Exploratory Data Analysis Report

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

Team IOTA

Team Lead: June Lemieux

Recorder: Rea Kelolli

Spokesperson: Mehmet Comert

Floater: Melody Rios

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.

1 Background and Research Question

Monitoring hospitalization strain is a critical goal for public health agencies, particularly in under-resourced or delayed-reporting regions. Our project uses social and structural determinants of health as a means of identifying counties experiencing elevated non-COVID inpatient hospital bed utilization—a proxy for healthcare system strain.

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 and Prediction: We hypothesize that counties with higher disability rates, older populations, limited insurance coverage, and poor access to care will show higher hospital bed utilization, even after adjusting for geography and environmental exposures. Prior work has shown that social determinants significantly shape health system demand (Artiga & Hinton, 2018; Magnan, 2017). Furthermore, analysis of COVID-19 hospitalization patterns has demonstrated how demographic and health characteristics correlate with system utilization, suggesting their potential value as predictive indicators (Garg et al., 2020).

Prediction: Counties with greater social vulnerability—such as those with lower income, higher disability, and fewer insured individuals—will exhibit higher average values of BED_UTIL_RATIO, even when controlling for geographic factors like metro vs. rural classification or ICU proximity.

We initially pursued a tract-level, engineered score. However, given the improved completeness, policy relevance, and continuous measurement of BED_UTIL_RATIO at the county level COVID hospitalization data (HHS's 2020 database) led us to revise the approach toward a predictive monitoring framework.

2 Data & Methods

2.1 Data Sources and Acquisition

The dataset merges two nationally recognized public sources. The first is the 2020 AHRQ SDOH database, which tracks access, income, insurance, and housing data at the county level. Its standardization and public availability make it ideal for surveillance and reproducible policy analysis (Magnan, 2017). The second dataset is the COVID-19 Healthcare Utilization file, which provides weekly averages for inpatient capacity. We derived the target variable using the formula:

BED_UTIL_RATIO = (Inpatient Beds Used - COVID Beds Used) / (Total Inpatient Beds - COVID Dedicated Beds).

This formula estimates non-COVID bed utilization and results in 3,239 county-level observations and 43 predictors after preprocessing.

2.2 Feature Selection

To reduce dimensionality and emphasize relevance, we selected the 20 features most correlated with BED_UTIL_RATIO using Pearson correlation and then added COUNTYFIPS, STATE, and REGION as identifiers. The final working dataset contains 23 features.

Top features include:

- IS_METRO_MICRO (r = 0.40)
- ACS_PCT_RENTER_HU_avg (r = 0.29)
- POS_DIST_MEDSURG_ICU_TRACT_avg (r = 0.28)
- ACS_PCT_RENTER_HU_COST_50PCT_avg (r = 0.25)

- ACS_TOT_POP_POV_sum (r = 0.24)

The table below displays the absolute Pearson correlation coefficients between various features and the target variable vs. BED_UTIL_RATIO.

Feature Absolute Correlation	**Coeff**
BED_UTIL_RATIO	1
IS_METRO_MICRO	0.404766
ACS_PCT_RENTER_HU_avg	0.286737
POS_DIST_MEDSURG_ICU_TRACT_avg	0.278767
ACS_PCT_RENTER_HU_COST_50PCT_avg	0.248769
ACS_PCT_RENTER_HU_COST_30PCT_avg	0.246322
ACS_TOT_POP_POV_sum	0.243171
ACS_PCT_OWNER_HU_COST_30PCT_avg	0.231916
CEN_POPDENSITY_TRACT_avg	0.224611
CEN_AREALAND_SQM_TRACT_avg	0.221157
ACS_MEDIAN_HH_INC_sum	0.197687
ACS_PCT_HH_ALONE_ABOVE65_avg	0.195666
ACS_PCT_AGE_ABOVE65_avg	0.188541
ACS_PCT_HU_NO_VEH_avg	0.174349
ACS_PCT_PUBL_TRANSIT_avg	0.167792
ACS_PCT_CHILD_1FAM_avg	0.161295
ACS_PCT_HH_INTERNET_avg	0.154939
POS_DIST_ED_TRACT_avg	0.154278
ACS_PCT_HU_MOBILE_HOME_avg	0.146036

ACS_PCT_DISABLE_avg	0.138186

Table 1_2.2 Feature Selection - Absolute Correlation with BED_UTIL_RATIO.

2.3 Missingness and Data Structure After Feature Selection

The dataset includes 3,239 county-level observations and originally contained 45 columns. After filtering for the top 20 correlated predictors and geographic identifiers, the working dataset retains 23 features. Among these:

- 20 are numeric and used for correlation/VIF/EDA
- 3 are identifiers (COUNTYFIPS, STATE, REGION)

The target variable BED_UTIL_RATIO has 2,421 non-null entries, meaning that data is missing for approximately 25% of counties. This is primarily because some counties in the AHRQ SDOH dataset did not have a matching COUNTYFIPS in the HHS COVID dataset used to derive bed utilization.

All selected predictor variables have minimal missingness, with fewer than 30 missing entries each. No features exceeded the standard 30% missingness threshold, and thus no variables were dropped based on missingness.

Among categorical features:

- STATE is fully complete.
- REGION has 96 missing values, but is retained for its stratification value and will be imputed with mode during modeling.
- The BED_UTIL_RATIO column was retained in both the target and correlation steps for transparency and traceability.

The following output provides a concise summary of the Data Frame’s structure and contents:

Range Index: 3239 entries, 0 to 3238				
Data columns (total 23 columns):				
	Column	Non-Null Count		Dtype
0	COUNTYFIPS	3239	non-null	int64
1	STATE	3239	non-null	object
2	REGION	3143	non-null	object
3	BED_UTIL_RATIO	2421	non-null	float64
4	IS_METRO_MICRO	2421	non-null	float64
5	ACS_PCT_RENTER_HU_avg	3234	non-null	float64
6	POS_DIST_MEDSURG_ICU_TRACT_avg	3234	non-null	float64
7	ACS_PCT_RENTER_HU_COST_50PCT_avg	3234	non-null	float64
8	ACS_PCT_RENTER_HU_COST_30PCT_avg	3234	non-null	float64
9	ACS_TOT_POP_POV_sum	3234	non-null	float64
10	ACS_PCT_OWNER_HU_COST_30PCT_avg	3234	non-null	float64
11	CEN_POPDENSITY_TRACT_avg	3234	non-null	float64
12	CEN_AREALAND_SQM_TRACT_avg	3234	non-null	float64
13	ACS_MEDIAN_HH_INC_sum	3234	non-null	float64
14	ACS_PCT_HH_ALONE_ABOVE65_avg	3234	non-null	float64

15	ACS_PCT_AGE_ABOVE65_avg	3234	non-null	float64
16	ACS_PCT_HU_NO_VEH_avg	3234	non-null	float64
17	ACS_PCT_PUBL_TRANSIT_avg	3234	non-null	float64
18	ACS_PCT_CHILD_1FAM_avg	3234	non-null	float64
19	ACS_PCT_HH_INTERNET_avg	3234	non-null	float64
20	POS_DIST_ED_TRACT_avg	3234	non-null	float64
21	ACS_PCT_HU_MOBILE_HOME_avg	3234	non-null	float64
22	ACS_PCT_DISABLE_avg	3234	non-null	float64
dtypes: float64(20), int64(1), object (2)				

Table 2_2.3_Missingness and Data Structure After Feature Selection - Data Frame Information.

3 Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics (Main Summary)

The selected dataset includes 3,239 county-level observations with 23 variables: 20 numeric and 3 categorical. Of the numeric variables, BED_UTIL_RATIO is the target, with 2,421 valid values. Descriptive statistics for the numeric variables show meaningful variability in healthcare access, housing burden, population characteristics, and infrastructure.

For example:

- BED_UTIL_RATIO has a mean and median of approximately 0.34, with values ranging from 0.0 to 0.9, reflecting considerable diversity in bed utilization across counties.

- ACS_PCT_RENTER_HU_avg averages 26.8%, with values extending as high as 82.0%, indicating variation in housing stability.
- POS_DIST_MEDSURG_ICU_TRACT_avg spans from 0.0 to 487.0 miles, with a right-skewed distribution, suggesting some rural counties are geographically isolated from intensive care facilities.

The following table presents descriptive statistics for the numerical features in our dataset. This provides a summary of the central tendency, dispersion, and range of each variable:

Features	count	mean	std	min	25%	50%	75%	max
COUNTYFIPS	3239	31504.7	16463.0	1001.0	19032.0	30033.0	46126.0	78030.0
BED_UTIL_RATIO	2421	0.3	0.2	0.0	0.2	0.3	0.5	0.9
IS_METRO_MICRO	2421	0.6	0.5	0.0	0.0	1.0	1.0	1.0
ACS_PCT_RENTER_HU_avg	3234	26.8	9.6	0.0	21.3	26.0	31.8	82.0
POS_DIST_MEDSURG_ICU_TRACT_avg	3234	12.6	17.7	0.0	4.7	7.3	16.5	487.0
ACS_PCT_RENTER_HU_COST_50PCT_avg	3234	18.6	7.8	0.0	14.3	18.8	23.3	75.0
ACS_PCT_RENTER_HU_COST_30PCT_avg	3234	39.7	12.0	0.0	34.0	41.0	46.8	100.0
ACS_TOT_POP_POV_sum	3234	98504.7	322472.6	0.0	9612.3	23553.0	63324.0	9884138.0
ACS_PCT_OWNER_HU_COST_30PCT_avg	3234	17.7	5.7	0.0	14.8	17.5	20.7	42.8
CEN_POPDENSITY_TRACT_avg	3234	891.2	2798.2	0.0	41.5	209.1	907.0	90386.2
CEN_AREALAND_SQM_TRACT_avg	3234	289.4	1297.0	0.0	27.9	72.0	173.3	36393.3
ACS_MEDIAN_HH_INC_sum	3234	53803.8	17732.2	0.0	44620.5	52881.5	61834.8	148523.0
ACS_PCT_HH_ALONE_ABOVE65_avg	3234	12.7	4.1	0.0	10.6	12.8	14.8	66.0
ACS_PCT_AGE_ABOVE65_avg	3234	18.8	5.7	0.0	16.0	18.8	21.7	49.9
ACS_PCT_HU_NO_VEH_avg	3234	6.0	4.6	0.0	3.8	5.3	7.3	89.0
ACS_PCT_PUBL_TRANSIT_avg	3234	0.9	3.1	0.0	0.0	0.3	0.7	59.3
ACS_PCT_CHILD_1FAM_avg	3234	28.6	11.4	0.0	22.1	28.2	34.5	97.0
ACS_PCT_HH_INTERNET_avg	3234	75.7	15.5	0.0	72.8	79.0	83.5	97.0
POS_DIST_ED_TRACT_avg	3234	8.0	11.4	0.0	4.0	5.7	8.5	361.0
ACS_PCT_HU_MOBILE_HOME_avg	3234	12.3	9.6	0.0	4.7	10.0	18.0	56.5
ACS_PCT_DISABLE_avg	3234	15.6	5.2	0.0	12.6	15.3	18.6	48.0

Table 3_3.1_Descriptive Statistics (Main Summary) - Feature Descriptive Statistics.

3.2 Skewness and Kurtosis

Several predictors exhibit strong skewness and/or kurtosis, indicating potential non-normality and the presence of outliers:

- ACS_TOT_POP_POV_sum: Skew = 13.74, Kurtosis = 314.4
- CEN_POPDENSITY_TRACT_avg: Skew = 17.2, Kurtosis = 436.2
- POS_DIST_ED_TRACT_avg: Skew = 15.3, Kurtosis = 377.6
- ACS_PCT_PUBL_TRANSIT_avg: Skew = 11.1, Kurtosis = 162.0

These variables may benefit from log or cube root transformation in future modeling phases. However, some degree of skew is expected in public health datasets due to rural divides.

The following table presents the skewness and kurtosis values for the features in our dataset. These statistics provide insights into the shape and tail behavior of each feature's distribution:

Features	**skewness**	**kurtosis**
COUNTYFIPS	0.17	-0.61
BED_UTIL_RATIO	0.13	-0.44
IS_METRO_MICRO	-0.55	-1.70
ACS_PCT_RENTER_HU_avg	0.30	2.21
POS_DIST_MEDSURG_ICU_TRACT_avg	11.04	215.93
ACS_PCT_RENTER_HU_COST_50PCT_avg	0.21	2.48
ACS_PCT_RENTER_HU_COST_30PCT_avg	-0.95	2.54
ACS_TOT_POP_POV_sum	13.74	314.35
ACS_PCT_OWNER_HU_COST_30PCT_avg	-0.22	2.15

CEN_POPDENSITY_TRACT_avg	17.23	436.21
CEN_AREALAND_SQM_TRACT_avg	17.11	367.58
ACS_MEDIAN_HH_INC_sum	0.28	3.51
ACS_PCT_HH_ALONE_ABOVE65_avg	0.51	11.21
ACS_PCT_AGE_ABOVE65_avg	-0.35	2.96
ACS_PCT_HU_NO_VEH_avg	6.70	84.15
ACS_PCT_PUBL_TRANSIT_avg	11.09	162.00
ACS_PCT_CHILD_1FAM_avg	0.34	1.88
ACS_PCT_HH_INTERNET_avg	-3.36	13.64
POS_DIST_ED_TRACT_avg	15.35	377.57
ACS_PCT_HU_MOBILE_HOME_avg	0.99	0.73
ACS_PCT_DISABLE_avg	-0.10	1.89

Table 4_3.2 Skewness and Kurtosis - Shape of Distribution Metrics.

3.3 Categorical Distributions

STATE includes 56 unique values, with Texas (n = 254), Georgia (n = 159), and Virginia (n = 133) being the most represented. REGION values are distributed as follows:

- South: 1,422 counties
- Midwest: 1,055 counties
- West: 449 counties
- Northeast: 217 counties
- Missing: 96 counties

These distributions confirm the geographic diversity of the dataset, reinforcing the need for regional variables in modeling and interpretation.

4 Visual Interpretation of Trends

4.1 Histogram of BED_UTIL_RATIO

This histogram shows a near-normal distribution with a slight right skew. Most counties cluster between 0.2 and 0.5, with fewer outliers beyond 0.6. The mean and median both hover around 0.34. This supports our decision to use this variable as a continuous target.

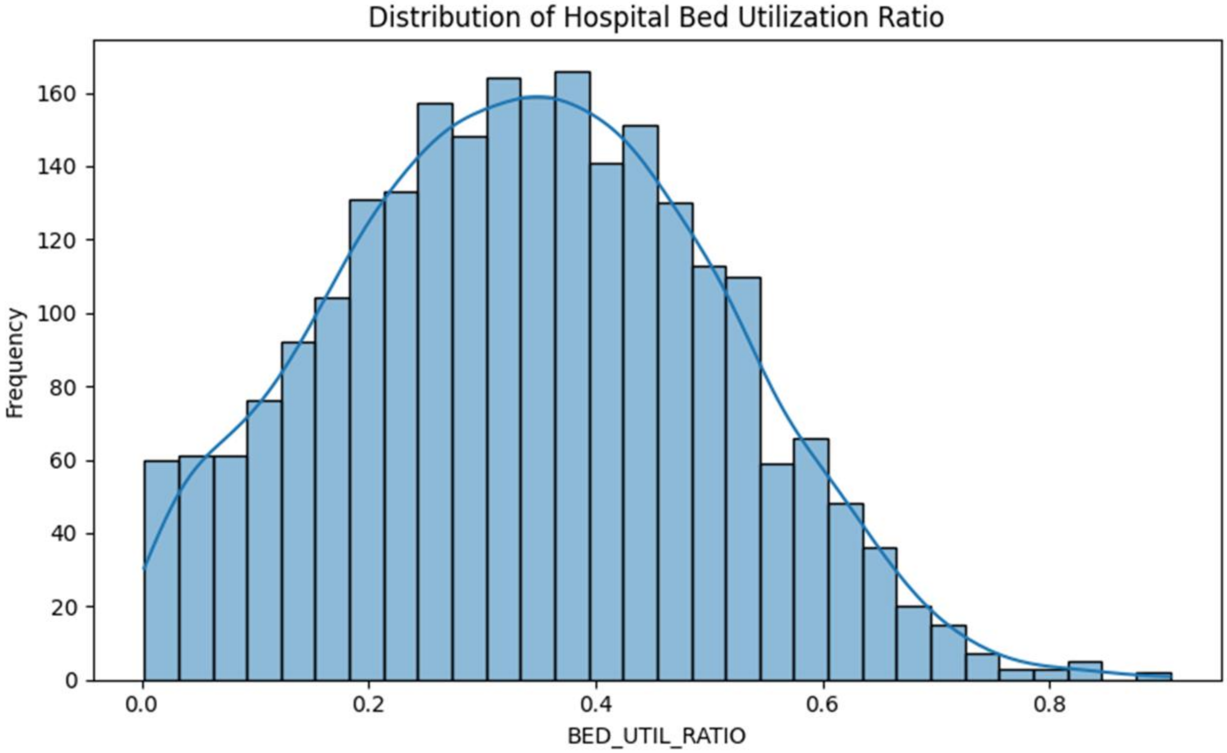


Figure 1_4.1 Histogram of BED_UTIL_RATIO – Distribution of Hospital Bed Utilization Ratio.

4.2 Boxplot by REGION

The following Boxplots reveal regional disparities. The South shows a wider spread and more high-end outliers, while the Midwest has more centralized utilization. The visible differences in the distribution of bed utilization across regions strongly suggest that

REGION is a significant factor influencing bed utilization. Therefore, retaining this feature for stratification in further analysis or modeling would be important to account for these regional variations.

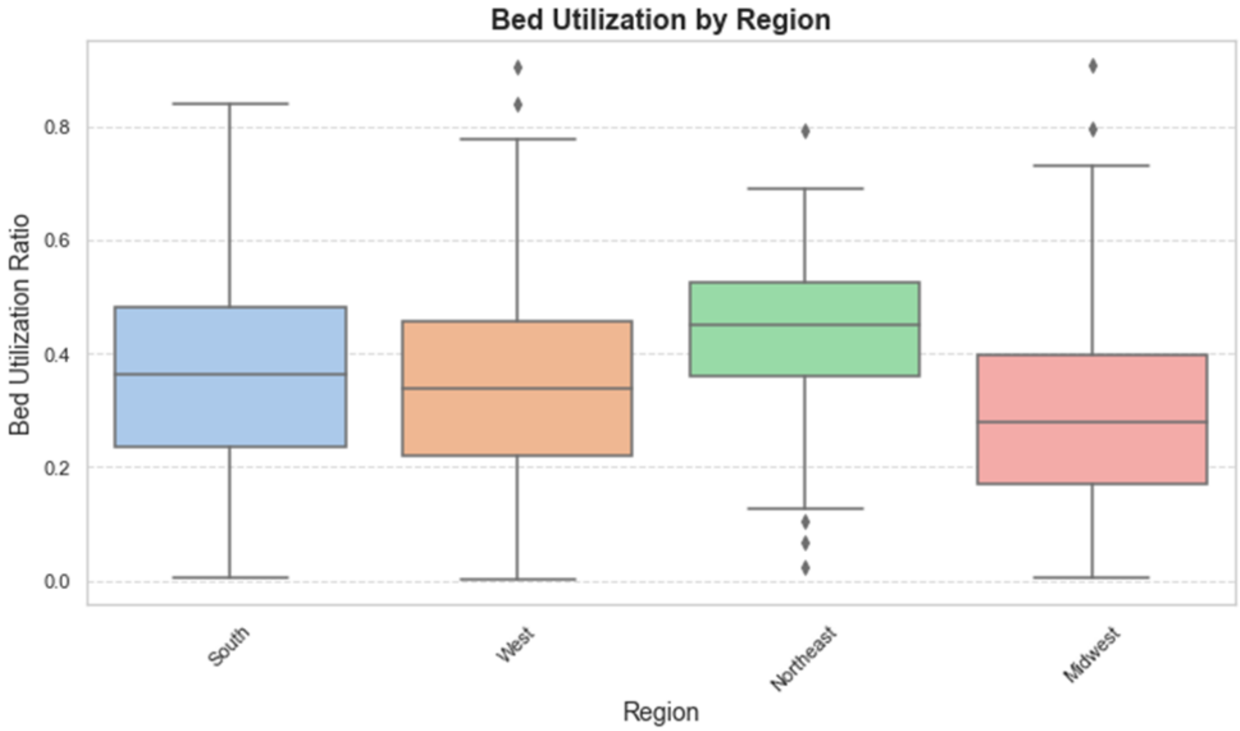


Figure 2_4.2 Boxplot by REGION - Bed Utilization by Region.

4.3 Bar Chart: Average Distance to Health Facilities by Region

The following Bar Chart plot reveals regional disparities in geographic access to healthcare. The West region exhibits the greatest average distances to both EDs and ICU facilities, followed by the South while the Northeast and Midwest show notably shorter distances. These patterns highlight the structural disadvantage faced by rural or sparsely populated areas, where hospitals and trauma facilities are less densely distributed. Such distance-related barriers can lead to delays in care, which may partially explain higher hospital bed utilization in regions with longer travel times. These findings support the inclusion of proximity features as predictive variables in future modeling.

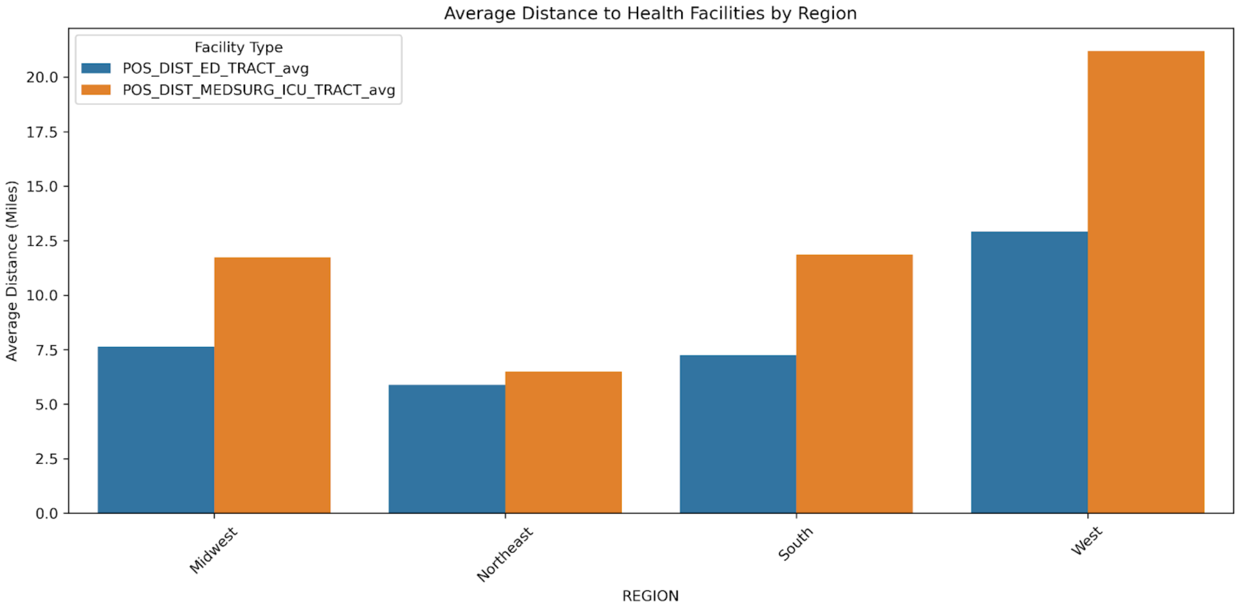


Figure 3_4.3 Bar Chart: Average Distance to Health Facilities by Region.

4.4 Pairwise Relationship of Vulnerability Indicators

A pairwise scatterplot matrix as seen in figure 4, was generated to explore relationships between three vulnerability-related predictors:

1. ACS_PCT_UNINSURED_avg (percentage of uninsured individuals)
2. ACS_PCT_DISABLE_avg (percentage of individuals with disabilities)
3. ACS_PCT_AGE_ABOVE65_avg (percentage of population aged 65 and over)

This visualization helps reveal potential interactions and collinearity:

- The strong positive correlation between age and disability confirms that older populations are more likely to have health conditions, aligning with theoretical expectations.
- The uninsured rate is weakly negatively correlated with both age and disability.

All three variables are right-skewed, consistent with previous descriptive statistics.

These findings validate the inclusion of these features in modeling efforts while also supporting the use of multicollinearity checks (e.g., VIF) to monitor shared variance.

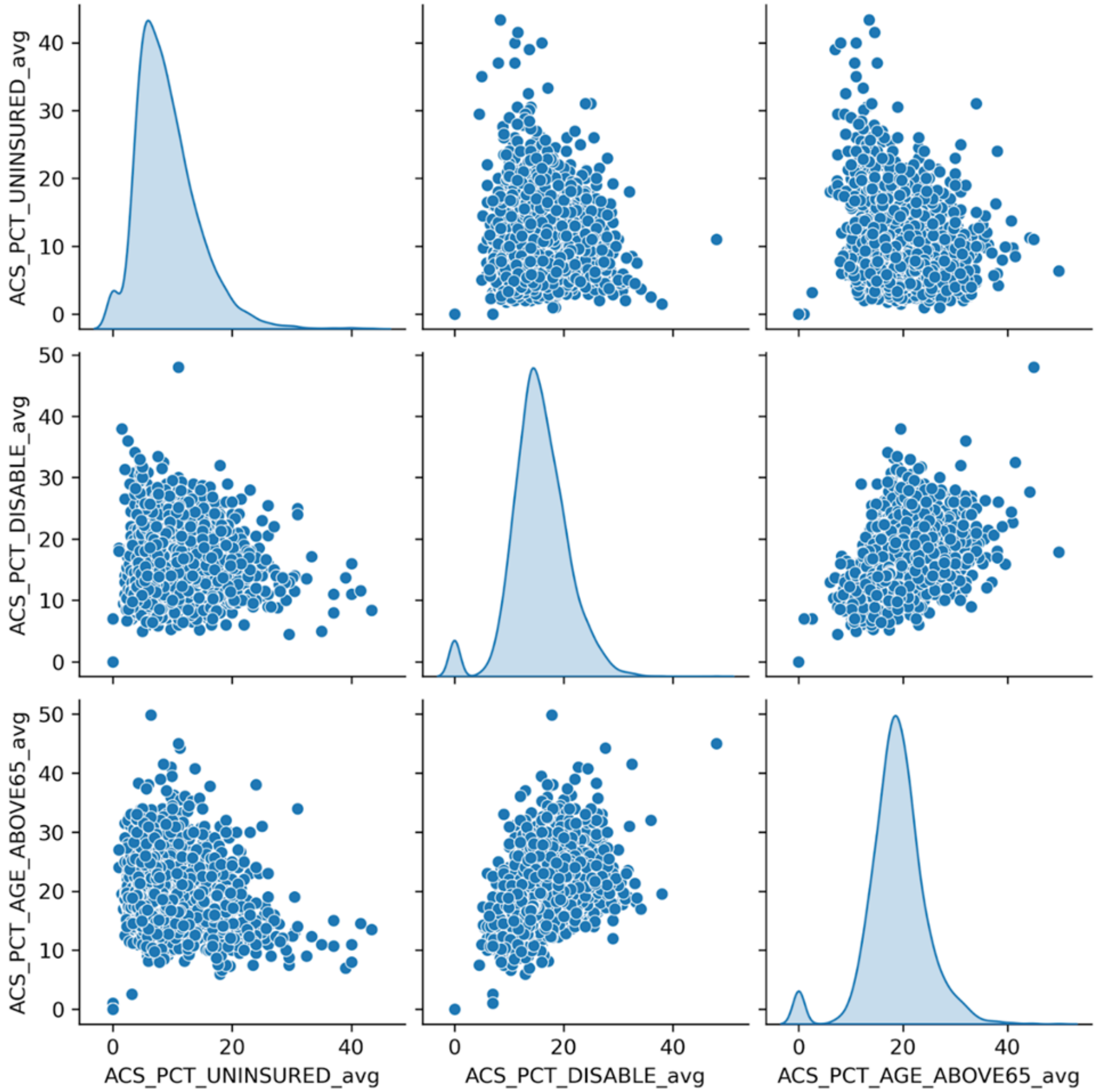


Figure 4_4.4 Pairwise Relationship of Vulnerability Indicators.

4.5 Bar Charts – Top 5 Counties by Hospitalization Risk in Each Region

This visualization emphasizes the geographic clustering of high-burden counties. In the Midwest, counties like Bottineau and Genesee top the list, while in the West, rural counties such as Polk and Whatcom show elevated utilization. The South exhibits consistently high BED_UTIL_RATIO values across multiple counties, suggesting broader systemic utilization in that region. These regional profiles reflect variation in hospital capacity, demand, and underlying health vulnerabilities. The visual also strengthens our hypothesis that social and geographic indicators such as access to ICU care, insurance coverage, and rental burden are tightly linked to observed strain on hospital systems.

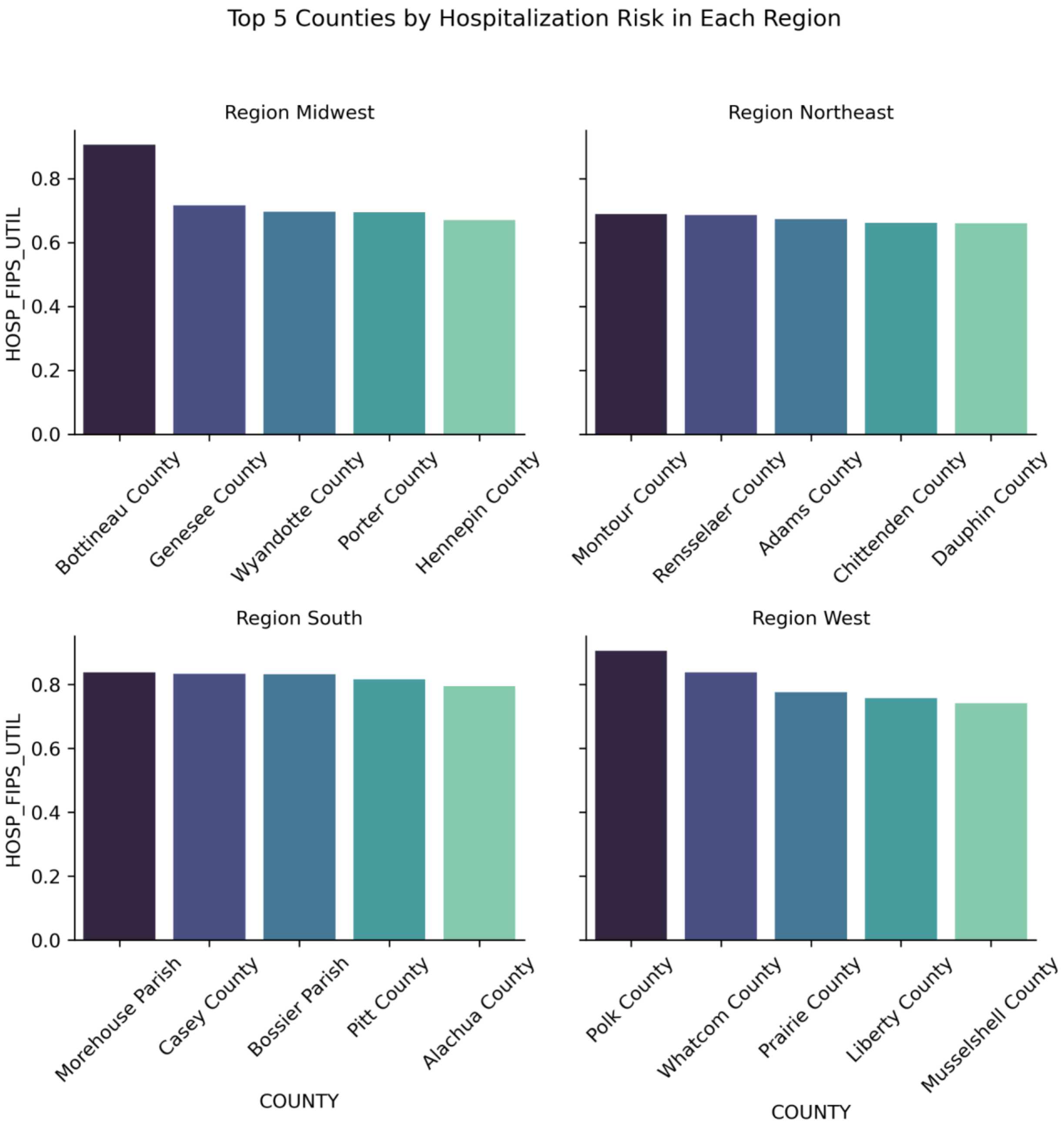


Figure 5_4.5 Bar Charts – Top 5 Counties by Hospitalization Risk in Each Region.

4.6 Correlation Heatmap

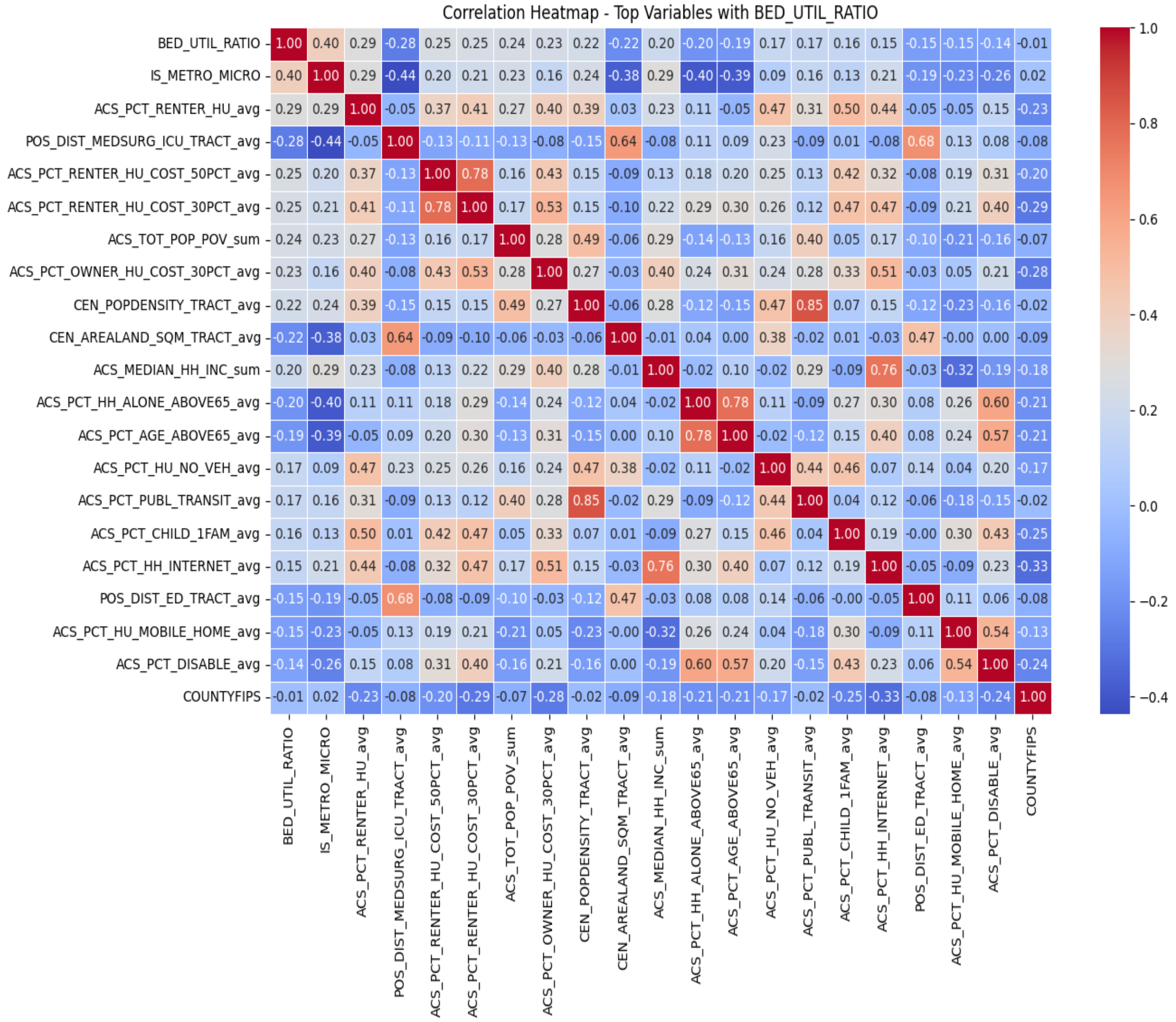
To evaluate relationships among predictors and with the outcome variable, we performed Pearson correlation and VIF analysis:

We computed pairwise Pearson correlations among all 21 numeric variables in the selected dataset, including the target (BED_UTIL_RATIO). The heatmap revealed several strong relationships, particularly:

- IS_METRO_MICRO (r = 0.40)
- ACS_PCT_RENTER_HU_avg (r = 0.29)
- POS_DIST_MEDSURG_ICU_TRACT_avg (r = 0.28)
- IS_METRO_MICRO (r = 0.40)
- ACS_PCT_RENTER_HU_avg (r = 0.29)
- ACS_TOT_POP_POV_sum (r = 0.24)

ACS_PCT_RENTER_HU_COST_50PCT_avg and ACS_PCT_RENTER_HU_COST_30PCT_avg were moderately correlated both with the target and with each other. Their correlation with BED_UTIL_RATIO is around 0.25, and the correlation between themselves is very high (~0.83), indicating potential multicollinearity.

Features related to population density, rental burden, and access appear to have notable correlation with target. Moderate collinearity is visible in related housing and poverty metrics.



	BED_UTIL_RATIO	IS_METRO_MICRO	ACS_PCT_RENTER_HU_avg	POS_DIST_MEDSURG_ICU_TRACT_avg	ACS_PCT_RENTER_HU_COST_50PCT_avg	ACS_PCT_RENTER_HU_COST_30PCT_avg	ACS_TOT_POP_POV_sum	ACS_PCT_OWNER_HU_COST_30PCT_avg	CEN_POPDENSITY_TRACT_avg	CEN_AREALAND_SQM_TRACT_avg	ACS_MEDIAN_HH_INC_sum	ACS_PCT_HH_ALONE_ABOVE65_avg	ACS_PCT_AGE_ABOVE65_avg	ACS_PCT_HU_NO_VEH_avg	ACS_PCT_PUBL_TRANSIT_avg	ACS_PCT_CHILD_1FAM_avg	ACS_PCT_HH_INTERNET_avg	POS_DIST_ED_TRACT_avg	ACS_PCT_HU_MOBILE_HOME_avg	ACS_PCT_DISABLE_avg	COUNTYFIPS
BED_UTIL_RATIO	1.00	0.40	0.29	-0.28	0.25	0.25	0.24	0.23	0.22	-0.22	0.20	-0.20	-0.19	0.17	0.17	0.16	0.15	-0.15	-0.15	-0.14	-0.01
IS_METRO_MICRO	0.40	1.00	0.29	-0.44	0.20	0.21	0.23	0.16	0.24	-0.38	0.29	-0.40	-0.39	0.09	0.16	0.13	0.21	-0.19	-0.23	-0.26	0.02
ACS_PCT_RENTER_HU_avg	0.29	0.29	1.00	-0.05	0.37	0.41	0.27	0.40	0.39	0.03	0.23	0.11	-0.05	0.47	0.31	0.50	0.44	-0.05	-0.05	0.15	-0.23
POS_DIST_MEDSURG_ICU_TRACT_avg	-0.28	-0.44	-0.05	1.00	-0.13	-0.11	-0.13	-0.08	-0.15	0.64	-0.08	0.11	0.09	0.23	-0.09	0.01	-0.08	0.68	0.13	0.08	-0.08
ACS_PCT_RENTER_HU_COST_50PCT_avg	0.25	0.20	0.37	-0.13	1.00	0.78	0.16	0.43	0.15	-0.09	0.13	0.18	0.20	0.25	0.13	0.42	0.32	-0.08	0.19	0.31	-0.20
ACS_PCT_RENTER_HU_COST_30PCT_avg	0.25	0.21	0.41	-0.11	0.78	1.00	0.17	0.53	0.15	-0.10	0.22	0.29	0.30	0.26	0.12	0.47	0.47	-0.09	0.21	0.40	-0.29
ACS_TOT_POP_POV_sum	0.24	0.23	0.27	-0.13	0.16	0.17	1.00	0.28	0.49	-0.06	0.29	-0.14	-0.13	0.16	0.40	0.05	0.17	-0.10	-0.21	-0.16	-0.07
ACS_PCT_OWNER_HU_COST_30PCT_avg	0.23	0.16	0.40	-0.08	0.43	0.53	0.28	1.00	0.27	-0.03	0.40	0.24	0.31	0.24	0.28	0.33	0.51	-0.03	0.05	0.21	-0.28
CEN_POPDENSITY_TRACT_avg	0.22	0.24	0.39	-0.15	0.15	0.15	0.49	0.27	1.00	-0.06	0.28	-0.12	-0.15	0.47	0.85	0.07	0.15	-0.12	-0.23	-0.16	-0.02
CEN_AREALAND_SQM_TRACT_avg	-0.22	-0.38	0.03	0.64	-0.09	-0.10	-0.06	-0.03	-0.06	1.00	-0.01	0.04	0.00	0.38	-0.02	0.01	-0.03	0.47	-0.00	0.00	-0.09
ACS_MEDIAN_HH_INC_sum	0.20	0.29	0.23	-0.08	0.13	0.22	0.29	0.40	0.28	-0.01	1.00	-0.02	0.10	-0.02	0.29	-0.09	0.76	-0.03	-0.32	-0.19	-0.18
ACS_PCT_HH_ALONE_ABOVE65_avg	-0.20	-0.40	0.11	0.11	0.18	0.29	-0.14	0.24	-0.12	0.04	-0.02	1.00	0.78	0.11	-0.09	0.27	0.30	0.08	0.26	0.60	-0.21
ACS_PCT_AGE_ABOVE65_avg	-0.19	-0.39	-0.05	0.09	0.20	0.30	-0.13	0.31	-0.15	0.00	0.10	0.78	1.00	-0.02	-0.12	0.15	0.40	0.08	0.24	0.57	-0.21
ACS_PCT_HU_NO_VEH_avg	0.17	0.09	0.47	0.23	0.25	0.26	0.16	0.24	0.47	0.38	-0.02	0.11	-0.02	1.00	0.44	0.46	0.07	0.14	0.04	0.20	-0.17
ACS_PCT_PUBL_TRANSIT_avg	0.17	0.16	0.31	-0.09	0.13	0.12	0.40	0.28	0.85	-0.02	0.29	-0.09	-0.12	0.44	1.00	0.04	0.12	-0.06	-0.18	-0.15	-0.02
ACS_PCT_CHILD_1FAM_avg	0.16	0.13	0.50	0.01	0.42	0.47	0.05	0.33	0.07	0.01	-0.09	0.27	0.15	0.46	0.04	1.00	0.19	-0.00	0.30	0.43	-0.25
ACS_PCT_HH_INTERNET_avg	0.15	0.21	0.44	-0.08	0.32	0.47	0.17	0.51	0.15	-0.03	0.76	0.30	0.40	0.07	0.12	0.19	1.00	-0.05	-0.09	0.23	-0.33
POS_DIST_ED_TRACT_avg	-0.15	-0.19	-0.05	0.68	-0.08	-0.09	-0.10	-0.03	-0.12	0.47	-0.03	0.08	0.08	0.14	-0.06	-0.00	-0.05	1.00	0.11	0.06	-0.08
ACS_PCT_HU_MOBILE_HOME_avg	-0.15	-0.23	-0.05	0.13	0.19	0.21	-0.21	0.05	-0.23	-0.00	-0.32	0.26	0.24	0.04	-0.18	0.30	-0.09	0.11	1.00	0.54	-0.13
ACS_PCT_DISABLE_avg	-0.14	-0.26	0.15	0.08	0.31	0.40	-0.16	0.21	-0.16	0.00	-0.19	0.60	0.57	0.20	-0.15	0.43	0.23	0.06	0.54	1.00	-0.24
COUNTYFIPS	-0.01	0.02	-0.23	-0.08	-0.20	-0.29	-0.07	-0.28	-0.02	-0.09	-0.18	-0.21	-0.21	-0.17	-0.02	-0.25	-0.33	-0.08	-0.13	-0.24	1.00

Figure 6_4.6 Correlation Heatmap with Hospital Bed Utilization.

4.7 U.S. Choropleth Map of BED_UTIL_RATIO by County

This choropleth map illustrates spatial clustering of higher hospital bed utilization in counties across the South and parts of the West. It highlights the utility of incorporating spatial and SDOH variables to monitor geographic vulnerability.

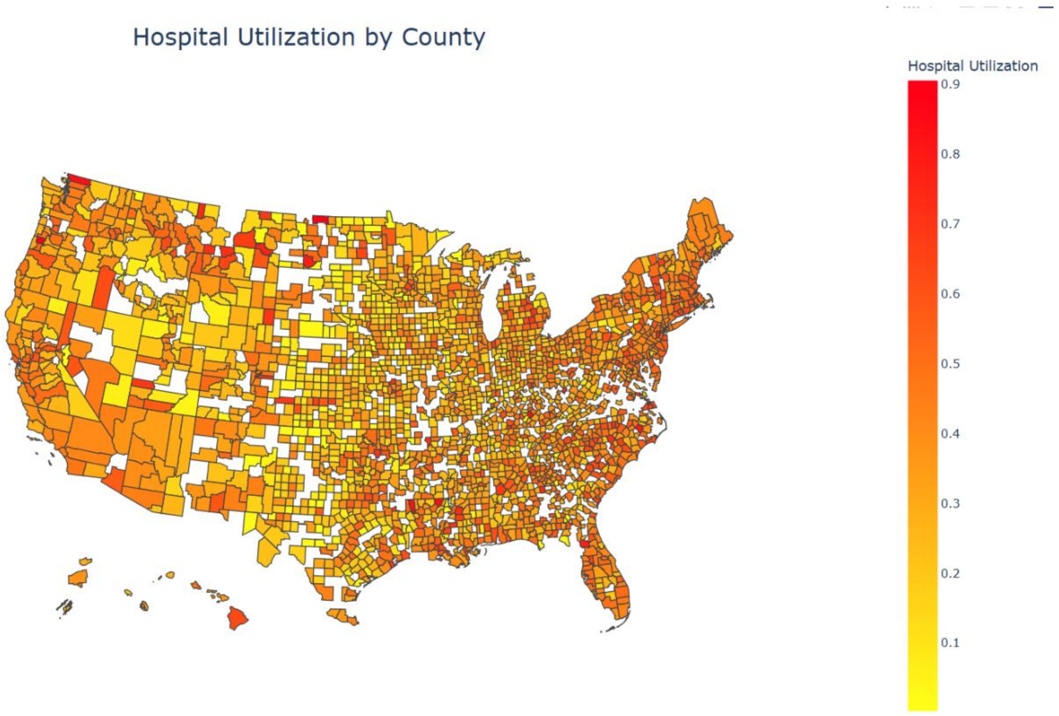


Figure 7_4.7 U.S. Choropleth Map of BED_UTIL_RATIO by County.

5 Collinearity and Feature Redundancy

To assess multicollinearity, we computed Variance Inflation Factors (VIF) after standardizing all 20 numeric predictors as seen in figure 8.

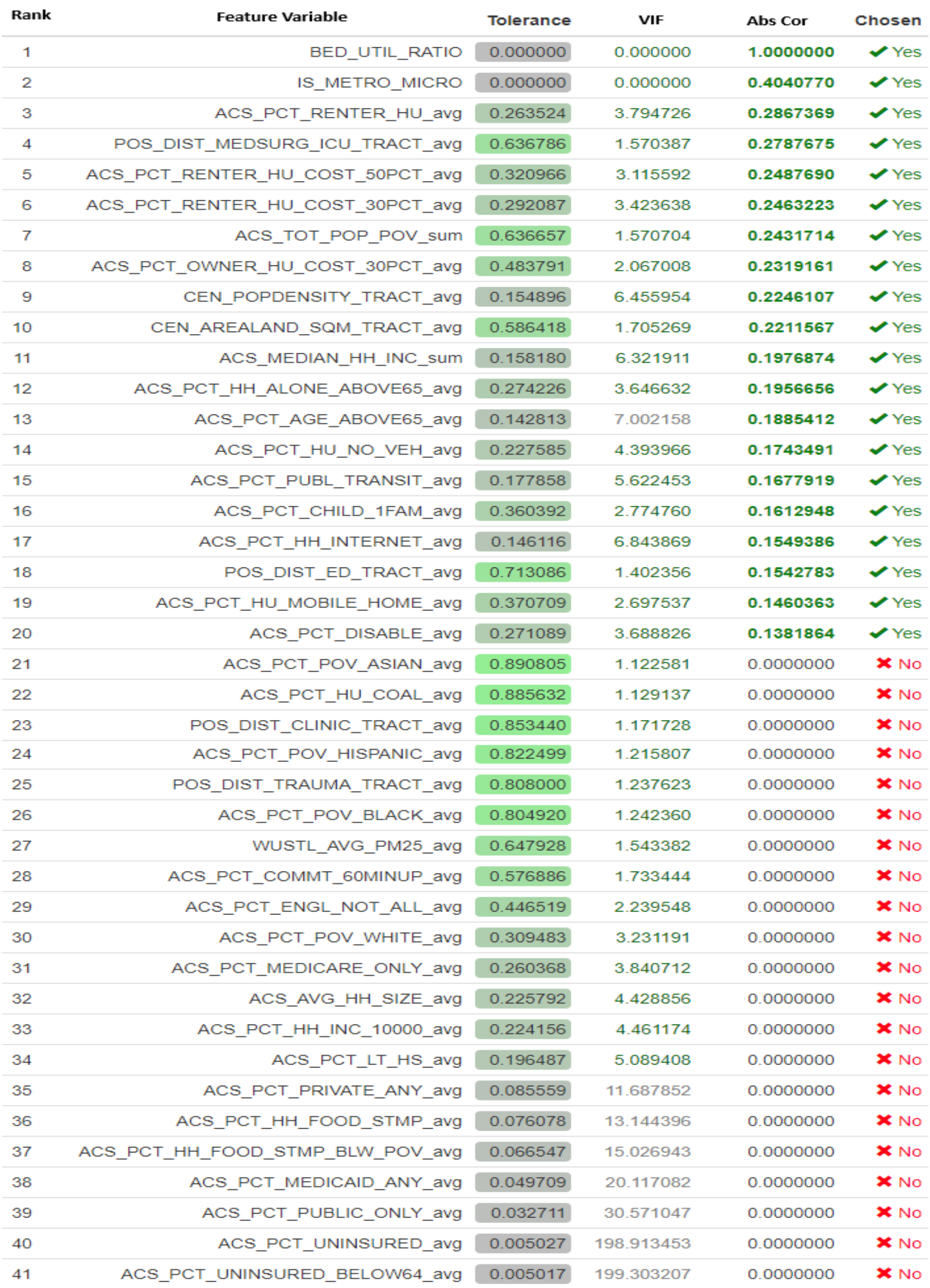
- 34 variables had VIF < 10 in the full dataset
- In our refined set of 23, all 20 numeric features used in modeling had VIF below 10
- Most VIF values were in the 1–6 range, indicating low to moderate correlation among predictors.

For instance:

- IS_METRO_MICRO: VIF = 1.86
- ACS_TOT_POP_POV_sum: VIF = 1.57

- CEN_POPDENSITY_TRACT_avg: VIF = 6.48

This analysis confirms that multicollinearity is not a major concern within the selected features, and the variables are suitable for use in penalized or regularized models like Ridge or Lasso.



Rank	Feature Variable	Tolerance	VIF	Abs Cor	Chosen
1	BED_UTIL_RATIO	0.000000	0.000000	1.0000000	✓ Yes
2	IS_METRO_MICRO	0.000000	0.000000	0.4040770	✓ Yes
3	ACS_PCT_RENTER_HU_avg	0.263524	3.794726	0.2867369	✓ Yes
4	POS_DIST_MEDSURG_ICU_TRACT_avg	0.636786	1.570387	0.2787675	✓ Yes
5	ACS_PCT_RENTER_HU_COST_50PCT_avg	0.320966	3.115592	0.2487690	✓ Yes
6	ACS_PCT_RENTER_HU_COST_30PCT_avg	0.292087	3.423638	0.2463223	✓ Yes
7	ACS_TOT_POP_POV_sum	0.636657	1.570704	0.2431714	✓ Yes
8	ACS_PCT_OWNER_HU_COST_30PCT_avg	0.483791	2.067008	0.2319161	✓ Yes
9	CEN_POPDENSITY_TRACT_avg	0.154896	6.455954	0.2246107	✓ Yes
10	CEN_AREALAND_SQM_TRACT_avg	0.586418	1.705269	0.2211567	✓ Yes
11	ACS_MEDIAN_HH_INC_sum	0.158180	6.321911	0.1976874	✓ Yes
12	ACS_PCT_HH_ALONE_ABOVE65_avg	0.274226	3.646632	0.1956656	✓ Yes
13	ACS_PCT_AGE_ABOVE65_avg	0.142813	7.002158	0.1885412	✓ Yes
14	ACS_PCT_HU_NO_VEH_avg	0.227585	4.393966	0.1743491	✓ Yes
15	ACS_PCT_PUBL_TRANSIT_avg	0.177858	5.622453	0.1677919	✓ Yes
16	ACS_PCT_CHILD_1FAM_avg	0.360392	2.774760	0.1612948	✓ Yes
17	ACS_PCT_HH_INTERNET_avg	0.146116	6.843869	0.1549386	✓ Yes
18	POS_DIST_ED_TRACT_avg	0.713086	1.402356	0.1542783	✓ Yes
19	ACS_PCT_HU_MOBILE_HOME_avg	0.370709	2.697537	0.1460363	✓ Yes
20	ACS_PCT_DISABLE_avg	0.271089	3.688826	0.1381864	✓ Yes
21	ACS_PCT_POV_ASIAN_avg	0.890805	1.122581	0.0000000	✗ No
22	ACS_PCT_HU_COAL_avg	0.885632	1.129137	0.0000000	✗ No
23	POS_DIST_CLINIC_TRACT_avg	0.853440	1.171728	0.0000000	✗ No
24	ACS_PCT_POV_HISPANIC_avg	0.822499	1.215807	0.0000000	✗ No
25	POS_DIST_TRAUMA_TRACT_avg	0.808000	1.237623	0.0000000	✗ No
26	ACS_PCT_POV_BLACK_avg	0.804920	1.242360	0.0000000	✗ No
27	WUSTL_AVG_PM25_avg	0.647928	1.543382	0.0000000	✗ No
28	ACS_PCT_COMMT_60MINUP_avg	0.576886	1.733444	0.0000000	✗ No
29	ACS_PCT_ENGL_NOT_ALL_avg	0.446519	2.239548	0.0000000	✗ No
30	ACS_PCT_POV_WHITE_avg	0.309483	3.231191	0.0000000	✗ No
31	ACS_PCT_MEDICARE_ONLY_avg	0.260368	3.840712	0.0000000	✗ No
32	ACS_AVG_HH_SIZE_avg	0.225792	4.428856	0.0000000	✗ No
33	ACS_PCT_HH_INC_10000_avg	0.224156	4.461174	0.0000000	✗ No
34	ACS_PCT_LT_HS_avg	0.196487	5.089408	0.0000000	✗ No
35	ACS_PCT_PRIVATE_ANY_avg	0.085559	11.687852	0.0000000	✗ No
36	ACS_PCT_HH_FOOD_STMP_avg	0.076078	13.144396	0.0000000	✗ No
37	ACS_PCT_HH_FOOD_STMP_BLW_POV_avg	0.066547	15.026943	0.0000000	✗ No
38	ACS_PCT_MEDICAID_ANY_avg	0.049709	20.117082	0.0000000	✗ No
39	ACS_PCT_PUBLIC_ONLY_avg	0.032711	30.571047	0.0000000	✗ No
40	ACS_PCT_UNINSURED_avg	0.005027	198.913453	0.0000000	✗ No
41	ACS_PCT_UNINSURED_BELOW64_avg	0.005017	199.303207	0.0000000	✗ No

Figure 8_5 Collinearity and Feature Redundancy.

6 Discussion and Key Takeaways

The exploratory analysis supports our hypothesis that counties with indicators of social vulnerability—including lower insurance coverage, older populations, high rental burden, and greater distance to care—tend to experience higher levels of hospital bed utilization. These findings reinforce prior work on the role of social determinants in shaping healthcare system strain during crises (Artiga & Hinton, 2018; Garg et al., 2020).

Multicollinearity was evaluated and found to be moderate. Variance Inflation Factor (VIF) diagnostics were used to ensure that selected predictors are statistically stable and interpretable. Through this process, we confirmed the suitability of 23 features for downstream modeling.

Spatial and regional characteristics also emerged as critical. For example, counties in the South—with higher rental burdens and greater distances to ICU facilities tend to exhibit elevated BED_UTIL_RATIO values. Visualizations such as choropleth maps, boxplots by region, and scatterplots of access metrics provide additional evidence of geographic disparities and potential clustering, supporting the case for future spatial modeling or stratified interventions.

Key conclusions:

- BED_UTIL_RATIO is a usable continuous outcome variable, showing moderate right skew and meaningful geographic variation.
- The 23 selected features demonstrate both theoretical relevance and statistical validity for modeling healthcare system stress.

- Use of publicly available datasets enhances reproducibility, transparency, and scalability of this work across policy, academic, and local government settings.

7 Next Steps

In the upcoming modeling phase, we will begin imputing missing values using mean for numeric fields and mode for categorical features such as REGION. STATE will be encoded using frequency methods, while REGION will undergo one-hot encoding.

Dimensionality reduction techniques like PCA will be applied, followed by supervised modeling using Lasso or Ridge Regression to evaluate performance. Model assessment will include residual analysis, calibration plots, and ROC curves to determine utility and fairness.

8 References

1. Agency for Healthcare Research and Quality (AHRQ). (2020). AHRQ SDOH Database 2020. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
2. Artiga, S., & Hinton, E. (2018). Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity. Kaiser Family Foundation. https://www.kff.org/disparities-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ or https://www.cdc.gov/mmwr/volumes/69/wr/pdfs/mm6915e3-H.pdf
3. Garg, S., Kim, L., Whitaker, M., et al. (2020). Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 — COVID-

NET, 14 States, March 1–30, 2020. MMWR Morb Mortal Wkly Rep, 69(15), 458–464. https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e3.htm or https://www.cdc.gov/mmwr/volumes/69/wr/pdfs/mm6915e3-H.pdf

4. HHS. (2020). COVID-19 Reported Patient Impact and Hospital Capacity by Facility. https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u

5. Magnan, S. (2017). Social Determinants of Health 101 for Health Care: Five Plus Five. National Academy of Medicine. https://nam.edu/social-determinants-of-health-101-for-health-care-five-plus-five/ or https://nam.edu/wp-content/uploads/2017/10/Social-Determinants-of-Health-101.pdf

9 Appendix A - Changes to the Data Dictionary

Over the course of our exploratory data analysis, we determined that 40+ feature variables would be too challenging to manage and not improve the outcome of our project. As previously mentioned, we looked at the VIF, tolerance and absolute correlation measures to identify which feature variables to include. An updated data dictionary can be found at this github location https://github.com/lemieuxjm-cap/Iota-Capstone/blob/main/reports/codebook.md.

In addition to reducing the number of feature variables, we also determined that the existing variable names were challenging to read and were not helpful nor meaningful identifiers. Therefore, as part of our EDA, we established a set of more representative variable names that we will incorporate starting this week. Our data dictionary now identifies the original variable name and an associated modified name.

1.