

|  |
| --- |
| **ISSS602 DATA ANALYTICS LAB** |
| **Assignment 2: Be Customer wise or Otherwise** |
| **FoodOnline’s Customer Segmentation & its Analysis** |

**Prepared By**

***Raveena Chakrapani***

***(raveenac.2021@smu.mitb.edu.sg)***

**27th February 2022**

Contents

[1. OVERVIEW 4](#_Toc96875520)

[2. OBJECTIVE 4](#_Toc96875521)

[3. DATA PREPARATION 4](#_Toc96875522)

[3.1 Raw Data Description 4](#_Toc96875523)

[3.2 Loading Data into JMP Pro 4](#_Toc96875524)

[3.3 Checking for Completeness and Accuracy 6](#_Toc96875525)

[3.4 Deriving New Variables 6](#_Toc96875526)

[3.5 Excluding the unnecessary variables 7](#_Toc96875527)

[3.6 Extracting Records 7](#_Toc96875528)

[4. DATA ANALYSIS 8](#_Toc96875529)

[4.1 DATA WRANGLING 8](#_Toc96875530)

[4.1.1 Deriving Frequency Measure 8](#_Toc96875531)

[4.1.2 Deriving Recency Measure 10](#_Toc96875532)

[4.1.3 Deriving Diners 11](#_Toc96875533)

[4.1.4 Joining Data Tables 12](#_Toc96875534)

[4.2 INTERACTIVE DATA EXPLORATION 13](#_Toc96875535)

[4.2.1 Distribution Analysis 13](#_Toc96875536)

[4.2.2 Multivariate Analysis 14](#_Toc96875537)

[5. SEGMENTATION USING VARIOUS CLUSTERING METHODS 16](#_Toc96875538)

[5.1 DETERMINING SUITABLE CLUSTER VARIABLES 16](#_Toc96875539)

[5.2 WORKING WITH HIERARCHICAL CLUSTERING 17](#_Toc96875540)

[5.2.1 Deriving Diners 17](#_Toc96875541)

[5.2.2 Reason for Selection / Rejection 17](#_Toc96875542)

[5.3 WORKING WITH K MEANS CLUSTERING 18](#_Toc96875543)

[5.3.1 Overview 18](#_Toc96875544)

[5.3.2 Performing K-Means Clustering 18](#_Toc96875545)

[5.3.3 Analysing K-Means Clustering Report 19](#_Toc96875546)

[5.3.4 Interpreting clustering results using Parallel Coordinates 20](#_Toc96875547)

[5.3.5 Verifying the Optimal Cluster 21](#_Toc96875548)

[5.4 WORKING WITH GAUSSIAN MIXTURE MODEL 22](#_Toc96875549)

[5.4.1 Overview 22](#_Toc96875550)

[5.4.2 Performing Gaussian Mixture Model 22](#_Toc96875551)

[5.4.3 Analysing Gaussian Mixture Model 22](#_Toc96875552)

[5.4.4 Interpreting clustering results using Parallel Coordinates 23](#_Toc96875553)

[5.5 WORKING WITH LATENT CLASS ANALYSIS 24](#_Toc96875554)

[5.5.1 Overview 24](#_Toc96875555)

[5.5.2 Binding Clustering Variables 24](#_Toc96875556)

[5.5.2 Performing Latent Class Analysis 25](#_Toc96875557)

[5.5.3 Analysing Latent Class Analysis Report 26](#_Toc96875558)

[6. INTERPRETATION OF ANALYSIS RESULTS 27](#_Toc96875559)

[6.1 Reasons for Selecting / Rejecting a clustering technique 27](#_Toc96875560)

[6.2 Mapping the cluster to Extracted Data 28](#_Toc96875561)

[7. MANAGERIAL RECOMMENDATIONS 31](#_Toc96875562)

[8. REFERENCES 32](#_Toc96875563)

# 1. OVERVIEW

To scale efficiently and effectively, expansion-stage companies need to focus their efforts on a specific subset of customers, who are most similar to their best current customers, not a broad universe of potential customers. Customer segmentation comes to rescue. It is the marketing strategy that helps any company to improve the whole product, to focus on marketing message, to pursue higher percentage opportunities, to get higher quality revenue. It can have numerous other ancillary benefits too. To carry out customer segmentation i.e. to identify segments in the data, cluster analysis technique is used to convert very large observations (in form of rows and columns) into meaningful taxonomies, groups, or clusters, based on combinations of input variables. Food Online is one of the leading discounted restaurant reservation platform in Singapore. The business model of the company aims to connect empty tables to empty stomachs, through time-based discounts. Let us deep dive into its customer data and perform customer segmentation to gain tangible insights.

# 2. OBJECTIVE

The purpose of this study is to perform suitable cluster analysis by grouping FoodOnline’s customers, namely the diners into homogeneous segments based on their booking, dinning behaviours. We can achieve this by considering desirable cluster variables and metrics. After deriving desired variables, let us try to understand the data by conducting Exploratory Data Analysis (EDA) to understand the pattern of customers and their booking.

# 3. DATA PREPARATION

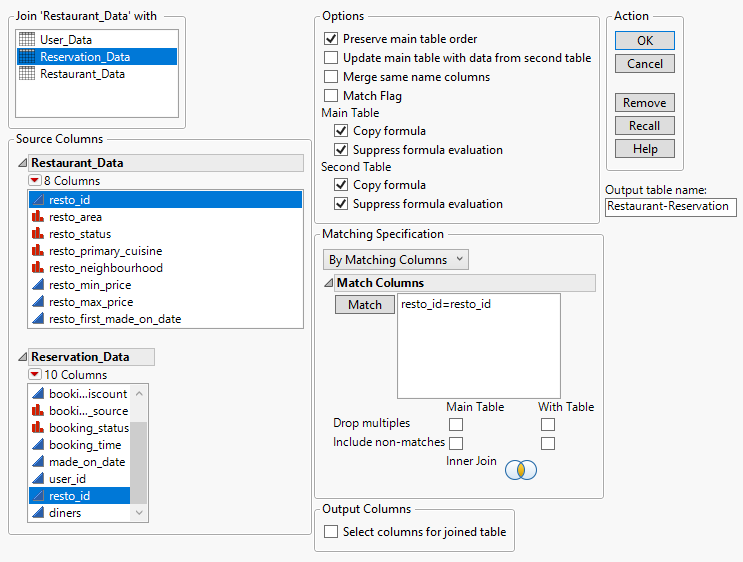
## 3.1 Raw Data Description

For this analysis, following data are used

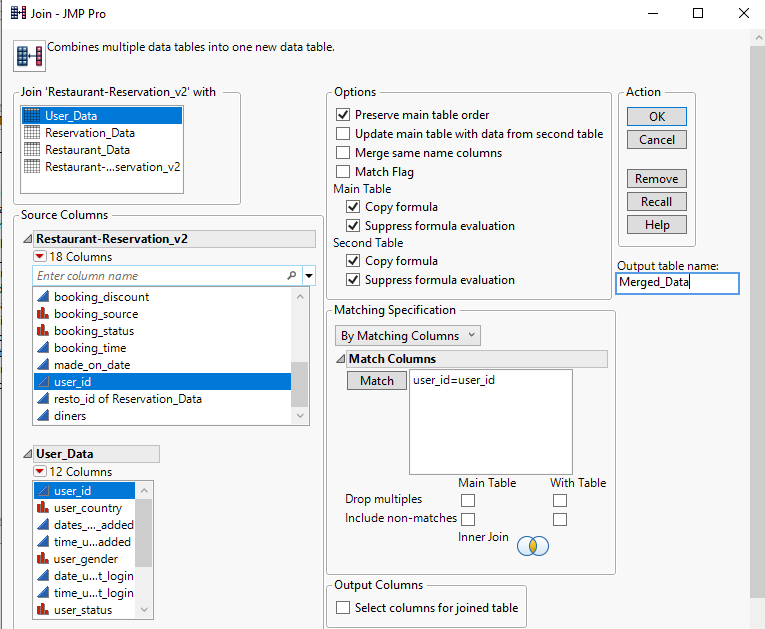
* *UserDataSheet.xls* consists of details information of the platform’s users.
* *Participating\_Restaurants\_Datasheet*.xls provides information of the participating restaurants.
* *Reservation\_Datasheet*.xls provide details information of each reservation made by the platform user.

## 3.2 Loading Data into JMP Pro

* Let us load these three different excel workbooks in JMP Pro software and merge it into a single data file using Join technique. Primary key variables, which are to be used for this joining, are user\_id and restaurant\_id.
* Import all three files into JMP pro. Open *Reservation\_Datasheet* table and click on
* **Tables 🡪 Join 🡪 Join Dialog window appears**
* Select *Reservation\_Data* from ‘Join ‘Restaurant\_Data with’ pane.
* From the source columns pane select *resto\_id* from both the tables and click on Match.
* Provide an output table name and click OK.



This *Restaurant-Reservation* is the intermediate table. With the above steps, this table should be joined with *User\_Data* table by matching the column ‘*user\_id*’.

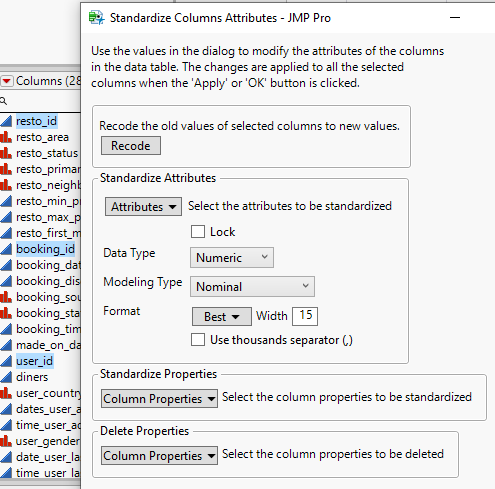


Final table obtained is named as *Merged Data.* Delete all redundant columns such as *user\_id* of reservation data and resto\_id of reservation data. Same columns from two different tables can be deleted. Let us use this final merged data to perform data feature engineering.

## 3.3 Checking for Completeness and Accuracy

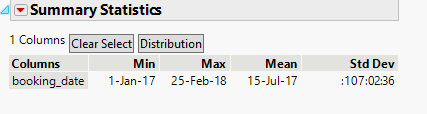
Once the data is loaded and integrated, let us check for accuracy and completeness. By accuracy, let us check whether datatypes in the imported dataset matches the original dataset. By completeness, we check whether no. of rows and columns, which are imported, matches the original dataset. This final merged data contains 520,586 records. Let us convert all id columns from continuous to nominal datatype.

Select *user\_id*, resto\_id and *booking\_id* **and Right click 🡪 Standardise Attributes 🡪 ‘Select All’ from the drop down of Standardise attributes pane 🡪 Nominal in Modelling Type 🡪 OK**



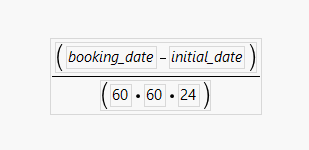
## 3.4 Deriving New Variables

Let us consider Booking Date. It ranges from 1 Jan 2017 to 25 Feb 2018. It can be seen by selecting the **Booking Date column 🡪 Cols 🡪 Column Viewer 🡪 Show Summary**



Based on these values let us number the days from 1 Jan 2017 to 25 Feb 2018 by creating a new column called Day. Before that, let us create a column called initial\_date and enter the value as 1-Jan-17. Now create a new column called Day and insert a formula.

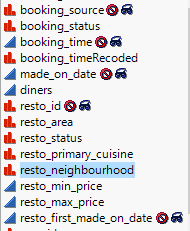
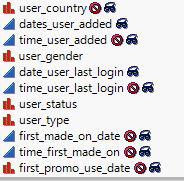
**Right click on any column name 🡪 Insert columns 🡪 Rename it as Day 🡪Right click again 🡪 Formula 🡪 Enter the below formula**



We are finding the difference between booking date and initial date, which is the day number from Day1 of 2017 to Last day of 2018. JMP Pro calculates the result in terms of seconds. It is converted to days by dividing it by (60 \* 60\* 24 = 86,400) as 1 day contains 86,400 seconds.

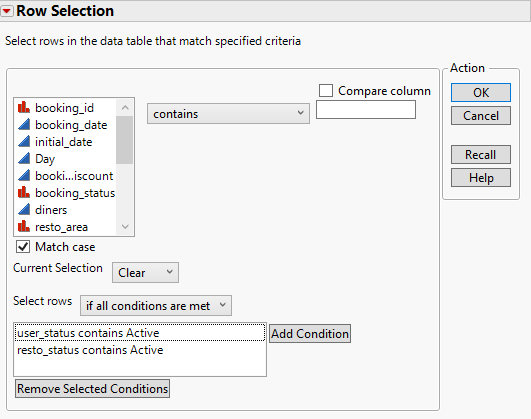
## 3.5 Excluding the unnecessary variables

While doing analysis, to avoid selecting the unwanted columns accidentally, let us hide and exclude the unnecessary columns. Here let us hide these columns shown in below fig. Select these columns on Columns Pane 🡪Right click 🡪 Hide/ Unhide and select Exclude/Unexclude options.

## 3.6 Extracting Records

Let us filter the required data. Select 🡪 Rows 🡪 Row Selection 🡪 Select Where and type the conditions as shown in figure. Click on OK button & save these filtering conditions into main data table by selecting red drop down left to Row Selection -> Save Script 🡪 Data Table 🡪 Fitering\_Conditions (Name of the script)



Save these records into new data table.

Right click ‘Selected’ in Rows Pane 🡪 Data View. Rename it as Extracted\_Data

# 4. DATA ANALYSIS

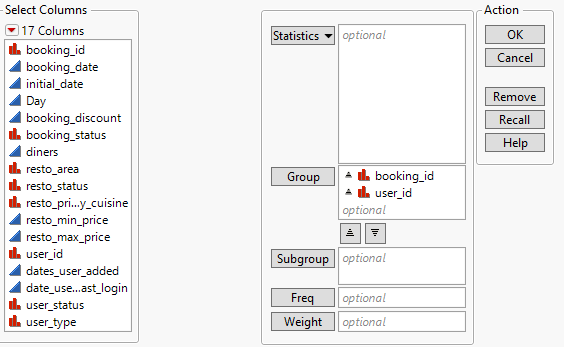
## 4.1 DATA WRANGLING

From this Extracted Data table let us derive new variables from the existing variables by using **Summary** function of JMP.

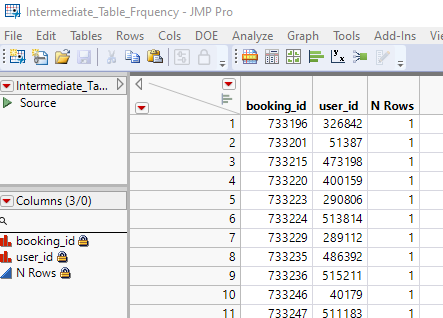
### 4.1.1 Deriving Frequency Measure

Let us derive frequency by performing summary count twice. First, we will calculate the frequency count by summary the transaction records by using both user\_id and booking\_id. This is because each row in the transaction dataset represent a single booking.

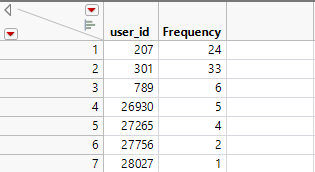
**Select Tables 🡪 Summary. Select booking\_id and user\_id from Select Columns pane and click on Group Button. Click OK.**



The intermittent table appears with these columns.



From this table, let us group the values again by user\_id based by following the steps mentioned before. Rename the column N Rows as Frequency and save the table as Frequency.

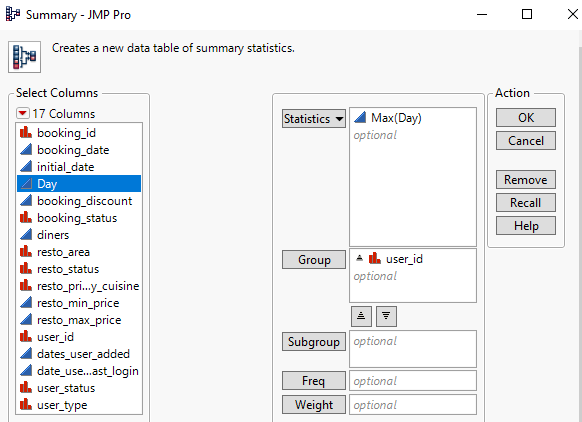


This shows that the user with user\_id 207 has reserved a table 24 times.

### 4.1.2 Deriving Recency Measure

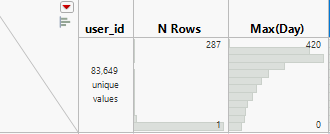
Let us derive the new measure called Recency which means how recently the user has booked a table.

* Select **Table 🡪 Summary**
* Select Day in Select Columns pane and select Max in Statistics Pane.
* Similarly Select *user\_id* in Columns pane and click on Group Button.
* Click OK.



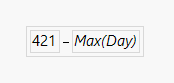
In this new summary table, double click on the empty column after the last. New column will be created and rename it as Recency.

Click on the small icon of Show /Hide header graphs in the table and view the max number of days



Here it shows 420 as the maximum day.

Right click the column and select Formula Type this formula in the dialog box which appears and select OK Save the data table as Recency.

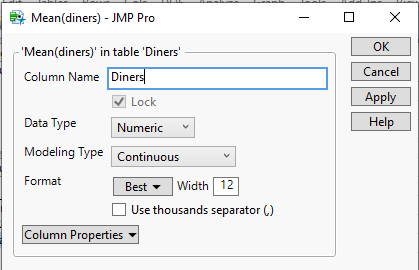


Notice that values are updated in Recency column. Records with value 1 indicate booking made a day ago whereas records with value 261 indicate a booking made 261 days ago.

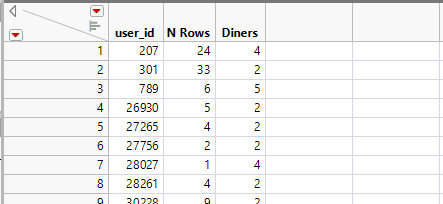
### 4.1.3 Deriving Diners

Let us group by user\_id and find average of Diners by following the steps used above for Recency and Frequency.

* Select Diners and select Mean in Statistics Pane and group by user\_id.
* Notice that in the newly derived column Mean(Diners), the values are in decimal. Let’s convert to whole number by steps shown below
* **Right click that column 🡪 Column Info 🡪 Change Format to Fixed Dec 🡪 Enter 0 in Dec.**
* Rename the column to Diners
* Click OK and Save the table as Diners.



We can infer from the Diners table that the customer with user id 207 books table for 4 on an average.



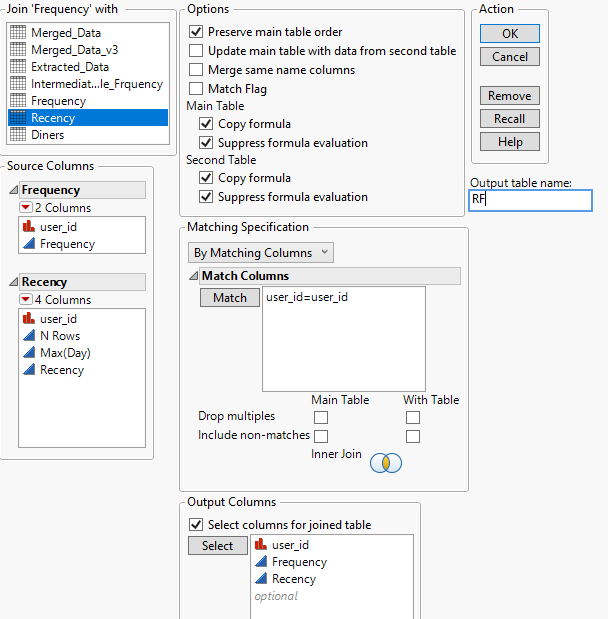
### 4.1.4 Joining Data Tables

Let us combine the data tables of Frequency, Recency and Diners by using the below steps:

Open Frequency Table 🡪 Table 🡪 Summary 🡪 Join.

Select Recency Table and use user\_id columns to match and select user\_id, Frequency and Recency for the output columns.

Type RF in Output Table Name and click OK.



Similarly using the same steps, join this RF table with Diners Table and rename the Final table as RF&Diners.

## 4.2 INTERACTIVE DATA EXPLORATION

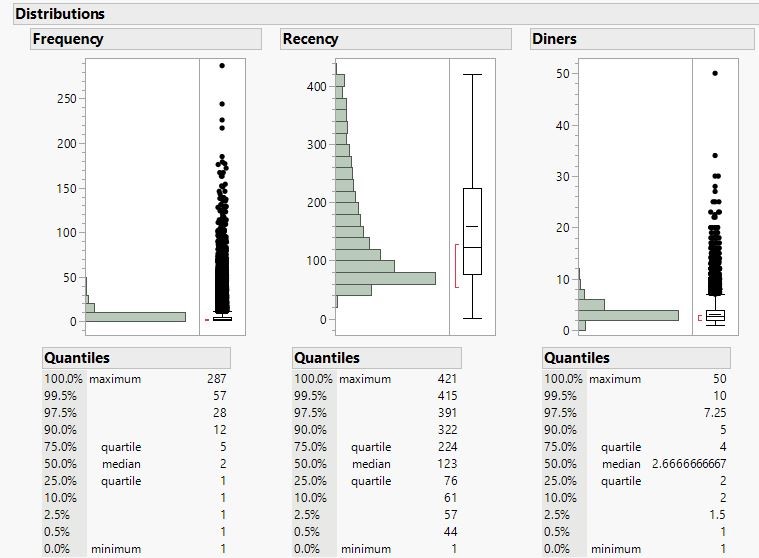
Before clustering, let us examine the distribution of the input variables to examine the outliers , skewness of data distribution and most importantly to examine variables that are highly correlated. Most popularly known as Multicollinearity.

### 4.2.1 Distribution Analysis

Let us start with exploring and analyzing the input variables by using the **Distribution Analysis** platform of JMP Pro.

• From the menu bar of RF&Diners data table, select **Analyze** -> **Distribution**.

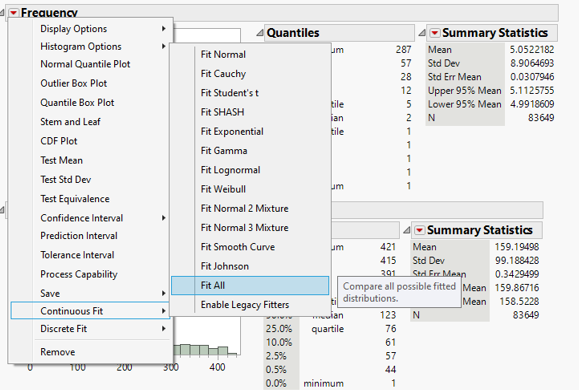
Select Recency, Frequency , Diners and select Y column



The distribution analysis results reveal that close to 25% of the records with 1 value in Frequency column. This is because there are many customers who book only once. This shows that our data is right skewed.

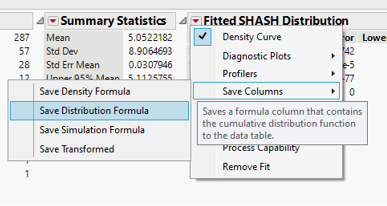
So let us perform transformation. Before that, let us save this distribution into data table as *DistBeforeTransf.*

In the distribution chart, select red button next to Frequency 🡪 Continuous Fit 🡪 Fit All



The system selects the best fit automatically. Repeat the same for Recency and Diners.

Save the distribution, formula in our data table by following the steps as shown in fig. This will be used in our analysis especially for the clustering methods, which are sensitive to outliers.



Save this distribution as *DistAfterTransf.*

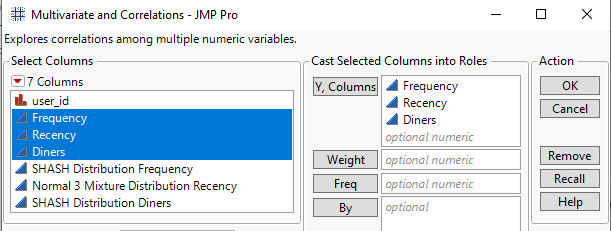
### 4.2.2 Multivariate Analysis

Let us use the Multivariate platform of JMP to investigate the correlation of the input variables.

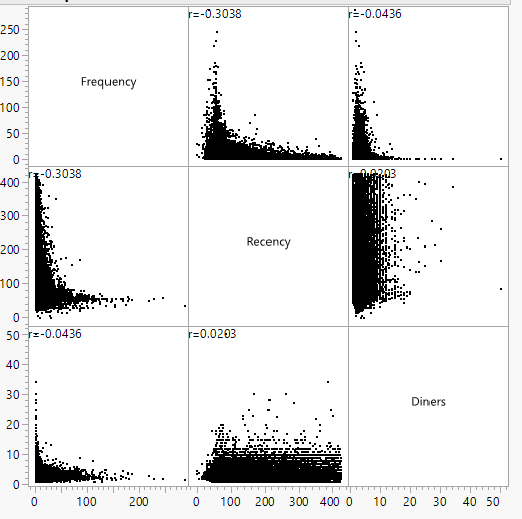
From the menu bar of RF&Diners, select **Analyze** -> **Multivariate Methods** -> **Multivariate**.

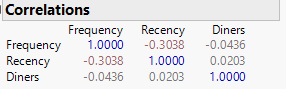
The **Multivariate and Correlation** platform dialog window appears.

Select Frequency, Recency and Diners from the Select Columns pane and click on OK.



The scatterplot matrix & correlation values are is shown below





The analysis results reveal that most of the selected variables are not strongly correlated. A highly correlated pair will have a correlation value greater than +/-0.80. In view of this, we will use all these variables in the cluster analysis.

# 5. SEGMENTATION USING VARIOUS CLUSTERING METHODS

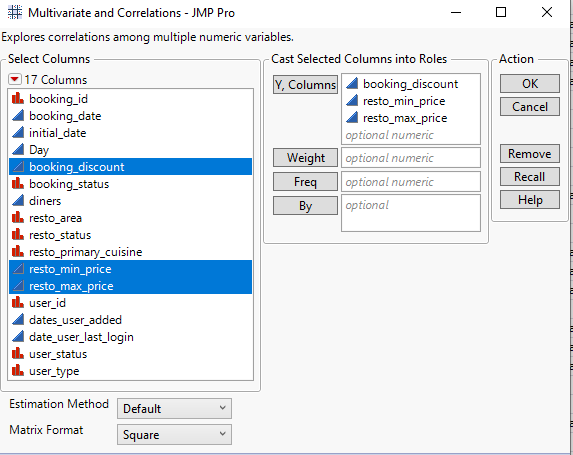
## 5.1 DETERMINING SUITABLE CLUSTER VARIABLES

Selecting correct cluster variables play an important role in clustering as the characteristics of the entire dataset is evaluated by keeping these variables as a reference while performing clustering.

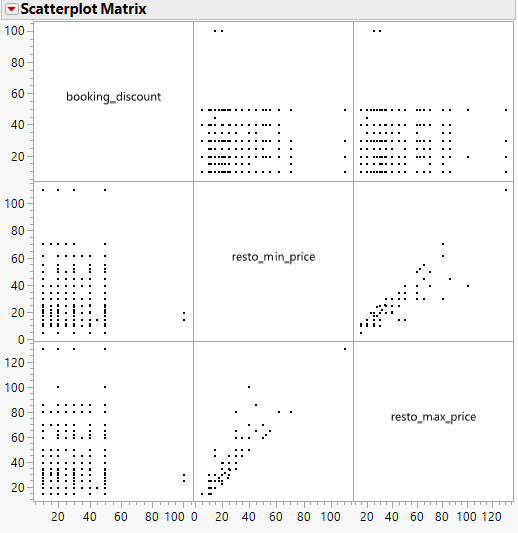
Already we have chosen 3 cluster variables such as Recency, Frequency and Diners. Let us look into some more continuous variables. Most of the clustering techniques except Latent Class Analysis can be performed only with continuous variables.

Let us perform Multivariate Analysis and choose the correct variables. In *Extracted\_Data* data table select Analyze 🡪 Multivariate Methods 🡪 Multivariate

Choose variables such as booking\_discount, resto\_min\_price and resto\_max\_price. Select Y columns and click OK.



Let us infer from the scatterplot matrix of these chosen variables



The plot clearly shows that not all chosen variables are continuous. They exhibit discrete characteristics. Hence, let us drop all these variables for clustering and proceed with already derived and chosen variables of *Frequency*, *Recency* and *Diners*.

## 5.2 WORKING WITH HIERARCHICAL CLUSTERING

### 5.2.1 Deriving Diners

**Hierarchical Clustering** refers to clustering methods that attempt to break data into a hierarchy of clusters. The procedures are characterised by the tree-like structure established in the course of the analysis. Generally, this can be done in two directions, namely: agglomerative (Many 🡪 One) and divisive (One 🡪 Many). This algorithm is based on distance measures such as Euclidean distance or city-block distance.

### 5.2.2 Reason for Selection / Rejection

One downside of hierarchical clustering is that if they have large storage requirements, they can be computationally intensive. Two factors affecting the choice of hierarchical clustering are the number of observations and the number of clustering variables.  For analytics data with more than 10,000 observations or/and 50 clustering variables, it is not wise to use hierarchical clustering because the distance matrix is going to be too large. In view of this statement, our final RF&Diners table has around 80K records.

Hence, let us not perform Hierarchical clustering for this dataset.

## 5.3 WORKING WITH K MEANS CLUSTERING

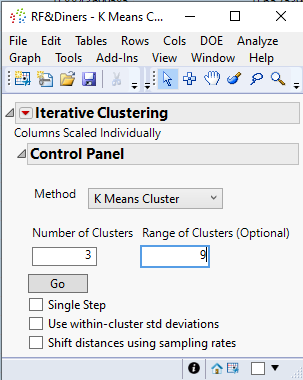
### 5.3.1 Overview

K-means clustering differs from hierarchical clustering by the way it clusters its data. The algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis. The clustering process starts by randomly assigning objects to a number of clusters. The objects are then successively reassigned to other clusters to minimize the within-cluster variation, which is the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster.

### 5.3.2 Performing K-Means Clustering

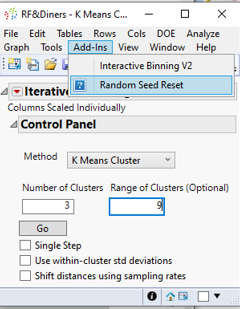
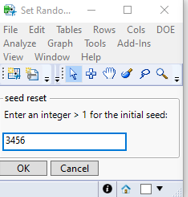
From the *RF&Diners* data table, click **Analyze** -> **Clustering** -> **K Means Cluster**. From **Options** dropdown list, select **KMeans** .The Cluster dialog window appears.

From the **Select Columns** pane, select *Recency, Frequency and Diners.* From the **Cast Selected Columns into Roles** pane, click on the Y, Columns button.



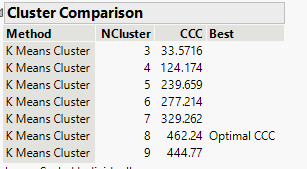
This K-Means Iterative Clustering window shows that the system generates three clusters centres randomly and asks the user to provide Range of Clusters (upper limit). Type 9 and let us select random seed value as 3456 to get the same results every time we run this script. Click OK in the Set Random Seed window and Click Go in the Iterative Clustering window.

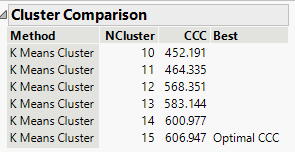
It can be set by selecting Add-Ins 🡪 Random Seed Set. Further, this Add-In can be downloaded from JMP Community site (<https://community.jmp.com/t5/JMP-Add-Ins/Random-Seed-Reset/ta-p/21973>) and get it installed in the system before performing the above step.

### 5.3.3 Analysing K-Means Clustering Report

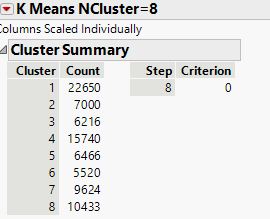
The K Means Cluster report window appears. As indicated in the Cluster Comparison report, the optimal number of clusters is 8 because it gives the largest CCC values. Let us also perform iterative clustering for different levels.





The optimal CCC increases with increase in the no. of clusters, which is ideally not convincing. So, lets confirm from the first comparison chart that Cluster 8 is the optimal one. Save the clusters and rename the column as K-8Cluster.

Let us examine the clustering report of **K Mean NCluster=8**.



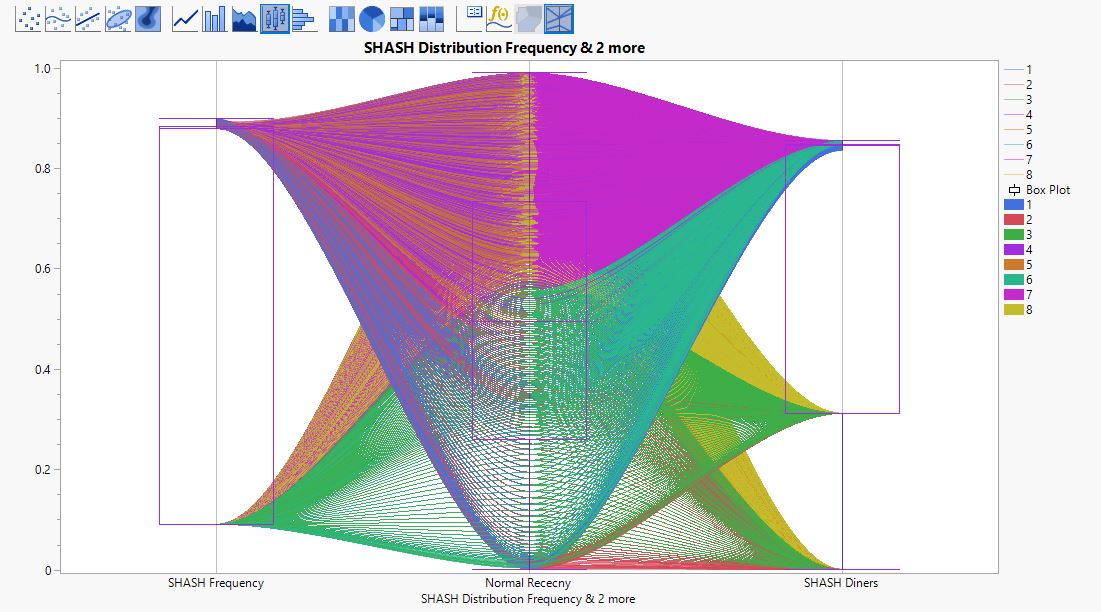
The report reveals that the minimum and maximum cluster sizes are 5520 and 22650 respectively. They are Cluster 6 and Cluster 1.

### 5.3.4 Interpreting clustering results using Parallel Coordinates

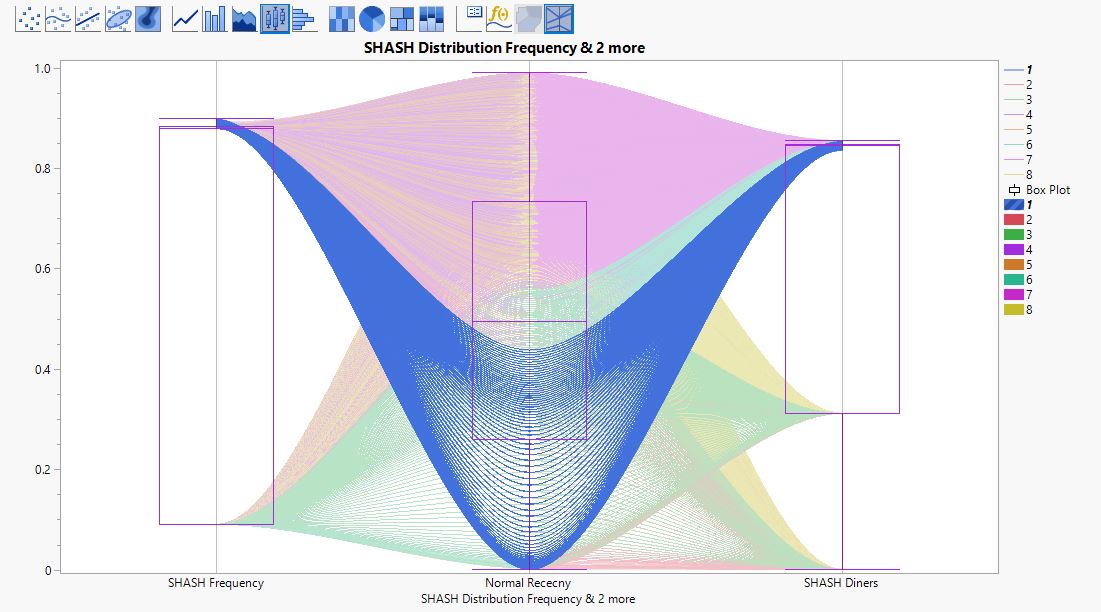
Let us use a data visualization technique called parallel coordinates to reveal the property of each cluster.

From the menu bar, select Graph -> Graph Builder.

* From the **Variables** pane, click on **Ctrl** key and select *SHASH Distribution Frequency, Normal 3 Mixture Distribution Recency, SHASH Distribution Diners.*
* Drag-and-drop the selected variables on **X**.
* From the **Variables** pane, click on *K-8Cluster*.
* Drag-and-drop on **Color** pane.
* Hover the mouse over the parallel co-ordinates
* Right-click and select Add -> Boxplot from the context menu.



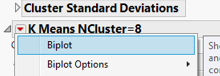
Select Cluster 1 in the legend and that part of curve is highlighted.

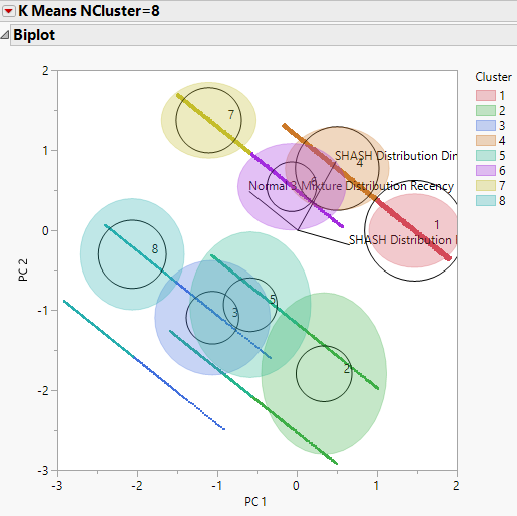


This plot shows that Cluster 1 has good set of customers who have high frequency , very recent and come with lot of diners.

### 5.3.5 Verifying the Optimal Cluster

The optimal cluster of 8 which we obtained through K Means can be verified using Biplot.





This biplot reveals that how clusters are formed and its organisation. Since the clusters are not overlapping to a large extent, this analysis through K Means Clustering hold good for Optimal 8 Cluster.

## 5.4 WORKING WITH GAUSSIAN MIXTURE MODEL

### 5.4.1 Overview

This is one of the advanced clustering techniques used for Segmentation.

Differences between K Means & Gaussian Mixture are as follows:

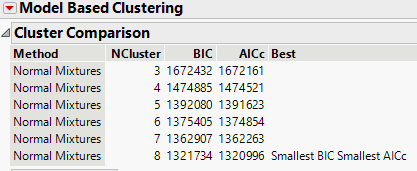
* K Means assumes sphericity, meaning that the clusters are shaped like multi-dimensional spheres. In practice, this roughly means that we’re assuming that all our features have equal variance, which is often not true.
* K-means performs hard classification whereas the Gaussian Mixture performs soft classification. In other words, K-means tells us what data point belong to which cluster. On the other hand, Normal Mixture algorithm provide us with the probabilities that a given data point belongs to each of the possible clusters.

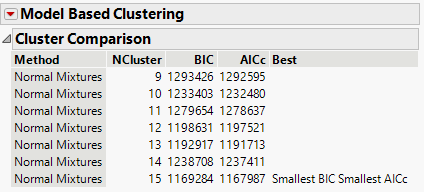
### 5.4.2 Performing Gaussian Mixture Model

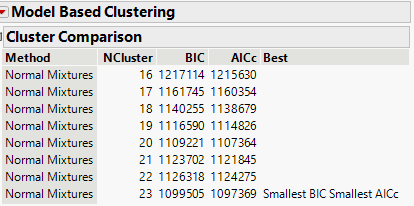
* In the RF&Diners Table, select Analyze 🡪 Clustering 🡪 Normal Mixtures
* Similar to the steps performed in K Mean Clustering select desired variables and provide desired seed and provide diff. max. no. of clusters iteratively.
* Repeat the iteration with different range of clusters (3-8) , (9-15) , (16-23). Set random seed as 3456 for each iteration.

### 5.4.3 Analysing Gaussian Mixture Model

The GMM report window appears. The optimal cluster would have small BIC and AIC.

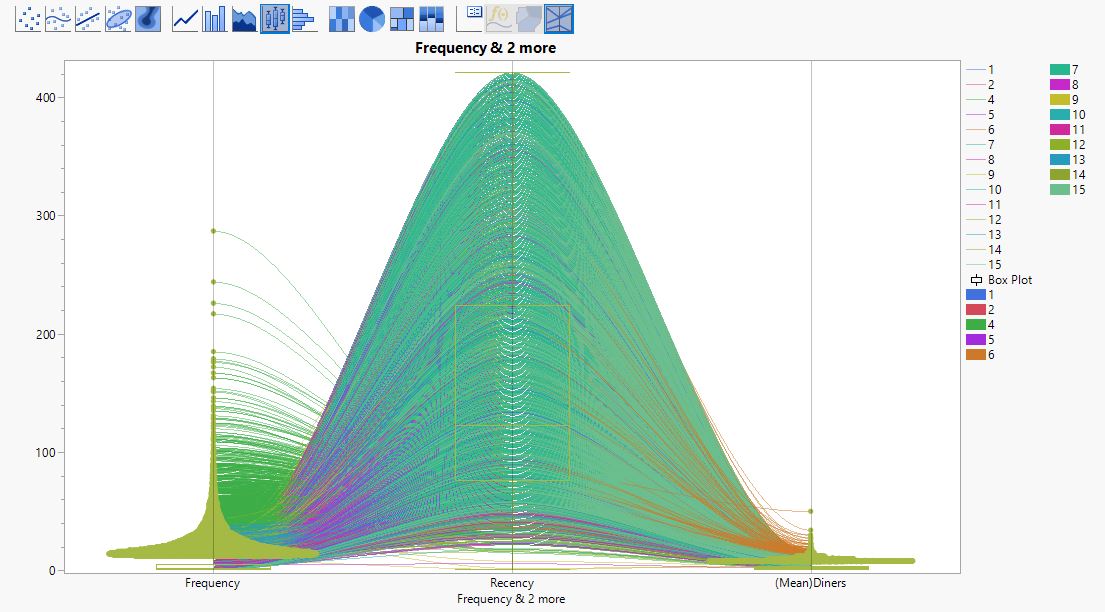






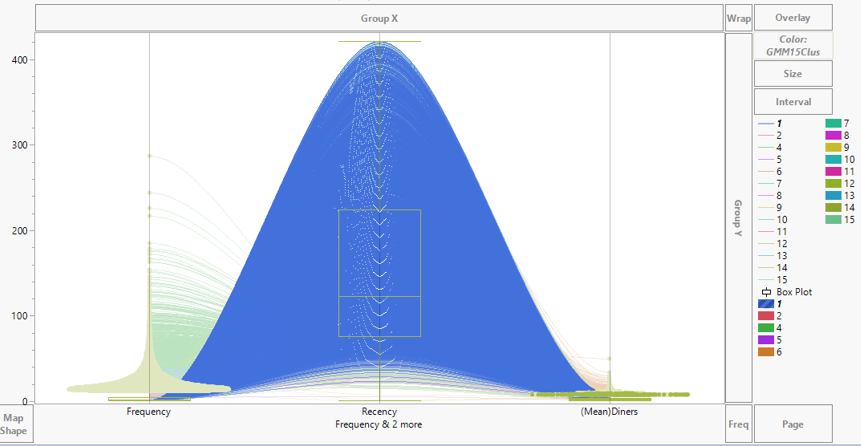
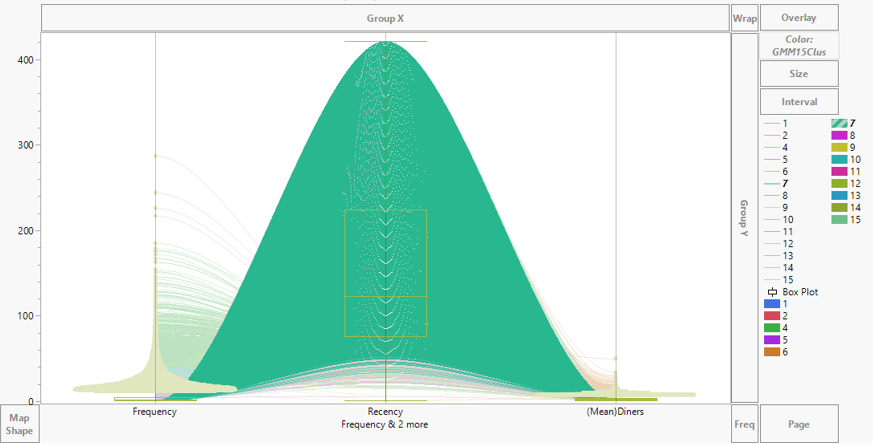
These iterations reveal that Best or Optimal Cluster changes with increase in max.no. of clusters provided. So, let us consider Cluster 15 as an optimal one for this type of clustering.

### 5.4.4 Interpreting clustering results using Parallel Coordinates



This graph reveals that almost all clusters are grouped together. Let’s analyse the clusters separately.

For eg. Cluster1, Cluster7 and Cluster 8 exhibit similar properties. It can be inferred from below figures.

Cluster 7

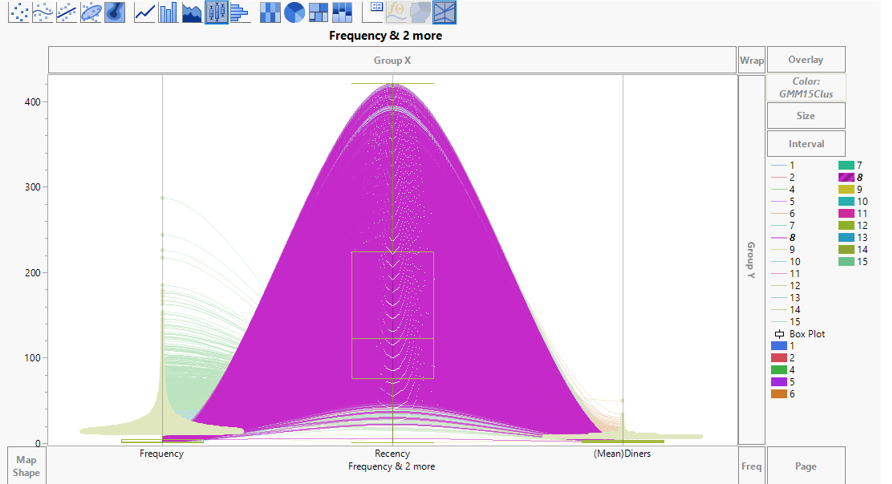
Cluster 7

Cluster 8

Cluster 1

Cluster 7

Cluster 8



Cluster 8

Cluster 7

Cluster 8

Most likely, this clustering method can be eliminated due to inefficiency of separating the clusters individually.

## 5.5 WORKING WITH LATENT CLASS ANALYSIS

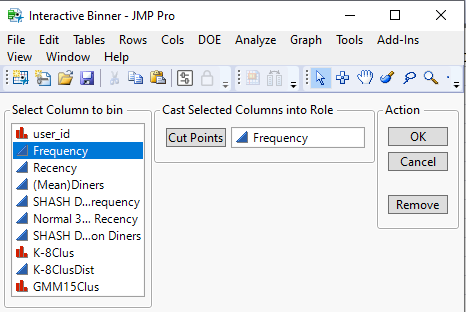
### 5.5.1 Overview

It is also one of the advanced clustering technique, in which clustering variables can be of categorical type unlike all previous methods, which uses continuous variables for clustering.

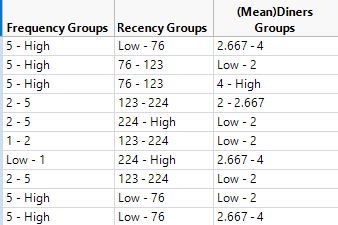
### 5.5.2 Binding Clustering Variables

For Latent Class Analysis, the input variables must be in categorical data type. Let us use the Interactive binning add-in of JMP to convert the input variables into categorical data type. Similar to Random seed ad in, Interactive Binning can also be downloaded from JMP Community and can be installed in the system.

* Let us bin values in *Frequency* field into four equal percentile.
* From the menu bar, select Add-Ins -> Interactive Binning V2
* The Interactive Binner dialog window appears.
* Select Frequency from Select Column to Bin Pane and click on Cut Points.
* Click OK.

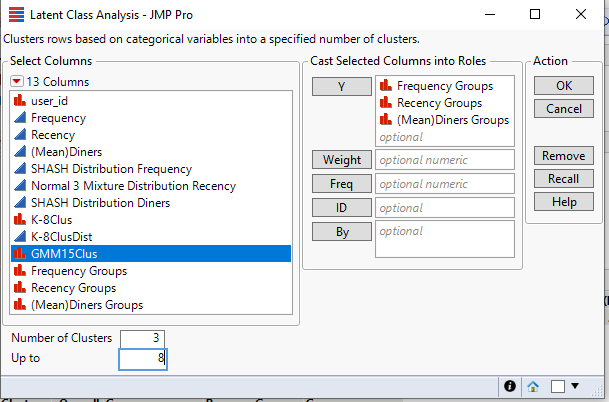


* In the window that appears click on the red triangle in front of **Interactive Binning**, select **Set Cutpoints At Percentiles…**
* Enter 25 in the percentile textbox, which allows JMP Pro to create four categories, and each category consists of 25% of the data points.
* Let us save the newly created bins as a new field in the data table.
* Click on the red triangle in front of **Interactive Binning**, select **Save Group Column** from the context menu.
* Similarly, perform the same steps for Recency and (Mean)Diners.
* The values are categorised into groups as shown in the below figure



### 5.5.2 Performing Latent Class Analysis

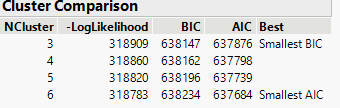
* From the menu bar, select **Analysis** -> **Clustering** -> **Latent Class Analysis**.
* From the **Select Columns** pane, select all grouped categories .Click on **Y** button
* For Number of Clusters, keep it as default i.e. 3.
* For **Up to** option, type 8.
* Click on OK Button



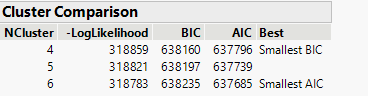
### 5.5.3 Analysing Latent Class Analysis Report

The Latent Class Analysis report appears.

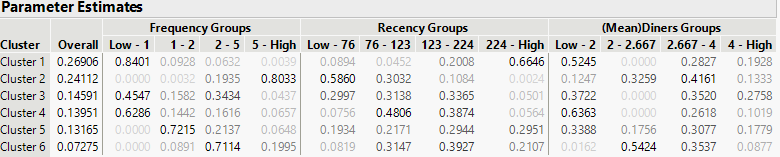
Firstly, let us focus our attention on the **Cluster Comparison** report.



Though we have provided maximum upto 8 clusters, this analysis computed only till 6. Let us try for other range of intervals such as 7-14.



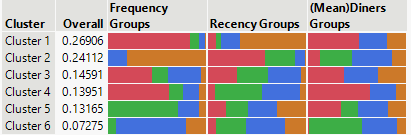
Even then, the number of clusters did not increase more than six. From these iterations, let us conclude Cluster 6 is the optimal one and let us focus on its parameter estimates.



The **Overall** column reveals that Cluster 1, 2,3,4,5 and 6 accounts for 0.27, 0.24, 0.15, 0.14, 0.13, and 0.07 of the total data points respectively.

The remaining columns show the conditional probabilities of the clustering variables.

For Cluster 1, for example, the probability of being a Low-1 Frequency Group given the customer is in Cluster 1 is 0.8401 and for 224 High Recency Group is 0.6646. They sum to 1.0 for the four cells. The table also reveals that the proportion of a Low-1 Frequency Group is relatively larger than the other three cells.



The horizontal stack bar chart in the report provide similar information visually. Since the clusters formed didn’t show clear and distinct characteristics, this may not be very useful for our further analysis.

Click on the red triangle in front of **Latent Class Model for 6 Clusters**, then select **Save Cluster Only** from the context menu.

# 6. INTERPRETATION OF ANALYSIS RESULTS

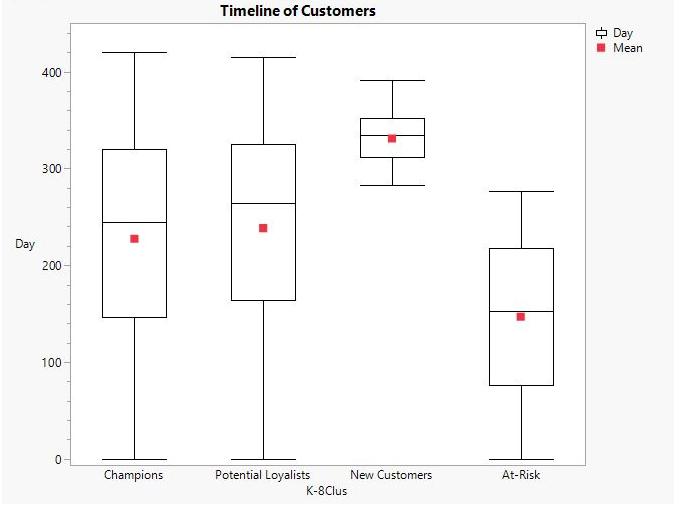
### 6.1 Reasons for Selecting / Rejecting a clustering technique

Now that we have performed different clustering methods, at the end of each section we have justified for selecting or rejecting any method.

* With more than 10,000 observations or/and 50 clustering variables, it is not wise to use hierarchical clustering because the distance matrix is going to be too large. Hence it can be rejected.
* In Gaussian Mixture Modelling, we noticed that clusters don’t exhibit individual characteristics. This clustering method can be eliminated due to inefficiency of separating the clusters individually. They were all present in one region.
* In Latent Class, we observed that with the given data, when we try to perform modelling with different interval of clusters, Optimal or Best cluster is arbitrary. And the smallest BIC and Smallest AIC don’t match to the same cluster even after trying multiple values. Also, the clusters formed in the parameter estimates section didn’t show clear and distinct characteristics .So, it is not recommended to use this type for our analysis.
* This boils down our selection to K-Means clustering for which we were able to get an optimum cluster and it is verified using Biplot in that section which clearly shows us to proceed the interpretation with K means Cluster.

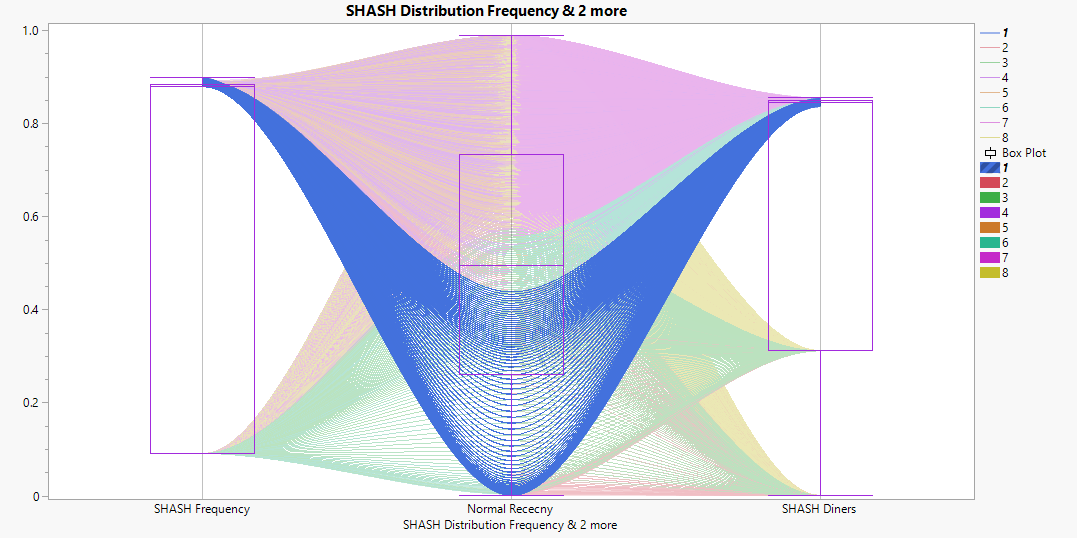
### 6.2 Mapping the cluster to Extracted Data

Using the steps mentioned before, let us join the *RF&Diners* Table with *Extracted\_Data* using user\_id as match column. Let’s do customer profiling by categorising the classes of clusters.

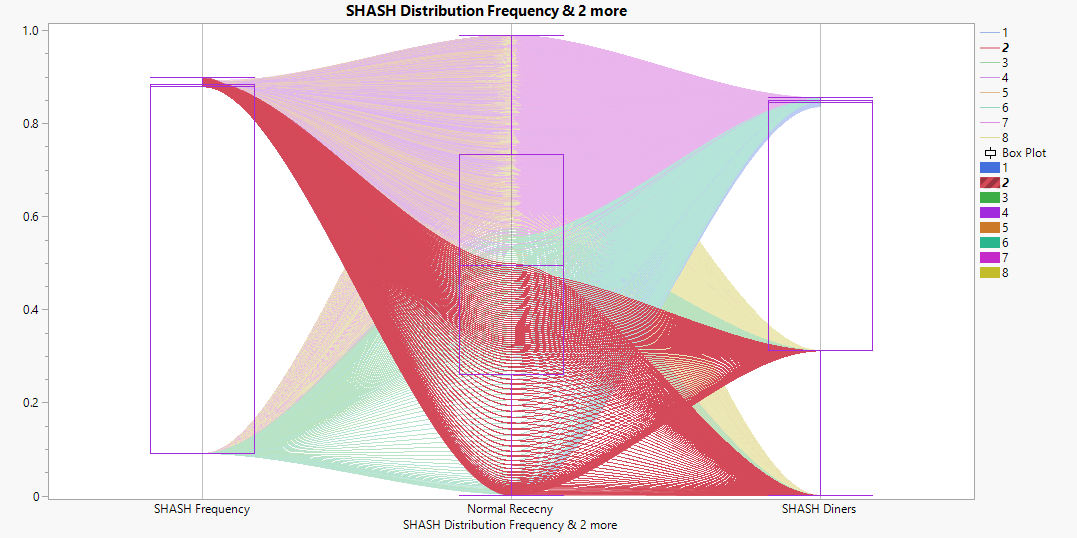
The above graph reveals that Food Online Customers can be classified into the following (Pushpa Makhija, 2021)

* Champions
* Potential Loyalists
* New Customers
* At- Risk Customers

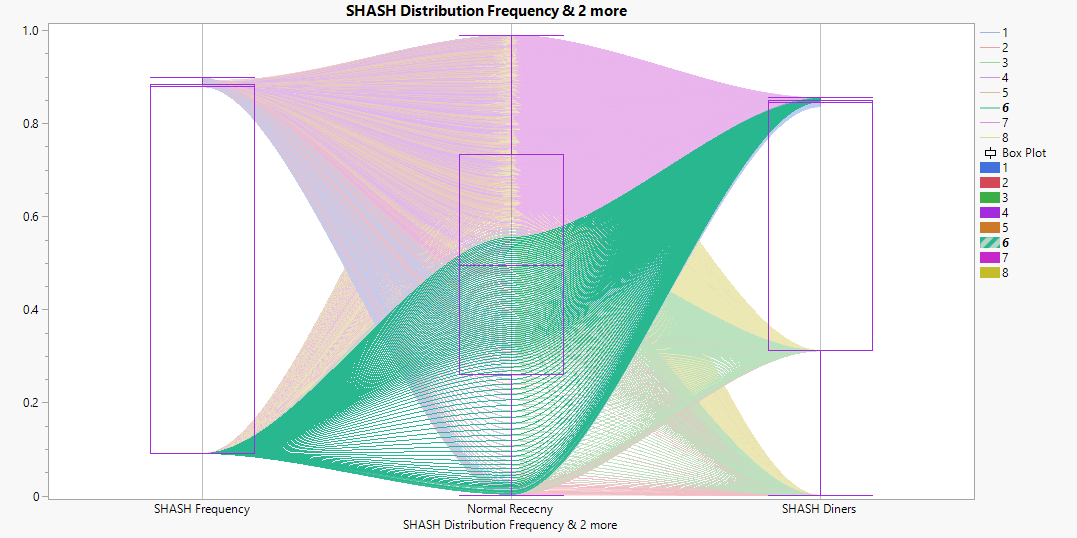
1. **Champions (Cluster 1)** are the customers who booked table most recently, visits frequently, and most often, they book table for more no. of people, which indirectly states that they spend more.



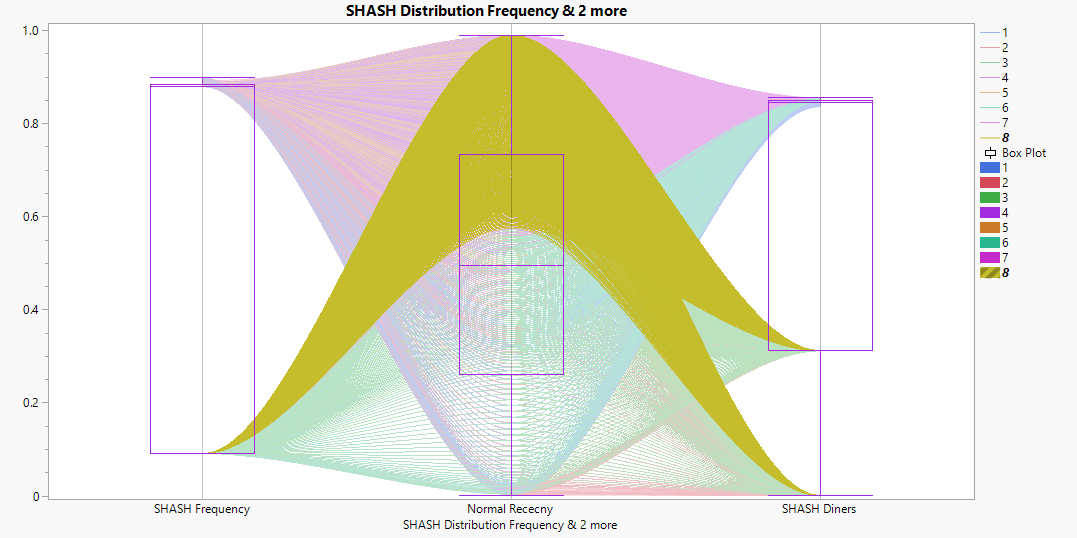
1. **Potential Loyalists (Cluster 2)** are the recent customers with average frequency and who accompany average no. of diners which states that they spend an average amount.



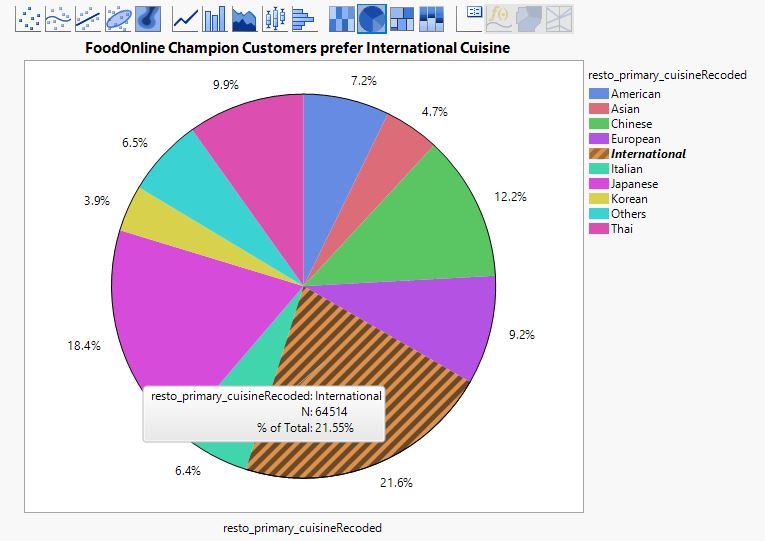
1. **New Customers (Cluster 6)** are the customers who have started booking table recently and they are less frequent in booking.



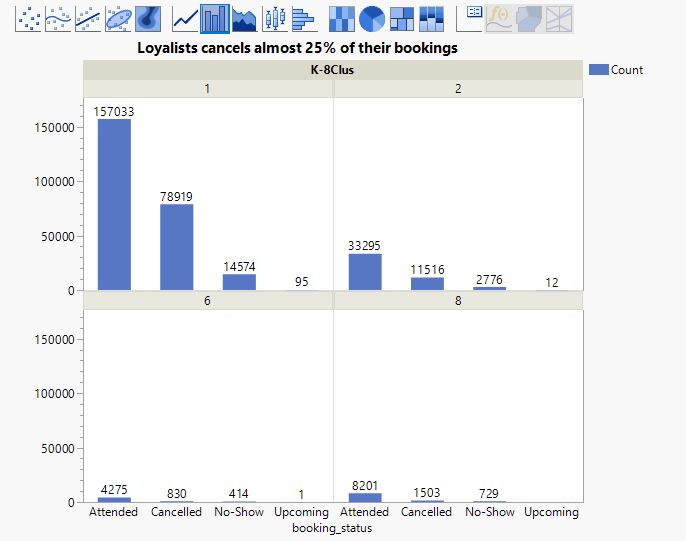
1. **At-Risk (Cluster 8)** Customers are the one who neither book table frequently nor booked it recently. In addition, they accompany low to medium no. of diners.



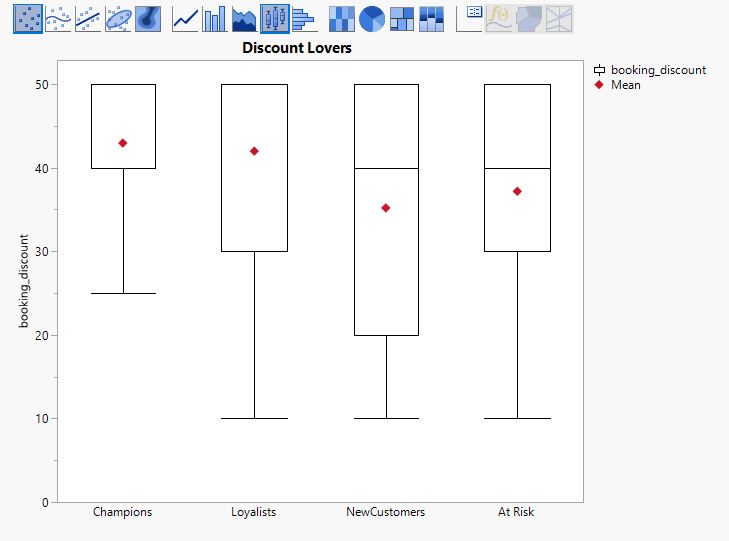
1. Among the Champion Customers, **International cuisine** is most preferred (21.6%) followed by Japanese cuisine (18.4%)



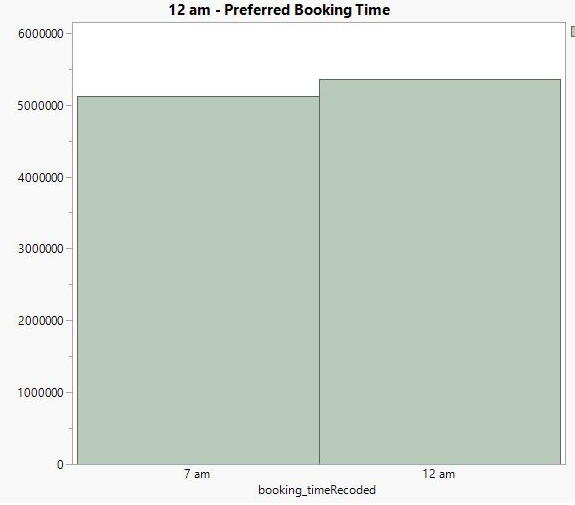
1. This graph shows that **potential loyal customers** (Cluster 2 ) have cancelled almost **25%** of their bookings.



1. It is interesting to notice that the users who utilise **50% discount** are still fall under the category **At-risk**. Though they enjoy discounts, they tend to book tables rarely and their accompanying diners are also less in number. By comparing mean of various customers it is understood that **New customers haven’t provided or haven’t utilised much offers.** It is still lesser than other customers.



1. The below graph shows that among the Champions, 12 am booking time has the slight edge over 7 am users. People are very active in midnight.



# 7. MANAGERIAL RECOMMENDATIONS

The following Interpretations and recommendations are made:

1. Users are categorised as Champions, Potential Loyalists, New Customers and At-risk customers.
2. Champions customers tend to book table most frequently and recently. They also accompany more diners which states that they spend more.
3. Loyal Customers who accompany less no. of diners visit on an average frequency as 25% of their bookings are cancelled.
4. International Cuisine is the most preferred one by the FoodOnline Customers.
5. Customers prefer 12 am as booking time than 7 am. They are all seem to be night owl rather early bird.
6. New Customers haven’t offered much discounts yet.

To further, substantiate the analysis:

1. Important characteristics such as geographical segmentation can be inferred if more accurate data for region is mentioned.( Sarvari, P.A., Ustundag, A. and Takci, H.,2016)
2. Close to 78% of gender, data is missing because of which unable to segment the data using gender, which is again a vital decision maker. ( Friedmann, Enav & Lowengart, Oded, 2018)
3. RFM Analysis could not be done due to the absence of proper pricing values of customers. It is one of the popular marketing technique and that would add more credibility to the segmentation.

# 8. REFERENCES

* Friedmann, Enav & Lowengart, Oded. (2018). Gender segmentation to increase brand preference? The role of product involvement. Journal of Product & Brand Management. 28. 10.1108/JPBM-06-2018-1917. Retrieved from <https://www.researchgate.net/publication/329591440_Gender_segmentation_to_increase_brand_preference_The_role_of_product_involvement/link/5e6dd899458515e5557c999d/download>
* Pushpa Makhija. (2021) RFM Analysis for Customer Segmentation. Retrieved from <https://clevertap.com/blog/rfm-analysis/>
* Sarvari, P.A., Ustundag, A. and Takci, H. (2016), "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis", [Kybernetes](https://www.emerald.com/insight/publication/issn/0368-492X), Vol. 45 No. 7, pp. 1129-1157. <https://doi.org/10.1108/K-07-2015-0180>