RWorksheet_Ceniza#4c

Zydrick Ceniza

2023-11-21

RWorksheet_Ceniza#4c

1. Use the dataset mpg

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

```
library(readr)
mpg <- read_csv("mpg.csv")
```

```
## New names:
## Rows: 234 Columns: 12
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (6): manufacturer, model, trans, drv, fl, class dbl (6): ...1, displ, year,
## cyl, cty, hwy
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
mpg
```

```
## # A tibble: 234 x 12
##     ...1 manufacturer model      displ  year   cyl trans drv     cty   hwy fl
##    <dbl> <chr>        <chr>      <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <chr>
##  1     1 audi         a4           1.8  1999     4 auto~ f        18    29 p
##  2     2 audi         a4           1.8  1999     4 manu~ f        21    29 p
##  3     3 audi         a4           2    2008     4 manu~ f        20    31 p
##  4     4 audi         a4           2    2008     4 auto~ f        21    30 p
##  5     5 audi         a4           2.8  1999     6 auto~ f        16    26 p
##  6     6 audi         a4           2.8  1999     6 manu~ f        18    26 p
##  7     7 audi         a4           3.1  2008     6 auto~ f        18    27 p
##  8     8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p
##  9     9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p
## 10    10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p
## # i 224 more rows
## # i 1 more variable: class <chr>
```

b. Which variables from mpg dataset are categorical?

```
categorical <- sapply(mpg, function(x) is.factor(x) || is.character(x))
cat("Categorical Variables:", names(mpg)[categorical])
```

```
## Categorical Variables: manufacturer model trans drv fl class
```

c. Which are continuous variables?

```
continuous <- sapply(mpg, function(x) is.numeric(x) && !is.factor(x) && !is.character(x))
cat("Continuous Variables:", names(mpg)[continuous])
```

```
## Continuous Variables: ...1 displ year cyl cty hwy
```

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

Show your answer.The manufacturer has the most model is the dodge and caravan 2wd has the most models in the manufacturer of dodge

```
md<-factor(mpg$model)
summary(md)
```

```
##            4runner 4wd                   a4            a4 quattro
##                      6                    7                     8
##             a6 quattro                altima     c1500 suburban 2wd
##                      3                    6                     5
##                  camry           camry solara           caravan 2wd
##                      7                    7                    11
##                  civic               corolla              corvette
##                      9                    5                     5
##      dakota pickup 4wd            durango 4wd         expedition 2wd
##                      9                    7                     3
##           explorer 4wd        f150 pickup 4wd           forester awd
##                      6                    7                     6
##     grand cherokee 4wd             grand prix                   gti
##                      8                    5                     5
##            impreza awd                  jetta        k1500 tahoe 4wd
##                      8                    9                     4
## land cruiser wagon 4wd                malibu                maxima
##                      2                    5                     3
##        mountaineer 4wd                mustang          navigator 2wd
##                      4                    9                     3
##             new beetle                passat         pathfinder 4wd
##                      6                    7                     4
##    ram 1500 pickup 4wd           range rover                sonata
##                     10                    4                     7
##                tiburon      toyota tacoma 4wd
##                      7                    7
```

```
cat("The most models is the caravan 2wd that has:",max(summary(md)))
```

```
## The most models is the caravan 2wd that has: 11
```

```
num1a<-factor(mpg$manufacturer)
num1a1<-summary(num1a)
num1a1
```

```
##       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
```

```
cat("The model that has the most variation is the dodge that have:",max(num1a1))
```

```
## The model that has the most variation is the dodge that have: 37
```

a. Group the manufacturers and find the unique models. Show your codes and result

```
num1a<-factor(mpg$manufacturer)
num1a1<-summary(num1a)
num1a1
```

```
##       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
```

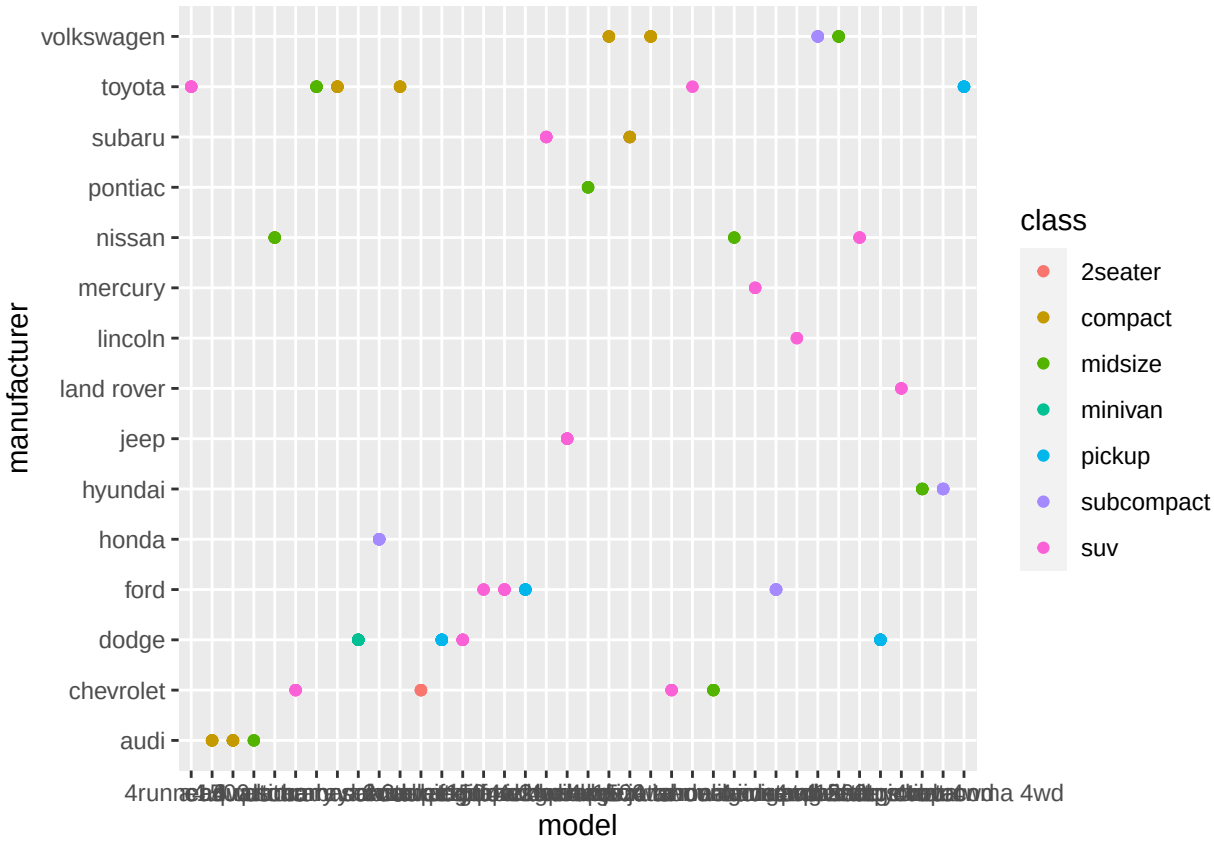
b. Graph the result by using plot() and ggplot(). Write the codes and its result.

```
library(ggplot2)
```

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

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

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

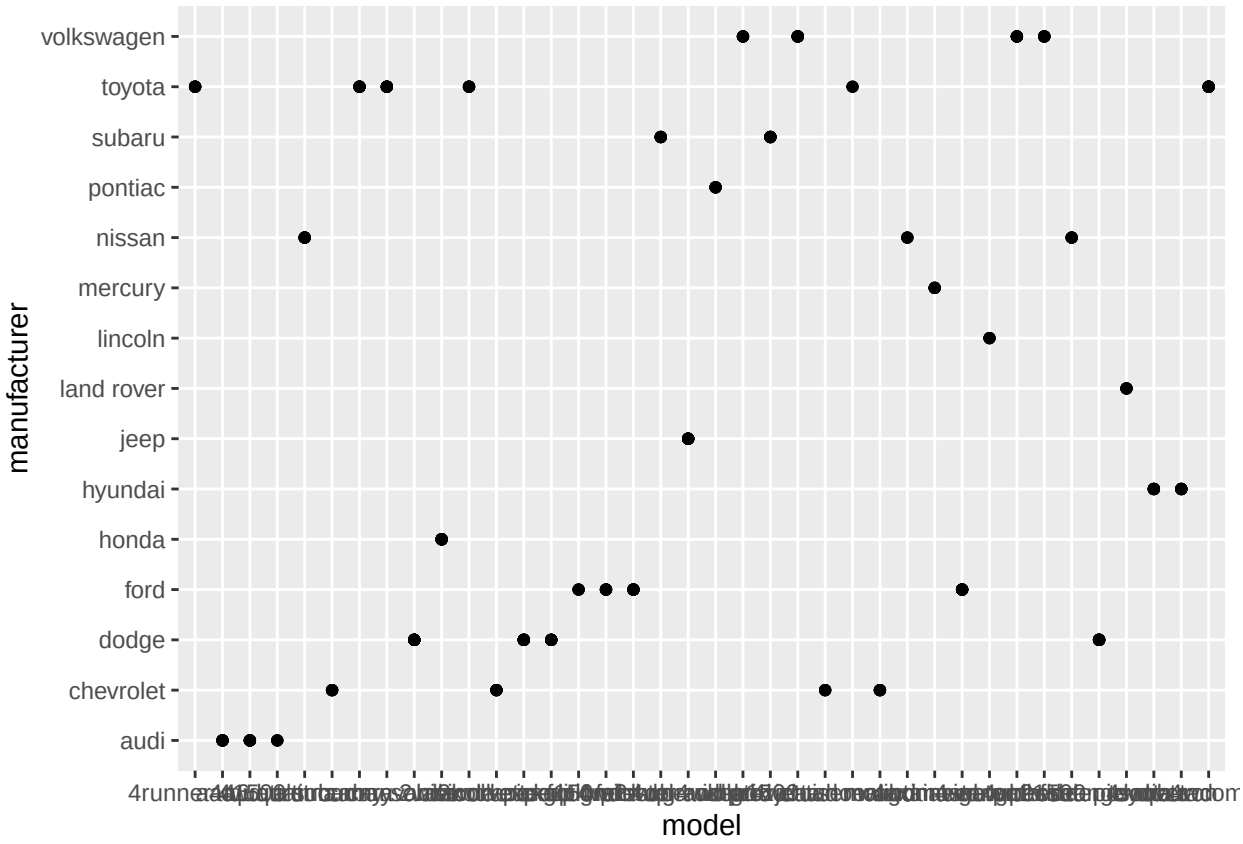


Same dataset will be used. You are going to show the relationship of the model and the manufacturer.

a. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?

The plot shows the model and manufacturer in a black and white color

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

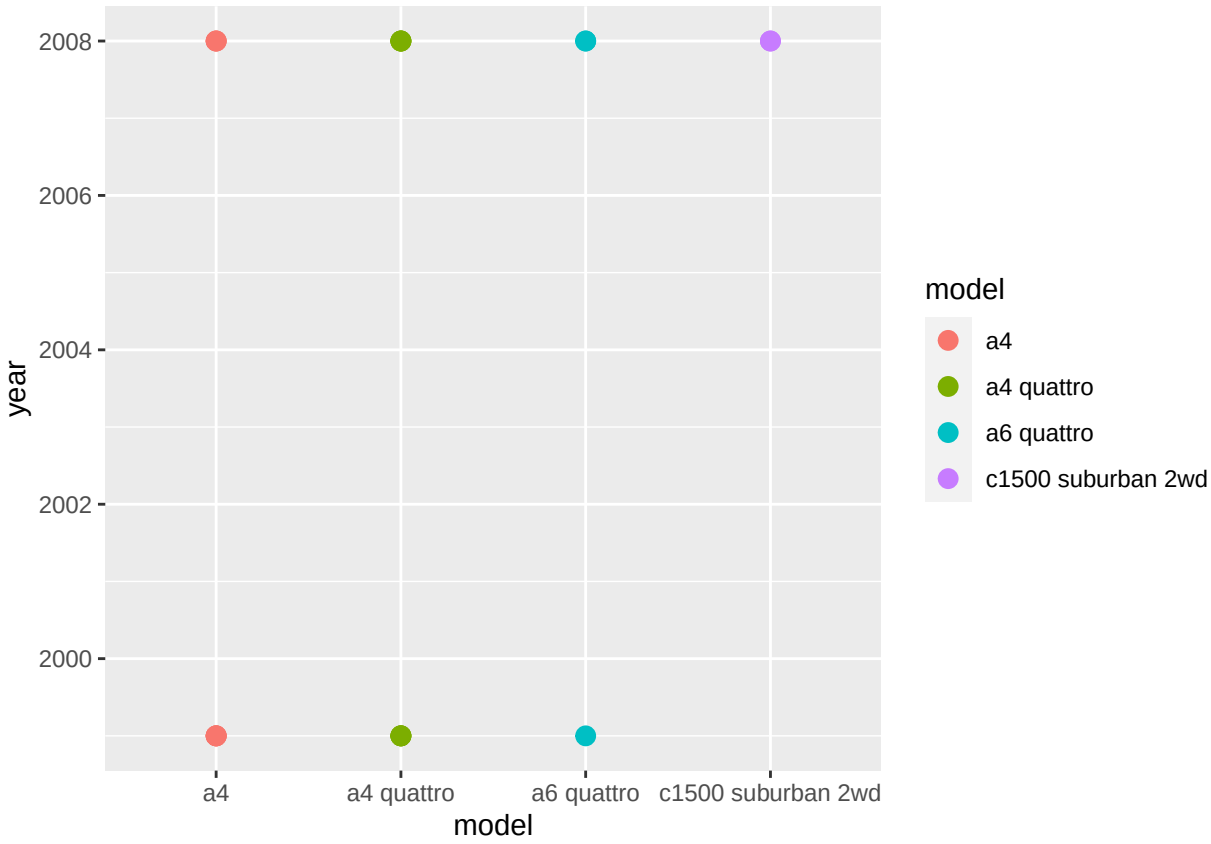


b. For you, is it useful? If not, how could you modify the data to make it more informative?

Yes,it is useful in order make a comparison to make decisions and soltuions.

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

```
num3mpg<-mpg[1:20,]
ggplot(num3mpg, aes(x = model, y = year, color=model )) + geom_point(size=3)
```



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

```
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
```

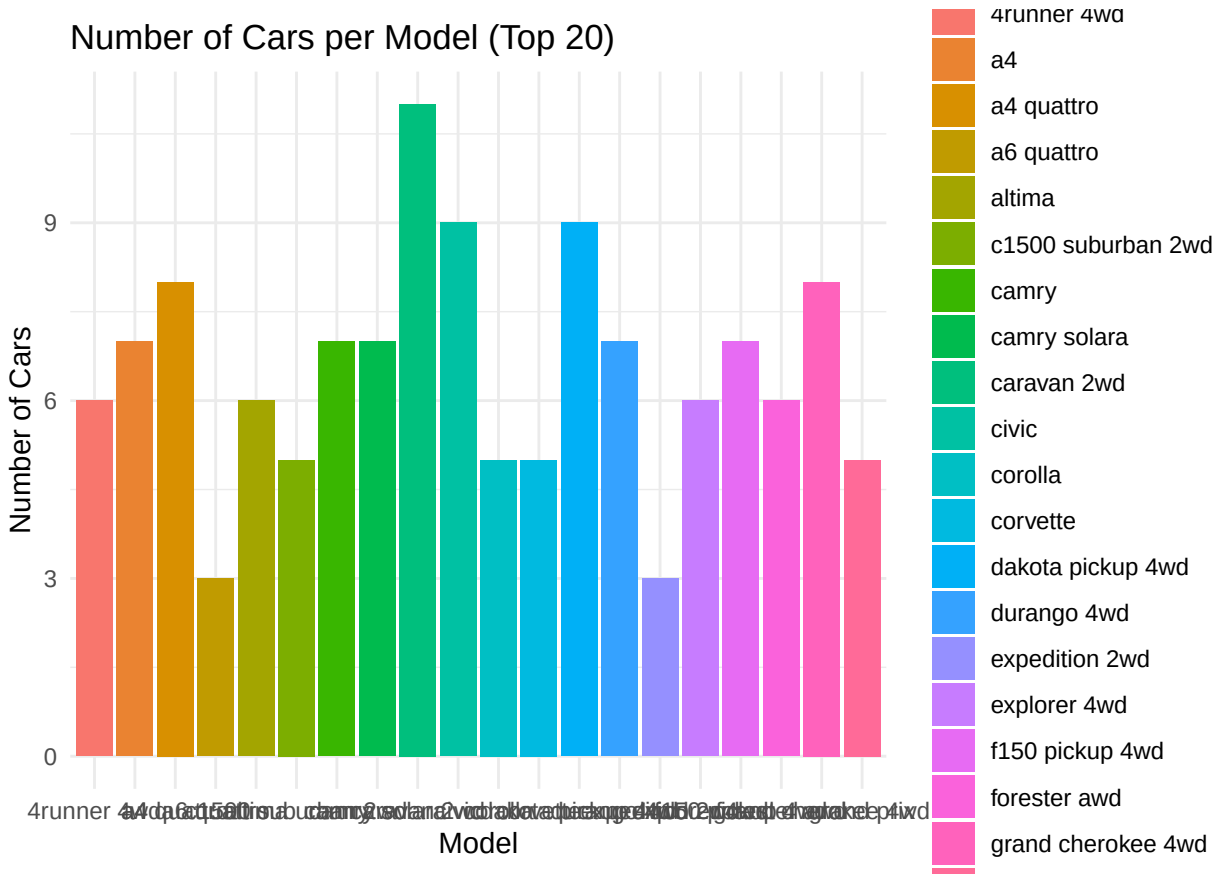
```
num4<-mpg %>%
 group_by(model) %>%
 summarise(count =n())


num4
```

```
## # A tibble: 38 x 2
##    model              count
##    <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
```

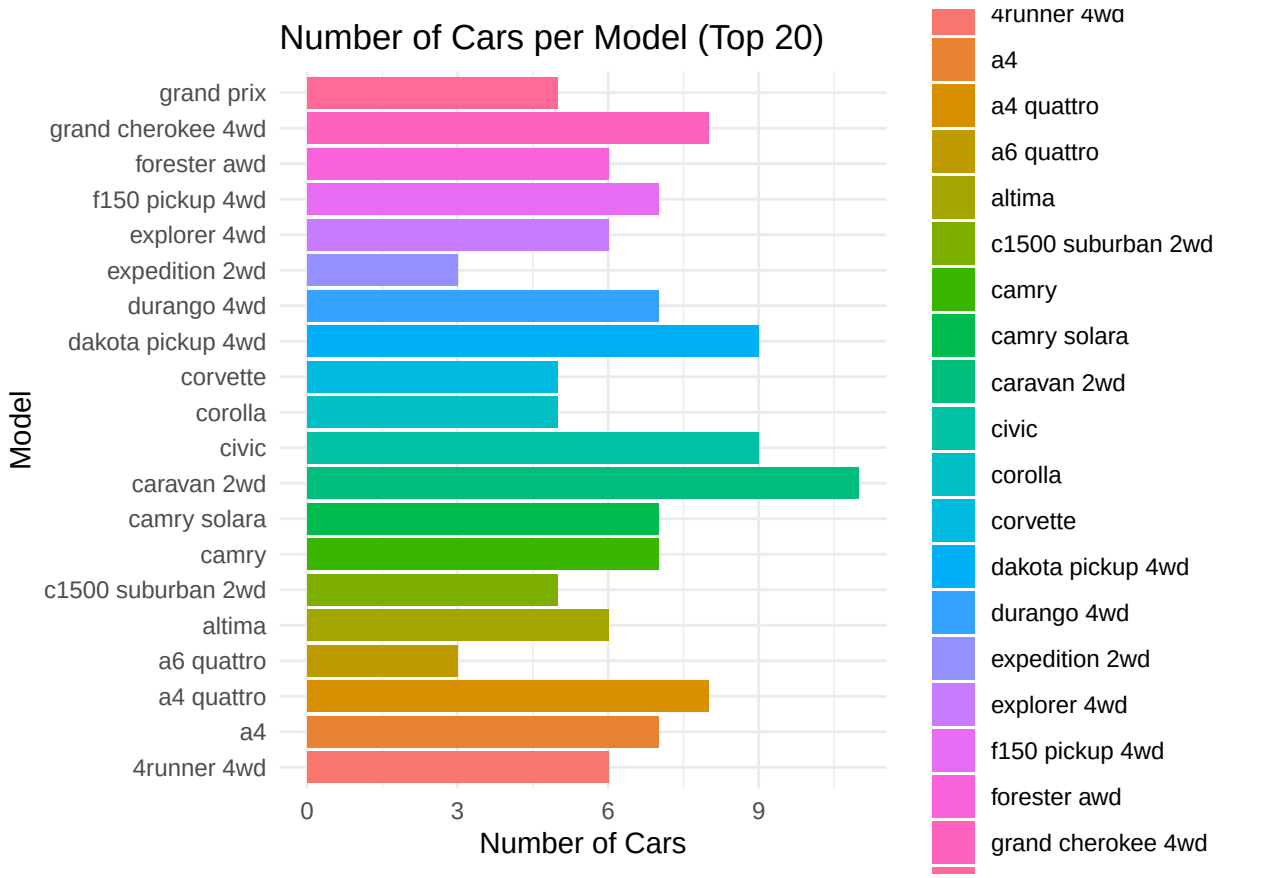
a. Plot using geom_bar() using the top 20 observations only. The graphs shoudl have a title, labels and colors. Show code and results.

```
ob20<-num4[1:20,1:2]

ggplot(ob20, aes(x = model, y = count, fill = model)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Number of Cars per Model (Top 20)",
       x = "Model",
       y = "Number of Cars") +
  scale_fill_hue()
```



b. Plot using the geom_bar() + coord_flip() just like what is shown below. Show codes and its result.

```
library(ggplot2)

ggplot(ob20, aes(x = model, y = count, fill = model)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Number of Cars per Model (Top 20)",
       x = "Model",
       y = "Number of Cars") +
  scale_fill_hue()
```

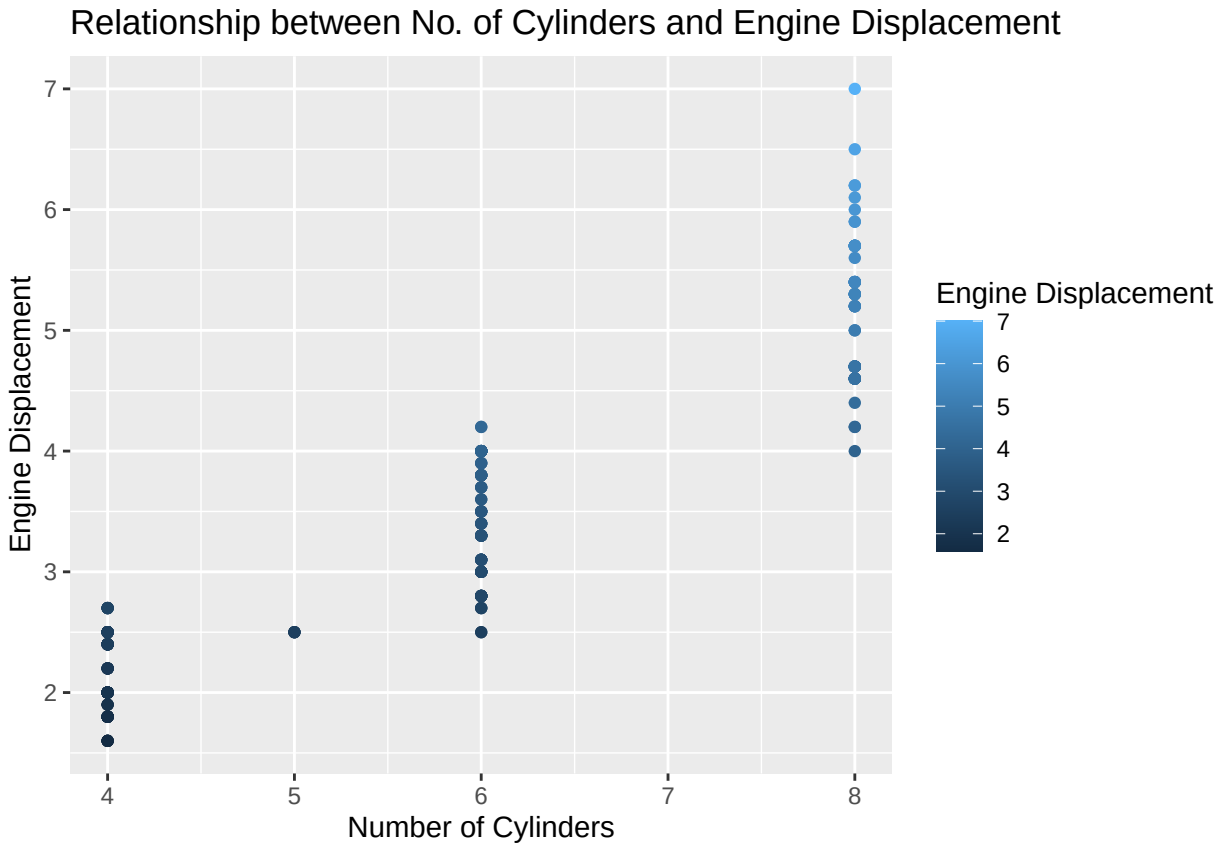


5. Plot the relationship between cyl - number of cylinders and displ - engine displacement using geom_point with aesthetic color = engine displacement. Title should be "Relationship between No. of Cylinders and Engine Displacement".

a. How would you describe its relationship? Show the codes and its result.

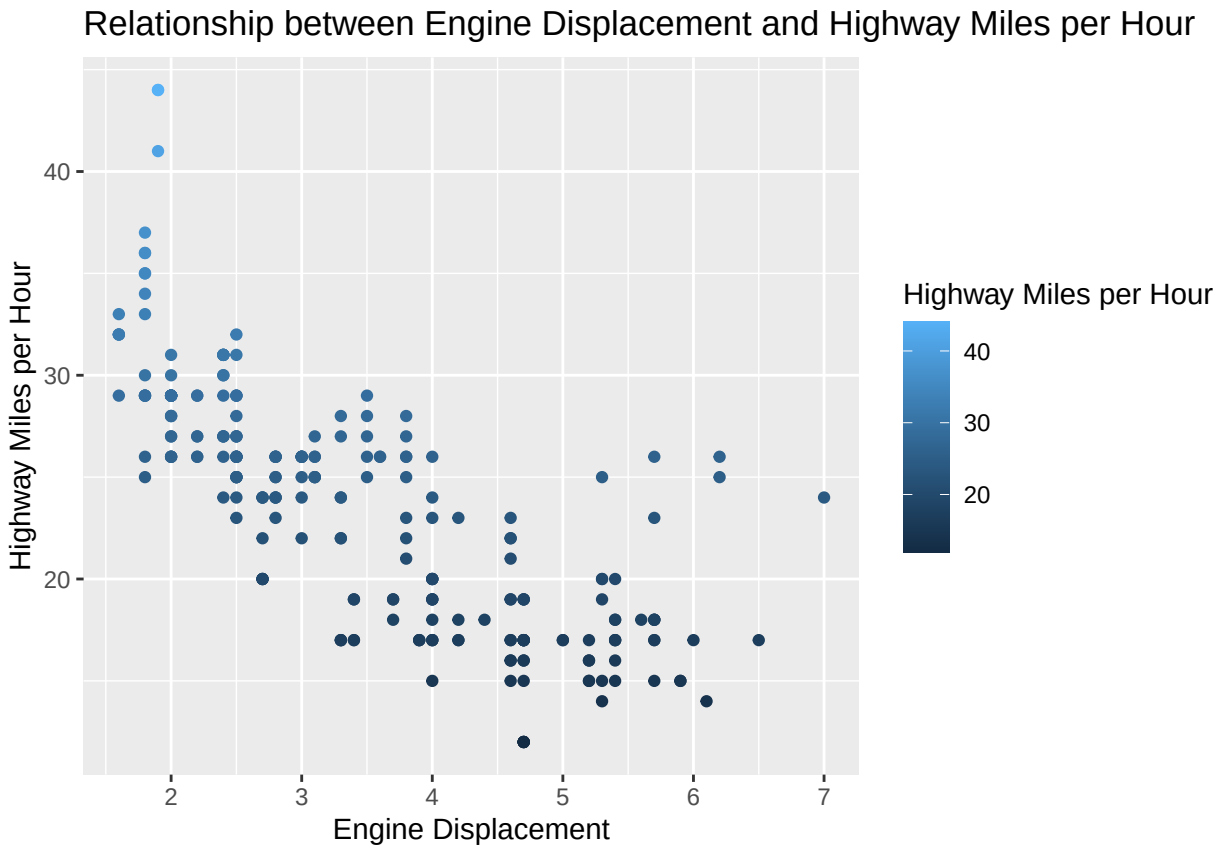
```
library(ggplot2)

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

```
      y = "Engine Displacement",
      color = "Engine Displacement")
```



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?

```
library(ggplot2)
ggplot(mpg, aes(x = displ, y = hwy, color = hwy)) +
  geom_point() +
  labs(title = "Relationship between Engine Displacement and Highway Miles per Hour",
       x = "Engine Displacement",
       y = "Highway Miles per Hour",
       color = "Highway Miles per Hour")
```

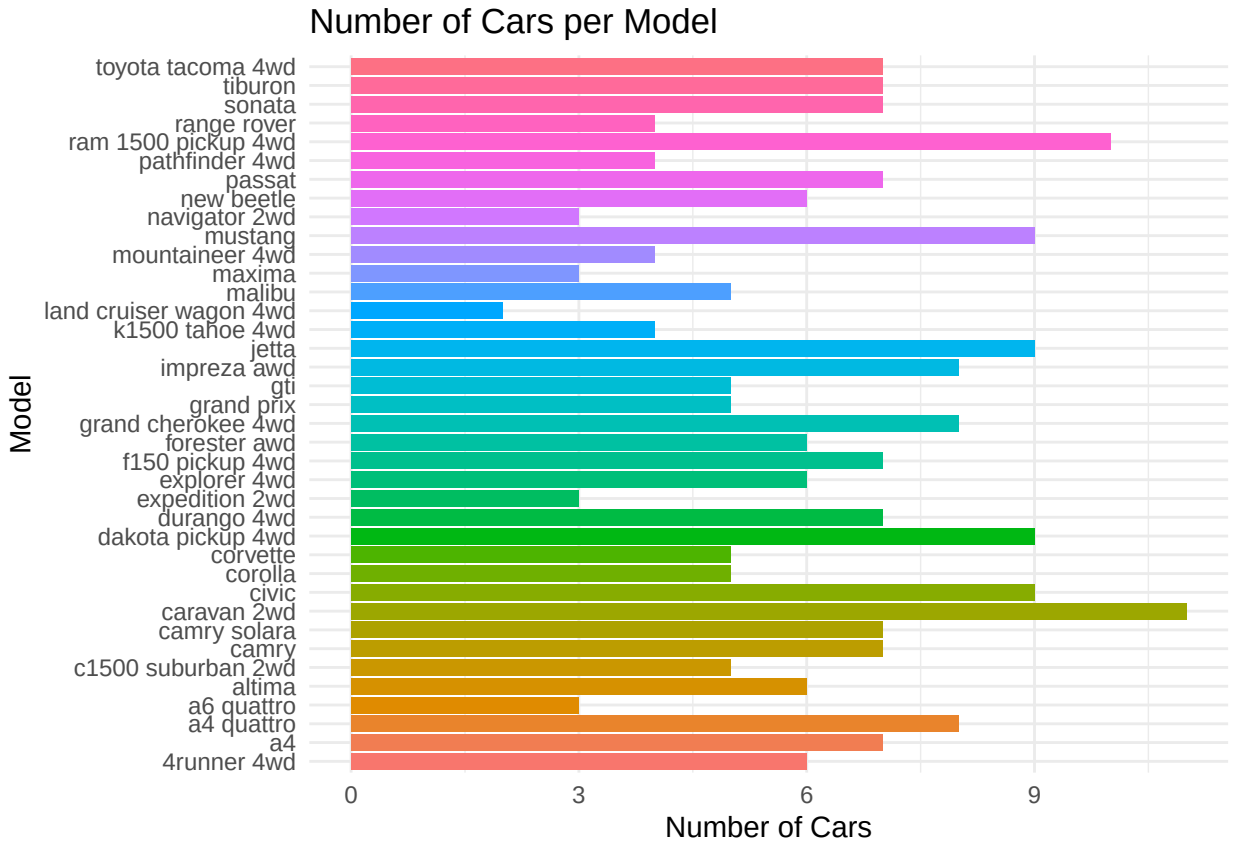


6. Import the traffic.csv onto your R environment.

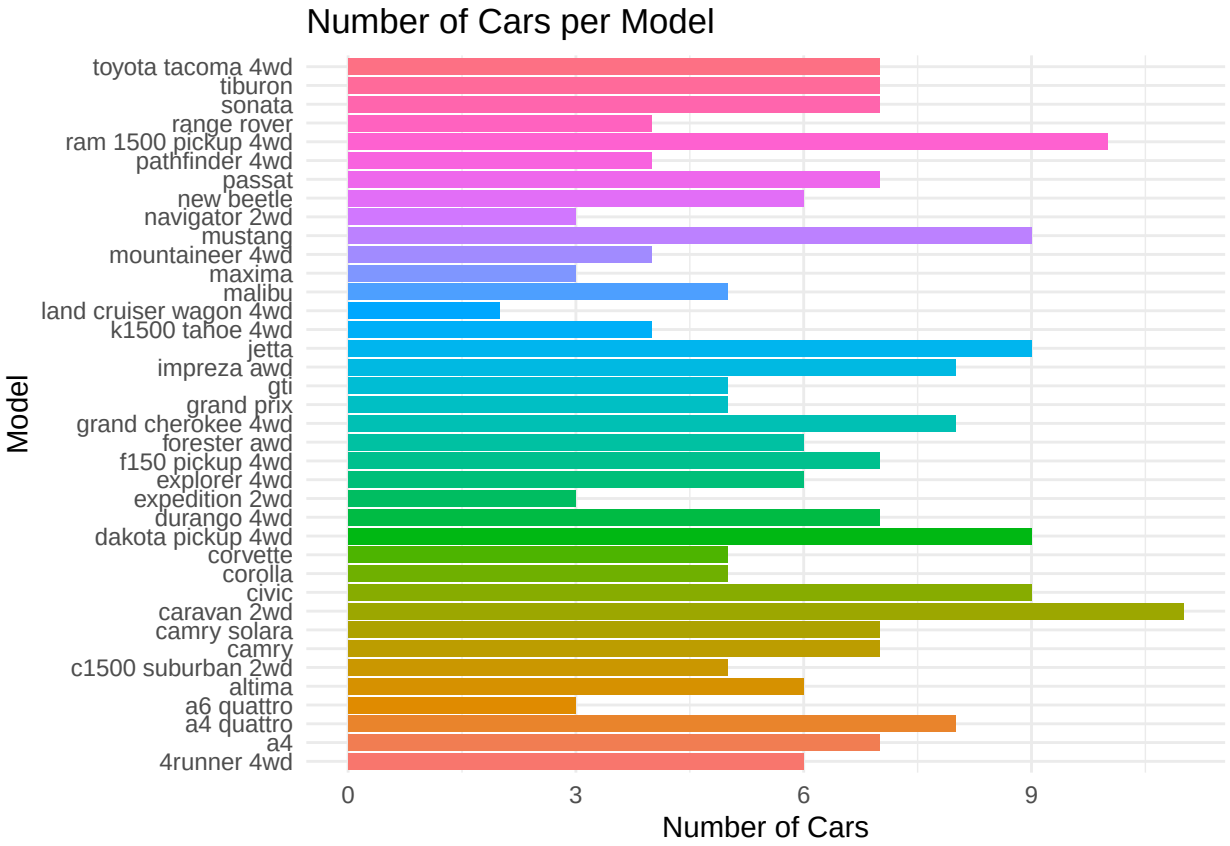
```
library(ggplot2)
traffic<-ggplot(num4, aes(x = model, y = count, fill = model)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Number of Cars per Model",
       x = "Model",
       y = "Number of Cars") +
  scale_fill_hue()+ guides(fill=FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
traffic
```




```
save(traffic,file = "traffic.csv")
load(file="traffic.csv")
traffic
```



a. How many numbers of observation does it have? What are the variables of the traffic dataset the Show your answer.

```
cat("The number of observations are:\n")
```

```
## The number of observations are:
```

```
nrow(traffic$data)
```

```
## [1] 38
```

```
cat("The variables of traffic dataset are:\n" )
```

```
## The variables of traffic dataset are:
```

```
colnames(traffic$data)
```

```
## [1] "model" "count"
```

b. subset the traffic dataset into junctions. What is the R codes and its output?

c. Plot each junction in a using geom_line(). Show your solution and output.

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

```
library(readxl)
alexa_file <- read_excel("alexa_file.xlsx")
```

a. How many observations does alexa_file has? What about the number of columns? Show your solution and answer.

```
cat("The number of observations in alexa_file are:\n ")
```

```
## The number of observations in alexa_file are:
##
```

```
nrow(alexa_file)
```

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

```
cat("The number of columns in alexa_file are:\n ")
```

```
## The number of columns in alexa_file are:
##
```

```
ncol(alexa_file)
```

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

b. group the variations and get the total of each variations. Use dplyr package. Show solution and answer.

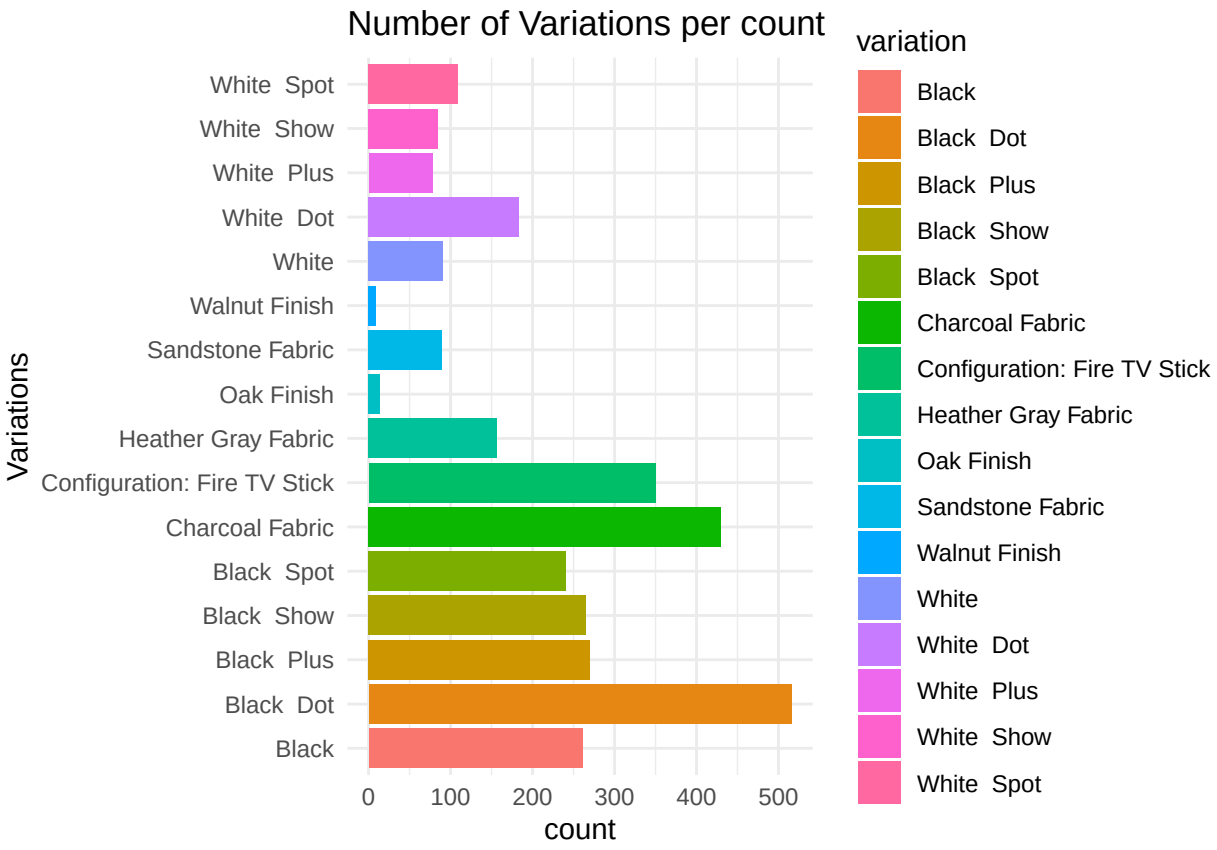
```
library(dplyr)
num7<-alexa_file %>%
 group_by(variation) %>%
 summarise(count = n())
num7
```

```
## # A tibble: 16 x 2
##    variation                    count
##    <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
```

c. Plot the variations using the ggplot() function. What did you observe? Complete the details of the graph. Show solution and answer.

```
library(ggplot2)
ggplot(num7, aes(x=variation,y=count,fill=variation))+geom_bar(stat="Identity")+coord_flip() +
  theme_minimal() +
```

```
  labs(title = "Number of Variations per count",
       x = "Variations",
       y = "count")
```



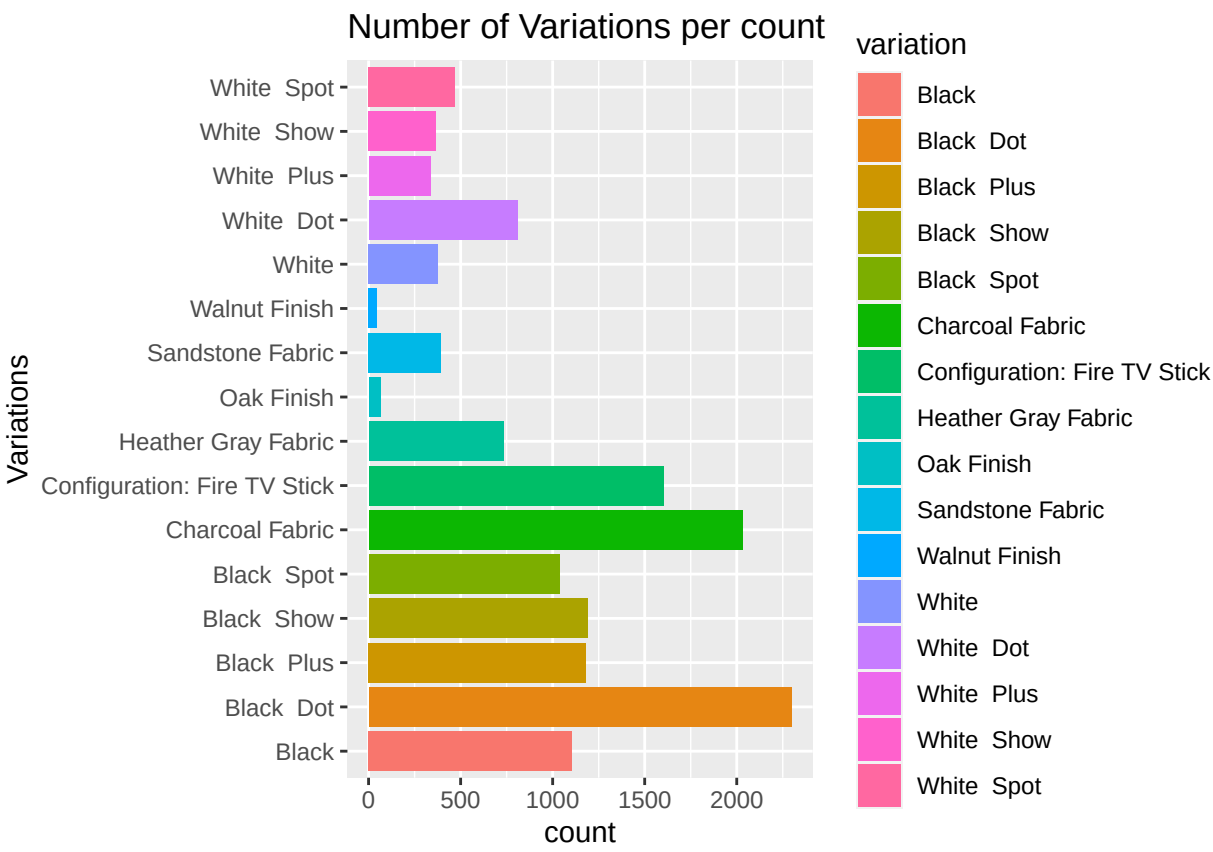
d. Plot a geom_line() with the date and the number of verified reviews. Complete the details of the graphs. Show your answer and solution.

```
library(ggplot2)

ggplot(alexa_file, aes(x = verified_reviews, y = date)) +
  geom_line(color = "skyblue") +
  geom_point(color = "violet", size = 2) +
  labs(title = "Verified Reviews Over Time",
       x = "Number of Verified Reviews",
       y = "Date") +
  theme_minimal()
```

e. Get the relationship of variations and ratings. Which variations got the most highest in rating? Plot a graph to show its relationship. Show your solution and answer

```
The Highest rating is the Black Dot
```

```
ggplot(alexa_file, aes(x=variation ,y=rating,fill=variation))+
  geom_bar(stat="Identity")+
  coord_flip() +
  labs(title = "Number of Variations per count",
       x = "Variations",
       y = "count")
```


Number of Variations per count
Variations
White Spot
White Show
White Plus
White Dot
White
Walnut Finish
Sandstone Fabric
Oak Finish
Heather Gray Fabric
Configuration: Fire TV Stick
Charcoal Fabric
Black Spot
Black Show
Black Plus
Black Dot
Black
0
500
1000
1500
2000
count
variation
Black
Black Dot
Black Plus
Black Show
Black Spot
Charcoal Fabric
Configuration: Fire TV Stick
Heather Gray Fabric
Oak Finish
Sandstone Fabric
Walnut Finish
White
White Dot
White Plus
White Show
White Spot