Coding Assignment 8: Data Visualizaton in R

Practice with Data Viz in R

Overview: For this assignment, you will practice working creating data visualizations with R.

Directions: Using one of the data files of your choice, create **two** data visualizations. At least one of the data visualizations should include at least two variables in your data. You may use any of the types of data visualizations we reviewed in class on 3/20, making modifications for a different analysis (different variables) or you can build a different kind of visualization. You may use online galleries for inspiration and for code templates to edit. But you still should not be using generative AI (e.g., ChatGPT, Bing, etc.). I encourage you to use `ggplot2` but you may also experiment with plotting functions in Base R. Experiment with different themes, colors, etc.

Data sets to choose from (same as before or you may pick a new dataset, as long as it is **not** what you are using for your project): - SleepStudy.csv (used in class previously) - colleges.csv (used in class previously) - billionaires.csv (new - available on github) - parking_citations.csv (new - availabe on githib)

Read in the data and load the required libraries. [2 POINTS]

Load required libraries

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

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Read in the data

```
# billionaires <- read.csv("/Users/samakheir/DSC201/DSC201/data/billionaires.csv")
billionaires <- read.csv("data/billionaires.csv") # edited by Kristin
str(billionaires)
```

```
## 'data.frame':    2614 obs. of  22 variables:
##  $ name                 : chr  "Bill Gates" "Bill Gates" "Bill Gates" "Warren Buffett" ...
##  $ rank                 : int  1 1 1 2 2 2 3 3 3 4 ...
##  $ year                 : int  1996 2001 2014 1996 2001 2014 1996 2001 2014 1996 ...
##  $ company.founded      : int  1975 1975 1975 1962 1962 1990 1896 1975 1975 1976 ...
##  $ company.name         : chr  "Microsoft" "Microsoft" "Microsoft" "Berkshire Hathaway" ...
##  $ company.relationship : chr  "founder" "founder" "founder" "founder" ...
```

```
## $ company.sector        : chr   " Software" " Software" " Software" " Finance" ...
## $ company.type          : chr   "new" "new" "new" "new" ...
## $ demographics.age       : int   40 45 58 65 70 74 0 48 77 68 ...
## $ demographics.gender    : chr   "male" "male" "male" "male" ...
## $ location.citizenship   : chr   "United States" "United States" "United States" "United States" ..
## $ location.country.code  : chr   "USA" "USA" "USA" "USA" ...
## $ location.gdp          : num   8.10e+12 1.06e+13 0.00 8.10e+12 1.06e+13 ...
## $ location.region        : chr   "North America" "North America" "North America" "North America" ..
## $ wealth.type           : chr   "founder non-finance" "founder non-finance" "founder non-finance"
## $ wealth.worth.in.billions: num   18.5 58.7 76 15 32.3 72 13.1 30.4 64 12.7 ...
## $ wealth.how.category    : chr   "New Sectors" "New Sectors" "New Sectors" "Traded Sectors" ...
## $ wealth.how.from.emerging: chr   "True" "True" "True" "True" ...
## $ wealth.how.industry    : chr   "Technology-Computer" "Technology-Computer" "Technology-Computer"
## $ wealth.how.inherited   : chr   "not inherited" "not inherited" "not inherited" "not inherited" ..
## $ wealth.how.was.founder : chr   "True" "True" "True" "True" ...
## $ wealth.how.was.political: chr   "True" "True" "True" "True" ...
```

For each data visualization:

(1) Make your plot easy to understand. [3 POINTS FOR EACH DATA VIZ]

- **Provide a good title - include information about your sample or population (who is covered by your dataset)**
- Label your axes.
- Ask yourself if a legend is needed or if it works (can you map plot info to the legend easily?).
- Don't rely on variable names if they are not sufficiently clear.

(2) Discuss the success of your plot. [3 POINTS FOR EACH DATA VIZ FOR COMPLETION]

- What works well for telling a data story?
- What design choices did you make and why?
- What might make it better but was too challenging or you were unsure how to proceed?

(3) Describe the results. [3 POINTS FOR EACH DATA VIZ FOR COMPLETION]

- How do you interpret the plot? What do the data findings tell you?
- Why is this particular plot choice helpful for what you want to learn?
- What are possible limitations?

Data Visualization #1: Age vs. Worth in Billions Among Billionaires

Step 1

```
plot1 <- ggplot(data = billionaires, aes(x = demographics.age, y = wealth.worth.in.billions)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Age vs. Worth in Billions Among Billionaires",
       x = "Age (Years)",
       y = "Net Worth (Billions USD)") +
  theme_minimal()
print(plot1)

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Age vs. Worth in Billions Among Billionaires
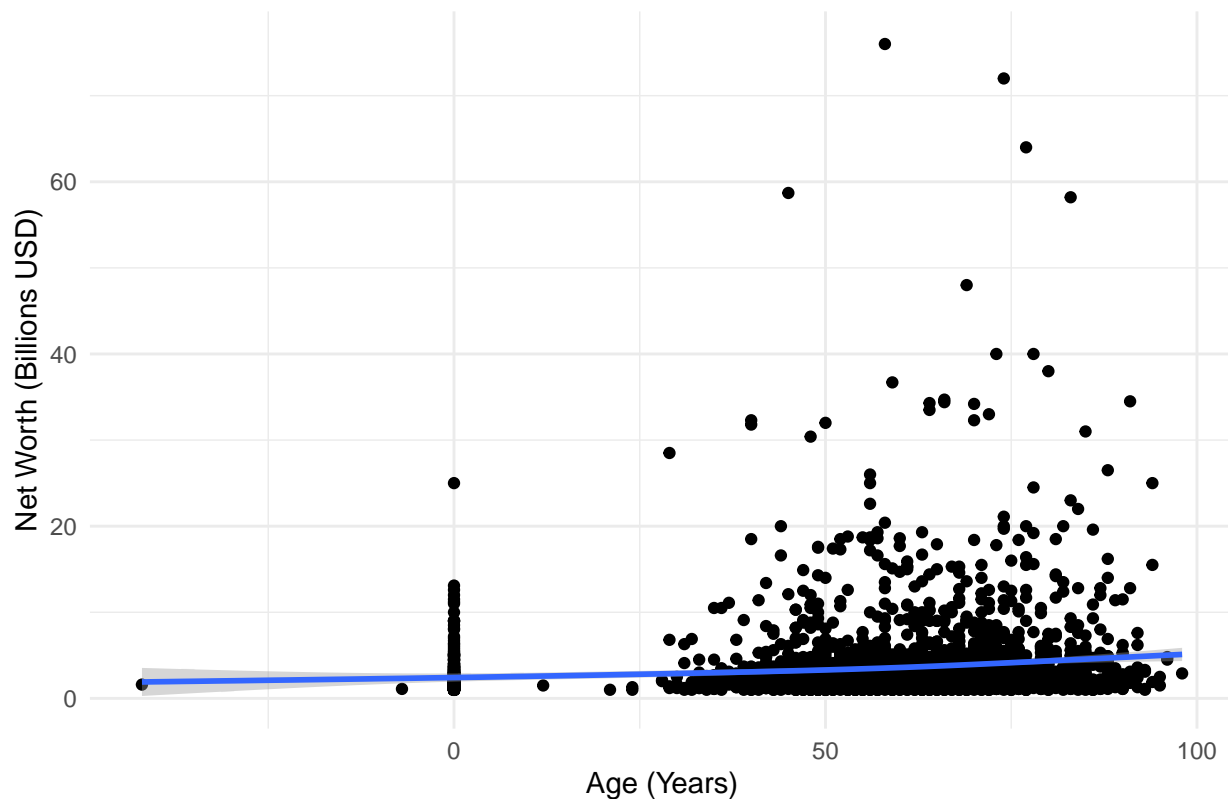
Step 2 Discuss the success of your plot:

The scatter plot is successful in that it presents the relationship between age and net worth among billionaires in this data set. Adding in the smoothed line helps visualize the trend that more older individuals are billionaires than younger.

I chose to use a scatter plot because it's suitable for showing individual data points and their distribution. It is really easy to see where data condenses versus not. The smoothed line also helps identify any underlying trends.

Thinking statistically, maybe adding confidence intervals to the smoothed line could provide more information about the uncertainty of the trend. I am not really sure how to do that but I am sure with a quick google search I could.

Step 3

The plot shows a general trend that older billionaires tend to have higher net worth. However, there's still a good amount of younger billionaires listed in the plot. It makes sense that older people are more likely to be billionaires because they have had more time to work, inherit money, and gain financial experience. Additionally, there are plenty of born billionaires which was cool and amusing to see right away.

I thought this plot choice was helpful because it allows me to visually explore the relationship between age and net worth among billionaires. It is one of the best plots to see if there's any correlation between two variables.

A limitation I noticed was that the plot doesn't account for other factors that could influence net worth, such as the industry or source of wealth. Additionally, outliers or influential data points might skew the trend observed in the plot, like I mentioned with the born billionaires.

Data Visualization #2: Distribution of Wealth Types Among Billionaires

Step 1

```r
plot2 <- ggplot(billionaires, aes(x = wealth.type)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Wealth Types Among Billionaires",
       x = "Wealth Type",
       y = "Number of Billionaires") +
  theme_minimal()
print(plot2)
```
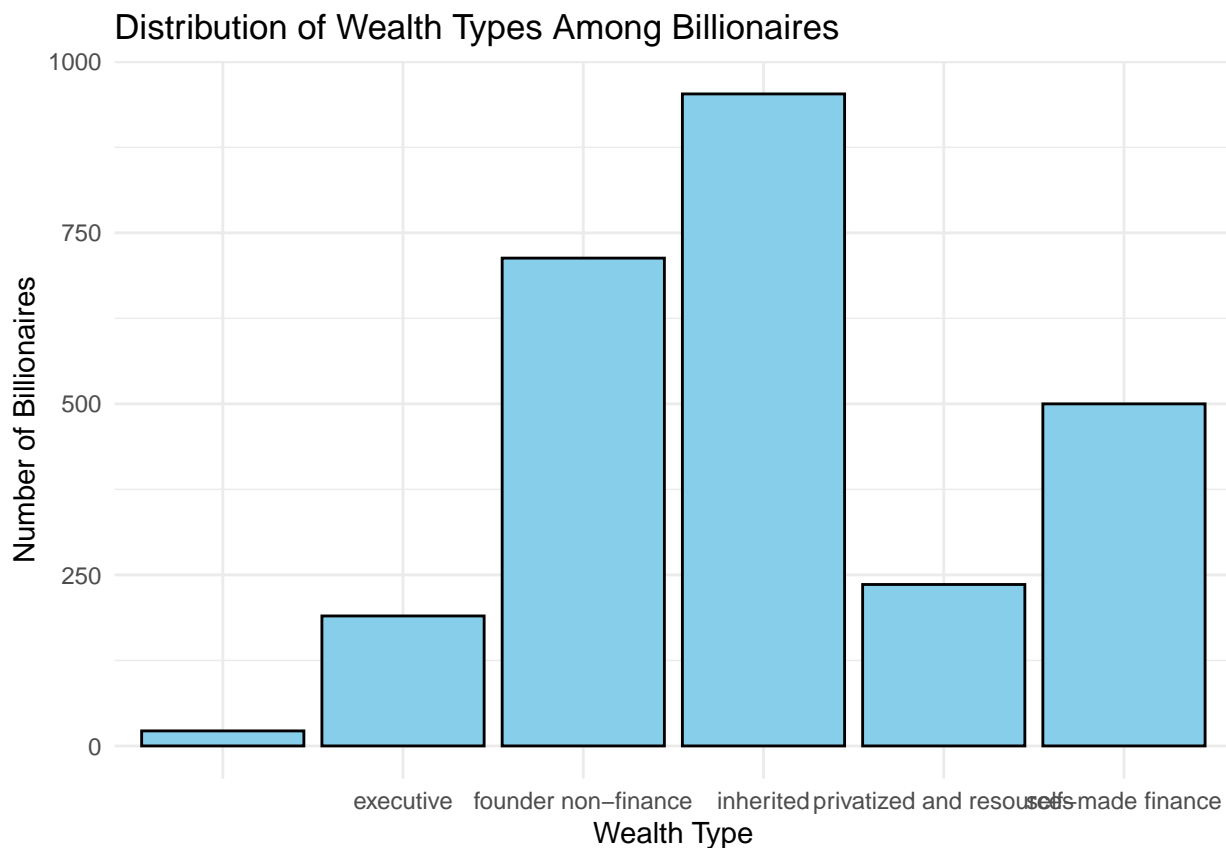

Step 2 Discuss the success of your plot:

This plot effectively displays the distribution of wealth types among billionaires. I do not think a legend is needed because there is only one variable mapped to the x-axis.

Design-wise, I chose to make the bars a pretty sky blue color as it makes the plot visually appealing and adds some contrast on the black and grey in the plot.

One possible improvement that I found challenging was that I could be sort the bars by frequency so that they either ascend or descend.

Step 3

I found this plot very interesting as it shows me that the most billionaires INHERIT their money. I guess that is not shocking but the findings also state that the runner ups to inheriting are being founders of companies but not making their own money in particular and then self-made financial workers. A bar plot was useful in helping me learn this because it instantly shows you the number one way people become billionaires and then the least effective way of becoming a billionaire. Possible limitations include the fact that this data does not encompass all billionaires, it does not show what sectors of business these selfmade and founding billionaires come from, and it also does not display what age these billionaires are. Somoenme who inherits money can become a billionaire at the age of 70 whereas another billionaire could be young, worked hard, and in their 40s.

Run all chunks after finalizing and knit to a PDF, HTML or Doc (and convert to a PDF if knitting to HTML or Doc) following the naming convention:

DSC201_602_SP24_codingassignment8_unityID

(For example, DSC201_602_SP24_codingassignment8_keporte2)

Points: 20 points

Due: March 27 at 11:59 PM