Analytics Practicum Supervisor Meeting 02

MINUTES AUGUST 24, 2016 1630 - 1815 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

1620 - 1640 ADDITIONAL DATASETS SOURCED ALL MEMBERS

DISCUSSION	Facility datasets - For additional datasets, geographical location of schools' sub-point, **we should be more specific on what entails "schools"** - Prof recommends childcare centers might be more important than JC or secondary schools, as most schools have their own libraries. Attractiveness for usage can be more about tuition centers, parents want to kill-time and hence they will visit the libraries more often - **We should think more about private school without a fully-functional library, hence they will be more likely to go to the public libraries** - **We should contrast with the reality, and include some intuitive theories in our proposal. We need to be more specific and pull out the main source when we discuss the project proposal with them** - **To pull out the tuition centers and private school listing from the Yellow Pages** Transport datasets - Incomplete to just say "5 MRT stops nearby". In network science we have something called "**centrality**", not all stations play the same role, some stations are interchanges and some stations only have 1 line. For example, Raffles Place MRT's traffic volume is greater than a station that is not an interchange. Since they have more commuters and more centralized, - Each bus stop may serve a collection of bus services; a bus stop that has more bus services holds more weight than one with fewer services - **The search radius is one aspect of the transport datasets, we need to calculate the centrality of the MRT stations and bus stops (no. of bus services as a proxy)** Geographical datasets - **If a library is located in a particular zone, we may want to determine the catchment area of the library. Some libraries may be located in downtown, some in the heartlands, we want to know the land-use zoning within this area.** - Bowei says to include the population data for each subzone in our analysis. **We can determine penetration rate of each subzone by integrating the population data. For example, if within a subzone we have 2 areas that have the same residential building but have different penetration rate, we may want to explore the reason of the disparity.**

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
As per bolded in the discussion segment	All members	Sponsor Meeting 01

1640 - 1650 UNDERSTANDING COLUMN HEADING OF DATA (LOCALE PLANNING ADZID IN PATRON DATASET) ALL MEMBERS

DISCUSSION	- To ignore the last 4 digits of the values (the running number of buildings in the subzone), and **match the data with the masterplan subzone data** - Either use the planning zone or we use the individual block, but - Current they have the postal code but they do not have the x-y coordinates - In the future we might have another set of data, with the postal code fit inside

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
As per bolded in the discussion segment	All members	Sponsor Meeting 01

1650 - 1710 RFM ANALYSIS ALL MEMBERS

DISCUSSION	- Recency as one of the indicator, intuitively we value a person with a more recent last transaction. - **Look at the distribution of each of the RFM variable and discover any patterns. Discover any records with errors.** - Prof says we should not use 2 variables if we want to conduct cluster analysis. **Before we do cluster analysis, we need to find out the distribution of the variables. If the variables are highly skewed, we need to transform the variables before we conduct the analysis. Also to normalize/standardize the variables.** - Prof recommends us to use JMP Pro 12 to conduct the cluster analysis. - **Asked about the Patron_UID '0', Prof mentions that we need to list down all the anomalies and discuss with the sponsor about the origins of the anomalies.** - We are able to determine the number of books borrowed per transactions in the data, and we should record down these anomalies and decide if (1) drop them from the analysis (2) discuss with the sponsor - **Need to conduct data exploration and check for errors and anomalies.** Some patrons may not even have a geographical subzone assigned. - We did not check the distribution before we standardize the data. Prof recommends using the Summary function instead of the Tabulate function. - Showed cluster analysis to the Prof, Prof showed us how to use JMP Pro 12 to perform cluster analysis.

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
As per bolded in the discussion segment	All members	Sponsor Meeting 01

1710 - 1740 HUFF'S MODEL ALL MEMBERS

DISCUSSION	Huff's Model - **When we design the application, we will check for the p-value of the newly added variable for the specified significance level** - Set a control for the p-value, if it is greater than threshold level, then we will take in the variable - Can also set the threshold level for the p-value - **Distance decay function – we can use the actual distance values to derive the function** - **Need to fine-tune the schools and institution parts** Determining distance between subzone & library - **Prof recommends to use a centroid**, but also lists some limitations with this method, that there may be neighboring subzones with a patron residential address in the middle of 2 libraries, and we wrongly assign the patron to the further library instead - Assumption of the centroid method is that population is uniformly spread out across each subzone - In the proposal, we need to take previous done-work into consideration, to look at areas where we can improve on

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
As per bolded in the discussion segment	All members	Sponsor Meeting 01

1740 - 1750 REGARDING TECHNOLOGY ALL MEMBERS

DISCUSSION	- We need to decide whether to use Apache Spark or SpatiaLite - People in NLB hope that we can give them something that they do not have to bother with the backend issues - Allow users to upload specific file types like .csv, but we do not need to do an extended list

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
-	-	-

1750 - 1815 SPONSOR MEETING SCHEDULE & FEEDBACK ALL MEMBERS

DISCUSSION	- To include the Gantt chart in the project proposal - The people who worked initially with the SLA team are no longer there. And the current group of sponsor is unable to use any insights from the SLA model. The current group has no idea what they can expect from this project. They want to be able to visualize (if we have our library distribution, I want to be able to visualize – when I select this library, where are the patrons coming from? We should be able to do it concurrently – **visualize the current catchment area of the NLB libraries** - **When end-user select a library, show catchment area visually, and also show additional information such as number of bus stops, number of MRTs.** - Sponsor meeting to be scheduled next week, to **get out the visualization before the meeting (if possible).** See whether NLB prefers this or prefer the Huff's Model proposal - If the sponsor decides to change the idea, we need to change the proposal. Document out the time and reason for any changes made. 4 people, 1 director of corporate analytics, real contact person. 2 persons from the corporate planning and library side.

ACTION ITEMS	PERSON RESPONSIBLE	DEADLINE
As per bolded in the discussion segment	All members	Sponsor Meeting 01

OBSERVERS	-
SPECIAL NOTES	Next Supervisor Meeting (03) will be tentatively scheduled after the Sponsor Meeting 01, which will be scheduled next week. All members will present on their findings in the Team Meeting prior.