My first assignment

Maitri Shah

5/15/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     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()
```

```
tinytex::install_tinytex()
```

```
## The directory /usr/local/bin is not writable. I recommend that you make it writable. See https://gitl
```

This is my first assignment for DA 5020

QUESTION 1

1) Difference between .R and .rmd

R script or .R is a text file that contains the commands that we enter in the command line of the R console. It does not display outputs. Moreover to create a .R file, we input our codes in R Script.

A .rmd file is created when we input our codes in R Markdown. It displays plain text, commands as well as output in a single file and can be knitted to HTML/PDF/Word. The codes are typed in chunks.

2) str() vs summary()

Both the commands are used after we analyze data to understand it in a better way. The str() command is more concise and is used to see the structure of the data. It shows us the number of rows, columns, values of columns and their respective heads.

summary() gives a broader understanding of the data. It gives a statistical summary of the data and may show you minimum, maximum, mean and median

For e.g

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
#Gives the rows, columns, values etc. from the data set

summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
#Gives min, max, quartiles, mean, median etc.
```

QUESTION 2

```
nrow(mtcars)
```

```
## [1] 32
```

```r
ncol(mtcars)
```

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

```r
#OR
dim(mtcars)
```

```
## [1] 32 11
```

QUESTION 3

```r
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

```r
tail(mtcars, 5)
```

```
##                 mpg cyl  disp  hp drat    wt qsec vs am gear carb
## Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
## Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
## Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
## Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
## Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
```

QUESTION 4

```r
#Categorical column names
CategoricalVariables <- "cyl, gear"
CategoricalVariables
```

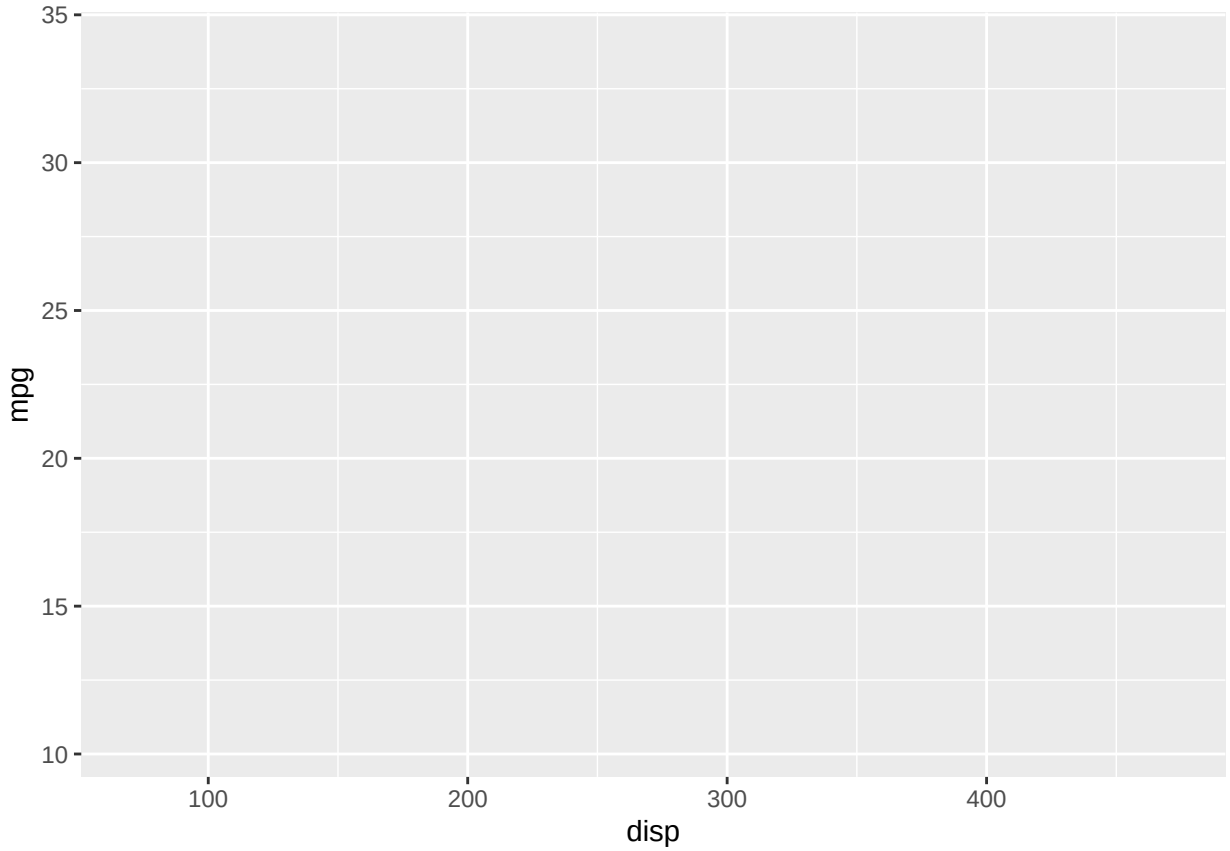
```
## [1] "cyl, gear"
```

```r
#Continuous column names
ContinuousVariables <- "mpg, displ"
ContinuousVariables
```

```
## [1] "mpg, displ"
```

QUESTION 5

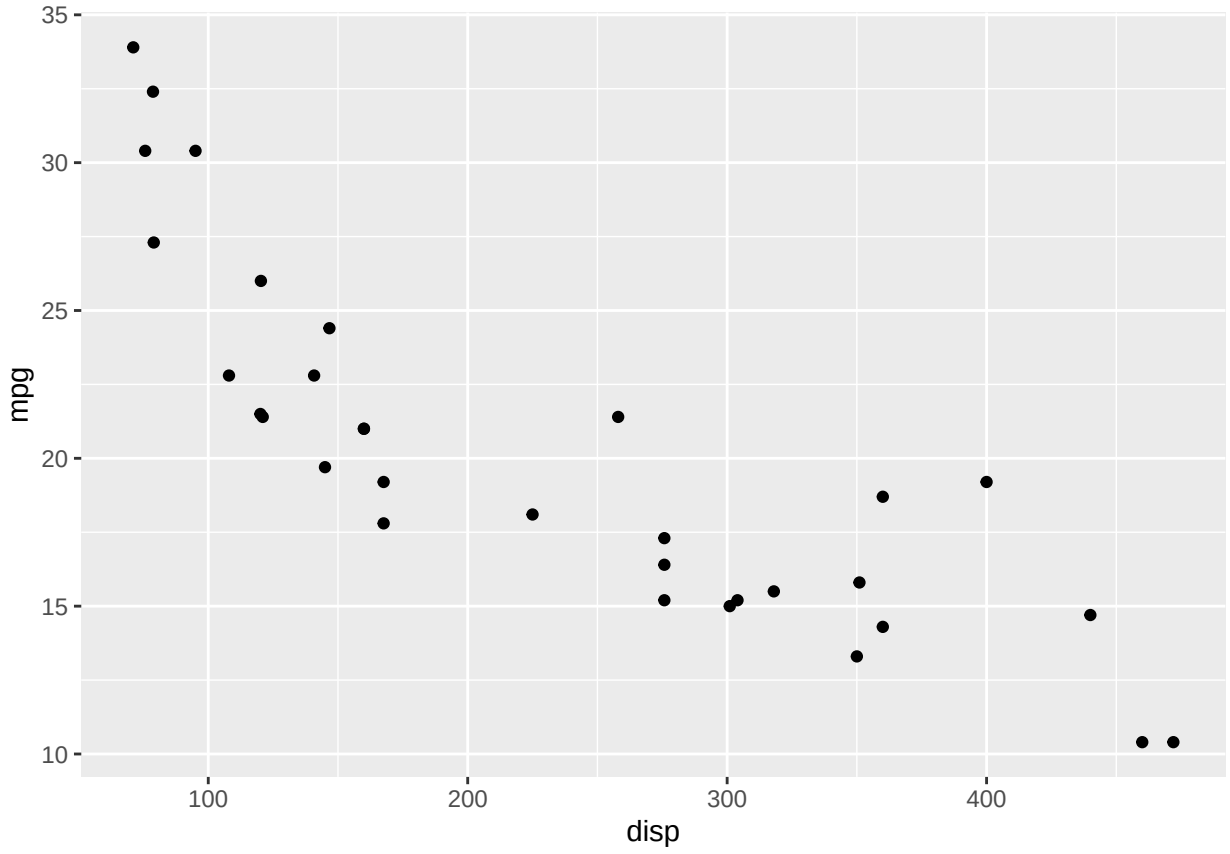
```r
ggplot(mtcars, aes(x= disp, y=mpg))
```



This code displays a blank plot with disp on the x-axis and mpg on the y-axis. Since there are no geoms added in the code, we cannot see a relationship between the two. It can be modified by using geom function

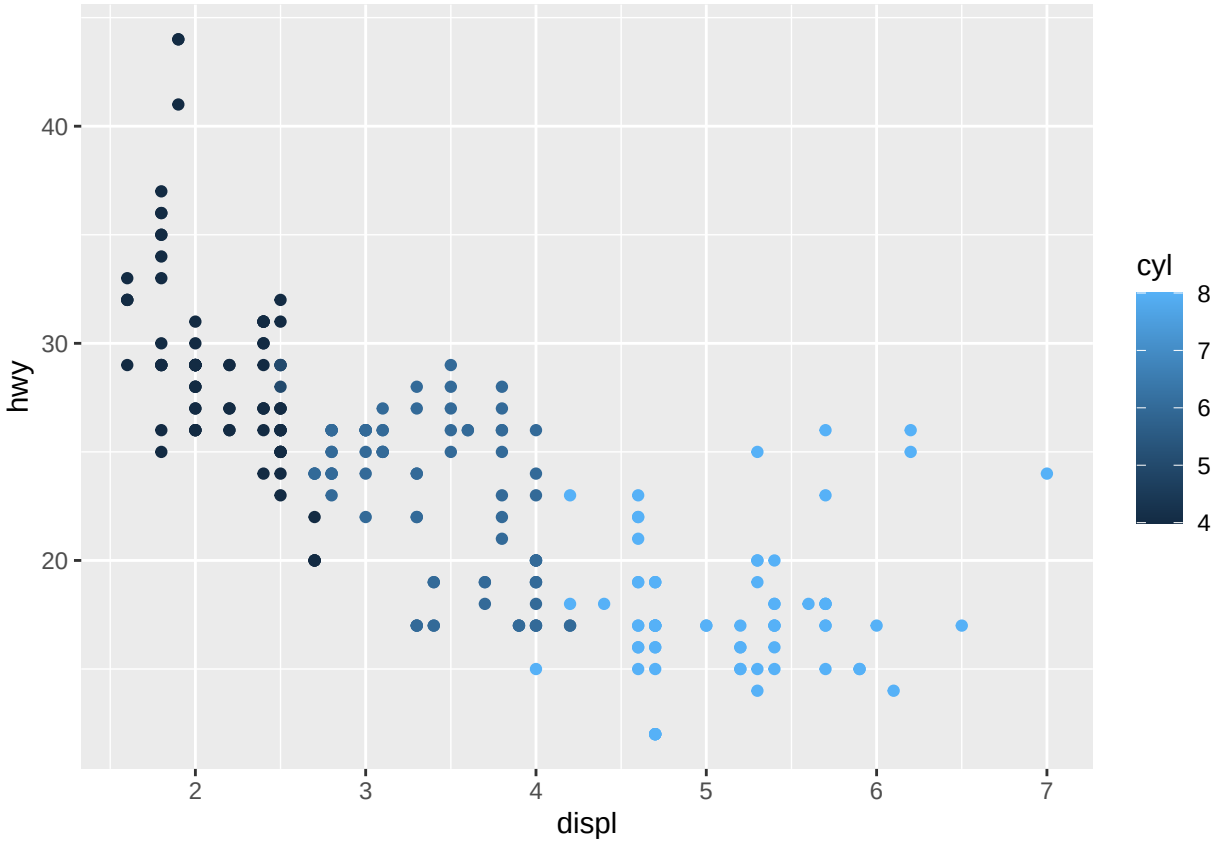
The modified code is as follows:

```
ggplot(mtcars, aes(x= disp, y=mpg))+
  geom_point()
```



QUESTION 6

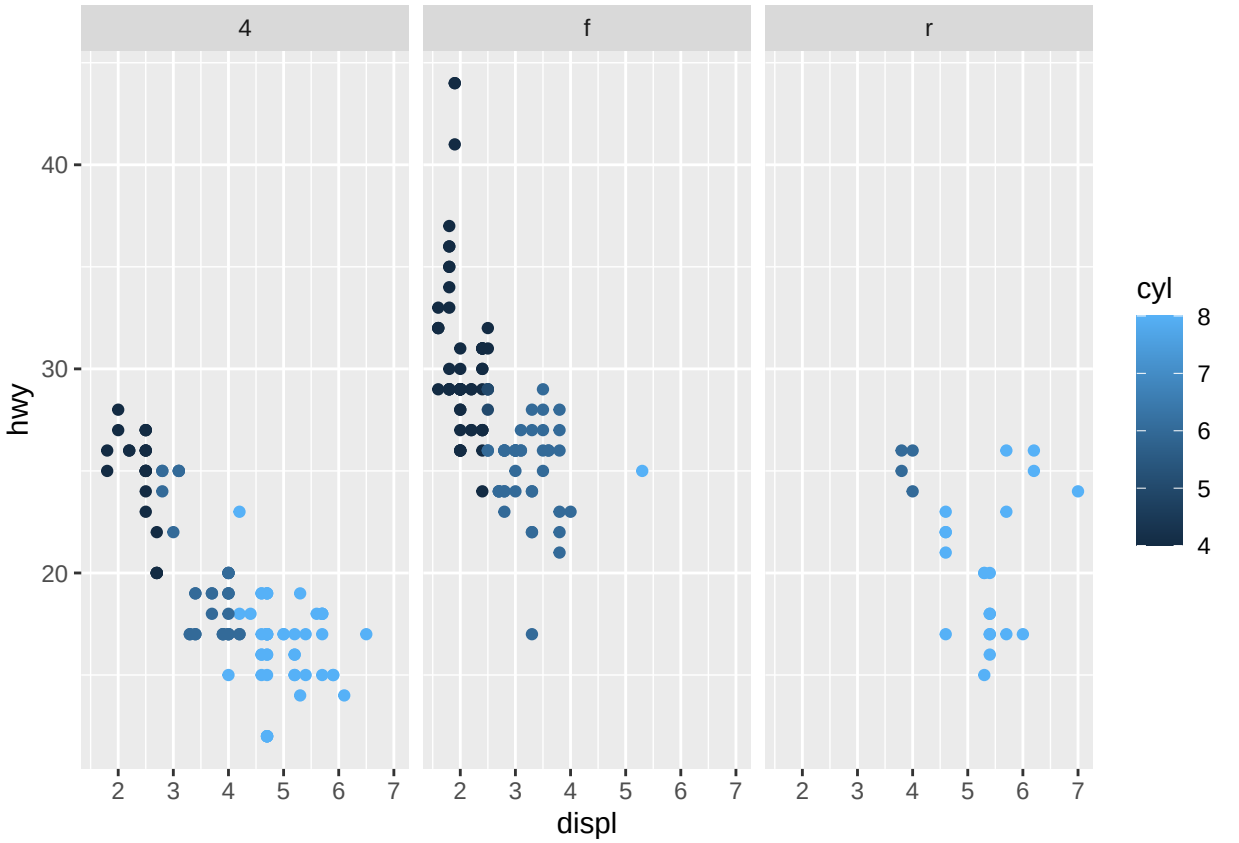
```
ggplot(data= mpg)+
  geom_point(mapping= aes(x= displ, y= hwy, color= cyl))
```



Explanation: This plot shows that cars with more number of cylinders (cyl) also have bigger engines sizes in litres (displ) and low fuel efficiency (hwy) thereby using more fuel. As the cylinder number goes down, the fuel efficiency increases and engine size decreases

QUESTION 7

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



Explanation

Rear wheel drive vehicles with higher number of cylinders have a better fuel efficiency inspite of having larger engines sizes compared to front and 4wd wheel drive ones. The 4wd wheel drive vehicles have varying number of cylinders but have a low to medium fuel efficiency and the front wheel drive ones, barring the two outliers have mid size engines and medium to high fuel efficiency.