Alejandro Naranjo Torres

Data Analytics: Applying Machine Learning to Housing, Income and Cost of Living Data

# Abstract

The focus of the following report will be dealing with three different datasets. The first data set is *Income In The Past 12 Months* data from the 2015 to 2021(except for 2020) (Census) and this data is from the American Community Survey. For a better understanding of what the America Community Survey (ACS) is, as described by the Census, “premier source for detailed population and housing information about our nation” (American Community Survey, n.d.). This is a place where in between the decades of when the official census is conduct data points are released for counties across The United States. For the second dataset, another ACS dataset, is *Selected Housing Characteristics* from the years of 2015 to 2021 (Selected Housing Characterisitcs ). This dataset describes housing characteristics for each state and can be filtered by county. The last dataset that is being used is the *Cost of Living Dataset* Provided by the Federal Reserve Bank of Atlanta (Fenderal Reserve Bank of Atlanta); this dataset describes living expenses someone in America would have to pay. The great thing about the cost-of-living data is that it is free and contains data based on county if available. Initial beliefs before diving into the work is if we are able to combine all three datasets together in a meaningful way we can predict; which state someone may be from based on expenses, income and housing; we can predict the total cost of living using information about the county; and clustering different states to see what separates each of them. Due to limited resources the focus of the report will be on three different states; the first is California, the second is New York and lastly is Washington. The reasoning behind selecting these states is because I am from California, I got to school in New York and Washington is the state I am moving to once I graduate. The motivation behind focusing on these datasets is because we are starting to face issues with the cost of living; people aren’t making enough; wages aren’t going up; but everything that we need to live continues to increase.

# Data Description

For the housing and income data it was a pretty easy choice with where the data would be pulled from the reason for that is because it’s public. This is publicly collected data and the data that is provided is full of useful information as well as information that may not be needed. For the cost of living data the goal was to find data that was available based off of county. There are a lot of resources out there which provide users API to access their databases with this data; however, it’s always locked behind pay walls. The second best option was to use The Federal Reserve Bank of Atlanta Cost of Living Database (CLD) (Fenderal Reserve Bank of Atlanta). The CLD consisted of large amounts of data that reflect how much the average person would pay for rent, healthcare, transportation, raising children, and much more. The data was also separated by county so this was another reason why the data was selected. On the same page as where the CLD can be downloaded is a Manual on how to operate and extract the necessary data.

For the two ACS datasets pulled from the census data tables, <https://www.census.gov/data/tables.html> , each had the necessary information for creating the models. The housing characteristics dataset (Selected Housing Characterisitcs ) has information on how many homes were in a county, vacancy, rental units, home owners and much more. For the ACS Income dataset (Census), the most important features that this dataset has is the information on the distribution of the income bracket for each bracket. This is important to have because certain counties may have more households in a higher bracket or might have more households in the lower brackets.

There will be four different machine learning (ML) models that will be created. Each of them will require different features as each will serve a different purpose.

### Linear Regression

For the linear regression, the goal is to look at two different features. The first feature is the *Average Rent Cost by County,* and the second feature is *Estimate Housing Occupancy Total Housing Units.* The reason behind this is to see if there are any correlations between the number of housing units a county has and how it impacts the average rent price. There will be three different models built for this and each one will represent one state, the data entries will reflect the counties of that state from the years 2015 to 2021.

### KMeans Clustering

Combining both the income bracket distribution and the estimated cost of living essentials, we can confidently predict where a certain entry is from based on the given data. The reason for this is the income distribution and the average expense cost enough to determine this information. Will we be able to confidently distinguish certain states and what else can we conclude from this confidentiality? If we are able to build a successful model, we can make the argument some states stick out from the rest when it comes to cost of living and income distribution.

### Random Forest for Predicting State Based on All Three Datasets

For this model the real goal is to use most of the features we have selected to use as the final dataset. Kmeans is a great method for modeling and looking for groups within the data, but it has it’s limitations. I want to be able to compare the results of the Kmeans model to this random forest model which will have more data.

### Random Forest for Predicting Total Household Expenses

For the last model the goal is to take the income distribution from each county and to generate “household” data points instead of using it as a distribution value. What this means is if an income bracket has a value representing the percentage of a given total, I will take ten percent of a predefined sample size and then generate an “income” for that bracket range. The goal for this is to simulate real data entries and to see how confidently we can predict the estimated cost of living. This will be done for each state.

# Data Description

## Income Dataset

To get a better understanding of the distribution of the income brackets for each state we can break it down by averaging all the counties. Household data is only going to be considered because it simplifies the process and doesn’t increase the complexity of it. A household is defined as a group of people related or unrelated living together. The reason for this is it gives us a better understanding of the actual data and what we are looking at. If we look at the three figures down below we can get a quick glimpse into what the data looks like. Each figure has the same y-axis and starts from the very top with the lowest bracket (under $10,000) and at the bottom we have the highest bracket ($200,000 and above).

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure Income Distribution For California (Left) 2018, 2019, and 2021

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generated

Figure Income Distribution For NY (Left) 2018, 2019 and 2021

Chart, bar chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

Figure Income Distribution For WA (Left) 2018, 2019, and 2021

­

Looking at each country through each year we see that they maintain a steady bracket distribution. We do see an increase in the lower bracket for each state. Each of the distribution for each year and state seems to be skewed to the right and this can give us an idea of what that might be. California has two major rejoins known for having high cost of living (Los Angeles, and The Silicon Valley), New York has New York City, and Washington has the Greater Seattle Region.

For the cleaning of the income data there was a lot of data that wasn’t required. The first part was a lot of data entries were missing and the reason for this is because there were columns included to indicate any annotations made to the data; these columns were dropped. Second, according to the data notes many of the entries that were denoted with a “(X)” were not displayed because the information was too small and could be ignored. Third was a lot of null date that was denoted by the use of “null”. These columns were easy to get rid of and were removed from consideration of the final merged dataset. Other issues with the dataset were all the unrelated metadata that was inside the large dataset. Originally the dataset had a total of 1,147 different features; all of these features were not needed and the only valuable data that was pulled was the income distribution data for each county. After all the data cleaning was done, we were left with 12 different variables to be used later.

Much of the data that came with the datasets was left behind because it wasn’t deemed worthy and not related to what the goal of the models were trying to achieve. Another issue encountered with the data was the different column names! One year a column would be named a certain way and the next year it was updated causing issues when merging the data together.

## Housing Dataset

It’s difficult to select a certain variable to focus on for housing characteristics as there a lot of information that can be used.

Graphical user interface, application

Description automatically generatedA picture containing scatter chart

Description automatically generated

Figure Occupied Housing In California 2021 and 2019

Chart

Description automatically generated with medium confidenceChart

Description automatically generated with medium confidence

Figure Occupied Housing For NY 2021 and 2019

Chart

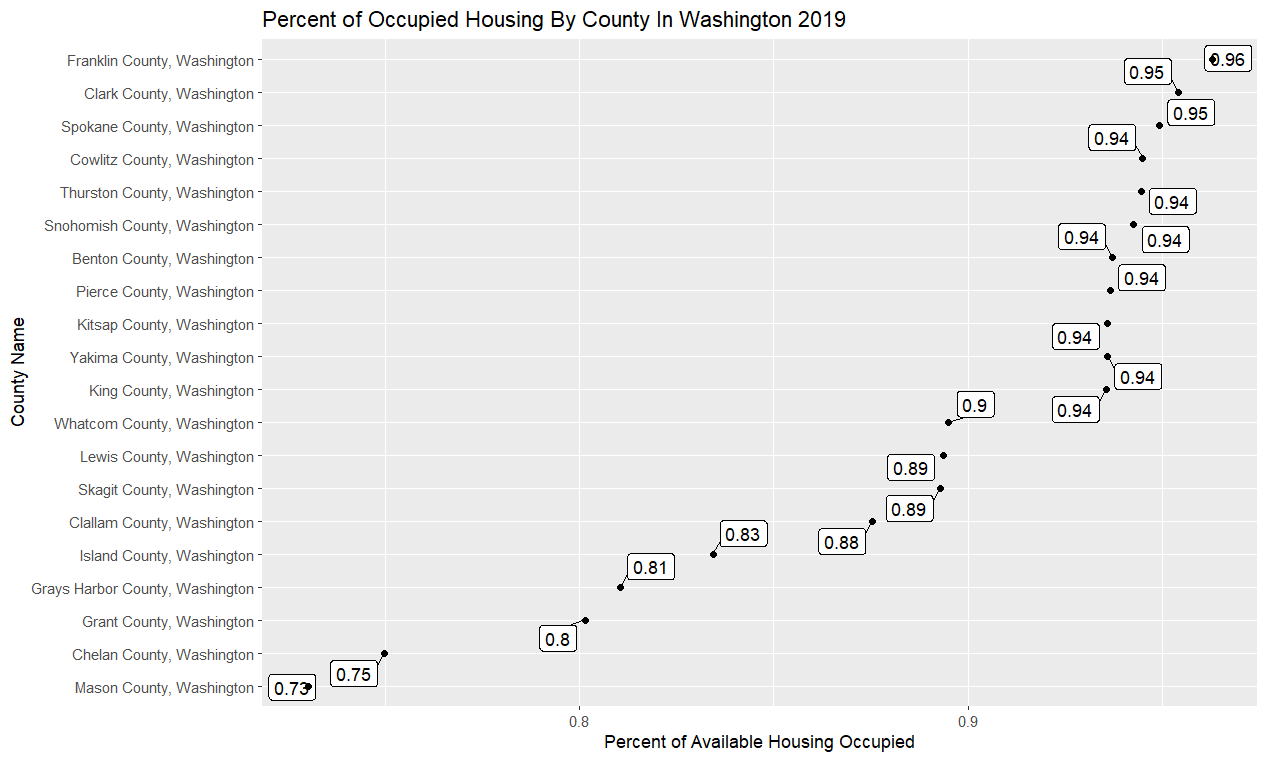
Description automatically generated

Figure Occupied Housing In Washington 2021 and 2019

## Cost of Living Dataset