Revision Exercises

Question 1

The table below is a small part of a data set that describes the fuel economy (in miles per gallon) of 1998 model motor vehicles.

Make and Model	Vehicle Type	Transmission Type	Number of Cylinders	City MPG	Highway MPG
:					
BMW 3181	Subcompact	Automatic	4	22	31
BMW 3181	Subcompact	Manual	4	23	32
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	16	30
.					

a) What are the individuals in this data set?
b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?
c) Present City MPG data in a well-labelled bar graph.
d) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

Question 2:

Company E is attempting to estimate when they should change tyres on their company cars. They feel that tyre tread is affected by the distance travelled.

The datafile tyres.csv contains data about distance and tyre ware.

a) Load the file tyres.csv into R
b) Produce a well labelled and presented scatter graph of the data, displaying the trend line
c) If the tyre travels 16 thousand km's what would the forecasted tread be?
d) Government regulations state that minimum tread is 4mm. What is the forecasted maximum distance you could travel on one set of tyres?

Question 3

You have been asked to analyse data about the Old Faithful Geyser.

a) Load the faithful data set that is built in to R.
b) Plot the data and then briefly describe the dataset
c) Summarise the data and calculate the
    - Minimum Value
    - 1st Quartile
    - Median
    - 3rd Quartile
    - Maximum Value
d) Calculate the standard deviations of the waiting time between eruptions. Can we assume a constant spread across the groups make a brief note.
e) You want to test the hypothesis that the resulting eruptions last on average three minutes.
    i) What should you choose as the null hypothesis and alternate hypothesis?

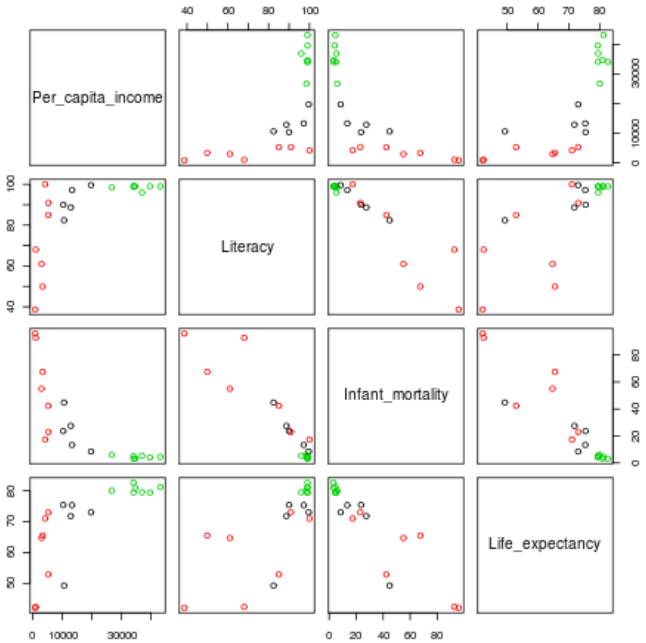
ii) Does your choice of hypothesis result in a one or a two tailed test? Explain you answer.
iii) Give R commands to compute the t statistic and the resulting p-value
iv) Would you reject the null hypothesis at the = 0:05 level? Explain you answer.

Question 4

Consider the scenario where countries need to be classified into three groups: developed, emerging and underdeveloped. To analyse their similarity and assign them to the groups, the following attributes should be taken into account:

- per capita income;
- literacy;
- infant mortality;
- life expectancy;

a) Import data from the countries.csv file
b) Apply k-means to the data, and store the clustering result (ensure you set the correct number of clusters)
c) Print the components of for the k-Means operation (**print**)
d) Plot the clusters and their centres for the first two dimensions: per capita income and literacy.
e) Plot the clusters for all dimensions displayed in a single graph (similar to below)



f) Briefly discuss the clusters that were generated and the countries that are in each cluster are they homogeneous did the clustering algorithm work?