Lab Journal Questions 1

Jaechul Jung

1/28/2022

1. What does it mean to “Tidy” your data?

Tidying a data means to make a data where every variable is stored in its own column, every row corresponds to a single case, each value of the variable is stored in a cell of the table. Variables of a tidy data set should not contain units and there should be no extraneous information.

1. What does it mean to “Transform” your data?

For the purpose of this class, transforming a data would mean to literally transform the set of data in a way that is easier to visualize and explain the cause and effect.

1. How important is the “data visualization” skill for your career? (Provide examples)

Data visualization is super important to keep everyone in the loop of what the presenter is trying to convey. Examples include the report of the number of COVID cases by each month using barplot or the percentage of people’s response to a certain survey using pie chart.

1. Choose a visualization and describe the data and mapping/encoding used. (Attach the image and source)

The data represents the number of people vaccinated across the US where each state has different tone in color depending on the percentage of people vaccinated. UI provides a way to single out a state, choose the age range, and many more to interact with.

Graphical user interface

Description automatically generated

<https://www.microsoft.com/en-us/ai/ai-for-health-covid-data?rtc=1>

1. What is the difference between the following command lines

git config user.email “[email@example.com](mailto:email@example.com)” and

git config —global user.email “[email@example.com](mailto:email@example.com)”

Using -global means you want to set the e-mail address for every repository in the computer whereas the one without it sets the e-mail address for a single repository.

1. What are the five key dplyr functions described in the chapter 5?
2. filter & slice: pick specific observations (i.e. specific rows)
3. arrange: reorder the rows
4. select: pick variables by their names (i.e. specific columns)
5. mutate: add new calculated columns to a data frame
6. summarize: aggregate many rows into a single row
7. What are cumulative and rolling aggregates?

They are the result of running sums, products, mins, and maxes over the set of data cumulatively.

1. What are ranking functions?

By default it ranks the value of a data from its smallest. Two equal numbers are obviously ranked the same x and then the following small number will have the rank x + 2.

1. Imaging that you are already assigned to work with a team, what does that team need to do to make that project a success? Provide concrete examples.
2. There needs a good communication among team members.
3. Each team member needs to be brave enough to say they’re behind schedule if it’s the case.
4. The team needs a plan so they know the process, risks, resources, requirements, etc. to make their project a success.
5. What is the most important difference between generic software product development and custom software development? What might this mean in practice for users of generic software products?

Generic software product development is developing a product in a way that meets the demand of the general population whereas custom software development focuses on a particular client’s need and cannot be generalized to the public. It means that not all clients will be satisfied with the generic software products. In that case, they’ll have to look for a custom software product.