G54DMA - Lab 4: Data Analysis in R

Question Sheet

This question sheet presents a series of exercises designed to make you use R for data analysis purposes. You will be applying the concepts you learned during the first two labs (data frames, reading data from files, etc.) to define and analyse data and answer questions. You will also be introduced to R's database repository.

In this lab session, you will learn to:

- **Use R's built-in datasets**
- **Use basic R and statistics to analyse data**
- **Use the library *dplyr* to manipulate and transform data**

To begin, start your R editor (RStudio or your editor of choice). I recommend that you revise Lecture 4 and read the Instruction sheet carefully. Once you have finished reading them, start working on these exercises.

A. The Iris Dataset

In this section, we will carry out simple data analysis on the ***Iris*** dataset. For each question, **provide two answers: one answer using basic R functions and one using *dplyr*.** Remember to load the ***Iris*** dataset to your R session before attempting any of the following exercises.

1. Obtain centrality measures (minimum, maximum, mean, median, 1st quartile, 2nd quartile and 3rd quartile) from all the relevant attributes from the *Iris* dataset.
2. Obtain the same information as in the previous section, but grouping the data into the three classes of plants.
   a. Which plants have a higher mean sepal length?

   b. Which plants have the sample with the smaller petal width?
   c. What is the median petal length of the virginica samples?
   d. What is the class of the plant with the highest sepal width?

3. Select a random sample of the *Iris* database with a number of instances between 75 and 100.
   a. What is the correlation coefficient between the sepal length and the width? What does it say about their relationship?
   b. What is the correlation coefficient between the sepal length and the petal length? What does it say about their relationship?
   c. What is the correlation coefficient between the petal length and the petal width? What does it say about their relationship?
4. Create a new column in the dataset, called *petal.ratio*, that calculates the ratio between the petal length and the petal width of all instances.
   a. What is the species of the sample with the highest ratio?
5. Create a new column in the dataset *small.setosa*, which gives a value of *true* to setosa plants with petal widths of less than 0.3 cm and *false* to any other samples.
   a. Calculate the mean, standard deviation and median of the Petal Width grouping samples by *small.setosa* values.

B. The Island Dataset

In this section, we will be using the ***island*** dataset. You can load the dataset from RStudio. This dataset has the areas in thousands of square miles of the landmasses which exceed 10,000 square miles. Using the dataset and the statistics we have seen in class, answer the following questions. You may use basic R or *dplyr* functions as you prefer.

1. How many island locations have been considered?
2. What is the average area of the islands? And the median?
3. What is the size of the biggest island? And its name?
4. And the size of the smallest island? And its name?
5. How can you obtain both measurements (minimum and maximum area) using only one function?
6. Calculate the dispersion measurements for all the areas.

7. How many islands are larger than the third quantile area?

C. The Chocolate Bar Dataset

In this section, we will analyse the ***Chocolate bars***[1] dataset. You can download the dataset from Moodle (*chocolate-bar-dataset.csv*). You will notice that the dataset has some missing values. This is ok. We will tackle how to pre-process our dataset to fix these issues in Lecture 6 (*Data Pre-processing and Data Mining*). For the time being, we will focus on describing our data without changing or fixing it. You may use basic R or *dplyr* functions as you prefer.

1. How many reviews have been recorded?
2. How many attributes for each chocolate bar are collected?
3. What is the data type of each attribute?
   a. X:
   b. Company:
   c. Name:
   d. Ref:
   e. Review.Date:
   f. Cocoa.Percent:
   g. Company.Location:
   h. Rating:
   i. Broad.Bean.Origin:
   j. Bean.Type
4. Obtain the mode from all of the nominal attributes in the dataset.
5. Obtain the mean, median, Q1 and Q3 from all relevant attributes in the dataset.
6. How many distinct companies have been considered?
7. How many distinct company locations have been collected?
8. Are there more samples from companies based in the UK or in Peru?
9. Are there chocolate bars from Spanish companies? How many?
10. Are there chocolate bars from Hungarian companies with a Rating over 3?
11. What is the most common bean origin?

[1] http://flavorsofcacao.com

12. What is the bean origin of the chocolate bar with the highest rating? Where is the company based? If there is more than one chocolate bar with maximum rating, report all bean origins.
13. Obtain the minimum, maximum and mean of the rating.
14. How many Belgian bars are rated 4? And how many French bars?
15. Obtain the centrality and dispersion measurements for the Rating of chocolate bars whose companies are in U.S.A.
16. On average, which country has the highest-rated bars?
17. How many reviews have been carried out each year since 2006?
18. Which year has the most reviews? Which year has the fewest reviews?

D. The Pulitzer Prize Dataset

Finally, in this section we will work with the ***Pulitzer.cvs*** dataset, which you can download from Moodle. This dataset stores how many Pulitzer finalists and winners different publications have had since 1990. Load the dataset and answer the following questions. You may use basic R or *dplyr* functions as you prefer.

1. How many journals have been considered?
2. How many attributes have been collected per publications?
3. What is the type of the data collected?
    a. Newspaper:
    b. Daily Circulation 2014:
    c. Daily Circulation 2003:
    d. Change in Daily Circulation:
    e. Winners and Finalists 1990-2003:
    f. Winners and Finalists 2004-2014:
    g. Winners and Finalists 1990-2014:
4. Calculate centrality and dispersion measurements from all of the appropriate attributes.
5. What is the total circulation of all of the journals considered?
6. How many publications have experience a positive increase in their circulation?
7. What is the publication with more winners and finalists overall?
8. And the publication with more winners and finalists between 1990 and 2003?
9. What is the range in the circulation in 2004? And in 2013?
10. How many journals experienced a decrease in circulation of over 40%?

11. What is the journal with most finalist and winners?
12. How many journals had over 30 candidates or winners between 1990-2003 and between 2004-2014. Which journals are these?
13. How many journals with a daily circulation of over 1 million in 2004 have received more than 30 nominations or prizes since 2004?