

**Be Customer Wise or Otherwise:  
Combining data mining and interactive visual analytics to analyze large and  
complex customer resource management (CRM) data**

Kam Tin Seong, Singapore Management University  
Aditya Hridaya Misra, Nanyang Technological University  
Ji Jun Yao, SAS

## ABSTRACT

In this competitive world, there are more companies, such as our project sponsor, a global logistics company, are exploring the potential use of data mining techniques to make informed and intelligent marketing strategies. We conducted a customer segmentation study using a comprehensive set of data extracted from their existing operational system which contains both transaction records and customer profiles. This paper aims to report on our experience gained in using the interactive visual analytics and data mining techniques of SAS® JMP to perform customer segmentation analysis in combination with RFM (Recency, Frequency and Monetary), a method used for determining the Customer Lifetime Value (CLV). We share our views on how interactive visual analytics and data mining techniques can empower everyday data analysts to gain useful insights and formulate informed decisions by demonstrating the interactive data visualization techniques of JMP such as histogram and parallel plots.

## INTRODUCTION

The project sponsor, a global market leader in the logistics industry, specializes in providing innovative and customized logistics services. However, the company lacks the insight about its customer segments and profile, in order to improve its decision making in marketing campaign, pricing strategy and the other key focus areas. While it had a rich repository of transaction data, the company lacks the expertise to analyze the data and identify the customer target list. This paper thus aims to demonstrate how we can perform customer segmentation analysis for the purpose of coming up with recommendations that can help the company in its decision making processes.

## MOTIVATION

The customer transaction data of a logistics company was analyzed. Data mining, clustering in particular, and visualization techniques were used to find meaningful relationships and patterns within the existing data. These mining and visualization techniques enable a company to target its customers more efficiently and improve its marketing campaign response rate. It also equips the company with the much needed customer intelligence that would enable it to gain a much better insight and visibility of its customers. Clustering analysis is a commonly used method to perform market segmentation by many companies. However, if data analysts make use of a clustering technique only, it would be difficult for the business users to understand and apply the results. On the other hand, using a pure marketing technique such as RFM without leveraging upon the data mining algorithms like clustering, would not add a significant value. Therefore, the focus of this paper of this paper shall be on a methodology that aims to bridge this gap between the two i.e. RFM and data mining, through visualization.

## RFM ANALYSIS

In order to strengthen the results of customer segmentation, we decided to use data mining through visualization coupled with the market segmentation and scoring methodology of RFM. RFM stands for Recency (how recently did the customer purchase), Frequency (how often do customers purchase) and Monetary (how much do customers spend). The reason for the methodology being so popular is because it is much easier to apply and understand compared with machine learning techniques such as clustering analysis or classification. A business user without a strong statistical background can easily and conveniently apply it and find the group of customers that he needs to focus on, and thereon view the required set of results within minutes. It can be used to target special offers to welcome new customers, encourage the small spending customers to spend more, reactivate lapsed customers and encourage marketing initiatives.

By applying RFM methodology, we were able to identify customers who have made a purchase lately, bought often or spent much (or satisfying all three) along with the rest. Mailing people who are not very regular can sometimes prove to be completely futile as there might be different reasons associated to their lack of response (the customers may have moved their location or changed interest or switched to a competitor). Nevertheless, RFM helps in customer development as it forms the basis of a good customer retention strategy.

## DATA PREPARATION

Three tables (Transaction details, Customer master table and Industry code directory) were selected from the existing database:

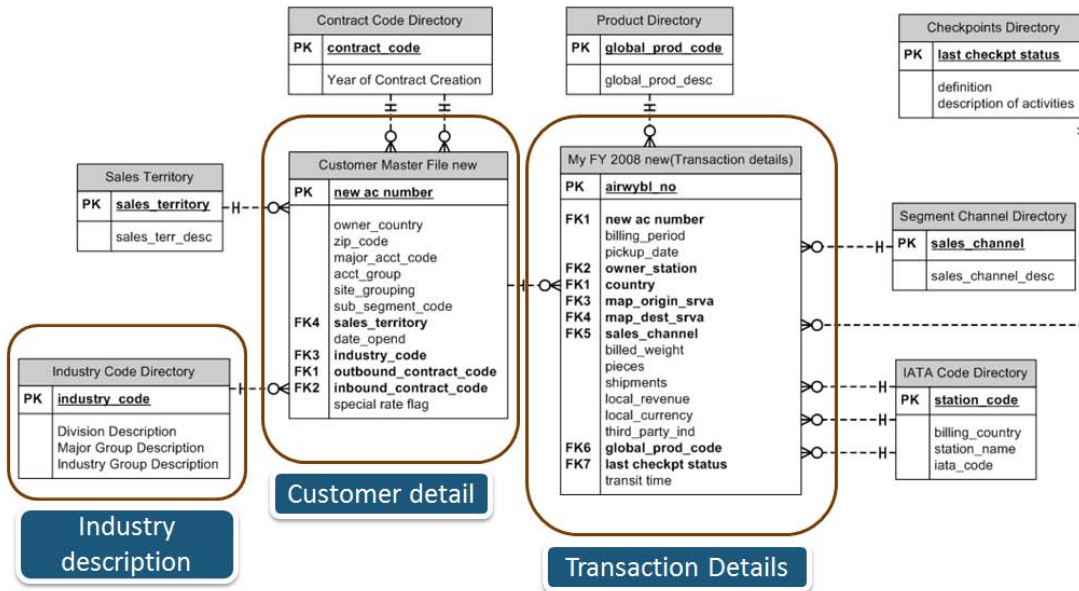


Figure 1. Selection of tables

The transaction table contains 2,507,495 records, where each record represents a single transaction. This dataset contains transactions over twelve months. In order to ensure the accuracy and completeness of the results, the entire dataset was extracted without sampling. Before proceeding with the analytics, transactions for each customer had to be aggregated for every account number, which is the unique identifier for every customer. The following ETL process was carried out on the selected tables, as shown in Figures 2 and 3:

- **Step 1:** A subset of the variables was selected and combined into a single table via inner joins. Three variables namely 'pickup\_date', 'local\_revenue' and 'count' were selected from the transaction details to derive the RFM variables (Recency, Frequency and Monetary). The other numeric variables such as 'billed\_weight', 'pieces' and 'shipments' are also amongst the important variables because these would have a direct impact on the business cost. For two transactions with similar revenue, the one with a higher billed weight could incur a much higher transportation cost and thus result in a lower profit margin (Since one transaction could contain more than one shipment and one shipment could contain more than one piece of product). However in this study, we shall focus only on the three RFM variables while the other numeric variables will not be considered as it is beyond the scope of this discussion.

The variables 'division\_description', 'major\_group\_description' and 'industry\_group\_description' were also selected from the table 'Industry Description' because these variables could be potentially used to enrich the clustering analysis by visualizing customers in different clusters based on their industry codes.

- **Step 2:** Aggregation: As part of the next step, the data was transformed to aggregate all the transactions performed by a particular customers under a single account number.

## ETL process step 1 & 2 (Select & Aggregate)

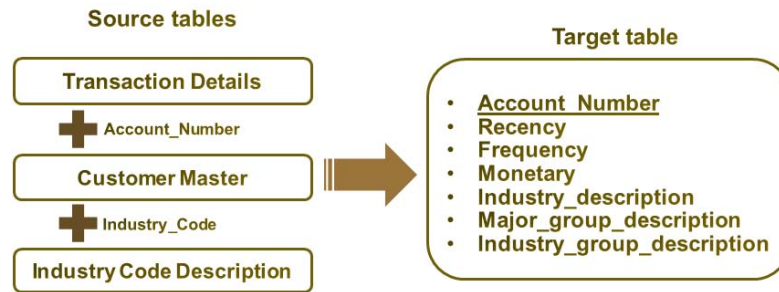


Figure 2. Data Transformation (Step 1 & 2)

- **Step 3:** Values for Recency, Frequency and Monetary variables are divided into five equal frequency bins. For each variable, the top 20% was given a score of 5 followed by 4,3,2,1. Using this scoring system, each customer was assigned a score ranging from 111 (lowest) to 555 (highest), with a total of 125 (5\*5\*5) combinations.

## ETL process step 3 (Transformation)

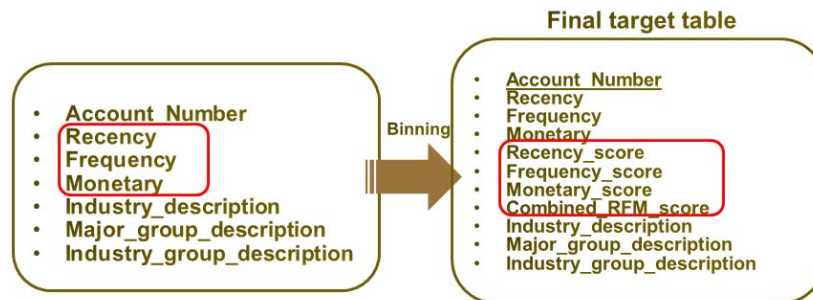


Figure 3. Data Transformation (Step 3)

The final table has a total of 11 columns, and 20,431 unique customer accounts. However, there are some records with a Monetary value of 0. Further investigation (something which is beyond the scope of this discussion) is required in order to find out the underlying reasons for this. After filtering out these records, there are still 18,595 records.

## SOFTWARE USED

We used the SAS® Enterprise Guide to extract, transform and load the source data into the target table. Apart from the easy-to-use, point-and-click interface, SAS Enterprise Guide also provides us with a comprehensive set of predefined tasks to perform quick data management, statistical analysis and ad-hoc reporting.

Due to the availability of limited hardware resources and the large size of the dataset, we decided to use JMP – a lightweight data mining software under the SAS® Suite for our mining tasks. Apart from being lightweight, JMP also provides the users with its in-memory processing features. Thus, it drastically improved the turnaround time of our results. SAS® JMP makes data mining much easier and intuitive by allowing the users to explore and interact with the results through a comprehensive set of visualization techniques. For example, parallel plot, histogram and tree map.

OBSERVATION OF RFM SCORE

Figure 4 shows the cut points for Recency and number of records for each group. Customers are equally divided into 5 groups based on the original values of Recency, Frequency and Monetary. They are generated using JMP Interactive Binner (an additional Add-In that can be downloaded from the JMP official website).

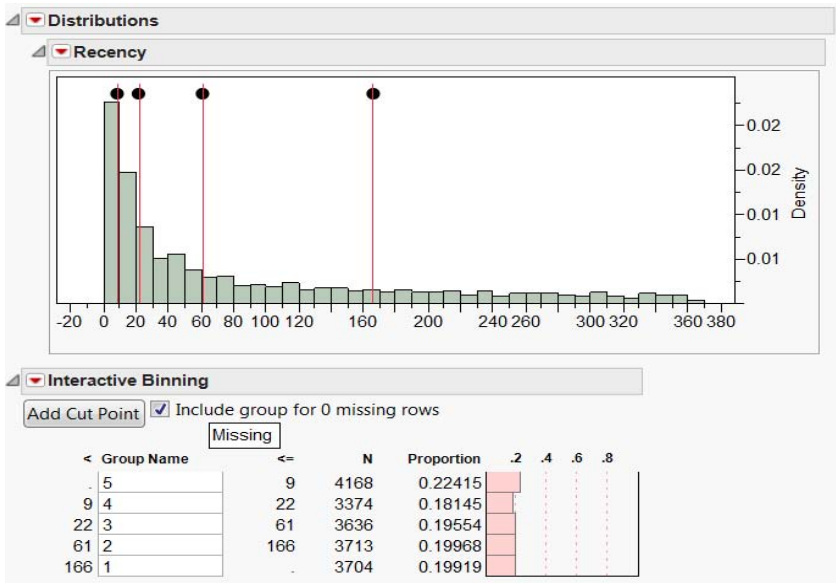


Figure 4. Binning for variable Recency

Figures 5 and Figure 6 show the distribution of the combined RFM score. As observed, customers with a score of 111 and 555 have the highest frequency. Based on Figure 6, it can be inferred that there is a cyclical pattern within the RFM scores. Upon digging deeper into the histogram, we discovered that all the various spikes are scores with same values for the last two digits (e.g. 544, 533, 455, 344, 211 and etc.). This pattern refers to a strong correlation between Frequency and Monetary.

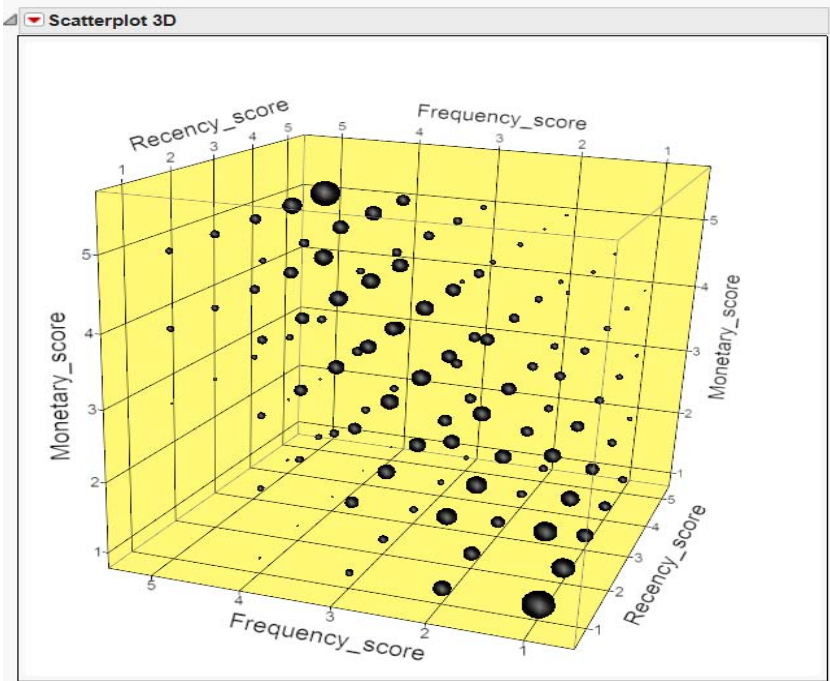


Figure 5. RFM Scoring 3D scatterplot

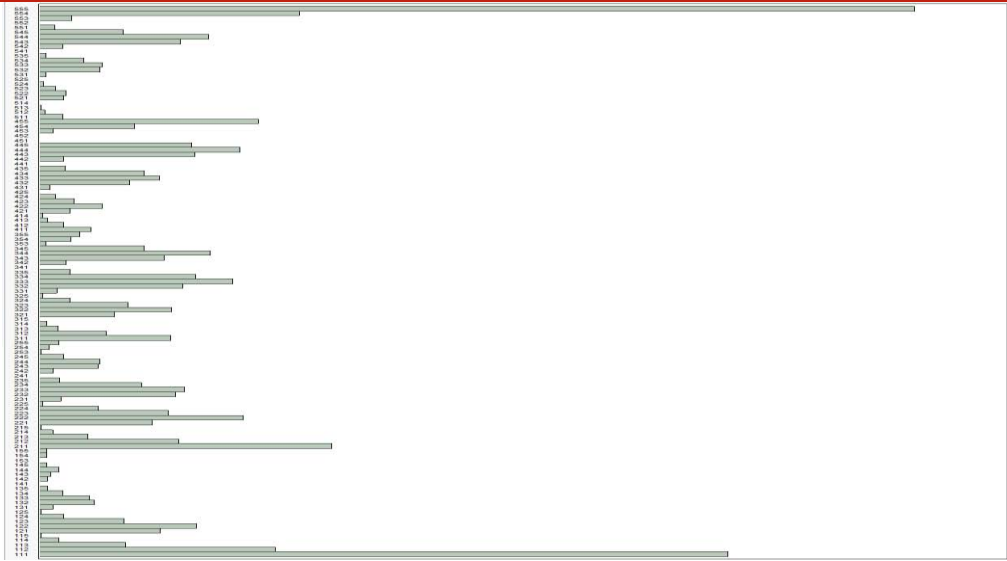


Figure 6. RFM Scoring Histogram

It can be concluded from the conventional RFM methodology that the customers with the highest RFM score are those who should be targeted by the company as they have purchased recently, regularly and have spent a lot of money. The 111s, on the other hand, will be considered too expensive to be included in the mailing list constantly. In general, most of the time, a company trims the customers with a RFM score of 111 from its target list. However, by doing so, the company may potentially suffer from a huge loss (in terms of customers) as more than half of the customers have only a few transactions with a low monetary value. Also, when applying the conventional RFM approach, there will be a total of 125 subclasses with different RFM scores. In most cases, they are too overwhelming as it is impractical to drill down into each subclass and try to differentiate them one by one. Regrouping these subclasses into larger groups will be rather subjective and challenging.

Conventional RFM method also fails to identify and reason out certain anomalies that require further exploration. Customers with the same RFM score may differ significantly and may thus potentially demand different actions and strategies from the company. Figure 7 shows a parallel plot of the original RFM values using a subset of customers who have a RFM score of 555. In this group, the customer with the highest monetary value has generated more than \$10 million in terms of revenue while the average for the bucket is only around \$150K. Having identified the limitations of the conventional RFM analysis, we shall now explore the application of statistical modeling (k-mean clustering analysis in particular) in order to improve the results of RFM analysis.

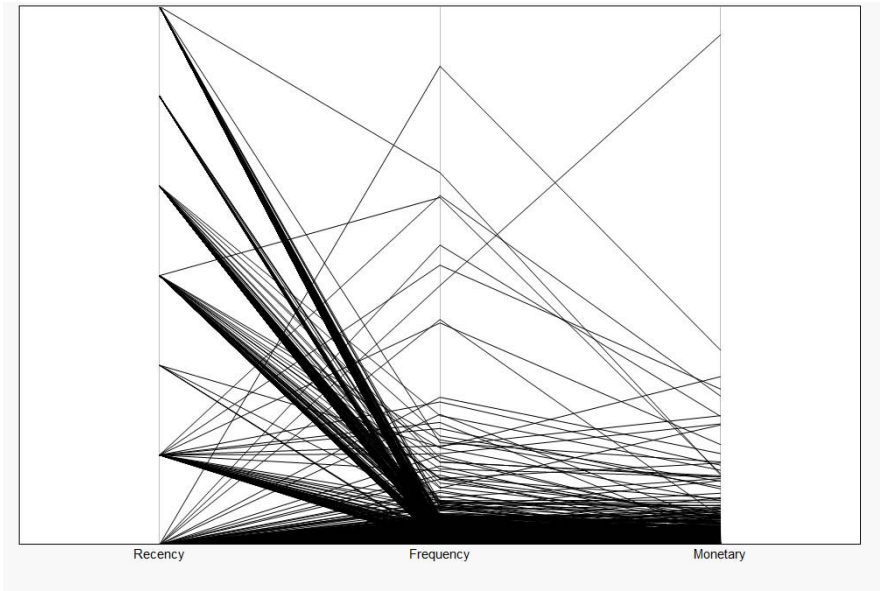


Figure 7. Parallel plot of customers with RFM score of 5-5-5



K-MEANS CLUSTERING USING SAS JMP

Before proceeding with the application of Recency, Monetary and Frequency as the input variables for clustering analysis, the required values were rescaled from 0 to 100 using the formula in figure 8. Rescaling the values will not only make the results easier to interpret but also allow the analyst to make a comparison between different variables in case they have different units and scales. Analysis done on variables with different scales of measurement can lead to clusters that are not very robust to add/delete variables or observations.

$$\left( \frac{(\text{Recency} - \text{Col Minimum}(\text{Recency}))}{(\text{Col Maximum}(\text{Recency}) - \text{Col Minimum}(\text{Recency}))} \right) * 100$$

Figure 8. Rescaling formula for Recency

In addition to rescaling, the distribution of the variables was examined too. It is important to understand the distribution of variables and identify the outliers before applying k-mean clustering. A variable that is highly skewed with many outliers will have to be normalized because K-means clustering is highly susceptible to noise and outliers.

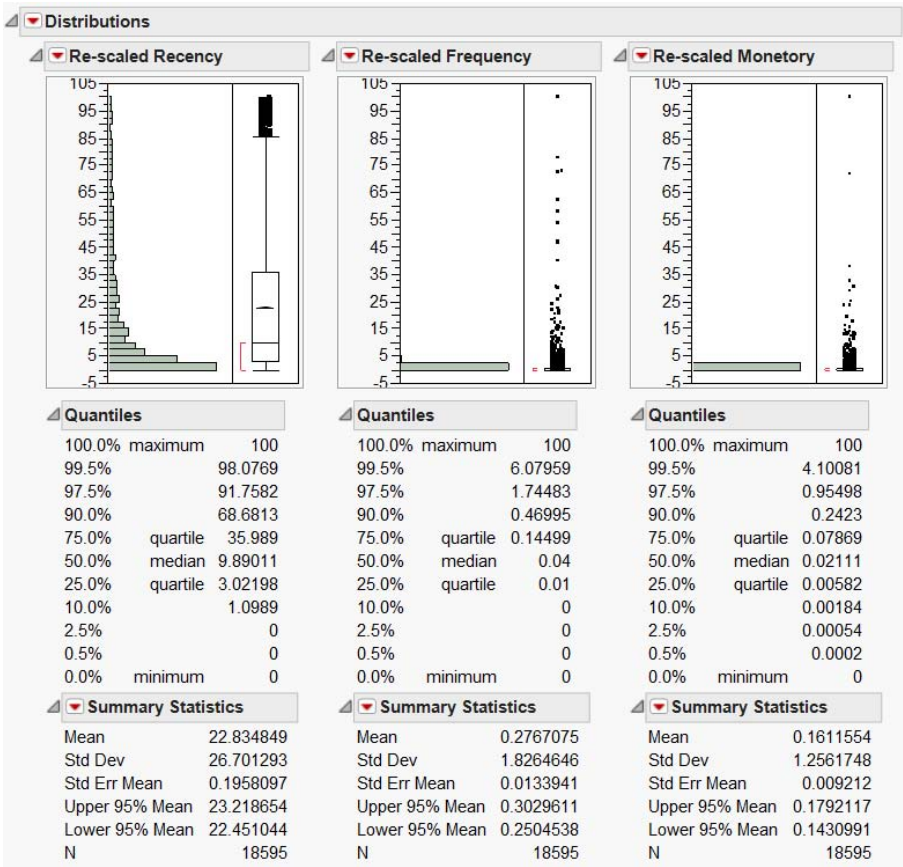


Figure 9. Distribution analysis result of rescaled RFM variables

Based upon the result shown above in Figure 9, the three rescaled RFM variables were found to be highly skewed. For both Frequency and Monetary, there were some outliers with extremely high values and thus required additional steps before we could have conducted the clustering analysis.

K-means clustering analysis in JMP provides an out-of-box data transformation functionality – ‘Johnson Transform’, which efficiently balances highly skewed variables and brings outliers closer to the center of the rest of the values.

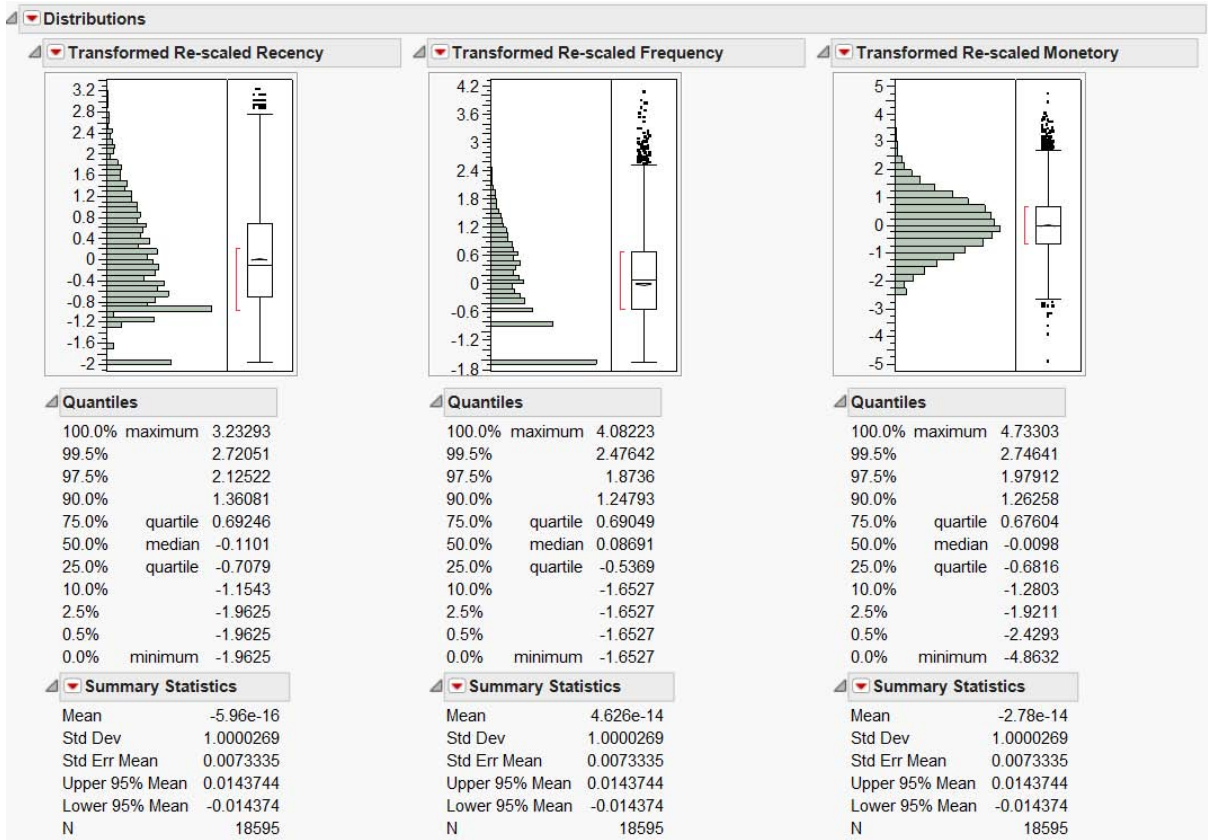


Figure 10. Distribution analysis result of transformed rescaled RFM variables

Based on Figure 10, the RFM variables are much more balanced after applying the Johnson Transform function with the same mean (close to 0) and standard deviation (close to 1). Once the variables were transformed, we conducted the K-mean clustering analysis. Below are some snapshots of the results.

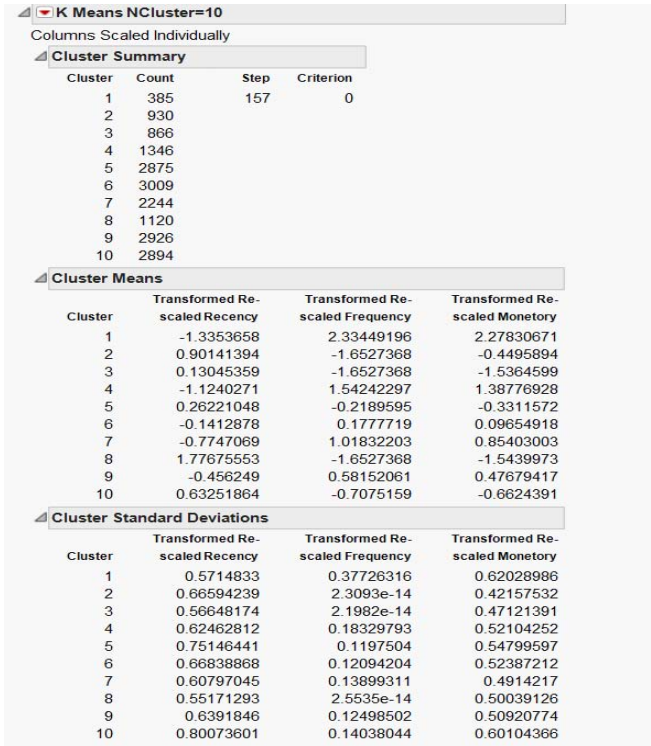


Figure 11. Summary of K-means clustering analysis

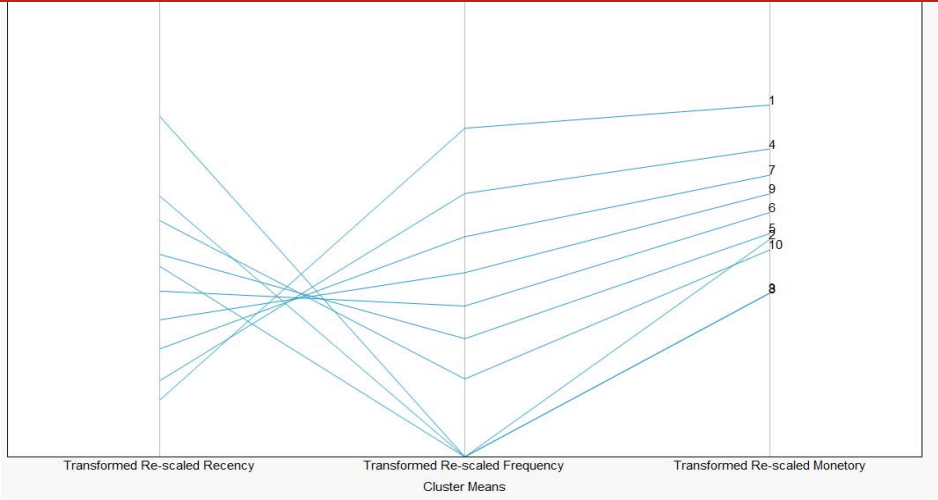


Figure 12. Parallel plot of the cluster means

*\*Cluster 3 and 8 share similar mean values for Frequency and Monetary. The labels for cluster3 and 8 are therefore overlapped.*

Having generated the clusters, we used the parallel plots, in order to identify the correlation between different variables, which allowed us to visualize a number of variables at the same time. By adopting an approach of data mining coupled with visualization, it became much more intuitive and a lot more easier for us to interpret and differentiate patterns and infer results.

Based on Figure 12, we concluded that there is a positive correlation between Frequency and Monetary (on the basis coherence of the lines) and a negative correlation between Recency and Frequency. Also, for the variable Recency, we need to take note that the smaller the value, the more recent the customer is. Since all the clusters follow a similar pattern, it can be generalized that a customer with a relatively recent transaction tends to contribute a higher amount of transactions and monetary value.

Also, another pattern can be observed from this parallel plot is cluster 2, 3 and 8 have extreme low Frequency values thus making them slightly different clusters when compared with the rest. After drilling down into these three clusters, we found out that all the customers of cluster 2, 3, 8 have the same Frequency value of 1, which makes up more than 15% of the company's customer base.

Besides parallel plot, scatterplot matrix and 3D scatterplot can also be used to figure out the influence of individual variables on the result of the clustering analysis by plotting the members of different clusters with different colors and markers.

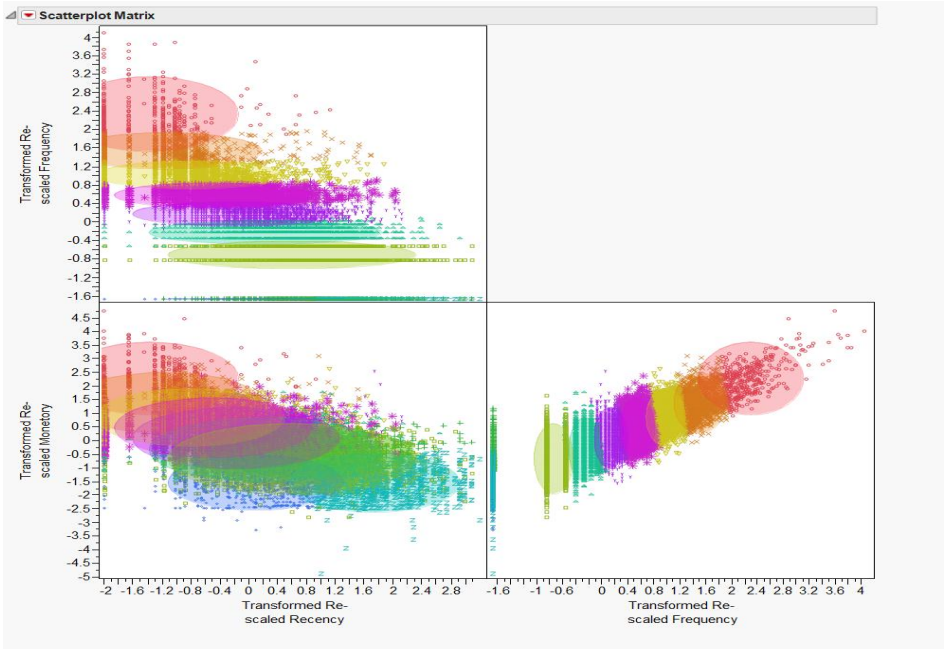


Figure 13. Scatterplot Matrix of the clusters



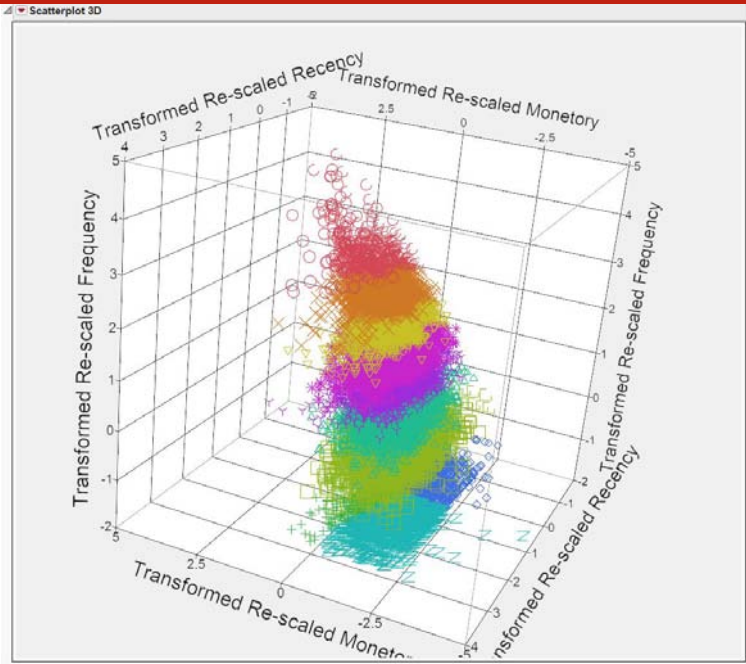


Figure 14. 3D Scatterplot of the clusters

Based on Figure 13 and 14, the clusters are mainly determined by the value of the variable Frequency. Also, Figure 13 re-emphasizes the point that there is a stronger correlation between Monetary and Frequency as compared to Recency.

Based on Figure 15, cluster 2, 3, and 8 represent customers with long Recency, very low Frequency and low Monetary. These are customers who are not requesting the services actively, on average once a year. By and large these customers hardly had transactions for the last five months and even they use the services, the sales are very low. They represent the least loyal and low value customers. Cluster 1, on the other hand are those customers with high value. By drilling down into the original values, these customers requested the services practically twice a day. They are the customers the company should try to retain as long as possible.

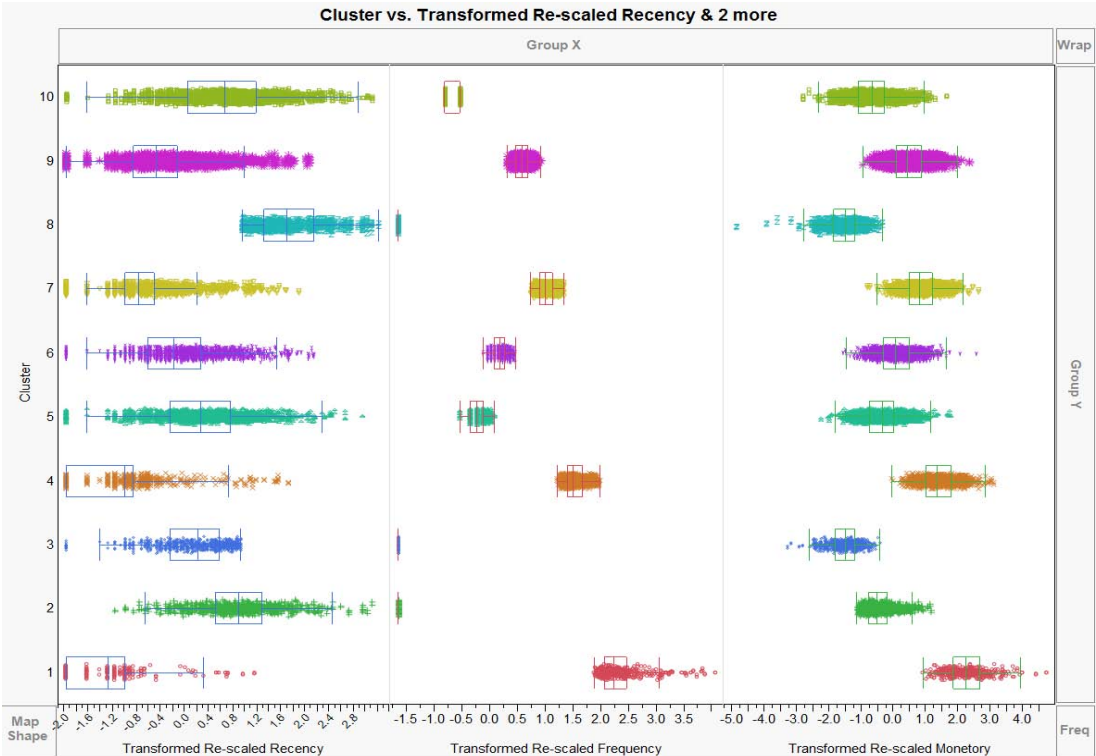


Figure 15. Comparison of the clusters using Graph Builder

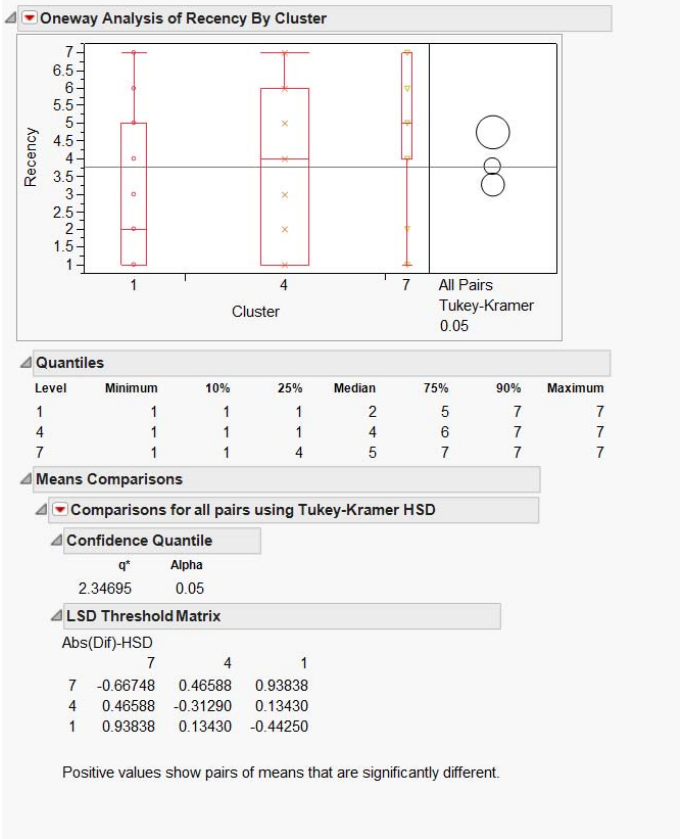


Figure 16. Oneway analysis of Recency by Cluster for Customers with RFM score of 555

The oneway analysis in Figure 16 is generated using a subset of data with the same RFM score of 555. Based on the result of LSD threshold matrix, all the pairs are positive thus implying that the cluster 1, 4, 7 are significantly different. By using clustering analysis, three clusters, namely Cluster 1, 4 and 7, were identified but by using RFM, they will be treated as one group. This result highlights the fact that the clustering analysis successfully identifies and covers the gaps encountered by the conventional RFM approach.

CONCLUSION

Instead of simply dividing the customers into large, non-segmented buckets using the RFM score, k-means clustering using RFM variables as the inputs allows us to benefit from both the conventional market segmentation as well as statistical modeling.

Interactive visual analytics and data mining techniques can empower everyday data analysts to gain insights and formulate informed decision. The best combination is of an “intelligent” analyst who not only has deep industry knowledge of the field but also a good understanding of data mining algorithms and techniques to apply and use. That way, useful and relevant findings and recommendations can be communicated to the decision makers of the business, and realize the full potential of data mining to support the decision making process.

**CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the authors at:

**Kam Tin Seong**

Singapore Management University  
80 Stamford Road  
Singapore 178902  
Email: [tskam@smu.edu.sg](mailto:tskam@smu.edu.sg)

**Aditya Hridaya Misra**

Nanyang Technological University  
50 Nanyang Avenue  
Singapore 639798  
Email: [adityahmisra@gmail.com](mailto:adityahmisra@gmail.com)  
Work phone: +65 85225432

**Ji Jun Yao**

SAS Institute Pte Ltd  
20 Anson Road Level 8  
Twenty Anson  
Singapore 079912  
Email: [junyao.ji@sas.com](mailto:junyao.ji@sas.com)  
Work phone: +65 63988960

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.