problemset_1a.R

Aaryan Agarwal

2023-09-19

```
# PROBLEM SET 1A
# This assignment is done in a group of three
# Prajwal Kaushal, Sumukha Sharma, Aaryan Agarwal

################################################################################
#Question 2: Data Visualization
################################################################################
#Question 2.1: Visualization Basics
################################################################################
library(ggplot2)
```

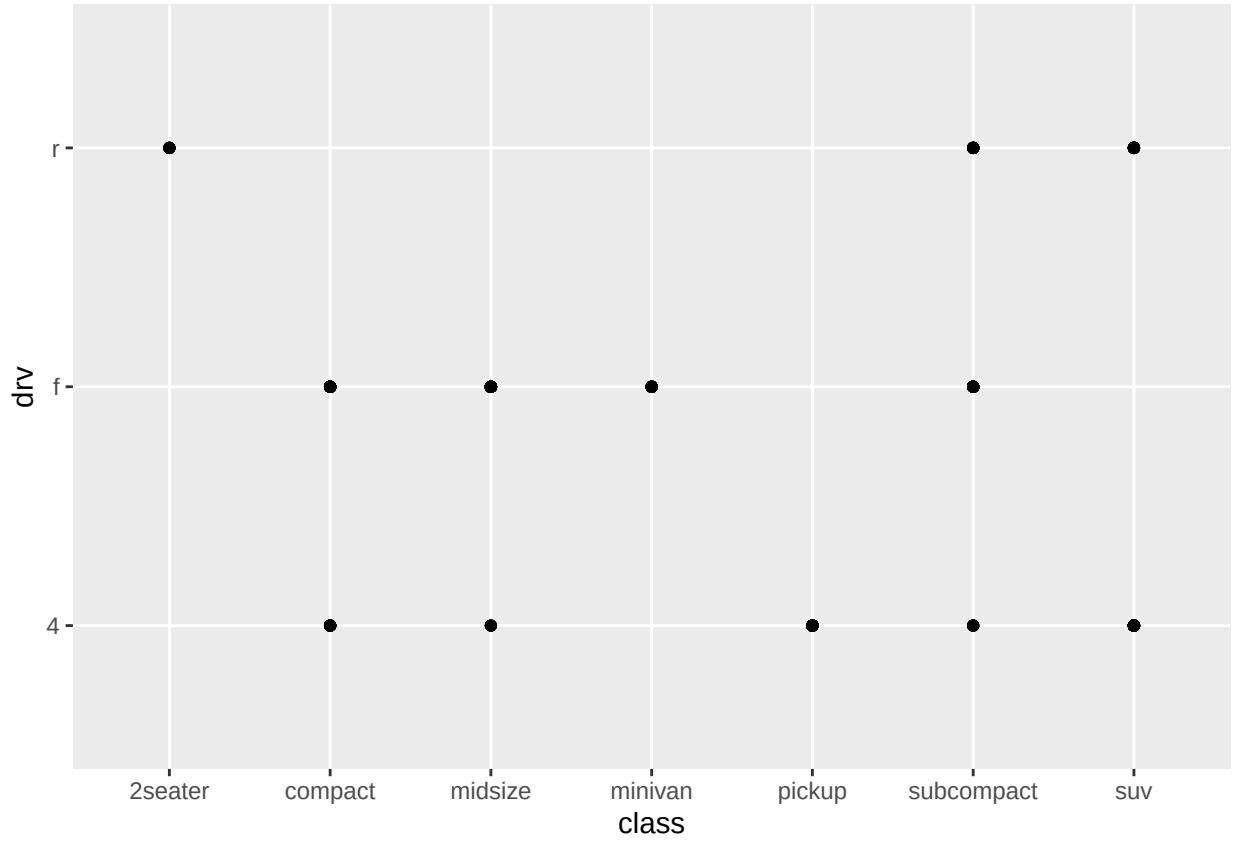
```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
# 1. Run ggplot(data = mpg). What do you see and why?
ggplot(data = mpg)
```

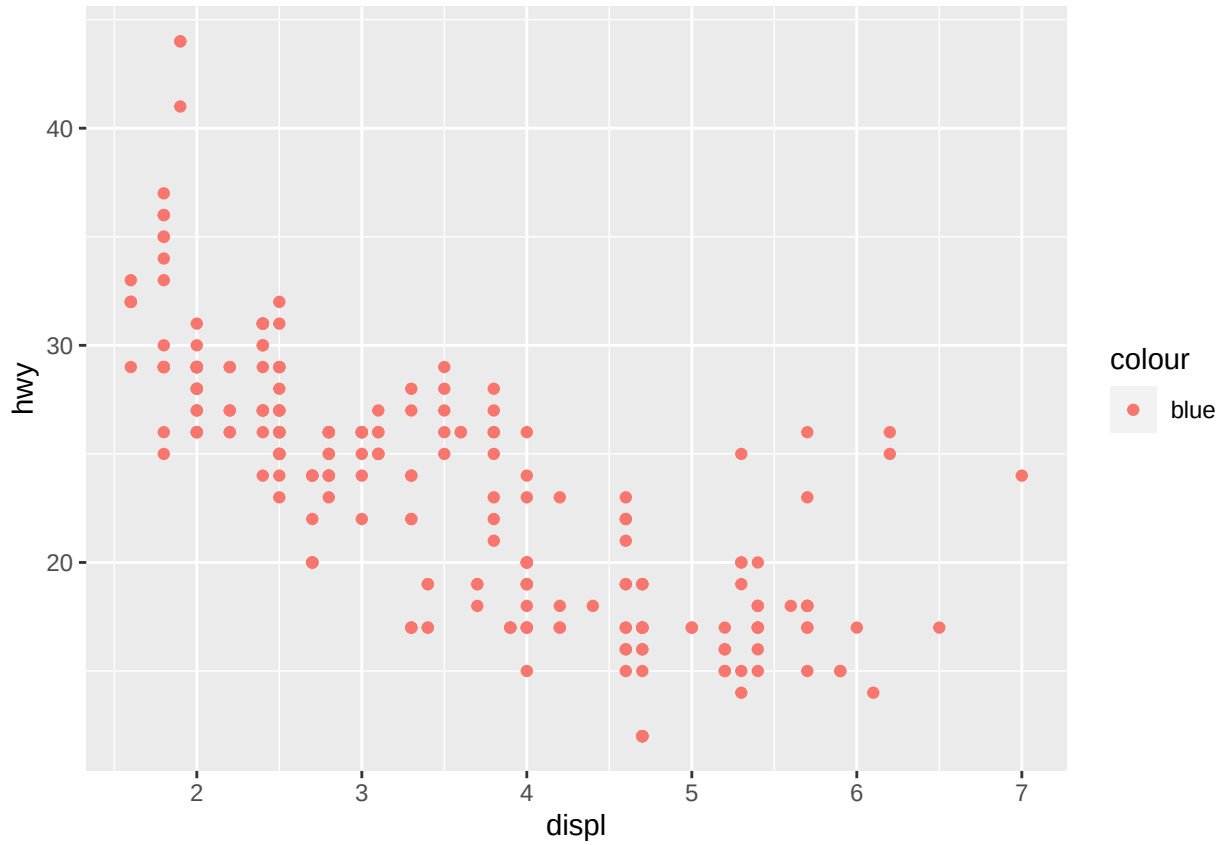
```
# Answer: It just creates the background of the plot, because we have to use
# the geom function and specify the layers for it to work correctly.

# 2. What does the drv variable describe? Read the help for ?mpg to find out.
?mpg
```

```
## starting httpd help server ... done
```

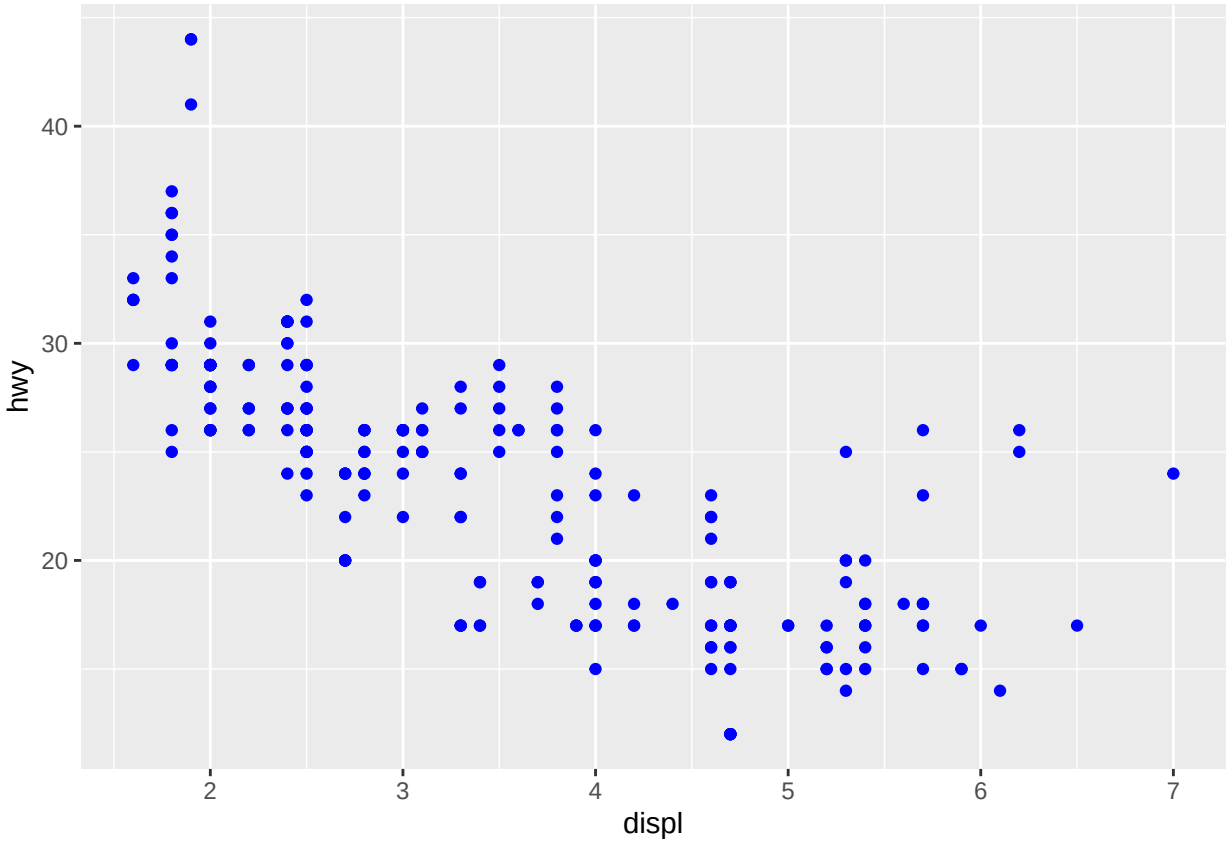
```
# Answer: It describes the type of drive, it has 3 categories:
# f = front wheel drive
# r = rear wheel drive
# 4 = 4 wheel drive

# 3. What happens if you make a scatterplot of class vs drv? Why is the plot not useful?
ggplot(mpg, aes(x = class, y = drv)) + geom_point()
```




```
# Answer: This code gives a non-useful plot, there are two reasons for that, the first
# is both of them are categorical variables and scatterplots are usually good for continuous
# variables. The second point is that there are limited amount of categories in both
# drv and class variable and many points will overlap.

################################################################################
# Question 2.2: Aesthetic Mappings
################################################################################
# 1. What's gone wrong with this code? Why are the points not blue?
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```



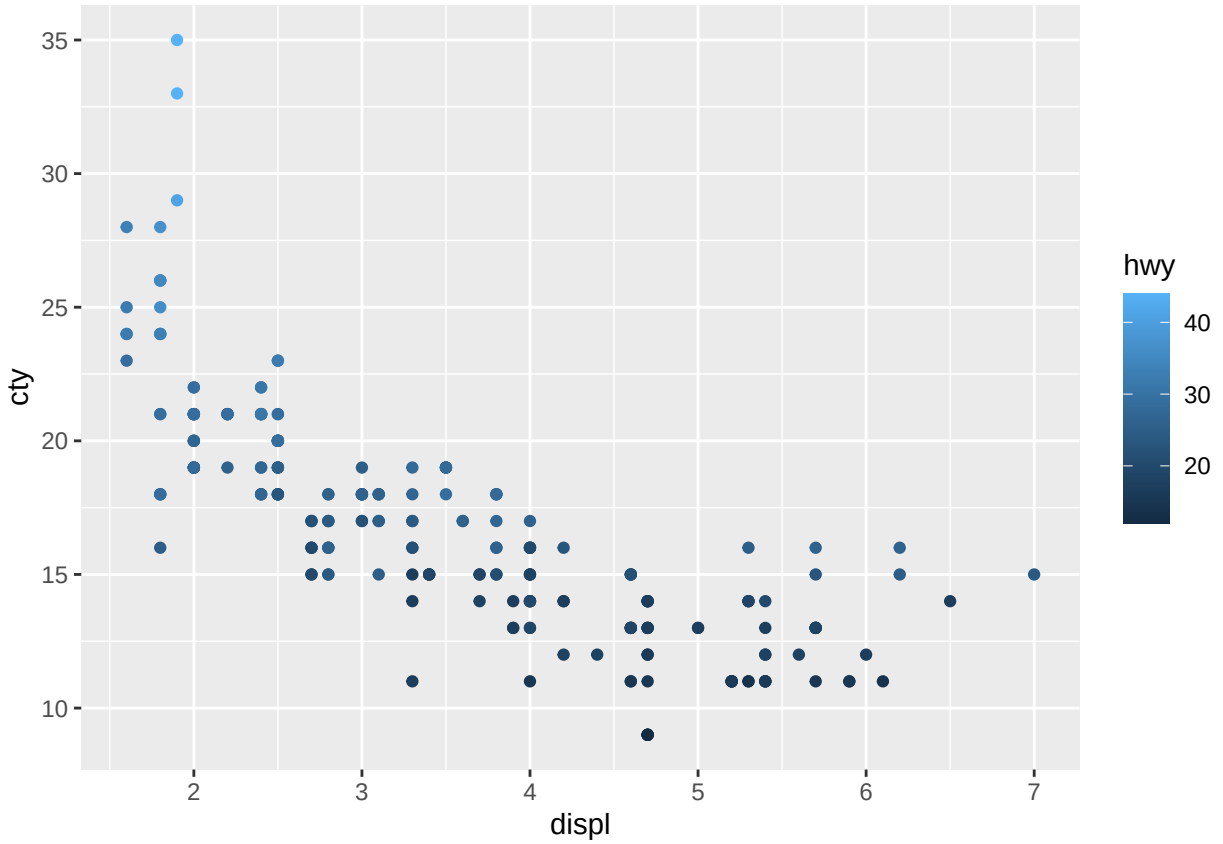

```
# Answer: The issue with the code is the placement of the color aesthetic.
# When you place the color = "blue" inside the aes() function, you're essentially telling
# ggplot2 to create a color mapping based on a constant value "blue",
# not to color the points blue.
# The correct code is:
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue") # color is outside aes
```



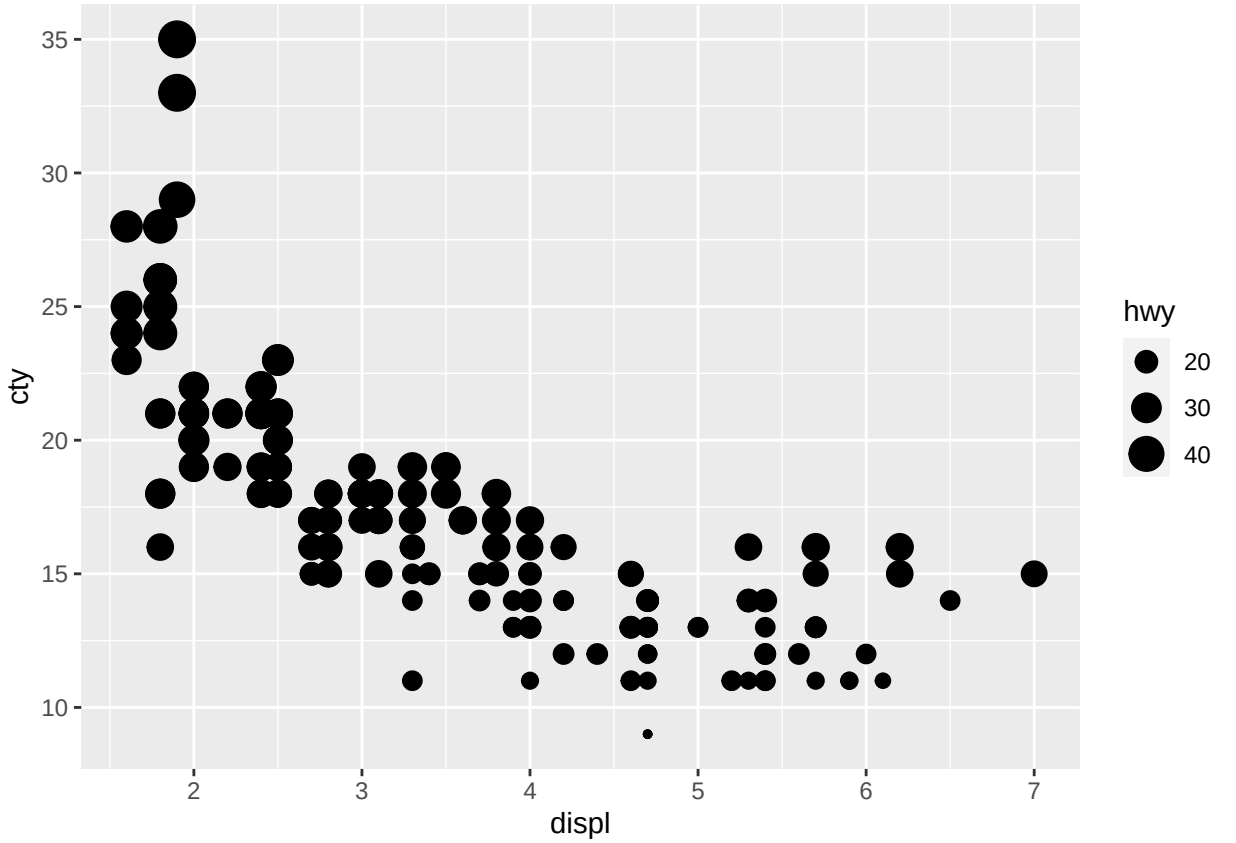

```
# 2. Which variables in mpg are categorical? Which variables are continuous? (Hint: type
# ?mpg to read the documentation for the dataset). How can you see this information
# when you run mpg?

# Answer: Categorical variables in the mpg dataset are:
# manufacturer model trans (transmission type)
# drv (type of drive train)
# fl (fuel type)
# class (class of car)
# To see the structure of the mpg dataset in R, including the types of each variable,
# we can use the str()
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

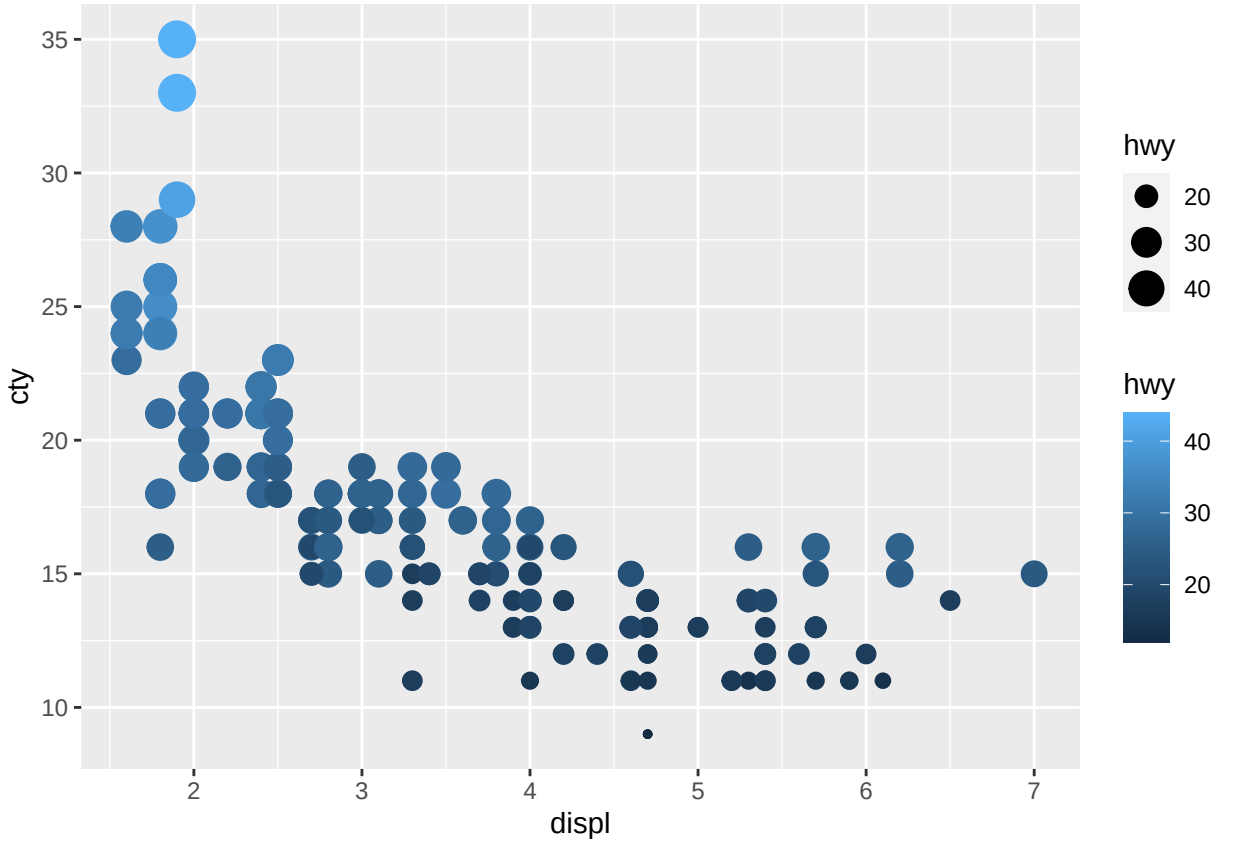
```r
# 3. Map a continuous variable to color, size, and shape. How do these aesthetics behave
# differently for categorical vs. continuous variables?

# Answer: Mapping to color:
ggplot(data = mpg, aes(x = displ, y = cty, color = hwy)) +
  geom_point()
```



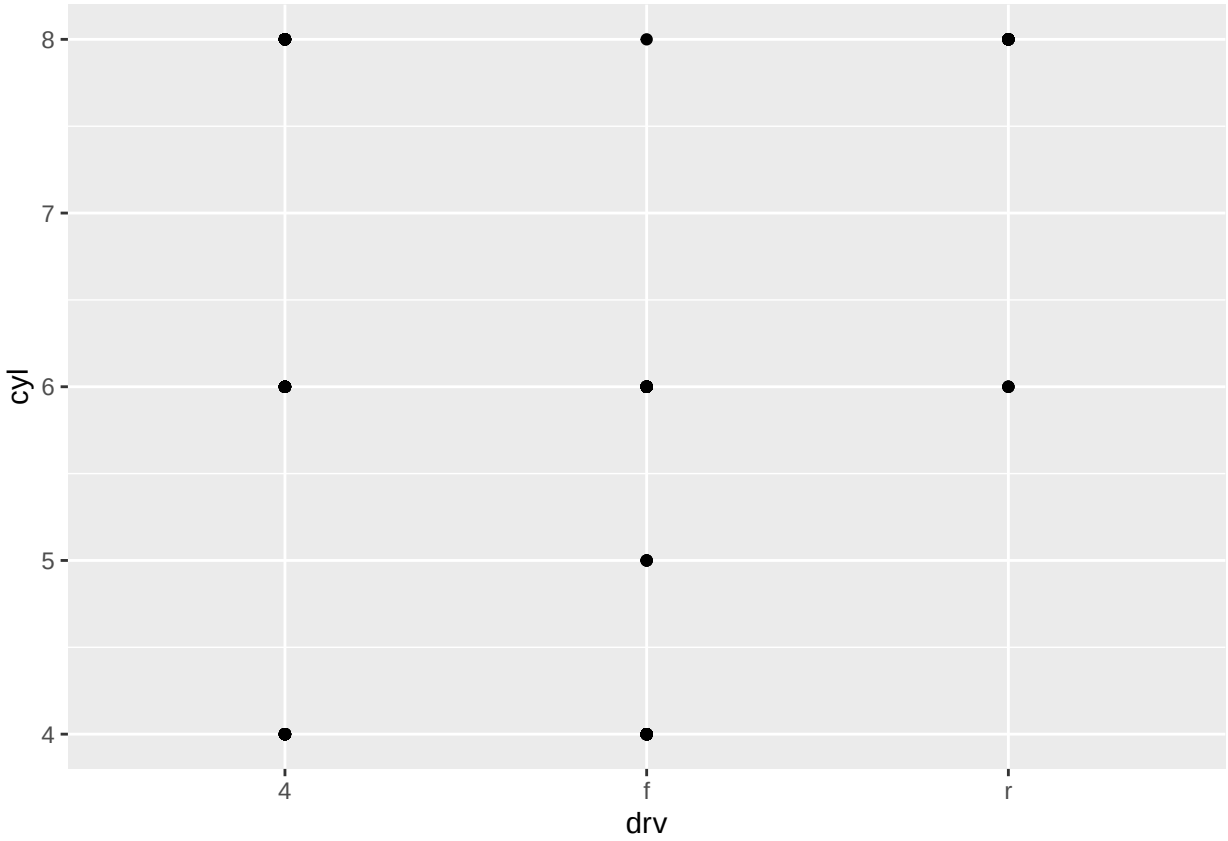

```r
# For a continuous variable mapped to color: A color gradient is used.
# By default, low values are mapped to one end of the color spectrum
# and high values to the other end (often red).

# Mapping to size:
ggplot(mpg, aes(x = displ, y = cty, size = hwy)) +
  geom_point()
```



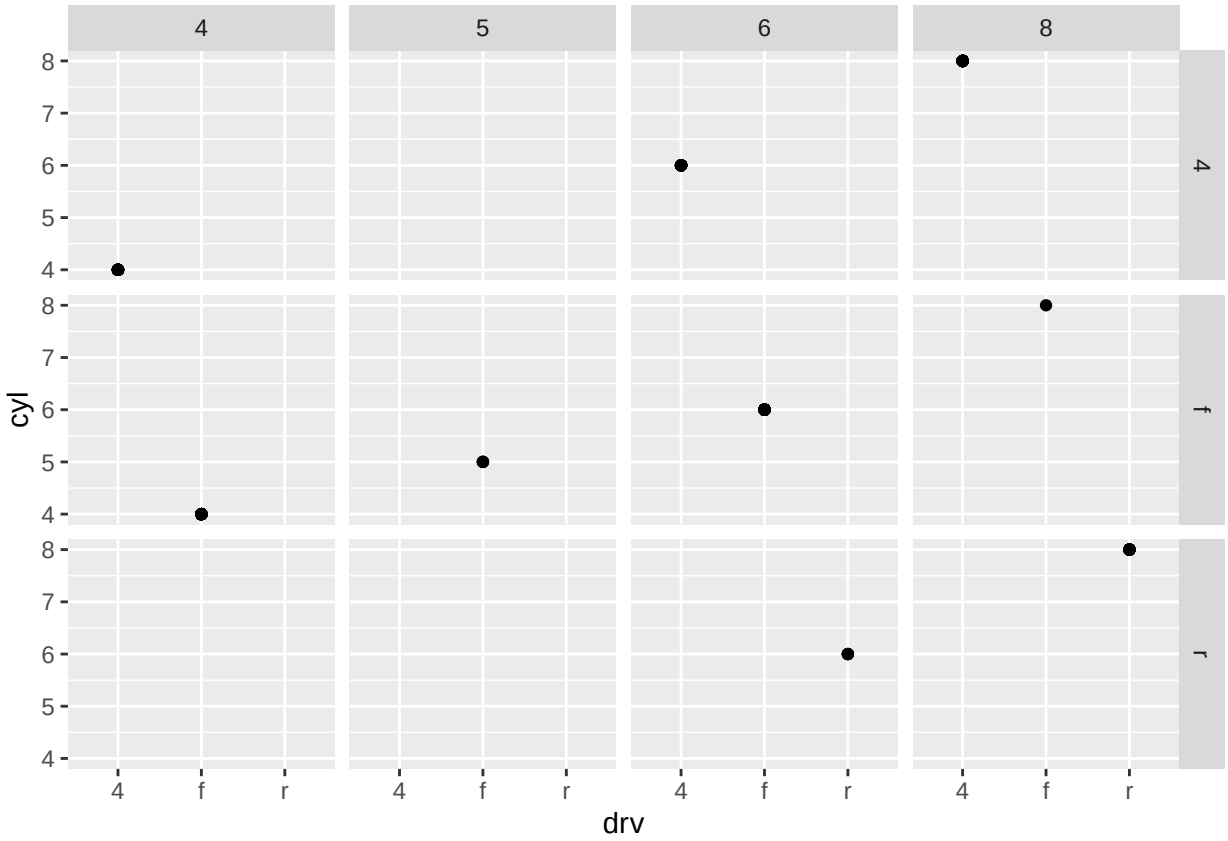

```
# For a continuous variable mapped to size:
# Point sizes vary continuously based on the values of the variable.
# Larger sizes correspond to higher values, and smaller sizes correspond to lower values.

# Mapping to shape:
#ggplot(mpg, aes(x = displ, y = cty, shape = hwy)) +
#  geom_point()

# shows an error because shape cannot handle a continuous scale by default.

# 4. What happens if you map the same variable to multiple aesthetics?
# Answer:
ggplot(mpg, aes(x = displ, y = cty, color = hwy, size = hwy)) +
  geom_point()
```



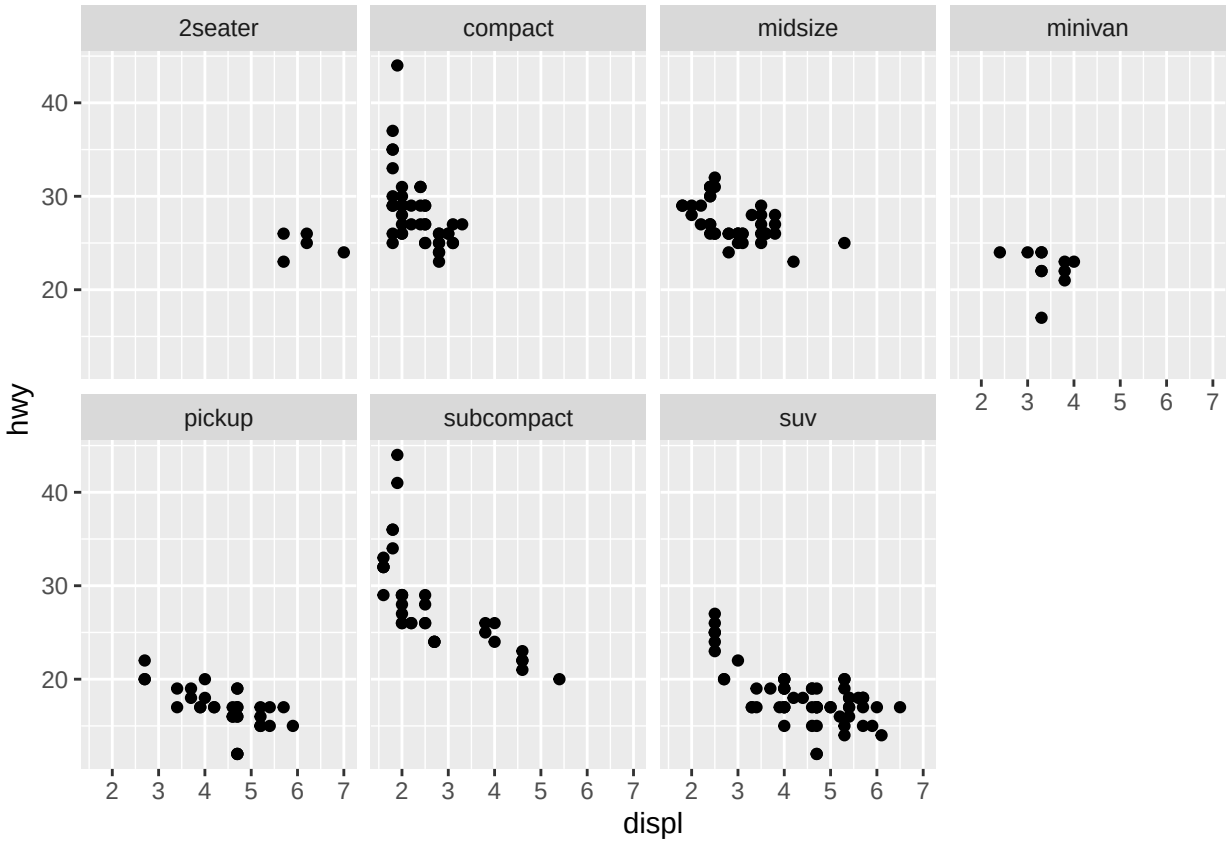

```
# This can tell the trends in the data but can also make the plot more complex
# and harder to interpret.

################################################################################
# Question 2.3: Facets
################################################################################

# 1. What happens if you facet on a continuous variable?

# Answer: When we facet on a continuous variable, the variable's range gets split into
# intervals, and a separate plot is created for each interval. The continuous variable
# is converted to a categorical variable, and the plot contains a facet for each distinct value.

# 2. What do the empty cells in plot with facet_grid(drv ~ cyl) mean? How do they
# relate to this plot?
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))
```




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



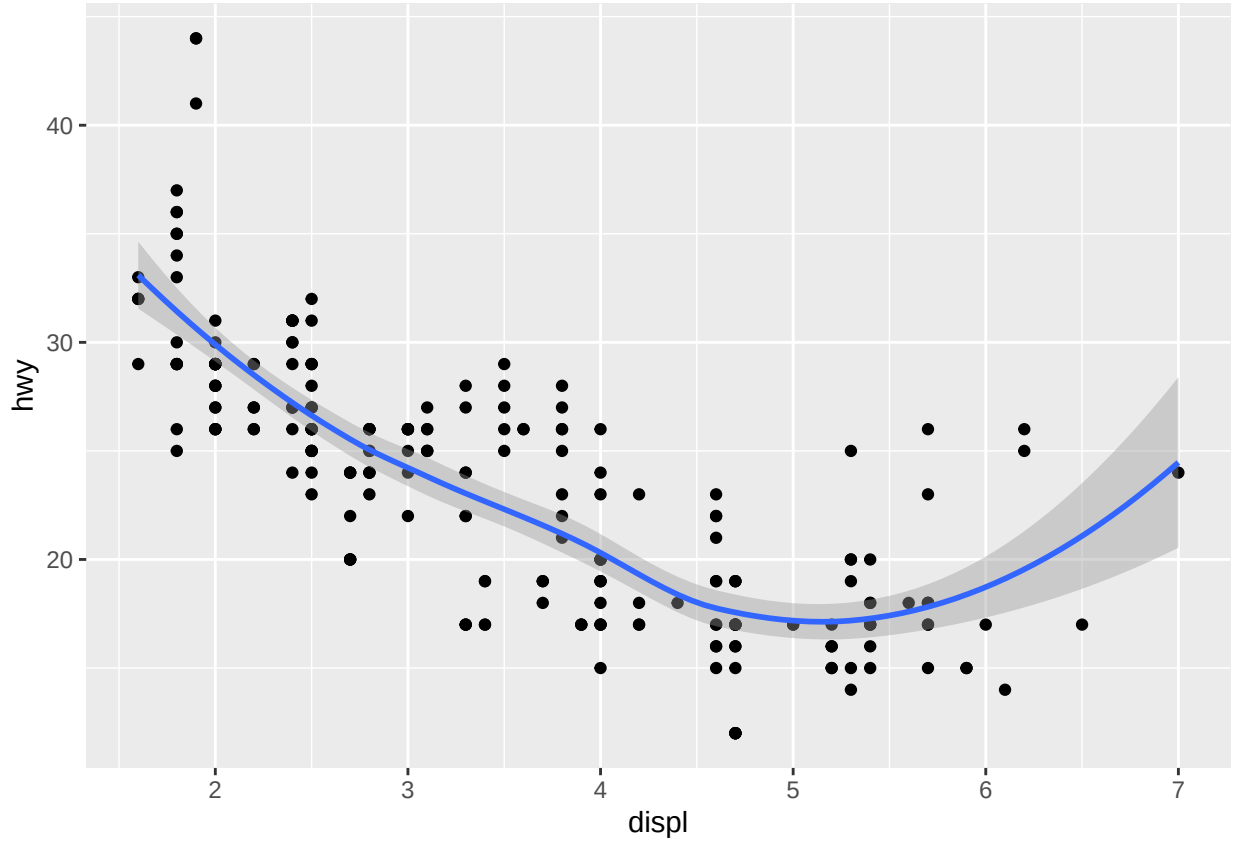

```
# Answer: When we use facet_grid(drv ~ cyl) with the mpg dataset, we creating a matrix
# of plots where the rows represent the unique values of the drv variable
# and the columns represent the unique values of the cyl variable.
# The empty cells in this facet plot represent combinations of drv and cyl for which
# there are no observations in the mpg dataset.

# 3. Take the following faceted plot:

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




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

# Answer:
# Advantages: It breaks the data into multiple smaller plots based
# on a categorical variable this allows us to easily compare patterns
# within each subset of data without overlap. In colors, it is difficult to distinguish properly.
# Also, with a large number of categories, color can become hard to read.
# Faceting distinguishes between groups better.
# Disadvantages: Consumes more space, Harder to compare points between different categories.

################################################################################
# Question 2.4: Geometric Objects
################################################################################

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

# Answer:
# line chart: geom_line()
# boxplot: geom_boxplot()
# histogram: geom_histogram()
# area chart: geom_area()
```

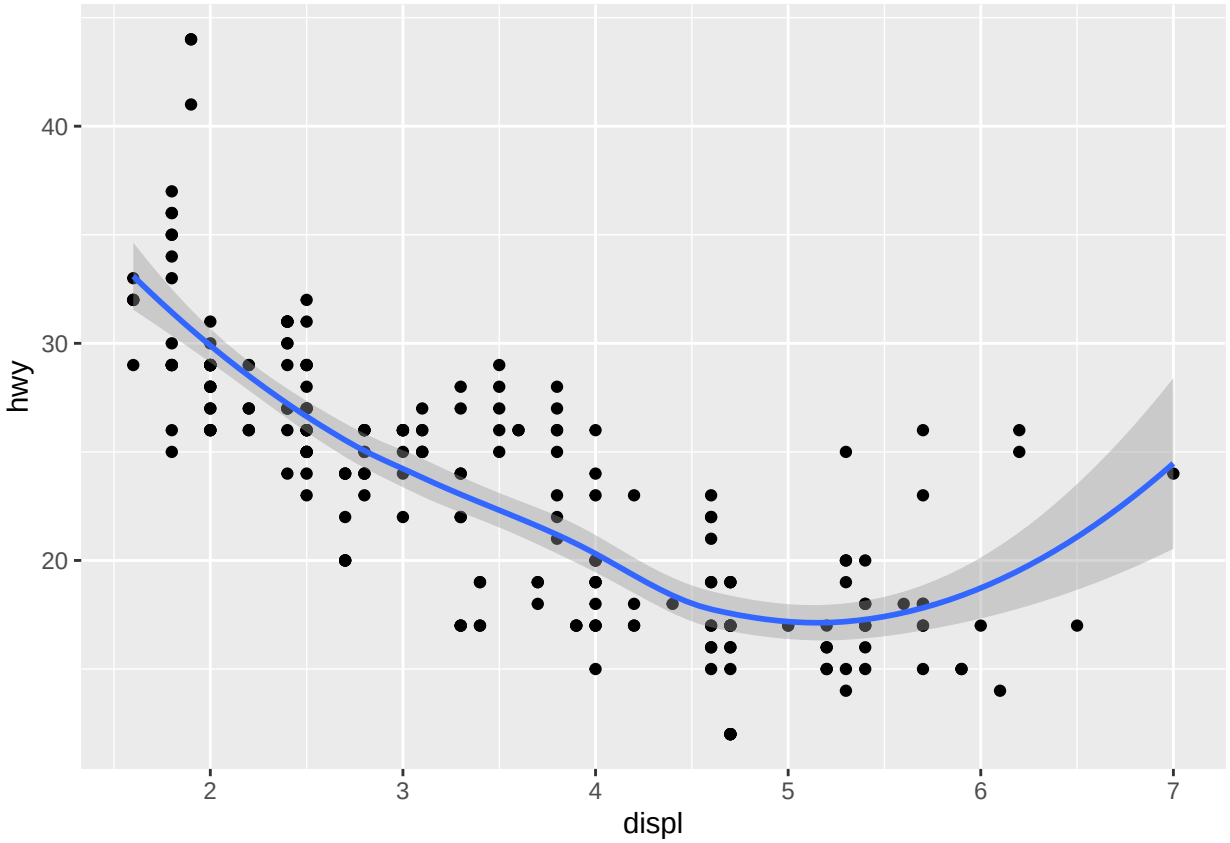
```
# 2. Will these two graphs look different? Why/why not?
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

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




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

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



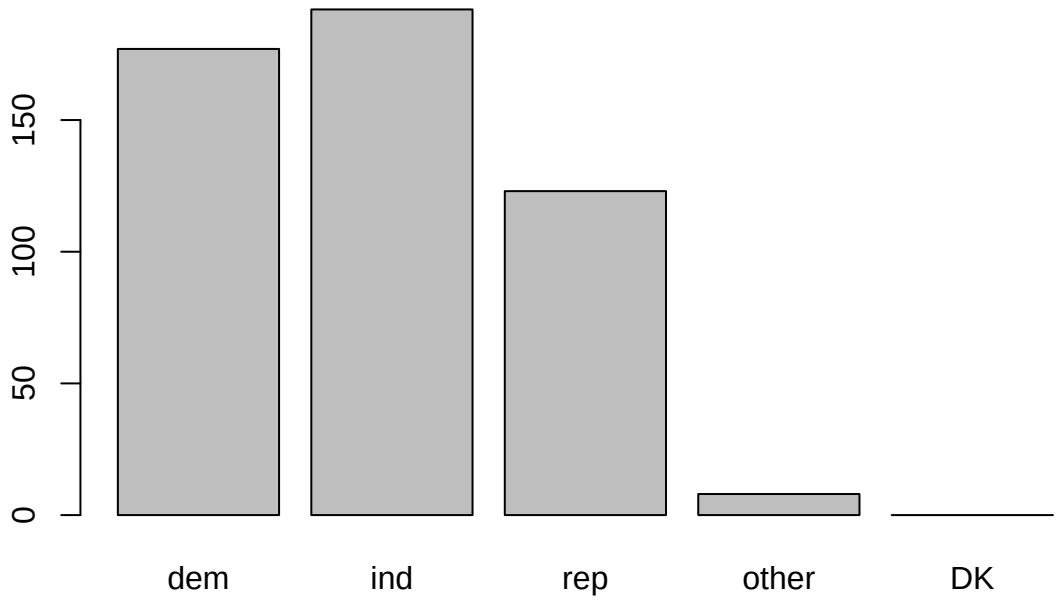

```
# Answer: Both the graphs will do the same thing and will look the same.
# The only difference is in the code structure.
# In the first code, aesthetic is defined globally and
# both the geom_point() and geom_smooth()
# will inherit these global settings.
# In the second code, the aesthetic is defined locally
# for both geom_point() and geom_smooth()
# Thus, both these codes give the same output graphs.


################################################################################
# Part 2: Your project
################################################################################

# Question 3

#An interesting data set we came across was the General Social Survey (GSS)
#dataset. It is a high-quality survey which gathers data on American society
#and opinions, and it conducted since 1972. The data set can be accessed in
#R using the library infer. The dataset present in R is a sample of the
#original dataset of 500 entries from the original with a span of years
#1973-2018. It includes demographic markers and some economic variables.
#It contains of 11 variables namely year (year the respondent was surveyed),
#age (age of the respondent at the time of the survey), sex (gender of the
#respondent which is self-identified by them), college
#(whether the respondent has a valid college degree or no),
```

```
#partyid (respondents political party affiliation),
#hompop (number of people in the respondents house),
#hours (number of hours the respondent works while he was being surveyed),
#income (total family income of the respondent), class
#(subjective socioeconomic class identification), finrela
#(opinion of family income) and weight (survey weight). The data set consists
#of just 500 rows of data.
#We can use this dataset to generate the average number of people living
#in each household in a certain year. We can chart out the slope of the '
#increase or the decrease in the number of people in each household.
#We can determine how much an average worker works each week and
#the average salary they get for each hour. We can group the previous
#result based on the class of the individual. We can determine which political
#party is likely to succeed in that area during a specific year. The literacy
#rate of the area can be determined on whether a person has achieved a degree
#or not. Many such inferences can be made through this dataset by various
#statistical methods. We can group the dataset based upon the years by
#splitting the dataset and can determine many inferences according to the year.
#Same can be done by splitting the dataset by class or political party
#preferences.

library(infer)
```

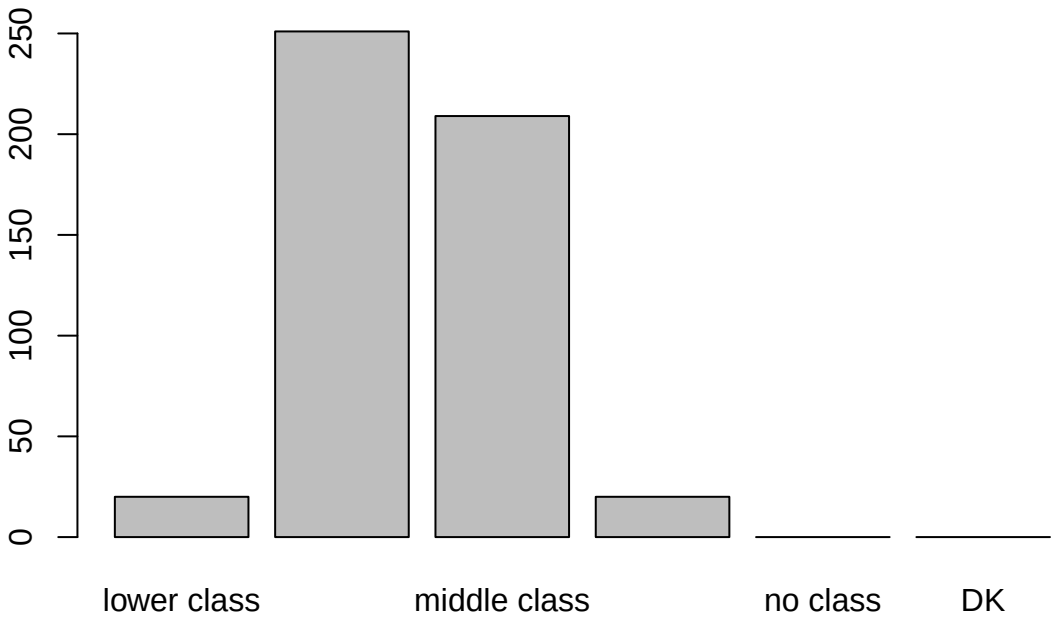
```
## Warning: package 'infer' was built under R version 4.2.2
```

```
library(ggplot2)
data=gss
plot(data['sex'])
```


```
#plot(data)
plot(data['partyid'])
```




```
plot(data['class'])
```


0
50
100
150
200
250
lower class
middle class
no class
DK