Aim:

This project aims to generate actionable insights by applying K-Means clustering to retail data. The goal is to conduct a modified RFM (Recency, Frequency, Monetary) analysis to better understand consumer behavior. The dataset, sourced from an online retail store, provides the foundation for identifying distinct customer segments—specifically, churned customers, loyal customers, and newly acquired ones. Based on these groupings, targeted marketing strategies can be developed and deployed to boost customer engagement, increase spending, and improve profit margins.

Introduction:

K-means Clustering is an unsupervised machine algorithm, used for clustering data into groups based on their similarity. It is often used in pattern recognition and market segmentation. It is different in a sense that it reveals patterns in group behavior rather than individual predictions. It allows profiling based on common patterns. For this process an ideal number of centroids (effectively the number of groups needed to capture variation in behaviors within data) is generated typically though a technique called elbow method. All the data points are assigned to the closted centroids based on their Euclidean distance. Next, the centroids are updated: the arithmetic mean of all points assigned to a cluster is calculated, and the centroid is moved to this new center. This process is iterative, with each step refining the centroid's position to better represent its cluster. The algorithm continues updating and reassigning until either the centroids stop moving significantly or a maximum number of iterations is reached—this is known as convergence.. In our analysis, each one of these clusters will contain customers with signature purchasing habits, i.e long time loyal customers, newly acquired, and old customers (churned).

Dataset:

Description:

The Dataset was acquired from Chen, D. (2015). *Online Retail* [Data set]. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/352/online+retail. The dataset consists of **541,910 entries** and includes the following columns: *Invoice Number*, *Description*, *Quantity*, *Invoice Date*, *Unit Price*, *CustomerID*, and *Country*. For the purposes of this analysis, only *Quantity*, *CustomerID*, and *Invoice Date* were utilized.

- *Description* refers to the specific item sold
- *Unit Price* indicates the cost of a single item
- *CustomerID* is a unique identifier assigned to each customer
- *Invoice Date* denotes the date and time when the transaction occurred

These selected features are sufficient for analyzing customer behavior patterns and segmenting consumers through clustering.

Data Cleaning:

The dataset contained missing values, particularly within the CustomerID column. The Pandas method .isnull().any() was used to check for missing values, and since it returned True, data cleaning was necessary. The .dropna() method was applied to remove rows containing missing values. While there are various ways to handle missing data (such as imputing with averages), this approach wasn't appropriate here. Columns like CustomerID and Quantity contain unique or specific information, and averaging would not produce meaningful or accurate results — it would introduce randomness rather than insight. Additionally, K-Means clustering does not handle missing or zero values well, which further justified the need for thorough cleaning. A total of 135,080 rows were dropped because they contained at least one missing value in the relevant columns

Data Preparation:

The data frame contained dates and times of when each invoice was generated. However, to perform analysis on it, conversions had to be made into the datetime format of the pandas library. In various analyses, this is usually enough but for K-means clustering the conversion had to be made into a format which would assign float or int values to all the data points including dates and times. For that purpose each data point was calculated as the difference of days from the default date of 12:00 am on Jan/1/1970. For instance a transaction that took place on Jan 3 1970 at noon would have been assigned a value of 2.5.

Since the main goal of the analysis is to understand patterns in consumer behavior the CustomerId were aggregated for the sum of all sales and invoice dates that showed their first purchase at the store all the way through the most recent purchases that they made. Furthermore the data frame includes the count which is corresponding to the number of times customers shopped at the store.

To Perform K-means Clustering, the data has to be normalized, so that values are relatively close to each other. Without normalizing (scaling) the bigger quantities will have a larger impact and that may not necessarily be meaningful. In our case the scaling was done on a 1 - 100 scale. The spread of values was so large doing normalizing it on a 10 scale felt too narrow and might lose information so instead scaling was done on a 1 - 100 scale.

Approach and Methodology

K-Means clustering is an unsupervised machine learning technique that groups data points based on similarity, using **Euclidean distance** as the measure of closeness. Each group of data points is called a **cluster**, and each cluster is represented by a **centroid**—essentially the average or center of the group. The centroid serves as the prototype of its cluster.

The algorithm begins by randomly selecting initial centroids. It then calculates the distance between each data point and each centroid, assigning each point to the nearest one. This process is called an **iteration**, and it repeats as centroids are updated based on the new cluster

assignments. The centroid is recalculated as the **arithmetic mean** of all the points assigned to it. Iterations continue until there is no significant change in the centroids' positions or until a pre-defined number of iterations is reached. This state is known as **convergence**. Once the algorithm has converged, the **characteristics of each centroid** can be analyzed to interpret the behavior of the points within each cluster.

Initially, K-Means was applied **without aggregating data by Customer ID**, which resulted in clusters that were too uniform and offered limited insights. To address this, the data was **engineered to aggregate by the customer**, creating a more meaningful structure for analysis. Each customer's transaction history was summarized with the following features:

- Quantity_sum: Total quantity of items purchased
- InvoiceDate_min: Date of their first recorded purchase
- InvoiceDate_max: Date of their most recent purchase
- InvoiceDate_count: Number of purchase transactions made

This aggregation allowed us to capture **customer-level behavior**, enabling deeper analysis and more actionable segmentation.

To determine the optimal number of clusters, the **elbow method** was used. This involves fitting K-Means with different values of k (the number of clusters) and measuring **inertia**, which reflects how tightly grouped the data points are within each cluster. Inertia is plotted against k, and the ideal cluster count is identified at the "elbow" point—where adding more clusters results in only marginal improvements. For this dataset, the elbow point occurred at **k = 3**.

To support interpretation and visualization, **Principal Component Analysis (PCA)** was applied alongside K-Means. PCA is a statistical technique used to reduce the dimensionality of data while preserving as much variance as possible. It transforms the original features into a smaller set of **orthogonal (uncorrelated)** components called **principal components**. In this case, PCA was used to project the data onto a **2-dimensional space**, making it easier to visualize the clusters.

The table below shows the contribution (or loadings) of each original feature to the first two principal components:

Feature	PC1	PC2
Quantity_sum	–0.00490	0.022455
InvoiceDate_min	0.869517	–0.49303
InvoiceDate_max	0.493835	0.868847
InvoiceDate_count	–0.00659	0.038992

- **PC1** (x-axis) is dominated by InvoiceDate_min, followed by InvoiceDate_max. This component primarily separates customers based on how long they've been active.
- **PC2** (y-axis) contrasts customers who **started early and stopped** with those who **started later and are still active**, reflecting a shift in customer behavior over time.

In this analysis, the K-Means algorithm was **implemented manually** rather than using the built-in function from scikit-learn. This approach provided greater flexibility and a deeper understanding of the mathematical operations involved. It also allowed for easier modification and tuning based on the structure of the dataset.

- A **function to generate initial centroids** was created by randomly sampling from the dataset.
- A **get_labels function** calculated the Euclidean distance between data points and centroids, assigning each point to the nearest cluster.
- Finally, a **new_centroids function** recalculated the mean position of each cluster, updating the centroids accordingly.

This iterative process formed the basis of the clustering logic and led to meaningful segmentation of customer behavior.

Results:

Based on the analysis, we identified **three clusters** (labeled 0, 1, and 2), each representing a distinct type of consumer behavior. The data revealed a clear structure in purchasing patterns,

as reflected in the plotted cluster graph, indicating that meaningful and actionable insights could be extracted.

	0	1	2
Quantity_sum	1.262336	1.854834	1.365533
InvoiceDate_min	13.896044	8.989647	71.002866
InvoiceDate_max	20.916808	88.575649	83.900723
InvoiceDate_count	1.280555	2.443249	1.497564

Cluster 0 (47.7% of customers)
Date range: **13.9 – 20.9** (on a 1–100 scaled timeline)*

This group represents customers who were active early on but have since stopped shopping at the store. They also have the **lowest total purchases**, as indicated by their Quantity_sum values. A likely reason for their departure could be the availability of more competitive products or more convenient alternatives elsewhere. To re-engage this group, a targeted **survey campaign**—offering incentives for completion—should be launched. Insights from these surveys can help identify the reasons for churn, allowing the store to make adjustments that could win these customers back.

Cluster 1 (34.4% of customers)
Date range: **8.99 – 88.58** (on the same scale)*

This cluster includes customers who have been **consistently loyal and frequent shoppers**. They began purchasing early and continue to shop regularly. Their high Quantity_sum and InvoiceDate_count indicate that they are among the **heaviest spenders**. For this group, a **reward or loyalty program** could further boost engagement—especially if tailored to products they frequently purchase. Analyzing the common traits and preferences of this group could also help shape strategies to encourage customers from other clusters to adopt similar behavior.

Cluster 2 (17.9% of customers)
Date range: **71.0 – 83.9***

This segment consists of **recently acquired customers** who are still in the early stages of forming loyalty to the store. Marketing strategies for this group should focus on **predicting their preferences**, offering **personalized discounts**, and fostering long-term engagement. Techniques such as **referral bonuses**, **automated reminders**, and **personalized interactions from store staff** can be useful in building strong relationships and encouraging repeat visits.

***Note:**

Due to the randomized nature of K-Means clustering (particularly in centroid initialization), the exact composition of clusters — including their sizes, percentages, and specific values — may vary slightly with each run. However, the **overall patterns and behavioral trends remain consistent**. If there are any discrepancies between the summary presented in this report and the cluster results in the Jupyter notebook, this variability is the reason.

Conclusion:

Given that the largest segment of customers belongs to the **churned group** and the smallest to the **newly acquired**, the store stands to benefit greatly from **investing in retention and re-engagement strategies**. Understanding the reasons behind customer loss and acting on them could lead to substantial improvements in customer lifetime value and overall profitability.

Cluster Sizes

Cluster	Number of Customers
0	2,084
2	1,506
1	782

While this analysis offers strong initial segmentation based on transaction behavior, future work could explore deeper behavioral insights—such as product-level preferences, sensitivity to marketing strategies, or patterns related to seasonal shopping trends. The current dataset is limited in that it does not include demographic or product-specific data, which may be essential for answering more targeted questions about why certain customers are retained and others are not. Expanding the analysis in this direction could offer even more valuable guidance for marketing and strategic planning.