Problem Solving Task 1 Instructions

SIT741 Problem Solving Task 1

Jnit Chair: Sergiy Shelyag

Due: 27 August 2021

Problem Solving Task 1 contributes to 20% of your final SIT741 mark. The full mark is 100. It must be completed individually and submitted to CloudDeakin before the due date: 8 pm, 27/08/2021 (Week 6 Friday).

Learning goals

In this assignment, you will work on a real-world problem to consolidate your learning in the first five weeks, including organise your data as tidy data and perform simple statistical analyses. This activity also serves as scaffolding for the upcoming Assignment 2.

Please start early so that you can identify any skill/knowledge gap and seek support from the teaching staff and other students.

Background

In Australia, we have experienced extreme heat in the year 2019. With the inevitable rise of extreme weather events, it is crucial that we better understand its potential impact on our everyday life.

In November 2016, a storm in Victoria triggered an unexpected surge of emergency department visits at the local public hospitals. Some consequences of this weather event were captured in this news article:

```
http://bit.ly/2gC8j6U
```

Apart from such storms, various weather events may affect the demand for care at our emergency departments (EDs). In SIT741, you will use publicly available data to understand the relationship

between weather patterns and ED demands. Your analysis could provide crucial knowledge for resource planning at our health care systems.

Assignment 1 will focus on the analysis of ED demand data.

Task 1: Obtaining ED demand data (16 points)

First, let's find data measuring ED demands. We will use the *emergency departments admissions and attendances* data set provided by the Department of Health of Western Australia:

```
http://data.gov.au/dataset/emergency-department-admissisons-and-attendances
```

Task 1.1 Download the data set using the link below **(4 points)**.

```
http://bit.ly/2nkCUEh
```

Task 1.2 Answer the following questions:

- How many rows and columns are in the data? **(1 point)**
- How many hospitals are in the data? **(1 point)**
- What data types are in the data? (Use data type selection tree and provide detailed explanation) **(2 points for data types, 2 points for explanations)**
- What time period does the data cover? **(1 point)**
- What's the difference between "Attendance" and "Admissions"? **(3 points)**
- What do the variables `Tri_1`, `Tri_2`, … represent? **(2 points)**

Hint: You may need to consult the relevant dataset description (see the link above).

Task 2: Tidy data (20 points)

Task 2.1 Cleaning up columns

You may notice that the ED csv file has two rows of heading. This is quite common in data generated by BI reporting tools. Let's clean up the column names.

```
ed_data_link <- 'govhack3.csv'
top_row <- read_csv(ed_data_link, col_names = FALSE, n_max = 1)
second_row <- read_csv(ed_data_link, n_max = 1)

column_names <- second_row %>%
  unlist(., use.names=FALSE) %>%
  make.unique(., sep = "__") # double underscore

column_names[2:8] <- str_c(column_names[2:8], '0', sep='__')

daily_attendance <-
  read_csv(ed_data_link, skip = 2, col_names = column_names)
```

Now print out a list of healthcare facilities (hospitals) in the data set. **(1 point)**

Task 2.2 Tidying data

1. Now we have a data frame. Answer the following questions for this data frame.

- Does each variable have its own column? **(1 point)**
- Does each observation have its own row? **(1 point)**
- Does each value have its own cell? **(1 point)**

2. Use spreading and/or gathering (or their pivot_wider and pivot_longer new equivalents) to transform the data frame into tidy data **(6 points)**. The key is to put data from the same measurement source in a column and to put each observation in a row. Please answer the following questions.

- How many spreading (or pivot_wider) operations do you need? **(1 point)**
- How many gathering (or pivot_longer) operations do you need? **(1 point)**
- Explain the steps in detail. **(3 points)**

3. Are the variables having the expected variable types in R? Clean up the data types. **(3 points)**
4. Are there any missing values? Fix the missing data. Justify your actions. **(2 points)**

Task 3: Exploratory Data Analysis (20 points)

It is often a good idea to visually check your data before fitting a model. The purpose is to understand the distribution of different measurements and relations between them.

Task 3.1 Select a hospital

Task 3.1 Select a hospital

Select a hospital and create a dataset for only the selected hospital. **(1 point)**

Print out the hospital's name **(1 point)**, the total number of ED attendances **(1 point)**, and the total number of admissions **(1 point)**.

Check if the total number of ED attendances corresponds to the total number of triaged patients and print the difference **(2 points)**.

Task 3.2 For the hospital selected, if we want to compare the volume of ED demands across the year, which plot can we use? Show your plot and explain what the plot shows. (Hint: Which variable measures the ED demands?) **(3 points)**

Task 3.3 How do the ED demands change during a week? Show it visually using violin plots **(2 points)**, describe the results **(2 points)** and provide your interpretation **(2 points)**.

Task 3.4 Use `skimr` and `fitdistrplus` libraries to answer the following questions. Which distributions are appropriate for modelling the ED demand? **(1 point)** Which variables meet the assumptions for the Poisson distribution and why? **(2 points)** To reduce the dependence between consecutive days, randomly sample 150 records out of the whole dataset (all records for the selected hospital) for modelling **(2 points)**.

Task 4: Fitting distributions (20 points)

As you may have seen in the previous step, although we are dealing with count data, a Poisson distribution may not provide a good fit. Actually, unconditional Poisson distribution is too restrictive for most real-world applications. In this task, we will fit a couple of distributions to the Triage 3 attendance using the same sample of Task 3.4.

Task 4.1: Fitting distributions **(4 points)**

Fit a Poisson distribution and a negative binomial distribution on `Tri_3`. You may use functions provided by the package `fitdistrplus`.

Task 4.2: Compare distributions **(6 points)**

Compare the log-likelihood of two fitted distributions

Compare the log-likelihood of two fitted distributions.

Which distribution fit the data better? Why?

Task 4.3: Try other distributions (research question 1) **(10 points)**

Find which distributions R stats library includes. Try to fit some of them to different Triage variables. Analyse and explain the results. Write a short report (200 words).

Task 5: Research question 2 (15 points)

There are more than one ways to fit a distribution to a set of numbers. Produce a short literature review on different distribution fitting methods, showing the pros and cons of each method. **5 points** will be given to relevance of the literature. **7 points** will be given for the quality of comparative analysis of distribution fitting methods. **3 points** will be given for the quality of presentation.

Task 6: Ethics question (7 points)

During your work, have you identified any issues that have ethical implications? **(2 points)** Does it concern security or privacy? **(2 points)** How was the risk mitigated? **(3 points)**

Task 7: Reflection (2 points)

Answer the following questions:

1. What help did you receive from other students? What did you learn from them? **(1 point)**

Reflect in ePortfolio **Download** **Print**

Alternative formats

Activity Details

You have viewed this topic

Due 27 August at 8:00 PM Starts 22 July, 2021 Ends 03 September, 2021

Last Visited 14 August, 2021 11:35 PM