

**Unsupervised learning Project on**

**Customer Segmentation**

**Project performed by**

**Sunil Kumar**



**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.
* **os** Used in provides functions for creating and removing a directory folder.

**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.

**%cd :** magical command finds present working directory.

**os.chdir():** Changes your directory to your data directory.

**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.

**IsNull():** 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.

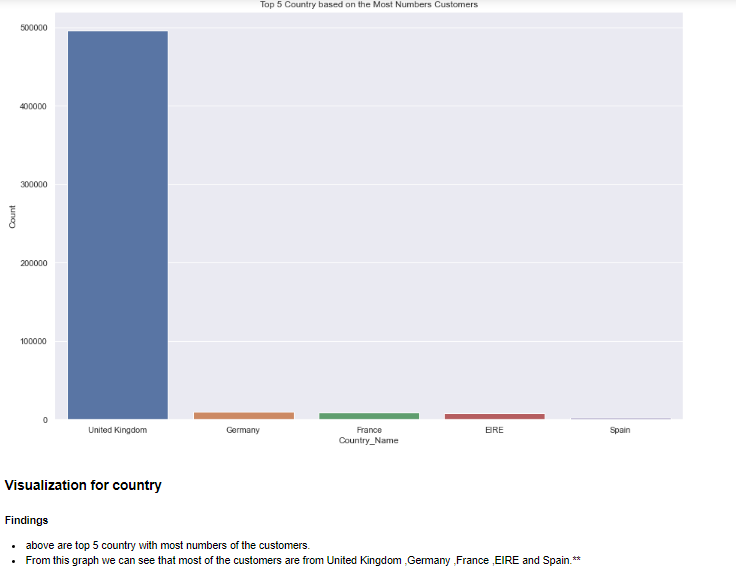
A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and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

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

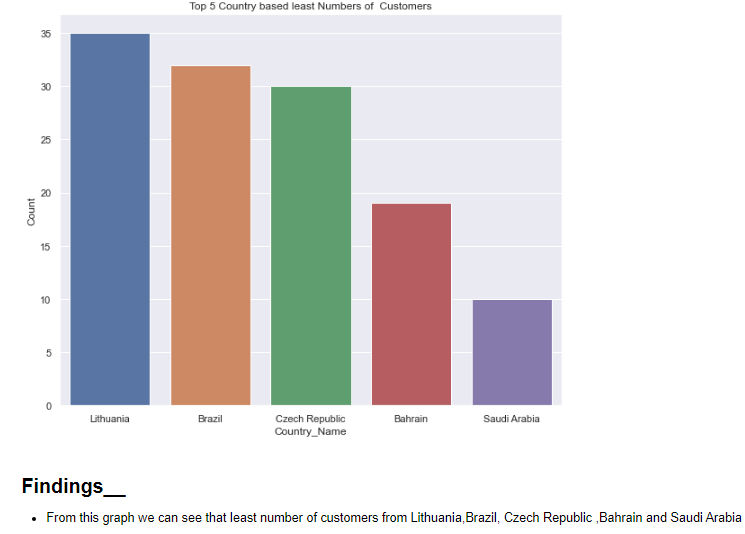
* **Getting the dataset.**
* **Importing libraries.**
* **Importing datasets.**
* **Finding Missing Data and treating those.**
* **Findings duplicates and treating those.**
* **Encoding Categorical Data if any.**
* **Feature scaling and feature engineering and feature selection.**
* **Removing outliers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

**Data Visualization**

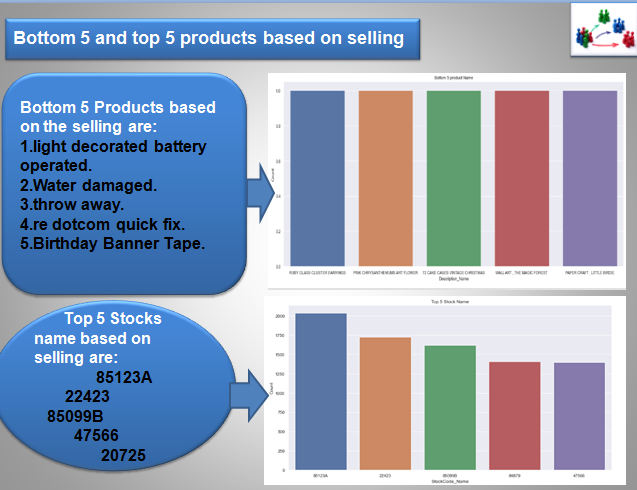
This step of visualization involves checking the data in pictorial form. Apart from this, visualization helps in easy understanding of the data and representation of most of the data information in rightly and quickly.



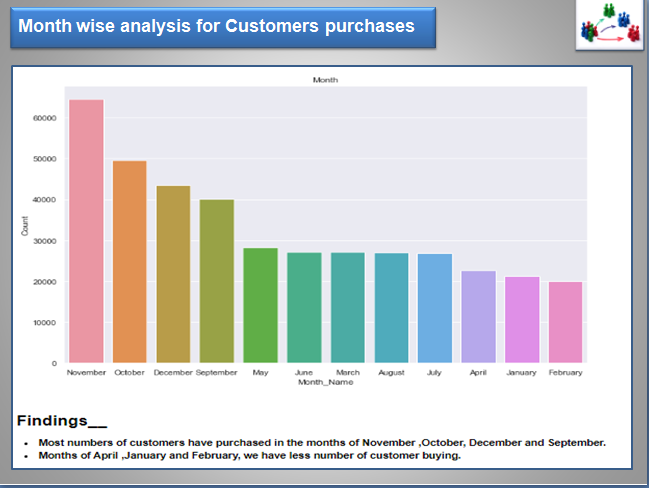
**This above pictorial representation is for top 05 countries based on most numbers of customers.**

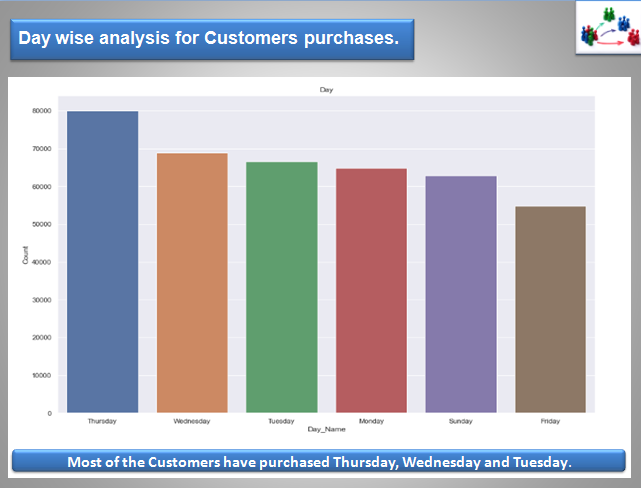


**This above pictorial representation is telling that we have least numbers of the customers from Lithuania,Brazil,Czech Republic, Bahrain and Saudi Arabia.**



**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.

DBSCAN.

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

Elbow Method.

Silhouette score.

DBSCAN.

Dendrogram.

**Procedure performed and Observations**

 Started with data loading and importing the libraries and then started with exploring the data and looking into columns and rows. Looked for the missing values in the dataset and imputed accordingly.

Did the EDA for the dataset and tried to find major trends and insights from data regarding the purchases. I have made few observations while performing the project on the given dataset.

1)There were (541909, 8) rows and columns out of which I saw that there were null values in the description and CustomerID columns, which were of float and object data types.

2) I have removed the null values as imputing them with mode would not be meaningful. After removal of the null values I had (406829, 8) observation and variables respectively.

I did see the overall data distribution and found few points as below:

* In quantity we have values in negative and as well as in Unit Price.
* Found positively skewed distribution of the dataset.

3) I started exploring further country wise, monthly basis, day basis and hourly basis and as per time zone my findings were below:

* Countries with top customers are: United Kingdom, Germany ,France ,EIRE and Spain.
* Most numbers of customers have purchased in the months of November, October, December and September.
* Most of the customers have purchased the items in Thursday, Wednesday and Tuesday.
* I have seen that afternoon timings are popular for the purchasing items.
* Especially 11-12-13-14-15 gave the more numbers of customers purchasing.
* Once I have seen the data and for its major minor trends, I then started with modeling techniques which are below:

4) Used RFM model for to find out the valuable customers based on Recency, Frequency and Monetary values by calculating first RFM score from actual data that I had then I got RFM scores and converted into the quintile bins assigning 1,2,3,4 value where 444 means highest score and 111 combination means least important customer.

* While using this model I have seen few points that there were customers which were having more Recency and more Monetary, more Recency and less monetary.
* Similarly, like for these combinations I have checked for each customer and found the best set of customers after setting the threshold to 5 and 8 respectively given 1263 customers and 2587 customers with threshold of setting to 8.

5) Once I was done with RFM model preparation, I then started with K-mean clustering and hierarchical clustering and DBCSCAN. To cluster the same set of customers and tried with 2 features and 3 features which were (RFM) Recency, Frequency and Monetary.

* I checked the cluster formation with the help of Silhouette score and elbow method and DBSCAN and dendrogram.
* I found DBSCAN performing good to find out the optimal clusters whereas K-mean clustering is not proven that well with elbow method and silhouette scores. After using all methods I have seen that most of the time optimal numbers of clusters were 2.

**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”.