# CUSTOMER SEGMENTATION ANALYSIS USING RFM AND K-MEANS IN PYTHON

### DATA CLEANING

The aim of this analysis is to classify customers based on their loyalty to the store. We will estimate the retency, frequency and monetary capacity of the customers and classify them into similar groups. To start with, we show in table 1 that the data contains 8 variables with approximately 541909 observation. We can see also from table 1 that only Description and CustomerID variables are missing each with 1454 and 135080 Description and customer ID respectively. To cater for the missing values, we have the option to replace with the averages, however the variables in concern are non-scalar (strings and nominal), thus we consider dropping all observations whose description and customer ID are missing as they are cogent identifier which will render the information provided by that observation baseless. Visual inspection of the data also made it evident that there are negative values (possibly due to imputation error) in the *Quantity* and *Unit Price* variables which we remove alongside the missing values. Thus the new dimension of the data is 524878 observation for the 8 variables. An overview of the data is shown in table 2.

Table 1: Variable Description

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Missing** |
| InvoiceNo | object | 0 |
| StockCode | object | 0 |
| Description | object | 1454 |
| Quantity | int64 | 0 |
| InvoiceDate | datetime | 0 |
| UnitPrice | float64 | 0 |
| CustomerID | float64 | 135080 |
| Country | object | 0 |

Table 2: Data overview

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **InvoiceNo** | **StockCode** | **Description** | **Quantity** | **InvoiceDate** | **UnitPrice** | **CustomerID** | **Country** |
| 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 01-12-10 8:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 01-12-10 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 01-12-10 8:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 01-12-10 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 01-12-10 8:26 | 3.39 | 17850 | United Kingdom |

#### DESCRIPTIVE ANALYSIS

Table 3

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Quantity** | **UnitPrice** |
| count | 524878 | 524878 |
| mean | 10.62 | 3.92 |
| std | 156.28 | 36.09 |
| min | 1 | 0 |
| 0.25 | 1 | 1.25 |
| 0.5 | 4 | 2.08 |
| 0.75 | 11 | 4.13 |
| max | 80995 | 13541.33 |

Table 3 above shows the Descriptive analysis of the quantitative variables in the data which include *Quantity* of product ordered and the *Unit Price* attached to each it is evident that the average quantity ordered is approximately 11 pieces while the average price for each product is $4 each. The standard deviation of both variable is 156.28 and 36.09 respectively which indicates that the distribution of data is non-normal and requires some transformation which will be catered for in the upcoming sections of this project.

Table 4: Countries with >1000 customers

|  |  |
| --- | --- |
| **Country** | **Customers** |
| United Kingdom | 349203 |
| Germany | 9025 |
| France | 8326 |
| EIRE | 7226 |
| Spain | 2479 |
| Netherlands | 2359 |
| Belgium | 2031 |
| Switzerland | 1841 |
| Portugal | 1453 |
| Australia | 1181 |
| Norway | 1071 |

DATA TRANSFORMATION (COHORT ANALYSIS)

Table 5: Active customers for each monthly cohort index

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CohortIndex** | | | | | | | | | | | | |
| **MonthlyCohort** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** |
| **01-12-10** | 885 | 324 | 286 | 340 | 321 | 352 | 321 | 309 | 313 | 350 | 331 | 445 | 235 |
| **01-01-11** | 417 | 92 | 111 | 96 | 134 | 120 | 103 | 101 | 125 | 136 | 152 | 49 |  |
| **01-02-11** | 380 | 71 | 71 | 108 | 103 | 94 | 96 | 106 | 94 | 116 | 26 |  |  |
| **01-03-11** | 452 | 68 | 114 | 90 | 101 | 76 | 121 | 104 | 126 | 39 |  |  |  |
| **01-04-11** | 300 | 64 | 61 | 63 | 59 | 68 | 65 | 78 | 22 |  |  |  |  |
| **01-05-11** | 284 | 54 | 49 | 49 | 59 | 66 | 75 | 27 |  |  |  |  |  |
| **01-06-11** | 242 | 42 | 38 | 64 | 56 | 81 | 23 |  |  |  |  |  |  |
| **01-07-11** | 188 | 34 | 39 | 42 | 51 | 21 |  |  |  |  |  |  |  |
| **01-08-11** | 169 | 35 | 42 | 41 | 21 |  |  |  |  |  |  |  |  |
| **01-09-11** | 299 | 70 | 90 | 34 |  |  |  |  |  |  |  |  |  |
| **01-10-11** | 358 | 86 | 41 |  |  |  |  |  |  |  |  |  |  |
| **01-11-11** | 323 | 36 |  |  |  |  |  |  |  |  |  |  |  |
| **01-12-11** | 41 |  |  |  |  |  |  |  |  |  |  |  |  |

Table 6: Retention Rate of customers for each monthly cohort

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CohortIndex** | | | | | | | | | | | | |
| **MonthlyCohort** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** |
| **01-12-10** | 1 | 0.37 | 0.32 | 0.38 | 0.36 | 0.4 | 0.36 | 0.35 | 0.35 | 0.4 | 0.37 | 0.5 | 0.27 |
| **01-01-11** | 1 | 0.22 | 0.27 | 0.23 | 0.32 | 0.29 | 0.25 | 0.24 | 0.3 | 0.33 | 0.37 | 0.12 |  |
| **01-02-11** | 1 | 0.19 | 0.19 | 0.28 | 0.27 | 0.25 | 0.25 | 0.28 | 0.25 | 0.31 | 0.07 |  |  |
| **01-03-11** | 1 | 0.15 | 0.25 | 0.2 | 0.22 | 0.17 | 0.27 | 0.23 | 0.28 | 0.09 |  |  |  |
| **01-04-11** | 1 | 0.21 | 0.2 | 0.21 | 0.2 | 0.23 | 0.22 | 0.26 | 0.07 |  |  |  |  |
| **01-05-11** | 1 | 0.19 | 0.17 | 0.17 | 0.21 | 0.23 | 0.26 | 0.1 |  |  |  |  |  |
| **01-06-11** | 1 | 0.17 | 0.16 | 0.26 | 0.23 | 0.34 | 0.1 |  |  |  |  |  |  |
| **01-07-11** | 1 | 0.18 | 0.21 | 0.22 | 0.27 | 0.11 |  |  |  |  |  |  |  |
| **01-08-11** | 1 | 0.21 | 0.25 | 0.24 | 0.12 |  |  |  |  |  |  |  |  |
| **01-09-11** | 1 | 0.23 | 0.3 | 0.11 |  |  |  |  |  |  |  |  |  |
| **01-10-11** | 1 | 0.24 | 0.12 |  |  |  |  |  |  |  |  |  |  |
| **01-11-11** | 1 | 0.11 |  |  |  |  |  |  |  |  |  |  |  |
| **01-12-11** | 1 |  |  |  |  |  |  |  |  |  |  |  |  |

Table 7: Overview of RFM values

|  |  |  |  |
| --- | --- | --- | --- |
| **CustomerID** | **Recency** | **Frequency** | **Monetary** |
| 12346 | 325 | 1 | 77183.6 |
| 12347 | 2 | 182 | 4310 |
| 12348 | 75 | 31 | 1797.24 |
| 12349 | 18 | 73 | 1757.55 |
| 12350 | 310 | 17 | 334.4 |
| 12352 | 36 | 85 | 2506.04 |
| 12353 | 204 | 4 | 89 |
| 12354 | 232 | 58 | 1079.4 |
| 12355 | 214 | 13 | 459.4 |
| 12356 | 22 | 59 | 2811.43 |

Table 7 above shows the overview of the recency, frequency and monetary value of each customers (identified by their customer ID) as return from the RFM model fitting in python.

Table 8: RFM metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Recency** | **Frequency** | **Monetary** |
| Count | 4338 | 4338 | 4338 |
| mean | 92.05947 | 90.52374 | 2048.688 |
| std | 100.0123 | 225.507 | 8985.23 |
| min | 0 | 1 | 3.75 |
| 0.25 | 17 | 17 | 306.4825 |
| 0.5 | 50 | 41 | 668.57 |
| 0.75 | 141.75 | 98 | 1660.598 |
| max | 373 | 7676 | 280206 |

Having dropped customers without ID’s it is evident from the metrics that there are only 4338 unique customers who have purchased the varying product with an average recency of 92 days, frequency of 90 and monetary value of $2048.69.

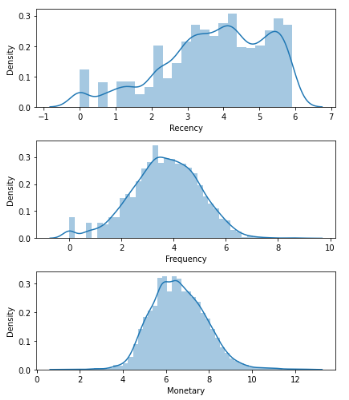
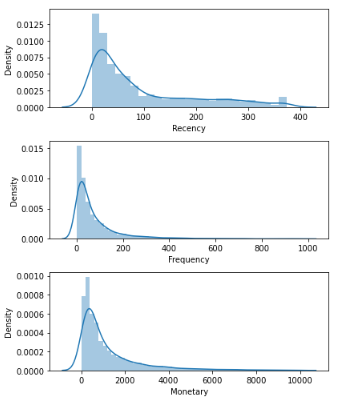
Table 9: Classifying customers based on RFM scores

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CustomerID** | **Recency** | **Frequency** | **Monetary** | **R** | **F** | **M** | **RFMGroup** | **RFMScore** | **Comment** |
| 14646 | 1 | 2076 | 280206 | 1 | 1 | 1 | 111 | 3 | VIP |
| 18102 | 0 | 431 | 259657.3 | 1 | 1 | 1 | 111 | 3 | VIP |
| 17450 | 8 | 336 | 194390.8 | 1 | 1 | 1 | 111 | 3 | VIP |
| 16446 | 0 | 3 | 168472.5 | 1 | 4 | 1 | 141 | 6 | VIP |
| 14911 | 1 | 5670 | 143711.2 | 1 | 1 | 1 | 111 | 3 | VIP |
| 12415 | 24 | 714 | 124914.5 | 2 | 1 | 1 | 211 | 4 | VIP |
| 14156 | 9 | 1395 | 117210.1 | 1 | 1 | 1 | 111 | 3 | VIP |
| 17511 | 2 | 963 | 91062.38 | 1 | 1 | 1 | 111 | 3 | VIP |
| 16029 | 38 | 241 | 80850.84 | 2 | 1 | 1 | 211 | 4 | VIP |
| 12346 | 325 | 1 | 77183.6 | 4 | 4 | 1 | 441 | 9 | Standard |
| 16684 | 4 | 277 | 66653.56 | 1 | 1 | 1 | 111 | 3 | VIP |
| 14096 | 4 | 5111 | 65164.79 | 1 | 1 | 1 | 111 | 3 | VIP |
| 13694 | 3 | 568 | 65039.62 | 1 | 1 | 1 | 111 | 3 | VIP |
| 15311 | 0 | 2366 | 60632.75 | 1 | 1 | 1 | 111 | 3 | VIP |
| 13089 | 2 | 1814 | 58762.08 | 1 | 1 | 1 | 111 | 3 | VIP |

The RFM scores in the second to the last column is a sum of individual recency, frequency and monetary value of each client. While classifying the client on three stages (VIP, Standard and Economy), the overview of classification above shows on 1 standard customer among several VIPs. The full table is attached in the output result.

K-MEANS CLUSTERING

Figure 1: Distribution of data before (LHS) and after (RHS) transformation.



The plots on LHS shows the distribution of the recency, frequency and monetary capacity of customer, it is evident that the three distributions are highly skewed, however after applying the log transformation, the distribution are approximately normal as show on the RHS.

After normalizing the data, the K-means cluster is conducted using 5 clusters as it is believed to capture better classification with sum of square distances below 4000 (figure 2).

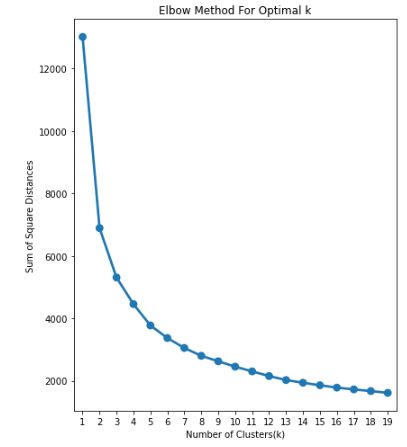


Figure 2

Table 10: Associating clusters to each customer ID

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Customer**  **ID** | **Recency** | **Frequency** | **Monetary** | **R** | **F** | **M** | **RFMGroup** | **RFMScore** | **RFM\_Loyalty\_Level** | **Cluster** | **Color** |
| **14646** | 1 | 2076 | 280206 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **18102** | 1 | 431 | 259657.3 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **17450** | 8 | 336 | 194390.8 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **16446** | 1 | 3 | 168472.5 | 1 | 4 | 1 | 141 | 6 | VIP | 3 | red |
| **14911** | 1 | 5670 | 143711.2 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **12415** | 24 | 714 | 124914.5 | 2 | 1 | 1 | 211 | 4 | VIP | 3 | red |
| **14156** | 9 | 1395 | 117210.1 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **17511** | 2 | 963 | 91062.38 | 1 | 1 | 1 | 111 | 3 | VIP | 3 | red |
| **16029** | 38 | 241 | 80850.84 | 2 | 1 | 1 | 211 | 4 | VIP | 3 | red |
| **12346** | 325 | 1 | 77183.6 | 4 | 4 | 1 | 441 | 9 | Standard | 0 | green |
| **…** | **…** | **…** | **…** | **.** | **.** | **.** | **…** | **.** | **…** | **.** | **…** |

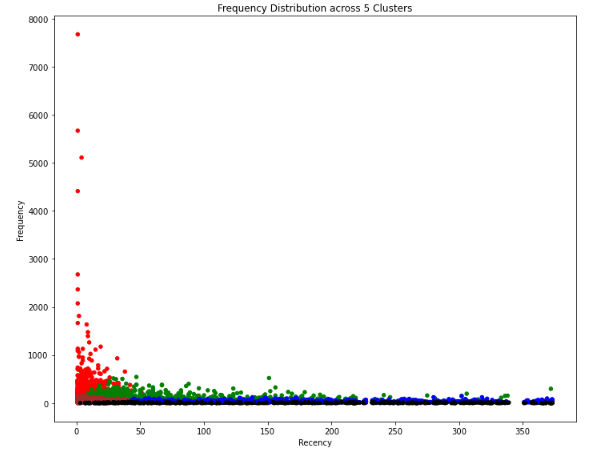


Figure 3

Figuratively (Table 10) we can see that the third cluster (red colour) are majorly dominated by the VIPs which are on the upper part of the distribution (figure 4) while the last cluster (black colour) are highly dominated by the economy class customers (as shown Figure 4)

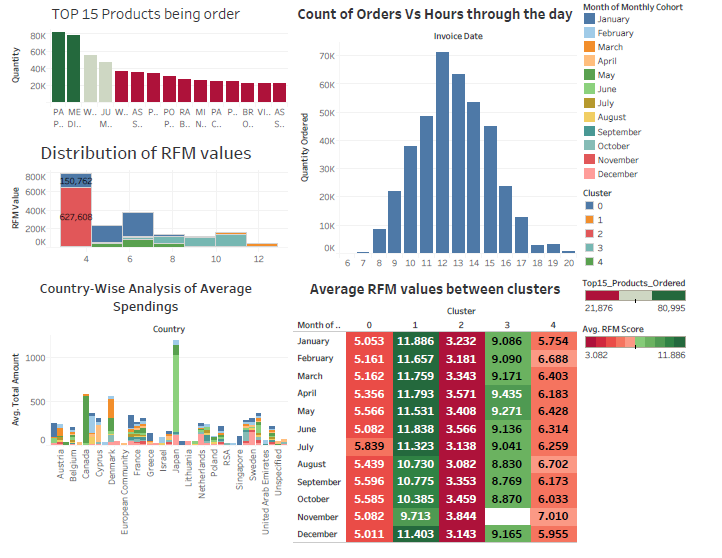


Figure 4: Dashboard for showing cluster performances

The full dashboard (whose overview is shown in figure 4) is available in the .twbx file attach to the submission.