**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Team Member’s Name:**  1.Mr.Pratik Ghodke Email Id :- pratikghodke777@gmail.com  2.Mis.Madhuri Bonela Email Id :- [madhuribonela91998@mail.com](mailto:madhuribonela91998@mail.com)  **Contribution:**  **1.Mr.Pratik Ghodke**   1. Importing libraries 2. Importing data set 3. Inspecting data set 4. Processing of data set 5. Cleaning of data set 6. Indentification and removal of null values 7. Replacing the null values if any null values with median 8. Check the discrete and continuous features 9. Handling outliers 10. Removing outliers by standard methods and plotting graphs 11. Feature Engineering 12. EDA -Most Frequent Values 13. Visualization of the country with most count in number of orders 14. Visualization of the country with highest in quanity number of orders 15. Visualization of the product with high in quanity of orders 16. Visualization of the product that has made most of the revenue 17. Visualization of the customer if with most number of cancellations 18. Visualization of the country with highest number of cancellations 19. Visualization of data with AUTOVIZ 20. Visualization of unitprice by country ie., the top 20 21. Visualization of Average CustomerID by country the top 20 22. Visulaization of Average unitprice by quantity 23. Visualization of Average CustomerID by quantity Ie., the top 20 24. RFM modelling 25. Log Transformation 26. Normal Distribution for RFM 27. RFM - loyalty level of customers 28. Clustering prerequisites 29. K- Means Clustering 30. Silhouette plots for clusters 31. Fitting the model on n\_clusters is 2 and displaying their RFM values 32. Fitting the model on n\_clusters is 5 and displaying their RFM values 33. Fitting the model on n\_ clusters is 4 and displaying their RFMvalues 34. Hierarchical clustering   35Analysis of the model of clusters when n value is different  36.PPT Documentation  37.Technical Documentation  **2.Mis.Madhuri Bonela**   1. Importing libraries 2. Importing data set 3. Processing of data set 4. Cleaning of data set 5. Data exploration of data set 6. Identification and removal of null values 7. Replacing the null values with median 8. Check the discrete and continuous features   9. Handling outliers  10. Removing outliers by standard methods and plotting graphs  11. Feature Engineering  12. EDA -Most Frequent Values  13. Visualization of the country with most count in number of orders  14. Visualization of the country with highest in quanity number of orders  15. Visualization of the product with high in quanity of orders  16. Visualization of the product that has made most of the revenue  17. Visualization of the customer if with most number of cancellations  18. Visualization of the country with highest number of cancellations  19. Visualization of data with AUTOVIZ  20. Visualization of unitprice by country ie., the top 20  21. Visualization of Average CustomerID by country the top 20  22. Visulaization of Average unitprice by quantity  23. Visualization of Average CustomerID by quantity Ie., the top 20  24. RFM modelling  25. Log Transformation  26. Normal Distribution for RFM  27. RFM - loyalty level of customers  28. Clustering prerequisites  29. K- Means Clustering  30. Silhouette plots for clusters  31. Fitting the model on n\_clusters is 2 and displaying their RFM values  32. Fitting the model on n\_clusters is 5 and displaying their RFM values  33. Fitting the model on n\_ clusters is 4 and displaying their RFMvalues  34. Hierarchical clustering  35. Analysis of the model of clusters when n value is different  36. PPT Documentation  37. Technical Documentation |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/pratikghodke777/Online-Retail-Customer-Segmentation>  Drive link:- https://drive.google.com/drive/folders/1jhmhPI\_bT6tuwDCLWs7er7d-YFtEYA4D?usp=sharing |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Customer segmentation is the process of classifying customers based on their shared behaviour or other attributes. The groups should be homogeneous within them and should also be heterogeneous to each other. The main goal is to identify customers that are most profitable and loyal and the ones who churned out, to prevent further loss of customers by redefining company policies. Having a large number of customers, each with different needs it is difficult to find which customer is most important for business and target them with an appropriate strategy .Effective decisions and stratergies to be taken are mandatory for a company to grow and result in good revenue outputs. In these days competition is huge and all companies are moving forward with their own different strategies. We should analyise data and take a proper decision.  Every person is different from one another and we don’t know what he/she buys or what their likes and requirements. But, with the help of machinelearning technique one can sort out the data and can find the targetgroup by applying several algorithms to the dataset. Without this, It will bevery difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset.  Here, The customer segmentation using various techniques to know the customer likes and their purposes so that we can apply it on business area to develop the business and it's growth interms of business.  The existing methods are storing customer information through paperwork and as a digital data we are storing and analysing it to predict customer benifits and needs and to know how much sales were sold and who are the regular customers and which products are having more demand in sales and what is the profit of business so ,all the required data is out from the data for business improvement so ,storing and data analysis is playing vital role to predict the business up and downs . 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. * The goal is to identify the customers behaviour and their needs by analysis of data and apply them on business field inorder to increase sales and profits and to calculate their RFM where recently purchased customers are analyzed , frequently purchasing costomer are analysed and also monetered .   **Approach**:  **Imported Libraries**  In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.  **Missing Value Imputation**  We will now check for missing values in our dataset. After checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.  **Cleaning and manipulation the data set**  We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models.  **EDA Analysis.**  The exploratory data analysis that we performed on our train dataset helped us to realize how different features in our dataset influence the target variable**.**  **1.Identification and removal of null values from data set**  **2.Updated heat map by removing all the highly correlation between data features**  **3.EDA by alternative method by using Autoviz**  **4.Model building prerequisites**  **5.Model building**  **Model Building** **1. RFM - RECENCY, FREQUENCY, MONETARY****2. CLUSTERING****2.1 K-MEANS CLUSTERING****2.2 HIERARCHICAL CLUSTERING** **Conclusion:**  **RFM QUANTILE CUT**  1. Lost poor customers  2. Lost loyal customers  3. Good customers  4. Best customers  **K-MEANS (2 CLUSTERS)**  1. Best customers  2. Lost poor customers  **K-MEANS (4 CLUSTERS)**  1. Losing loyal customers  2. Best customers  3. Lost poor customers  4. Recently visited average customers  **K-MEANS (5 CLUSTERS)**  1. Lost poor customer  2. Best customers  3. Recently visited average customers  4. Losing loyal customers  5. Average customers  **AGGLOMERATIVE (2 CLUSTERS)**  1. Average customers  2. Best customers  EDA Observations  **1.United Kingdom has most number of orders With a count of more than 16000.**  **2.First country with high quantity of orders is from Netherlands with a count of 80 above**  **3.The product with high quantity of order is Paper Craft Little Birdie with a count of 80000.**  **4.Product that made most of the profit in terms of revenue is Paper Craft Little Bride. With more than the count of 160000.**  **5.The customer id with most number of cancellations is 149110.**  **6. Country with highest number of cancellation is United Kingdom with a Count of more than 7000.** |