More practice of ggplot2

Math 241, Week 2

```
# it's good practice to check that all the packages required are loaded and installed
libs <- c('tidyverse','dplyr','ggplot2','ggmap','knitr','viridis','mdsr')
for(l in libs){
  if(!require(l,character.only = TRUE, quietly = TRUE)){
    message( sprintf('Did not have the required package << %s >> installed. Downloading now ... ',l))
    install.packages(l)
  }
  library(l, character.only = TRUE, quietly = TRUE)
}
```

Goals of this in-class activity:

- Getting you even more accustomed to using `ggplot2`.

Notes:

- When creating your graphs, consider context (i.e. axis labels, title, ...)!
- If I provide partially completed code, I will put `eval = FALSE` in the chunk. Make sure to change that to `eval = TRUE` once you have completed the code in the chunk.
- Be prepared to ask for help from me, Tory, and your classmates! We scratch the surface of `ggplot2` in class. But I encourage you to really dig in and make your graphs your own (i.e. don't rely on defaults).

Problem 1: US counties

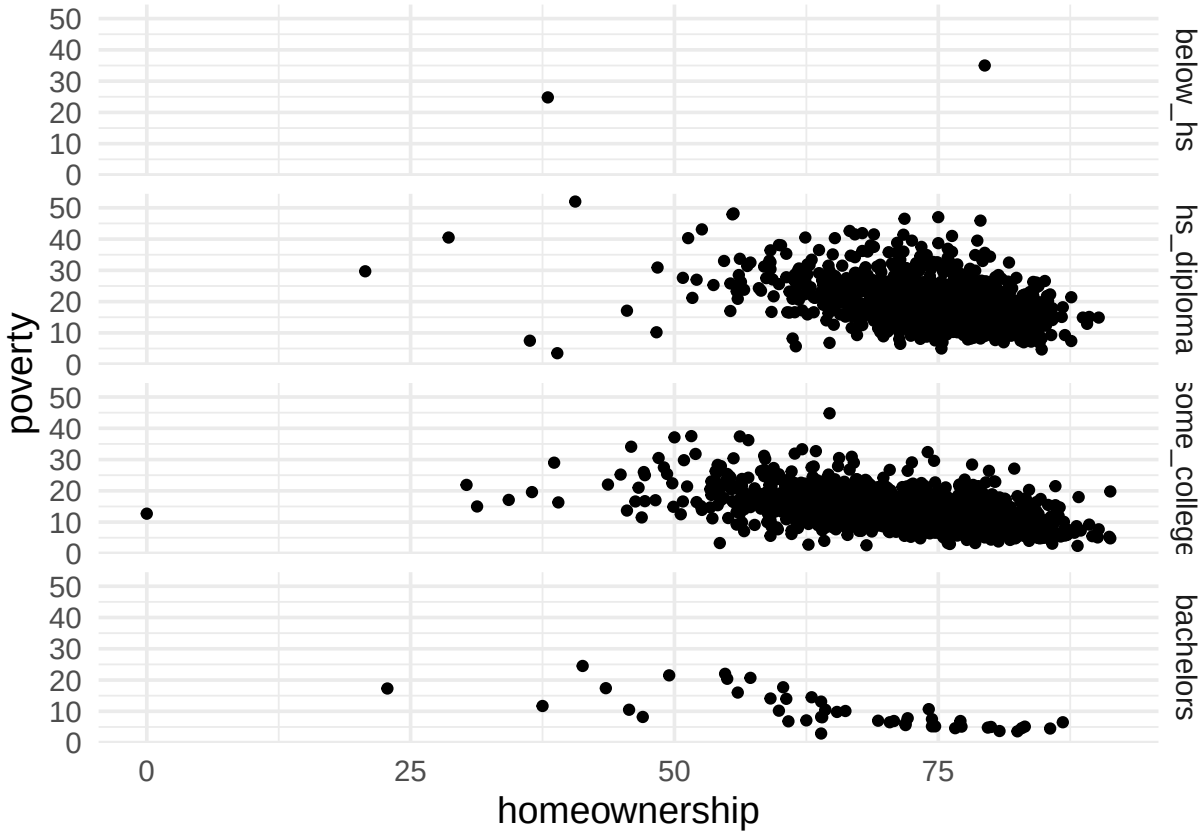
The following questions use the county dataset in the `openintro` package. You can find out more about the dataset by inspecting its documentation with ?county and you can also find this information here.

a. What does the following code do? Does it work? Does it make sense? Why/why not?

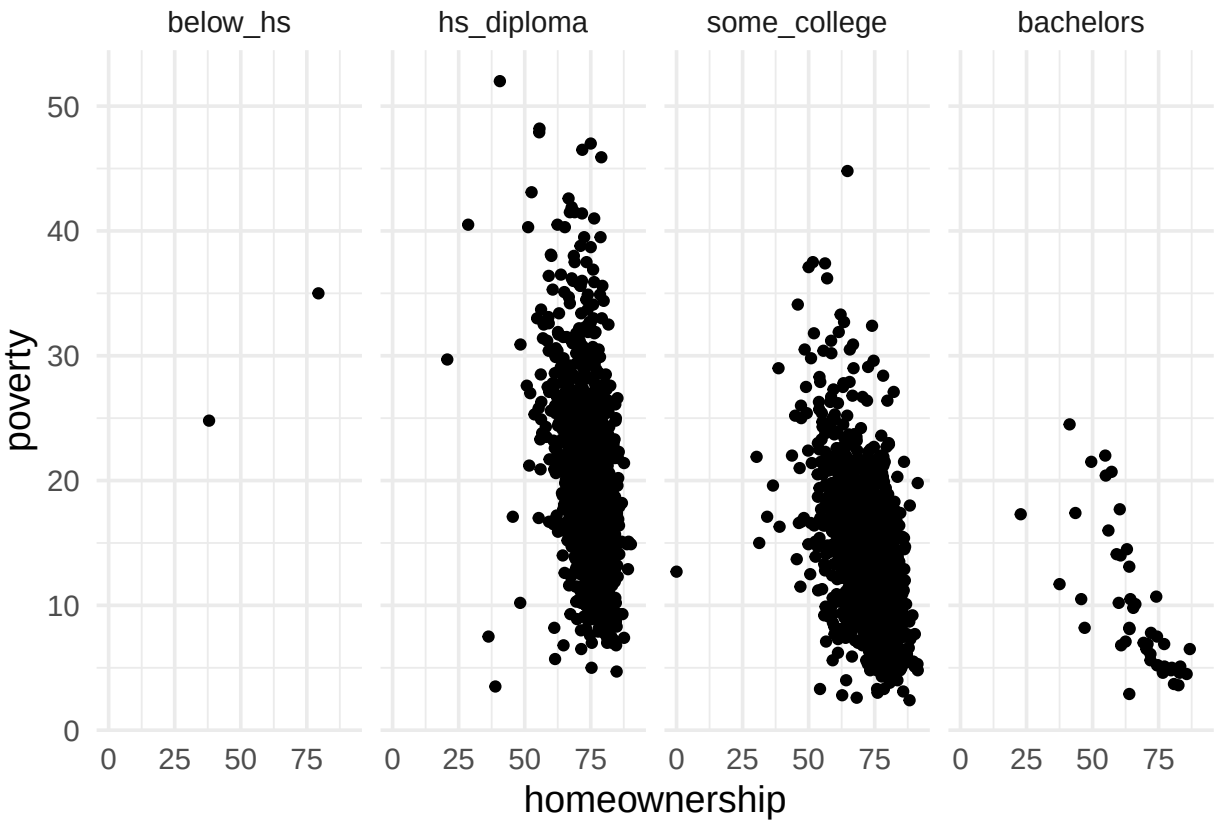
```
ggplot(county) +
  geom_point(aes(x = median_edu, y = median_hh_income)) +
  geom_boxplot(aes(x = smoking_ban, y = pop2017))
```

b. Which of the following two plots makes it easier to compare poverty levels (`poverty`) across people from different median education levels (`median_edu`)? What does this say about when to place a faceting variable across rows or columns?

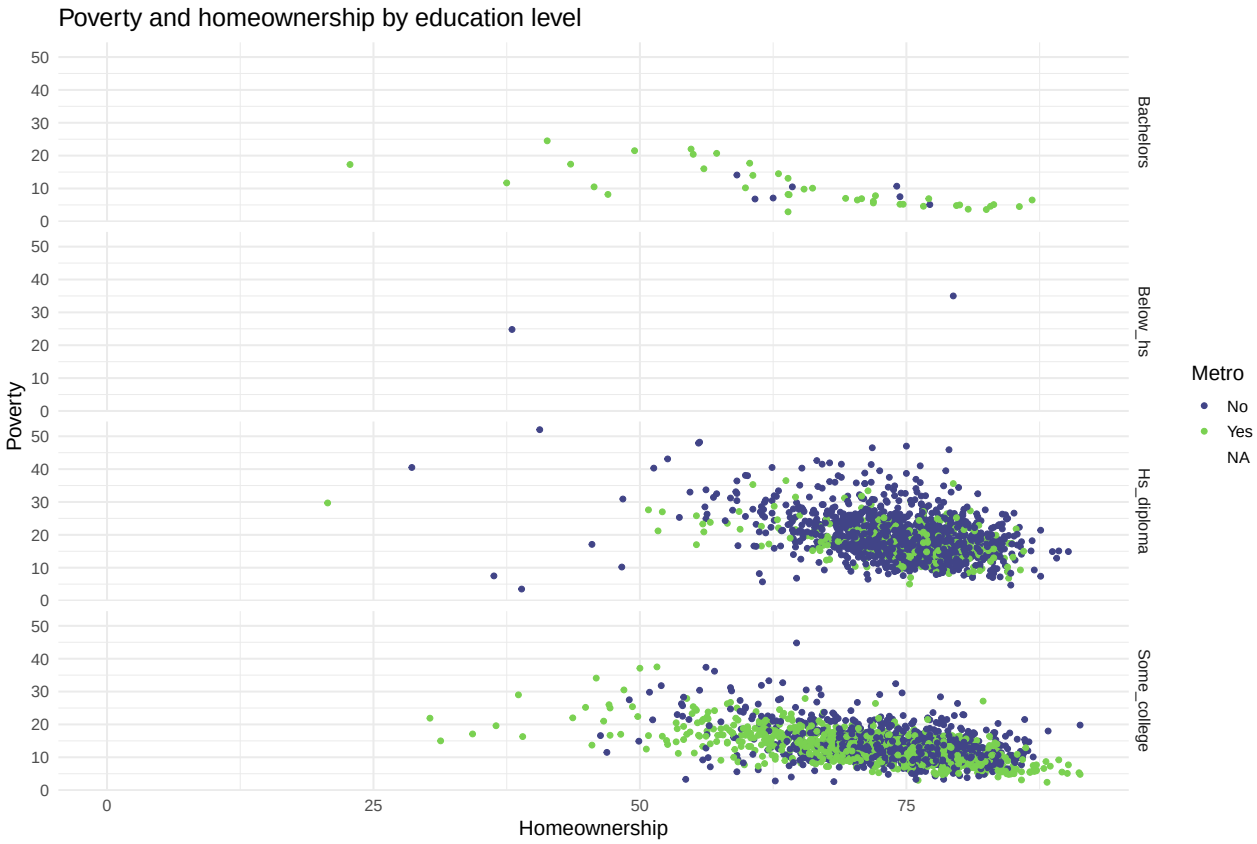
```
ggplot(county %>% filter(!is.na(median_edu))) +
  geom_point(aes(x = homeownership, y = poverty)) +
  facet_grid(median_edu ~ .)
```




```
ggplot(county %>% filter(!is.na(median_edu))) +
  geom_point(aes(x = homeownership, y = poverty)) +
  facet_grid(. ~ median_edu)
```

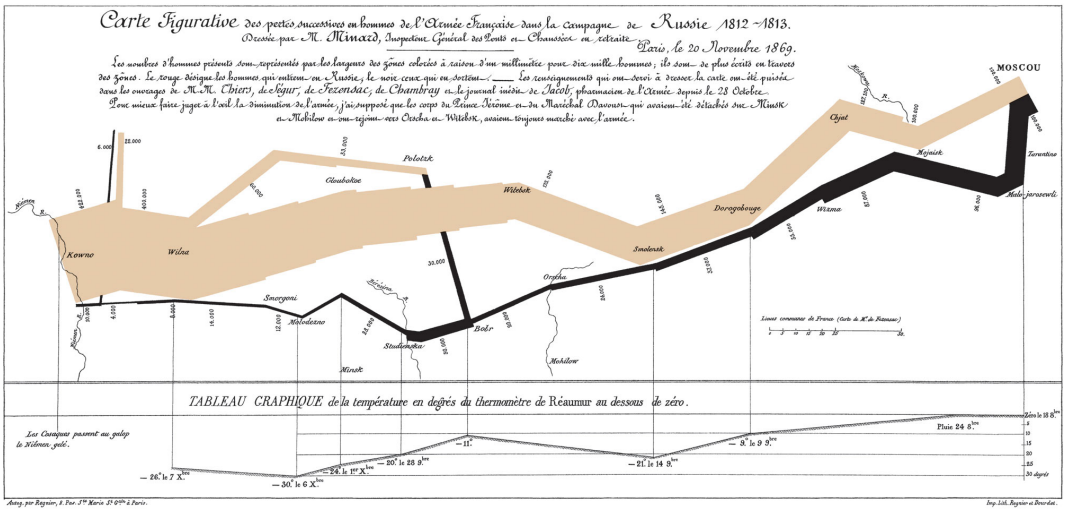


c. Recreate the R code necessary to generate the following graphs. Note that wherever a categorical variable is used in the plot, it's metro.



Problem 2: Napoleon's March

The instructions for this exercise are simple: recreate the Napoleon's march plot by **Charles John Minard** in `ggplot2`.



Edward Tufte has said that Minard's plot "may well be the best statistical graphic ever drawn" because it manages to pack a ton of information into one dense figure. The plot contains six variables, each mapped to a different aesthetic:

Information	Aesthetic
Size of Napoleon's Grande Armée	Width of path
Longitude of the army's position	x-axis
Latitude of the army's position	y-axis
Direction of the army's movement	Color of path
Date of points along retreat path	Text below plot
Temperature during the army's retreat	Line below plot

The data is provided as three separate text documents: cities, temperatures, and troops.

```
library(tidyverse)
library(lubridate)
library(ggmap)
library(ggrepel)
library(gridExtra)
library(pander)

cities <- read.table("../data/minard/cities.txt",
                     header = TRUE, stringsAsFactors = FALSE)

troops <- read.table("../data/minard/troops.txt",
                     header = TRUE, stringsAsFactors = FALSE)

temps <- read.table("../data/minard/temps.txt",
                    header = TRUE, stringsAsFactors = FALSE) %>%
  mutate(date = dmy(date))  # Convert string to actual date
```

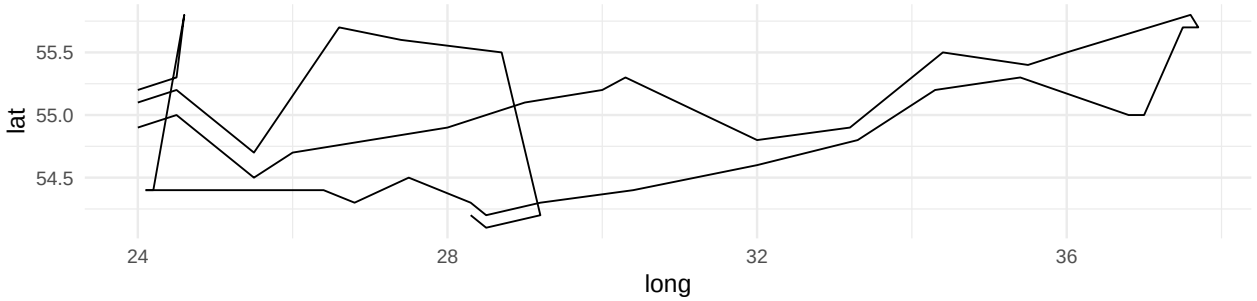
Getting Started The troops data includes five variables about troop movement: location, number of survivors, direction (advancing or retreating) and group (since Napoleon had generals commanding different elements of the army).

```
troops %>% head() %>% pandoc.table(style = "rmarkdown")
```

long	lat	survivors	direction	group
24	54.9	340000	A	1
24.5	55	340000	A	1
25.5	54.5	340000	A	1
26	54.7	320000	A	1
27	54.8	3e+05	A	1
28	54.9	280000	A	1

Each of these variables maps well into ggplot's aesthetic-based paradigm. If we include just geographic and group information (so there are separate lines for the different divisions), we get a basic skeleton of the original plot:

```r
ggplot(troops, aes(x = long, y = lat, group = group)) +
  geom_path()
```



Searching for "Napoleon's march in ggplot2" online will yield many resources to recreate the visualization using `ggplot2`. This activity involves using these online resources while citing them correctly. The goals are: (1) to navigate and reference web materials, (2) to understand and replicate `ggplot2` code, (3) to explain the code's function in your words, and (4) to add your personal touch to the final output. Here's what to focus on:

- Cite resources used, including those you read or get code snippets from. A list with links and a brief usage summary is sufficient.
- Explain your code's workings in your own words, rather than interpreting the visualization.
- Personalize your visualization, like altering colors or labels, to resemble the original more closely. Highlight what changes you made and why.