Data Visualisation

Yifan Jin

05/02/2020

Introduction

Prerequisites

```
library(tidyverse)
```

```
## -- Attaching packages --------------- tidyverse 1.3.0 --
```

```
## 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 ------------------ tidyverse_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
```

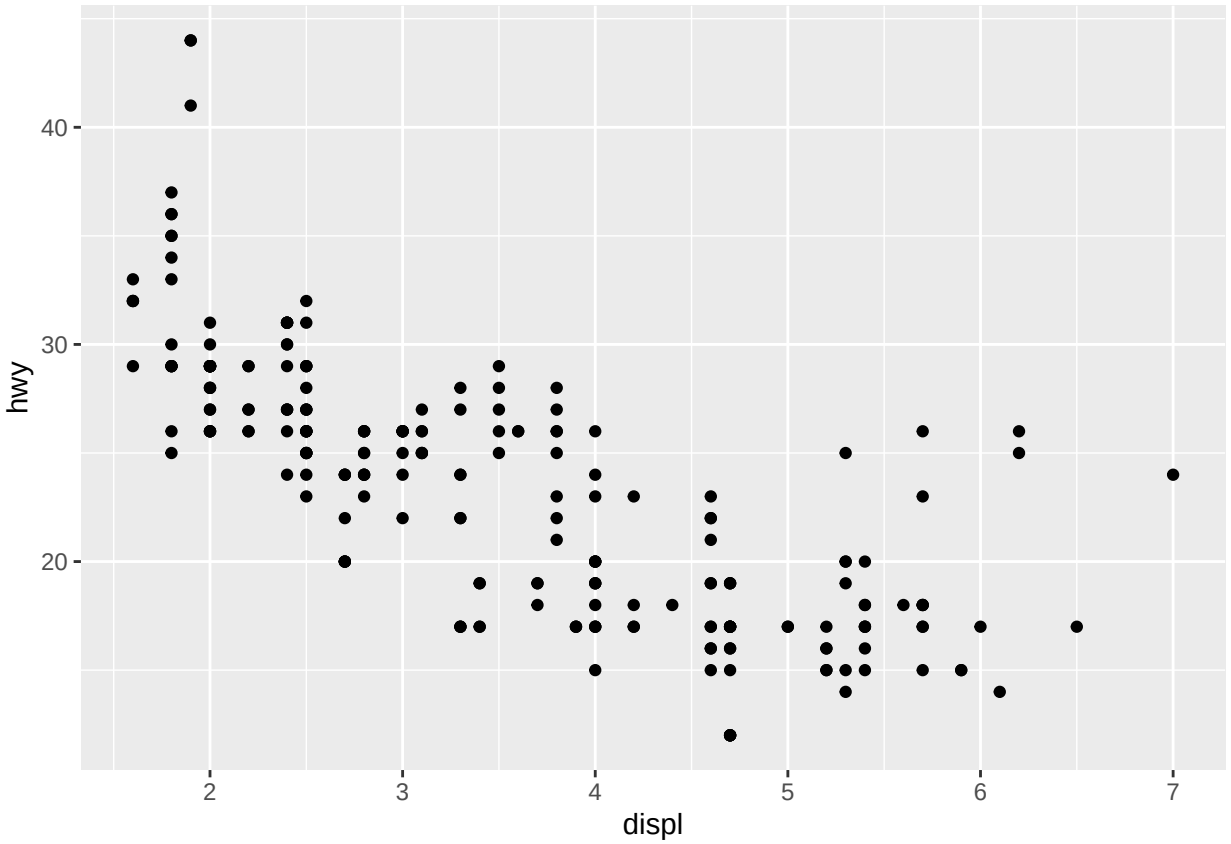
Variables descriptions:

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

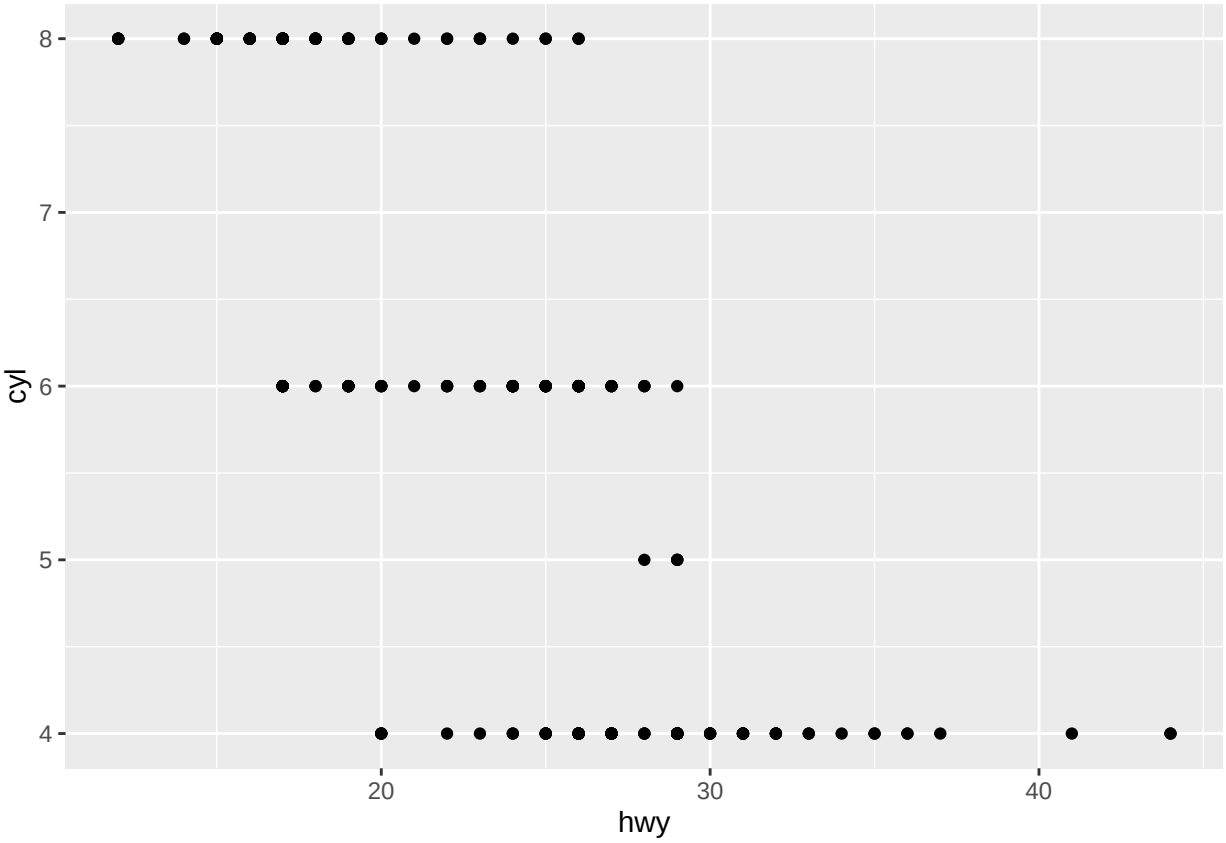
```
?mpg
```

`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