RWorksheet #6

Grace Anne E. Capanang

2022-11-26

Use the dataset **mpg**

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
data("mpg")
as.data.frame(data(mpg))
```

```
##   data(mpg)
## 1       mpg
```

```
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ 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        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [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         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [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" ...
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## 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
```

```
glimpse(mpg)
```

```
## Rows: 234
## Columns: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "audi", "~
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 quattro", "~
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0, 2.0, 2.~
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, 200~
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, 8, ~
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "auto~
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4", "4~
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17, 1~
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25, 2~
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p~
## $ class        <chr> "compact", "compact", "compact", "compact", "compact", "c~
```

#1. How many columns are in mpg dataset? How about the number of rows? Show the codes and its result.

```
mpgNumOfCol <- ncol(mpg)
mpgNumOfCol
```

```
## [1] 11
```

```
mpgNumOfRows <- nrow(mpg)
mpgNumOfRows
```

```
## [1] 234
```

```
mpgColRow <- glimpse(mpg)
```

```
## Rows: 234
## Columns: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "audi", "~
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 quattro", "~
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0, 2.0, 2.~
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, 200~
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, 8, ~
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "auto~
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4", "4~
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17, 1~
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25, 2~
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p~
## $ class        <chr> "compact", "compact", "compact", "compact", "compact", "c~
```

```
mpgColRow
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

#The number of rows in mpg dataset is 234, and the number of columns is 11.

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

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

```
##
##       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
```

#The manufacturer named 'Dodge' has the most models in the data set, consisting of 37 models.

```
mostNumberModel <- table(mpg$model)
mostNumberModel
```

```
##
##            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
```

```
anotherMeth <- mpgColRow %>%
  group_by(model) %>%
  count()
anotherMeth
```

```
## # A tibble: 38 x 2
## # Groups:   model [38]
##    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
## # ... with 28 more rows
```

#The model who has the most variations is the 'caravan 2wd', having 11.

#2a. Group the manufacturers and find the unique models. Copy the codes and result.

```
groupResult <- mpgColRow %>%
  group_by(manufacturer, model) %>%
  distinct() %>% count()
groupResult
```

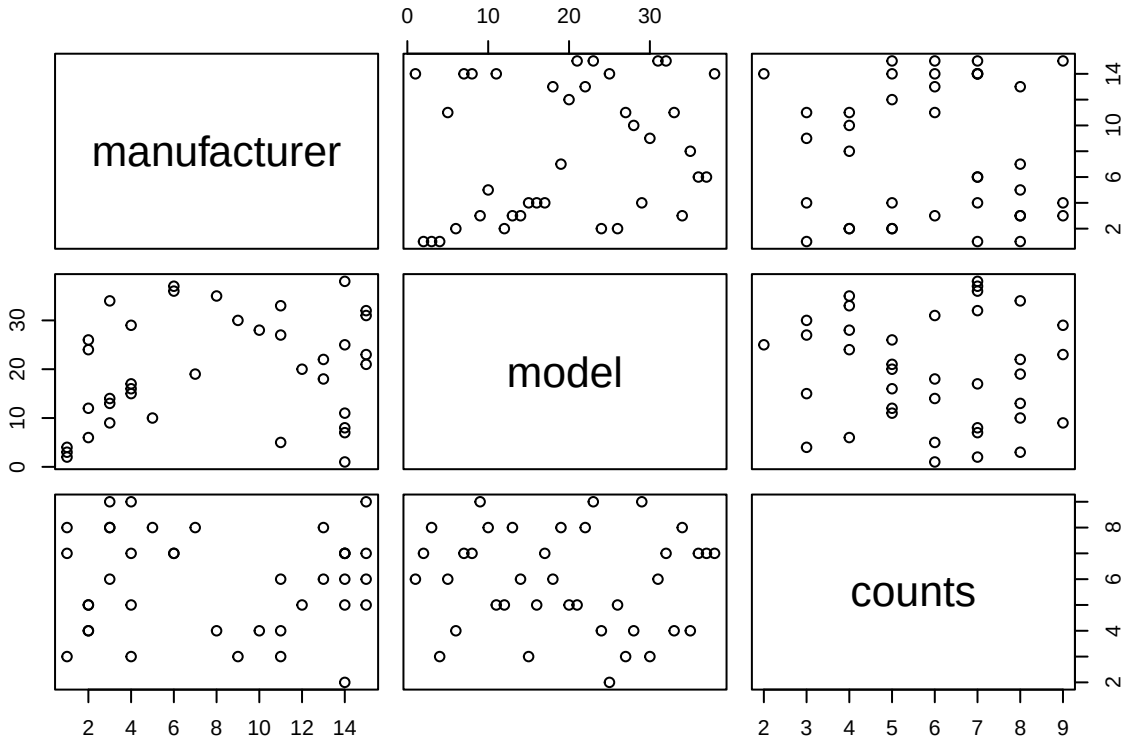
```
## # A tibble: 38 x 3
## # Groups:   manufacturer, model [38]
##    manufacturer model                  n
##    <chr>        <chr>              <int>
##  1 audi         a4                     7
##  2 audi         a4 quattro             8
##  3 audi         a6 quattro             3
##  4 chevrolet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

```
## 10 dodge         durango 4wd             6
## # ... with 28 more rows
```

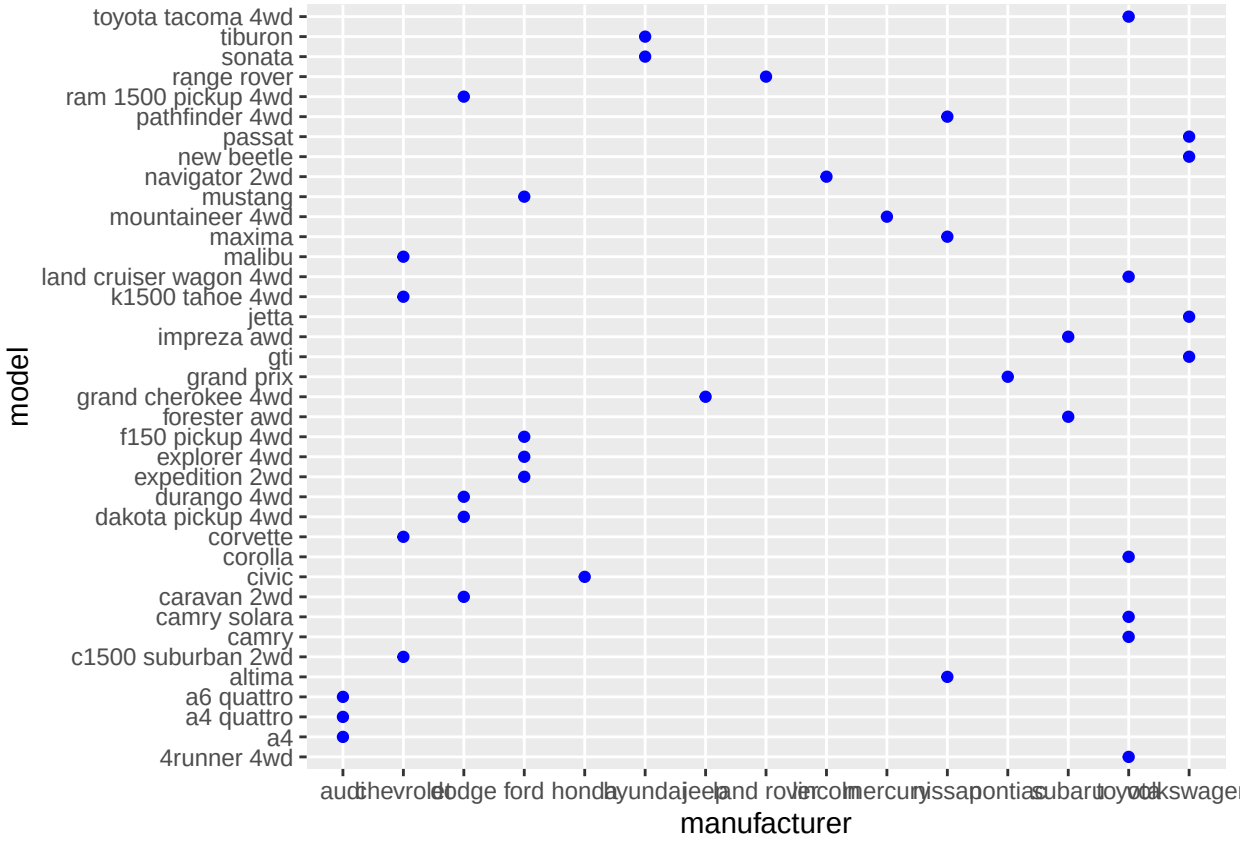
```
colnames(groupResult) <- c("manufacturer", "model", "counts")
```

#2b. Graph the result by using plot() and ggplot(). Write the codes and its result.

```
plot (groupResult)
```

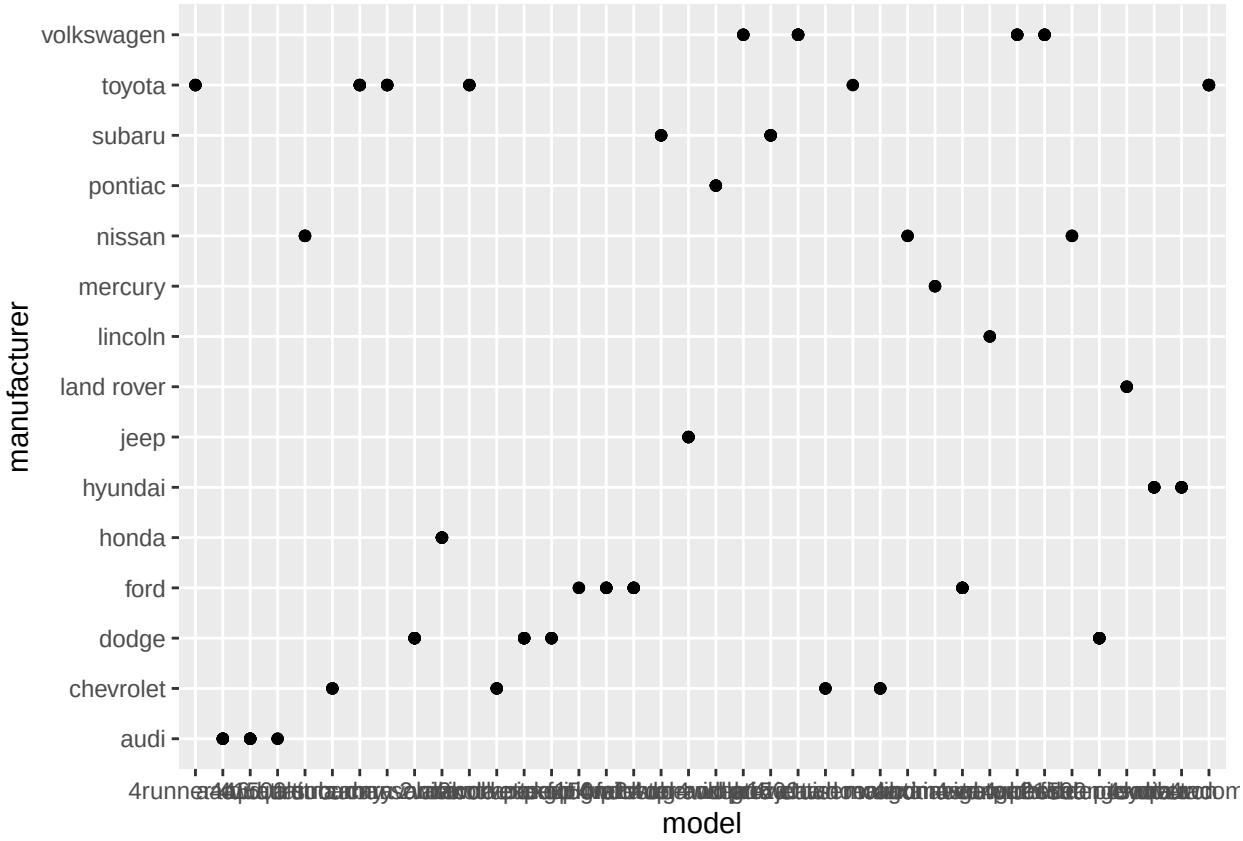



```
ggplot(groupResult,
       aes(x = manufacturer, y = model)) +
  geom_point (color = 'blue')
```



#3a. Same dataset will be used. You are going to show the relationship of the modeland the manufacturer. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?

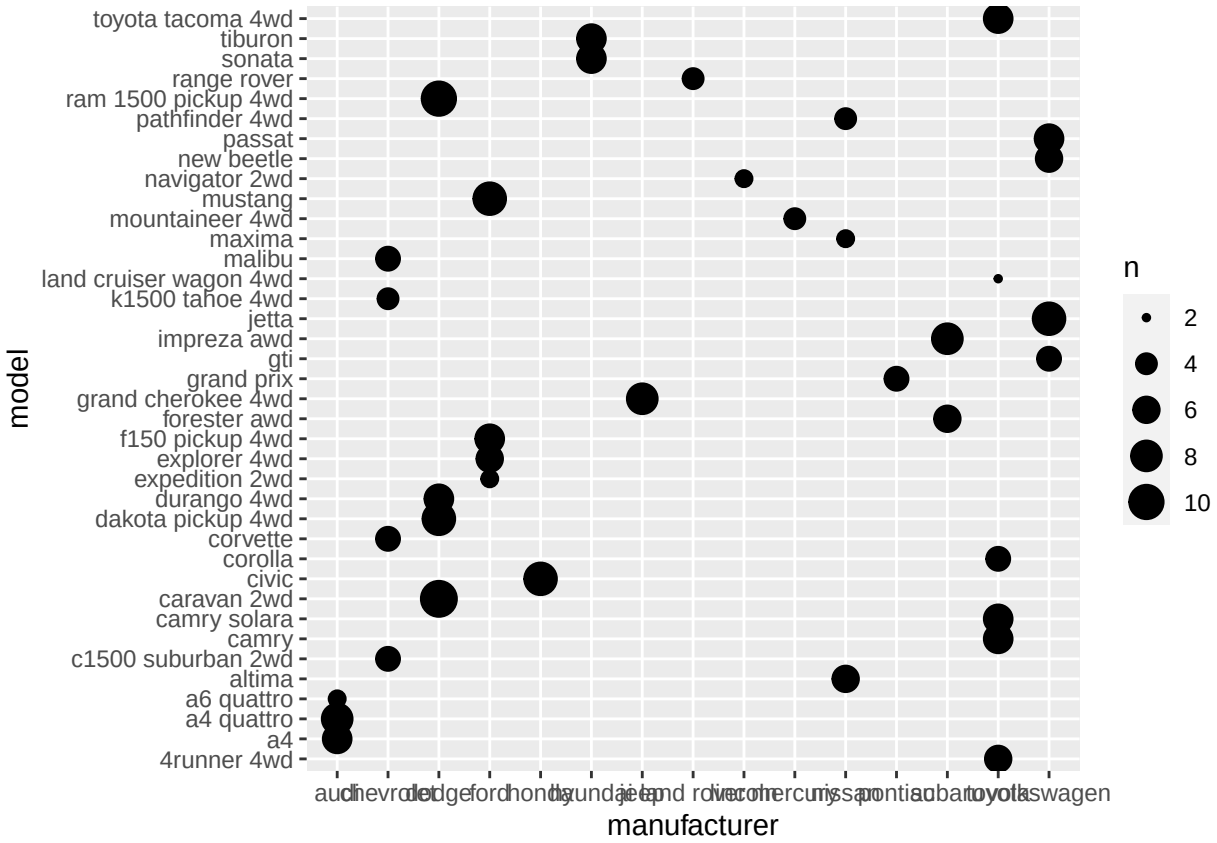
```
ggplot(mpg, aes(model, manufacturer)) + geom_point()
```



#The plot shows a scatter plot displaying the relationship between car manufacturer and model of the the mpg data set; a variety of car models and what manufacturer they belong into.

#3b. For you, is it useful? If not, how could you modify the data to make it more informative?

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



#To make it look more informative and neat, putting the model in the x-axis part is better because there are many of them and if you put them in the y-axis the names become incomprehinsible, thus changing their axis position is a good option. # Using geom_count() than geom_point() delivers more information because the size of the dot serves as a guide to the models' count to the manufacturer they are categorized into.

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

```
mpgModel <- mpgColRow %>%
  group_by(model) %>% count()
mpgModel
```

```
## # A tibble: 38 x 2
## # Groups:   model [38]
##    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
```

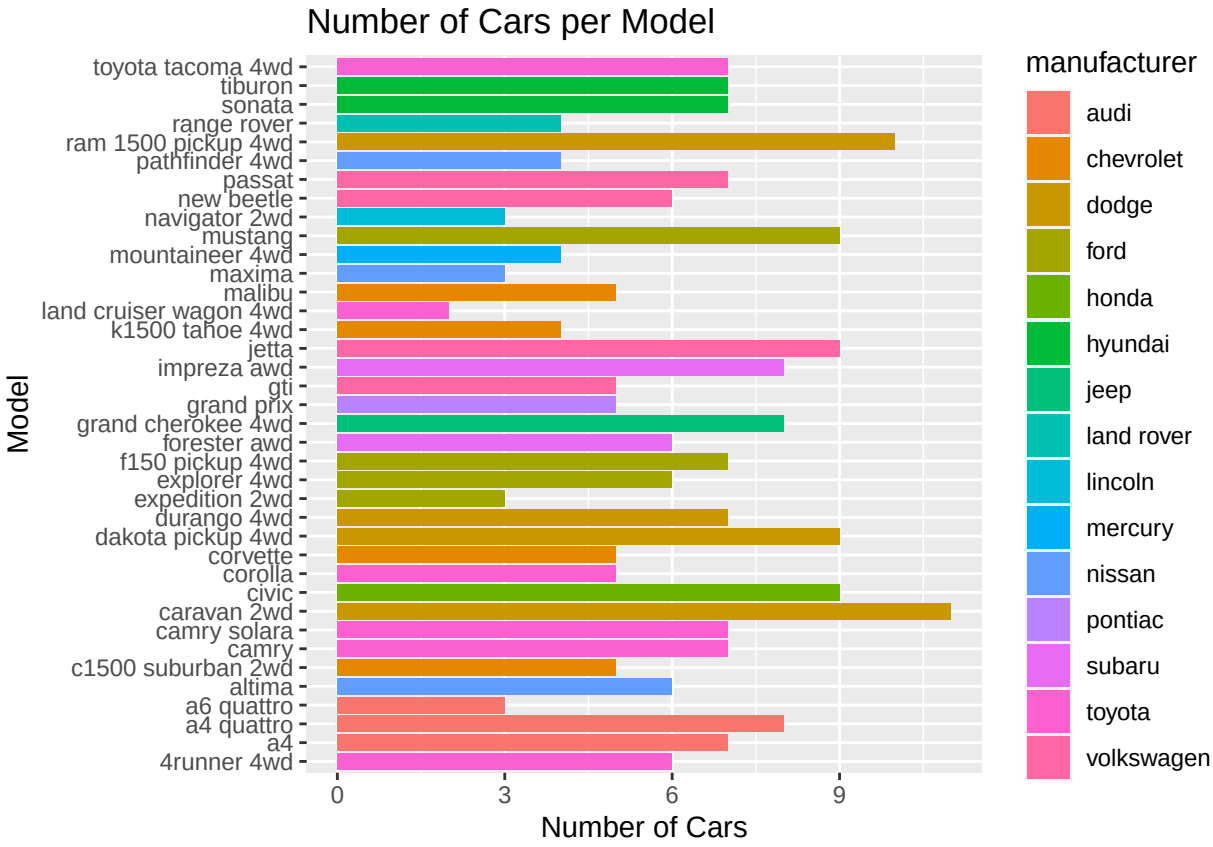
```
## # ... with 28 more rows
```

```
colnames(mpgModel) <- c("model", "counts")
```

#4a. Plot using the geom_bar() + coord_flip() just like what is shown below. Show codes and its result.

```
qplot(model,data = mpgColRow,
      main = "Number of Cars per Model",
      xlab = "Model",
      ylab = "Number of Cars",
      geom = "bar",
      fill = manufacturer) + coord_flip()
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```



#4b. Plot using the geom_bar() + coord_flip() just like what is shown below. Show codes and its result.

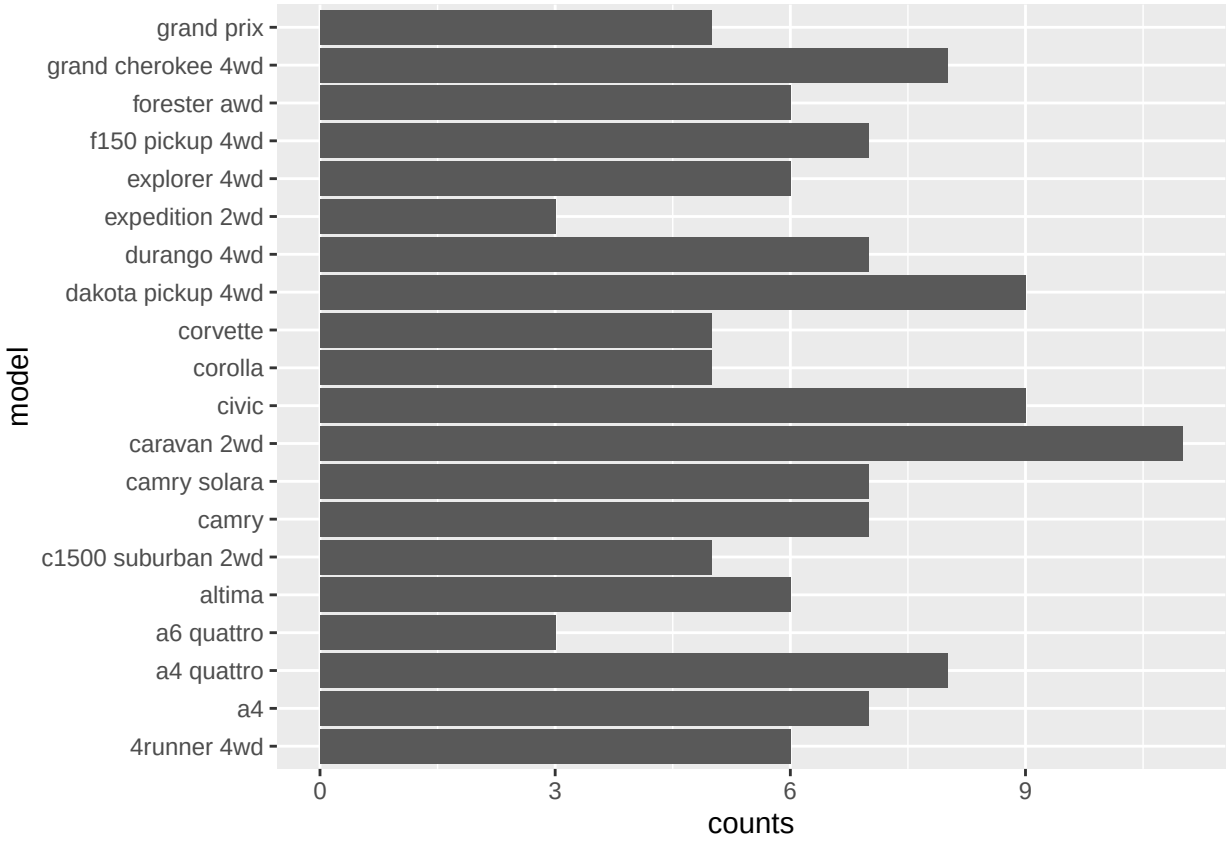
```
firstTwenty <- mpgModel[1:20,] %>%
  top_n(2)
```

```
## Selecting by counts
```

```
firstTwenty
```

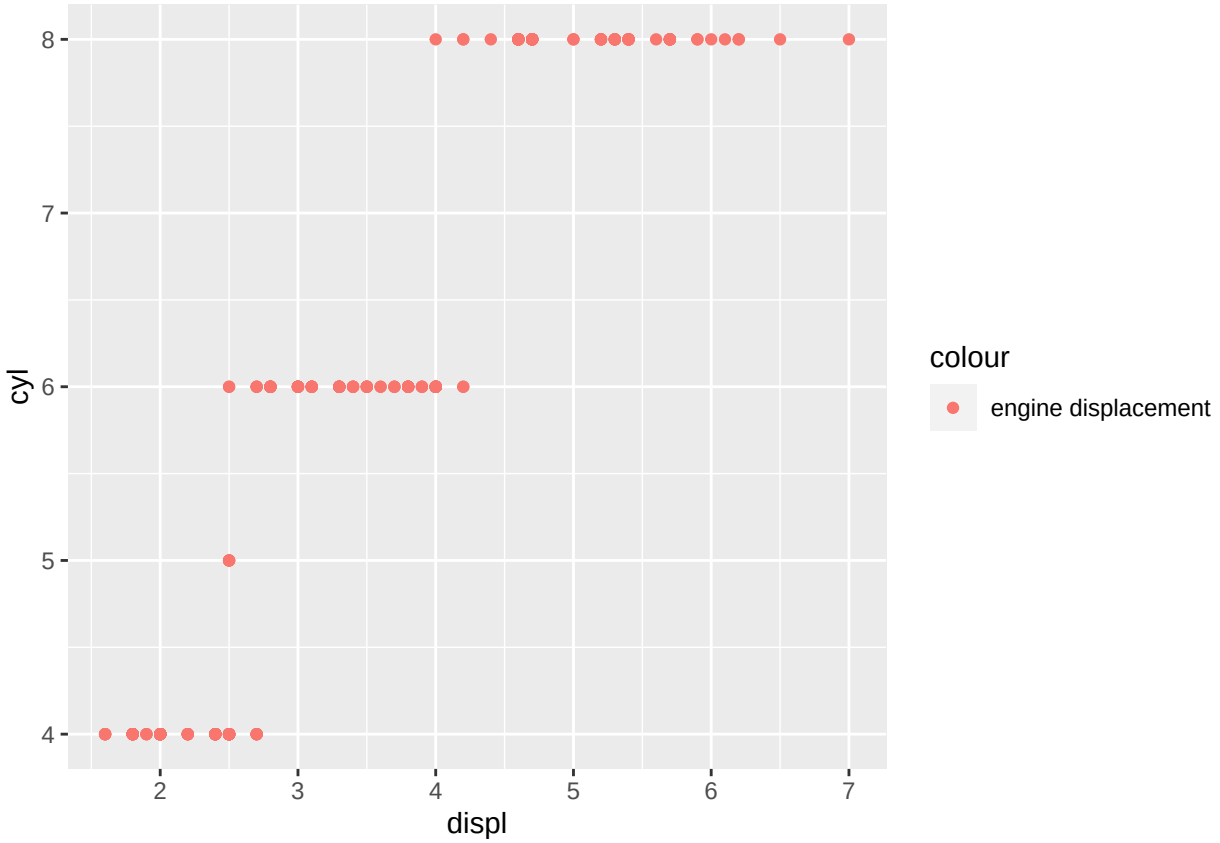
```
## # A tibble: 20 x 2
## # Groups:   model [20]
##    model              counts
##    <chr>               <int>
##  1 4runner 4wd             6
##  2 a4                      7
##  3 a4 quattro              8
##  4 a6 quattro              3
##  5 altima                  6
##  6 c1500 suburban 2wd      5
##  7 camry                   7
##  8 camry solara            7
##  9 caravan 2wd            11
## 10 civic                   9
## 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
```

```
ggplot(firstTwenty,
       aes(x = model, y = counts)) +
       geom_bar (stat = "Identity") + coord_flip()
```



#5a. Plot the relationship between cyl - number of cylinders and displ - engine displacement using geom_point with aesthetic colour = engine displacement. Title should be "Relationship between No. of Cylinders and Engine Displacement". Show the codes and its result.

```
ggplot(data = mpgColRow,
       mapping = aes(x = displ, y = cyl,main = "Relationship Between No. of Cylinders and Engine Displa
       geom_point (mapping = aes(colour = "engine displacement"))
```

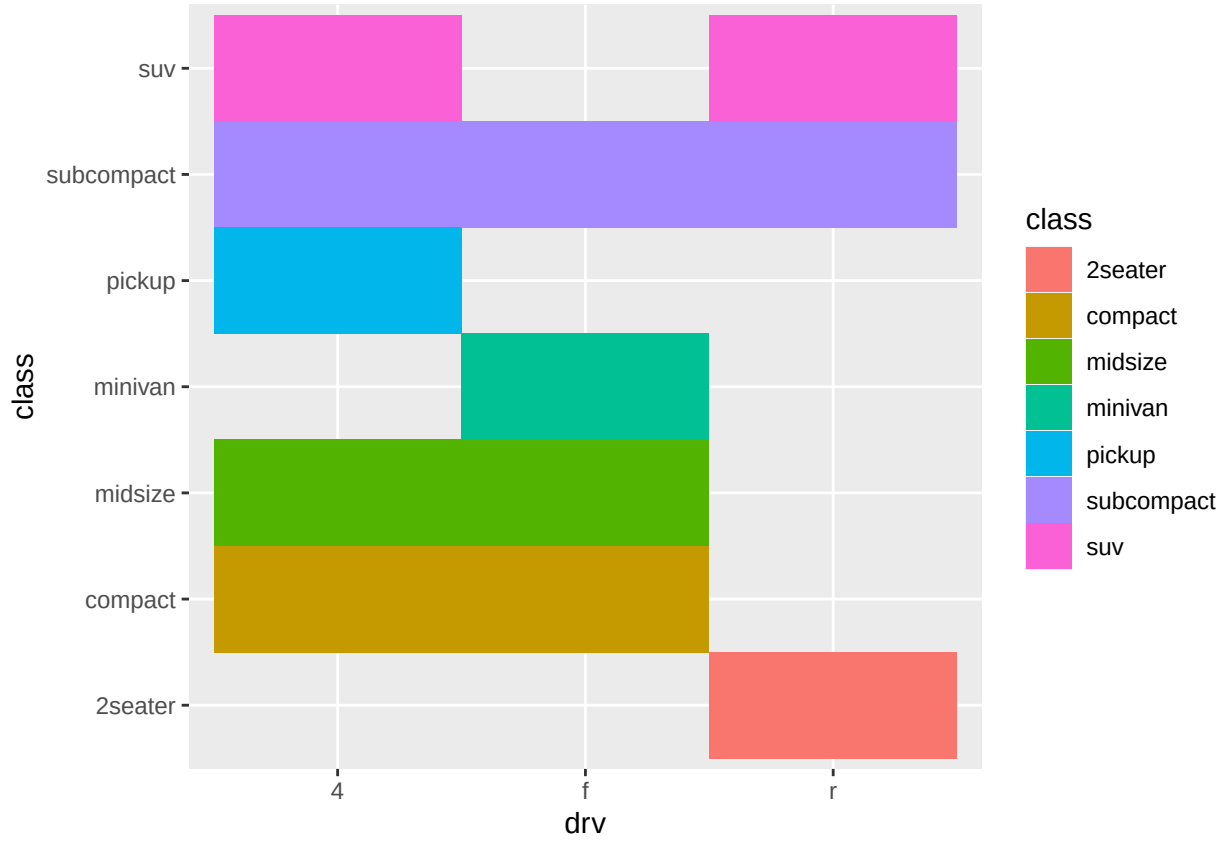


#5b. How would you describe its relationship?

#Engine displacement is how the car's engine can powers the car by how much air and fuel it can displace. And larger cylinders means more air. #Hence, the relationship between the number of cylinders and the engine displacement is that the greater the number of the cylinder, the bigger the engine displacement, the more power it can create. If the model has smaller number of cylinder, the engine displacement is small too, and the less power it create but also the lesser the fuel it consume.

#6a. Get the total number of observations for drv - type of drive train (f = front-wheel drive, r = rear wheel drive, 4 = 4wd) and class - type of class (Example: suv, 2seater, etc.). Plot using the geom_tile() where the number of observations for class be used as a fill for aesthetics. Show the codes and its result for the narrative in #6.

```
ggplot(data = mpgColRow,
       mapping = aes(x = drv, y = class)) +
  geom_tile (aes(fill=class))
```



#Claire de Lune

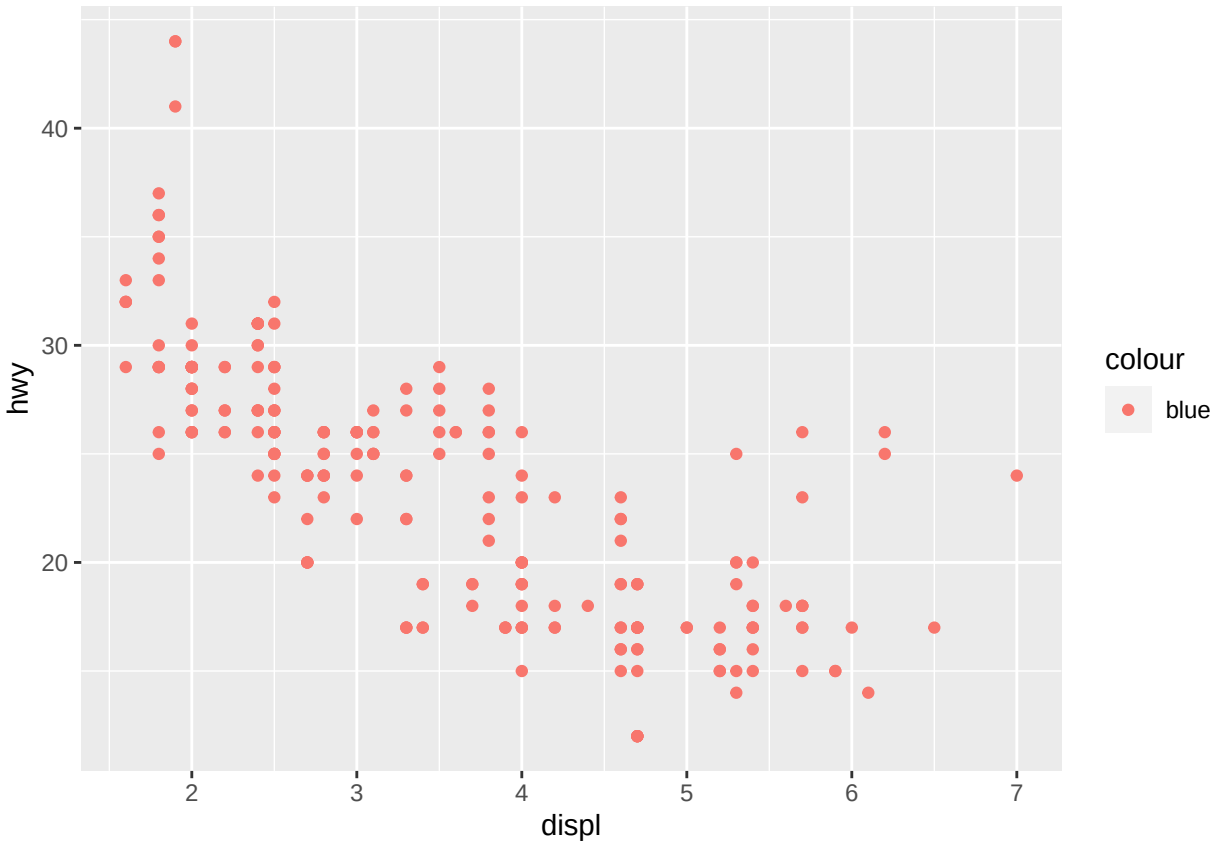
#6b. Show the codes and its result for the narrative in #6.

#The data shows a tile graph wherein most of the car's class have a 4-wheel drive as type of drive train, and the least of the class' drive train have is the rear-wheel drive.

#7. Discuss the difference between these codes. Its outputs for each are shown below.

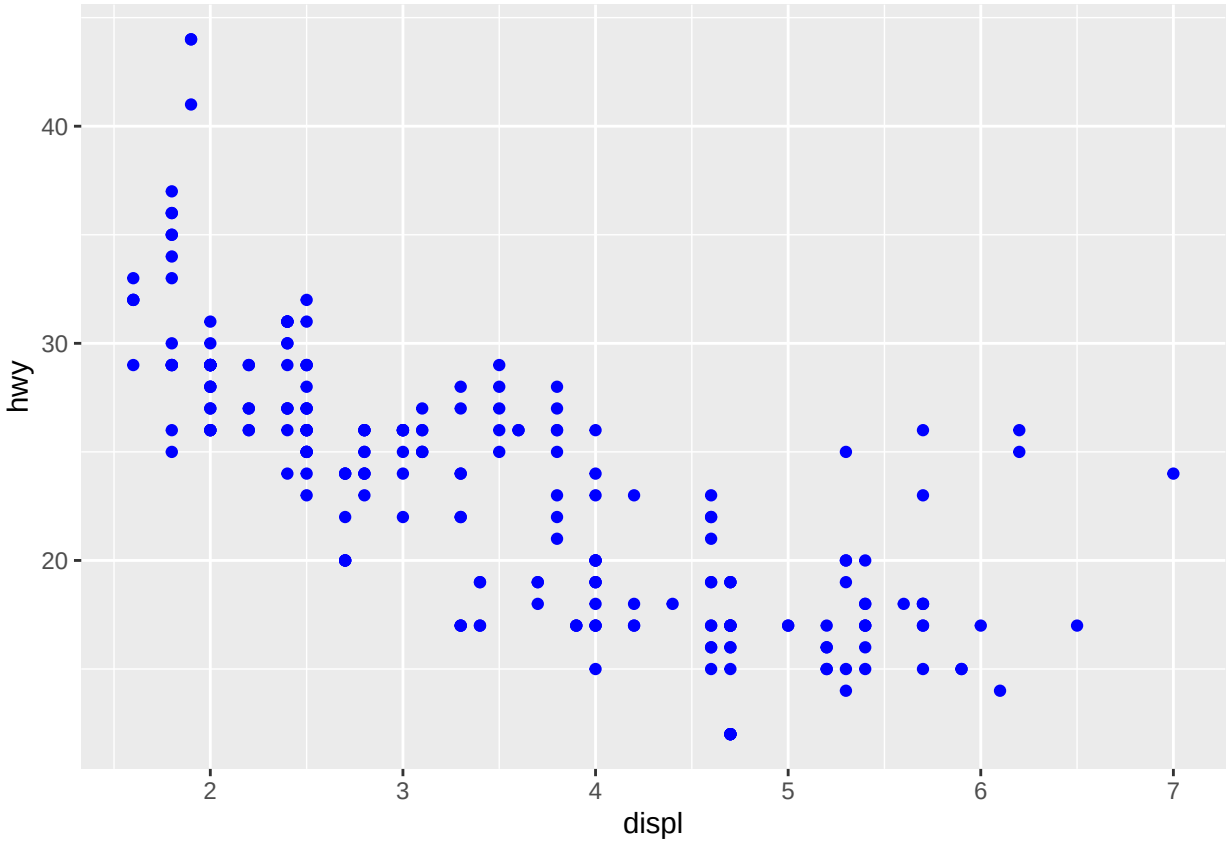
#1st code

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = "blue"))
```



#2nd code

```
ggplot(data = mpg) +
   geom_point(mapping = aes(x = displ, y = hwy), colour = "blue")
```



#The difference between the two code is the color of the geom point. In the first code the colour did not apply to the graph. #While in the second code the colour 'blue' is successfully used as a guide to the graph. #In the code, the colour must be not inside the aesthetic in order for it to be successfully use in the graph we'll make.

#8. Try to run the command ?mpg. What is the result of this command?

```
?mpg
```

```
## starting httpd help server ... done
```

```
    knitr::include_graphics("mpg1.jpg")
```

```
  knitr::include_graphics("mpg2.jpg")
```

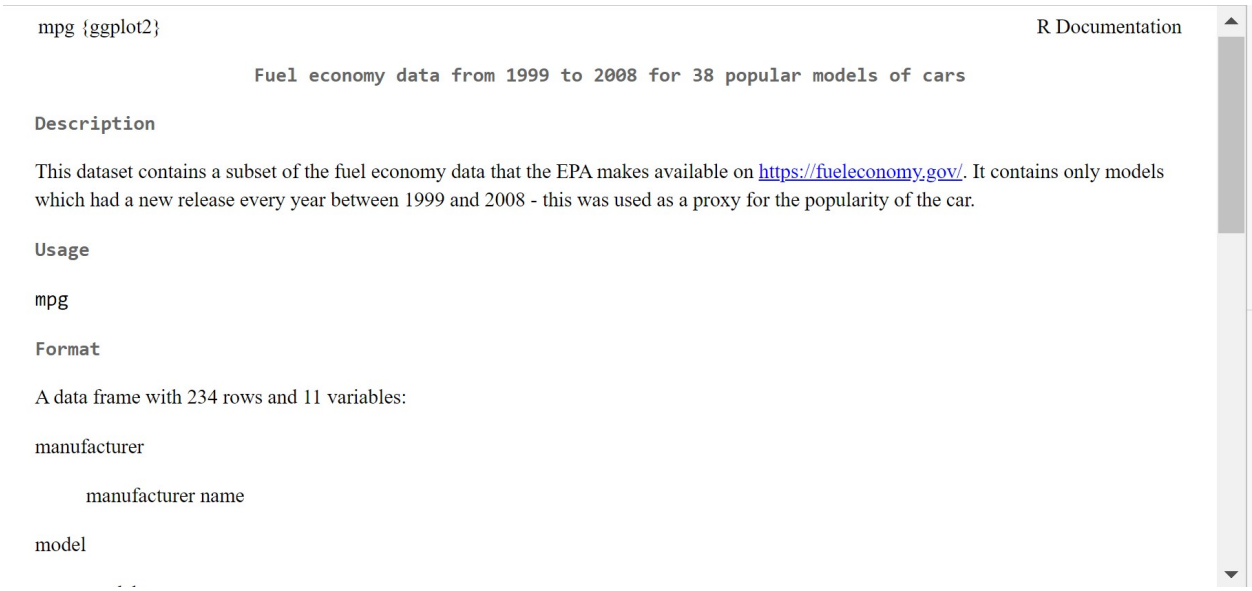
```
  knitr::include_graphics("mpg3.jpg")
```

#8a. Which variables from mpg dataset are categorical?

#The manufacturer, model, trans, drv, fl, and class are the categorical variable from the dataset because they are characters which values are limited and specific.

#8b. Which are continuous variables?

#The displ, year, cyl, cty, and hwy are the continous variable because they are numeric and have infinite values which you cannot categorize.



mpg {ggplot2} R Documentation

Fuel economy data from 1999 to 2008 for 38 popular models of cars

Description

This dataset contains a subset of the fuel economy data that the EPA makes available on https://fueleconomy.gov/. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

Usage

```
mpg
```

Format

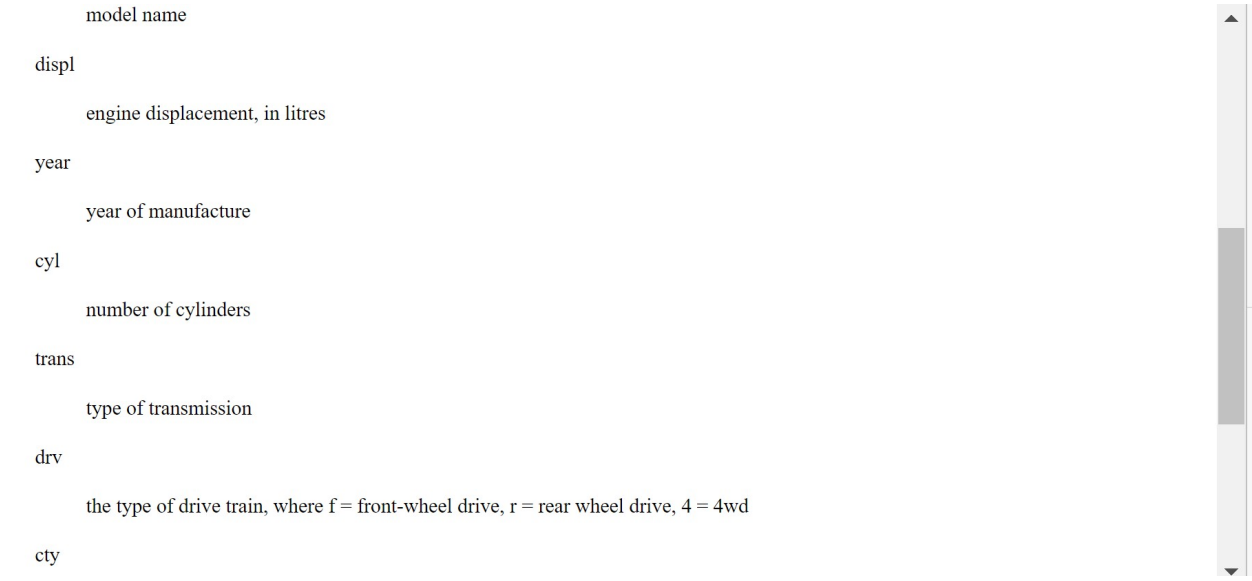
A data frame with 234 rows and 11 variables:

manufacturer

: manufacturer name

model

Figure 1: A caption



: model name

displ

: engine displacement, in litres

year

: year of manufacture

cyl

: number of cylinders

trans

: type of transmission

drv

: the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd

cty

Figure 2: A caption

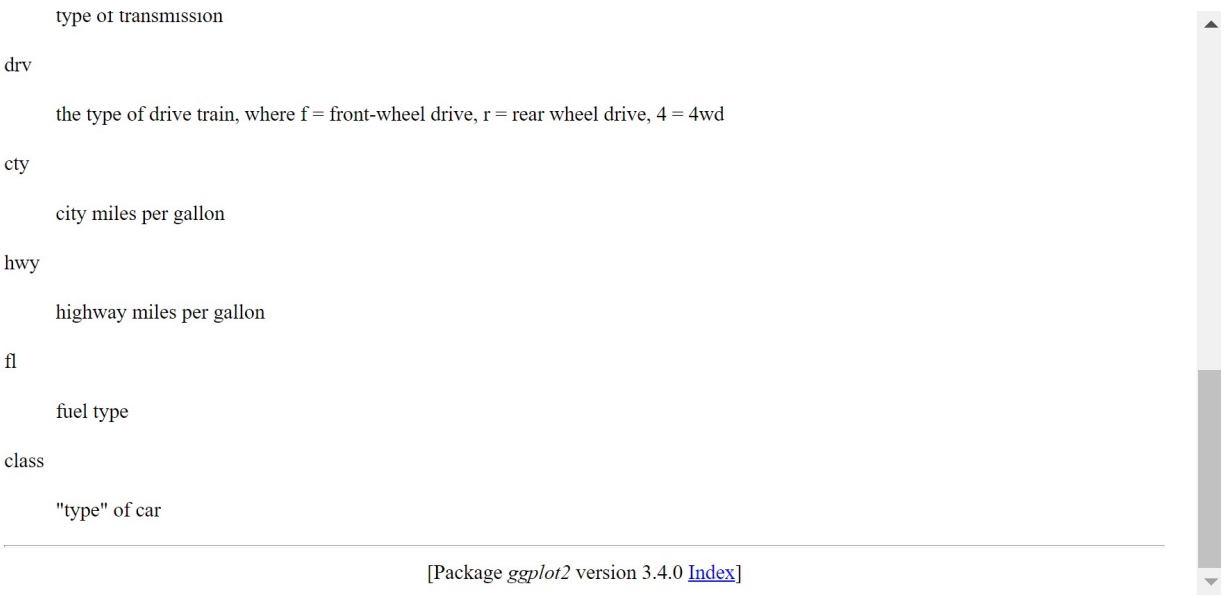
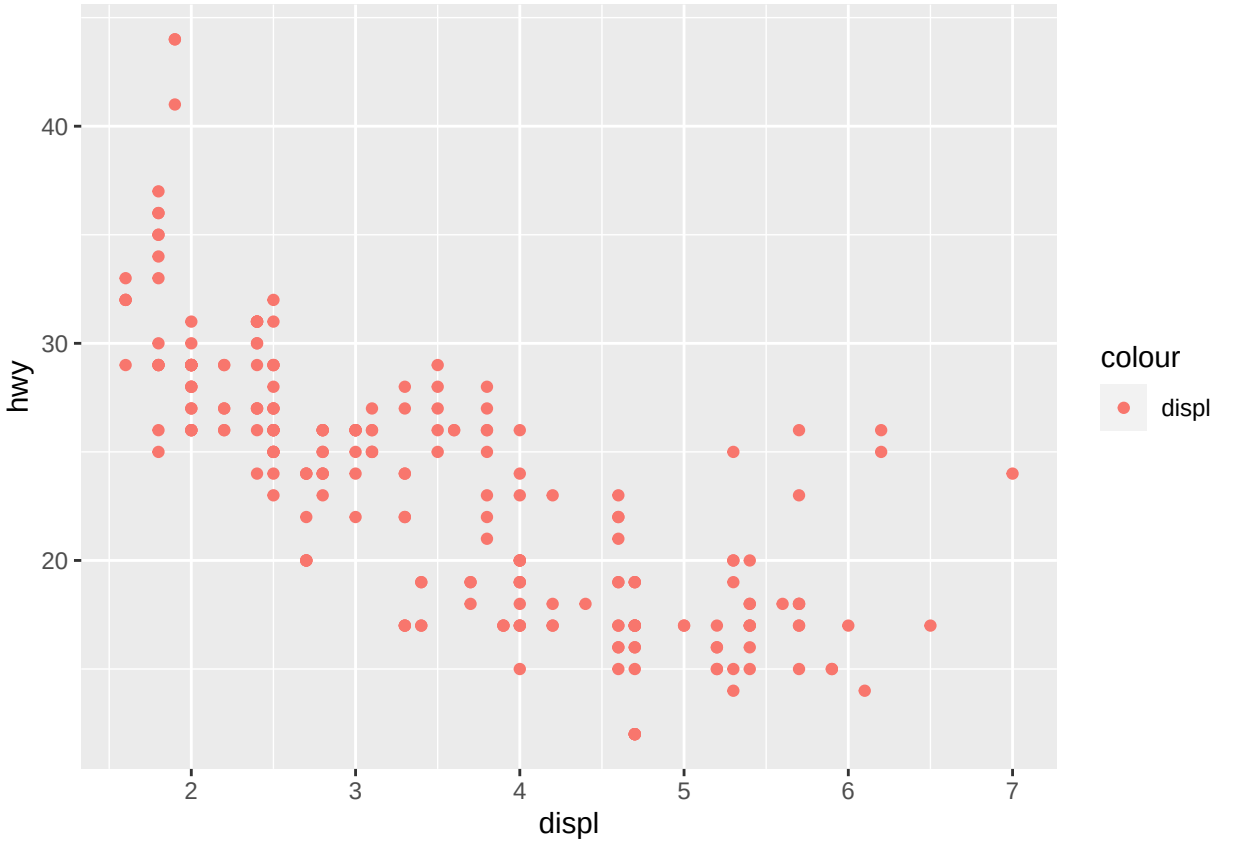


Figure 3: A caption

#8c. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon). Mapped it with a continuous variable you have identified in #5-b. What is its result? Why it produced such output?

```
ggplot(data = mpgColRow,
       mapping = aes(x = displ, y = hwy)) +
  geom_point (mapping = aes(colour = "displ"))
```

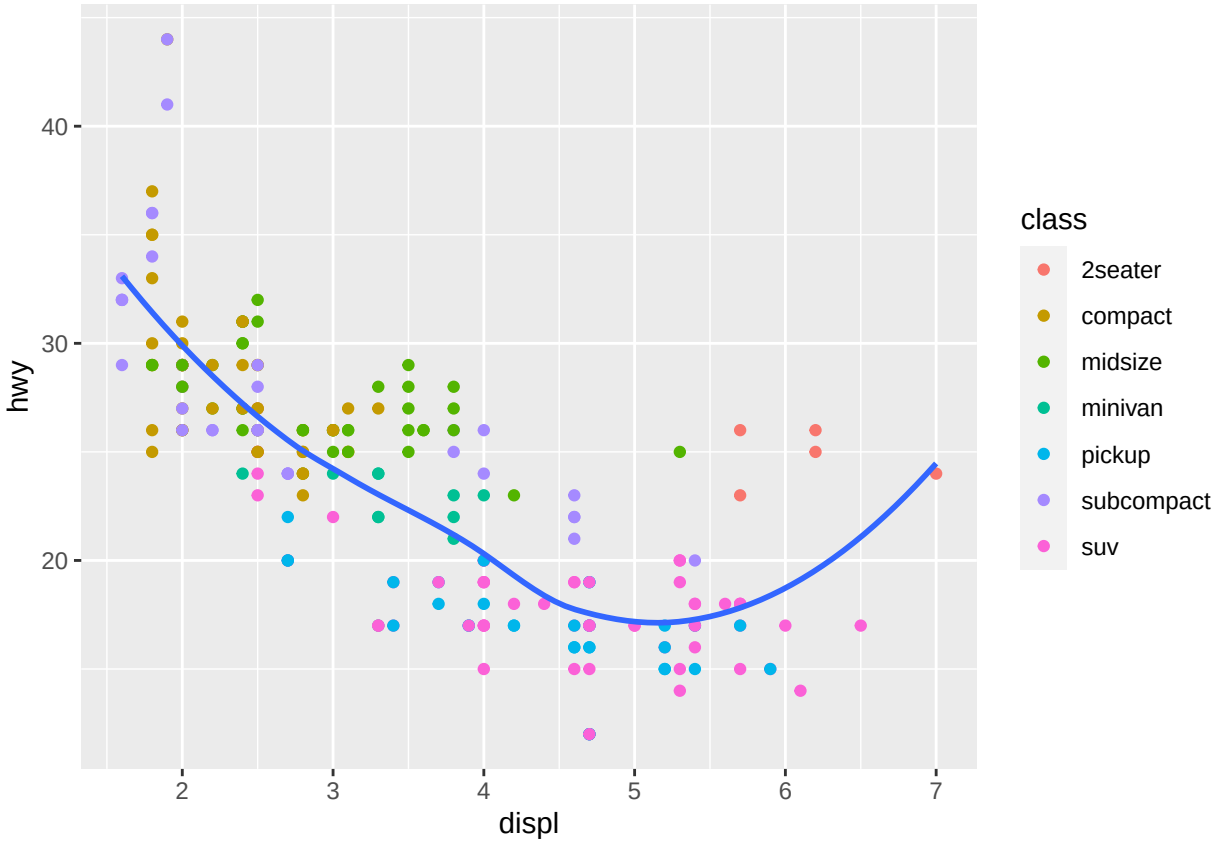


#The plot reveals the relationship between the car's engine displacement and the highway per gallons it consume.

#9. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon). Mapped it with a continuous variable you have identified in #5-b. What is its result? Why it produced such output?

```
ggplot(data = mpgColRow, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping=aes(color=class)) +
  geom_smooth(se = FALSE)
```

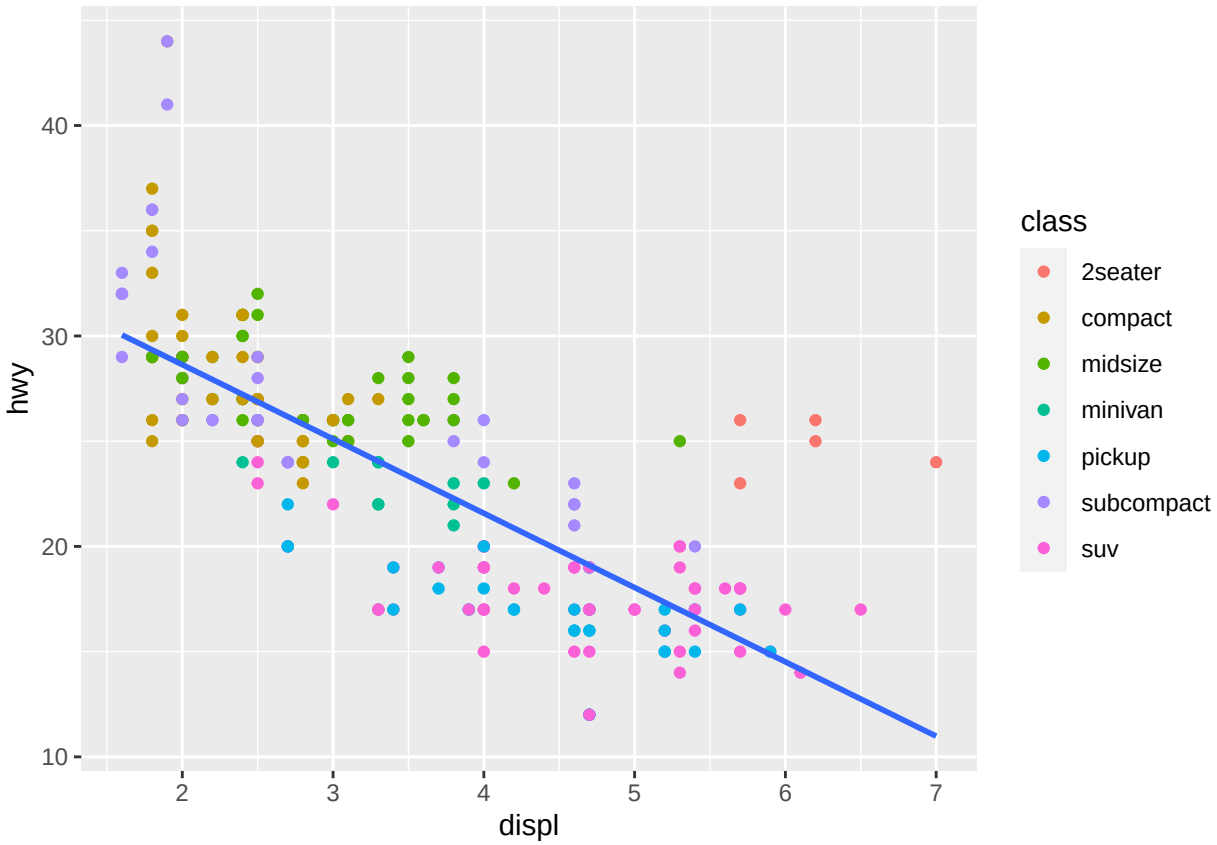
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



#10. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon). Mapped it with a continuous variable you have identified in #5-b. What is its result? Why it produced such output?

```
ggplot(data = mpgColRow, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping=aes(color=class)) +
  geom_smooth(se = FALSE, method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


hwy
40
30
20
10
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv