Format for the Assignment

You can format the assignment based on the following sections:

Note:

1. Here the page number mentioned are just random numbers and do not depict the length of a section.
2. You can add a new section or sub-section as per your approach.
3. In section 5, you can create sub sections:

   For example

   5.1 Handling Missing data:

   (Mention what was the problem, in which column and how you have handled the problem. Also give the small sample of before or after output).

   5.2 Detect Outliers:

   Some of the sample sections are given below for your reference on the next page. Please do not copy paste the same.

Executive Summary

Data contains all the information of transactions which includes distinct customers who are visiting the store and the detailed information of items purchased. The characteristics of customers and the items are categorized according to the data produced by the business. This characterization is performed by using the unsupervised algorithm which analyses data and groups them to clusters. The nodes are segregated into their respective clusters. Considering these different groups in the customer category, there were groups which bring good revenue to the business while purchasing a few costly items or by purchasing many budgeted items. These customers were contributing largely to the revenue. Few customers are loyal who consistently bought several specific items, which resulted in consistent revenue and few customers were purchasing a specific group of items in large quantities, but they were not visiting the business often. We assume these groups are small business vendors and are potential customer who can bring more interesting patterns of item purchase and revenue. The remaining clusters had the customers who were visiting the store occasionally and there is a need to attract those groups by providing offers and having them visit the store often.

The analysis was also performed on items and the groups that were quite interesting. The more popular groups which brought the highest revenue to the business had the least average price per visit, which indeed indicates that those were the items which were household daily use items and fall into the category of fast-moving consumer goods with least price. The next popular group/cluster has a better average price per visit, but the trend remains as popular groups. Few groups of items were purchased largely in quantity and the remaining groups had the least average prize since the items were returned often. There are few recommendations tonavigate investment into quickly consumed groups and to monitor their availability in the store.Further monitoring these groups and understanding developments according to the recommendations must be maintained to improve the business growth. The profiling detailed every category and appropriate recommendations were provided to evaluate the current business processes.

Objective

From the “OnlineRetail” dataset we will be considering 2000 Records each to create the Customer and Product Datasets, which further must go through the cleaning process and data analysis. An unsupervised algorithm (k-means) is performed on the data to identify indigenous clusters in accordance with Customers and Items. The number of clusters is identified using Elbow and Gap Statistics methods. The characteristics of the clusters are studied and comparedwith each other and relevant information is provided for good business practices.

Data Summary

For our analysis, we will be using the "OnlineRetail" dataset. It consists of 541,909 records which are from December 2011 to December 2012, of the various products bought by customers from multiple countries. The following are the attributes of the data set:

Attributes	Description
Invoice Number (InvoiceNo)	It is generally a 6-digit number that uniquely identifies the transaction made by a customer.
Product Item ID (StockCode)	The code is alphanumeric, consisting of 1 - 6 characters. It is used to uniquely identifies the product item.
Description	It is text that is used to describe the stock code. It mainly provides details for the product item.
Quantity	It indicates the number of products brought or returned by the customer.
Unit Price	It is a positive float, that indicates the cost of a single product. But for some records the Unit Price is negative, it was done to "adjust for bad debt" as mentioned in the Description.
CustomerID	It is generally a 6-digit number that uniquely identifies a customer.
Country	It is text that describes the location where the product was bought.
InvoiceDateTime	It consists of the data and time when the product was purchased. In general, the records are between December 2011 to December 2012.

Table 1: Attributes of the "OnlineRetail" Dataset

Observations:

The following table briefly describes the observation made on each of the attributes in the "OnlineRetail" data set:

Attributes	Description
Invoice Number	• All records follow the same pattern, i.e. The code is numeric, consisting of 5 characters. • There are 9,292 records containing Invoice Number 0. • Majority of the records with Invoice Number 0 have a negative value for Quantity.
StockCode	• The code follows 2 distinct patterns: 1. The code is numeric, consisting of 5 characters. 2. The code is alphanumeric, consisting of 5 numeric characters and a single letter. • Apart from the above-mentioned patterns there are 15 unique stock codes that don't follow the above patterns, but they are used to indicate Discount, Bank Charges, Amazon Fee, Samples, Postage, etc.
Description	• For most of the records it displays the title of the product. • It provides describes for the 15 unique stock codes as mentioned in the previous observation. • There are 1,454 records that have no description.
Quantity	• There are 10,624 records that have a negative value. • The Maximum Quantity of a product brought and not returned by a customer is 12,540
Unit Price	• There is 1 record that has a negative value. • The records are between 0.00 to 9.99.
CustomerID	• All records follow the same pattern, i.e. The code is numeric, consisting of 6 characters. • There are 135,080 records containing Customer ID 0. • There are 1,719 records with Customer ID 0 and have a negative value for Quantity. • There are 386 records with Customer ID 0 and Invoice No 0.
Country	• There are 38 distinct counties in the dataset. • Majority of the purchases is done in the United Kingdom. • The least number of purchases is done in Lebanon, RSA and Brazil.
InvoiceDateTime	• All records follow the same pattern of when the product was purchased. • The records are between December 2011 to December 2012.

Table 2: Observations made on the Attributes of the "OnlineRetail" Dataset

Limitations:

Invoice Number 0, CustomerID 0 and a few StockCode items are not clearly defined in the data set. These attributes are interlinked with other attributes in the dataset like Quantity, UnitPrice,etc. so if we remove them from the dataset we would not get accurate results.

Since Invoice Number 0 mainly contains data on the items that were returned by the customer,so if we remove Invoice Number 0 we would not be able to get the actual quantity purchased by the customer.

To overcome this, we summed the quantity of items purchased by a customer, so if a customerpurchased and returned a product then the aggregated quantity would be 0.

Outcome:

From the "OnlineRetail" dataset we will created 2 datasets or tables in MySQL, **ProductCluster** and **CustomerCluster.** Each table will consist a total of **2000 records** from the "OnlineRetail" dataset and the data will be order by the highest revenue earned. Each table will be utilized thefollowing attributes:

1. **Invoice Number**
2. **StockCode**
3. **Quantity**
4. **Unit Price**
5. **CustomerID**

Design/Method/Approach

The following steps were performed for our analysis:

1. **Dataset Selection:**
    - We begin by selecting the "OnlineRetail" dataset that contains all the transactiondetails.
    - Then we review the source data to understand the content, its structure and itsinterconnectivity with other attributes within the same dataset.
    - Then we check if the data is correctly formatted and consistent. To do so we perform various analytical checks on the data, such as maximum/minimum,

sum andcount.

2. **Feature Selection:**
    - Then select and engineer the appropriate features to support the analysis.
    - We use the RFM model to characterize the Customer.
3. **Sample Selection:**
    - We utilize the above features to select the top 2000 customers and top 2000products based on revenue.
4. **Outlier Removal:**
    - We plot the filtered dataset using 'ggpair'. Then we identify and remove the outliers.
5. **Normalize Data:**
    - The features selected from the filtered dataset vary in magnitude, this in turn couldcause issues when we use the k-means algorithm in the next step.
    - If we apply the algorithm to the filtered dataset it would take only the magnitude ofthe features into consideration and neglect the units.
    - The K-mean algorithm uses the Euclidean distance between two data points. Hencethis measure is sensitive to magnitude. (Garbade, 2019)
    - Hence, we should scale and equally weigh all features
6. **Clustering:**
    a) To determine appropriate number of clusters for the analysis we use:
        - Elbow Method
        - Gap Statistics
    b) Performed clustering using k-means algorithm.
7. **Profiling:**
    - De-normalize the data since we have made the clusters for the datasets.
    - Based on clustering results, we create customer and product profile use tableau,after combining the results with the metadata from the original dataset.