Name: Chioma Sandra Achugonye
Course: Digital Health Exam

PROBLEM DESCRIPTION

The problem statement here is to understand the distribution of ICU beds statewide and countywide across the United States of America. With such knowledge, stakeholders such as the US government will be able to effectively manage resources in terms of providing ICU beds to the most vulnerable states and counties. Additionally, stakeholders will also want to understand the number and distribution of older people who need intensive care.

I used a dataset from Kaggle but compiled by the Kaiser Health Network, comprising the number of ICU beds across different states and counties in the USA. It was evaluated by Kaiser Health News using the number of ICU beds reported by each hospital in their annual financial cost report filed with the Centers for Medicare and Medicaid Services. KHN included beds reported in coronary care, burn intensive care units, surgical intensive care units, and general intensive care units. The total number of ICU beds reported per county was then matched with the county population figures from a survey conducted by the Census Bureau. This dataset focuses on the population aged 60 and above in different counties, including the percentage of county residents aged 60 and above. This is because a significant number of people within that age bracket are weaker, have existing health conditions, and are more likely to be hospitalized compared to the younger population.

However, the survey does not include Veteran Affairs hospitals because those hospitals do not file cost reports. The number of ICU beds filed on the cost reports was therefore less than the number of ICU beds identified by the American Hospital Association's annual survey of hospital beds. This dataset was compiled four years ago in the wake of the COVID-19 pandemic.

PATHWAY TO SOLUTION

To be able to address the challenges faced by the stakeholders, I decided to approach the data analysis in a methodological manner as stated below.

- Define the Problem:

The first step I took in solving the above-mentioned problem was to clearly define the problem statement. This involves understanding the objectives, identifying the stakeholders, and defining the scope of the analysis. For example, in the aforementioned data, the objectives were to access the ICU bed distribution across US and correlate it with other parameters such as population.

- Data Collection:

Once the problem was defined, the next step I took was to collect and understand the relevant data. This involved collecting data from sources such as Kaggle which contains a lot of healthcare records associated with ICU bed and population

distribution across the US. Furthermore, I ensured the quality and reliability of the data collected to avoid inaccurate or incomplete data which can lead to erroneous conclusions.

- Data Preprocessing:

Before starting my analysis, I preprocessed the collected data. This included tasks such as cleaning the data to remove errors or inconsistencies, transforming the data into a suitable format for analysis, and handling missing values. Data preprocessing is crucial for ensuring the accuracy and reliability of the analysis results. I did this by first viewing the data on an Excel sheet where I found out that some observations on the last column (residents aged 60/each ICU bed) were recorded as null and the actual integer value =0. I changed all the 'nulls' to 0.

- Exploratory Data Analysis (EDA):

Exploratory data analysis involves examining the data to understand its characteristics, identify patterns, and uncover insights. This may include visualizing the data through charts or graphs, calculating summary statistics, or conducting hypothesis testing. EDA helps to gain a deeper understanding of the data and guide subsequent analysis steps.

- Data Analysis Techniques:

During the data analysis process, I utilized various analysis techniques to extract insights and solve the problem at hand. This involved descriptive statistics to summarize the data and correlation matrix to understand relationships between variables. This allowed me to draw conclusions about the data I had.

- Interpretation and Evaluation:

After performing the analysis, it's essential to interpret the results and evaluate their validity and relevance. This involves assessing the robustness of the analysis methods used, considering potential biases or limitations in the data, and critically evaluating the implications of the findings. Because I have the domain knowledge, I was able to make valuable interpretations of the data to the stakeholders as presented in my final interactive HTML file.

- Communication and Visualization:

Effective communication of the analysis findings is crucial for driving action and decision-making and this is what I will present on my exam day. This involved creating visualizations such as charts, graphs, and interactive HTML files to present the results in a clear and understandable manner.

- Implementation and Action:

Finally, the insights gained from the data analysis need to be translated into action. This may involve implementing changes such as allocation of more resources to areas with low ICU beds and areas with elderly populations.

- Conclusion:

Solving a data analysis problem involves a systematic and iterative process that begins with defining the problem and ends with implementing actionable solutions. By following a structured approach and leveraging appropriate data analysis techniques, stakeholders can unlock the value of data to drive positive outcomes and

address complex challenges effectively. Continuous learning and adaptation are essential in navigating the evolving landscape of data analysis and decision-making.

IMPLEMENTATION

My implementation phase is divided into the following steps.

- Programming language, file type and presentation medium
- Practical steps
- Errors encountered

Programming language: I used R as the programming language of choice. This is because it is the language, I learnt in the Digital Health Course. Also, I find it to be a powerful tool in terms of data visualization. Additionally, my data storage file is Excel CSV file type and my final output file was HTML file generated from R markdown.

Practical steps: I employed different steps during the data analysis process as shown below.

- **Setting working directory** – Firstly, I set the working directory by using setwd () and pasting the file part into the brackets.

- **Importing ICU data** – I imported the ICU data by using the function read.csv () because the ICU data was as a .CSV file.

- **Package installation** – When it comes to installation of packages needed, it was done as the data analysis was going on. However, at the end of my analysis I placed all the packages I at the beginning of my codes.

- **Initial data exploration** – In this section took an overview of my data by using functions like head (), summary (). In addition, I used dim (), glimpse () to look at the dimension of my data frame. To check for the classes of the variables, I used sapply (ICU, class). Furthermore, I checked for missing values using is.null() function and length/unique names of state and county using length(unique()) function.

- **Correlation** – In order to understand the relationship between the variables, I employed the correlation matrix by using corrplot () function.

- **Data analysis** – I deployed ggplot2, tidyverse, dplyr packages to plot bar chats representing the various parameters I was interested in providing a solution to the challenges of the stakeholders. This analysis was done at the state and county level.

- **Data presentation** – The final deliverable of my entire data analysis process was a dynamic HTML file which was the product of R markdown. To generate

this dynamic HTML file, I transferred my codes into an ". Rmd" environment. Afterwards, I ran the individual block of codes then clicked on "knit" function to generate the dynamic HTML file.

- **Storage** – All my data outputs and the codes I used were stored in my GitHub page and the link provided for direct access to them by the Professor.

Errors encountered: During the process of data analysis, I encountered several error messages which is shown in the table below and what I did to solve it.

Errors	Issue/Solution
Error in `mutate()`: ℹ In argument: `State = fct_reorder(State, n)`. **Caused by error in `fct_reorder()`:** ! could not find function "fct_reorder"	Forecast package not installed. I installed it.
Error in `geom_text()`: ! Problem while computing aesthetics. ℹ Error occurred in the 1st layer. **Caused by error in `compute_aesthetics()`:** ! Aesthetics are not valid data columns. ✖ The following aesthetics are invalid: ✖ `label = n` ℹ Did you mistype the name of a data column or forget to add `after_stat()`?	Mis-typed one of the variables. I corrected the variable spelling.
Error in `select()`: ! Can't subset columns with `Total_population`. ✖ `Total_population` must be numeric or character, not a <tbl_df/tbl/data.frame> object. Run `rlang::last_trace()` to see where the error occurred.	I wrote `Total_population` instead of `Total.population`. Wrong spelling and I corrected it.
Error in x[j] : invalid subscript type 'closure'	I used a wrong subscripting type - "head" instead of "head ()".
Error in lowest_ICU_beds[1:5,] : incorrect number of dimensions	Wrong dimension. I fixed the dimension by making sure i check my dataframe and i selected a wrong dataframe.
Error: unexpected ']' in "lowest_ratio_population_above_60 <- ICU[order(ICU$Residents.Aged.60..Per.Each.ICU.Bed, decreasing = FALSE,]"	Omitted a bracket. I added the bracket
ℹ The package "maps" is required for `map_data()` ✖ Would you like to install it?	I installed "maps" package.
Error: object 'County' not found	Wrong selection of column. I selected the right one.
could not find function "theme_map"	Install the package "ggthemes"
Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0. ℹ Please use `linewidth` instead	Replaced "size" with "linewidth"