US Cost of Living Analysis

By: Alison Kam, Zhe Jiang, Carl Ge

Contributions

Alison Kam: Introduction, Data Description, Comparative Cost Analysis (Preliminary Analysis, Methodology, Results)
Zhe Jiang: Affordability Index Development
Carl Ge: Factor Effects Analysis on Affordability Index

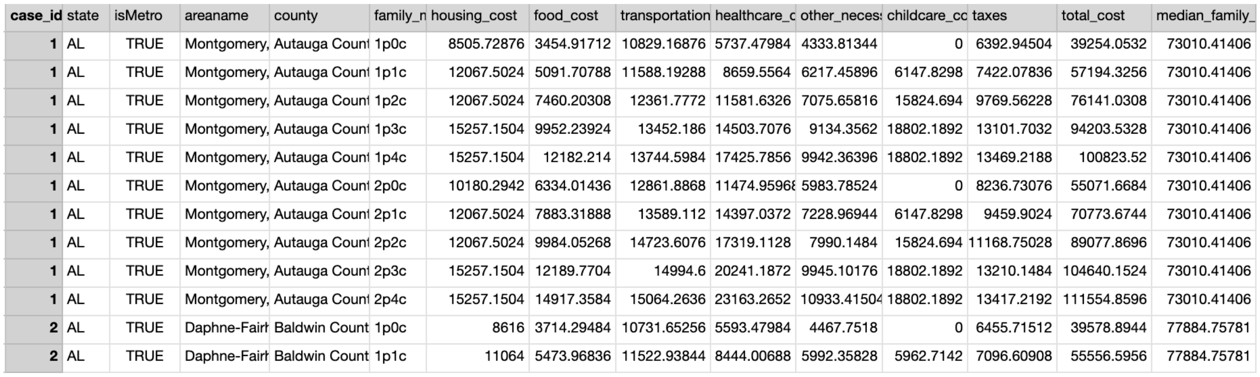
Introduction & Motivation

Over the past few years, everyone has felt a definite increase in prices, whether that be in housing, food, transportation, or any other costs. While Cost-of-Living Adjustments (COLA) have been implemented by Social Security to keep up with growing costs, many families are still left struggling to budget their money to cover all expenses. 2020 in particular saw a great increase in prices in the United States, partially due to the spread of COVID-19; the cost of living increased faster than the cost of living of the official CPI in that year. Many people could not afford to pay the increasing costs in their geographical area and around 8.94 million people moved homes. Therefore, we aim to analyze the affordability of living in different counties in the United States by taking into account different factors such as household size, income, and various costs. We will quantify affordability from the following three approaches: Comparative Cost Analysis (decide the most significant factors), Affordability Index Development (create index to rank the counties), and Impact of Household Composition on Affordability. Focusing on these approaches, we ask the following questions about the US Cost of Living.

Research Questions

1. How do living costs vary across different U.S. counties, and what are the key drivers of these differences?
2. How does household composition (with vs. without children) impact the affordability of living in different counties?
3. Can we develop an Affordability Index to rank counties based on a combination of living costs and median family income?

Data Description



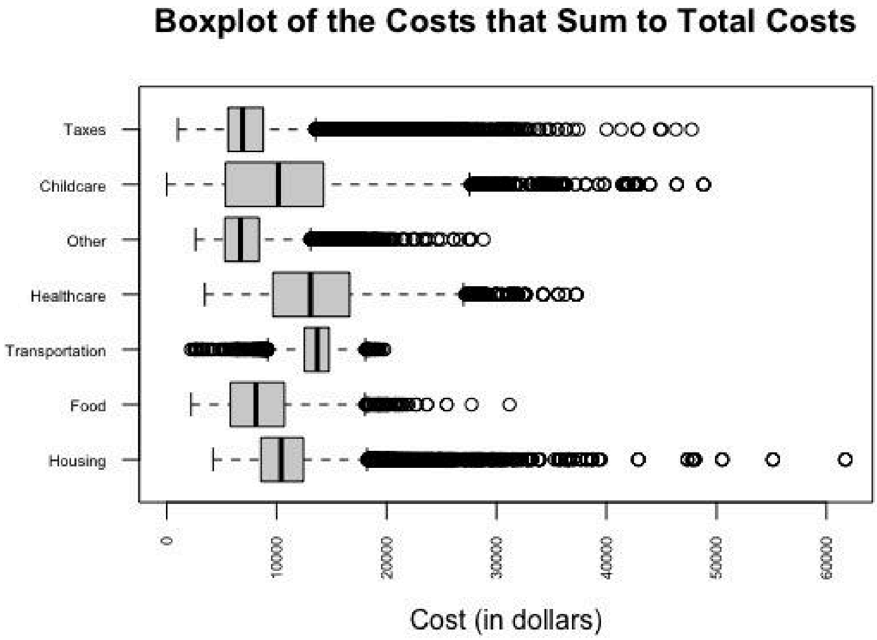
case_id	state	isMetro	areaname	county	family_n	housing_cost	food_cost	transportation	healthcare_c	other_neces:	childcare_cc	taxes	total_cost	median_family_
1	AL	TRUE	Montgomery,	Autauga Count	1p0c	8505.72876	3454.91712	10829.16876	5737.47984	4333.81344	0	6392.94504	39254.0532	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	1p1c	12067.5024	5091.70788	11588.19288	8659.5564	6217.45896	6147.8298	7422.07836	57194.3256	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	1p2c	12067.5024	7460.20308	12361.7772	11581.6326	7075.65816	15824.694	9769.56228	76141.0308	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	1p3c	15257.1504	9952.23924	13452.186	14503.7076	9134.3562	18802.1892	13101.7032	94203.5328	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	1p4c	15257.1504	12182.214	13744.5984	17425.7856	9942.36396	18802.1892	13469.2188	100823.52	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	2p0c	10180.2942	6334.01436	12861.8868	11474.95968	5983.78524	0	8236.73076	55071.6684	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	2p1c	12067.5024	7883.31888	13589.112	14397.0372	7228.96944	6147.8298	9459.9024	70773.6744	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	2p2c	12067.5024	9984.05268	14723.6076	17319.1128	7990.1484	15824.694	11168.75028	89077.8696	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	2p3c	15257.1504	12189.7704	14994.6	20241.1872	9945.10176	18802.1892	13210.1484	104640.1524	73010.41406
1	AL	TRUE	Montgomery,	Autauga Count	2p4c	15257.1504	14917.3584	15064.2636	23163.2652	10933.41504	18802.1892	13417.2192	111554.8596	73010.41406
2	AL	TRUE	Daphne-Fairl	Baldwin Count	1p0c	8616	3714.29484	10731.65256	5593.47984	4467.7518	0	6455.71512	39578.8944	77884.75781
2	AL	TRUE	Daphne-Fairl	Baldwin Count	1p1c	11064	5473.96836	11522.93844	8444.00688	5992.35828	5962.7142	7096.60908	55556.5956	77884.75781

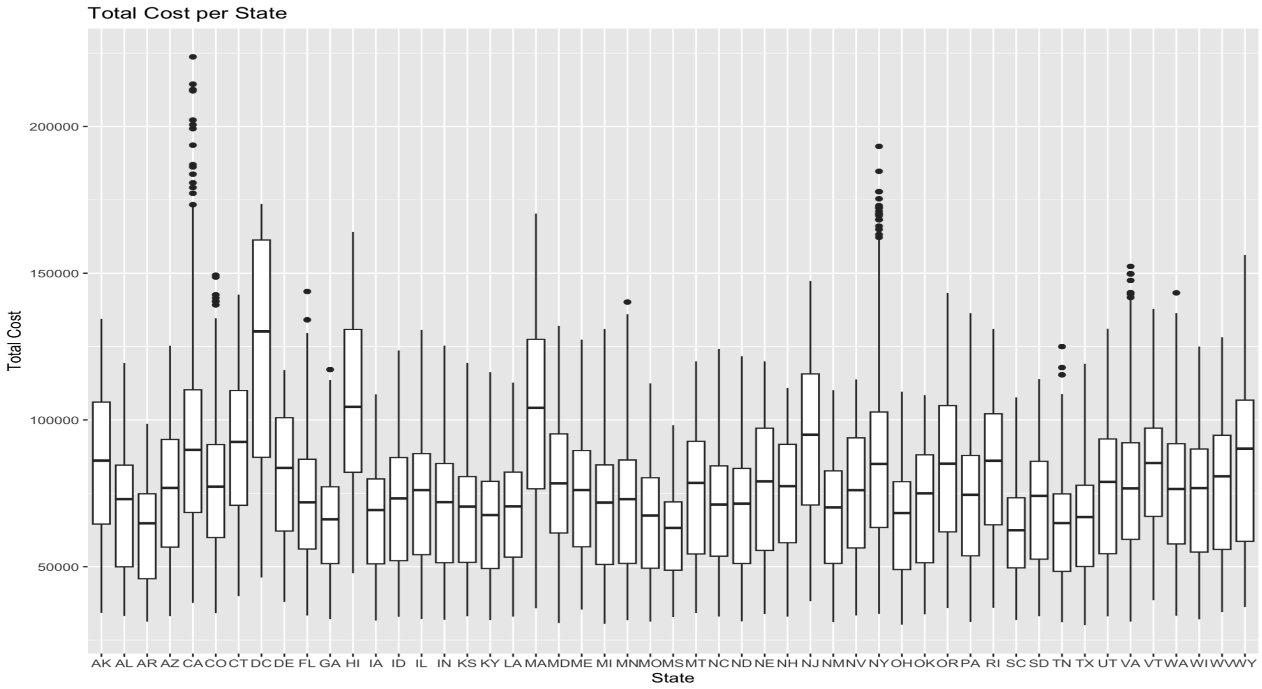
Above is a preview of the dataset we will be using to perform our analysis. The dataset includes 31,430 observations and 15 different variables. It samples 1877 counties in the US, with 10 households (8 with children and 2 without) sampled in each county. The variables in the dataset include case_id, state, isMetro, areaname, county, family_member_count, housing_cost, food_cost, transportation_cost, healthcare_cost, other_necessities_cost, childcare_cost, taxes, total_cost and median_family_income. case_id is a unique identifier for each areaname, in which a number is assigned corresponding with a unique county. state is the state where the county is located, and isMetro indicates whether or not the county is part of a metropolitan area. areaname lists the name of the area formatted as City, State, Metropolitan Area (if isMetro is true). County gives the name of each county. family_member_count is the number of family members in the household, with p representing parent and c representing children. The number of parents range from 1 to 2, and the number of children ranges from 0 to 4. housing_cost, food_cost, transportation_cost, healthcare_cost, other_necessities_cost, childcare_cost are the estimated annual cost for the family type in the county of housing, food, transportation, healthcare, other necessities, and childcare respectively. taxes is the estimated annual tax amount. total_cost sums the housing, food, transportation, healthcare, other necessities, childcare, and tax costs as the total estimated annual cost of living for the family type in the county. Last, median_family_income gives the median annual income for the county. The data is taken from the Family Budget Calculator by the Economic Policy Institute (EPI) and is based on 2020 dollars.

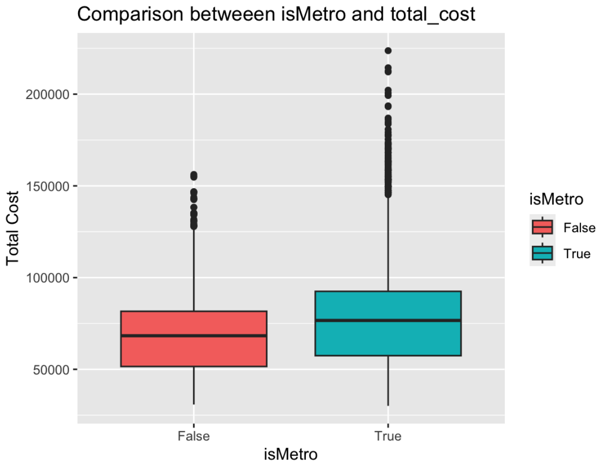
Comparative Cost Analysis:

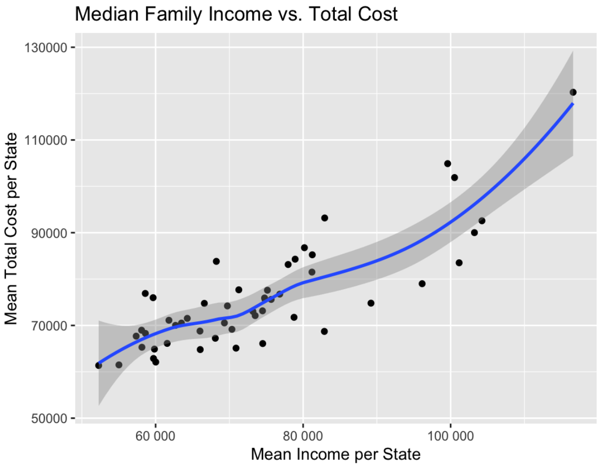
Preliminary Analysis:

With comparative cost analysis, we aim to determine the significant factors that contribute to the total cost of living across counties. Before applying methodology, we conduct preliminary analysis on the data. We first take the summary statistics of the data. Since the total_cost is the sum of the various costs, a larger mean or median of a cost variable may suggest that it is a cost that weighs heavier on the total cost. Our findings from looking at the summary statistics can be displayed using boxplots of the cost variables. From this, we see that transportation has the highest median value, with healthcare just below it. On the other hand, taxes and other necessities costs had the lowest median value. While transportation has the highest median, it is the only cost that has many values that lie below the minimum. All of the other costs have its outliers above the maximum, most notably housing, taxes, and childcare. Housing has the greatest spread. Transportation has the smallest Interquartile Range (IQR), so a large majority of people have around the same transportation cost. This gives us an idea of the weight of costs when taking into account all of the data observations before looking specifically at the differences between the states and counties.





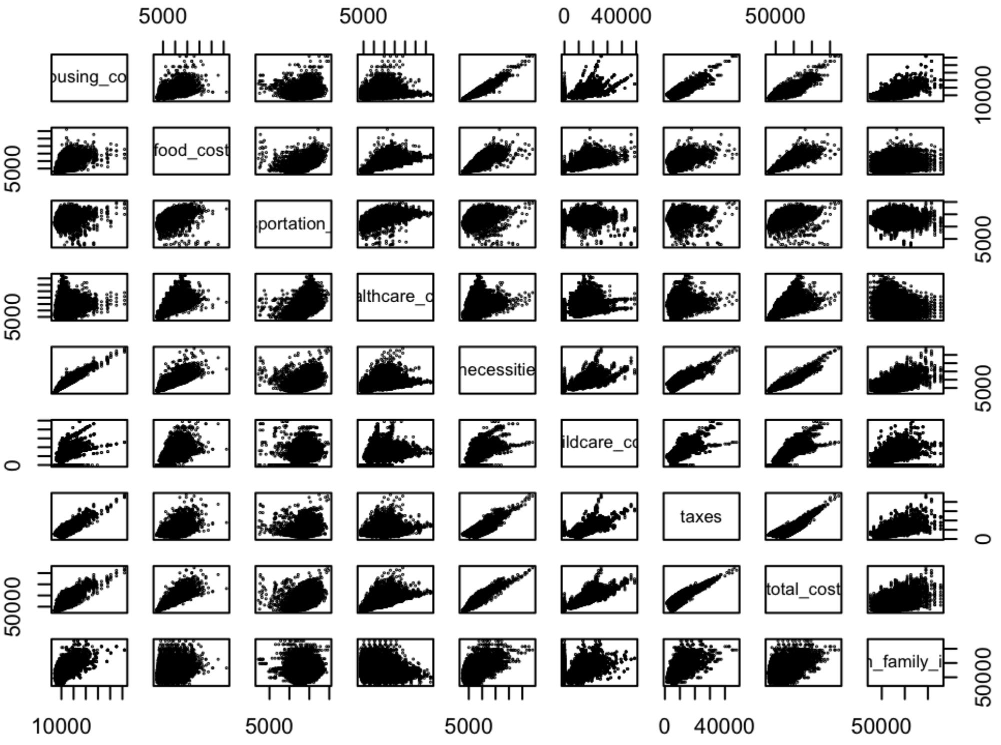




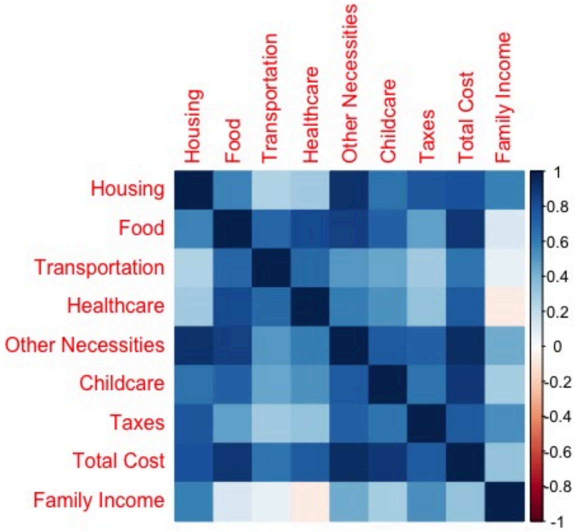
Now, we look at the total cost against each state. From the Total Cost per State boxplot, we see that DC has the highest median total cost while AR, MS, and SC have some of the lowest median total costs. The states CA and NY have many outliers beyond the max, indicating that some families have total costs far from the state median.

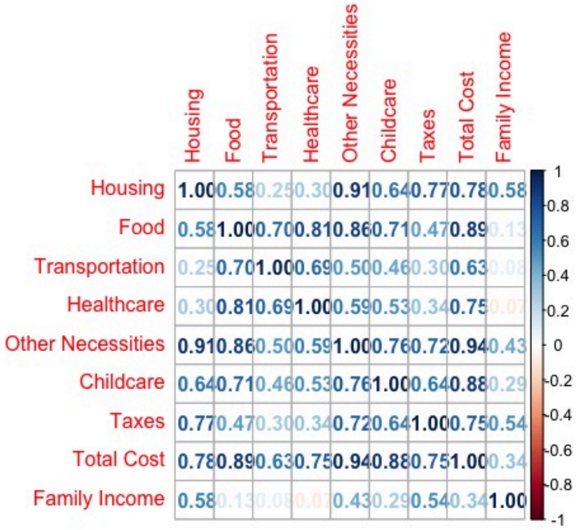
We also check if there is correlation between isMetro and median_family_income with total_cost. Based on the first graph, it is hard to determine at this point if being in a metropolitan area is correlated with the total cost of living. From the trend line in the second graph, it seems possible that median family income and total cost has a positive correlation.

Methodology



We create a scatter plot matrix and correlation matrix to look at the relationships between the quantitative variables. From the scatter plot matrix, we look for linear relationships. In the correlation matrix, we look for higher numerical values to indicate strong correlation. If there is strong correlation between two variables, it can be said that the two costs are related.





	Housing	Food	Transportation	Healthcare	Other Necessities	Childcare	Taxes	Total Cost	Family Income
Housing	1.00	0.58	0.25	0.30	0.91	0.64	0.77	0.78	0.58
Food	0.58	1.00	0.70	0.81	0.86	0.71	0.47	0.89	0.13
Transportation	0.25	0.70	1.00	0.69	0.50	0.46	0.30	0.63	0.08
Healthcare	0.30	0.81	0.69	1.00	0.59	0.53	0.34	0.75	0.07
Other Necessities	0.91	0.86	0.50	0.59	1.00	0.76	0.72	0.94	0.43
Childcare	0.64	0.71	0.46	0.53	0.76	1.00	0.64	0.88	0.29
Taxes	0.77	0.47	0.30	0.34	0.72	0.64	1.00	0.75	0.54
Total Cost	0.78	0.89	0.63	0.75	0.94	0.88	0.75	1.00	0.34
Family Income	0.58	0.13	0.08	0.07	0.43	0.29	0.54	0.34	1.00

All of the costs have a positive correlation coefficient, which means that as one cost increases, the other costs will as well.

Results

The scatter plot matrix and correlation matrix show us that most of the costs have a strong correlation with one another, where we define a strong correlation as one between the values 0.5 and 1. Looking against total cost, the costs that have the highest correlation are other necessities, childcare, and food. Interestingly, median family income has a relatively low correlation with total cost, indicating that if a family has a higher income, it is not necessarily true that they will have greater total costs. Looking at family income against all costs, there is a low correlation, supporting the result that there is a low correlation between income and total cost. Except for family income, other necessities costs have a high correlation with all other costs. From this, we can conclude that food, other necessities, and childcare contribute the most to total cost. Transportation has the lowest correlation with total cost, indicating that transportation costs tend to not increase as total costs increase.

Affordability Index Development:

In this section, we aim to define affordability by using a numerical value to quantify it.

Idea:

We define affordability by developing a measure by dividing the median family income by the total cost. Since the range of each costs vary, we use normalization to re-scaling by using the following equation:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

Because of the different ranges for each predictors, instead of directly normalizing the total cost, we are rescaling each element and summing them up to become the new normalized total cost. Those costs include housing cost, food cost, transportation cost, healthcare cost, other necessities cost, childcare cost and taxes. After rescaling the family median income, we can develop a value of the ratio between normalized family median income and normalized total cost. We call this value as "Affordability Index" and calculate its value by:

$$\text{Affordability Index} = \frac{housing\_cost^{*} + transportation\_cost^{*} + \ldots + taxes^{*}}{median\_family\_income^{*}} \quad \text{(* - value after normalization)}$$

Note that the family median income is all the same in one county regardless of the family size, so it is a measure that reflects overall economical standing. Instead of calculating the index for each family size, we are averaging the total cost for each family size after normalization to get the index. Thus, the affordability index is a scale between zero and one that measures the affordability in each county regardless of family size.

Procedure:

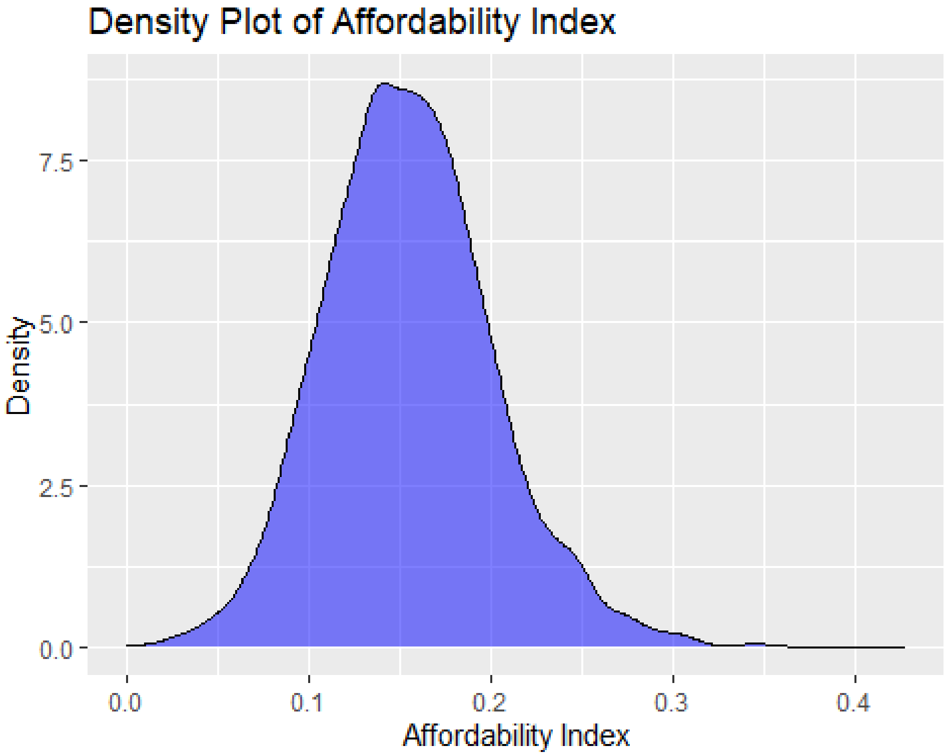
1. Remove all the NA. value for the data set.
2. Create a function of normalization and apply it to all the numeric predictors.
3. Add all the normalization costs together and take the average of the cost for each county.
4. Calculate the affordability index and sort them in descending order.

This is the Top 5 county with affordability Index:

	case_id	County	Total	Income	Index
1838	1865	Clinton County	1.977843	0.8468594	0.4281732
2955	2984	Asotin County	2.552528	1.0000000	0.3917685
2856	2885	Gloucester County	2.575763	0.9175812	0.3562367
2091	2118	Lucas County	1.951594	0.6792613	0.3480547
1216	1226	Worcester County	2.337415	0.8095681	0.3463519

Data Visualization:

Because the data set is very large even disregarding the family size, we choose to create a density plot to explore its overall distribution. By using ggplot2, we get:



From the density plot, the overall distribution is slightly right skewed, indicating there are a few countries where the median family income is significantly higher compared to the cost of living, which results in a high affordability index. Furthermore, there is a balance between income and cost for the majority of the county since the peak is not near zero.

Factor Effect Analysis on Affordable Index:

Idea:

Through the previous section, we get the Affordability Index of different families. We now have a keen interest in the factors that influence the Affordability Index. Since the various costs of each family generally depend on the number of people in the family, we decided to observe changes in the Affordability Index through "family_total_count" in the data.

Model Assumption:

First, we assume that family size affect the Affordability Index at the same time, so the model can be listed:

$$Y = \beta 0 + \beta 1 X1 + \epsilon ,$$

where Y represents the Affordability Index as the response, X_1 is the variable for the count of family members ("family_total_count"). For the coefficient parts, β_0 is the intercept of the model, representing the baseline level of the affordability index when both. β_1 is the coefficient for X_1, which measures the change in the Affordability Index for each additional family member.

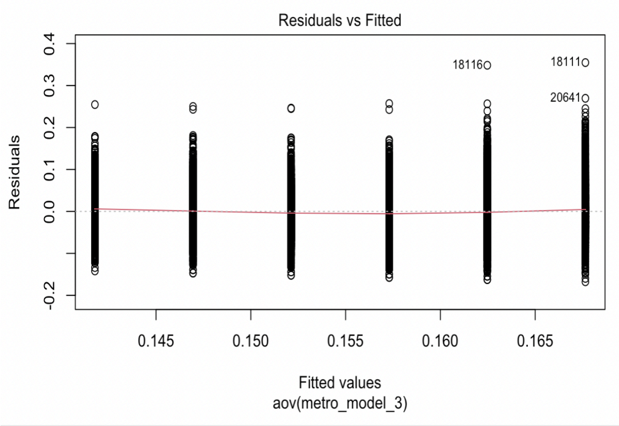
Hypothesis Test:

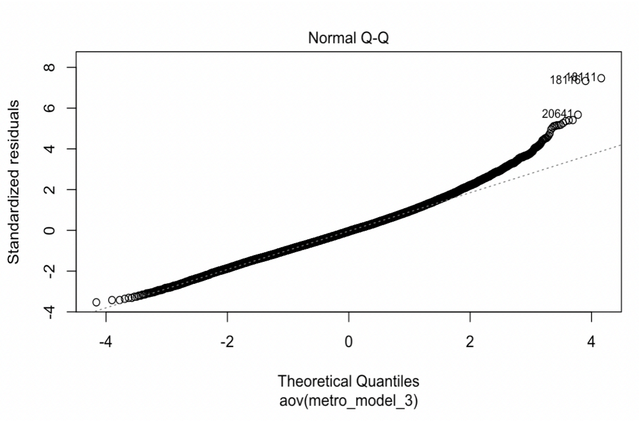
Since we do not know whether the variable of family size is significant, we need to perform a General Linear Test. We assume the "family_total_count" is not significant as Null Hypothesis and it is significant as Alternative Hypothesis.

H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$ (Decision Rule: P-value < α, reject H_0.)

We are taking α level = 0.05, we find the p-value for the F-statistic here is extremely small (<2.2e-16), where α are greater than this p-value. Thus, we reject the null hypothesis, which means the variable "family_total_count" is significant.

Model Diagnostics:



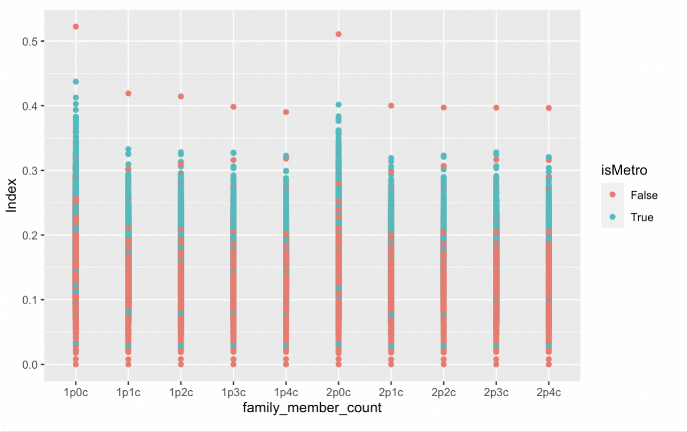


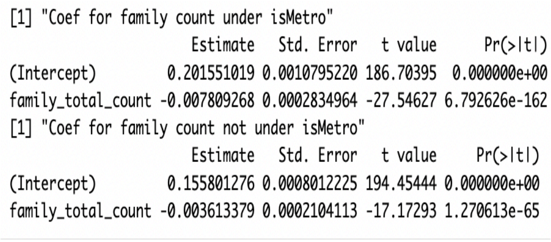
According to the Residuals vs Fitted plot, we can see that most of the points are around zero, and the spread of the variance is changing subtly while the fitted values increase, the red line is keeping horizontal all the way in the plot as well. Thus, we conclude that the model meets the equal variance assumption. For the Normal Q-Q plot, except for the points slightly off the fitted line at the upper portion of the plot, the whole pattern is approximately a line that is close to the fitted line. This suggests the model possesses normality.

Pairwise Comparison:

We will stratify the data by different factors to further explore the relationship between household composition and Affordability Index.

First, we are interested in the impact of different households on the Affordability Index in urban areas. We divide the data into two subsets according to urban areas and follow the steps to fit the model. We also visualized the distribution of the data.

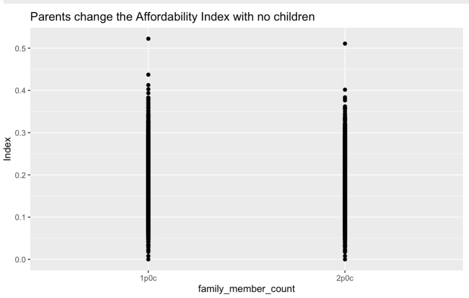


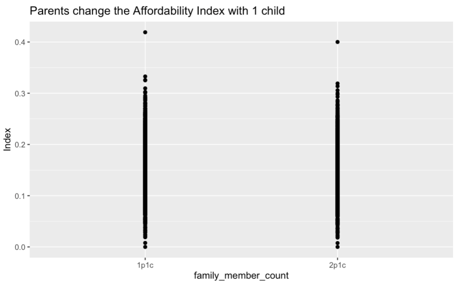


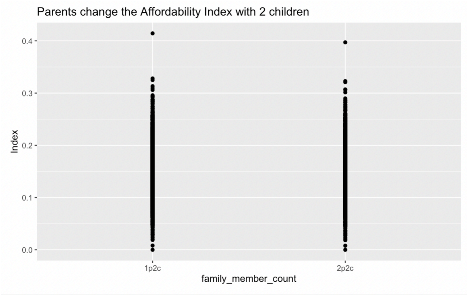
```
[1] "Coef for family count under isMetro"
                      Estimate   Std. Error   t value      Pr(>|t|)
(Intercept)        0.201551019 0.0010795220 186.70395  0.000000e+00
family_total_count -0.007809268 0.0002834964 -27.54627 6.792626e-162
[1] "Coef for family count not under isMetro"
                      Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)        0.155801276 0.0008012225 194.45444 0.000000e+00
family_total_count -0.003613379 0.0002104113 -17.17293 1.270613e-65
```

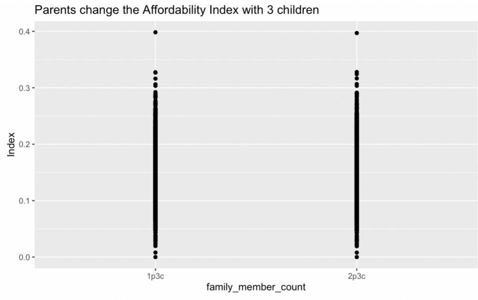
As shown in the figure, under each family composition, families living in cities have a higher Affordability Index than families living in non-urban areas. We can also see from the summary table on the right that the initial Affordability Index data of urban households is 0.05 higher than that of non-urban households. Although the slope of the index is smaller (-0.007 < -0.004)for the number of people in urban households, compared with the larger initial difference, the Index of urban households is generally higher than that of non-urban households regardless of the number of people and combinations of household composition.

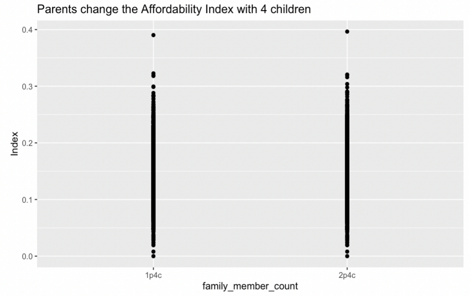
Next, we turned our attention to the main impact of the number of parents on the Affordability Index. Before examining the effects, we split the data into five groups based on the number of children in order to observe deeper changes.

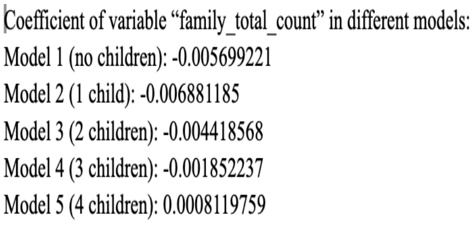








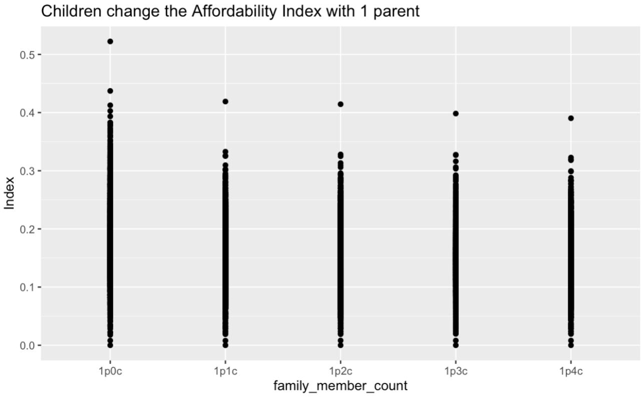


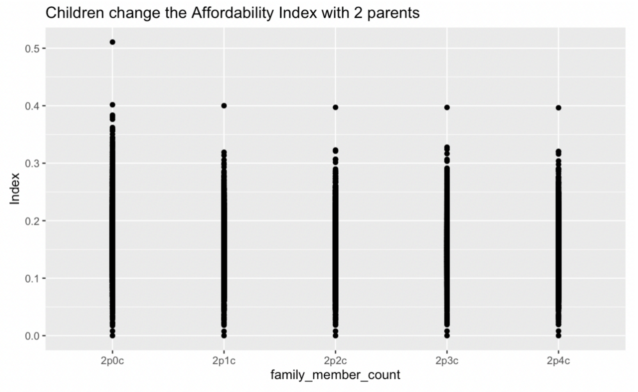


Coefficient of variable “family_total_count” in different models:
Model 1 (no children): -0.005699221
Model 2 (1 child): -0.006881185
Model 3 (2 children): -0.004418568
Model 4 (3 children): -0.001852237
Model 5 (4 children): 0.0008119759

By visualizing five different models, we know that families with only one parent generally have a larger Affordability Index. But looking closely at the five charts, we find that as the number of children in the family increases, the Index of a family containing two parents gradually approaches that of one parent.

Then, continuing our research, we stratified the data by the number of parents to see the impact of different numbers of children on the Affordability Index.





Coefficient of variable “family_total_count” in different models:
Model 6 (1 parent): -0.006272881
Model 7 (2 parents): -0.004467747

From the above two figures, we can see that an increase in the number of children will reduce the Affordability Index. But when we compare groups with different numbers of parents, we find that a greater number of parents slows down the slope of the total number of families variable on the Affordability Index.

Findings:

1. Generally speaking, the Affordability Index of urban households is higher than that of non-urban households.
2. When there are many children in a family, more parents will have a lighter burden than one parent (the Affordability Index becomes higher). However, when a family has no children or a small number of children, a person's Affordability Index will often be higher.

Sources

US Cost of Living Dataset
Family Budget Calculator