

**Unsupervised learning Project on**

**Customer Segmentation**

**Project performed by**

**Rahul Deshmukh**



**Table of Contents**

* Summary............................................................................1)
* Problem statement..............................................................2)
* Technical Work...................................................................3)

1. Library used.
2. Function and Method used.

* Step involved………………………………............................5)
* Data Visualization………………………………………………6)
* Modeling.
* Procedure performed and Observations.............................6)
* Challenges.

**Summary:**

Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviors, so you can market to those customers more effectively.

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

* Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).
* Select the best communication channel for the segment, which might be email, social media posts, radio advertising, or another approach, depending on the segment.
* Identify ways to improve products or new product or service opportunities.
* Establish better customer relationships.
* Test pricing options.
* Focus on the most profitable customers.
* Improve customer service.
* Up sell and cross-sell other products and services.

**Problem Statement**

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

**Technical Work**

* ***Library used***
* ***Functions and methods used.***

**Library used**

* **Pandas** This library is used in data analysis and manipulation and importing files.
* **NumPy** which stands for Numerical Python is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. NumPy is a Python package. It stands for 'Numerical Python'.
* **Matplotlib** Used infor data visualization, graphs & plotting.
* **Seaborn** Used inmaking data visualization in a more colorful and meaningful way.
* **Sklearn:** useful and robust library for machine learning in Python.

**Common functions and methods used**

**pd.read\_csv ():** used in to import data in CSV format. This function has a number of arguments, but the only essential argument is file, which specifies the location and filename.

**head ():** Used in checking the first rows of the dataset.

**tail ():** Used in fetching the last rows of the dataset.

**shape ():** Used in getting the total no. Of columns and rows, a dataset contains.

**Type ():** Used in knowing the data structure of the dataset.

**df.dtypes:** Used in understanding the data types of the dataset.

**Is Null():** used in checking if any null values are available.

**Sum ():** used in giving sum of the values.

**value\_counts() :** Used in providing counts of particular values present in data.

**sort\_values() :**function sorts a data frame in Ascending or Descending order of passed Column. It's different than the sorted Python function since it cannot sort a data frame and particular column cannot be selected.

**describe ()** : This method is used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

**unique ():** function is used to find the unique elements of an array.

**GroupBy:** It allows you to split your data into separate groups to perform computations for better analysis

**Sum():** The Python sum() function calculates the total of all numerical values in an iterable.

**Mean():**It returns mean of the data set passed as parameters. Arithmetic mean is the sum of data divided by the number of data-points. It is a measure of the central location of data in a set of values which vary in range.

**Info():** The info() method prints information about the DataFrame. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

**astype():** The astype() function is used to cast a pandas object to a specified data type.

.

**Steps involved:**

* **Setting directory/path.**
* **Loading dataset.**
* **Data Pre-Processing.**
* **Data Visualizations.**
* **Modeling.**
* **Finding Optimal K.**

**Setting directory/path**

Before we proceed with python to understand the data, it is important to let python understand first that where our data available is so that, we can perform the operation on data using python. The setting directory is similar to that and helps in understanding the presence of the data.

**os.chdir(your path)**

**Loading dataset**

Once the directory is set now we need to load the dataset which is commonly done using:

**Pd.read\_csv(‘file\_name’)**

**Data pre-processing**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

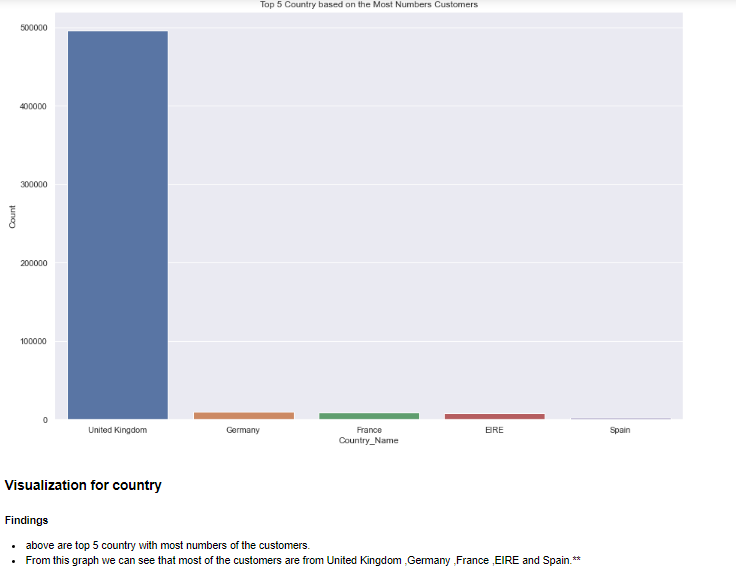
A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

**Data Pre-processing** involves in below:

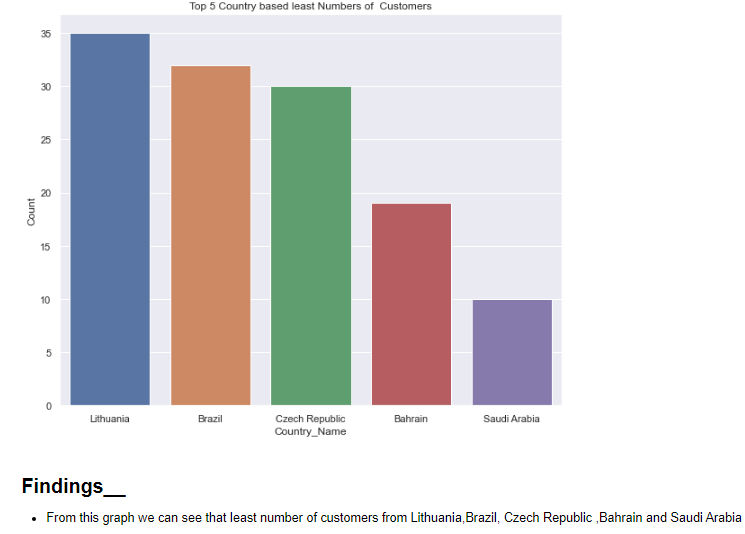
* **Getting the dataset.**
* **Importing libraries.**
* **Importing datasets.**
* **Finding Missing Data and treating those.**
* **Findings duplicates and treating those.**
* **Encoding Categorical Data if any.**
* **Feature scaling and feature engineering and feature selection.**
* **Removing outliers if any.**
* **Data exploratory analysis. etc**

**Data Visualization**

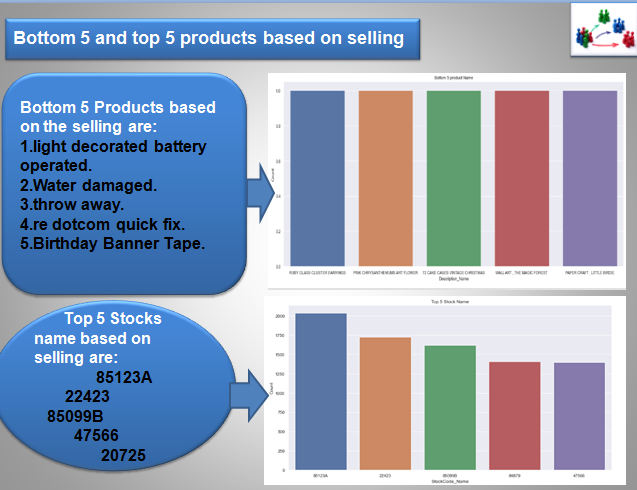
This step of visualization involves checking the data in pictorial form. Apart from this, visualization helps in easy understanding of the data and representation of most of the data information in rightly and quickly.



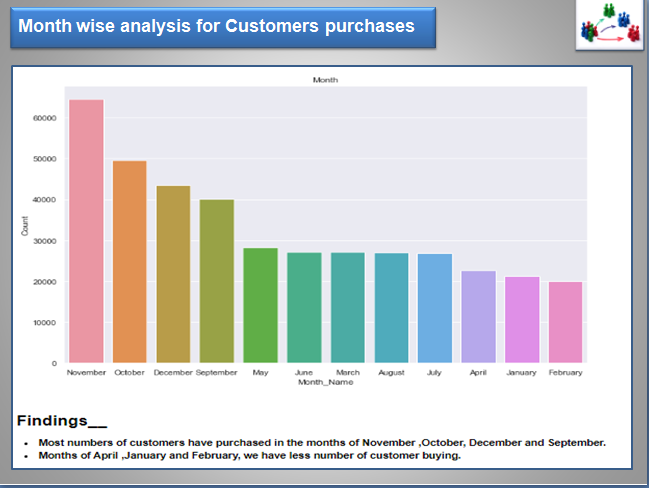
**This above pictorial representation is for top 05 countries based on most numbers of customers.**

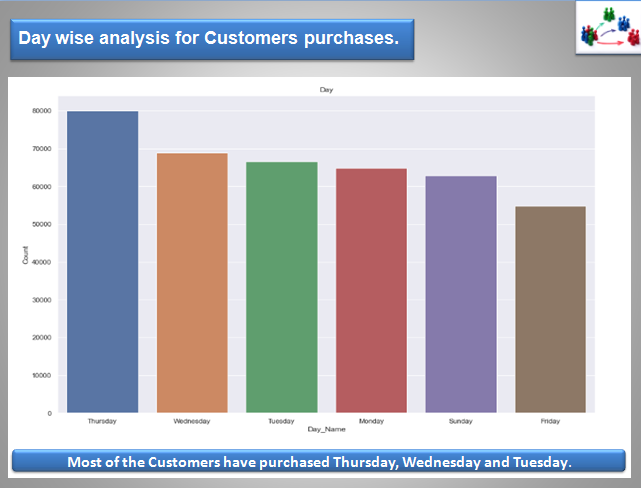


**This above pictorial representation is telling that we have least numbers of the customers from Lithuania,Brazil,Czech Republic, Bahrain and Saudi Arabia.**



**This above pictorial representation is telling the Bottom 5 product and top 5 stocks based on selling.**

****



**Modelling**

Before we proceed with model building, it is important that we have the right amount of data and we understand the features/variables very well. Almost every model takes the numerical data for further processing in the model part so it is important to convert the entire data into numerical columns so that we can process it further.

In case it requires suitable feature engineering or feature selection we should perform these earlier in the process before we actually start with modeling or training models.

 Next in this process, we would need some of the machine learning libraries and suitable models for the prediction After fitting the data look for the scores if we need to improve the scores, look for the scope of improvement if required.

**Models used on data**

RFM model

K-mean clustering.

Hierarchical Clustering.

**Method used for Optimal-K-Finding**

Elbow Method.

Silhouette score.

Dendrogram.

**Observations/Inferences/Conclusion:**

**1. Three Clusters (Customer Segments):**

Carefully examining the three-cluster classification, I observe following groups of customers:

**High value customer:**

'Cluster 2' is the high value customer segment for the online retails store as the customers in this group place the highest value orders with a very high relative frequency than other members. They are also the ones who have transacted the most recently.

**Medium value customer:**

'Cluster 0' appears to be the medium valued customer segment. These customers place an order of a considerable amount, though not as much as high valued customers, but still quite higher than low valued customers. Also, their orders are relatively more frequent than the lowest value segment.

**Low value customer:**

It is quite evident that 'Cluster 1' has customers who rarely shop and when they order, their orders are pretty low valued. Apart from the numbers, the visualization of clusters in Silhouette Analysis show that all three customer segments are quite distinct with very less overlap between them. The general trend resonated in these 3 clusters is that high monetary value is correlated with high frequency of orders and more recent ones.

**2. Five Clusters (Customer Segments):**

*In five clusters, we find the following customer segments:*

**Overall high valued customers:**

'Cluster 0' is the typical high value customer who has shopped recently and shops regularly for high value orders.

**High monetary value but less frequent:**

'Cluster 1' represents a peculiar customer segment who place quite a high valued order but do not do so frequently or have not done much recently. But these customers do hold a lot of promise if targeted to improve sales.

**Medium value - low frequency - recent customers:**

The customers from 'Cluster 4' have recently placed medium valued orders but do not do so frequently.

**Medium value - low frequency - older customers:**

The customers from 'Cluster 3' happen to place medium valued orders quite a long time ago and they do not do so frequently.

**Low valued customers:**

'Cluster 2' is the segment of customers who have not shopped in the longest time, nor do they shop frequently and their orders are of the lowest values. The visualization of clusters in Silhouette Analysis shows some overlap between the customer segments. However, the dataset does not distinguish between wholesale and retail customers, it is quite likely that high value frequent clients are the wholesale dealers and medium/ low valued ones are individual retail purchasers.

Summary of Conclusion:

The customer segments thus deduced can be very useful in targeted marketing, scouting for new customers and ultimately revenue growth. After knowing the types of customers, it depends upon the retailer policy whether to chase the high value customers and offer them better service and discounts or try and encourage low/ medium value customers to shop more frequently or of higher monetary values.

Summary:

This study started with importing dataset, analyzing dataset after this I have done preprocessing, I have checked for the null values as our dataset contains many null values in Customer id feature and we have to segment the customers, without customer id we are unable to segment customers therefore I have removed all the rows without Customer id.

After that I have done some exploratory data analysis (EDA) I came to know about top customers, Worst customers, periodical purchasing stats, most revenue generated weekday, purchase stats of country, top and lease purchasing country, top sold product, most revenue generated product, Customer stats, etc.

After that I have done some feature engineering to build RFM model (recency, frequency and monetary value) I have extracted and analyzed RFM score then I have created customer segment in 3 category bronze, silver and gold.

After that I have done data preprocessing for clustering with the help of log transformation, I have reduced Skewness of data then I have scaled data, after scaling I have extracted Silhouette Score Based on the inertia and silhouette score, I came to know that optimal number of clusters is 3.

Then I have implemented Kmeans clustering and plotted different graphs to visualize clusters. after that I have merged cluster column to data and used classification model for prediction. I have used Logistic Regression, Random Forest Classifier and XGBoost and done evaluation of it, my Optimal model was Random Forest as I got train Accuracy of 1.00 and test accuracy of 0.98 with it.

**Challenges.**

Some of the challenges I have faced in this project are mentioned below:

Null handling for Description column and Customer ID.

Looking for few values which were negative like in Total Price.

Applying Log transformation would be right or not.

Handling skewed dataset.

Binning of quintile for the customers based on 1,2,3,4 score.

Getting Silhouette score for every sample.

Finding optimal “K”.