Analyzing the Relationship Between Demographic, Geographic, and Economic Factors and Cost-Burdened Housing in Georgia

Authors:

Jason Withrow, Ken Wong, Rajath Prabhakar, William Hudson

Client:

The Carl Vinson Institute of Government (CVIOG) at the University of Georgia: Scott King, Taylor Hafley, James Byars

Department:

Department of Statistics, University of Georgia

Course:

STAT 5020W: Statistical Capstone Course II

Date: April 30, 2025

Introduction

The Carl Vinson Institute of Government (CVIOG) at the University of Georgia provides research, consulting, and training services to Georgia's state and local governments. It attempts to support policy development and decision-making by providing data-driven insights to help improve public services and the well-being of Georgia residents. But, according to a 2023 article by Chris Dowd, there is one such problem compromising the well-being of Georgia residents: housing insecurity (Dowd, 2023).

According to Dowd, in Athens, more than half of its residents are housing burdened, spending more than 30% of their income on housing. For those earning minimum wage or close to minimum wage, the situation is much more dire; they must rely on low-income housing options such as Section 8. This is not only an issue of supply, but an issue of corporate greed. GA Code 44-7-19, passed in 1981, prohibits counties and municipalities from enacting rent control, effectively allowing landlords to raise rents without limits(Georgia Code, 2024, § 44-7-19).

Consequently, since the 2008 housing crash, hedge funds and private equity investors have been buying up starter homes and apartments, especially in Atlanta and other big cities, leaving properties vacant to appreciate in value or to be transformed into "luxury rentals" (Perry, 2023). In metro Atlanta alone, these bulk buyers have accumulated more than 65,000 single-family homes over the past decade, with 11 companies owning over 1,000 homes each.

Even though these corporate buyers did not create the housing crisis, they greatly exacerbated it. Metro Atlanta home values rose across the board from 2012 to 2022. The same analysis found that they climbed more sharply in places where investors bought more houses. In fact, in the 30 ZIP codes with the most investor-owned properties, home values appreciated at nearly twice the annual rate as the 30 ZIP codes where investors own the least. As of this writing, the six largest single-family rental firms own 63% of the homes in metro Atlanta.

Project Focus

The goal of this study is to analyze the housing burden in Georgia by investigating the demographic, economic, and geographic features that contribute to the challenges of housing affordability and access. Specifically, we will examine how factors such as income, race, education, and location influence housing affordability and access across the state. Understanding these relationships will inform policies with the goal of improving housing affordability and accessibility across the state of Georgia. This project is being

conducted in collaboration with the CVIOG, who concentrate on improving the effectiveness of government through research and analysis. Our findings will help CVIOG support local governments and communities across Georgia in addressing the housing burden crisis.

Research Questions

This study addresses multiple research questions as part of a broader investigation into housing affordability and accessibility in Georgia, with a primary focus on the first research question:

1. What is the relationship between demographic, geographic, and economic features collected by the Census Bureau's American Community Survey and the percentage of owner- and renter-occupied housing that is classified as cost-burdened (e.g., 30% of their income dedicated to housing costs)?
2. What demographic, geographic, and economic features collected in the Census Bureau's American Community Survey influence the housing mix (e.g., owner-occupied, renter-occupied, owner-vacant, and renter-vacant) in each geographical area in Georgia?
3. Is there a statistically driven and reliable index or indices that provides a comprehensive understanding of housing availability, accessibility, and affordability that highlights a geographic area's housing security or insecurity, similar to the concept of food deserts?

Data Description

The data for this project is sourced from the ACS which is conducted by the U.S. Census Bureau. The ACS collects data continuously throughout the year, with approximately 3.5 million households surveyed annually. This large sample size allows the survey to generate reliable estimates for all types of communities in Georgia. Households are selected using a random sampling method, ensuring that all areas are represented. Data collected from the sampled households are then weighted by the Census Bureau to account for demographic and geographic differences, allowing the survey to produce accurate estimates that are representative of the entire population of Georgia. The demographic data collected by the survey focuses on topics such as income, housing, education, employment, transportation, and more. Contrary to the decennial census where everyone is counted the ACS surveys a sample of households each year and uses that data and their own methods to extrapolate that data to represent total populations of the United States.

The two main estimates used in this analysis are 1–Year and 5-Year estimates. The 1–Year estimates are straightforward and are only estimates for that year. On the other hand the 5–Year estimates are a bit

different in that they represent a 5 year period such as 2017–2021. Rather than summing data from each year, the Census Bureau averages it to produce an estimate that reflects a typical year within the period.

The data are split into multiple tables each focusing on specific demographic information. For example, separate tables detail cost-burden for owner-occupied and renter occupied housing and are further broken down by various race categories, age groups, and other demographic features. In addition to using the ACS data as presented by the U.S. Census Bureau, a separate aggregated dataset was created specifically for this analysis. The dataset we used was constructed using the 2023 ACS 5-Year estimates and contains 159 rows corresponding to each county in Georgia. The dataset includes a variety of variables (listed in Table 1) that capture demographic, economic, and housing-related information across geographic regions.

Table 1: Description of Variables Used in the Analysis

Variable Name	Description
GEOID	Geographic identifier for the county (FIPS Code).
NAME	Name of the county.
MedHouseIncome	Median household income in the county.
Pct_CostBurdened_Owners	Dependent Variable: Percentage of owner-occupied housing units classified as cost-burdened.
Pct_CostBurdened_Renter	Dependent Variable: Percentage of renter-occupied housing units classified as cost-burdened.
Cost Burden by Age	
Pct_CostBurdened_15_24	Percentage of all individuals aged 15-24 classified as cost-burdened.
Pct_CostBurdened_25_34	Percentage of all individuals aged 25-34 classified as cost-burdened.
Pct_CostBurdened_35_64	Percentage of all individuals aged 35-64 classified as cost-burdened.
Pct_CostBurdened_65_plus	Percentage of all individuals aged 65 and over classified as cost-burdened.
Cost Burden by Age (Renters)	
Pct_RenterCostBurdened_15_24	Percentage of renters aged 15-24 classified as cost-burdened.
Pct_RenterCostBurdened_25_34	Percentage of renters aged 25-34 classified as cost-burdened.
Pct_RenterCostBurdened_35_64	Percentage of renters aged 35-64 classified as cost-burdened.
Pct_RenterCostBurdened_65_plus	Percentage of renters aged 65 and over classified as cost-burdened.
Cost Burden by Race (Owners)	
Pct_CostBurdened_White_Owners	Percentage of White homeowners classified as cost-burdened.
Pct_CostBurdened_Black_Owners	Percentage of Black homeowners classified as cost-burdened.
Pct_CostBurdened_Other_Owners	Percentage of homeowners of Other races (Native, Asian, Pacific Islander combined) classified as cost-burdened.
Cost Burden by Race (Renters)	
Pct_CostBurdened_White_Renters	Percentage of White renters classified as cost-burdened.
Pct_CostBurdened_Black_Renters	Percentage of Black renters classified as cost-burdened.
Pct_CostBurdened_Other_Renters	Percentage of renters of Other races (Native, Asian, Pacific Islander combined) classified as cost-burdened.
County Characteristics	
isUrban	Indicator variable denoting whether the county is urban (e.g., 1) or rural (e.g., 0).
Unemployment_Rate	Unemployment rate in the county.
Pct_Bachelors_Higher	Percentage of the county population (typically aged 25+) with a bachelor's degree or higher education level.

Note: Variables derived from the 2023 American Community Survey (ACS) 5-Year Estimates. Cost-burdened defined as spending 30% or more of household income on housing costs. The specific definition/coding for `isUrban` should be detailed in the text. The 'Other' race category combines Native, Asian, and Pacific Islander populations.

When exploring census data, there are many tables with row values marked "not computed." This is often a result of incomplete information provided by individuals or households. The official census documentation advises removing rows with "not computed" values, but the aggregated dataset did not include any of such values. Instead, as a result of the computations used to obtain certain columns in the aggregated set, new NA values were introduced. Figure 1 shows a histogram of the missingness present in the new aggregated dataset. These NA values function as zero, but were labeled as NA due to special circumstances in which calculations involved dividing by zero. Thus, these NA values were replaced by zeroes, and there is no further

missingness present.

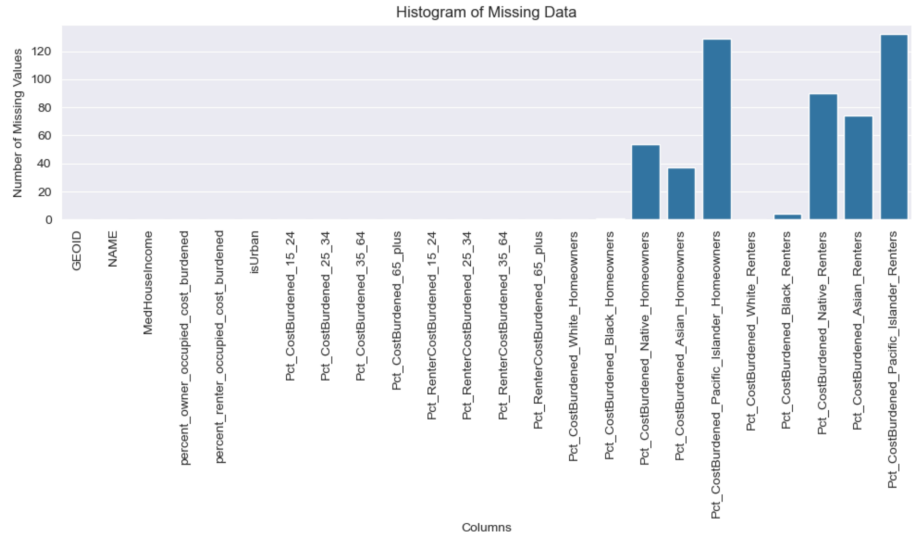

Figure 1: Histogram of Missingness

The analysis was conducted using *R* (R Core Team, 2024) with an emphasis on the use of the *tidycensus* (Walker & Herman, 2024) package, which simplifies accessing the ACS data by connecting to Census Bureau API. The API allows for users to directly retrieve data into R. This package facilitates data wrangling tasks, such as filtering and aggregating variables. Along with that *tidycensus* integrates with visualization libraries in R, allowing for seamless creation of maps and charts to illustrate housing burden trends across the various regions of Georgia. Instead of summing up data from each year, the Census Bureau averages the data to produce an estimate that reflects a typical year within the 5-year period

Exploratory Data Analysis (EDA)

To provide proper context for the forthcoming analysis, it is important to first consider the summary statistics of the overall demographics of Georgia. The plot of median income in Georgia (Figure 2) highlights the income disparity across the state, with the highest median incomes centered around metro Atlanta, while the more rural regions display notably lower median incomes.

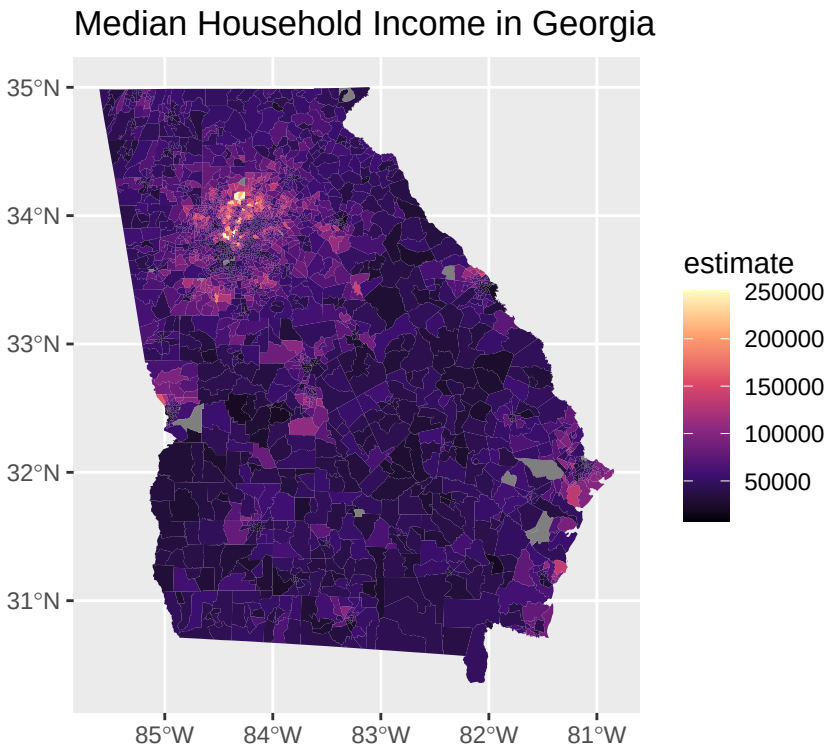

Figure 2: Median Household Income (2020)

Next, the plot of housing cost burden by county (Figure 3) reveals the distribution of cost-burdened housing across different geographic regions in Georgia. Figure 3 displays that housing cost burden is distributed rather evenly across the state, regardless of the median income of the respective area. The concentration of higher median incomes in metropolitan areas is expected; however, areas of higher median income do not appear to correspond with lower levels of housing burden. This suggests that rising housing costs in cities may create affordability challenges even for higher earning households. The data was processed using *dplyr* (Wickham et al., 2023) for data manipulation and tools from *tigris* (Walker, 2024) for spatial plots. Additionally, *ggplot2* (Wickham, 2016) was used for data visualization. Lastly, *sf* (Pebesma & Bivand, 2023) was used for managing geospatial objects. The packages *dplyr* and *ggplot2* come from *tidyverse* (Wickham et al., 2019), a collection of R packages for data manipulation, visualization and other tasks.

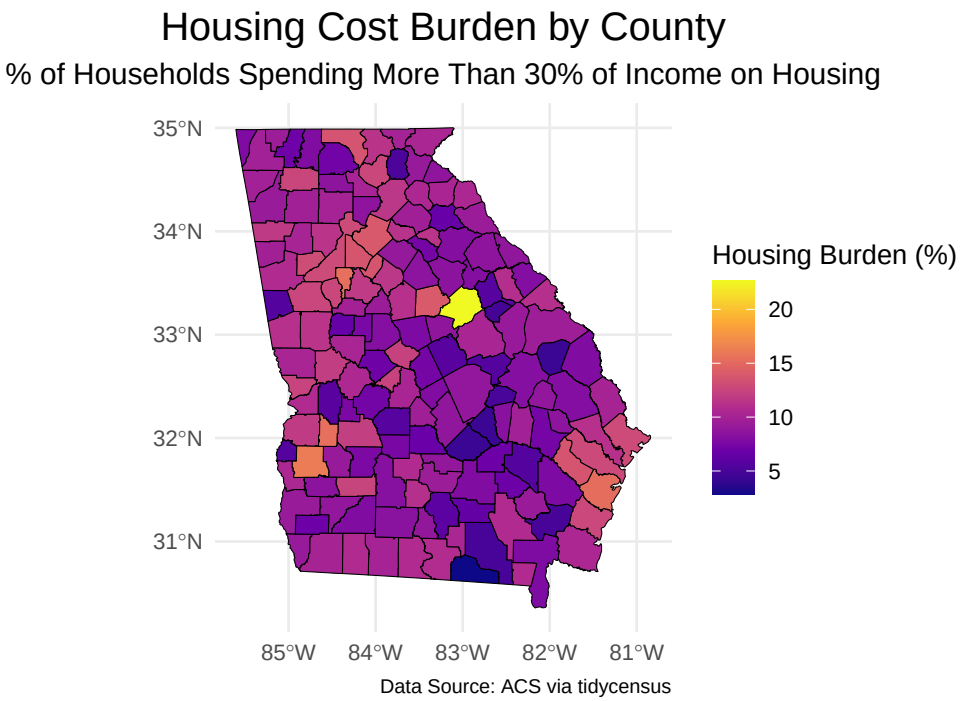

Figure 3: Housing Cost Burden by County (ACS 2017-2021)

When examining population and income by race (Figures 4 & 5), economic and demographic inequities become more pronounced. White populations make up the largest demographic, appearing spread out across the state, with a notable absence in urban Atlanta. Black populations also make up a large proportion of the population, but tend to concentrate in urban Atlanta, as well as central and southern regions of Georgia. Hispanic and Asian populations, on the other hand, make up lower proportions of the population and are centered primarily around metro Atlanta. The plot of median income by race displays the greatest median household incomes in White households, particularly in suburban areas, while Black and Hispanic households consistently report lower median household incomes. The Asian population tends to have greater median household incomes on average in comparison to Black and Hispanic populations, but this wealth is generally centered around metro Atlanta. The income and population disparities present in the plots emphasize the greater vulnerability to housing insecurity for Black and Hispanic households.

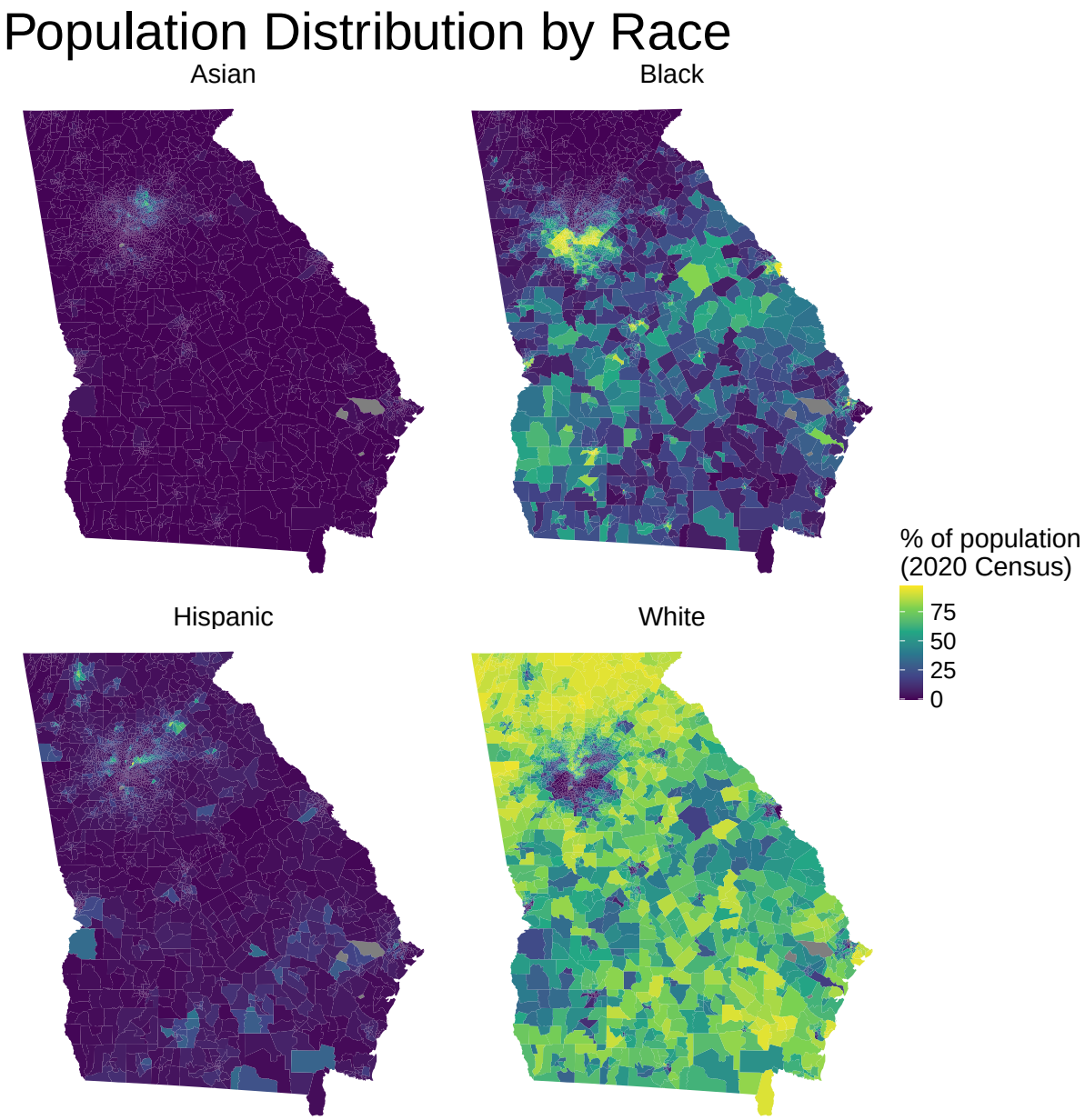

Figure 4: Population Distribution by Race (2020 ACS 1-year)

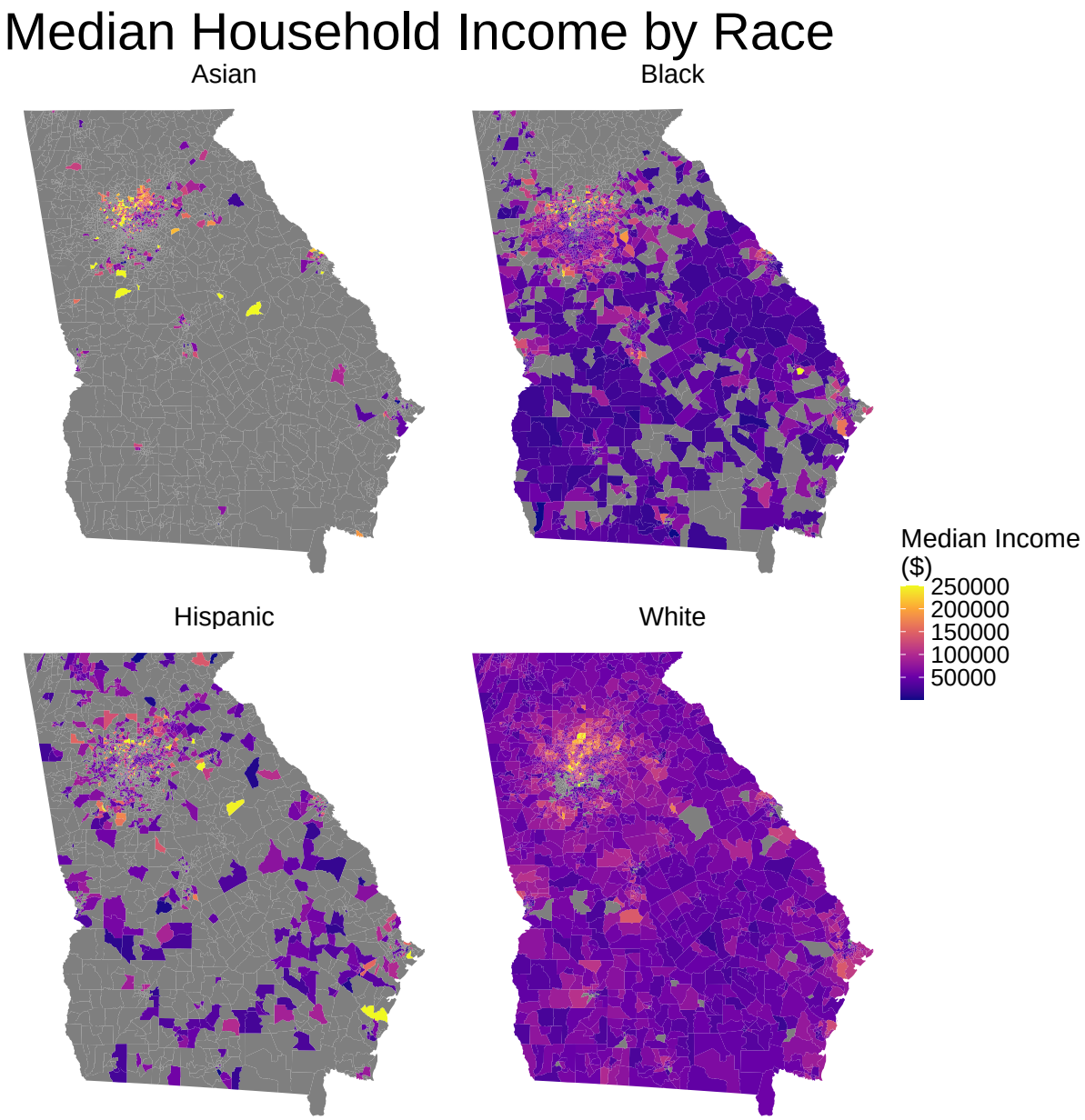

Figure 5: Median Household Income by Race in Georgia (2021 ACS 5-year)

A closer examination of the plots of burdened homeowners and renters by race (Figures 6 and 7) accentuates the racial inequity present in Figures 5. These plots display the mean percentage of cost-burdened owners and renters across Georgia's 159 counties.Among homeowners, Black homeowners and renters account for the greatest mean percentage of cost burdened housing. In contrast White homeowners have the lowest mean percentage and White renters and Other renters have about the same mean. These findings suggest that there may exist systemic barriers, such as wage gaps and discriminatory housing practices, that disproportionately affect minority communities. Particularly, renters face rising challenges due to increasing rental costs and a lack of affordable housing options in both rural and metropolitan areas.

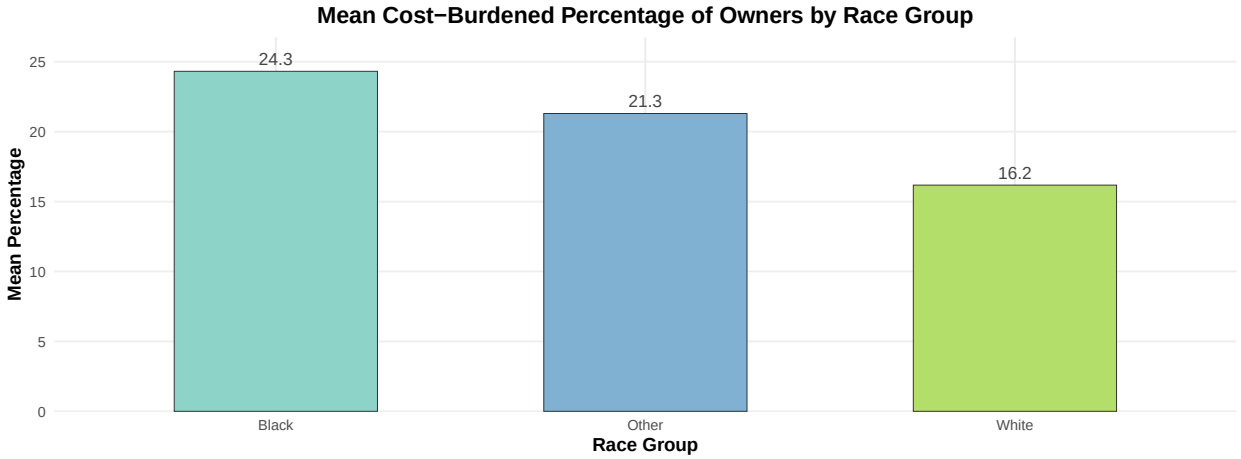

Figure 6: Mean Percentage of Cost-Burdened Owner-Occupied Households by Race in Georgia (ACS 5-Year Estimates, 2019-2023)

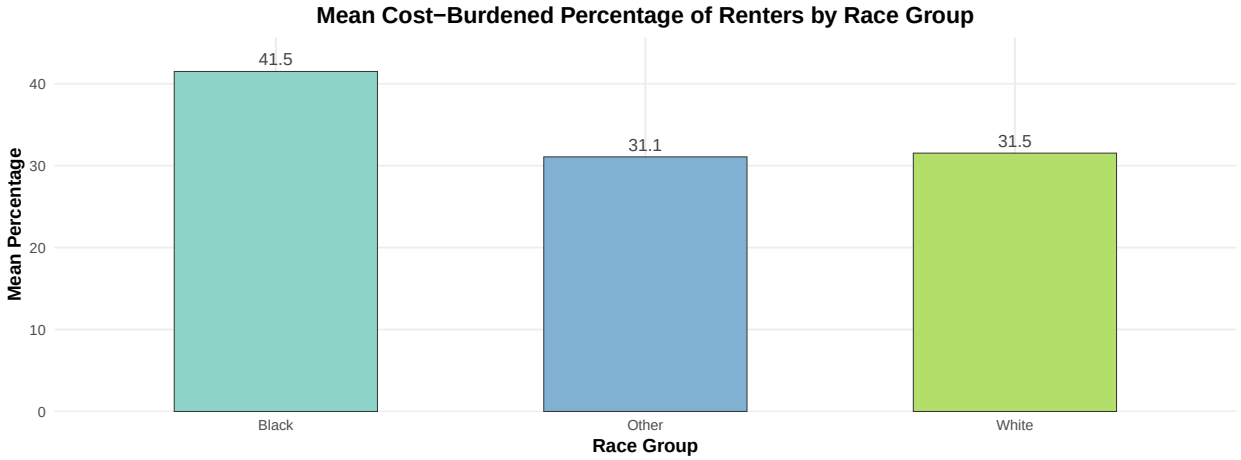

Figure 7: Mean Percentage of Cost-Burdened Renter Occupied Households by Race in Georgia (ACS 5-Year Estimates, 2019-2023

The addition of age adds another layer of complexity to the analysis. As seen in the plots of the mean percentage of cost-burdened homeowners and renters by age across Georgia's 159 counties (Figure 8 & 9), households of those aged 15–24 experience the greatest levels of cost burden, for both renters and homeowners. A simple explanatory hypothesis is that lower ages correlate with a lower income, specifically amongst entry-level workers. Homeowners over the age of 65 also face greater levels of housing burden. This is possibly a result of fixed retirement incomes, as well as rising property taxes.

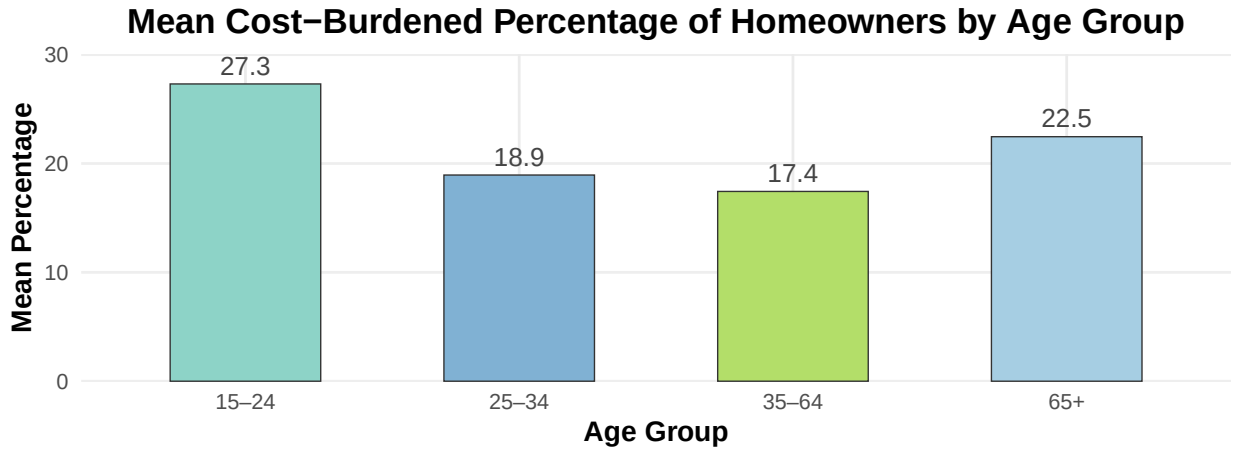

Figure 8: Mean Percentage of Cost-Burdened Owner-Occupied Households by Age in Georgia (ACS 5-Year Estimates, 2019-2023)

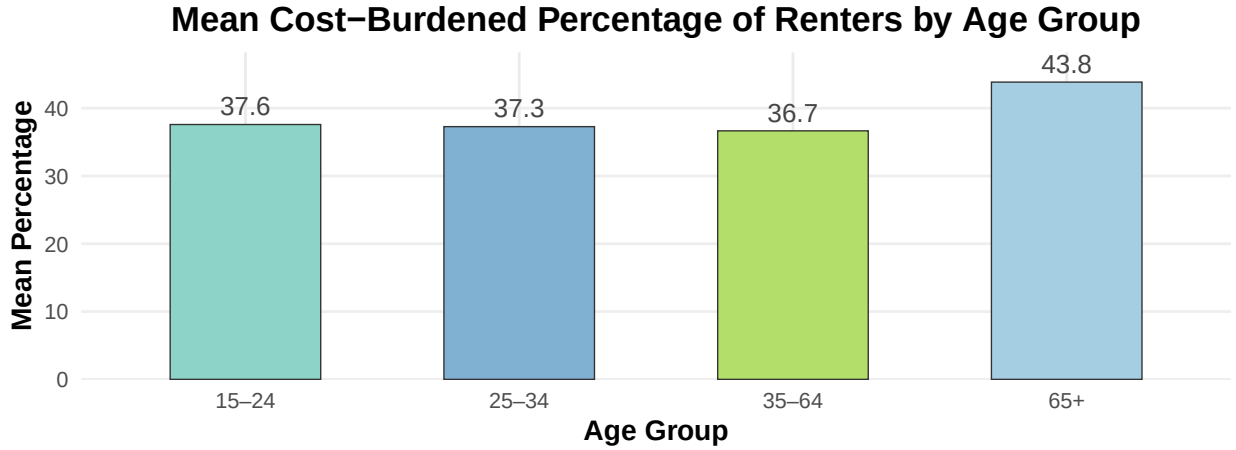

Figure 9: Mean Percentage of Cost-Burdened Renter-Occupied Households by Age in Georgia (ACS 5-Year Estimates, 2019-2023)

To better understand the nature of the selected data, a correlation heatmap was created (Figure 10) using only the continuous features that will be included in the regression models. Most of the variables do not exhibit strong correlations with one another. However, there is a notably strong positive correlation of 0.81 between Pct_Bachelors_Higher and MedHouseIncome. This relationship is expected, as it is well documented that individuals with a bachelor's degree or higher tend to have higher incomes on average.

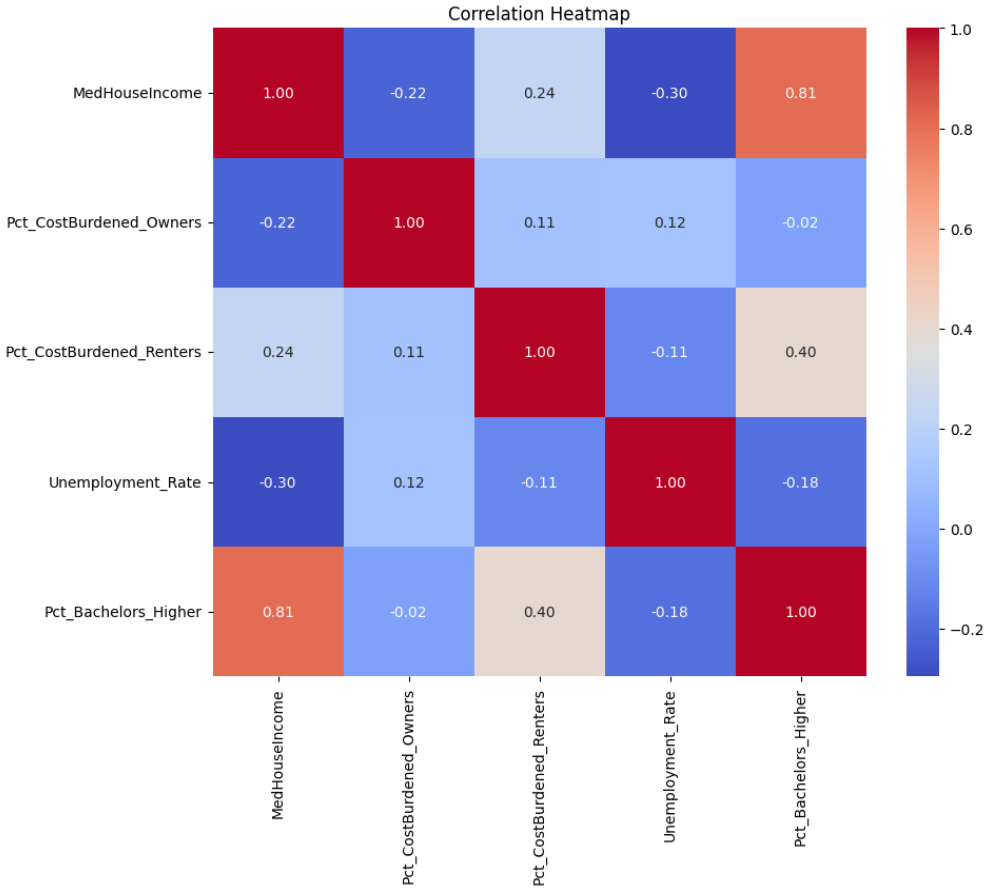

Figure 10: Correlation Heatmap

The initial analysis of the census datum reveal that rising housing cost burden in Georgia may come as a result of a complex, multivariate interaction of variables such as income, race, geography, age, etc. While less populated areas face housing cost challenges due to low incomes, it is evident that highly populated areas are also becoming increasingly impacted by housing costs. Minorities, mainly Black and Hispanic households, are disproportionately affected by housing cost burden, indicating potential systemic inequities. Younger and older renters/homeowners are also vulnerable to these challenges, suggesting that focused intervention may be necessary to tackle the growing issue of housing affordability. The package gridExtra (Auguie, 2017) was used for arranging multiple plots on a single panel.

Methods

To answer the questions posed by clients, several tools were utilized to better understand the relationships between economic, geographic, and demographic predictor variables in the context of predicting cost-burden percentage for both owner-occupied households and renter-occupied households. The analysis was conducted at the county level for the state of Georgia, meaning each observation in the dataset represented a specific county, allowing for the assessment of regional trends and differences.

Multiple Linear Regression

The analysis began with the application of a multiple linear regression model to predict the cost-burden percentage for owner-occupied households. The model equation is:

$$\begin{aligned}\text{Pct\_CostBurdened\_Owners}_i = \beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i \\ + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i + \beta_4 \cdot \text{Unemployment\_Rate}_i + \epsilon_i\end{aligned} \tag{1}$$

where *Pct_CostBurdened_Owners* is the dependent variable and *isUrban*, *MedHouseIncome*, *Pct_Bachelors_Higher*, and *Unemployment_Rate* are the predictor variables defined in Table 1.

To address potential heterodescasticity present within the model, the Breusch-Pagan test conducted as a diagnostic check. Upon detecting evidence of heterodescasticity, a log-transformation on the dependent variable was used to stabalize the variance in the residuals. Additionally, cross-fold valuation was employed to compare performance between the original multiple linear regression model and the log-transformed model. Root means squared error and prediction accuaracy were metrics used to evaluate performance for model selection. The log-transformed model equation is:

$$\begin{aligned}\log(\text{Pct\_CostBurdened\_Owners}_i) = \beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i \\ + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i + \beta_4 \cdot \text{Unemployment\_Rate}_i + \epsilon_i\end{aligned} \tag{2}$$

where *log(Pct_CostBurdened_Owners)* is the dependent variable and *isUrban*, *MedHouseIncome*, *Pct_Bachelors_Higher*, and *Unemployment_Rate* are the predictor variables defined in Table 1.

Once the conditions were met — including checking for assumptions for model-fitting (such as linearity,

independence, normality of residuals, and constant variance within residuals) — backward selection utilizing AIC criterion was used to assess variable selection. After deciding on a final model, the significance of each predictor on the dependent variable was evaluated using p-values at a 5% level of significance. Additionally, multicollinearity was checked using Variance Inflation Factors (VIF). The effects, whether positive or negative, as well as the extent to which the owner-occupied cost-burden percentage was affected, were analyzed in terms of the coefficient values of the predictors. Model performance was measured using R-squared and adjusted R-squared values, and possible interaction effects between predictor variables were tested to determine any combined influences on cost burden.

The second model applied was a multiple linear regression model analyzing the effects of the previously mentioned predictors on the cost-burden percentage for renter-occupied households. After fitting the model, conditions were checked similarly to the previous analysis. The model failed the constant variance condition, due to the presence of a statistically significant p-value at a 5% level of significance using the Breusch-Pagan test. To attempt to fix the problem of heteroscedasticity present within the model, an outlier point was removed from the data set and the model was refitted. This did not fix the problem of heteroscedasticity, due to the presence of a more statistically significant p-value using the Breusch-Pagan test. The renter cost-burden model equation is:

$$\begin{aligned} \text{Pct\_CostBurdened\_Renters}_i = \beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i \\ + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i + \beta_4 \cdot \text{Unemployment\_Rate}_i + \epsilon_i \end{aligned} \tag{3}$$

where *Pct_CostBurdened_Renters* is the dependent variable and *isUrban*, *MedHouseIncome*, *Pct_Bachelors_Higher*, and *Unemployment_Rate* are the predictor variables defined in Table 1.

A log-transformation was applied to the original multiple linear regression model, without removing the outlier point. The log-transformed model failed the Breusch-Pagan test due to a statistically significant p-value at a 5% level of significance, indicating heteroscedasticity. Additionally, the normality condition was violated, as evidenced by skewness in the histogram of residuals and the failure of the Shapiro-Wilk test, where the p-value was below a 5% level of significance, indicating a rejection of the null hypothesis of normality. An attempt was made at refitting the log-transformed model by removing the outlier point, but it still resulted in the same conclusion and did not improve on checking regression assumptions. The log-transformed renter cost-burden model equation is:

$$\log(\text{Pct\_CostBurdened\_Renters}_i) = \beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i + \beta_4 \cdot \text{Unemployment\_Rate}_i + \epsilon_i \tag{4}$$

where *log(Pct_CostBurdened_Renters)* is the dependent variable and *isUrban*, *MedHouseIncome*, *Pct_Bachelors_Higher*, and *Unemployment_Rate* are the predictor variables defined in Table 1.

Weighted Least Squares Regression

To address this problem, an ordinary least squares model was deemed inappropriate due to the potential for biased and inaccurate coefficient estimates. Instead, a weighted least squares regression model was applied, which adjusts for heteroscedasticity by assigning different weights to different observations based on a chosen factor. The *isUrban* variable was selected as the weighting factor, as going from a rural county to a urban county significantly increased cost-burden for renter-occupied households. This approach also had a significant impact on the coefficients for the other factors in the model. The application of this method stabilized the heteroscedasticity in the residuals, and the model met all necessary regression assumptions. The weighted least squares model equations are:

$$\text{Pct\_CostBurdened\_Renters}_i = \beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i + \beta_4 \cdot \text{Unemployment\_Rate}_i + \epsilon_i \tag{5}$$

$$\text{Var}(\epsilon_i) = \sigma_j^2, \quad \text{for} \quad j \in \{\text{Urban}, \text{Non-Urban}\} \tag{6}$$

where:

- *isUrban*, *MedHouseIncome*, *Pct_Bachelors_Higher*, and *Unemployment_Rate* are defined in Table 1.
- σ_j^2 represents the distinct error variance for urban and non-urban counties.

After ensuring conditions were met, backward selection was used incorporating BIC criterion for variable selection. Then, after getting a final model, the significance of each predictor variable on the cost-burden percentage for renter-occupied households was evaluated using p-values and confidence intervals. Multicollinearity was checked using Variance Inflation Factors (VIF) to ensure that variables were not correlated.

Model fit was assessed using R-squared and adjusted R-squared values. Additionally, possible interaction effects between predictor variables were tested to determine any combined influences on cost burden.

At the request of the client, separate models were analyzed for predicting cost-burden percentages for renter-occupied and owner-occupied households at the county level. This approach was taken to capture the distinct effects of various factors, as these effects can differ significantly between renters and owners.

Analysis of Variance (ANOVA)

Before conducting statistical analyses, the assumptions underlying analysis of variance (ANOVA) were carefully examined to ensure the validity of the results. Specifically, we assessed the assumptions of normality, independence, and homogeneity of variance. Normality within groups was evaluated using diagnostic plots and the Shapiro-Wilks test, which indicated that the distribution of cost-burdened housing percentages did not significantly deviate from normality. However, Levene's test for homogeneity of variance revealed violations of the equal variance assumption across groups.

Given that normality was reasonably satisfied but homogeneity of variance was not, we selected Welch's ANOVA as the appropriate statistical test. Unlike the standard one-way ANOVA, Welch's ANOVA is robust to unequal variances while still assuming normality, making it a suitable alternative under these conditions.

For the demographic analysis, four separate one-way Welch's ANOVAs were conducted to compare differences in cost-burdened housing percentages across age and race groups, for both owner-occupied and renter-occupied housing. Age groups were categorized as 15–24, 25–34, 35–64, and 65+. For race, categories included White, Black, and Other.

Results

Multiple Linear Regression

The resulting model uses a linear model to predict the log of cost burden of owner-occupied housing using median household income and percent of population with a bachelor's degree as predictors. The model was found to be statistically significant with a p-value of approximately 0.0001. The equation is as follows:

$$\log(\text{Pct\_CostBurdened\_Owners}) = 3.167 - 0.000006972 \cdot \text{MedHouseIncome} + 0.009859 \cdot \text{Pct\_Bachelors\_Higher} \tag{7}$$

For every $1,000 increase in median household income, there's a 0.7% decrease in cost-burdenedness for owner-occupied households. For every 1% increase in people with a bachelor's degree or higher, there's a 0.9908% increase in cost-burden for owner-occupied households. These results suggest that higher median household income is associated with reduced housing cost burden, while a higher percentage of individuals with a bachelor's degree or higher is associated with increased housing cost burden among owner-occupied households.

Table 2: Coefficients for Final Log Model for Owner Occupied Cost Burden

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.167	0.063	50.427	0
MedHouseIncome	-0.00001	0.00000	-4.363	< 0.001
Pct_Bachelors_Higher	0.010	0.003	3.537	0.001

Table 3: Summary Statistics for Final Log Model for Owner Occupied Cost Burden

	Statistic	Value
1	R-squared	0.109
2	Adj. R-squared	0.097
3	F-statistic	9.517
4	p-value	< 0.001

Weighted Least Squares Regression

A similar multiple linear regression model was fit to predict the percentage of renter-occupied cost-burdened households based on the same predictors as the owner-occupied model. However, after examining the residuals through a Breusch-Pagan test, there is evidence of heteroscedasticity, BP = 11.52, p = 0.021.

Table 4: Breusch-Pagan Test Results

	Method	Statistic	df	p.value
BP	studentized Breusch-Pagan test	11.519	4	0.021

Several transformations, such as removing residual outliers and log transforming the data, were performed to address challenges with non-constant variance in the residuals, but strong evidence of heteroscedasticity remained. Thus, a weighted least squares (WLS) model was deployed instead. It was hypothesized that variability of percent renter-occupied cost-burdened households may differ across urban and rural counties, which is reflected by the chosen weighting function. (Several different weighting options were tested, and weighting based on urban or rural county provided the most reliable performance, while meeting all underlying assumptions). It should also be noted that the unemployment rate predictor was removed to reduce multicollinearity and achieve a better BIC (decreased from 1151.61 to 1148.88). Cross validation and backward selection were employed to produce the final model.

$$\begin{aligned} \text{Pct\_CostBurdened\_Renters} = 28.705 &+ 9.429 \cdot \text{isUrban} \\ &- 0.00015 \cdot \text{MedHouseIncome} \\ &+ 0.351 \cdot \text{Pct\_Bachelors\_Higher} \end{aligned} \tag{8}$$

where:

$$Var(\epsilon_i) = \begin{cases} \sigma^2 & \text{if } isUrban_i = 0 \\ \theta^2\sigma^2 & \text{if } isUrban_i = 1 \end{cases} \tag{9}$$

Thus, the model minimizes the expression:

$$\sum_{i=1}^{n} \left(\frac{1}{\text{Var}(\epsilon_i)} \right) \left(y_i - \Big[\beta_0 + \beta_1 \cdot \text{isUrban}_i + \beta_2 \cdot \text{MedHouseIncome}_i + \beta_3 \cdot \text{Pct\_Bachelors\_Higher}_i \Big] \right)^2 \tag{10}$$

where:

$$w_i = \begin{cases} \frac{1}{\sigma^2} & \text{if } isUrban_i = 0 \\ \frac{1}{\theta^2 \sigma^2} & \text{if } isUrban_i = 1 \end{cases} \tag{11}$$

This weighting criterion ensures that the isUrban factor level with larger residual variance gets less weight in the estimation.

Specifically, weighting by urban or rural counties allows the model to fix a baseline standard deviation of 1.00 for rural counties, while the standard deviation for urban areas was assigned to 0.65. This effectively handles heteroscedasticity by reducing residual variation within urban counties, in comparison with rural counties. Examining the residuals for the WLS model with a Breusch-Pagan test, the heteroscedasticity was properly handled, BP = 0.10235, p-value = 0.797.

Table 5: Breusch-Pagan Test Results

	Method	Statistic	df	p.value
BP	studentized Breusch-Pagan test	0.066	1	0.797

Table 3 shows the coefficients, standard errors, t-values, and p-values for the final WLS model. All of the predictors are statistically significant. As a result, the null hypothesis (there is no correlation between individual predictors and the percentage of renter-occupied cost-burdened households) can be rejected. Additionally, when median household income and education level are held constant, urban counties display a 9.43% increase in renter-occupied housing cost burden, relative to rural counties. This supports the original hypothesis that urban counties are associated with greater percentages of renter-occupied cost-burdened households. Furthermore, every $1000 increase in median household income correlates with a 0.153% decrease in renter-occupied housing cost burden, when all other variables are held constant. In practical

terms, greater median incomes are linked with a slightly lower percentage of renter-occupied cost-burdened households. Another notable finding is that every 1% increase in the population holding bachelor's degrees correlates with a 0.35% increase in renter-occupied housing cost burden. This finding provides support against a logical hypothesis that greater percentages of formal higher education in a population correlates with a lower percentage of households being labeled as cost-burdened.

Table 6: Weighted Least Squares Model for Rent Burden

	Dependent variable:
	Pct_CostBurdened_Renters
Urban (isUrban = 1)	9.429
	(1.479)
	t = 6.377
	p = <0.001
Median Household Income	−0.0002
	(0.0001)
	t = −2.971
	p = <0.001
Percent with Bachelor's or Higher	0.351
	(0.094)
	t = 3.741
	p = <0.001
(Intercept)	38.134
	(2.235)
	t = 17.066
	p = <0.001
Observations	159
Log Likelihood	−559.308
Akaike Inf. Crit.	1,130.616
Bayesian Inf. Crit.	1,148.877
Note:	GLS model weighted by varIdent(form = 1 | isUrban)

Welch's ANOVA Results

A Welch's one-way ANOVA was conducted to examine mean differences in the percentage of cost-burdened housing across age groups (15–24, 25–34, 35–64, and 65+) for both owner-occupied and renter-occupied housing. The analysis was performed at a significance level of $\alpha = 0.05$.

For owner-occupied housing, this analysis revealed a statistically significant effect of age group on the percentage of cost-burdened housing, $F = 23.43$, $p < .0001$. This shows that the mean percentage of cost-burdened housing differs meaningfully between at least one of the age groups, based on statistically significant evidence. Pairwise comparisons using the Games-Howell test indicated that the percentage of cost-burdened housing for the age groups 15-24 differed significantly with the age groups 25-34 and 35-64 with p-values of .00504 and .00027 respectively. Age group 65+ also differed significantly with age groups 25-34 and 35-64 with p-values of .00367 and <.001 respectively. Age groups 25-34 and 35-64 did not differ significantly with a p-value of .43913. Age groups 15-24 and 65+ did differ significantly with a p value .00504. Due to age groups 15-24 and 65+ having higher means than age groups 25-34 and 35-64 indicates that ages 15-24 and 65+ are more like to experience cost-burden and since age 15-24 has the highest mean its the most likely to experience it.

For renter-occupied housing, Welch's ANOVA also revealed a significant effect of age group on the percentage of cost-burdened housing, $F = 7.51$, $p < 0.001$. This indicates that the mean percentage of cost-burdened housing differs significantly between at least one of the age groups. Pairwise comparisons using the Games-Howell test indicated that the percentage of cost-burdened housing for age group 65+ differed significantly with the age groups 15-24, 25-34, and 35-64 with p values .0431, .0021, and < 0.001 respectively. No significant differences were observed between other age groups. This suggests that the 65+ renters are more likely to experience cost-burdened housing compared to younger renters as they had a higher mean (see Tables 7, and 8,9,10 for details).

Table 7: Welch's ANOVA Results for Cost Burden by Age Group (Owners)

	Source	DF	F_value	p_value
1	Between Groups	3	23.429	<0.001
2	Residuals	330.1600		

Table 8: Games Howell Test Results for Age Groups (Owners)

Comparison	P_Value
15-24 vs 25-34	0.005
15-24 vs 35-64	< 0.001
15-24 vs 65+	0.175
25-34 vs 35-64	0.439
25-34 vs 65+	0.004
35-64 vs 65+	< 0.001

Table 9: Welch's ANOVA Results for Cost Burden by Age Group (Renters)

	Source	DF	F_value	p_value
1	Between Groups	3	7.5073	<0.001
2	Residuals	340.1400		

Table 10: Games Howell Test Results for Age Groups (Renters)

Comparison	P_Value
15-24 vs 25-34	0.999
15-24 vs 35-64	0.974
15-24 vs 65+	0.043
25-34 vs 35-64	0.979
25-34 vs 65+	0.002
35-64 vs 65+	< 0.001

Welch's ANOVA was also performed to examine differences in the percentage of cost-burdened housing across race groups for both owner-occupied and renter-occupied housing. As with the age groups, Welch's ANOVA was used due to violations of the assumption of equal variances. The analysis was conducted at a significance level of $\alpha = 0.05$.

For owner-occupied housing, the analysis indicated a significant effect of race group on the percentage of cost-burdened housing, F = 31.20, p < .001. This suggests that the mean percentage of cost-burdened housing differs significantly across race groups. Pairwise comparisons using the Games-Howell test showed that the Black group differed significantly from the White group with a p-value <.0001 and did not differ significantly from the Other group with a p-value of .365. Additionally, the White group was found to differ significantly from the Other group with a p value of .031. Also, the Black group had the highest mean followed by the Other group which indicates that Black owners and Other owners are more likely to experience cost-burden than White owners.

For renter-occupied housing, Welch's ANOVA also showed a significant effect of race group on the percentage of cost-burdened housing, F = 18.08, p < .001. This suggests that the mean percentage of cost-burdened housing differs significantly across race groups for renters. Pairwise comparisons using the Games-Howell test indicated that the Black group differed significantly from the White and Other groups with p-values <.0001 and .0011 respectively. It also showed the White group did not differ significantly from the Other group with a p-value of .9832. These results indicate that Black renters are more likely to experience cost-burden than the White and Other groups as they have a higher mean (see Tables 11,12,13,14 for details)

Table 11: Welch's ANOVA Results for Cost Burden by Race Group (Owners)

	Source	DF	F_value	p_value
1	Between Groups	2	31.201	<0.001
2	Residuals	235.1700		

Table 12: Games Howell Test Results for Race Groups (Owners)

Comparison	P_Value
White vs Black	<0.001
White vs Other	0.031
Black vs Other	0.365

Table 13: Welch's ANOVA Results for Cost Burden by Race Group (Renters)

	Source	DF	F_value	p_value
1	Between Groups	2	18.078	<0.001
2	Residuals	267.9100		

Table 14: Games Howell Test Results for Race Groups (Renters)

Comparison	P_Value
White vs Black	<0.001
White vs Other	0.9832
Black vs Other	0.0011

Conclusion

This analysis of housing cost burden in Georgia revealed several meaningful insights into the geographical, economic, and demographic factors that influence cost burden for both owner-occupied and renter-occupied households. Remarkable findings include that urban counties exhibit a 9.43% increase in cost-burden for renter-occupied households as compared to rural counties. Median household income and education levels are also key drivers for influencing cost-burden for owner-occupied households: higher household income levels correlate with a slight reduction in cost-burden (0.0007%), while unexpectedly higher education levels correlate with a slight increase in cost-burden (0.9908%). Meanwhile, for renter-occupied households, higher education levels also correlate with a slight increase in cost-burden (0.351%). These findings suggest that while higher education may lead to higher paying jobs, rising housing costs in urban counties may be offesetting gains in earning potential.

The analysis revealed that both age and race are significantly associated with differences in the mean percentage of cost-burdened housing for owner-occupied and renter-occupied households. For owner-occupied housing, younger adults (15–24) and older adults (65+) were significantly more likely to be cost-burdened compared to middle-aged groups (25–34) and (35-64), with the 15–24 age group showing the highest cost burden. Racial disparities were also present, with Black and Other homeowners experiencing higher cost burdens than White homeowners. For renter-occupied housing, adults aged 65+ were significantly more likely to be cost-burdened compared to all other age groups. Looking back into race, Black renters showed a significantly higher cost burden than both White and Other groups, highlighting a notable disparity.

These methodological approaches, including both weighted least squares regression and Welch's ANOVA, effectively addressed challenges such as heterodescasticity and unequal variances, allowing for more robust

results. These models suggest need for intervention from policy-makers, such as expanding affordable housing options for more cost-burdened individuals in urban areas, implementing rent stabilization measures, and addressing wage discrepancies. Further research could explore additional variables or ways to handle spatially correlated data, based on county level, to further refine these insights.

In summary, this analysis provides practical evidence for policymakers to address housing insecurity in Georgia, emphasizing urgent need for equitable solutions specifically tailored to most affected demographics (i.e. urban renters, Black households, and vulnerable age groups). By addressing these disparities, Georgia can move toward a better housing future.

References

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. https://CRAN.R-project.org/package=gridExtra

Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Walker, K. (2024). *Tigris: Load census TIGER/line shapefiles*. https://CRAN.R-project.org/package=tigris

Walker, K., & Herman, M. (2024). *Tidycensus: Load US census boundary and attribute data as 'tidyverse' and'sf'-ready data frames*. https://CRAN.R-project.org/package=tidycensus

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Appendix

The R code for this project is available on Google Drive:

[Click here to view the R script] (https://drive.google.com/file/d/1XfGuoredDW2b447PLKGmWomd1ExPw4Op/view?usp=share_link)