Capstone Project - 4

Customer Segmentation

Name: Rudrajit Bhattacharyya

Cohort: Zanskar Pro

What’s inside?

1. Defining the problem statement
2. Defining the data
3. Data Cleaning/Pre-processing
4. Feature Engineering
5. EDA
    a. Which products are the most and least sold ones
    b. Which countries had the most and least number of customers
    c. Which day, month and time had the most number of purchases
    d. Distribution of the numerical variables
6. RFM Segmentation and Analysis
7. Building Clustering Models
8. Evaluating and Comparing all the Models
9. Conclusion

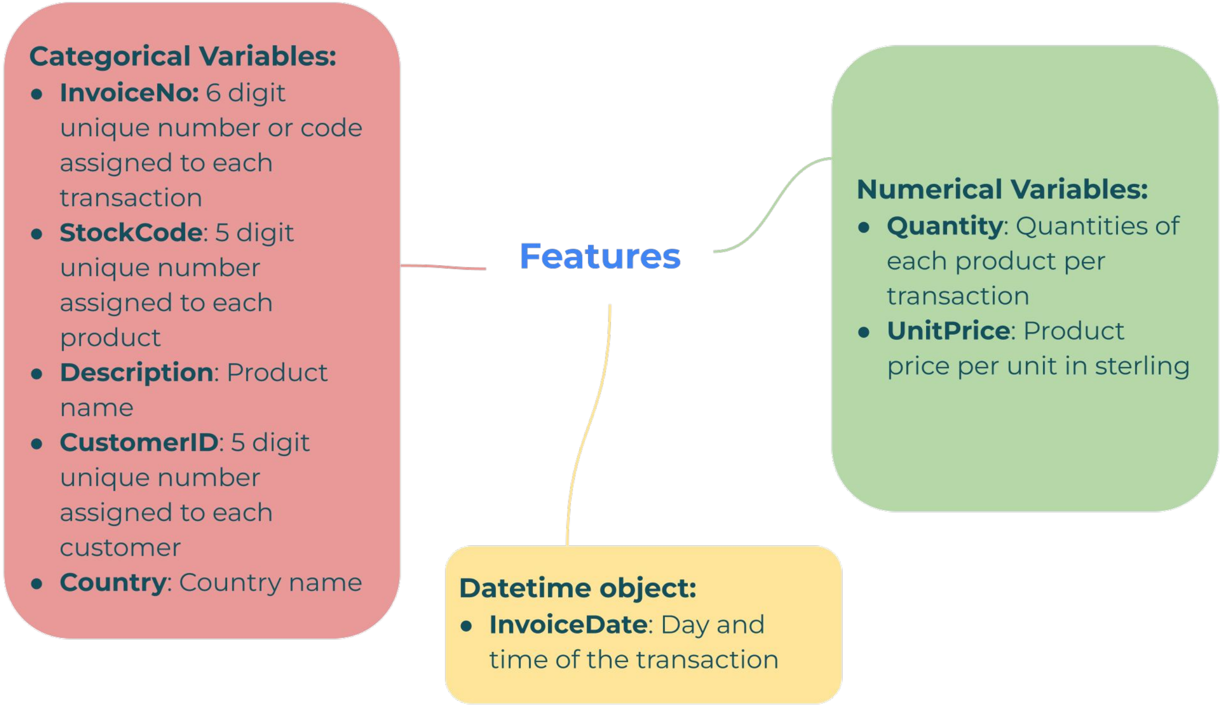
Defining the problem statement

Customer segmentation is a way to split customers into groups based on certain characteristics that those customers share. Customer segmentation will allow marketers to better tailor their marketing efforts to various audience subsets.

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

The dataset contains 541909 records of transactional data with 8 features.

Data Summary



Features

Categorical Variables:

- **InvoiceNo:** 6 digit unique number or code assigned to each transaction
- **StockCode**: 5 digit unique number assigned to each product
- **Description**: Product name
- **CustomerID**: 5 digit unique number assigned to each customer
- **Country**: Country name

Numerical Variables:

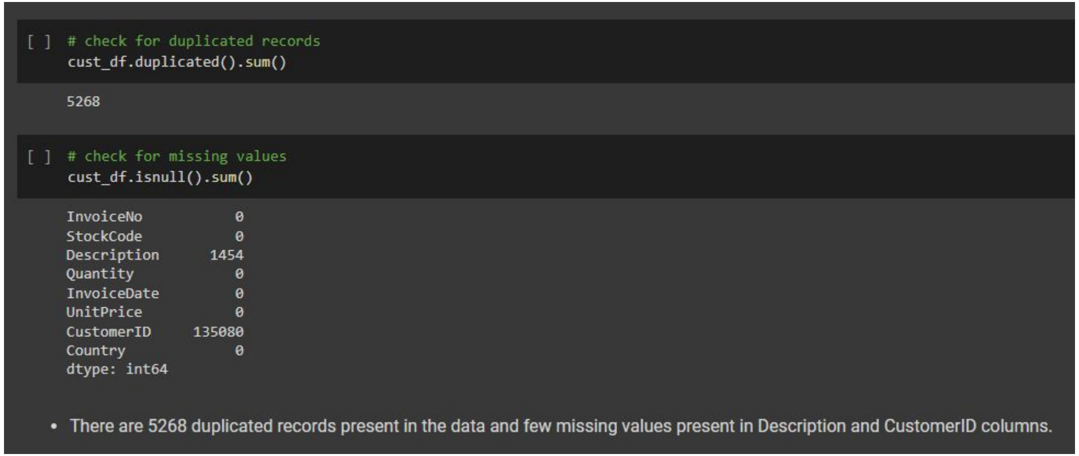
- **Quantity**: Quantities of each product per transaction
- **UnitPrice**: Product price per unit in sterling

Datetime object:

- **InvoiceDate**: Day and time of the transaction

Data Cleaning

- The dataset had 5268 duplicated records and few missing values present in the 'Description' and 'CustomerID' columns.
- These values were dropped from the dataset before proceeding further.

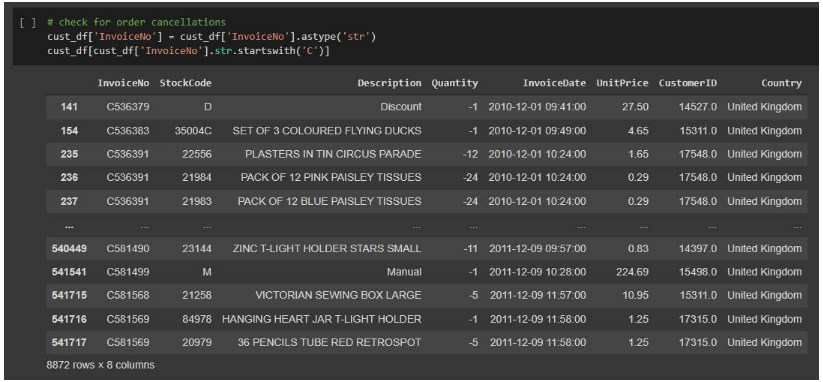


```
[ ]  # check for duplicated records
     cust_df.duplicated().sum()

     5268

[ ]  # check for missing values
     cust_df.isnull().sum()

     InvoiceNo           0
     StockCode           0
     Description      1454
     Quantity            0
     InvoiceDate         0
     UnitPrice           0
     CustomerID     135080
     Country             0
     dtype: int64
```

- There are 5268 duplicated records present in the data and few missing values present in Description and CustomerID columns.

Data Cleaning

- The dataset had 8872 records where the orders are cancelled and the quantity contains a negative value. We had to drop these values even before moving ahead.
- Interestingly there were few zero values present in ‘UnitPrice’ column which does not make sense as no store will be giving out items for free. Hence, we will consider only the records where ‘UnitPrice’ > 0.



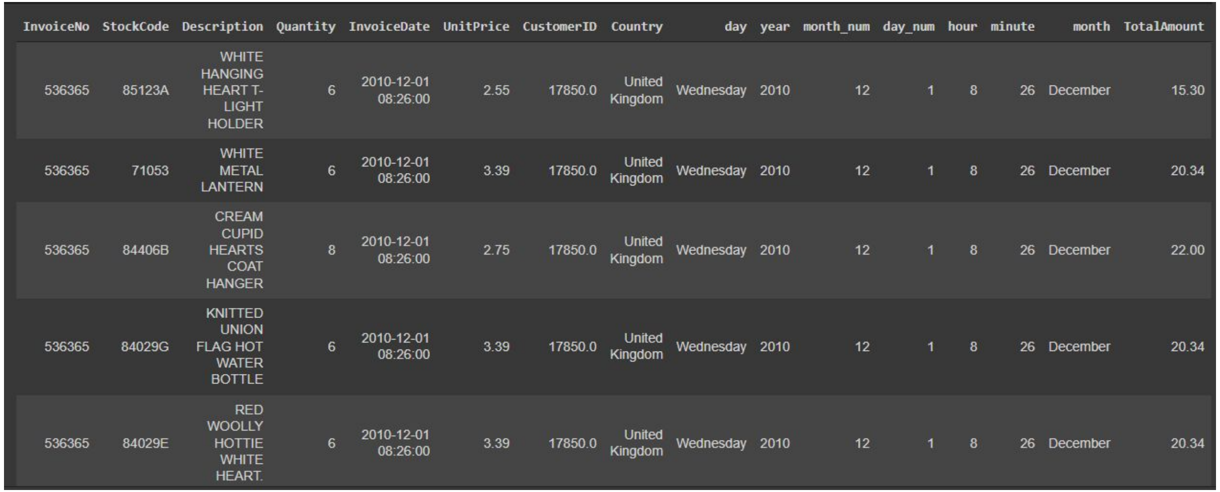
```
[ ] # check for order cancellations
    cust_df['InvoiceNo'] = cust_df['InvoiceNo'].astype('str')
    cust_df[cust_df['InvoiceNo'].str.startswith('C')]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

8872 rows × 8 columns

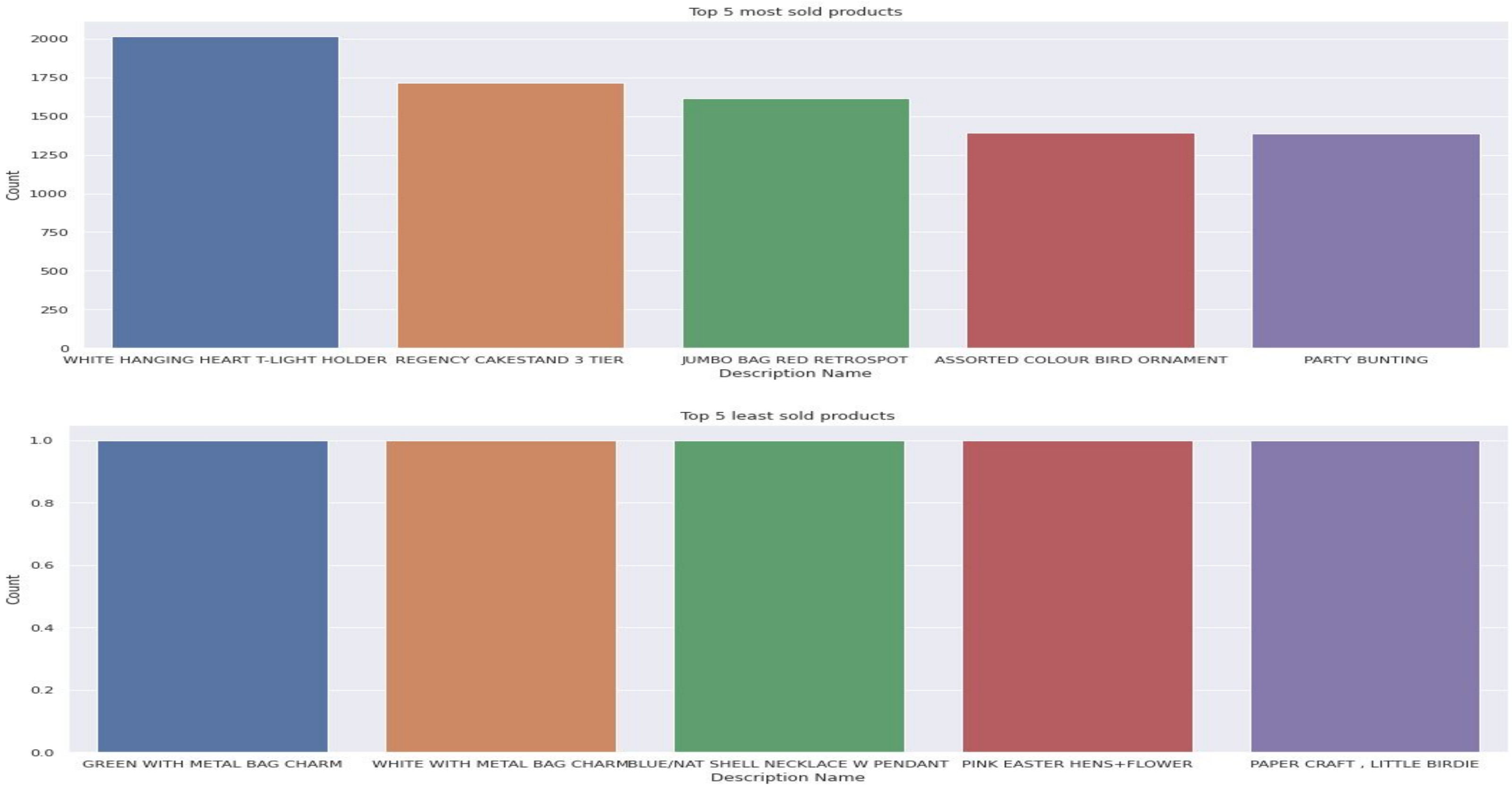
Feature Engineering

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used to make machine learning work well on new tasks, it might be necessary to design and train better features.

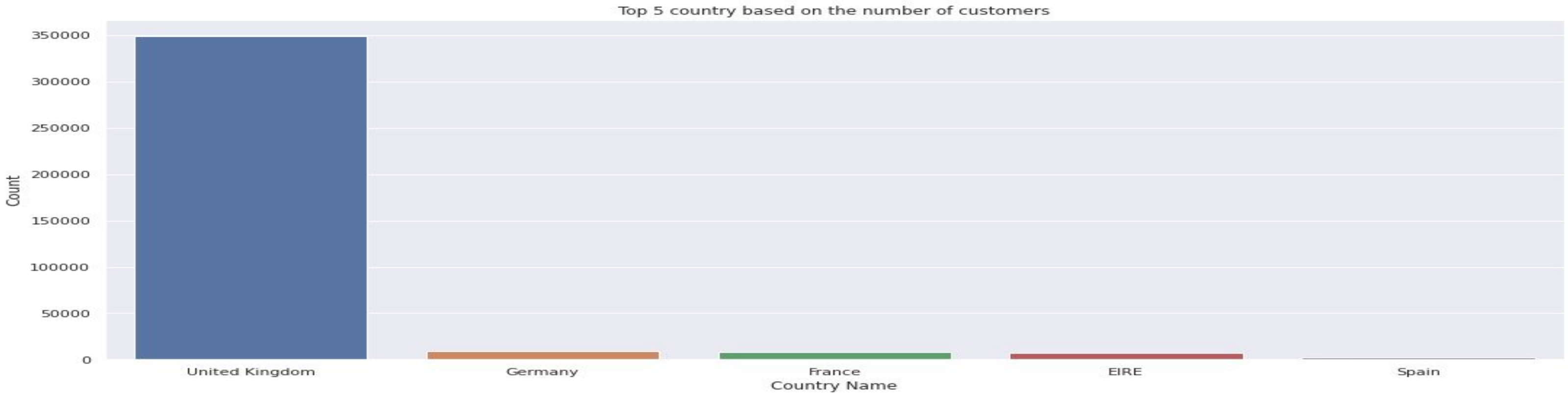


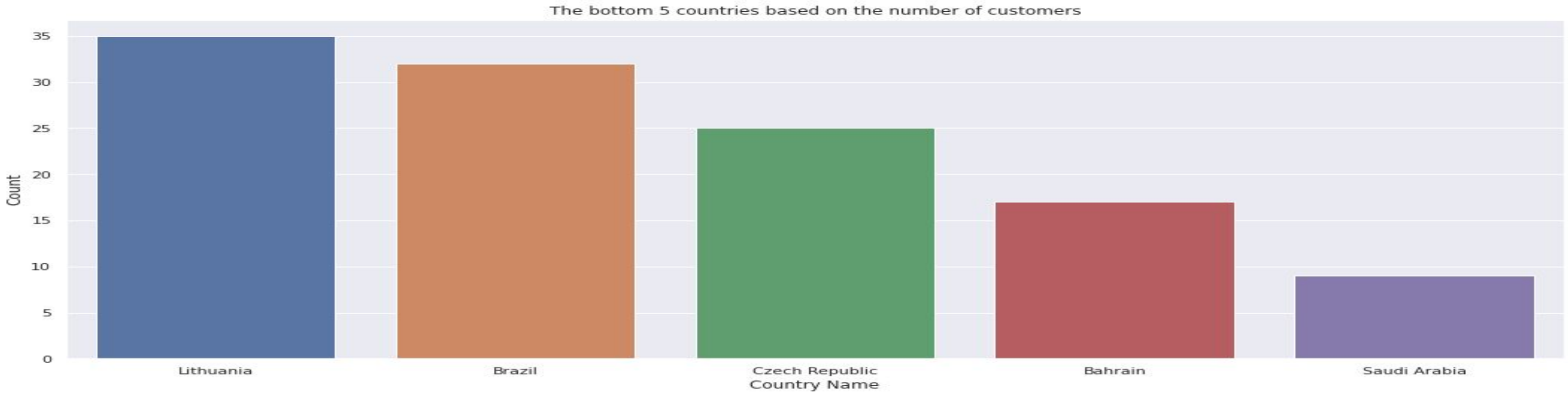
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	day	year	month_num	day_num	hour	minute	month	TotalAmount
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	15.30
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	22.00
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34

Exploratory Data Analysis

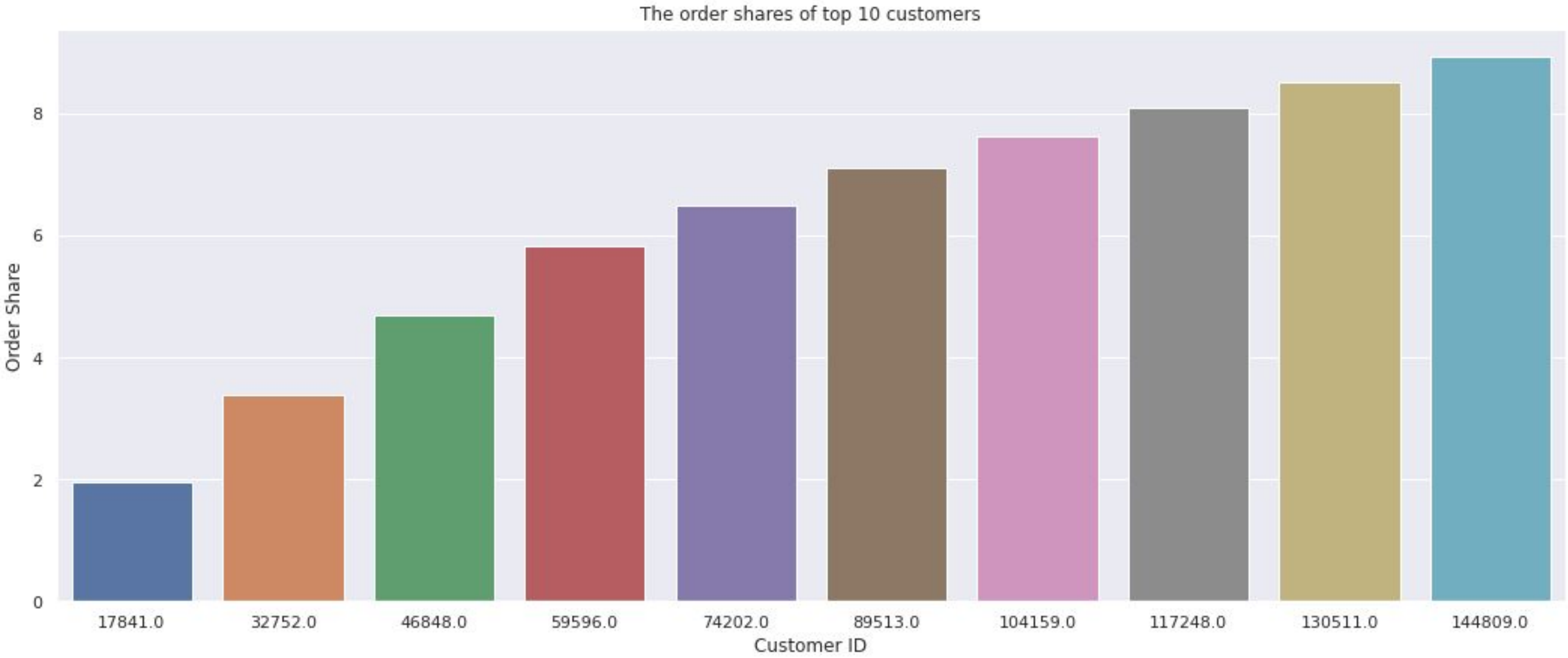


Exploratory Data Analysis

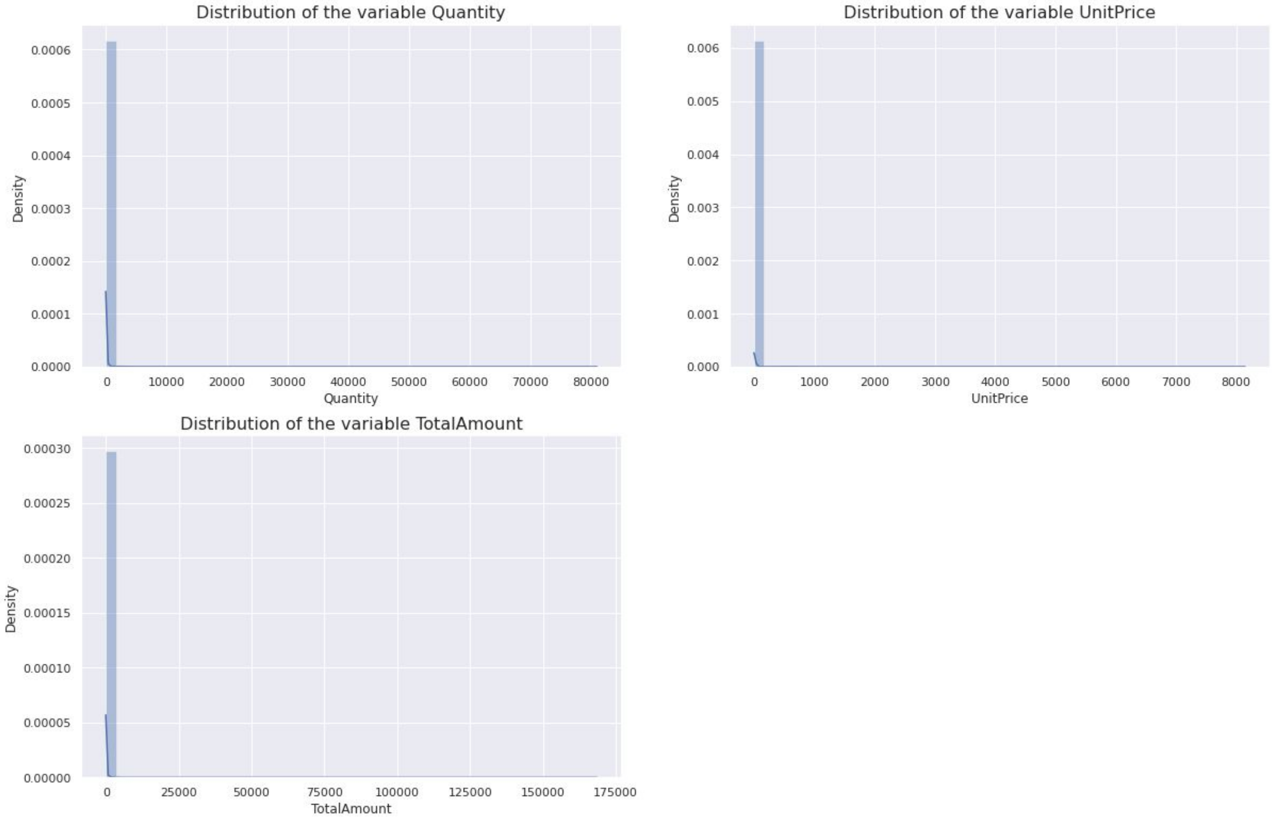




Exploratory Data Analysis



Exploratory Data Analysis

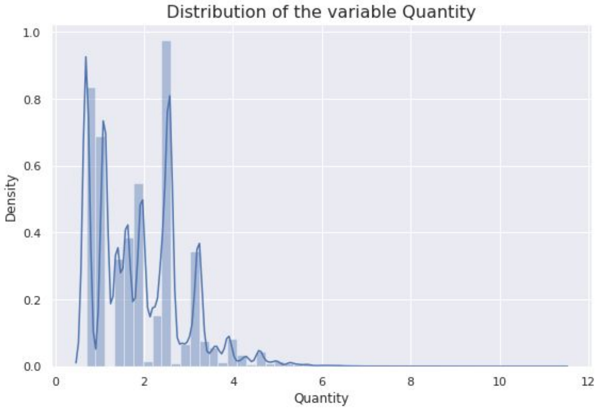


Distribution of the variable Quantity

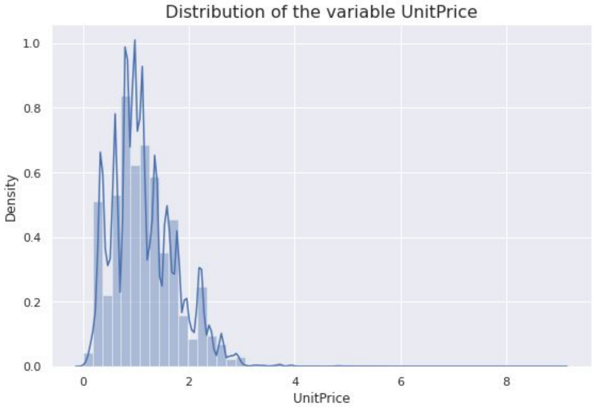
Distribution of the variable UnitPrice

Distribution of the variable TotalAmount

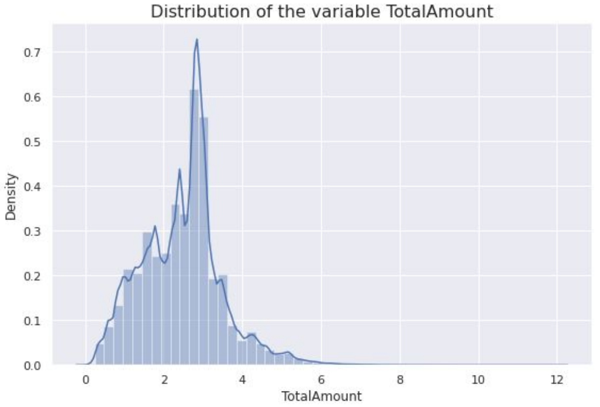
Exploratory Data Analysis



Distribution of the variable Quantity

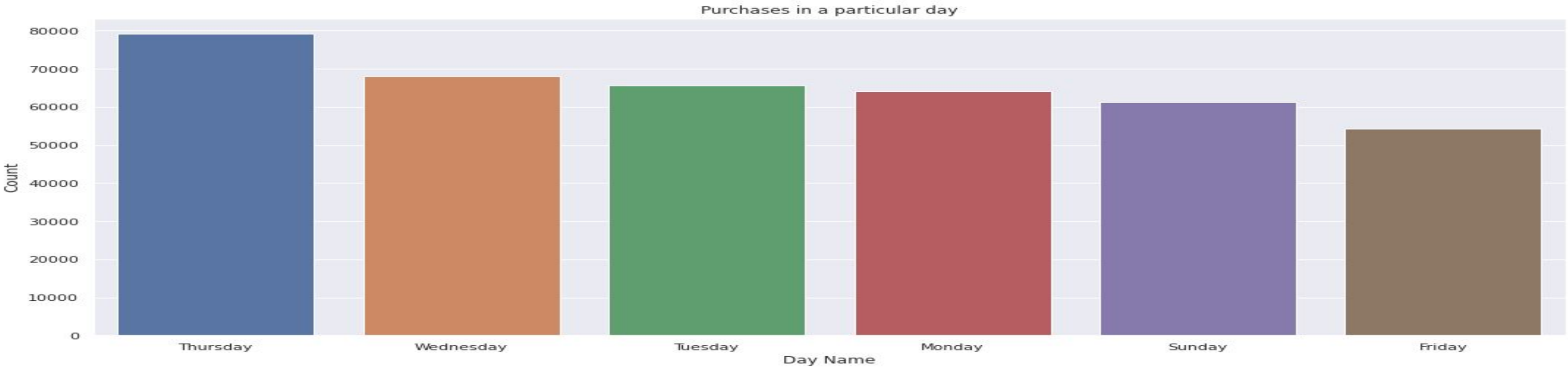


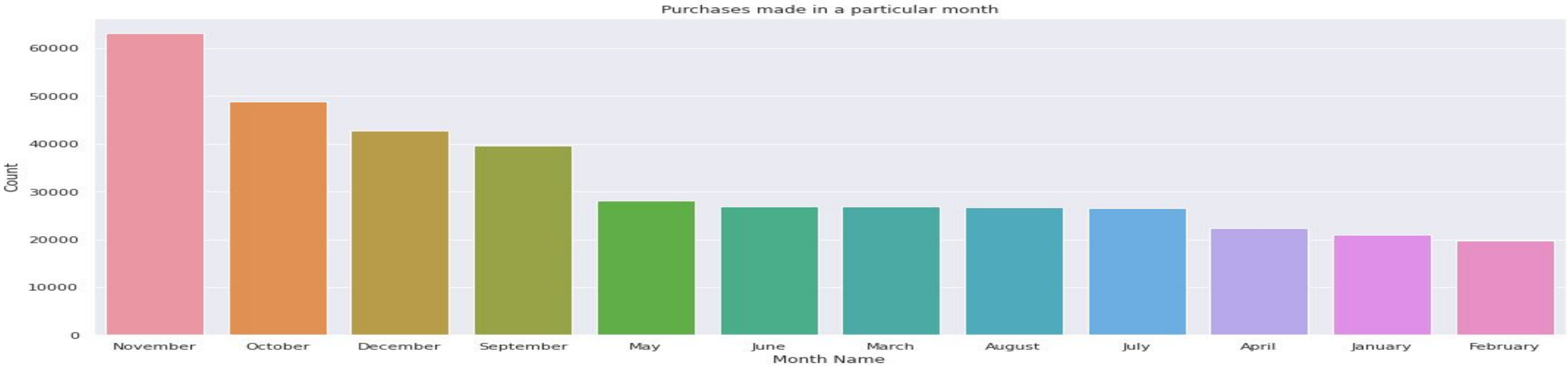
Distribution of the variable UnitPrice



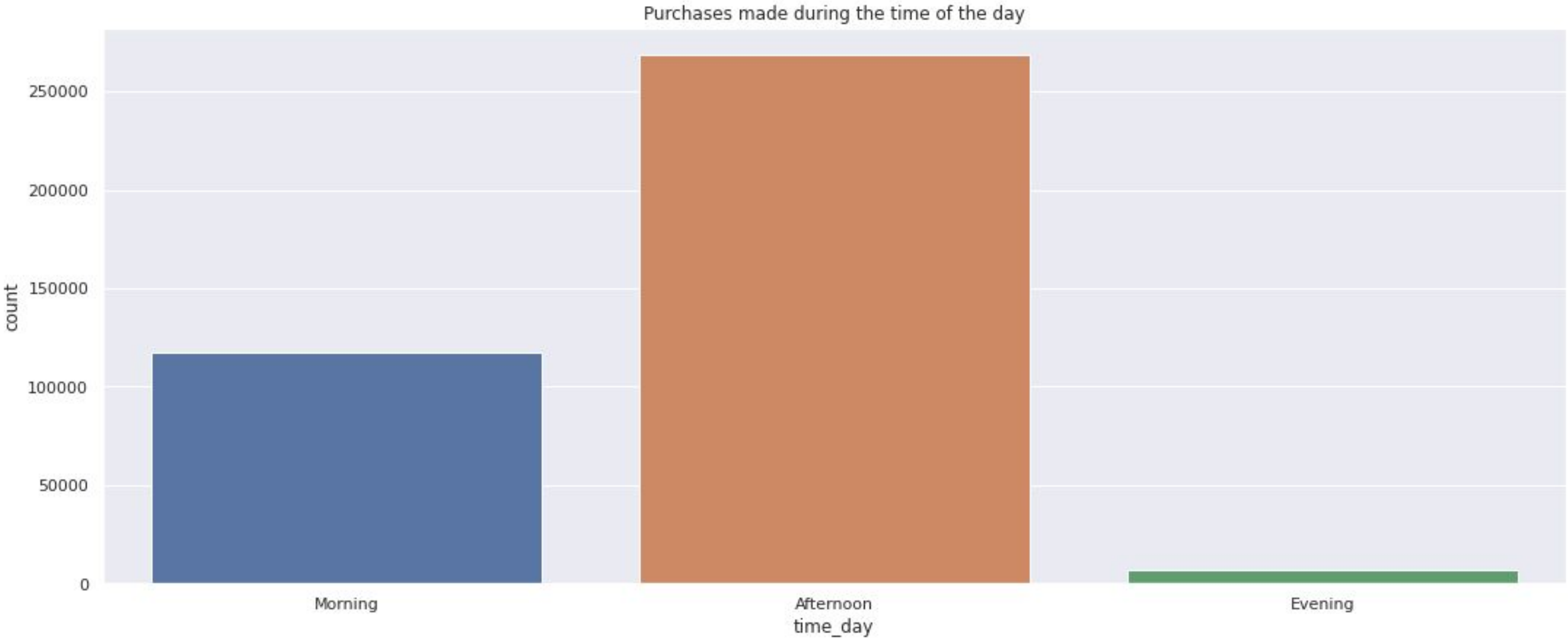
Distribution of the variable TotalAmount

Exploratory Data Analysis





Exploratory Data Analysis

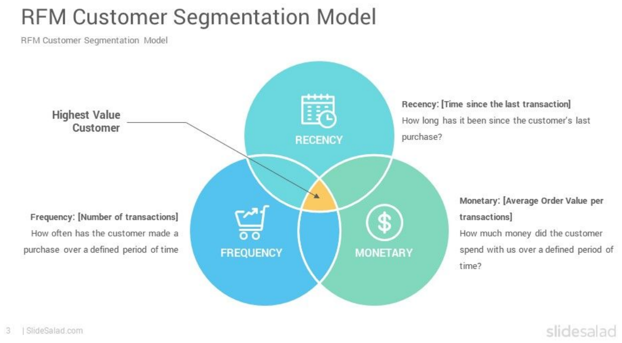


RFM Segmentation

Recency, frequency, monetary value is a marketing analysis tool used to identify a company’s or an organization’s best customers by measuring and analyzing spending habits.

The RFM model is based on three quantitative factors:

- Recency: How recently a customer has made a purchase
- Frequency: How often a customer makes a purchase
- Monetary: How much money a customer spends on purchases



RFM Customer Segmentation Model

RFM Customer Segmentation Model

Recency: [Time since the last transaction]
How long has it been since the customer’s last purchase?

Frequency: [Number of transactions]
How often has the customer made a purchase over a defined period of time

Monetary: [Average Order Value per transactions]
How much money did the customer spend with us over a defined period of time?

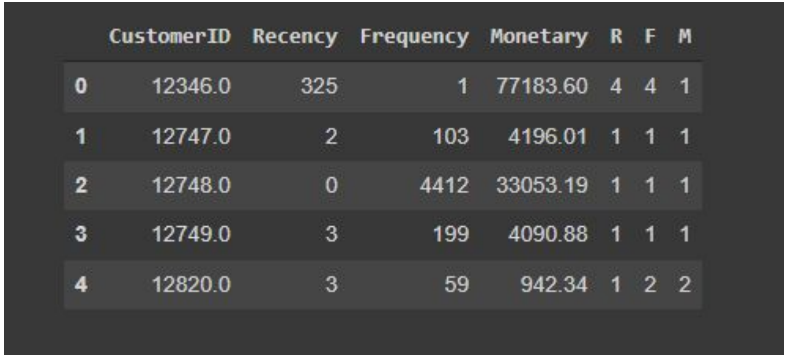
3 | SlideSalad.com

slidesalad

RFM Analysis

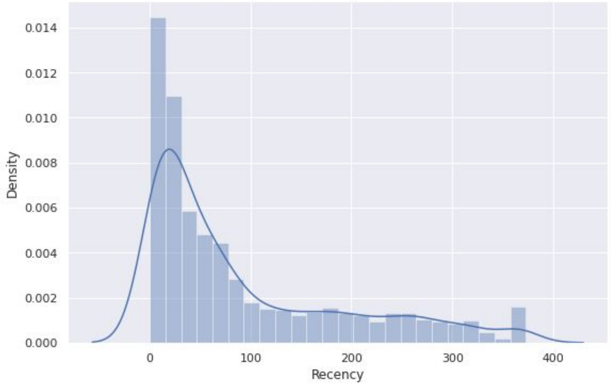
RFM analysis numerically ranks a customer in each of these three categories, generally on a scale of 1 to 5. The best customer would receive a top score in every category.

RFM analysis allows a comparison between potential customers or clients. It gives organizations a sense of how much revenue comes from repeat customers vs new customers, and which levers they can pull to try to make customers happier so they can become repeat customers.

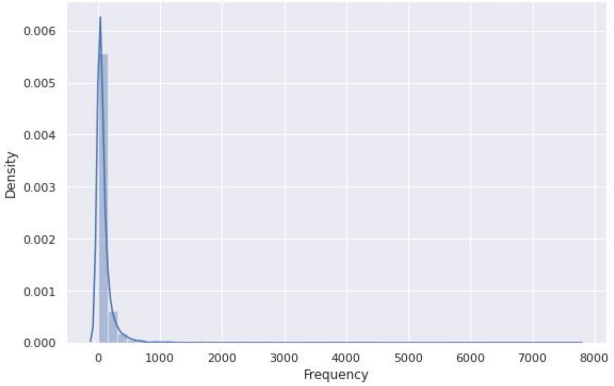


	CustomerID	Recency	Frequency	Monetary	R	F	M
0	12346.0	325	1	77183.60	4	4	1
1	12747.0	2	103	4196.01	1	1	1
2	12748.0	0	4412	33053.19	1	1	1
3	12749.0	3	199	4090.88	1	1	1
4	12820.0	3	59	942.34	1	2	2

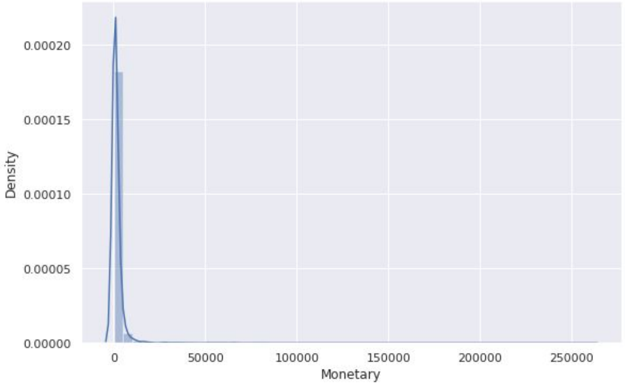
RFM Distribution



Distribution of the variable Recency

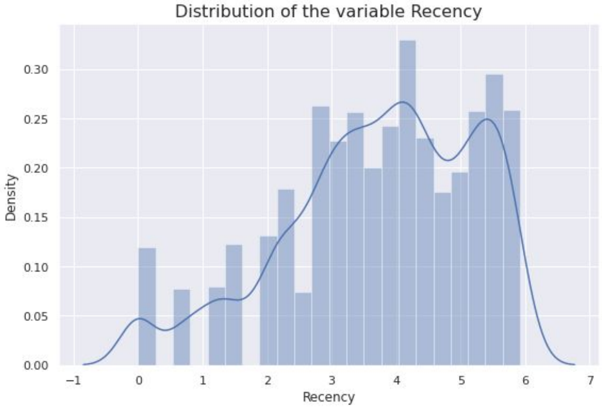


Distribution of the variable Frequency

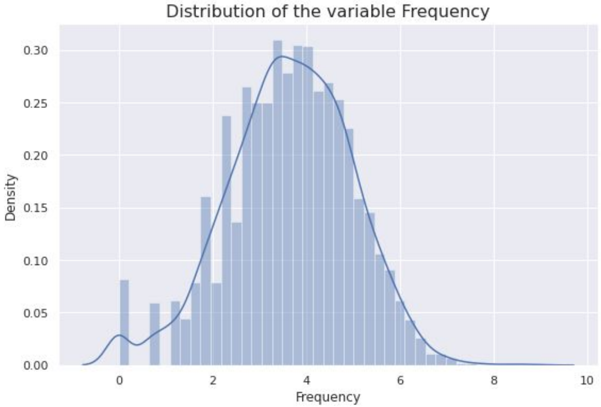


Distribution of the variable Monetary

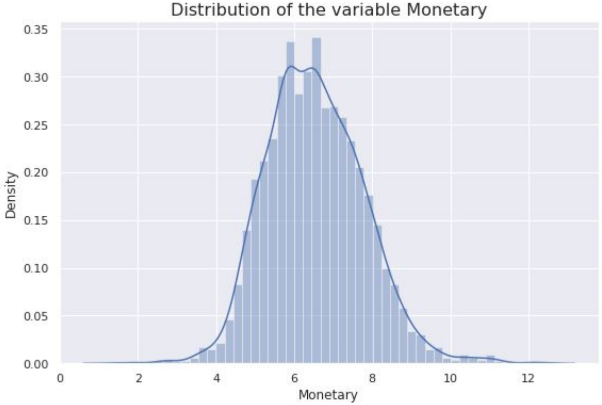
RFM Distribution



Distribution of the variable Recency



Distribution of the variable Frequency

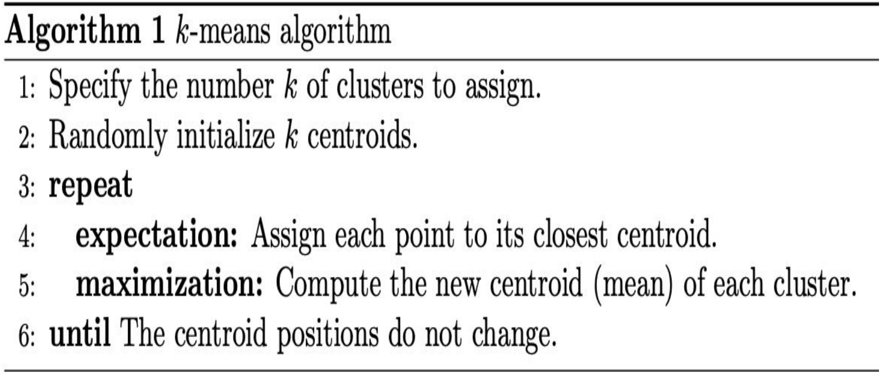


Distribution of the variable Monetary

K-Means Clustering

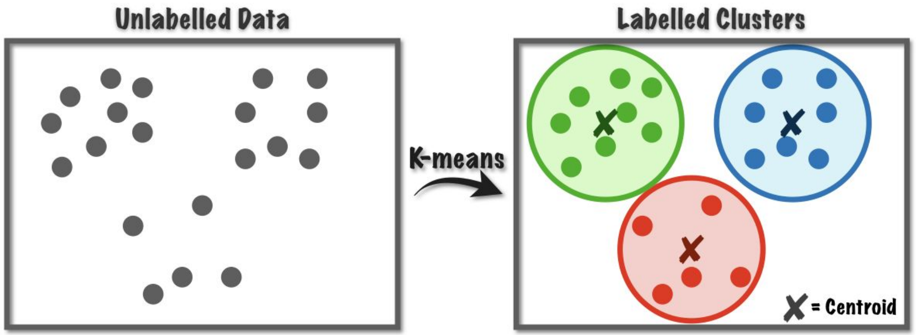
The K-Means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters are there. In this method, the data points are assigned in such a way that the sum of the squared distances between the data points and the centroid is as small as possible.

K-Means implements the Expectation-Maximization strategy to solve the problem.



Algorithm 1 k-means algorithm

```
Specify the number $k$ of clusters to assign.
Randomly initialize $k$ centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.
```

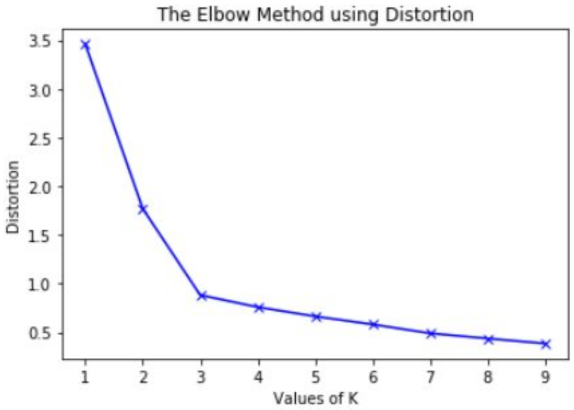


Evaluation Metrics

Silhouette Score: Silhouette score is used to evaluate the quality of clusters created using clustering algorithms in terms of how well samples are clustered with other samples that are similar to each other. The silhouette score is calculated for each sample of different clusters.

$$Silhouette - score = \frac{b_i - a_i}{\max\left(bi, a_i\right)}$$

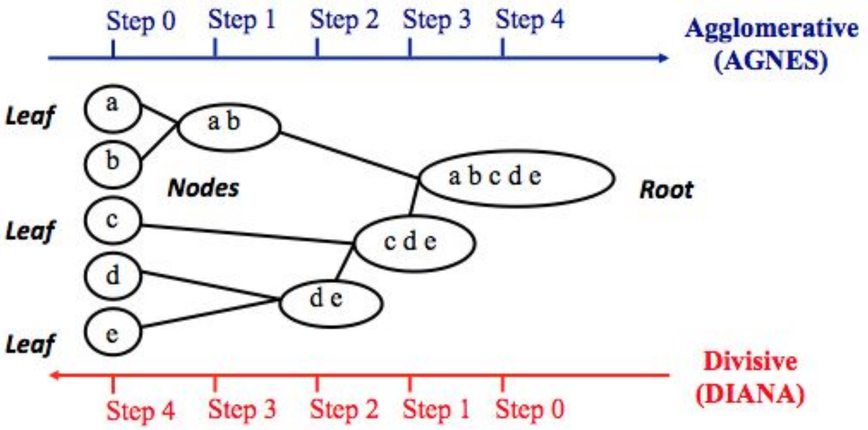
Elbow Method: A commonly used method for finding optimal K value is the Elbow method. For different values of K, the elbow method calculates WCSS which is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an elbow. The point at which the graph changes rapidly is the optimal K value.

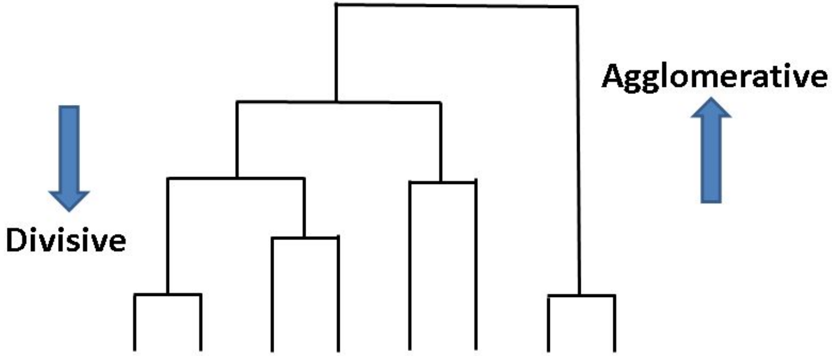


Hierarchical Clustering

Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from other clusters, and the objects within each cluster are broadly similar to each other.

The main output of hierarchical clustering is a dendrogram, which shows the hierarchical relationship between the clusters.

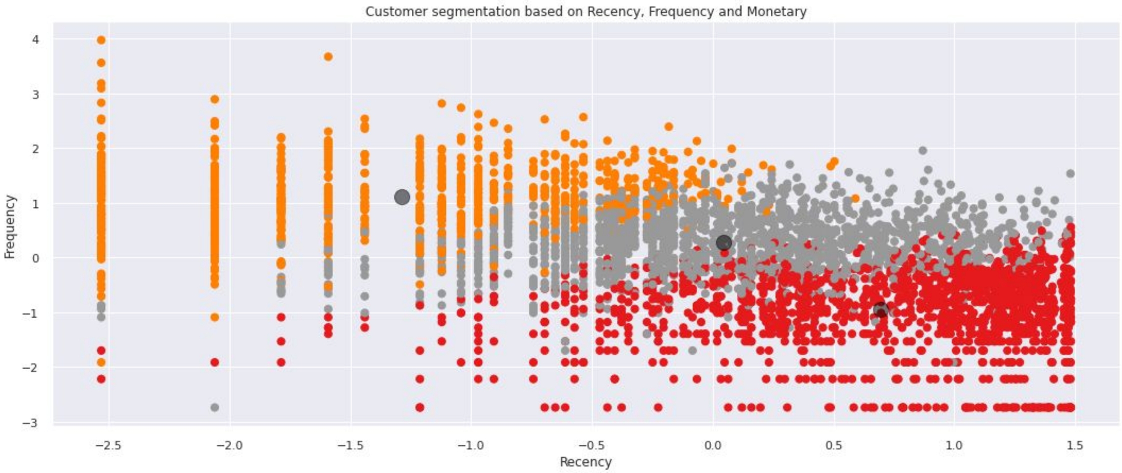




Final Results

The final clustering derived from the K-Means model applied on RFM data is as follows:

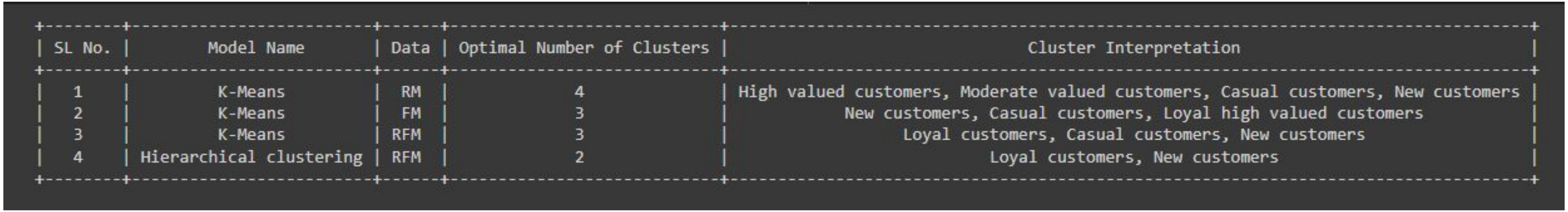
- **Loyal Customers (Orange Points)**: These are the most frequent buyers with significant monetary contribution.
- **Casual Customers (Grey Points):** Not so frequent or recent buyers but contribute moderately in terms of money.
- **New Customers (Red Points)**: Very recent buyers who could be the new customers.



Summary Table

Additional to the K-Means model implemented on RFM data, we have used K-Means separately on Recency, Monetary and Frequency, Monetary along with hierarchical clustering on RFM data to compare the clustering performance of each method.

We can observe from the results that the optimal number of clusters for every method is around 2 or 3. Thus we can consider our K-Means model applied on RFM data as our final model.



```
+--------+-------------------------+------+---------------------------+--------------------------------------------------------------------------------+
| SL No. |        Model Name       | Data | Optimal Number of Clusters |                             Cluster Interpretation                             |
+--------+-------------------------+------+---------------------------+--------------------------------------------------------------------------------+
|   1    |         K-Means         |  RM  |             4             | High valued customers, Moderate valued customers, Casual customers, New customers |
|   2    |         K-Means         |  FM  |             3             |          New customers, Casual customers, Loyal high valued customers          |
|   3    |         K-Means         | RFM  |             3             |                Loyal customers, Casual customers, New customers                |
|   4    | Hierarchical clustering | RFM  |             2             |                          Loyal customers, New customers                          |
+--------+-------------------------+------+---------------------------+--------------------------------------------------------------------------------+
```

Conclusion

Let us summarize the insights we have discovered or derived from our customer segmentation task:

- The dataset had a lot of duplicated records and missing values
- UK had the most number of customers
- Only 10 customers account for 9% of the orders
- Most purchases were recorded during the festive months, Oct to Dec
- RFM method was used to segment customers and it is based on 3 quantitative factors, Recency, Frequency and Monetary
- RFM analysis gives us a sense of how much revenue comes from repeat customers vs new customers
- K-Means clustering was used to identify different customer segments
- We used the silhouette score and elbow method to determine the optimal number of clusters which was found out to be 3.
- The final model resulted in 3 clusters i.e, Loyal Customers, Casual Customers, New Customers

Thank You !