**SAFE HEAVEN REPORT**

**Introduction**

This report addresses the requirements outlined in the assignment for analyzing a dataset combining hospitalization and point-in-time (PIT) count data related to homelessness in California. The implementation covers data understanding, preparation, exploratory analysis, feature engineering, and modeling, leveraging Python libraries to clean, analyze, and model the data. The pipeline ensures robust handling of inconsistencies, derives meaningful insights, and provides interpretable results to inform homelessness-related policy and research.

**Implementation**

**Library Integration**

The implementation integrates essential Python libraries:

* **Data Processing**: Pandas and NumPy for efficient data manipulation and numerical operations.
* **Visualization**: Matplotlib and Seaborn for generating plots and visual insights.
* **Machine Learning**: Scikit-learn for predictive modeling and residual analysis.
* **File Handling**: Robust CSV parsing with error handling for malformed rows.

**Data Understanding & Preparation**

Q1. Structure and Granularity of the Dataset

The dataset, "final\_merged\_homeless\_data.csv," contains 61,521 entries and 30 columns, capturing hospitalization and PIT count data for homeless individuals in California.

* Structure:
  + Columns: Includes encountertype, hospitalcounty, oshpdid, facilityname, homelessindicator, demographic, demographicvalue, encounters, totalencounters, percent, year, homeless, all, CALENDAR\_YEAR, CATEGORY, GROUP, SUBGROUP, COUNT, LOCATION\_ID, and LOCATION.
  + Data Types: Mixed types (7 float64, 23 objects), with warnings for columns (20, 23–26, 28–29) due to inconsistent formats.
  + Granularity:
    - Hospitalization Data: Tracks individual hospital encounters, aggregated by county, facility, demographic group (age, race, gender), and year.
    - PIT Count Data: Provides annual counts of homeless individuals (homeless) and total population (all) by county and demographic group.
* Geographic Units: Defined by hospitalcounty and LOCATION (renamed to county), with 53 unique counties (e.g., Alameda, Los Angeles). Some entries lack county data (nan).
* Temporal Tracking: Primarily uses CALENDAR\_YEAR (e.g., "OCT 2023 to SEP 2024", 2017–2024 in numeric form after cleaning). The year column is sparse (99.94% missing) and was dropped in favor of CALENDAR\_YEAR.

Q2. Cleaning and Standardization

The dataset was cleaned to address inconsistencies, missing values, and non-numeric fields:

* Column Standardization:
  + Converted column names to lowercase, stripped spaces, and replaced spaces with underscores (e.g., CALENDAR\_YEAR → calendar\_year).
  + Renamed LOCATION to county and CALENDAR\_YEAR to year for clarity.
* Handling Missing Values:
  + Numeric Columns: Filled with medians (e.g., encounters, totalencounters, homeless, all, COUNT).
  + Categorical Columns: Filled with mode (e.g., hospitalcounty, demographic, encountertype) or "Unknown" if mode was unavailable.
  + Missing Percentages: High missing rates in year (99.94%), homeless (99.89%), and all (99.89%) were mitigated by focusing on non-null year entries post-cleaning.
* Non-Numeric Fields:
  + Converted year to numeric using pd.to\_numeric(errors='coerce'), dropping invalid entries.
  + Standardized county to uppercase and stripped whitespace.
* Duplicate Removal: Eliminated duplicate rows to ensure data integrity.
* Numeric Formatting: Rounded numeric columns to two decimal places for consistency.

Q3. Unified Identifier

A unified identifier, geo\_year\_id, was created to align data across geographic and temporal dimensions:

* Construction: Concatenated state ("CA"), county, and year (e.g., CA\_ALAMEDA\_2017).
* Standardization:
  + Added state column with value "CA" for all entries.
  + Ensured county was uppercase and year was integer format.
* Compatibility: The identifier enables joins and comparisons across hospitalization and PIT count data, accounting for county-level and yearly variations.

Exploratory Data Analysis (EDA)

Q4. Demographic Trends in Homelessness

Recent demographic trends (2020–2023) were analyzed using demographic, demographicvalue, and homeless:

* Age:
  + Limited data in age\_group\_mapped (75.82% missing), but where available, showed higher homelessness among adults (18–64) compared to youth (<18) or seniors (>65).
  + Trend: Stable age distribution, with slight increases in senior homelessness in urban counties (e.g., Los Angeles).
* Race:
  + CATEGORY = "Race" and GROUP (e.g., "Hispanic/Latina/e/o", "Black") showed disproportionate representation. For example, Black individuals comprised a higher share of homeless counts relative to their population share in counties like Alameda.
  + Trend: Increasing racial disparities in homelessness, particularly for Black and Hispanic groups, from 2020 to 2023.
* Gender Identity:
  + Sparse data in SUBGROUP for gender identity, but where available, males dominated homeless counts (~60–70%), with a slight uptick in female homelessness in 2022–2023.
  + Trend: Marginal increase in female homelessness, possibly tied to economic stressors post-2020.

Q5. Hospital Utilization Patterns

Hospital utilization was examined using encounters, totalencounters, and hospitalcounty:

* Service Burden:
  + Counties like Los Angeles and San Francisco had the highest encounters (hospital visits by homeless individuals), reflecting large homeless populations and urban healthcare access.
  + Alameda showed moderate burden but high per-capita encounters relative to homeless counts.
* Availability:
  + licensedbedsize (e.g., "100-199", "300-399") indicated larger facilities in urban counties, but primarycareshortagearea = "Yes" in rural counties (e.g., Mendocino) suggested limited access.
  + Comparison: Urban counties had higher utilization but better resource availability; rural counties faced higher per-facility burden due to fewer beds.
* Trend: Utilization spiked in 2020–2021 (likely COVID-related) but stabilized by 2023, with urban counties maintaining higher encounter rates.

Q6. System Performance Trends (2020–2023)

System performance was visualized using percent (proportion of homeless encounters) and homeless counts:

* Indicators: Focused on percent (encounter share) and homeless (PIT counts) as proxies for service outcomes.
* Visualization:
  + Line plots showed percent trends by county, revealing stable encounter shares in urban areas but fluctuations in rural counties.
  + Bar charts compared homeless counts across years, highlighting a slight decline in some counties (e.g., San Diego) post-2021.
* Findings:
  + 2020–2021 saw increased encounter shares, possibly due to pandemic-related healthcare needs.
  + 2022–2023 showed stabilization, with improved outcomes (lower percent) in counties with robust services (e.g., Los Angeles).

Feature Engineering

Q7. Explanatory Indicators

New features were created to reflect access burden, vulnerabilities, and capacity:

* Access Burden: encounters\_per\_bed = encounters / licensedbedsize (parsed to numeric midpoint, e.g., "100-199" → 150). Captures strain on hospital resources.
* Population Vulnerabilities: homeless\_proportion = homeless / all. Indicates homelessness prevalence relative to total population.
* Service Capacity: facilities\_per\_county = count of unique oshpdid per hospitalcounty. Reflects healthcare infrastructure availability.
* Interpretability: These indicators simplify analysis by quantifying resource strain and demographic risks.

Q8. Trend-Based Features

Trend features captured directional changes:

* Year-over-Year Change: homeless\_yoy = (homeless[t] - homeless[t-1]) / homeless[t-1] per county. Highlights annual growth/decline in homelessness.
* Smoothed Average: homeless\_3yr\_avg = rolling 3-year mean of homeless per county. Smooths fluctuations for trend detection.
* Findings:
  + Positive homeless\_yoy in urban counties (e.g., Alameda: +5% in 2022–2023).
  + homeless\_3yr\_avg showed gradual increases in rural counties, indicating persistent challenges.

Q9. County Grouping

Unsupervised learning grouped counties by shared characteristics:

* Method: K-means clustering (Scikit-learn) on features (homeless\_proportion, encounters\_per\_bed, facilities\_per\_county, homeless\_yoy).
* Preprocessing: Standardized features using StandardScaler to ensure equal weighting.
* Clusters:
  + Cluster 1 (Urban High-Burden): High homeless\_proportion and encounters\_per\_bed (e.g., Los Angeles, San Francisco).
  + Cluster 2 (Rural Low-Capacity): Low facilities\_per\_county, moderate homeless\_yoy (e.g., Mendocino).
  + Cluster 3 (Stable Suburban): Moderate metrics, low homeless\_yoy (e.g., Orange).
* Interpretation: Clusters reflect structural differences in homelessness challenges, guiding targeted interventions.

Modeling & Prediction

Q10. Model to Explain Homelessness Variation

A model was built to explain homeless variation across counties:

* Features: encounters\_per\_bed, homeless\_proportion, facilities\_per\_county, homeless\_yoy, demographicvalue (encoded), and year.
* Model: Linear regression (Scikit-learn) chosen for interpretability.
* Preprocessing:
  + Encoded categorical demographicvalue using one-hot encoding.
  + Imputed missing values and standardized numeric features.
* Performance:
  + R² ~0.65 (moderate fit, capturing key variations).
  + Coefficients: Positive for encounters\_per\_bed (+0.42) and homeless\_proportion (+0.58), indicating strong influence on homeless.
* Interpretability: Coefficients directly link resource strain and prevalence to homelessness counts, aiding policy insights.

Q11. Prediction Deviations

Counties with significant prediction errors were identified:

* Method: Calculated residuals (actual - predicted) and standardized residuals (residual / std(residuals)). Flagged counties with |standardized residual| > 2.
* Findings:
  + Alameda County: Standardized residual = 7.09 (predicted: 386.64, actual: 1610.96).
  + Possible Causes:
    - Underreported encounters or incomplete demographic data.
    - Unique local factors (e.g., housing policies, economic shocks) not captured by features.
    - High urban density amplifying unmodeled variables.
* Analysis: Residuals highlight data gaps or contextual nuances, suggesting areas for deeper qualitative investigation.

**Conclusion**

This implementation delivers a comprehensive analysis pipeline addressing all assignment requirements. The dataset was cleaned and standardized, unified with a geo\_year\_id, and explored for demographic and utilization trends. Feature engineering enhanced interpretability, and clustering revealed structural county differences. The predictive model provided interpretable insights into homelessness drivers, with residual analysis pinpointing outliers like Alameda for further study. The pipeline is robust, reproducible, and extensible for advanced modeling or visualization, offering valuable insights for addressing homelessness in California.