TEAM V
ANLY482 SUPERVISOR MEETING
MINUTES 8

Date	8 March 2017
Time	17:20 - 18:00
Venue	SIS Meeting Room 4.3
Attendees	Team V: Andrew, Sarah Supervisor: Prof Kam
Agenda	1. Update supervisor of revised EDA progress 2. Update supervisor of cluster analysis progress

S/N	Item Discussed	Remarks
1	Revised EDA	- Andrew showed the revised EDA for bookings breakdown by monetary value. The intervals now are all equal and the charts are separated. - Andrew also showed the distribution table to show the niche and mass markets.
2	Cluster Analysis	- For this analysis, we filtered users with at least one booking. Hence, instead of 15K users, 5K user profiles are used. - 3 new columns are created following the RFM model. The columns are for the users' recency, frequency and monetary (money spent). - For the monetary value, an average is taken. However, Prof Kam suggested that we do one for total amount spent as well. - Some users did not input that date of birth, so Andrew asked Prof Kam if we should include it in the analysis. Prof Kam said that age is not a crucial factor. - He suggested that, we could instead do an EDA on the clusters after the analysis. This will help us find out the demographics of the users. - Andrew asked Prof Kam if we should click on the default standardisation option before the

		analysis is done. Prof Kam said that if standardisation is done beforehand, this would not be needed. - Prof Kam asked for a live demo of the distribution table for the 3 cluster analysis variables. From the distribution, frequency and monetary are very skewed, whereas recency seems fine. Prof Kam suggested to do a $\log_{10}$ to transform it. - Comparing the results of both the K-means and Hierarchical cluster results, Prof Kam suggested we use K-means because it would converge better. He said Hierarchical is more suited for dataset of a few hundred. In this case, we have a few thousands. - In the presentation slides, "cluster size" should be stated as "number of clusters". Prof Kam asked us to change this. - In JMP, they would state the recommended number of clusters. So we should start with following the recommended one. In this case, the recommended is 20. - Going back to the presentation, the hierarchical clustering results of all 4 variables including age shows the R-square results for the monetary variable become insignificant which means it does not contribute much to the cluster analysis. - In this case, Prof Kam said it supports the reason to remove the age variable from the analysis. - Going back to JMP, Prof Kam said that we can choose a range of cluster sizes. He demonstrated to us using a cluster size range of 3 to 24. - For this, cluster size of 7 seems optimal. It shows the value increasing very slightly from cluster of 5 and 6 before it dips to a cluster size of 8. - Prof Kam teached that after cluster analysis is done, under the red arrow menu, click on parallel coordinate plot to see how each cluster fared. This can be used to inspect the clusters. - There is only one sub cluster that contains one value. We can consider that an outlier. - To determine the optimal number of clusters, use the CCC value. - At the end of the day cluster analysis is to see the homogeneity of the group. If there are too many clusters with only one data inside, then we can single them out as outlier.

S/N	Action Item	Action By	Deadline
1	Continue with cluster analysis and profile the clusters	Andrew, Sarah	By 14 Mar 2017