Kelsey Matsik
April 23, 2024
Data Visualization Final Project

Objective

When creating each of my plots, I challenged myself to create effective and powerful visualizations that revealed insights about the data rather than just graphics that appeared aesthetically pleasing. In the past, I have gravitated towards just making plots look as "pretty" as possible, and by the end, the plots weren't revealing much information at all. To create meaningful visualizations, I started simple and left my plots in their default color and theme settings. Once I was happy with the plot, I added in color, grouping, themes, and other aesthetic elements.

I arrived at each of my final plots through trying first trying several different types of plots for the relationships between variables I wanted to capture. Visualizing my data proved to be more challenging than I originally thought, as many of my numeric variables had values 1-5 that corresponded to categorical values, such as "very poor" (1) to "very good" (5). My data also contained several variables that were very similar; for example, mother's education and father's education. When building visualizations for these types of variables, I created interactive plots (Shiny Apps) where the user could toggle between the two similar variables and the plot would react accordingly. I also did a fair amount of data cleaning beforehand with the intention of creating meaningful visualizations in mind.

Accessibility

Another consideration when creating my plots was accessibility. To make sure my plots were accessible, I used the plasma color palette from the viridis package and adjusted my axes, title, and text labels to be large enough to be read from far away or across different devices. I chose to use the plasma palette because it is color-blind friendly, and its colors do a good job of clearly distinguishing groups in variables.

Aesthetic Choices and Storytelling Organization

I chose to organize my project into three sections while exploring how different variables affected high school students' health and academic success. The focus points of the three sections were alcohol consumption, academic grades, and personal relationships. My first plot is a scatterplot with a regression line. I chose to include the linear regression line with confidence bands to guide the viewer to the relationship between age and alcohol consumption. I also added the correlation coefficient to the top right of the plot to accomplish this same objective. My second plot is just a facet grid of my first plot grouped by gender. I chose to show the facet grid of this plot because I thought it was interesting that there was a

strong association between age and alcohol consumption for males, but not so much for females.

In all three of my Shiny Apps, the user can select from a dropdown menu to change one variable. I chose to start with a default graph in each of my plots rather than including a 'Generate' button to make plots because I prefer the user to immediately see a visualization upon opening the app. My third plot is a Shiny App featuring a heatmap showing the relationship between health status and alcohol consumption level. I included three forms of user input: a slider to change the bin width, a drop down to change the "high" fill color of the gradient, and a checkbox to group by sex (facet grids). The user has the option to change bin width from 0.5 to 2. I chose this range because it allows the user to see the data at a more granular level while still being visually appealing and relevant. The fill color input was mostly for fun, but it provides the user with some more optionality when viewing high and low values by hue. The dropdown colors are all similar colors to the plasma palette to be consistent with my project theme. I included a checkbox to group by sex to let the user draw comparisons in the data, with the default being set to all the data.

My fourth plot is a Shiny App featuring a ridgeline plot of students' final grades and their alcohol consumption levels. The app includes a dropdown menu to switch the view from weekday alcohol consumption from weekend alcohol consumption. I chose to switch between these two variables because I thought it would be insightful to see how much students drink on the weekends versus the weekdays and how that affects their academic performance. My fifth plot is an interactive (plotly) scatter plot that shows the relationship between absences and final grades, grouped by the total hours a student spent studying per week. The interactive element of my plot is hover text; the user can hover over any data point, representative of one student, and see how many hours they spent studying per week. Because of the hover element, I was not going to include a legend with the same information, but peer review suggested that this would be helpful in understanding the data, so I did. My sixth plot is another plotly object showing how a student's health status relates to their academic performance throughout the year. The plot animates over grades from three points in time: first semester grade, second semester grade, and final grade. Rather than assigning the plasma color palette to the five distributions, I chose to highlight the grade distributions of students reporting very good. I did this because the difference in the distributions would be difficult to see otherwise, and I wanted to guide the viewer to the comparison I was trying to make. I highlighted only the distribution students reporting very good health because this was the most drastic difference compared to other students in terms of grades. My seventh plot is a facet grid boxplot visualizing the relationship between free time and final grade, grouped by whether or not students participated in extracurricular activities. I chose to group by extracurriculars/no extracurriculars because I thought it would be interesting to see whether or not students spend their free time doing school activities and if those students tend to earn higher grades.

My eighth plot is a Shiny App featuring a bar graph displaying counts of students interested in attending college based on their parent's highest level of education. The user can

select the dropdown menu to view by father's education level instead of mother's education label. I gave the user the option to toggle between these two variables because they are very similar, and both of them together provide a picture of parents' education levels. For the students who answered "Yes" to wanting to go to college, I chose blue, and for those who answered "No," I chose gray, intentionally making the counts of students who *did* want to attend college pop. This draws the viewer to the most important comparison I wanted to make. I also chose to display the counts of each bar directly on the plot so the user didn't have to look back and forth between the bars and axis, minimizing visual load. My ninth and final plot is an alluvial plot showing the relationship between satisfaction with family relationships and having a romantic partner. I chose an alluvial plot to display these variables because I wanted a way to visualize the relationship between two categorical variables. I also played with the idea of including another factor variable, number of family members, but this plot wasn't very insightful and risked overcrowding so I decided against it.

I chose the visuals in my final project very intentionally. Before each final plot, there were iterations of different plots with the same variables, and then the same plot with different aesthetic themes. I included all of these plots in the Appendix of my code submission.

Sources

- Chat GPT
    - *Below is a list of ChatGPT prompts I used to help clean data and create plots for my project.
    - View duplicate rows
    - I have multiple columns I want to change in R in the same way. I want to change each 'yes' value to 1, and each 'no' to 0. Show me how to do this to multiple columns at the same time.
    - Error in train_continuous(x, self$range) :
    - df <- df %>% mutate(drinking = mean(weekday.alc, weekend.alc))
    - What are some plots I can make in rstudio that visualize the relationship between a factor variable and a binary variable?
    - what are some more creative ones?
    - replace values in the study time df equal to 1 with <2
    - show me how to replace values 2 3 and 4 as well in one statement
    - apply the viridis color plot to my plot
    - change the transparency of my points
    - R is not doing anything with my case_when() statements. Why?
    - Change the scale of my y-axis in my plotly plot
    - increment by 10 not 20 (the default)
    - change the size of the title, x and y axis, axis labels, and animation frame slider
    - Add text labels to my bars in geom_bars
    - No, I want the labels to represent the values, "yes" and "no" of the factor(higher.edu) variable.
    - I want to move the labels to the right so they are not overlapping the bars
    - remove the legend
    - tilt x-axis labels
    - change facet labels
    - change size of facet labels
    - create a frequency variable for my data. My variables are romantic (binary), health.status (factor), and famrel (binary). My goal is to make an alluvial plot using ggalluvial with frequency as my y variable
    - change the values of each variable in the alluvial plot
    - change the size of the annotation
    - change the text labels to counts of each bar rather than the values "Yes" and "No"
    - make the legend title and values bigger
- https://stackoverflow.com/questions/40001518/add-title-to-the-plotly-legend
- https://tidyr.tidyverse.org/reference/gather.html

- https://ggplot2.tidyverse.org/reference/geom_text.html
- https://rpkgs.datanovia.com/ggpubr/reference/stat_cor.html
- https://stackoverflow.com/questions/70494677/how-do-i-display-a-correlation-coefficient-in-a-scatterplot