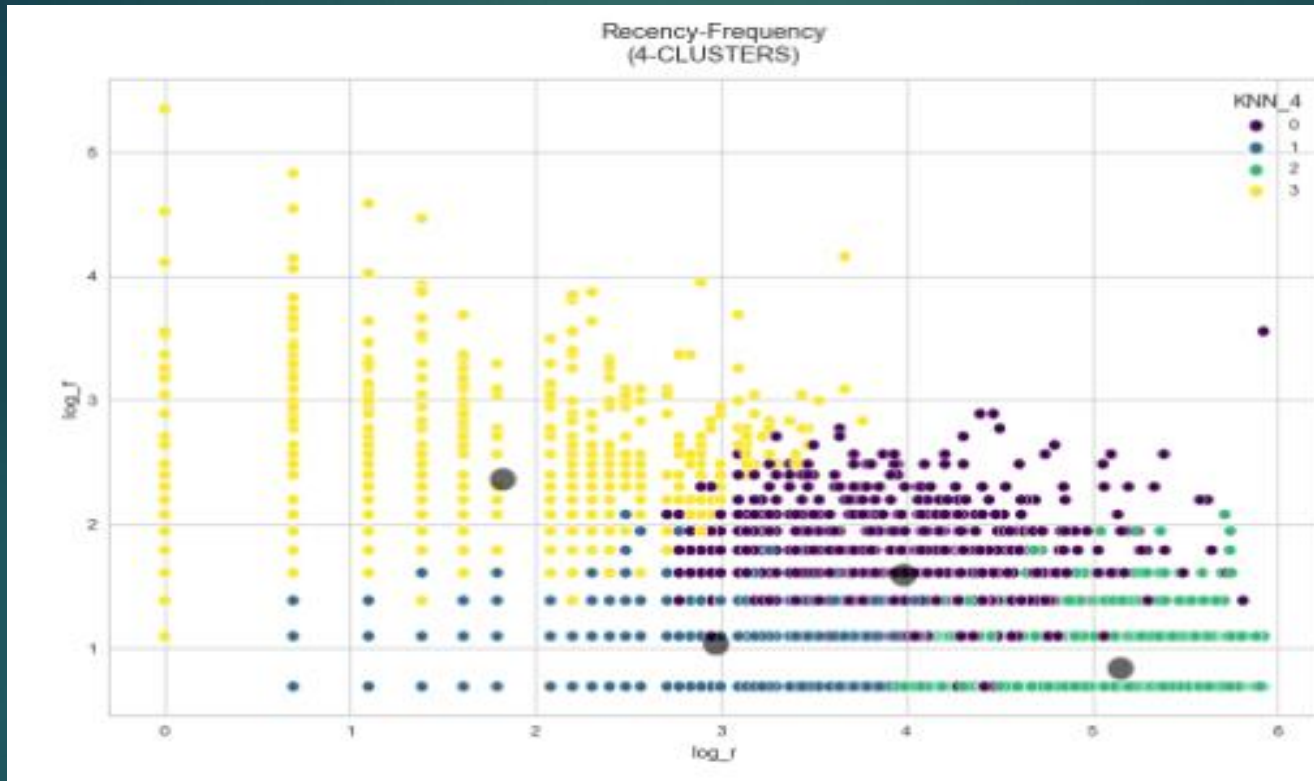
When the separation of borders in terms of monetary-recency values is considered, a significant divergence is observed.

- Visualize the Clusters (Frequency-Monetary)



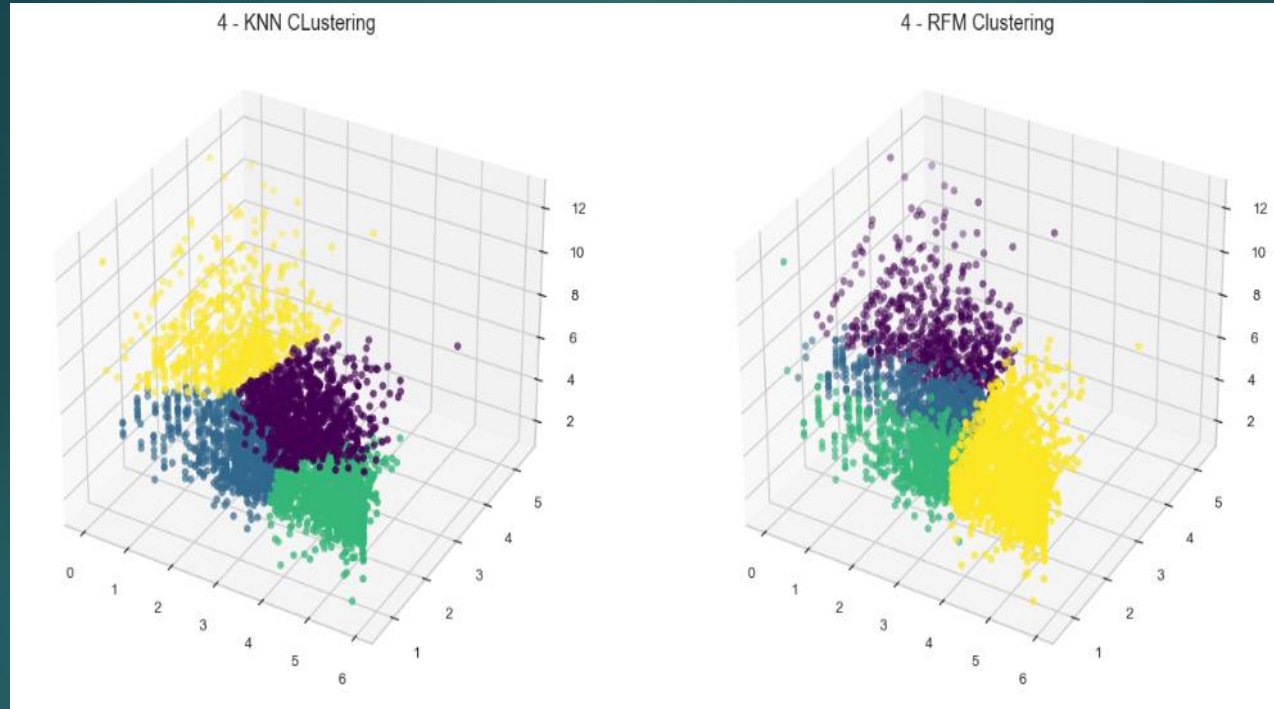
We can say that the 3rd group, which has high frequency and turnover, diverges better. For other groups, turnover is a bigger factor for divergence.

- Visualize the Clusters (Recency-Frequency)



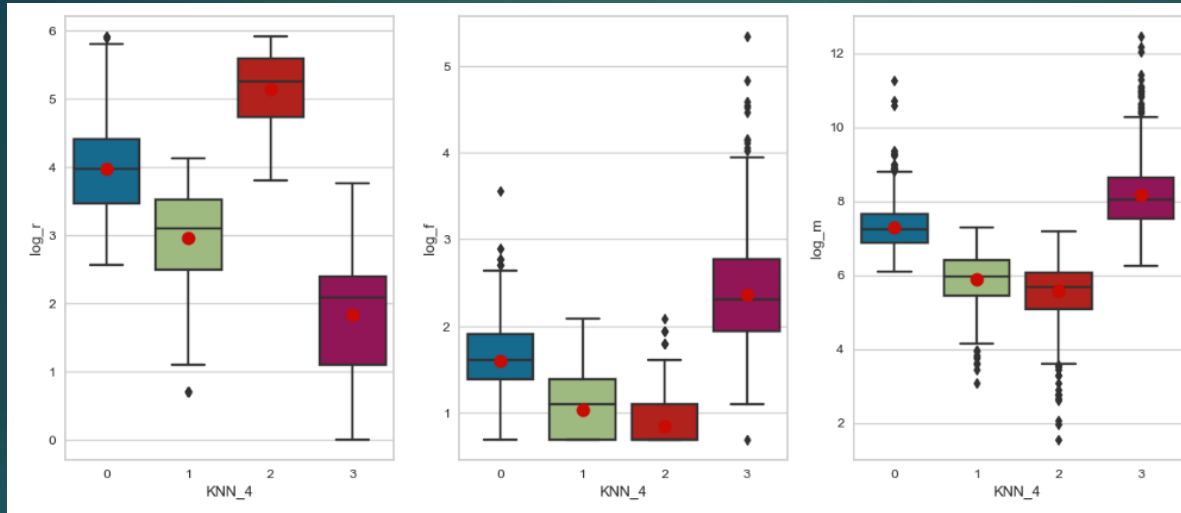
According to Recency-Frequency, when the segments are examined, it is observed that frequency is a serious distinguishing feature, as expected.

•KNN vs RFM



When we look at the results of KNN and RFM, we see that there are different classifications from each other. In RFM, we decided on the classes and their properties. In KNN, the algorithm gave us its own conclusions.

•Assigning the Label



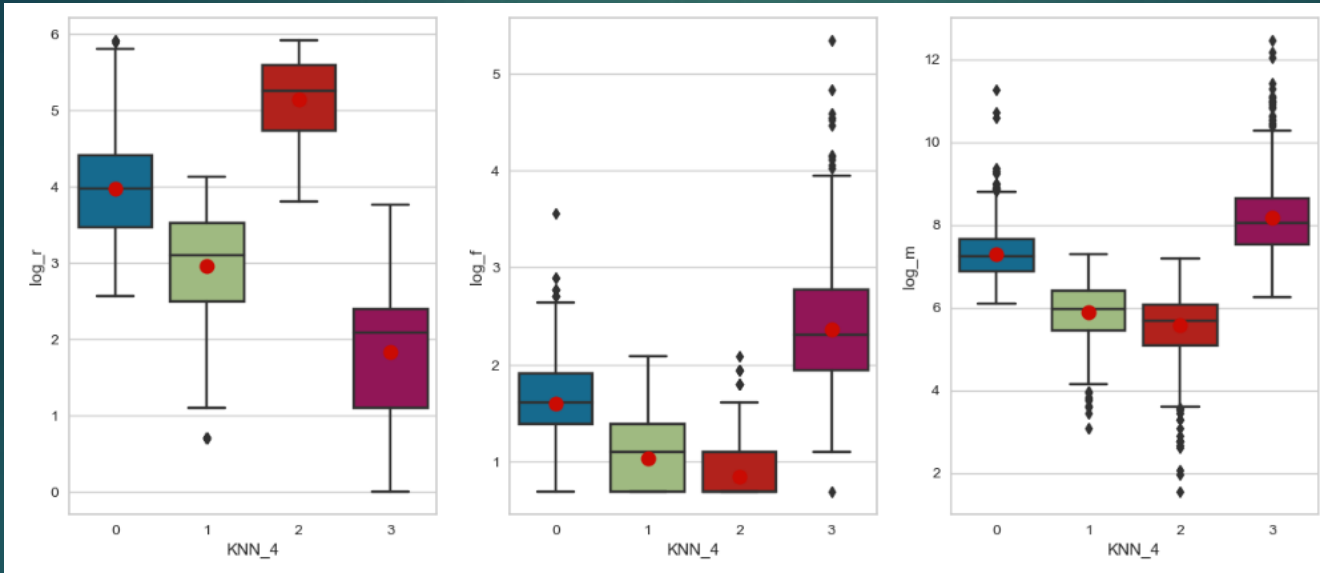
Loyal customers: medium-recently, medium frequency, high turnover customers.

Risky customers: mid-term, low-frequency, mid-turnover customers

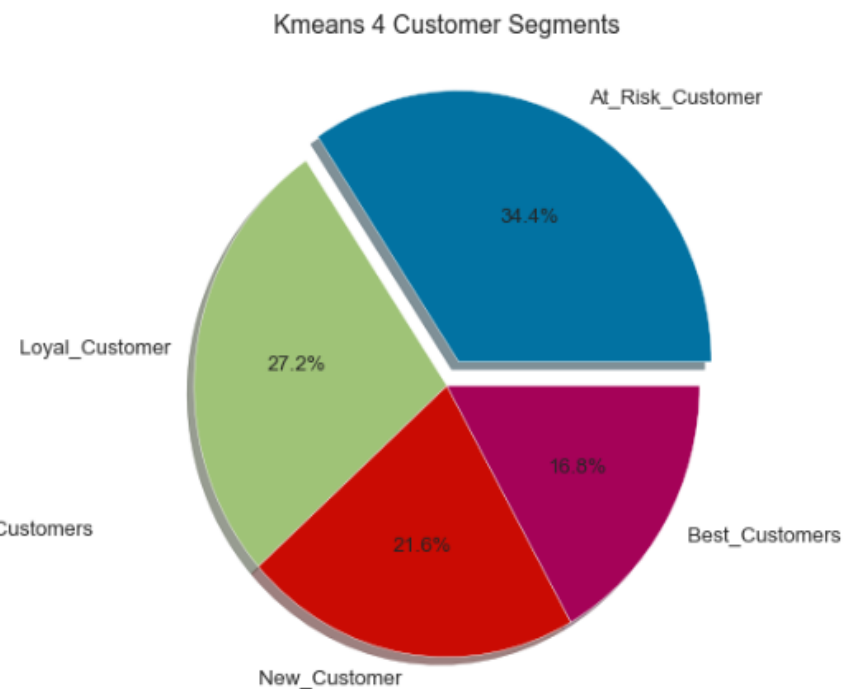
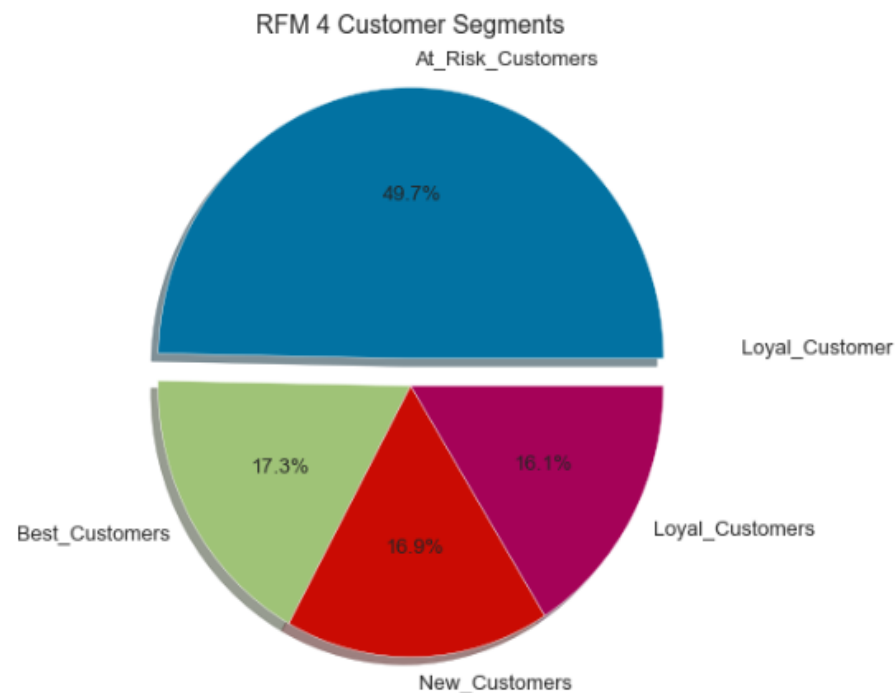
New customers: low-frequency, medium-turnover customers who have arrived very recently.

Customers that need to be regained : Customers from a long time ago with high frequency and high turnover

- Group features



It is seen that the customers are clearly separated in terms of their last arrival dates, and the 1st and 2nd groups cover each other in terms of arrival frequency and turnover.



v. Conclusion

As a result of our investigations, we see that there are differences between RFM and K-Means analyses. As a result of the 4 different grouping domains obtained by the K-means algorithm as a result of its examination according to the locations on the space, we decided to continue with the RFM results. When the 3D graphs of the grouping made as a result of RFM are examined, we can say that the separation is enough.

Create Cohort & Conduct Cohort Analysis

In this analysis, we will try to draw insights by classifying customers during the first time they shop.

- A cohort analysis try to find these questions' answers

- How much effective was a marketing campaign held in a particular time period?
- Did the strategy employ to improve the conversion rates of Customers worked?
- Should I focus more on retention rather than acquiring new customers?
- Are my customer nurturing strategies effective?
- Which marketing channels bring me the best results?
- Is there a seasonality pattern in Customer behavior?
- Along with various performance measures/metrics for your organization.

- Future Engineering
 - Extract the Month of the Purchase

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	TotalPrice	date	InvoiceMonth	CohortMonth	CohortIndex
276704	573865	22678 FRENCH BLUE METAL DOOR SIGN 3	3	2011-11-01 12:00:00	1.250	15311	3.750	2011-11-01	2011-11	2010-12	12
135111	556112	48184 DOORMAT ENGLISH ROSE	2	2011-06-09 09:28:00	7.950	13265	15.900	2011-06-09	2011-06	2011-03	4
294363	575694	22953 BIRTHDAY PARTY CORDON BARRIER TAPE	1	2011-11-10 16:38:00	1.250	12748	1.250	2011-11-10	2011-11	2010-12	12
27115	540502	21166 COOK WITH WINE METAL SIGN	48	2011-01-09 10:50:00	1.690	13183	81.120	2011-01-09	2011-01	2011-01	1
177925	562283	22396 MAGNETS PACK OF 4 RETRO PHOTO	12	2011-08-04 10:43:00	0.390	15249	4.680	2011-08-04	2011-08	2011-03	6

We assign each customer to the month they first came

•Cohort Customer Numbers

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12	815.000	289.000	263.000	304.000	293.000	323.000	291.000	278.000	289.000	325.000	299.000	405.000	218.000
2011-01	358.000	76.000	93.000	84.000	119.000	99.000	90.000	87.000	108.000	117.000	127.000	43.000	NaN
2011-02	340.000	64.000	66.000	97.000	98.000	86.000	87.000	96.000	90.000	104.000	25.000	NaN	NaN
2011-03	419.000	64.000	109.000	83.000	94.000	69.000	111.000	96.000	119.000	38.000	NaN	NaN	NaN
2011-04	277.000	58.000	56.000	60.000	56.000	61.000	61.000	73.000	20.000	NaN	NaN	NaN	NaN
2011-05	256.000	48.000	44.000	44.000	53.000	58.000	68.000	23.000	NaN	NaN	InvoiceMonth		
2011-06	214.000	38.000	31.000	51.000	51.000	69.000	21.000	NaN	NaN	NaN	2010-12	815	
2011-07	169.000	30.000	33.000	39.000	47.000	18.000	NaN	NaN	NaN	NaN	2011-01	647	
2011-08	141.000	32.000	32.000	34.000	17.000	NaN	NaN	NaN	NaN	NaN	2011-02	679	
2011-09	276.000	63.000	83.000	32.000	NaN	NaN	NaN	NaN	NaN	NaN	2011-03	880	
2011-10	324.000	79.000	36.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-04	784	
2011-11	297.000	35.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-05	962	
2011-12	34.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-06	889	
											2011-07	859	
											2011-08	834	
											2011-09	1146	
											2011-10	1230	
											2011-11	1505	
											2011-12	560	

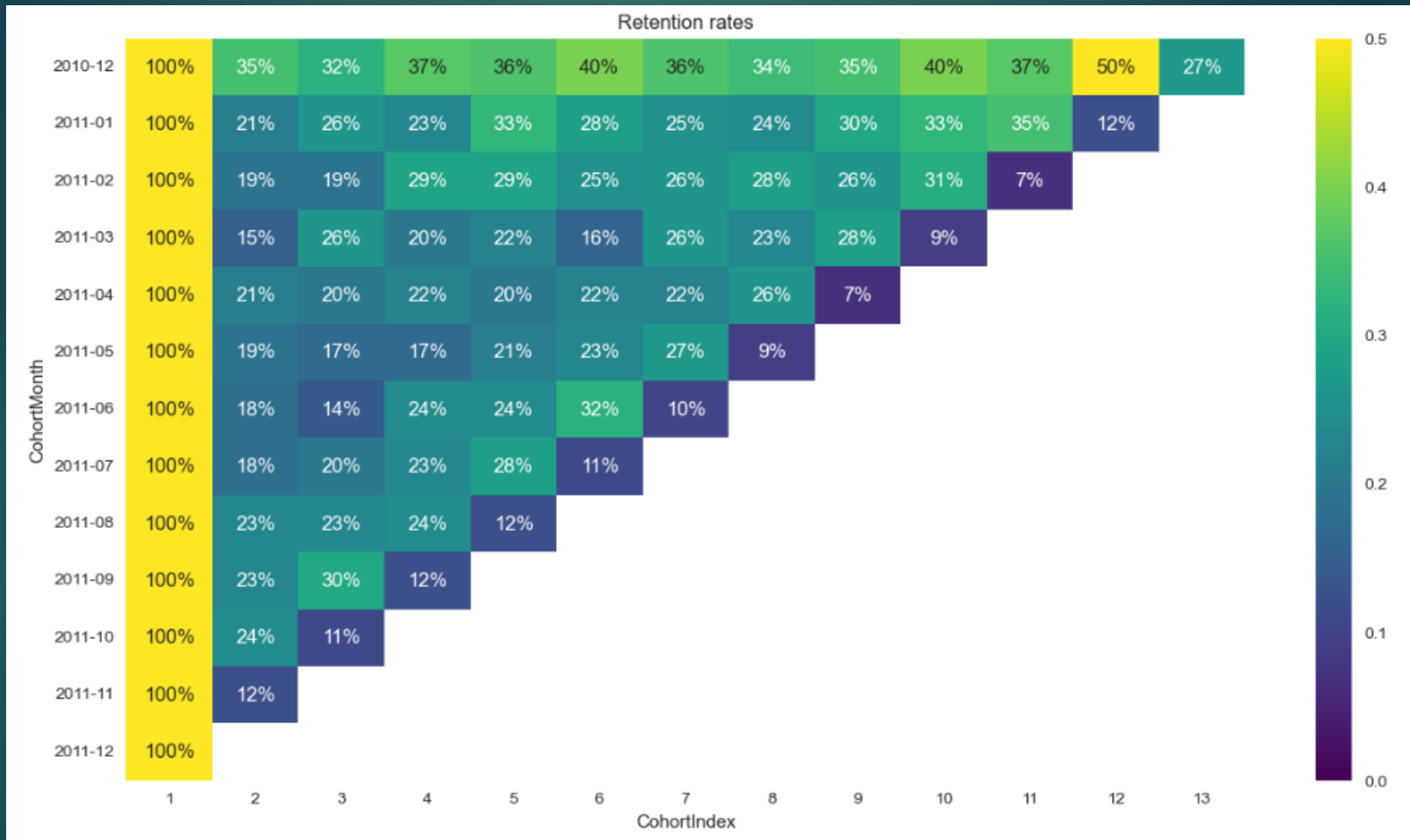
In this table, we see the frequency of the customers after the first month they came.

•Cohort Customer Numbers with Ratios

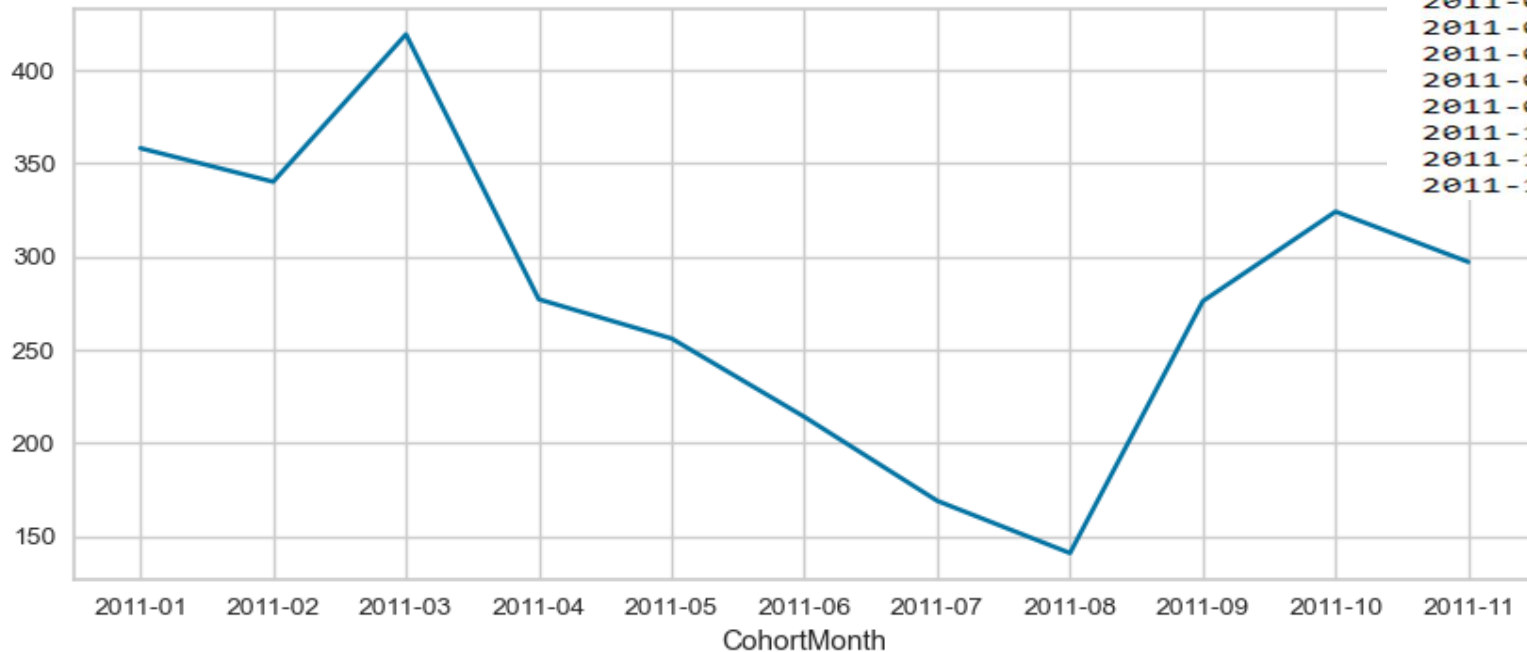
CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12	1.000	0.350	0.320	0.370	0.360	0.400	0.360	0.340	0.350	0.400	0.370	0.500	0.270
2011-01	1.000	0.210	0.260	0.230	0.330	0.280	0.250	0.240	0.300	0.330	0.350	0.120	NaN
2011-02	1.000	0.190	0.190	0.290	0.290	0.250	0.260	0.280	0.260	0.310	0.070	NaN	NaN
2011-03	1.000	0.150	0.260	0.200	0.220	0.160	0.260	0.230	0.280	0.090	NaN	NaN	NaN
2011-04	1.000	0.210	0.200	0.220	0.200	0.220	0.220	0.260	0.070	NaN	NaN	NaN	NaN
2011-05	1.000	0.190	0.170	0.170	0.210	0.230	0.270	0.090	NaN	NaN	InvoiceMonth		
2011-06	1.000	0.180	0.140	0.240	0.240	0.320	0.100	NaN	NaN	NaN	2010-12	815	
2011-07	1.000	0.180	0.200	0.230	0.280	0.110	NaN	NaN	NaN	NaN	2011-01	647	
2011-08	1.000	0.230	0.230	0.240	0.120	NaN	NaN	NaN	NaN	NaN	2011-02	679	
2011-09	1.000	0.230	0.300	0.120	NaN	NaN	NaN	NaN	NaN	NaN	2011-03	880	
2011-10	1.000	0.240	0.110	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-04	784	
2011-11	1.000	0.120	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-05	962	
2011-12	1.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-06	889	
											2011-07	859	
											2011-08	834	
											2011-09	1146	
											2011-10	1230	
											2011-11	1505	
											2011-12	560	

In the table we can see which cohort has a more loyal shopping habit, but the fact that the data is only one year old may mislead us. In the 9th, 10th and 11th months, the number of unique customers increased significantly, correspondingly high return for each cohort appear in these months.

- Visualize Cohort Customer Numbers with Ratios



- New Customers Gained Monthly



InvoiceMonth	
2010-12	815
2011-01	647
2011-02	679
2011-03	880
2011-04	784
2011-05	962
2011-06	889
2011-07	859
2011-08	834
2011-09	1146
2011-10	1230
2011-11	1505
2011-12	560

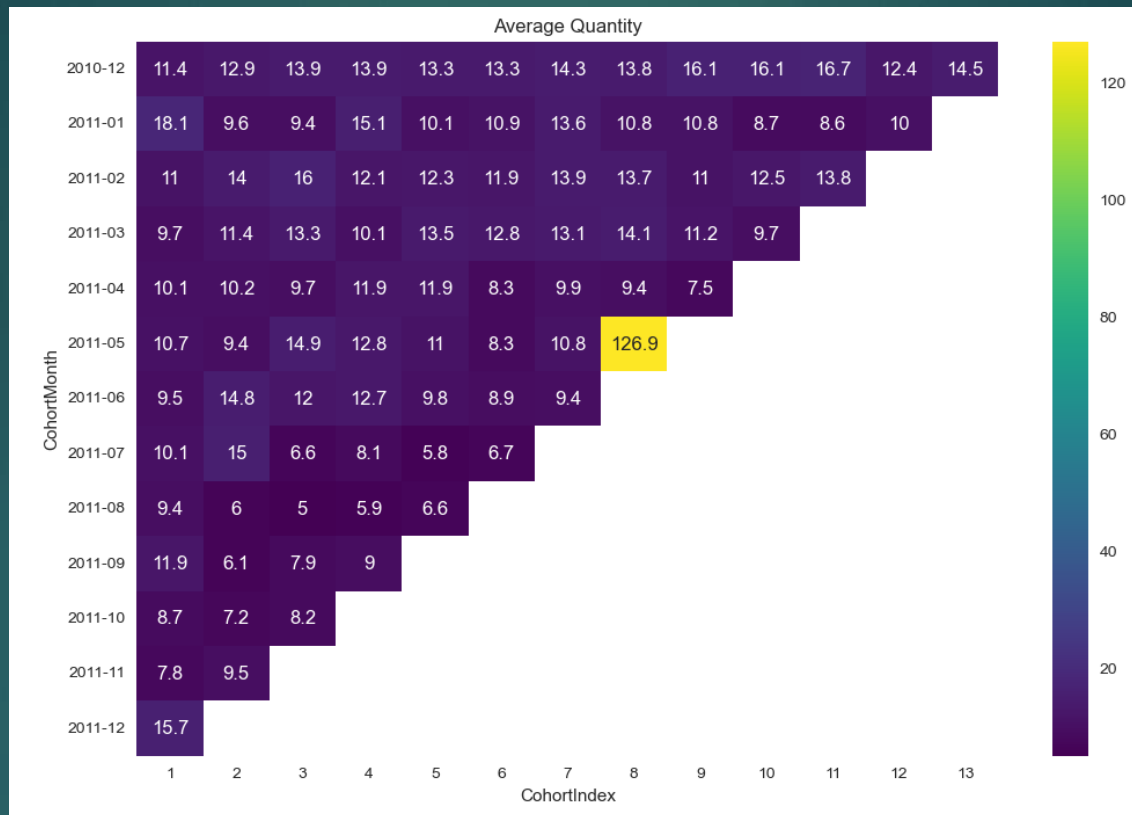
In the table we can see which cohort has a more loyal shopping habit, but the fact that the data is only one year old may mislead us. In the 9th, 10th and 11th months, the number of unique customers increased significantly, correspondingly high return rates for each cohort appear in these months.

•Average Quantity Sold

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12	11.400	12.900	13.900	13.900	13.300	13.300	14.300	13.800	16.100	16.100	16.700	12.400	14.500
2011-01	18.100	9.600	9.400	15.100	10.100	10.900	13.600	10.800	10.800	8.700	8.600	10.000	NaN
2011-02	11.000	14.000	16.000	12.100	12.300	11.900	13.900	13.700	11.000	12.500	13.800	NaN	NaN
2011-03	9.700	11.400	13.300	10.100	13.500	12.800	13.100	14.100	11.200	9.700	NaN	NaN	NaN
2011-04	10.100	10.200	9.700	11.900	11.900	8.300	9.900	9.400	7.500	NaN	NaN	NaN	NaN
2011-05	10.700	9.400	14.900	12.800	11.000	8.300	10.800	126.900	NaN	NaN	NaN	NaN	NaN
2011-06	9.500	14.800	12.000	12.700	9.800	8.900	9.400	NaN	NaN	NaN	NaN	NaN	NaN
2011-07	10.100	15.000	6.600	8.100	5.800	6.700	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08	9.400	6.000	5.000	5.900	6.600	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09	11.900	6.100	7.900	9.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10	8.700	7.200	8.200	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11	7.800	9.500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12	15.700	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

As we know from the shopping site, bulk purchases are made.
Your customs here may be misleading for us.

- Visualize Avarage Quantity Sold



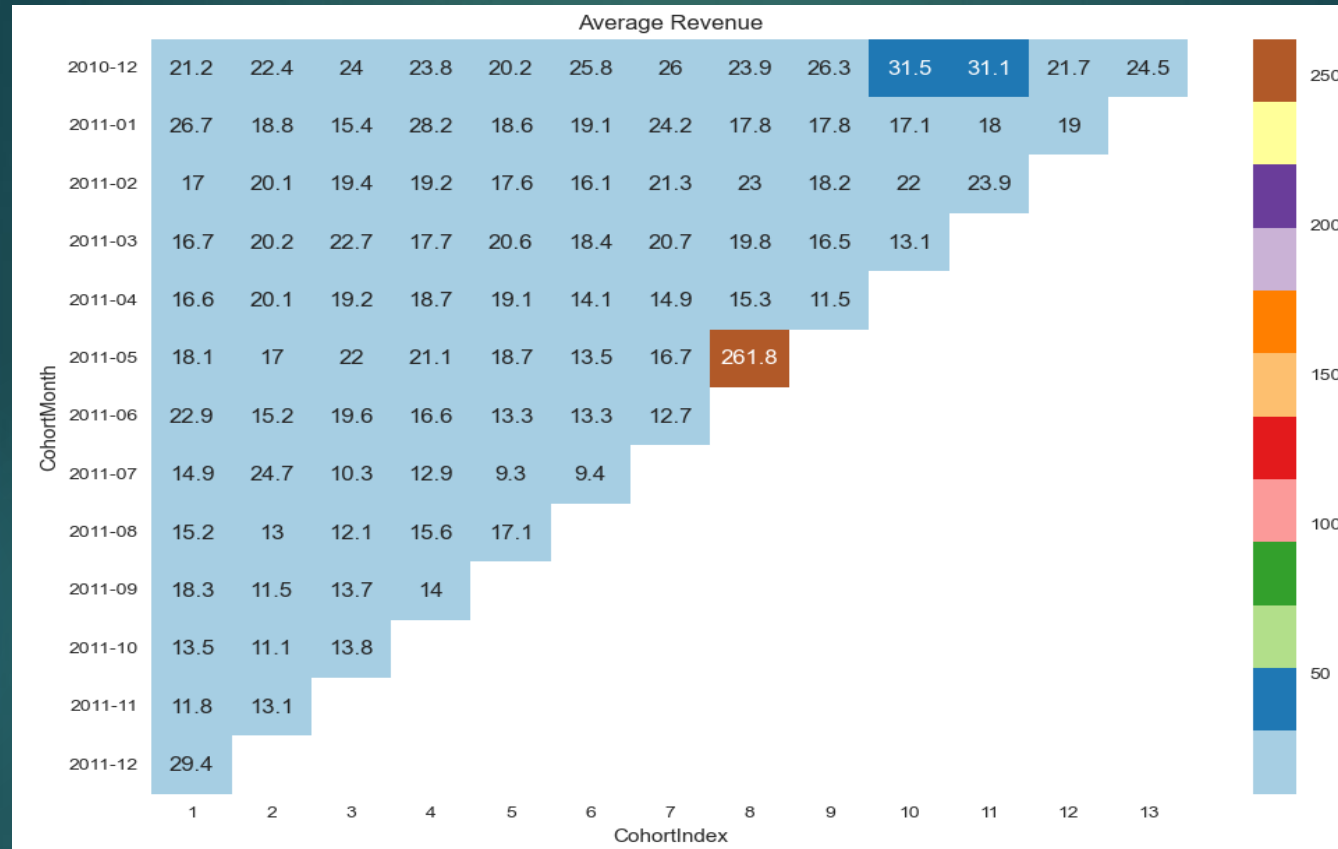
The extreme value here was realized by the purchases of very high volumes of the same product by a few customers. This difference is not significant.

•Average Total Price

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12	21.200	22.400	24.000	23.800	20.200	25.800	26.000	23.900	26.300	31.500	31.100	21.700	24.500
2011-01	26.700	18.800	15.400	28.200	18.600	19.100	24.200	17.800	17.800	17.100	18.000	19.000	NaN
2011-02	17.000	20.100	19.400	19.200	17.600	16.100	21.300	23.000	18.200	22.000	23.900	NaN	NaN
2011-03	16.700	20.200	22.700	17.700	20.600	18.400	20.700	19.800	16.500	13.100	NaN	NaN	NaN
2011-04	16.600	20.100	19.200	18.700	19.100	14.100	14.900	15.300	11.500	NaN	NaN	NaN	NaN
2011-05	18.100	17.000	22.000	21.100	18.700	13.500	16.700	261.800	NaN	NaN	NaN	InvoiceMonth	
2011-06	22.900	15.200	19.600	16.600	13.300	13.300	12.700	NaN	NaN	NaN	NaN	2010-12	21.165
2011-07	14.900	24.700	10.300	12.900	9.300	9.400	NaN	NaN	NaN	NaN	NaN	2011-01	24.605
2011-08	15.200	13.000	12.100	15.600	17.100	NaN	NaN	NaN	NaN	NaN	NaN	2011-02	20.220
2011-09	18.300	11.500	13.700	14.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-03	19.646
2011-10	13.500	11.100	13.800	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-04	19.788
2011-11	11.800	13.100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-05	22.051
2011-12	29.400	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011-06	22.328
												2011-07	20.743
												2011-08	21.717
												2011-09	22.529
												2011-10	19.101
												2011-11	16.973
												2011-12	30.367

In this table, we can say that new customers do not have a high average turnover as old customers. Although this is not true for all months, it appears to be so in general.

•Visualize Avarage Total Price



Conclusion

- First of all, we started to examine our data, we completed the necessary cleaning processes, and during this time we started to get to know the data.
- With the RFM analysis, we assigned points to the Recency, Frequency, Monetary Value values of the customers and classified them with our own business information.
- Then, with K-Means Clustering, we had 4 segments in the number and shape suggested by the algorithm. Here, since the formation of segments is the algorithm's own classification rather than our business knowledge, we have reviewed the classes and named them appropriately with our business knowledge.
- In the cohort analysis, on the other hand, we assigned customers to classes within the first months of our acquisition and observed their later coming habits. This allowed us to examine if there are campaigns, discounts or seasonal effects that we have made in the customer's shopping cycle.

Thanks for listening