An Exploration and Analysis of Housing Cost Burden Data in California from 2006-2010

Jack Wright

The data used to conduct this report was sourced from the United States Government’s open data website, and the data was originally traced back to the State of California’s data portal. The dataset compiled multiple demographic variables and numerical variables on housing cost burden in CA. Housing cost burden, interchangeably called rent burden, is when a person’s housing costs exceed 30% of their income. This has become a serious issue across the world recently, however in California it has been a large issue before most people were aware of it. The relevant categorical variables in the dataset include a race demographic variable, a region of California variable, a county variable, a tenure variable for owning vs renting, and variables on the total number of households, the number of rent burdened households, and the % of households that are rent burdened. The dataset also contains additional variables that were cleaned or omitted because of issues or their lack of usefulness.

The rest of the report is organized as follows: section 1 details the data preprocessing and cleaning necessary to ensure the dataset is workable and useable for advanced exploratory data analysis. Section 2 is an overview of the exploratory data analysis results where each relevant variable is examined and relationships between the variables are analyzed. Section 3 will include visualizations made using Power BI.

**Section 1: Data Preprocessing and Cleaning**

All original data preprocessing and cleaning was conducted using R. When initially viewing the dataset, it contained 521,263 rows of data, and 26 variables. Among these, many were categorical variables that had equivalent numerical counterparts. This included variables specifying the race of the household, the geography the household was in, and the region and county the household was in. This allowed me to omit the 4 redundant numerical variables from the dataset. Next, there were variables that were found to be irrelevant to data analysis, including a variable stating where the data was sourced from, a variable stating the years the data included (every piece of data was aggregated from 2006-2010), lower level and upper level confidence interval variables, standard error and random standard error variables, a California population decile variable, a variable comparing local rank to statewide rank, and lastly a variable stating the completion date of the dataset. These variables were all irrelevant and unhelpful, and as a first step it became necessary to omit and ignore them for future data analysis in R and PowerBI. Additionally, there were many missing values for the number of households and percent variables, so I chose to exclude those observations from the dataset. I next found that there were a few observations for the percentage variable where the percentage was >100, so I excluded these observations from the dataset too. After making these changes the dataset was left 6 variables and 173,971 observations, a significant number. With the new dataset, there are still some missing variables for the region variable, but only for observations for the state totals. There were more missing observations for the county variable as well, specifically for the state totals and the regional totals. Because of this, I did not exclude either of these variables’ missing values. Due to the numerous diverse regions in CA and the structure of the dataset the data was subsetted by the region variable for analysis.

**Section 2: Exploratory Data Analysis**

The relevant variables that can be used to conduct an exploratory data analysis are as follows: a total households and percentage burdened households variable, the race of household variable, the region variable, the geoname variable (for state totals), and the tenure variable.

There are 14 regions which are: Bay Area, Butte, Central/Southeast Sierra, Monterey Bay, North Coast, Northeast Sierra, Northern Sacramento Valley, Sacramento Area, San Diego, San Joaquin Valley, San Luis Obispo, Santa Barbara, Shasta, and Southern California. All associated screenshots and visualizations can be found in the Data Exploration and Visualization Report Image Analysis Index.

The mean percent of total housing cost burdened households in the state of California was 35.59. The mean percent of rent burdened households for each region is listed from highest (worst) to lowest (best): San Diego (38.24), Southern California (38.05), Bay Area(34.22), San Luis Obispo (34.09), Sacramento Area (33.88), Monterey Bay (33.2), Santa Barbara (31.3), Shasta (31.25), Butte (31.1), San Joaquin Valley (31.1), North Coast (29.75), Northern Sacramento Valley (27.24), Northeast Sierra (25.23), Central/Southeast Sierra (22.87). Over 20% of the population being rent burdened is a substantial figure, meaning none of the regions in California would qualify as having a stable and affordable housing market from 2006-2010. Additionally, below are the histograms for the state total and the 3 highest and lowest regions. For the state and 3 highest regions, the largest histogram column is right above 20%, and there is a second peak above 40%. This means that for most areas or demographics the percentage of rent burdened households is 20, and there are many areas where the percentage is 40 or above. For the 3 lowest regions the highest bar is the first one, meaning most areas in those regions have a mode percentage of rent burdened households that is 10 or below. For the lowest regions the histogram bars still extend to 100%, meaning there are still areas or demographics where the percentage of rent burdened households is over 90.

California:

A graph of a graph

Description automatically generated

San Diego:

A graph of a graph

Description automatically generated

Southern California:

A graph of a social level

Description automatically generated with medium confidence

Bay Area:

A graph of a bar graph

Description automatically generated

Northern Sacramento Valley:

A graph of a bar graph

Description automatically generated

Northeast Sierra:

A graph of a number of people

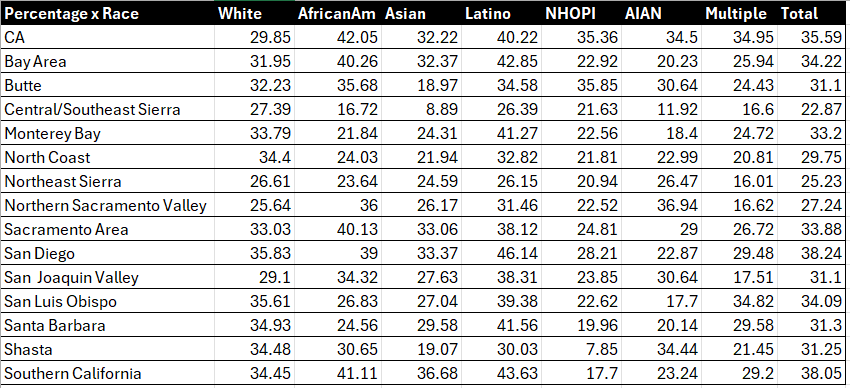
Description automatically generated with medium confidence

Central/Southeast Sierra:

A graph of a number of gray bars

Description automatically generated with medium confidence

The conducted bivariate analysis was done with rent burdened percent across different races subsetted again by location. For the state totals, African American households had the highest mean percentage rent burdened, followed by Latino, households of more than 1 race, Native Hawaiians or Pacific Islanders (NHOPI), Native Americans or Alaskans (AIAN), Asians, and lastly whites had the lowest mean percentage of households that were housing cost burdened. Below is a table showing the means for each race across each region and the state total.



Additionally, below are the boxplots for the 3 highest and lowest regions along with the state totals. Unsurprisingly, for most regions either Latino or African American households experienced the highest rent burden rates, as rising housing costs and a lack of new housing development generally affects marginalized groups and lower income families the most. The general patterns for each race is that regions on the coastline and southern regions tend to have higher percents of rent burdened households across each race. These regions are more densely populated and some reasons for these trends include general higher cost of living by the coast, and the space limitations that come with that, as well as the undeveloped land that mountains, and state and national parks take up. The boxplots below corroborate the results from the previous histograms in that there is generally a large range in the percentages for rent burdened households. Surprisingly NHOPI had the lowest percentage of rent burdened households in any region of California with 7.85% in Shasta. It is possible that this is due to the relatively low number of NHOPI households statewide, and that those who are in California and Shasta in particular are of higher income.

**California:**

A graph of a number of rectangular objects

Description automatically generated

**San Diego:**

A graph with rectangles and lines

Description automatically generated

**Southern California**

A graph of a number of rectangular objects

Description automatically generated with medium confidence

**Bay Area:**

A graph with a number of rectangular objects

Description automatically generated with medium confidence

**Central/Southeast Sierra:**

A graph of a number of people

Description automatically generated with medium confidence

**Northeast Sierra:**

A graph with a number of columns

Description automatically generated with medium confidence

**Northern Sacramento Valley:**

A graph of a number of people

Description automatically generated with medium confidence

Next is an analysis of tenure and the percentage of rent burdened households subsetted once again by region. The tenure variable looks at the type of housing the household is under- renting or owning. We would expect a renting household to be more likely to be rent burdened compared to an owner-occupied household. Below are the boxplots for the state of California and the 3 highest and lowest regions. As can be seen in the boxplots below, on average more renter-occupied households are rent burdened than owner-occupied households. This is true for each of the regions except Central/Southeast Sierra. This is not surprising; however, it is interesting that renters are less likely to be rent burdened in Central/Southeast Sierra than owners.

California:

A graph showing a couple of squares

Description automatically generated with medium confidence

San Deigo:

A graph with a row of squares

Description automatically generated

Southern California:

A graph with a row of squares

Description automatically generated

Bay Area:

A graph with a row of squares

Description automatically generated with medium confidence

Northern Sacramento Valley:

A graph with a chart and a diagram

Description automatically generated with medium confidence

Northeast Sierra:

A graph with a row of squares

Description automatically generated

Central/Southeast Sierra:

A graph with a row of squares

Description automatically generated

Next is the multivariable analysis of the percentage of rent burdened households across different races and tenure subsetted by region. Once again, below are the boxplots of the California totals and the 3 highest and lowest states. Statewide across most different races the average percentages of rent burdened households were lower for owner occupied households than renter occupied households. The same is true for most of the regions below except for the Bay Area, where owner-occupied households and renter-occupied households had very similar percentages of rent burden.

California:

A graph of different colored rectangles

Description automatically generated

San Diego:

A graph of different colored rectangles

Description automatically generated

Southern California:

A graph of colored lines

Description automatically generated with medium confidence

Bay Area:

A graph of different colored lines

Description automatically generated with medium confidence

Northern Sacramento Valley:

A graph of different colored squares

Description automatically generated

Northeast Sierra:

A graph with different colored squares

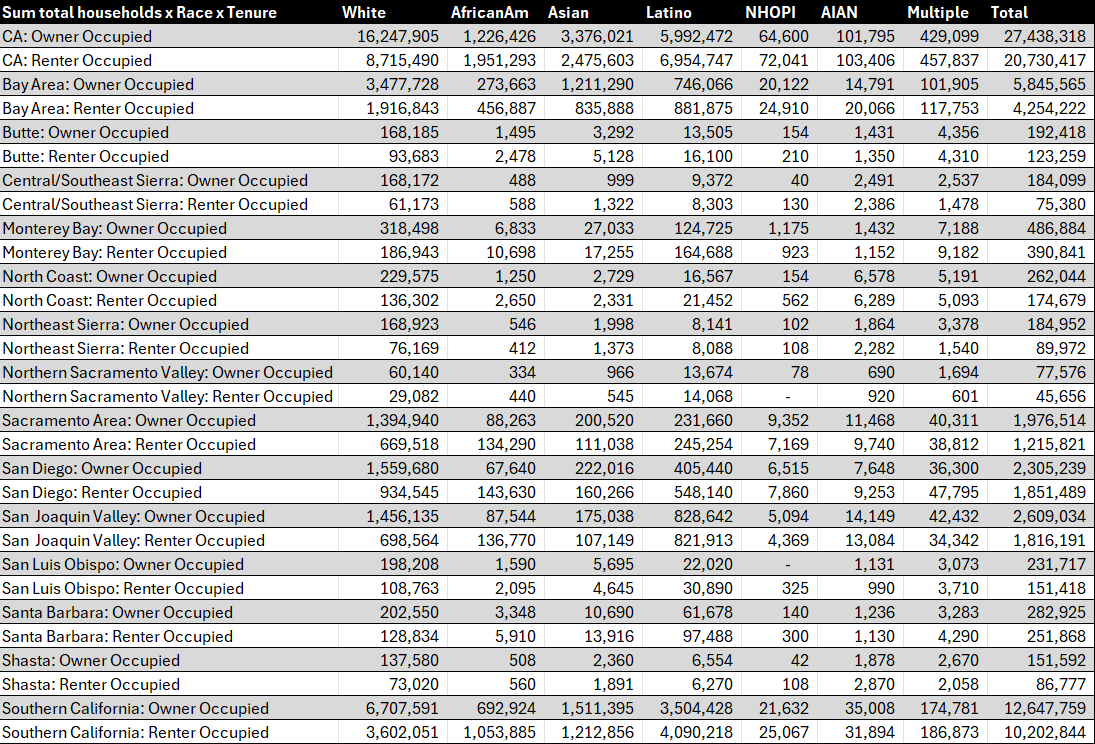
Description automatically generated

Central/Southeast Sierra:

A graph of different colored squares

Description automatically generated

Below is the table with the number of total households across each region and race for Owner-occupied and Renter-occupied households. Mortgage paying owner-occupied households and rent-paying renter occupied households were omitted because of the number of missing values. For state totals, there are more owner-occupied White households and Asian households than renter-occupied. While for every other race the opposite is true, additionally, there are nearly double white owner-occupied households as renter-occupied ones. For the three highest rent burdened regions, there are more owner-occupied households than renter-occupied ones, but they are relatively similar. For the 3 lowest however, far more households are owner occupied than renter occupied, and for both Northeast Sierra and Central/Southeast Sierra, there are twice as many owner-occupied households as renter-occupied ones. Southern California is by far the most populous region in CA, followed by the Bay Area, the San Joaquin Valley, San Diego, and the Sacramento Area. The least populated are the Northern Sacramento Valley, Shasta, and the Central/Southeast Sierra region. Native Hawaiian or Pacific Islander have the fewest households in CA, which is unsurprising, and in the San Luis Obispo region there are no NHOPI owner-occupied households. Housing cost burden and a lack of housing are huge issues right now, and although the data is from 2006-2010, it is unlikely the situation in CA has improved. Some ways to combat housing cost burden include changing zoning laws to make it possible to build more housing developments and encourage housing density. Other options are to make it cheaper to build housing in denser areas where more housing is most needed. Some drawbacks are that CA has to be aware of the height of their buildings given the fault line they lie on, as well where new developments are built because forest fires can often spread and force people out of their homes. However, housing cost burden is a solvable issue, and because it is so prevalent statewide across all demographic groups, it is necessary that legislative and cultural changes take place so that less households are rent burdened.



**Section 3: Power BI Dashboards**

A map of the united states

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A close-up of a map

Description automatically generated