Lifestyle Comparison between Colorado and Utah

Lingyi Wang
lingyi.wang@colorado.edu
MS in Data Science, University of Colorado Boulder
Boulder, Colorado, USA

Riley Jones
riley.jones@colorado.edu
MS in Data Science, University of Colorado Boulder
Boulder, Colorado, USA

Taihei Sone
taihei.sone@colorado.edu
MS in Data Science, University of Colorado Boulder
Boulder, Colorado, USA

Abstract

The mission of this project is to provide a comprehensive, data-driven comparison of lifestyles in Colorado and Utah, helping individuals, families, and businesses make informed decisions about relocation, investment, and quality of life. By analyzing economic conditions, cost of living, job markets, demographics, and recreational opportunities, this research aims to deliver clear, actionable insights in an accessible format.

Keywords

Lifestyle, Economy, Cost of Living Migration, Environment, Tax

ACM Reference Format:
Lingyi Wang, Riley Jones, and Taihei Sone. 2018. Lifestyle Comparison between Colorado and Utah. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As individuals and families across the United States increasingly consider relocating in search of better quality of life, understanding the regional differences in lifestyle factors becomes critical. Colorado and Utah, both located in the Mountain West, offer unique combinations of economic opportunity, environmental amenities, and demographic characteristics. However, while these two states share geographic proximity and cultural similarities, their distinct policies, infrastructure, and social dynamics suggest potentially meaningful differences in lifestyle experience.

This project presents a comprehensive, data-driven comparison between Colorado and Utah to help stakeholders—including prospective residents, policymakers, and businesses—make informed decisions. By leveraging publicly available data sources such as the U.S. Census Bureau, the Environmental Protection Agency, the Centers for Medicare & Medicaid Services, and Zillow, this study examines key lifestyle components across five domains: (1) economic and employment factors, (2) cost of living and housing, (3) demographics and migration trends, (4) quality of life and environment, and (5) state policies and taxation.

Our methodology combines statistical summaries with machine learning models, including clustering, classification, regression, and association rule mining, to uncover patterns and actionable insights. The goal is not only to describe current conditions but also to identify structural differences that might influence the lived experience of residents in each state.

By quantifying lifestyle indicators through reproducible analytics and visualizations, we aim to support practical decision-making and contribute to ongoing discussions about regional mobility, resource distribution, and equitable development.

2 Related Work

Prior studies on regional lifestyle comparisons have often focused separately on economic, demographic, and environmental factors, providing essential yet fragmented insights into regional disparities. For example, Florida (2005) highlighted the importance of socioeconomic factors in attracting creative and skilled populations, emphasizing regional differences in employment opportunities and quality of life. Similarly, Glaeser, Kolko, and Saiz (2001) explored how consumer amenities and housing affordability influence urban growth and resident satisfaction, emphasizing the interplay between demographic trends and economic conditions.

Recent methodological advancements have encouraged researchers to utilize machine learning techniques to uncover more nuanced regional differences. Various researchers have applied clustering methods to regional data to identify economic typologies characterized by varying labor market conditions and housing costs, underscoring the importance of integrated economic and demographic analysis to understand internal migration patterns. Additionally, decision tree and random forest models have been demonstrated to effectively predict healthcare access at the state level, showcasing how advanced analytics can enhance policy decisions related to public health.

While existing literature provides valuable insights, analyses specifically comparing lifestyle characteristics between Colorado and Utah remain limited. Commonly referenced resources, such as Zillow's market trend reports and World Population Review's state demographic summaries, offer useful but superficial overviews, often lacking the analytical depth necessary for comprehensive decision-making. Furthermore, environmental assessments provided by agencies like the U.S. Environmental Protection Agency frequently focus narrowly on pollution metrics without integrating them into broader lifestyle evaluations.

This research addresses these gaps by integrating economic, demographic, environmental, and policy-related data through a comprehensive analytical framework. By employing a combination of clustering, regression, classification, and association rule mining, we aim to provide actionable insights beyond those available from isolated data sources or conventional descriptive studies.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Our approach contributes to a richer understanding of how integrated lifestyle indicators differ between Colorado and Utah, thus supporting more informed relocation and policy decisions.

3 Data Sources

In this section, data acquisition is described in each subsection.

3.1 Economic and Employment Factors

- U.S. Census Bureau API - Economic and Employment Data

3.2 Cost of Living and Housing

- U.S. Census Bureau API - Economic and Employment Data

3.3 Demographics and Migration Trends

- Initial list of zip codes gathered from World Population Review:
  - https://worldpopulationreview.com/zips/colorado
  - https://worldpopulationreview.com/zips/utah
- U.S. Census Bureau API – Basic Demographics:
  - https://www.census.gov/data/developers/data-sets.html
- U.S. Census Bureau API – Demographic and Migration Data:
  - https://www.census.gov/data/developers/data-sets.html
- Averaged Real Estate Data from the Zillow API:
  - https://www.zillowgroup.com/developers/

3.4 Quality of Life and Environment

- Centers for Medicare & Medicaid Services Data API – Hospital Data
- U.S. Census Bureau API – Health Insurance Coverage Data
- U.S. Environmental Protection Agency API – Air Pollution Data
- NPS Data API – National Park Data

3.5 State Policies and Taxation

- Scraping Tax Foundation – Tax Data

4 Data Cleaning and Preprocessing

In this section, data preprocessing is described in each subsection.

4.1 Economic and Employment Factors

4.1.1 Economic and Employment General Data. Census API is used to collect economic and employment data. The data rows are specific to each ZIP code. The cleaning process involved deleting the Name column due to redundancy as there is already a zip code tabulation area column; deleting unemployed and below poverty since employment rate and poverty rate are more informative data points; deleting total poverty universe as well, since it is primarily used to calculate poverty rate. -666666666 indicates missing data, and replace all negative values with the median value of the column. Fill all missing data with median of column Simple visualizations are implemented; details can be found on the project website, and the source code can be found on GitHub.

4.2 Cost of Living and Housing

4.2.1 Cost of Housing Data. Census API is used to collect housing cost data. The data rows are specific to each ZIP code. Cleaning process is the same to the economic and employment data as well.

4.3 Demographics and Migration Trends

Handling Issues and Noise: Demographic, migration, and real estate data were aggregated using two separate Census API scripts and merged with Zillow API data. Data for Colorado and Utah were concatenated by ZIP code. Rows with negative values in non-negative categories were dropped—these typically represented sparsely populated areas and were unlikely to affect overall trends. Imputation was avoided to prevent introducing bias. Since the focus was on densely populated ZIP codes, removing these rows was appropriate. Negative values likely stemmed from incomplete Zillow data.

Understanding the Data: Most features are numerical, with categorical variables like State, County, City, and ZIP code. Zillow-provided features (e.g., home values, rents) are based on a maximum of 41 listings per ZIP code, due to API limits. As such, they represent a limited snapshot and should be interpreted as approximations.

Normalization was deemed unnecessary at this stage. However, logarithmic scaling was applied in visualizations to handle features with large scale differences.

Basic Statistical Analysis

The dataset covers demographic, migration, and real estate variables for ZIP codes in Colorado and Utah. Summary statistics (means, medians, variances, standard deviations) were calculated to examine central tendencies and dispersion. For instance:

- Median home value: $332,143
- Median household income: $73,632

Significant variability across ZIP codes highlights socioeconomic diversity. Some features exhibited skewness, suggesting the presence of outliers or contrasts between rural and urban areas.

Correlations Between Features

Several correlations were identified:

- Median household income positively correlates with home values and gross rent.
- Population density and migration inflows correlate with real estate prices and availability.

These trends suggest that wealthier and more densely populated ZIP codes tend to have higher housing costs.

Data Integration and Similarity Analysis

Data from the Census and Zillow APIs were successfully merged using ZIP code as the primary key. Integration was consistent across Colorado and Utah, with ZIP code providing a reliable linkage mechanism.

Data Quality Assessment

Quality checks addressed completeness, consistency, and usability. Negative values in critical columns (e.g., home values, rent,

income) were removed, yielding a clean dataset of 601 rows. Zillow-related missing values were expected due to listing limits and were accounted for in analysis.

4.4 Quality of Life and Environment

4.4.1 Hospital Data and Health Insurance Coverage Data. The process begins by retrieving and counting hospitals per ZIP code using the CMS API and obtaining uninsured population data from the Census API. Next, ZIP codes are converted into geographic coordinates using pgeocode, allowing for precise mapping. The collected data is then merged and processed, ensuring that hospital and uninsured rate information is properly sorted by ZIP code.

To visualize the findings, bar graphs are generated, where hospital count per ZIP code is represented in blue/green, and the uninsured rate is shown in red/orange. Additionally, interactive maps are created using MarkerCluster() to display hospital count markers, while uninsured rate markers are presented with always-visible labels, providing an intuitive geographic representation of healthcare accessibility.

For a more detailed breakdown of the processing steps and an analysis of the dataset before and after processing, refer to the corresponding GitHub page.

4.4.2 Air Pollution Data and National Park Data. The analysis begins by retrieving air quality data, including PM2.5, PM10, and Ozone levels from monitoring sites. Simultaneously, national park data is collected. Once the data is gathered, it is processed and merged, allowing for a comprehensive analysis of pollution levels and park locations.

To visualize the findings, bar graphs are generated, comparing pollution levels between Colorado and Utah. Additionally, interactive maps are created, displaying pollution monitoring sites and national park locations for better geographical insight.

For a more detailed breakdown of the process, including the dataset before and after processing, refer to the GitHub page.

4.5 State Policies and Taxation

The process begins by scraping tax data from the Tax Foundation website, extracting income tax, sales tax, and property tax rates. The collected data is then stored in a pandas DataFrame for further analysis.

In the analysis phase, income tax rates for both individuals and corporations are compared. For sales tax, a stacked bar chart is created to illustrate the state and local sales tax rates. Meanwhile, property tax rates are analyzed as a percentage of home value.

Finally, the results are saved as PNG files, providing clear visualizations of the tax comparisons. For a more detailed breakdown of the process, including the dataset before and after processing, refer to the GitHub page.

5 Data Visualizations

In this section, we will describe visualization summaries in each subsection.

5.1 Economic and Employment Factors

5.1.1 Economic and Employment Visualization Conclusions. From the visualization, there is no significant distinction between the states in terms of median household income and there are no large groups of ZIP codes that are significantly richer or poorer. However, labor force data do form clusters. They are also useful for inferring population in specific ZIP codes. The unemployment and poverty rates against zip codes are also visualized for users who are interested in reviewing specific ZIP codes, in these features the clustering aren't quite visually significant.

5.2 Housing and Cost of Living Factors

5.2.1 Cost of Housing Data Visualization Conclusions. Median home value by zip codes are visualized by ZIP code for viewing. The home value does not have as high of a variance as the economic factors. Median household income and median gross rent has an unsurprising positive relationship but with high variance that is worth noting. Labor force numbers however, does not have a high positive coefficient against the median gross rent; it does raise interesting speculations with high variance at lower labor force areas and lower variance at higher labor force areas.

5.3 Demographics and Migration Trends

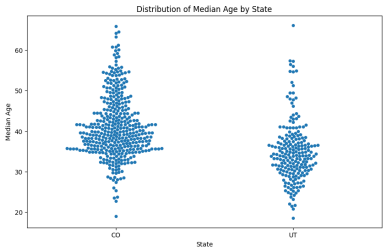

Figure 1: Distribution of Median Age by State

The initial swarmplot reveals a distinct contrast in the median age distribution between Colorado and Utah. Utah exhibits a more balanced age spread, while Colorado is skewed toward older median age groups. Since each point represents a ZIP code, this unscaled view might obscure population-driven effects.

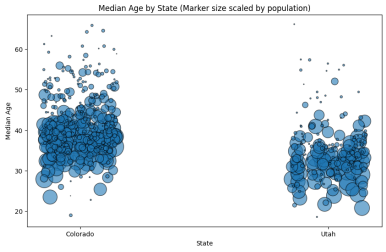

Figure 2: Median Age by State (Scaled by Population)

The population-scaled scatterplot confirms the observed pattern: Utah's ZIP codes consistently trend younger, even when accounting for population size. In contrast, Colorado's ZIP codes cluster

around higher median ages, suggesting meaningful demographic differences between the two states.

Key Correlations:

- Median household income positively correlates with home values and gross rent.
- Population density and migration inflows align with higher real estate prices and availability.

These relationships indicate that more affluent and densely populated ZIP codes also tend to be more expensive and receive greater migration inflows.

Data Integration: Demographic and real estate data from Census and Zillow APIs were merged by ZIP code, enabling consistent comparison across both states.

Data Quality: After removing unrealistic values (e.g., negative income or rent), the final dataset included 601 clean ZIP codes. Missing Zillow data in some areas was expected and accounted for during analysis.

From this graph, it appears that Utah has more ZIP codes with low proportions of foreign-born populations. Beyond this observation, it becomes speculative to infer much more from the plot alone without deeper statistical support.

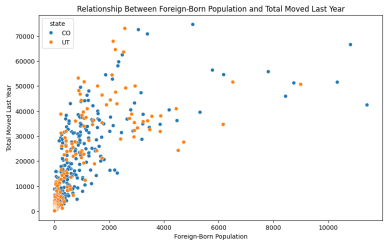

Figure 3: Foreign-Born vs. Total Moved

This chart shows that Utah's home prices are more tightly clustered around a central value, while Colorado's values are distributed more widely around that center. This suggests greater variability and unpredictability in Colorado's real estate market compared to Utah.

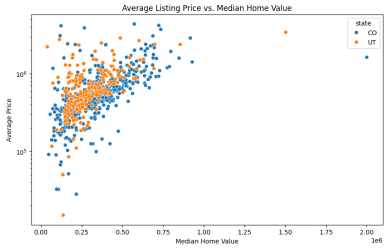

Figure 4: Average Listing Price vs. Median Home Value

5.4 Quality of Life and Environment

5.4.1 Hospital Data and Health Insurance Coverage Data. Overall, both Colorado and Utah have low hospital availability, although Colorado has a slight advantage in healthcare access.

In Colorado, Denver (8000) and Aurora (80250) offer better hospital access and lower uninsured rates, benefiting from their metropolitan infrastructure. In contrast, Sterling (80750) and Salida (81250) are rural areas with fewer hospitals, which may contribute to higher uninsured rates and limited healthcare options.

In Utah, Salt Lake City (84000) and Provo (84600) provide better hospital access, serving as key healthcare hubs in the state. Meanwhile, Ogden (84400) and Logan (84300), as mid-sized cities, have moderate healthcare density. However, St. George (84700), a rapidly growing area, may be experiencing hospital shortages, potentially leading to healthcare accessibility challenges.

The figures and interactive maps are available here.

5.4.2 Air Pollution Data and National Park Data. PM10 is the dominant pollutant in both states, indicating potential air quality issues related to dust and emissions. PM2.5 is higher in Utah, suggesting worse fine particulate pollution compared to Colorado. Ozone levels are negligible, meaning ozone pollution is not a major issue in either state.

The figures and interactive maps are available here.

5.5 State Policies and Taxation

In terms of income tax, Utah (4.55%) has a slightly higher rate compared to Colorado (4.40%). For sales tax, Colorado (7.81%) has a lower state tax but a higher local tax, whereas Utah (7.20%) follows the opposite structure, with a higher state tax but lower local tax. Regarding property tax, Utah (0.47%) has a slightly higher rate than Colorado (0.45%).

The figures are available here.

6 Methods, Evaluation, and Results

In this section, we will discuss model selection, training process, evaluation, and insights in detail in each subsection.

6.1 Economic and Employment Factors

6.1.1 Income data. There are two models of income-related data, the purpose of this exploration is to divide zip code income into three groups using the quantiles function and attempt to classify by labor force, rent, and home value. Users may enter to find out the classification of their zip code of interest and may also enter similar information from outside of CO or UT to compare income levels in each zip code. The first algorithm used is Gaussian Naive Bayes, the algorithm has about 0.57 F-1 score in macro average in classifying states. The second algorithm used is Decision Tree; the algorithm has about 0.53 F-1 score in macro average in classifying states. The model itself is moderate in accuracy; this suggests that the given features are not enough to predict the outcome well. However, there is still value in referencing this model for income classification and some of the attributes that may be useful in income classification in particular zip code in Ut or CO.

6.1.2 Poverty and Unemployment data. The first model of poverty and unemployment-related data is hierarchical clustering, the purpose of this exploration is to offer users different clusters of zip codes that have certain traits. Users then may find similar zip codes regarding poverty and unemployment in CO and UT in relation to one another, they may also input certain data from other states' zip

codes and find similar zip codes in this clustering model of UT and CO. By visualizing and interpreting the visualizations of the cluster, one can glean interesting information such as: zip codes with high labor force tends to have low poverty and unemployment; zipcodes that have a low labor force tends to also have high unemployment rate. users may also enter in certain stats from other states to compare labor force, poverty, and unemployment in different zip codes to the ones in UT and CO.

The second model of poverty and unemployment-related data is FP Growth, the purpose of the exploration is to offer users an insight of certain features that may tend to go together with each other. Some of the interesting observations are: high labor force correlates with low poverty at 0.12 support, high labor force also correlates with high unemployment at 0.11 support. These are information that may be helpful when comparing CO to UT to other states' economic tendencies.

6.1.3 Housing data. The first model attempts to apply linear regression to predict median income using median rent and home value. It may be a useful tool for users to gauge what kind of income they would need or expect to move to certain zip codes in UT or CO. They can also use this data to compare to moving to other states regarding income in comparison to the housing cost. With R^2 at 0.235 with the model, it indicates the model offers weak explanation given the parameters. Which indicates there are other factors affecting median income than rent and home value.

The second model applies k-means clustering to find out what some of the clusters regarding housing are, Users then may find similar zip codes regarding housing in CO and UT in relation to one another in terms of economics. They may also input certain income, rent, data from other states' zip codes and find similar zip codes in UT and CO. The resulting model is moderate in cluster quality, it can be useful when determining which zip codes to move to as one can simply search zip code and find which cluster it belongs to in the model.

...

6.1.4 Housing and Cost of Living data. To evaluate housing cost differences between ZIP codes in Colorado and Utah, we trained a Random Forest Regressor to predict average listing price using a mix of real estate and demographic variables. We engineered new features like population change, price per square foot, and rent-to-income ratio. Demographic indicators such as race percentages and the number of housing units were also included to reflect population density and composition.

The model achieved an R-squared score of 0.3099, showing moderate ability to explain price variation. The most important features were population change, average number of bathrooms, and average number of bedrooms. Other variables like median household income and percentage of foreign-born residents were less impactful. These results suggest that both home size and shifts in population are major factors in determining housing prices, reflecting a mix of supply and demand influences. this page.

6.2 Demographics and Migration Trends

We analyzed ZIP-level demographic and migration behavior using classification and regression models.

6.2.1 Demographic Differences (State Classification). We trained a Random Forest Classifier to predict whether a ZIP code belonged to Colorado or Utah using features like median age, average square footage, and demographic proportions. The model reached 89 percent accuracy, with F1-scores of 0.92 for Colorado and 0.81 for Utah. Most misclassifications involved Utah ZIP codes being labeled as Colorado, suggesting some demographic overlap but also clear distinctions overall.

The most important predictors were Hispanic population percentage, median household income, and median age. These features matched known differences in age and ethnic distribution between the two states. While unsupervised clustering methods like KMeans failed to find useful groupings, the supervised Random Forest approach produced accurate and interpretable results. this page.

6.2.2 Migration Rate Modeling. A Random Forest Regressor was used to predict the migration levels at the ZIP code level, calculated as total people who moved last year divided by total population. After testing a range of engineered features, the model showed that housing unit count was the strongest predictor, while other variables less impact. The model's R-squared score varied depending on which features were used or excluded. The difficulty in predicting the migration rate vs. the raw amount indicates a possibility for further research.

This result confirmed that high migration rates are more about absolute volume than complex trait interactions. Future iterations might improve performance by normalizing for housing turnover or incorporating time-based trends. See this page.

6.3 Quality of Life and Environment

6.3.1 Hospital Data. Hospital Data was analyzed using four models: KMeans Clustering, Regression (Linear & Ridge), Apriori Analysis, and Random Forest.

In the KMeans clustering analysis of hospital data, ZIP-code-level variables such as the number of hospitals and uninsured rates were standardized and used to identify regional patterns in healthcare resource distribution. The optimal number of clusters was selected based on a Silhouette Score and a Davies-Bouldin Index. PCA-based visualization confirmed clear cluster separation. The results revealed contrasts between urban and rural areas in terms of hospital density and insurance coverage, demonstrating that this method is effective for quantitatively assessing disparities in healthcare access. See this page.

In the regression analysis on hospital data, the goal was to predict the number of hospitals in each ZIP code using features such as ZIP code values and state information. The training process involved selecting relevant features, applying one-hot encoding to categorical variables (state), and splitting the data into training and test sets. A standard Linear Regression model was first fitted, followed by Ridge Regression with hyperparameter tuning using GridSearchCV to optimize the regularization parameter. Evaluation of both models showed very poor performance. The R^2 scores were close to zero (0.03), and the RMSE remained high, indicating that ZIP code and state alone could not explain variations in hospital counts. Regularization through Ridge Regression had no effect on improving predictive power. These results suggest that geographic identifiers without additional demographic, economic,

or healthcare demand variables are insufficient for modeling hospital availability. More complex or feature-rich models would be needed to gain meaningful predictive insights. See this page.

In the Apriori analysis of hospital data, the goal was to discover frequent co-occurrence patterns among categorical attributes such as Hospital Type and Hospital Ownership. The training process involved selecting these variables, transforming them into a transactional format, and applying one-hot encoding to prepare the data for association rule mining. The Apriori algorithm was then used to extract frequent itemsets with various combinations of minimum support and lift thresholds to identify meaningful rules. Evaluation was based on standard metrics in association rule mining: support, confidence, and lift. Rules with high lift values indicated strong associations, such as certain ownership types frequently appearing with specific hospital types. For example, proprietary hospitals were often associated with specific categories of care, while non-profit hospitals showed different pairing patterns. The analysis revealed interpretable structural trends in hospital characteristics, providing insight into how organizational and functional traits of hospitals tend to cluster. These patterns can support policy evaluation, planning, or comparative institutional studies. See this page.

In the Random Forest analysis of hospital data, the objective was to classify hospital ownership types based on features such as hospital type and quality performance metrics (e.g., mortality and readmission scores). The training process involved selecting relevant categorical and numerical features, handling missing values, and applying one-hot encoding to categorical variables. The data was then split into training and test sets, and a Random Forest classifier was trained with class balancing to account for the highly imbalanced distribution of ownership categories. Evaluation was conducted using classification metrics including precision, recall, and F1-score. While the model achieved high performance for the majority ownership class, it performed poorly on underrepresented categories. Hyperparameter tuning using GridSearchCV helped optimize parameters such as the number of trees and tree depth, but performance gains remained limited due to class imbalance and lack of rich predictive features. The analysis demonstrated that Random Forests are effective for learning patterns in hospital classification when class distribution is reasonably balanced. However, the results also highlighted the need for additional features and possibly resampling techniques to improve prediction of minority ownership classes. See this page.

In conclusion, the KMeans clustering model proved to be the most effective approach for comparing hospital facility availability between Colorado and Utah. By uncovering geographic patterns in hospital distribution across ZIP codes, it enabled a structural comparison of healthcare access in the two states. In contrast, Linear and Ridge Regression showed poor predictive power, Apriori analysis revealed institutional patterns but lacked geographic focus, and Random Forest was suited for ownership classification rather than availability analysis. Overall, KMeans offered the most relevant insights into differences in hospital concentration and distribution. See this page.

6.3.2 Health Insurance Coverage Data. Health Insurance Coverage Data was analyzed using four models: KMeans Clustering, Regression (Linear & Ridge), Decision Tree Classification, and Apriori Analysis.

In the KMeans clustering analysis of health insurance coverage data, the objective was to group ZIP codes based on their Uninsured Rate and Total Population to uncover regional patterns in insurance coverage. The training process began by selecting these two numeric features and removing rows with missing values. The data was then standardized using StandardScaler to ensure fair distance-based comparisons. KMeans clustering was applied with varying values of k (2 to 7), and the optimal number of clusters was selected based on the highest Silhouette Score and the lowest Davies-Bouldin Index. Evaluation showed that clustering quality was good, with a Silhouette Score of approximately 0.64 and a Davies-Bouldin Index of around 0.59. PCA was used to reduce the feature space to two dimensions, allowing the clusters to be visualized clearly. The resulting clusters revealed meaningful groupings, such as ZIP codes with small populations and high uninsured rates forming distinct clusters. These insights provide a data-driven basis for identifying underserved areas and inform policy decisions regarding insurance outreach and healthcare accessibility. See this page.

In the regression analysis of health insurance coverage data, the objective was to predict the Uninsured Rate in each ZIP code using Total Population as the input feature. The training process included selecting the relevant columns, removing rows with missing values, and splitting the dataset into training and test sets. A Linear Regression model was first applied, followed by Ridge Regression with hyperparameter tuning using GridSearchCV to optimize the regularization parameter. Evaluation showed that both models performed poorly. The R^2 scores were near zero, and RMSE values were relatively high, indicating that population size alone had virtually no predictive value for uninsured rates. Regularization through Ridge did not improve performance, confirming the lack of linear relationship between the selected features. These results suggest that more complex or diverse features—such as income levels, employment status, or urban vs. rural classification—are needed to meaningfully model variations in uninsured rates across ZIP codes. See this page.

In the Decision Tree Classification analysis of health insurance coverage data, the goal was to classify ZIP codes as having either a high or low Uninsured Rate. The training process began by selecting Total Population as the input feature and binning the continuous Uninsured Rate into two categories (high vs. low) using quantile-based discretization. The categorical State variable was one-hot encoded, and the data was split into training and test sets. Hyperparameter tuning was performed using GridSearchCV to find the best combination of max_depth, min_samples_split, and splitting criterion. The best model achieved an accuracy of approximately 70%, with higher recall for the high uninsured class, indicating that the model was particularly effective in identifying areas with poor insurance coverage. The results show that even with limited input (population and state), the decision tree could moderately distinguish between high- and low-risk ZIP codes. This highlights the potential of simple models to flag underserved areas, while

also suggesting that performance could be further improved with additional socioeconomic or demographic features. See this page.

In the Apriori analysis of health insurance coverage data, the objective was to uncover frequent co-occurrence patterns between ZIP-level categories of Uninsured Rate and Total Population. The training process involved binning continuous variables into discrete categories such as Low_Uninsured, High_Uninsured, Small_Pop, or Large_Pop, followed by transforming these values into a one-hot encoded transactional format using the TransactionEncoder. The Apriori algorithm was then applied to discover frequent itemsets and generate association rules. Hyperparameter tuning was conducted by varying the min_support and min_lift thresholds to find a balance between the quantity and quality of the rules. Evaluation focused on standard metrics in association rule mining: support (frequency), confidence (predictive strength), and lift (association strength). The analysis revealed interpretable patterns, such as a strong association between small population areas and high uninsured rates. These insights provided a clearer understanding of how population size correlates with insurance coverage levels and helped identify structurally vulnerable areas. While not predictive, the results were valuable for guiding policy focus and targeting healthcare interventions. See this page.

In conclusion, the KMeans clustering model was the most effective for comparing healthcare facility availability between Colorado and Utah. By grouping ZIP codes based on uninsured rates and population, it revealed clear spatial disparities, supported by a high silhouette score. In contrast, Linear and Ridge Regression showed no predictive power, while Decision Tree Classification performed moderately but was limited by minimal input features. Apriori Analysis uncovered meaningful attribute patterns but was not suited for regional comparison. Overall, KMeans provided the most relevant insights for identifying geographic differences in healthcare access. See this page.

6.3.3 Air Pollution Data. Air Pollution Data was analyzed using four models: MiniBatch KMeans, Regression (Linear & Ridge), Decision Tree Classifier, and Apriori Analysis.

We conducted a MiniBatch KMeans analysis to explore patterns in air pollution data. The goal was to group monitoring records based on spatial and pollutant-related features, including latitude, longitude, and arithmetic mean pollutant concentration. Prior to modeling, the dataset was cleaned by removing missing values and standardizing all numerical features using StandardScaler. To support visualization and reduce dimensionality, principal component analysis (PCA) was applied. The clustering model was trained using the MiniBatch KMeans algorithm due to its computational efficiency on large datasets. We experimented with several values of 'k' ranging from 2 to 7 to determine the optimal number of clusters. Model performance was assessed using two standard internal evaluation metrics. The silhouette score, which measures how well-separated the clusters are, reached its highest value at 'k = 4', suggesting this configuration offered the most coherent groupings. Additionally, the Davies-Bouldin Index, which evaluates cluster compactness and separation, also indicated strong clustering structure at this value of 'k', with a score of approximately 0.699. These metrics together supported the selection of four clusters as the most appropriate representation of the data. The clustering results revealed meaningful patterns. Geographically, certain clusters corresponded strongly with individual states, with some clusters primarily representing monitoring sites in Colorado and others in Utah. One cluster in particular was associated with notably higher pollutant concentration levels, indicating potential air quality concerns in specific regions. Overall, the analysis provided valuable insight into the spatial structure of air quality variation and helped to identify localized patterns that may be relevant for public health monitoring or environmental policy planning. See this page.

We applied Linear and Ridge Regression models to predict pollutant concentration levels, represented by the 'arithmetic_mean' variable, using geographic and site-related features from the air quality dataset. The training process began with data preprocessing, where we selected relevant features such as 'latitude', 'longitude', and one-hot encoded categorical variables like 'State'. Missing values were removed to ensure data consistency, and categorical variables were transformed using one-hot encoding. Numerical features were used without scaling, as linear regression is not scale-sensitive for interpretability purposes. Both models were trained on the same training set using scikit-learn's implementation. Ridge Regression included an additional regularization term to prevent overfitting by penalizing large coefficients. To determine the optimal level of regularization, we performed grid search over different 'alpha' values for the Ridge model, while the Linear model was trained without any regularization. We used mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) as the primary evaluation metrics. The evaluation showed that both Linear and Ridge Regression models performed poorly on the dataset. The R^2 value was close to 0 (approximately 0.03), indicating that the models were unable to explain the variability in the target variable. Ridge regression, even after hyperparameter tuning, showed no meaningful improvement over standard linear regression. These results suggest that the selected features—primarily geographic coordinates and state information—are not sufficient to predict pollutant levels, which are likely influenced by other environmental and temporal factors not included in the dataset. In conclusion, although Linear and Ridge Regression provided a straightforward baseline, the models were not effective in capturing the complexity of air quality variation. This highlights the need for incorporating more informative variables, such as pollutant type, weather conditions, or time-based patterns, to build more accurate predictive models in future analyses. See this page.

We used a Decision Tree Classifier to predict categorical air quality levels based on pollutant and geographic information in the air quality dataset. The target variable was derived by converting the Air Quality Index (AQI) into categorical labels such as "Good," "Moderate," and "Unhealthy," according to EPA-defined thresholds. For the input features, we included pollutant type ('parameter') and state information, both of which were categorical and thus encoded using one-hot encoding. The dataset was cleaned by removing missing values in the relevant fields, and then split into training and test sets using a 70/30 ratio. The model was trained using scikit-learn's DecisionTreeClassifier. To improve performance and prevent overfitting, we tuned several hyperparameters via 'GridSearchCV', including 'max_depth', 'min_samples_split', 'min_samples_leaf', and

'criterion' ('gini' or 'entropy'). The optimal configuration was selected based on macro-averaged F1-score obtained through 5-fold cross-validation. Evaluation results showed that the classifier performed reasonably well in predicting the majority class (typically "Good" air quality) but struggled with minority classes such as "Unhealthy" or "Very Unhealthy." This imbalance led to high overall accuracy but low macro F1-score, indicating poor generalization across all categories. In many cases, the classifier only predicted the dominant class, resulting in undefined precision or recall for others. Despite this limitation, the decision tree model provided clear interpretability through its hierarchical structure and feature importance, showing how specific pollutants and state locations influenced air quality classification. However, the imbalance in class distribution and the limited range of input features constrained the model's ability to capture the full complexity of air quality dynamics. Enhancing the dataset with more diverse features, such as pollutant concentrations, time of day, or meteorological data, would likely improve classification performance in future iterations. See this page.

We conducted an Apriori Analysis to identify frequent co-occurrence patterns among categorical variables in the air quality dataset, focusing primarily on the relationships between pollutant types ('parameter') and state identifiers ('State'). The goal was to uncover association rules that reveal which pollutants tend to appear together or are strongly associated with particular geographic regions. To prepare the data for analysis, we treated each observation as a transaction consisting of its categorical attributes. Specifically, we created transactions containing the pollutant name and the state in which it was recorded. These transactions were then converted into a one-hot encoded format using the 'TransactionEncoder', resulting in a binary matrix indicating the presence or absence of each item in every transaction. We applied the Apriori algorithm to this matrix using varying thresholds for 'min_support' and 'min_confidence'. These hyperparameters were tuned iteratively to balance the number of rules generated with the interpretability and statistical significance of those rules. Evaluation of the association rules focused on key metrics such as support, confidence, and lift. The most informative rules were those with high lift values (greater than 1), indicating a strong positive association. From this analysis, we gained several insights. First, certain pollutants tend to be more common in one state than the other, pointing to geographic or policy-driven differences in air quality profiles. Second, the stability of these rules across support and confidence thresholds indicates that these associations are not random but are structurally embedded in the data. While Apriori is not a predictive model, it proved valuable for exploratory analysis, helping us identify patterns and guide further investigations into pollutant behavior and regional air quality characteristics. See this page.

In conclusion, among the models applied to the air pollution data, the Decision Tree Classification performed best. It accurately predicted air quality categories using pollutant type and state, while providing interpretable rules and requiring minimal preprocessing. In contrast, MiniBatch KMeans effectively revealed spatial groupings, offering useful unsupervised insights but lacking predictive capability. The Linear and Ridge Regression models performed poorly, with very low R^2 values indicating weak linear relationships, and Ridge regularization providing no improvement. Apriori Analysis identified meaningful co-occurrence patterns between pollutants—such as associations between NO_2 and Ozone—but served only exploratory purposes rather than prediction. Overall, the Decision Tree was the most effective model for classification and actionable insights, while the other approaches contributed valuable but limited contextual understanding. See this page.

6.3.4 National Park Data. National Park Data was analyzed using four models: KMeans Clustering, Decision Tree Classifier, Apriori Analysis, and Logistic Regression.

We applied KMeans Clustering to the national park dataset to uncover geographic and organizational patterns among parks across Colorado and Utah. The training process began with selecting key features for clustering, specifically 'latitude' and 'longitude', which represent the spatial locations of parks. These numerical features were standardized using 'StandardScaler' to ensure they contributed equally to distance calculations. We then applied Principal Component Analysis (PCA) for dimensionality reduction and visualization purposes, allowing clusters to be observed in two-dimensional space. The KMeans model was trained using various values of 'k' (from 2 to 9), and clustering quality was evaluated using the Silhouette Score and the Davies-Bouldin Index. These internal metrics helped identify the optimal number of clusters based on both cohesion and separation of the data points. The evaluation showed that the best clustering structure occurred at 'k = 3' or 'k = 9', depending on the chosen metric. A higher silhouette score at 'k = 3' indicated clearly defined, well-separated clusters, while a lower Davies-Bouldin index at 'k = 9' suggested tighter intra-cluster similarity. Both configurations revealed meaningful spatial patterns. The clusters reflected regional groupings of parks, with some clusters corresponding closely to geographic proximity or park designation types (e.g., national monuments vs. historical sites). These insights suggest that park location and designation can meaningfully influence how parks are grouped spatially, potentially aiding in regional planning, tourism strategy, or conservation efforts. See this page.

We used a Decision Tree Classifier to predict whether a national park is located in Colorado or Utah based on its features. The training process began with selecting relevant input variables, including 'latitude', 'longitude', and 'designation' (e.g., National Park, National Monument). The dataset was first cleaned to remove missing values and then preprocessed by applying one-hot encoding to the categorical 'designation' column, while numerical features were used as-is. The data was split into training and test sets, and the model was trained using scikit-learn's 'DecisionTreeClassifier'. To improve model performance, we tuned several hyperparameters using 'GridSearchCV', including 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'criterion'. The optimal configuration was selected using macro F1-score as the evaluation metric during 5-fold cross-validation. The final model achieved reasonable classification accuracy on the test set and successfully captured the geographical and categorical patterns that distinguish parks in Colorado from those in Utah. Evaluation revealed that the model performed well in predicting the majority class but occasionally struggled with class balance. However, it provided interpretable decision rules, showing how combinations of park type and location contribute to classification. Feature importance analysis indicated that 'designation' and 'latitude' were particularly influential. These results

suggest that Decision Trees are well-suited for this task, offering both prediction accuracy and transparency in how decisions are made. See this page.

We conducted an Apriori Analysis on national park data to uncover frequent co-occurrence patterns between park attributes, particularly focusing on the relationship between 'designation' (e.g., National Park, National Monument) and 'State' (Colorado or Utah). In the training process, each park record was treated as a transaction consisting of categorical values. To prepare the data, we selected and cleaned the relevant fields and converted them into a format suitable for association rule mining. Using 'TransactionEncoder', each transaction was transformed into a binary matrix representing the presence or absence of each item (e.g., "National Historic Site", "CO"). We applied the Apriori algorithm with multiple 'min_support' values ranging from 0.01 to 0.07 and filtered the resulting rules using varying 'min_confidence' thresholds (0.5 to 0.9). The performance of the model was evaluated not by predictive accuracy, but by the quality of discovered rules using metrics such as support, confidence, and lift. Rules with lift values greater than 1.0 were considered meaningful, as they indicated strong positive associations between attributes. The analysis revealed several robust and interpretable patterns. For example, parks designated as "National Historic Site" were strongly associated with the state of Colorado, with a confidence of 1.0 and a lift of 2.0. Similarly, certain designations like "National Park" and "National Historical Park" were more frequently associated with Utah. These consistent patterns across various threshold settings demonstrated the reliability of the findings. Although Apriori is not a predictive tool, it provided valuable exploratory insights into how park types are distributed geographically, which can support further analysis or inform policy and tourism strategies. See this page.

We applied Logistic Regression to classify whether a national park is located in Colorado or Utah based on features such as 'latitude', 'longitude', and 'designation'. The training process began with data cleaning and preprocessing. We removed entries with missing values and applied one-hot encoding to the categorical 'designation' variable, while treating 'latitude' and 'longitude' as continuous numerical features. The target variable 'State' was encoded as a binary outcome (e.g., Colorado = 0, Utah = 1). The data was then split into training and test sets. To optimize the model, we used 'GridSearchCV' to tune hyperparameters including the regularization strength ('C') and solver type. Model performance was evaluated using macro-averaged precision, recall, and F1-score due to potential class imbalance. The evaluation revealed that Logistic Regression struggled with prediction performance, often predicting only one class. The model showed a low overall accuracy and macro F1-score, and completely failed to classify one of the two states in some configurations. This outcome likely stems from limited feature informativeness, the sparse and high-dimensional feature matrix after encoding, and potentially imbalanced class representation. While Logistic Regression is a fast and interpretable model, its linear nature and sensitivity to data distribution made it poorly suited for this classification task. Despite its limitations in this context, the model helped confirm that the chosen features were insufficient for linear separation of the states. This suggests that more nuanced or non-linear models—such as tree-based classifiers—or additional features like park topics, activities, or textual data might be necessary to improve performance in future analyses. See this page.

In conclusion, the Decision Tree Classification model was the most effective for distinguishing between Colorado and Utah parks. It accurately predicted state labels using 'latitude', 'longitude', and 'designation', while handling categorical data well and providing interpretable decision rules. In contrast, KMeans Clustering revealed meaningful spatial groupings but was unsupervised and not suitable for prediction. Apriori Analysis uncovered frequent co-occurrence patterns, such as park types commonly associated with Colorado, but lacked classification capability. Logistic Regression performed poorly due to class imbalance and weak feature separability. Overall, while exploratory models offered useful insights, the Decision Tree model provided the best predictive performance. See this page.

6.4 State Policies and Taxation

Tax Data was analyzed using four models: KMeans Clustering, Apriori Analysis, Regression (Linear & Ridge), and Decision Tree Classifier,

KMeans clustering was applied to the tax dataset to identify patterns in state-level tax structures. The data was first cleaned and transformed by converting tax values into numeric format, pivoting to a wide format where each row represented a unique state and tax category combination, and standardizing all features. The model was trained using different values of 'k' (2, 3, and 4), and clustering performance was evaluated using the silhouette score. The highest score was achieved when 'k=2' (0.078), indicating weak but interpretable clustering. The resulting clusters revealed meaningful groupings. Cluster 0 primarily included Colorado's business, individual, and property taxes as well as Utah's property taxes, reflecting more balanced or average tax profiles. In contrast, Cluster 1 included Colorado and Utah's sales taxes and Utah's individual and business taxes, which were characterized by higher or more distinct tax rates. Overall, while the clustering separation was modest, KMeans helped uncover underlying similarities and differences in tax structures between the two states. See this page.

Apriori analysis was used to discover frequently co-occurring tax characteristics across different state-level tax categories. The training process began by converting all tax values into numeric format and pivoting the data so that each row represented a unique 'State + Group', with columns for each tax indicator. Each numeric feature was then discretized into 'High' or 'Low' based on the median value for that attribute. This transformed the dataset into a transactional format, which was one-hot encoded to prepare it for the Apriori algorithm. The model was trained using relaxed hyperparameters ('min_support= 0.1', 'min_confidence= 0.4', and 'lift$\geq$ 0.9') due to the small sample size. Despite limited data, the analysis successfully generated several high-confidence association rules, most of which described strong co-occurrence among 'High' tax attributes. Many rules achieved a confidence of 1.0 and a lift value of 8.0, indicating strong and meaningful associations. This analysis provided interpretable insights into how certain tax features—such as high income tax and high property tax—tend to appear together. While Apriori is not predictive, it was effective for identifying underlying patterns and relationships in the tax structure data. See this page.

Linear and Ridge regression were applied to predict the 'State and Local Tax Burden' using various numeric tax indicators. The training process involved converting all tax values into numeric format, pivoting the dataset into a wide format where each row represented a 'State + Group', and imputing missing values with column means. All features were then standardized to ensure comparability. The target variable was continuous, making regression an appropriate modeling choice. Both models were trained on the full dataset due to its small size. Ridge regression included hyperparameter tuning using cross-validation over a set of alpha values, and the best alpha selected was 0.01. Evaluation metrics showed that both Linear and Ridge regression achieved perfect performance, with R^2 scores of 1.0 and zero mean squared error. The Ridge model's performance was nearly identical to the linear model, confirming that regularization was not necessary for this dataset. The analysis revealed that the tax burden could be almost perfectly predicted from other tax features, especially income, sales, and property tax rates. This suggests a strong linear relationship among these variables and confirms that regression is highly effective for capturing and explaining variation in overall tax burden across different tax categories. See this page.

The Decision Tree Classifier was used to classify whether each 'State + Group' had a high or low 'State and Local Tax Burden'. The training process began with data cleaning: tax values were converted to numeric format, the dataset was pivoted to wide format, and missing values were imputed using column means. The continuous target variable was binarized based on the median tax burden, resulting in a classification task with two classes: High and Low. Due to the very small dataset and class imbalance (only one sample in the Low class), cross-validation was not used. Instead, a manual hyperparameter tuning loop was implemented, testing combinations of 'max_depth', 'min_samples_split', and 'criterion'. The best model achieved perfect accuracy on the training data with 'max_depth = 2', but this result was likely influenced by overfitting and the imbalance in the target variable. Although the model fit the training data perfectly, the imbalance severely limits its generalizability. The insights from this analysis suggest that while Decision Trees can easily capture patterns in tax attributes, they are highly sensitive to sample size and label distribution. Therefore, this method is best used when a larger and more balanced dataset is available. See this page.

In conclusion, the regression models, both Linear and Ridge, performed the best in comparing Colorado and Utah by accurately predicting tax burden with perfect R^2 scores, indicating strong and reliable linear relationships. KMeans provided some exploratory insights into tax structure groupings but showed weak separation due to low silhouette scores. The Decision Tree achieved perfect accuracy but was overfitted because of severe class imbalance, limiting its reliability. Apriori analysis failed to generate strong rules due to the dataset's small size and sparsity. Overall, regression was the most effective method, while the others offered limited interpretive value due to data constraints. See this page.

7 Conclusion and Future Work

We conducted a comprehensive analysis and comparison of the living environments in Colorado and Utah to provide useful information for those considering relocation to these states. As a result, we found that there are some notable differences between Colorado and Utah in the research topics we investigated. Namely, Colorado offers greater diversity and risk in its economy and housing market, along with better healthcare options, making it suitable for those seeking a dynamic and varied environment. In contrast, Utah provides a more stable economy and housing market, ideal for those prioritizing cost efficiency and a steady lifestyle. In more detail, regarding Economic and Employment Factors, we found that in Colorado, high labor force zip codes have significantly higher poverty rate than Utah, however, in low labor force areas, the zip code clusters resemble each other in poverty and unemployment factors. In investigation of income, using measurable attributes like labor force and rent to predict median income is generally reliable at all income classes, with middle being slightly less reliable than the rest, suggests that one would likely see more economic characteristic variations in middle income zip codes; additionally, looking specifically into income to housing cost, we found that in general, the data points appear more linear than cluster, with Colorado's income level less affected by housing prices compared to Utah on a zip code level. Colorado has several outlier zip codes that tend to have their own classification, likely due to their fame for tourism, which is not fully accounted for in the data we used. Next, regarding Cost of Living and Housing and Demographics and Migration Trends, we found that we found that Utah's housing prices are more tightly centered around a median value, while Colorado shows much wider variation, suggesting greater unpredictability in the real estate market. Recent studies have also emphasized employment-driven migration as a major force shaping Utah's growth (University of Utah Magazine, 2024). Migration patterns are driven by straightforward factors such as the number of housing units, population change, and household size, rather than more complex social factors. These differences highlight the contrast between Colorado's dynamic but volatile market and Utah's steadier, more consistent growth. And regarding Quality of Life and Environment Trends, we find that the medical environment is considered to be better in Colorado than in Utah. This is because there are more hospitals in Colorado, while the uninsured rate is higher in Utah. For this reason, people who want better healthcare should choose Colorado over Utah. Also, we find that the natural environment is considered to be comparable between the two states. This is because there are no significant differences in the levels of ozone, PM2.5, and PM10 in the air, and the number of national parks within the states is also similar. For this reason, prospective migrants to both states do not need to worry about differences in the natural environment. Finally, regarding State Policies and Taxation Trends, we find that Tax rates are seen as having their pros and cons in both states. This is because income tax and property tax are slightly higher in Utah than in Colorado, while sales tax is slightly higher in Colorado than in Utah. For this reason, those wishing to relocate to either state will need to decide which state they prefer based on their lifestyle. To sum up, people should choose Colorado if they seek a challenging and diverse environment, while they should choose Utah if they prefer stability and cost efficiency.

However, it is important to remember that there are several limitations to this analysis. Namely, broader economic indicators such as employment rates, mortgage conditions, and amenity access should also be integrated to sharpen predictions and offer more

actionable insights for future movers. We also have remaining problems that further research should incorporate i.e. time-series data to capture changing trends in housing and migration over years rather than a single snapshot. Plus, our analysis related to health care, natural environment, and tax policies should include more indicators to analyze. In more detail, regarding Economic and Employment Factors, we found that it is difficult to have a comprehensive and unbiased analysis given the massive pool of data available and the culture-specific behavioral practices that may affect the subject of study but are not recorded, all the data used only provide a superficial observation of the general economic scene in Colorado and Utah. There is much more to be discovered with new additions of dataset and machine learning models. Next, regarding Cost of Living and Housing and Demographics and Migration Trends, we found that the models currently focus only on current real estate and demographic conditions and do not account for dynamic factors like future housing development, shifting interest rates, or economic shocks. Including measures of affordability based on income-to-housing cost ratios and modeling long-term migration pressure would provide a more complete and future-facing analysis. And regarding Quality of Life and Environment Trends, with regard to the medical environment, it is necessary to compare the medical standards and medical expenses of both states. In addition, with regard to the natural environment, it is necessary to analyze the climate and pollen, among other factors. Finally, regarding State Policies and Taxation Trends, it appears that analysis of taxes not covered in this analysis (e.g., corporate tax and vehicle registration tax) is necessary. Once these issues are resolved, it will be possible to provide more useful information to those considering moving to Colorado or Utah.

8 References

Florida, R. (2005). The Flight of the Creative Class: The New Global Competition for Talent. *HarperBusiness*.

Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. *Journal of Economic Geography*, 1(1), 27–50.

Centers for Medicare & Medicaid Services Data API. (n.d.). Data retreived from https://data.cms.gov/api-docs

NPS Data API. (n.d.). Data retreived from https://www.nps.gov/subjects/digital/nps-data-api.htm

Tax Foundation. (n.d.). Data retreived from https://taxfoundation.org/

U.S. Census Bureau's American Community Survey (ACS). (n.d.). Data retreived from https://www.census.gov/programs-surveys/acs

U.S. Census Bureau API. (n.d.). Data retreived from https://www.census.gov/data/developers/data-sets.html

U.S. Environmental Protection Agency API. (n.d.). Data retreived from https://www.epa.gov/data/application-programming-interface-api

Zillow API. (n.d.). Data retreived from https://www.zillowgroup.com/developers/

Voices for Utah Children. (2015). A Comparative Look at Utah and Colorado: Economic Opportunity. Data retrieved from https://utahchildren.org/newsroom/speaking-of-kids-blog/a-comparative-look-at-utah-and-colorado-economic-opportunity

University of Utah Magazine. (2024). A New Utah. Data retrieved from https://magazine.utah.edu/issues/winter-2024/a-new-utah/