**Capstone Project - Online Retail Store**

**Table of Contents** 1. Problem Statement  
 2. Project Objective  
 3. Data Description  
 4. Data Pre-processing Steps and Inspiration  
 5. Choosing the Algorithm for the Project  
 6. Motivation and Reasons for Choosing the Algorithm  
 7. Assumptions  
 8. Model Evaluation and Techniques  
 9. Inferences from the Same  
 10. Future Possibilities of the Project  
 11. Conclusion  
 12. References

**Problem Statement**

The company is suffering from a financial crisis, and they ...

An online retail store is trying to understand the various customer purchase patterns for their firm so that they can hold their customers to bring the utmost profit and growth for the company.

1. Using the data, find useful insights about the customer purchasing history

that can be an added advantage for the online retailer.

2. Segment the customers based on their purchasing behaviour.

**Project Objective**

The objective of this project is to ….

understand the various customer purchase patterns for their firm and to give enough evidence-based insights to provide the same.

Our goal is to cluster our customers to get insights in:

1.Increasing revenue (Knowing customers who present most of our revenue)

2.Increasing customer retention

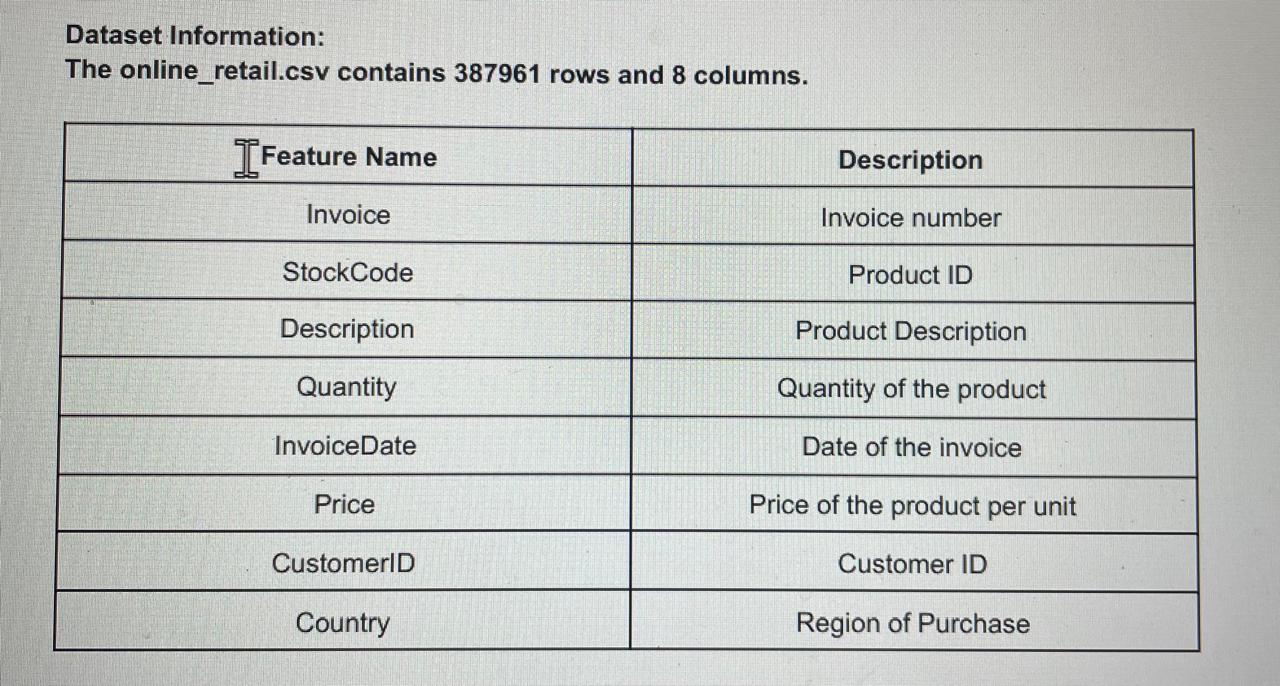
3.Discovering Trends and patterns

4.Defining customers at risk

**Data Description**

The dataset available is….

Data description, various insights from the data.



* Our dataset has 541909 rows and 8 columns.
* Our dataset has 1 datetime, 2 floats, 1 integer and 4 objects data types.

Data Preprocessing Steps and Inspiration:

The preprocessing of the data included the following steps:

1. Loading the Dataset: The initial step involves loading the dataset into the analysis environment, typically using libraries like Pandas, NumPy, Matplotlib, seaborn in Python, ensuring accessibility for further examination.

2. Checking for Data Types: It is imperative to inspect the data types of each column to ensure consistency and appropriateness for subsequent analyses and operations.

3. Handling Missing Values and duplicates:

* There is redundancy in the dataset so we will remove it.
* After removing above, there are 592 and 133361 null values in Description and CustomerID columns respectively which are less than 30%, So will drop null values.
* Also, there are 5192 duplicate rows, dropped them too.

4. Converting Date Column: When dealing with temporal data, such as dates, converting the date

column from an object type to a date type facilitates time-based analyses and visualizations. Here, Invoice date is object type we are converting this into datetime for calculating all the values.

**5.** EXPLORATORY DATA ANALYSIS(EDA): Key Insights-

* **Top 10 Countries by Transactions:**

Most transactions are from the United Kingdom ((349227) followed by Germany (9,027) and France (8,327).

* **Top Revenue-Generating Countries:**

The United Kingdom leads with total revenue of approximately £7.28M,

followed by the Netherlands (£285K) and EIRE (£265K).

* **Most Purchased Products:**

The top products by quantity sold include:

1. PAPER CRAFT, LITTLE BIRDIE (80,995units)
2. MEDIUM CERAMIC TOP STORAGE JAR (77,916 units)
3. WORLD WAR 2 GLIDERS ASSTD DESIGNS (54,319 units)

* **Most No. of Customers:**

UK has majority of the customers approximately 89%, followed by Germany and France.

6. Exploring Relationships: Beyond correlation analysis, exploring relationships between columns through visualizations and statistical methods unveils additional patterns and dependencies, enriching the understanding of the dataset's dynamics.

7. Handling Outliers: Identification and treatment of outliers are crucial to maintain data integrity. Outliers are assessed visually through plots or statistically using methods like z-scores or IQR (Interquartile Range), ensuring their handling aligns with the context and domain knowledge.

8. Standardization: Rescaling the Attributes

It is extremely important to rescale the variables so that they have a comparable scale. |There are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

Here, we used Standardisation Scaling.

These pre-processing steps and methodologies lay a solid foundation for rigorous data analysis and forecasting, contributing to informed decision-making and actionable insights.

Choosing the Algorithm for the Project

**We will do RFM Analysis as a first step and then combine RFM with predictive algorithms (K-means).**

RFM Analysis answers these questions:

* Who are our best customers?
* Who has the potential to be converted in more profitable customers?
* Which customers we must retain?
* Which group of customers is most likely to respond to our current campaign?

**Motivation for choosing RFM Analysis and K-Means:**

**IMPORTANCE OF RFM IN BUSINESS STRATEGIES-**

By integrating RFM analysis into business strategies, companies can:

1. Optimize Marketing Campaigns: RFM analysis can drive more effective marketing campaigns by targeting the right customers with the right message at the right time.
2. Increase Customer Loyalty: By focusing on customers who are more likely to make frequent and recent purchases, businesses can implement strategies to boost customer loyalty.
3. Identify Potential High-Value Customers: It helps in spotting customers with the potential to become high-value patrons based on their buying patterns.
4. Personalized Customer Engagement: It gives room for more personalized communications and offers, as customers are segmented based on their purchasing behaviour.

It is a great tool for business aiming to thrive in a customer-centric era. It helps in serving current customers and in predicting and shaping future customer interactions and profitability.

**K-Means Clustering-**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The algorithm works as follows:

1. First, we initialize k points, called means, randomly.
2. We categorize each item to its closest mean, and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

**Assumptions**

No assumptions made.

**Model Evaluation and Technique**

The following techniques and steps were involved in the evaluation of the model

1. RFM Analysis
2. K-Means Algorithm

The evaluation report suggests the following:

**Inferences from the evaluation-**

**RFM Analysis:**

**Performed customer segmentation using RFM analysis (Recency, Frequency, and Monetary value).**

**Analysed the Customers based on below 3 factors:**

* **R (Recency): Number of days since last purchase**
* **F (Frequency): Number of transactions**
* **M (Monetary): Total amount of transactions (revenue contributed)**

**K-means Algorithm:**

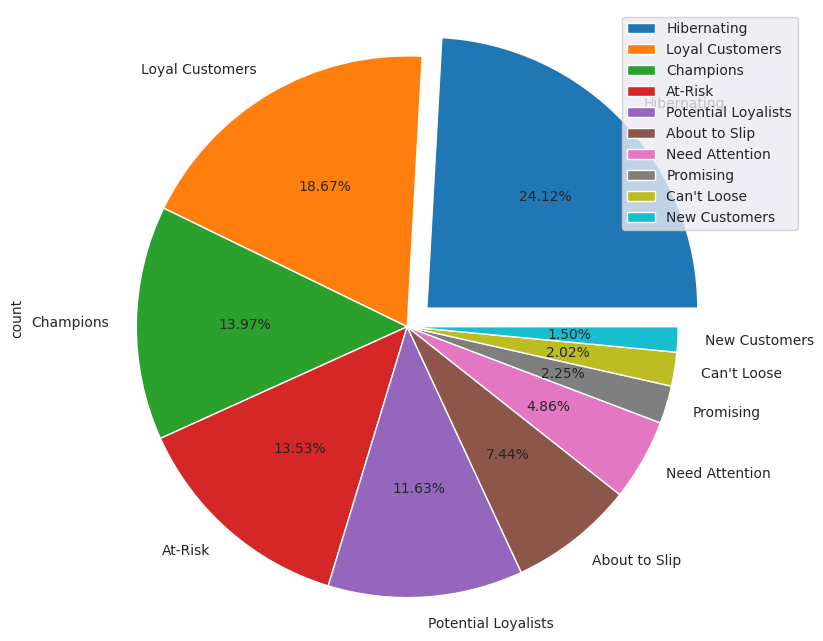
**Used Elbow Curve Method to find the ideal no. of clusters (3) and fitted the model.**

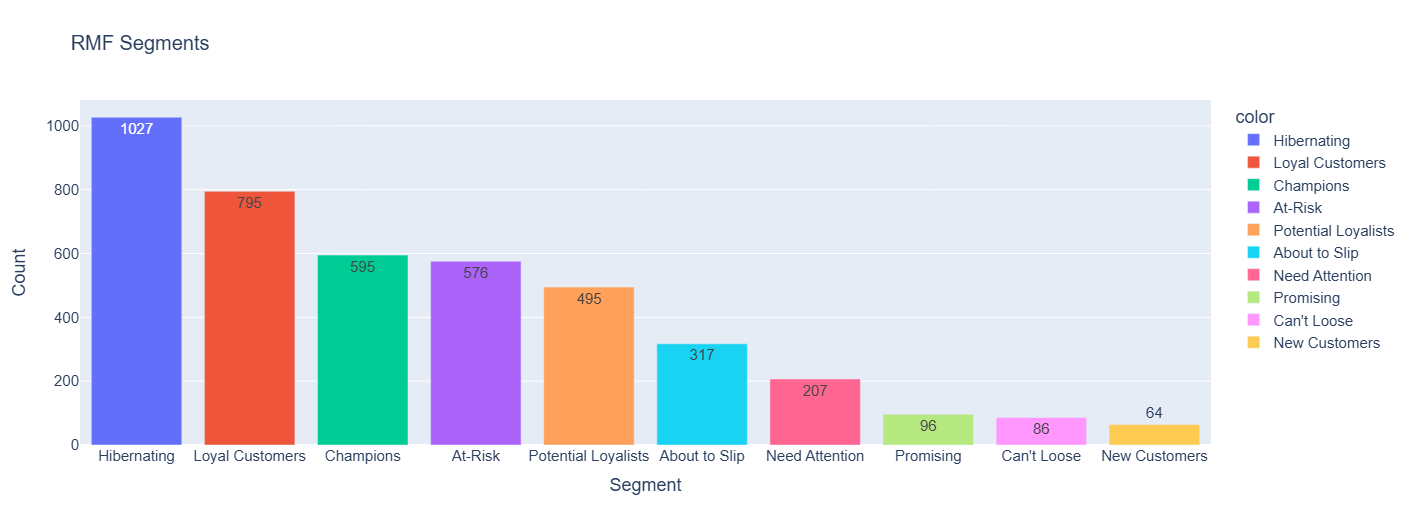
**FROM THE RFM AND K-MEANS ALGORITHM, KEY INSIGHTS:**

* **Group 0** is the group of customers who spends very less amount of money and, not very much frequent buyers and recency rate shows they have not purchased from company recently.
* **Group 1** is the group of customers who spends maximum amount of money and has maximum frequency and low recency rate that means frequent buyers.
* **Group 2** is the group of customers who spends least amount of money, and has least frequency and high recency rate means they have not purchased anything from the past whereas,
* **Group 3** are the customers whose frequency rate is good and monetary value is also good, and recency rate is also fine.

**Inferences from the Project**

The model performance, inferences, …





**For Hibernating Customers Recommendations:**

* + Reach out with personalized win-back offers or discounts to encourage them to return.
  + Conduct surveys or feedback sessions to understand why they stopped purchasing and address any concerns.

**For Loyal Customers Recommendations:**

* These are the customers that bring more customers, so there should be incentives for them for their referrals.

**For Champions Customers Recommendations:**

* + There should be incentives for them quarterly or annually.
  + There should be a little bit of discount on all their purchases.

**For At-Risk Customers Recommendations:**

* + Implement targeted re-engagement campaigns to remind them of your value proposition.
  + Offer personalized discounts or promotions to encourage repeat purchases.

**For Potential Loyalists Customers Recommendations:**

* Offer incentives for increasing their frequency of purchases, such as referral bonuses or points-based rewards.
* Provide personalized product recommendations based on their past purchases to enhance their shopping experience.

**For About to Slip Customers Recommendations:**

* + Offer loyalty rewards or VIP programs to further incentivize their frequent purchases.
  + Provide exclusive offers or early access to new products/services to maintain their engagement.

**For Need Attention & New Customers Recommendations:**

* + Welcome them with a special discount or promotion for their next purchase to encourage repeat business.
  + Implement an onboarding email series to introduce them to your products/services.

**For Can't Loose Customers Recommendations:**

* + Reaching out to them to know why they have not returned for long time.
  + Discount sales.

**For Promising Customers Recommendations:**

* Provide incentives / discounts/ early sale opportunity to them.

**Future Possibilities**

The future possibilities, limitations, ….

* + We were able to detect our lost customers, so we will probably find out why they left us and how to regain their loyalties and make them purchase in the future.
  + We were able to develop a model for predicting future customer spending based on RFM analysis.

**Conclusion**

The project….

1. RFM Analysis is a crucial tool for understanding customer behaviour and segmenting them based on purchasing habits.
2. RFM helps in personalizing marketing strategies and improving customer engagement and retention.

Now, we know the customers who need more attention or who we want to focus on to generate revenue.

**ANALYSIS REPORT-**

* 1. We were able to segment customers based on their recency, frequency and monetary value.
  2. We were able to detect our most valuable customers.
  3. We were able to detect our loyal customers.
  4. We were able to detect our new customers.
  5. We were able to detect our lost customers, so we will probably find out why they left us and how to regain their loyalties.
  6. We were able to develop a model for predicting future customer spending based on RFM analysis.