Stat 145, Thu 4-Feb-2021 -- Thu 4-Feb-2021

Thursday, February 4th 2021

Topic:: Data

Read:: Lock5 1.2

New R commands:

 nrow(): nrow(ssurv)

 filter(): filter(ssurv, speedtickets >= 3)

 subset(): subset(ssurv, select=c(oncampus, speedtickets))

 sort(~cds, data=ssurv) does not work, though it feels like it should

    with(ssurv, sort(cds))   does what the above appears to intend

    sort(ssurv$cds)   this also works

 gf_histogram(): gf_histogram(~speedtickets, data=ssurv)

relationships between variables:

 tally:

   tally(~selfhandedness | dadhandedness, data=ssurv)

   tally(~selfhandedness | dadhandedness, data=filter(ssurv, dadhandedness!=""))

 addmargins():

   addmargins(tally(~selfhandedness | dadhandedness, data=ssurv))   or

   tally(~selfhandedness | dadhandedness, data=ssurv) %>% addmargins()

 gf_bar()

   gf_bar(~selfhandedness|dadhandedness, data=filter(ssurv, dadhandedness!=""))

 gf_histogram()

   gf_histogram(~speedtickets | oncampus, data=ssurv)

   gf_histogram(~speedtickets | oncampus, data=filter(ssurv, oncampus!=""))

Complete these tasks in groups. Discuss details of how to do these things, and record the answers you find, along with specific commands that assist you. If some sidelight question comes up, feel free to write about it and the modifications to commands that would help get at their answers.

To record your work, point your browser at this link to Etherpad

`https://pad.disroot.org/`

When prompted for a name of a "pad" to open, use the name

`s145-04feb2021-g`X

but replace the X with your group number. Use hyphens not underscores, and no capital letters. Make sure you get this right, or I will not be able to find your work. Use the "users" icon at the far right of the tool bar at the top to add your first name. Delete the initial text, beginning with "Welcome to Etherpad!" and concluding with a website. Replace this with a list that includes the full names of all users in this day's Group X.

The tasks/questions:

1. Load a package called `NHANES`, whose primary purpose is to make available a data frame also named `NHANES`. Though the data frame is already named, but make your own copy of it and give it a convenient name, perhaps using a command like

   ```
   nh <-   NHANES    # makes a copy called "nh"
   ```

   after which you can do all your work using either this copy or the original. (Note: Following this, `help(nh)` does not give you the information that `help(NHANES)` does.) Determine how many cases are in the data frame, how many variables, what is the target population, and whether this can be considered a *simple random sample* from that population.

2. One of the variables in this data frame is named `Poverty`. Is it categorical or quantitative? How is it measured? What proportion of people in this data set live below the poverty line?

3. What is the largest number of pregnancies (see variable `nPregnancies` among respondents?

4. One quantitative variable in this data frame is `BMI`. Does it work to create a frequency table on this variable? Do you see any drawbacks to using `tally()` on a quantitative variable? Does your observation apply to all quantitative variables? (How does it work out if you use it on `nPregnancies`, for example?)

5. Execute the command

   ```
   gf_histogram(~ BMI, data=NHANES)
   ```

   which produces a histogram. How is this like a bar chart? How does it differ from a bar chart? Can you describe in detail the process the software follows in order to produce this histogram from the `BMI` data?