Final Report

Nake Shrivatsa Aparanji1

1 University of Memphis, Memphis, TN - 38152, USA  
nspranji@memphis.edu

**Abstract.** In the wake of the Covid-19 pandemic, our healthcare systems face unprecedented challenges. The need for informed decision-making has never been more crucial. Leveraging a blend of data science techniques, statistical analysis, and machine learning algorithms, this analysis addresses crucial questions related to hospitalizations, ICU bed availability, and inpatient bed utilization at the state level. The project identifies geographical hotspots, assesses resource needs, explores age-related impacts, and establishes correlations between key variables.

1. Introduction

This project focuses on unraveling the complex patterns of Covid-19 impact on healthcare systems across U.S. states, utilizing the 'Covid-19 Reported Patient Impact and Hospital Capacity by State' dataset provided by the U.S. Department of Health and Human Services. Scheduled for weekly updates, the dataset's August 28, 2023 version serves as the foundation for my analysis. I aim to address key questions pertaining to Covid-19 hospitalizations, ICU bed availability, and inpatient bed utilization at the state level.

The questions I chose to guide this exploration are diverse yet pivotal -

* Identifying Hotspots: Are there geographical clusters with notably high Covid-19 hospitalizations?
* Resource Allocation: Which states might require additional resources to effectively combat the pandemic?
* ICU Bed Utilization: Are there regional variations in the utilization of ICU beds?
* Age and Hospitalizations: Does age influence Covid-19 hospitalizations, and how do pediatrics and adults differ?
* Bed Occupancy and Deaths: Is there a correlation between occupied beds and recorded deaths?
* Predictive Modeling: Can we predict future patient admissions, considering state variations?
* Temporal Correlations: Is there a link between suspected cases one day and confirmed cases the next?

In my analysis of Covid-19 impact on U.S. healthcare systems, California stood out with the highest hospitalizations, while North Carolina, South Carolina, Florida, Georgia, Virginia, and Pennsylvania formed a notable cluster of significant hospitalization numbers. California and Michigan were identified as states requiring additional staff in the upcoming week. Rhode Island showed the highest inpatient bed utilization, suggesting a need for resource optimization, while Wyoming exhibited the least utilization. In ICU bed utilization, Arizona and Maryland led in adults, while California and Texas topped the list for pediatric cases. Patients aged '80+' had the highest hospitalization rates, contrasting with the lowest rates observed in the '18-19' age group. Notably, a strong positive correlation emerged between “deaths\_covid” and “inpatient\_beds\_used”, signifying a close association, while “suspected” and “confirmed” Covid-19 cases displayed a weak positive correlation.

Drawing inspiration from Related Work, the Design section details the use of Pearson correlation coefficients, visualization with ggplot, and significance testing. The Evaluation section answers specific questions through visualizations and statistical analysis, providing insights into regional variations and correlations.

1. Related Work

JMIR Public Health Surveillance made this research on ‘Estimating Covid-19 Hospitalizations in the US with Surveillance Data Using a Bayesian Hierarchical Model’. The objective of the study was to develop a method that utilizes surveillance data to estimate monthly hospitalization rates for Covid-19 over an extended period. Researchers estimated monthly Covid-19 hospitalization rates for different age groups across 50 states from May 2020 to April 2021. They used data from the Covid-19 Associated Hospitalization Surveillance Network (COVID-NET) and implemented a Bayesian Network for analysis. The analysis was done separately for 6 age groups and a covariate selection was done using LASSO and spike & slab techniques. Validation was done by assessing sensitivity to covariate selection. This approach was concluded to have the potential to monitor Covid-19 burden.

1. Design

(I have a couple of questions that analyses correlation between features) Correlation is a statistical measure of how related or un-related two or more variables/features are. Correlation coefficient is measure that quantifies the correlation concept, it tells us to what degree two or more variables are related/un-related. Although there are many correlation coefficient types, I will be using the Pearson correlation coefficient. There are a couple of reasons for this – firstly, Pearson correlation coefficient is the most widely used type and secondly, R implements this by default. Pearson coefficient takes values between -1 and +1, the closer the value is to -1 or +1 stronger the relationship (- meaning negative association while + meaning positive association). To visualize the correlation concept, I used a correlation plot (created using the “corrplot” package in R), that represents the correlation matrix of variables. Along with that I took help of heat map which uses color intensity and shape to convey the strength and direction of relationships between pairs of variables. Positive correlations are typically represented by one color gradient, while negative correlations are represented by another.

(To answer questions that required comparisons between columns/variables, visualization is a must) “ggplot” is a versatile and powerful data visualization package in R, providing a declarative syntax for creating a wide range of static and dynamic visualizations. The package allows users to construct plots layer by layer, specifying data aesthetics, geometries, and themes. I chose this package due to its flexibility and modularity in communicating insights effectively through visually appealing graphics in R. I have created bar plots, scatter plots and dot plots using “ggplot”.

Significance testing is done in multiple occasions enabling me to assess the validity of observed relationships and draw meaningful conclusions about population parameters. At the core of this process lies the concept of p-values, representing the probability of obtaining results as extreme or more extreme than the observed ones, assuming the null hypothesis is true. A p-value below a chosen significance level (I chose 0.05, i.e., with 95% confidence) indicates that the observed data provides sufficient evidence to reject the null hypothesis in favor of the alternative. This rigorous statistical approach ensures a systematic and objective evaluation of hypotheses, allowing for confident interpretations of the data and reinforcing the robustness of my analytical findings.

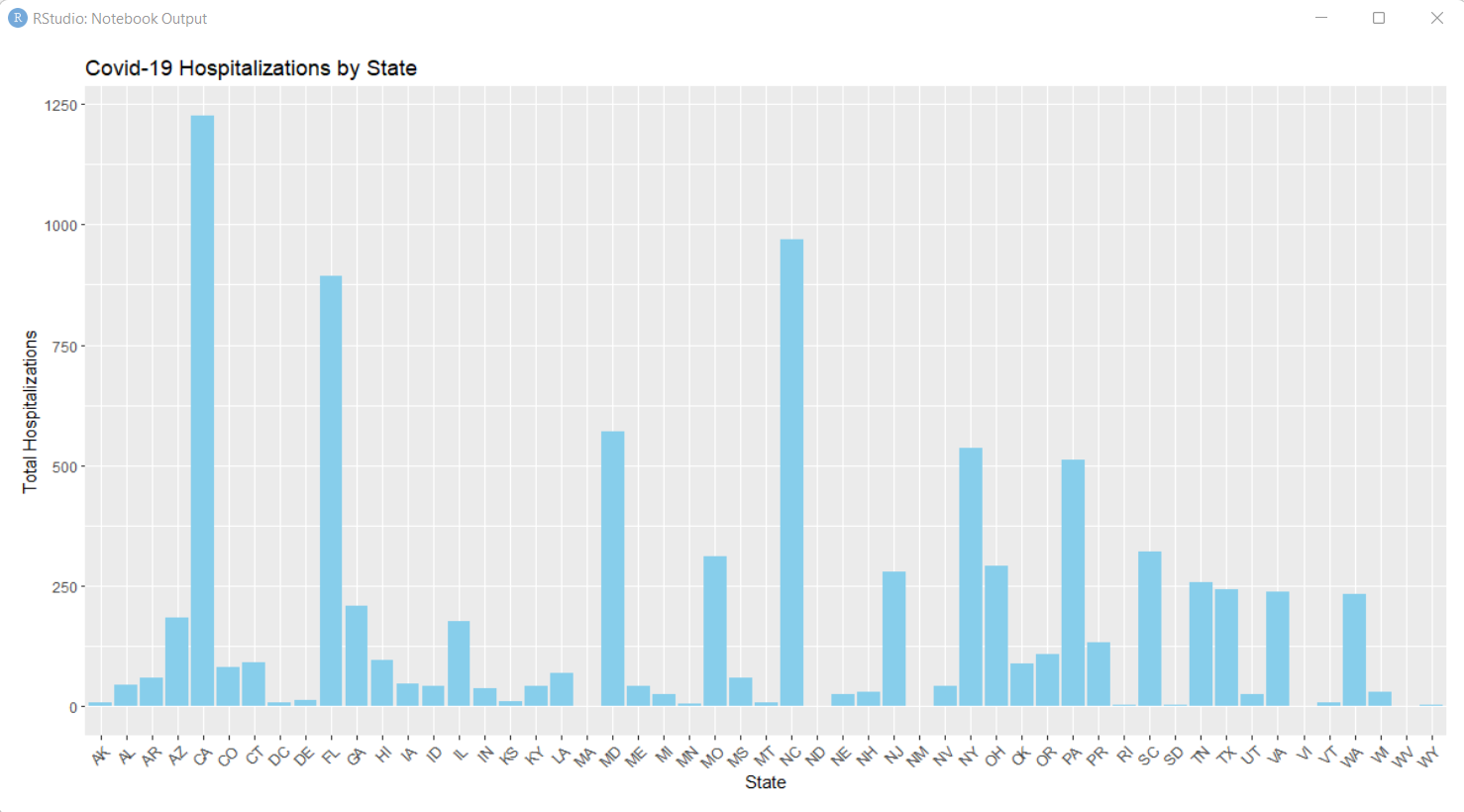
(To explore the relationship between age and Covid-19 hospitalizations) Mean and sum counts were instrumental in summarizing data by age group, providing valuable insights into the potential impact of age on Covid-19 hospitalization rates. The mean counts offered a representative average for each age group, allowing for a subtle understanding of the central tendency within the data. Simultaneously, sum counts provided a cumulative measure, reflecting the overall magnitude of hospitalizations within each age category. By examining both metrics, I gained a comprehensive view of the distribution and scale of Covid-19 hospitalizations across different age groups. Visualization using box plot was done to understand the relationship between pediatrics and adults’ Covid-19 hospitalizations. Box plot provides a concise summary of key statistical measures, allowing for a quick and effective comparison of different columns/variables. The plot consists of a rectangular "box" that represents the interquartile range (IQR) of the data, with a line inside marking the median. Whiskers extend from the box to the minimum and maximum values within a specified range, determined by a factor of the IQR. Outliers, individual data points significantly distant from the rest, are typically shown as dots or other symbols. This visualization is valuable for identifying the central tendency, spread, and skewness of a dataset, making it particularly useful for comparing distributions and detecting variations between groups/columns (which is exactly what I desired).

(On designing a predictive model to predict the patients getting admitted in the near future, with respect to each state) Pre-processing was the major step in designing a predictive model.

* From Appendix 7.1 – (a), the code snippet employs the train\_test\_split function to partition a dataset into training and testing sets, a fundamental step in machine learning model development. The dataset, denoted by variables x and y, represents the features and corresponding labels, respectively. I have specified test\_size=0.2, meaning that 20% of the data will be reserved for testing, while the remaining 80% will be used for training the machine learning model. This division helps evaluate the model's performance on unseen data, providing a more accurate assessment of its generalization capabilities. The random\_state=42 (can be any value) parameter ensures reproducibility, ensuring that the same split is achieved every time the code is executed.
* In the Appendix 7.1 – (b), I have addressed the missing values in the target variable through a process of imputation using the “SimpleImputer” class, a common technique in data pre-processing for machine learning. The target variable (y), is subjected to imputation to replace any missing values. The imputation strategy employed here is the use of the mean value of the non-missing entries within the target variable. The “SimpleImputer” is configured with the strategy parameter set to 'mean', indicating that the missing values will be replaced with the mean of the available values. The “fit\_transform” method is applied to the training set (y\_train), and the “transform” method is used for the test set (y\_test). The reshaping operations (values.reshape(-1, 1)) are employed to meet the input requirements of the imputer.
* In Appendix 7.1 – (c), code snippet initiates the construction of a column transformer for preprocessing in a machine learning pipeline. This transformer plays a vital role in preparing the dataset for model training. I have specified two distinct operations within the transformer: (a) Numerical Scaling - The ('num', StandardScaler(), x.select\_dtypes(include='number').columns) configuration designates that numerical features, identified by their data type, will undergo standard scaling. Standard scaling transforms numerical variables to have a mean of 0 and a standard deviation of 1, ensuring that their magnitudes do not unduly influence my machine learning model. (b) One-Hot Encoding for Categorical Variables: The ('cat', OneHotEncoder(handle\_unknown='ignore'), ['state']) configuration indicates that categorical variable 'state' will undergo one-hot encoding. This process expands the categorical variable into a binary matrix, with each unique category represented by a separate column. The 'handle\_unknown' parameter ensures that the transformer gracefully handles any previously unseen categories during model application.
* In Appendix 7.1 – (d), I have constructed a Linear Regression model using the scikit-learn library in Python. This pipeline streamlines the data pre-processing steps and the application of the linear regression algorithm into a cohesive workflow. The pre-processor represents a series of preprocessing steps, which includes handling missing values, encoding categorical variables, and scaling features. The subsequent component of the pipeline, the LinearRegression() regressor, is responsible for fitting the linear regression model to the pre-processed data. The Pipeline is then trained on the training data (X\_train and y\_train), where X\_train represents the feature matrix, and y\_train is the target variable. The fit method orchestrates the training process, optimizing the model parameters to minimize the difference between the predicted and actual values. Following the training phase, I have applied the trained model to the test set (X\_test) to make predictions (y\_pred).

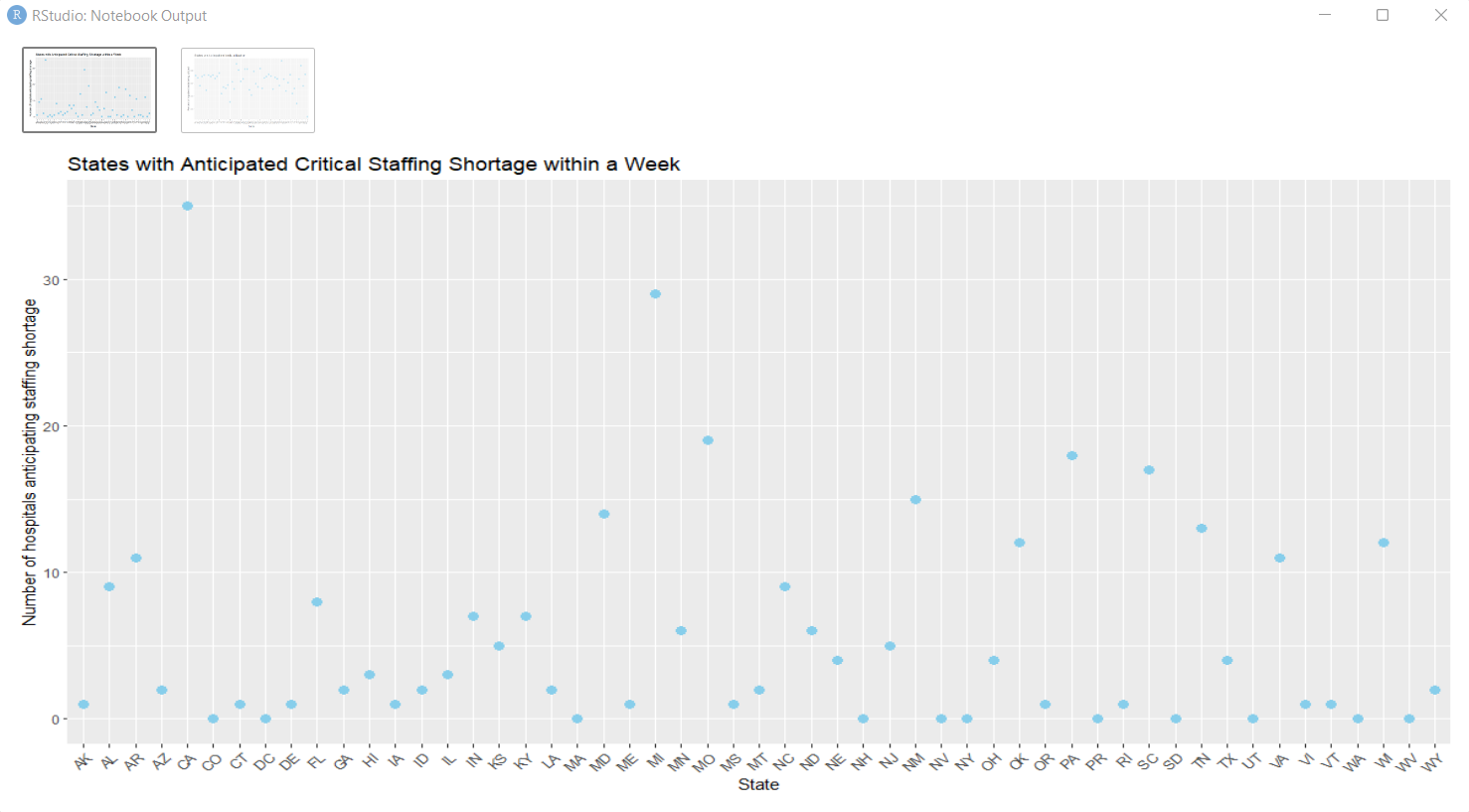
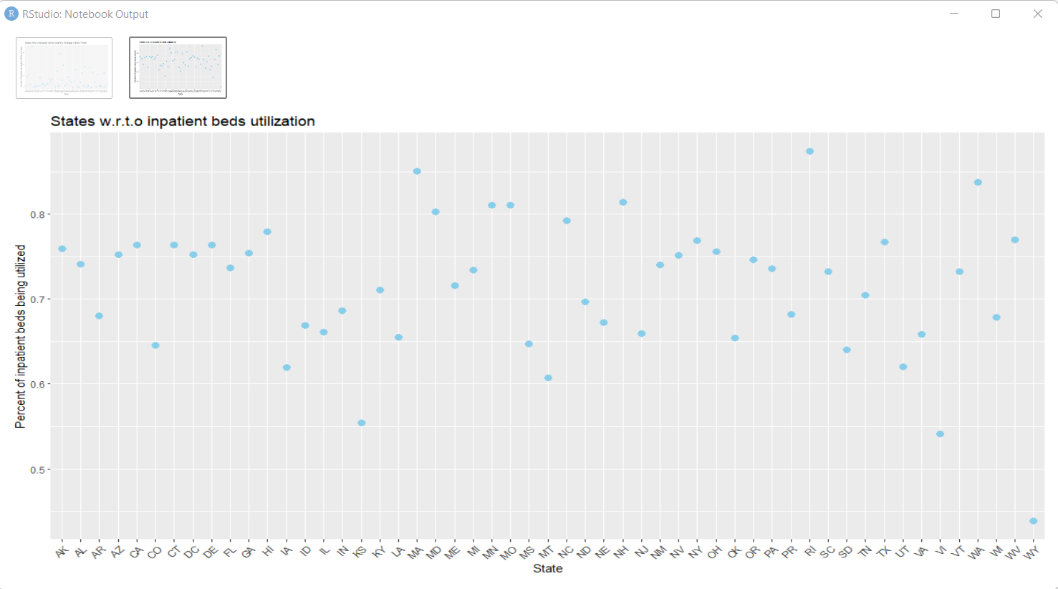
1. Evaluation

Question 1: Are there any geographical clusters of high Covid-19 hospitalizations?



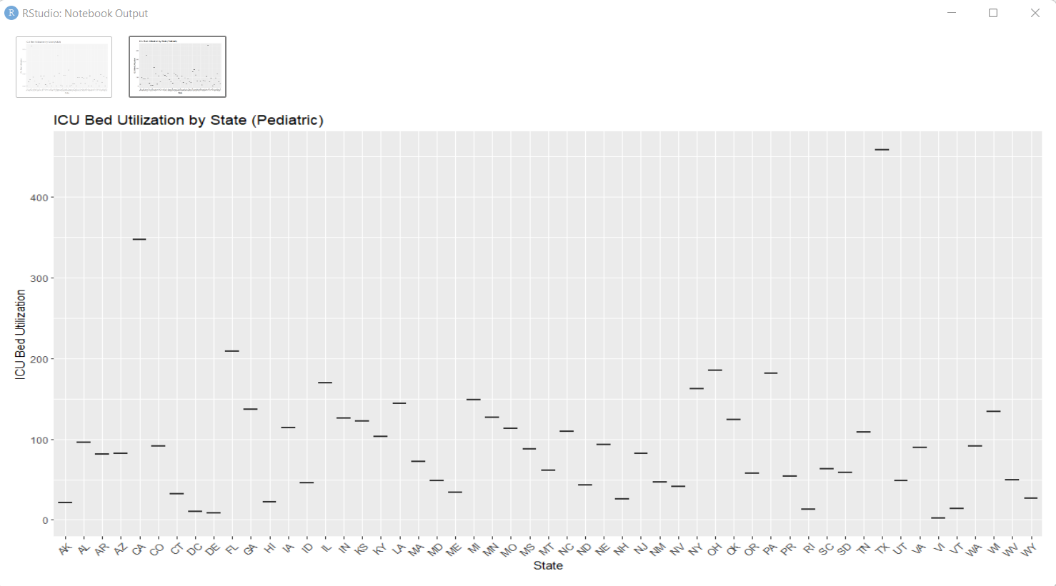
From the bar plot, it is clear that CA has the highest hospitalizations but the adjacent states have very less count. So, I believe it is not necessary that the states adjacent to a state with lot of Covid-19 hospitalizations will also have too many cases. And to answer the question, NC, SC, FL, GA, VA and PA form a good cluster of states with abundant number of hospitalizations.

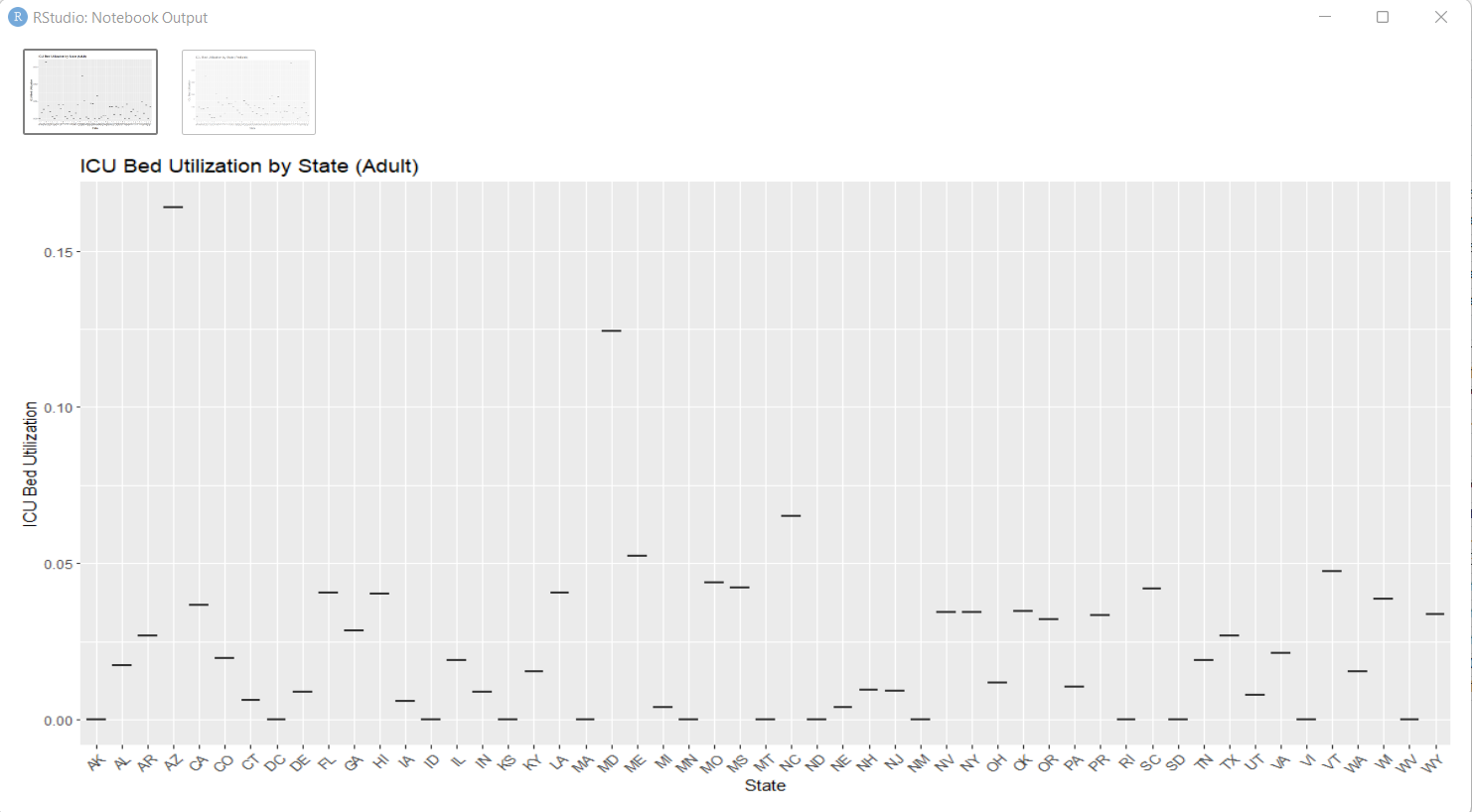
Question 2: Which regions state might need more resources to deal with the pandemic?



I have considered staff and inpatient beds as the resources needed to deal with the pandemic. Clearly, CA and MI are the states that needs more staff within the next week. RI has the highest percentage of inpatient beds utilized, while WY has the least. So, some beds from WY can be shifted to RI as backup.

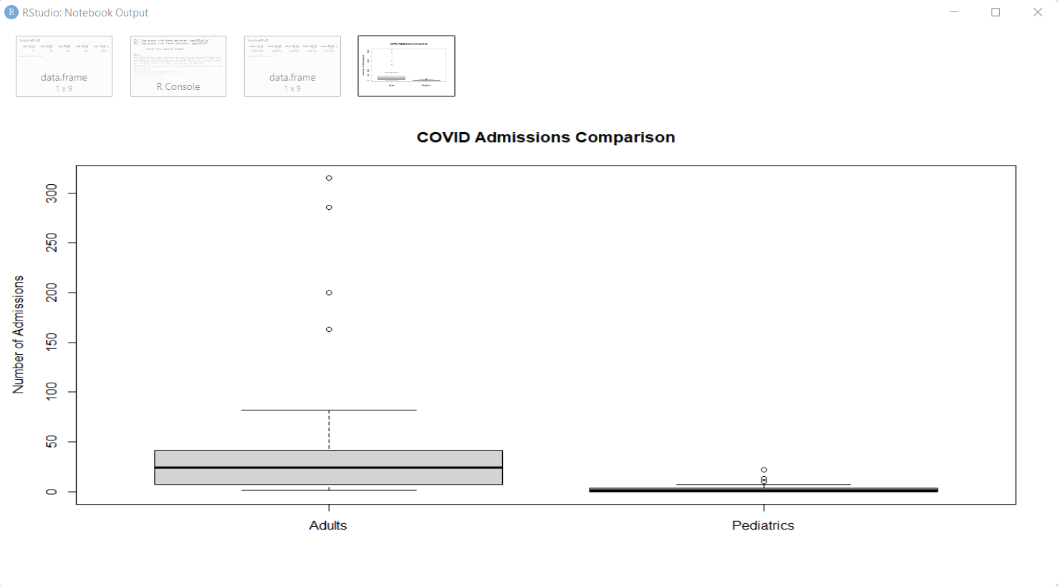
Questio 3: Any regional differences in ICU bed utilizations?



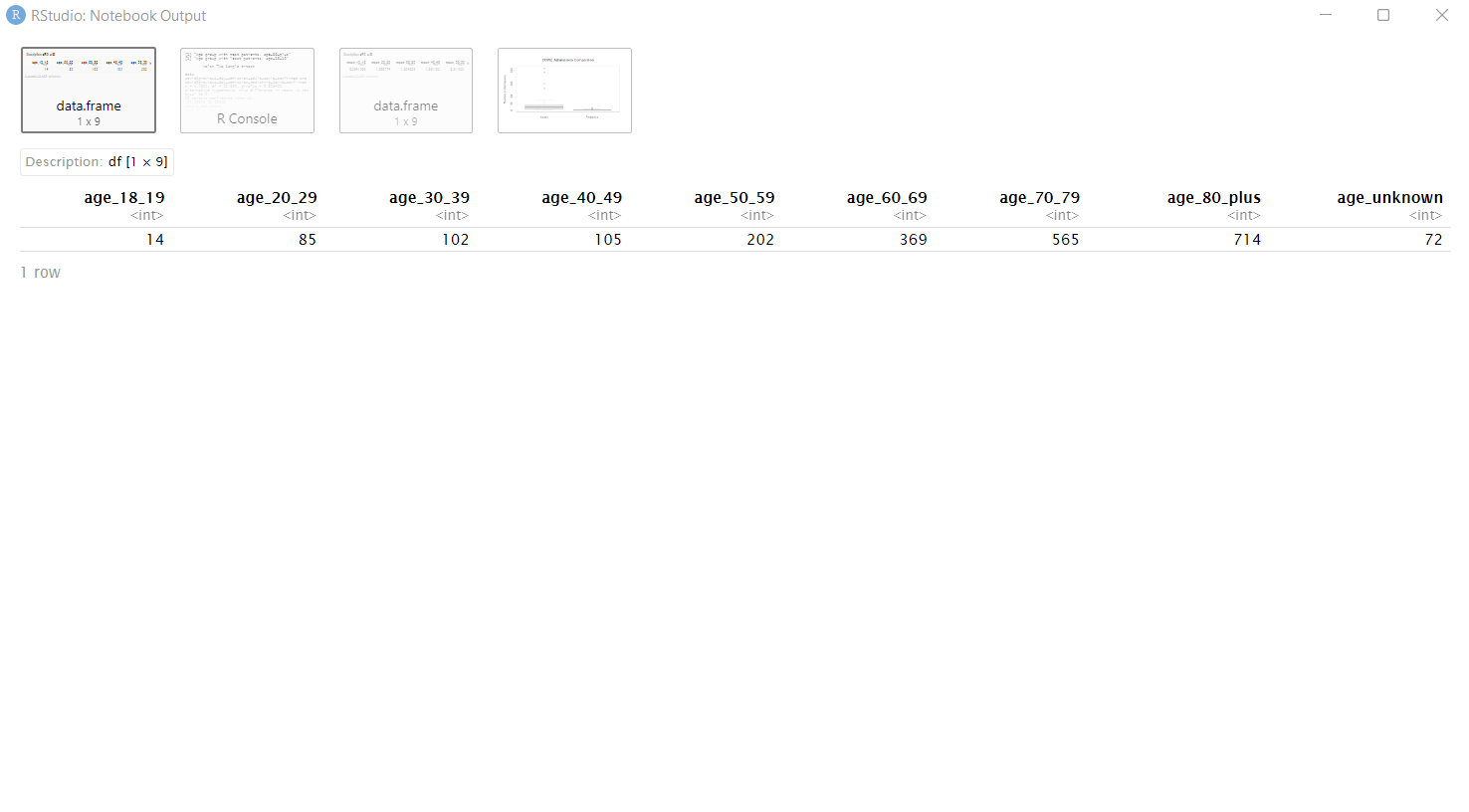


I have considered bed utilization for both adult and pediatrics. From the plot, AZ and MD are on top of the list when it comes to adult ICU beds utilization, and, CA and TX are on top of the list when it comes to pediatric ICU beds utilization.

Question 4: Do age have a say in Covid-19 hospitalizations? How different do pediatrics and adults fare?

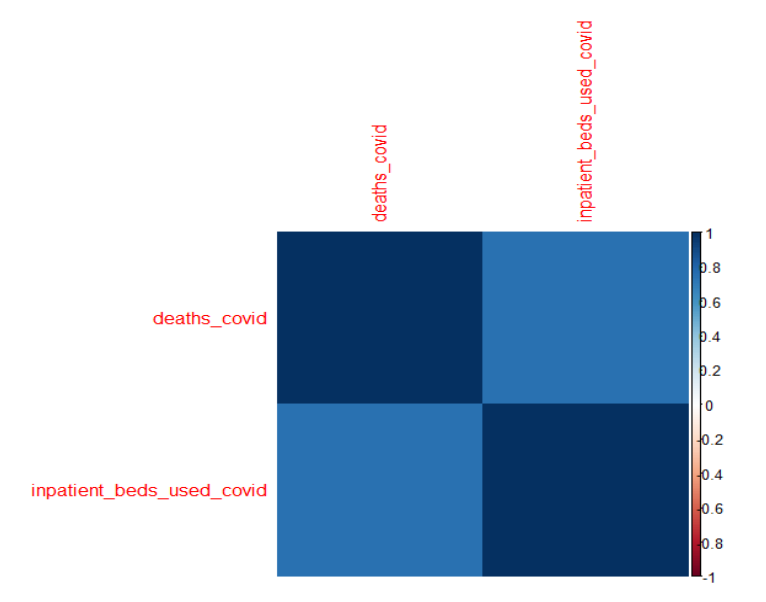


Here, it is clear that there are lot of outliers for number of admissions in adults, mainly due to the abundant number of admissions of adults when compared to pediatrics.



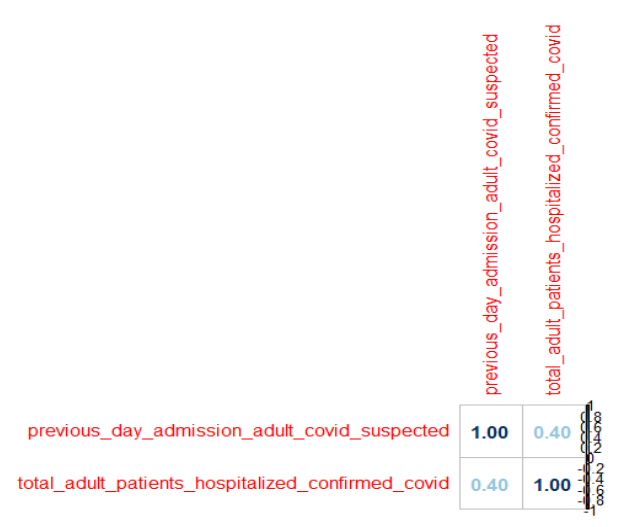
Clearly, “Age group with most patients is 80 plus" and "Age group with least patient is between 18 & 19". Additional output showing “mean counts” and results of t-testing can be found in Appendix 7.2 – 1 & 2.

Question 5: Any Correlation between beds occupied and deaths recorded?



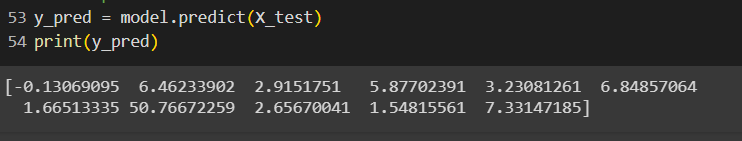
There is a strong positive correlation between beds occupied and deaths recorded as the correlation between the two variables is 0.740344. Additional visualization of “corplot” can be found in Appendix 7.2 - 4.

Question 6: Any correlation between suspected Covid-19 the previous day and the confirmed cases the subsequent day?

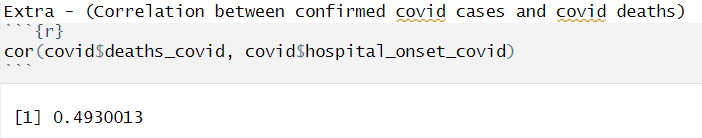


There is a weak positive correlation between suspected Covid-19 the previous day and the confirmed cases the subsequent day, as the correlation coefficient between the two variables is 0.399077. Additional output about that shows the significance testing can be found in Appendix 7.2 – 3.

Question 7: A predictive model to predict the patients getting admitted in the near future, with respect to state.



Question 8 (not included in the initial project proposal): Correlation between confirmed covid cases and covid deaths



There is a moderately positive correlation, meaning the variables are weakly associated.

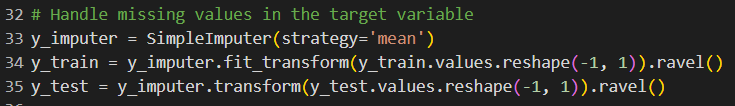
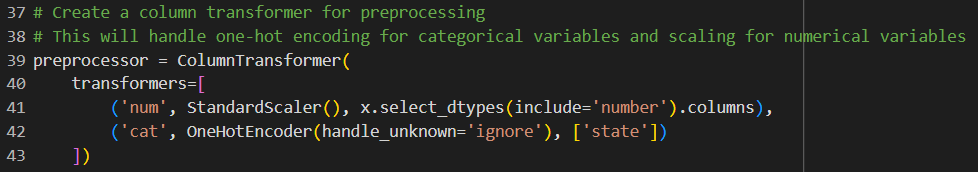
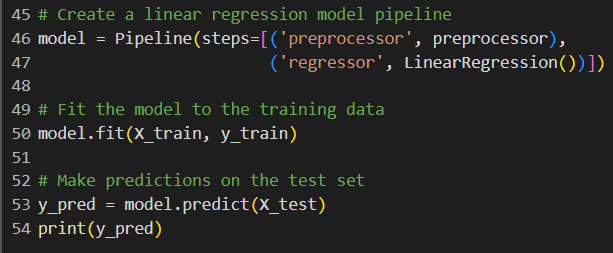
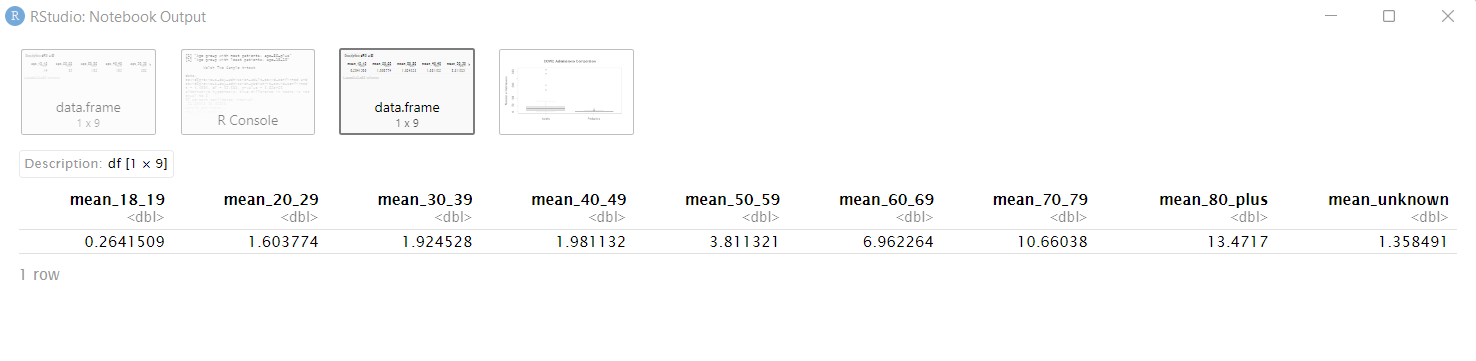
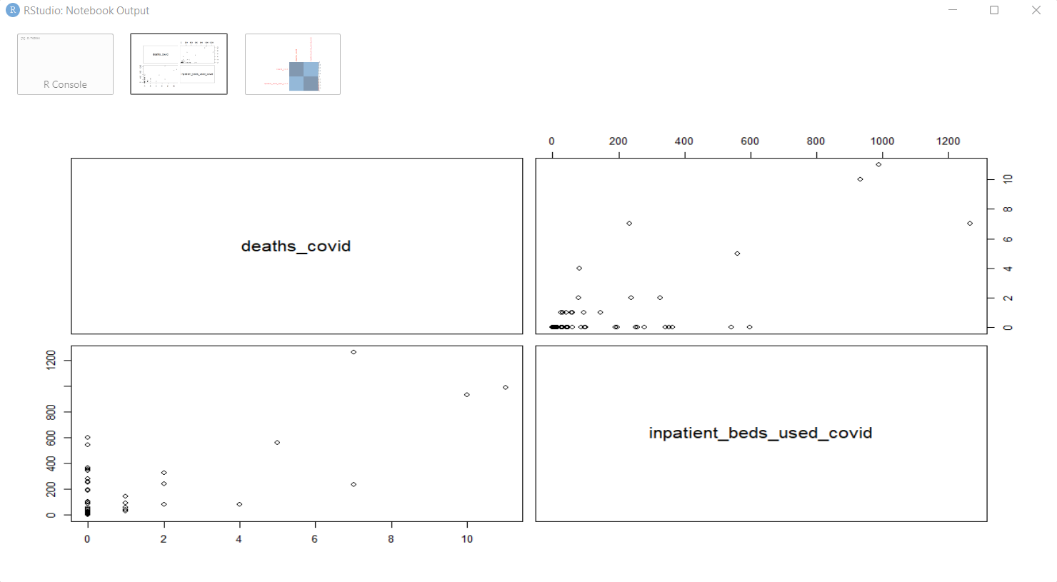
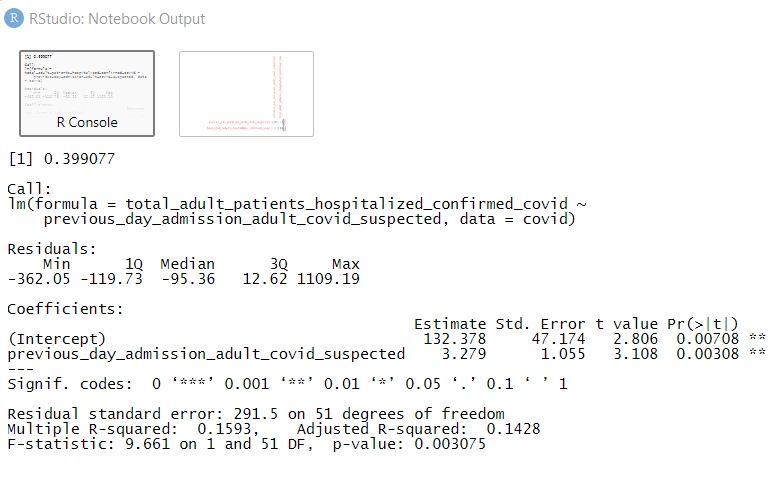
1. Conclusion

This report encapsulates a comprehensive exploration of the Covid-19 impact on U.S. healthcare, showcasing California's prominence, regional clusters, resource requirements, age differentials, and notable correlations, thereby contributing valuable insights for informed decision-making in the ongoing public health crisis.

1. Future Works

* Machine Learning for Causation: Move beyond correlation analysis to explore causal relationships. Implement machine learning algorithms designed for causal inference to identify factors contributing directly to increased hospitalizations or other outcomes.
* Temporal Analysis Enhancement: Expand the temporal analysis by incorporating more recent data beyond the August 28, 2023 version. This can provide a dynamic understanding of how Covid-19 impacts evolve over time, aiding in the identification of emerging trends and potential shifts in healthcare demands.

.

1. Appendix
   1. Code Snippets
2. 
3. 
4. 
5. 
   1. Outputs & Visualization
6. 
7. 

References

1. [https://www.ncbi.nlm.nih.gov/pmc/artices/PMC9169704/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9169704/)
2. <https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/9psv-r5iz>
3. Dataset: - <https://drive.google.com/file/d/1Od_Hr9tUdJkYxSOBbbwZ6RvKJgV22EnW/view>