Scatterplot 1: Basic Plotting and Data Analysis

Code:

```
perchsd <-
        ggplot(midwest, aes(x=poptotal, y=perchsd)) #init ggplot with data +
        geom_point() + #add scatterplot points
        geom_smooth(method="lm", se=FALSE) + #add linear regression
        coord_cartesian(xlim=c(55, 90), ylim=c(0, 1500000)) + #zoom in scatterplot
        labs(title="Population vs Percent High School Degree", subtitle="From midwest
        dataset", x = "Population", y="Percent High School Degree", caption="Midwest
        Demographics") #add labels to plot
percollege <-
        ggplot(midwest, aes(x=poptotal, y=percollege)) +
        geom_point(color = "purple") +
        geom_smooth(method="lm", se=FALSE, color = "black") +
        coord_cartesian(xlim=c(0, 1500000), ylim=c(0, 50)) +
        labs(title="Population vs Percent College", subtitle="From midwest dataset", x =
        "Population", y="Percent College", caption="Midwest Demographics")

percwhite <-
        ggplot(midwest, aes(x=poptotal, y=percwhite)) +
        geom_point(color =  "aquamarine4 ") +
        geom_smooth(method="lm", se=FALSE, color =  "darkorchid4") +
        coord_cartesian(xlim=c(0, 1500000), ylim=c(60, 100)) +
        labs(title="Population vs Percent White", subtitle="From midwest dataset", x =
        "Population", y="Percent White", caption="Midwest Demographics")

percasian <-
        ggplot(midwest, aes(x=poptotal, y=percasian)) +
        geom_point(color="cadetblue3") +
        geom_smooth(method="lm", se=FALSE, color="cadetblue4") +
        coord_cartesian(xlim=c(0, 1500000), ylim=c(0, 3)) +
        labs(title="Population vs Percent Asian", subtitle="From midwest dataset", x =
        "Population", y="Percent Asian", caption="Midwest Demographics")
```

7. In writing, reflect on the visualizations you've just made.

a. What do you notice about the relationship between population and education level?

It seems like there is no large correlation between population and high school degree. The line of best fit in the visualization slightly increases, however, the scatterplot shows a large variation in percentage despite similar population levels. College on the other hand has a steeper increase for the line of best fit, and there seems to be a slightly more noticeable trend in percent increasing as the population increases. Despite this, I would still say there is not a large correlation between population and college either, as many points on the scatter plot do not seem to correspond with the line.

b. What do you notice about the relationship between population and percent white? What about the relationship between population and percent Asian? How do the two compare? What do you think accounts for any differences you perceive?

It seems like the relationship between population and percent white is that is the population increases, the percent white decreases. With percent Asian on the other hand, as the population increases, the percent Asian increases as well. In both scatterplots, the points tend to loosely follow the line of best fit. It is important to note however, that while percent Asian increases, it only increases by almost 2%. Keeping all of this in mind, it would make sense that in order for one to increase the other has to decrease, since the percentages of each race in relation to the population is dependent on the percentages of all other races in relation to the population.

<u>Scatterplot 2: Color, Theme, and Reflecting on Aesthetic Choices</u>

Code:

```
percollege_state <-
    percollege +  #use percollege data
    geom_point(aes(col=state)) + #color points based on state
    scale_color_brewer(palette="Set1") + #set color palette
    theme_classic() #set theme
```

2. What do you notice about the results? What do you think accounts for any differences you perceive?

It seems as though each state has differing patterns since some do not have much increase in population while others do. Generally, when looking at things state by state, if the population increases then the percent college will also slightly increase, but if there is not much increase in the population in the state then there is no correlation between percent college and population. For the states that do not see an increase in population, there is still variation in percent college, which could be attributed to other factors that are not dependent on population levels, like socioeconomic status or value of higher education for people in that area.

3. Which color palette did you choose to use, and why?

I chose the color palette “Set1”, because after looking at the color palettes, this palette seemed to have the best variation in colors where each color was noticeably different than the next. This makes it easier to differentiate each state from each other when looking at the clustering points on the scatterplot as they starkly contrast each other.

4. Explore some possible themes you could use. Which theme did you ultimately settle on, and why?

I chose the classic theme. I wanted the focus of the plot to be easily identifying trends in the points as opposed to looking into the specifics of the data and numbers, so to me the gridded lines on the graph were not necessary. I liked how the classic theme did not add extra colors or lines on the scatterplot and simply produces a white background, making the plot simpler and easier to look at and causes our eyes to focus mainly on the colored points and the patterns between them.

Scatterplot 3: Telling your own data story with the 'midwest' data

Code:

```
table(midwest$state) #WI least, IL most

library(tidyverse)

IL <- midwest %>% filter(state == "IL")

WI <- midwest %>% filter(state == "WI")

#create new data frame with filtered data
IL_WI_dem <- data.frame(State = c("IL", "IL", "IL", "IL", "IL", "WI", "WI", "WI", "WI", "WI"), Race = c("White", "Black", "Amerindian", "Asian", "Other", "White", "Black", "Amerindian", "Asian", "Other"), Percent = c((sum(IL$popwhite) / sum(IL$poptotal) * 100), (sum(IL$popblack) / sum(IL$poptotal) * 100), (sum(IL$popamerindian) / sum(IL$poptotal) * 100), (sum(IL$popasian) / sum(IL$poptotal) * 100), (sum(IL$popother) / sum(IL$poptotal) * 100), (sum(WI$popwhite) / sum(WI$poptotal) * 100), (sum(WI$popblack) / sum(WI$poptotal) * 100), (sum(WI$popamerindian) / sum(WI$poptotal) * 100), (sum(WI$popasian) / sum(WI$poptotal) * 100), (sum(WI$popother) / sum(WI$poptotal) * 100)))

IL_WI_plot <- ggplot(IL_WI_dem, aes(x=State, y=Percent, fill=Race)) + geom_bar(stat="identity", width = 0.5) + theme_bw() + scale_fill_brewer(palette="Set3") + labs(title="Illinois and Wisconsin Racial Demographics", subtitle="From midwest dataset")
```

2. In writing, briefly describe what you think is the significance of your data visualization. What story is the data telling through your graphic? What are the advantages and disadvantages of your visualization?

For my data visualization, I took the states with the most and least entries in the Midwest data set, calculated the percentage of each race out of the state's entire population, and put it on a bar graph by state. This graphic depicts the difference in racial demographics for Illinois and Wisconsin so that the percentage of each race and diversity of both states can be easily compared. The advantages are that the difference in percentage by race per state and between the two states is easy to see, which states we are talking about is clear, and the color associated with each race is clear. Disadvantages are that the racial demographics between only two states that are in the same region of the U.S. does not tell us much, the exact numerical percentage of each race is difficult to decipher, and the way race is separated could be more comprehensive (meaning that I think a few races have been squished into "other" that are usually on their own… but this is likely due to them being such a small fraction of the Midwest's demographics).

Scatterplot #4

Code:

```
plot_troops <- ggplot() +
  geom_path(data = troops, aes(x = long, y = lat, color = direction, size = survivors, group = group), lineend = "round") +
  geom_point(data = cities, aes(x = long, y = lat), color = "brown4", size = 1) +
  geom_text(data = cities, aes(x = long, y = lat, label = city), vjust = 1.25, color = "brown4", family = "Helvetica") +
  scale_color_manual(labels = c("advance", "retreat"), values = c("burlywood3", "darkseagreen4")) +
  scale_size(range = c(0.5, 15)) +
  labs(title = "Napoleon's 1812 Russian Campaign", subtitle = "A Variation of Minard's Plot", color = "movement (line color)", size = "# of survivors (line size)", x = "Longitude of Position", y = "Latitude of Position") + theme_classic()
```

2. In writing, briefly describe the choices you made when making your plot. What did you choose to represent and not represent? What were you happiest about being able to visualize? Was there anything you wanted to show but couldn't figure out how to?

When first looking at the plot, it was difficult for me to understand what I was looking at and I couldn't interpret what anything meant. It wasn't until I thoroughly read the example plot that I understood the information being communicated. I wanted my plot to be able to be understood immediately without needing to read something separate. In order to do this, I added explicit titles to the x and y axis, the legends, and overall title. I also made sure that no colors were repeated and that the colors contrasted for accessibility, so that it's clear which visuals and legends correspond to each other, and to differentiate visuals that may overlap with each other. In addition, I adjusted the range in line size so that the differences in size could be more drastic and easier to understand. I chose to omit information on the temperature, mostly because while I think it is interesting, I don't think it contributes and is as relevant as all the other information and makes the plot too busy. Everything I wanted to show I was able to figure out, especially since a lot of it was what I put into practice for all the previous plots.

```
ggsave(percasian, file = "plot1.pdf", width=5, height=3)
ggsave(perchsd, file = "plot2.pdf", width=5, height=3)
ggsave(percollege, file = "plot3.pdf", width=5, height=3)
ggsave(percwhite, file = "plot4.pdf", width=5, height=3)
ggsave(percollege_state, file = "plot5.pdf", width=5, height=3)
ggsave(IL_WI_plot, file = "plot6.pdf", width=5, height=3)
ggsave(plot_troops, file = "plot7.pdf", width=10, height=5)
```