Topics in macroecon

LIXINYU|180e203e

22 06, 2018

```
suppressMessages(library("tidyverse"))
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
## Warning: package 'readr' was built under R version 3.3.2
```

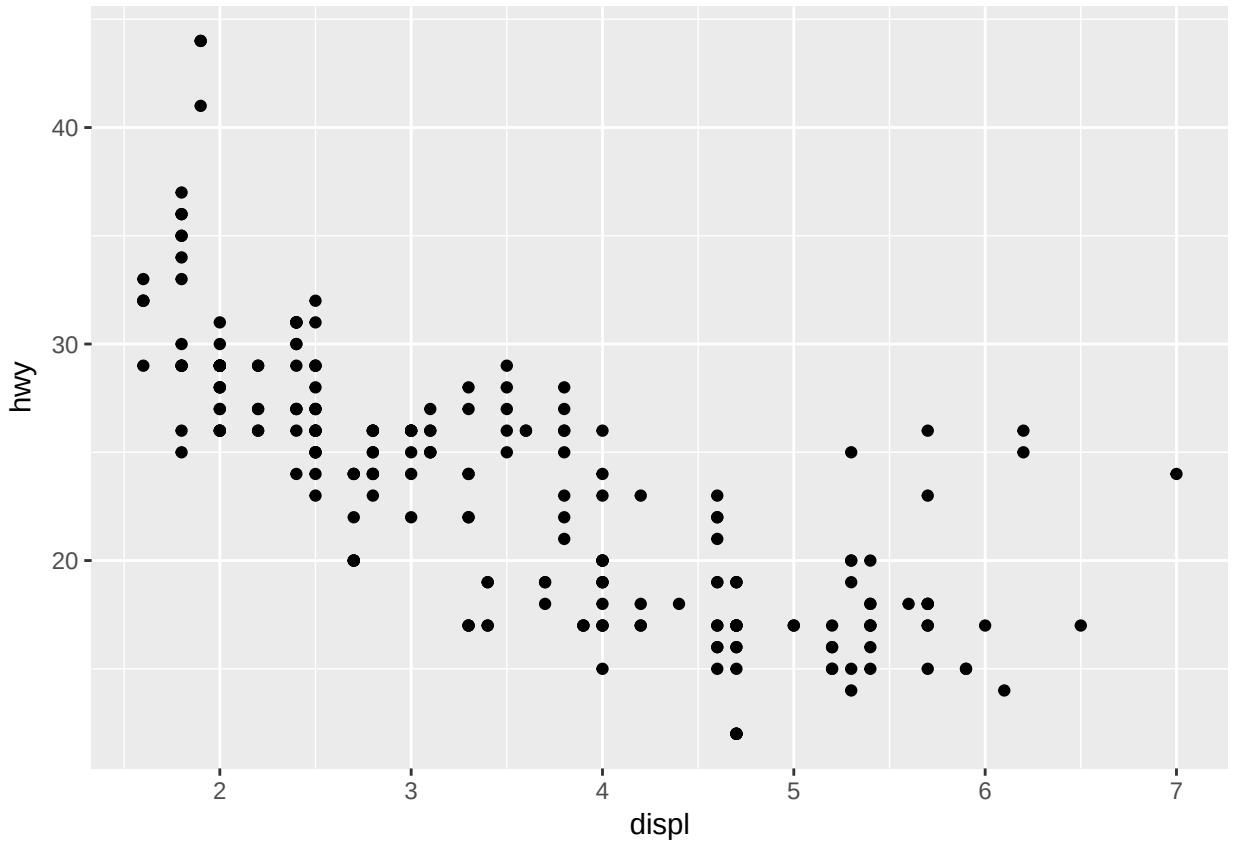
```
## Warning: package 'purrr' was built under R version 3.3.2
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
library("tidyverse")
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4        1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4        1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4        2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4        2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4        2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4        2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4        3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~  1.80  1999     4 manual~ 4        18    26 p
##  9 audi         a4 quat~  1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~  2.00  2008     4 manual~ 4        20    28 p
## # ... with 224 more rows, and 1 more variable: class <chr>
```

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



3.2.4 Excersice problem

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

```
nrow(mpg)
```

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

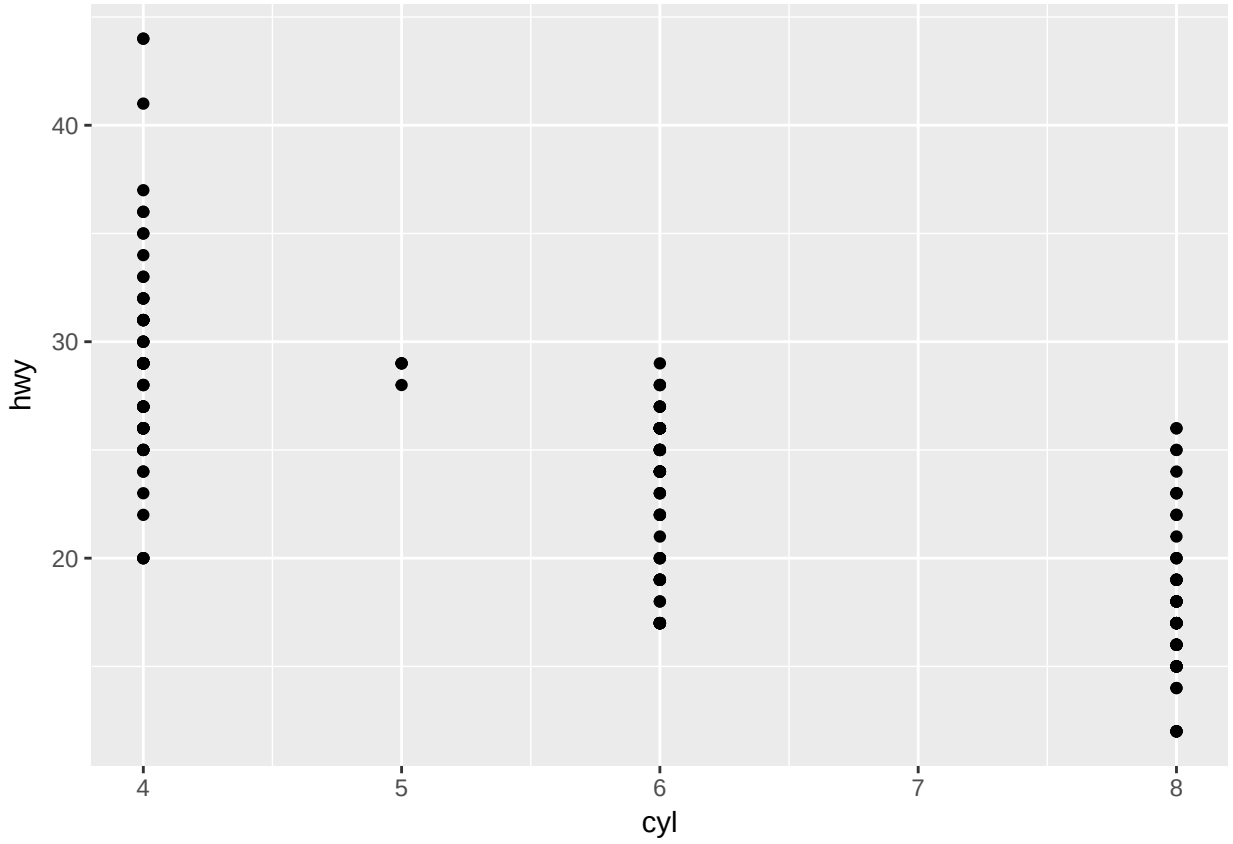
```
ncol(mpg)
```

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

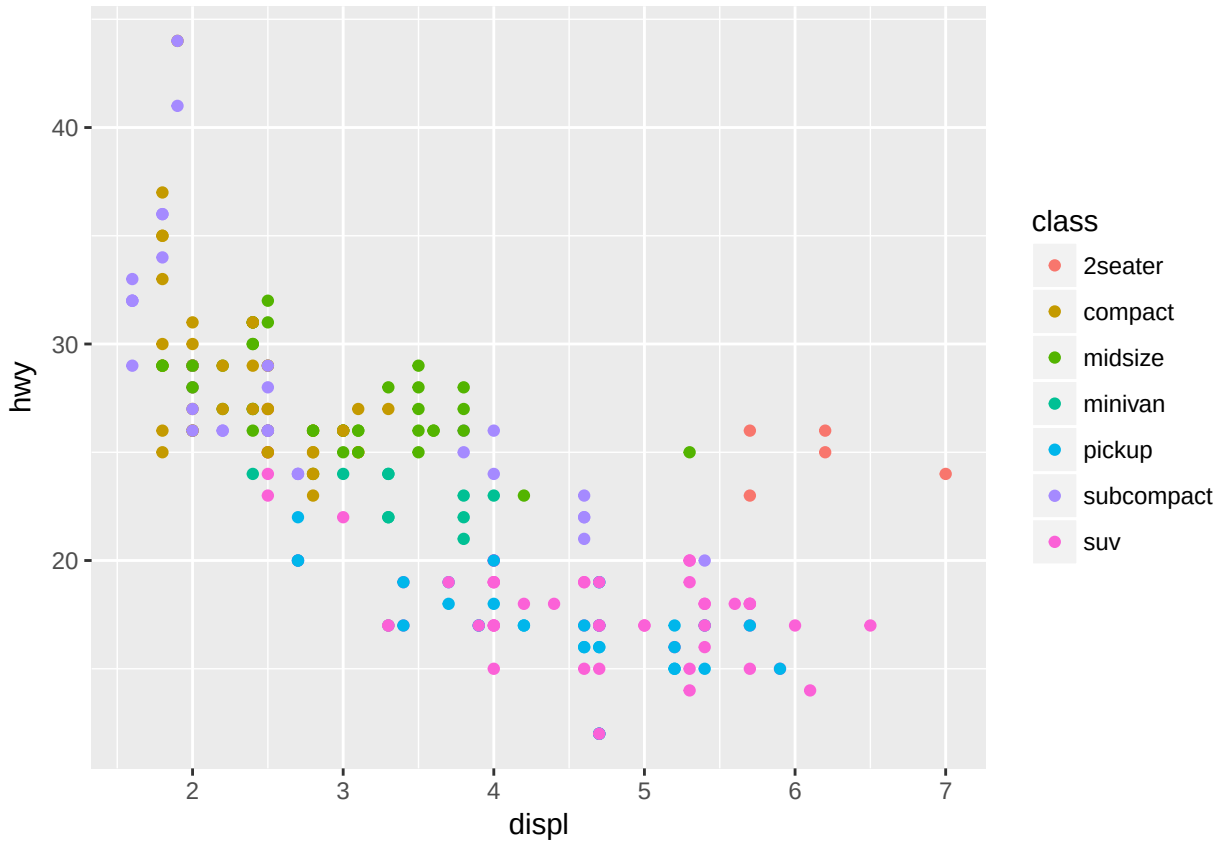
```
?mpg
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4        1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4        1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4        2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4        2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4        2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4        2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4        3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~  1.80  1999     4 manual~ 4        18    26 p
##  9 audi         a4 quat~  1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~  2.00  2008     4 manual~ 4        20    28 p
## # ... with 224 more rows, and 1 more variable: class <chr>
```

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

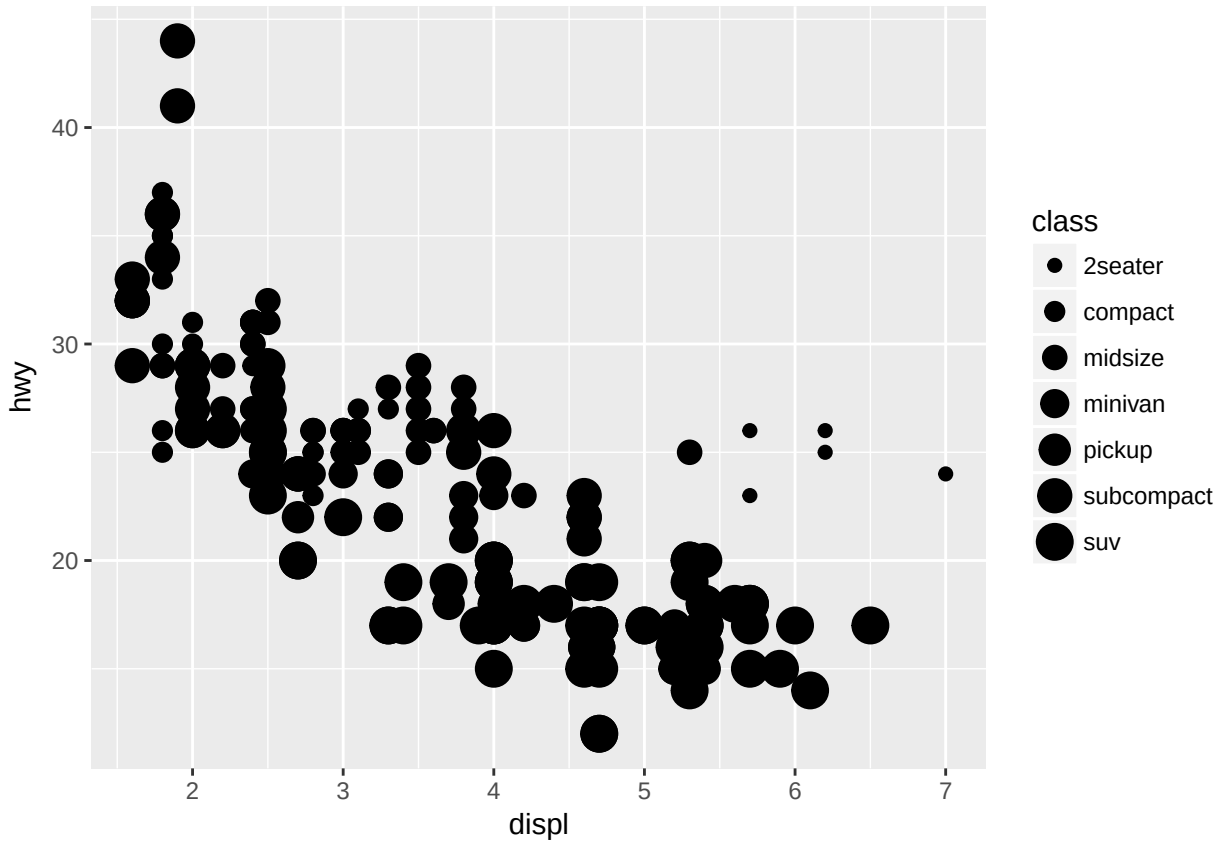



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class))
```

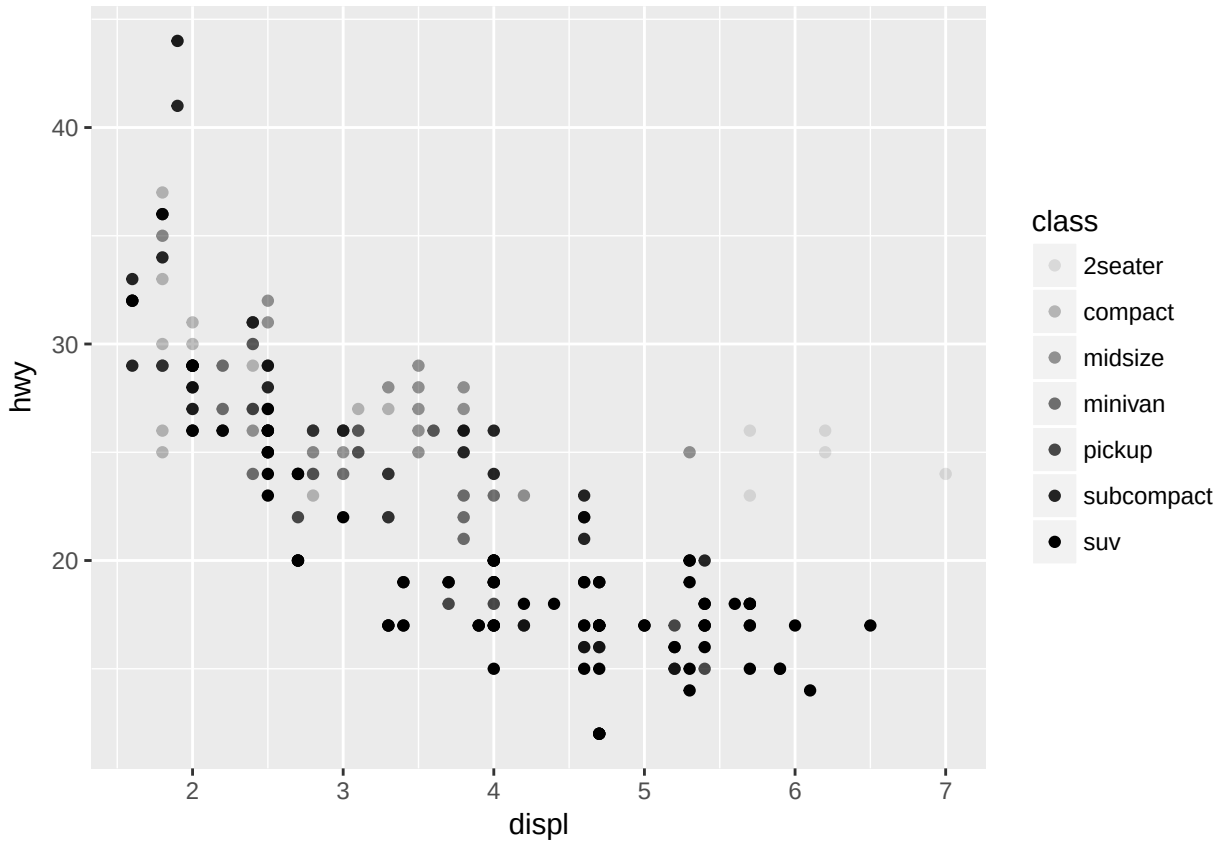



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = class))
```

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



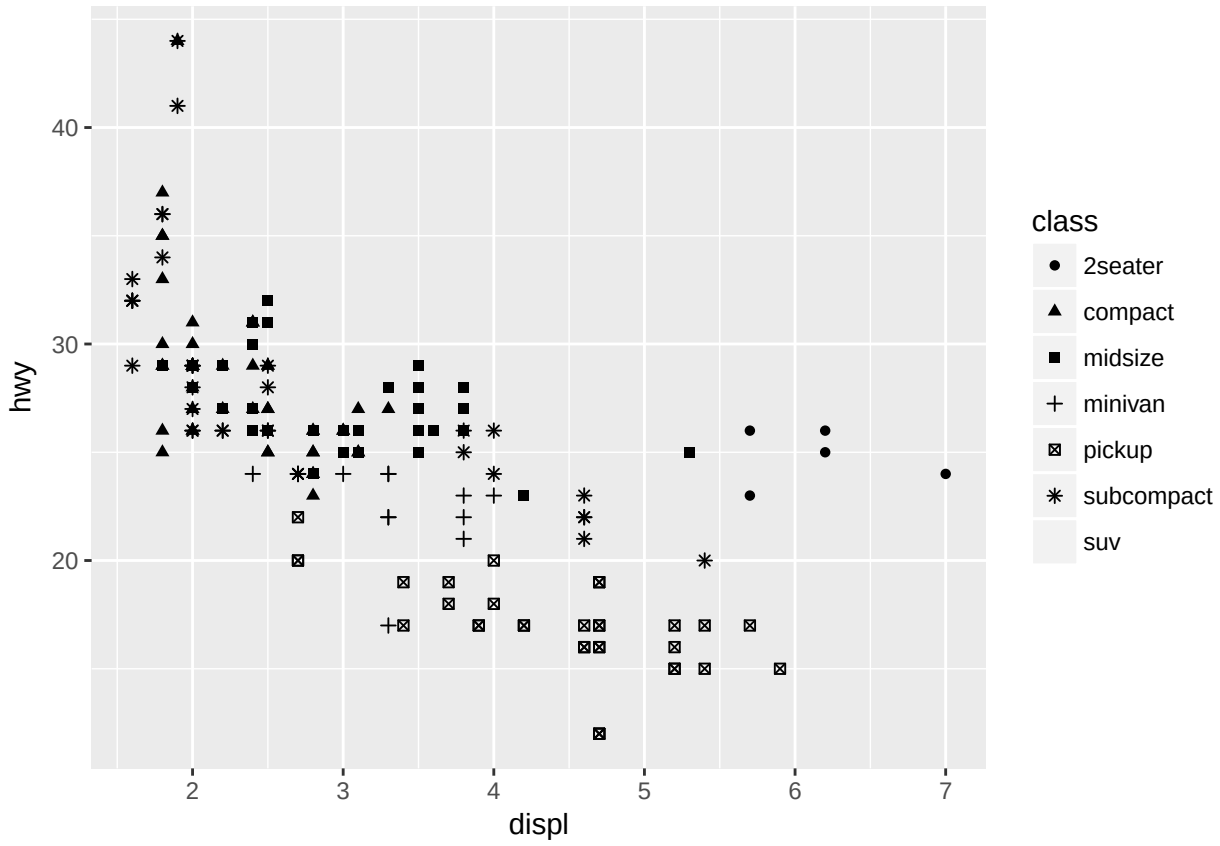

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class))
```



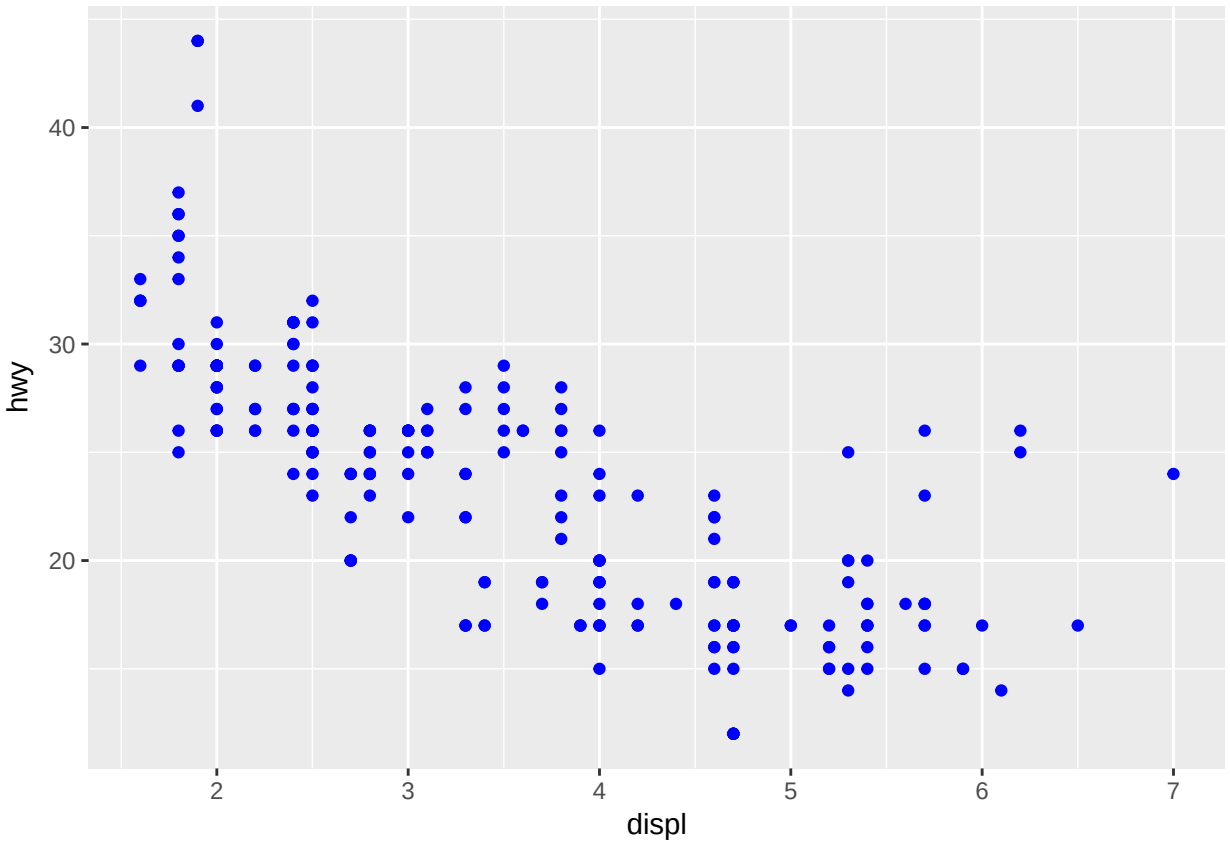

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = class))
```

```
## 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).
```




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

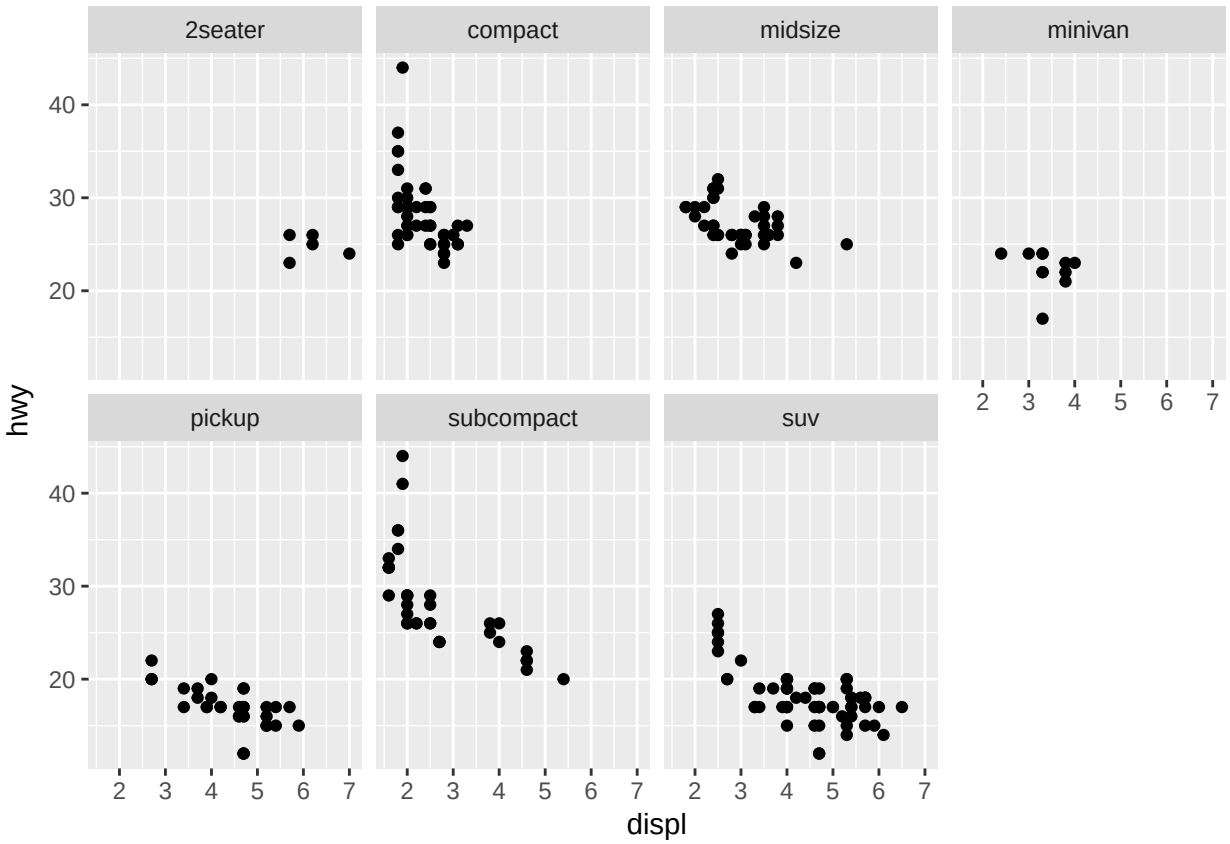


#3.3.1 Exercise problem#

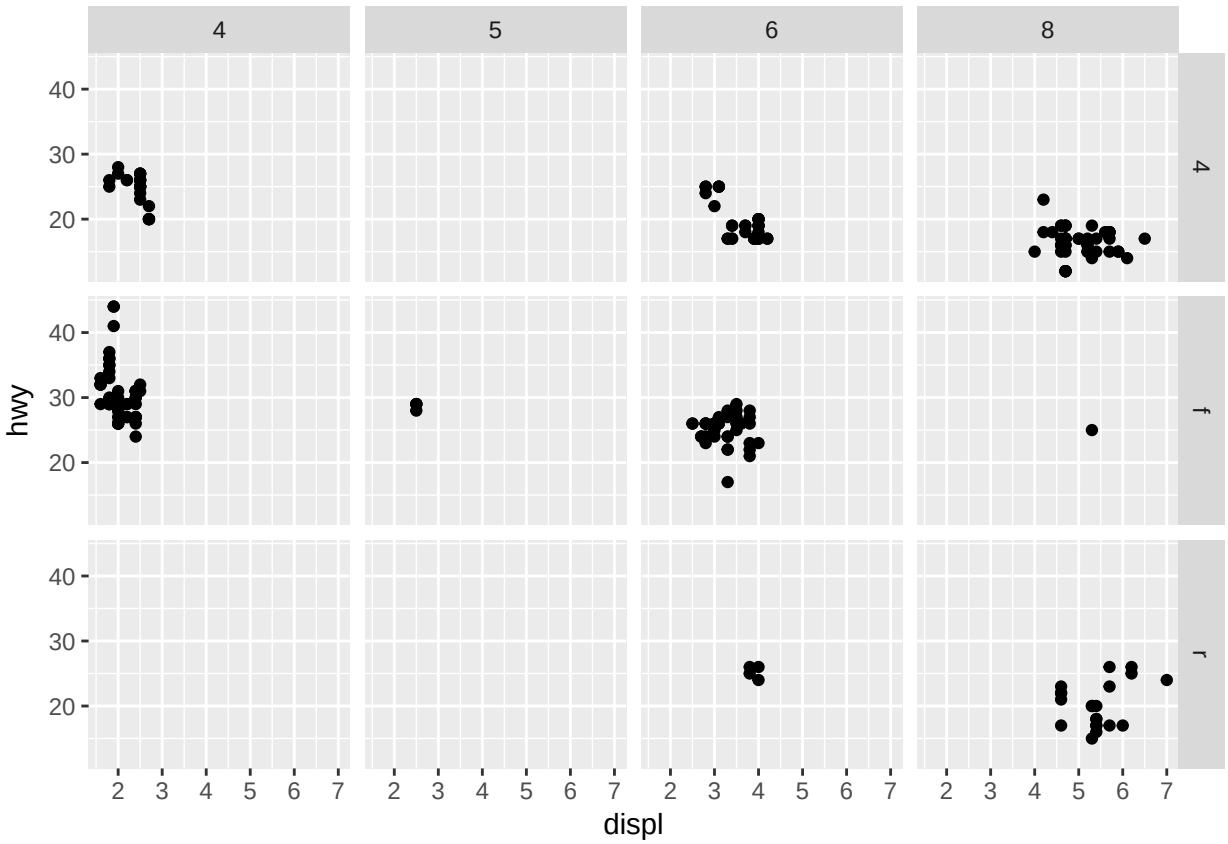
```
# In the first question, code:color='blue', supposed be outside of brackets#
```

3.5 Facets

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

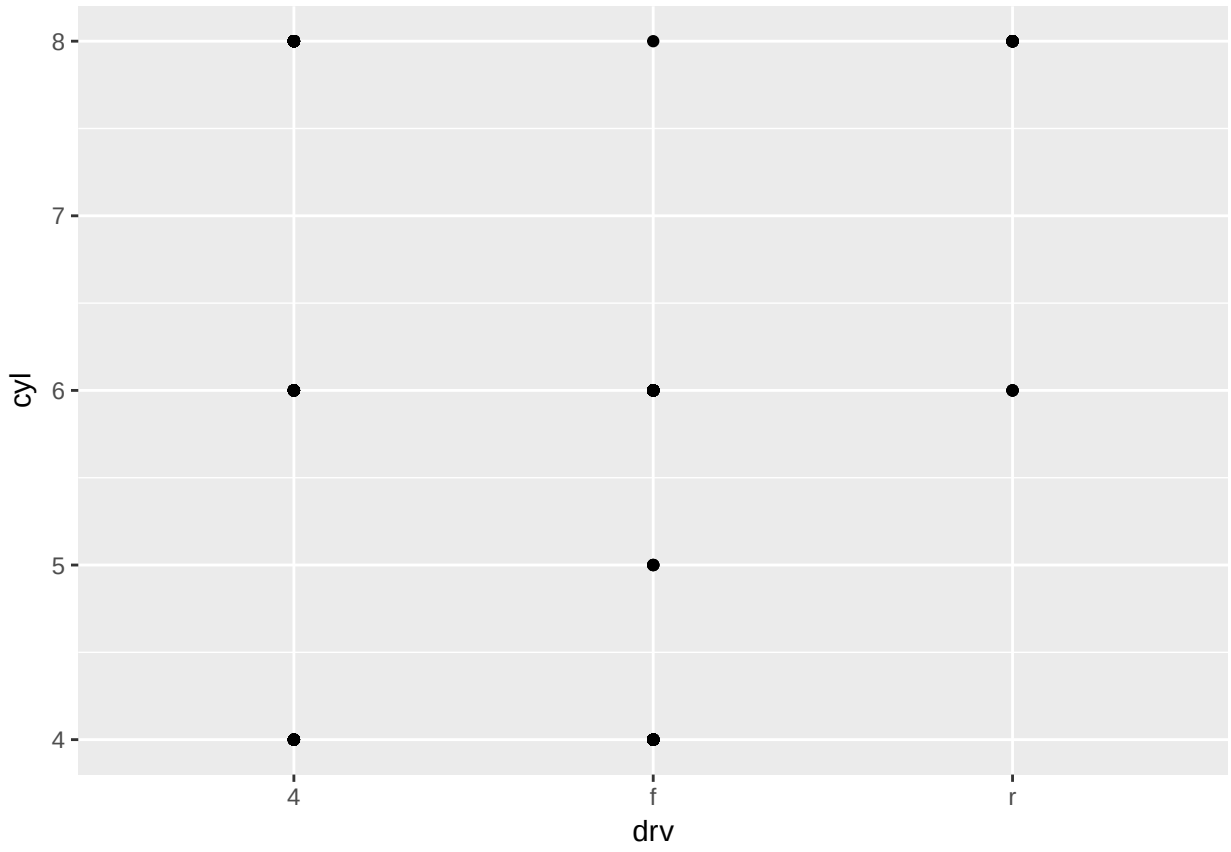



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ cyl)
```

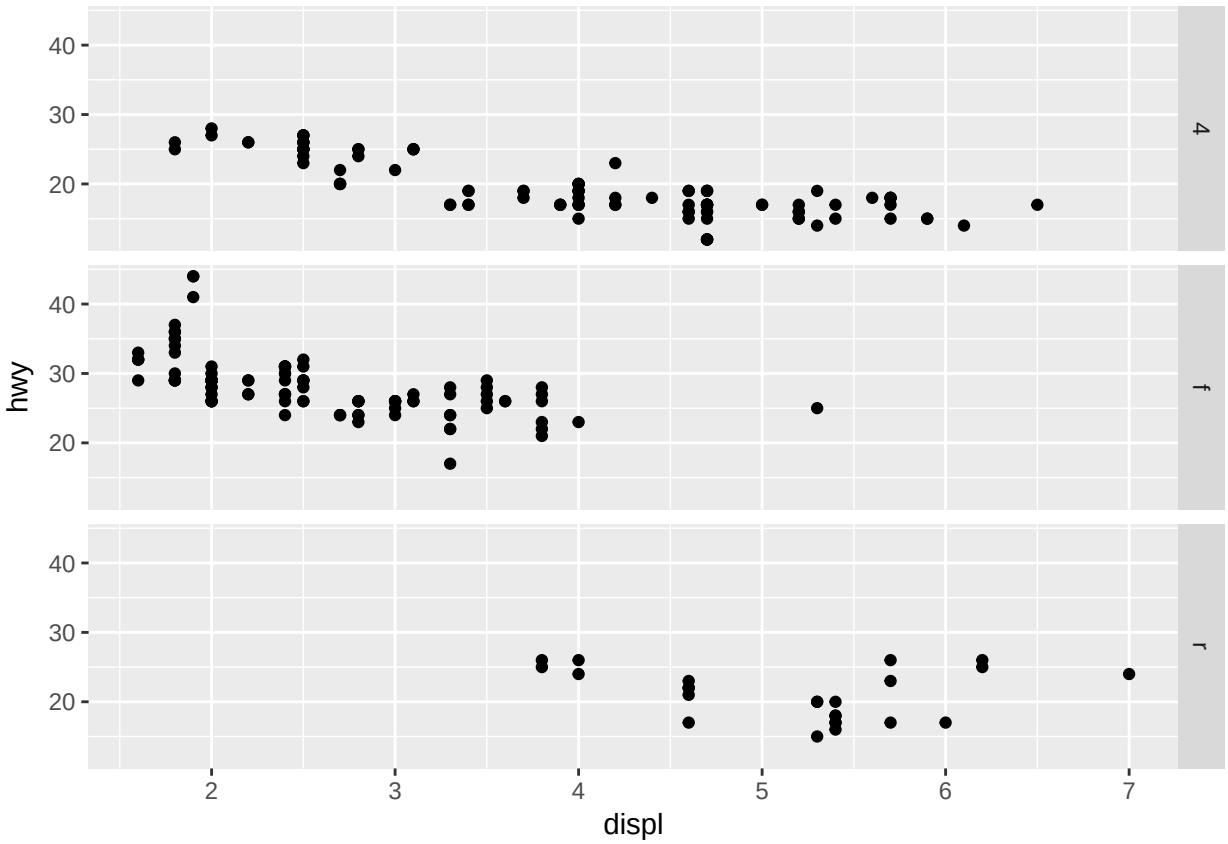


#3.5.1 EXercise#

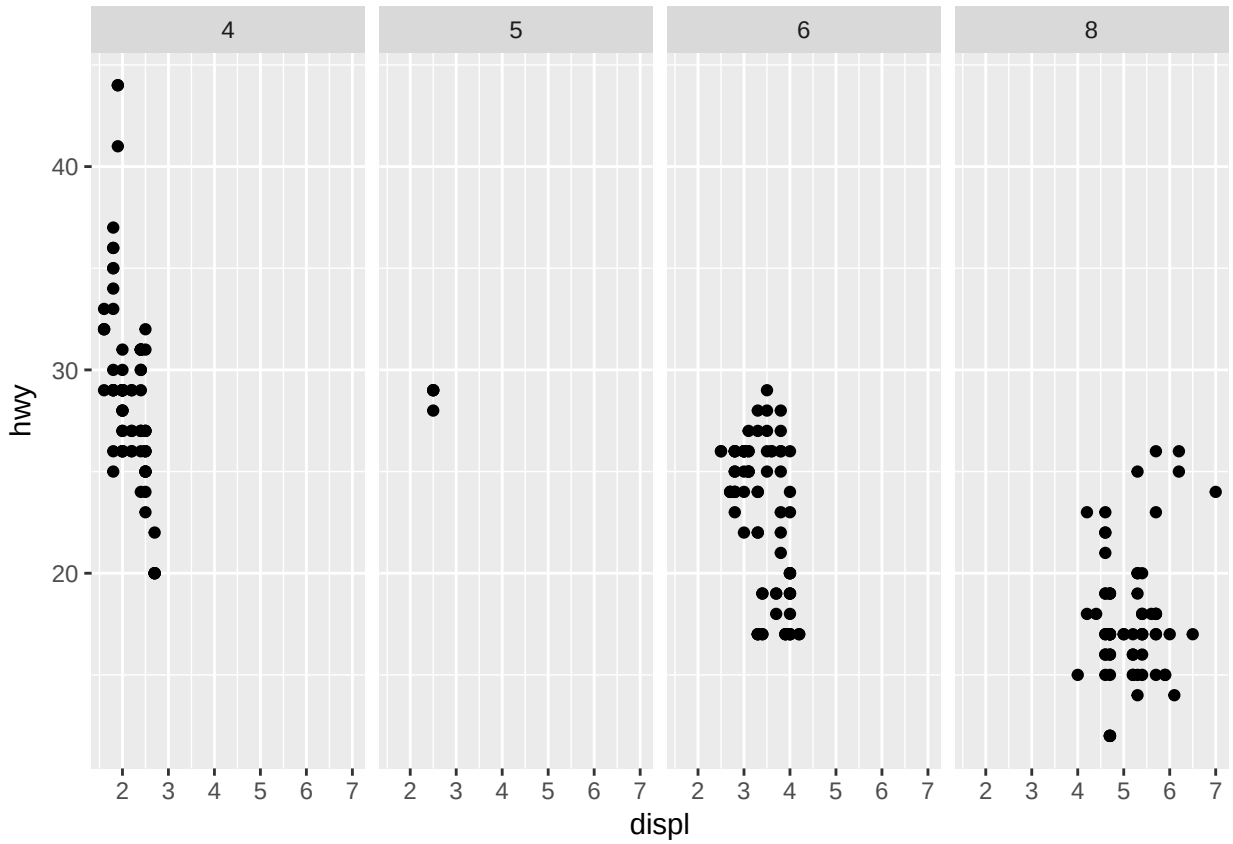
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))
```




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

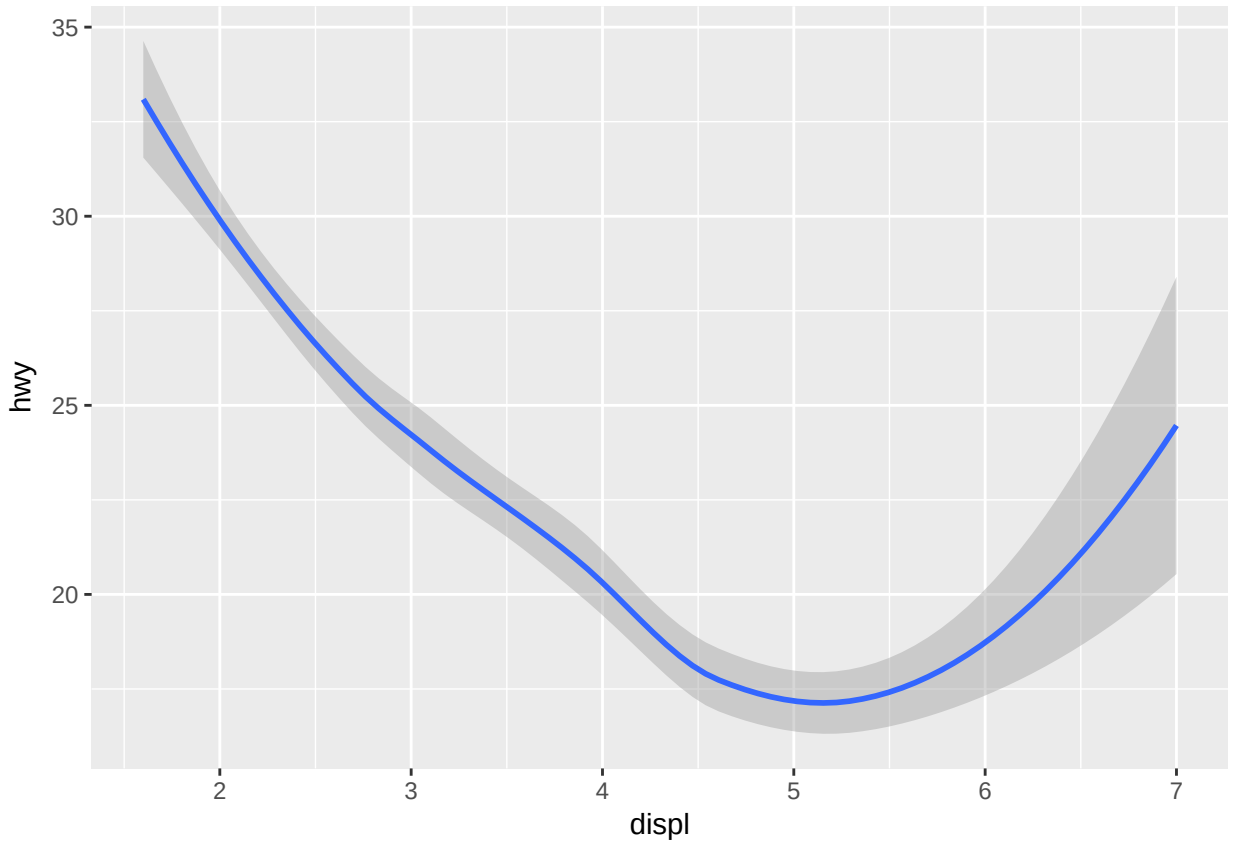



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```

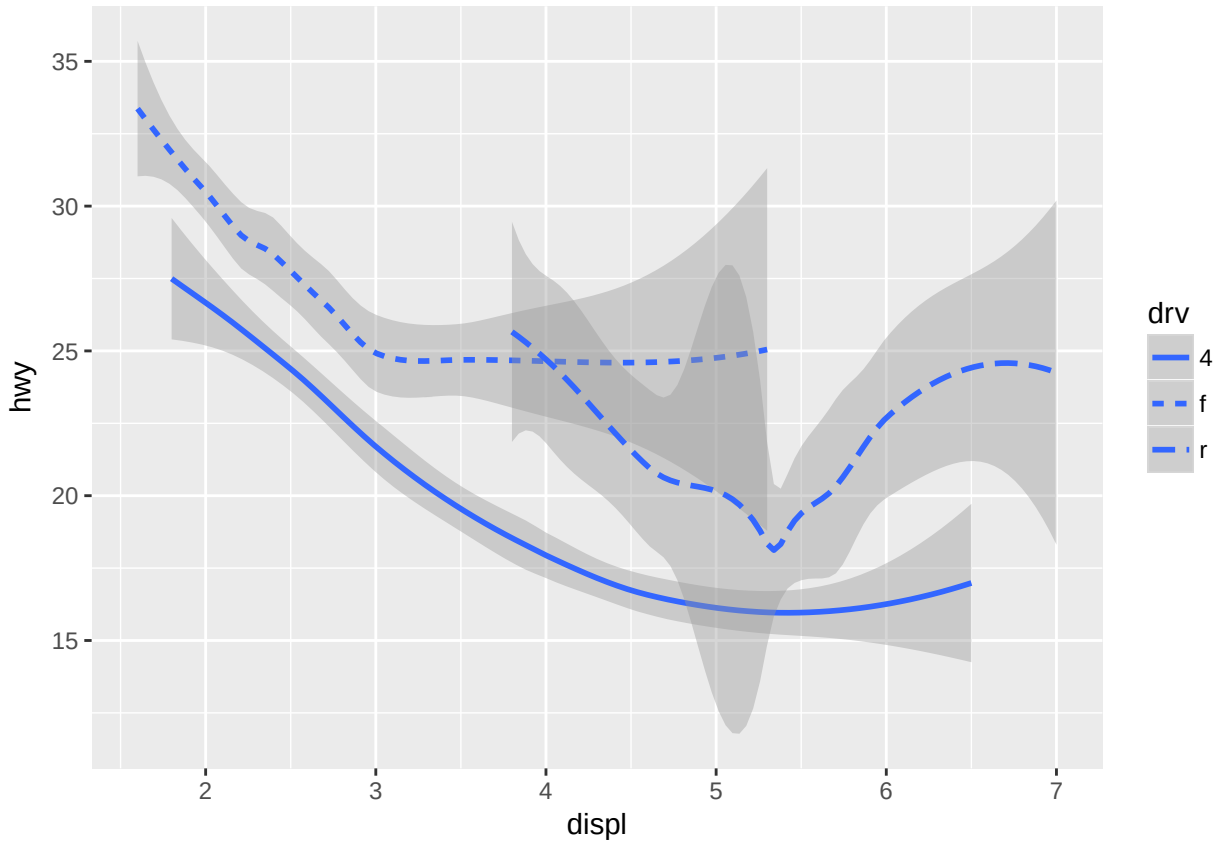


#3.6#

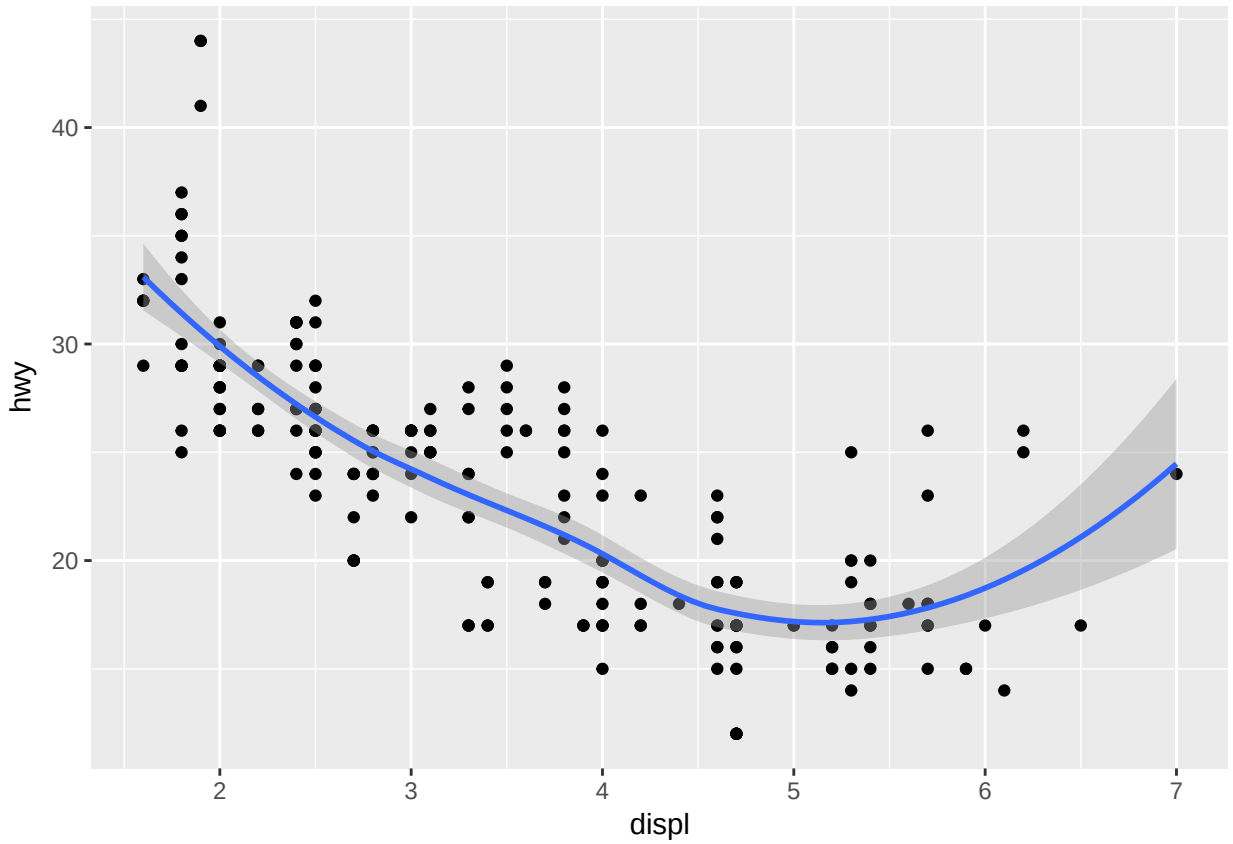
```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```




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

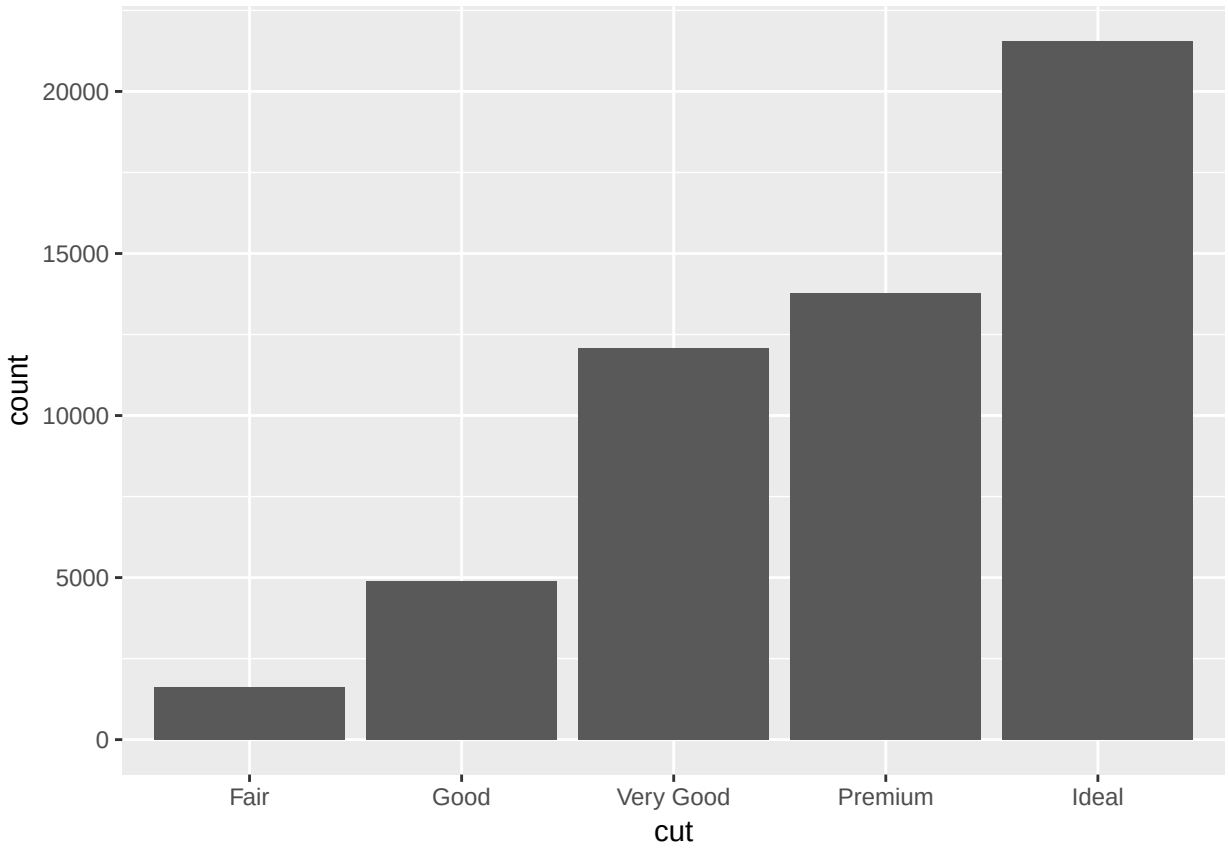



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

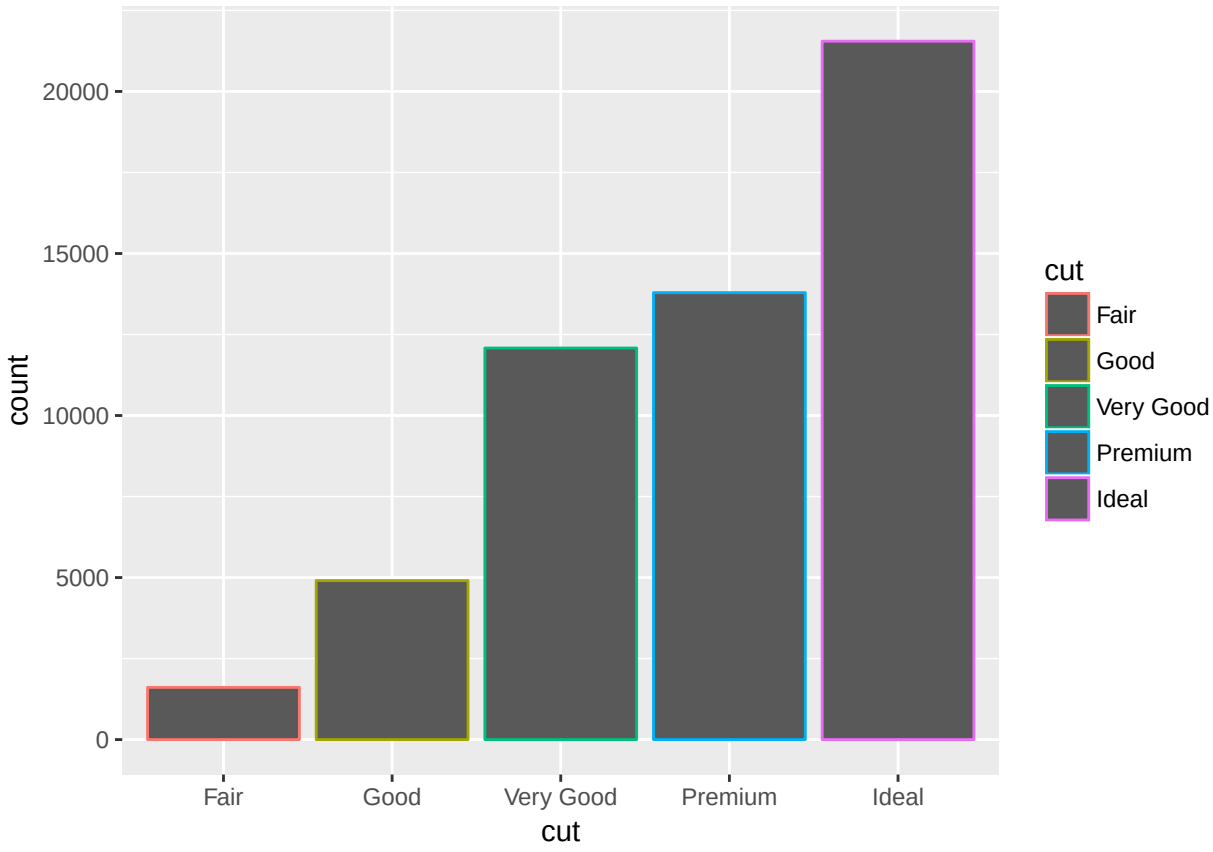


#3.7&3.8#

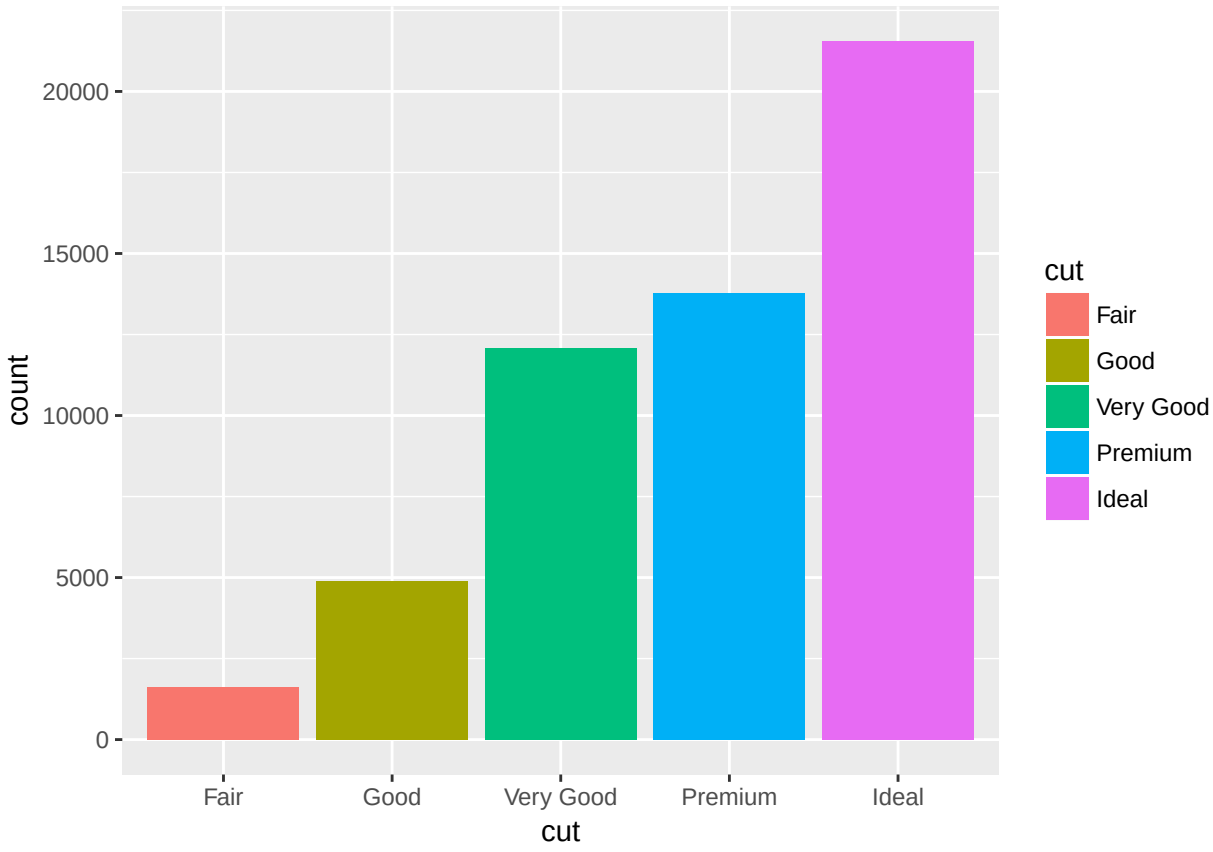
```
ggplot(data= diamonds)+ geom_bar(mapping = aes(x= cut))
```



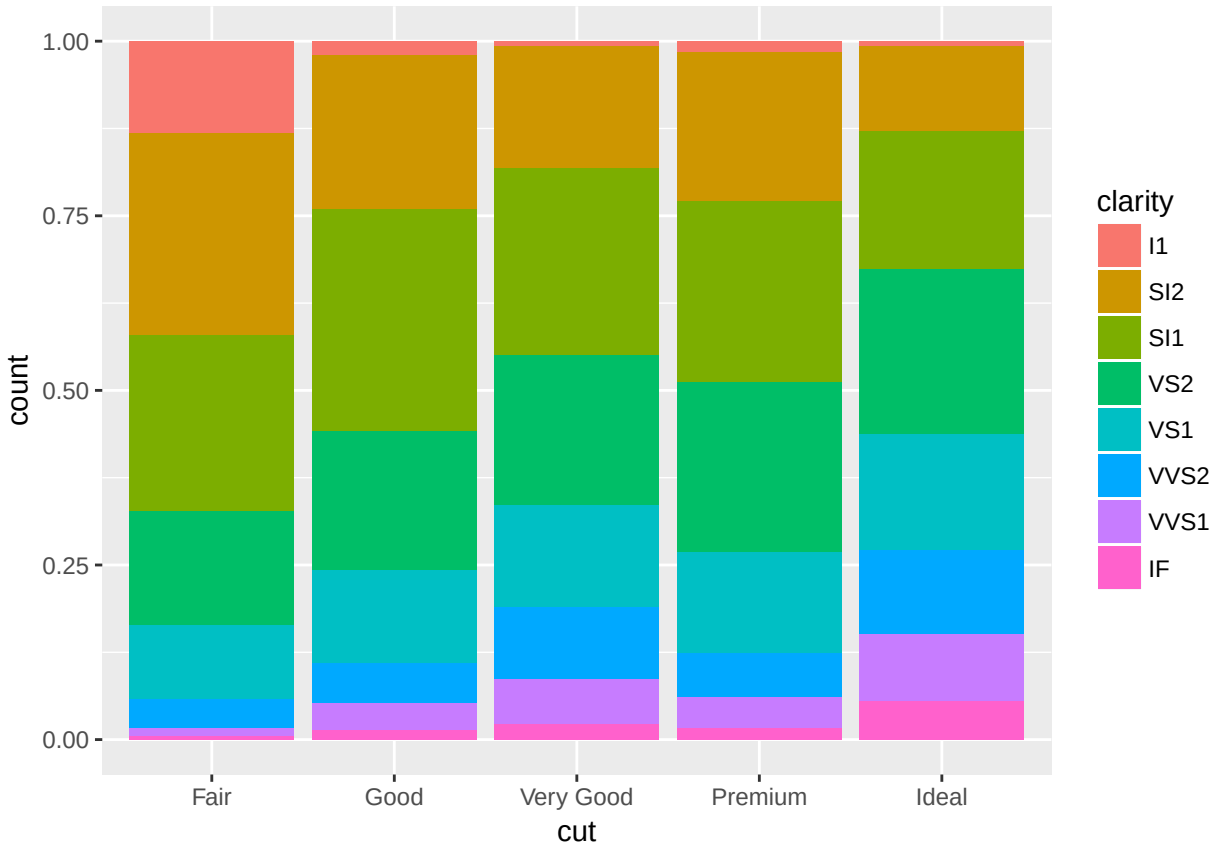

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, colour = cut))
```



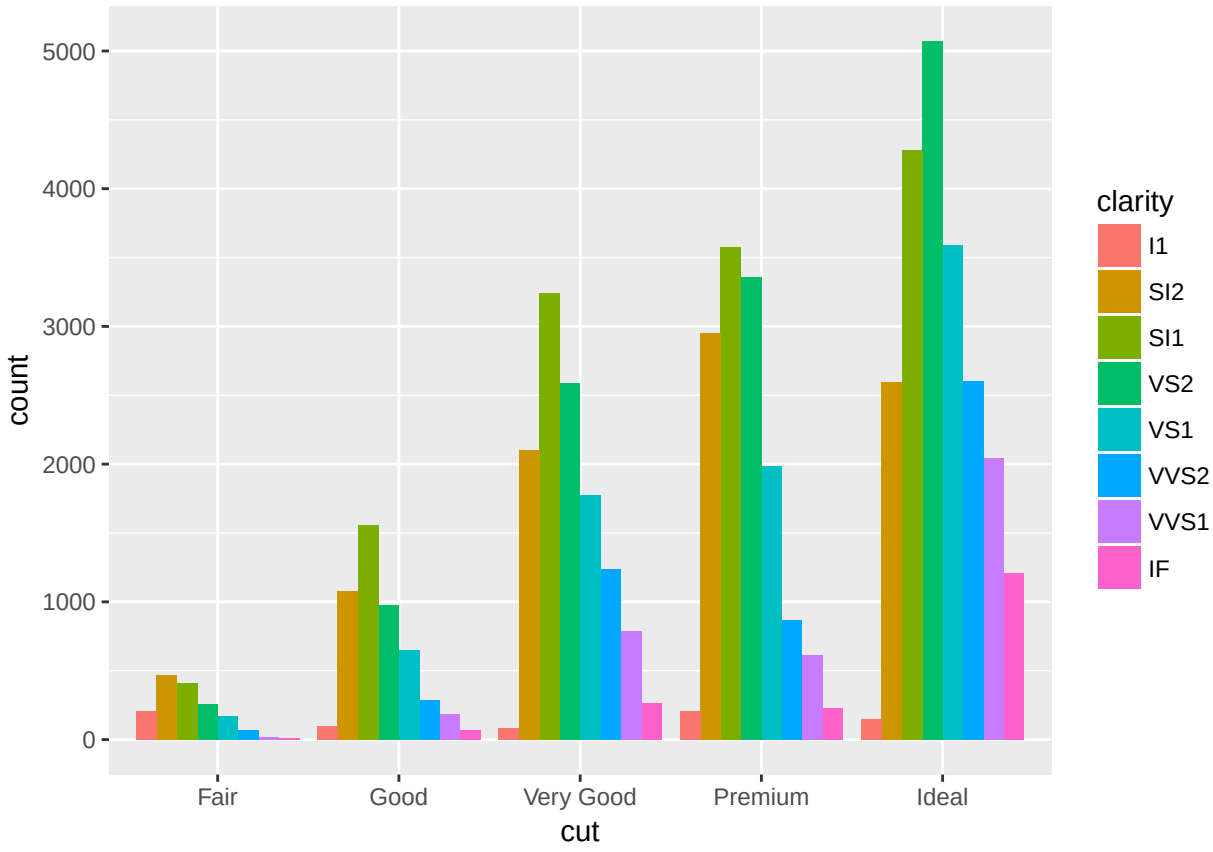

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = cut))
```




```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity), position = "fill")
```




```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity), position = "dodge")
```



#5#

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.3.2
```

```
nycflights13::flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515        2.      830
##  2  2013     1     1      533            529        4.      850
##  3  2013     1     1      542            540        2.      923
##  4  2013     1     1      544            545       -1.     1004
##  5  2013     1     1      554            600       -6.      812
##  6  2013     1     1      554            558       -4.      740
##  7  2013     1     1      555            600       -5.      913
##  8  2013     1     1      557            600       -3.      709
##  9  2013     1     1      557            600       -3.      838
## 10  2013     1     1      558            600       -2.      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
View(flights)
library(tidyverse)
```

5.1&5.2

```
View(flights)
filter(flights, month == 1, day == 1)
```

```
## # A tibble: 842 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515        2.      830
##  2  2013     1     1      533            529        4.      850
##  3  2013     1     1      542            540        2.      923
##  4  2013     1     1      544            545       -1.     1004
##  5  2013     1     1      554            600       -6.      812
##  6  2013     1     1      554            558       -4.      740
##  7  2013     1     1      555            600       -5.      913
##  8  2013     1     1      557            600       -3.      709
##  9  2013     1     1      557            600       -3.      838
## 10  2013     1     1      558            600       -2.      753
## # ... with 832 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
jan1 <- filter(flights, month == 1, day == 1)
(dec25 <- filter(flights, month == 12, day == 25))
```

```
## # A tibble: 719 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013    12    25      456            500       -4.      649
##  2  2013    12    25      524            515        9.      805
##  3  2013    12    25      542            540        2.      832
##  4  2013    12    25      546            550       -4.     1022
##  5  2013    12    25      556            600       -4.      730
##  6  2013    12    25      557            600       -3.      743
##  7  2013    12    25      557            600       -3.      818
##  8  2013    12    25      559            600       -1.      855
##  9  2013    12    25      559            600       -1.      849
## 10  2013    12    25      600            600        0.      850
## # ... with 709 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
filter(flights, month == 11 | month == 12)
```

```
## # A tibble: 55,403 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013    11     1        5           2359        6.      352
##  2  2013    11     1       35           2250      105.      123
##  3  2013    11     1      455            500       -5.      641
##  4  2013    11     1      539            545       -6.      856
##  5  2013    11     1      542            545       -3.      831
##  6  2013    11     1      549            600      -11.      912
##  7  2013    11     1      550            600      -10.      705
##  8  2013    11     1      554            600       -6.      659
##  9  2013    11     1      554            600       -6.      826
## 10  2013    11     1      554            600       -6.      749
## # ... with 55,393 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
nov_dec <- filter(flights, month %in% c(11, 12))
```

5.2.4 Excercise

```
filter(flights, arr_delay >= 120, dep_delay >= 120)
```

```
## # A tibble: 8,482 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      848           1835      853.     1001
##  2  2013     1     1      957            733      144.     1056
##  3  2013     1     1     1114            900      134.     1447
##  4  2013     1     1     1815           1325      290.     2120
##  5  2013     1     1     1842           1422      260.     1958
##  6  2013     1     1     1856           1645      131.     2212
##  7  2013     1     1     1934           1725      129.     2126
##  8  2013     1     1     1938           1703      155.     2109
##  9  2013     1     1     1942           1705      157.     2124
## 10  2013     1     1     2006           1630      216.     2230
## # ... with 8,472 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

5.3

```r
arrange(flights, year, month, day)
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515        2.      830
##  2  2013     1     1      533            529        4.      850
##  3  2013     1     1      542            540        2.      923
##  4  2013     1     1      544            545       -1.     1004
##  5  2013     1     1      554            600       -6.      812
##  6  2013     1     1      554            558       -4.      740
##  7  2013     1     1      555            600       -5.      913
##  8  2013     1     1      557            600       -3.      709
##  9  2013     1     1      557            600       -3.      838
## 10  2013     1     1      558            600       -2.      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```r
arrange(flights, desc(arr_delay))
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     9      641            900    1301.      1242
##  2  2013     6    15     1432           1935    1137.      1607
##  3  2013     1    10     1121           1635    1126.      1239
##  4  2013     9    20     1139           1845    1014.      1457
##  5  2013     7    22      845           1600    1005.      1044
##  6  2013     4    10     1100           1900     960.      1342
##  7  2013     3    17     2321            810     911.       135
##  8  2013     7    22     2257            759     898.       121
##  9  2013    12     5      756           1700     896.      1058
## 10  2013     5     3     1133           2055     878.      1250
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

11 Data import

```r
library(tidyverse)
heights <- read_csv("heights.csv")

read_csv("The first line of metadata
  The second line of metadata
  x,y,z
  1,2,3", skip = 2)
```

```
## # A tibble: 1 x 3
##       x     y     z
##   <int> <int> <int>
## 1     1     2     3
```

```
read_csv("# A comment I want to skip
  x,y,z
  1,2,3", comment = "#")
```

```
## # A tibble: 1 x 3
##       x     y     z
##   <int> <int> <int>
## 1     1     2     3
```

```
read_csv("1,2,3\n4,5,6", col_names = c("x", "y", "z"))
```

```
## # A tibble: 2 x 3
##       x     y     z
##   <int> <int> <int>
## 1     1     2     3
## 2     4     5     6
```

```
read_csv("a,b,c\n1,2,.", na = ".")
```

```
## # A tibble: 1 x 3
##       a     b c
##   <int> <int> <chr>
## 1     1     2 <NA>
```

11.2.2 Excercise

```
#The first three of questions, they all got a problem with Value of column names should be indentical t
```

11.3

```
str(parse_logical(c("TRUE", "FALSE", "NA")))
```

```
##  logi [1:3] TRUE FALSE NA
```

```
str(parse_integer(c("1", "2", "3")))
```

```
##  int [1:3] 1 2 3
```

```r
str(parse_date(c("2010-01-01", "1979-10-14")))
```

```
##  Date[1:2], format: "2010-01-01" "1979-10-14"
```

```r
parse_integer(c("1", "231", ".", "456"), na = ".")
```

```
## [1]   1 231  NA 456
```

```r
x <- parse_integer(c("123", "345", "abc", "123.45"))
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 2 parsing failures.
## row # A tibble: 2 x 4 col     row   col expected               actual expected   <int> <int> <chr>
```

```r
x
```

```
## [1] 123 345  NA  NA
## attr(,"problems")
## # A tibble: 2 x 4
##     row   col expected               actual
##   <int> <int> <chr>                  <chr>
## 1     3    NA an integer             abc
## 2     4    NA no trailing characters .45
```

```r
problems(x)
```

```
## # A tibble: 2 x 4
##     row   col expected               actual
##   <int> <int> <chr>                  <chr>
## 1     3    NA an integer             abc
## 2     4    NA no trailing characters .45
```

```r
parse_double("1.23")
```

```
## [1] 1.23
```

```r
parse_double("1,23", locale = locale(decimal_mark = ","))
```

```
## [1] 1.23
```

```r
parse_number("$100")
```

```
## [1] 100
```

```
parse_number("20%")
```

```
## [1] 20
```

```
parse_number("It cost $123.45")
```

```
## [1] 123.45
```

11.3.4 Dates, date-times, and times

```
parse_datetime("2010-10-01T2010")
```

```
## [1] "2010-10-01 20:10:00 UTC"
```

```
parse_datetime("20101010")
```

```
## [1] "2010-10-10 UTC"
```

```
parse_date("01/02/15","%m/%d/%y")
```

```
## [1] "2015-01-02"
```

```
parse_date("01/02/15","%d/%m/%y")
```

```
## [1] "2015-02-01"
```

```
parse_date("01/02/15","%y/%m/%d")
```

```
## [1] "2001-02-15"
```

```
library(hms)
parse_time("01:10 am")
```

```
## 01:10:00
```

```
parse_time("20:10:01")
```

```
## 20:10:01
```

11.4.2

```
challenge <- read_csv(readr_example("challenge.csv"))
```

```
## Parsed with column specification:
## cols(
##   x = col_integer(),
##   y = col_character()
## )
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 1000 parsing failures.
## row # A tibble: 5 x 5 col     row col   expected               actual             file
## ... ................. ... ..................................................................
## See problems(...) for more details.
```

```
problems(challenge)
```

```
## # A tibble: 1,000 x 5
##      row col   expected               actual             file
##    <int> <chr> <chr>                  <chr>              <chr>
##  1  1001 x     no trailing characters .23837975086644292 '/home2/180e203e~
##  2  1002 x     no trailing characters .41167997173033655 '/home2/180e203e~
##  3  1003 x     no trailing characters .7460716762579978  '/home2/180e203e~
##  4  1004 x     no trailing characters .723450553836301   '/home2/180e203e~
##  5  1005 x     no trailing characters .614524137461558   '/home2/180e203e~
##  6  1006 x     no trailing characters .473980569280684   '/home2/180e203e~
##  7  1007 x     no trailing characters .5784610391128808  '/home2/180e203e~
##  8  1008 x     no trailing characters .2415937229525298  '/home2/180e203e~
##  9  1009 x     no trailing characters .11437866208143532 '/home2/180e203e~
## 10  1010 x     no trailing characters .2983446326106787  '/home2/180e203e~
## # ... with 990 more rows
```

```
challenge <- read_csv(
  readr_example("challenge.csv"),
  col_types = cols(
    x = col_integer(),
    y = col_character()
  )
)
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not a multiple of vector leng
```

```
## Warning in rbind(names(probs), probs_f): 1000 parsing failures.
## row # A tibble: 5 x 5 col     row col   expected               actual             file
## ... ................. ... ..................................................................
## See problems(...) for more details.
```

```
write_csv(challenge, "challenge.csv")
```

12

```
df <- tibble::tibble(
  a = rnorm(10),
  b = rnorm(10),
  c = rnorm(10),
  d = rnorm(10)
)

df$a <- (df$a - min(df$a, na.rm = TRUE)) /
  (max(df$a, na.rm = TRUE) - min(df$a, na.rm = TRUE))
df$b <- (df$b - min(df$b, na.rm = TRUE)) /
  (max(df$b, na.rm = TRUE) - min(df$b, na.rm = TRUE))
df$c <- (df$c - min(df$c, na.rm = TRUE)) /
  (max(df$c, na.rm = TRUE) - min(df$c, na.rm = TRUE))
df$d <- (df$d - min(df$d, na.rm = TRUE)) /
  (max(df$d, na.rm = TRUE) - min(df$d, na.rm = TRUE))
```

19

```
y<-2
if (y < 0 && debug) {
  message("Y is negative")
}

if (y == 0) {
  log(x)
} else {
  y ^ x
}
```

```
## [1]  1.063382e+37 7.167183e+103            NA            NA
## attr(,"problems")
## # A tibble: 2 x 4
##     row   col expected               actual
##   <int> <int> <chr>                  <chr>
## 1     3    NA an integer             abc
## 2     4    NA no trailing characters .45
```

selecet

```
rm(list = ls())
library(tidyverse)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:hms':
##
##     hms

## The following object is masked from 'package:base':
##
##     date
```

```
data_00 <- read_fwf(file="data_00.dat",
                    fwf_cols(year      = c(1, 4),
                    serial    = c(5,9),
                    month     = c(10,11),
                             hwtfinl   = c(12,21),
                             cpsid     = c(22,35),
                             asecflag  = c(36,36),
                             hflag     = c(37,37),
                             asecwth   = c(38,47),
                             pernum    = c(48,49),
                             wtfinl    = c(50,63),
                             cpsidp    = c(64,77),
                             asecwt    = c(78,87),
                             age       = c(88,89),
                             sex       = c(90,90),
                             race      = c(91,93),
                             educ      = c(94,96),
                             schlcoll  = c(97,97),
                             indly     = c(98,101),
                             classwly  = c(102,103),
                             wkswork1  = c(104,105),
                             wkswork2  = c(106,106),
                             fullpart  = c(107,107),
                             incwage   = c(108,114)),
                  col_types = cols(year       = "i",
                                   serial     = "n",
                                   month      = "i",
                                   hwtfinl    = "d",
                                   cpsid      = "d",
                                   asecflag   = "i",
                                   hflag      = "i",
                                   asecwth    = "d",
                                   pernum     = "i",
                                   wtfinl     = "d",
                                   cpsidp     = "d",
                                   asecwt     = "d",
                                   age        = "i",
                                   sex        = "i",
                                   race       = "i",
                                   educ       = "i",
                                   schlcoll   = "i",
                                   indly      = "i",
                                   classwly   = "i",
                                   wkswork1   = "i",
                                   wkswork2   = "i",
```

```
                                    fullpart   = "i",
                                    incwage    = "n"))
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 22296336 parsing failures.
## row # A tibble: 5 x 5 col     row col     expected   actual                                file
## ... ................. ... ..........................................................................
## See problems(...) for more details.
```

```
data_00$hwtfinl = data_00$hwtfinl/10000
data_00$wtfinl = data_00$wtfinl/10000
data_00$asecwt = data_00$asecwt/10000
```

merge cpi data (see Acemoglu and Autor's Data Appendix)

```
data_cpi <- read_csv(file = "data_cpi.csv", col_names = c("year","cpi"), col_types=cols(year = "D", cpi
data_cpi$year <- year(data_cpi$year)
data_cpi <- data_cpi %>%
  mutate(price_1982 = ifelse(year == 1982, cpi, 0)) %>% # the base year is 1982 (see Acemoglu and Autor
  mutate(price_1982 = max(price_1982)) %>%
  mutate(cpi = cpi/price_1982) %>%
  select(year, cpi)
data_00 <- data_00 %>%
  left_join(data_cpi, by = "year")
data_00 <- data_00 %>%
  mutate(educ = ifelse(educ == 999, NA, educ)) %>%
  mutate(classwly = ifelse(classwly == 99, NA, classwly)) %>%
  mutate(wkswork2 = ifelse(wkswork2 == 999, NA, wkswork2)) %>%
  mutate(incwage = ifelse(incwage == 9999999 | incwage == 9999998, NA, incwage)) %>%
  mutate(race = ifelse(race == 999, NA, race))
data_00 <- data_00 %>%
  mutate(wkswork = ifelse(year >= 1976, wkswork1, NA)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 1, 7, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 2, 20, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 3, 33, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 4, 43.5, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 5, 48.5, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 6, 51, wkswork))
data_00 <- data_00 %>%
  group_by(year) %>%
  mutate(top_incwage = max(incwage, na.rm = TRUE)) %>%
  mutate(incwage = ifelse(incwage == top_incwage, 1.45*incwage, incwage)) %>%
  ungroup()
data_00 <- data_00 %>%
  mutate(rwage = incwage/cpi/wkswork) %>%
  mutate(lrwage = log(rwage))
data_00 <- data_00 %>%
  mutate(dfemale = (sex == 2)) # female
```

```
data_00 <- data_00 %>%
  mutate(deduc_1 = ifelse(educ < 70, 1, 0)) %>%                  # highshool dropout
  mutate(deduc_2 = ifelse(educ >= 80 & educ < 110, 1, 0)) %>%  # some college
  mutate(deduc_3 = ifelse(educ >= 110 & educ < 123, 1, 0)) %>% # 4 years college
  mutate(deduc_4 = ifelse(educ >= 123, 1, 0))                    # more than college
data_00 <- data_00 %>%
  mutate(drace_1 = (race == 200)) %>% # black
  mutate(drace_2 = (race > 200)) # nonwhite other
```

create experience variable: check the IPUMS website for variable definition

```
data_00 <- data_00 %>%
  mutate(exp = ifelse(educ == 10, age - 8.5, NA)) %>%
  mutate(exp = ifelse(educ == 11, age - 7, exp)) %>%
  mutate(exp = ifelse(educ == 12, age - 8, exp)) %>%
  mutate(exp = ifelse(educ == 13, age - 9, exp)) %>%
  mutate(exp = ifelse(educ == 14, age - 10, exp)) %>%
  mutate(exp = ifelse(educ == 20, age - 11.5, exp)) %>%
  mutate(exp = ifelse(educ == 21, age - 11, exp)) %>%
  mutate(exp = ifelse(educ == 22, age - 12, exp)) %>%
  mutate(exp = ifelse(educ == 30, age - 13.5, exp)) %>%
  mutate(exp = ifelse(educ == 31, age - 13, exp)) %>%
  mutate(exp = ifelse(educ == 32, age - 14, exp)) %>%
  mutate(exp = ifelse(educ == 40, age - 15, exp)) %>%
  mutate(exp = ifelse(educ == 50, age - 16, exp)) %>%
  mutate(exp = ifelse(educ == 60, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 70, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 71, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 72, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 73, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 80, age - 19, exp)) %>%
  mutate(exp = ifelse(educ == 81, age - 19, exp)) %>%
  mutate(exp = ifelse(educ == 90, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 91, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 92, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 100, age - 21, exp)) %>%
  mutate(exp = ifelse(educ == 110, age - 22, exp)) %>%
  mutate(exp = ifelse(educ == 111, age - 22, exp)) %>%
  mutate(exp = ifelse(educ == 120, age - 23.5, exp)) %>%
  mutate(exp = ifelse(educ == 121, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 122, age - 24, exp)) %>%
  mutate(exp = ifelse(educ == 123, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 124, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 125, age - 27, exp))
```

sample selection (see Katz and Murphy (1992) and Acemoglu and Autor (2011)'s Data Appendix)

```
data_00 <- data_00 %>%
  filter(rwage >= 67) %>%                                                    # real wage more than 6
  filter(age >= 16 & age <= 64) %>%                                          # age equal or above 16
  filter(fullpart == 1) %>%                                                  # work more than 35 hou
  filter(wkswork >= 40) %>%                                                  # work more than 40 wee
  filter(classwly != 10 | classwly != 13 | classwly != 14) %>%               # not self-employed
  filter(!((year >= 1992 & year <= 2002) & (indly >= 940 & indly <= 960))) %>%  # not in military
  filter(!(year >= 2003 & indly == 9890)) %>%
  filter(schlcoll == 5 | year < 1986) %>%                                    # no school attendance
  filter(exp >= 0)
```