conclude

segment

loyalty

stats

process

about

Optimizing Customer Segmentation & Targeted Marketing Strategies through Advanced Analytics and Machine Learning

George Mason University

Guided By: Professor Eddy Zhang
AIT-614-001
Big Data Essentials

04 - 18 - 2024

Team 3

- Sardar Rohan Singh – Team Lead
  Mail: rsardar3@gmu.edu

- Bhargav Patel
- Monisha Jaganathan
- Saketh
- Sai Saketh

conclude

segment

loyalty

stats

process

Objectives

- **Decoding Customer Behavior**: Utilize EDA, advanced analytics, and ML to understand customer behavior intricacies.
- **Targeted Marketing Strategies**: Drive targeted marketing efforts based on insights, optimizing customer segmentation.
- **Optimizing Retail Operations**: Refine retail strategies to overcome market challenges, leveraging data-driven decision-making.

about

conclude
segment
loyalty
stats
process

Analytical Objectives

- ➢ *Use RFM Analysis for Customer Segmentation:*
  - o **Recency**: How recently did the customer purchase?
  - o **Frequency**: How often does the customer make purchases?
  - o **Monetary Value/Margin**: How much does the customer spend?

- ➢ *Utilize clustering algorithms (e.g., K-Means) for Segmentation*
  - o K-Means identifies customer segments by analyzing similarities in purchasing behavior.
  - o It optimizes segmentation by minimizing differences within clusters and maximizing differences between clusters, enabling tailored marketing strategies for diverse customer groups.

- ➢ Segment the best customers/profitable customers for the retail store.

about

conclude
segment
loyalty
stats
process

Timeline

➢ **Project Kickoff**	March 2
➢ **Literature Review and Gap Analysis**	March 9
➢ **Dataset Acquisition and Initial Data Analysis**	March 16
➢ **Development Environment Setup**	March 23
➢ **Data Cleaning and Preparation**	March 30
➢ **Data Model Implementation Phase 1**	April 6
➢ **Mid-Project Review and Adjustments**	April 13
➢ **Final Data Model Refinement**	April 18

about

conclude
segment
loyalty
stats

Description of the Dataset

The Dataset is based on the UK retain chain store. It is sourced from a publicly available repository for authenticity.

Dataset Link: https://doi.org/10.24432/C5CG6D

8 attributes and 541909 records:

1. ***Invoice No:*** A unique identifier for each transaction made by the customer.
2. ***Stock Code****:* An identifier for the product or item purchased in the transaction.
3. ***Description****:* A textual description of the product or item purchased.
4. ***Quantity****:* The quantity of the product or item purchased in the transaction.
5. ***Invoice Date****:* The date and time when the transaction occurred.
6. ***Unit Price****:* The price of a single unit of the product or item purchased.
7. ***Customer ID***: A unique identifier for each customer making the transaction.
8. ***Country****:* The country where the transaction took place.

process
about

conclude
segment
loyalty
stats

Data Processing & Analytics Approach

Data 1

Kaggle

Retail chain billing data

Imputation 2

Manipulation

Clean the data ready for analyzing

Classify 3

Segmentation

Classify the data based on the objectives

process
about

conclude
segment
loyalty
stats

Data Processing

- **Data Loading**: Load the raw data into the Spark Data Frame.
- **Data Cleaning**: Handle missing values, outliers, and inconsistencies in the data.
- **Feature Engineering**: Create new features or modify existing ones to extract valuable insights.
- **Scaling**: Standardize the features to bring them to the same scale, ensuring fair comparison.
- **Vectorization:** Assemble the features into a single vector for machine learning algorithms.

process
about

conclude
segment
loyalty
stats

Analytics Approach

- **Exploratory Data Analysis (EDA):** Understand the data distribution and relationships between variables.
- **RFM Analysis**: Calculate Recency, Frequency, and Monetary Value/Margin metrics to segment customers.
- **K-Means Clustering**: Apply K-Means algorithm to cluster customers based on RFM metrics.
- **Cluster Analysis**: Analyze and interpret the characteristics of each customer cluster.
- **Visualization**: Visualize the clusters to gain insights and communicate findings effectively.

process
about

conclude
segment
loyalty
stats

Software Platform

- **Python**: A versatile programming language for data analysis, machine learning, and web development.
- **Benefits**: A rich ecosystem of libraries and frameworks. Simplicity and readability, facilitate faster development.
- **Usage**: Developing data preprocessing scripts, machine learning models, and analytical tools.

process
about

Hardware Platform

- **AWS**: Leading cloud computing platform offering scalable infrastructure and resources.
- **Components**: Can utilize AWS EC2 instances for deployment and management.
- **Storage**: Storing and accessing datasets and intermediate results using AWS S3 buckets.

Benefits:

- **Scalability**: Easily scale resources to meet growing demands.
- **Reliability**: AWS infrastructure is highly reliable and fault-tolerant.

conclude
segment
loyalty
stats

Hardware Platform

Integration with Databricks:

- Seamless integration with Databricks allows for leveraging the power of AWS for data processing and analytics.
- Databricks clusters can be provisioned on AWS EC2 instances, and data can be stored in S3 buckets for efficient processing.

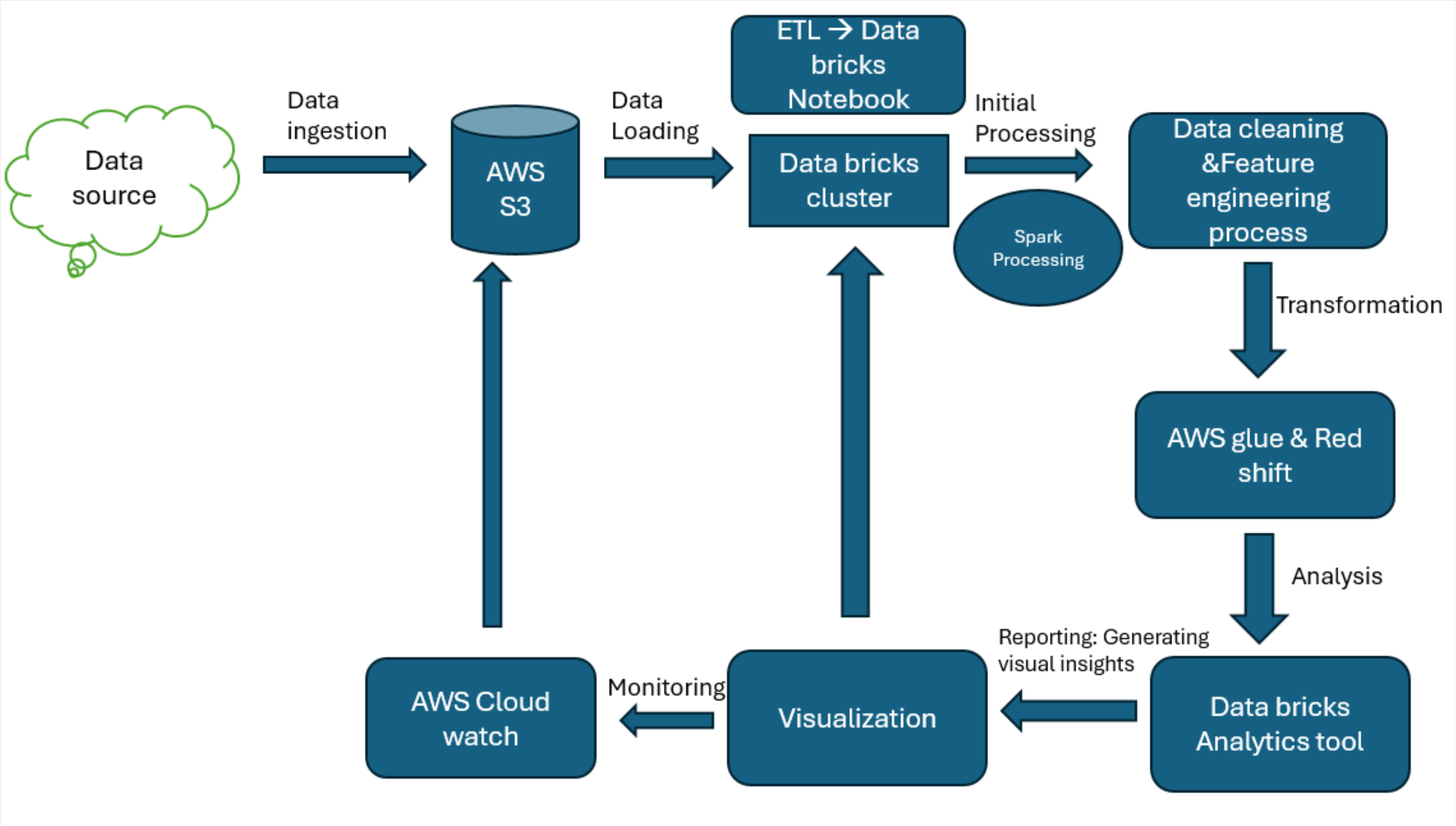
process
about

Note: In our project, we mainly focused on using Databricks and PySpark to process the data and perform analysis and visualization.

Considering AWS integration with Databricks was a learning process to understand how to Built the ETL Pipeline.

conclude
segment
loyalty
stats

Architecture Framework



process
about

conclude
segment
loyalty
stats

Integrating Databricks with AWS

Step 1:

- ➢ Databricks Cluster: Provision a Databricks cluster on AWS EC2.
- ➢ ETL Scripts: Write Python or Spark scripts for data processing in Databricks notebooks.
- ➢ Spark Processing: Utilize Spark for distributed data processing and transformation.

process
about

conclude
segment
loyalty
stats

Integrating Databricks with AWS

Step 2: Data Storage on AWS

- AWS S3: Store raw and processed data in AWS S3 buckets.
- AWS Glue: Use AWS Glue for data cataloging and metadata management.
- AWS Redshift: Store processed data in AWS Redshift for analytics.

process
about

Integrating Databricks with AWS

Step 3: Analytics and Monitoring

- Databricks Analytics: Perform advanced analytics and machine learning on Databricks.
- Visualization: Visualize insights using Databricks notebooks and visualization libraries.
- Monitoring: Monitor Databricks clusters and AWS resources with AWS CloudWatch.

conclude

segment

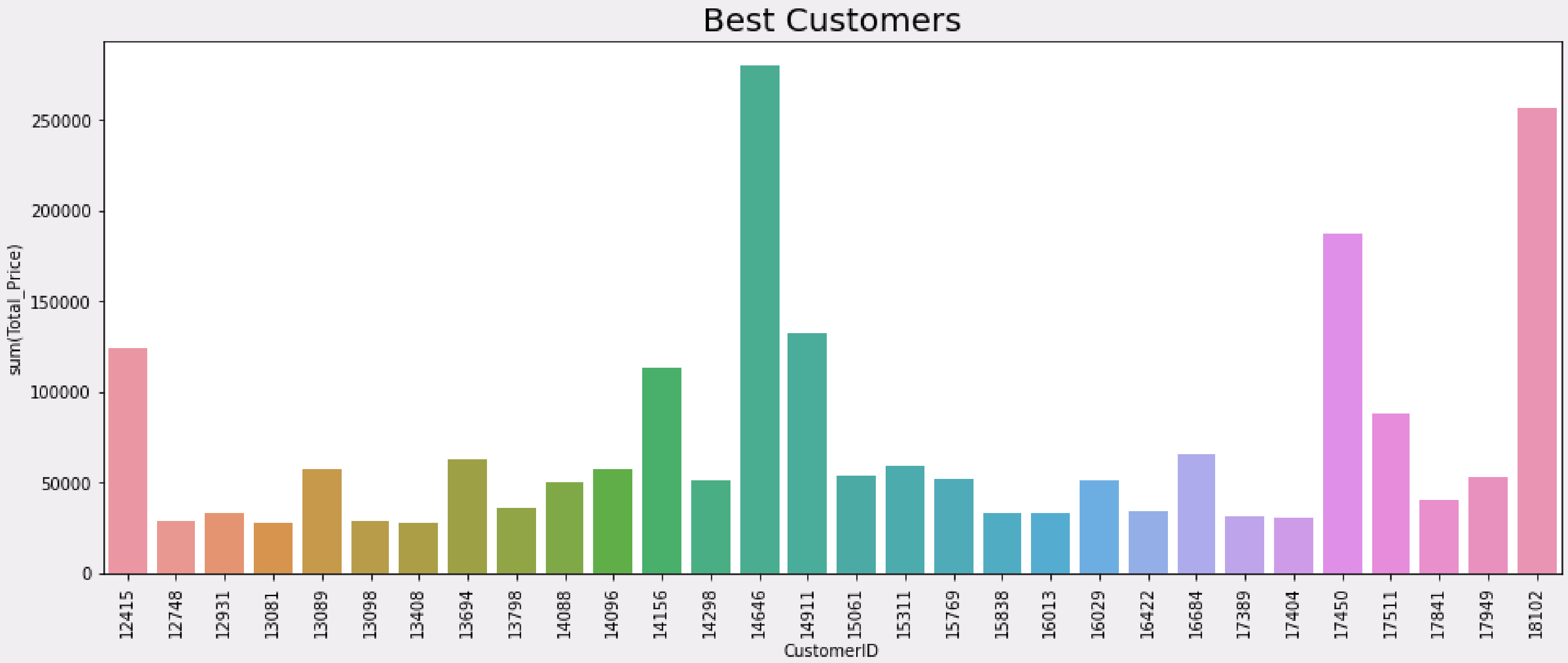
loyalty

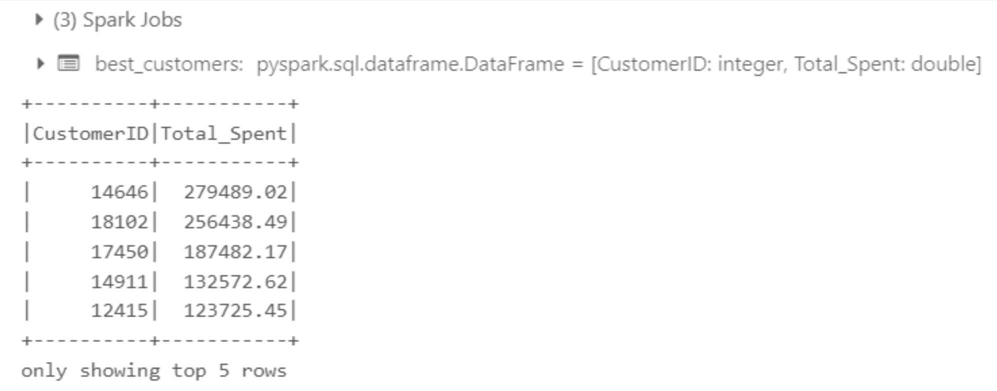
stats

process

about

Best 30 Customers





▸ best_customers: pyspark.sql.dataframe.DataFrame = [CustomerID: integer, Total_Spent: double]

```
+----------+-----------+
|CustomerID|Total_Spent|
+----------+-----------+
|     14646|  279489.02|
|     18102|  256438.49|
|     17450|  187482.17|
|     14911|  132572.62|
|     12415|  123725.45|
+----------+-----------+
only showing top 5 rows
```

conclude
segment
loyalty

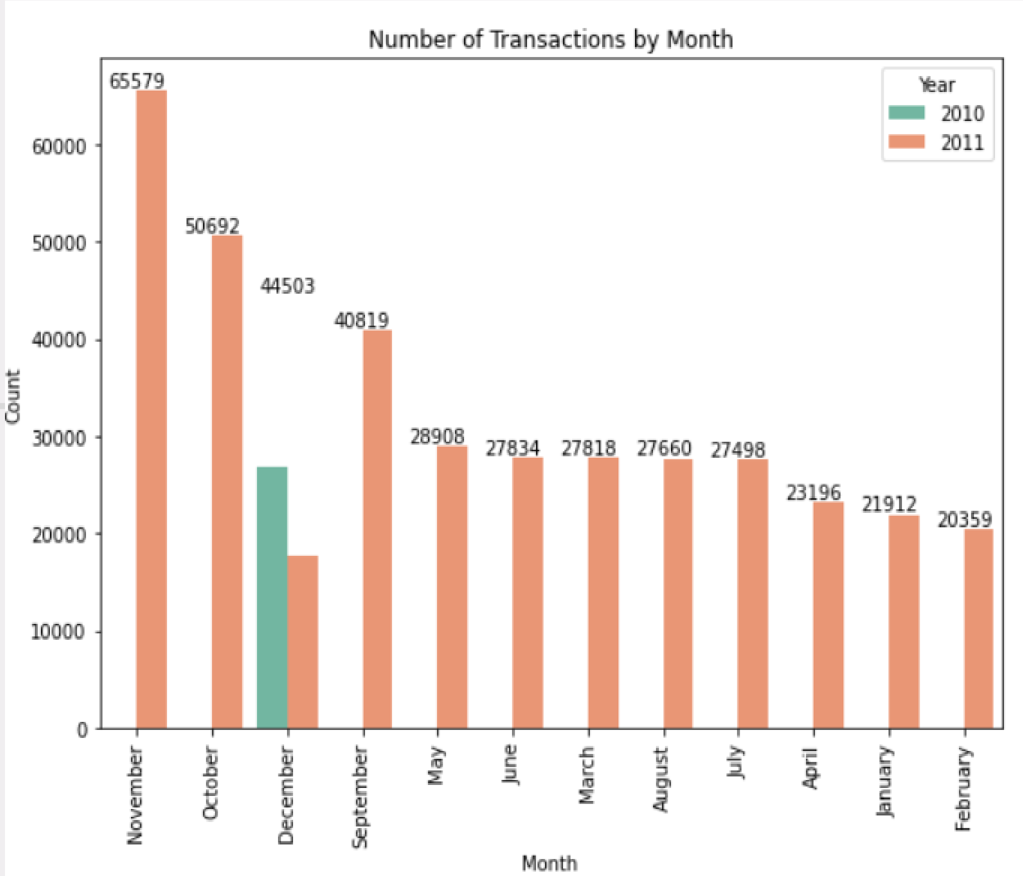
Periodical Purchasing Stats – Timeline

Month
Weekday
Time period

stats
process
about

Periodical Purchasing Stats – timeline

Monthly sales data

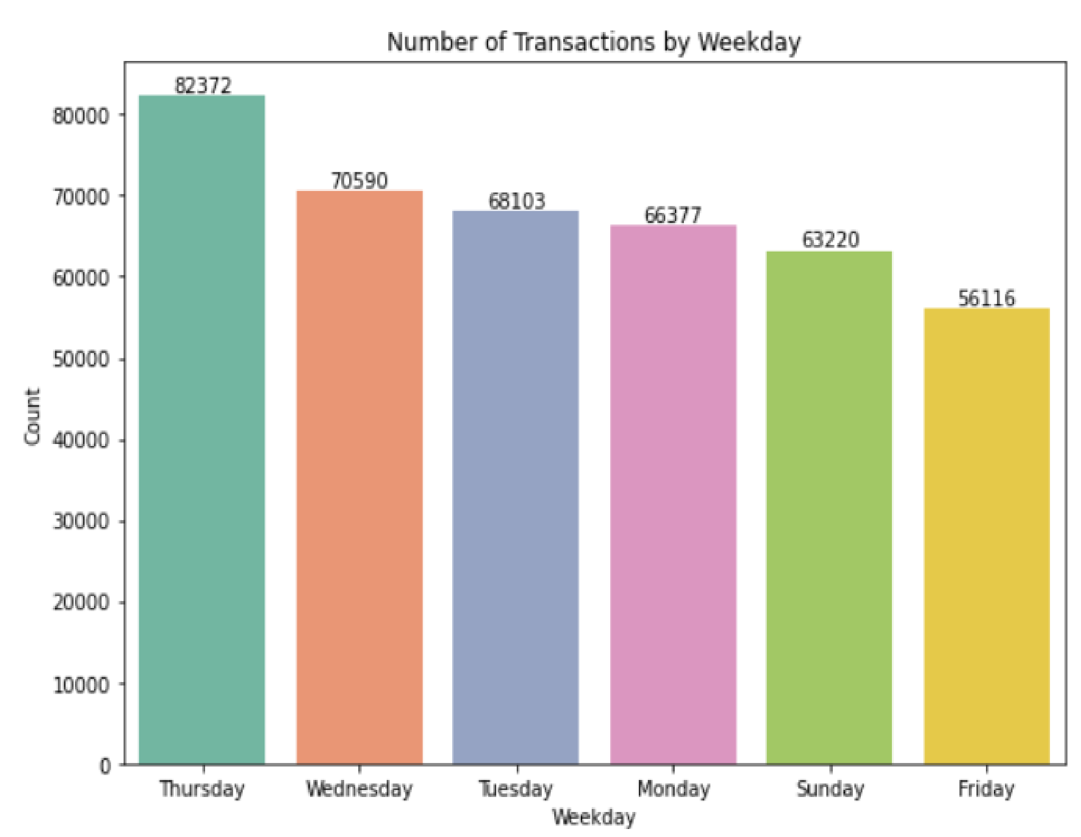


conclude
segment
loyalty
stats
process
about

conclude
segment
loyalty

Periodical Purchasing Stats – timeline

weekday
sales data



stats
process
about

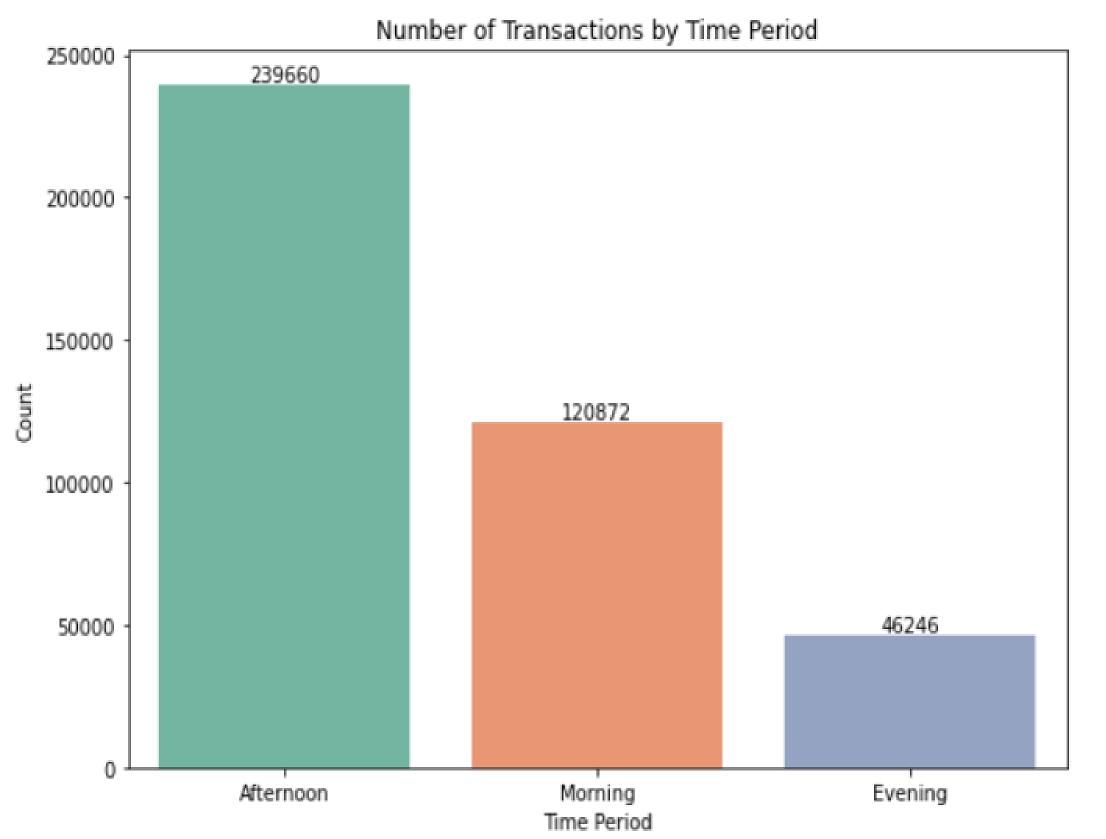
conclude

segment

loyalty

Periodical Purchasing Stats – timeline

Time period sales data

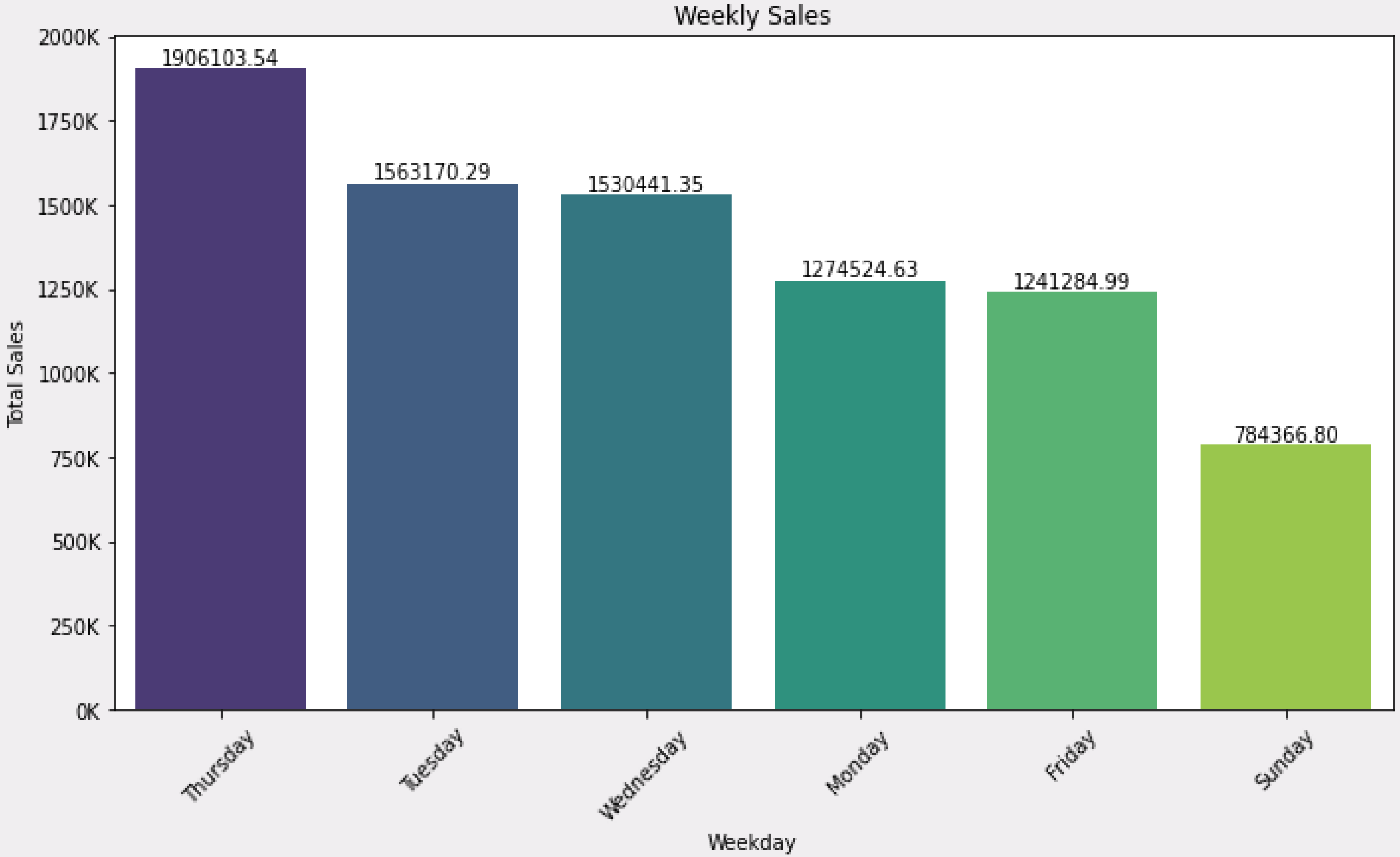


stats

process

about

Most Revenue Generated Week-Day



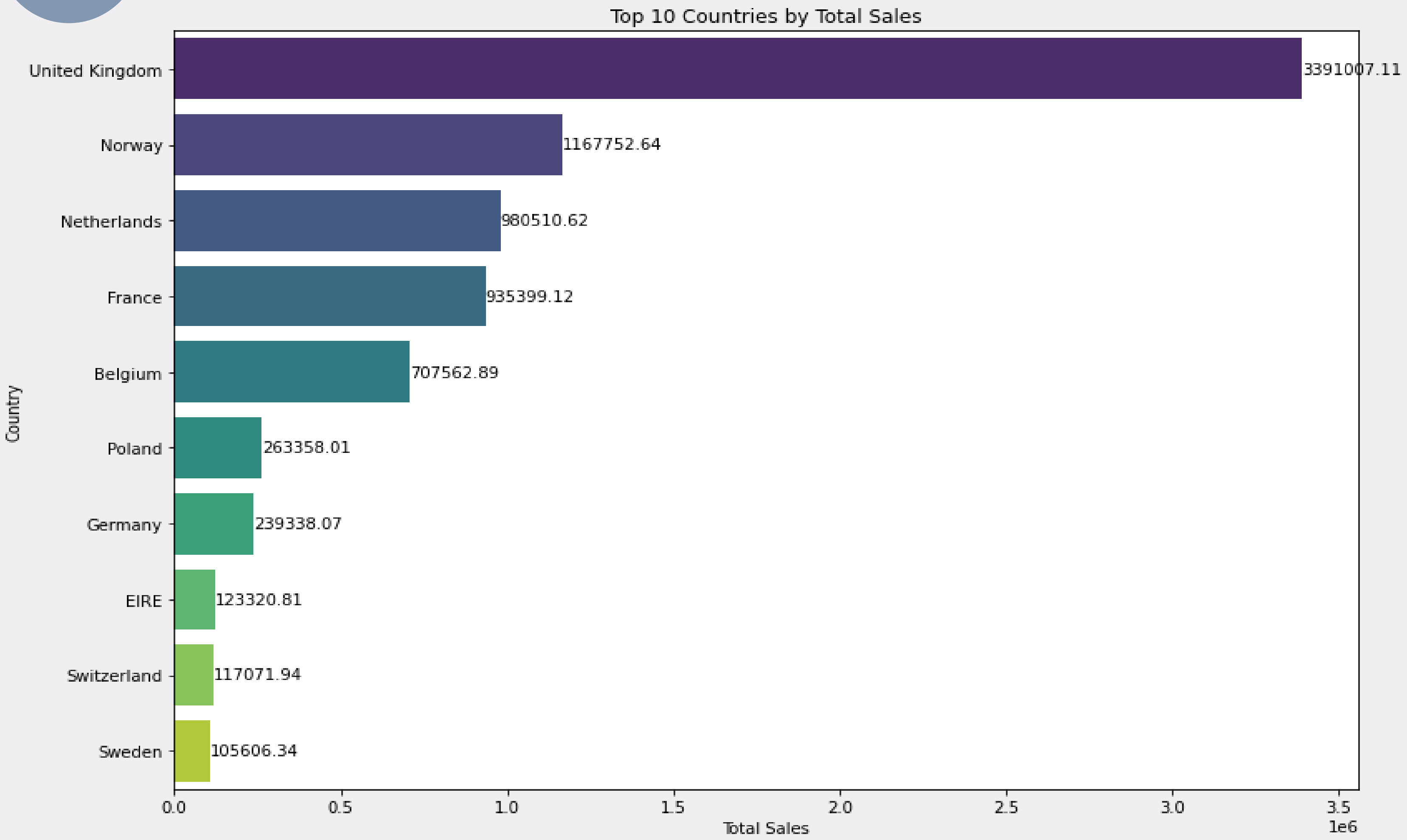
conclude

segment

loyalty

Country Sales Information

UK is anyway its primary location
Norway, France, Netherlands are in order to focus on

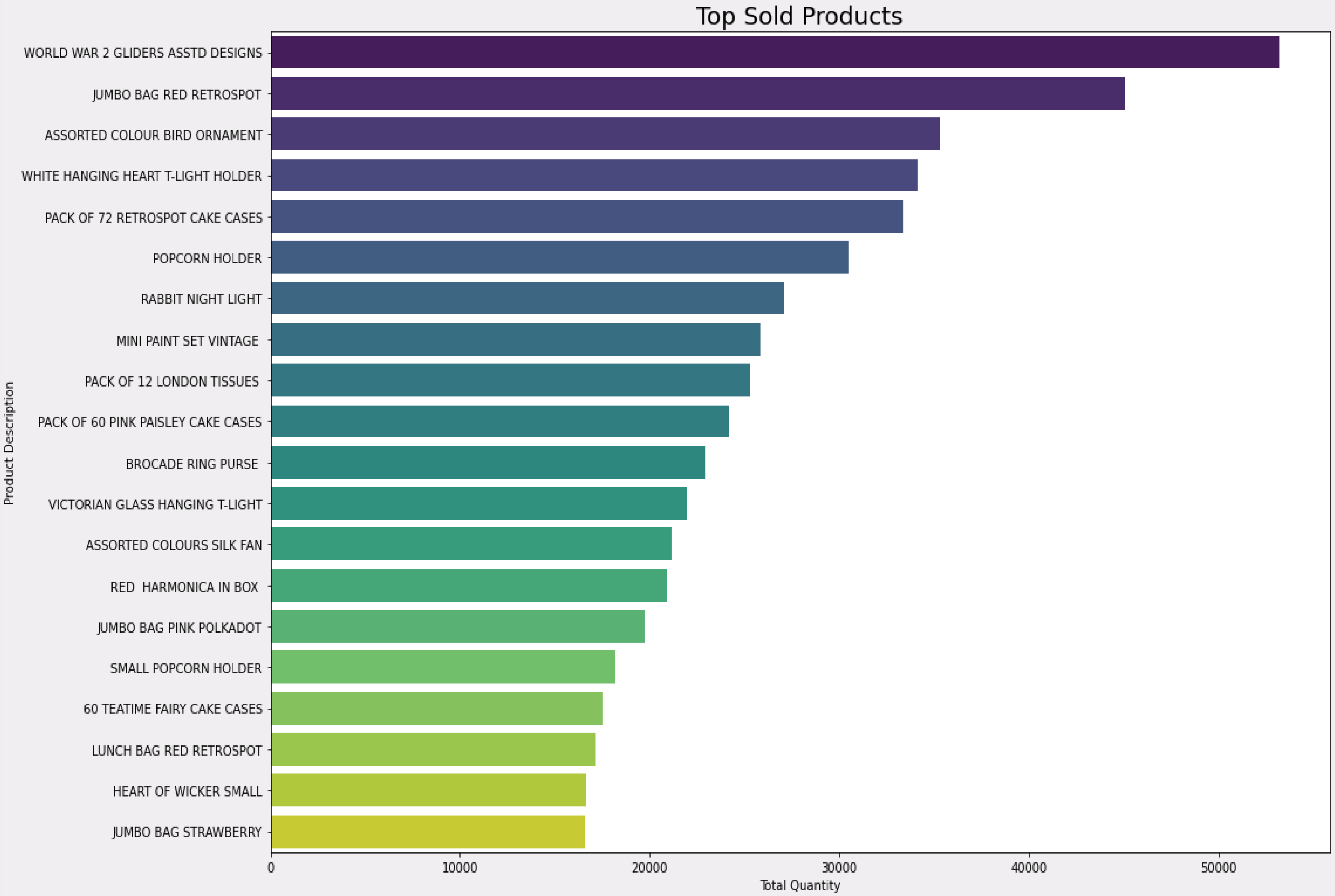


stats

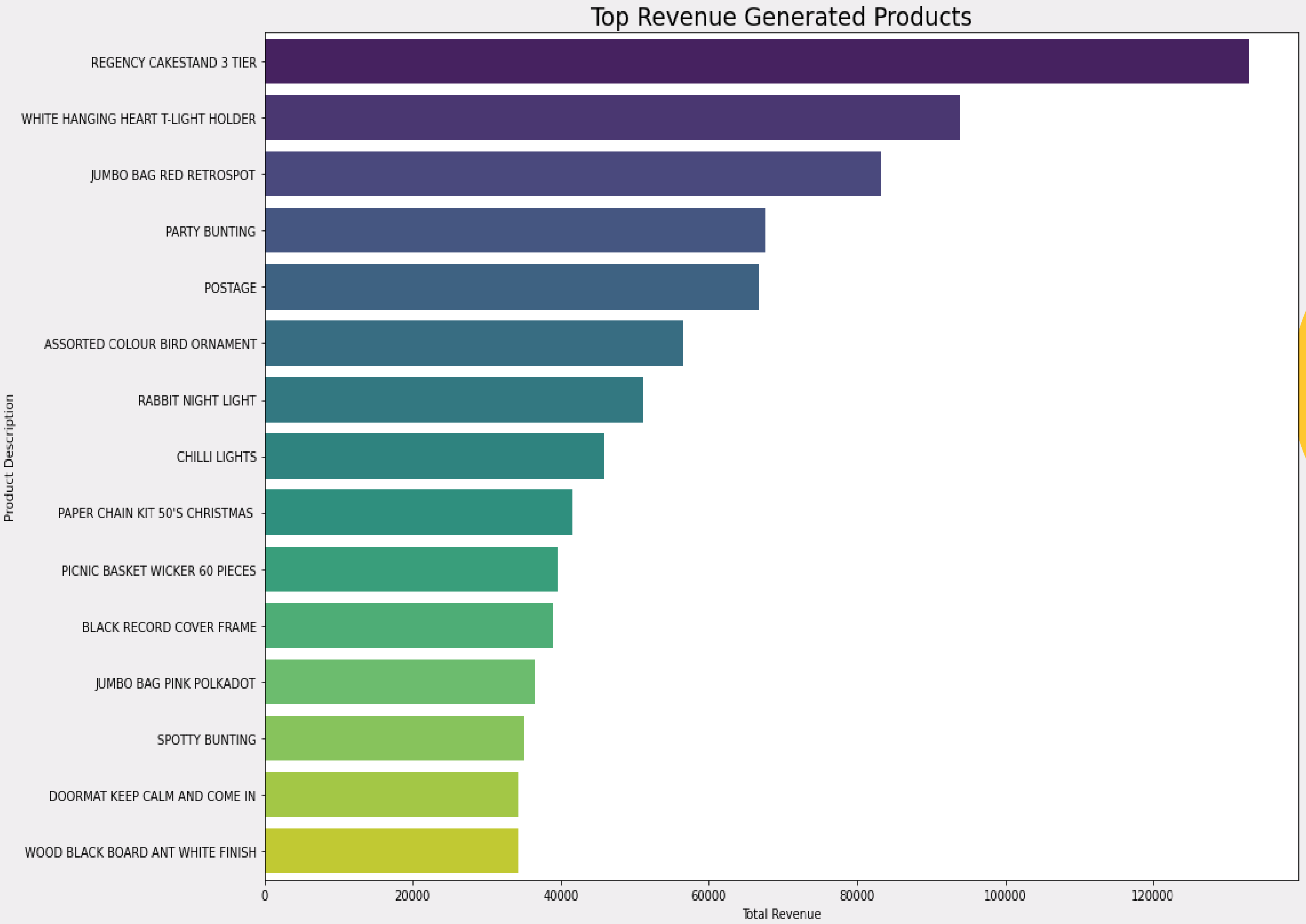
process

about

Top Sold Products



Top Revenue Generated Products



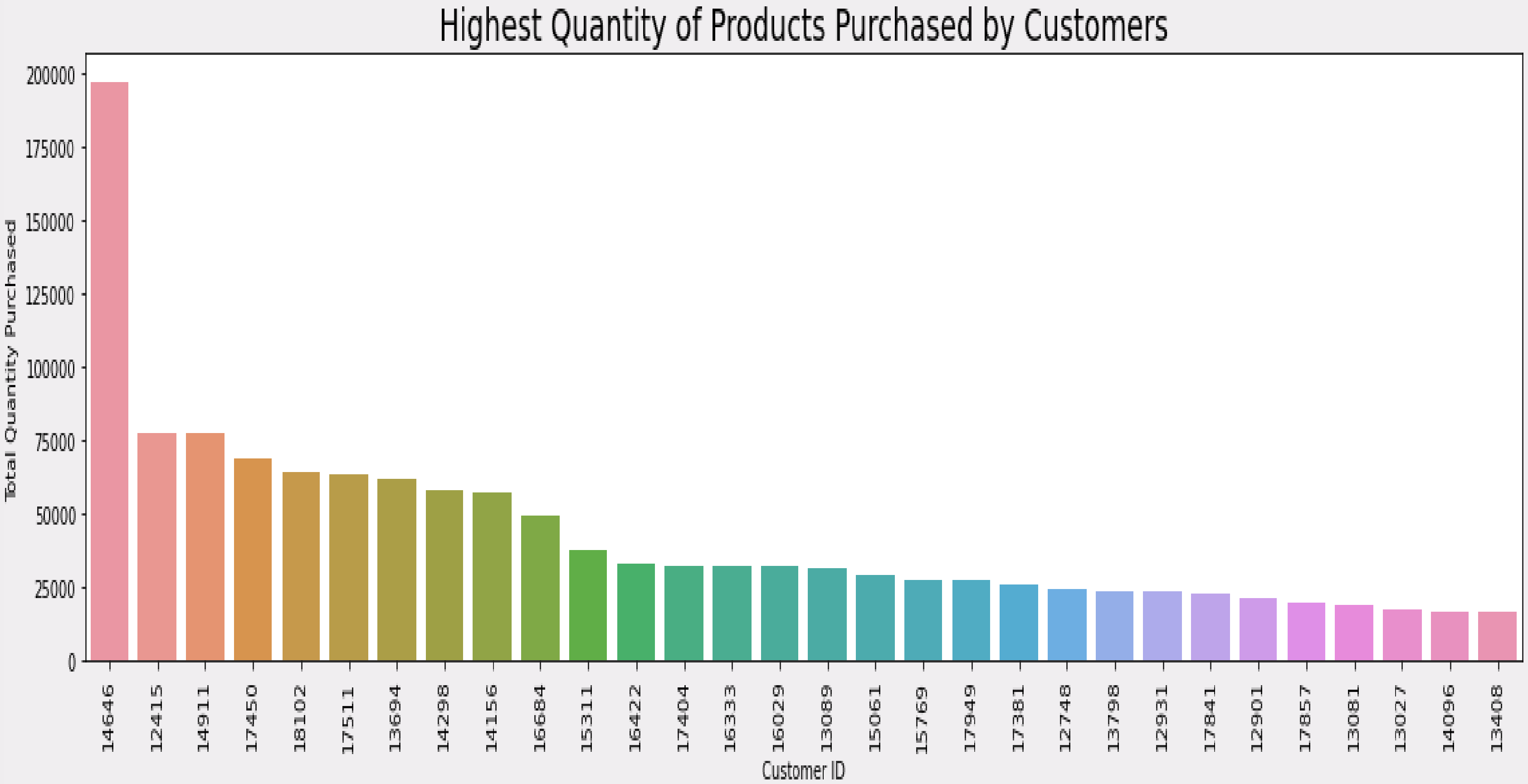
conclude
segment
loyalty
stats
process
about

conclude
segment
loyalty

Highest Sales Information

Based on the customer ID, we have plotted the quantity of products our customers buy.

Based on these sales, we can segment our customers

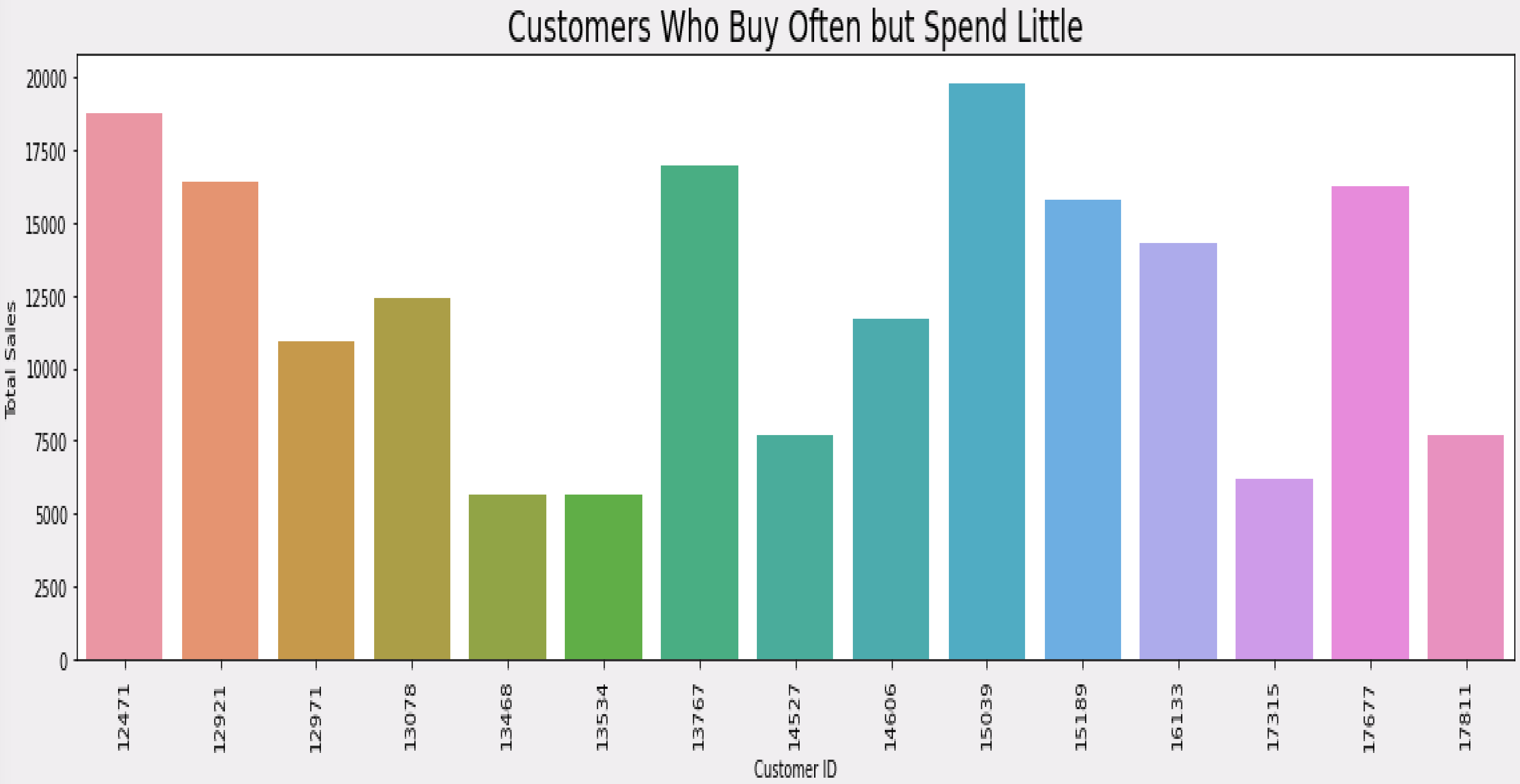


stats
process
about

Customers buying often but spend less

Considering 20,000 is less payment from Daily Customers

Below are the Top 15 customers who buy



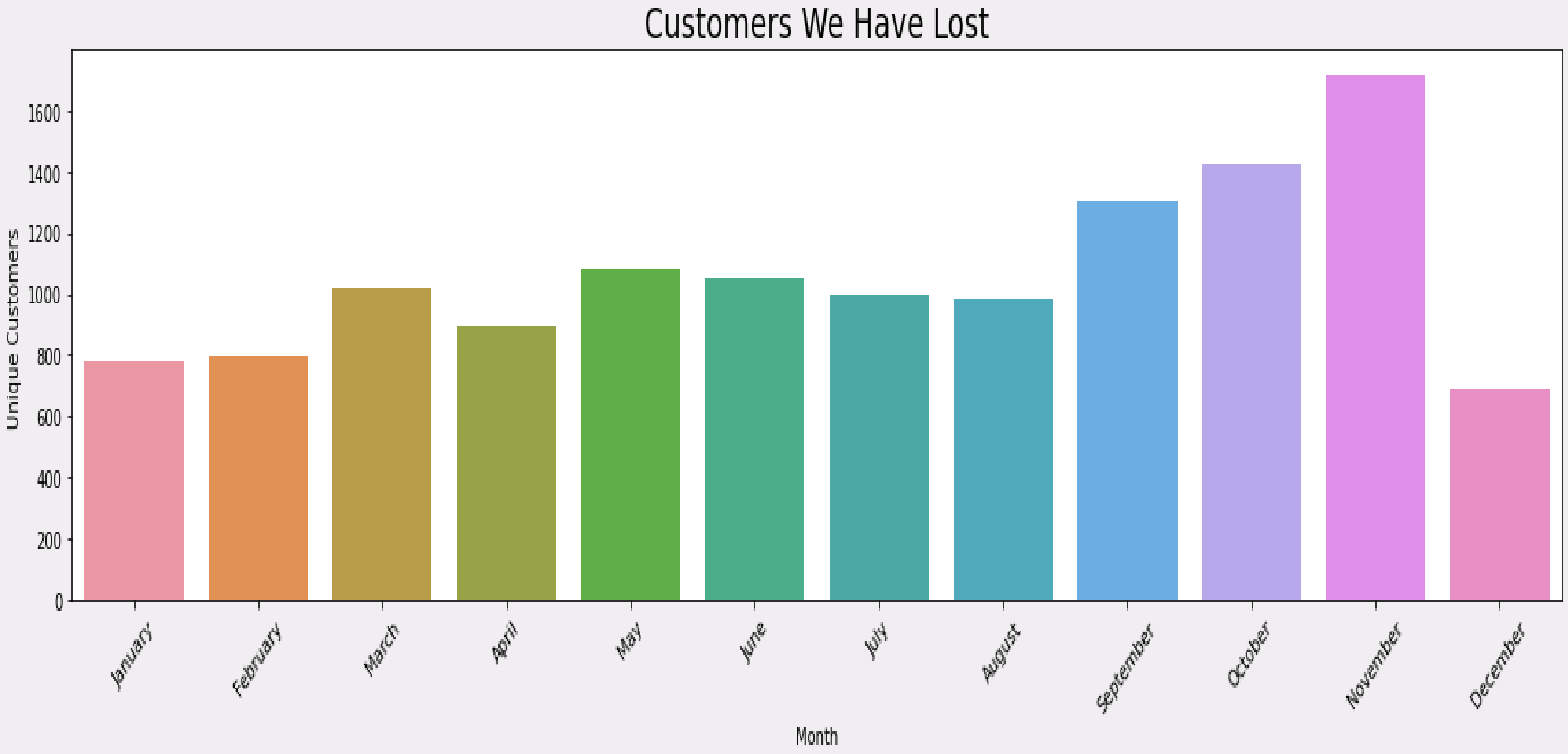
conclude
segment
loyalty
stats
process
about

Customers lost monthly

Considering the year 2011, We have analyzed number of customers lost monthly

Below We can see that , More customers were lost by end of the year.

With this company can focus on year end sales and prepare strategies



conclude
segment

Customers Loyalty segmentation

01 **Gold**

>1000 sales

02 **Silver**

500 < 1000

03 **Bronze**

<500

loyalty
stats
process
about

Why Segment Customers

Targeting customers

Create and communicate targeted marketing messages

Test pricing options

Greatest advantage to select a price that generates slightly less revenue if that price also generates more new customers.

Best communication channel

Might be email, social media posts, radio advertising, or another approach, depending on the segment.

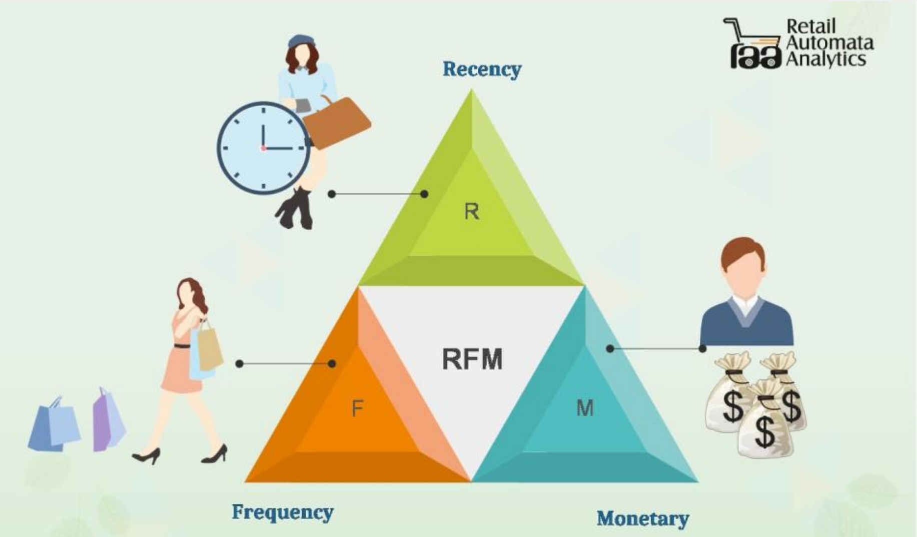
Analyze profitable customers

Focus on the customers who are more profitable because 80% of income comes from 20% of customer

conclude
segment
loyalty
stats
process
about

Segmentation with RFM Analysis

RFM stands for Recency - Frequency - Monetary Value.
Theoretically we will have segments like below:



Low Value: Customers who are less active than others.

Mid Value: Fairly frequent and generates moderate revenue.

High Value: High Revenue, Frequency and low Inactivity.

conclude
segment
loyalty
stats
process
about

Overview of RFM

Recency

The freshness of customer activity like last purchase

E.g. Number of days since last order

Frequency

The frequency of customer transactions

E.g. Average number of days between transactions

Monetary Value

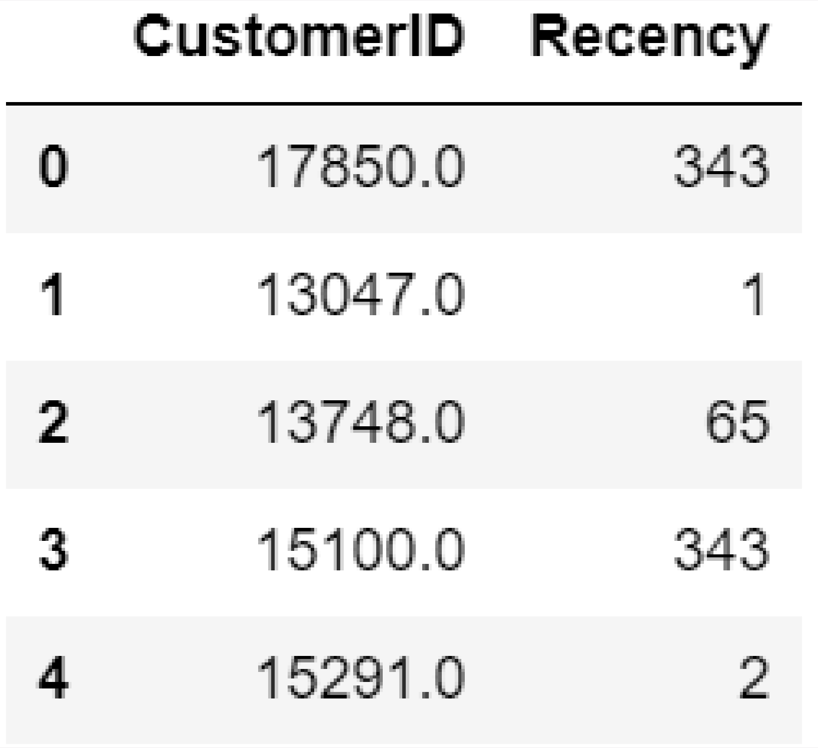
Total revenue that a customer contributes

E.g. Total or average transactional value

conclude
segment
loyalty
stats
process
about

Calculate Recency values

We calculate how recently a customer has made a purchase



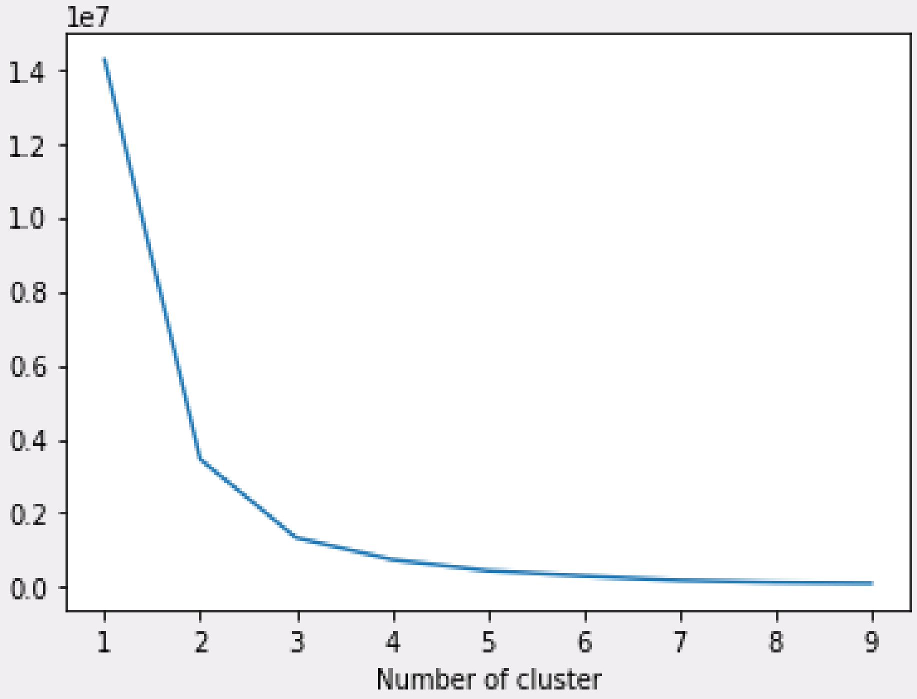
	CustomerID	Recency
0	17850.0	343
1	13047.0	1
2	13748.0	65
3	15100.0	343
4	15291.0	2

conclude
segment
loyalty
stats
process
about

Elbow Method

We are going to apply K-means clustering to assign a recency score.
But we should tell how many clusters we need for K-means algorithm.
To find it out, we will apply Elbow Method.

Elbow Method simply tells the optimal cluster number for optimal inertia.

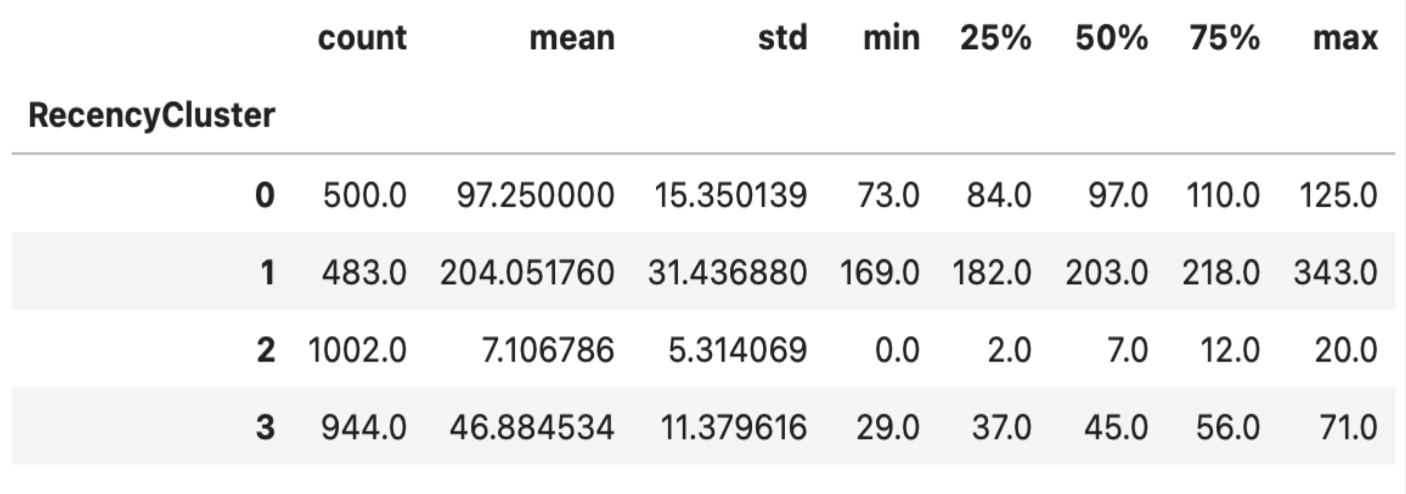


Here it looks like 3 is the optimal one.

Based on business requirements, we can go ahead with less or more clusters. Here we select 4

conclude

Building 4 clusters for recency



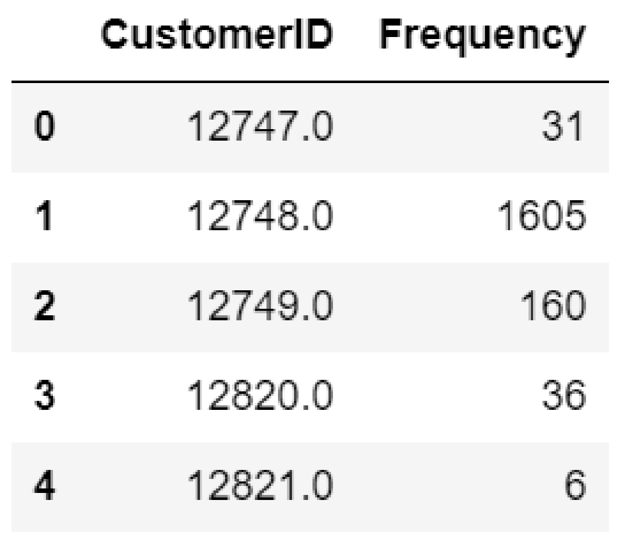
	count	mean	std	min	25%	50%	75%	max
RecencyCluster								
0	500.0	97.250000	15.350139	73.0	84.0	97.0	110.0	125.0
1	483.0	204.051760	31.436880	169.0	182.0	203.0	218.0	343.0
2	1002.0	7.106786	5.314069	0.0	2.0	7.0	12.0	20.0
3	944.0	46.884534	11.379616	29.0	37.0	45.0	56.0	71.0

segment
loyalty
stats
process
about

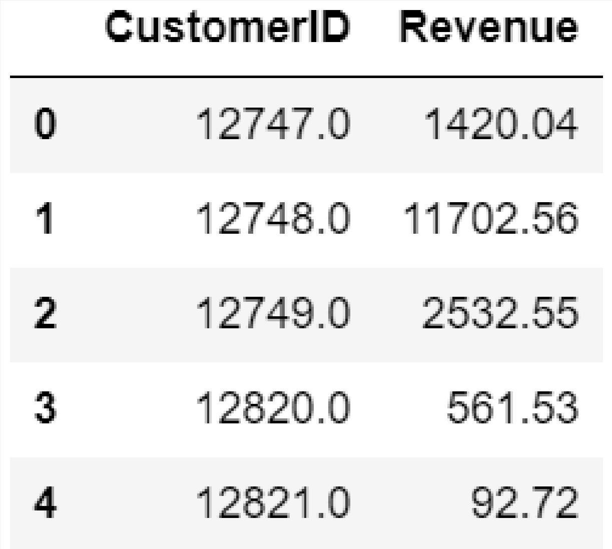
Cluster 2 has customers with the best(low) recency [recent visits are with 20 days] and Cluster 1 has customers with high recency value [recent visits are 169 days to 343 days ago

Calculate Frequency and Monetary values

We applied same method for frequency and revenue values



	CustomerID	Frequency
0	12747.0	31
1	12748.0	1605
2	12749.0	160
3	12820.0	36
4	12821.0	6

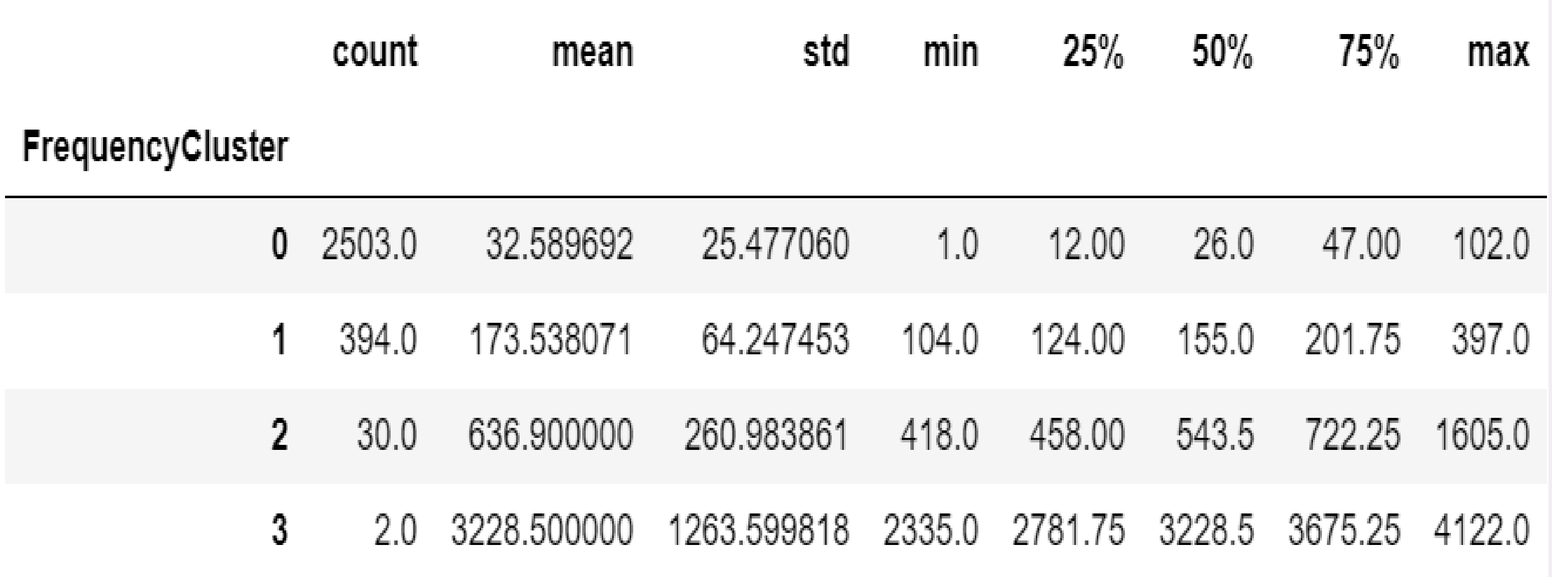


	CustomerID	Revenue
0	12747.0	1420.04
1	12748.0	11702.56
2	12749.0	2532.55
3	12820.0	561.53
4	12821.0	92.72

conclude
segment
loyalty
stats
process
about

Building 4 clusters for Frequency

To create frequency clusters we need to find total number of orders for each customer. First we try to calculate this and see how frequency looks like in customer database



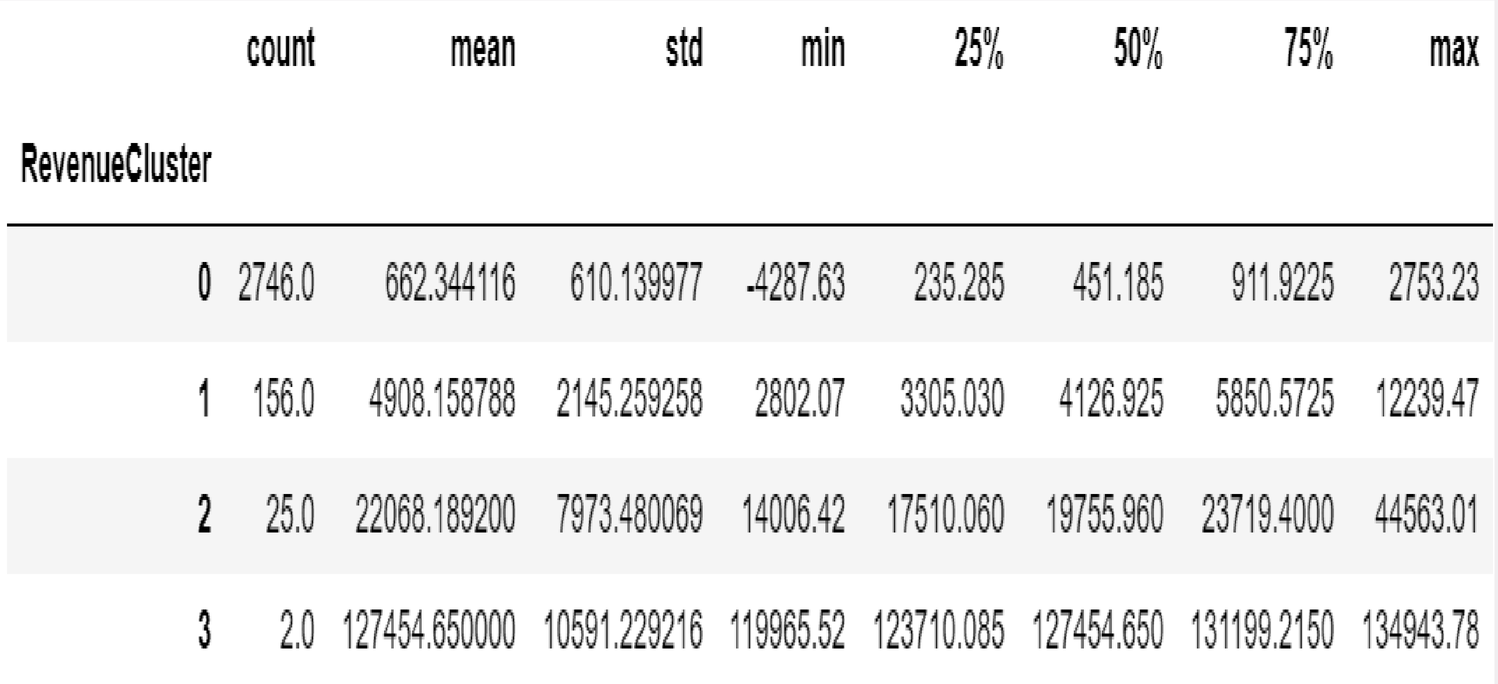
FrequencyCluster	count	mean	std	min	25%	50%	75%	max
0	2503.0	32.589692	25.477060	1.0	12.00	26.0	47.00	102.0
1	394.0	173.538071	64.247453	104.0	124.00	155.0	201.75	397.0
2	30.0	636.900000	260.983861	418.0	458.00	543.5	722.25	1605.0
3	2.0	3228.500000	1263.599818	2335.0	2781.75	3228.5	3675.25	4122.0

Cluster 3 has customers with higher frequency than other clusters.

Note: High frequency number indicates better customers

conclude
segment
loyalty
stats
process
about

Building 4 clusters for Revenue

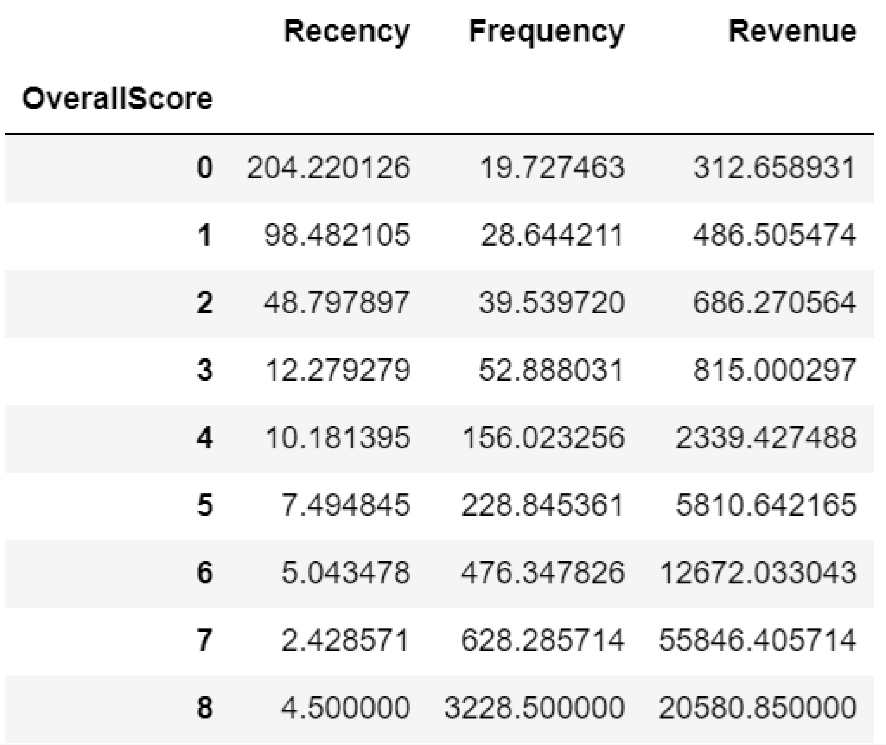


	count	mean	std	min	25%	50%	75%	max
RevenueCluster								
0	2746.0	662.344116	610.139977	-4287.63	235.285	451.185	911.9225	2753.23
1	156.0	4908.158788	2145.259258	2802.07	3305.030	4126.925	5850.5725	12239.47
2	25.0	22068.189200	7973.480069	14006.42	17510.060	19755.960	23719.4000	44563.01
3	2.0	127454.650000	10591.229216	119965.52	123710.085	127454.650	131199.2150	134943.78

Note: We see how our Revenue clusters have different characteristics.

We can say from the table, that Cluster 3 has more revenue-generating customers.

Overall Score for RFM clusters



OverallScore	Recency	Frequency	Revenue
0	204.220126	19.727463	312.658931
1	98.482105	28.644211	486.505474
2	48.797897	39.539720	686.270564
3	12.279279	52.888031	815.000297
4	10.181395	156.023256	2339.427488
5	7.494845	228.845361	5810.642165
6	5.043478	476.347826	12672.033043
7	2.428571	628.285714	55846.405714
8	4.500000	3228.500000	20580.850000

To keep things simple, better we name these scores:

0 to 2 : Low Value
3 to 4 : Mid Value
5 to 8 : High Value

The scoring above clearly shows us that customers with score 8 are our best customers whereas 0 are customers with less RFM

Tracking Customer Movements:

- Customers with high RFM scores (7 and 8) are considered ‘Loyal’.
- Customers with medium RFM scores (4,5 and 6) are considered ‘Average’.
- Customers with low RFM scores (1,2 and 3) are considered ‘Low Engagement’.

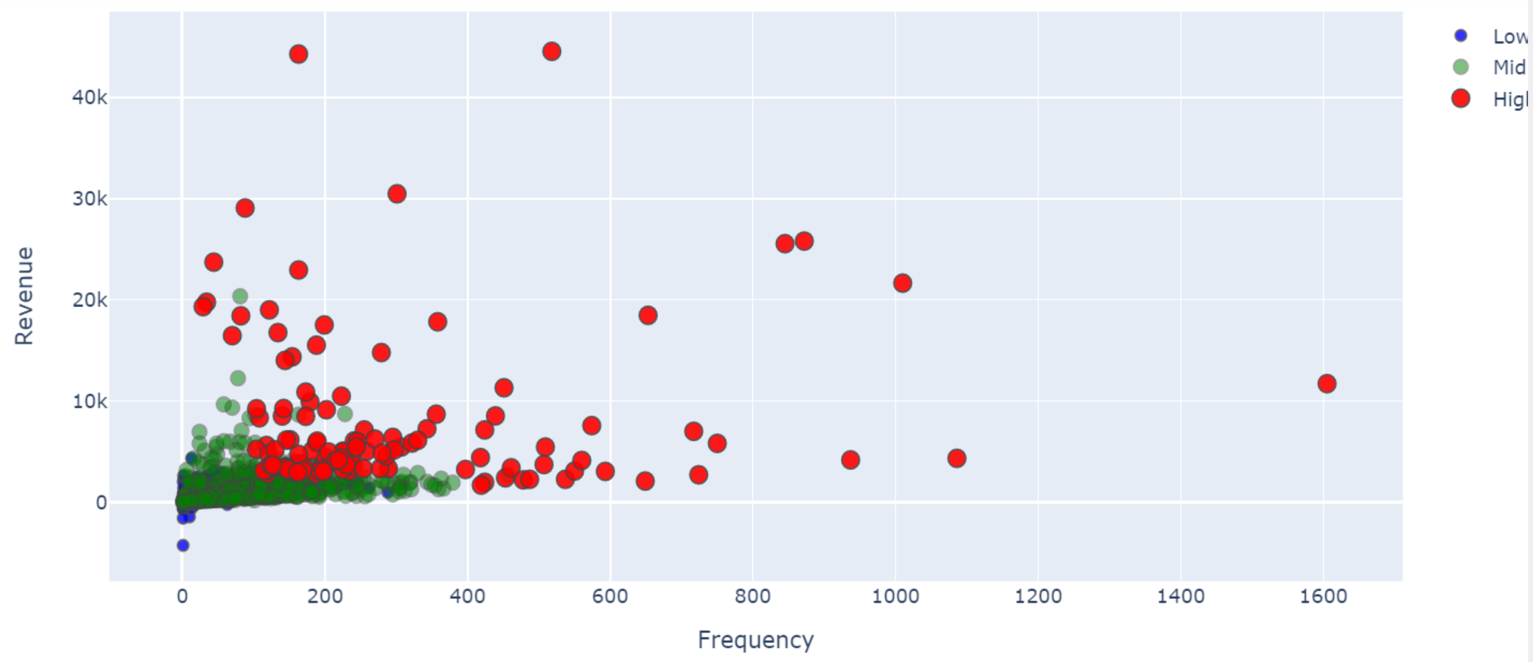
conclude
segment
loyalty
stats
process
about

conclude

K-Means Clustering in Customer Segmentation using RFM Analysis

segment
loyalty
stats
process
about

Customer Segmentation Clustering for Revenue vs Frequency

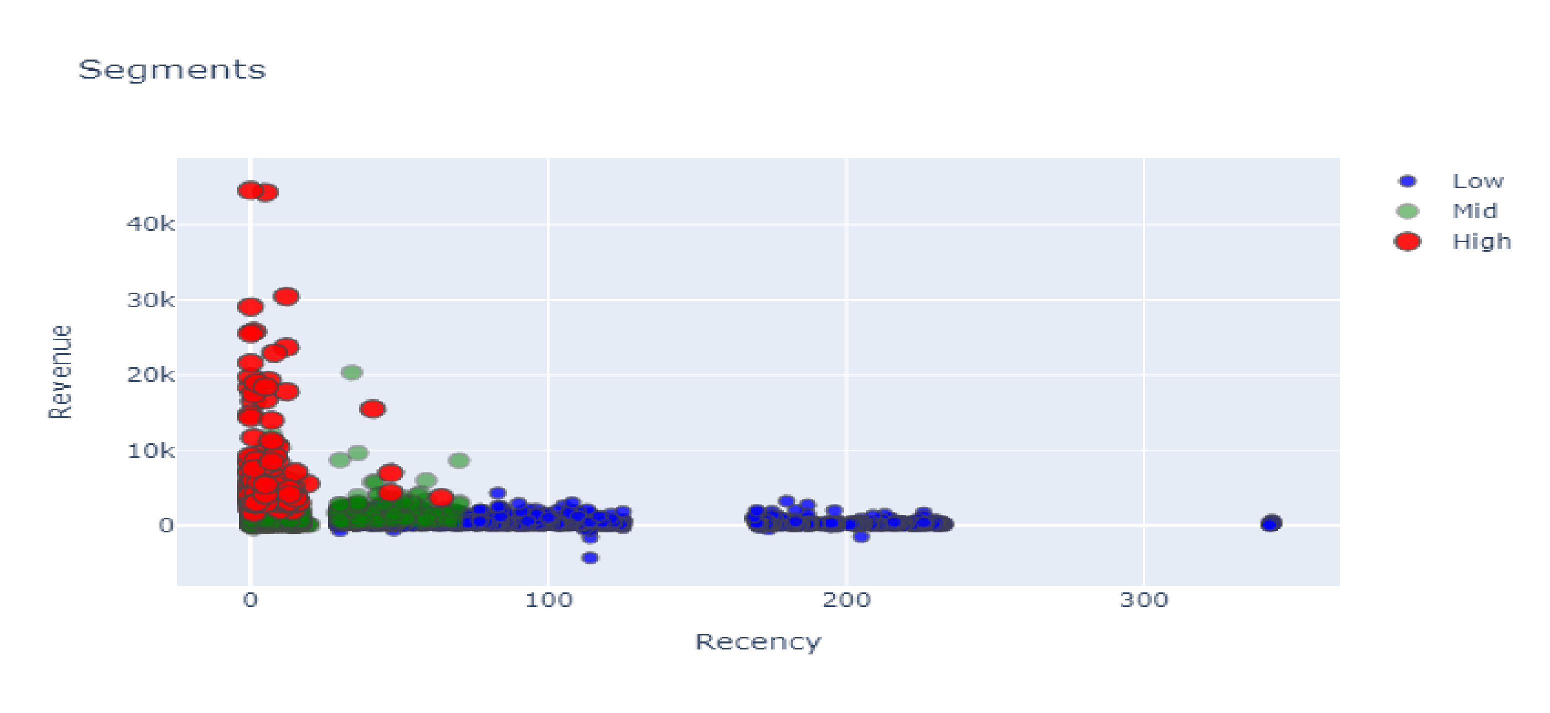


Low segment : Customers purchasing less generating less revenue negative revenue because of returning items

Mid segment : More Customers often spend less than 10K

High segment : Customers are scattered in this segment as very few customer generate revenue more than 20k

Customer Segmentation Clustering for Revenue vs Recency

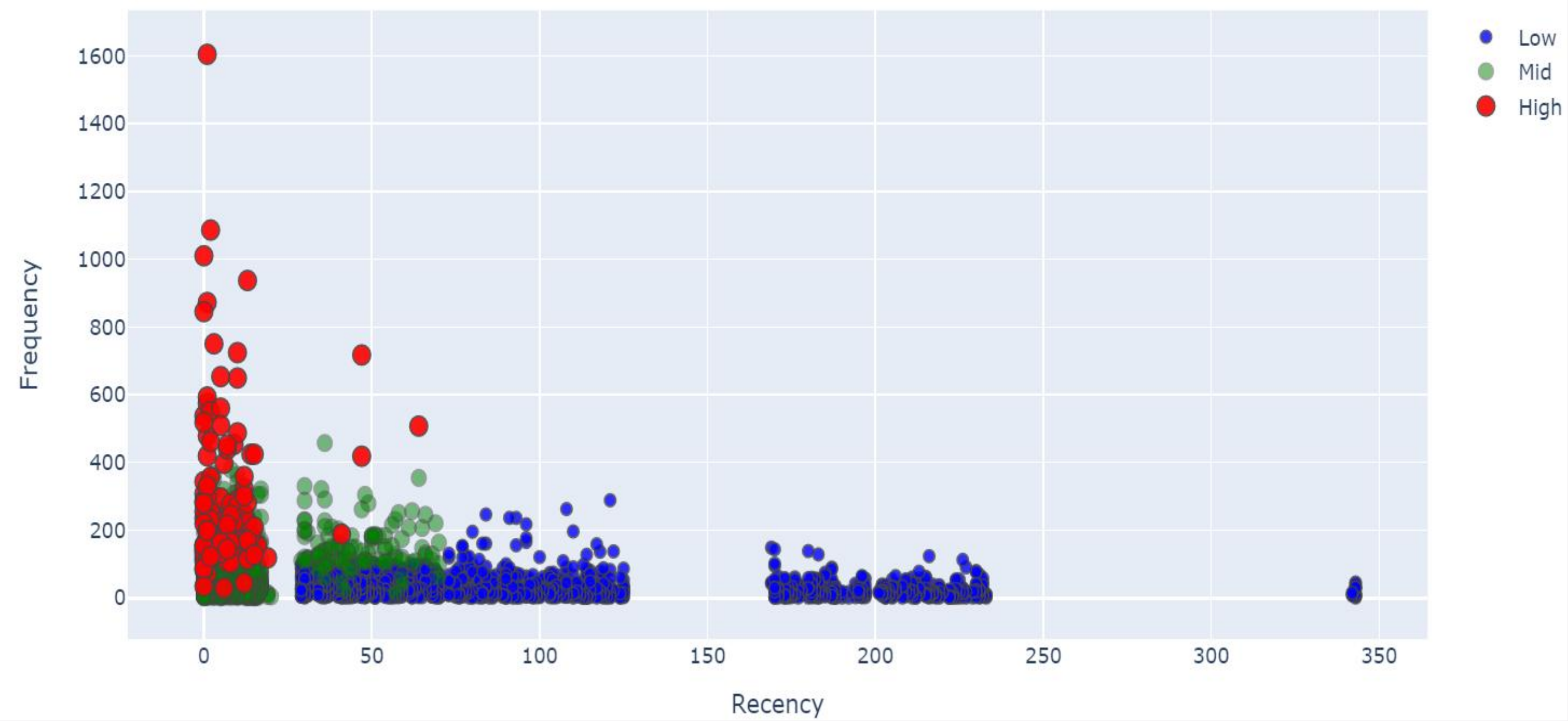


Low segment : Revenue is low, and recency is high for these customers

Mid segment : Revenue is low to moderate, and so is the recency for these customers

High segment : Revenue is high, and recency is low for these customers

Customer Segmentation Clustering for Frequency vs Recency



Low segment : Customers with more recency generate less number of transactions

Mid segment : Customers with mid recency generate moderate purchases

High segment : Customers with less recency are purchasing more items

conclude

Strategy Insights

The main strategies are quite clear:

High Value: Improve Retention

Mid Value: Improve Retention + Increase Frequency

Low Value: Increase Frequency

segment
loyalty
stats
process
about

Addressing Objectives

- **Behavioral Insights:** Uncover patterns and trends in customer behavior through advanced analytics like RFM and clustering.

- **Targeted Campaigns**: Segment customers based on behavior, demographics, and preferences to craft tailored marketing strategies.

- **Operational Optimization:** Optimize inventory, pricing, and product placement using data-driven insights, enhancing marketing efficiency.

conclude
segment
loyalty
stats
process
about

conclude

Conclusion

- Successfully developed a customer segmentation system using Databricks.
- Real-time analytics with PySpark processed large datasets efficiently.
- RFM analysis effectively segmented customers to enhance marketing and sales strategies.
- Demonstrated the transformative potential of big data tools in retail.

Lessons Learned

- Big Data Complexity: Mastery of tools like Databricks and AWS requires understanding their complexities.
- Data Quality: Preliminary data cleaning is crucial for reliable analytical outcomes.
- ETL Pipeline: Despite the challenges of not having AWS access, Understanding the integration process and associated AWS features with Databricks using PySpark is interesting and very important skill gained through this project and course.

segment
loyalty
stats
process
about

References

- Johnson, P. (2020). Effective Big Data Management and Opportunities for Implementation. Journal of Data Management, 22(4), 15-29.
- Smith, A. (2019). Using Databricks for Scalable Customer Segmentation. International Journal of Customer Relations, 18(3), 204-218.
- Databricks Community. (2018). Databricks Guide. Retrieved from https://docs.databricks.com/
- Amazon Web Services. (2021). AWS Official Documentation. Retrieved from https://aws.amazon.com/documentation/
- Retail Data Analysis. (2021). Retrieved from https://example.com/retaildata

Dataset Reference

- UCI Machine Learning Repository. (2021). Online Retail II. Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

end

segment

loyalty

stats

process

about

Thank you for listening

end

segment

loyalty

stats

process

about