MONASH University

Faculty of Information Technology
Semester 2, 2020

FIT5145 Introduction to Data Science
Assignment 1: Description

Due Date: 11:55pm, Friday 11 September 2020

The aim of this assignment is to investigate and visualise data using various data science tools. It will test your ability to:

- Using R,
  - read data files and extract related data from those files;
  - wrangle and process data into the required formats**;**
  - use various graphical and non-graphical tools to perform exploratory data analysis and visualisation; and
- communicate your findings in your report.

Tasks:

- There are two tasks (**A & B**) in this assignment. Each task has separate data set files.
- You need to use **R** to complete the tasks.
- You need to use **R Markdown** to communicate
  - your answers,
  - the code you used to complete the tasks, and
  - your explanation of the steps you took and any issues that arose

It is crucial that the R Markdown report you submit clearly identifies which questions you are answering, and explains how you are processing the data and why you are processing the data in that way. **It is not adequate for you to just answer the questions for each task or just supply the code you used.**

The data supplied for each task **will also have to be wrangled** in order to answer the questions. The supplied data is not guaranteed to be "clean" and without faults. This may require you to

- examine the data,
- filter the data,
- deal with missing or inconsistent values or formats,
- deal with any outliers or exceptional values,
- merge or divide the values or data sets,
- sort the data, and/or
- any other pre-processing steps that are required in order to be able to analyse the data.

Your report must explain why and how you are performing this data wrangling, including identifying any issues you find with the data.

Task A: Investigating the size of the Indigenous Australian Population

In this task, you are required to visualise the relationship between the distribution and age of Indigenous Australians and gain insights into relations and trends over time. The data files used in this task were originally downloaded from the Australian Bureau of Statistics (ABS). **We have extracted the data from the original files and put it into a simpler format.** Please download the data from Moodle:

- **IndigAusPopData_by_region (Data1):** This file contains yearly data regarding the estimated resident population of Indigenous Australians, grouping by indigenous regions, between 2016 to 2031.
- **IndigAusPopData_by_state (Data2):** This file contains yearly data regarding the estimated resident population of Indigenous Australians, grouping by state or territory, between 2006 and 2031.

A1. Investigating the Distribution of Indigenous Australians

Indigenous Australians are part of Australian society everywhere, but some parts of the country have larger populations than others. For Data1, Australia is segmented into **regions** (titled "Indigenous regions") and the expected Indigenous population for each region is indicated. This data also divides each region's population into different age groups.

1. Use R to read, wrangle and analyse the data in Data1. Make sure you describe any complications you encounter and the steps you take when answering the following questions.

a. What **regions** have the maximum and minimum **total** Indigenous populations in 2016 and 2031?

b. What **region/s** have the maximum and minimum growth or decay rates of their **total** Indigenous population between 2016 and 2031?

Calculate these rates as the percentage difference between the 2016 and 2031,
e.g., if 2031 population = 5500 & 2016 population = 5000,
then rate = (5500 – 5000) / 5000 = 500/5000 = 0.1, so 10% growth

c. Plot and describe the growth or decay of the **total** Indigenous populations for the **capitals of the 8 state/territories** across **all time periods**.

For these calculations, you will need to work out the growth/decay rates for each time period, where the total population of the capital in time period N is compared to that in time period N+1.
e.g., if 2017 population = 5050 and 2016 population = 5000,
then rate = (5050 - 5000) / 5000 = 50/5000 = 0.01, so 1% growth for 2016-2017

A2. Investigating the Ages of Indigenous Australians

On average, the lifespan of Indigenous Australians is lower than that of the overall Australian population, due to a variety of socio-economic factors. Data1 and Data2 give separate populations for different ages or age groups, but because this is about living populations, not when they die, we can't use it to calculate average lifespans. Instead, let's look at how many children are in the populations. Make sure you describe any complications you encounter and the steps you take when answering the following questions.

1. Using **Data1**, which **region** has the highest percentage of **children** in its **total** 2016 population?

   For this, calculate this as a percentage of the total population for a region. The ABS commonly considers children to be under 15 years of age.

2. Data2 includes estimated populations measured for the years 2006-2016 and projected estimates predicted for the years 2016-2031. Data1 just uses projected estimates. Using **Data2** only, calculate and discuss which **state or territory** has the highest percentage of **children** in its **total** 2006, 2016 and 2031 populations.

3. Use R to build a Motion Chart comparing the total Indigenous Australian population of **each region** to the percentage of Indigenous Australian children in **each state/territory**. Use the **region populations calculated from Data1** and the **child percentage values calculated from Data2.** The motion chart should show the population on the x-axis, the percentage on the y-axis, the bubble size should depend on the population.

   Hint: an example of how to construct an R motion chart can be found on Moodle. You will have to install the 'googleVis' package and may have to allow Flash to work on your browser (see https://community.rstudio.com/t/gvismotionchart-from-googlevis-is-not-working-any-suggestion/6109/9 for advice on allowing Flash for Chrome). If you cannot get the example script to work, contact your tutor.

4. Using the Motion Chart, answer the following questions, supporting your answers with relevant R code and/or Motion Charts
   a. Which **region**'s population overtakes that of another region **in the same state/territory**? In which **year/s** does this happen?

   b. Is there generally a relationship between the Indigenous Australian **population size** and **percentage of children in the population**? If so, what kind of relationship? Explain your answer.

   c. **Colour** is commonly used in data visualisation to help understand data. Which aspect of this data would you use colour for in your plot and why?

   d. Are there any other interesting things you notice in the data or any changes you would recommend for the Motion Chart?

B: Exploratory Analysis on Australian Immunisation rates

In this task, you are required to do some exploratory analysis on data relating to the Australian childhood immunisation rates. This data was originally prepared and released through the Australian Government's Australian Institute of Health and Welfare. **We have extracted the data from the original files and put it into a simpler format.** Please download the data from Moodle:

- **AusImmunisationData (Data3):** This file contains yearly data regarding the number of 1, 2 and 5 year-old Australian children fully or partially immunised in various Primary Health Network (PHN) areas.

COLUMN	DESCRIPTION
State	State or territory for the PHN area
PHN code	Identification number for PHN area relating to the data
PHN area name	Description of PHN area
Reporting Year	Financial period examined
Age group	Age group of children
Number of registered children	Number of children registered in the age group
Number fully immunised	Number of children in the age group who were fully immunised, according to government objectives
Number not fully immunised	Number of children in the age group who were not fully immunised, according to government objectives
Number of registered IndigAus children	Number of Indigenous Australian children in the age group
Number IndigAus fully immunised	Number of Indigenous Australian children in the age group who were fully immunised, according to government objectives
Number IndigAus not fully immunised	Number of Indigenous Australian children in the age group who were not fully immunised, according to government objectives
Interpret with caution	This area's eligible population is between 26 and 100 registered children.

Use R to read, wrangle and analyse the data from Data3. Make sure you describe any complications you encounter and the steps you take when answering the following questions.

B1. Values and Variables

1. How many **PHN areas** does the data cover?

2. What are the possible values for **'PHN code'**?

3. For each row, calculate the percentage of **Australian children** that are fully immunised (this is the immunisation rate). What are the average, maximum and minimum immunisation rates? Calculate the same for the group that are **Indigenous Australian children**. Do all of those values seem statistically reasonable to you?

B2. Variation in rates over Time, Age and Location

Generate **boxplots (or other plots)** of the **immunisation rates** versus **year and age** to answer the following questions:

1. Have the immunisation rates improved over time? Are the **median** immunisation rates increasing, decreasing or staying the same?

2. How do the immunisation rates vary with the **age** of the child?

Generate **boxplots** (or other plots) of the **immunisation rates** versus **locations** and answer the following questions:

3. What is the **median** rate per **state/territory?**

4. Which **states or territories** seem most consistent in their immunisation rates?

Assessment Resources

You will need the following resources in order to complete this assessment item.

- A R-Markdown file containing all R code that you have written to wrangle and process, analyse and plot the data for Tasks A and B. It must clearly contain your answers to all the questions and any explanations about how you completed the tasks.
- You may need to review the **FIT citation style** guide **(https://guides.lib.monash.edu/c.php?g=219786&p=1453281)** to make sure you're familiar with appropriate citing and referencing for this assessment. Also review the Monash University library's guide on **citing and referencing (http://www.monash.edu/library/skills/resources/tutorials/citing)** for help.

Development instructions:

- **Use R version 3. No other programming languages or statistical software is allowed**.
- If you use any external R packages that are not available from CRAN, you must include installation directions.
- Make sure that your code works independent of where the working directory is placed. Staff need to be able to run your code on any OS.
- Make sure your code for Tasks A and B can be run independently from each other. For instance, if they both require the same package, then they both contain the relevant `library` line.
- All data files must be expected to be in the same directory as your R-Markdown file, not a subdirectory or elsewhere.
- Do not change the names of the unmodified data files.
- Do not include any data files in your submission for Tasks A and B.
- Make sure it is very clear where you are answering which question in your work.
- Make sure you explain what and why you process and analyse the data as you do.

How to Submit

Once you have completed your work, take the following steps to submit your work.

1. Please ensure you **name the file containing your work for Tasks A and B** correctly using the following format:
   LastName_StudentNumber_Assessment#AB.Rmd
   e.g., *Finn_21872187_Assessment1AB.Rmd*

2. Upload your assignment in the assignment link provided on Moodle

Penalties

- Late submission

For all assessment items handed in after the official due date, and without an agreed extension due to special considerations, a **5%** penalty applies to the student's mark for **each day after the due date** (including weekends, and public holidays) for up to 7 days. Assessment items handed in after **7 days** will not be considered.

Assessment Criteria & Grading

The following outlines the criteria which you will be assessed against.

- Demonstrated understanding of the tasks
- Ability to wrangle and process using appropriate R code
- Ability to analyse data using appropriate R code
- Ability to visualise the data using appropriate R code
- Ability to interpret the analysis and visualisations to complete the tasks
- Written communication skills, including explaining the considerations and steps required to complete the tasks, citing sources when required