Take Home exam, R (tidyverse), 2019.

*Submission must be before 5pm, April 17th*.

Download the examR.zip file from the Resources folder on CAMBRO extract it and save the folder examR in your R\_Course folder. **On completion, zip your examR folder and rename it with your CODE.** Submit the zip file to Angelica .

**Q1. (50%)** You and I are writing a paper on some analysis of back-pain data in the Sage database . I’ve started our collaboration , in **Q1.Rmd**, which you will find in the examRcode folder. Open and examine the Q1.Rmd file. Note that we are using a different file containing a little more data than the one you used during the course.

Firstly, check that the .Rmd file will ‘knit’, so that you know you are accessing the data correctly, using the relative path and file name in the readData chunk of the Q1.Rmd file.

This question comprises the following numbered list of tasks, each with an indicated mark. There is further explanation of the requirements of each task in Q1.rmd and these requirements must be met to achieve full marks.

1. Add your CODE as your name in the “author” specification and modify the title appropriately. (1 Mark)
2. Modify variable names as indicated in Q1.Rmd (3 Marks)
3. Change the type of the 11 variables to factors (4 Marks)
4. Modify levels as indicated in Q1.Rmd (6 Marks)
5. Include a summary() (2 Marks)
6. Present ggplots of a boxplot and histogram of bmi for men and women (6 Marks)
7. Prepare tables of Categorical variables as indicated in Q1.Rmd (8 Marks)
8. Prepare regression plots and compare as indicated in Q1.Rmd ( 8 Marks)
9. Generate and comment on diagnostics for the regressions. (4 Marks)
10. Show boxplots of the bmi of Urban, Chinese women for all wealth categories and use colour to show smokers and non-smokers. (6 Marks)
11. Save the cleaned up df in the ExamData folder to be loaded in the next question (1 Mark)
12. Save your completed, **knit-able** .Rmd file in the examRcode folder (where you found it). (1 Mark)

[At the completion of your exam you must re-name the examR folder with your CODE, zip it, and submit it.]

**Q2. (10%) Use the file, Q2.Rmd in the examR folder**, with comments, to do the following:

1. Add your CODE as your name in the “author” specification.
2. Load the saved file from Q1. (1 Mark)
3. Use functions from the ggplot2 package with the SAGE backPain data set (from Q1), to produce boxplots of bmi against eworked, using fill in the aesthetics to distinguish men and women. ( 3 Marks)
4. Use a facet grid of (country ~ depression). (3 Marks)
5. Selectively remove NA’s. (2 Marks)
6. Briefly comment. (1 Mark)

**Q3. (20% ) ANSWER EITHER: Q3A or Q3B – do not do both parts!**

**Q3A – (20%) Mapping**

1. Use Q3a.Rmd to:
2. Download a WDI of a social determinant ( NOT gnp per capita) and plot it on a world map, using country names for the join. ( 9 Marks)
3. There will be problems with different country names in the two data sets: the map and WDI. Show how to change the name for the USA and Russia to match in the two data-frames before joining. (5 Marks)
4. Get a second WDI with a health measure and plot it on a world map; put both maps on a single figure, one above the other. (6 Marks)

OR:

**Q3B. (20%)** **Breast cancer data**. The RMarkdown file Q3B.Rmd in the ExamRCode folder outlines and provides hints on some of the steps to do the following:

1. Load the BreastCancer dataset from the mlbench package. (1 Marks)
2. Undertake some exploratory analysis with Tables (3 Marks)
3. Create some exploratory visualizations (4 Marks)
4. Do a logistic regression of probability of malignancy vs Cl.thickness (4 Marks)
5. Prepare a ggplot of 4. (3 Marks)
6. Report on the outcomes - (acknowledging the age of the data. Look **briefly** at more recent literature and comment. (5 Marks)

Website for data: <http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)>

**Q4. (20%)** **Programming**. The somewhat surprising result that the estimators of the mean and the standard deviation of a sample drawn from a population which is normally distributed with unknown mean and standard deviation are independent, can be proved using moment generating functions.

Using simulation we can **demonstrate** that it does indeed appear to be so, though we cannot prove it in this fashion. At the same time, this kind of simulation helps improve our intuition about such results.

Undertake a simulation for this purpose. Create a RMarkdown file which does the following:

1. Assign values to *m*, the number of samples and *n*, the size of the sample. (1 Mark)
2. Assign values to the population mean, *mu* and standard deviation, *sdev*. (1 Mark)
3. Sample using rnorm in a loop over m to generate the *m\**n random numbers and store them in a matrix initialized with code: df<- matrix(, nrow = n, ncol = m) (4 Marks)
4. Then: df <- as.data.frame(df)
5. Compute the *m* sample means and standard deviations. (using sapply twice to generate two vectors) . (3 Marks)
6. Compute the correlation of the vectors from part 5. (1 Mark)
7. Do a simple scatter plot of the standard deviations vs means using ggplot2 (3 marks)
8. Test the standard definition of independence: (4 marks)

where Y is the sample mean and *S* is the sample standard deviation.

1. Test the speed of your simulation using the function Sys.time(). ( 3 Marks)

Save your file in the ExamRcode folder as Q4.Rmd.