Assignments

This page will contain all the assignments you submit for the class.

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ``` ```{r} ``` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Rachael Villari, Elizabeth Stoner, and Halle Wasser

This assignment is due on Canvas on Monday 9/20 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
library(knitr)
library(datasets)
```

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```
dat <- USArrests
```

Answer: Renaming the dataset provides a short-hand way of calling it as I code instead of trying to remember the name of the dataset.

Problem 2

Use this command to make the state names into a new variable called State.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```
names(dat)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Answer: The variables are Murder, Assault, UrbanPop, Rape, and state.

Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: It is a quantitative/continuous variable.

What R Type of variable is it?

```
typeof("Murder")
```

```
## [1] "character"
```

Answer: It is a character data type.

Problem 4

What information is contained in this dataset, in general? What do the numbers mean?

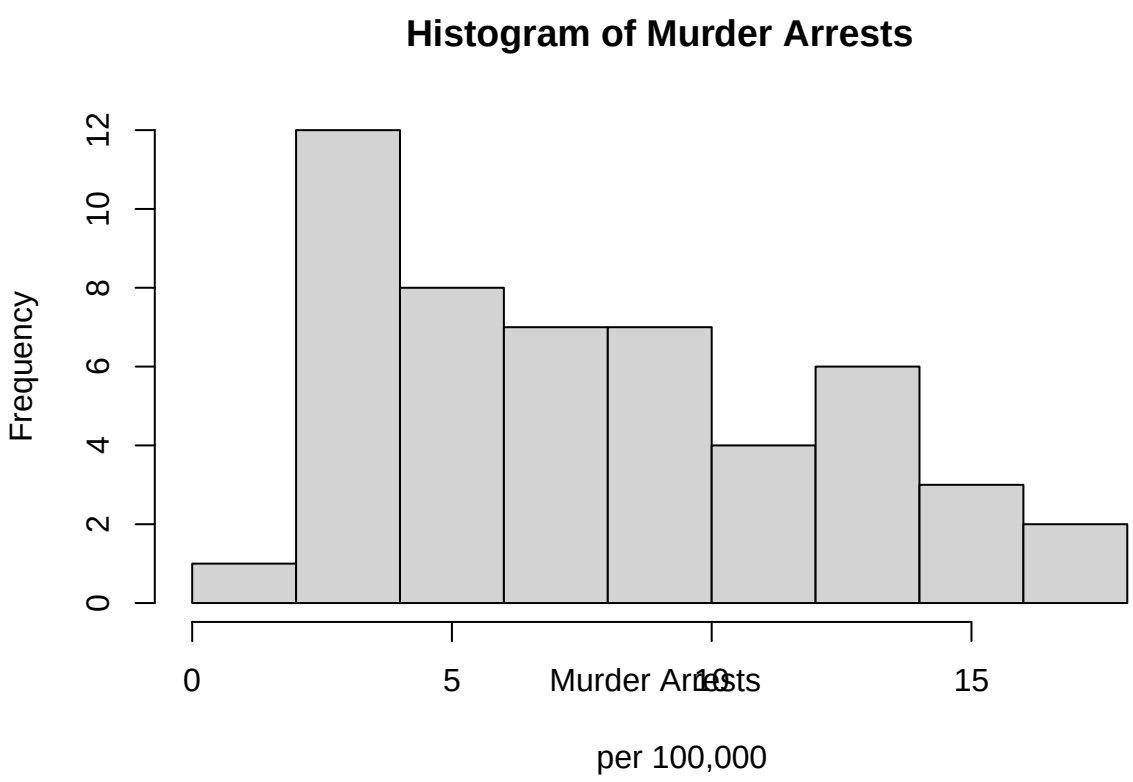
```
?USArrests
```

Answer: This dataset contains 50 observations/instances on 5 variables: Murder (murder arrests per 100,000), Assault (assault arrests per 100,000), UrbanPop (percent urban population), and Rape (rape arrests per 100,000), and state.

Problem 5

Draw a histogram of `Murder` with proper labels and title.

```
hist(dat$Murder, main="Histogram of Murder Arrests", xlab="Murder Arrests \n
     per 100,000", ylab="Frequency")
```



Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
summary(dat$Murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

Answer: The mean number of murder arrests is 7.7888 per 100,000, and the median is 7.25 murder arrests per 100,000. According to HS, the mean and median are both measures of central tendency. However, the mean is described to be the value that each observation would receive if the values were redistributed among the dataset, and the mean on a histogram is the "balancing point". The median is the value that is at the center if all the values were put in an ordered list. According to HS, the interquartile range (IQR) is a measure of spread, and the "quartiles. . . are the three values which divide the distribution into even fourths. The IQR is obtained by subtracting the 1st quartile from the 3rd quartile which is why R gives us both these values.

Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

```
summary(dat$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

```
summary(dat$Rape)
```

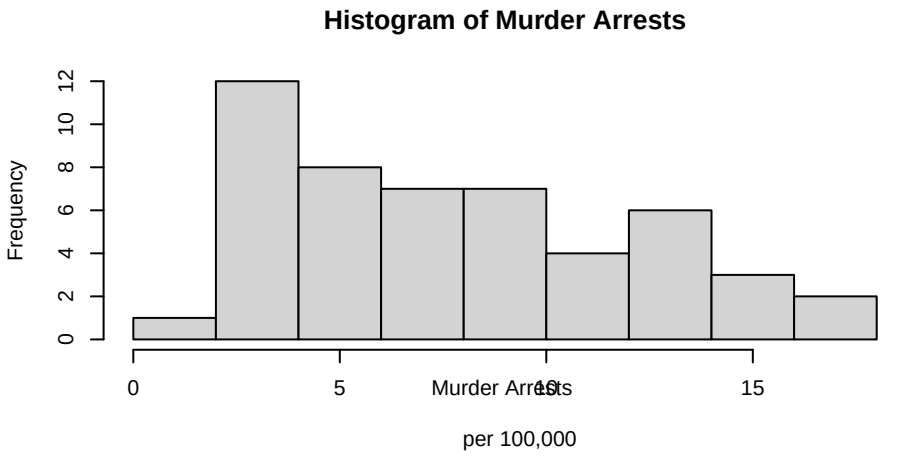
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

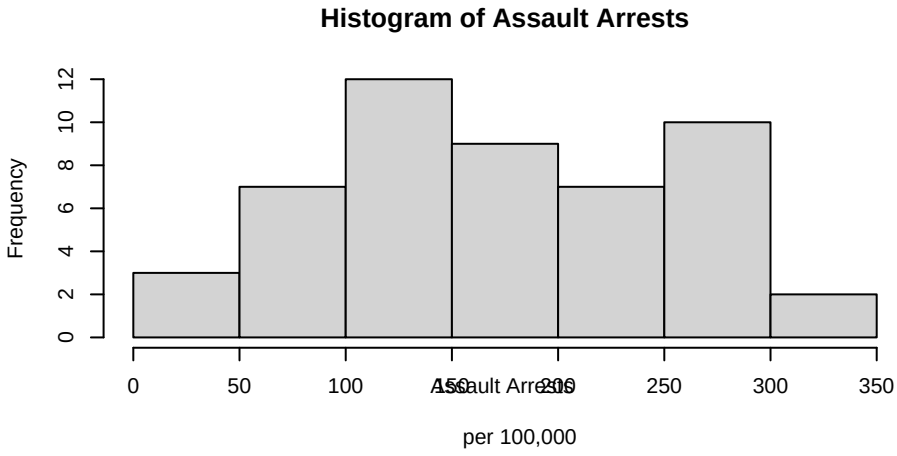
Answer: The mean number of assault arrests is 170.8 per 100,000, and the median is 159 assault arrests per 100,000. The mean number of rape arrests is 21.23 per 100,000, and the median is 20.1 rape arrests per 100,000.

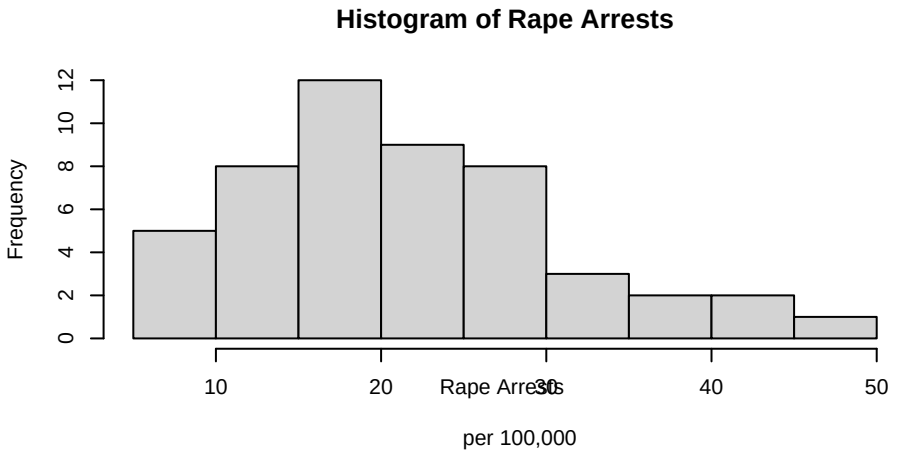
What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: There are a lot of graphical features in R. Par() allows us to search and set them.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Histogram of Murder Arrests", xlab="Murder Arrests \n
     per 100,000", ylab="Frequency")
hist(dat$Assault, main="Histogram of Assault Arrests", xlab="Assault Arrests \n
     per 100,000", ylab="Frequency")
hist(dat$Rape, main="Histogram of Rape Arrests", xlab="Rape Arrests \n
     per 100,000", ylab="Frequency")
```







What can you learn from plotting the histograms together?

Answer: Murder and rape arrests are right skewed, which is good. What I thought was interesting were how the assault arrests are pretty normally distributed. Additionally, I see the average assault arrests per 100,000 were much greater than the average murder or rape arrests per 100,000 which makes sense.

Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
library('maps') # load maps library into R session
library('ggplot2') # load ggplot2 into R session

ggplot(dat, aes(map_id=state, fill=Murder)) + # call ggplot function, use dat
                                              # dataframe, and set up plot aesthetics
  geom_map(map=map_data("state")) + # create map of all states
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat) # expand plot limits
```

What does this code do? Explain what each line is doing.

Answer: This code is creating a heat map of the murder arrests per 100,000 in each state. See comments for a description of each line of code.

Assignment 2

Problem 1

```
#load library
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.5
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(ggplot2)
# Read the data
dat <- read.csv(file = 'dat.nsduh.small.1.csv')
```

What are the dimensions of the dataset?

```
nrow(dat)
```

```
## [1] 171
```

```
ncol(dat)
```

```
## [1] 7
```

```
names(dat)
```

```
## [1] "mjage"     "cigage"    "iralcage"  "age2"      "sexatract" "speakengl"
## [7] "irsex"
```

Answer: This dataset contains 171 rows/observations and 7 columns.

Problem 2

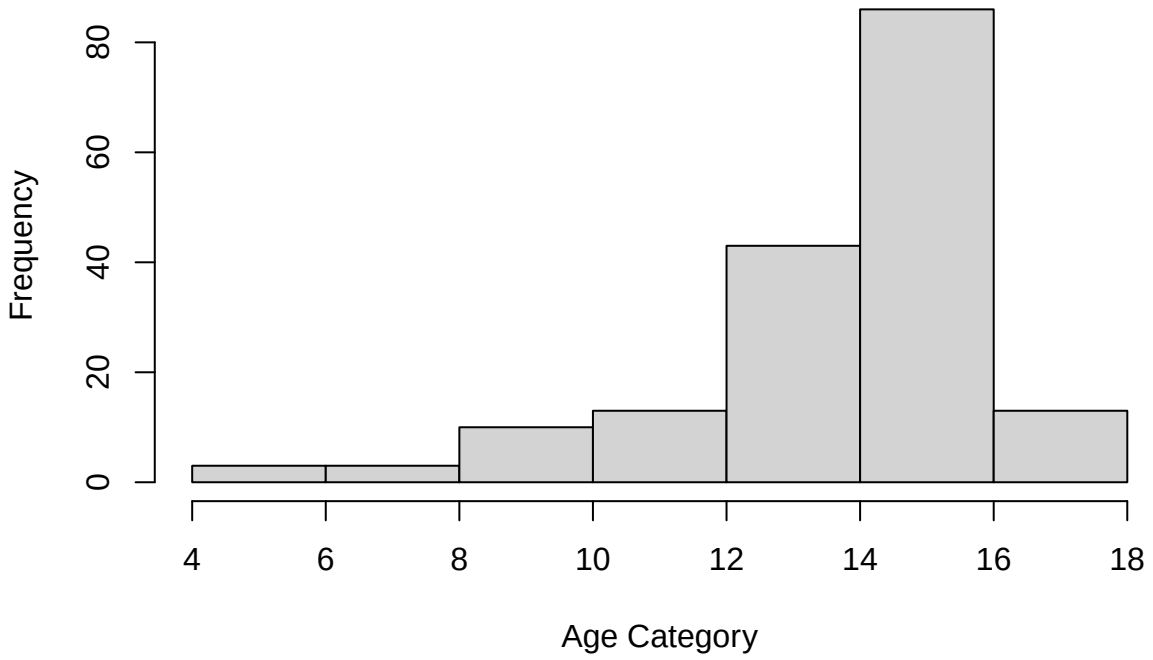
What is this dataset about? Who collected the data, what kind of sample is it, and what was the purpose of generating the data?

Answer: This dataset is a sample of the 2019 National Survey of Drug Use and Health (NSDUH). According to their website (https://nsduhweb.rti.org/respweb/about_nsduh.html), NSDUH is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA) and the US Dept. of Health and Human Services, and it serves as a sample to inform policymakers of tobacco, alcohol, and drug use in the US population as well as other health related concerns. Their website also states that they survey approximately 70,000 people each year ages 12 and up. The variables of the provided dataset are as follows: 'mjage', 'cigage', 'iralcage', 'age2', 'sexatract', 'speakengl', and 'irsex'. 'Mjage', 'cigage', and 'iralcage' refer to the age at which participants first started using marijuana or hashish, cigarettes, and alcohol, respectively. The 'age2' variable is the age of the respondent at the end of the questionnaire. 'Sexatract', 'speakengl', and 'irsex' are the respondent's sexual attraction/orientation, English speaking level, and gender.

Problem 3: Age and gender

What is the age distribution of the sample like?

```
hist(dat$age2, xlab = 'Age Category', main = 'Histogram of Age')
```




```
summary(dat$age2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   13.00   15.00   13.98   15.00   17.00
```

```
dat %>% group_by(age2) %>% count()
```

```
## # A tibble: 13 x 2
## # Groups:   age2 [13]
##     age2     n
##    <int> <int>
##  1     4     2
##  2     6     1
##  3     7     1
##  4     8     2
##  5     9     7
##  6    10     3
##  7    11     6
##  8    12     7
##  9    13    27
## 10    14    16
## 11    15    62
## 12    16    24
## 13    17    13
```

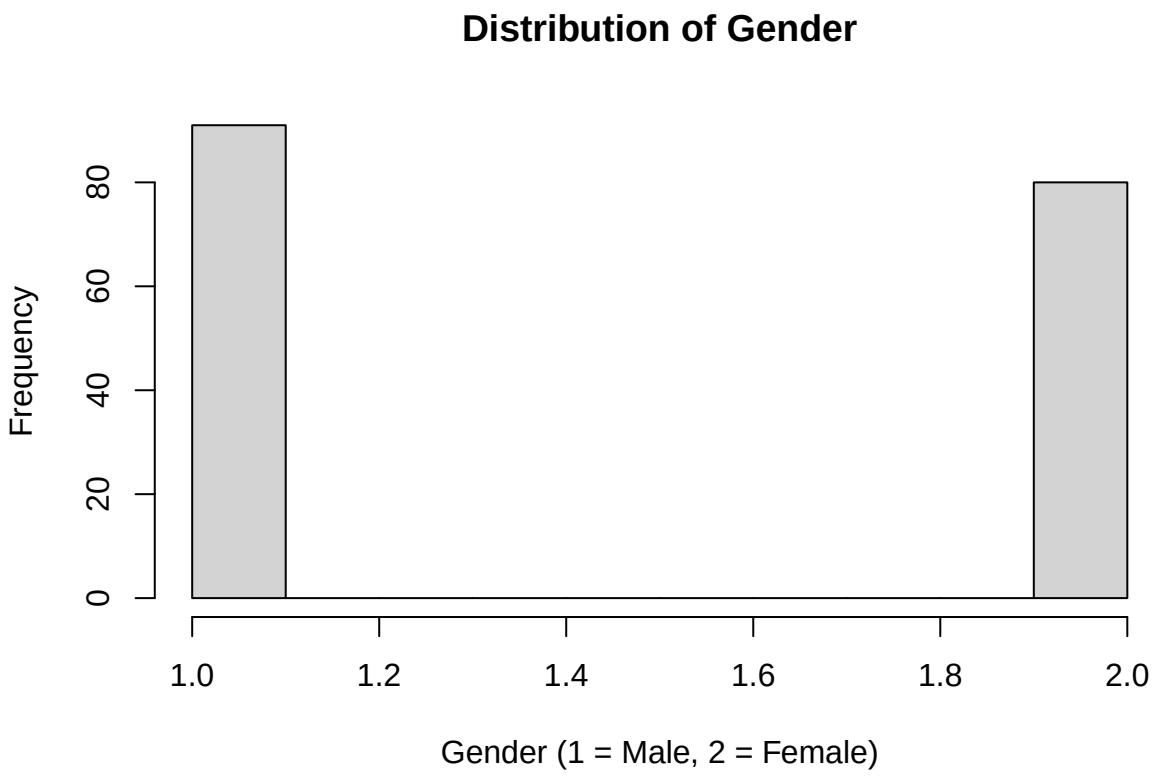
Answer: It appears to be left skewed, and it looks like there's a peak at 15 which the codebook describes as being 35-49 years old. This means that the majority of the people within this dataset are not teenagers

Do you think this age distribution representative of the US population? Why or why not?

Answer: No, because this is a very small sample size (171 people), and it wouldn't be representative of the millions of people that live in the US. Additionally, the NSDUH only surveys people aged 12 and older meaning children under 12 would be excluded from the age distribution. Also there are not many teens or people in their early twenties represented by this dataset.

Is the sample balanced in terms of gender? If not, are there more females or males?

```
hist(dat$irsex, xlab = 'Gender (1 = Male, 2 = Female)', main = 'Distribution of Gender')
```

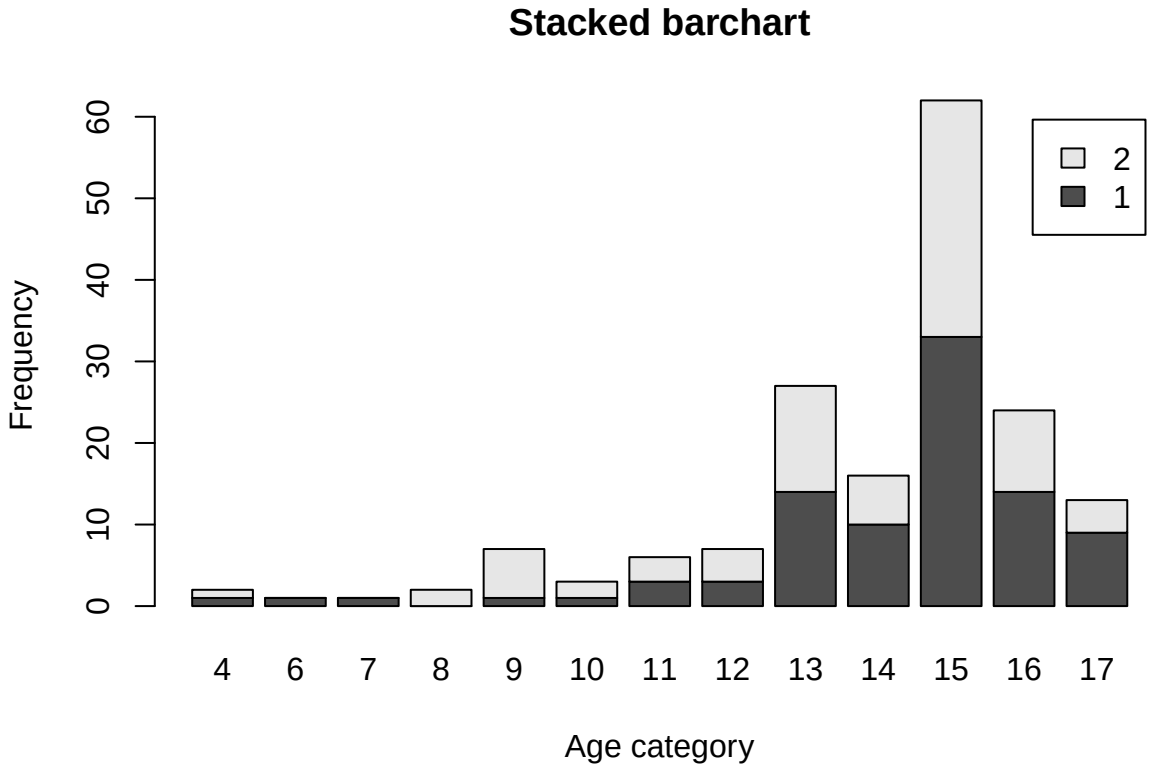



```
dat %>% count(irsex)
```

```
##   irsex  n
## 1     1 91
## 2     2 80
```

Answer: No, the sample is not balanced in terms of gender. There are more males (91 total) than females (80 total).

```
tab.agesex <- table(dat$irsex, dat$age2)
barplot(tab.agesex,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agesex),
        beside = FALSE) # Stacked bars (default)
```



Answer: It's difficult to be certain by just looking at this bar plot, but visually it appears as though this dataset contains males who are older and females who are younger. For example, the age categories of 14, 16, and 17 look to have more men than women while categories 8, 9, 10, 11, and 12 seem to have more females than males. This might also be because there are more males than females in this dataset. However, it is clear that the main age categories in this dataset are 13-17.

Problem 4: Substance use

For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?

```
summary(dat$mjage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   14.00   16.00   15.99   17.50   35.00
```

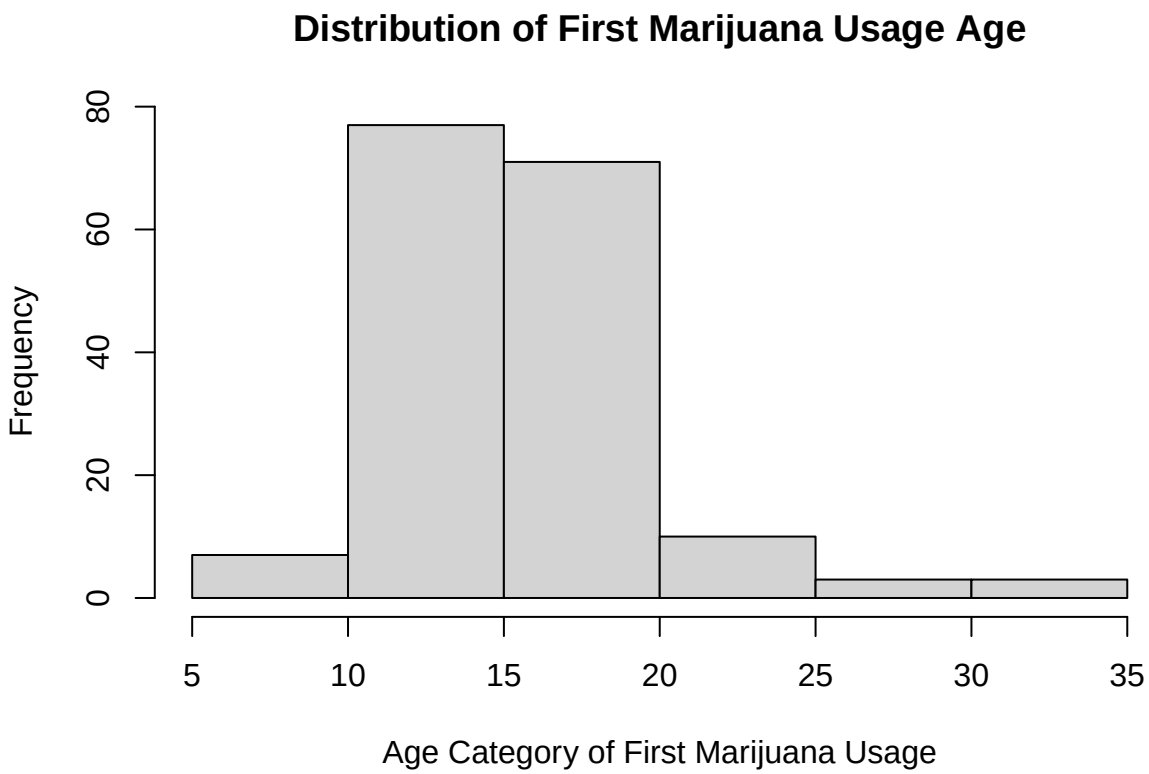
```
summary(dat$cigage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   15.00   17.00   17.65   19.00   50.00
```

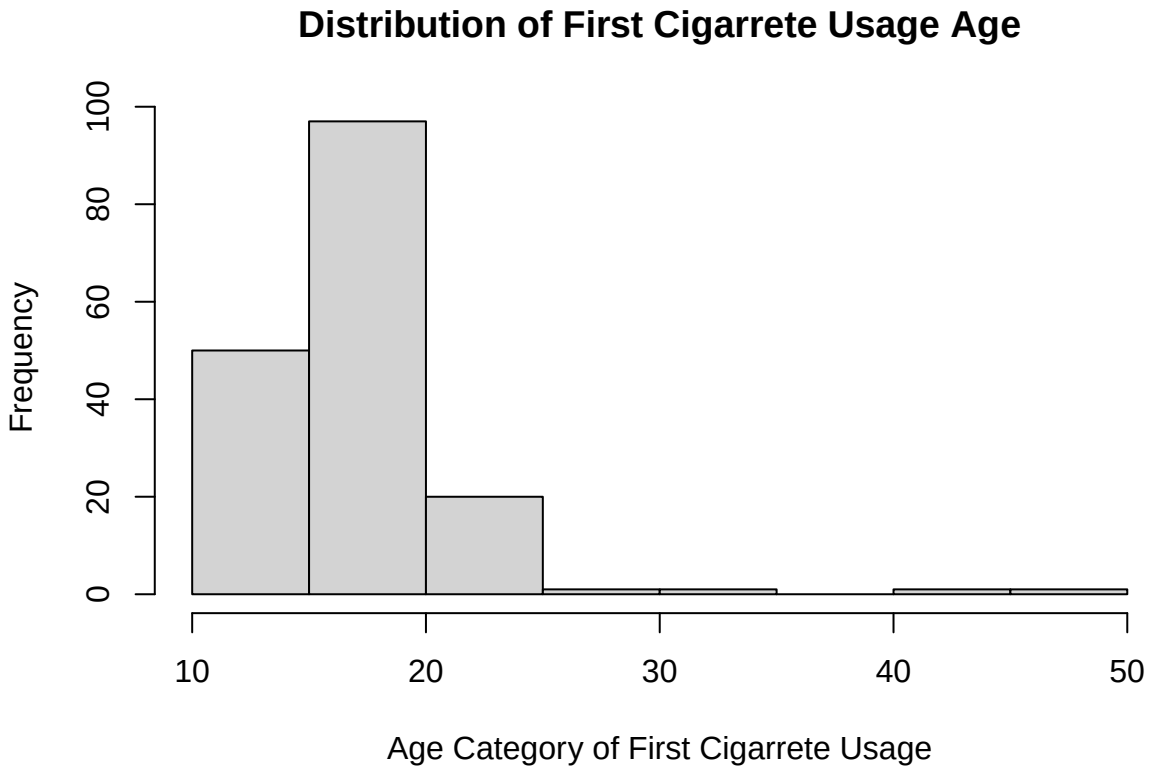
```
summary(dat$iralcage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   13.00   15.00   14.95   17.00   23.00
```

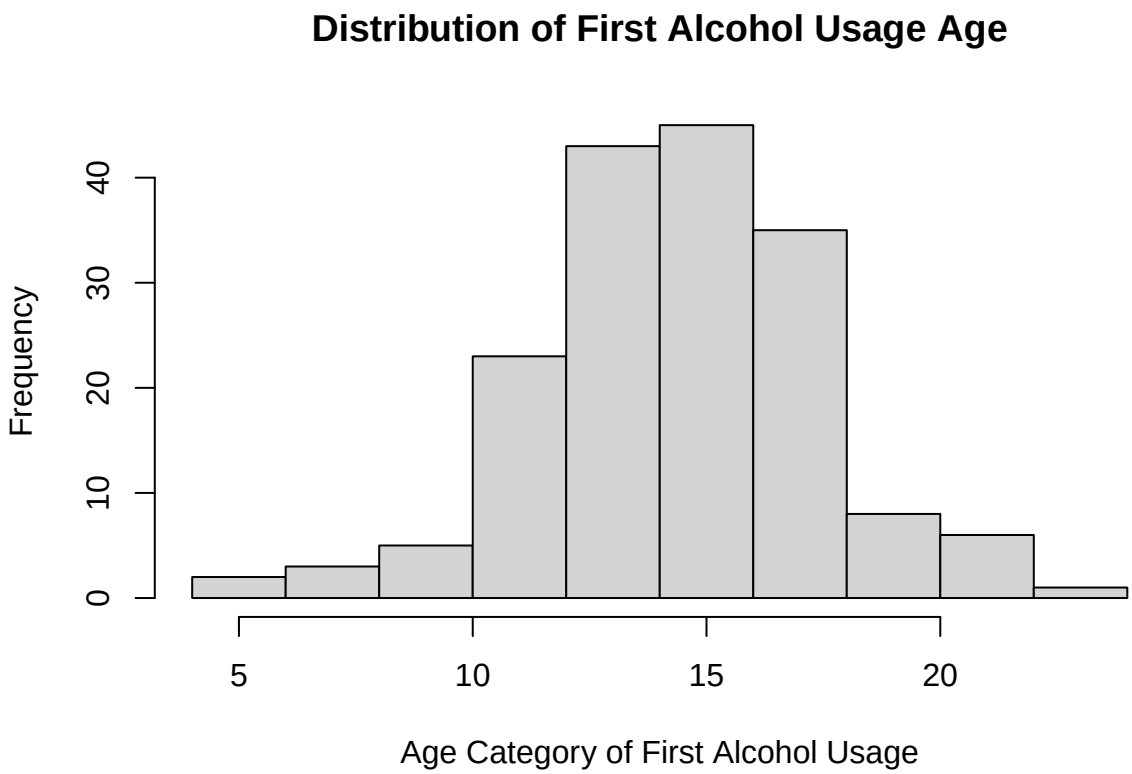
```
hist(dat$mjage, xlab = 'Age Category of First Marijuana Usage', main = 'Distribution of First Marijuana
```




```
hist(dat$cigage, xlab = 'Age Category of First Cigarrete Usage', main = 'Distribution of First Cigarret
```



```
hist(dat$iralcage,  xlab = 'Age Category of First Alcohol Usage', main = 'Distribution of First Alcohol
```



Answer: From the summary statistics and the histograms, it looks like people tend to use alcohol earlier than cigarettes or marijuana, mainly around the age category of 15 (35-49 years old).

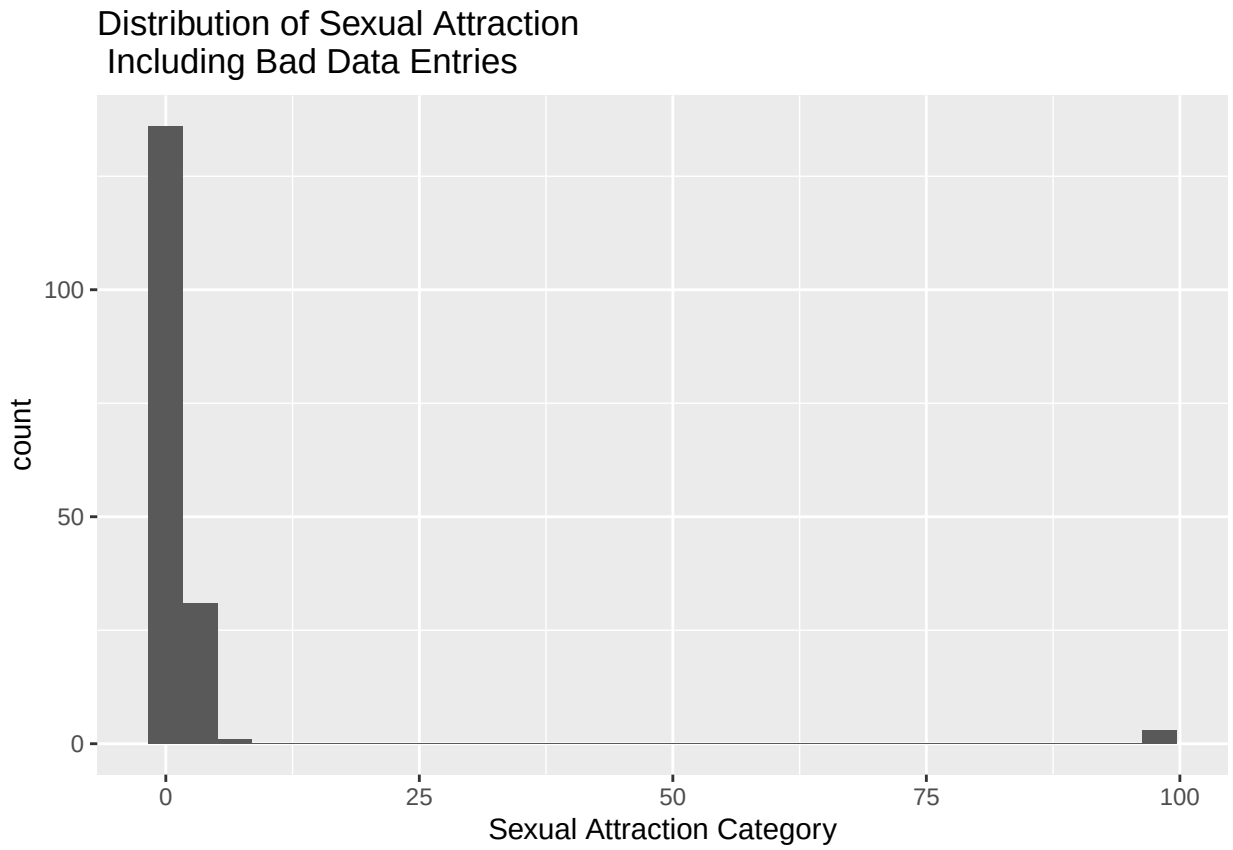
Question 5

What does the distribution of sexual attraction look like? Is this what you expected?

What is the distribution of sexual attraction by gender?

```
ggplot(dat, aes(x=sexatract)) +
  geom_histogram() +
  xlab('Sexual Attraction Category') +
  ggtitle('Distribution of Sexual Attraction \n Including Bad Data Entries')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



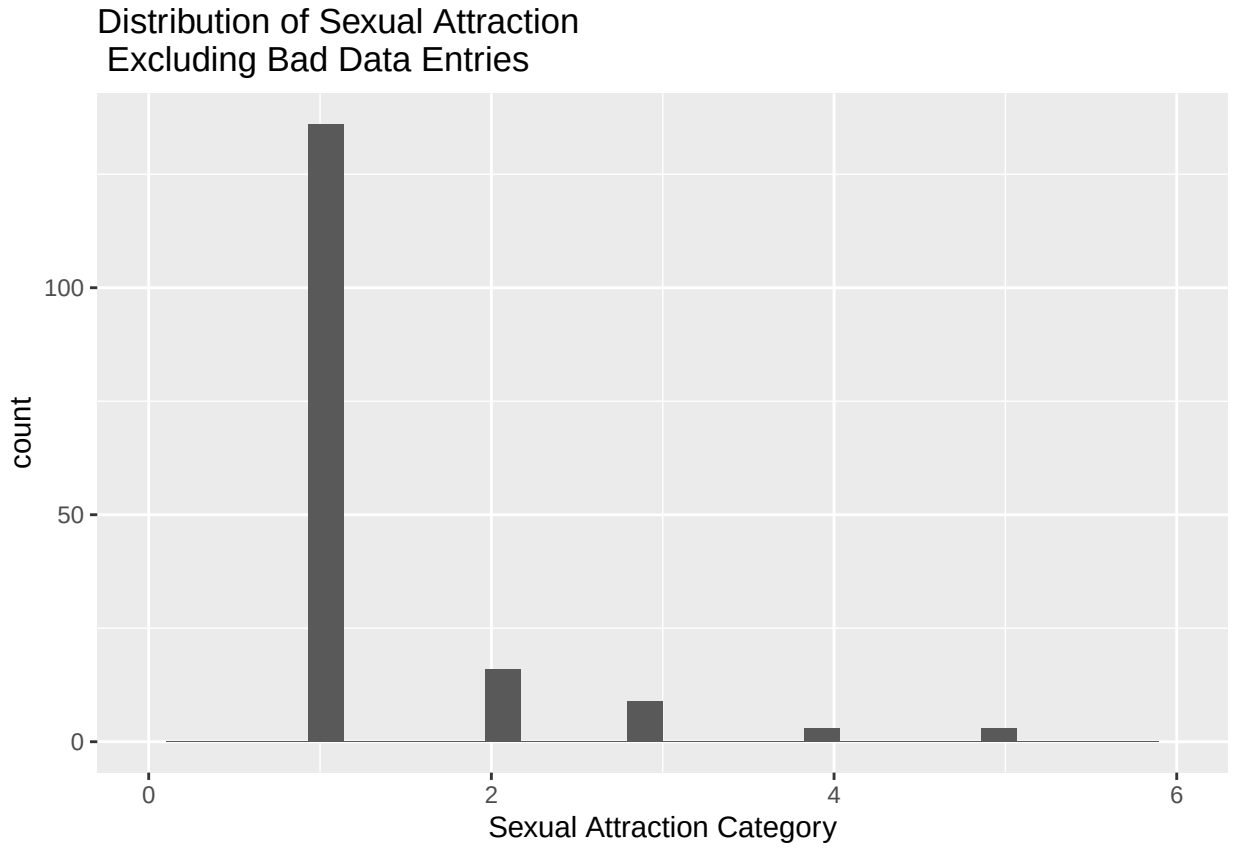
Answer: This histogram includes the entries where people chose not to answer or the entries were blank, answered, or bad data for whatever reason. I expected to see these types of entries in this dataset.

```
ggplot(dat, aes(x=sexatract)) +
  geom_histogram() +
  xlim(0, 6) +
  xlab('Sexual Attraction Category') +
  ggtitle('Distribution of Sexual Attraction \n Excluding Bad Data Entries')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

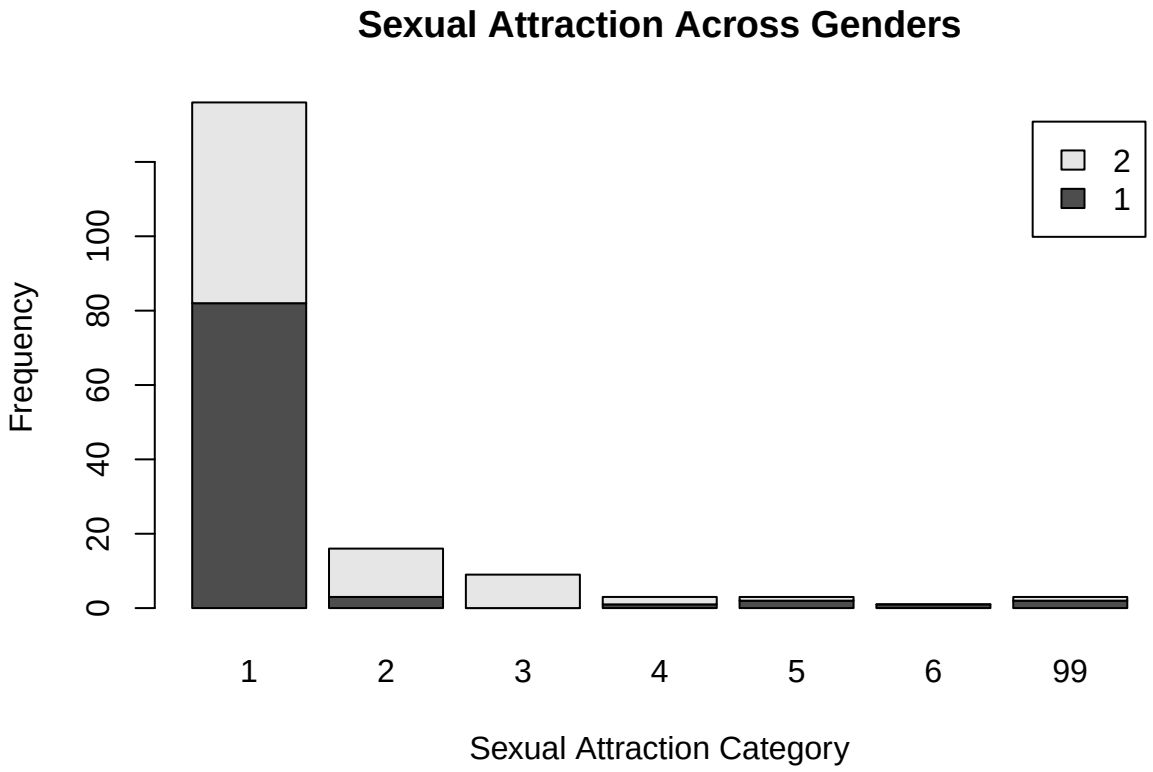
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Answer: I generated this histogram to isolate the sexual attraction categories by excluding the bad data entries. This clearly shows the majority of people in this dataset are only attracted to the opposite sex which is what I expected since heterosexuality is the majority sexual orientation in society. It is followed by mostly opposite sex attraction and attraction to both sexes. I would say none of this is a surprise given my sense of sexual orientations within the US.

```
tab.attsex <- table(dat$irsex, dat$sexatract)
barplot(tab.attsex,
        main = "Sexual Attraction Across Genders",
        xlab = 'Sexual Attraction Category',
        ylab = 'Frequency',
        legend.text = rownames(tab.attsex),
        beside = FALSE)
```



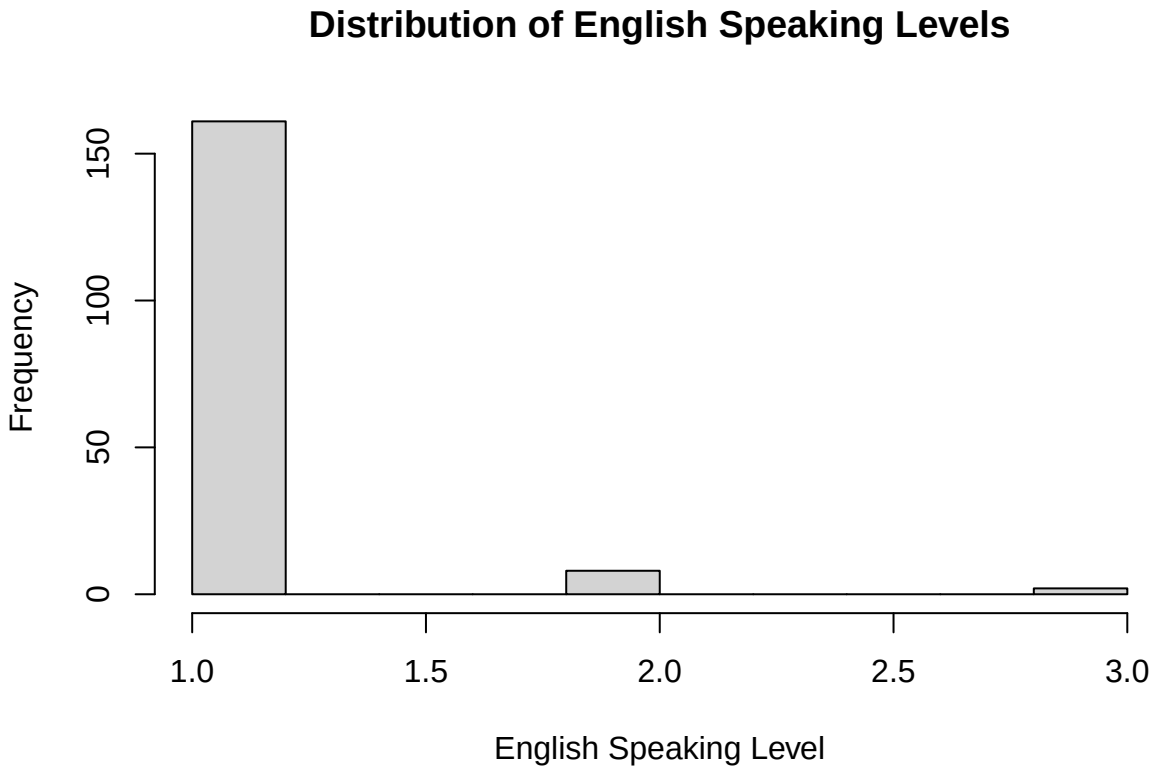
Answer: This plot shows that the males mostly comprised the opposite sex attraction category as well as the strictly same sex attraction and unsure sexual attraction categories. They were also the majority sex that skipped this question altogether. It's interesting that females were the majority in the mostly opposite sex attraction and the both sex attraction categories.

###Problem 6: English speaking

What does the distribution of English speaking look like in the sample? Is this what you might expect for a random sample of the US population?

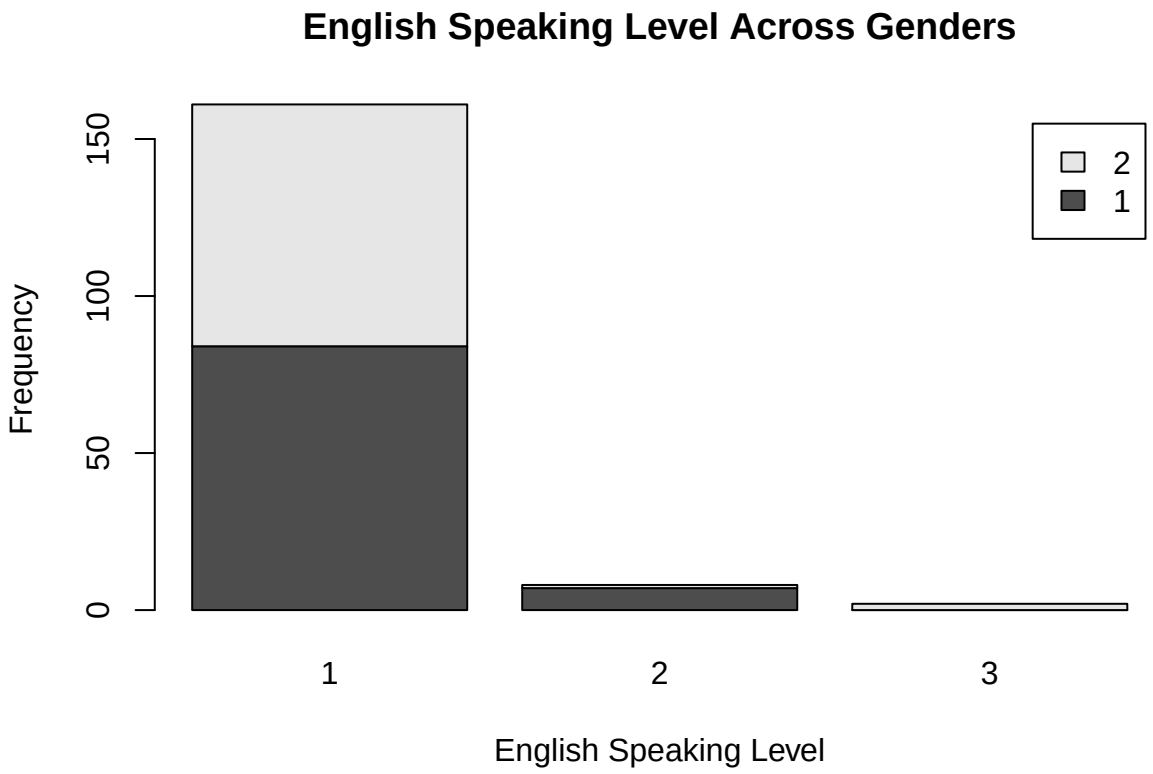
Are there more English speaker females or males?

```
hist(dat$speakengl,
     xlab = 'English Speaking Level',
     main = 'Distribution of English Speaking Levels')
```



Answer: The vast majority of people in this survey speak English very well followed by people who speak it well and those who don't speak it at all. I would expect this from a random sample of the US population since the dominant language is English although we are a melting pot country with a variety of people groups.

```
tab.englsex <- table(dat$irsex, dat$speakengl)
barplot(tab.englsex,
        main = "English Speaking Level Across Genders",
        xlab = 'English Speaking Level',
        ylab = 'Frequency',
        legend.text = rownames(tab.englsex),
        beside = FALSE)
```




```
dat %>% group_by(irsex, speakengl) %>% count()
```

```
## # A tibble: 5 x 3
## # Groups:   irsex, speakengl [5]
##   irsex speakengl     n
##   <int>     <int> <int>
## 1     1         1    84
## 2     1         2     7
## 3     2         1    77
## 4     2         2     1
## 5     2         3     2
```

Answer: From these statistics and plot, it appears as though there are more male English speakers in this dataset, but one has to remember there are more males in the dataset overall. Additionally, it's interesting that there are no males who rated themselves as not speaking English at all.

Exam 1

Load the data into an R data frame.

```
df <- read.csv("fatal-police-shootings-data.csv")
library(ggplot2)
library(tidyverse)
library(tidyr)
```

Problem 1 (10 points)

a. Describe the dataset. This is the source: https://github.com/washingtonpost/data-police-shootings . Write two sentences (max.) about this.

Answer: According to the codebook, this dataset consists of fatal shootings (observations) of civilians by police officers on duty since 2015. The dataset includes information (variables/columns) of the victims such as age, race, gender, threat level, etc.

b. How many observations are there in the data frame?

```
nrow(df)
```

```
## [1] 6594
```

Answer: There are 6594 observations in this dataset.

c. Look at the names of the variables in the data frame. Describe what "body_camera", "flee", and "armed" represent, according to the codebook. Again, only write one sentence (max) per variable.

```
names(df)
```

```
##  [1] "id"                     "name"
##  [3] "date"                   "manner_of_death"
##  [5] "armed"                  "age"
##  [7] "gender"                 "race"
##  [9] "city"                   "state"
## [11] "signs_of_mental_illness" "threat_level"
## [13] "flee"                   "body_camera"
## [15] "longitude"              "latitude"
## [17] "is_geocoding_exact"
```

"Body_camera" indicates 'true' or 'false' if the officer wore a body camera and recorded a portion of the shooting. "Flee" is if the victim was fleeing by foot, car, or not fleeing, and "armed" tells if the victim's arm is determined or unknown or if they were not armed.

d. What are three weapons that you are surprised to find in the "armed" variable? Make a table of the values in "armed" to see the options.

```
table(df$armed)
```

```
##
##                                    air conditioner
##                        207                          1
##                air pistol                Airsoft pistol
##                         1                          3
##                        ax                    barstool
##                        24                          1
##                baseball bat     baseball bat and bottle
##                        20                          1
```

```
## baseball bat and fireplace poker          baseball bat and knife
##                                 1                               1
##                             baton                         BB gun
##                                 6                              15
##               BB gun and vehicle                 bean-bag gun
##                                 1                               1
##                      beer bottle                    binoculars
##                                 3                               1
##                     blunt object                          bottle
##                                 5                               1
##                    bow and arrow                     box cutter
##                                 1                              13
##                            brick               car, knife and mace
##                                 2                               1
##                          carjack                           chain
##                                 1                               3
##                       chain saw                       chainsaw
##                                 2                               1
##                            chair               claimed to be armed
##                                 4                               1
##               contractor's level                  cordless drill
##                                 1                               1
##                         crossbow                        crowbar
##                                 9                               5
##                        fireworks                       flagpole
##                                 1                               1
##                       flashlight                    garden tool
##                                 2                               2
##                      glass shard                        grenade
##                                 4                               1
##                              gun                    gun and car
##                             3798                              12
##                     gun and knife                gun and machete
##                                22                               3
##                     gun and sword                gun and vehicle
##                                 1                              17
##               guns and explosives                         hammer
##                                 3                              18
##                      hand torch                        hatchet
##                                 1                              14
##                  hatchet and gun                       ice pick
##                                 2                               1
##                incendiary device                          knife
##                                 2                             955
##                knife and vehicle               lawn mower blade
##                                 1                               2
##                           machete                machete and gun
##                                51                               1
##                     meat cleaver                metal hand tool
##                                 6                               2
##                     metal object                      metal pipe
##                                 5                              16
##                       metal pole                      metal rake
##                                 4                               1
```

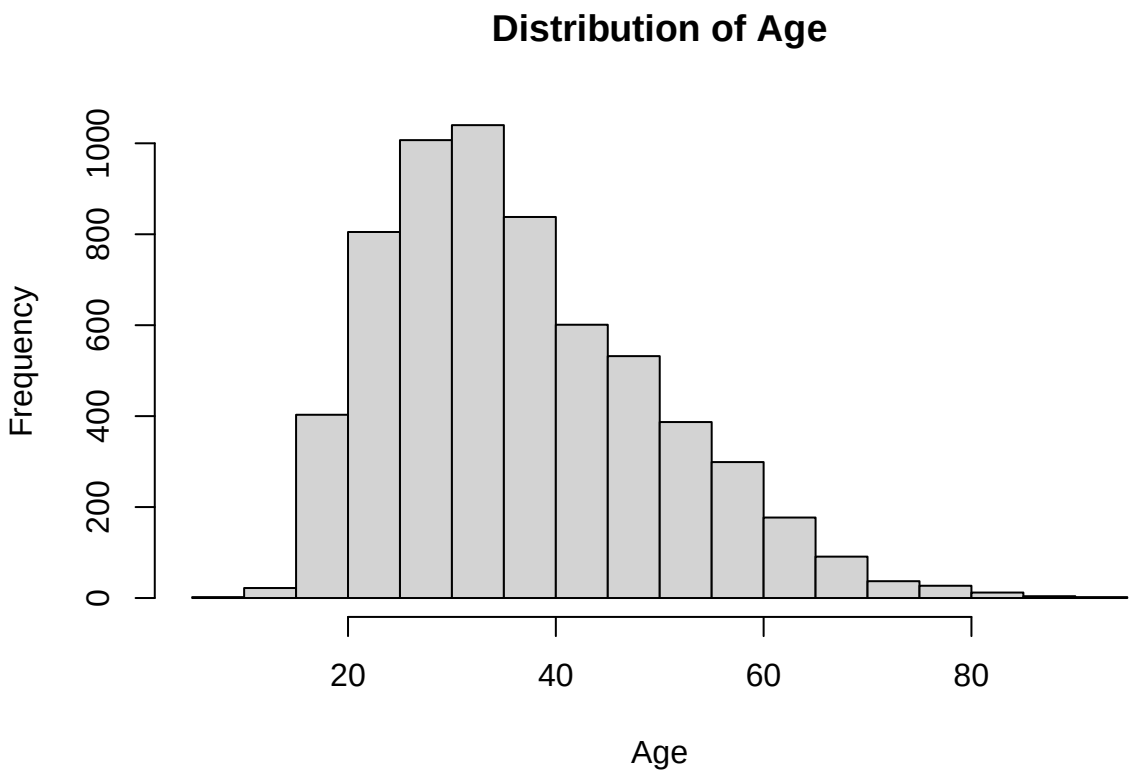
```
##                   metal stick                     microphone
##                             3                              1
##                    motorcycle                      nail gun
##                             1                              1
##                           oar                     pellet gun
##                             1                              3
##                           pen                    pepper spray
##                             1                              2
##                      pick-axe                   piece of wood
##                             4                              7
##                          pipe                       pitchfork
##                             7                              2
##                          pole                   pole and knife
##                             3                              2
##                 railroad spikes                           rock
##                             1                              7
##                   samurai sword                       scissors
##                             4                              9
##                     screwdriver                    sharp object
##                            16                             14
##                           shovel                          spear
##                             7                              2
##                          stapler                straight edge razor
##                             1                              5
##                           sword                          Taser
##                            23                             34
##                       tire iron                    toy weapon
##                             4                            226
##                         unarmed                    undetermined
##                           421                            188
##                    unknown weapon                       vehicle
##                            82                            213
##                   vehicle and gun                vehicle and machete
##                             8                              1
##                    walking stick                     wasp spray
##                             1                              1
##                           wrench
##                             1
```

Answer: It's kind of funny to see an air conditioner, binoculars, and microphone as weapons.

Problem 2 (10 points)

a. Describe the age distribution of the sample. Is this what you would expect to see?

```
hist(df$age, main = 'Distribution of Age', xlab = 'Age')
```



Answer: The age distribution is right skewed and unimodal with the peak being at around early to mid-30s. I would somewhat expect to see this because it's probable that police don't view older adults as threats typically, but I'm not exactly sure of this.

b. To understand the center of the age distribution, would you use a mean or a median, and why? Find the one you picked.

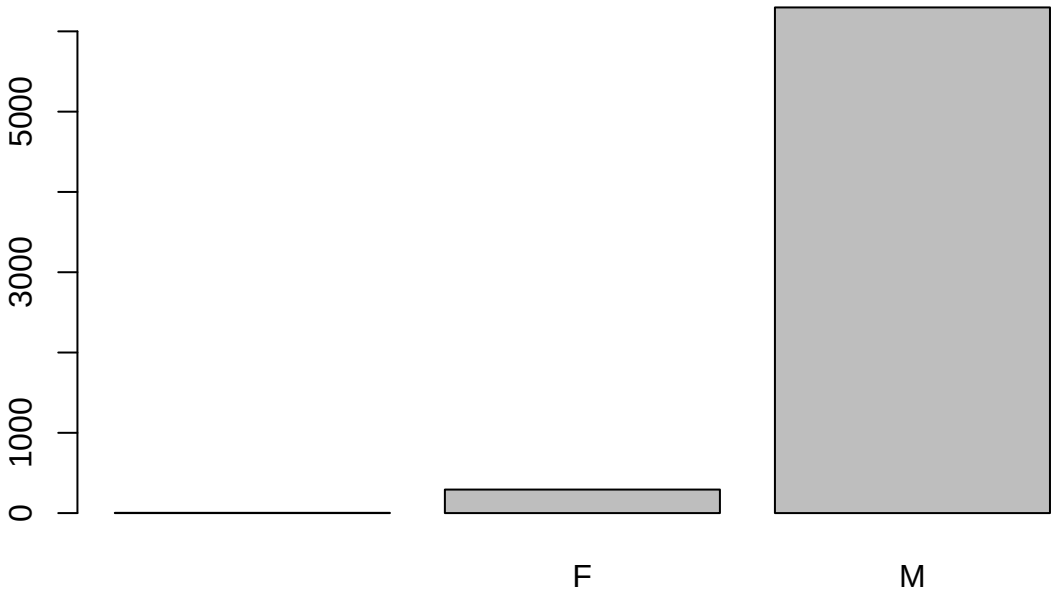
```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   37.12   45.00   91.00     308
```

I would use the median (35 years old) because the age distribution is right skewed, so if the mean were to be used, it wouldn't give an accurate representation of centrality.

c. Describe the gender distribution of the sample. Do you find this surprising?

```
barplot(table(df$gender), main = "Distribution of Gender")
```




```
df %>% group_by(gender) %>% count()
```

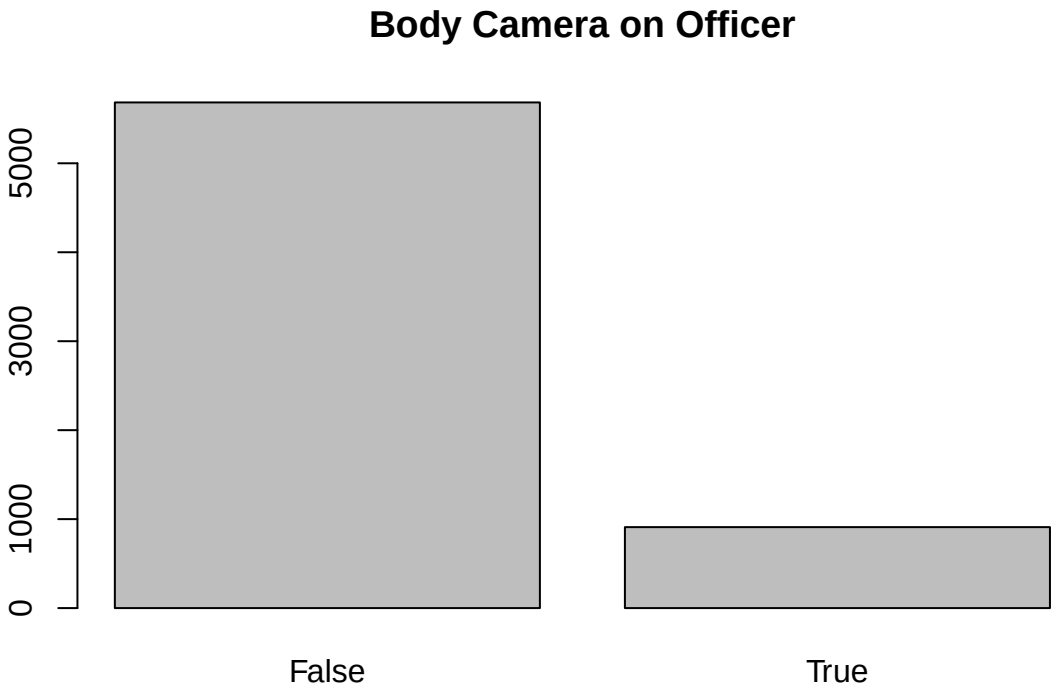
```
## # A tibble: 3 x 2
## # Groups:   gender [3]
##   gender     n
##   <chr>  <int>
## 1 ""         3
## 2 "F"      293
## 3 "M"     6298
```

Answer: The gender distribution is very unbalanced with males making up the majority (6298/6594 = 95.5%) of the dataset. This surprises me because although I expected more males in the dataset, I didn't think they would make up about 96% of the dataset. It looks like the gender column also contains some null values.

Problem 3 (10 points)

a. How many police officers had a body camera, according to news reports? What proportion is this of all the incidents in the data? Are you surprised that it is so high or low?

```
barplot(table(df$body_camera), main = 'Body Camera on Officer')
```



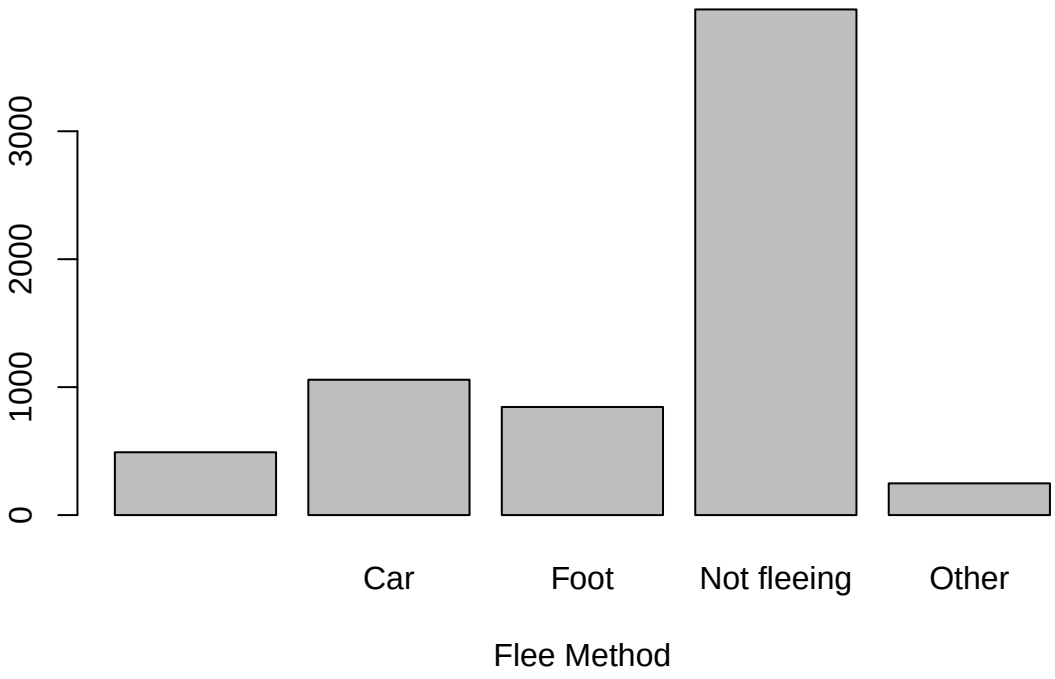

```
df %>% group_by(body_camera) %>% count() %>% mutate(proportion = n/nrow(df))
```

```
## # A tibble: 2 x 3
## # Groups:   body_camera [2]
##   body_camera     n proportion
##   <chr>       <int>      <dbl>
## 1 False        5684      0.862
## 2 True          910      0.138
```

Answer: Very few, only 13.8%, officers had a body camera. This doesn't surprise me because when police are on duty, I would imagine they don't wear cameras on their person.

b. In how many of the incidents was the victim fleeing? What proportion is this of the total number of incidents in the data? Is this what you would expect?

```
barplot(table(df$flee), main = 'Distribution of Flee Methods', xlab = 'Flee Method')
```




```
df %>% group_by(flee) %>% count() %>% mutate(proportion = n/nrow(df))
```

```
## # A tibble: 5 x 3
## # Groups:   flee [5]
##   flee              n proportion
##   <chr>         <int>      <dbl>
## 1 ""              491     0.0745
## 2 "Car"          1058     0.160
## 3 "Foot"          845     0.128
## 4 "Not fleeing"  3952     0.599
## 5 "Other"         248     0.0376
```

Answer: Victims fleeing by car or foot make up 28.9% of the dataset. This number excludes null or 'other' values. This is expected because I imagine most people don't flee from police.

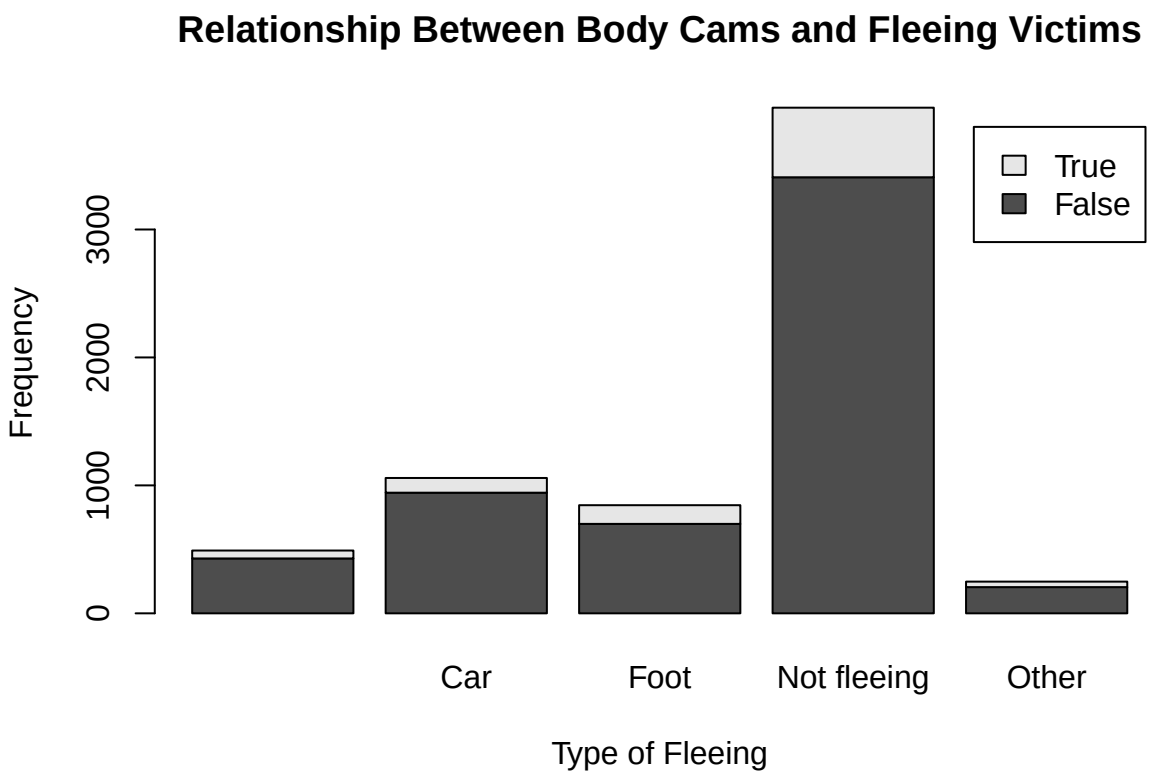
Problem 4 (10 points)

a. Describe the relationship between the variables "body camera" and "flee" using a stacked barplot. What can you conclude from this relationship?

Hint 1: The categories along the x-axis are the options for "flee", each bar contains information about whether the police officer had a body camera (vertically), and the height along the y-axis shows the frequency of that category).

Hint 2: Also, if you are unsure about the syntax for barplot, run ?barplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.

```
tab.camflee <- table(df$body_camera, df$flee)
barplot(tab.camflee,
        main = "Relationship Between Body Cams and Fleeing Victims",
        xlab = 'Type of Fleeing',
        ylab = 'Frequency',
        legend.text = rownames(tab.camflee),
        beside = FALSE)
```




```
df %>% group_by(body_camera, flee) %>% count()
```

```
## # A tibble: 10 x 3
## # Groups:   body_camera, flee [10]
##    body_camera flee              n
##    <chr>       <chr>         <int>
##  1 False       ""              429
##  2 False       "Car"           943
##  3 False       "Foot"          699
##  4 False       "Not fleeing"  3408
##  5 False       "Other"         205
##  6 True        ""               62
##  7 True        "Car"           115
##  8 True        "Foot"          146
##  9 True        "Not fleeing"   544
## 10 True        "Other"          43
```

Answer: It's important to note that these variables contain some null and 'other' values, and the 'true' and 'false' in the legend indicates whether or not the officer had a body camera. It looks like a body camera was used mostly on victims who didn't flee followed by the victims that fled on foot. This makes sense since the majority of the victims in this dataset didn't flee.

Extra credit (10 points)

a. What does this code tell us?

```
mydates <- as.Date(df$date)
head(mydates)
(mydates[length(mydates)] - mydates[1])
```

Answer: This tells us the time difference in days between the oldest and the most recent fatal shooting in this dataset.

b. On Friday, a new report was published that was described as follows by The Guardian: "More than half of US police killings are mislabelled or not reported, study finds." Without reading this article now (due to limited time), why do you think police killings might be mislabelled or underreported? **Answer: It might be because the ones that are recorded have missing values which causes them to be excluded from the analysis.**

c. Regarding missing values in problem 4, do you see any? If so, do you think that's all that's missing from the data? **Answer: Yes, there were some, and I think there are more missing values in the dataset.**

Assignment 3

Collaborators: Elizabeth Stoner and Halle Wasser.

This assignment is due on Canvas on Wednesday 10/27/2021 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

Submit your responses as either an HTML file or a PDF file on Canvas. Also, please upload it to your website.

Save the file (found on Canvas) crime_simple.txt to the same folder as this file (your Rmd file for Assignment 3).

Load the data.

```
library(readr)
library(knitr)
library(ggplot2)
library(tidyverse)
dat.crime <- read_delim("crime_simple.txt", delim = "\t")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   R = col_double(),
##   Age = col_double(),
##   S = col_double(),
##   Ed = col_double(),
##   Ex0 = col_double(),
##   Ex1 = col_double(),
##   LF = col_double(),
##   M = col_double(),
```

```
##   N = col_double(),
##   NW = col_double(),
##   U1 = col_double(),
##   U2 = col_double(),
##   W = col_double(),
##   X = col_double()
## )
```

This is a dataset from a textbook by Brian S. Everitt about crime in the US in 1960. The data originate from the Uniform Crime Report of the FBI and other government sources. The data for 47 states of the USA are given.

Here is the codebook:

R: Crime rate: # of offenses reported to police per million population

Age: The number of males of age 14-24 per 1000 population

S: Indicator variable for Southern states (0 = No, 1 = Yes)

Ed: Mean of years of schooling x 10 for persons of age 25 or older

Ex0: 1960 per capita expenditure on police by state and local government

Ex1: 1959 per capita expenditure on police by state and local government

LF: Labor force participation rate per 1000 civilian urban males age 14-24

M: The number of males per 1000 females

N: State population size in hundred thousands

NW: The number of non-whites per 1000 population

U1: Unemployment rate of urban males per 1000 of age 14-24

U2: Unemployment rate of urban males per 1000 of age 35-39

W: Median value of transferable goods and assets or family income in tens of $

X: The number of families per 1000 earning below 1/2 the median income

We are interested in checking whether the reported crime rate (# of offenses reported to police per million population) and the average education (mean number of years of schooling for persons of age 25 or older) are related.

1. How many observations are there in the dataset? To what does each observation correspond?

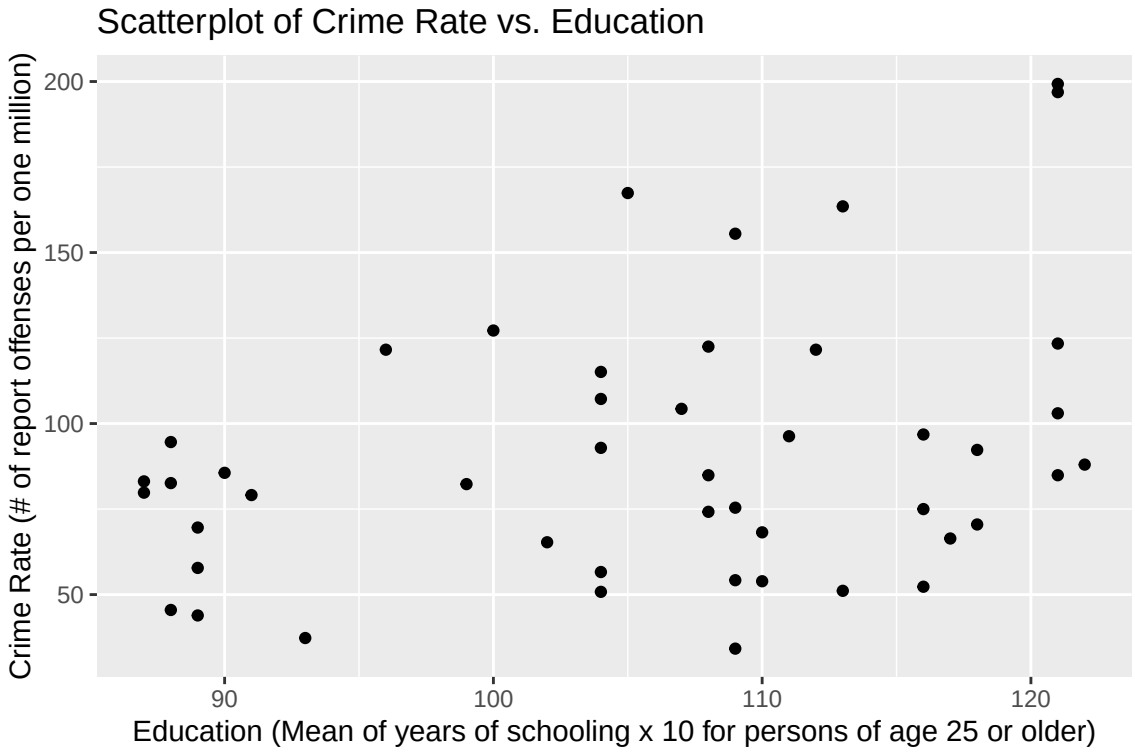
```
nrow(dat.crime)
```

```
## [1] 47
```

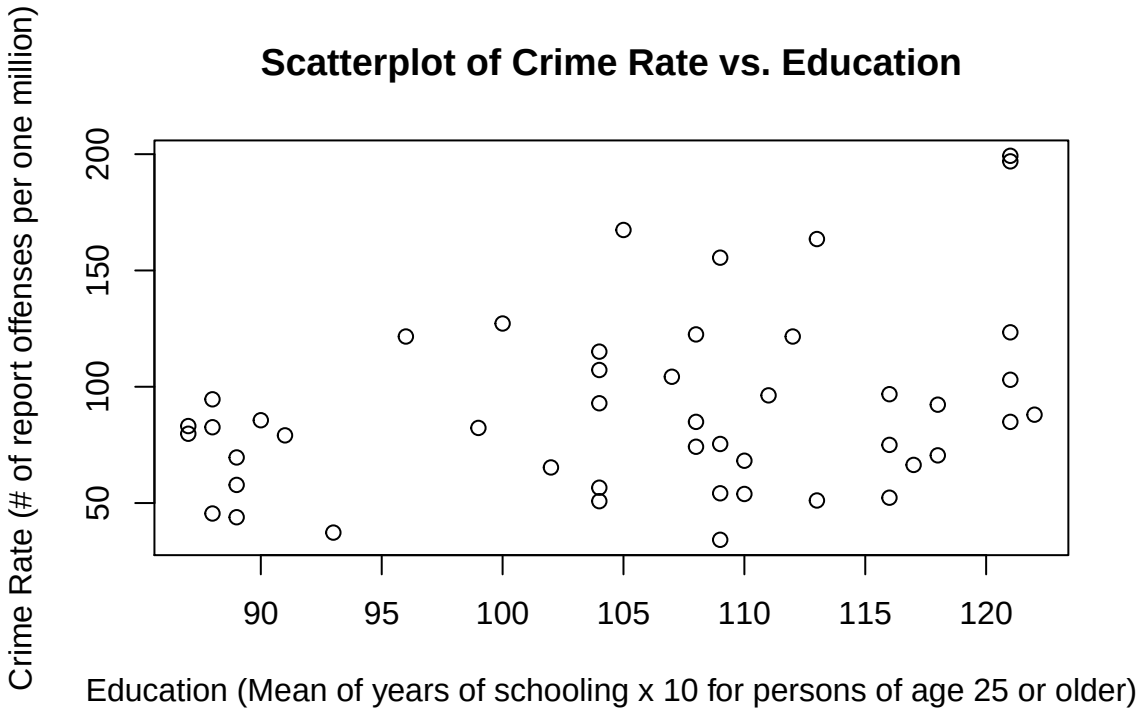
Answer: There are 47 rows in the dataset. Each row corresponds to a US state, and the columns are various attributes of each state such as population, education, crime rate, etc.

2. Draw a scatterplot of the two variables. Calculate the correlation between the two variables. Can you come up with an explanation for this relationship?

```
dat.crime %>% ggplot(aes(x=Ed, y=R)) +
  geom_point() +
  ggtitle("Scatterplot of Crime Rate vs. Education") +
  xlab("Education (Mean of years of schooling x 10 for persons of age 25 or older)") +
  ylab("Crime Rate (# of report offenses per one million)")
```




```
plot(dat.crime$Ed, dat.crime$R, main = "Scatterplot of Crime Rate vs. Education",
     xlab = "Education (Mean of years of schooling x 10 for persons of age 25 or older)",
     ylab = "Crime Rate (# of report offenses per one million)")
```




```r
cor(x=dat.crime$Ed, y=dat.crime$R)
```

```
## [1] 0.3228349
```

Answer: Based on the correlation, it looks like there is a weak positive relationship between crime rate and average education. This is somewhat expected since people with more education may be more inclined to report incidences of crime, but I'm unsure of this.

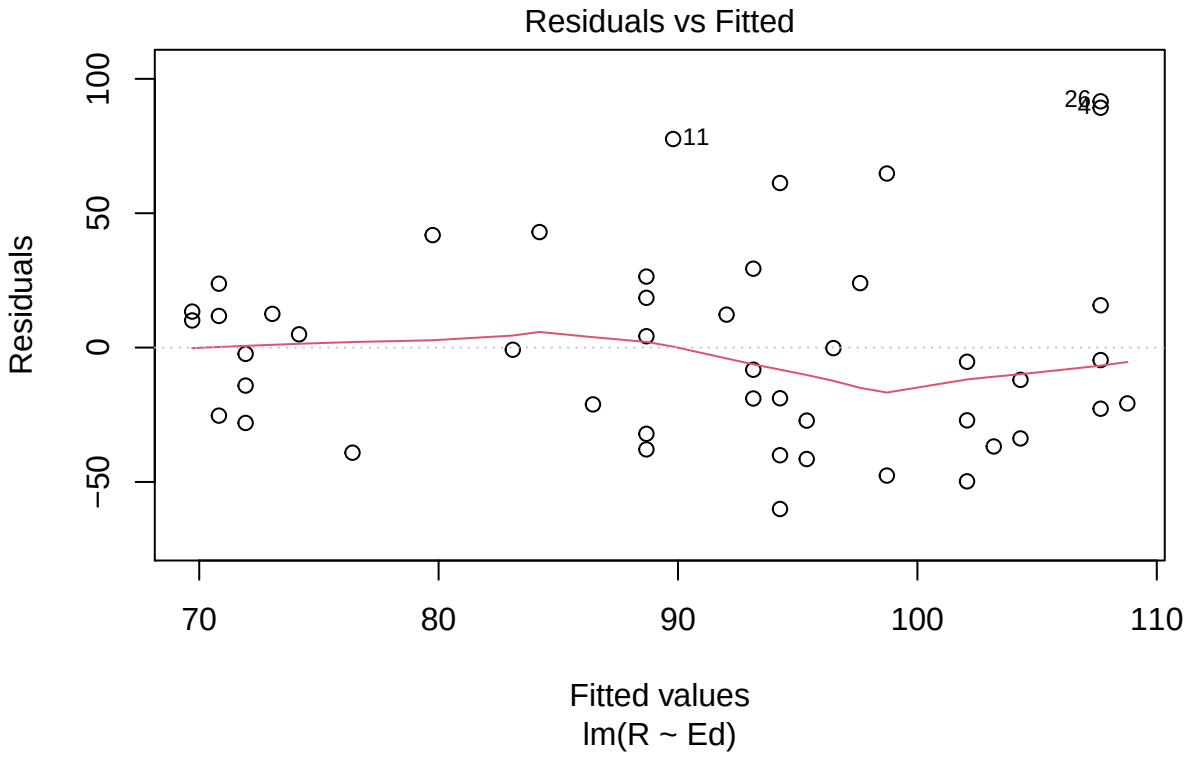
3. Regress reported crime rate (y) on average education (x) and call this linear model `crime.lm` and write the summary of the regression by using this code, which makes it look a little nicer `{r, eval=FALSE} kable(summary(crime.lm)$coef, digits = 2)`.

```r
# Remember to remove eval=FALSE above!
crime.lm <- lm(R ~ Ed, data = dat.crime)
kable(summary(crime.lm)$coef, digits = 2)
```

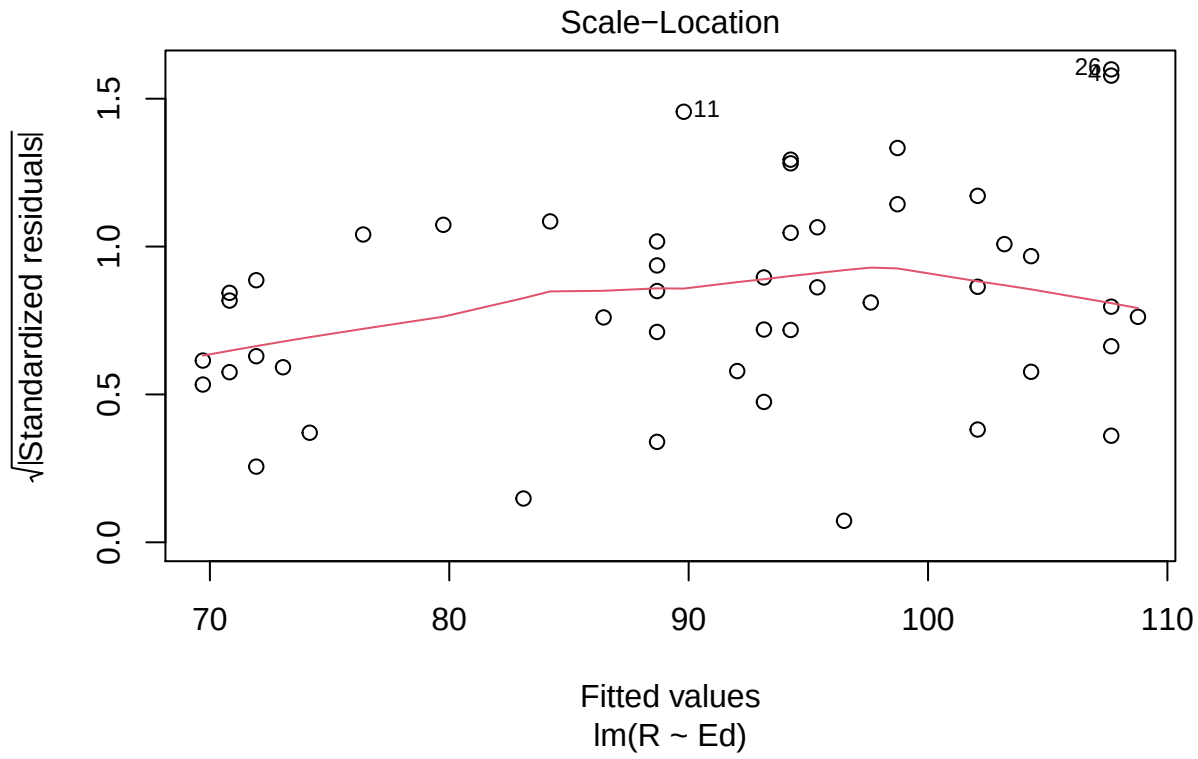
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.40	51.81	-0.53	0.60
Ed	1.12	0.49	2.29	0.03

4. Are the four assumptions of linear regression satisfied? To answer this, draw the relevant plots. (Write a maximum of one sentence per assumption.)

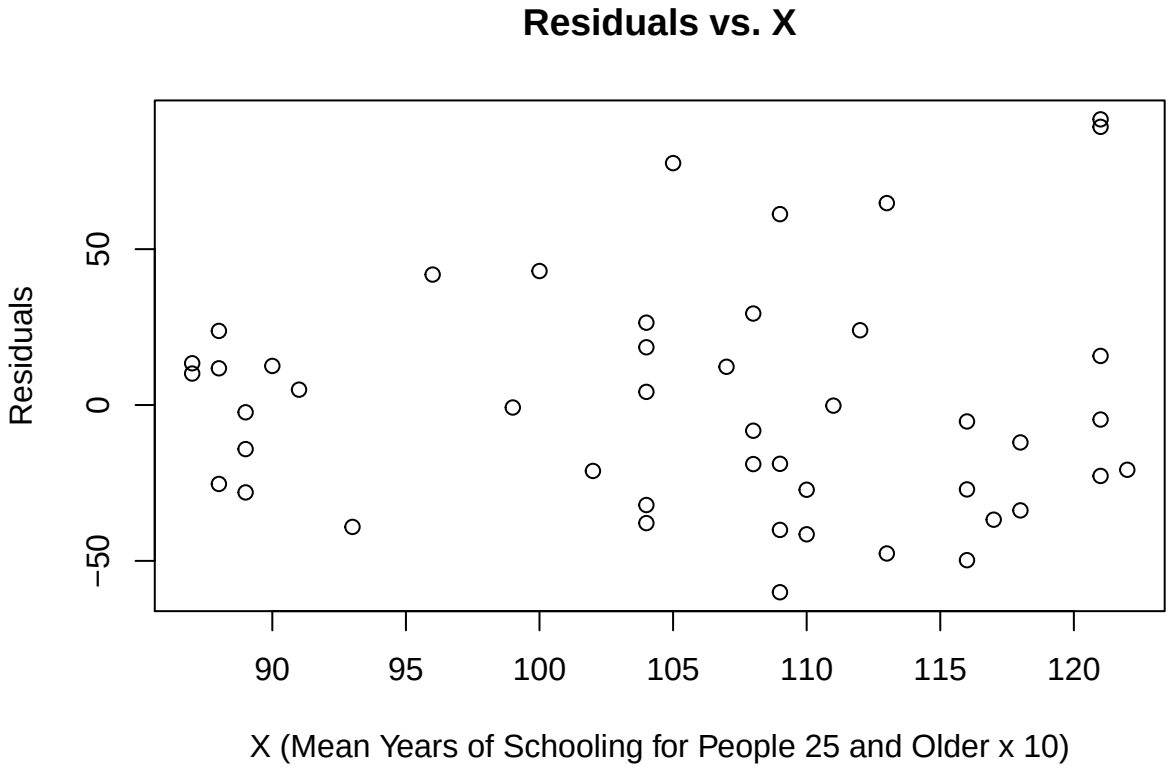
```r
#residuals vs fitted plot
plot(crime.lm, which = 1) #equal variance assumption met (roughly same variance throughout given the am
```




```
plot(crime.lm, which = 3)
```

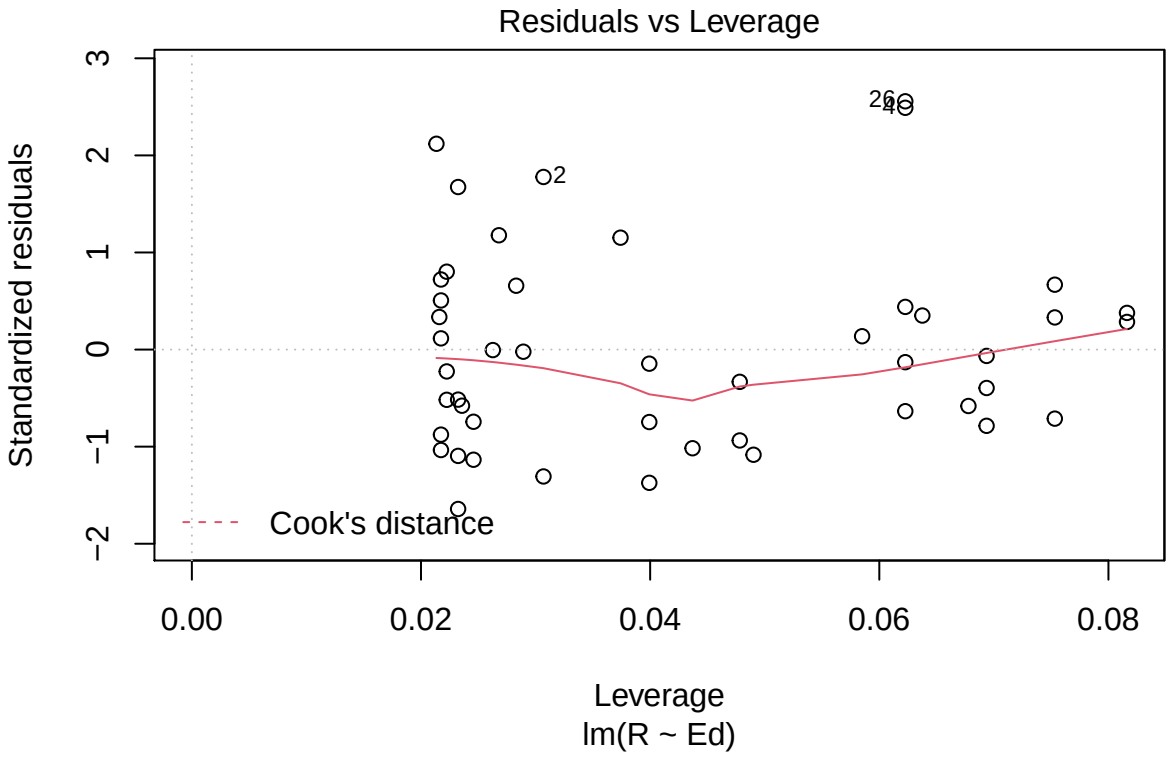



```
#residuals vs x plot
plot(dat.crime$Ed, crime.lm$residuals, main = "Residuals vs. X", xlab = "X (Mean Years of Schooling for
```

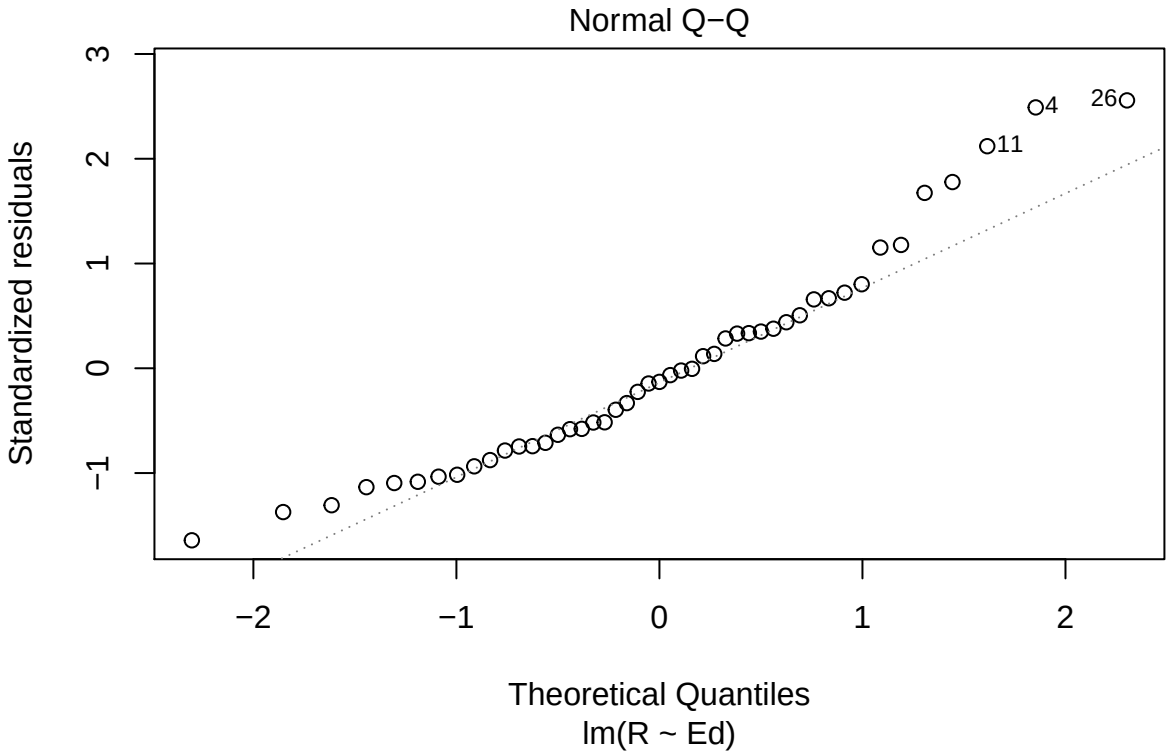


Answer: The equal variance assumption (roughly same variance throughout given the amount of data, although subjective), the linearity assumption (roughly horizontal direction with no distinguishable pattern), and the independence assumptions are met. Even in the more sensitive approach using the scale location plot, these assumptions are somewhat met. There are some residuals that are closer together at the beginning and fan out throughout the plot, and the line is not completely horizontal. Also the errors appear independent since there is no pattern in the Residuals vs. X plot which meets the independence assumption. As mentioned before, this is subjective, but given the size of the dataset, I believe these assumptions are satisfied.

```
plot(crime.lm, which = 5)
```




```
plot(crime.lm, which = 2)
```



Answer: There appears to be no large residuals that have leverage in the model, but the QQ plot indicates a violation of the normality assumption since as the quantiles increase, the standardized residuals tend to curve upwards. This will affect the t tests in the model because

the p-values will be too optimistic.

5. Is the relationship between reported crime and average education statistically significant? Report the estimated coefficient of the slope, the standard error, and the p-value. What does it mean for the relationship to be statistically significant?

Yes, it is statistically significant. The estimated coefficient is 1.12, standard error is 0.49, and its p-value is 0.03. The p-value is the probability of observing any value greater than or equal to the t value (2.29 in this case) which is how many standard deviations away the estimate is from 0. An estimate is considered to be statistically significant if the p-value is less than a certain cutoff (alpha value), usually 0.05, in which case the null hypothesis is rejected.

6. How are reported crime and average education related? In other words, for every unit increase in average education, how does reported crime rate change (per million) per state?

For every unit increase in Ed (mean of years of schooling x 10 for persons of age 25 or older), there is, on average, an increase in the number of reported crimes (per million per state) by 1.12.

7. Can you conclude that if individuals were to receive more education, then crime will be reported more often? Why or why not?

This conclusion cannot be obtained from this analysis because this regression does not model a causal relationship. There are many variables for which to control that could affect the amount of crime recorded, so we cannot say that more education causes more crimes to be reported. This analysis merely shows association between the variable, not causation. However, it must be taken into consideration that not all the assumptions were met, so one needs to be cautious about statistical inference using this model.

Exam 2

Data description: This dataset provides (simulated) data about 200 police departments in one year. It contains information about the funding received by the department as well as incidents of police brutality. Suppose this dataset (sim.data.csv) was collected by researchers to answer this question: **"Does having more funding in a police department lead to fewer incidents of police brutality?"** d. Codebook: - funds: How much funding the police department received in that year in millions of dollars. - po.brut: How many incidents of police brutality were reported by the department that year. - po.dept.code: Police department code

Problem 1: EDA (10 points)

Describe the dataset and variables. Perform exploratory data analysis for the two variables of interest: funds and po.brut.

```
#load libraries
library(knitr)
library(ggplot2)
library(tidyverse)
```

```
dat <- read.csv(file = 'sim.data.csv')
dim(dat)
```

```
## [1] 200   3
```

```
sum(is.na(dat$po.dept.code))
```

```
## [1] 0
```

```
sum(is.na(dat$funds))
```

```
## [1] 0
```

```
sum(is.na(dat$po.brut))
```

```
## [1] 0
```

Answer: This dataset set contains 200 observations (rows), and each of them represents a police department. The columns (3 total) are attributes of each police department, and they include funds (funding received in millions of dollars for that year), po.brut (number of reported incidents of police brutality) for that year, and po.dept.code (id number for the department, most likely to protect privacy and confidentiality). There are no missing (n/a) values in this dataset.

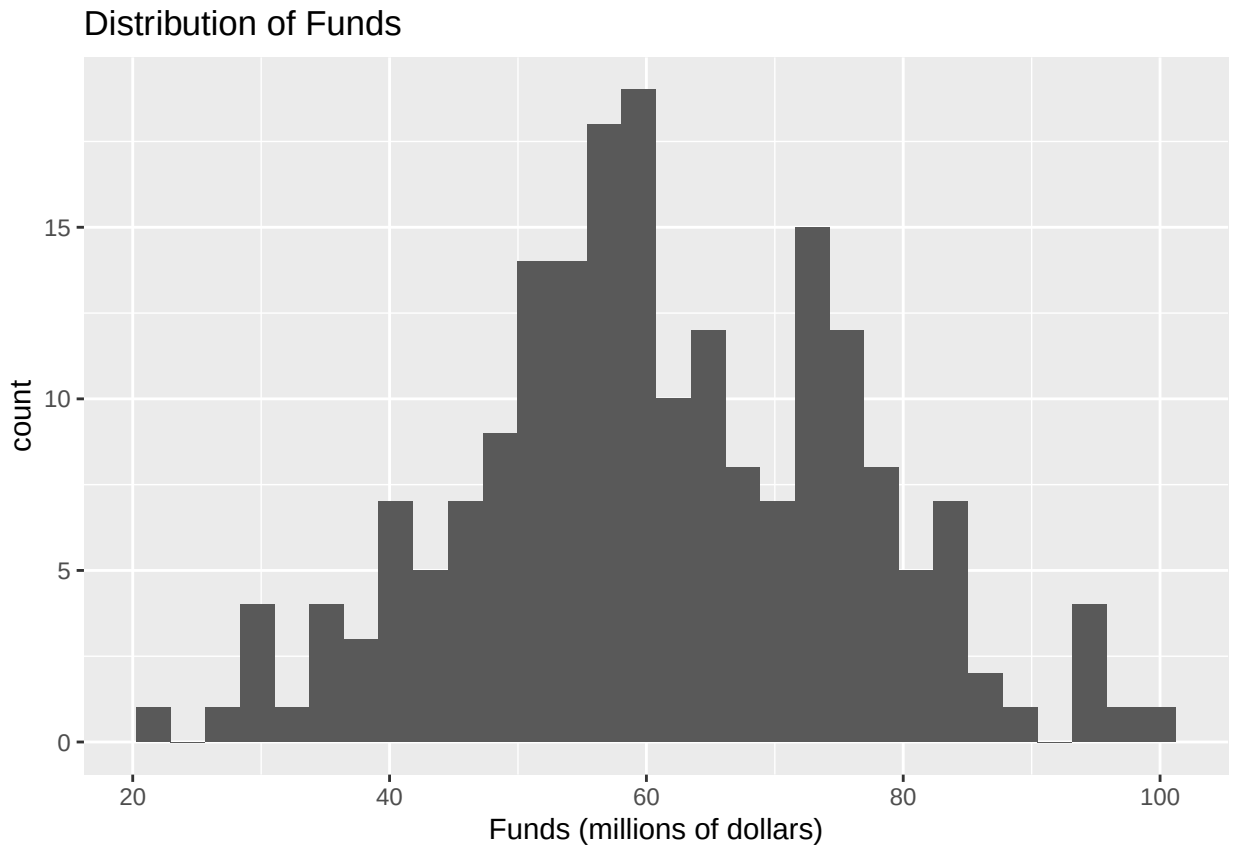
```
summary(dat)
```

```
##   po.dept.code        funds          po.brut
##  Min.   :  1.00   Min.   :21.40   Min.   : 0.00
##  1st Qu.: 50.75   1st Qu.:51.67   1st Qu.:14.00
##  Median :100.50   Median :59.75   Median :19.00
##  Mean   :100.50   Mean   :61.04   Mean   :18.14
##  3rd Qu.:150.25   3rd Qu.:72.17   3rd Qu.:22.00
##  Max.   :200.00   Max.   :99.70   Max.   :29.00
```

Answer: It's interesting that in this dataset, police departments received at least $21.4 million for that year, and average is $61.04 million. It's also interesting to that that at least one department reported 0 incidences of police brutality, and the highest number of incidences reported for that year was 22. It's concerning that the average number of incidences is 18.14.

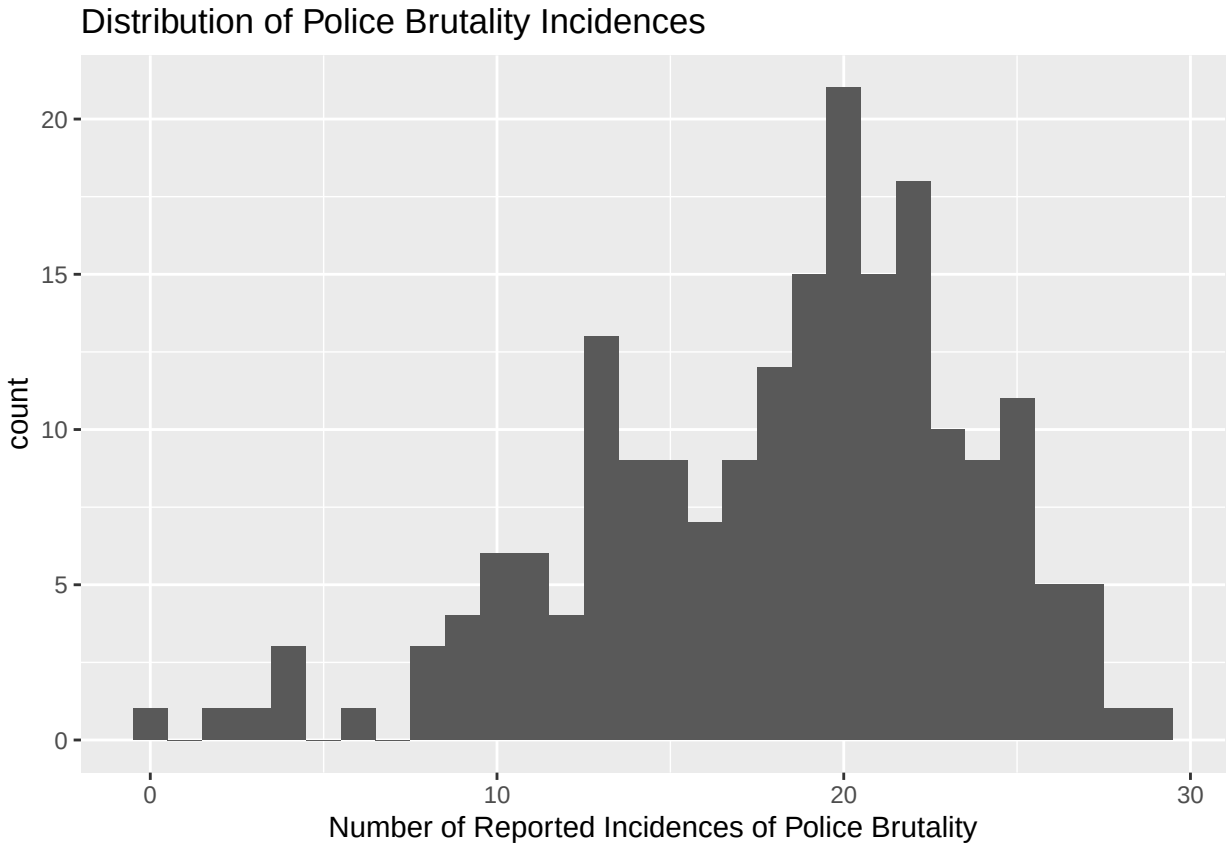
```
dat %>% ggplot(aes(x=funds)) + geom_histogram() +
                               xlab("Funds (millions of dollars)") +
                               ggtitle("Distribution of Funds")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



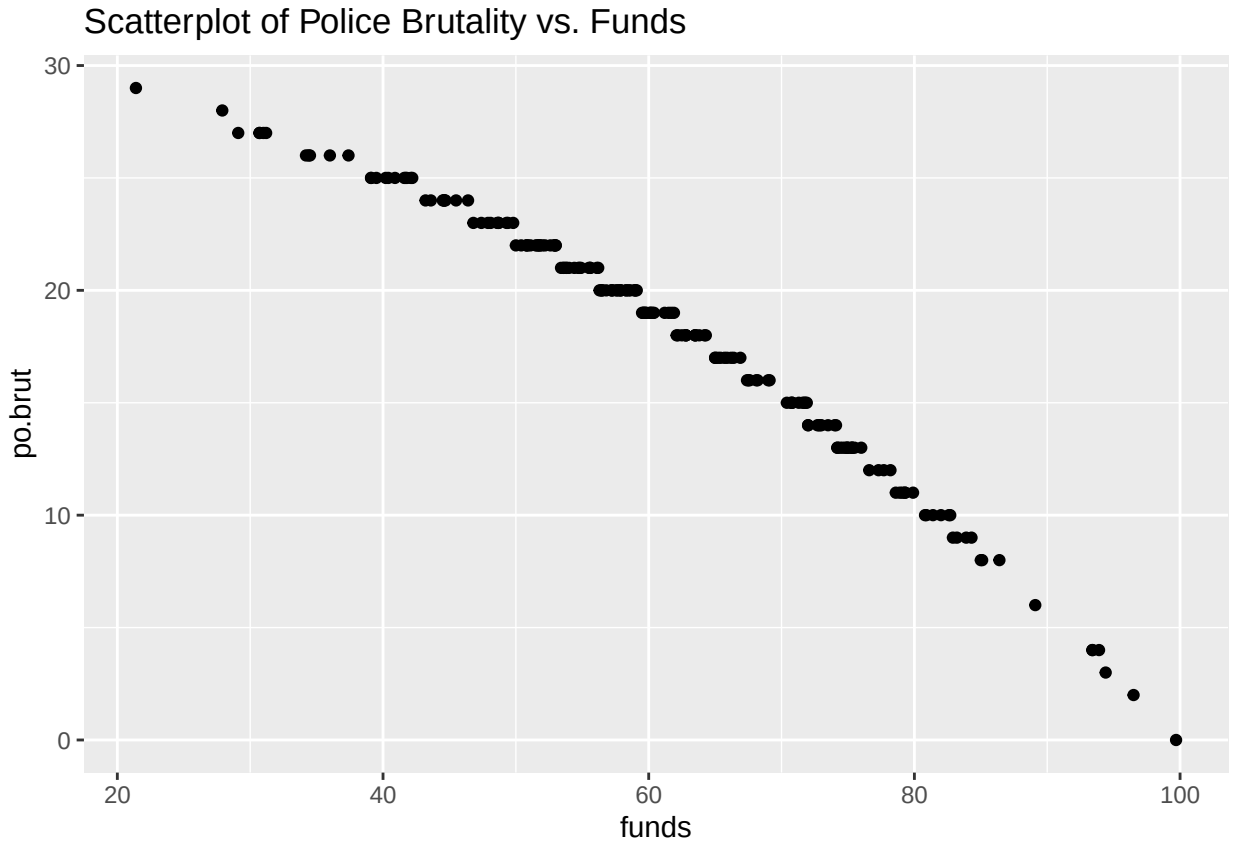

```
dat %>% ggplot(aes(x=po.brut)) + geom_histogram() +
                                 xlab("Number of Reported Incidences of Police Brutality") +
                                 ggtitle("Distribution of Police Brutality Incidences")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Answer: These histograms show possible outliers in both funding and number of police brutality incidences. The distribution of funds looks to be possibly bimodal, and the distribution of police brutality looks to be slightly left skewed. These are subjective observations of the graph, but if given a larger sample, these graphs will be more representative of the population distribution of these variables.

```
dat %>% ggplot(aes(x=funds, y=po.brut)) + geom_point() + ggtitle("Scatterplot of Police Brutality vs. Fu
```




```
cor(dat$funds, dat$po.brut)
```

```
## [1] -0.9854706
```

Answer: There looks to be a strong negative relationship between funds and po.brut, but it's concerning that the relationship doesn't look linear.

Problem 2: Linear regression (30 points)

a. Perform a simple linear regression to answer the question of interest. To do this, name your linear model "reg.output" and write the summary of the regression by using "summary(reg.output)".

```
# Remember to remove eval=FALSE!!
reg.output <- lm(po.brut ~ funds, data = dat)
summary(reg.output)
```

```
##
## Call:
## lm(formula = po.brut ~ funds, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9433 -0.2233  0.2544  0.5952  1.1803
##
## Coefficients:
```

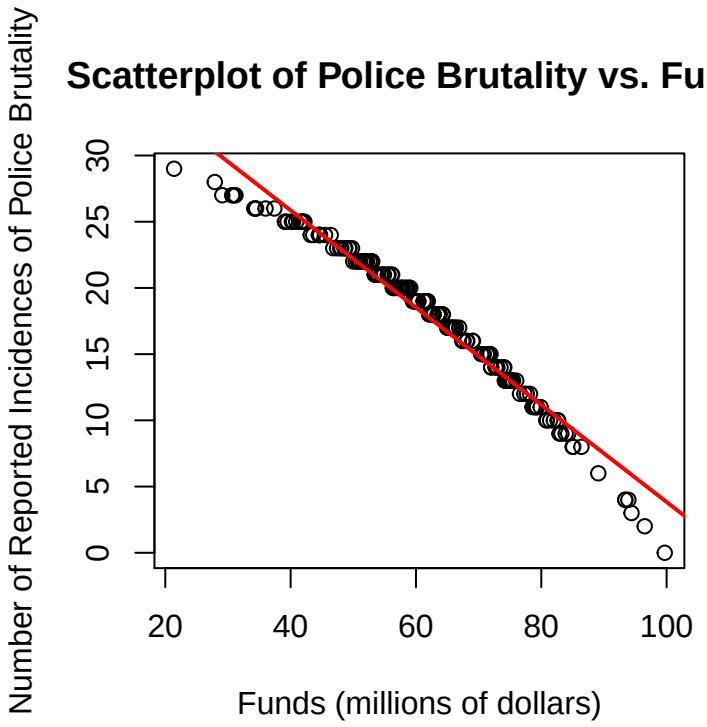
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.543069   0.282503  143.51   <2e-16 ***
## funds       -0.367099   0.004496  -81.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9464 on 198 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.971
## F-statistic:  6666 on 1 and 198 DF,  p-value: < 2.2e-16
```

b. Report the estimated coefficient, standard error, and p-value of the slope. Is the relationship between funds and incidents statistically significant? Explain.

Answer: The estimated coefficient is -0.367099, standard error is 0.004496, and the p-value is <2e-16, which is almost 0. Since the p-value is less than alpha (0.05), we are able to reject the null hypothesis that there is no association between police brutality and funds which means this relationship is statistically significant. Additionally the R-squared of the model is very high which indicates this model fits the data well.

c. Draw a scatterplot of po.brut (y-axis) and funds (x-axis). Right below your plot command, use abline to draw the fitted regression line, like this:

```
# Remember to remove eval=FALSE!!
plot(dat$funds, dat$po.brut, xlab = "Funds (millions of dollars)", ylab = "Number of Reported Incidences
abline(reg.output, col = "red", lwd=2)
```



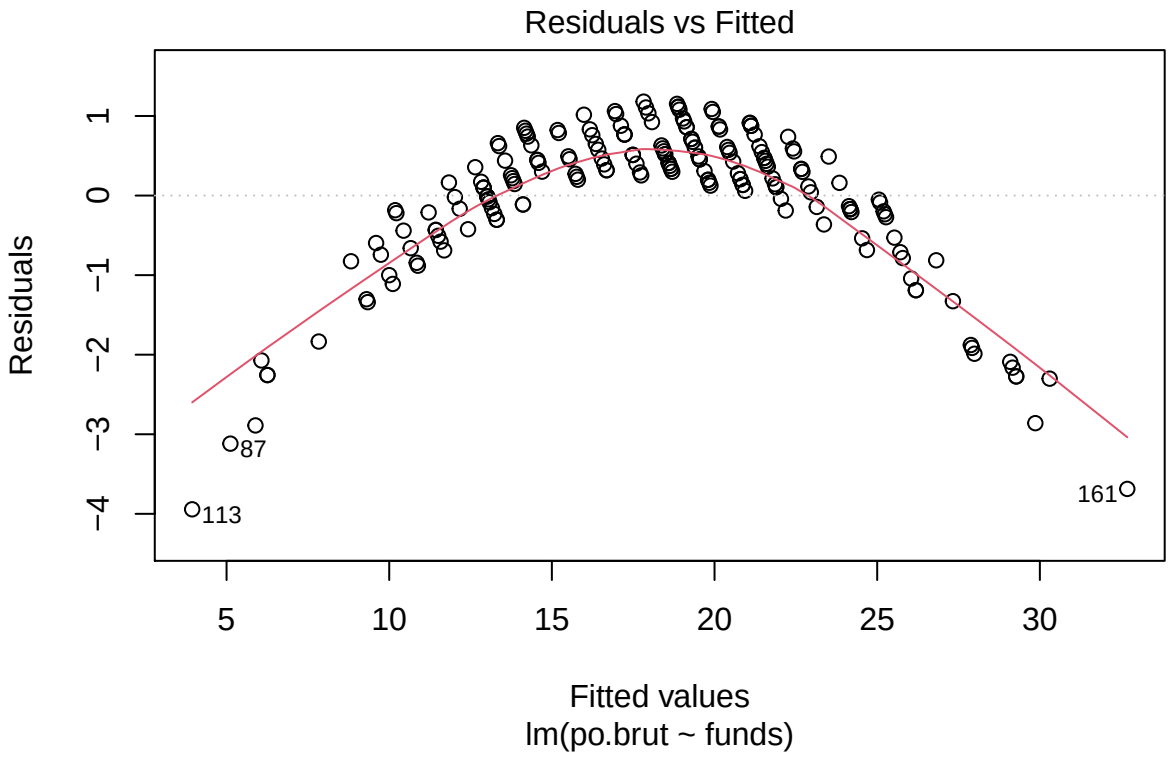
Does the line look like a good fit? Why or why not?

Answer: No, it does not show a good fit because the data looks curved, and since this is a linear model, it does not fit the data properly. The R-squared is high because the linear fit does model quite a bit of the data, but it fails to capture the overall trend. If there were

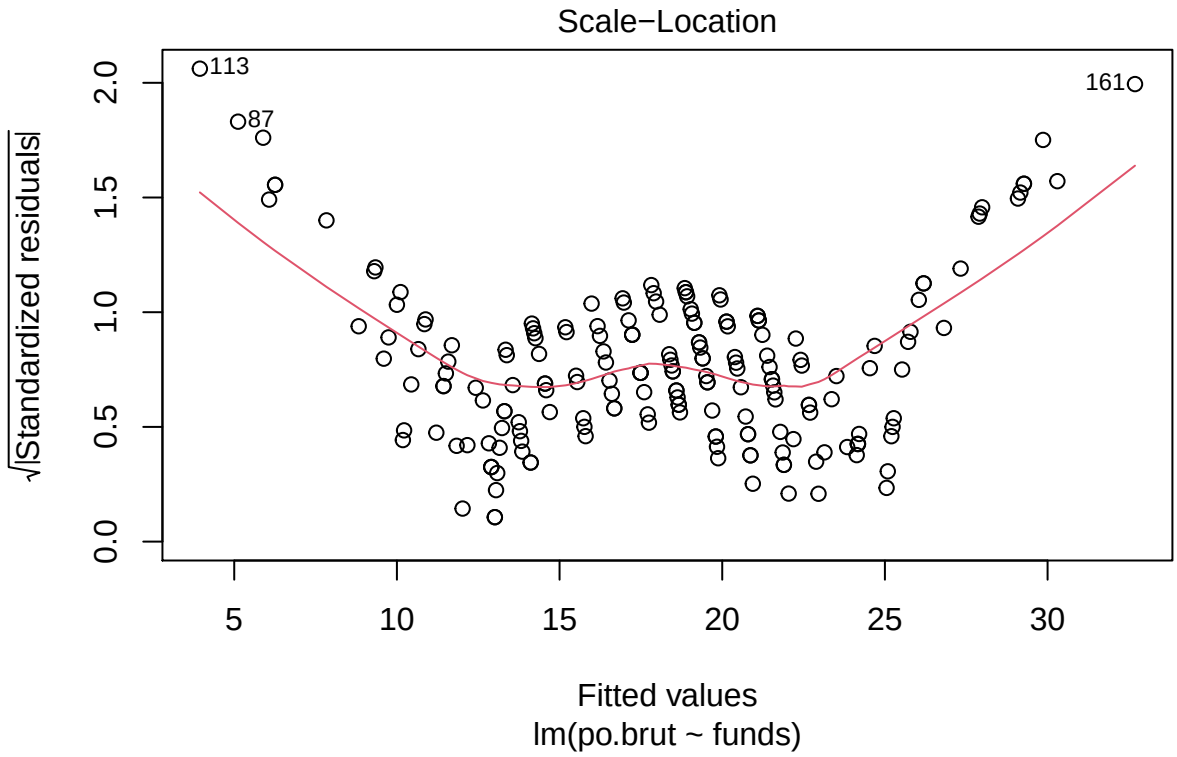
more observations, this might be the case. Perhaps transformation would be appropriate to apply to capture the fit of the data, but this runs the risk of overfitting. Given the size of the dataset, it would be most beneficial to split the data into test and training data to cross validate.

d. Are the four assumptions of linear regression satisfied? To answer this, draw the relevant plots. (Write a maximum of one sentence per assumption.) If not, what might you try to do to improve this (if you had more time)?

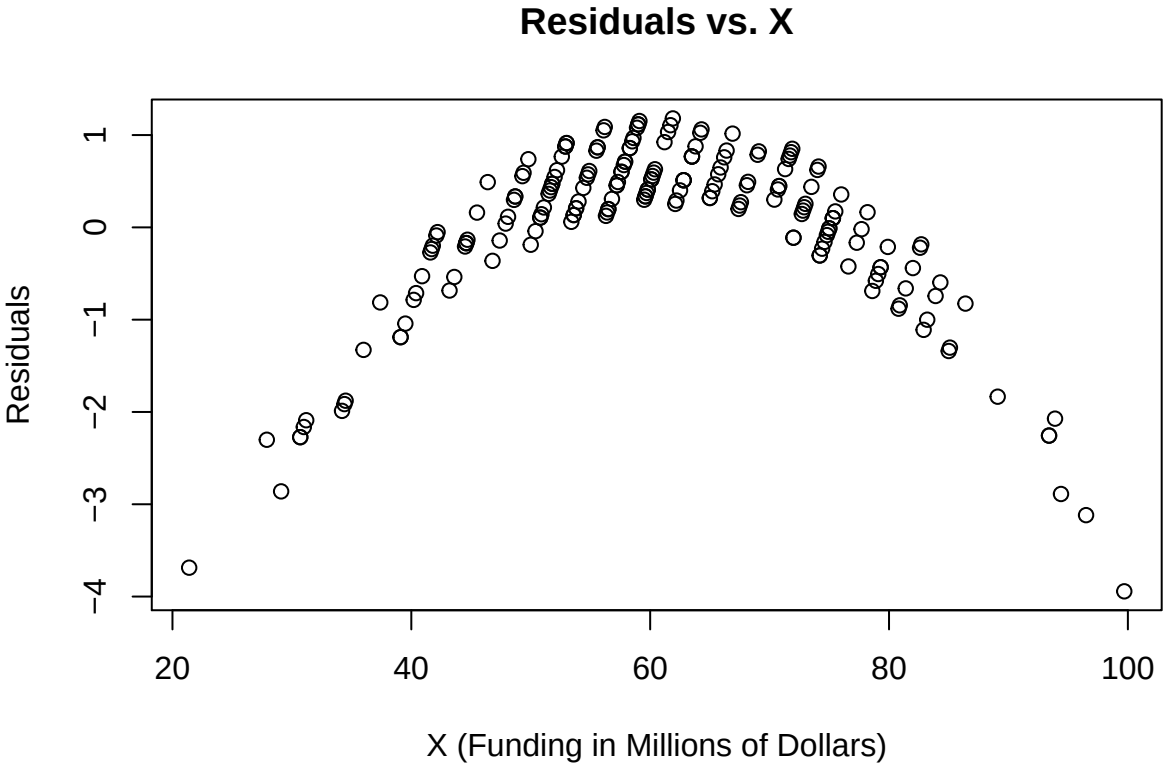
```
#residuals vs fitted plots
plot(reg.output, which = 1)
```




```
#scale location plot
plot(reg.output, which = 3)
```



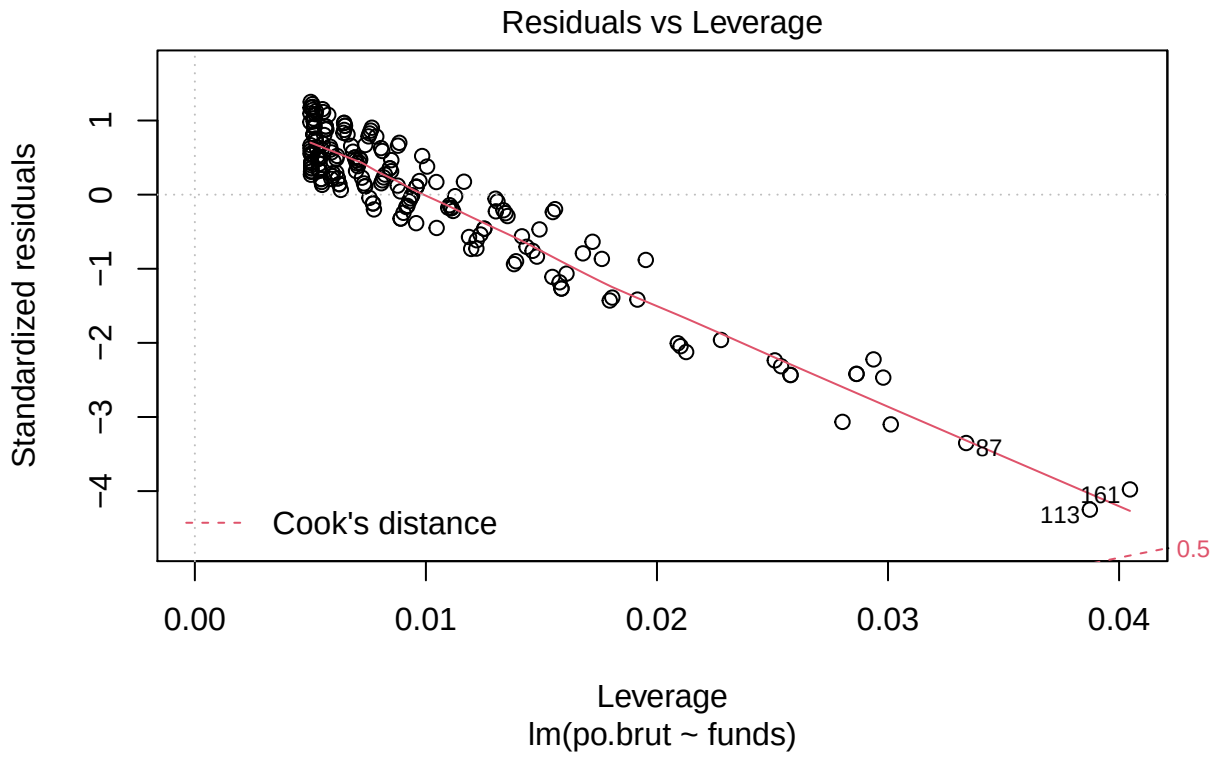

```
#residuals vs x plot
plot(dat$funds, reg.output$residuals, main = "Residuals vs. X", xlab = "X (Funding in Millions of Dolla
```



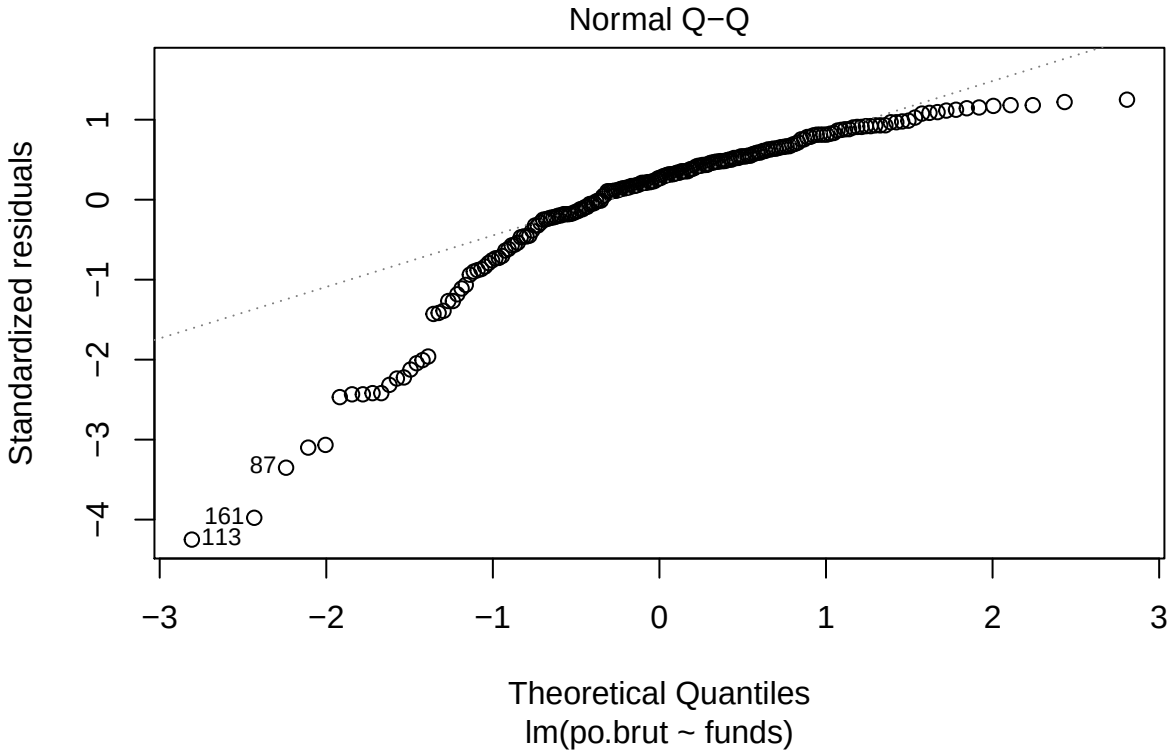
Answer: The linearity assumption is not met because the scatterplot of po.brut vs. funds shows a curved relationship over a linear one, and the residuals vs. x and residuals vs. fitted plots

show a distinguishable curved pattern. The independence assumption is not met because there is a pattern shown in the residuals vs. x plot. The equal variance assumption is not satisfied since there are significant trends in the residuals vs. fitted and scale location plots.

```
plot(reg.output, which = 5)
```




```
plot(reg.output, which = 2)
```



Answer: The normality assumption is violated because the QQ plot does not show a straight line like we want. Since none of the assumptions are met, I would apply a transformation to the model to see if it will help satisfy the assumptions, and to choose which transformation, I would use the Box-Cox method.

e. Answer the question of interest based on your analysis.

Answer: Since none of the assumptions are met, the p-value obtained by the model are optimistic which means the model can't be used for inference. If the assumptions were met, there would be an associated decrease of reported incidences of police brutality by 0.367099 when funding increases by $1 million. It is possible that it works well for prediction, but as mentioned above, the best way to make sure is to partition the data into test and train. However, with a larger sample size or with transformations, it's possible the model can be used for statistical inference.

Problem 3: Data ethics (10 points)

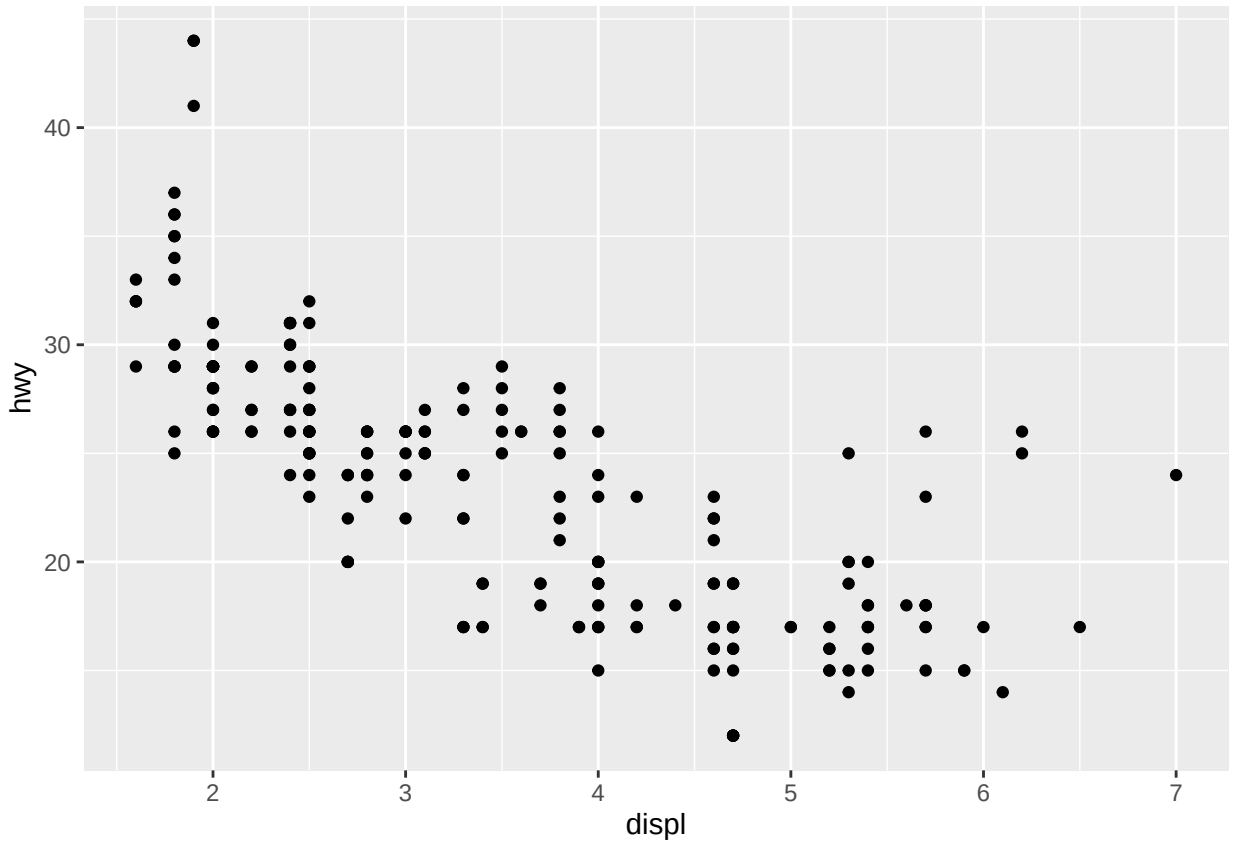
Describe the dataset. Considering our lecture on data ethics, what concerns do you have about the dataset? Once you perform your analysis to answer the question of interest using this dataset, what concerns might you have about the results?

Answer: This dataset contains the funding in millions of dollars for the given year as well as the number of reported incidences of police brutality report by each police department. As discussed in class, the databases of police departments are prone to bias due to the biases of officers. This shows up in predictive policing where the predicted area of crime is over-represented. Since the incidences of police brutality are reported by each department, this data may be biased and therefore produce biased results (garbage in, garbage out). This data might not completely capture the true amount of police brutality occurring since it may be underreported. It would be interesting if a separate organization is able to cross check this data with a different database of police brutality that is not recorded by police. This would hold police departments accountable in recording data and it would provide a level of fairness.

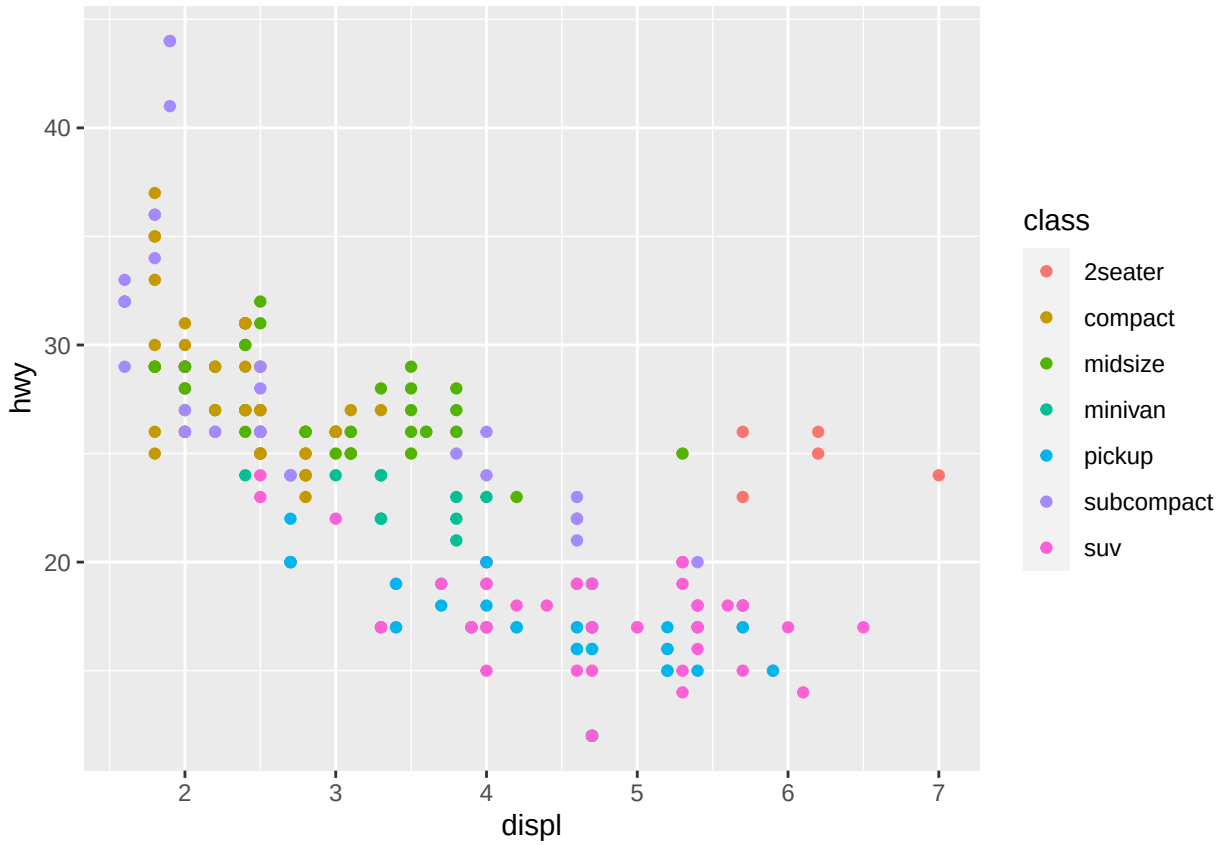
Assignment 4

```
#install and load tidyverse library
#install.packages("tidyverse")
library(tidyverse)
```

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy)) #plot a scatterplot of hwy vs displ
```

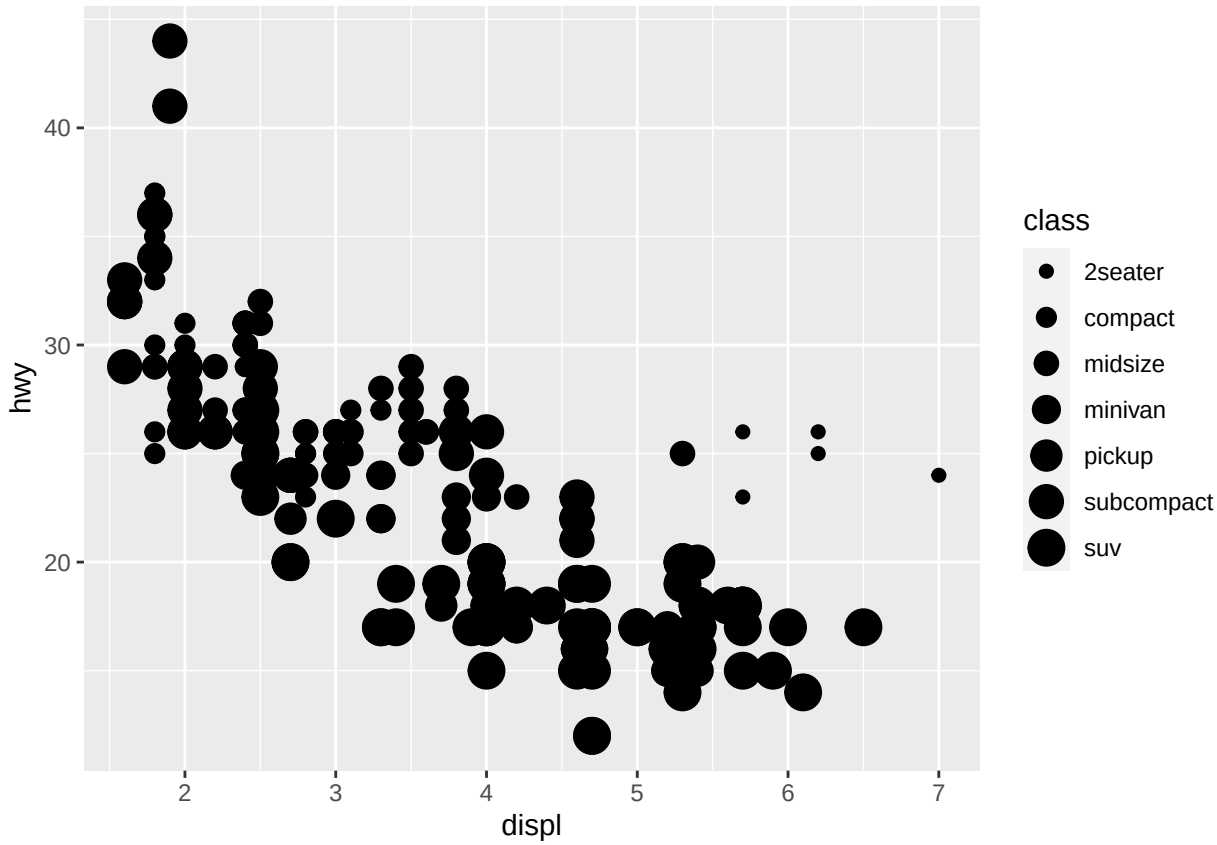



```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy, color = class)) #plot a scatterplot of hwy vs displ by c
```



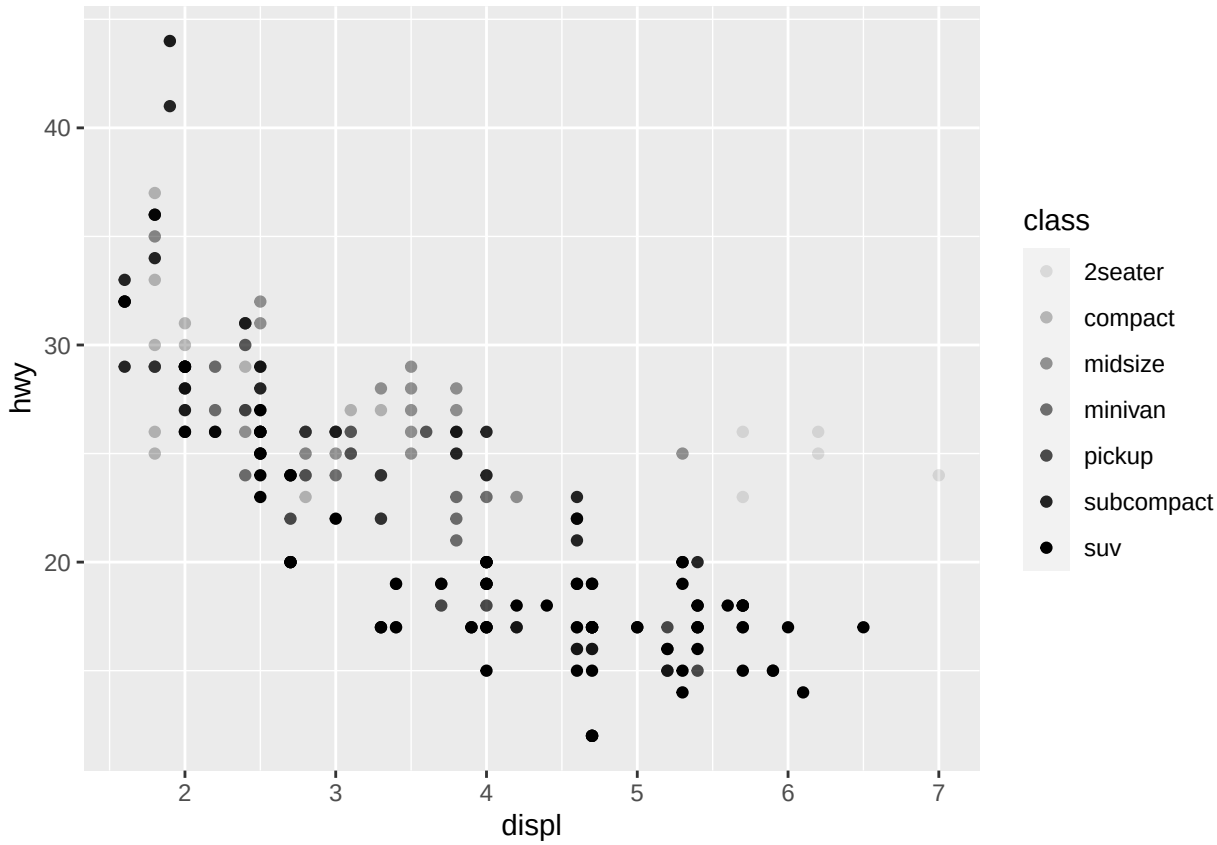

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy, size = class)) #plot a scatterplot of hwy vs displ by cl
```

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




```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class)) #plot a scatterplot of hwy vs displ with
```

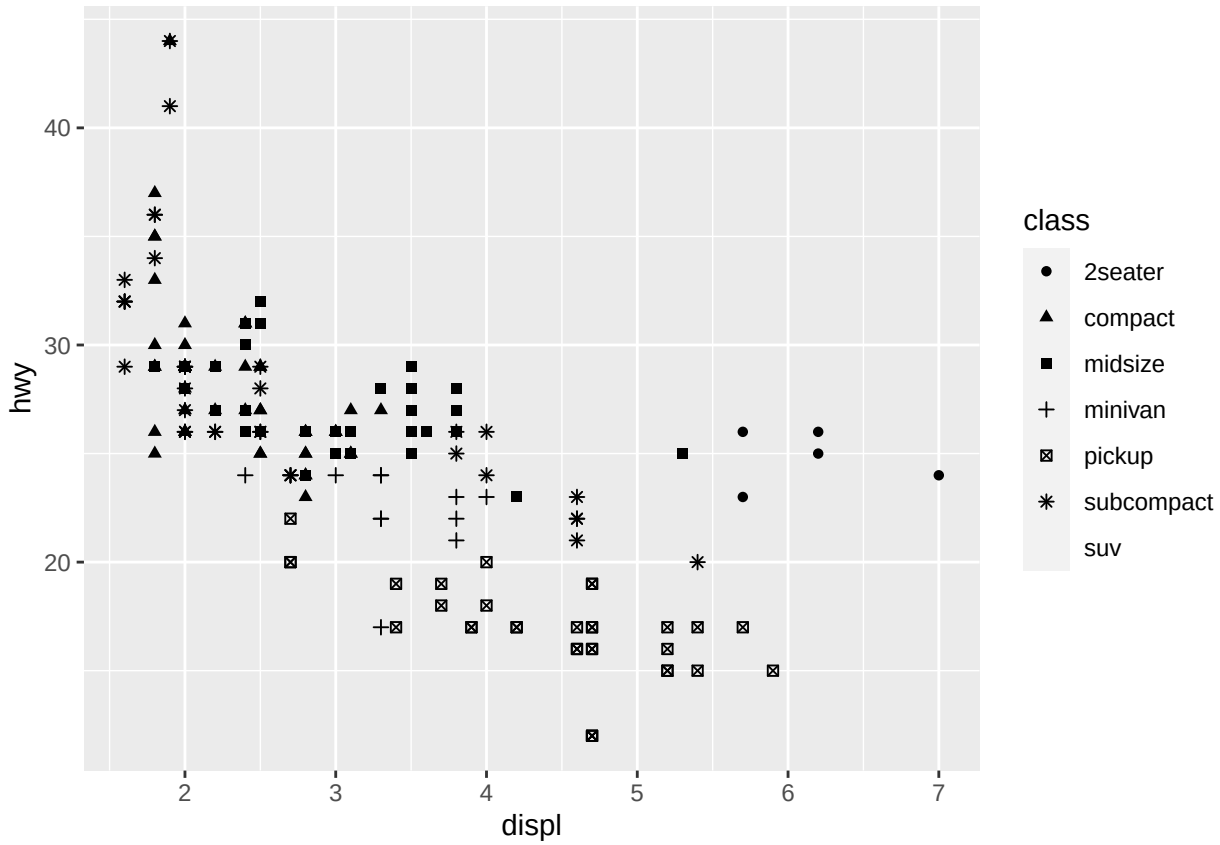
```
## Warning: Using alpha for a discrete variable is not advised.
```



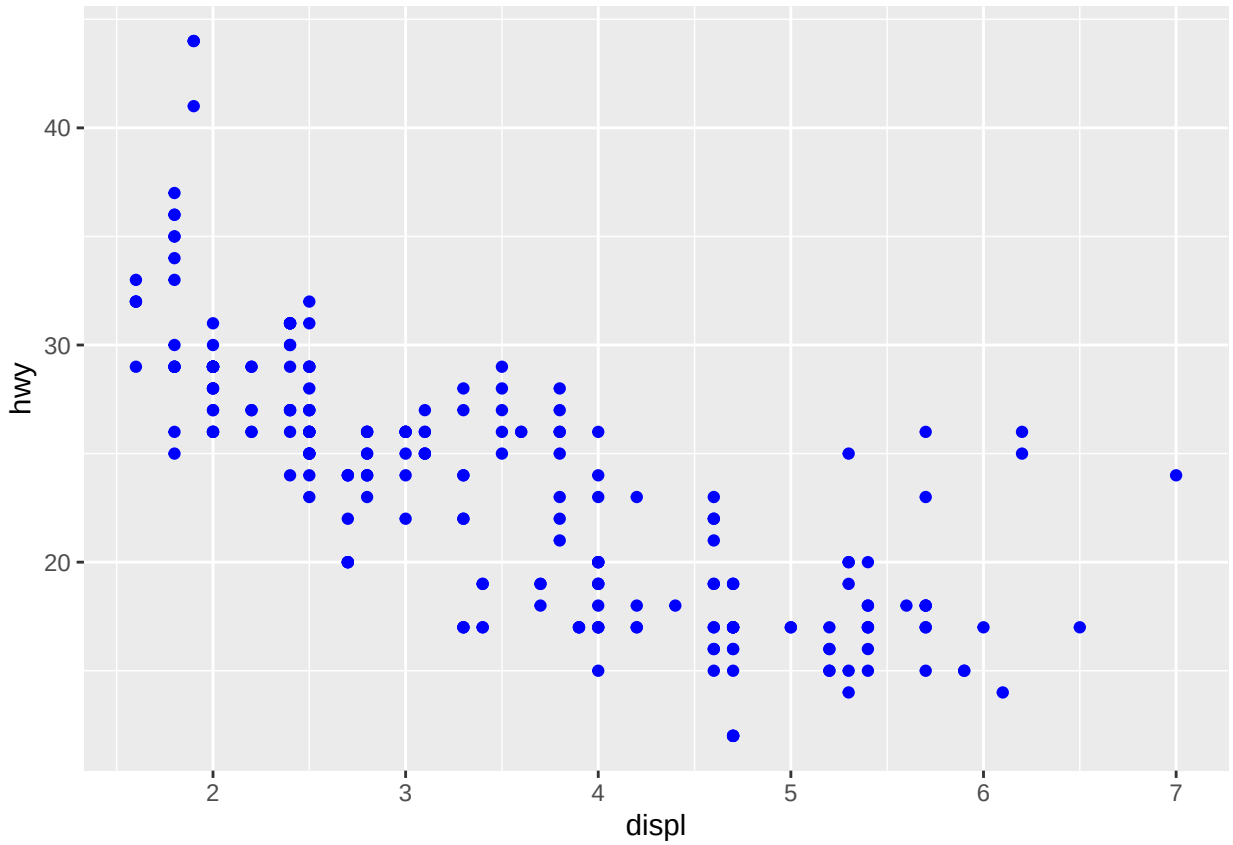

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy, shape = class)) #plot a scatterplot of hwy vs displ with
```

```
## 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.
```

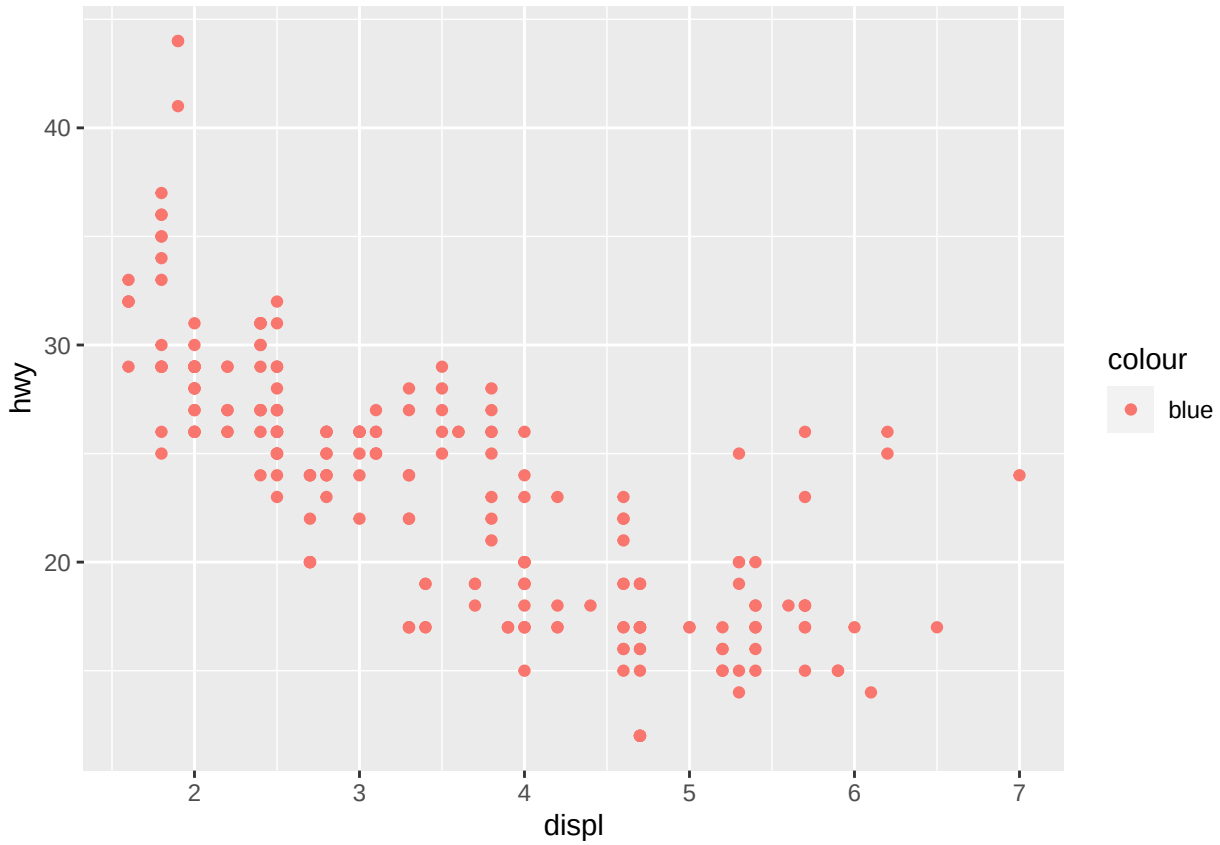
```
## Warning: Removed 62 rows containing missing values (geom_point).
```




```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue") #plot a scatterplot of hwy vs displ wit
```

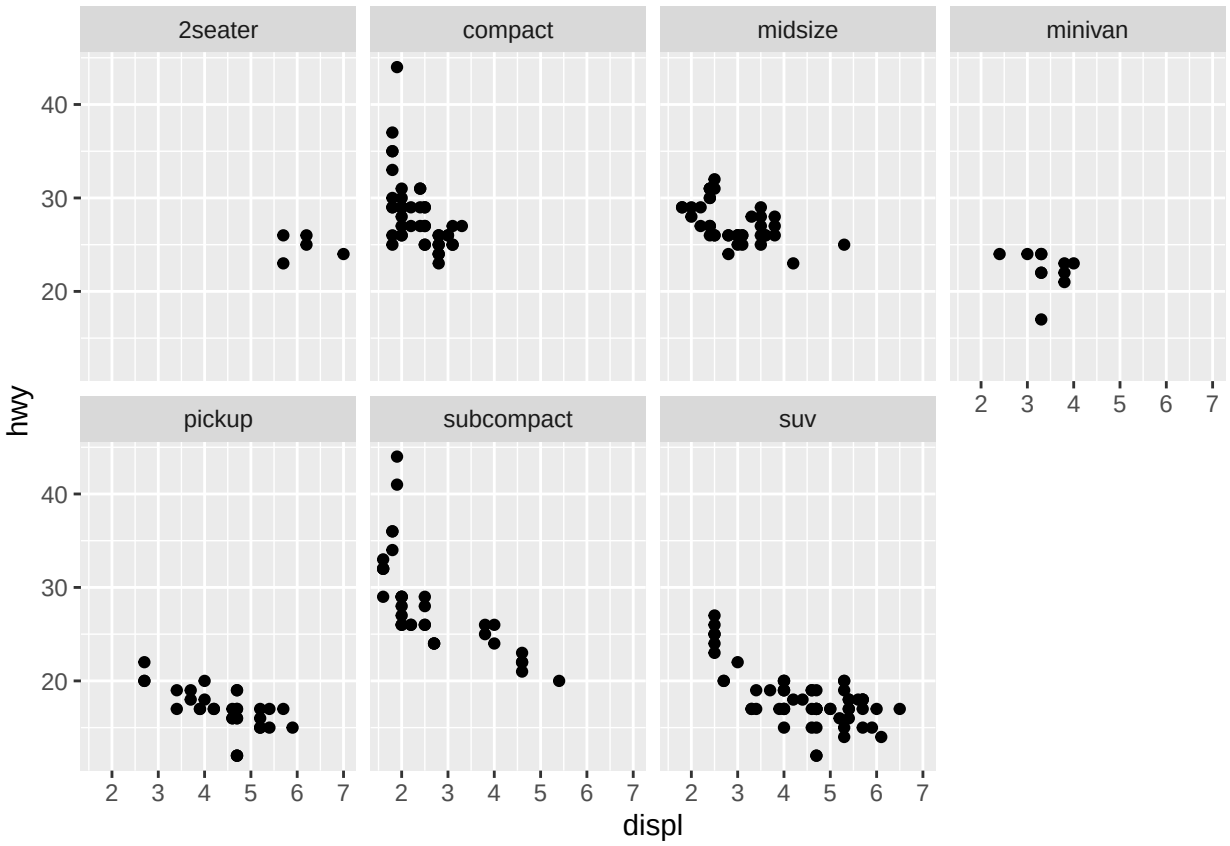



```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue")) #plot a scatterplot of hwy vs displ wit
```

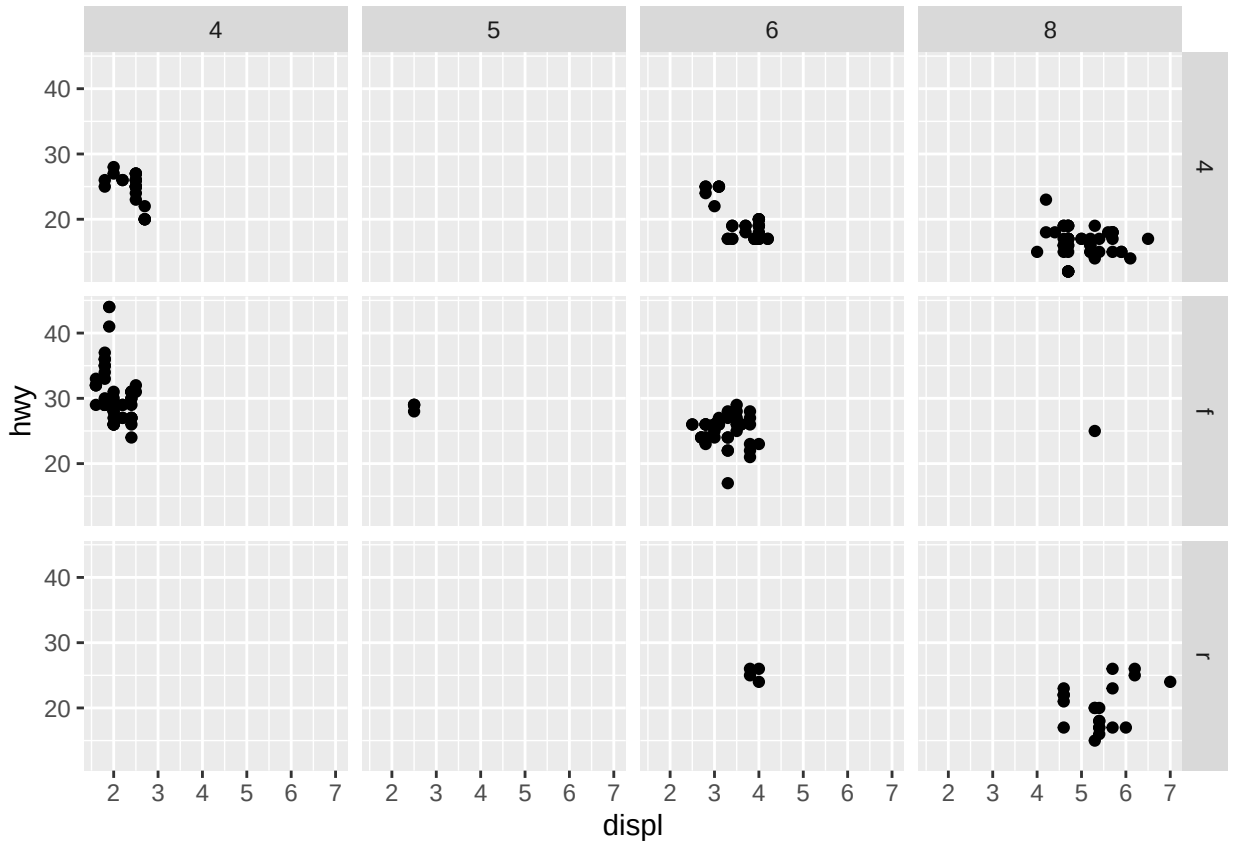



```
#do not put "+" at the beginning of a line, always at the end
#ggplot(data = mpg) #call ggplot function and load data as argument
#+ geom_point(mapping = aes(x = displ, y = hwy))
```

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy)) + #plot a scatterplot of hwy vs displ
  facet_wrap(~ class, nrow = 2) #create multiple scatterplots of hwy vs displ across different class ca
```

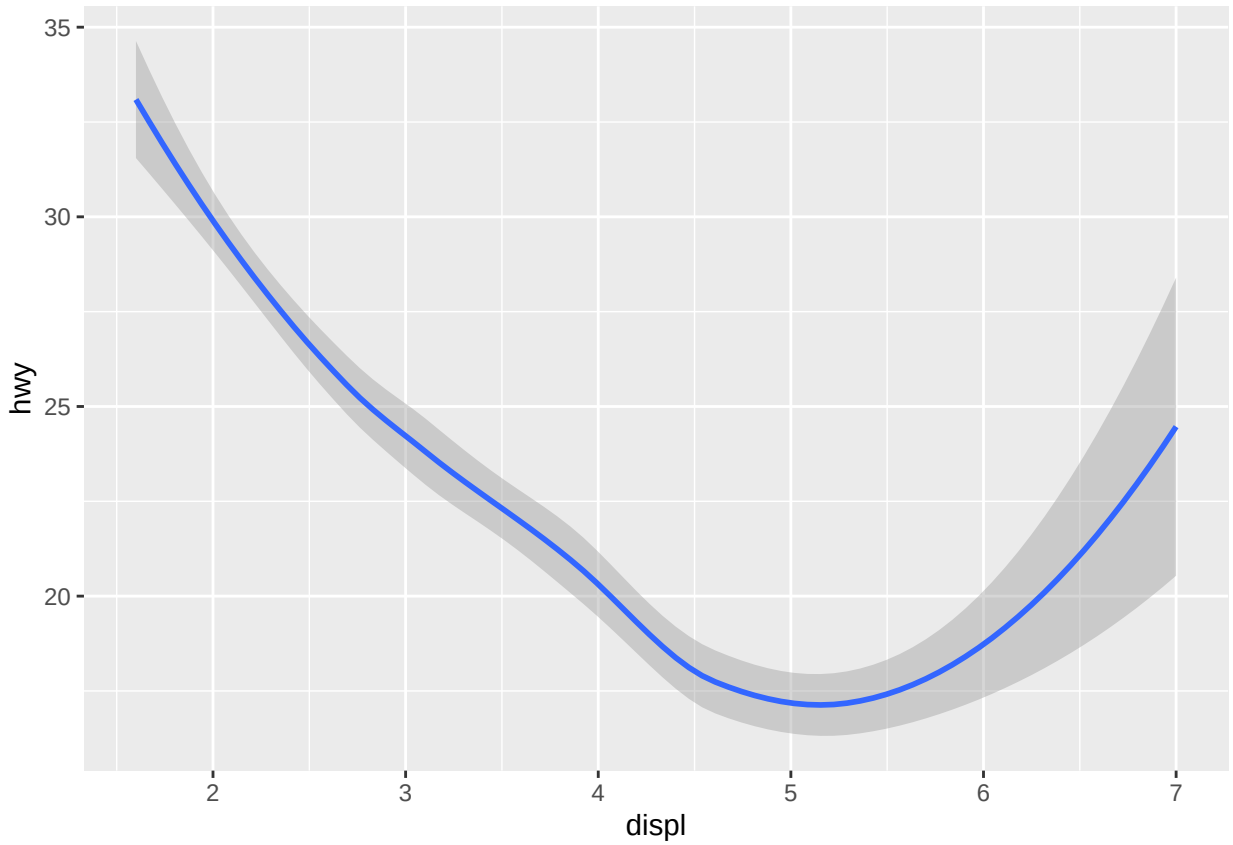



```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy)) + #plot a scatterplot of hwy vs displ
  facet_grid(drv ~ cyl) #create multiple scatterplots of hwy vs displ across different categories (drv
```



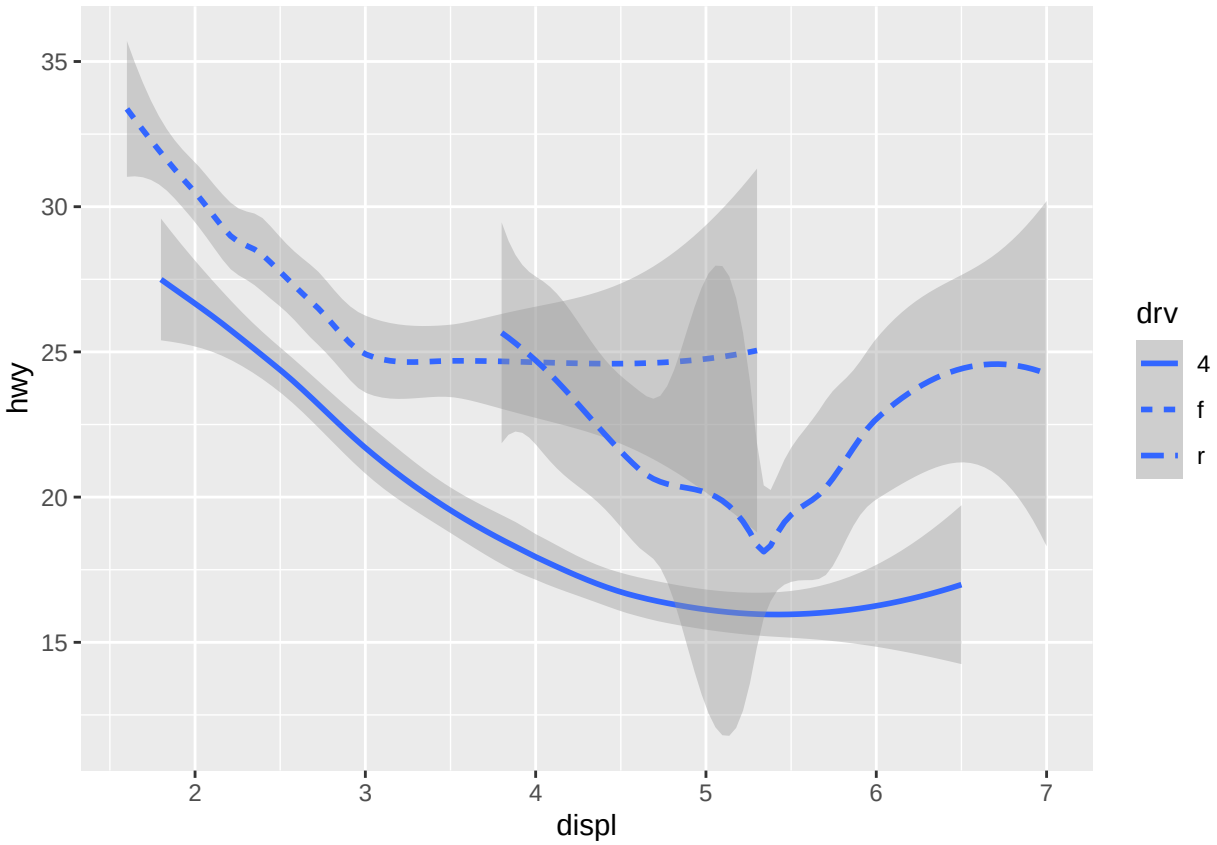

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_smooth(mapping = aes(x = displ, y = hwy)) #plot a line of best fit of hwy vs displ
```

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



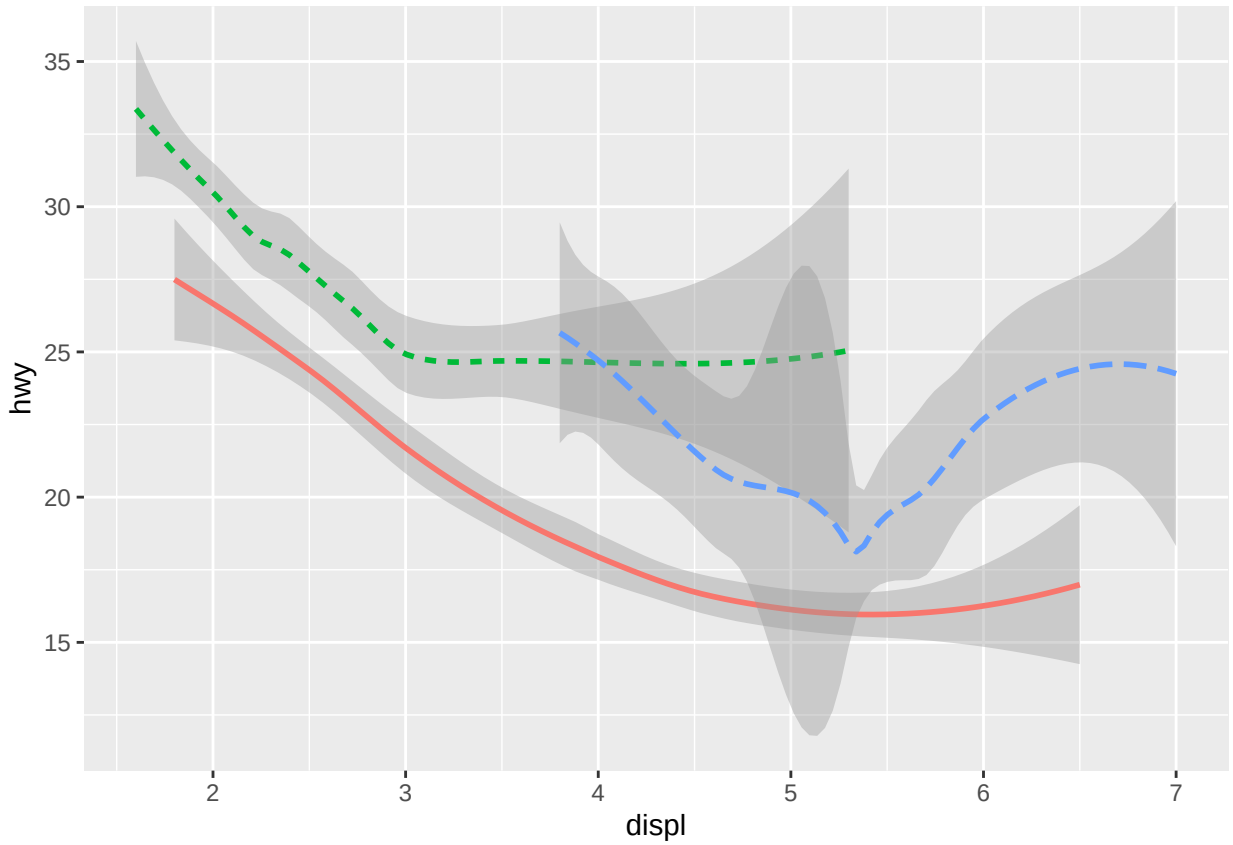

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv)) #plot a line of best fit of hwy vs dis
```

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



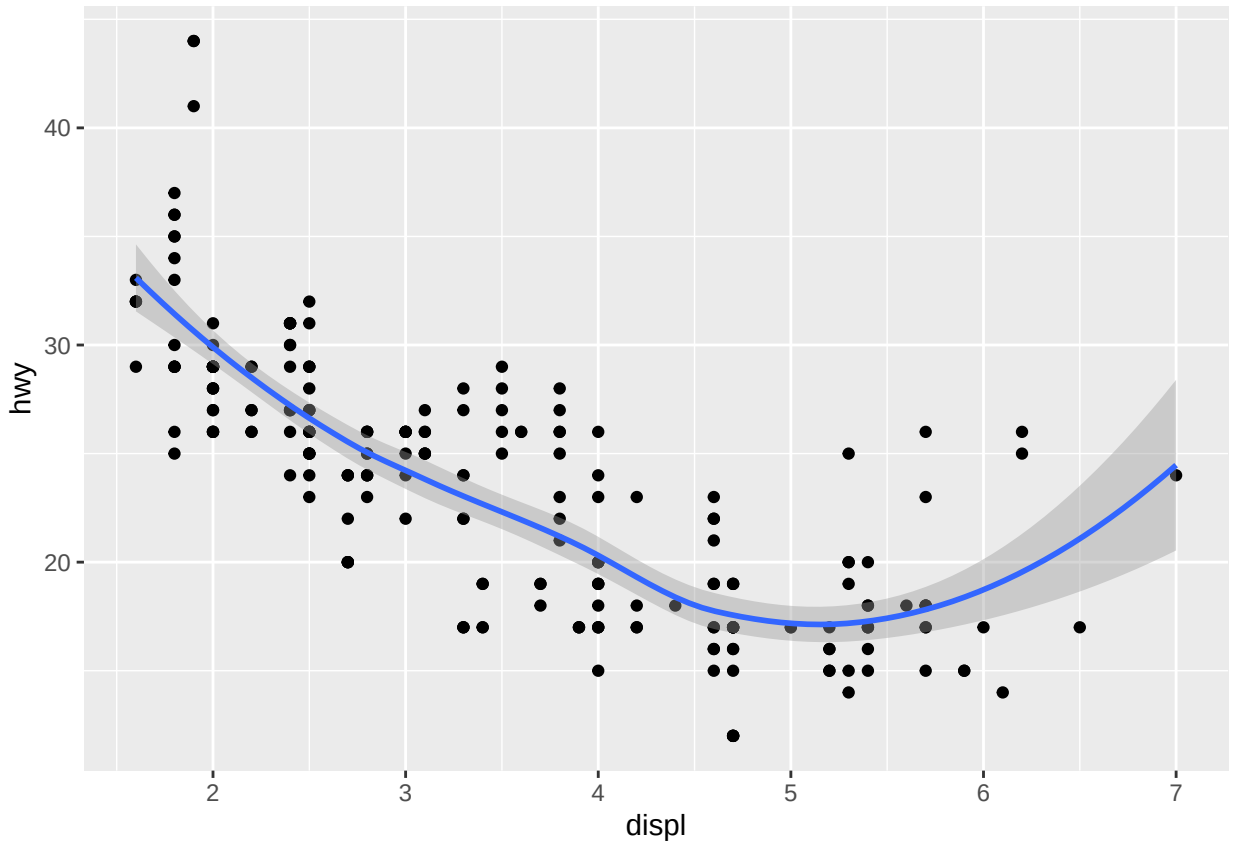

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv, color = drv), show.legend = FALSE) #plo
```

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



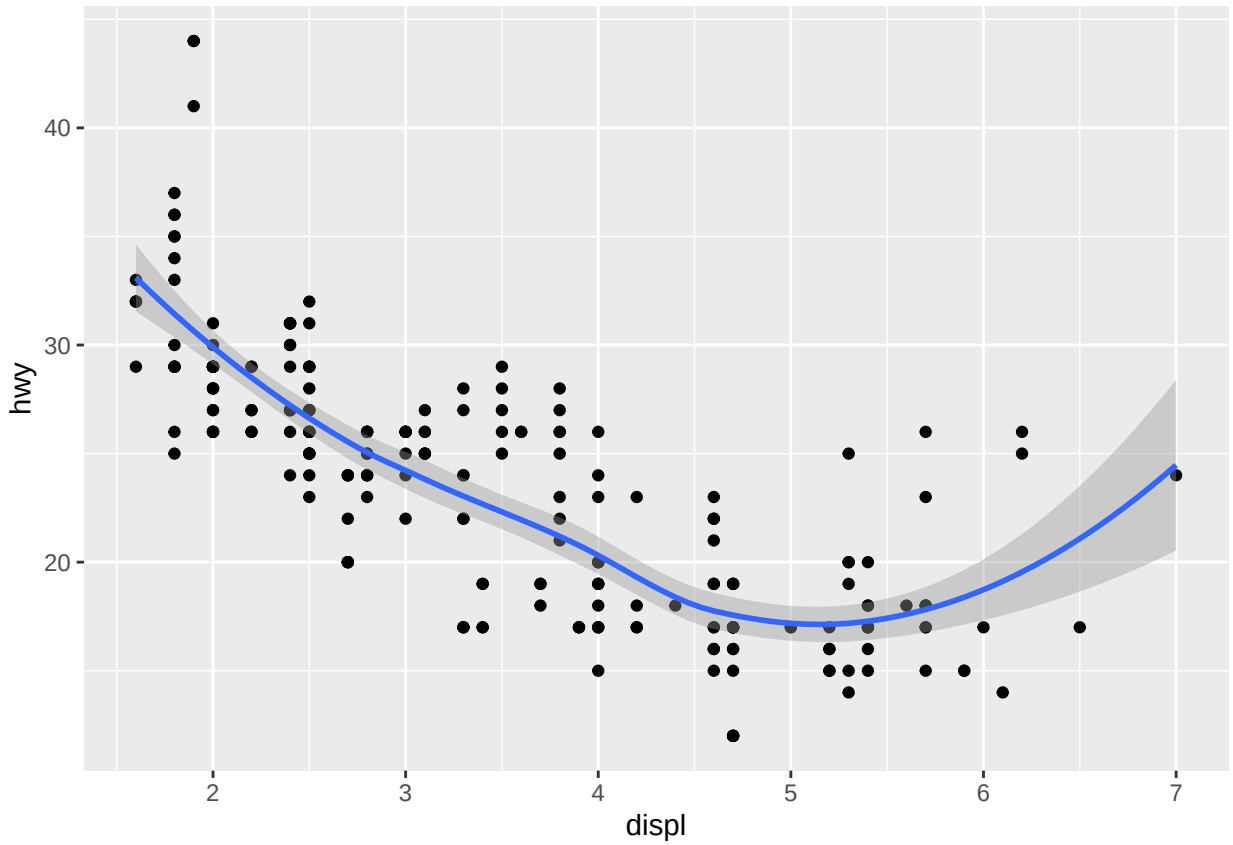

```
ggplot(data = mpg) + #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy)) + #plot a scatterplot of hwy vs displ
  geom_smooth(mapping = aes(x = displ, y = hwy)) #adding a line of best fit on top of scatterplot
```

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



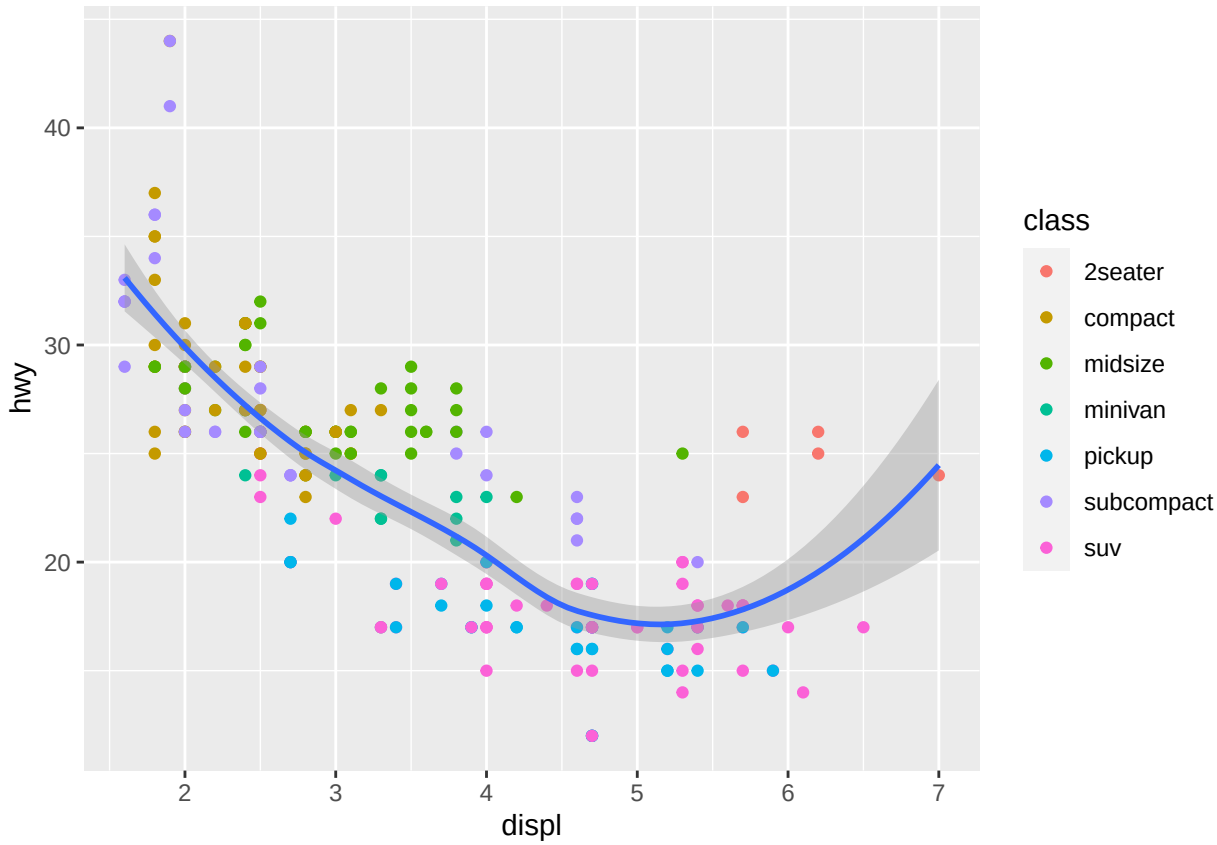

```
#another way of writing the previous function
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) + #creates aesthetics of the plot
  geom_point() + #scatterplot of hwy vs displ
  geom_smooth() #line of fit over scatterplot
```

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



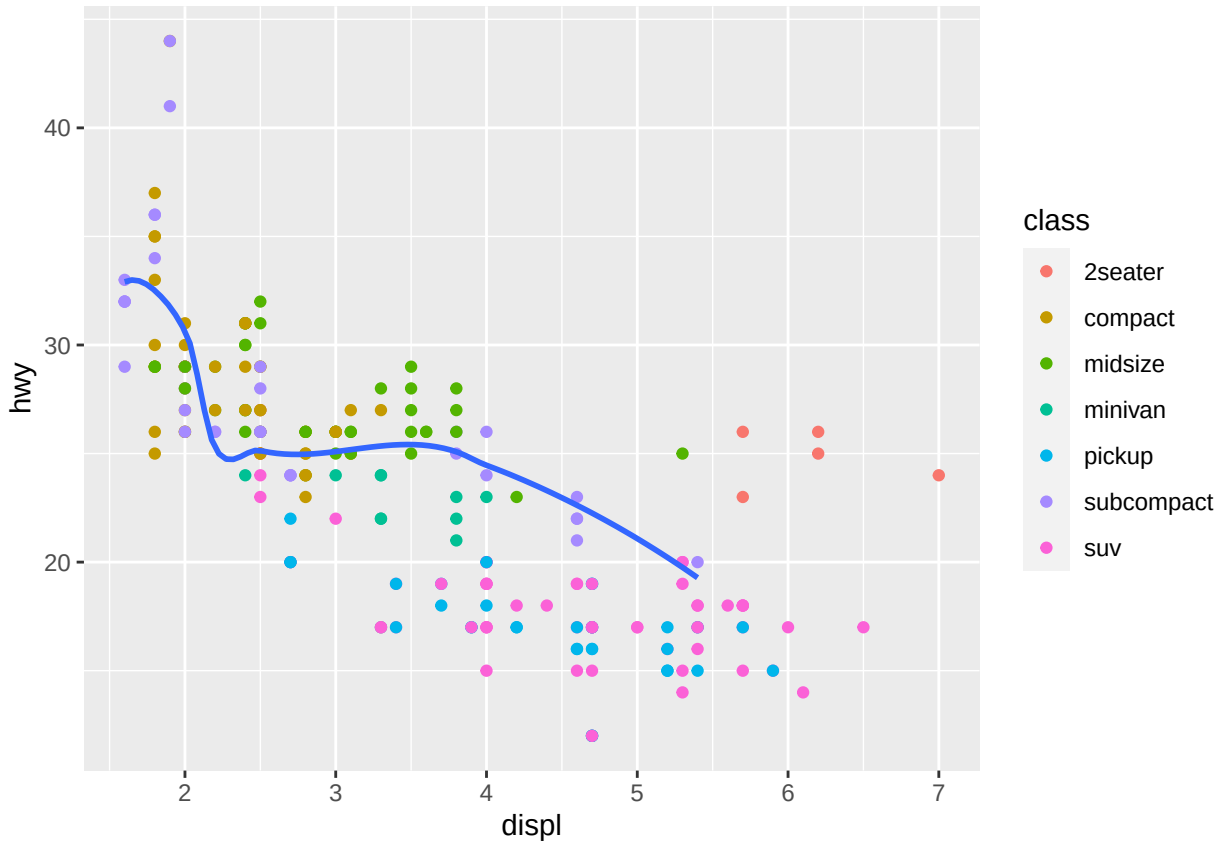

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) + #creates aesthetics of the plot
  geom_point(mapping = aes(color = class)) +#scatterplot of hwy vs displ with color dividing classes
  geom_smooth() #line of fit over scatterplot
```

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

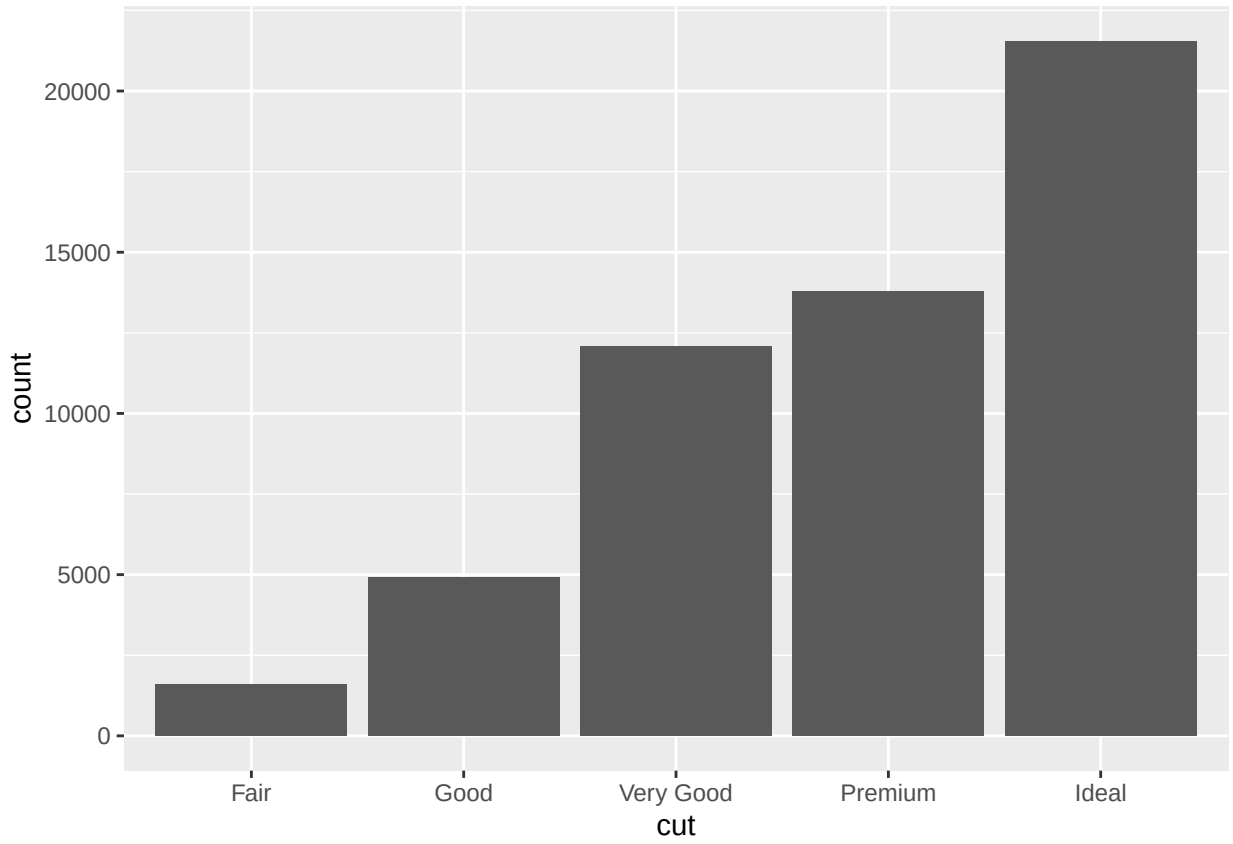



```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) + #creates aesthetics of the plot
  geom_point(mapping = aes(color = class)) +#scatterplot of hwy vs displ with color dividing classes
  geom_smooth(data = filter(mpg, class == "subcompact"), se = FALSE) #line of fit over scatterplot for
```

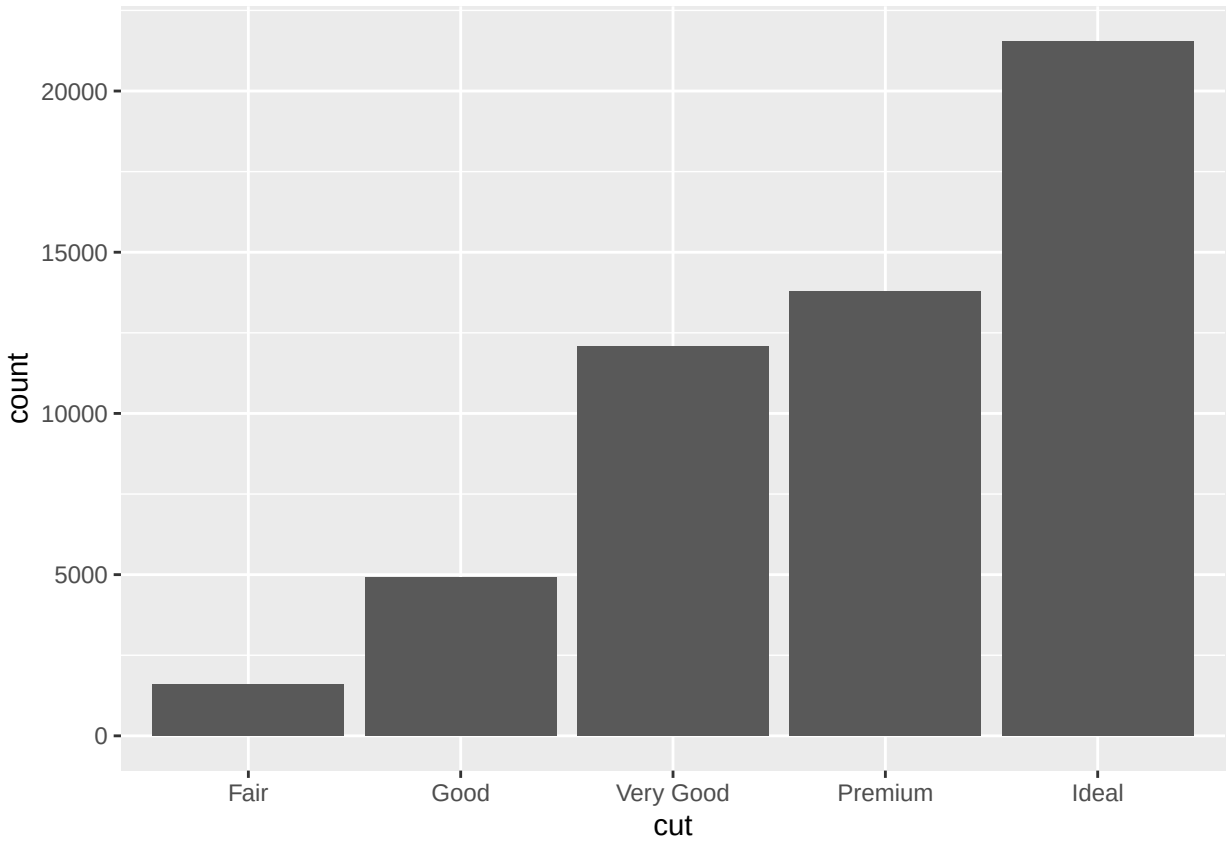
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



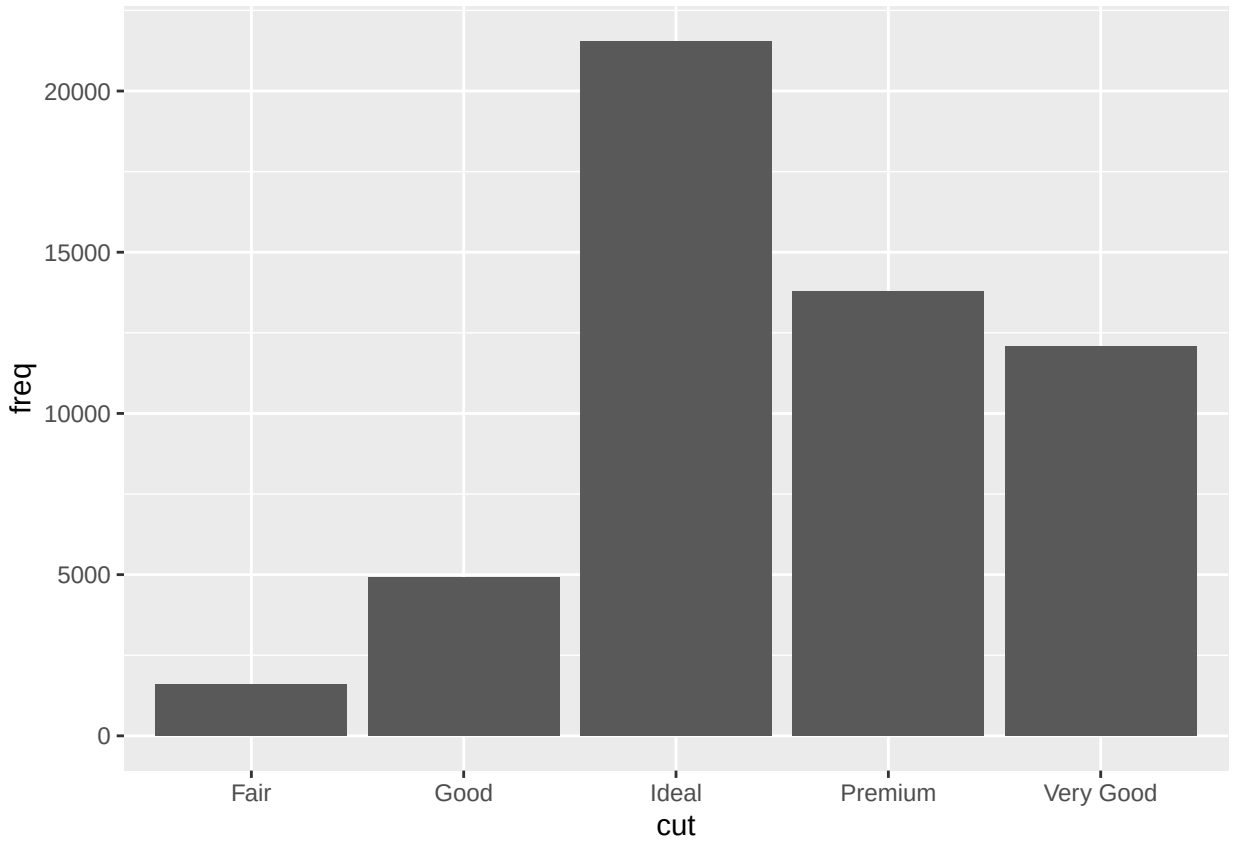

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut)) #bar chart of diamond cuts
```



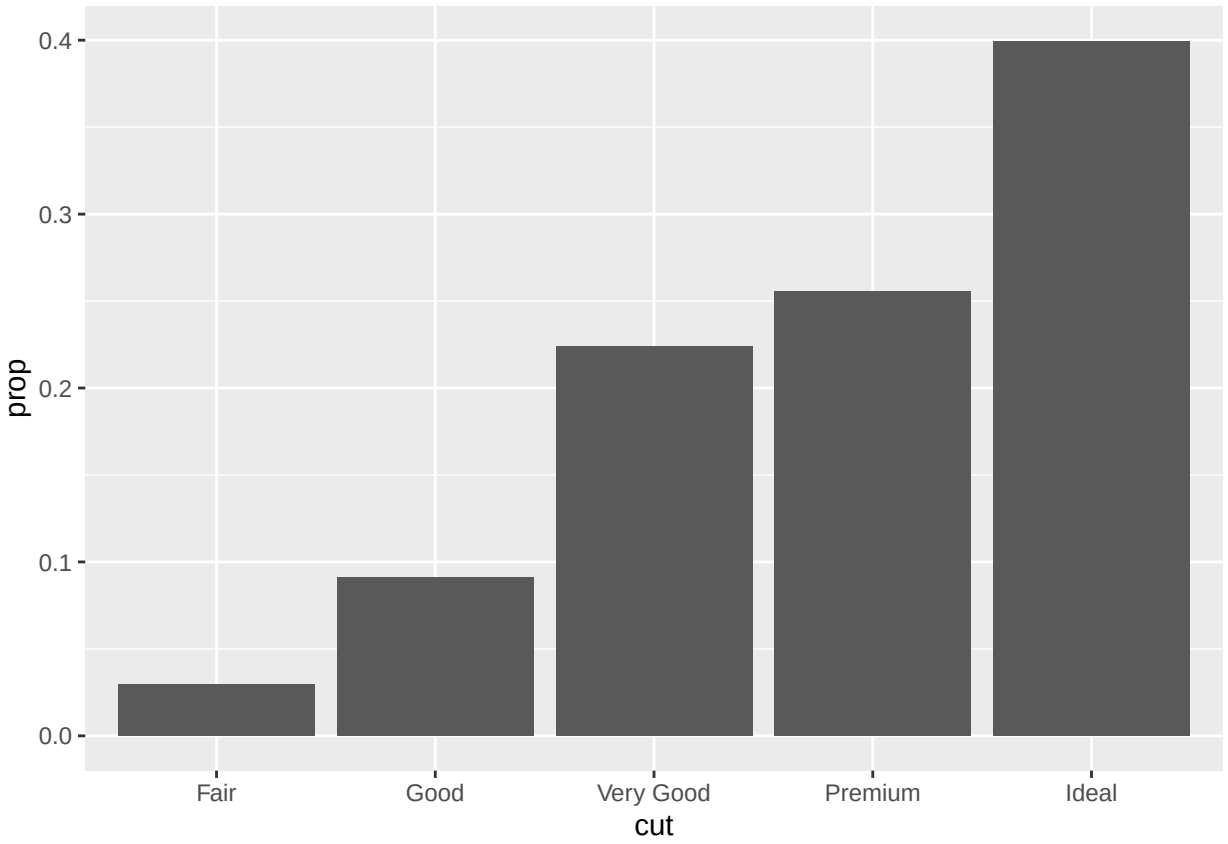

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  stat_count(mapping = aes(x = cut)) #bar chart of diamond cuts using stat_count
```



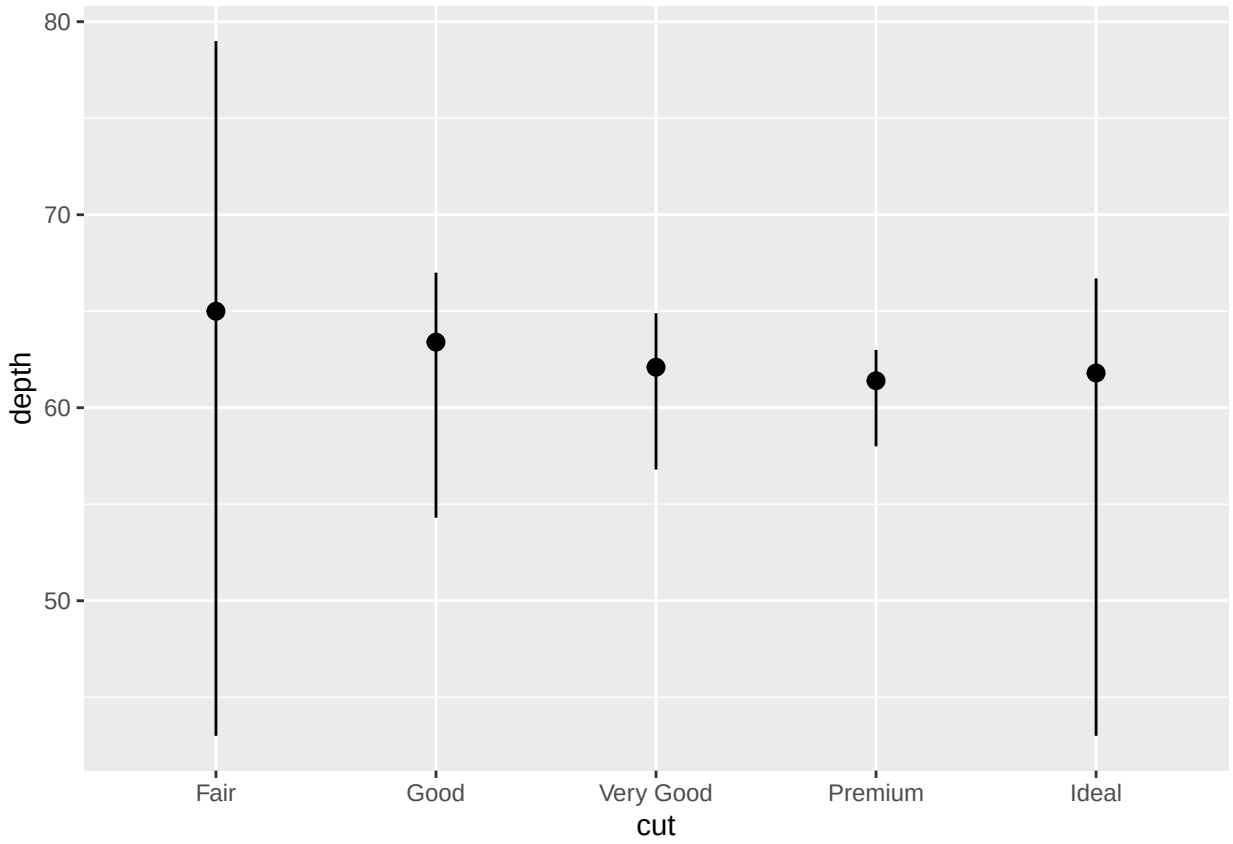

```
demo <- tribble( #create data
  ~cut,          ~freq,
  "Fair",        1610,
  "Good",        4906,
  "Very Good",   12082,
  "Premium",     13791,
  "Ideal",       21551
)

ggplot(data = demo) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, y = freq), stat = "identity") #bar chart of diamond cuts and frequenc
```



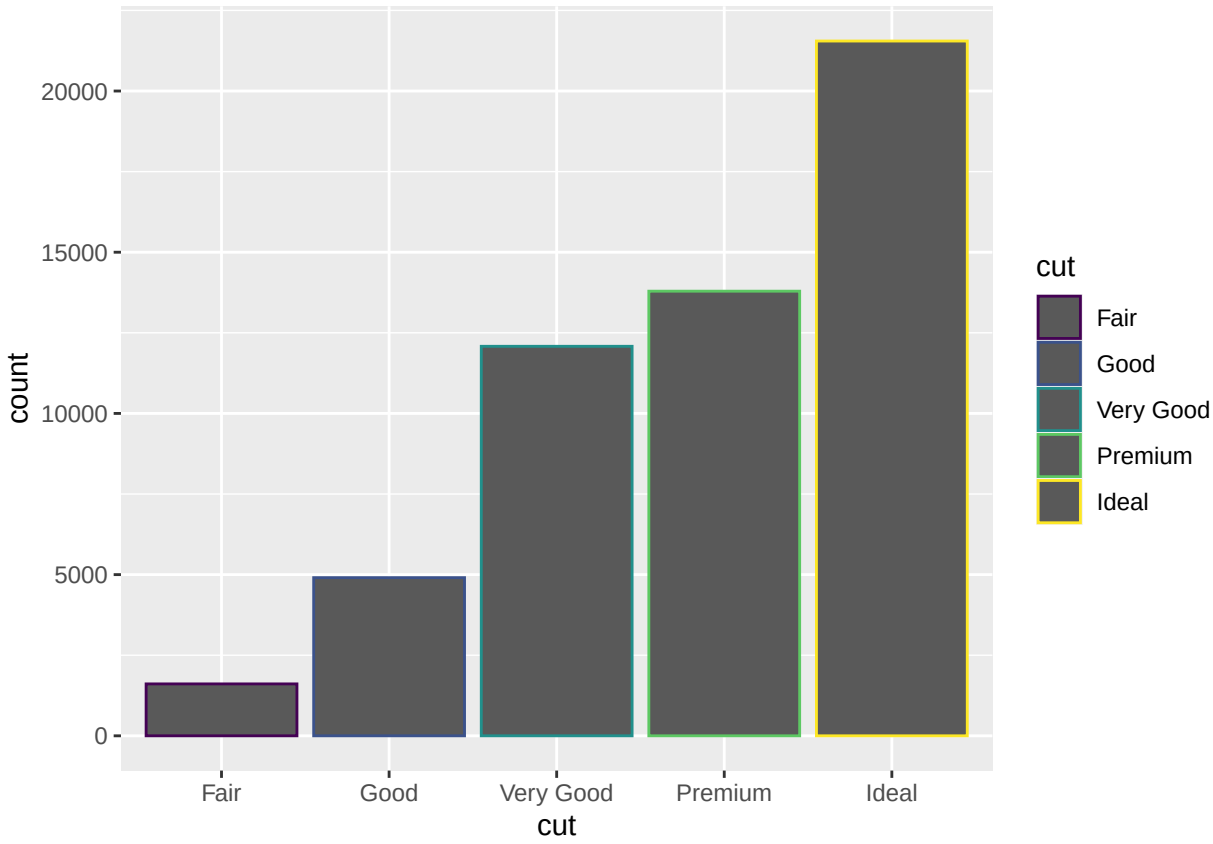

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, y = stat(prop), group = 1)) #bar chart of diamond cuts by proportion
```



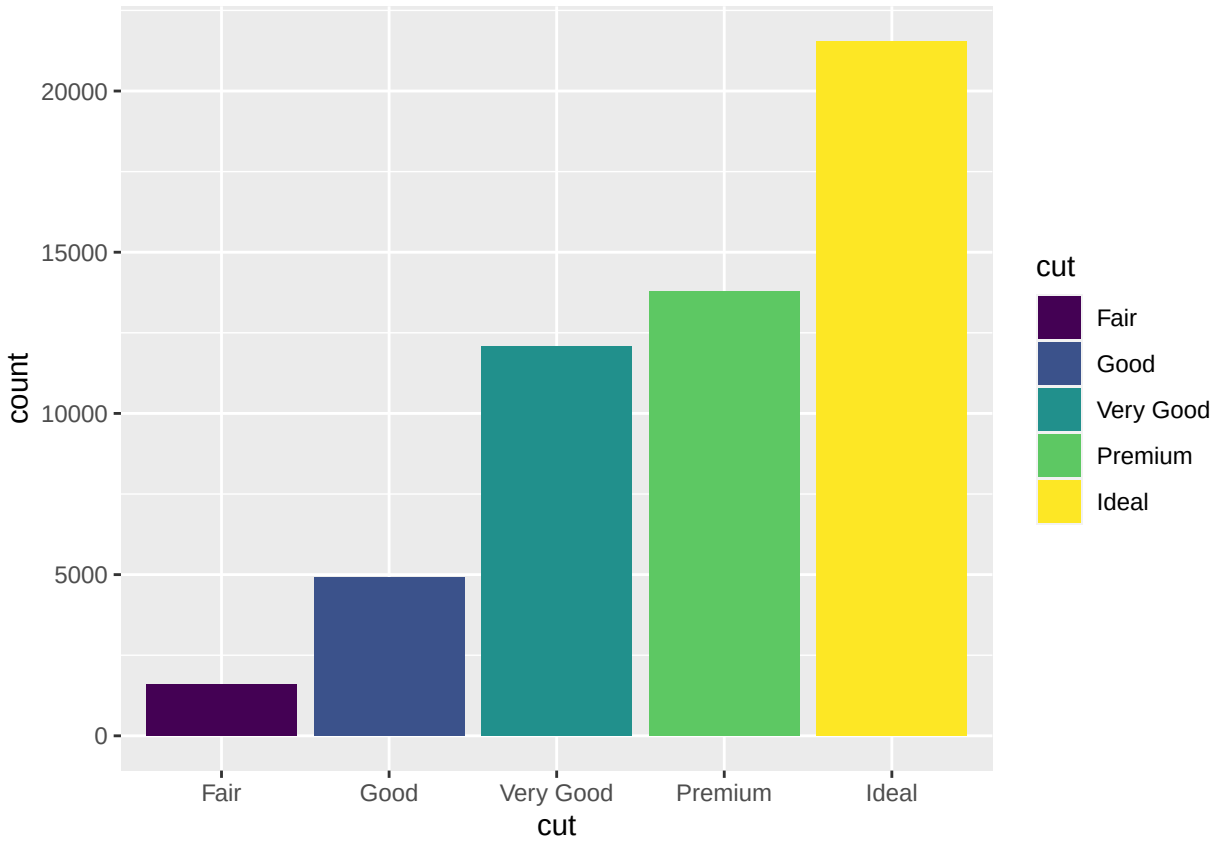

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  stat_summary( #plot summary stats of each diamond cut category
    mapping = aes(x = cut, y = depth), #aesthetics of the plot
    fun.min = min, #min of each category
    fun.max = max,#max of each category
    fun = median#median of each category
  )
```



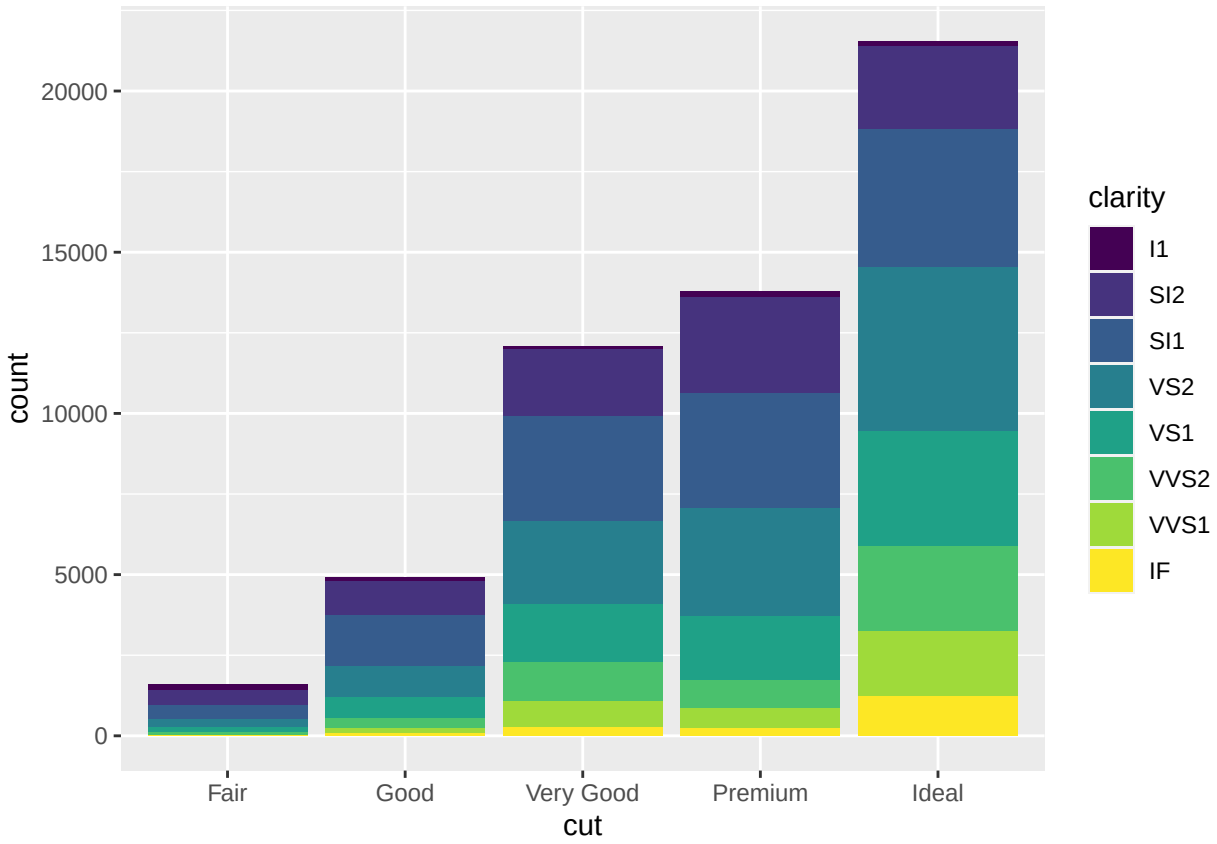

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, colour = cut)) #bar chart of diamond cuts with bars outlined in color
```



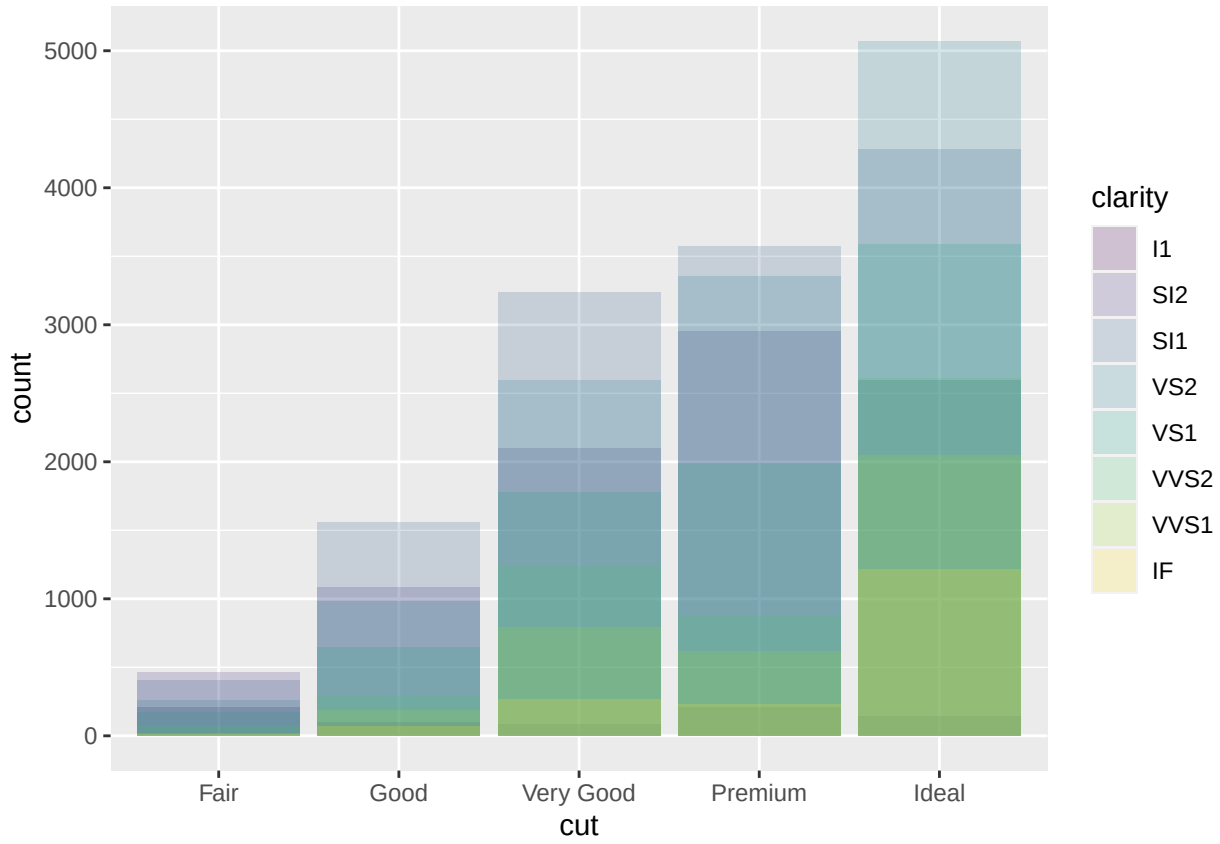

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, fill = cut)) #bar chart of diamond cuts with bars filled with color
```



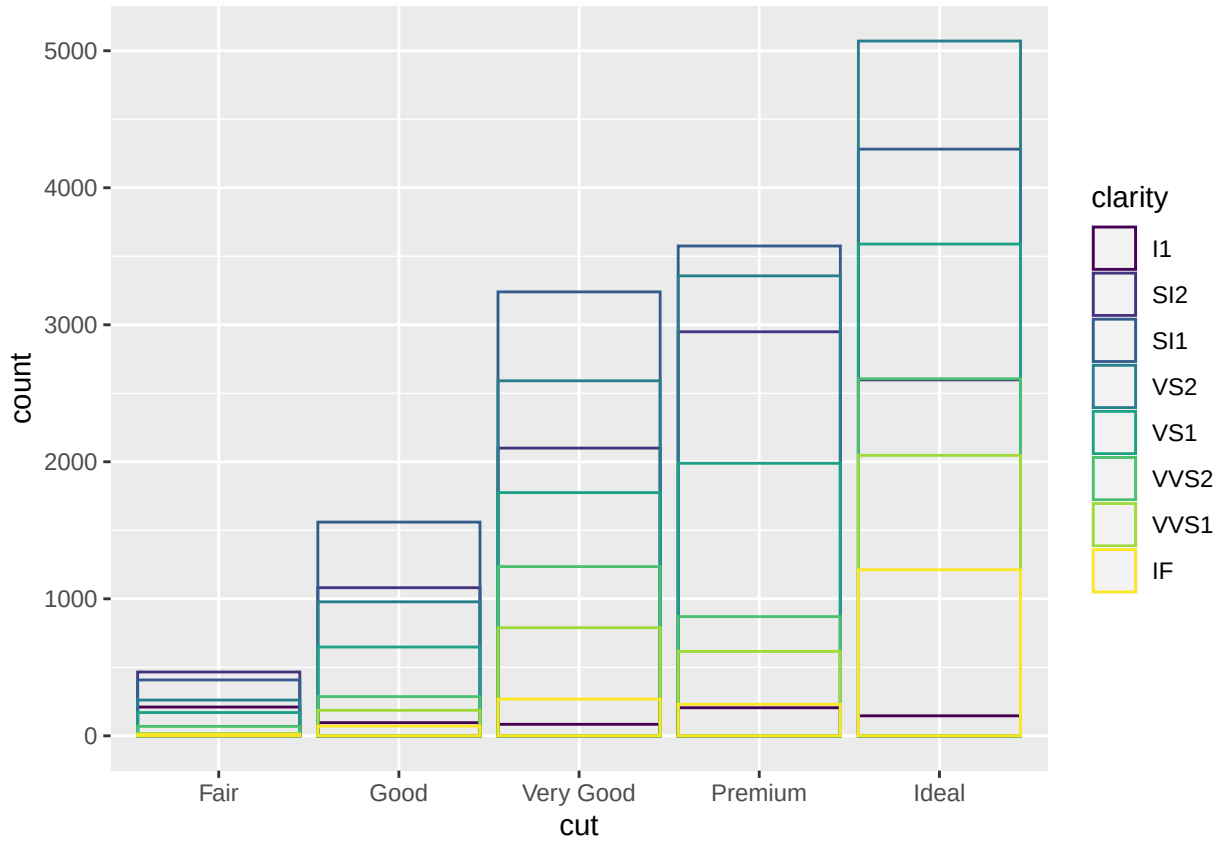

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, fill = clarity)) #bar chart of diamond cuts with bar color filling fo
```



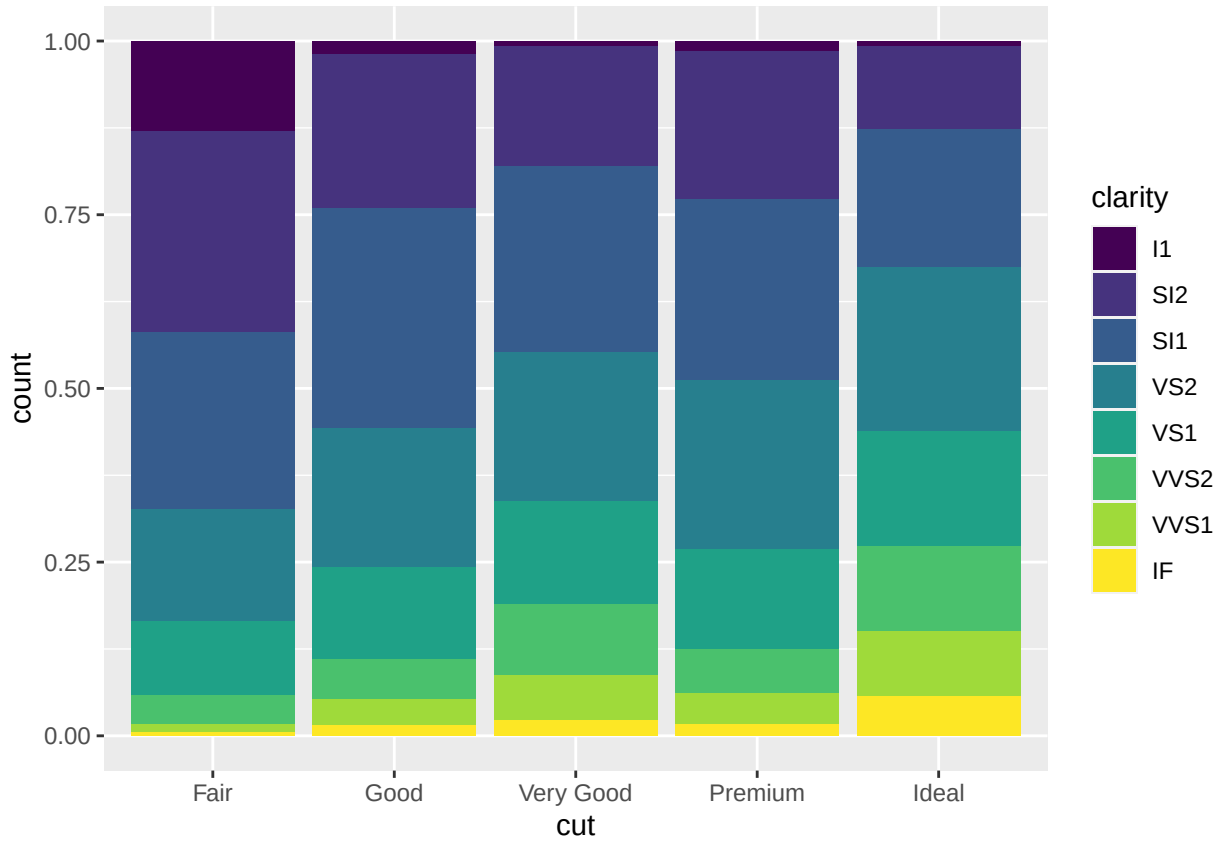

```
ggplot(data = diamonds, mapping = aes(x = cut, fill = clarity)) + #call ggplot function and load data a
  geom_bar(alpha = 1/5, position = "identity") #barchart of cuts and divided into semitransparent clari
```



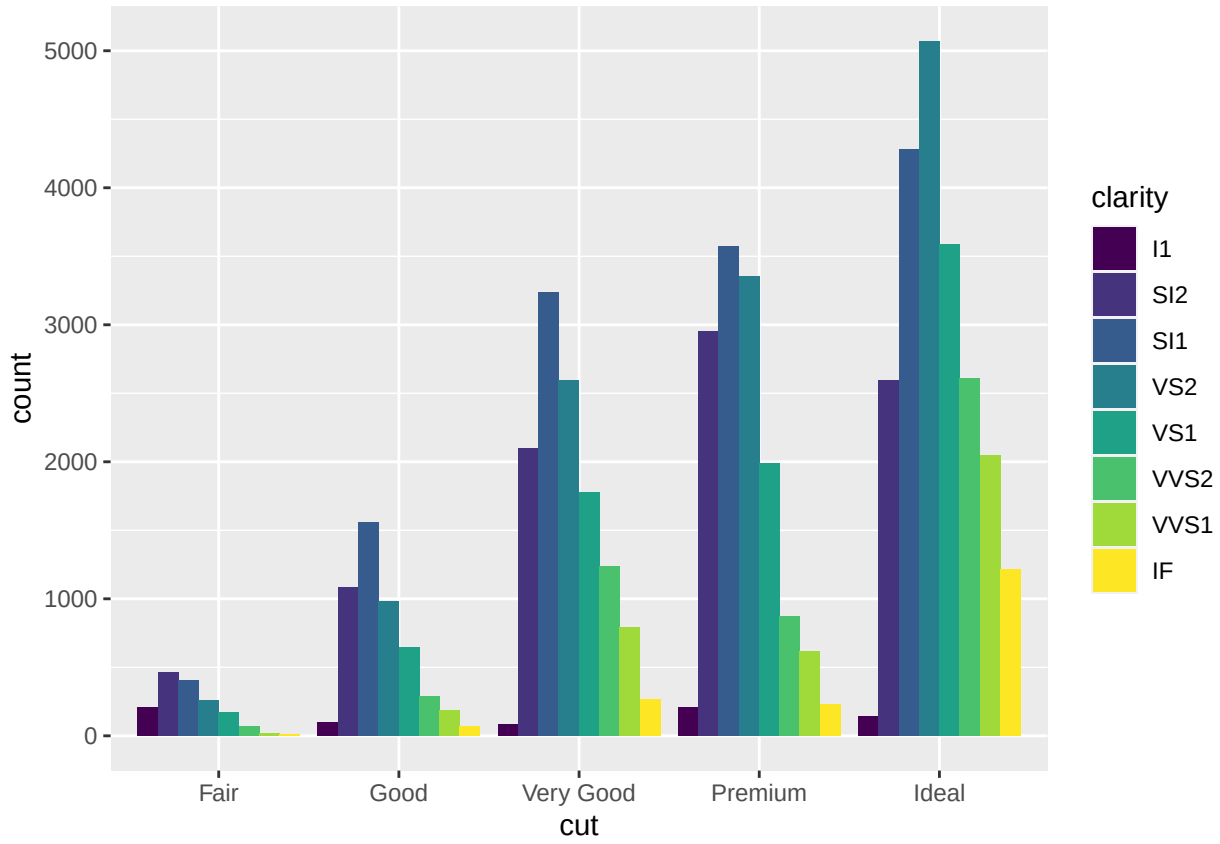

```
ggplot(data = diamonds, mapping = aes(x = cut, colour = clarity)) + #call ggplot function and load data
  geom_bar(fill = NA, position = "identity") #barchart of cuts and divided into transparent clarity cat
```



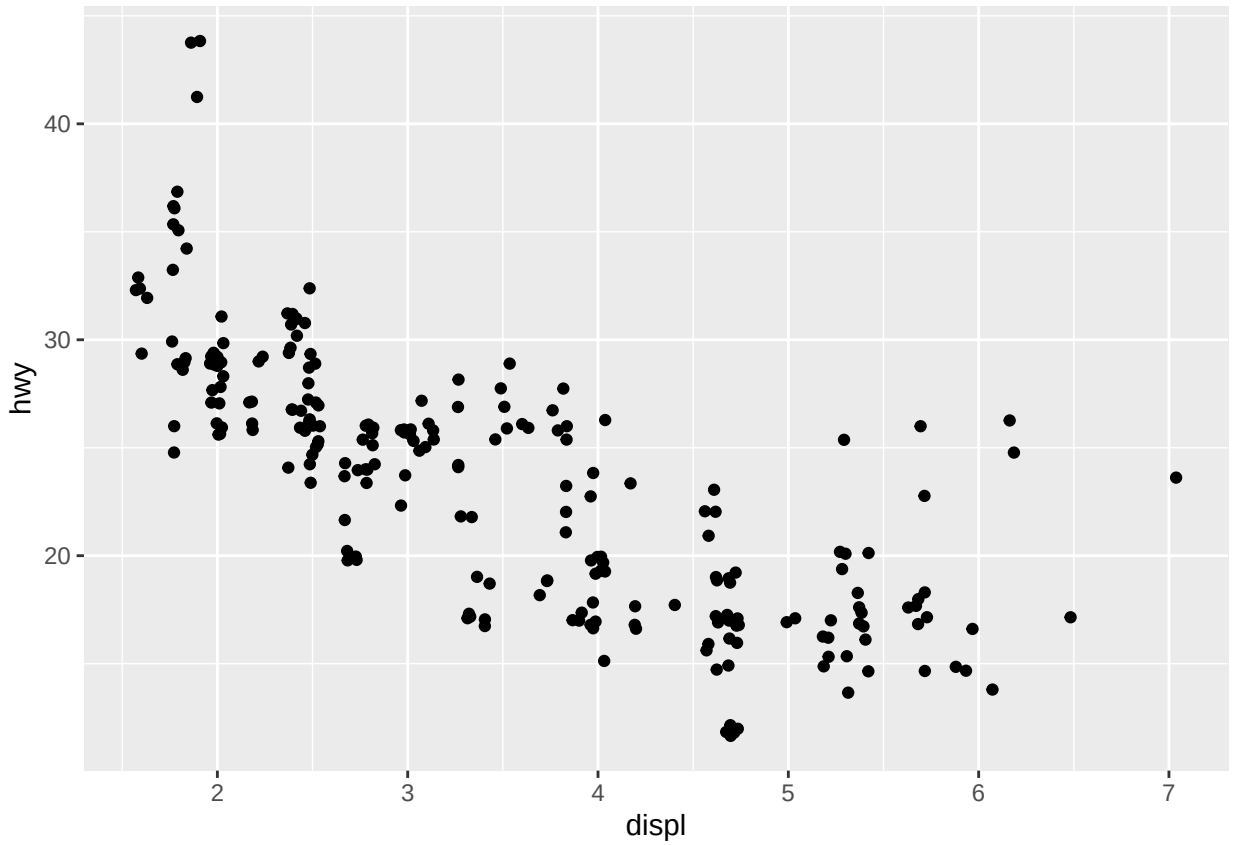

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, fill = clarity), position = "fill") #bar chart of diamond cuts with a
```



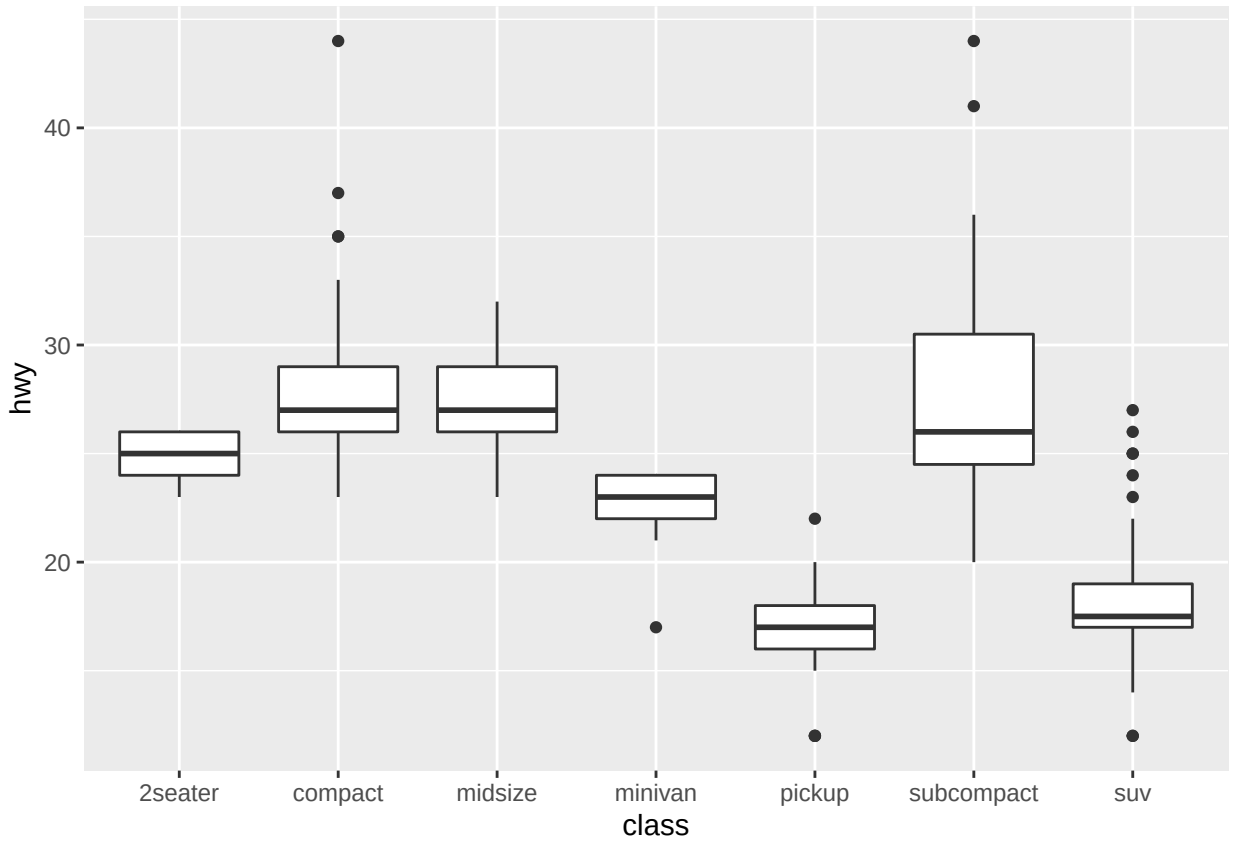

```
ggplot(data = diamonds) + #call ggplot function and load data as argument
  geom_bar(mapping = aes(x = cut, fill = clarity), position = "dodge") #bar chart of diamond cuts with
```



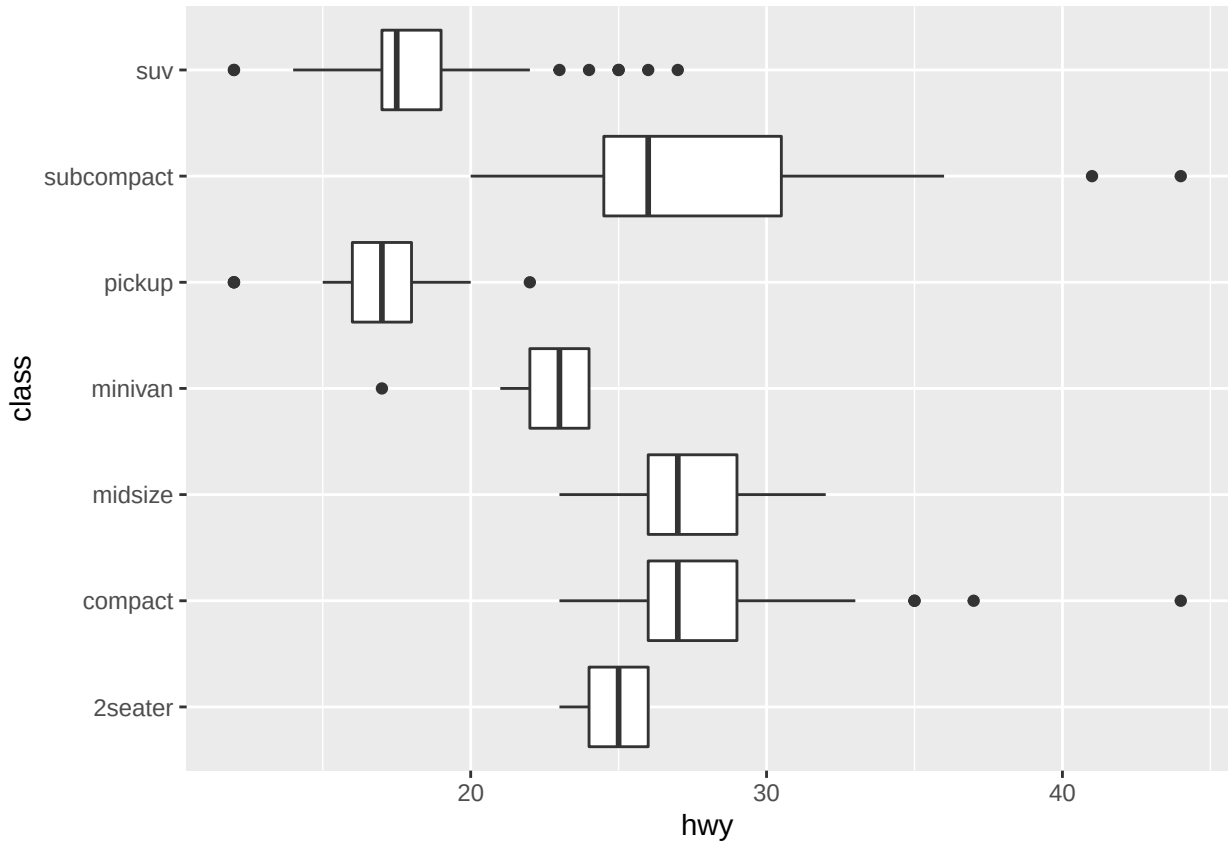

```
ggplot(data = mpg) +  #call ggplot function and load data as argument
  geom_point(mapping = aes(x = displ, y = hwy), position = "jitter") #scatterplot of hwy vs displ with
```



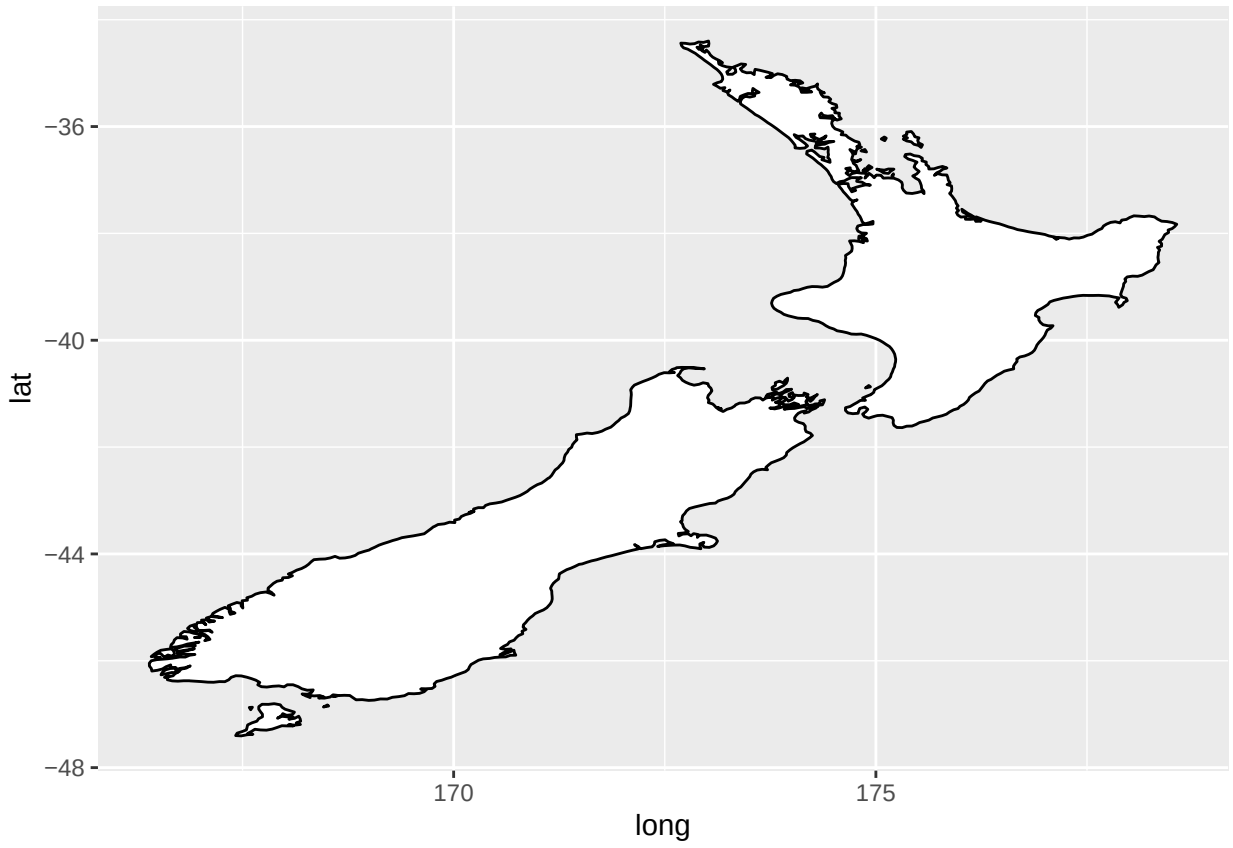

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) + #call ggplot function and load data as argument
  geom_boxplot() #boxplot of class and hwy
```



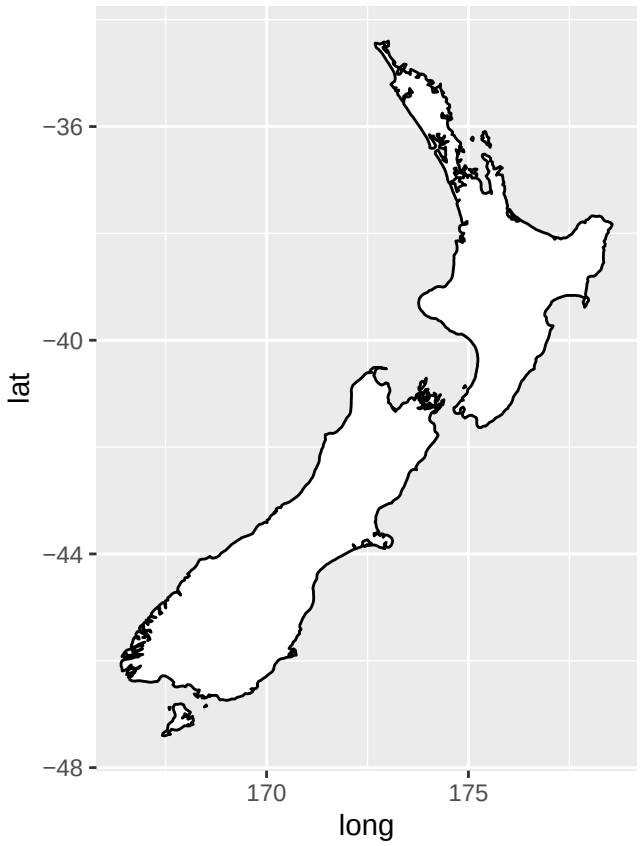

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) + #call ggplot function and load data as argument
  geom_boxplot() + #boxplot of class and hwy
  coord_flip() #flips x and y of plot
```



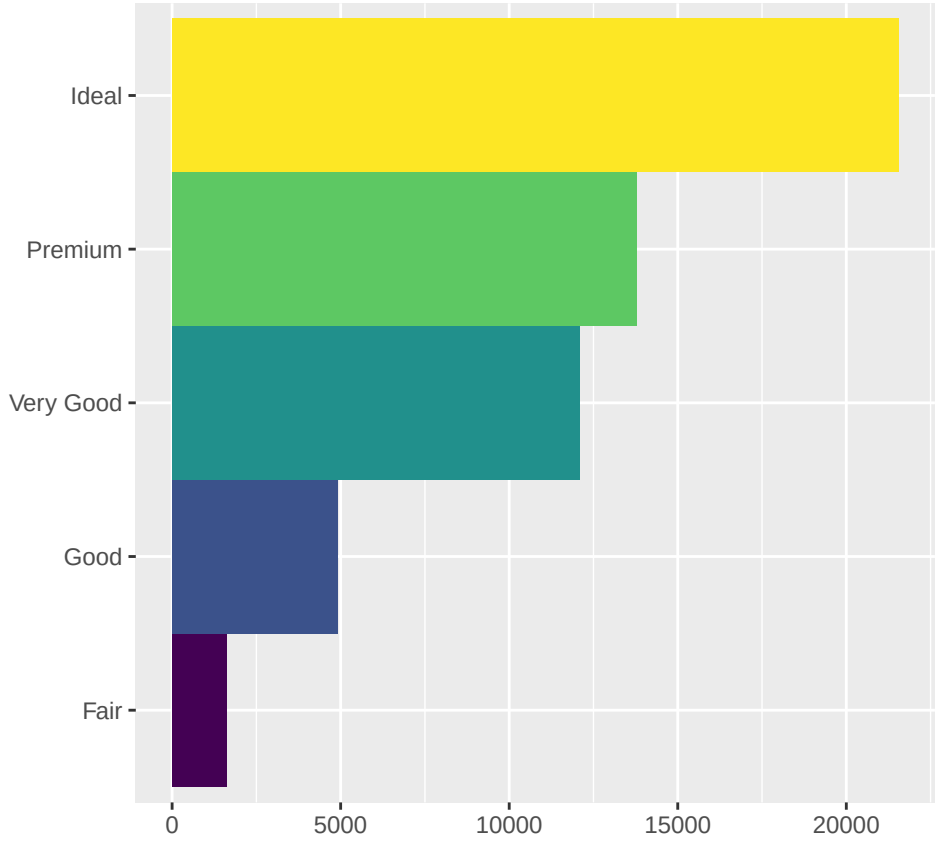

```
nz <- map_data("nz") #create dataframe

ggplot(nz, aes(long, lat, group = group)) + #call ggplot with data and set aesthetics
  geom_polygon(fill = "white", colour = "black") #create map with white fill and black outline
```



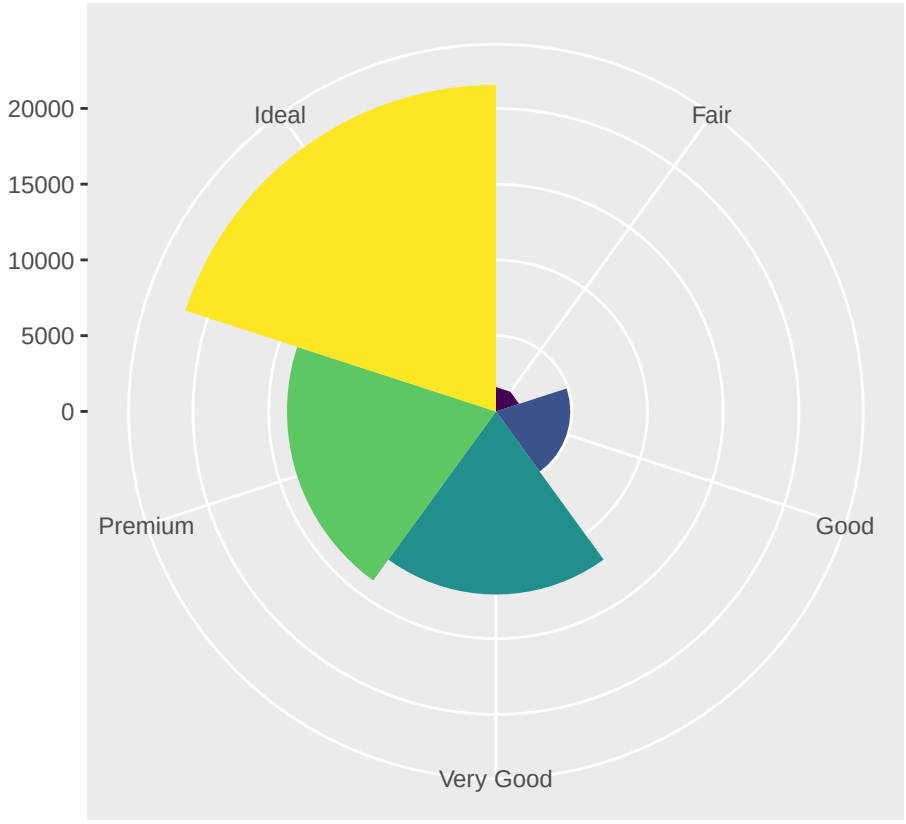

```
ggplot(nz, aes(long, lat, group = group)) + #call ggplot with data and set aesthetics
  geom_polygon(fill = "white", colour = "black") + #create map with white fill and black outline
  coord_quickmap() #sets aspect ratio
```




```
bar <- ggplot(data = diamonds) + #creates bar, calls ggplot and loads data
  geom_bar( #bar chart
    mapping = aes(x = cut, fill = cut), #barchart of cut
    show.legend = FALSE, #don't show legend
    width = 1
  ) +
  theme(aspect.ratio = 1) + #color graph with theme
  labs(x = NULL, y = NULL) #null x and y labels

bar + coord_flip() #flip x and y
```

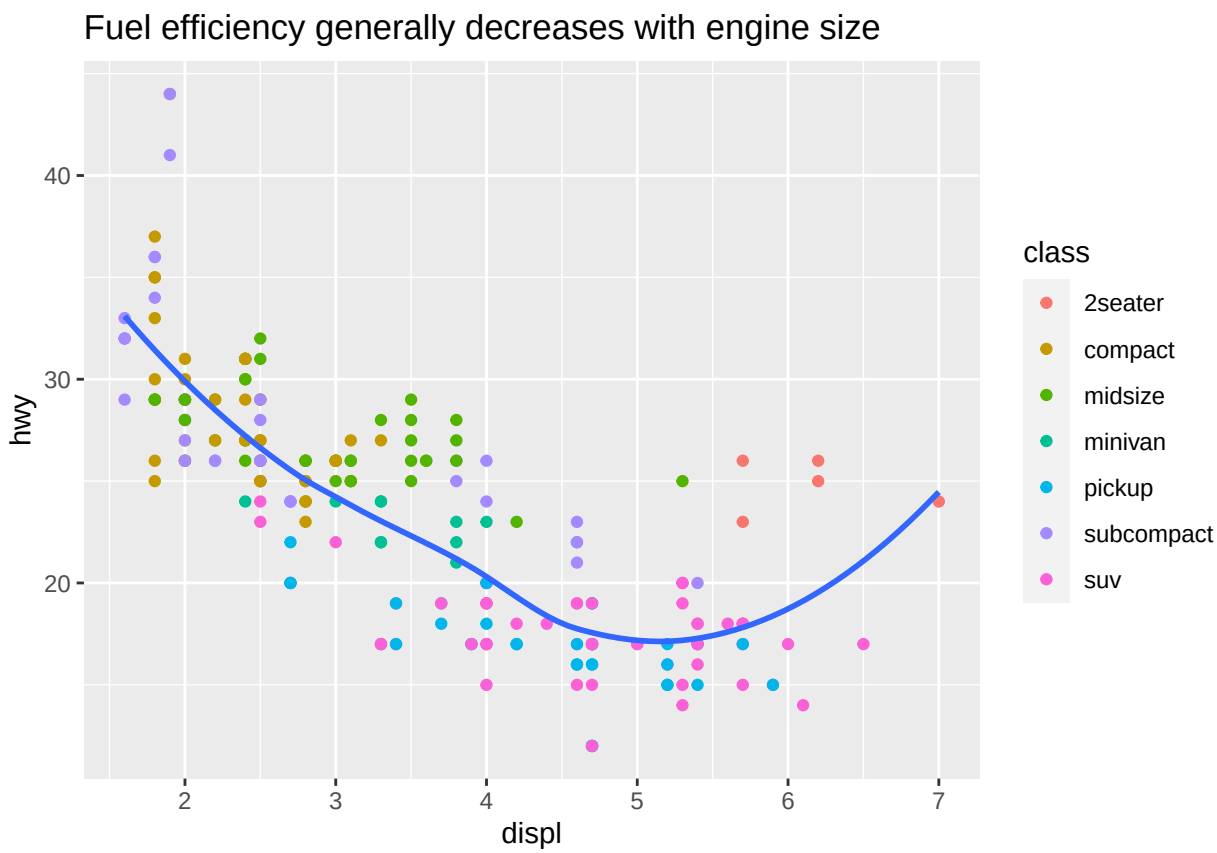



```
bar + coord_polar() #use polar coordinates
```



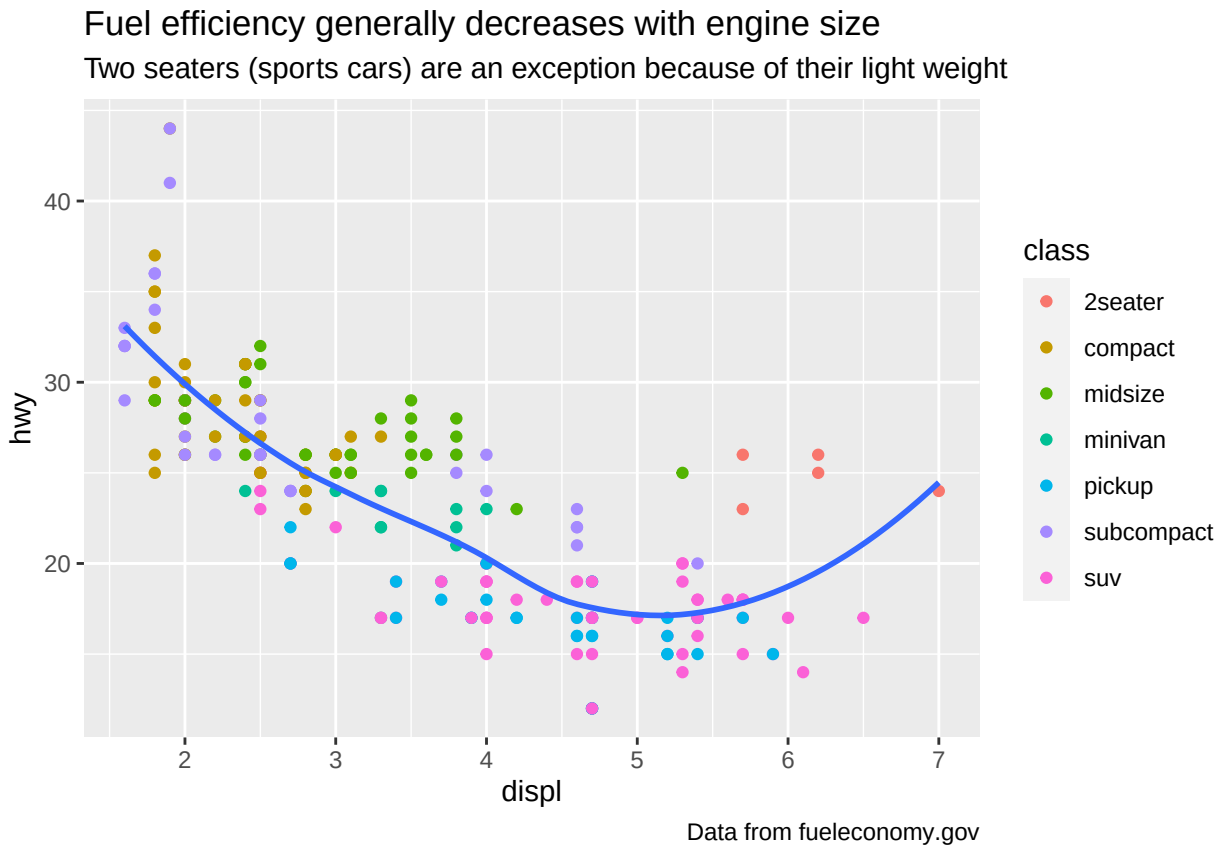

```
ggplot(mpg, aes(displ, hwy)) + #call ggplot, load data, and create aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with color = class
  geom_smooth(se = FALSE) + #line of fit without standard error
  labs(title = "Fuel efficiency generally decreases with engine size") #add title
```

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



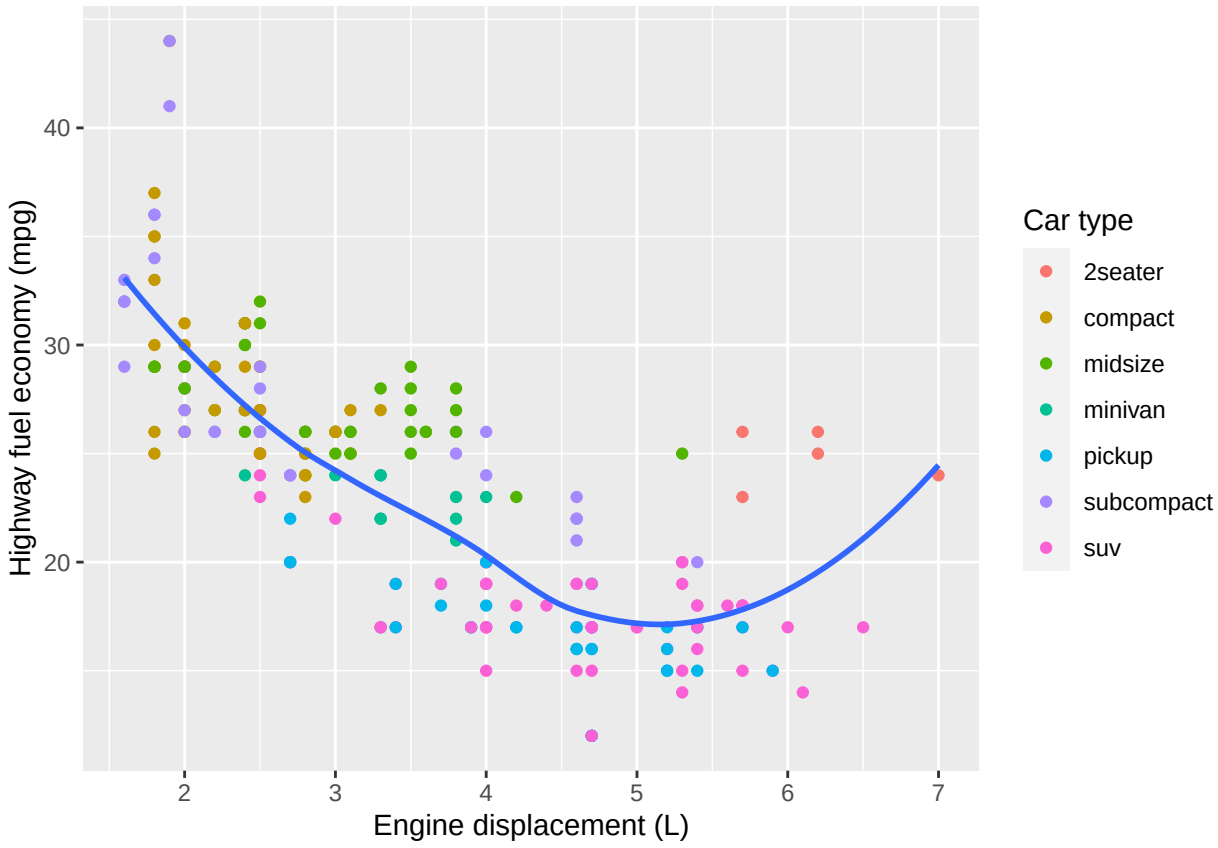

```
ggplot(mpg, aes(displ, hwy)) + #call ggplot, load data, and create aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with color = class
  geom_smooth(se = FALSE) + #line of fit without standard error
  labs(
    title = "Fuel efficiency generally decreases with engine size",
    subtitle = "Two seaters (sports cars) are an exception because of their light weight",
    caption = "Data from fueleconomy.gov"
  ) #add title, subtitle, and caption
```

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

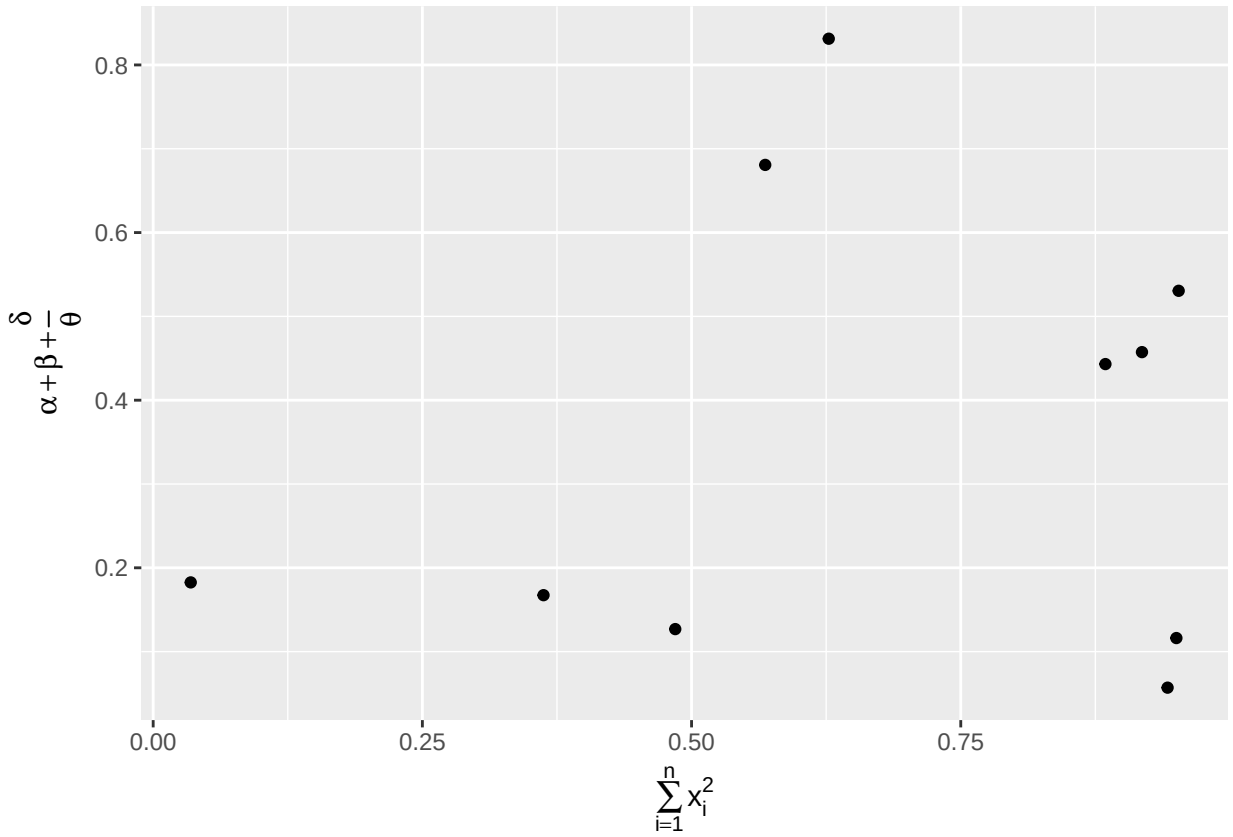



```
ggplot(mpg, aes(displ, hwy)) + #call ggplot, load data, and create aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with color = class
  geom_smooth(se = FALSE) + #line of fit without standard error
  labs(
    x = "Engine displacement (L)",
    y = "Highway fuel economy (mpg)",
    colour = "Car type"
  ) #add x, y, and legend labels
```

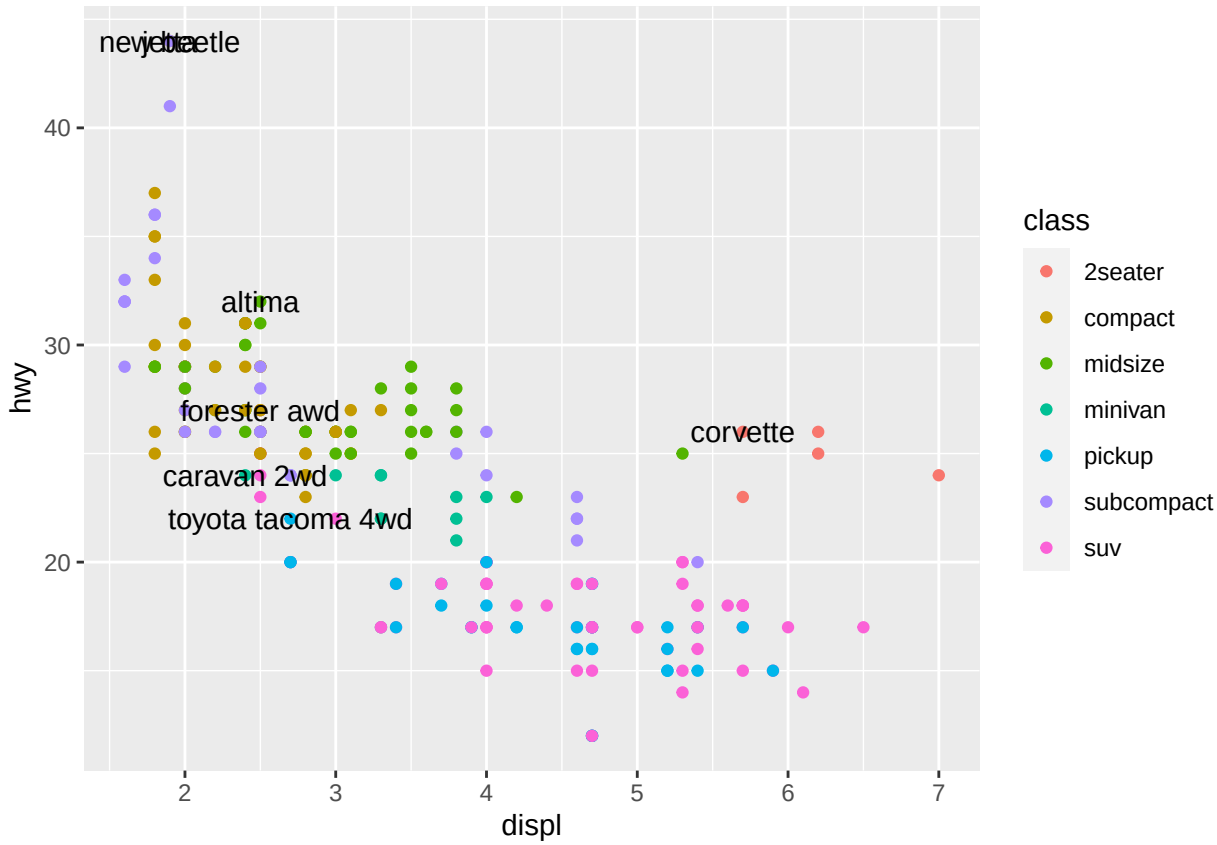
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



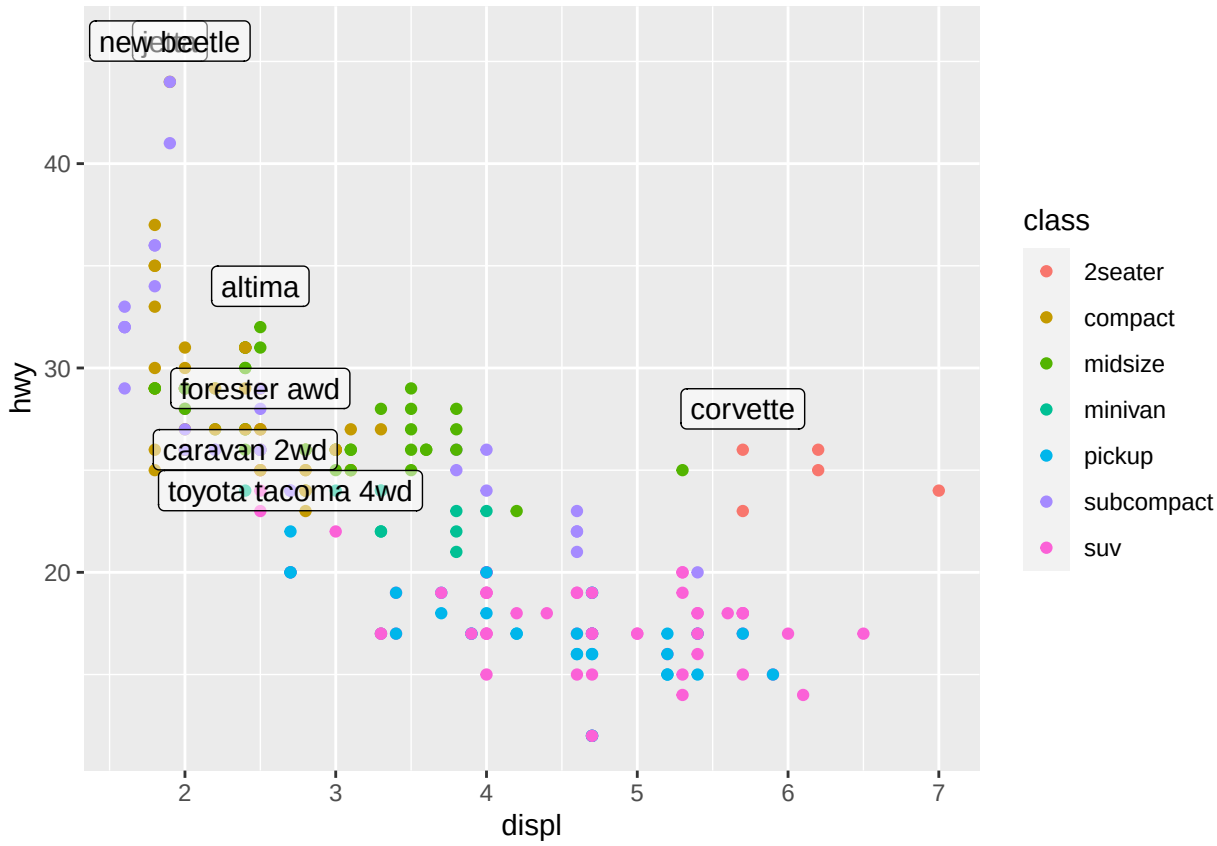

```
df <- tibble(
  x = runif(10),
  y = runif(10)
) #create data
ggplot(df, aes(x, y)) + #ggplot, load data
  geom_point() + #scatterplot of y vs x
  labs(
    x = quote(sum(x[i] ^ 2, i == 1, n)),
    y = quote(alpha + beta + frac(delta, theta))
  ) #create labels for mathmetical expressions for x and y
```



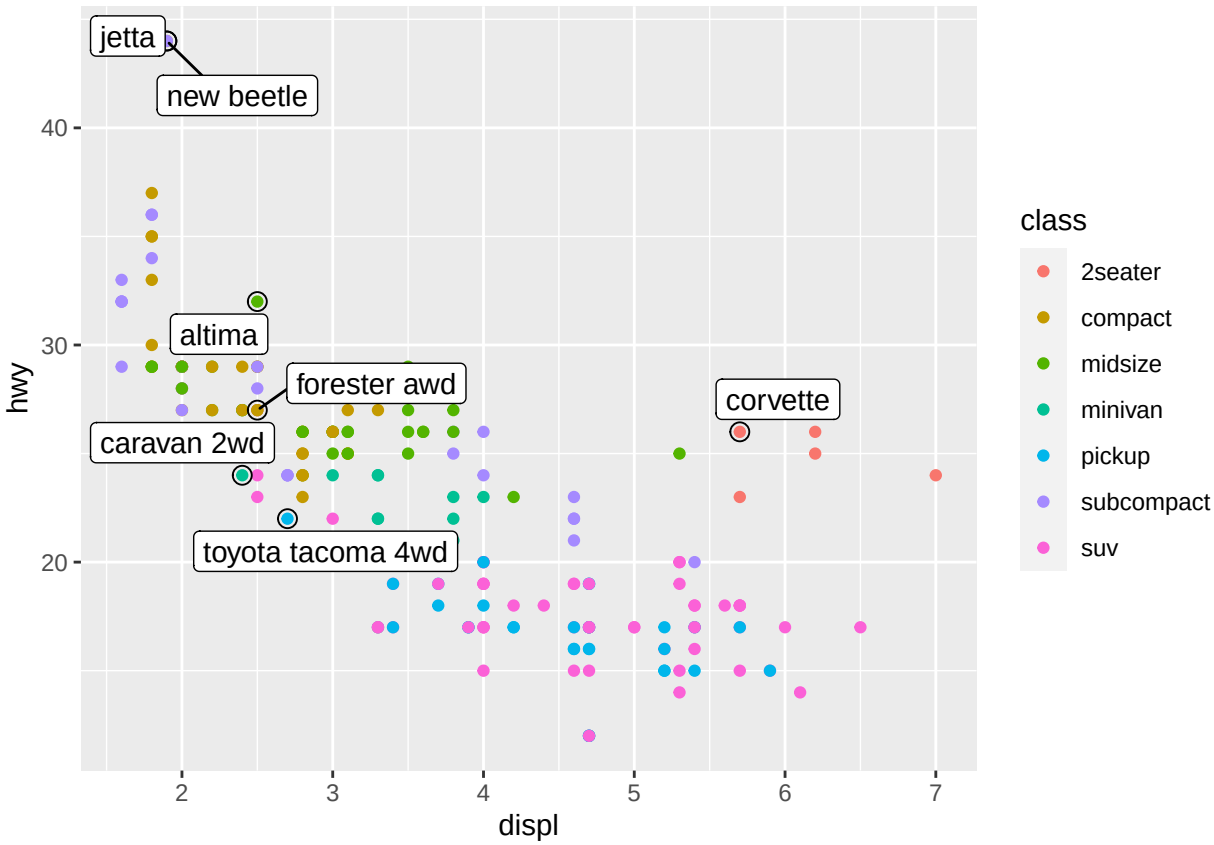

```
best_in_class <- mpg %>% #create best_in_class data
  group_by(class) %>%
  filter(row_number(desc(hwy)) == 1) #filter

ggplot(mpg, aes(displ, hwy)) + #ggplot, load data, and set aesthetics
  geom_point(aes(colour = class)) + #scatterplot of hwy vs displ with color = class
  geom_text(aes(label = model), data = best_in_class) #create labels for points in best_in_class
```



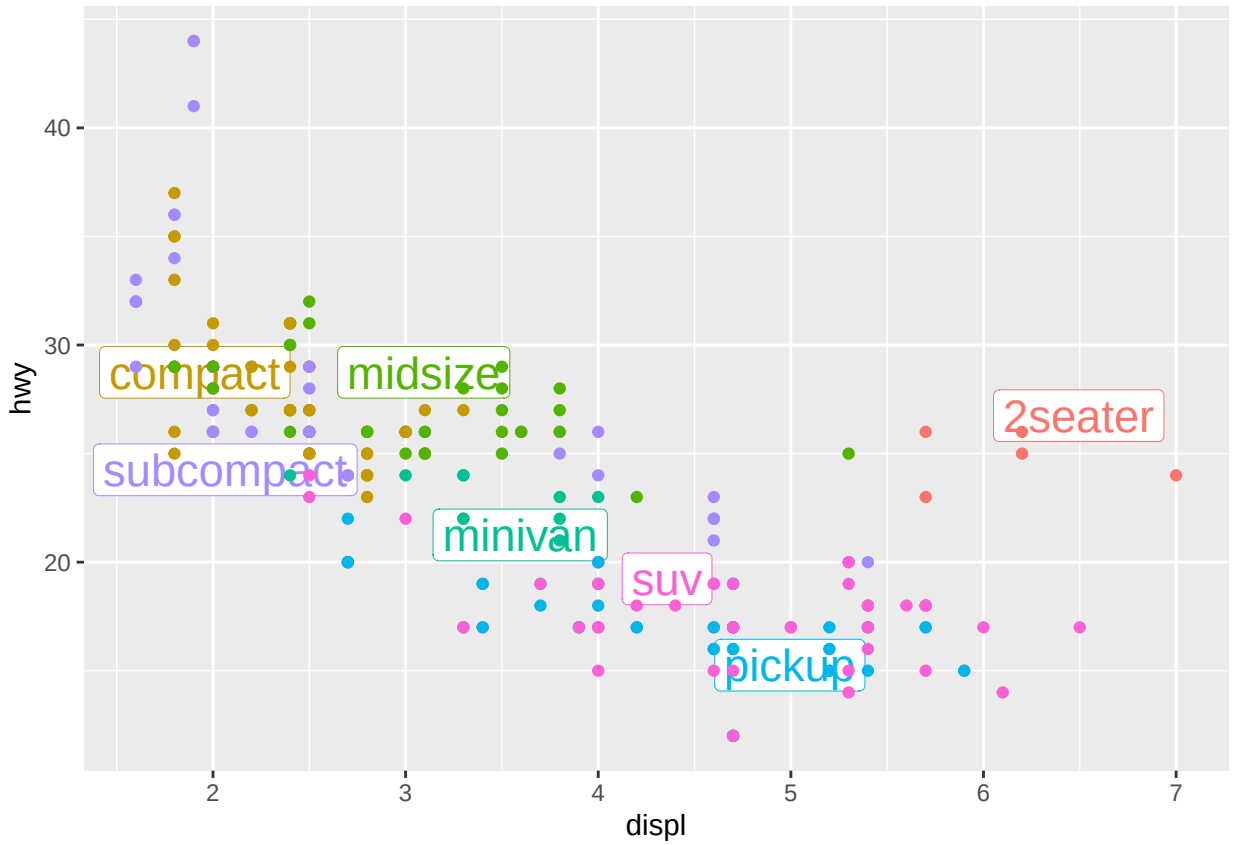

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, load data, and set aesthetics
  geom_point(aes(colour = class)) + #scatterplot of hwy vs displ with color = class
  geom_label(aes(label = model), data = best_in_class, nudge_y = 2, alpha = 0.5) #create labels for poi
```



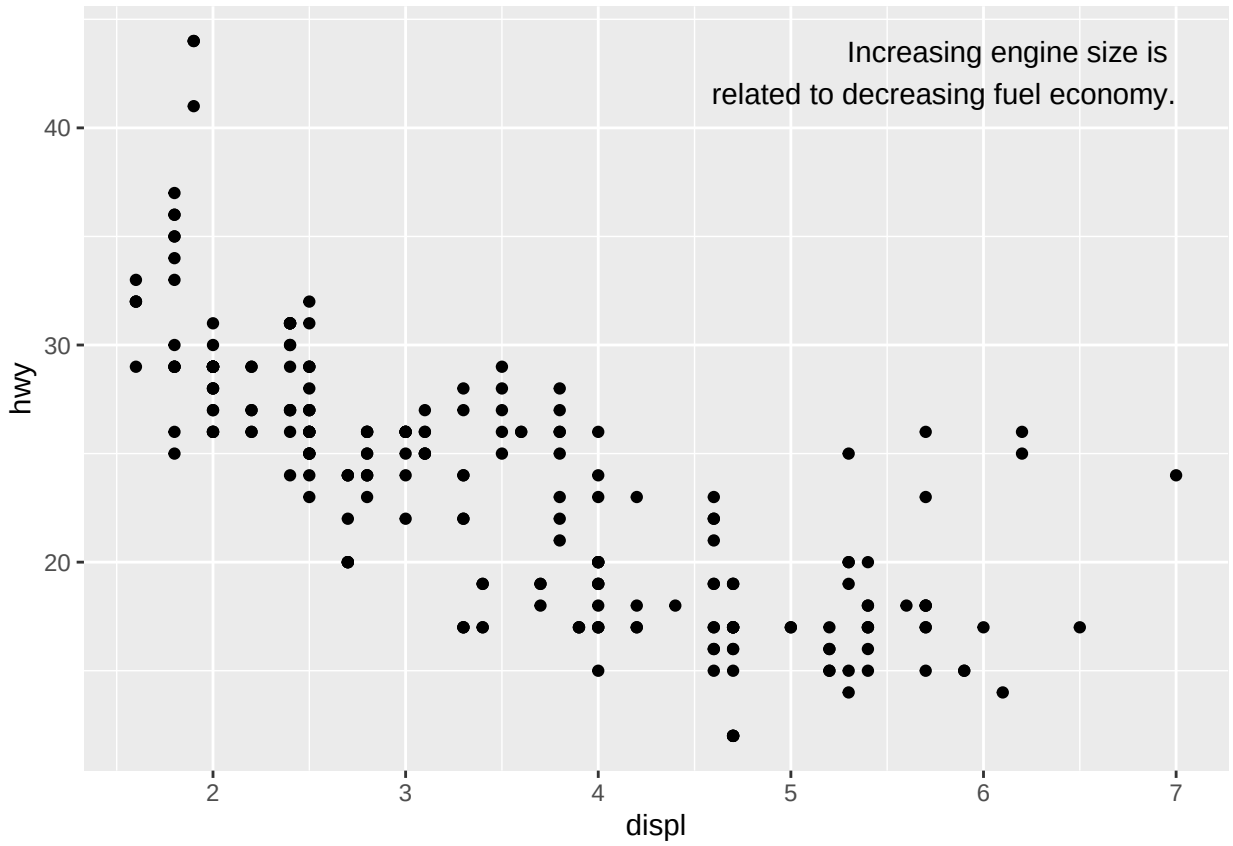

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, load data, and set aesthetics
  geom_point(aes(colour = class)) + #scatterplot of hwy vs displ with color = class
  geom_point(size = 3, shape = 1, data = best_in_class) + #adjust size and data
  ggrepel::geom_label_repel(aes(label = model), data = best_in_class) #create labels for points in best
```



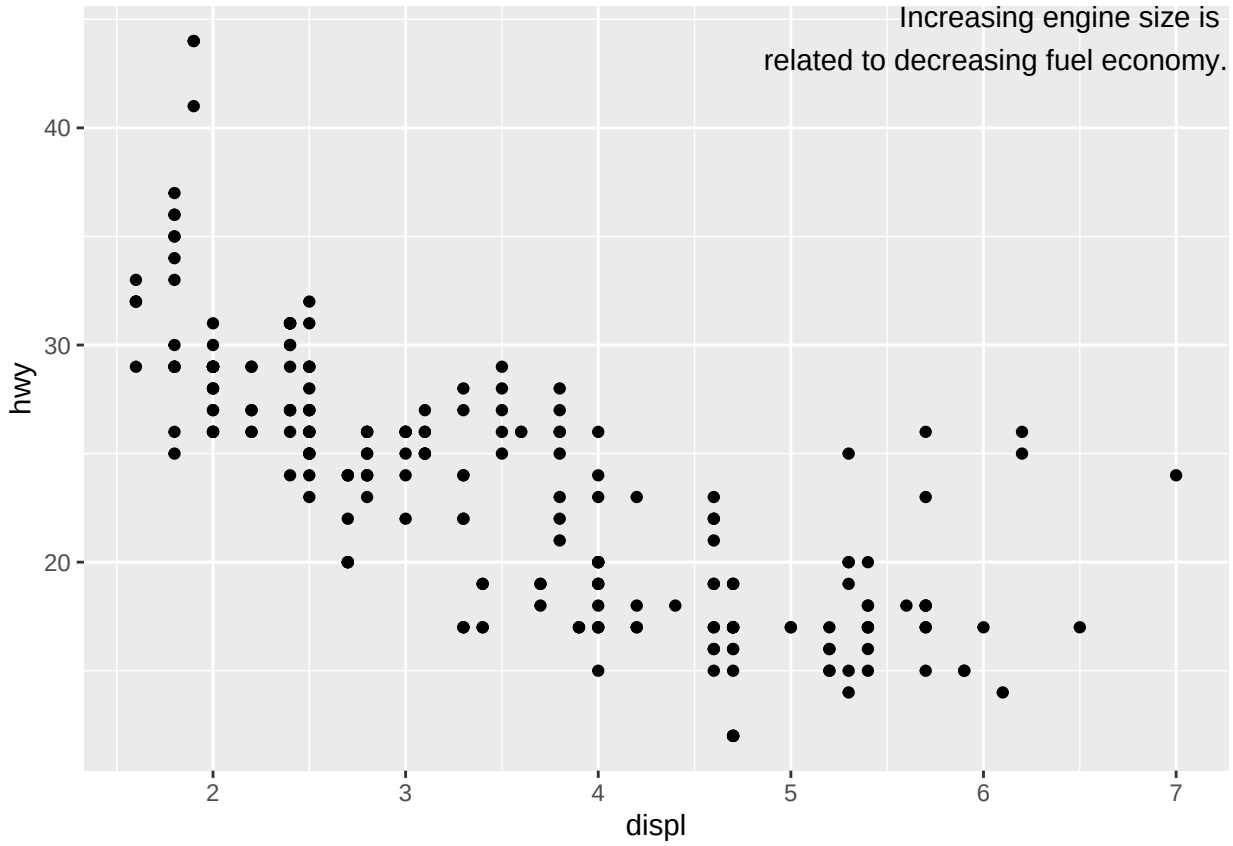

```
class_avg <- mpg %>% #create class_avg dataframe
  group_by(class) %>%
  summarise(
    displ = median(displ),
    hwy = median(hwy)
  )
#> `summarise()` ungrouping output (override with `.groups` argument)

ggplot(mpg, aes(displ, hwy, colour = class)) + #ggplot, data load, and set aesthetics
  ggrepel::geom_label_repel(aes(label = class), #create labels
    data = class_avg,
    size = 6,
    label.size = 0,
    segment.color = NA
  ) +
  geom_point() + #scatterplot of hwy vs displ
  theme(legend.position = "none") #color with theme
```




```
label <- mpg %>% #create label dataframe
  summarise(
    displ = max(displ),
    hwy = max(hwy),
    label = "Increasing engine size is \nrelated to decreasing fuel economy."
  )

ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point() + #scatterplot of hwy vs displ
  geom_text(aes(label = label), data = label, vjust = "top", hjust = "right") #create text label
```




```
label <- tibble( #create label dataframe
  displ = Inf,
  hwy = Inf,
  label = "Increasing engine size is \nrelated to decreasing fuel economy."
)

ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point() + #scatterplot of hwy vs displ
  geom_text(aes(label = label), data = label, vjust = "top", hjust = "right") #create text labels on th
```

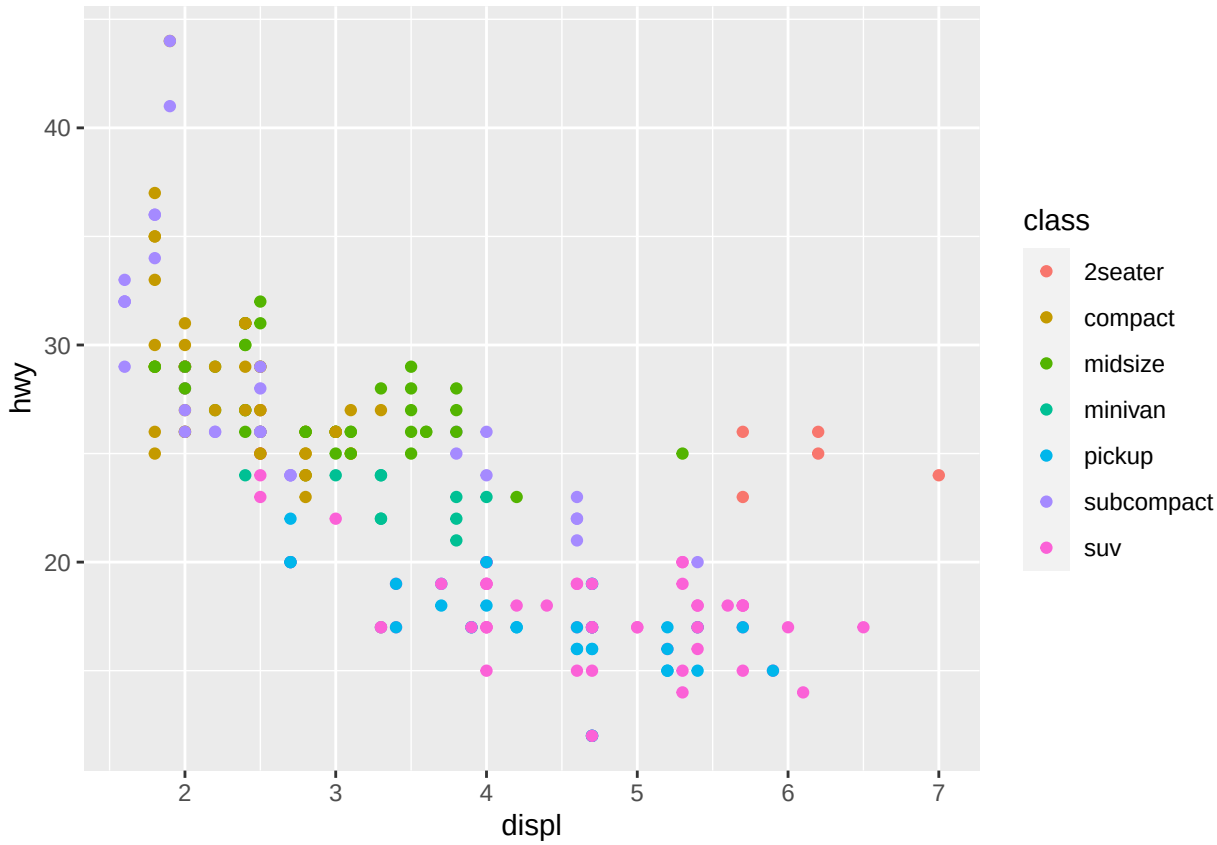



```
#add line breaks given the number of characters chosen per line
"Increasing engine size is related to decreasing fuel economy." %>%
  stringr::str_wrap(width = 40) %>%
  writeLines()
```

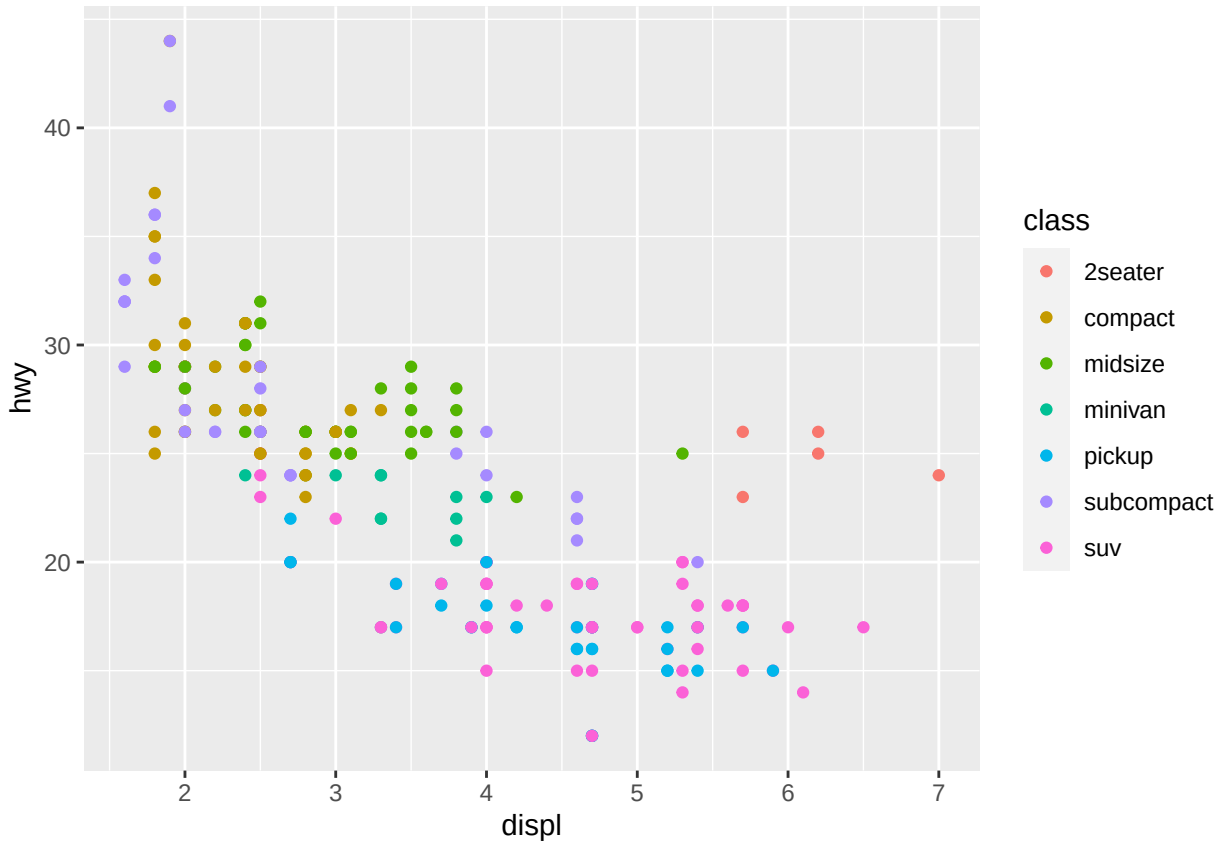
```
## Increasing engine size is related to
## decreasing fuel economy.
```

```
#> Increasing engine size is related to
#> decreasing fuel economy.
```

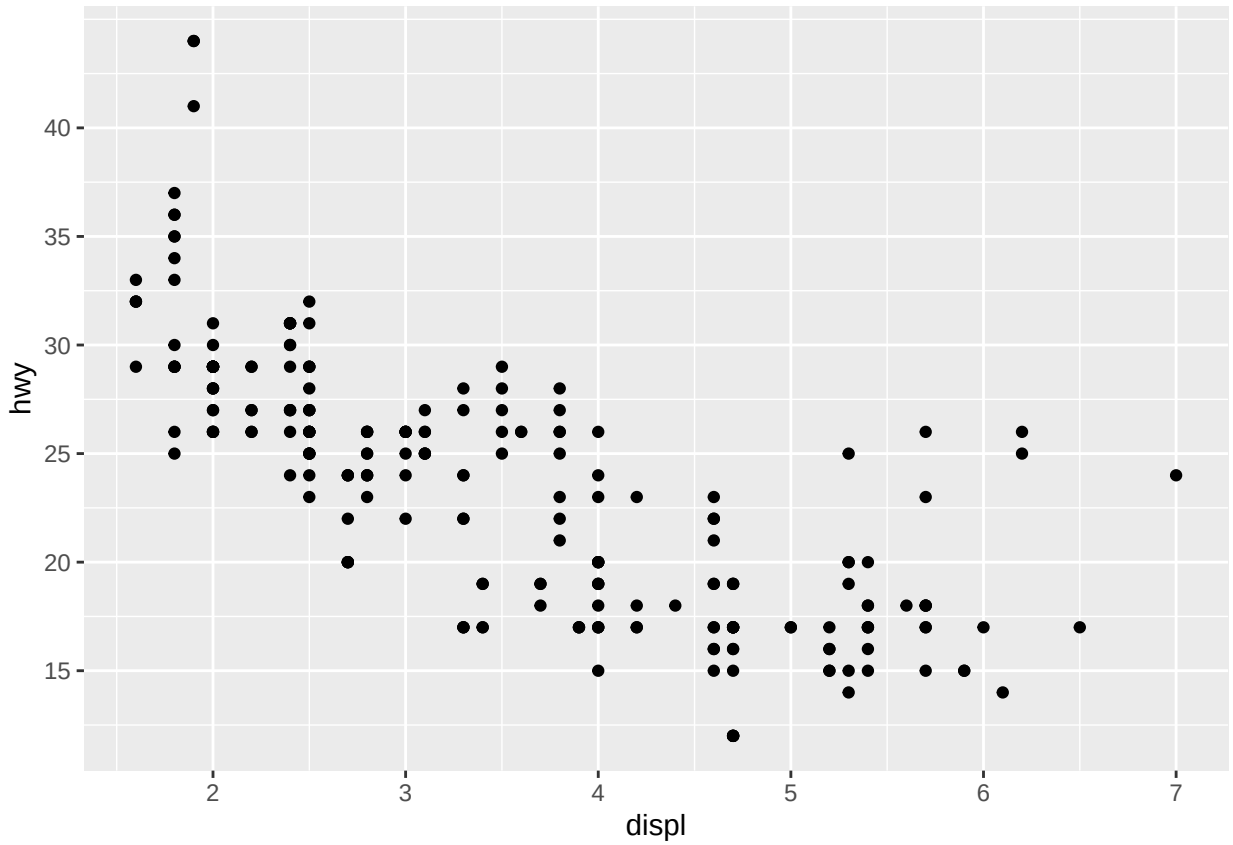
```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point(aes(colour = class)) #scatterplot of hwy vs displ with class color differentiation
```



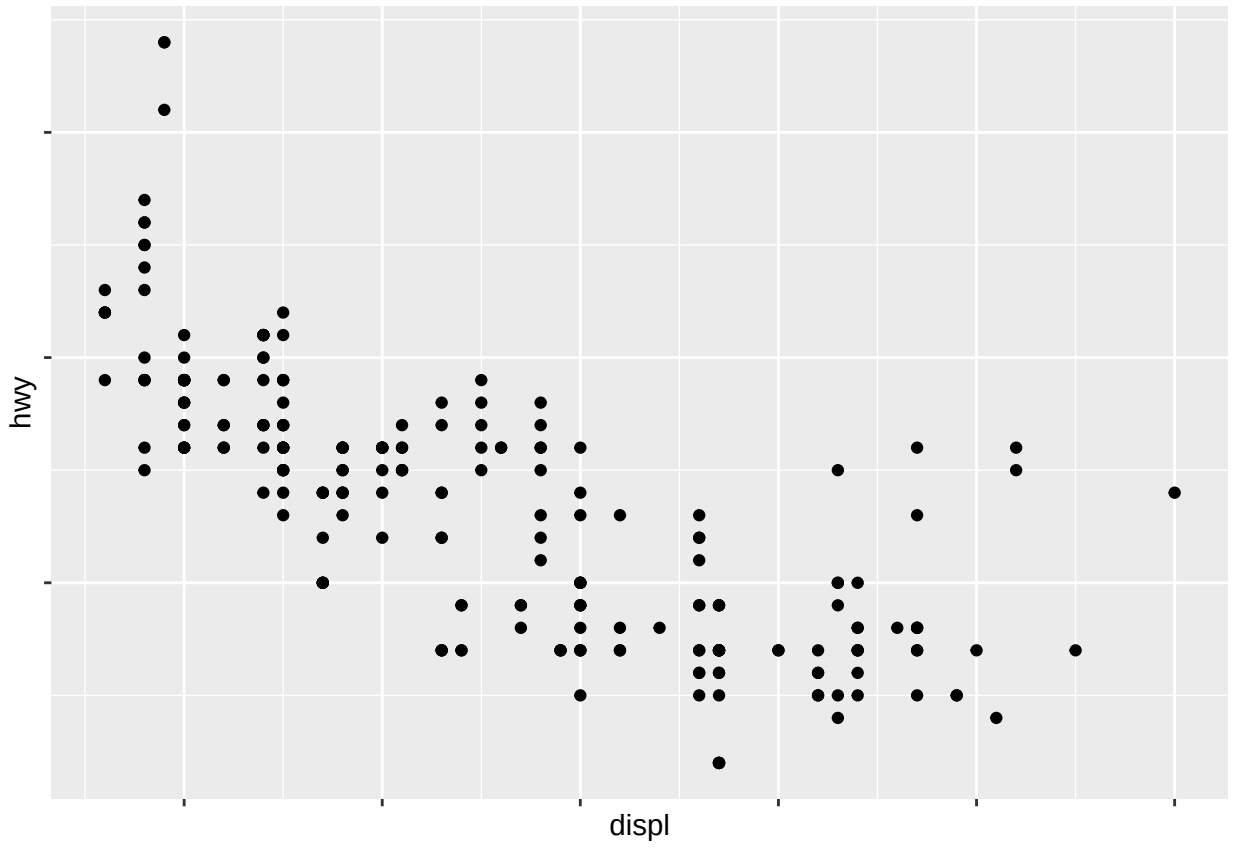

```
#default scales of the above plot (behind the scenes)
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(colour = class)) +
  scale_x_continuous() +
  scale_y_continuous() +
  scale_colour_discrete()
```



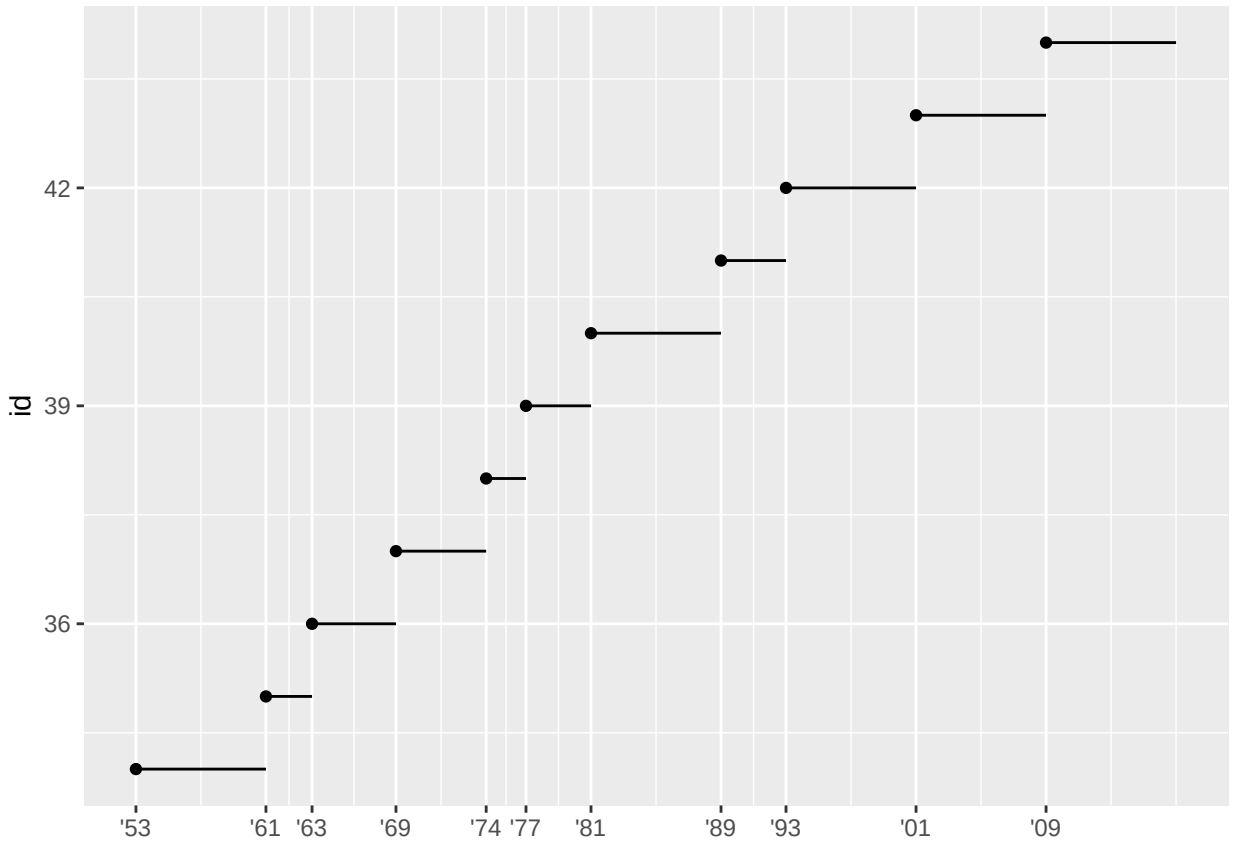

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point() + #scatterplot of hwy vs displ
  scale_y_continuous(breaks = seq(15, 40, by = 5)) #controls position of ticks in plot
```



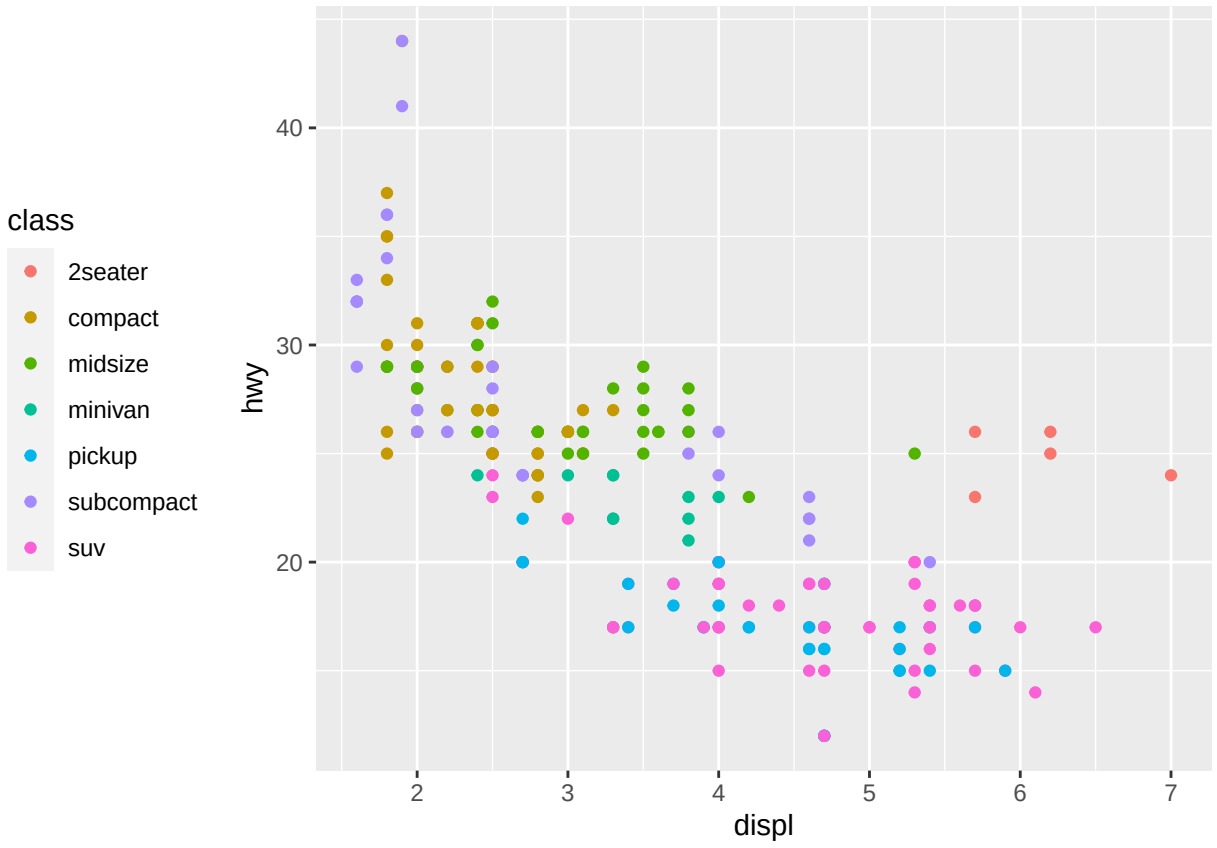

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point() + #scatterplot of hwy vs displ
  scale_x_continuous(labels = NULL) + #no labels for ticks on x axis
  scale_y_continuous(labels = NULL) #no labels for ticks on y axis
```



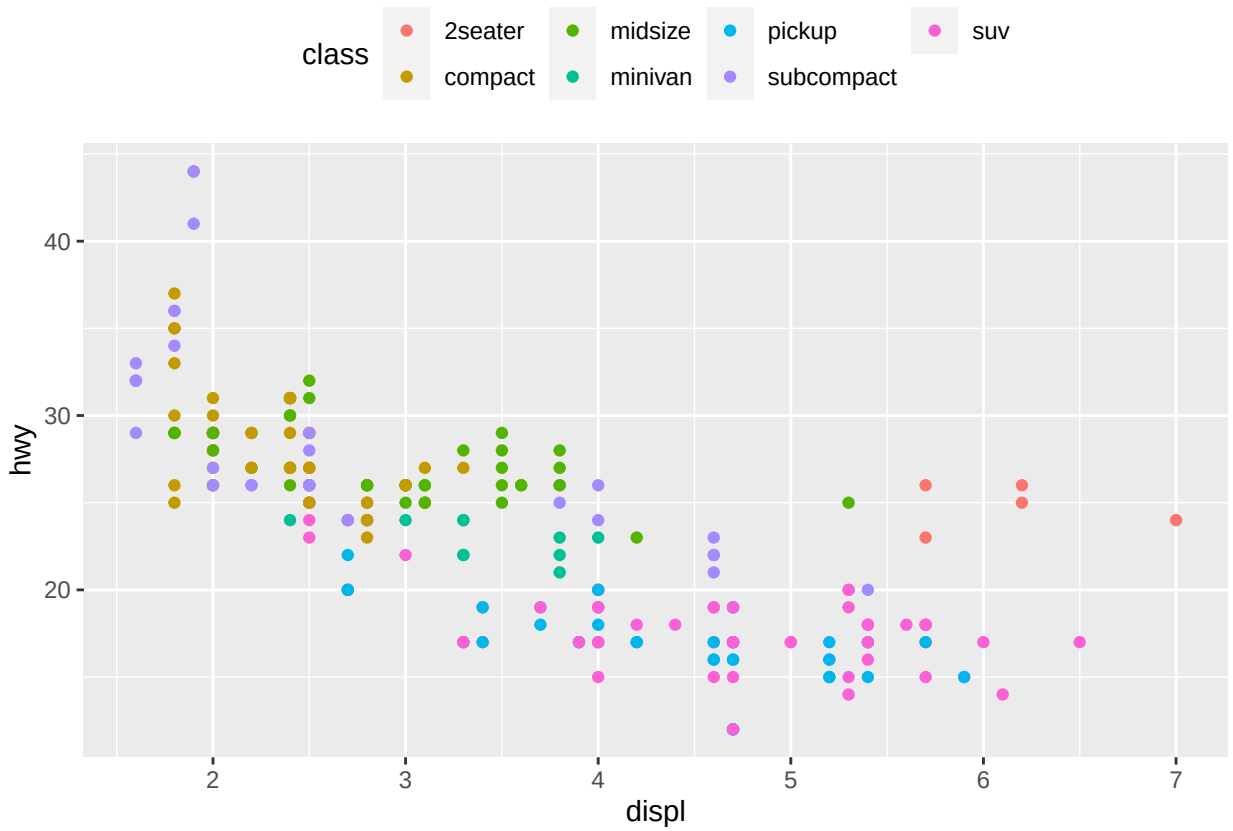

```
presidential %>%
  mutate(id = 33 + row_number()) %>%
  ggplot(aes(start, id)) +
    geom_point() + #scatterplot of id vs start
    geom_segment(aes(xend = end, yend = id)) + #scale of plot
    scale_x_date(NULL, breaks = presidential$start, date_labels = "'%y") #controls ticks and labels of
```



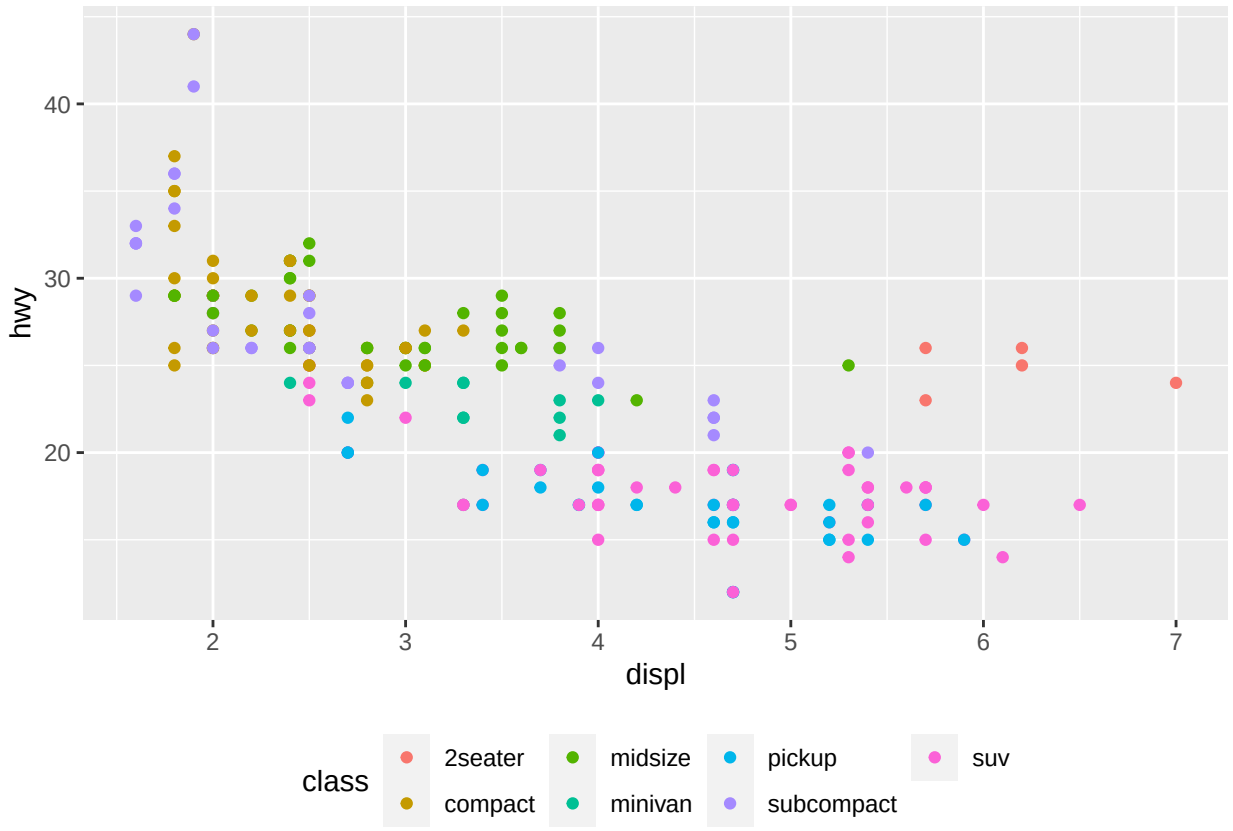

```
base <- ggplot(mpg, aes(displ, hwy)) +  #ggplot, data load, and set aesthetics
  geom_point(aes(colour = class)) #scatterplot of hwy vs displ with class differentiation

base + theme(legend.position = "left") #legend on left
```



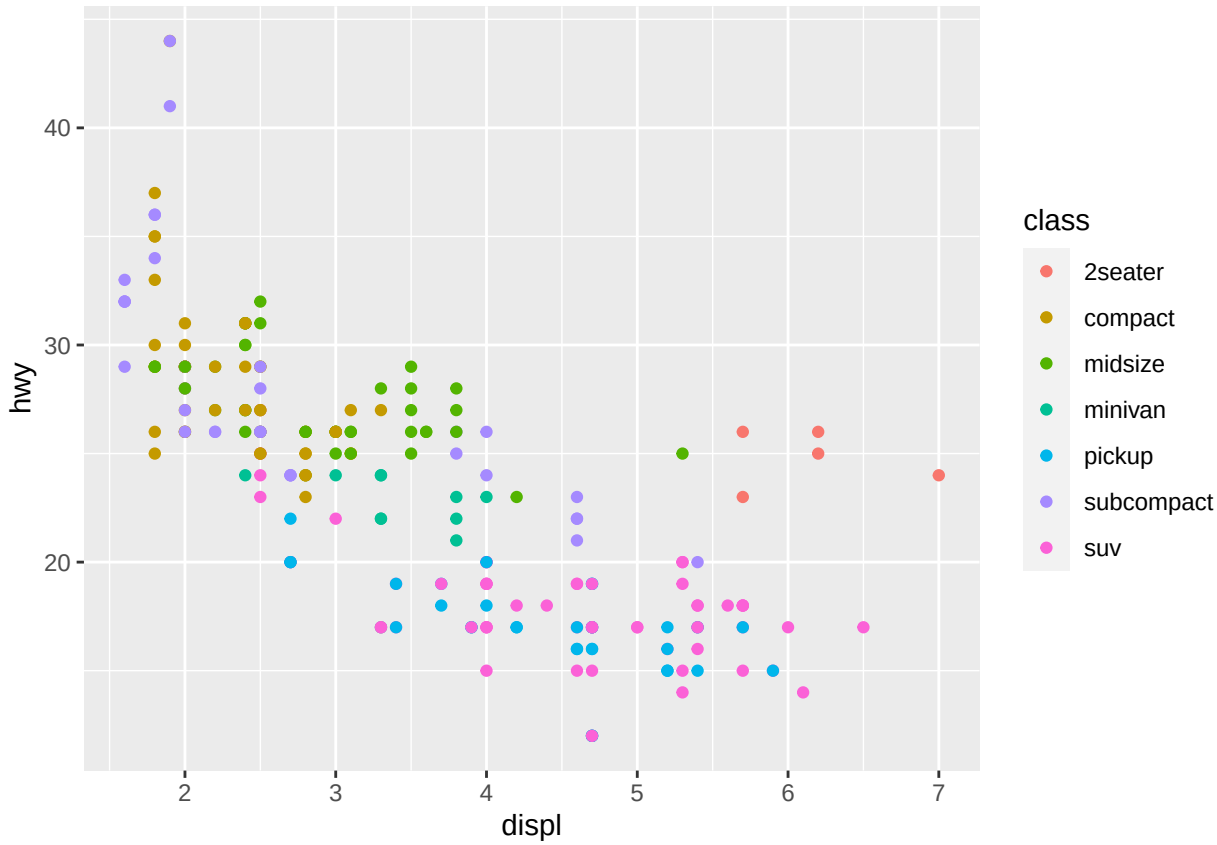

```
base + theme(legend.position = "top") #legend on top
```




```
base + theme(legend.position = "bottom") #legend on bottom
```

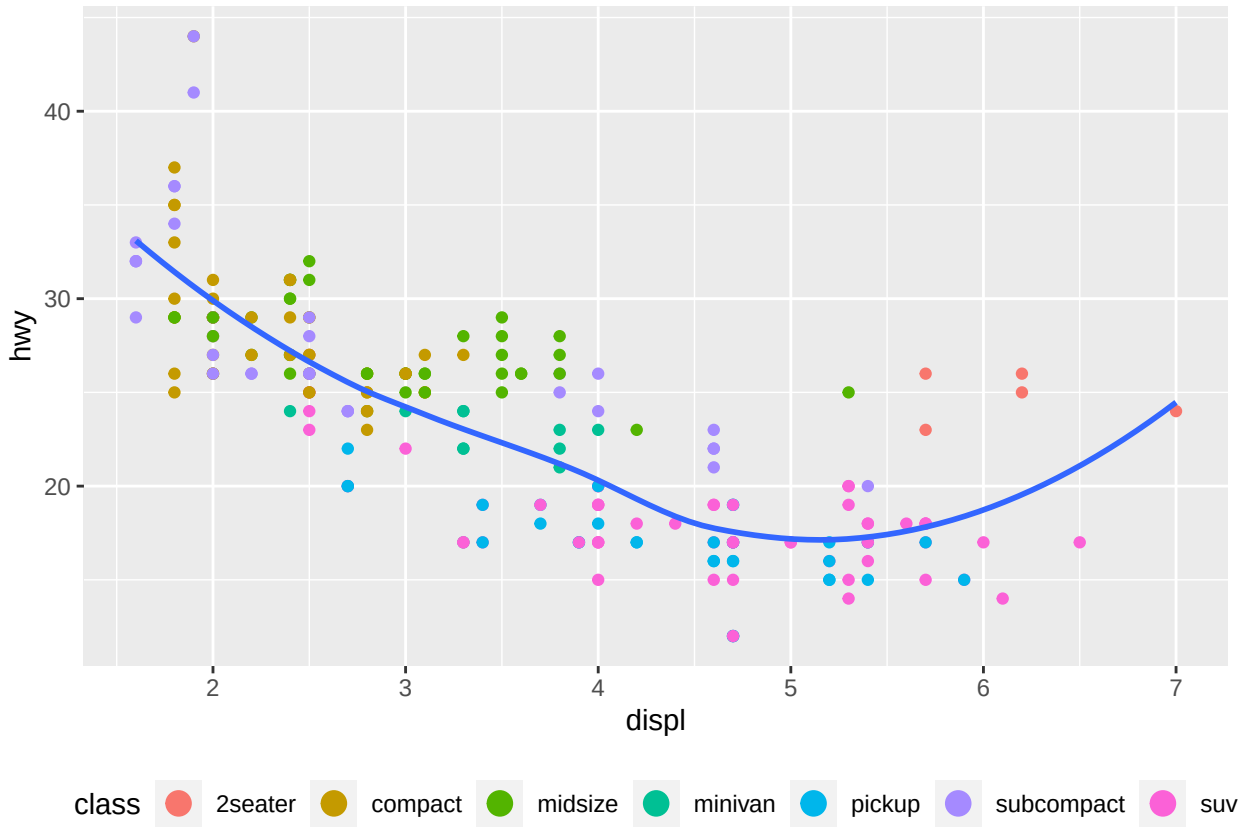



```
base + theme(legend.position = "right") # the default
```



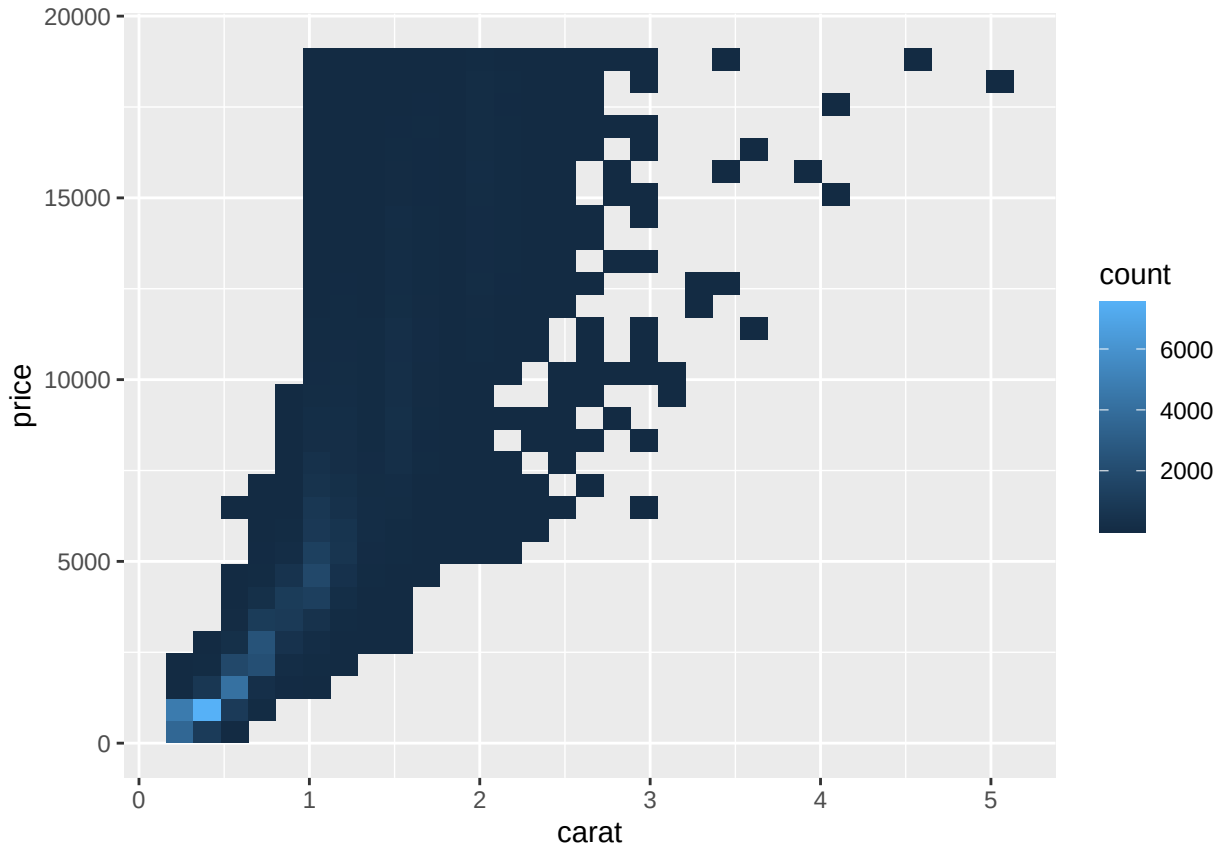

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, and set aesthetics
  geom_point(aes(colour = class)) + #scatterplot of hwy vs displ with class differentiation
  geom_smooth(se = FALSE) + #line of fit without standard error
  theme(legend.position = "bottom") + #position of legend at bottom
  guides(colour = guide_legend(nrow = 1, override.aes = list(size = 4))) #layout of legend in 1 row
```

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

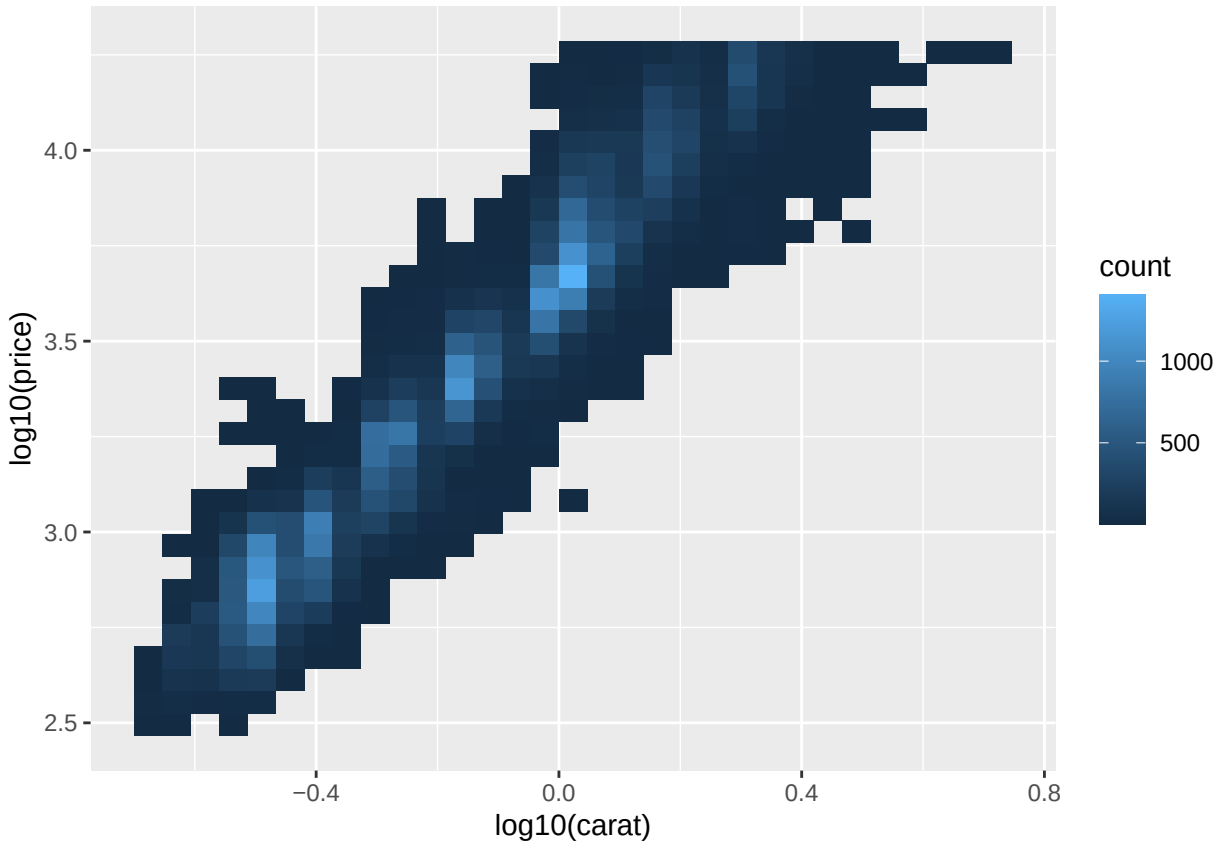



```
#> `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

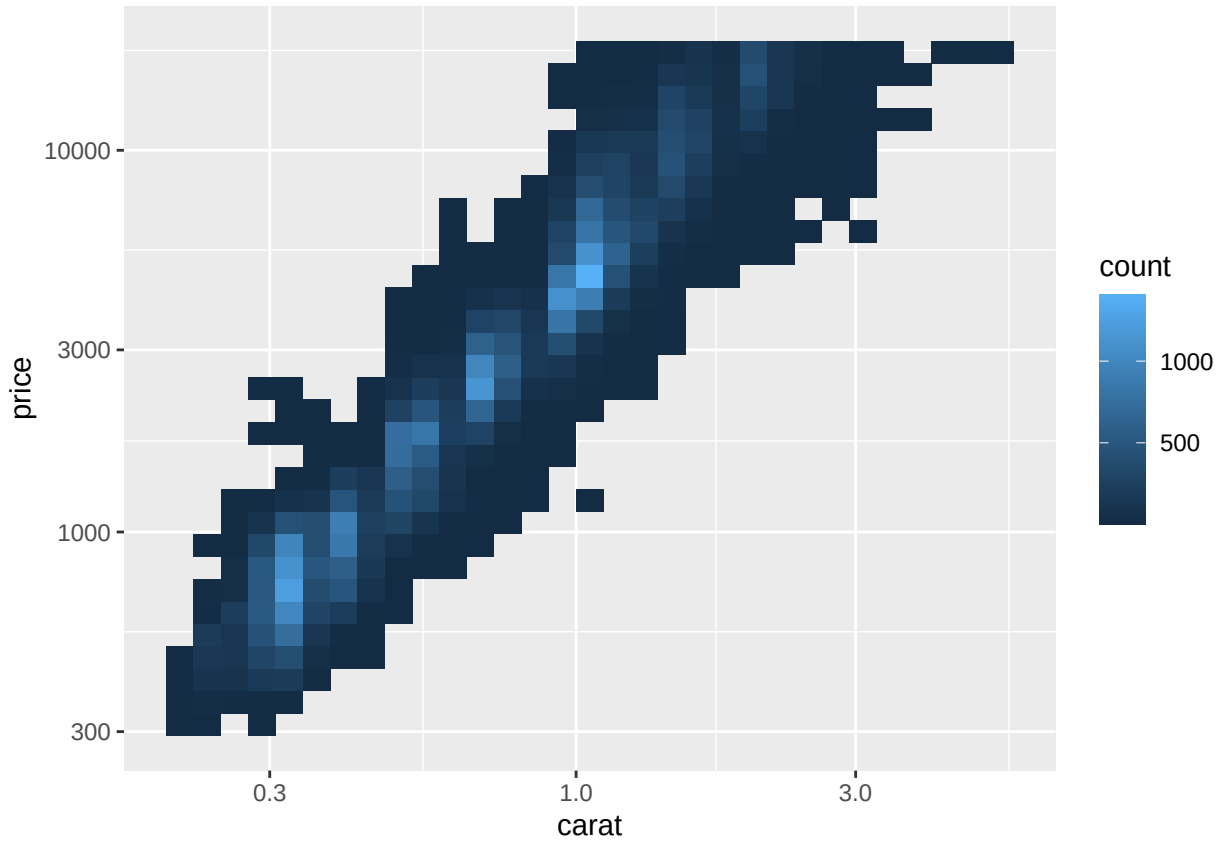
```
ggplot(diamonds, aes(carat, price)) + #ggplot, data load, and set aesthetics
  geom_bin2d() #heatmap of 2d bins
```



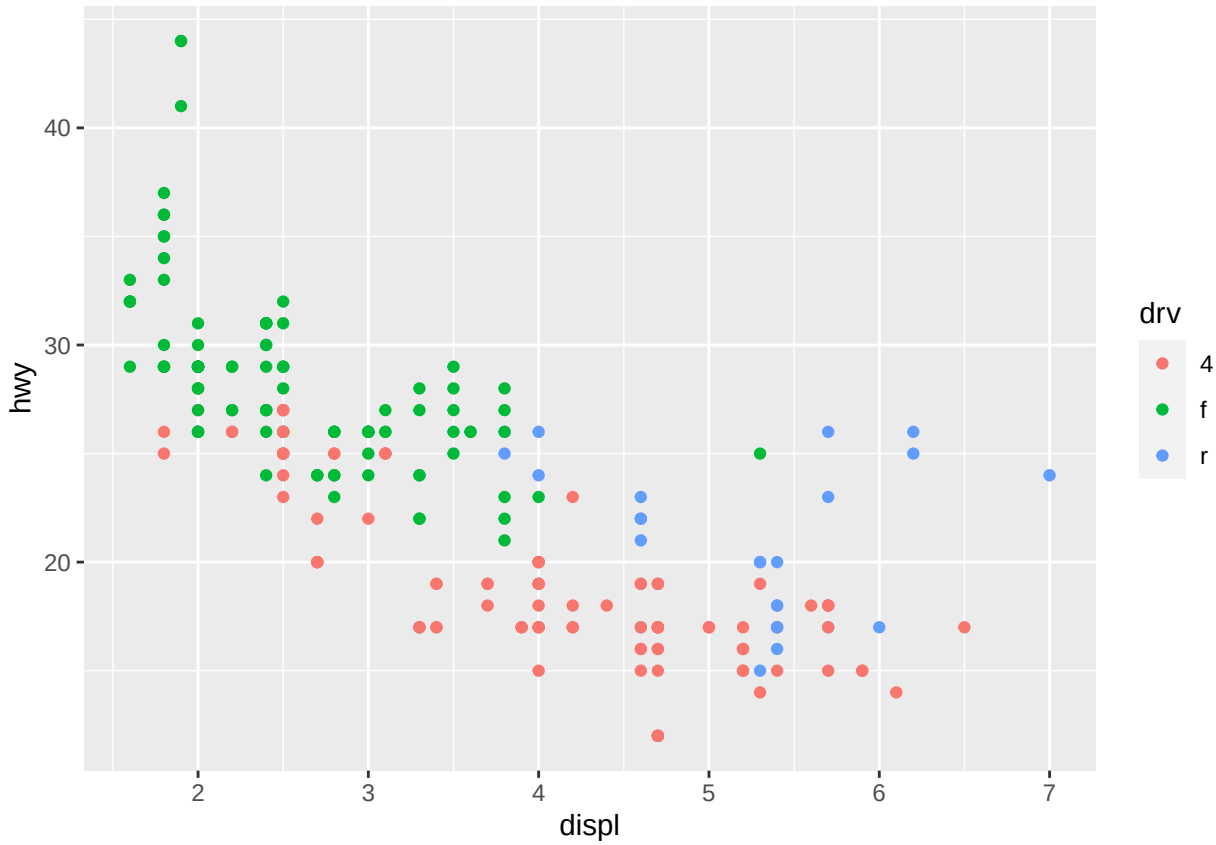

```
ggplot(diamonds, aes(log10(carat), log10(price))) + #ggplot, data load, and log of aesthetics
  geom_bin2d() #heatmap of 2d bins
```



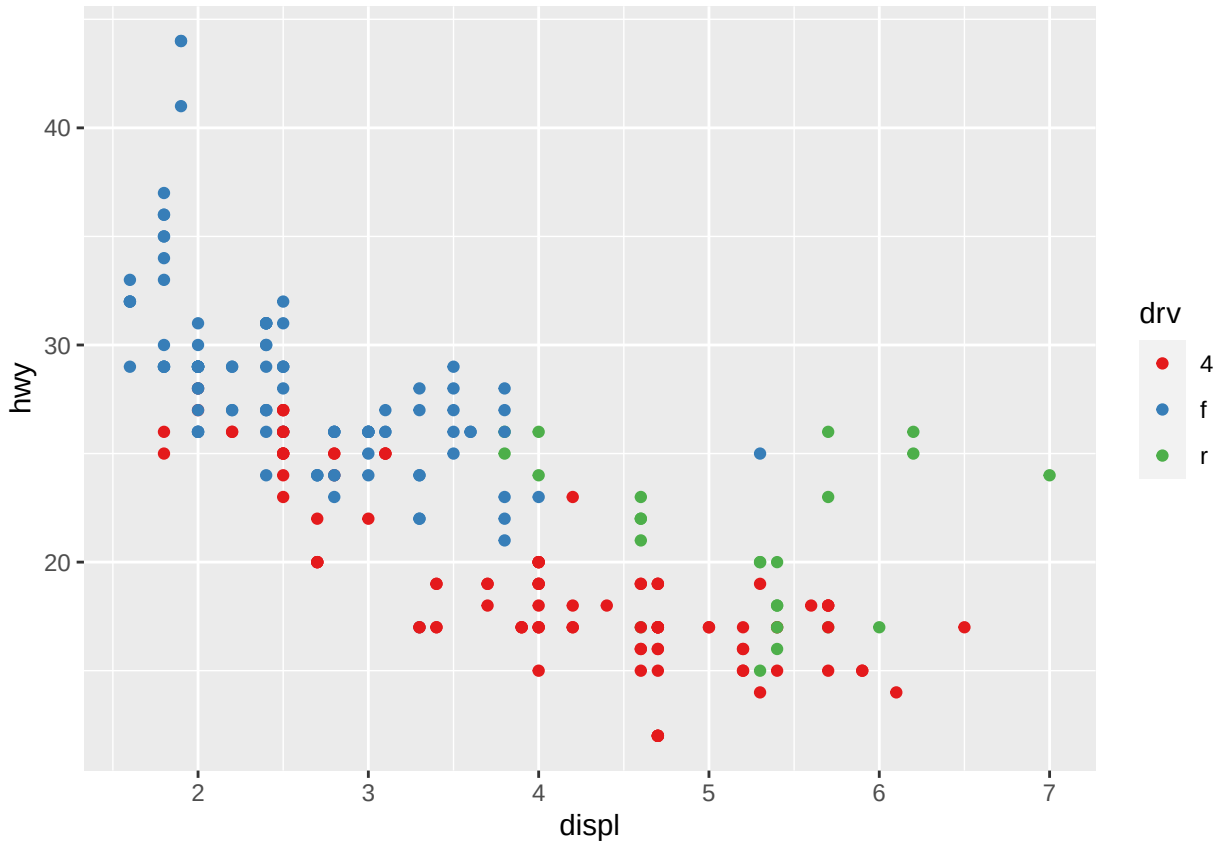

```
ggplot(diamonds, aes(carat, price)) + #ggplot, data load, set aesthetics
  geom_bin2d() + #heatmap of 2d bins
  scale_x_log10() + #log10 of x scale
  scale_y_log10() #log10 of y scale
```



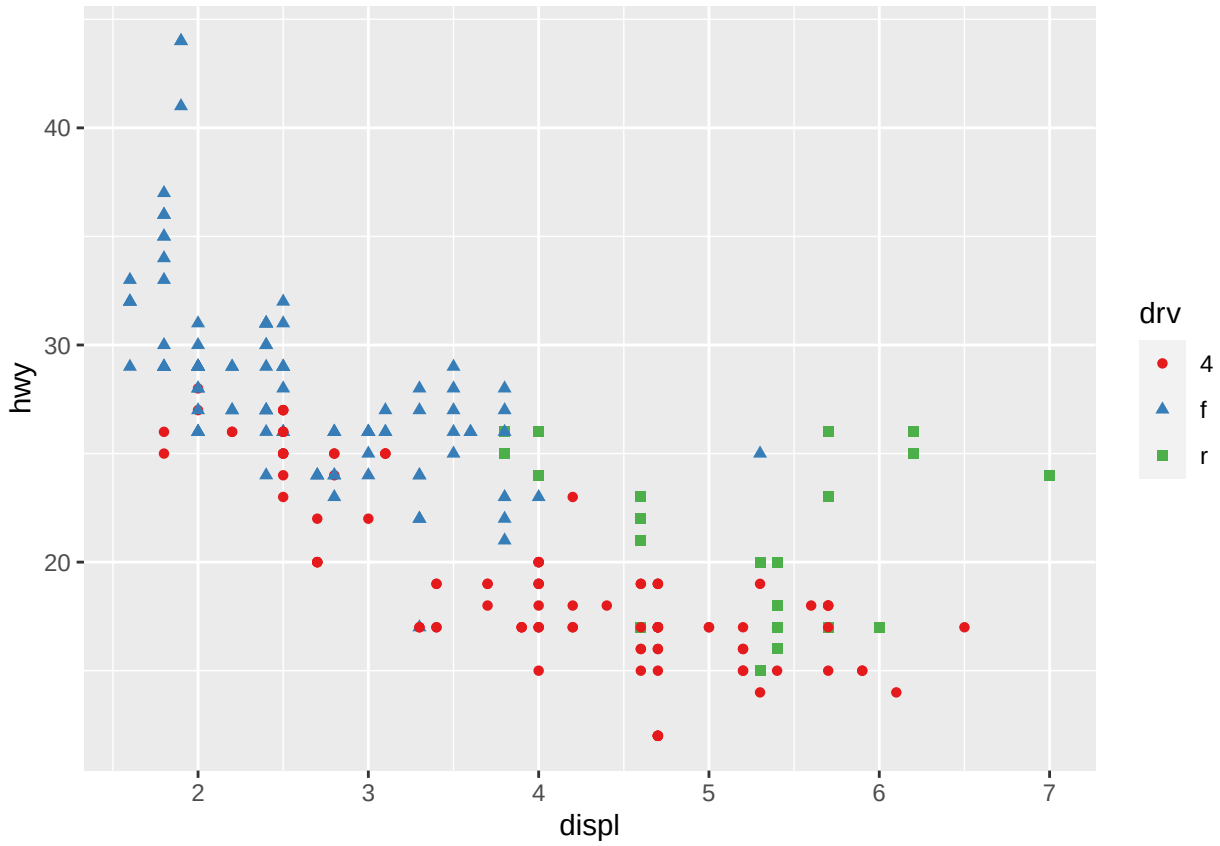

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, set aesthetics
  geom_point(aes(color = drv)) #scatterplot of hwy vs displ with drv color category differentiation
```



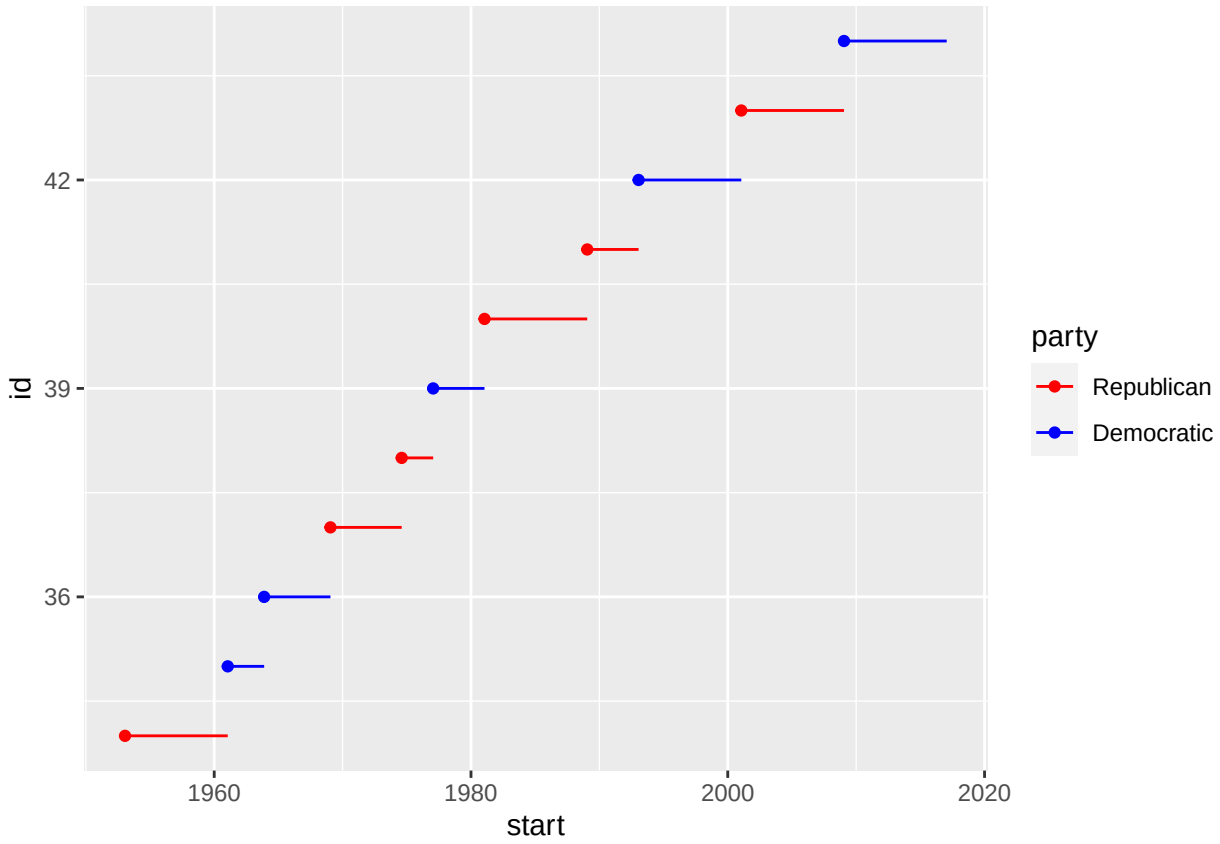

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, set aesthetics
  geom_point(aes(color = drv)) + #scatterplot of hwy vs displ with drv color category differentiation
  scale_colour_brewer(palette = "Set1") #scale by color
```



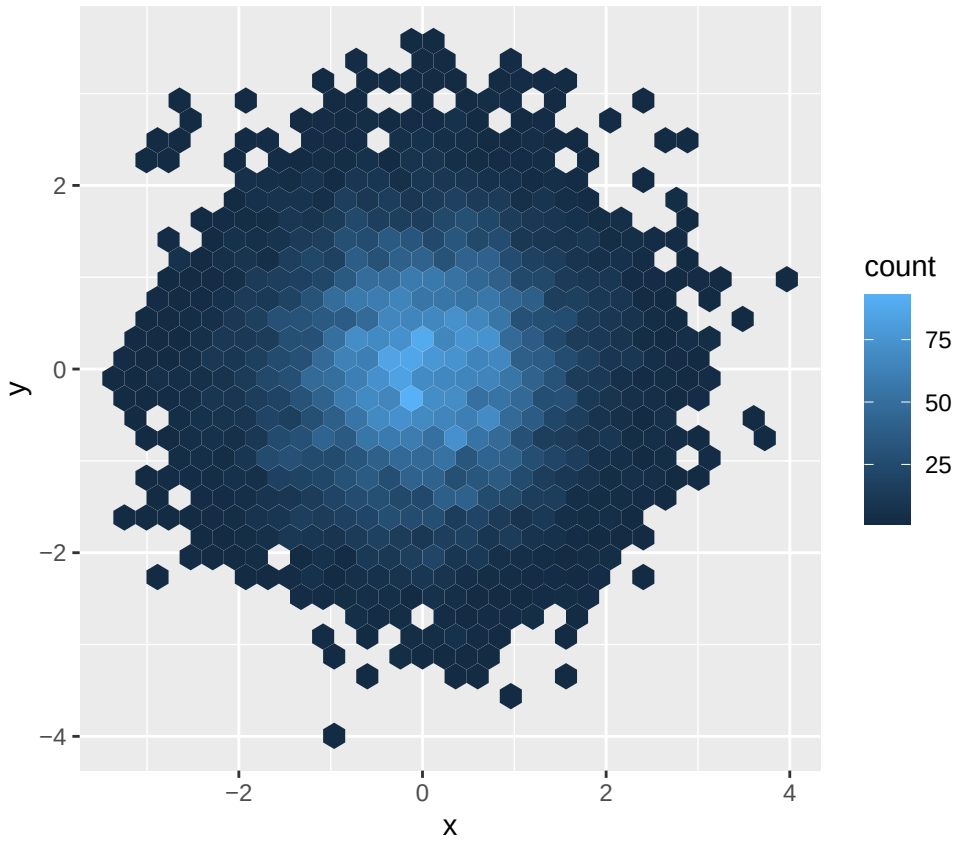

```
ggplot(mpg, aes(displ, hwy)) + #ggplot, data load, set aesthetics
  geom_point(aes(color = drv, shape = drv)) + #scatterplot of hwy vs displ with drv color category and
  scale_colour_brewer(palette = "Set1") #scale by color
```



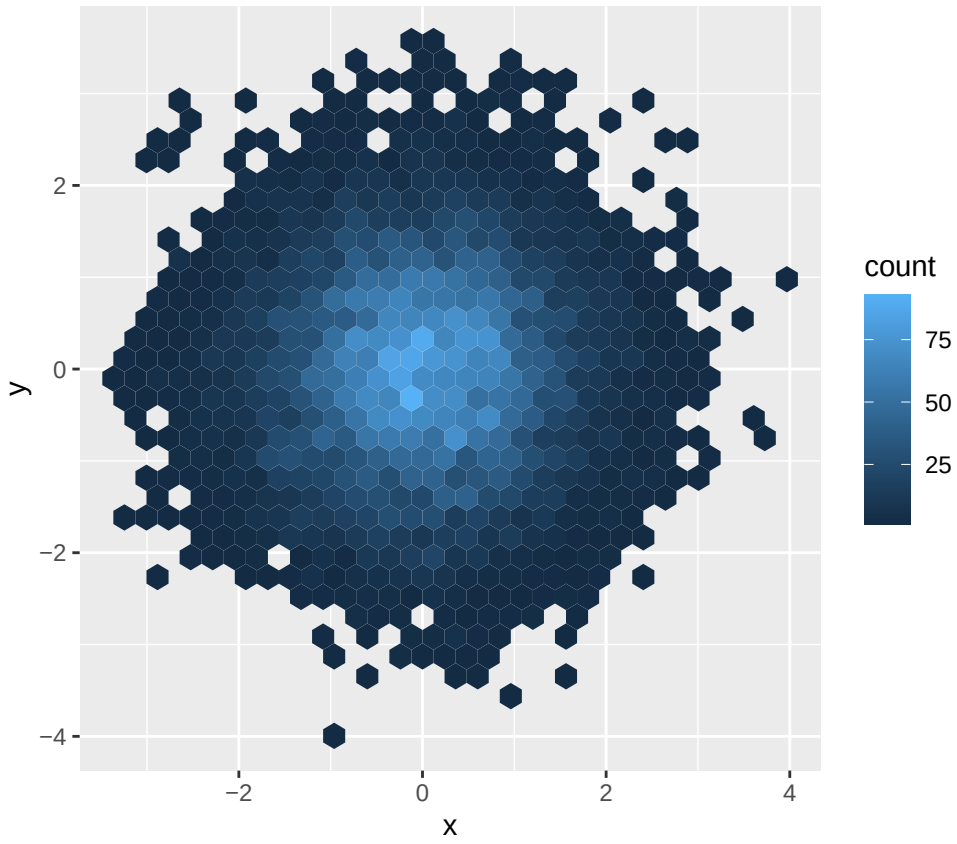

```
presidential %>%
  mutate(id = 33 + row_number()) %>% #select columns and rows
  ggplot(aes(start, id, colour = party)) + #ggplot, data load, set aesthetics
    geom_point() + #scatterplot
    geom_segment(aes(xend = end, yend = id)) + #segment plot
    scale_colour_manual(values = c(Republican = "red", Democratic = "blue")) #color scale for certain p
```




```
df <- tibble( #create data
  x = rnorm(10000),
  y = rnorm(10000)
)
ggplot(df, aes(x, y)) + #ggplot, data load, set aesthetics
  geom_hex() + #create hexagonal shape
  coord_fixed() #fixed aspect ratio
```

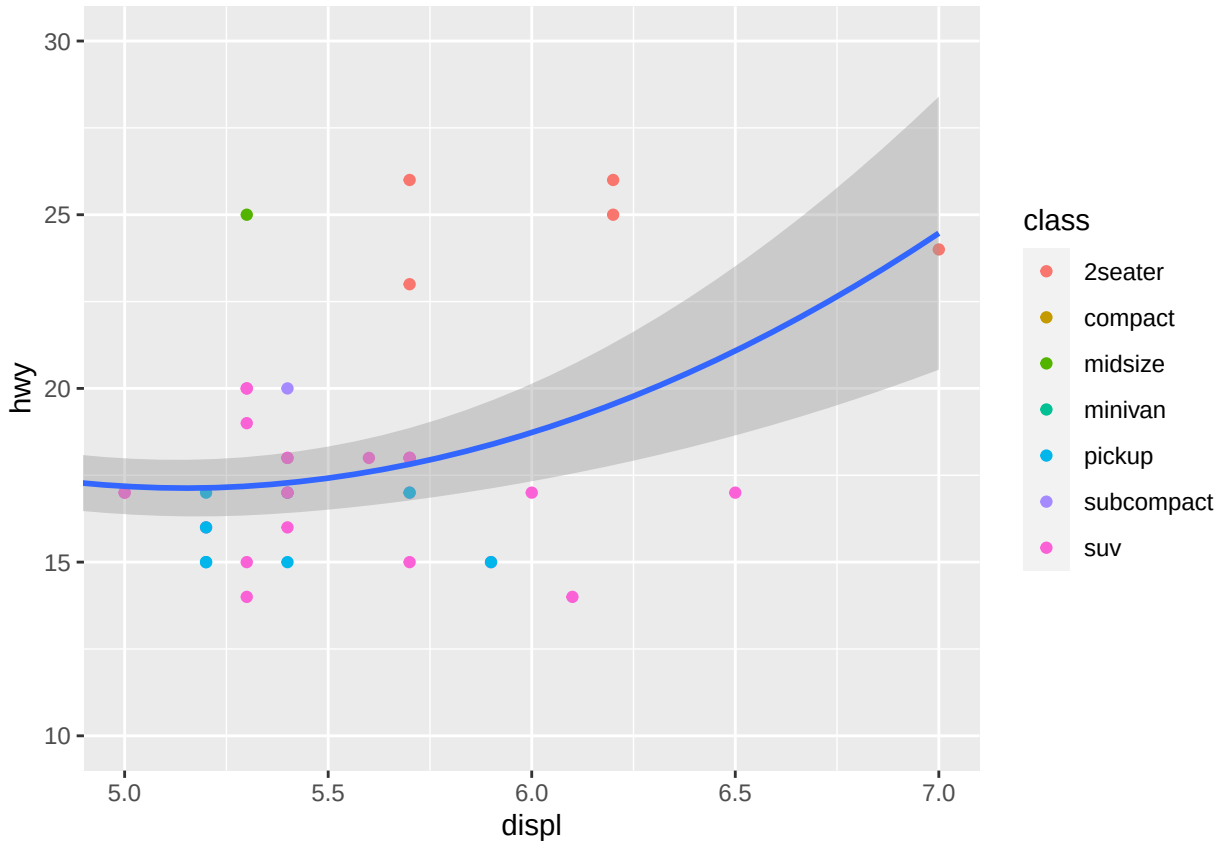



```
ggplot(df, aes(x, y)) + #ggplot, data load, set aesthetics
  geom_hex() + #create hexagonal shape
  #viridis::scale_fill_viridis() + #contrasting colors in shape
  coord_fixed() #fixed aspect ratio
```



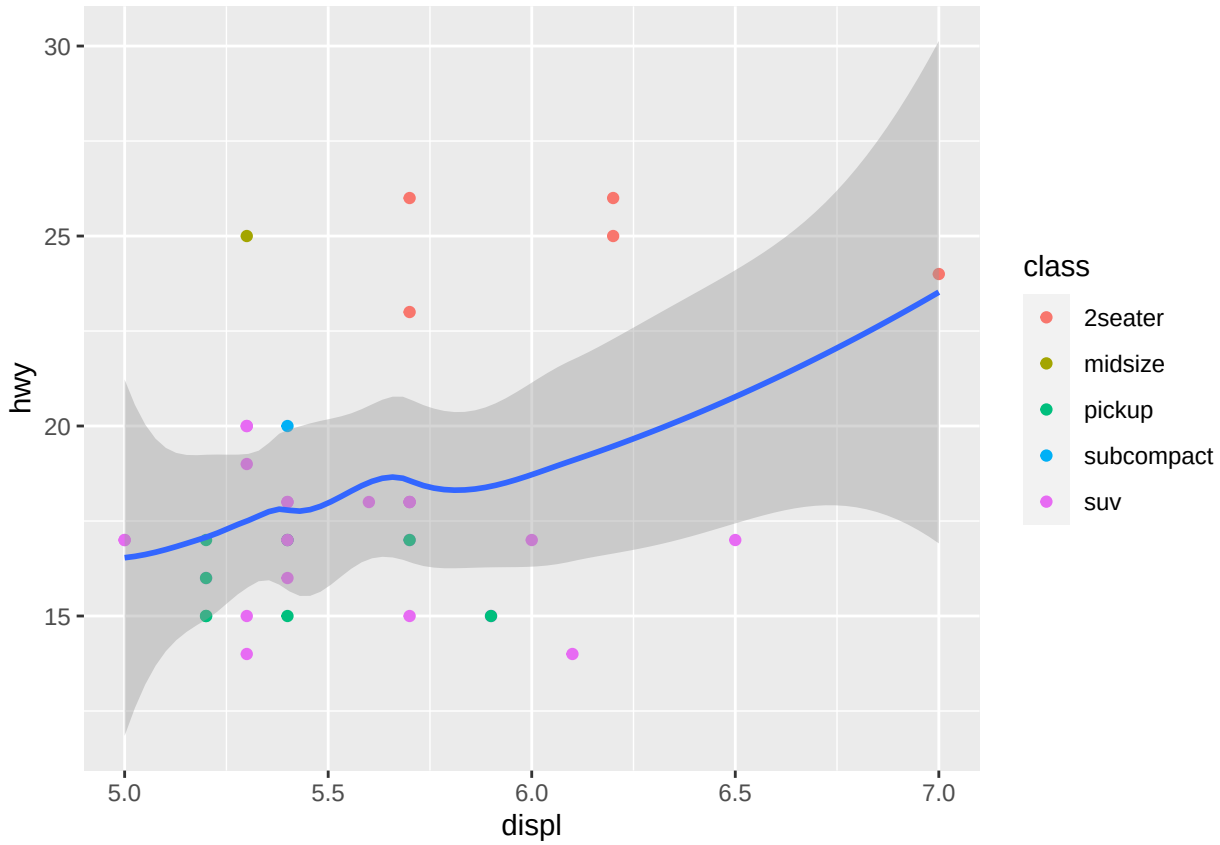

```
ggplot(mpg, mapping = aes(displ, hwy)) + #ggplot, data load, set aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with class color category differentiati
  geom_smooth() + #line of fit on top of scatterplot
  coord_cartesian(xlim = c(5, 7), ylim = c(10, 30)) #adjust x and y limits
```

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

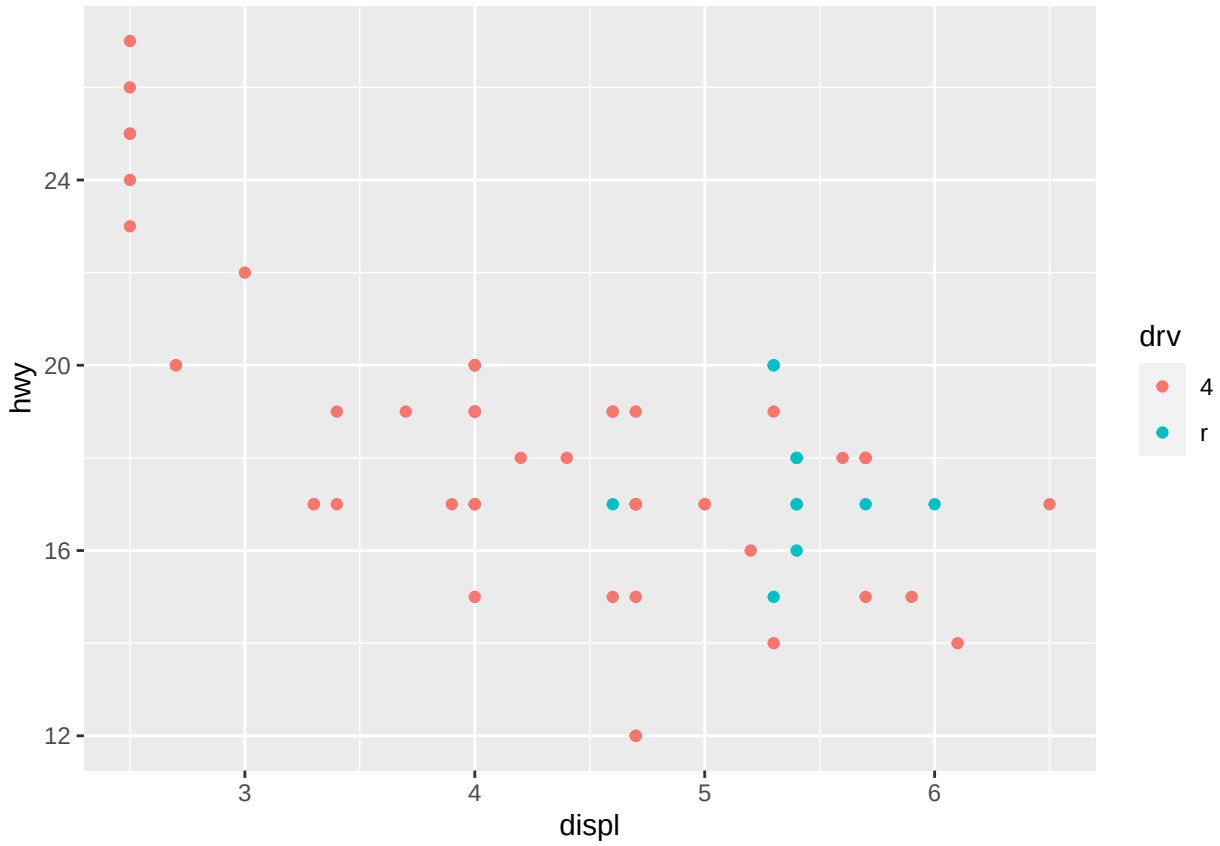



```
mpg %>%
  filter(displ >= 5, displ <= 7, hwy >= 10, hwy <= 30) %>% #filter rows
  ggplot(aes(displ, hwy)) + #ggplot, data load, set aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with class color category differentiati
  geom_smooth() #line of fit on top of scatterplot
```

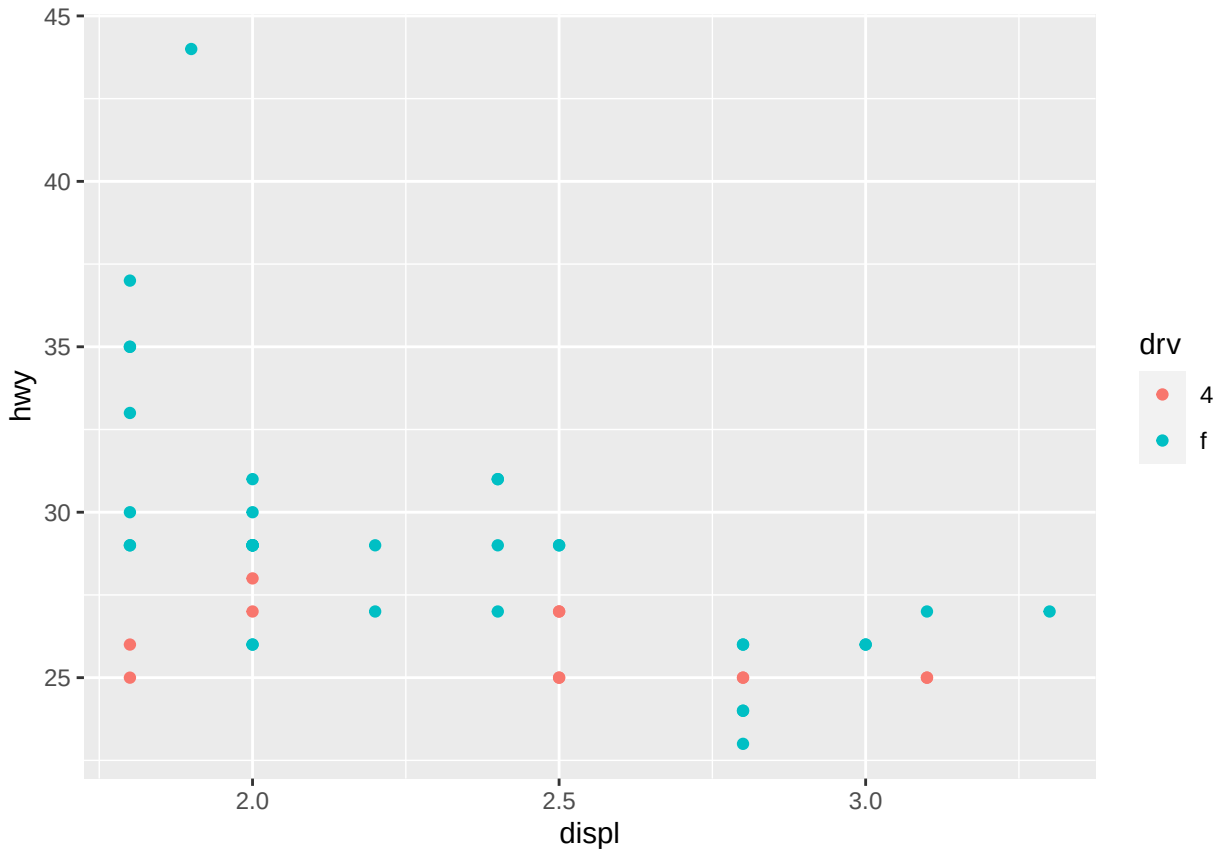
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



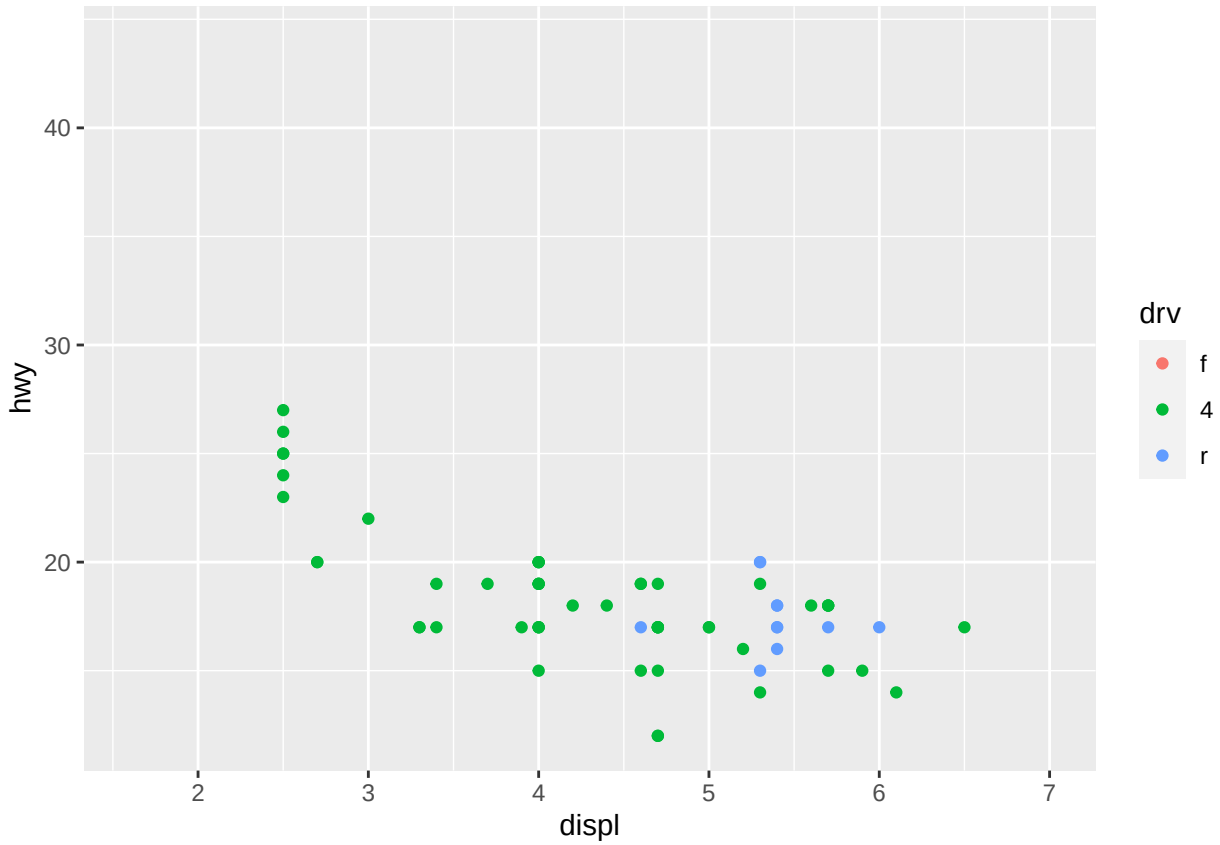

```
suv <- mpg %>% filter(class == "suv") #create suv dataframe
compact <- mpg %>% filter(class == "compact") #create compact dataframe

ggplot(suv, aes(displ, hwy, colour = drv)) + #ggplot, data load, set aesthetics
  geom_point() #scatterplot of hwy vs displ with drv color category differentiation
```



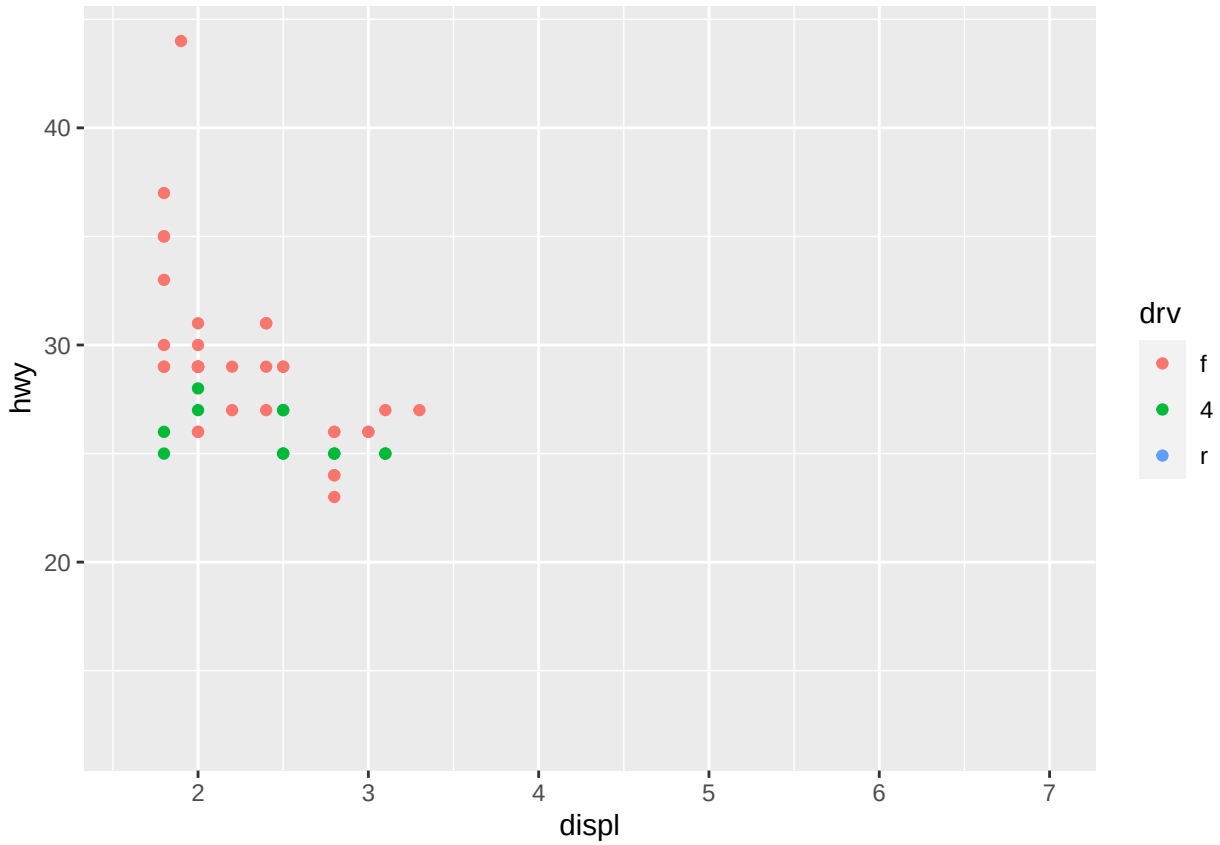

```
ggplot(compact, aes(displ, hwy, colour = drv)) + #ggplot, data load, set aesthetics
  geom_point() #scatterplot of hwy vs displ with drv color category differentiation
```




```
x_scale <- scale_x_continuous(limits = range(mpg$displ)) #create x_scale dataframe
y_scale <- scale_y_continuous(limits = range(mpg$hwy)) #create y_scale dataframe
col_scale <- scale_colour_discrete(limits = unique(mpg$drv)) #create col_scale dataframe

ggplot(suv, aes(displ, hwy, colour = drv)) + #ggplot, data load, set aesthetics
  geom_point() + #scatterplot of hwy vs displ with drv color category differentiation
  x_scale + #adjust x scale
  y_scale + #adjust y scale
  col_scale #adjust color scale
```

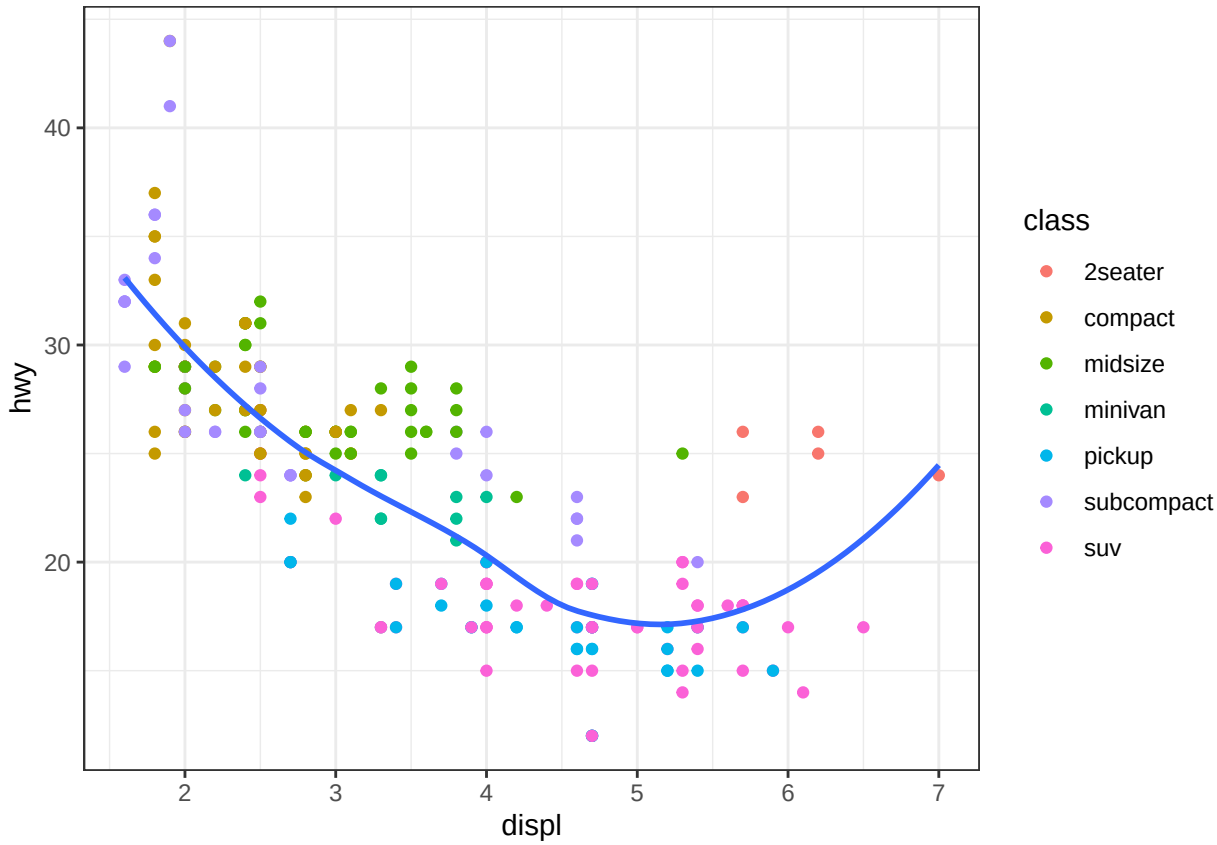



```
ggplot(compact, aes(displ, hwy, colour = drv)) + #ggplot, data load, set aesthetics
  geom_point() + #scatterplot of hwy vs displ with drv color category differentiation
  x_scale + #adjust x scale
  y_scale + #adjust y scale
  col_scale #adjust color scale
```

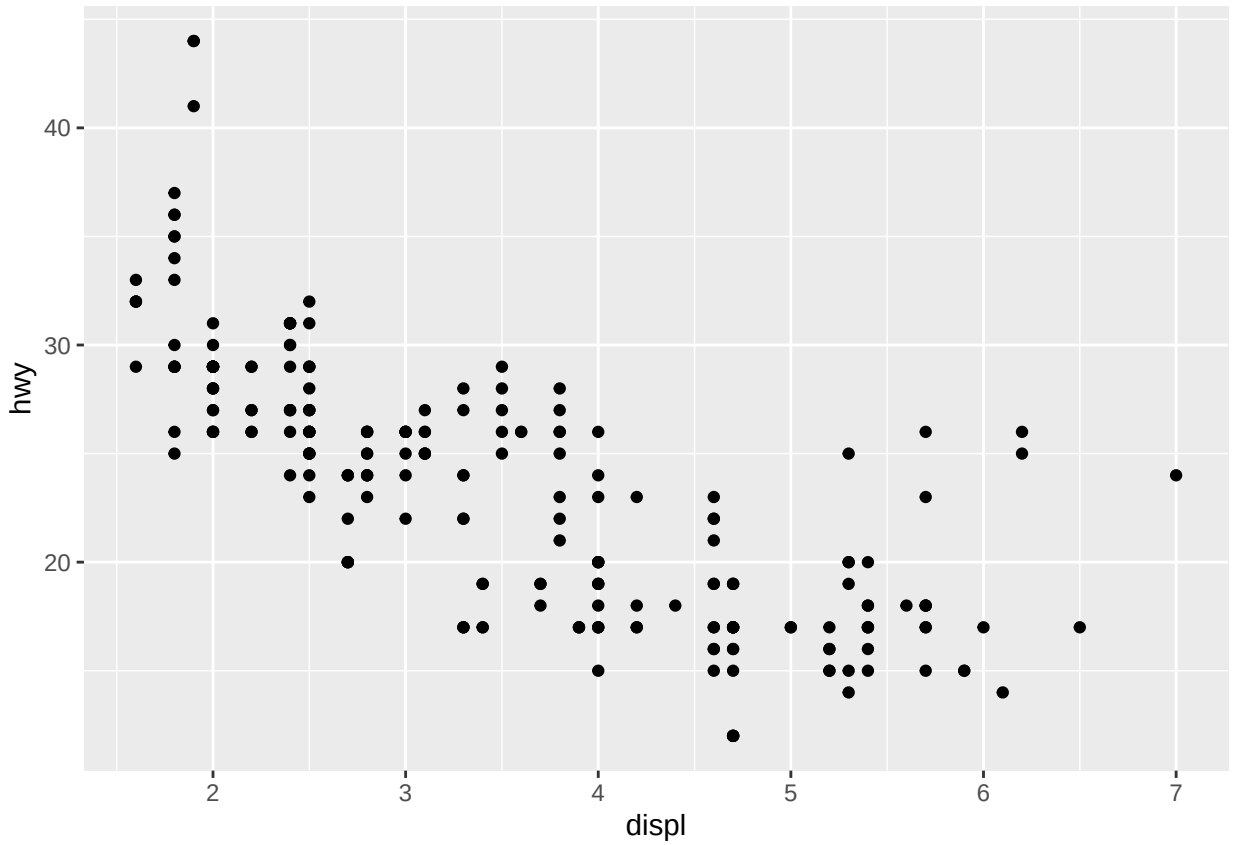



```
ggplot(mpg, aes(displ, hwy)) +  #ggplot, data load, set aesthetics
  geom_point(aes(color = class)) + #scatterplot of hwy vs displ with class color category differentiati
  geom_smooth(se = FALSE) + #line of best fit without standard error
  theme_bw() #adjust theme
```

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




```
ggplot(mpg, aes(displ, hwy)) + geom_point() #create scatterplot of hwy vs displ
```




```
ggsave("my-plot.pdf") #save plot
```

```
## Saving 6.5 x 4.5 in image
```