

RWorksheet_Pastor#6

Vincent Pastor

2022-11-25

1. How many columns are in mpg dataset? How about the number of rows? Show the codes and its result.
- It has 11 columns and 234 rows.

```
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)

data(mpg)
data2 <- as.data.frame(mpg)
nrow(mpg)
```

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

```
ncol(mpg)
```

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

2. Which manufacturer has the most models in this data set? Which model has the most variations?
- The manufacturer that has most model are the Dodge while the most variations of model are the caravan2wd.

```
mostmod <- data2 %>% group_by(manufacturer, model) %>%
  count()
mostmod
```

```
## # 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  5
## 5 chevrolet    corvette         5
## 6 chevrolet    k1500 tahoe 4wd   4
## 7 chevrolet    malibu           5
```

```
## 8 dodge      caravan 2wd      11
## 9 dodge      dakota pickup 4wd  9
## 10 dodge     durango 4wd      7
## # ... with 28 more rows
```

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

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

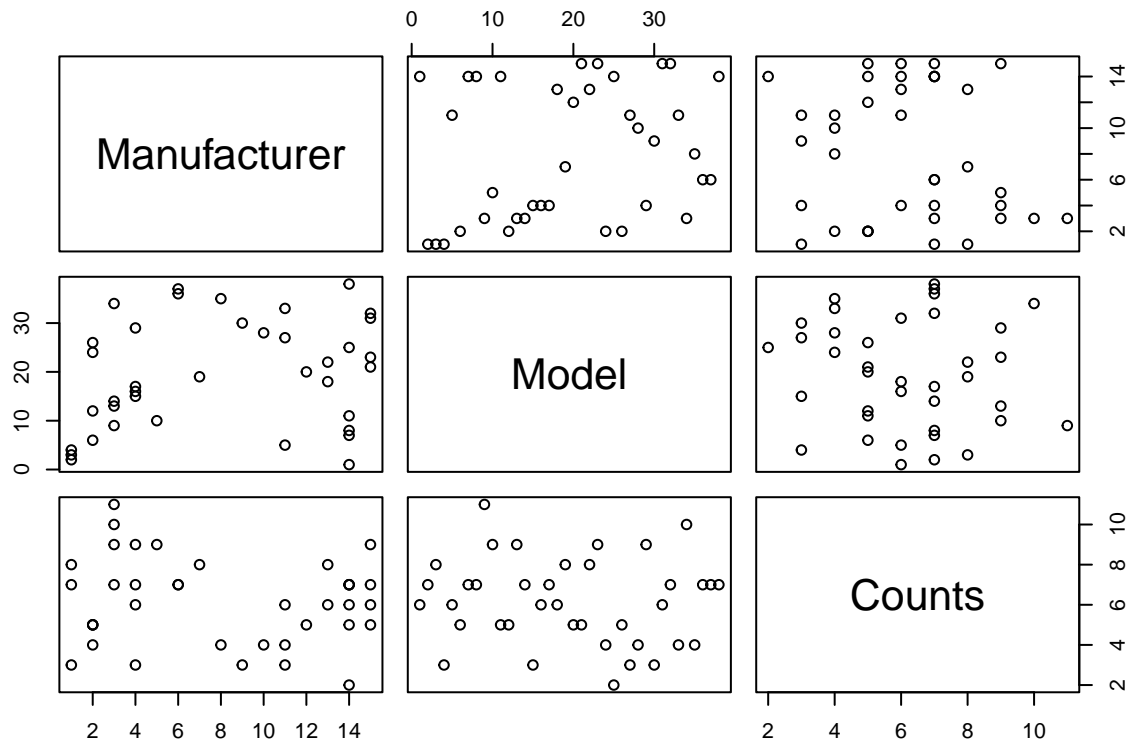
```
uniquemods <- data2 %>% group_by(manufacturer, model) %>%
  distinct() %>% count()
uniquemods
```

```
## # 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
```

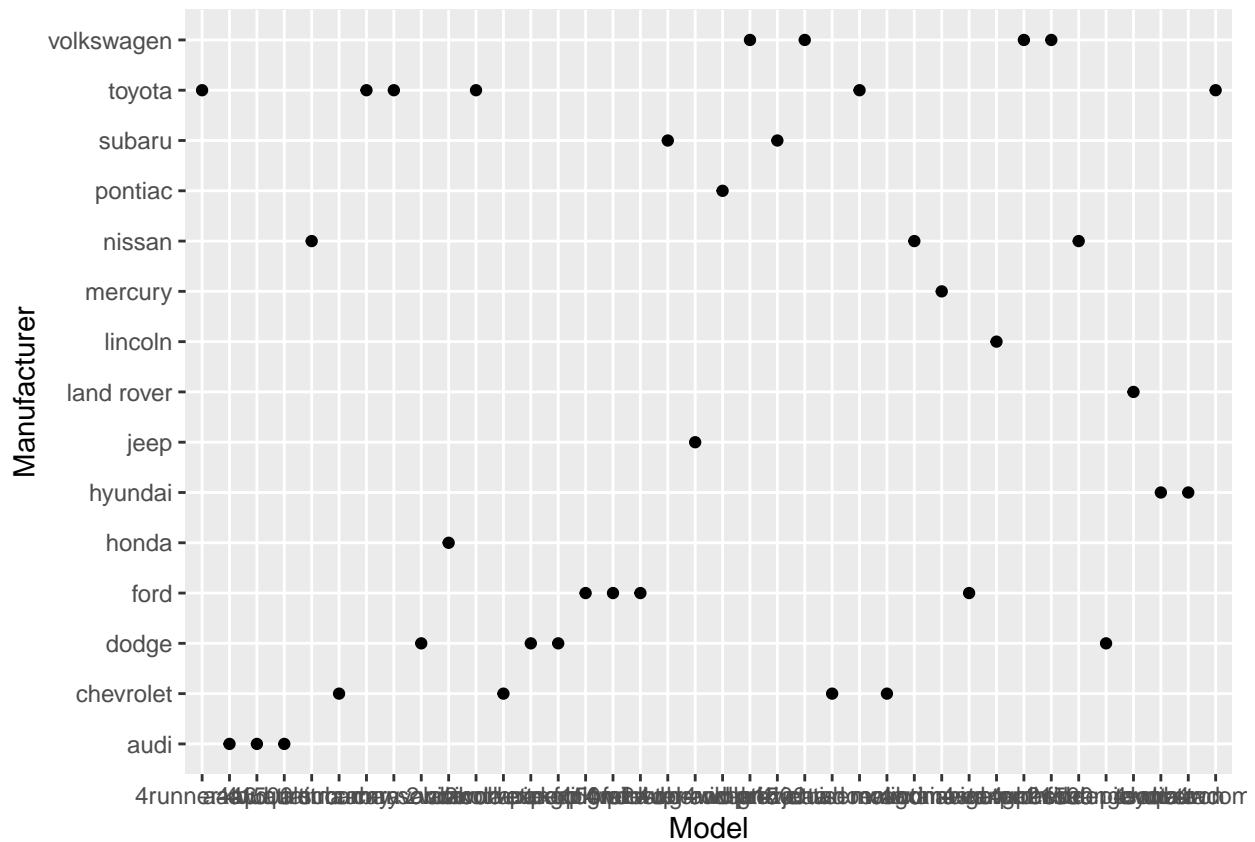
```
colnames(uniquemods) <- c("Manufacturer", "Model", "Counts")
```

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

```
plot(mostmod)
```



```
ggplot(mostmod, aes(Model, Manufacturer)) + geom_point()
```



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

```
mpgdata <- mpg
mod_fact <- data2 %>% group_by(manufacturer, model) %>%
  distinct() %>% count()
mod_fact
```

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

```
## # 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
```

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

`ggplot(mpg, aes(model, manufacturer)) + geom_point()` ### - It shows the 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?

-

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

```
mpgdat <- data2 %>% group_by(model) %>% count()
mpgdat
```

```
## # 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

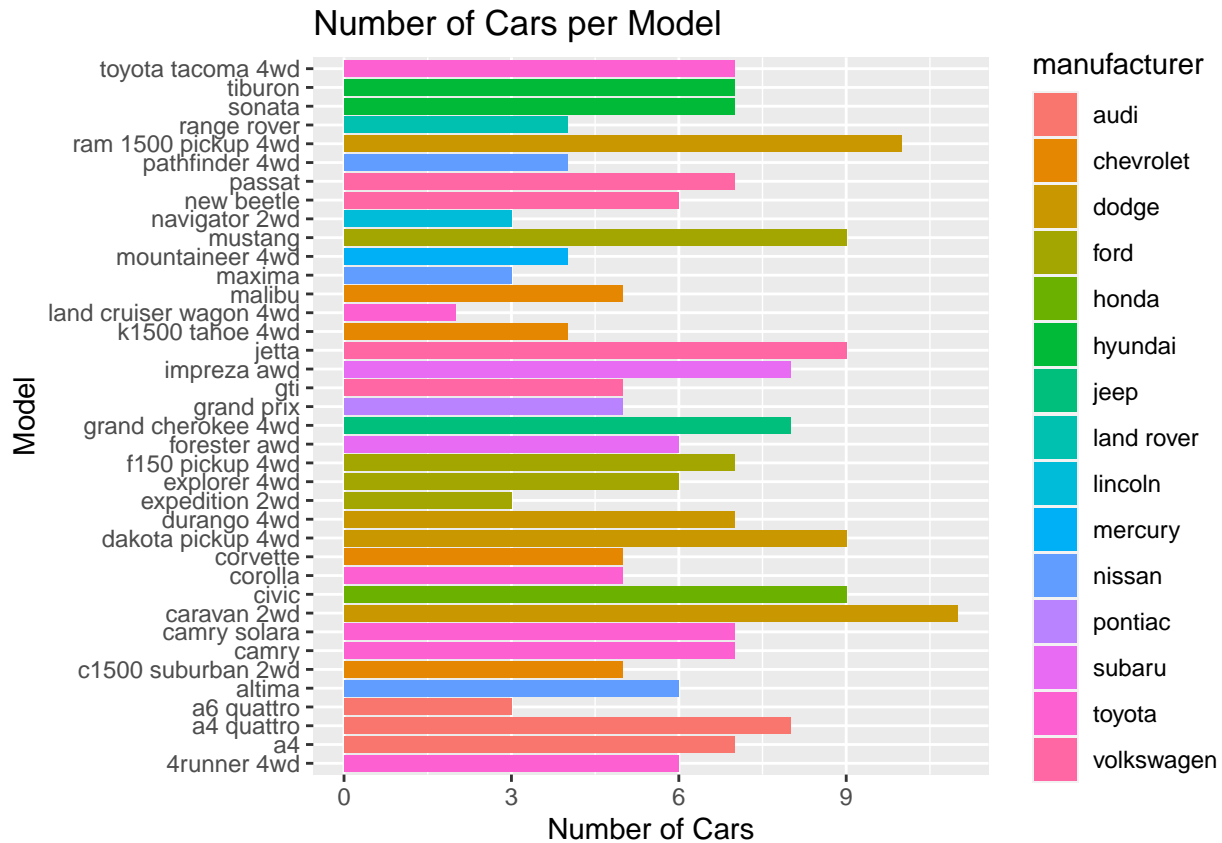
colnames(mpgdat) <- c("Model", "Counts")
mpgdat
```

```
## # A tibble: 38 x 2
## # Groups:   Model [38]
##   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 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.

```
mpg2 <- mpgdat[1:20,] %>% top_n(2)
```

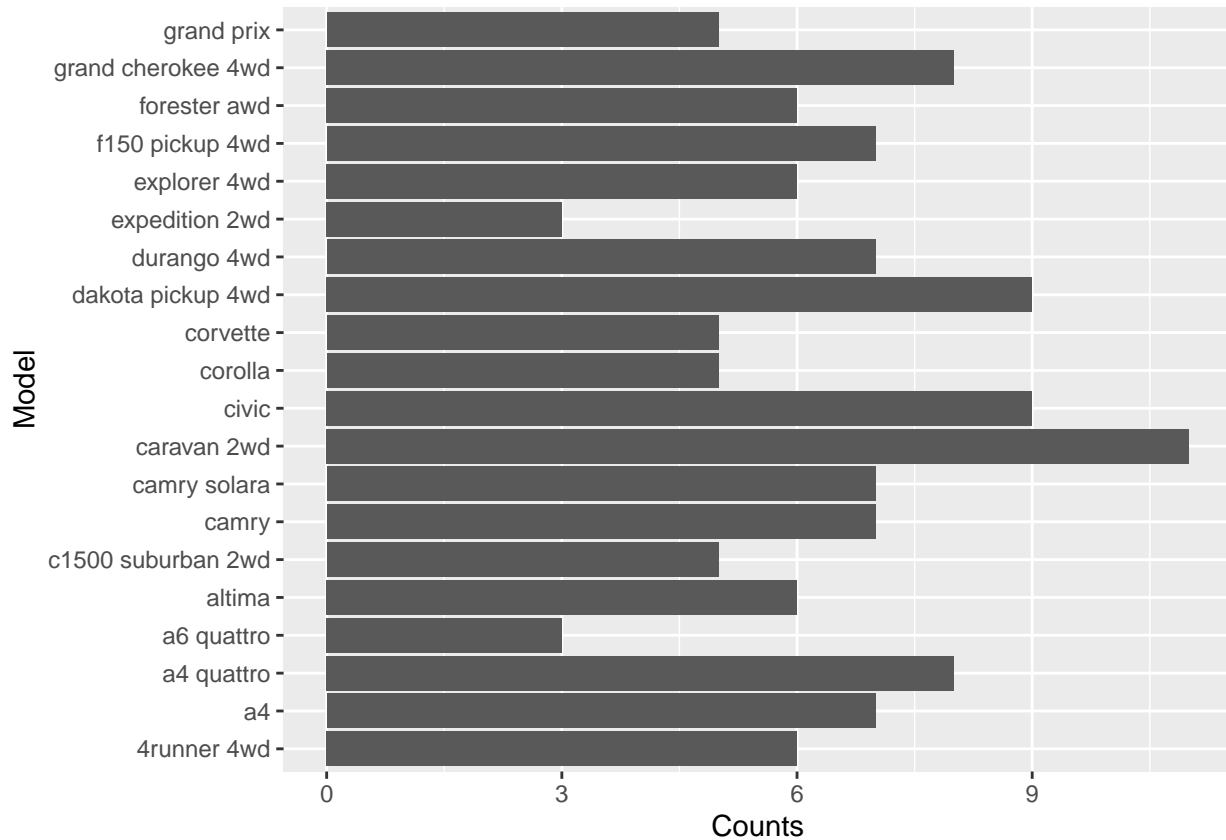
Selecting by Counts

```
mpg2
```

```
## # 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
```

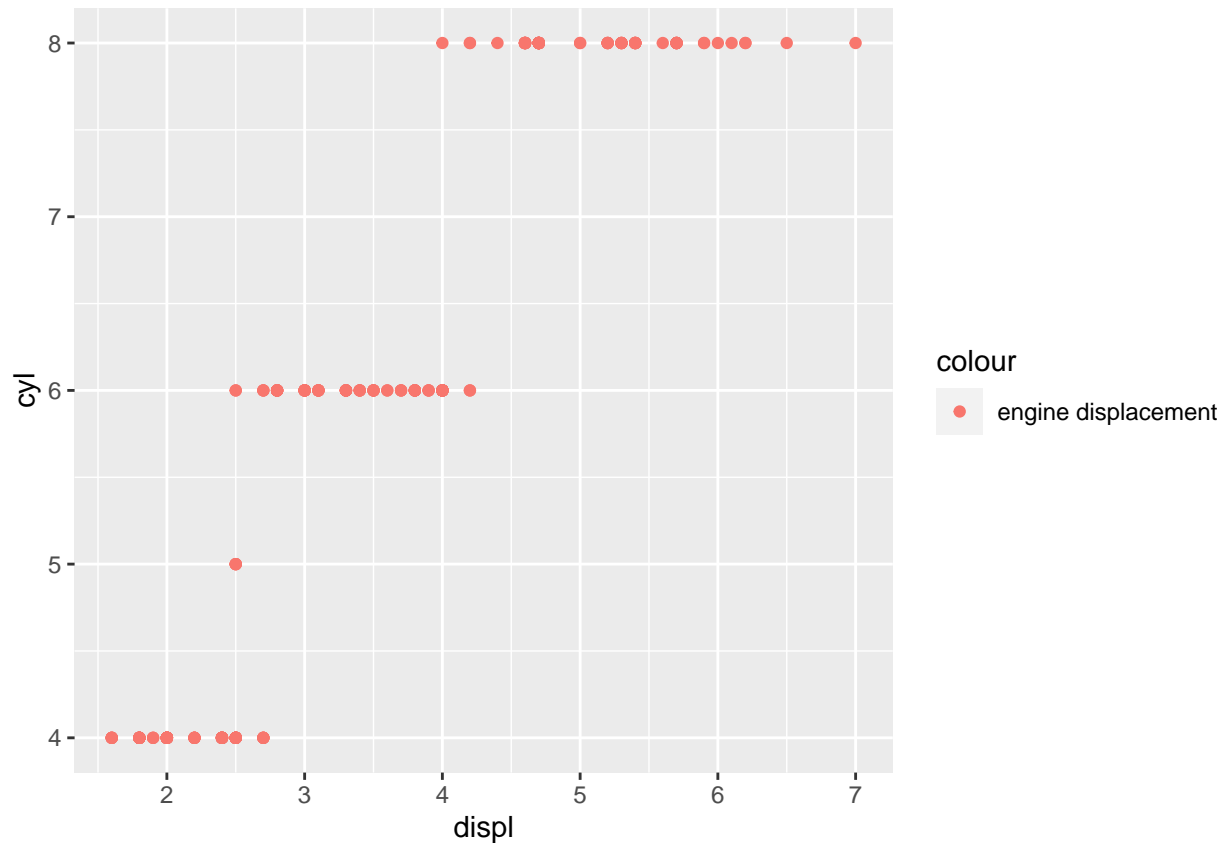
```
ggplot(mpg2, aes(x = Model, y = Counts)) +
  geom_bar(stat = "identity") + coord_flip()
```



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"))
```



b. How would you describe its relationship?

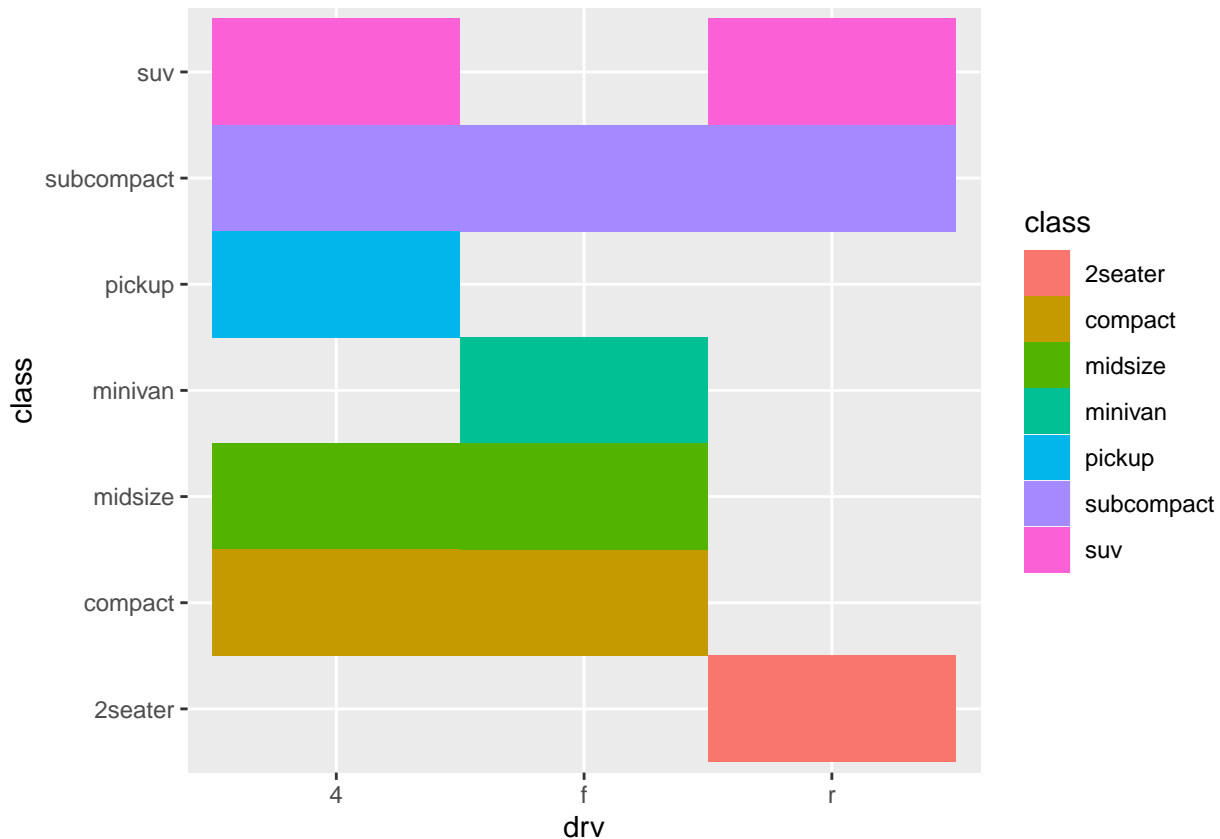
- The scatter plot shows that the “cyl” is a number of cylinders in y and the “displ” is an engine displacement x.

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_tile(aes(fill=class))
```

b. Interpret the result.

- By adding color to the geom_tile you can see the different type of drive train and the class.

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

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

```
?mpg
```

- The result are directed you to help option showing the data mpg.

a. Which variables from mpg dataset are categorical?

- The manufacturer trans ,model, drv, fl, class

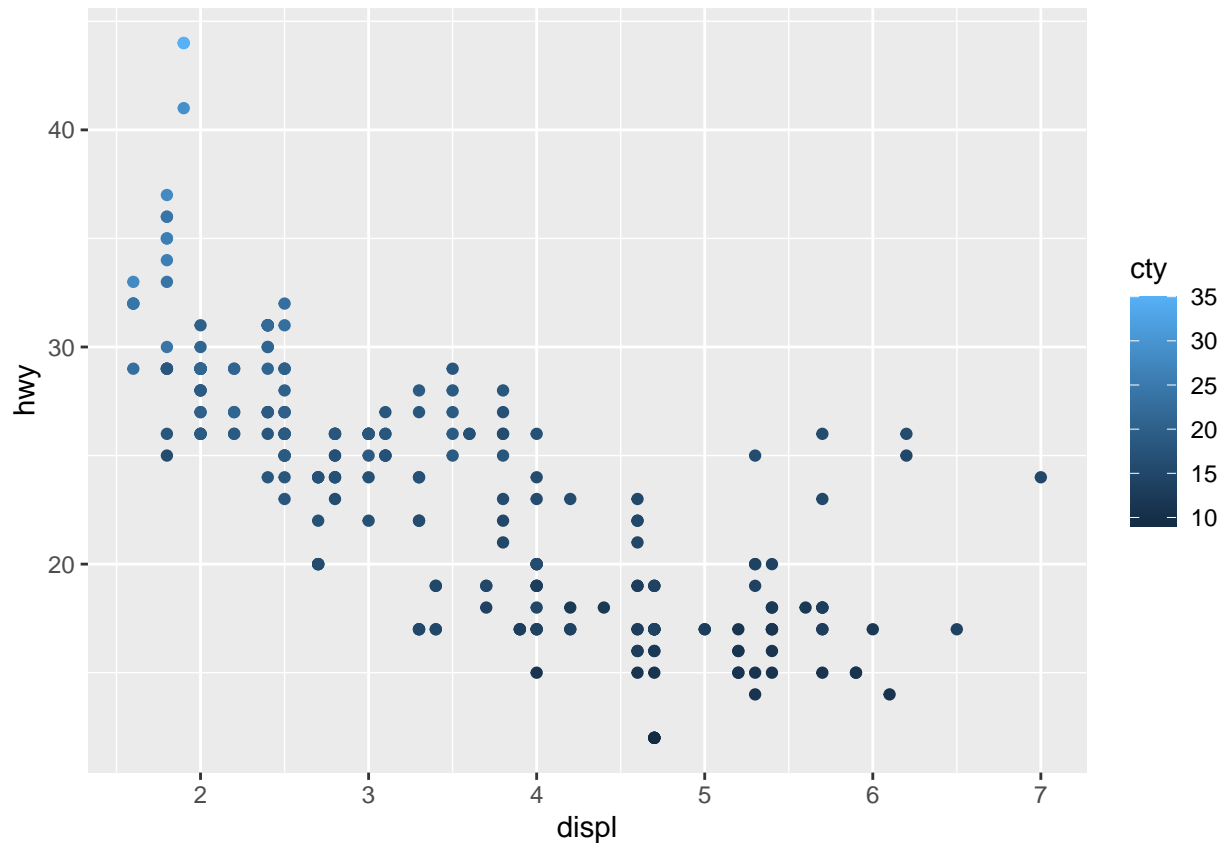
are the categorical variables from the data-set of mpg.

b. Which are continuous variables?

- The continuous variable of the mpg data-set are the dsipl, cyl etc.

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 is its result? Why it produced such output?

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

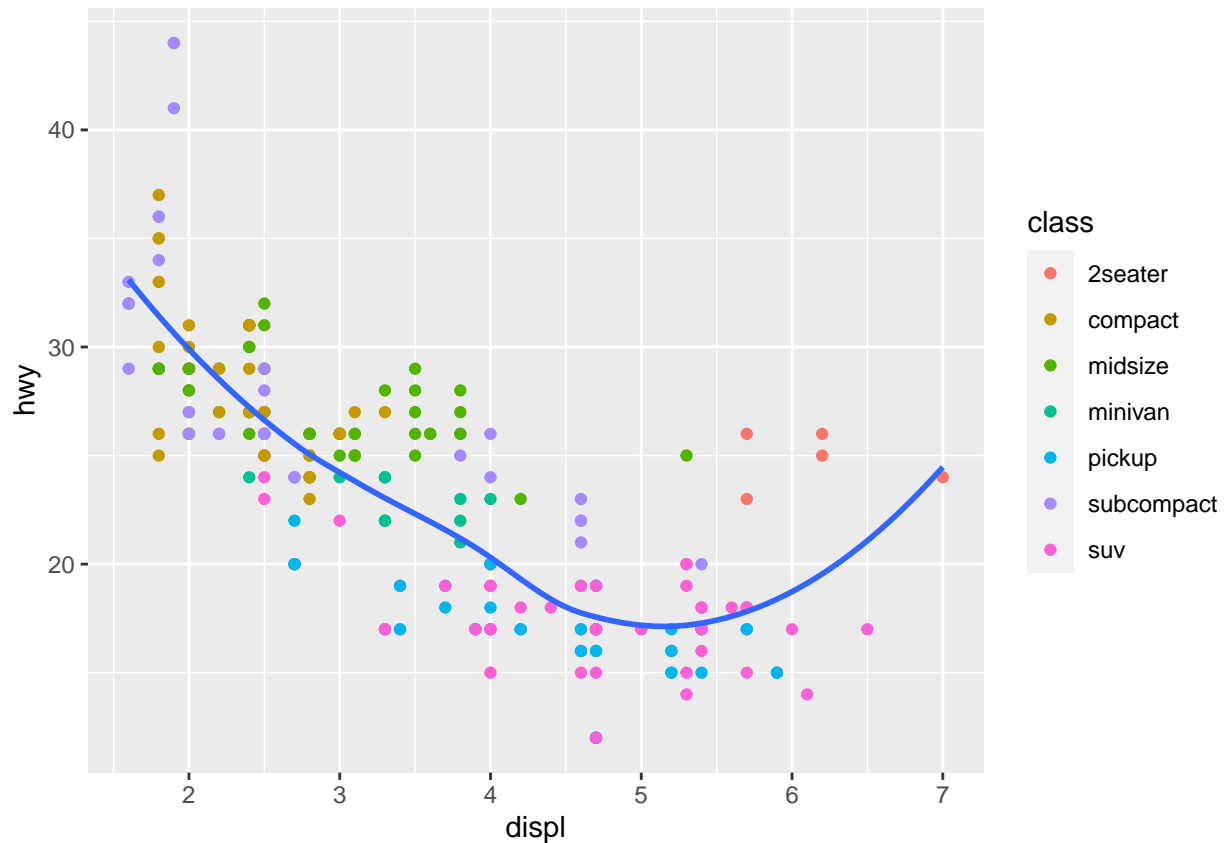


- The data shows that they are in the positive rate using the displ for hwy and cty scattered plot.

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 these = 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, method = lm)
```

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

