SMU
SINGAPORE MANAGEMENT UNIVERSITY

2803: Supervisor Briefing: Cluster Analysis Review

Date/Time 28 March 2018, 2PM – 3PM
Attendees Arushi, Shubhangi, Tanushree

Sr. No.	Notes	Actors	Follow up Action
1	Ran RFM Variables and Clusters with Prof. : Frequency Comments: - Problem with current Frequency measure→ Average time b/w bookings. - We have frequency values like 0. But then that doesn't make sense. - Make definition clear first - One Way: 3 months/6months/year→ Within that time period, how many times does the customer come? How many times the customer booked/the days since the customer joined the eatigo application.	Shubhangi	Remember to update the frequency variables

	- For everyone who were there from earlier, total number of bookings/365. - If the base is not meaningful, you can consider dividing by one month or 3 months or the period you think meaningful		
2	Recency: Calculated as last date of year- Max booking date. Variable passed		
3	Monetary: Currently calculating the average number of diners across the bookings. Look at Total Number averaged number of diners	Shubhangi	Revise the monetary values
4	Notes on the Paper: - Remember to define how you've calculated each - And talk about the cut-off q	Shubhangi	Keep in mind
5	Notes on Standardization: - Only use standardization if the data range is very big Our data is not that wide. Therefore, we need to fix skewness not standardization. Therefore we should do transformation. Log Transformation:	Arushi	Figure out the standardization method and complete by next meeting

	When you log, range becomes wider, then you standardize		
6	Notes on using different data types: - We have lots of variables of different types of data, some proportions, some continuous and numeric. Need to know how to deal with it. Prof Comments Option #1: - Use RFM Analysis - Then Profile. Identify common booking behaviour Option #2: - Standardization doesn’t always have to be Z-Score - Look at the variable that has maximum values - And can be scaled up	All	Decide on which option we would like to go ahead with
7	Using Clustering: K-Means: - Use - Step #2: Based on CCE (statistical method to choose the best number of clusters), decide on the number of clusters.	All	By next meeting have the distributions of variables standardized and transformed ready. Have the initial clustering done

	CCE should not be negative. Two possibilities of negative: - Lots of outliers: If there’s a cluster with only 1 variable, then it’s wrong. It should ideally have equal number in each cluster - Data Skewed Deciding on important variables: - Look at cluster mean and cluster SD. - But since it’s standardized, it’s not reliable to use it to interpret. - Save the cluster. - Go back to the main table, map out distribution of cluster inputs and cluster itself. - Remove the cluster. - Then at the menu on top of graphs, find data filter, then click cluster. - Use that for interpretation. Interpretation #2: - Graph→ Parallel Plot, put all cluster inputs. - Click Data Filter, Cluster. Then look at the parallel plot, colored based on clusters - Uncheck Include and show.		with parallel plots ready to show prof. In case any questions email prof with the doubts.

	Question 2. Outlier is important Clustering Technique #2: If there are outliers, use Normal Mixtures. If the clusters are equal sizes, then use this. Look at BIC and AIC and find the smallest. Rerun the iteration just in case and expand the range. Then check the AIC and BIC. If the small BIC is in one cluster and small AIC is another cluster then compare both the clusters.		