Non-COVID Hospital Bed Utilization Prediction Using Social and Geographic Determinants of Health

Mehmet Comert, Rea Kelolli, June Lemieux, Melody Rios[1]

May 11, 2025

[1] * Each author contributed equally to the design, coding and development, analysis, and writing of this project.

Abstract

This study developed a predictive model for non-COVID hospital bed utilization using social and geographic determinants of health. Using county-level 2020 data from the AHRQ SDOH database and the COVID-era hospitalization dataset, using feature engineering and machine learning, we trained and evaluated several regression models to estimate the Bed Utilization Ratio, a proxy for healthcare system strain excluding COVID-19 impacts. After extensive model comparison, an XGBoost Regressor trained on the full feature set without dimensionality reduction emerged as the best-performing model, achieving a Test R^2 of 0.3617 and RMSE of 0.1378.

Key predictors included population poverty, housing cost burden, single-parent household rates, disability prevalence, and distance to intensive care facilities. The model reveals how structural vulnerabilities rather than clinical metrics alone can predict healthcare utilization. These insights offer valuable support for public health agencies to target interventions, allocate resources more equitably, and anticipate systemic strain in under-resourced communities.

1 Introduction

1.1 Background and Motivation

Hospital bed availability is a fundamental measure of healthcare system readiness and resilience. Managing this capacity efficiently is crucial not only during crises but also in everyday circumstances to ensure timely patient care and optimal resource use. Traditional forecasting models often focus on demographic indicators, disease prevalence, and seasonal illness trends.

However, these approaches can miss broader structural factors such as poverty, transportation access, and housing instability that also heavily influence hospitalization patterns. While many models focus on demographics and disease burden, social determinants of health (SDOH) and geographic access are often overlooked (Kreuter et al., 2021, 329-344).

The social determinants of health (SDOH) have gained increasing recognition as critical predictors of health outcomes. Individuals facing socioeconomic hardship or geographic isolation may delay seeking preventive care, leading to higher rates of hospitalization when conditions become acute. This project addresses a critical gap by developing a predictive model that integrates SDOH and geographic access indicators to forecast non-COVID hospital bed utilization across U.S. counties. By doing so, we aim to support **public health agencies** in identifying vulnerable regions, guiding strategic investments, and proactively strengthening healthcare system resilience.

1.2 Research Question

Can social and geographic determinants of health (e.g., disability, poverty, healthcare access, advanced age, and transportation access) predict non-COVID hospital bed utilization at the county level in 2020?

1.3 Hypothesis & Prediction

We hypothesize that counties with higher rates of poverty, single-parent households, housing cost burdens, disability prevalence, older adult populations, and limited access to emergency and intensive care facilities will exhibit higher non-COVID hospital bed utilization. We expect these social, housing, and geographic determinants of health (SDOH) to serve as strong

predictors of system strain, even when accounting for population density and regional differences.

1.4 Significance

Understanding the community-level drivers of hospital strain can empower policymakers to intervene before critical capacity thresholds are reached. By combining predictive modeling with structural vulnerability indicators, this project supports a shift from reactive to proactive public health planning. Our approach provides a replicable framework that can be integrated into community health assessments, hospital capacity planning, and grant allocation strategies focused on health equity.

2 Literature Review

2.1 Hospital Capacity Planning Models

Traditional hospital capacity planning has relied primarily on demographic projections and historical utilization patterns. Artiga & Hinton (Artiga & Hinton, 2018) emphasized that overlooking social and economic barriers can lead to underestimating system vulnerabilities. Additionally, they fail to capture delayed care behaviors common in uninsured or low-income populations. Recognizing the limitations of purely demographic models is essential for building more comprehensive forecasts.

2.2 Social Determinants of Health

A growing body of research highlights the importance of social determinants in healthcare utilization. Garg et al. (Garg et al., 2020, 458–464) found that income level, education, and

insurance status significantly predicted preventable hospitalizations, while Magnan (Magnan, 2017) demonstrated that food insecurity and housing instability were associated with increased emergency department visits.

2.3 Geographic Access and Structural Barriers

Geographic barriers, including rurality, distance to medical facilities, and lack of transportation infrastructure, exacerbate disparities in access to care. Rural residents often experience delayed or foregone care due to these structural barriers, resulting in more severe illness upon presentation and increased reliance on inpatient services. Addressing geographic access factors alongside socioeconomic vulnerabilities provides a more holistic understanding of healthcare demand patterns.

2.4 Machine Learning Approaches in Healthcare Forecasting

Recent advances in machine learning have enabled the development of more flexible and powerful predictive models. Ensemble learning methods, particularly Random Forest and XGBoost, are well-suited to high-dimensional healthcare data containing nonlinear relationships and missing values. Studies have demonstrated that these models can outperform traditional regression approaches when predicting complex outcomes such as hospital admissions, length of stay, and resource utilization (Rahmatinejad et al., 2024; Zeleke et al., 2024).

3 Data and Methods

3.1 Data Sources

Two primary datasets were integrated for this study. The first was the 2020 Social Determinants of Health (SDOH) database curated by the Agency for Healthcare Research and Quality (AHRQ), which provides a wide array of county-level indicators including poverty rates, household structures, disability prevalence, and transportation access. The second was the COVID-19 Hospitalization Dataset released by the U.S. Department of Health and Human Services (HHS), from which we computed a non-COVID `BED_UTIL_RATIO`. This ratio was calculated as:

$$Bed\_util\_ratio = \frac{(Inpatient\ Beds\ Used - COVID\ Beds\ Used)}{Total\ Inpatient\ Beds}$$

This measure reflects the proportion of hospital beds occupied by non-COVID patients, providing a baseline assessment of routine healthcare system demand.

3.2 Data Preprocessing

To support the development of our hospital bed utilization prediction model, we implemented a targeted preprocessing strategy that prioritized data completeness, clarity, and model readiness. This approach was designed to balance data integrity with analytical flexibility while maintaining consistency across features and records.

We began by removing rows that contained missing values in the `Region` field, and excluded the District of Columbia, which did not align with state-based comparisons and introduced geographic outlier effects. In terms of feature selection, we eliminated three specific variables: `State`, a geographic label that offered no predictive value; `Land_area_sqmi`, which was redundant given population density measures; and `Pct_renter_cost_30pct_plus`, a highly collinear housing cost burden feature identified through VIF analysis.

Records with missing predictor values were removed to ensure consistency in the modeling dataset. However, we retained rows that contained information in key fields like `Is_Metro_Micro` and `Bed_util_ratio`, allowing us to preserve cases that could be imputed or predicted in later steps. Categorical fields such as `Region` were transformed using one-hot encoding, while `Is_Metro_Micro` was encoded as a binary variable. To prepare for distance-based imputation and maintain numerical stability, we standardized all features using z-score normalization.

Missing values in the `Is_Metro_Micro` field were imputed using a K-Nearest Neighbors (KNN) method with five neighbors, applied after scaling. This allowed us to retain the feature without introducing bias or unnecessary exclusion. The modeling dataset was then split into training and testing sets using a feature-aware partitioning approach that considers the number of predictors, resulting in a well-balanced sample for validation.

To preserve the flexibility of future model application, we maintained a separate version of the dataset containing counties with missing `Bed_util_ratio` values. This ensures that the trained model can be used later to generate predicted scores for regions with incomplete

outcome data. The full technical details of this preprocessing workflow including variable scaling, encoding, imputation, and dataset partitioning are documented in Appendix A.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted on the cleaned and unscaled dataset to preserve interpretability and support sound feature selection prior to modeling. This phase of the analysis helped reveal meaningful relationships, potential data issues, and regional patterns in the target variable.

The distribution of the target variable, `Bed_util_ratio`, was moderately right-skewed. However, no transformation was applied, as tree-based models such as XGBoost are robust to non-normality in the outcome. To address potential collinearity, we used Variance Inflation Factor (VIF) analysis and removed variables that exceeded accepted thresholds—particularly housing cost burden features that exhibited strong overlap. For example, `Pct_renter_cost_30pct_plus` was excluded due to its high VIF and conceptual redundancy.

Correlation analysis and scatterplot visualization revealed strong positive relationships between ACS_TOT_POP_POV_sum (later renamed `Total_population_poverty`), `Pct_single_parent`, and `Bed_util_ratio`. These patterns were reinforced by the correlation heatmap and aligned with our expectations regarding the influence of structural socioeconomic factors on hospital bed utilization. Geographic analysis also indicated that counties in the South and in more rural areas tended to have higher utilization ratios. This may reflect regional disparities in healthcare access, infrastructure, or underlying population needs.

Full EDA outputs, including correlation plots, summary statistics, and visualizations, are provided in Appendix C.

3.4 Exploratory Clustering Analysis

As an exploratory complement to our supervised modeling, we performed unsupervised clustering to better understand how counties group together based on structural characteristics. We used Principal Component Analysis (PCA) for dimensionality reduction and applied K-Means clustering with k=3, which yielded three distinct county profiles.

Cluster 1 was characterized by lower rates of single-parent households, fewer renters, and low housing cost burden. These counties reflect relatively stable housing conditions, higher rates of homeownership, and more traditional household structures.

Cluster 2 showed high percentages of mobile home residency and disability, along with lower rates of household internet access. These traits are indicative of structurally vulnerable, often rural or isolated communities with limited digital infrastructure and greater social service needs.

Cluster 3 included counties with high population density, high median household income, and widespread internet access. These areas likely represent urban or suburban communities with stronger local economies and better access to technology and services.

While clustering was not used in the final predictive model, it provided valuable context for understanding population-level variation across U.S. counties. These profiles can inform regional strategies for targeted planning and resource allocation. Visualizations and loadings for these clusters are included in Appendix H.

4 Modeling and Evaluation

4.1 Modeling Approach

We evaluated four machine learning models:

- Support Vector Regression (SVR)
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost Regressor

The dataset was divided into training and testing sets using a dynamic ratio calculated based on the number of predictors. This resulted in a split approximately equivalent to 80% training and 20% testing, ensuring optimal generalization capacity for model evaluation, and hyperparameter tuning via GridSearchCV. Parallel pipelines were tested with and without principal component analysis (PCA) for dimensionality reduction.

4.2 Model Performance

Across all models, those trained without PCA consistently outperformed their PCA-reduced counterparts, confirming that even lower-variance features contributed meaningful predictive signals. Among the models, XGBoost Regressor trained on the original feature set achieved the best test set performance with a Test R^2 of 0.3617 and a Test RMSE of 0.1378.

Model	Train R^2	Test R^2	Train RMSE	Test RMSE	Notes
SVR (No PCA)	**0.1721**	**0.1542**	**0.5691**	**0.6271**	**Underfit**

SVR (PCA)	**0.3062**	**0.1796**	**0.521**	**0.6176**	**Weak fit**
Random Forest (No PCA)	**0.4768**	**0.3524**	**0.1185**	**0.1388**	**Competitive**
Random Forest (PCA)	**0.6505**	**0.1695**	**0.3698**	**0.6214**	**Overfit**
Gradient Boosting (No PCA)	**0.5018**	**0.3558**	**0.1156**	**0.1384**	**Strong**
Gradient Boosting (PCA)	**0.2566**	**0.1395**	**0.5393**	**0.6325**	**Poor fit**
XGBoost (No PCA)	**0.4956**	**0.3617**	**0.1163**	**0.1378**	**Best Performance**
XGBoost (PCA)	**0.2421**	**0.1393**	**0.5445**	**0.6326**	**Weak fit**

4.3 Hyperparameter Tuning

The final XGBoost model was tuned using GridSearchCV, evaluating learning rate, tree depth, and number of estimators. The optimal hyperparameters identified were:

- Learning Rate: 0.05
- Max Depth: 3
- Number of Estimators: 100

Expanded hyperparameter tuning grids were explored for SVR and XGBoost models; however, further tuning did not yield significant improvements in test set performance compared to the original tuned XGBoost model.

4.4 Feature Importance

Feature importance analysis using the final XGBoost model revealed that `Total_population_poverty` remained the most dominant predictor of hospital bed utilization, with an F-score of 114—more than twice the importance of any other variable. This strong signal reinforces the critical role of structural economic hardship in driving baseline strain on healthcare systems.

The next most influential features were `Pct_owner_cost_30plus` and `Pct_single_parent`, followed by `Pct_disabled` and `Distance_to_medsurge_icu`. These variables collectively reflect the burden of housing costs, household structure, physical vulnerability, and spatial access to critical care. Their prominence supports the conclusion that social and geographic barriers—rather than clinical metrics alone—significantly influence hospital utilization patterns.

Additional contributors included `Pct_renter_occupied`, `Pct_homes_no_vehicle`, and `Pct_hh_65_alone`, highlighting the importance of mobility constraints and aging-related vulnerability. Geographic region indicators such as `Region_South` and `Region_West` appeared in the top 20 but showed limited standalone influence. Similarly, the urban-rural classification (`Is_Metro_Micro`) had the lowest relative importance in the final model, suggesting that SDOH impacts extend across both metro and non-metro areas.

These findings reaffirm the interpretability and public health relevance of the model's predictors. A full breakdown of feature importance is provided in Appendix E.

4.5 Final Model Evaluation

Diagnostic plots of the final XGBoost confirmed that residuals were approximately normally distributed with mild heteroskedasticity. Predicted versus actual plots showed a strong positive association, indicating that the model was able to capture meaningful variation in hospital bed utilization across counties. This final recommended model explains approximately 36% of the variance of bed utilization ratio, which is a strong result considering the complexity and noise typically present in healthcare and social determinants of health data (Yang et al., 2023).

4.5.1 Predicted vs. Actual Bed Utilization Ratio (XGBoost Test Set)

The scatterplot identified in Figure G.1 in Appendix G compares the predicted and actual values of the non-COVID hospital bed utilization ratio for counties in the test set. Each point represents a county. The red dashed line indicates the ideal 1:1 correspondence, where predictions perfectly match actual values.

The plot shows a clear positive trend, confirming that the model captures meaningful variance in hospital bed utilization. Most points cluster around the diagonal, suggesting accurate predictions overall. Some deviation is observed among higher-utilization counties, likely reflecting local factors or noise not captured in the model. The absence of major bias above or below the line supports good generalization across the target range.

4.5.2 Residual Analysis by Region

The histogram identified in Figure G.2 in Appendix G shows the distribution of residuals from the test set. The approximately symmetric shape suggests no extreme skew or bias in model predictions. Residuals are centered near zero, supporting the conclusion that the XGBoost

model produces balanced predictions overall, without significant over or underestimation. Additionally, the boxplot identified in Figure G.3 in Appendix G compares the residuals by region.

5 Discussion

5.1 Key Insights

The results of this study confirm that non-clinical indicators, especially socioeconomic factors, are central to understanding patterns of hospital bed utilization. Population poverty, housing cost burden, and household composition (particularly single-parent rates) emerged as the most influential features in the model. While geographic factors such as population density and metro status contributed to prediction performance, their relative influence was considerably lower.

We also found that reducing dimensionality through PCA diminished model accuracy. This reinforces the value of retaining full feature sets in social data modeling, as even features with limited individual variance may offer important predictive signals when combined. Moreover, the use of composite indices—constructed during feature engineering—helped amplify meaningful relationships while reducing multicollinearity.

In addition to model performance metrics, both residual analysis and unsupervised clustering helped reveal patterns of regional variability and potential model limitations tied to unmeasured geographic context. Unsupervised clustering, in particular, identified distinct geographic groupings based on structural characteristics. For example, counties in the Southeast

and rural West consistently clustered together as socially and economically vulnerable, aligning with areas where prediction residuals were higher. These results suggest the presence of unmeasured or region-specific factors not fully captured in feature importance analysis alone. (See Appendix G, Figure G.1.)

XGBoost outperformed all other models in terms of generalization to unseen data, demonstrating the value of ensemble tree-based methods in capturing nonlinear effects in complex, noisy datasets. Achieving a Test R^2 of 0.3617 is particularly meaningful given the inherent variability and unmeasured confounders in SDOH datasets.

5.2 Policy Implications

This model offers an evidence-based foundation for proactive capacity planning and resource allocation by public health agencies. Predictions from the final model can be used to:

- Identify high-risk counties for system strain before actual surges occur.
- Support equitable distribution of healthcare investments, such as emergency preparedness funding, mobile health units, or targeted outreach programs.
- Justify grant applications or regional planning efforts by demonstrating model-based risk.
- Integrate structural indicators of vulnerability into ongoing health needs assessments.

As a flexible and transparent framework, the model can be adapted over time with updated data to refine predictions and monitor trends in baseline system strain.

5.3 Limitations

There are several limitations to this study.

1. While the model provides meaningful predictions, it does not establish causality between SDOH factors and hospital bed utilization.
2. The analysis is restricted to county-level data, which may obscure disparities within large or heterogeneous counties.
3. While the COVID-related components of hospitalization were excluded from the target variable, 2020 remained a unique year with lingering pandemic effects on both care-seeking behavior and hospital operations. This limits the model's ability to generalize across different years or predict future bed utilization under non-pandemic conditions.
4. Certain potentially important variables—such as primary care provider density or outpatient utilization—were not available in our dataset.
5. Data quality and completeness varied across counties, and preprocessing steps like imputation or the removal of rows could introduce bias.

5.4 Future Work

Future research may extend this work in several directions. Incorporating multi-year or seasonal data could allow for time-series modeling to monitor changes in baseline system demand. More granular analysis at the census tract or ZIP code level could improve targeting of interventions within counties. Additionally, causal modeling frameworks such as directed acyclic graphs (DAGs) or instrumental variable approaches could help clarify pathways of influence. Finally,

integrating environmental exposures such as pollution levels or climate vulnerability—may offer a more complete picture of systemic risk.

6 Conclusion

This study confirms that structural community-level vulnerabilities particularly poverty, housing cost burden, family structure, disability status, and access to critical care can meaningfully predict hospital bed utilization at the county level, even in the absence of a pandemic. Our final model, based on XGBoost and a carefully engineered feature set, enables public health stakeholders to move beyond reactive capacity reporting and proactively identify at-risk regions.

The XGBoost model trained on our carefully engineered feature set provides an optimal balance between predictive accuracy, interpretability, and generalizability for hospital bed utilization prediction. The selected approach offers several advantages:

- Best predictive accuracy among all tested models
- Robust performance on unseen test data
- Effective handling of missing data and outliers
- Preservation of feature interpretability
- Clear and defensible methodology

Public health agencies can use this model to identify areas at higher risk of system strain and prioritize interventions accordingly. While predictive accuracy is not perfect, the insights

generated provide a replicable, transparent framework for monitoring vulnerability and advancing equity. This work affirms that hospital system stress is not just a clinical issue but a structural one and addressing it begins with understanding the communities most at risk.

Appendices

The following appendices provide technical documentation, supplementary visuals, and extended results that support the findings and methodology described in the main report. These materials are intended for readers who wish to explore the project's data pipeline, exploratory analysis, modeling diagnostics, and clustering insights in greater detail. While the main body of the report is designed to be accessible to public health stakeholders, the appendices offer transparency and reproducibility for data science practitioners, technical reviewers, and future collaborators.

Appendix A: Data Cleansing, Wrangling, and Merging Datasets

This project used two datasets: Social Determinants of Health from the Agency for Healthcare Research and Quality and COVID-19 Reported Patient Impact and Hospital Capacity by Facility from HealthData.gov. The data cleaning, wrangling and merging was done using MSSQL Server and the scripts and steps involved can be found in the PreliminaryMergedData section of development_test of the GitHub repository; (https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/development_test/data/PreliminaryMergedData). Data cleaning was performed on both datasets, including replacing empty strings and '-999999' with zeros, and verifying the validity of zip code and FIPS values. Grouping and aggregation was needed in both datasets to have mergeable records. The FIPS-level bed utilization rate was calculated on both datasets as `Bed_util_ratio`. The datasets were merged into a single dataset.

Appendix B: Data Preprocessing Technical Details

To properly prepare the merged dataset for analysis, a number of preprocessing, encoding and imputation steps were performed. This includes:

- Defining a function for train-test split calculation
- Dropping unneeded features
- Removing records with missing values
- Splitting the dataset into training and test subsets
- Encoding `Is_Metro_Micro` to 0 and 1
- Performing one-hot encoding on Region into `Region_Midwest`, `Region_West`, `Region_South`, and `Region_Northeast` features
- Scaling data using StandardScalar
- Imputing data using KNNImputer
- Exporting the resulting dataset for future use

The resulting dataset was evaluated by reviewing the distribution of the training and test sets, comparing the distribution of values before and after binary encoding, performing Shapiro-Wilk testing and Q-Q normality check on `Bed_util_ratio` values, as well as reviewing the distribution of those values. The evaluation identified that the target values were not normally distributed, but this was something to keep in mind as the analysis progressed.

The Python notebook for this appendix is called `Appendix_B_Data_Processing_Technical_Details.ipynb` which can be found at https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/final_delivery/Appendix

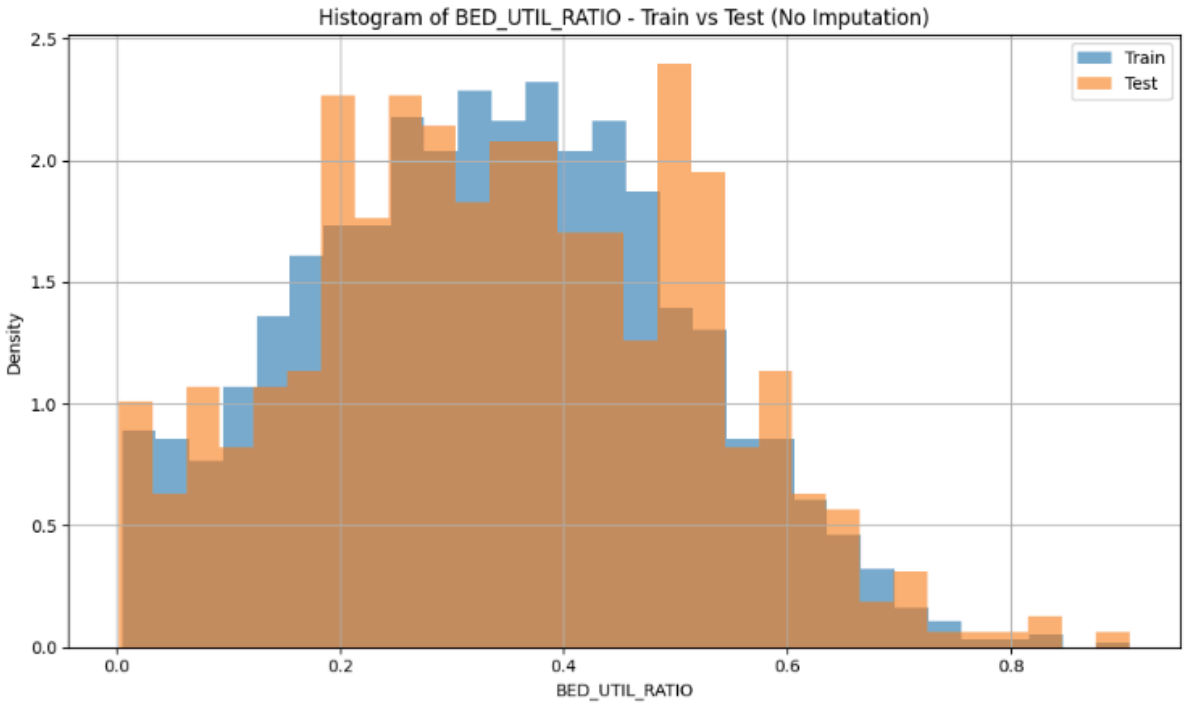

Figure B.1 Histogram of BED_UTIL_RATIO after train-test splitting showing uniform distribution between training and test sets but also showing skewness

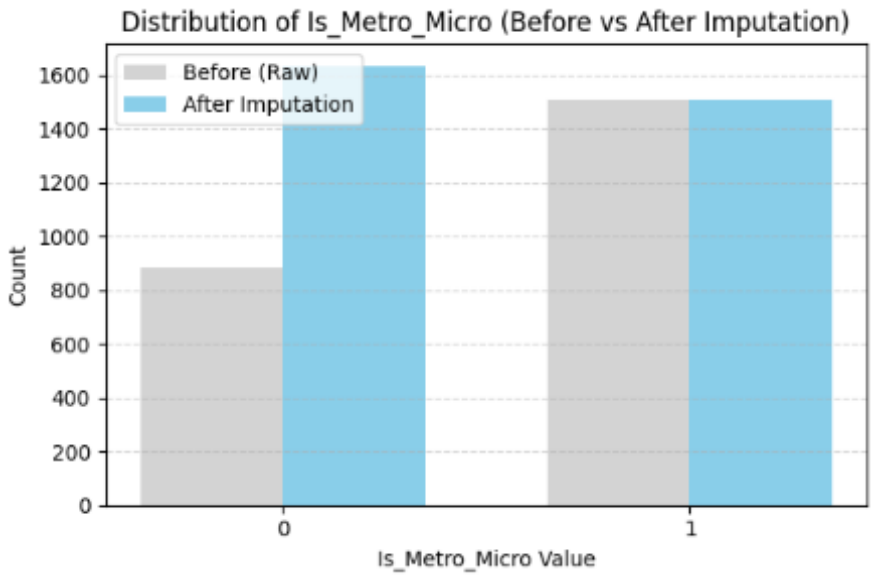

Figure B.2 Balanced distribution before and after Imputation

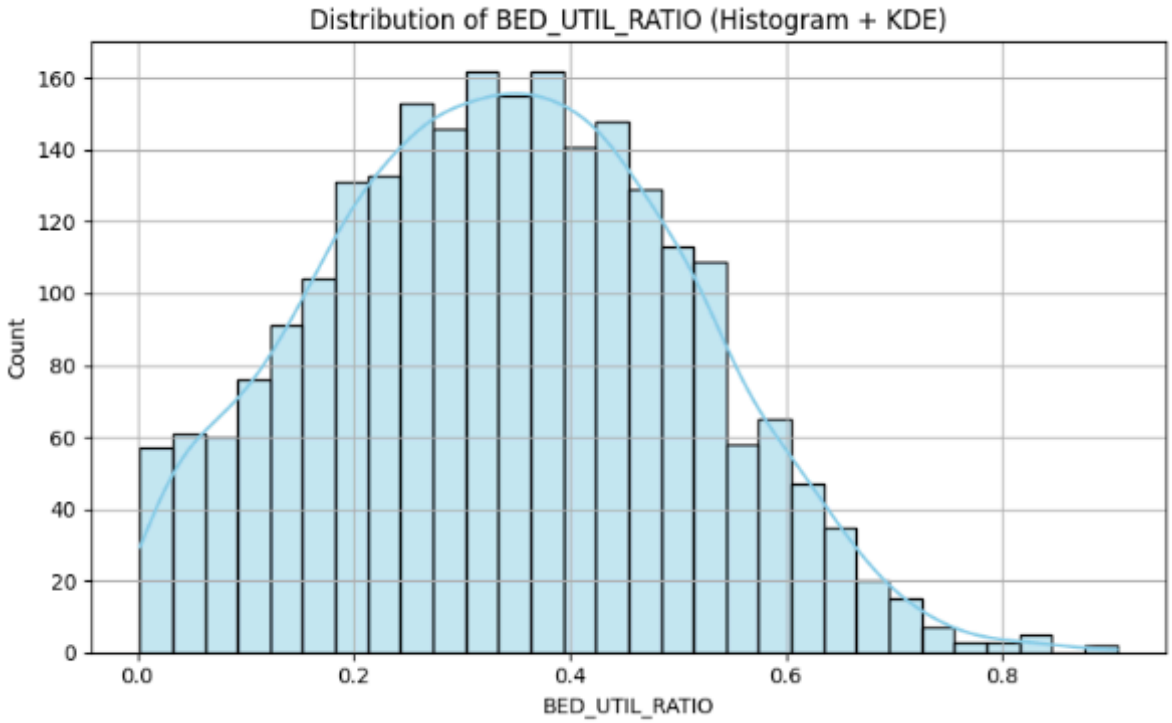

Figure B.3 Skewed distribution of target values

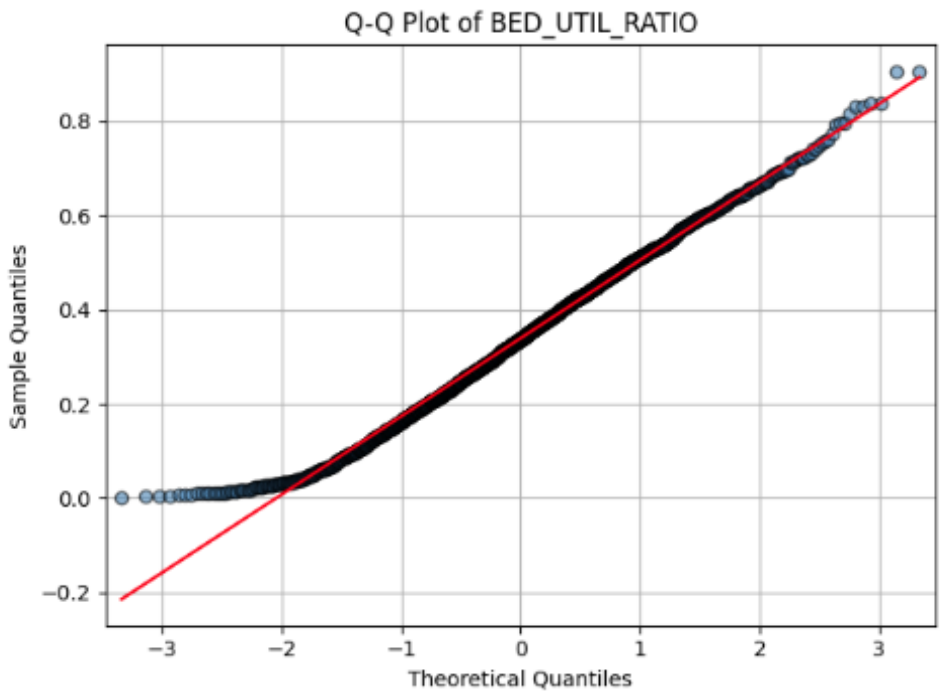

Figure B.4 Q-Q Plot showing skewed data

Appendix C: Exploratory Data Analysis (EDA) Results

Summary Statistics

- Descriptive statistics were computed for all numeric features
- Features showed varied ranges and distributions, indicating the need for normalization or transformation prior to modeling
- Skewness was also observed in several variables

```
--- Main Descriptive Statistics ---
                           count     mean       std     min      25%      50%      75%        max
CountyFIPS                3239.00 31504.66  16462.99 1001.00 19032.00 30033.00 46126.00   78030.00
Pct_renter_occupied       3234.00    26.78      9.63    0.00    21.33    26.00    31.80      82.00
Distance_to_medsurge_icu  3234.00    12.57     17.67    0.00     4.67     7.30    16.50     487.00
Pct_renter_cost_50pct_plus 3234.00   18.63      7.79    0.00    14.28    18.84    23.29      75.00
Pct_renter_cost_30pct_plus 3234.00   39.68     12.04    0.00    34.00    41.00    46.76     100.00
Total_population_poverty  3234.00 98504.68 322472.55    0.00  9612.25 23553.00 63324.00 9884138.00
Pct_owner_cost_30plus     3234.00    17.71      5.73    0.00    14.83    17.50    20.70      42.75
Population_density        3234.00   891.21   2798.16    0.00    41.52   209.10   907.02   90386.23
Land_area_sqmi            3234.00   289.39   1296.98    0.00    27.86    72.00   173.32   36393.25
Median_hh_income          3234.00 53803.78  17732.18    0.00 44620.50 52881.50 61834.75  148523.00
Pct_hh_65_alone           3234.00    12.68      4.14    0.00    10.62    12.75    14.80      66.00
Pct_age_65plus            3234.00    18.82      5.72    0.00    16.00    18.82    21.67      49.86
Pct_homes_no_vehicle      3234.00     5.96      4.55    0.00     3.75     5.31     7.25      89.00
Pct_public_transit        3234.00     0.91      3.06    0.00     0.00     0.25     0.72      59.30
Pct_single_parent         3234.00    28.57     11.43    0.00    22.11    28.16    34.53      97.00
Pct_hh_no_internet        3234.00    75.69     15.46    0.00    72.76    79.00    83.50      97.00
Distance_to_ED            3234.00     8.01     11.38    0.00     4.00     5.67     8.50     361.00
Pct_mobile_homes          3234.00    12.29      9.60    0.00     4.69    10.00    18.00      56.50
Pct_disabled              3234.00    15.58      5.20    0.00    12.63    15.33    18.60      48.00
Is_Metro_Micro            2421.00     0.63      0.48    0.00     0.00     1.00     1.00       1.00
Bed_util_ratio            2421.00     0.34      0.17    0.00     0.22     0.34     0.46       0.91
```

Figure C.1 Main Descriptive Statistics

```
--- Skewness and Kurtosis ---
                          skewness  kurtosis
CountyFIPS                    0.17     -0.61
Pct_renter_occupied           0.30      2.21
Distance_to_medsurge_icu     11.04    215.93
Pct_renter_cost_50pct_plus    0.21      2.48
Pct_renter_cost_30pct_plus   -0.95      2.54
Total_population_poverty     13.74    314.35
Pct_owner_cost_30plus        -0.22      2.15
Population_density           17.23    436.21
Land_area_sqmi               17.11    367.58
Median_hh_income              0.28      3.51
Pct_hh_65_alone               0.51     11.21
Pct_age_65plus               -0.35      2.96
Pct_homes_no_vehicle          6.70     84.15
Pct_public_transit           11.09    162.00
Pct_single_parent             0.34      1.88
Pct_hh_no_internet           -3.36     13.64
Distance_to_ED               15.35    377.57
Pct_mobile_homes              0.99      0.73
Pct_disabled                 -0.10      1.89
Is_Metro_Micro               -0.55     -1.70
Bed_util_ratio                0.13     -0.44
```

Figure C.2 Skewness and Kurtosis

Distribution Analysis

- Histogram was used to examine the distribution of `Bed_util_ratio`

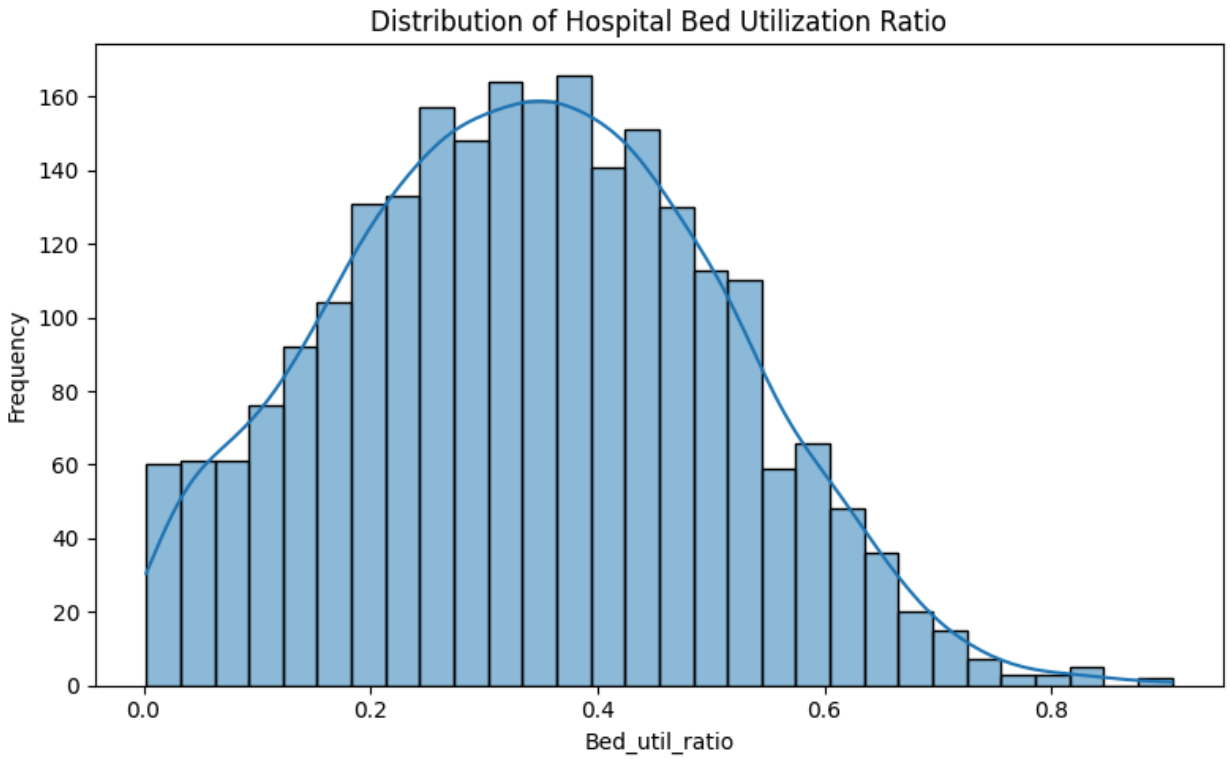

Figure C.3 Distribution of Hospital Bed Utilization Ratio

Heatmap of Hospital Utilization Across the U.S.

- Spot geographic patterns of strain or surplus
- Identify outliers or regional clusters of high utilization

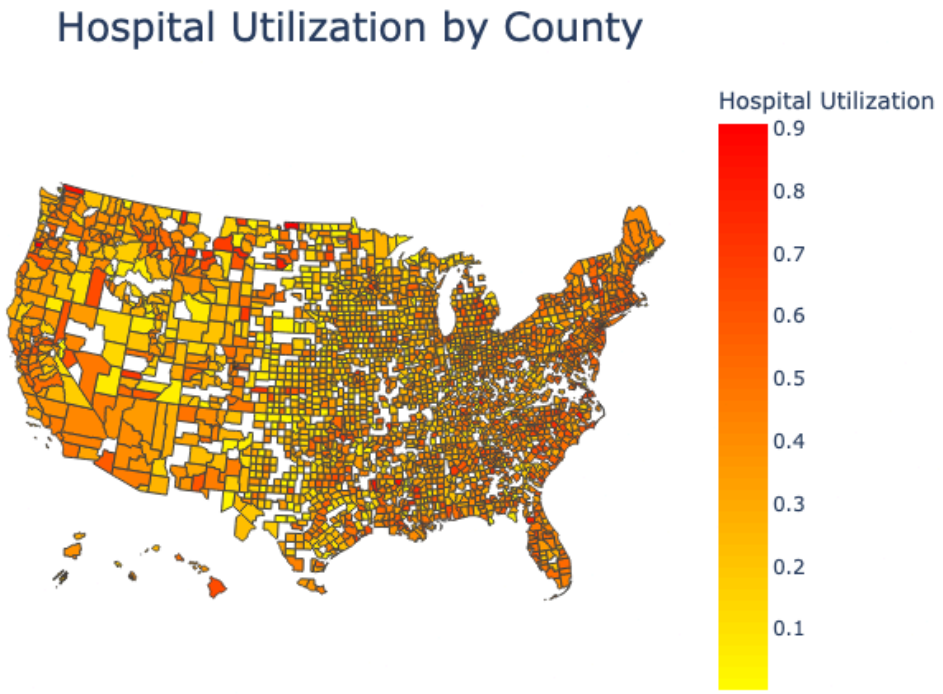

Figure C.4 Choropleth Map of Hospital Utilization by County

Appendix D: Full Model Comparison Results

Presents complete performance metrics (R^2, RMSE, MAE) for all models with and without PCA, and compares pipeline performance across preprocessing strategies. The Python notebook for this appendix is called `Appendix_D_Full_Model_Comparison_Results.ipynb` which can be found at

https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/final_delivery/Appendix

Model	XGBoost	XGBoost
Characteristics	No PCA	PCA
CV Strategy	GridSearchCV (Kfold)	GridSearchCV (Kfold)
splits	5	5
candidates	8	8
Shuffle	TRUE	TRUE
random_state	42	42
Estimators	[100, 200]	[100, 200]
max depth	[3, 5]	[3, 5]
learning rate	[0.05, 0.1]	[0.05, 0.1]
Best learning rate	0.05	0.05
Best max depth	3	3
Best estimators	100	100
Train RMSE	0.1163	0.5445
Train R^2	0.4956	0.2421
Train MAE	0.0917	0.3083
Test RMSE	0.1378	0.6326
Test R^2	0.3617	0.1393
Test MAE	0.1057	0.3668
Conclusion	Best Performance	Weak fit

Table D.1 Identifying XGBoost without PCA to be the best model

Model	Random Forest	Random Forest
Characteristics	No PCA	PCA
CV Strategy	GridSearchCV (Kfold)	GridSearchCV (Kfold)
splits	5	5
candidates	12	12
Shuffle	TRUE	TRUE
random_state	42	42
Estimators	[100, 200]	[100, 200]
max depth	[5, 10, None]	[5, 10, None]
min sample split	[2, 5]	[2, 5]
Best max depth	5	10
Best min sample split	2	5
Best estimators	200	200
Train RMSE	0.1185	0.3698
Train R^2	0.4768	0.6505
Train MAE	0.0939	0.1987
Test RMSE	0.1388	0.6214
Test R^2	0.3524	0.1695
Test Mae	0.1063	0.3608
Conclusion	Competitive	Overfit

Table D.2 Identifying Random Forest without PCA as an alternative to XGBoost without PCA

Model	Gradient Boosting	Gradient Boosting
Characteristics	No PCA	PCA
CV Strategy	GridSearchCV (Kfold)	GridSearchCV (Kfold)
splits	5	5
candidates	8	8
Shuffle	TRUE	TRUE
random_state	42	42
Estimators	[100, 200]	[100, 200]
max depth	[3, 5]	[3, 5]
learning rate	[0.05, 0.1]	[0.05, 0.1]
Best learning rate	0.05	0.05
Best max depth	3	3
Best estimators	100	100
Train RMSE	0.1156	0.5393
Train R^2	0.5018	0.2566
Train MAE	0.0917	0.3068
Test RMSE	0.1384	0.6325
Test R^2	0.3558	0.1395
Test MAE	0.1061	0.3686
Conclusion	Strong	Poor fit

Table D.3 Identifying Gradient Boosting without PCA as a strong model option

Model	SVR	SVR
Characteristics	No PCA	PCA
CV Strategy	GridSearchCV (Kfold)	GridSearchCV (Kfold)
splits	5	5
candidates	30	6
Shuffle	TRUE	TRUE
random_state	42	42
Regulation Parameters (C)	[0.1, 1, 10]	[1, 10, 100]
Tube Width (epsilon)	[0.01, 0.1]	[0.001, 0.01]
Kernel	['linear', 'rbf']	['rbf']
Best C	0.1	1
Best epsilon	0.01	0.01
Best kernel	rbf	rbf
Train RMSE	0.5691	0.521
Train R^2	0.1721	0.3062
Train MAE	0.3212	0.2748
Test RMSE	0.6271	0.6176
Test R^2	0.1542	0.1796
Test MAE	0.3659	0.3581
Conclusion	Underfit	Weak fit

Table D.4 Identifying that SVR was not a good match for our data

Appendix E: Hyperparameter Tuning Details

Documents the parameter tuning grids and selected configurations for all models, including expanded XGBoost tuning tests.

Learning Rate	N Estimators	Max Depth
0.05	100	3
0.05	100	5
0.05	200	3
0.05	200	5
0.1	100	3
0.1	100	5
0.1	200	3
0.1	200	5

The Python notebook for this appendix is called `Appendix_E_Hyperparameter Tuning Details.ipynb` which can be found at

https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/final_delivery/Appendix

Appendix F: Feature Importance Analysis

Socioeconomic distress indicators continue to dominate the top features, with poverty having exceptionally high influence. `Total_population_poverty` had an F-score of 114, which was more than double that of the next most important feature. Housing cost burden

(`Pct_owner_cost_30plus`) and family structure (`Pct_single_parent`) were nearly tied for second place in importance, followed by disability rates and distance to medical/surgical ICU facilities. The Python notebook for this appendix is called `Appendix_F_Top_20_Features_by_F_score.ipynb` which can be found at https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/final_delivery/Appendix

XGBoost's feature importance analysis revealed the following top contributors:

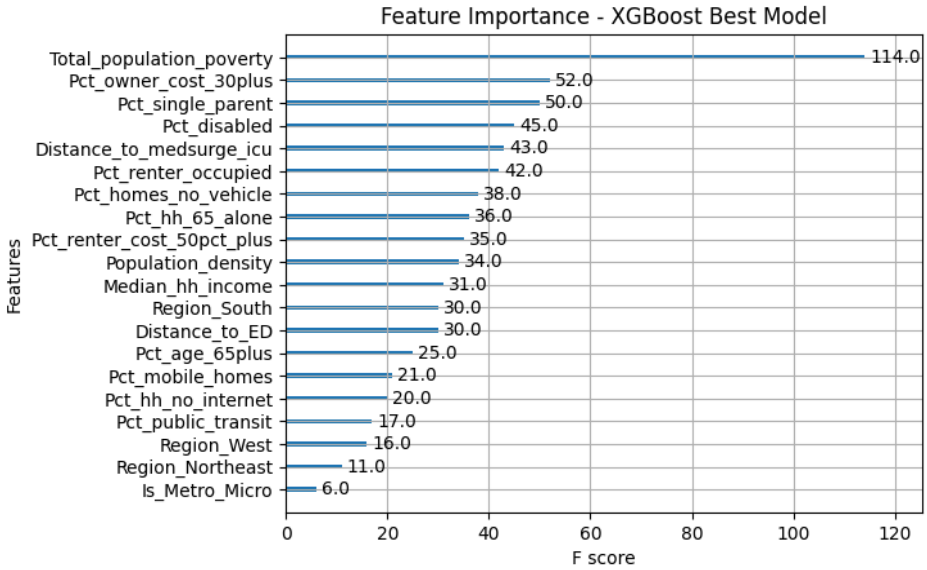

Figure F.1 Feature importance by F score demonstrating weight of feature Total_population_poverty

Feature Importance by Category:

Socioeconomic Indicators

These dominated the model's predictions:

`Total_population_poverty`, `Pct_owner_cost_30plus`, and `Pct_single_parent` were the top 3 contributors.

Median_hh_income, Pct_disabled, and Pct_renter_occupied also held moderate-to-high importance.

Housing Burden & Access Variables

Housing cost (Pct_owner_cost_30plus, Pct_renter_cost_50pct_plus)

Transportation limitations (Pct_homes_no_vehicle, Pct_public_transit, Pct_hh_no_internet)

Aging households and disability (Pct_hh_65_alone, Pct_age_65plus, Pct_disabled)

Geographic Access Factors

Distance_to_medsurge_icu and Distance_to_ED emerged as important, reinforcing spatial barriers to care.

However, Is_Metro_Micro had the lowest F-score, suggesting structural vulnerability transcends metro status.

Region Indicators

Encoded region indicators (Region_South, Region_West, Region_Northeast) appeared in the mid-to-lower range of the importance list.

This suggests regional variation exists but is overshadowed by more direct structural indicators.

Appendix G: Diagnostic Plots and Residual Analysis

This appendix includes diagnostic plots evaluating the final XGBoost model. The Python notebook for this appendix is called `Appendix_G_Diagnostic_Plots_and_Residual_Analysisipynb.ipynb` which can be found at https://github.com/lemieuxjm-cap/Iota-Capstone/blob/main/final_delivery/Appendix/Appendix_G_Diagnostic_Plots_and_Residual_Analysisipynb.ipynb. Visuals include:

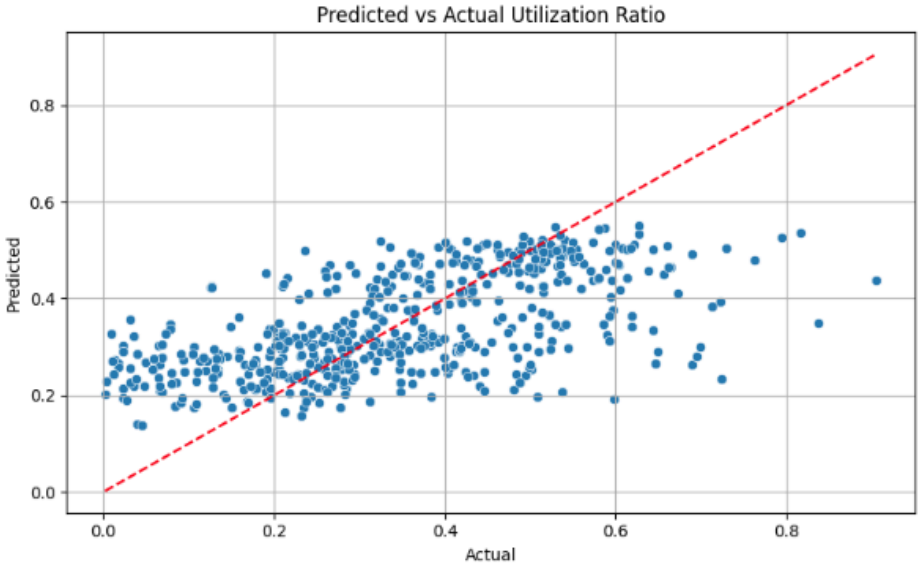

Figure G.1 Predicted vs. Actual Utilization Ratio

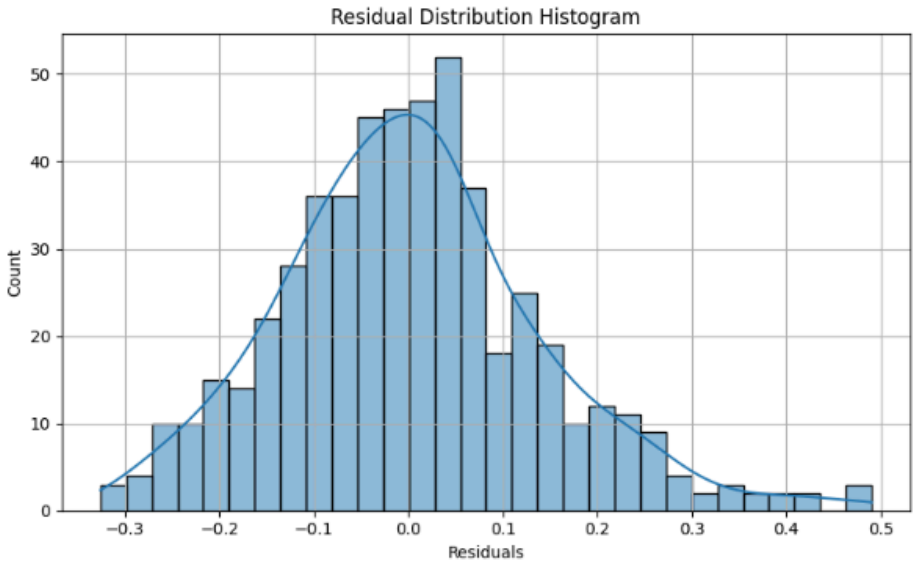

Figure G.2 Residual Distribution Histogram

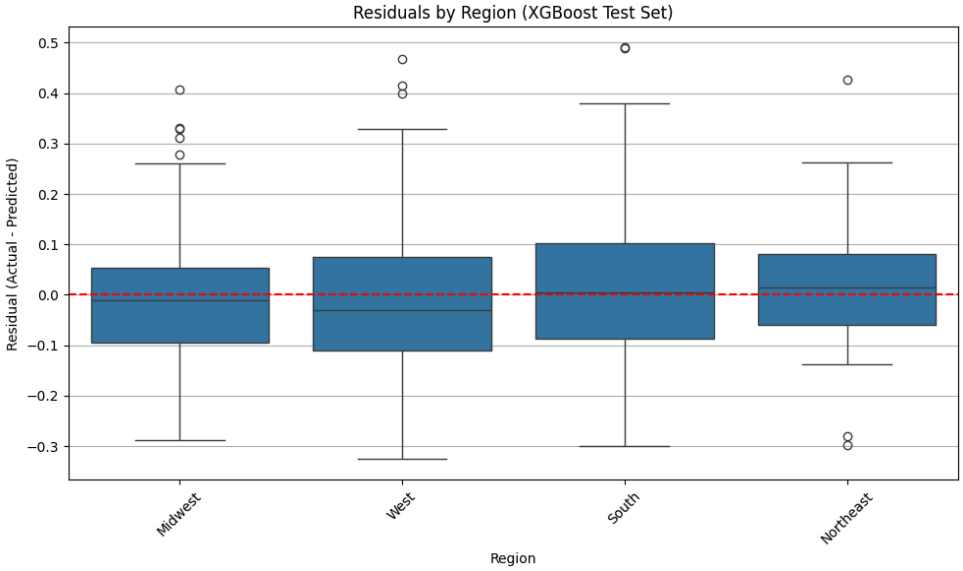

Figure G.3 Residuals by region illustrating overall balanced actual-to-predicted values with slight underpredicting in the West region

These plots assess prediction accuracy, residual distribution symmetry, potential regional bias, normality of residuals, and variance behavior across predicted values.

Appendix H: Clustering Analysis Results

This appendix presents the results of unsupervised clustering using PCA and K-Means. It includes a county-level cluster map, silhouette score evaluation, visualizations of the first two PCA components, and detailed cluster profiles highlighting key structural traits (e.g., poverty, disability, internet access). These analyses support interpretation of regional patterns in vulnerability and complement residual-based model evaluation.

Cluster	Features (Loadings)	Key Traits	Interpretation
0	Pct_single_parent (-0.51) Pct_renter_occupied (-0.49) Pct_renter_cost_50pct_plus (-0.49)	Lower single-parent households, lower renters, and few cost-burdened renters	Higher homeownership rates, lower housing cost burden, and more traditional household structures
1	Pct_mobile_homes (0.88) Pct_disabled (0.84) Pct_hh_no_internet (-0.82)	High % of mobile homes, residents with disabilities, and lower households with no internet	Rural or isolated counties with structural access barriers, higher vulnerability due to disability, and poor digital access
2	Population_density (1.10) Median_hh_income (1.04) Pct_hh_no_internet (0.88)	High population density, high median household income, higher % with internet access	Reflects stronger economies, likely urban or suburban, with some areas of digital divide

Figure H.1 Summary of clusters

Appendix H.1.: County-Level Clusters Based on SDOH Features (K-Means, k=3)

This map displays the results of a K-Means clustering algorithm applied to U.S. counties using selected social determinants of health (SDOH) features. Each county is assigned to one of three clusters, represented by distinct colors. The Python notebook for this appendix is called

`Appendix_H_Clustering_Analysis_Results.ipynb` which can be found at

https://github.com/lemieuxjm-cap/Iota-Capstone/tree/main/final_delivery/Appendix

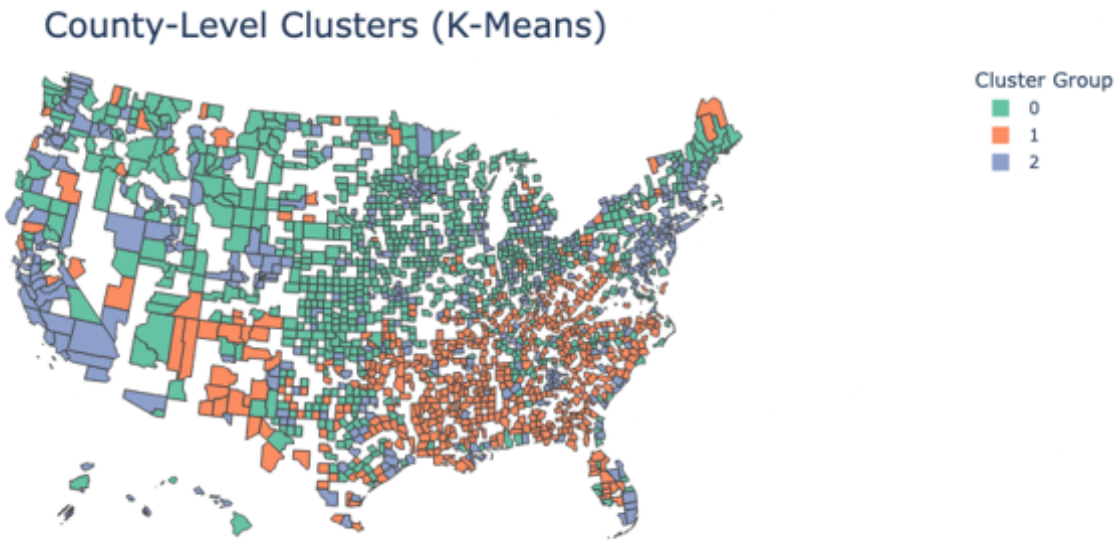

Figure H.2 Counties associated with cluster groups after exploratory data analysis using K-Means(k=3)

The clustering reveals clear spatial patterns:

- Cluster 0 (teal) captures a mixed profile, including rural Western counties with mobile home prevalence and access-related barriers.
- Cluster 1 (orange) represents structurally vulnerable counties, concentrated in the Southeast, with higher poverty and disability rates, and limited healthcare infrastructure.
- Cluster 2 (periwinkle) consists largely of urban or suburban counties with stronger economic indicators, higher rates of internet access, and lower disability prevalence.

Findings:

The clustering highlights how counties group into structurally similar profiles that transcend simple geographic labels. These clusters help explain regional disparities observed in residual analysis and suggest that unmeasured social and infrastructural differences—not captured fully in linear models—may influence healthcare demand patterns. Integrating this type of unsupervised learning provides valuable context for interpreting model performance and informing targeted public health interventions.

Appendix I: Code and GitHub Repository

https://github.com/lemieuxjm-cap/Iota-Capstone/

Direct link to the GitHub repository containing source code, modeling scripts, final datasets, and visualizations to support full reproducibility

Appendix J: Data Dictionary

https://github.com/lemieuxjm-cap/Iota-Capstone/blob/main/final_delivery/data/ReadMe_for_DataDictionary.md

This link will provide access to the data dictionary in the GitHub repository.

References

Artiga, S., & Hinton, E. (2018, May 10). *Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity*. KFF. Retrieved April 26, 2025, from https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/

Garg, S., Kim, L., Whitaker, M., & et al. (2020, April 17). Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 —

COVID-NET, 14 States, March 1–30, 2020. *Morbidity and Mortality Weekly Report*, *69*(15), 458-464. http://dx.doi.org/10.15585/mmwr.mm6915e3

Kreuter, M. W., Thompson, T., McQueen, A., & Garg, R. (2021, April). Addressing Social Needs in Health Care Settings: Evidence, Challenges, and Opportunities for Public Health. *Annual Review of Public Health*, *42*(1), 329-344. https://doi.org/10.1146/annurev-publhealth-090419-102204

Magnan, S. (2017, October 9). *Social Determinants of Health 101 for Health Care: Five Plus Five* (Issue https://doi.org/10.31478/201710c) [Discussion Paper]. Social Determinants of Health 101 for Health Care: Five Plus Five. Retrieved April 26, 2025, from https://nam.edu/perspectives/social-determinants-of-health-101-for-health-care-five-plus-five/

Rahmatinejad, Z., Dehghani, T., Hoseini, B., Rahmatinejad, F., Lotfata, A., Reihani, H., & Eslami, S. (2024). A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department. *Nature Scientific Reports*, *14*(3406), 1. https://doi.org/10.1038/s41598-024-54038-4

Yang, M. Y., Kwak, G. H., Pollard, T., Celi, L. A., & Ghassemi, M. (2023, August 29). *Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit* [AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society 2023 Proceeding]. Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit. https://doi.org/10.1145/3600211.3604719

Zeleke, A. J., Palumbo, P., Tubertini, P., Miglio, R., & Chiari, L. (2024). Comparison of nine machine learning regression models in predicting hospital length of stay for patients

admitted to a general medicine department. *Informatics in Medicine Unlocked*, *47*(101499), 1. https://doi.org/10.1016/j.imu.2024.101499