Foundations Data Analytics

Project 2

Prof. Sivarit (Tony) Sultornsanee

Team Members

Rushikesh Ghatage

Omkar Narkar

Parth Deshmukh

Sai Mahitha Etikala

Akshaya Murugan

Index

Introduction:

This project embarks on the exploration of customer behavior within the eCommerce realm through the lens of RFM (Recency, Frequency, Monetary) analysis. Serving as a strategic tool, RFM analysis facilitates customer segmentation based on recency, frequency, and monetary metrics. The ultimate goal is to empower businesses with insights for targeted marketing and improved customer engagement.

This methodology aspires not only to elucidate the intricacies of customer engagement but also to furnish businesses with a meticulously crafted blueprint for precision in marketing endeavors and the elevation of customer experiences. As we embark on this odyssey, the ensuing sections will systematically unveil the methodological intricacies, conduct in-depth analyses, and expound upon the strategic implications, presenting a holistic panorama of how RFM analysis can serve as a transformative catalyst, sculpting success within the intricate landscape of eCommerce.

Data Source:

The foundational dataset for this analysis originates from an eCommerce platform. It encompasses diverse variables that capture customer interactions, purchases, and related attributes. Prior to analysis, preprocessing steps are undertaken to handle missing values and ensure data uniformity.

The cornerstone of this analytical endeavor lies in the comprehensive dataset sourced from an eCommerce platform. This dataset proves to be a rich tapestry, woven with diverse variables meticulously designed to encapsulate various facets of customer interactions, purchases, and associated attributes. These variables range from transactional details to customer-specific information, providing a holistic view of the eCommerce landscape under examination.

Before delving into the substantive analysis, a critical preliminary phase involves preprocessing steps. These steps are undertaken with precision to address missing values and ensure a uniform and standardized structure across the dataset. By navigating through this meticulous preprocessing phase, we lay a robust foundation for the subsequent analyses, aiming for data integrity and reliability in the pursuit of meaningful insights.

Data Preprocessing:

In this dataset, 'Description' values are consistently linked with specific 'StockCodes.' To address missing 'Description' entries, a mode-based imputation is performed. For each 'StockCode,' the most frequent 'Description' is determined, and any missing values in 'Description' are replaced with their corresponding modes. This approach leverages the sequential association between 'StockCodes' and 'Description' items to enhance the accuracy of imputing missing information in the dataset. Also, for RFM analysis we removed the null customer IDs from the data set. Items with 0 unit-price are also excluded from the RFM analysis.

For payment analysis, we generated random data for payment methods and updated those values in a dataframe. Refund and return data are not available, so we considered canceled items as return items. Based on these, we conducted return and refund analysis. Data for profitability analysis and customer satisfaction analysis is not available.

Analysis:

Having meticulously prepared the dataset, we now embark on the analytical phase, aiming to get meaningful insights from the intricate tapestry of eCommerce interactions. Our analysis unfolds in several layers, each shedding light on distinct facets of customer behavior and transactional dynamics.

Exploratory Data Analysis (EDA):

In this initial stage, we employ statistical and visual methods to unravel patterns and trends within the dataset. EDA serves as our compass, guiding us through the vast sea of information to identify outliers, understand distributions, and discern potential correlations.

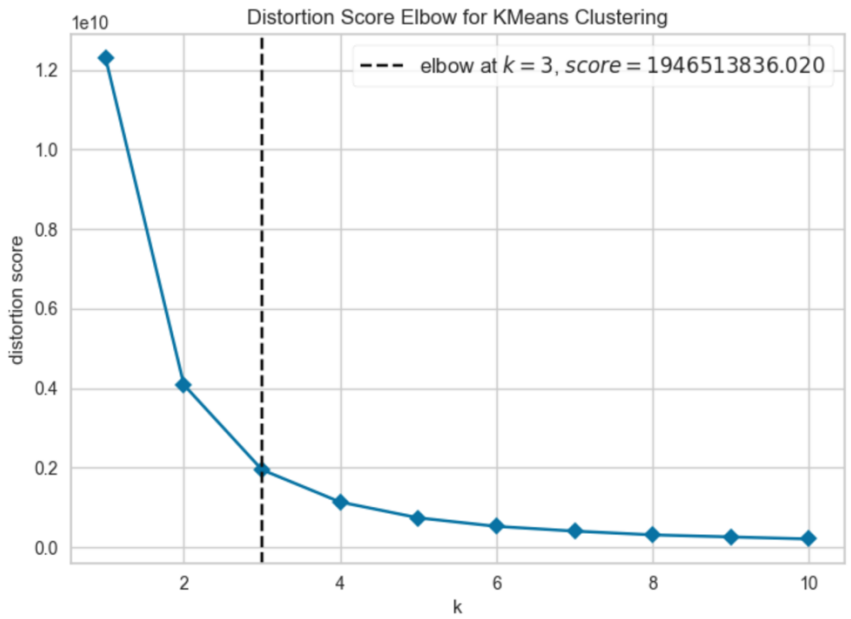
RFM Analysis:

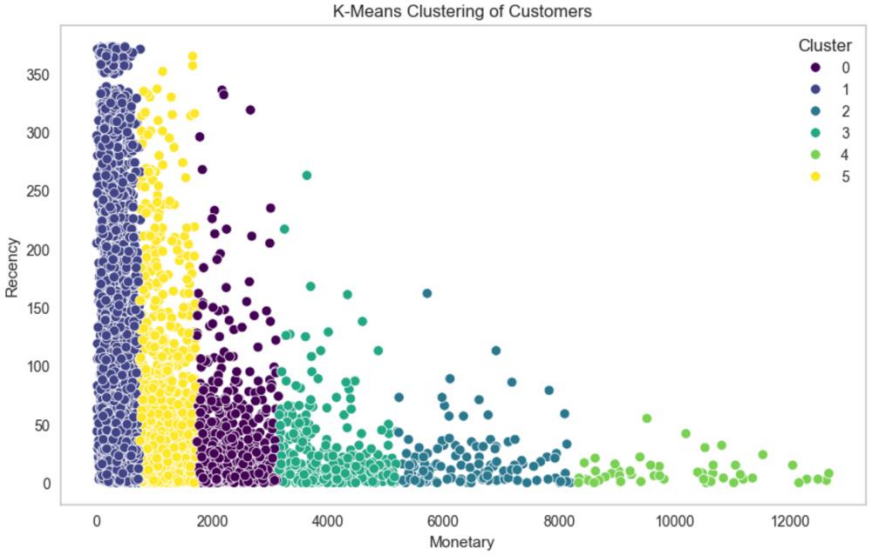
This segmentation aims to categorize customers into distinct groups, each offering unique insights into purchasing behaviors and engagement levels.

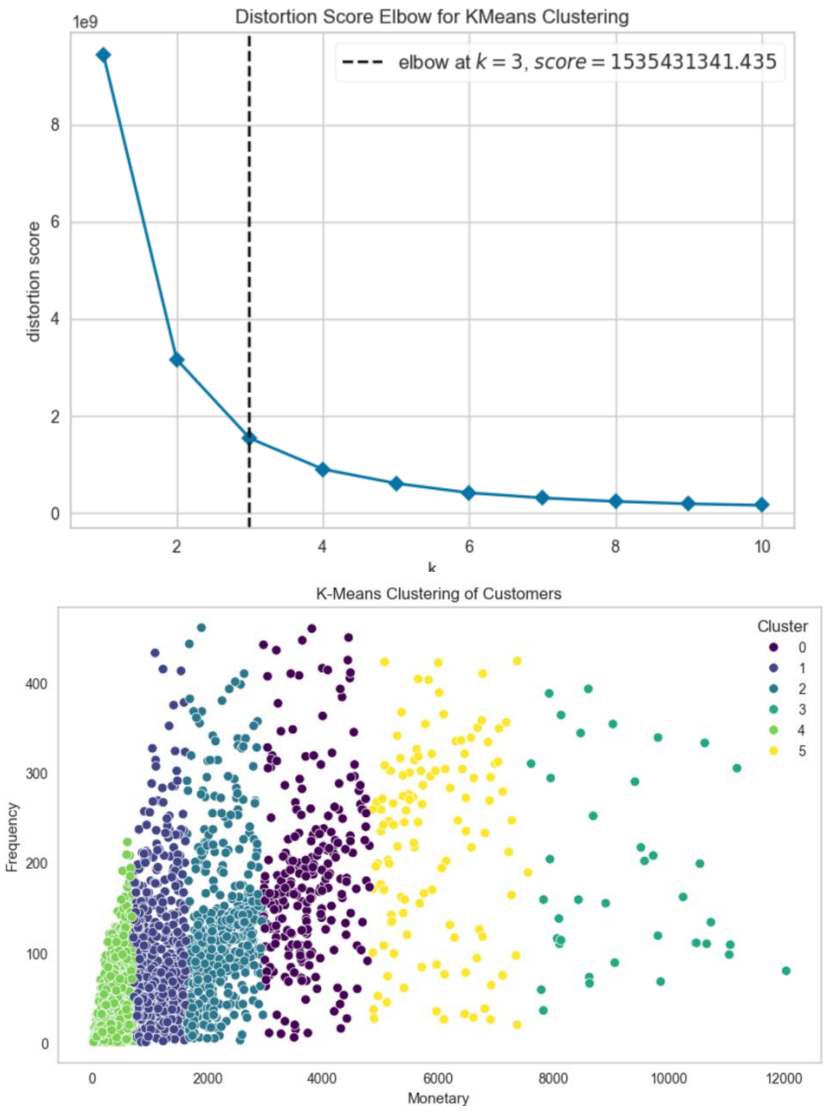
The analysis unfolds in a structured manner, leveraging key Python libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn. Initial steps involve data manipulation and visualization, setting the stage for the core RFM calculation. This entails computing recency, frequency, and monetary metrics, followed by a quartile-based scoring system for customer segmentation. We created 6 customer segments based on RFM score and based on those segments provided some marketing strategies for each segment

For KMean cluster implementation, two dataframes, df_fm (Frequency and Monetary) and df_rm (Recency and Monetary), were created for clustering. Z-score normalization was applied to df_fm and df_rm to handle outliers, and Isolation Forest was employed to further remove anomalies. The resulting dataframe, df_fm and df_rm, underwent K-means clustering after eliminating outliers, with the optimal number of clusters determined using the Elbow method.

K-Means clustering is employed for in-depth customer segmentation, with the elbow method guiding the determination of the optimal cluster count. Outlier management using Z-score normalization ensures the robustness of the analysis.


Distortion Score Elbow for KMeans Clustering
1e10
elbow at $k = 3$, $score = 1946513836.020$
distortion score
k
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2
4
6
8
10


K-Means Clustering of Customers
Cluster
0
1
2
3
4
5
Recency
Monetary
0
50
100
150
200
250
300
350
0
2000
4000
6000
8000
10000
12000



The narrative then delves into RFM scoring, emphasizing quartile-based ranking and the assignment of unique RFM scores to each customer. Visual representations, including pie charts and bar charts, offer a comprehensive view of the segmented customer landscape and geographical nuances.

Payment analysis extends the analysis scope, exploring payment methods and their relationships with order amounts. The technical storytelling unfolds through time analyses, return and refund assessments, profitability evaluations, and sentiment analysis of customer feedback.

Impact of Preprocessing on Insights:

Reflecting on the preprocessing steps, we assess how the handling of missing values and standardization has influenced the analytical outcomes. This reflection ensures transparency in understanding the dataset's reliability and the impact of data preparation on subsequent findings.

Marketing Strategies based on customer Segmentation:

1. Best Customer:

- Strategy: Offer exclusive loyalty rewards or a VIP program.

- Best customers are those who consistently spend a significant amount. To retain their loyalty, providing exclusive rewards, early access to new products, or a VIP program can make them feel valued and encourage them to buy more items.

2. Lost Customer:

- Strategy: Implement win-back campaigns with special discounts or incentives.

- Lost customers are those who haven't engaged recently. To re-engage them, win-back campaigns with special discounts, incentives, or personalized offers can bring them back to make a purchase.

3. Almost Lost Customer:

- Strategy: Provide targeted promotions or personalized offers.

- Almost lost customers may be showing signs of disengagement. Offering targeted promotions or personalized offers based on their past preferences can help retain their interest and prevent them from becoming fully lost.

4. Loyal Customer:

- Strategy: Acknowledge loyalty with exclusive access, early releases, or loyalty programs.

- Loyal customers are valuable assets. Recognizing their loyalty with exclusive access to new products, early releases, or participation in a loyalty program reinforces their commitment and encourage them to keep buying.

5. Big Spending Customer:

- Strategy: Offer premium products, exclusive bundles, or personalized services.

- Big spenders are willing to spend more money. Providing them with premium product options, exclusive bundles, or personalized services caters to their preferences and enhances their overall shopping experience.

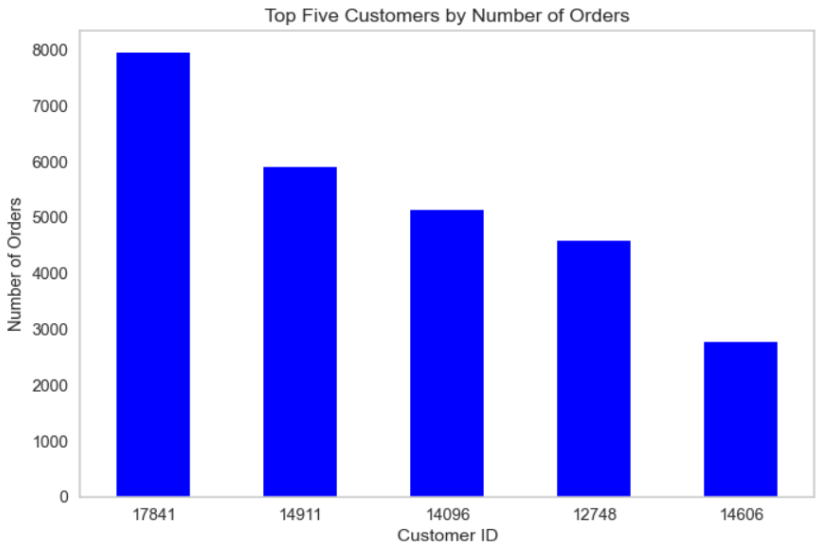
6. Visitor Customer:

- Strategy: Provide incentives for more frequent visits or purchases.

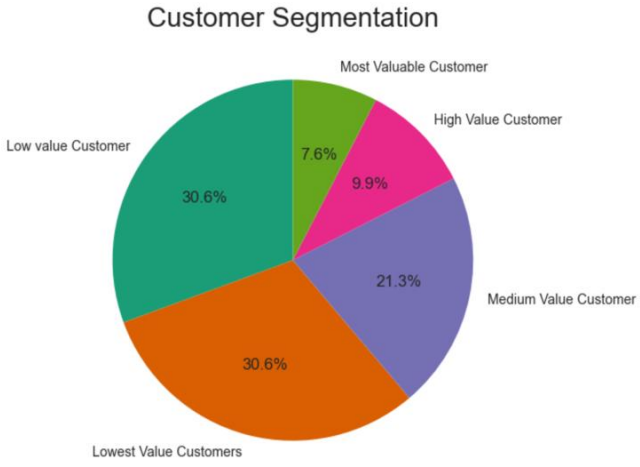
- Visitor customers are occasional shoppers. Encouraging more frequent visits or purchases with incentives such as discounts, loyalty points, or special offers can help convert them into regular customers.

Summary of Results:

The results of our analysis offer a comprehensive understanding of eCommerce customer behavior, distilling actionable insights and strategic recommendations for businesses. Through RFM segmentation, we unveil distinct customer groups based on recency, frequency, and monetary metrics, enabling tailored marketing strategies. Exploring temporal patterns guides strategic timing for promotions, while insights into monetary dynamics across segments optimize revenue generation. Geographical patterns, payment method dynamics, and other facets add layers to this understanding, painting a holistic portrait of the eCommerce customer landscape. As a result, businesses are equipped with nuanced intelligence to refine their marketing strategies, optimize promotional timing, and enhance customer engagement with precision in the dynamic eCommerce environment.

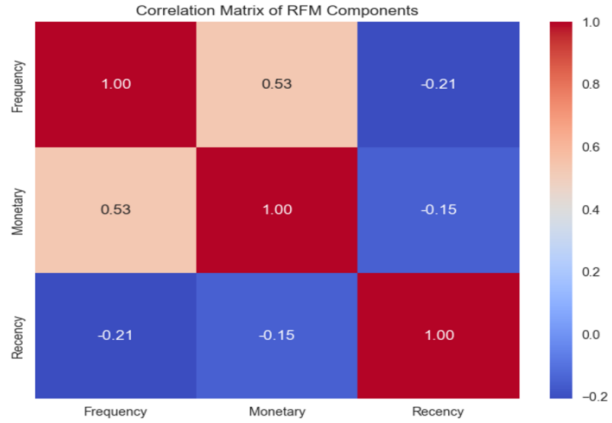



```
#Pie chart visualization
segment_counts = rfm_df['Customer Segment'].value_counts()
plt.figure(figsize=(8,6))
plt.pie(segment_counts, labels=segment_counts.index, autopct='%1.1f%%', startangle=90, colors= sns.color_palette("Da
plt.title('Customer Segmentation',size = 20)
plt.show()
```

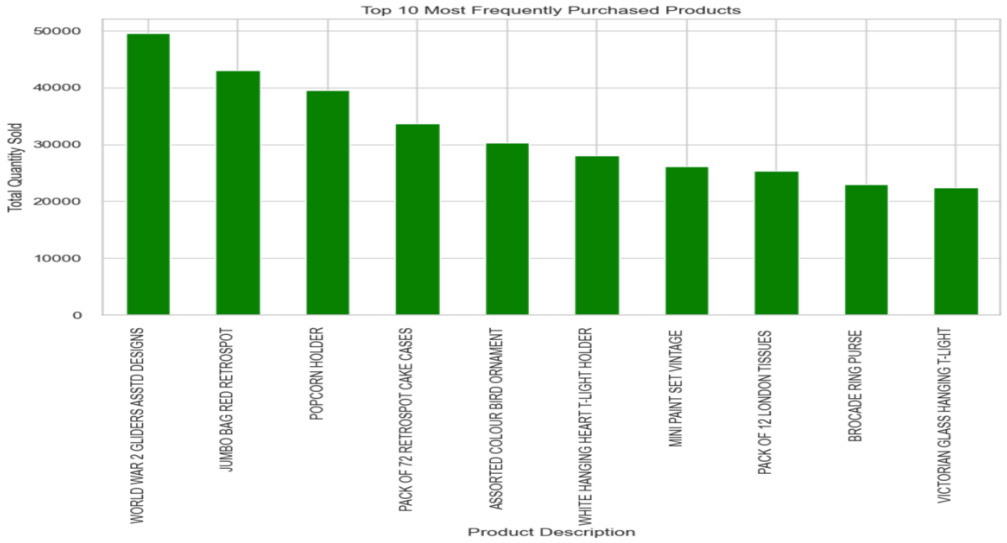


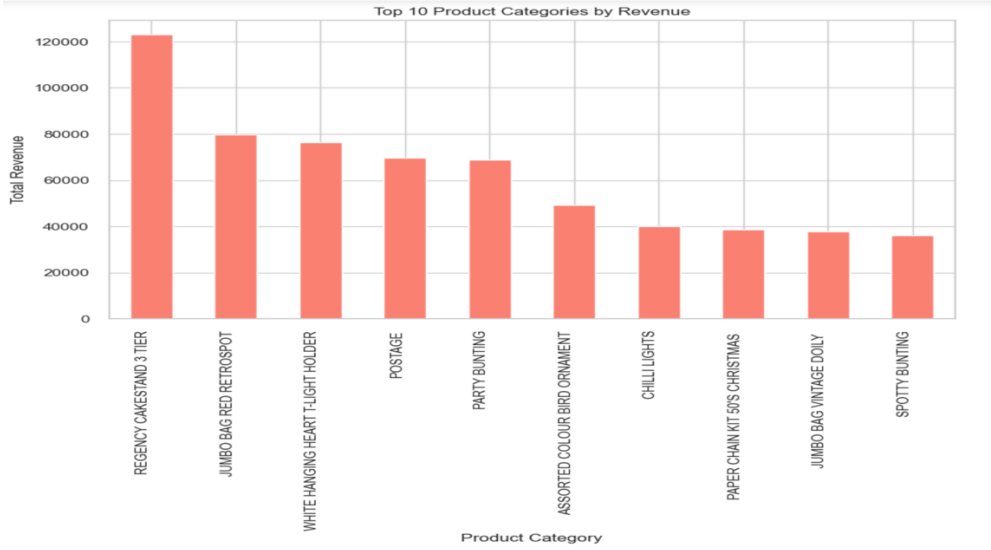
```
# correlation between frequency , recency and monetary
cor_matrix = rfm_df[['Frequency','Monetary','Recency']].corr()
print(cor_matrix)

plt.figure(figsize=(8, 6))
sns.heatmap(cor_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix of RFM Components')
plt.show()
```

```
             Frequency  Monetary   Recency
Frequency     1.000000  0.531348 -0.206486
Monetary      0.531348  1.000000 -0.150159
Recency      -0.206486 -0.150159  1.000000
```



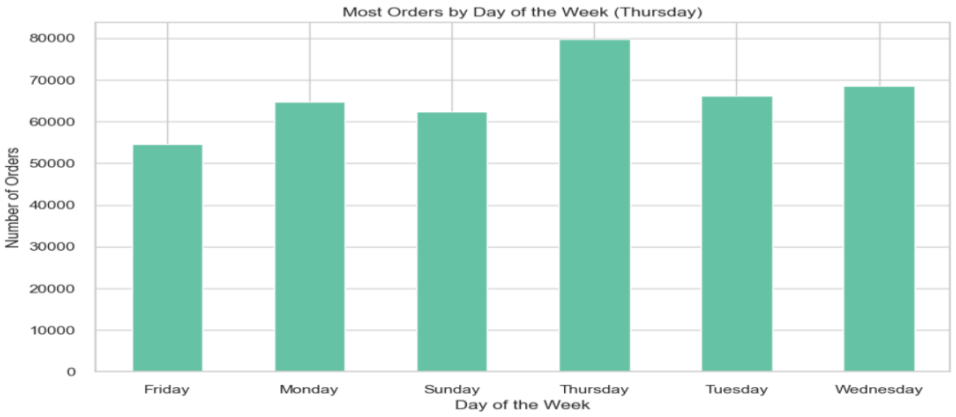
```
#Visualization

sns.set(style="whitegrid")
# Visualization for Top 10 most frequently purchased products
plt.figure(figsize=(10, 6))
top_products.plot(kind='bar', color='green')
plt.title('Top 10 Most Frequently Purchased Products')
plt.xlabel('Product Description')
plt.ylabel('Total Quantity Sold')
plt.xticks(rotation=90, ha='right')
plt.show()

# Visualization for Product category generating the highest revenue
plt.figure(figsize=(10, 6))
revenue_by_category.head(10).plot(kind='bar', color='salmon')
plt.title('Top 10 Product Categories by Revenue')
plt.xlabel('Product Category')
plt.ylabel('Total Revenue')
plt.xticks(rotation=90, ha='right')
plt.show()
```

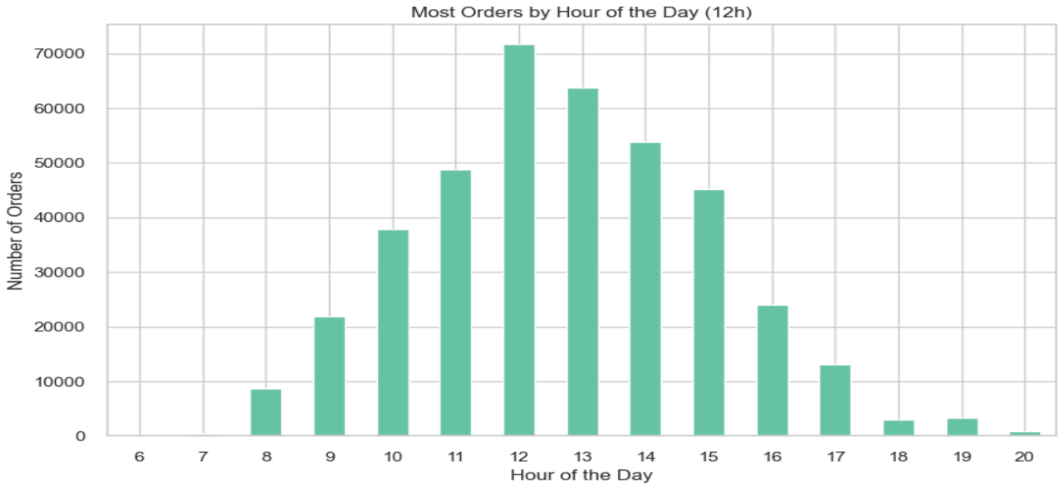


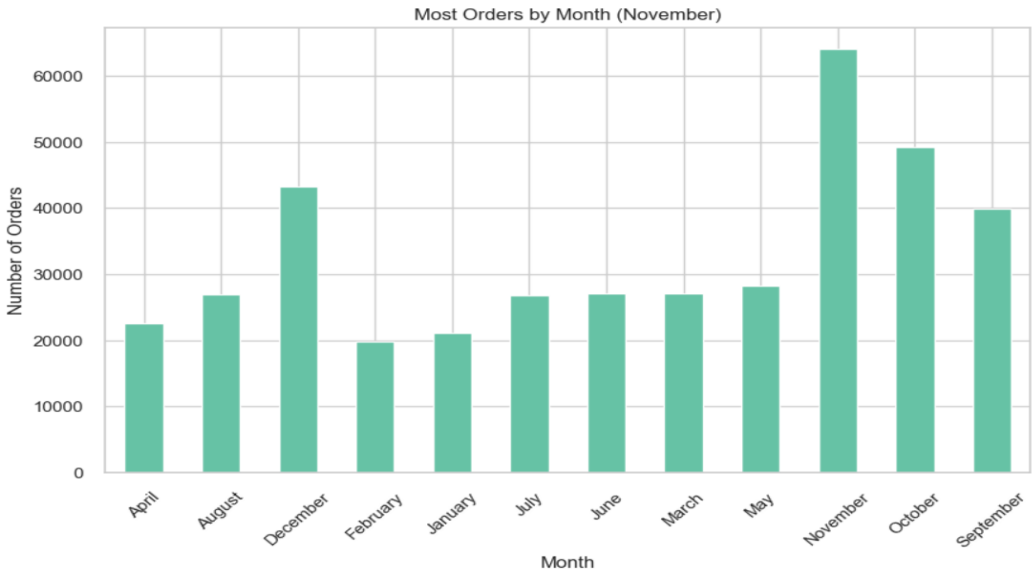


```
#4 Time Analysis
df1['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df1['DayOfWeek'] = df1['InvoiceDate'].dt.day_name()
df1['HourOfDay'] = df1['InvoiceDate'].dt.hour

most_orders_day = df1['DayOfWeek'].value_counts().idxmax()
most_orders_hour = df1['HourOfDay'].value_counts().idxmax()
most_orders_month = df1['InvoiceDate'].dt.month_name().value_counts().idxmax()

monthly_order_trends = df1.resample('M', on='InvoiceDate').size()
print("Day of the week with most orders:", most_orders_day)
print("Hour of the day with most orders:", most_orders_hour)
print("Month with most orders:", most_orders_month)
```

```
Day of the week with most orders: Thursday
Hour of the day with most orders: 12
Month with most orders: November
```



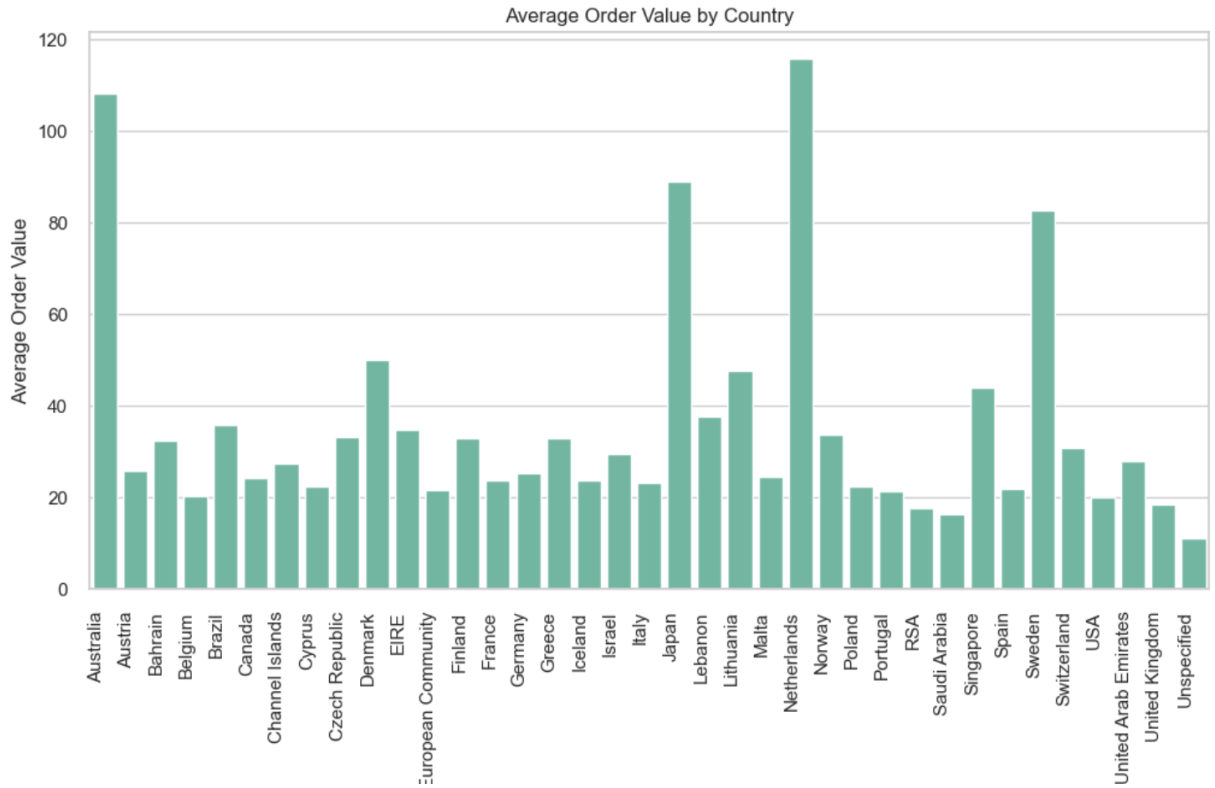





```
#5 Geographical Analysis

top_countries = df1['Country'].value_counts().head(5)

average_order_value_by_country = df1.groupby('Country')['TotalPrice'].mean()
average_order_value_by_country.rename({'TotalPrice': 'AverageValue'}, inplace=True)

print(f"\n\nTop 5 countries with the highest number of orders:\n{top_countries}")
print(f"\n\nAverage order value by country:\n{average_order_value_by_country.reset_index()}\n\n")


plt.figure(figsize=(12, 6))
sns.barplot(x=average_order_value_by_country.index, y=average_order_value_by_country.values)
plt.title("Average Order Value by Country")
plt.xlabel("Country")
plt.ylabel("Average Order Value")
plt.xticks(rotation=90, ha='right')
plt.show()
```

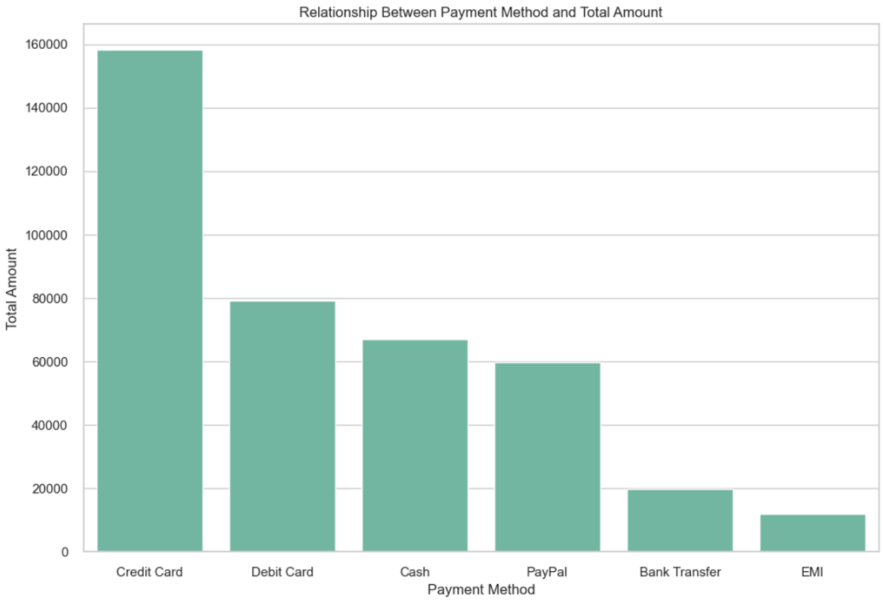
```
Top 5 countries with the highest number of orders:
Country
United Kingdom    352616
Germany             9039
France              8336
EIRE                7223
Spain               2477
Name: count, dtype: int64


Average order value by country:
                Country  TotalPrice
0             Australia  108.083612
1               Austria   25.624824
2               Bahrain   32.258824
3               Belgium   20.283772
4                Brazil   35.737500
5                Canada   24.280662
6        Channel Islands   27.363507
7                Cyprus   22.256535
8         Czech Republic   33.069600
9               Denmark   49.882474
10                EIRE   34.605110
11   European Community   21.670833
12              Finland   32.913985
13               France   23.744026
14              Germany   25.319798
15               Greece   32.831172
16              Iceland   23.681319
17               Israel   29.339390
18                Italy   23.064960
19                Japan   88.842673
20              Lebanon   37.641778
21            Lithuania   47.458857
```

21	Lithuania	47.458857
22	Malta	24.335625
23	Netherlands	115.701607
24	Norway	33.767918
25	Poland	22.226212
26	Portugal	21.228278
27	RSA	17.584386
28	Saudi Arabia	16.213333
29	Singapore	43.881521
30	Spain	21.802325
31	Sweden	82.645178
32	Switzerland	30.659397
33	USA	20.002179
34	United Arab Emirates	27.974706
35	United Kingdom	18.381439
36	Unspecified	11.005455



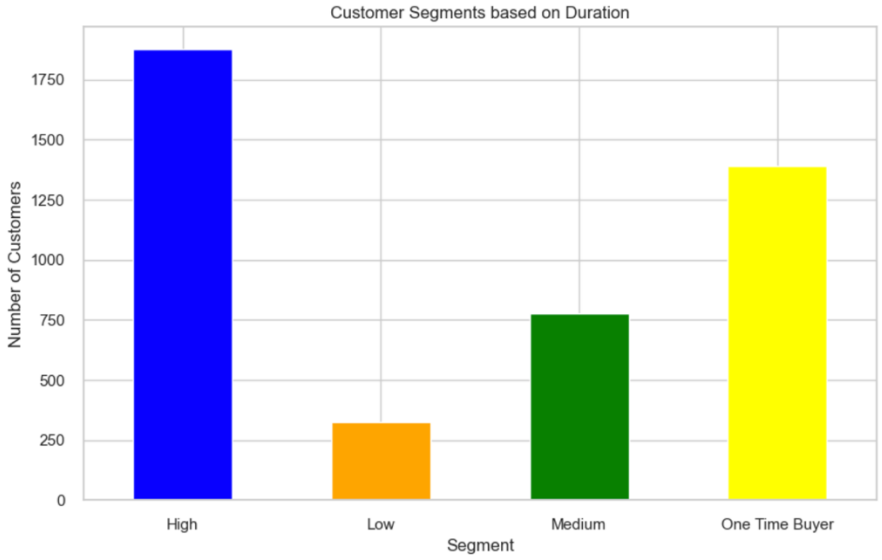
```python
#6 Payment Analysis

common_payment_methods = df1['PaymentMethod'].value_counts()
print(f"Most Common Payment Methods:\n{common_payment_methods}")
common_payment_methods = common_payment_methods.reset_index()

plt.figure(figsize=(12, 8))
sns.barplot(x='PaymentMethod', y='count', data=common_payment_methods)
plt.title('Relationship Between Payment Method and Total Amount')
plt.xlabel('Payment Method')
plt.ylabel('Total Amount')
plt.show()
```

```
Most Common Payment Methods:
PaymentMethod
Credit Card      158408
Debit Card        79165
Cash              67049
PayPal            59845
Bank Transfer     19786
EMI               11859
Name: count, dtype: int64
```

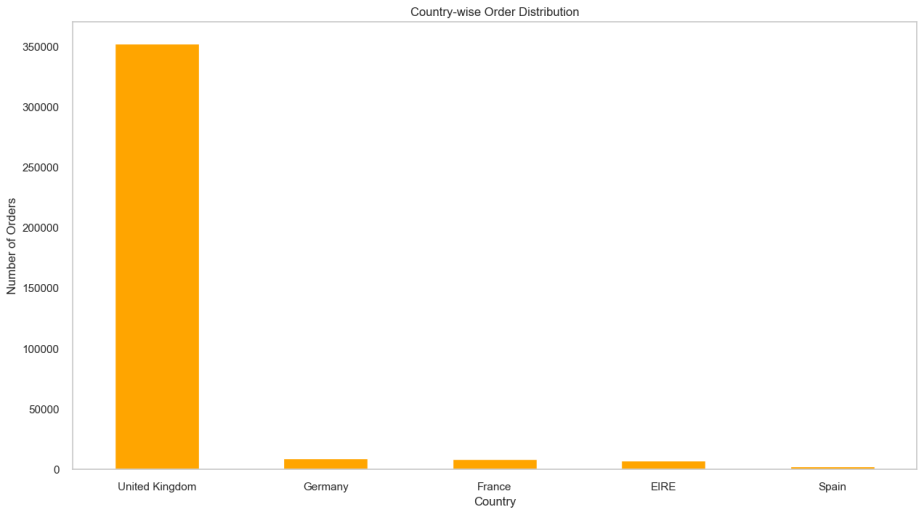


```
#7 Customer Behaviour
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
customer_duration = df.groupby('CustomerID')['InvoiceDate'].agg(['min', 'max'])
customer_duration['Duration'] = (customer_duration['max'] - customer_duration['min']).dt.days

average_duration = customer_duration['Duration'].mean()
x = customer_duration['Duration'].max()
y = customer_duration['Duration'].min()
print(f"Maximum Duration between two Purchases: {x}\nMinimum Duration between two Purchases: {y}")
print(f"Average Customer Lifespan: {average_duration:.2f} days")

def customer_seg(duration):
    if duration == 0:
        return 'One Time Buyer'
    elif 0 < duration <= 30:
        return 'Low'
    elif 30 < duration <= 150:
        return 'Medium'
    else:
        return 'High'

customer_duration['DurationSegment'] = customer_duration['Duration'].apply(customer_seg)
plt.figure(figsize=(10, 6))
customer_duration['DurationSegment'].value_counts().sort_index().plot(kind='bar', color=['blue', 'orange', 'green','
plt.title('Customer Segments based on Duration')
plt.xlabel('Segment')
plt.ylabel('Number of Customers')
plt.xticks(rotation=0)
plt.show()
```

```
Maximum Duration between two Purchases: 373
Minimum Duration between two Purchases: 0
Average Customer Lifespan: 133.42 days
```



```
# Country wise Analysis
country_order_distribution = df1['Country'].value_counts().head(5)

sns.set_palette("Dark2")
plt.figure(figsize=(15, 8))
country_order_distribution.plot(kind='bar',color = 'Orange')  # Use country_order_distribution instead of country_or
plt.title('Country-wise Order Distribution')
plt.xlabel('Country')
plt.ylabel('Number of Orders')
plt.xticks(rotation=0, ha='center')
plt.grid(False)
plt.show()
```



Result:

Our analysis goes beyond data, manifesting tangible results through compelling visual narratives that highlight correlations, trends, and outliers. Businesses now possess a nuanced understanding of eCommerce customer behavior, thanks to meticulous RFM segmentation. This segmentation not only forms the foundation for tailored marketing strategies but also serves as a visual guide for data-driven decision-making.

Temporal patterns pinpoint strategic windows for promotions, optimizing the impact of marketing initiatives. Insights into monetary dynamics across customer segments enable businesses to fine-tune revenue optimization strategies for both high-value and lower-value customer segments.

Incorporating geographical patterns and payment method dynamics adds additional layers to our understanding, offering businesses a comprehensive view of customer behavior in the eCommerce realm. These visual narratives are not only descriptive but also prescriptive, providing a roadmap for businesses to refine marketing strategies, optimize promotional timing, and foster customer engagement with precision.

Conclusion:

The project concludes by synthesizing the key findings and their implications. It underscores the importance of RFM analysis in deciphering customer behavior and emphasizes the role of Python libraries in orchestrating a comprehensive analysis.

Limitation:

While our analysis provides valuable insights, it is essential to acknowledge certain limitations. The scope and representativeness of the dataset may influence the generalizability of findings, especially if it lacks diversity. Assumptions in RFM segmentation, such as uniform metric significance, pose potential challenges to accuracy. Temporal patterns may be context-dependent and influenced by external factors, impacting their sustainability. External influences on monetary dynamics, regional variations, and payment method dynamics add layers of complexity. The analysis primarily identifies correlations, not causation, and the dynamic nature of the eCommerce landscape implies findings may be time-sensitive. Recognizing and addressing these limitations enhances the report's credibility, offering a nuanced understanding within defined constraints.

Future Work:

The analysis paves the way for future exploration. Potential avenues for further research and enhancement of the analysis are outlined, ensuring a dynamic and evolving approach to understanding eCommerce customer behavior.

In the future landscape of eCommerce analytics, advancements include exploring sophisticated segmentation methods, real-time analysis, and incorporating external data for a nuanced grasp of customer behavior. Enhanced analytical capabilities are sought through the creation of predictive models, personalized marketing strategies, and the integration of qualitative insights from customer feedback. Further expansion into the analysis of cross-channel behavior, a commitment to ethical considerations, and the establishment of continuous monitoring frameworks contribute to a refined and adaptive approach. Benchmarking against industry standards provides a contextual evaluation of performance and areas for enhancement, collectively aiming to elevate the depth and agility of eCommerce analytics.