# Challenge_4: Intro to Visulization: Univariate and Multivariate Graphs

AUTHOR
Muskan Dhar

PUBLISHED
June 28, 2023

**Make sure you change the author's name in the above YAML header.**

## Setup

If you have not installed the following packages, please install them before loading them.

```
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   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ 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(readxl)
library(haven) #for loading other datafiles (SAS, STATA, SPSS, etc.)
library(stringr) # if you have not installed this package, please install it.
library(ggplot2) # if you have not installed this package, please install it.
```

## Challenge Overview

In this challenge, we will practice with the data we worked on in the previous challenges and the data you choose to do some simple data visualizations using the `ggplot2` package.

There will be coding components and writing components. Please read the instructions for each part and complete your challenges.

## Datasets

- Part 1 the ESS_Polity Data (created in Challenge#3) ⭐⭐
- Part 2: the Australia Data (from Challenge#2) ⭐⭐
- Part 3: see [Part 3. Practice plotting with a dataset of your choice (25%)]. For online platforms of free data, see Appendix: sources for data to be used in Part 3.

Find the `_data` folder, then read the datasets using the correct R command.

## Part 1. Univariate and Multivariate Graphs (45%)

We have been working with these two data in the previous three challenges. Suppose we have a research project that studies European citizens' social behaviors and public opinions, and we are interested in how the countries that respondents live in influence their behavior and opinion. In this challenge, let's work with the combined dataset *ESS_Polity* and create some visualizations.

1. **Read the combined data you created last time. (2.5%)**

```
ESS_polity<- read.csv("ESS_Polity.csv")
head(ESS_polity)
```

| | idno | year.x | male | age | edu | eth_major | income_10 | cntry | vote |
|---|---|---|---|---|---|---|---|---|---|
| | <int> | <int> | <int> | <int> | <int> | <int> | <int> | <chr> | <int> |
| 1 | 15906 | 2010 | 0 | 14 | 1 | 1 | 2 | GR | 3 |
| 2 | 21168 | 2010 | 0 | 14 | 1 | 1 | 2 | IE | 3 |
| 3 | 40 | 2010 | 0 | 14 | 1 | NA | 8 | LT | 3 |
| 4 | 2108 | 2010 | 0 | 14 | 1 | 1 | NA | RU | 3 |
| 5 | 519 | 2010 | 0 | 14 | 1 | 1 | NA | IL | 2 |
| 6 | 2304 | 2010 | 0 | 14 | 1 | 1 | NA | ES | 3 |

6 rows | 1-10 of 20 columns

2. **Suppose we are interested in the central tendencies and distributions of the following variables. At the individual level:** *age, male, edu, income_10,* and *vote*. **At the country level**: *democ*.

   (1) Recode the "vote" column: if the value is 1, recode it as 1; if the value is 2, recode it as 0; if the value is 3, recode it as NA. **Make sure to include a sanity check for the recoded data. (2.5%)**

```
ESS_polity$vote <- ifelse(ESS_polity$vote == 1, 1,
                   ifelse(ESS_polity$vote == 2, 0,
                   ifelse(ESS_polity$vote == 3, NA, ESS_polity$vote)))

unique(ESS_polity$vote)
```

```
[1] NA  0  1
```

\(2\) For each of the five variables (*age, edu, income_10,* *vote,* and *democ)*, please choose an appropriate type of univariate graph to plot the central tendencies and distribution of the variables. Explain why you choose this type of graph to present a particular variable (for example: "I use a histogram to plot *age* because it is a continuous numeric variable"). **(25%)**

**(Note: You should use at least two types of univariate graphs covered in the lecture.)**

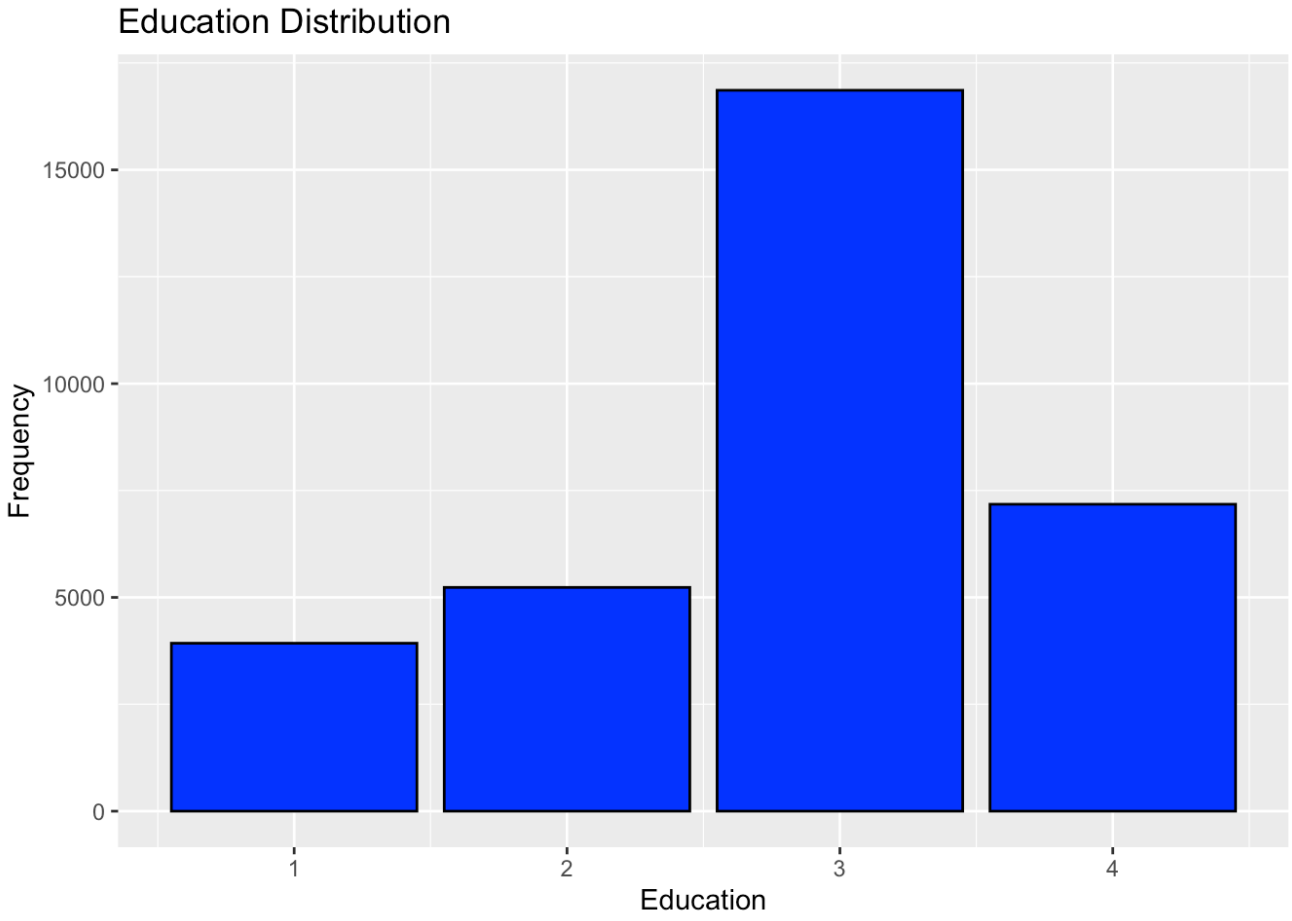
```
::: {.cell}

```{.r .cell-code}
ggplot(ESS_polity, aes(x = age)) +
  geom_histogram(binwidth = 5, color = "black",fill ="pink",na.rm = TRUE) +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```

::: {.cell-output-display}
![](challenge_4_Spring24_files/figure-html/unnamed-chunk-4-1.png){width=672}
:::
:::
```

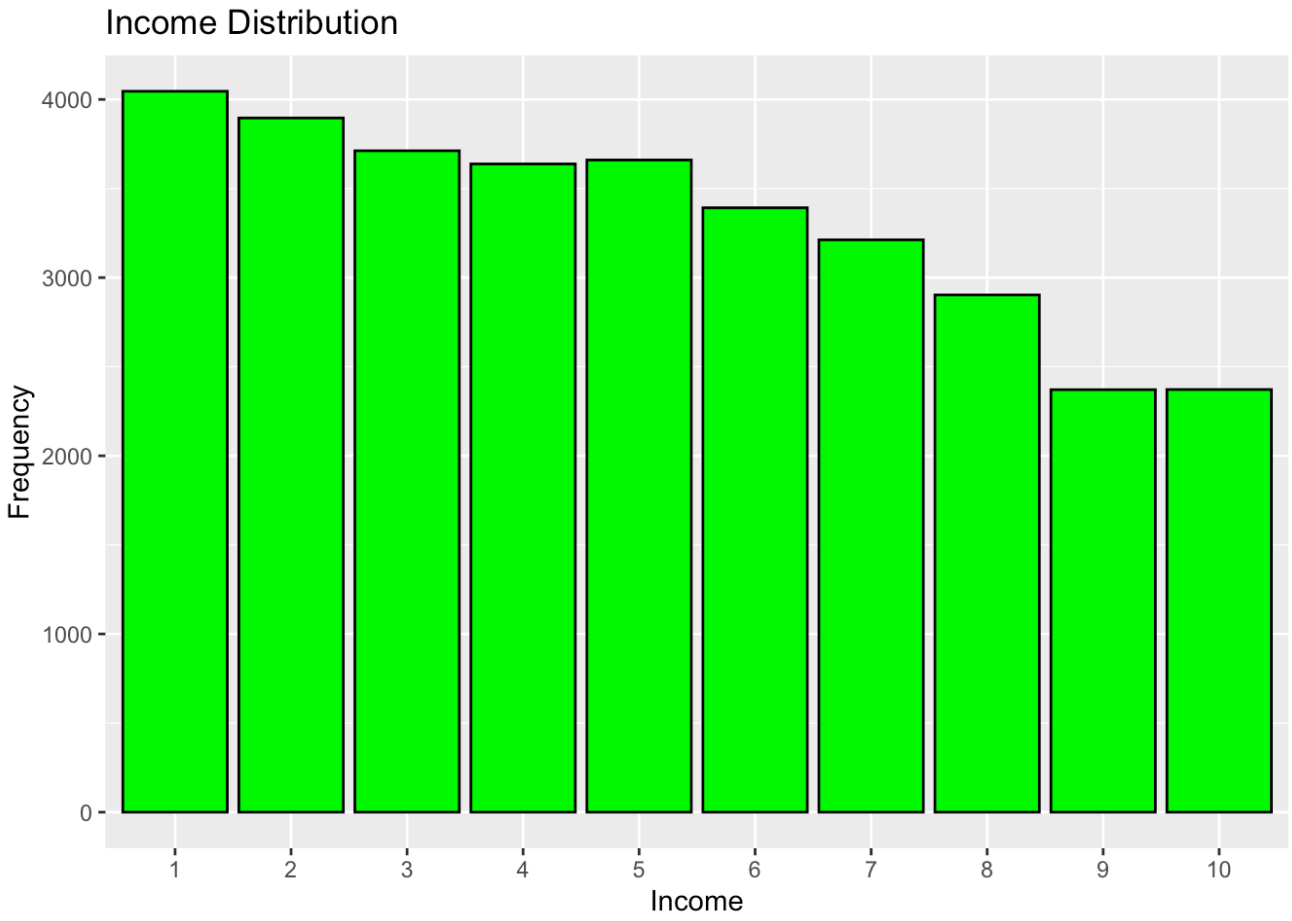
Since Age is a continuous variable , I chose histogram for effectively visualizing the frequency distribution of the column.

```r
ggplot(na.omit(ESS_polity), aes(x = edu)) +
  geom_bar(color = "black",fill="blue") +
  labs(title = "Education Distribution", x = "Education", y = "Frequency")
```



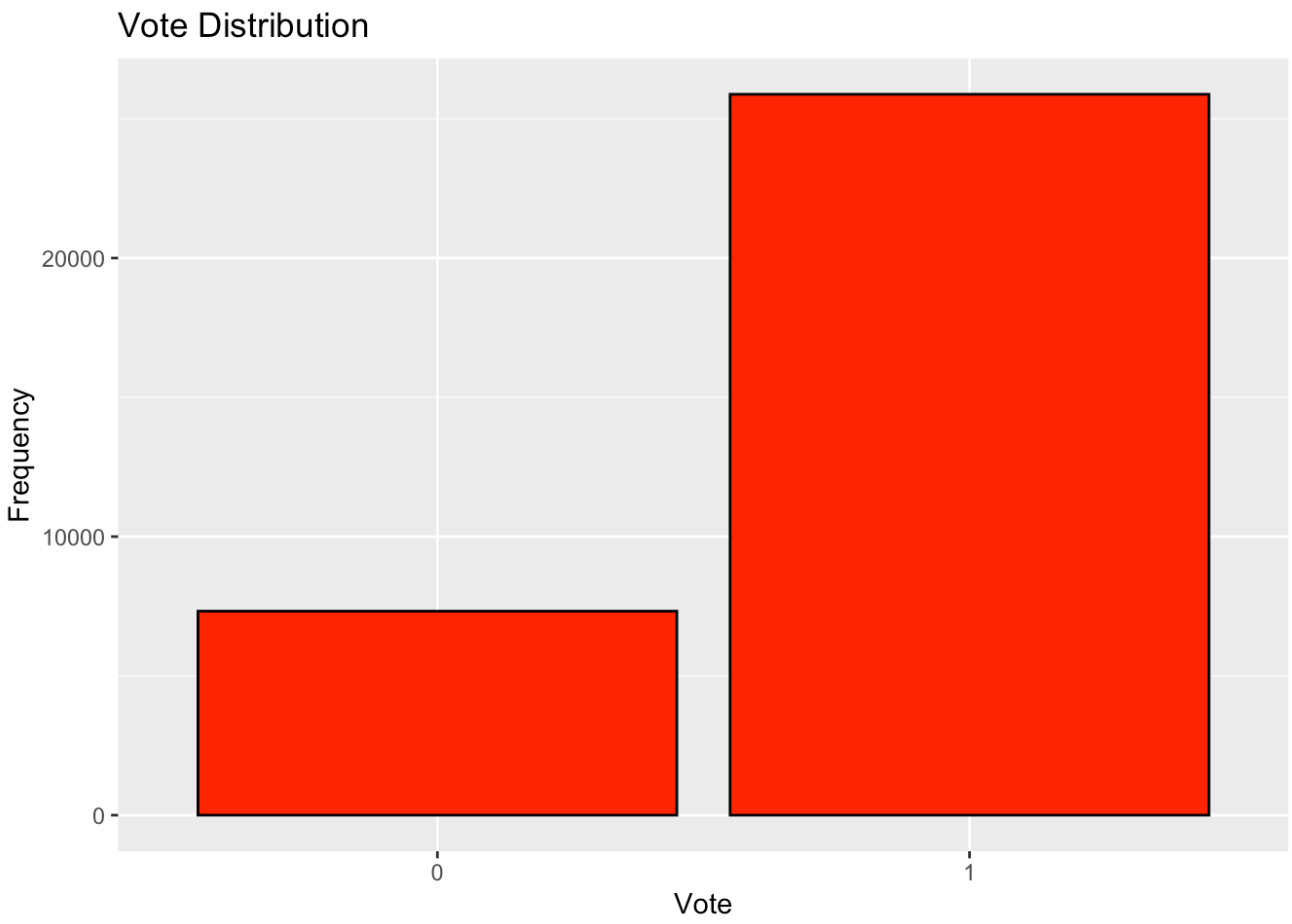

Since edu is a categorical variable, I use barplot to get a clear view of its distribution .

```r
ggplot(na.omit(ESS_polity), aes(x = as.factor(income_10))) +
  geom_bar(fill = "green", color = "black") +
  labs(title = "Income Distribution ", x = "Income", y = "Frequency")
```



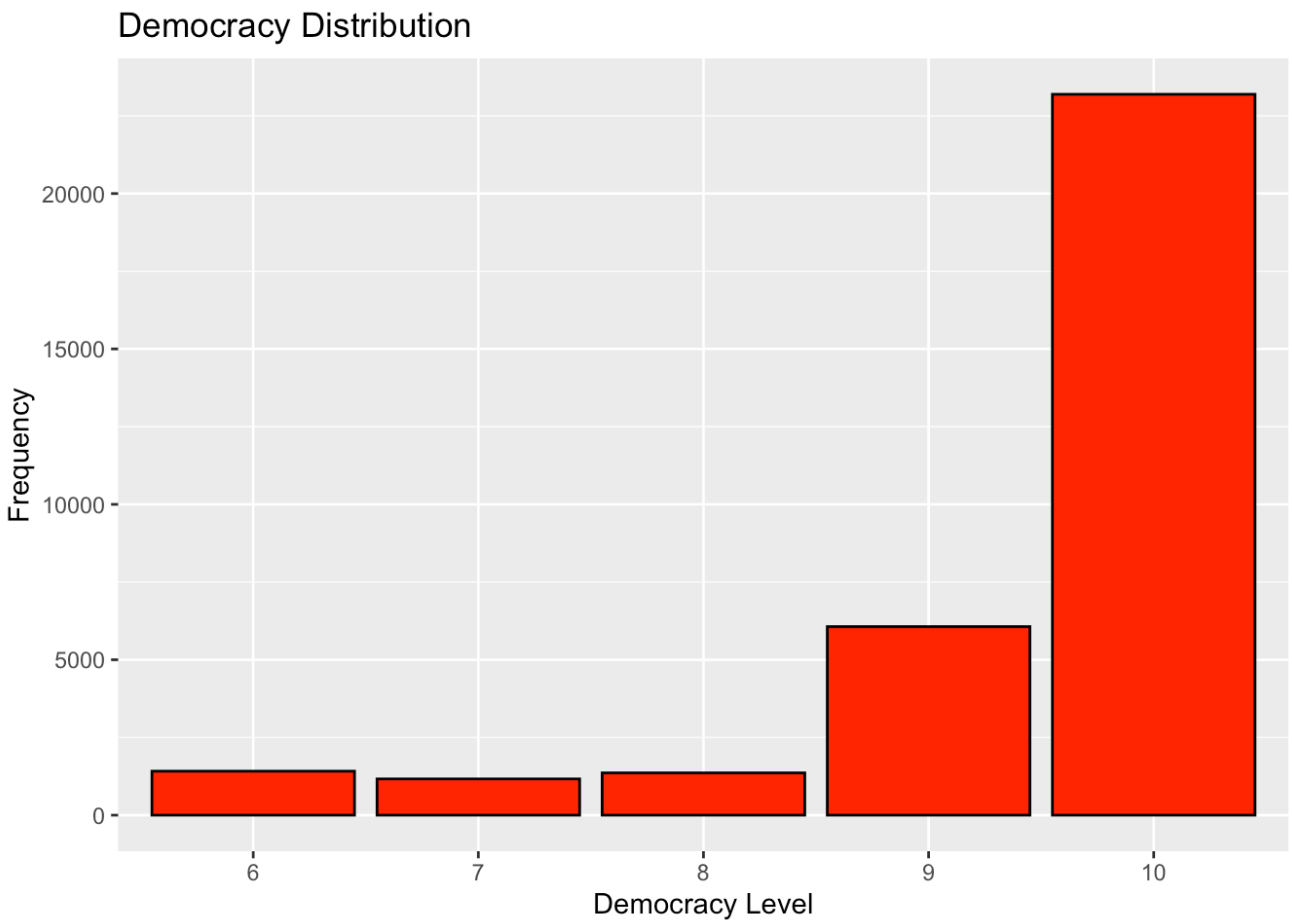

Since Income is a categorical variable, I have used a bar plot to see the frequency of each category.

```
ggplot(na.omit(ESS_polity), aes(x = as.factor(vote))) +
geom_bar(fill = "red", color = "black") +
labs(title = "Vote Distribution", x = "Vote", y = "Frequency")
```




Since Vote is a categorical variable, I have used a bar plot to see the frequency of each category.

```
ggplot(na.omit(ESS_polity), aes(x = as.factor(democ))) +
geom_bar(fill = "red", color = "black") +
labs(title = "Democracy Distribution", x = "Democracy Level", y = "Frequency")
```

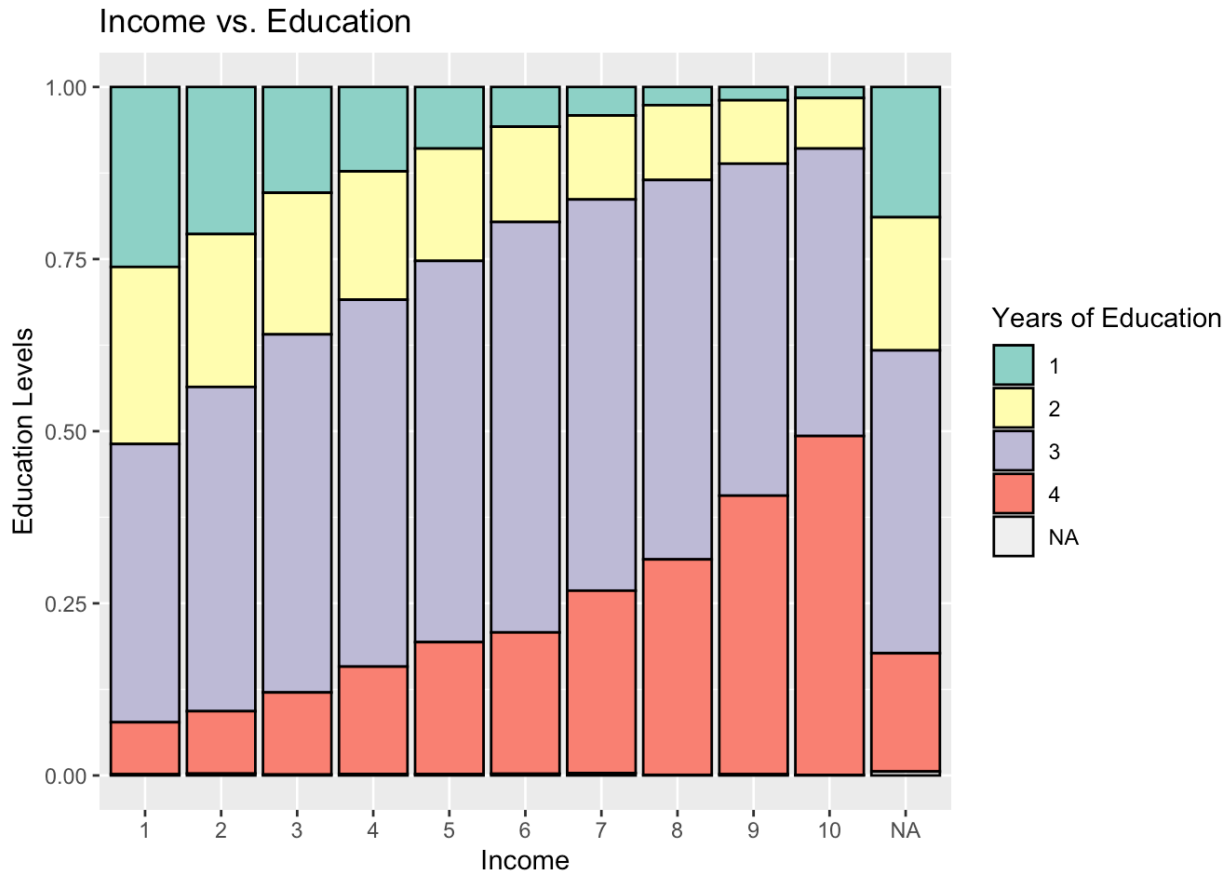



Since democ is a categorical variable, I have used a bar plot to see the frequency of each category.

3. **Suppose we want to test two hypotheses on the relationships of two pairs of variables. Please use the appropriate type of graphs we learned to visualize these two pairs of variables. Briefly describe the graph you plot, and answer: Does the graph we create from the data support the hypothesis?**

   (1) Hypothesis#1: The more years of education (edu) a person completed, the higher income (income_10) they earn. **(7.5%)**

```
ggplot(ESS_polity, aes(x = as.factor(income_10), fill = as.factor(edu))) +
  geom_bar(position = "fill", color = "black") +
  scale_fill_brewer(palette = "Set3", name = "Years of Education") +
  labs(title = "Income vs. Education", x = "Income", y = "Education Levels")
```




The grouped bar plot indicates that there is a positive relationship between years of education and income levels.

```
\(2\) Hypothesis#2: There is a gender disparity (male) in voting behavior (vote).
(Either men are more likely to vote, or women are more likely to vote). **(7.5%)**


::: {.cell}

```{.r .cell-code}
ESS_polity_clean <- ESS_polity[!is.na(ESS_polity$vote) & !is.na(ESS_polity$male), ]
ggplot(ESS_polity_clean, aes(x = as.factor(male), fill = as.factor(vote))) +
  geom_bar(position = "fill", color = "black") +
  scale_fill_manual(values = c("0" = "blue", "1" = "lightblue"), name = "Vote") +
  labs(title = "Gender Disparity in Voting Behavior", x = "Gender (0 = Female, 1 =
Male)", y = "Proportion")
```


::: {.cell-output-display}
![](challenge_4_Spring24_files/figure-html/unnamed-chunk-10-1.png){width=672}
:::
:::
```

Since both males and females have similar voting patterns, we can conclude that there is no disparity.

## Part 2. Comparing between Partial and Whole, and among Groups (30%)

In this part, we will use the clean version of the Australian public opinion poll on Same-Sex Marriage to generate graphs and plots. **You may need to do the data transformation or mutation needed to help graphing.**

1. Read in data. **(2.5%)**

```
#type of your code/command here.

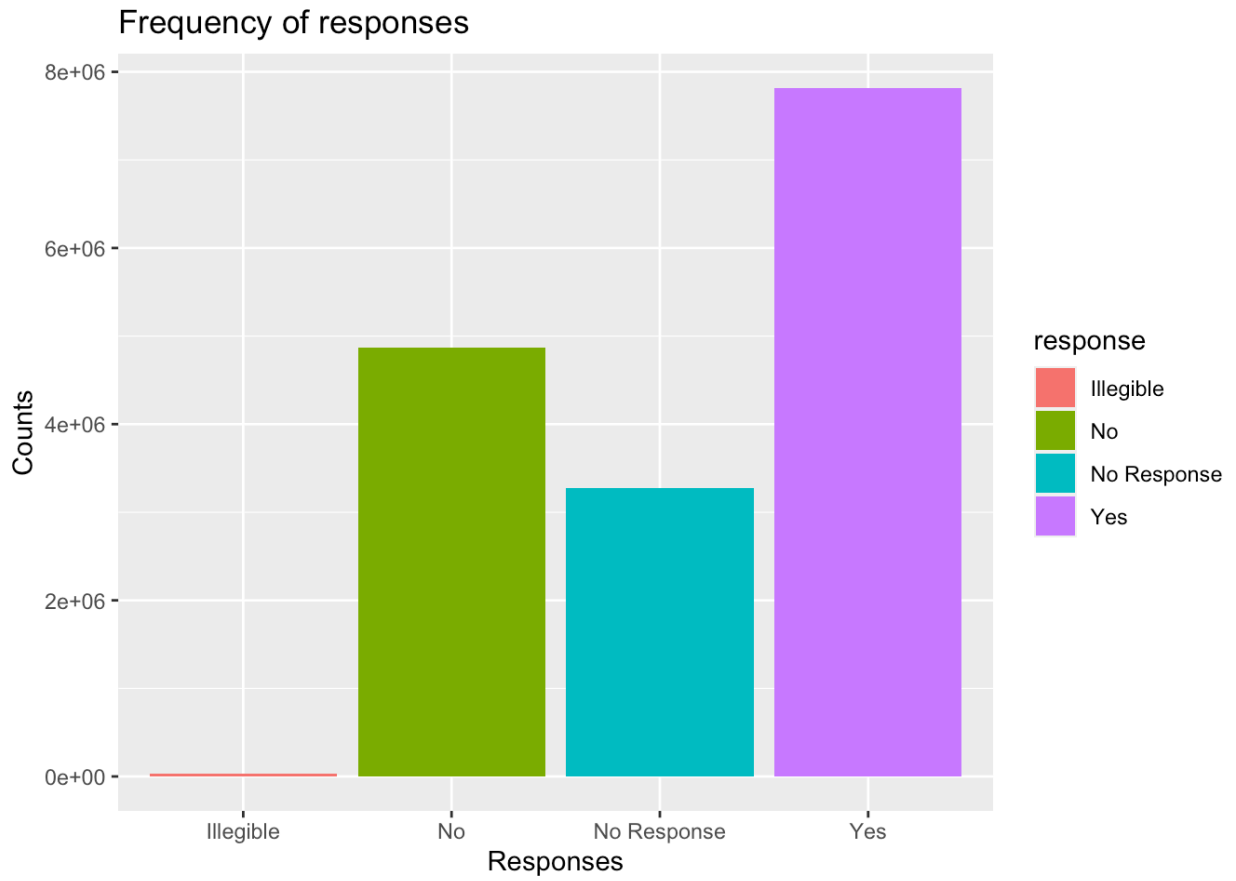
A_data <- read.csv("australian_data.csv", check.names = FALSE)
```

2. Use a barplot to graph the Australian data based on their responses: yes, no, illegible, and no response. The y-axis should be the count of responses, and each response should be represented by one individual bar (so there should be four bars). **(7.5%)**

   (you can use either geom_bar() or geom_col())

```
count_df <- data.frame(
  response = c("Yes", "No", "Illegible", "No Response"),
  count = c(sum(A_data$Yes), sum(A_data$No), sum(A_data$Illegible), sum(A_data$`

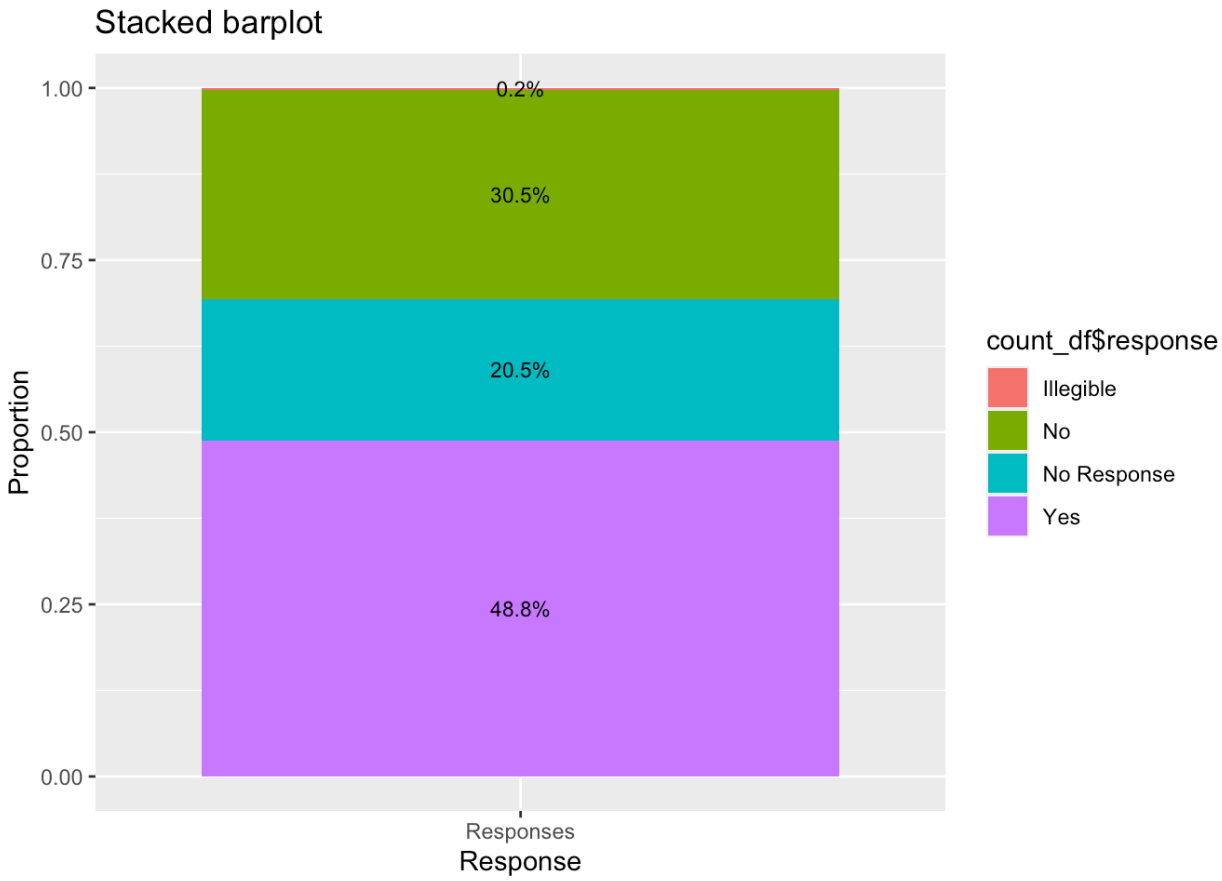
ggplot(count_df, aes(x = response, y = count, fill = response)) + geom_col() +
  labs(title = "Frequency of responses", x = "Responses", y = "Counts")
```



3. The previous graph only shows the difference in amount. Let's create a stacked-to-100% barplot to show the proportion of each of the four responses (by % of the total response). **(7.5%)**

   (you can use either geom_bar() or geom_col())

```
count_responses <- sum(count_df$count)
ratios <- count_df$count / count_responses
new <- data.frame(response = "Responses",proportion = ratios)
ggplot(new, aes(x = response, y = proportion, fill = count_df$response, label =
geom_col(position = "fill") +
labs(title = "Stacked barplot", x = "Response", y = "Proportion") +
geom_text(position = position_fill(vjust = 0.5), size = 3)
```

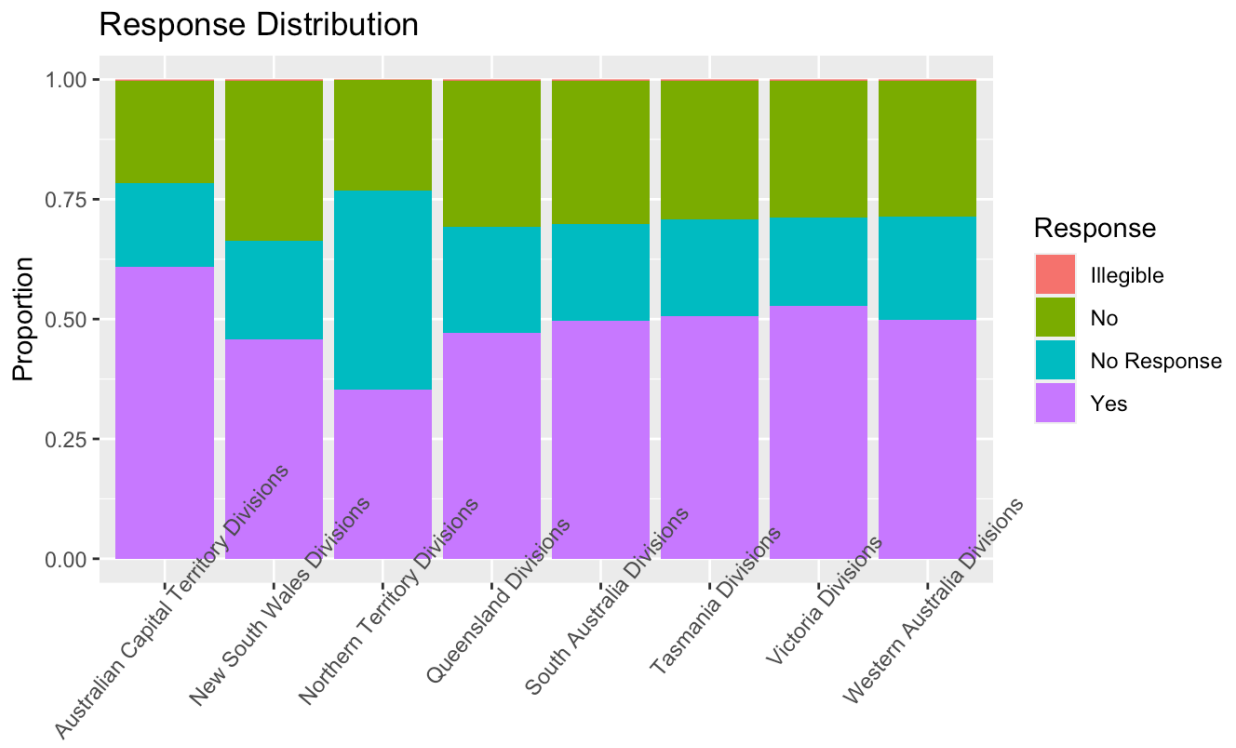



4. Let's see if there's a relationship between Division and Response - that is, are certain divisions more likely to respond one way compared to other divisions? Again, we will use barplot(s) to present the visualization. **(12.5%)**

   (you can use either geom_bar() or geom_col())

```
long_data <- data.frame(
  District = rep(A_data$District, times = 4),
  Division = rep(A_data$Division, times = 4),
  Response = rep(c("Yes", "No", "Illegible", "No Response"), each = nrow(A_data)
  Count = c(A_data$Yes, A_data$No, A_data$Illegible, A_data$`No Response`)
)
ggplot(long_data, aes(fill = Response, x = Division, y = Count)) +
  geom_bar(stat = "identity", position = "fill") +
```

```
  labs(title = "Response Distribution", x = NULL, y = "Proportion") +
  theme(axis.text.x = element_text(angle = 50))
```




## Part 3. Practice plotting with a dataset of your choice (25% of the total grade)

In this part, you will choose data of your interests for graphing and plotting. This data can be tidy/ready-to-be-used or raw data that needs cleaning. If the data is very large (for example, more than 20 columns), you should definitely subset the data by selecting less than 10 variables of your interests to avoid taking too much room in your R memory.

1. Include a link to the data page (this page should include the introduction or description and the link to download this dataset). **(2%)** https://www.kaggle.com/datasets/shreyanshverma27/online-sales-dataset-popular-marketplace-data

2. Read the data you choose and briefly answer the following questions. (Optional: you may need to subset, clean, and transform the data if necessary). **(8%)**

```
   sales <- read.csv("Online Sales Data.csv")
head(sales)
```

| | Date | Product.Category | Product.Name |
|---|---|---|---|
| | <chr> | <chr> | <chr> |
| 1 | 1/1/24 | Electronics | iPhone 14 Pro |

| | Date | Product.Category | Product.Name |
|---|---|---|---|
| | <chr> | <chr> | <chr> |
| 2 | 1/2/24 | Home Appliances | Dyson V11 Vacuum |
| 3 | 1/3/24 | Clothing | Levi's 501 Jeans |
| 4 | 1/4/24 | Books | The Da Vinci Code |
| 5 | 1/5/24 | Beauty Products | Neutrogena Skincare Set |
| 6 | 1/6/24 | Sports | Wilson Evolution Basketball |

6 rows | 1-4 of 9 columns

(1) What is the structure (dimension) of the data;

```
print(dim(sales))
```

```
[1] 240   8
```

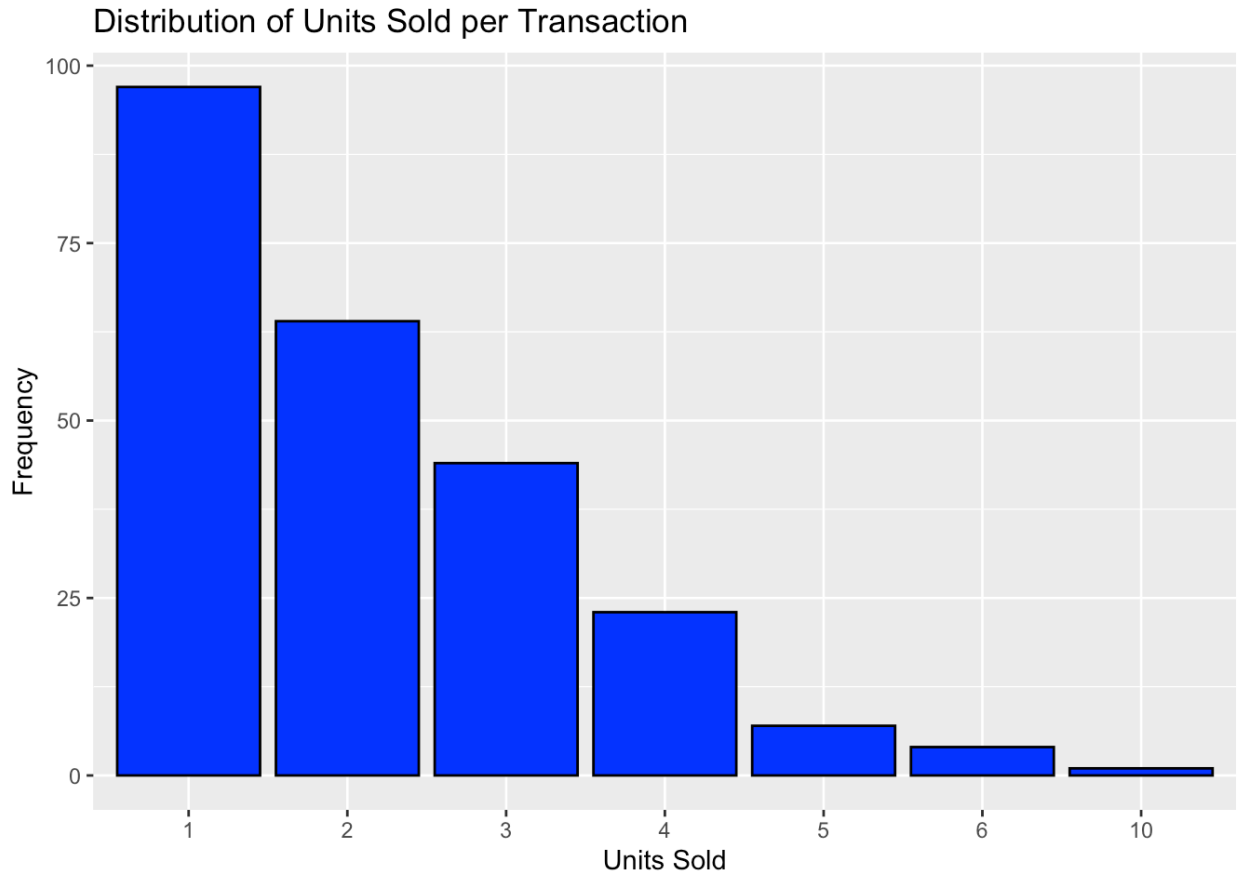
There are total 240 rows and 8 columns.

```
\(2\) What is the unit of observation?
The unit of observation is each individual sale transaction. Each row represents a
unique transaction, which includes details such as the date of the transaction, the
product category, the product name, the number of units sold, the unit price, the
total revenue, the region of the sale, and the payment method used.

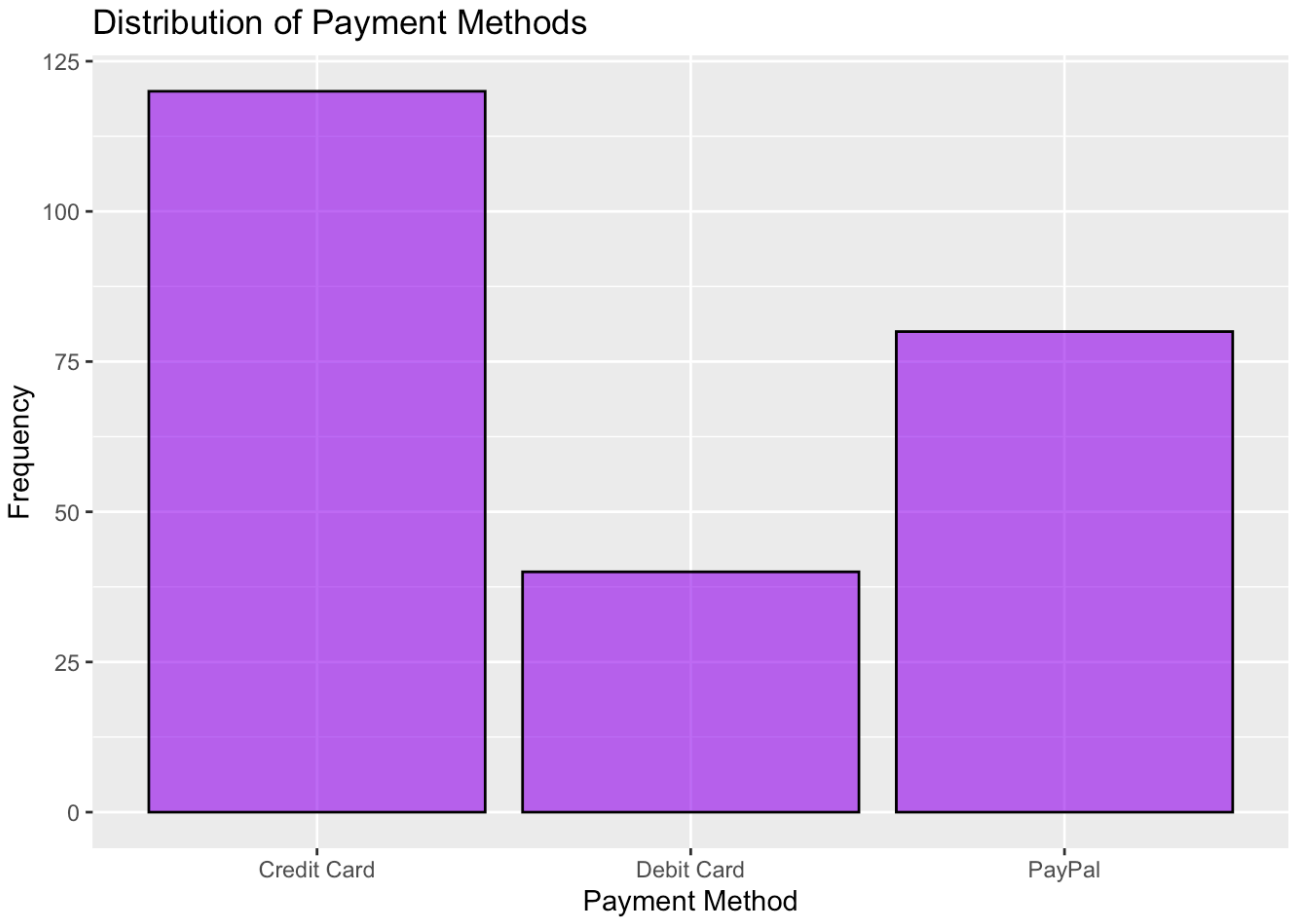
\(3\) What does each column mean in this data?
Date: The date on which the transaction occurred.
Product.Category: The category to which the product belongs.
Product.Name: The name of the product sold.
Units.Sold: The number of units sold.
Unit.Price: The price per unit of the product.
Total.Revenue: The total revenue generated from the transaction.
Region: The region where the sale occurred.
Payment.Method: The method of payment used for the transaction.
```

3. Choose two columns/variables of your interests. Plot one univariate graph for each of the variables. **(5%)**

```
sales$Units.Sold <- as.factor(sales$Units.Sold)
ggplot(sales, aes(x = Units.Sold)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Distribution of Units Sold per Transaction", x = "Units Sold", y
```

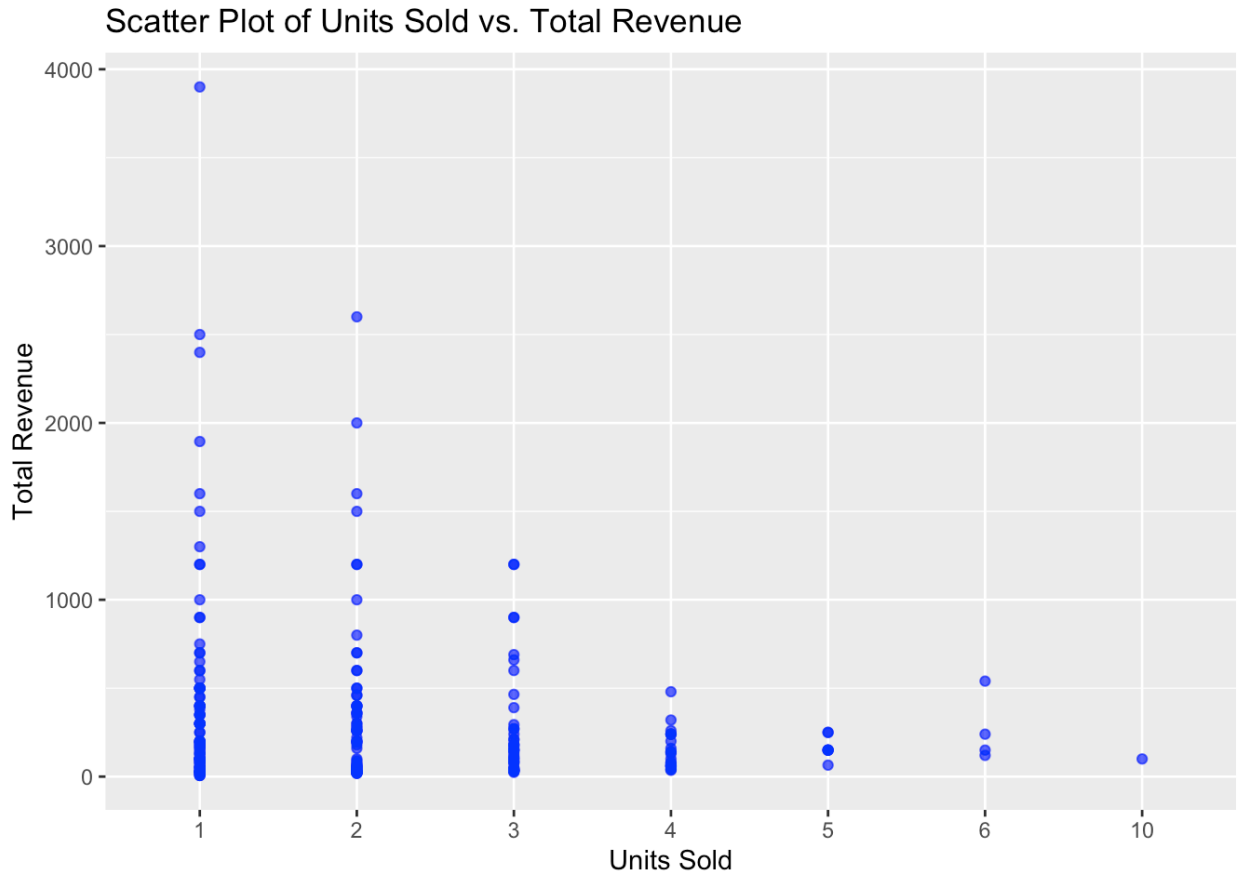



```
ggplot(sales, aes(x = Payment.Method)) +
  geom_bar(fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Payment Methods", x = "Payment Method", y = "Frequer
```




4. Choose a pair of variables that may be correlated and make a graph (scatter plot or barplot) using them. Based on the visual evidence, do you see any potential correlation between the two variables **(10%)**

```
ggplot(sales, aes(x = Units.Sold, y = Total.Revenue)) +
geom_point(color = "blue", alpha = 0.7) +
labs(title = "Scatter Plot of Units Sold vs. Total Revenue", x = "Units Sold",
```




There is a potential positive correlation between Units.Sold and Total.Revenue. As the number of units sold increases, the total revenue tends to increase.

## Appendix: sources for data to be used in Part 3

**Here are some online sources and popular Online Dataset Hub:**

1. Many US governments (usually at the federal and state levels),  bureaus, and departments have open data archives on their websites, allowing the public to access, download, and use them. Just use Google to search for them.

2. **The Harvard Dataverse Repository** is a free data repository open to all researchers from any discipline, inside and outside the Harvard community, where you can share, archive, cite, access, and explore research data. Each individual Dataverse collection is a customizable collection of datasets (or a virtual repository) for organizing, managing, and showcasing datasets.

3. **Inter-university Consortium for Political and Social Research (ICPSR)** of the University of Michigan-Ann Arbor provides leadership and training in data access, curation, and methods of analysis for the social science research community.

4. **UN: https://data.un.org/**

5. **OECD Data**:  economic and development data of the most developed countries in the world.

6. The five sources above are mainly for social science data; **there exists another very big community and open data archives for machine-learning and data science: Kaggle.**