Intro to R Programming: Class Test 1

Please log in using your own credentials. During the test you are not allowed to talk to or otherwise communicate with other students (email, instant messaging, etc.), or access the internet/course material. The only material you can use is the R reference manual provided as well as the R help within RStudio.

Only at the end of the test when specified by the invigilators, please log on to Moodle and upload your script by clicking on the link *Upload your class test script*. Please upload one R script only including your code and comments. You do not need to upload R output. Please do not upload Word documents or Zip files. The name of the file submitted **must be** your student ID number.

Please make sure you regularly save your R script, just in case RStudio crashes. For all parts in the test give the R code which can be used to answer the questions below.

You have to attempt all questions within the time of one hour

Task 1

The data file `houseprices.csv` contains data on property sales in and around Glasgow between July 1st and December 31st 2014. It contains the following columns:

Day	Day of the month of the transaction
Month	Month of the transaction (integer)
Address	Address of the property
Lon	Longitude of the property (degrees)
Lat	Latitude of the property (degrees)
Price	Price

Crown copyright material reproduced with the permission of Registers of Scotalnd

1. **[1 mark]** Read the data file `houseprices.csv` correctly into a data frame called `houseprices`.
2. **[1 mark]** What is the average house price in August 2014?
3. **[2 marks]** Create a dataset called `houseprices.summer` including the transactions occurred between July 15th and August 15th. How many transactions occurred in that period?
4. **[1 mark]** Which house sold for the lowest price?
5. **[1 mark]** Transform the column `Lon` to include the longitude of the properties expressed in radians, i.e. divide the longitude by 180° and multiply by π. Repeat the same process for `Lat`.
6. **[4 marks]** Create a new variable `Dist2University` which contains the distance to the University in kilometres. Consider two locations with longitudes λ_1 and λ_2 and latitudes ϕ_1 and ϕ_2 expressed in radians. Define

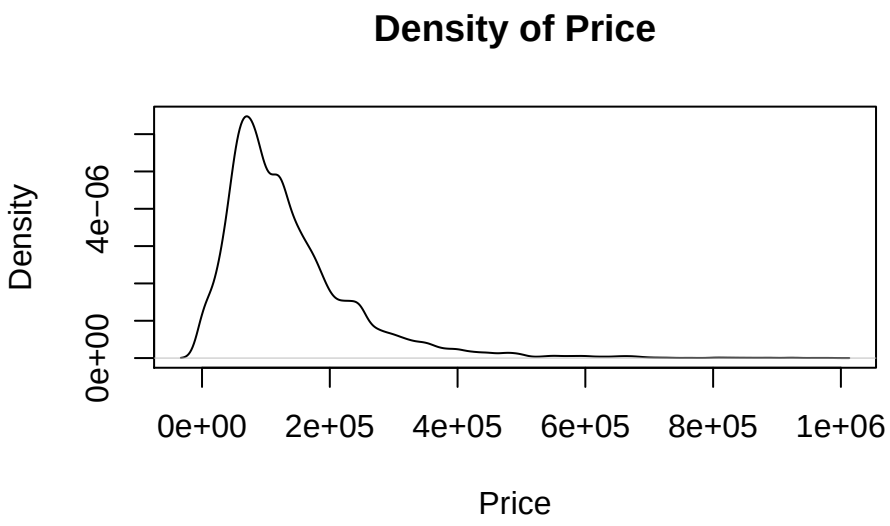
$$\begin{aligned}
\Delta\lambda &= \lambda_2 - \lambda_1 \\
\Delta\phi &= \phi_2 - \phi_1 \\
\alpha &= \sin\left(\frac{\Delta\phi}{2}\right)^2 + \cos(\phi_1)\cos(\phi_2)\sin\left(\frac{\Delta\lambda}{2}\right)^2 \\
d &= 12742\tan^{-1}\left(\frac{\sqrt{\alpha}}{\sqrt{1-\alpha}}\right),
\end{aligned}$$

then d gives the distance in kilometres between the two locations. The longitude and the latitude of the University are (in degrees):

$$\lambda = -4.2886^\circ, \qquad \phi = 55.8711^\circ$$

Hint: $tan^{-1}(\frac{a}{b})$ can be calculated in R using the function `atan2(a,b)`.

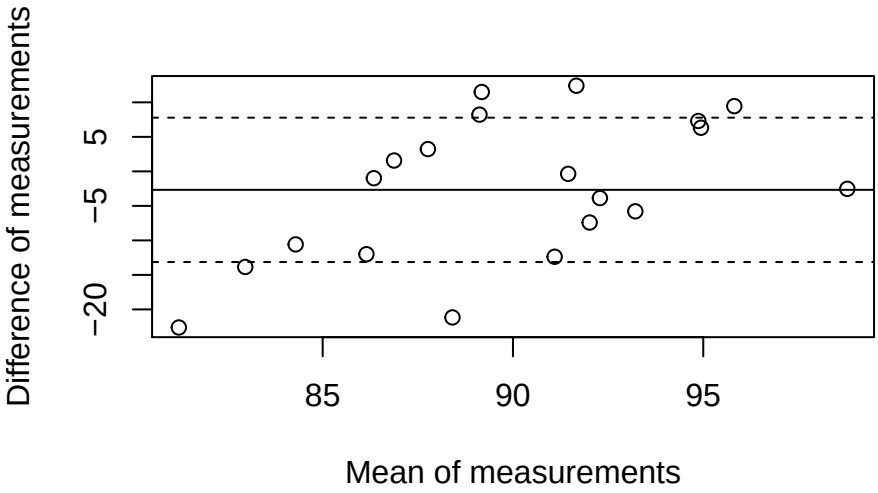
7. **[1 mark]** What was the average price of properties which are within 1km of the University?
8. **[2 marks]** Plot the density of the price of properties which cost less than one million. Include the title `Density of Price` and the label for the x-axis `Price`. Your plot should like like the following:



Task 2

The file `hearth.txt` contains measurements of the transmitral volumetric flow (`MF`) by Doppler echocardiography and left ventricular stroke volume (`SV`) by cross-section echocardiography in 21 patients without aortic valve disease.

1. **[1 mark]** Read the data correctly into R and store it in the data frame `hearth`.
2. **[1 mark]** Remove all rows containing missing values from `heart`.
3. **[2 marks]** Add the column `difference` including `MF - SV` and the column `mean` including (`MF`+ `SV`)/2 to the data frame `hearth`.
4. **[2 marks]** Suppose we are interested in informally assessing whether there is a systematic difference between the two techniques. This is often done using what is called the Bland-Altman plot, which is a scatter plot of `mean` against `difference`. Produce such a plot and label the x-axis `Mean of measurements` and the y-axis `Difference of measurements`.
5. **[2 marks]** Add a solid horizontal line at the average value of `difference` and a pair of dashed lines one standard deviation above and below the solid horizontal line. Your plot should look similar to the one below.



Task 3

The file `potus.txt` contains the results of the last presidential election for each county in the United States. The dataset contains the following columns:

County	Name of the county
State	Name of the state
VotesTrump	Number of votes cast for Donald Trump
VotesClinton	Number of votes cast for Hillary Clinton
PercTrump	Vote share (percent) for Donald Trump
PercClinton	Vote share (percent) for Hillary Clinton
PercWhite	Percentage of the population which is Caucasian
HIncome	Median household income.

1. [**1 mark**] Read the file `potus.txt` into R and store into a data frame called `potus`.
2. [**2 marks**] What is the average median household income in counties in which Donald Trump received at least three times as many votes as Hillary Clinton?
3. [**1 mark**] How many votes have been cast for Hillary Clinton in the state of California?
4. [**1 mark**] Add a column to `potus` named `Hillary.Wins` which is `TRUE` for counties where Hillary Clinton has more votes than Donald Trump and `FALSE` otherwise.
5. [**3 marks**] Create a plot of `HIncome` against `PercWhite`. Counties in which Hillary Clinton obtained more votes than Donald Trump should be plotted in red, the others in blue.
6. [**1 mark**] Add a legend to the plot from part 5. Your final plot should look similar to the one below.