Introduction to Visualization in R

Pramod Duvvuri

3/22/2019

Introduction to ggplot2

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------ tidyverse 1.2.1 -

## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0

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

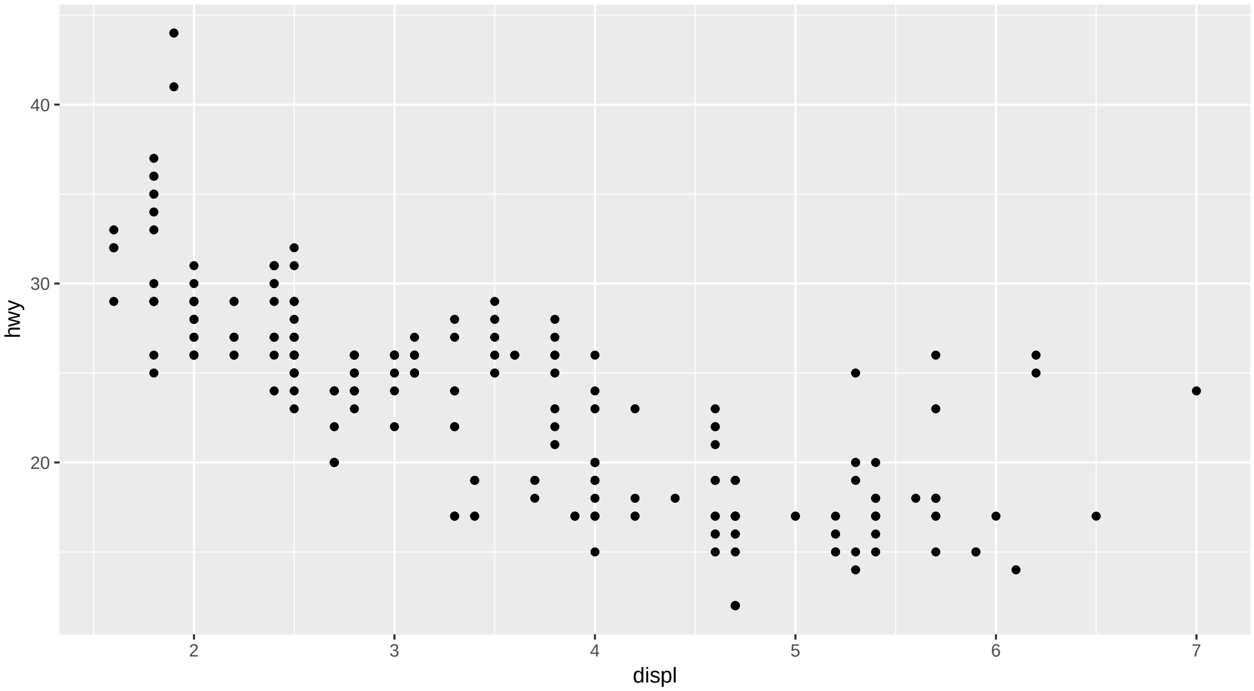
```r
mpg
```

```
## # 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 q~   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 q~   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 q~   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

```r
attach(mpg)
```

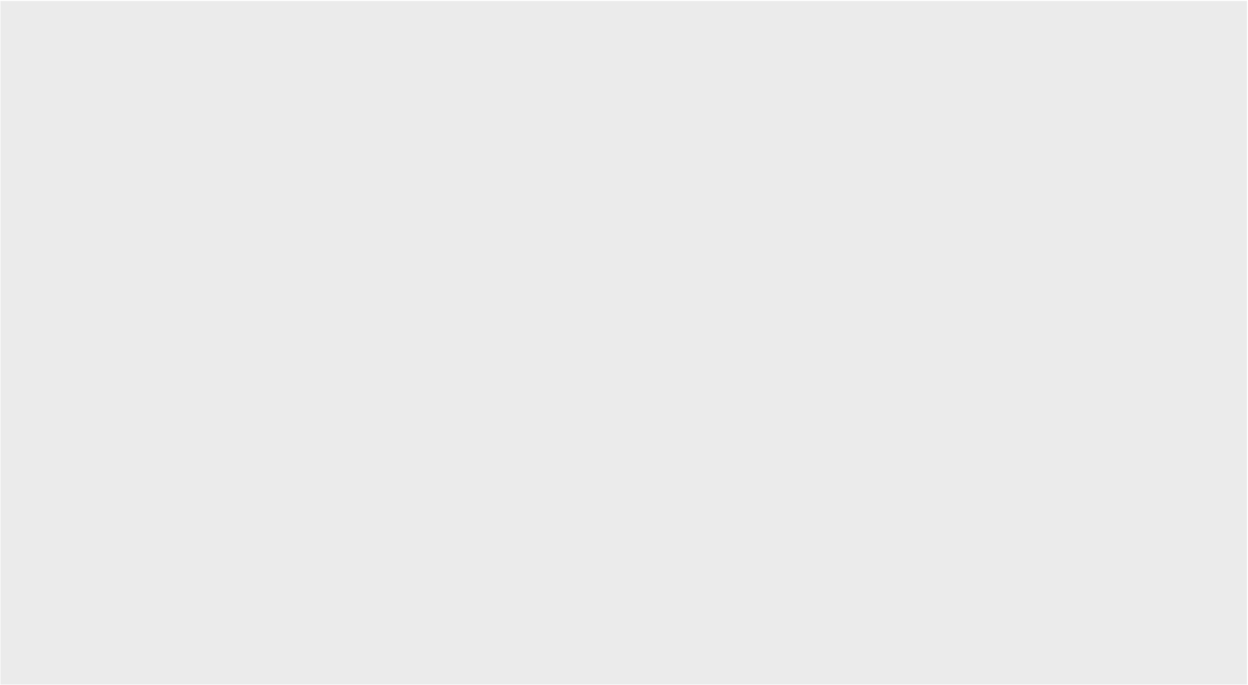
```r
# ggplot2 TEMPLATE
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

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



Exercises-1

```r
# Question-1
ggplot(data = mpg)
```



```r
# Question-2
nrow(mpg) # Rows
```

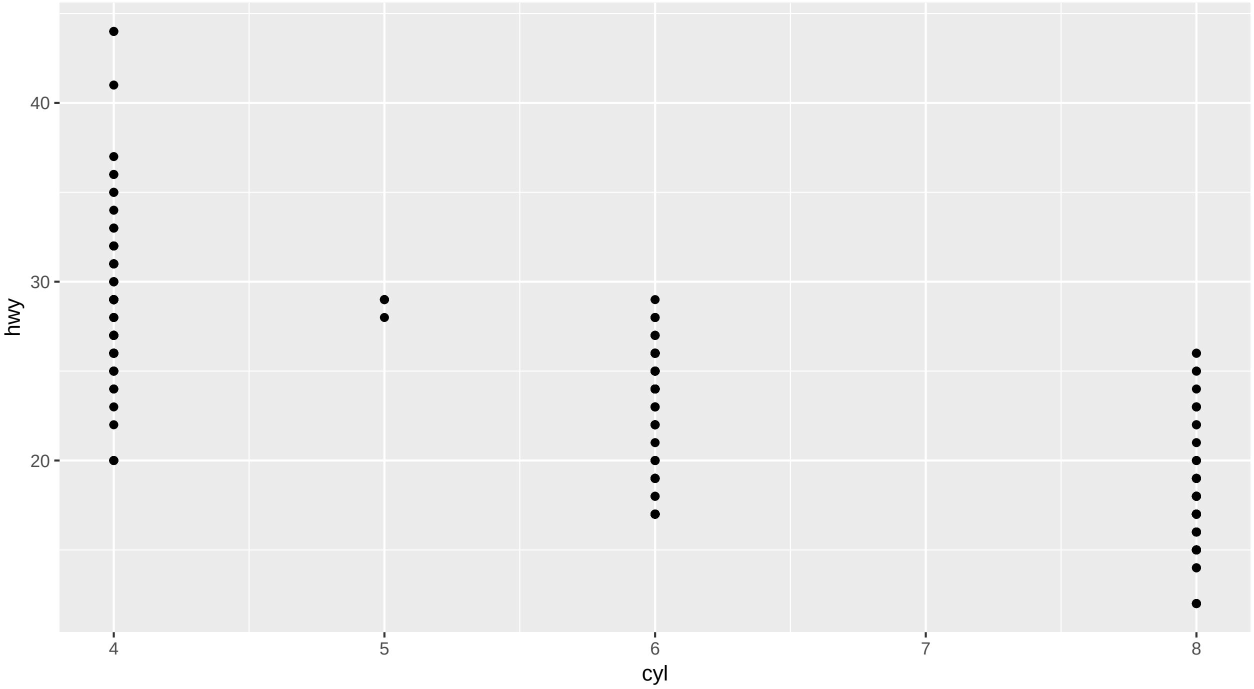
```
## [1] 234
```

```r
ncol(mpg) # Columns
```

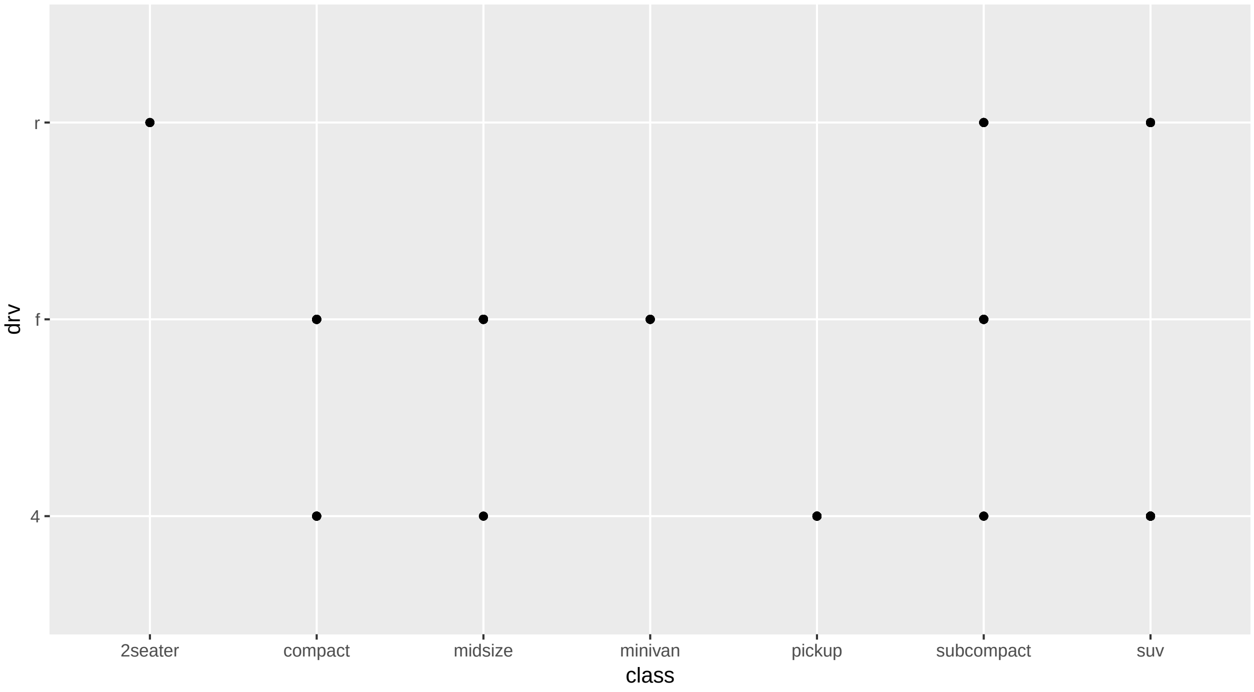
```
## [1] 11
```

```r
# Question-3
?mpg
# Answer: f = front-wheel drive, r = rear wheel drive, 4 = 4wd
```

```r
# Question-4
ggplot(data = mpg) + geom_point(mapping = aes(x = cyl, y = hwy))
```

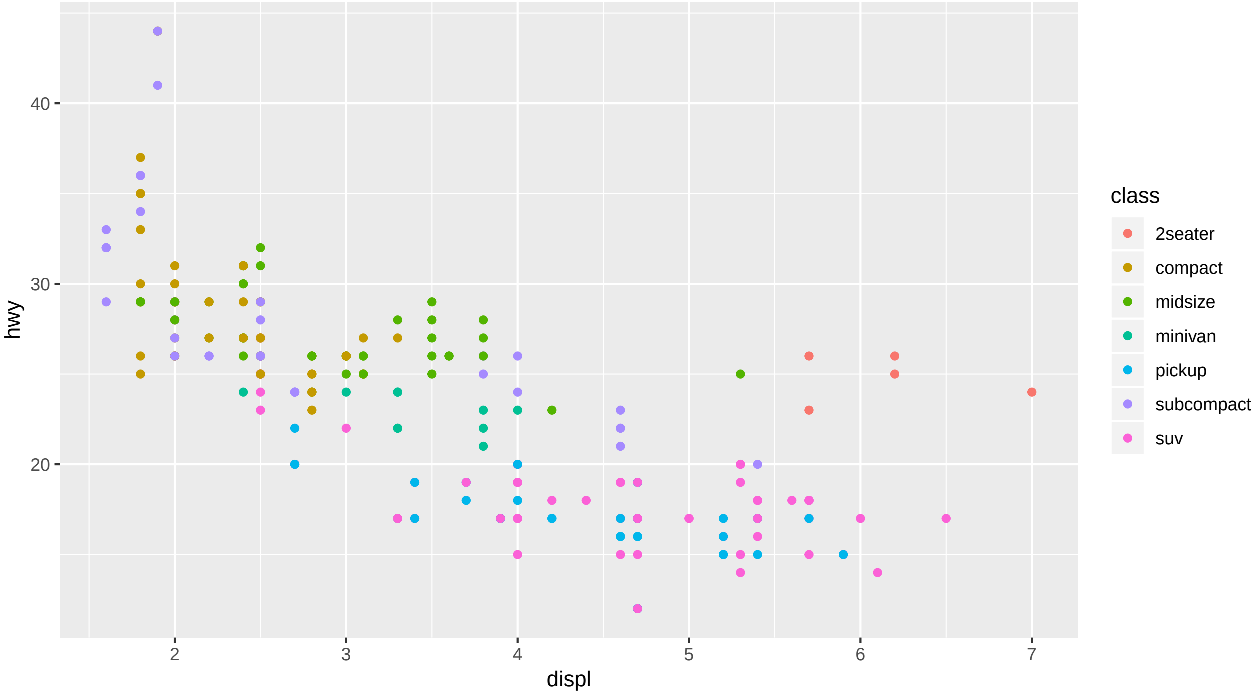


```r
# Question-5
ggplot(data = mpg) + geom_point(mapping = aes(x = class, y = drv))
```



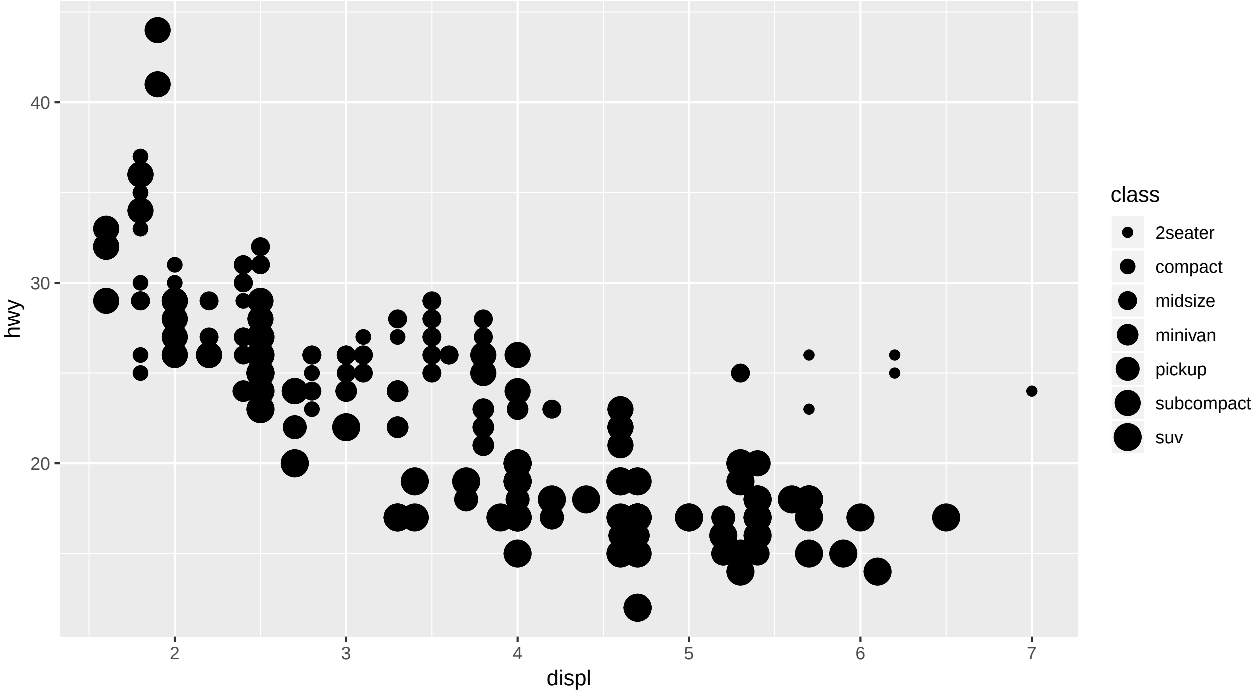
Aesthetic Mappings

```r
# Adding aesthetic to the plot (color/size/shape could be used here)
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              color = class))
```



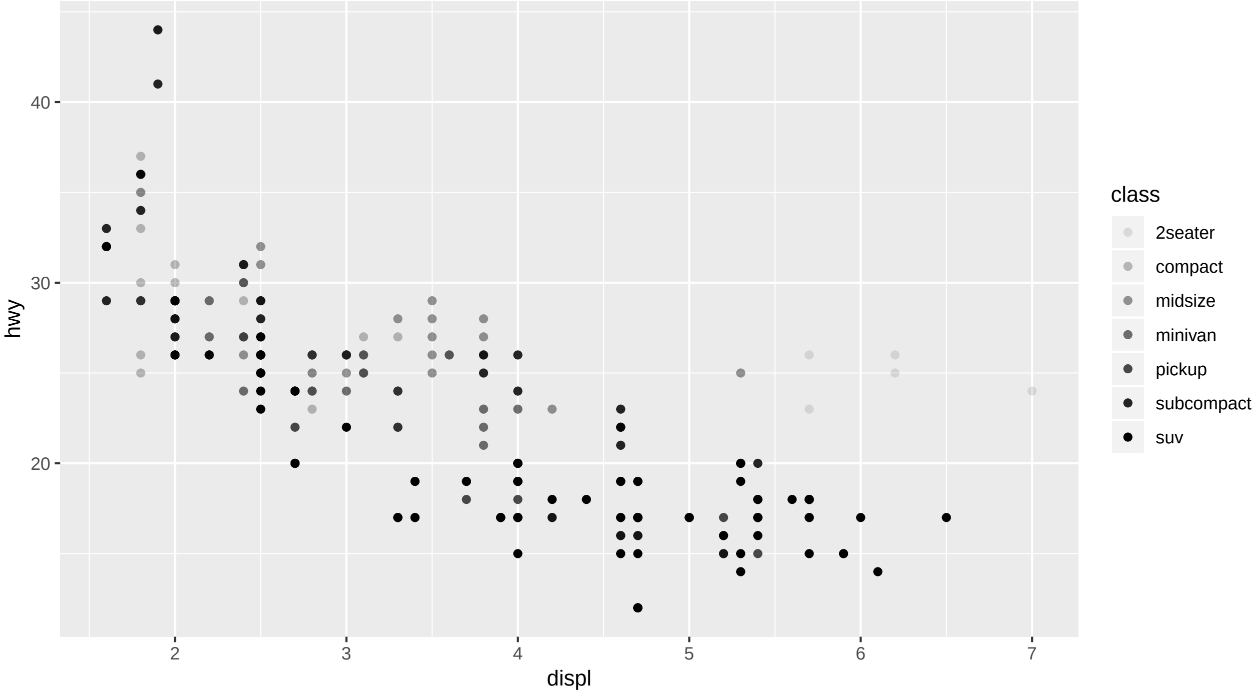
```r
# Using size as aesthetic instead of color
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              size = class))
```

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



```
# Alpha and Shape Aesthetics
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              alpha = class))
```

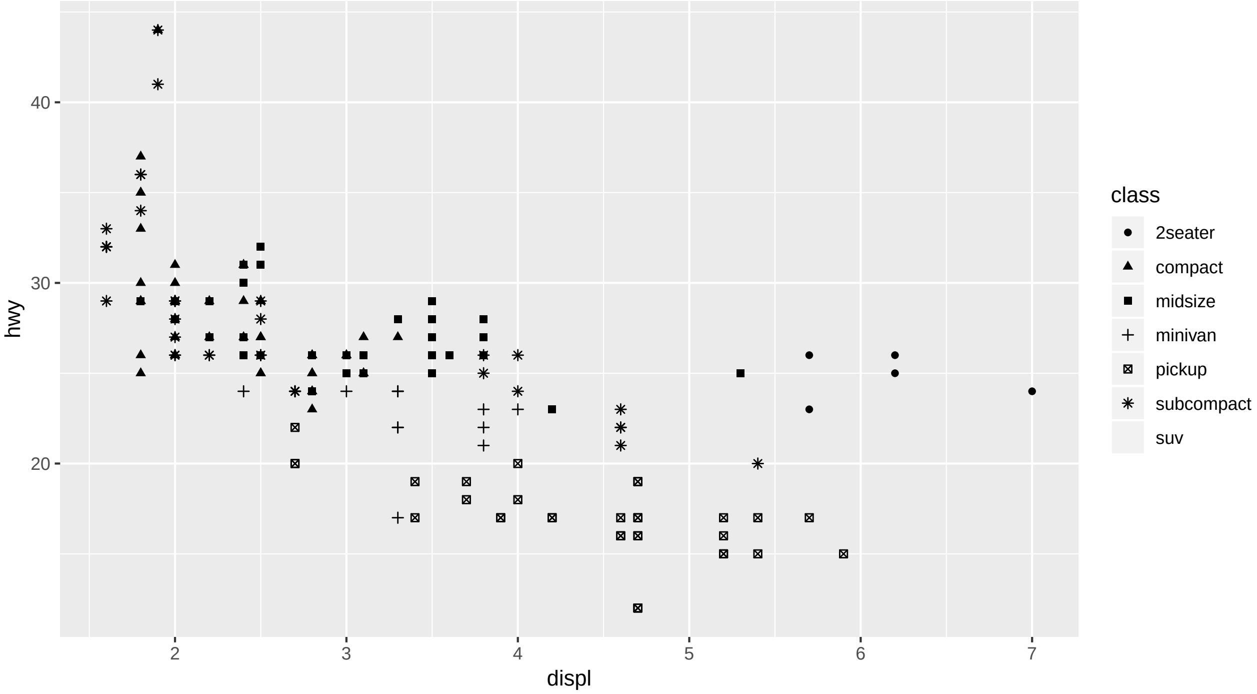
```
## Warning: Using alpha for a discrete variable is not advised.
```



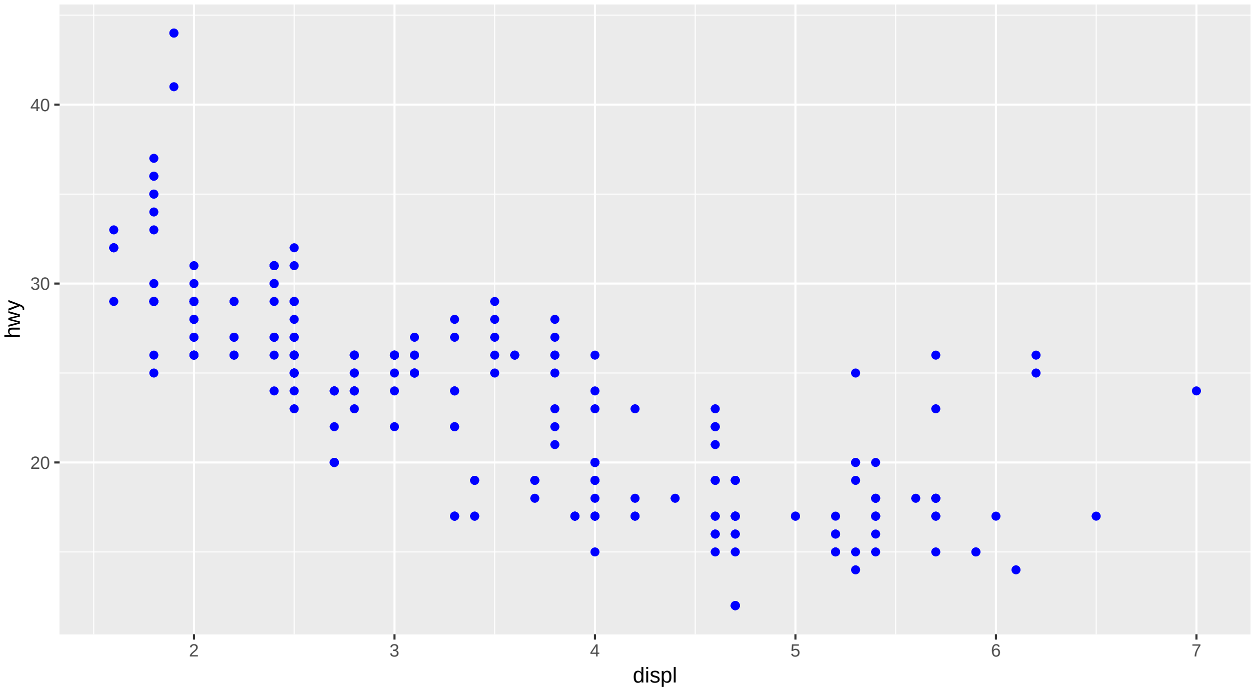
```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              shape = class))
```

```
## 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).
```



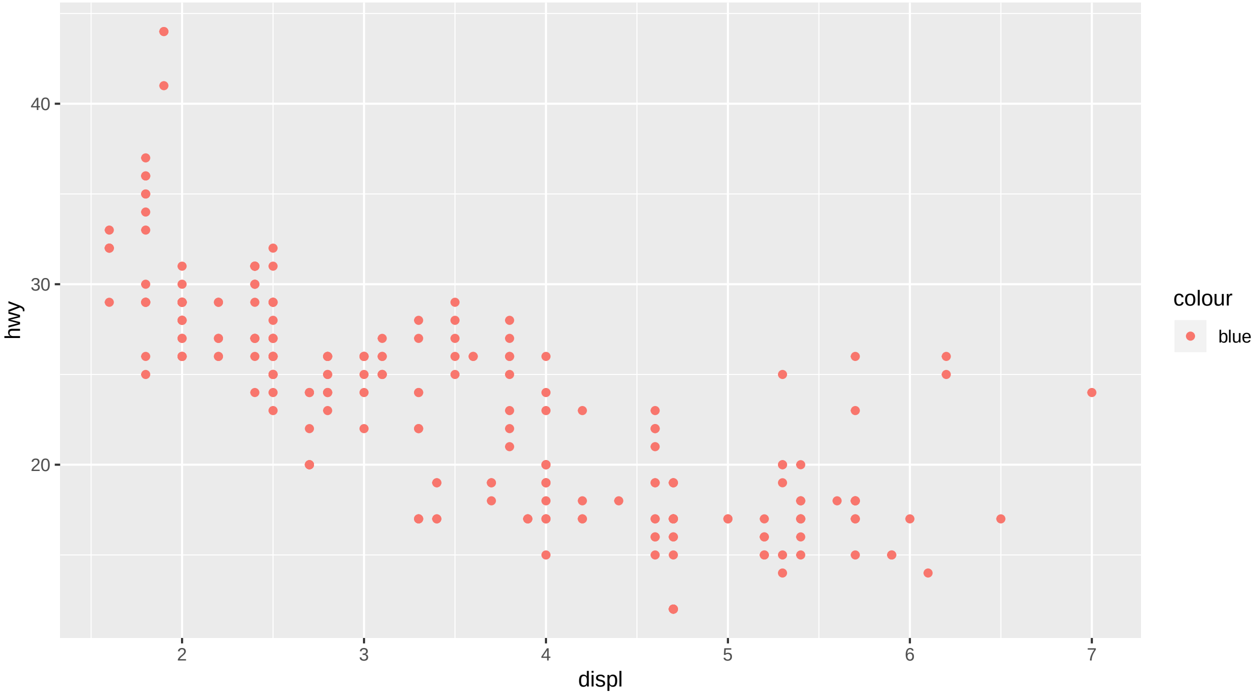
```
# Using a different color for the data points
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy),
                                              color = "blue")
```



Exercises-2

Q1. What's gone wrong with this code? Why are the points not blue?

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



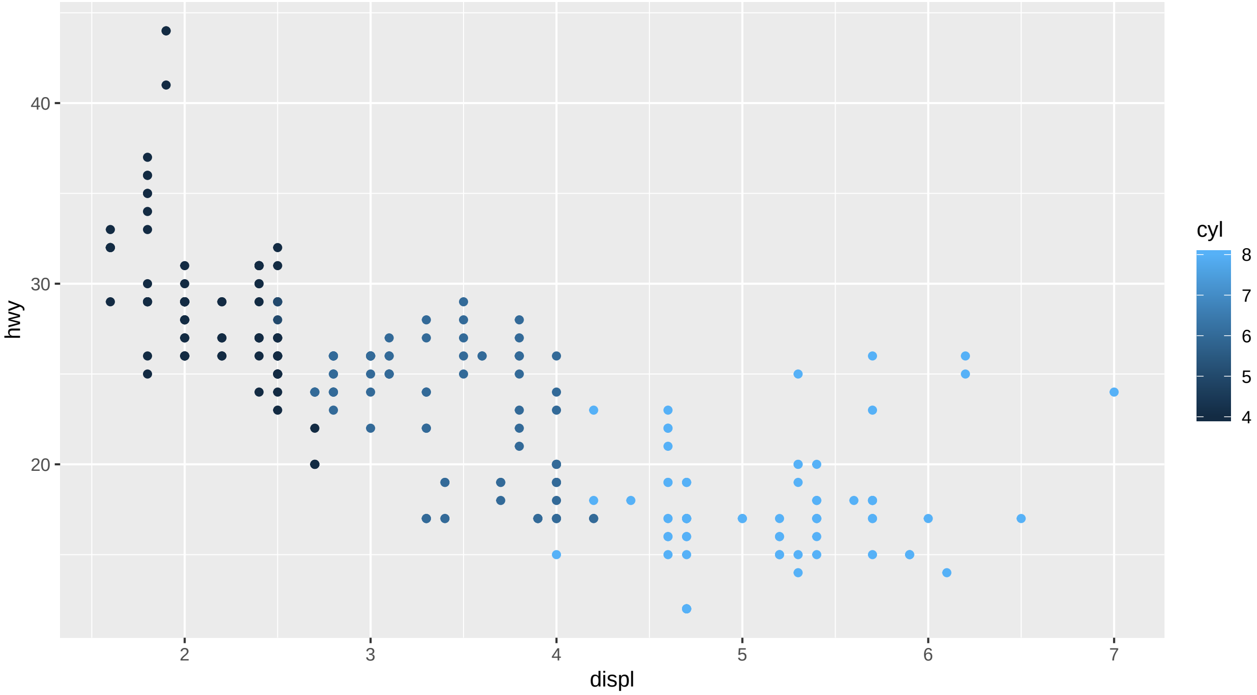
Q2. Which variables in mpg are categorical? Which variables are continuous? (Hint: type ?mpg to read the documentation for the dataset). How can you see this information when you run mpg?

```
summary(mpg)
```

```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

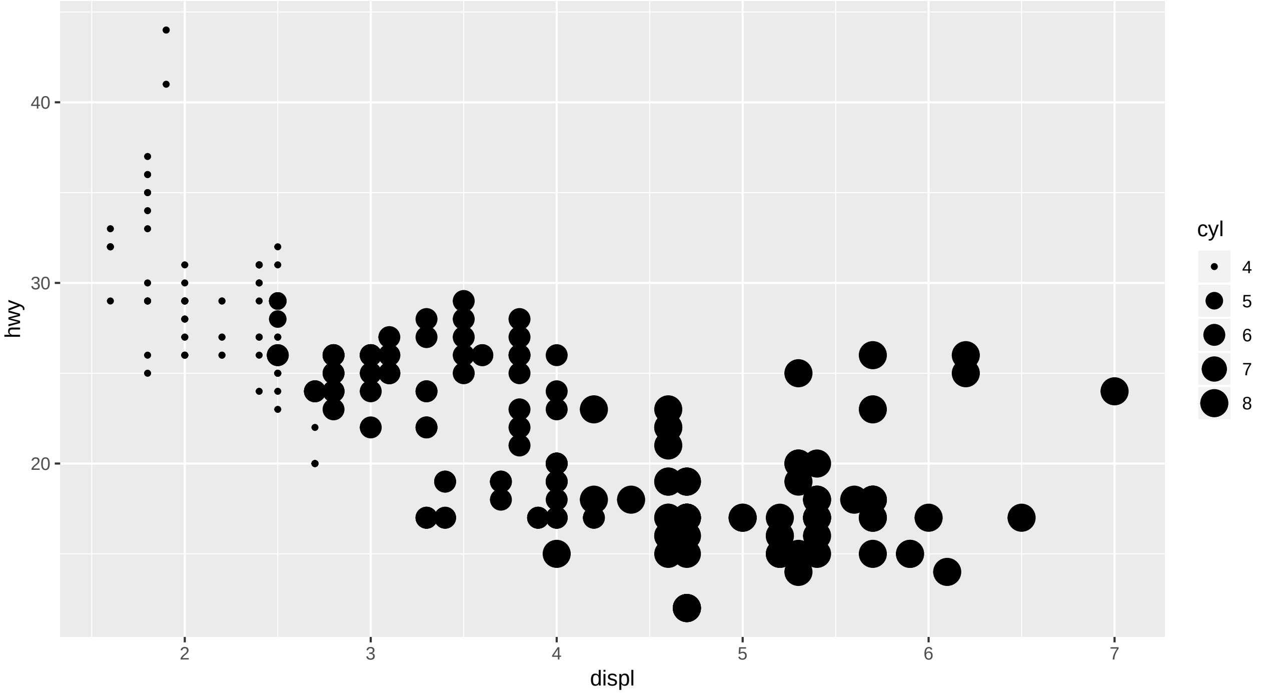
Q3. Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

```
# Mapping a continuos variable to color
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              color = cyl))
```



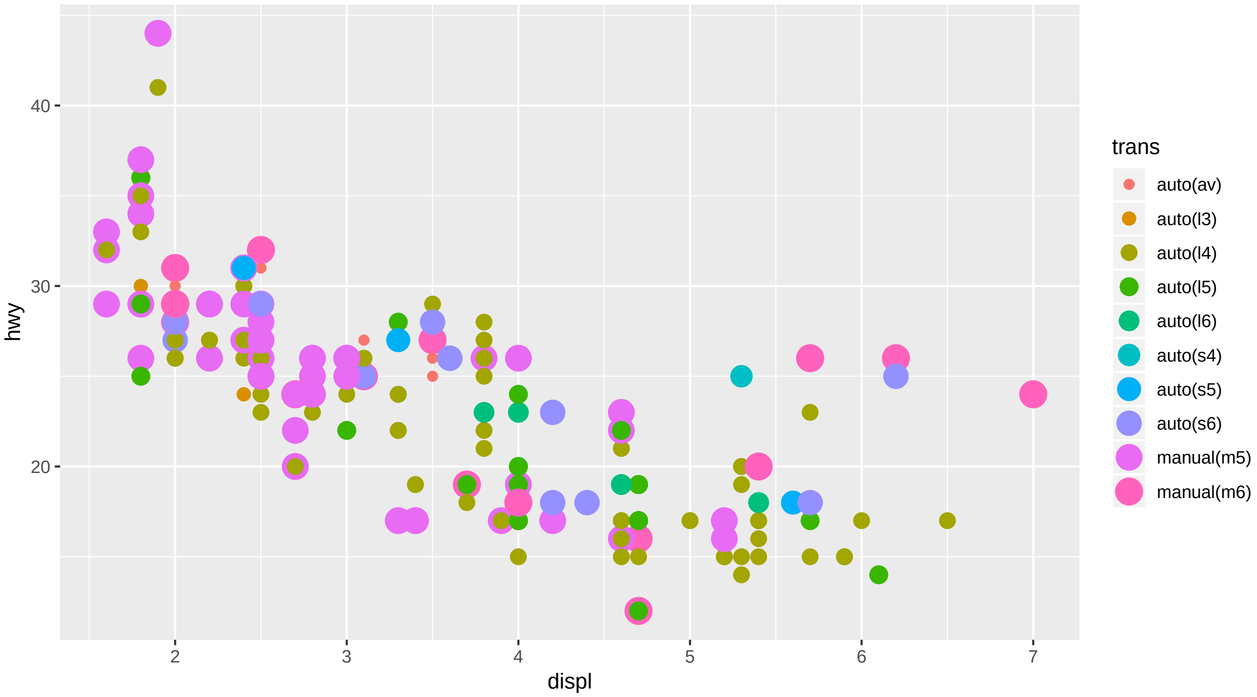
```
# Shape cannot be mapped, the line below throws an error
#ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
#                                              shape = cty))
```

```
# Mapping a continuos variable to size
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              size = cyl))
```

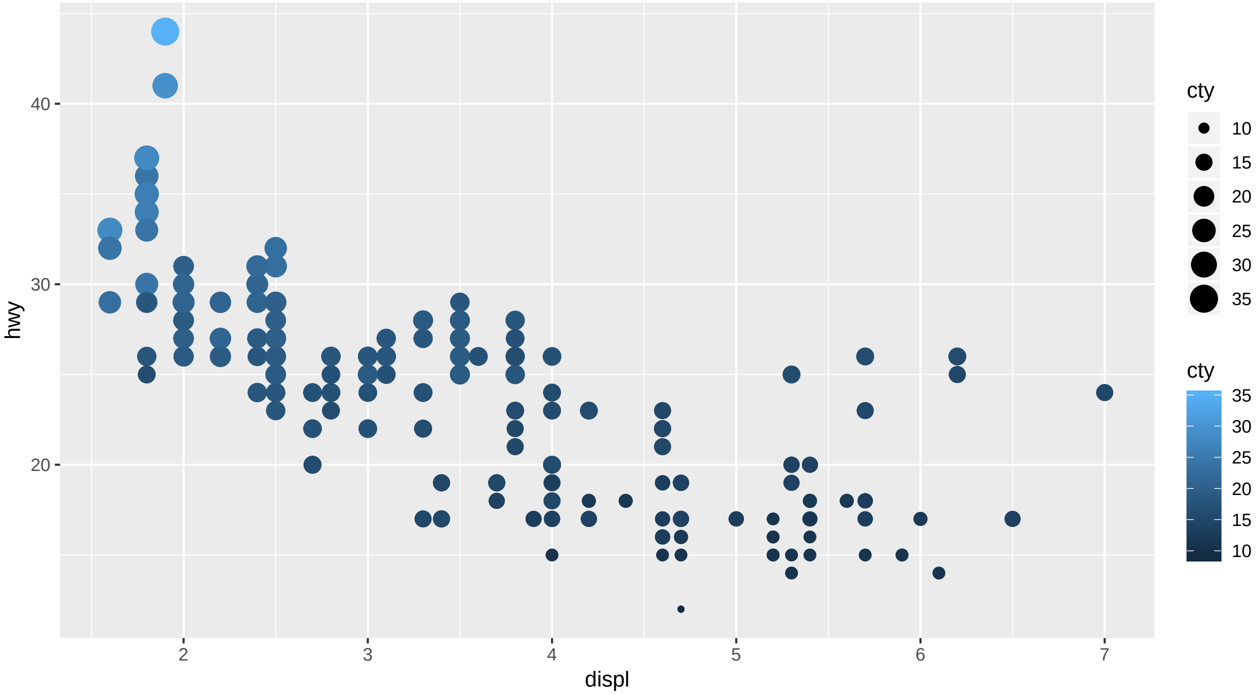


```
# Mapping same variable to multiple aesthetics (Discrete Variable)
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              color = trans, size = trans))
```

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



```
# Mapping same variable to multiple aesthetics (Continuos Variable)
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy,
                                              color = cty, size = cty))
```

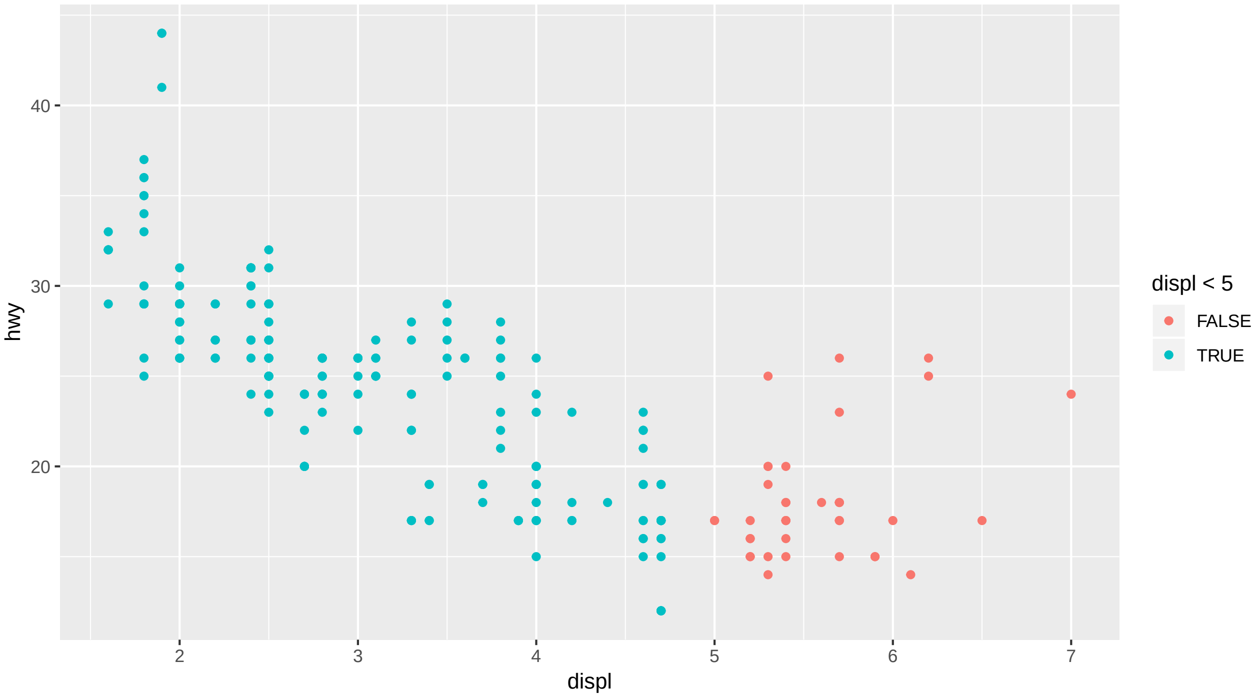


Q. What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom_point)

```
?geom_point # stroke aesthetic
```

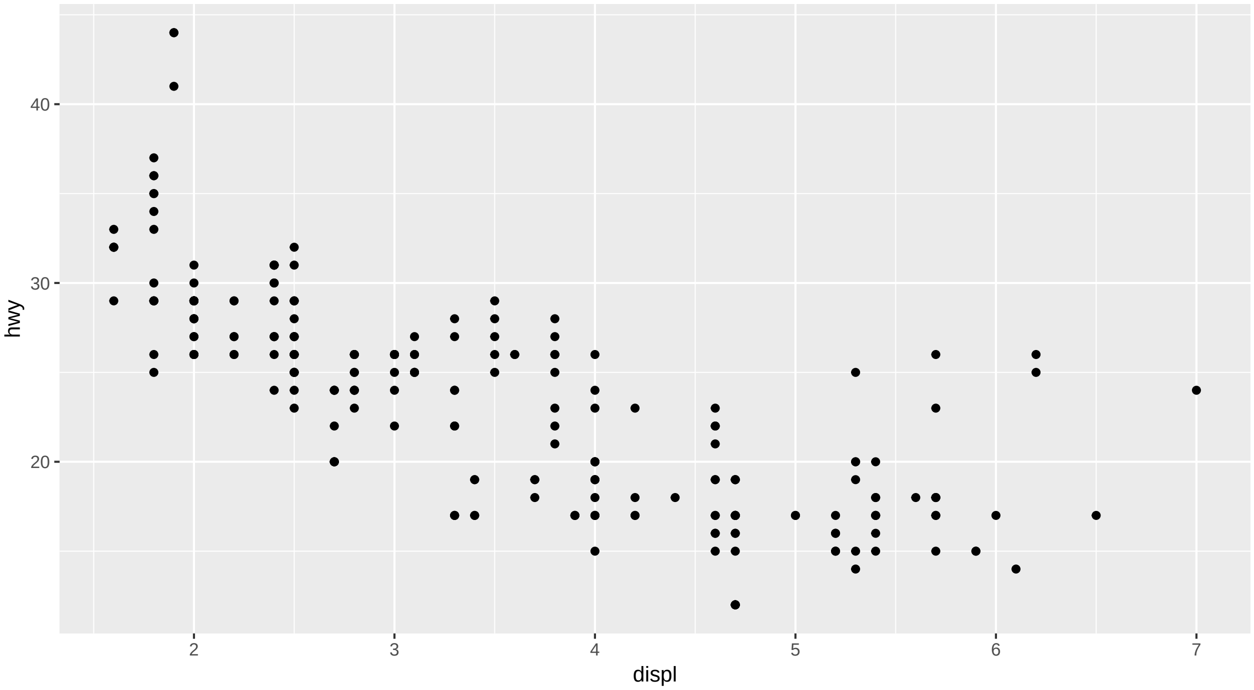
Q. What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you'll also need to specify x and y.

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

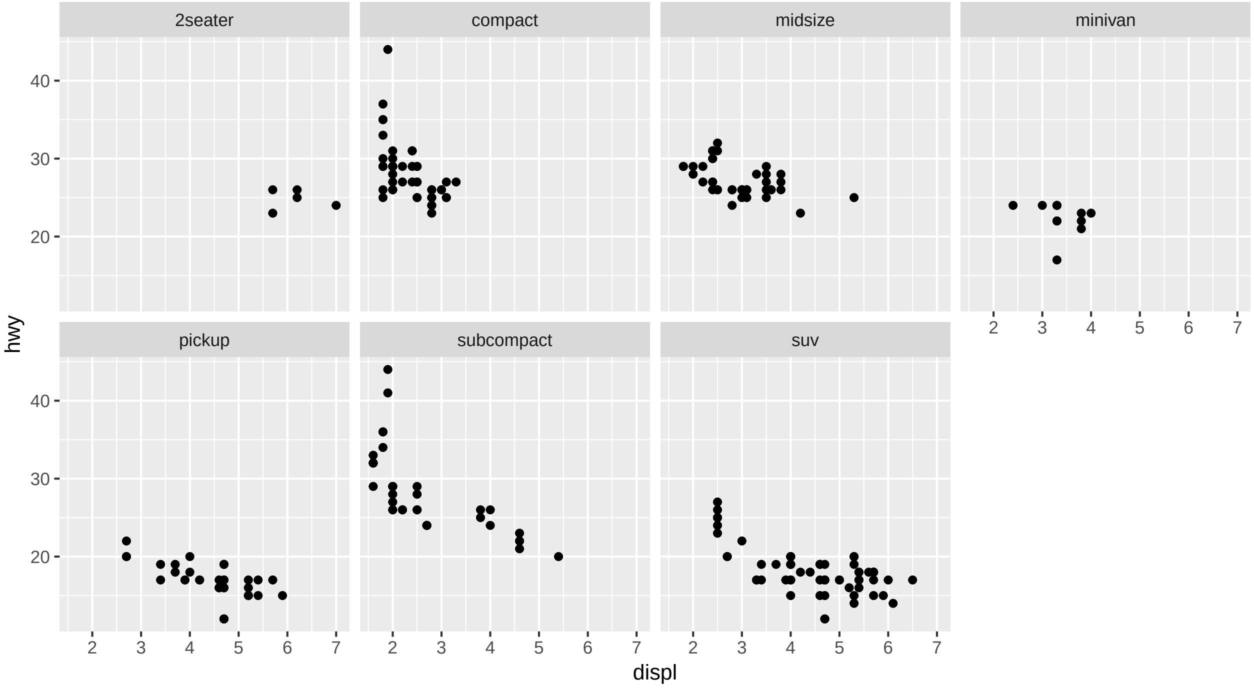


Facets

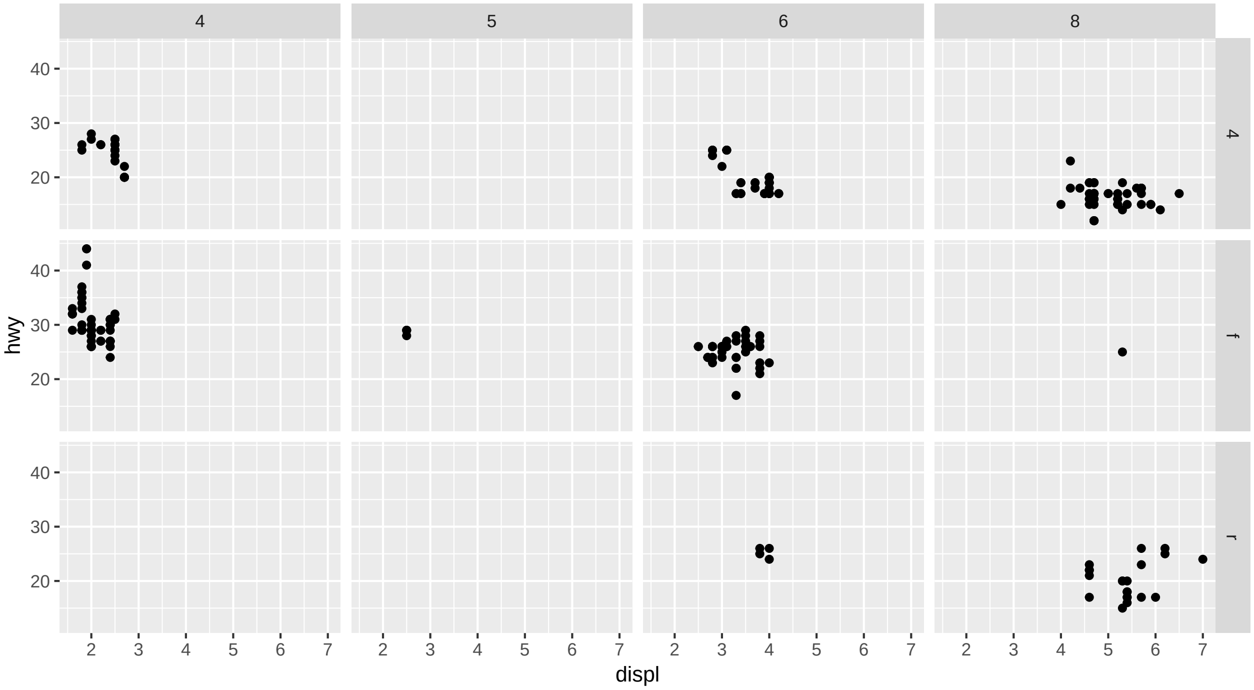
```
# Original Plot
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))
```



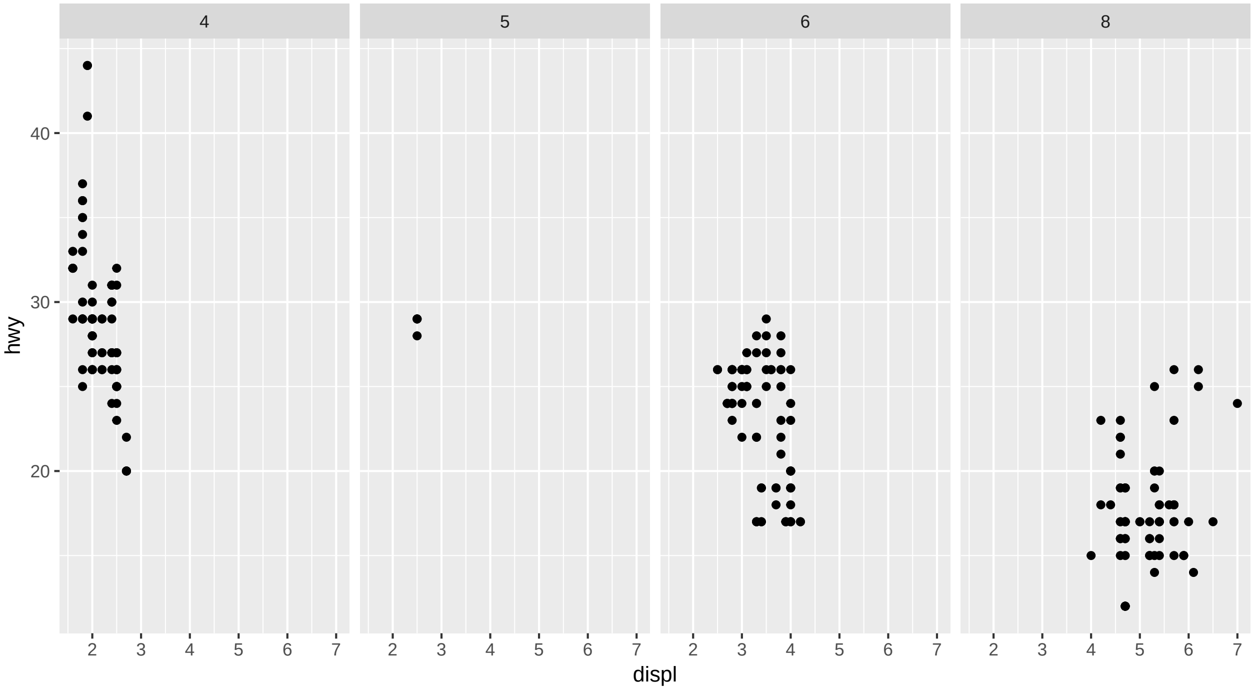
```r
# Using facet_wrap()
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```



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



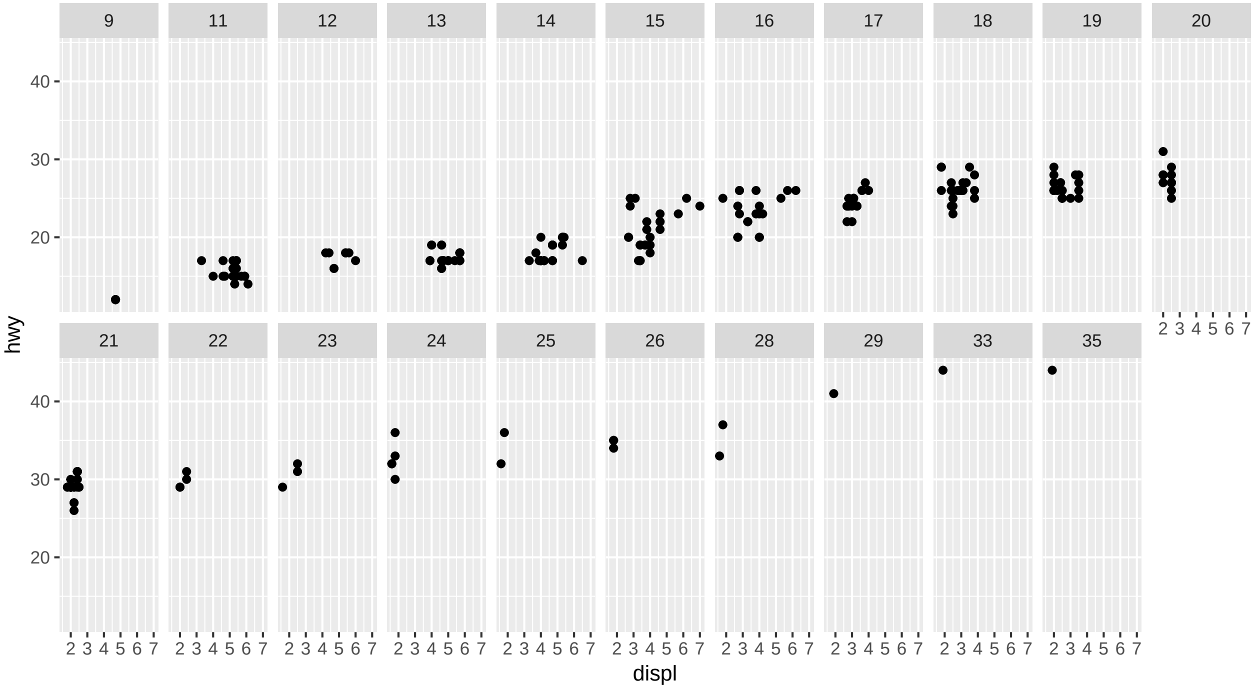
```r
# Using facet_grid()
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```



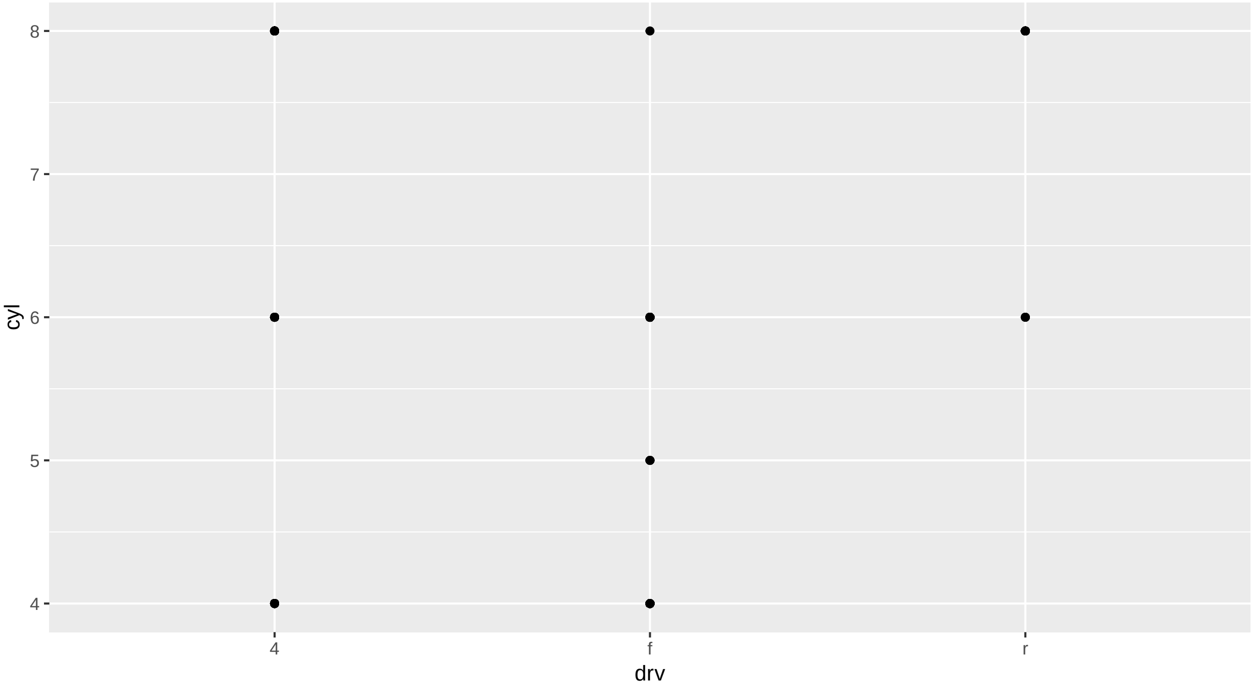
Exercises-3

Q1. What happens if you facet on a continuous variable?

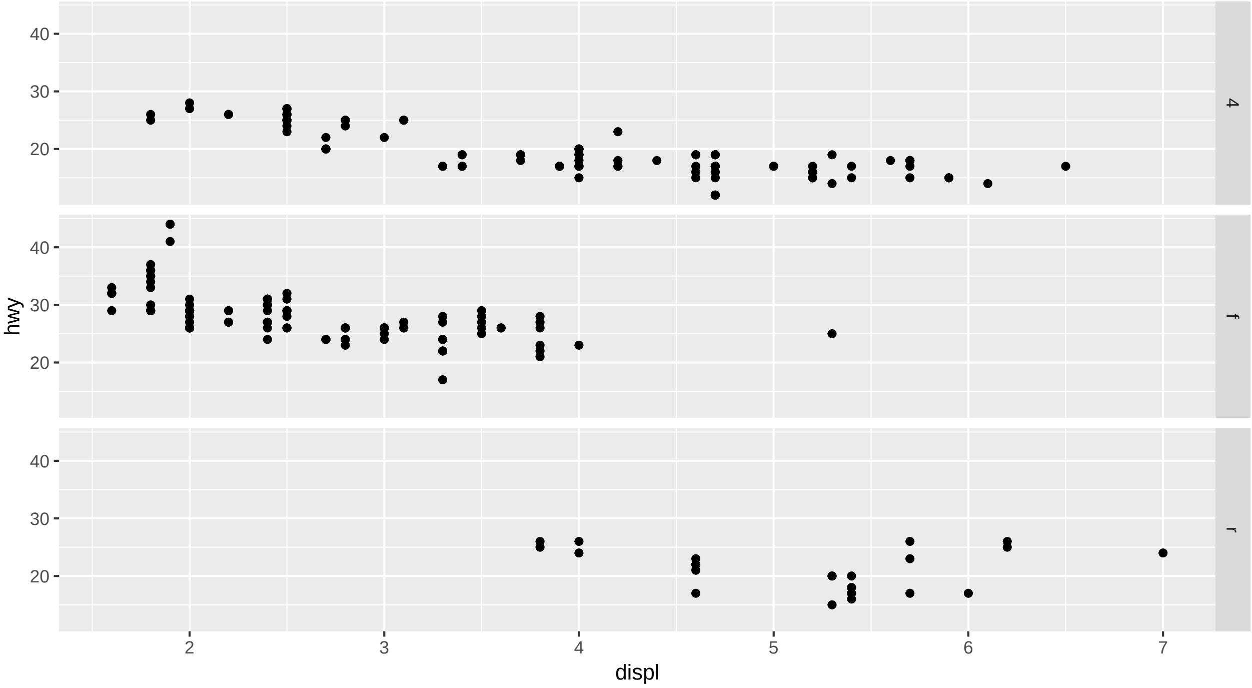
```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ cty, nrow = 2)
```



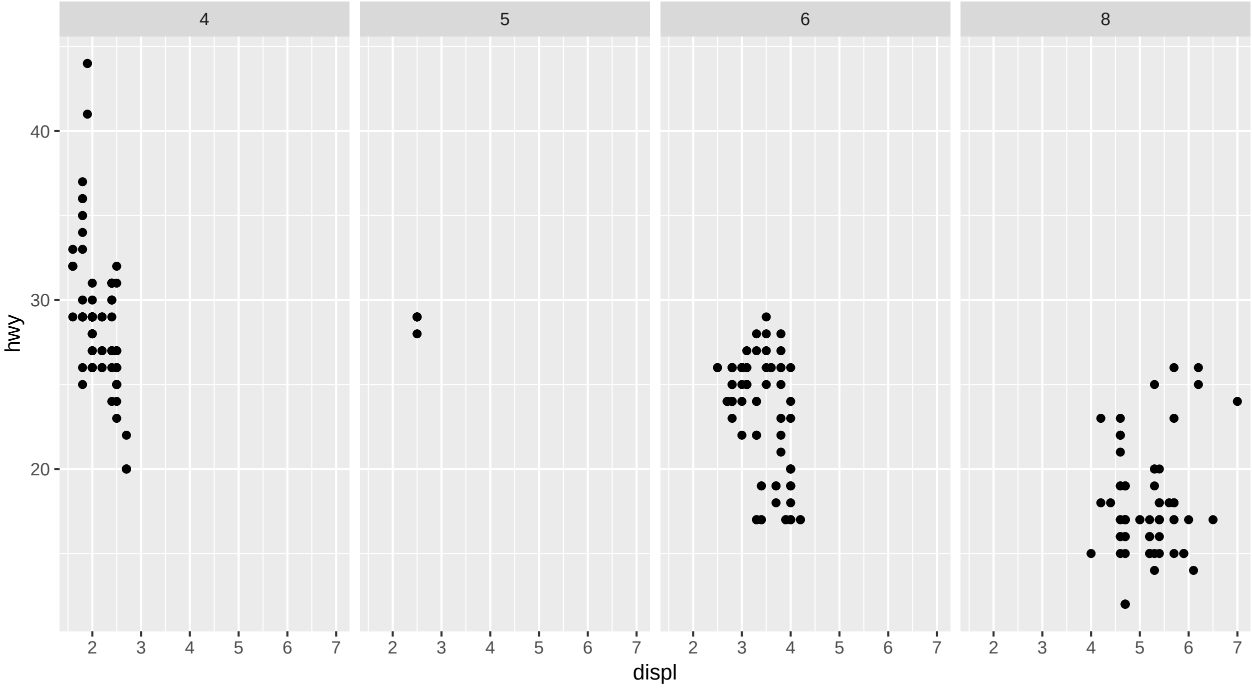
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))
```



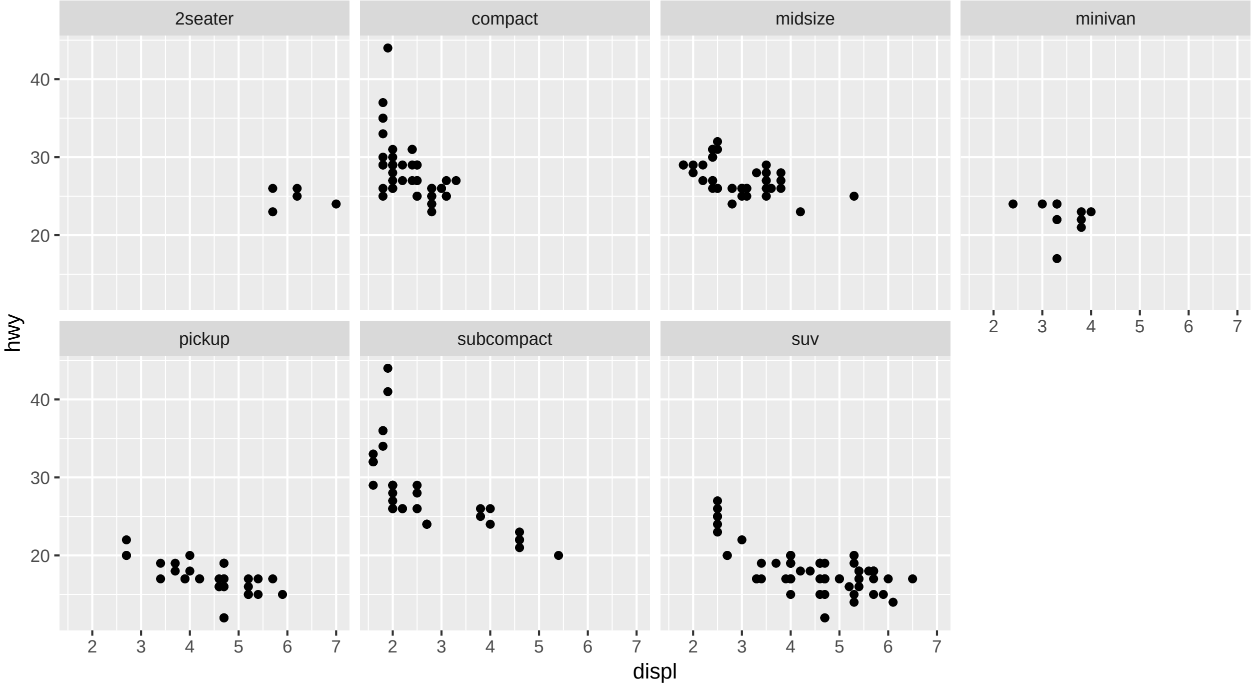
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ .)
```



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



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```

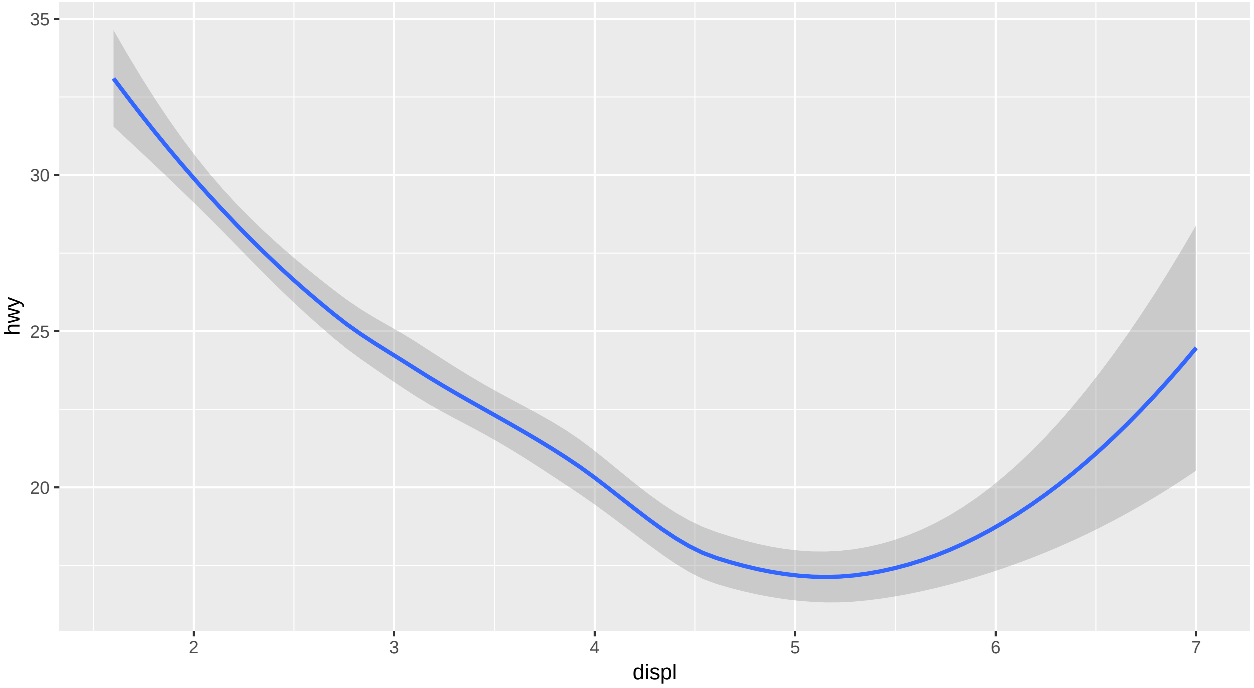


```
?facet_wrap
```

Geometric Objects

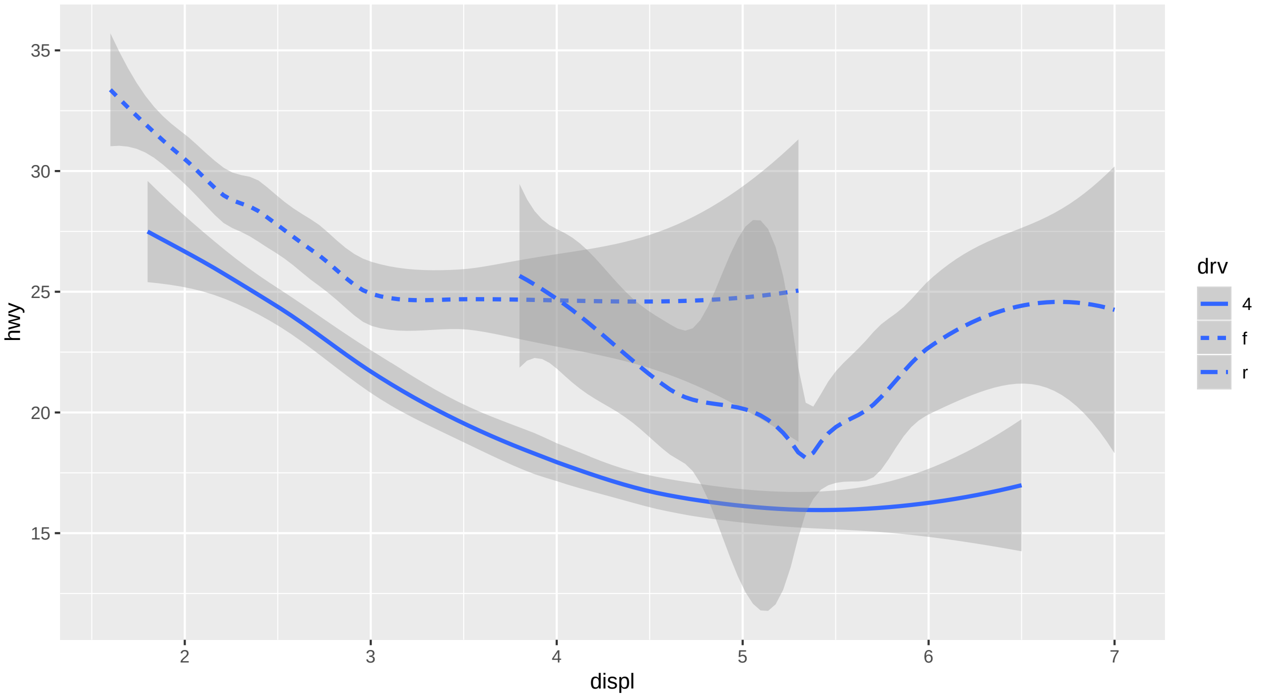
```
# Using geom_smooth()
ggplot(data = mpg) + geom_smooth(mapping =
                                   aes(x = displ, y = hwy))
```

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



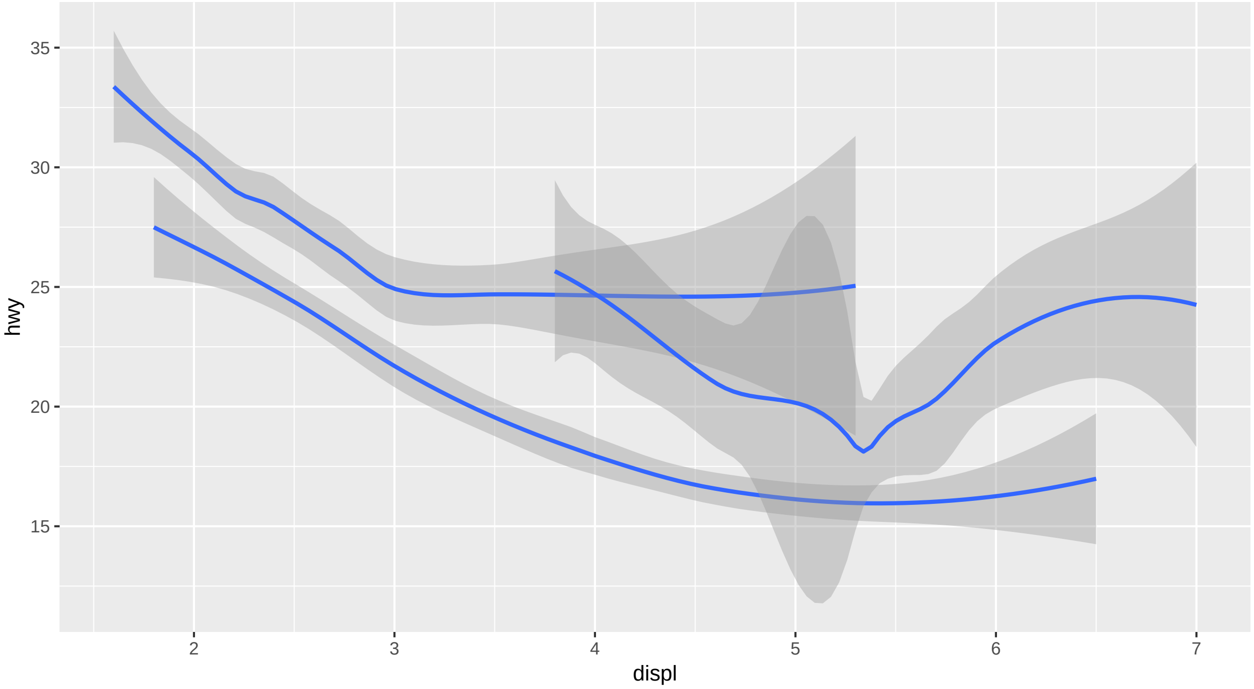
```
# Different linetype aesthetic
ggplot(data = mpg) + geom_smooth(mapping =
                                   aes(x = displ, y = hwy, linetype = drv))
```

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



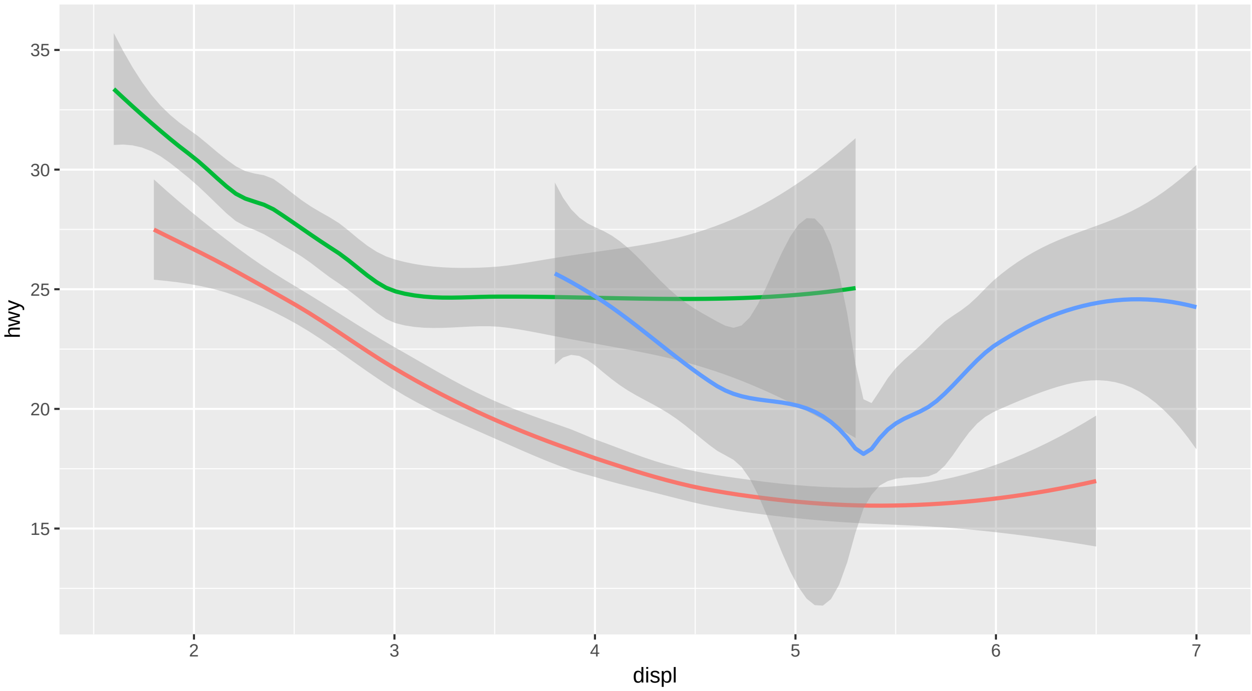
```r
ggplot(data = mpg) + geom_smooth(mapping = aes(x = displ, y = hwy, group = drv))
```

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



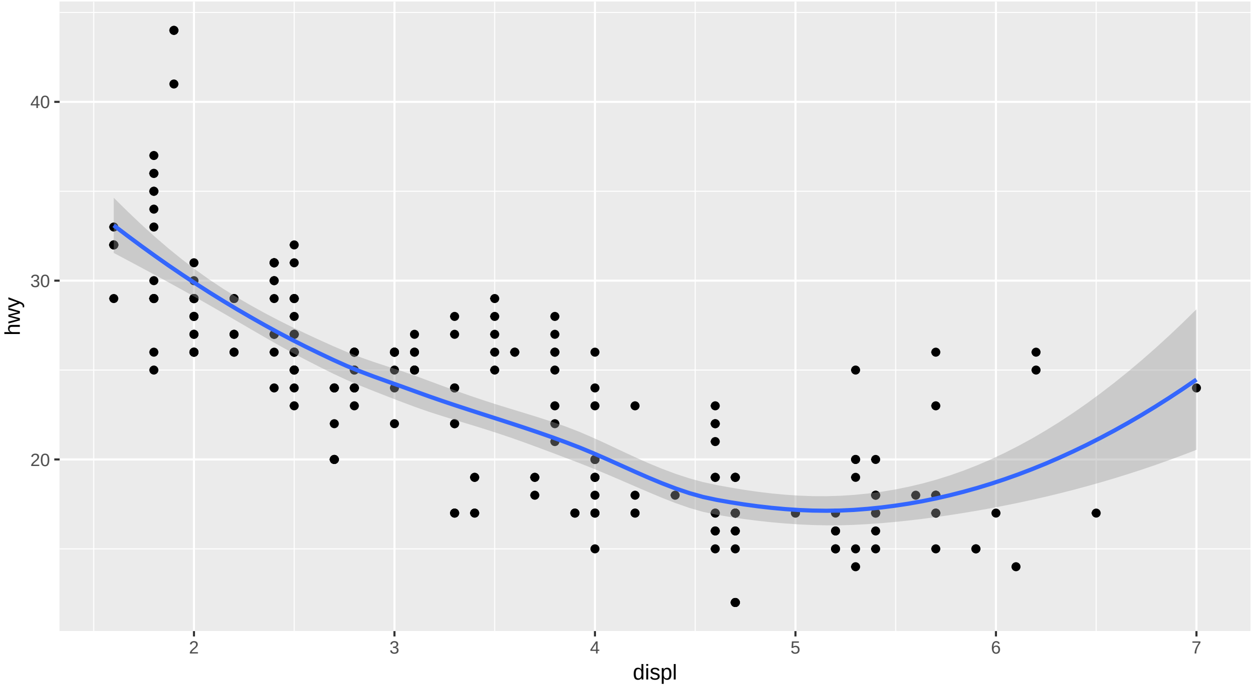
```r
# Disabling Legend
ggplot(data = mpg) + geom_smooth(mapping = aes(x = displ,
                                               y = hwy, color = drv),
                                               show.legend = FALSE)
```

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



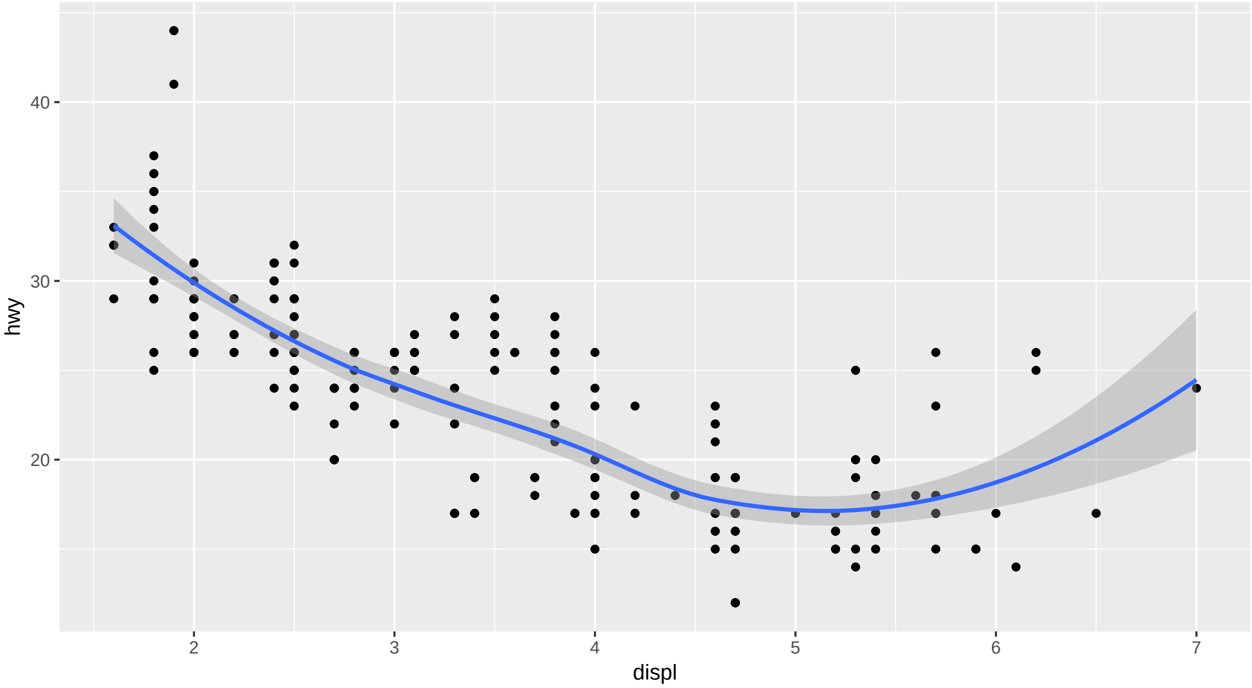
```r
# Multiple geoms in a Plot
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

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



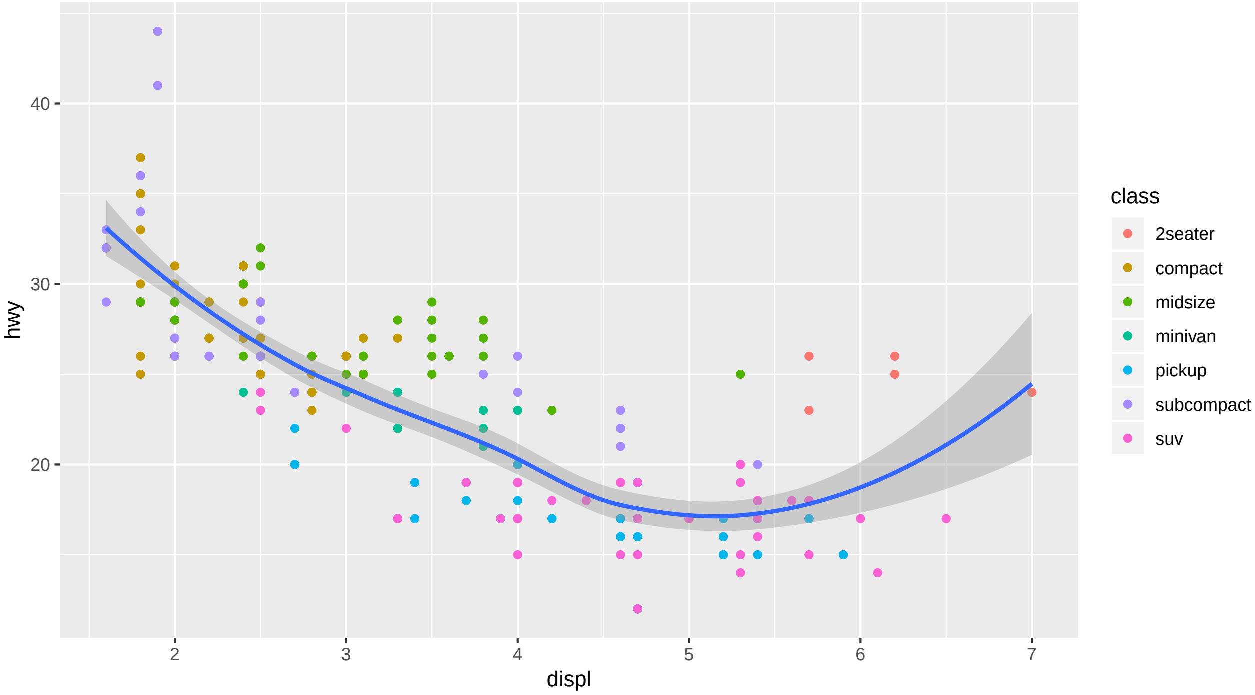
```r
# Global Mappings
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() + geom_smooth()
```

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



```
# Local Mappings
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping = aes(color = class)) +
  geom_smooth()
```

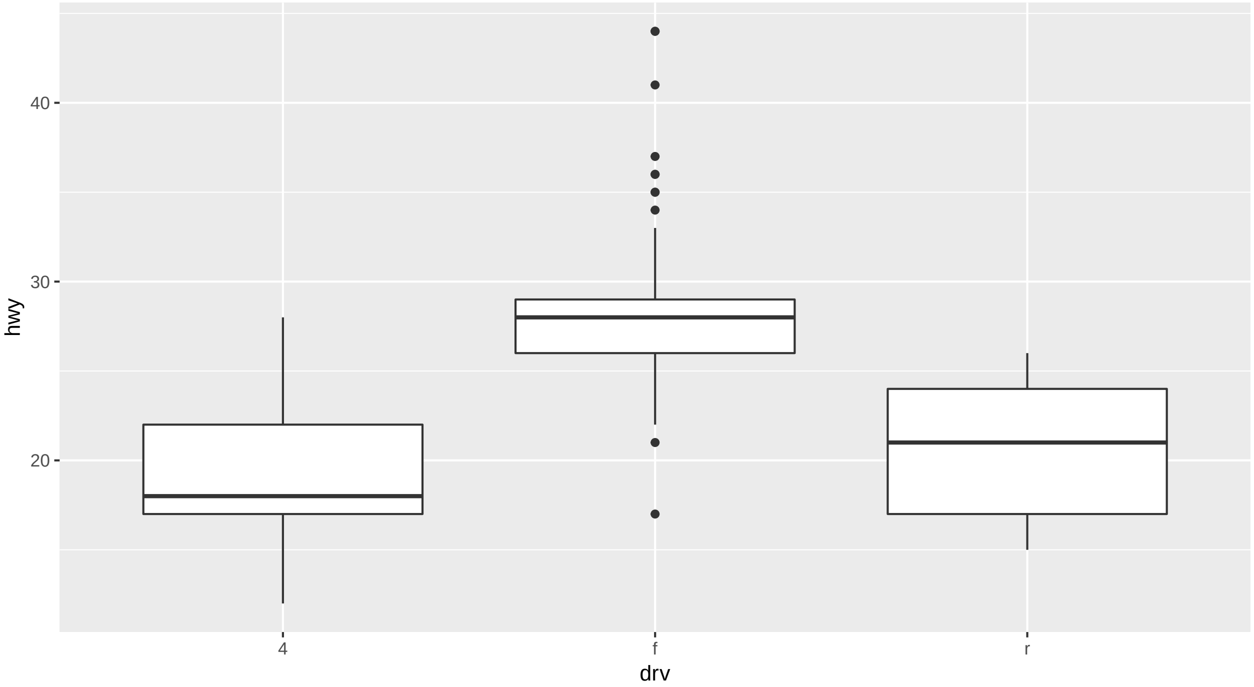
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Exercises-4

Q1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart? Ans. geom_boxplot(), geom_histogram()

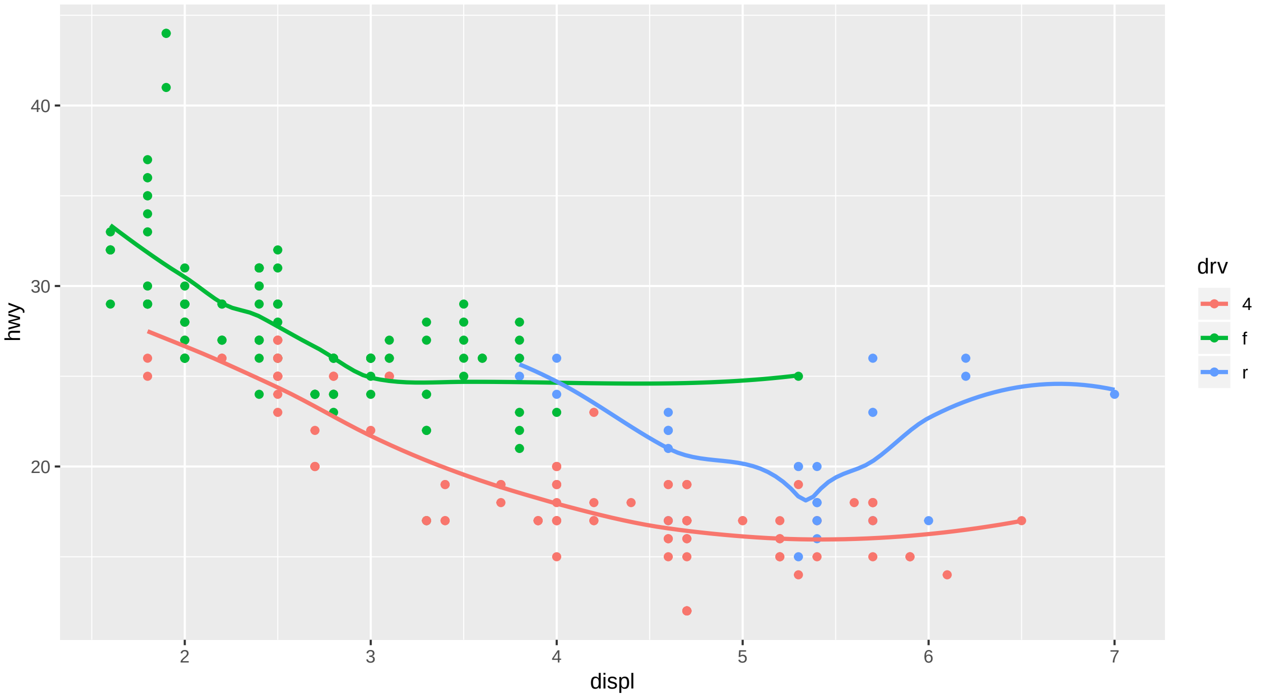
```
ggplot(data = mpg, mapping = aes(x = drv, y = hwy)) + geom_boxplot()
```



Q2. Recreate the plots

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

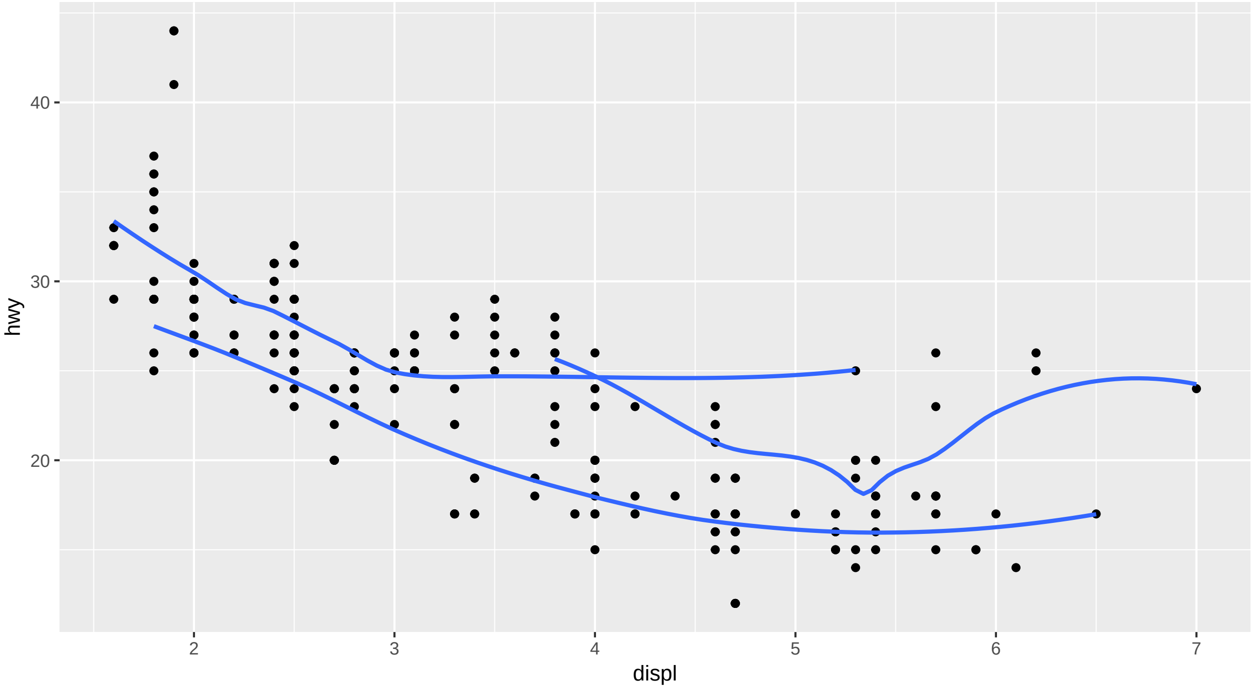
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Q3. Recreate the plots

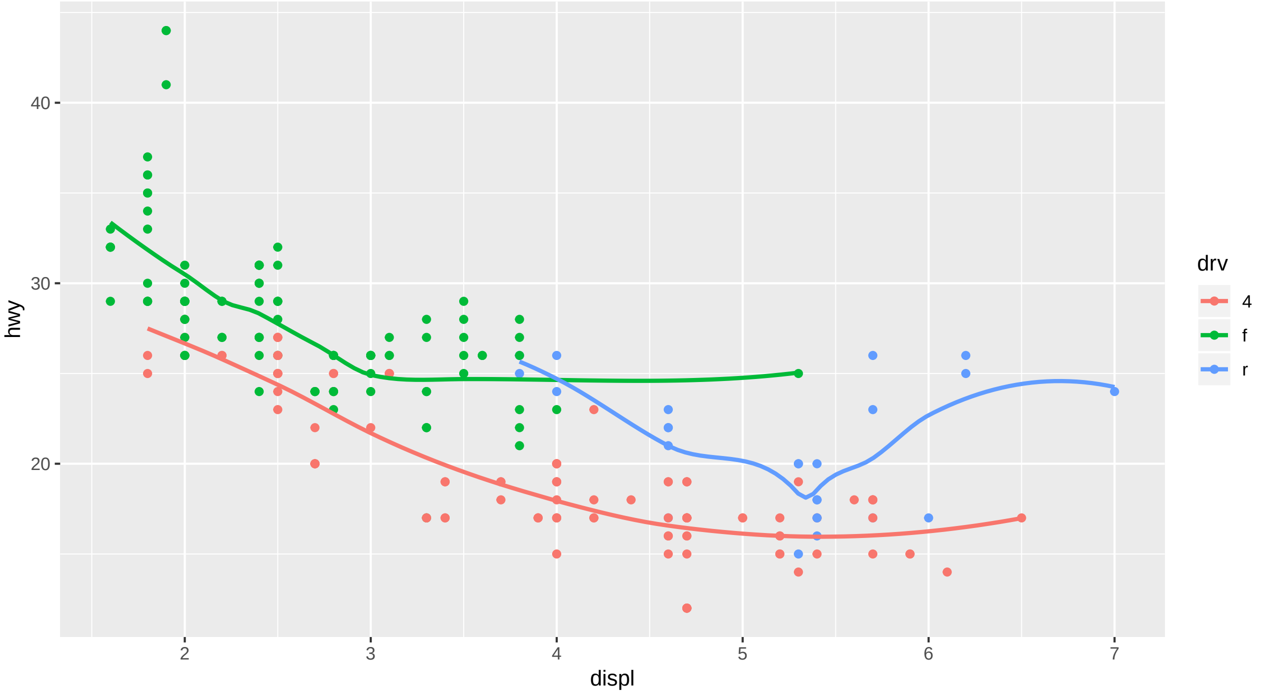
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
        geom_point() +
        geom_smooth(mapping = aes(group = drv), se = FALSE)
```

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



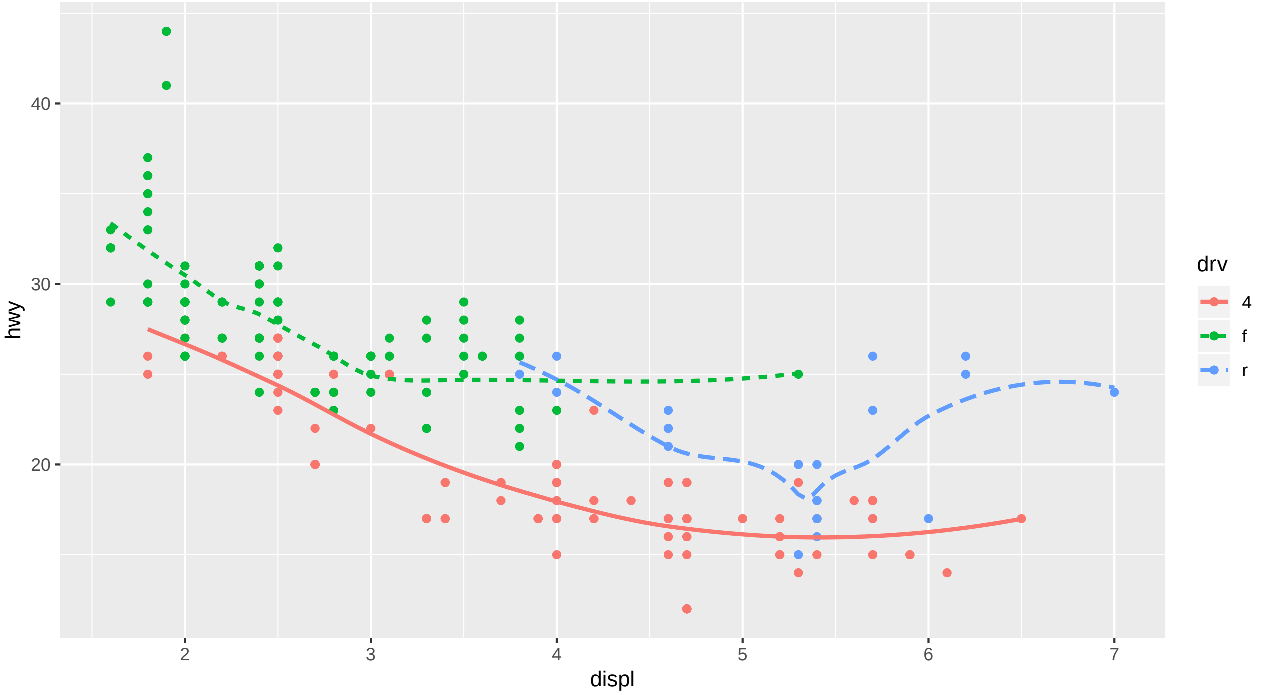
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
        geom_point() +
        geom_smooth(mapping = aes(group = drv), se = FALSE)
```

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



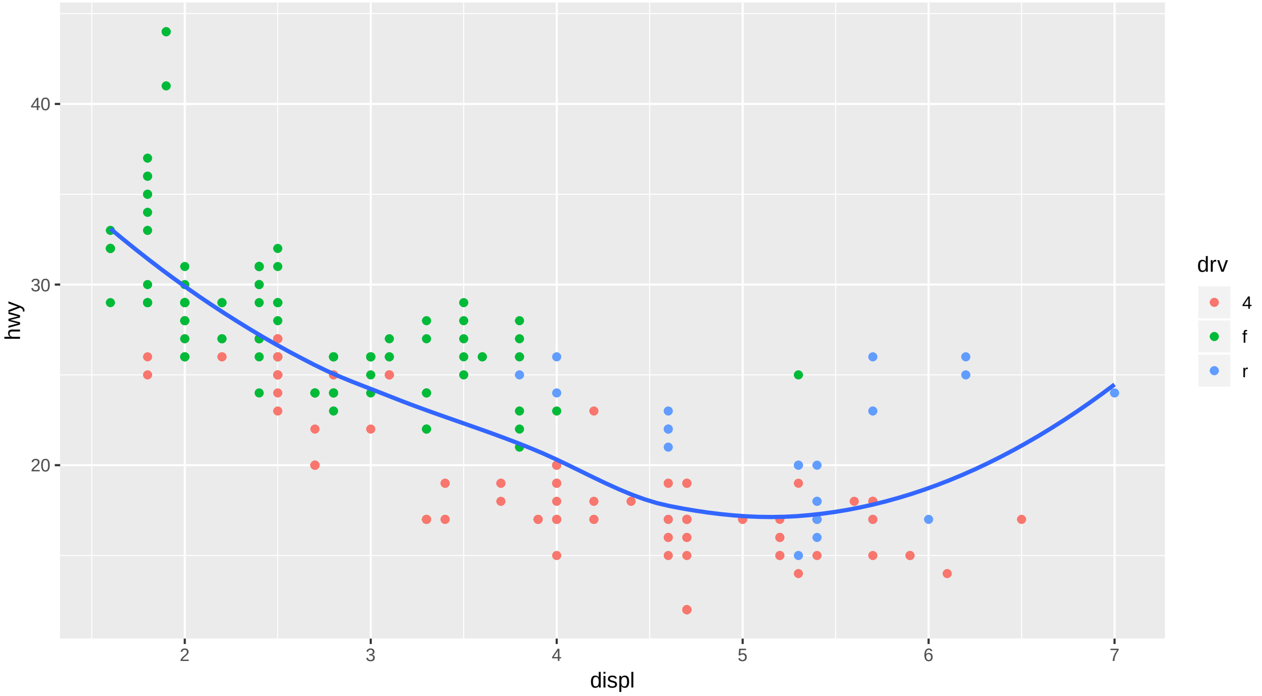
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
        geom_point() +
        geom_smooth(mapping = aes(group = drv, linetype = drv), se = FALSE)
```

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

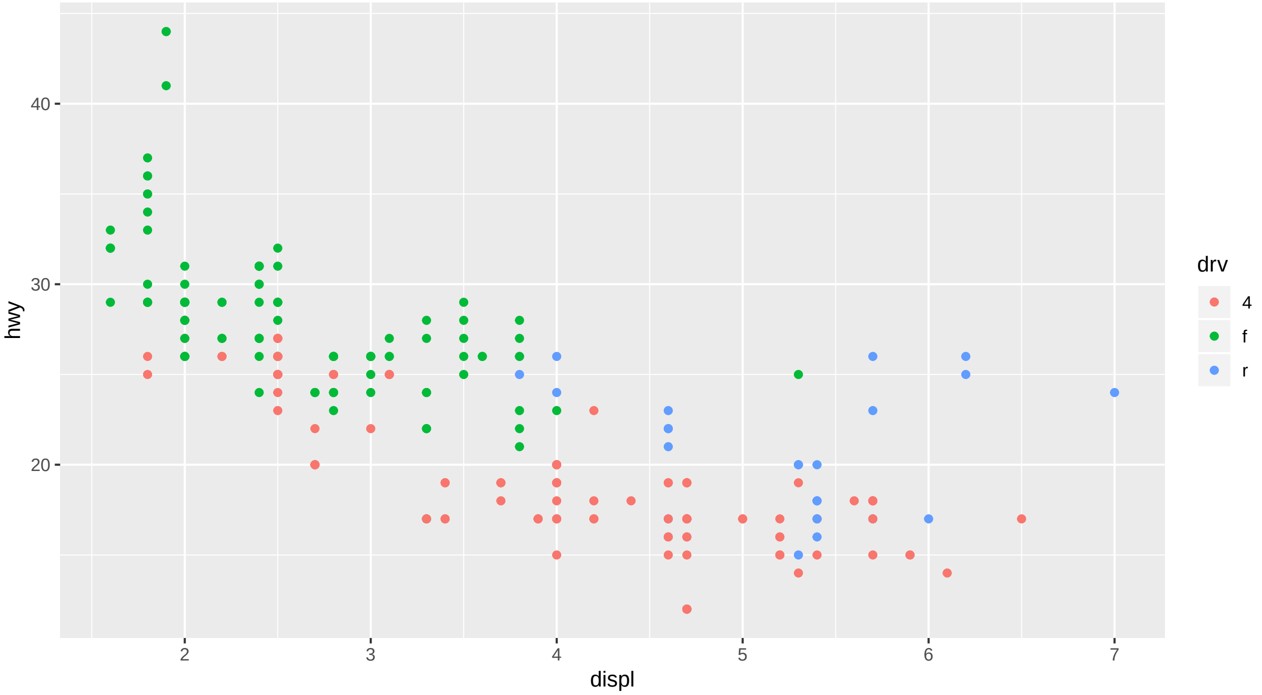


```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
        geom_point(mapping = aes(color = drv)) +
        geom_smooth(se = FALSE)
```

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



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

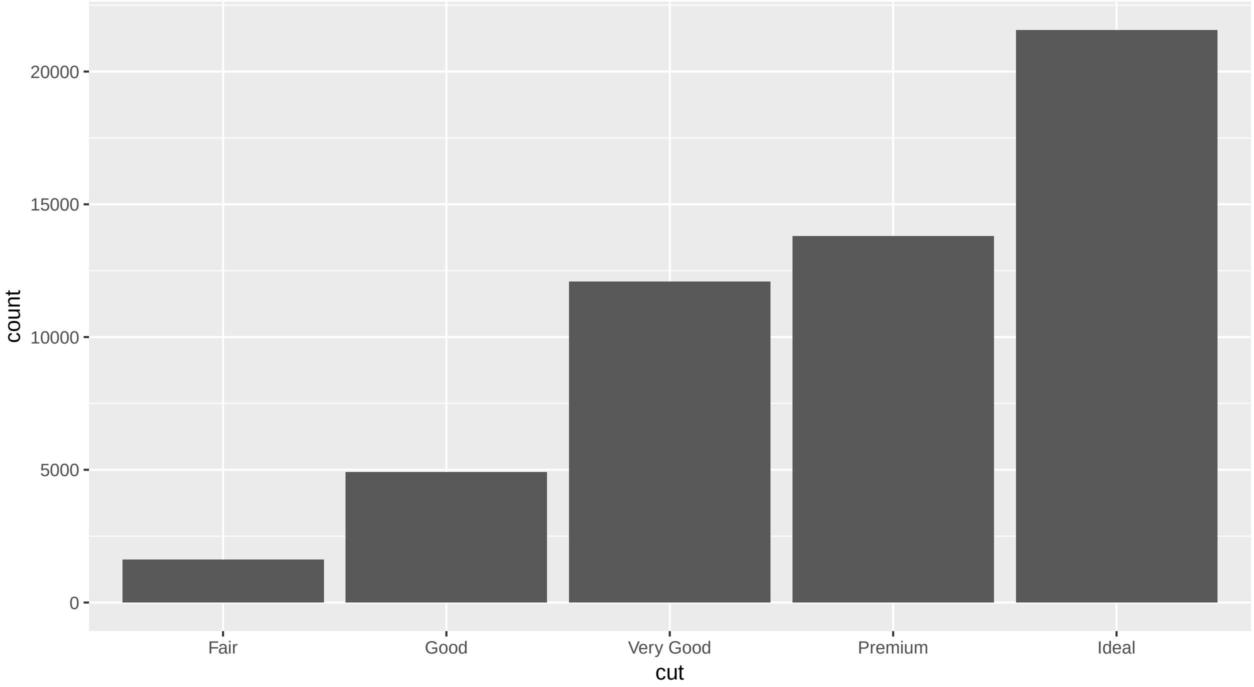


Statistical transformations

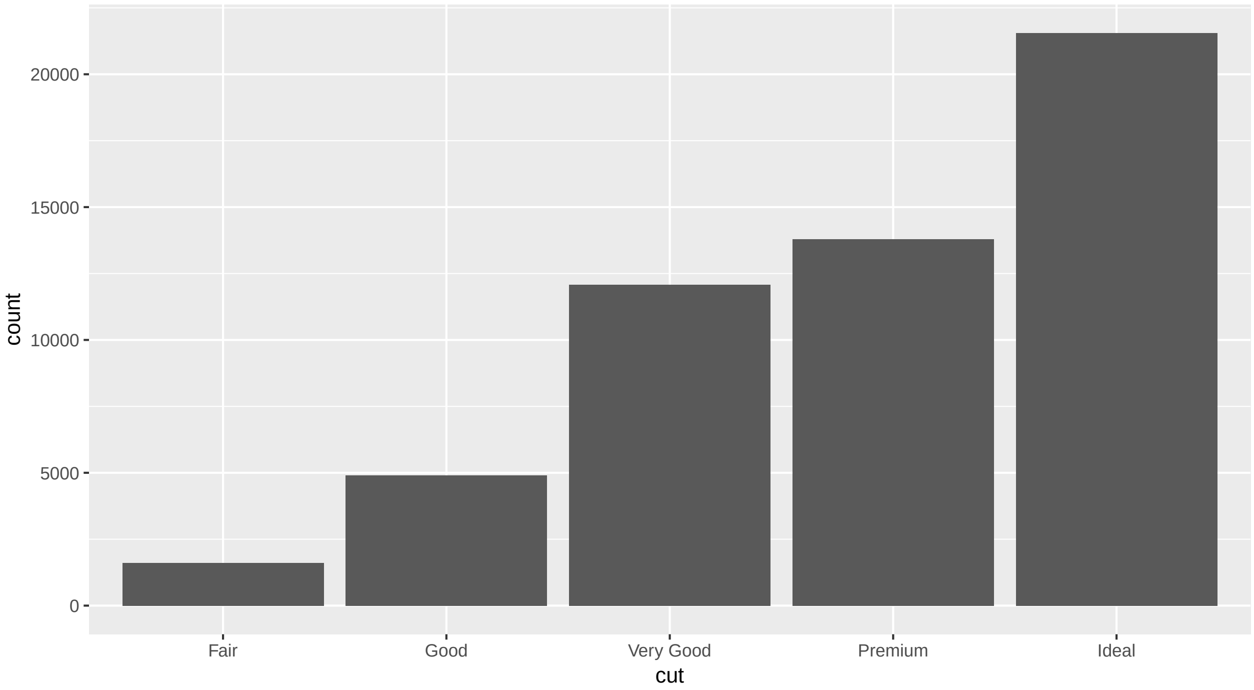
```
# Bar Charts
# Using the diamonds dataset
attach(diamonds)
nrow(diamonds)
```

```
## [1] 53940
```

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```

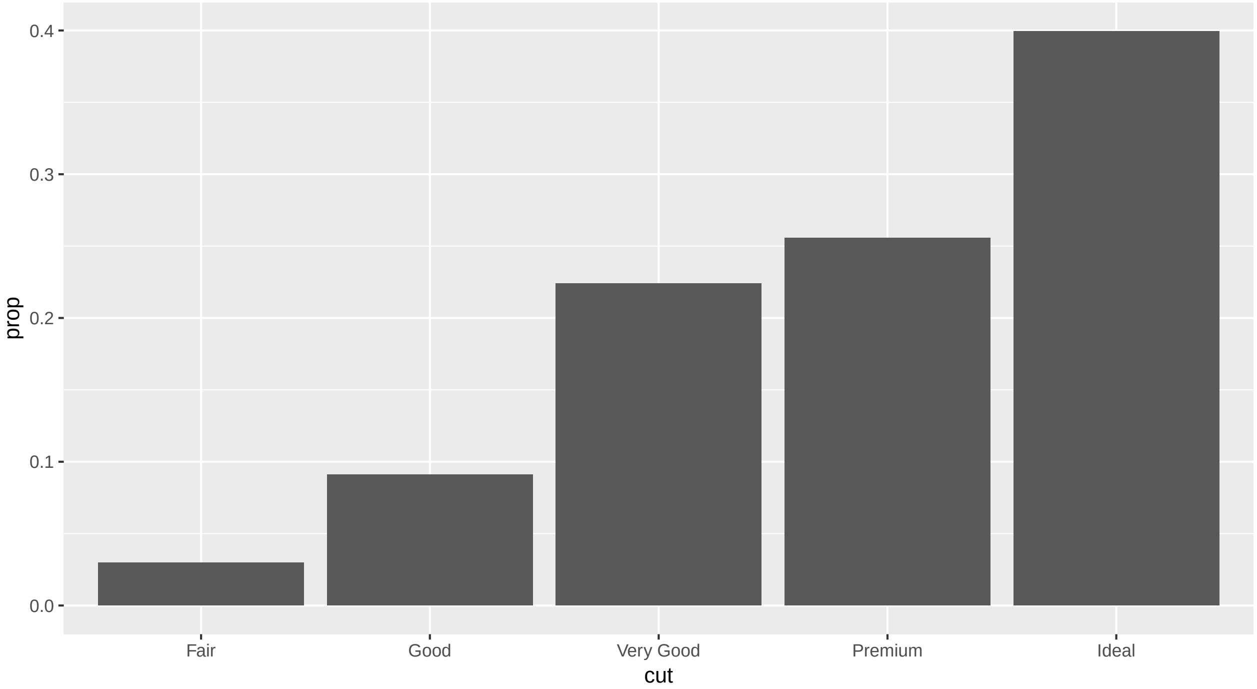


```
# Using stat_count() instead of geom_bar()
ggplot(data = diamonds) +
  stat_count(mapping = aes(x = cut))
```

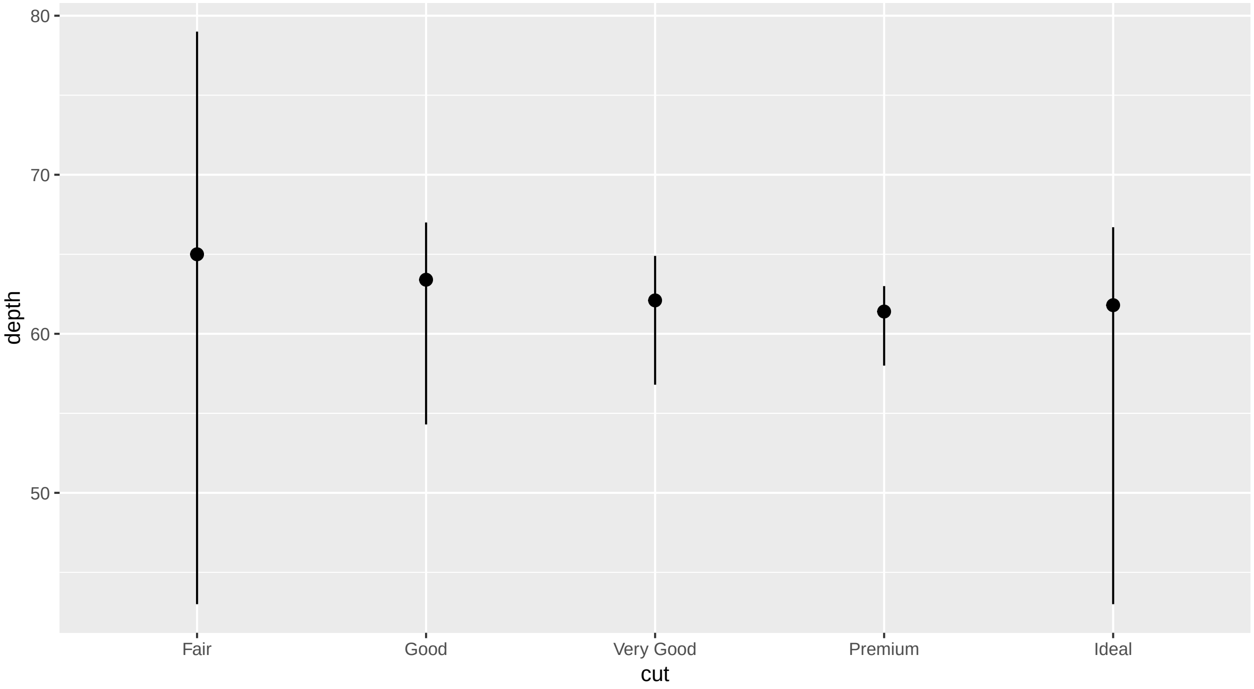


Every geom function has a stat equivalent and vice-versa that could be used to produce the same plot in two different ways.

```
# Displaying Proportion instead of frequency
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, y = ..prop.., group = 1))
```



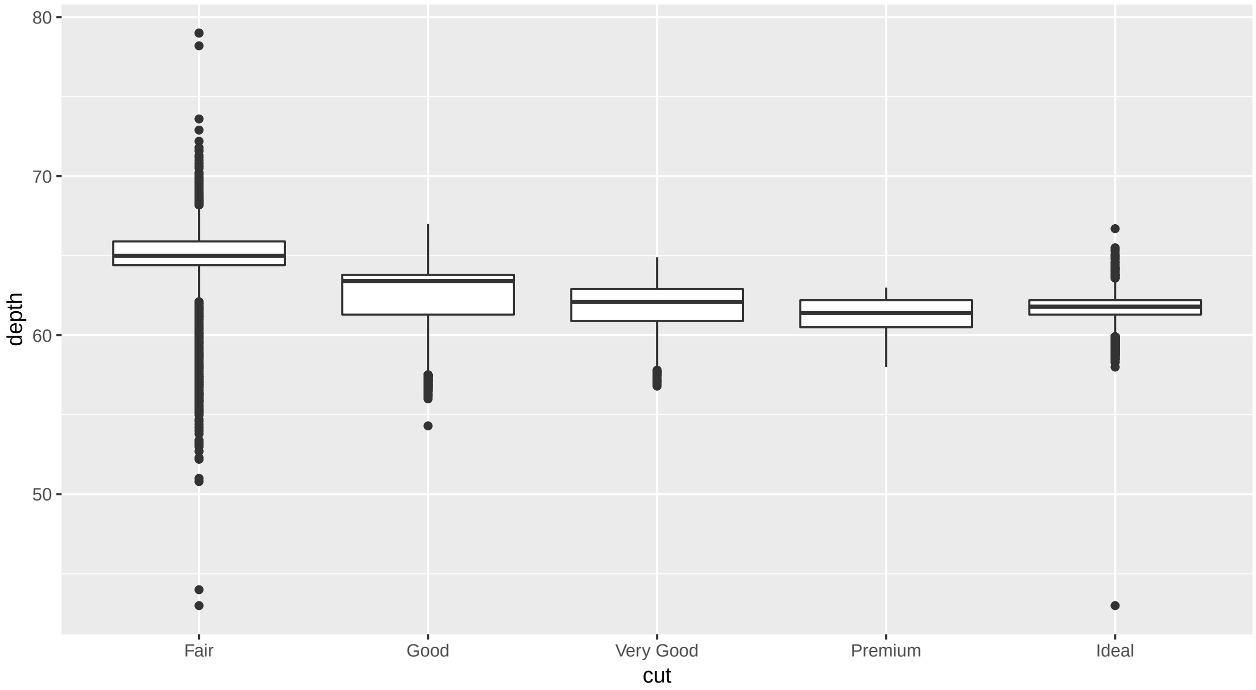
```
ggplot(data = diamonds) +
  stat_summary(
    mapping = aes(x = cut, y = depth),
    fun.ymin = min,
    fun.ymax = max,
    fun.y = median
  )
```



Exercises-5

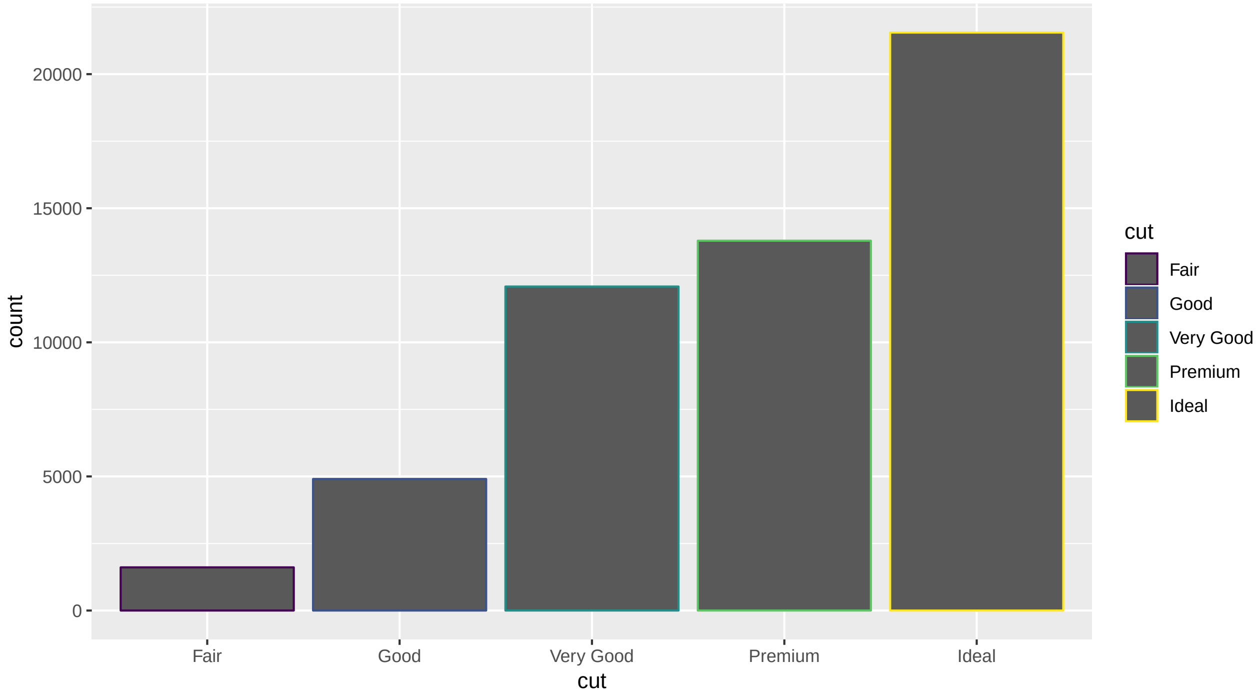
Q1. What is the default geom associated with stat_summary()? How could you rewrite the previous plot to use that geom function instead of the stat function?

```
ggplot(data = diamonds) + geom_boxplot(mapping = aes(x = cut, y = depth))
```

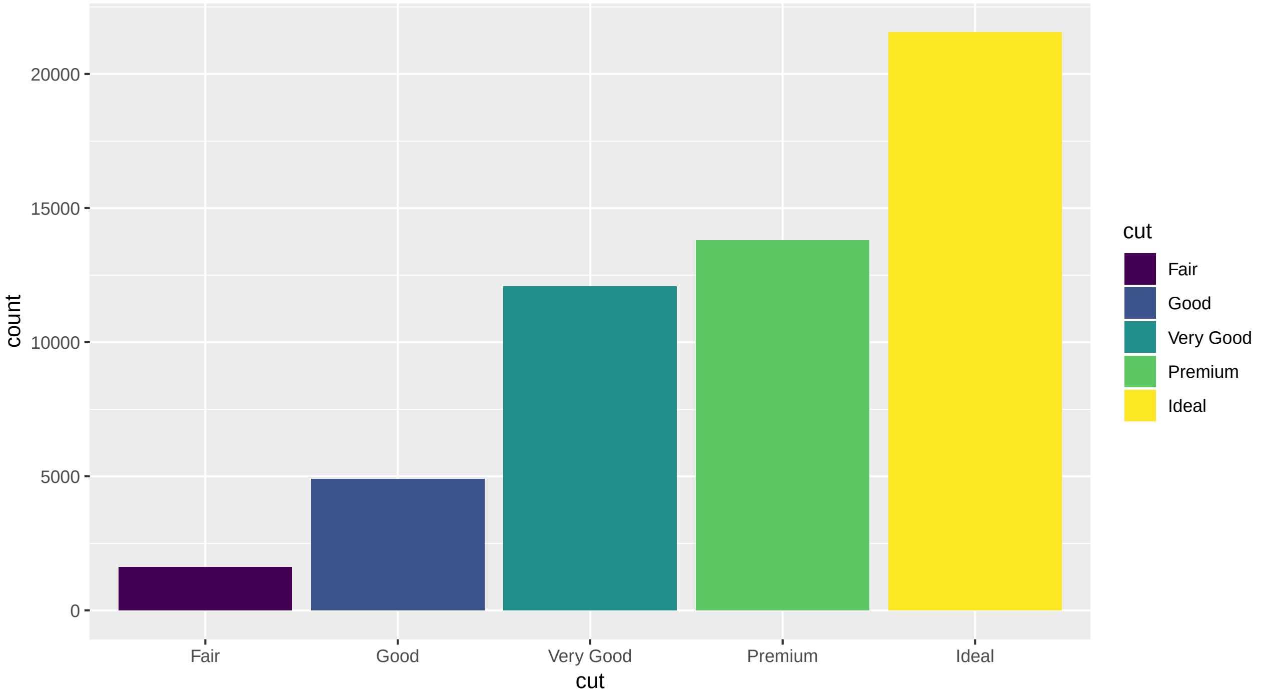


Position Adjustments

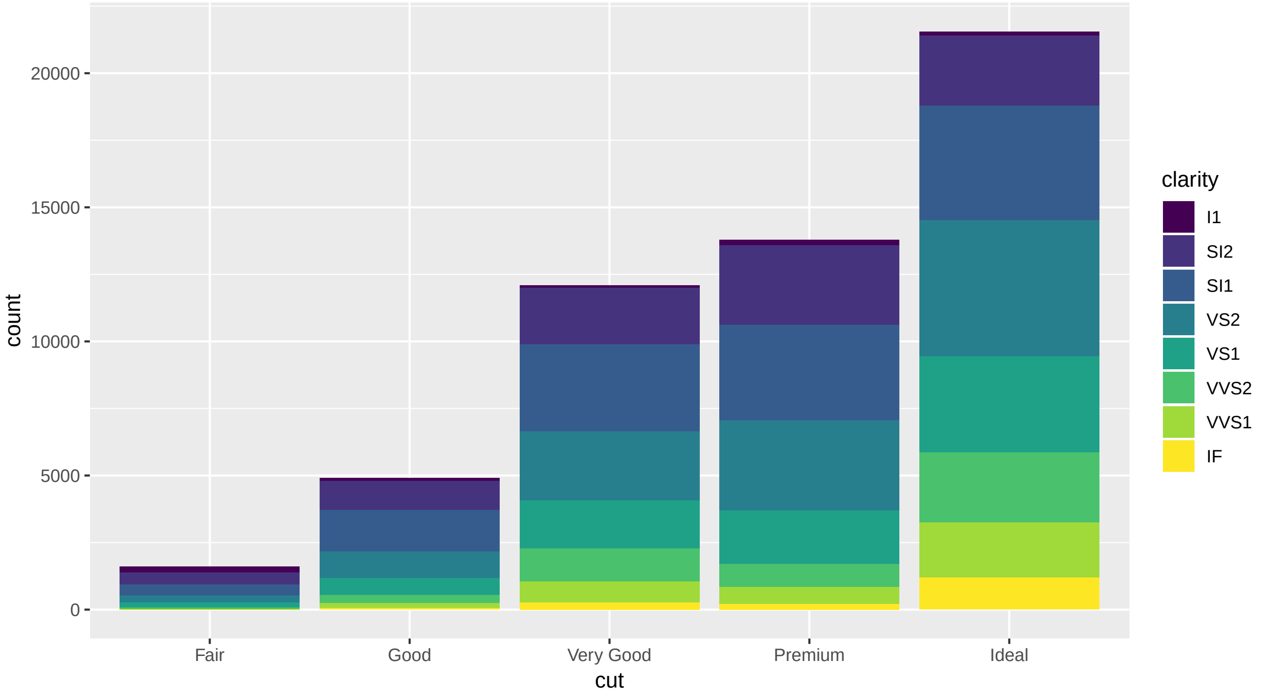
```r
# Using Colour aesthetic
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, colour = cut))
```



```r
# Using fill aesthetic
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = cut))
```



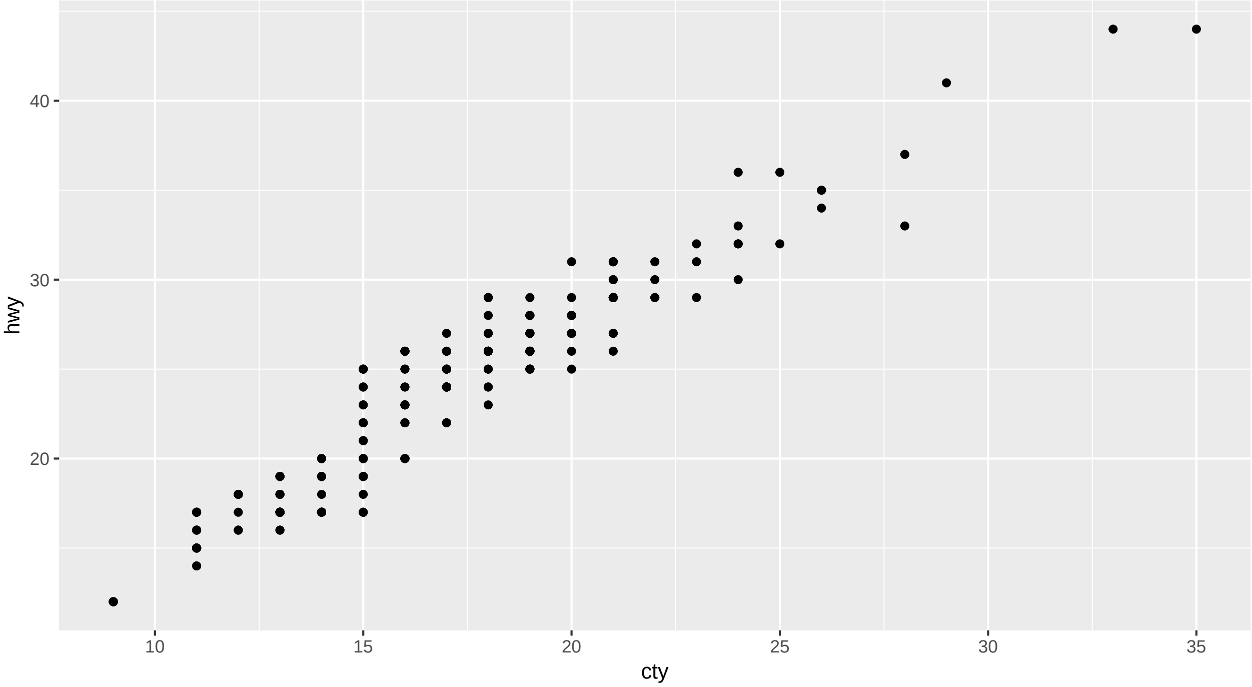
```r
# Stacked Bar Chart
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity))
```



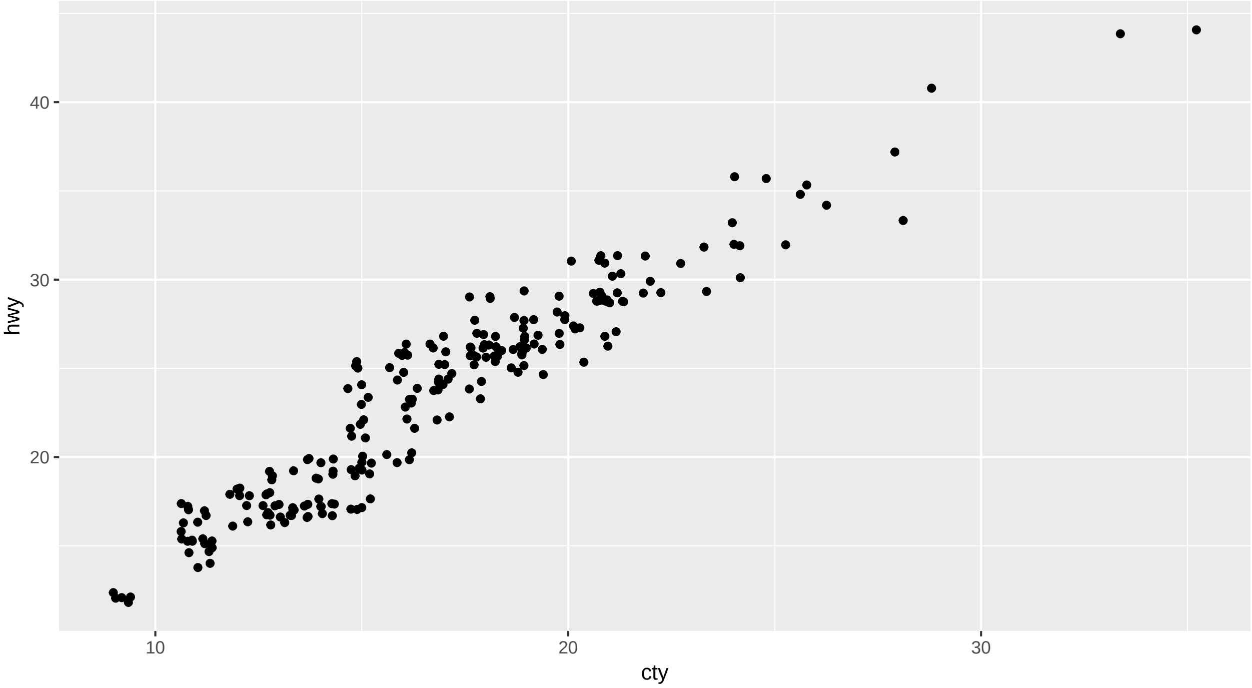
Exercises-6

Q.What is the problem with this plot? How could you improve it?

```
# Original Plot
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point()
```



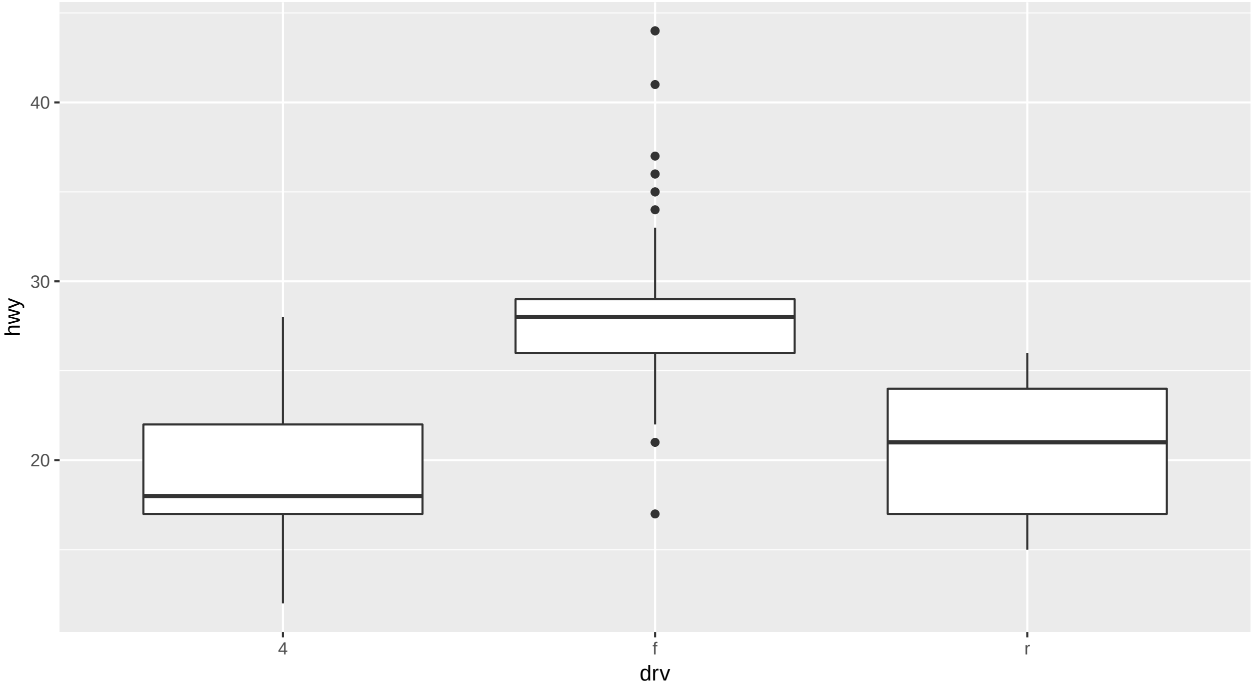
```
# Adding jitter
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point(position = "jitter")
```



The above plot can also be plotted using the geom_jitter function. This jitter avoids overplotting and adds a small amount on random noise to each point.

Q. What's the default position adjustment for geom_boxplot()? Create a visualisation of the mpg dataset that demonstrates it.

```r
ggplot(data = mpg, mapping = aes(x = drv, y = hwy)) + geom_boxplot()
```



```r
# Grammar of Graphics Template
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
     mapping = aes(<MAPPINGS>),
     stat = <STAT>,
     position = <POSITION>
  ) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```