Assignment 1

Giancarlo Samayoa, Guanhao Hu, Junyi Hua, Qing Yin

January 29, 2024

Part I

In order to analyse labour actions and build a predictive model, a dataset would need to contain information on labour action events such as location, duration, success, date, and size. Ideally these data would be sorted by individual labour action events and contain information dating back multiple periods of study.

Additionally, for greater accuracy these data would ideally include information on working conditions in the region of labour action, geo-located economic data, and information on average wages adjusted for inflation for each labour group in question. In terms of tidiness the data would also ideally have each individual event listed separately with any action involving multiple events listing sub events in separate rows.

Part II

The actual LAT data is more lacking than we would like. We initially thought the LAT data would include more information, such as the wages of workers participating in labor action and some geo-economic information. To this end, if necessary, we may consider introducing some new data to add more control variables to our model.

Moreover, the actual LAT is more complex than our ideal. We may need to put more effort into data cleaning. For example, in the data that records the city where labor action events take place, many events take place in multiple cities at the same time. Therefore, in the column "city," there may be multiple cities placed in the same row. This will put additional demands on our data cleaning. Therefore, before we process the data, this reminds us that we need to understand the meaning of each column in advance and then clean it according to our requirements.

Part III

The project will largely consist of data cleaning, modeling, and visualization, the most time consuming step being data modeling and the least time consuming steps being the remaining two steps. In order to efficiently and fairly distribute the workload, all members of the group will participate in each step of the project (i.e. data cleaning, modeling, and visualization) and share these models and their results with the group on a weekly basis.

GianCarlo Samayoa will be responsible for leading and organizing the weekly meetings. At these meetings, group members will propose and agree upon the specific steps that we shall take to complete this project. Furthermore, all group members will be responsible for presenting the results of their assigned tasks from the prior week. In the meetings these results will be presented via google slides, Zoom, Discord, or other video conferencing applications. Lines of communication will be kept via a Discord channel and a Whatsapp group chat.

Edison Hu will create a repository from his GitHub account to store all the collaborative coding files and both input and output files. Other group members can create a project environment in RStudio linking to Edison's GitHub repository then they can pull and push all the files. Our group members can view and write codes at their local environment and upload to GitHub. We will use one same branch under the repository and only Edison will be in charge of setting the directory and paths in order to maintain consistency. We will ask our group members to avoid pushing changes at the same time and comment on their editing history when pushing to the repository.