STAT 486: Final Report

Madison Wozniak

Table of Contents

Introduction

Problem Statement

Food affordability in the Unites States has reached an all-time high in recent years, and is a struggle that many individuals and families are no stranger to. This is a critical issue for single mothers in particular, as it impacts their household's stability, health, and overall quality of life. In California especially, the average cost of living remains just barely achievable for many and women-headed households face unique economic challenges because of it. The ability to understand and predict the median income of women-headed households in California can provide valuable insights into financial realities, and help identify common trends.

Aim of this Report

This report aims to understand if food affordability and socioeconomic factors are accurate tools in predicting the median income of women-headed households in California. By looking at data that includes information on food costs, regional demographics and economic indicators, this analysis seeks to accomplish the following:

1. Understand the relationship between income and food affordability.
2. Develop predictive machine learning models on median income.
3. Address how these insights can influence policy changes and improve economic outcomes in the future.

Exploratory Data Analysis

To fully understand the important factors that may influence median income, we must first look at the raw data. This section gives an overview of the key variables by identifying relevant distributions and variations within the data. These summary statistics act as a foundation for deciding what potential predictors of median incme could be.

The dataset includes the following features:

- `race_eth_name` (string) Name of race/ethnic group
- `geotype` (string) Type of geographic unit place (PL), county (CO), region (RE), state (CA)
- `geoname` (string) Name of geographic unit
- `county_name` (string) Name of county the geotype is in
- `region_code` (string) Metropolitan Planning Organization (MPO)- based region code
- `cost_yr` (numeric) Annual food costs
- `median_income` (numeric) Median household income
- `affordability_ratio` (numeric) Ratio of food cost to household income
- `LL95_affordability_ratio` (numeric) Lower limit of affordability 95% confidence interval
- `UL95_affordability_ratio` (numeric) Upper limit of affordability confidence interval
- `rse_food_afford` (numeric) Relative standard error of affordability ratio
- `CA_RR_Affordability` (numeric) Ratio of affordability rate to California affordability rate
- `ave_fam_size` (numeric) Average family size - combined by race
- `latitude` (numeric) Latitude coordinate of region
- `longitude` (numeric) Longitude coordinate of region
- `affordability_per_person` (numeric) Affordability ratio relative to family size

With the variables defined above, the next step is to delve into the statistical properties of each, including the underlying distributions and variations of each. It is important to be aware of any potential outliers or missing data before selecting predictive features for future models.

	reg_code	cost_yr	med_income	afford_ratio	LL95_afford	UL95_afford	rse_afford	afford_decile
count	295362	264366	101836	101836	99445	99445	99445	26715
mean	11.388	7602.632	38038.998	0.325	0.113	0.769	53.362	5.804
standard deviation	3.341	1435.062	27050.591	0.42	0.116	3.599	174.806	2.685
minimum	1	3095.425	2500	0.021	0	0.041	0.684	1
Q2	9	6667.128	22417	0.153	0	0.229	12.574	4
median	14	7460.84	33103	0.227	0.092	0.351	25.596	6

	reg_code	cost_yr	med_income	afford_ratio	LL95_afford	UL95_afford	rse_afford	afford_decile
Q4	14	8325.618	46418	0.349	0.174	0.561	53.824	8
maximum	14	16872.05	25000	4.852	0.851	108.784	5227.124	10

Table 2.1: Numeric Data Summary Satistics

	latitude	longitude	afford_per_person	CA_RR_Affordability	ave_fam_size
count	291717	291717	101733	101836	280593
mean	35.218	-116.352	0.097	1.219	3.294
standard deviation	2.655	8.745	0.124	1.578	0.636
minimum	26.13	-123.73	0.009	0.08	1.36
Q2	33.97	-119.67	0.05	0.576	2.88
median	34.06	-118.26	0.069	0.852	3.25
Q4	36.75	-117.32	0.101	1.312	3.61
maximum	48.92	-71.35	0.996	18.214	7.2

Table 2.2: Numeric Data Summary Satistics

Tables 2.1 and 2.2 act as a good baseline, but sometimes it can be easier to understand relationships with graphs. Below is a correlation matrix that visualizes the how strongly certain features are correlated with one another.

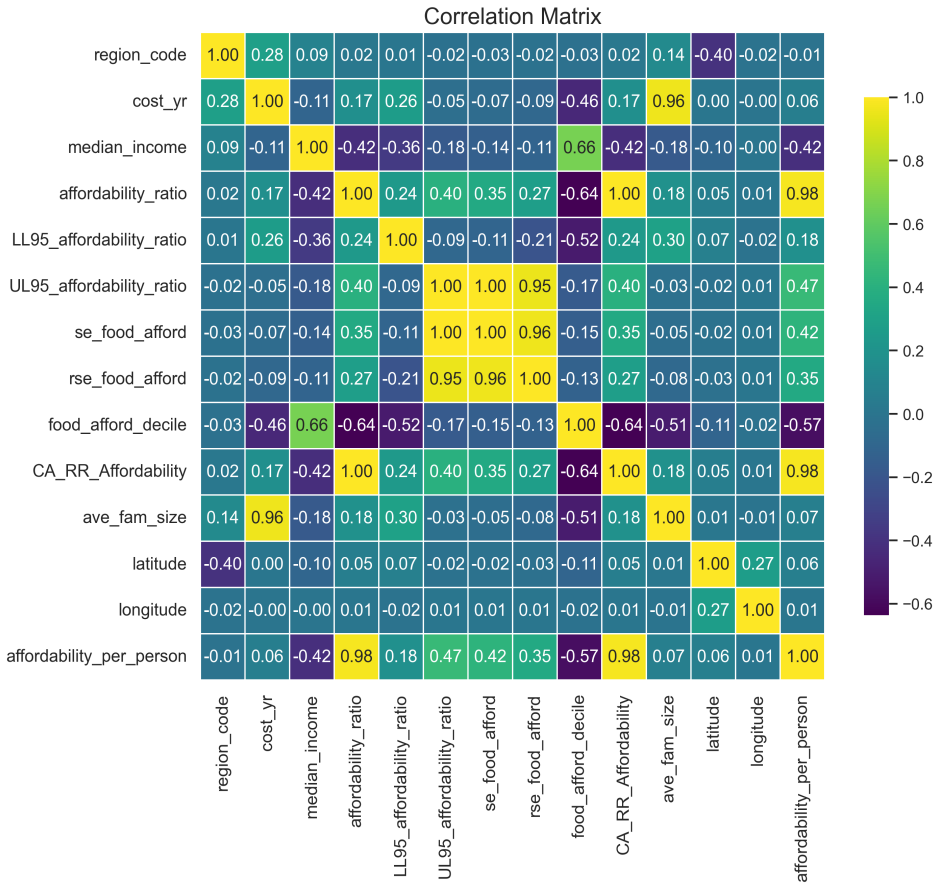

Figure 1: Correlation Matrix for Numeric Features

It can be noted from the matrix, that there is a strong, almost perfect correlation between several variables. This is expected because many variables such as `UL96_affordability_ratio`, `LL95_affordability_ratio`, `se_food_afford`, `rse_food_afford`, `CA_RR_Affordability` and `affordability_per_person`are all calculated from information taken from `affordability_ratio`. Some more interesting strong correlations to take note of are between `cost_yr` and `ave_fam_size`, `med_income` and `food_afford_decile`.

There are also some important relationships between income and average family size, and racial/ethnic group as illustrated below in Figures 2 and 3. Figure 3 reveals a particularly interesting trend: while the spread of the points suggests that most women-headed households earn enough to afford

food for their families, the red line superimposed on the graph highlights the unfortunate reality. The median annual income in this dataset is $33,103.00. This is significantly below the median annual household income in the United States which stood at $80,610.00 as of 2022.

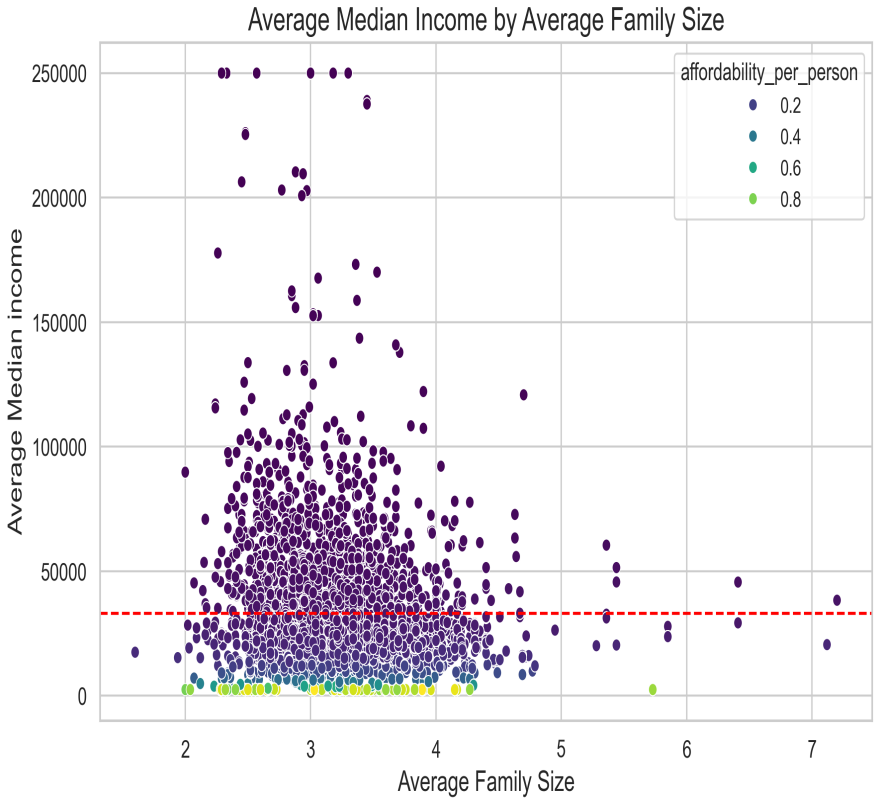

Figure 2: Scatterplot of Income Against Family Size

	race_eth_name	geotype	geoname	county_name
count	295371	295371	295371	295236
unique	9	4	1581	58
top	AIAN	PL	Franklin CDP	Los Angeles
frequency	32819	298431	990	123966

Table 3: Categorical Data Summary Satistics

present summary stats/plots that give an overview of the data

present any key findings from EDA

Methods

Feature Engineering

anything that was done, and how it improve the model

include dimension reduction features

list or table of all models used

for each:

- brief description
- key hyperparameters used
- high-level model results

Discussion on Model Selection

mentioned patterns observed across models

Detailed Discussion on Best Model

Conclusion and Next Steps