Homework 4: Data Import and Enhanced Graphics

YOUR NAME HERE

2026-01-05

```
# Load your libraries here
```

Part I: Importing data

1. Read in the `High School and Beyond` data set. Use the R function `names` to get a list of variable names in this data set.
2. Read in the `Crime Data` data set. Use R functions to create a table that shows how many states are in each division (using the `division` variable).
3. When reading in the Excel data set on crime, what do the arguments `sheet=1` and `col_names=TRUE` mean?
4. Read in the `countyComplete` data set. 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.

Part II: Improving your plots

1. 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.

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.

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.

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.

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.

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.

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

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`.

e. Add the mean to each group.

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.

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). Include a linear trend line.

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.