**Rohan Shivlani rms477**
**George Thomas gkt20**

# Analyzing the Relationship Between Hospital Bed Capacity and COVID-19 Outcomes in New York State

Through the work we have done by analyzing datasets related to hospital bed capacity and weekly COVID-19 hospitalizations and fatalities across New York State, we aimed to identify potential patterns that may emerge from both sets. We used data from the website, healthdata.gov, and our data management and analysis techniques helped us to make potential predictions about the number of hospitalizations and fatalities based on statewide hospital bed availability. Using our research question, "What is the connection between hospital bed capacity and COVID-19 hospitalizations and fatalities, especially as it relates to New York State?", we developed a thorough and comprehensive project that incorporated various topics from throughout the course to give more insight and potentially solve this fundamental question.

We believe this project is crucial because it addresses and tackles two significant issues that may impact healthcare providers. Ever since the COVID-19 pandemic, our healthcare systems all across the country have been pushed to the brink and have been consistently working to save lives. In our case, finding a real, tangible connection between bed availability and COVID-19 statistics is imperative. Our project can help hospitals and the State of New York to make quick and life-saving decisions based on bed availability at various hospitals statewide. It can also encourage the state to provide more support for hospitals that tend to face a shortage more frequently than their counterparts, thereby alleviating the significant strain faced by the healthcare sector.

After searching online, we found prior research from the NIH that addressed this topic early in the COVID-19 pandemic. This resource was a helpful reference, as it showed us that this

topic is very important in the healthcare sector. It also showed us that there is certainly data in this field that can be parsed through and used to make life-changing decisions.

The entire project that we produced was simultaneously intended to showcase the skills and techniques that we have learned throughout the semester in a way that was positive and beneficial for society. After being assigned the project, we decided to split the work fairly and helped each other whenever it was necessary. One of us was assigned to deal with the more earlier parts of the course, ranging from the usage of Data Series and DataFrames to Regex as well. The other teammate was in charge of using SQL to create a database and creating a linear regression model. This project was exciting for both of us because it allowed us to put pen to paper and truly practice coding in a real high-stakes environment. It allowed us to simulate what it would be like when working for a company that needed a certain project done in a specified time period. As with any real-world project, there were times where we had to check-in, whether it be at the planning stage with a Project Proposal, or while we were coding in the form of an Interim Report. Moreover, the publication of a Final Report only serves to strengthen this simulation of the workforce.

Our project revolved around two datasets from the government website mentioned above. The hospital bed dataset included information on the number of staffed beds, occupied beds, ICU bed availability, and facility details for hospitals across New York State. On the other hand, the COVID-19 dataset contained weekly counts of new admissions, current hospitalizations, and fatalities. These datasets enabled us to examine the capacity strain at different hospitals and compare it with COVID-19 outcomes.

At the onset of the project, our game plan involved converting data from the government repository into DataFrames using the versatile programming language Python. Specifically, we

accessed the CSV files for both datasets and converted them into Pandas DataFrames, calling them df_hospital_bed_capacity and df_weekly_covid19_hospitalizations_and_fatalities. Upon naming and creating our DataFrames, we cleaned our values to make sure they were cohesive and consistent across both DataFrames. We started by fixing casing issues in both DataFrames and made sure that the titles in the df_hospital_bed_capacity were more in line with those in the df_weekly_covid19_hospitalizations_and_fatalities DataFrame. Subsequently, to prepare for the future plotting of our data, we edited the "As of Date" columns in both DataFrames using the pd.to_datetime() method. Since the reporting frequency differed, we aggregated daily bed-capacity data into weekly summaries to match our df_weekly_covid19_hospitalizations_and_fatalities DataFrame. Both datasets contain fields "As of Date" and "Facility PFI," and we aligned the bed capacity and COVID-19 outcomes at the hospital level by merging both DataFrames into df_combined. We further addressed missing values by replacing them with their respective means. We also eliminated outliers that could have compromised our analysis. The addition of more features, such as occupancy rate, peak hospital strain, and available beds, among others, helped to bolster our project and the information we wished to present. Once df_combined was truly cleaned and ready, we decided to take advantage of the plots we have learned about during this semester to make a time-series graph and a collection of bar graphs. These graphs were meant to highlight potential connections between hospital bed capacity and COVID-19 hospitalizations and fatalities, building a foundation for the more meaningful graphical interpretations we made later in the project. All of these engineered features formed the basis of our SQL tables, allowing us to build a meaningful dataset for regression modeling.

After downloading the data and loading it into DataFrames, we transferred the data into SQL tables using SQLite. We created tables that track daily and weekly bed capacity, as well as COVID-19 outcomes on a weekly basis. SQLite further provided the opportunity to join databases into a single, unified table. This gave us the opportunity to fully leverage the structured data, allowing us to connect and relate data describing available hospital beds and data highlighting COVID-19 patients in New York State. Eventually, we then made our data into another DataFrame so we could start using linear regression. To prepare for modeling, we also used z-score normalization to make our data correctly scaled. This ensured no single datapoint was seen as more important than the others.

To execute the modeling phase, we utilized a Linear Regression model to predict weekly COVID-19 fatalities based on four key capacity metrics: total acute care beds, ICU availability, new admissions, and the acute care occupancy rate. We split the data into training (80%) and testing (20%) sets to ensure robust evaluation.

Our analysis generated two primary visualizations that provided critical insights into the data structure. First, the correlation heatmap revealed a weak but positive correlation between the Acute Care Occupancy Rate and fatalities, providing initial evidence that fuller hospitals may correlate with higher mortality. Second, the scatter plot highlighted a significant characteristic of our dataset: Zero-Inflation. The vast majority of data points were clustered at 0 fatalities, indicating that for most hospitals in most weeks, COVID-19 fatalities were fortunately rare events.

The Linear Regression model yielded a Root Mean Squared Error (RMSE) of 0.32 and an R^2 score of -0.02. The near-zero R^2 score indicated that same-week hospital bed capacity is a poor predictor of COVID-19 fatalities in a linear framework. We attributed this to two factors:

the lag effect (fatalities occur weeks after admission, whereas our model used same-week data) and the zero-inflated nature of the data (93.8% of weeks had zero deaths).

However, despite the low predictive power, the model coefficients revealed an important trend supporting our hypothesis. The coefficient for Acute Care Occupancy Rate was positive (+0.19), while the raw number of 'Available Beds' had a coefficient near zero. This suggested that simply having empty beds does not automatically save lives; rather, the strain ratio (Occupancy Rate) served as the meaningful risk signal, confirming that high hospital occupancy correlated with worse outcomes.

Unfortunately, there were several issues that we had to account for in the current data management practices that were available to us. One such issue was the overall complexity of the data we were dealing with. One of our sources listed thousands of data points, while the other only listed hundreds. This also tied into the topic that not all datasets were standardized, leading to headaches and more data scrutiny. Furthermore, another issue we encountered were incomplete data points, which could have severely impacted our work. Despite all these obstacles, through the methods we learned in class, we found ways to overcome them and deliver accurate results.

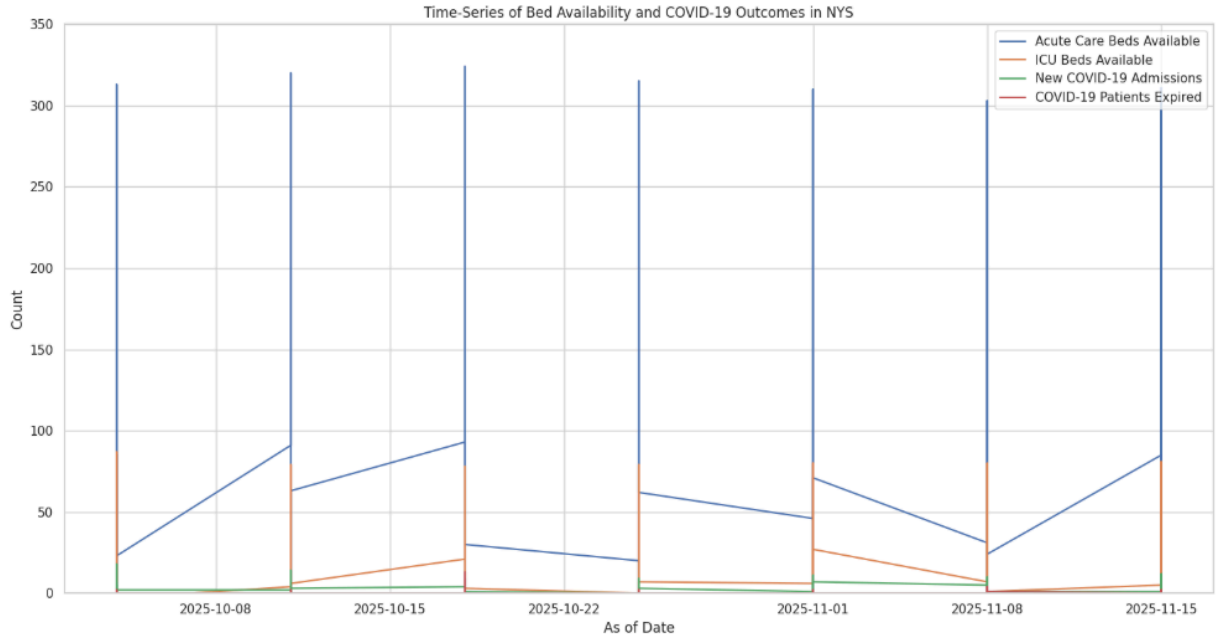
In the end, we were very excited to be working with these datasets because we got the chance to serve our community through the skills we have learned in this class. From using Pandas to extract data from the aforementioned government repository, to modeling the data using linear regression, we showcased the knowledge we have gained and refined during the course of this semester in a way that aims to help save lives and improve governance. Specifically, our finding that high occupancy rates correlated with increased fatalities even when

beds were technically available suggests that policymakers should focus on staffing and load-balancing proactively, rather than relying solely on bed count metrics.

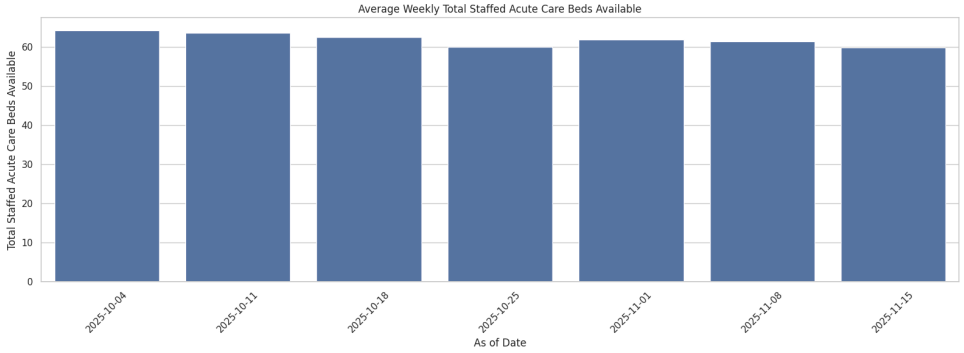
**References/Links:**

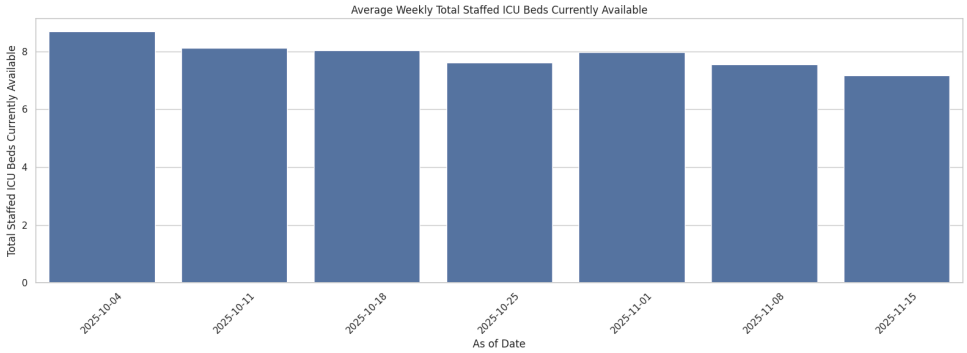

- https://healthdata.gov/State/New-York-State-Statewide-Hospital-Bed-Capacity/cvrn-b3j2/about_data
- https://healthdata.gov/State/New-York-State-Statewide-Weekly-COVID-19-Hospitali/ji5e-24mb/about_data
- https://health.data.ny.gov/Health/New-York-State-Statewide-Weekly-COVID-19-Hospitali/vgyq-b7tb/about_data
- https://pmc.ncbi.nlm.nih.gov/articles/PMC8025594/

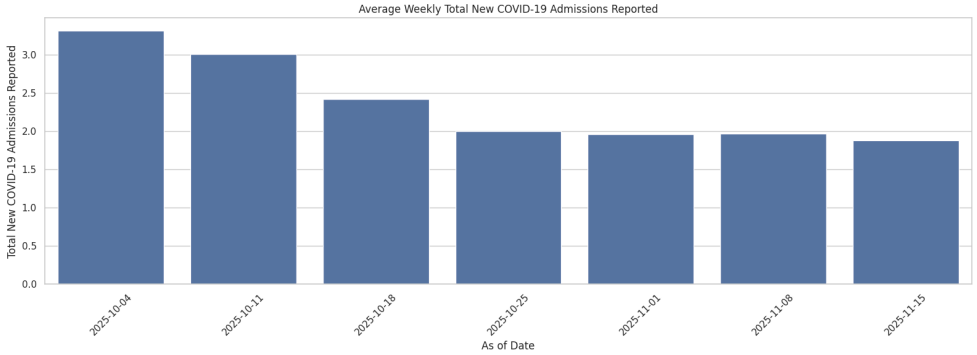
## Time Series Graph:

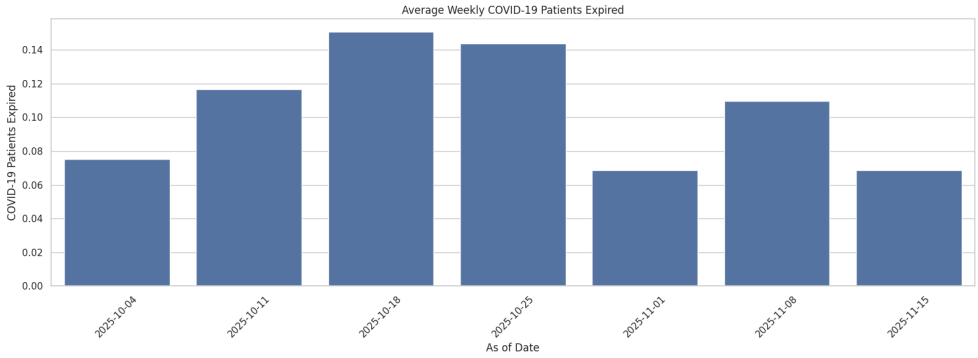



## Collection of Bar Graphs:

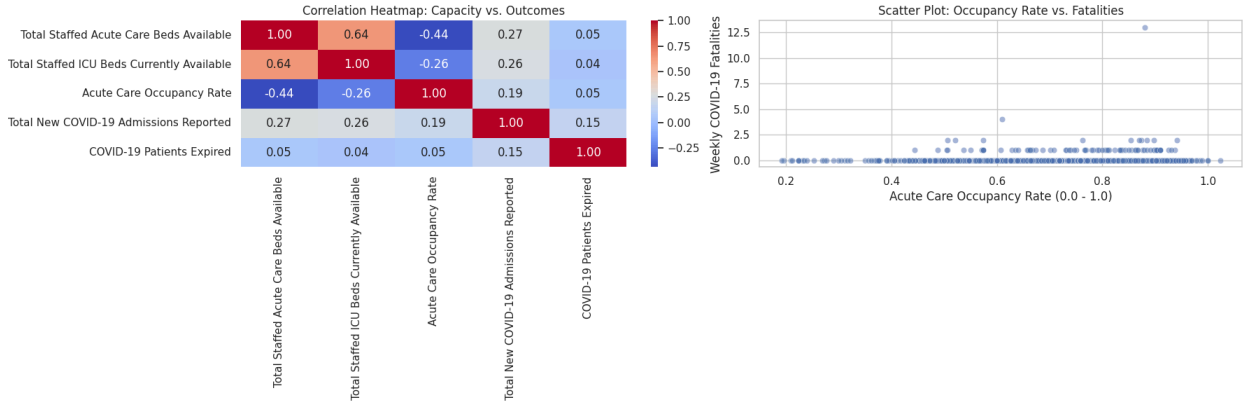





# Heatmap and Scatterplot:



## Scatter Plot: