Homework 4: Data Import and Enhanced Graphics

Dr. D - Solutions

2026-01-05

```
library(dplyr)
library(ggplot2)
library(readxl)
```

Part I: Importing data

Go to Dr. D's data website to find the three data files used in the questions below. Download them to your `data` folder and import them in the first three questions below.

1. Read in the `High School and Beyond` data set. The High School and Beyond (HS&B) Longitudinal Study was the second study conducted as part of NCES' National Longitudinal Studies Program. This program was established to study the educational, vocational, and personal development of young people, beginning with their elementary or high school years and following them over time as they take on adult roles and responsibilities

Use the R function `names` to get a list of variable names in this data set.

```
hsb2 <- read.delim(here::here("data/hsb2.txt"), header = TRUE, sep = "\t")
names(hsb2)
```

```
 [1] "id"      "gender"  "race"    "ses"     "schtyp"  "prog"    "read"
 [8] "write"   "math"    "science" "socst"
```

2. Read in the `Crime Data` data set. This data set contains State and regional level information on crime and murder rates.

Use R functions to create a table that shows how many states are in each division (using the `division` variable).

```
crime <- read_excel(here::here("data/Crime_Data.xlsx"), sheet = 1, col_names = TRUE)
table(crime$division)
```

```
East North Central East South Central    Middle Atlantic           Mountain
                 5                   4                   3                  8
       New England            Pacific     South Atlantic West North Central
                 6                   5                   9                  7
West South Central
                 4
```

3. When reading in the Excel data set on crime, what do the arguments `sheet=1` and `col_names=TRUE` mean?

Read from the first sheet and the first row contains column names.

4. Read in the `countyComplete` data set. This data set contains data on characteristics of all counties in the United States.

Use an appropriate function to determine how many observations and variables are contained in the `countyComplete` data set. Write the answer in a complete sentence.

```
county <- read.csv(here::here("data/countyComplete.csv"), header = TRUE, sep = ",")
```

This data frame contains 3116 observations (rows) and 56 variables (columns).

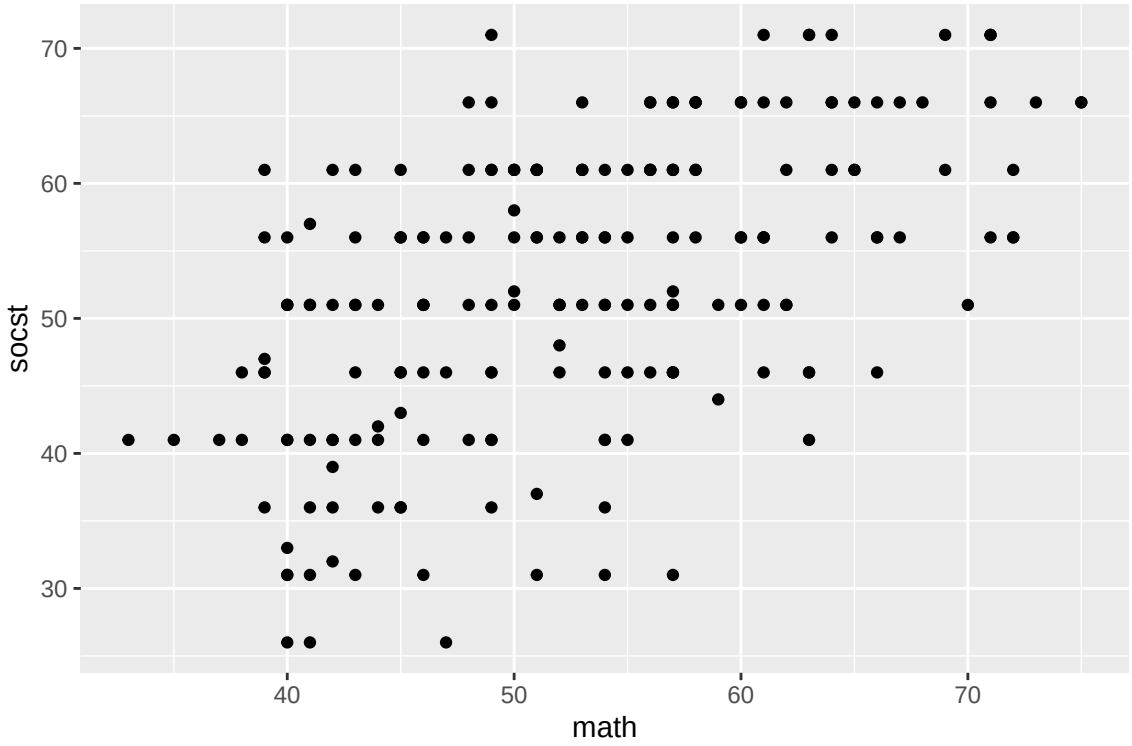
Part II: Improving your plots

This section demonstrates how you can build a nice graph in stages. Each question will have you adjust one thing on the plot. Instead of retyping the entire code each time, copy the code from the prior question and then make the requested adjustment. Small changes bit by bit helps you solidify learning what each piece of code does.

1. **Let's examine the relationship between math (`math`) and social studies (`socst`) scores using the HS & Beyond (`hsb2`) data set.**

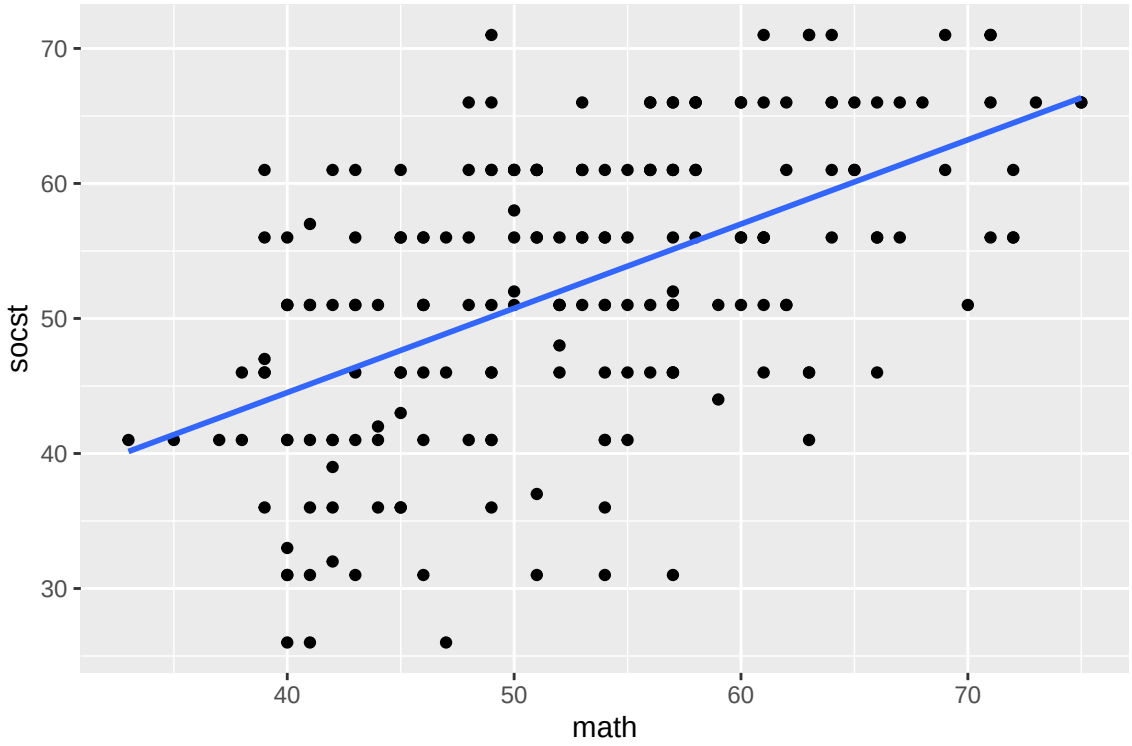
a. Start with a basic scatterplot. Create a scatterplot showing math score on the y-axis and social studies score on the x-axis.

```
ggplot(hsb2, aes(x = math, y = socst)) +
  geom_point()
```



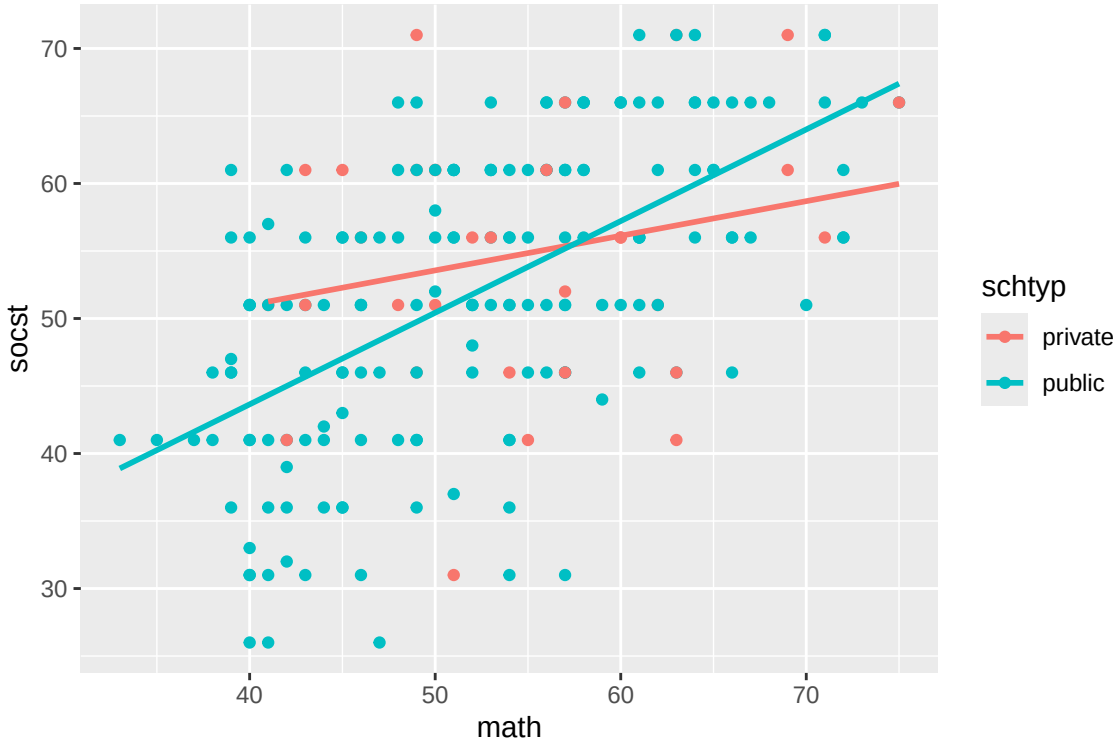
b. Add a trend line. Add a linear trend line to your plot to help clarify the overall relationship between math and social studies scores. Use a linear model line, but hide the confidence band.

```
ggplot(hsb2, aes(x = math, y = socst)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



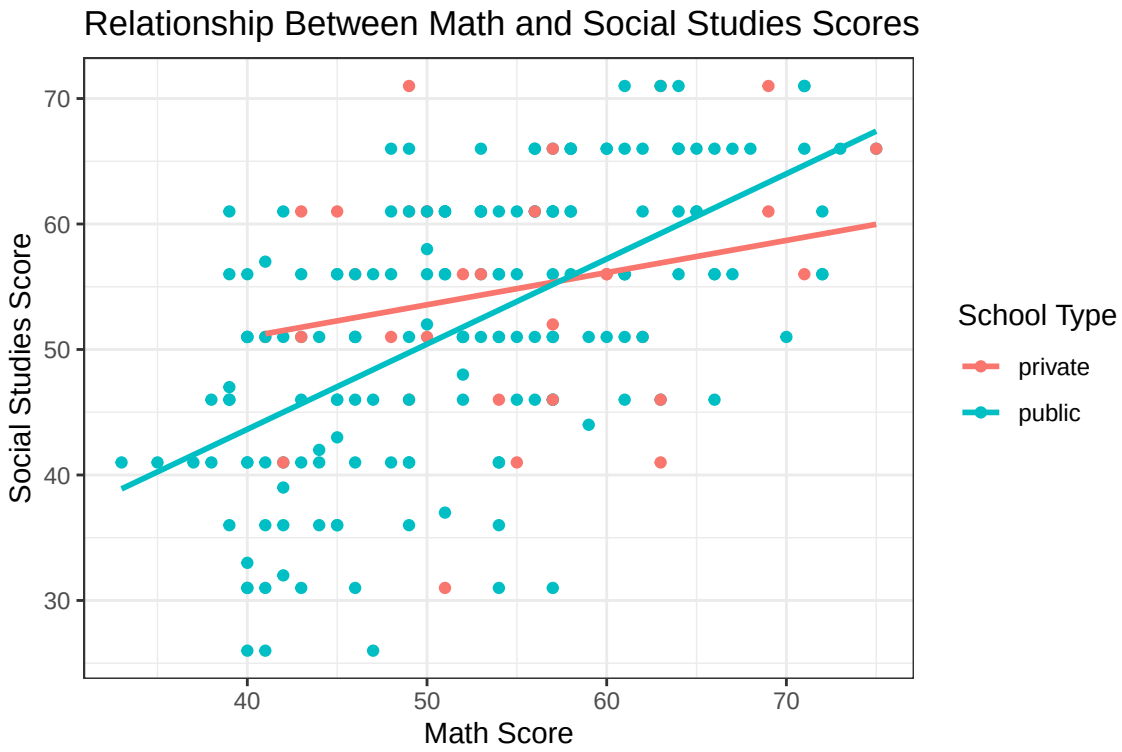
c. Group by school type. Check for a different relationship between math and social studies scores depending on school type. Color the points and lines by the `schtyp` variable.

```
ggplot(hsb2, aes(x = math, y = socst, color = schtyp)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



d. Improve communicability. Now refine the plot so that it clearly communicates results. Add a title, axis and legend labels, and apply a clean theme.

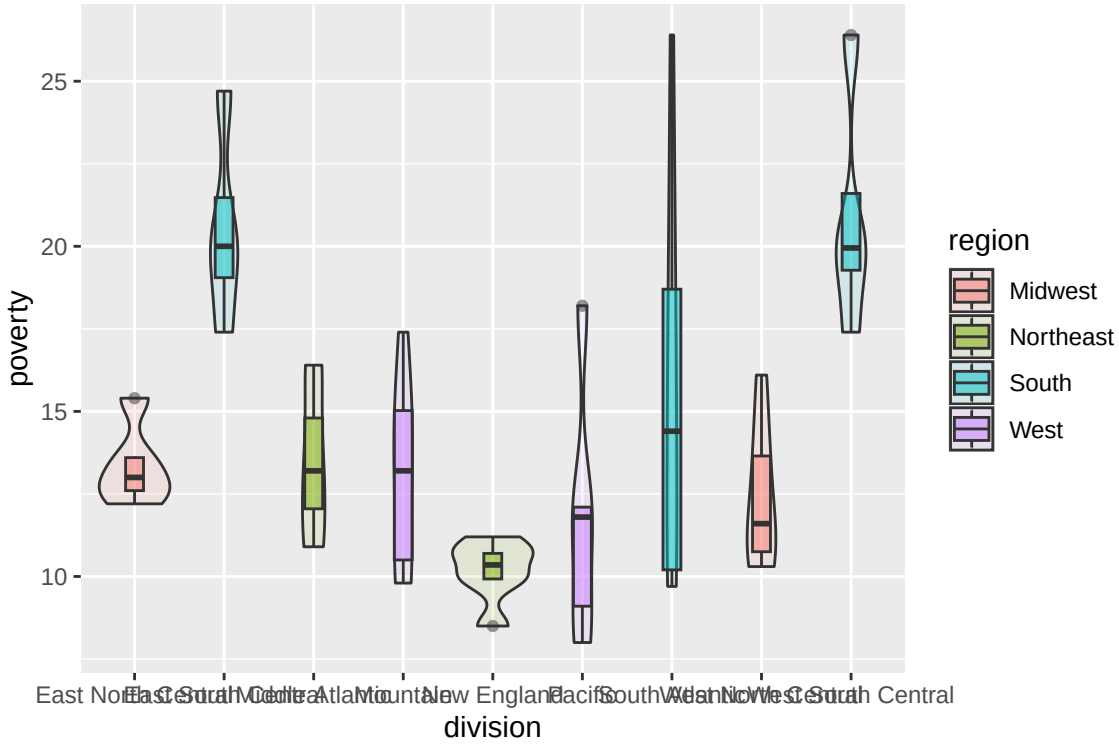
```
ggplot(hsb2, aes(x = math, y = socst, color = schtyp)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Relationship Between Math and Social Studies Scores",
    x = "Math Score", y = "Social Studies Score") +
  scale_color_discrete(name="School Type") +
  theme_bw()
```



2. Let's explore the distribution of poverty rates in states around the USA.

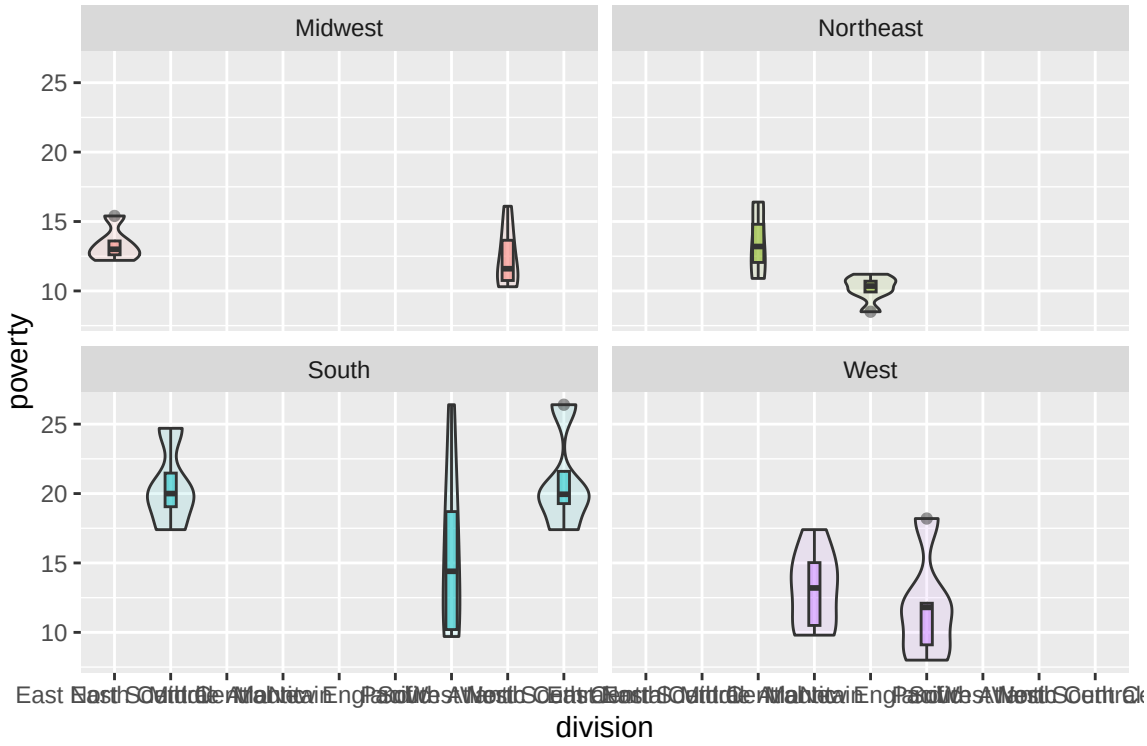
a. Using the `crime` data, create a boxplot with violin plot of the distribution of `poverty` rate on the y axis, `division` on the x axis, and filling the boxes by `region`. Adjust transparency of boxes and violins as necessary.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2)
```



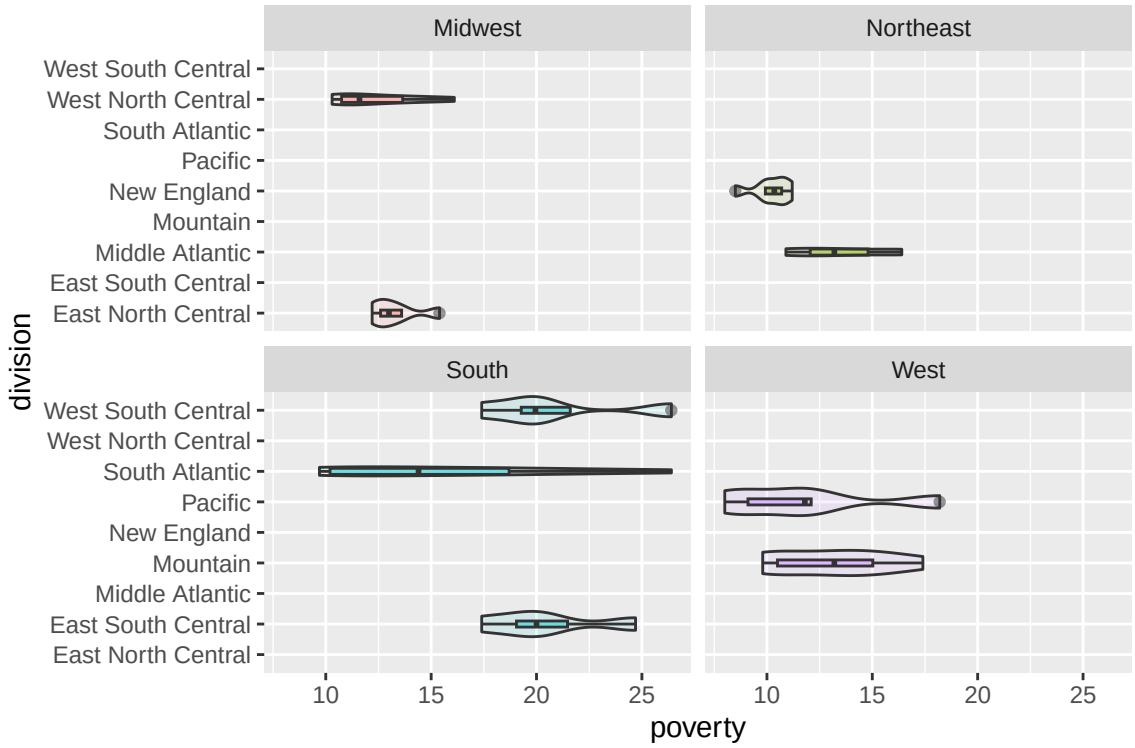
b. These are multiple divisions in a region, so let's facet on `region` so that each region is in it's own panel. Also get rid of the legend since the name of the region is in the panel already.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2) +
  facet_wrap(~region) +
  scale_fill_discrete(guide = "none")
```



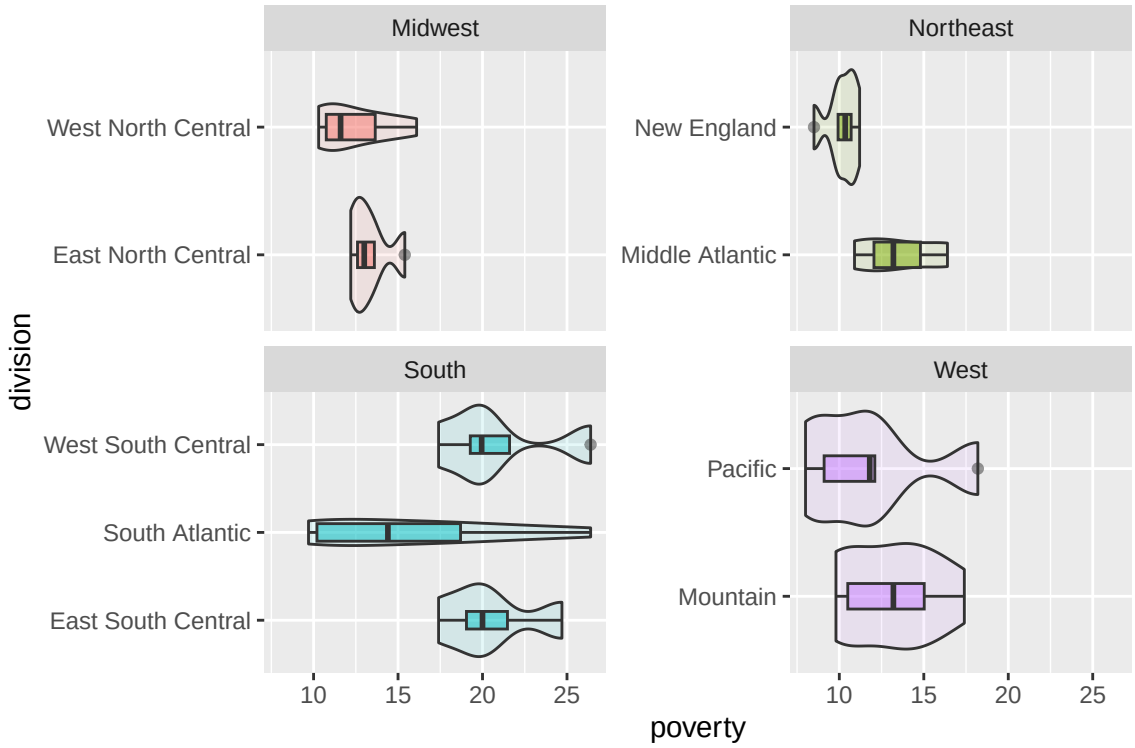
c. Improve readabilty. Let's try to fix the overlapping of the x axis labels using `coord_flip()` to rotate the plots horizontally.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2) +
  facet_wrap(~region) +
  scale_fill_discrete(guide = "none") +
  coord_flip()
```



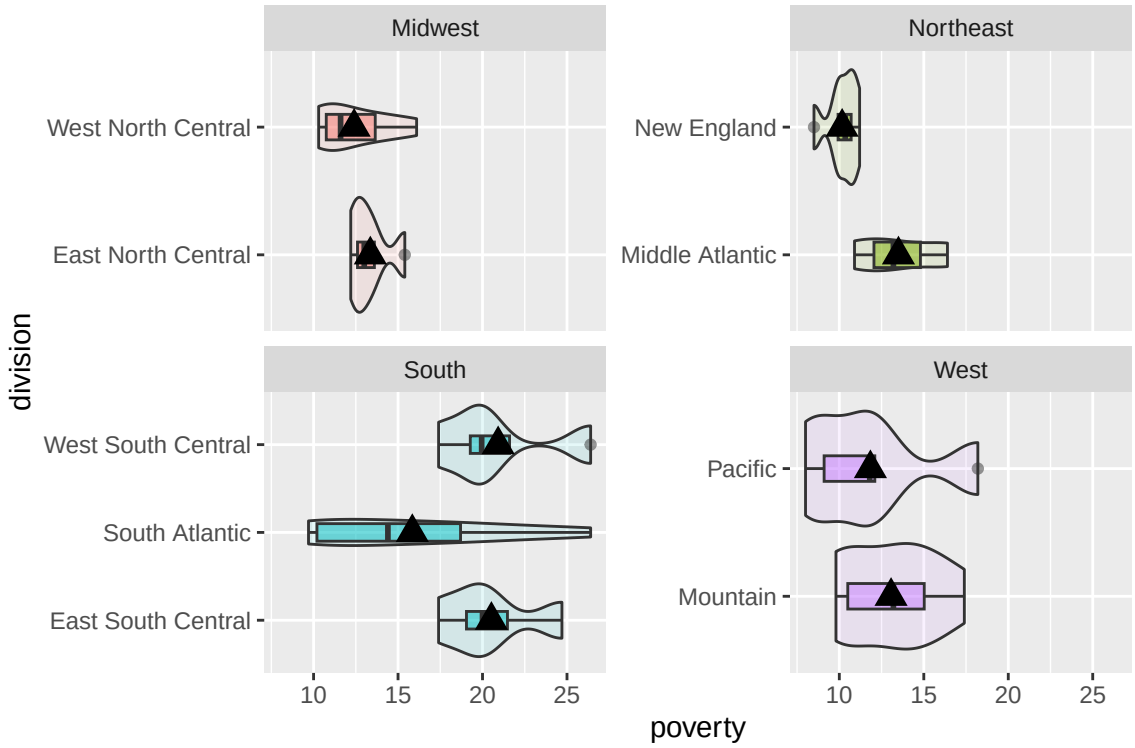
d. Simplify. Since division is nested within region we don't need ALL the levels of division to show up in each panel. Let's "free" our y axis labels by adding `scales="free_y"` to the `facet_wrap`.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2) +
  facet_wrap(~region, scales = "free_y")+
  scale_fill_discrete(guide = "none") +
  coord_flip()
```



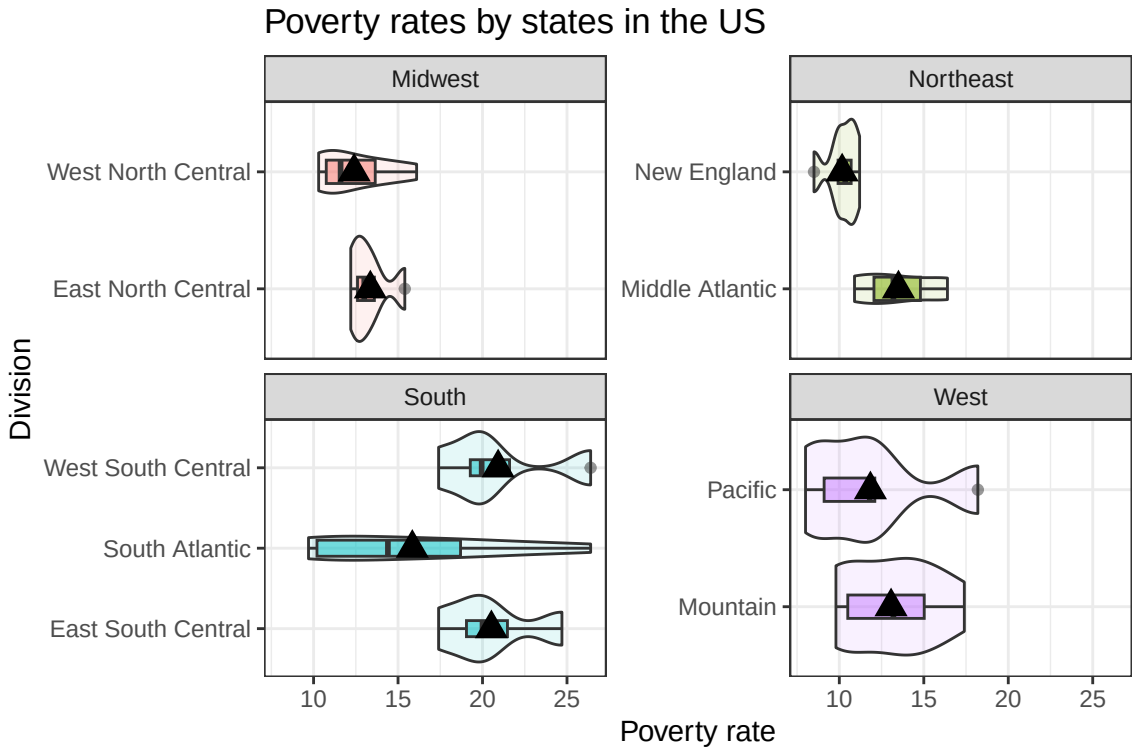
e. Add the mean to each group.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2) +
  facet_wrap(~region, scales = "free_y")+
  scale_fill_discrete(guide = "none") +
  coord_flip() +
  stat_summary(fun="mean", geom="point",
               shape=17, size=4)
```



f. Polish. Finish up by adding a plot title, axis labels and a clean theme. Double check your axis labels before you submit - coord_flip makes this backwards sometimes.

```
ggplot(crime, aes(y=poverty, x = division, fill = region)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.2) +
  facet_wrap(~region, scales = "free_y")+
  scale_fill_discrete(guide = "none") +
  coord_flip() +
  stat_summary(fun="mean", geom="point",
               shape=17, size=4)  +
  theme_bw() +
  labs(
    title = "Poverty rates by states in the US",
    y = "Poverty rate",
    x = "Division"
  )
```



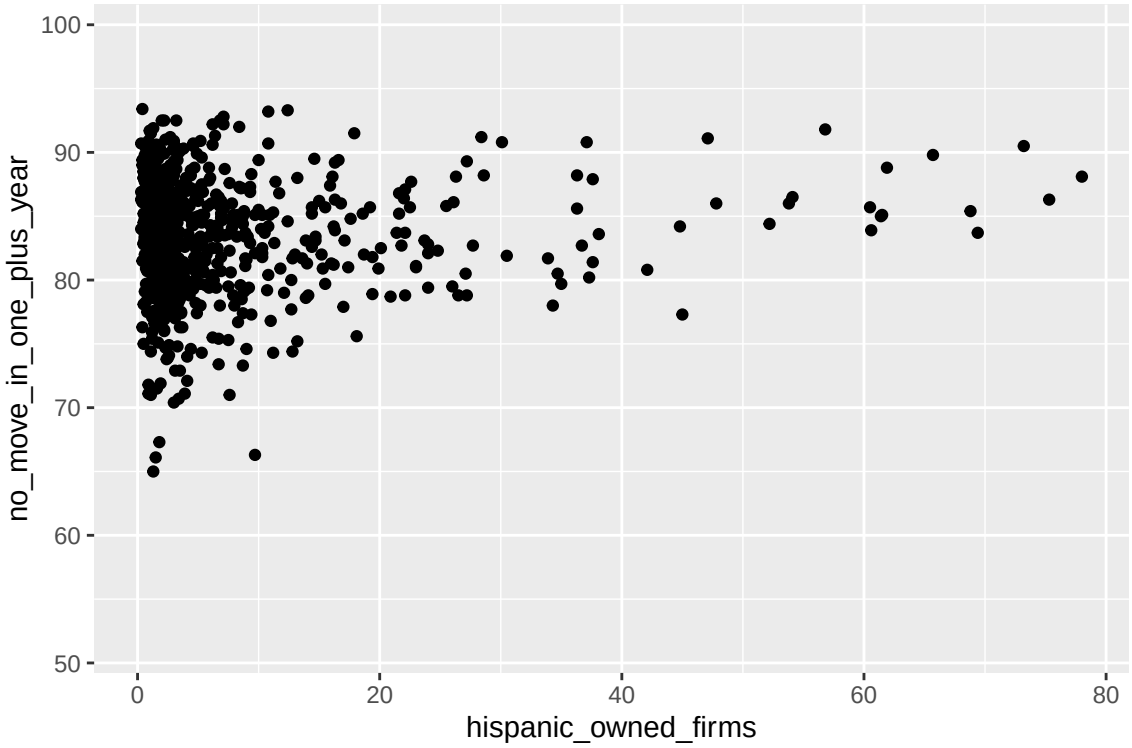
Part III: Plotting aggregated data

This section will ask you to create graphics on data that needs pre-processing using the dplyr verbs prior to plotting. Try to chain your commands together so that each question starts with a data set and ends with a plot.

3. Using the `countyComplete` data set, you will summarize business ownership characteristics and then visualize the results.

a. Use `dplyr` verbs to 1) Group the data by `state`, and then, 2) calculate the average percentage of Hispanic-owned firms (`hispanic_owned_firms`) within each state, and the standard error of this average using `se = number of counties used to calculate each state's average. (_hint: these can be done in the same` mutate `statement with the function` n()‘_)

```
ggplot(county, aes(x=hispanic_owned_firms, no_move_in_one_plus_year)) + geom_point()
```



Challenge Question

Using the `countyComplete` data set, create a scatterplot to investigate the relationship between the proportion of residents in a county earning a bachelors degree (`bachelors`) income (`per_capita_income`) for west coast states only (CA, OR, WA).

In your visualization, counties should be colored according to the percentage of residents who speak a foreign language at home (`foreign_spoken_at_home`). This variable should be binned into the following ranges before plotting: <10%, 10-20%, 20%+ (*hint: use* ***`case_when`***).

Present the results using separate panels for each state, apply a color scale that is appropriate for categorical groupings, and format the plot so that it is publication-ready with a clear title, labeled axes, and an informative legend.

```
library(viridisLite)
county%>%
  filter(state=="California" | state=="Oregon" | state=="Washington") %>%
  mutate(fb = case_when(
    foreign_spoken_at_home < 10 ~ "<10%",
    foreign_spoken_at_home >= 10 &  foreign_spoken_at_home < 20 ~ "[10-20%)",
```

```
  foreign_spoken_at_home >=20 ~ "20+%"
)) %>%
ggplot(aes(x=bachelors, y=per_capita_income, col=fb)) +
geom_point() + geom_smooth(method = "lm", se=FALSE) +
facet_wrap(~state) +
scale_color_discrete(name = "Percent speaking \n foreign language \n at home",
                     palette = plasma(3))   +
theme_bw()  +
xlab("% with a BS") + ylab("Per Capita Income") +
ggtitle("Per Capita Income by Bachelors Degree % for West Coast")
```