DATA WRANGLING 422

GROUP PROJECT DIARY

Group members:

Kim Van Nguyen - 68569443

Lianyin Liu - 17899438

Swapna Josmi Sam

Ancy John

Lab 7 - 13 September 2019

Lianyin Liu took the 3pm lab session to do searching for sources because she couldn’t join the group at 10am session.

Swapna, Ancy and Kim went to 10am session and also looked for datasets. We used Google to find dataset and also searched on the World Health Organization, United Nations and Stats NZ websites. Kim found files about health issues such as tuberculosis, malaria. Ancy found files about special species on biology sites. It was easy to find 2 datasets related to each other but we wanted to work on 3 datasets so that made the finding process much harder. We also had to consider what variable we could used as a primary key to join 3 datasets together. At the end of the day, we still could not find our datasets.

Lab 8 - 20 September 2019

All group members presented on the day and kept finding suitable datasets.

Kim found “public expenditures on education” dataset from the UNESCO Institute of Statistics source that includes government expenditure of all countries around the world over time. Kim found “public expenditures on education” interesting, and would be useful to see what the expenditure is for, for example teaching staff, facilities, infrastructures, curriculums, etc. and whether they are effective on the outcomes of education such as number of enrollment or labour and unemployment rate. Therefore, we change our focus to look for these related datasets.

Lab 9 - 27 September 2019

We found “Teaching staff” dataset and “Enrollment in education” dataset from the United Nations website. We finally had 3 datasets:

- Education at the primary, secondary and tertiary levels
- Teaching staff at the primary, secondary and tertiary levels
- Public expenditure on education

Things we would like to find out from these datasets:

- Student-teacher ratio
- Relationship between public expenditure on education and quality of education

Lab 10 - 4 October 2019

After asking lecturer Thomas more details about the requirement for the group project, we understood that the more sources we had the higher mark we could get. Because the “Teaching staff” dataset is from the same source as “Public expenditures on education” dataset, whereas we want to have different sources for each data, so we decided to not take the “Teaching staff” dataset. Then, Kim found “Labour and Unemployment” dataset from World Bank website to be the third dataset. Our datasets finally are:

- Enrollment in Education:
  http://data.un.org/
- Labour and Unemployment:
  https://databank.worldbank.org/source/education-statistics-%5e-all-indicators#
- Government Expenditure on Education:
  http://data.uis.unesco.org/

Pairwise presentation and feedback of other groups:

We discussed on our projects works with the pair group. Kim and Swapna did 1 minute feedback presentation about the other group. We always worked together as a team, and did contributions.

After talking with other group, we decided to do scraping and API which we had not yet thought of for our project. Swapna is confident with these so she volunteered to do that part.

Lab 11 - 11 October 2019

Group work: wrangling data. Ideas were shared on deciding the final data model and helped each other when we were stuck with errors while coding because of which we were able to save our time. So we can tell that group work can reduce the extra efforts which we face when we do it individually.

Ancy wrangled “Enrolment” and “Labour and Unemployment rate” datasets.

Lianyin wrangled “Expenditure” dataset.

Kim did joining and plot some graphs.

Swapna did the web scraping and API for New Zealand schools enrolment.Created Google API key and tried to plot the NZ map using the ggmap package which extracts longitude and latitude from Google Map.

We shared our work in Github:https://github.com/kimvan13/datawrangling

Lab 12 - 18 October 2019

Final public presentation:

- Kim did Introduction, Choice of topic, Analysis (datasets and sources), Libraries (in R that we used), Challenges and Conclusions.
- Ancy, Lianyin and Swapna talked about their works, the original data, how they wrangled it, and the outputs of the datasets they worked on.

22 October 2019

We assigned the work for writing a report as below:

- Kim did Data sources, Targets, Difficulties, Techniques.
- Swapna did Introduction.
- Swapna, Ancy and Lianyin did the Outcomes of their work and Conclusions.