Estudo R for Data Science - Cap 2

Bruno de A. Machado

9/20/2020

Visualização de Dados

R for Data Science

Neste segundo capítulo os autores nos ensinam como visualizar dados utilizando `ggplot2` e apresenta o conceito de **gramática dos gráficos** e como criar plots em camadas.

Importando o pacote tidyverse para que nos fornece o `ggplot2` entre outros :

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

Usando o primeiro gráfico para responder algumas perguntas :

1. Carros com motores maiores consomem mais combustível que carros com motores menores ?

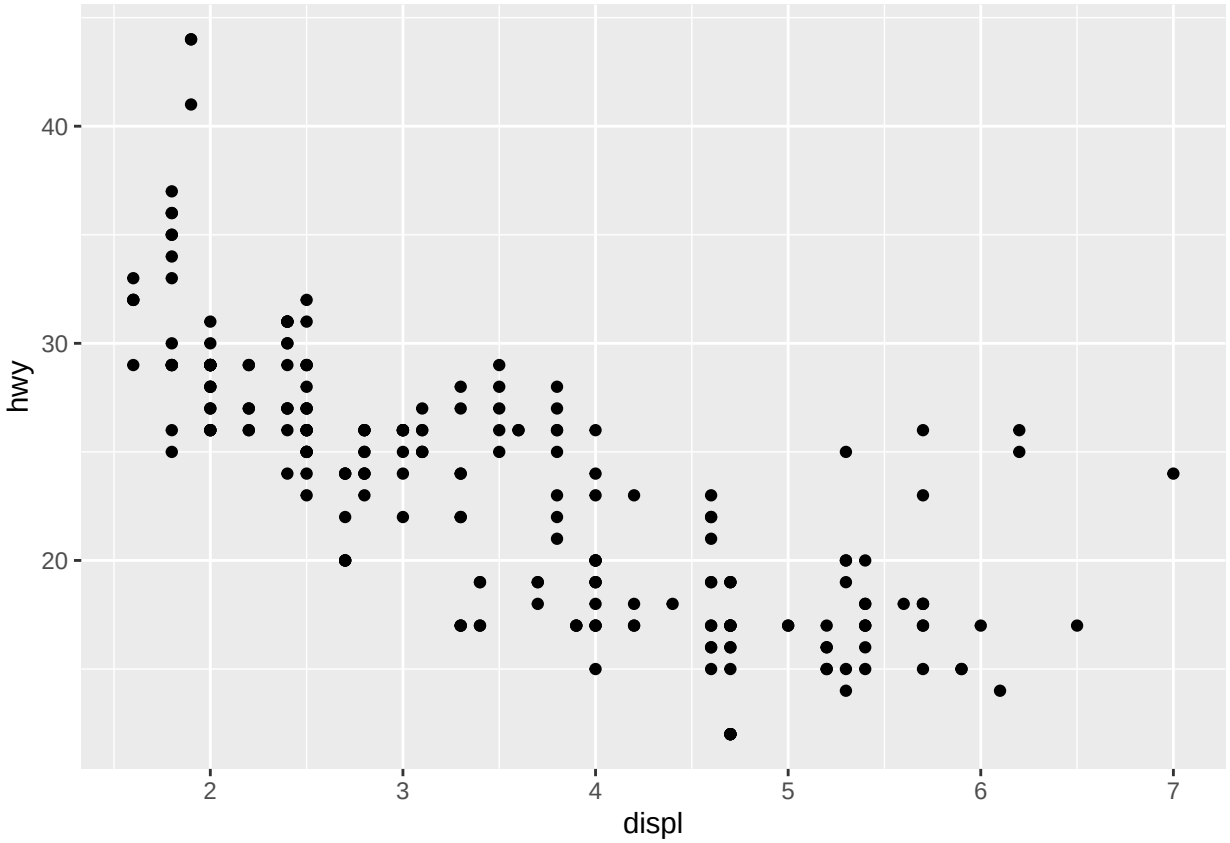
- `displ` : engine size em litros
- `hwy` : eficiencia em miles per gallon (mpg), um carro com baixa eficiencia consome mais combustivel .

  Resposta: Podemos observar essa relação que é de conhecimento comum que carros com menor eficiência com relação a combustível `hwy` tem os maiores motores `displ`

2. Qual o relacionamento entre `engine size (displ)` e `efficience (hwy)` ?

Resposta: Relaçao negativa entre as duas variáveis, ou seja, carros com engines maiores são menos eficientes e usam mais combustível

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



Exercícios 3.2.4: 1. Run ggplot(data = mpg). What do you see?

Resposta: essa primeira parte só monta a área do gráfico

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

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

Resposta: O dataframe `mpg` tem 11 colunas e 234 observações ou linhas

```
ncol(mpg)
```

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

```
nrow(mpg)
```

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

Outra forma de observar é utilizando a função `glimpse()` que oferece além do número de linhas e colunas, tipos das variáveis e exemplos de observações:

```
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,...
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, ...
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, ...
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "a...
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4",...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25...
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p",...
```

```
## $ class        <chr> "compact", "compact", "compact", "compact", "compact",...
```

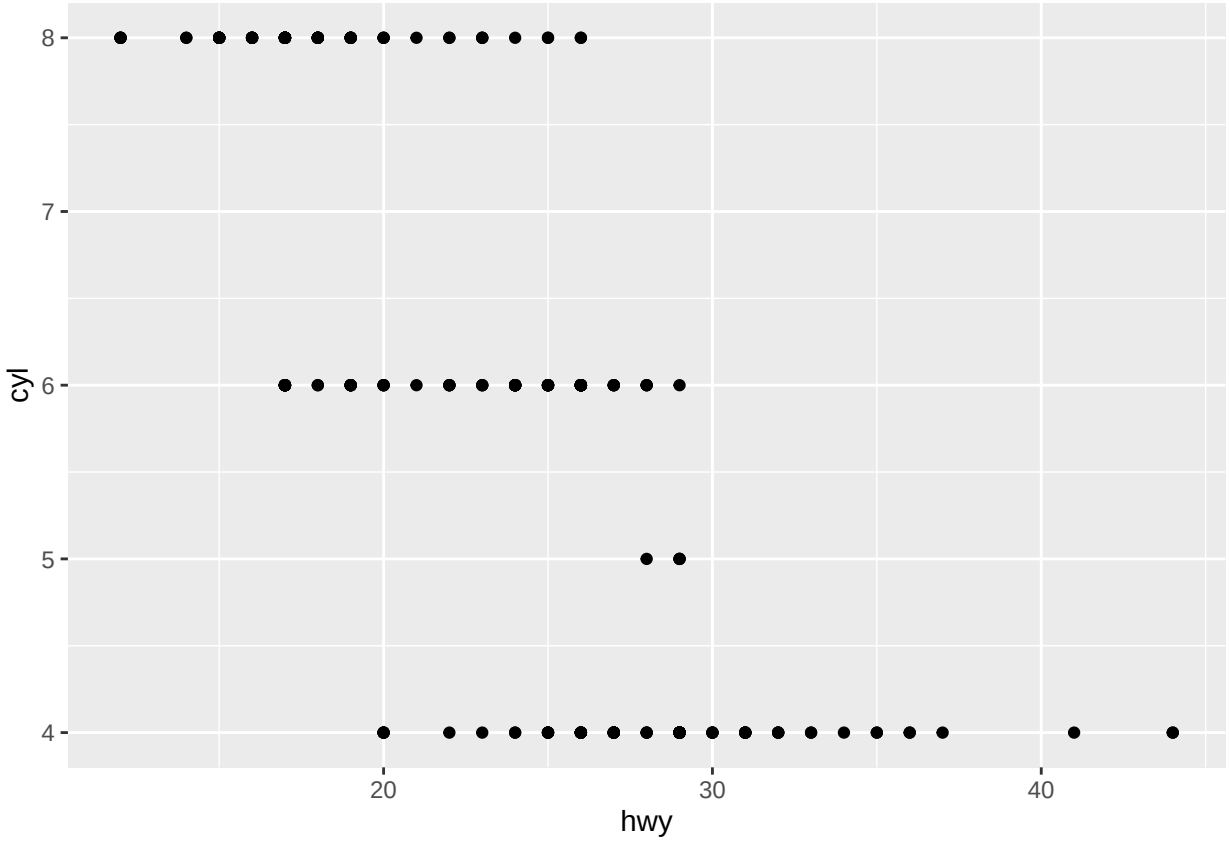
3. What does the drv variable describe? Read the help for ?mpg to find out.

Resposta: `drv`: the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd

```
?mpg
```

4. Make a scatterplot of hwy vs cyl.

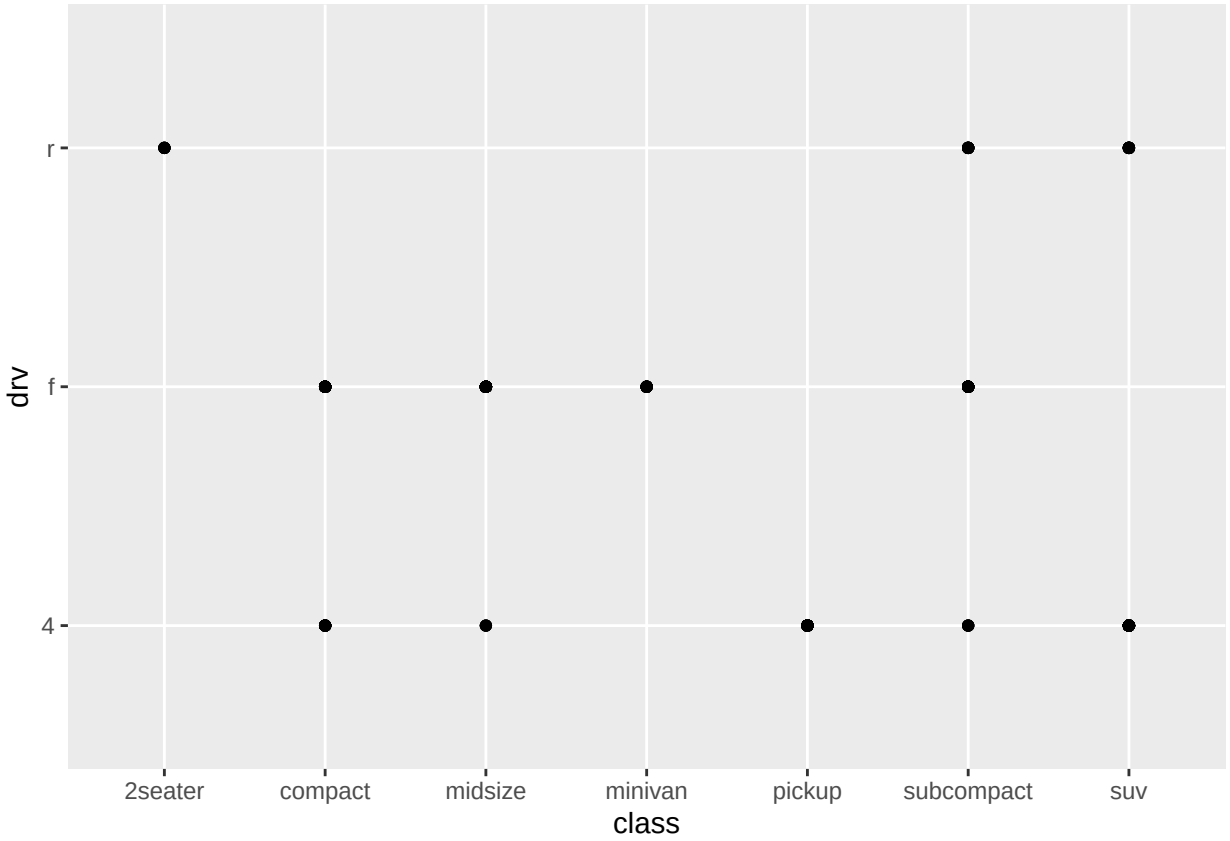
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = hwy, y= cyl))
```



5. What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

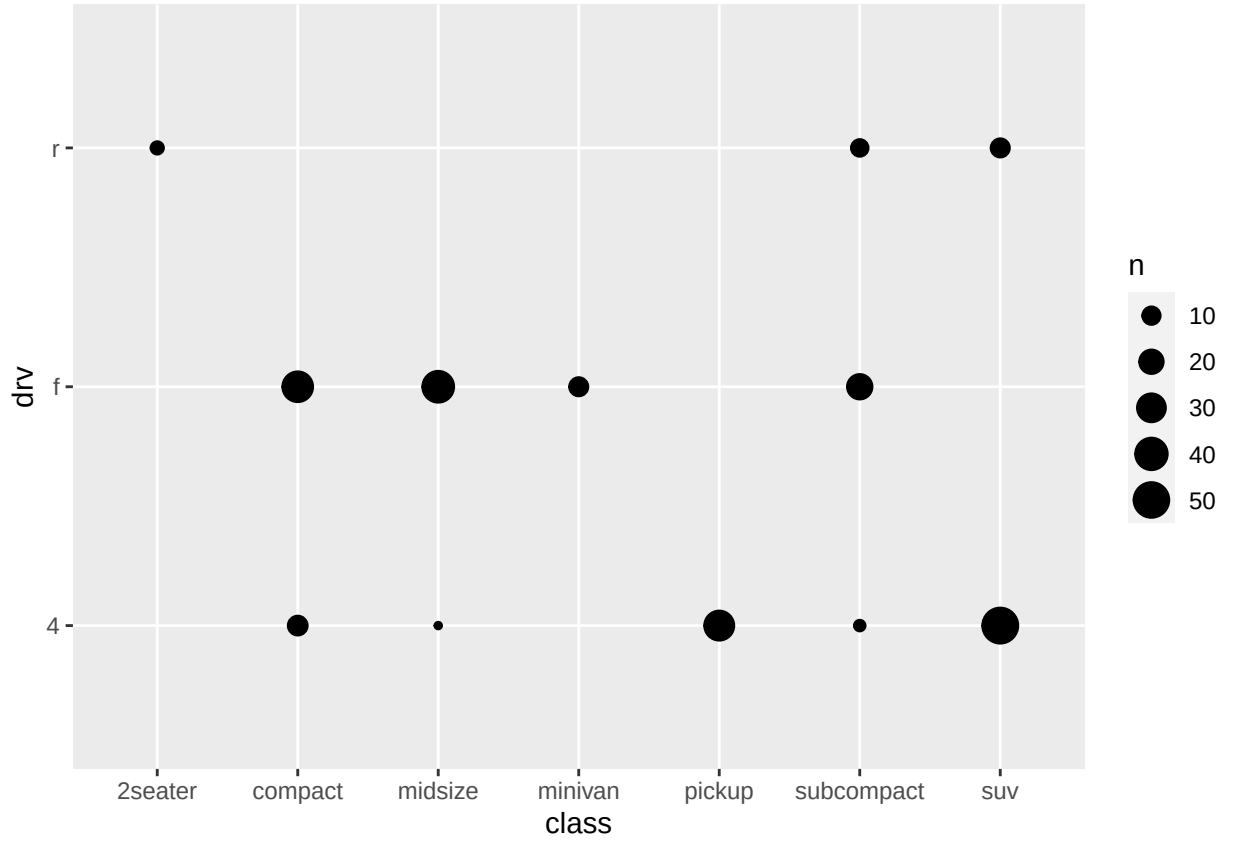
Resposta : Utilizando a função `glimpse` percebemos que `drv e calss` são do tipo `<chr>` que significa categórico, ou seja plotar um scatterplot de dados categóricos não fornece nenhuma informação útil.

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



Na documentação do ggplot tem a função `geom_count`, plotando esse gráfico ele nos fornece algumas informações mais interessantes

```
ggplot(data = mpg, aes(x = class, y = drv))+
  geom_count()
```

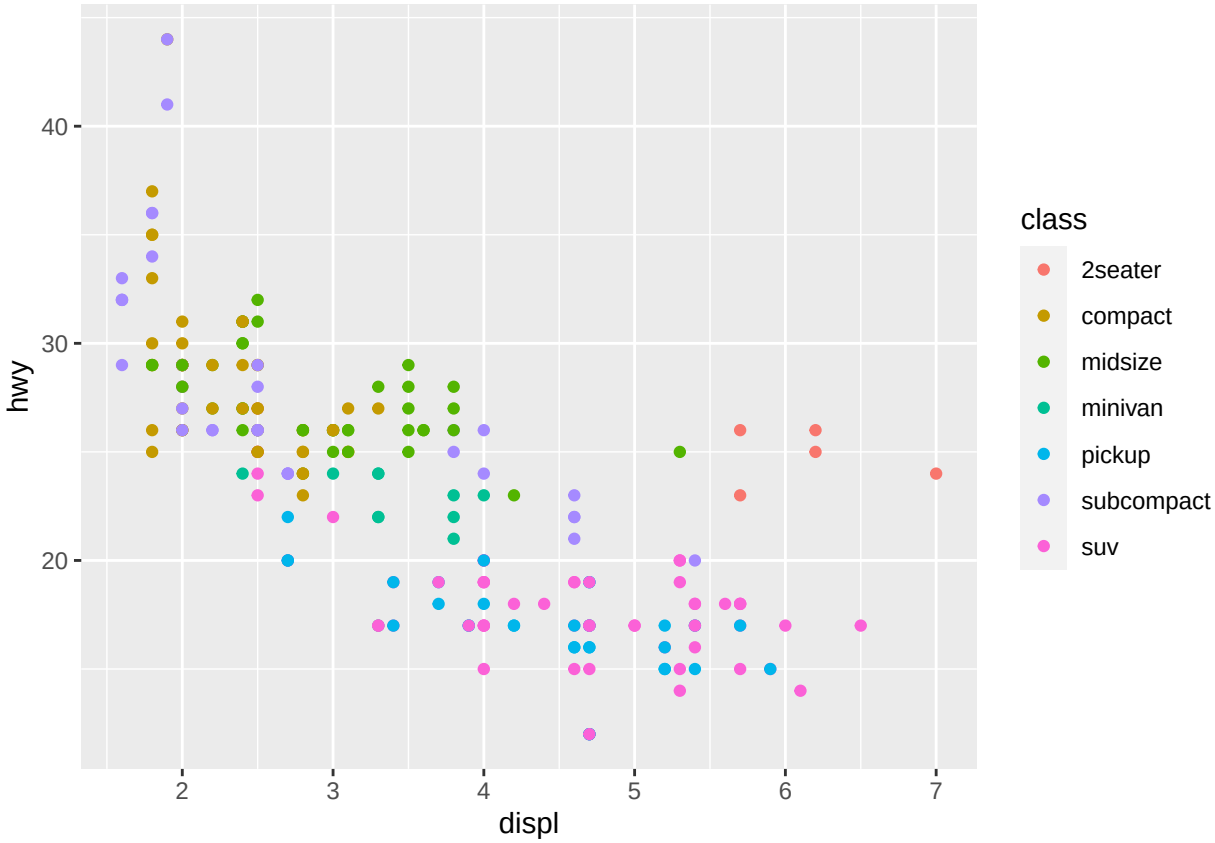


Aesthetic Mapping

Aesthetic é uma propriedade visual do objeto no plot, podemo adicionar funções que destacam o tamanho, shape ou cor.

Utilizando **aesthetic** `color` mapeando a classe o `ggplot` automaticamente cria uma legenda e atribuiu uma cor específica para cada tipo de classe, processo chamado de **scaling**

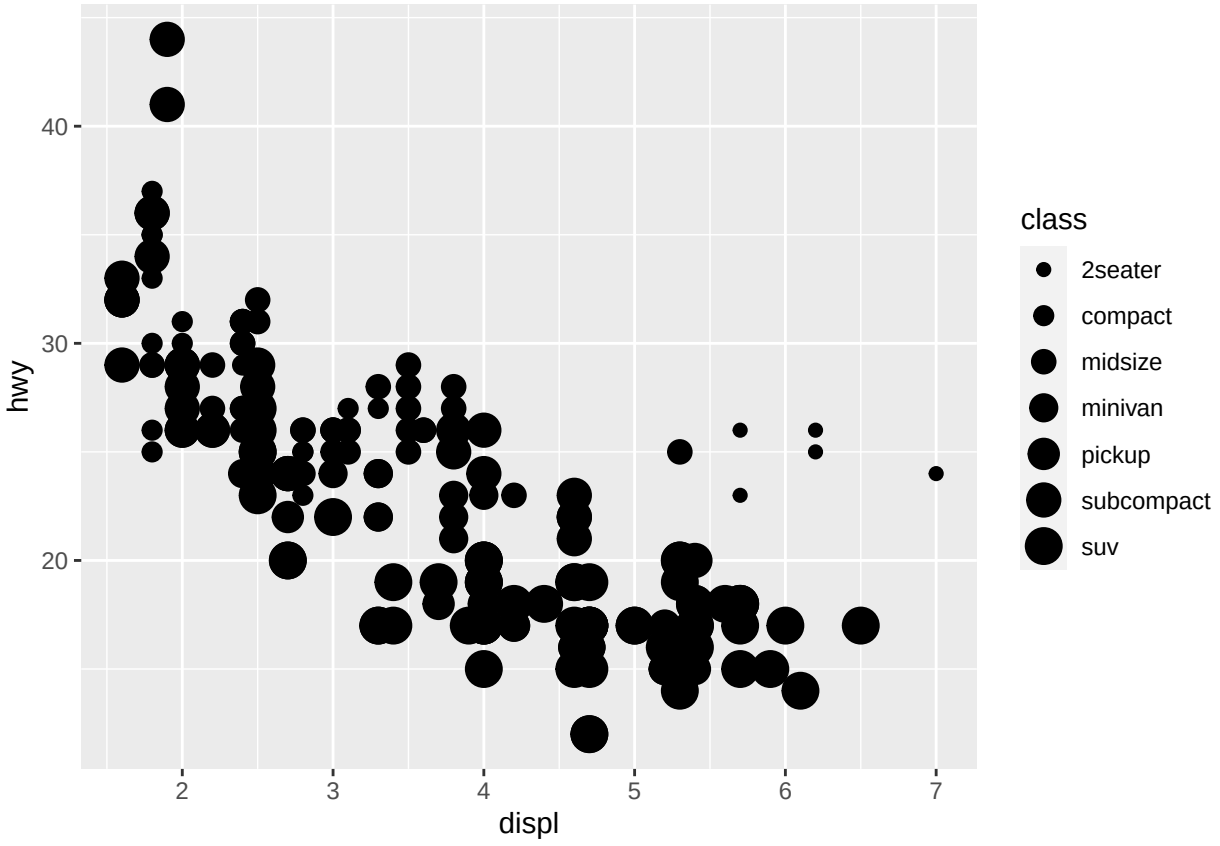
```
ggplot(data = mpg)+
  geom_point(mapping = aes(x= displ, y= hwy, color=class))
```



Além de `color` podemos utilizar `size`, outras opções são `alpha` e `shape` .

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

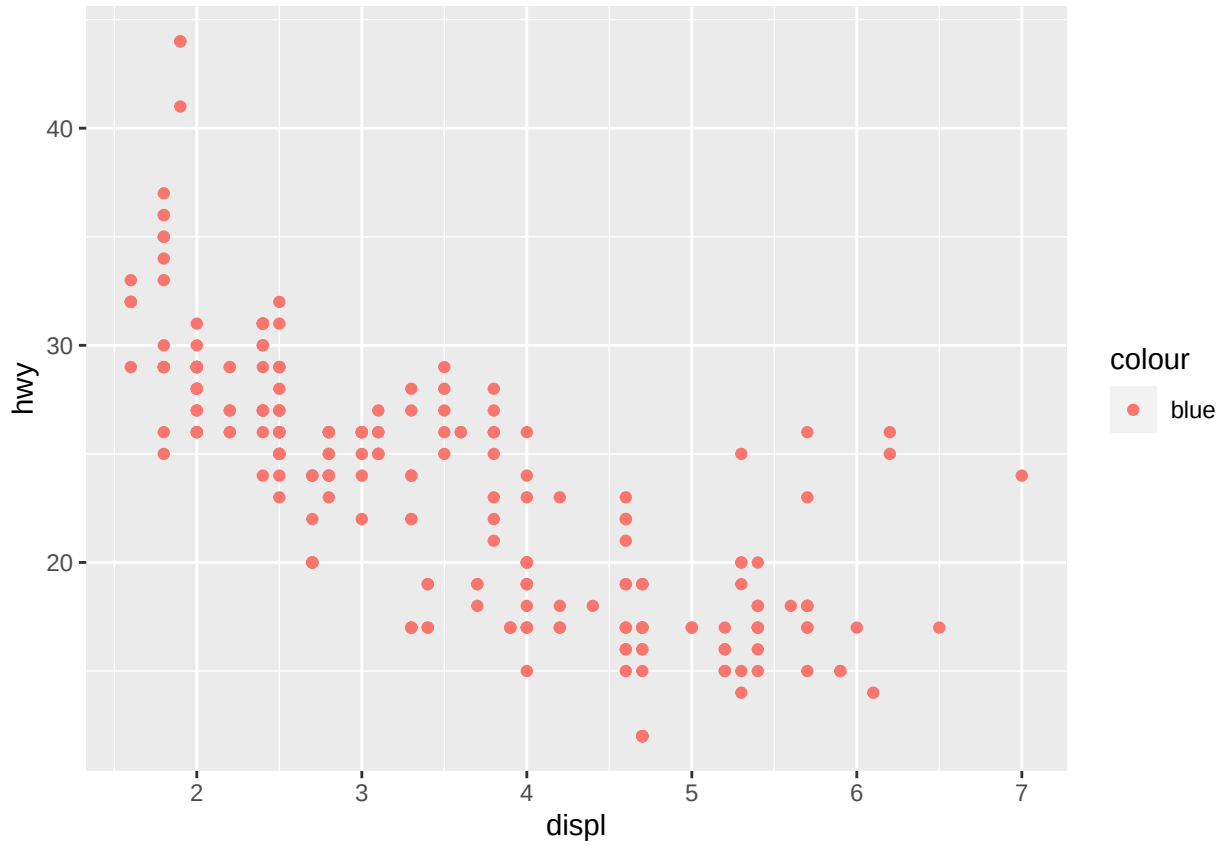
```
## Warning: Using size for a discrete variable is not advised.
```



Exercícios 3.3.1: 1. What's gone wrong with this code? Why are the points not blue?

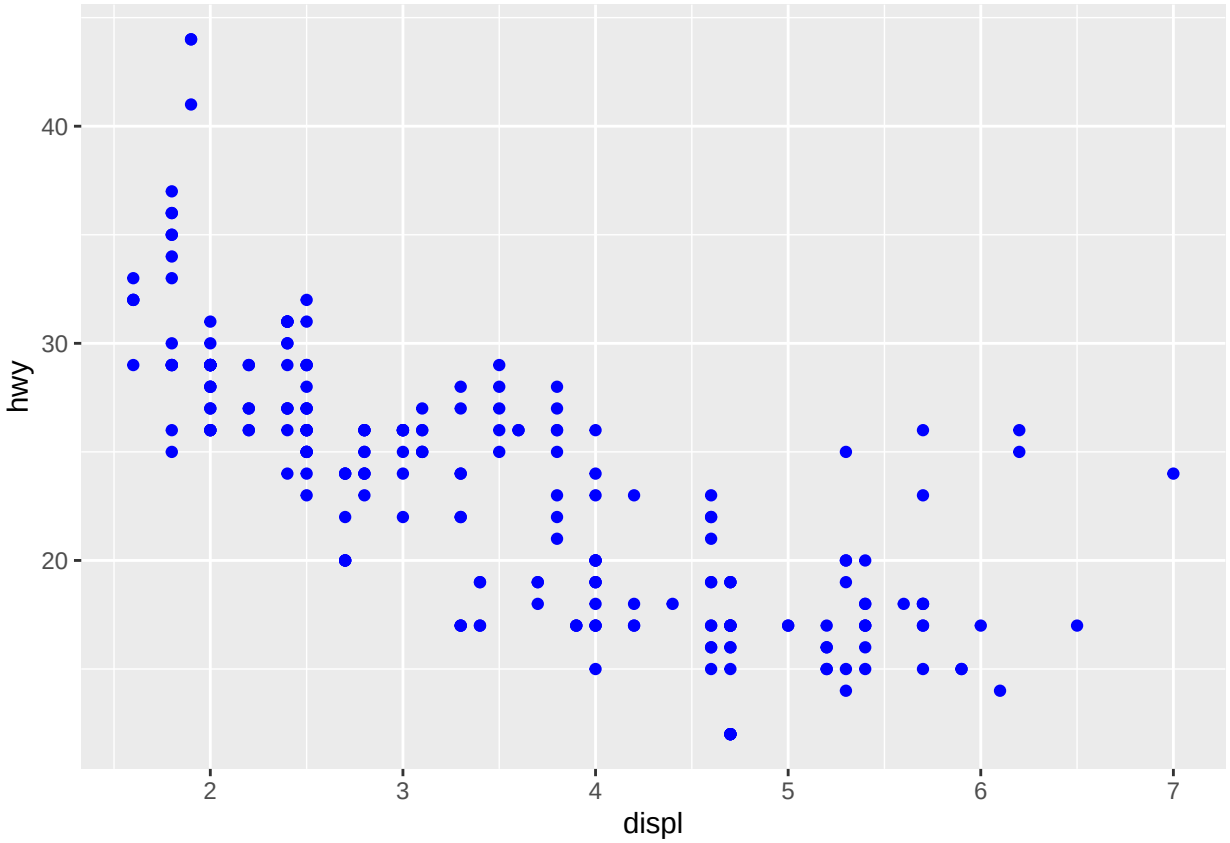
Resposta : `blue` é intepretado como uma variável por isso o ggplot não identificou a cor, para modificar a cor do gráfico pode-se adicionar color fora do `aes`

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



Fix: *Incluindo a cor fora do aes*

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



2. 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?

```
glimpse(mpg)
```

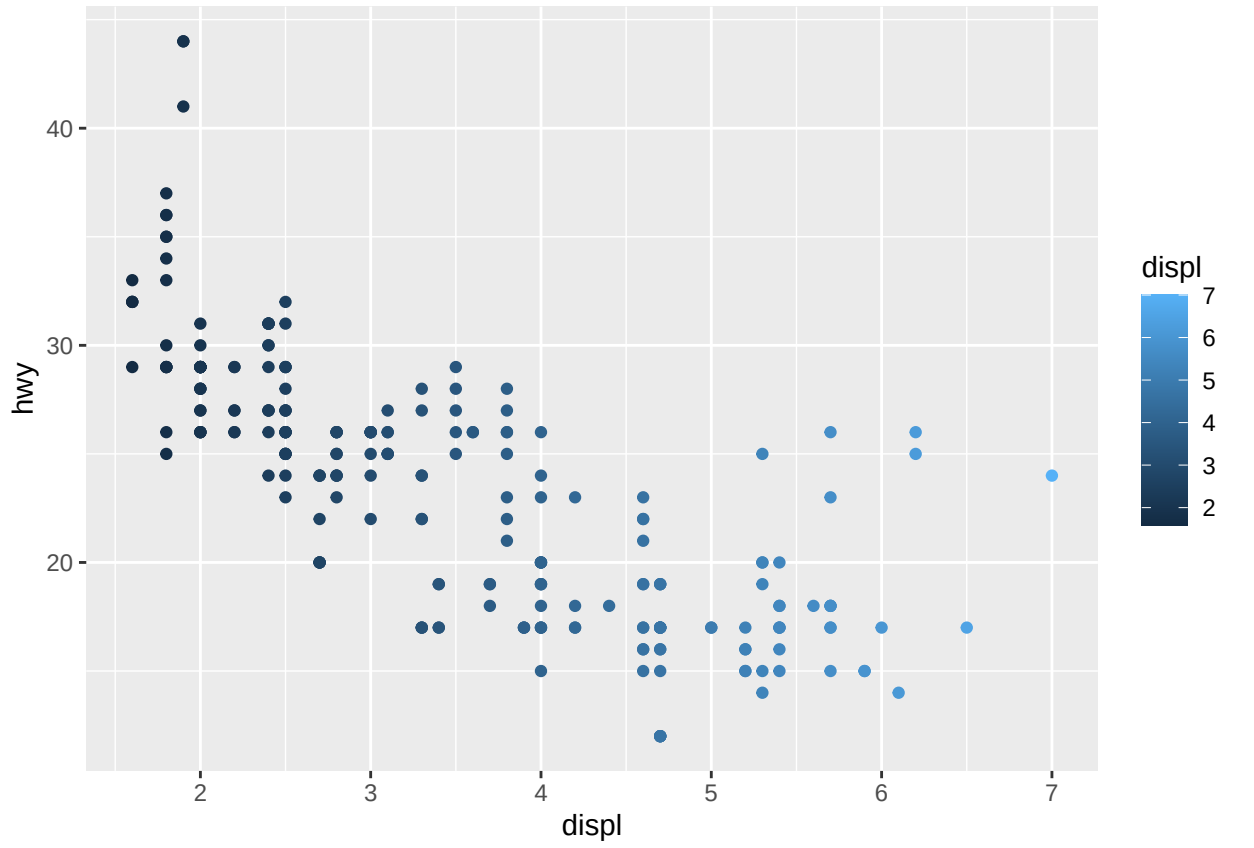
```
## Rows: 234
## Columns: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "audi"...
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 quattro"...
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0, 2.0,...
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, ...
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, ...
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "a...
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4",...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25...
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p",...
## $ class        <chr> "compact", "compact", "compact", "compact", "compact",...
```

Categórica : manufacturer, model, trans, drv, fl, class

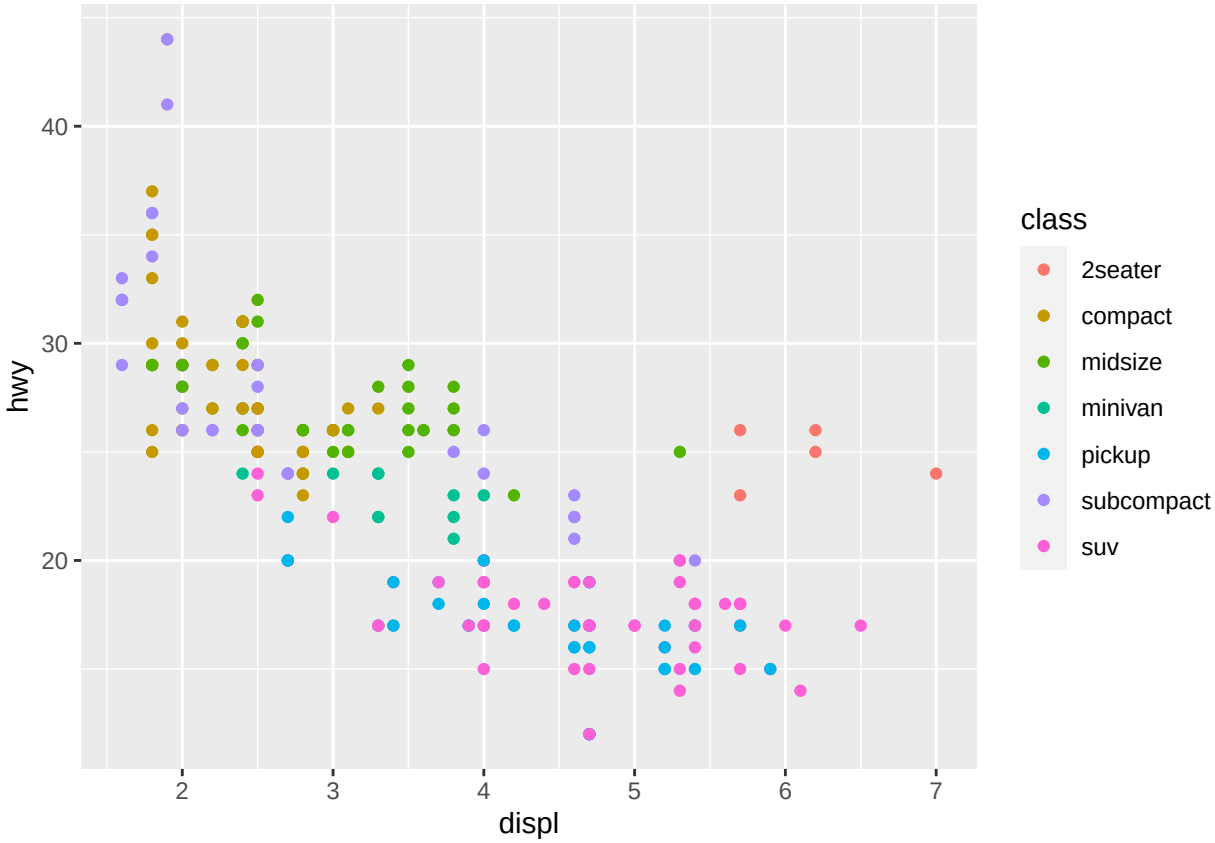
Contínua : displ, year, cyl, cty, hwy

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

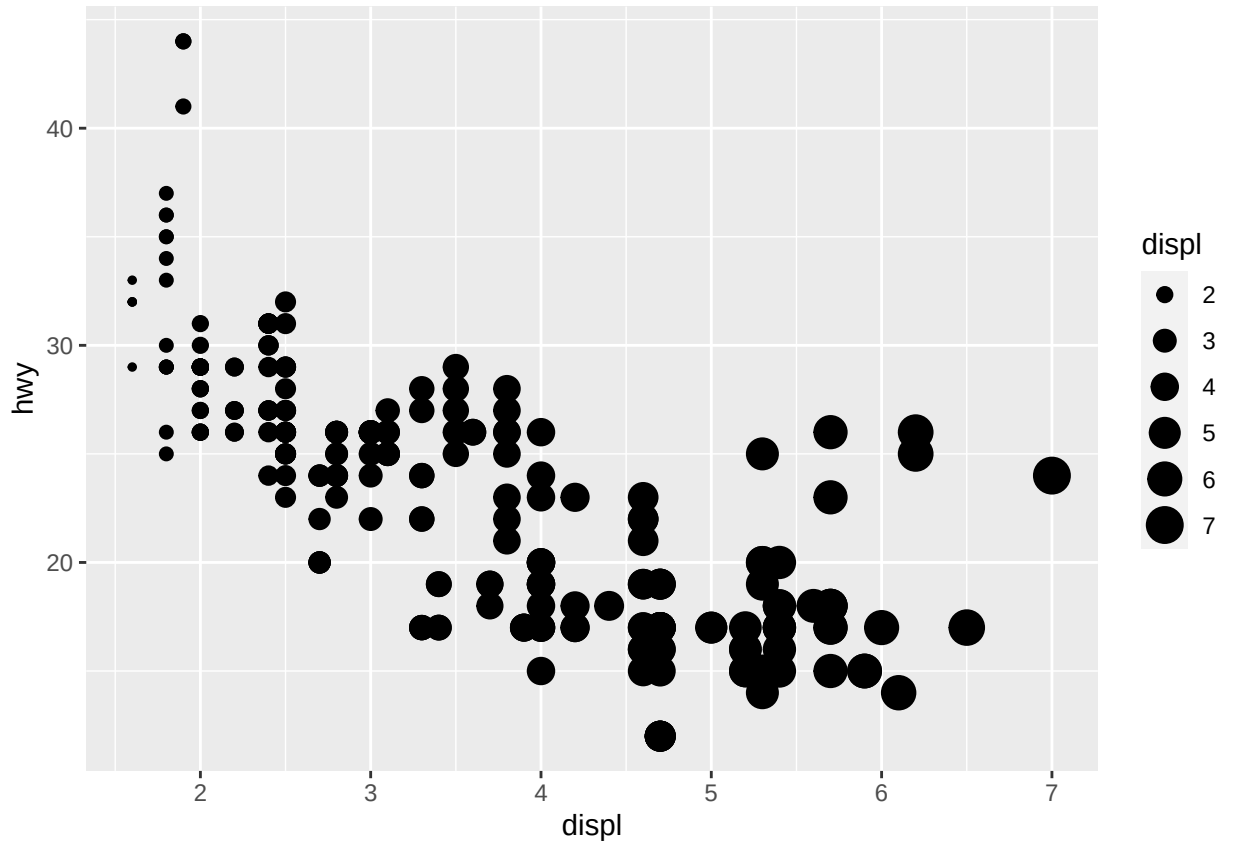
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ,  y = hwy, color = displ))
```



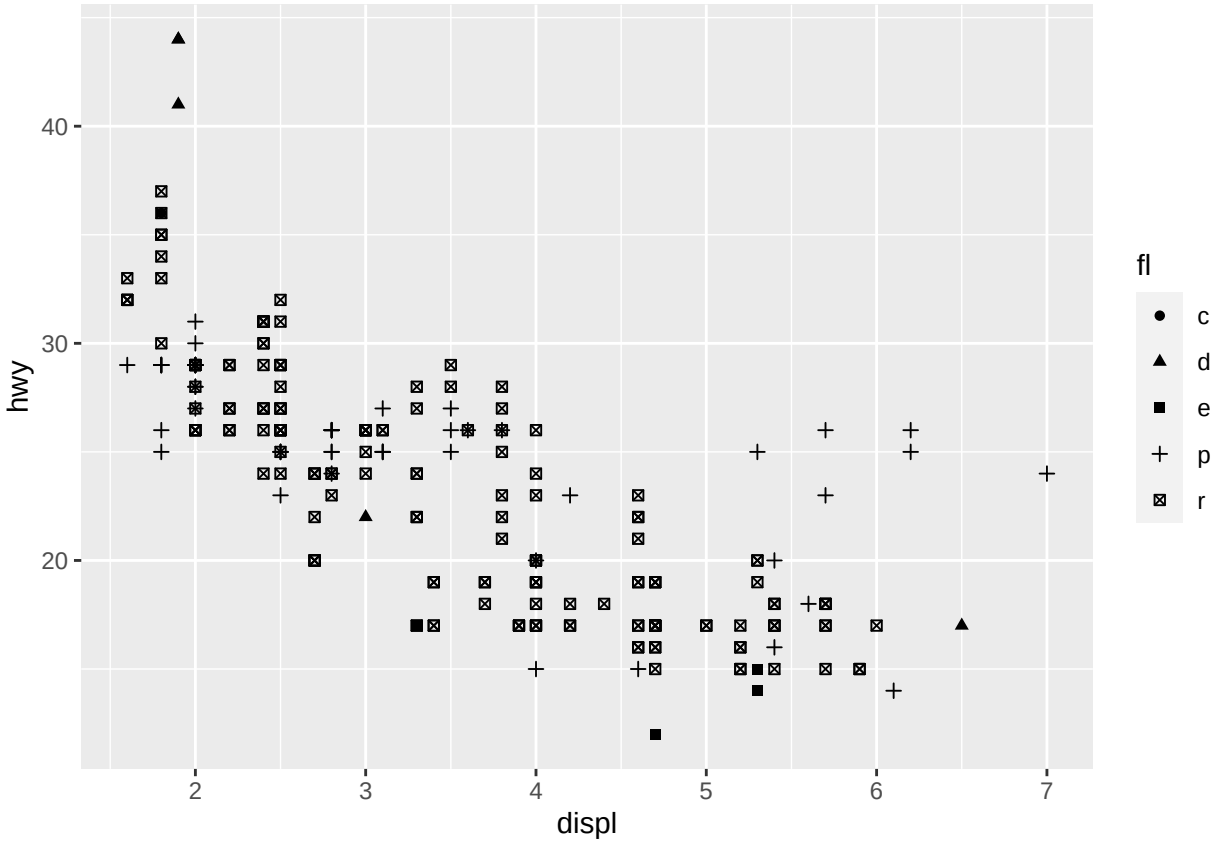

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ,  y = hwy, color = class))
```




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

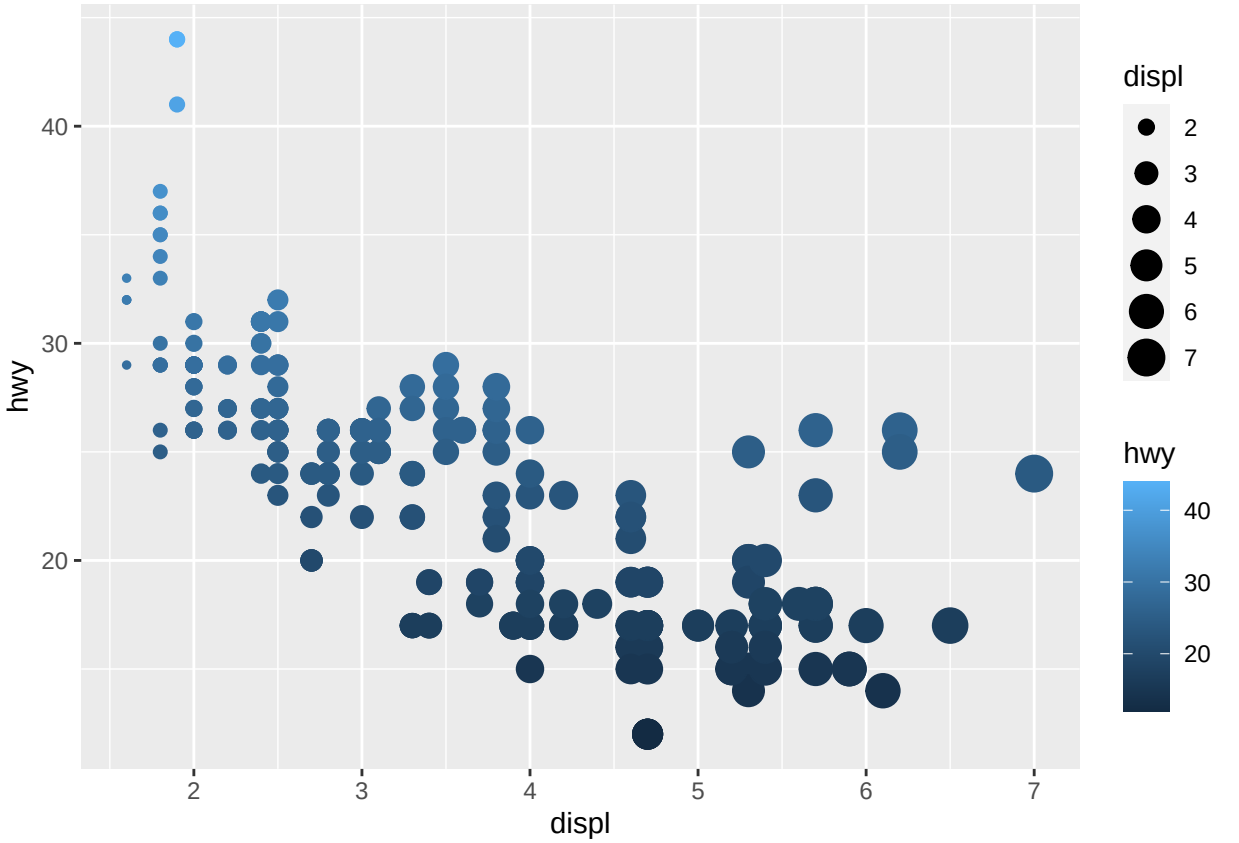



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ,  y = hwy, shape = fl))
```



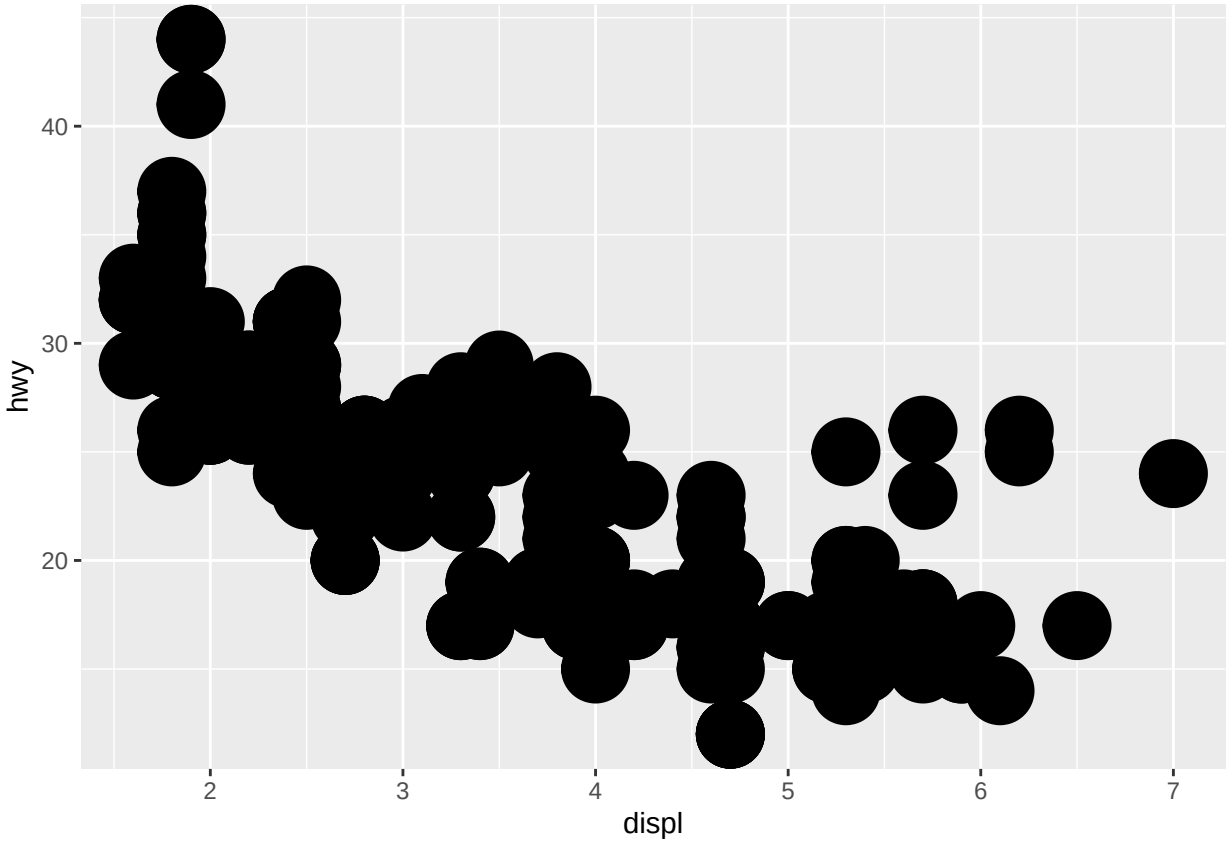
4. What happens if you map the same variable to multiple aesthetics?

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



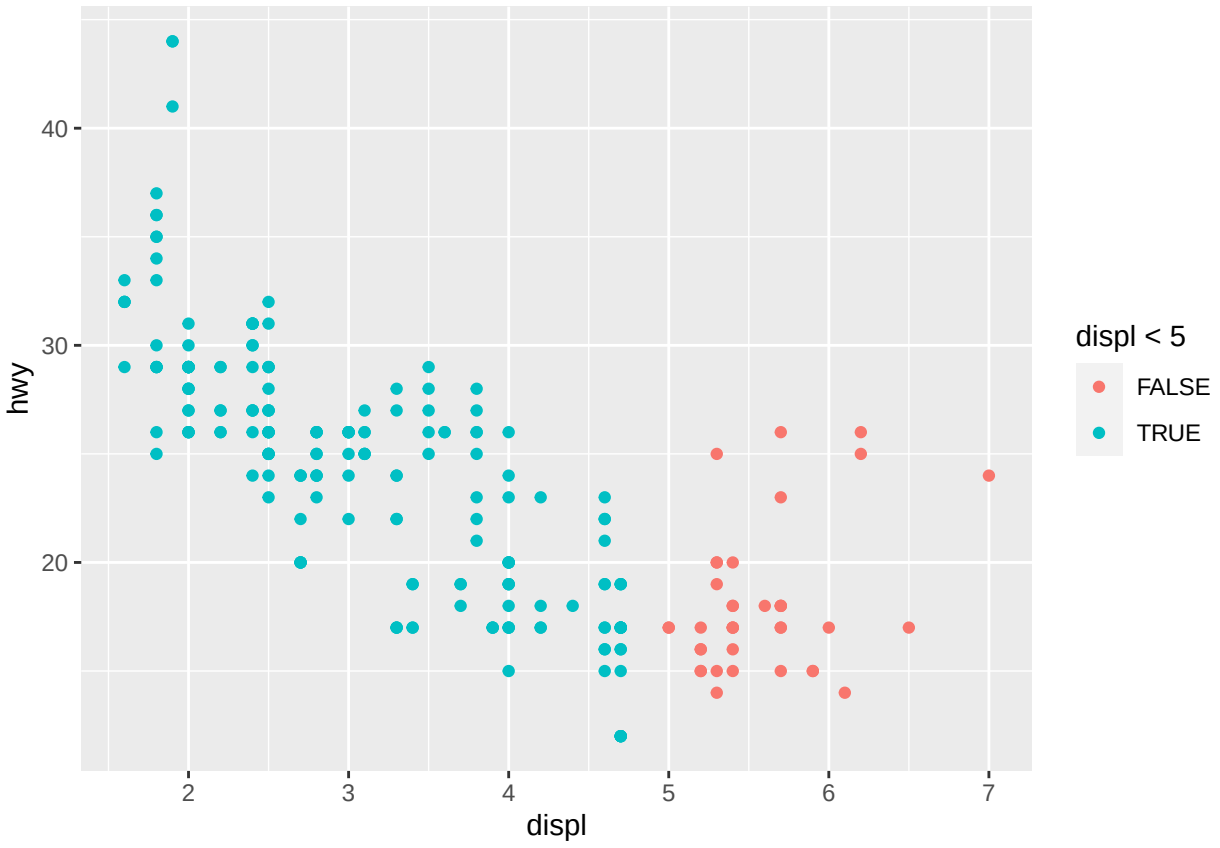
5. What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom_point)

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ,  y = hwy, stroke =8 ))
```



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.

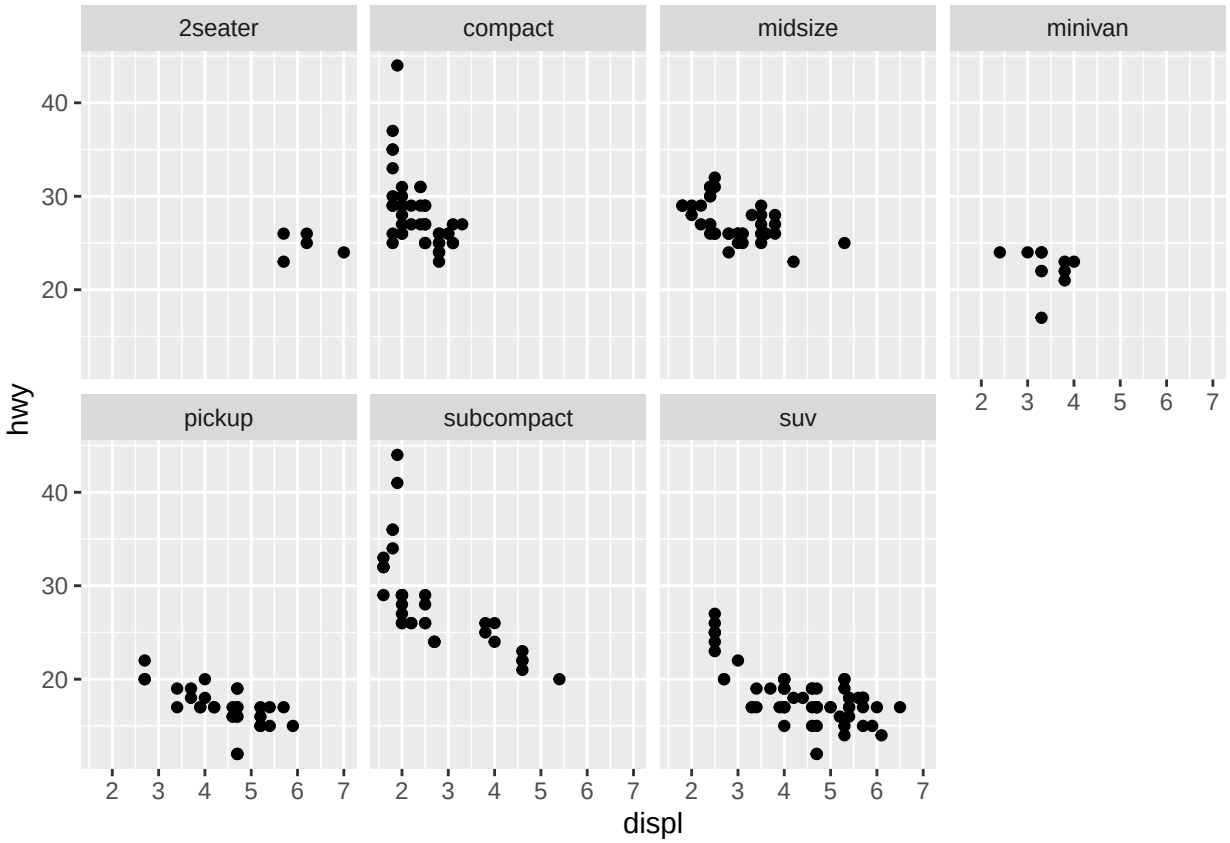
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ,  y = hwy, color= displ < 5))
```



Facet

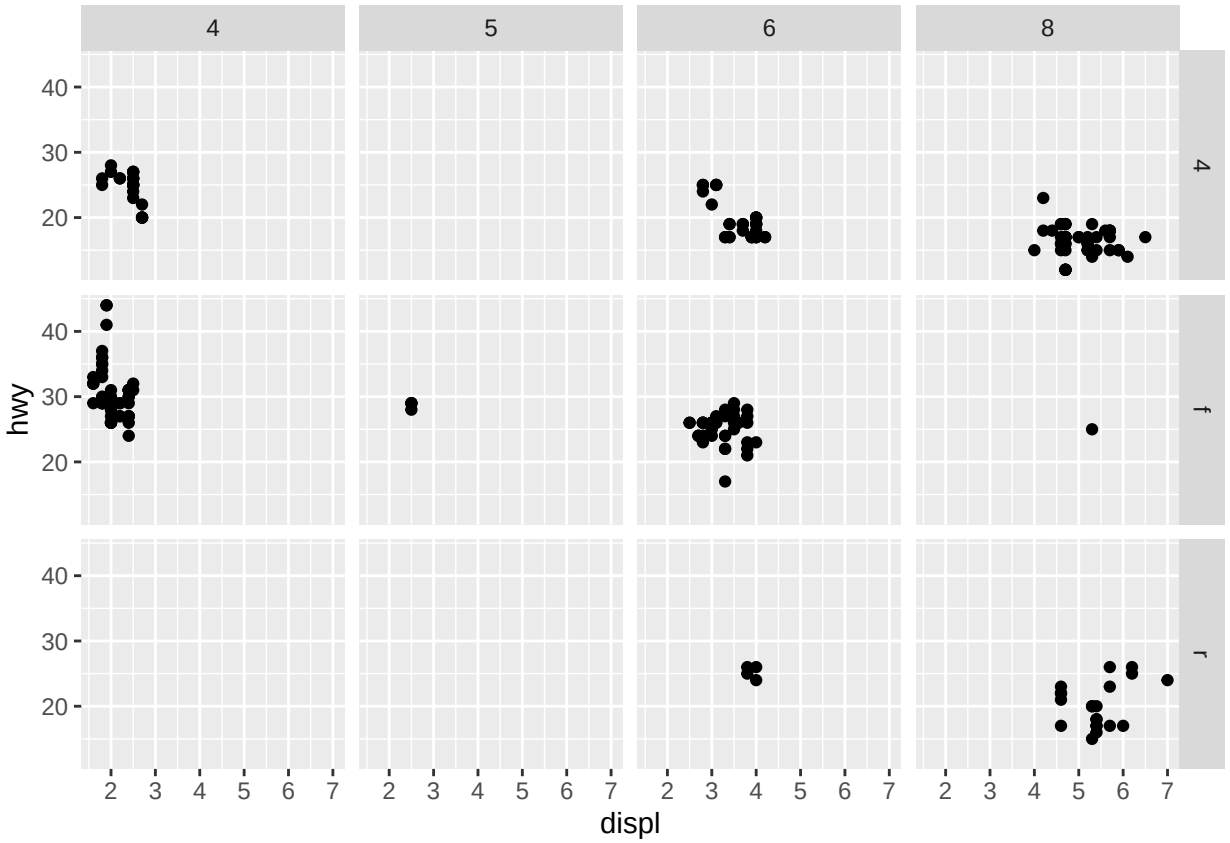
O autor destaca que outra forma de observar variáveis categóricas é dividir o gráfico em `facet`, ou seja, criar subplots para cada tipo da variável. A função utilizada é a `facet_wrap()`.

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

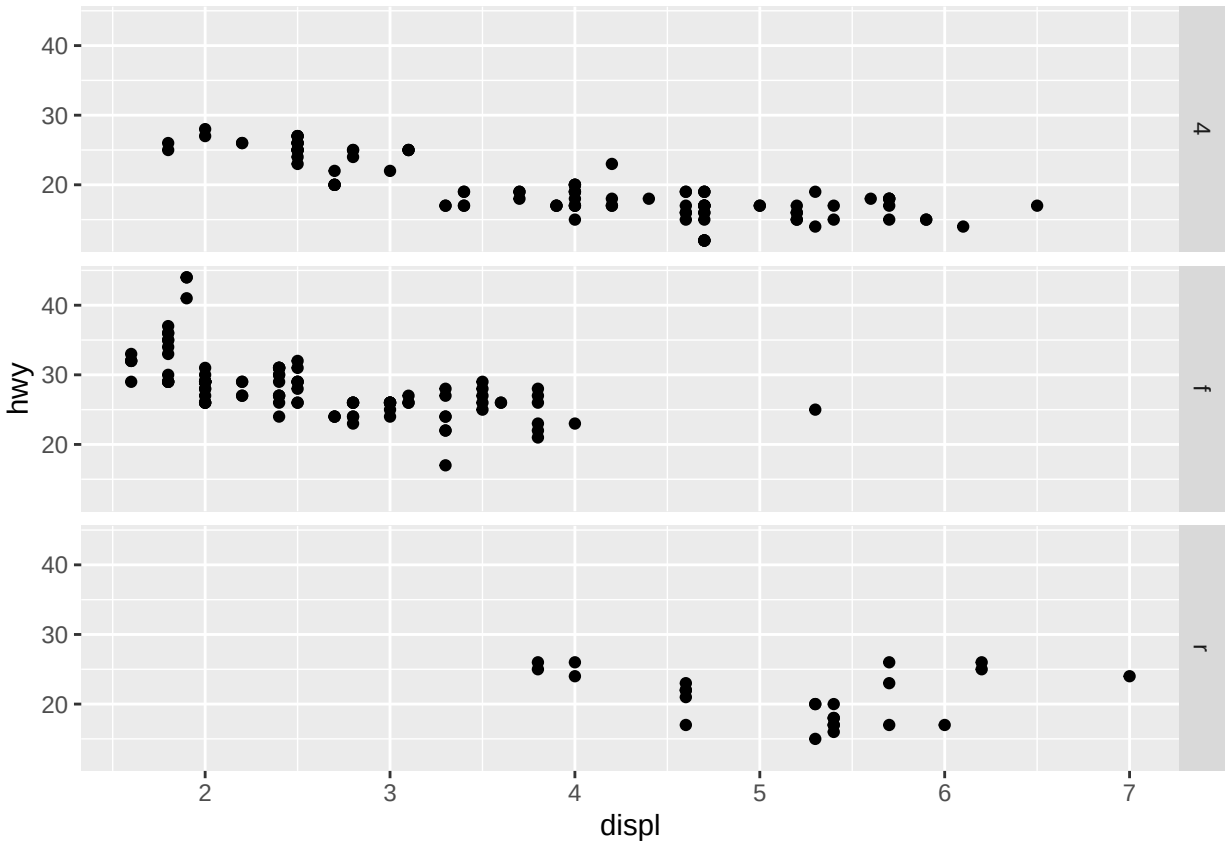


Para combinar duas variáveis com `facet` utilizamos o `facet_grid()`

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



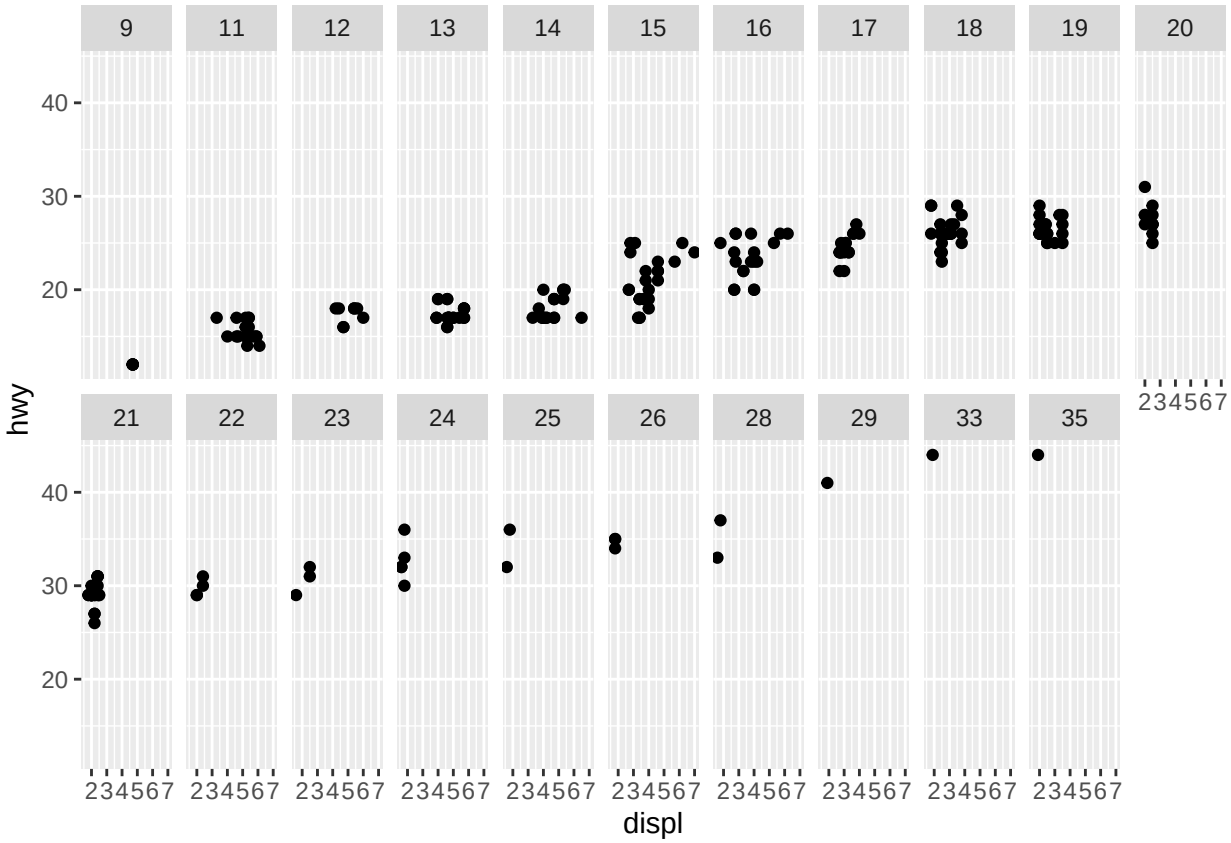

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ , y = hwy)) +
  facet_grid(drv ~ .)
```



Exercícios 3.5.1: 1. What happens if you facet on a continuous variable?

Resposta : Na minha opinião , não é gráfico muito últil ter uma variável continua no facet, pois ele irá dividir o gráfico em muitas camadas.

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

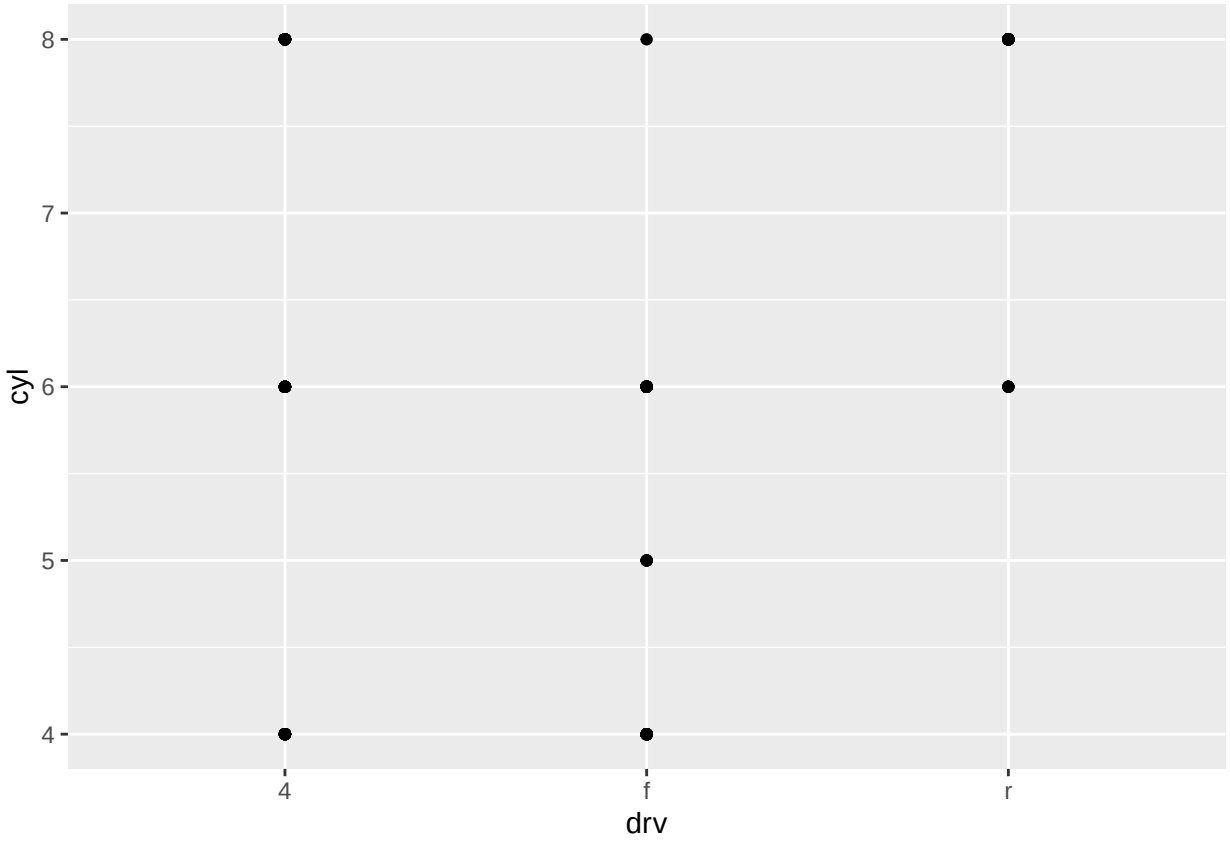



```
#glimpse(mpg)
```

2. What do the empty cells in plot with `facet_grid(drv ~ cyl)` mean? How do they relate to this plot?

Resposta :

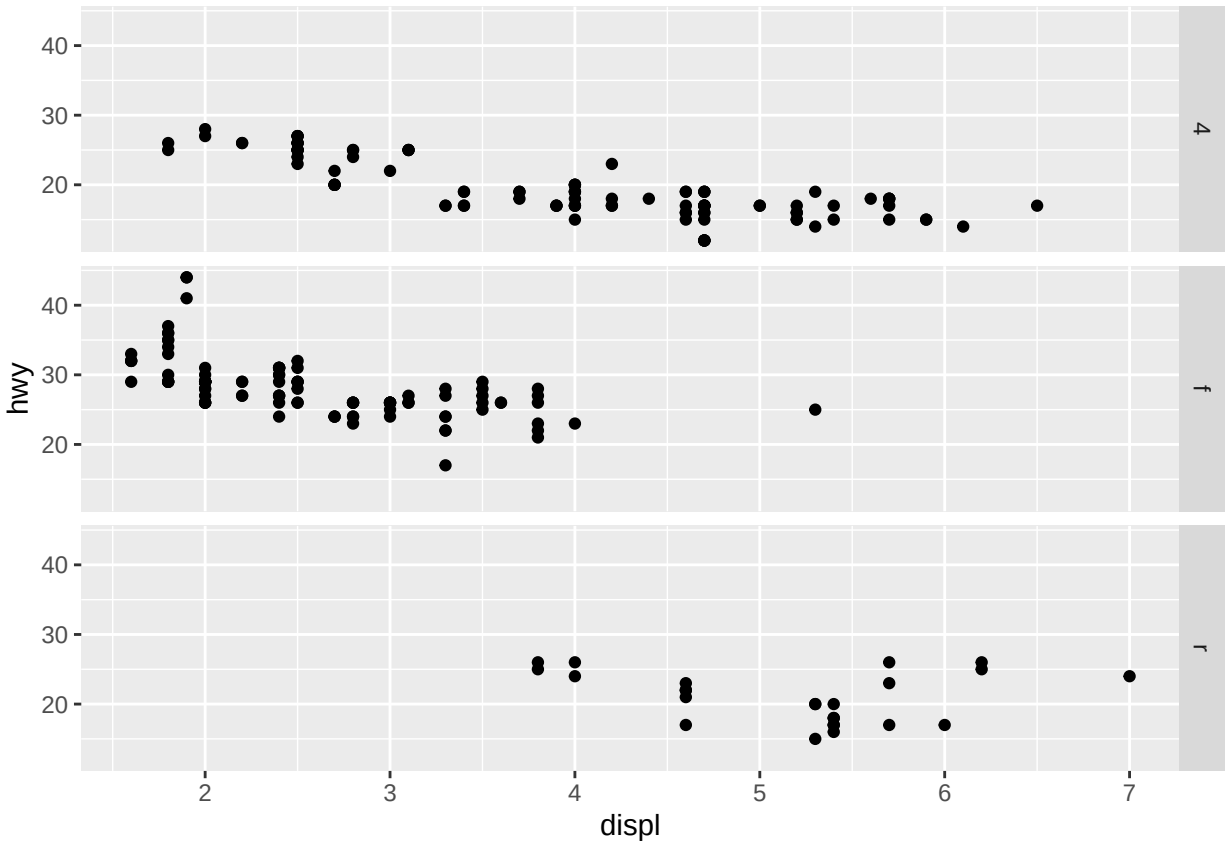
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))
```



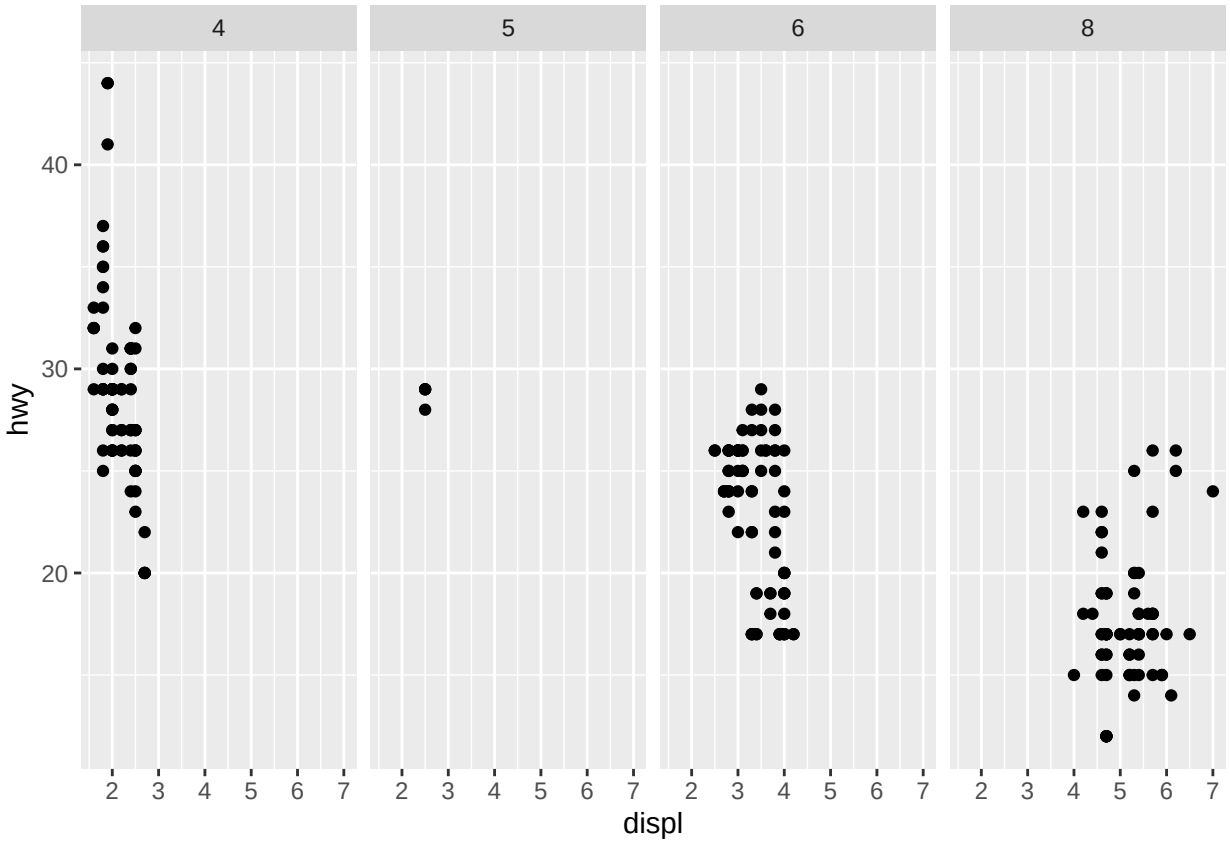
3. What plots does the following code make? What does . do?

Resposta : Define em qual axis e qual variável você quer observar.

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

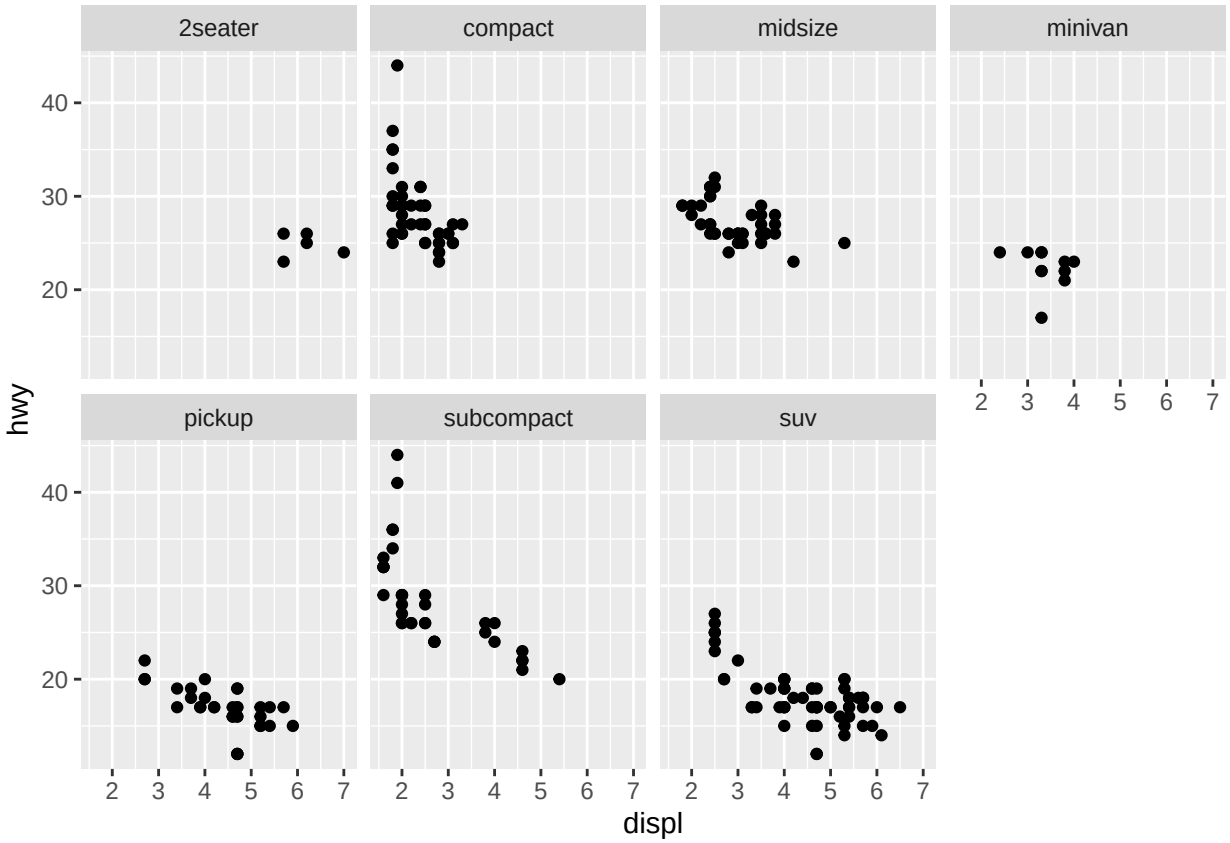



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```



4. Take the first faceted plot in this section:

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



What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

Resposta : A vantagem de usar facet em relação a aes é porque você pode observar cada classe de forma isolada, já com color as cores se misturam e não fica claro cada grupo, outro detalhe é se tivermos muitos grupos/classes complica mais a identificação de cada grupo. Casos de datasets muito grandes ou com diversas classes devemos podemos tentar filtrar por grupos, ou então teremos um facet com muitas camadas, e caso utilize colour não terá cores para todos os tipos.

5. Read `?facet_wrap`. What does nrow do? What does ncol do? What other options control the layout of the individual panels? Why doesn't facet_grid() have nrow and ncol arguments?

Resposta : ncol e nrow especifica o número de colunas ou linhas que irá dividir seu subplot, outras opções para controlar o layout são : scales, shrink.

6. When using facet_grid() you should usually put the variable with more unique levels in the columns. Why?

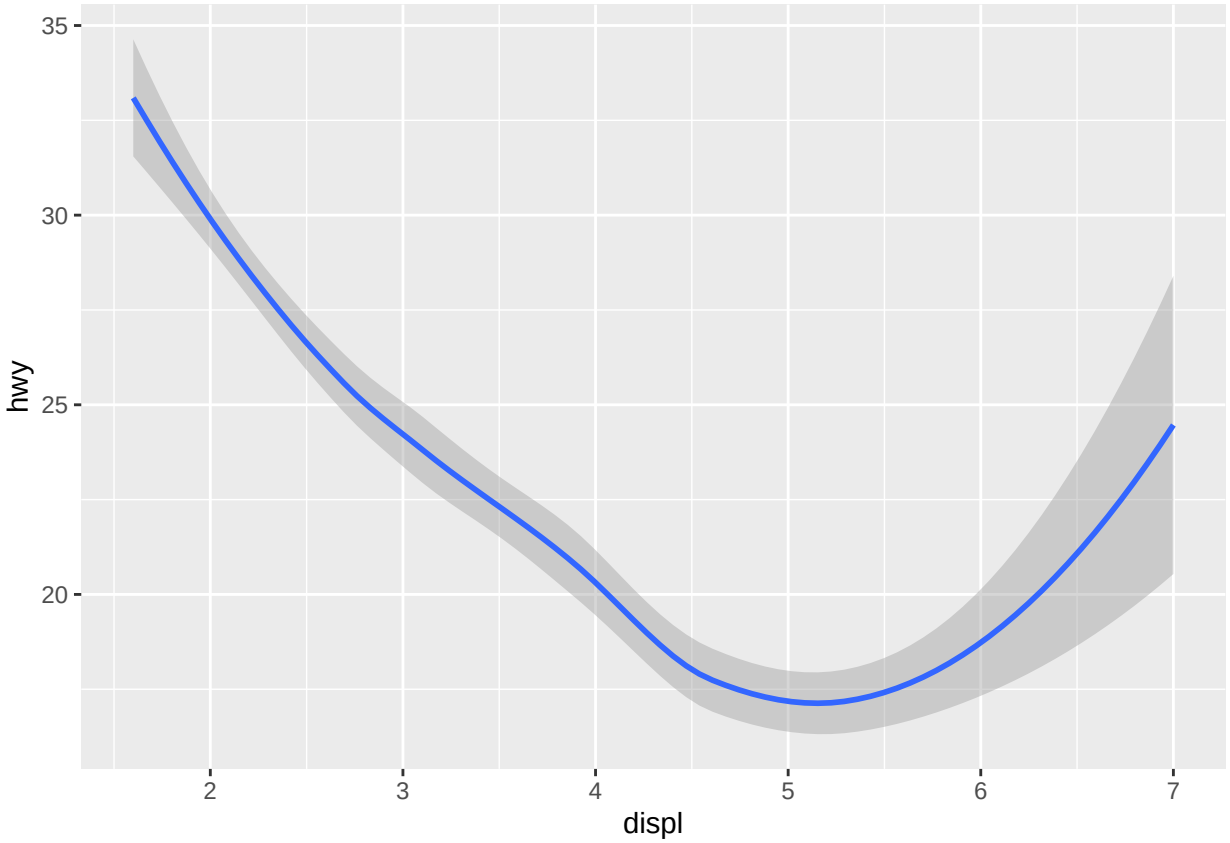
Resposta : Para não ter diversas subdivisoes no plot

Geometric Objects

geoms são normalmente conhecidos por tipos de gráficos como bar chart, line chart, , boxplot, scatterplot, etc. Cada um desses tipos de gráficos tem seu proprio `geom` .

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

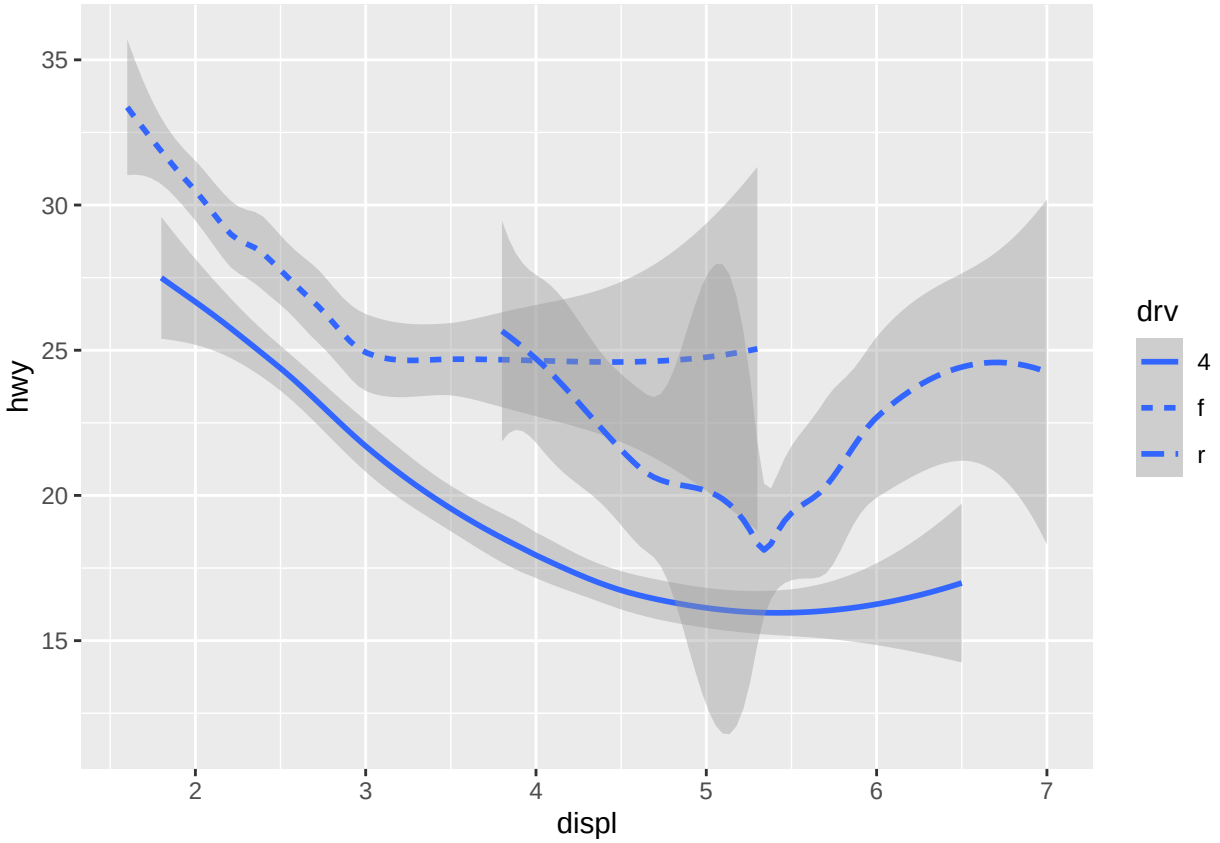
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Adicionando `linetype` , com isso o ggplot irá separar os carros em 3 categorias de `drv` :

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

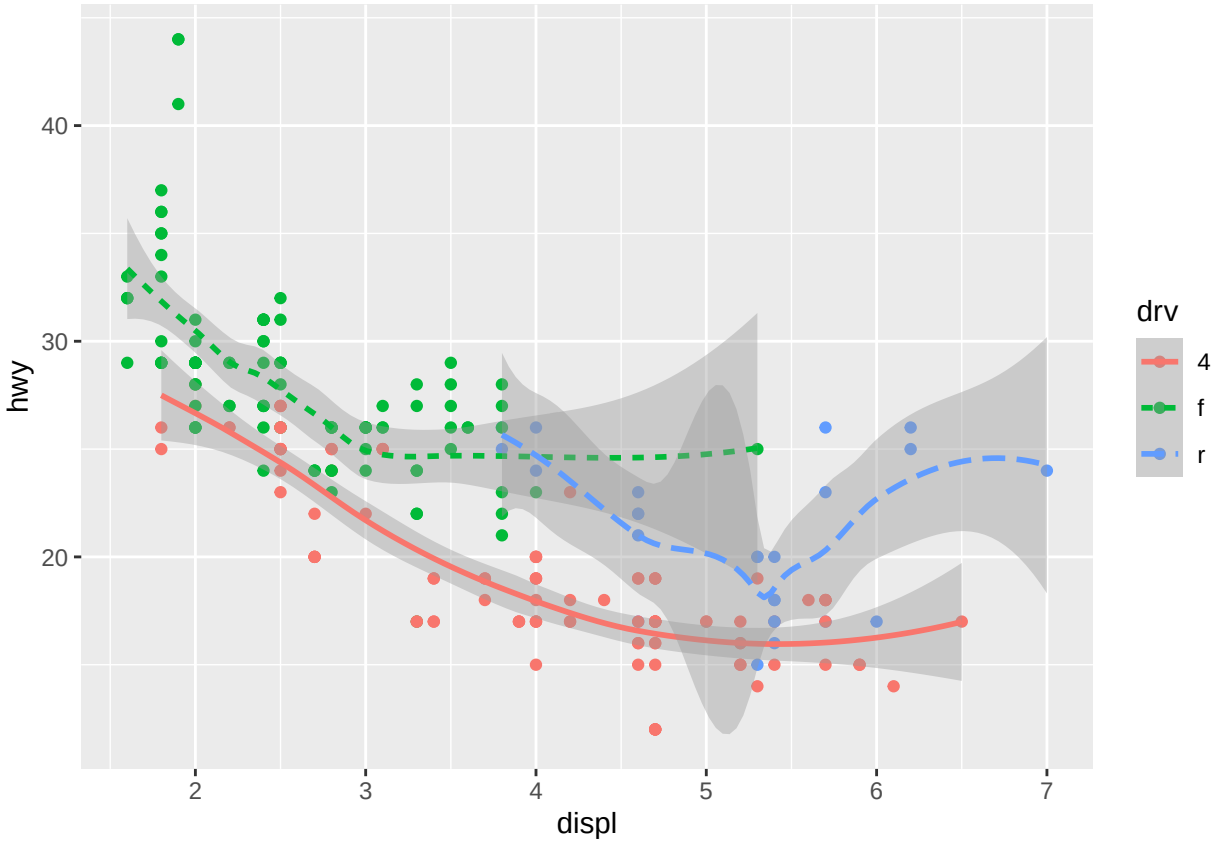
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Podemos adicionar uma camada de `geom` em cima da outra, combinado o gráfico de linhas com o scatterplot

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

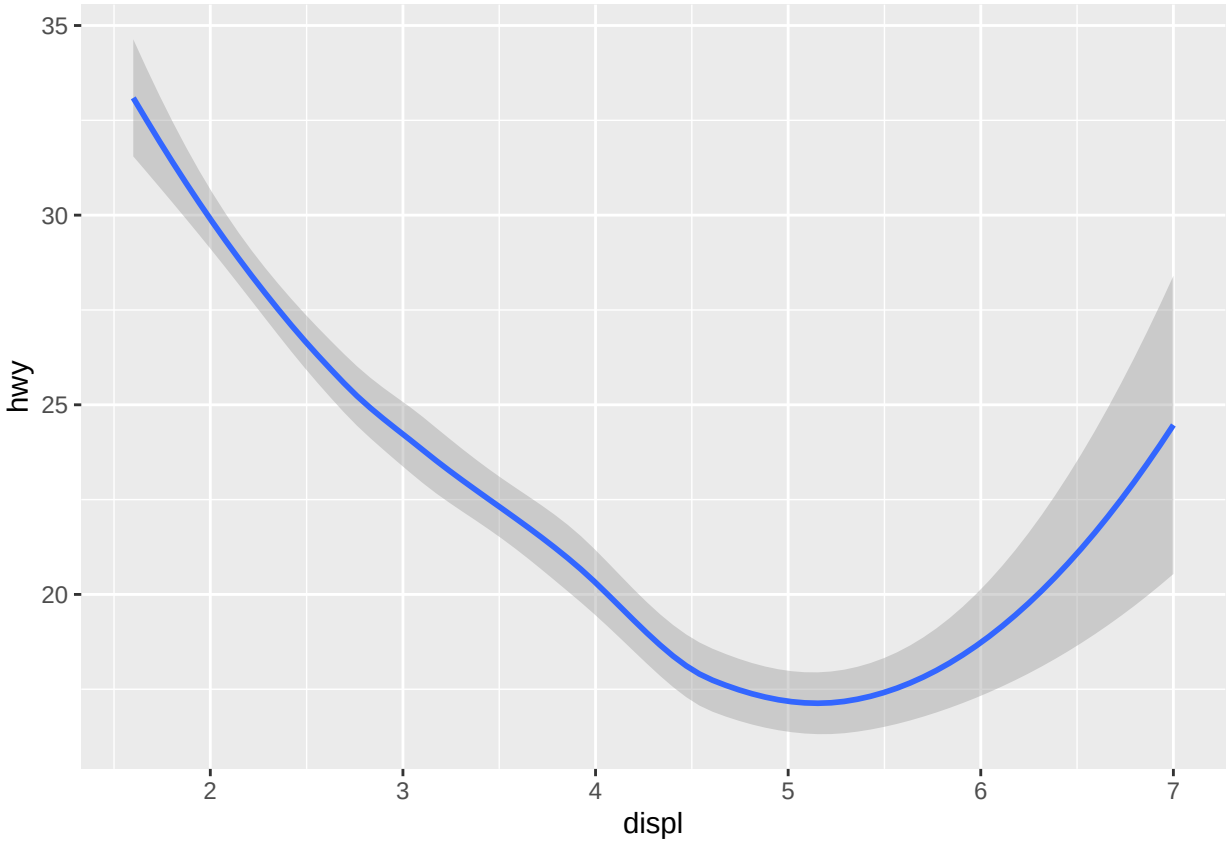
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Uma das opções do `geom_smooth` é utilizar as opções de `group` e `color` para plotar variáveis categóricas :

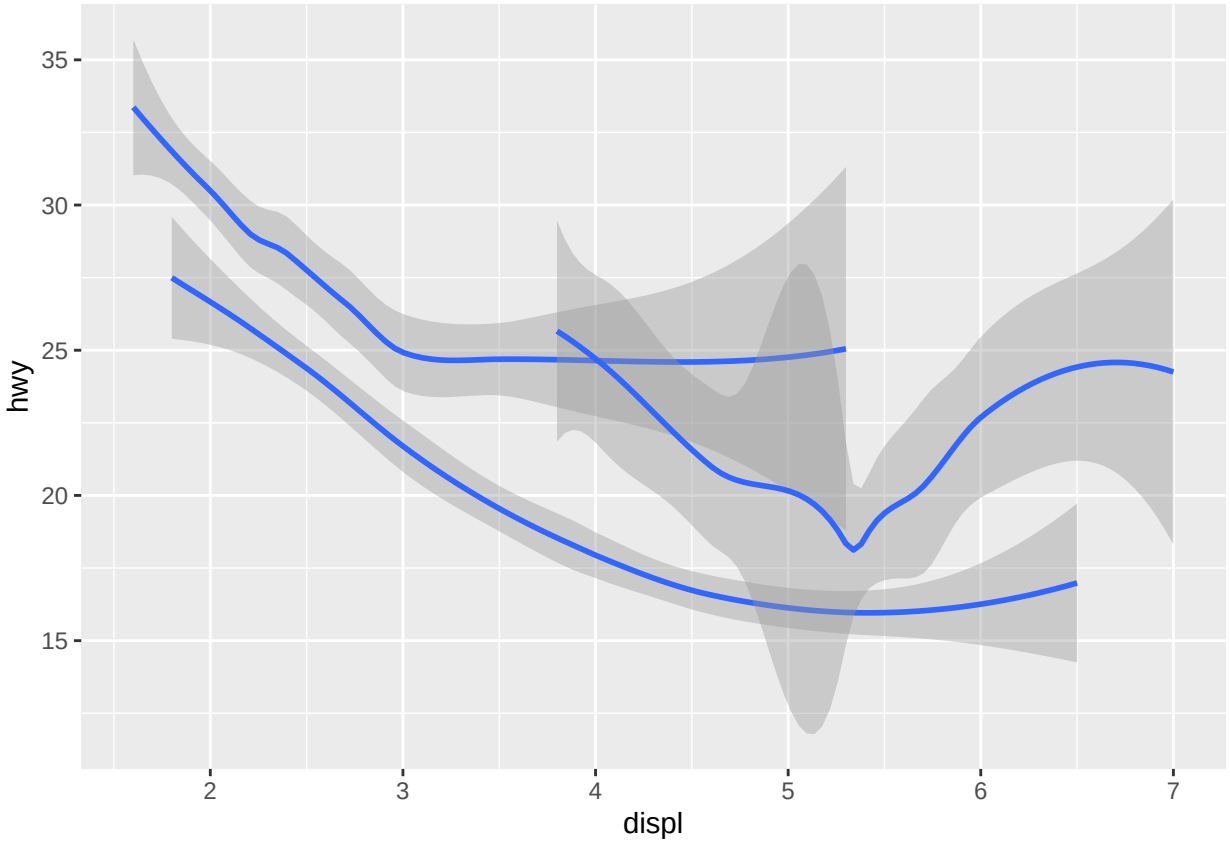
```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

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



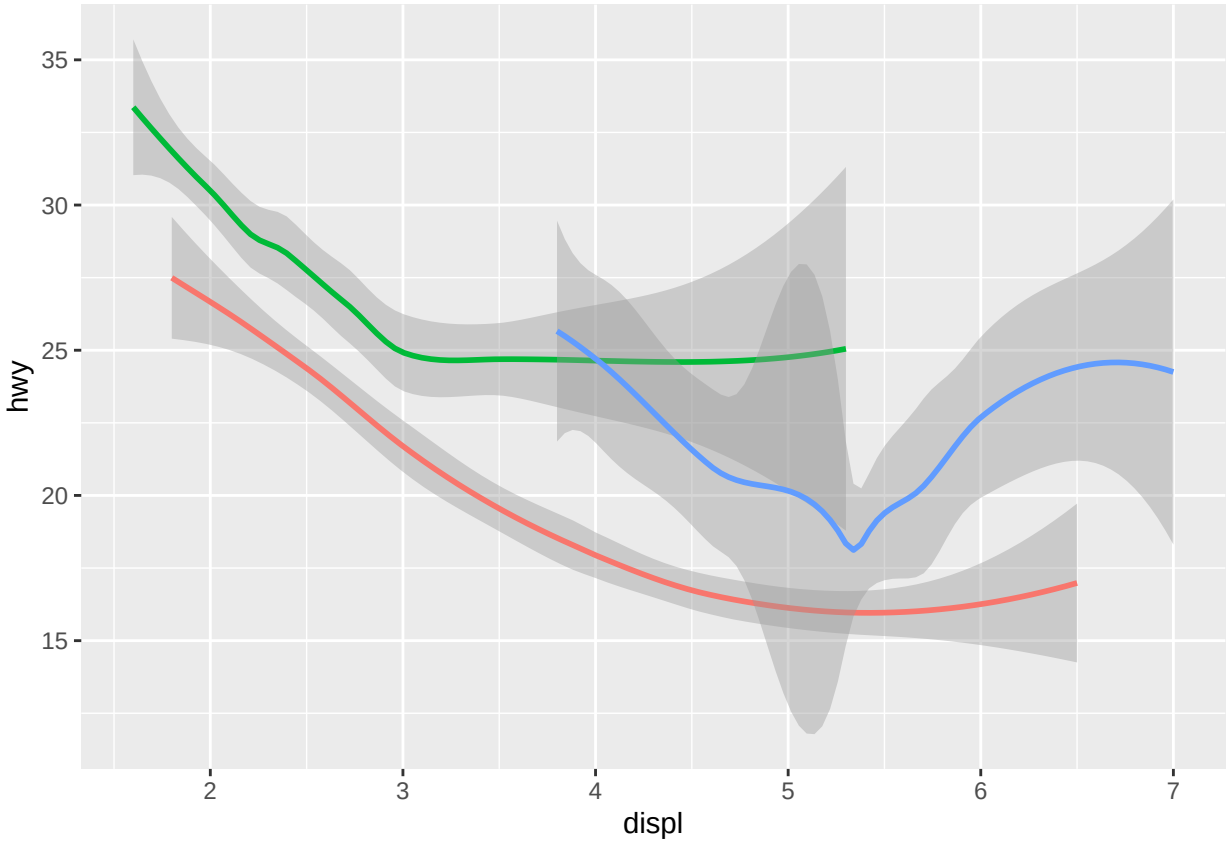

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, group = drv))
```

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




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

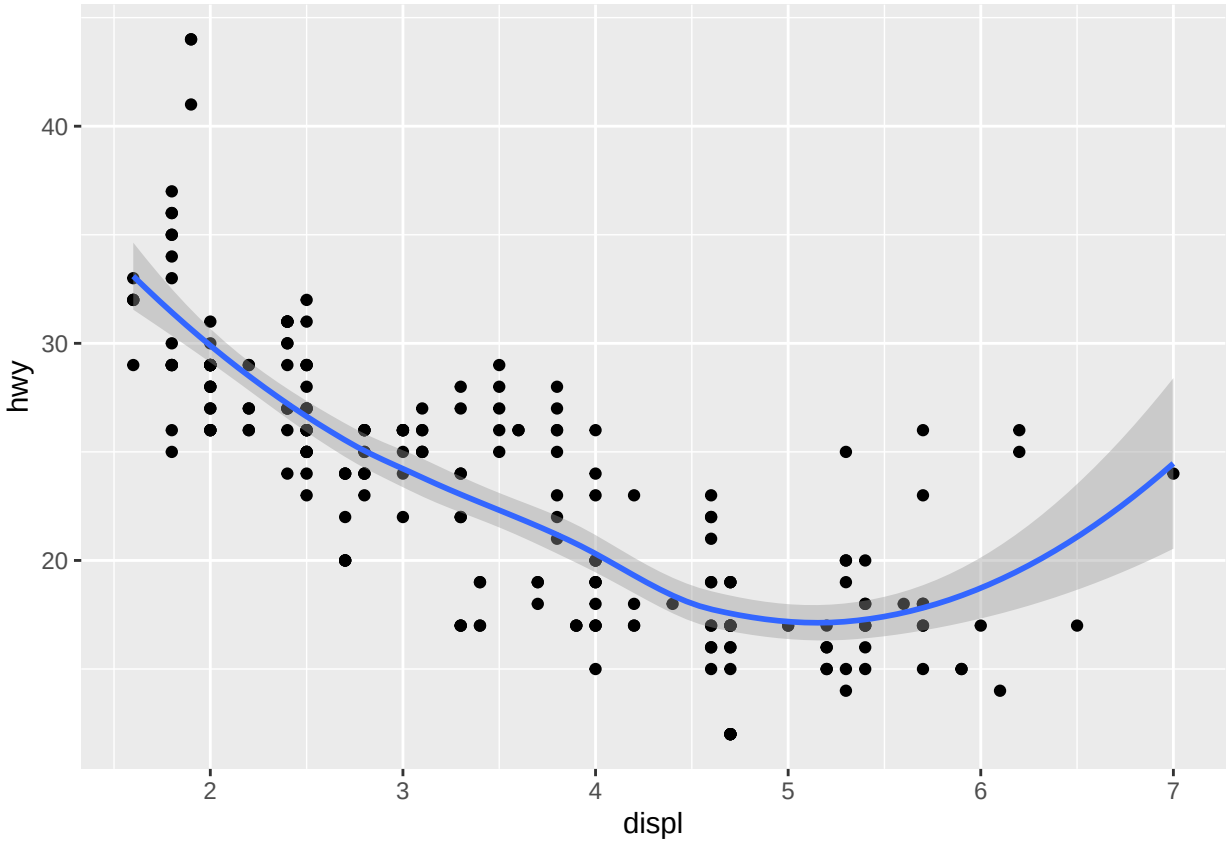
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Para evitar que se duplique informação no gráfico adicionando geoms podemos tratar o `mappting` na parte **global** do ggplot

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

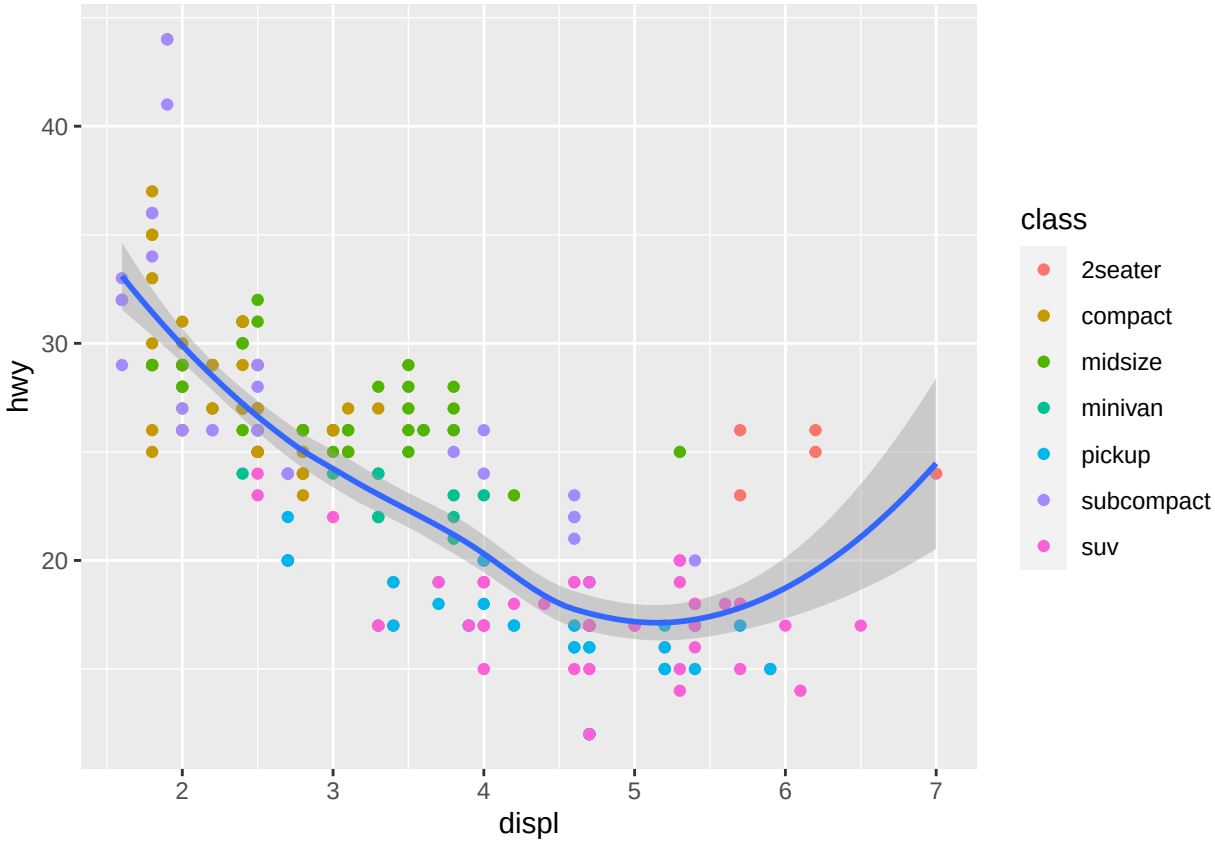
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Sobrescrevendo o mapping, neste caso o color para variável class será aplicado somente a camada de scatterplo:

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

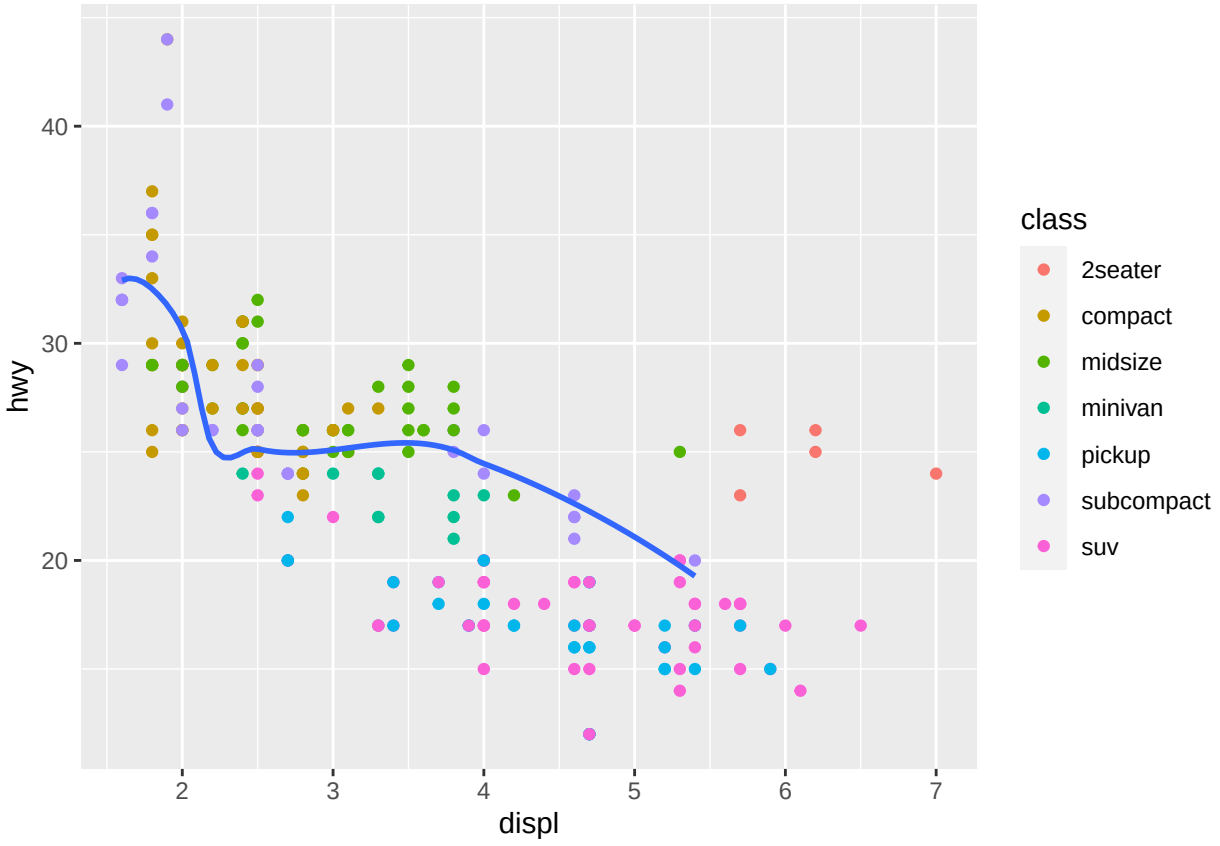
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Filtrando os dados para alguma camada de plot :

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

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



Exercícios 3.6.1: 1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?

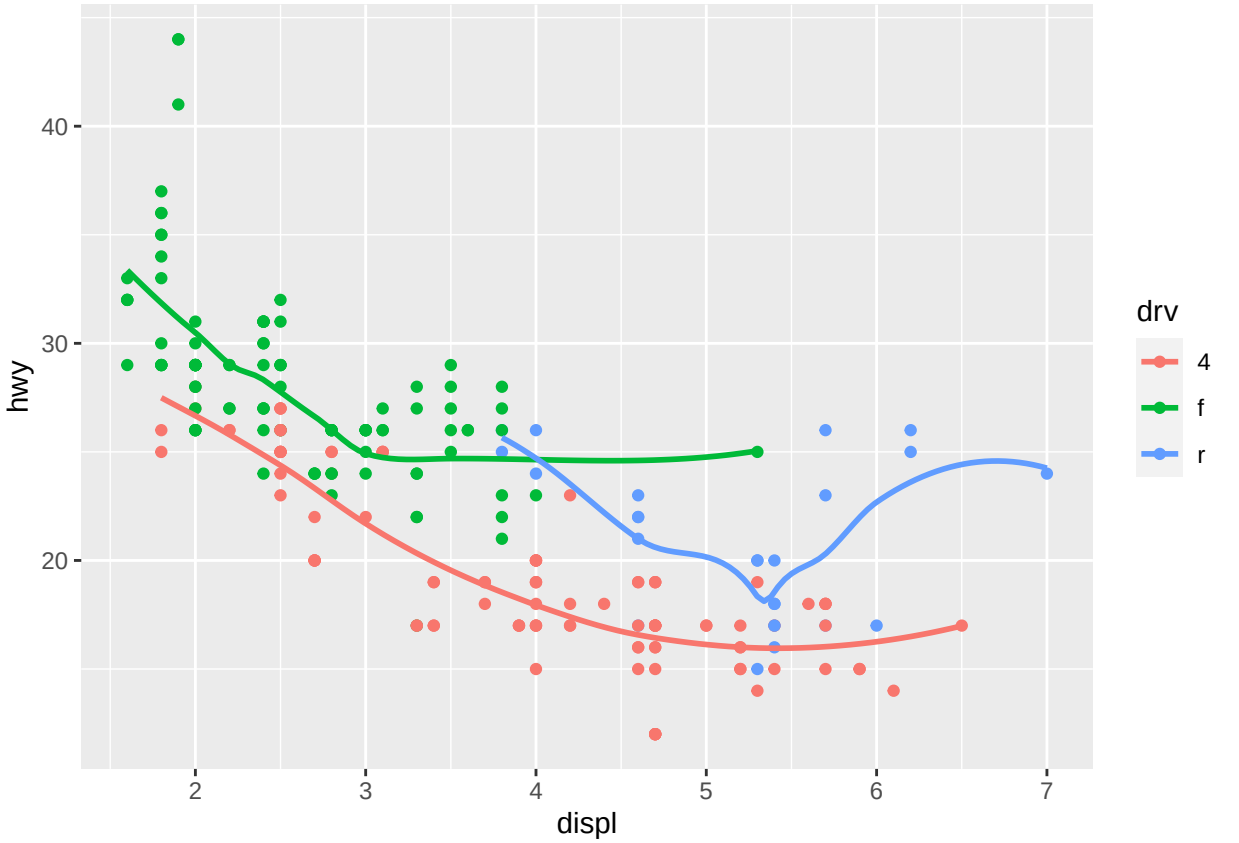
Resposta : line chart : geom_line() boxplot : geom_boxplot() histogram: geom_histogram() area chart: geom_area()

Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions.

Resposta : Na parte global irá criar um gráfico com x e y destacando cores por drv, adicionado um scatterplot e um smooth. *So depois de rodar que identifiquei que o se=FALSE removeu o intervalo de confiança do smooth*

```
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'
```



3. What does show.legend = FALSE do? What happens if you remove it? Why do you think I used it earlier in the chapter? Resposta : remove e legenda, se remoer irá criar uma legenda.

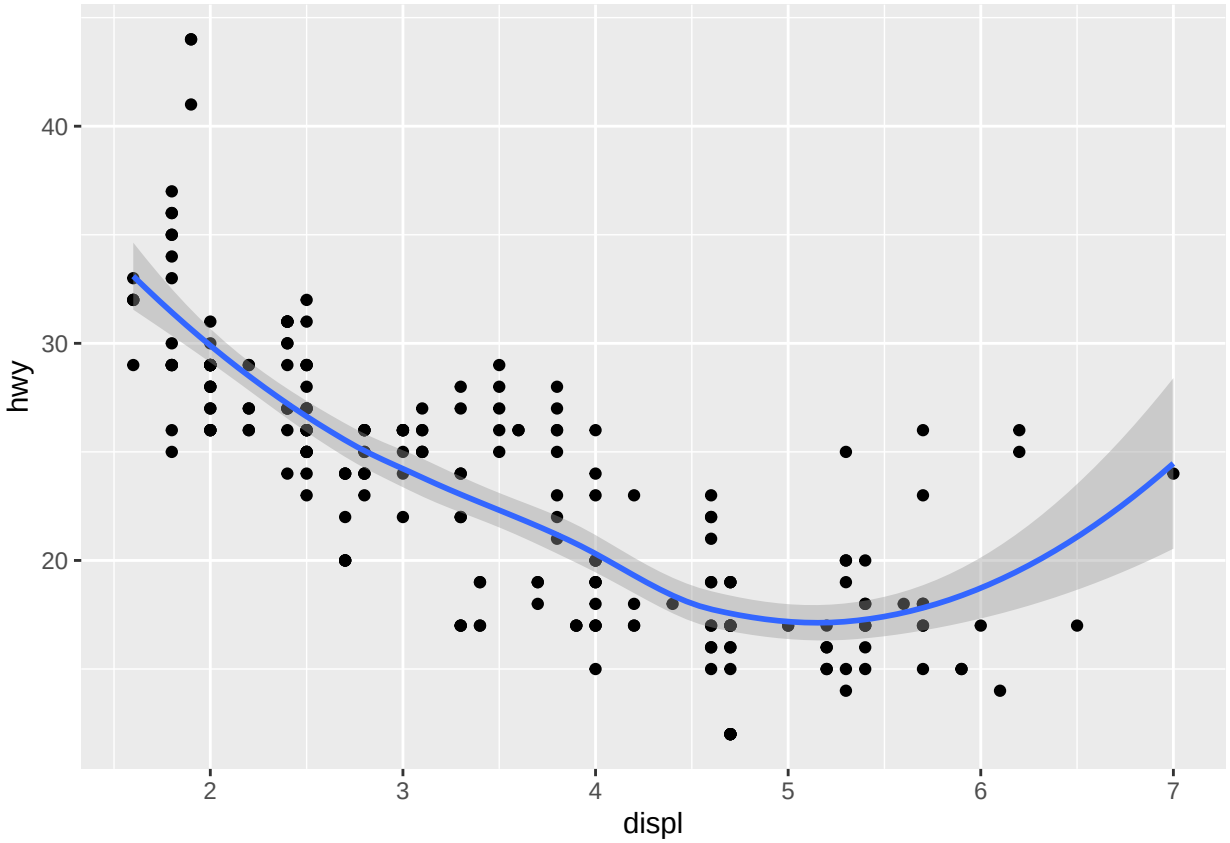
4. What does the se argument to geom_smooth() do?

Resposta : Habilita ou não o intervalo de confiança no gráfico

5. Will these two graphs look different? Why/why not? Resposta : O resultado será o mesmo, mas com informação duplicada

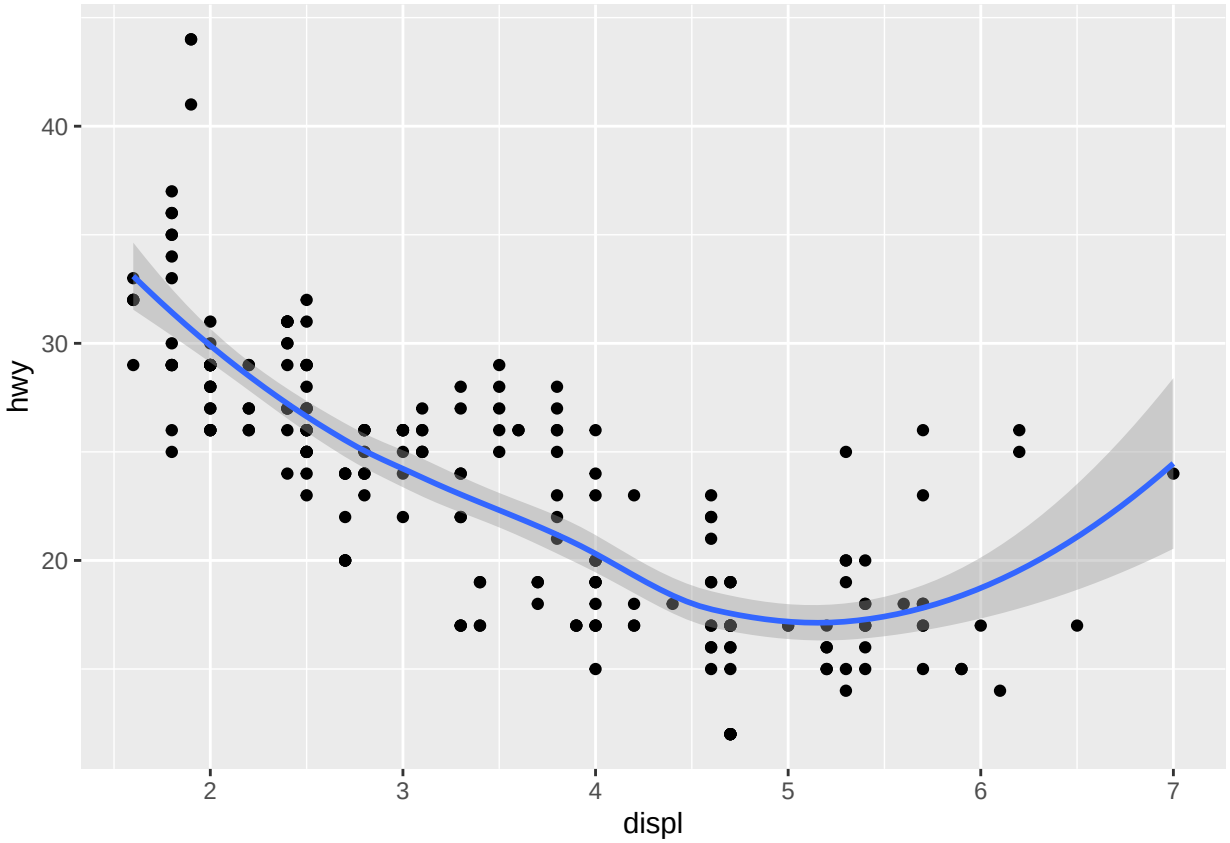
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

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




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

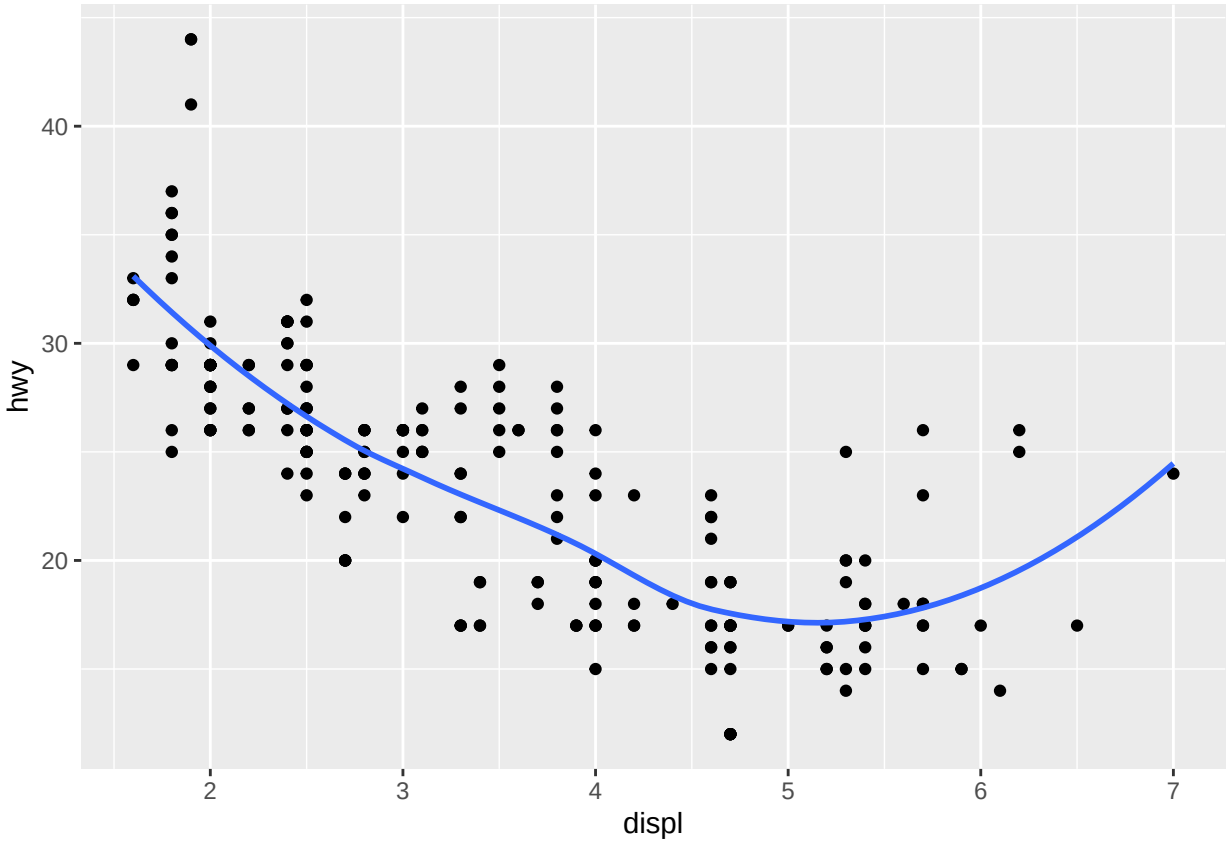
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



6. Recreate the R code necessary to generate the following graphs.

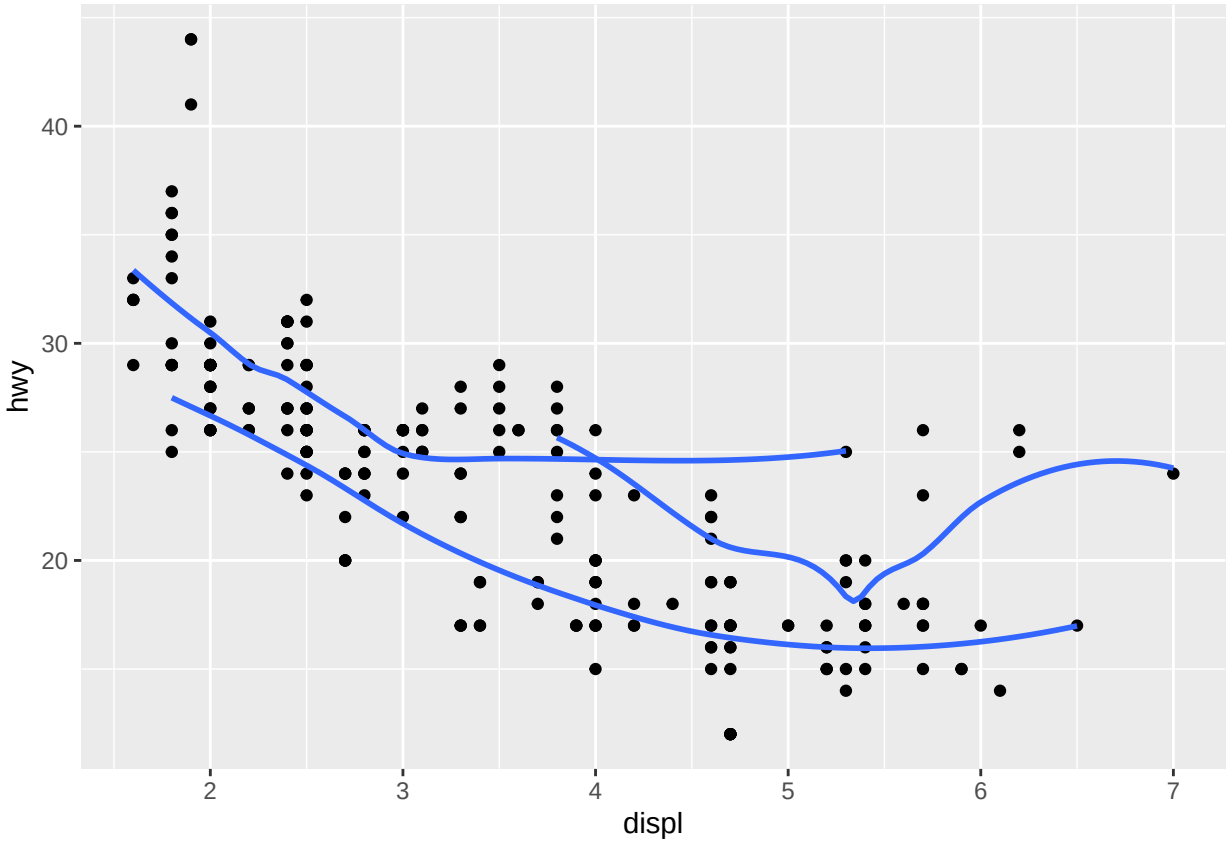
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth(se=FALSE)
```

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



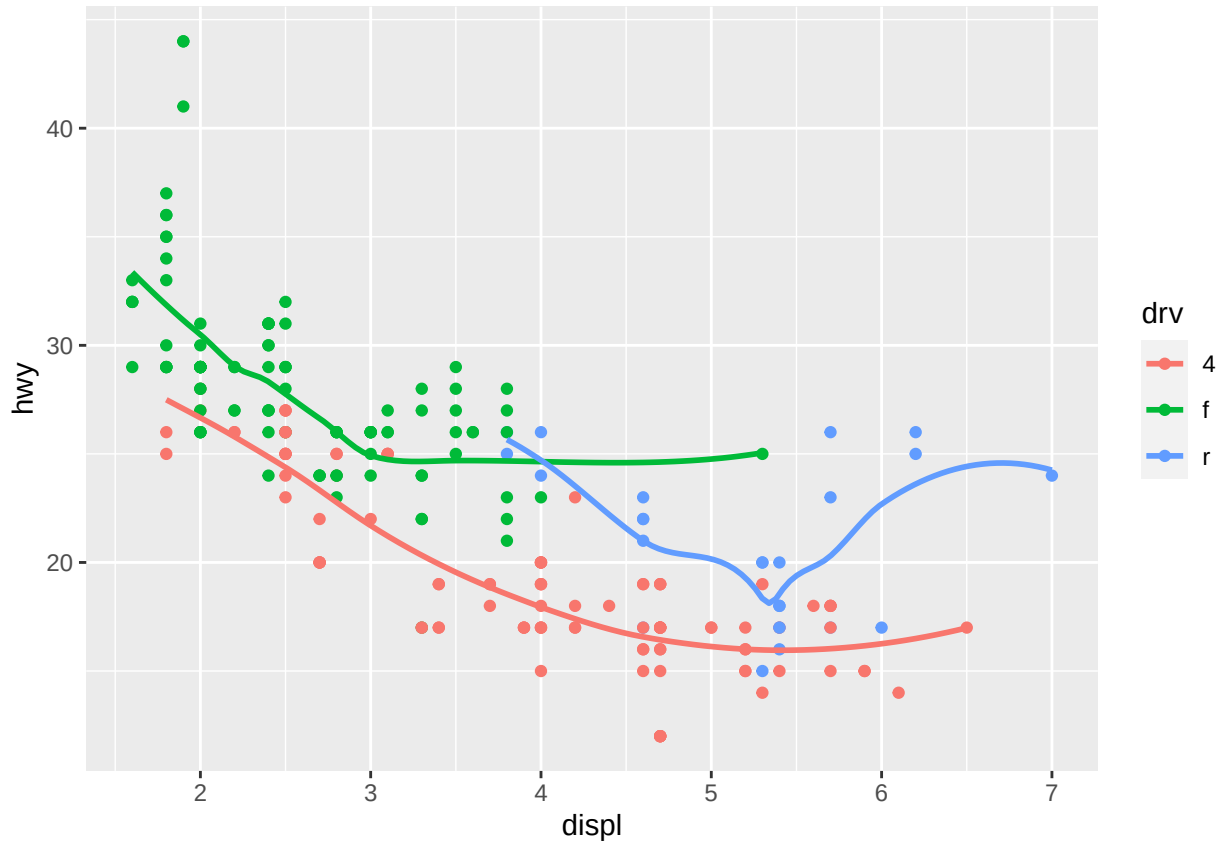

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, group = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

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



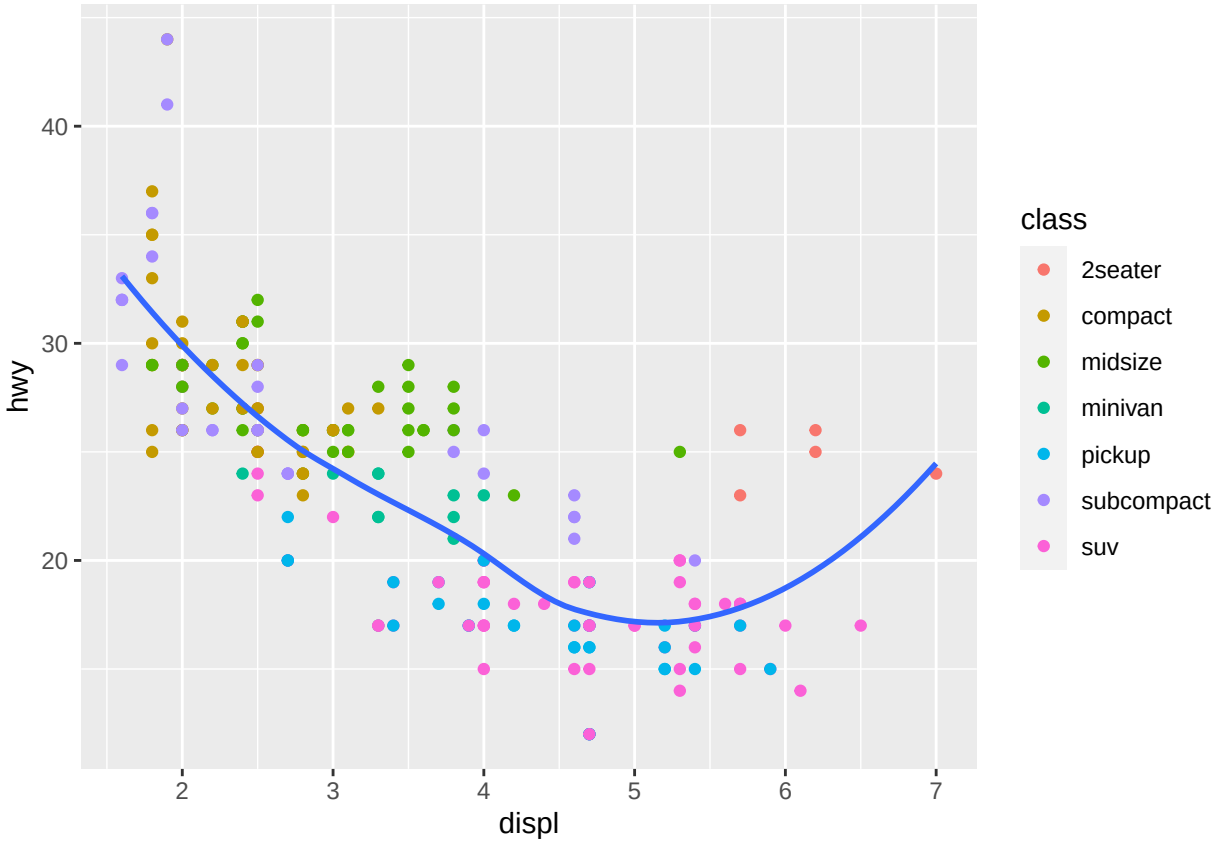

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

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




```
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'
```

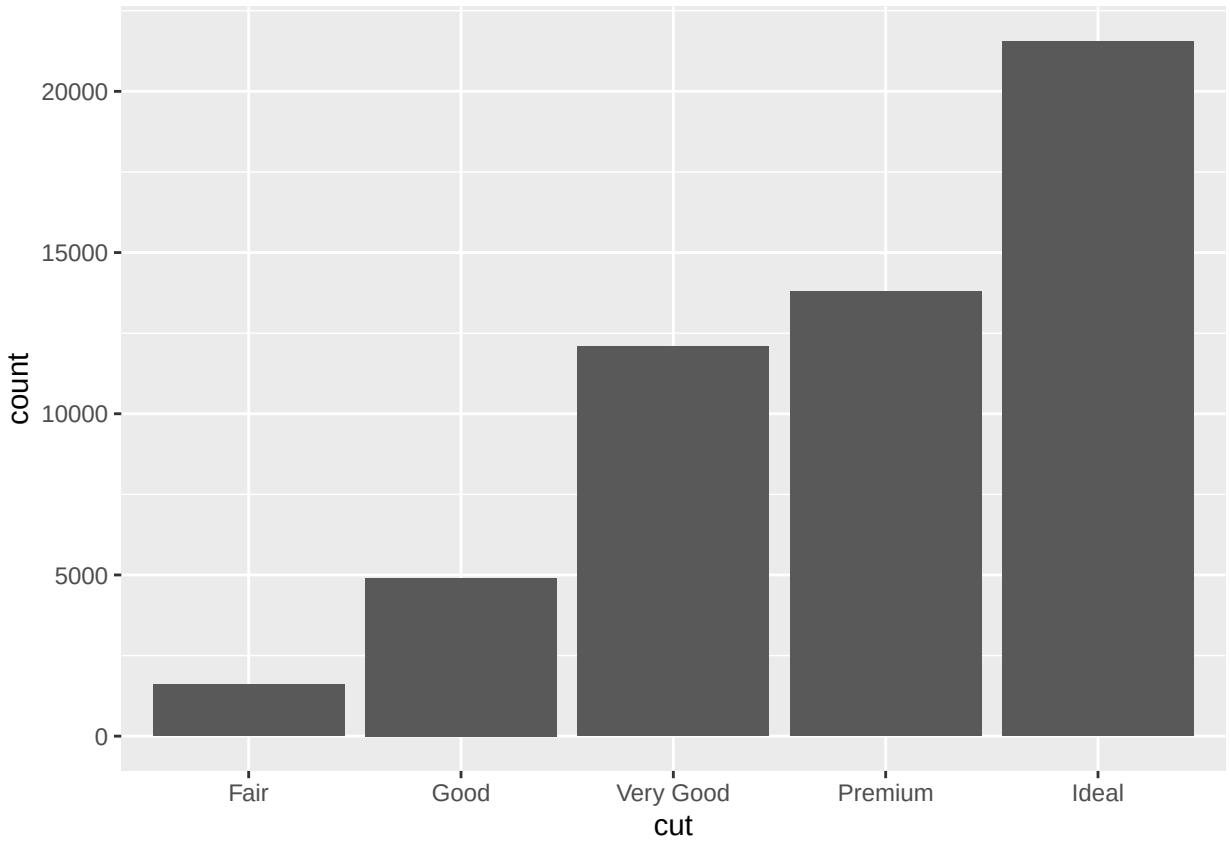


Transformações Estatísticas

Nesta parte do capítulo somos apresentados ao dataset `diamonds` que contém cerca de ~54k informações sobre diamantes, detalhes como preço, cor, claridade, etc.

Utilizando `geom_bar` mapeando o eixo X agrupando pela variável `cut` podemos ver o numéro total de diamantes, como o geom_bar usa `stat_count()`, podemos até usar essa função no lugar `de geom_bar()`.

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



Como podemos ver nesse gráfico no eixo X é computado a quantidade por agrupamento, outros gráficos apresentam valores computados como :

- Bar chart, histograms and frequência compute te counts
- smoothers realiza o fit nos dados
- boxplots computa a distribuição estatística

Referências e Links :

- ggplot galery