Acosta-Worksheet6

ACOSTA, MELBOURNE BSIT2A

2022-12-11

```
library(ggplot2)
```

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

```
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
```

```
library(tinytex)


data("mpg")
#1. How many columns are in mpg dataset? How about the number of rows?
# show the codes and its result.
ncol(mpg)
```

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

```
nrow(mpg)
```

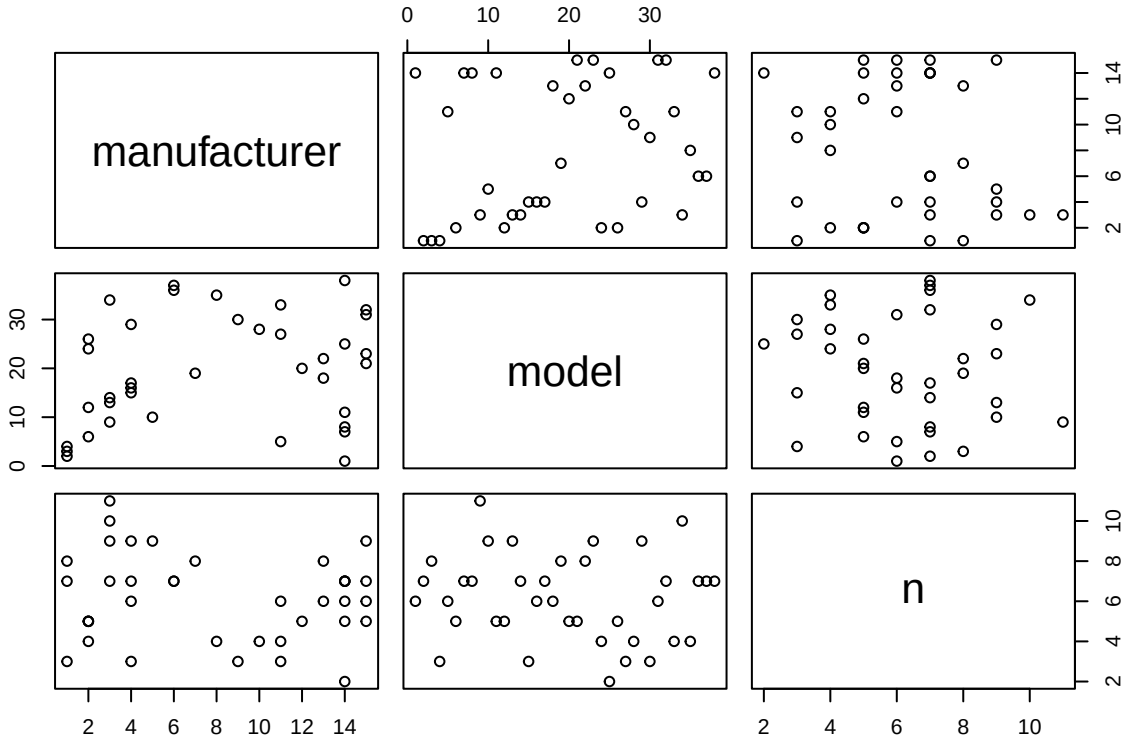
```
## [1] 234
```

```
# There are 11 columns and 234 rows in mpg.
```

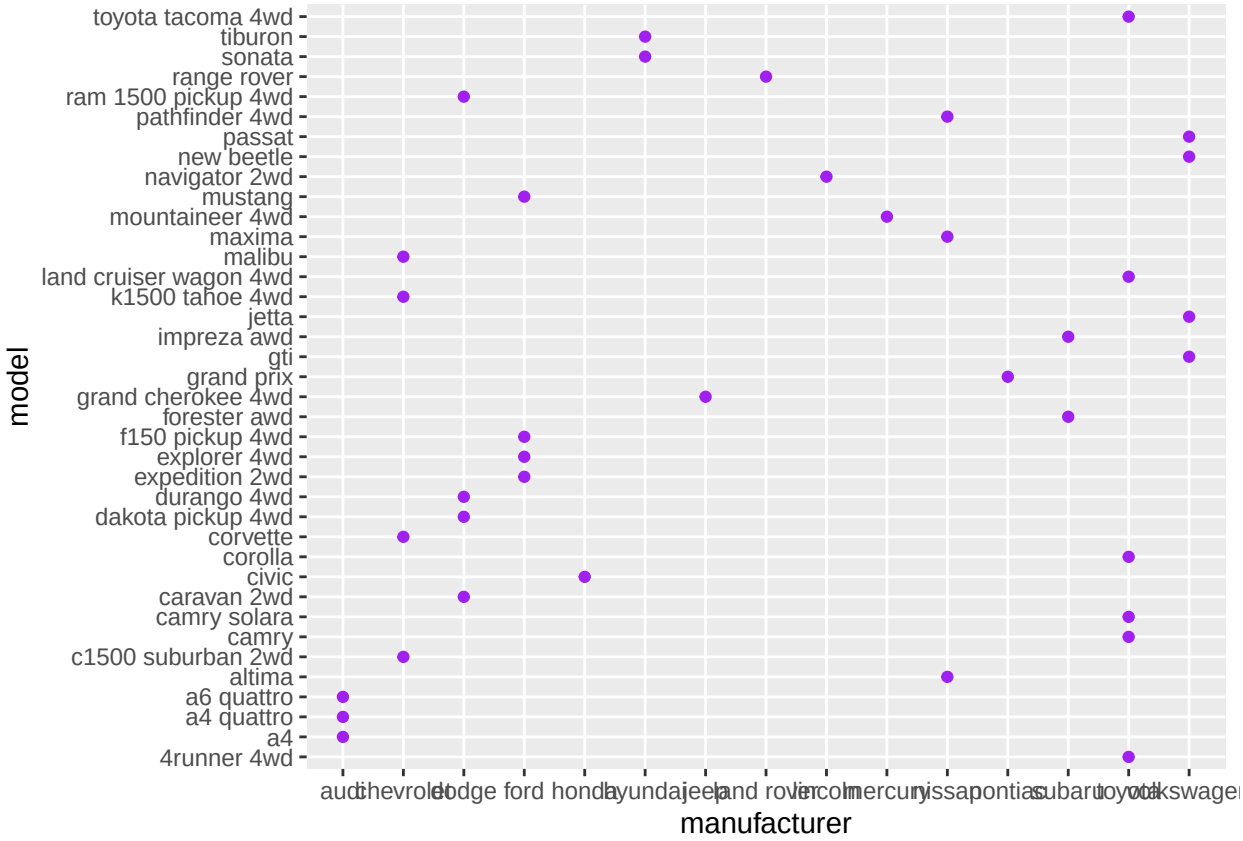
```
#2. Which manufacturer has the most models in this data set? Which model has the most
#variations?
# The manufacturer with the most model is Dodge. While the model with the most variation is
# caravan 2wd
```

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

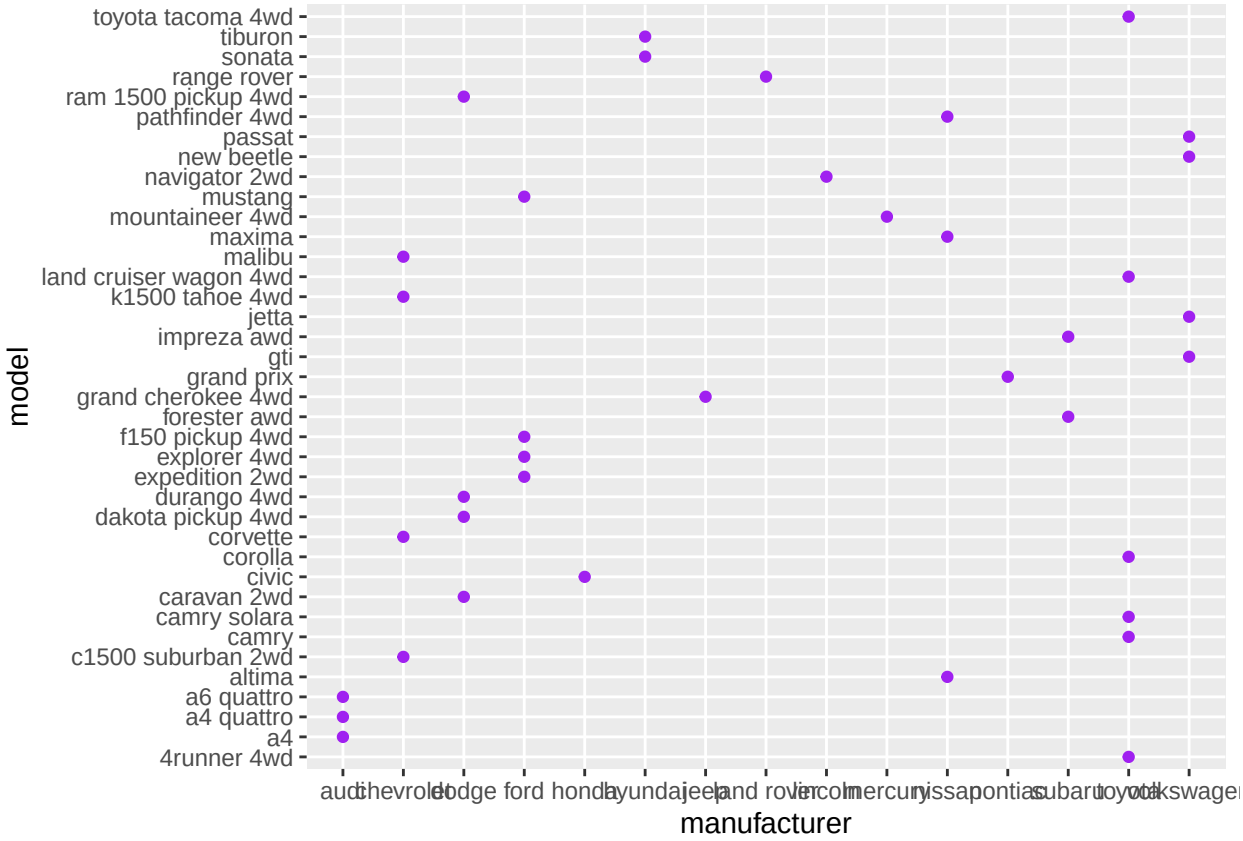
```
#B. Graph the result by using plot() and ggplot(). Write the codes and its result.
d <- mpg %>% group_by (manufacturer, model) %>% count()

plot(d)
```



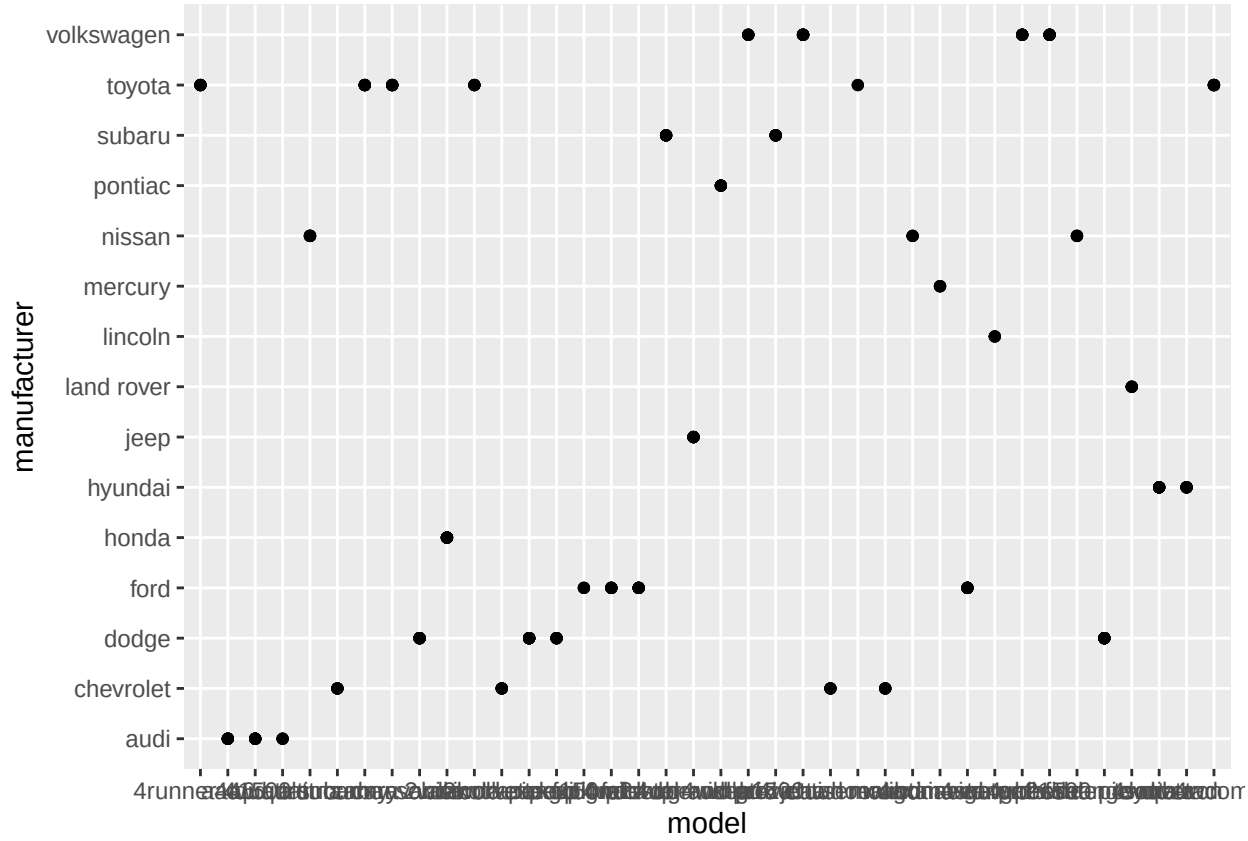

```
ggplot(d, aes(x = manufacturer, y = model)) + geom_point(color = 'purple')
```




```
#3. Same dataset will be used. You are going to show the relationship of the modeland
#the manufacturer.
ggplot(d, aes(x = manufacturer, y = model)) + geom_point(color = 'purple')
```




```
#A. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?
ggplot(mpg, aes(model, manufacturer)) + geom_point()
```




```
# The plot shows the relationship of the model and manufacturer.
```

```
#B. For you, is it useful? If not, how could you modify the data to make it more
#informative?
#Yes, it is useful as it makes the data easier to understand.
```

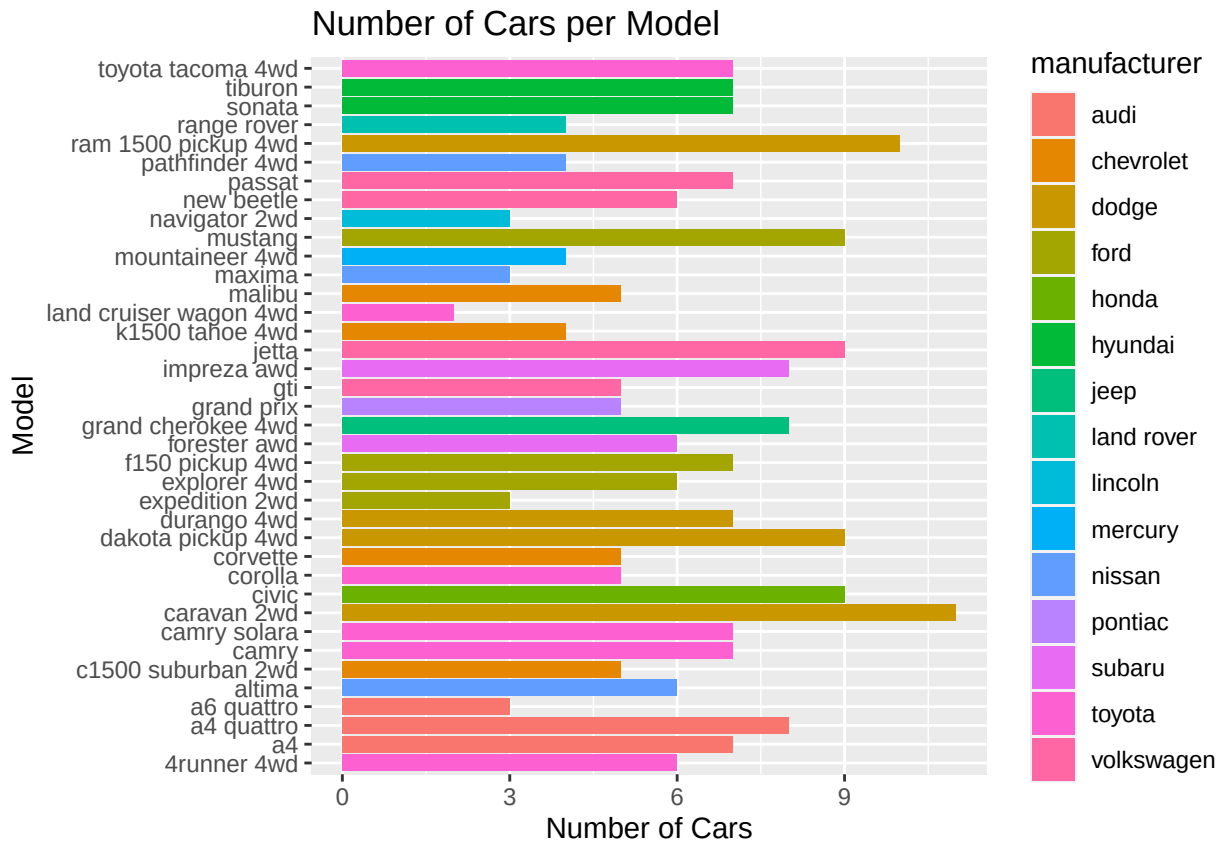
```
#4. Using the pipe (%>%), group the model and get the number of cars per model. Show
#codes and its result.
grouped_model <- mpg %>% group_by (model) %>% count()
grouped_model
```

```
## # 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
```

```
#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()
```

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




```
#B. Use only the top 20 observations. Show code and results.
top <- model[1:20,] %>% top_n(2)
```

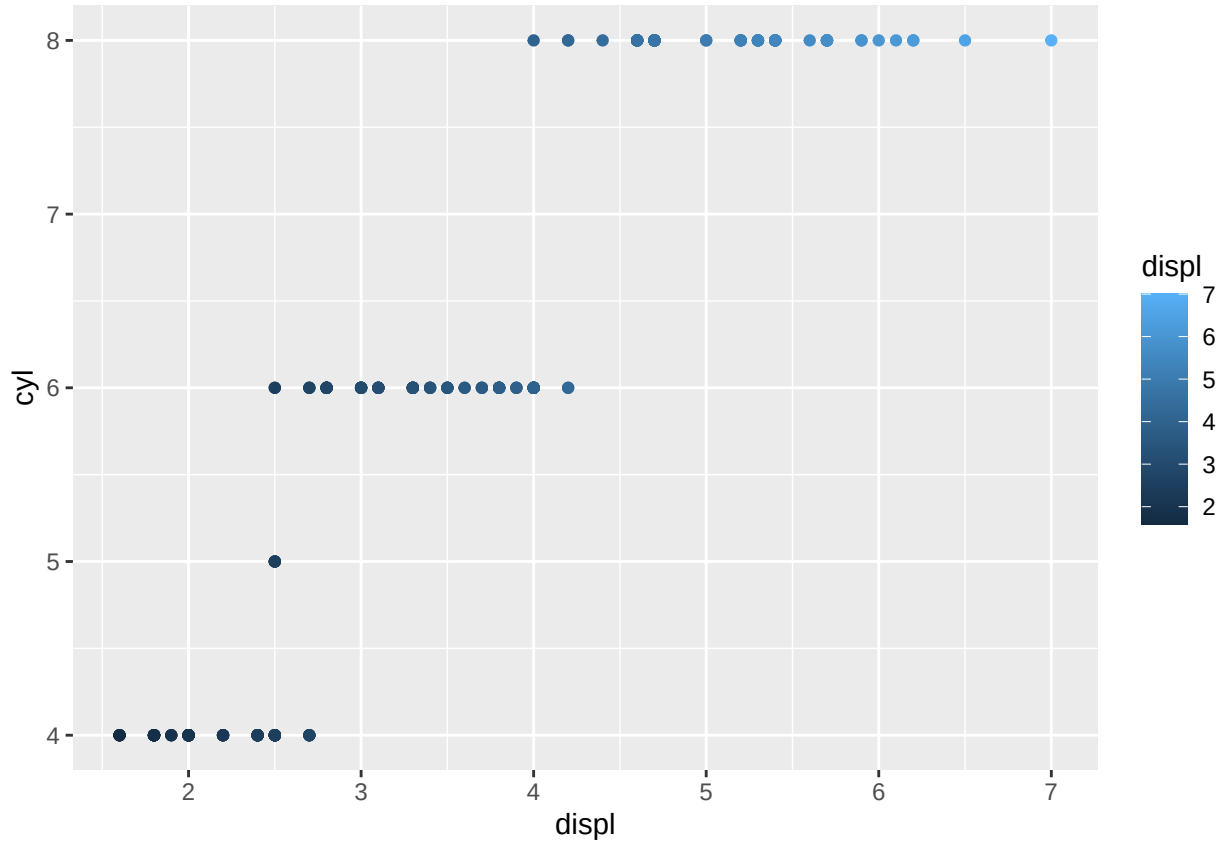
```
## Selecting by n
```

```
top
```

```
## # A tibble: 20 x 2
## # Groups:   model [20]
##    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
## 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
```

```
#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".

ggplot(data = mpg, mapping = aes(x = displ, y = cyl)) +
  geom_point(mapping=aes(color=displ))
```



```
#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.
front_wheel <- subset(mpg, drv == 'f')
nrow(front_wheel)
```

```
## [1] 106
```

```
front_wheel
```

```
## # A tibble: 106 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(l5)  f        18    29 p     comp~
##  2 audi         a4       1.8  1999     4 manual(m~ f        21    29 p     comp~
##  3 audi         a4       2    2008     4 manual(m~ f        20    31 p     comp~
##  4 audi         a4       2    2008     4 auto(av)  f        21    30 p     comp~
##  5 audi         a4       2.8  1999     6 auto(l5)  f        16    26 p     comp~
##  6 audi         a4       2.8  1999     6 manual(m~ f        18    26 p     comp~
##  7 audi         a4       3.1  2008     6 auto(av)  f        18    27 p     comp~
##  8 chevrolet    malibu   2.4  1999     4 auto(l4)  f        19    27 r     mids~
##  9 chevrolet    malibu   2.4  2008     4 auto(l4)  f        22    30 r     mids~
## 10 chevrolet    malibu   3.1  1999     6 auto(l4)  f        18    26 r     mids~
## # ... with 96 more rows
```

```
rear_wheel <- subset(mpg, drv == 'r')
nrow(rear_wheel)
```

```
## [1] 25
```

```
rear_wheel
```

```
## # A tibble: 25 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 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        14    20 r     suv
##  2 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        11    15 e     suv
##  3 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        14    20 r     suv
##  4 chevrolet    c1500 sub~   5.7  1999     8 auto~ r        13    17 r     suv
##  5 chevrolet    c1500 sub~   6    2008     8 auto~ r        12    17 r     suv
##  6 chevrolet    corvette     5.7  1999     8 manu~ r        16    26 p     2sea~
##  7 chevrolet    corvette     5.7  1999     8 auto~ r        15    23 p     2sea~
##  8 chevrolet    corvette     6.2  2008     8 manu~ r        16    26 p     2sea~
##  9 chevrolet    corvette     6.2  2008     8 auto~ r        15    25 p     2sea~
## 10 chevrolet    corvette     7    2008     8 manu~ r        15    24 p     2sea~
## # ... with 15 more rows
```

```r
n4 <- subset(mpg, drv == '4')
nrow(n4)
```

```
## [1] 103
```

```r
n4
```

```
## # A tibble: 103 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 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  2 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
##  3 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
##  4 audi         a4 quattro   2    2008     4 auto~ 4        19    27 p     comp~
##  5 audi         a4 quattro   2.8  1999     6 auto~ 4        15    25 p     comp~
##  6 audi         a4 quattro   2.8  1999     6 manu~ 4        17    25 p     comp~
##  7 audi         a4 quattro   3.1  2008     6 auto~ 4        17    25 p     comp~
##  8 audi         a4 quattro   3.1  2008     6 manu~ 4        15    25 p     comp~
##  9 audi         a6 quattro   2.8  1999     6 auto~ 4        15    24 p     mids~
## 10 audi         a6 quattro   3.1  2008     6 auto~ 4        17    25 p     mids~
## # ... with 93 more rows
```

```r
suv <- subset(mpg, class == 'suv')
nrow(suv)
```

```
## [1] 62
```

```r
suv
```

```
## # A tibble: 62 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 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        14    20 r     suv
##  2 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        11    15 e     suv
##  3 chevrolet    c1500 sub~   5.3  2008     8 auto~ r        14    20 r     suv
##  4 chevrolet    c1500 sub~   5.7  1999     8 auto~ r        13    17 r     suv
##  5 chevrolet    c1500 sub~   6    2008     8 auto~ r        12    17 r     suv
##  6 chevrolet    k1500 tah~   5.3  2008     8 auto~ 4        14    19 r     suv
##  7 chevrolet    k1500 tah~   5.3  2008     8 auto~ 4        11    14 e     suv
##  8 chevrolet    k1500 tah~   5.7  1999     8 auto~ 4        11    15 r     suv
##  9 chevrolet    k1500 tah~   6.5  1999     8 auto~ 4        14    17 d     suv
## 10 dodge        durango 4~   3.9  1999     6 auto~ 4        13    17 r     suv
## # ... with 52 more rows
```

```r
com <- subset(mpg, class == 'compact')
nrow(com)
```

```
## [1] 47
```

```
com
```

```
## # A tibble: 47 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 37 more rows
```

```
m_size <- subset(mpg, class == 'midsize')
nrow(m_size)
```

```
## [1] 41
```

```
m_size
```

```
## # A tibble: 41 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         a6 quattro   2.8  1999     6 auto~ 4        15    24 p     mids~
##  2 audi         a6 quattro   3.1  2008     6 auto~ 4        17    25 p     mids~
##  3 audi         a6 quattro   4.2  2008     8 auto~ 4        16    23 p     mids~
##  4 chevrolet    malibu       2.4  1999     4 auto~ f        19    27 r     mids~
##  5 chevrolet    malibu       2.4  2008     4 auto~ f        22    30 r     mids~
##  6 chevrolet    malibu       3.1  1999     6 auto~ f        18    26 r     mids~
##  7 chevrolet    malibu       3.5  2008     6 auto~ f        18    29 r     mids~
##  8 chevrolet    malibu       3.6  2008     6 auto~ f        17    26 r     mids~
##  9 hyundai      sonata       2.4  1999     4 auto~ f        18    26 r     mids~
## 10 hyundai      sonata       2.4  1999     4 manu~ f        18    27 r     mids~
## # ... with 31 more rows
```

```
two_seater <- subset(mpg, class == '2seater')
nrow(two_seater)
```

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

```
two_seater
```

```
## # A tibble: 5 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 chevrolet    corvette   5.7  1999     8 manual(~ r        16    26 p     2sea~
## 2 chevrolet    corvette   5.7  1999     8 auto(14) r        15    23 p     2sea~
```

```
## 3 chevrolet    corvette   6.2  2008     8 manual(~ r        16    26 p     2sea~
## 4 chevrolet    corvette   6.2  2008     8 auto(s6) r        15    25 p     2sea~
## 5 chevrolet    corvette   7    2008     8 manual(~ r        15    24 p     2sea~
```

```
mini_van <- subset(mpg, class == 'minivan')
nrow(mini_van)
```

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

```
mini_van
```

```
## # A tibble: 11 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 dodge        caravan 2~   2.4  1999     4 auto~ f        18    24 r     mini~
##  2 dodge        caravan 2~   3    1999     6 auto~ f        17    24 r     mini~
##  3 dodge        caravan 2~   3.3  1999     6 auto~ f        16    22 r     mini~
##  4 dodge        caravan 2~   3.3  1999     6 auto~ f        16    22 r     mini~
##  5 dodge        caravan 2~   3.3  2008     6 auto~ f        17    24 r     mini~
##  6 dodge        caravan 2~   3.3  2008     6 auto~ f        17    24 r     mini~
##  7 dodge        caravan 2~   3.3  2008     6 auto~ f        11    17 e     mini~
##  8 dodge        caravan 2~   3.8  1999     6 auto~ f        15    22 r     mini~
##  9 dodge        caravan 2~   3.8  1999     6 auto~ f        15    21 r     mini~
## 10 dodge        caravan 2~   3.8  2008     6 auto~ f        16    23 r     mini~
## 11 dodge        caravan 2~   4    2008     6 auto~ f        16    23 r     mini~
```

```
p_up <- subset(mpg, class == 'pickup')
nrow(p_up)
```

```
## [1] 33
```

```
p_up
```

```
## # A tibble: 33 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 dodge        dakota pi~   3.7  2008     6 manu~ 4        15    19 r     pick~
##  2 dodge        dakota pi~   3.7  2008     6 auto~ 4        14    18 r     pick~
##  3 dodge        dakota pi~   3.9  1999     6 auto~ 4        13    17 r     pick~
##  4 dodge        dakota pi~   3.9  1999     6 manu~ 4        14    17 r     pick~
##  5 dodge        dakota pi~   4.7  2008     8 auto~ 4        14    19 r     pick~
##  6 dodge        dakota pi~   4.7  2008     8 auto~ 4        14    19 r     pick~
##  7 dodge        dakota pi~   4.7  2008     8 auto~ 4         9    12 e     pick~
##  8 dodge        dakota pi~   5.2  1999     8 manu~ 4        11    17 r     pick~
##  9 dodge        dakota pi~   5.2  1999     8 auto~ 4        11    15 r     pick~
## 10 dodge        ram 1500 ~   4.7  2008     8 manu~ 4        12    16 r     pick~
## # ... with 23 more rows
```

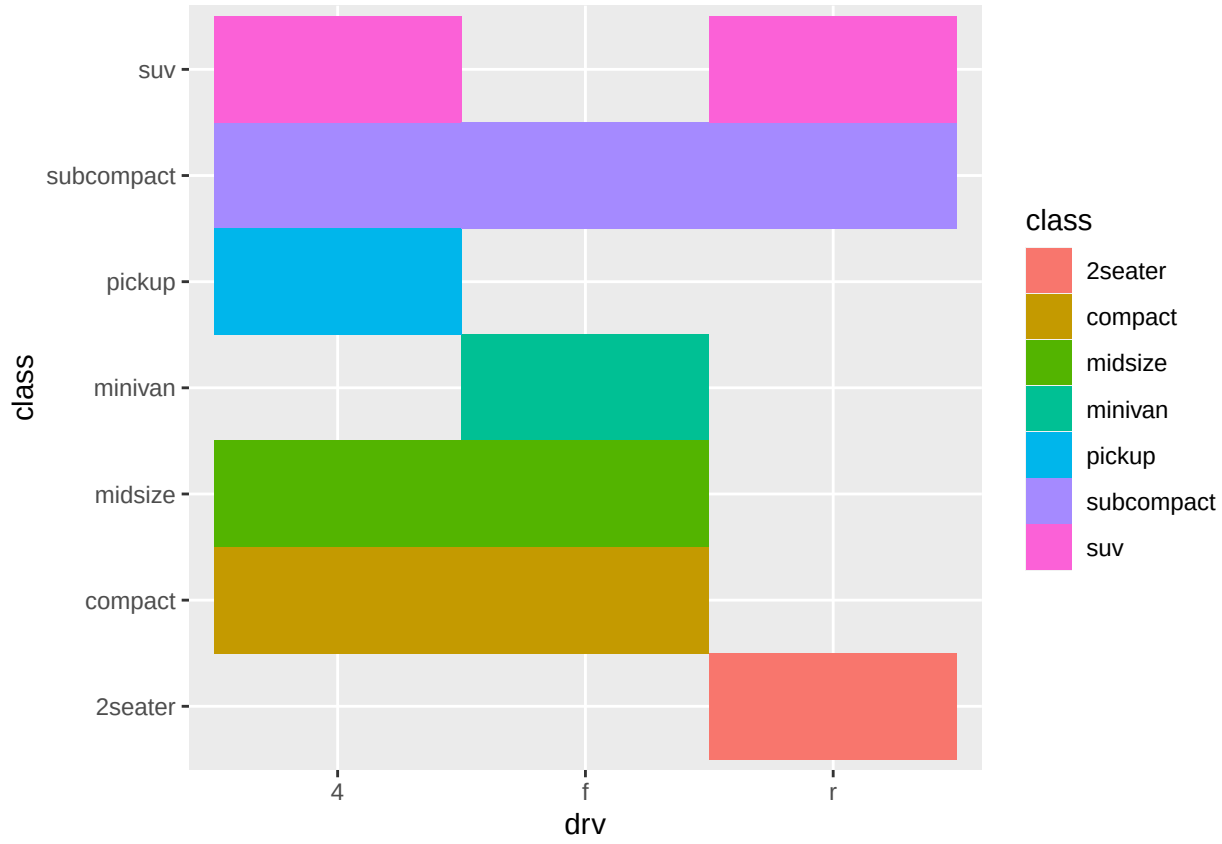
```
sub_com <- subset(mpg, class == 'subcompact')
nrow(sub_com)
```

```
## [1] 35
```

```
sub_com
```

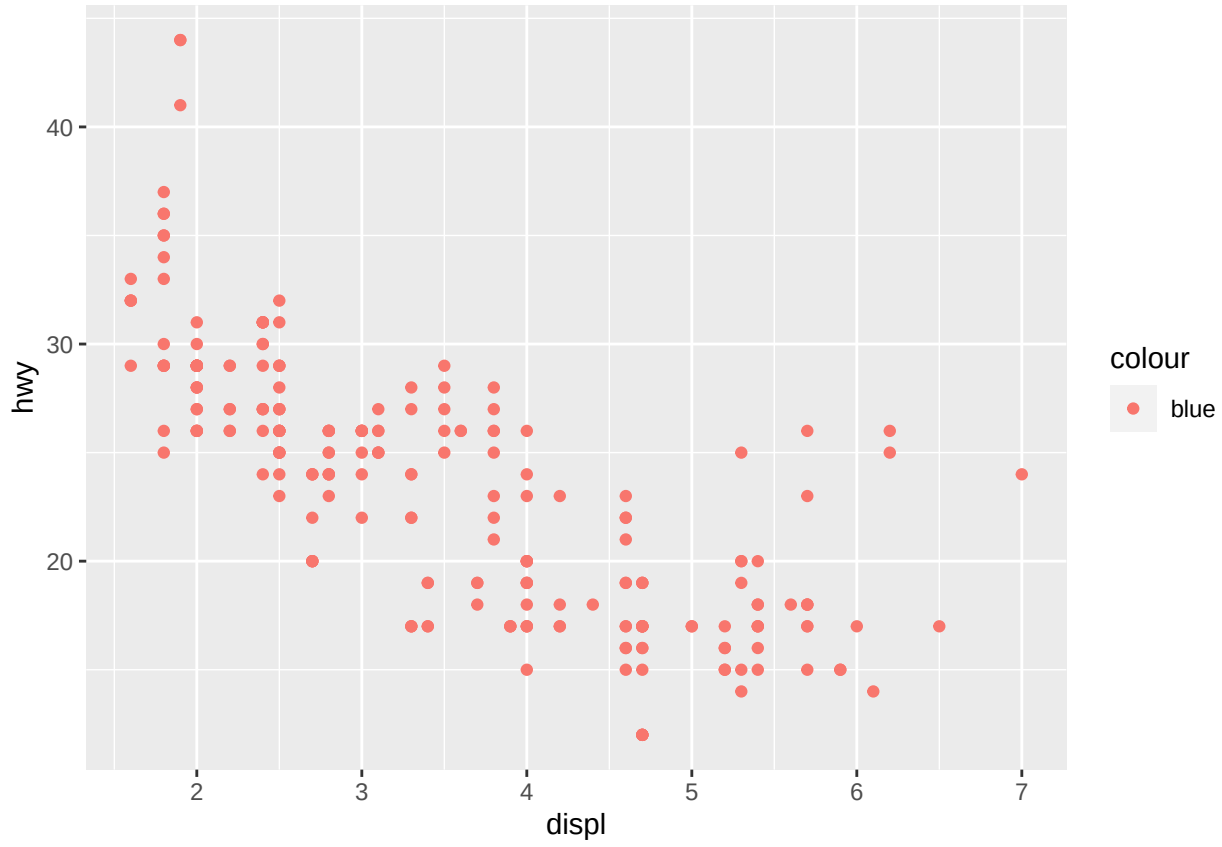
```
## # A tibble: 35 x 11
##    manufacturer model   displ  year   cyl trans    drv     cty   hwy fl    class
##    <chr>        <chr>   <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
##  1 ford         mustang   3.8  1999     6 manual(~ r        18    26 r     subc~
##  2 ford         mustang   3.8  1999     6 auto(14) r        18    25 r     subc~
##  3 ford         mustang   4    2008     6 manual(~ r        17    26 r     subc~
##  4 ford         mustang   4    2008     6 auto(15) r        16    24 r     subc~
##  5 ford         mustang   4.6  1999     8 auto(14) r        15    21 r     subc~
##  6 ford         mustang   4.6  1999     8 manual(~ r        15    22 r     subc~
##  7 ford         mustang   4.6  2008     8 manual(~ r        15    23 r     subc~
##  8 ford         mustang   4.6  2008     8 auto(15) r        15    22 r     subc~
##  9 ford         mustang   5.4  2008     8 manual(~ r        14    20 p     subc~
## 10 honda        civic     1.6  1999     4 manual(~ f        28    33 r     subc~
## # ... with 25 more rows
```

```
#a. Show the codes and its result for the narrative in #6.
ggplot(mpg, aes(drv, class)) +
  geom_tile (aes(fill = class))
```

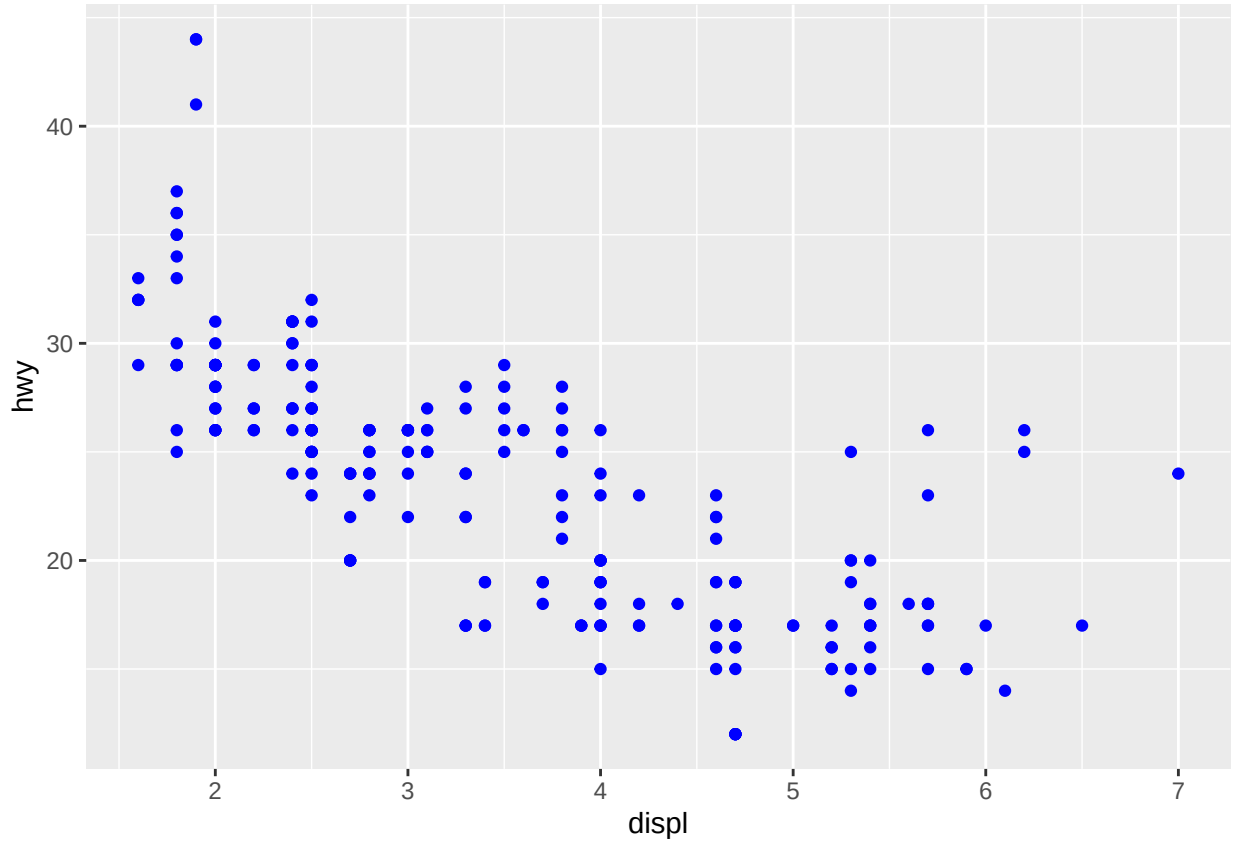


```r
#b. Interpret the result.
# the plot shows the relationship between class and drv.
```

```r
#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"))
```




```r
#Code #2
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), colour = "blue")
```




```
# the colour in the first code is inside the parentheses which indicates the legend.
#While the second one is outside the parentheses that contains the x and y which
#indicates the color of the geom_point.
```

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

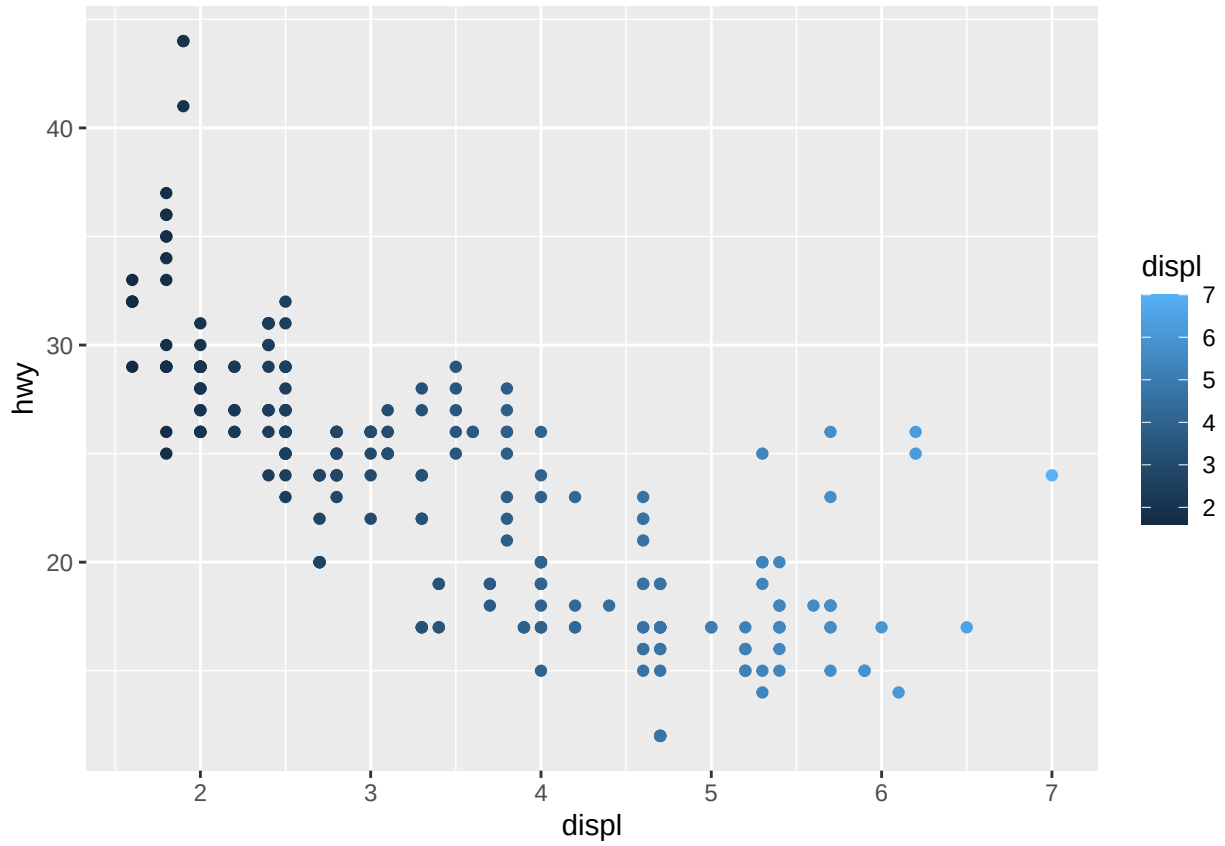
```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           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
```

```
# it shows the data in mpg.
```

```
#a. Which variables from mpg data set are categorical?
#Categorical variables in mpg include: manufacturer, model, trans (type of transmission),
#drv (front-wheel drive, rear-wheel, 4wd), fi (fuel type), and class (type of car).
```

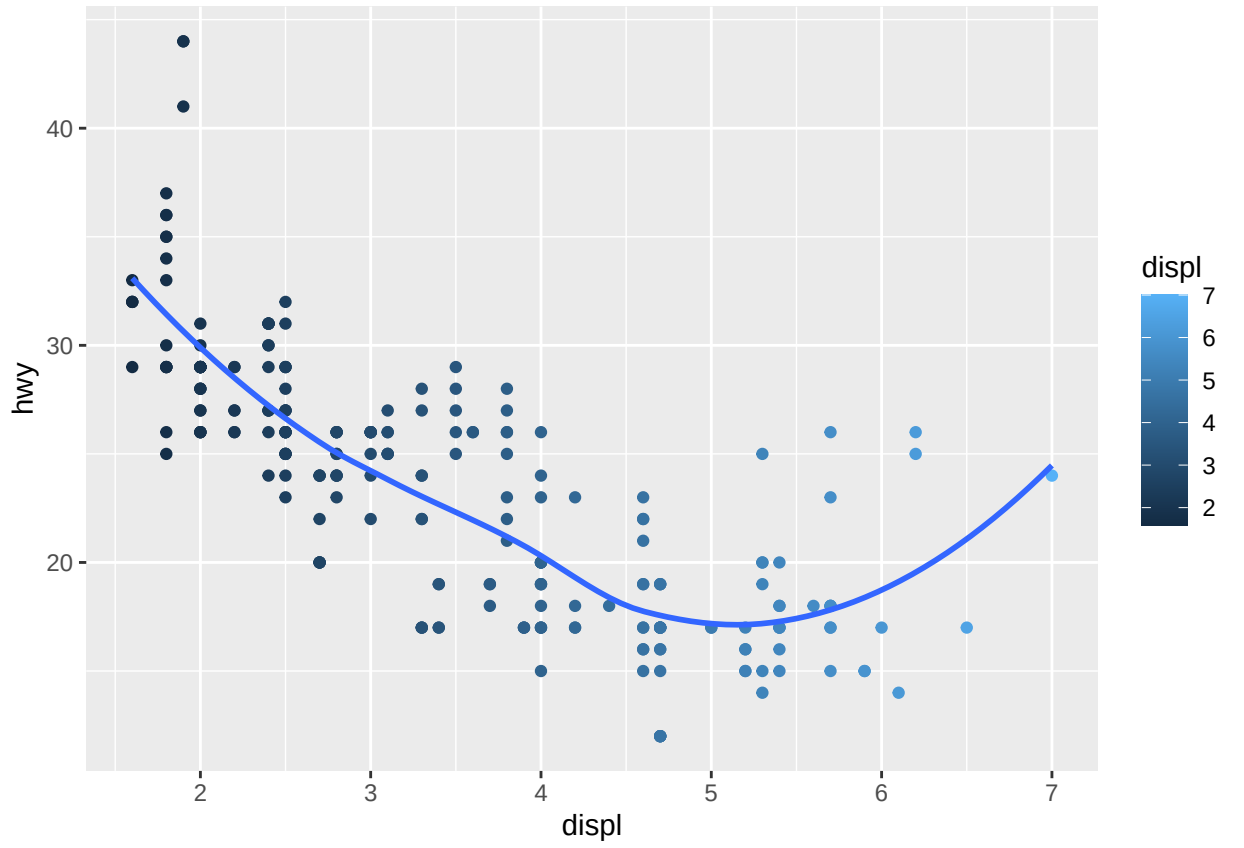
```
#b. Which are continuous variables?
#Continuous variables in mpg include: displ (engine displacement in litres),
#cyl (number of cylinders), cty (city miles/gallon), and hwy (highway gallons/mile)
```

```
#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. What
#result? Why it produced such output?
ggplot( data = mpg) +
  geom_point(mapping = aes(x = displ , y = hwy, col = displ))
```



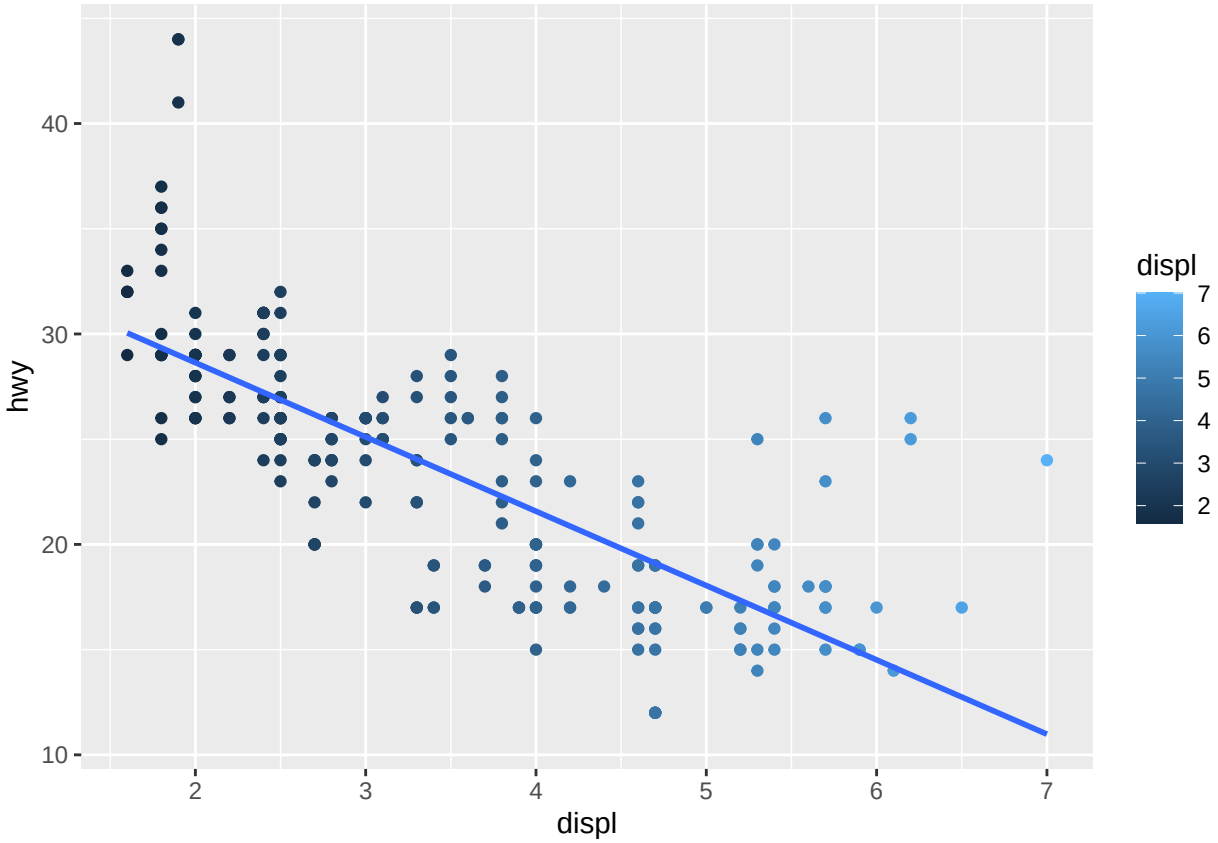

```
#9. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon) using g
#  geom_smooth() with se = FALSE. Default method is "loess".
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping=aes(color=displ)) +
  geom_smooth(se =FALSE)
```

```
## `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)) +
  geom_point(mapping=aes(color=displ)) +
  geom_smooth(se =FALSE,method = lm)
```

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


hwy
40
30
20
10
2
3
4
5
6
7
displ
displ
7
6
5
4
3
2