Worksheet-4c in R

Missy Key Sadsad BSIT-2B

2023-11-22

1. Use the dataset mpg

a. Show your solutions on how to import a csv file into the environment.

```
#a
mpg <-read.csv(file = "mpg.csv",header = TRUE, sep = "," )
View(mpg)

#b
str(mpg)
```

```
## 'data.frame':    234 obs. of  12 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
#c
#based on given its displ, cty,hwy
```

2. Which manufacturer has the most models in this data set? Which model has the most variations? Show your answer.

```
#a
manufacturers <- table(mpg$manufacturer)
manufacturers #dodge
```

```
##
##       audi  chevrolet      dodge       ford      honda    hyundai       jeep
##         18         19         37         25          9         14          8
## land rover    lincoln    mercury     nissan    pontiac     subaru     toyota
##          4          3          4         13          5         14         34
## volkswagen
##         27
```

```r
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.3.2
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
#b
models <- mpg%>%count(mpg$model)
models #caravan 2wd
```

```
##                  mpg$model  n
## 1              4runner 4wd  6
## 2                       a4  7
## 3               a4 quattro  8
## 4               a6 quattro  3
## 5                   altima  6
## 6       c1500 suburban 2wd  5
## 7                    camry  7
## 8             camry solara  7
## 9              caravan 2wd 11
## 10                   civic  9
## 11                 corolla  5
## 12                corvette  5
## 13       dakota pickup 4wd  9
## 14             durango 4wd  7
## 15          expedition 2wd  3
## 16            explorer 4wd  6
## 17         f150 pickup 4wd  7
## 18            forester awd  6
## 19      grand cherokee 4wd  8
## 20              grand prix  5
## 21                     gti  5
## 22             impreza awd  8
## 23                   jetta  9
## 24         k1500 tahoe 4wd  4
## 25  land cruiser wagon 4wd  2
## 26                  malibu  5
## 27                  maxima  3
## 28         mountaineer 4wd  4
## 29                 mustang  9
## 30           navigator 2wd  3
## 31              new beetle  6
## 32                  passat  7
## 33         pathfinder 4wd  4
## 34    ram 1500 pickup 4wd 10
## 35            range rover  4
## 36                 sonata  7
## 37                tiburon  7
## 38      toyota tacoma 4wd  7
```

```
unique_models <- mpg %>%group_by(manufacturer)%>%distinct(model)
unique_models
```

```
## # A tibble: 38 x 2
## # Groups:   manufacturer [15]
##    manufacturer model
##    <chr>        <chr>
##  1 audi         a4
##  2 audi         a4 quattro
##  3 audi         a6 quattro
##  4 chevrolet    c1500 suburban 2wd
##  5 chevrolet    corvette
##  6 chevrolet    k1500 tahoe 4wd
##  7 chevrolet    malibu
##  8 dodge        caravan 2wd
##  9 dodge        dakota pickup 4wd
## 10 dodge        durango 4wd
## # i 28 more rows
```

```
library(ggplot2)
```

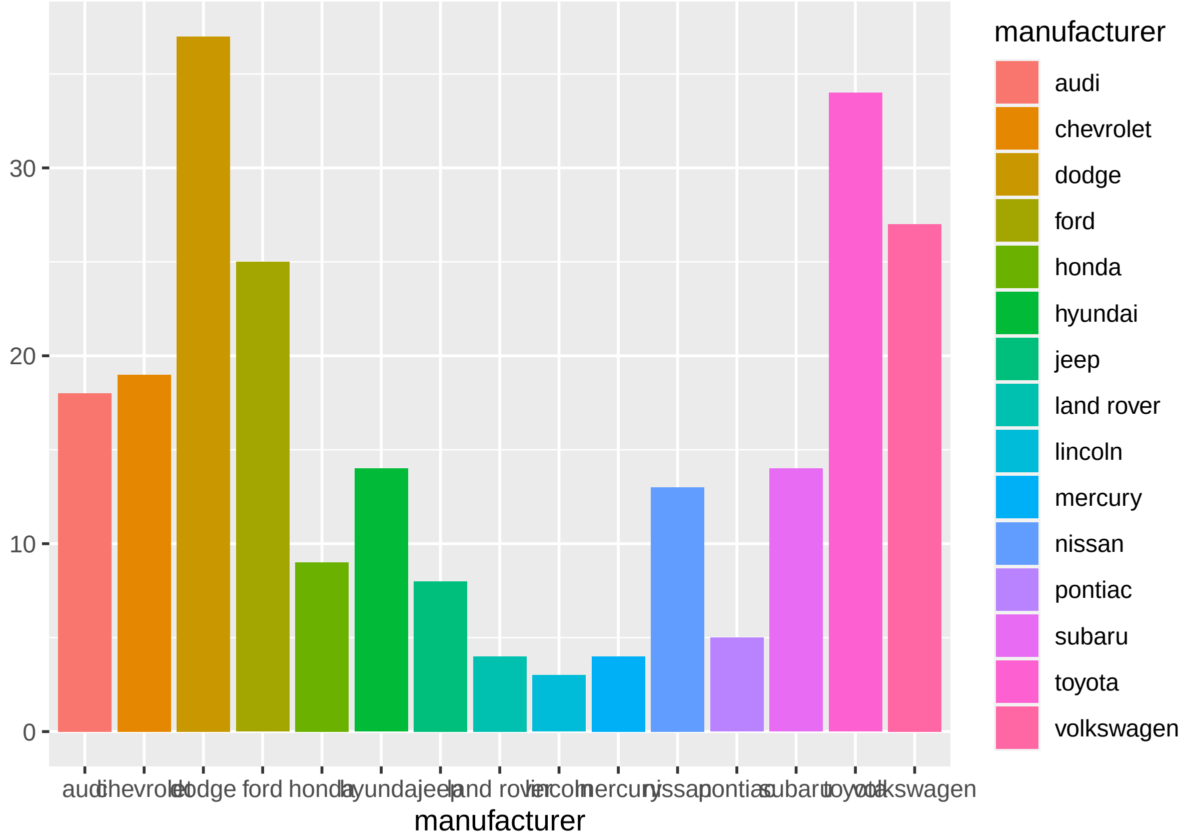
```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     mpg
```

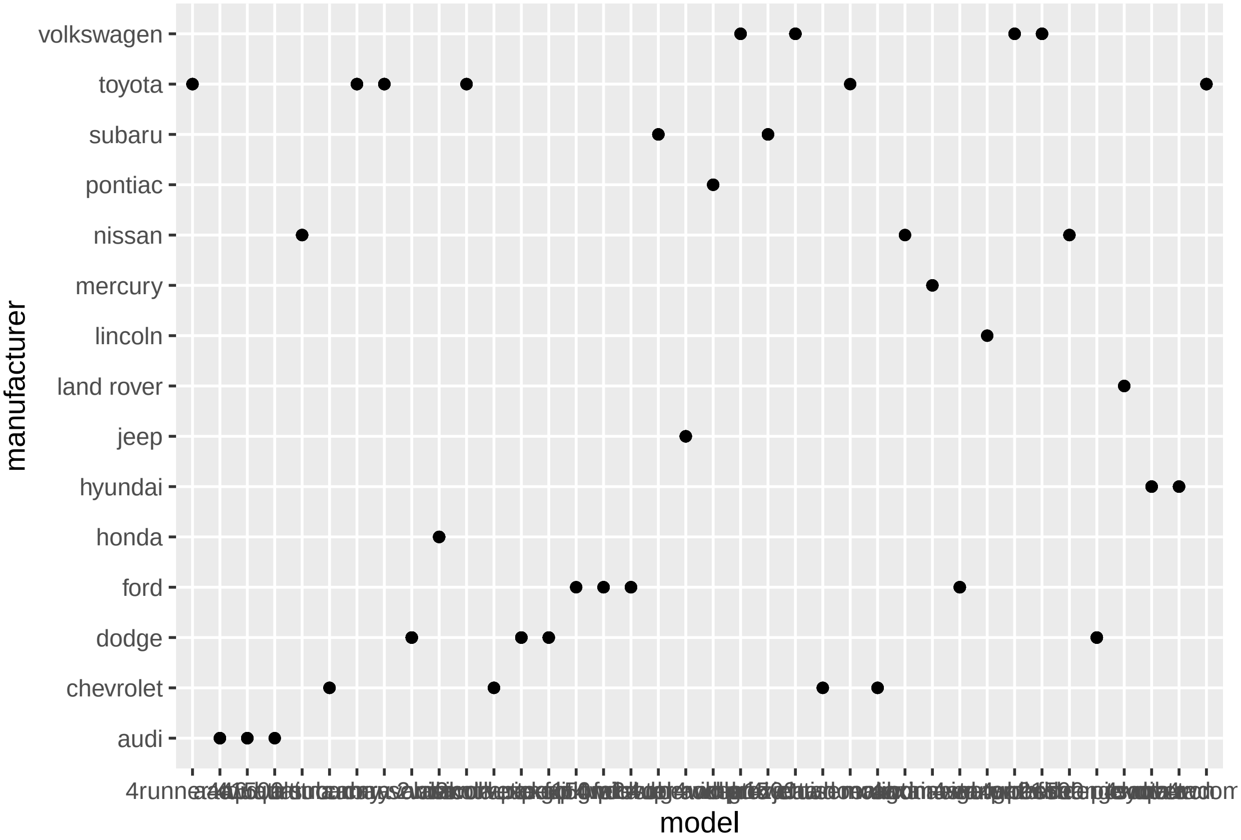
```
qplot(manufacturer, data = mpg,
    geom = "bar", fill = manufacturer)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



2. Same dataset will be used. You are going to show the relationship of the modeland the manufacturer.

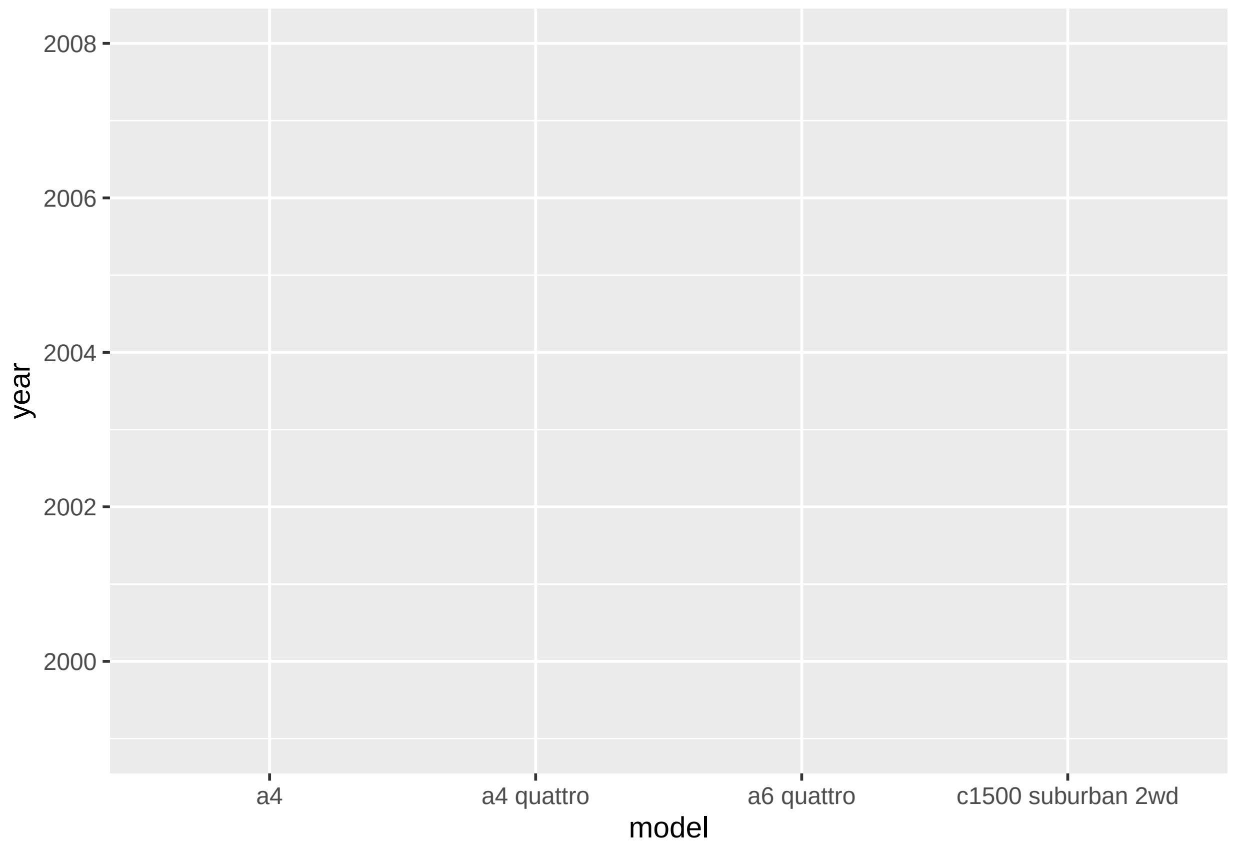
```
#a
ggplot(mpg, aes(model, manufacturer)) + geom_point()
```



```
#b
#based on the graph, it created a simple scatterplot where each point represent the manufacturer and it
```

3. Plot the model and the year using ggplot(). Use only the top 20 observations. Write the codes and its results.

```
#3.
top_20_observations <- head(mpg, 20)

ggplot(top_20_observations, aes(x = model, y = year))
```

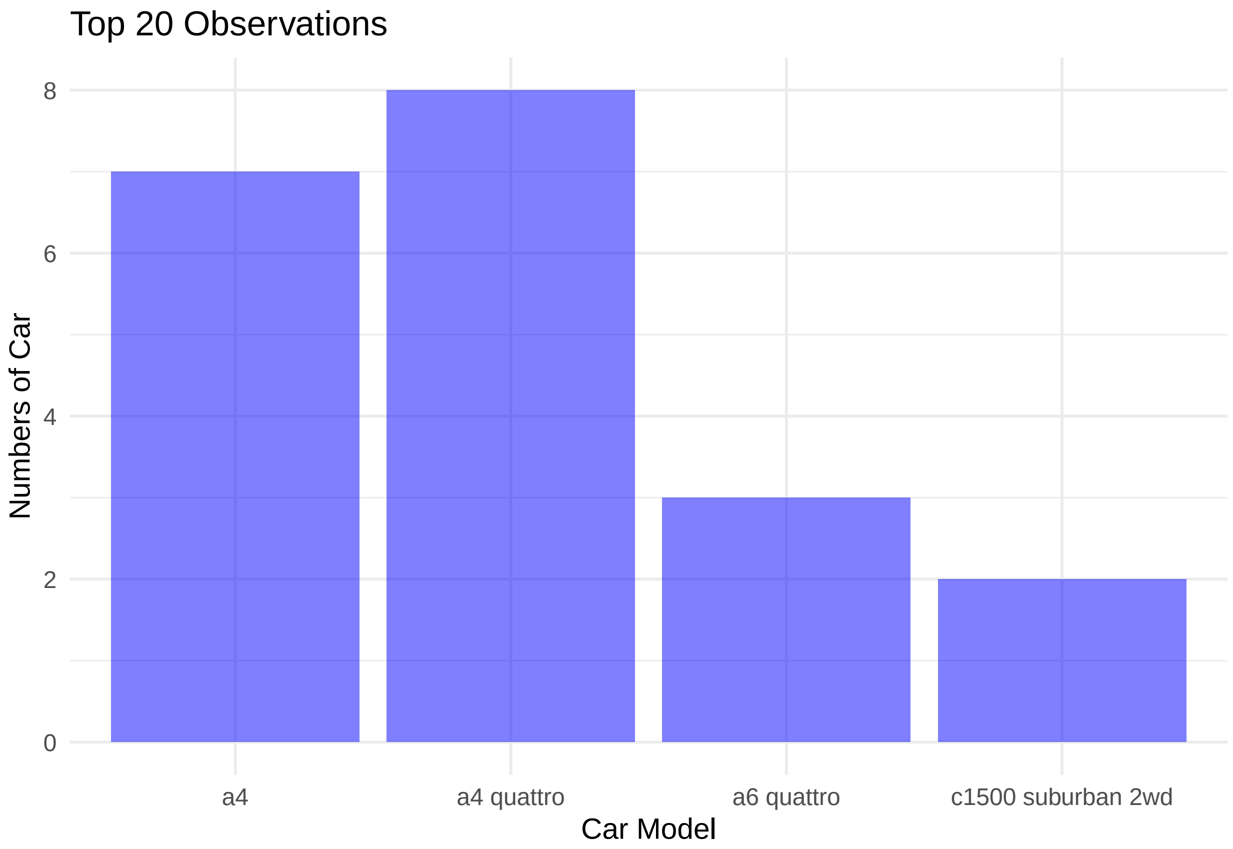


4. Using the pipe (%>%), group the model and get the number of cars per model. Show codes and its result

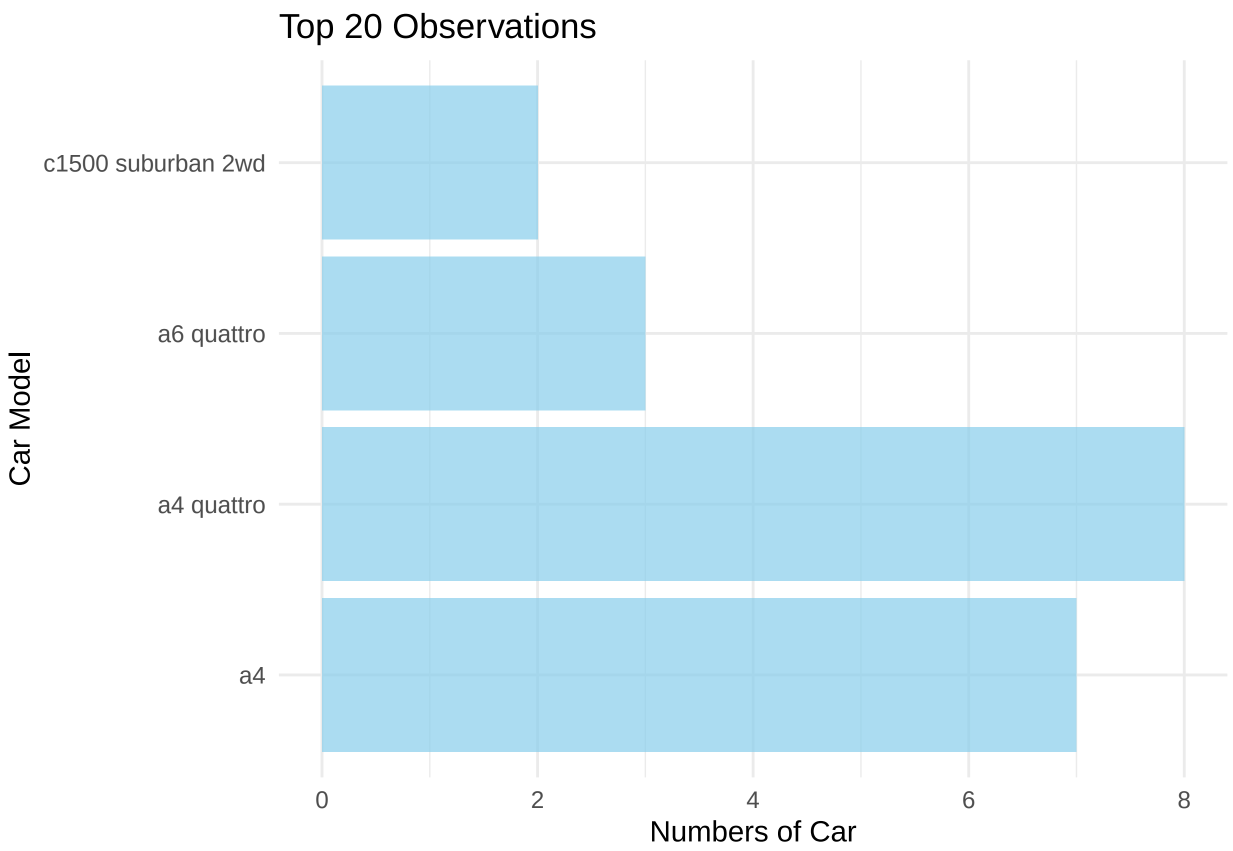
```
library(dplyr)

models_group <- mpg %>%
group_by(model)%>%
summarise(number_of_cars = n())
models_group
```

```
## # A tibble: 38 x 2
##    model              number_of_cars
##    <chr>                       <int>
##  1 4runner 4wd                     6
##  2 a4                              7
##  3 a4 quattro                      8
##  4 a6 quattro                      3
##  5 altima                          6
##  6 c1500 suburban 2wd              5
##  7 camry                           7
##  8 camry solara                    7
##  9 caravan 2wd                    11
## 10 civic                           9
## # i 28 more rows
```

```r
#a
ggplot(top_20_observations, aes(x = model)) +
  geom_bar(fill = "blue", alpha = 0.5) +
  labs(title = "Top 20 Observations",
       x = "Car Model",
       y = "Numbers of Car") +
  theme_minimal()
```

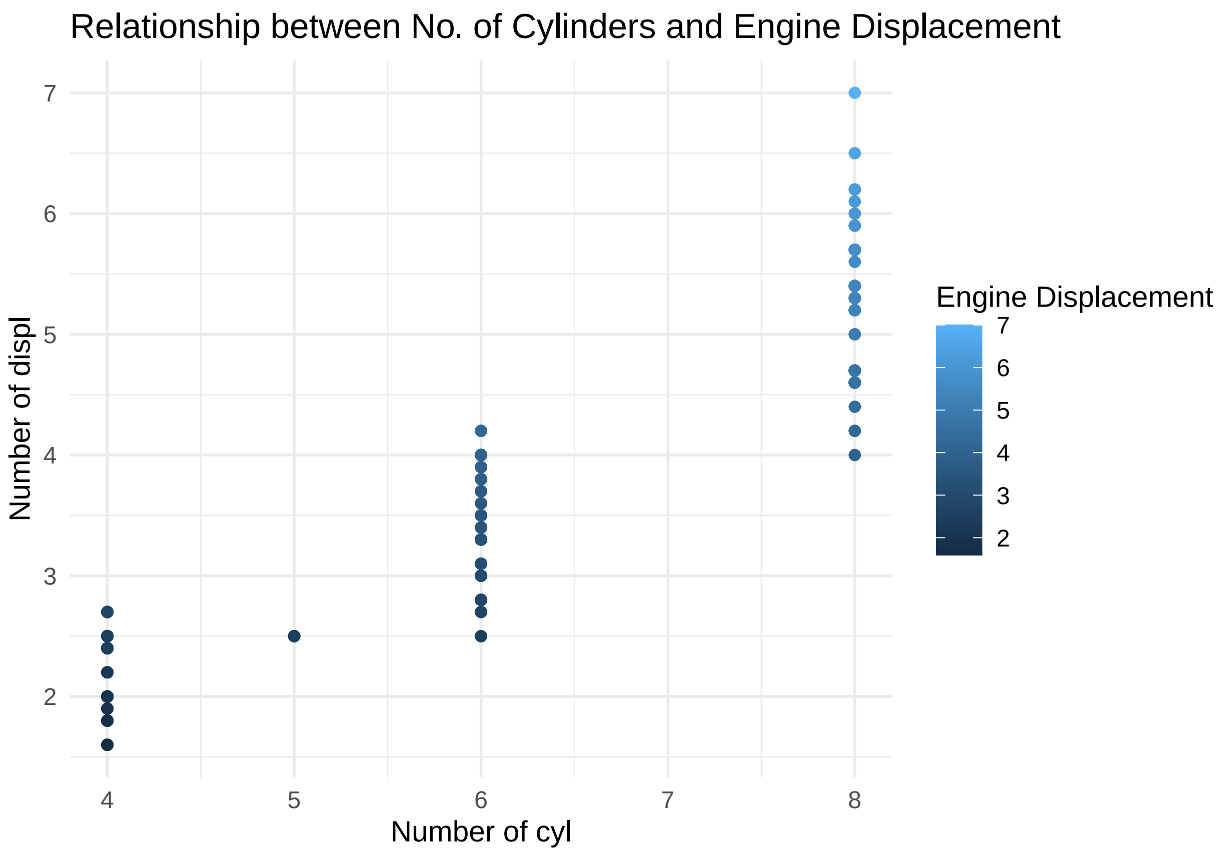


```r
#b
ggplot(top_20_observations, aes(x = model)) +
  geom_bar(fill = "skyblue", alpha = 0.7) +  # You can customize the color and transparency
  labs(title = "Top 20 Observations",
       x = "Car Model",
       y = "Numbers of Car") +
  theme_minimal() +
  coord_flip()
```



5. Plot the relationship between cyl - number of cylinders and displ - engine displacement using geom_point with aesthetic color = engine displacement. Title should be "Relationship between No. of Cylinders and Engine Displacement".

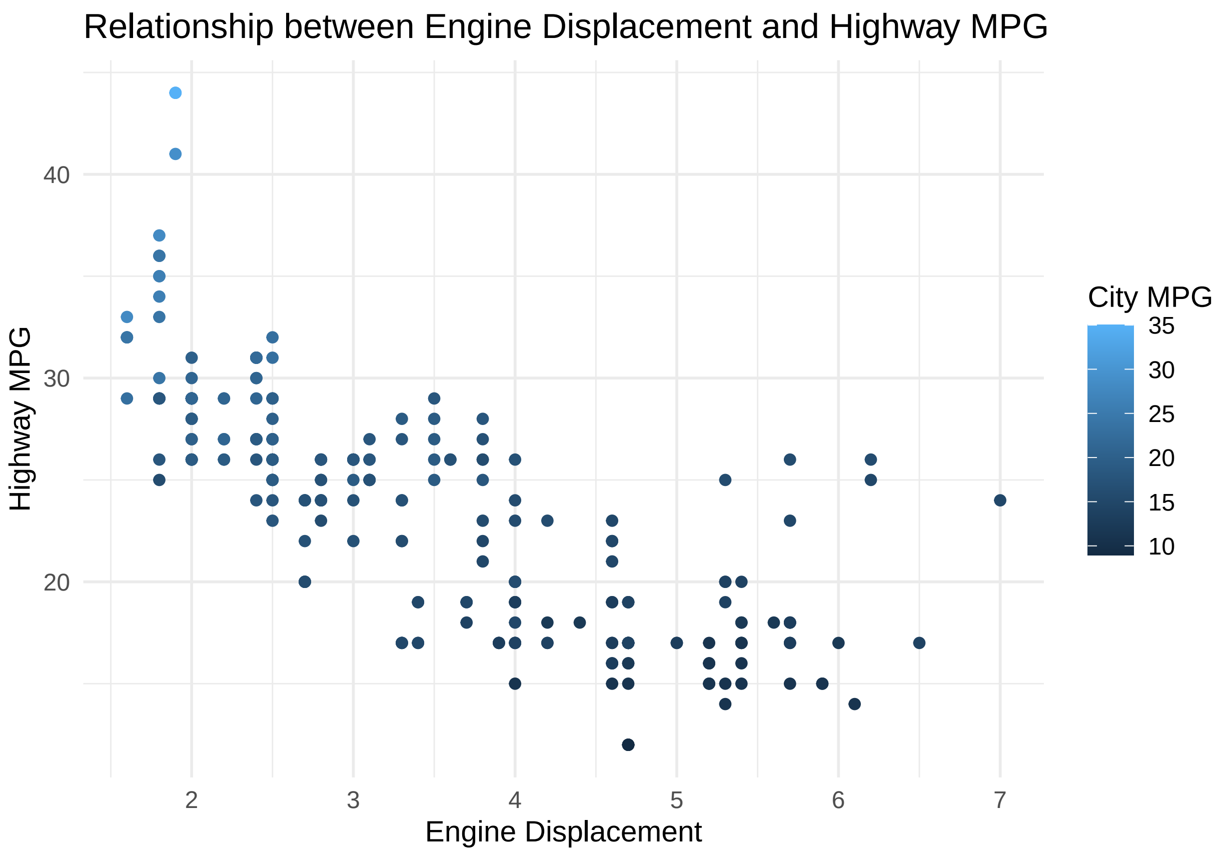
```r
ggplot(mpg, aes(x = cyl, y = displ, color = displ)) +
  geom_point() +
  labs(title = "Relationship between No. of Cylinders and Engine Displacement",
       x = "Number of cyl",
       y = "Number of displ") +
  scale_color_continuous(name = "Engine Displacement") +
  theme_minimal()
```



```
#a
#this code will generate a scatter plot with a trendline, allowing you to observe any patterns in the r
```

6. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon). Mapped it with a continuous variable you have identified in #1-c. What is its result? Why it produced such output?

```
ggplot(mpg, aes(x = displ, y = hwy, color = cty)) +
geom_point() +
  labs(title = "Relationship between Engine Displacement and Highway MPG",
       x = "Engine Displacement",
       y = "Highway MPG",
       color = "City MPG") +
  theme_minimal()
```



6. Import the traffic.csv onto your R environment.

```
#a
traffic_data <- read.csv("traffic.csv")
View(traffic_data)

num_traffic_obv <-nrow(traffic_data)
num_traffic_obv
```

```
## [1] 48120
```

```
str(traffic_data)
```

```
## 'data.frame':    48120 obs. of  4 variables:
##  $ DateTime: chr  "2015-11-01 00:00:00" "2015-11-01 01:00:00" "2015-11-01 02:00:00" "2015-11-01 03:0
##  $ Junction: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Vehicles: int  15 13 10 7 9 6 9 8 11 12 ...
##  $ ID      : num  2.02e+10 2.02e+10 2.02e+10 2.02e+10 2.02e+10 ...
```

```
#The variable of traffic dataset is DateTime, Junction, Vehicles, and ID.

 #b
```

7. From alexa_file.xlxs, import it to your environment

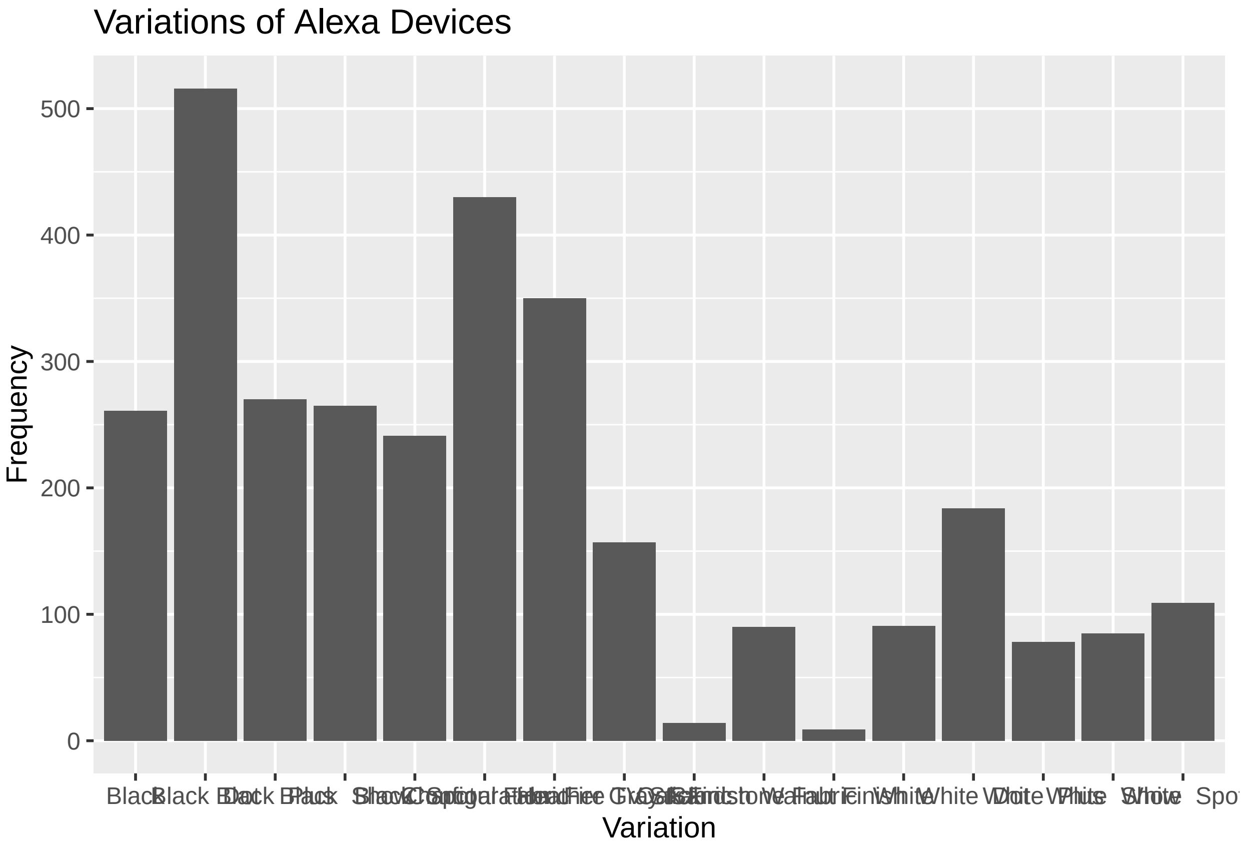
```r
library(readxl)
alexa_file <- read_excel("~/GitHub/RWorksheets_Sadsad/Worksheet#4/RWorksheet#4b/alexa_file.xlsx")
View(alexa_file)

#a
nrow(alexa_file)
```

```
## [1] 3150
```

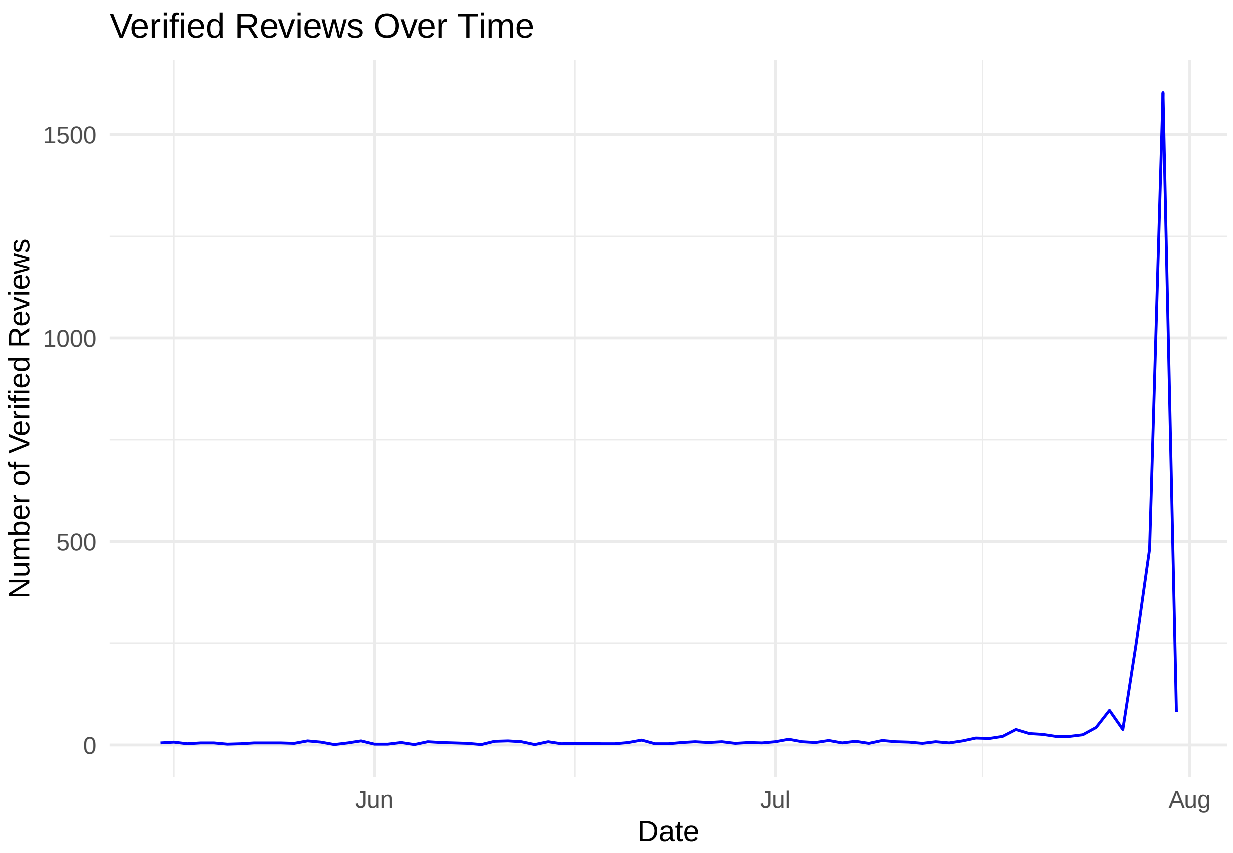
```r
ncol(alexa_file)
```

```
## [1] 5
```

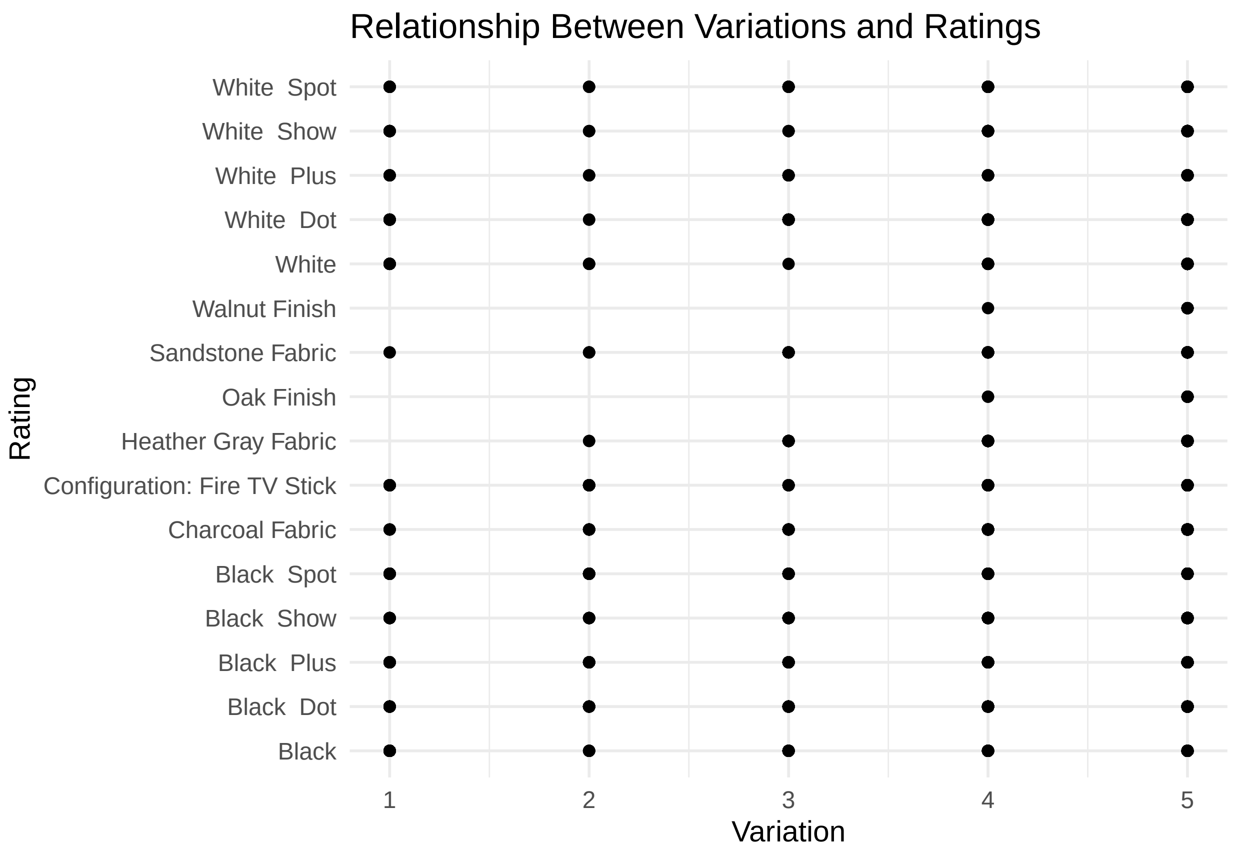
```r
#b
alexa_data <- alexa_file%>%
  group_by(variation) %>%
  summarise(Frequency = n())

View(alexa_data)

ggplot(alexa_data, aes(x = variation, y = Frequency )) +
  geom_bar(stat = "identity")  +
   labs(
    title = "Variations of Alexa Devices",
    x = "Variation",
    y = "Frequency"
  )
```



```
#b
#Each bar represents a variation, and its height indicates how frequently it appears in the data. This

#c
summary_reviews <- alexa_file %>%
  group_by(date) %>%
  summarize(NumVerifiedReviews = n())
#e
ggplot(summary_reviews, aes(x =  date, y = NumVerifiedReviews )) +
  geom_line(color = "blue") +
  labs(
    title = "Verified Reviews Over Time",
    x = "Date",
    y = "Number of Verified Reviews"
  ) +
  theme_minimal()
```



```
#d
ggplot(alexa_file, aes(x = rating, y = variation)) +
  geom_point() +
   labs(
    title = "Relationship Between Variations and Ratings",
    x = "Variation",
    y = "Rating"
  ) +
  theme_minimal()
```



```
#the highest variations rating is Walnut Finish and Oak Finish
```