|  |  |  |  |
| --- | --- | --- | --- |
| Customer segment | R | F | M |
| High value | 1 | 1-3 | 1 |
| Loyal customers | Any | 2-3 | Any |
| $Potential loyalists | Any | 1-2 | Any |
| Recent customers | 1 | Any | Any |
| At risk | 3-4 | 1 | 1 |
| Hibernating | 3-4 | 4 | 4 |
| Can’t lose them | 4 | 1-2 | 1 |
| Need attention | Which does not fall under any | | |

We applied RFM analysis during our time processing the first dataset. We used the data under the header ‘from\_totally\_fake\_account’ to serve as the Customer ID, representing each individual customer. The header ‘not\_happened\_yet\_date’ shows the recency of the serving dates. The parameter ‘monopoly\_money\_amount’ contains data on the value of each transaction. The first step was to generate several random dates to make the transactions realistic. Then we used quantiles to segment the customers. We set a score from one to four representing the level of recency, frequency, and monetary values. Finally, we obtain an RFM score which is a comprehensive number capturing the summation of the three quantiles above, indicating the loyalty of the customers. We then segmented the customers into eight different levels, making a plot to communicate this visually.

 R\_Score: Score based on recency of purchases (1 is most recent, 4 is least recent).

F\_Score: Score based on the frequency of purchases (1 is the highest frequency, 4 is lowest).

M\_Score: Score based on monetary value (1 is highest spending, 4 is lowest).

RFM\_Score: Combined score from each RFM metric, where a score of "111" represents a top customer (most recent, frequent, and highest spending).

High value: Buy recently, buy often, and spend the most. Typically, RFM scores "111" through "131".

Loyal Customers: Buy on a regular basis. Responsive to promotions. Scores like "X3X", "X2X".

Potential Loyalists: Recent customers with average frequency. Scores like "X1X", "X2X".

Recent Customers: Bought most recently, but not often. RFM scores "1XX".

At Risk: Used to purchase frequently but haven’t purchased recently. High RFM scores like "311", "411".

Can’t Lose Them: Made big purchases and often, but haven’t returned for a long time. Scores like "411", "421".

Hibernating: Low recency, frequency, and monetary scores. Scores like "344", "444".

Additionally, we perform a methodology called K-means clustering on the dataset. This dataset contains four columns and we chose the ‘monopoly\_money\_amount’ column to cluster since it contains the transaction amounts. We choose the number of K as three since the data set is huge and the iteration of computation is complex. We plot the elbow curve to visualize the data. The X-axis and Y-axis represent the number of clusters and the inertia respectively. We cluster the data and count the data points of each cluster. Then the distribution of the data points can be observed obviously.

|  |  |  |  |
| --- | --- | --- | --- |
| Customer segment | R | F | M |
| High value | 1 | 1-3 | 1 |
| Loyal customers | Any | 2-3 | Any |
| $Potential loyalists | Any | 1-2 | Any |
| Recent customers | 1 | Any | Any |
| At risk | 3-4 | 1 | 1 |
| Hibernating | 3-4 | 4 | 4 |
| Can’t lose them | 4 | 1-2 | 1 |
| Need attention | Which does not fall under any | | |