Model Selection and Development Report

Project Title: Hospital Bed Utilization Monitoring Using SDOH Data (County-Level, 2020)

Team IOTA

Team Lead: Mehmet Comert

Recorder: Melody Rios

Spokesperson: June Lemieux

Floater: Rea Kelolli

Recap Background & Question

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

Hypothesis & Prediction:
We hypothesize that communities with more residents who are older, uninsured, living with disabilities or poverty, or lacking access to nearby care are more likely to experience higher hospital bed usage—even when we factor in things like geography and environmental conditions. Previous research has shown that these social and demographic factors play a big role in shaping healthcare demand, especially during the COVID-19 pandemic, where they strongly correlated with hospitalization patterns.

Introduction

Hospital bed utilization is a critical indicator of healthcare system stress, particularly during periods of widespread illness, policy disruption, or resource strain. However, real-time data on hospital capacity can be limited or delayed, especially in rural or under-resourced regions (HHS, 2020), making it challenging to develop timely interventions based on comprehensive health datasets (AHRQ, 2020). As a result, public health stakeholders and policymakers need tools to proactively monitor and anticipate healthcare strain using stable and accessible indicators.

This project explores whether community-level social and structural determinants of health (SDOH)—such as poverty rates, population density, healthcare access, and insurance coverage—can serve as proxies for predicting hospital bed utilization. We selected the 2020 AHRQ SDOH database and the U.S. Department of Health and Human Services' COVID-19 Hospitalization Dataset (2020) due to their completeness, wide geographic scope, and ease of integration at the county level (AHRQ, 2020; HHS, 2020). The BED_UTIL_RATIO target variable was derived from the HHS dataset using the formula:

BED_UTIL_RATIO = (Inpatient Beds Used – COVID Beds Used) / Total Inpatient Beds.

This measure improves interpretability and policy relevance for stakeholders focused on systemic strain. The merged dataset provides a high-resolution snapshot of how local health determinants relate to system burden.

In this report, we present our Exploratory Data Analysis (EDA) to examine the distribution of key variables, detect missingness, assess skewness and multicollinearity, and visualize relationships between features and the target variable. These results will guide model building and feature selection in future phases of the project. We also emphasize the importance of using publicly available datasets for analytical transparency, policy relevance, and equity-focused research. This approach not only supports reproducibility, but also enables diverse stakeholders—including local governments, researchers, and healthcare administrators—to engage with and act on the insights produced.

Methods:

Our project investigates whether social and geographic determinants of health can predict county-level non-COVID hospital bed utilization in 2020. Based on our exploratory data analysis

(EDA), we initially hypothesized that counties with higher proportions of older adults, uninsured individuals, residents with disabilities, and limited healthcare access would experience higher hospital bed utilization.

Data Cleaning and Filtering: Our goal was to assess whether SDOH factors could be used to predict the continuous outcome Bed_util_ratio. To start, we performed extensive preprocessing and filtering steps on a combined dataset containing 3,239 county-level records.

We first excluded counties with missing Region values and removed District of Columbia due to its outlier characteristics. Redundant or collinear variables, such as State and Land_area_sqmi, were also dropped following early modeling diagnostics. We retained two versions of the dataset: one with complete target values for modeling (n = 2,391) and one with missing targets for future prediction or policy simulation (n = 751).

Target Normality Assessment: The binary variable Is_Metro_Micro was imputed using KNN imputation and later re-coded as 0 or 1. The Region variable was one-hot encoded, and all continuous predictors were scaled using StandardScaler.

We used a stratified train-test split based on a calculated ratio (~78% train / 22% test) and examined the distribution of the target variable, Bed_util_ratio. This variable was substantially right-skewed, with a Shapiro-Wilk $W = 0.9919$, $p < 0.001$, rejecting the assumption of normality. We applied a Box-Cox transformation to stabilize variance and improve distributional symmetry, which improved the shape but still fell short of full normality.

Dimensionality Reduction with PCA: Principal Component Analysis (PCA) is applied to reduce dimensionality while retaining structure. PCA reduced the 21 predictor variables to 16 principal components while preserving over 95% of the variance. The first few components captured over 40% of the total variance, with loadings highlighting the influence of population

density, median household income, internet access, and disability rates—suggesting strong patterns of structural disadvantage and healthcare access barriers.

Using the reduced dataset, we trained a Support Vector Regression (SVR) model with an RBF kernel:

- Cross-validated R^2: 0.2989 ± 0.0262
- Test RMSE (Box-Cox): 0.8684
- Final RMSE (Original Scale): 0.8051
- Test R^2 (Box-Cox): 0.3143

Residual diagnostics for this model indicated mild deviation from normality (Shapiro-Wilk W = 0.9933, p = 0.0195), and plots showed a generally symmetric distribution centered around zero.

To benchmark against PCA, we trained an SVR on the full scaled feature set (no dimensionality reduction), which yielded comparable results:

- Cross-validated R^2: 0.2929 ± 0.0301
- Final RMSE (Original Scale): 0.8075
- Test R^2: 0.3249
- Residuals: Shapiro-Wilk W = 0.9949, p = 0.0811 (likely normal)

Unsupervised Clustering with K-Means: Lastly, K-Means clustering is used on the first two PCA components to identify latent subgroups of counties. The elbow method suggested k=3 as optimal, and the resulting Silhouette Score was 0.357, indicating moderately distinct clusters.

Visual inspection of the PCA biplot with loadings showed interpretable separations based on regional and structural variables, suggesting potential for future segmentation or profiling.

Overall, our pipeline successfully integrated imputation, transformation, dimensionality reduction, and nonlinear regression. The moderate R^2 values suggest that while SDOH indicators contain predictive value, hospital bed utilization may also depend on unmeasured temporal or geographic factors. Future steps will explore tree-based methods, binary classification thresholds, and integration of spatial clustering to improve generalization.

Cross Validation

Throughout our modeling progression, we employed cross-validation—typically 5-fold for regression models and stratified k-fold for classification tasks—to estimate generalization error and monitor for overfitting. For each model, we tracked metrics such as RMSE, R^2, and ROC AUC, which helped us iteratively refine our approach and compare models on a consistent basis.

Results and Findings:

PCA Analysis

We aimed to reduce dimensionality while preserving ≥95% of the variance in the original feature space. A scree plot was generated to visualize the cumulative explained variance by each principal component (PC). The plot showed diminishing returns after the first 10 components, with the first two PCs alone capturing nearly 40.7% of the variance.

- PC1 explains 23.84% of the variance
- PC2 explains 16.84% of the variance
- First 10 PCs explain approximately 83.47% of the variance
- All 16 components were needed to retain ≥95% of the variance (down from 21 original predictors)

Original shape (X_train_scaled): (1864, 21)

Reduced shape after PCA (train): (1864, 16)

Reduced shape after PCA (test): (527, 16)

PCA loadings showed that features like Population Density, Median Household Income, and Pct Public Transit Use contributed heavily to PC1, while variables such as Pct Single-Parent Households and Pct Homes with No Vehicle were strong contributors to PC2.

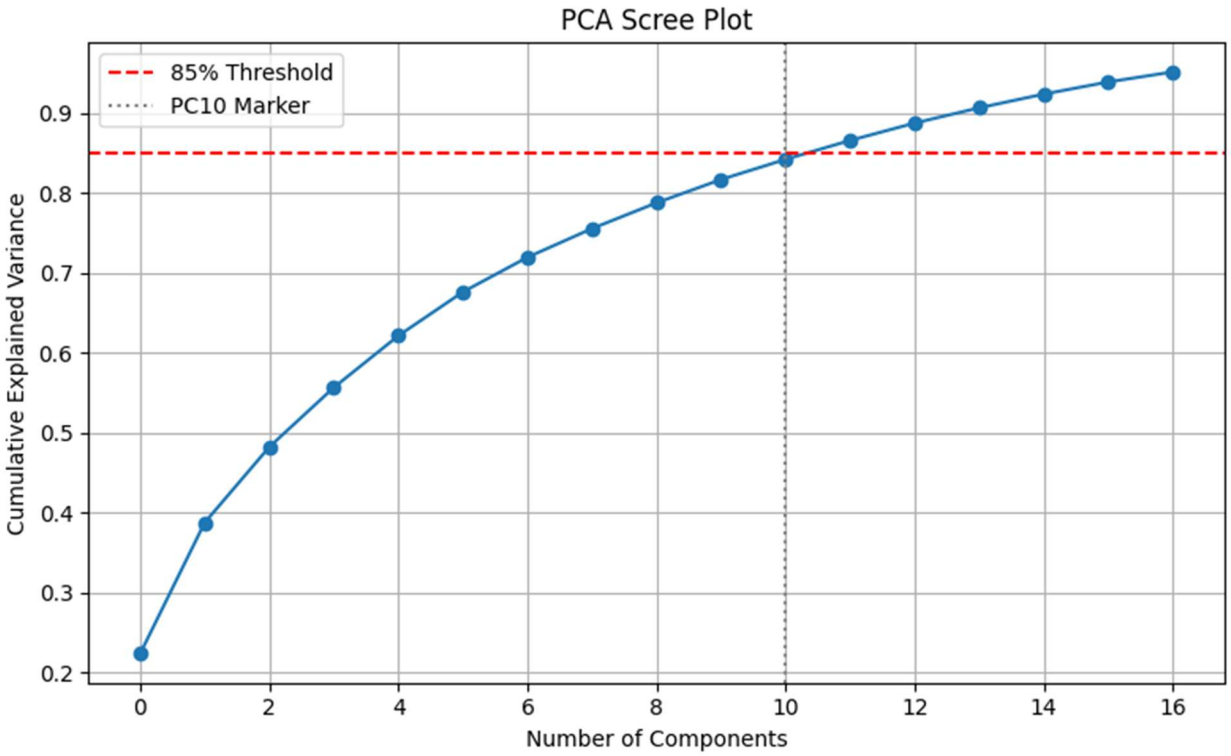

Figure 1: Cumulative Explained Variance by Principal Components

Box-Cox Transformation Analysis

To address the right-skewed distribution of the Bed_util_ratio target, a Box-Cox transformation was applied. A small positive shift was added to make all values strictly positive, satisfying Box-Cox assumptions.

- Lambda value: 0.7989
- Histogram of transformed target: Showed a visibly more symmetric distribution
- Shapiro-Wilk Test on transformed target:
    - $W = 0.9930$, $p < 0.0001$
    - Indicates the transformed distribution is still not perfectly normal, though improved

While the transformation helped reduce skewness, it did not fully normalize the distribution. The transformation nevertheless made the target more suitable for modeling, especially with algorithms that are sensitive to target shape (e.g., SVR).

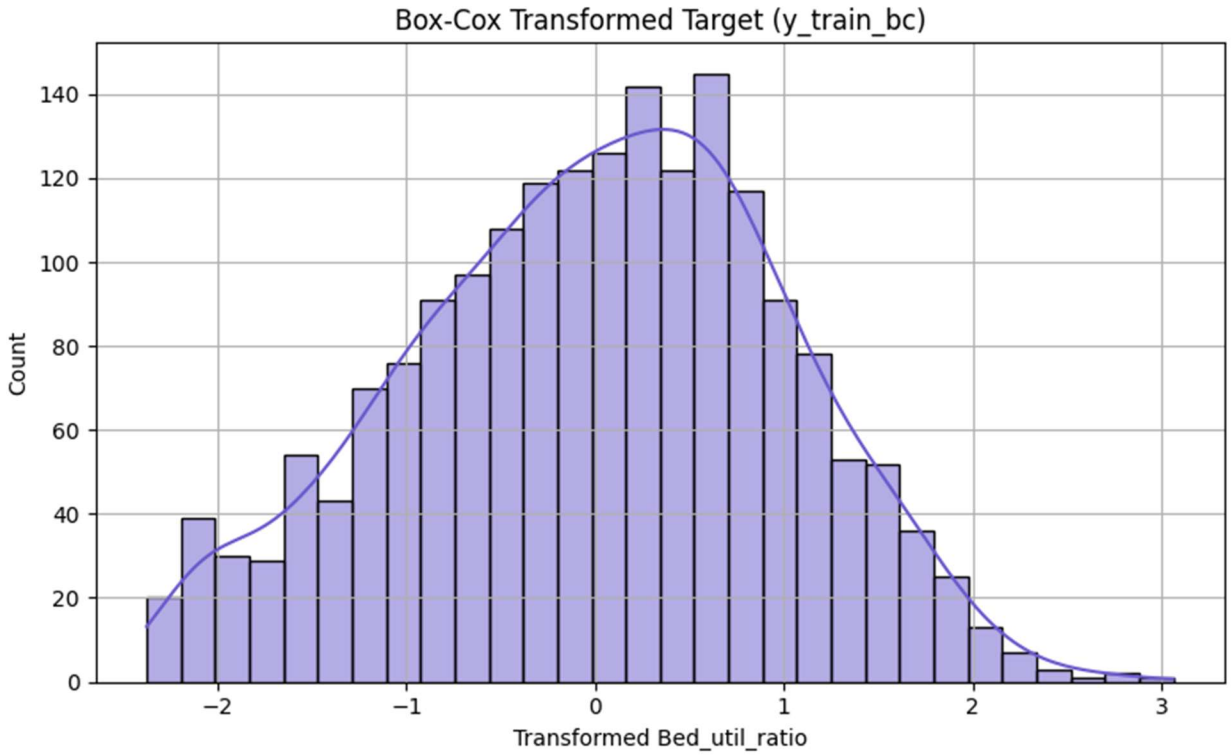

Figure 2: Box Cox Transformation for PCA showing improved distribution

Key Findings for PCA & Box-Cox Transformation

1. The Box-Cox transformation improved symmetry but did not produce a fully normal distribution, as confirmed by the Shapiro-Wilk test.
2. `Lambda ≈ 0.7989 suggests a meaningful transformation was applied.`

3. The PCA dimensionality reduction retained ≥95% of variance using 16 components.
4. The first 2 PCs explained over 40% of the variance, pointing to dominant structural dimensions (e.g., income, internet access, population density).
5. There is a diminishing return after the first 10 components, suggesting a tradeoff point for alternative analyses aiming to reduce model complexity.
6. Overall, this preprocessing pipeline successfully improved data quality and interpretability, providing a strong foundation for downstream regression and clustering analyses.

This preprocessing approach provides a good balance between dimensionality reduction and information preservation while attempting to normalize the target distribution.

PCA Loadings:

The PCA biplot illustrates how the original features contribute to the first two principal components, which together explain approximately 40.7% of the total variance (PC1 = 23.84%, PC2 = 16.84%).

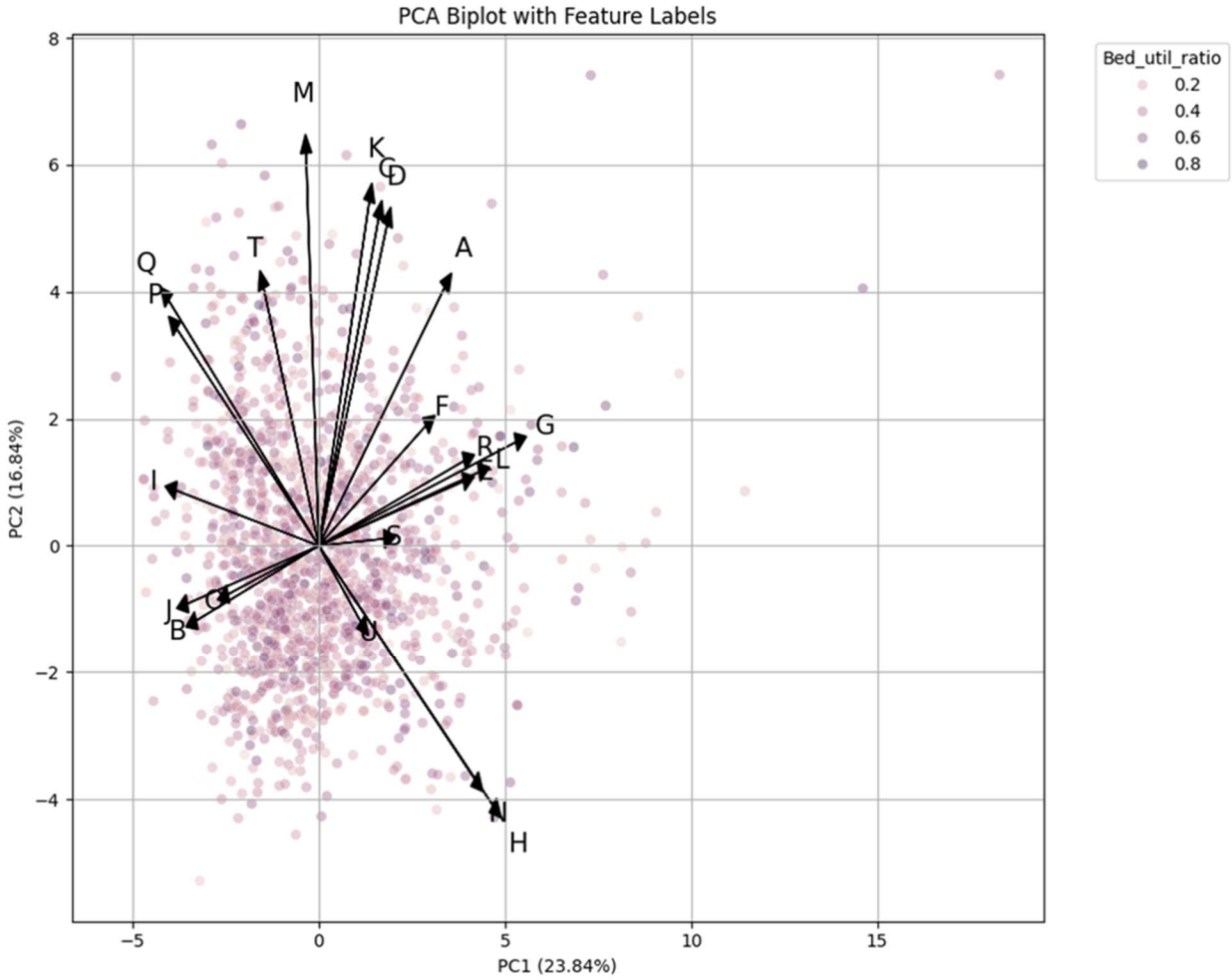

Figure 3: PCA Plot with Loadings Describing all Initial Features

PC1 (Horizontal Axis) primarily captures a socioeconomic and structural disadvantage dimension. Features with strong positive loadings on PC1 include:

- Pct_renter_occupied (A)
- Pct_renter_cost_50pct_plus (C) and 30pct_plus (D)
- Pct_single_parent (M)
- Pct_disabled (Q)

These suggest that PC1 may be interpreted as a latent factor capturing housing burden and household vulnerability.

In contrast, Median household income (H) and Pct_hh_no_internet (N) load in opposite directions, reinforcing that PC1 captures disparities in access and economic capacity.

PC2 (Vertical Axis) seems to reflect population structure and transportation access. Features with strong contributions here include:

- Pct_ag_65plus (K) and Pct_hh_65_alone (I) (indicating aging population)
- Pct_public_transit (L) and Distance_to_ED (O)

These patterns suggest that PC2 may relate to age-driven healthcare needs and transit accessibility.

Figure 4: Legend for PCA Feature Loadings

While the bed utilization ratio (Bed_util_ratio) gradient does not align strongly with any single component, it appears more elevated in the upper-right quadrant, where housing stress and older age groups co-occur. This supports the hypothesis that a combination of socioeconomic stressors and healthcare access barriers is linked to higher hospital bed usage.

Together, these components capture complex multidimensional variation across counties and offer an interpretable basis for clustering and prediction.

Using PCA to Identify Clusters (K-Means Clustering) and Silhouette Score:

We applied K-Means clustering to the PCA-reduced dataset, to explore whether meaningful groupings exist within counties based on their social and geographic characteristics,

We began by visualizing the total within-cluster sum of squared errors (SSE) using the elbow method, which suggested an inflection point at k = 3. This indicated that three clusters may be a reasonable choice for capturing underlying structure without overfitting.

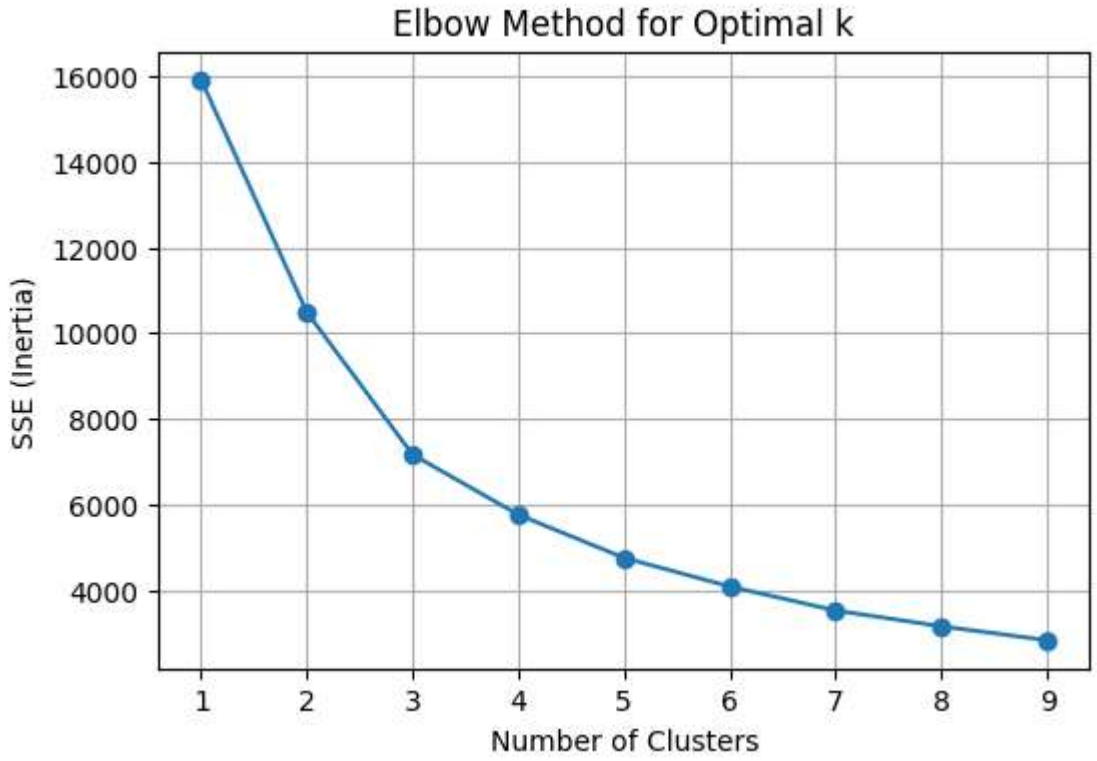

Figure 5: Identifying SSE with Elbow Method

We then applied K-Means clustering with k = 3 to the first two principal components, which together accounted for approximately 40.68% of the total variance (PC1: 23.84%, PC2: 16.84%). The resulting scatter plot of the clusters revealed fairly distinct groupings, although some overlap exists.

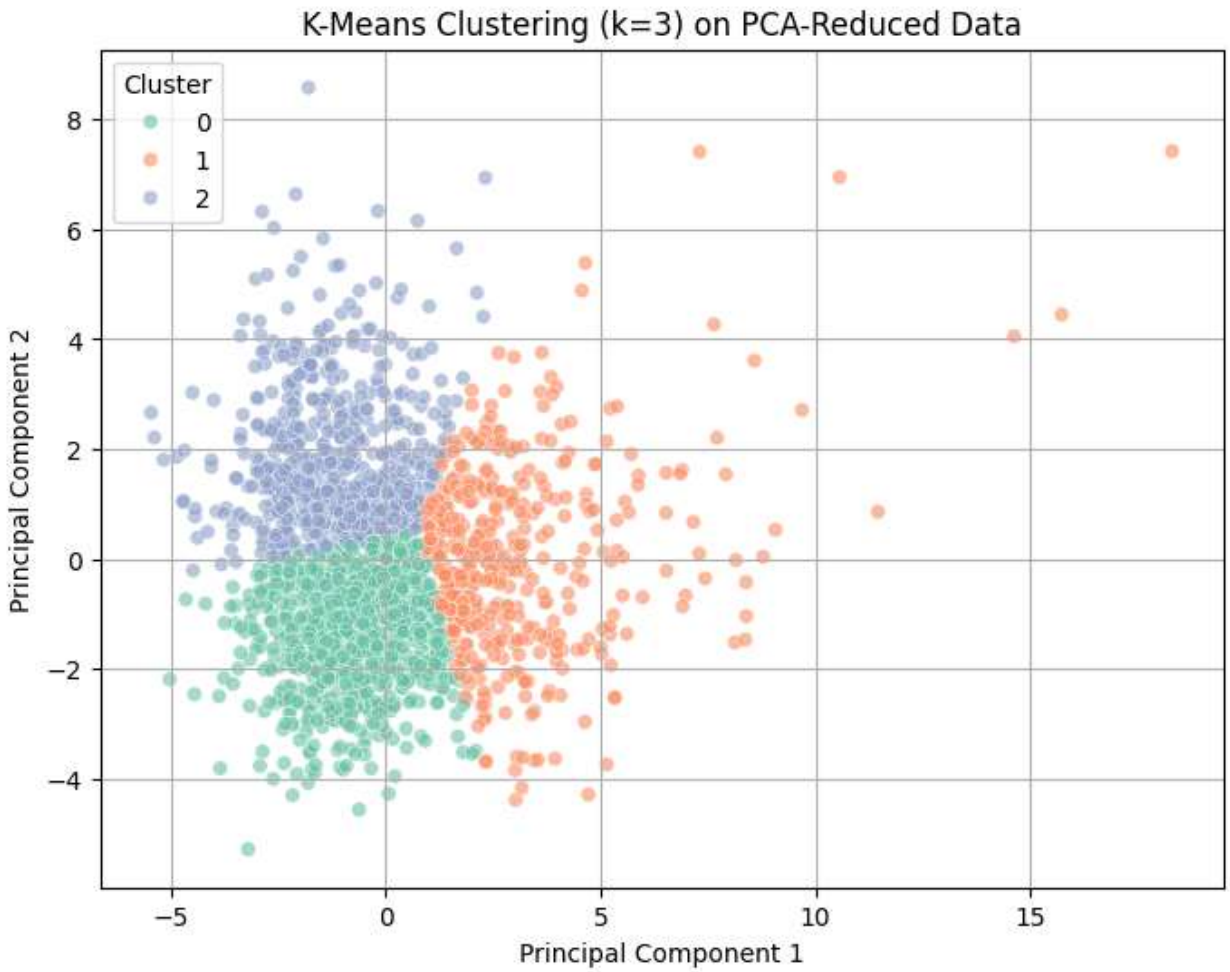

Figure 6: K-Means Clustering Analysis on K=3 for PCA

To evaluate the strength and separation of these clusters, we computed a **Silhouette Score** of 0.357. This score reflects moderate clustering structure indicating that while some separation exists, the clusters are not completely distinct. Values near 0.5 or higher typically suggest well-defined clusters, so this result points to some ambiguity in boundary placement and potential overlap in community characteristics.

These clusters offer a foundation for future profiling of communities by shared risk characteristics, potentially informing regionally tailored public health strategies. However, the modest silhouette score also cautions that more nuanced patterns may require additional features or alternative clustering strategies.

SVM Regression on Full Feature Set (No PCA)

For comparison against the SVM Regression with PCA, we trained a Support Vector Regression (SVR) model using the full scaled dataset without dimensionality reduction. This model leverages all 21 original features following standard scaling and preprocessing.

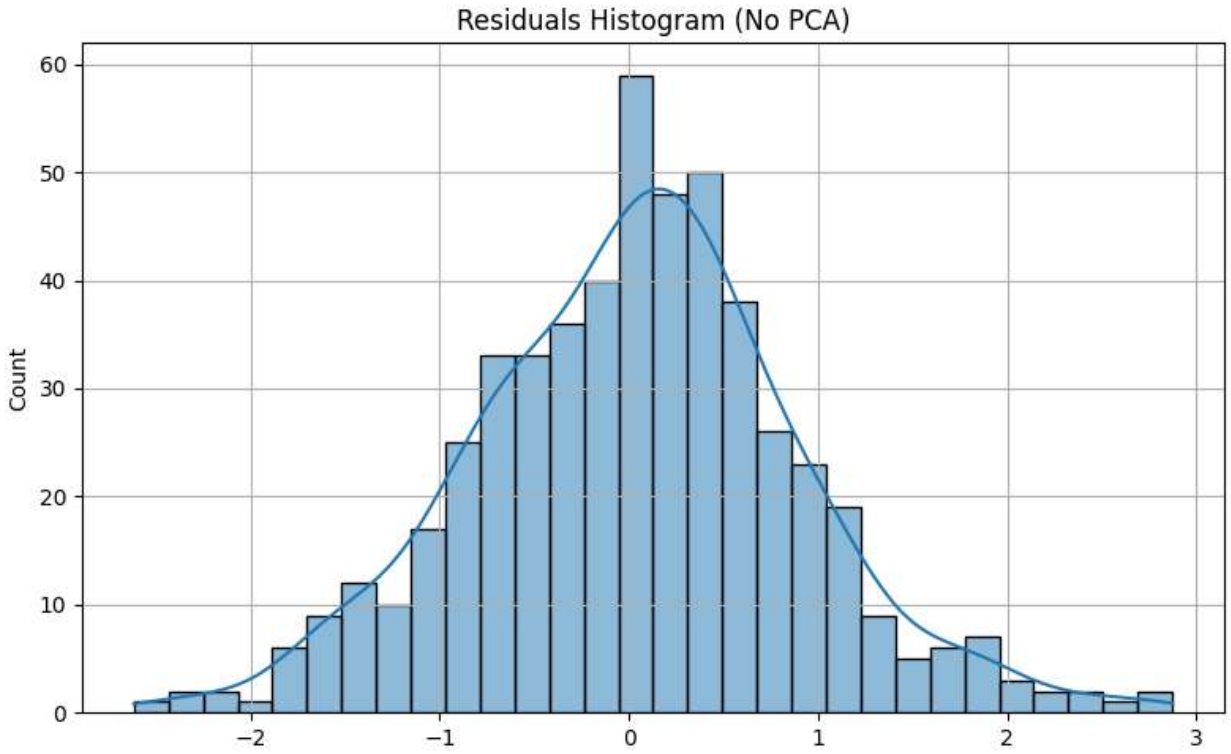

Figure 7: Histogram of SVM Regression Residuals(No PCA) for Full Feature Data

Model Performance:

- Cross-validated R^2 (train): 0.2929 ± 0.0301
- Test RMSE (Box-Cox scale): 0.8617
- Test MAE (Box-Cox scale): 0.6679
- Test R^2 (Box-Cox scale): 0.3249

After inverse-transforming predictions back to the original scale using the Box-Cox lambda, the Final RMSE (Original Scale) was 0.8075. This version dropped 68 rows due to NaN predictions after inverse transformation.

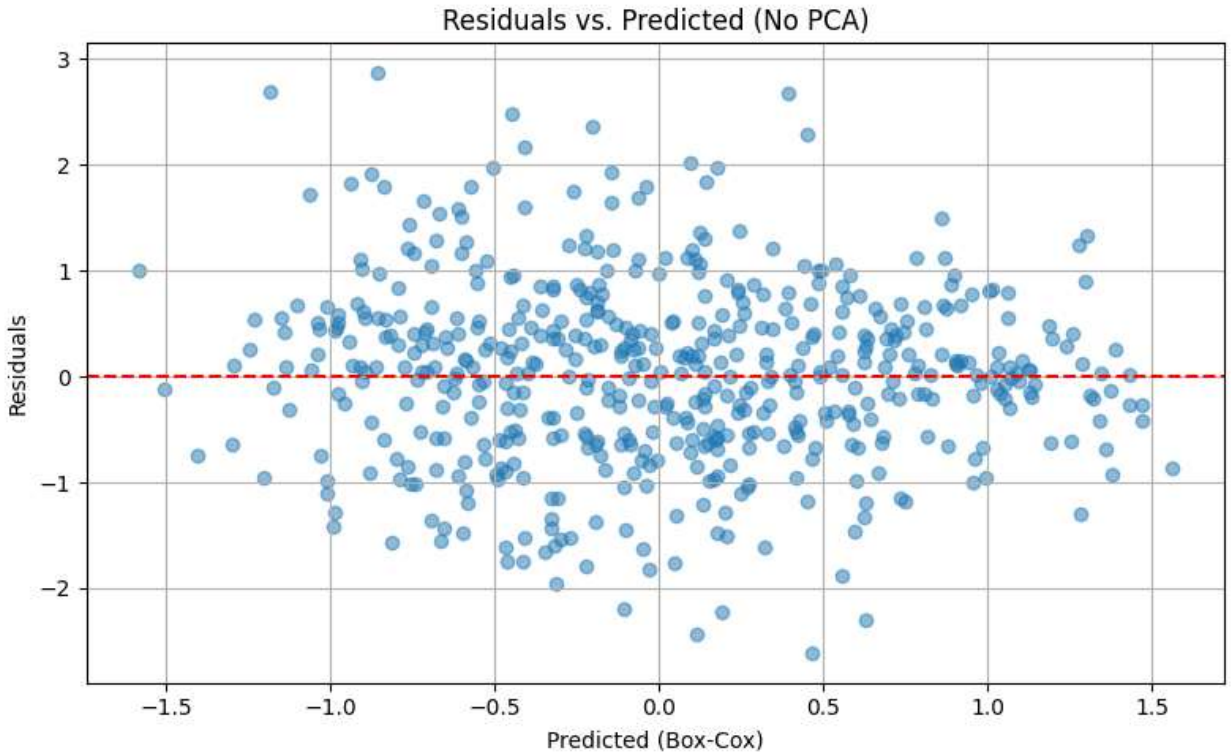

Figure 8: Residuals vs. Predicted Values (No PCA)

Residual Diagnostics:

- Shapiro–Wilk W = 0.9949, p = 0.0811 → Residuals are likely normal.
- The residuals histogram shows an approximately normal distribution.
- The residuals vs. predicted plot shows no strong pattern, though some mild heteroscedasticity may be present.
- The predicted vs. actual plot shows a modest linear trend, suggesting the model was able to capture part of the structure but not all variance.

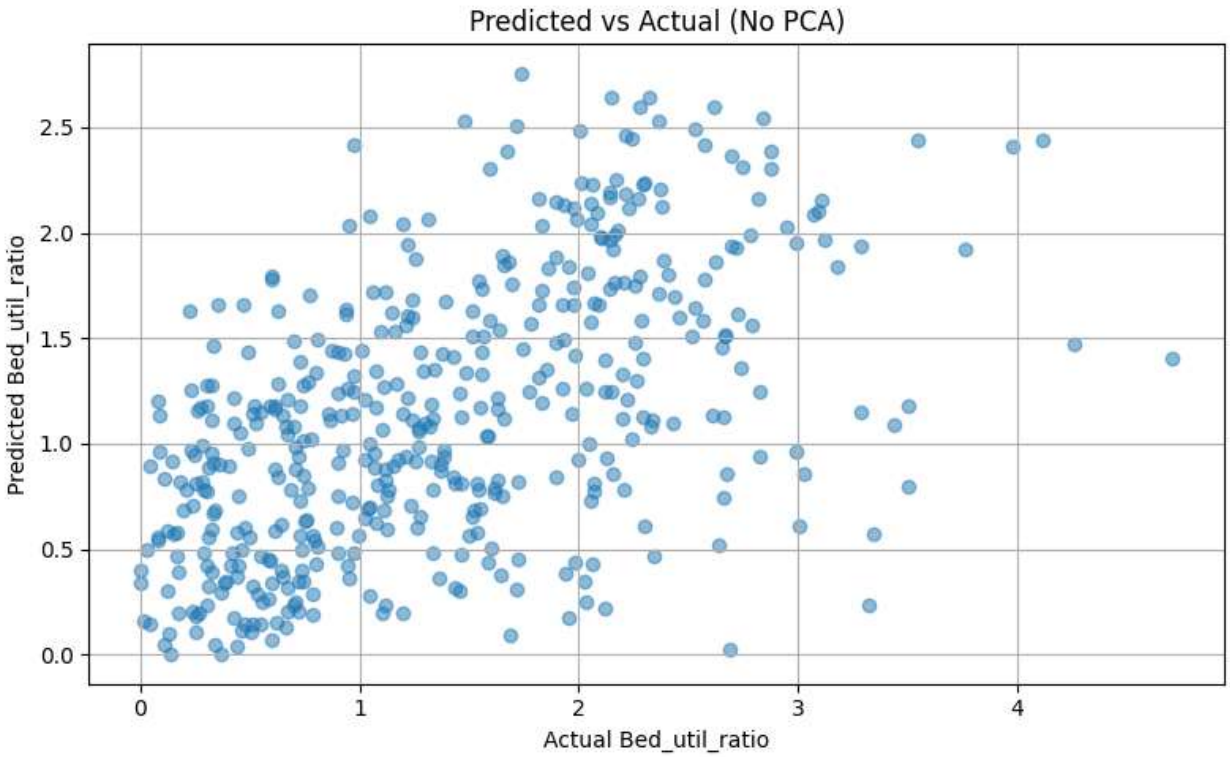

Figure 9: Predicted vs. Actual Bed Utilization Ratio (No PCA)

While this model produced results similar to the PCA-based SVR model, it achieved slightly better R^2 and RMSE values on the test set. The relatively normal distribution of residuals and absence of major structure in residual plots indicate a reasonable model fit. However, the R^2 score suggests the model only modestly explains variance in the target, reinforcing the challenge of predicting hospital bed utilization with the current features.

SVM Regression on PCA Reduced Feature Set:

To evaluate whether dimensionality reduction could improve model performance or stability, we trained a Support Vector Regression (SVR) model on features reduced via Principal Component Analysis (PCA). PCA compressed the original 21 variables into 16 orthogonal components, preserving over 95% of the total variance while addressing multicollinearity and simplifying the model structure.

Before modeling, we applied a Box-Cox transformation to the target variable (Bed_util_ratio) to mitigate its right skew and better satisfy modeling assumptions. While the transformation improved the symmetry of the distribution, the Shapiro-Wilk test still rejected normality ($W = 0.9933$, $p = 0.0195$), suggesting that the target was only partially normalized.

The SVR model trained on the PCA-reduced feature set indicates the following performance on the Box-Cox scale:

- Cross-validated R^2 (train): 0.2989 ± 0.0262
- Test RMSE: 0.8684
- Test MAE: 0.6691
- Test R^2: 0.3143

These results indicate moderate predictive power, capturing about 31% of the variance in the target variable. The low standard deviation in cross-validation scores suggests the model generalizes reasonably well without high sensitivity to random variation.

The histogram of residuals showed a roughly bell-shaped distribution, although slightly skewed.

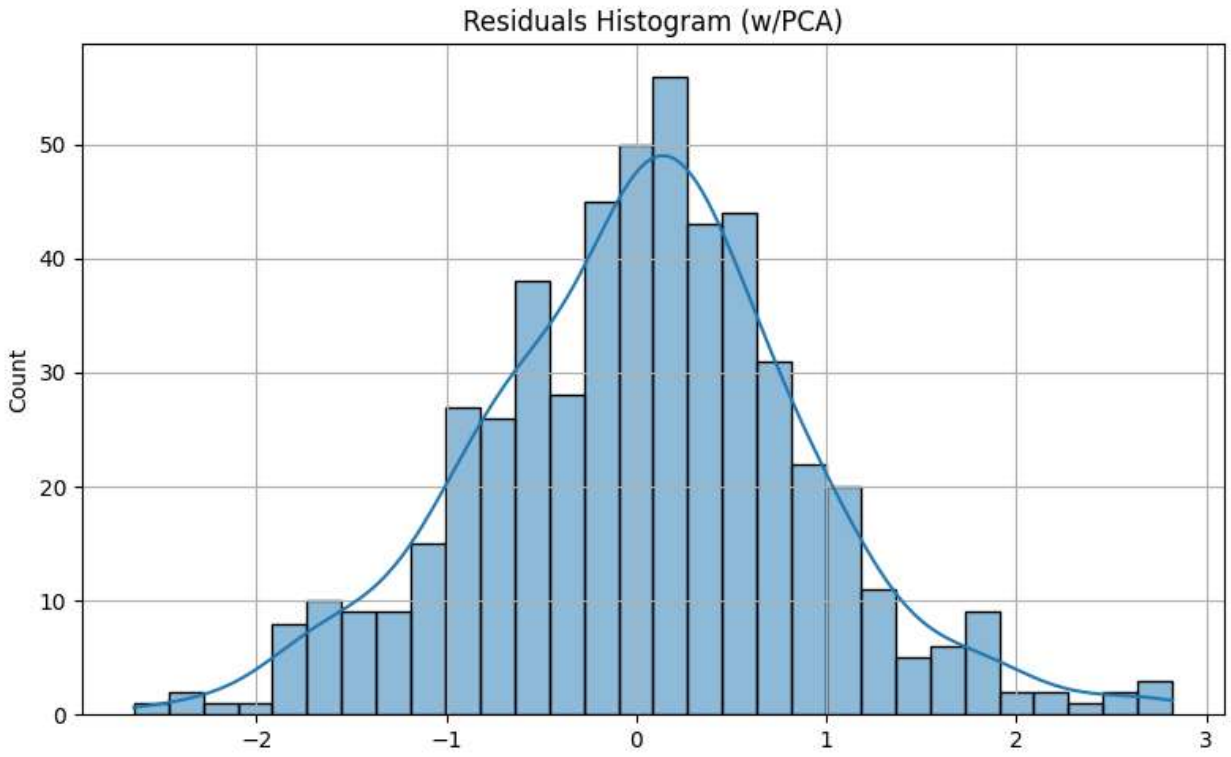

Figure 10: Histogram of SVM Regression Residuals(w/ PCA) for Reduced Feature Data

Residual diagnostics further supported model consistency. The residuals vs. predicted plot did not reveal strong systematic bias, though mild funneling at the extremes indicated slight heteroscedasticity. These patterns suggest residuals are approximately but not perfectly normally distributed.

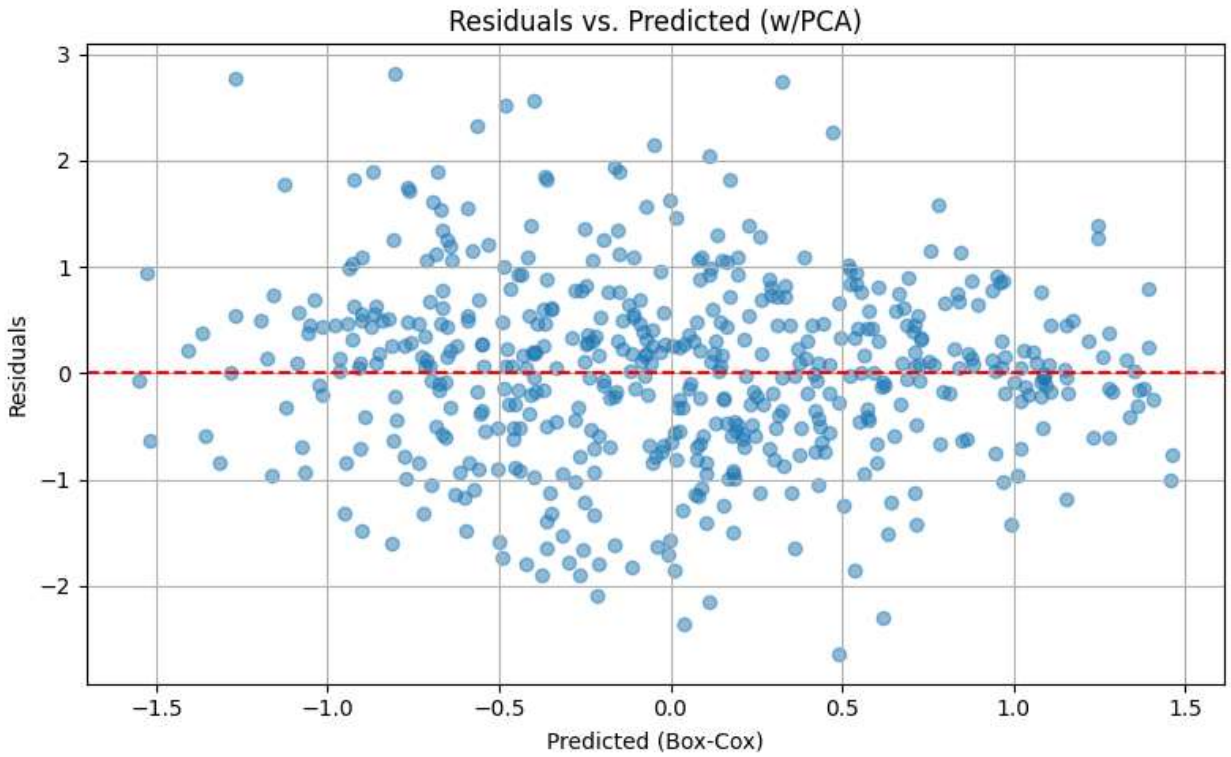

Figure 11: Residuals vs. Predicted Values (w/ PCA)

After applying the inverse Box-Cox transformation to return predictions to the original target scale:

- Final RMSE: 0.8051
- 71 predictions dropped due to NaNs

The predicted vs. actual plot (in original scale) demonstrated moderate alignment, with scatter visible especially at the lower and upper ends of the utilization spectrum. This confirms that the model's general structure is sound but that finer-grained prediction accuracy is still limited.

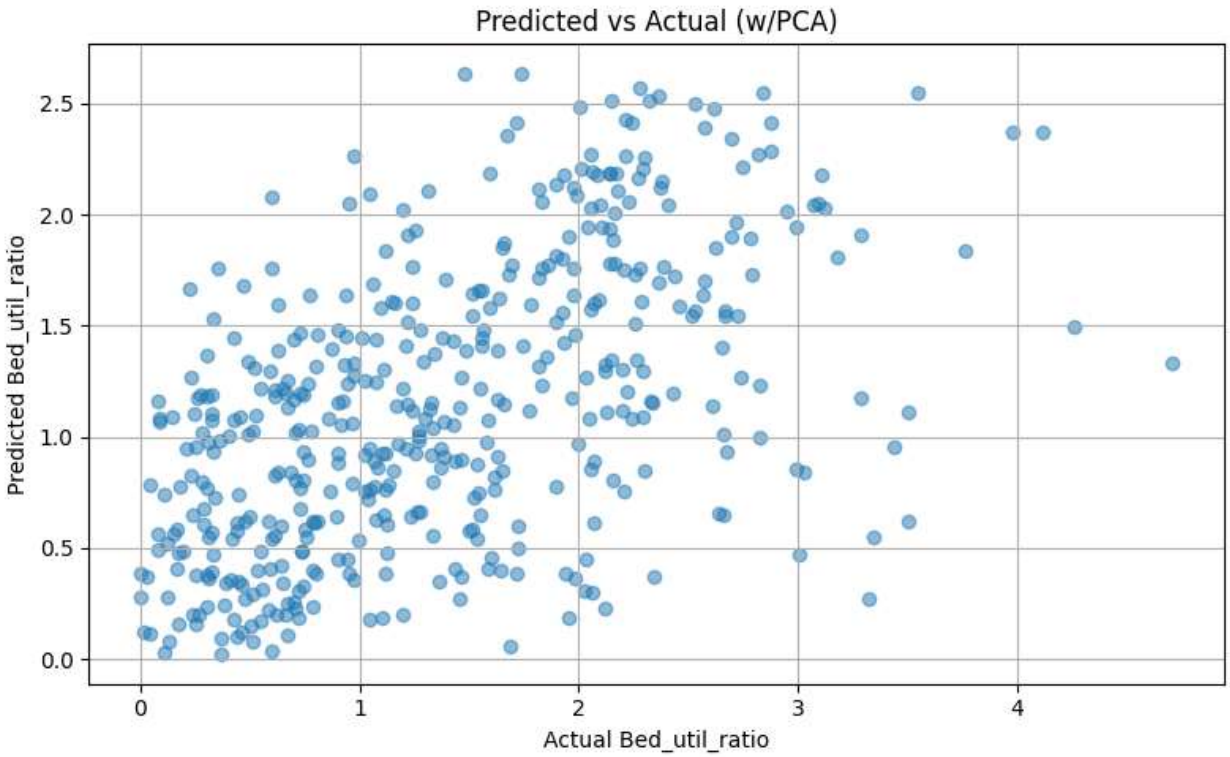

Figure 12: Predicted vs. Actual Bed Utilization Ratio (w/ PCA)

Finally, the SVM model on PCA-reduced data performed similarly to the model trained on the full feature set, offering comparable R^2 and slightly lower error. This reinforces that the predictive signal is preserved within the PCA-transformed space. While overall model performance remains modest, dimensionality reduction supports a more stable and interpretable modeling pipeline—particularly useful when pairing with clustering or subgroup analysis.

Conclusion:

This project examined whether social and geographic determinants of health could predict non-COVID hospital bed utilization across U.S. counties in 2020. After extensive data cleaning, transformation, and dimensionality reduction via PCA, we built multiple Support Vector Regression models both with and without dimensionality reduction. While both versions of the SVM model showed modest predictive power (R^2 ~0.31), results suggest that the relationship

between community-level SDOH features and non-Covid hospital bed utilization is complex and potentially influenced by unmeasured factors.

Clustering analysis revealed patterns in community similarity, but with only moderate cohesion (Silhouette Score = 0.357), indicating that the influences on utilization are broadly distributed rather than localized. Preprocessing improvements such as removal of redundant fields like state and land area helped boost model performance and reduce multicollinearity.

Future work could explore ensemble methods, feature interactions, or temporal modeling. Overall, while the results provide some predictive value, they also underscore the limitations of using SDOH alone to forecast real-time healthcare system strain.

Appendix A - Data Dictionary

Our data dictionary can be found at this github location:

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

Appendix B - Github Repository

Our Github repository can be found at this location:

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

Appendix C - Github Repository - M05 Initial Model Report

Our GitHub repository contains all code, data preprocessing steps, and modeling scripts used in the M05 Initial Modeling phase of the project. This includes implementation of PCA, K-Means clustering, Box-Cox transformation, SVR modeling (with and without PCA), residual diagnostics, and supporting visualizations. The repository serves as a central resource for reproducibility and transparency of our analysis.

All materials can be found at this location:

https://github.com/lemieuxjm-cap/Iota-Capstone/blob/main/code/M05_Initial_Modeling.ipynb