### **University of Waterloo**

### 

### **Report on Group Project:**

### **Inflation - The Influence of Household Size and Geographic Location on Affordability**

**Course:**Foundations of Data Science - 0048-13 - Winter 2025  
*Delina Ivanova & Xuan Zhang*

**Group 17:**Faiza Qayoom, Dhruv Sethi, Eric Ramalheiro, Miguel Morales, Lucas Hele, Jack Hao

**Date:**April 14, 2025

**GitHub Repository (public):**<https://github.com/miguelmog10/data-science-group17>

## **1. Introduction**

This report discusses a group project examining the influence of household size, family structure, and geographic location (metropolitan vs. non-metropolitan areas) on the total cost of living and affordability. The hypothesis guiding this study suggests that the total cost of living is significantly impacted by how household size and structure interact with geographic location. Additionally, the relationship between the cost of living and household size is not necessarily linear, but may vary depending on the region and family dynamics. The group project was divided into distinct sections to explore various aspects of this hypothesis.

### **Hypothesis Overview**

*The affordability of different regions is significantly influenced by the comparison between the total cost of living and the average family income. This study will explore how household size, family structure, and geographic location impact the total cost of living, including whether the relationship between cost and household size is linear or nonlinear.*

1. **Metropolitan vs. Non-Metropolitan Areas:**
   * Metropolitan areas typically have higher living costs, particularly for housing, transportation, and goods. Non-metropolitan areas generally exhibit a lower cost of living relative to income. This disparity is largely due to differences in housing markets, public transportation availability, and the concentration of goods and services in urban areas.
2. **Key Expense Categories:**
   * Housing, transportation, and healthcare are the primary contributors to the cost of living. In metropolitan areas, housing tends to be the most significant expense, driven by high demand and limited supply. Transportation costs also tend to be higher in cities with less access to public transit. Non-metropolitan areas may see higher costs in some categories, but they often experience lower housing and transportation costs relative to income.
3. **Cost of Living and Household Size:**
   * The total cost of living may not be linearly related to household size—it could be **sub-linear** or **super-linear**. As household size increases, some costs (e.g., food, utilities) may scale less than proportionally, while others (e.g., housing) may increase at a faster rate. The impact of family structure (e.g., parents vs. children) on cost should also be considered, as different family dynamics may influence expenses differently.
4. **Exploring Linearity (or Non-Linearity) of Costs as a Function of Household Size:**
   * One of the central questions of this study is whether the total cost of living scales linearly with household size, or if it is sub-linear (costs increase less than proportionally with more members) or super-linear (costs increase more than proportionally). Additionally, the analysis will explore the following:
     + **Is the total cost a linear function of the number of household members, or is it sub-linear or super-linear?**
     + **Considering cost as a function of 1 variable (total household members) or as a function of 2 variables (e.g., parents and children).** Additionally, we will explore whether households follow the nuclear family model or if there are other family structures in the data (which could affect the relationship between size and cost).
     + **Does the function vary over location?** We will assess whether the relationship between household size and cost differs across metropolitan and non-metropolitan areas and whether we need to normalize the data by location.
     + **If we analyze individual costs separately (housing, transportation, healthcare), will this change the conclusions?** For example, housing may exhibit different scaling behavior compared to transportation or groceries.
5. **Regional Variation:**
   * The relationship between household size and cost may vary depending on geographic location. Metropolitan areas may see more pronounced increases in housing costs as household size grows, while non-metropolitan areas may experience more stable or lower increases. Normalizing for regional differences will be necessary to understand the broader patterns of affordability.

### **Hypothesis Conclusion:**

This hypothesis will guide an investigation into how cost-of-living patterns differ based on household size, structure, and regional location. It includes exploring whether costs increase linearly with household members or follow a different scaling pattern (sub-linear or super-linear), and how this relationship is affected by different family dynamics and geographical factors.

### **Research Questions:**

* **Metropolitan vs. Non-Metropolitan Areas:** How do cost-of-living patterns differ in urban and rural areas?
* **Key Expense Categories:** How do housing, transportation, and healthcare costs compare between metropolitan and non-metropolitan regions?
* **Cost of Living and Household Size:** Is the total cost of living directly proportional to household size, or does it vary?
* **Regional Variation:** How does the cost-of-living pattern change depending on geographic location?

**Project Team & Work Division:** The project was divided into five parts: Miguel handled Metropolitan vs. Non-Metropolitan Areas, Dhruv focused on Key Expense Categories, Faiza worked on Cost of Living and Household Size, Eric explored Linearity of Costs with Household Size, and Jack studied Regional Variation.

**Tools for the Project:** The analysis utilized Pandas, NumPy, Seaborn, Matplotlib, Statsmodels, Scikit-learn, GeoPandas, and Plotly for data manipulation, statistical analysis, machine learning, and visualization.

## **2. Data Preparation**

### 2.1 Package Requirements

This study was conducted using the Python Anaconda distribution (version 3.12.7), which offers a robust environment for data science and statistical analysis. While most required libraries are included by default, the geopandas library was additionally installed to support spatial data processing and geographic visualization.

### 2.2 Imported Modules

A range of Python libraries were employed to support the analysis. Pandas and numpy were used for data manipulation, while matplotlib and seaborn facilitated the creation of visualizations to identify patterns and regional differences in cost structures. Statsmodels enabled statistical regression and model diagnostics. Machine learning tasks such as linear regression modeling were handled using scikit-learn. Geopandas was essential for working with spatial data and visualizing geographic trends in affordability.

### 2.3 Reading the Data

The primary dataset contains detailed cost-of-living information for various U.S. counties and metropolitan areas. Each record includes location-based attributes such as state, county, and metro designation, along with household structure data (e.g., "1p2c" represents one parent with two children). It provides a breakdown of costs in categories like housing, food, transportation, healthcare, childcare, and taxes, as well as total annual cost and median family income, allowing for affordability comparisons.

*Note: The original dataset was obtained from kaggle:* [*US Cost of Living Dataset (1877 Counties)*](https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties)

### 2.4 Inspecting the Data

The dataset consists of 31,430 entries and 15 columns, including both numerical (e.g., cost categories and family income) and categorical (e.g., state, area name, household composition) variables. Descriptive statistics confirmed that the cost and income ranges align with known U.S. regional disparities. A check for duplicates found no exact matches, confirming data integrity. Categorical data revealed 51 states, 2,561 area names, and 1,877 counties, with household types spanning 10 combinations of parents and children and all 10 being equally common.

### 2.5 Initial Observations

The dataset, though large (31,430 rows), has a few inconsistencies, such as missing values in the median\_family\_income field (0.03% of entries). Some entries for childcare costs were marked as $0, likely for households without children. There are 51 states, including Washington D.C., and 1,877 unique counties (96% coverage of U.S. counties). Further investigation is required to address potential data inconsistencies and missing information, particularly in the family\_member\_count and areaname fields.

### 2.6 Rough Work

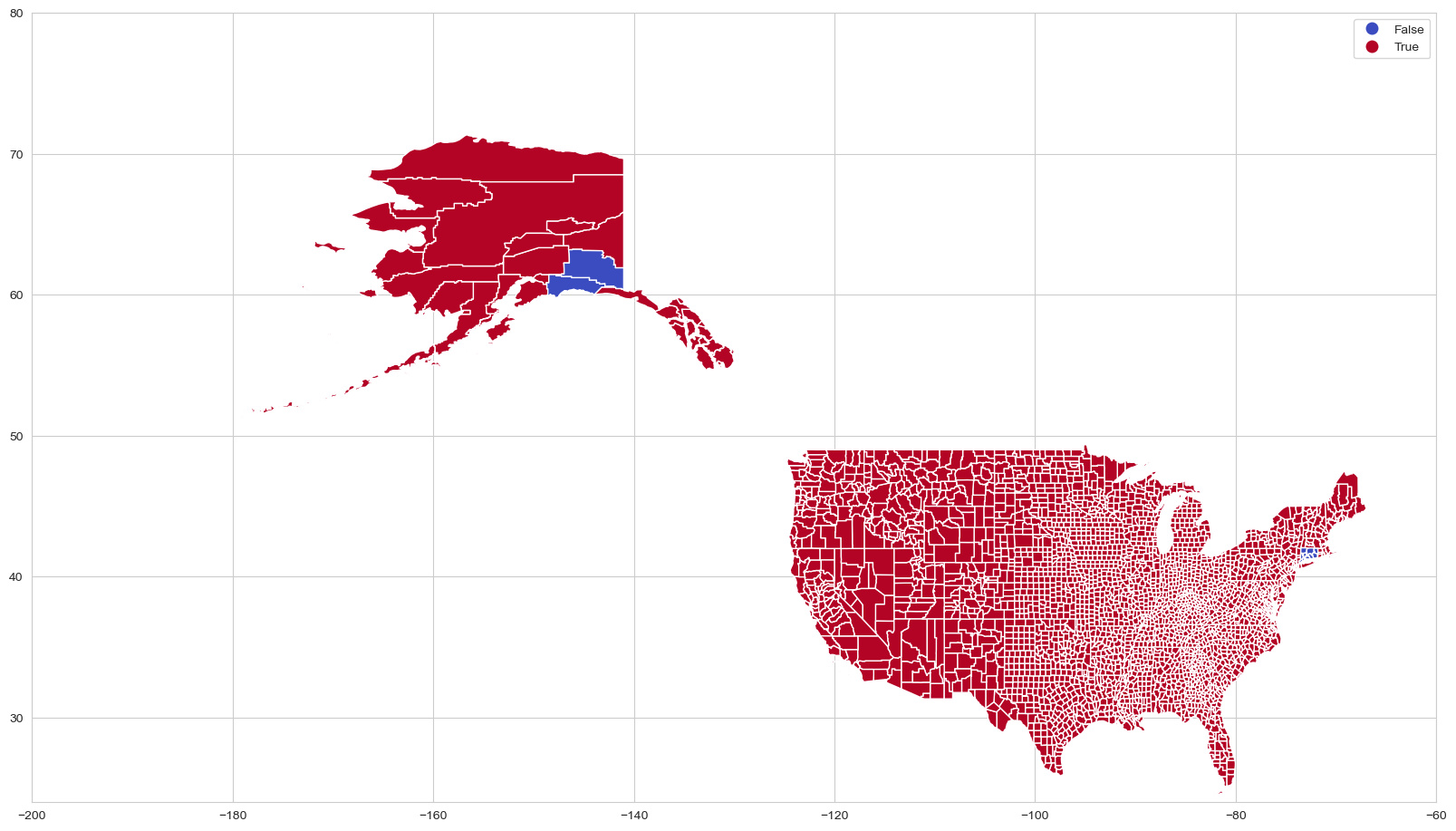
The analysis revealed discrepancies in the case\_id field, where it was found that case\_id uniquely identifies combinations of areaname and county, though some case\_id values were missing. The dataset also includes instances where childcare costs are marked as 0 for families without children. An investigation into the case\_id field revealed that it corresponds to unique combinations of area name and county, with 10 rows per case\_id, indicating a structure that represents different family types within each location.

### 2.7 Dealing with the "10 Null Values"

The missing median\_family\_income values were imputed using the mean income of the corresponding geographic region (state, metro status, and area name). This approach ensured that missing data was handled consistently, and after imputation, the dataset was free of null values in the median\_family\_income column, maintaining data integrity for further analysis.

### 2.8 Data Preparation for Counties and States

To streamline geographic analysis, a new "state\_county" column was created to uniquely identify county-state pairs. The FIPS (Federal Information Processing Standards) reference file was used to match counties with their FIPS codes, revealing some missing values that were addressed. A check confirmed no missing FIPS codes in the dataset. However, some U.S. counties (notably in Alaska, Connecticut, and Puerto Rico) were absent. After deduplication, the dataset covered 96% of U.S. counties, ensuring the completeness of the geographic data for further analysis.



### 2.9 Data Enrichment and Feature Engineering

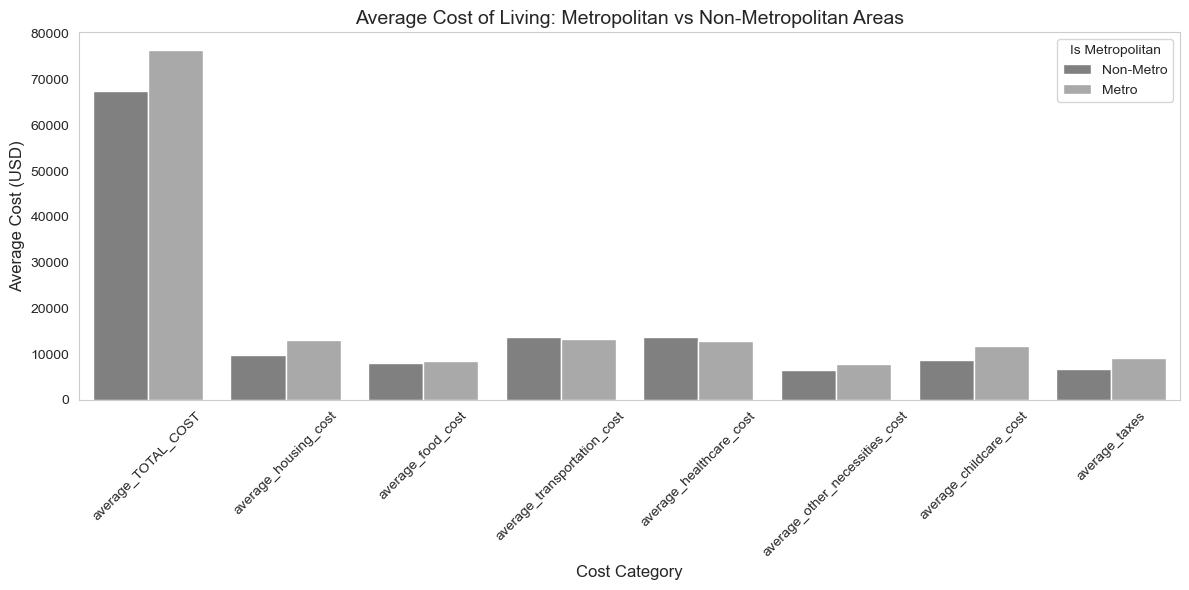
Several new columns were created to enhance the dataset, including the mapping of state codes to full state names, family composition details, and total family size. Percentages for each cost category (housing, food, etc.) relative to total cost were calculated, offering a clearer view of how each expense contributes to a household’s budget. An affordability metric, representing the total cost as a percentage of median family income, was introduced to assess financial strain. After cleaning and validation, the enriched dataset provides more detailed insights, making it easier to analyze affordability across different regions. Refer below table with featured columns

| **Item** | **New Column** | **Type** | **Description** |
| --- | --- | --- | --- |
| 1 | state\_name | string | state name (from state code, e.g. Alabama from AL) |
| 2 | parent\_count | int | # of parents in the household (from family\_member\_count) |
| 3 | children\_count | int | # of children in the household (from family\_member\_count) |
| 4 | family\_size | int | Total #of people in the household (from family\_member\_count) |
| 5 | Housing\_% | float | Percentage of housing cost relative to the total cost |
| 6 | Food\_% | float | Percentage of food cost relative to the total cost |
| 7 | Transportation\_% | float | Percentage of transportation cost relative to the total cost |
| 8 | Healthcare\_% | float | Percentage of healthcare cost relative to the total cost |
| 9 | Other\_% | float | Percentage of other costs relative to the total cost |
| 10 | Childcare\_% | float | Percentage of childcare cost relative to the total cost |
| 11 | Taxes\_% | float | Percentage of taxes relative to the total cost |
| 12 | Affordability metric | float | Total cost as a percentage of the median income |
| 13 | MSA | string | Metropolitan Statistical Area (used for geographical area analysis) |
| 14 | Delta Cost of Living | int | Cost of Living - total\_cost *(Note: dropped since data was correct)* |

## **3. Analysis**

### 3.1 Metropolitan vs. Non-Metropolitan Areas:

Metropolitan areas generally have higher living costs compared to non-metropolitan areas, primarily driven by higher housing, food, childcare, and tax expenses. This is due to factors such as increased demand for housing, limited space, and the concentration of services and goods in urban centers. Our analysis of the dataset's 'isMetro' column showed that the average total cost of living in metropolitan areas is $76,489.60, while in non-metropolitan areas, it is lower at $67,588.02. Although metropolitan areas incur higher costs in most categories, transportation and healthcare costs were lower, likely due to better public transit systems and healthcare infrastructure available in urban regions.

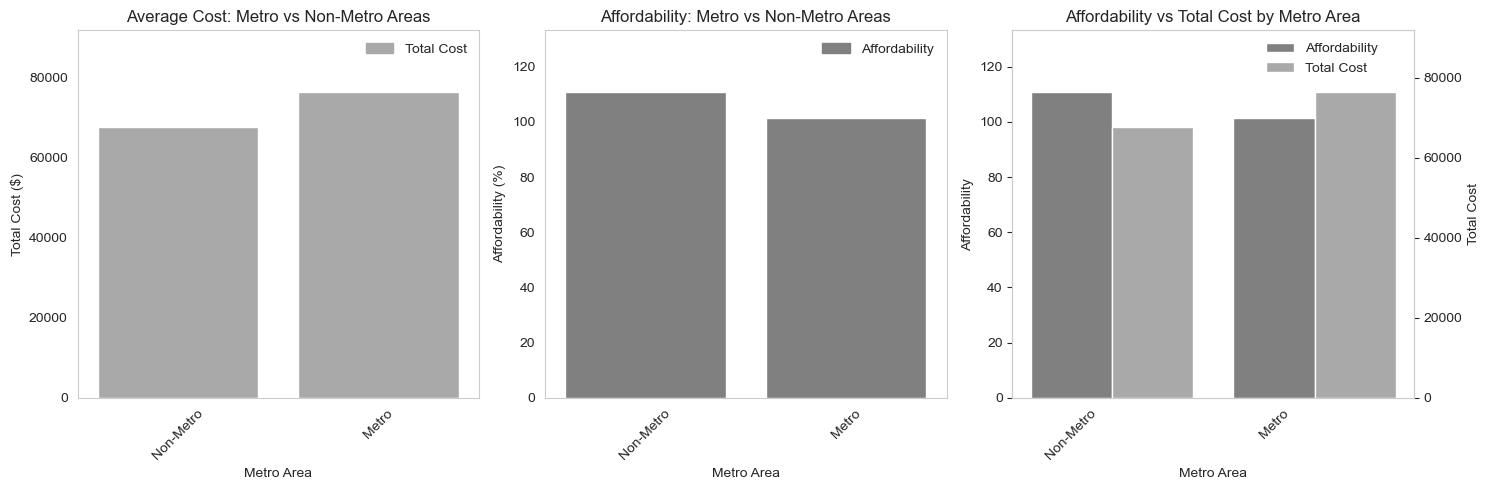
For better visualization, we formatted these cost averages into currency format and compared costs across both metropolitan and non-metropolitan areas. We then plotted these values using a bar chart, which clearly illustrated the differences in costs. 

The bar plot shows that in metropolitan areas, expenses related to housing, food, other necessities, Childcare Costs and taxes tend to be higher, while transportation and healthcare costs are more affordable than in non-metropolitan areas. This insight helps us understand how location influences the distribution of living costs across different household expenses.

The analysis also extended to the affordability metric, providing insights into how the total cost of living compares to the median family income across metropolitan and non-metropolitan areas. By calculating affordability as a percentage of the median family income, we were able to quantify the relative expense of living in these areas. The findings revealed that although the average total cost of living is higher in metropolitan areas ($76,489.60) compared to non-metropolitan areas ($67,588.02), the affordability metric indicates that metropolitan areas are actually more affordable on average. This is because the median family income in metropolitan areas ($77,582.63) is higher than in non-metropolitan areas ($62,826.03), leading to a lower affordability metric (101.44 for metropolitan areas vs. 110.97 for non-metropolitan areas).

This suggests that even though living costs are higher in metropolitan areas, higher incomes may offset these costs, making metropolitan living relatively more affordable. We visualized these insights through several bar charts, comparing both total cost and affordability metrics across metropolitan and non-metropolitan areas. The results showed that while the total cost is higher in metropolitan areas, their higher median income makes them more affordable overall.

For clarity, the visualizations included three graphs: one comparing the average total cost of living, one showing the affordability percentage, and another illustrating the relationship between affordability and total cost for both areas. These findings highlight an important dynamic where higher living costs in metropolitan areas are counterbalanced by higher salaries, providing a more favorable affordability scenario than non-metropolitan areas.



#### Conclusions of Metro vs Non-Metro

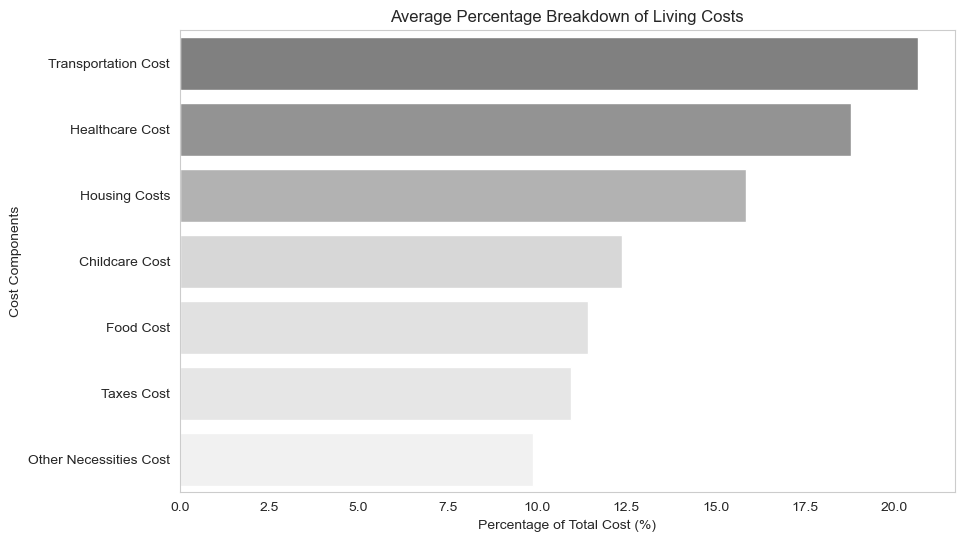
We can observe that the average total cost of living (in absolute terms) is higher in metropolitan areas compared to non-metropolitan areas, If we take into consideration the median family income, via our affordability metric, we can see that the average total cost of living based on average income is actually slightly lower in metropolitan areas! (The affordability metric is lower in metropolitan areas).

*Remembering:*

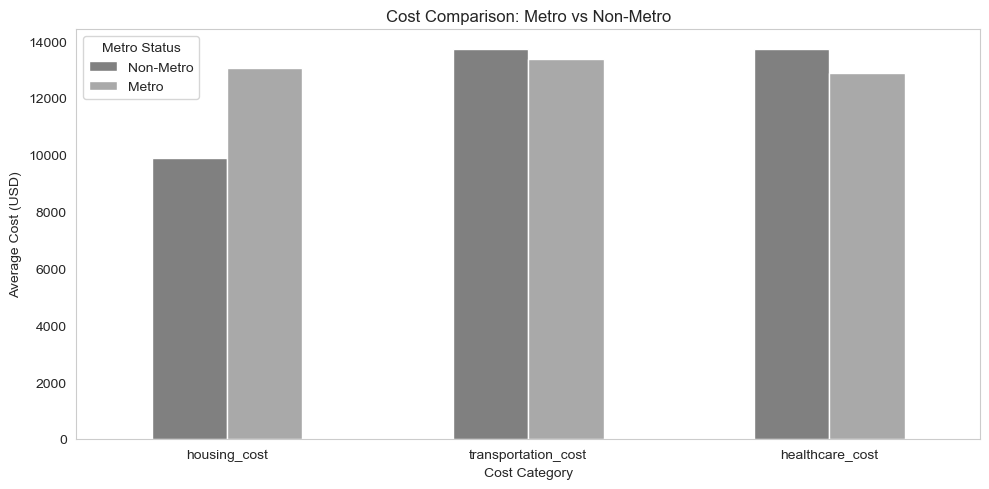
* *Affordability = 100 means that the total cost is equal to the family income.*
* *Affordability > 100 means that the total cost is greater than the family income.*
* *Affordability < 100 means that the total cost is less than the family income.*

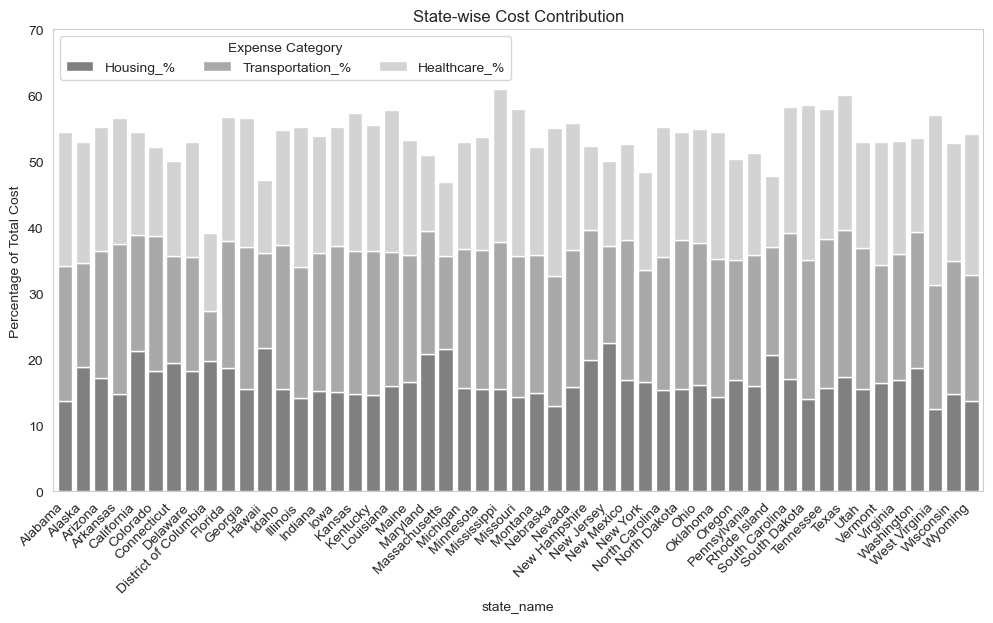
In other words, on average it seems that the cost of living is more affordable in metropolitan areas than in non-metropolitan areas. This is a very interesting finding, and it might be related to the fact that metropolitan areas usually have higher salaries than non-metropolitan areas (and even though they are more expensive, the higher salaries compensate for it)

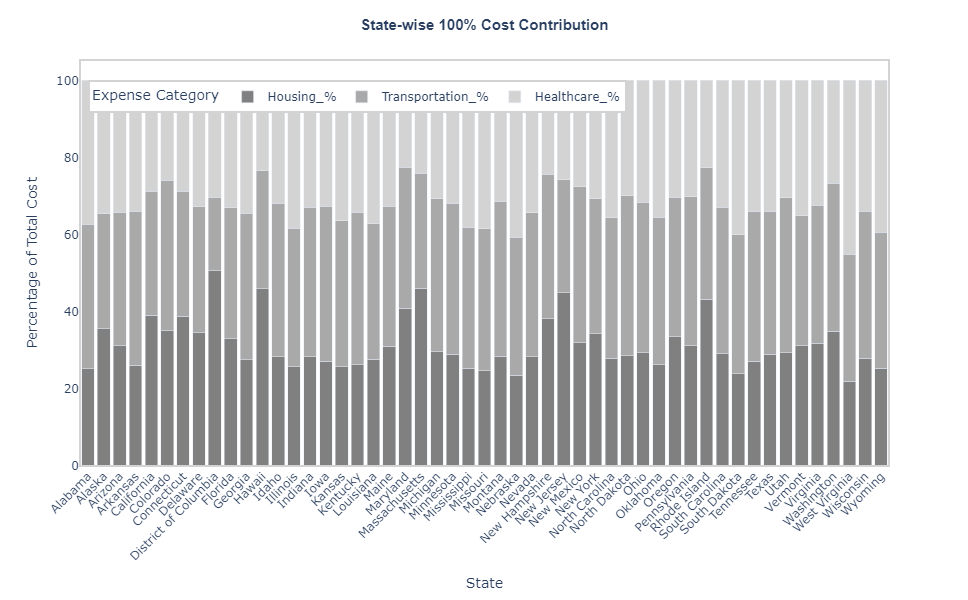
### 3.2 Key Expense Categories:

The hypothesis posits that housing, transportation, and healthcare are the primary contributors to the cost of living, with metropolitan areas typically experiencing higher housing costs due to demand and limited supply, and elevated transportation costs due to less public transit access. Non-metropolitan areas, on the other hand, tend to have lower housing and transportation costs relative to income. The objective is to compare these costs between metropolitan and non-metropolitan regions. The analysis will include bar charts showing the sum of contributors across the U.S., comparisons between metro and non-metro regions, and state-level breakdowns of contributors as a percentage of total cost, ultimately highlighting the key cost differences between these areas.

We calculated and visualized the average percentage breakdown of living costs across various expense categories, such as housing, food, transportation, healthcare, and others. By mapping the components to user-friendly labels, we made the data more accessible. We normalized the values and created a color palette that highlights larger expenses with darker shades. A barplot was generated to show the relative contributions of each cost factor, offering a clear view of how living costs are distributed across different categories. This visualization helps in understanding the proportion of total costs spent on essential living expenses.

In the analysis of cost contributions across metropolitan and non-metropolitan areas, some surprising findings have emerged. **Observation 1**: Contrary to the initial assumption that housing costs would dominate, the data reveals that transportation costs (21% of total costs) are the highest, followed closely by healthcare costs(19% of total costs), with housing costs (16% of total costs) being the lowest among the three. This is a surprising result, as it defies the expectation that housing would be the most significant cost in urban settings, often driven by high demand and limited supply. Instead, transportation appears to be a more pressing concern, especially in regions with limited access to efficient public transit systems.

**Observation 2:** When comparing housing, transportation, and healthcare costs between metropolitan and non-metropolitan areas, the data shows that transportation and healthcare costs are quite similar in both regions, but housing cost is notably higher in metropolitan areas due to higher demand and limited supply. A state-wise comparison using a pivot table reveals the average percentage of total costs spent on housing, transportation, and healthcare in each state. The stacked bar chart visualizes these contributions, demonstrating that while housing costs dominate in many states, transportation and healthcare expenses remain relatively consistent across regions.

We normalized the state-wise cost data to ensure each row adds up to 100%, then created an interactive 100% stacked bar plot to show the proportion of total costs spent on housing, transportation, and healthcare for each state. The plot is color-coded and has a clean, customizable layout, making it easy to compare cost contributions across states.

In this analysis we compared the cost of living of top 3 cost contributors - Transportation , Healthcare and Housing to identify any patterns and there is a clear trade-off across categories—states with low housing often have higher healthcare or transportation expenses, showing that total cost of living varies in how it's distributed across different needs. High housing costs dominate urban and coastal states with dense populations and limited land (e.g., DC, MA, HI). Transportation costs are typically higher in rural or spread-out states where residents depend more on driving. Healthcare costs are highest in states with lower housing costs, possibly reflecting older populations, fewer providers, or poorer public health infrastructure.

**Housing Observations:**

1. The District of Columbia has the highest housing cost percentage at 51%, significantly above all other states. This suggests extremely high urban housing expenses.
2. Northeastern and coastal states like Massachusetts (46%), Hawaii (46%), and New Jersey (45%) also report high housing cost shares, aligning with high property values and rent prices in these areas.
3. Midwestern and Southern states like West Virginia (22%), Nebraska (23%), and South Dakota (24%) have the lowest housing percentages, indicating more affordable housing markets

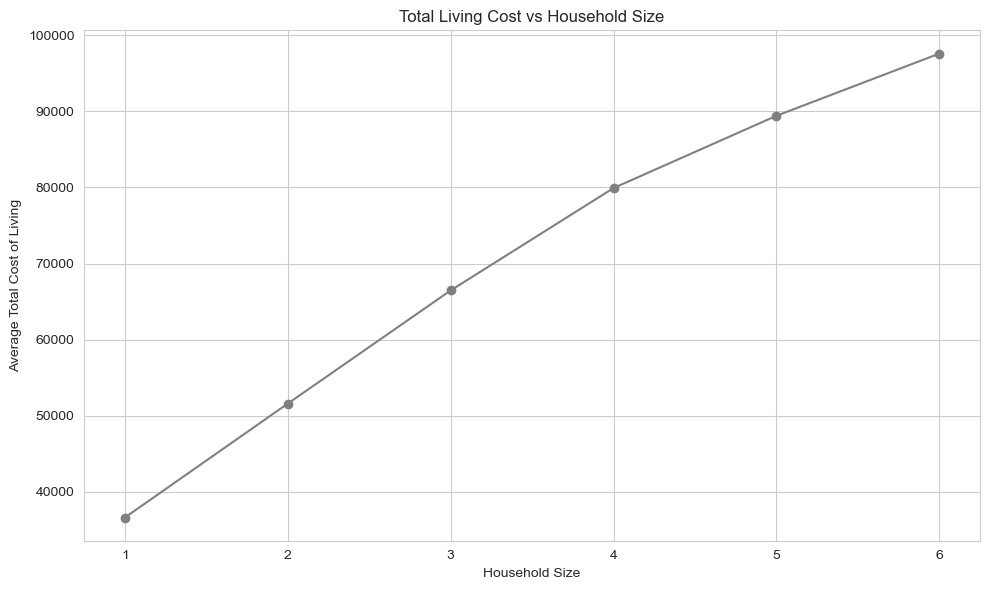
**Transportation Observations:**

1. North Dakota (42%), New Mexico, Utah, Kentucky, Arkansas, and Michigan (40%) all have some of the highest transportation costs, likely reflecting greater reliance on personal vehicles and longer travel distances.
2. District of Columbia (19%) has the lowest transportation percentage, possibly due to dense urban infrastructure and public transit availability reducing reliance on cars.
3. Most states have transportation costs clustered between 35–40%, suggesting it’s a significant cost across the board, especially in states with less urban density.

**Healthcare observations:**

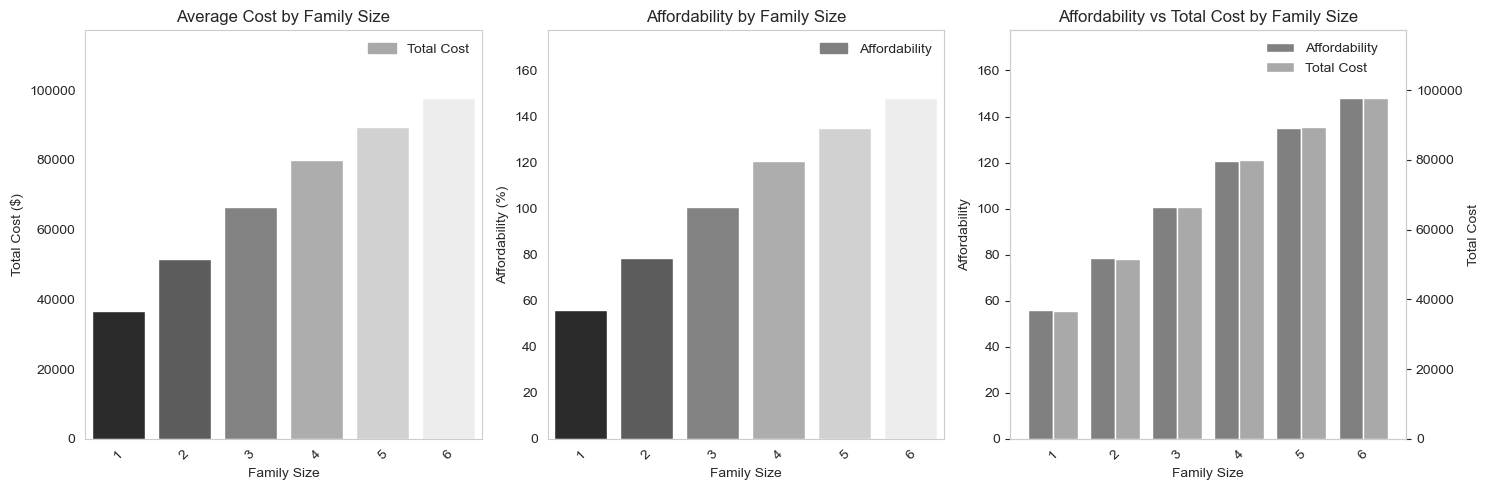
1. West Virginia (45%), Nebraska (41%), and South Dakota (40%) report the highest healthcare cost shares, potentially due to older populations or limited access leading to higher out-of-pocket expenses.
2. Maryland (22%) and Hawaii (23%) are among the lowest, possibly benefiting from more effective public health programs or employer-based coverage.
3. Healthcare percentages are generally higher in lower-housing-cost states, suggesting that where people save on housing, they might spend more on healthcare.

### **3.3 Cost of Living and Household Size**

To explore how total living costs vary by household size, we grouped the dataset by family\_member\_count and calculated the average for each major cost category, including food, transportation, housing, healthcare, childcare, and other necessities. By summing these values, we derived the overall average total cost of living for each household configuration.

The results revealed a clear upward trend: as household size increases, so does the total cost of living. For instance, a single parent with no children (1p0c) has an average cost of just over $31,000, while a two-parent, four-child household (2p4c) faces costs nearing $89,000. These findings were visualized in a line chart, clearly highlighting how expenses scale with family size—an important insight for policy planning, budgeting tools, and understanding financial pressure on larger families.

The analysis examines how household size impacts the total cost of living across key categories like housing, food, transportation, utilities, healthcare, and education. As household size increases, total expenses also rise, though the rate of increase becomes steeper once families reach three or more children. This growth is driven by higher housing costs for larger spaces, as well as increased food and transportation expenses. While utilities and education costs rise more gradually, they still add up. The data highlights the compounding financial challenges faced by larger families, emphasizing the need for strategic financial planning and targeted policies to support them. To assess affordability, the analysis compares household costs to median income, with affordability metrics indicating when costs exceed income. By grouping data by family size and visualizing average costs, affordability, and their relationship, the study provides valuable insights into the financial strain on larger families, helping to inform policy decisions and financial support programs.



In conclusion, as family size increases, the overall cost of living also rises—often at an accelerating rate. While some expenses, like food and utilities, scale gradually, major costs such as housing, transportation, and healthcare increase more significantly with each additional family member. This pattern underscores the growing financial burden placed on larger households and highlights the importance of thoughtful financial planning and supportive policy measures to address the unique challenges they face.

### 3.4 Exploring Linearity (or Non-Linearity) of Costs as a Function of Household Size

#### Overview: Exploring Linearity of Costs Relative to Household Size

This section investigates whether different categories of household expenses demonstrate a linear relationship with household size. The analysis covers eight cost variables: seven subcategories (e.g., housing, food, transportation) and the overall total cost. Household size is tested in two ways: as a single variable (family\_size) and as two separate variables (parent\_count and children\_count). Linear regression is applied across various segments of the dataset, including the full dataset, metro versus non-metro areas, and by state. The performance of each model is evaluated using R² scores.

#### Key Results from the Entire Dataset

The results for the full dataset show that **food costs** have the highest linearity with household size, with an R² score above 0.90 regardless of whether one or two variables are used. This indicates a strong and consistent relationship between the number of family members and the amount spent on food. **Total cost** and **healthcare cost** also display reasonably strong linear trends, particularly when both parent and child counts are considered separately. In contrast, **transportation, childcare, and other necessities** exhibit moderate linearity, with R² scores improving slightly when the independent variable is split into two components. However, **housing costs and taxes** show weak linear relationships with household size. Their low R² values suggest that other complex factors, such as geography, income level, or housing policy, may play a more significant role in determining these expenses.

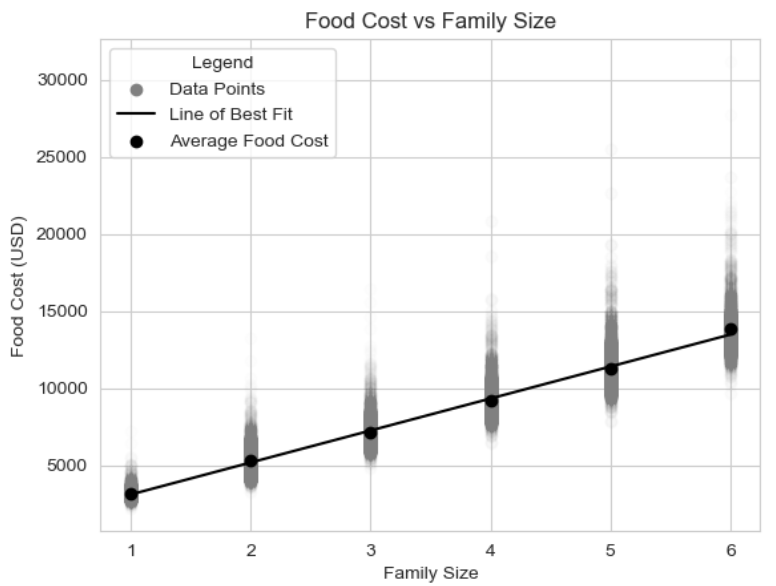
#### Metro vs. Non-Metro Comparisons

When the data is divided into metro and non-metro segments, notable differences emerge. For most categories, **non-metro areas yield higher R² scores** than metro areas, indicating that household size is a more reliable predictor of cost in rural or less urbanized regions. Specifically, the R² score for **housing in non-metro areas** is substantially higher (around 0.45–0.50) than in metro areas (less than 0.25), where cost variability appears greater. Similarly, the total cost of living is more predictable in non-metro areas, reflecting more uniform cost structures in these regions.

#### State-Level Analysis

Stratifying the dataset by state reveals even greater improvements in model performance. For example, when cost models are run separately for each state, **R² scores for housing and taxes increase significantly**, suggesting that localized cost differences are critical to accurately modeling these expenses. The average R² score for housing costs across states jumps from 0.22 (in the national model) to over 0.56, and similar trends are observed for taxes and other categories. These findings imply that **cost structures vary meaningfully by state**, and a one-size-fits-all model underrepresents these variations.

#### Metro vs. Non-Metro Within States

A more granular breakdown—analyzing metro and non-metro areas within each state—leads to further incremental improvements in model accuracy. In this configuration, **transportation, childcare, healthcare, and housing** all show better R² scores when both independent variables are used. However, food, total cost, and other necessities appear to benefit less from this increased model complexity. These results suggest that while breaking out household size into parent and child counts can enhance accuracy, the impact varies by expense type.

#### Visualizing Linearity: Food Cost vs. Family Size

To provide a more intuitive understanding, a scatter plot of **food cost vs. family size** is used to visualize the linear trend. A regression line is fitted to the data, and the average food cost for each family size is plotted to show alignment with the model's predictions. The regression achieves an R² of approximately 0.90 and a low RMSE of around $1,011. This indicates that food costs not only follow a clear linear pattern but also cluster closely around the predicted values, making it one of the most consistent and reliable cost categories in the dataset.

#### 

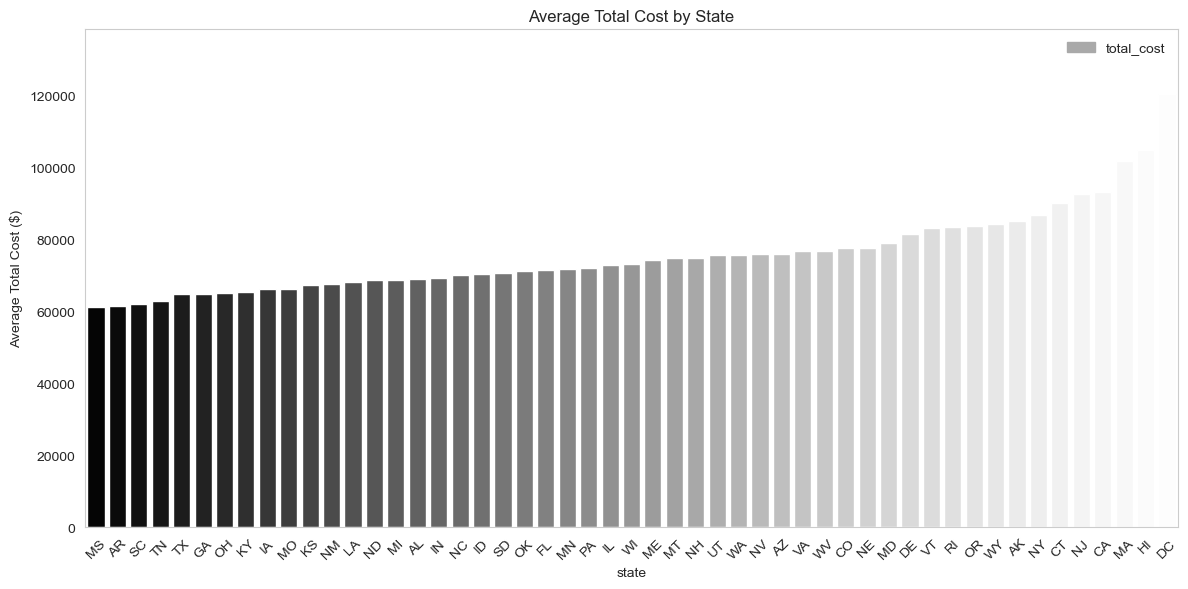
#### Conclusions and Insights

The analysis concludes that **some costs scale linearly with household size**, while others do not. **Food, healthcare, and total cost** demonstrate strong linear relationships, while **housing and taxes** appear to follow more complex, non-linear patterns. Breaking household size into **parent and child components** offers marginal improvements in many cases, particularly for categories like healthcare and childcare. Additionally, segmenting the data by **location—especially by state and metro status—improves the model's accuracy**, highlighting the regional variability in cost structures. This analysis underscores the importance of **context and household composition** when assessing cost-of-living dynamics and offers valuable insights for budgeting, policymaking, and regional economic planning.

### 3.5 Exploring Regional Variation:

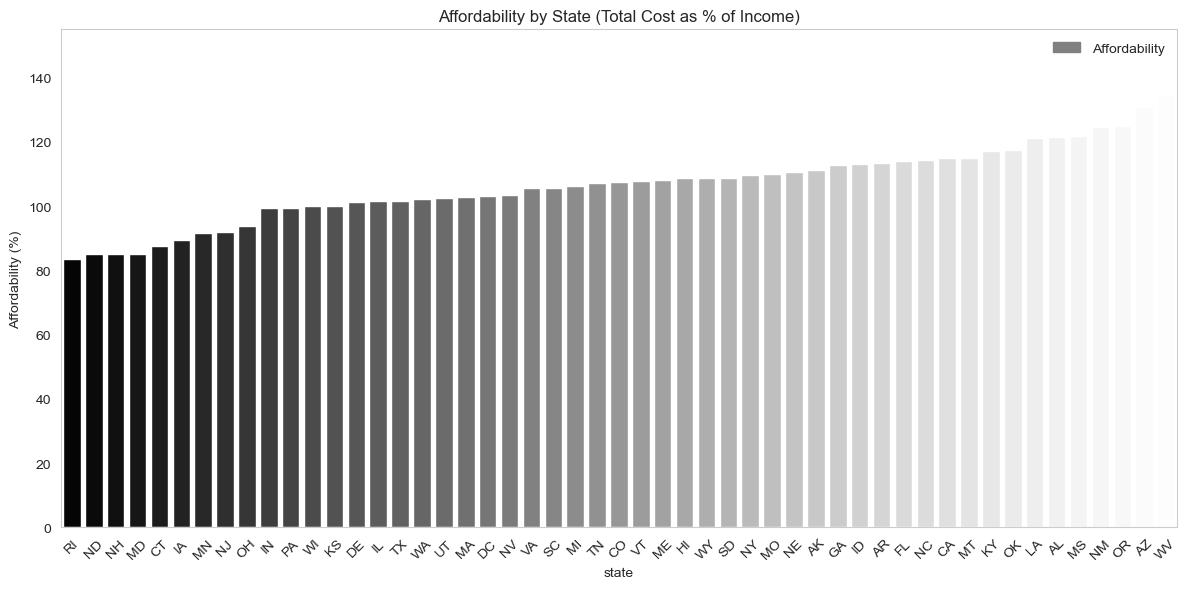
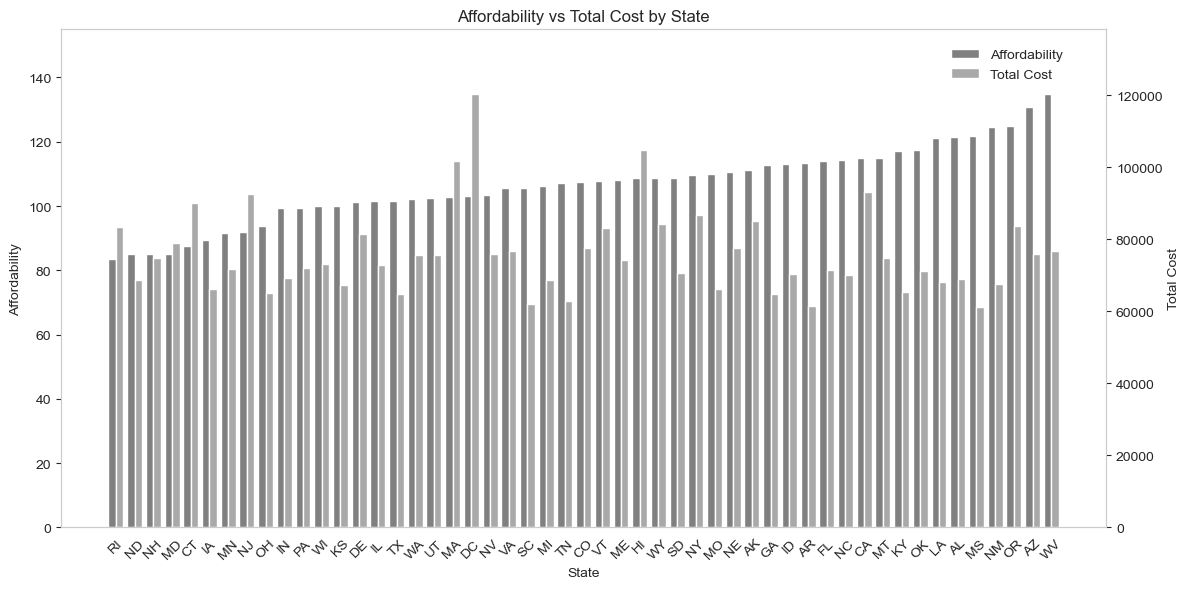
We can extend our analysis by looking at regional differences. A regional analysis could be done by looking at certain groupings of the data based for instance on:

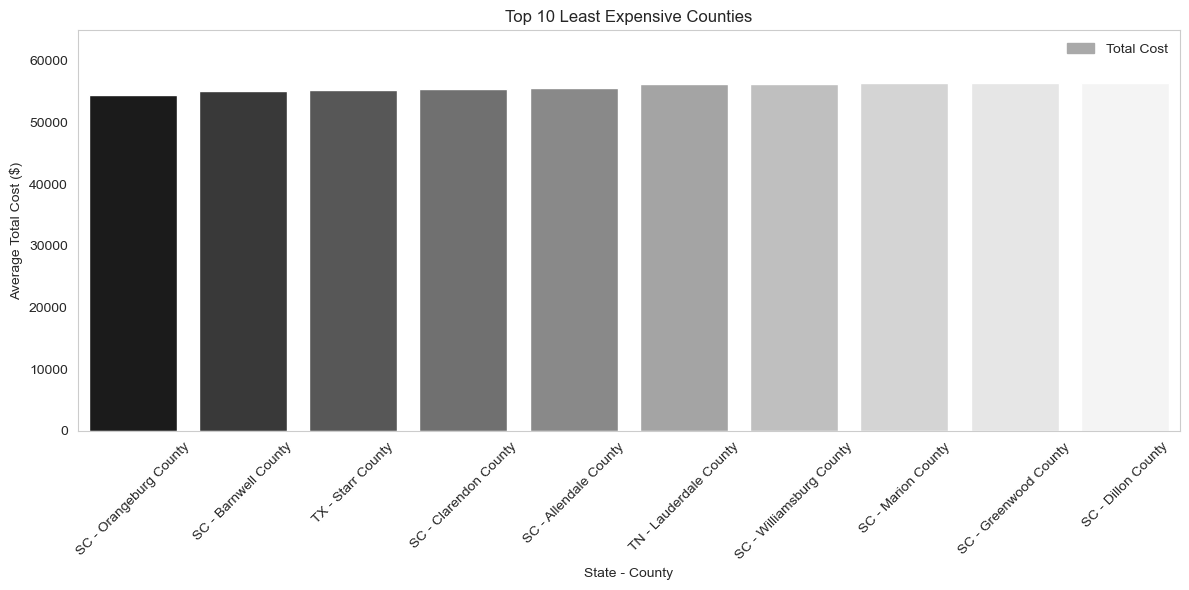
* Metropolitan vs Non-Metropolitan areas (already done in section 2.3)
* State-level Analysis
* County-level Analysis
* Metropolitan Statistical Areas (MSA) Analysis

While Section 2.3 covered disparities between metropolitan and non-metropolitan areas, this section focuses on more granular geographic levels: state, county, and Metropolitan Statistical Areas (MSAs). In each of the following sections, The results were sorted and displayed in two ways—by average total cost and by affordability, also, a dual-axis chart compares the affordability and the cost side-by-side. To improve readability, monetary values were formatted in standard currency format.

#### 3.5.1 State-Level Analysis:

*Visualizing the Affordability vs Total Cost by States:*

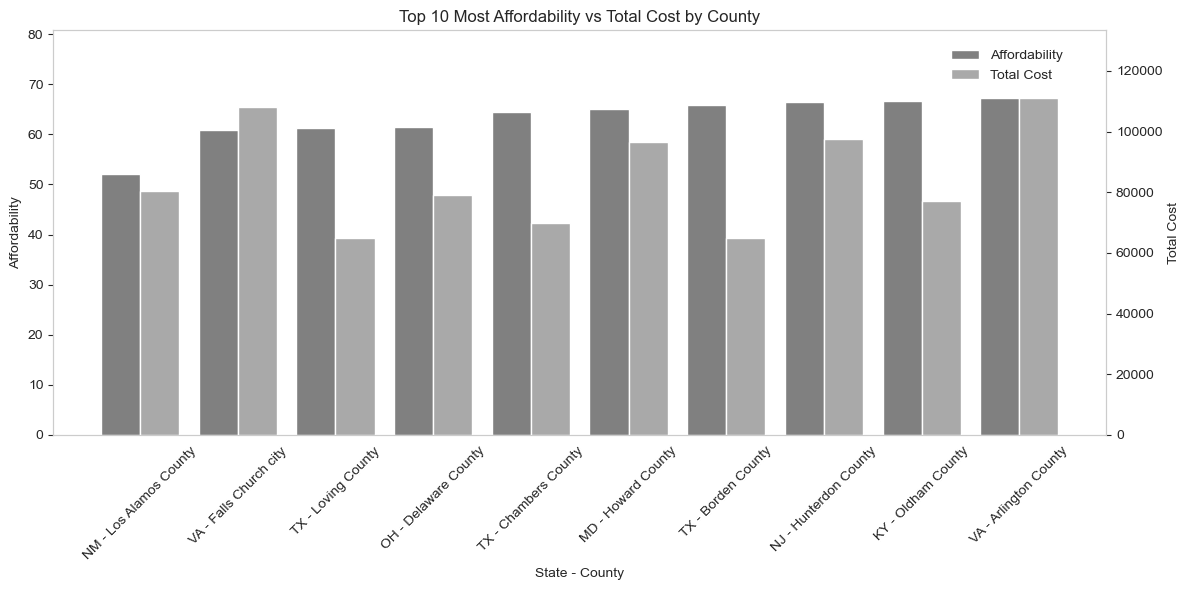
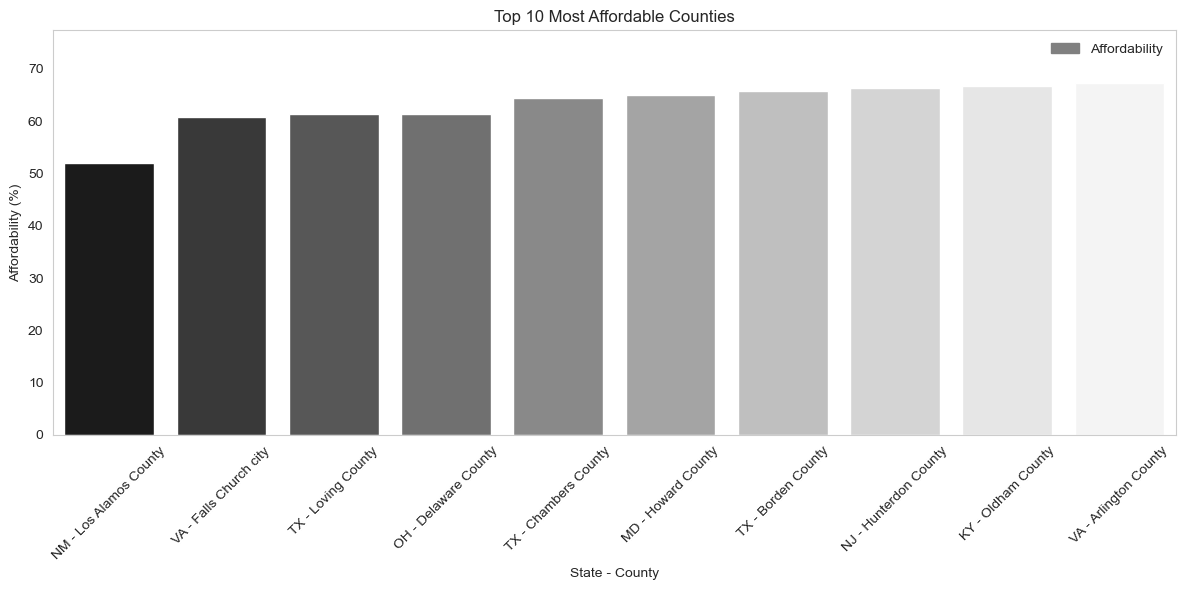
We begin with a state-level analysis, grouping the data by state and calculating the average total cost, median family income, and affordability for each. The results highlight which states present the highest and lowest cost burdens relative to income where households face the greatest financial strain. This sets the stage for more detailed geographic analysis in the following sections.



##### 

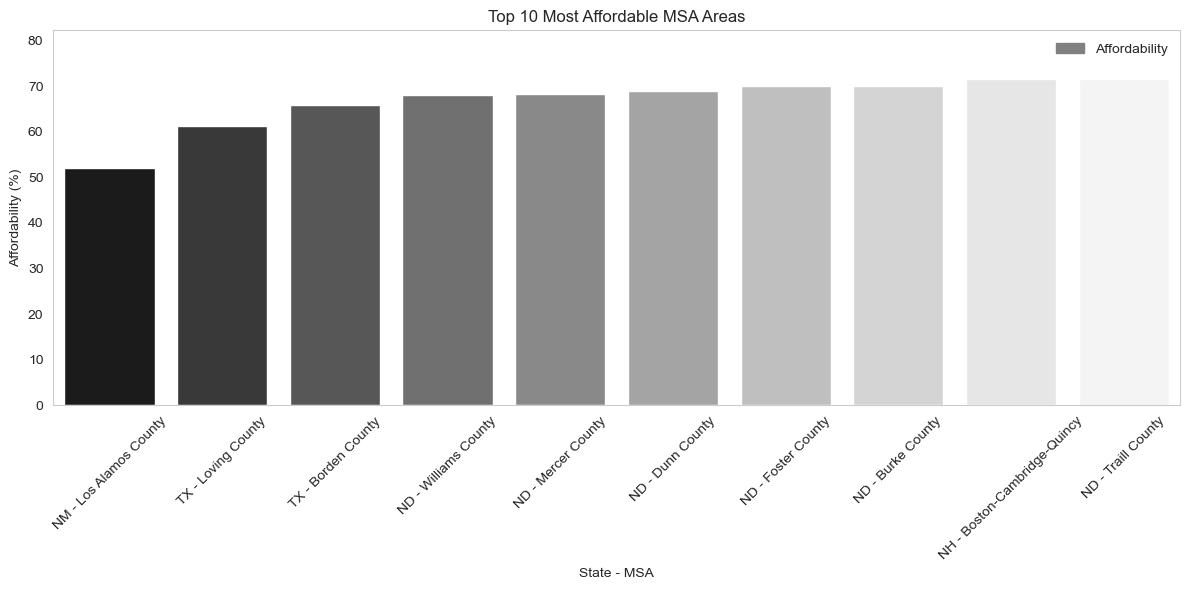
#### 

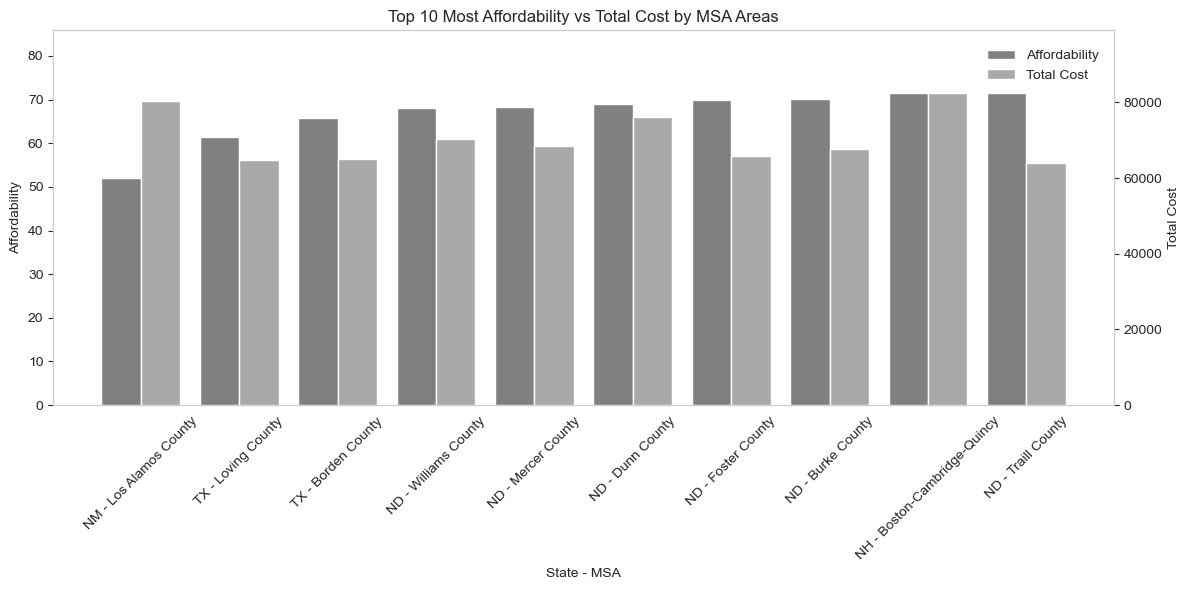
#### 3.5.2 County-Level Analysis

*Visualizing the Top 10 Most Affordable Counties vs Total Cost by County as an example. (In our notebook code, you will find more details on the top 10 least affordable Counties vs Total Cost by County as well)*

Next, we conducted a **county-level analysis** to identify localized affordability patterns. By grouping the data by both state and county, we calculated the average total cost, median family income, and affordability for each county. We then highlighted the top 10 **most expensive** and **least expensive** counties based on total cost, as well as the **least** and **most affordable** counties based on affordability metrics. Additionally, we created a **dual-axis plot** to compare both affordability and average total cost side-by-side. To enhance clarity, we combined state-county names into a single label. This analysis helps pinpoint areas where families may be under the most financial pressure or enjoying relatively affordable living conditions.

#### 3.5.3 Metropolitan Statistical Areas (MSA) Analysis

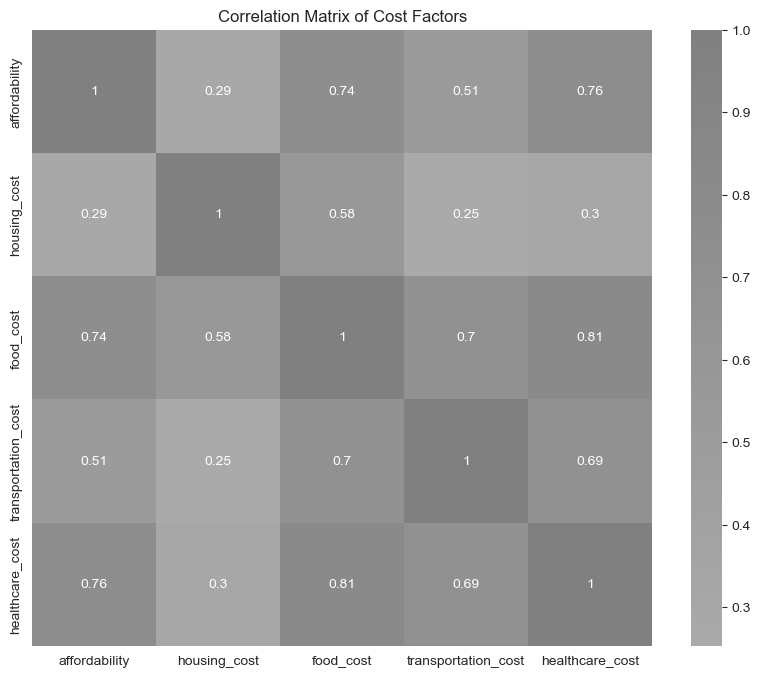
*Now we visualize the top 10 most affordable MSAs and their total costs as an example. (In our notebook code, you will find more details on the top 10 least affordable MSAs and their total cost as well)*

In this section, we focus on analyzing the data at the MSA level, where the Metropolitan Statistical Areas (MSAs) are used to assess affordability, total cost, and median family income. After grouping the data by both state and MSA, we identified the top 10 **most expensive MSAs**, **least affordable MSAs**, **least expensive MSAs**, **most affordable MSAs** and a **dual-axis plot** to compare both affordability and average total cost side-by-side. By examining the affordability metrics (total cost as a percentage of income) and total costs in different MSAs, we gain insights into the economic realities faced by families in urban areas across the country.

#### Key Findings and Insights about Regional Variations:

1. State Variations: The data shows significant differences in affordability between states. Some states have costs that represent a much higher percentage of median income than others. The dataset shows significant variation in living costs across states.
2. County Variations: Even within states, there can be substantial variation at the county level, with some counties being significantly more expensive than others in the same state.
3. Geographic Patterns: Certain metropolitan statistical areas (MSAs) stand out as particularly affordable or unaffordable, often reflecting regional economic conditions.
4. Affordability: Some areas with lower absolute costs may actually be less affordable when considering local incomes, while some high-cost areas remain relatively affordable due to higher incomes.

### 3.6 Individual Expense Categories

The analysis breaks down living costs into categories like housing, food, and transportation, using user-friendly labels for clarity. The average percentage for each cost component is calculated and sorted, with a color palette highlighting larger expenses in darker shades. A barplot is created to visually compare these categories, making it easy to understand the distribution of living costs. The grid is removed for a clean presentation, emphasizing key insights into expense allocation.

Then we use the correlation matrix to identify relationships between cost factors. For example, food and healthcare costs have a strong correlation, suggesting they often rise together. Affordability is moderately linked to housing costs, but food and healthcare play a bigger role. The heatmap visually highlights these connections, helping to better understand how different costs influence overall affordability.

In conclusion, our regional analysis revealed that affordability and cost of living vary significantly across states, counties, and Metropolitan Statistical Areas (MSAs). State-level results highlighted stark contrasts in both average total costs and how those costs relate to local incomes. County level insights showed even more localized disparities, with some counties offering affordable living while others impose substantial financial burdens on families. MSA-level analysis further emphasized how urban areas can differ dramatically in affordability—some offering relatively balanced costs and incomes, while others reflect high expenses that outpace local earnings. Through dual-axis charts and side-by-side comparisons, we clearly illustrated where affordability gaps are most pronounced. This comprehensive regional breakdown underscores the importance of tailoring economic and policy responses to specific geographic needs, recognizing that affordability challenges are not uniformly distributed but heavily influenced by local conditions.

## **4. Overall Conclusions**

The findings of this study strongly support the initial hypothesis that affordability across regions is closely tied to the relationship between the total cost of living and median family income. While **metropolitan areas** exhibit higher absolute costs—particularly for **housing and goods**—they also tend to offer **higher incomes**, resulting in **better average affordability metrics** when compared to non-metropolitan areas. This indicates that **urban living, while expensive, may be more financially sustainable** due to higher earning potential, challenging the common assumption that rural areas are always more affordable.

Additionally, the **impact of household size** on cost of living is **non-linear**. For smaller families, costs increase moderately, but for larger households—especially those with three or more children—costs rise at an accelerating rate, particularly in housing, transportation, and healthcare. This confirms that **some expenses scale super-linearly**, placing a disproportionate financial burden on larger families. On the other hand, categories like food and utilities show more **sub-linear behavior**, benefiting from shared usage among household members.

The analysis also revealed that the **scaling behavior of costs varies by region**. For instance, **housing costs scale more steeply with family size in metropolitan areas**, while **transportation costs are more variable in rural regions**. Segmenting costs by region and household structure improved our understanding of these patterns, highlighting that **a one-size-fits-all approach to affordability is ineffective**.

Finally, breaking costs down by category showed that **housing, transportation, and healthcare are the dominant drivers** of total living expenses, but their impact varies by location and family type. The hypothesis that affordability is shaped by the interaction between **family structure, household size, and geographic location** was validated. These findings underscore the importance of **location-specific, data-driven policy interventions** and financial planning strategies that account for the complex, non-linear dynamics of cost-of-living pressures across America.

**5. References**

Asaniczka. (2023). \*US cost of living dataset - 3,171 counties\* [Data set]. Kaggle. <https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties>

Qayoom, Sethi, Ramalheiro, Morales, Hao**.** (2025). **Data science group17 (Version latest) [Source code]. GitHub.** [**https://github.com/miguelmog10/data-science-group17**](https://github.com/miguelmog10/data-science-group17)

University of Waterloo. (2025). **Foundations of Data Science** **(Course No. 0048-13, Winter 2025)** [University course].