Appendix C: Data & Methods

The purpose of this research was to determine whether racial inequalities exist in the cost of homeowners insurance and, if inequalities do exist, where in the country they are starkest and whether these inequalities can be explained by third variables. The units of observation for this study are ZIP codes; this means that this study examines inequalities in the cost of homeowners insurance at the neighborhood - not individual - level.

The dependent variable in this analysis is the average homeowners insurance premium in a ZIP code. To calculate average homeowners insurance premiums, we use proprietary data purchased from Quadrant Information Services in August 2024. The Quadrant data consists of roughly six "test quotes" (216,151 quotes) for nearly every residential ZIP code in the United States (32,978 ZIPs) built from insurer rate filings in each state. These test quotes represent what a homeowner with certain characteristics would be charged for a given policy to insure a given home. We fix over 150 characteristics of the homeowner, home and policy (for example by taking an average credit score, a $350,000 replacement value, and a 20-year old home) to best represent a typical homeowner for each test quote. These data offer the unique advantage of allowing for like-to-like comparisons of what a typical homeowner in any given location would pay.

The quotes come from roughly six of the largest home insurers in each state (and DC), representing 57 percent of the national homeowners insurance market. All of the insurers were private except for Citizens Property Insurance Corporation in Florida, which was included due to its large share of the Florida market. The test quotes were averaged within each ZIP, weighted by state-level insurer market share, to obtain a single average premium measure for each ZIP. All national and state-level averages are weighted by the number of households in a ZIP.

Another possible dataset to calculate average homeowners insurance premiums in each ZIP code is the supplemental data from the *Analyses of U.S. Homeowners Insurance Markets, 2018-2022: Climate-Related Risks and Other Factors* report released by the Federal Insurance Office in January 2025. This dataset is a first-of-its-kind data collection effort conducted by the National Association of Insurance Commissioners (NAIC) and contains ZIP-code level data on key variables like average premiums paid each year from 2018-2022. However, the data call was voluntary and not every state participated. The dataset only contains 77 percent of the ZIP codes in the Quadrant dataset. Even if every state did participate, the Quadrant data is preferable because all quotes assume identical homes, policyholder profiles, and coverage levels. This offers the strongest like-with-like comparisons and means that the racial premium gap estimated in this analysis is based on premium inequalities between geographic location alone.

The independent variable in this analysis is the racial composition of a ZIP code. Data on racial composition come from the 2024 American Community Survey (ACS) 5-year estimates Table 25003. This table contains information on the count of households (number of owner-occupied units plus the number of renter-occupied units) for a given geographic unit (in this case ZIP Code Tabulation Areas, known as ZCTA) broken down by race and ethnicity: White Alone, Black or African American, American Indian or Alaska Native Alone, Asian Alone, Native Hawaiian or Pacific Islander Alone, Some Other Race Alone, Two or More Races, White Alone Not Hispanic or Latino, and Hispanic or Latino. In this analysis we primarily focus on White Alone Not Hispanic or Latino ("white"), Black or African American ("Black"), and Hispanic or Latino ("Hispanic") due to small counts and clustering in certain states for the other six groups. However, the racial premium gap is estimated in the regression models for Asian Alone ("Asian"), American Indian or Alaska Native Alone ("AIAN"), and the remaining groups (collapsed into "Other").

We categorize a ZIP code as the same race as the racial group with the largest number of households in that ZIP. For example, if white households make up 37 percent of the households in a ZIP - and no other racial group has more than 37 percent - the ZIP code is classified as "white." If two racial groups are tied for the largest number of households, the race of the ZIP code is classified as "Tied" and then collapsed into the "Other" category. Measuring race in this way results in 28,756 "white," 1,364 "Hispanic," and 1,444 "Black" ZIP codes.

For the second finding, which explored the geography of the racial premium gap, we restrict the analysis to states with at least five predominantly Black ZIP codes (for the Black racial premium gap) and at least five predominantly Hispanic ZIP codes (for the Hispanic racial premium gap). Twenty-eight states have at least five predominantly Black ZIPs and 17 states have at least five predominantly Hispanic ZIPs. Restricting the analysis in this way ensures that the state-level racial premium gaps that we calculate are not just driven by random noise in one or two ZIP codes.

Our third finding examines whether the association between a ZIP code's predominant race and average homeowners insurance premium is actually due to a third factor. For example, Black and Hispanic communities on average have a much greater level of housing density (see **Appendix Figure D-2**), which might prove to explain the higher premiums in these communities. To control for this and other factors, we construct a series of regression models. This methodological approach does not prove a causal link between neighborhood racial composition and the cost of homeowners insurance. Rather, the approach preempts concerns that the disparities found in the first two findings are simply the result of "race-neutral" third factors. Our findings pose the question to the insurance industry: *if not race or the variety of risk factors that we control for, then what is causing the racial premium gap?*

Three regression models are estimated, each one adding control variables to the previous specification. All models are multi-level, with a random intercept for state, to account for the fact that ZIP codes within the same state resemble each other more closely than ZIP codes between states (as insurance is regulated at the state-level). The Quadrant data is set-up so that each ZIP code only belongs to one state. In all models, premiums are log-transformed to account for their right-skewed distribution. All models are weighted by the number of households in a ZIP code (scaled by the mean number of households).

To avoid dropping any observations, missing values are imputed using Multivariate Imputation by Chained Equations (MICE). Most of the missing data stem from the two control variables from the Federal Insurance Office dataset, claim frequency and claim severity (described below), which have missing values for 23 percent of ZIP codes. Missing data on the remaining variables are minimal, if any. The dataset is imputed 25 times using the Predictive Mean Matching (PMM) method. The coefficients, standard errors, test statistics, and p-values are all pooled following Rubin's rules.

Model 1 regresses logged insurance premiums on the ZIP code's predominant race and the random intercept for state. All models assumes that the random intercept is Normally distributed around mean 0 with a variance of 2 and the residuals are Normally distributed around mean 0 with a variance of 2. The formula for Model 1 for ZIP code i in state j and predominant race k in {1=Black; 2=Hispanic; 3=Asian; 4=AIAN; 5=Other} is:

$$\ln(\text{Premium})_{ij} = \beta_0 + \sum_{k=1}^{5} \beta_k \text{Race}_{k,ij} + u_j + \epsilon_{ij}$$

Model 2 adds a control for environmental risk. The environmental risk variable comes from the Federal Emergency Management Agency's National Risk Index (FEMA NRI). For each census tract, the Expected Annual Loss (EAL) is summed from the following twelve weather events typically covered by homeowners insurance: avalanche, cold wave, hail, heat wave, hurricane, ice storm, lightning, strong wind, tornado, volcanic activity, wildfire, and winter weather. We then calculate a ZIP code's environmental risk as the average EAL of the census tracts in that ZIP (weighting by number of residents) and log-transform to account for its right-skewed distribution. The formula for Model 2 is:

$$\ln(\text{Premium})_{ij} = \beta_0 + \sum_{k=1}^{5} \beta_k \text{Race}_{k,ij} + \beta_6 \ln(\text{Environmental Risk})_{ij} + u_j + \epsilon_{ij}$$

Model 3 adds four additional variables that insurance companies frequently cite as reasons for charging higher insurance premiums in certain neighborhoods. Housing age was calculated as the median year housing units in a ZIP (ZCTA) were constructed, subtracted from 2024. Housing age was obtained from the 2024 ACS. Housing density was calculated as the number of housing units in a ZIP (ZCTA) divided by the land area of the ZIP (ZCTA). The number of housing units was obtained from the 2024 ACS and the land area obtained from the Census TIGER/Line shapefiles. Annual claim frequency in a ZIP code was obtained from the Federal Insurance Office dataset discussed above, and is defined as, "[t]he percentage of claims insurers have paid relative to the policies they have issued." Annual claim severity in a ZIP code was also obtained from the FIO dataset, and is defined as, "[t]he average amount insurers have paid for each claim." Both claim frequency and claim severity are averaged over all years in the FIO data (2018-2022). Claim severity is in units of $1,000s and housing density was log-transformed. The formula for Model 3 is:

$$\begin{aligned}\ln(\text{Premium})_{ij} = \beta_0 + \sum_{k=1}^{5} \beta_k \text{Race}_{k,ij} + \beta_6 \ln(\text{Environmental Risk})_{ij} + \beta_7 \text{Housing Age}_{ij} + \\ \beta_8 \ln(\text{Housing Density})_{ij} + \beta_9 \text{Claim Frequency}_{ij} + \beta_{10} \text{Claim Severity}_{ij} + u_j + \epsilon_{ij}\end{aligned}$$

As discussed in the **Findings** section, the effect of race is significant ($p < 0.001$) and substantial across all three models, raising concerns about racial bias in how insurance companies price premiums. **Appendix D** below includes tables with the descriptive statistics of the variables in the models (**Figure D-1**), how the averages of the control variables vary by the predominant race of a ZIP Code (**Figure D-2**), and a formal regression output table (**Figure D-3**). All analysis was conducted in R version 4.5.0.

Appendix D: Regression Tables

Figure D-1. Descriptive Statistics for Variables in Regression Models

	Mean/Proportion	Std. Dev.	Min.	Max.
Dependent Variable				
Logged Annual Insurance Premium	7.93	0.49	6.56	10.84
Independent Variables				
ZIP Predominant Race				
White (Baseline)	0.8720		0	1
Black	0.0438		0	1
Hispanic	0.0414		0	1
Asian	0.0055		0	1
AIAN	0.0105		0	1
Other	0.0268		0	1
Control Variables				
Logged Environmental Risk (EAL)	12.21	1.36	4.96	17.20
Housing Age (yrs)	48.06	16.85	3.00	86.00
Logged Housing Density (units/sq. mile)	3.88	2.36	-7.67	11.58
Claim Frequency (%)	5.52%	3.78%	0.00%	68.77%
Claim Severity (in $1,000s)	$19.24	$20.41	$0.00	$1,150.70
Weight Variable				
Households (Scaled)	1.00	1.46	0.00	11.65

Source: CFA analysis of Quadrant Information Services, American Community Survey, FEMA, and Federal Insurance Office data.

Note: N = 32,978 ZIP codes. Means/proportions and standard deviations pooled across imputed datasets. Minimums and maximums from unimputed data. Means weighted by number of households in ZIP.

Figure D-2. Averages of Control Variables in Predominantly White, Black, and Hispanic ZIP Codes

	Number ZIPs	Environmental Risk (EAL)	Housing Age (yrs)	Housing Density (units/ sq. mile)	Claim Frequency (%)	Claim Severity ($)
White	28756	$442,608	47.97	471.36	5.47%	$18,989
Black	1444	$479,005	49.69	1517.97	7.92%	$16,176
Hispanic	1364	$434,765	47.99	1956.10	5.31%	$20,056

Source: CFA analysis of Quadrant Information Services, American Community Survey, FEMA, and Federal Insurance Office data.

Note: N = 32,978 ZIP codes. Averages weighted by number of households in ZIP and pooled across imputed datasets.

Figure D-3. Multi-level Regression Modeling of Annual Average Premium (Logged) by ZIP Code Predominant Household Race

	Dependent variable:		
	Logged Annual Home Insurance Premium		
	Model 1	Model 2	Model 3
Black (ref: White)	0.115***	0.149***	0.098***
	(0.005)	(0.005)	(0.005)
Hispanic	0.091***	0.136***	0.112***
	(0.005)	(0.004)	(0.004)
Asian	-0.081***	-0.024*	-0.040***
	(0.011)	(0.010)	(0.010)
AIAN	0.071*	0.123***	0.200***
	(0.035)	(0.033)	(0.031)
Other	0.050	0.036	0.022
	(0.068)	(0.062)	(0.061)
Logged Environmental Risk (EAL)		0.084***	0.099***
		(0.001)	(0.001)
Housing Age (years)			0.003***
			(0.0001)
Logged Housing Density (units/sq. mile)			0.024***
			(0.001)
Claim Frequency (%)			0.137**
			(0.052)
Claim Severity (1,000s)			0.002***
			(0.0001)
Constant	7.822***	6.809***	6.350***
	(0.067)	(0.062)	(0.064)
Number of Clusters (States)	51	51	51
sd(State)	0.482	0.438	0.444
Log-likelihood	-22295	-19682	-18216
AIC	44606	39381	36457
BIC	44674	39457	36566
Observations	32978	32978	32978
Note:	*p<0.05; **p<0.01; ***p<0.001		

Source: CFA analysis of Quadrant Information Services, American Community Survey, FEMA, and Federal Insurance Office data.

*Note: N = 32,978 ZIP codes. Fixed effect parameters, standard errors, and p-values pooled across 25 imputed datasets. State random effect standard deviation, log-likelihood, AIC, and BIC stable across all imputed datasets and reported values are from representative first imputed dataset. All models weighted by number of households in ZIP. Table built using stargazer package in R. * p < 0.05; ** p < 0.01; *** p < 0.001.*

References

45. The market share in each state varies based on the insurers included, ranging from 35 percent to 82 percent coverage. The 57 percent figure represents the average market coverage across all states weighted by the number of homeowners in each state.

46. National Association of Insurance Commissioners. "2023 Market Share Reports for Property/Casualty Groups and Companies by State and Countrywide." August 2024. https://naic.soutronglobal.net/Portal/Public/en-GB/DownloadImageFile.ashx?objectId=10643&ownerType=0&ownerId=4923.

47. Data on the number of households in each ZIP (ZCTA) come from the 2024 ACS 5-year estimates Table B25003. We define number of households as the number of owner-occupied units plus the number of renter-occupied units in a ZIP.

48. The dataset was deleted from the Federal Insurance Office's website in 2025 as part of a broader data suppression effort by the Trump administration. However, CFA saved a local copy and made it available for download here: https://consumerfed.org/federal-insurance-report-details-rising-homeowners-insurance-costs-from-2018-to-2022/.

49. Weiland, Ethan and Michael DeLong. "Homeowners Deserve Transparency on Rising Insurance Costs." Consumer Federation of America (blog), June 12, 2025. https://consumerfed.org/homeowners-deserve-transparency-on-rising-insurance-costs/.

50. These data are available to download at: https://data.census.gov/table/ACSDT5Y2024.B25003?q=B25003

51. Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

52. https://www.fema.gov/flood-maps/products-tools/national-risk-index

53. ACS Table B25037 (https://data.census.gov/table/ACSDT5Y2024.B25037).

54. ACS Table B25001 (https://data.census.gov/table/ACSDT5Y2024.B25001). For more information and to download TIGER/Line shapefiles see: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html.

55. Hlavac, Marek. 2022. "stargazer: Well-Formatted Regression and Summary Statistics Tables." R package version 5.2.3. https://CRAN.R-project.org/package=stargazer