Predicting Median Income in Los Angeles County

Tommy Ngo, Natalia Fullmer, Jayden Tran, Megha Yaramada
University of California, Davis
Dr. Xiao Hui Tai

1 Introduction

Income inequality has emerged as one of the most significant socioeconomic challenges in the United States, especially in large metropolitan cities and countries. As one of the nation's most diverse and populous regions, Los Angeles County is home to over 10 million residents spread across 88 incorporated cities and 123 communities. Despite being one of the biggest economic hubs in California, the county exhibits stark contrasts in median household income across neighborhoods, with some areas characterized by wealth and prosperity. In contrast, others face persistent poverty and economic stagnation. To address these disparities requires a comprehensive understanding of the factors that might contribute to income inequality.

Motivated by both the practical and theoretical interest in understanding the relationship between socioeconomic variables and economic outcomes, this study aims to answer the question: What are the most significant socioeconomic factors influencing median household income in Los Angeles County? By addressing this question, the study hopes to provide insights into the underlying motivations for income disparities, enabling policymakers, urban planners, and community organizations to develop evidence-based policies to address economic inequality.

As mentioned, income inequality is not just a local issue; it has profound implications at the national and global levels. Prior studies have demonstrated that income inequality contributes to a range of social and economic problems, including reduced access to education, healthcare, and housing, as well as increased political and social instability. Understanding the specific drivers of income disparities in Los Angeles County offers valuable insights into how these patterns emerge in other urban regions, providing a potential framework for broader applications.

Prior research suggests that interactive variables such as education, employment, taxation, housing costs, and crime influence income levels. Higher levels of education and employment are generally associated with increased income, while high crime rates and heavy taxes can act as obstacles to financial growth. We hypothesize that among these variables, education will play the most significant role in predicting median household income in Los Angeles County. Areas with higher educational attainment are likely to exhibit higher incomes due to increased access to better-paying jobs and more stable employment opportunities. To investigate this relationship, we will employ predictive modeling using publicly available socioeconomic data regarding LA County. By using a combination of both linear models and clustering methods, we aim to uncover patterns and relationships across the dataset. The study also incorporates exploratory data analysis to explore common challenges such as multicollinearity and to identify the most significant predictors of income.

The findings from this analysis have both theoretical and practical implications. Theoretically, it contributes to the growing body of literature on the determinants of income inequality by focusing on a specific, highly diverse metropolitan area. On the other hand, the insights gained have practical benefits for informed policy decisions, such as investments in education, adjustments to tax policies, and the allocation of resources to economically disadvantaged communities. In doing so, this research aims to connect academic insights with real-world applications, laying the groundwork for tackling one of Los Angeles County's most significant socioeconomic challenges.

2 Related Studies

The challenges of understanding and addressing income inequality have become a significant topic for researchers across many disciplines. A notable study by Silva et. al. (2017) titled “Socioeconomic, scientific and technological indicators as parameters for prediction model” proposed a predictive model for socioeconomic development using Artificial Neural Networks (ANN), focusing on various socioeconomic, scientific, and technological indicators. The study accomplished this by selecting countries, choosing indicators, and developing an ANN-based prediction model through training and validating with historical data. Firstly, the country selection process was based on the relationship between normalized Gross Domestic Product (GDP) and Human Development Index (HDI). Using a sigmoidal normalizing function, the GDP for each country was scaled into an interval from 0 to 1. Then the mean between normalized GDP and HDI was calculated, and countries with the highest values were selected. Besides using common socioeconomic indicators such as wealth and health quality, the study also employed the scientific development status of countries through the number of yearly patents and published articles. A prediction and classification network were then trained and tested on the data using a split of 70% train, 15% test, and 15% validation subsets. The collected data consists of time series data from various sources from 1980 to 2015 for the 20 selected countries. The models yielded satisfactory results, with most indicators having an absolute error of less than 8%. However, the error rate for patent applications was higher at 25.33%, which may be due to the heterogeneity of the data set and the peculiar characteristics of this indicator. Overall, the study shows the effectiveness of using an ANN model for predicting a country’s development status. However, the authors suggest implementing a model that uses multiple indicators in tandem to capture a more comprehensive and accurate prediction.

In another study titled “Systematic comparison of household income, consumption, and assets to measure health inequalities in low- and middle-income countries,” health inequality in low and middle-income countries (LMICs) was investigated in relation to socioeconomic factors. Poirer (2024) assessed socioeconomic status (SES), household income, consumption, and total assets on health quality in 22 different countries using the Living Standards Measurement Study (LSMS) survey created by the World Bank. The LSMS survey was developed in 1980 with the purpose of time series data collection of households categorized as low to middle-income. The survey contains detailed and representative information regarding household living conditions and socioeconomic indicators such as income, consumption, education, health, employment, housing, and access to services. To assess the relationship, the author first compiled SES measures by aggregating household income and consumption data and calculating asset indices using polychoric principal component analysis (PCA). Then, the study simulated income distributions based on asset index rankings for each country and year by using a hybrid income proxy. Health quality was standardized using the World Health Organization (WHO) child growth standards, with stunting, underweight, and mortality rates as the main indicators. After, health inequalities were quantified using the following metrics: the Concentration Index (CI), the Relative Index of Inequality (RII), and the Slope Index of Inequality (SII). Each of these metrics was calculated for each SES measure. The study found that using asset indices and the hybrid income proxy showed larger health inequalities compared to using income or consumption, especially in wealthier low- and middle-income countries. However, there were no clear patterns of change over time. This highlights how the way socioeconomic status is measured can change how we observe patterns and their magnitude when analyzing inequality.

The two studies we review demonstrate the importance of indicator selection when assessing socioeconomic relationships. Silva et al.'s work highlights the potential of predictive models using socioeconomic data, emphasizing the effectiveness of machine learning techniques like artificial neural networks (ANNs) for capturing complex relationships among socioeconomic indicators and development status. Poirer's study built upon this, exploring the impact of different SES measures on health inequalities. However, both do not explore how specific indicators, such as crime rates or educational attainment, impact income predictions at a more granular level such as within a single region or county. To address this gap, our study aims to test how variables such as educational level, crime rates, and age distribution collectively contribute to income predictions in LA County, a highly diverse and economically complex region. This approach will expand on previous research by highlighting which specific indicators are most significant when predicting income in an urban setting.

3 Material and Methods

3.1 Data Collection

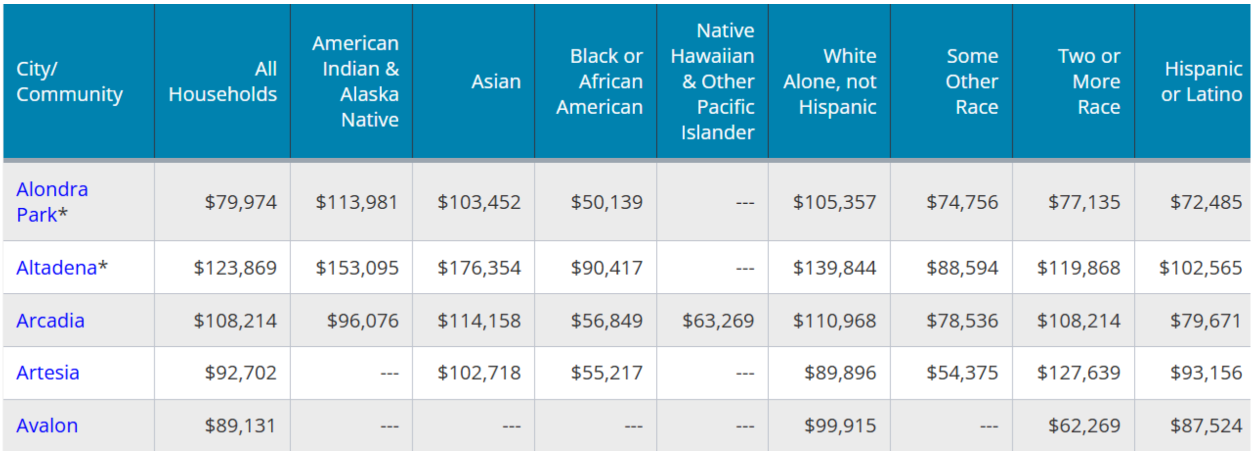
To examine the effects of socioeconomic factors on income, we rely on data taken from the American Communities Survey conducted by the United States Census each year from 2018 to 2022. The survey data is available on the US Census online database. The website does not separate county data by its respective cities; however, community-wise data is also available. For convenience, all communities within Los Angeles County were separately extracted, compiled, and organized into tables by demographic category, courtesy of the Los Angeles Almanac (LAA).

We select 40 variables of interest. Some labels instead only provided the estimated population proportion in percentage form rather than a number, and some provided both, so some variables are duplicated. The purpose of including both absolute and relative counts is to adjust for population size while showing the scale of a phenomenon. Variables of interest are measured in the survey by total population estimate and then subdivided into demographic labels and their respective categories. For example, we can extract an estimate of how many people belong to the "30 to 35 years old" category underneath the "age" label. For more general variables involving totals or averages, summary indicators are provided. The data is further organized into different websites.

Because the LA Almanac does not have data readily available to download in table format, we utilize the requests library in Python to extract the necessary data. Additionally, the website does not have a specified Application Programming Interface (API) to request data directly from the server in a structured format, so we performed HTML-based web scraping. This then requires us to structure the data into a table, which organizes the data into a configuration readable by computers for modeling.

To do this, we first requested the HTML file from each endpoint using Python 3.14. The data is then parsed using BeautifulSoup4, a library specifically designed for HTML and XML website scraping. Using the developer tool on the website, we identified relevant tags to access the main data in the format of a table, as seen in Figure 1. Then, a loop was created to iterate through the rows. The data was then transformed into a data frame for preprocessing, which included removing missing values and standardizing the column names. This was repeated for 7 different websites from the LAA data collection.

We then merge the variables of interest into one table on the community name. Adding a column for estimated community-wise median household income as the target variable, we can use this data to train linear regression models that predict median household income based on the variables of interest representative of demographic and socioeconomic features specific to each community.



City/ Community	All Households	American Indian & Alaska Native	Asian	Black or African American	Native Hawaiian & Other Pacific Islander	White Alone, not Hispanic	Some Other Race	Two or More Race	Hispanic or Latino
Alondra Park*	$79,974	$113,981	$103,452	$50,139	---	$105,357	$74,756	$77,135	$72,485
Altadena*	$123,869	$153,095	$176,354	$90,417	---	$139,844	$88,594	$119,868	$102,565
Arcadia	$108,214	$96,076	$114,158	$56,849	$63,269	$110,968	$78,536	$108,214	$79,671
Artesia	$92,702	---	$102,718	$55,217	---	$89,896	$54,375	$127,639	$93,156
Avalon	$89,131	---	---	---	---	$99,915	---	$62,269	$87,524

Figure 1: Example Table from Los Angeles Almanac Data Collection

3.2 Exploratory Data Analysis

Since our main objective is to apply regression analysis, we check some basic assumptions by looking at an overview of the data. We first plot each variable of interest against the target variable to assess general linearity. We observe that plots with more apparent correlations are education levels, and working and retirement populations. Also, there appears to be a stronger correlation between median income and our percentage data across the board than with raw quantitative data. This is likely since the percentage data is proportional and therefore normalized, providing a size-independent representation of the data. To build upon this, we calculate the Spearman Correlation value for each variable of interest. Notably, the percentages of Graduate Degrees and Grades 9-12 within a community have the strongest correlation with a Spearman value of 0. 74 and -0. 74, respectively. This further highlights the effectiveness of percentage data over absolute count, as indicators measured in percentage tend to have a much stronger correlation with the target variable.

Furthermore, we evaluate the constant variance of each indicator using the Breusch-Pagan Test. The test concluded that there is non-constant variance as most of the variables have low p-values. However, since we are looking at predictive modeling, the validity of the data still holds. We also assess multicollinearity amongst variables using the Variance Inflation Factor. As expected, there is a substantial amount of multicollinearity between related variables, such as different education levels or age distributions. To confirm this, we calculate the Spearman Correlation between Graduate Degrees (%) and all other percentage indicators for education. We conclude that there is a highly correlated linear relationship between these variables, with Undergraduate Degree (%) showing 0. 9363 and Grade 9-12 (%) showing a -0. 9177 correlation with Graduate Degree (%). Similarly, there is also a moderate to strong correlation between age distribution groups, with the strongest correlation being the population of 0-15 (%) and 65+ (%) at -0. 6773 correlation value.

Finally, a map of LA County was plotted based on tax rates to illustrate a general distribution of income. As seen in Figure 2, there is a clear pattern of lower tax rates for cities nearing the perimeter or coastline of the County. This can be explained by the fact that coastal cities tend to have higher property taxes, which do not need higher tax rates. That said, this can still be an indication of wealth since property ownership requires higher income or total household assets.

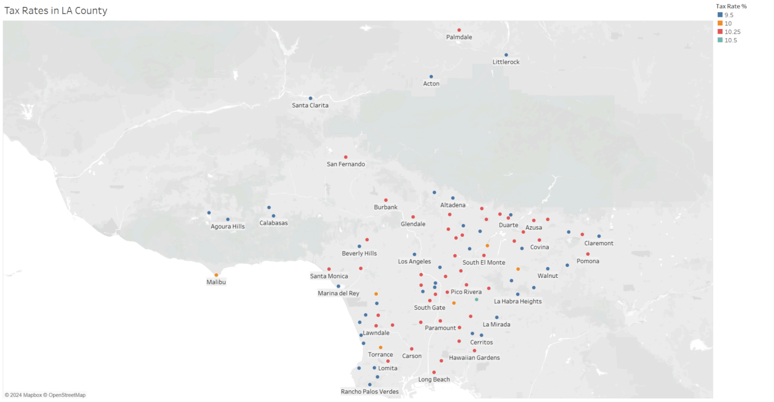


Figure 2: Map of LA County with Tax Rate

3.3 Statistical Methods

We trained the data using ridge regression and lasso regression. By comparing the two, we can quantify the effect of each input variable on median income in different ways. Ridge regression shrinks coefficients of less important predictors, reducing their impact. This is particularly helpful when multicollinearity is present as the model is stabilized and all predictors are retained. On the other hand, lasso regression sets those coefficients to zero, similar to feature selection. This simplifies the model further by identifying only the most important predictors, which makes the model easier to interpret. Z-score feature scaling was also applied to our data, which allows our predictors to have equal weighting and the hyperparameter of the models to be applied uniformly.

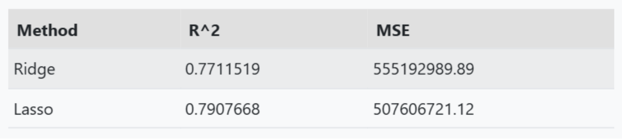
To supplement the predictive model, we implemented the k-Nearest Neighbor (kNN) algorithm. kNN does not assume linearity and classifies data points based on the closest neighbors in the feature space. The goal of using K-nearest neighbors in this study is to reveal geographical patterns: by grouping cities with similar characteristics and mapping these groups, we can highlight spatial trends on a map that are not apparent in regression models and are supported by quantitative analysis.

Before applying the k-Nearest Neighbor algorithm, we preprocessed the data by also using Z-score normalization. This standardization method made sure that all of the features contributed equally to the Euclidean distance calculations by rescaling them to have a mean of zero and a standard deviation of one. The optimal number of neighbors for the model was determined through cross-validation, which tested different k-values to balance between underfitting and overfitting. To determine the number of clusters, we use the Within-Cluster Sum of Squares (WCSS) value for our data to create an Elbow graph. The graph indicates that 3 clusters would be the optimal number, balancing bias and variance. We then fitted our data with the model and calculated the Silhouette Score of 0.26178, suggesting that there is a moderate strength in clustering. Additionally, we used silhouette scores and Principal Component Analysis (PCA) to validate and visualize the clustering patterns generated by kNN.

4 Results

The results of our analysis provide a comprehensive understanding of the factors influencing median household income across Los Angeles County. Using both linear regression and clustering methods, we identify significant predictors and geographic patterns in our data set.

Comparing predictive linear models, the results reveal that Lasso regression is the most effective linear method for predicting income in this study in regards to the R-squared statistic and Mean Squared Error (MSE), signifying that the Lasso model's predictions are closer to the actual values than the Ridge model:



Method	R^2	MSE
Ridge	0.7711519	555192989.89
Lasso	0.7907668	507606721.12

Figure 3: Method comparison

The Lasso model identifies household size as the strongest predictor of median household income. Larger households are positively correlated with income, i.e. the larger a household, the larger the predicted income. Aptly, households with children under the age of 18 are a significant predictor since this adds to the household size. Education is the second strongest predictor: the association between median household income and proportions of education level becomes more positive the higher the level of education (except for the “some college” variable). Similarly, populations above the age of 55 tend to have a positive relationship with median household income, both in absolute and relative count, while the younger population has a negative relationship. Overall, the Lasso model suggests that larger, older households with higher levels of education will have a higher median household income.

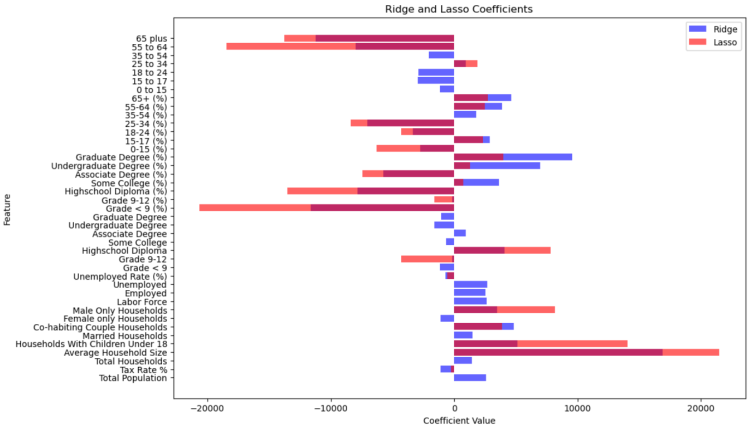
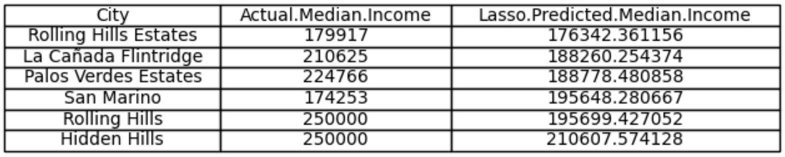


Figure 4: Coefficient Comparison

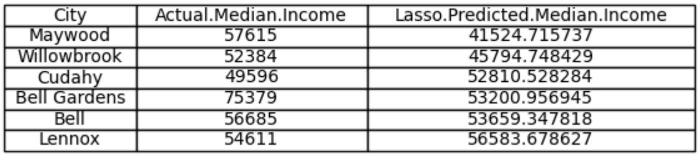
Both the Lasso and Ridge models assign significant weight to these predictors, but expectedly the impact is less in the Ridge model because weights of the coefficients are distributed across all variables of interest for stability.

Geographic trends also highlight disparities in income distribution and similarities in demographics across the county. The model predicts that communities located near the perimeter of Los Angeles County had higher median income (Figure 4). In contrast, communities located in the southern and central regions of Los Angeles County are at the lower end of the income spectrum (Figure 5).



City	Actual.Median.Income	Lasso.Predicted.Median.Income
Rolling Hills Estates	179917	176342.361156
La Cañada Flintridge	210625	188260.254374
Palos Verdes Estates	224766	188778.480858
San Marino	174253	195648.280667
Rolling Hills	250000	195699.427052
Hidden Hills	250000	210607.574128

Figure 5: 6 highest earning communities



City	Actual.Median.Income	Lasso.Predicted.Median.Income
Maywood	57615	41524.715737
Willowbrook	52384	45794.748429
Cudahy	49596	52810.528284
Bell Gardens	75379	53200.956945
Bell	56685	53659.347818
Lennox	54611	56583.678627

Figure 6: 6 lowest earning communities

To confirm local patterns that may contribute to geographic placement, we perform k-Nearest Neighbor analysis. Figure 8 illustrates that communities located more inland are placed into Cluster 2. The distribution of Cluster 0 and Cluster 1 across the map is less apparent, so we reduce the amount of clusters from 3 to 2 in Figure 9. (The unclear distribution of Cluster 0 in particular is supported in Figure 7, where the data points are not as close together as Cluster 1 and 2.) Here, by comparing the map of clusters to income, it becomes clearer that the algorithm separates communities located around the perimeter from those that are inland. This further enforces our hypothesis that the geographic positioning of a community is also an indicator of wealth.

Including both the proportion of age and education, plus the absolute count may cause issues to arise. Using both informs us on the size of a community *and* the relative size of the community population falling underneath a certain variable of interest. However, this could unnecessarily inflate the complexity of the model and lead to false predictions: looking at Figure 6, the model predicts that the community of Bell Gardens has a median household income of $53,000 per year, while the true estimate was $75,379 per year. This could be due to redundancy in the model. Figure 7 suggests that the data points in Cluster 0 are weakly correlated from their broad spread, perhaps a result of overcomplexity.

On the other hand, the kNN algorithm groups Bell Gardens with the remaining 5 lowest-earning communities in Los Angeles County based on our variables of interest: Maywood, Willowbrook, Cudahy, Bell, and Lennox. All 6 communities share similar average household sizes, age distribution, and levels of education. Additionally, on a geographical map, Bell Gardens is directly bordered by Maywood, Cudahy, and Bell. This information suggests that shared demographic features and geographical proximity shape household income with the exception of Bell Gardens. Here, Lasso regression provides a more overall view between income and variables of interest while kNN assists in finding local patterns in the data where linear relationships may not apply.

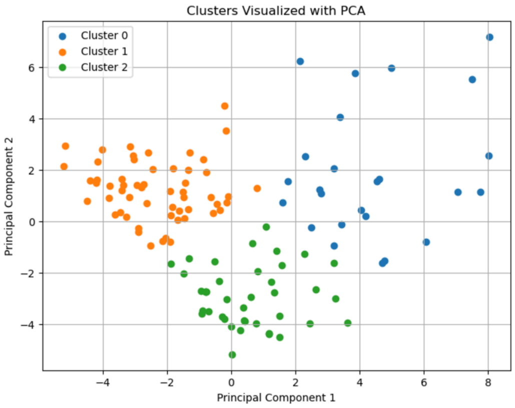


Figure 7: kNN Cluster Visualization with PCA

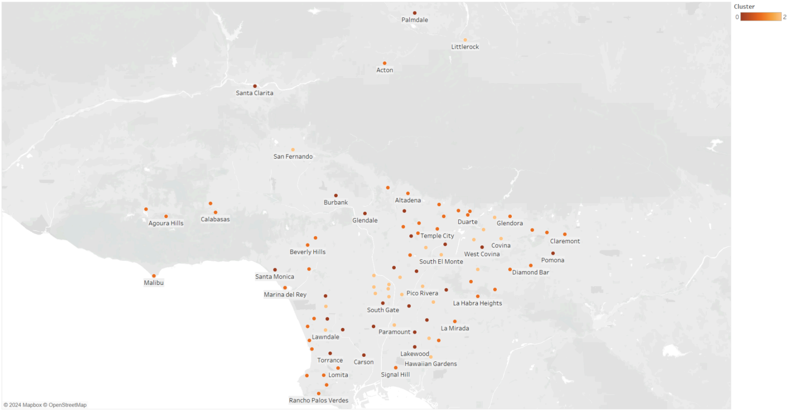


Figure 8: LA County map based on Cluster Assignment

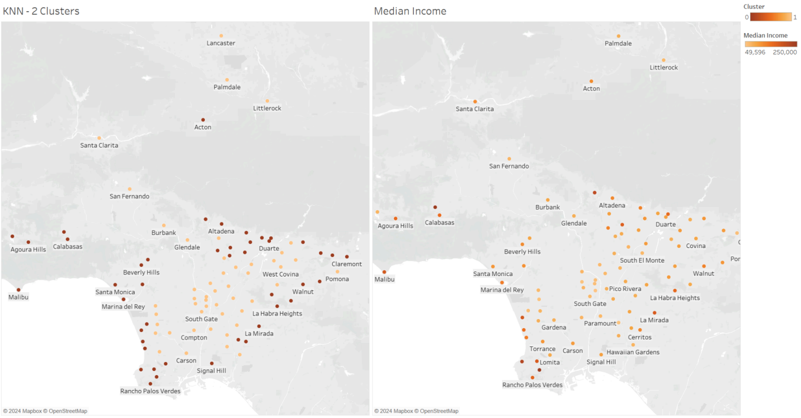


Figure 9: Reduced Clusters vs. Median Income

5 Discussion and Conclusions

5.1 Discussion

This study aimed to predict median household income across various towns, cities, and communities in Los Angeles County using socioeconomic and demographic-based variables, as well as identifying significant predictors that may be influencing the median income of a community. Educational attainment, household size, and proportion of elderly residents were identified as the strongest positive predictors of median household income. We expected education to be an important indicator of median household income because of the mobility education provides in the job market. A larger household size would be able to pool more resources together, which aligns with Poirier (2024), where total household asset and consumption is also a significant contributor to health service access. Furthermore, a larger elderly population would have more time to accumulate more wealth, such as from investments such as property and stocks, as well as retirement-based savings such as a Roth IRA, 401(k), and social security

We were also able to identify various geographical trends, revealing pronounced disparities in income distribution across the county. Cities around the perimeter of the county border tended to have a higher income than those further inland. Additionally, the cities on the extremely low end of predicted income such as Cudahy, Bell, and Lennox were clustered together in the southern region of the county.

These results have several important implications for policymakers. They can decide on policies and infrastructure that could focus on fostering economic stability for families, as well as support for the aging population. Moreover, the stark geographical disparity necessitates the need for more redistribution of economic opportunities, informing policymakers and urban planners to invest in developing high-paying industries, improving access to higher education, and enhancing infrastructure for lower-income areas to create more equitable economic growth. Urban planners can also use these findings to improve Los Angeles County's public transportation infrastructure, which might help with income inequality since it would allow lower-income neighborhoods easier access to more important industries and services.

5.2 Limitations

Since this study is only confined to Los Angeles County, it could have limited implications for other regions in the country because the demographic makeup of a region and its correlation to income may not be exactly the same as what the model is trained on. Additionally, several statistical assumptions can bring the reliability of interpretation into question. The assumption of heteroscedasticity (assumption of constant variance) is indicated by our use of the constant variance test, the Breusch-Pagan Test, which can affect the reliability of coefficient interpretations. Multicollinearity among predictors also posed challenges, resulting in the use of regularization methods such as Ridge and Lasso regression to stabilize our models. While these methods were effective for predictive purposes, interpretation and oversimplification of relationships between variables may bring problems, as seen in the case of Bell Gardens.

5.3 Future Research

If further research can be conducted, performing a similar study but with a larger geographical area such as including more counties near Los Angeles, the Bay Area, and even the State of California can allow for broader generalizations. Incorporating other models such as Random Forests or Gradient Boosting Machines may also capture relationships between socioeconomic variables. The application of Principal Component Analysis may also simplify overlapping variables and improve model interoperability and performance. Beyond statistical improvements, exploring more variables such as race, housing/renting prices, and health-related data can help paint a better picture of what variables impact median household income and income inequality.

5.4 Conclusion

Ultimately, this study provides a deeper understanding of the factors associated with median housing income in Los Angeles County. Highlighting the roles of education, household size and structure, and proportion of the elderly population in median household income, can bring valuable insights for addressing income inequality. By identifying key socioeconomic variables that contribute to income disparities, the study can inform policy and identify target regions in the highest need of aid to combat poverty, deploy intervention strategies, and open more opportunities for education and higher standards of living. These findings can also showcase the importance of policies that consider socioeconomic variables and regional dynamics to address income inequality to create a more equitable world.

7 References

1) Los Angeles Almanac. (n.d.). Median income by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/employment/em12.php
2) Los Angeles Almanac. (n.d.). Age distribution by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/population/po09a.php
3) Los Angeles Almanac. (n.d.). Household information by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/population/po30.php
4) Los Angeles Almanac. (n.d.). Unemployment rates by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/employment/em03.php
5) Los Angeles Almanac. (n.d.). Crime rates by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/crime/cr03.php#:~:text=Other%20Los%20Angeles%20County%20cities
6) Los Angeles Almanac. (n.d.). Tax rates by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/government/gx11.php
7) Los Angeles Almanac. (n.d.). Educational levels by cities. Retrieved November 4, 2024, from https://www.laalmanac.com/education/ed33aa.php
8) U.S. Census Bureau. (n.d.). When to use 1-year to 5-year estimates. Retrieved November 4, 2024, from https://www.census.gov/programs-surveys/acs/guidance/estimates.html
9) Los Angeles Almanac. (n.d.). *Income thresholds and their categories standardized for state and metropolitan area; adjusted for household size and inflation*. Retrieved November 4, 2024, from https://www.laalmanac.com/employment/em720.php
10) Pew Research Center. (2024, September 16). *Are you in the American middle class?* Retrieved November 18, 2024, from https://www.pewresearch.org/short-reads/2024/09/16/are-you-in-the-american-middle-class/
11) Pew Research Center. (2024, May 31). *The state of the American middle class*. Retrieved November 18, 2024, from https://www.pewresearch.org/race-and-ethnicity/2024/05/31/the-state-of-the-american-middle-class/
12) Poirier, M. J. (2024). Systematic comparison of household income, consumption, and assets to measure health inequalities in low- and middle-income countries. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-54170-1
13) Silva, L. F., da Silva, M. C., Alves, A. J., Reis, M. R., Bulhoes, J. S., Costa, R. E., Silva, B. C., Aleixo, E. L., Gomes, V. M., & Calixto, W. P. (2017). Socioeconomic, scientific and technological indicators as parameters for prediction model. 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 1–6. https://doi.org/10.1109/chilecon.2017.8229733

8 Code Repository

The complete code, data, and documentation for this project are available at the following GitHub repository: https://github.com/mpngo/Income-Prediction-LAcounty