Graphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Student Name

***Sanatkumar Rajmogali Ippalpalli***

Title of Project Report

***CAPSTONE Project 1 – Customer Segmentation***

Table of Contents

[Introduction 4](#_Toc84543759)

[Eckovation theme & Question 6](#_Toc84543760)

[Prerequisites before starting coding 7](#_Toc84543761)

[program DEVELOPMENT steps 9](#_Toc84543762)

[Dataset requirements 9](#_Toc84543763)

[Technique – Eda and kmeans 9](#_Toc84543764)

[PROGRAM / CODE DEVELOPMENT 10](#_Toc84543765)

[Analysis 23](#_Toc84543766)

[CONCLUSION 24](#_Toc84543767)

[REFERENCES 25](#_Toc84543768)

[Figure 1 Import libraries and datasets/modules 10](#_Toc84543769)

[Figure 2 Load Datasets and converting Pandas Dataframe 10](#_Toc84543770)

[Figure 3 Dataset Information 11](#_Toc84543771)

[Figure 4 Dataset Information 11](#_Toc84543772)

[Figure 5 Dataset Information – Removal of NA, Duplicate & Unique Information 12](#_Toc84543773)

[Figure 6 Dataset Preparation -Code for drop & encoding parameters 13](#_Toc84543774)

[Figure 7 results after dataset preparation 16](#_Toc84543775)

[Figure 8 RFM application 17](#_Toc84543776)

[Figure 9 RFM Columns 18](#_Toc84543777)

[Figure 10 Seaborn Visualization 18](#_Toc84543778)

[Figure 11 RFM Model Parameters 19](#_Toc84543779)

[Figure 12 Clusters Optimum Elbow Method 21](#_Toc84543780)

[Figure 13 Clusters Centroid K Means 22](#_Toc84543781)

***EXECUTIVE SUMMARY***

It is very important for any business dataset, to ensure the results produced by any kind of machine algorithm / technique are valid and applicable to any desired business optimization outcomes and goals.

Science and technology improved many technologies and has guided numerous innovative features which advanced the digitization with respect to data (excel, text format) in the charts that we use for quick understanding / judgement.

Exploratory Factor Analysis (EFA) can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EFA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including [machine learning](https://www.ibm.com/cloud/learn/machine-learning).

As part of CAPSTONE project, to find by using dataset which explore / extract top features interact were coded and implemented to capture the details as mentioned in the question.

# Introduction

The objective is to showcase a few of the basics of data science (data cleaning, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses: ***Customer Segmentation***.

Customer segmentation (often abbreviated as CSAT) is a method of dividing customers into groups or clusters on the basis of common characteristics. This activity :

* help in identifying the most potential customers.
* help managers to easily communicate with a targetted group of the audience.
* help in selecting the best medium for communicating with the targetted segment.
* improves the quality of service, loyalty, and retention.
* Improve customer relationship via better understanding needs of segments.
* provides opportunities for upselling and cross-selling.
* help managers to design special offers for targetted customers, to encourage them to buy more products.
* helps companies to stay a step ahead of competitors.
* helps in identifying new products that customers could be interested in.

Customers play an important role and are essential in keeping a product or service relevant; it is, therefore, in the best interest of the business how successful the organization is at providing products and/or services to the marketplace.

Science and technology improved many technologies and has guided numerous innovative features which advanced measurement of multiple attributes (features) indicate recency, frequency, monetary (RFM) provided.

Hence Eckovation includes this CAPSTONE project on Customer Segmentation in the courseware for students to understand, implementation / execute the code themselves on the Online Retail dataset available. This guided project objective is to showcase a few of the basics of data science (data cleaning/transform, Normalization, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses: Customer Segmentation.

This report includes the 5W1H about the theme of development of code and running the code with database available over the internet of Online Retail Customer Segmentation. At the end of the report, the conclusions share the features extracted like Recency, Frequency & Monetary Value (RFM) and useful for next course of activities to gain advantages in the development.

# Eckovation theme & Question

**Theme : Customer Segmentation**

Grouping customers into sections based on their common characteristics is called Customer Segmentation. These clusters allow the companies to target the customers with the correct marketing message and tailor their offers for a specific group. This not only helps them boost their sales, but also helps them build customer relations and understand them in a better way.

In this project, our aim will be to perform customer segmentation on Online Retail Dataset ([https://archive.ics.uci.edu/ml/datasets/Online+Retail#](https://archive.ics.uci.edu/ml/datasets/Online+Retail) ) to understand the customers. Given this dataset, our task is to:

a) Load the dataset and perform a descriptive analysis on it (Total number of entries, the column types, unique/non-null entries for each attribute, unique stock items, visualizing various attributes using bar charts/pie-charts and so on).

b) Perform data cleaning. Specifically, given the dataset, handle the entries that either have missing information or have attribute values that are not feasible such as negative quantity.

c) Perform data pre-processing for the required attribute fields.

d) Since this database has no additional attribute information for the customer, we will use RFM model (refer: <https://clevertap.com/blog/rfm-analysis/> ) for segmentation. Modify the database to include RFM model attributes.

e) Now once you have your database ready, perform data clustering on this dataset by assuming a fixed number of clusters.

f) Find the optimal number of clusters that the customers can be divided into.

# Prerequisites before starting coding

1. Who - Software needed?
2. What - Version / Release of software?
3. Any Prerequisites
4. How - to install the software
5. Which -libraries are needed to execute the problem statement
6. Where – dataset requirements, path location to include in the code
7. When – to use the above feature extraction
8. Who – Software neeed?

Python

1. What- Version / Release of software?

Python version 3.6 (latest version of python)

1. Any Prerequisites

RAM space availability & hard disk space availability

Admin rights to install the software

1. How - to install the software
2. The following url [https://www.python.org/downloads/](https://www.python.org/downloads/%20) can be referred to download python.
3. Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/>
4. Which -libraries are needed to execute the problem statement
5. Sklearn (scikit-learn) ( pip install -U scikit-learn)
6. Numpy (pip install numpy)
7. Matplotlib (pip install matplotlib)
8. Scipy (pip install scipy)
9. Seaborn (pip install seaborn)
10. Where – dataset requirements, path location to include in the code
11. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: [https://www.pythoncentral.io/add-python-to-path-python-is-not- recognized-as-an-internal-or-external-](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/) [command/](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/).
12. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.
13. When – to use the above feature extraction
14. When – to use the above feature data segregation and in depth visual analysis to find out relevant attributes for business excellence.

# program DEVELOPMENT steps

* Dataset requirement
* Technique selections
* Program / code development
* Analysis

### Dataset requirements

To perform customer segmentation on Online Retail Dataset ([https://archive.ics.uci.edu/ml/datasets/Online+Retail#](https://archive.ics.uci.edu/ml/datasets/Online+Retail) ) to understand the customers with attribute features Recency, Frequncy, Monetary Value.

### Technique – Eda and kmeans

Let us hop to the inscribing carving!

### PROGRAM / CODE DEVELOPMENT

As explained step by step during the lecture by mentor, we would approach steps and understand the basics with brief explanation as needed.

#### Step 1: Import the relevant libraries and applicable datasets/modules

Text

Description automatically generated

Figure Import libraries and datasets/modules

#### Step 2: Load dataset and convert Pandas Dataframe

Download the csv dataset and copy into the disk as Pandas Dataframe. Estimate shape of

training and testing dataset to understand available features in the dataset.

Text

Description automatically generated with medium confidence

Figure Load Datasets and converting Pandas Dataframe

Understand, Clean and Transform Dataset

The main purpose of EFA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

The Training dataset has 25 features and we need to check if all features are numerical (integer) format before proceeding to machine learning model. This is called cleaning / prepare the dataset for transformation.

A picture containing text

Description automatically generated

Figure Dataset Information

Based on the figure above, there are multiple variables which needs to be analysed before setting for dataset cleaning / preparation. Each parameter uniqueness to be studied beforehand using describe function in pandas dataframe.

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Figure Dataset Information

#### Step 3: Dataset Preparation

Dataset to be checked if there are any missing values in our dataset or not. It can occur when no information is provided for one or more items or for a whole unit.  issing Data is a very big problem in real-life scenarios. Missing Data can also refer to as NA(Not Available) values in pandas. There are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

* [isnull()](https://www.geeksforgeeks.org/python-pandas-isnull-and-notnull/)
* [notnull()](https://www.geeksforgeeks.org/python-pandas-isnull-and-notnull/)
* [dropna()](https://www.geeksforgeeks.org/python-pandas-dataframe-dropna/)
* [fillna()](https://www.geeksforgeeks.org/python-pandas-dataframe-fillna-to-replace-null-values-in-dataframe/)
* [replace()](https://www.geeksforgeeks.org/python-pandas-dataframe-replace/)
* [interpolate()](https://www.geeksforgeeks.org/python-pandas-dataframe-interpolate/)
* isna()
* nunique()
* duplicated()

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

Figure Dataset Information – Removal of NA, Duplicate & Unique Information

Chart, pie chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Figure 6 Dataset Preparation -Code for drop & encoding parameters

There are quite a few parameters which need to be cleaned in order to prepare the data for use in a machine learning model.

There are also a couple of columns that are unnecessary, such as 'Unnamed:0' and 'id' which shall be dropped. The functions below will be used to perform the dataset preparation.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, table

Description automatically generated

Text, application

Description automatically generated

Table

Description automatically generated with low confidence

Figure 7 results after dataset preparation

#### Step 4: Visualization of Attributes

A picture containing text

Description automatically generated

**Chart, bar chart

Description automatically generated**

A picture containing calendar

Description automatically generated

**Chart, bar chart

Description automatically generated**

#### Step 5: RFM Model

Graphical user interface, text, application

Description automatically generated

Figure RFM application

Graphical user interface, text, application

Description automatically generated

Figure 9 RFM Columns

Table

Description automatically generated

We will use ***Matplotlib and Seaborn*** library for the data visualization.

Text

Description automatically generated

**Graphical user interface, application, table, Excel

Description automatically generated**

Figure 10 Seaborn Visualization

Table

Description automatically generated

A picture containing table

Description automatically generated

Text

Description automatically generated with medium confidence

Figure 11 RFM Model Parameters

A picture containing text

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

**Chart, line chart

Description automatically generated**

Figure 12 Clusters Optimum Elbow Method

Graphical user interface, text, application, email

Description automatically generated

A picture containing graphical user interface

Description automatically generated

**Chart

Description automatically generated with medium confidence**

Figure 13 Clusters Centroid K Means

### Analysis

Though we had got recency, frequency, monetary value of UK based customers having more than 90% overall customers.

This is first attempt, the improvements in the code with time with multiple attempts may be checked and justified in future.

# CONCLUSION

In this guided project, we built a code for development of RFM model on the Online Retail dataset

As it’s first attempt, the improvements in the code with time with multiple attempts may be checked and justified for the conclusions.

This entire program runs within few minutes.

# REFERENCES

1. <https://www.ibm.com/cloud/learn/exploratory-data-analysis>
2. <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>
3. <https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python>
4. <https://towardsdatascience.com/customer-segmentation-analysis-with-python-6afa16a38d9e>
5. <https://colab.research.google.com/github/khalidmeister/Customer-Segmentation-using-Python/blob/master/Customer%20Segmentation%20in%20Python.ipynb#scrollTo=lLREAo-yXZjt>
6. [https://archive.ics.uci.edu/ml/datasets/Online+Retail#](https://archive.ics.uci.edu/ml/datasets/Online+Retail)
7. <https://clevertap.com/blog/rfm-analysis/>