PS2

Vivian Yeh

Question 1: Measures of center and spread

1. **Estimate Q1 (the median of the first half of the data), the median, and Q3 (the median of the second half of the data) from the histogram**

   Q1: 0

   The median: 10

   Q3: 30

2. **Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning**

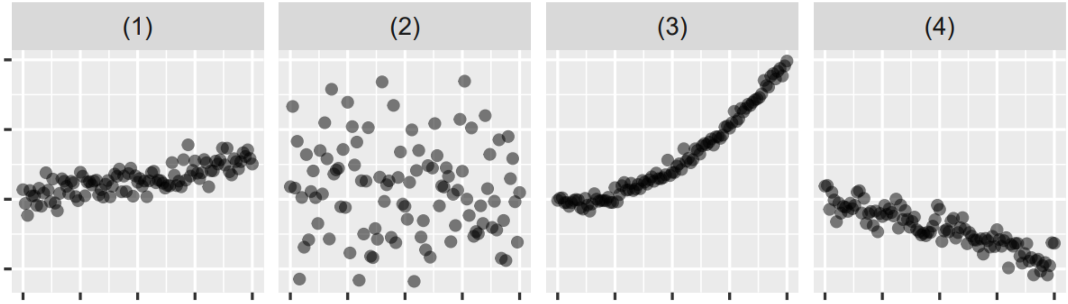
   I expect the mean of this data set to be larger than the median because the distribution of data is skewed to the right. The mode is on the left side of the data set, so the mean would probably larger than the median.

3. **Would the sample standard deviation or the IQR do a better job of measuring the spread of this distribution? Why?**

   The IQR would do a better job of measuring the spread of this distribution because this method is not affected by extreme outliers. The sample standard deviation is sensitive to the extreme outliers, and an extreme large value will cause the standard deviation to me much larger. Therefore, IQR is the better way to measure the spread of this skewed distribution which has outliers.

Question2: Associations in scatterplots

Indicate which of the plots show (a) a positive association, (b) a negative association, or (c) no association. Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.



(1) positive association / linear

(2) no association

(3) positive association / linear

(4) negative association / linear

Question3: Practice with ggplot I

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

```
library(openintro)
```

```
Loading required package: airports
```

```
Loading required package: cherryblossom
```

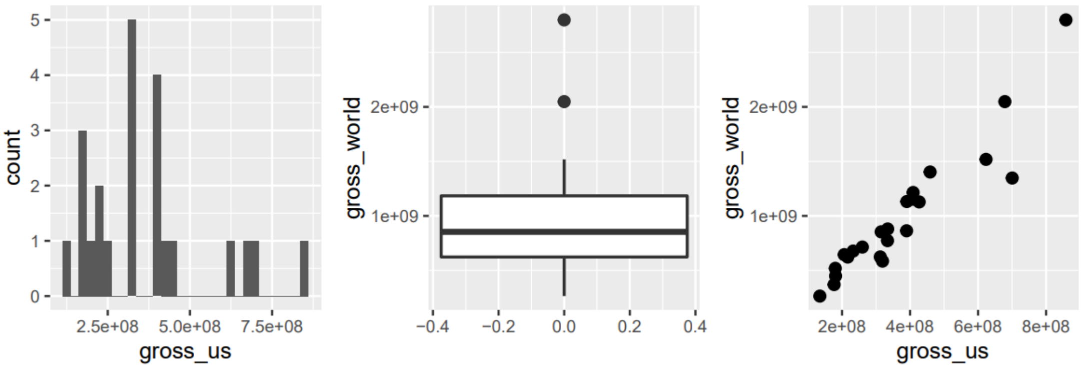
```
Loading required package: usdata
```

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
v ggplot2 3.3.6     v purrr   0.3.4
v tibble  3.1.8     v dplyr   1.0.10
v tidyr   1.2.0     v stringr 1.4.1
v readr   2.1.2     v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

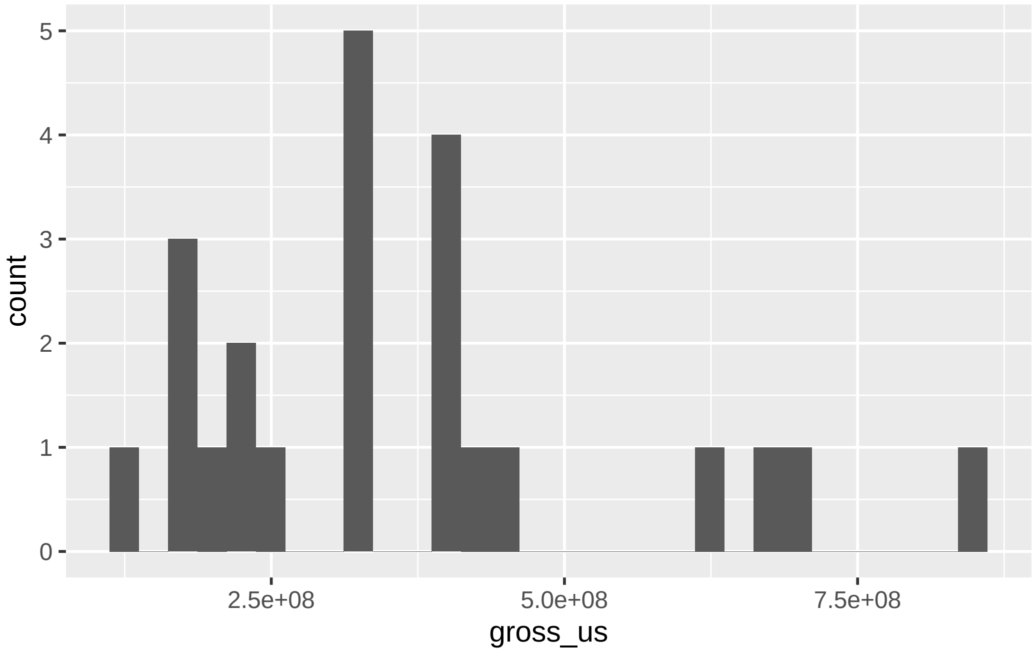
```
data(mcu_films)
```



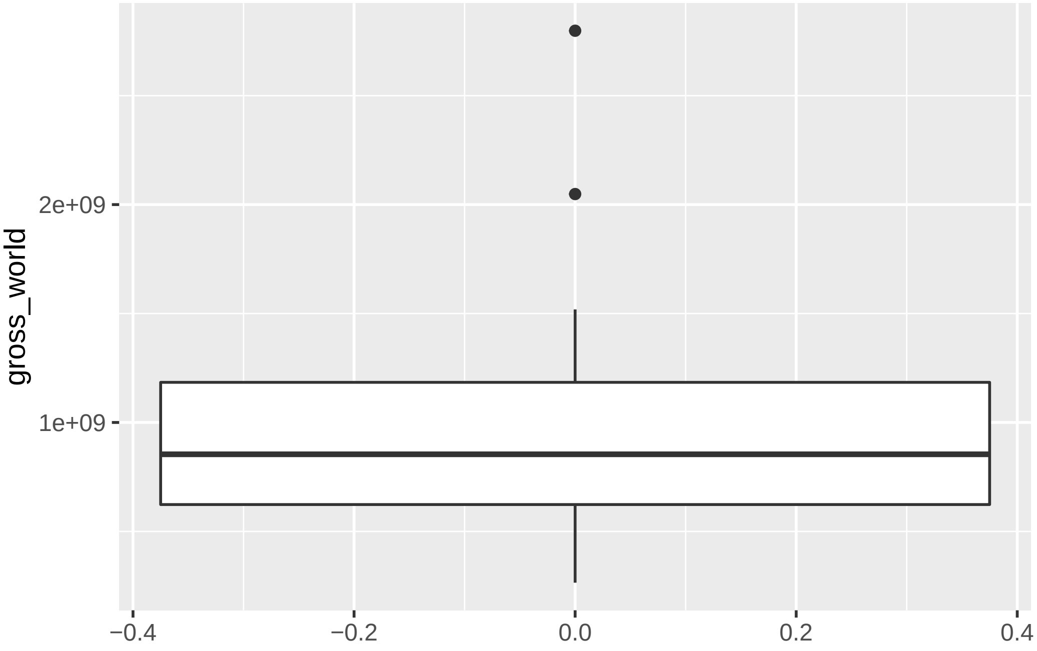
1.

```
mcu_films %>%
  ggplot(aes(x = gross_us)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

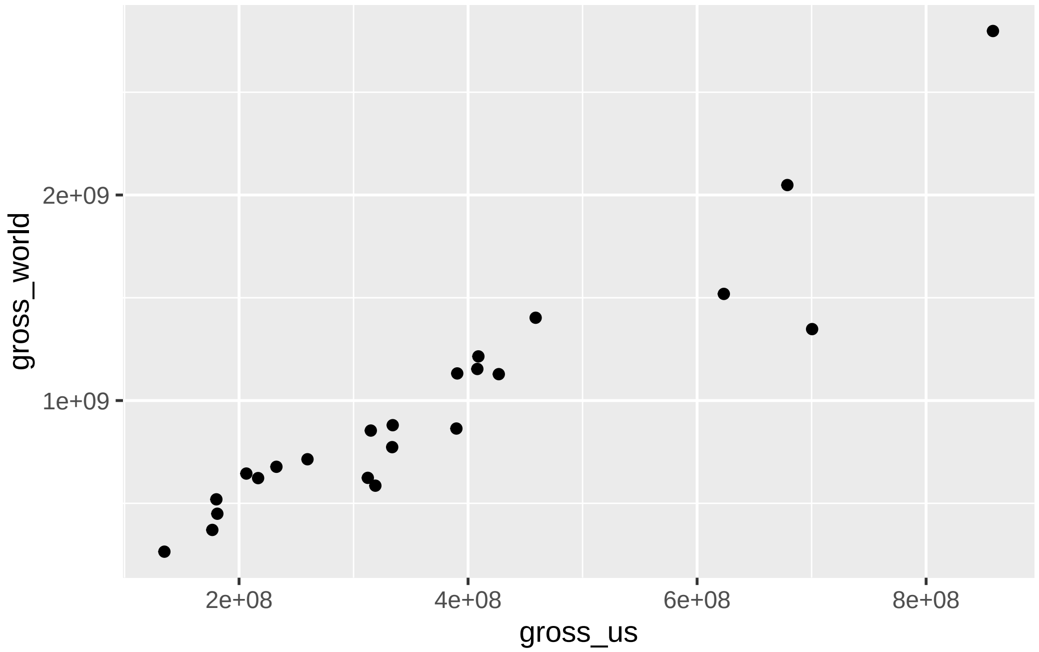


2.
```
mcu_films %>%
  ggplot(aes(y = gross_world)) +
  geom_boxplot()
```

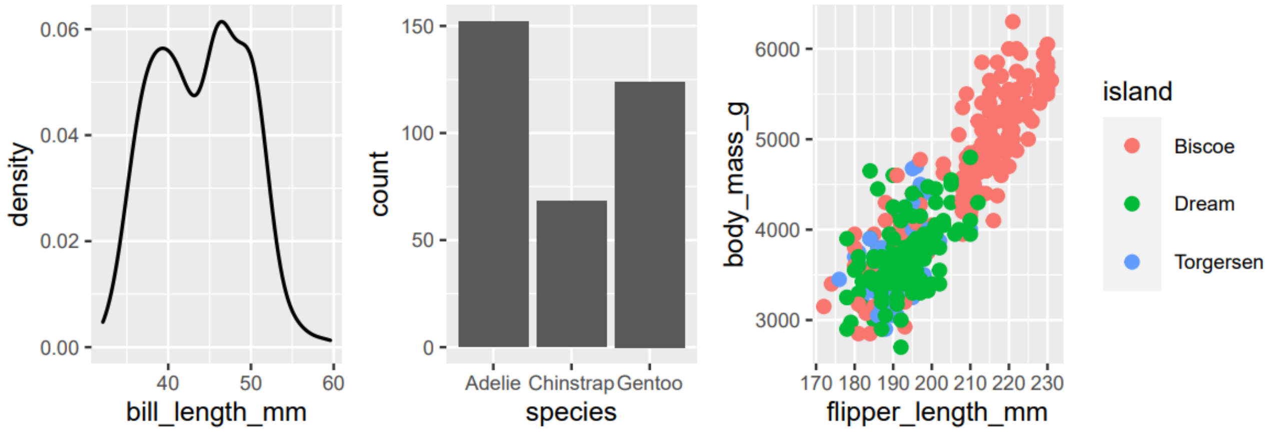


3.

```
mcu_films %>%
  ggplot(aes(x = gross_us,
             y = gross_world)) +
    geom_point()
```

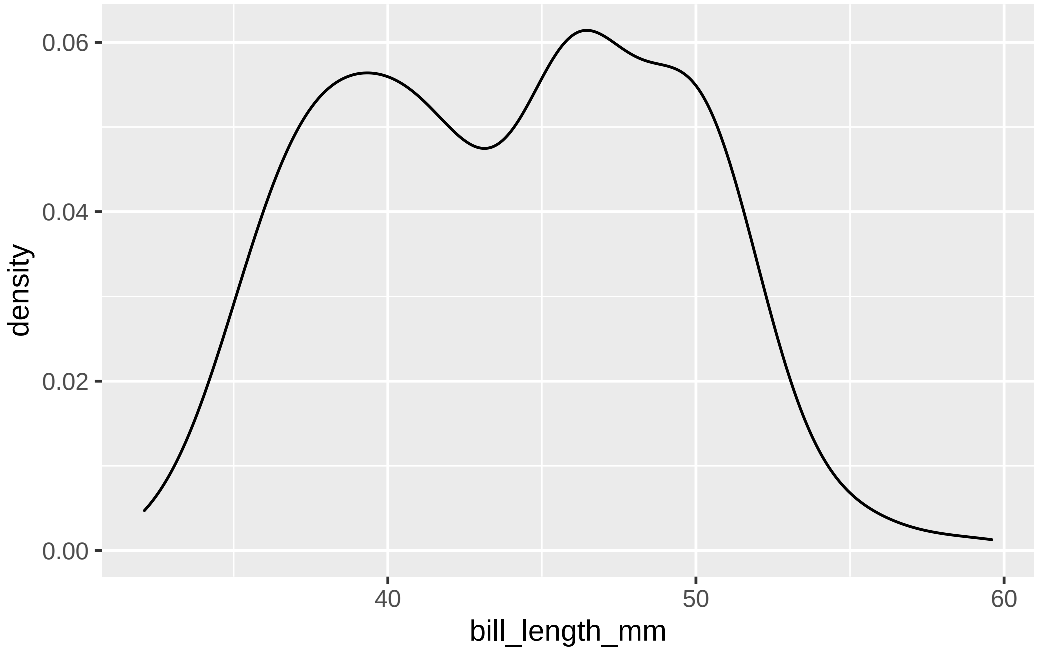


Question4: Practice with ggplot II

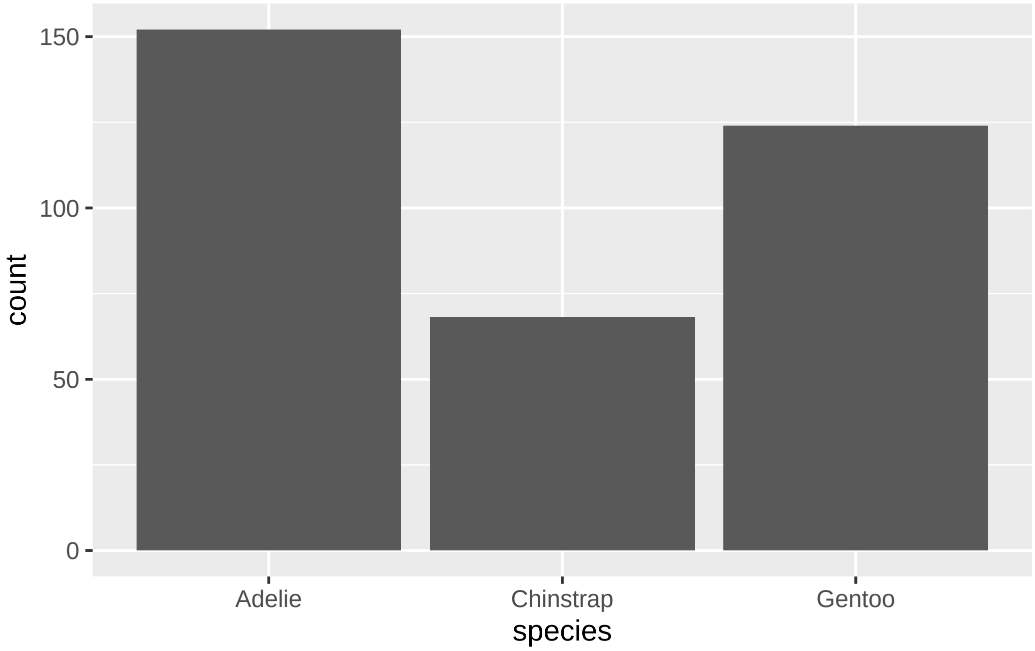


1.
```
library(palmerpenguins)
data(penguins)
penguins %>% ggplot(aes(x = bill_length_mm)) + geom_density()
```

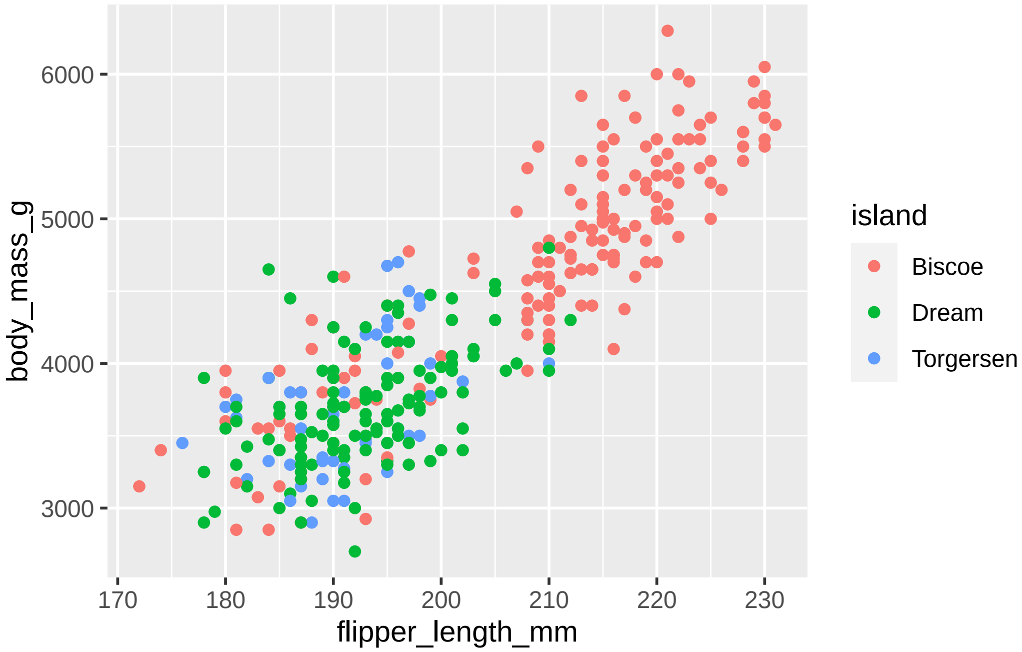
```
Warning: Removed 2 rows containing non-finite values (stat_density).
```



2.
```
penguins %>%
  ggplot(aes(x = species)) +
    geom_bar()
```



3.
```
penguins %>%
  ggplot(aes(x = flipper_length_mm,
             y = body_mass_g,
             color = island)) +
    geom_point()
```



Question 5: Numerical summaries with an interactive tutorial

“Yes, I have completed the tutorial”

Question 6: Review

Response Conservative Liberal Moderate

Apply for citizenship 57 101 120

Guest worker 121 28 113

Leave the country 179 45 126

Not sure 15 1 4

a. What percent of these Tampa, FL voters identify themselves as conservatives? Is this a joint, marginal, or conditional proportion?

```
(57 + 121 + 179 + 15) / 910
```

```
[1] 0.4087912
```

Marginal proportion

b. What percent of these Tampa, FL voters are in favor of the citizenship option? Is this a joint, marginal, or conditional proportion?

```
(57 + 101 + 120) / 910
```

```
[1] 0.3054945
```

Marginal proportion

c. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option? Is this a joint, marginal, or conditional proportion?

```
57 / 910
```

```
[1] 0.06263736
```

Joint proportion

d. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view? Are these joint, marginal, or conditional proportions?

```
con <- (0.063 / 0.409)

mod <- ((101/910) / (175/910))

lib <- ((120/910) / (363/910))
```

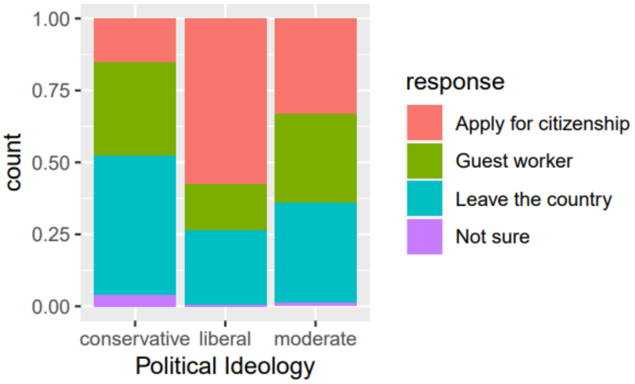
FL voters who identify themselves as conservatives are also in favor of the citizenship option = 0.154

FL voters who identify themselves as moderates are also in favor of the citizenship option = 0.577

FL voters who identify themselves as liberals are also in favor of the citizenship option = 0.331

These are conditional proportions

e. Based on the stacked bar chart shown below, do political ideology and views on immigration appear to be associated? Explain your reasoning.



Based on the stacked bar chart shown, political ideology and views on immigration appear to be associated because in the chart, there are visibly much more people who identified themselves as liberal apply for citizenship than other two groups of people who identified themselves as other political ideology. People who have conservative political ideology tend to leave the country rather than applying for citizenship and being guest workers. People who identified themselves as moderate have more balanced choices on views on immigration.

f. Conjecture other possible variables that might explain the potential relationship between these two variables.

Education background and levels, the level of salary, the category of jobs, and residency location may cause the different choice of being either a conservative, liberal, or moderate advocate.