Using Machine Learning to Analyze Cost of Living Trends With Housing, and Income Data

Alejandro Naranjo [1] (narana2@rpi.edu),
[1]Rensselaer Polytechnic Institute, Troy, NY, United States,

Abstract

With the recent pandemic and the rising inflation this year it is clearly that who are found to be in lower income bracket are drastically affected. With the large income disparity in The United States, it becomes a concern as to what it would look like in the future if things are not improved or changed. By taking income data, housing data, and the cost-of-living data we want to analyze the trends that are occurring for the cost of living based on the counties in The United States.

Problem Area

The goal of the poster is to presented findings about what we can successfully predict when it comes to using income, housing, and cost of living data. There are four different models that have been constructed:

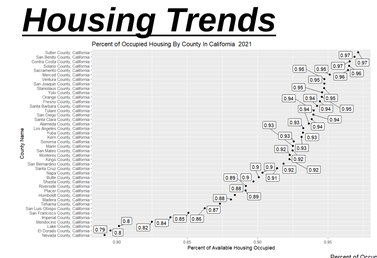
- Linear Regression Model
  - *Average Rent Cost ~ Estimate Number of Units*
- Kmeans Clustering
  - Clustering based on the different state data.
- Random Forest Model
  - Predicting which state each data point is from.
- Random Forest Model
  - Predicting *Estimated Total Cost* based off of County and Income.

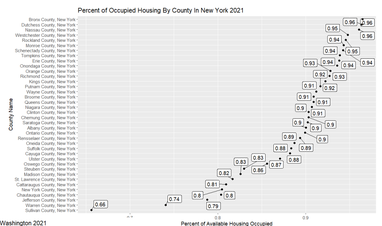
These four different models will be built to prove or disprove our hypothesis and to make future predictions about the cost of living.

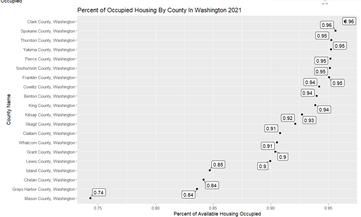
Where The Data Came From

For the three different data sets each one came from a public government entity. The first data set *Income In the Past 12 Months 2015 – 2021* is from the American Community Survey (ACS). The second data set comes from the ACS and it is the *Selected Housing Characteristics* 2015 to 2021. The last data set is Cost of Living Database from the Federal Reserve Bank of Atlanta.

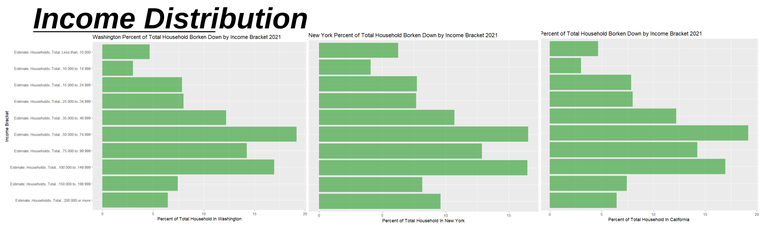
Housing Trends



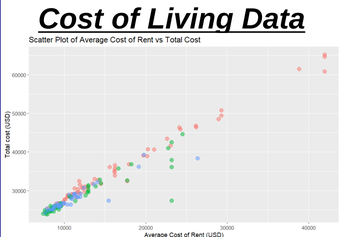


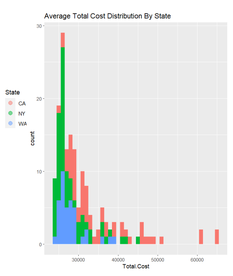


Income Distribution



Cost of Living Data

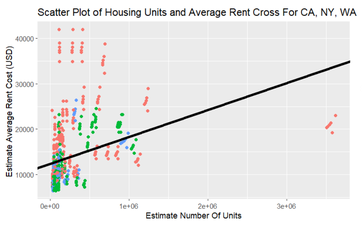




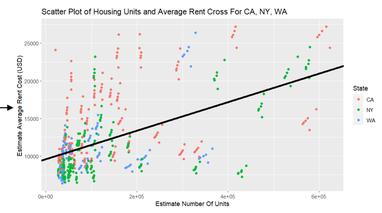

```
Average.Cost.of.Insurance.With.Employer Average.Transportation.Expense Average.Misc.Expense     Total.Cost
 Min.   :1552                            Min.   : 2431                  Min.   :    0          Min.   :23839
 1st Qu.:1593                            1st Qu.: 9934                  1st Qu.: 6868          1st Qu.:29940
 Median :1593                            Median :10329                  Median : 5360          Median :28311
 Mean   :1593                            Mean   :10223                  Mean   : 5508          Mean   :30854
 3rd Qu.:1770                            3rd Qu.:10748                  3rd Qu.: 6806          3rd Qu.:32296
 Max.   :1770                            Max.   :11364                  Max.   :10639          Max.   :65213
```

Model Building

Linear Regression

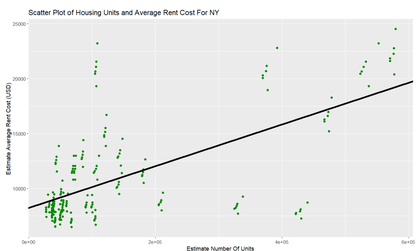


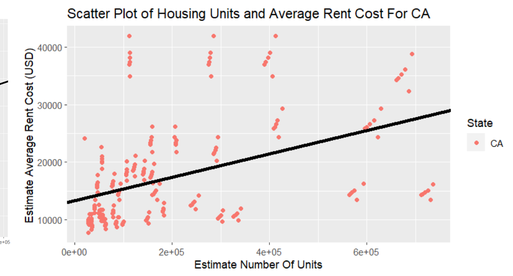
Applying IQR Fence For Outlier detection



Results: R^2 = 0.128

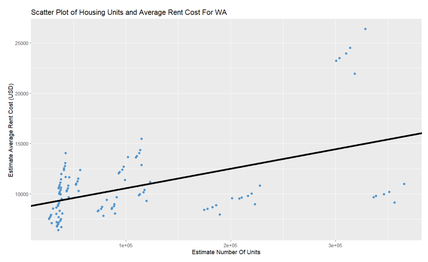
Results: R^2 = 0.28





Results: R^2 = 0.425

R^2 = 0.191



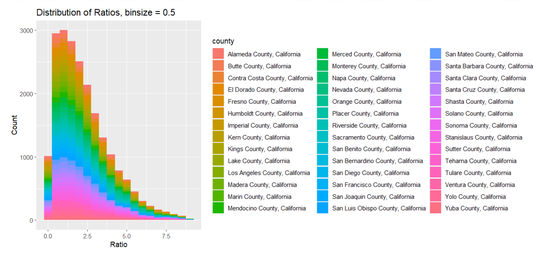
R^2 = 0.243

Five different models were built each using the same formula *Average Rent Cost ~ Estimate Number of Units*

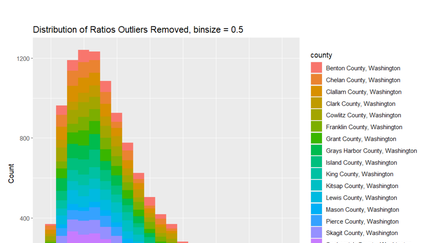
To better improve the R^2 value we use IQR Fence Outlier detection and create a model for each state.

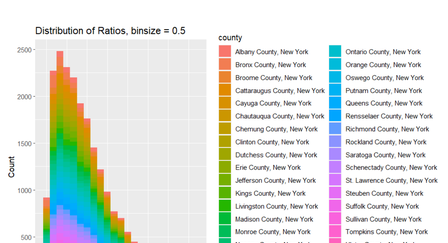
Random Forest

This one is special because we generated data points based off of the income distribution per county. Each entry is assigned a random income based on the income bracket.

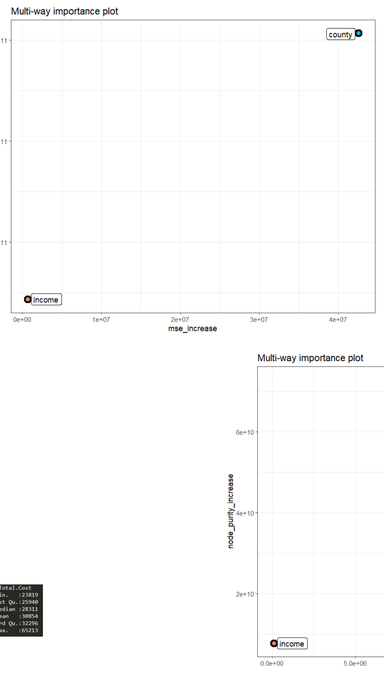


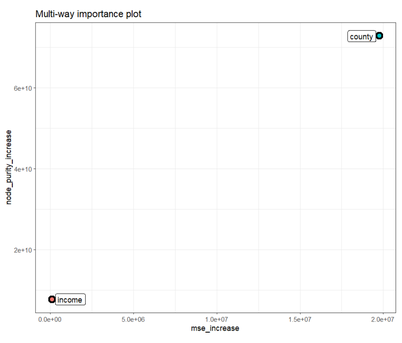
Since we generated the data, we need to relook at the distribution. We're looking at the ratio of the (Total Cost / Income)





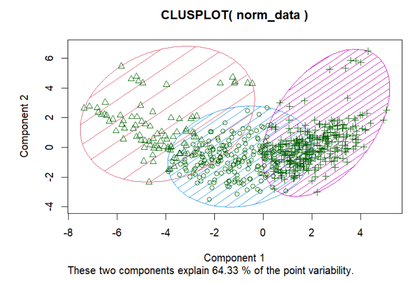
Results		
California	New York	Washington
Number of trees: 5506245	Number of trees: 50	Number of Trees: 50
Mean of Squared Residuals: 5506245	Mean of Squared Residuals: 2791977	Mean Squared Residuals: 2336829
% Var Explained: 93.78	% Var Explained: 88.93	% Var Explained: 81.37

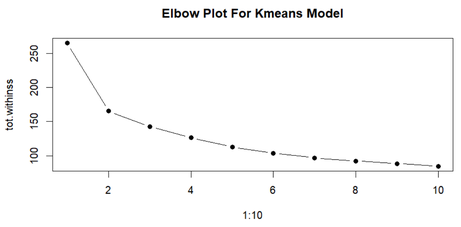




KMeans

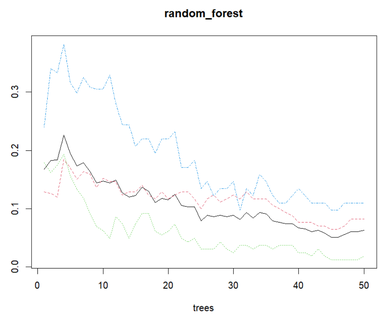
For the Kmeans model we know how many dependent variables we want so we set it to n = 3. While the elbow plot shows we need 4 clusters for the best error rate.



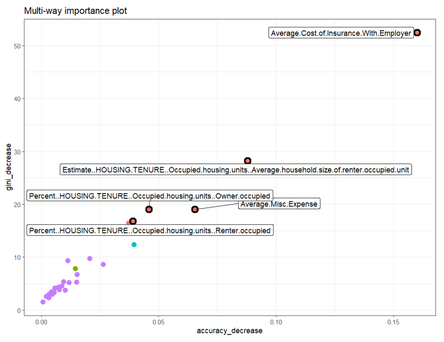


Results			
	1	2	3
CA	48	98	97
NY	37	143	53
WA	4	63	50
Accuracy: 40.6 %			

Random Forest



The Model: State~.
For this model we used all the income distribution data, some of the housing characteristics and all the cost-of-living features to predict which state someone lives. We were very successful in predicting where someone is from based off housing, income, and cost data.



The model had an accuracy of 97.2%. There was no need to tune the model as we already have a high rating. The next step is to analysis the importance variables and see what we can adjust so we don't overfit our model.

Predicted Results			
	CA	NY	WA
CA	72	0	3
NY	0	69	2
WA	0	0	30
Accuracy: 97.2%			

Conclusion

We came in wanting to predict certain features about the cost of living and where someone was from based on certain housing characteristics, their household income, and cost of living. Two out of the four models we were successful in achieve what we wanted. With both random forest models, we can predict their cost of living based off income and county they are from. For the other we can predict someone's household state.

The linear regression and Kmeans models can both be improved by changing the independent variables of each model.

Glossary:
Random Forest, Kmeans, Linear Regression– These are machine learning methods meant for large data analysis and creating a mathematical capable of making predictions.
R – A program to process data and perform statistical analysis
Package (P) or Library (R): software package to be loaded to perform extra tasks
Df, dataframe – Data manipulation structure in R & python pandas

Resources:
American community Survey - https://www.census.gov/programs-surveys/acs
Income Data – https://data.census.gov/table?q=Income&g=0100000US$0400000&tid=ACSST1Y2021.S1901
Housing Data - https://data.census.gov/table?q=Rent&g=0100000US$0400000&tid=ACSDP1Y2021.DP04
Importance Library – R - https://cran.r-project.org/web/packages/randomForestExplainer/index.html
GGPlot2 – R Package - https://ggplot2.tidyverse.org/