CUSTOMER SEGMENTATION: ONLINE-RETAIL

Group 9

(Meghann Sandhu, Aleena Varghese, Priyadarshini Venkatesh)

University Of SanDiego, California

AAI-500

Dr. Rakesh Das

24-June-2024

DECLARATION

I declare that this report, entitled " CUSTOMER SEGMENTATION: ONLINE-RETAIL" is the result of my own work and that all sources I have used or quoted have been indicated and acknowledged by means of complete references.

ABSTRACT

This report presents an in-depth analysis of customer segmentation for an online retail dataset using the Recency, Frequency, and Monetary (RFM) model. By applying the RFM model, we segmented customers into five distinct clusters and provided tailored marketing strategies for each segment. The analysis includes exploratory data analysis, data cleaning, feature engineering, model selection, and a detailed model analysis. The insights gained from this segmentation can help in devising effective marketing strategies, improving customer engagement, and increasing overall revenue.

INTRODUCTION

Customer segmentation is a key strategy for understanding the diverse needs and behaviors of a customer base in online retail. This process allows businesses to identify different types of customers, understand their buying patterns, and create tailored marketing strategies to enhance engagement and sales. In this report, we explore the segmentation of an online retail dataset using the RFM model, which considers the recency, frequency, and monetary value of customer purchases.

ABOUT THE DATASET

ONLINE RETAIL

The Online Retail dataset available on the UCI Machine Learning Repository was compiled by Daqing Chen in 2015. It contains transactional data from a UK-based online retail store. Here are some key details about this dataset.

SOURCE

The dataset is available on the UCI Machine Learning Repository at this link: Online Retail Dataset.

DESCRIPTION

The dataset includes information about individual transactions made by customers. Each row in the dataset represents a single transaction, and the columns provide details about the items purchased, the quantities, prices, customer IDs, invoice numbers, and other relevant information.

FEATURES:

- InvoiceNo: A unique identifier for each transaction.
- StockCode: A unique identifier for each product.
- Description: Description of the product.
- Quantity: The quantity of each product purchased in a transaction.
- InvoiceDate: The date and time when the transaction was generated.
- UnitPrice: The price per unit of each product in sterling.
- CustomerID: A unique identifier for each customer.
- Country: The country where each customer resides.
- Purpose: This dataset is commonly used for retail analytics, customer segmentation, market basket analysis, and other data analysis tasks related to retail and e-commerce.

The Online Retail dataset is valuable for studying customer behaviour, sales patterns, and other aspects of online retail operations.

DATA CLEANING

Data cleaning is a crucial step to ensure the quality and accuracy of the analysis. The data cleaning process involved the following steps:

- **Removing Duplicates**: Identifying and removing duplicate entries to ensure each transaction is unique.

- **Handling Missing Values**: Dealing with missing values by either filling them with appropriate values or removing the entries if they are not significant.
- **Correcting Data Types**: Ensuring that all data types are consistent and appropriate for analysis.
- **Removing Outliers**: Identifying and removing outliers that could skew the results. This included filtering out transactions with negative values or extremely high values that do not represent typical customer behaviour.
- **Validating Data Integrity**: Checking the consistency and validity of the data, such as ensuring that all transaction dates are within a reasonable range and all monetary values are positive.

	nulls	missing_ratio
CustomerID	135080	24.926694
Description	1454	0.268311
InvoiceNo	0	0.000000
StockCode	0	0.000000
Quantity	0	0.000000
InvoiceDate	0	0.000000
UnitPrice	0	0.000000
Country	0	0.000000

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:27:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

Fig1(a,b): dataset before cleaning

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	397884.000000	397884	397884.000000	397884.000000
mean	12.988238	2011-07-10 23:41:23.511023360	3.116488	15294.423453
min	1.000000	2010-12-01 08:26:00	0.001000	12346.000000
25%	2.000000	2011-04-07 11:12:00	1.250000	13969.000000
50%	6.000000	2011-07-31 14:39:00	1.950000	15159.000000
75%	12.000000	2011-10-20 14:33:00	3.750000	16795.000000
max	80995.000000	2011-12-09 12:50:00	8142.750000	18287.000000
std	179.331775	NaN	22.097877	1713.141560

Fig2: dataset after cleaning

EXPLORATORY DATA ANALYSIS

The first step in our analysis involved exploring the dataset to understand its structure and identify any patterns or anomalies. Key areas of focus included:

- **Market Representation and Country**: The first step in our analysis was understanding the market representation and distribution of customers across different countries. This step helps in identifying the geographical diversity and potential market segments based on location.

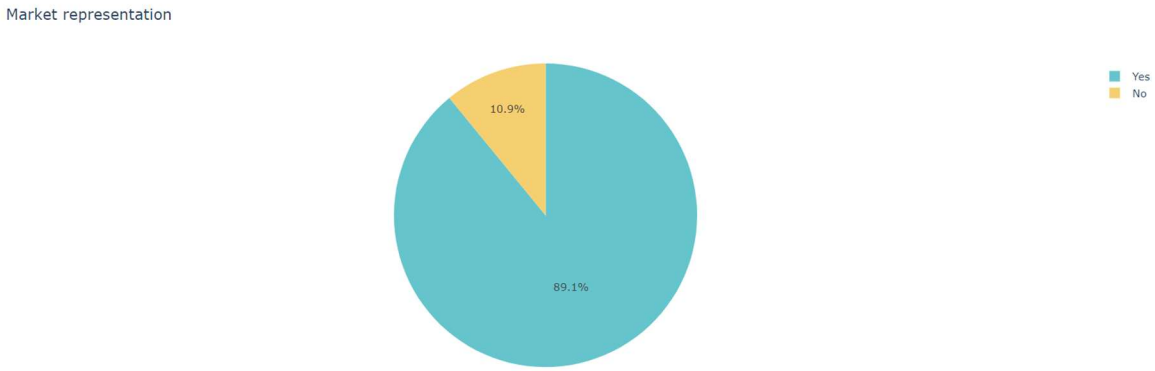

Fig3: market representation by country(Yes: Customer is from UK)

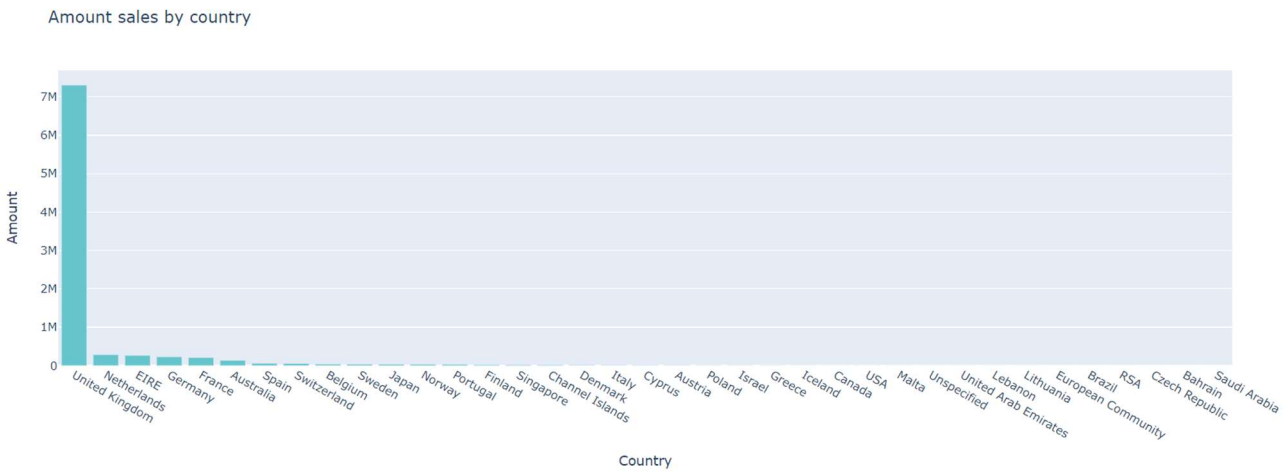

Fig4: amount sales by country

- **Top Customers and Products**: Identifying top customers and products provides insights into who the best customers are and which products are performing well. This information is crucial for inventory management, promotional strategies, and personalized marketing.

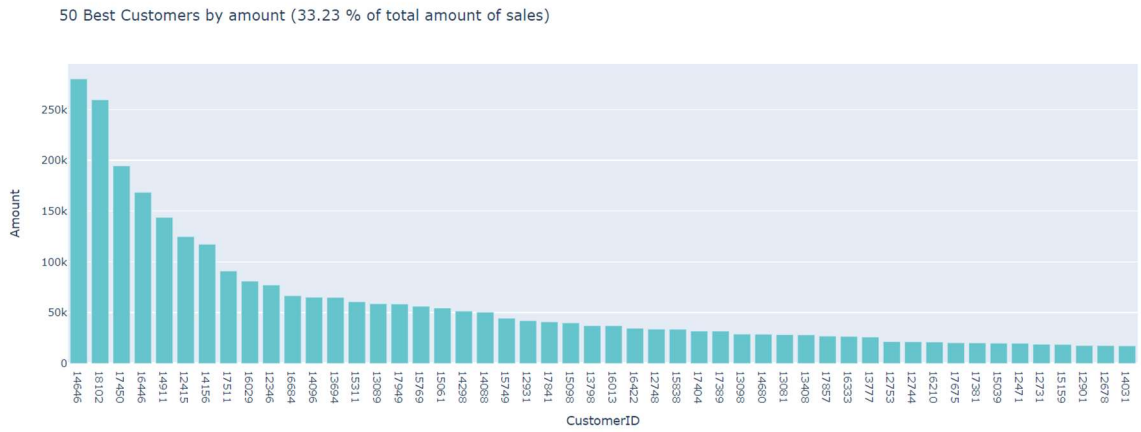

Fig5: 50 best customers by amount of sales

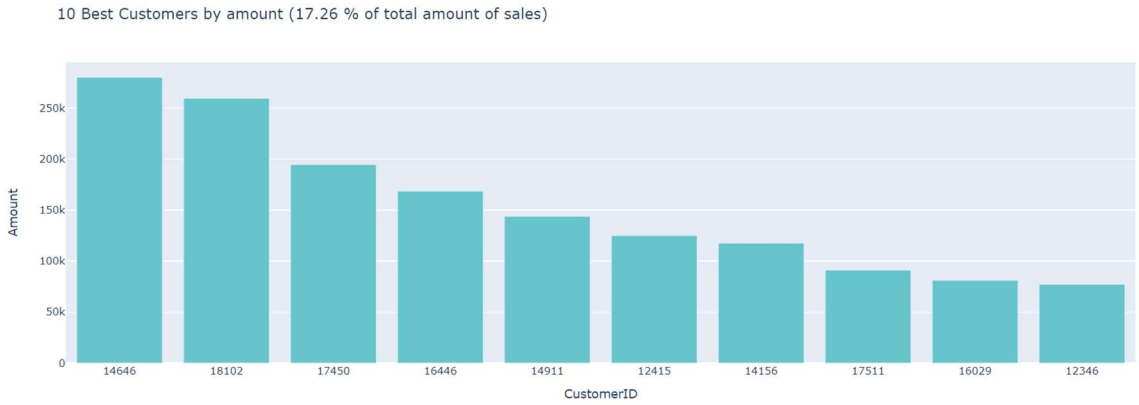

Fig6: 10 best customers by amount of sales

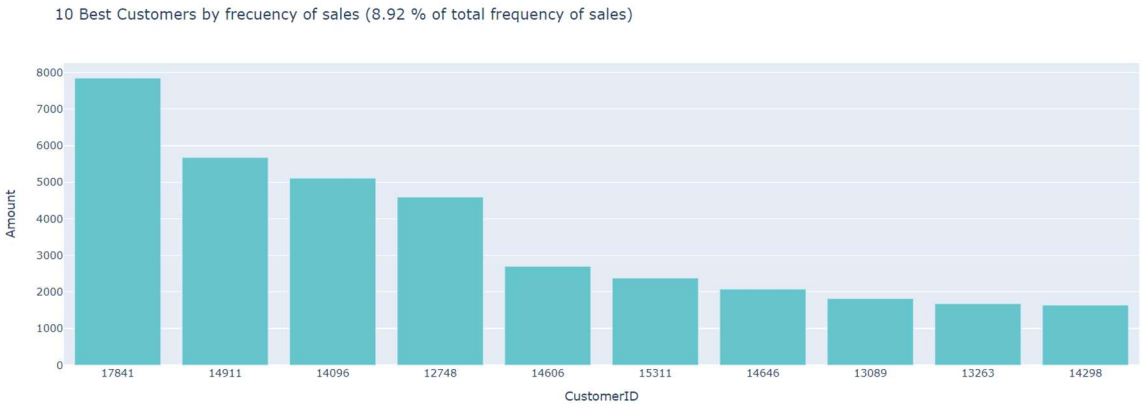

Fig7: 10 best customers by frequency of sales

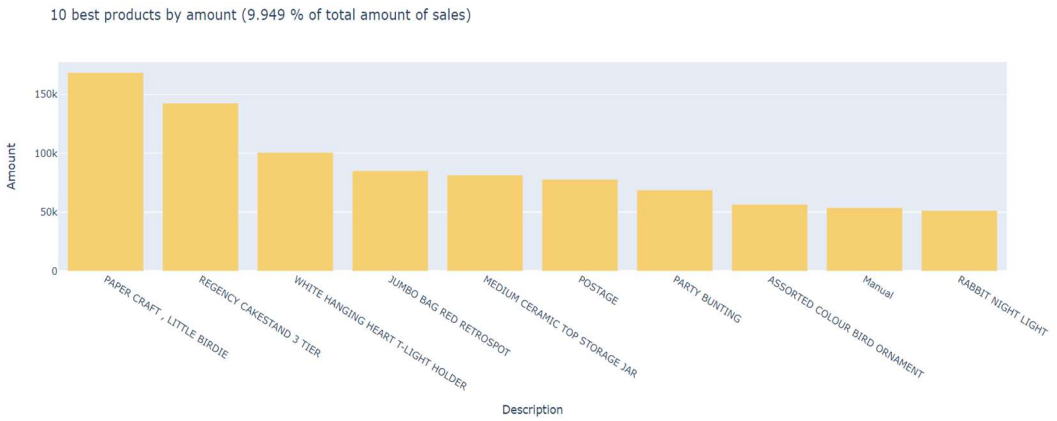

Fig8: 10 best products by amount of sales

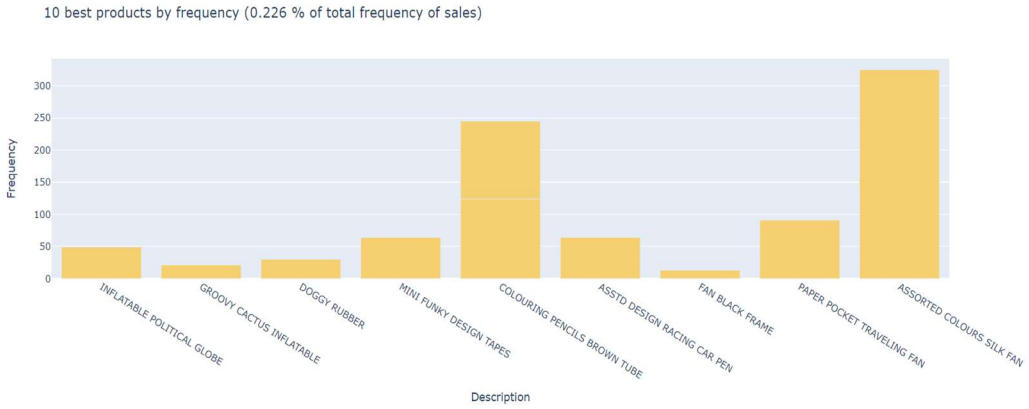

Fig9: 10 best products by frequency

FEATURE ENGINEERING

To prepare the data for the RFM model, we performed feature engineering, which involved:

- **Calculating Recency**: Recency was calculated by determining the number of days since the customer's last purchase.
- **Calculating Frequency**: Frequency was determined by counting the number of transactions made by each customer within the analysis period.
- **Calculating Monetary Value**: Monetary value was calculated by summing the total amount spent by each customer.

- **Normalizing the Data**: To address skewness in the data, we applied a log transformation, which helps in stabilizing the variance and making the data more suitable for clustering.

The data was skewed, and a log transformation was applied to normalize it for better analysis.

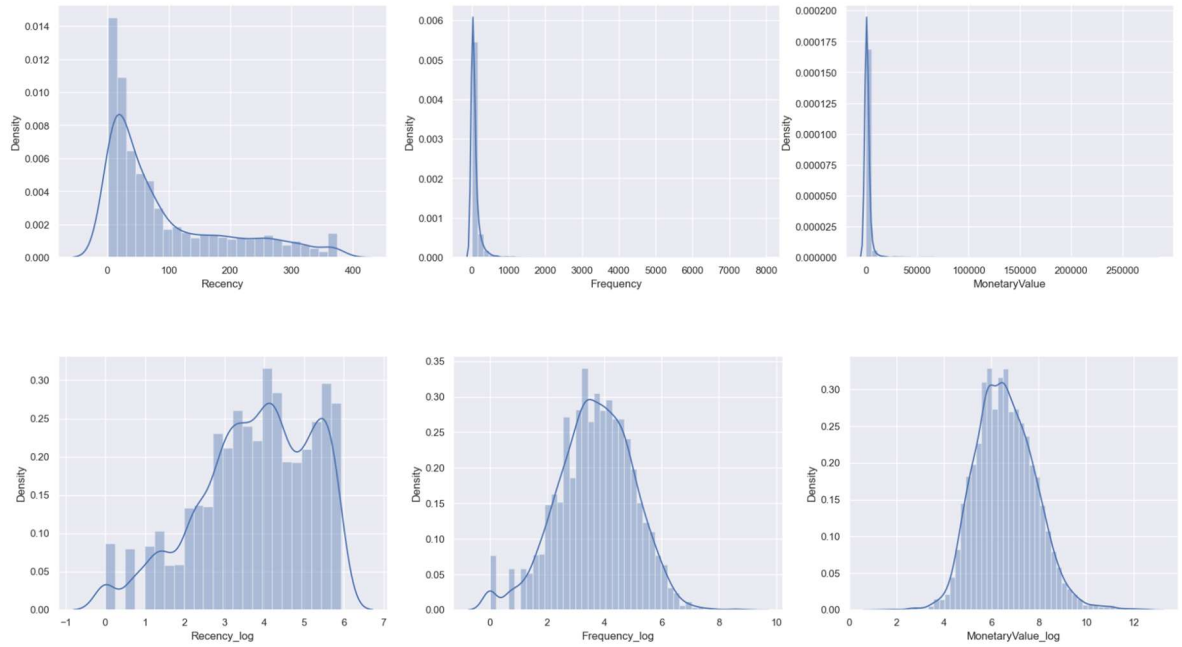

Fig10:data before and after applying log transformation (skewed data was normalized)

MODEL SELECTION

To determine the optimal number of clusters, we used the elbow method. This method involves plotting the inertia for each cluster count from 1 to 15 and identifying the "elbow point," where the rate of decrease in inertia slows down significantly. Based on this heuristic, we selected k=5k = 5k=5 for our model, as this point indicated a balance between the number of clusters and the variance explained.

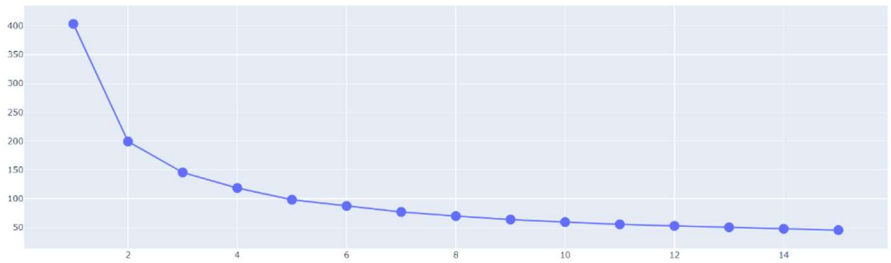

Fig11: using the elbow heuristic We decide to use k = 5 for our model.

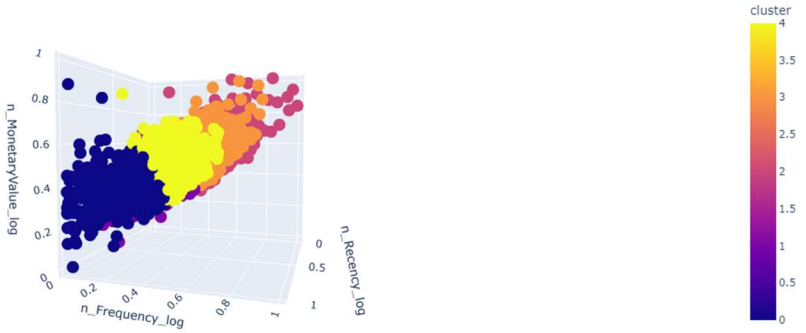

Fig12: 3D scatter plot

MODEL ANALYSIS (RESULTS)

The RFM model segmented the customers into five distinct clusters:

- **Cluster 0**: Customers in this cluster have high recency, frequency, and monetary value. These are the most valuable and highly engaged customers, contributing significantly to revenue. They make frequent purchases and spend a considerable amount per transaction.
- **Cluster 1**: These customers exhibit moderate recency, frequency, and monetary value. They are somewhat engaged but not as much as Cluster 0. They make purchases regularly and spend a moderate amount per transaction.
- **Cluster 2**: This cluster includes customers with low recency but high frequency and monetary value. They may not buy often but tend to spend a lot when they do. They are loyal customers who make significant purchases occasionally.
- **Cluster 3**: Customers in this cluster have moderate values across all three metrics. They are neither highly engaged nor disengaged, falling somewhere in between. Their purchasing behavior is consistent but not outstanding.
- **Cluster 4**: These customers have high recency but low frequency and monetary value. They are the least engaged, making infrequent purchases and spending relatively little per transaction. They may need re-engagement strategies to increase their activity.

	n_Recency			n_Frequency			n_MonetaryValue		
	mean	min	max	mean	min	max	mean	min	max
cluster									
0	0.619670	0.168901	1.000000	0.001886	0.000000	0.010706	0.001545	0.000000	0.275443
1	0.095597	0.005362	0.217158	0.002591	0.000000	0.012108	0.001461	0.000011	0.024072
2	0.005788	0.000000	0.018767	0.034290	0.000000	1.000000	0.026576	0.000415	1.000000
3	0.053130	0.016086	0.152815	0.020948	0.000892	0.208514	0.011788	0.000766	0.445788
4	0.283725	0.096515	0.994638	0.009454	0.000765	0.069080	0.004926	0.000469	0.158923

Fig13: 5 segmented clusters

CONCLUSION

The RFM model provides valuable insights into customer behaviour and allows for the development of targeted marketing strategies. By understanding the distinct needs and behaviours of different customer segments, businesses can enhance customer satisfaction, loyalty, and overall revenue. Tailored strategies for each cluster can include loyalty programs for Cluster 0, targeted promotions for Cluster 1, high value offers for Cluster 2, regular promotions for Cluster 3, and reactivation campaigns for Cluster 4.

SCOPE

The scope of this encompasses several key business objectives:

- Market Segmentation: Understanding the distinct segments within the customer base to tailor marketing efforts effectively.
- Customer Retention: Identifying high-value customers and developing strategies to retain them, thereby increasing their lifetime value.
- Targeted Marketing Campaigns: Designing personalized marketing campaigns for different customer segments to enhance engagement and conversion rates.
- Resource Allocation: Optimizing the allocation of marketing resources by focusing efforts on the most profitable customer segments.
- Customer Lifetime Value: Enhancing customer lifetime value by implementing strategies that encourage repeat purchases and increase average order value.
- Business Growth: Identifying new opportunities for growth by understanding customer behavior and preferences.

REFERENCES

- Chen,Daqing. (2015). Online Retail. UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.
- https://www.kaggle.com/code/mgmarques/customer-segmentation-and-market-basket-analysis

APPENDIX

Detailed RFM model calculations and transformations.

```
snapshot_date = df_ori['InvoiceDate'].max() + timedelta(days=1)
print(snapshot_date)
```

```
2011-12-10 12:50:00
```

```
#Extract features for each customer
data_process = df_ori.groupby(['CustomerID']).agg({
        'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
        'InvoiceNo': 'count',
        'Amount': 'sum'})

#renaming
data_process.columns = ['Recency', 'Frequency', 'MonetaryValue']

data_process = data_process
data_process[:3]
```

CustomerID	Recency	Frequency	MonetaryValue
12346	326	1	77183.60
12347	2	182	4310.00
12348	75	31	1797.24

```
fig, axes = plt.subplots(1, 3, figsize=(22, 5))
for i, feature in enumerate(list(data_process.columns)):
    sns.distplot(data_process[feature], ax=axes[i])
```

The data is skewed. Using log transformation we can improve the quality of the data for future analysis

```
data_process['Recency_log'] = data_process['Recency'].apply(math.log)
data_process['Frequency_log'] = data_process['Frequency'].apply(math.log)
data_process['MonetaryValue_log'] = data_process['MonetaryValue'].apply(math.log)
data_process[:3]
```

CustomerID	Recency	Frequency	MonetaryValue	Recency_log	Frequency_log	MonetaryValue_log
12346	326	1	77183.60	5.786897	0.000000	11.253942
12347	2	182	4310.00	0.693147	5.204007	8.368693
12348	75	31	1797.24	4.317488	3.433987	7.494007

```
scaler = MinMaxScaler()
#scaler = StandardScaler()
data_process_normalized = pd.DataFrame(scaler.fit_transform(data_process))
#renaming
data_process_normalized.columns = ['n_'+ i for i in data_process.columns]
data_process_normalized.describe()
```

	n_Recency	n_Frequency	n_MonetaryValue	n_Recency_log	n_Frequency_log	n_MonetaryValue_log
count	4338.000000	4338.000000	4338.000000	4338.000000	4338.000000	4338.000000
mean	0.245406	0.011563	0.007318	0.635951	0.410325	0.469546
std	0.268135	0.029159	0.032081	0.241793	0.147873	0.112364
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.045576	0.002039	0.001084	0.487888	0.315929	0.392678
50%	0.134048	0.005098	0.002394	0.663683	0.414097	0.462699
75%	0.378016	0.012618	0.005917	0.836532	0.513518	0.543051
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000