HW 1 - Data visualization

Harrison Kane

! Important

This homework is due Thursday, Sep 15 at 11:59pm ET.

Getting Started

- Go to the sta199-f22-1 organization on GitHub. Click on the repo with the prefix `hw-01`. It contains the starter documents you need to complete the homework assignment.
- Clone the repo and start a new project in RStudio. See the Lab 0 instructions for details on cloning a repo and starting a new R project.

Packages

```
library(tidyverse)
library(openintro)
```

Guidelines + tips

As we've discussed in lecture, your plots should include an informative title, axes should be labeled, and careful consideration should be given to aesthetic choices.

Remember that continuing to develop a sound workflow for reproducible data analysis is important as you complete this homework and other assignments in this course. There will be periodic reminders in this assignment to remind you to knit, commit, and push your changes to GithHub. You should have **at least 3 commits with meaningful commit messages** by the end of the assignment.

ℹ Note

Note: Do not let R output answer the question for you unless the question specifically asks for just a plot. For example, if the question asks for the number of columns in the data set, please type out the number of columns. You are subject to lose points if you do not.

Workflow + formatting

Make sure to

- Update author name on your document.
- Label all code chunks informatively and concisely.
- Follow the Tidyverse code style guidelines.
- Make at least 3 commits.
- Resize figures where needed, avoid tiny or huge plots.
- Turn in an organized, well formatted document.

Exercises

Data 1: Duke Forest houses

ℹ Note

Use this dataset for Exercises 1 and 2.

For the following two exercises you will work with data on houses that were sold in the Duke Forest neighborhood of Durham, NC in November 2020. The `duke_forest` dataset comes from the **openintro** package. You can see a list of the variables on the package website or by running `?duke_forest` in your console.

Exercise 1

Suppose you're helping some family friends who are looking to buy a house in Duke Forest. As they browse Zillow listings, they realize some houses have garages and others don't, and they wonder: **Does having a garage make a difference?**

Luckily, you can help them answer this question with data visualization!

- Make histograms of the prices of houses in Duke Forest based on whether they have a garage.

  - In order to do this, you will first need to create a new variable called `garage` (with levels `"Garage"` and `"No garage"`).
  - Below is the code for creating this new variable. Here, we `mutate()` the `duke_forest` data frame to add a new variable called `garage` which takes the value `"Garage"` if the text string `"Garage"` is detected in the `parking` variable and takes the test string `"No garage"` if not.

```
duke_forest |>
  mutate(garage = if_else(str_detect(parking, "Garage"),   "Garage", "No garage")) |>
  ggplot(aes(x = price, fill = garage)) +
  geom_histogram(binwidth = 500000, show.legend = FALSE) +
  facet_wrap(~garage) +
  labs(
    x = "Price", y = "Count",
    title = "Distirbution of House Price By Garage/No Garage"
  )
```

- Then, facet by `garage` and use different colors for the two facets.
- Choose an appropriate binwidth and decide whether a legend is needed, and turn it off if not.
- Include informative title and axis labels.
- Finally, include a brief (2-3 sentence) narrative comparing the distributions of prices of Duke Forest houses that do and don't have garages. Your narrative should touch on whether having a garage "makes a difference" in terms of the price of the house.

Based on the distributions above, having a garage doesn't "make a difference" in the pricing of the house. While there are more houses without a garage priced between $500k and 1 million, there are roughly an equal amount of houses priced below and above this range.

Now is a good time to render, commit, and push. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceding.

Exercise 2

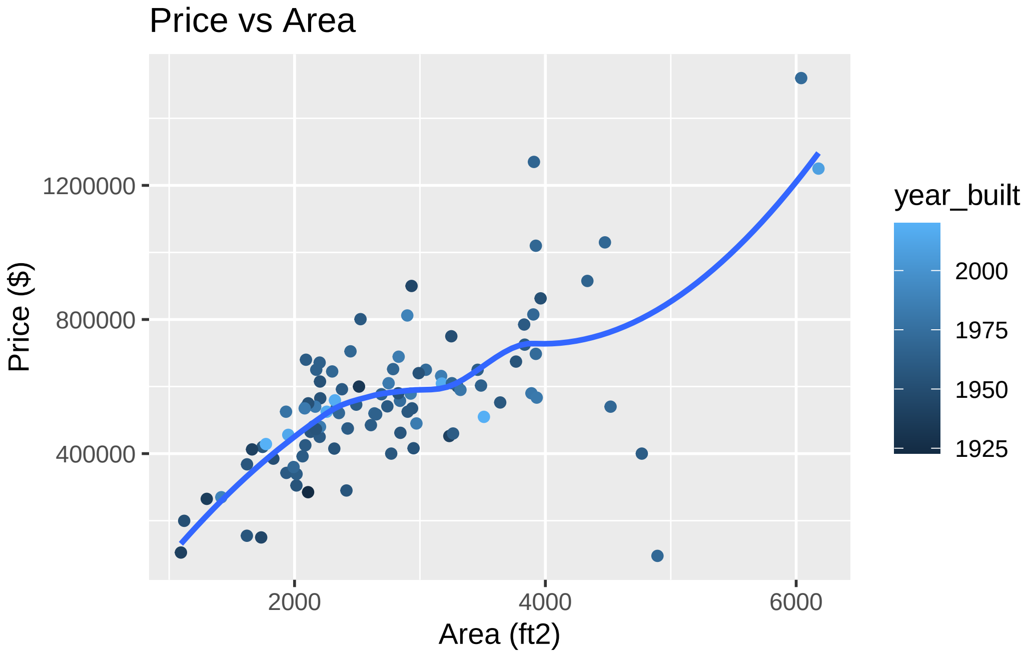
It's expected that within any given marker larger houses will be priced higher. It's also expected that the age of the house will have an effect on the price. However in some markets new houses might be more expensive while in others new construction might mean "no character" and hence be less expensive. So your family friends ask: "In Duke Forest, do houses that are bigger and more expensive tend to be newer ones than those that are smaller and cheaper?"

Once again, data visualization skills to the rescue!

- Create a scatter plot to exploring the relationship between `price` and `area`, conditioning for `year_built`.
- Use `geom_smooth()` with the argument `se = FALSE` to add a smooth curve fit to the data and color the points by `year_built`.
- Include informative title, axis, and legend labels.

```
ggplot(duke_forest, aes(x = area, y = price)) +
  geom_point(aes(color = year_built)) + geom_smooth(se = FALSE) +
  labs(x = "Area (ft2)", y = "Price ($)",
       title = "Price vs Area")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



- Discuss each of the following claims (1-2 sentences per claim). Your discussion should touch on specific things you observe in your plot as evidence for or against the claims.
  - Claim 1: Larger houses are priced higher. Supported - best fit line has positive slope.
  - Claim 2: Newer houses are priced higher. Not supported - different color points throughout whole plot.
  - Claim 3: Bigger and more expensive houses tend to be newer ones than smaller and cheaper ones. Not supported - different color points throughout whole plot.

Now is a good time to render, commit, and push.

Make sure that you commit and push all changed documents and your Git pane is completely empty before proceding.

Data 2: BRFSS

> **Note**
>
> Use this dataset for Exercises 3 to 5.

> The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.
>
> Source: cdc.gov/brfss

In the following exercises we will work with data from the 2020 BRFSS survey. The originally come from here, though we will work with a random sample of responses and a small number of variables from the data provided. These have already been sampled for you and the dataset you'll use can be found in the `data` folder of your repo. It's called `brfss.csv`.

```r
brfss <- read_csv("data/brfss.csv")
```

Exercise 3

- How many rows are in the `brfss` dataset? What does each row represent? 2000 participants
- How many columns are in the `brfss` dataset? Indicate the type of each variable. 4 variables
- Include the code and resulting output used to support your answer.

```r
glimpse(brfss)
```

```
Rows: 2,000
Columns: 4
$ state          <chr> NA, "CO", "MN", "VA", "UT", "KS", "UT", "TX", "OR", "OH~
$ general_health <chr> "Fair", "Good", "Very good", "Excellent", "Very good", ~
$ smoke_freq     <chr> "Not at all", "Some days", "Every day", "Not at all", "~
$ sleep          <dbl> 6, 7, 6, 8, 7, 10, 7, 6, 8, 8, 8, 6, 9, 8, 7, 7, 8, 6, ~
```

Now is a good time to render, commit, and push.

Exercise 4

Do people who smoke more tend to have worse health conditions?

- Use a segmented bar chart to visualize the relationship between smoking (`smoke_freq`) and general health (`general_health`). Decide on which variable to represent with bars and which variable to fill the color of the bars by.
- Pay attention to the order of the bars and, if need be, use the `fct_relevel` function to reorder the levels of the variables.
  - Below is sample code for releveling `general_health`. Here we first convert `general_health` to a factor (how R stores categorical data) and then order the levels from Excellent to Poor.

```
brfss |>
  mutate(
    general_health = as.factor(general_health),
    general_health = fct_relevel(general_health, "Excellent", "Very good", "Good", "Fair",
    smoke_freq = fct_relevel(smoke_freq, "Every day", "Some days", "Not at all")
  ) |>
  ggplot(aes(x = general_health, fill = smoke_freq)) +
  geom_bar() +
  labs(x = "General Health", y = "Count",
       title = "General Health from Smoking Frequency")
```

- Include informative title, axis, and legend labels.
- Comment on the motivating question based on evidence from the visualization: Do people who smoke more tend to have worse health conditions? I guess? Would probably be easier to facet by smoke_freq and look at each individual distribution.

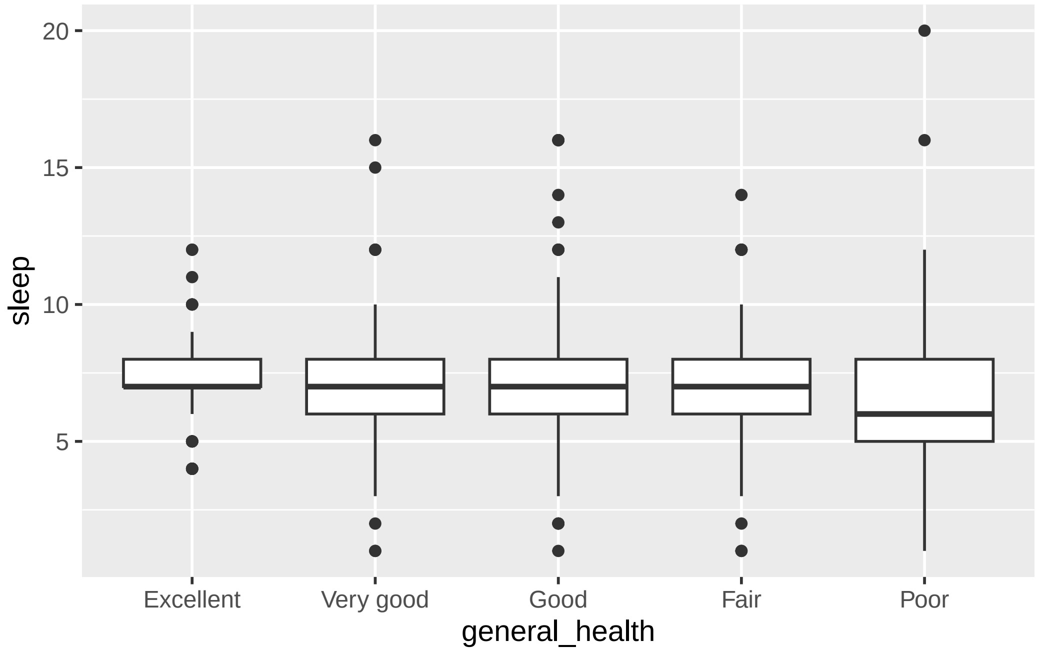
Now is a good time to render, commit, and push.

Exercise 5

How are sleep and general health associated?

- Create a visualization displaying the relationship between `sleep` and `general_health`.
- Include informative title and axis labels.
- Modify your plot to use a different theme than the default.
- Comment on the motivating question based on evidence from the visualization: How are sleep and general health associated? People with poor general have a larger variability in sleep length.

```
brfss |>
  mutate(
    general_health = as.factor(general_health),
    general_health = fct_relevel(general_health, "Excellent", "Very good", "Good", "Fair",
  ggplot(aes(x = general_health, y = sleep)) + geom_boxplot() +
  theme()
```



Now is a good time to render, commit, and push.

Exercise 6

(a) Fill in the blanks:

  - The gg in the name of the package ggplot2 stands for Grammar of Graphics.

    - If you map the same continuous variable to both `x` and `y` aesthetics in a scatterplot, you get a straight diagonal line. (Choose between "vertical", "horizontal", or "diagonal".)

(b) Code style: Fix up the code style by spaces and line breaks where needed. Briefly describe your fixes. (Hint: You can refer to the Tidyverse style guide.)

```
ggplot(data = mpg,mapping = aes(x = drv, fill = class)) +
  geom_bar() +
  scale_fill_viridis_d()
```

(c) Read `?facet_wrap`. What does `nrow` do? What does `ncol` do? What other options control the layout of the individual panels? Why doesn't `facet_grid()` have `nrow` and `ncol` arguments? nrow and ncol specify the number of rows adn columns in the 2d wrapping.

Render, commit, and push one last time.

Make sure that you commit and push all changed documents and your Git pane is completely empty before proceding.

Wrap up

Submission

- Go to http://www.gradescope.com and click Log in in the top right corner.
- Click School Credentials Duke Net ID and log in using your Net ID credentials.
- Click on your STA 199 course.
- Click on the assignment, and you'll be prompted to submit it.
- Mark all the pages associated with exercise. All the pages of your homework should be associated with at least one question (i.e., should be "checked"). If you do not do this, you will be subject to lose points on the assignment.
- Select the first page of your PDF submission to be associated with the "Workflow & formatting" question.

Grading

- Exercise 1: 7 points
- Exercise 2: 9 points
- Exercise 3: 5 points
- Exercise 4: 9 points

- Exercise 5: 7 points
- Exercise 6: 8 points
- Workflow + formatting: 5 points
- Total: 50 points