Alejandro Naranjo Torres

Data Analytics

Contents

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 in place between the periods of when the official census is conducted. 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.

All the following graphs, datasets, and R code can be found on the following github link:

Git Hub: https://github.com/nara343/Data-Analytics-Project-6000

Data Description

For the housing and income data it was a pretty easy choice 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 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 which 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 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 this dataset has distribution information on income per household. 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.

The data that will be used for all the models will be data from 2015 to 2021 excluding 2020 because the Census was conducted that year and ACS varies from the Census.

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 household 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 as our goal is to predict the estimated total cost of living by county and having all three states combined can lead to over fitting that is not needed.

Data Analysis

Income Dataset

Housing Dataset

It's difficult to select a certain variable to focus on for housing characteristics as there is a lot of information that can be used. A data point that can be used to visualize how the data looks is by focusing on the Occupied Housing broken down by county. For each state there are two years being shown as it's to get an idea how the states are changing from the previous year to the next.

Figure 4 Occupied Housing In California 2021 and 2019

Looking at *Figure 4*, a cool thing to look at is for larger counties such as *Santa Clara, San Francisco, San Diego, and Los Angeles* are sitting closer to the middle and from the previous years 2019 there occupied housing rating has gone up. These larger cities are not expected to have a near perfect rating of 1 due to the rapidly expanding and growing number of houses every year.

Figure 5 Occupied Housing For NY 2021 and 2019

For New York, *Figure 5,* looking at 2019 to 2021 we can see that the lowest number of occupied housing has gone up from what is was at 2019. The lowest ratio was held by Sullivan County, New York but in 2021 that value increased by .10. This could be due to people leaving certain areas with a higher cost of living and moving to more rural areas. This can be supported by the occupied housing percentage for New York City. In 2019 it had a 86% and jumping to 80% in 2021 and compared to some of the larger cities in California we saw an increase for some of the larger cities. The reasoning for this could be correlated to the worldwide pandemic (2020 – 2021) and how it affects major cities as white collard workers moved to a remote style of work.

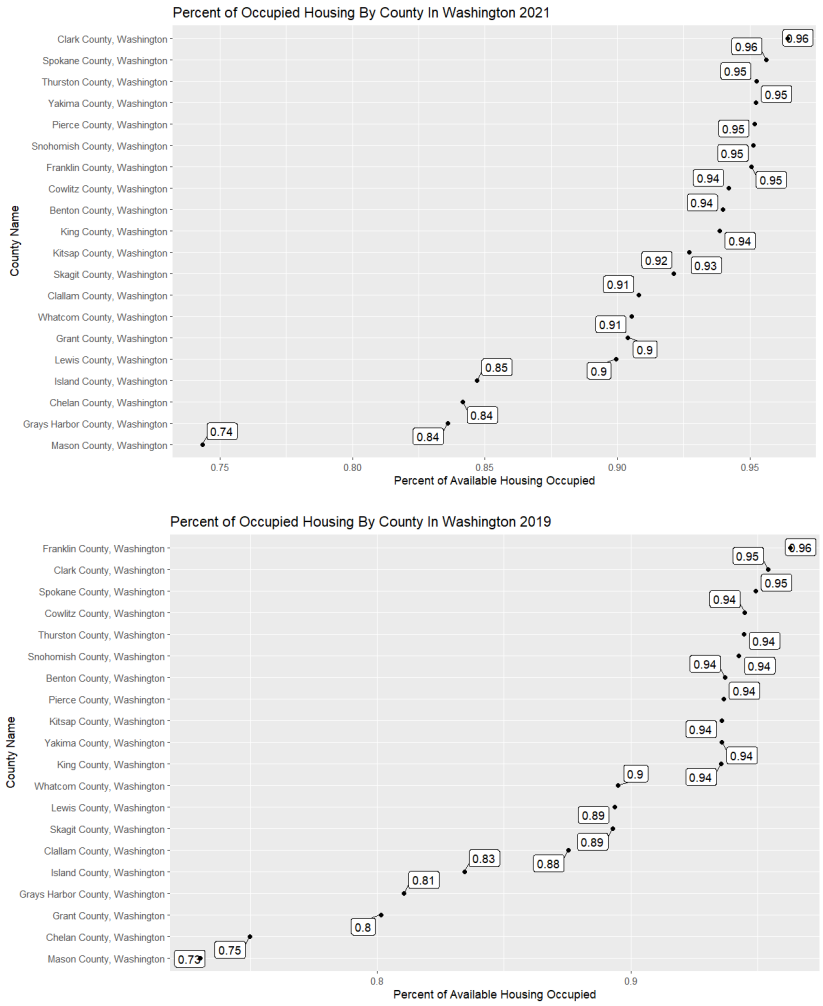


Figure 6 Occupied Housing In Washington 2021 and 2019

Looking at, *Figure 6*, we can see the percentage of occupied housing has gone up for most counties and from the year 2019 to 2021 we can see that only 4 counties now set below a ration of 80%. The rest of the counties continue to see a growth outside Mason County which has only seen a 1% increase from 2019 to 2021.

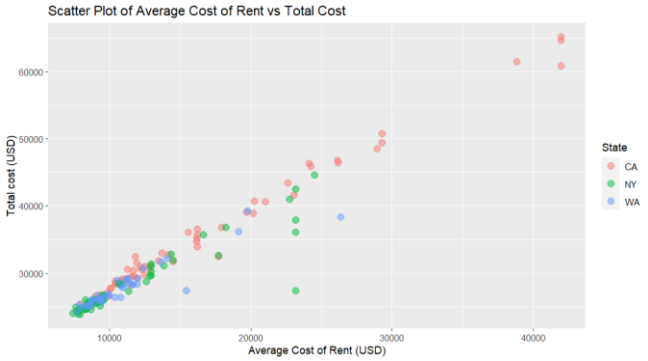
This housing occupied is valuable data as it can have an affect on the average rent price or the average price a homeowner is paying a month. The reason behind this is housing can drive up the cost of living and it can be an indicator of a country seeing growth which can lead to increased costs. From personal experience I was grow up in Morgan Hill, California and this was a small town which was not as thriving or full of traffic; now after being gone for five years due to school (2018-2023) whenever I visit the town is full of new expensive housing, more luxurious apartments, prices of food has gone up, and everything has become harder to exist without worrying about money.

For the data cleaning that was required, similar to the ACS Income dataset there was a lot of similar issues. There were “Annotation…” columns full of null or missing data so those were the first to be dropped; second the dataset was full of lots of metadata about housing characteristics. Given this was a housing characteristics dataset it was expected to see a lot of this information; however, for the goal of the project it was not needed and the data was dropped. For the remaining dataset the data was clean and there was no need to alter any of the existing information.

Cost of Living Dataset

For the cost-of-living dataset the data was pulled from the Federal Reserve Bank of Atlanta (Fenderal Reserve Bank of Atlanta) and from there individual data points were pulled from the database. There are 4 different variables pulled: *Cost of Rent, Cost Of Insurance With Employer, Transportation Expense, and Misc Expens*e. Before the data could be used for any data analysis it first had to be retrieved. The data was stored as data entries broken down by county and each county had various data points. In order to reduce the complexity of the data the average was taken for each county for each individual state. This meant each county data point in the models represents the average for that county. As for which year the data represents varies depending on what the category is. For *Cost of Insurance With Employer* this data was representing values from 2014 and we wanted data points from 2015 to 2021. In order to account for the inconsistent yearly data I used the average inflation rate between those years and the target year to get an estimate of what that would look like. This is assuming the changes in price are following the inflation rate, and the following rates were used for each year; 2021 at 4.7% , 2020 at1.2%, 2019 at 1.81%, 2018 at 2.44%, 2017 at 2.13%, 2016 at 1.26% and 2015 at 0.12% (Current US Inflation Rates: 2000 - 2022, n.d.)

The following data represents 2021 and some of the data was estimated by applying the rate of inflation if 2021 data was not available. All the cost data is yearly cost data and is not broken down into monthly data.



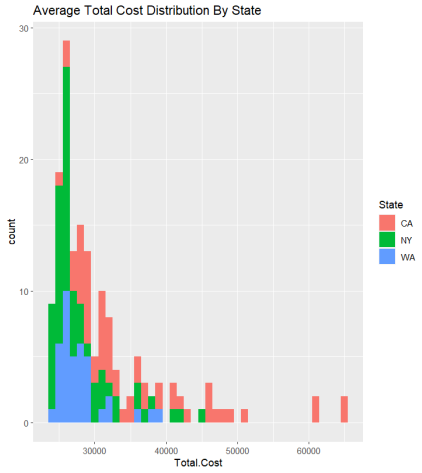


Figure 7 Scatter Plot of Average Cost of Rent vs Total Cost

Figure 8 Average Total Cost Distribution By State

For the scatter plot on the left, *Figure 7*, we can see the average cost vs the total cost (summation of all the expenses) and it has been filled by the state. By looking at the data we can see California has more counties with high cost of living and has a overall higher total cost of living. For Washington state we have much fewer counties in the upper region for average cost of rent this can be supported by the Average Total Cost distribution by looking at the graph on the right, *Figure 8.* On the right graph, we can see that both California and New York are the reason why the distribution has a longer right tail. From both of these graphs alone we can see that California is overall more expensive than both of the other states.

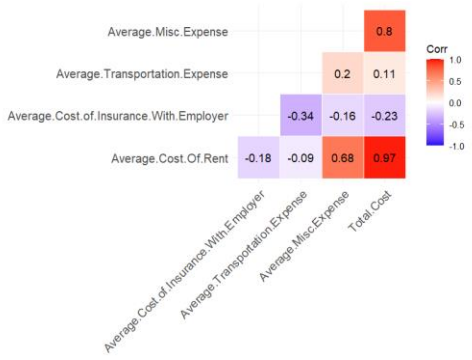


Figure 9 Correlation Matrix

Figure 9, we see the correlation matrix for our data, and we can understand how each affects each other. Starting with the higher correlation we can see that *Average Misc Expense and Total Cost* highly positively correlated and the same can be said for *Average Cost of Rent and Total Cost* having a very high positive correlation. These values are going to play an important role in the models and can be a dominating factor when it comes to variable importance and capturing the variance of the data.

Average.Cost.of.Insurance.With.Employer	Average.Transportation.Expense	Average.Misc.Expense	Total.Cost
Min. :1552	Min. : 2414	Min. : 0	Min. :23819
1st Qu.:1593	1st Qu.: 9932	1st Qu.: 4860	1st Qu.:25940
Median :1593	Median :10329	Median : 5360	Median :28311
Mean :1652	Mean :10222	Mean : 5588	Mean :30854
3rd Qu.:1770	3rd Qu.:10748	3rd Qu.: 6006	3rd Qu.:32296
Max. :1770	Max. :13364	Max. :10639	Max. :65213

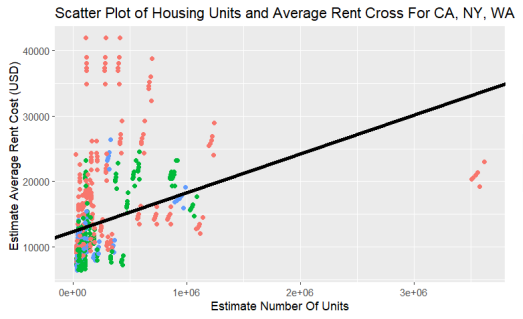
Figure 10 Summary Statistics For Cost Of Living

Looking at *Figure 10* we get a quick glance at the data and how it looks. Each feature has consistent data and there exist outliers but those are handled case by case. The only feature which is concerning is the *Average Misc Expense* as there are a few values which are 0.

Modeling

Linear Regression

For the Linear Regression Model 5 variations of the same model were built with different small changes. The base model that was created was looking to see if the *Estimate Number of Units* in a county can be used to predict the *Average Rent Cost.*



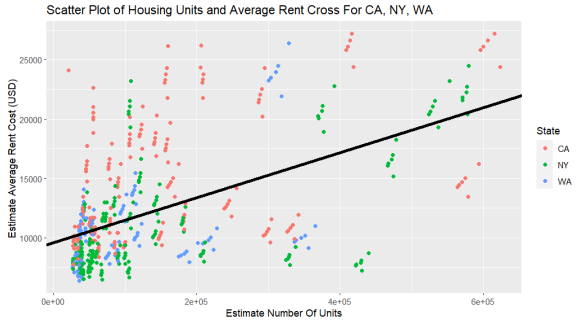


Figure 11 Average Rent Cost ~ Estimate Number of Units Unfiltered Outliers *Figure 12 Average Rent Cost ~ Est Number of Units filtered Outliers*

The model formulation was *Average Rent Cost ~ Estimate Number of Units* and for the first two models we are including all the data and we are not separating it by state. Looking at Figures *11 & 12* we can see the effects of the outliers in our dataset. For the model including the outlier data (*Figure 11)* we get a an R^2 value of 0.1276. This is a really bad score as it does not really show how the two variables can be used to explain the dependent variable. In order to confirm this we need to remove the outlier data to improve the model and the method we are going to be using is the "Fences" method. After removing the outlier data, *Figure 2,* we improved our R^2 value and we can now see the *Estimate Number of Units* explains .28 of the *Average Rent Cost*. The results were improved but due to the large amounts of data it can be difficult for *Estimate Number of Units* to capture most of the variance. To further test this we will

be building three separate models for each state and continuing to filter out outliers and *Figures 13,14, & 15* are the results.

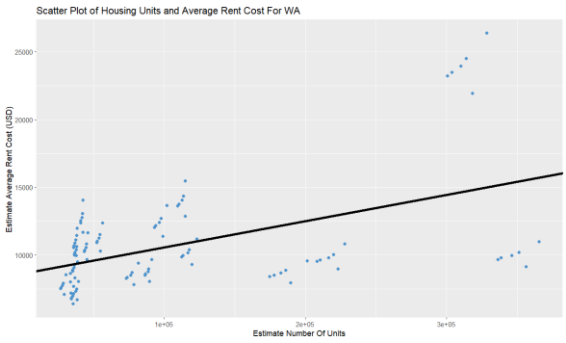


Figure 13 Linear Regression Results For Washington

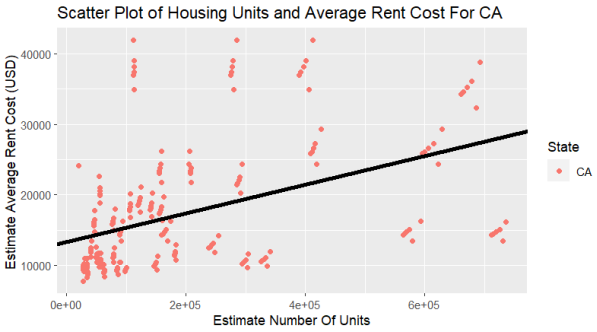


Figure 14 Linear Regression Results For California

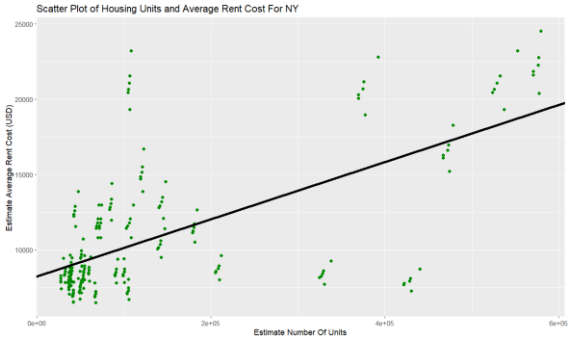


Figure 15 Linear Regression Results For New York

The following are the R^2 results for each state:

- Washington: 0. 243
- California: 0.191
- New York: 0.425

The results of the individual models for each state is interesting as I expected a much larger R^2 value for each model. The belief was the more housing there is meaning a certain county is growing and that would be a reason for the average rent to go up in price. The models for Washington and California only explained less than 0.25 of the variance while the New York model had a score of 0.425. This could be due to the data that was available as we can see with California the data points are a lot more scattered and grouped together far apart.

We cannot conclude using *Estimate Number of Units*, as the independent variable, would explain the variance of our dependent variable.

KMeans Clustering

For the clustering model the goal is to use income distribution data paired with cost-of-living data to see if we can form clusters to predict which state each data entry is from. The only variable in cost of living that was removed from the model was the *Average Total Cost* value for each county. The reason for this is motivated by *Figure 7* and how we can see California is sitting above the other two states and if this is enough to predict the state.

Since we already know the number of clusters that are needed the model was built using that information and the *nstart* was set equal to 20. Before the model was created all the numerical data was normalized to optimize the model the best we could. The following function call was used to create the model: *kmeans(norm_data, center=3, nstart=20).*

In order to visualize the results we are using a table to check the accuracy of the model.

```
table <- table(final_dataset$State, df_cluster$cluster)

#     1   2   3
# CA  48  98  97
# NY  37 143  53
# WA   4  63  50
```

Figure 16 Kmeans Results

Looking at the results we can see that there are a lot of incorrect predictions indicating the data could not be clustered properly to separate each state. The results in *Figure 16* show an accuracy score of 40.6 which is 7% better than taking a random guess at 33%. Taking a look at *Figure 18* we can visually see how the data is grouped together but there is a lot of overlap between them, and the model was not able to form the clusters correctly.

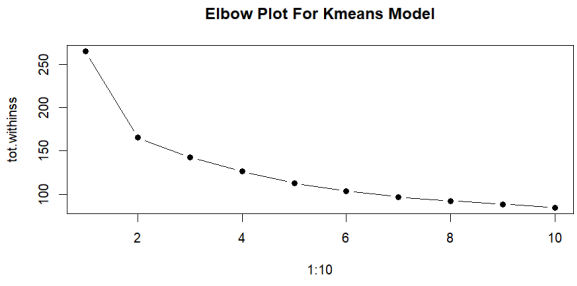


Figure 17 Elbow Plot For Kmeans

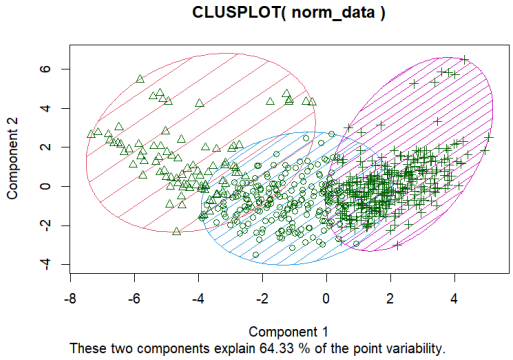


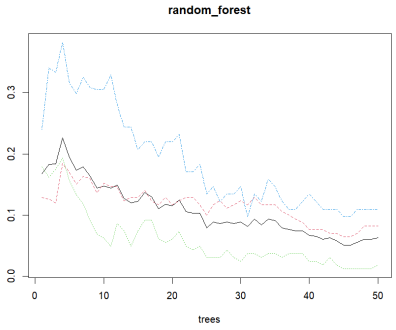
Figure 18 Cluster Plot

If we didn't know how many dependent variables we had then the best number of clusters to use would have been 4, *Figure 17.* Of course, this was not needed as we were building the model with the intention of fitting our data to those three states. The results from the model show we are not able to properly cluster the states based off the data we used. Even though we did perform better than randomly guessing the data we choose only explains so much of the actual dependent variable.

Random Forest for Predicting State Based on All Three Datasets

To further test the idea proposed for the Kmeans model we will the income data alongside the rest of the selected housing data and cost of living data to building a Random Forest model to predict the state. For this model this required having a training and testing set so the split was a 70/30.

First Focusing on the results of the model we will look at how well the model classified the testing data the information can found down below, *Figure 19.*




```
# No. of variables tried at each split: 5
# OOB estimate of  error rate: 6.24%
# Confusion matrix:
#    CA  NY WA class.error
# CA 157   6  8  0.08187135
# NY   3 161  0  0.01829268
# WA   8   1 73  0.10975610

tableCheck <- table(pred_w_random, testing$State)
```

```
tableCheck <- table(pred_w_random, testing$State)
  # pred_w_random CA NY WA
  # CA 72  0  3
  # NY  0 69  2
  # WA  0  0 30
Accuracy: 0.972
```

Figure 19 Random Forest Plot and Results

On the left side of *Figure 19* we see a summary of the model and what it tells us about the classifications. We can see the error margins for each of the states and it's low and the estimated error is 6.24%. On the right side of the figure we see the prediction results using our testing set and we get an accuracy of 97.2% which is a 2.8% error rate which is much lower than the estimated error. At around 30 trees we start to see the error rate flatten out. For us this means the independent variables selected do a good job of explaining the dependent variable. We can confidently predict each state based off of income data, housing data, and cost of living data from each of it's counties.

To take a deeper look at the results of the model we will look at the variable importance and see which variables had the most impact on our model. In order to visualize the importance data, we will be using the Random Forest Explainer package (randomForestExplainer: Explaing and Visualizing Random Forest In Terms of Variable Importance, n.d.)

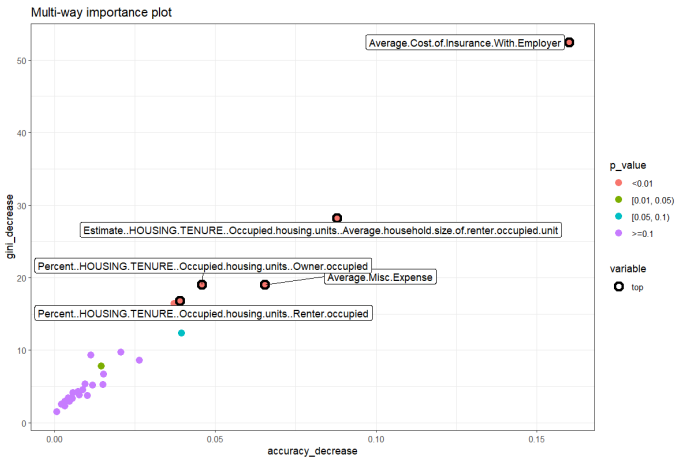


Figure 20 Variable Importance For RF Model

Looking at the plot above we see a plot of *gini_decrease* and *accuracy_decrease*. Each of these variables resembles the decrease in accuracy when the variable is permuted and the decrease in the Gini Index of node impurity when split on the variable. There are only five different variables which are labeled, and these are our five most important variables; *Average Cost of Insurance with Employer, …Average Household Size of Renter Occupied Unit, …Housing Units Owner Occupied, Average Misc. Expense and …Housing Unit Renter Occupied.* When considering building another model we can only focus on including these five variables as these are the most important to the model we built.

Random Forest for Predicting Total Household Expenses

For the following models, three different models will be trained on generated data. As mentioned, before we will be taking the income distribution for each county and generating samples based off of a sample size. If we have County A with a 10% distribution with 10 income brackets and we want 100 samples from that county there will be 10 generated data points from each bracket and the income assigned will be a randomly generated value of that bracket range. We will combine this data and our model formulation will be as follows: *Average Total Cost ~ County + Income*. Before the model was built some data cleaning was required because our distribution was heavily right skewed. *Figure 21* down below shows the distribution after the outlier data points were removed using IQR Fences.

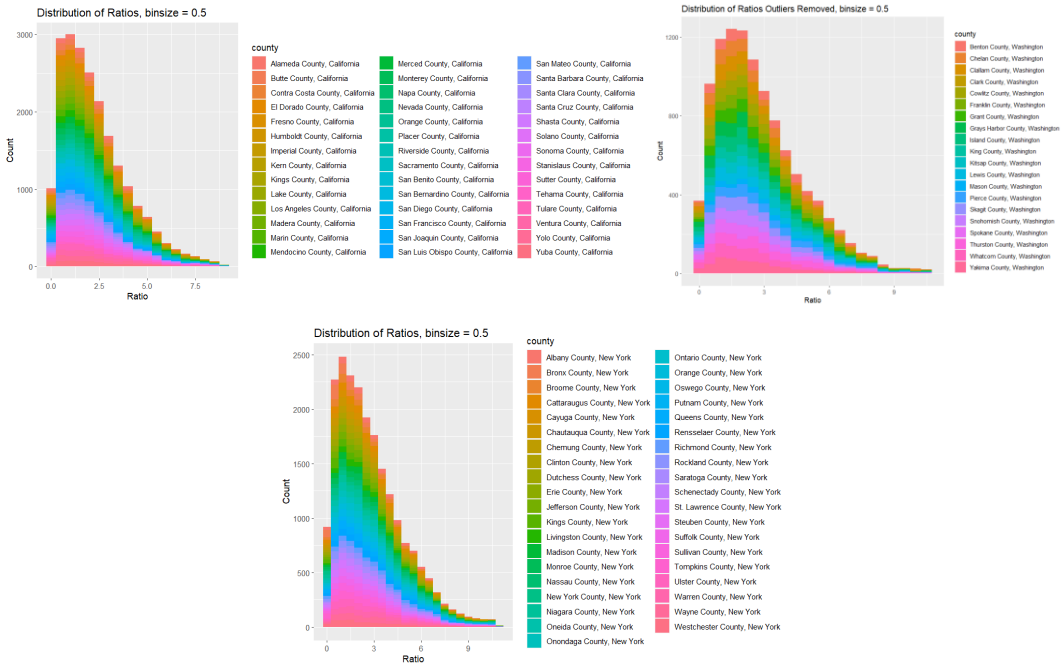


Figure 21 Distribution After Generating Data

To keep the graphs clean we plotted the income distribution using a ratio of the generated income value to *Average Total Cost* and used bin sizes of 0.5. Looking at the graphs we can see each state has very similar distributions of ratios and this can further be tied to how wage disparities exist and how it affects the common people. In 2021 The United States median household income is $70, 784 (Census Gov) and in 2019 the median income for individuals was $31,133 (Data Commons Placer Explorer) with the total cost of living ranging from a minimum of $23K to maximum of $65K (Figure 10) most will probably be able to cover their expenses; however, we aren't accounting for taxes, debts, or any other expense that isn't listed. Another major thing that isn't considered with the following model is children are excluded. This is just based off meeting living expenses. Even though *Figure 21* looks good on paper we are not including the cost of children or other costs such as utilities (unavailable through CLD), groceries (unavailable through CLD), etc. .

Focusing on the models that were created below are the regression results for each model:

```
California                                  New York                                  Washington
# Type of random forest: regression        # Type of random forest: regression       # Type of random forest: regression
# Number of trees: 5506245                 # Number of trees: 50                     # Number of trees: 50
# No. of variables tried at each split: 1  # No. of variables tried at each split: 1 # No. of variables tried at each split: 1
#                                          #                                         #
# Mean of squared residuals: 5506245       # Mean of squared residuals: 2791977      # Mean of squared residuals: 2336829
# % Var explained: 93.78                   # % Var explained: 88.93                  # % Var explained: 81.37
```

Figure 22 Results of Random Forest Model Average Total Cost ~ County + Income

The results above are broken down into three sections starting from the left most we have the results for California, New York and Washington. Respectively the model built for California both County and Income explained the variance very effectively with a score of 93.74 while the other two were trending down. New York had a rating of 88.93 and Washington had the lowest rating of 81.37. This proves our hypothesis that we can predict *Average Total Cost* using both *County names* with a households income. The only concern with the California model is any potential overfitting.

There were two dependent variables but understanding how important each variable is will give us some insight on which variable played a bigger role. The three plots of variable importance can be found down below, *Figure 23.* In each plot we see the use of the *node_purity_increase (increase by splits on the variable) and mse_increase (increase of mse after variable is permuted)* (randomForestExplainer: Explaing and Visualizing Random Forest In Terms of Variable Importance, n.d.).

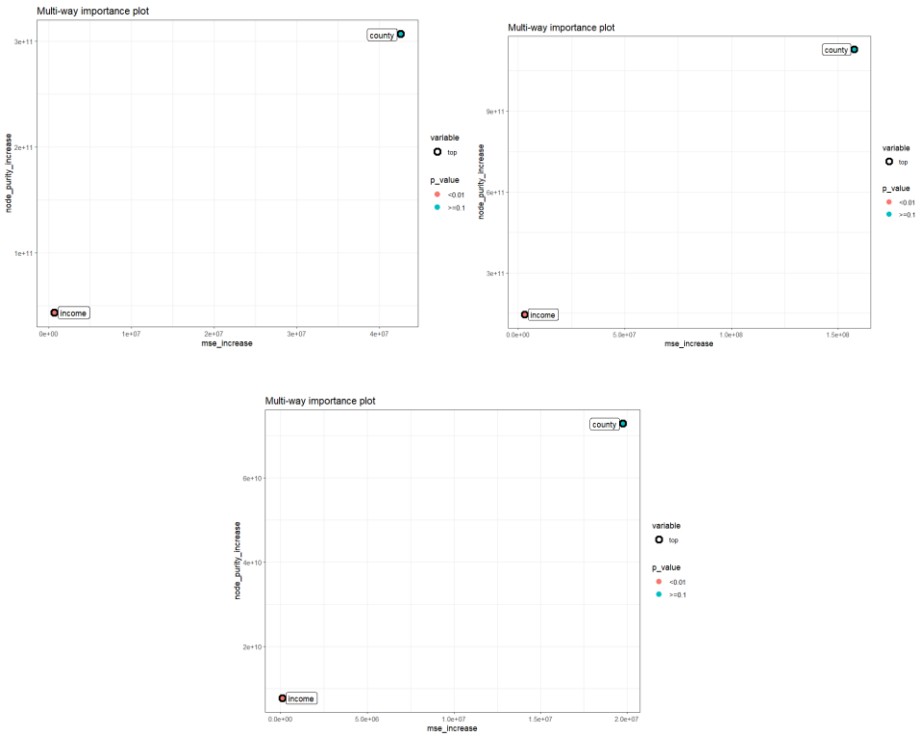


Figure 23 Variance Importance for Each Model

Understanding which variables were more important can give us insight into what can be improved in the future or if we can form different models using the same data or change the variable relationships. For each of the models the county name played the role of most importance for explaining the variance of the dependent variable.

By building this model we can conclude that we can explain the variance confidentiality with just the name of the county and the income of household and successfully predict what might be the total cost for a household.

Conclusion

In conclusion, we have built four different models each with the goal of finding different variables that could explain the variance of our independent variables. Two of the models (both Random Forest models) were successful while the other two (Kmeans and Linear Regression) lacked the variance explanation. This does not mean that the two models which underperformed are not good, it shows that the independent variables selected were not good at explaining the dependent variables and changes can be made in the future. For the linear regression we can improve it by including additional information such as the name of the county represented by a value. For the Kmeans model we can focus on reducing the amount of features we used to build the model and to use different features that may explain variance. As for the two Random Forest models, the only concern is to worry about overfitting and the next step is to continue training different models and using other techniques such as K-Fold validation.

From the very start of this project the goal was to look at housing, income, and cost of living data to analysis the trends here in America. The reason for that is the large wealth disparity and the lack of wages for certain communities. Before any models were built the plan for what was going to be investigated changed a lot after the data analysis process began; the reason for that was because of the data that was available. The most important piece of data needed was the cost of living but in order to get accurate and recent datasets were behind pay walls which required resorting to public data that did lack some depth. The original plan was to build two linear regression model one of them was a multivariate, clustering using Kmeans, and a Random Forest model but that changed once I started working with the data.

Working with Census data can be very overwhelming given the fact that it contains lots of data and having to filter through it was difficult. For future changes what would be done differently is what data is used; there's lot of data out there and potentially paying to access higher quality data to improve some of the models. For the models, I would explore different types of methods such as multivariate linear regressions, XGBoosting or bootstrapping data to improve variance scores.

References

American Community Survey. (n.d.). Retrieved from United States Census Bureau:
https://www.census.gov/programs-surveys/acs

Census Gov. (n.d.). *Income in the United States: 2021.* Retrieved from Census Bureau: https://www.census.gov/library/publications/2022/demo/p60-276.html#:~:text=Real%20median%20household%20income%20was,and%20Table%20A%2D1) .

Census. (n.d.). *Income In The Past 12 Months* . Retrieved from United States Census Bureau: https://data.census.gov/table?q=Income&g=0100000US$0400000&tid=ACSST1Y2021.S1901

Current US Inflation Rates: 2000 - 2022. (n.d.). Retrieved from U.S. Inflation Calculator: https://www.usinflationcalculator.com/inflation/current-inflation-rates/

Data Commons Placer Explorer. (n.d.). Retrieved from datacommons: https://datacommons.org/place/country/USA?utm_medium=explore&mprop=income&popt=Person&cpv=age%2CYears15Onwards&hl=en

Fenderal Reserve Bank of Atlanta. (n.d.). *Cost of Living Database (CLD)* . Retrieved from Atlantafed: https://www.atlantafed.org/economic-mobility-and-resilience/advancing-careers-for-low-income-families/cost-of-living-database

randomForestExplainer: Explaing and Visualizing Random Forest In Terms of Variable Importance. (n.d.). Retrieved from https://cran.r-project.org/web/packages/randomForestExplainer/index.html

Selected Housing Characterisitcs . (n.d.). Retrieved from United States census Bureau : https://data.census.gov/table?q=Rent&g=0100000US$0400000&tid=ACSDP1Y2021.DP04