CAPSTONE PROJECT- 2
CUSTOMER PURCHSE BEHAVIOR ANALYSIS

Table of Contents

1.Problem Statement

Improving Customer Experience and Sales through Understanding of Customer Purchase Patterns in Online Retail Store. The online retail store is facing challenges in retaining customers and increasing sales despite offering a wide range of products. There is a need to gain a deeper understanding of the customer purchase patterns to improve the customer experience and increase sales.

2.Project Objective

The objective of this project is to:

- Extract valuable insights from the customer purchasing data to inform and enhance the online retailer's sales strategy.
- Classify customers into distinct groups based on their purchasing behaviour to better understand and target each segment.

3.Data Description

This dataset online_retail.csv contains sales information for multiple retail stores across the country. **The data includes the following variables:**

1.Invoice: Invoice number (numeric)

2.StockCode: Code (Object)

3.Description: Product Description (Text)

4.Quantity: Quantity of the product (numeric)

5.InvoiceDate: Date of the invoice (DateTime)

6.Price: Price of the product per unit (numeric)

7.CustomerID: Customer ID (numeric)

8.Country: Region of Purchase (Quantitative)

The following insights from the data:

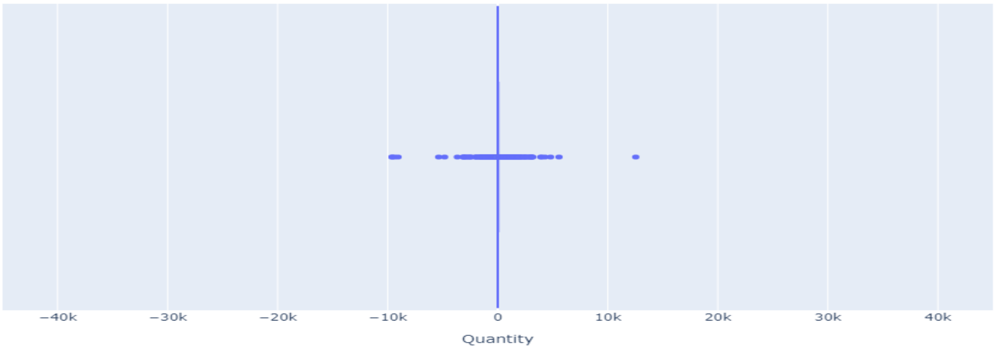
The dataset consists of 5,41,909 rows observations and 8 variables. There are 1,35,080 missing values and 5268 duplicate values in the data. The distribution of variables is as follows:

1.Invoice: 25,900 distinct values

2.StockCode: 4,070 distinct values with 85123A is the most selling item

3.Description: 4,223 distinct values

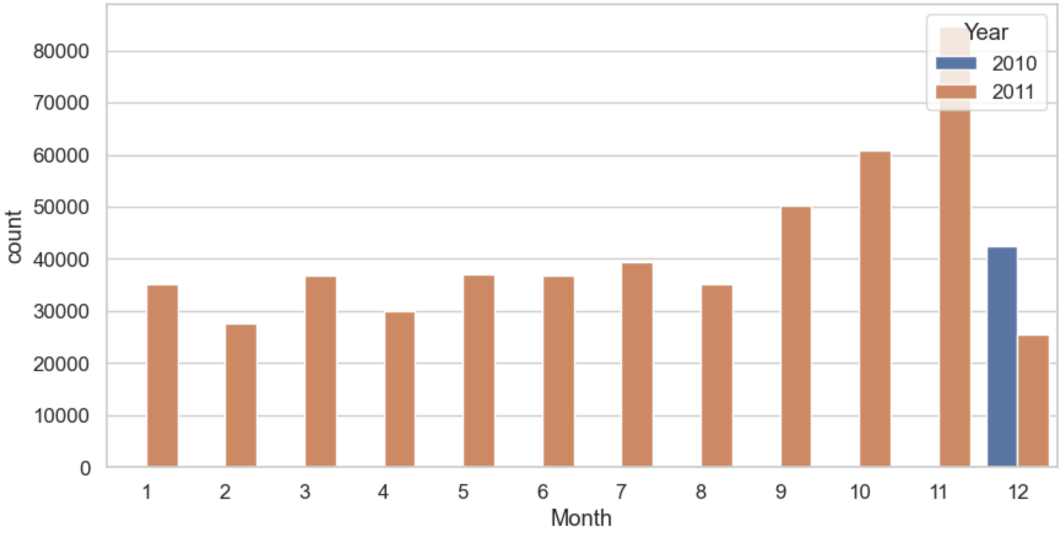
4.Quantity: 722 distinct values



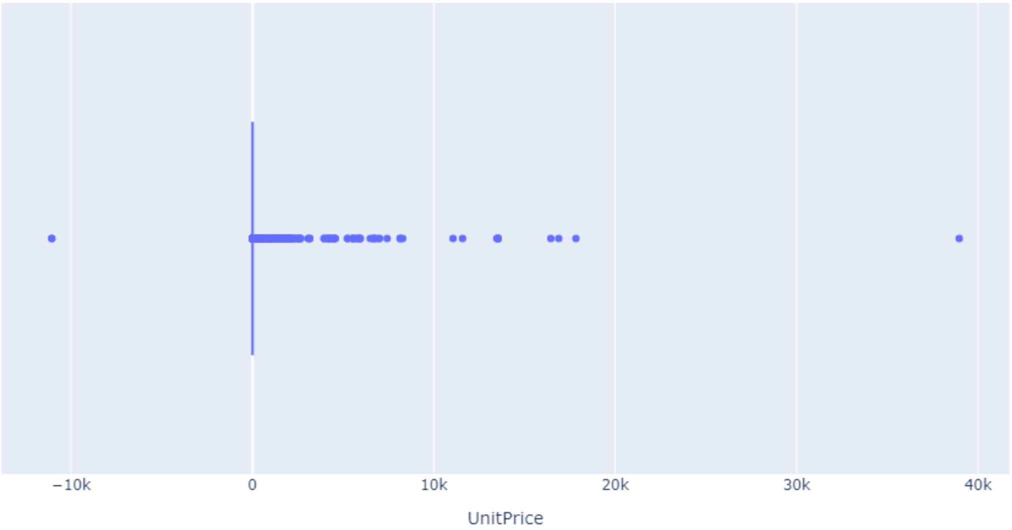
10624 values of InvoiceNo having quantity<=0

2 outliers (74125 and 80885)

5.InvoiceDate: (2010-12-01 08:26:00, 2011-12-09 12:50:00) - dec2010 to dec 2011 with unique 23260 dates



6.Price: 1630 distinct values



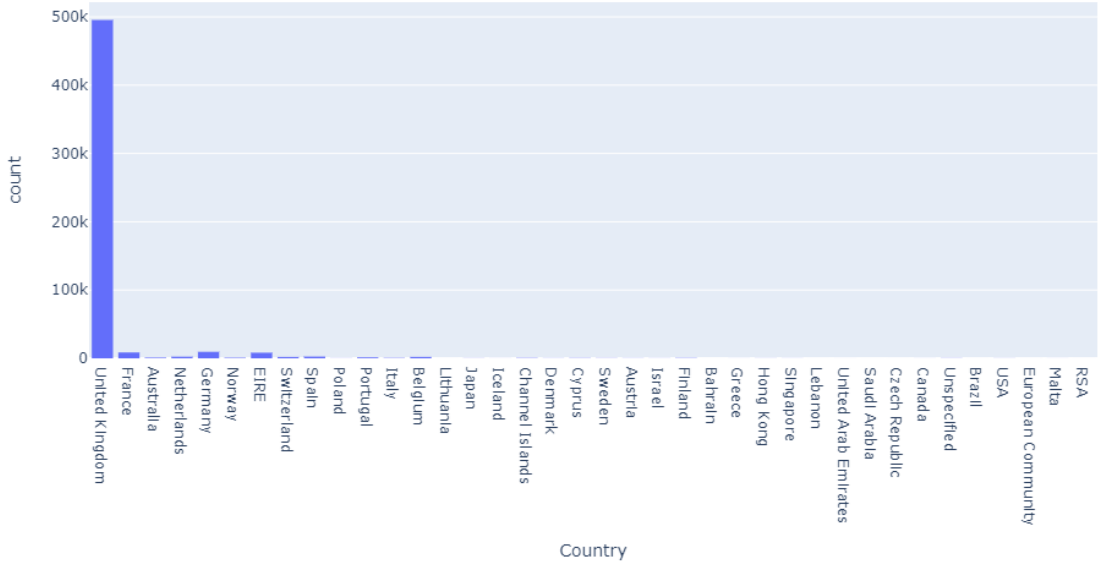
Having outliers: 38,970 and -11,062

2517 InvoiceNo having unit price is less than equal to zero

7.CustomerID: 4,372 distinct customers and having 135080 missing values

12346 starting Id and 18287 is ending ID

8.Country: 38 countries



91% of customers from UK, 2% from Germany, 2% from France

Country is having 0.03 with Quantity and 0.02 with Unit Price

This dataset provides valuable information for analysing sales trends and customer behaviour, and can be used to improve inventory management and marketing strategies.

4. Data Pre-processing Steps and Inspiration

The pre-processing of the data included the following steps:

1. Data Cleaning:
   - Handling Missing data:
     Check null values
     Drop null values as there are 135080 null values.
   - Handling Duplicate values:
     Check duplicated values
     Drop duplicate values -6435 duplicates are there.
   - To Keep data in which Quantity>0 and UnitPrice>0

2. Data Transformation:
   - Convert the datatype of the variables:
     Previously InvoiceDate variable is Object it is then converted as DateTime respectively.
     Converting CustomerID from int to String as a quantitative variable
   - Splitting Date variable:
     Date variable is spited into week, month and year variables
   - Adding new variable Year_Month with year month format.
   - Adding new column Amount which is product of Quantity and UnitPrice.
   - Creating a dataframe focussed on customerId and aggregated by Amount variable for customer segmentation
     df_amount is dataframe and amount_df.csv file is saved.
   - Creating another dataframe consists of customerID, Year_Month and Amount variable for customer behavior analysis
     Df_amount_month dataframe and amount_df_month.csv file saved

5.Choosing the Algorithm for the Project

K-means is an unsupervised learning algorithm that partitions a set of data points into K clusters, where K is specified by the user. The algorithm aims to minimize the sum of squared distances between each data point and the mean of the cluster it belongs to.

The steps in the algorithm include:

- Import the necessary libraries: pandas and KMeans from sklearn.cluster.
- Load the data into a Pandas dataframe.
- Train the KMeans clustering model with 3 clusters using the fit method.
- Calculate the distance between each data point and all 3 centroids using the fit_transform method.
- Assign the cluster labels to each data point and store the result in the 'labels' column of the dataframe.
- Replace the integer values in the 'labels' column with meaningful string labels using the replace method.

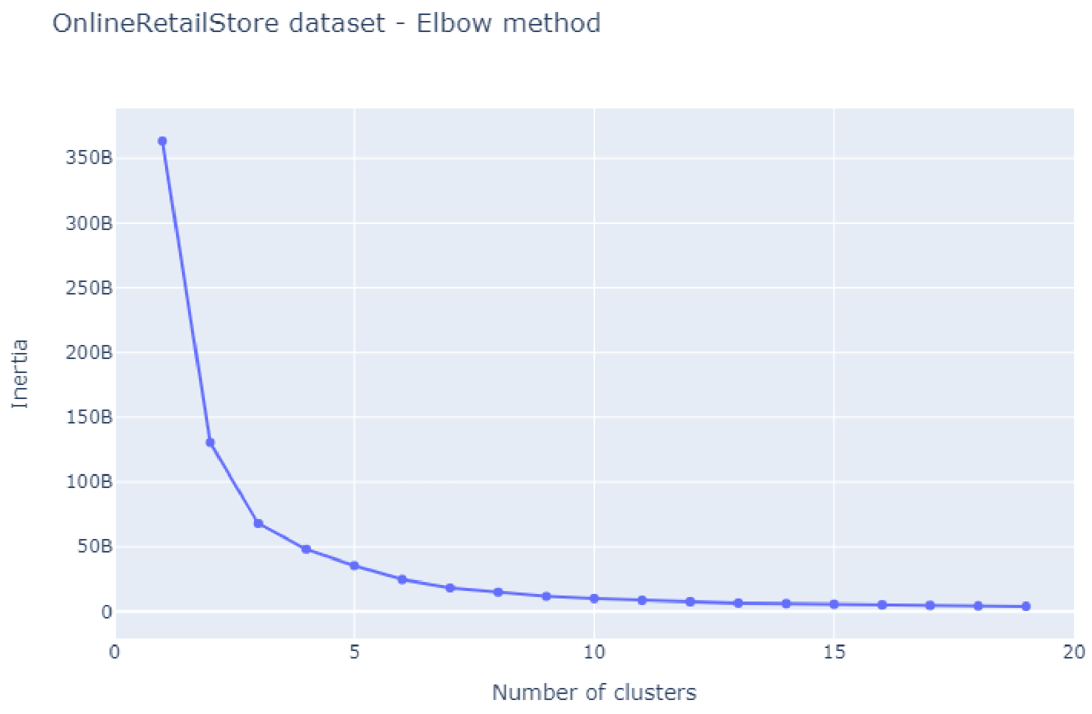
Here is a brief explanation of the KMeans algorithm that is used in this code:

- Initialize K cluster centroids, either randomly or using some heuristic.
- Assign each data point to the nearest centroid.
- Recalculate the centroids as the mean of all data points in the cluster.
- Repeat steps 2 and 3 until the centroids no longer change or a maximum number of iterations is reached.

6. Motivation and Reasons for Choosing the Algorithm

I have chosen K-Means Clustering algorithm rather than Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise, Gaussian Mixture Model (GMM) because it is a widely used algorithm for customer segmentation. It partitions the data into K clusters, where K is specified by the user or from elbow method. K-Means is fast and efficient.

The no of clusters(k) is decided from elbow method. It is a popular approach to determining the optimal number of clusters in a KMeans clustering algorithm. The basic idea behind the elbow method is to plot the value of the within-cluster sum of squares (WCSS) as a function of the number of clusters and choose the number of clusters at the "elbow point" of the plot. The "elbow point" is the point at which the WCSS starts to decrease at a slower rate, indicating that adding more clusters will not significantly improve the model's fit to the data.



From the above visualization k=3 provides the elbow point.

7. Assumptions

- The clusters are spherical
- The clusters are and equally sized

8. Model Evaluation and Techniques

1. Internal evaluation: Measures the quality of the clusters based on the similarity between the observations within each cluster. Examples of internal evaluation metrics are Silhouette Score, Calinski-Harabasz Index, etc.

   The score metrics:

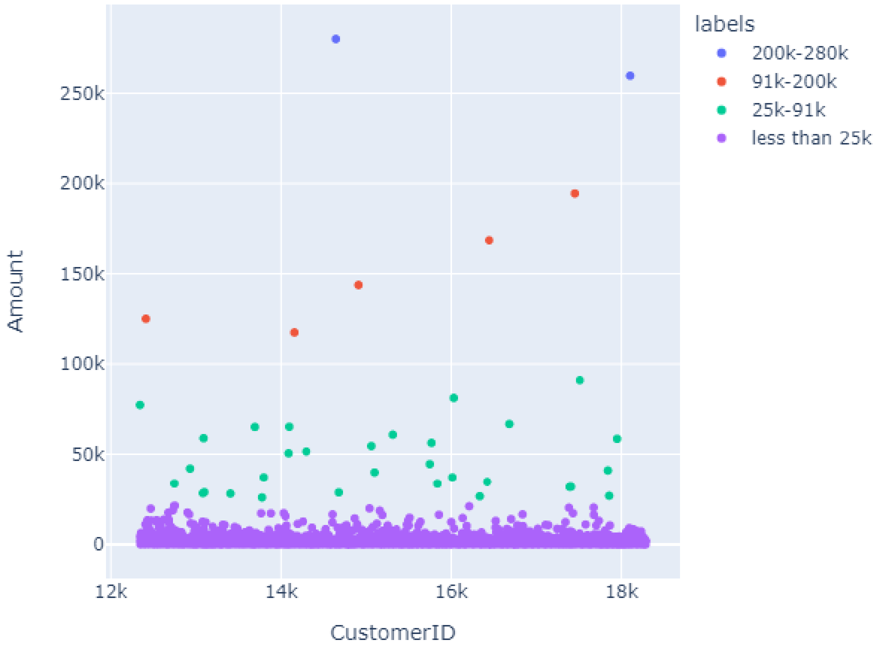
   - The Silhouette Score is 0.93255, which is high, indicating that the observations within each cluster are highly similar.
   - The Calinski-Harabasz Index is 9416.02, which is also high, indicating that there is a good separation between the clusters.
   - The Davies-Bouldin Index is 0.4282, which is relatively low, indicating that the clusters are well-separated.

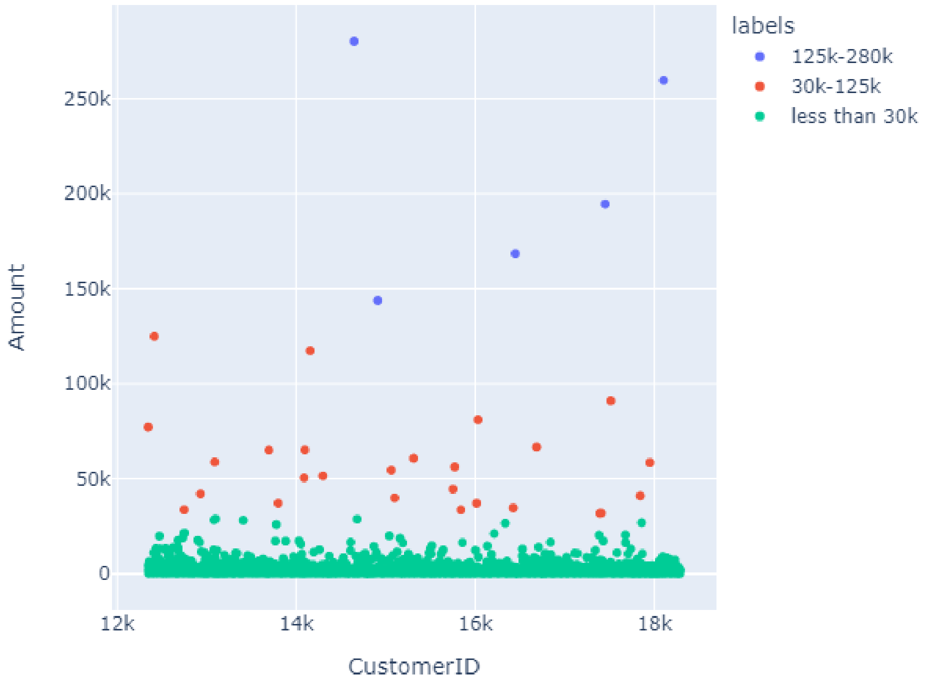
   The score metrics explanation:

   - The Silhouette Score of the model, which provides a measure of the quality of the clusters. The higher the score, the better the clusters are considered to be.
   - The Calinski-Harabasz Index measures the ratio of between-cluster variance to within-cluster variance. Higher values of the Calinski-Harabasz Index indicate that there is a good separation between the clusters, while lower values indicate that the clusters are poorly separated.
   - The Davies-Bouldin Index (DBI) is an internal evaluation metric used to assess the quality of a clustering model. The DBI measures the similarity between each pair of clusters, where a lower value indicates that the clusters are well-separated and a higher value indicates that the clusters are highly similar. A value close to zero indicates that the clustering solution is good, while a value close to N-1 indicates a poor clustering solution.

2. Visual inspection: Plotting the data and the clusters in a 2D or 3D space to visually inspect the quality of the clusters.

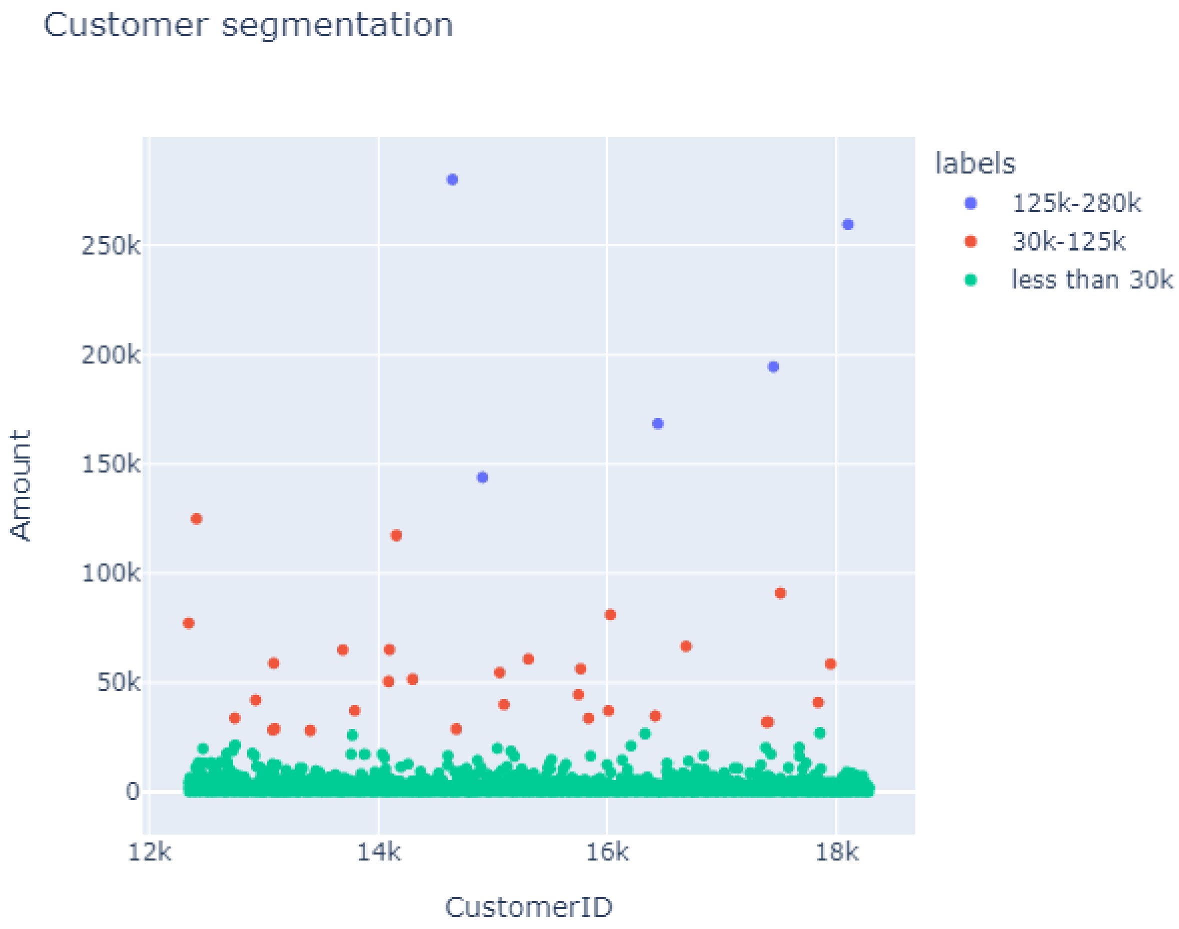
   The following visuals explain customer segmentation into 3 clusters and 4 clusters





As 3 cluster segmentation makes more sense with good performance metrics.

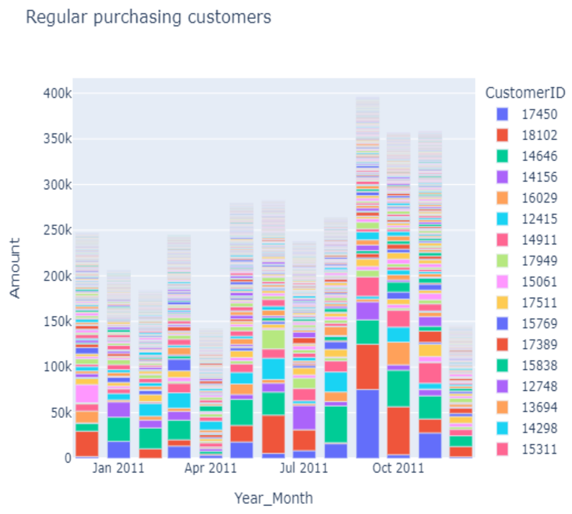
9. Inferences from the Same

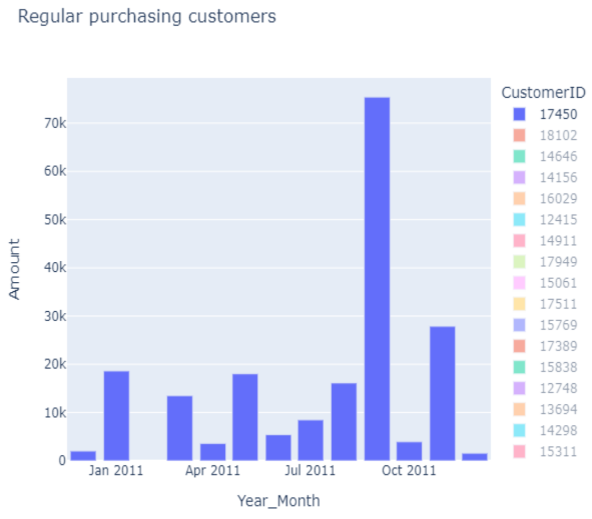


- The majority of customers fall into the lower end of the spending range, with sales less than 30k.
- A minority of customers, 6 in total, are driving the highest level of sales, with spending ranging from 125k to 280k.
- A slightly larger group of customers, 30 in total, are contributing to moderate sales, with spending between 30k to 125k.

This customer spending data can be used to create effective customer segments and target marketing strategies. Focusing efforts on the group of customers with lower spending levels can be used to tailor product offerings, promotional campaigns, and overall customer experience to better meet the needs of each segment. By effectively targeting each group, businesses can drive sales growth and improve customer loyalty.

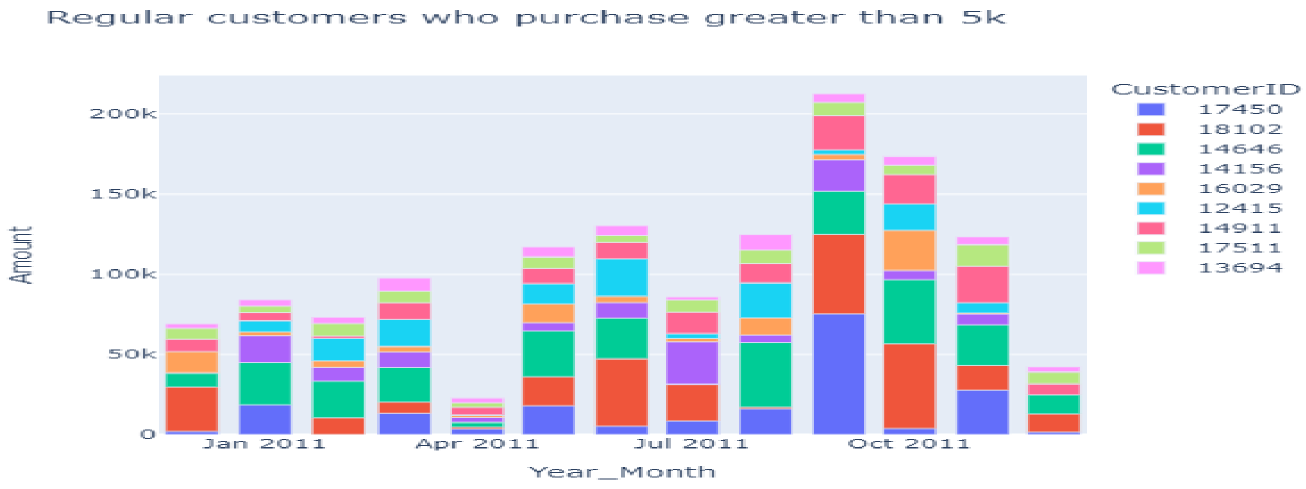
Interactive visualization is available in customersegmentation.html page





From regular purchasing customers visualization:

- Regular customers play a crucial role in the success and stability of the business by providing a consistent source of revenue. Their minimum sales being 150k and the peak sales crossing 400k in the best month.
- Personalized rewards and incentives can enhance customer loyalty and increase the lifetime value of regular customers, particularly those who make purchases over 1,000 or 5,000.



This visualisation is of regular customers who purchase greater than 5k

- These customers are most promising and alone purchase half of the sales, making their understanding and retention a high priority.
- Understanding their purchasing habits allows the business to make informed decisions on inventory and promotions.
- January, February, and April are the low sales months for the business, but they present an opportunity to improve revenue through targeted promotions, limited-time offers, and other sales driving techniques.

10. Future Possibilities of the Project

The future possibilities:

- Customer Retention: The data and insights from this project can be used to create targeted retention strategies for both regular customers and those with lower spending levels. By understanding the drivers of customer behavior, the business can take steps to improve customer loyalty and reduce churn.
- Marketing Campaigns: The data from this project can inform the development of more effective marketing campaigns, including personalized email and direct mail campaigns, as well as targeted social media advertising.
- Collaboration with Other Business Units: The insights from this project can be shared with other business units, such as product development and customer service, to inform and improve their efforts. By working together, different departments can create a cohesive and effective customer experience.

These are just a few potential directions for the project, and there may be others based on the specific goals and needs of the business. By continuing to analyze and understand customer spending patterns, businesses can make informed decisions that drive sales growth and improve customer loyalty.

11. Conclusion

The analysis of customer spending patterns has provided valuable insights into the behaviour of a business's customers. The majority of customers tend to have lower spending levels, while a small minority drives the highest level of sales. By identifying and targeting specific customer segments, businesses can improve their sales growth and increase customer loyalty through personalized marketing strategies and product offerings.

Regular customers play a crucial role in the success and stability of the business by providing a consistent source of revenue. Their retention and understanding should be a top priority for the business, as they drive half of the sales. Offering personalized rewards and incentives can enhance their loyalty and increase their lifetime value to the business. Additionally, the low sales months of January, February, and April present an opportunity for the business to improve its revenue through targeted promotions and limited-time offers. By understanding the spending patterns of its customers and utilizing this information effectively, a business can make informed decisions that will lead to its success.

12. References

1. Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." Journal of Relationship Marketing 6.3-4 (2008): 9-39.
2. Kansal, Tushar, et al. "Customer segmentation using K-means clustering." 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS). IEEE, 2018.
3. Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." Sustainability 14.12 (2022): 7243.
4. Syakur, M. A., et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." IOP conference series: materials science and engineering. Vol. 336. IOP Publishing, 2018.
5. Sial, Ali Hassan, Syed Yahya Shah Rashdi, and Abdul Hafeez Khan. "Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python." International Journal 10.1 (2021).
6. Podo, Luca, and Paola Velardi. "Plotly. plus, an Improved Dataset for Visualization Recommendation." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.