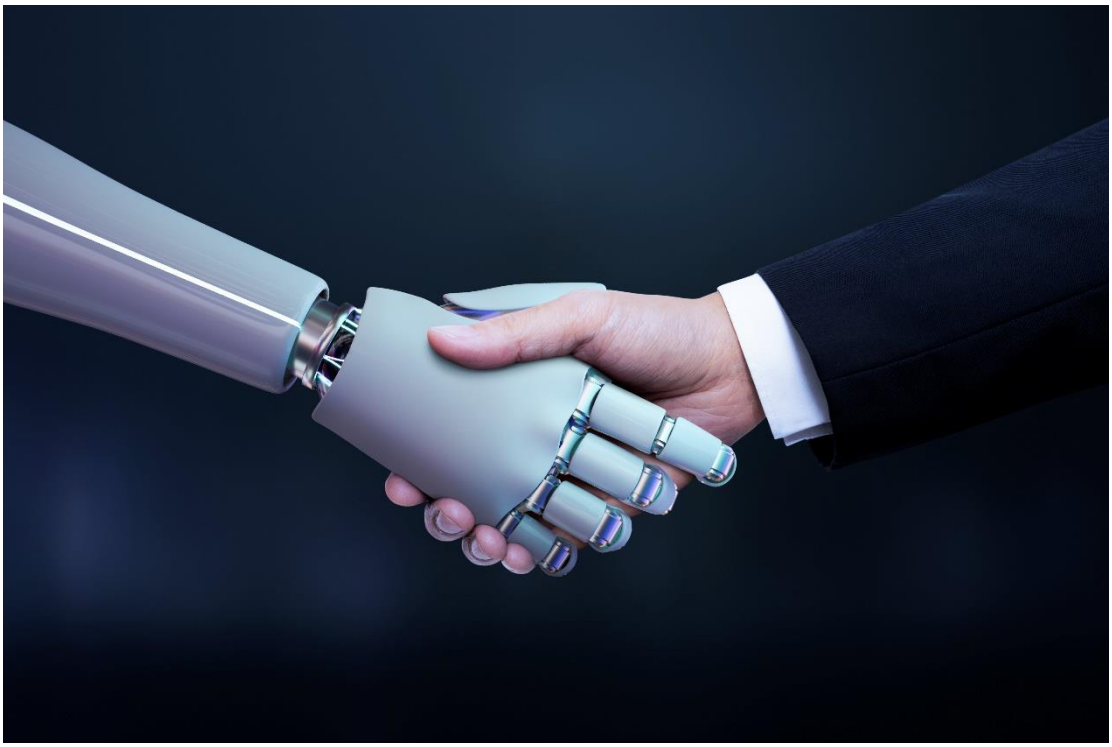
Big Data Analysis with Machine Learning in E-commerce



CETM 24 Data Science Fundamentals

Mahmoud Al-Mubaideen

Student ID: 219217301

Supervised by: Dr. Basel Barakat

University of Sunderland

Word Count 2089

1. Introduction:

Big data has been a hot topic in both academia and the e-commerce industry in recent years. According to research, firms who include big data analytics (BDA) into their value chain see productivity gains of 5–6 percent compared to those that don't. **(McAfee and Brynjolfsson, 2012).**

Emerging internet-based technologies (e.g., real-time customer assistance, dynamic pricing, tailored offers, or increased engagement) deliver dramatic benefits to e-commerce enterprises **(Riggins, 1999).** BDA has the potential to amplify these effects by providing the information necessary to make well-informed decisions **(Jao, 2013).**

In particular, "big data enables retailers to analyze each user's activity and connect the dots to find the most efficient methods to convert one-time consumers into repeat purchasers with machine learning applications" in the context of e-commerce **(Jao, 2013).** When combined with Big data analytics (BDA), e-commerce companies may use data more effectively, achieve a greater conversion rate, improve decision making, and empower customers **(Miller 2013).**

Nothing works better together than eCommerce and new technology; in fact, without technical advancement, we wouldn't have eCommerce at all. The most innovative eCommerce technology is likely to be one based on machine learning (ML) and Big Data Analysis (BDA).

To put it simply, Machine Learning is the ability of a machine to mimic human logic. Machine learning is a computer's capacity to find how to complete tasks automatically without being prompted to do so. Fundamentally, Big Data Analysis assists Machine Learning in determining how to handle a data collection.

The benefits of ML and BDA for eCommerce are amazing. The truth is that it can calculate data and forecasts considerably faster than an entire team of humans! It includes everything from greater insights and forecasts to improved customer service and remarketing, all of which led to considerable revenue increase.

2. Supervised vs Unsupervised learning models:

2.1. Supervised Learning:

The explanatory and dependent variables are clearly separated in Supervised Learning models. Models are trained to explain dependent variables using explanatory factors. In other words, we already know what the model's output attributes are. [Gupta, B., 2017].

2.2. Unsupervised Learning:

No target qualities or model outputs exist in unsupervised learning: explanatory and dependent variables are not differentiated. The models are built to discover the data's fundamental structure. [Gupta, B., 2017].

3. Machine Learning Algorithms:

We will go through different machine learning algorithms:

3.1. Clustering Techniques:

When a dataset is broken down into separate, distinct clusters, the method of clustering can be used.

Let's pretend there are 1000 rows of client information in our database. Clustering may be used to separate customers into various groups or segments depending on a variety of characteristics. Demographic data and shopping patterns are only two examples of what may be included in customer data.

A clustering approach that does not rely on any prior knowledge is called unsupervised learning. The method determines the output rather than training on past input-output data. [Gupta, B., 2017].

3.2. Neural Networks:

A neural network (also known as an Artificial Neural Network) is inspired by the human nervous system and how it absorbs and processes complicated information. Neural networks, like people, learn by example and are tailored to a certain purpose.

In order to find patterns in large amounts of data, neural networks are employed. These networks may then make predictions and classify data points. Neural networks are usually organised into layers. [Gupta, B., 2017].

3.3. Decision Trees:

As the name implies, decision trees are a tree-shaped visual depiction of how to make a certain conclusion by setting out all possibilities and their likelihood of occurrence.

Decision trees are relatively simple to comprehend and interpret. At each node of the tree, one may analyze the ramifications of picking that node or choice [Gupta, B., 2017].

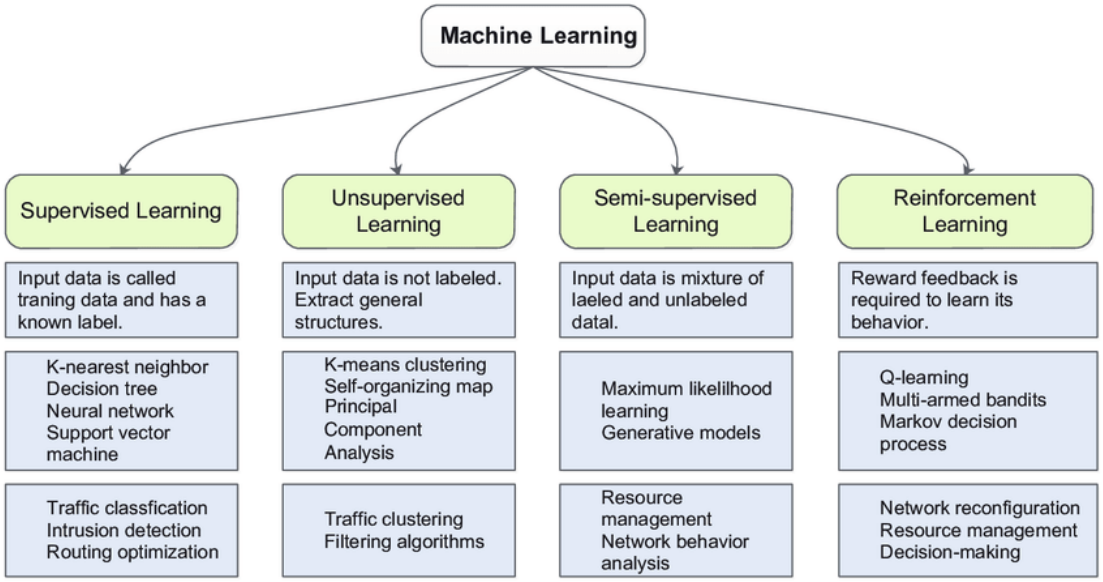


Figure (1): Table of All Machine Learning Algorithms [Liu, Y., 2020.]

4. Anomaly Detection:

Machine learning, time series, neural network anomaly detection, supervised and unsupervised outlier identification algorithms and so on are used in big data management and data science to identify fraud and other abnormal occurrences such as skewed data sets. [Valcheva, S., 2020].

Real-world datasets are full with anomalies and outliers. Data corruption, experimental error, or a blunder by a person can all result in anomalies. Analyses may be affected by the presence of anomalies, hence the dataset must be free of anomalies in order to train a robust data science model [Kumar, S., 2021].

4.1. What are Anomalies?

In statistical analysis, anomalies are data points that are out of the ordinary and do not follow the expected pattern. In contrast to the dataset's typical behaviour, these data points or observations are outliers. [Kumar, S., 2021].

Anomaly detection is a technique for detecting abnormalities in a dataset that is performed unsupervised [Kumar, S., 2021]. Anomaly can be roughly divided under the following categories:

- Outliers: are short/small aberrant patterns that arise in data collecting in an ad hoc manner.
- Change in Events: A systematic or abrupt departure from prior typical behavior.
- Drifts are long-term changes in data that are slow and unidirectional.

Anomalies identification is extremely important for detecting fraudulent transactions, sickness diagnosis, and dealing with case studies with high-class imbalance.

Techniques for detecting anomalies can be utilized to create more robust data science models. [Kumar, S., 2021].

4.2. Type of Anomaly Detection Algorithms:

We will go through different unsupervised machine learning techniques for detecting anomalies and evaluate one of the techniques on the Olist Company dataset:

4.2.1. Isolation Forest:

Isolation Forest is an unsupervised anomaly detection approach that detects outliers in a dataset using a random forest algorithm (decision trees). [Pedregosa *et al.*, 2011] The method attempts to split or divide the data points so that each observation is separated from the rest.

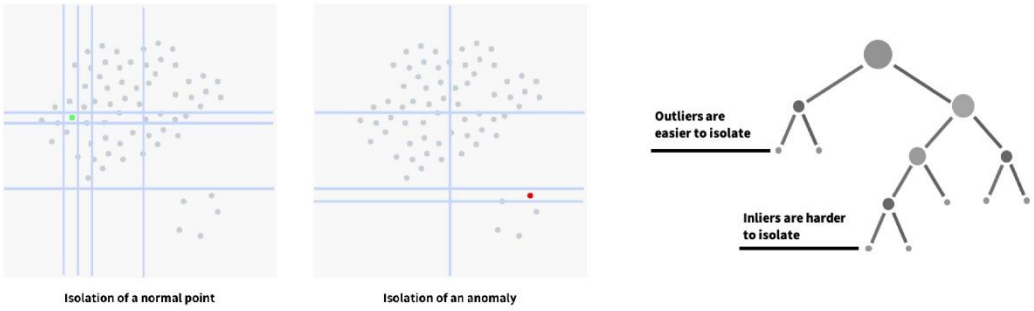


Figure (2): Partitioning of Anomaly and Regular data point [Pedregosa *et al.*, 2011].

Figure(2) show that normal data points require a disproportionately greater number of divisions than aberrant data points [Pedregosa *et al.*, 2011].

4.2.2. Local Outliner Factor:

Another anomaly identification approach is the Local Outlier Factor, which considers the density of data points to determine whether or not a point is an abnormality. The local outlier factor computes an anomaly score, which indicates how separated the place is from the surrounding neighborhood. The local outlier component generates an anomaly score, which reflects how isolated the location is from its surroundings. [Pedregosa *et al.*, 2011].

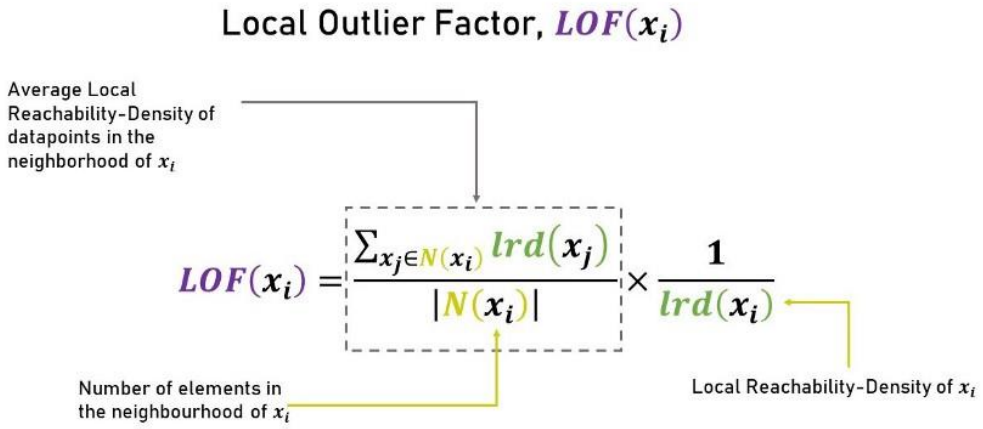


Figure (3): Local Outlier Factor Formulation. [Mahto, P., 2020]

4.2.3. K-means:

K-means is a widely used clustering technique in data mining. It generates k groups from a set of things in order to make the constituents of a group more similar. To recap, cluster algorithms are meant to create groups with individuals that are more similar. Clusters and groups are interchangeable in this context. The K-means algorithm clusters data objects based on feature similarity. One of the most significant advantages of k-means is its ease of implementation. K-means has been effectively implemented in the majority of the common computer languages used in data science [Valcheva, S., 2020].

4.2.4. Why is k-means clustering used in this report?

It's really easy to understand and understand. K-means is more efficient than Hierarchical clustering for numerous variables in the dataset. In addition, an instance can alter the cluster and operate on unlabeled numerical data while redetermining the cluster centre. The greatest results are obtained when datasets are well distinct (thoroughly segregated) from one another since it is rapid, robust, and simple to interpret.

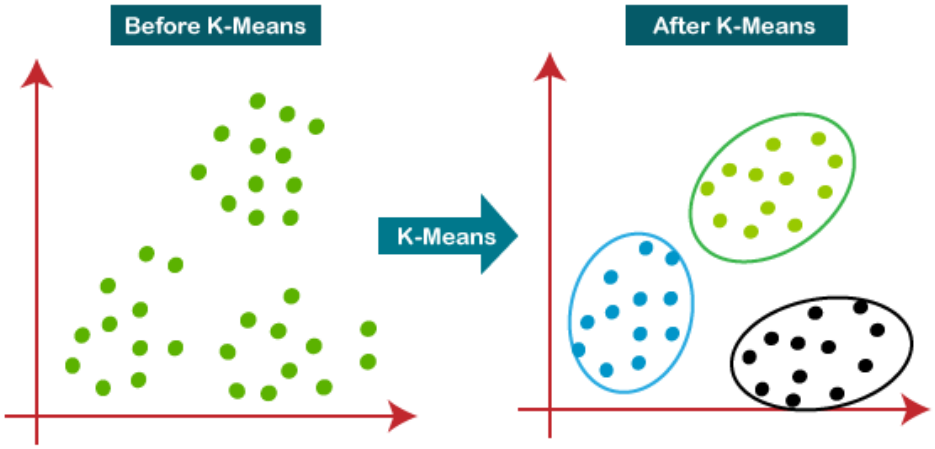


Figure (4): K-Means Effect before and after [Sharma, P., 2021].

5. Dataset Overview:

This project makes use of the Online Retail dataset, which can be found at (https://archive-beta.ics.uci.edu/ml/datasets/online+retail+ii) This dataset covers all transactions for a UK-based and registered non-store internet retailer that occurred between 01/12/2009 and 09/12/2011. The firm mostly provides one-of-a-kind all-occasion giftware. Many of the company's clients are wholesalers. However, in the analysis of this dataset, this research only used data from 2010 to 2011.

5.1. Dataset Preparation:

```
## Observations: 541,910
## Variables: 8
## $ Invoice     <chr> "536365", "536365", "536365", "536365", "536365", ...
## $ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "...
## $ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL...
## $ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 6, 3, 3, 3, 32, 6, 6, 8...
## $ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12...
## $ Price       <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1....
## $ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 1...
## $ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdo...
```

Consider the following data summaries to get an overview:

```
##    Invoice            StockCode         Description
##  Length:541910      Length:541910      Length:541910
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity          InvoiceDate                      Price
##  Min.   :-80995.00   Min.   :2010-12-01 08:26:00   Min.   :-11062.06
##  1st Qu.:     1.00   1st Qu.:2011-03-28 11:34:00   1st Qu.:     1.25
##  Median :     3.00   Median :2011-07-19 17:17:00   Median :     2.08
##  Mean   :     9.55   Mean   :2011-07-04 13:35:22   Mean   :     4.61
##  3rd Qu.:    10.00   3rd Qu.:2011-10-19 11:27:00   3rd Qu.:     4.13
##  Max.   : 80995.00   Max.   :2011-12-09 12:50:00   Max.   : 38970.00
##
##    CustomerID       Country
##  Min.   :12346    Length:541910
##  1st Qu.:13953    Class :character
##  Median :15152    Mode  :character
##  Mean   :15288
##  3rd Qu.:16791
##  Max.   :18287
```

As seen in the dataset overview, numerous actions must be conducted to prepare the dataset before clustering the data. Then, before to clustering, some columns must be recoded to factor and alter the date and time to date type alone, as seen in the code below:

```
##     Invoice         StockCode      Description          Quantity
##  576339 :   542   85123A :  2035   Length:397885      Min.   :    1.00
##  579196 :   533   22423  :  1723   Class :character   1st Qu.:    2.00
##  580727 :   529   85099B :  1618   Mode  :character   Median :    6.00
##  578270 :   442   84879  :  1408                      Mean   :   12.99
##  573576 :   435   47566  :  1396                      3rd Qu.:   12.00
##  567656 :   421   20725  :  1317                      Max.   :80995.00
##  (Other):394983   (Other):388388
##   InvoiceDate            Price            CustomerID
##  Min.   :2010-12-01   Min.   :   0.001   17841  :  7847
##  1st Qu.:2011-04-07   1st Qu.:   1.250   14911  :  5675
##  Median :2011-07-31   Median :   1.950   14096  :  5111
##  Mean   :2011-07-10   Mean   :   3.117   12748  :  4595
##  3rd Qu.:2011-10-20   3rd Qu.:   3.750   14606  :  2700
##  Max.   :2011-12-09   Max.   :8142.750   15311  :  2379
##                                          (Other):369578
##           Country
##  United Kingdom:354321
##  Germany       :  9040
##  France        :  8342
##  EIRE          :  7236
##  Spain         :  2484
##  Netherlands   :  2359
##  (Other)       : 14103
```

5.2. RFM Data Preparation:

To do the RFM Analysis, the data collection must be processed further as follows:

1- Use the day after the latest InvoiceDate as a reference date.
2- To obtain recency data, find the most recent transaction date and compute the day to the reference date per client.
3- Determine the number of transactions per client and label them as frequency values.
4- Total expenditure per client expressed in monetary terms.

```
## # A tibble: 6 x 4
##   CustomerID recency frequency monetary
##   <fct>        <dbl>     <int>    <dbl>
## 1 12346          326         1   77184.
## 2 12347            3         7    4310
## 3 12348           76         4    1797.
## 4 12349           19         1    1758.
## 5 12350          311         1     334.
## 6 12352           37         8    2506.
```

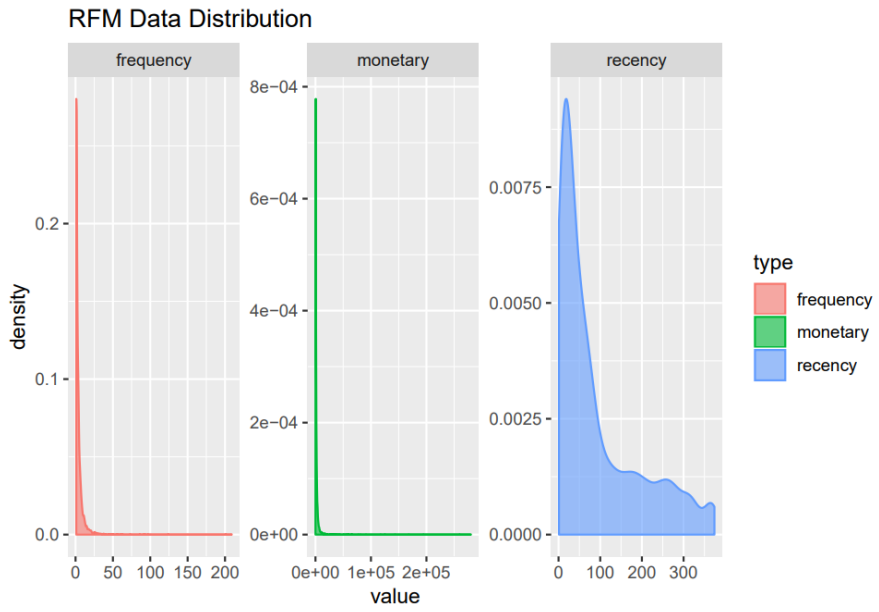


Figure (5): RFM Data Distribution

5.3. Analysis and Machine Learning Algorithms:

The K-Means approach was utilized in this project to find groupings across all customers. K-Means clustering is a form of unsupervised learning technique that groups point depending on their distance.

Transformation of Data As observed in the RFM Data Distribution graph, the data is substantially skewed, particularly in terms of frequency and monetary value. The log transformation can be used to make more sense of the data:

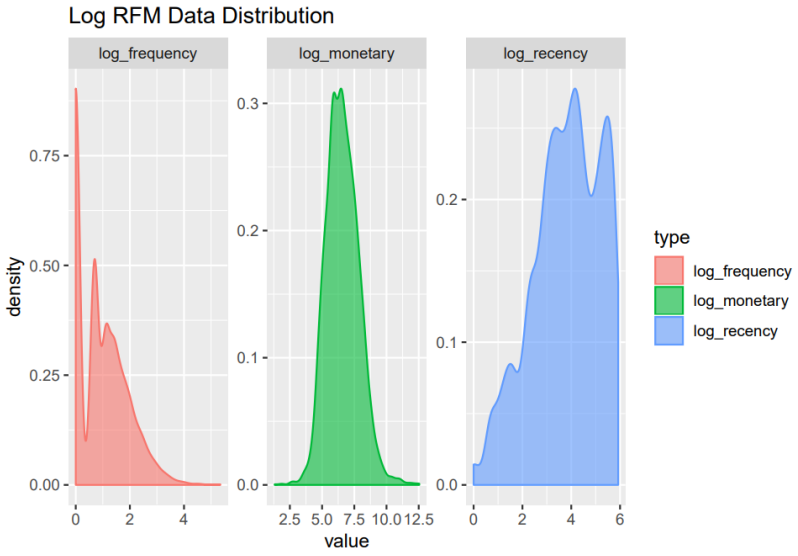


Figure (6): Log RFM Data Distribution

Furthermore, because the distance is employed in the K-Means approach, the unit scale is significant. As a result, standardization must be performed prior to clustering by determining the z-score of features. This may be accomplished by calculating using this formula:

$$z = \frac{x - \mu}{\sigma}$$

A z-score: indicates how many standard deviations a data point is above or below the mean. **[Khan Academy. 2022.]**

The scale function in R may be used to perform this computation.

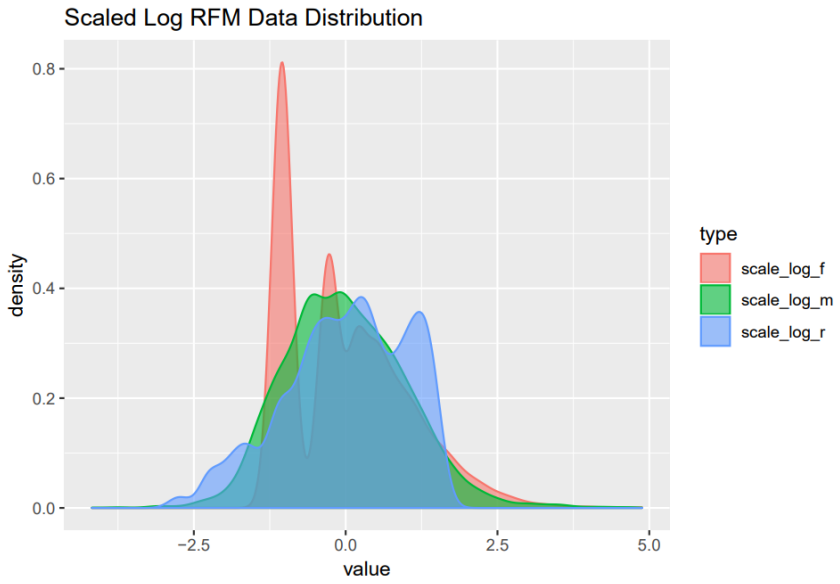


Figure (7): Scaled Log RFM Data Distribution

After standardization, all of the characteristics are on the same scale. So, now that the data has been preprocessed, the clustering procedure may be carried out using this scaled dataset.

5.3.1. Clustering with K-Means:

The initial stage in clustering is determining the appropriate number of clusters. Create an elbow curve and then select the best optimal cluster based on that curve. The elbow curve is created by visualizing the Sum Square Error (SSE) from the K-Means algorithm. This figure is the sum square value of each cluster's real point distance to the center.

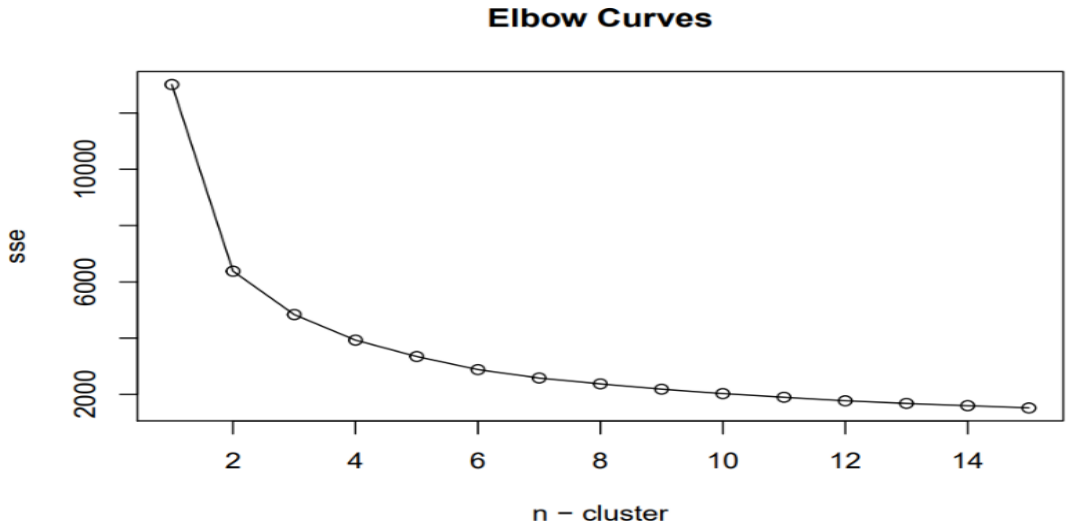


Figure (8): Elbow Curves Clustering

The most optimal cluster was shown as the elbow of the curve someplace SSE considerably decreased but not too much. In this scenario, 4 was chosen as the best cluster. After determining the cluster numbers, a model may be constructed and an actual cluster created, as seen below:

cluster	total	average_recency	average_frequency	average_monetary
1	833	22.18	1.90	483.53
2	1239	72.74	4.13	1743.12
3	762	11.13	13.01	7653.64
4	1504	190.56	1.27	343.64

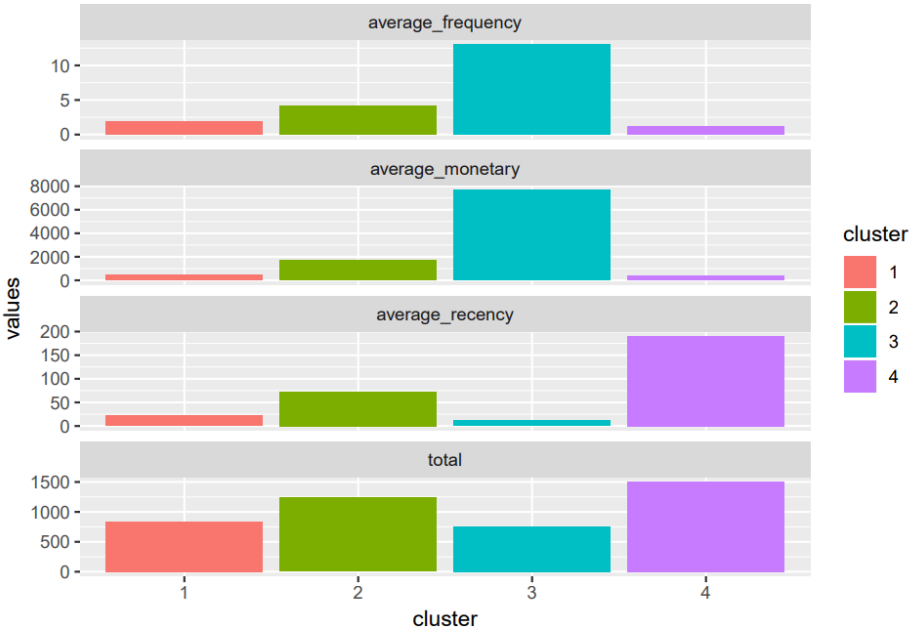


Figure (9): 4 FRM Clustering

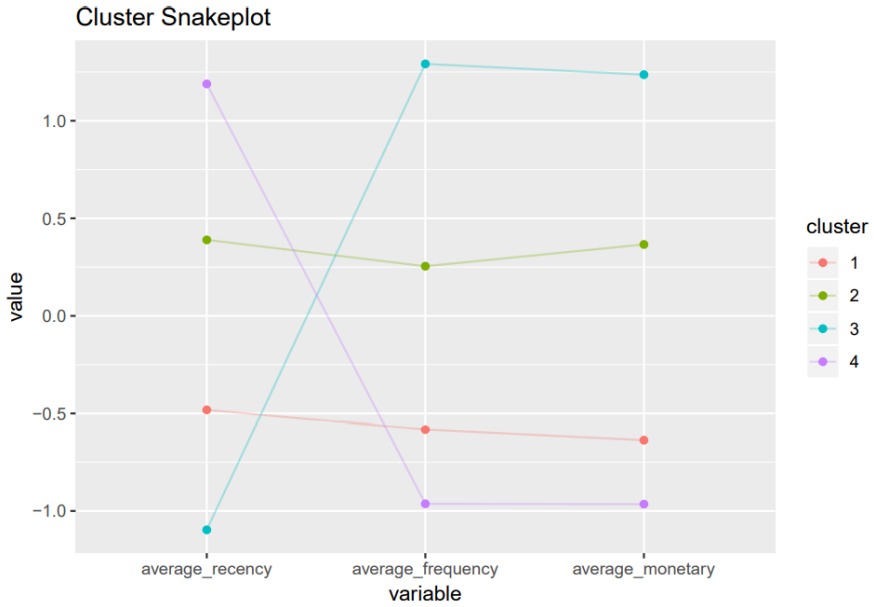


Figure (10): RFM Snakeplot Clustering

5.3.2. Results:

As a result, this initiative developed four different types of clients. Cluster 1 represents the most valued customer. Customers in this group have spent the most money and spent it the most frequently. Cluster 2 is less common and has a lower worth of money than Cluster 1. However, they have not lately transacted. Cluster 3 has lately had transactions; however, they were not frequent and only spent a tiny amount of money. This Cluster may be a new client who has just completed the purchase. Finally, cluster 4 has become our losing client with the least frequent and monetary value and has not transacted in a long time.

The following is an overview of the detail for each cluster:

cluster	total	average_recency	average_frequency	average_monetary
1	833	22.18	1.90	483.53
2	1239	72.74	4.13	1743.12
3	762	11.13	13.01	7653.64
4	1504	190.56	1.27	343.64

Finally, a cluster has been successfully constructed, and each Customer may be classified according to their recency, frequency, and monetary amounts. Furthermore, this cluster may be employed as a foundation for providing various treatments to provide greater advantage to the firm. Further analysis may be performed and other variables introduced, such as tenure or the number of days between the customer's initial transaction and the last day of their transaction. More extensive analysis may also be performed by specifying a time range, such as RFM for yearly, monthly, or weekly, to observe how our clients perform throughout that time period.

6. Conclusion

Big data analytics has emerged as the new frontier of innovation and competitiveness in the e-commerce business as a result of the information revolution's issues and potential. By translating data into insights for good decision making and problem resolution, big data analytics is fast delivering value to e-commerce firms. Data, sources, talents, and systems are all taken into consideration in order to get a competitive advantage. Leading e-commerce giants like Google, Amazon, eBay, ASOS, Netflix, and Facebook have already adopted big data analytics and achieved enormous success with this strategy. An in-depth investigation and development of taxonomies of the most significant elements of big data analytics.

7. Appendix

```
# Prepare Required
   library(tidyverse)
   library(readxl)
   library(GGally)
   library(lubridate)

 # Import Dataset
   files <- tempfile()
   download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/00502/online_retail_II.xlsx", files)
   df <- read_excel(files, sheet = 'Year 2010-2011', col_names = TRUE)
   df <- df %>% rename(CustomerID = `Customer ID`) # Rename CustomerID Column Names
   glimpse(df)

   Summary(df)

   # Remove Rows that have Negative Values of Quantity and Price
   clean_df <- df %>%
     filter(Quantity > 0 & Price > 0) %>%
     drop_na()
   # Recode Dataset
   Recode_df <- clean_df %>%
     mutate(Invoice = as.factor(Invoice), StockCode = as.factor(StockCode),
            InvoiceDate = date(InvoiceDate), CustomerID = as.factor(CustomerID),
            Country = as.factor(Country))
   summary(Recode_df)

   # Get Analysis Reference Date (One Day After Last Transaction)
   Max_Date <- date(max(Recode_df$InvoiceDate)) + 1
   # Calculate RFM Values
   rfm_df <- Recode_df %>%
     group_by(CustomerID) %>%
     summarise(recency = as.numeric(Max_Date - max(InvoiceDate)),
               frequency = n_distinct(Invoice), monetary = sum(TotalSpend))
   head(rfm_df)


   rfm_df %>%
     gather(type,value,recency:monetary) %>%
     ggplot(aes(x = value, color = type, fill = type)) +
     geom_density(alpha = 0.6) +
     facet_wrap(~type, nrow = 1, scales="free") +
     labs(title = 'RFM Data Distribution')


   log_rfm <- rfm_df %>%
     mutate(log_recency = log(recency), log_frequency = log(frequency), log_monetary = log(monetary))
   log_rfm %>%
     gather(type,value,log_recency:log_monetary) %>%
     ggplot(aes(x = value, color = type, fill = type)) +
     geom_density(alpha = 0.6) +
     facet_wrap(~type, nrow = 1, scales="free") +
     labs(title = 'Log RFM Data Distribution')

   scale_df <- log_rfm %>%
     mutate(scale_log_r = scale(log_recency), scale_log_f = scale(log_frequency), scale_log_m = scale(log_monetary))
   scale_df %>%
     gather(type,value,scale_log_r:scale_log_m) %>%
     ggplot(aes(x = value, color = type, fill = type)) +
     geom_density(alpha = 0.6) +
     labs(title = 'Scaled Log RFM Data Distribution')

```

```
62.
63.    # Iterate from 1 to 15 to find the most optimum cluster by elbow curve
64.    set.seed(100)
65.    used_var = c("scale_log_r","scale_log_f","scale_log_m")
66.    sse <- sapply(1:15,
67.                  function(k)
68.                  {
69.                    kmeans(x=scale_df[used_var], k, nstart=25)$tot.withinss
70.                  }
71.    )
72.    plot(sse, type = "o", xlab = "n - cluster", main = 'Elbow Curves')
73.
74.
75.    # Calculate Cluster group with 4 cluster
76.    segment_4 <- kmeans(x=scale_df[used_var], 4, nstart=25)
77.    cluster <- as.factor(segment_4$cluster)
78.    rfm_clustered4 <- cbind(scale_df,cluster)
79.    rfm_clust4_summary <- rfm_clustered4 %>%
80.      group_by(cluster) %>%
81.      summarise(total = n_distinct(CustomerID),
82.                average_recency = round(mean(recency),2),
83.                average_frequency = round(mean(frequency),2),
84.                average_monetary = round(mean(monetary),2)
85.      )
86.    rfm_clust4_summary %>% knitr::kable()
87.
88.    rfm_clust4_summary %>%
89.      gather(key = 'measure', value = 'values',c(5,4,3,2)) %>%
90.      ggplot(aes(x = cluster , y = values, fill = cluster)) +
91.      8
92.    geom_col() +
93.      facet_wrap(~measure, ncol = 1, scales="free_y")
94.
95.    rfm_clustered4 %>%
96.      group_by(cluster) %>%
97.      summarise(average_recency = round(mean(scale_log_r),2),
98.                average_frequency = round(mean(scale_log_f),2),
99.                average_monetary = round(mean(scale_log_m),2)) %>%
100.       ggparcoord(columns = 2:4, groupColumn = 'cluster',
101.                  showPoints = TRUE,
102.                  alphaLines = 0.3, title = "Cluster Snakeplot")
103.
104.
105.
106.
107.     rfm_clust4_summary %>% knitr::kable()
```

8. Reference:

- Liu, Y., 2020. [online] Research Gate. Available at: <https://www.researchgate.net/profile/Yiming-Liu-10> [Accessed 12 May 2022].
- Gupta, B., 2017. 10 Machine Learning Algorithms every Data Scientist should know. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/10-machine-learning-algorithms-every-data-scientist-know/> [Accessed 9 May 2022].
- Valcheva, S., 2020. Anomaly Detection Algorithms: in Data Mining (With Comparison). [online] Blog For Data-Driven Business. Available at: <https://www.intellspot.com/anomaly-detection-algorithms/> [Accessed 12 May 2022].
- Kumar, S., 2021. 5 Anomaly Detection Algorithms every Data Scientist should know. [online] Medium. Available at: <https://towardsdatascience.com/5-anomaly-detection-algorithms-every-data-scientist-should-know-b36c3605ea16> [Accessed 11 May 2022].
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., 2011. *Scikit-learn: Machine Learning in Python*. [online] Jmlr.csail.mit.edu. Available at: <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html> [Accessed 14 May 2022].
- Mahto, P., 2020. *Local Outlier Factor: A way to Detect Outliers*. [online] Medium. Available at: <https://medium.com/mlpoint/local-outlier-factor-a-way-to-detect-outliers-dde335d77e1a> [Accessed 12 May 2022].
- Sharma, P., 2021. *K Means Clustering Simplified in Python | K Means Algorithm*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/> [Accessed 14 May 2022].
- McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D., 2012. Big data. The management revolution. Harvard Business Review.
- Riggins, F. J. 1999. A framework for identifying web-based electronic commerce opportunities. Journal of Organizational Computing and Electronic Commerce.
- Jao, J., 2013. Why big data Is A must In ecommerce. Available at: http:// www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce [Accessed 8 May March, 2022].
- Miller, G., 2013. 6 ways To use “big data” To increase operating margins By 60 %. Available at: http://upstreamcommerce.com/blog/ 2012/04/11/6-ways-big-data-increase-operating-margins-60-part-2 [Accessed 8 May, 2022]
- Daqing, C., 2019. *UCI Machine Learning Repository*. [online] Archive-beta.ics.uci.edu. Available at: <https://archive-beta.ics.uci.edu/ml/datasets/online+retail+ii> [Accessed 14 May 2022].
- Khan Academy. 2022. | *Khan Academy*. [online] Available at: <https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/z-scores/a/z-scores-review> [Accessed 23 May 2022].