Supervisor Meeting 5

Date & Time: 08 Feb 10:30am - 12:00pm

Venue: SIS MR 4.6

Attendees: Prof. Kam Tin Seong (Supervisor), Wang Sijia, Ren Mengxi, Wang Tianjing

Absentees: Null

Agenda:

1. Consult prof on the problem of wrong matching
    a. 178/8670
2. Update prof on R learning progress
    a. Shiny free to deploy?

Details:

1. Prod Kam suggested that the wrong matched data should be excluded from our analysis for the time being because 2% is not that big proportion. Rerun all the analysis after excluding. Make sure our sponsor is aware of the problem.
2. Study on the individuals who come to library most frequently. What business value can we bring to our sponsors by doing this? Prof Kam said that don't guess what they are looking for but exploring the usage level and who are using the library. This allows them to gain insights. Separate terms is more useful than the whole 6 months. Look at the distributions and see is it true that most of the students have low frequency to come to school.
3. First thing is to see the overall distribution. Next is to see whether this pattern applies to all school. (distribution by school, stack to have a better view) SOB students averagely go 16 days, SIS is less only 12 days, SOL have the highest with 30 days. This is able to show that how different school distribution like. Some are more skewed than others (compare diamond and the middle line -> SOL is closer). DIamond is the mean and middle line is median
4. Next thing is to check whether they are really statistically significant :No. Of rows/ School→ one way analysis → quantiles , SOL highest, how diff is every school, -> compare mean- all pairs -> confirm that SOL is statistically significant higher than all other schools with a low p value, compare the difference
5. Next the same logic can be used to compare within schools, to see whether there is a difference between different year of study. Row->Data Filter->select a specific school to study (to unfilter: row -> clear row states).
6. In order to study years, a new column need to be derived which includes like Year 1, Year 2 etc. (ordinal) According to the formula suggested by prof which uses '2016-Admission Year' as a benchmark, there will be some year 0/5/6 students. We may consider consider combine year 0s with year 1s, and year 5,6 with year 4.
7. Through the demonstration from prof we get to see that the more senior students tend to spend less and less time in lib, and this effect is more obvious for SOL students compared to SIS.

8. Can draw comparison between Term 2 & Term 3A. But main analysis should base on Term 2 because some SOA students are compulsory to take Term 3A so the date may be bias.
9. At per user level, it may be interesting to study how many students are doing single visit a day and how many are multiple visits, how frequent they visit the library during the week and the day.
10. Next step for the segmentation, use student behaviour (frequency, the day of week, the hour of the day) as the indicator. Try to come up with more meaningful segments instead of differentiating by schools.
11. An error has been founded in our current logic of defining alumni. The students with a graduation year of 2016 actually graduates in May 2016, so they are still studying as current enrolled students during the first half of the year. Therefore we need to refine our formula to exclude this group of people from alumni.
12. In terms of R, what kind of technical to use depends on what kind of report we expect. Static report in HTML or PDF format can be achieved using standard R codes. In this case data preparation and exploration using JMP need to be translated using R. So for future use our sponsor just need to rerun R script and generate updated report. Another method for interactive report can use Shiny, which is not necessary for our project.
13. Some R library recommended: dplyr, tidyr, rmarkdown, knit, ggplot, rcolorbrewer
14. Confirmed that there will be a 30 minutes time slot during week 8 for midterm review. Result expected include: clean set of data, data preparation & exploration fully completed, show useful insights.

Action Plan:

Item	Person in charge	Deadline
Inform supervisor about our mitigation about the data problem	Sijia	Feb.10
Clean out wrong-matched data and rerun exploratory analysis	Mengxi, Tianjing	Feb.14
Prepare for midterm presentation	All	Feb.16