Homework 1

Prepare your answers as a **single PDF file**.

**Group work**: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.

**Due**: check on Canvas.

1. Use the in-built dataset, mtcars, for this problem. Write code to (**for each question, give (1) the code and (2) the output)**:

1. Get number of rows (Hint: nrow)

> nrow(mtcars)

[1] 32

1. Get number of columns (Hint: ncol)

> ncol(mtcars)

[1] 11

1. Get datatype of the disp column (Hint: class)

> class(mtcars$disp)

[1] "numeric"

1. What is the unit of the disp column? (Hint: see help)

help(mtcars)

|  |  |  |
| --- | --- | --- |
| [, 3] | disp | Displacement (cu.in.) |

1. Show first 10 rows

> mtcars[1:10,]

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160.0 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108.0 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258.0 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360.0 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225.0 105 2.76 3.460 20.22 1 0 3 1

Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 0 0 3 4

Merc 240D 24.4 4 146.7 62 3.69 3.190 20.00 1 0 4 2

Merc 230 22.8 4 140.8 95 3.92 3.150 22.90 1 0 4 2

Merc 280 19.2 6 167.6 123 3.92 3.440 18.30 1 0 4 4

1. Show every other row (i.e., 1st, 3rd, 5th, …) (Hint: seq)

> mtcars[seq(1,32,by=2),]

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 0 1 4 4

Datsun 710 22.8 4 108.0 93 3.85 2.320 18.61 1 1 4 1

Hornet Sportabout 18.7 8 360.0 175 3.15 3.440 17.02 0 0 3 2

Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 0 0 3 4

Merc 230 22.8 4 140.8 95 3.92 3.150 22.90 1 0 4 2

Merc 280C 17.8 6 167.6 123 3.92 3.440 18.90 1 0 4 4

Merc 450SL 17.3 8 275.8 180 3.07 3.730 17.60 0 0 3 3

Cadillac Fleetwood 10.4 8 472.0 205 2.93 5.250 17.98 0 0 3 4

Chrysler Imperial 14.7 8 440.0 230 3.23 5.345 17.42 0 0 3 4

Honda Civic 30.4 4 75.7 52 4.93 1.615 18.52 1 1 4 2

Toyota Corona 21.5 4 120.1 97 3.70 2.465 20.01 1 0 3 1

AMC Javelin 15.2 8 304.0 150 3.15 3.435 17.30 0 0 3 2

Pontiac Firebird 19.2 8 400.0 175 3.08 3.845 17.05 0 0 3 2

Porsche 914-2 26.0 4 120.3 91 4.43 2.140 16.70 0 1 5 2

Ford Pantera L 15.8 8 351.0 264 4.22 3.170 14.50 0 1 5 4

Maserati Bora 15.0 8 301.0 335 3.54 3.570 14.60 0 1 5 8

1. What is the mean mpg value

> mean(mtcars$mpg)

[1] 20.09062

1. Show all rows where the number of cylinders is 6

> which(mtcars$cyl==6,)

[1] 1 2 4 6 10 11 30

1. Show all rows where its mpg is lower than the mean mpg value

mn <-mean(mtcars$mpg)

> mn

[1] 20.09062

> which(mtcars$mpg < mn)

[1] 5 6 7 10 11 12 13 14 15 16 17 22 23 24 25 29 30 31

1. What is the horsepower of the car with the highest mpg (code should show only the horsepower value)?

> max(mtcars$mpg)

[1] 33.9

mtcars$hp[which.max(mtcars$mpg)]

[1] 65

2. Consider the answer posted to Quora.com to “[Why is R great for Data Science?](https://www.quora.com/Why-is-R-great-for-Data-Science/answer/Dan-Reznik). Answer **one** of the following questions.

The author lists 5 parts of the R ecosystem, the 5th being “community”. Write 4-5 sentences about any one online community where members discuss R. (Include the URL, how active is the community, what types of people post here, how “friendly” it is to newcomers, etc.)

According to “Why is R great for Data Science”, one of the main 5 benefits of using R is its’ pleasurable/social element of R users. By examining some communities, which discuss R language I found a Stack overflow “Questions tagged [r]” and realized that community is friendly in deed. There are lots of questions asked and answered in kindly manner, people try to help each other and make improvements. There are also plenty of tutorials for humans to get familiar with new features of R as well as plenty of code snippets on how to do the code correctly or/and fix your bugs.

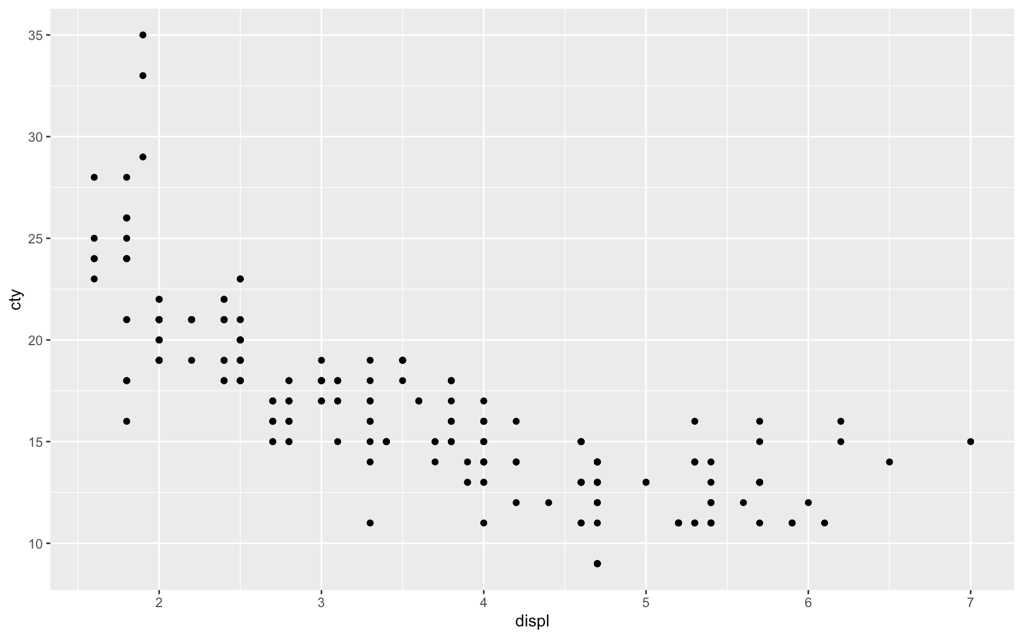
https://stackoverflow.com/questions/tagged/r?tab=Newest

**OR** (if you know Python)

The author says “Note that in python, data frame manipulation will require numpy and pandas external packages (and the syntax is more cumbersome)”. Do you agree with this statement? Justify your answer in 4-5 sentences.

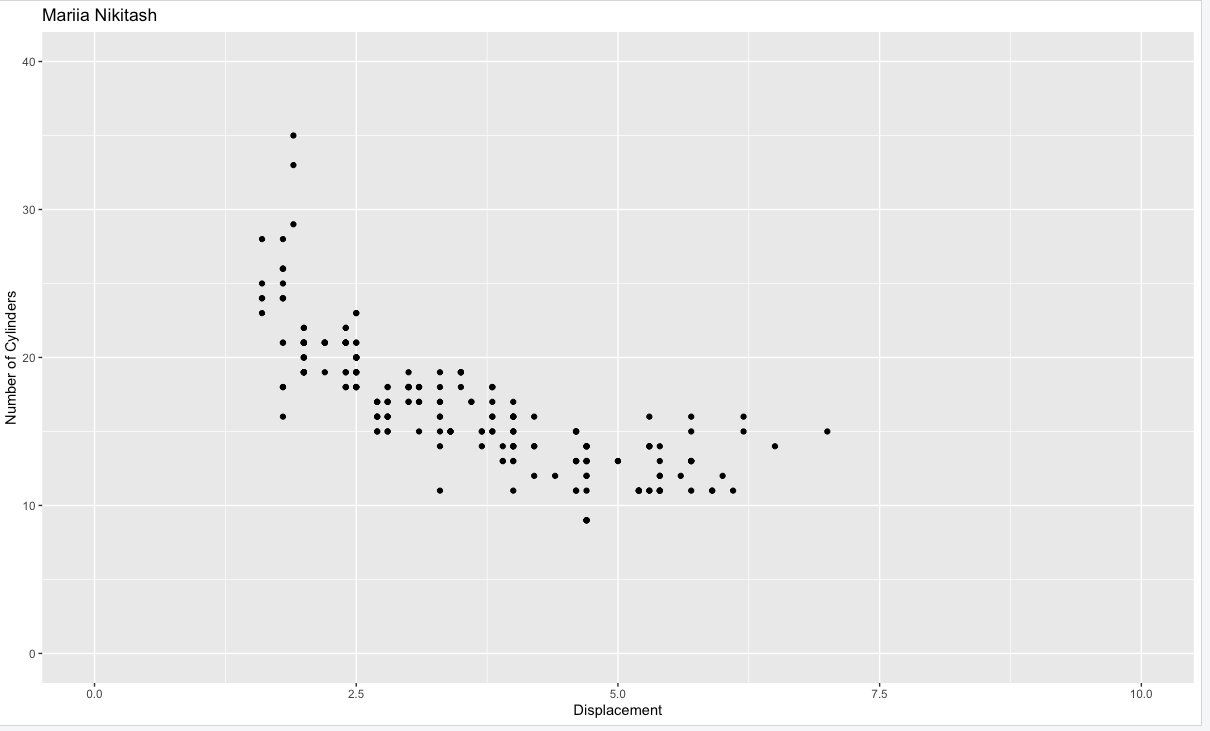
3. Installing ggplot2 also installs some datasets, including the mpg dataset (see help(mpg) for a description of the data). Generate the following graphs from the mpg dataset. All plots should use **ggplot**. Include **both** the R code and paste the plot as an image.

1. Plot a scatterplot of variables displ and cty.

ggplot(mpg, aes(displ,cty)) + geom\_point()**

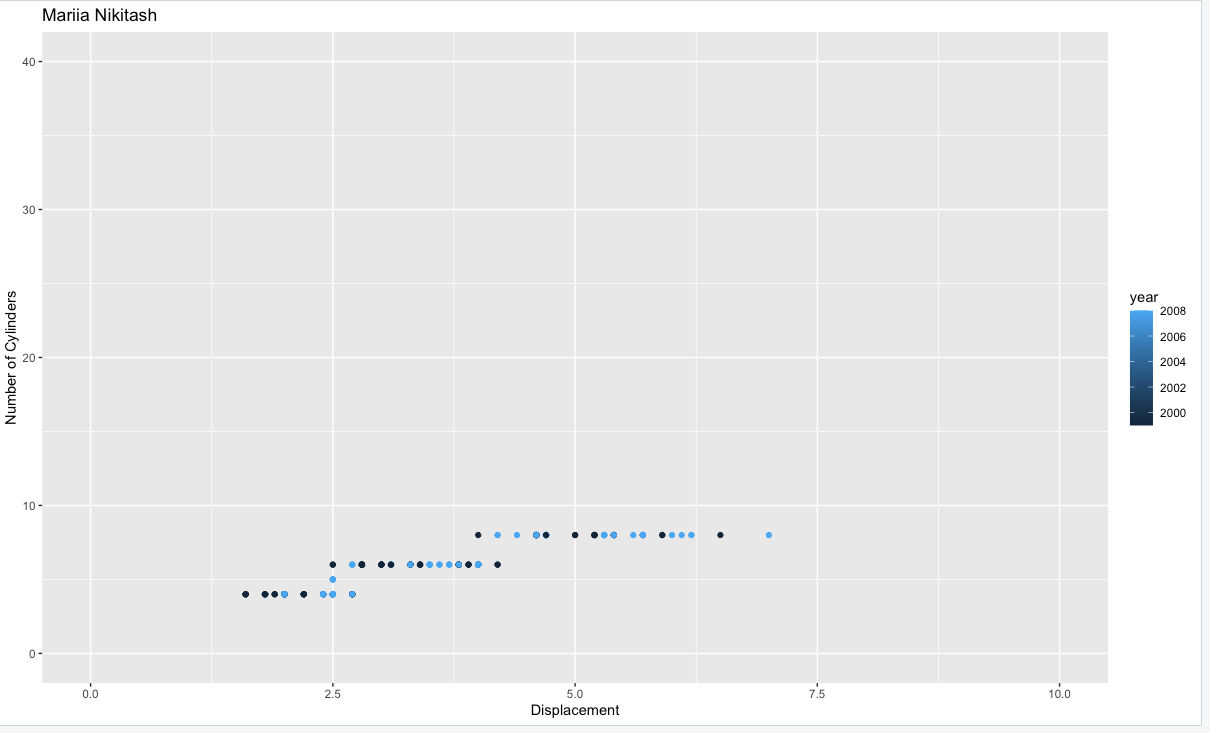
1. Redraw the previous scatterplot but also add all these:
   * more descriptive x and y-axis labels,
   * a title that should be the names of all group members, and
   * set cty limits to (0,40) and displ limits to (0,10).

ggplot(mpg)+geom\_point(mapping=aes(displ,cty))+labs(x="Displacement",y="Number of Cylinders",title="Mariia Nikitash")+xlim(0,10)+ylim(0,40)



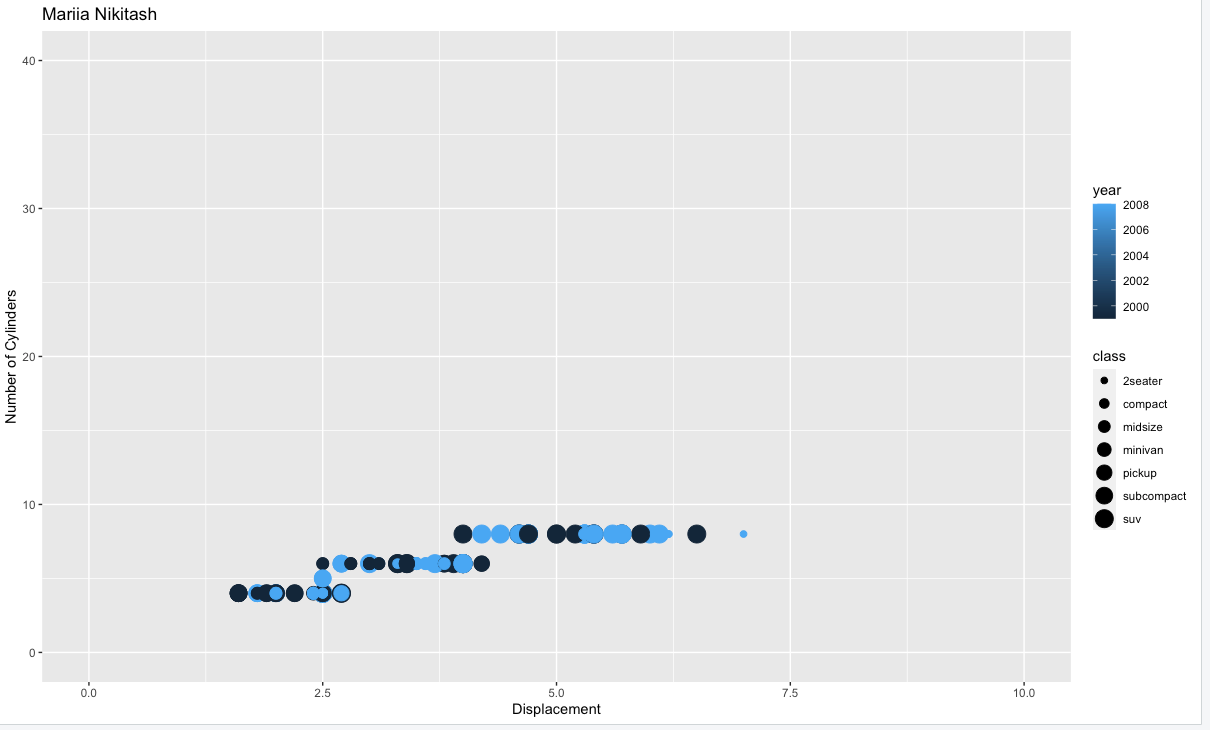
1. Plot a scatterplot of variables displ and cty. Show variable year also.

ggplot(mpg)+geom\_point(mapping=aes(x = displ, y = cyl, color=year))+labs(x="Displacement",y="Number of Cylinders",color="year",title="Mariia Nikitash")+xlim(0,10)+ylim(0,40)



1. Plot a scatterplot of variables displ and cty. Show variables year and class also.
   * Hint: There are different ways of doing this using the multiple “aesthetics” of geom\_point

ggplot(mpg)+geom\_point(mapping=aes(x = displ, y = cyl, color=year, size=class))+labs(x="Displacement",y="Number of Cylinders",color="year",title="Mariia Nikitash")+xlim(0,10)+ylim(0,40)

**