“Customer Segmentation and Predicting Lifetime Value in E-Commerce”

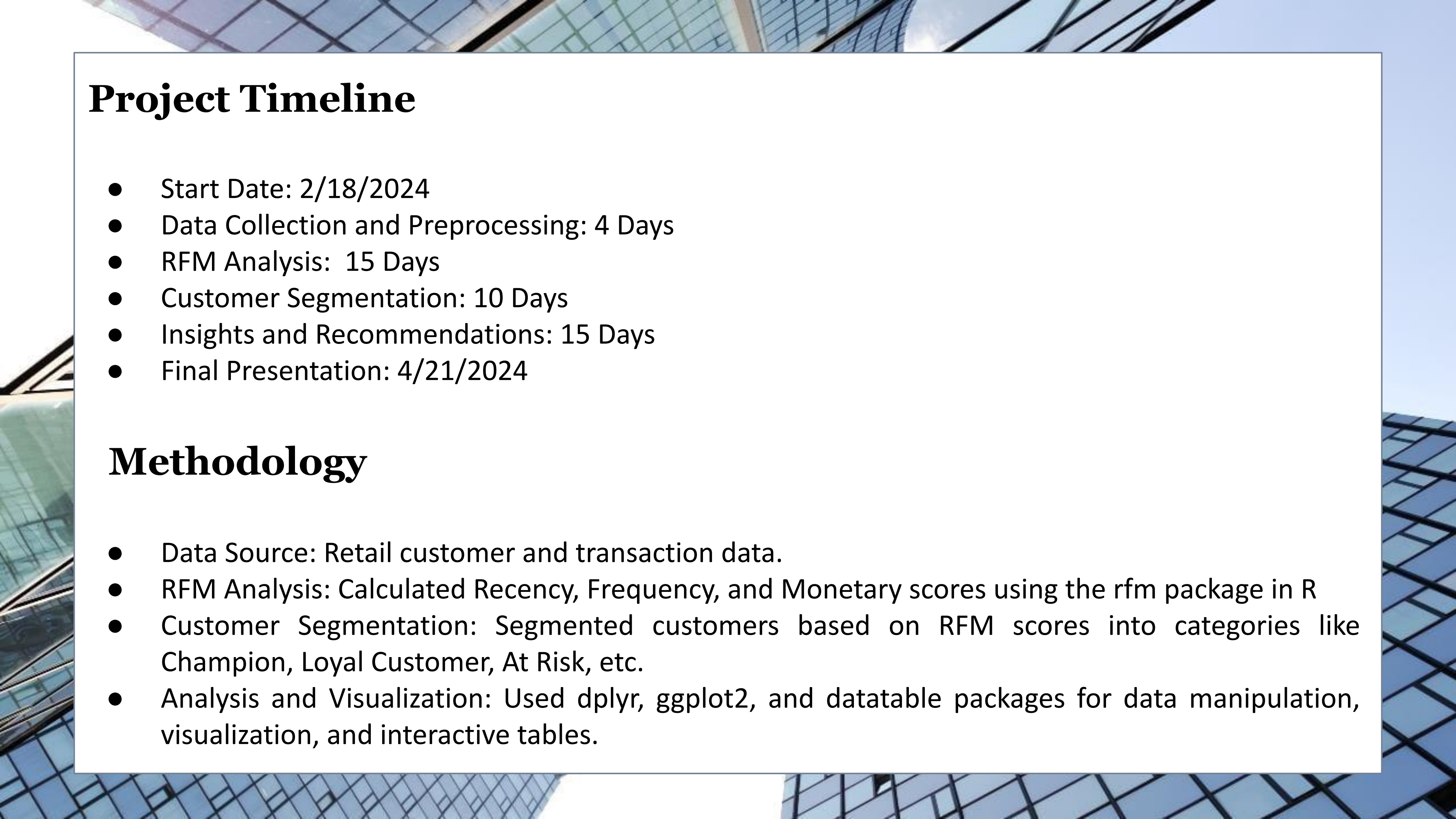
CSP 571 - Project Presentation

CONTENTS

- Tasks and Changes
- Data processing
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Customer Segmentation
- RFM Feature Engineering
- K-means Clustering
- RFM Analysis
- Conclusion
- Executive Summary
- Key Research Issue
- Future Work
- References

Overview of project team

1. Satwika Shaganti (Group Leader)- CWID: A20554319
2. Bhavana Polakala - CWID: A20539792
3. Lakshmi Sindhu Pulugundla - CWID: A20553567
4. Sai Venkata Vamsi Krishna Yelike - CWID: A20543669



Project Timeline

- Start Date: 2/18/2024
- Data Collection and Preprocessing: 4 Days
- RFM Analysis: 15 Days
- Customer Segmentation: 10 Days
- Insights and Recommendations: 15 Days
- Final Presentation: 4/21/2024

Methodology

- Data Source: Retail customer and transaction data.
- RFM Analysis: Calculated Recency, Frequency, and Monetary scores using the rfm package in R
- Customer Segmentation: Segmented customers based on RFM scores into categories like Champion, Loyal Customer, At Risk, etc.
- Analysis and Visualization: Used dplyr, ggplot2, and datatable packages for data manipulation, visualization, and interactive tables.

Tasks and Changes

- Initial Task: Identify the most valuable customers and segment the customer base
- Change 1: Adjusted the customer segmentation criteria to handle overlapping conditions
- Change 2: Included an analysis of median monetary value by segment to understand revenue contribution
- Additional Task: Create an interactive table to display customer segments for targeted marketing strategies.

Data Cleaning

1. **Handling Missing Values:**

   In this, some missing values are identified while working with the dataset, and in the place of missing values, some assumption values are replaced.

2. **Removing Duplicate Records & Correcting Inaccurate Data:**

   In this dataset, duplicate and some inaccurate values are identified these are fixed with relevant data from dataset itself

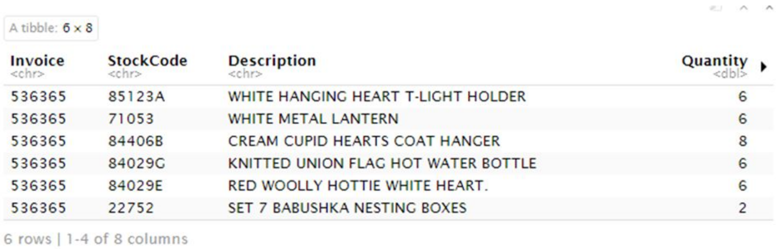
3. **Handling inconsistent data:**

   Resolving Inconsistencies in Categorical Data.

4. **Removing Irrelevant Data:**

   In this dataset, some of irrelevant data is removed

Dataset output



A tibble: 6 × 8

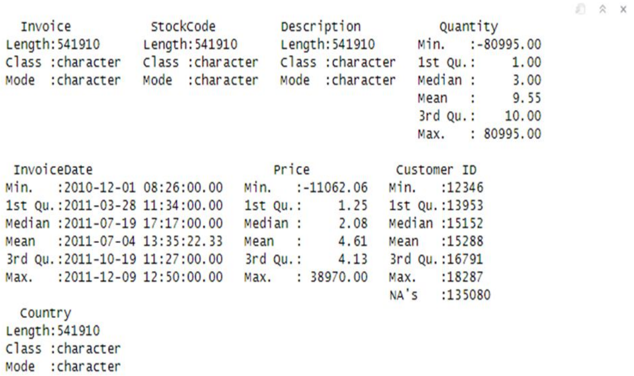
Invoice <chr>	StockCode <chr>	Description <chr>	Quantity <dbl>
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6
536365	71053	WHITE METAL LANTERN	6
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6
536365	22752	SET 7 BABUSHKA NESTING BOXES	2

6 rows | 1-4 of 8 columns

View of first Few rows of Data

1. The data consists of **541,909 rows and 8 columns/features**, namely - “Invoice”, “StockCode” ,“Description”, “Quantity”, “InvoiceDate”, “Price”, “Customer ID”, “Country”.
2. We can also see that the values of Invoice, which are suppose to be num, are categorised as character. Also the InvoiceDate feature values are in a bad format, which we will have to change during the data cleaning part.
3. Also, we can observe that the column names consist of spaces, which may cause errors in further analysis. Hence, we will also change the column names to make it easy in the future steps.

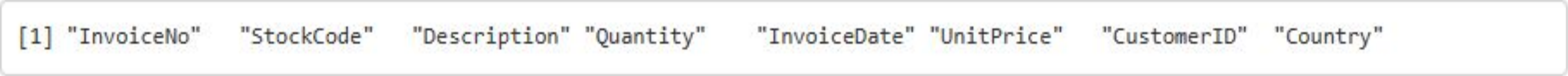
Summary of the Data



```
  Invoice            StockCode          Description           Quantity
Length:541910      Length:541910      Length:541910      Min.   :-80995.00
Class :character   Class :character   Class :character   1st Qu.:     1.00
Mode  :character   Mode  :character   Mode  :character   Median :     3.00
                                                         Mean   :     9.55
                                                         3rd Qu.:    10.00
                                                         Max.   : 80995.00

 InvoiceDate                         Price             Customer ID
Min.   :2010-12-01 08:26:00.00   Min.   :-11062.06   Min.   :12346
1st Qu.:2011-03-28 11:34:00.00   1st Qu.:     1.25   1st Qu.:13953
Median :2011-07-19 17:17:00.00   Median :     2.08   Median :15152
Mean   :2011-07-04 13:35:22.33   Mean   :     4.61   Mean   :15288
3rd Qu.:2011-10-19 11:27:00.00   3rd Qu.:     4.13   3rd Qu.:16791
Max.   :2011-12-09 12:50:00.00   Max.   : 38970.00   Max.   :18287
                                                     NA's   :135080
  Country
Length:541910
Class :character
Mode  :character
```

Changing Column names:

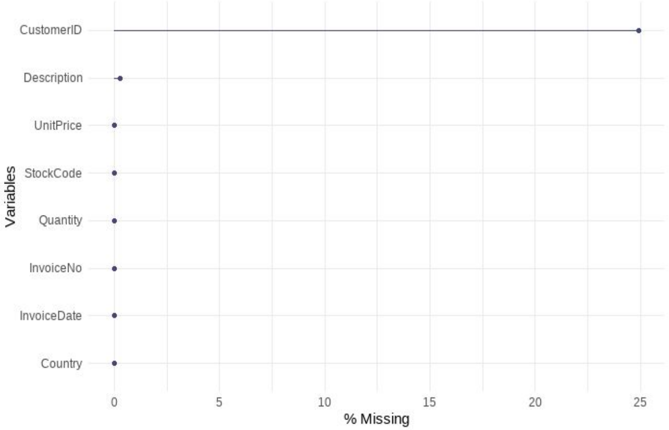


```
[1] "InvoiceNo"   "StockCode"   "Description" "Quantity"    "InvoiceDate" "UnitPrice"   "CustomerID"  "Country"
```

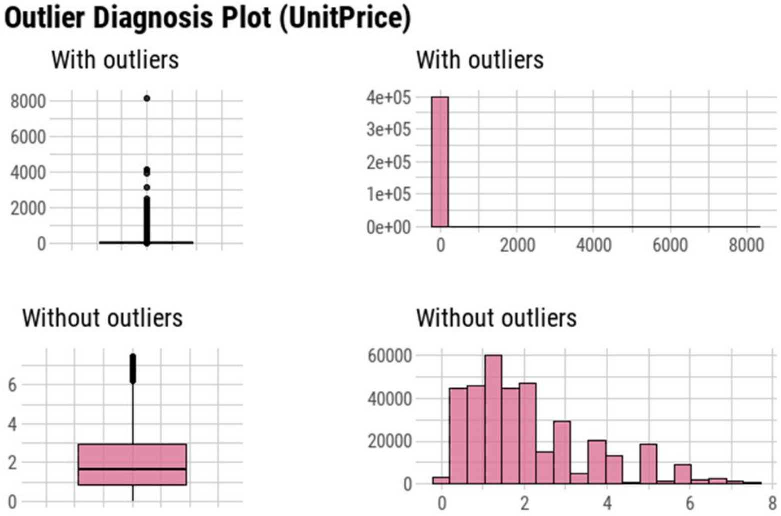
Data processing

Data processing, also known as data cleansing or data scrubbing, is a crucial process in the field of data analysis and data science. It involves detecting and correcting errors or inconsistencies in data to improve its quality, reliability, and accuracy. The goal of data cleaning is to ensure that data is correct, complete, relevant, and properly formatted for analysis or other downstream tasks.

We can see that the total number of missing values in the dataset are 136534, of which 1454 are from the `Description` column while the remaining 135080 are from `CustomerID` column.

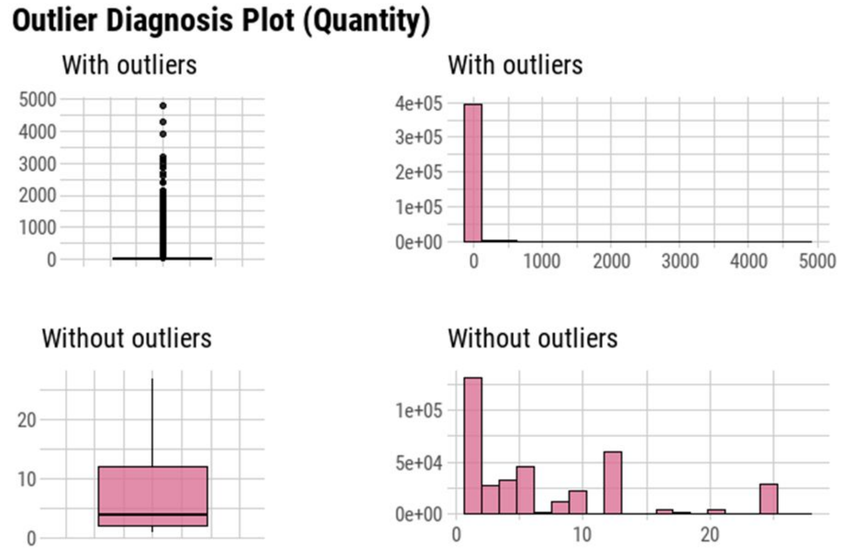


Outliner Diagnosis plot of Unit price



This graph gives us clear picture of outliner’s of Unit Price, and also in addition a comparison is given between with outliner’s and without outliner’s.

Outliner Diagnosis plot of Quantity

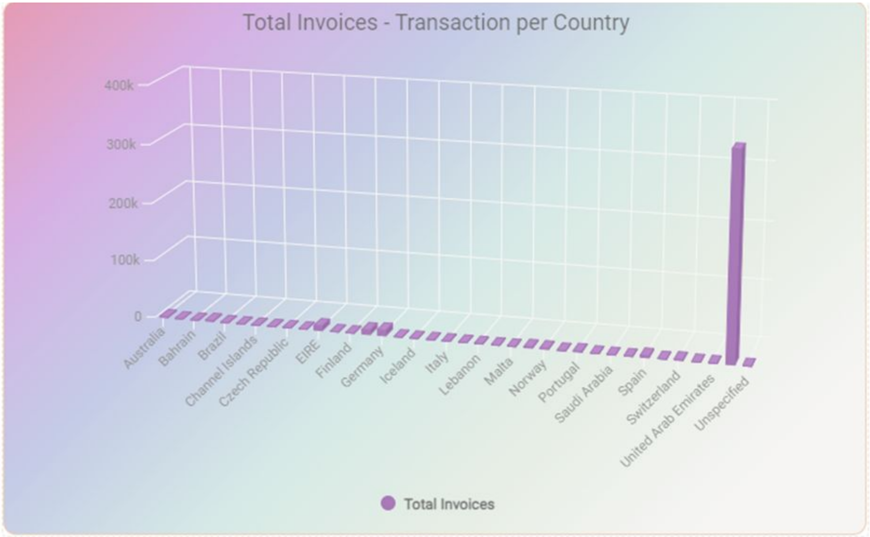


This graph gives us clear picture of outliner's of Quantity, and also in addition a comparison is given between with outliner's and without outliner's.

Exploratory Data Analysis (EDA)

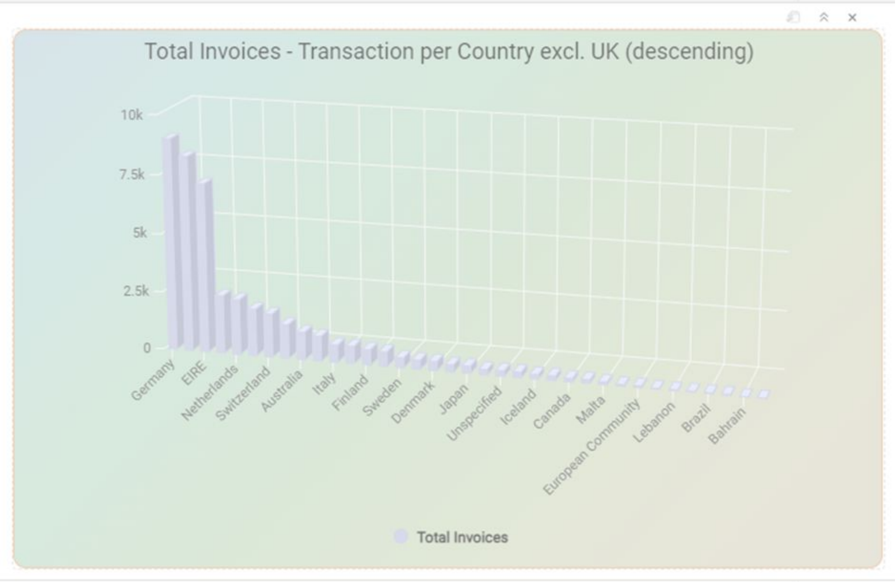
- EDA is an important factor to conduct an initial investigation inside the dataset
- Observe common patterns, spot anomalies, and retrieve useful information about the data in a graphical way
- By conducting an in-depth exploration of the data, we can identify key factors that may influence the store's performance, customer behavior, and potential areas for improvement or growth.

Total Invoices - Transaction per Country



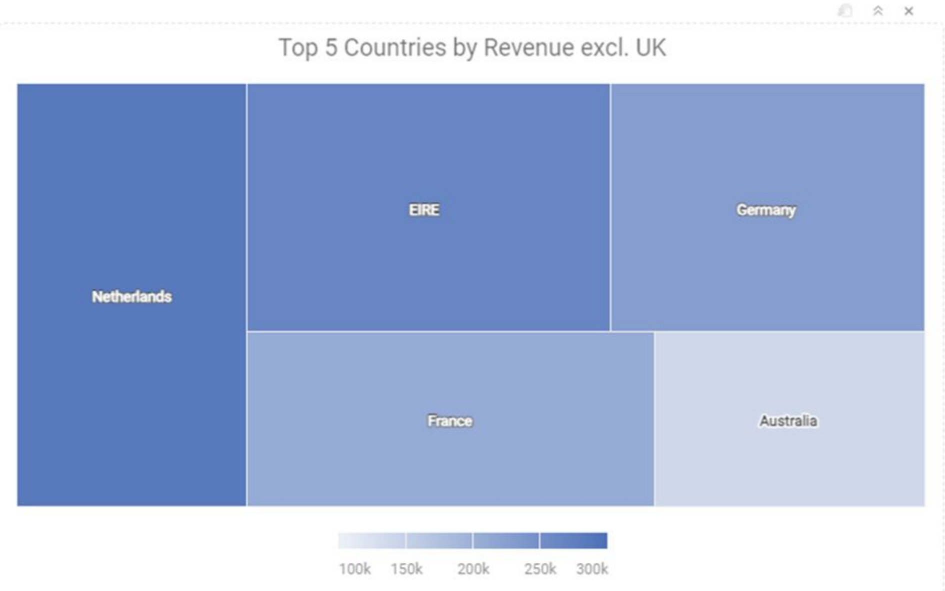
- This chart provides a clear overview of the store's customer base and highlights the countries with the highest transaction volumes.
- United Kingdom has the highest number of transactions (store is based in UK)

Transactions Outside the UK



•Germany and France emerge as the countries with the highest transaction volumes outside the UK, suggesting that these markets play a significant role in the store's international operations.

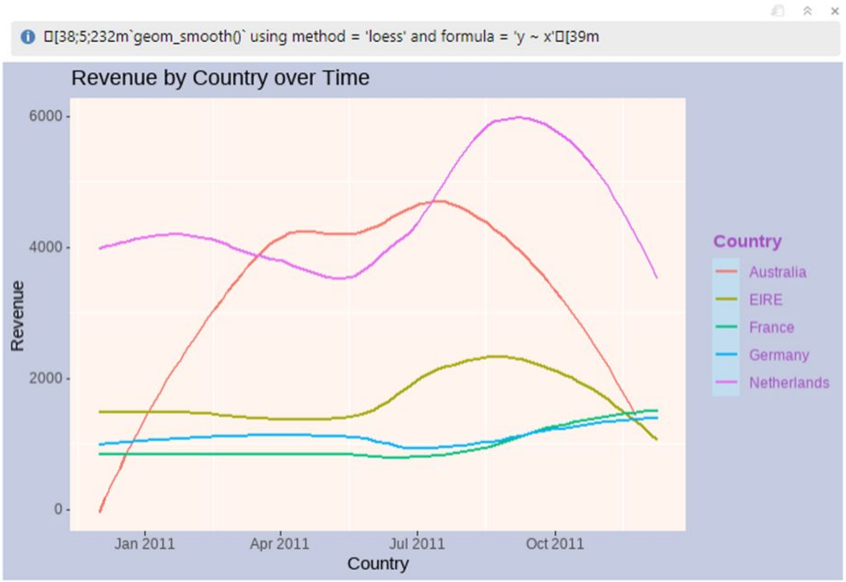
Top 5 Countries by Revenue excl. UK



- Treemap of top 5 revenue countries ex-UK: Netherlands, Ireland, Germany, France, Australia

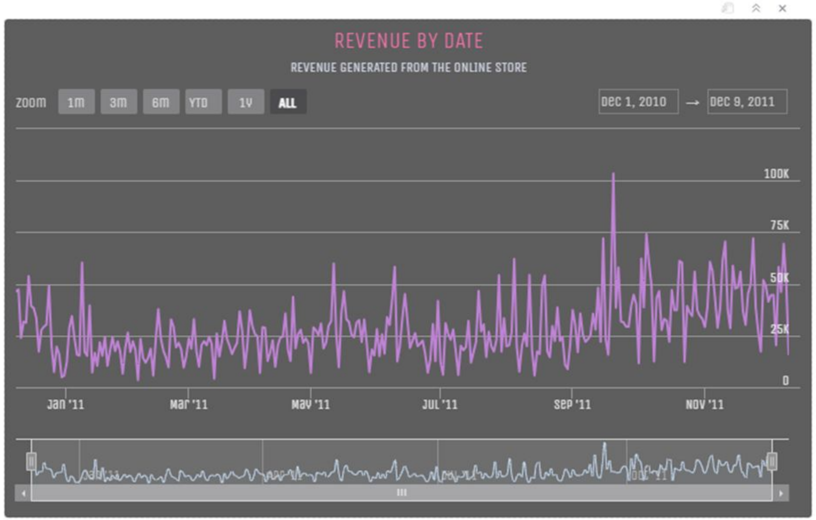
Revenue by Country over Time

- Line chart: Revenue by country over time
- Sharp drop for Netherlands and Australia after summer 2011 (investigate causes)
- Stock chart: Overall revenue by date, steady 2011 increase, peak around Sept 20th



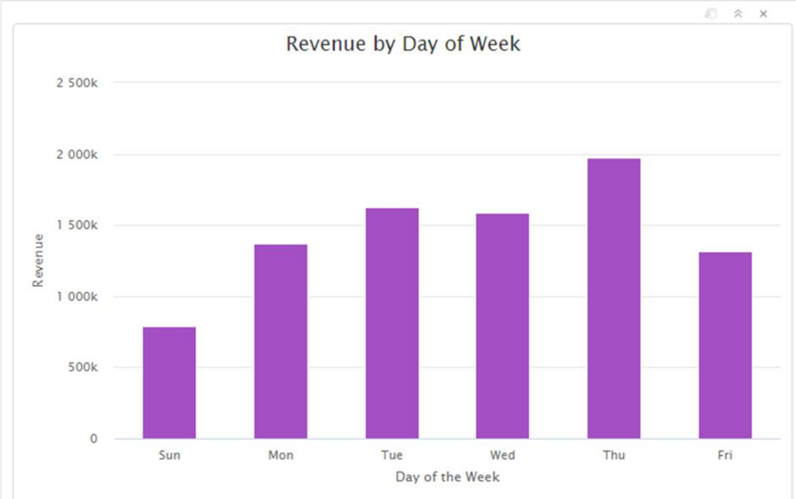
```
□[38;5;232m`geom_smooth()` using method = 'loess' and formula = 'y ~ x'□[39m
```

Revenue by Date



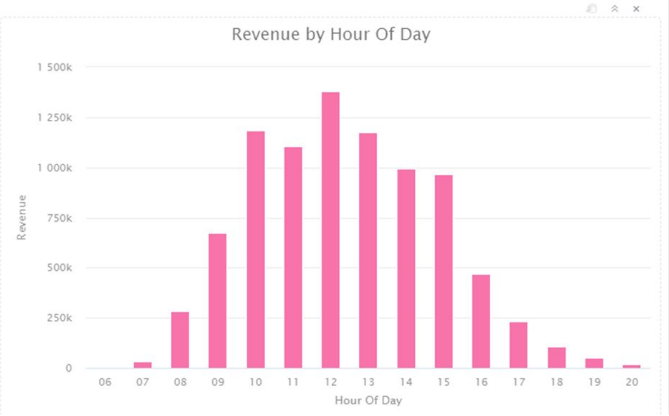
This stock chart provides an overall view of the store's revenue pattern throughout the year, allowing us to identify peak periods, seasonal trends, or potential external influences on revenue generation.

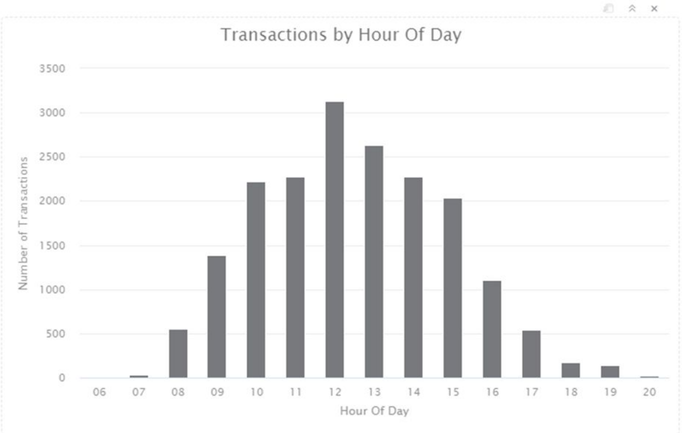
Revenue by Day of Week



This column chart illustrates the total revenue generated on different days of the week, revealing that customers tend to make fewer purchases on Sundays compared to other days.

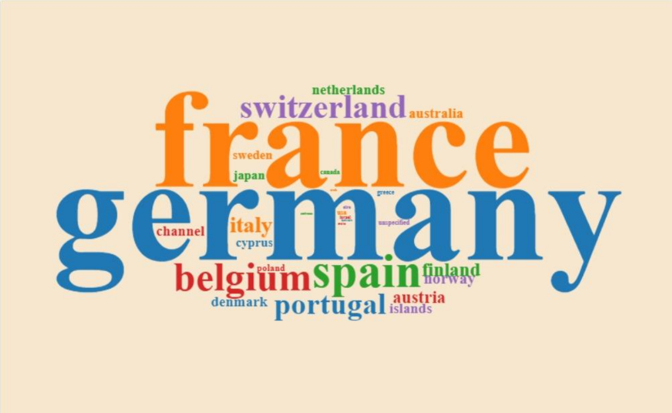
Revenue and Transactions by Hour Of Day





- These column charts provide insights into customer behavior and purchasing patterns by examining the revenue and number of transactions at different hours of the day
- Both charts show a peak around noon (12:00), indicating that the store experiences the highest level of customer activity and revenue generation during midday hours

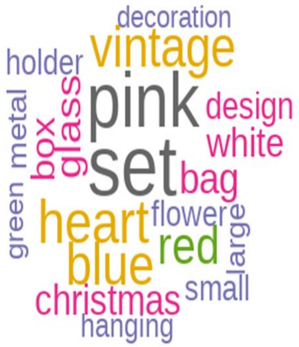
Wordcloud - Top Countries (excluding UK)

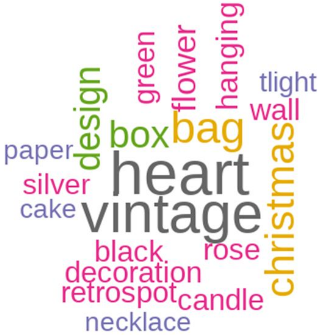


- Wordclouds offer a visually appealing and intuitive way to identify frequently occurring words or themes within textual data.
- In this case, the wordcloud visualizes the most frequent countries (excluding the United Kingdom) among the store's customer base.

Wordcloud - Product Descriptions

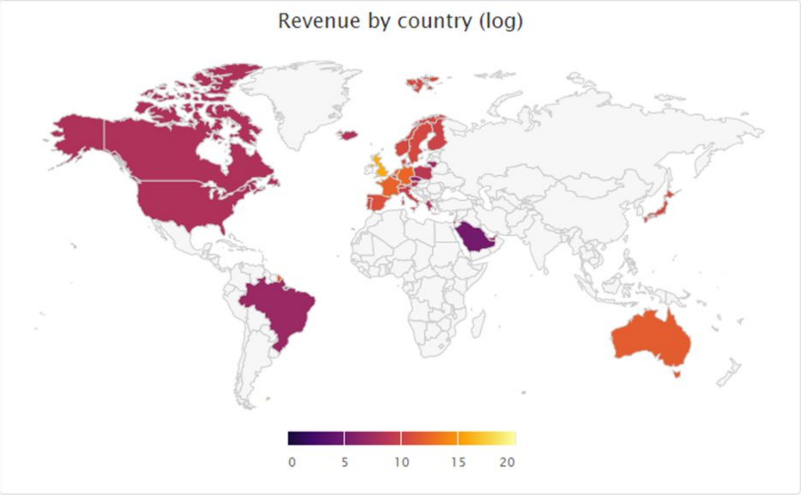
- The wordclouds for product descriptions provide insights into the types of products offered by the store.
- The first wordcloud displays the 20 most common words found in the product descriptions
- The second wordcloud removes these common words, focusing on more meaningful terms that could represent potential product categories, such as decorations, Christmas-related items, bags, vintage designs, and flowers/presents





World Map - Revenue by Country

- Colored by log of revenue (darker = higher)
- Top revenue sources: UK, parts of Europe, Australia, Japan
- Informs market prioritization, resource allocation, expansion strategies



CUSTOMER SEGMENTATION

Segmentation is the process of splitting a dataset into discrete groups according to shared traits like behavior, preferences, or demographics. As a part of segmentation we have performed RFM analysis and K-means clustering.

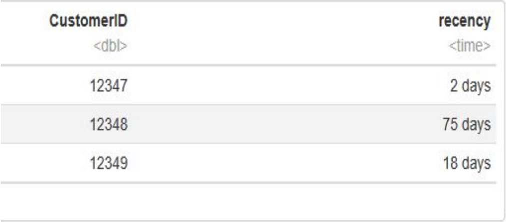
Recency, Frequency, Monetary (RFM) analysis uses monetary value, frequency, and recency to segment customers. Similar data points are grouped together using clustering techniques like K-means clustering.

RFM Feature Engineering

RFM analysis is a widely used technique for client identification and segmentation. RFM analysis is a technique to divide up customers into groups according to how they transact. It is essential, particularly for the e-commerce sector.

Recency: The term describes how many days a customer spent on their most recent purchase prior to the reference date. The likelihood of a client visiting a store increases with decreasing recency value.

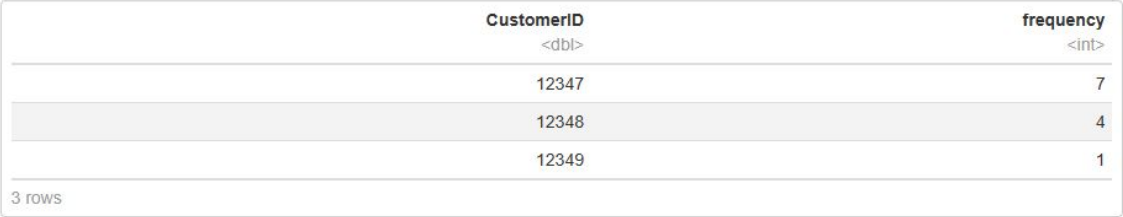
Here we can observe from the output that we have built our recency feature, which allows us to see how many days have passed since each particular customer's last purchase at the store.



CustomerID	recency
<dbl>	<time>
12347	2 days
12348	75 days
12349	18 days

RFM Feature Engineering

Frequency: It is the time interval between a customer's two following purchases. The frequency of consumer visits to the business increases with its worth.



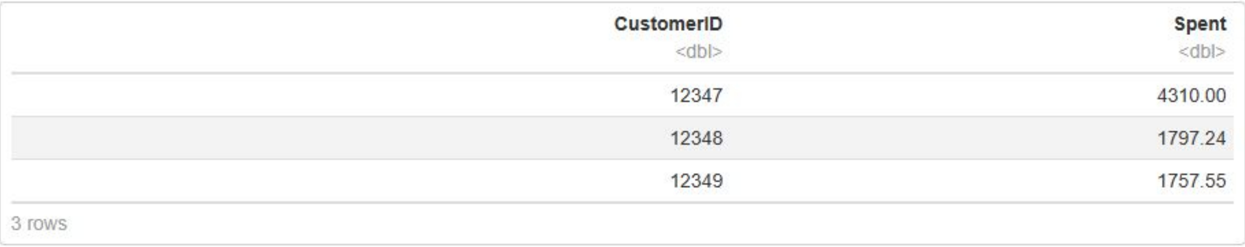
CustomerID	frequency
<dbl>	<int>
12347	7
12348	4
12349	1

3 rows

Here, we can observe the frequency of purchases for each customer. The first row indicates that customer with ID 12347 made a total of 7 purchases and so on.

RFM Feature Engineering

Monetary: This describes the total amount of money a client spends over a given time frame. The more the value, the greater the profit earned by the business.



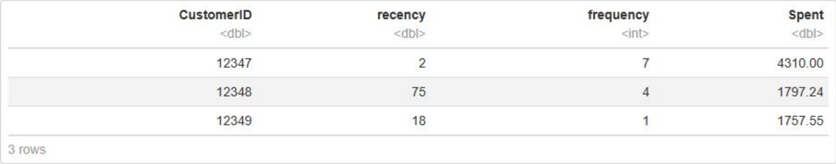
CustomerID <dbl>	Spent <dbl>
12347	4310.00
12348	1797.24
12349	1757.55

3 rows

Here we can observe that the monetary contains two variables and each row represents a customer and their “customerID” and the amount “Spent ”.

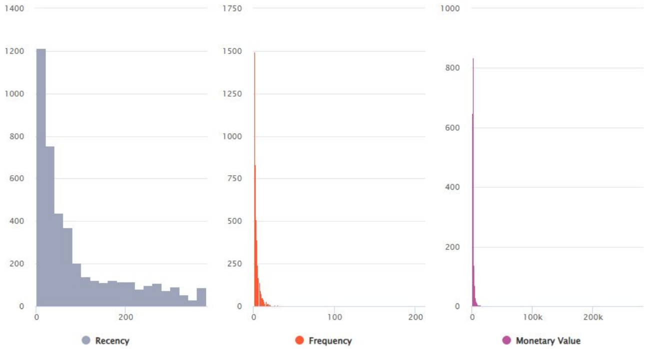
RFM Feature Engineering

RFM Complete Table: After completing our RFM engineering, we can combine the three data frames to create a comprehensive data frame that will be used for the store's customers' RFM clustering analysis. There is also a plot to depict the RFM analysis.



CustomerID <dbl>	recency <dbl>	frequency <int>	Spent <dbl>
12347	2	7	4310.00
12348	75	4	1797.24
12349	18	1	1757.55

3 rows



We can observe the various distributions that exist for the three RFM attributes from the charts.

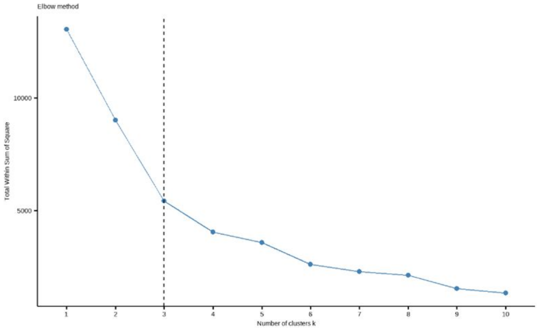
K-means Clustering

An iterative method of clustering is K-means clustering. The K-Means algorithm divides the client base for RFM values using the Euclidean distance metric. Each data point is repeatedly assigned to one of the K clusters based on feature similarity, with the number of groups K being predetermined.

We can bring out many features to the same scale by scaling the data. Scaling the data is the best scenario for the k-means technique.Finding the ideal number of clusters (k) for this data set of customers is necessary before we can begin the actual clustering. The most widely used method of counting the total amount of clusters are Elbow Method, Silhouette Method and Gap Statistic Method. The gap statistic method is a statistical approach, while the elbow and silhouette approaches are direct methods.

K-means Clustering

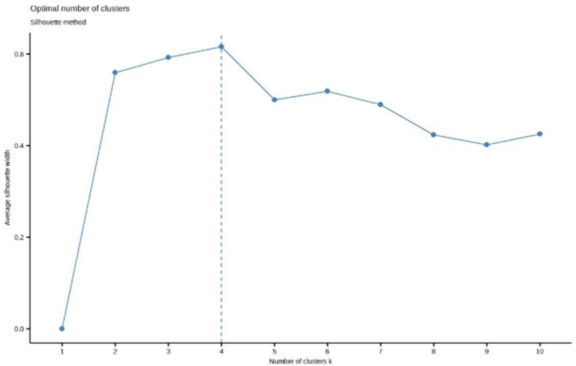
Elbow Method: The elbow method is a way for finding out how many clusters is the ideal number for K-means clustering. Plotting the within-cluster sum of squares (WCSS) versus the total number of clusters is needed to determine the "elbow"-shaped point at which the WCSS decreases at a slower rate.



Here we could utilize 3, 4, 5, or even 6 clusters depending on the slope and coming closer to zero, according to the Elbow method chart. But the way it functions better is subjective.

K-means Clustering

Silhouette Method: This method evaluates the quality of clustering by calculating the average distance between the clusters. For every data point, silhouette coefficients are computed. These coefficients are shown in silhouette plots for various values of k, or the number of clusters, which usually ranges from 1 to 10.

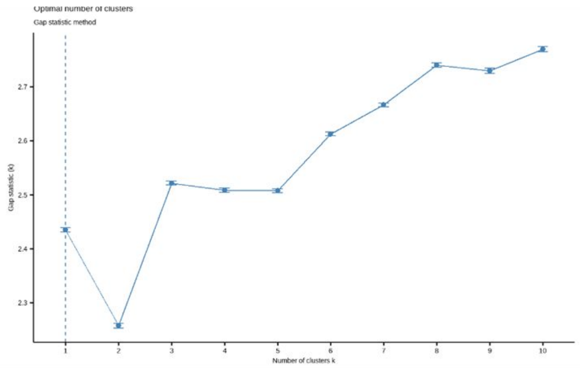


This plot shows that when the number of clusters, k equals 4, the highest average silhouette coefficient is obtained. The data may be best clustered into four different clusters when the silhouette coefficient peaks at k = 4.

K-means Clustering

Gap Statistic Method: The intra-cluster variation for various choices of k is compared with the expected intra-cluster variation under the null distribution using the gap statistic. The ideal value of k would be one that maximizes the gap statistic, but in real-world datasets with poorly defined clusters, it might be simpler to select the value of k that corresponds to the point at which the statistics rate of increase slows down.

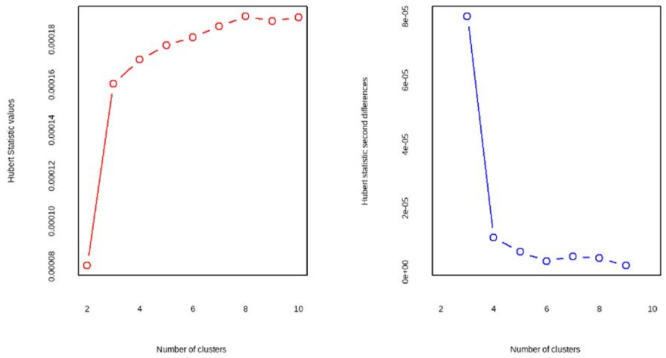
Here we can observe the trend in the graph, where rise in the gap statistic begins to slow down around about k = 3. This observation points to a key transition in the quality of the clustering, meaning that further cluster additions may not significantly enhance the overall structure .

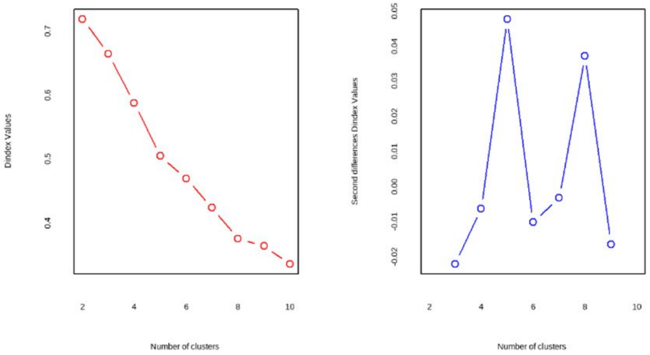


K-means Clustering

Calculating the ideal number of clusters: The subjective nature of selecting the ideal k number of clusters is one of the most discussed drawbacks of k-means clustering. We use the NbClust software in figuring out how many clusters would be most beneficial for our data.

Calculating the ideal number of clusters:

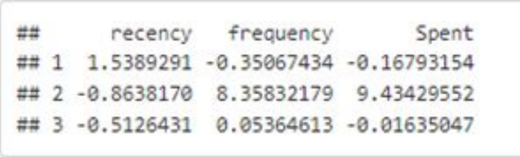




K-means Clustering

Hubert Index: It is a visual way for figuring out how many clusters there are. We look for a significant peak in the Hubert index plot, which shows increase in the measure's value.k = 3 is the optimal number as determined by the majority rule.

Clusters – Visualizing: Following the previously mentioned tests, we are able to determine that k = 3 appears to be the most appropriate ideal number of clusters for our data set.

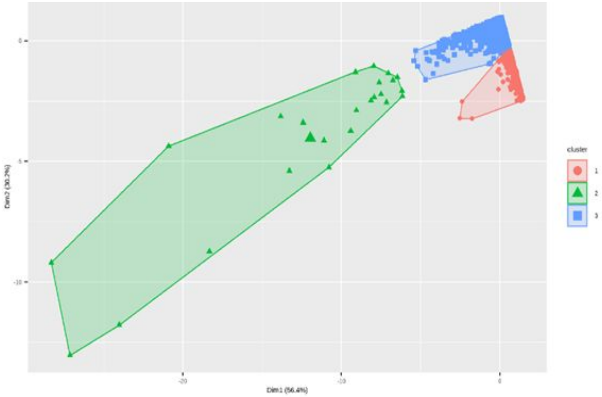


```
##      recency   frequency       Spent
## 1  1.5389291 -0.35067434 -0.16793154
## 2 -0.8638170  8.35832179  9.43429552
## 3 -0.5126431  0.05364613 -0.01635047
```

The data show that Cluster 3 has the most recent transactions, but what's really intriguing is that, when compared to the other two clusters, Cluster 2 has the largest transaction amount and frequency.

K-means Clustering

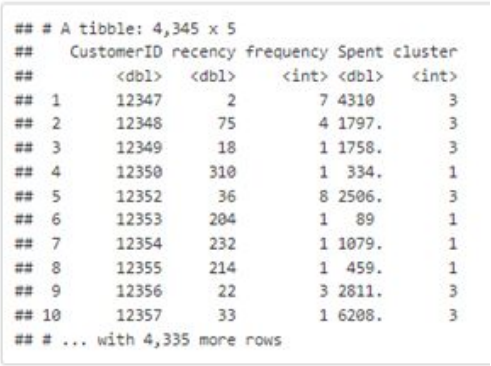
Clusters – Visualizing: Let's also see how many customers there are in each cluster. Below we get the relatively effectively the score (between_SS / total_SS) is 58.2%.



The total sum squares of the data points that are shared by the clusters make up this ratio. Also, there is Visualization of the outcomes of the clustering method below.

K-means Clustering

Clusters – Visualizing:

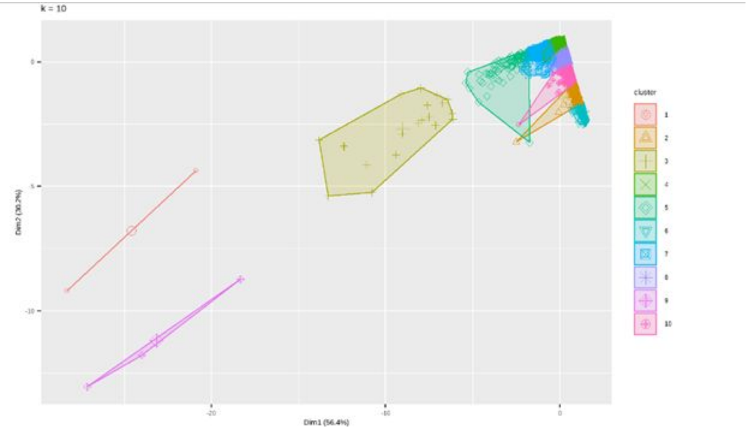


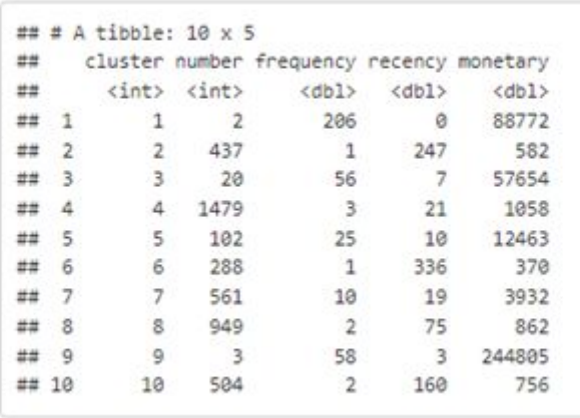
```
## # A tibble: 4,345 x 5
##    CustomerID recency frequency Spent cluster
##         <dbl>   <dbl>     <int> <dbl>   <int>
##  1      12347       2         7 4310        3
##  2      12348      75         4 1797.       3
##  3      12349      18         1 1758.       3
##  4      12350     310         1  334.       1
##  5      12352      36         8 2506.       3
##  6      12353     204         1   89        1
##  7      12354     232         1 1079.       1
##  8      12355     214         1  459.       1
##  9      12356      22         3 2811.       3
## 10      12357      33         1 6208.       3
## # ... with 4,335 more rows
```

Thus, it is evident that cluster 2 has produced the highest revenue by spending the most on store merchandise. We have a variety of groups in cluster 1 that may prove to be important future customers for the company because they either make large purchases or have recently done a transaction that opened up new possibilities.

K-means Clustering

Clusters – Visualizing: For K = 10: When we do the same process to k=10 as done for k=3, we check for the number of customers in each cluster and as anticipated, the ratio rises to 89.4% as the number of clusters grows. Below is the visualization of the outcomes of the clustering method.



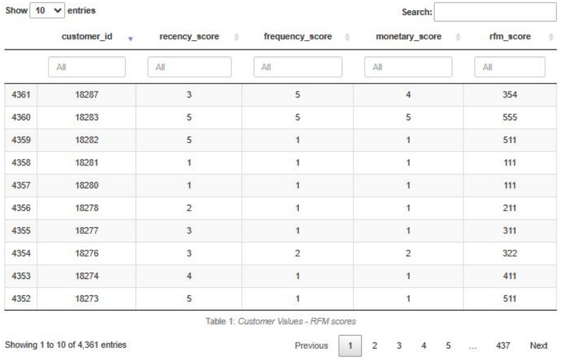


```
## # A tibble: 10 x 5
##    cluster number frequency recency monetary
##      <int>  <int>     <dbl>   <dbl>    <dbl>
## 1        1      2       206       0    88772
## 2        2    437         1     247      582
## 3        3     20        56       7    57654
## 4        4   1479         3      21     1058
## 5        5    102        25      10    12463
## 6        6    288         1     336      370
## 7        7    561        10      19     3932
## 8        8    949         2      75      862
## 9        9      3        58       3   244805
## 10      10    504         2     160      756
```

RFM Analysis

We calculated RFM scores for each customer using the rfm package in R and Quantified customer behavior.

Identifying Top 20% Customers:



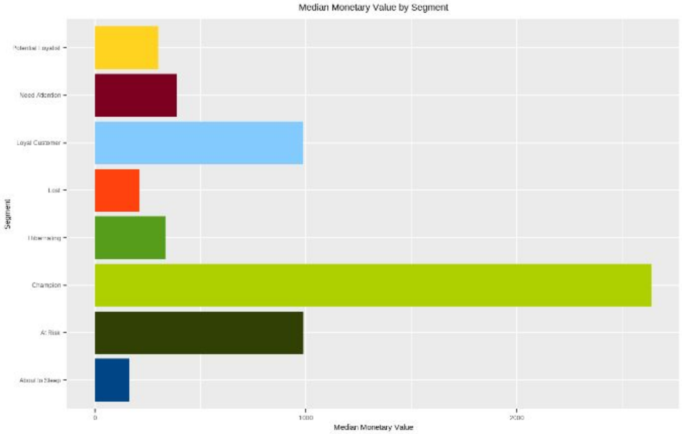
	customer_id	recency_score	frequency_score	monetary_score	rfm_score
	All	All	All	All	All
4361	18287	3	5	4	354
4360	18283	5	5	5	555
4359	18282	5	1	1	511
4358	18281	1	1	1	111
4357	18280	1	1	1	111
4356	18278	2	1	1	211
4355	18277	3	1	1	311
4354	18276	3	2	2	322
4353	18274	4	1	1	411
4352	18273	5	1	1	511

Table 1: *Customer Values - RFM scores*

Showing 1 to 10 of 4,361 entries

Applied Pareto principle (80/20 rule): 20% of customers contribute to 80% of revenue and Identified the top 20% of customers based on their RFM scores. These "vital few" customers are the most important for the business to retain.

Plot of the number of customers in each segment



- We notice that the median monetary value is primarily generated by our Champion Customers.
- To mitigate this risk, the store should concentrate on retaining these customers through targeted marketing campaigns, promotions, and perhaps offering discounts to strengthen their relationship.

Conclusion

In order to identify critical business aspects and segment customers into useful data that can be used to improve customer relationship management, we analyzed the online retail data in this project. More precisely, we used the k-means approach to identify three significant the customer clusters. In order to motivate action, we divided the customers into more focused groups and conducted an RFM analysis. Other information that might be utilized for a more thorough classification of customer such as age, gender, ethnicity and so on. The distinctive nature of the goods or services the organization offers will have impact on the strategy they choose and how their consumers are categorized and ultimately, have an effective future.

In conclusion, regardless of the outcome, we should always remember that clustering is a model before making any business decisions. Although it can typically produce value, it should always be regarded as such.

Executive Summary

- This project focuses on customer segmentation and predicting customer lifetime value (CLV) in the e-commerce industry.
- The objective is to enhance marketing strategies and improve customer relationship management.
- Customer segmentation and predicting customer lifetime value (CLV) are critical aspects of e-commerce strategy aimed at optimizing marketing efforts, improving customer retention, and driving business growth.
- As a part of this, Exploratory data analysis and customer segmentation has been done using the RFM feature engineering and K -means Clustering techniques which have given desirable details and representations.
- By addressing the research issues and implementing the recommended strategies, e-commerce businesses can gain deeper insights into their customer base and drive sustainable growth.

Key Research Issue

The main problem addressed is the requirement for precise CLV prediction models and efficient client segmentation in e-commerce. Understanding consumer behavior and value is necessary before allocating resources and putting tailored marketing into practice.

Findings:

1. Customer Segmentation: Employing machine learning techniques such as clustering algorithms (e.g., K-means, hierarchical clustering) is effective in grouping customers based on behavior, demographics, and purchasing patterns.

2. CLV(Customer Life Value) Prediction: Advanced predictive analytics models (e.g., RFM analysis, machine learning regressions) can forecast CLV accurately, considering historical data and customer attributes.

Future Work

➔ **Enhanced Segmentation Techniques:** Investigate more sophisticated segmentation methods incorporating real-time data and unstructured data sources (e.g., social media interactions).

➔ **Dynamic CLV Models:** Develop dynamic CLV models that adapt to changing customer behaviors and market trends.

➔ **Personalized Marketing Strategies:** Implement personalized marketing campaigns based on segmented customer groups and CLV predictions.

➔ **Customer Journey Analysis:** Explore the complete customer journey to refine segmentation and improve CLV estimations.

➔ By implementing these recommendations, e-commerce businesses can optimize marketing efforts, enhance customer satisfaction, and ultimately improve profitability.

References

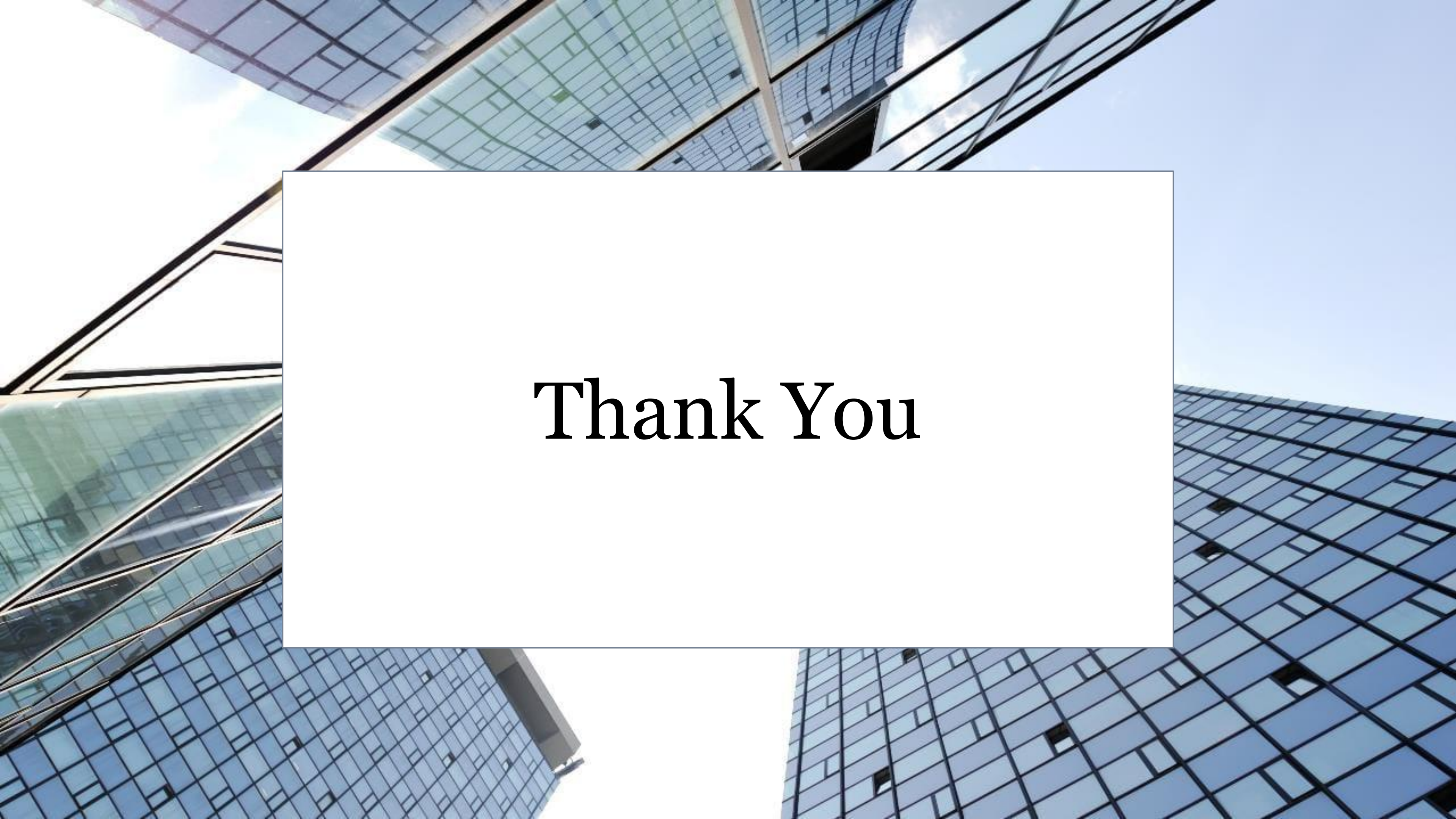
1. Gurram Venkata Sai Lakshmi Tejaswani et al, “Customer Segmentation using Machine Learning in R”, International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 06 | June 2021.

2. Nilay Jha, et al, “Customer Segmentation and Churn Prediction in Online Retail”, Part of the Lecture Notes in Computer Science book series (LNAI,volume\12109).

3. Mahmoud SalahEldin Kasem et al, “Customer Profiling, Segmentation, and Sales Prediction using AI in Direct Marketing” Multimedia Department , Assiut University, Egypt.

4. Smith, J., & Smith, A. (2017). Customer Segmentation in E-commerce: A Review of Techniques and Best Practices. Journal of Marketing Analytics, 5(2), 65-80.

5. Brown, P., & Langer, A. (2015). Customer Lifetime Value: A Review and Theoretical Framework. Journal of the Academy of Marketing Science, 43(3), 253-270.



Thank You