Analytics Practicum Supervisor Meeting 07

MINUTES NOVEMBER 2, 2016 1600 - 1700 SMU SIS BUILDING MEETING ROOM 4-3

MEETING CALLED BY	Prof Kam
TYPE OF MEETING	Project Briefing
FACILITATOR	-
NOTE TAKER	Chong Xin
TIMEKEEPER	Chong Xin
ATTENDEES	Chong Xin, Bowei, Hui Min

Agenda topics

1600 - 1630 SHINY R VISUALISATIONS ALL MEMBERS

DISCUSSION	- Trying to build all the features before integrating the Huff's Model in Shiny R code. Going through the features: - User choose a library - User chooses a buffer and then the amenities to be filtered - Bowei mentions to include the data report in the buffer circle of the library - R has a geo-sphere library, and it calculates the haversine distance. - Hui Min mentions a problem with the library markers: when selected, the library markers will cumulatively add to the selected list. Prof mentions that since Leaflet has integrated the layer control with R, so we can rely on that instead. - Prof mentions that we can just replicate the Tableau visualizations in the Shiny R. He also mentions that NLB is not going to invest in Tableau. So NLB expects us to deliver something, or implement in QlikView. We should not map Tableau's features entirely out, we should only pick out the good points and integrate with the Shiny R designs. - Bowei mentions adding new (additional) data files in will be very complicated, since there is no database. Initially we wanted to allow them to add additional layers. Profs mention that there is an example he knows, which reads the csv and if it follows the same structure, then a new layer will be added. - Bowei mentions about the "&>&" operator, and mentions that it will be hard to convert the file back to normal R script. It limits the flexibility of the code. Is there a source where he can understand how this thing works? Bowei mentions the GitHub source is not comprehensive enough. Prof mentions the best thing to communicate is to use the Google Groups for Shiny R. - We are talking about a dataframe in R, and we can read several columns in 1 dataframe. Hence the '~' denotes the collection of only 1 column. - Bowei asked about concatenating the LAT and LNG, Prof is unsure about possible solutions. - Clarified that drawing the buffer is working, but to calculate the point-in-polygon is not working. - Currently we are using R code, and it is independent from Shiny. Prof is not too sure about the point-in-polygon for R code at the moment. Hui Min asks if the current project has a wide-enough scope for Analytics Practicum. - Prof says it is fine.

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
- As per mentioned above	Hui Min & Bowei	9th Nov

1630 - 1700 CALCULATING ATTRACTIVENESS ALL MEMBERS

DISCUSSION	- Went through the 2 different measures of demand: (1) proportion of books borrowed (2) proportion of TXNs. - From there, we will calculate the optimal alpha and beta combinations, and then perform regression analysis on the explanatory properties of the different amenities value on the alpha value - Huff models demand->total no. books borrowed by subzone - No. of transactions by subzone for whole year - Total unique patrons - Most of patterns are explained by distance - Information that we lack is alpha and beta, after substituting in the variables we have - After we get the alpha and beta, we can substitute these values into the equation and compare the results for alpha and beta - We will get the postal code data

- We haven’t run the huff model yet, due to the problem of the denominator, there are 5 different terms. There are a large number of alphas and betas, making it difficult to model it.
- Prof suggested multiple regression, but Chong Xin said that it has nothing to do with regression, and is instead finding out the possible combinations of data.
- We need to find out the distance decay, Prof suggested we apply it as a regression and use the least square method, minimizing the difference between expected and predicted demand
- Use probability as the regression, and alpha as experimental variables, plug in the rest of the values.
- Alpha for the library will not change as it depends on the amenities for each library.
- Alpha and beta should give us a range of values, some higher, some lower
- Beta should be more or less constant. If results from analysis does not match probability, we should change alpha
- For testing, start with one of them first. If we test it and it works, we can work with another set of variables first. If we do that, the alpha and beta values will change and we will have to repeat the regression
- Fit model, add variables like collection size, and run, giving us a regression model. Estimate for amenities are alpha value, Estimate for distance is beta value. The relevant attributes are collection size and distance. To be sure, we need to calibrate the model. Technically we should exclude everything except collection size.
- But r square is very low, so something is not being explained by the fit result
- Profiler->Collection size is very big, should have normalized
- Analysis->look at distribution of variables (collection size, etc)
- There are cases of zero night population in subzone. Subzone with zero population should not be borrowing books. Reason is due to in the data, boon lay area population is zero. But there are people borrowing book from boon lay. Maybe we should remove these areas, as they do not explain the regression result and they are anomalous.
- Log base 10 transform collection size, then regress again. Result: nothing much changed.
- Prof told us to send him the data for him to look into it.

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
- As per mentioned above	Chong Xin	9th Nov

OBSERVERS	-
SPECIAL NOTES	- Prof mentioned that he will get back to us on the feedback for our midterm report.