DATA422 Project Dairy

Week8

2020-09-17 Lab 8

Group meeting

Narrowed down the scope of the project to social media websites, such as, Twitter, Linkedin and FB.

Week9

2020-09-24 Lab 9

Group meeting

We feel that the topics we chose before would result in too much workload. We decided to switch on to Covid 19 topic. We have a new group member, Anoop.

2020-09-29 before the lecture

We discussed regarding the project. We proposed another possible topic: Donald Trump's posts on Twitter.

Week10

2020-10-01 Lab 10

Applied for authentication, get access to Twitter API. We also tried to get authentication for TripAdvisor, but they do not allow us to do so. At meanwhile, we thought about if NZ Stat website would possibly have the dataset, which we need for Covid-19 topic. We browsed the dataset from NZ Stat website, looking for dataset that we are interested in and with the right level of cleanliness. We downloaded several datasets from the website. However, we noticed that the datasets sourced from Stat NZ are not suitable for our project for two reasons: the data is very tidy and there is no much wrangling work relevant; the data doesn't contain enough information regarding Covid 19. Therefore, we have decided to get data from other sources to get data regarding the impact of Covid 19 on such as economics, education etc.

A couple of days later, we received an email from Twitter customer service team, regarding the API we applied and asking for more detail, plus we already gathered a set of datasets that would fit our project with a good level of data wrangling required and wide range of potential subtopics to choose from, we decide to go ahead based on the dataset we already acquired.

Data Source:

WHO Coronavirus Disease Dataset:

https://covid19.who.int/

World Bank Open Data:

https://data.worldbank.org/

Spatiotemporal data for 2019-Novel Coronavirus Covid-19 Cases and deaths:

https://data.humdata.org/dataset/2019-novel-coronavirus-cases

Note: Writing code with comments to make the code more understandable.

Our project is divided into two parts, one is the impact of Covid on education, the other is the impact of Covid on the economy. Shanshan Liu and Xue Yang are responsible for economic part. Jiangwei Wang and Xiaoxi Guo are responsible for education part.

Week11

2020-10-08 Lab 11

Shanshan Liu: Wrote data wrangling using R and got final datasets Summary_by_areas.csv and Year_Quarterly_GDP_by_country.csv. Using the final data for visualization.

Xue Yang: Wrote data wrangling using R and got final datasets GDP_History_2010_2021_by_country.csv and RealGDP_by_areas.csv. Using the final data for visualization.

Jiangwei Wang and Xiaoxi Guo: Wrote data wrangling using R and got final dataset Covid_pop_edu.csv. Using the final data for visualization, such as bar chart, line chart, map etc.

Jiangwei Wang and Xiaoxi Guo: Extract unemployment data, then combine population and Covid 19 data into one data frame using Julia. They generated a Unemployment.csv.

Shanshan Liu and Xue Yang: Complete visualization task based on the tidied Unemployment data frame in Julia.

Next week, we are going to give a presentation regarding our project during the lab time. We are instructed to cover the following points:

- Data source
- Process (job flows)
- Challenge (success, failure)
- Delivered dataset
- Ethics/Legality (privacy, intended use, likelihood to cause harm)

Discussion with other groups and Emil.

Questions for our project:

Q: how we combine data frame? A: by country code

Q: where we get the education data and what it looks like? How combine the education data with the others A: we sourced the data from the WHO Open data, the data regarding education is about the status of education institutes (we are going to do data wrangling to remove duplicates before combining data).

Q: Challenges?

Different formats combine dataset from small countries. Data selection (for example, small countries don't have values for some variables).

Q: Ethics? Privacy?

A: the data we sourced are all from open data sources

One group will do project on HIV cases worldwide. Their major concern is about missing data strategies.

The other group will do project regarding game (4 data sources). Data source ready, code ready, API ready. Challenge: the coin prices dataset is huge. Keyword is another problem. Twitter API is not available directly.

Week12

Prepared the presentation slides

Shanshan Liu: prepared introduction, data source and economic part of data visualization.

Xue Yang: prepared data wrangling processing and delivered datasets about economic part.

Jiangwei Wang: prepared data wrangling processing and delivered datasets about education part.

Xiaoxi Guo: prepared data visualization about education part, Ethics/Legality and intended use part.

Presentation day. This Monday and Wednesday, we had group meetings to discuss the presentation, preparing ppt slides and rehearsal.

Week13

Xue Yang created a shared OneDrive folder for easy organization project.

Shanshan Liu was responsible for group diary keeping.

Jiangwei Wang and Xue Yang organized and merged R and Julia files according to the topic, and finally got 2 R files and one Julia file.

Xiaoxi Guo was responsible for writing up Document Description.

Project report:

Shanshan Liu is responsible for Data Sources, Why you chose those data sources, What target you chose and What difficulties you had to overcome to wrangle the data sources into the target data model.

Xue Yang is responsible for What techniques you did use (a) Economy.

Jiangwei Wang is responsible for What techniques you did use (b) Population, Education and Unemployment including Julia and R.

Shanshan Liu and Xiaoxi Guo are responsible for Data Visualization.

Xiaoxi Guo is responsible for What you managed to achieve and what you failed to do, Legality and Ethics and Intended use.

Shanshan Liu and Xue Yang are responsible for grammar checking and structure adjustment.