Preprocessing and Feature Engineering Report

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

Team IOTA

Team Lead: Rea Kelolli

Recorder: Mehmet Comert

Spokesperson: Melody Rios

Floater: June Lemieux

Recap Background & Question

Research Question:
Can social and geographic determinants of health (e.g., disability, insurance, poverty, healthcare 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 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.

Methods

The initial preprocessing strategy included cleaning procedures for both the SDOH dataset and the COVID Hospitalization dataset. The exploratory data analysis (EDA) revealed rows containing five null values in selected variables which we removed during our analysis. We chose to impute the “Bed_util_ratio” column because it contained 818 null values and determined that retaining this data would be more beneficial than dropping the corresponding rows. We implemented MICE and Random Forest imputation methods among others before selecting K-Nearest Neighbors (KNN) imputation as our final approach. The method enabled us to predict values from tracts with comparable healthcare utilization characteristics thus providing a suitable balance between precision and speed.

Preprocessing Methods & Justification:

We began preprocessing by merging and aligning the AHRQ SDOH and HealthData.gov datasets using the CountyFIPS identifier to ensure consistent geographic referencing across sources.

To address missing values, we applied variable-specific strategies based on data type and importance. For the "Region" variable, we removed 96 rows with missing values due to their small proportion and then applied one-hot encoding, as the variable consisted of only four distinct categories. For "State", we used target encoding, assigning a numeric value based on each state's frequency in the dataset to preserve geographic weight without introducing high cardinality. For "Is_Micro_Metro", we applied binary encoding, reflecting its true/false structure in a format compatible with modeling.

We removed rows containing exactly five missing values across critical SDOH features to achieve a balance between data retention and quality control.

The preprocessing choices were made to maximize the geographic and demographic coverage of our dataset while keeping the feature distributions clean, interpretable and model ready.

Feature Engineering Methods & Justification:

To quantify hospitalization risk at the county level, we engineered our target variable, Bed_Util_Ratio, using operational hospital capacity data from the HHS COVID-19 Impact dataset. The ratio was calculated using the following formula:

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

This metric estimates non-COVID inpatient bed utilization, providing a more stable indicator of baseline healthcare system strain, independent of pandemic-specific surges. We focused on non-COVID usage to capture structural disparities in hospital access and demand driven by chronic health conditions and population-level vulnerabilities.

This engineered target serves as a continuous variable suitable for regression modeling and allows us to explore how social determinants influence underlying hospitalization patterns.

Results

In the preprocessing stage, we identified that the dataset contained both a County and CountyFIPS column. These two variables effectively conveyed the same geographic information- however, CountyFIPS is a unique identifier, noted by a numerical number, while County is a text field that may not be unique as multiple states can have the same County name. Since County lacks uniqueness and does not provide additional value beyond what is captured in CountyFIPS, we dropped the County column to reduce redundancy.

Our first preprocessing step for the data was to perform a train-test split. Instead of picking an arbitrary split, we used the amount of parameters as part of our calculation to determine the split. For our data, a train-test split of 79% training and 21% testing ratio was the given data and features.

During the data extraction phase, we observed that the target variable, Bed_util_raito, contained missing values. Although imputation is a common technique for handling missing data, it is not appropriate to impute the target variable. Imputing these missing outcomes introduces artificially

generated values that do not reflect actual observations, thereby risking bias in the model's training process and compromising the validity of performance evaluation.

In addition to encoding, we addressed missing values in the Is_Metro_Micro column. Since the feature contained missing values, we employed KNN imputation to estimate them. This approach is particularly effective for categorical and binary variables, as it represents existing patterns in the data. Previously we encoded the True and False values to be either 1 or 0. Our calculator approximated the binary classification by assigning a value greater than or equal to 0.5 to 1 and a value of lower than 0.5 to 0.

While we did not impute the missing values for our target variable, Bed_util_ratio, we did observe the distribution of the numbers using various methods. To better understand the distribution of our target variable, we conducted both visual and statistical analyses. Based on the histogram in Figure 1, it is evident that both the training and test sets are skewed. Furthermore, after performing the Shapiro-Wilk test, we observed a Shapiro-Wilk statistic of 0.9919 and a p-value of 2.82e-10. Since the p-value is significantly less than .05, we reject the null hypothesis and conclude that Bed_util_ratio is not normally distributed.

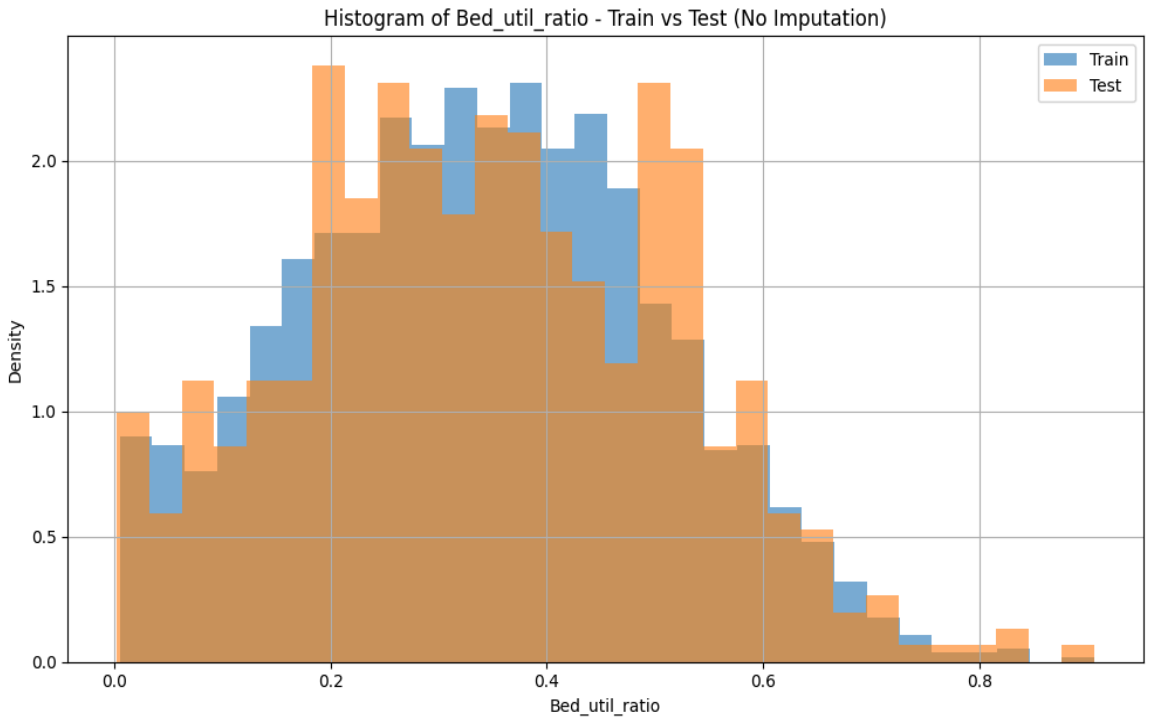

Figure 1 Histogram of BED_UTIL_RATIO – Train vs Test(no Imputation) indicating skewed distribution

Next, we assessed the distribution and normality of the target variable. We visualized its shape using both a histogram with a KDE overlay in Figure 2. The histogram with KDE reveals a unimodal, right-skewed distribution, where most values cluster between 0.3 and 0.4, and a long tail extends toward higher values above 0.6.

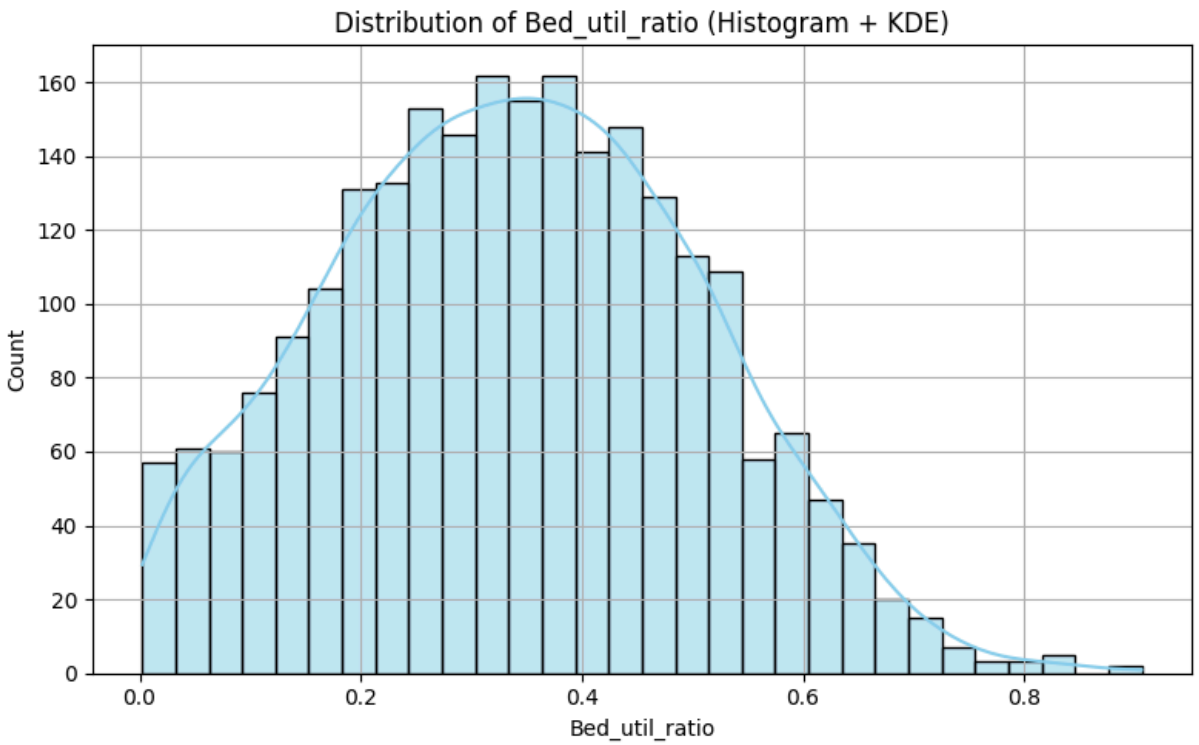

Figure 2 Distribution of Bed_util_ratio(Histogram +KDE) indicating skewed distribution

The Q-Q plot, chart 3, of the Bed_util_ratio provides additional insights into its distributional characteristics. While the central proportion of the data closely follows the expected normal line, deviations become more apparent at the tails. Specifically, the points curve away from the diagonal line on both ends, indicating mild skewness and heavier tails than would be expected under normal distribution. This suggested that while the variable approximate normality in the center, it departs from normality in the extremes, reinforcing our findings from the histograms and statistical testing.

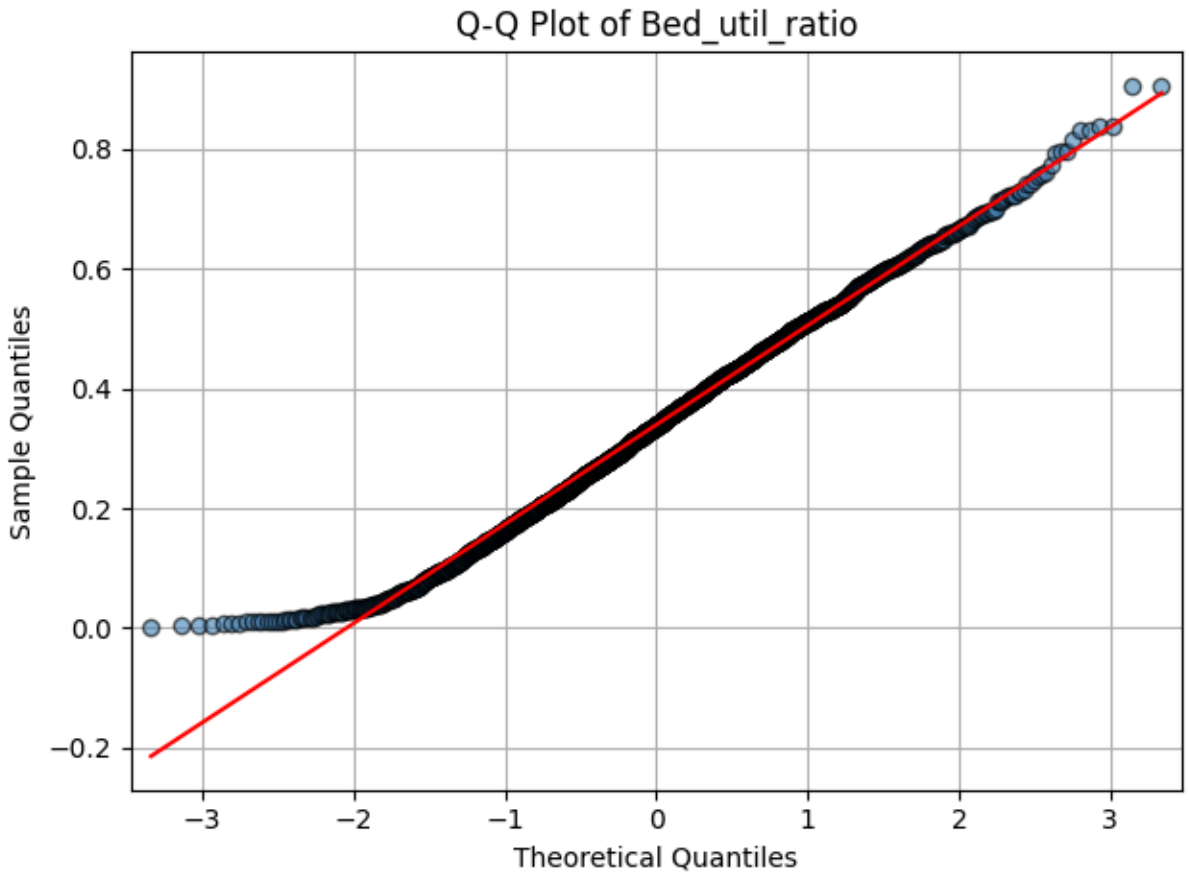

Figure 3 Q-Q Plot of Bed_util_ratio indicating skewed distribution

Based on the results of the initial Shapiro-Wilk test, histogram and Q-Q plot, a Box-Cox transformation was applied to address right-skewness in the target (Bed_util_ratio) in an attempt to bring the distribution to a more normal state. We did have values that were zero, so a small offset of 1e-4 was applied to make this transformation method work. Figure 4 is a histogram of the transformed data.

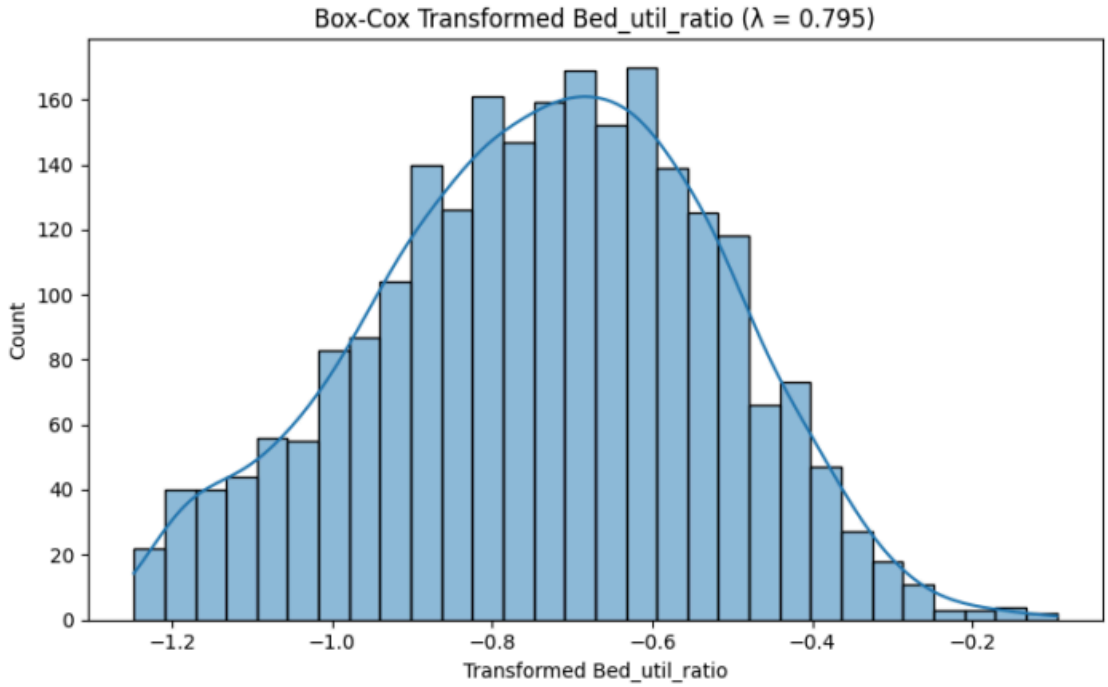

Figure 4 Post-transformation distribution indicating improved distribution

After the transformation, the distribution appeared more normal. Using the transformed data, we developed preliminary Lasso and Ridge regression models, created histograms, performed the Shapiro-Wilk test again, and also performed the Breusch-Pagan test for homoscedasticity. The histograms (see Figure 5) appeared fairly normal. However, Shapiro-Wilk testing did not support the hypothesis that the distribution was normal as a result of either regression model. Additionally, the Breusch-Pagan test results indicated significant heteroscedasticity in both models ($p < 0.001$), meaning the variance of residuals was not constant across predicted values.

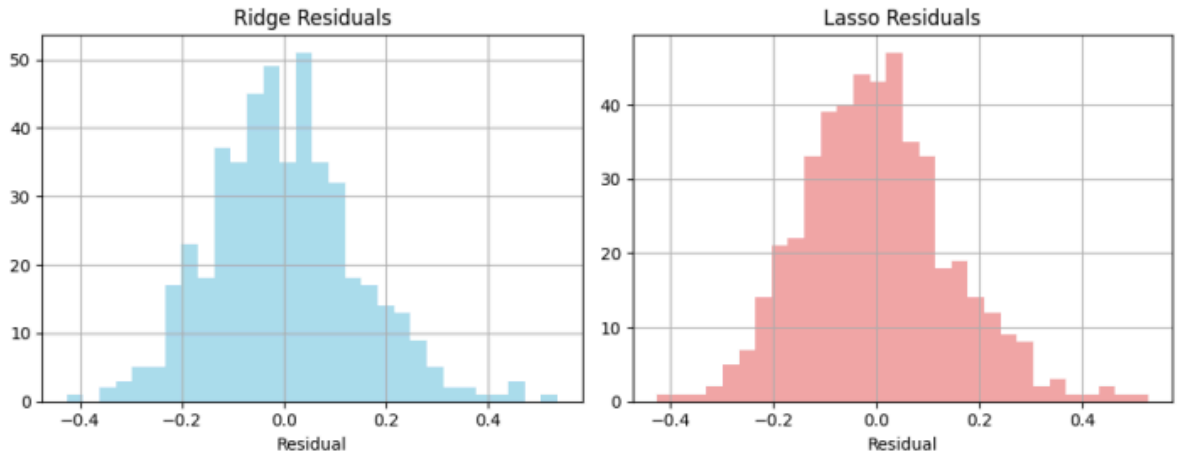

Figure 5 Histograms of post-transformed Ridge and Lasso regression results

Based on the results of the post-transformation diagnostic testing, we have determined that there are violations of linear regression assumptions which compromise the interpretability and inferential reliability of the models, particularly for coefficient-based policy conclusions.

Discussion and Next Steps

Our analysis aims to answer whether social and geographic determinants of health can predict non-COVID hospital bed utilization at the county level in 2020. Initially, we hypothesized that counties with higher proportions of older adults, uninsured populations, individuals with disabilities, and limited healthcare access would experience higher non-COVID hospital bed utilization. During preprocessing and exploratory analysis, the complexity of this relationship was supported and it also highlighted the substantial skewness in the target variable, Bed_util_ratio. Although we tried normalizing the target using Box-Cox transformation and fit regularized linear regression models (Lasso and Ridge), our diagnostic tests–such as the Shapiro-Wilk test and Breusch-Pagan test–revealed key assumptions violations. It also included non-normality and heteroscedasticity, where these findings prompted us to reevaluate our

modeling strategy, since linear models may not offer reliable inference or interpretable coefficients under these conditions.

Because regression model results may not be reliable, we are pivoting from regularized regression toward an unsupervised modeling strategy that better handles variance instability and potential non-linear patterns. Principal Component Analysis (PCA) will be used to reduce multicollinearity and uncover latent factor structures across socioeconomic and geographic indicators. These principal components will then be used to construct cluster profiles, offering actionable insights into regional hospitalization risk patterns for stakeholders. This shift prioritizes robustness, interpretability, and practical alignment with real-world policy applications.

Conclusion

Our findings emphasize the complexity of modeling hospital bed utilization using social determinants of health. Although our initial regression-based approach revealed valuable insights, it also highlighted key limitations due to skewed distributions and violations of model assumptions. Shifting to an unsupervised model, we aim to uncover meaningful patterns that align more closely with real-world needs. This transition not only strengthens the robustness of our analysis but also enhances our ability to generate actionable insights for public health decision-makers.

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 - Table of Figures