Homework 01 - Data visualization

Jemimah Osei (jko35)”

Invalid Date

Setup

Load packages and data:

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(viridis)
```

```
Loading required package: viridisLite
```

```
library(scales)
```

```

Attaching package: 'scales'

The following object is masked from 'package:viridis':

    viridis_pal
```

```
The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

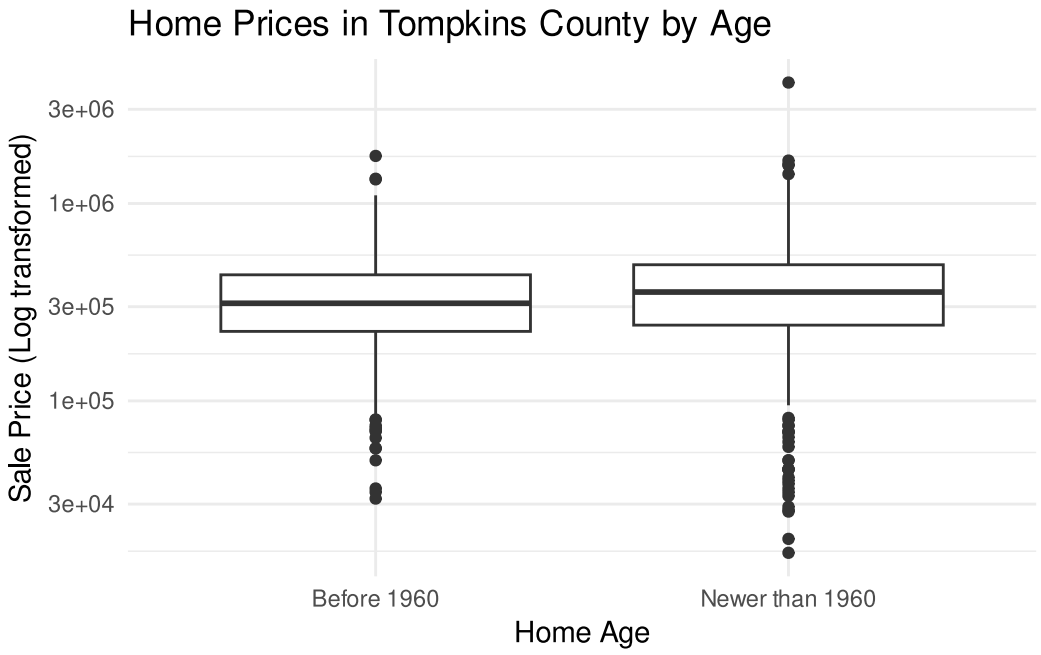
Exercises

Exercise 1

```
tompkins <- read_csv("data/tompkins-home-sales.csv")
```

```
Rows: 1247 Columns: 12
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (2): town, municipality
dbl  (9): price, beds, baths, area, lot_size, year_built, hoa_month, long, lat
date (1): sold_date

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tompkins <- tompkins |>
  mutate(home_age = if_else(year_built < 1960, "Before 1960", "Newer than 1960"))

ggplot(tompkins, aes(x = home_age, y = price)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Home Prices in Tompkins County by Age",
    x = "Home Age",
    y = "Sale Price (Log transformed)"
  ) +
  theme_minimal()
```



The boxplot shows that newer homes (built after 1960) have a higher median sale price and exhibit greater variability with more outliers. Thus while home age may have some impact on price, other factors likely play an important role in determining home value.

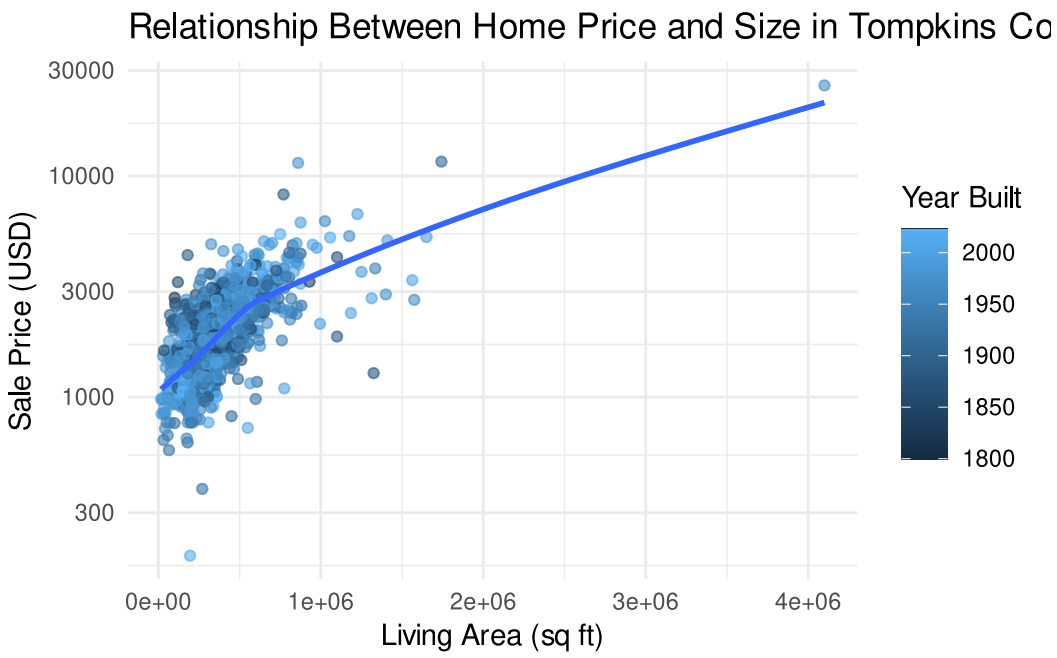
Exercise 2

```
ggplot(tompkins, aes(x = price, y = area, color = year_built)) +
  geom_point(alpha = 0.6) +
  geom_smooth(se = FALSE) +
  scale_y_log10() +
  labs(
    title = "Relationship Between Home Price and Size in Tompkins County",
    x = "Living Area (sq ft)",
    y = "Sale Price (USD)",
    color = "Year Built"
  ) +
  theme_minimal()
```

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

```
Warning: The following aesthetics were dropped during statistical transformation:
colour.
ℹ This can happen when ggplot fails to infer the correct grouping structure in
  the data.
```

```
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```



1.The scatter plot shows a positive correlation between area and price (i.e., points generally trend upward), then this claim **"Larger houses are priced higher"** is supported.

2.Newer homes are generally at higher price levels, thus the claim is supported.

3.The scatter plot reveals that larger and more expensive homes are generally represented by newer colors on the year_built scale. Thus that bigger and more expensive houses tend to be newer ones than smaller and cheaper properties.

Exercise 3

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

```
Rows: 2000 Columns: 4
── Column specification ────────────────────────────────────────────────────
Delimiter: ","
chr (3): state, general_health, smoke_freq
dbl (1): sleep

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
num_rows <- nrow(brfss)
num_columns <- ncol(brfss)
column_types <- sapply(brfss, class)
num_rows
```

```
[1] 2000
```

```r
num_columns
```

```
[1] 4
```

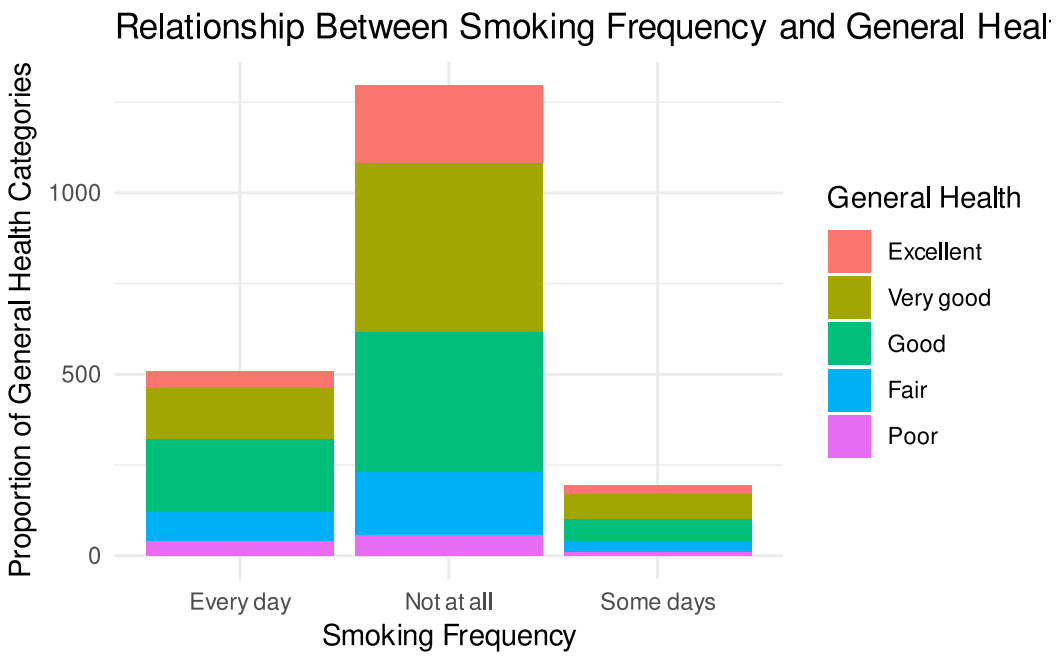
```r
column_types
```

```
         state general_health     smoke_freq          sleep
   "character"    "character"    "character"      "numeric"
```

The number of rows in the dataset corresponds to the total survey responses, with each row representing one respondent's data. The number of columns indicates the variables measured, and the `column_types` show the data type of each variable.

Exercise 4

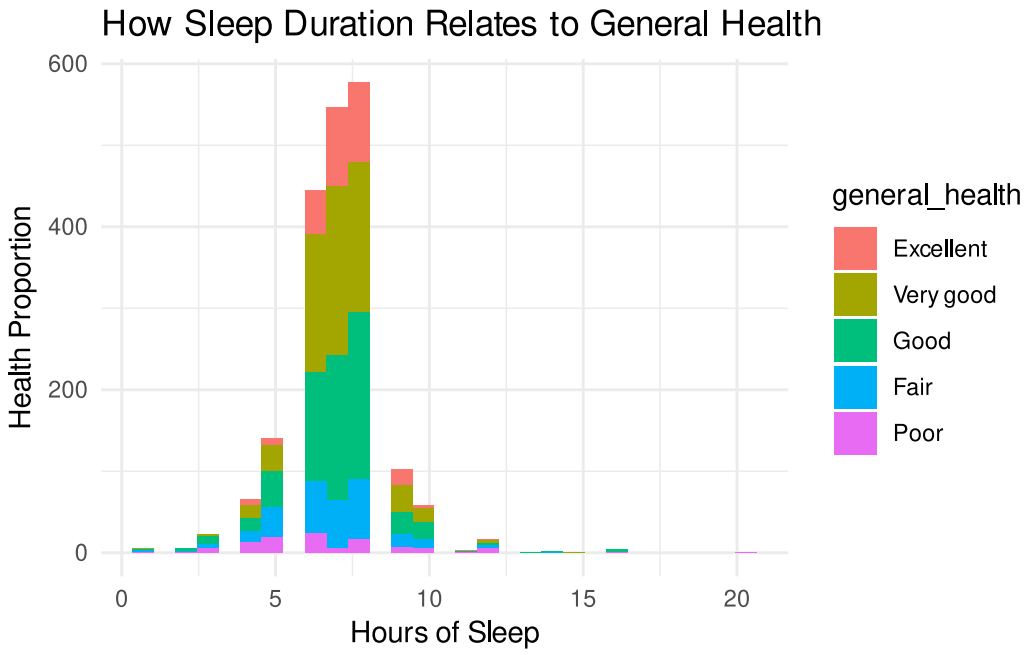
```r
brfss <- brfss |>
  mutate(
    general_health = as.factor(general_health),
    general_health = fct_relevel(general_health, "Excellent", "Very good",
                                 "Good", "Fair", "Poor")
  )

ggplot(brfss, aes(x = smoke_freq, fill = general_health)) +
  geom_bar() +
  labs(
    title = "Relationship Between Smoking Frequency and General Health",
    x = "Smoking Frequency",
    y = "Proportion of General Health Categories",
    fill = "General Health"
  ) +
  theme_minimal()
```



The chart indicates a clear pattern where individuals who smoke more frequently tend to report worse health, with higher proportions in the "Fair" and "Poor" health categories. This suggests a strong association between increased smoking frequency and poorer health outcomes.

Exercise 5

```
ggplot(brfss, aes(x = sleep, fill = general_health)) +
  geom_histogram(binwidth = .7) +
  labs(
    title = "How Sleep Duration Relates to General Health",
    x = "Hours of Sleep",
    y = "Health Proportion"
  ) +
  theme_minimal()
```



A larger proportion of individuals who get less sleep tend to report poorer health, with "Fair" and "Poor" health categories more prevalent in those with shorter sleep durations.

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 = penguins, mapping = aes(x=species, fill=island))+
  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: Controls the number of rows in a plot. **ncol**: Controls the number of columns in a plot.
`scales` is for axis scaling and `labeller` is for facet label customization labels.
For `facet_grid()` , the nrow and ncol arguments are unnecessary since the number of unique values of the variables specified in the function determines the number of rows and columns

For the above code I added spaces around the equal sign and after commas and break up the functions after the plus sign