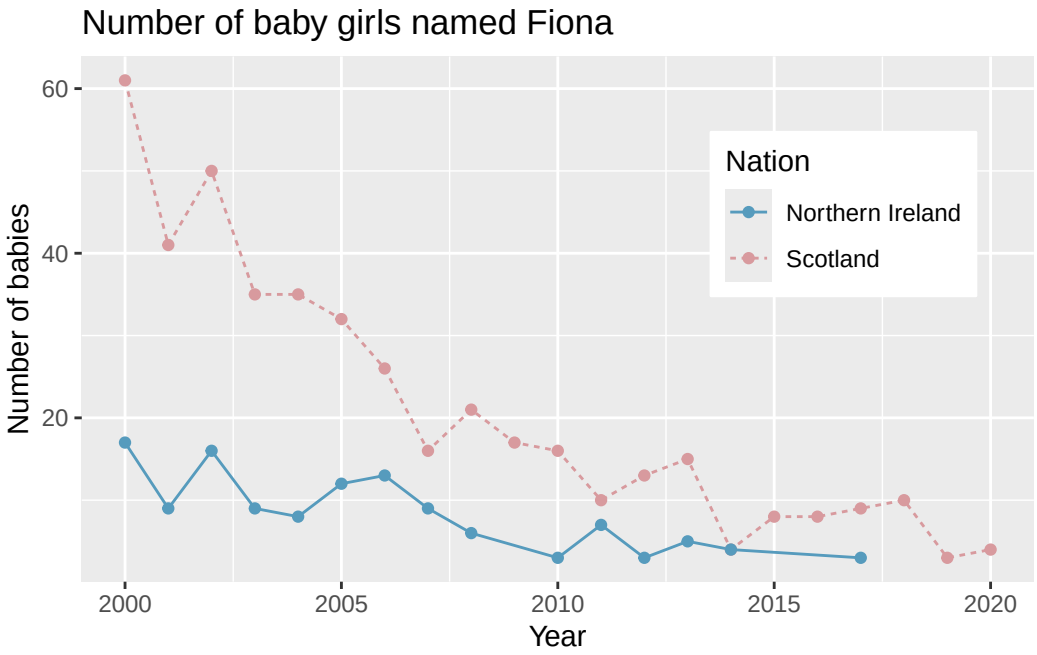
Stat 20: Problem Set 2

due Thursday, June 27 at 11:59pm

Questions 1-4 (A Grammar of Graphics)

Question 1 - UK Baby names.

The visualization below shows the number of baby girls born in the United Kingdom (comprised of England & Wales, Northern Ireland, and Scotland) who were given the name "Fiona" over the years.[1]



a. List the variables you believe were necessary to create this visualization.

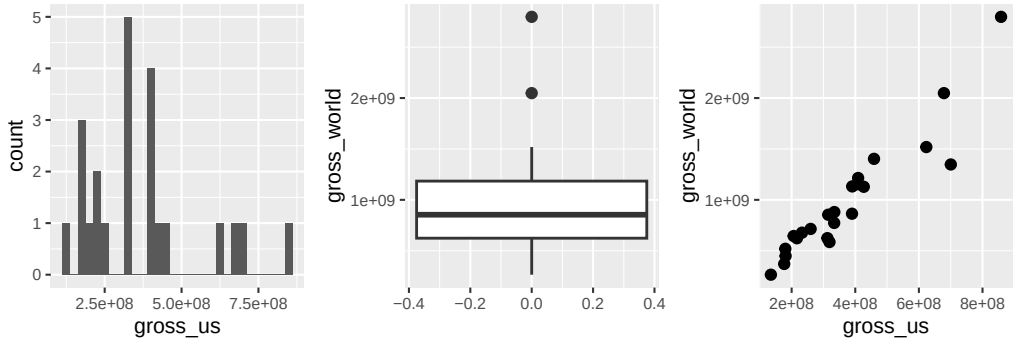
b. List the aesthetic mappings of each of the variables you noted in **part a** as per the *Grammar of Graphics*.

c. Identify the type of each variable in the *Taxonomy of Data*.

[1] The `ukbabynames` data used in this exercise can be found in the **ukbabynames** R package.

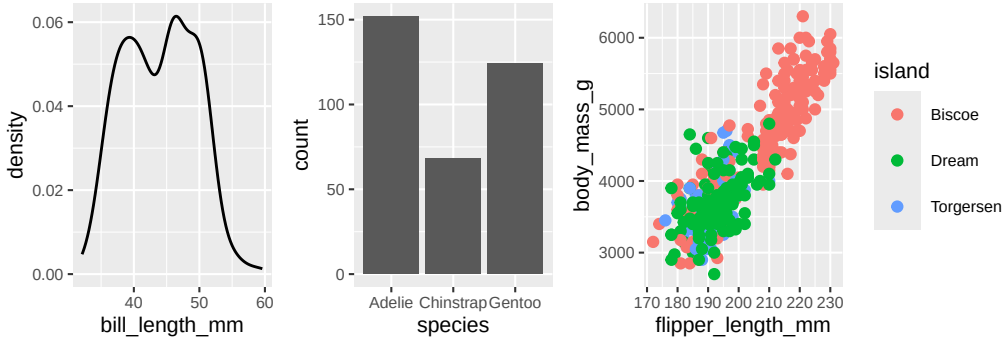
Question 2 - Practice with `ggplot2` I

The following three plots come from a data set called `mcu_films` *that is inside the* ***`openintro`*** *package.* Please write out the `ggplot2` code that will produce each one.



Question 3 - Practice with `ggplot2` II

The following three plots come from a data set called `penguins` *that is inside the* ***`stat20data`*** *package.* Please write out the `ggplot2` code that will produce each one.



Question 4 - Adele

One general rating system for music/movies is to rate the art from 1 stars to 5 stars; the more stars, the better. Sometimes there are half stars given as well.

Consider now the following data set (call it `Adele`), which tracks reviews of singer-songwriter Adele's last two studio albums, *25* (2015) and *30* (2021) by major music critic outlets. Data obtained from Metacritic. Where necessary, review scores have been translated from 1 to 5 stars using the following scale:

> **0-10:** 0.5 stars **11-20:** 1 stars **21-30:** 1.5 stars **31-40:** 2 stars **41-50:** 2.5 stars **51-60:** 3 stars **61-70:** 3.5 stars **71-80:** 4 stars **81-90:** 4.5 stars **91-100:** 5 stars

`Outlet`	`Review_25`	`Review_30`
Rolling Stone	5	5
The Telegraph (UK)	5	5
Pitchfork	4	4.5
Los Angeles Times	4	4.5
Consequence	4	5

a. With the column names given in the data set above, write code with `ggplot()` that would visualize the reviews of the two albums using a two-variable plot. Make sure to label your axes and give the plot a title. What is the geometry and aesthetic mapping(s) of the plot you chose?

b. Then **sketch the plot out** as you have coded it, and plot the points in the data set on the graph.

c. Did you run into any issues plotting your points? If so, what were they?

d. Now write code with `ggplot()` that would visualize the reviews of the two albums in such a way that would resolve the issues you should have identified in **part c**.

Questions 5-6 (A Grammar of Graphics)

Question 5

Provide the name of the `ggplot2` layer that does each of the following.

a. allows you to set x and y limits.
b. allows you to make add an annotation to the plot
c. allows you to add or modify axis, legend, and labels.

Question 6

For each of the following statements determine whether an aesthetic mapping or a setting, or neither has been applied.

a. Consider the `penguins` data. When plotting the relationship between `bill_length_mm` and `bill_depth_mm`, I decide to color all of the points red.

b. Consider the `penguins` data. When plotting the relationship between `bill_length_mm` and `bill_depth_mm`, I decide to color the points by species of penguin.

c. Consider the `penguins` data. I make a histogram of the `bill_length_mm` variable; and the frequencies (counts) of each of the histogram bins is mapped to the y-axis.

d. Consider a dataset called `images` based off of the shoebill image activity from the second day of class. Each row is one image, and there are three columns: `red`, `blue` and `green`, which each take values from 0 to 255, where 0 indicates no saturation of color and 255 indicates full saturation of color. Here is some code I wrote for a `ggplot` which involves use of the blue color. Mapping, or setting?

```
ggplot(data = images, mapping = aes(x = red)) +
  geom_histogram(binwidth = 25, color = "blue")
```

e. Consider the `flights` data. I make side-by-side boxplots of the `dep_delay` variable, grouped by airline `carrier`. The `carrier` variable is mapped to the y-axis.

Questions 7-9 (Data Pipelines)

Question 7 - Multiple Choice

Consider a made up dataset called `Students` with two columns: `ID` and `Birth_Year`. Which of the following lines of code would create a new column `ID_3000` which reads `TRUE` for rows whose `ID` number is greater than 3000 before arranging the data frame in ascending order by `Birth_Year`? *Circle the correct choice.*

```
Students %>%
    mutate(ID_3000 = (ID < 3000)) %>%
    arrange(Birth_Year)

Students %>%
  filter(ID_3000 = (ID > 3000)) %>%
  arrange(Birth_Year)

Students %>%
  mutate(ID_3000 = (ID > 3000)) %>%
  arrange(Birth_Year)

Students %>%
  mutate(ID_3000 = (ID > 3000)) %>%
  select(ID_3000, Birth_Year) %>%
  arrange(Birth_Year)
```

Question 8 - more on the `penguins` dataset in the `stat20data` library

Write your code answers in the spaces provided.

a. Extract a data frame from the original `penguins` data frame that excludes the Adelie penguins.

b. Then, with the new, extracted data frame, create a column that has the value `TRUE` for penguins with bill lengths between 40 and 50 mm and `FALSE` otherwise.

c. Using the new column you created, calculate the proportion of penguins in the data frame that have bill lengths between 40 and 50 mm.

d. Consider a new metric called `bill_size` that's the sum of the length and depth. What is the average bill size and it's standard deviation among each species, broken out among each of the island? You may end up with potentially nine pairs of statistics. Sort your resulting data structure in decreasing order by average bill size.

e. What are the total number of penguins in the data set belonging to each species-island combination? Why may have you not gotten nine pairs of statistics in the last question?

Question 9 - Air Quality data

You can access the `airquality` data directly just by typing "airquality" in a pipeline. The data descriptor reads "*daily air quality measurements in New York, May to September 1973.*" Write your code answers below.

a. Calculate the overall: mean, median, variance, standard deviation and IQR of the daily temperatures in New York City.

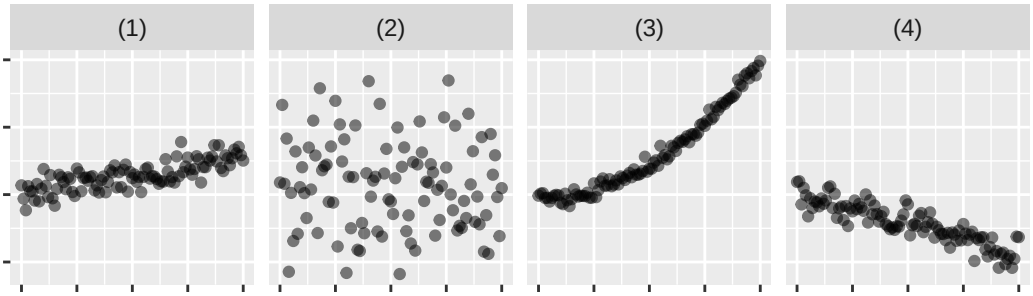
b. Then, calculate all of these statistics by month.

Questions 10-11 (Summarizing Numerical Associations)

Question 10 - Associations in scatterplots

For each of the four plots, indicate if they show, between the two variables:

a. a positive association
b. a negative association
c. no association.
d. if **part a** or **part b** is true: whether the association is linear or nonlinear.



Question 11 - `mtcars` dataset

There is a data set built into R called `mtcars` that includes several measures on different types of cars. Learn more about the data set using `?mtcars`. Write your code answers below.

a. Summarize the association between the fuel efficiency (measured in miles per gallon) and the weight of the car using a scatter plot, the correlation coefficient, and a linear model. Since we seek to explain the fuel efficiency, put that one on the y.

b. Repeat **part b** but use the horsepower of the car instead of the weight. Compare the scatter plots: why does one of them have a higher correlation coefficient than the other?

c. What is the better way to compare the strength of the linear relationship between these two pairs of variables (mpg and wt; mpg and hp): the correlation coefficients or the slopes of the linear models? Why?

d. Which car has the lowest fuel efficiency given its weight?

e. Visualize the relationship between number of forward gears and the number of cylinders. Address any overplotting that might occur, and title the plot with a claim about the strength of the association between the two variables.

Question 12 (Data Pipelines + Grammar of Graphics)

Use the `russian_influence_on_us_election_2016` dataset within the `MASS` library to generate the following plot as closely as possible. Write the code you used below.