Customer Segmentation For An Online Retail

Zhaoqi Fan

Outline

- Problem Identification
- Data Wrangling
- Exploratory Data Analysis
- Clustering
- Conclusion & Recommendation

Problem Identification

Two-year transnational data has been collected for a UK-based and registered non-store online retail. The manager is interested in targeting customers for making business development strategies.

Data source:

- the dataset was scraped from [Kaggle](#) and [UCI Machine Learning Repository](#)

Constrains:

- The data is mainly from the UK. The scarce data from other countries may bring noise into the analysis.

Column	Non-Null	Count	Type
Customer ID	809561	non-null	float64
Invoice	1044848	non-null	object
InvoiceDate	1044848	non-null	object
Price	1044848	non-null	float64
Quantity	1044848	non-null	int64
StockCode	1044848	non-null	object
Description	1040573	non-null	object
Country	1044848	non-null	object

Problem Identification

How many groups we can segment the customers by using numerical features?

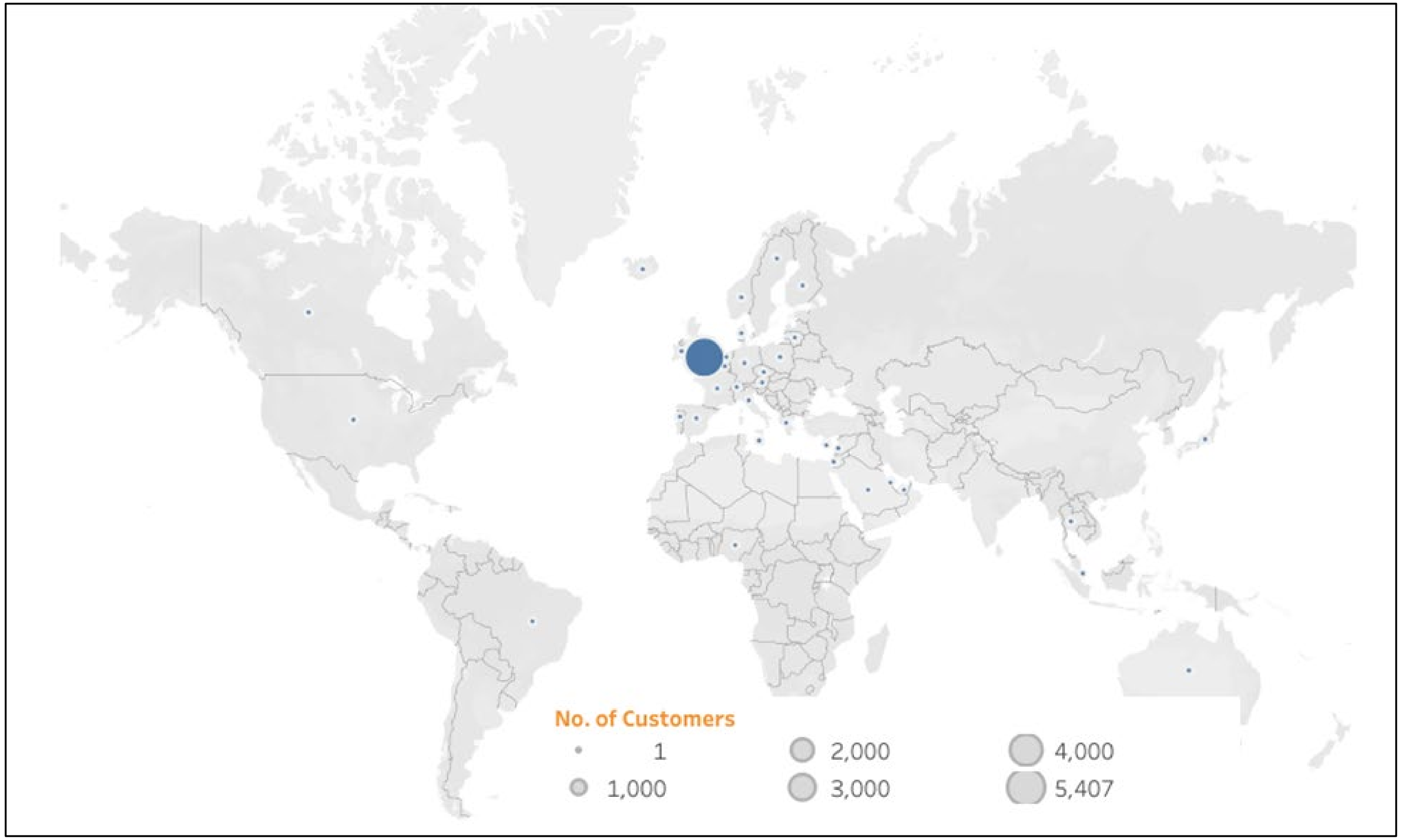
Data Wrangling

- ***Missing data***: 235287 rows containing missing customer ID are dropped.
- ***Duplicates***: 11676 duplicated rows are dropped.

	Customer ID	Invoice	InvoiceDate	Price	Quantity	StockCode	Description	Country
255587	12346	C514024	2010-06-30 11:22:00	12.94	-1	M	manual	United Kingdom
255589	12346	C514024	2010-06-30 11:22:00	12.94	-1	M	manual	United Kingdom
485600	12356	534804	2010-11-24 12:24:00	1.95	1	22629	spaceboy lunch box	Portugal
485615	12356	534804	2010-11-24 12:24:00	1.95	1	22629	spaceboy lunch box	Portugal

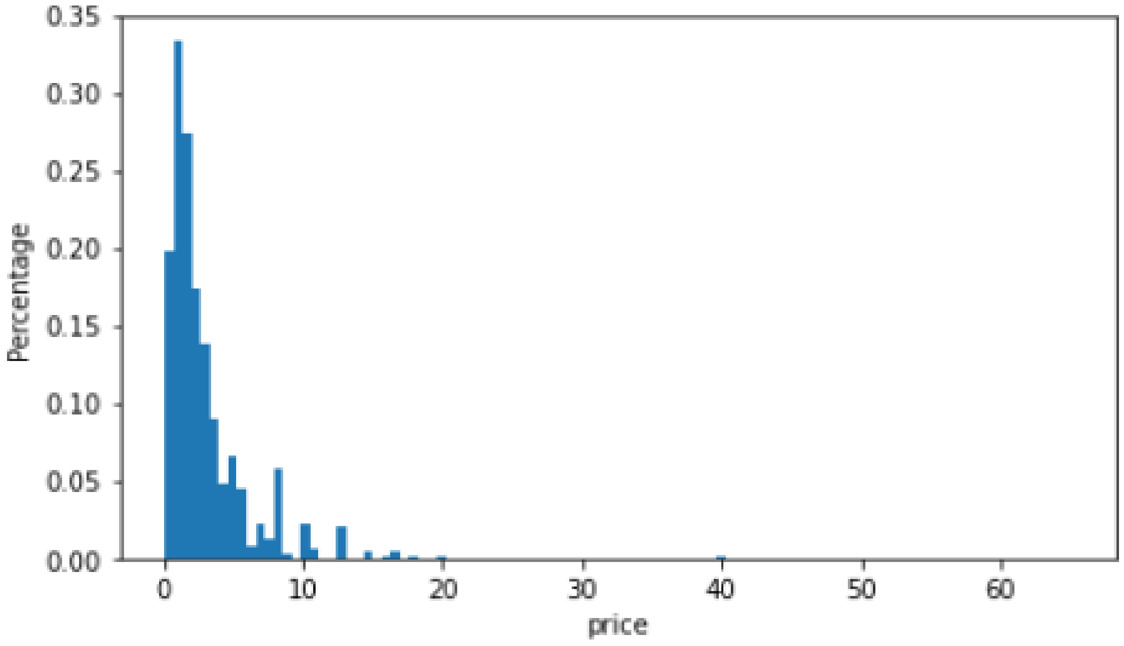
Data Wrangling

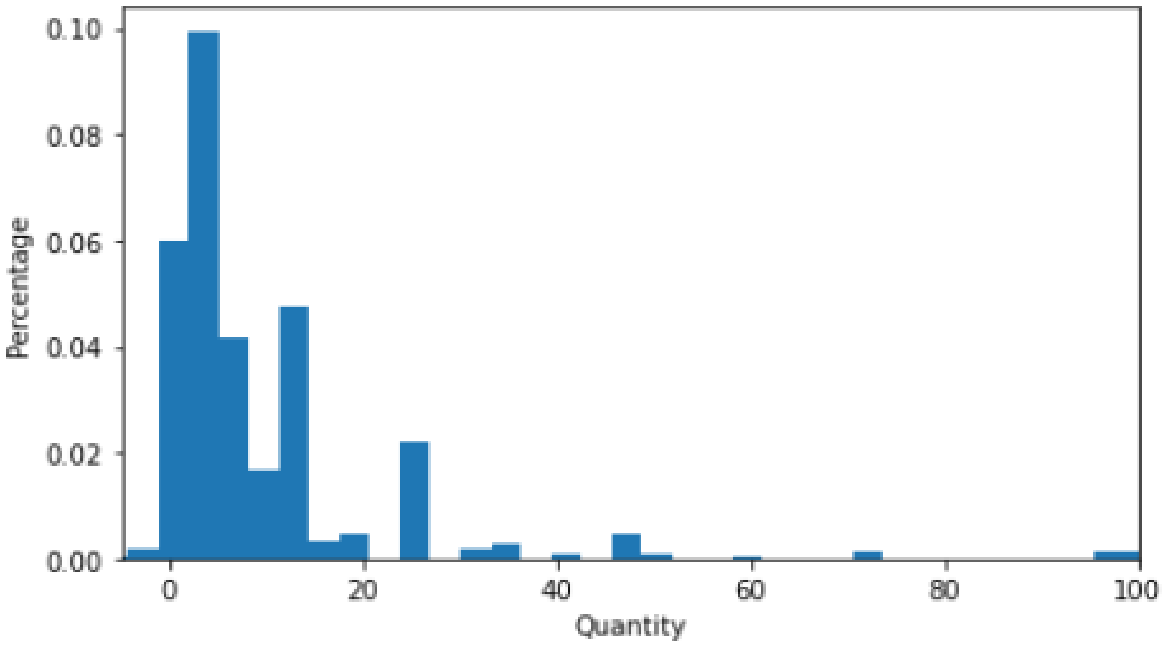
- ***Categorical feature***
  - The stock code and description are not considered.
  - Most of data are collected from the UK.



Data Wrangling

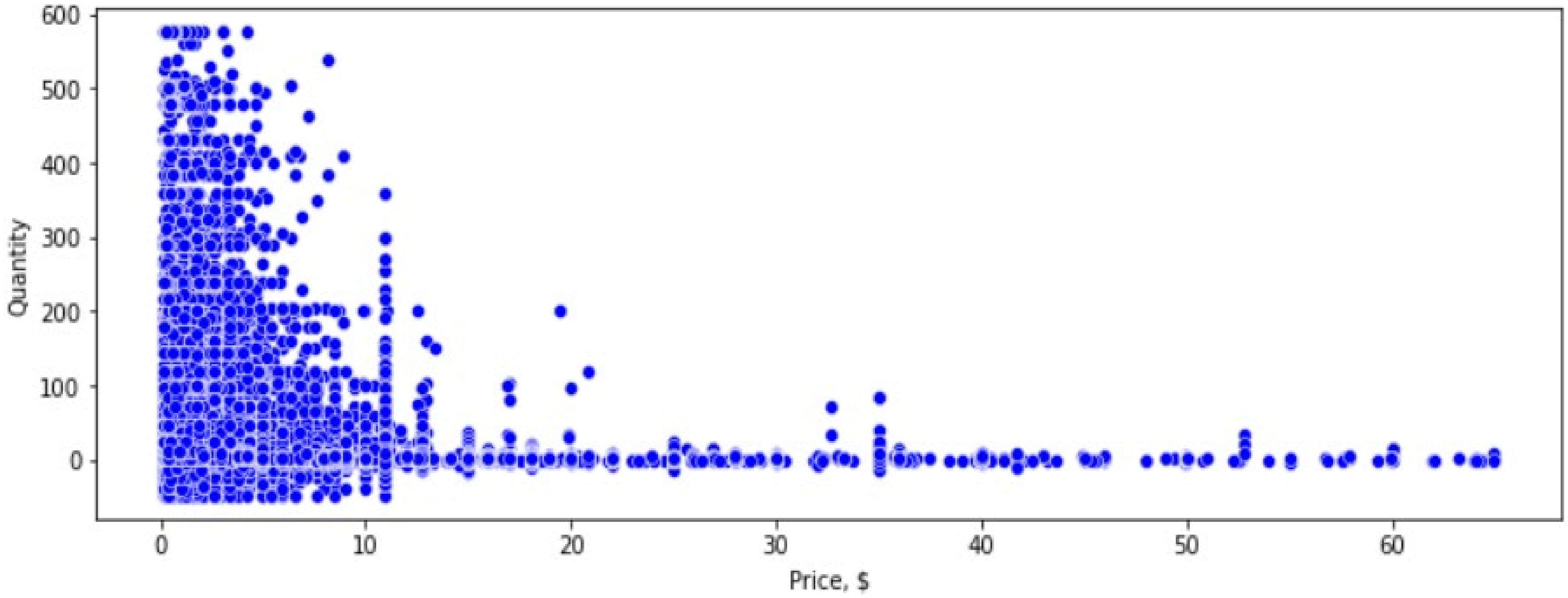
- ***Numerical feature statistics***
  - The distribution is extremely skewed.
  - It suggests a transformation in modeling.





Exploratory Data Analysis

- ***Price vs quantity***
  - no clear correlation between price and quantity.

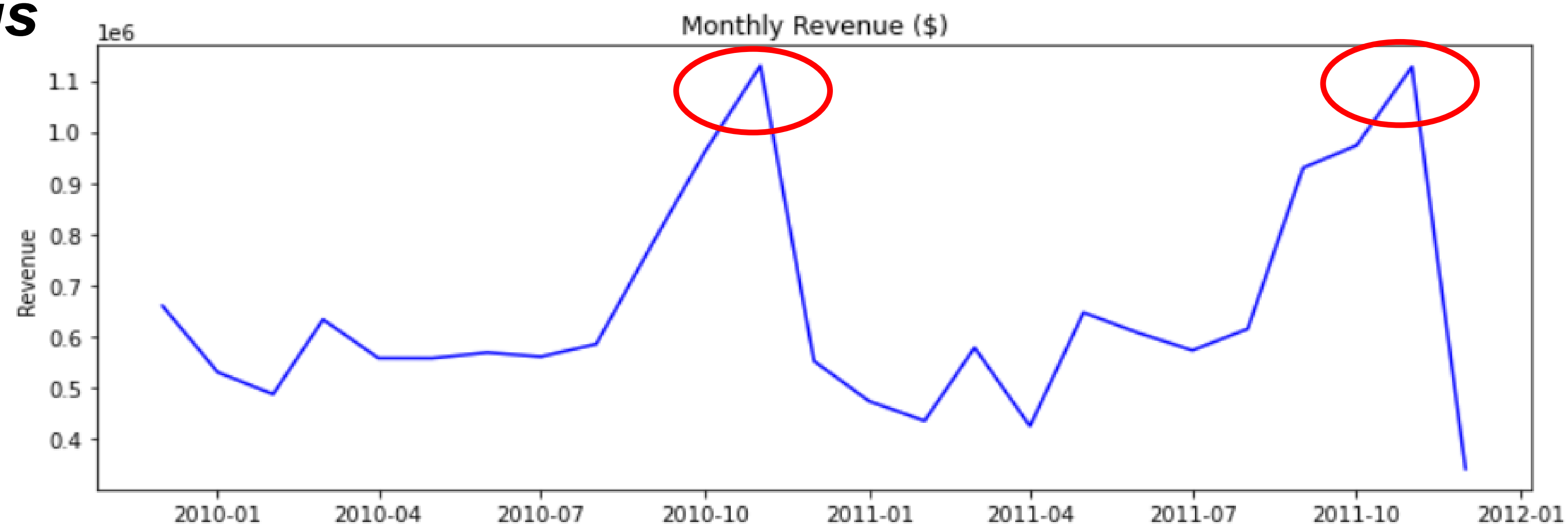


Exploratory Data Analysis

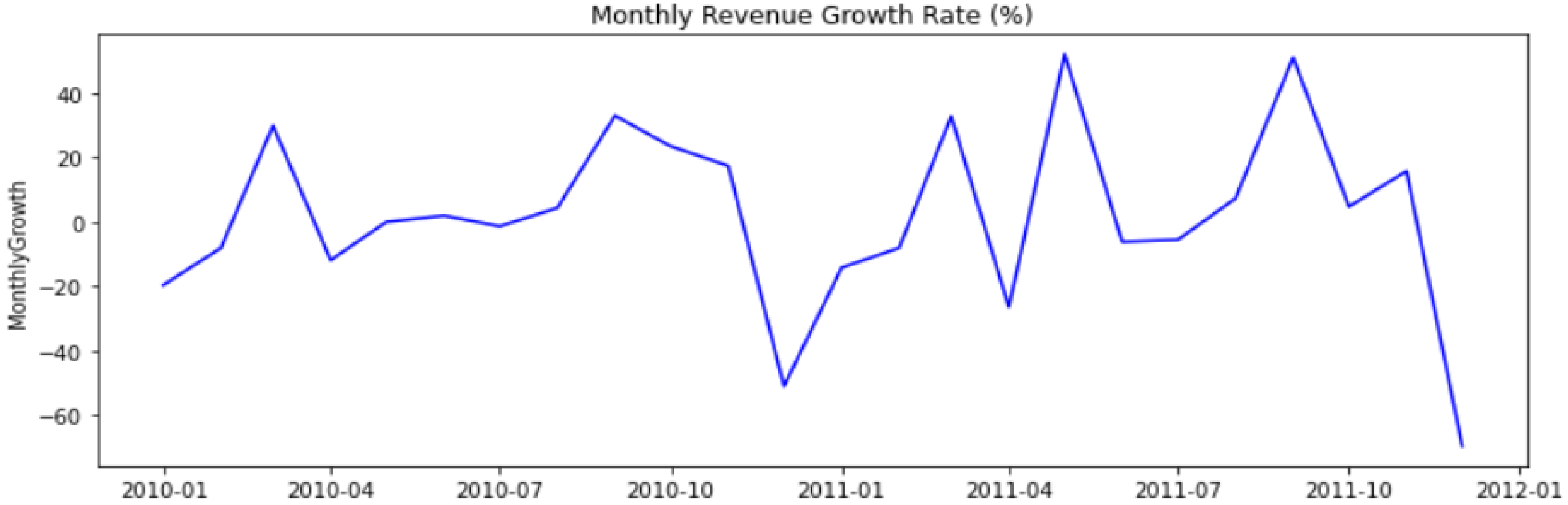
- ***Revenue Analysis***

Monthly revenue

- Seasonality
- Oct. – Nov.
- Jan. – Jun.



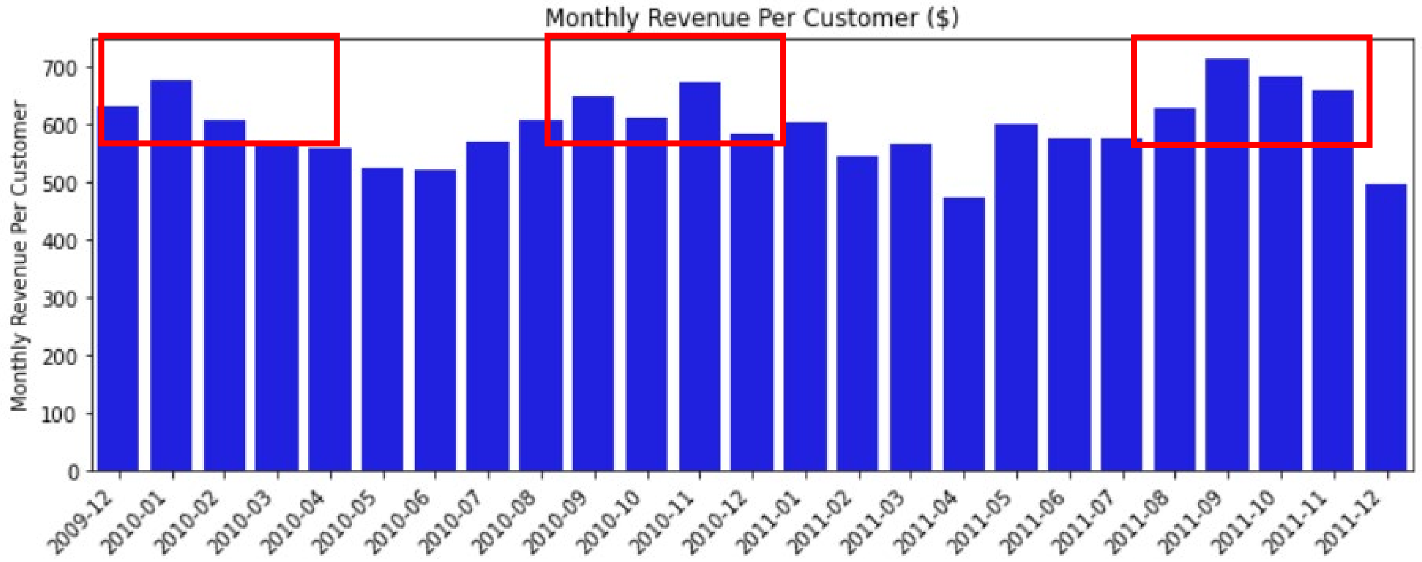
Monthly revenue growth rate



Exploratory Data Analysis

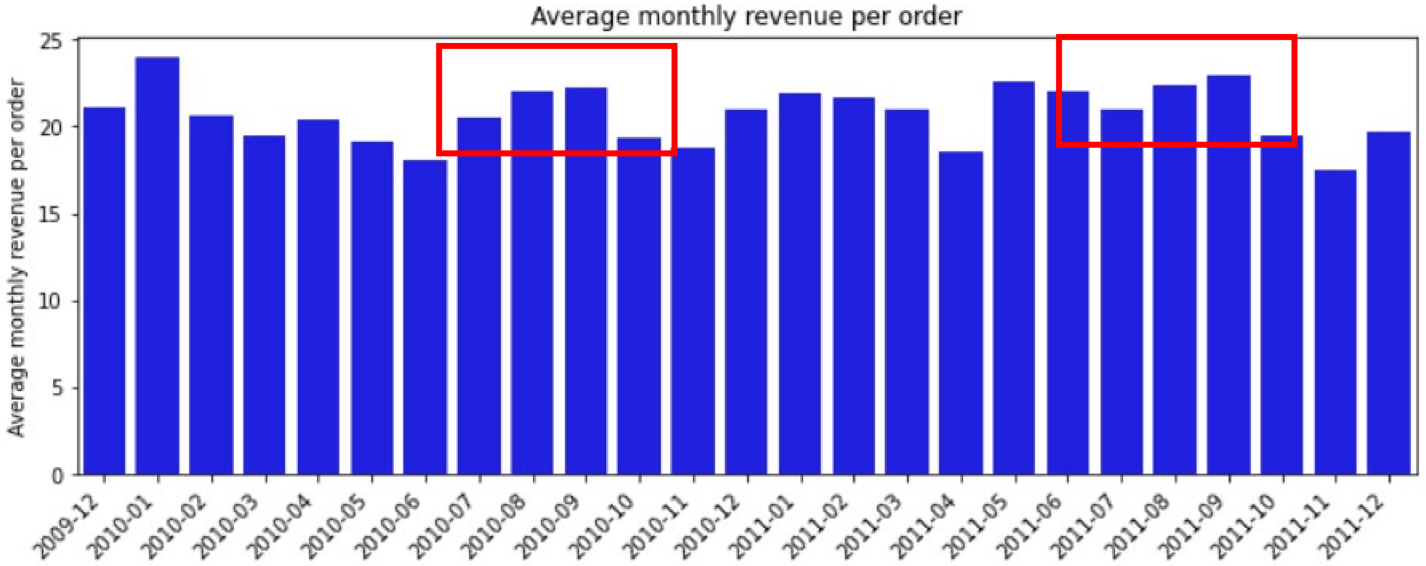
- ***Revenue Analysis***

Monthly revenue per customer



  - Seasonality

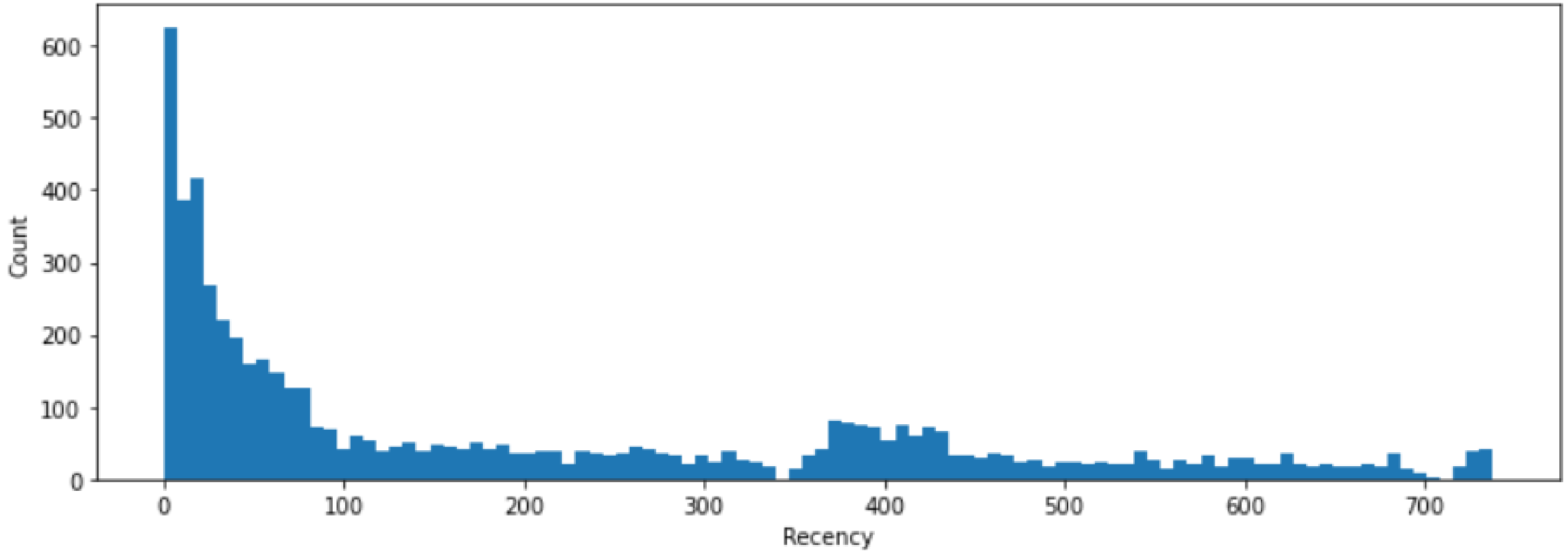
Monthly revenue per order



Clustering

- ***Recency (R) calculation and clustering***

To calculate recency, we need to find out the most recent purchase date of each customer and see how many days they are inactive.

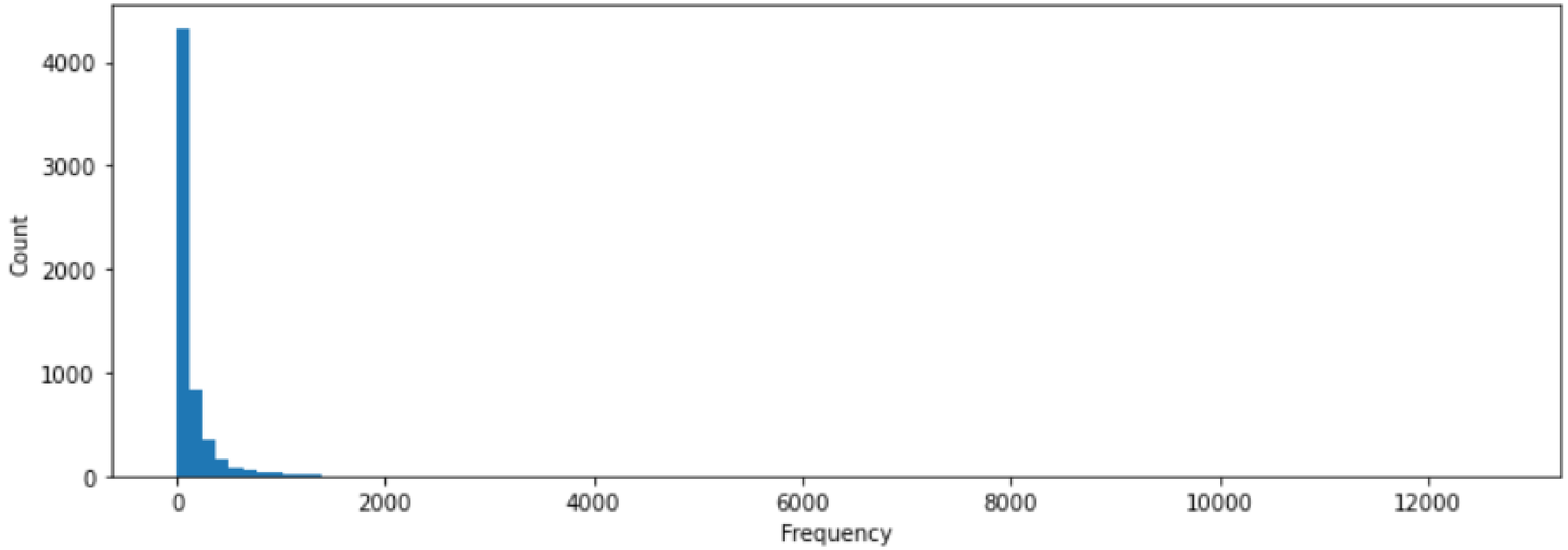


Clusters	count	mean	std	min	25%	50%	75%	max
0	673.0	621.983655	65.012623	516.0	568.0	618.0	674.00	738.0
1	1133.0	409.353045	49.093665	310.0	378.0	407.0	441.00	515.0
2	968.0	209.136364	52.880052	123.0	164.0	205.0	254.25	309.0
3	3165.0	35.881201	31.345800	0.0	9.0	27.0	57.00	122.0

Clustering

- ***Frequency (F) calculation and clustering***

To create frequency clusters, we need to find the total number of orders for each customer.

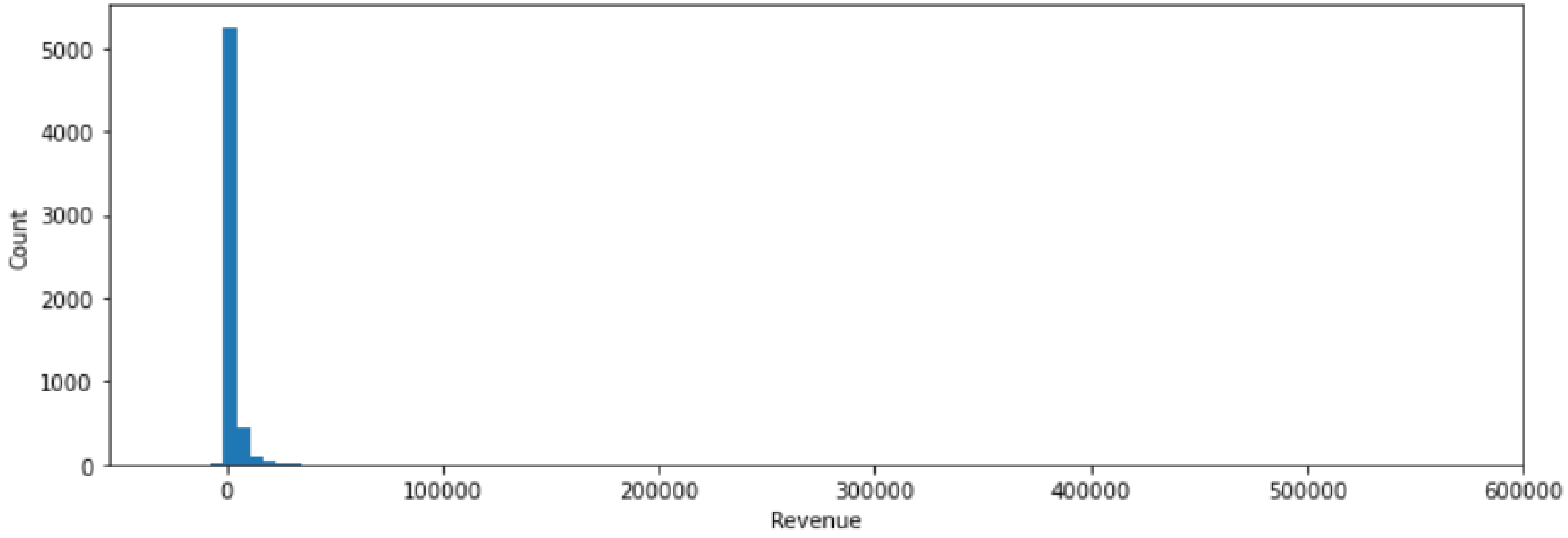


Clusters	count	mean	std	min	25%	50%	75%	max
0	5515.0	79.655485	85.588707	1.0	19.00	46.0	110.00	385.0
1	408.0	689.620098	306.629750	387.0	465.75	589.0	821.25	2077.0
2	14.0	3790.714286	1445.820844	2352.0	2728.00	3244.0	4463.75	6660.0
3	2.0	12040.000000	845.699710	11442.0	11741.00	12040.0	12339.00	12638.0

Clustering

- ***Monetary (M) calculation and clustering***

To create monetary clusters, we need to use the revenue.



Clusters	count	mean	std	min	25%	50%	75%	max
0	5858.0	1821.652251	2830.594995	-25111.09	317.0125	800.135	2052.270	21535.90
1	71.0	41740.448493	22088.363188	21893.53	25539.2050	33480.820	52250.470	111739.36
2	8.0	195182.235000	62923.301710	124961.98	142827.5300	179256.230	239982.390	296063.44
3	2.0	546861.340000	33261.270611	523342.07	535101.7050	546861.340	558620.975	570380.61

Clustering

- ***RFM calculation and clustering***

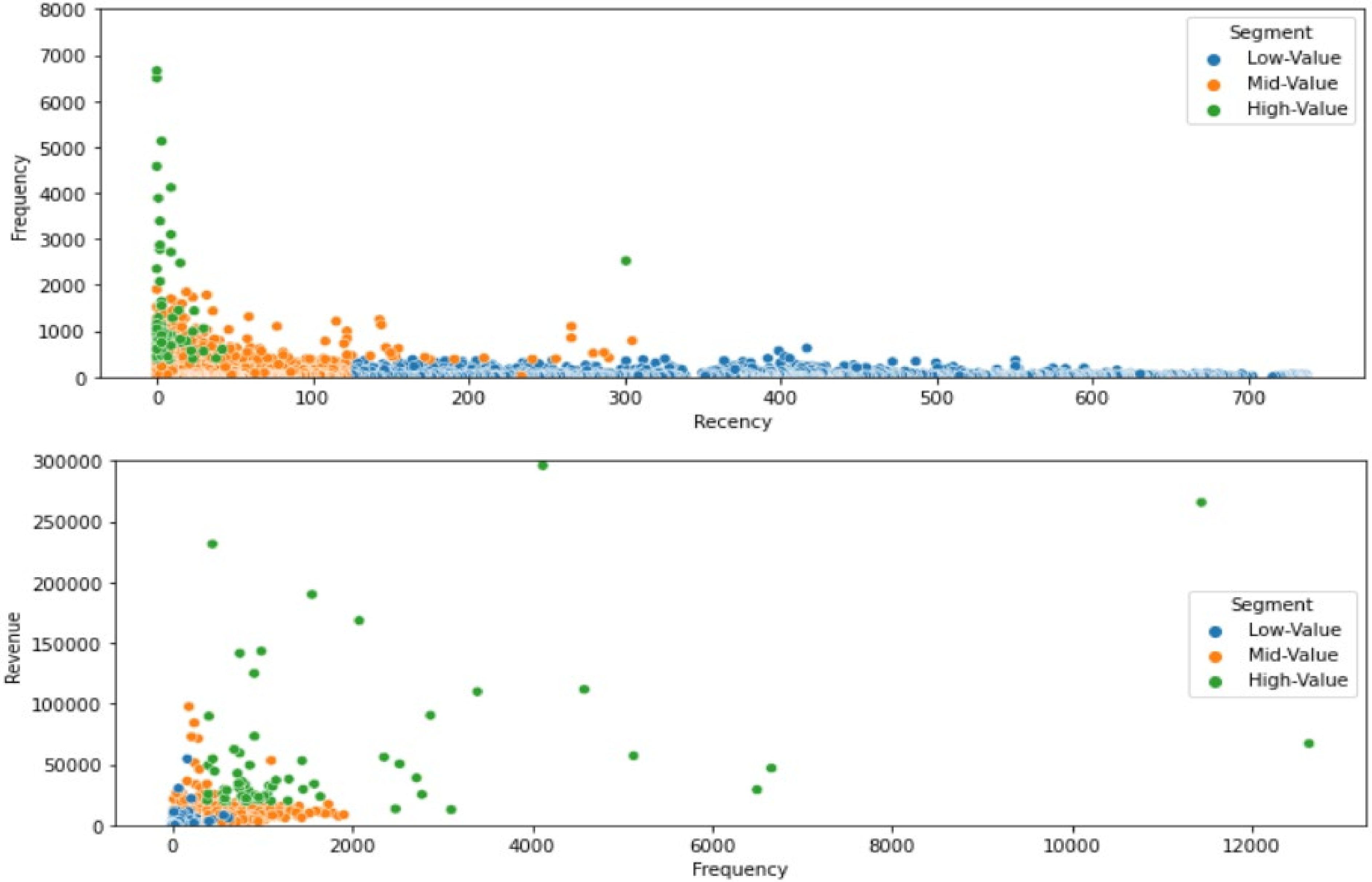
RMF values are calculated by adding up recency, frequency, and monetary scores.

RFM value	Recency	Frequency	Revenue
0	621.494815	23.662222	275.126803
1	408.922872	49.085106	748.954301
2	209.935857	64.652997	1168.003012
3	39.865217	114.338768	2105.548216
4	20.097765	633.480447	9776.187405
5	14.553191	974.808511	32971.483468
6	4.384615	2747.000000	103344.460769
7	2.400000	5155.000000	231052.704000
8	0.500000	7663.500000	394549.990000

Segmentation:

- 0 to 2: Low Value: good candidates for improving retention
- 3 to 4: Mid Value: good candidates for improving retention and increasing frequency
- 5+: High Value: good candidates for increasing frequency

Clustering


Segment
Low-Value
Mid-Value
High-Value
Frequency
Recency
Revenue
Frequency

Clustering

- ***RFM + Price + Quantity***

Scaling

	Price	Quantity	Recency	Frequency	Revenue
count	5939.000000	5939.000000	5939.000000	5939.000000	5939.000000
mean	19.835088	18.512167	201.784812	134.334905	2742.884541
std	405.031023	83.275333	211.727459	348.092835	13679.955199
min	0.151333	-16.000000	0.000000	1.000000	-25111.090000
25%	2.323423	5.155944	24.000000	20.000000	321.365000
50%	3.010753	9.075472	95.000000	52.000000	823.530000
75%	3.962383	13.409903	380.000000	140.000000	2143.280000
max	25111.090000	3255.074627	738.000000	12638.000000	570380.610000

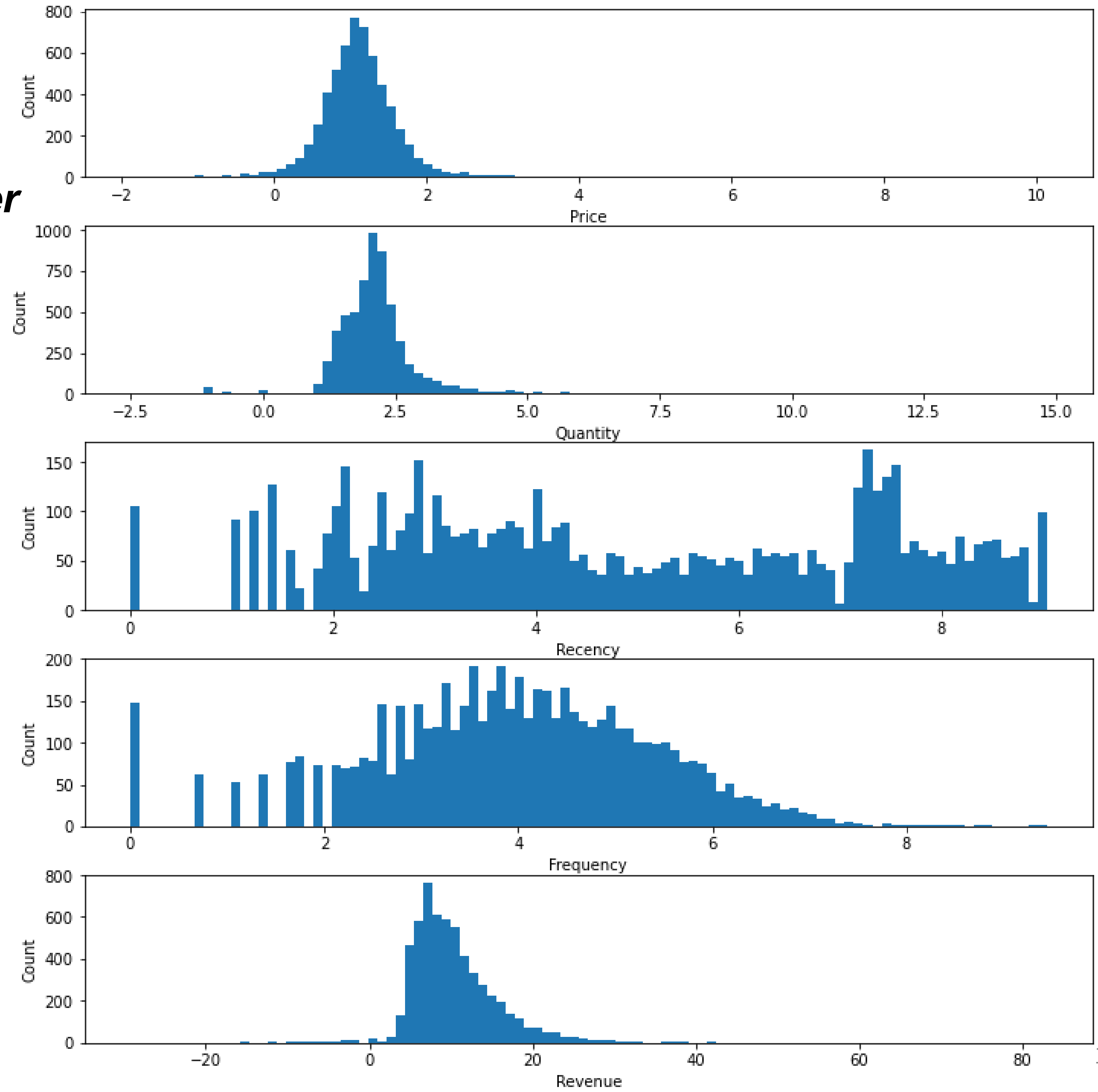
Logarithmic Transformation

Cubic root transformation

- ***Distribution of features after transformation***

- ***Silhouette analysis***

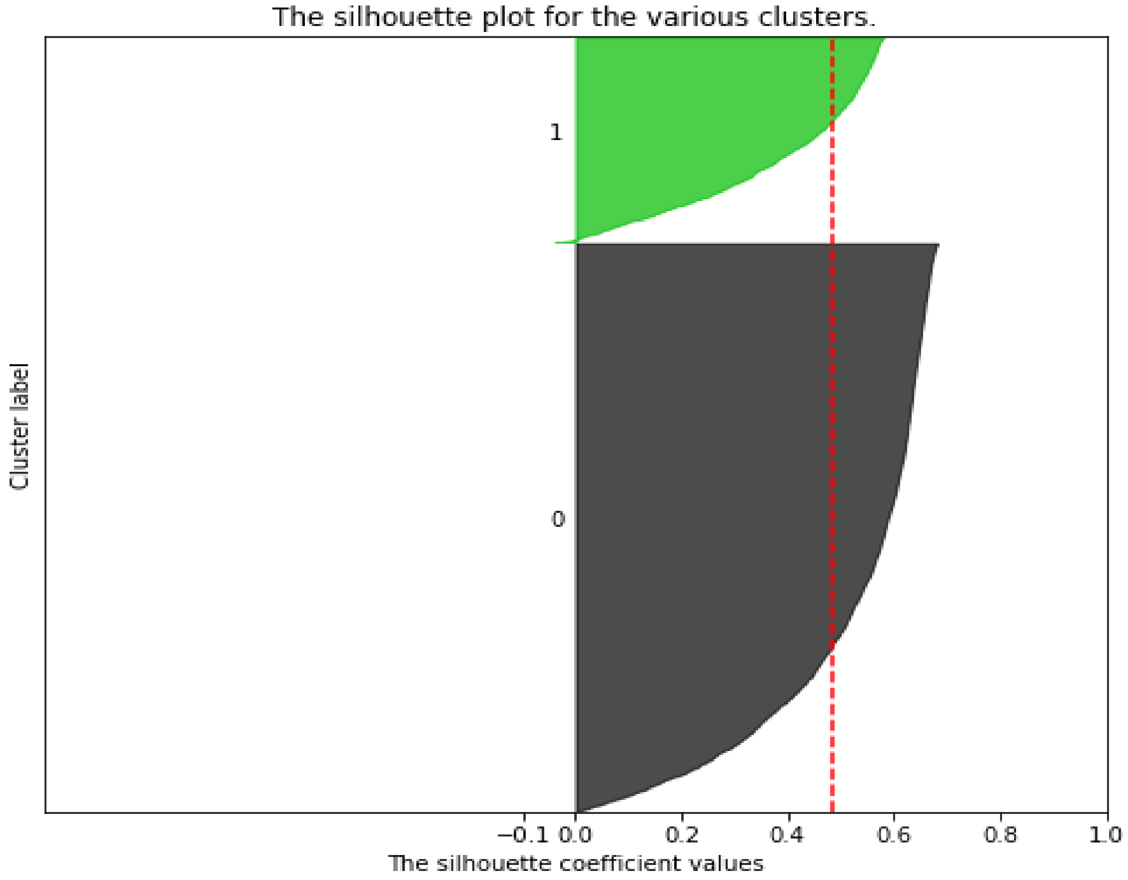
Clusters	Silhouette score
2	0.485
3	0.392
4	0.350
5	0.374

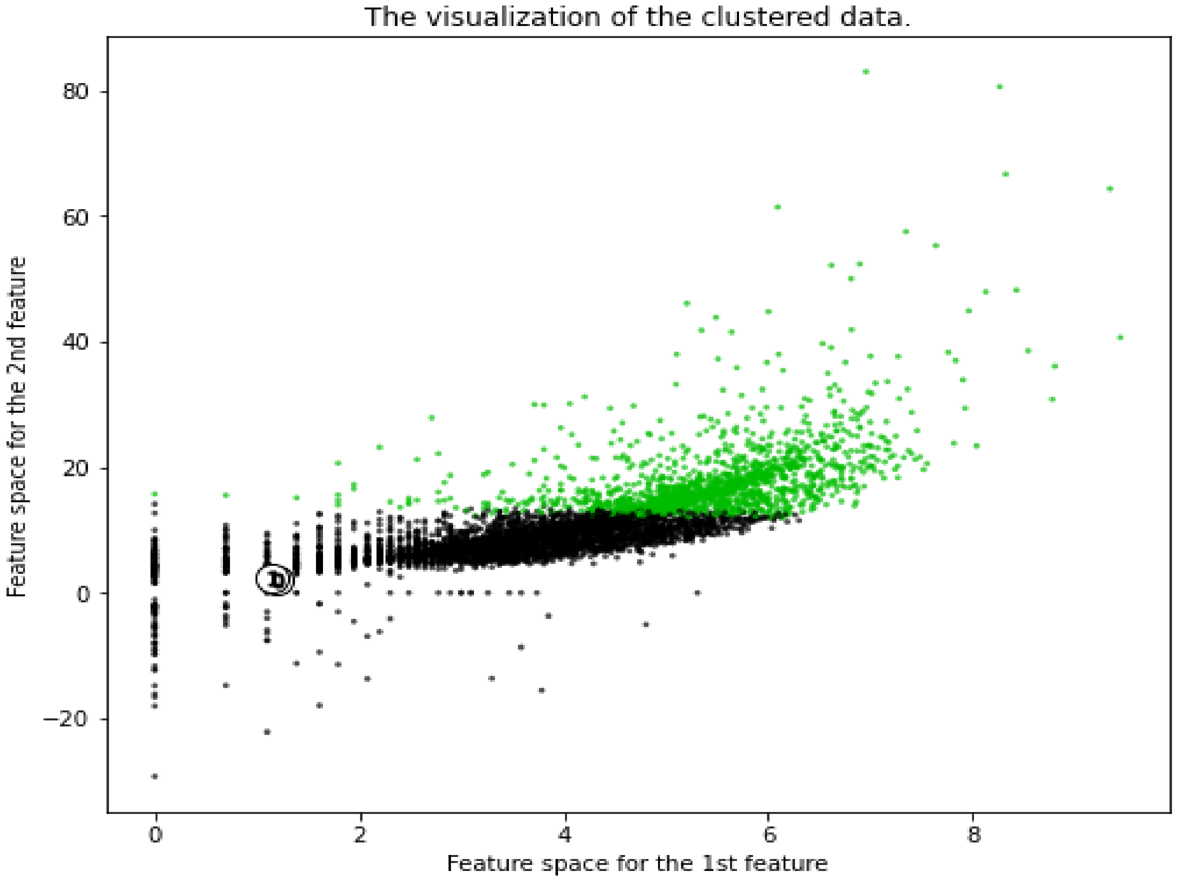


Clustering

- ***Silhouette analysis***

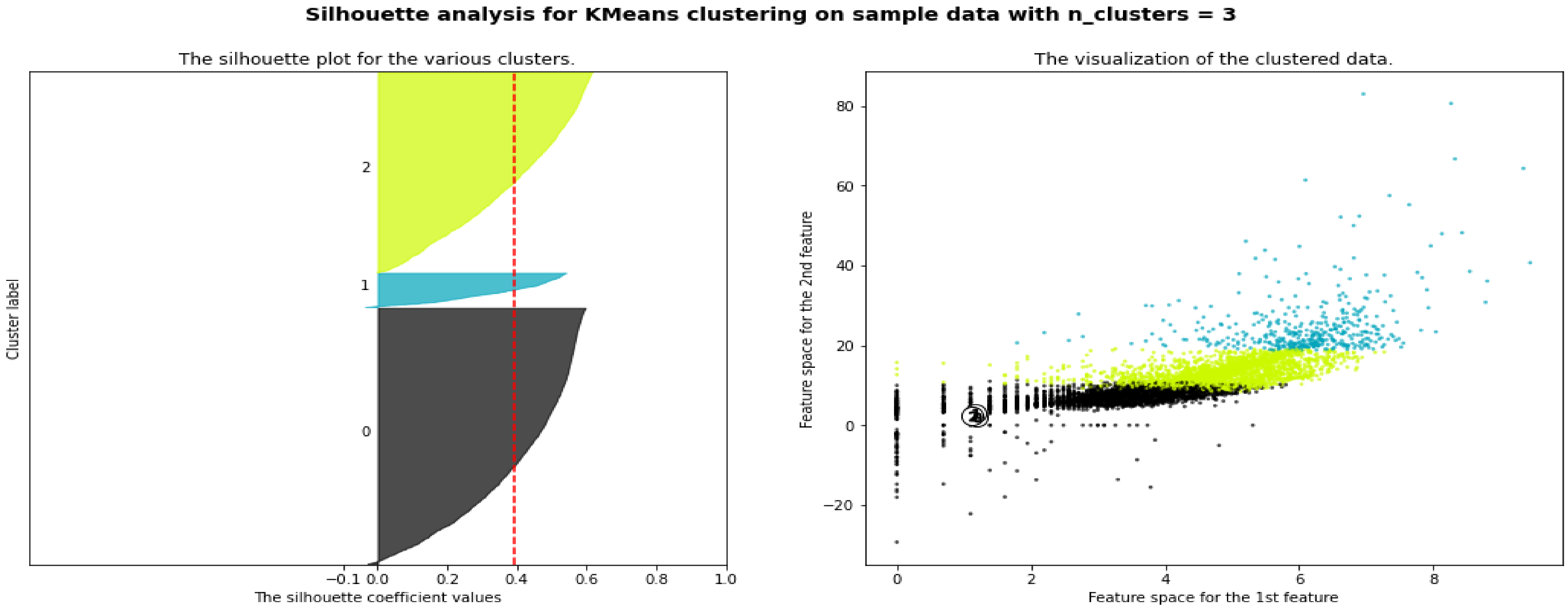
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2





Clustering

- ***Silhouette analysis***



Summary and conclusion

- The revenue analysis and customer analysis reveal the seasonality associated with sales and customer retention rates.
- The Elbow method and silhouette analysis are used to determine the best clusters of KMeans algorithm.
- Two or three clusters of customers are recommended. With various business plans or more information about customers, corresponding strategies may be developed for either two or three clusters

Recommendations & Future Work

- Using natural language processing tools to interpret the stock code and description features, which may reveal customers' preference for goods and further split subsets for customer segmentation.
- Collecting and importing more data into the modeling. For example, the cost of goods can be used to estimate profit. High revenue doesn't mean high profit. It is critical for a business to identify which goods have the highest profit and which group of customers contribute the highest profits.

Thank you