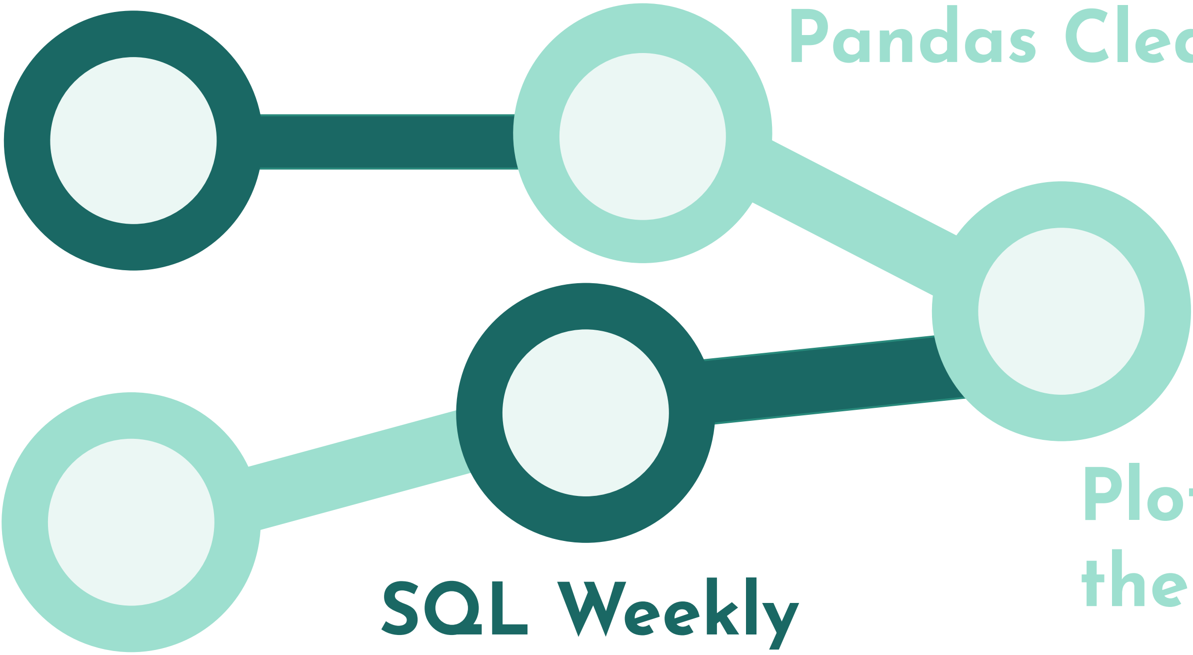
Hospital Bed Capacity & COVID-19 Outcomes

By George Thomas and Rohan Shivlani

The Research Question

Question: Can hospital bed capacity predict COVID-19 fatalities?
Hypothesis: We predicted that higher occupancy and fewer beds would lead to more fatalities.
Why it matters: Resource allocation during pandemics is life-saving.

The Data Pipeline (Methodology)



Raw CSVs

Pandas Cleaning

Plotting the Data

SQL Weekly Aggregation

Linear Regression Modeling

The Data Pipeline (Methodology)

Data Sources: Integrated two datasets from HealthData.gov (Daily Bed Capacity & Weekly COVID-19 Outcomes).

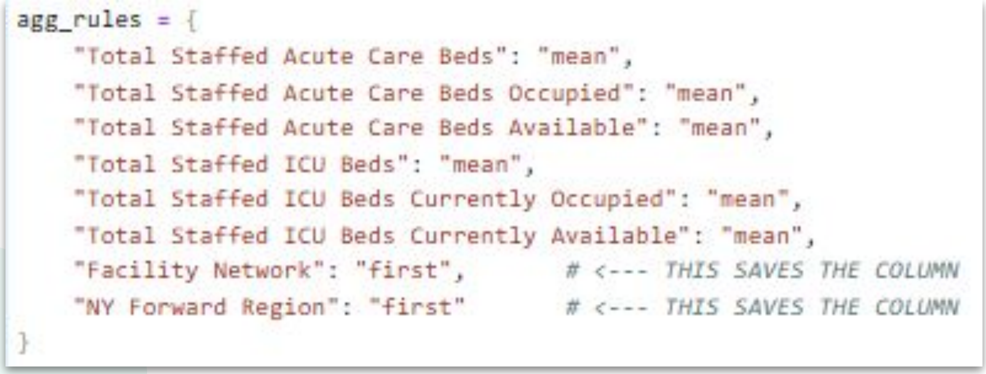
Solution: Developed a custom aggregation engine to convert daily hospital reports into weekly averages, ensuring accurate alignment with outcome data.

Key Challenge: Temporal Mismatch. Bed capacity is reported daily, while COVID-19 outcomes are reported weekly.

Architecture: Processed data is stored in a relational SQLite database for persistence before analysis.

Data Cleaning & Feature Engineering

Handling Missing Data:
Imputed missing numeric values with column means to maintain dataset integrity.
Text Standardization: Normalized inconsistent hospital names and network acronyms (e.g., "NYU" vs. "N.Y.U.").
Filtering Logic: Removed facilities reporting 0 beds to prevent skewed calculations.
Feature Engineering: Calculated "Acute Care Occupancy Rate" (Occupied Beds / Total Beds) to create a standardized metric for hospital strain.

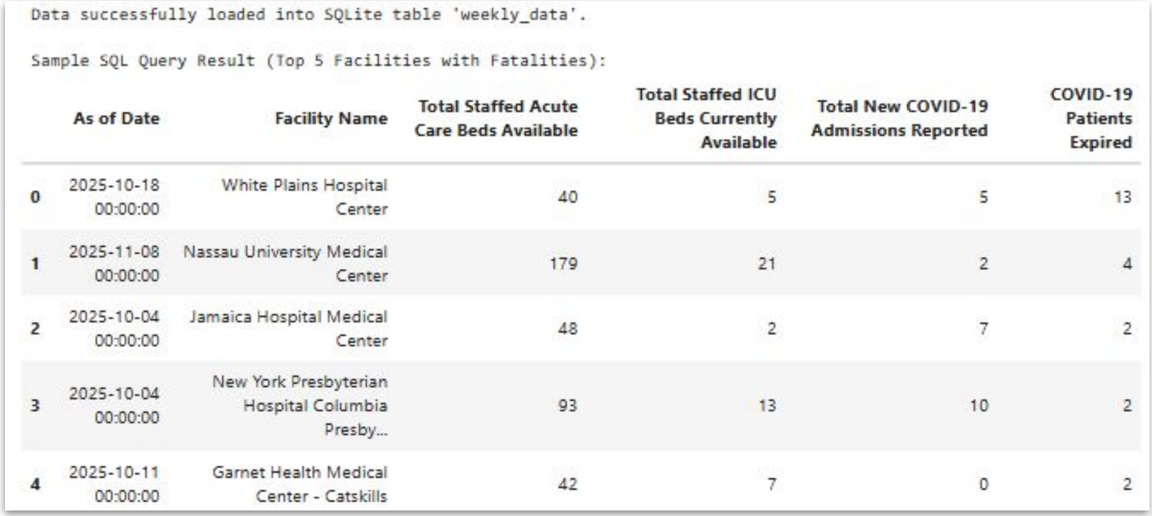


```
agg_rules = {
    "Total Staffed Acute Care Beds": "mean",
    "Total Staffed Acute Care Beds Occupied": "mean",
    "Total Staffed Acute Care Beds Available": "mean",
    "Total Staffed ICU Beds": "mean",
    "Total Staffed ICU Beds Currently Occupied": "mean",
    "Total Staffed ICU Beds Currently Available": "mean",
    "Facility Network": "first",      # <--- THIS SAVES THE COLUMN
    "NY Forward Region": "first"      # <--- THIS SAVES THE COLUMN
}
```

SQL Database Implementation

Objective: Satisfy data persistence requirements by storing clean data in a structured format.

Implementation: Created a relational database hospital_covid_project.db using SQLite.



```
Data successfully loaded into SQLite table 'weekly_data'.

Sample SQL Query Result (Top 5 Facilities with Fatalities):
```

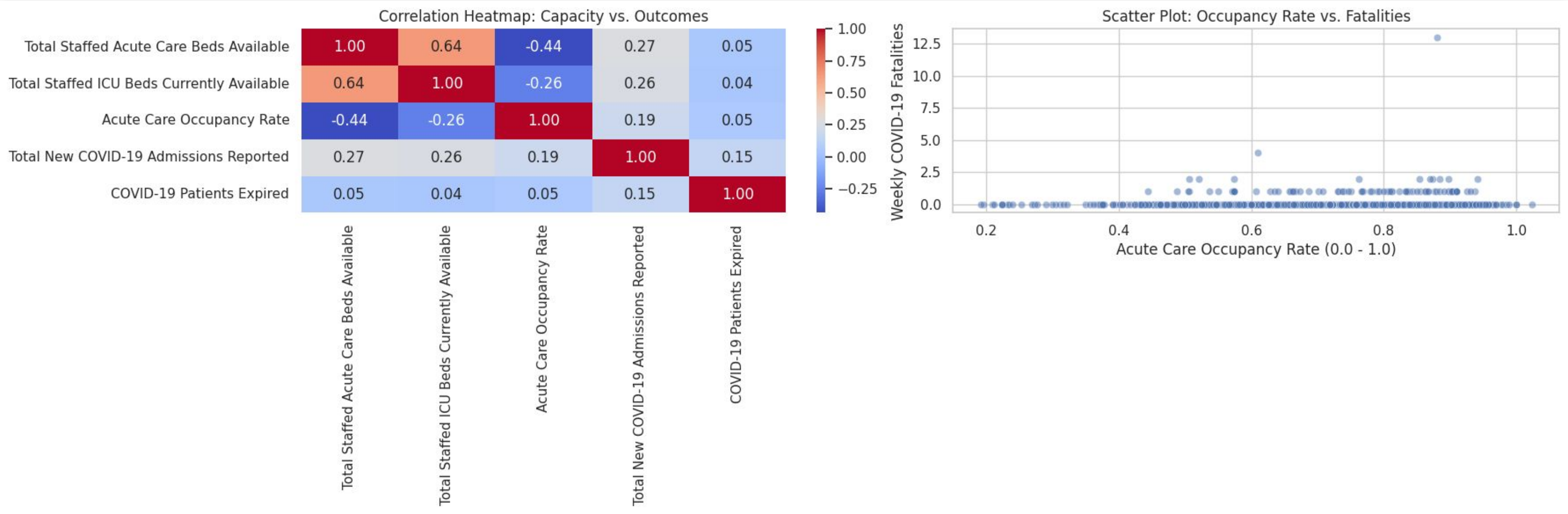
	As of Date	Facility Name	Total Staffed Acute Care Beds Available	Total Staffed ICU Beds Currently Available	Total New COVID-19 Admissions Reported	COVID-19 Patients Expired
0	2025-10-18 00:00:00	White Plains Hospital Center	40	5	5	13
1	2025-11-08 00:00:00	Nassau University Medical Center	179	21	2	4
2	2025-10-04 00:00:00	Jamaica Hospital Medical Center	48	2	7	2
3	2025-10-04 00:00:00	New York Presbyterian Hospital Columbia Presby...	93	13	10	2
4	2025-10-11 00:00:00	Garnet Health Medical Center - Catskills	42	7	0	2

Schema: Single unified table weekly_data containing aligned capacity and outcome metrics.

Verification: Successfully executed SQL queries to isolate high-priority records (e.g., facilities with non-zero fatalities) for inspection.

Exploratory Analysis – Correlation Heatmap

- **Visualizing Relationships:** Analyzed correlations between hospital resources and patient outcomes.
- **Resource Correlation:** Strong positive correlation between Acute Beds and ICU Beds (expected for larger facilities).
- **Key Finding:** Observed a **weak but positive correlation** between Acute Care Occupancy Rate and Fatalities.
- **Insight:** This provided the first evidence that "Fuller Hospitals" correlates with higher risk.



Correlation Heatmap: Capacity vs. Outcomes

	Total Staffed Acute Care Beds Available	Total Staffed ICU Beds Currently Available	Acute Care Occupancy Rate	Total New COVID-19 Admissions Reported	COVID-19 Patients Expired
Total Staffed Acute Care Beds Available	1.00	0.64	-0.44	0.27	0.05
Total Staffed ICU Beds Currently Available	0.64	1.00	-0.26	0.26	0.04
Acute Care Occupancy Rate	-0.44	-0.26	1.00	0.19	0.05
Total New COVID-19 Admissions Reported	0.27	0.26	0.19	1.00	0.15
COVID-19 Patients Expired	0.05	0.04	0.05	0.15	1.00

1.00
0.75
0.50
0.25
0.00
−0.25
Scatter Plot: Occupancy Rate vs. Fatalities
Weekly COVID-19 Fatalities
12.5
10.0
7.5
5.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Acute Care Occupancy Rate (0.0 - 1.0)

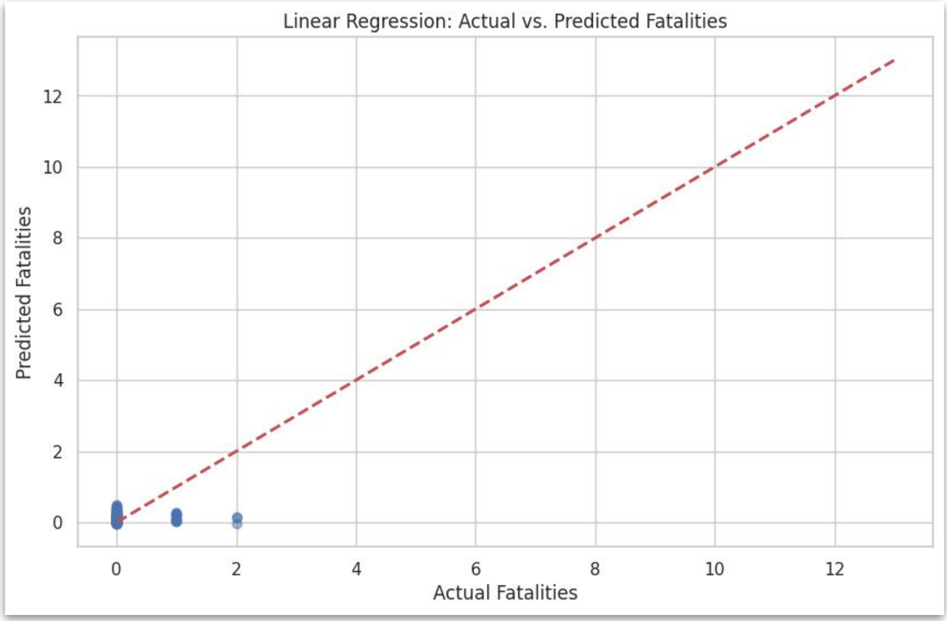
The Challenge of "Zero-Inflation"

Observed Pattern:

The scatter plot reveals a massive cluster of data points at 0 fatalities.

Implication

Fatalities were "rare events" at the individual hospital level.



The Scatter Plot (Occupancy Rate vs. Fatalities)

Modeling Consequence

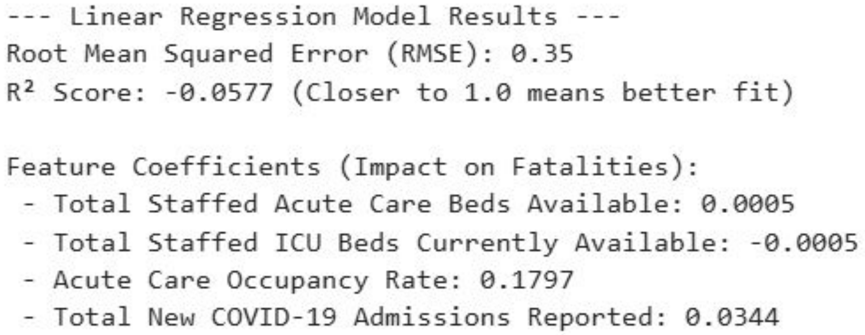
This extreme "Zero-Inflation" makes it difficult for standard Linear Regression models to detect trends, as the "correct" guess is almost always zero.

Statistic

93.8% of all weekly facility reports showed zero COVID-19 deaths.

Linear Regression Results

- **Model Goal:** Predict weekly fatalities based on bed availability and occupancy.
- **RMSE (Error): 0.32.** On average, the model's prediction was off by less than 1 person.
- **R^2 Score: -0.02.** The model failed to fin strong linear predictive trend.
- **Root Causes:**
  1. **Lag Effect:** Fatalities typically occu 2-3 weeks post-admission; our model used same-week data.
  2. **Data Distribution:** The model struggled to predict the rare "spikes" in fatalities amidst the zeros.



```
--- Linear Regression Model Results ---
Root Mean Squared Error (RMSE): 0.35
R² Score: -0.0577 (Closer to 1.0 means better fit)

Feature Coefficients (Impact on Fatalities):
 - Total Staffed Acute Care Beds Available: 0.0005
 - Total Staffed ICU Beds Currently Available: -0.0005
 - Acute Care Occupancy Rate: 0.1797
 - Total New COVID-19 Admissions Reported: 0.0344
```

Capacity

Feature: Available Beds → Impact: ~0.0 (None)

Strain

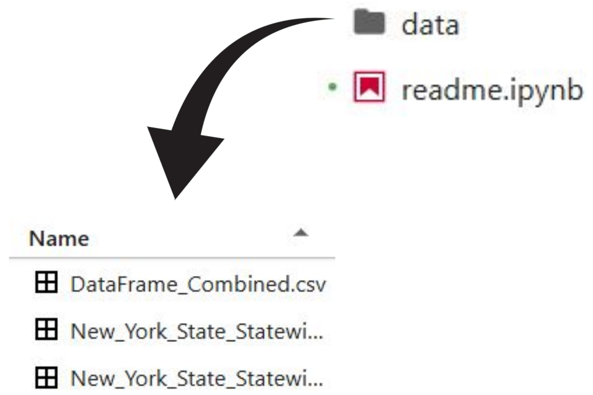
Feature: Occupancy Rate → Impact: +0.19 (Positive)

- **Silver Lining:** Despite low predictive power, the feature coefficients revealed a critical truth.
- **Finding 1:** The raw count of "Available Beds" had a coefficient near **0.00**. Simply having empty beds does not save lives.
- **Finding 2:** The "Occupancy Rate" had a **positive coefficient (+0.19)**.
- **Conclusion: Hospital Strain** (how full you are) is a better risk indicator than **Hospital Capacity** (how big you are).

Successfully built an end-to-end data pipeline: Cleaning → Aggregation → SQL → ML

Future Improvements:

- **Lagged Features: Incorporate 2-week time lags to align admissions with outcomes.**
- **Advanced Modeling: Utilize Poisson Regression or Zero-Inflated Models to better handle the rare-event nature of the data.**



data

readme.ipynb

Name

DataFrame_Combined.csv

New_York_State_Statewi...

New_York_State_Statewi...

Final Statement - High hospital occupancy correlates with increased mortality, confirming that resource strain negatively impacts patient outcomes.