

BUS5DWR - Assignment 3 (40%)

The purpose of this assignment is to develop and assess your skills in R programming including summarising, wrangling and plotting data. Using the tidyverse package is recommended but not compulsory. Please read through the entire assignment and understand the submission format and marking rubrics before starting.

Part 1 [22 marks]

The spreadsheet titled 'obesity.xlsx' records the prevalence of obesity among adults in each country in the world. Each sheet is supposed to record the information of each country in each year including the average obesity rate and its 95% confidence interval of female, male and both groups. There were three years in the given dataset (2006, 2011, and 2016). Another sheet called Continent with information about countries in each continent is also given.

You will see that it is far from being ready for analysis and needs to be 'wrangled'. Additionally, a few errors have been deliberately introduced so these will need to be corrected by applying your R code.

- 1.1. Explain why the data in its current form is not considered to be in 'tidy' format. (1 mark)
- 1.2. Write a function that takes a year and outputs a dataframe with one or more rows, where each row shows the obesity rate of Male and Female in a country of that given year along with its 95% confidence interval values. The returned dataframe should have 6 columns (Country, Year, Sex, Rate, MinCI, MaxCI).
 - a) Load the data from the worksheet of the given year into a dataframe. (1 mark)
 - b) Drop the column 'Both sexes'. (1 mark)
 - c) Add a new column named 'Year' filled all rows with the given year. (1 mark)
 - d) Use the gather() function to transform the data in the two columns Male and Female to become rows. After this step, your dataframe now should have 4 columns: Country, Year, Sex, Rate. (2 marks)
 - e) Split the Rate column into 3 columns named Rate, MinCI, MaxCI. Your dataframe should have 6 columns after this step: Country, Year, Sex, Rate, MinCI, MaxCI (2 marks)
 - f) Check and make necessary changes to make sure the data type of Rate, MinCI and MaxCI is numeric. Print the summary of the dataframe. (1 mark)
 - g) Find and display rows with any invalid data, e.g. the rate value is not in the range of MinCI and MaxCI, MinCI is larger than MaxCI, etc. If they exist, change the MinCI and MaxCI values in these rows into NA. Print the summary of the dataframe. (2 marks)
 - h) Return the dataframe
- 1.3. Apply the function to each of the three years in the data to obtain three datasets then combine the rows to form a single dataframe. Print the numbers of rows of the dataframe. (2 marks)
- 1.4. Query the dataframe obtained in 1.3 to print the average obesity rates of Female and Male of each year. (2 marks)
- 1.5. Sort the dataframe obtained in 1.3 and display the country name, year, sex and rate in descending order of obesity rates. Write the result to a csv file. (2 marks)
- 1.6. Load the data from the Continent worksheet into a dataframe, keeping only two columns, "Country or area" and "Continent". (1 mark)

- 1.7. Check if there is any country in the dataframe obtained in (1.3) not in the country list loaded in (1.6). Display the country names if any (no duplicates). (2 marks)
- 1.8. Display the average obesity rate of Female in Europe and North America. (2 marks)

Part 2 [18 marks]

The online hospitality company Airbnb has made publicly available a number of datasets. This part of the assignment makes use of a subset of the Melbourne dataset. The dataset is given in the AirBnBMel6500.tsv file.

It consists of a number of parameters related to properties available for lodging in the Melbourne metropolitan area and can be visualised at <http://insideairbnb.com/melbourne>.

Write R code to answer the following.

- 2.1. Load the dataset from the given file into a dataframe. Change the column name to remove spaces. Observe the data and report whether the type of each column is appropriate or not to the data. (2 marks)
- 2.2. How many listings and unique locations in the dataframe? (1 mark)
- 2.3. Keep only the listings that have the last review in 2019 in a dataframe. Remove all the others. Print the number of remained listings and unique locations. (2 marks)
- 2.4. Display the number of listings of the three most popular property type, excluding listings with missing property type. (2 marks)
- 2.5. Remove the country name (Australia) in the location column. (2 marks)
- 2.6. Find the average price of listings in Carlton. (1 mark)
- 2.7. Find the top ten locations that have the highest average price. Display the name and the average price in its descending order. (2 marks)
- 2.8. Display the listing ID and location of listings that its transport description mentions both words university and supermarket with upper or lower case or mixed in any order. (2 marks)
- 2.9. Suppose somebody wants to choose a listing based on the following criteria. Write a function that inputs a listing id and returns a score that is the sum of points as below: (3 marks)
 - a. Points for price: (200 minus price) but not less than zero
 - b. (Review score rating minus 100)
 - c. Points by popularity based on the number of reviews: 100 if at least in the first quartile, 0 if less than the median value, 50 otherwise.
- 2.10. Which listing ID has the highest score according to the above criteria? (1 mark)

Submission Guidelines

Your submission to this assignment will consist of two files.

- 1) A single .Rmd file comprising all the codes to do the described requirements for all parts in the given order. Please include introduction to the code for each subquestion and explain your steps.
- 2) A generated report as a HTML file from your notebook.

Note: When writing your code, keep the data files in the same directory as your notebook so that you do not specify directories in your code. This will make your code easier to assess.

Marks will be deducted if your submission does not follow the guidelines.

Marking Rubrics

Please note the following as it shows how marks may be deducted.

Marks will be deducted if the R code does not work easily on the marker's R studio installation and if you need to be offered an opportunity to show the marker that it does work on your installation. This means all references to directories have to be removed and packages being used are to be specified clearly. It will be assumed that the tidyverse and readxl have been installed.

For each coding question:

- Full mark will be given for non-error and correct answer,
- Half of the mark will be given for something close,
- 0.5 mark will be deducted for not having or having an incomplete introduction to the code, including the question number (if it is in the middle of a code chunk, put it in a comment).