Work Sheet 6

Wendy Nalaza

2022-11-25

Use the dataset mpg

```
library(ggplot2)
data(mpg)
as.data.frame(data(mpg))
```

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

data(mpg)

```
data("mpg")
str("mpg")
```

```
##  chr "mpg"
```

```
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
```

```
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~
```

Example. graph using ggplot()

ggplot(mpg, aes(cty, hwy)) + geom_point()

1. How many columns are in mpg dataset? How about the number of rows? Show the codes and its result. *ANSWER: The total of columns are 11 columns and 234 rows*

```
ROW <- nrow(mpg)
COLUMN <-ncol(mpg)
ROW
```

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

```
COLUMN
```

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

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

```
## # A tibble: 15 x 2
##    manufacturer     n
##    <chr>        <int>
##  1 dodge           37
##  2 toyota          34
##  3 volkswagen      27
##  4 ford            25
##  5 chevrolet       19
##  6 audi            18
##  7 hyundai         14
##  8 subaru          14
##  9 nissan          13
## 10 honda            9
## 11 jeep             8
## 12 pontiac          5
## 13 land rover       4
## 14 mercury          4
## 15 lincoln          3
```

ANSWER: Dodge and has 37 models

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

```
DATAmpg <- mpg
Manufacturer2 <- DATAmpg %>% group_by(manufacturer, model) %>%
  distinct() %>% count()
Manufacturer2
```

```
## # 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
```

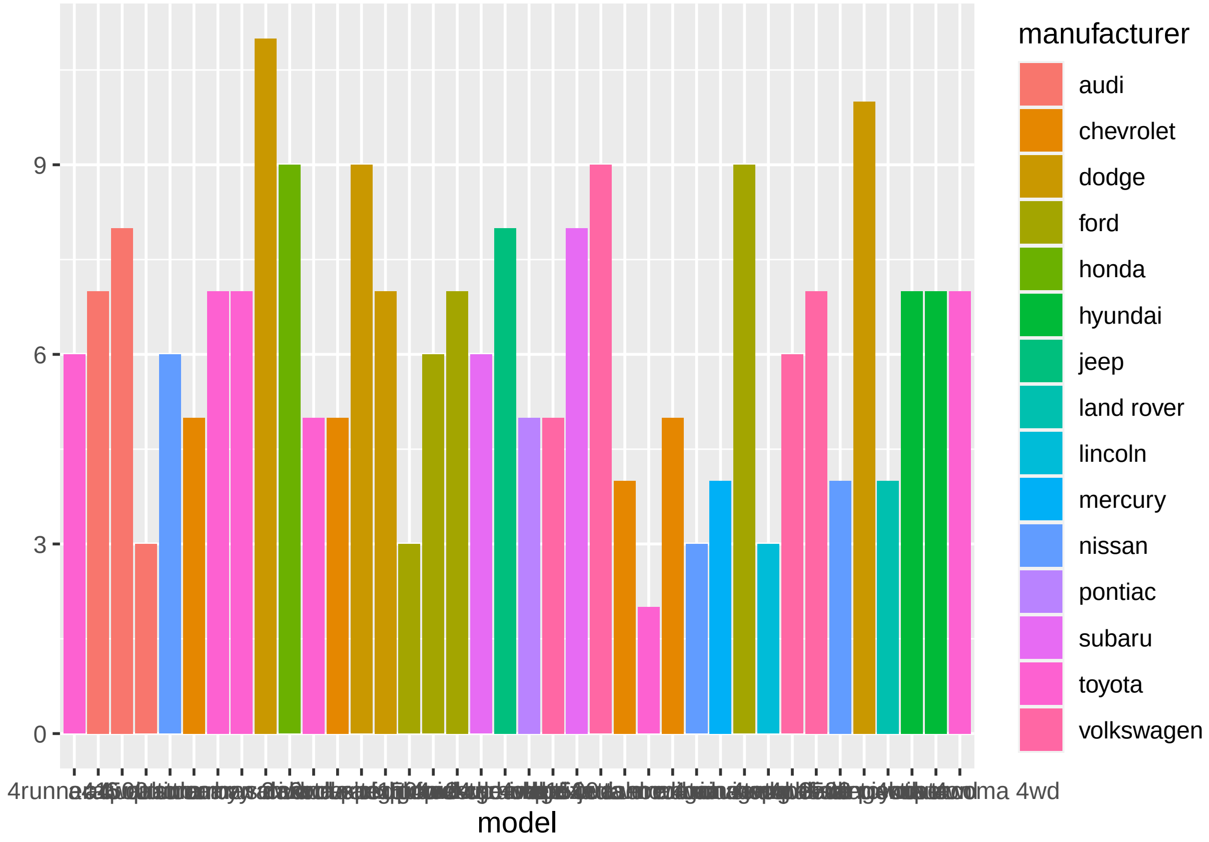
```
colnames(Manufacturer2) <- c("Manufacturer", "Model","Counts")
Manufacturer2
```

```
## # A tibble: 38 x 3
## # Groups:   Manufacturer, Model [38]
##    Manufacturer Model              Counts
##    <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
```

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

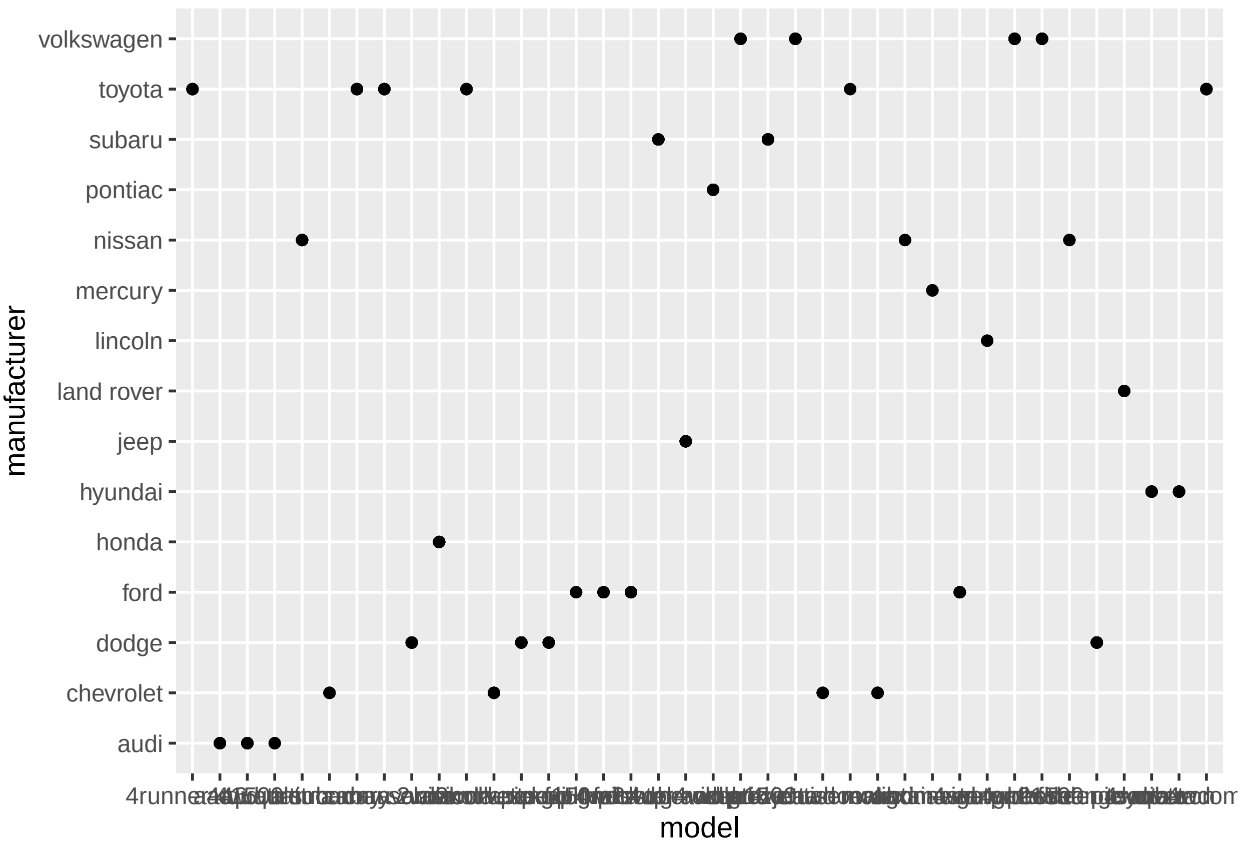
- **plot**

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



- **ggplot**

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



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

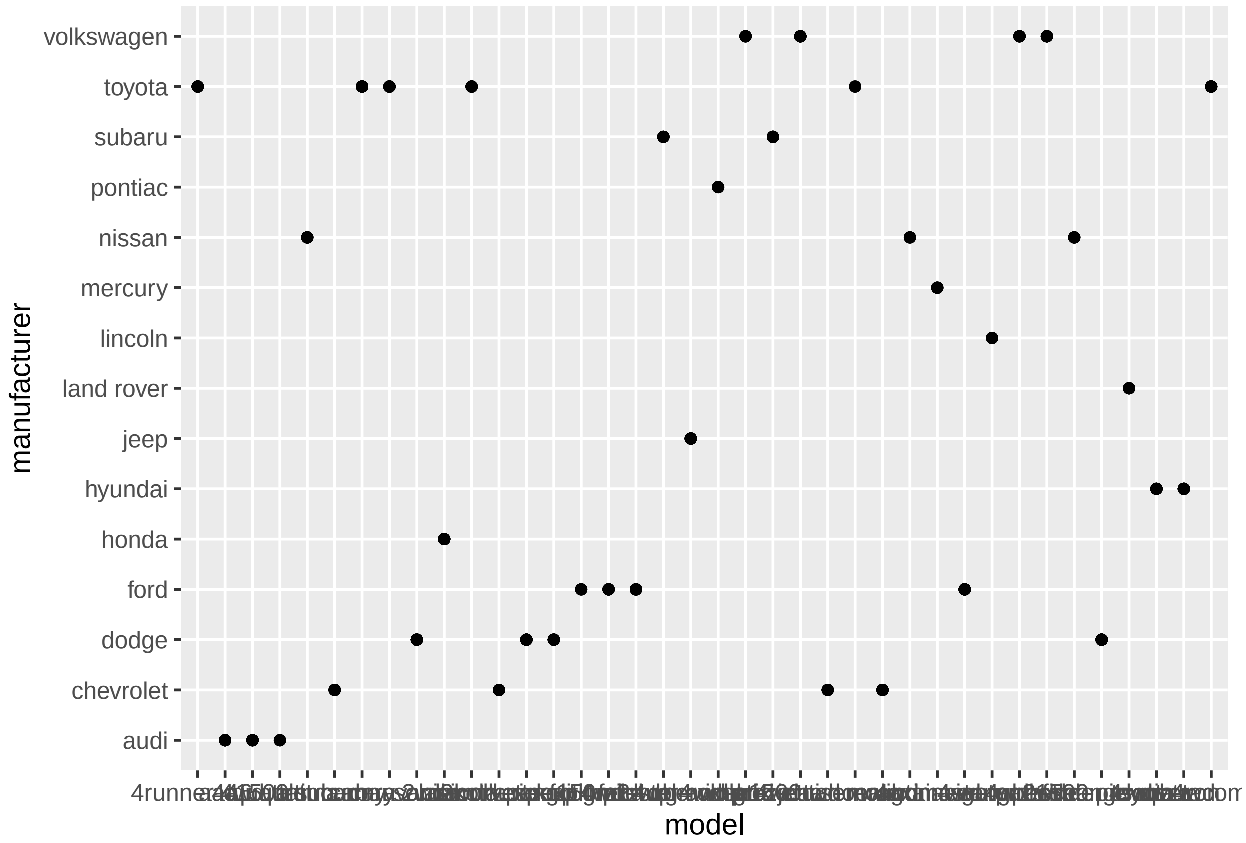
```
DATAmpg <- mpg
Manufacturer3 <- DATAmpg %>% group_by(manufacturer, model) %>%
  distinct() %>% count()

Manufacturer3
```

```
## # 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(Manufacturer3) <- c("Manufacturer", "Model")
Manufacturer3
```

```
## # A tibble: 38 x 3
## # Groups:   Manufacturer, Model [38]
##    Manufacturer Model                ‘‘
##    <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
```

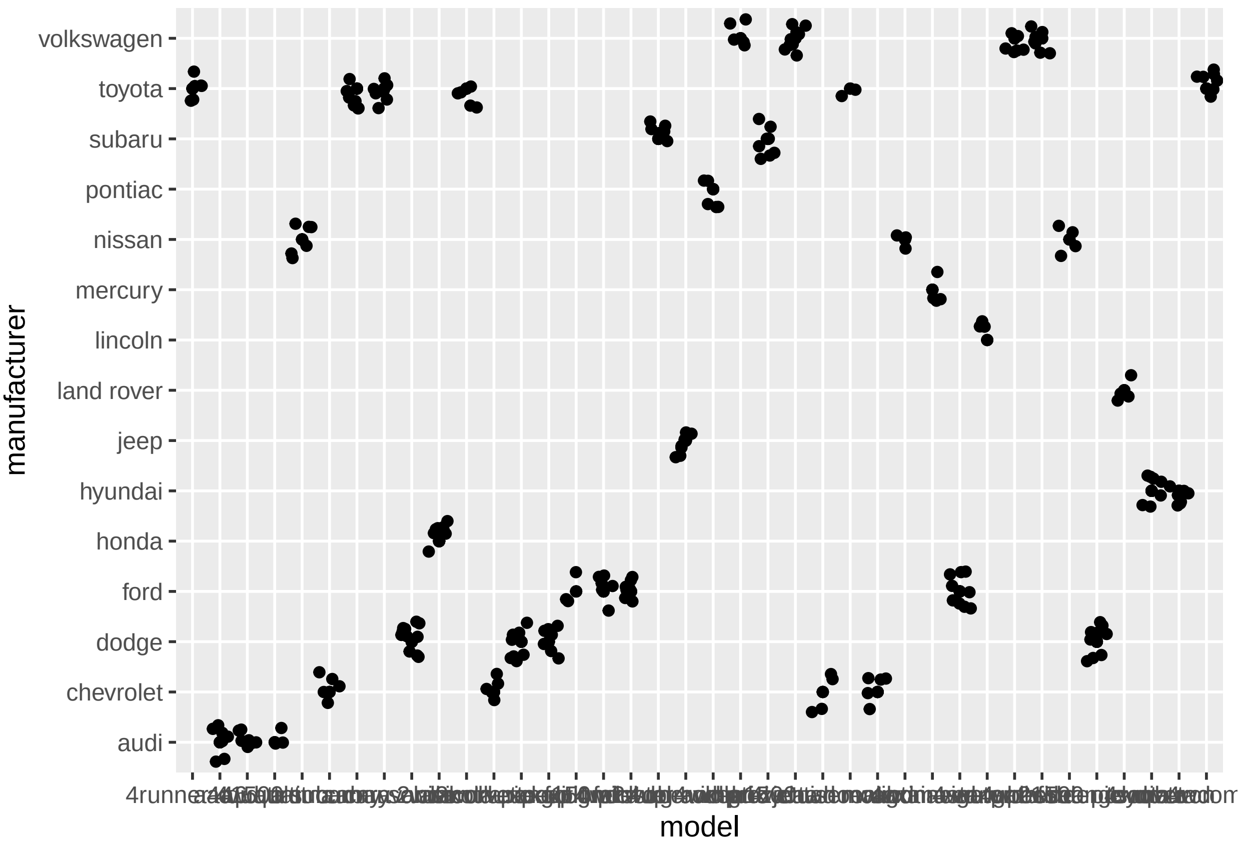
a. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?



ANSWER: Geometric point graph of mpg(model and manufacturer)

b. For you, is it useful? If not, how could you modify the data to make it more informative? *ANSWER: Yes, it is helpful since you can track down and edit each model's data directly from the manufacturer.* + **to modify the data:**

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



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

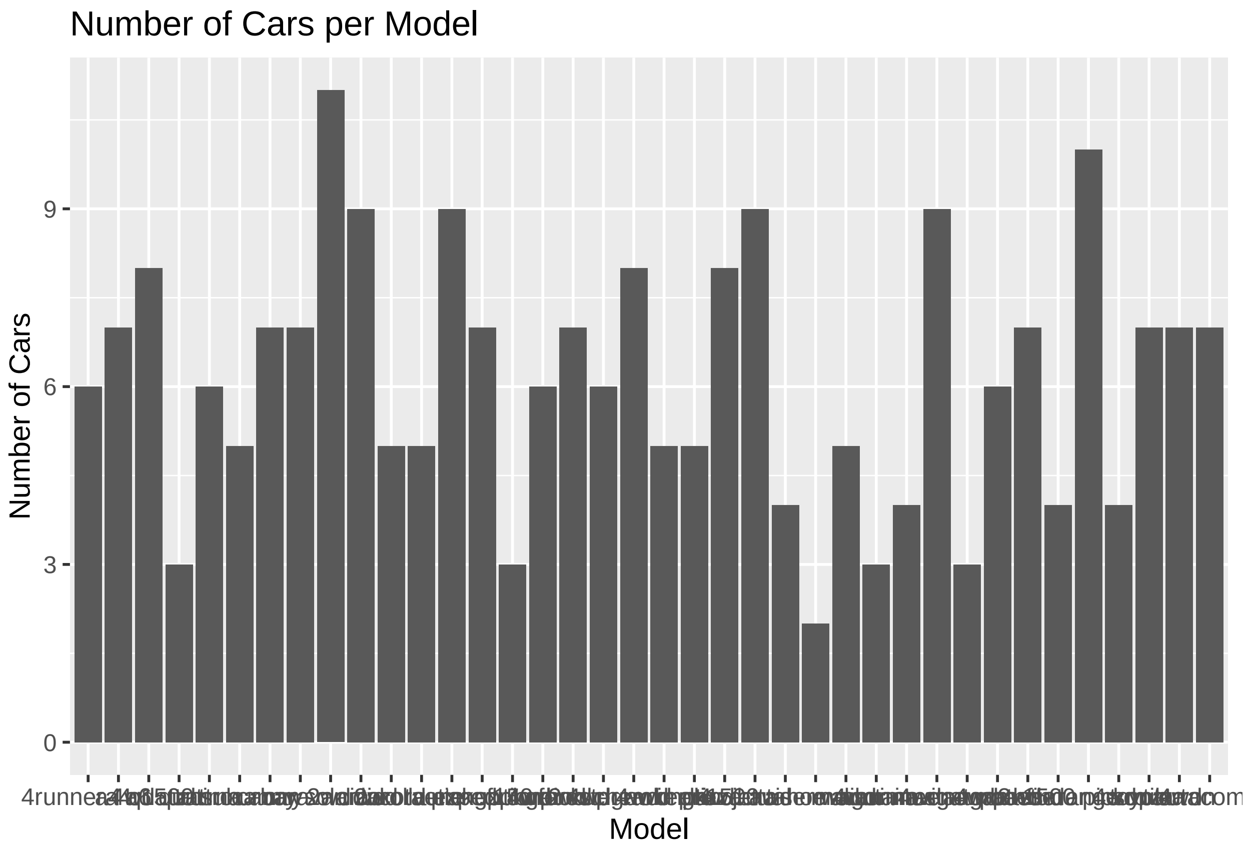
```
DATAmpg4 <- Manufacturer2 %>% group_by(Model) %>% count()
DATAmpg4
```

```
## # A tibble: 38 x 2
## # Groups:   Model [38]
##    Model                   n
##    <chr>               <int>
##  1 4runner 4wd             1
##  2 a4                      1
##  3 a4 quattro              1
##  4 a6 quattro              1
##  5 altima                  1
##  6 c1500 suburban 2wd      1
##  7 camry                   1
##  8 camry solara            1
##  9 caravan 2wd             1
## 10 civic                   1
## # ... with 28 more rows
```

```
colnames(DATAmpg4) <- c("Model","Counts")
```

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

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



b. Use only the top 20 observations. Show code and results.

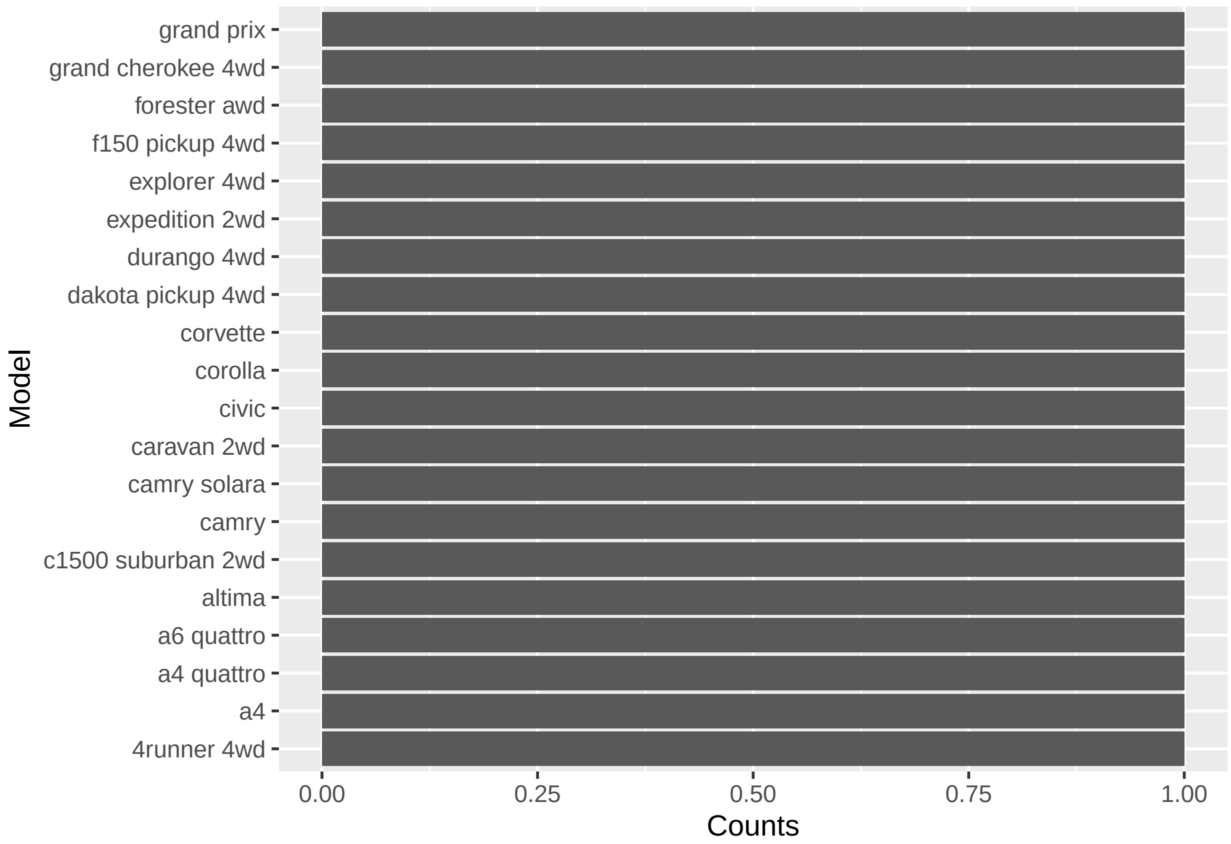
```
Top_Data <- DATAmpg4[1:20,]%>%top_n(2)
```

```
## Selecting by Counts
```

```
Top_Data
```

```
## # A tibble: 20 x 2
## # Groups:   Model [20]
##    Model              Counts
##    <chr>               <int>
##  1 4runner 4wd             1
##  2 a4                      1
##  3 a4 quattro              1
##  4 a6 quattro              1
##  5 altima                  1
##  6 c1500 suburban 2wd      1
##  7 camry                   1
##  8 camry solara            1
##  9 caravan 2wd             1
## 10 civic                   1
## 11 corolla                 1
## 12 corvette                1
## 13 dakota pickup 4wd       1
## 14 durango 4wd             1
## 15 expedition 2wd          1
## 16 explorer 4wd            1
## 17 f150 pickup 4wd         1
## 18 forester awd            1
## 19 grand cherokee 4wd      1
## 20 grand prix              1
```

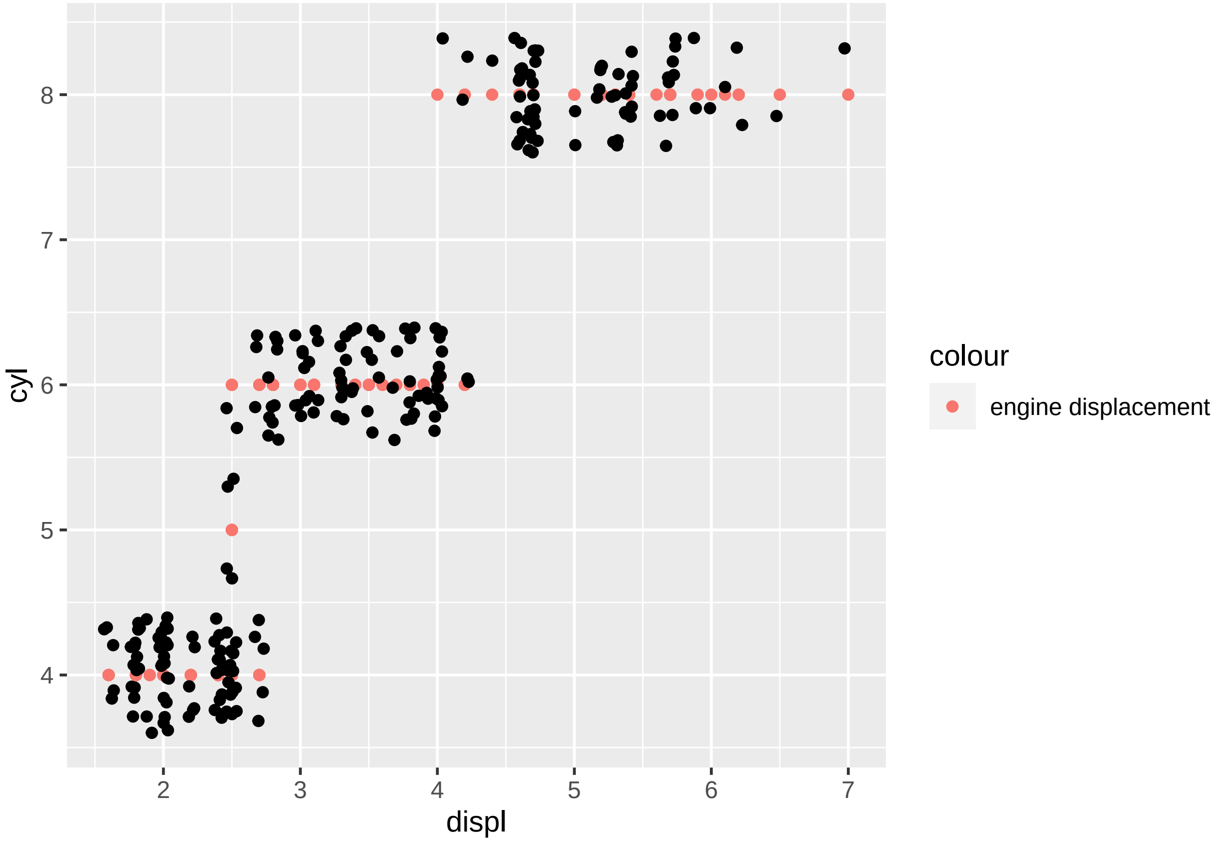
```
ggplot(Top_Data,aes(x = Model,y =Counts)) + geom_bar(stat = "Identity") + coord_flip()
```



5. 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".

a. Show the codes and its result.

```
ggplot(data = mpg , mapping = aes(x = displ, y = cyl,
      main = "Relationship between No of Cylinders and Engine Displacement")) +
  geom_point(mapping=aes(colour = "engine displacement")) + geom_jitter()
```



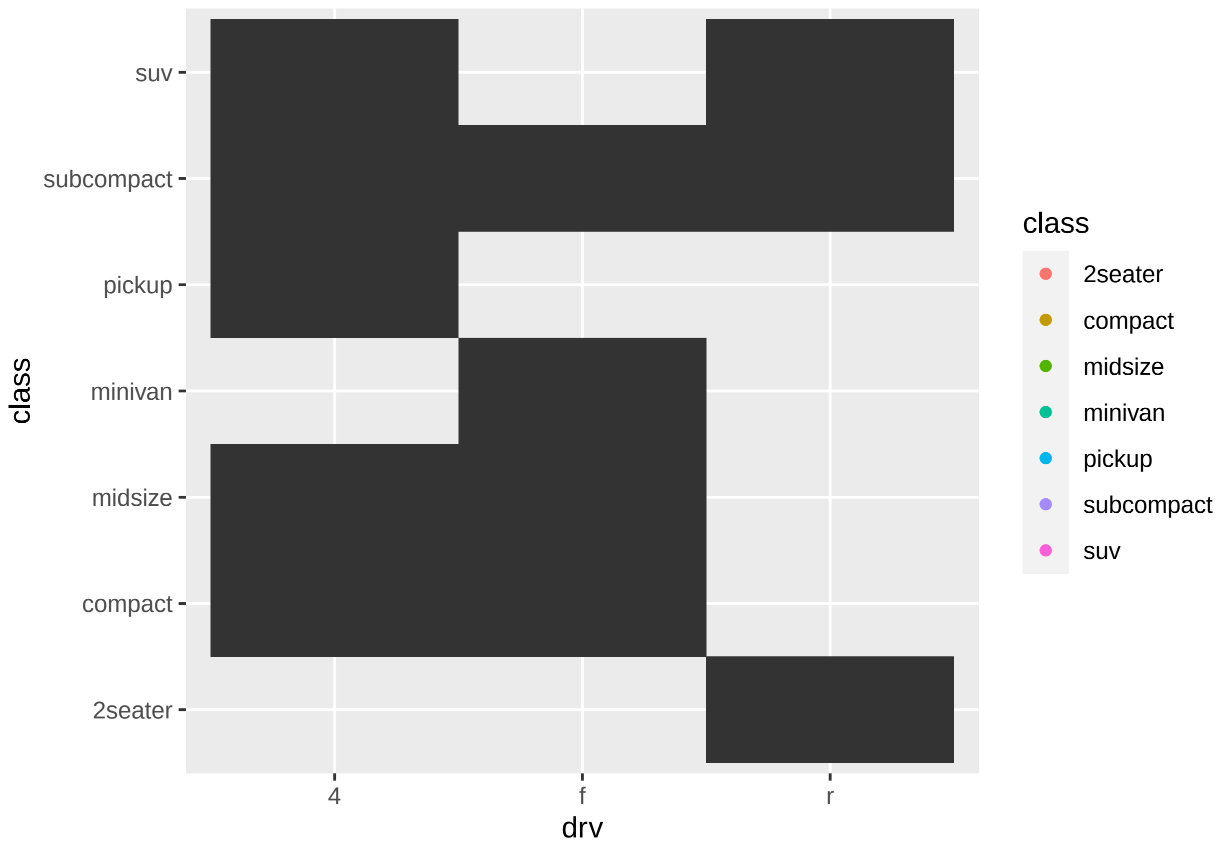
b. How would you describe its relationship?

ANSWER: So according to the data, by making cyl into y, the graph is scattered, and the pink color indicates the engine displacement, as you can see from the dots in a straight horizontal position.

6.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.

a. Show the codes and its result for the narrative in 6.

```
ggplot(data = mpg, mapping = aes(x = drv, y = class)) +
  geom_point(mapping=aes(color=class)) +
  geom_tile()
```



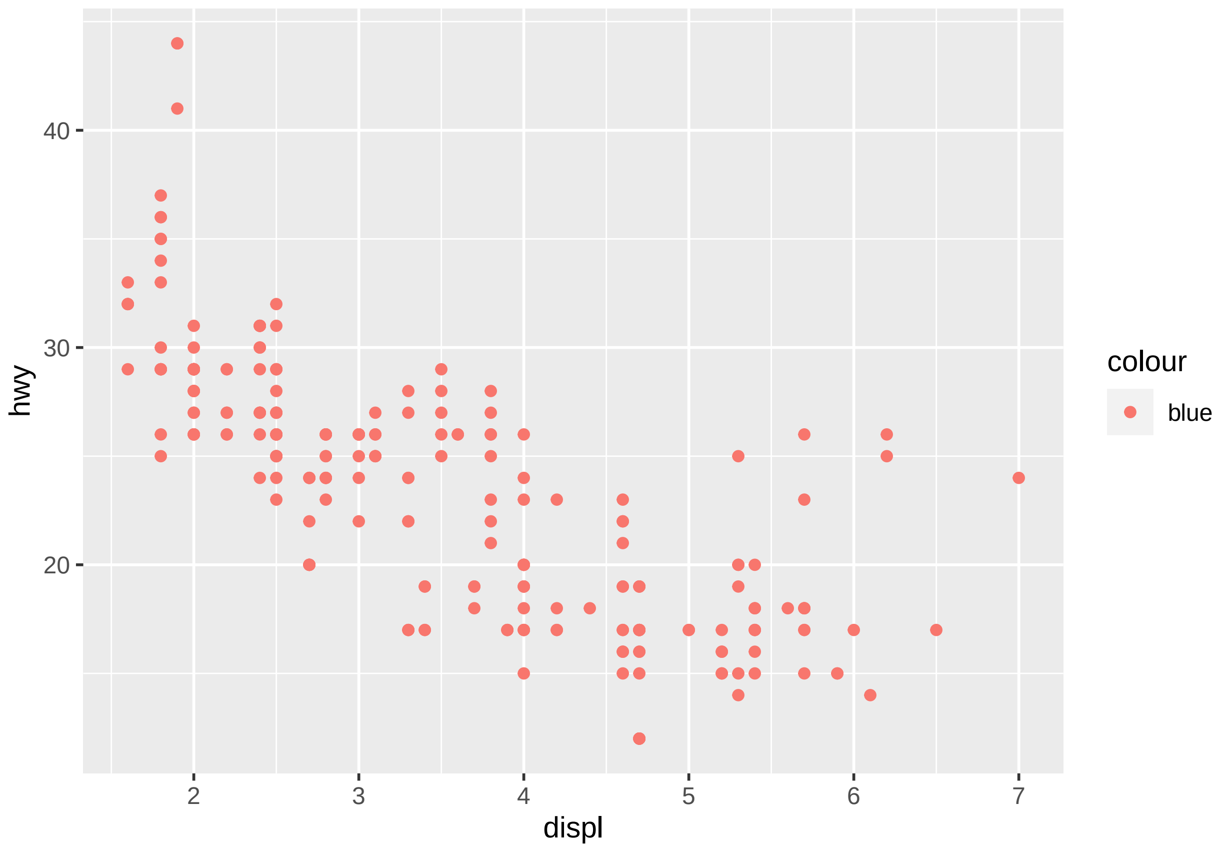
b. Interpret the result:

ANSWER : Areas covered with black are "mapped" using the mapping geometric point graph, with y as class and x as drv.

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

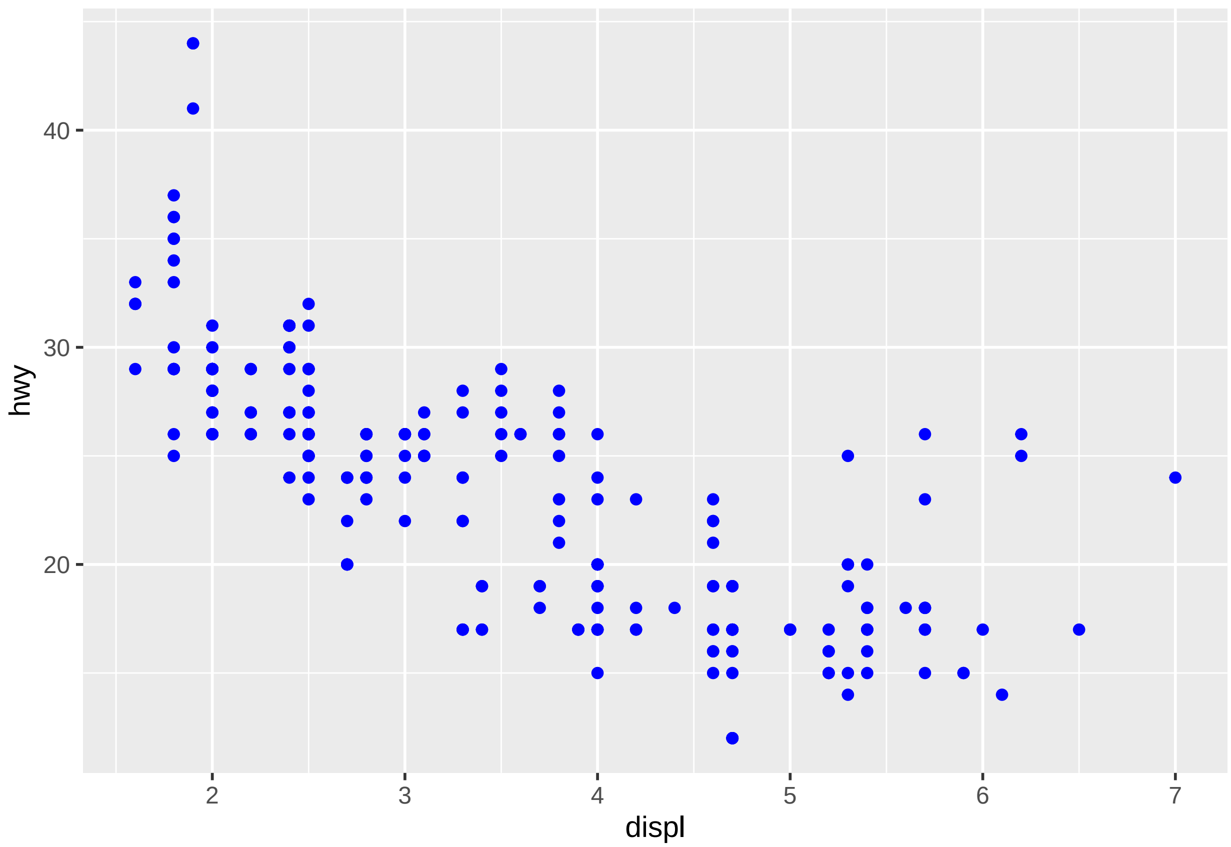
- **Code 1**

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



- **+ Code 2**

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



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

```
?mpg
```

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

The result of the command are the server website and the data of mpg.

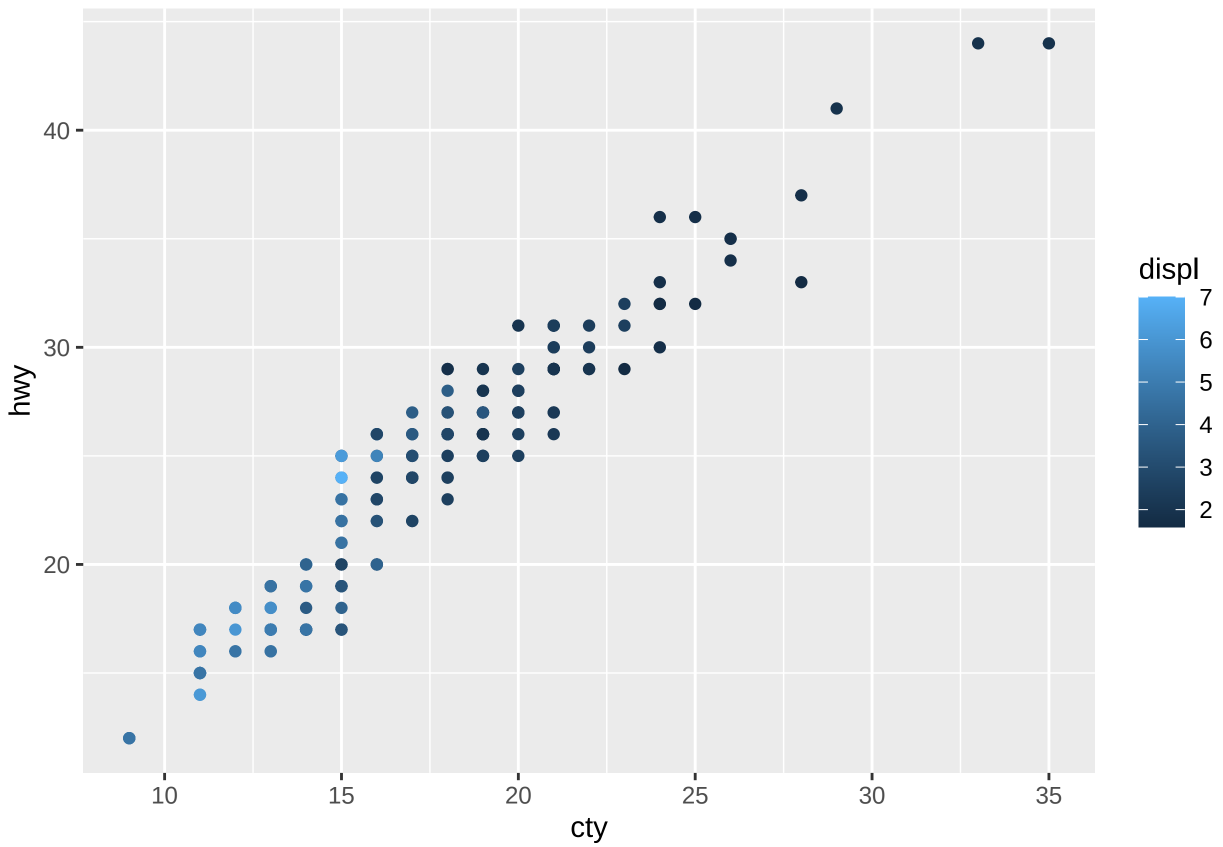
a. Which variables from mpg dataset are categorical?

ANSWER: Categorical variables in mpg which include: the manufacturer, model, trans (type of transmission), drv (front-wheel drive, rear-wheel, 4wd), fl (fuel type), and class (type of car).

b. Which are continuous variables? *ANSWERS: Continuous varibles in R were also known as doubles or integers.*

c. 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.

```
ggplot(mpg, aes(x = cty, y = hwy, colour = displ)) + geom_point()
```

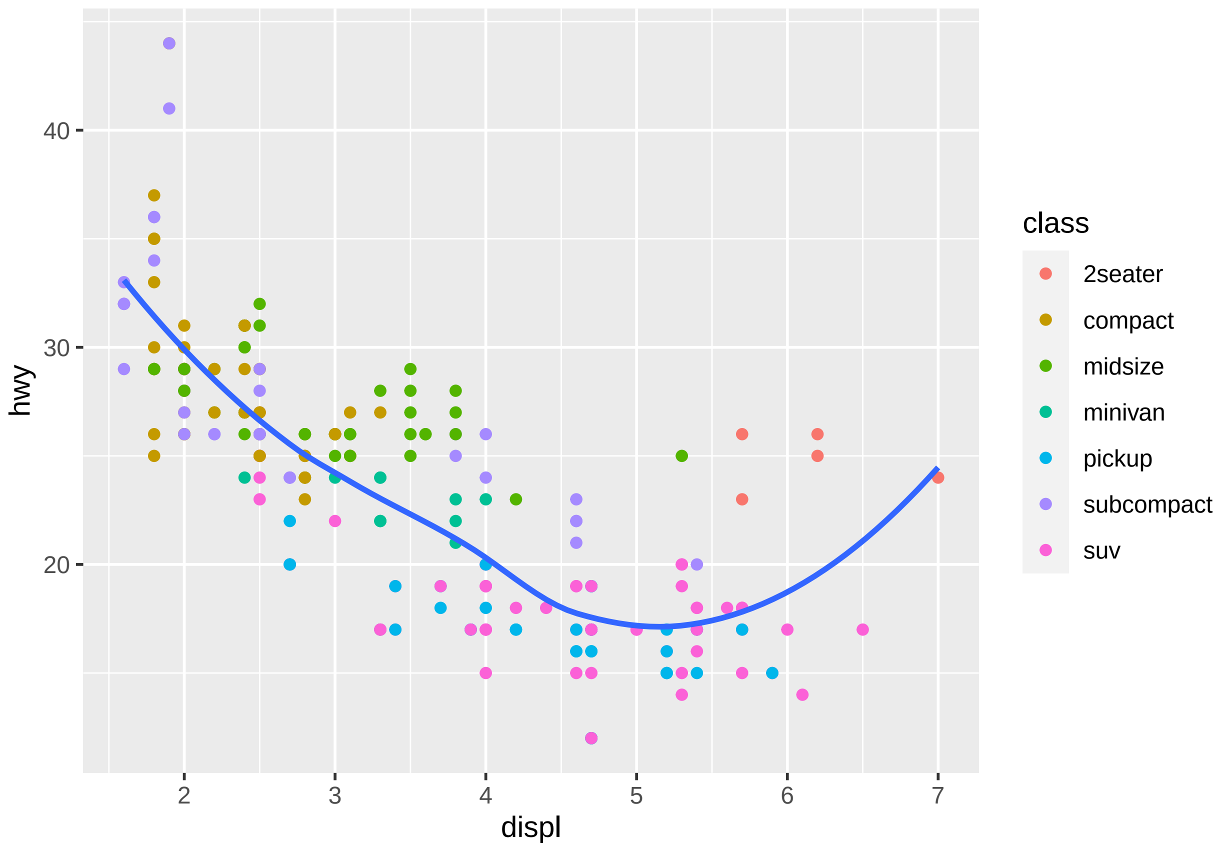


What is its result? Why it produced such output? *ANSWER: Data tracks the cty by showing (city miles per gallon) in a color with a blue hue or variation of blue.*

9.Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon) using geom_point(). Add a trend line over the existing plot using geom_smooth() with se = FALSE. Default method is "loess".

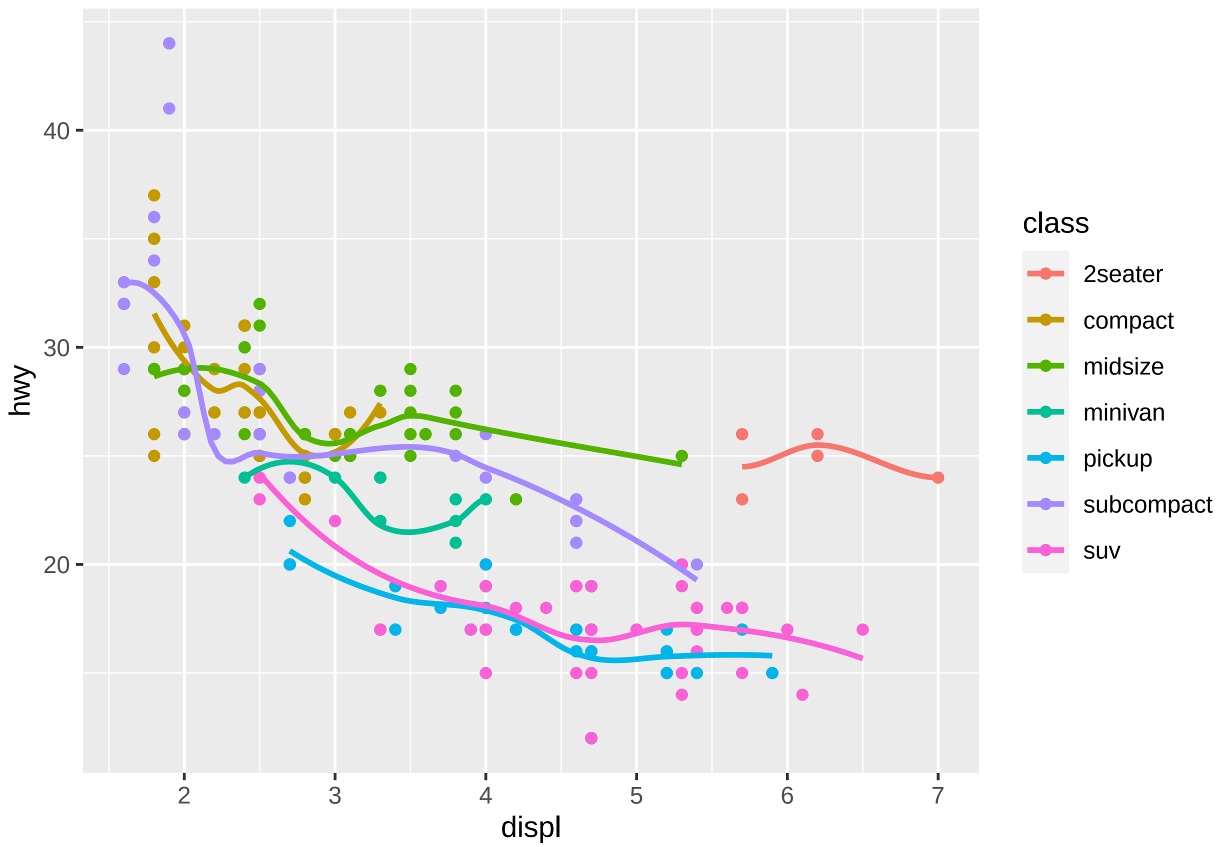
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping=aes(color=class)) +
  geom_smooth(se = FALSE)
```

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



10. Using the relationship of displ and hwy, add a trend line over existing plot. Set the se = FALSE to remove the confidence interval and method = lm to check for linear modeling

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


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