**University of Waterloo**

**Report on Group Project:**

**Hypothesis Testing and Statistical Inference**

**The Cost-of-Living Differences Between Metro and Non-Metro Counties across the United States**

**Course:**  
Statistics for Data Science - 0049-14 - Summer 2025  
*Sergiy Nokhrin & Amanda Easson*

**Group 4:**Jack Hao, Duc Vu Hoang, Krishna Shah, Joginder Singh

**Date:**August 11, 2025

**GitHub Repository (public):**<https://github.com/howexg9/Statistics-for-Data-Science-Group4>

**1. Introduction**

This report investigates the cost of living differences between metro and non-metro counties across the United States. The project includes hypothesis testing to assess whether significant cost disparities exist, followed by predictive modeling using regression and classification techniques to estimate total cost of living based on various cost components.

**Objective**

* Determine whether there is a statistically significant difference in total cost of living between metro and non-metro counties.
* Build predictive models to estimate total cost of living using key cost components such as housing, food, transportation, healthcare, childcare, taxes, and others.

**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.

**3. Part 1: Hypothesis Testing and Statistical Inference**

**3.1 Hypothesis Statement** - H₀: There is no significant difference in total cost of living between metro and non-metro counties. - H₁: There is a significant difference in total cost of living between metro and non-metro counties.

**3.2 Data Visualization and Distribution Check** - Histograms and Q-Q plots were used to visualize the distribution of total costs for metro and non-metro groups. - Boxplots highlighted potential outliers in metro areas.

**3.3 Normality and Transformation** - The Shapiro-Wilk test and Q-Q plots suggested non-normality. - Box-Cox transformation was applied to improve normality.

**3.4 Outlier Detection and Removal** - The IQR method identified extreme values. Comparisons were re-run after removing outliers.

**3.5 Welch’s T-Test** - Welch’s t-test showed a statistically significant difference in means (p < 0.001), both with and without Box-Cox transformation and outlier removal.

**3.6 Cohen’s d Effect Size** | Comparison | Cohen’s d | Interpretation | |————|———–|—————-| | Original | 0.40 | Small to medium | | No Outliers | 0.38 | Small to medium | | Box-Cox | -6.16 | Not interpretable (distorted scale) | | Box-Cox No Outliers | -3.16 | Not interpretable (distorted scale) |

**4. Part 2: Predictive Modeling**

**4.1 Box-Cox Transformed Target Variable** - The target variable (total\_cost) was transformed using Box-Cox to improve linear model assumptions.

**4.2 OLS Model Fitting and Assumption Checks** - Residuals were checked for normality, homoscedasticity, and independence. - Q-Q plots, residual vs. fitted, and histogram plots confirmed assumptions were reasonably met.

**4.3 Multicollinearity and Correlation Analysis** - Correlation matrix showed strong linear relationships between some predictors. - VIF analysis revealed multicollinearity (VIF > 5), especially in housing and taxes.

**4.4 Ridge Regression to Handle Multicollinearity** - RidgeCV with cross-validation selected the best alpha (~1.59). - Ridge provided more stable coefficients than OLS.

**4.5 Performance Comparison: OLS vs Ridge** | Model | RMSE | R² | Best Alpha | |——-|——|—–|————-| | OLS | 0.0866 | 0.9925 | - | | Ridge | 0.0866 | 0.9925 | 1.59 |

**4.6 Feature Importance** - Standardized Ridge coefficients indicated housing\_cost, taxes, and childcare\_cost were the most influential predictors of total\_cost.

**4.7 Model Cross-Validation** - K-Fold cross-validation confirmed that Ridge generalizes slightly better than OLS due to regularization.

**4.8 Classification Using Naïve Bayes** - Total cost was binned into tiers (Low, Medium, High). - Naïve Bayes classification achieved reasonable accuracy with highest influence from housing\_cost and childcare\_cost.

**5. Conclusion**

* There is a statistically significant and practically meaningful difference in cost of living between metro and non-metro counties.
* Ridge regression performed comparably to OLS in predictive accuracy, while offering better generalizability and coefficient stability.
* Key predictors of total cost are housing, taxes, and childcare.
* The model can be extended to classification tasks, such as tier prediction.

**6. 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>

Jack (Xianguo) Hao, Duc Vu Hoang, Krishna Shah, Joginder Singh**.** (2025). **Statistics for Data Science Group 4 (Version latest) [Source code]. GitHub.** <https://github.com/howexg9/Statistics-for-Data-Science-Group4>

University of Waterloo. (2025). **Statistics for Data Science (Course No. 0049-14, Summer 2025)** [University course].