US COST OF LIVING ANALYSIS PROJECT

BY

Rohil Bhingarde

October 7th, 2023

Table of Contents

I. Executive Summary……..... Page 3

II. Evaluation of Existing Solutions……..…….. Page 3

III. Data Description and Preprocessing……... Page 5

IV. Modeling and Evaluation……...… Page 9

V. Summary……... Page 12

VI. Appendix…….……. Page 13

VII. Reference……....……. Page 17

VIII. Code……..……..…… Attached

I. Executive Summary

With our dataset, US Cost of Living, we were aiming to answer three business questions. Firstly, we wanted to see which counties have the most affordable places to live, food, transportation, healthcare, childcare, and other things people need. Next, we investigated how family size affects the estimated budget and found counties where bigger families have higher costs. Finally, we used the dataset to compare living standards and economic security in different US countries. Though it is possible that there are many factors outside of the dataset that impact the cost of living in the United States, we were interested to see how the different variables interacted from county to county.

The dataset consisted of information from 1,877 counties and metropolitan areas across the United States. The estimates in the dataset are based on the Economic Policy Institute's (EPI) household budget calculator and have a relatively high degree of confidence. Among this information were community-specific estimates for ten family types, made up of one or two adults with zero to four children. These estimates included the following fifteen variables: “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”.

While this dataset provides a relatively comprehensive perspective for analysis, we also recognize that it still has limitations, such as it may be time-sensitive, not all areas were captured or it failed to include all types of costs. Future research may explore these regions, as well as past trends. In summary, the results of our analysis have practical applications for household planning, local finance development, and other areas, to help them better understand and plan for the economic challenges in the future.

II. Evaluation of Existing Solutions

When identifying solutions by other teams on our same US Cost of Living dataset, we chose to look at the work of two users, Monika Noor Karima and Eugenio Schiavoni. We selected these specific solutions because we believe they produced the closest outcome to what we would be able to create without existing Python coding knowledge. Had we selected one of the more advanced solutions, it would have been more difficult to evaluate the modeling approach and find potential feedback.

Karima's work with the US Cost of Living dataset began with preprocessing the data to remove any missing values. We believed that there could have been an opportunity to further clean the data in the preprocessing phase by separating the family_member_count variable to have one column with parents and another with kids. This could have allowed for further analysis regarding how cost of living impacts children specifically across the various counties in the dataset.

Regarding the models Karima created for the analysis of the dataset, she found the counties with the highest and lowest living costs. According to her findings, the county with the highest total expenses was San Francisco, while the county with the lowest total expenses was Flint, Michigan. Karima next explored the correlation between income and each cost-related variable through the use of scatter plots, and she found that family income is most strongly correlated with food costs. She then looked at the impact family size had on the estimated budget. It was concluded that the more family members, children specifically, in a family, the greater the average estimated budget. According to Karima's analysis, the number of parents did not impact the average estimated budget but the number of children did. We believed that this was a possible shortfall in her analysis, so creating a similar, better model would help determine if this remained consistent.

The final step of Karima's analysis was creating a dataframe to decide if a county was affordable or not. Through the creation of a new "affordability" variable, each county was assigned either a "1.0," indicating that the county is considered affordable for the respective family size, or a "0.0". We believe that Karima's work with the US Cost of Living dataset provided a good foundation for basic knowledge and analysis of the data, but there are instances where the analysis could have been improved.

The second analysis of the US Cost of Living dataset that we looked at echoed some of the same analysis of Karima. Eugenio Schiavoni began his analysis by looking at the distribution of median family income through a histogram. Following this, Schiavoni looked at the count of records per state and the relationship between housing cost and median family income, which unsurprisingly had a strong, positive correlation. He continued to create various visualizations, we assumed, to familiarize himself with the dataset. In addition to the already listed visualizations, Schiavoni also made boxplots of the distribution of housing costs by state. The states that saw the greatest variance were California, New York, Texas, and Virginia. This did not come as a surprise because all of these states either have a major metropolitan area or are right outside of one.

The next visualization presented a stacked bar chart showing the distribution of metropolitan areas by state. Although this graph was easy to interpret, it did not contribute anything useful to the viewer. When considering this in our own graphs, we believed there was a more effective way to convey the same results. Similar to this visualization, the following output from Schiavoni's output showed information that was useful to the viewer but was not necessarily needed. The bar chart displayed the top ten areas with the highest frequency, which proved to be somewhat confusing. Had the code not been readily available in front of us, it is nearly impossible to tell what the frequency is counting. The stacked bar chart of necessity cost by state also presented a shortfall in the analysis, as the actual output of the visualization was not reflective of the title.

Schiavoni presented useful information to the viewer and gave a complete understanding of the dataset with his bar chart of median_family_income by state, line chart of the evolution of median_family_income across all cases, and correlation heatmap between all numerical values. As opposed to the previously discussed visualizations that left gaps for us as the viewers, the output of this code helped us gain a well-rounded understanding of all variables included and how they interact with one another.

In conclusion, we believed that Karima's and Schiavoni's solutions to the US Cost of Living dataset provided the most complete, comparable analysis to what we were able to generate with our existing Python. Both analyses offered insight into effective visualizations and coding skills to convey effective findings in the dataset. Overall, we believed it was important to look at both the positives and negatives in these existing solutions, and we are able to move forward feeling confident in our work done with the US Cost of Living dataset.

III. Data Description and Preprocessing

Data Description

This dataset we used provides a detailed analysis of the cost of living for households in different regions of the US state. For the original data set, each row of data represents the geographic location and approximate financial description of a different household. The first five columns of data give geographic information about the families, including the "state" in which they are located, which is indicated by a two-letter abbreviation, such as "AL" for Alabama, "isMetro" uses a boolean value to return whether the area is a metropolitan area or not, and more specific area

name and county to provide detailed geographic location for each row of data. The same case_id indicates the same area_name and county.

The next column focuses on the composition of the family, which is represented by family_member_count. For example, “1p0c” represents an adult childless family, where "p" stands for adult and "c" stands for child. The subsequent columns are the core data of this dataset, which mainly classify the annual expenditures of households, including housing_cost, food_cost, transportation_cost, healthcare_cost, other_necessities_cost, childcare_cost, and taxes. The sum of these data constitutes the household's total annual cost of living as total_cost. In addition, the dataset records an important measure of household income levels in each region, called "median_family_income."

Data Structure

Using well-structured data ensures the consistency and reliability of data in the first place. Structured data simplifies the process of data cleansing, and allows data to be easily accessed, retrieved, understood, and used by humans and software. Structured data should be stored clearly defined with data attributes and names. This data makes it efficient to analyze with analytical tools and algorithms while decreasing the likelihood of errors in the subsequent processing of the data.

To explore the data quality of our current dataset, we used the "info" function. Fortunately, as we can see from Figure 1, the output shows that our current dataset exhibits very high data quality: it contains 31,430 complete records without any missing values. The field data type matches the content of the variable, for example, the costs of the various categories are floating point numbers, which helps to decrease the subsequent process of converting the data type.

Data Cleaning

Although the current data already has a relatively high integrity, it is still necessary to perform data cleaning. The first step is to identify and clean up the missing values, we use the "isna()" function to return a boolean DataFrame of missing values. Then we use the "sum()" function to count the number of missing values in each column so that we can visually determine the distribution of missing values. Although there are many ways to handle missing values, such as removing them completely, calculating the mean, or median, or padding the data before and after. In this case, we choose the "dropna()" function to remove all rows containing missing values to maintain the integrity of the data. The second step is to convert the field containing the boolean variable "isMetro" to an integer (1/0) to represent true or false. These are intended to make the

subsequent calculations more uniform and easier. The final step is to create a new variable to infer the size of the family by calculating the sum of the parents and children, allowing us to easily analyze the family as a whole, rather than as individuals. So far the preparation of the data is complete.

Data Basic Statistics

Linear Regression

Before constructing a model for predictive analysis, we need to have an understanding of the relationship between the variables and the research topic as well as their basic statistical characteristics. We elected to use regression analysis and three visualizations to assist in further understanding.

First, we wanted to focus on the impact of the different cost of living factors mentioned in the variables on the total household costs. To explore whether there is a significant relationship between each of the independent variables and the total household costs. We focus on the following key economic indicators: household size, housing costs, food costs, transportation costs, health care costs, child care costs, taxes, and median household income. We build the model in Python using "statsmodels.api" package, using the key economic indicators just mentioned as explanatory variables "Y", and then setting total expenditures as the response variable "X", and adding an intercept term to the explanatory variable dataset “X” using the "sm.add_constant" function. After adding the intercept term in X, we fitted the OLS model to our variable to calculate the optimal coefficients. From the results in Figure 2, each expenditure significantly affects the total expenditure.

Heatmap

Next, we used correlation heatmaps to represent the interrelationships between these economic indicators just mentioned. The method of calculation was to compute the Pearson correlation coefficients between the indicators using Python's “pandas” library, and then generate a heatmap using the “matplotlib” and “seaborn” libraries to visualize the strength of the correlation between these indicators. The Heatmap demonstrates the correlation between the variables, with blue representing negative correlation and red representing positive correlation, and the temperature of each color representing the magnitude of the correlation coefficient. The values in the cells are their correlation coefficients.

From the Heatmap in Figure 3, we can see that family_size shows a strong positive correlation with food_cost, transportation_cost, healthcare_cost, and total_cost all showing strong positive correlations. This may reflect the fact that the larger the number of household members, the higher the expenditures on food and healthcare and total cost naturally become. It also suggests that larger households and those with higher food expenditures are also likely to spend more on transportation. The high positive correlation between housing_cost and total_cost suggests that housing expenditures are an important component of total household costs. There is also a high correlation between taxes and housing expenditures, which may indicate that the level of taxes and housing payments are closely connected. It is worth noting that the correlation between median_family_income and other cost indicators is low, and that households at different income levels may have different expense structures.

Boxplot

The boxplot(Figure 4) shows us the variations in the distribution of cost of living across household size categories. Each box represents a household size category and depicts us five key statistical points. Points beyond the box represent outliers, and these usually point to unusually high costs of living.

The boxplot of small households shows a relatively concentrated distribution of their cost of living, as most small households are tightly clustered around the median, which means that their cost of living fluctuates at a low level. In contrast, the boxplots for medium-sized households show a wider distribution of costs, suggesting that medium-sized households have more significant differences in their cost of living. Medium-sized households experience greater variability in their daily expenses. In the boxplots for large households, we observe significantly more outliers, which represent unusually high costs of living. This implies that there may be other influences in large households that are not captured in the current model.

Barplot

Finally, we wish to use bar charts to view the average cost of living in different regions of the United States. We use the average performance of key economic indicators across regions to analyze the impact of geographic differences on household economics. We first used the "map" function to map each state in the dataset to its corresponding region based on state affiliation. Then, we grouped the data by region using the "groupby" function and calculated the average cost for

each region. Finally, we generated a bar chart to visualize these average costs using the "matplotlib" library.

Each bar cluster in the bar chart represents a region, and different colors indicate different types of costs. The height of the bars reflects the average of the corresponding costs. The NORTHEAST and WEST regions have relatively high-cost averages, while the MIDWEST and SOUTH regions have relatively low total costs. As with the results shown in the previous heat maps, housing expenses are high in areas with high total costs. Other costs are also high on average in the NORTHEAST and WEST regions, but there is not a significant difference between regions for food costs and transportation costs.

IV. Modeling and Evaluation

Following the data description and preprocessing, we explored various algorithms to assess how the independent variables influence or predict the overall cost of living for households in the US. We chose a mix of algorithms for this task including Ordinary Least Squares, Random Forest Regressor, and XGB Regressor. To do the evaluation, we create a list that contains several independent variables which include family_size, housing_cost, food_cost, transportation_cost, healthcare_cost, childcare_cost, taxes, median_family_income, and create a subset DataFrame X from this selection. Concurrently, we create a series Y to represent our dependent variable, total_cost. The OLS regression model results show that there is a significant relationship between most explanatory variables and the total household costs. In particular, the coefficients on the housing cost and food cost indicators are greater than one, clearly showing that increases in these expenses are strongly and positively associated with increases in total household costs. However, the coefficient on median household income is the negative coefficient, which may indicate that as income increases, households can manage their day-to-day expenses more effectively. It is also interesting to note that the model shows a very high degree of fit (R-squared value is 1), which is very rare in actual cases, which could mean that the model is overfitting the data, and also a side note that these costs and total expenditures are all closely related. (Figure 2)

Moving to the Random Forest Regressor and XGBRegressor, we began with an empty list to hold the algorithm abbreviations. We split the data to 70% on training, 30% on testing, and ensuring reproducibility with a random state of 42, then implemented a for loop to iterate through each algorithm and output both its types and the corresponding accuracy measurement. The result for the

Random Forest Regressor shows an MAE (Mean Absolute Error) of approximately 652.63, an MSE (Mean Squared Error) of about 1,311,672.85, and an R-squared value of approximately 0.997. Similarly, the XGB Regressor yields an MAE of approximately 682.58, an MSE of about 1,284,554.49, and an R-squared value of approximately 0.997. Both regression models indicate a very high level of predictive accuracy with an R-squared value of approximately 0.997. This coefficient of determination indicates that the independent variable included in the models can explain most of the changes in the output i.e., the overall cost of living for households.

```
RandomForestRegressor

MAE:  652.6301978001275

MSE 1311672.852677671

R_squared 0.997236329987052
XGBRegressor

MAE:  682.5798030726183

MSE 1284554.4900783629

R_squared 0.9972934678666406
```

After running the regression model, we use SHAP analysis to determine the importance of each independent variable in the model. SHAP value assigns a numerical value to each independent variable in the model, reflecting the strength and direction of that feature's influence on the model's predictions which shows the significance of each feature contributing to the model's output. We utilized both bar chart and summary plots from SHAP to extract insights into our model, the bar chart gives the average impact of each independent variable on the model's output, and the summary plot provides more details of how each independent variable affects the prediction.

The bar chart shows that 'food_cost' has the highest mean absolute SHAP value which implies it is the most influential factor in determining the cost of living, with an average impact of +7319.31 on the model's output. The following factor is 'taxes' which also has a significant positive impact on the cost of living (+4544.62). The next in line are 'Childcare_cost', 'family_size', and 'housing_cost', with significant contributions of +3250.4, +2206.57, and +1830.48, respectively. 'Healthcare_cost' and 'transportation_cost' are also significant but not as much as the top factors, with mean SHAP values of +1569.16 and +362.59. Furthermore, the 'region_South' and 'region_West' have a minor impact on the target variable, with mean SHAP values of +18.18 and

+17.83, respectively. At the bottom of the chart, there is a collective representation of three additional independent variables that were not individually as impactful; this combination bar has a mean SHAP value of +25.56 affects on the outcome. (Figure 6)

In the summary plot, every dot represents the SHAP value of a specific independent variable and its corresponding data point, spread over the possibility of the effects on the model. Each independent variable impact on the output is quantified by the horizontal axis, where placement to the right indicates a positive influence and placement to the left is a negative one. Additionally, the color indicates the value of independent variables, with one color representing the higher values and another representing the lower values. The order of the independent variable is based on the sum of the SHAP value which identified 'food_cost' as having the highest overall impact on the predicted total cost of living. Similarly, 'taxes' represent a deeper positive effect which has a more definitive positive relationship with the total cost of living. Specifically, higher taxes generally corresponded to higher costs in other areas, such as food and housing, while elevated food costs could reflect broader economic conditions. As we move down the plot, 'childcare_cost', 'family_size', and 'housing_cost' all show a significant effect on the model predictions of the cost of living. The summary plot highlights the complexity of healthcare_cost' and 'transportation_cost' are significant indicators as indicated by a tiger clustering of SHAP value. The regional variables such as 'region_South' and 'region_West' have minor but noticeable effects; their SHAP value clustering is close to the zero line. Moreover, this plot also accounts for the aggregated influence of three less impactful variables which the combination represents a moderate increase in the predictive outcome. (Figure 7)

Building on our comprehensive analysis result, we then focus on our most impactful variables as indicated by SHAP, narrowing down to 'family_size', 'houseCost_to_income_ratio', and 'taxes' for further regression modeling with the Random Forest and XGB Regressors. These models also got 70% of the data on training, the remaining 30% was used for testing, ensuring consistency with a random state set to 42 and iterating through both Random Forest Regressor and XGB Regressor algorithm respectively. As a result, both models achieved high predictive accuracy with R-squared values of approximately 0.967 for Random Forest and 0.969 for XGB, suggesting a strong explanatory power over the variance in household cost of living. The MAE and MSE for Random Forest were 2871.01 and 15532185.06, respectively; the XGB model performed slightly better, with MAE and MSE somewhat lower at 2739.70 and 14621897.16, respectively. These

results confirm our SHAP analysis conclusions by the chosen independent variables in the model's ability to accurately predict the cost of living.

```
RandomForestRegressor

MAE:  2871.0136035467863

MSE 15532185.056609195

R_squared 0.9672739784246656
XGBRegressor

MAE:  2739.7049254286676

MSE 14621897.165171377

R_squared 0.9691919378789462
```

V. Summary

In our Python project, we conducted a thorough analysis of the US Cost of Living dataset, with the primary objective of unraveling key insights related to county-wise affordability, the impact of family size on budget estimates, and a comparative assessment of living standards across diverse regions. The dataset, comprising data from 1,877 counties and metropolitan areas, featured fifteen variables derived from the Economic Policy Institute's household budget calculator. Our analysis addressed three core business questions, exploring the nuances of cost factors such as housing, food, transportation, healthcare, and childcare. We were cognizant of the dataset's limitations, including potential time-sensitivity and exclusions, and acknowledged these in our evaluation.

In evaluating existing solutions, we delved into the works of Monika Noor Karima and Eugenio Schiavoni. By identifying strengths and weaknesses in their analyses, we refined our approach to create a more sophisticated model. Moving to data description and preprocessing, we emphasized the importance of structured data for efficient analysis. Our data cleaning efforts ensured a dataset of high quality without missing values, setting the stage for subsequent in-depth analyses.

Utilizing various data analysis techniques, including basic statistics, linear regression, correlation heatmaps, box plots, and barplots, we gained a nuanced understanding of the dataset's characteristics and interrelationships between variables. Our modeling phase involved the application of Ordinary Least Squares (OLS), Random Forest Regressor, and XGBRegressor

algorithms. OLS regression highlighted significant relationships but suggested potential overfitting, while Random Forest and XGB Regressor models exhibited impressive predictive accuracy with R-squared values around 0.997.

To further dissect the model's intricacies, we employed SHAP (SHapley Additive exPlanations) analysis, unveiling the significance of each variable. Food cost emerged as the most influential factor, followed by taxes, childcare cost, family size, and housing cost. The resulting insights contribute to a comprehensive understanding of factors influencing the cost of living. Our conclusion emphasizes the practical applications of our analysis for household planning, local finance development, and broader economic insights. The provided visual appendix, including OLS regression results, correlation heatmaps, and SHAP analysis, enhances the transparency and interpretability of our findings. This project equips stakeholders with actionable insights, facilitating informed decision-making in household planning, local finance, and broader economic analysis.

VI. Appendix

Figure 1: Data structure information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31430 entries, 0 to 31429
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   case_id                 31430 non-null  int64
 1   state                   31430 non-null  object
 2   isMetro                 31430 non-null  bool
 3   areaname                31430 non-null  object
 4   county                  31430 non-null  object
 5   family_member_count     31430 non-null  object
 6   housing_cost            31430 non-null  float64
 7   food_cost               31430 non-null  float64
 8   transportation_cost     31430 non-null  float64
 9   healthcare_cost         31430 non-null  float64
 10  other_necessities_cost  31430 non-null  float64
 11  childcare_cost          31430 non-null  float64
 12  taxes                   31430 non-null  float64
 13  total_cost              31430 non-null  float64
 14  median_family_income    31420 non-null  float64
dtypes: bool(1), float64(9), int64(1), object(4)
memory usage: 3.4+ MB
```

Figure 2: OLS Regression Result

OLS Regression Results

Dep. Variable:	total_cost	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	4.647e+17
Date:	Thu, 07 Dec 2023	Prob (F-statistic):	0.00
Time:	04:51:01	Log-Likelihood:	1.5055e+05
No. Observations:	31420	AIC:	-3.011e+05
Df Residuals:	31411	BIC:	-3.010e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0004	0.000	-3.088	0.002	-0.001	-0.000
family_size	8.989e-05	4e-05	2.246	0.025	1.14e-05	0.000
housing_cost	1.3623	5.76e-09	2.36e+08	0.000	1.362	1.362
food_cost	1.3623	1.41e-08	9.65e+07	0.000	1.362	1.362
transportation_cost	1.0000	1.07e-08	9.36e+07	0.000	1.000	1.000
healthcare_cost	1.0000	5.11e-09	1.96e+08	0.000	1.000	1.000
childcare_cost	1.0000	3.58e-09	2.79e+08	0.000	1.000	1.000
taxes	1.0000	7.36e-09	1.36e+08	0.000	1.000	1.000
median_family_income	-5.466e-10	9.5e-10	-0.575	0.565	-2.41e-09	1.32e-09

Omnibus:	1010.671	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2780.464
Skew:	0.051	Prob(JB):	0.00
Kurtosis:	4.454	Cond. No.	7.99e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.99e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 3: Household Economic Indicators Correlation Heat Map

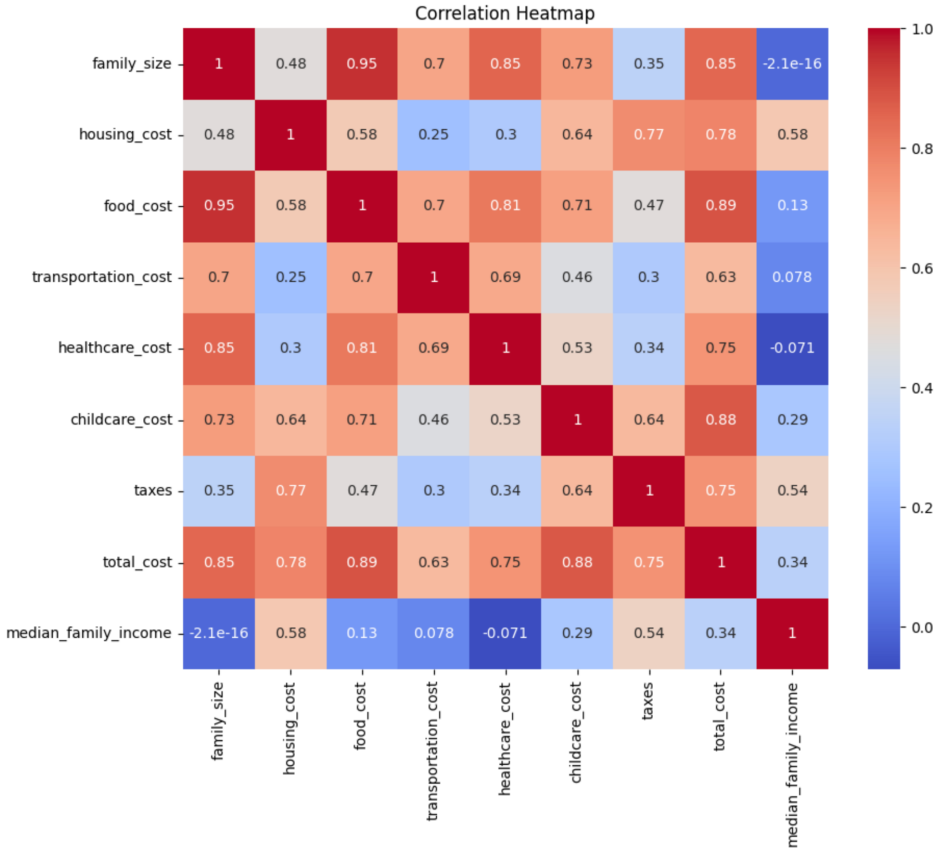


Figure 4: Cost of Living Variation by Three Different Family Size

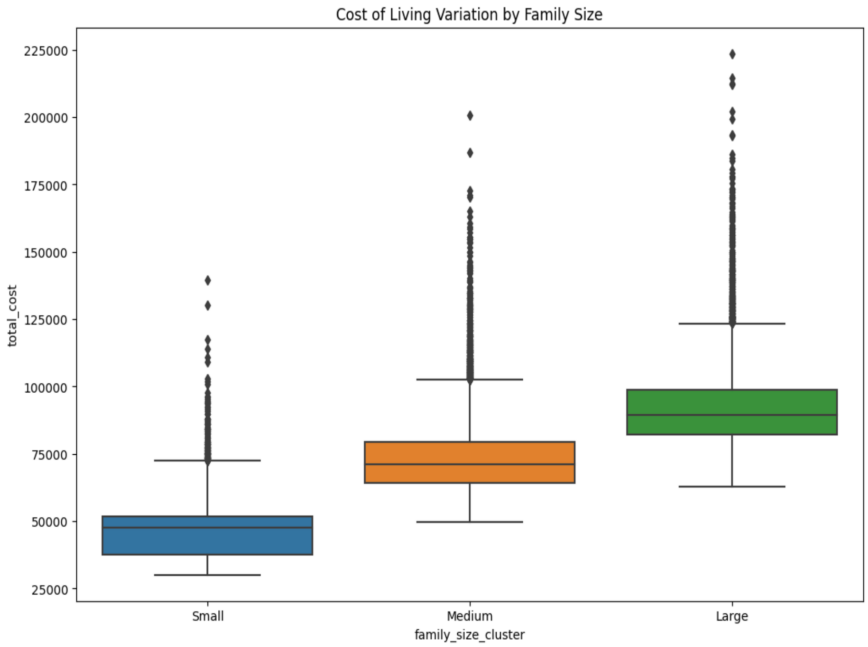


Figure 5: Average Cost of Living by Region

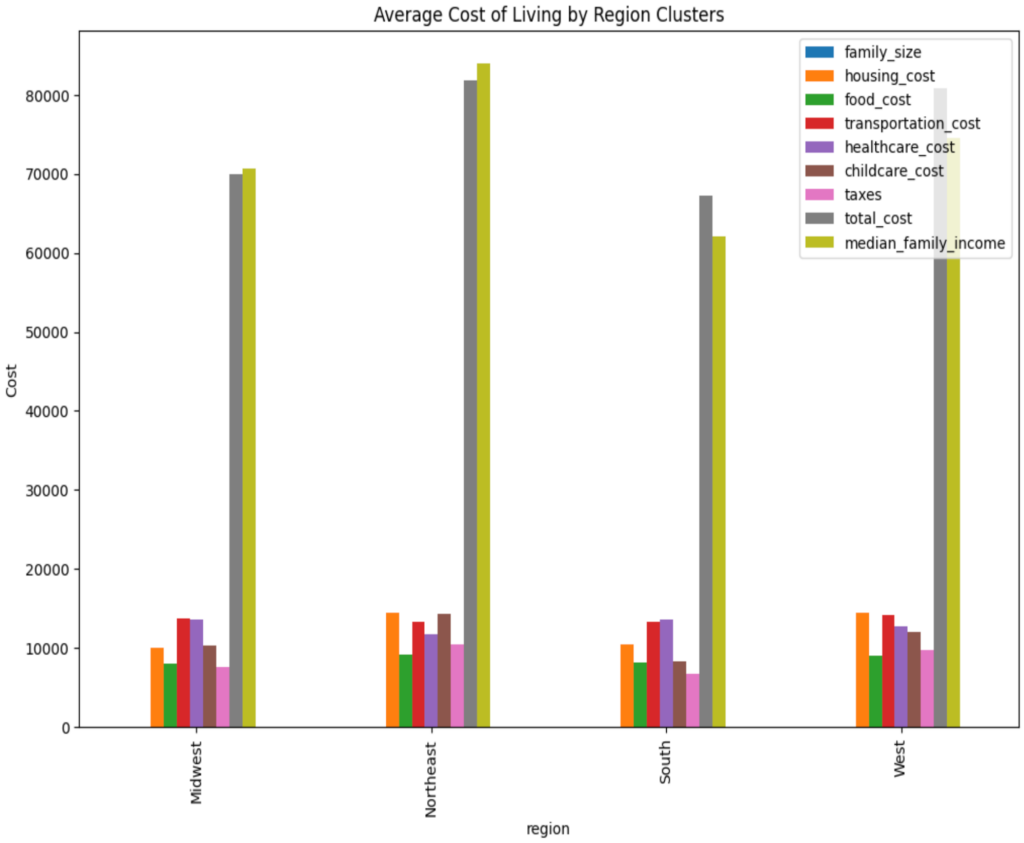


Figure 6: SHAP bar chart

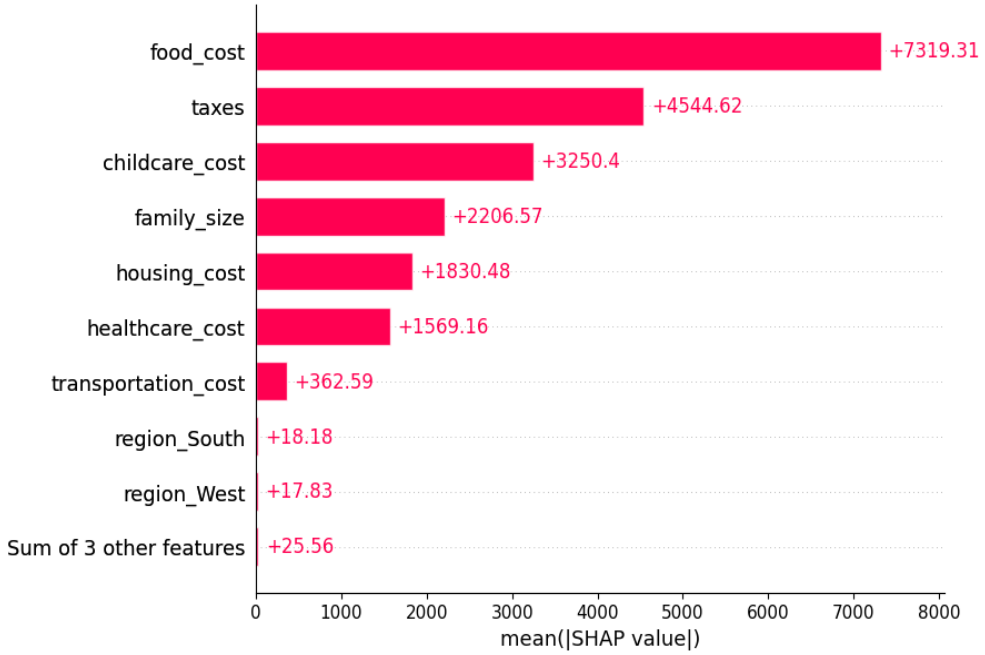
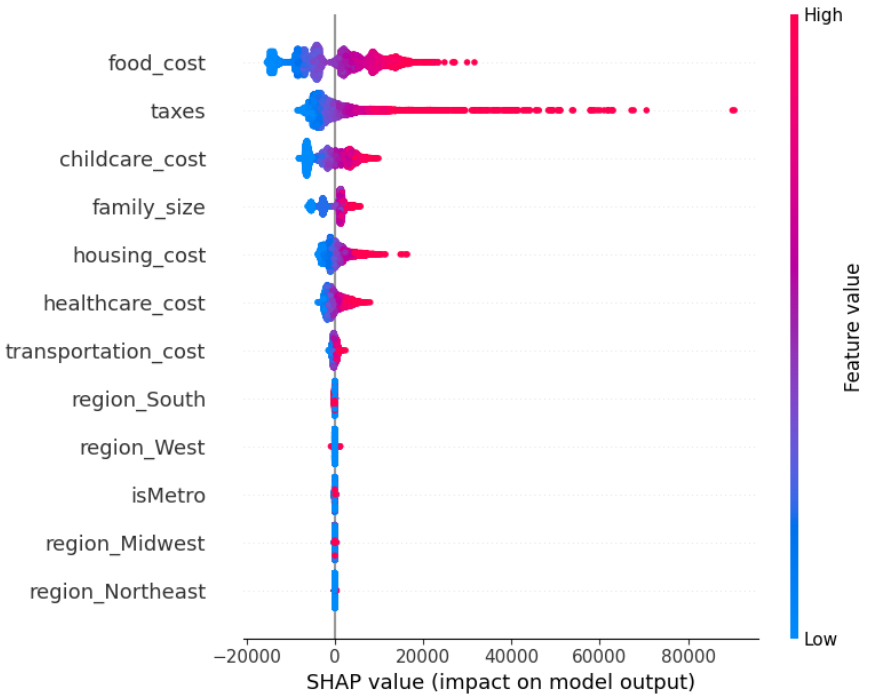


Figure 7: SHAP summary plot



VII. Reference

Dataset: https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties

VIII. Code

See attached Python notebook