Data Visualisation

Yifan Jin

05/02/2020

Introduction

Prerequisites

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -----------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

We can employ function by using `package::function()` eg. `ggplot2::ggplot()` or directly using function eg. `ggplot()`

First steps

Think about the relationship between variables, think about the mean of one group is smaller or greater than other, think about one variable is positively or negatively correlated with other.

Example: The data frame `mpg`

```
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(l~ f        18    29 p     comp~
## 2 audi          a4         1.8  1999     4 manual~ f        21    29 p     comp~
## 3 audi          a4         2    2008     4 manual~ f        20    31 p     comp~
## 4 audi          a4         2    2008     4 auto(a~ f        21    30 p     comp~
## 5 audi          a4         2.8  1999     6 auto(l~ f        16    26 p     comp~
## 6 audi          a4         2.8  1999     6 manual~ f        18    26 p     comp~
## 7 audi          a4         3.1  2008     6 auto(a~ f        18    27 p     comp~
## 8 audi          a4 quat~   1.8  1999     4 manual~ 4        18    26 p     comp~
## 9 audi          a4 quat~   1.8  1999     4 auto(l~ 4        16    25 p     comp~
```

```
## 10 audi          a4 quat~    2    2008    4 manual~ 4        20    28 p      comp~
## # ... with 224 more rows
```

Variables descriptions:

1. `displ`, a car's engine size, in litres

2. `hwy`, a car's fuel efficiency on the highway, in miles per gallon (mpg). A car with low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance

To learn more information about `mpg`, use `?mpg` to get more.

Creating a ggplot

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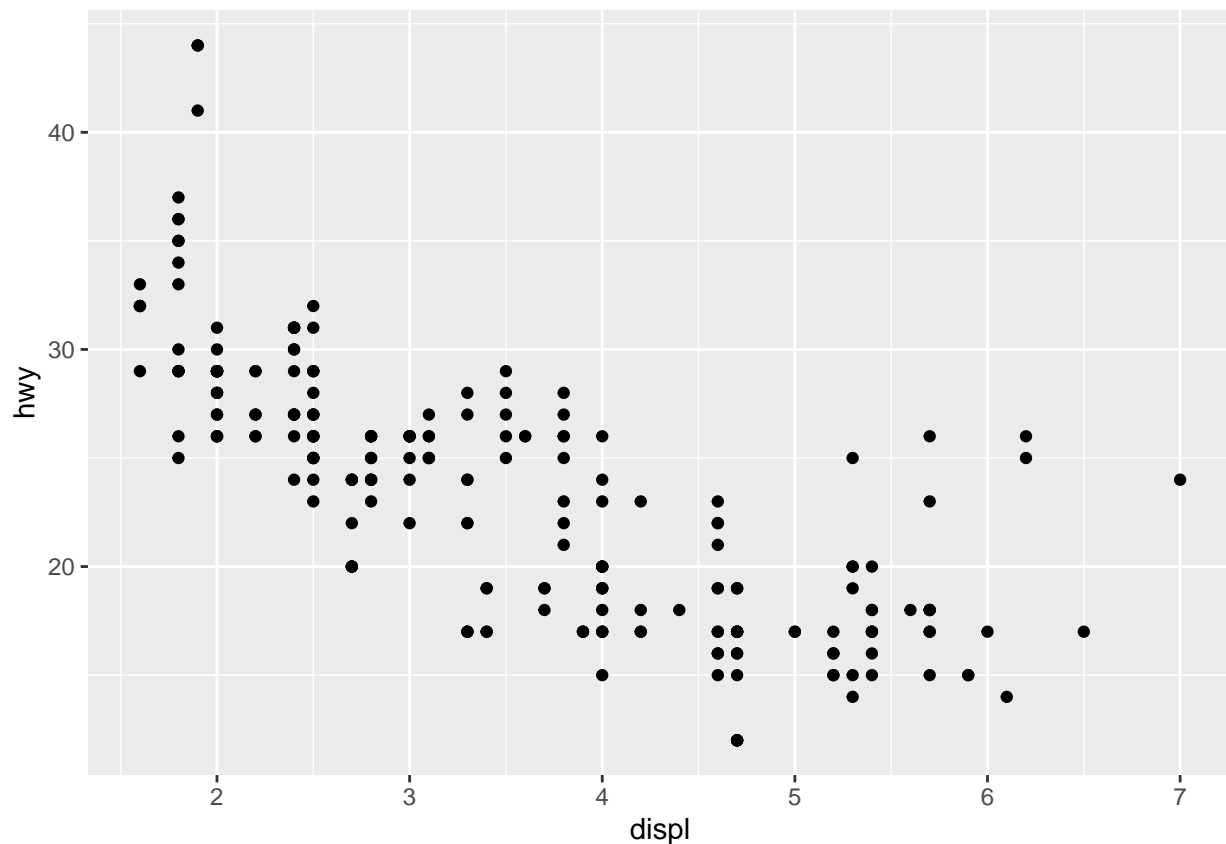

We can see that there is a negative relationship between engine size (`displ`) and fuel efficiency (`hwy`). We can form a hypothesis test to formally test the relationship.
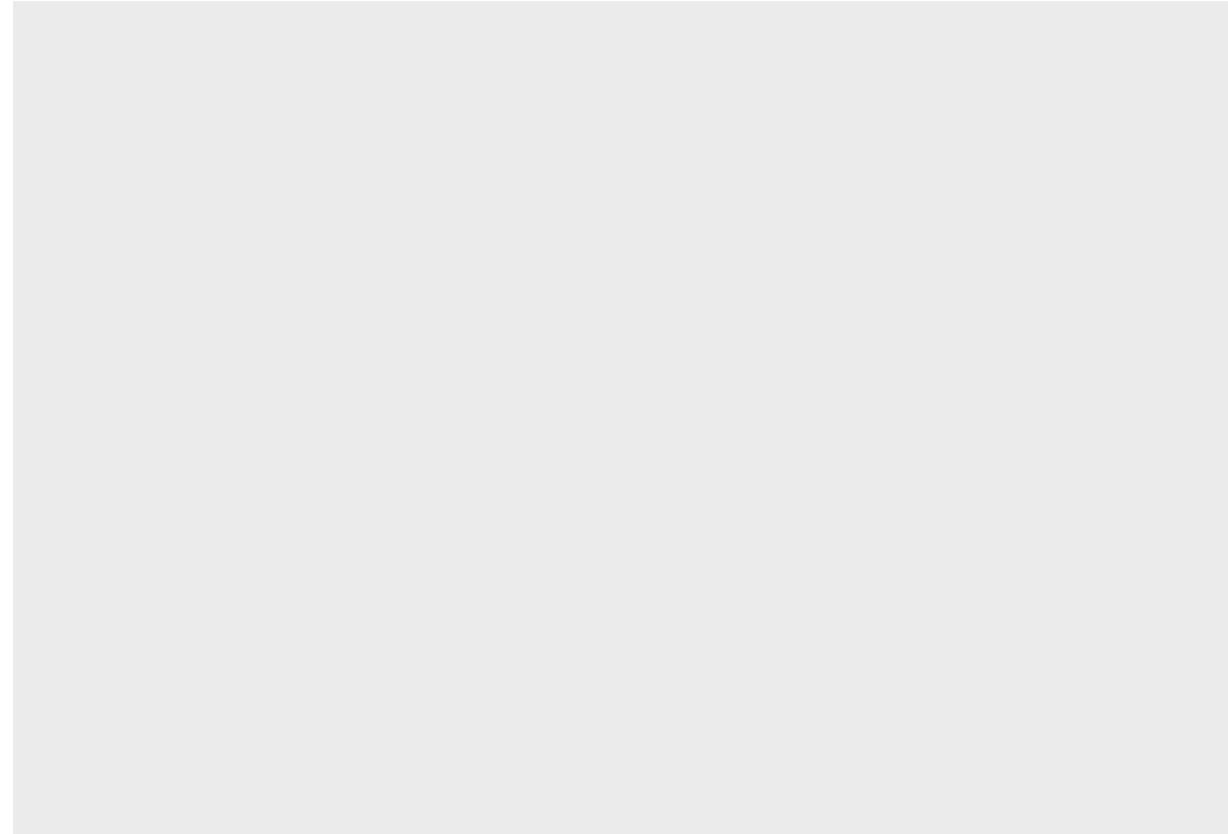
A graphing template

```
# ggplot(data = <DATA>) +
#   <GEOM_FUNCTION>(mapping = #    aes(<MAPPINGS>))
```

Exercise

1. Run ggplot(data=mpg). What do you see?

```
ggplot(data=mpg)
```


Answer: We get nothing

2. How many rows are in mpg? How many columns

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

Answer:

We have 234 rows of observations and 11 variables

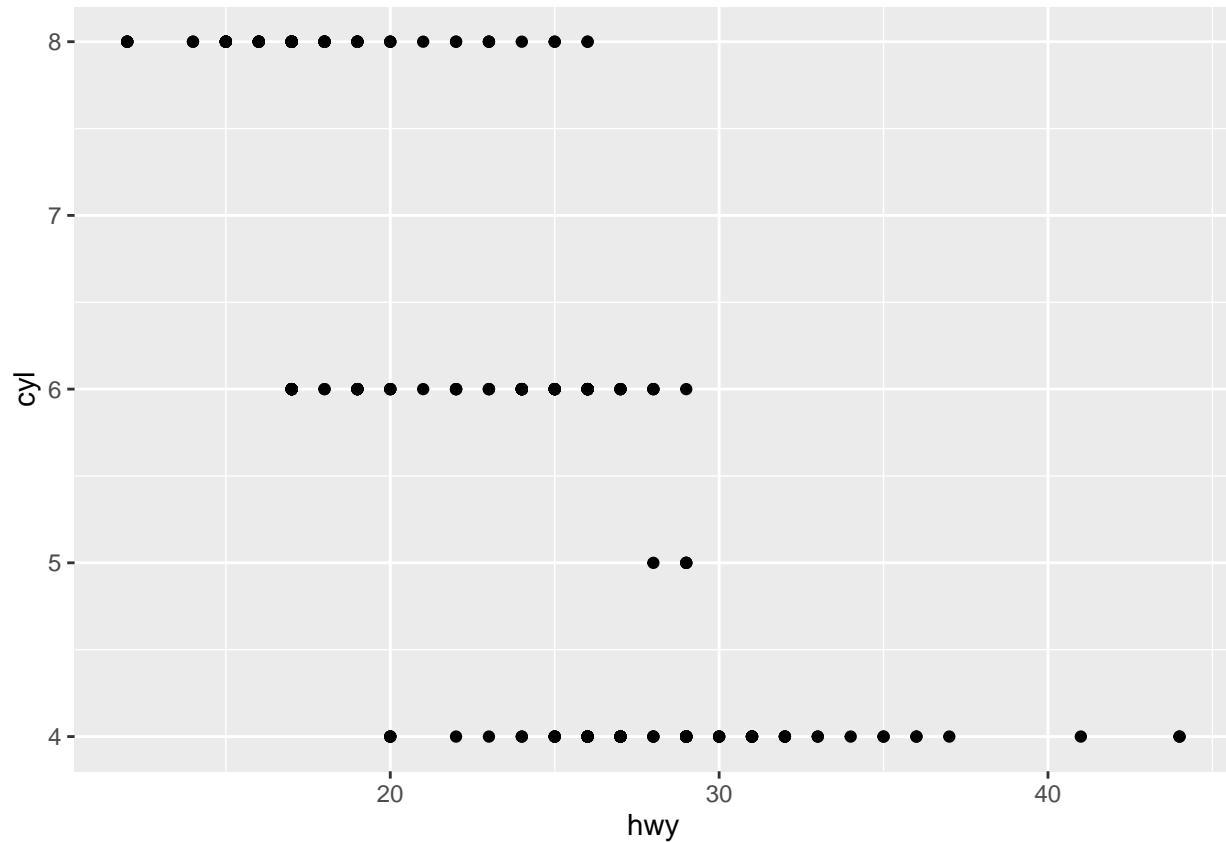**3. What does the drv variable describe? Read the help for out**

3

```
?mpg
```

drv: f = front-wheel drive, r = rear wheel drive, 4 = 4wd

4. Make a scatterplot of `hwy` vs `cyl`

Answer:

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


**5. What happens if you make scatterplot of `class` vs `drv`? Why is not useful?**

Answer:

It can be seen that `cyl` only take certain values which is discrete variable. Therefore, it looks strange. We can not see any information about the realtionship between two variables, hence, it doesn't make much sense.

Aesthetic mappings

The greatest value of a picture is when it forces us to notice what we never expected to see
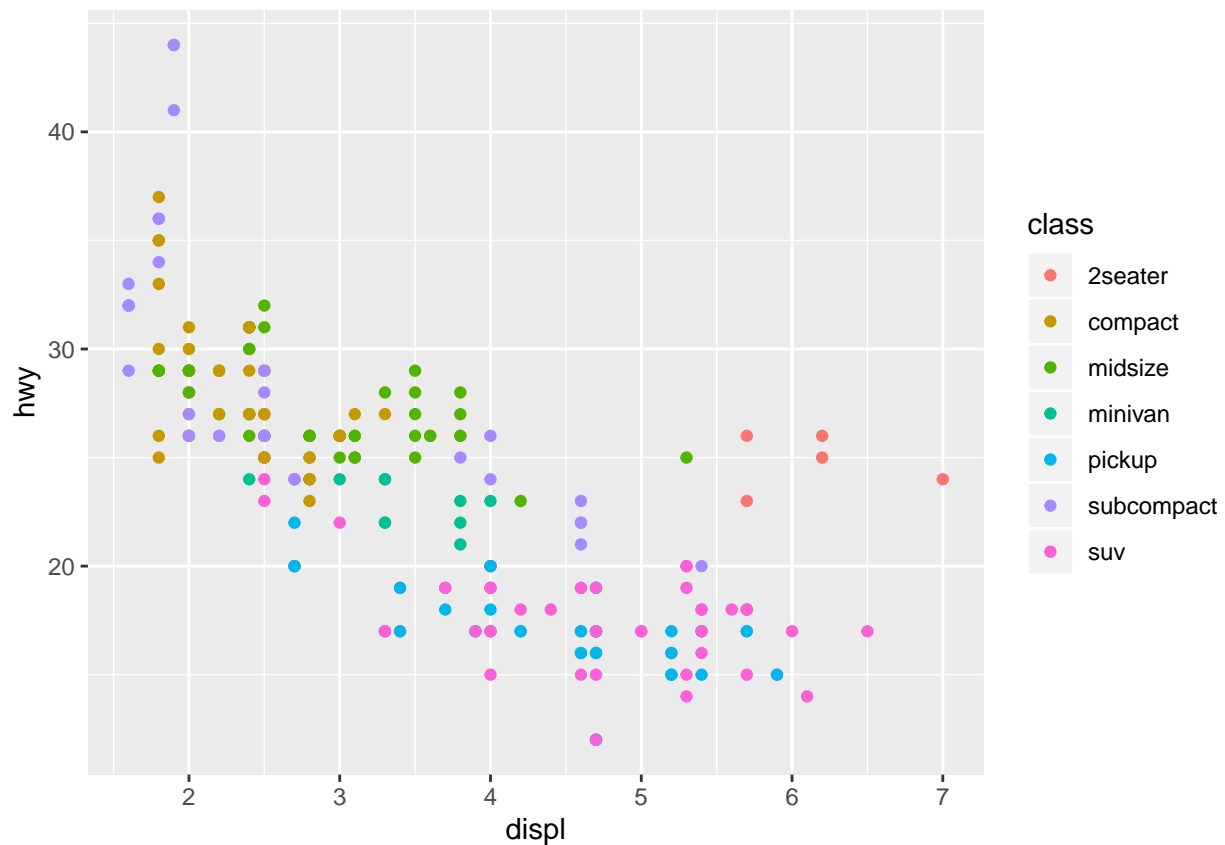
```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy))
```
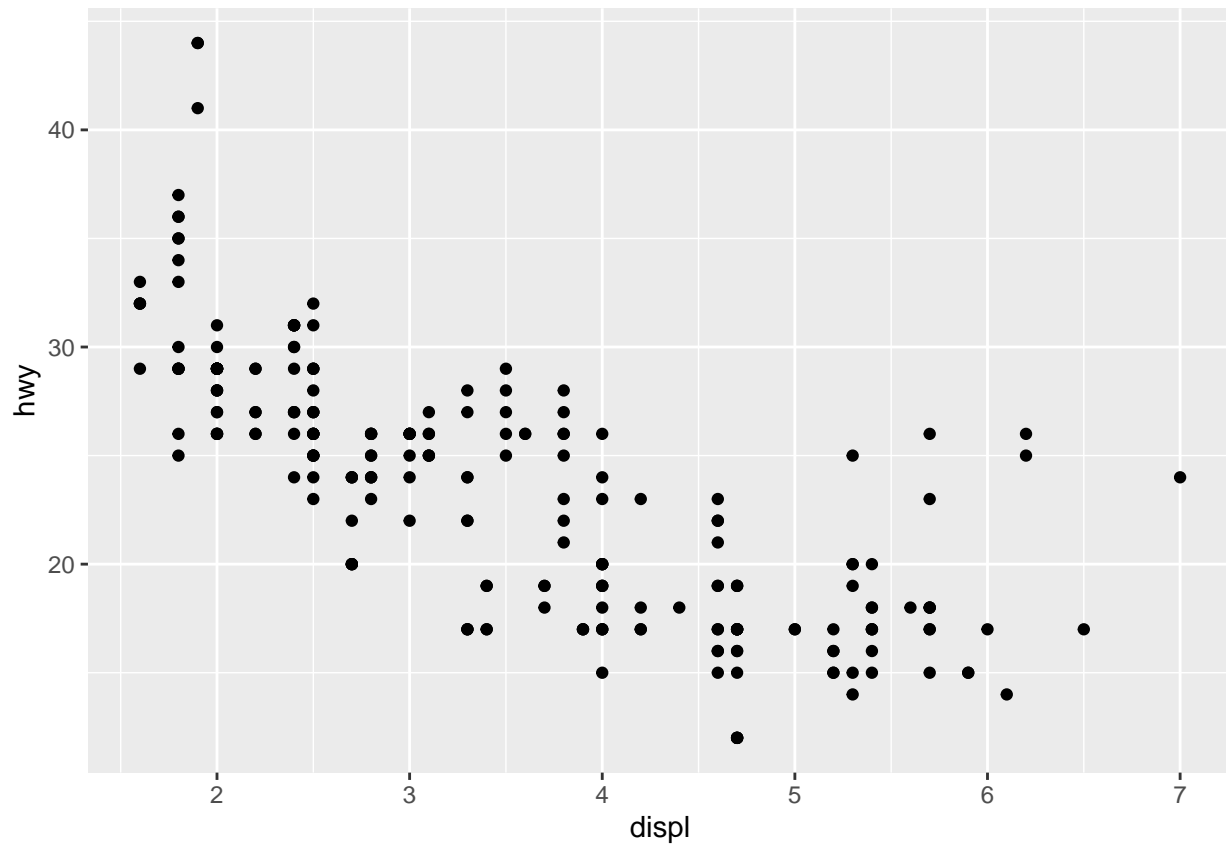
There are some points on the right which is outside the general trend. There may be other factors that affect hwy.

We can add `class` variable

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy,color=class))
```
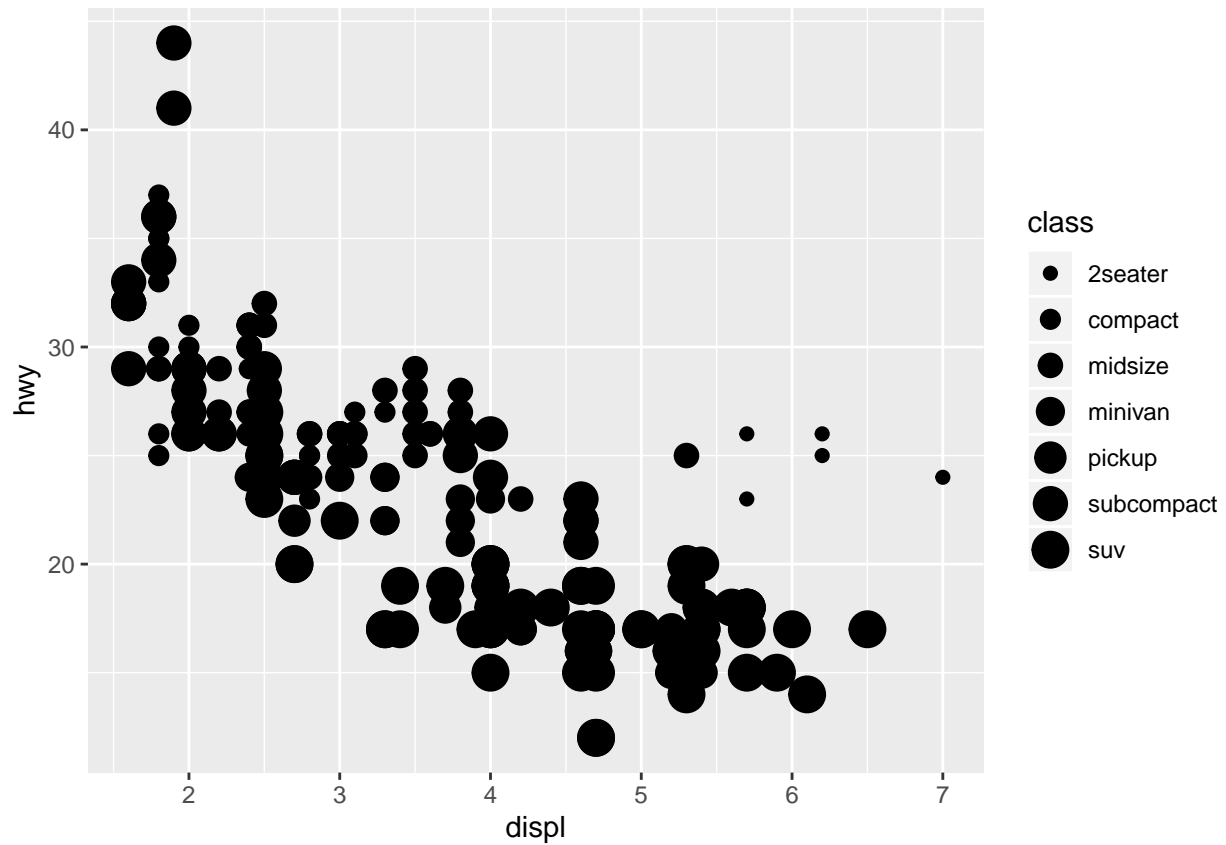
The colors reveal that many of the unusual points are two-seater cars.

Not only the color can be used for categorical variable, but also the size.

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy,size=class))
```
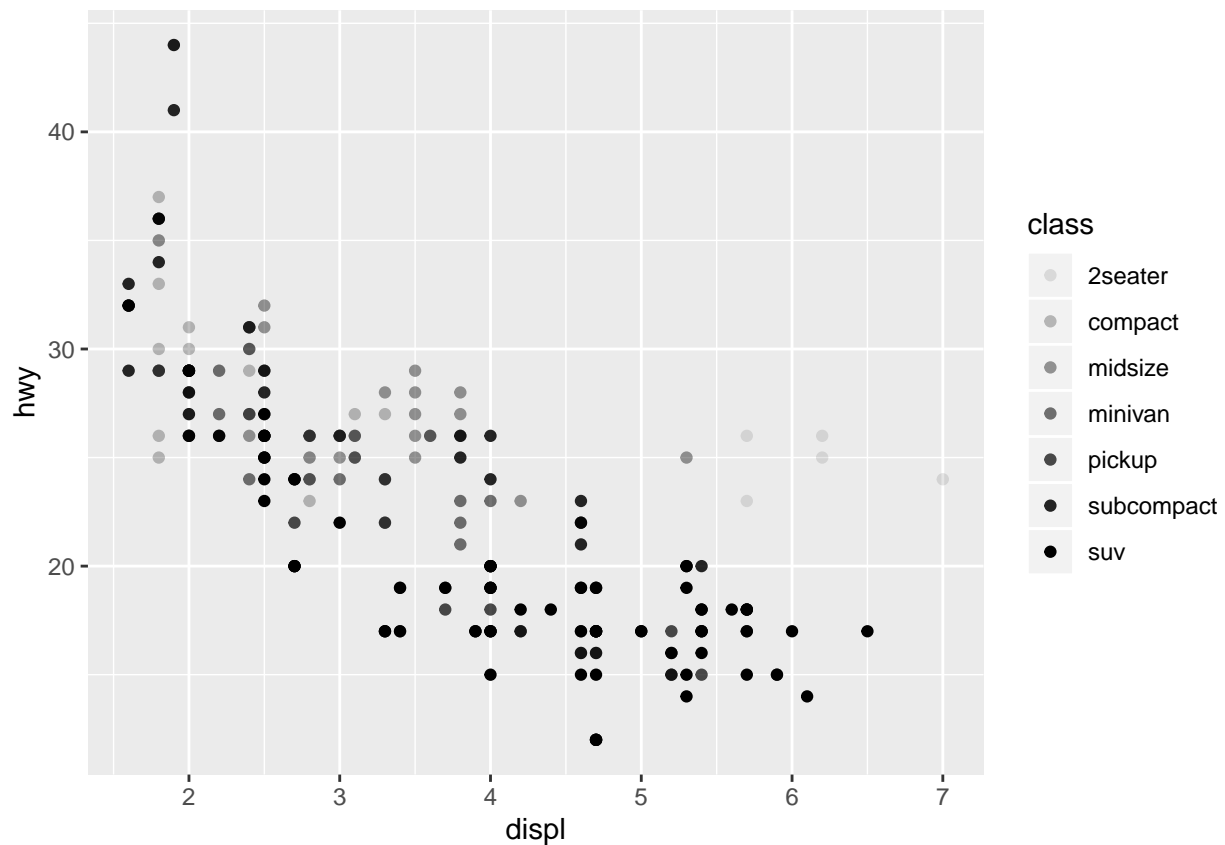
```
## Warning: Using size for a discrete variable is not advised.
```

We could also mapped `class` to the alpha aesthetic, which controls the transparency of the points, or to the shape aesthetic, which controls the shape of the points.

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy,alpha=class))
```

```
## Warning: Using alpha for a discrete variable is not advised.
```

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy,shape=class))
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 7. Consider
## specifying shapes manually if you must have them.
```
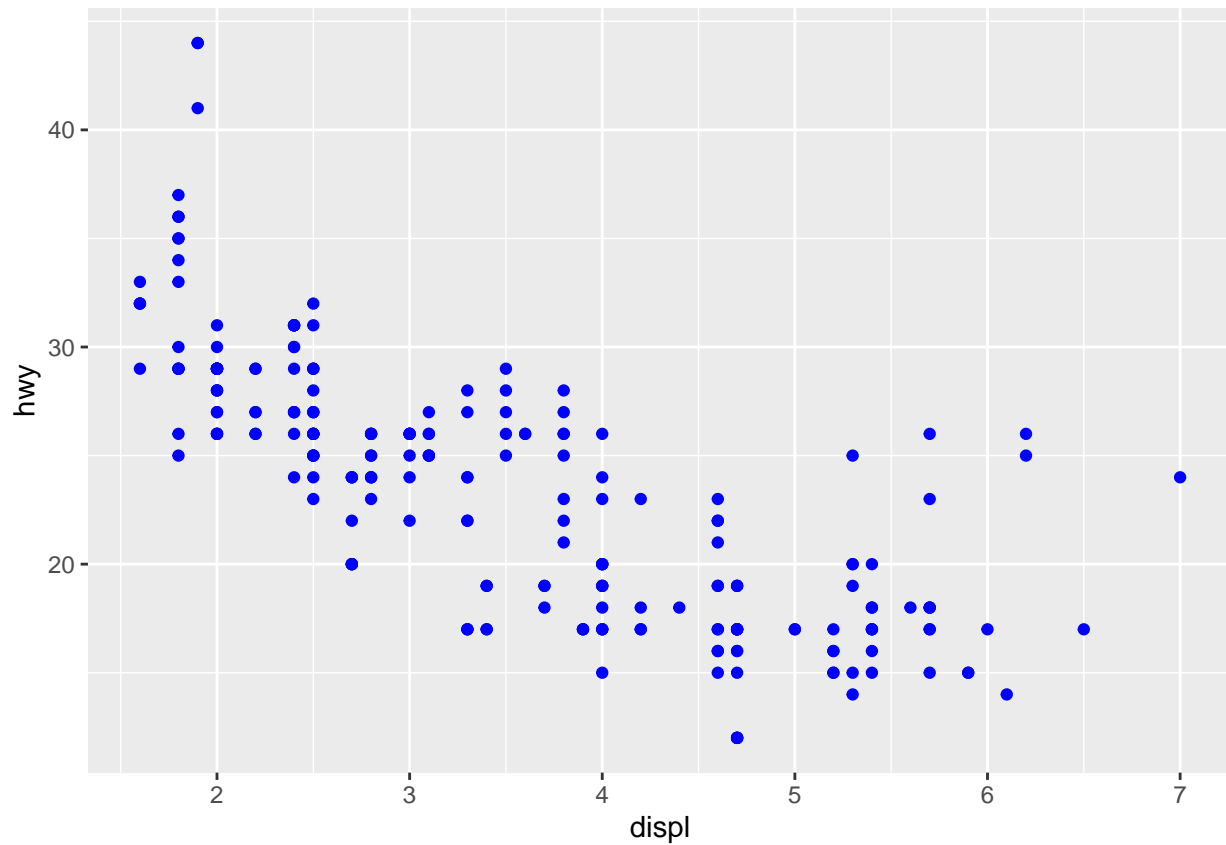
```
## Warning: Removed 62 rows containing missing values (geom_point).
```
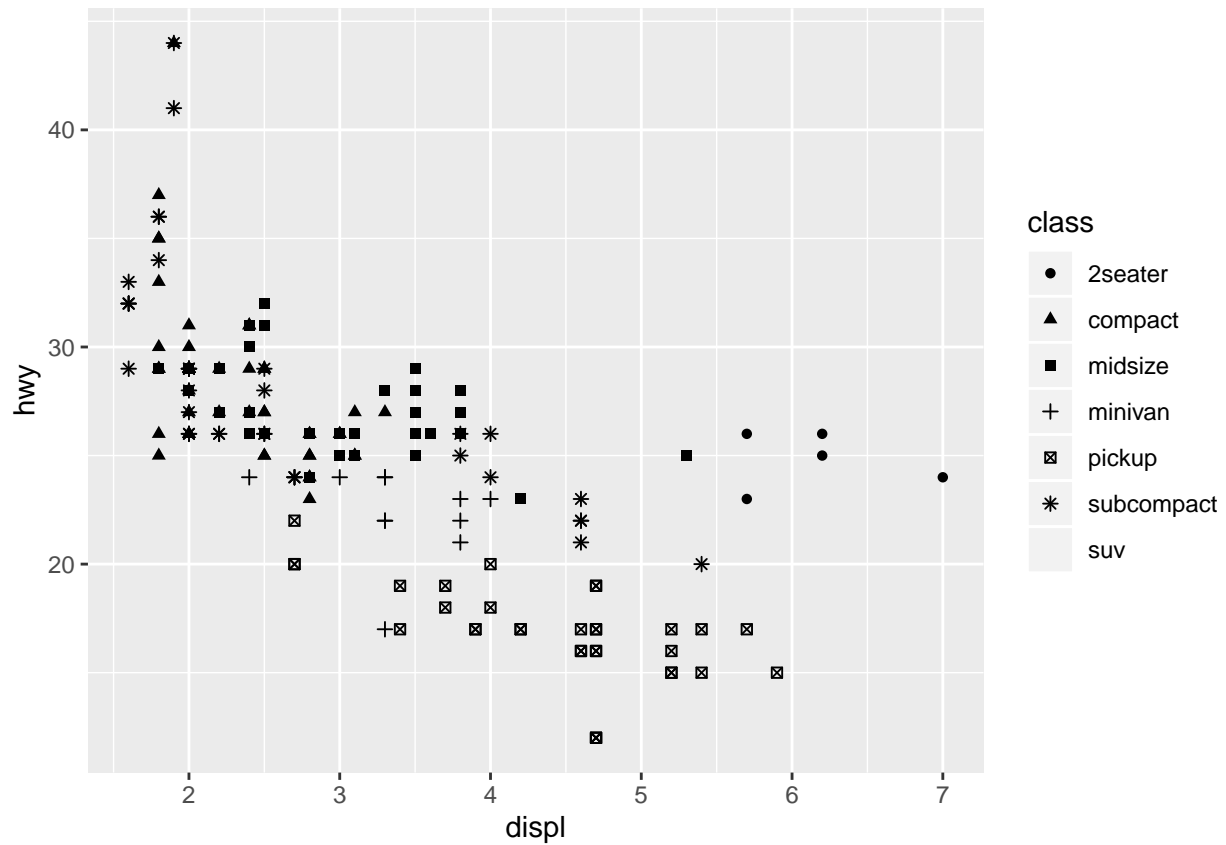
However, for shape method, maximum shape in a graph is 6, there are seven. the additional groups are unplotted.
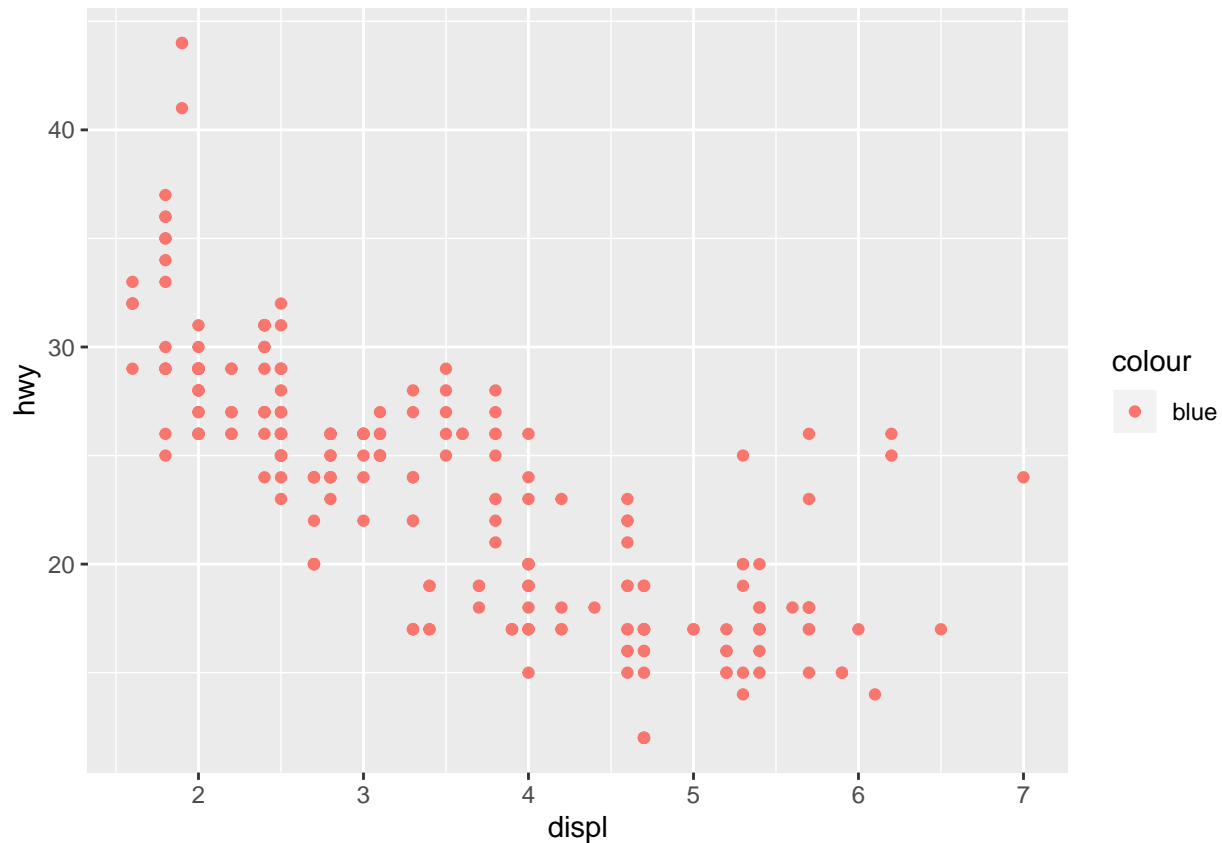
We also can directly add color to the plot.

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

Exercise

1. What is gone wrong with this code? Why are the points not blue?

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

Answer

It shouldn't inside the aes bracket.

2. Which variables in `mpg` are categorical? Which variables are continuous? use?mpghow can you see this information when you run `mpg`

```
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(l~ f        18    29 p     comp~
##  2 audi         a4         1.8  1999     4 manual~ f        21    29 p     comp~
##  3 audi         a4         2    2008     4 manual~ f        20    31 p     comp~
##  4 audi         a4         2    2008     4 auto(a~ f        21    30 p     comp~
##  5 audi         a4         2.8  1999     6 auto(l~ f        16    26 p     comp~
##  6 audi         a4         2.8  1999     6 manual~ f        18    26 p     comp~
##  7 audi         a4         3.1  2008     6 auto(a~ f        18    27 p     comp~
##  8 audi         a4 quat~   1.8  1999     4 manual~ 4        18    26 p     comp~
##  9 audi         a4 quat~   1.8  1999     4 auto(l~ 4        16    25 p     comp~
## 10 audi         a4 quat~   2    2008     4 manual~ 4        20    28 p     comp~
## # ... with 224 more rows
```
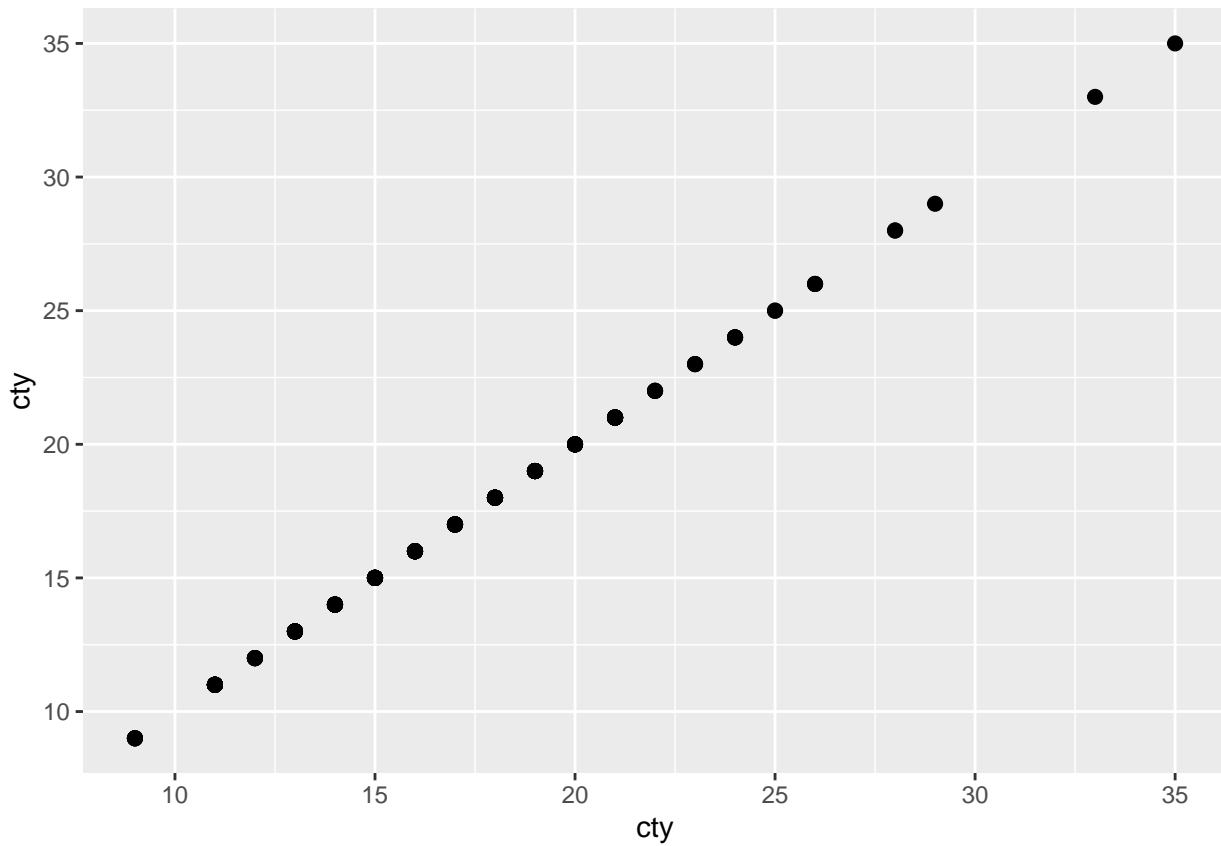
We can see this by seeing data structure

3. Map a continuous variable to `color`, `size`, and `shape`. How do these aesthetics behave differently for categorical vs. continuous variables?

Answer Nothing

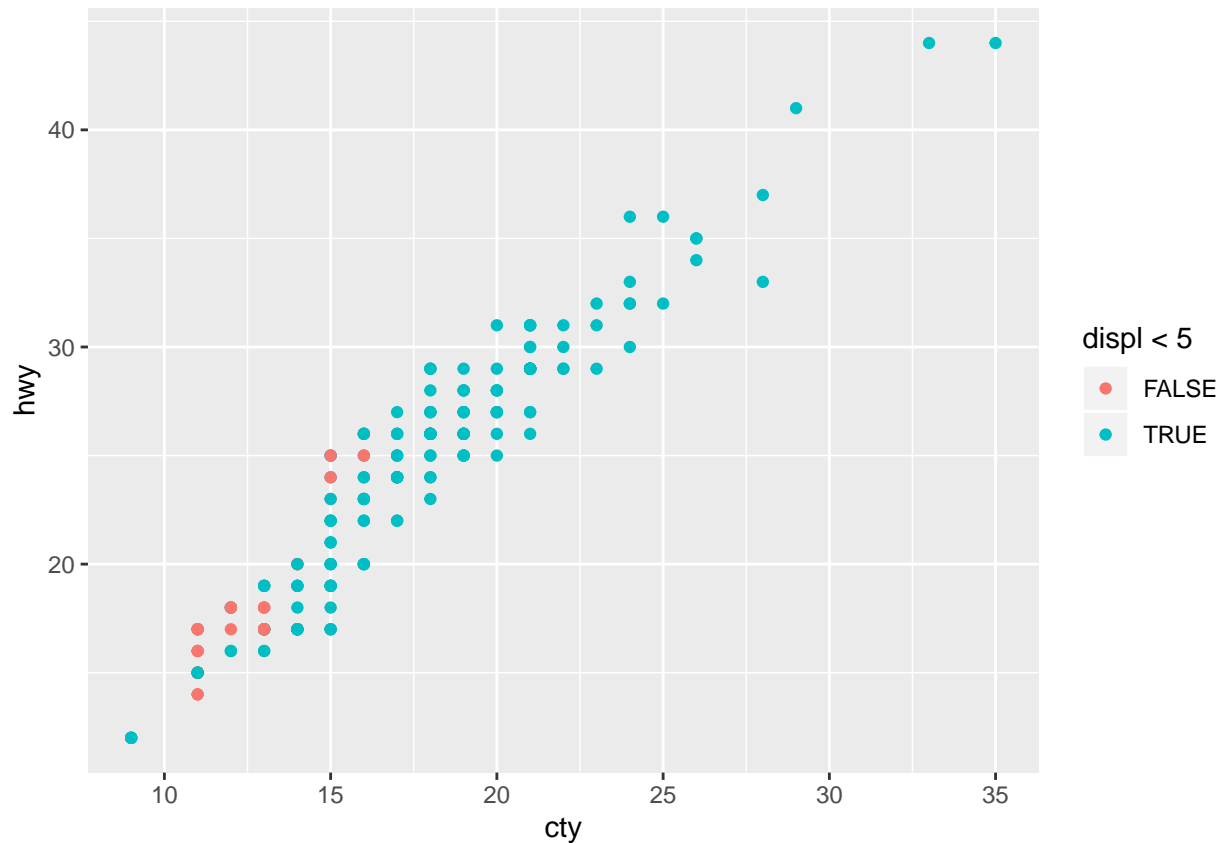**4. What happens if you map the same variable to multiple aesthetics**

```
ggplot(data=mpg)+geom_point(aes(x=cty,y=cty),stroke=1)
```


Answer It will form a 45 degree line.

6. What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you'll also need to specify x and y.

```
ggplot(data=mpg)+geom_point(aes(x=cty,y=hwy,color=displ<5))
```

Answer

It distinguishs data point with the displ with <5 and >5.

Common problems

Remember to use pairness and put everything on th right position.

Facets

Facets is particularly useful for categorical variables, is to split your plot into facets.
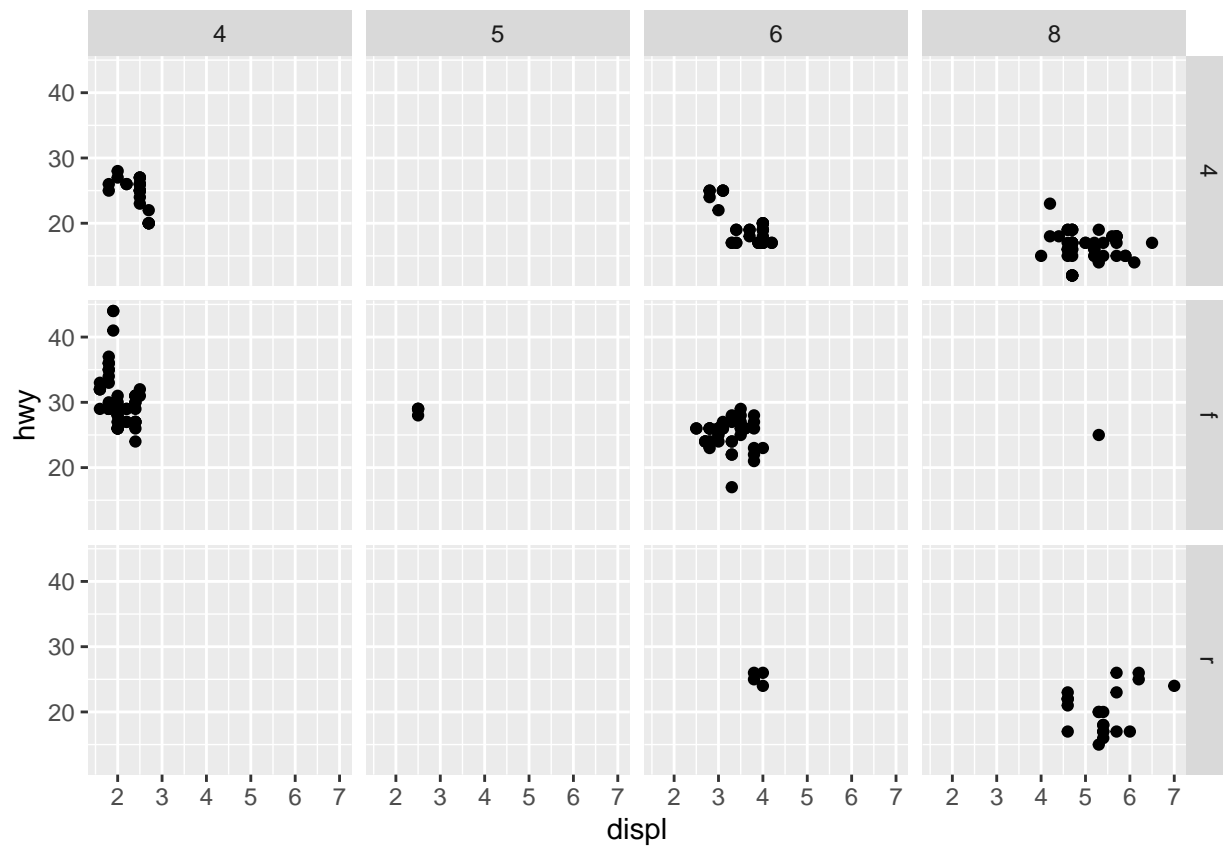
`facet_wrap`: first argument is `~`, which is followed by a variable name. the variable you pass to `facet_wrap` must be discrete

```
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy))+
  facet_wrap(~cyl,nrow=2)
```
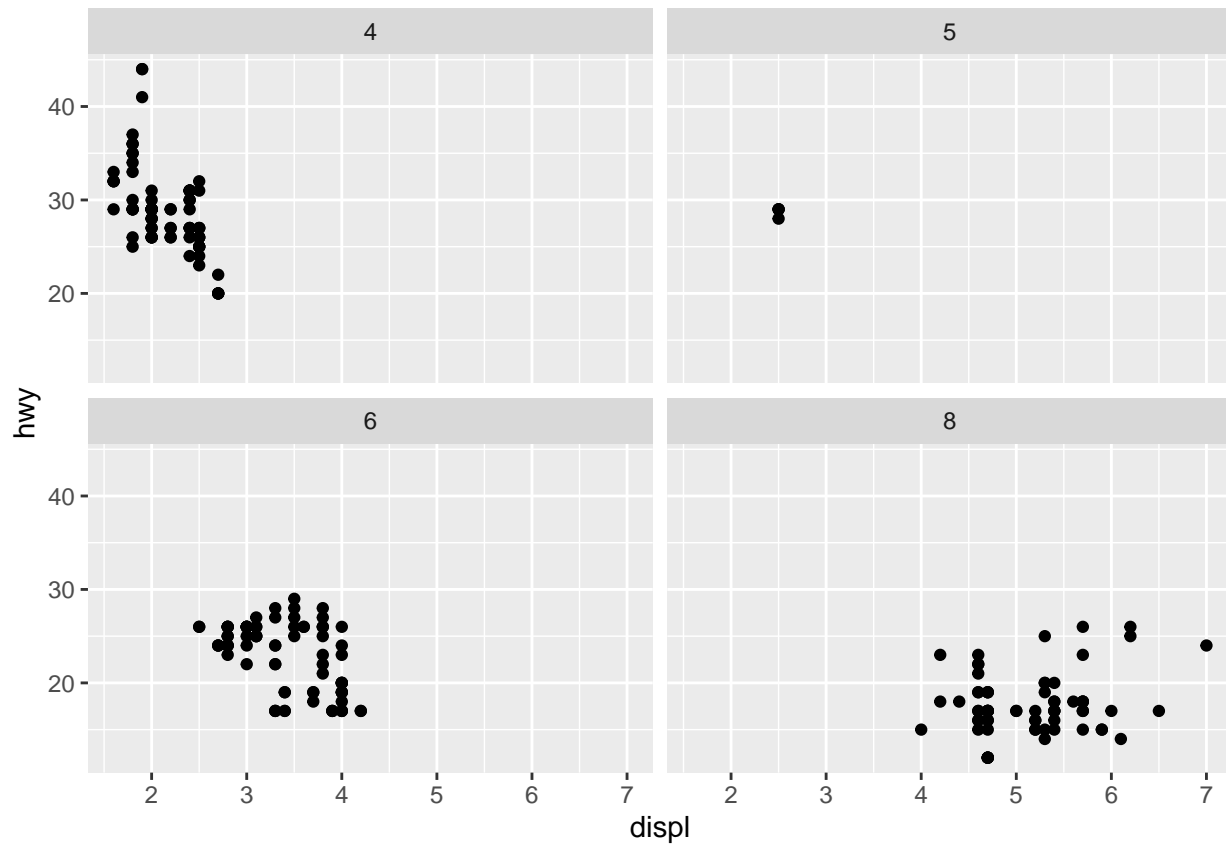
facet_grid() is contain two discrete variables ~

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy))+
  facet_grid(drv~cyl)
```

If you prefer to not facet in the rows or columns dimension.

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy))+
  facet_grid(.~cyl)
```