Lab 4 – What's the Explanation?

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please submit your lab report as a **pdf** to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.
- Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- In this lab, we're going to look at 2 different datasets stored in the mosaicData package on R.
- Each case will involve identifying unusual variation, or an unusual association, followed by some data investigation to make sense of what we observed!

STEP 0

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 4 first: https://stat212-learnr.stat.illinois.edu/
- **Install** the **mosaicData** package, and then **library** it as well (install the same way we installed tidyverse!)
- Next, open the two datasets we will use for this investigation: `View(Births2015)` and `View(SAT)`
- Remember to also `library(tidyverse)` so that you can use the ggplot function to visualize the data.
- Coding Tip: Remember that R is CaSe AnD sYmBoL_SeNsItIvE. Be aware of capitalized and non-capitalized letter when writing data names and variable names.

Part 1 – Explaining variation in number of births. Take a look at the `Births2015` data spreadsheet in the viewer window. Each row of this data represents one day in the Year 2015. The `births` variable represents the number of births recorded in the United States on that day.

Question 1 (5pts): Create a histogram of the births variable (using `ggplot2`) and also report the results from the `summary` function when summarizing that variable.

Include the image of your histogram

- Use a fill color of your choice
- Define a border color to better define the bins
- Add an appropriate title
- Adjusting number of bins is optional

Include the numeric summary output

Include your R code for this question

Briefly describe what you see. How would you describe the shape of this distribution? Is it what you expected, or is it difficult for you to explain?

Question 2 (5pts): Next, let's try creating a scatterplot with `births` on the y axis and `date` on the x axis. This will help us see if the time of year might explain some of the variation we see in number of births.

Include the image of your scatterplot

- Add an appropriate title

Include your R code for this question

Briefly describe what you see. What trends do you see between time of year and number of births? Does time of year explain most of the variability in births, or do you think there is still a lot of variability leftover?

Question 3 (5pts): Go back to the data viewer and notice what other variables we have in this dataset. Try to find a variable that explains why there are 2 distinct modes to the `births` variable.

Include an image of your visualization

- This visualization should have your "best predictor" on the x axis and the births variable on the y axis
- Add an appropriate title

Include your R code for this question

Briefly describe in context why this variable explains the bimodal variability in `births`. *Feel free to do some internet searching for some insight if you're not sure!*

Part 2 – Explaining variation in SAT Scores. Take a look at the `SAT` data spreadsheet in the viewer window. Each row of this data represents one state in the US. The variable `sat` represents the average SAT score of students in that state for the year. This data comes from the 1994-95 school year.

Question 4 (5pts): Create a histogram of the `sat` variable, and also report the results from the `summary` function when summarizing that variable.

Include the image of your histogram

- Use a fill color of your choice
- Define a border color to better define the bins
- Add an appropriate title
- Adjusting number of bins is optional

Include the numeric summary output

Include your R code for this question

Briefly describe what you see. Is this an expected amount of variability across state averages, or is this more or less variability than you expected?

Question 5 (5pts): Consider these three variables in the dataset:

- `expend`: Expenditure per pupil in thousands of dollars.
- `ratio`: average student/teacher ratio
- `salary`: estimated average annual teacher salary in thousands of dollars

Create a scatterplot with each of the three as the predictor variable (x axis) and `sat` listed as the response variable (y axis).

Include images of all 3 scatterplots

- Add an appropriate title for each

Include your R code for this question

Briefly describe what you see. What is the relationship between these variables? Do these variable relationships seem expected to you or surprising?

Question 6 (5pts): One variable we would like to investigate further as a possible predictor is `frac`, which represents the percentage of eligible students in the state taking the SAT. While we could leave it as a numeric variable for this stage, it will be a lot easier to use and see the results if we switch it to a binary variable!

First: look at the data viewer and sort the data by `frac` (click on the column header). We're looking for a noticeable gap where we can separate the High percentage states and Low percentage states.

At what percentage might you sensibly choose to use as a cut-off (i.e., where there is a large gap)?

Second: Create a new variable in the `SAT` data frame called `frac_bin` which will now label each state as "High" if above the cut-off value and "Low" if below the cut-off value. Use an `ifelse` function to create this new variable and be sure to assign it to `SAT$frac_bin`.

Include the ifelse code you used to complete this.

Question 7 (5pts): Now recreate the scatterplot with the `expend` variable on the x axis and `sat` on the y axis, but add a color aesthetic with `frac_bin` (if you were unsuccessful with Q6, just use `frac`) .

Include an image of your scatterplot

Include your R code for this question

Briefly describe what you see.

- Is there any association between the percent of eligible students in a state taking the SAT and the state's avg score?
- If only focusing on states with a high fraction of students taking the SAT or only focusing on states with a low fraction, is there an association between expenditure and SAT scores?
- Do you have any ideas as to why the fraction of students taking the SAT in a state might explain score differences?