RWorksheet_Conlu#4c

Jasper Conlu

2023-12-20

```
#1. Use the dataset mpg
#A data frame with 234 rows and 11 variables:
#' \describe{
#' \item{manufacturer}{manufacturer name}
#' \item{year}{year of manufacture}
#' \item{model}{model name}
#' \item{displ}{engine displacement, in litres}
#' \item{cyl}{number of cylinders}
#' \item{trans}{type of transmission}
#' \item{drv}{the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd}
#' \item{cty}{city miles per gallon}
#' \item{hwy}{highway miles per gallon}
#' \item{fl}{fuel type}
#' \item{class}{"type" of car}
#' }
"mpg"
```

```
## [1] "mpg"
```

```
#A.
#1st download the mpg.csv file
#2nd upload the mpg file in the posit cloud or r studio by clicking the upload in the file/plot tab
#3rd click the mpg.csv file in the files/plot tab and click import data set

library(openxlsx)
library(readr)
mpg <- read_csv("mpg.csv",show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
spec(mpg)
```

```
## cols(
##   ...1 = col_double(),
##   manufacturer = col_character(),
##   model = col_character(),
##   displ = col_double(),
##   year = col_double(),
##   cyl = col_double(),
##   trans = col_character(),
##   drv = col_character(),
##   cty = col_double(),
##   hwy = col_double(),
##   fl = col_character(),
##   class = col_character()
## )
```

```
head(mpg)
```

```
## # A tibble: 6 x 12
##    ...1 manufacturer model displ  year   cyl trans drv     cty   hwy fl    class
##   <dbl> <chr>        <chr> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <chr> <chr>
## 1     1 audi         a4      1.8  1999     4 auto~ f        18    29 p     comp~
## 2     2 audi         a4      1.8  1999     4 manu~ f        21    29 p     comp~
## 3     3 audi         a4      2    2008     4 manu~ f        20    31 p     comp~
## 4     4 audi         a4      2    2008     4 auto~ f        21    30 p     comp~
## 5     5 audi         a4      2.8  1999     6 auto~ f        16    26 p     comp~
## 6     6 audi         a4      2.8  1999     6 manu~ f        18    26 p     comp~
```

```
#B.
str(mpg)
```

```
## spc_tbl_ [234 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1        : num [1:234] 1 2 3 4 5 6 7 8 9 10 ...
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : num [1:234] 1999 1999 2008 2008 1999 ...
##  $ cyl         : num [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : num [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : num [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   manufacturer = col_character(),
##   ..   model = col_character(),
##   ..   displ = col_double(),
##   ..   year = col_double(),
##   ..   cyl = col_double(),
##   ..   trans = col_character(),
##   ..   drv = col_character(),
##   ..   cty = col_double(),
##   ..   hwy = col_double(),
##   ..   fl = col_character(),
##   ..   class = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#spc_tbl_ [234 × 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
#$ ...1        : num [1:234] 1 2 3 4 5 6 7 8 9 10 ...
#$ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
#$ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
#$ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
#$ year        : num [1:234] 1999 1999 2008 2008 1999 ...
#$ cyl         : num [1:234] 4 4 4 4 6 6 6 4 4 4 ...
#$ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
#$ drv         : chr [1:234] "f" "f" "f" "f" ...
#$ cty         : num [1:234] 18 21 20 21 16 18 18 18 16 20 ...
#$ hwy         : num [1:234] 29 29 31 30 26 26 27 26 25 28 ...
#$ fl          : chr [1:234] "p" "p" "p" "p" ...
#$ class       : chr [1:234] "compact" "compact" "compact" "compact" ...





#C.
#the continuous variables are displ, year, cyl, cty, hwy




#2.

manufacturer <- table(mpg$manufacturer)
manufacturer
```

```
## 
##       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 
```

```
#dodge

library(dplyr)
```

```
## 
## 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
```

```
model <- mpg%>%count(mpg$model)
model
```

```
## # A tibble: 38 x 2
##    `mpg$model`            n
##    <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
#caravan 2wd


#A.
unique_mods <- mpg %>%group_by(manufacturer)%>%distinct(model)
unique_mods
```

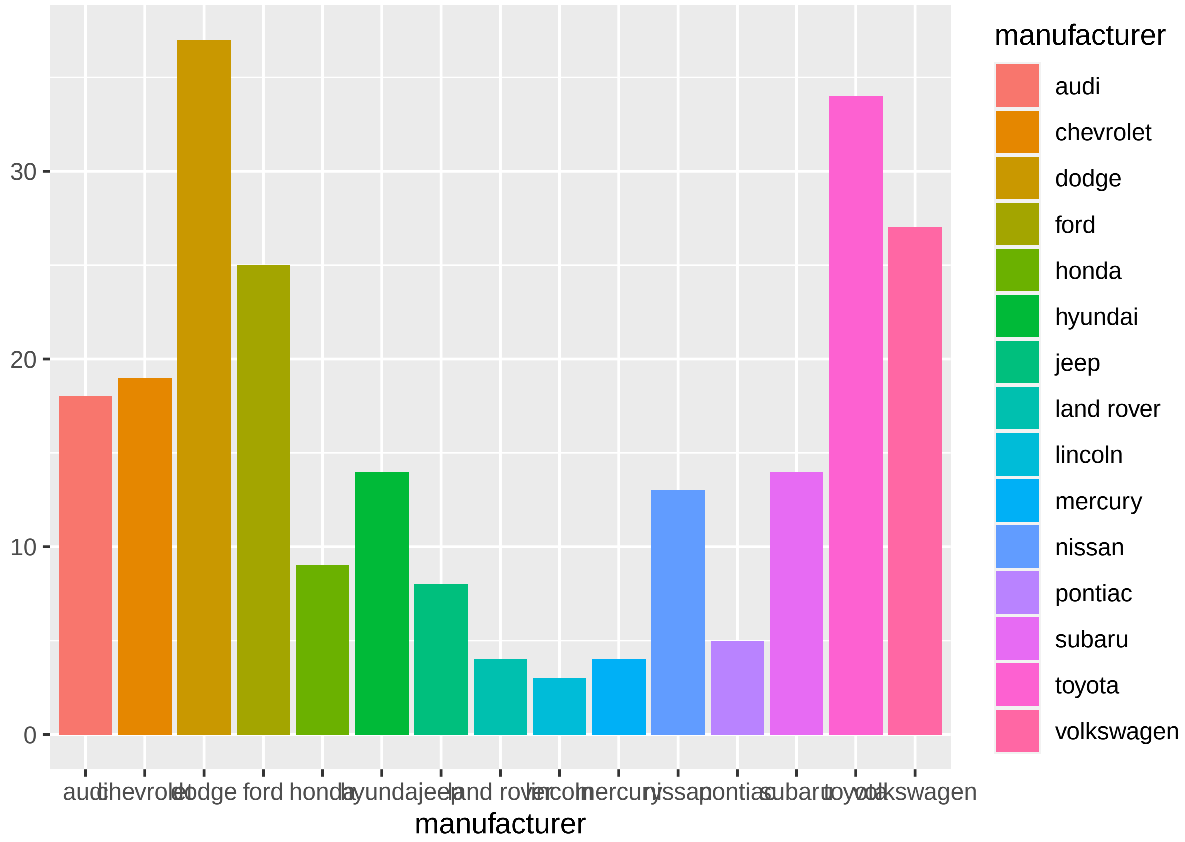
```
## # A tibble: 38 x 2
## # Groups:   manufacturer [15]
##    manufacturer model
##    <chr>        <chr>
##  1 audi         a4
##  2 audi         a4 quattro
##  3 audi         a6 quattro
##  4 chevrolet    c1500 suburban 2wd
##  5 chevrolet    corvette
##  6 chevrolet    k1500 tahoe 4wd
##  7 chevrolet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

```r
#B.
library(ggplot2)
```

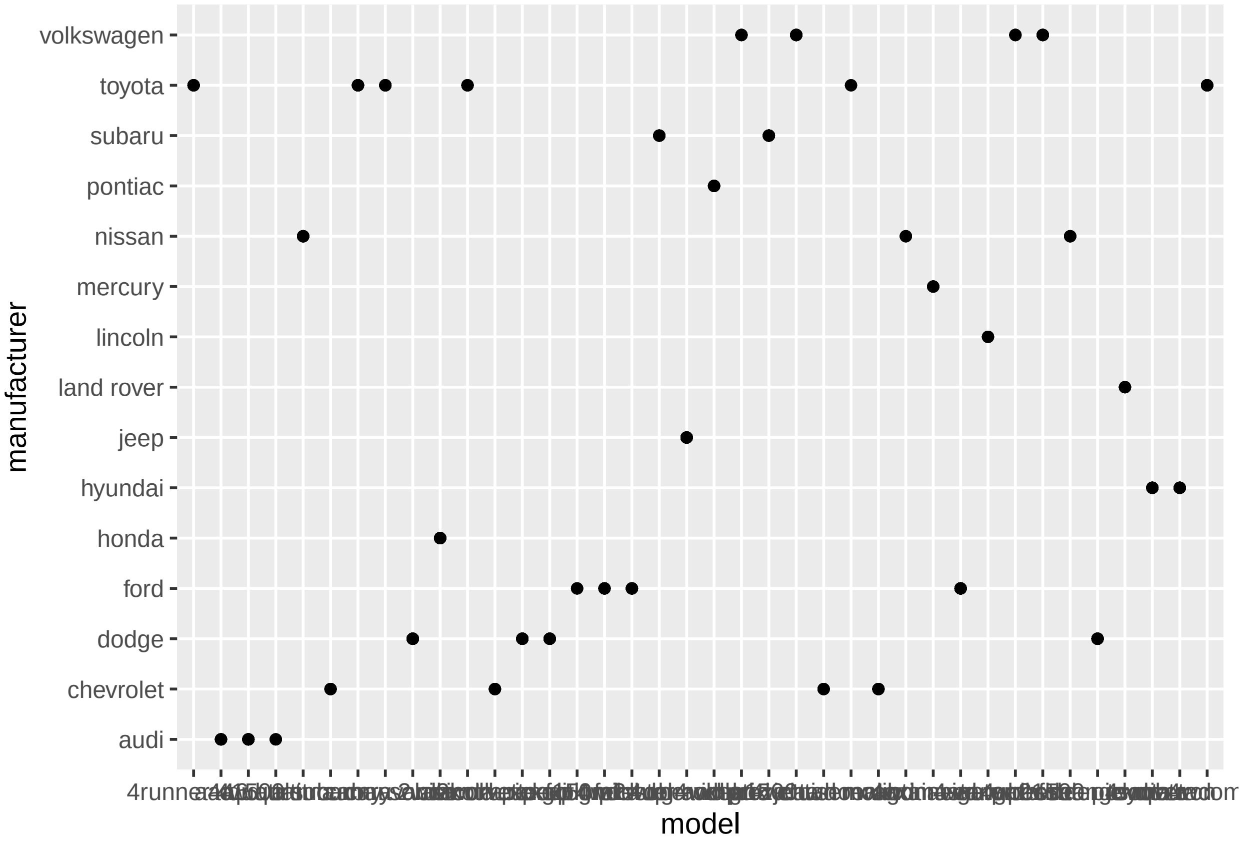
```
##
## Attaching package: 'ggplot2'
##
## The following object is masked _by_ '.GlobalEnv':
##
##     mpg
```

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

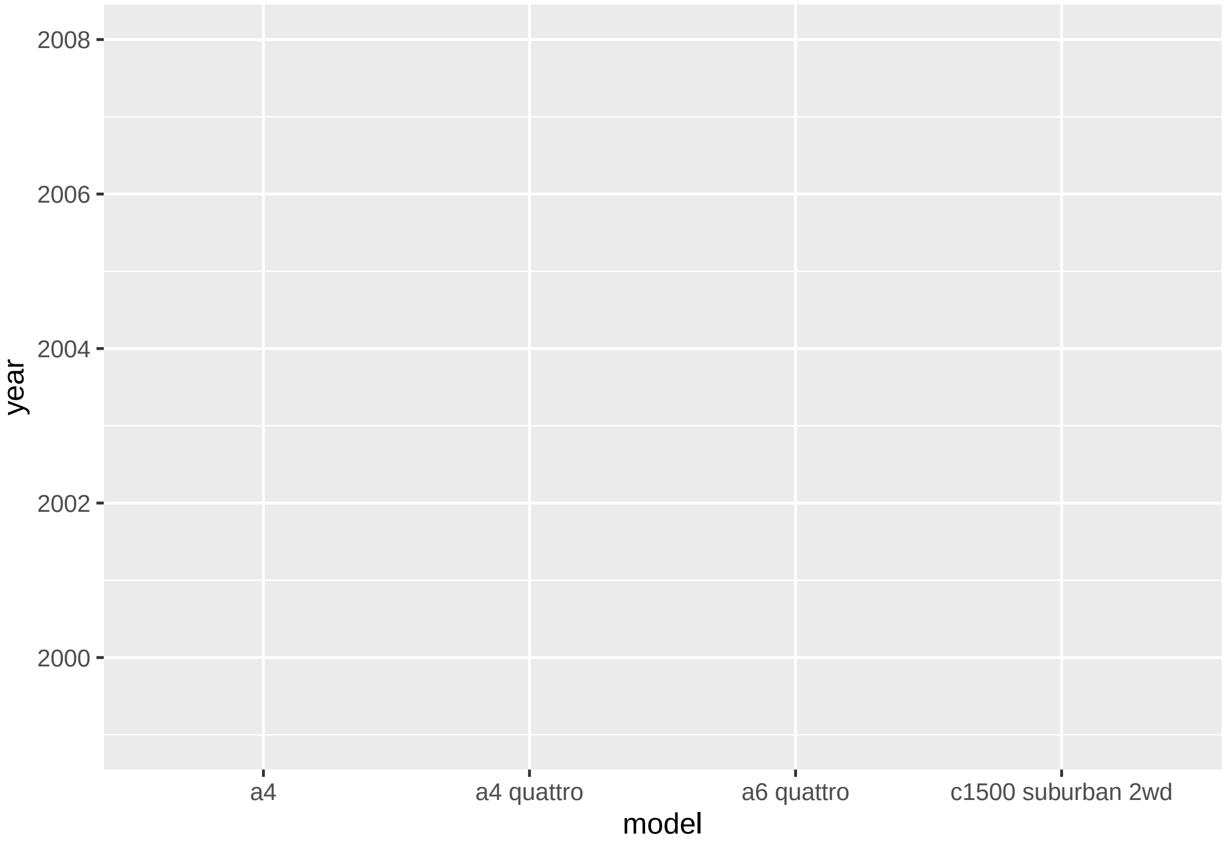
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
#2.part 2

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



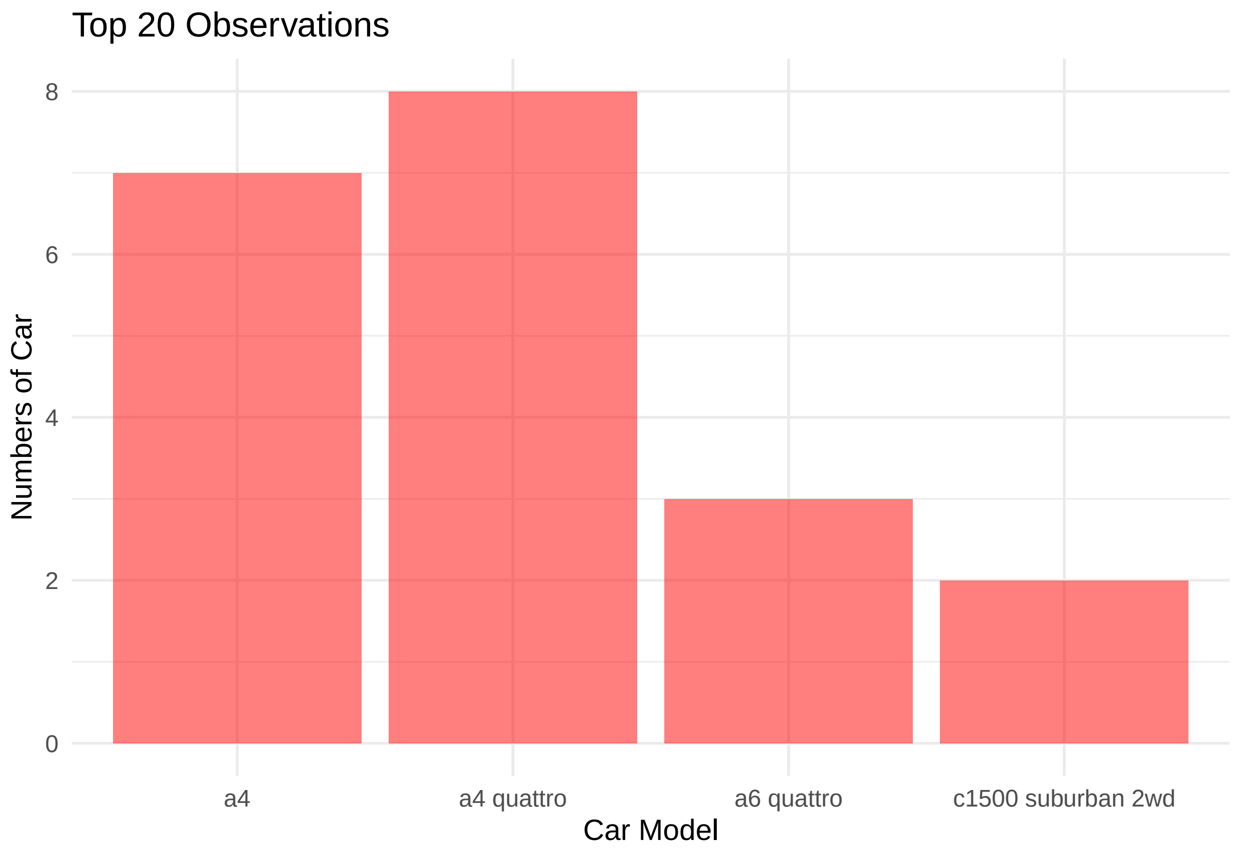
```
#B.
#rates a scatter plot illustrating the relationship between car models and their respective manufacture

#The current plot may lack usefulness due to potential overlap of data points, making it challenging to
#we can add Jitter the Points along the x-axis to alleviate overlap and improve visual clarity.
#aggregate the data to present summaries.


top20 <- head(mpg, 20)
ggplot(top20, aes(x = model, y = year),
       labs(title = "Scatter Plot of Model and Year (Top 20 Observations)",
            x = "Car Model",
            y = "Year"))
```



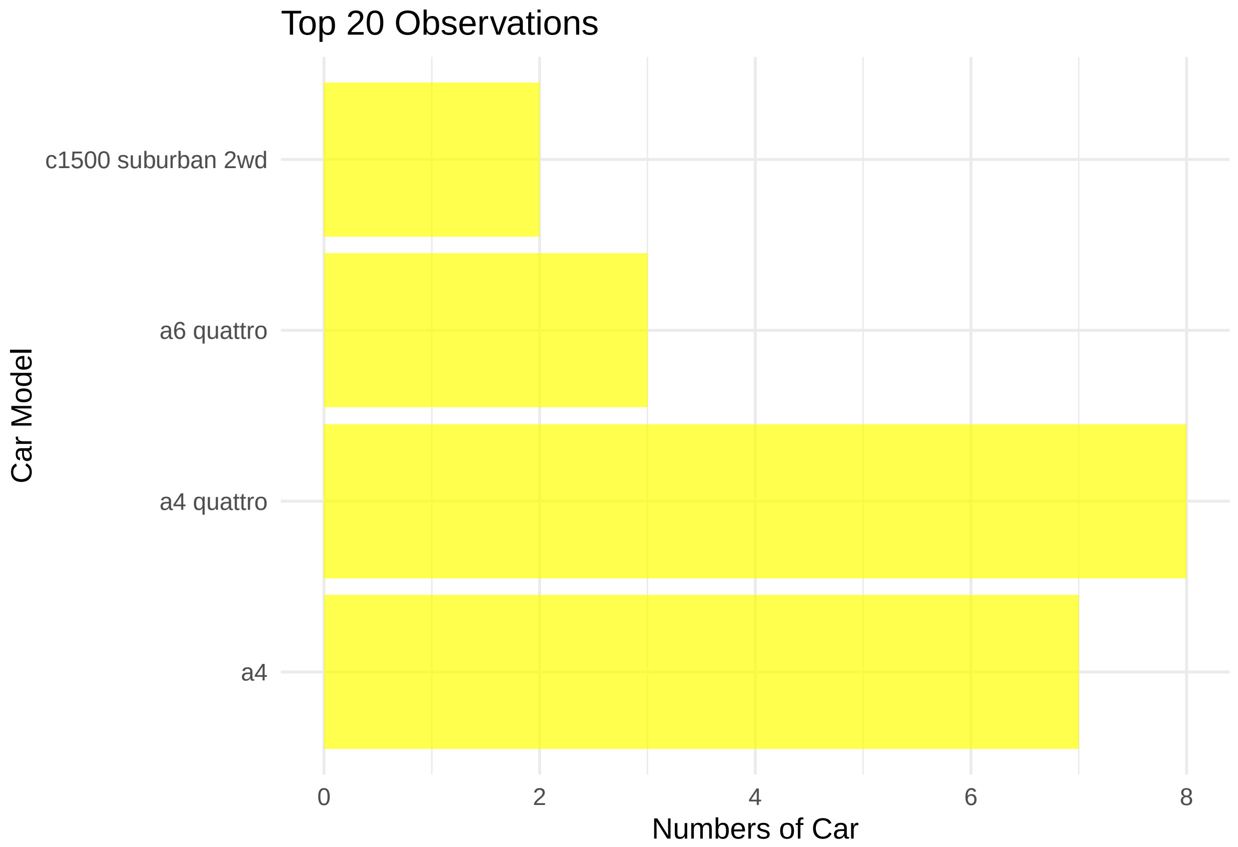
```
#4.
library(dplyr)
group_of_models <- mpg %>%
  group_by(model)%>%
  summarise(number_of_cars = n())
group_of_models
```

```
## # 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
```

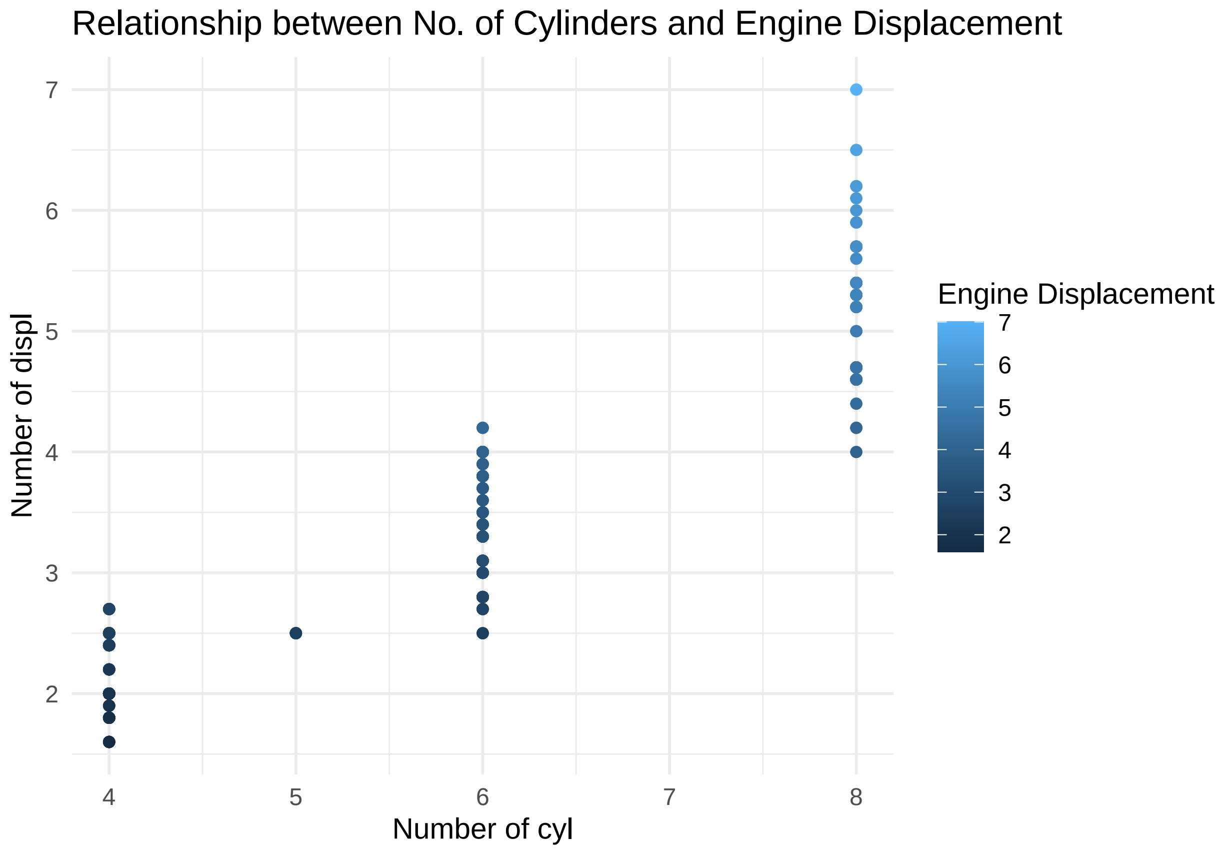
```
#A
ggplot(top20, aes(x = model)) +
  geom_bar(fill = "red", alpha = 0.5) +
  labs(title = "Top 20 Observations",
       x = "Car Model",
       y = "Numbers of Car") +
  theme_minimal()
```



```
#B
ggplot(top20, aes(x = model)) +
  geom_bar(fill = "yellow", alpha = 0.7) +
  labs(title = "Top 20 Observations",
       x = "Car Model",
       y = "Numbers of Car") +
  theme_minimal() +
  coord_flip()
```



```
#5.
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()
```



```
#This scatter plot depicts the distribution of automobile models between manufacturers. Each data point




#6
#A
traffic_data <- read.csv("traffic.csv")
traffic_obv <-nrow(traffic_data)
traffic_obv
```

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

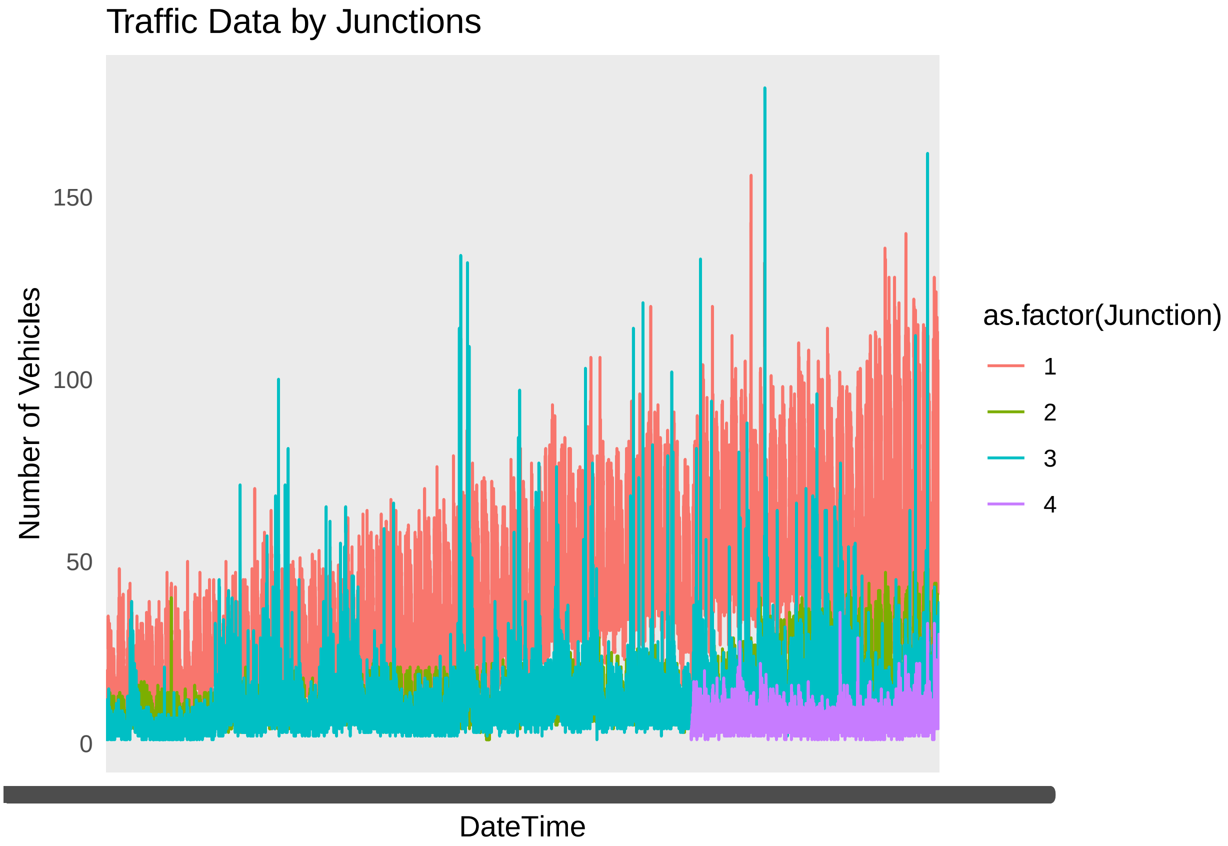
```
str(traffic_data)
```

```
## '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 variables of traffic dataset is DateTime, Junction, Vehicles, and ID.

#B.
junctions_sub <- traffic_data %>%
  select(DateTime, Junction, Vehicles)

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- 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
```

```
ggplot(junctions_sub, aes(x = DateTime, y = Vehicles, color = as.factor(Junction), group = Junction)) +
  geom_line() +
  labs(title = "Traffic Data by Junctions",
       x = "DateTime",
       y = "Number of Vehicles") +
  theme_minimal()
```



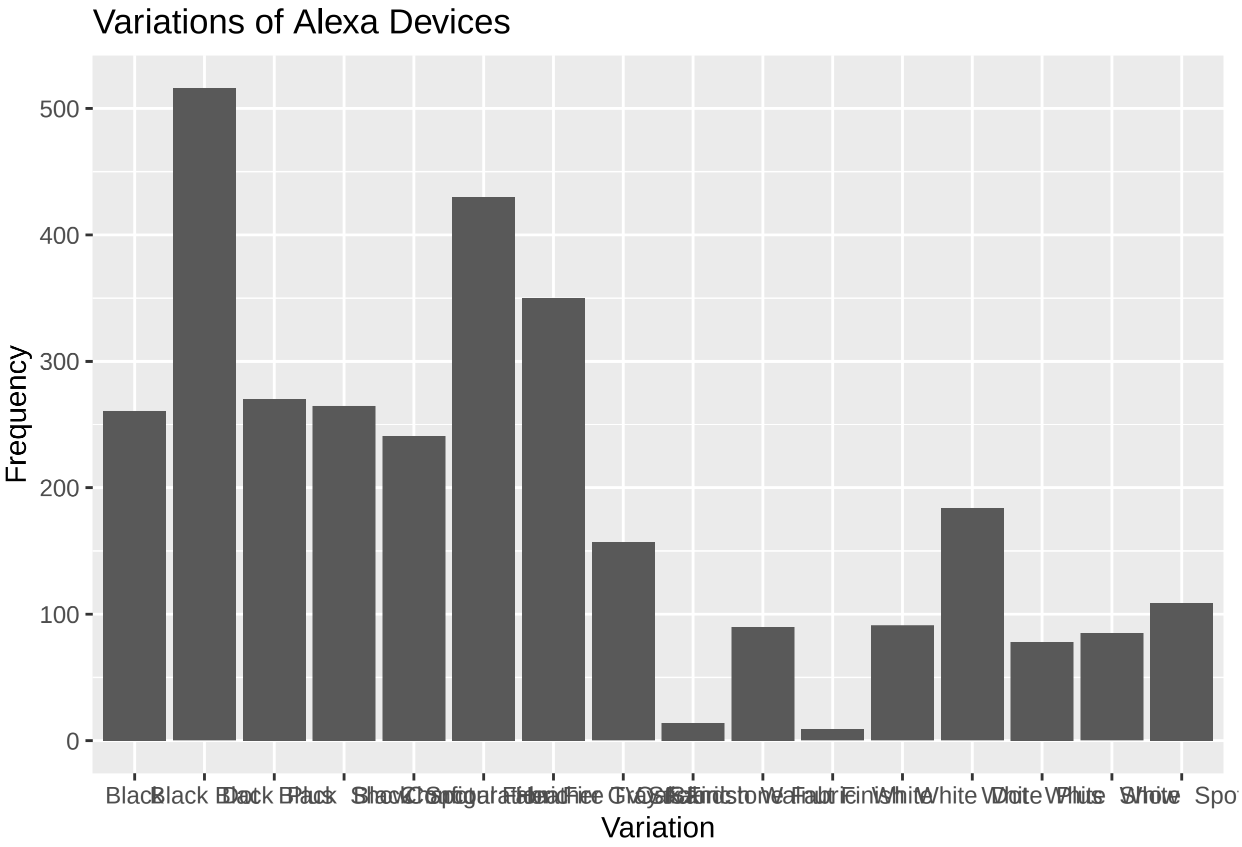
```
#7.
library(readxl)
alexa_file <- read_excel("/cloud/project/RWorksheet_Conlu#4b/alexa_file.xlsx")
#View(alexa_file)


#A.
nrow(alexa_file)
```

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

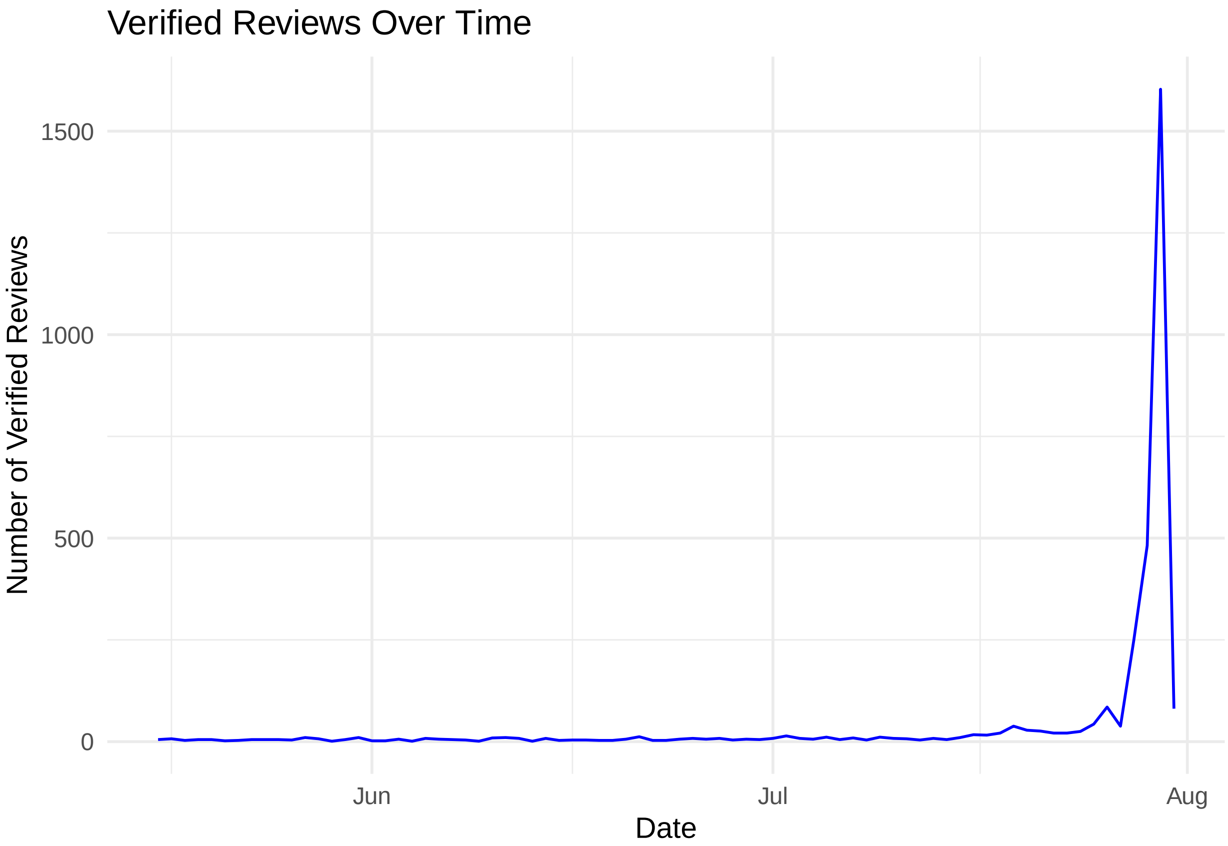
```
ncol(alexa_file)
```

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

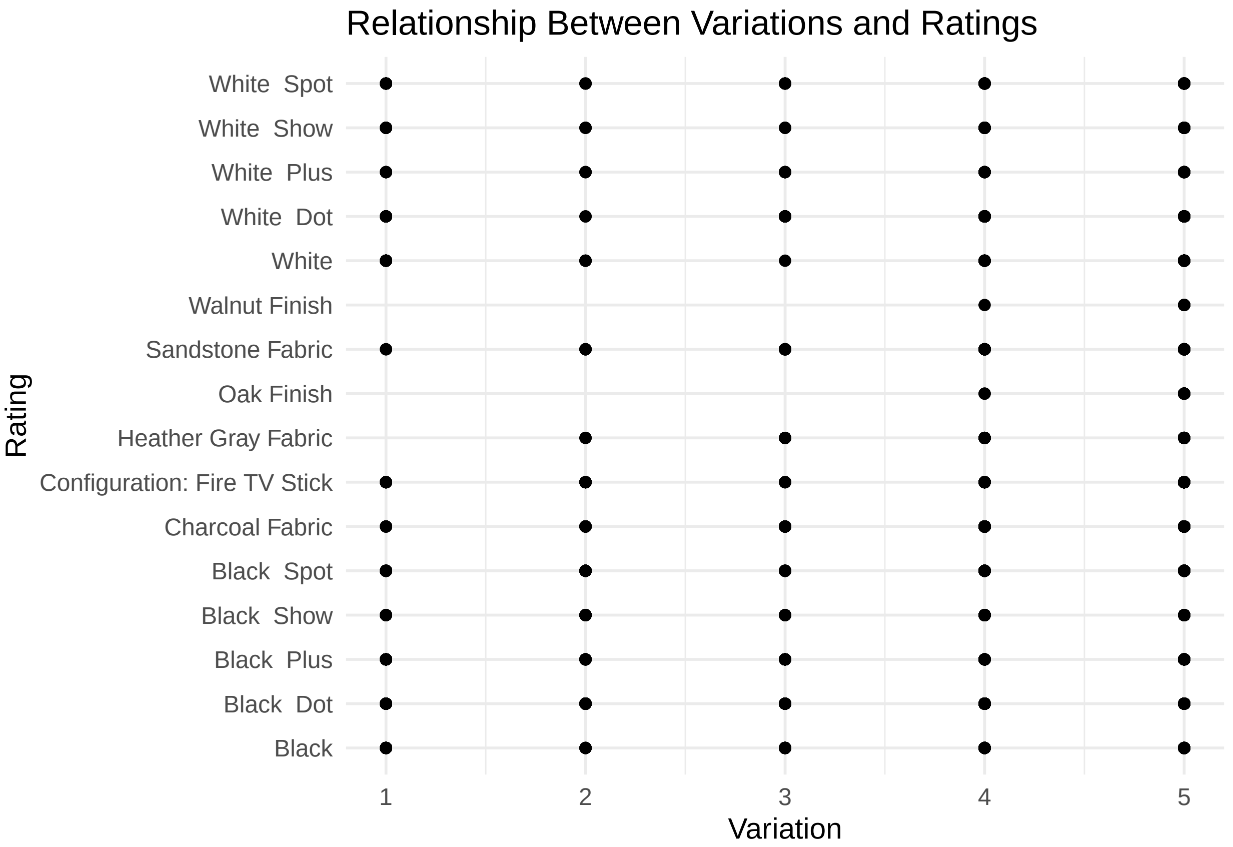
```
#B.
alexaData <- alexa_file%>%
  group_by(variation) %>%
  summarise(Frequency = n())
#View(alexaData)


#C
library(dplyr)
ggplot(alexaData, aes(x = variation, y = Frequency )) +
  geom_bar(stat = "identity") +
  labs(
    title = "Variations of Alexa Devices",
    x = "Variation",
    y = "Frequency")
```



```
#Each bar represents a different variant, and its height shows how frequently it appears in the data.

#D.

summary_reviews <- alexa_file %>%
  group_by(date) %>%
  summarize(NumVerifiedReviews = n())

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()
```



```
#E.
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
```