**UNIVERSITY OF WESTMINSTER**

**Msc Big Data Technologies**

Module Code: 7BUIS010W.2

Module Title: Data Warehousing and Business Intelligence

Module Leader: Dr. Panagiotis Chountas

Assignment: Coursework 2

Student ID: W1848650

Student Name: Mohammadamin Shams

**Table of Contents**

[1 Abbreviations 2](#_Toc125311015)

[2 Introduction 3](#_Toc125311016)

[3 Data Understanding 3](#_Toc125311017)

[4 Perform RFM Segmentation 7](#_Toc125311018)

[5 Customer segmentation with k-means 9](#_Toc125311019)

[6 Review of RFM Segmentation Result 11](#_Toc125311020)

[7 Data Mart Design 12](#_Toc125311021)

[8 References 12](#_Toc125311022)

# Abbreviations

|  |  |
| --- | --- |
| RFM metrics | Recency Frequency Monetary |

# Introduction

This report is about analysing data from a UK-based non-store online retail company selling mainly all-occasions gifts. Many customers of the company are wholesalers. All the transactions occurred between 01/12/2010 and 09/12/2011.

The raw data contains information about each of the products ordered in different transactions. In the next parts all the features are explained.

Goal of this report is to create groups of customers based on their activity and understand each group behaviour towards the company, so the analysing team would figure out which actions need to be done for each of these groups to make them engage more or don’t do any actions if they are bad customers.

To put customers into segments based on their activity, first we need to capture the data and understand the dataset, do the preliminary steps to analyse the raw data. After that because the aim is to get insights out of the raw data, it is needed to curate the dataset and get rid of noisy and uncleaned information, which won’t add any useful information for the customer analysis.

After creating a clean data set and got rid of the uncleaned noisy data, the analysis of customers will begin. First it is needed to use definitions used by analysts that suits the data to analyse the customers. In this report recency frequency monetary properties are used to investigate the customers behaviour. Recency is the duration between their last order till the latest order of all customers, frequency is the number of purchases they made and monetary is all of money they spent for their orders. Based on the data set these properties should be created.

After that there is two different ways to create groups of customers, first one is creating a scoring system to score each of these RFM values based on their value, then define groups of customers based on their score on RFM values and at last take actions for different segments based on their score.

The other way to do the segmentation of customers after creating RFM values, is using unsupervised machine learning algorithms to label them into different groups. Then analyse the outcome and use that labelling to see which of those groups of customers, are good or bad or great and take actions to keep them engaged or get rid of them. In this report the second way with the help of k-means clustering is chosen to put customers into segments.

In the last part of the report an evaluation of the segments is provided, and data mart design for this company has been done too. All the coding parts for the above goals is done by Python in the Google Collab environment.

# Data Understanding

First, it is needed to import libraries for our analysis in Python (libraries and all the code is provided in the appendix) and then we need to capture the data set of the retail company and understand features of our table to understand the data. For capturing the data Pandas library is used to read the raw file. Another good tool in the python is profile report that can give us a good idea of the dataset and its variables, the report is done with PyCharm environment because of issues that google collab has with profile report function and the detailed report of dataset is provided in the appendix with html file. The first rows of the data are in the figure below with the code that will help us to do this part and in the table after the figure, features of the dataset with a description of each feature is provided.

A picture containing table

Description automatically generated

Figure Sample of the data

|  |  |
| --- | --- |
| Feature | Description |
| InvoiceNo | Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation. |
| StockCode | Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product. |
| Description | Product (item) name. Nominal. |
| Quantity | The quantities of each product (item) per transaction. Numeric. |
| InvoiceDate | Invoice Date and time. Numeric, the day and time when each transaction was generated. |
| UnitPrice | Unit price. Numeric, Product price per unit in sterling. |
| CustomerID | Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer. |
| Country | Country name. Nominal, the name of the country where each customer resides. |

Table Features in the table explained

Now we can explore the data set and see different properties of the data set. we can see the number of columns and rows, also get a good info for every column and data type, and see if it has null values or not, with the snip of the code below. We can see the numeric values details with describe function in the figure too.

Graphical user interface, text

Description automatically generated with medium confidence

Figure Data description

We can see, we have null values and also there is minus values in unit price and quantity, and they need to be curated. Another thing needs to be done is understand how many null values and how many duplicates is in our dataset. Next figure is showing how to do it with the help of isnull function to see the number of null values and for duplication, we use duplicated function on our dataset to see all of them in our table.

Graphical user interface, application

Description automatically generated

Figure Null values and duplication

Now we can explore each column to see the range of values and get an insight of the distribution of some of the columns on the dataset. All of the analysis can be created with different codes but also all of them can be get at once with profile report and in the html file in the appendix all of the different analysis on features and exploration is provided but in the next steps we try to them with different codes.

In the figures below we examine some of the features and the code to how to do it and then we talk about insights that we can get from our distribution analysis on our features.

Table

Description automatically generatedText

Description automatically generatedGraphical user interface, table

Description automatically generated with medium confidenceText

Description automatically generated with low confidence

Figure Distribution analysis

We can see the countries that orders come from and since the company is in UK, 90 percent of the transactions are from UK too. Also, we can see top 10 items being bought and the unique number of all of them is 4223 based on their unique description but if we compare it to unique stock code which is 4070 it seems that some descriptions are redundant and some of them are used in the wrong way. There is 25900 unique transactions and at last in the distribution of unit prices we have negative values which may refer to orders that returned to the company.

Other steps like corelation and other distribution and visualizations are done after cleaning the data set. With the help of SQLite and connecting to it and using some commands we clean the data. Also, these steps can be done with the help of Python too, for example using drop duplicates and drop nulls functions to get rid of duplicates and null values but SQLite is used to first create the table and then use SQL commands to get rid of Null values, duplicates, and also invoice numbers that starts with c which means cancelled orders and descriptions with question mark(‘?’) is also dropped and unit prices below 0 and quantity below 0 are dropped too. To do this cleaning process first we created a table then we selected from that table all the columns with putting desired constraints on our rows to exclude null values, and other cleaning goals mentioned above with a WHERE clause, in the fig below the code is provided.

Table

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidence

Figure Cleaning the data with SQLite

Now we can do other analysis that we couldn’t do on the dataset like distribution, visualization because of some of the values weren’t cleaned and contained noises that prevented us from getting other pre-processing steps done.

We can make sure the data is cleaned by checking the null and duplicated values which we can see there is non in our dataset with figure below.

Graphical user interface, text, application

Description automatically generated

Figure Check to see the data cleaning step

Now we can see the distribution of some of the features with the help of seaborn library and distribution plot.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure Distribution plot

Another important step is understanding the correlation of numeric columns to each other for further analysis. This can be done with correlation function and for a better visualization heatmap is used to understand correlation better with the help of seaborn library and heatmap function in this library.

Text

Description automatically generatedChart, treemap chart

Description automatically generated

Figure Corelation analysis

We can see that these columns are not correlated but these are not the values need to be used in our RFM because we didn’t group them by the customer id, we do that after explaining RFM metrics.

# Perform RFM Segmentation

After cleaning the data and getting it ready for our customer behaviour analysis, first we need to define metrics to evaluate our customers and put them in different segments based on their behaviour on our defined metrics. RFM analysis is used because it will allow us to compare clients. This analysis gives a sense of how much revenue come from old and new customers to take some actions to engage new ones and also make old one interested too.

The recency, frequency, monetary value (RFM) model is based on three quantitative factors namely recency, frequency, and monetary.

Recency factor is based on the how much time has elapsed since customers last activity or transaction with the company. This information can be used to remind recent customers to revisit the business soon and make them back with offers.

Frequency factor is based on how many times a customer interacted with the business in a period of time. It can be helpful to analyse it with recent purchases customer makes on the past and keep reminding them on the time that we analysed they would need to buy that product again.

Monetary factor is based on how much money customer spends on the company. This will help us to know the customers who spend a lot and make some arrangements and offers to keep them engaged more and respect them because they help a lot to business runs smoothly.

For analysing our customers, we need to create these features based on the clean data set we have. There is two different ways to do that, one is using SQLite again with the help of GROUP BY expression on customer ID and for frequency use a count function and for recency (to get the actual recency first we need to get customers last order and then find recency) use the maximum of order dates and for monetary use sum function on all of the purchases (unit price\*quantity) for a customer. This is one way to create our table for RFM analysis or we can use Python to do this objective on the clean data with the help of group by function in python. Both ways are used but the SQLite is showed in the figure below. Also, we need to know we haven’t created RFM metrics yet. We gather the data above for each customer and we only need to change the last order date value and find the last order date of the customer and then subtract it from the last date of all the transactions and to avoid having 0, we add one day to it and we have the recency in this way, but monetary and frequency is created with SQLite and we need to change the variables name only to create a more understandable table. In the figure all of the steps to get the final RFM table is provided.

Table

Description automatically generated with low confidenceGraphical user interface

Description automatically generated with medium confidence

Figure Creating RFM table

We can again do some data analysis processes on our RFM table before using it on the k-means algorithm. We can see the distribution analysis and correlation analysis in the figures below. Like before all of them are created with the help of seaborn library and displot() for distribution and for the correlation visualization heatmap function is used.

Table

Description automatically generatedChart

Description automatically generatedTable

Description automatically generatedChart, histogram

Description automatically generated

Figure RFM metrcis Distribution

A picture containing table

Description automatically generatedChart, bar chart, treemap chart

Description automatically generated

Figure RFM correlation

From the corelation analysis we can see that frequency and recency corelate together and they can be good candidates to use for k-means, but we use all 3 for k-means but at the end we will see these two are well grouped after k-means performed.

# Customer segmentation with k-means

After creating our RFM metrics, it is time to create our segments and find meaning out of them. One way is using scoring system and based on the monetary, frequency and recency, create a score for each of the metrics and based on the range of scores we can create multiple segments. For example, if we score them from one to four, we can create 64 different segments. After that based on the score that each of these values will get, we can define group of customers, for example if they have a high score in recency monetary and frequency, they can be called gold customers or if they got a low score, we could put them on group of bad customers. Based on the meaning of the RFM metrics we can create different groups and use those groups with different actions to maintain them and keep them engaged in the company. By reading the coursework description and searching online I found out this is not the objective of report, and we need to use k-means instead of our scoring system to put customers in different segments.

Another way to create segments of customers is using unsupervised machine learning algorithms like k-means to label our data based on their features. K-means algorithm will get the desired columns and based on its algorithm put data points in different groups which data points in a group are like each other. But we need to mention the number of groups that we want so our data be labelled based on number of k. To find k, we will put range of 2 to 11 for k and see the elbow curve which will give us info about the best number of k, and after that specific number choosing more groups, can’t give us more data. Also, we need to use another method called silhouettes score too to find k, based on both of these we will find the best k value of groups and after that we label our data into k groups.

Before showing the code and use k-means to find the number of k and then create our segments, we need to know that k-means algorithm is sensitive on scale of the data we put in it. To get a better result 2 steps can be done. One to eliminate outliers because k-means is sensitive to range of the data, and it based on the data points being dense and close to each other. Second thing is using standardising methods to scale the data before k-means. Both approaches are done to get better results. Figure below shows the boxplot to see outliers and also a function created to get rid of data out of 10 percentile and 90 percentile band in each feature (RFM).

Chart, box and whisker chart

Description automatically generatedGraphical user interface, text, application

Description automatically generatedChart, box and whisker chart

Description automatically generated

Figure Boxplot after and before removing outliers

After removing outliers, data needs to be scaled. Now we can perform k-means but first with the help of silhouette score and distortion score elbow the number of k is found and its 4 based on these two values.

Text

Description automatically generatedChart, line chart

Description automatically generated

Figure Standardising the data and finding the right number of clusters

After finding the number of clusters, we can use k-means on our dataset to label them and another column is added to dataset called label to put customers in our four groups.

Text

Description automatically generated with medium confidenceTable

Description automatically generated with low confidence

Figure K-means with 4 segments

After the k-means we can see our dataset is grouped into four groups the label zero have 2092 customers inside it and label 2 has the 992 and 1 has 841 and 3 has 313. Also we can see the mean of recency, monetary, frequency for each of the groups to understand each segments situation.

In the next parts the visualization of the data is provided in 2d and 3d. Also, boxplot of each label in the features, monetary, recency and frequency is provided with the code to know how to create them, after that the meaning that can be got out of these visualizations and segments with k-means is provided in the next title.

Chart, box and whisker chart

Description automatically generatedApplication, Word

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure Segments visualization

# Review of RFM Segmentation Result

Based on figure 14 and 15 we can see we have 4 groups of customers. The labels are used for different groups to describe the situation of each group.

Label 2 has 992 customers inside it and based on the figures and mean of the values on fig 14 we can see that this group has high recency, low frequency and lowest monetary. Putting all of this together we can see these are not our good customers, the last time they bought products is not recent and its been a while, they don’t have a lot of orders in this period of time and also they are not spending a lot of money. It’s safe to assume these are the customers that we lost and maybe come up with deals to get back their attention towards the company again.

Label 0 has 2024 customers inside it and majority of customers are from this segment. They have the second highest recency which means, it’s been a while since their last order. Their frequency is better than label 2 but its still not high enough and they are spending more than label 2 in the products they buy. This group are important because they are the majority of customers and we can say they are not the worst customers but with good measures and offers and other marketing techniques, we need to get their attention so at least some of them become the company better customers. This group have potential and also based on the number of members needs to take right actions to get a better result from them.

Label 1 is 841 of customers and they have a good recency is not the lowest but its second lowest ands its good. They buy products frequent, and they also spend good money on the products. This group is one of the good groups which means, they even can become better, or they may lose interest. Analysists should work on them and find ways to not lose them based on their needs with offers related to them.

Label 3 is the company best customers; they are 313 customers not a lot, but they buy products frequent and pay more than anyone on the products. Also, they have the best recency. They should be the company’s highest priority because they contribute a lot to the business and they deserve to be treated better than the others.

# Data Mart Design

In the last part we need to create a data mart for our dataset based on the findings in the previous parts. To create a data mart first we need to find our business process and then find our dimensions and find some metrics to measure the business process. Based on the data we gathered, we evaluated the customers based on their activity towards the company. Those activities measured were frequency, monetary and recency. So, we can say the main business process is customers and we want to evaluate them.

One of the dimensions of every data mart is time and based on the time of the purchase, recency can be measured so we don’t need to create another feature for capturing recency because we can create it based on last order and its not wise to use more space than its needed.

Another dimension is products which will give info of all products with their ids and their price. Also, description of every product in the dataset.

In the fact table we measure our customers based on their orders, so our primary key is not only the customer id, it’s a combination of customer id and order id and time. In the fact table we measure the total price and quantity of the products being sold so we could find monetary, and frequency can be found by the number of orders grouped by customer id and use a count function.

So our main goal or business process is measuring the customers behaviours and that makes it our fact. Dimensions are time and products. Measures are quantity and price and total price of every order customer made. In the figure below data mart design is provided.

Graphical user interface, application

Description automatically generated

Figure Data mart design

# References

Golfarelli, M. and Rizzi, S. (2011). Data Warehouse Testing. *International Journal of Data Warehousing and Mining*, 7(2), pp.26–43.

Investopedia. (n.d.). Recency, Frequency, Monetary Value (RFM) Definition. [online] Available at: https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp#:~:text=RFM%20analysis%20allows%20a%20comparison [Accessed 4 May 2022].

matplotlib.org. (n.d.). Choosing Colormaps in Matplotlib — Matplotlib 3.5.0 documentation. [online] Available at: https://matplotlib.org/3.5.0/tutorials/colors/colormaps.html [Accessed 4 May 2022].

Chandana, D. (2021). Exploring Customers Segmentation With RFM Analysis and K-Means Clustering. [online] Web Mining [IS688, Spring 2021]. Available at: https://medium.com/web-mining-is688-spring-2021/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-118f9ffcd9f0#:~:text=RFM%20%2B%20K%2DMeans%20extended&text=The%20fewer%20number%20of%20days [Accessed 4 May 2022].