RWorksheet_Tubat#4c

2023-11-21

```
library(readxl)
```

```
mpgDataset <- read.csv(file = "/cloud/project/worksheet#4/mpg.csv")
```

```
str(mpgDataset)
```

```
## '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" ...
```

```
#4b. The variables that are categorical are: manufacturer, model, trans, drive, class, fl, year, cyl
```

```
#The variables that are continuous are displ, cty, hwy.
class(mpgDataset$hwy)
```

```
## [1] "integer"
```

```
uniqueManufacturer <- mpgDataset$manufacturer
numCars <- table(mpgDataset$manufacturer)
numCars
```

```
##
##       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
```

```
uniqueModels <- unique(mpgDataset$model)
uniqueModels
```

```
##  [1] "a4"                    "a4 quattro"            "a6 quattro"
##  [4] "c1500 suburban 2wd"    "corvette"              "k1500 tahoe 4wd"
##  [7] "malibu"                "caravan 2wd"           "dakota pickup 4wd"
## [10] "durango 4wd"           "ram 1500 pickup 4wd"   "expedition 2wd"
## [13] "explorer 4wd"          "f150 pickup 4wd"       "mustang"
## [16] "civic"                 "sonata"                "tiburon"
```

```
## [19] "grand cherokee 4wd"     "range rover"            "navigator 2wd"
## [22] "mountaineer 4wd"        "altima"                 "maxima"
## [25] "pathfinder 4wd"         "grand prix"             "forester awd"
## [28] "impreza awd"            "4runner 4wd"            "camry"
## [31] "camry solara"           "corolla"                "land cruiser wagon 4wd"
## [34] "toyota tacoma 4wd"      "gti"                    "jetta"
## [37] "new beetle"             "passat"
```

```
library(dplyr)
```

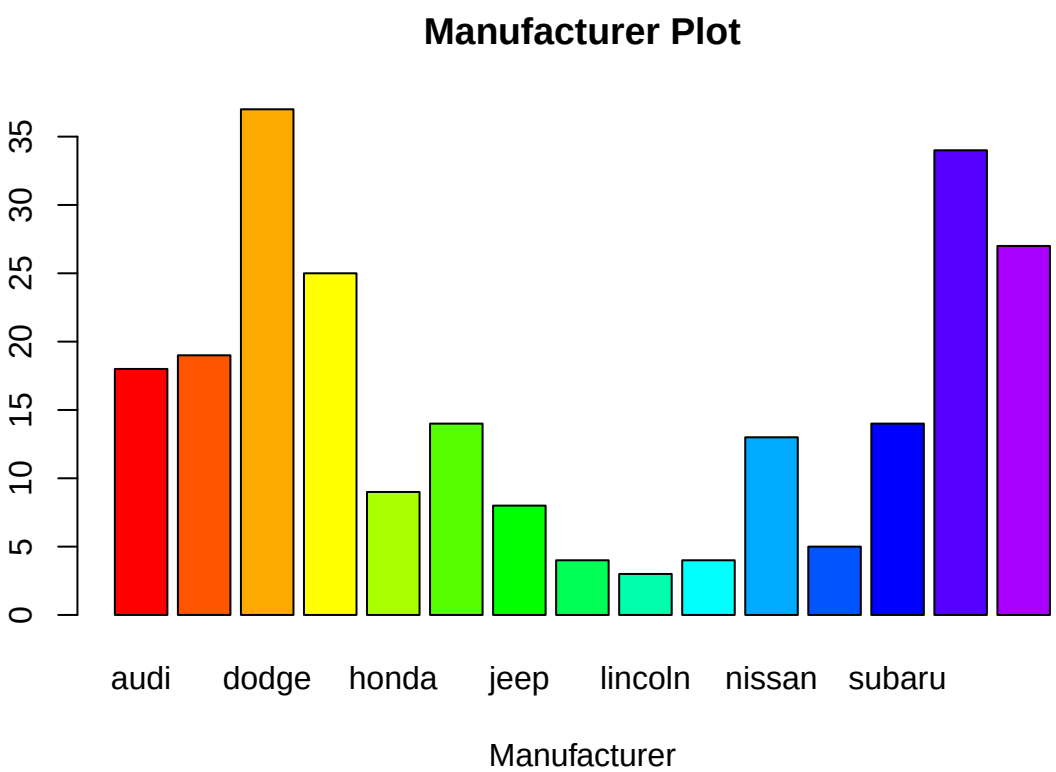
```
##
## Attaching package: 'dplyr'
```

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

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

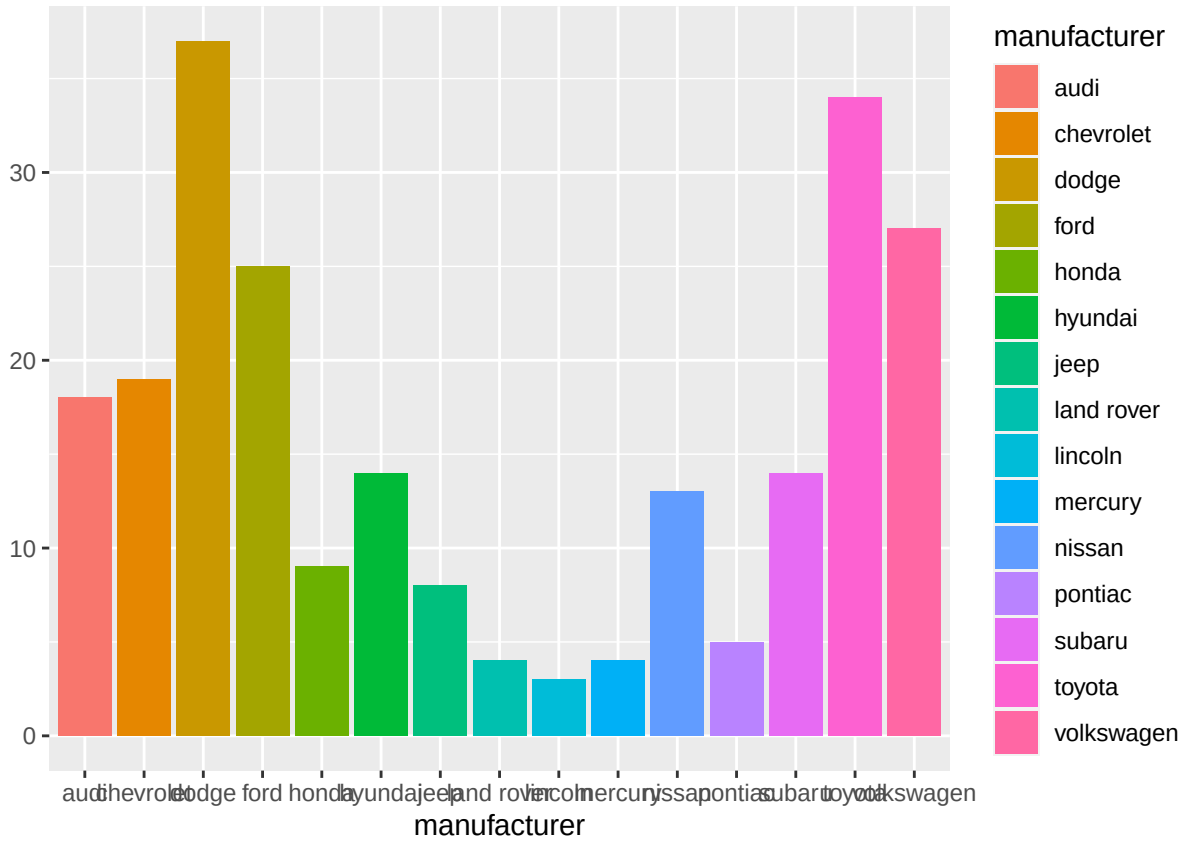
```
library(ggplot2)
```

```
barplot(numCars, col = rainbow(18), xlab = "Manufacturer", main = "Manufacturer Plot")
```

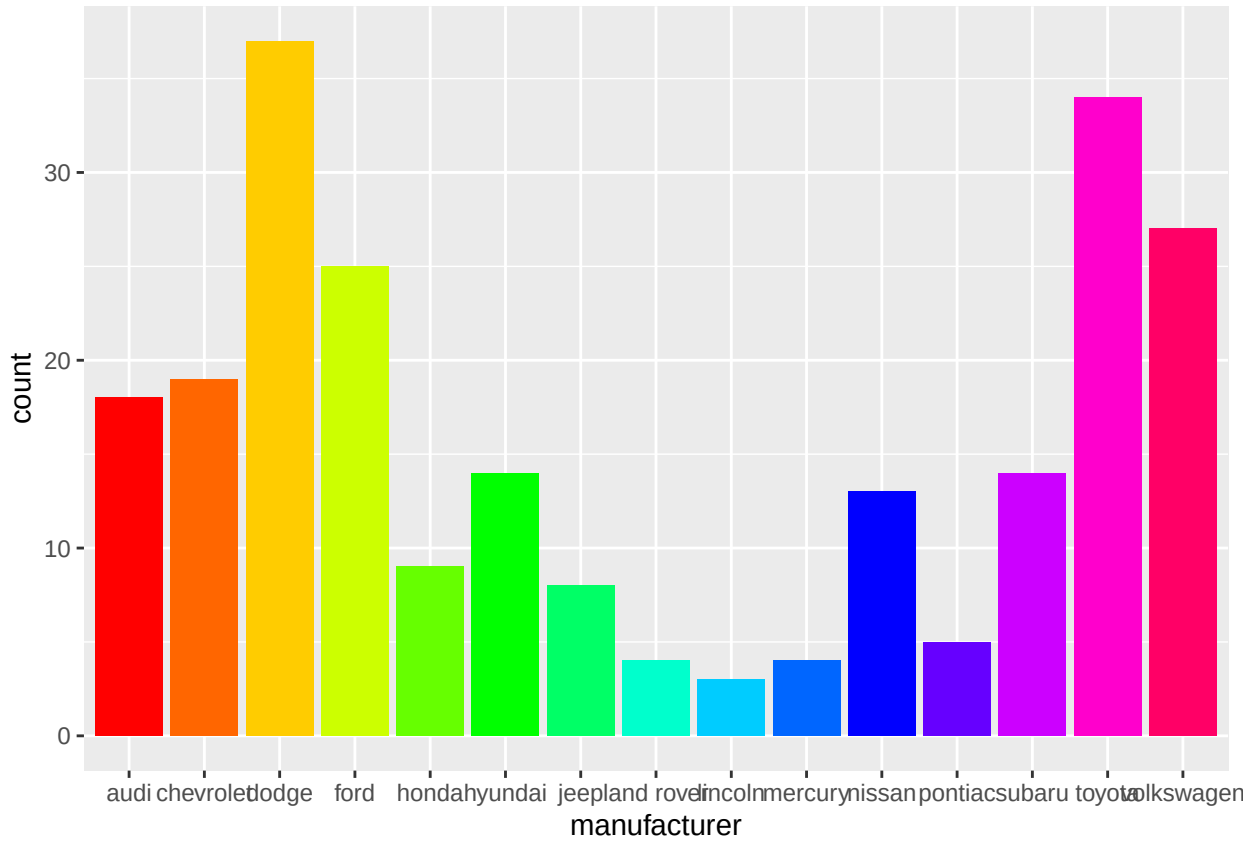



```
qplot(manufacturer, data = mpg, geom = "bar", fill = manufacturer)
```

```
## 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.
```



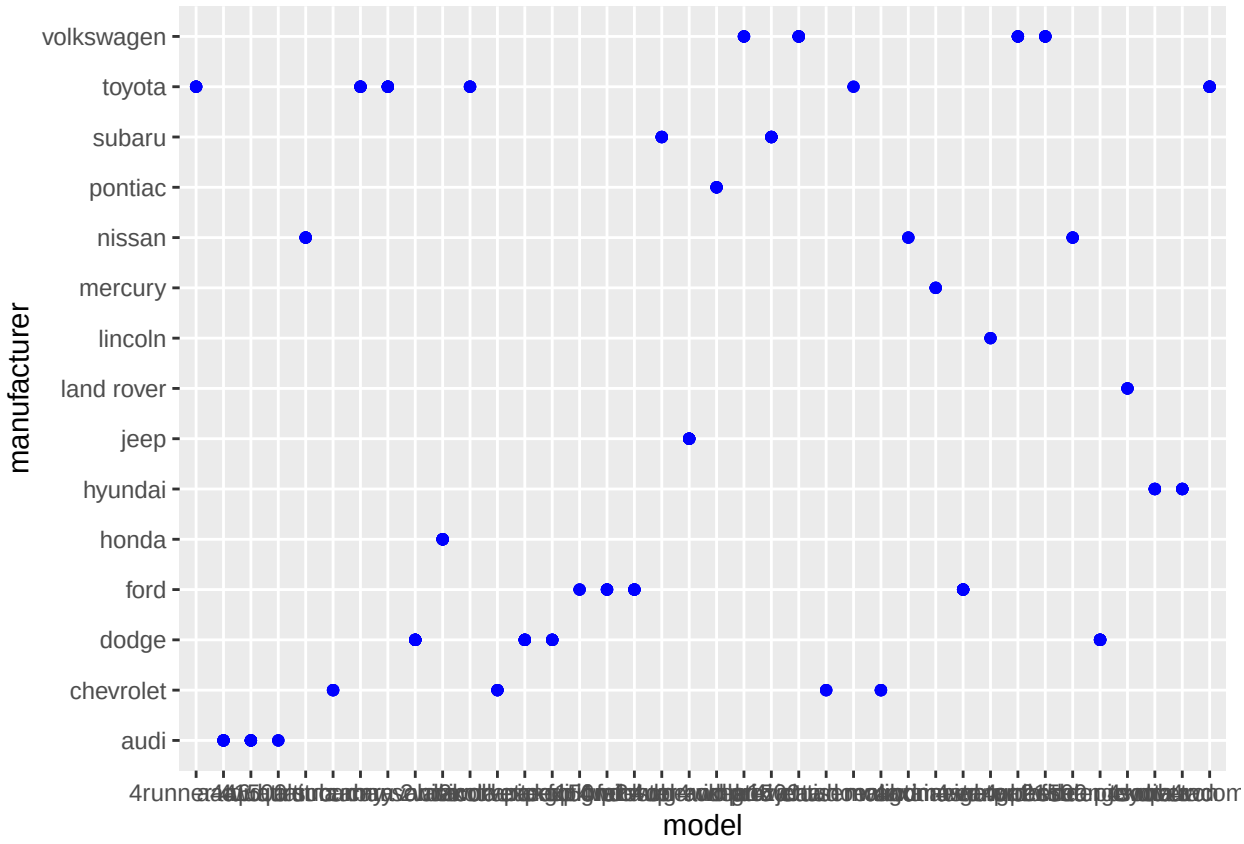

```
ggplot(mpg, aes(manufacturer), fill = manufacturer)+
  geom_bar(fill = rainbow(15))
```




```
groupedManufacturer <- group_by(mpgDataset, manufacturer)
summarize(groupedManufacturer)
```

```
## # A tibble: 15 x 1
##    manufacturer
##    <chr>
##  1 audi
##  2 chevrolet
##  3 dodge
##  4 ford
##  5 honda
##  6 hyundai
##  7 jeep
##  8 land rover
##  9 lincoln
## 10 mercury
## 11 nissan
## 12 pontiac
## 13 subaru
## 14 toyota
## 15 volkswagen
```

```
ggplot(mpg,aes(x =model, y = manufacturer)) +
geom_point(color = "blue")
```




```
#It shows a scatterplot with the manufacturer on the y axis and model type at the x axis

#2b.
#The data shown in this scatterplot is not useful, In order to make it more useful I would put the manu
```

```
newModel <- head(mpg$model, 20)
newModel
```

```
##  [1] "a4"                 "a4"                 "a4"
##  [4] "a4"                 "a4"                 "a4"
##  [7] "a4"                 "a4 quattro"         "a4 quattro"
## [10] "a4 quattro"         "a4 quattro"         "a4 quattro"
## [13] "a4 quattro"         "a4 quattro"         "a4 quattro"
## [16] "a6 quattro"         "a6 quattro"         "a6 quattro"
## [19] "c1500 suburban 2wd" "c1500 suburban 2wd"
```

```
newYear <- head(mpg$year, 20)
newYear
```

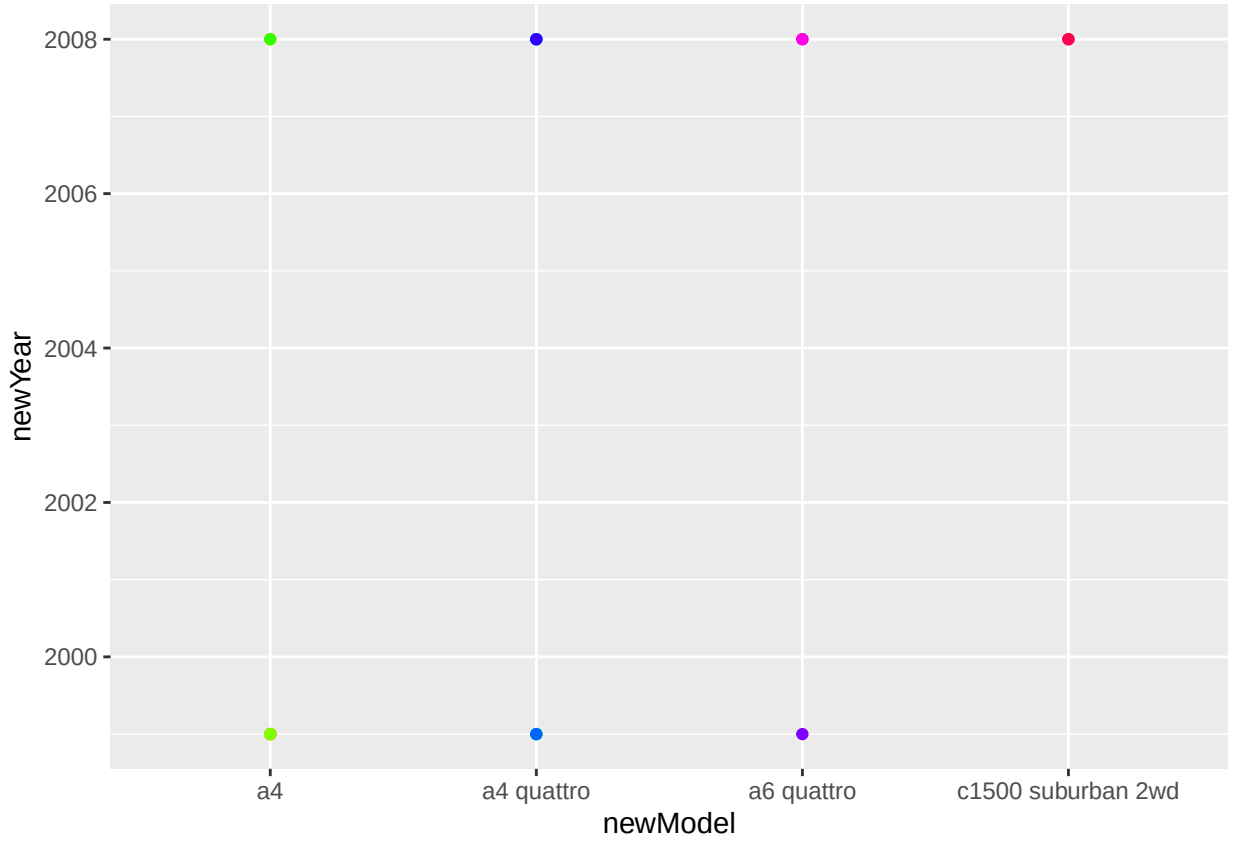
```
##  [1] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 2008 1999 1999 2008 2008
## [16] 1999 2008 2008 2008 2008
```

```
newMpgDF <- data.frame(newModel, newYear)
newMpgDF
```

```
##              newModel newYear
## 1                  a4    1999
## 2                  a4    1999
## 3                  a4    2008
```

```
## 4                  a4    2008
## 5                  a4    1999
## 6                  a4    1999
## 7                  a4    2008
## 8           a4 quattro    1999
## 9           a4 quattro    1999
## 10          a4 quattro    2008
## 11          a4 quattro    2008
## 12          a4 quattro    1999
## 13          a4 quattro    1999
## 14          a4 quattro    2008
## 15          a4 quattro    2008
## 16          a6 quattro    1999
## 17          a6 quattro    2008
## 18          a6 quattro    2008
## 19 c1500 suburban 2wd    2008
## 20 c1500 suburban 2wd    2008
```

```
ggplot(newMpgDF,aes(newModel, newYear))+
  geom_point(color = rainbow(20))
```

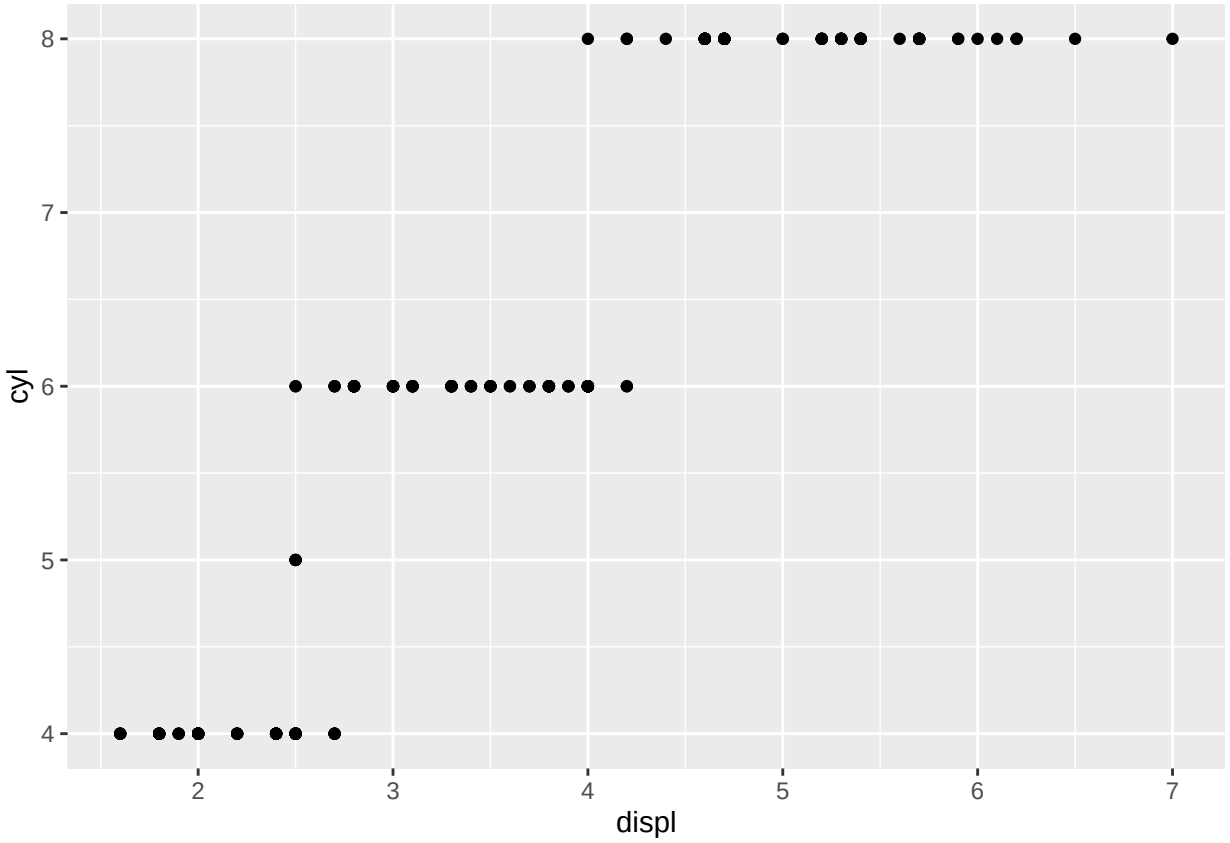



```
group<- mpgDataset %>%
  group_by(model)

summarise(group)
```

```
## # A tibble: 38 x 1
##    model
```

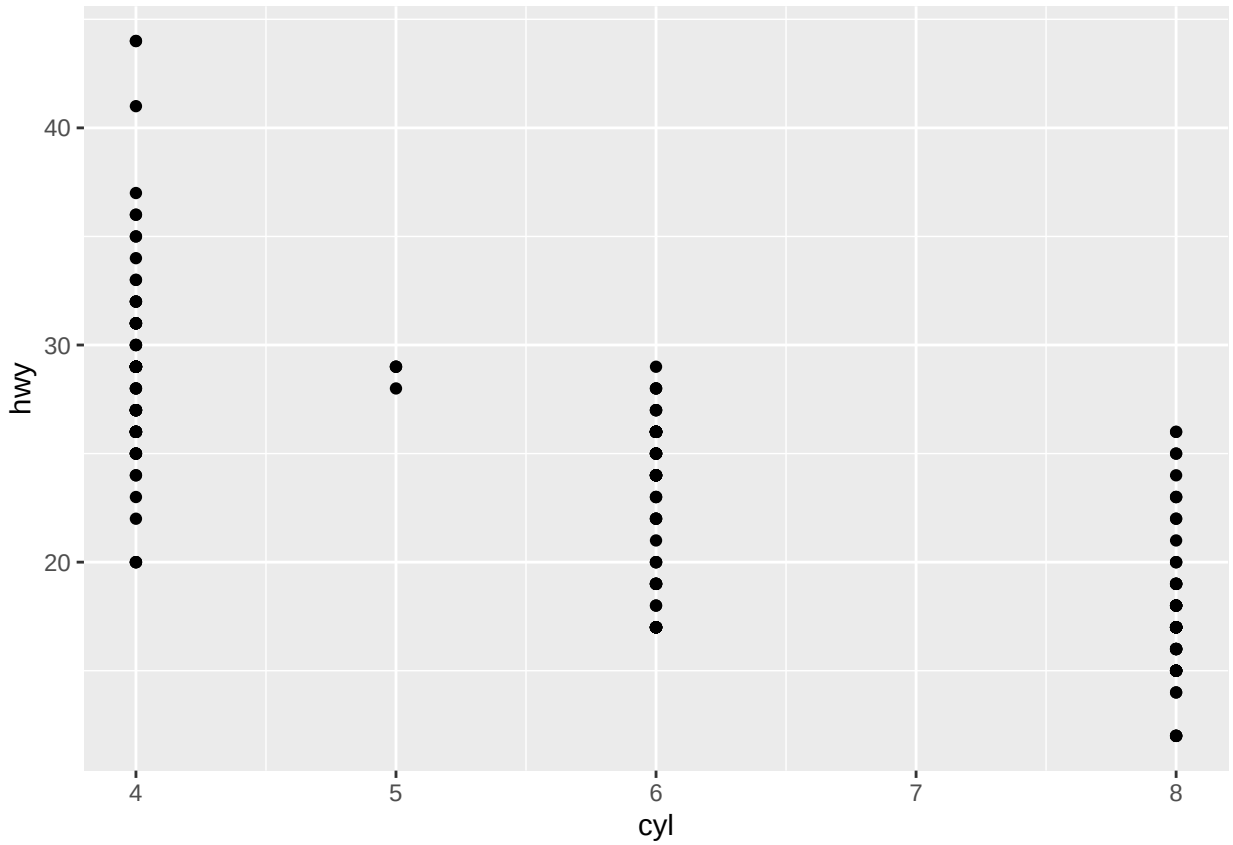
```
##    <chr>
##  1 4runner 4wd
##  2 a4
##  3 a4 quattro
##  4 a6 quattro
##  5 altima
##  6 c1500 suburban 2wd
##  7 camry
##  8 camry solara
##  9 caravan 2wd
## 10 civic
## # i 28 more rows
```

```
ggplot(mpg, aes(displ,cyl))+
  geom_point()
```




```
#The higher number of cylinders, the higher the displacement.
```

```
ggplot(mpg, aes(cyl,hwy))+
  geom_point()
```




```
#The result is that the lower the number of cylinders, the higher the highway mileage it has.
```

```
trafficDS <- read.csv("/cloud/project/worksheet#4/traffic.csv")

dim(trafficDS)
```

```
## [1] 48120     4
```

```
#The traffic dataset has 48120 number of observations
```

```
gpJunction <- group_by(trafficDS, Junction)
gpJunction
```

```
## # A tibble: 48,120 x 4
## # Groups:   Junction [4]
##    DateTime            Junction Vehicles          ID
##    <chr>                  <int>    <int>       <dbl>
##  1 2015-11-01 00:00:00        1       15 20151101001
##  2 2015-11-01 01:00:00        1       13 20151101011
##  3 2015-11-01 02:00:00        1       10 20151101021
##  4 2015-11-01 03:00:00        1        7 20151101031
##  5 2015-11-01 04:00:00        1        9 20151101041
##  6 2015-11-01 05:00:00        1        6 20151101051
##  7 2015-11-01 06:00:00        1        9 20151101061
##  8 2015-11-01 07:00:00        1        8 20151101071
##  9 2015-11-01 08:00:00        1       11 20151101081
## 10 2015-11-01 09:00:00        1       12 20151101091
## # i 48,110 more rows
```

```
summarise(gpJunction)
```

```
## # A tibble: 4 x 1
##   Junction
##      <int>
## 1        1
## 2        2
## 3        3
## 4        4
```

```
ggplot(trafficDS, aes(x = DateTime, y = Vehicles, group = Junction, color = as.factor(Junction))) +
  geom_line() +
  labs(title = "Junction-wise Traffic Plot", x = "Date and Time", y = "Number of Vehicles")
```


```
alexaData <- read_xlsx("alexa_file.xlsx")
alexaData
```

```
## # A tibble: 3,150 x 5
##    rating date                variation          verified_reviews      feedback
##     <dbl> <dttm>              <chr>              <chr>                    <dbl>
##  1      5 2018-07-31 00:00:00 Charcoal Fabric    Love my Echo!                1
##  2      5 2018-07-31 00:00:00 Charcoal Fabric    Loved it!                    1
##  3      4 2018-07-31 00:00:00 Walnut Finish      Sometimes while play~        1
##  4      5 2018-07-31 00:00:00 Charcoal Fabric    I have had a lot of ~        1
##  5      5 2018-07-31 00:00:00 Charcoal Fabric    Music                        1
##  6      5 2018-07-31 00:00:00 Heather Gray Fabric I received the echo ~       1
##  7      3 2018-07-31 00:00:00 Sandstone Fabric   Without having a cel~        1
##  8      5 2018-07-31 00:00:00 Charcoal Fabric    I think this is the ~        1
```

```
##  9      5 2018-07-30 00:00:00 Heather Gray Fabric looks great                  1
## 10      5 2018-07-30 00:00:00 Heather Gray Fabric Love it! I've listen~        1
## # i 3,140 more rows
```

```
dim(alexaData)
```

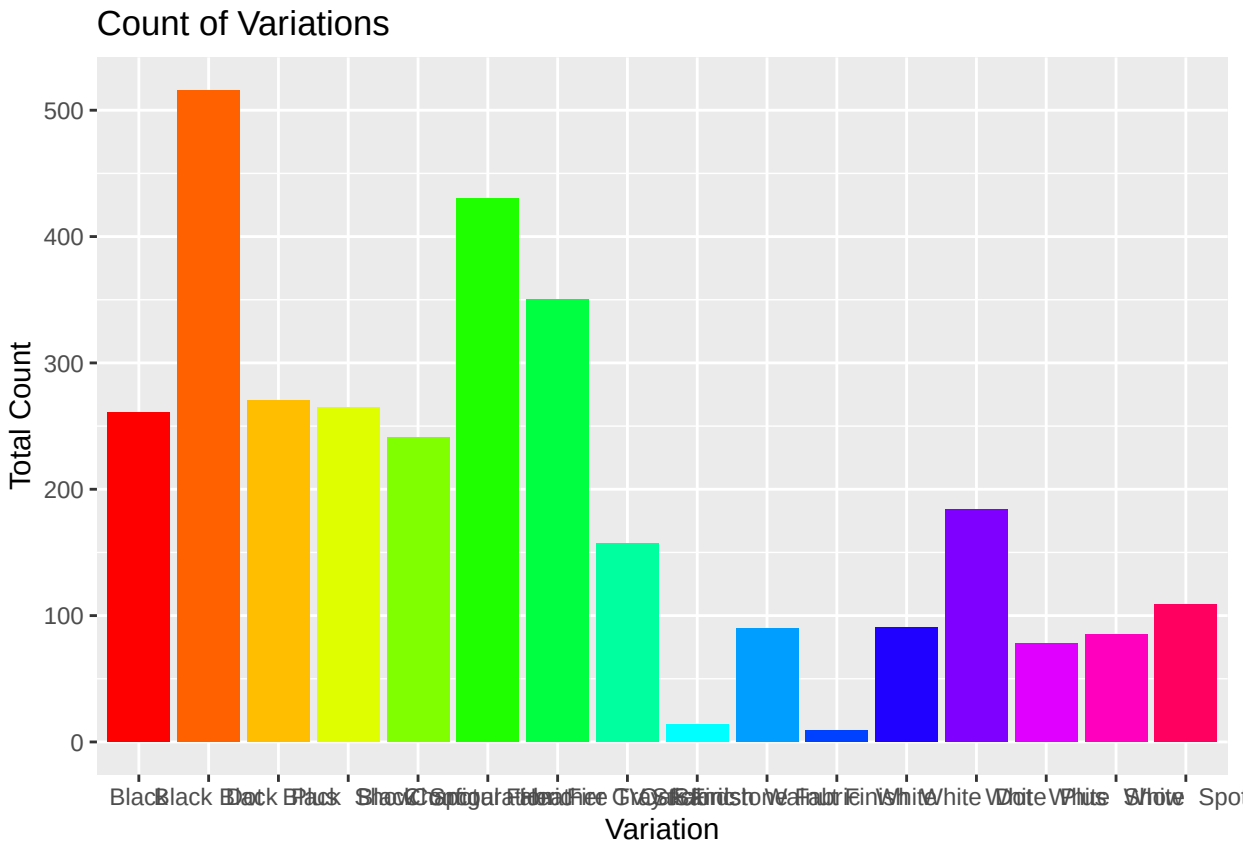
```
## [1] 3150    5
```

```
#Alexa data has 3150 number of observations, and 5 columns.
```

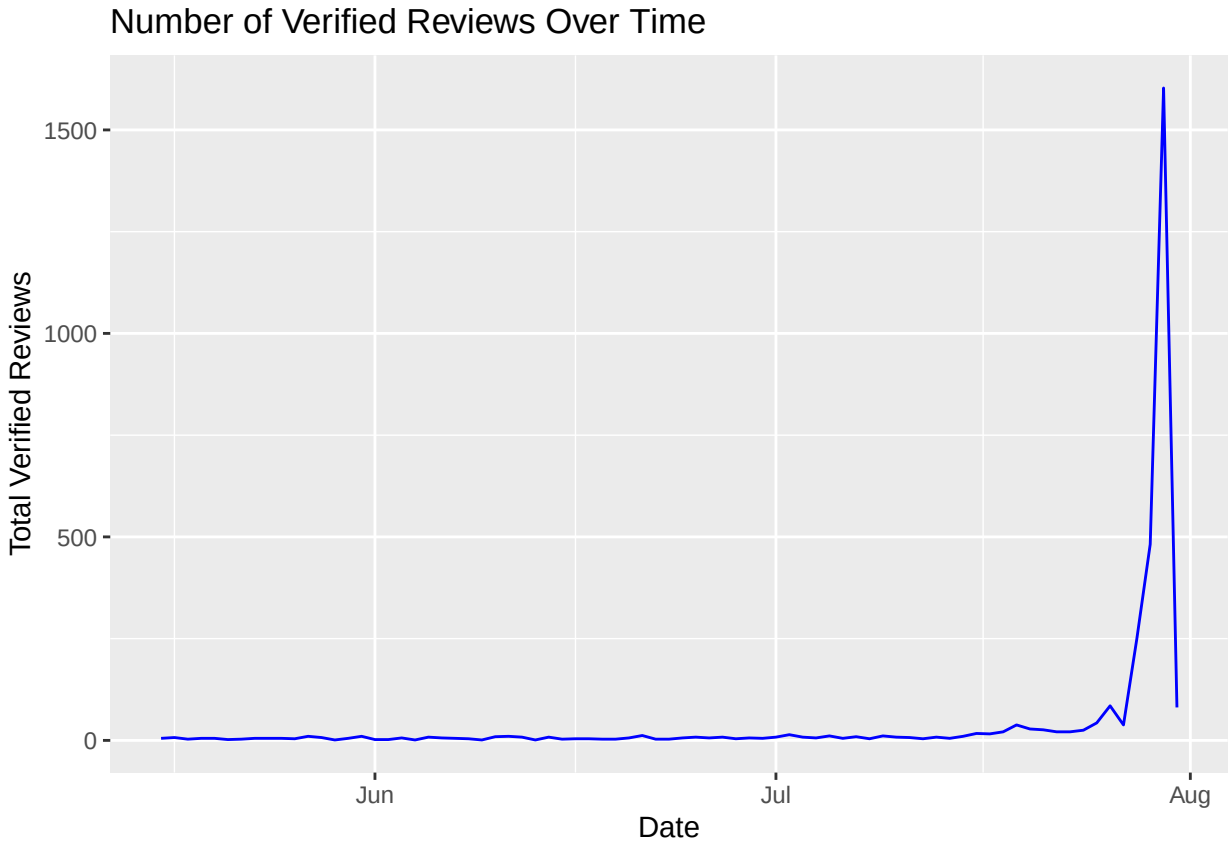
```
variations <- alexaData %>%
  group_by(variation) %>%
  summarise(totalCount = n())

variations
```

```
## # A tibble: 16 x 2
##    variation                    totalCount
##    <chr>                             <int>
##  1 Black                               261
##  2 Black  Dot                          516
##  3 Black  Plus                         270
##  4 Black  Show                         265
##  5 Black  Spot                         241
##  6 Charcoal Fabric                     430
##  7 Configuration: Fire TV Stick        350
##  8 Heather Gray Fabric                 157
##  9 Oak Finish                           14
## 10 Sandstone Fabric                     90
## 11 Walnut Finish                         9
## 12 White                                91
## 13 White  Dot                          184
## 14 White  Plus                          78
## 15 White  Show                          85
## 16 White  Spot                         109
```

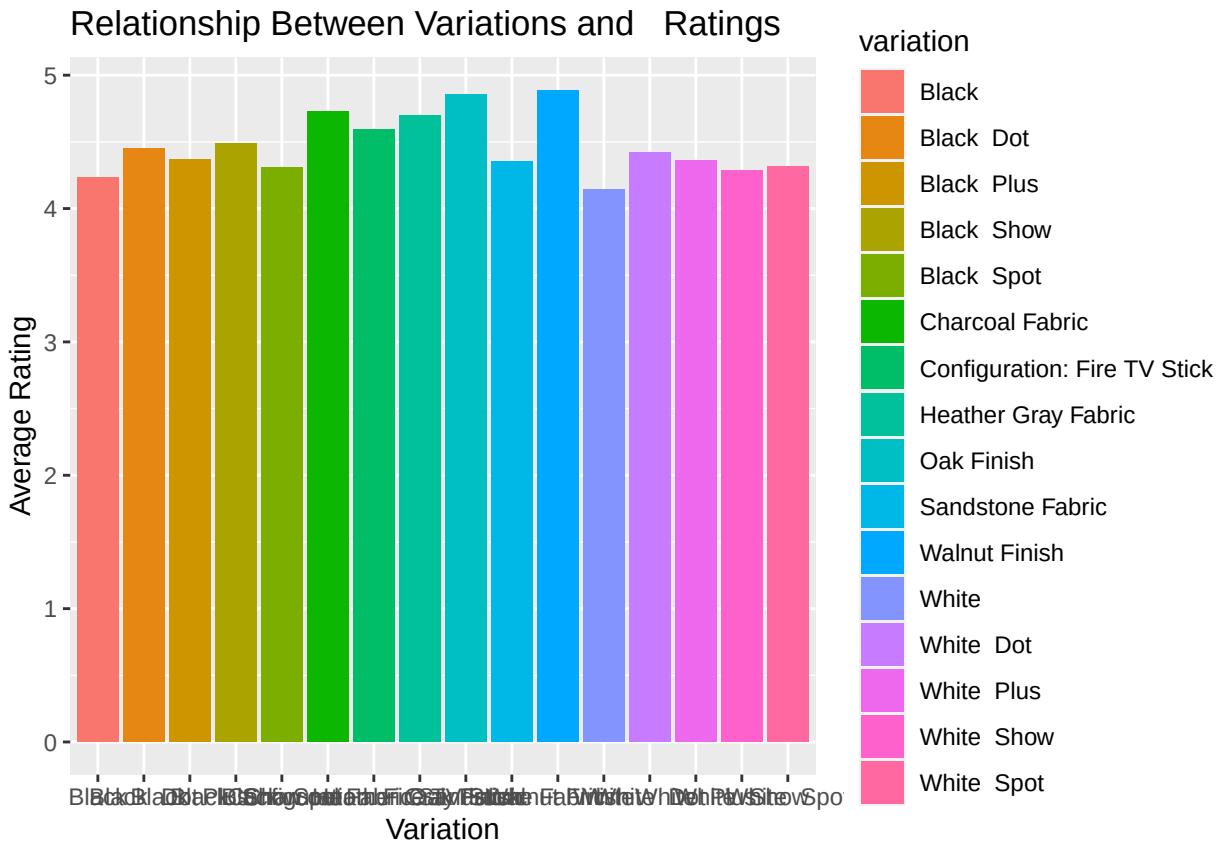
```
ggplot(variations, aes(x = variation, y = totalCount)) +
  geom_bar(stat = "identity", fill = rainbow(16)) +
  labs(title = "Count of Variations", x = "Variation", y = "Total Count")
```




```
alexaData$date <- as.Date(alexaData$date, format ="%d-%b-%y")

reviewDate <- alexaData %>%
  group_by(date) %>%
  summarise(totalReviews = n())

ggplot(reviewDate, aes(x = date, y = totalReviews)) +
  geom_line(color = "blue") +
  labs(title = "Number of Verified Reviews Over Time",
       x = "Date",
       y = "Total Verified Reviews")
```




```r
alexaData$rating <- as.numeric(alexaData$rating)

ratingsVariation <- alexaData %>%
  group_by(variation) %>%
  summarise(average_rating = mean(rating, na.rm = TRUE))

ggplot(ratingsVariation, aes(x = variation, y = average_rating, fill = variation)) +
  geom_bar(stat = "identity") +
  labs(title = "Relationship Between Variations and   Ratings", x = "Variation", y = "Average Rating")
```




```
#The variation with the highest ratings is the Walnut Finish.
```