Lab 4: Data Visualization

Code

AUTHOR
Maxim Dokukin

PUBLISHED
February 20, 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

```
library(tidyverse)
library(ggplot2)
```

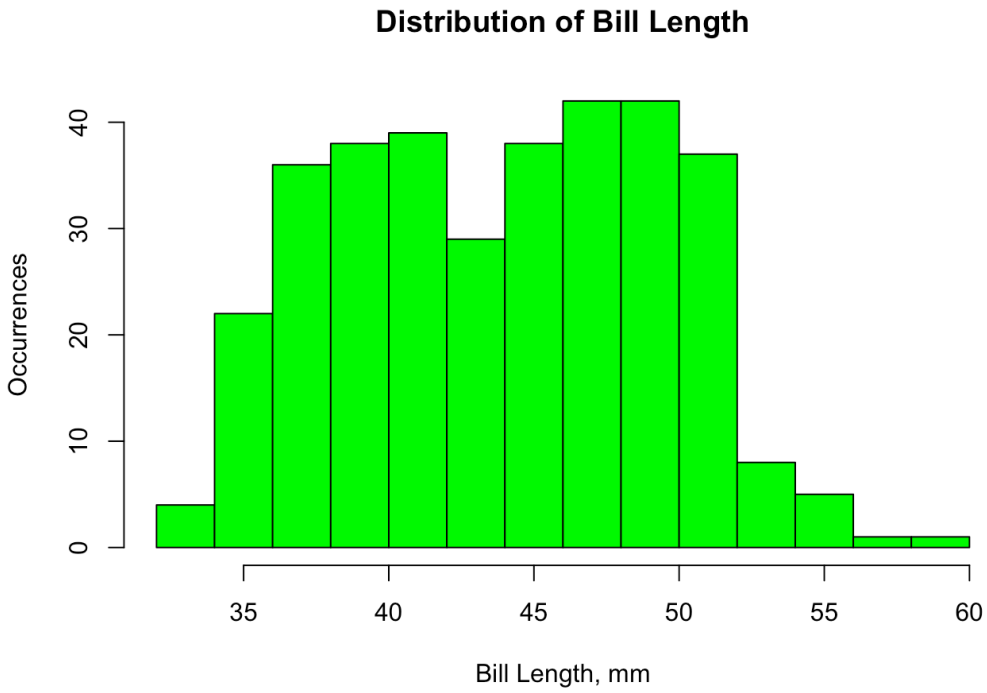
A. Basic visualizations

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.

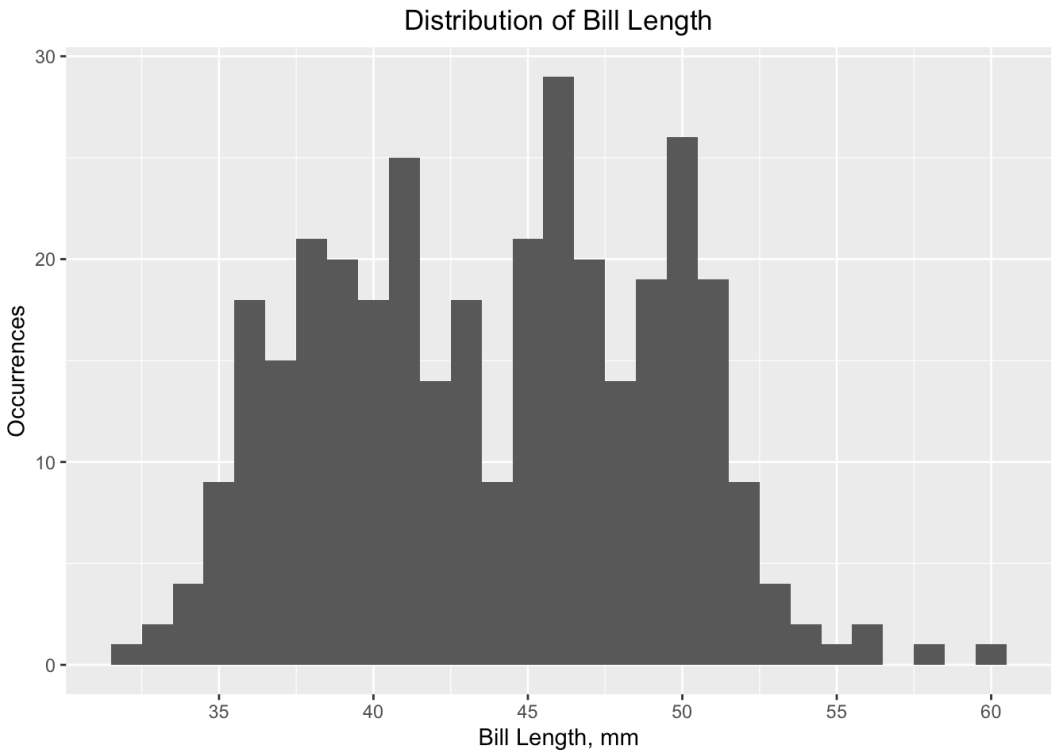
```
hist(
  penguins$bill_length_mm,
  main = 'Distribution of Bill Length',
  xlab = 'Bill Length, mm',
  ylab = 'Occurrences',
  col = 'green'
)
```



47-50mm is the most common occurrence of Bill Length.

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

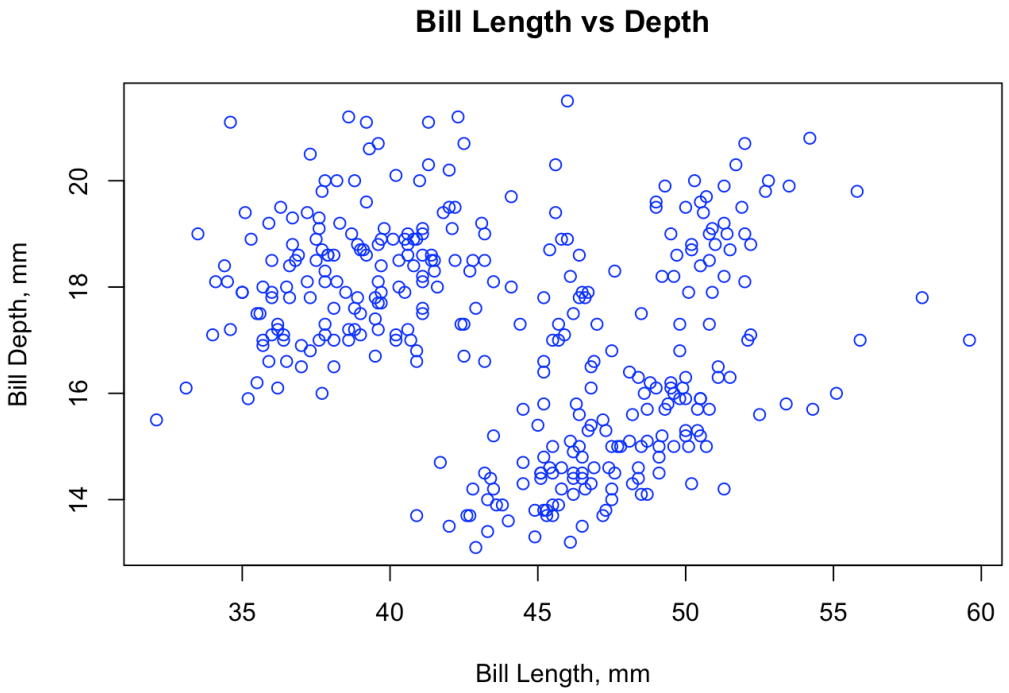
```
ggplot(data = penguins,
       aes(x = bill_length_mm)) +
  geom_histogram(binwidth=1) +
  xlab('Bill Length, mm') +
  ylab('Occurrences') +
  ggtitle('Distribution of Bill Length') +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks = seq(30, 60, by = 5))
```



Looks like bimodal distribution. 45-47mm is the most common occurence of Bill Length.

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.

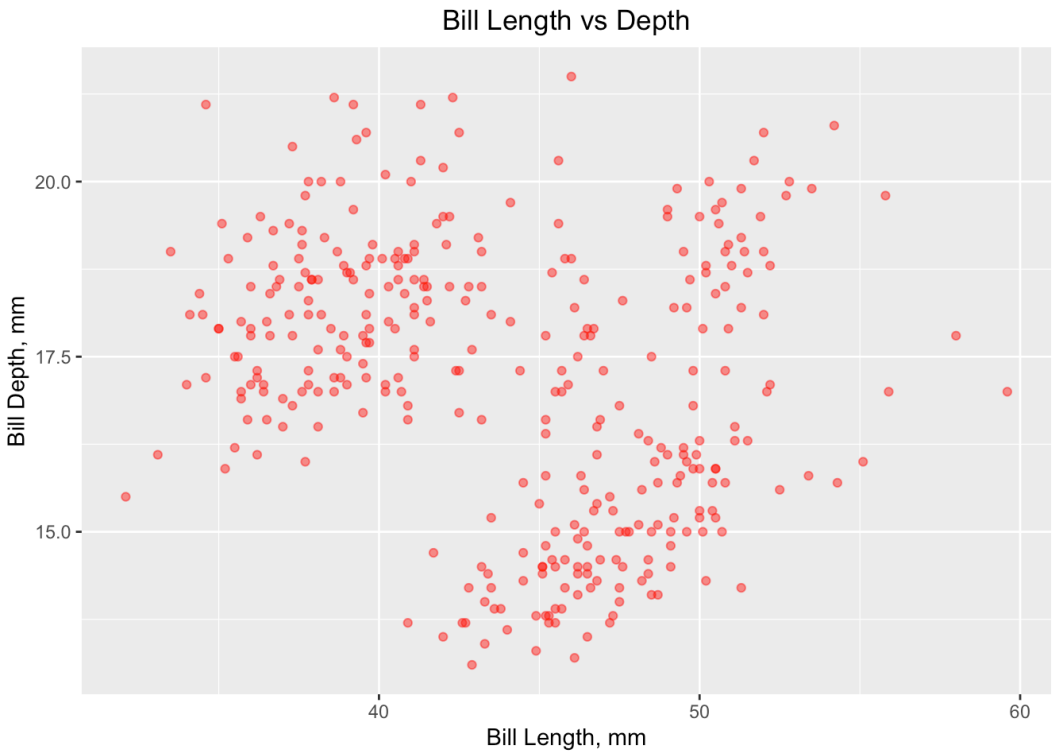
```
plot(
  x = penguins$bill_length_mm,
  y = penguins$bill_depth_mm,
  main = 'Bill Length vs Depth',
  xlab = 'Bill Length, mm',
  ylab = 'Bill Depth, mm',
  col = 'blue',
)
```



Data seems to be forming 3 clusters.

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.

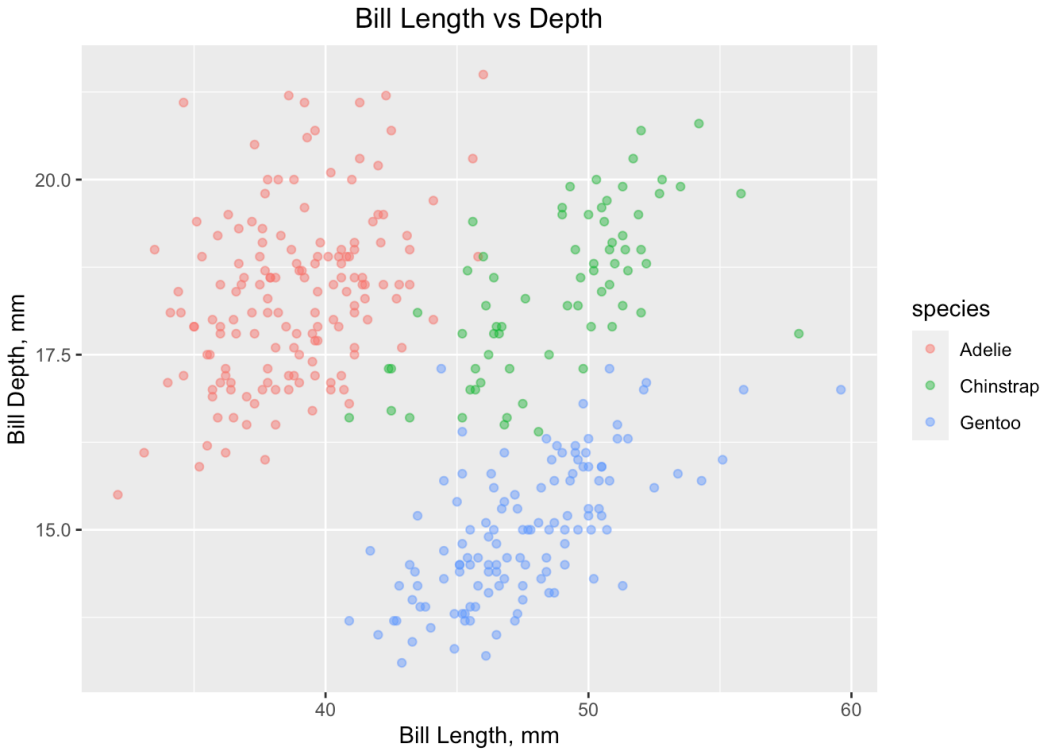
```
ggplot(data = penguins,
       aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point(color = 'red', alpha = 0.5)  +
  xlab('Bill Length, mm') +
  ylab('Bill Depth, mm') +
  ggtitle('Bill Length vs Depth')+
  theme(plot.title = element_text(hjust = 0.5))
```



Data seems to be forming 3 clusters.

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 = penguins,
       aes(x = bill_length_mm, y = bill_depth_mm, color=species
  geom_point(alpha = 0.5)  +
  xlab('Bill Length, mm') +
  ylab('Bill Depth, mm') +
  ggtitle('Bill Length vs Depth')+
  theme(plot.title = element_text(hjust = 0.5))
```



Data indeed forms 3 clusters, characterized by species.

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.

```
rent <- readr::read_csv('https://raw.githubusercontent.com/rforc
summary(rent)
```

```
head(rent)
```

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

```
sc_onebed <- rent |> filter(beds == 1, county == 'santa clara')
print(paste(nrow(sc_onebed), 'one bedroom listings in Santa Cla
```

```
[1] "12455 one bedroom listings in Santa Clara county"
```

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

```
sc_onebed_2018 <- rent |> filter(beds == 1, county == 'santa cl
print(paste('Median price for a 1 bedroom listing in Santa Clar
            median(sc_onebed_2018$price)))
```

```
[1] "Median price for a 1 bedroom listing in Santa Clara county
in 2018 is $ 2095"
```

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

```
county_medians <- rent |> filter(beds == 1, year == 2018) |>
                          group_by(county) |>
                          summarize(median_price = median(price

most_expesive_county <- county_medians[which.max(county_medians$

print(paste('Most expensive county:', most_expesive_county[, 'co
```

```
[1] "Most expensive county: san francisco"
```

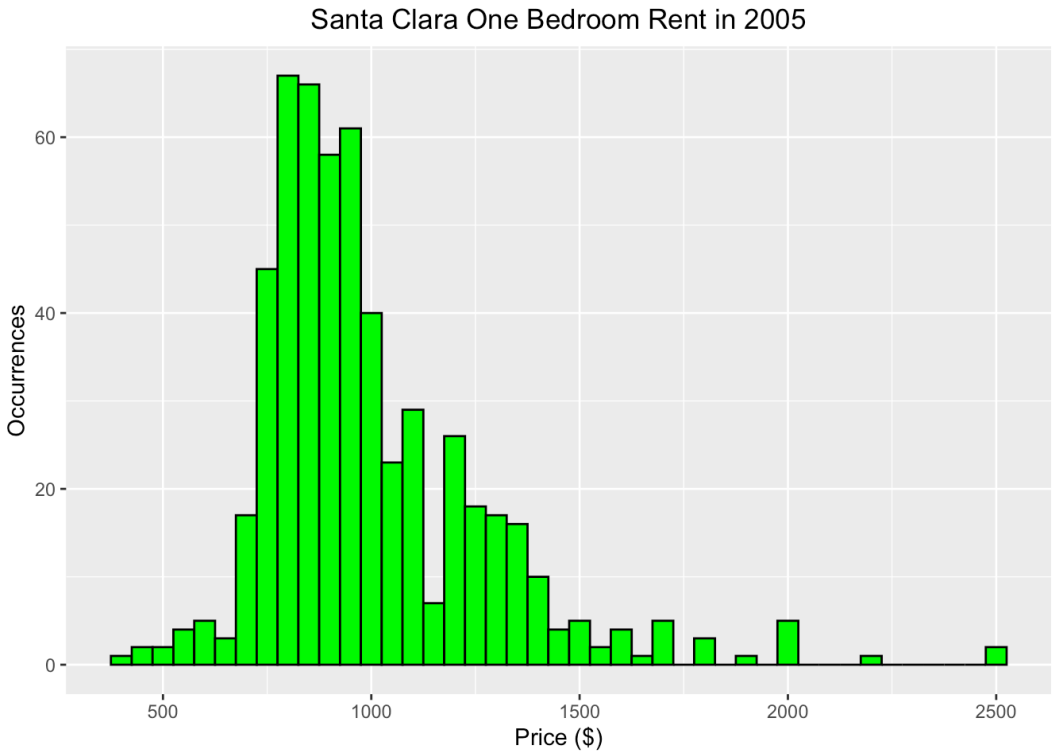
```
print(paste('With median price for a 1 bedroom in 2018: $',
            most_expesive_county[, 'median_price']))
```

```
[1] "With median price for a 1 bedroom in 2018: $ 3000"
```

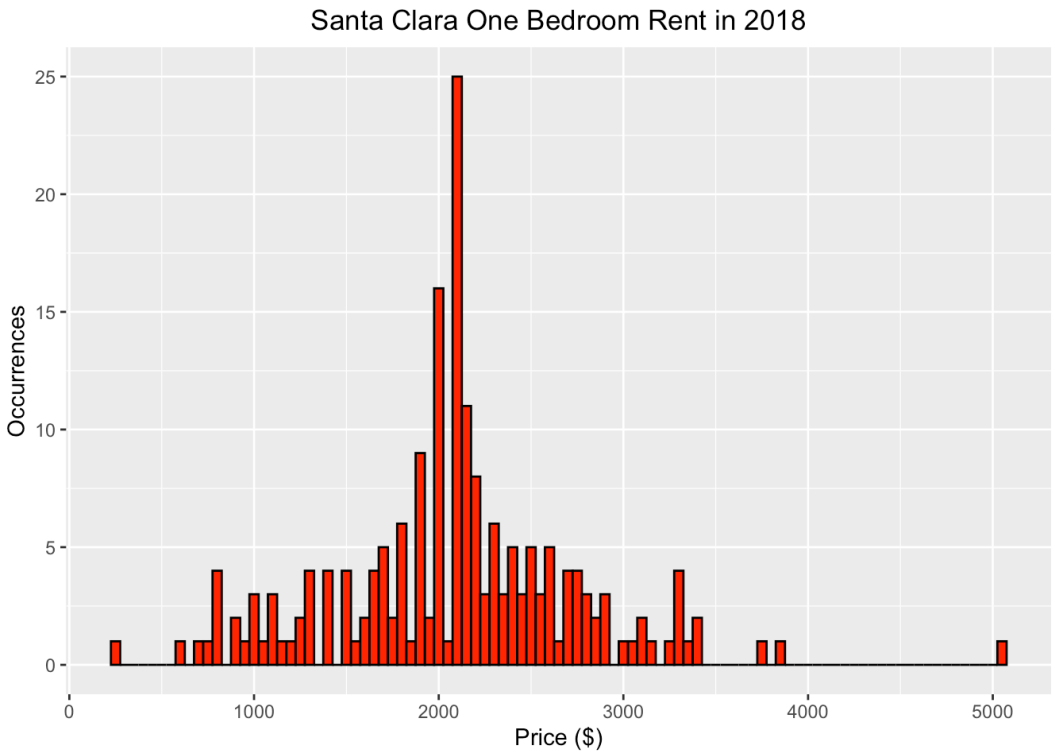
9. Create two histograms for the prices of 1 bedroom listings in Santa Clara county in 2005 and 2018. Compare and discuss.

```
sc_onebed_2005 <- rent |> filter(county == 'santa clara', beds =
sc_onebed_2018 <- rent |> filter(county == 'santa clara', beds =
```

```
ggplot(data = sc_onebed_2005,
       aes(x = price)) +
       geom_histogram(binwidth=50, fill='green', color = 'black
       xlab('Price ($)') +
       ylab('Occurrences') +
       ggtitle('Santa Clara One Bedroom Rent in 2005') +
       theme(plot.title = element_text(hjust = 0.5))
```



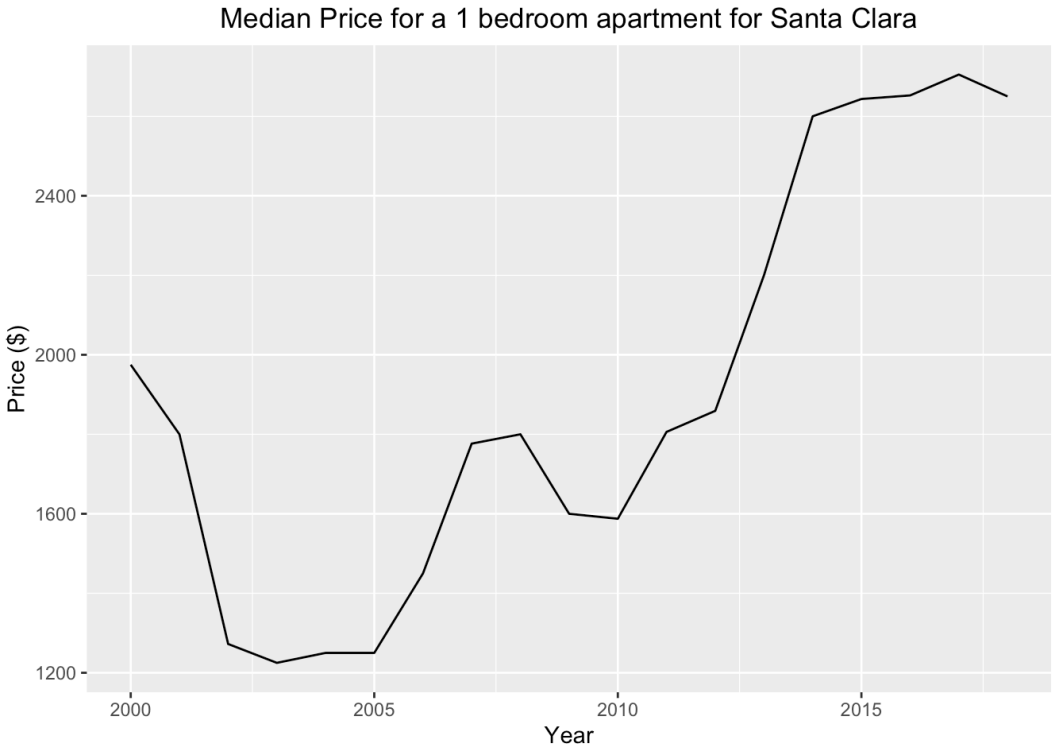

```
ggplot(data = sc_onebed_2018,
       aes(x = price)) +
       geom_histogram(binwidth=50, fill='red', color = 'black')
       xlab('Price ($)') +
       ylab('Occurrences') +
       ggtitle('Santa Clara One Bedroom Rent in 2018') +
       theme(plot.title = element_text(hjust = 0.5))
```



Average price shifted to the right from 2005 to 2008. It went up from $900 to $2000. Also, the price range is more wide, up to 2500 in 2005 and up to $5000 in 2018. It also seems like there is fewer listings.

10. 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.

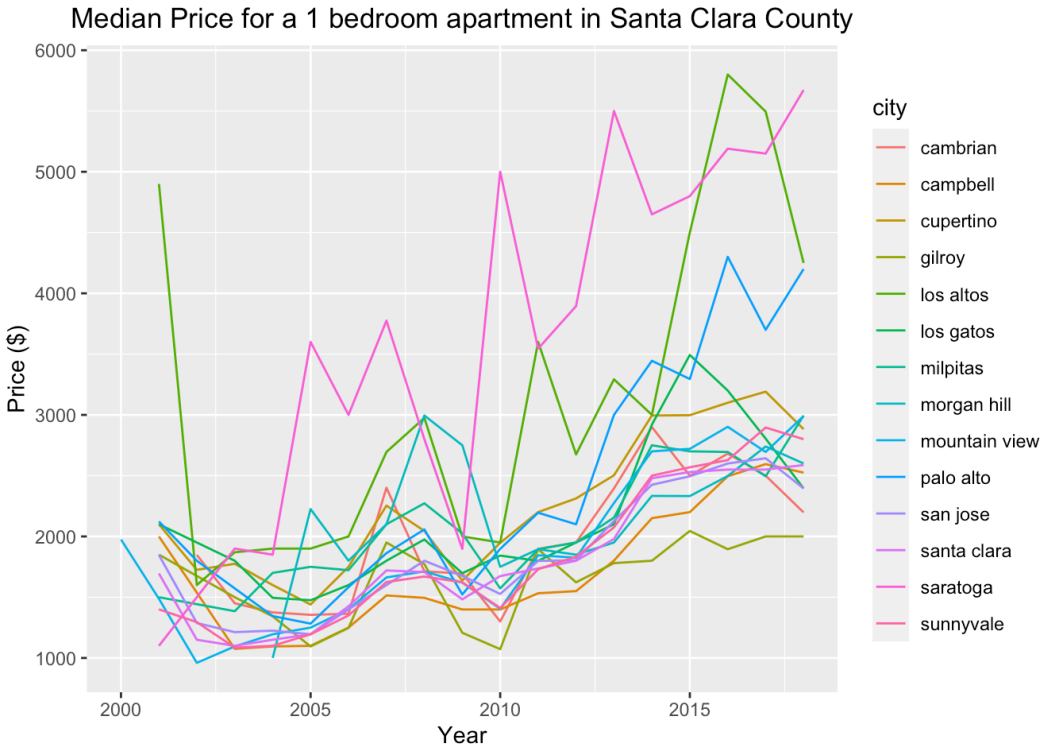
```
sc_median_byyears <- rent |> filter(county == 'santa clara',
                                    year >= 2000,
                                    year <= 2018) |>
                             group_by(year) |>
                             summarize(median_price = median(pr
ggplot(sc_median_byyears,
       aes(x = year, y = median_price)) +
       geom_line() +
       labs(title = 'Median Price for a 1 bedroom apartment for
            x = 'Year',
            y = 'Price ($)') +
       theme(plot.title = element_text(hjust = 0.5))
```



Average prices dipped in 2001-2005, but luckly grew significantly for the following years.

11. 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_median_byyears <- rent |> filter(county == 'santa clara',
                                    year >= 2000,
                                    year <= 2018) |>
                            group_by(city, year) |>
                            summarize(median_price = median(pr

ggplot(sc_median_byyears,
       aes(x = year, y = median_price, color = city)) +
       geom_line() +
       labs(title = 'Median Price for a 1 bedroom apartment in S
            x = 'Year',
            y = 'Price ($)') +
       theme(plot.title = element_text(hjust = 0.5))
```



There is too many colors, its hard to tell. Overall trend is growth. Gilroy is the cheapest city over the years. Saratoga is consistently expensive. That city along with Los Gatos have the highest amplitude of price fluctuations over the years.

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.).

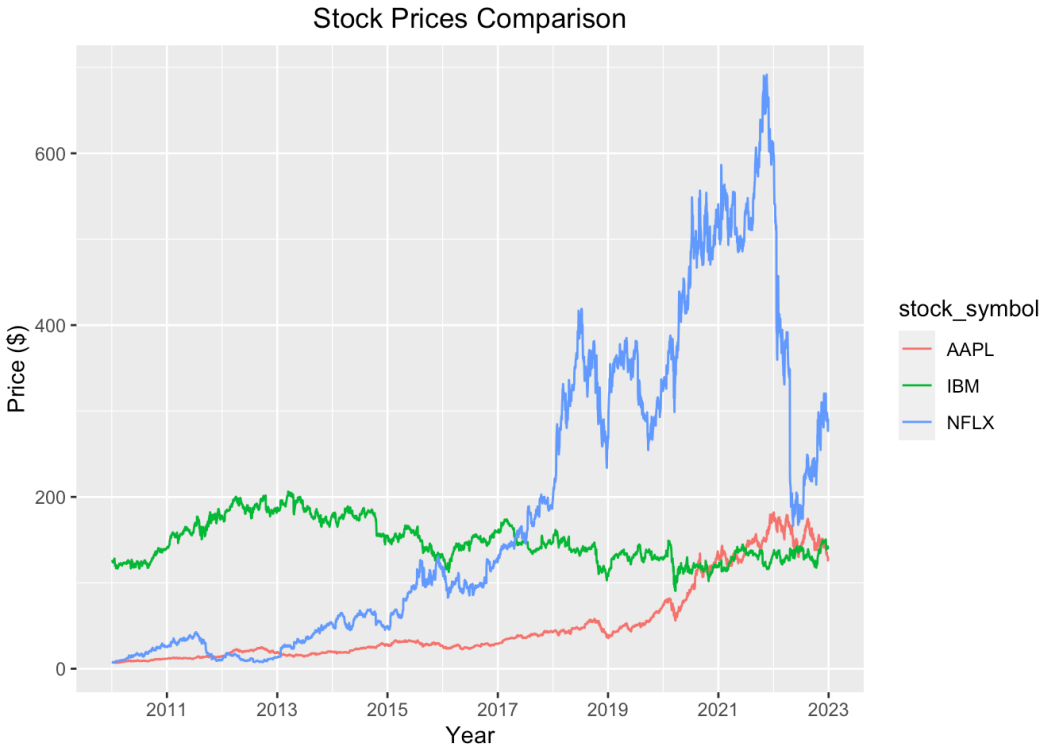
- At least one of your plots should display variation over time or location (or both) in some way.
- Each visualization should include a caption that fully explains how to understand your visualization (i.e. explain all the components, legends, etc.). A good guideline is that someone who has not read 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.

```
big_tech_stock_prices <- readr::read_csv('https://raw.githubuse

summary(big_tech_stock_prices)
```

```
head(big_tech_stock_prices)
```

1. Compare-analyze Apple, Netflix, and IBM overall stock performances during 2010-2023.

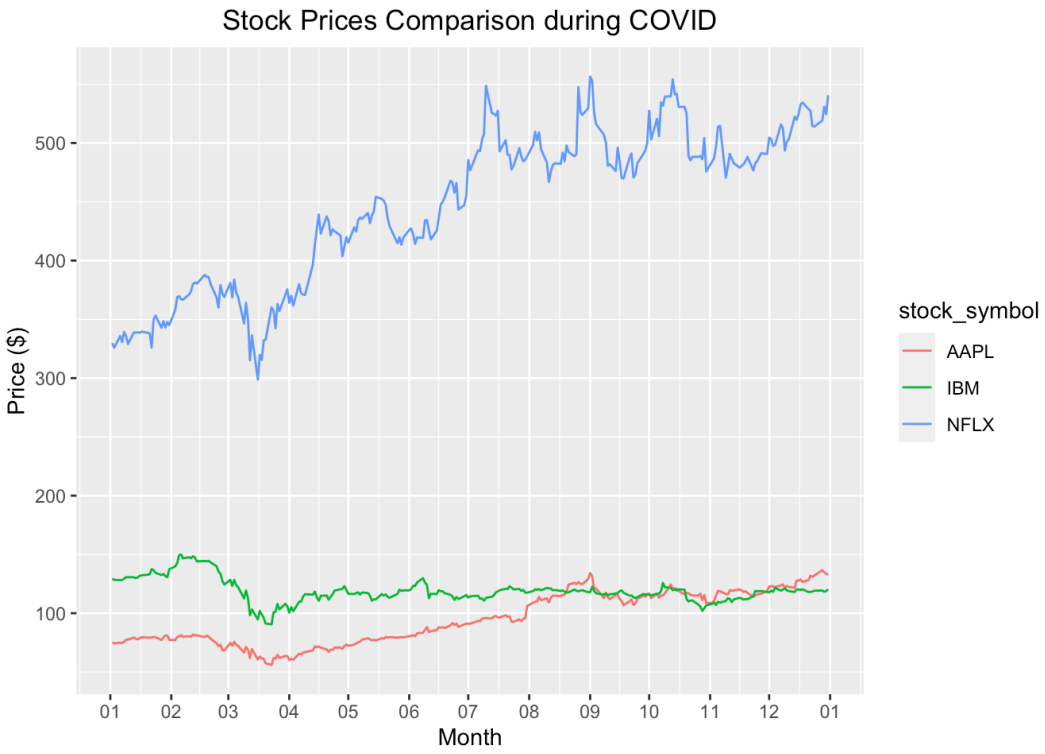
```
AAPL_IBM_NFLX_2010_2023 <- big_tech_stock_prices |> filter(stocl

ggplot(AAPL_IBM_NFLX_2010_2023,
       aes(x = date, y = close, color = stock_symbol)) +
       geom_line() +
       scale_x_date(date_breaks = '2 years', date_labels = '%Y'
       labs(title = 'Stock Prices Comparison',
            x = 'Year',
            y = 'Price ($)') +
       theme(plot.title = element_text(hjust = 0.5))
```



The graph displays a comparison of stock prices for Apple (AAPL), IBM, and Netflix (NFLX) over a period ranging from before 2011 to beyond 2023. AAPL's stock shows a significant upward trend with substantial growth. In contrast, IBM's stock price appears relatively stable with less volatility, while NFLX shows a more volatile pattern with a steep increase until it peaks and then a sharp decline towards the end of the period.

2. Compare-analyze how Apple, Netflix, and IBM during COVID crisis.

```
AAPL_IBM_NFLX_covid <- AAPL_IBM_NFLX_2010_2023 |>
                          filter(date >= as.Date('2020-01-01')

ggplot(AAPL_IBM_NFLX_covid,
       aes(x = date, y = close, color = stock_symbol)) +
       geom_line() +
       scale_x_date(date_breaks = '1 month', date_labels = '%m'
       labs(title = 'Stock Prices Comparison during COVID',
            x = 'Month',
            y = 'Price ($)') +
       theme(plot.title = element_text(hjust = 0.5))
```

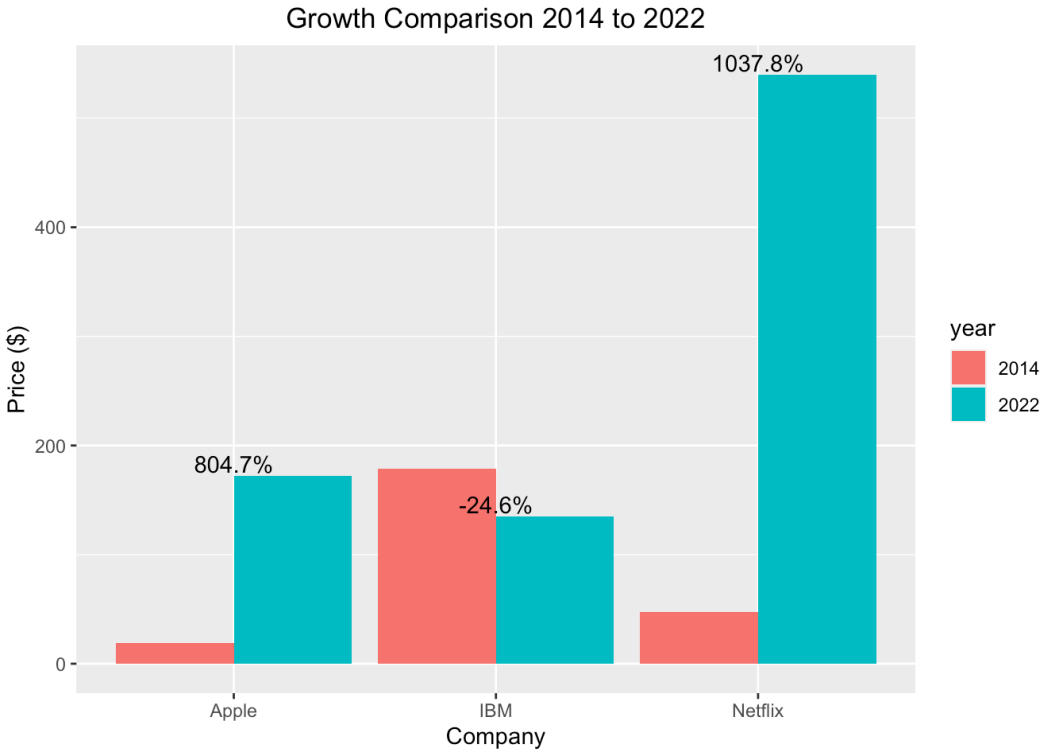


There is visible decline for all 3 companies in March-April 2020. Yet it is not as significant as I expected. Overall, Apple and Netflix grew, while IBM declined.

3. Establish which company performed best Jan 2014 to Jan 2022.

```
companies <- c('Apple', 'IBM', 'Netflix')
prices_2014 <- AAPL_IBM_NFLX_2010_2023[AAPL_IBM_NFLX_2010_2023$c
prices_2022 <- AAPL_IBM_NFLX_2010_2023[AAPL_IBM_NFLX_2010_2023$c
percent_changes <- (prices_2022$close -
                    prices_2014$close) /
                    prices_2014$close * 100

prices_change <- data.frame(
  company = companies,
  price_2014 = prices_2014$close,
  price_2022 = prices_2022$close,
  percent_change = percent_changes
)

prices_change <- prices_change |> pivot_longer(
                                      cols = starts_with('price
                                      names_to = 'year',
                                      names_prefix = 'price_',
```

```
                                    values_to = 'price')

ggplot(prices_change, aes(x = company, y = price, fill = year)) +
  geom_bar(stat = 'identity', position = position_dodge()) +

  geom_text(data = subset(prices_change, year == '2022'),
            aes(label = sprintf("%.1f%%", percent_change)),
            position = position_dodge(width = 0.9), vjust = -0.2

  labs(title = 'Growth Comparison 2014 to 2022', x = 'Company',
  theme(plot.title = element_text(hjust = 0.5))
```



The bar chart illustrates the growth comparison of stock prices between 2014 and 2022 for Apple, IBM, and Netflix. Apple's stock price shows an 804.7% increase over the period, while Netflix's stock price exhibits an even higher growth of 1037.8%. In contrast, IBM's stock price experienced a decline, with a -24.6% change, indicating a decrease in stock value over the same timeframe.