

Time Challenge: Prediction Homelessness trends using Socioeconomic Datasets

1. Challenge Overview

Objective:

Participants will predict short-term homelessness trends in California by analyzing multiple public datasets covering demographics, hospital visits, and fiscal funding. The goal is to develop models and derive insights that could support policymaking and resource allocation.

Problem Statement:

Forecast homelessness levels in California counties using heterogeneous data sources, build accurate models, and generate actionable recommendations based on trends in hospital utilization, funding allocations, and vulnerable demographics.

2. Datasets Provided

A summary of each dataset used in the challenge:

1. California Homelessness Demographics Dataset

- Source: [California Homeless Data Explorer](#)
- Content: Demographic breakdowns (age, race, gender) by county and year.
- Coverage: California counties from 2017–2023.
- Key Columns: Calendar Year, Location, Homeless Population by Demographics.

2. Hospital Encounters for Homeless Patients (CA)

- Source: [California Health & Human Services Open Data](#)
- Content: Hospital visits involving homeless individuals, categorized by age, gender, and race.
- Coverage: 2019–2020
- Key Columns: Facility Name, Hospital County, Demographic Group, Encounters.

3. Fiscal Information for National Profile (NCHE)

- Source: [NCHE Consolidated Profiles](#)
- Content: Funding amounts related to homeless education programs by year and state.
- Coverage: 2015–2020
- Key Columns: State, Yearly Funding, Funding Change, Funding Per Capita.

3. Challenge Structure

Duration: 3–4 hours

Team Size: 1–4 members

Tools Allowed: Python, Jupyter Notebooks, Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn, Plotly, etc.

Submission: Final .ipynb or .py notebook with all implementations and their reasons like why they are performing a step and what they got to know after result generation.

4. Instructions for Participants

- Perform independent preprocessing for each dataset.
- Conduct visual exploratory data analysis.
- Merge/align datasets thoughtfully for modeling.
- Train baseline models, then improve through feature engineering & tuning.
- Forecast future trends and simulate policy impact.
- Derive insights and give policy recommendations based on ML findings.

5. Challenge Question Set

◆ 1. Data Understanding & Preparation (6 Questions)

- Q1. Identify key variables in each dataset and describe their data types, granularity (county/year/etc.), and any categorical structure.
- Q2. Are there any missing or inconsistent values? Quantify them and suggest appropriate treatment methods per dataset.
- Q3. How is the “Location” column structured across the datasets? Normalize these entries to allow future merging.
- Q4. What transformations are necessary to ensure numeric fields (like counts, funding) are model-ready?
- Q5. Filter the datasets to only include relevant years with sufficient coverage across all datasets. How many valid years remain?
- Q6. Reformat any date or year columns to consistent formats and create usable time-based features.
-

◆ 2. Exploratory Data Analysis (6 Questions)

- Q7. What patterns do you observe in homelessness counts by race, age, and gender across California counties?
- Q8. Which counties have the highest hospital encounters relative to total homeless population?
- Q9. Explore funding trends per state in the fiscal dataset. Are some states consistently increasing homelessness-related budgets?
- Q10. Visualize county-wise distribution of homelessness. Are there spatial patterns (urban vs. rural)?
- Q11. Compute pairwise correlations within each dataset. Which variables appear to be most associated with homeless counts?
- Q12. Are there lagging or leading relationships (time shift) between fiscal funding and homelessness levels?

- ◆ **3. Feature Engineering (6 Questions)**

Q13. Create proportional or normalized features such as: homeless population per 10,000 residents or hospital encounters per homeless individual.

Q14. Derive change-over-time features such as YOY change in funding or encounter counts.

Q15. Engineer rolling averages or lag features to capture temporal movement in homelessness counts.

Q16. Encode categorical variables (like gender or race) appropriately for modeling.

Q17. Create a merged dataset by aligning similar counties across homelessness, hospital, and fiscal data using normalized identifiers.

Q18. Can you derive a “vulnerability index” for each county based on homelessness counts and hospital usage?

- ◆ **4. Baseline Modeling (4 Questions)**

Q19. Train a simple regression model (Linear/Ridge) to predict homeless count based on demographic features alone.

Q20. Repeat the above model using hospital data only. Which set of predictors is more effective?

Q21. Evaluate both models using MAE and RMSE. Which baseline model performs best?

Q22. Analyze residuals from your models. Are there counties or subgroups where predictions consistently fail?

- ◆ **5. Model Optimization & Advanced ML (6 Questions)**

Q23. Train a tree-based model (Random Forest or Gradient Boosting) using all datasets individually. Compare results.

Q24. Tune hyperparameters using grid/random search. Which configurations yielded the best model?

Q25. Extract and rank feature importances. Do the most important features align with your expectations?

Q26. Can a multi-source model (merged dataset) improve predictions? Combine all features across datasets and re-train.

Q27. Apply dimensionality reduction (e.g., PCA) to reduce multicollinearity. Does it improve results?

Q28. Evaluate your best model on a holdout set. Are errors evenly distributed across counties or demographics?

◆ **6. Bonus Forecasting & Insight Generation (4 Questions)**

Q29. Forecast homelessness for the next 12 months using your best model. How confident are you in this projection?

Q30. Visualize your forecast vs. actual values (where possible). Do you observe under/overestimation patterns?

Q31. Recommend three counties for increased resource allocation. Justify your picks based on model + analysis.

Q32. If you were to deploy this model in production, what external features (weather, housing cost) might improve it?

 Note: The *Integrated Forecasting* section is optional. However, completing it will add up to 150 bonus points to your total score. If you're confident in your core results and have time left, we encourage you to explore this!

◆ **7. Policy Framing & Reporting (3 Questions)**

Q33. What demographic groups consistently experience the highest homelessness rates? Provide data-backed evidence.

Q34. Based on all your work, what policy recommendations can you make to reduce homelessness in California?

Q35. Summarize your project's key takeaways in a one-page executive summary. Highlight models, insights, and policy impact.

6. Grading Rubric (Out of 1000 Points)

Category	Points	Evaluation Criteria
 Data Understanding & Preparation	100	Clean and transform the datasets, handle missing data, format columns, ensure consistency across datasets (e.g., locations, timeframes).
 Exploratory Data Analysis (EDA)	150	Use visualizations and statistics to uncover trends, identify relationships, generate meaningful insights from each dataset.
 Feature Engineering	150	Create useful features (lag, rolling averages, ratios, encodings), ensure models can utilize new features effectively, explain feature choices clearly.
 Baseline Modeling	100	Train simple regression models and compare their performance, interpret metrics, explain modeling decisions.
 Model Optimization & Advanced ML	200	Tune hyperparameters, use tree-based models (e.g., Random Forest, XGBoost), combine datasets, evaluate performance on holdout/test sets.
 Bonus: Forecasting & Strategic Insights	100	Use models to make future predictions and visualize them. Identify high-need counties or trends. Propose allocation or strategy shifts.
 Policy Framing & Executive Summary	100	Communicate findings clearly. Present recommendations for policymakers using data-driven arguments and visual storytelling.
 Timeliness & Clarity of Submission	100	Submitting on time, with clear code, comments, structured notebook format, and proper explanations.

7. Expected Deliverables

- A clean, logically structured notebook with:
 - Preprocessing steps for each dataset
 - Visual EDA with commentary
 - Feature engineering and modeling process
 - Final predictions/forecasts
 - Summary of findings & recommendations

8. Scoring Notes for Judges

- Use partial scoring if a task is only partially completed.
- Prioritize code explainability and visualization insight.
- Award bonus points for novel visualizations, smooth dashboards, or insightful forecasts.

9. Key Evaluation Guidelines

- **Is the problem clearly understood and addressed?**
- **Are the results data-driven and reproducible?**
- **How well are the insights communicated?**
- **Are datasets handled with care and accuracy?**
- **Is the forecasting method logical and grounded?**