Lab 4: Data Visualization

Jack Rellamas

02-29-2024

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking "Code" in the top right corner of this page.

Collaborators

INSERT NAMES OF ANY COLLABORATORS

```
# LOAD ANY RELEVANT PACKAGES HERE
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

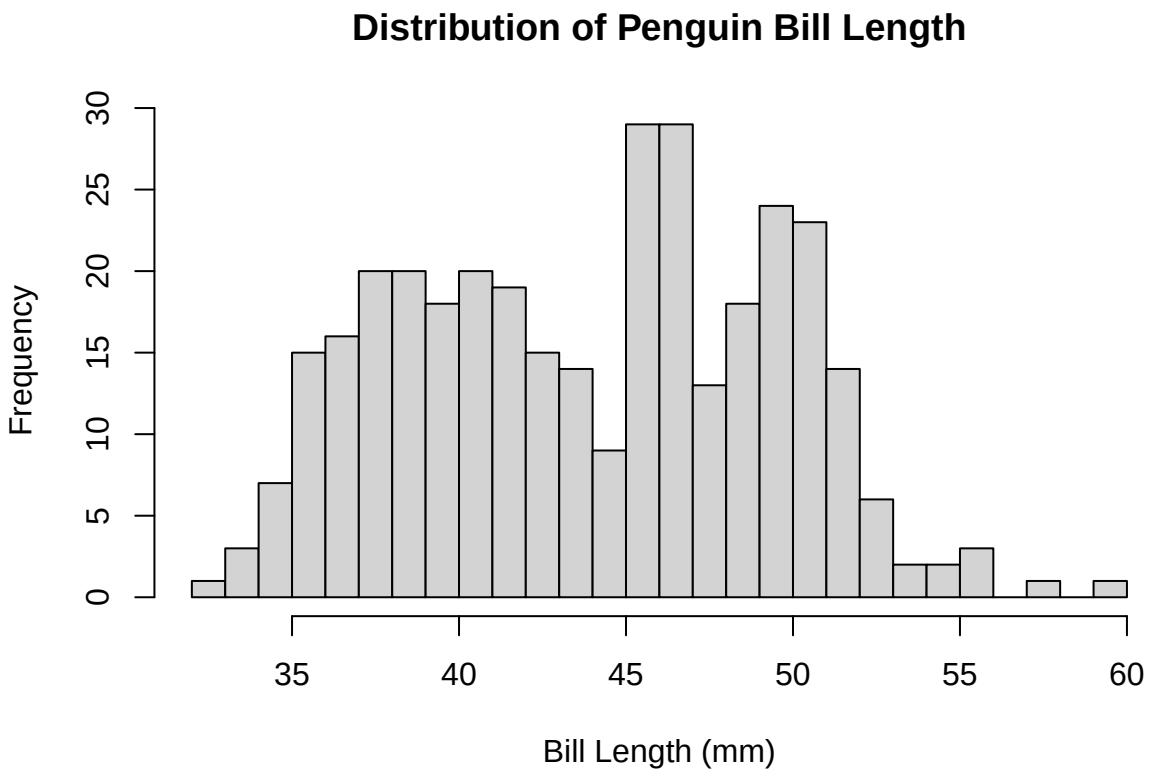
A. Basic visualizations

For this portion, we'll be using the `palmerpenguins` data. Use the following code to load the data.

```
library(palmerpenguins)
data(penguins)
```

1. Create and interpret a histogram of `bill_length_mm` using base R code. Be sure to use meaningful axis labels and titles.

```
hist(x = penguins$bill_length_mm,
     breaks = 20,
     main = "Distribution of Penguin Bill Length",
     xlab = "Bill Length (mm)")
```

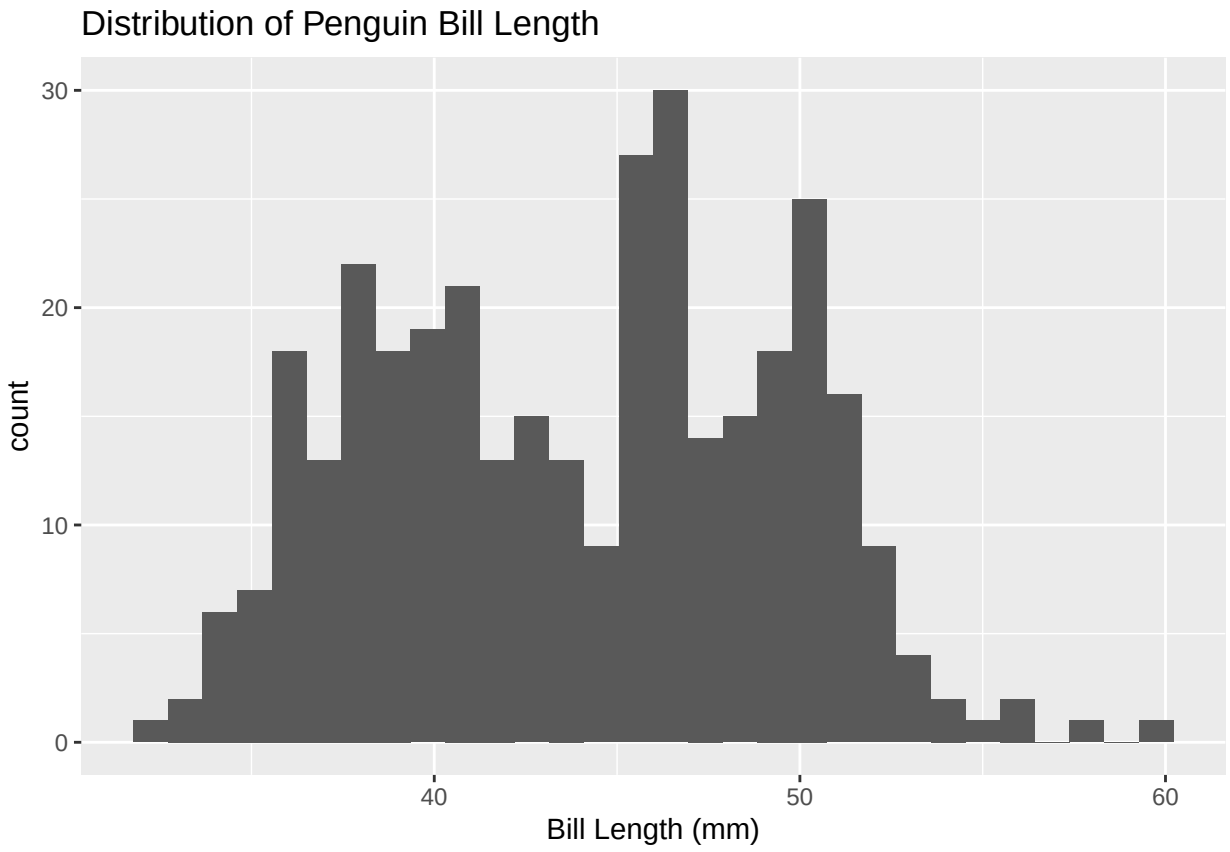


2. Create and interpret a histogram of `bill_length_mm` using ggplot2. Be sure to use meaningful axis labels and titles.

```
ggplot(data = penguins,
       aes(x = bill_length_mm)) +
  geom_histogram() +
  xlab("Bill Length (mm)") +
  ggtitle("Distribution of Penguin Bill Length")
```

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

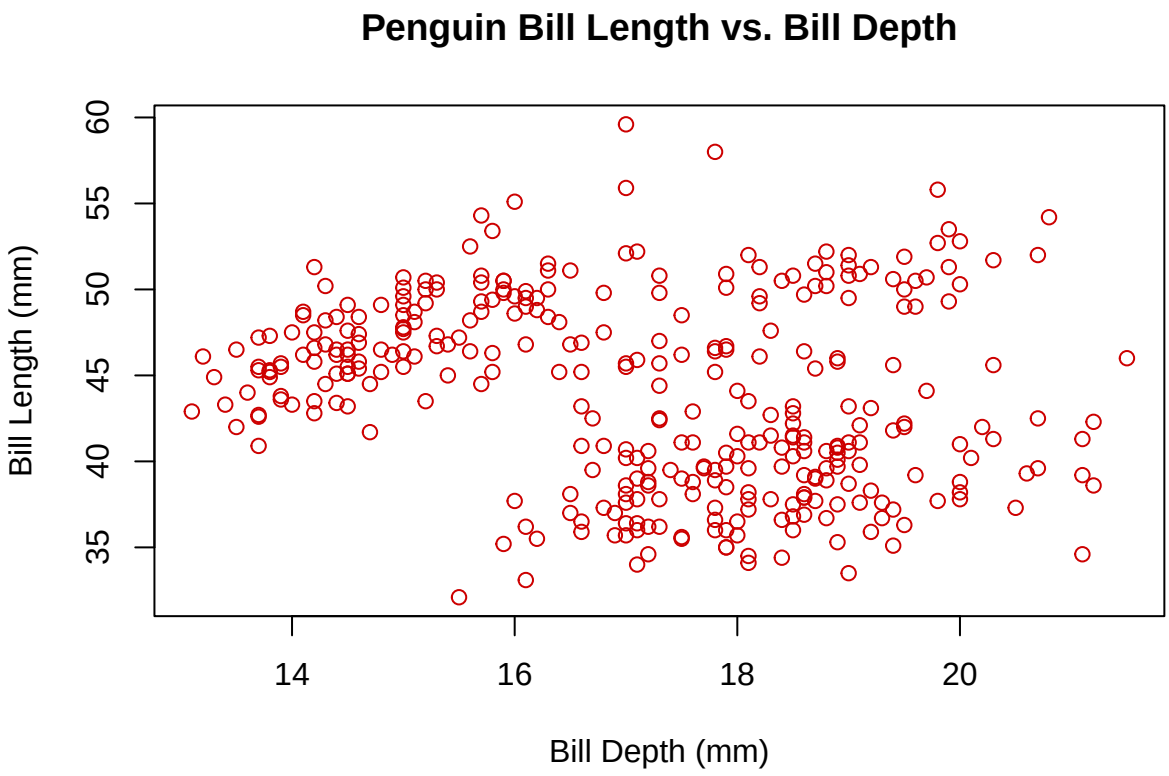
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



3. Create and interpret a scatterplot of `bill_length_mm` versus `bill_depth_mm` using base R code. Be sure to use meaningful axis labels and titles.

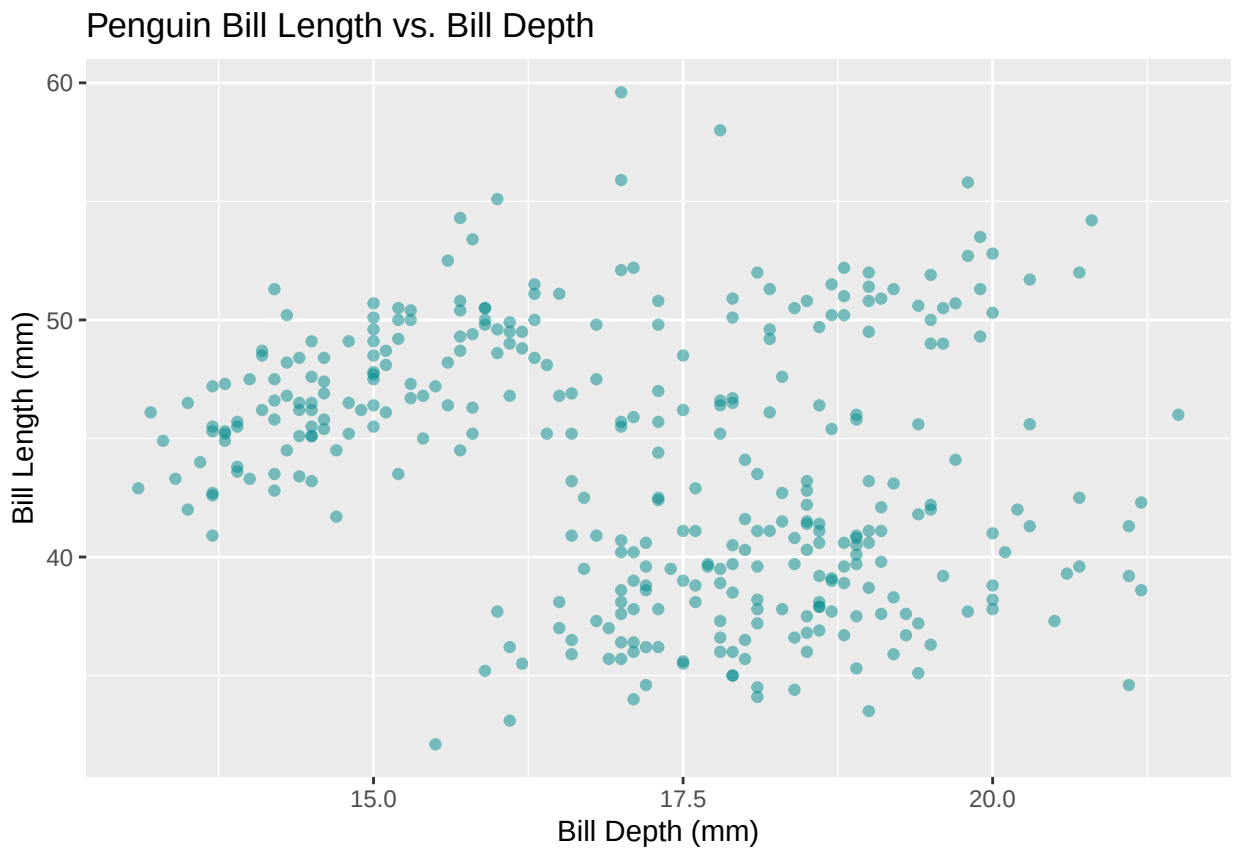
```
# filter for completed cases of bill_length vs bill_depth
clean_penguins <- penguins %>%
  filter(!is.na(bill_depth_mm) & !is.na(bill_length_mm))

plot(x = penguins$bill_depth_mm, y = penguins$bill_length_mm,
     main = "Penguin Bill Length vs. Bill Depth",
     xlab = "Bill Depth (mm)",
     ylab = "Bill Length (mm)",
     col = "red3",
     type = "p")
```



4. Create and interpret a scatterplot of `bill_length_mm` versus `bill_depth_mm` using ggplot2. Be sure to use meaningful axis labels and titles.

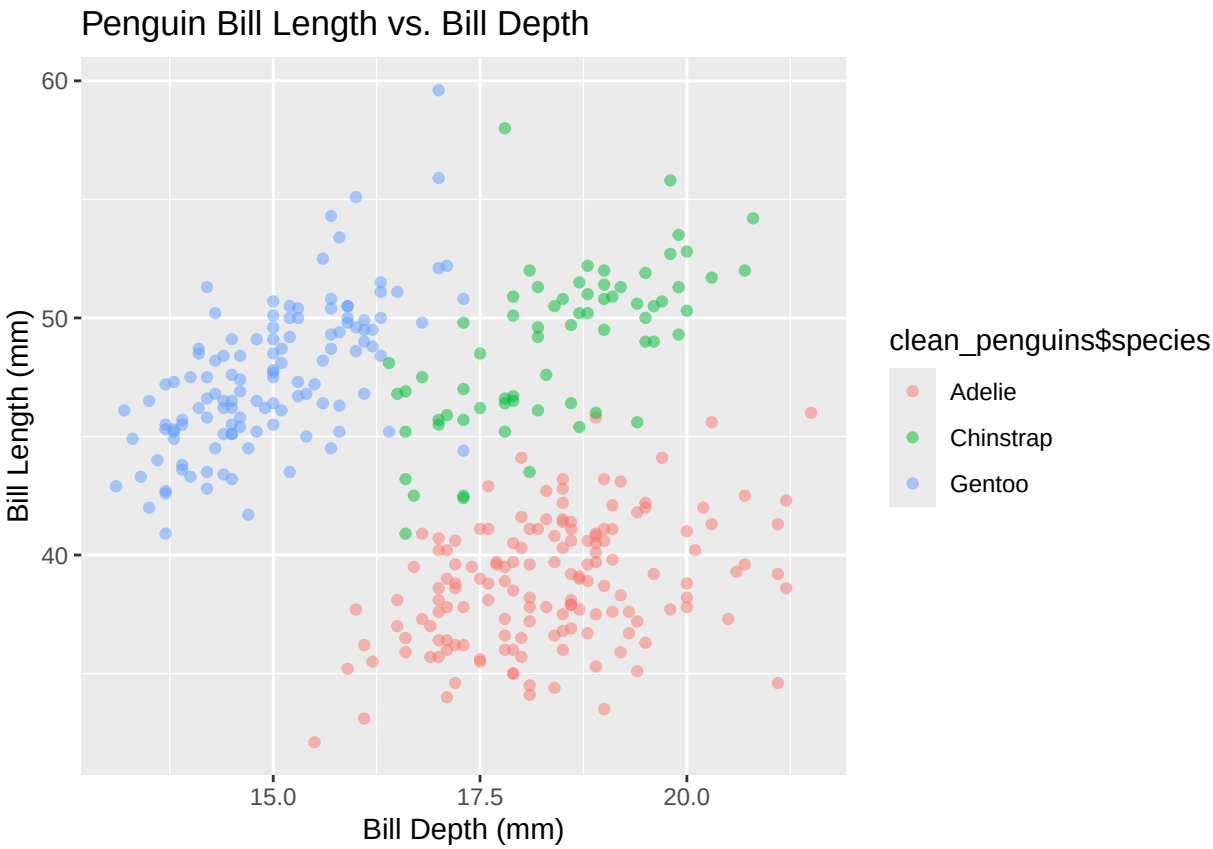
```
ggplot(data = clean_penguins,
       aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point(alpha = 0.5, color = "cyan4") +
  xlab("Bill Depth (mm)") +
  ylab("Bill Length (mm)") +
  ggtitle("Penguin Bill Length vs. Bill Depth")
```



5. Update your `ggplot2` scatterplot of `bill_length_mm` versus `bill_depth_mm` using ggplot2 so that the color of a point represents the corresponding penguin's species. What do you notice?

```
ggplot(data = clean_penguins,
       aes(x = bill_depth_mm, y = bill_length_mm, color = clean_penguins$species)) +
  geom_point(alpha = 0.5) +
  xlab("Bill Depth (mm)") +
  ylab("Bill Length (mm)") +
  ggtitle("Penguin Bill Length vs. Bill Depth")
```

```
## Warning: Use of ‘clean_penguins$species‘ is discouraged.
## i Use ‘species‘ instead.
```



B. Analyzing trends in San Jose rental prices

For this component, you will be exploring and visualizing data on Craigslist apartment rental postings in the Bay Area. The data are available here from Tidy Tuesday, as prepared by Dr. Kate Pennington. Note that you can use links within `read_csv()` to read online .csv files. I recommend saving a version of the unprocessed .csv on your machine in a `data` subfolder within your project folder so you will be able to work offline.

6. How many 1 bedroom listings from Santa Clara county are in this dataset?

```
sc_onebed <- filter(rent, beds == 1 & county == "santa clara")
table(sc_onebed$beds)
```

```
##
##     1
## 12455
```

There are 12455 1 bedroom listings from Santa Clara county in this dataset.

6. What is the median price for a 1 bedroom listing in Santa Clara county in 2018?

```
sc_onebed_2018 <- filter(sc_onebed, year == 2018)
summary(sc_onebed_2018)
```

```
##    post_id               date               year         nhood
##  Length:193         Min.   :20180104   Min.   :2018   Length:193
##  Class :character   1st Qu.:20180201   1st Qu.:2018   Class :character
##  Mode  :character   Median :20180315   Median :2018   Mode  :character
##                     Mean   :20180355   Mean   :2018
##                     3rd Qu.:20180524   3rd Qu.:2018
##                     Max.   :20180717   Max.   :2018
##
##      city              county              price           beds
##  Length:193         Length:193         Min.   : 258   Min.   :1
##  Class :character   Class :character   1st Qu.:1785   1st Qu.:1
##  Mode  :character   Mode  :character   Median :2095   Median :1
##                                        Mean   :2093   Mean   :1
##                                        3rd Qu.:2445   3rd Qu.:1
##                                        Max.   :5045   Max.   :1
##
##      baths            sqft         room_in_apt        address
##  Min.   :1.000   Min.   : 110.0   Min.   :0.00000   Length:193
##  1st Qu.:1.000   1st Qu.: 590.0   1st Qu.:0.00000   Class :character
##  Median :1.000   Median : 690.0   Median :0.00000   Mode  :character
##  Mean   :1.031   Mean   : 677.1   Mean   :0.01036
##  3rd Qu.:1.000   3rd Qu.: 769.2   3rd Qu.:0.00000
##  Max.   :2.000   Max.   :1621.0   Max.   :1.00000
##  NA's   :128     NA's   :45
##       lat             lon            title              descr
##  Min.   :37.33   Min.   :-121.9   Length:193         Length:193
##  1st Qu.:37.33   1st Qu.:-121.9   Class :character   Class :character
##  Median :37.33   Median :-121.9   Mode  :character   Mode  :character
##  Mean   :37.33   Mean   :-121.9
##  3rd Qu.:37.33   3rd Qu.:-121.9
##  Max.   :37.33   Max.   :-121.9
##  NA's   :177     NA's   :177
##    details
##  Length:193
##  Class :character
##  Mode  :character
##
##
##
##
```

The median price for a 1 bedroom listing in Santa Clara in 2018 is $2095.

6. Which county has the highest median price for a 1 bedroom listing in 2018?

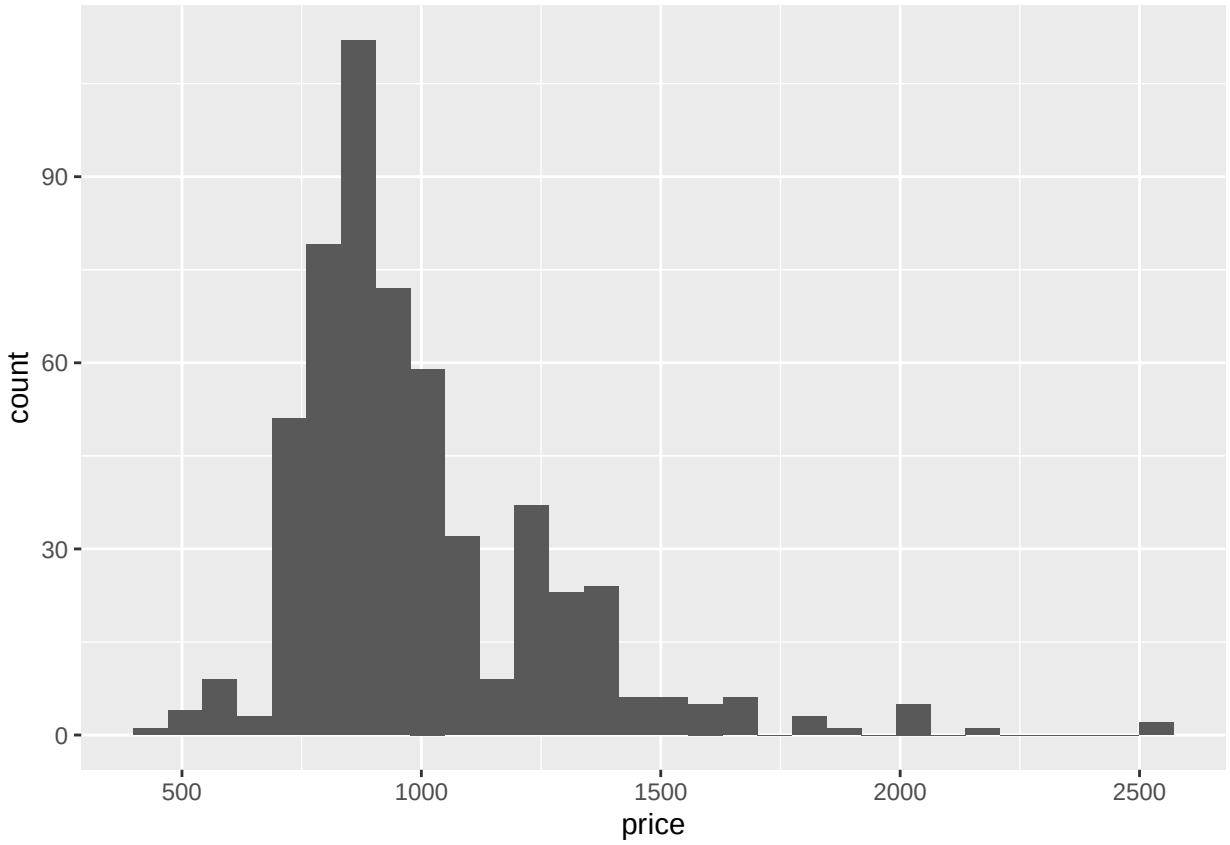
```
onebed_2018 <- rent %>% filter(beds == 1 & year == 2018)
onebed_2018 <- mutate(onebed_2018, county = factor(county))
summary_by_county <- aggregate(price~county, data=onebed_2018, summary) #  #https://www.tutorialspoint.
```

San Francisco is the county with the highest median price of rent at $3000.

7. Create two histograms for the prices of 1 bedroom listings in Santa Clara county in 2005 and 2018. Compare and discuss.

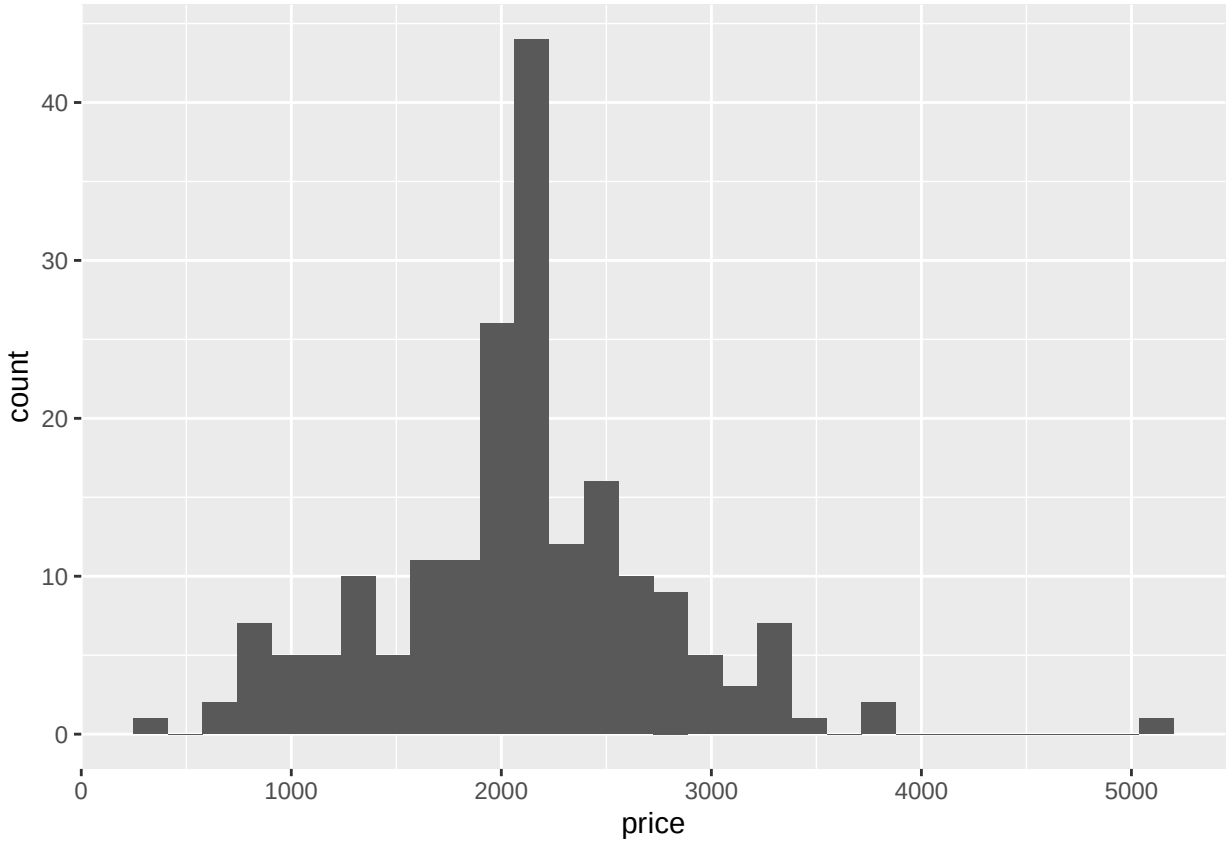
```
sc_1b_2005 <- rent %>% filter(beds == 1 & year == 2005 & county == "santa clara")
ggplot(data = sc_1b_2005,
       aes(x=price)) +
  geom_histogram()
```

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




```
sc_1b_2018 <- rent %>% filter(beds == 1 & year == 2018 & county == "santa clara")
ggplot(data = sc_1b_2018,
       aes(x=price)) +
  geom_histogram()
```

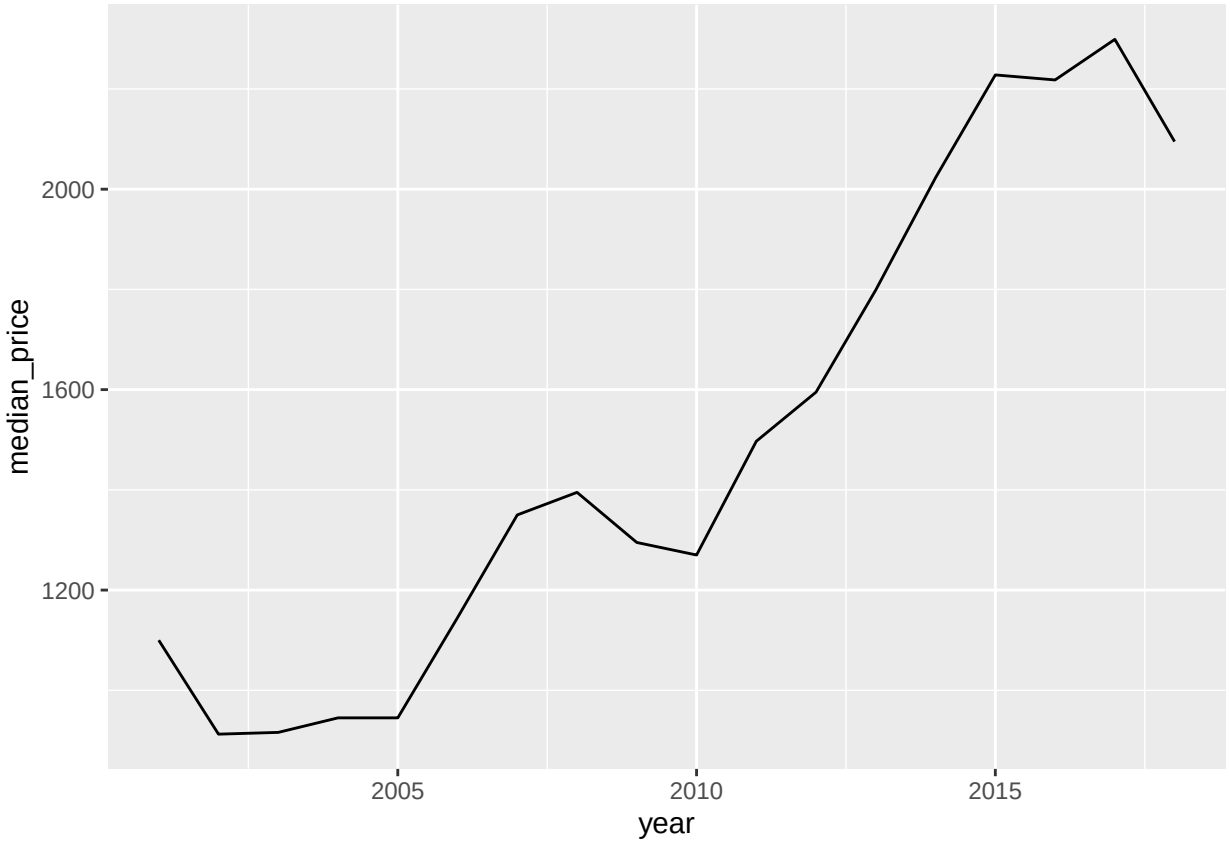
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The most common price for rent in 2005 was around $800. In 2018, the most common pricing for a one bedroom apartment in Santa Clara county had jumped to around $2000. The price distribution in 2005 is more right skewed meaning there were more apartment available on the affordable side of the market. The price distribution in 2018 is more normally distributed.

8. Create and interpret a line plot with year on the x-axis and median price for a 1 bedroom apartment for Santa Clara county on the y-axis from 2000 to 2018.

```
sc_median_price_change <- sc_onebed %>%
  group_by(year) %>%
  summarise(median_price = median(price))
ggplot(data = sc_median_price_change,
       aes(x = year, y = median_price)) +
  geom_line()
```



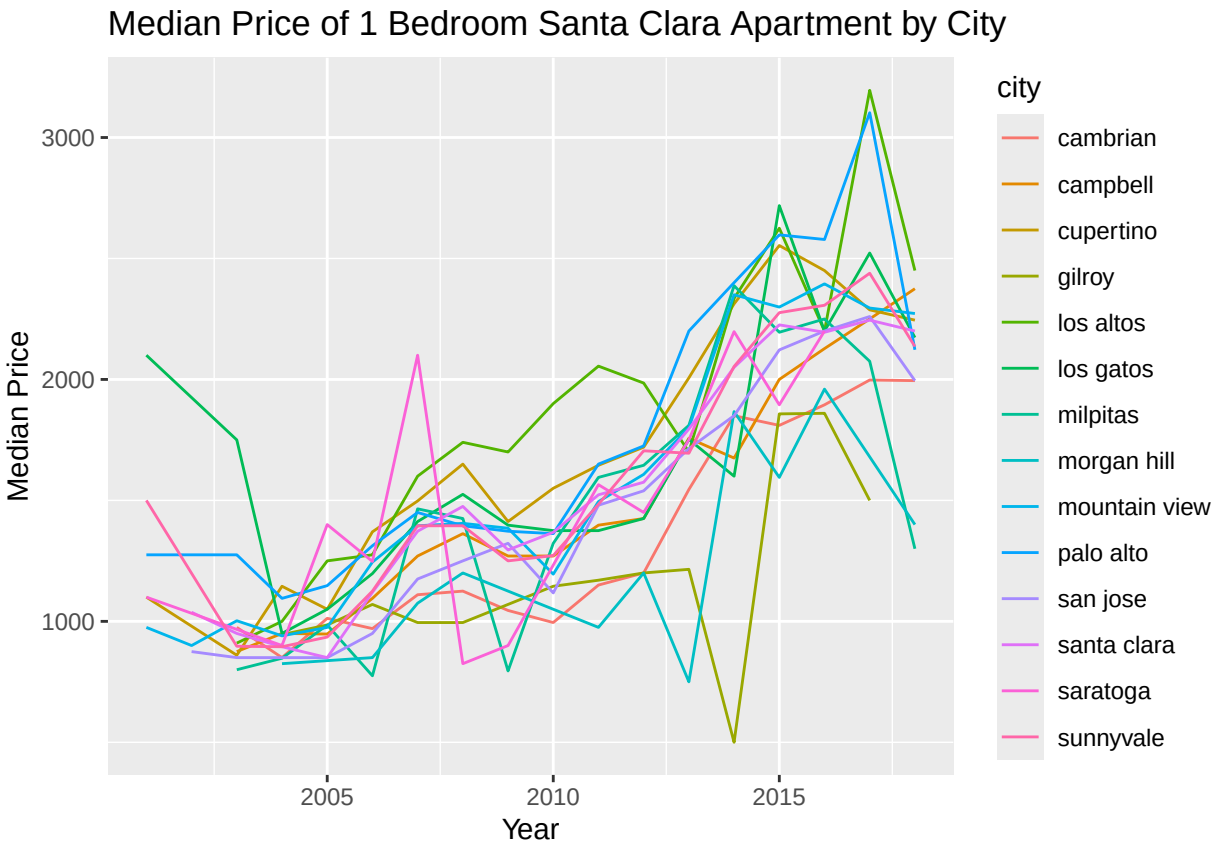
The median price of a one bedroom apartment in Santa Clara county has been rapidly and steadily been increasing since the beginning of the 2000's.

9. Create and interpret a single plot with year on the x-axis and median price for a 1 bedroom apartment on the y-axis, using a separate line for each city in Santa Clara county, for the years 2000 to 2018.

```
sc_1b_by_city <- sc_onebed %>%
  group_by(year, city) %>%
  summarise(median_price = median(price))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
ggplot(data = sc_1b_by_city, aes(x = year, y = median_price, color = city)) +
  geom_line() +
  labs(x = "Year", y = "Median Price",
       title = "Median Price of 1 Bedroom Santa Clara Apartment by City")
```



C. Open ended data visualization

For this part, choose a dataset that interests you and identify a set of questions that you would like to explore via data visualizations. In particular, you should create three visualizations that satisfy the following requirements.

Instructions

- Identify three research questions of interest that you want to study using this dataset.
- For each of your three research questions, generate a data visualization using your dataset. Discuss and interpret your findings.
- Your project should include at least two different types of visualizations (e.g. scatterplots, box plots, bar plots, histograms, line plots, etc.).
- At least one of your plots should display variation over time or location (or both) in some way.
- Each visualization should include a caption that fully explains how to understand your visualization (i.e. explain all the components, legends, etc.). A good guideline is that someone who has not read your report should be able to look at just a visualization and its caption and fully understand what that visualization is showing.
- Each visualization must be accompanied by at least one paragraph of text. This text should include an interpretation of your visualization as well as what is interesting about your visualization. A strong visualization will be accompanied by text explaining what patterns or insights it helps us glean from the data.

Using U.S. Monthly Count of Deaths 2014-2019 Dataset

```
# load dataset
# https://catalog.data.gov/dataset/monthly-counts-of-deaths-by-select-causes-2014-2019-da9df
# monthly_death_count (mdc)
mdc <- read_csv("/Users/rellamas/math_and_algos/R/data/Monthly_Counts_of_Deaths_by_Select_Causes__2014-
```

```
## Rows: 72 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Jurisdiction of Occurrence
## dbl (20): Year, Month, All Cause, Natural Cause, Septicemia, Malignant Neopl...
##
## 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.
```

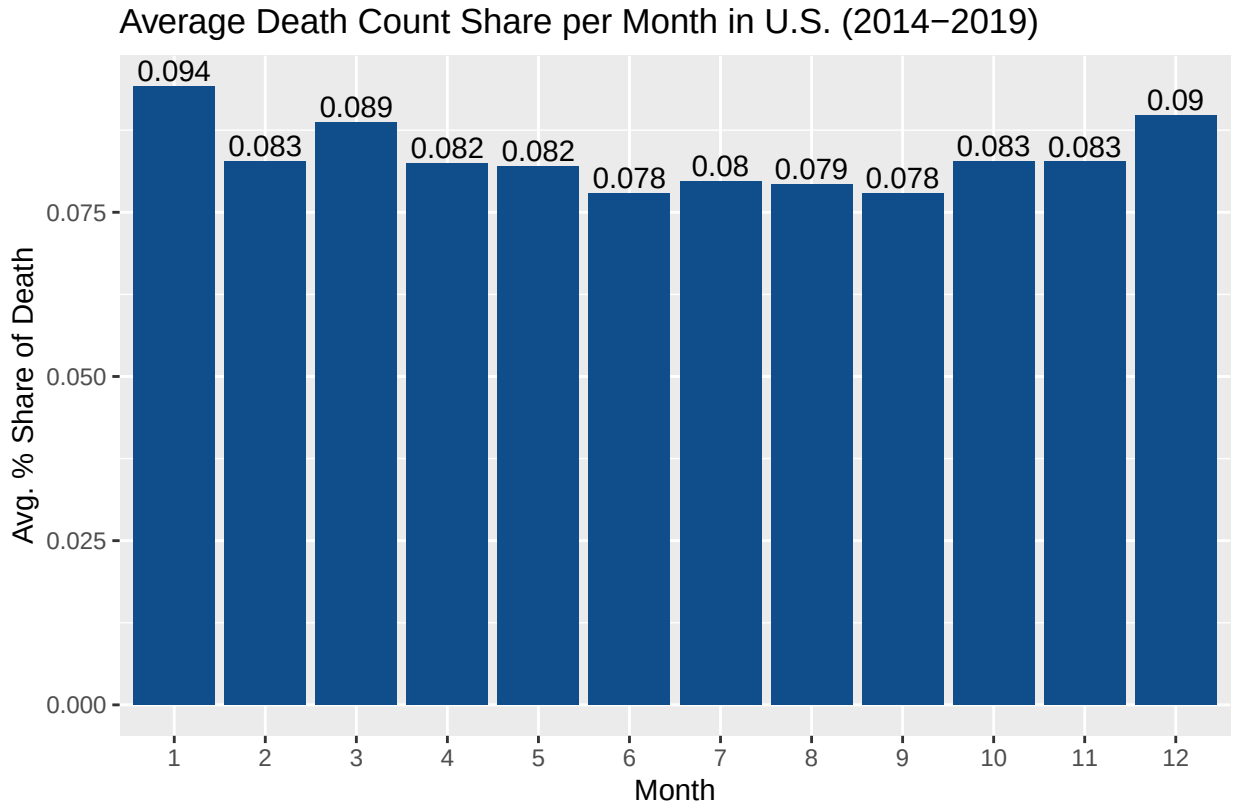
Question 1: Which are the deadliest months of the year? Which are the least deadly?

```
# turn Year and Month cols to factor types
mdc <- mdc %>%
  mutate(Year = factor(Year), Month = factor(Month))

# link all cause of deaths to months
all_vs_month <- aggregate(`All Cause`~Month, data = mdc, mean)

all_cause_sum <- sum(all_vs_month$`All Cause`)

# make bar plot
ggplot(data = all_vs_month,
       aes(x = Month, y = `All Cause` / all_cause_sum)) +
  geom_bar(stat = "identity", fill="dodgerblue4") +
  ylab("Avg. % Share of Death") +
  labs(title = "Average Death Count Share per Month in U.S. (2014-2019)",
         caption = "Noticable increase in deaths in winter months and a decrease in summer months.") +
  geom_text(aes(label = round(`All Cause`/all_cause_sum, digits = 3)), position=position_dodge(width=0.
  # https://intellipaat.com/community/16343/how-to-put-labels-over-geombar-for-each-bar-in-r-with-ggplo
  theme(plot.caption = element_text(hjust = 0.5))
```



Noticable increase in deaths in winter months and a decrease in summer months.

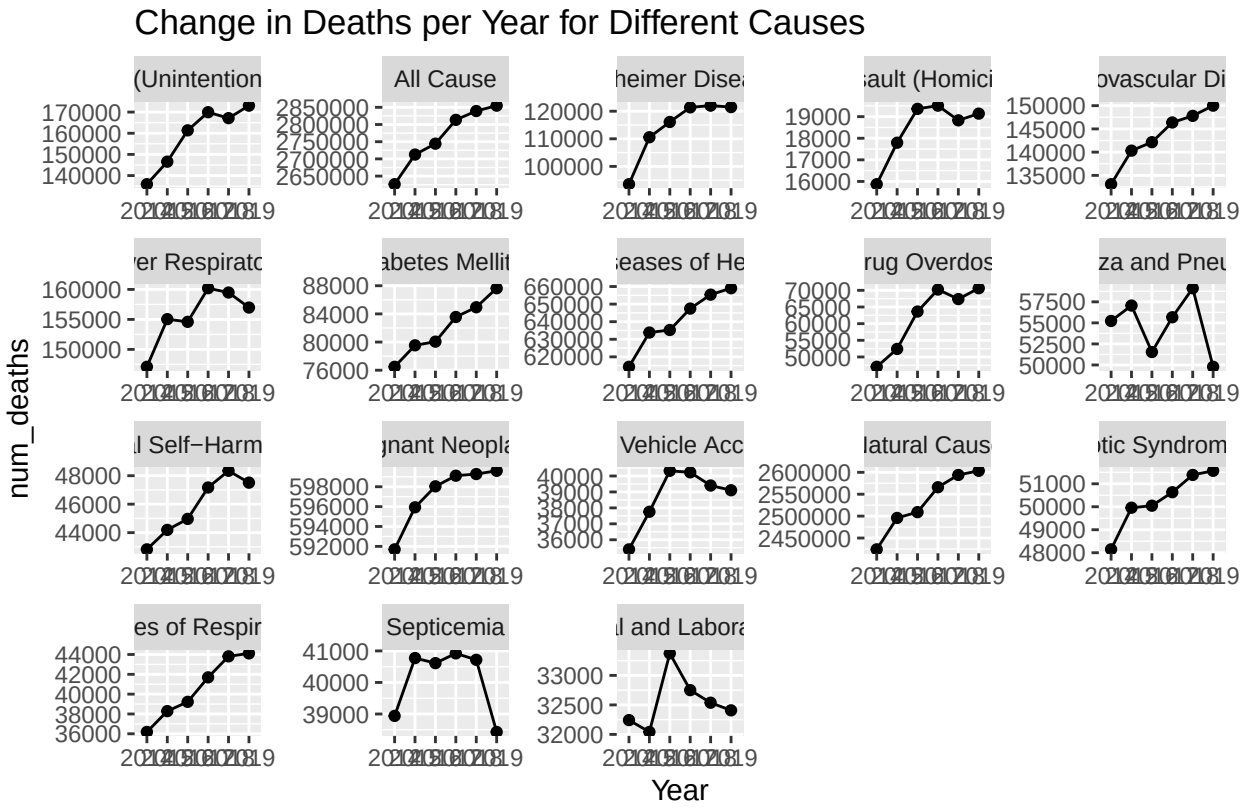
From this graphic we can see that the top 3 deadliest months are January, March, and December. There is a stretch in the year where people die less, which goes from April to September. Then, there is an uptick starting in October and the increased levels of all causes of death continues through March. It makes intuitive sense that the late autumn and winter months, where people are most prone to sickness, have a higher share of deaths during the year.

Question 2: Are there any causes of death that have decreased during the 2014-2019 period? Are there any that increased?

```
# make new df where all numeric columns (causes of death) are summed by Year
ydc <- mdc %>%
  group_by(Year) %>%
  summarise(across(where(is.numeric), sum))
# turn all other cols, except Year, are put into single factor col, and their numeric values are in own
ydc_reshape <- pivot_longer(ydc,
                        cols = -Year,      # all cols except "Year"
                        names_to = "cause_of_death",
                        values_to = "num_deaths")
# convert "cause_of_death" col to factor
ydc_reshape <- ydc_reshape %>%
  mutate(cause_of_death = factor(cause_of_death))

ggplot(data = ydc_reshape, aes(x = Year, y = num_deaths, group = 1)) +
  geom_line() +
```

```
  geom_point() +
  labs(title = "Change in Deaths per Year for Different Causes",
       caption = "Most causes in death saw a relatively steady increase throughout the 5 year observatio
  theme(plot.caption = element_text(hjust = 0.5)) +
  facet_wrap(~ cause_of_death, scales = "free")
```



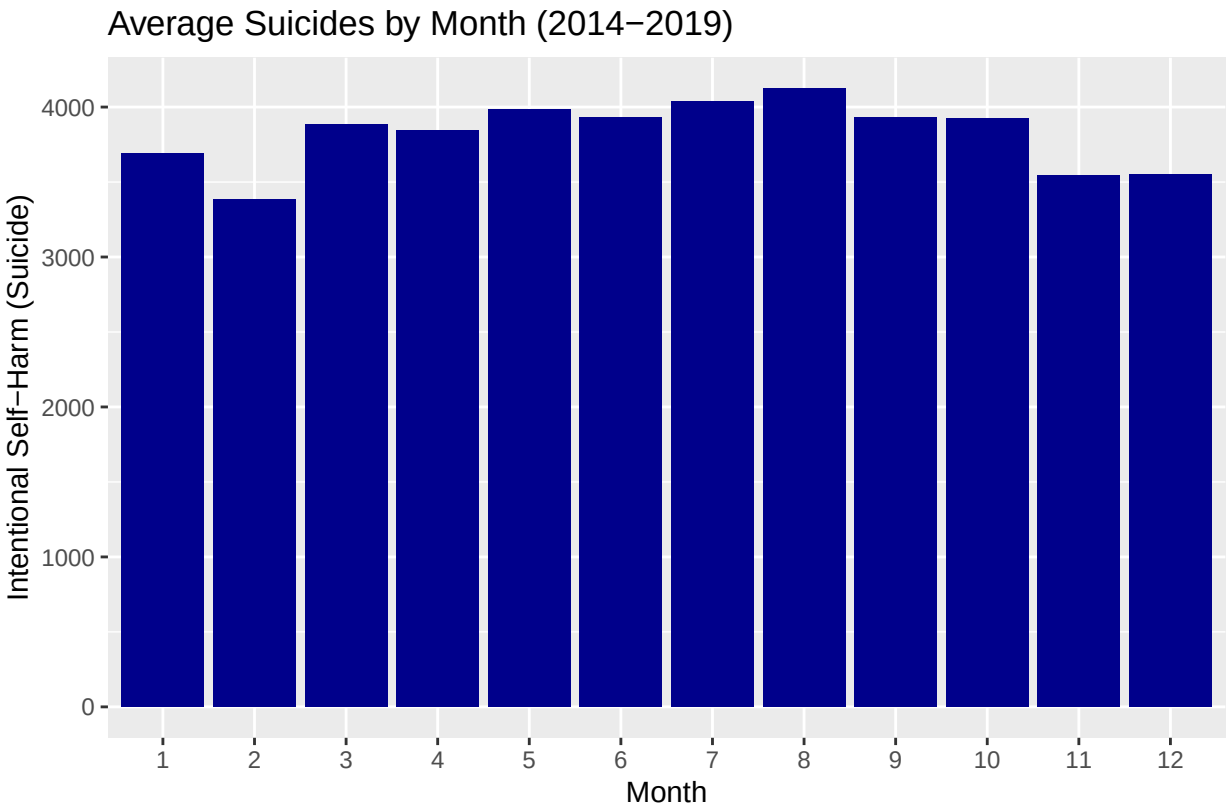
Most causes in death saw a relatively steady increase throughout the 5 year observation period.

Most of the causes of death had a relatively steady uptrend during the 5 year period of this dataset. The only exceptions being septicemia, abnormal clinical and laboratory findings, and influenza and pneumonia. It would be more interesting if the numbers were adjusted by the share of the population, which, I presume, is also steadily growing in the U.S.

Question 3: How much does seasonal depression affect self harm, drug overdose, and homicides?

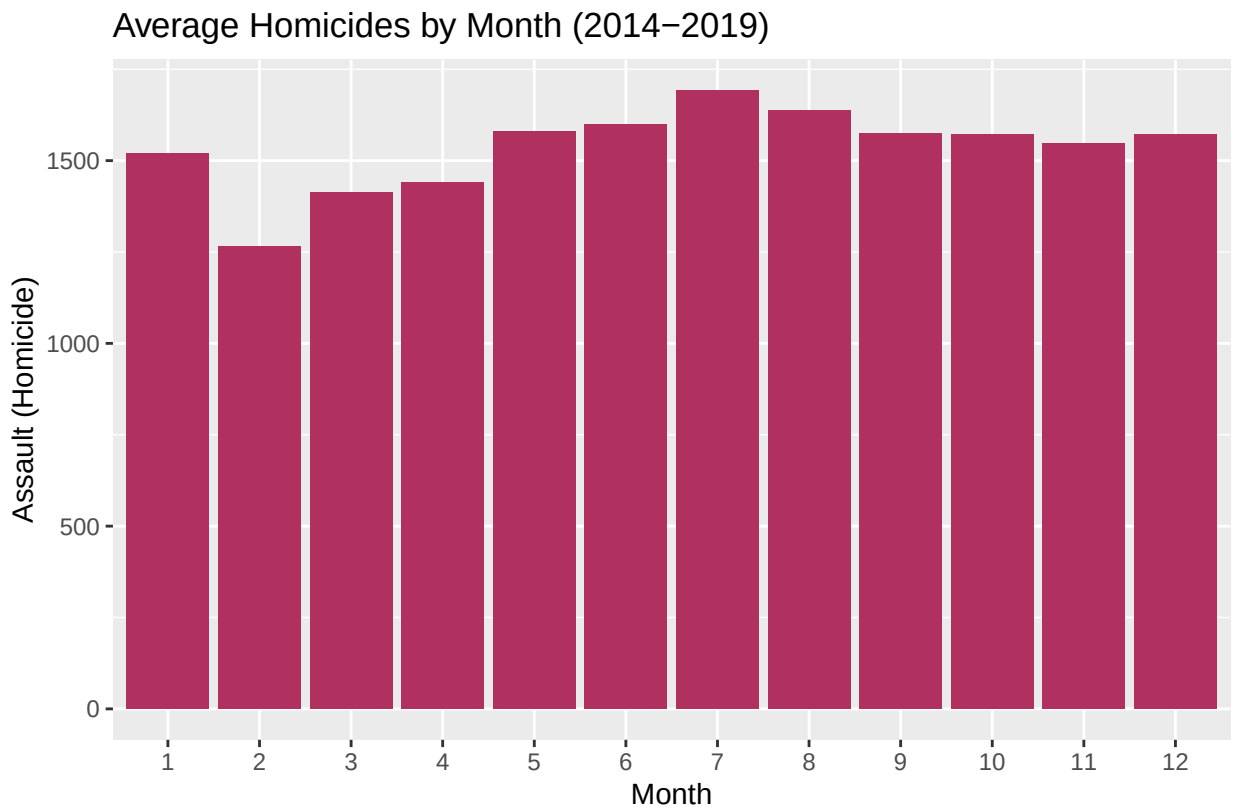
```
selfharm_month <- aggregate(`Intentional Self-Harm (Suicide)`~Month, data = mdc, mean)
homicide_month <- aggregate(`Assault (Homicide)`~Month, data = mdc, mean)
od_month <- aggregate(`Drug Overdose`~Month, data = mdc, mean)

ggplot(data = selfharm_month,
       aes(x = Month, y = `Intentional Self-Harm (Suicide)`)) +
  geom_bar(stat = "identity", fill="darkblue") +
  labs(title = "Average Suicides by Month (2014-2019)",
       caption = "There are more suicides in the spring and summer seasons.") +
  theme(plot.caption = element_text(hjust = 0.5))
```



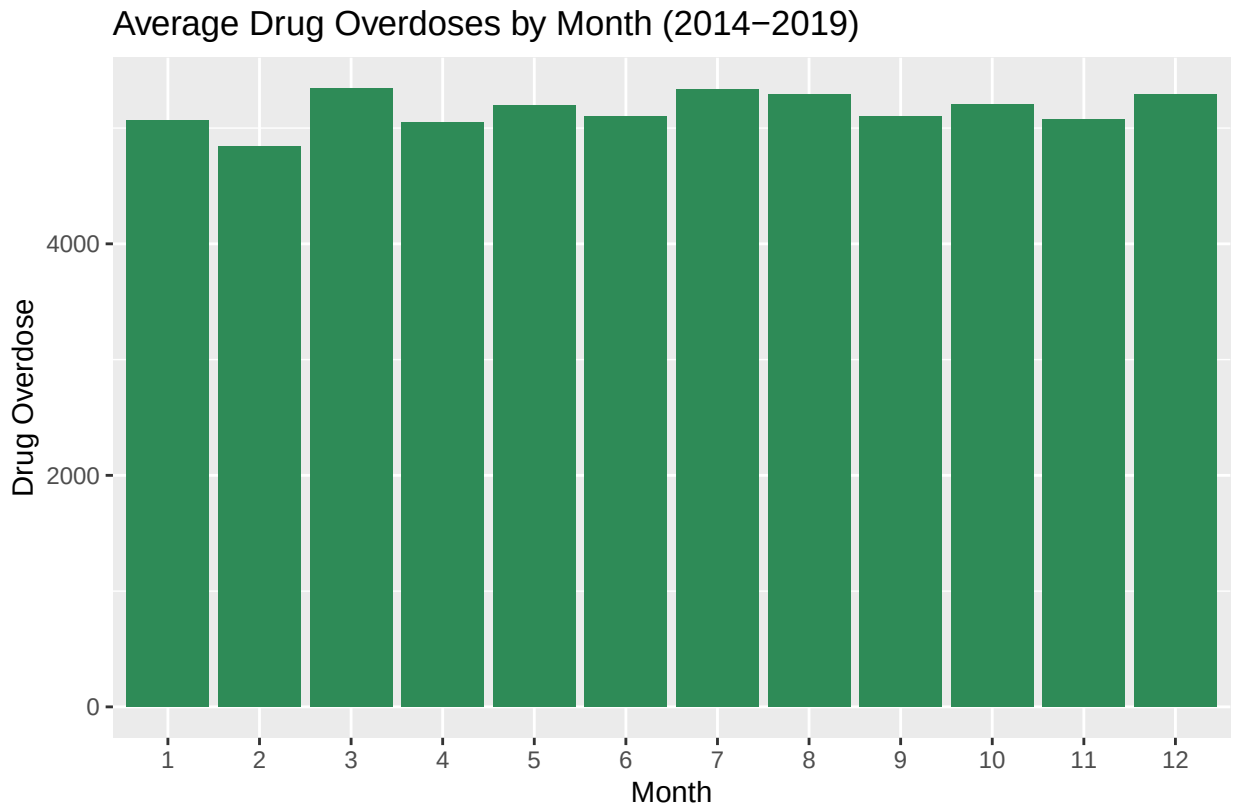
There are more suicides in the spring and summer seasons.

```
ggplot(data = homicide_month,
       aes(x = Month, y = `Assault (Homicide)`)) +
  geom_bar(stat = "identity", fill="maroon") +
  labs(title = "Average Homicides by Month (2014-2019)",
       caption = "There is a steady increase in the number of homicides starting in February and cappin
  theme(plot.caption = element_text(hjust = 0.5))
```



There is a steady increase in the number of homicides starting in February and capping out in July.

```
ggplot(data = od_month,
       aes(x = Month, y = `Drug Overdose`)) +
  geom_bar(stat = "identity", fill="seagreen4") +
  labs(title = "Average Drug Overdoses by Month (2014-2019)",
       caption = "Average number of drug overdoses is steady throughout the year.") +
  theme(plot.caption = element_text(hjust = 0.5))
```



Average number of drug overdoses is steady throughout the year.

These 3 causes did not follow my assumption that there would be an increase in self-harming and violent behavior in the winter. I assumed that seasonal depression would drive more people to destructive behavior. A possible explanation for the increase in homicides in the summer is that there are more people going out than in the colder months, so there are more opportunities for violence between people. There is also always a dip in February for every cause of death. I assume this is because February only has 28 days per month where the other months have 30 or 31 days.