CISC 839
Topics in Data Analytics
Competition 3

Unsupervised learning

Sahil Shethiya
20186685
19ss55@queensu.ca
School of Computing

Kamal Andani
20188032
19kaa3@queenu.ca
Department of Electrical and
Computer Engineering

Software Packages Used:

1. **Pandas:** Python has long been great for data managing and preparation, but less so for data analysis and modeling. Pandas is an open-source library that helps fill this gap, enabling you to carry out your entire data analysis workflow in Python.
    - To install type the following in Anaconda prompt (terminal):
      $ conda install pandas

2. **scikit-learn:** Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib.
    - To install type the following in Anaconda prompt (terminal):
      $ conda install scikit-learn

3. **Seaborn:** Seaborn is a matplotlib-based library of Python data visualization. It offers a high-level user interface to draw statistical graphics that are appealing and insightful.
    - To install type the following in Anaconda prompt (terminal):
      $ conda install seaborn

4. **Datetime:** In both basic and complex ways, the datetime module offers classes for manipulating dates and times. Although the date and time arithmetic is provided, effective attribute extraction for output formatting and manipulation is the objective of the implementation.
    - To install type the following in Anaconda prompt (terminal):
      $ conda install datetime

Description of analytics process:

- Firstly, we checked for the missing values and found out that two columns were containing missing values i.e. **CustomerID**, **Description**. Thus, we removed the records missing **CustomerID** but didn't remove the record with missing **Description** since that attribute isn't used during analysis.
- Secondly, looking at the unique values of the **Country** column we found out that most of the customers were from the UK thus we only kept only those data.
- We also found out the **InvoiceNo** attribute having value 'C' which means the order was canceled and cannot be used during the analytics process thus we removed those records as well.
- Furthermore, we also deleted the records in **Quantity** having values less than zero and records in **UnitPrice** having value zero.
- After dealing with missing values, we removed the outliers by finding the IQR for **Quantity**.
- For the analytics process, we only kept 5 important attributes i.e. **InvoiceNo**, **CustomerID, InvoiceDate, Quantity, UnitPrice.**
- We added one new attribute i.e. **TotalPrice** which is the multiplication of **Quantity and UnitPrice.**

- Next, the **InvoiceDate** has value with both date and time of purchase but we only kept the date as the time is not required during analysis.
- Now after handling Missing values, outliers, and feature extraction, we did the RFM Analysis.

RFM Analysis:

- **Recency:** To calculate, we grouped the data by CustomerID and find the difference between the reference date(09/12/2011) and the most recent purchase date for each customer respectively.
- **Frequency:** To calculate, we grouped the data by CustomerID and found the Unique InvoiceNo in each group.
- **Monetary:** To calculate, we grouped the data by CustoemrID and found the sum of TotalPrice for each customer respectively.

```
data_rfm.describe()
```

	Recency	Frequency	Monetary
count	3825.000000	3825.000000	3825.000000
mean	92.665098	4.051503	1169.564477
std	99.286524	6.662454	2240.842468
min	1.000000	1.000000	1.900000
25%	18.000000	1.000000	243.220000
50%	51.000000	2.000000	551.950000
75%	145.000000	4.000000	1293.350000
max	374.000000	203.000000	61295.620000

Figure 1

- We divided the recency and monetary into five levels using the qcut function from the pandas library. Furthermore, we divided the frequency into five levels manually.
- As stated in the competition report we added a new attribute **RFM_CELL** which consists of all the three rfm scores, and we also added the **RFM_SCORE** attribute which is the average of all three scores.

- Lastly, based on the **RFM_SCORE** we added a new attribute **CustomerLabel** which can have one of the following values:
    1 - Very rare customer
    2 - Customer at risk
    3 - Average customer
    4 - Promising customer
    5 - Loyal and best customer

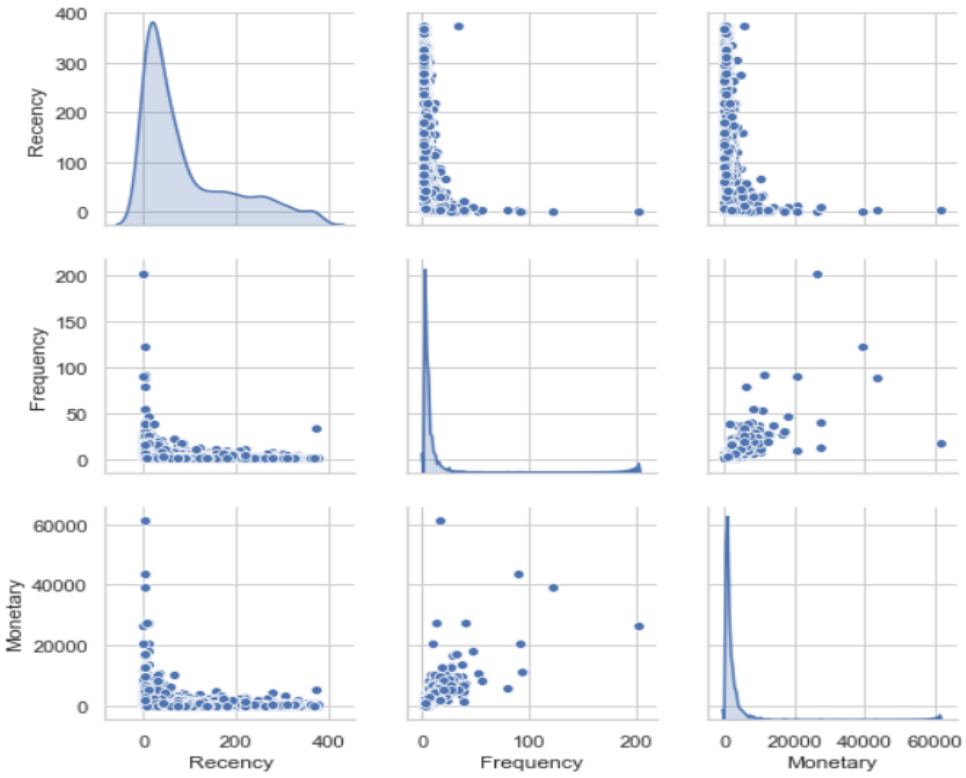

Figure 2

- After getting the RFM Score, we found out that the data is right-skewed as seen in figure 2, thus we applied **log** on the above three columns.
- Then we applied **StandardScaler** on log values to standardize it.

Clustering algorithm:

- Firstly, we have used the K-means clustering algorithm on the dataset. To get the optimal K value we have used the range from 1 to 15 and observed where the error remains less and therefore used the K equals 5 as it was more efficient.
- Lastly, using K-means with k=5 we assigned each customer to one of the five clusters.

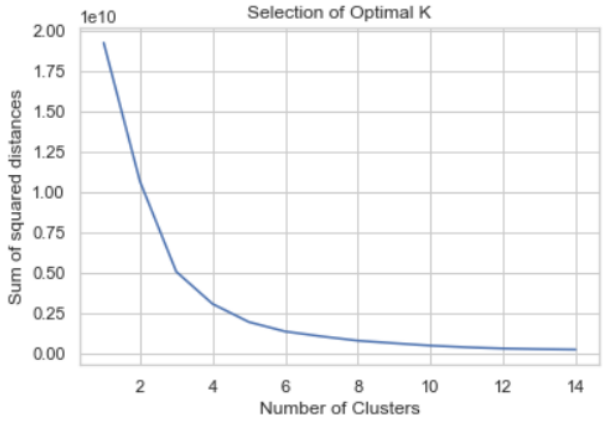

Figure 3

- After Clustering, we calculated the mean, median, minimum, maximum for each cluster for RFM parameters.

Cluster 1

	Mean	Minimum	Maximum	Median
Recency	33.792722	1.00	334.0	18.000
Frequency	8.145570	1.00	37.0	7.000
Monetary	2661.084748	1588.04	4891.4	2453.195

Cluster 2

	Mean	Minimum	Maximum	Median
Recency	15.747967	1.00	373.00	8.00
Frequency	20.455285	4.00	93.00	18.00
Monetary	7159.418374	4910.92	13833.57	6505.29

Cluster 3

	Mean	Minimum	Maximum	Median
Recency	3.333333	2.00	5.00	3.00
Frequency	76.333333	17.00	123.00	89.00
Monetary	47994.486667	39264.81	61295.62	43423.03

Cluster 4

	Mean	Minimum	Maximum	Median
Recency	5.666667	1.00	12.00	3.00
Frequency	54.777778	10.00	203.00	31.00
Monetary	21215.942222	16521.08	27597.28	20507.58

Cluster 5

	Mean	Minimum	Maximum	Median
Recency	108.269784	1.0	374.00	65.000
Frequency	2.325376	1.0	39.00	2.000
Monetary	515.449040	1.9	1581.23	394.545

Figure 4

- For calculating the customer ratio, we divided the total number of customers in each cluster by the total number of customers.

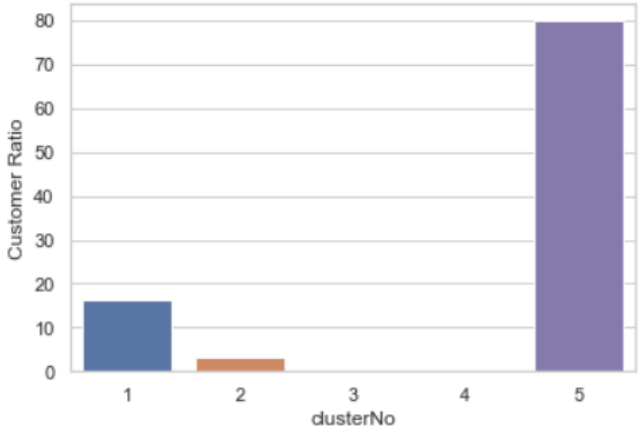

Figure 5

- Lastly, we have not applied any other approaches for the clustering algorithm as of now.