

288R Project Progress Report #1 (Group 5)

Predicting Poverty Susceptibility Using U.S. Census Microdata

Project Group #: 5

Authors: Marianne Sawires, Ingrid Altamirano, Vanessa Scott

Emails: masawire@ucsd.edu, ialtamir@ucsd.edu, vascott@ucsd.edu

Github Repo: <https://github.com/ingridalt/DSC-288R-CAPSTONE>

Background

Poverty is a multidimensional public health challenge influenced by interconnected demographic, employment, health, and structural factors. Leveraging large-scale census data with machine learning offers a scalable way to identify individuals at elevated poverty risk and uncover the key predictors driving economic vulnerability, creating opportunities to inform public support strategies and guide program development.

Dataset name and citation/link

American Community Survey 1-Year Estimates Public Use Microdata Sample

Census Bureau's FTP Server Site:

<https://www2.census.gov/programs-surveys/acs/data/pums/>

Census Bureau's MicroData API Details and Sample Variable Information:

https://www.census.gov/data/developers/data-sets/census-microdata-api.ACS_1-Year_PUMS.html#list-tab-78025481

Data Pipeline

To ingest the data, inside the [Load_Raw_Data.ipynb](#) notebook in function 'download_acs_1year_persons_data' we are hitting the direct Census Bureau's FTP Server to retrieve 1-year PUMS persons level zipped CSV files for the years 2018-2024, retrieving both our train and testing dataset. Next we are creating a local folder structure at 'data_persons_ca_1_yr/{year}' to store the results by year. Next we are opening the zip downloaded file and extracting all the contents into the designated year specific folder. Later, in the function 'build_master_common_columns' we are collecting all CSV

files, and generating a master CSV file that intersects column names across years. As we are generating this master file, we are adding the column 'year' along with adding only common columns shared across all 6 years. This generates the final 'persons_master.csv' file inclusive of all 275 columns. Handling of null values, missing values, feature augmentation and normalization was completed and discussed in the feature engineering component of this report.

EDA Description

We completed a very thorough exploratory data analysis of our California census data. It combines graphical and nongraphical, univariate, bivariate and multivariate analyses to assess data quality, understand variable distributions, identify anomalies, and surface relationships relevant for modeling poverty susceptibility with over 40 visualizations produced in our file [EDA_master](#). *Visualizations can be observed at the end of this document with their respective variable type.*

Univariate (Reference Figures 1- 3):

Univariate analysis was used to examine distributions of key variables such as income to poverty ratio (POVPIP), which is going to be our label in supervised learning models, employment status, educational attainment, age, and household composition. This included summary statistics and frequency tables to assess skewness, sparsity, and class imbalance. A null value audit across all 275 columns was completed, identifying columns with >60% missing values for removal. We created a [data dictionary](#), constructed from fetching Census API metadata and mapping all variable codes to human-readable labels. We identified and removed allocation flags (F-prefixed columns), person weight columns (PWG-prefixed), and redundant income variables that would leak target information.

Bivariate (Reference Figures 4 - 19):

Bivariate analysis was conducted to examine the relationship between the target variable, income-to-poverty ratio (POVPIP), and individual candidate predictors. This analysis was central to identifying features with meaningful associations to poverty susceptibility and to informing downstream feature selection. Pairwise correlations with POVPIP were computed to assess the direction and strength of linear relationships. Variables with consistently non-trivial correlations were flagged as high-priority candidates, while highly collinear features were noted for potential exclusion or consolidation. POVPIP distributions were compared across category levels using grouped means, medians, and frequency-based summaries. This enabled identification

of structural differences in poverty susceptibility across employment status, education levels, marital status, and household characteristics.

Multivariate (Reference Figures 20 - 25):

Multivariate analysis was used to examine relationships among multiple predictors and assess feature interdependencies. Correlation matrices were analyzed to identify highly correlated variables and potential multicollinearity, informing feature filtering and regularization strategies (flagging pairs with $|r| > 0.80$ for recommended features). We also completed Population Stability Index (PSI) drift analysis comparing feature distributions between base year and subsequent years to assess temporal stability. This analysis helped distinguish features with independent signals from redundant predictors prior to model development. A final recommended feature set was selected with 23 features combining correlation strength, domain relevance, and multicollinearity constraints.

EDA Summary:

In summary, the exploratory data analysis provided a structured foundation for feature selection by linking data quality assessment, univariate and bivariate relationships with POVPIP, and multivariate dependency analysis. Through iterative filtering based on missingness, variance, correlation strength, multicollinearity constraints, and temporal stability, we identified a [final recommended set of 23 features](#). These features span key domains including demographics, education, employment, income, household composition, health insurance coverage, and citizenship, and collectively balance predictive signal, interpretability, and stability over time. This feature set will serve as the input for baseline and advanced models in subsequent phases of the project.

Feature Engineering

Our feature selection and engineering process was guided both by prior literature and by findings from our exploratory data analysis. We follow an established approach for handling large census-like datasets that reduces high-dimensional inputs through correlation with the target variable, covariance checks among predictors, removal of highly null variables, and assessment of overall relevance, similar to the framework described by Li et al. (2022). During EDA, we identified several variables with more than 60% missing values and removed them from consideration. Although some of these variables may have been substantively interesting, their high degree of missingness reduced their reliability for model input and risked introducing noise.

An additional data limitation observed during EDA was the absence of explicit year identifiers across raw files. Because temporal context is central to understanding poverty trends, we engineered a year variable to label when each record was collected. Geographic information was also originally coded numerically, limiting interpretability. Based on EDA showing regional variation patterns, we transformed these codes into categorical regional labels to enable meaningful topographical distinctions across California (e.g., Bay Area vs. Inland Empire). Finally, the POVPIP variable was highly skewed in its continuous form, so we recoded it into interpretable outcome classes representing poverty risk ranges (e.g., at-risk vs. stable). This transformation supports a clearer and more policy-relevant classification framework.

Our feature inclusion also aligns with prior poverty prediction research using machine learning. Studies such as Okolie et al. (2025) demonstrate that demographic, employment, education, and access-related indicators are strong predictors of socioeconomic vulnerability. Consistent with this literature, we retain variables capturing educational attainment (SCHL), employment status and work intensity (ESR, WRK, WKL, WKHP, OCCP), health insurance coverage (HICOV, PRIVCOV, PUBCOV), demographic characteristics (AGEP, SEX, MAR, MSP, CIT, NATIVITY), race/ethnicity (RAC1P, HISP), and disability status (DIS).

We intentionally focused on identifying structural risk factors associated with poverty susceptibility rather than direct income proxies. All income variables (PINCP, PERNP, WAGP, RETP, SSP, SSIP, PAP, etc.) were excluded to prevent data leakage, as POVPIP is itself derived from income. Instead, our goal is to allow the model to learn patterns of vulnerability based on demographic, employment, education, health access, and disability-related characteristics.

To reduce multicollinearity, we prioritized summary recodes over component variables (e.g., DIS instead of multiple disability indicators, HICOV/PRIVCOV/PUBCOV instead of HINS1–7, ESR instead of NWLK/NWLA/NWAB). We also simplified race variables by retaining RAC1P while dropping redundant binary race flags.

What model(s) are you using? Justify

Modeling:

Our modeling roadmap closely mimics a similar project led at the Wentworth Institute of Technology in Boston where they assessed a similar dataset and predicted food insecurity in low income communities in the U.S.. This group used a Logistic Regression, Random Forest and XGBoost methods to assess their area of interest (Okolie et. al., 2025). We will be mirroring this approach and utilizing an initial model of multivariate logistic regression, random forest and XGBoost algorithm to assess predicting levels of poverty. The first model will be used as our baseline assessment, the random forest model will likely surpass the regression model and XGBoost offers a powerful approach in assessing poverty by penalizing and conducting regularization as

the model progresses as seen (Hu et. al., 2022). In addition we are interested in possibly exploring FT Transformer, an architecture that works for tabular data. This model works by embedding entire features, where the embeddings are then processed by the Transformer module and the final representation of the [CLS] token is used for prediction. (Shavitt, Yoav, et al. 2021). Given the complexity of our features, the FT transformer can learn more complex relationships that algorithms like XG Boost may not capture.

Baseline Model:

We trained a multinomial logistic regression baseline with L2 regularization on class-balanced training data, with the majority class downsampled to match the at-risk population. Evaluation on the 2024 test set yielded an overall accuracy of 70% and a weighted F1 score of 0.67, but a low macro F1 score of 0.29. Performance was strong for the majority class but weak for minority classes, highlighting the need for more expressive models (e.g. Random Forest, XGBoost) and improved class-balancing strategies. Coefficient analysis identified hours worked (WKHP), education tier, and employment status as the strongest predictors across classes. We noticed the baseline model struggled to find relationships for Class 2 which is near poverty, where precision, recall and F1-score came out to 0.0. This is something we must explore in our more advanced models.

Progress Report Details on your progress

Significant effort has been dedicated to dataset collection and preparation, exploratory data analysis, feature engineering and preprocessing, and even early training of a baseline model.

We downloaded ACS 1-Year PUMS person-level microdata for California across six years (2018, 2019, 2021–2024), totaling approximately 2.3 million records and 275 columns. The year 2020 was excluded due to COVID-related anomalies. We built an automated data ingestion pipeline ([1 Load Raw Data.ipynb](#)) to download, concatenate, and store multi-year Census data, and tested a DuckDB-based EDA workflow to support efficient analysis at scale without memory constraints. In addition, we constructed a complete data dictionary by programmatically retrieving variable metadata from the Census API and mapping all coded column names to human-readable labels.

EDA was conducted across multiple notebooks and consolidated into a master analysis ([EDA master.ipynb](#)) that integrates work from all team members. We defined a 4-class target variable derived from POVPIP for multi-class classification: Stable (>200), Low

Risk (150–200), Moderate Risk (100–150), and Deep Poverty (<100) to be used for multi-class classification. Univariate, bivariate, and multivariate analyses were performed across more than 12 demographic dimensions, resulting in 40+ visualizations and identification of high-signal predictors. A data cleaning pipeline reduced the feature space from 275 raw variables to 115 candidate features by removing allocation flags, person weights, identifiers, high-missingness columns (>60%), and income variables that would leak target information. This was further narrowed to a final recommended set of 23 features using correlation analysis and multicollinearity checks. PSI (Population Stability Index) drift analysis confirmed feature stability across years, supporting the suitability of 2024 as a test set.

Preprocessing for the baseline model was implemented in [3 Preprocessing Baseline.ipynb](#). This included binarizing categorical indicators (insurance, disability, sex, marital status), recoding education into four tiers, grouping occupations into top-10 categories, creating derived indicators (citizenship status, employment binary, born-in-California, mobility), handling missing values, applying one-hot encoding, and scaling continuous features. Data were split, with 2018–2023 used for training and 2024 held out for testing to better reflect real-world deployment conditions. Finally, we trained a class-balanced multinomial logistic regression baseline with L2 regularization.

Team member contribution

Vanessa – Currently, I am working on the completion of a portion of this milestone report and cleaning up our exploratory data analysis as a way to continue having a clear road map on our next steps. Previously, I worked on preprocessing categorical vs. numerical variables in the dataset. As any additional preprocessing items are identified, I will/can assist in that and will begin on model building.

Marianne – Assisted in organizing the GitHub repository and contributed to the abstract and progress report writeups. Conducted exploratory data analysis focused on missingness, distributions, variance, and correlations to support candidate feature identification. Plans to contribute to model development in the next phase.

Ingrid – Currently, I worked on building the multivariate logistic regression baseline model. This included preprocessing data and feature engineering for the baseline model and generating the dataset with the final interested features. Previously I assisted with setting up the GitHub Repository, initial data ingestion and pre-processing. I also completed the EDA for regional poverty trends.

Risks and Mitigation

Currently, we have assessed project risks at length and have not identified any barriers that would prevent successful completion of the Capstone Project. However, we have identified potential weak areas in our approach for which we may seek guidance from the teaching assistants and professor. One area of uncertainty is whether excluding variables with more than 60% missing values during preprocessing is sufficient, or whether this threshold could lead to the removal of potentially important features. Additionally, the POVPIP outcome variable exhibits substantial class imbalance, as the majority of Californians do not fall within poverty risk categories. We will need to apply strategies to address this imbalance and plan to explore multiple approaches for each model.

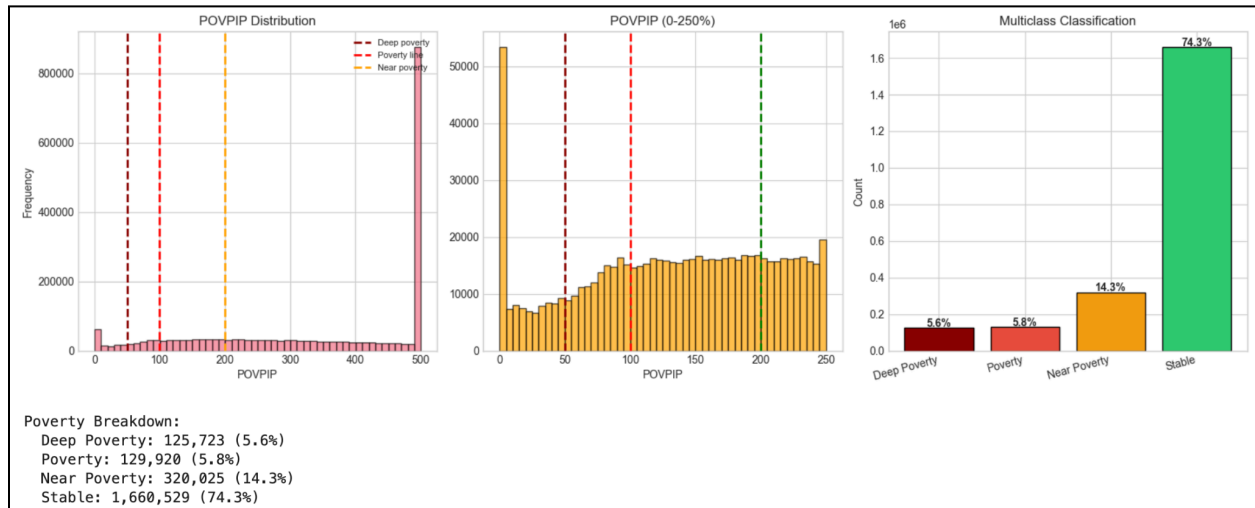
We also recognize the potential for bias introduced through our feature selection process. By excluding variables to reduce dataset dimensionality, it is possible that some meaningful predictors with high missingness were removed. Furthermore, because the model predicts poverty risk, variables related to race, gender, ethnicity, and other demographic characteristics may introduce or amplify bias. Our approach is to mitigate these risks by evaluating model performance across demographic subgroups and carefully assessing feature importance to ensure that sensitive variables are not disproportionately driving predictions. Overall, for any problem we encounter we have agreed as a team to face each issue together and seek help when necessary.

References

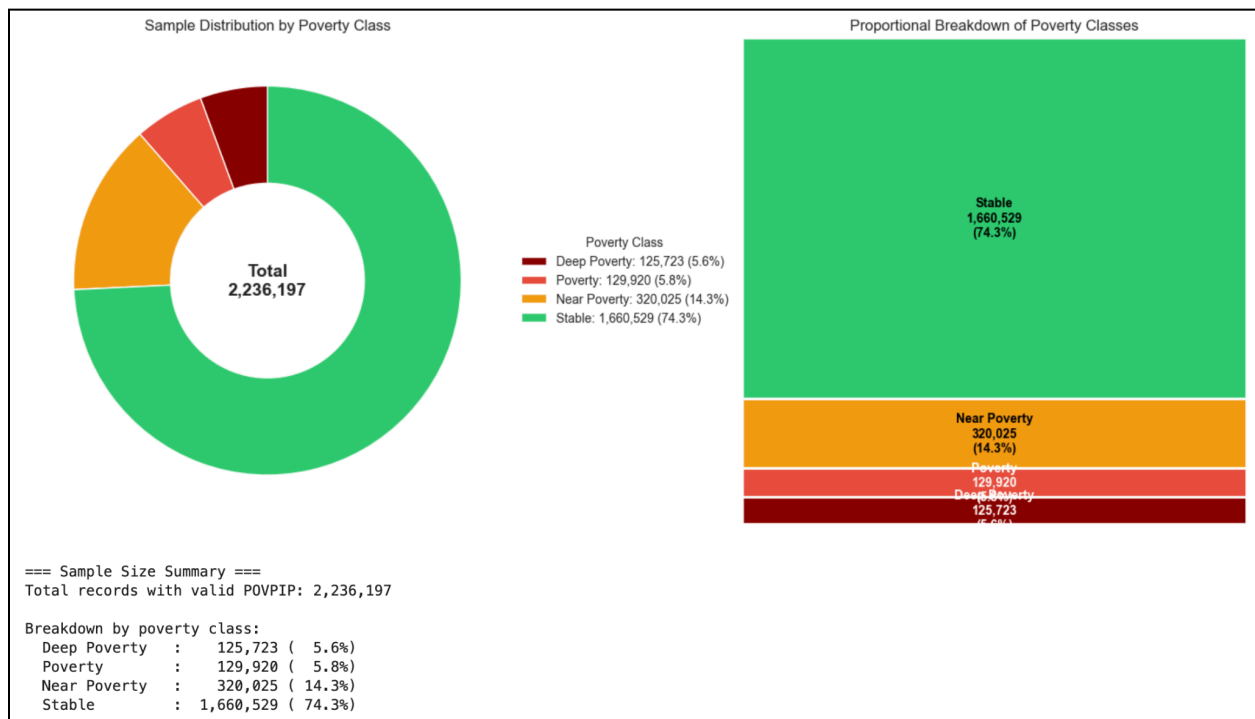
- Hu, Y., Yu, Z., & Chen, X. (2022). Integrating predictive analytics with social vulnerability assessment for food insecurity research. *Computers, Environment and Urban Systems*, 94, 101787. <https://doi.org/10.1016/j.compenvurbsys.2022.101787>
- Corral, P., Henderson, H., & Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 172, 103377. <https://doi.org/10.1016/j.jdeveco.2024.103377>
- Li, Q., Yu, S., Échevin, D., & Fan, M. (2022). *Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan*. *Socio-Economic Planning Sciences*, 81, 101195. <https://doi.org/10.1016/j.seps.2021.101195>
- Okolie, A., Lawal, O., Alumona, P., Obunadike, C., Ikhifa, M., & Akwabeng, P. (2025). Predicting food insecurity across U.S. census tracts: A machine learning analysis using the USDA Food Access Research Atlas. *International Journal of Science and Research Archive*, 17, 1156–1172. <https://doi.org/10.30574/ijrsra.2025.17.2.3156>
- Burger, R., & van der Laan, J. (2021). *Predicting transitions in and out of poverty using machine learning*. Statistics Canada. <https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2021001/article/00003-eng.pdf>

Figures

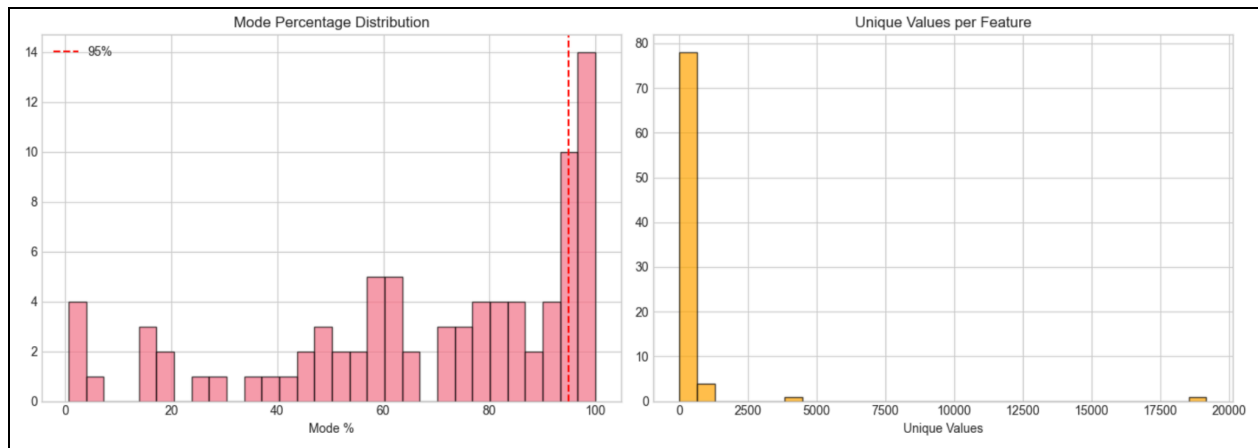
Univariate Figures (1-3)



^Figure 1: Distribution analysis of the target variable POVPIP, including histograms and bar plots of the 4-class poverty risk score (Stable, Near poverty, Poverty, Deep Poverty).

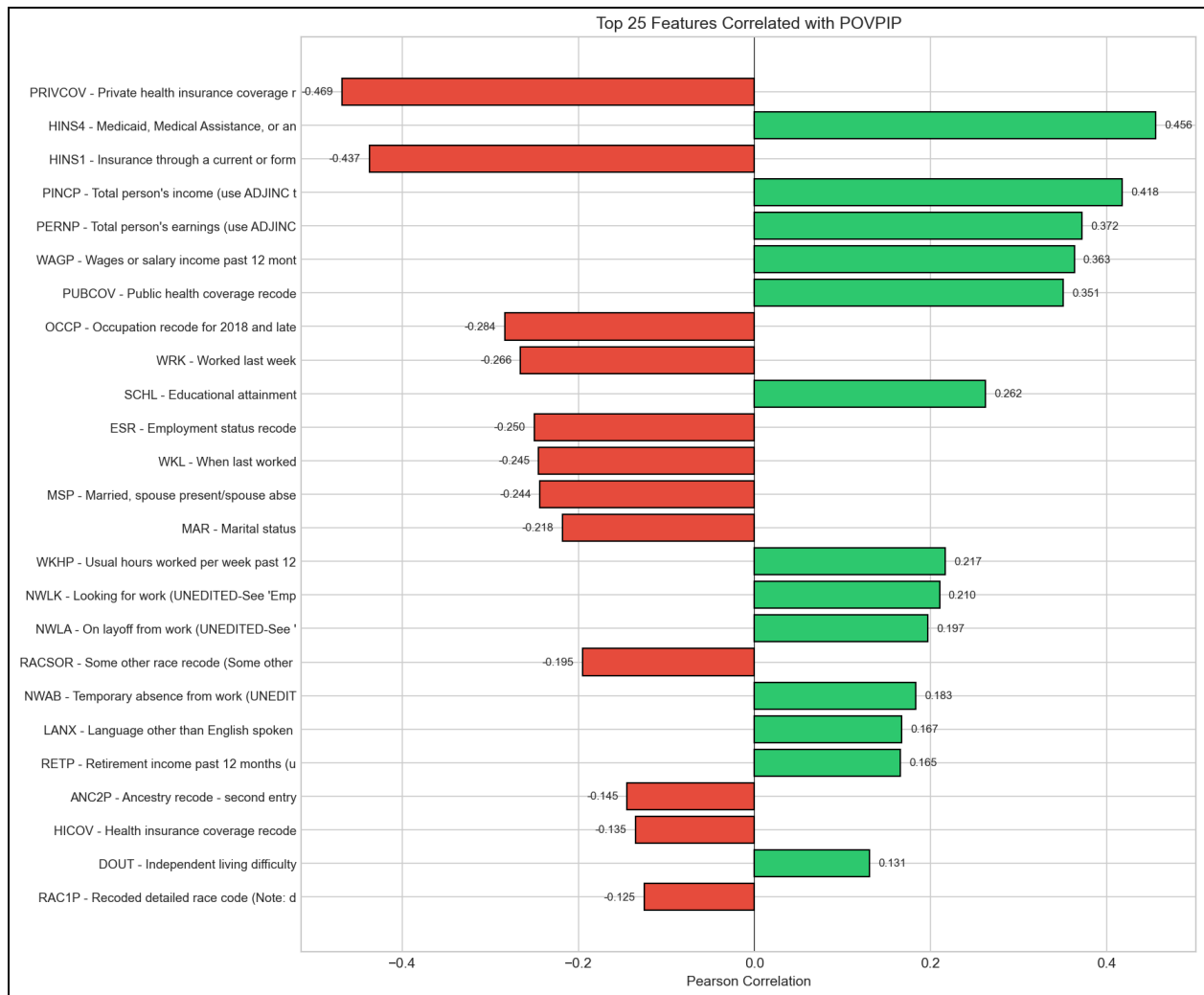


^Figure 2: Sample size overview via donut chart and stacked bar chart of label POVPIP.

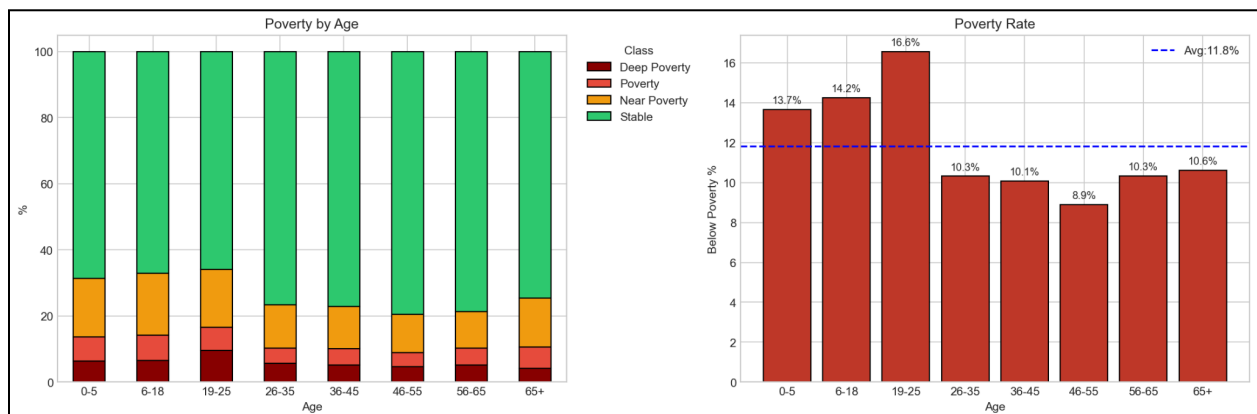


^Figure 3: Variance analysis across all features examining mode percentages and unique value counts.

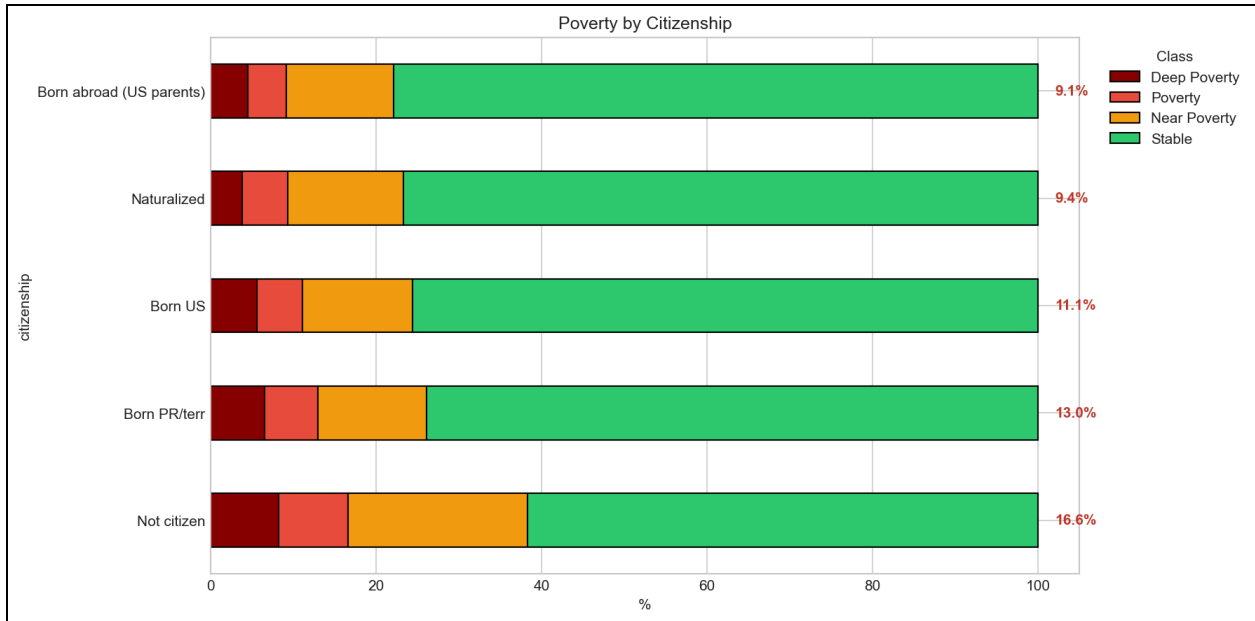
Bivariate Figures 4 - 19



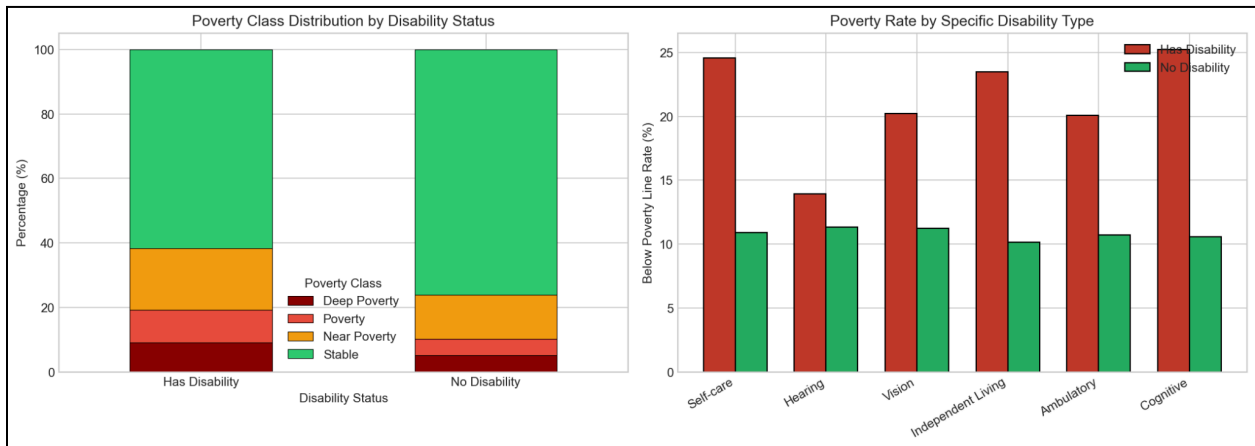
^Figure 4: Pearson correlation bar chart of the top 25 features correlated with POVPI.



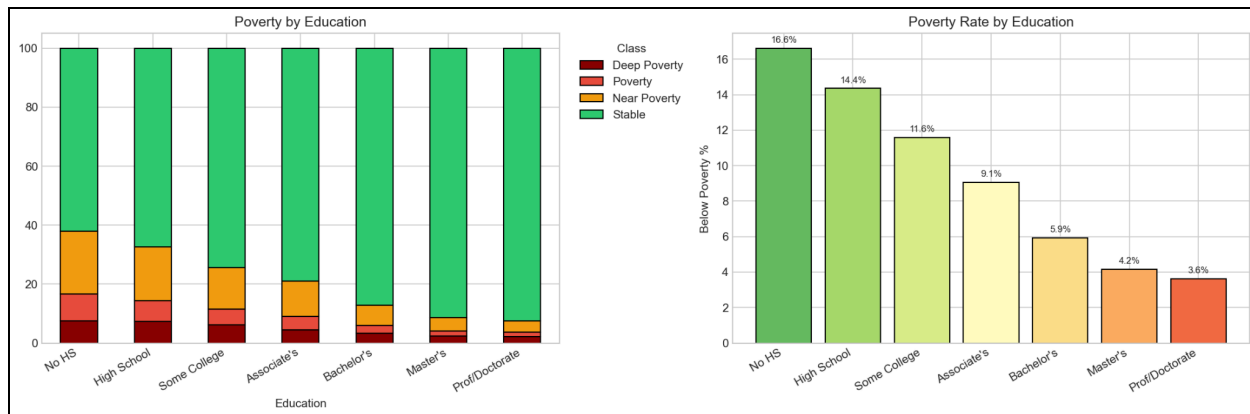
^Figure 5: Poverty class distribution by age.



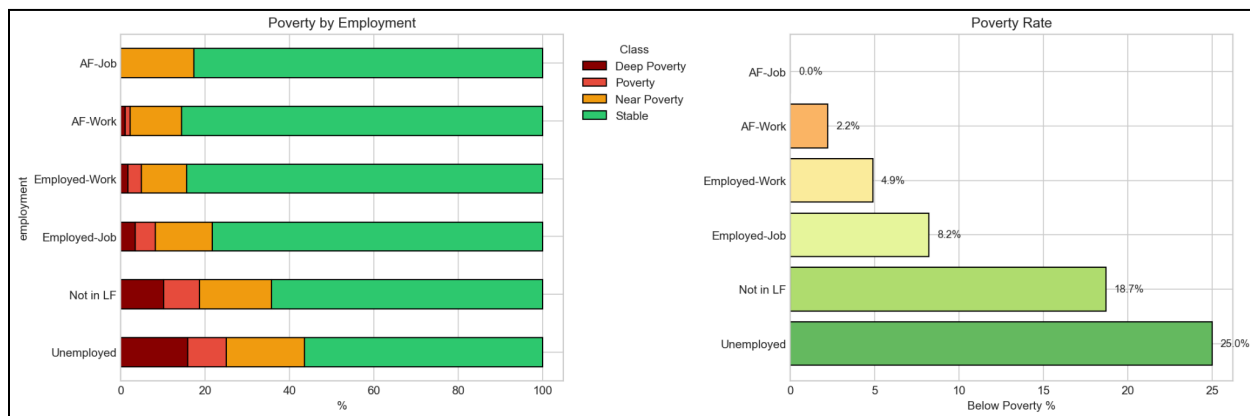
^Figure 6: Poverty class distribution by citizenship.



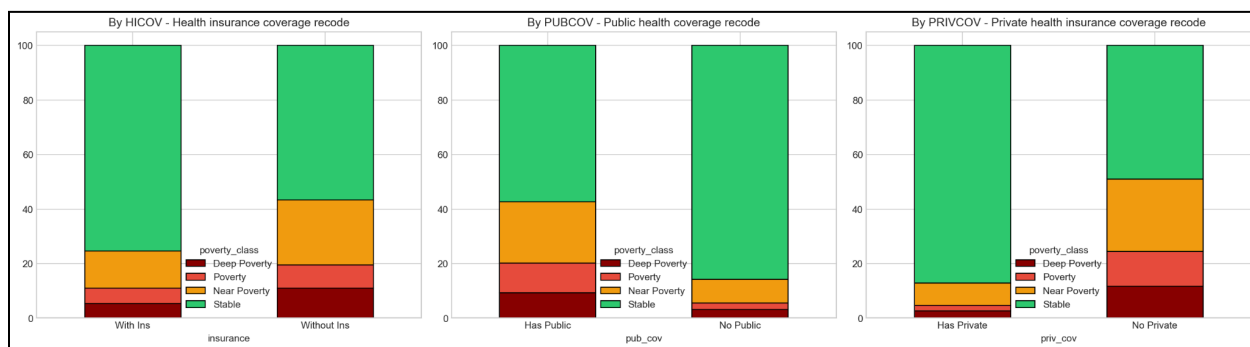
^Figure 7: Poverty class distribution by disability.



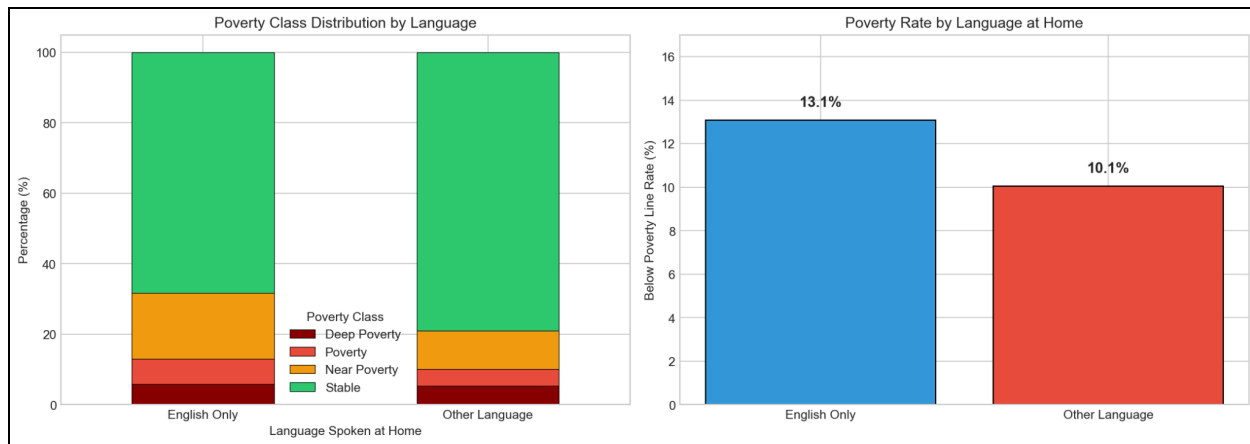
^Figure 8: Poverty class distribution by education.



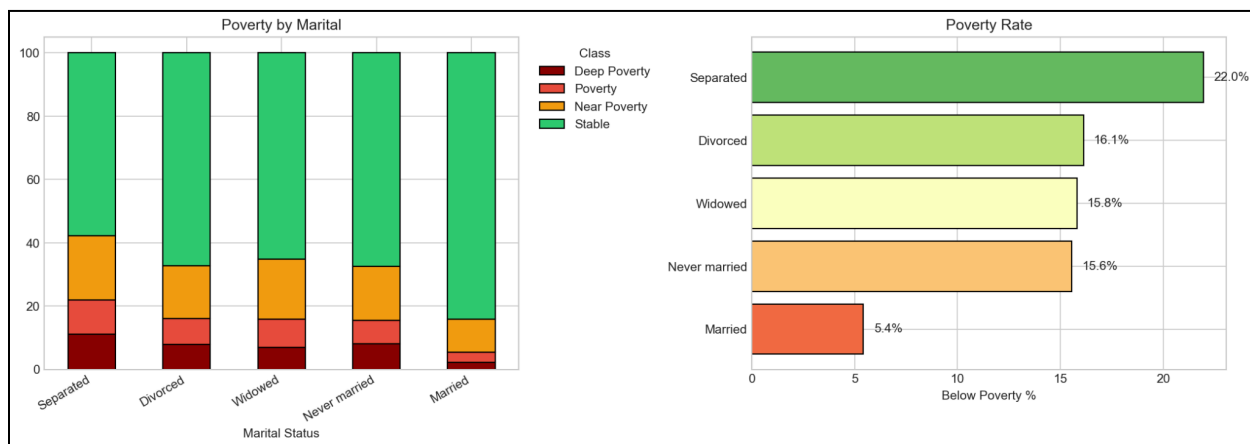
^Figure 9: Poverty class distribution by employment.



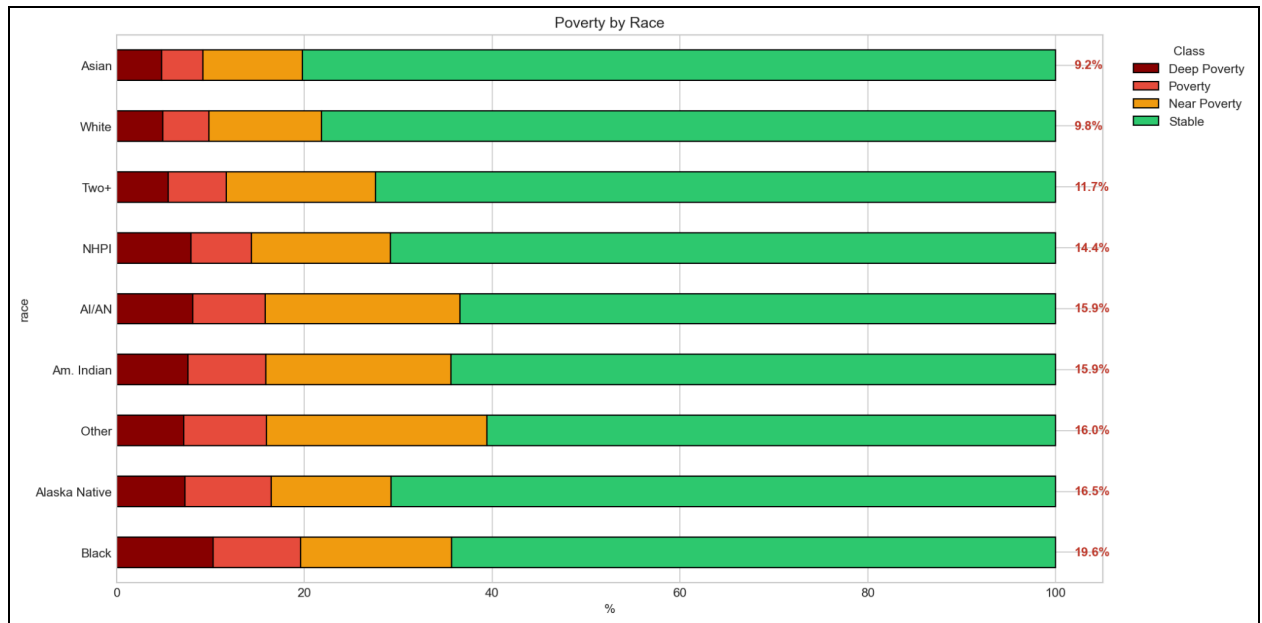
^Figure 10: Poverty class distribution by health insurance.



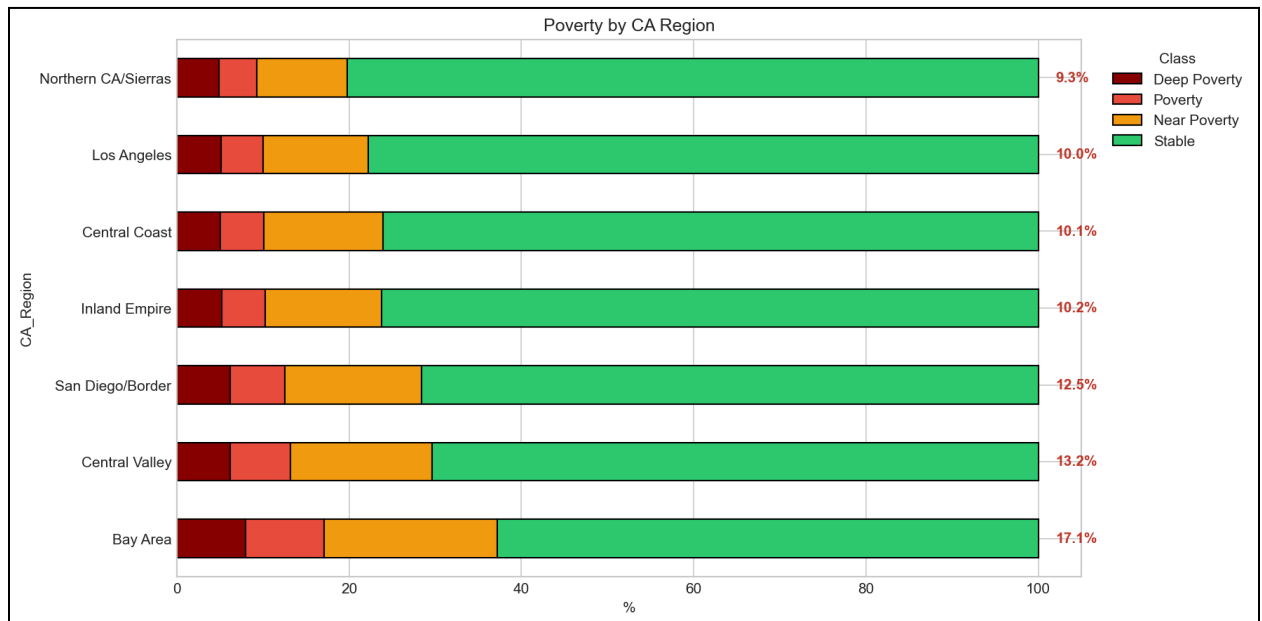
^Figure 11: Poverty class distribution by language spoken at home.



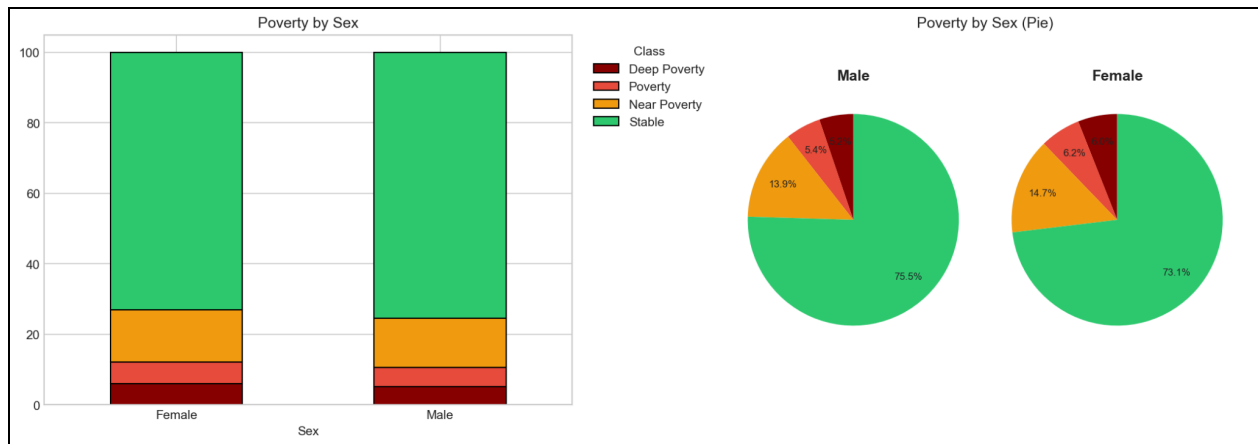
^Figure 12: Poverty class distribution by marital status.



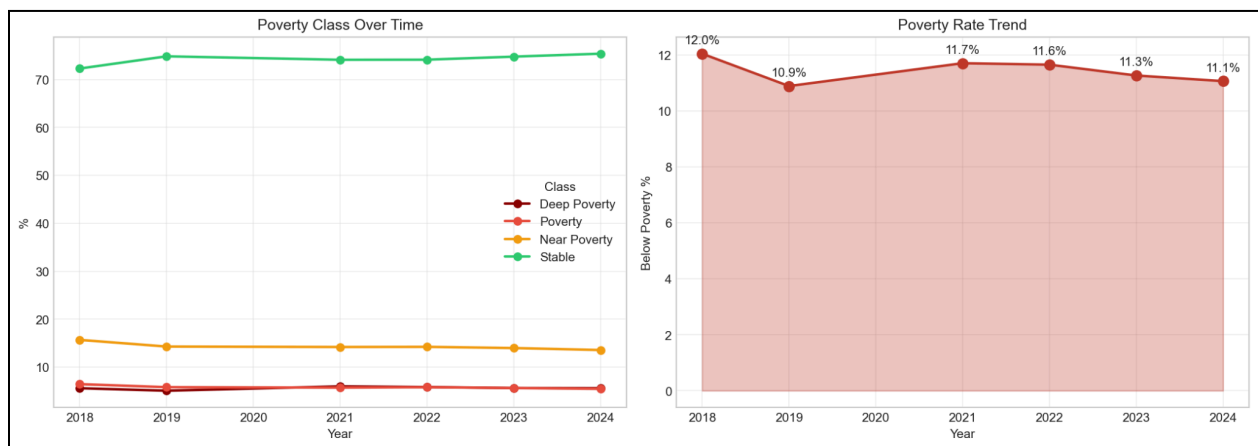
^Figure 13: Poverty class distribution by race.



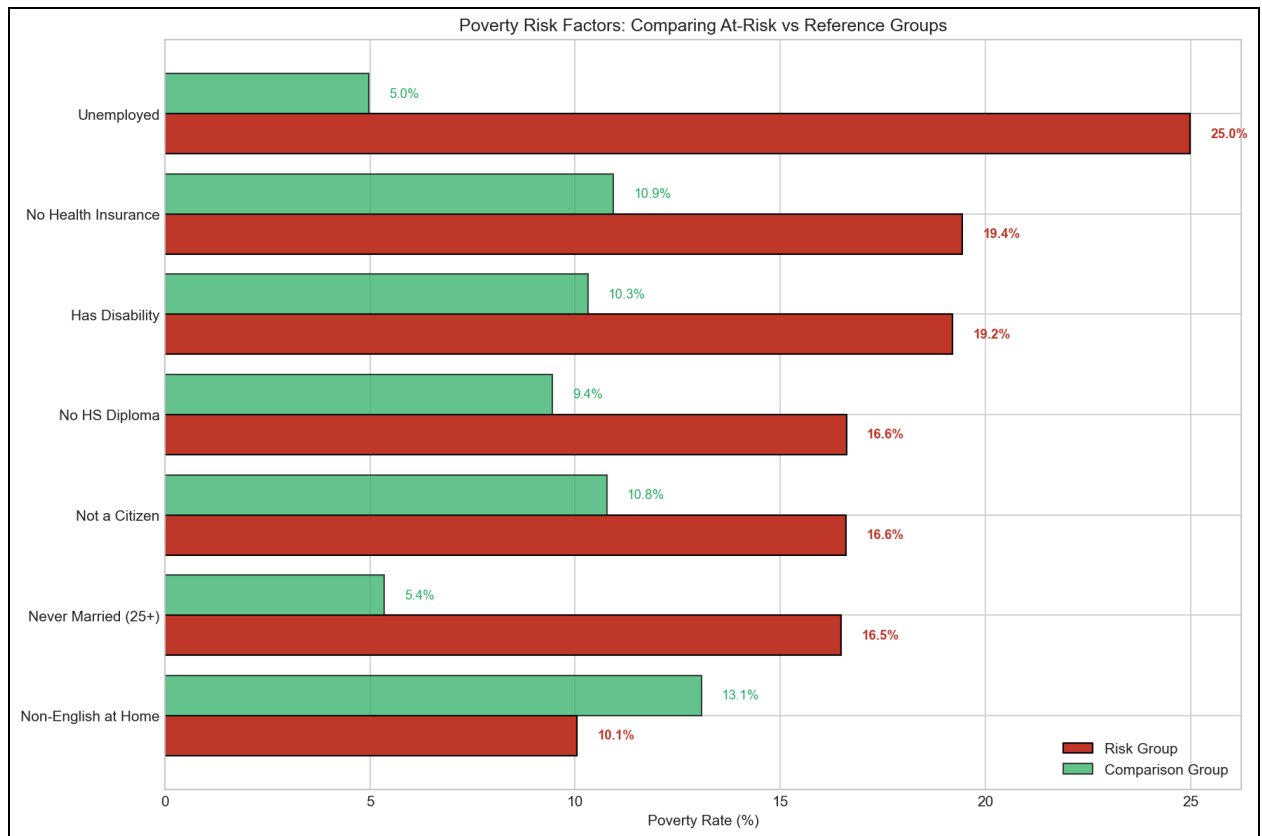
^Figure 14: Poverty class distribution by region.



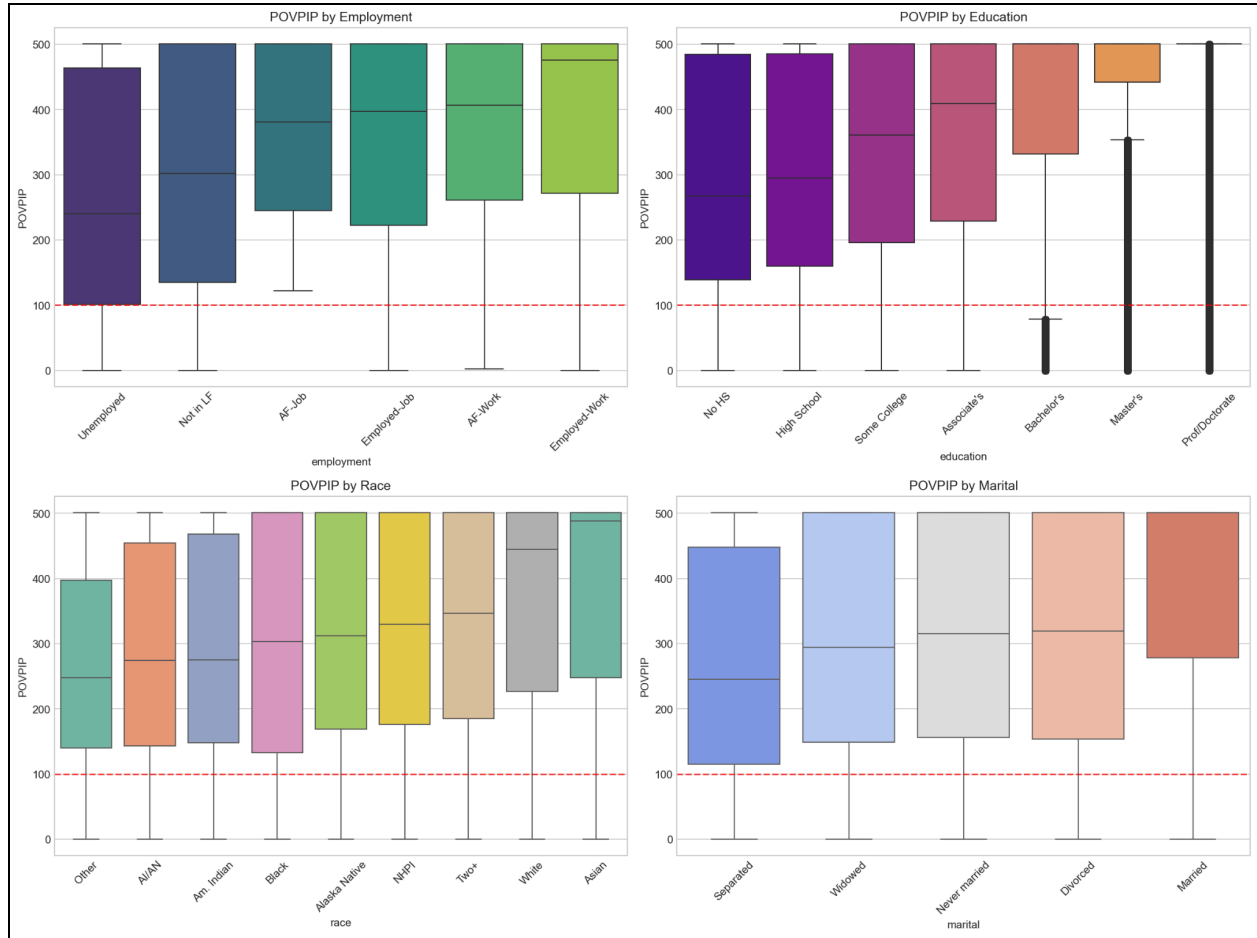
^Figure 15: Poverty class distribution by sex.



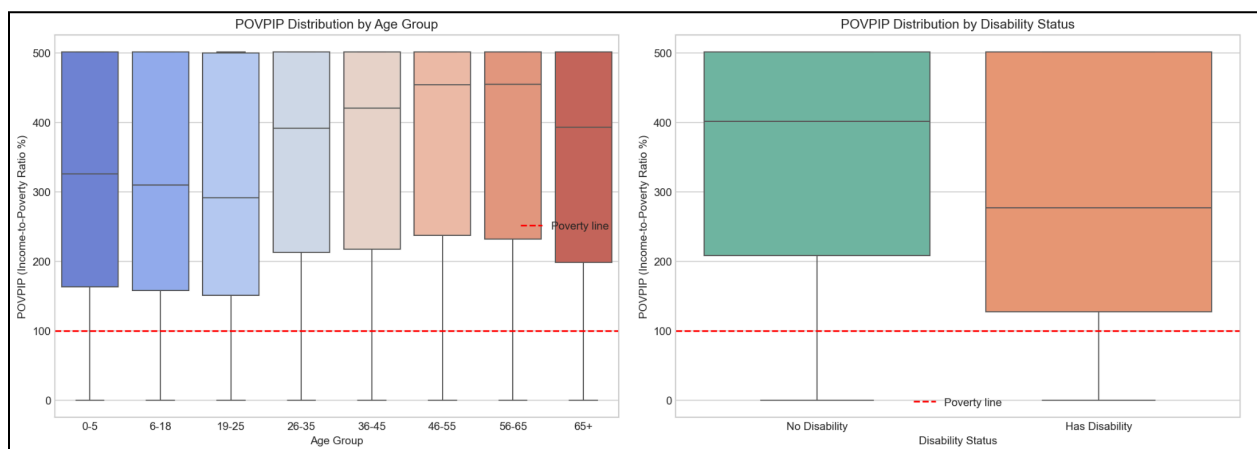
^Figure 16: Poverty class distribution over time (2020 anomaly excluded).



^Figure 17: Poverty risk factory.

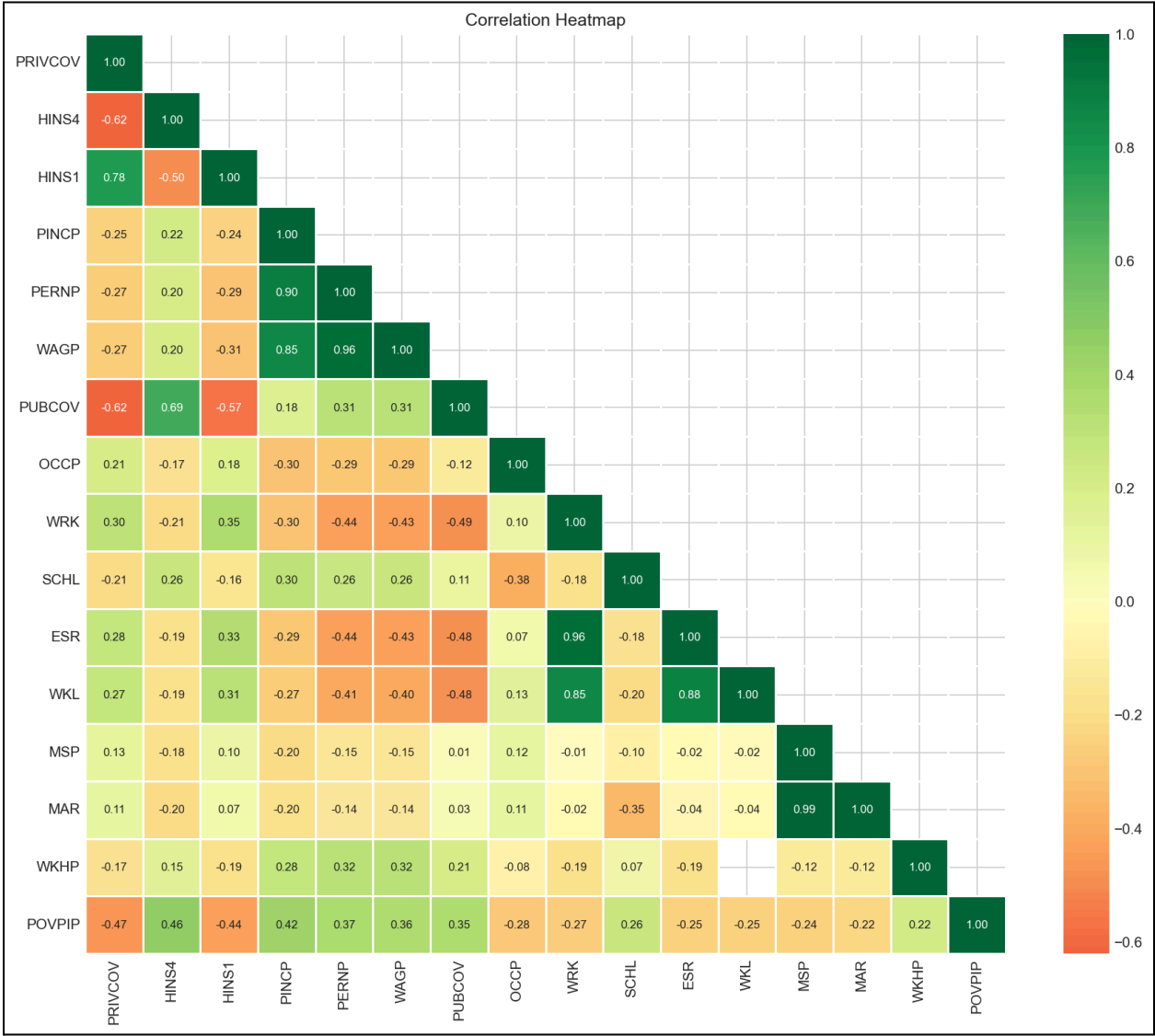


^Figure 18: Box plots of POVPIP distributions by employment, education, race, marital status.

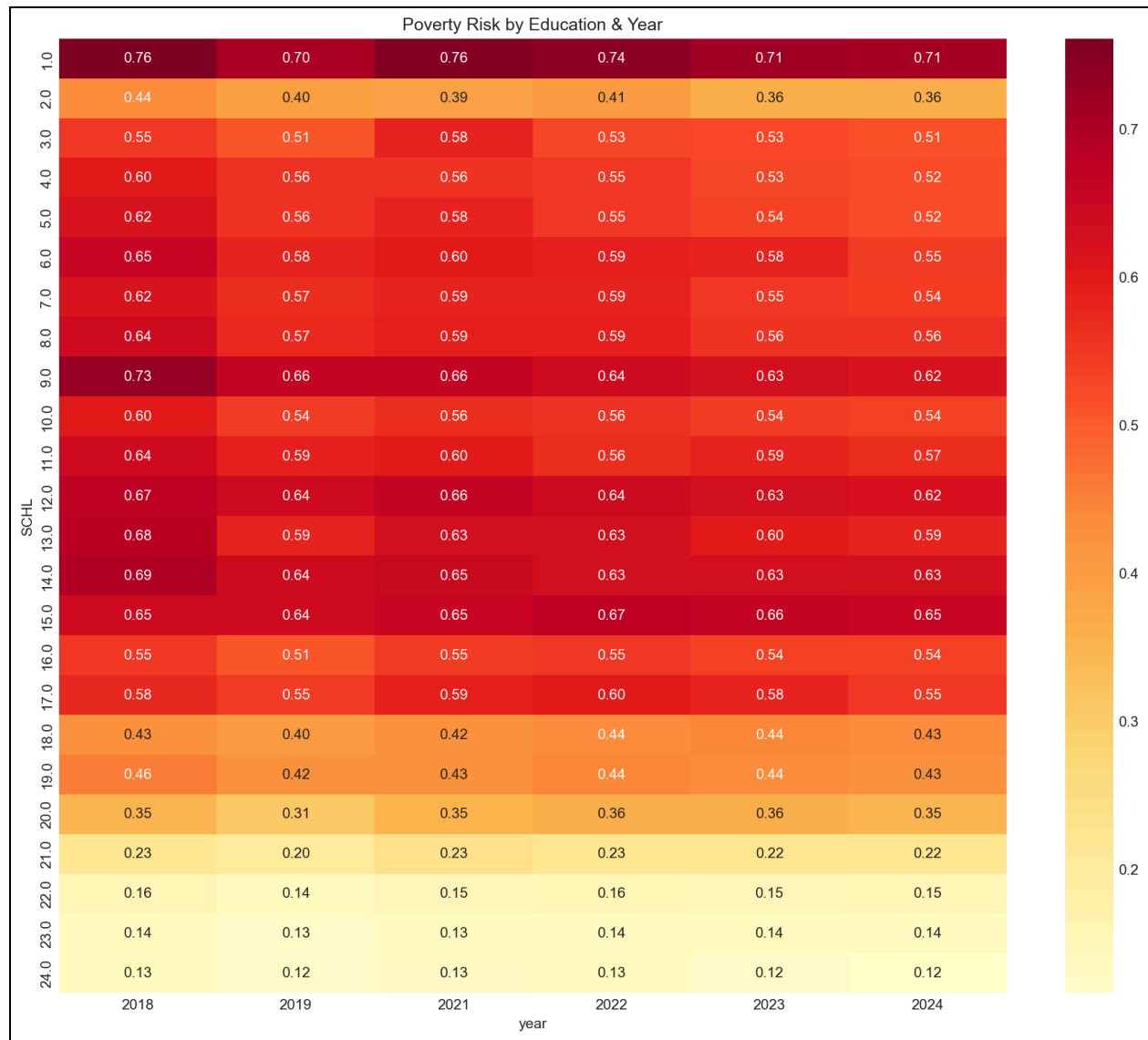


^Figure 19: Box plots of POVPIP distributions by age group, disability status.

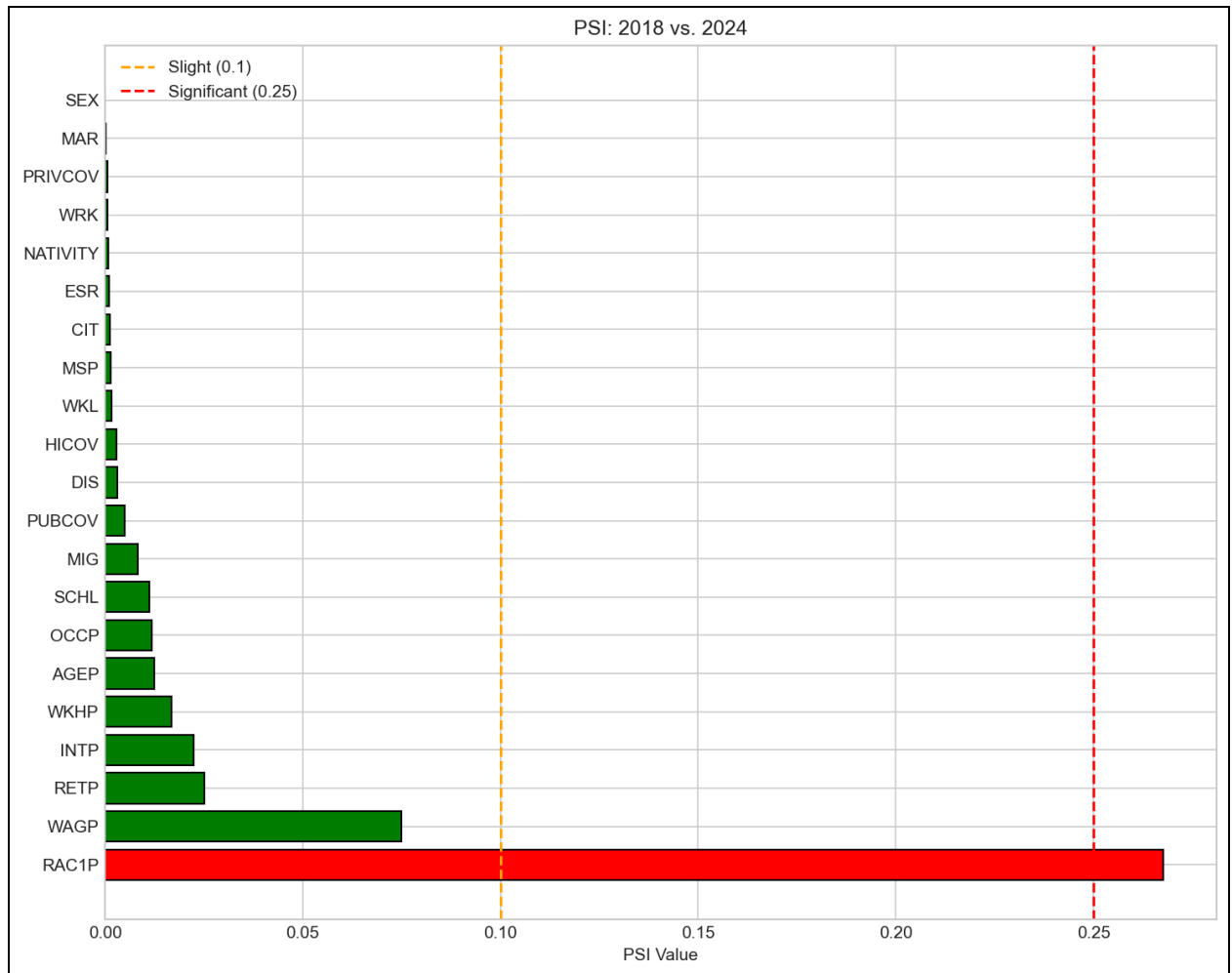
Multivariate Figures 20 - 25



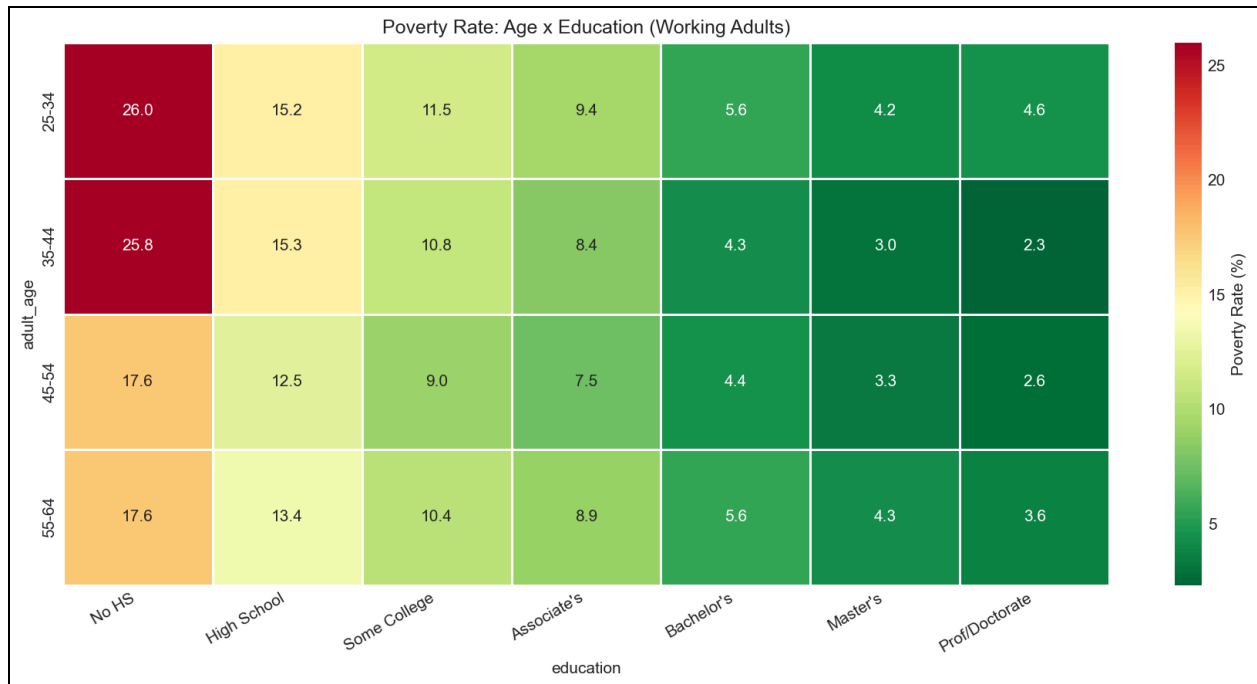
^Figure 20: Correlation heatmap among multiple features.



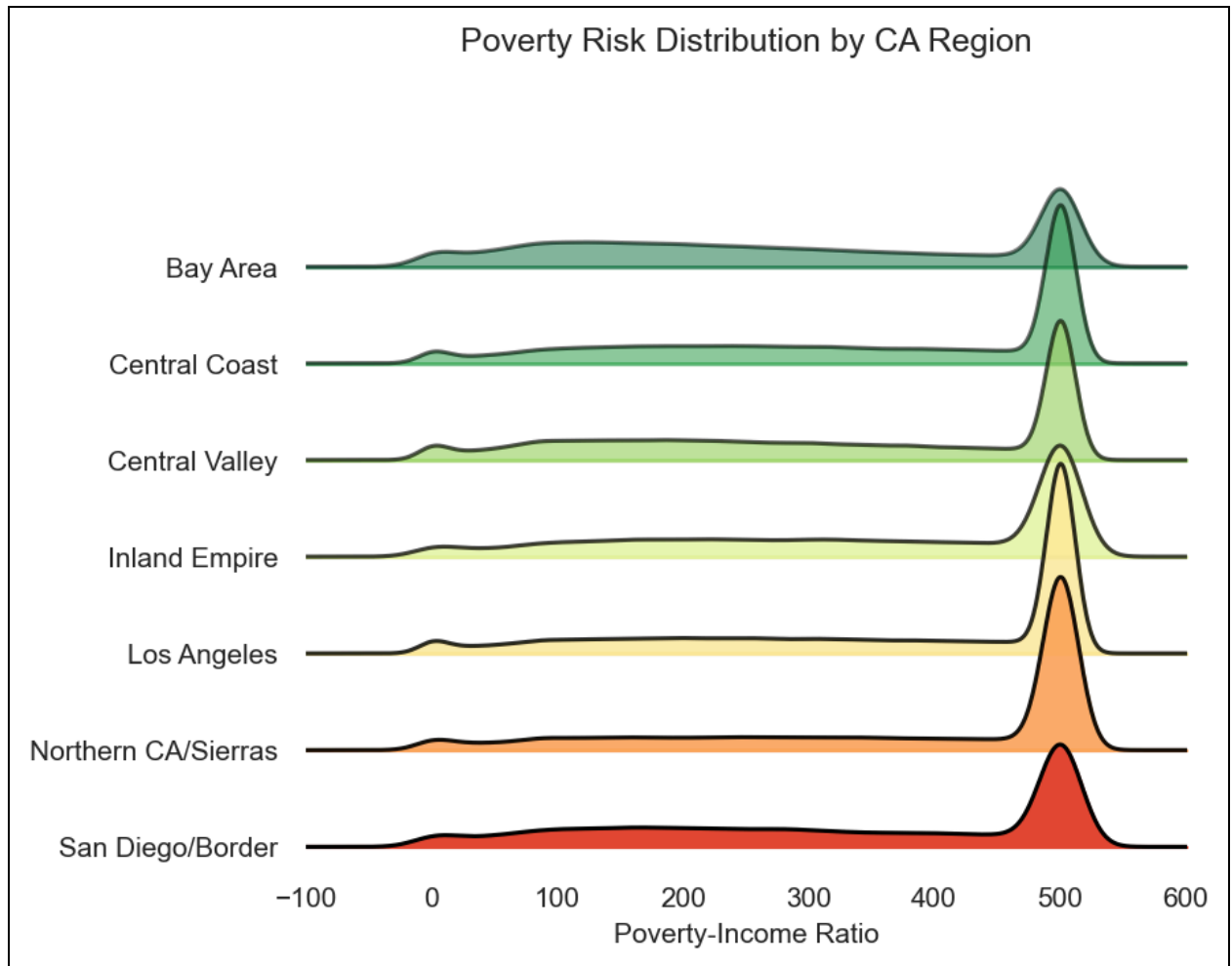
^Figure 21: Education vs. Poverty Risk heatmaps faceted by year (2018-2024).



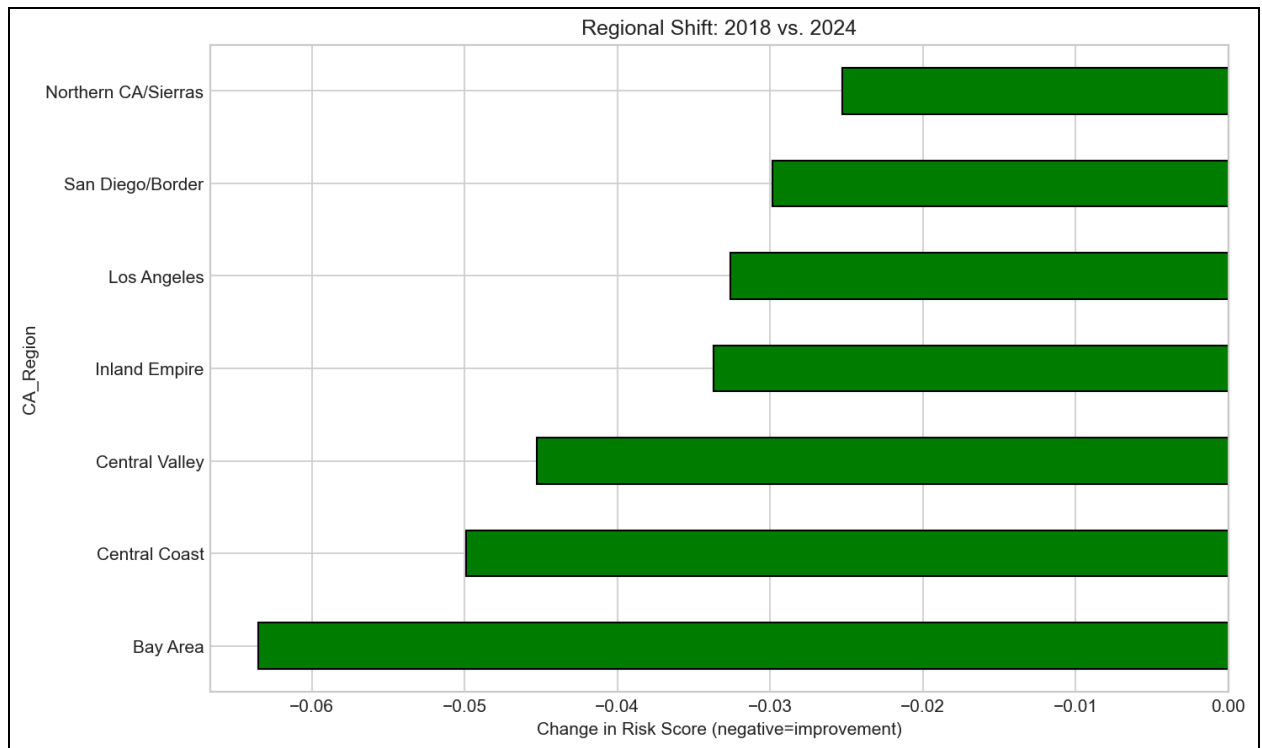
^Figure 22: Population Stability Index (PSI) drift analysis.



^Figure 23: Age x Education intersection heatmap showing poverty rates across combined demographic groups.



^Figure 24: Regional joyplot showing POVPIP density distributions across California regions.



^Figure 25: Regional shift analysis tracking poverty rate changes across years by region.