Data Analyst Assignment

Mahmoodur Rahman

09 June, 2023

Assignment Overview

In this assignment, you will connect to an SQLite database containing information about patients tested for MRSA, CDI, and COVID-19. You will run some analytics on the data and develop insights into the patient distribution, infection rates, and treatment effectiveness.

NOTE: Please remember to load in any packages you plan to use for your analysis.

Task 1: Connect to the Database

Your first task is to connect to the SQLite database infection.db using the RSQLite package. Once you've established a connection, list the tables in the database. You can find the SQL file located in the same zip file as this assignment.

Table 1: List of tables

Name of Dataset
CDI
COVID19
MRSA

Task 2: Load the Data into Data Frames

Next, write a query to select all data from each of the three tables: MRSA, CDI, and COVID19. Store the result of each query in a separate data frame.

```
# saving tables according to names followed by "_df"
for (df.name in DBI::dbListTables(con)) {
  assign(paste0(df.name, "_df"),
         DBI::dbReadTable(con,df.name))
}
```

Variable Description:

- patient_identifier: the unique patient id
- age: current age in years
- unit: the name of unit where patient was staying at when got tested, format: "floor level - facility name" (There are 10 different facilities in the tables)
- room: the room number where patient was staying at when got tested

- bed: the bed code where patient was staying at when got tested
- result: the test result: Positive or Negative
- treatment: the medicines patient was treated

Task 3: Analyze the Data

Now, perform some basic analysis on each data frame. This could include (but not limited to):

1. Count the number of patients in each facility.
2. Identify the unit with the highest number of positive results in each facility.
3. Calculate the infection rate for each disease (infection rate = number of positive results/total number of results).

Feel free to be creative during this section and come up with some unique ways to interpret and analyze the data.

Table 2: Number of patients in each facility

Facility Name	Number of patients in CDI dataframe	Number of patients in Covid-19 dataframe	Number of patients in MRSA dataframe
Blossom Community Hospital	549	435	175
East Valley Medical Clinic	250	1339	201
Featherfall Medical Center	890	697	798
Goldvalley Medical Clinic	605	560	569
Hyacinth General Hospital	1526	998	416
Kindred Medical Clinic	583	318	125
Magnolia Hospital	763	1883	613
Maple Valley Medical Center	535	633	364
Rosewood Hospital Center	988	1283	435
Ruby Valley Hospital	377	342	148
Summit Community Hospital	508	777	503
Wildflower Medical Center	426	735	653

Task 4: Develop Insights

Based on your analysis and visualization, develop some insights into the data. Discuss any patterns or trends you observed, any surprising results, and any limitations or potential improvements to your analysis.

```
# Number of patients attended largerly varies by facility and by disease itself
# In almost all facilities, the number of Covid-19 patients are more in number, compared to other two d
# Among the two diseases primarily caused by antimicrobila resistance, number of CDI patients are more
# In case of number of test-postive patients in a facility, we can see that it differs largely between
# We have a tie in number of positive patients in unit floor 2 nad floor 4 in Ruby Valley Hospital in c
# Interestingly MRSA patients are the least in numbers attended by the facilities, but the disease has
```

Table 3: Units with the highest number of positive results in each facility

Facility name	CDI Data		Covid-19 Data		MRSA Data	
Facility name	Units with most positive patients	Number of patients	Units with most positive patients	Number of patients	Units with most positive patients	Number of patients
Summit Community Hospital	Floor 1	38	Floor 5	51	Floor 1	46
East Valley Medical Clinic	Floor 2	19	Floor 4	168	Floor 3	24
Kindred Medical Clinic	Floor 2	35	Floor 1	25	Floor 1	15
Maple Valley Medical Center	Floor 2	35	Floor 1	50	Floor 1	29
Ruby Valley Hospital	Floor 2	39	Floor 2	30	Floor 4	27
Ruby Valley Hospital	Floor 2	39	Floor 4	30	Floor 4	27
Blossom Community Hospital	Floor 4	26	Floor 5	43	Floor 4	23
Magnolia Hospital	Floor 4	45	Floor 2	101	Floor 2	92
Rosewood Hospital Center	Floor 4	67	Floor 5	66	Floor 2	43
Wildflower Medical Center	Floor 4	19	Floor 3	61	Floor 4	115
Featherfall Medical Center	Floor 5	71	Floor 3	55	Floor 3	131
Goldvalley Medical Clinic	Floor 5	38	Floor 2	32	Floor 1	56
Hyacinth General Hospital	Floor 5	161	Floor 4	53	Floor 1	46

Table 4: Infection rate for each diseas

Diseases	Total Number of Tests	Number of Positive Tests	Infection Rate
CDI	8000	1222	0.15
Covid-19	10000	1721	0.17
MRSA	5000	1514	0.30

Submission Methods

You can submit your assignment as a RMarkdown file, PDF, HTML, or Word document.

To generate a PDF, HTML, or Word document from your RMarkdown file, click on the "Knit" button in the RStudio toolbar and select the output format you want.

If you choose to submit PDF, HTML, or Word document, please make sure we can see both the results and your code(with echo = TRUE in your RMarkdown chunks). This is crucial for our review process.

Alternative Submission Methods and Programming Languages

We acknowledge the diverse skills of our participants, so you are welcome to complete this assignment using the programming language of your choice. The provided RMarkdown file is simply a guide. You may choose to use Python, Julia, SQL, SAS, or any other language that you are comfortable with and that can effectively interact with the SQLite database.

The medium of your assignment submission can also vary. If you choose not to use the provided RMarkdown file, please ensure that your chosen medium is clear, well-organized, and accessible for the selection committee. Some examples of alternative submissions include:

Hosting a temporary website: Create a website to present your assignment. The website should clearly display your code, output, visualizations, and insights. Ensure that the website is live and accessible by the selection committee until the review process is complete.

Sharing a GitHub repo: You can push your code, along with any generated outputs and visualizations, to a public GitHub repository. Make sure to include a README file explaining the structure of your repo and how to run your code.

Sharing a Kaggle notebook: You can develop your assignment in a Kaggle notebook. Make sure the notebook is public and can be accessed by the selection committee.

Please include the link to your website, GitHub repo, Kaggle notebook, or any other chosen medium when you submit your assignment.