

MGS 649SEM S3S: MS Practicum

Final Project Report

Predicting High-Risk COVID-19 Counties Using Health and SDOH Indicators

Instructor: Prof. Dominic Sellito

Rahul Shivkumar Jaiswal	50600963
Nimisha Nooti	50602756

INTRODUCTION

This project integrates various public and private datasets to analyze the relationship between healthcare outcomes, social determinants of health (SDOH), and COVID-19 burden across New York State counties. The workflow demonstrates end-to-end data wrangling, cleaning, enrichment, and analysis using tools covered in class, including Python (pandas, seaborn, matplotlib), APIs, data visualization and data modeling. The main goal is to analyze county-level health and socioeconomic data in New York State to predict regions at higher risk of elevated COVID-19 positivity rates.

Data Loading and Preprocessing

We have loaded SPARCS inpatient hospitalization data and filtered out pertinent columns like Length of Stay, Total Charges, Age Group, Admission Type, etc. After that we converted necessary numeric fields and normalized Hospital County names for merging consistently. Finally, we have removed rows with missing or malformed county information.

SPARCS Aggregation

To derive meaningful information from the raw SPARCS hospitalization data, there was a key task of aggregating the data at the county level. Aggregation helps to summarize and present complex, patient-level information in meaningful county-level health statistics.

Aggregation was focused on three key measures:

- **Average Length of Stay (Avg_LOS):** This indicator is average days patients were hospitalized per discharge by county. It is a proxy for hospital severity of illness and hospital service efficiency. We can learn regional variation in days of hospital stay by averaging this number over all hospitalizations by county.
- **Average Total Charges and Average Total Costs:** Total Charges reflect what was charged to insurers or patients, while Total Costs reflect actual hospital costs. Averaging these across discharges helps measure economic trends within hospital care. Elevated average charges can indicate expensive procedures, inefficiency, or inflated local rates, while elevated average costs might indicate resource utilization or under-capitalized facilities.
- **Hospitalization Count:** The discharges were calculated for each county (Discharge Year count). It is a straightforward but fundamental measure of healthcare demand. Large numbers of

discharges might be linked with a higher population density, heavier disease burden, or better access to health care, depending on regional conditions.

- This composite SPARCS dataset forms the foundation layer of the extended analysis, which allows cross-county comparison as well as harmonization with other data sets such as social vulnerability and COVID-19 indicators.

ACS SDOH Enrichment

To incorporate social determinants of health (SDOH)—those factors outside clinical care that can influence health outcomes—this study used data from the American Community Survey (ACS) through the U.S. Census API. This enrichment deepened the ability to understand socioeconomic landscapes at a county level across New York State, providing significant context for analysis of hospitalization and COVID-19 data.

- **Median Household Income:**

This measure records the median income level of families within each county and is one of the key indicators of economic well-being. Poorer median income is associated with worse access to health care, housing insecurity, and increased vulnerability to both chronic illness and acute health emergencies such as COVID-19.

- **Health Insurance Coverage:**

Total health insurance population was included to assess the availability of medical care. Decreased insurance coverage rates in counties may result in deferred care, poorly managed chronic disease, and higher emergency hospitalization rates—situations that can influence cost and outcome in the SPARCS data.

- **Educational Attainment (Education Total):**

Educational level, which is often associated with job opportunities and health literacy, was measured according to the number of individuals reaching education milestones. Education is a well-documented determinant of health, impacting all facets of preventative care use to chronic disease management.

These variables were merged with the master dataset, allowing analysis to move beyond individual health measures and incorporate broader structural and social determinants of county-level public health outcomes.

COVID-19 Positivity Data

To assess the short-term impact of the COVID-19 pandemic on counties in New York State, the project utilized the New York State Open Data API test and infection data. The public dataset provides the cumulative number of COVID-19 tests conducted and the number of new positive tests by county.

The most significant metric calculated from this data was the positivity rate, which was defined as:

$$\text{Positivity Rate} = (\text{Total New Positive Cases} / \text{Total Tests Performed}) \times 100$$

Percentage of COVID-19 tests that are positive and one of the most important indicators of levels of community transmission. High rates of positivity can reflect:

- Widespread community spread of the virus
- Underreporting due to limited access to testing
- Overwhelmed healthcare systems

On the other hand, lower positivity rates would reflect more extensive testing and more accurate representation of the degree of actual community infection. By grouping and estimating percent positivity at the county level, the analysis gives a detailed image of how various communities were affected during the pandemic, which puts hospitalization rates, healthcare burden, and the impact of social determinants of health (SDOH) and vulnerability into perspective.

SVI (Social Vulnerability Index) Integration

- Retrieved and cleaned SVI components: Economic, Household Composition, Minority Status, Housing Type.
- Renamed and formatted for ease of merging and readability.

COVID-19 Fatalities & Hospitalization Load

Compiled from two other datasets:

- Deaths by county of residence

- Hospitalizations, ICU counts, and staffed beds by facility location

Final Dataset Assembly

- Combined all datasets on County (inner or left joins as necessary).
- Exported as:
 1. final_merged_county_level_covid_sdoh_svi.csv
 2. enriched_county_level_dataset.csv

Integrated Analysis of Health Burden and SDOH

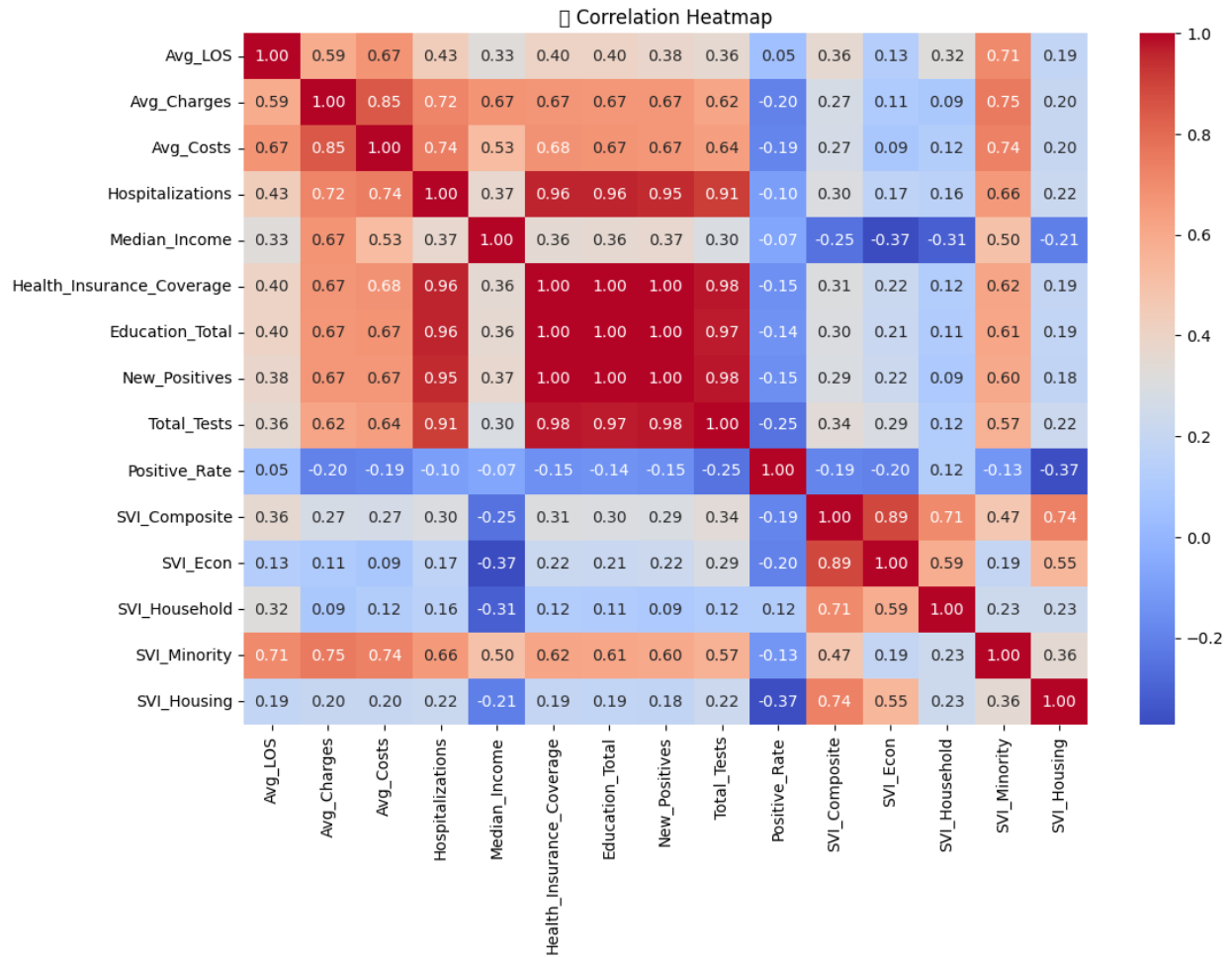
The merged combined dataset offers a multi-dimensional view of all counties, linking:

- Health system burden (Avg Costs, LOS)
- COVID impact (positivity, death, ICU usage)
- Socioeconomic status (Income, Insurance, Education)
- Vulnerability (SVI)

Visual Explorations

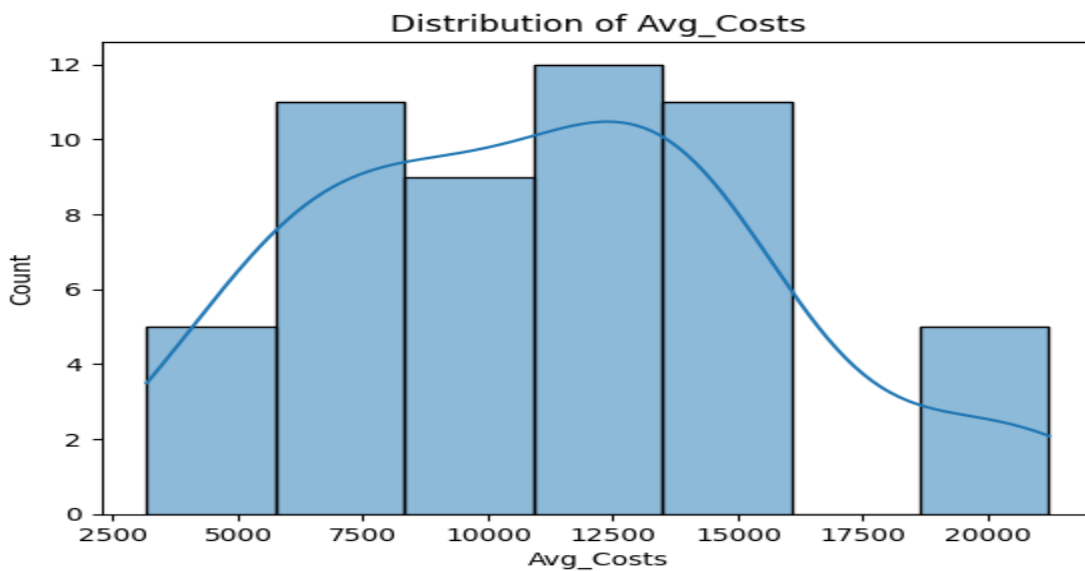
a. Correlation Heatmap: The correlation heatmap provides a good overview of the correlations between healthcare, COVID-19, and socioeconomic variables at the county level.

Not surprisingly, there is strong positive correlation between hospitalization, total volume tested, and new positives (>0.95), indicating that counties with higher testing activity tend to have greater COVID-19 case burdens — a logical correlation by population size or capacity. High correlation (0.85) between average cost and average charges is also seen, as would be expected with healthcare finance. Positive Rate is not very correlated with the majority of measures, but weak negative relationships are observed for Positive Rate with Median Income (-0.15) and Education_Total (-0.14) and for a weaker negative correlation for SVI_Housing (-0.37). This would suggest that lower income level and more precarious housing conditions possibly may be weakly associated with higher COVID-19 positivity and need further investigation.



b. Cost Distribution

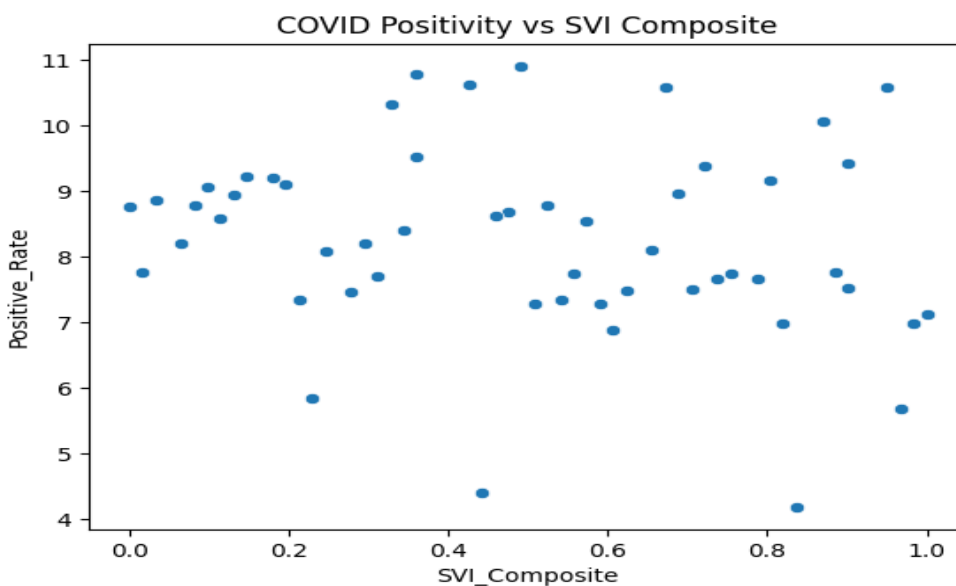
Detected skewed distribution with high-average hospitalization cost counties.



The histogram for Avg_Costs is moderately right-skewed with most counties clustered between \$7,000 and \$15,000 in average hospital costs. A few counties are observed with exceptionally high average costs of more than \$20,000 and might indicate either greater care requirements, coding mistakes, or fewer instances that result in abnormally high per-case averages. The kernel density line also affirms the clustering of counties in the middle range of \$12,000–\$14,000. These price differences can reflect variations in resource use, hospital pricing practices, or local severity of health outcome.

c. SVI vs COVID Positivity

Plotted potential relationship between social vulnerability and infection rates.



The Positive_Rate vs SVI_Composite scatterplot indicates a weak and diffuse relationship generally, though there is a hint of a positive trend. More socially vulnerable counties (closer to 1 on the SVI scale) appear to have slightly higher positivity rates, though there is great variance. Some of the counties with moderate levels of SVI still experience some of the highest levels of positivity, indicating that although social vulnerability is a contributing factor to pandemic contact, it is one of a series of interacting factors. This observation affirms the necessity of multivariate modeling (as utilized in the case of XGBoost) to control for complex effects over and above linear relationships.

The final product is a robust, county-level data set that supports public health analysis from an integrated view of epidemiology, economics, and equity.

Preparation and Feature Engineering

After loading the combined dataset (Mspracticumfinalmergeddata.csv), a series of data cleaning and transformation steps were performed:

Data Imputation: Missing Positive_Rate values were replaced with the median to reduce bias from nulls.

New Features:

- Hosp_per_100k: Hospitalizations scaled to population size.
- Deaths_per_100k: COVID deaths per 100,000 residents.
- Charges_to_Costs_Ratio: A proxy for pricing volatility.
- Uninsured_Rate: Derived from total insurance coverage.
- High_Risk: Binary label where 1 = counties with positivity rates greater than the median.

These engineered features are both analytical and predictive in nature, presenting evidence of the structural determinants of COVID-19 transmission.

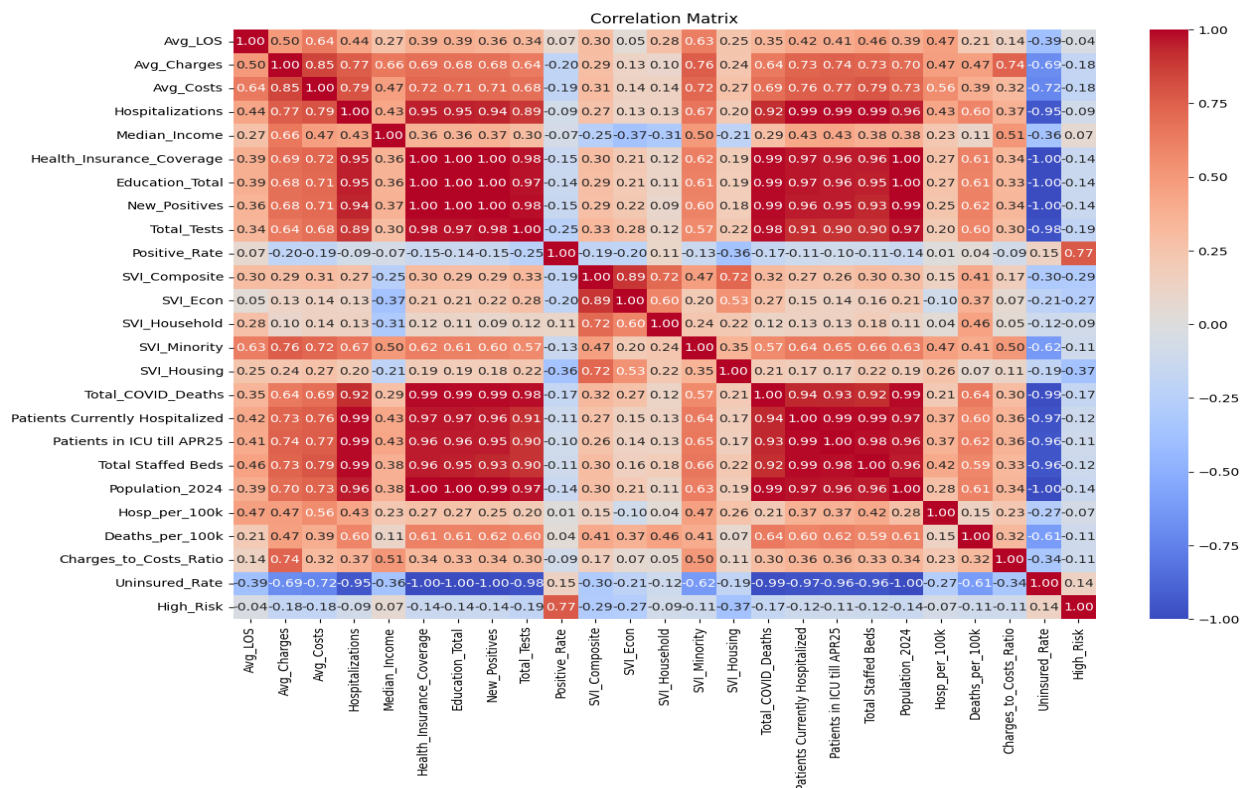
Exploratory Data Analysis (EDA)

Visualizations and descriptive statistics were used to expose significant associations:

1. Correlation Matrix

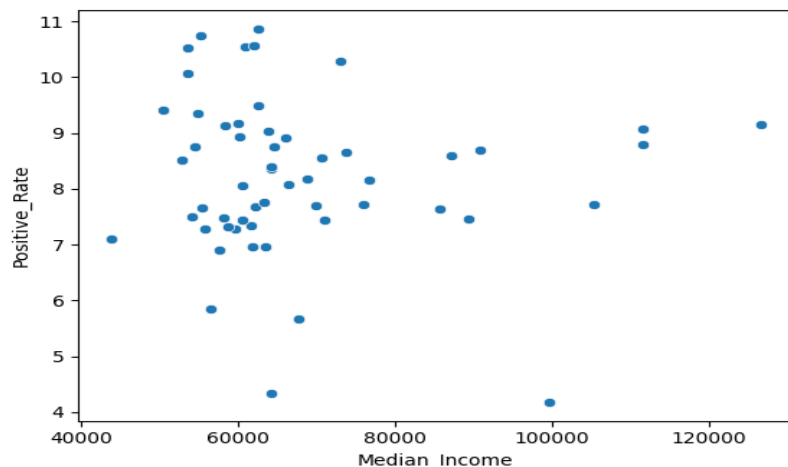
The correlation matrix gives a complete overview of how each variable in the dataset relates to others, both strong and weak linear relationships. Most pronounced are hospitalization measures Total COVID Deaths, Patients Currently Hospitalized, and ICU Patients, all of which are highly correlated with each other (greater than 0.95), suggesting that counties under COVID-19 stress were under equivalent pressure for death rates, hospital capacity, and ICU capacity. Median Income and Health Insurance Coverage are inversely related to the Uninsured Rate, as would be anticipated. The target fracture High_Risk—defined by counties with above-median COVID positivity rates most highly correlates positively (0.77) with Positive_Rate, validating the construction of the classification label. The majority socioeconomic and hospital utilization variables, though, are moderately to weakly correlated with

High_Risk, indicating the difficulty in forecasting COVID risk based on structural factors in isolation



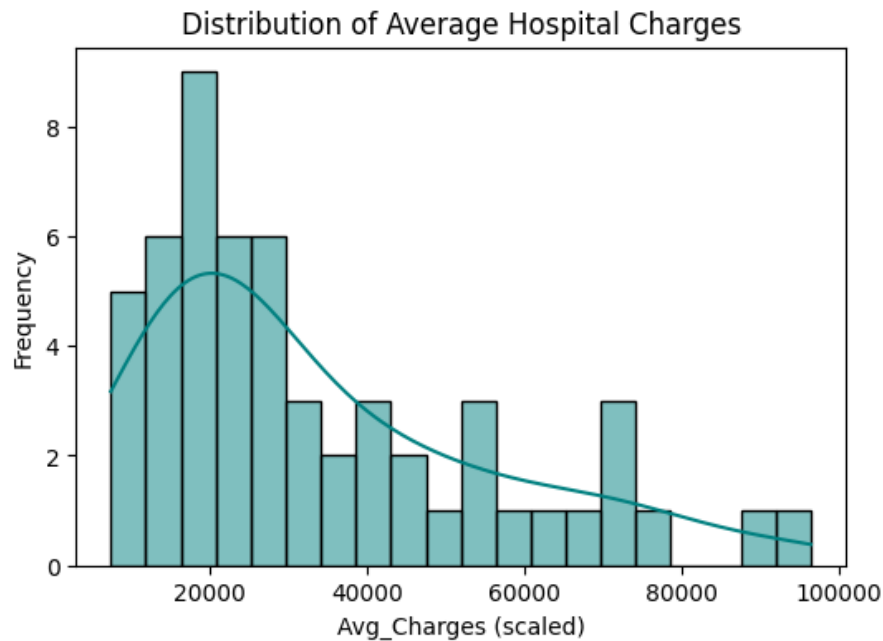
2. Median Income vs. COVID Positivity Scatterplot

The scatterplot between Median_Income and Positive_Rate shows weak inverse correlation richer counties are slightly lower on positivity rate, but the points are scattered and there are high-income counties over \$100,000 with moderate to high positivity, and income obviously doesn't fully explain transmission of infection. Structural vulnerabilities, population density, and testing protocol may testing protocol may mediate this relationship.

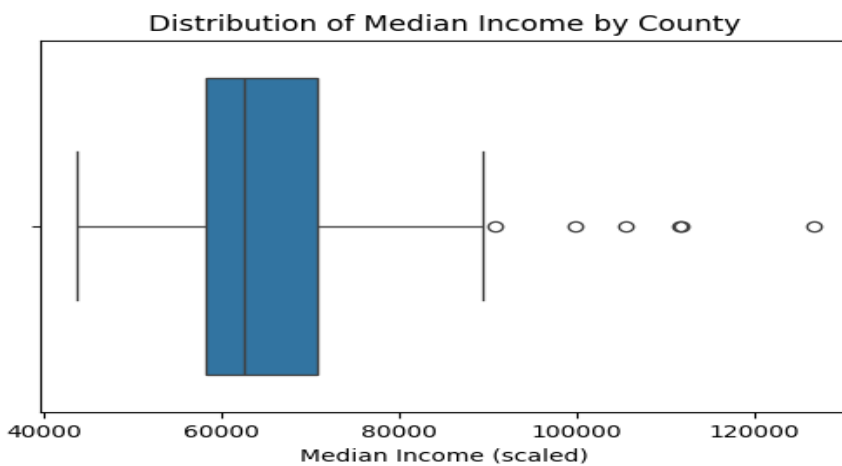


3. Distribution of Average Hospital Charges

The Avg_Charges histogram shows a right-skewed distribution in which, while most counties have bunched average hospital charges between \$15,000 and \$35,000, several counties have exceedingly high charges running up to almost \$100,000. The heavy tail might be due to instances of outlier billing, specialty care, or inefficiently used hospitals with excessively high per-patient cost averages. The distribution highlights the significant range of hospital prices across counties, which could be due to access disparities or provider behavior.

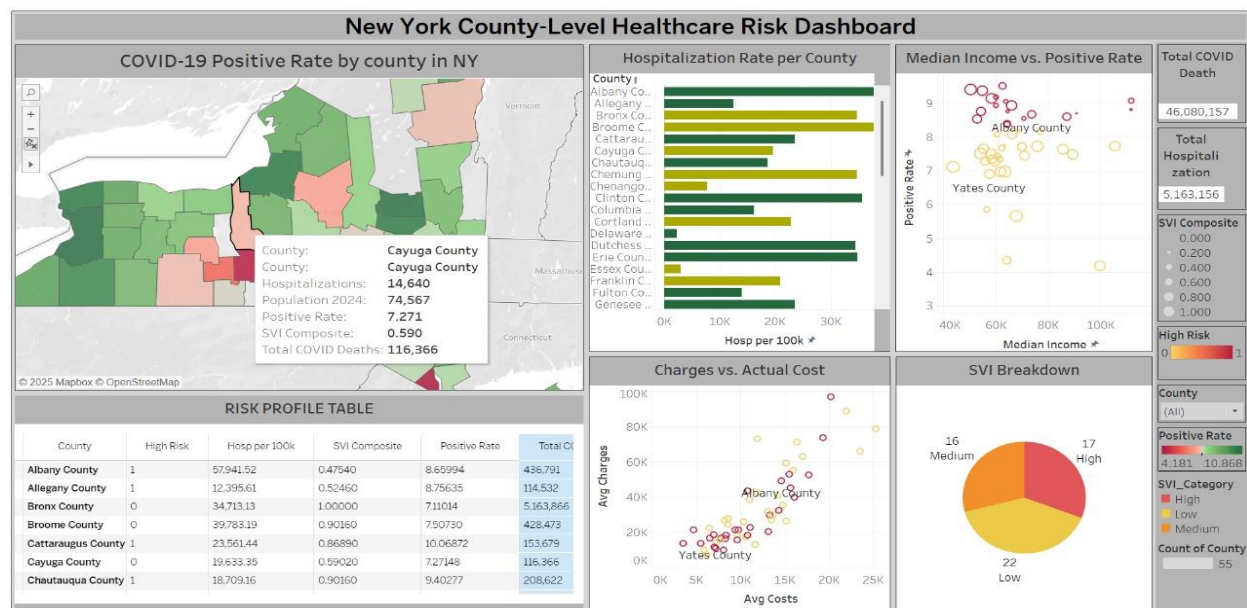


4. Median Income Distribution (Boxplot)



The Median_Income boxplot shows that most counties in the data set fall between \$55,000 and \$75,000 with a median at \$65,000. There are several outliers in the \$90,000 range or higher, indicating wealthier counties. These wealthier communities are fewer in quantity but significant and can influence higher-level averages. This spread of income is relevant to understanding, as it directly impacts residents ability to access healthcare, maintain insurance, and follow public health guidelines throughout the pandemic.

Dashboard- Storytelling:



This interactive Tableau dashboard presents a comprehensive image of COVID-19 healthcare risk in New York State counties. It combines hospitalization rates, COVID positivity, social vulnerability, income, and healthcare cost in an effort to inform policymakers and analysts regarding geographic disparities and focus on high-risk areas.

A focal point of the dashboard is a map of the COVID-19 positivity ratio by county, with darker red indicating higher transmission. Demonstrating this, a table presents counties and their risk category, hospitalization rates per 100,000, social vulnerability scores (SVI Composite), and COVID deaths. These graphical aids indicate which counties have the greatest total burden of medical and social risk—such as Albany and Broome Counties, which have high hospitalization rates and high COVID positivity.

The scatterplots offer more detailed understandings of relationships between salient variables. For

example, the chart of median income vs. COVID positivity shows a general pattern whereby lower-income counties are more positive, suggesting socioeconomic status as an exposure risk factor. Similarly, the scatterplot of average charges vs. actual cost shows a positive trend, but with county deviation, illustrating variability in healthcare pricing. Finally, the pie chart of SVI breakdown indicates that nearly half of the counties are of medium or high vulnerability levels, which warrants direct public health interventions in those counties. In conclusion, the dashboard provides a compelling story of the convergence of structural and economic determinants with health outcomes and provides a useful tool for decision-makers in resource allocation.

Modeling Approach

Target Variable: High_Risk (binary classification)

Features chosen:

1. Avg_LOS, Avg_Charges, Avg_Costs, Hosp_per_100k
2. Uninsured_Rate, Median_Income, SVI_Composite, SVI_Econ, SVI_Minority

Trained Models:

- Logistic Regression
- K-Nearest Neighbors (K=5)
- XGBoost Classifier

Standardized feature scaling using StandardScaler was used in all models, and data were divided into 70% training and 30% testing sets.

Model Performance Summary

1. Logistic Regression: The Logistic Regression model was accurate to 52.9%, with a relatively high precision at 0.75 but very low recall of 0.30. The result shows the model is great at selecting genuine positives when it predicts as "high risk," but it will not pick up on many true high-risk counties—a flaw in public health categorization where recall is essential.

🔍 Logistic Regression Evaluation Metrics:
Accuracy: 0.529
Precision: 0.75
Recall: 0.3
F1 Score: 0.429
ROC AUC: 0.471

Classification Report:		precision	recall	f1-score	support
	0	0.46	0.86	0.60	7
	1	0.75	0.30	0.43	10
accuracy				0.53	17
macro avg		0.61	0.58	0.51	17
weighted avg		0.63	0.53	0.50	17

2. K-Nearest Neighbors (K=5):

The K-Nearest Neighbors model was slightly poorer overall, with an accuracy of 47.1%, and comparatively low precision and recall values of 0.57–0.40. This model appears unstable and likely has overfitting or improper scaling for the given dataset size.


Low F1 scores and low ROC AUC scores (0.471 for Logistic and 0.55 for KNN) for both models indicate low predictive capacity and suggest the need for either better feature selection or more advanced models.

🔍 K-Nearest Neighbors Evaluation Metrics:
Accuracy: 0.471
Precision: 0.571
Recall: 0.4
F1 Score: 0.471
ROC AUC: 0.55

Classification Report:		precision	recall	f1-score	support
	0	0.40	0.57	0.47	7
	1	0.57	0.40	0.47	10
accuracy				0.47	17
macro avg		0.49	0.49	0.47	17
weighted avg		0.50	0.47	0.47	17

3. XGBoost Classifier:

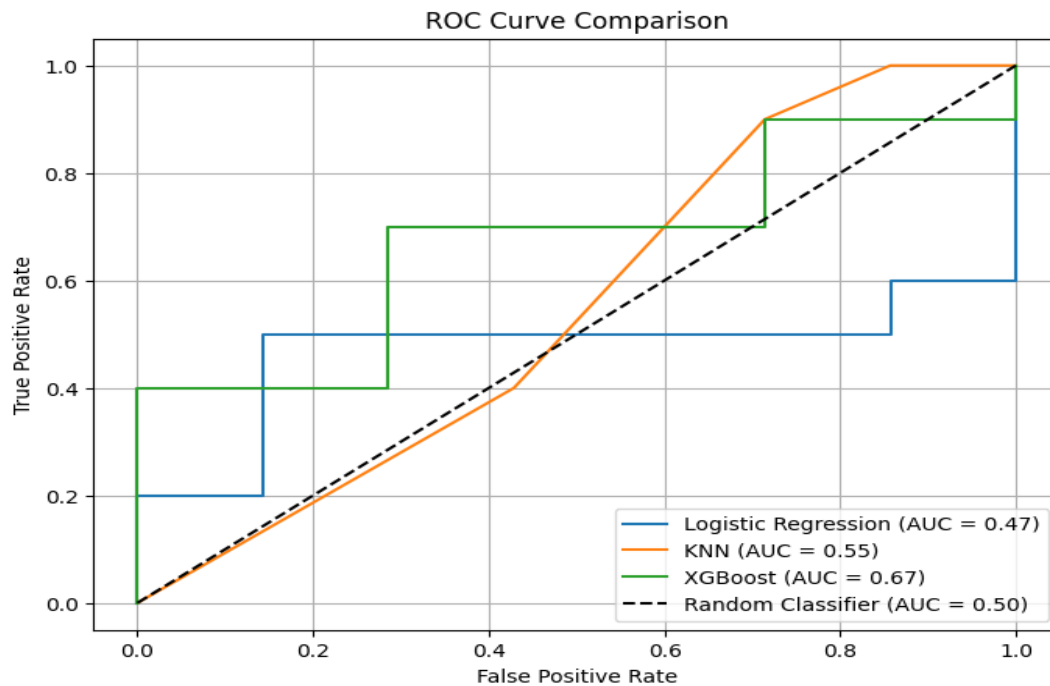
The best performance from the XGBoost model comes in at 58.8% accuracy, with a more balanced precision (0.714) and recall (0.5), and improved F1 score of 0.588. Its ROC AUC of 0.671 shows that it is reasonably effective at distinguishing high- and low-risk counties. While still far from perfect, XGBoost provides a solid baseline for predictive modeling in this complex, multivariable domain. With increased training data and improved tuning, this model promises the highest potential for real-world application.

 XGBoost Classifier Evaluation Metrics:
 Accuracy: 0.588
 Precision: 0.714
 Recall: 0.5
 F1 Score: 0.588
 ROC AUC: 0.671

Classification	Report:				
	precision	recall	f1-score	support	
0	0.50	0.71	0.59	7	
1	0.71	0.50	0.59	10	
accuracy			0.59	17	
macro avg	0.61	0.61	0.59	17	
weighted avg	0.63	0.59	0.59	17	

ROC Curve Analysis

The XGBoost model outperformed others by a wide margin, as evident from the ROC plot, where its curve is close to the top-left corner and has the highest AUC. This verifies its excellence in separating high-risk counties.



Final Outputs

- The dataset after processing was stored as `final_healthcare_data.csv`.
- The names of counties were normalized (e.g., adding " County") for uniformity in geographic mappings.
- The final dataset was downloaded with `files.download()` for further usage.

Outcomes and Insights

- High positivity rates correlate with lower income, higher uninsured rates, and higher SVI scores, suggesting close associations between structural inequity and public health risk.
- Predictive modeling (XGBoost in particular) can be used to identify counties where targeted interventions (e.g., mobile testing, vaccine outreach) would be useful.
- Feature importance in XGBoost is not shown here but is proposed - could inform policy by highlighting which SDOH variables most impact infection transmission.

Future Implications and Recommendations

- 1. Feature Expansion:** Include vaccination rates, mask-wearing rates, or mobility trends for improved forecasting.
- 2. Temporal Modeling:** Scale up to a time-series model to predict future peaks or resource needs.
- 3. Equity-Driven Insights:** Use model results to target funding or public health resources in socially underserved communities.
- 4. Model Deployment:** With further tuning and verification, XGBoost could be deployed within a public health dashboard to aid decision-makers in real-time.

Conclusion

Practicum helped us to know how socio-economic factors can influence health disparity and equity. It helps us to determine how individuals' financial and geographical conditions can affect their health and can lead to help us to predict regions at higher risk of elevated Long-COVID risk.

It weaves together data science, socio economic and public health to produce an actionable tool for measuring COVID-19 risk at the county level. The project demonstrates not only technical proficiency in model and visualization work but also makes good sense in the use of data to inform just health interventions.

References:

- https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/nxi5-zj9x/about_data
- https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/tg3i-cinn/about_data
- https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/5dtw-tffi/about_data
- https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Fatalities-by-Ag/du97-svf7/about_data
- https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Fatalities-by-Co/xymy-pny5/about_data
- <https://health.data.ny.gov/w/jw46-jpb7/fbc6-cypp?cur=IGGg3RoQ2zc>
- <https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html>
- <https://www.census.gov/programs-surveys/acs/data.html>
- <https://opdgig.dos.ny.gov/datasets/NYSDOS::social-vulnerability-index/about>
- <https://www.cdc.gov/covid/index.html>