2903: Supervisor Meeting

Date/Time 29 March 2018, 1:00PM – 1:30PM

Attendees : Shubhangi, Tanushree

Sr. No.	Notes	Actors	Follow up Action
1	Running clustering variables after 2803 pointers: - High Skew→ Recently, Monetary, Frequency, Don't Transform→ Resto, cuisine variety. Freq too small to transform Frequency → Multiply by 1,000 to be able to conduct a transformation	All	Work on clustering
2	**Notes on the Paper:** - Remember to define how you've calculated each - And talk about the cut-off q	All	
3	**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:** When you log, range becomes wider, then you don't need standardization	All	

4	**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	
5	**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. CCE should not be negative. Two possibilities of negative: - Lots of outliers:	All	

	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. Question 2. Outlier is important		
6	Clustering Technique #2:	All	

	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.		