LA County Homelessness Prediction Challenge

Executive Summary

This competition represents a new model for data science competitions focused on real-world policy impact. Using publicly available data from Los Angeles County's homelessness response efforts and the US Census, this challenge asks data scientists to identify and weight the socioeconomic factors that drive homelessness patterns across LA County.

Policy Impact Potential: Results can directly inform LA County's $1+ billion annual homelessness budget allocation, service delivery strategies, and intervention targeting. Winners' insights could influence policy decisions affecting over 75,000 individuals experiencing homelessness.

Strategic Value for Kaggle: This competition pioneers a new category of policy-focused data science challenges that emphasize interpretable analysis over predictive accuracy. It attracts socially-conscious data scientists, demonstrates Kaggle's commitment to social impact, and creates opportunities for government partnerships and real-world applications of competitive data science.

Innovation in Competition Format: Unlike traditional prediction contests, this challenge evaluates participants on factor analysis, policy actionability, and interpretable insights using the RICE framework - expanding Kaggle's competition formats to serve the growing policy analytics field.

Competition Overview

Background

Homelessness is one of the most pressing social challenges facing Los Angeles County, with over 75,000 individuals experiencing homelessness on any given night. Understanding the socioeconomic factors that contribute to homelessness patterns across different geographic areas is crucial for effective resource allocation, policy development, and intervention strategies.

This competition challenges data scientists to build analytical models that can identify the relationship between community characteristics and homelessness outcomes using real data from LA County's Service Planning Areas (SPAs) and the US Census.

The Challenge

Can you identify and weight the factors leading to homelessness in LA County using these publicly available datasets?

You'll work with comprehensive datasets spanning 8 Service Planning Areas (SPAs) in LA County, combining. You may choose not to study at the SPA level: it is at the analysts discretion.

- **Census socioeconomic indicators** (income, employment, disability, housing costs)
- **SPA homeless population characteristics** (demographics, health conditions, family composition)

Competition Objectives

Primary Goals

1. **Factor Analysis**: Identify which socioeconomic factors are most strongly associated with homelessness outcomes
2. **Relative Weighting**: Determine the comparative importance of different community characteristics
3. **Actionable Insights**: Generate findings that could inform policy and resource allocation decisions

Specific Tasks

Contestants may approach this challenge from multiple angles:

Factor Analysis Tasks:

- Identify which socioeconomic factors most strongly correlate with homelessness rates
- Determine the relative importance of correlated factors
- Consider investigating individual vs household factors

Cohort Analysis:

- Are contributing factors more dominant in one SPA than another?
- Are there population segments that span all SPAs?

Output Expectations: While technical analysis (correlation matrices, statistical tests, model diagnostics) is essential for rigor, your final deliverables must translate findings into policy-relevant insights. Focus on creating clear, interpretable results that can inform decision-making by non-technical stakeholders. Be sure to include your analysis artifacts in an appendix so that

future Data Scientists can replicate your findings. Include links to your hosted notebooks and datasets.

Data Overview

Dataset Structure

The competition provides **6 comprehensive data tables** covering LA County's 8 Service Planning Areas:

Homelessness Data (1 Table)

1. **SPA 2020 - Homeless Population Count**: Comprehensive homeless population characteristics including:
   - Geographical location within Los Angeles County (SPA level)
   - Demographics (age, gender, race/ethnicity)
   - Vulnerability indicators (chronic homelessness, disability status, health conditions)
   - Household composition (individuals vs. families, youth, veterans)
   - Shelter status (sheltered vs. unsheltered)
   - Special populations (domestic violence survivors, LGBTQ+ individuals)

Census Socioeconomic Data (5 Census tables with appended SPA categories)

1. **P1 - Total Population**: Baseline population counts by demographic groups
2. **B18107 - Disability Status**: Population with various disability types by age and employment
3. **B23025 - Employment Status**: Labor force participation, unemployment rates, employment by demographics
4. **B25014 - Housing Occupancy**: Housing unit counts, occupancy status, vacancy patterns
5. **B25070 - Housing Cost Burden**: Gross rent as percentage of household income, cost burden analysis

Geographic Coverage

The region under study is Los Angeles County. This region is divided into 8 Service Planning Areas (SPAs) for greater detail. LA County's 8 administrative regions are:

1. Antelope Valley
2. San Fernando Valley

3. San Gabriel Valley
4. Metro Los Angeles
5. West Los Angeles County
6. South Los Angeles County
7. East Los Angeles County
8. South Bay

Data Scale

- **Population Coverage**: ~10 million residents across 8 SPAs
- **Homeless Population**: ~75,000 individuals (2020 PIT count)
- **Feature Count**: 100+ variables spanning socioeconomic, housing, and homeless population characteristics
- **Geographic Granularity**: Census tract level data aggregated to SPA level

Technical Details

Understanding PIT and SPA Data Relationship

- **Point-in-Time (PIT) Count**: A federally mandated annual census of homeless individuals conducted on a single night in January of each year for each of the 8 Los Angeles SPAs
- **Service Planning Areas (SPAs)**: LA County's 8 administrative regions used for service delivery and planning
- **Data Structure**: PIT homeless counts are organized and reported by SPA geography, providing homeless population characteristics within each SPA boundary

Known Issues in the Available Data

- All original data sets are aggregated and not at the individual level. While unfortunate, this is immutable and fundamental to the issue facing a better understanding of homelessness
- Available Census data contains no such homeless designation. While officials state that the census includes homeless counts, the data does not appear to be publicly available
- Data features across the SPA and Census data are similar, but not identical. Significant care will need to be taken on combining data sets
- SPA and Census data is not at the same geographic level. To help, the author has appended each census table with the correct SPA in as much as this is possible

- Not all LA County Census data maps to a SPA geography. Multiple census tracts in the census could not easily be matched to their correct SPA

Data Processing

- All Census data aggregated from tract level to SPA level using official geographic crosswalks
- PIT data provided at SPA level with comprehensive demographic and vulnerability breakdowns
- Missing values handled consistently across all tables
- Geographic boundaries validated for consistency

Modeling Approaches

Consider these strategies for your analysis:

- **Correlation analysis**: Examine relationships between Census and PIT variables
- **Regression techniques**: Quantify factor importance and effect sizes
- **Feature selection**: Identify the most important socioeconomic drivers
- **Dimensionality reduction**: Group related factors into meaningful categories
- **Statistical testing**: Validate significance of identified relationships

Evaluation Criteria

RICE Framework Assessment

Reach (25%)

- **Population Impact**: How many people could be affected by your findings?
- **Geographic Scope**: Do insights apply across multiple SPAs or specific regions?
- **Stakeholder Relevance**: How many different agencies/departments could use your results?

Impact (35%)

- **Policy Actionability**: How directly can findings inform policy decisions?
- **Resource Allocation**: Potential to improve funding and service delivery efficiency
- **Intervention Targeting**: Ability to identify high-leverage intervention points
- **Measurable Outcomes**: Clarity of expected results from recommended actions

Confidence (25%)

- **Statistical Rigor**: Appropriate methodology and validation for findings
- **Data Quality**: Transparent handling of limitations and uncertainties
- **Reproducibility**: Clear documentation enabling replication
- **Evidence Strength**: Robustness of conclusions given available data

Effort (15%)

- **Implementation Feasibility**: How realistic are your recommendations given current resources?
- **Time to Value**: How quickly could insights be acted upon?
- **Complexity**: Are proposed interventions manageable for existing systems?

Submission Requirements

Required Deliverables

1. **Factor Analysis Model**: Working analytical model identifying key socioeconomic factors and their relative importance
2. **Analysis Report**: 3-5 page summary of methodology, findings, and recommendations
3. **Code Submission**: Complete, reproducible analysis code
4. **Data Outputs**: Factor importance rankings and key analytical results in specified format

Report Structure

Your analysis report should include:

Executive Summary (0.5 pages)

- Key findings and policy recommendations
- Most important predictive factors identified
- Model performance summary

Methodology (1-1.5 pages)

- Data processing approach
- Modeling strategy and rationale
- Validation methodology
- Handling of small sample size challenges

Results (1-2 pages)

- Model performance metrics
- Feature importance analysis
- Key relationships discovered
- Uncertainty analysis

Policy Recommendations (1 page)

- Actionable recommendations for LA County
- Resource allocation insights
- Intervention opportunities
- Limitations and caveats

Technical Specifications

- **Code**: R, Python, or other specified languages
- **Format**: Jupyter notebooks or R Markdown preferred
- **Documentation**: Clear comments and explanation of approach
- **Reproducibility**: All code should run without modification
- **File Naming**: Follow specified naming conventions

Data Access and Setup

Dataset Location

All competition data is provided in structured CSV format:

```
datasets/
├── census/           # 5 Census socioeconomic tables
├── SPA/              # SPA homeless population data
├── geographic/       # SPA-ZIP-Tract mapping files
└── processed/        # Pre-processed analysis-ready datasets
```

Getting Started

1. **Download Data**: All files available from competition platform
2. **Review Data Dictionary**: Comprehensive variable documentation provided separately

3. **Explore Starter Code**: Example data loading and EDA notebooks provided
4. **Validate Setup**: Ensure you can reproduce basic summary statistics

Support Resources

- **Data Dictionary**: Complete variable definitions and data sources
- **Starter Analysis**: Exploratory data analysis examples
- **Technical Forum**: Community discussion and Q&A
- **Office Hours**: Scheduled support sessions with organizers

Timeline

- **Competition Launch**: [Date]
- **Data Download Available**: [Date]
- **Clarification Deadline**: [Date] (last day for questions)
- **Submission Deadline**: [Date] 11:59 PM PST
- **Winners Announced**: [Date]

Ethical Considerations

Data Privacy

- All data uses publicly available aggregated statistics
- No individual-level information included
- Geographic data aggregated to protect privacy

Responsible Analysis

- Consider potential biases in data collection and representation
- Acknowledge limitations of findings
- Avoid stigmatizing language or conclusions
- Focus on systemic factors rather than individual characteristics

Policy Impact

- Recommendations should be evidence-based and actionable
- Consider unintended consequences of proposed interventions
- Acknowledge uncertainty and data limitations in policy suggestions

Questions and Support

For technical questions, data clarifications, or general support:

- **Competition Forum**: [Link to discussion platform]
- **Email Support**: [competition-support@email.com]
- **Office Hours**: [Schedule and connection details]

Good luck, and thank you for contributing to this important challenge!