University of Tehran

School of Electrical and Computer Engineering

Statistical Inference

Instructor: Dr. Hossein Vahabie

Assignment 1

Data Visualization in R

Teaching Assistants:

kasramoumeni@gmail.com (® Problems)

hamedgholami@ut.ac.ir

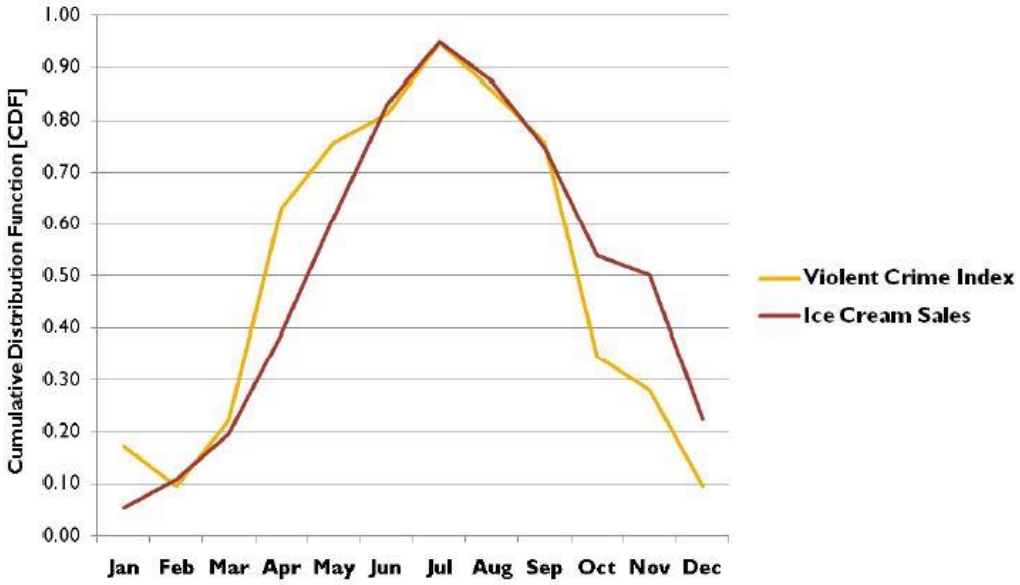
Winter 2022

Contents

Question 1

In each of the following parts, is there a potential confounding variable? If the answer is yes, find the confounding factor and explain.

a) In red, the plot shows the amount of ice cream sold in each month. In yellow, the plot shows the violent crime rate in each month. We see a strong relationship between these signals.

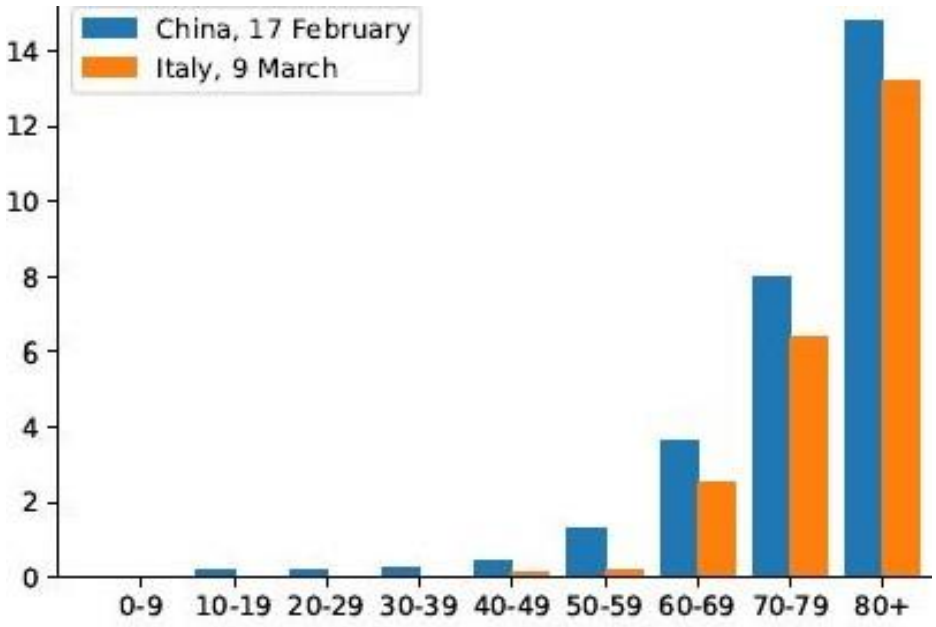


b) It is known that birth rates and the presence of storks are highly positively correlated, That is, as birth rates rise, so does the presence of storks. This has given rise to the urban myth that strokes deliver babies.

Question 2

The following graph shows COVID-19's case fatality rates (CFRs) in Italy and China by age groups.

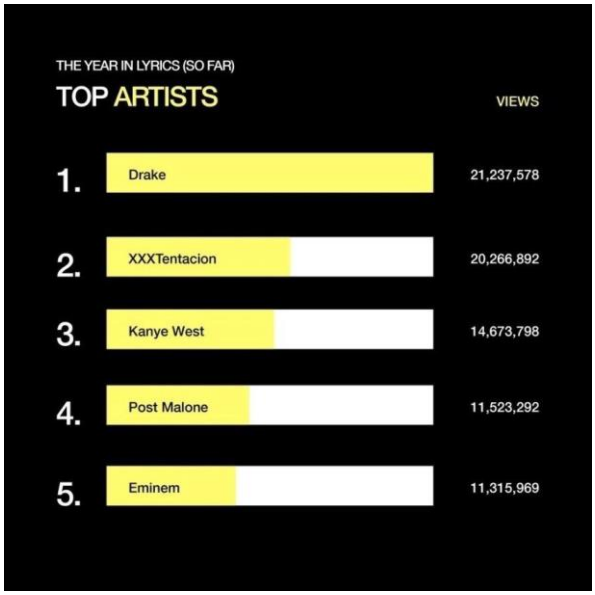
Ali saw this graph and thought with himself: in total, Italy should have a lower CFR than China because they have lower CFR in each of the age groups. Do you agree or disagree with his statement? Explain your reasoning.



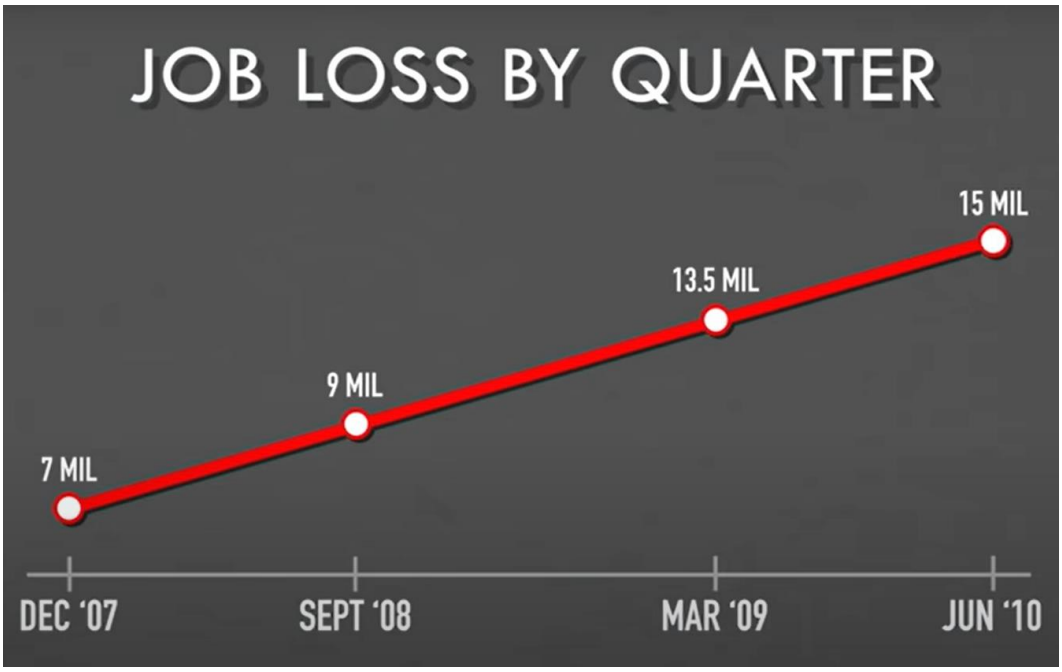
Question 3

For each of the following graphs explain what are the most concerning problems of that graph in regards to conveying the truth (if any)?

a) The following graph is from a site that ranks the singers by the number of times their songs have been played.

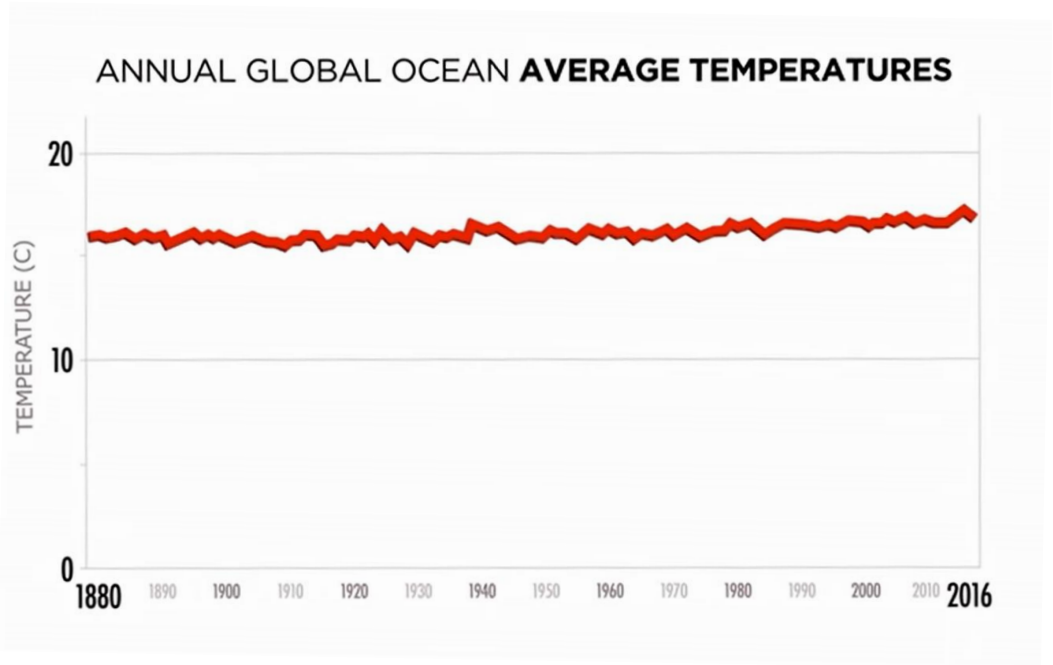


b) The following graph shows the rise in American unemployment from 2008 to 2010. This graph was originally shown by a major press channel.



c) Following is a visualization used by Tim Cook for iPhone sales.

d) The following graph shows the global ocean average temperature each year. Ocean temperature is one of the main symptoms of global warming (more important than air temperature).



Question 4

Explain when a result from an experimental sample can be generalized to the whole population? Also, when can we make a casual conclusion, that is, we can say the explanatory variable "caused" the change on the response variable? You can answer by filling in a table like the following table.

		Generalizability	
		no	yes
Causality	no		
	yes		

Question 5

For each of the following parts, explain (use more than three words) what is the most concerning potential source of sampling bias (If any).

a) Recently, The Heroes show became the first TV show in the IMDB list of top 250 TV shows of all time, then IMDB removed the show from the list.

b) A team of researchers investigated the street fights that led to the death of one of the parties of the conflict and found out that in 90% of the cases, it is the starter of the fight who is eventually killed.

c) A taxi driver believes that wherever the police are present, the traffic is heavier, so the presence of police is one of the factors in the city traffic.
1-

d) A parent is choosing a school for his child and wants to know the class sizes of a specific school, so he goes to the school and randomly asks the kids how many students are in their class, and then he averages those numbers.

R Useful Tutorial Sources

"r-graph-gallery" and "r-statistics" are good sources to get familiar to **GGPLOT2** package. Also, this is a good cheat sheet you can use.

Now, according to your ID, choose one of these datasets and continue:

ID's Last Digit	Question 6's Datasets	Question 7's Datasets
0-3	foods	foods [download here]
4-6	mpg	mtcars
7-9	diamonds (first 500 rows)	diamonds (first 500 rows)

Variables Guide Table

In these questions draw each desired diagram according to variables specified in below tables:

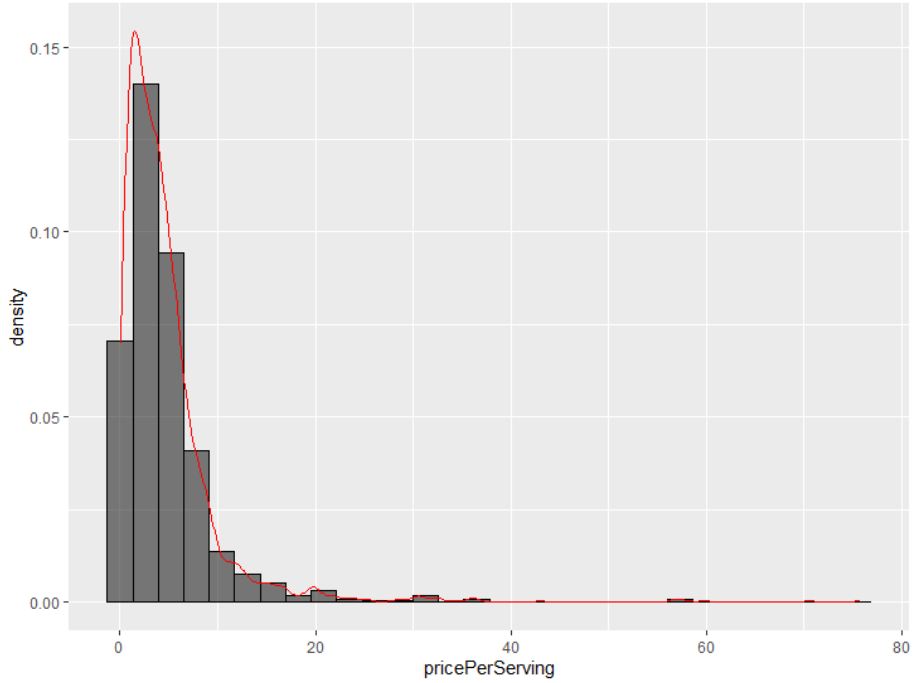
Dataset	Q6-A	Q6-B	Q6-C	Q6-D	Q6-E
Foods	pricePerServing	DishType	healthScore/ dishType	veryHealthy/ DairyFree	healthScore/ readyInMinutes
Mpg	hwy	fl	hwy/class	drv/fl	hwy/cty
Diamonds	depth	cut	Carat/cut	color/cut	z/y

Dataset	Q7-A	Q7-B	Q7-C	Q7-D
Foods	USA_gross_income/ worldwide_gross_income		dishType	healthScore/ dishType
Mtcars	Mpg/Wt		cyl	hwy/class
Diamonds	z/y		cut	Carat/cut

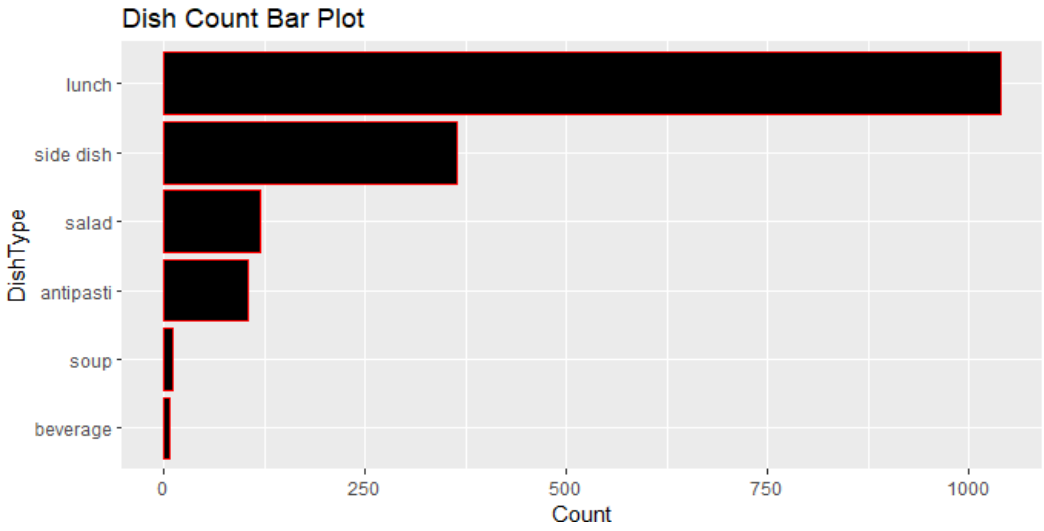
After drawing each plot, you should briefly explain the application of each plot and when it is necessary to use them!

Question 6®

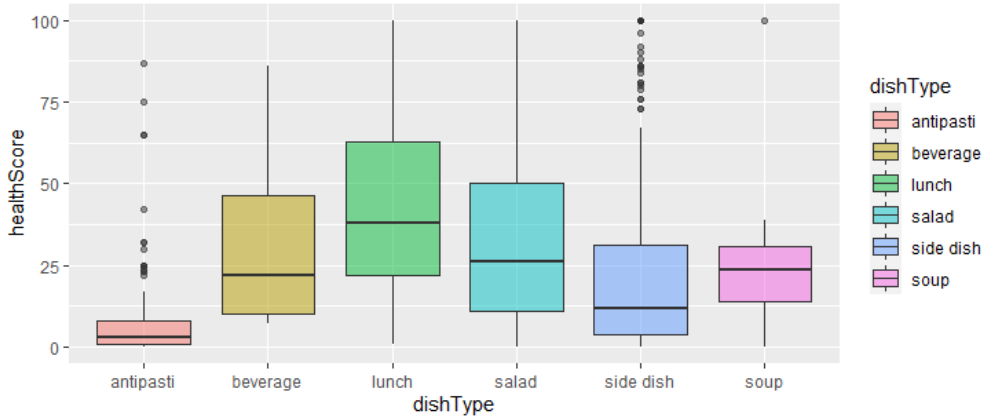
A) Plot the histogram for each variable specified in above table with an appropriate bin size, then overlay that with the density curve. Your diagram must be similar to the following figure:
(with appropriate axis labels, titles and legends)



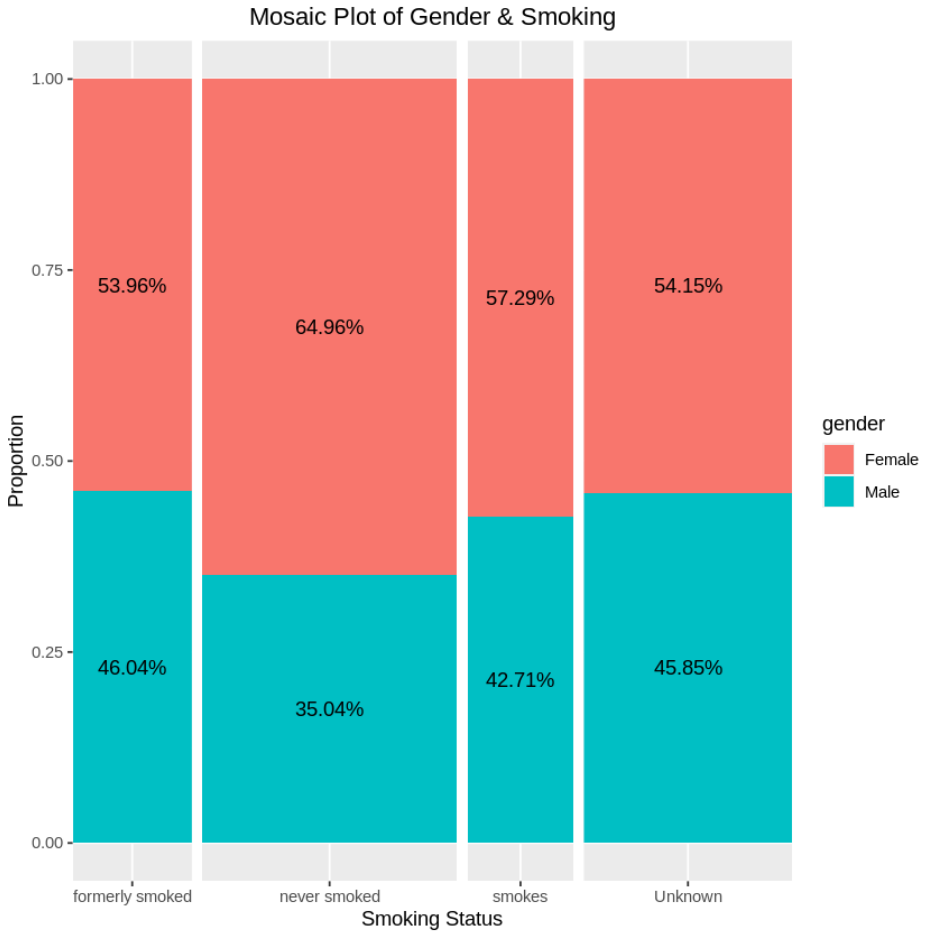
B) Sort the categories by their frequencies, then draw a horizontal barplot to show the result. Your output must be similar to the following image:
(You may draw your plots with any color but labels are necessary!)



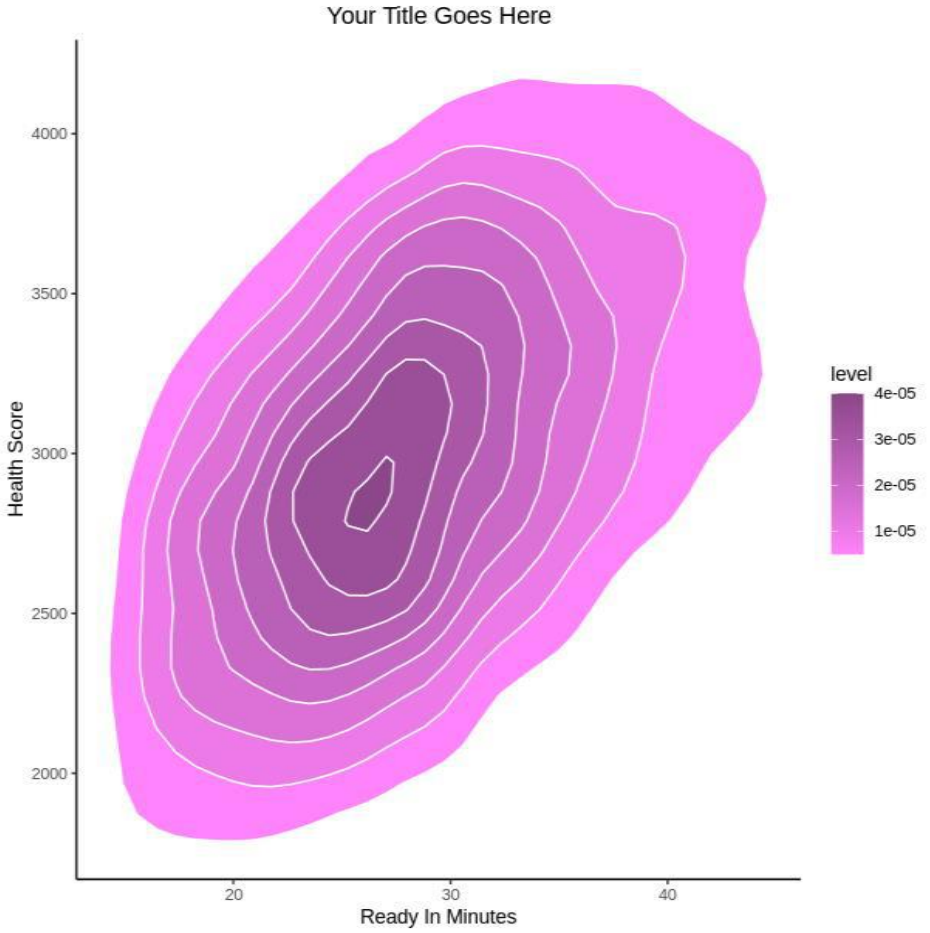
C) Draw the separate boxplots of the specified variables for each dataset. Your diagram must look like the following image. (titles, axis labels, legends,)



D) Draw mosaic plot of the specified variables for each dataset. Your diagram must look like the following image. (titles, axis labels, legends, percentage, etc.)



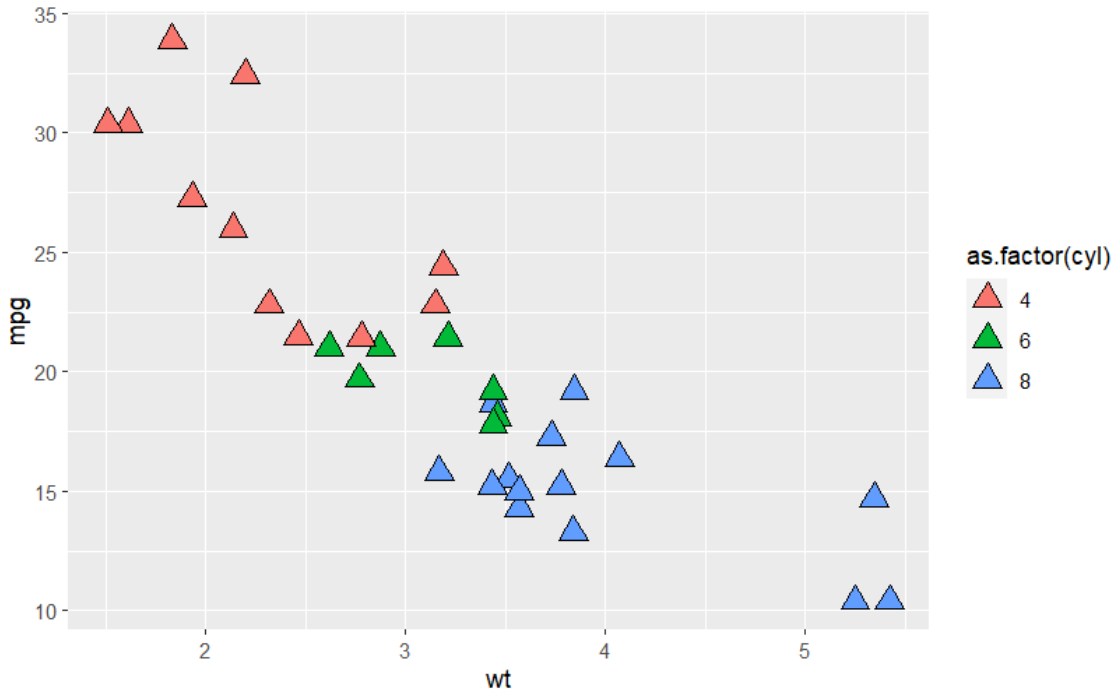
E) Draw 2D density plot of the specified variables for each dataset. Your diagram must look like the following image. (titles, axis labels, legends, percentage, etc.)



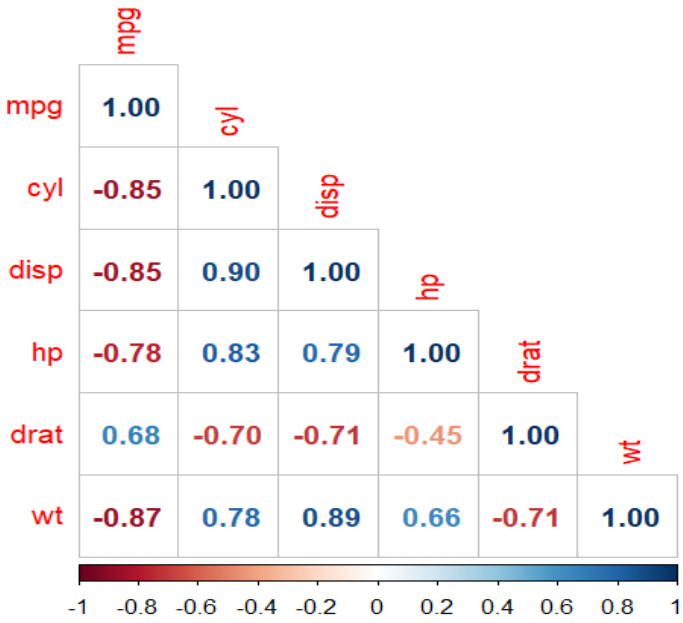
Question 7®

After drawing each plot, you should briefly explain the application of each plot and when it is necessary to use them!

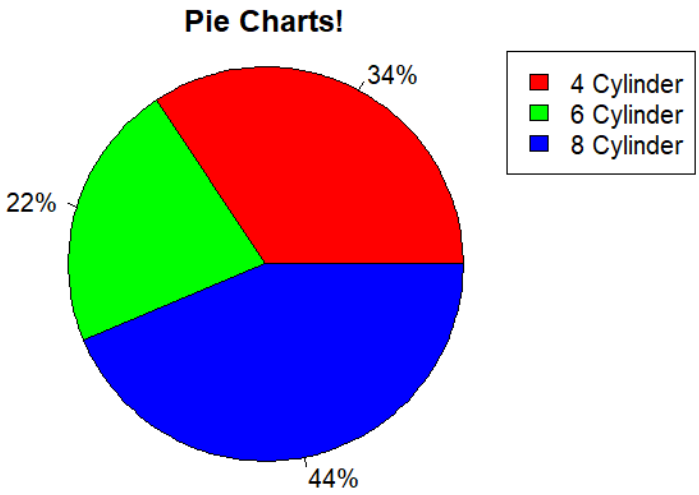
A) Draw scatterplot of each specified variables in the guide table.
Also, to make your plots more informative you can use color as the third dimension as it is specified below.
(You may draw your plot with any color or shape but your figures must have axis labels and titles)



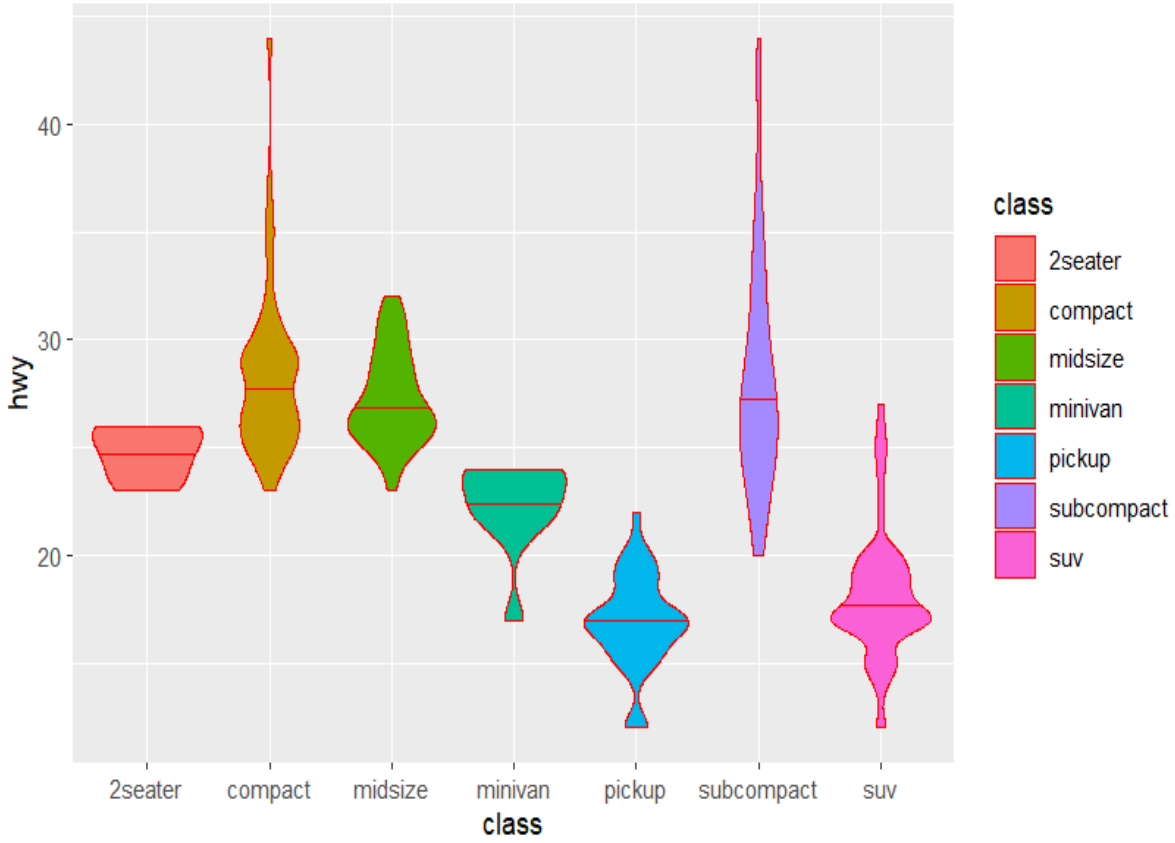
B) In regression, it is often necessary to choose the predictors which have the highest correlation with the response variable. In order to achieve this, we sometimes use “Correlogram”! Now, choose the first 6 columns of the *mtcars* dataset and draw the following figure:



C) Draw a pie chart according to the guide table’s specified variable. Please pay attention to all the details. Also, briefly explain what are the disadvantages of using pie charts? below chart is brought as an example.



D) Violin plots can bring many additional information to the observer compared to the ordinary boxplot. Discuss what these things are and then draw violin plot of each specified variables for each datasret. (your diagram must have titles, axis labels and legends)



Question 8®

In this question we want to draw qqplots for each vector of random distributions. So, first generate these random vectors with size = 50 with specife distributions and then Answer the questions:

- $X_1 \sim Exp(1)$
- $X_2 \sim N(0,1)$
- $X_3 \sim Cauchy(location = 0, scale = 0.5)$
- $X_3 \sim Uniform(0,1)$

A) Draw qqplot for each of the above vectors in separate subplots in a single figure.

B) Comparing the Diagrams how can you tell if a vector has nearly normal distribution?

C) How can qqplot be used in order to tell if two arbitrary vectors have the same distribution ?

Your plots should be something in this format.