# Assignment 3

Kazi Amit Hasan (ID: 20341105)*, Anirban Dey (ID: 20417747)*, Tahosina Monir (ID: 20441144)*
School of Computing, Queen's University
Kingston, ON, Canada
{kaziamit.hasan, 22rf58, tahosina.monir}@queensu.ca
*All authors contributed equally to this assignment.

## I. Used software packages and datasets

For the completion of this assignment, Python [1] was selected as the preferred programming language. Python can be obtained from the official Python website. To augment Python's data analysis capabilities, a variety of libraries were used. While some of these libraries are readily available as they are commonly used in the Python community, others may require installation using pip/pip3, which is a built-in library installer in Python. To install a library, one can execute the command "pip install name_of_library" in the command line, with "name_of_library" representing the specific library that needs to be installed.

### A. Software packages

The software packages used to the analysis are given below:

- Python 3.12.1 [1]
- Jupyter notebook [2]
- Pandas [3]
- Matplotlib [4]
- yellowbrick [5]
- scikit-learn [6]

### B. Datasets

The dataset provided for this assignment was a 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.

## II. How to run

The codes and datasets are available in our GitHub repository [1]. This repository is currently private. We will publish this once we finish the course. For grading purpose, we are providing all the codes and datasets into a zip file in OnQ submission. In order to run our codes, please run the followings step by step:

- pip install -r requirements.txt

Make sure you put all the data and notebook files in the same directory.

## III. Analytical Process

### A. Exploration

First of all, we loaded the excel file in a dataframe and performed basic statistics and data sanity check. We found that the attributes were already in their preferred data types. This reduces our work on transforming the data types into suitable form.

Upon examining the 'Quantity' and 'Price' features, we observe the presence of negative values. These values suggest that some products were returned. Furthermore, a comparison between the 75th percentile and the maximum values reveals a significant disparity, indicating the existence of outliers in our dataset. We also checked the unique values in our dataset and found that there are a total of 25900 unique invoices, 4070 unique products, 4372 unique customers and 38 different countries are present. This ensures the diversity of the dataset.

Upon checking the null values, we found that there's lot of customers without Customer ID. To be specific, 135080 and 1454 null values were present in customer id and description of the dataset.

We also plotted the correlation heatmap to get more insights about the data. We noticed that the cell at the intersection of Quantity and UnitPrice shows a very small negative correlation (-0.0012). This suggests that there's almost no linear relationship between these two variables, but where it exists, it's slightly negative, implying that as quantity increases, unit price might decrease very slightly, and vice versa. Also, the cell at the intersection of Quantity and CustomerID shows a negative correlation (-0.0036). This value is also very close to zero, indicating a negligible linear relationship between Quantity and CustomerID.

After checking the correlation, we dropped the rows which has negative quantity and price. Also, we removed the duplicate data from the dataset. After following these steps the dataset shape is (524878, 8).

### B. Preparation

We can call this RFM construction phase. The initial phase in constructing an RFM model involves attributing Recency, Frequency, and Monetary values to every customer. Therefore, RFM analysis can only be performed on customers who possess a customerID. The RFM metrics are:

- **Recency:** This metric represents the duration since the customer's last transaction, measured in days.
- **Frequency:** This signifies the total count of transactions made by a customer.
- **Monetary:** This refers to the cumulative amount spent by the customer across all transactions.

Each customer needs to be assigned a score from 1 to 5 for recency, frequency, and monetary value. The following describes the transformation of these columns into RFM scores

[1] https://github.com/AmitHasanShuvo/cisc_839_assignments

ranging from 1 to 5, with '5' being the highest and '1' the lowest.

- For monetary value, a higher score of '5' is assigned to a higher value, indicating more spending.
- In terms of recency, a lower value signifies more recent purchases, thus it is given a higher score of '5'.
- Frequency follows the same pattern as monetary value; the more frequent the purchases, the higher the score.

After following this step, we divided the customers into nine categories to understand their characteristics.

- **Champions:** This category represents best customers who buy often and spend a lot. They are recent shoppers.
- **Loyal Customers:** Buy frequently but might not have made a purchase very recently or might not spend as much.
- **Recent Customers:** Have made a purchase recently but might not buy often or spend a lot.
- **Potential Loyalists:** Recent customers with relatively frequent purchases, indicating they're becoming more engaged.
- **Big Spenders:** Customers who spend the most, regardless of how recent or frequent their purchases are.
- **At Risk:** Were once frequent shoppers but haven't made purchases recently.
- **Can't Lose Them:** High-value customers who have not made a purchase in a long time.
- **Hibernating:** Customers who haven't shopped in a long time and used to shop infrequently.
- **Lost:** The lowest engagement level across all RFM measures.

Upon reaching this step, we analyzed two ways of assigning the categories to customers. One way is summing the recency, frequency and monetary scores and treating the sum as a value. Another is treating each scores as string. **We analyzed the benefits of two ways and decided to treat the scroes as string.** The reason behind treating as string is the summing method may not differentiate well between customers with the same total score but different patterns of behavior. For example, a customer who scores "1" in recency (recent purchase), "5" in frequency (buys frequently), and "5" in monetary (spends a lot) will have the same total score as a customer with a "5" in recency (hasn't purchased recently), "3" in frequency, and "3" in monetary. These two customers have very different behaviors and potentially require different marketing strategies. We can handle this problem if we treat the scores as string.

After applying these categorization into main dataset, we ended up with some characteristics of customers.

According to figure 1, we can conclude the following:

1) Champions: These are the best customers, who have made purchases very recently (mean recency of 11 days), frequently (average frequency of almost 9 times), and have spent significantly (average monetary value of about 4832). Their spending ranges from 41.99 to an exceptional 280206.02. The large range in monetary values suggests that this group includes both regular and high-stakes spenders. This is the largest segment with 1269 customers. Company can design special offers for this category.
2) At risk: Customers who have not made a purchase recently (mean recency of 140 days), with their most recent purchase ranging from 72 to 366 days ago. They have a low to moderate purchase frequency (averaging about 3.78 times) and have spent an average of approximately 937, with individual expenditures ranging widely from 136 to 2011.88. There are 185 customers in this segment.
3) Big Spenders: Customers who have made purchases somewhat recently (mean recency of 91 days) and with a frequency ranging from once to six times (average frequency of about 3.39). They are characterized by their high spending (average monetary value of roughly 4340), with individual amounts ranging from 2055.51 to an impressive 77183.60. This segment contains 140 customers.
4) Hibernating: Customers in this segment have been inactive for a considerable time (mean recency of 217 days), with low purchase frequency (average of 1.10 times), and lower spending (mean monetary value of approximately 365). The expenditures in this group range from 3.75 to 2044.37, indicating some variation in the amount spent when they do purchase. This segment includes 1055 customers.
5) Lost: These customers have an average recency similar to Big Spenders (mean of 69 days), but their frequency (mean of 2.14) and monetary values (mean of 604) are lower. The range of spending is quite broad, from 6.20 to 2053.02, which suggests variability in the amount they've spent historically. This is the second-largest segment with 1505 customers.
6) Loyal Customers: This segment represents highly frequent shoppers (mean frequency of 8.55 times), who have made their most recent purchase within the last 33 to 372 days (mean recency of 77 days). They also tend to spend a lot (mean monetary value of about 3695), with individual spending ranging from 70.02 to 80850.84. There are 184 customers in this segment.

| | recency | | | frequency | | | monetary | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | min | max | mean | min | max | mean | min | max | count |
| customer_segement | | | | | | | | | | |
| At Risk | 139.978378 | 72 | 366 | 3.778378 | 3 | 5 | 936.808054 | 136.00 | 2011.88 | 185 |
| Big Spenders | 91.107143 | 14 | 336 | 3.385714 | 1 | 6 | 4340.542436 | 2055.51 | 77183.60 | 140 |
| Champions | 10.805359 | 0 | 32 | 8.993696 | 1 | 209 | 4832.127147 | 41.99 | 280206.02 | 1269 |
| Hibernating | 217.122275 | 72 | 373 | 1.096682 | 1 | 2 | 364.997072 | 3.75 | 2044.37 | 1055 |
| Lost | 68.902326 | 14 | 373 | 2.136213 | 1 | 5 | 604.133603 | 6.20 | 2053.02 | 1505 |
| Loyal Customers | 77.331522 | 33 | 372 | 8.554348 | 6 | 63 | 3695.440924 | 70.02 | 80850.84 | 184 |

Figure 1: Statistics of recency, frequency and monetary value among the customer segments

## IV. Modeling

We plan to use unsupervised machine learning to categorize customers into distinct groups or clusters based on their

Table I: Distribution of customers within each cluster (Ratio & Percentage)

| Clusters | Customers | Ratio | Percentage |
|---|---|---|---|
| 0 | 1877 | 0.52 | 52.08% |
| 1 | 819 | 0.23 | 22.72% |
| 2 | 908 | 0.25 | 25.19% |

purchasing patterns. The formation of these clusters will be influenced by three key factors - recency, frequency, and monetary values of their purchases. Firstly, we checked the outliers and removed them using IQR method.

### A. K Means clustering

Recency, Frequency, and Monetary values can vary significantly across customers, with each metric potentially operating on a different scale. Without scaling, a metric like Monetary, which typically exhibits higher numeric values, could disproportionately influence the clustering outcome, overshadowing the variance captured by Recency and Frequency. Scaling ensures that each metric contributes equally to the distance calculations that underpin the K-Means algorithm, facilitating a more balanced and meaningful segmentation of the customer base. We scaled the RFM features using stander scaler function. Following data scaling, the visualize_elbow_method

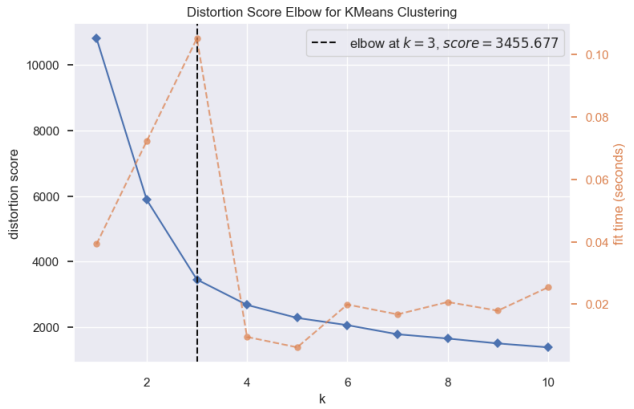


Figure 2: Distortion score elbow for KMeans clustering

function applies the elbow method to determine the optimal number of clusters. This method is instrumental in K-Means clustering as it provides a data-driven approach to selecting the number of customer segments. By plotting the within-cluster sum of squares (WCSS) against a range of cluster numbers and identifying the 'elbow' point where the rate of decrease sharply changes, we can infer the most appropriate number of clusters that balances between minimizing WCSS and avoiding overfitting through excessive segmentation. In this case, the elbow was identified at k=3 (shown in figure 2), suggesting an optimal segmentation into three distinct customer groups.

Table I represents the distribution of customers within each cluster (Ratio & Percentage).


Figure 3: Customer Distribution per Cluster

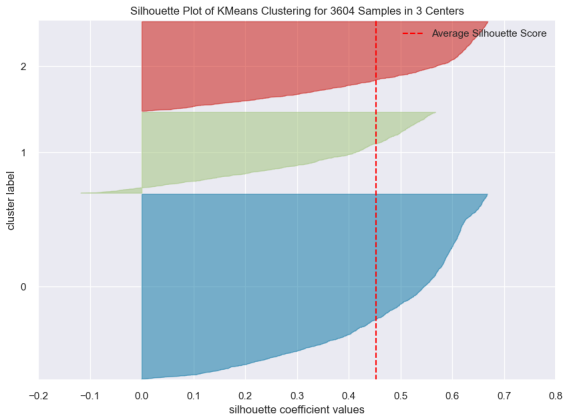


Figure 4: Silhouette plot of KMeans clustering.

### B. Silhouette method

The Silhouette score, which ranges from -1 to 1, measures the similarity of an object to its own cluster (cohesion) compared to other clusters (separation). A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. We also implemented that are got the score 0.46 which can be considered as optimal.

## V. Cluster Assessment and Explanation

To assess the clusters, we did cluster profiling and tried to label the clusters based on the data characteristics.

### A. Cluster 0: Infrequent, Lower-Mid Value Customers

This cluster consists of the largest number of customers who have interacted with the business relatively recently (average recency of 49 days). They have interacted infrequently (on average just over 2 times) and spent a lower-mid range of money (average of approximately 557).

After grouping the clusters with customer segments, we observed Cluster 0 is the largest cluster and is characterized by relatively recent interactions with the business, infrequent transactions, and moderate spending. Predominantly made up of 'Lost' customers (1210), suggesting many are at risk of churning despite recent engagement. 'Champions' (330) show

that there is a solid core of customers who, despite infrequent purchases, are potentially loyal and valuable. A significant number of 'Hibernating' customers (258) indicate a group that could be targeted to increase transaction frequency. The presence of a few 'At Risk' (77) and very few 'Big Spenders' and 'Loyal Customers' (1 each) implies opportunities for targeted promotions to increase spend and loyalty.

### *B. Cluster 1: Recently Engaged, Higher Value Customers*

Customers in this cluster are more recent and frequent in their purchases than those in Cluster 0 and spend considerably more.

After grouping the clusters with customer segments, we observed that the cluster is dominated by 'Champions' (494), showing a strong base of highly engaged and potentially loyal customers. A considerable number of 'Loyal Customers' (104) and 'Big Spenders' (76) underscore the higher value and loyalty within this group. 'At Risk' (48) customers may need immediate attention to retain their engagement level. The smaller presence of 'Lost' customers (97) indicates a higher retention rate within this cluster.

### *C. Cluster 2: Inactive, Lower Value Customers*

These customers are the least active both in terms of recent interactions and purchase frequency, and they also spend less on average compared to other clusters.

After grouping the clusters with customer segments, we observed that customers in this cluster are the least active and have the lowest average spend, indicating a segment with significant re-engagement potential. A majority are 'Hibernating' (654), which reflects the cluster's overall lower activity and engagement. 'Lost' customers (190) are the second-largest segment, reinforcing the need for reactivation strategies within this cluster. 'At Risk' (59) customers may be on the verge of becoming 'Lost' or 'Hibernating' and could benefit from timely re-engagement initiatives. The few 'Big Spenders' and 'Loyal Customers' (1 and 4, respectively) suggest that even among less active customers, there are opportunities to capitalize on their higher spending habits.

### *D. Observations*

1) The 'Champions' within Cluster 0 are engaging infrequently, which seems at odds with the typical definition of 'Champions' as recent, frequent, and high-spending customers. It's possible that the 'Champions' in Cluster 0 have high monetary scores compensating for their lower frequency. This may reflect a purchasing pattern where customers make significant purchases in fewer transactions. Alternatively, this could indicate a recent change in customer behavior where previously frequent customers have become less so, yet their historical data still categorizes them as 'Champions.'
2) Clusters contain a mix of RFM segments such as 'At Risk' and 'Champions,' suggesting complex behaviors that single RFM labels may not fully capture. This diversity may be due to the multidimensional nature of customer behavior and the limitation of using three variables to capture this complexity. Customers may shift between segments over time, or the segments may overlap due to close scoring thresholds. A deeper analysis could reveal lifecycle stages or transaction patterns that explain the presence of different segments within the same cluster.
3) 'Lost' customers are present in Cluster 1, which is otherwise characterized by recent and frequent purchases. The 'Lost' customers in Cluster 1 could be recent one-time purchasers with high monetary value who have not returned. Their classification as 'Lost' could be premature if not enough time has passed to observe their potential for repeat purchases. This underlines the dynamic nature of customer engagement and the need for a time-sensitive approach to segmentation.
4) Despite being labeled as inactive and lower value, Cluster 2 includes 'Big Spenders' and 'Loyal Customers.' This could occur if the cluster contains a few high-value transactions that are not recent or frequent enough to shift the overall cluster average but still qualify under the 'Big Spenders' criteria. As for 'Loyal Customers,' they may have been frequent in the past but have not made recent purchases, reflecting a need to update the segmentation model to account for changes in customer behavior over time.

## REFERENCES


[1] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[2] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds. IOS Press, 2016, pp. 87 – 90.

[3] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[4] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] B. Bengfort and R. Bilbro, "Yellowbrick: Visualizing the Scikit-Learn Model Selection Process," vol. 4, no. 35, 2019. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01075

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.