BUS5DWR – Data Wrangling and R
Assignment 02: Data wrangling with R
Marks: 60 (will be scaled to 30)
Assignment Type: Individual

Overview

Over the past few weeks, you have learned how to use R to wrangle business data. This assignment will provide you with an opportunity to demonstrate your R skill for data wrangling. Using the tidyverse package is recommended but not compulsory.

Please carefully read the entire assignment to make sure you understand the requirements and also the submission format and marking rubrics before starting.

Academic Integrity

Plagiarism occurs when you use words, ideas, or work products attributable to another identifiable person or source:

- without attributing the work to the source from which it was obtained
- in a situation in which there is a legitimate expectation of original authorship
- in order to obtain some benefit, credit, or gain which need not be monetary

Self-plagiarism refers to the re-submission of work as if it were original. You may not submit your own academic work for assessment when it has already been submitted for assessment at another time (including at another institution), without the express permission of the academic staff member who will assess it.

By submitting* this piece of work and signing this document, I declare that:

1. The work is my own individual work.
2. I have not previously submitted all or part of this work for assessment in any subject unless the subject coordinator for the current subject (or my research supervisor, if applicable) has given me written permission to reuse specific material and I have correctly referenced the material taken from my own earlier work.
3. I have read and agree to be bound by the Statutes, Regulations and Policies of the University relating to Academic Integrity available at http://www.latrobe.edu.au/students/academic-integrity; and
4. I may be subject to student discipline processes in the event of an act of academic misconduct by me, including an act of plagiarism or cheating.

I further grant to the University or any third party authorised by the University (www.latrobe.edu.au/text-match) the right to reproduce and/or communicate (make available online or electronically transmit) the work I have submitted for the purpose of detecting plagiarism.

Assignment Requirements

Part 1 [30 marks]

The online hospitality company Airbnb has made publicly available a number of datasets. This part of the assignment makes use of the detail data of Sydney listings which is available at the link below. You have to unzip it after downloading.

http://data.insideairbnb.com/australia/nsw/sydney/2022-06-06/data/listings.csv.gz

The data dictionary of this dataset can be found at

https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?usp=sharing

Write R code in an Rmd file to perform the following request.

1.1. Read in the dataset into a dataframe. Keep only the below columns. Make sure the numeric columns are in the correct data type. Display the summary of the dataframe.

id, name, description, host_name, neighbourhood_cleansed, property_type, room_type, accommodates, bathrooms_text, bedrooms, beds, amenities, price, number_of_reviews, number_of_reviews_ltm, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value.

(3 marks)

1.2. How many listings have their name column contain:

a. *Beautiful* (in upper or lower case or mixed).

b. Both *Quiet* and *Beautiful* (in upper or lower case or mixed).

(2 marks)

1.3. List the five neighbourhoods with the highest number of reviews in the last 12 months (list them along with the average review score rating).

(3 marks)

1.4. Display the average price and average review score rating of each property type in Woollahra.

(2 marks)

1.5. You are going to visit Sydney and want to find an accommodation. Let define four criteria based on:

a. the neighbourhood,

b. the maximum price,

c. the bath room type, and

d. the required amenities.

You must adjust the defined criteria so that the result has at least 10 and at most 20 listings for you to choose. Display only the name and columns related to the criteria.

(5 marks)

1.6. Draw a bar chart to show the number of listing of the top 10 hosts having the most listings. Write a short paragraph to describe the chart.

(7 marks)

1.7. Draw a histogram and a boxplot to show the distribution of the listing prices of all types of *private room.* Redraw both charts with outliers removed. Write a short paragraph to describe your insight about the price distribution.

(8 marks)

Part 2 [30 marks]

The given data file, **census.xlsx**, contains the information about the number of bedrooms in occupied private dwellings for local government areas in Melbourne for the years 2011 and 2016. You will see that the data is far from being ready for analysis and needs to be 'wrangled'. Additionally a few errors have been deliberately introduced into the data so these will need to be corrected. You are required to write R code to perform the following steps to have the clean data.

2.1. Read in the 2011 dataset into a dataframe. Show the structure of the dataframe. (2 marks)

2.2. Investigate and fix if there are any inconsistent values in the first column. (2 marks)

2.3. You can see that the second column contains both count and percentage. Let split the two values into 2 columns. Show the structure of the dataframe. (3 marks)

2.4. You can see that each suburb have 9 rows of data. Display the number of suburbs in the dataframe. (1 mark)

2.5. We only interest in the count value. Remove the percentage column. Then, transform data to have each statistic shown in one column. (3 marks)

2.6. Add the year column showing the year of the data. Make sure each column has an appropriate type. Rename and reorder columns to have a data frame with columns in the below order. Show the summary of the dataframe.

region, year, br_count_0, br_count_1, br_count_2, br_count_3, br_count_4_or_more, br_count_unstated, av_per_dwelling, av_per_household (3 marks)

2.7. How many regions do we have in the data? Remove if there are any duplicate regions. (2 marks)

2.8. Define a function that takes a year, applies all the steps from 2.1 to 2.7 to return a clean dataframe for that year. (4 marks)

2.9. Call the defined function to have two dataframes for 2011 and 2016. Then, combine them into one dataframe. (2 marks)

2.10. Investigate the combined dataframe in 2.11, report and fix it if there is any errors in the data. Show a summary of the combined dataframe. (4 marks)

2.11. How many houses with 2 or 3 bedrooms in 2016? (2 marks)

2.12. Which region has the largest decrease in the number of 3 bed room houses from 2011 to 2016? (2 marks)

Submission Guidelines

- You must submit **A SINGLE file (.Rmd)** comprising all the codes to answer all the questions of the two parts **in the given order**.
- ***Answer to each question is presented in ONE code chunk***.
- PUT the question number before the code chunk. DO NOT include the question description (to avoid a high Turnitin similarity score).
- When writing your code, keep the data files in the same directory as your notebook so that you DO NOT specify directories or file paths in your code. This allows us to run your code smoothly on our device.
- Marks will be deducted if your submission does not follow the guidelines.

Marking Rubrics

- For each question, the full mark will be awarded for a non-error and correct answer. Half of the mark will be given for something close.
- Marks will be deducted if the R code does not work smoothly on the marker's R studio installation, and we need to offer you an opportunity to show us that it does work on your installation. This means ***all absolute paths or references to your local machine's directories must be removed, and the packages*** must be specified clearly. It is assumed that the tidyverse and readxl have been installed on our device. If you use other packages, have commands to install and load them.
- Submissions with high similarity in the answers with another submission will be considered plagiarism/collusion. The only way to avoid plagiarism is no copying from anyone or letting anyone else see your answers.