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. Please upload a "group picture" of your group. You can be creative - an actual group photo, a screen capture of a zoom meeting, putting your profile pictures/avatars into one, ... Please put your names on the picture too. (You must answer this question even if you work by yourself).

2. Use the in-built dataset, `airquality`, for this problem. Write code to:
   a. Get number of rows (Hint: nrow)
      i.
```
> nrow(airquality)
[1] 153
```
   b. Get number of columns (Hint: ncol)
      i.
```
> ncol(airquality)
[1] 6
```
   c. Show first 10 rows

i.

```
> head(airquality, 10)
   Ozone Solar.R Wind Temp Month Day
1     41     190  7.4   67     5   1
2     36     118  8.0   72     5   2
3     12     149 12.6   74     5   3
4     18     313 11.5   62     5   4
5     NA      NA 14.3   56     5   5
6     28      NA 14.9   66     5   6
7     23     299  8.6   65     5   7
8     19      99 13.8   59     5   8
9      8      19 20.1   61     5   9
10    NA     194  8.6   69     5  10
```

d. Show the mean Wind

i.

```
> mean(airquality$Wind)
[1] 9.957516
```

e. Show all rows where Month == 6

i.

```
> airquality[airquality$Month==6, ]
   Ozone Solar.R Wind Temp Month Day
32    NA     286  8.6   78     6   1
33    NA     287  9.7   74     6   2
34    NA     242 16.1   67     6   3
35    NA     186  9.2   84     6   4
36    NA     220  8.6   85     6   5
37    NA     264 14.3   79     6   6
38    29     127  9.7   82     6   7
39    NA     273  6.9   87     6   8
40    71     291 13.8   90     6   9
41    39     323 11.5   87     6  10
42    NA     259 10.9   93     6  11
43    NA     250  9.2   92     6  12
44    23     148  8.0   82     6  13
45    NA     332 13.8   80     6  14
46    NA     322 11.5   79     6  15
47    21     191 14.9   77     6  16
48    37     284 20.7   72     6  17
49    20      37  9.2   65     6  18
50    12     120 11.5   73     6  19
51    13     137 10.3   76     6  20
52    NA     150  6.3   77     6  21
53    NA      59  1.7   76     6  22
54    NA      91  4.6   76     6  23
55    NA     250  6.3   76     6  24
56    NA     135  8.0   75     6  25
57    NA     127  8.0   78     6  26
58    NA      47 10.3   73     6  27
59    NA      98 11.5   80     6  28
60    NA      31 14.9   77     6  29
61    NA     138  8.0   83     6  30
```

f. What are the row indexes of the rows where Month==6? (Hint: which)

i.

```
> which(airquality$Month==6)
 [1] 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61
```

g. Show all rows where Month == 6 and Day < 10

```
> airquality[airquality$Month==6 & airquality$Day<10, ]
   Ozone Solar.R Wind Temp Month Day
32    NA     286  8.6   78     6   1
33    NA     287  9.7   74     6   2
34    NA     242 16.1   67     6   3
35    NA     186  9.2   84     6   4
36    NA     220  8.6   85     6   5
37    NA     264 14.3   79     6   6
38    29     127  9.7   82     6   7
39    NA     273  6.9   87     6   8
40    71     291 13.8   90     6   9
```

i.

h. Get the largest value of Wind

```
> max(airquality$Wind)
[1] 20.7
```

i.

i. On what Month and Day was this largest value of Wind observed?

```
> airquality[max(airquality$Wind), 5:6]
   Month Day
20     5  20
```

i.

For each question, give (1) the code and (2) the output.

3. Consider the answer posted to Quora.com to "Why is R great for Data Science?. Answer one of the following questions.

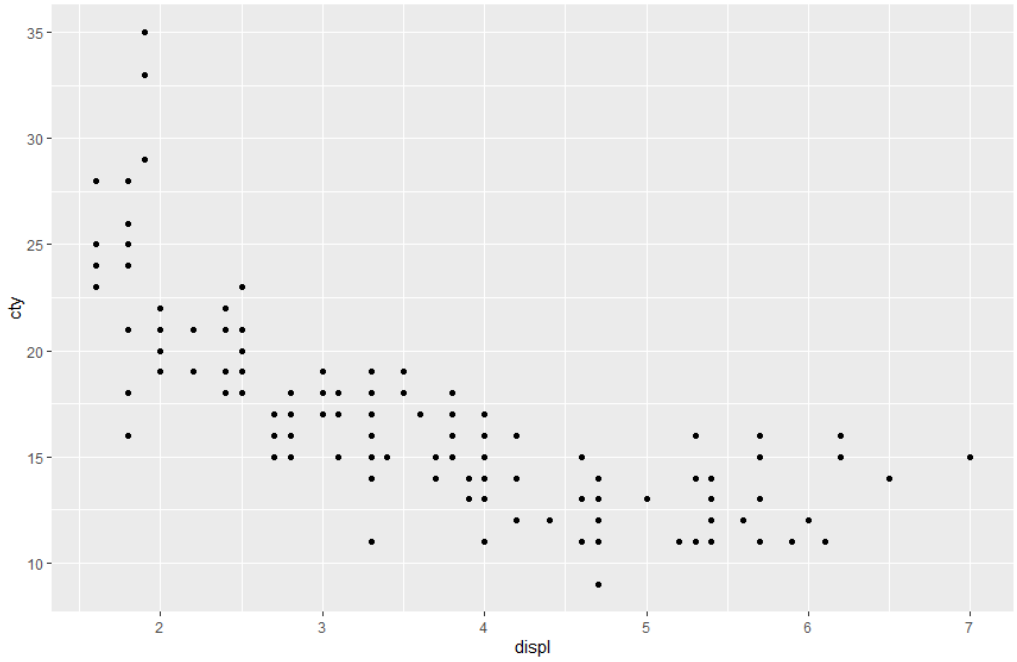
The author lists 5 parts of the R ecosystem, the 5th being "community". Write 4-5 sentences about any one online community where members discuss R. (Include the URL, how active is the community, what types of people post here, how "friendly" it is to newcomers, etc.)

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.

In response to the Python question:
I disagree that the syntax of numpy and pandas is "more cumbersome". This could be due to the fact that Python was the first language I learned and I became comfortable with using pandas at an early point. Accessing and replacing items in a dataframe using pandas has always seemed intuitive to me. That being said, I am a new user of R so I could feel differently after I gain a better understanding of manipulating data frames.

4. 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.

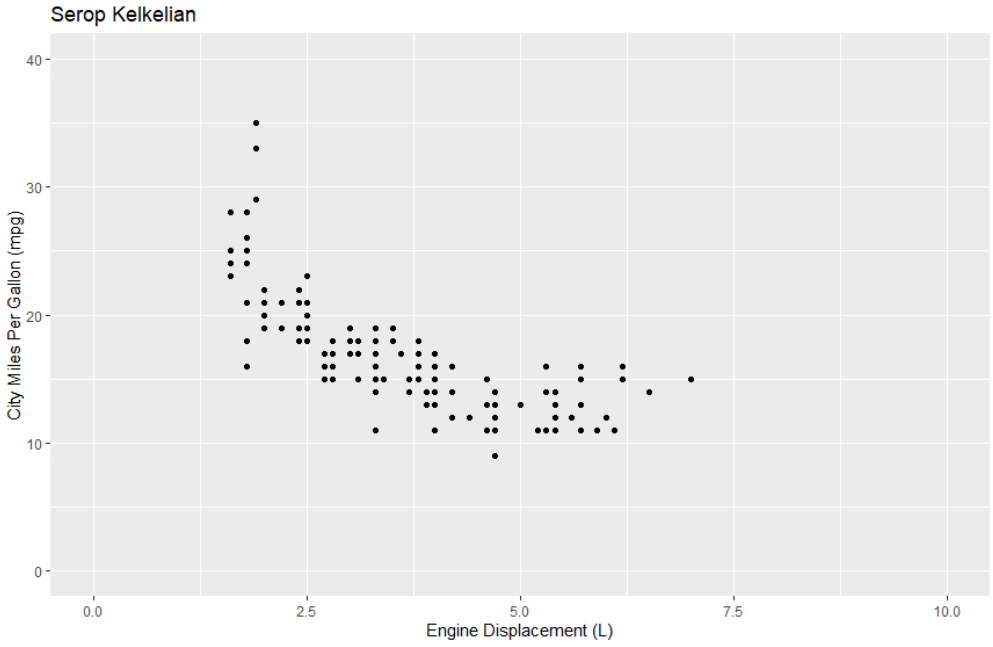
a. Plot a scatterplot of variables `displ` and `cty`.




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

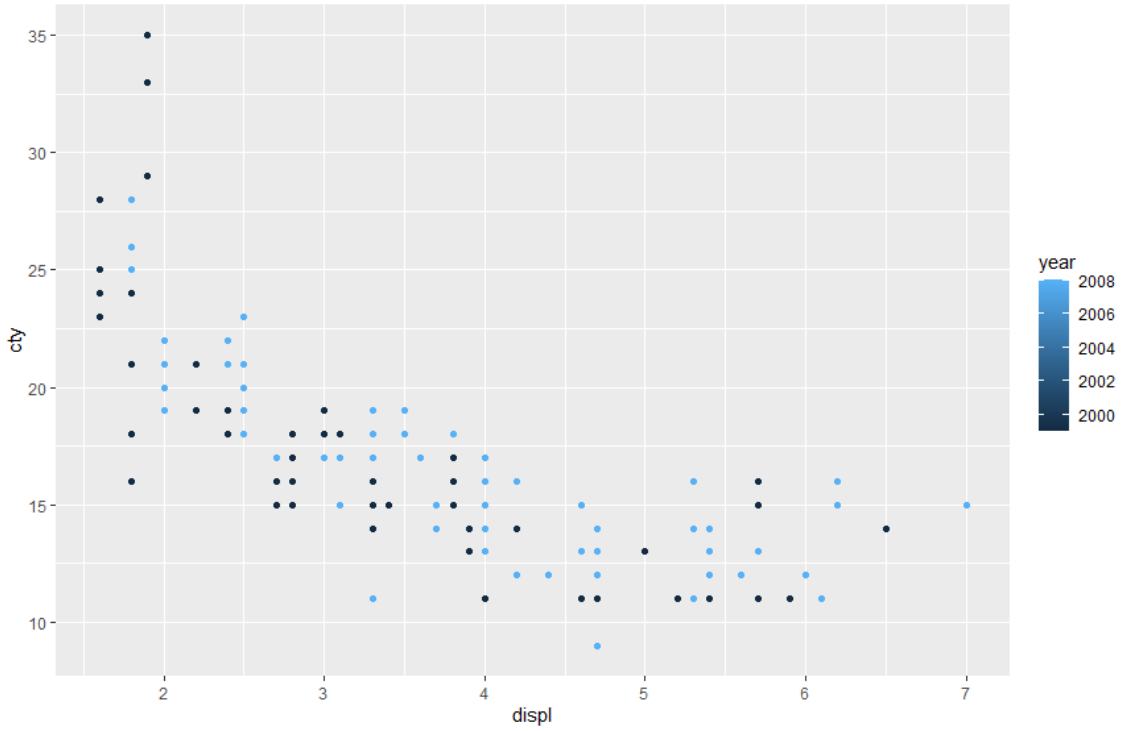
b. Redraw the previous scatterplot but also add all these:

- 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(data=mpg) + geom_point(mapping=aes(x = displ, y = cty)) + labs(title = "Serop Kelkelian", y = "City Miles Per Gallon (mpg)", x = "Engine Displacement (L)") + xlim(0, 10) + ylim(0, 40)`

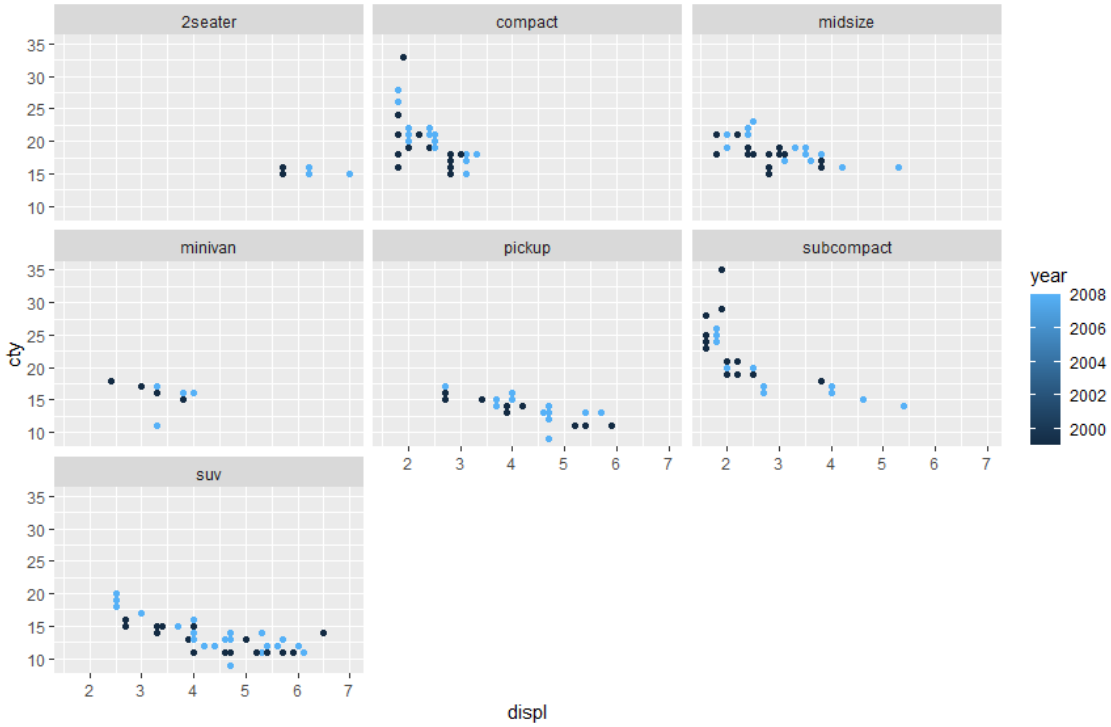
c. Plot a scatterplot of variables `displ` and `cty`. Show variable `year` also.



- `ggplot(data=mpg) + geom_point(mapping=aes(x = displ, y = cty, color = year))`

d. 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(data=mpg) + geom_point(mapping=aes(x = displ, y = cty, color = year)) + facet_wrap(~class)`