# Predicting Housing Affordability and Migration Patterns

Carly Carroll & Ryan Tsang

---

Final Project Report

CSC/DSP 461: Machine Learning

Professor Marco Alvarez

December 11, 2024

---

University of Rhode Island

Department of Computer Science and Statistics

Full code and materials for this project can be accessed on Github:

https://github.com/carlycarroll25/DSP461_FinalProject

# Abstract

Housing affordability and migration patterns are critical factors with a heavy influence on urban planning, economic development, and the well being of society. This project utilizes socioeconomic, demographic, and migration data to predict affordability scores and analyze their relationship with migration trends across US counties. Using a combination of different analyses, the project provides a comprehensive plan to explore these relationships. The findings aim to support decision making for individuals, families, policy makers, and planners by presenting insights into affordability and migration patterns, promoting development, and enhancing quality of life.

# 1 Introduction

Finding a place to live can be a challenging task, especially when external factors such as family and work come into play. As such, thorough and proper research is essential in order to find a location in the United States that caters to an individual's needs and financial capabilities. Housing affordability and migration patterns are two interconnected factors that can significantly impact the well being of society, urban planning, and economic development. Housing costs and local resources can heavily influence where individuals and families choose to settle down, while migration patterns are a reflection of broader socioeconomic trends. By integrating diverse datasets and applying various techniques, this project seeks to find the relationship between where individuals chose to live and why. With the goal of presenting actionable insights, this analysis isn’t only relevant to individuals and families, but also policy makers,  urban planners, and economists aiming to understand patterns and improve decision making.

Previous studies have made use of economic and demographic data to predict housing affordability with a focus on housing prices, living costs, and income amounts. Likewise, migration studies have used inflow and outflow data to understand regional appeal and population changes. This project combines fields aiming to provide a more holistic view of housing affordability and its role in migration patterns.

# 2 Problem Definition

Develop an affordability score for each county that includes housing costs, income levels, local demographics, and living expenses. The score will indicate how financially feasible a particular area is for individuals or families to live comfortably in a given area. Additionally, this project aims to investigate how housing affordability influences migration patterns by analyzing inflow and outflow data. This includes understanding which locations are gaining or losing residents and whether the affordability score correlates with these trends.

# 3 Data

## 3.1 Data Overview

The datasets used in this project incorporate a variety of socioeconomic, demographic, and migration related attributes for multiple counties across the United States. It serves as the foundation for analyzing housing affordability and migration patterns. The dataset provides a broad overview of trends at both county and state levels, despite not covering all counties nationwide. It includes approximately 2000 rows and numerous detailed columns allowing for a detailed and adequate exploration of the relationships between affordability and migration patterns.

## 3.2 Key Data Features

Highlights the most critical variables in the dataset that influence housing affordability, migration patterns, and overall living conditions. The dataset provides a structured overview of metrics related to affordability, income, migration, demographics, living costs, employment, and crime. These variables collectively form for analyzing how financial, societal, and economic factors affect the appeal and feasibility of living in different US counties. By focusing on these variables, the analysis aims to uncover meaningful insights and relationships that are relevant to affordability and migration trends.

### 3.2.1 Affordability Metrics

**AffordabilityScore:** A composite score indicating how affordable or financially feasible it is to live in a particular area, combining housing costs, income levels, and living expenses for each household
**HousingCostAvg:** The average cost of housing in each county, encompassing both rental and ownership expenses
**TotalLivingCost:** A measure of living expenses in each county including a variety of basic living needs

### 3.2.2 Income and Poverty Metrics

**median_family_income:** The median household income in each county, representing the financial capabilities of the average household
**Income:** The total household income reported for each county, providing broader financial insights
**Poverty:** The percentage of the population living below the poverty line, highlighting economic challenges in each county

### 3.2.3 Migration Metrics

**INflow:** The total number of individuals moving into each county, reflecting the areas attractiveness
**OUTflow:** The total number of individuals leaving each county, reflecting the areas reduced appeal
**NET in:** The total net migration, calculated as the difference between inflow and outflow migration, indicating whether the population in a particular county is overall growing or declining
**GROSS out:** The measure of the total number of people entering and leaving an area, providing context for migration trends

### 3.2.4 Demographics

**TotalPop:** The total population of individuals living in each county
**Hispanic, White, Black, Asian:** Demographic breakdown showing the racial and ethnic composition of the population

### 3.2.5 Cost Breakdown

**housing_cost:** The cost associated with owning or renting a home
**food_cost:** The cost of food for an average household in each county
**transportation_cost:** Expenses related to commuting and other transportation needs
**heathcare_cost:** Cost associated with medical care and health services
**other_necessities_cost:** Cost for items beyond the essentials

### 3.2.6 Employment

**Unemployment:** The percentage of the labor force that is unemployed
**Employed:** The number of employed individuals

### 3.2.7 Crime Rates

**crime_rate_per_100000:** Crime rates per 100,000 people, influencing the safety and livability in each county

# 4 Methods

Analyzing housing affordability and migration trends requires a structured approach, implementing a variety of analysis techniques. This will encompass exploratory data analysis, statistical modeling, and machine learning methods to uncover meaningful patterns and relationships. These methods were chosen to ensure that the analyses effectively evaluate the factors influencing affordability and migration. Each technique that will be used plays a crucial role in understanding any underlying structure and trends within the data.

## 4.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to gain initial insights into the dataset and understand relationships between key variables. Descriptive statistics were calculated to summarize the distributions of various variables and graphical techniques were used to visualize distributions and correlations between them.

### 4.1.1 Key Visualization Techniques

**Histograms:** Frequency distributions were plotted for key variables to understand their variability and potential skewness along with an overview of data spread and central tendency
**Barplots:** Distributions displaying the relationships between affordability scores across all US states
**Scatterplots:** Relationships between affordability scores and migration data were visualized to identify potential trends and patterns
**Heatmaps:** Correlation heatmap was generated to quantify relationships among various variables, highlighting positive and negative associations across them

### 4.1.2 Relevant Statistical Equations

**Correlation Coefficient (*r)***
*Quantifies the strength and direction of relationships between two variables*

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

**Mean (μ)**
*Represents the average value of a variable*

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

**Standard Deviation (σ)**

*Measures the spread of a piece of data around the mean*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

**Linear Regression**

*Simple linear relationships between two variables (x,y)*

$$y = \beta_0 + \beta_1 x + \epsilon$$

**Distribution Fit**

*Examines how well a variable's observed distribution aligns with a theoretical distribution*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

## 4.2 Data Clustering

Clustering techniques were used to group counties based on affordability related characteristics, allowing for the identification of patterns and trends. Two primary approaches were used in the analysis: K-means clustering and hierarchical clustering.

### 4.2.1 K-Means Clustering

The K-means algorithm was implemented to divide counties into distinct clusters. The optimal number of clusters was determined using the Elbow Method which plots the within cluster sum of squares (WCSS) against the number of clusters. The point of inflection on this curve was chosen as the optimal cluster amount.

$$WCSS = \sum_{k=1}^{K} \sum_{i \epsilon C_k} ||x_i z_k||_2^2$$

*Where $C_k$ represents cluster k and $z_k$ is its centroid*

When implementing K-means clustering, the algorithm groups data points into clusters based on the similarities within. To start, a few centroids or central points are randomly selected, and then each data point is assigned to the nearest centroid. The algorithm then recalculates the centroids based on the new groups and the process repeats until the clusters are distinct. This process is shown in **Figure 1** below.

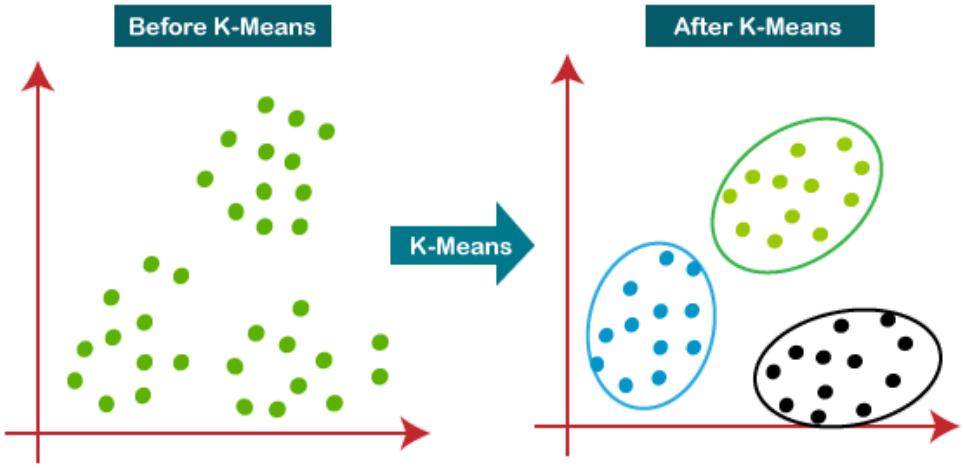



**Figure 1: Before and After K-means Implementation**

### 4.2.2 Hierarchical Clustering

Hierarchical clustering with Ward's linkage was applied to compare clustering results with K-means. This method minimizes the total variance within clusters. Two dendrograms were generated to visualize hierarchical clustering among the top and bottom 10 counties.

$$\Delta ESS = \frac{|A| \bullet |B|}{|A|+|B|} ||\mu_A - \mu_B||^2$$

*Ward's Linkage displaying the increase in ESS when two cluster A and B are merged*

Hierarchical clustering starts by treating each data point as its own cluster. It then identifies the two closest clusters and merges them into a larger cluster. This process then repeats until all data points are in one large cluster. There are two approaches to hierarchical clustering including agglomerative (bottom-up approach) and divisive (top-down approach) displayed in **Figure 2** below.

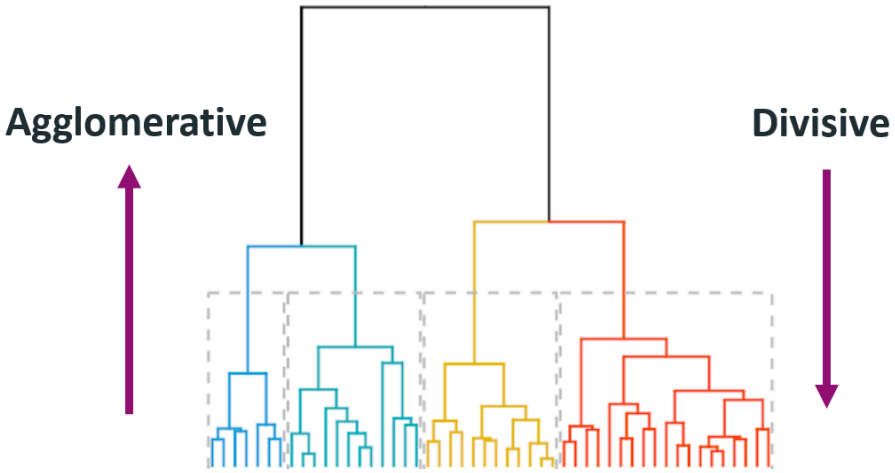



***Figure 2: Hierarchical Clustering Approaches***

#### 4.2.3 Visualization Techniques

**Pairplots:** Displays relationships between variables across clusters
**Silhouette Plots:** Highlights the distribution of silhouette scores for each cluster
**Dendrograms:** Visualizes hierarchical clustering structures with a focus on counties with extreme affordability scores

## 4.3 Predictive Modeling

Predictive modeling was utilized to develop insights into housing affordability scores and migration patterns. This involved regression and classification approaches with each tailored to address different aspects of the problem.

#### 4.3.1 Regression Models

To predict affordability scores, a random forest regressor was implemented. This learning method combines decision trees and aggregates their predictions to reduce overfitting and enhance performance using metrics such as mean absolute error and coefficient of determination.

**Mean Absolute Error (MAE)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

**Coefficient of Determination ($R^2$)**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

### 4.3.2 Classification Models

Migration patterns were predicted using various classification techniques including random forests and decision tree classifiers. These models categorized counties into net inflow, outflow, or neutral migration based on predictor variables.

**Hyperparameter Tuning:** A grid search was conducted to optimize various parameters. Additionally, cross validation was used during this process to maintain robustness
**Confusion Matrix:** Created and analyzed a confusion matrix to evaluate the accuracy and misclassification rates of the models

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

*Used to calculate the accuracy of a classification*

### 4.3.3 Bagging and Boosting

Bagging and boosting are ensemble methods designed to improve the accuracy and robustness of predictive models. These techniques were applied to migration classification to enhance the predictive performance and compare their results against the random forest classifier. These two methods differ in how they train the given data. Each method comes with different advantages and disadvantages and with that comes different results. Below in **Figure 3** displays the difference between bagging and boosting.

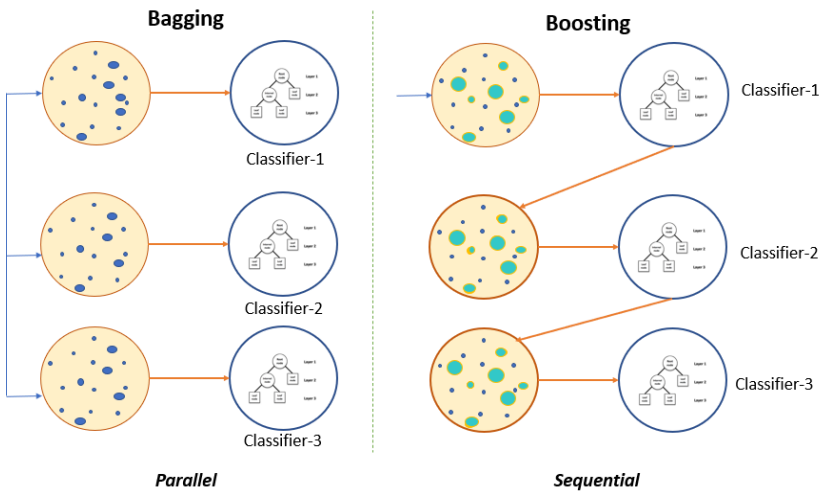



***Figure 3: Difference in Approaches Between Bagging and Boosting***

**Bagging:** Works by training multiple instances of the same base estimator on different bootstrapped subsets of the data. Each model is trained independently and their outputs are aggregated to produce the final prediction. With this method, it allows for reduced variance and overfitting. With the affordability data, the bagging classifier was implemented using 50 estimators with a base decision tree classifier. The performance was assessed using standard classification metrics including an F-1 score.

**Boosting:** With boosting, each model learns from the mistakes of the previous. Instead of training models independently, boosting builds a series of weak learners with each subsequent model focusing on correcting errors made by the previous one. This process typically reduces bias and improves the performance of the weaker models. Using the affordability data, an AdaBoost classifier was implemented using decision trees as the base learner. The number of estimators was set to 50 with adaptive adjustments made to sample weights during training.

### 4.3.4 Feature Selection

Recursive feature elimination was used to identify the most critical predictors for both regression and classification. Using this technique, the model was able to remove the most important features based on model performance. The combination of visualization techniques allow for the analysis to provide actionable insights into housing affordability and migration patterns.

**Feature Importance Plots:** Highlighted variables with the highest predictive contribution
**Predictive vs Actual Plots:** Displayed the alignment between the predicted and actual values for affordability and migration.

## 4.3 Neural Network Analysis

Neural networks were selected for analysis for their ability to capture complex, non-linear relationships between multiple variables. The experiment integrates neural network modeling to extract actionable insights from highly dimensional data. These neural network methods build upon the exploratory data analysis and advanced modeling techniques to ensure accurate predictions.

### 4.3.1 Neural Networks for Predictive Modeling

Neural networks were used to predict housing affordability scores and classify migration patterns. This methodology used multi-layer perceptrons for both regression and classification tasks due to their ability to model complex and non-linear relationships. For the regression, the target variable was based on the affordability score. The architecture of the neural network consisted of 2 hidden layers with 64 and 32 neurons, ReLU activation, and an $L_2$ regularization penalty. The model was trained using the Adam optimizer with a learning rate of 0.00005. **Figure 4** displays the general architecture of a neural network.

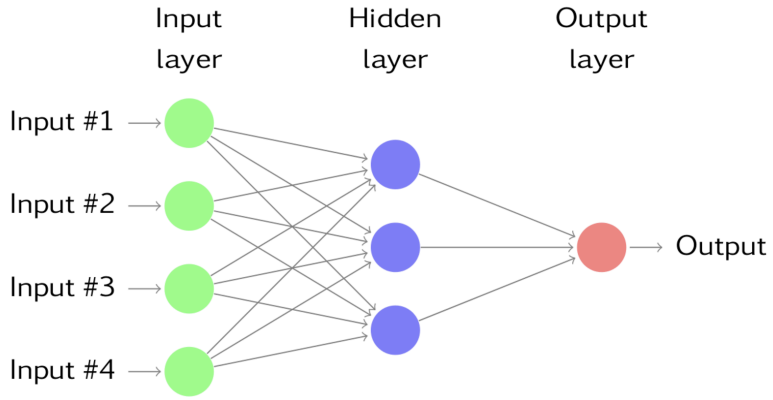



**Figure 4: Simple Artificial Neural Network (ANN) Architecture**

#### 4.3.2 Cross Validation

To ensure model robustness, k-fold cross validation was applied to both regression and classification methods. Regression results were validated using mean absolute error across folds and classification results were validated using average accuracy. Additionally, partial dependence plots were used to show how a single feature or a pair of features affect the predicted outcome.

**Partial Dependence Plot (PDP)**

$$f(x_s) = \frac{1}{n}\sum_{i=1}^{n} f(x_s, x_{-s}^{(i)})$$

*Where $x_s$ represents the feature of interest and $x_{-s}$ represent additional features*

# 5 Experiments

In this project, three experiments were conducted to explore the relationships between housing affordability and migration patterns across US counties. The first experiment focuses on clustering techniques to identify patterns among counties based on affordability related variables. The second experiment performs predictive modeling to develop an affordability score and evaluate how various factors influence this score. Lastly, the third experiment utilizes neural networks to refine predictions of affordability and migration patterns. These experiments aim to offer a comprehensive analysis of the relationships between affordability and migration patterns.

## 5.1 Experiment 1: Data Clustering and Pattern Analysis

### 5.1.1 Clustering with K-Means

The K-means clustering algorithm was applied to identify patterns in the affordability dataset by grouping counties into clustering based on their affordability. Using the elbow method (**Figure 5**), the optimal number of clusters was determined to be three. The within cluster sum of squares (WCSS) reveals a distinct “elbow” in the graph at three clusters, indicating that this is the point where adding more clusters will result in diminishing returns when explaining data variability.

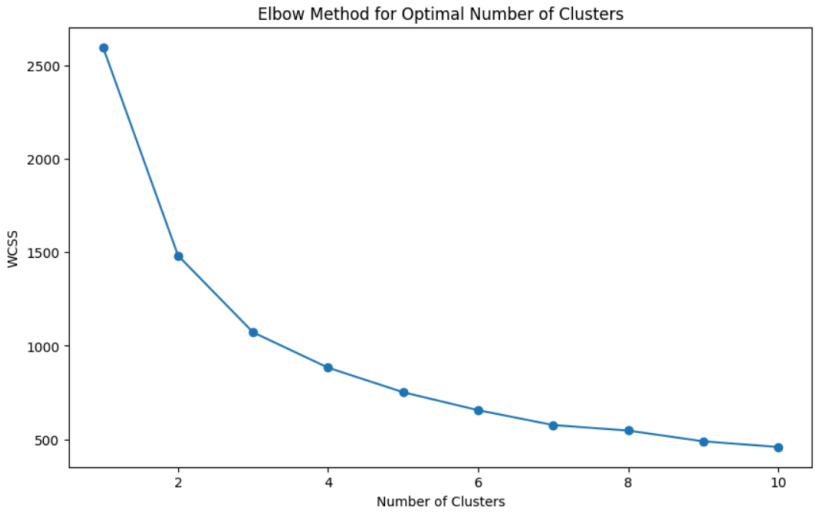



***Figure 5: Elbow Method Plot Identifying Optimal Cluster Amount***

Cluster assignments were then graphed and visualized, using the optimal cluster amount. The pairplot (**Figure 6**) displays how the clusters were distributed across various key features including housing costs, living costs, and income levels. In **Figure 7**, a scatter plot between HousingCostAvg and AffordabilityScore variables shows the separation of clusters and provides insights into affordability trends.

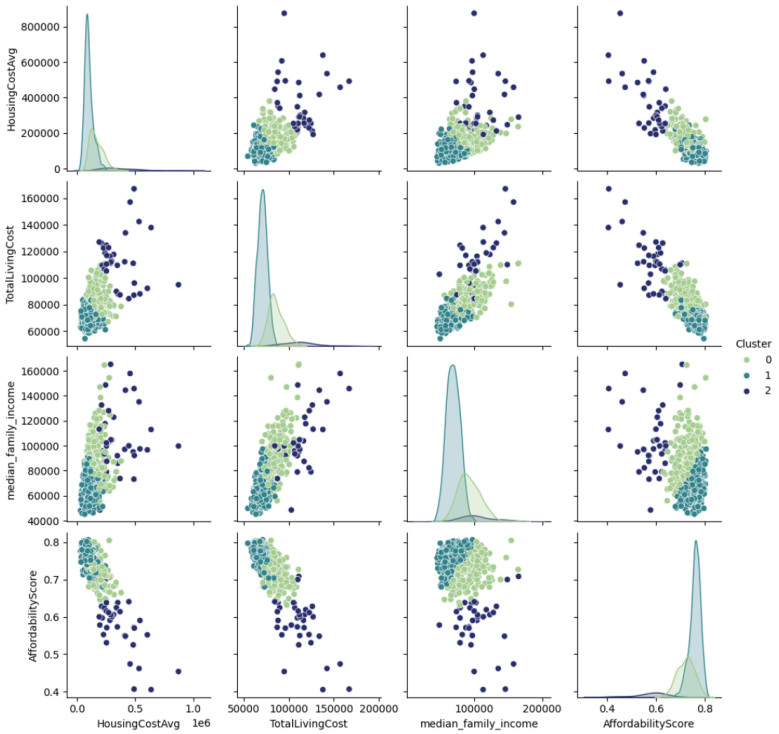



***Figure 6: Pairplot of Cluster Distribution***

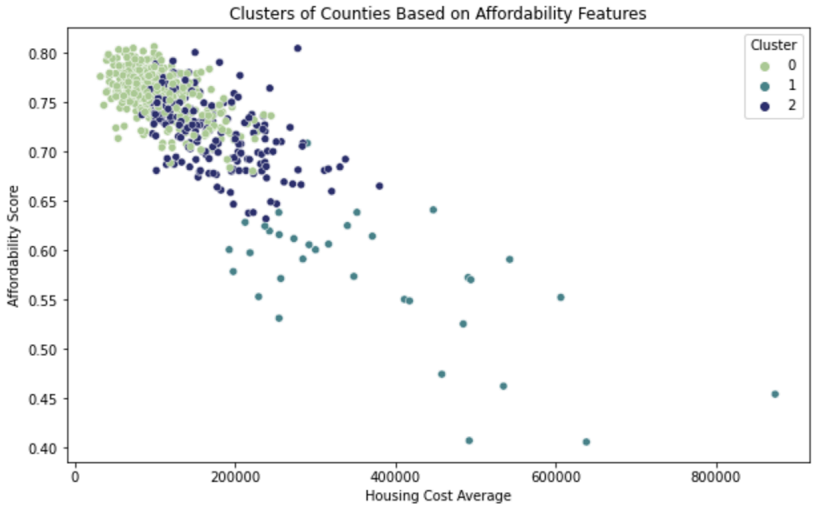



***Figure 7: Scatterplot Displaying Separation of Clusters***

### 5.1.2 Silhouette Analysis for Cluster Validation

To assess the quality of the clusterings, silhouette analysis was performed. The average silhouette score across the three clusters was 0.41, indicating that the clusters displayed moderate separation (**Figure 8**). The silhouette plot provides crucial insight into the separation and connection between each of the clusters. Cluster 1 exhibited the highest average silhouette score indicating the presence of well defined boundaries. Clusters 0 and 2 display greater variability, representing overlaps between certain counties.

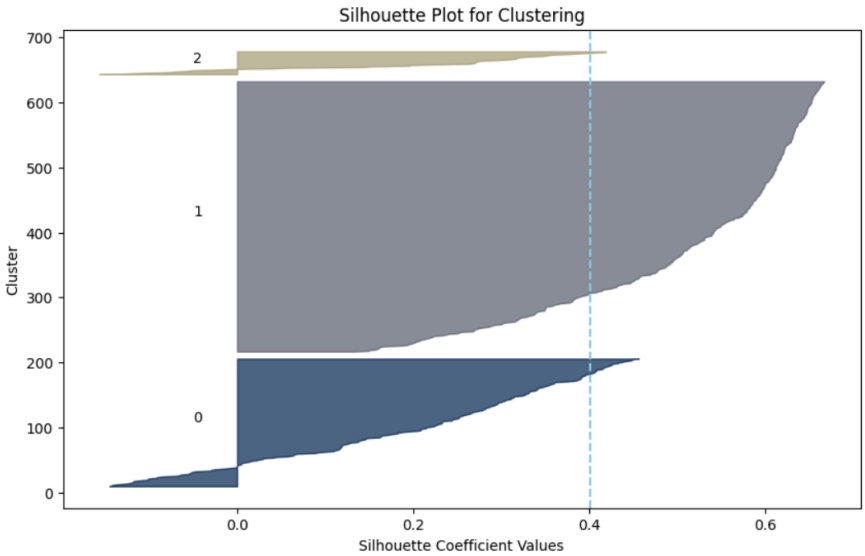



***Figure 8: Silhouette Plot Displaying Quality of Cluster Assignment***

### 5.1.3 Hierarchical Clustering

To validate the k-means clustering results and to explore alternative clustering methods, hierarchical clustering was performed using Ward's method. Two dendrograms were created to display the top and bottom ten counties based on affordability scores. The hierarchical clustering results aligned with the k-means results, further confirming the robustness of identified patterns.


***Figure 9: Dendrogram of Top Counties Based on Affordability Scores***

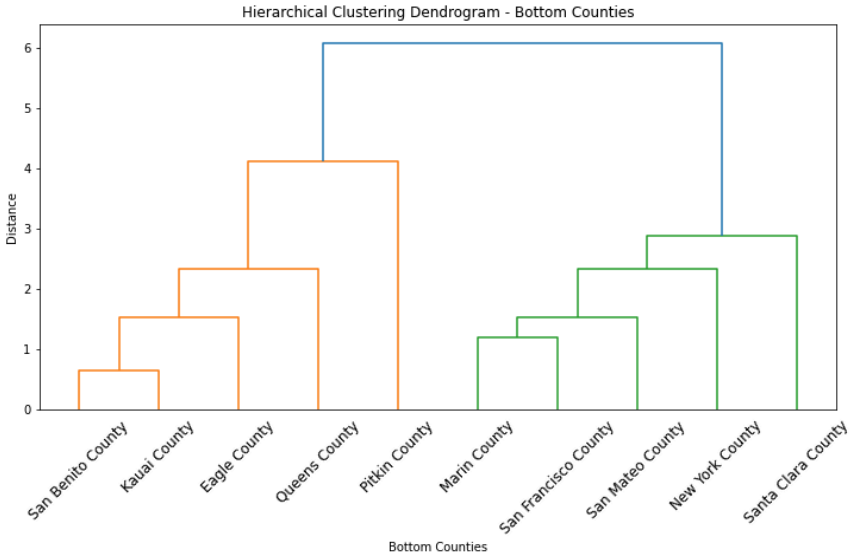



***Figure 10: Dendrogram of Bottom Counties Based on Affordability Scores***

The dendrogram for top counties (**Figure 9**) displays well defined groupings with counties being merged repeatedly based on their similarities in affordability metrics. Similarly, the dendrogram for bottom counties (**Figure 10**) displays the hierarchical structure among less affordable regions and highlights the connections between counties with high housing costs and low affordability scores.

### 5.1.4 Classification of Cluster Memberships

After refining the clustering model and finalizing cluster assignments to each county, a random forest classifier was trained on the affordability variables. This was done to analyze the predictability of cluster memberships.

Classification Report for Predicting Clusters:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 127 |
| 1 | 0.67 | 1.00 | 0.80 | 6 |
| 2 | 1.00 | 0.90 | 0.95 | 62 |
| accuracy | | | 0.97 | 195 |
| macro avg | 0.88 | 0.97 | 0.91 | 195 |
| weighted avg | 0.97 | 0.97 | 0.97 | 195 |

***Table 1: Classification Report for Predicting Clusters***

The classification report shows that in nearly all sectors, the model did well. The classification model achieved an impressive accuracy of 97% as shown in the classification report in **Table 1**. The model's satisfactory ability to predict cluster assignments emphasizes the separability of each of the clusters.

The feature importance analysis (**Table 2**) reveals each of the variables that contribute the most to the classification. TotalLivingCost was identified as the most critical variable with an importance of about 40%. The ranking shown below indicates that living costs and income levels play a vital role in distinguishing different counties, reinforcing the relevance of these particular factors in the affordability analysis.

Feature Importances:

| | Feature | Importance |
|---|---|---|
| 1 | TotalLivingCost | 0.402611 |
| 2 | median_family_income | 0.224901 |
| 3 | AffordabilityScore | 0.198250 |
| 0 | HousingCostAvg | 0.174238 |

***Table 2: Feature Importance Analysis***

#### 5.1.5 Summary of Findings

The clustering analysis was able to successfully identify meaningful patterns in the dataset , grouping the counties into distinct clusters based on the affordability metrics. From this, validation through silhouette analysis and hierarchical clustering confirmed the robustness of k-means clusters and the classification model demonstrated the predictive power of the key variables. These findings emphasized the vital importance of housing costs, living costs, and income levels in forming affordability patterns in counties across the United States.

This detailed analysis on data clustering provides a solid foundation for understanding the patterns of affordability trends. The integration of clustering, validation, and classification methods ensure a comprehensive evaluation of the dataset, highlighting the relationships between key variables and their impact on affordability.

## 5.2 Experiment 2: Predictive Modeling for Trend Analysis

#### 5.2.1 Regression for Migration Metrics

Expanding on the exploration of deterministic features, the regression model aimed to predict migration metrics including inflow, outflow, and net migration. The results for inflow prediction (**Table 3, Figure 11**) demonstrated a strong model performance with results suggesting near perfect predictions. Outflow prediction (**Table 3, Figure 12**) exhibited equally impressive results, reaffirming the model's robustness. For net migration (**Table 3, Figure 13**), the model showed a high accuracy and almost as impressive $R^2$ score. The consistently high $R^2$ scores help confirm the model's capability at capturing underlying relationships within the data.

```
Results for Inflow:
Mean Absolute Error: 564.99
R² Score: 1.00
Results for Outflow:
Mean Absolute Error: 504.46
R² Score: 1.00
Results for Net Migration:
Mean Absolute Error: 264.09
R² Score: 0.99
```

***Table 3: Results for Inflow (Top), Outflow (Middle), and Net Migration (Bottom) Regression***

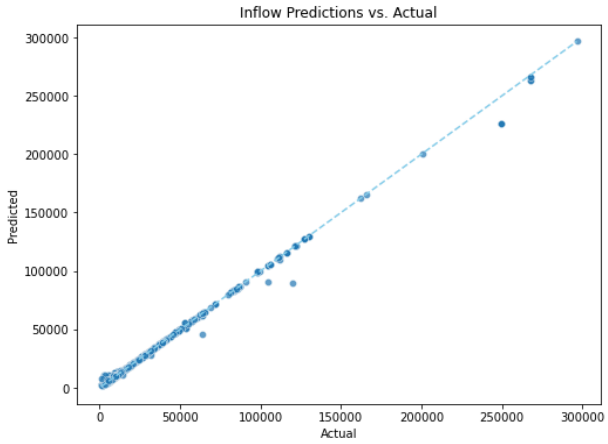



***Figure 11: Inflow Predictions vs Actual***

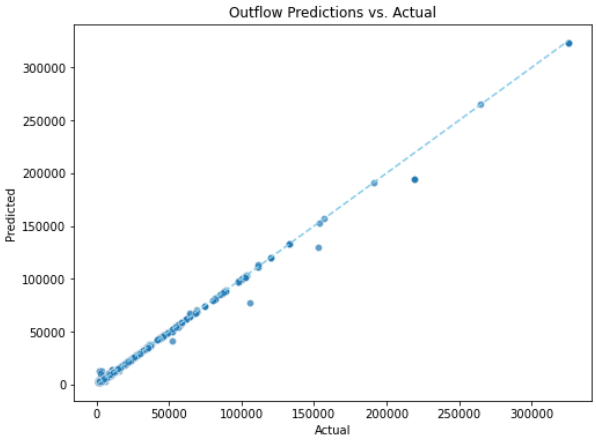



***Figure 12: Outflow Predictions vs Actual***

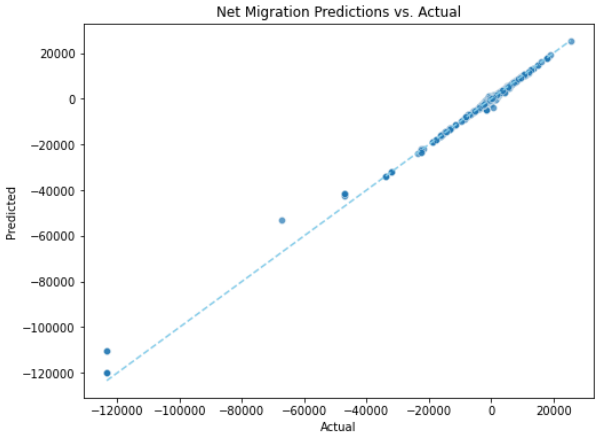



***Figure 13: Net Migration Predictions vs Actual***

Feature importance plots (**Figures 14, 15, and 16**) for each of these models revealed that TotalPop, Employed, and LivingCost variables were the most significant predictors for inflow, outflow, and net migration, respectively. These findings highlighted the crucial role of population size and economic stability in influencing migration trends.

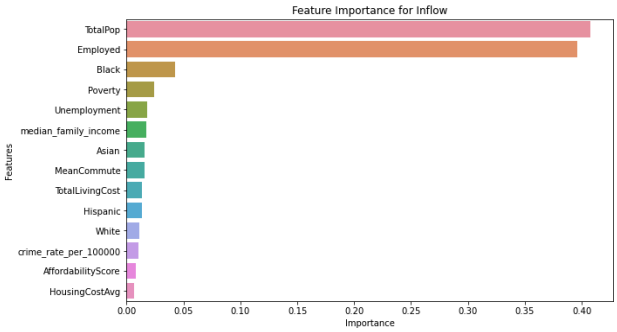



***Figure 14: Feature Importance Plot (Inflow)***

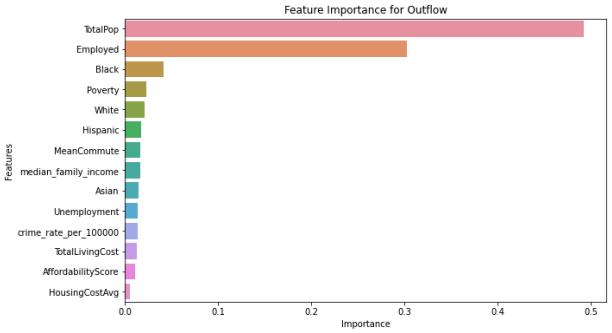



***Figure 15: Feature Importance Plot (Outflow)***

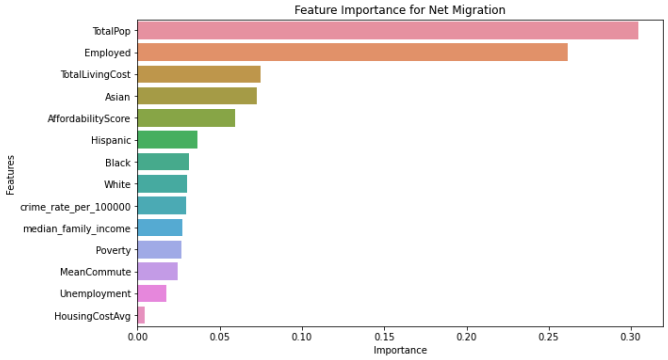



***Figure 16: Feature Importance Plot (Net Migration)***

### 5.2.2 Classification for Migration Patterns

The classification model aimed to categorize counties into two migration classes: Net Gain and Net Loss, using tuner random forest classifiers. Using hyperparameter tuning through grid search identified optimal parameters (Max Depth: None, Min Samples Split: 5, n-Estimators: 200) which resulted in a grid search accuracy of about 74.22%. This sufficiently high accuracy indicates that the model effectively generalized migration patterns.

The final tuned random forest classification model achieved an accuracy of 77%  on the test dataset, shown in **Table 4**, displaying a strong performance when identifying Net Gain counties with a precision of 0.75 and a recall of 0.88. However, Net Loss was comparatively less robust with a precision of 0.80 and a recall of 0.61,  indicating there were potential challenges with class imbalance.

Classification Report for Tuned Random Forest – Migration Class:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Net Gain | 0.76 | 0.88 | 0.81 | 112 |
| Net Loss | 0.80 | 0.61 | 0.69 | 83 |
| accuracy | | | 0.77 | 195 |
| macro avg | 0.78 | 0.75 | 0.75 | 195 |
| weighted avg | 0.77 | 0.77 | 0.76 | 195 |

***Table 4: Classification Report for Tuned Random Forest***

The confusion matrix (**Figure 17**) displays the model's performance, showing a higher rate of misclassifications for counties with Net Loss, where 32 instances were incorrectly categorized at Net Gain.

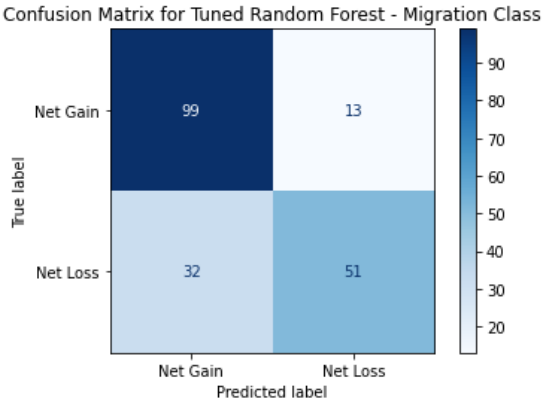



***Figure 17: Confusion Matrix for Tuned Random Forest***

Alternative classification techniques including bagging and boosting were performed for comparison. Bagging achieved an accuracy of 79% (**Table 5**), slightly outperforming the tuned random forest model while boosting had a lower average of 69% (**Table 6**). Both bagging and boosting methods exhibited similar trends with Net Gain predictions consistently outperforming the predictions for Net Loss.

Bagging – Migration Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Net Gain | 0.79 | 0.86 | 0.82 | 112 |
| Net Loss | 0.78 | 0.70 | 0.74 | 83 |
| accuracy | | | 0.79 | 195 |
| macro avg | 0.79 | 0.78 | 0.78 | 195 |
| weighted avg | 0.79 | 0.79 | 0.79 | 195 |

***Table 5: Classification Report for Bagging***

Boosting - Migration Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Net Gain | 0.72 | 0.76 | 0.74 | 112 |
| Net Loss | 0.65 | 0.60 | 0.63 | 83 |
| accuracy | | | 0.69 | 195 |
| macro avg | 0.68 | 0.68 | 0.68 | 195 |
| weighted avg | 0.69 | 0.69 | 0.69 | 195 |

***Table 6: Classification Report for Boosting***

### 5.2.3 Cross Validation

Cross validation metrics provided a robust validation of the predictive models used in this experiment. For the regression model, cross validation achieved an MAE of 0.098 which indicated that the regression models maintained low error rates across all data folds. This highlights the stability of the models in capturing the relationships between socioeconomic factors and migration metrics. Despite the high $R^2$ values achieved in the test set, the cross validation results suggested that overfitting was unlikely.

For classification, the cross validation accuracy achieved a percentage of 89.2% which exceeded the accuracy observed in the test set predictions (77%). This adequate performance confirms the neural networks ability to generalize well across the data's subsets. The performance in cross validation emphasizes the model's reliability despite challenges in accurately categorizing Net Loss in the test set. The consistency across the folds demonstrates that classification has identified meaningful patterns within the data, suggesting potential for refinement through approaches such as tuning.

Cross validation serves as a crucial measure of the model's ability to generalize across unseen and new data, offering insights in their overall performance and predictive power. The inconsistency between cross validation and test set results, particularly in classification, emphasizes the importance in addressing any class imbalances and refining the model further to achieve decent performance across all categories.

### 5.2.4 Summary of Findings

This experiment highlights the use of predictive modeling in analyzing migration trends and their relationship with various socioeconomic factors. The regression models performed exceptionally for inflow, outflow, and net migrations predictions, displaying low MAE values and high $R^2$ scores. The classification models demonstrated robust cross validation accuracy despite the challenges with imbalanced classes in test predictions. Feature importance and partial dependency plots enhanced the analysis by revealing the critical factors that are influencing migration decisions. These insights open the door for data driven policy interventions to address migration dynamics and regional disparities. Additionally, the evaluation of different classification techniques provided a comprehensive understanding of the strengths and limitations of various predictive approaches.

## 5.3 Experiment 3: Neural Network Models for Predictive Analysis

### 5.3.1 Affordability Regression Analysis

The first phase of this experiment involved using neural networks to predict the affordability score of each county based on socioeconomic and demographic variables. A multi-layer perceptron (MLP) regressor was trained on various variables such as housing costs, total living costs, median family income, population size, and crime rates. The structure of the MLP regressor included two hidden layers with 64 and 32 neurons and activated ReLU activation function to model any non-linear relationships. Additionally, the Adam optimizer was utilized to reduce the loss function with hyperparameters tuned for improved convergence.

When performed, the model showed poor results. Despite achieving a MAE of 0.14, the models coefficient of determination $R^2$ of -14.24 indicated poor generalization. This suggests that the neural network struggled to capture the foundational relationships in the data. However, the cross validation MAE results of 0.098 were able to demonstrate a consistent model performance across the data subsets, suggesting that there was potentially overfitting of the training set.

Through recursive feature elimination (RFE), the regression model identified the variables HousingCostAvg, TotalLivingCost, median_family_income, TotalPop, and crime_rate_per_100000 as the most significant predictors. Collectively, these features highlighted the economic and demographic influences on affordability. Additionally, partial dependency plots (PDP) were used to interpret the model by visualizing the marginal impact of HousingCostAvg and TotalLivingCost on affordability predictions. These plots illustrate that affordability scores and rising housing and living costs are negatively correlated, providing insights into the relationships obtained by the model.

### 5.3.2 Classification for Migration Patterns

The classification analysis aimed to categorize counties into three migration categories: Net Gain, Net Loss, and Neutral, using inflow, outflow, and socioeconomic variables. The classification report for migration predictions (**Table 7**) highlights a mixed model performance. The model achieved an overall accuracy of 59% on the test set and had strong precision and recall for the Net Gain class with a precision of 0.59 and a recall of 0.96. However, the model struggled to identify Net Loss counties  with a precision of 0.67 and a recall of 0.10 which indicates challenges when distinguishing counties with population losses.

Classification Report for Migration Prediction:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Net Gain | 0.59 | 0.96 | 0.73 | 112 |
| Net Loss | 0.67 | 0.10 | 0.17 | 83 |
| accuracy | | | 0.59 | 195 |
| macro avg | 0.63 | 0.53 | 0.45 | 195 |
| weighted avg | 0.62 | 0.59 | 0.49 | 195 |

***Table 7: Classification Report for Migration Prediction***

The confusion matrix (**Figure 18**) further explains the model's performance by describing the true and predicted class distributions for Net Loss and Net Gain. While the Net Gain was predicted correctly consistently, the Net Loss class showed significant misclassifications when comparing the true and predicted labels.

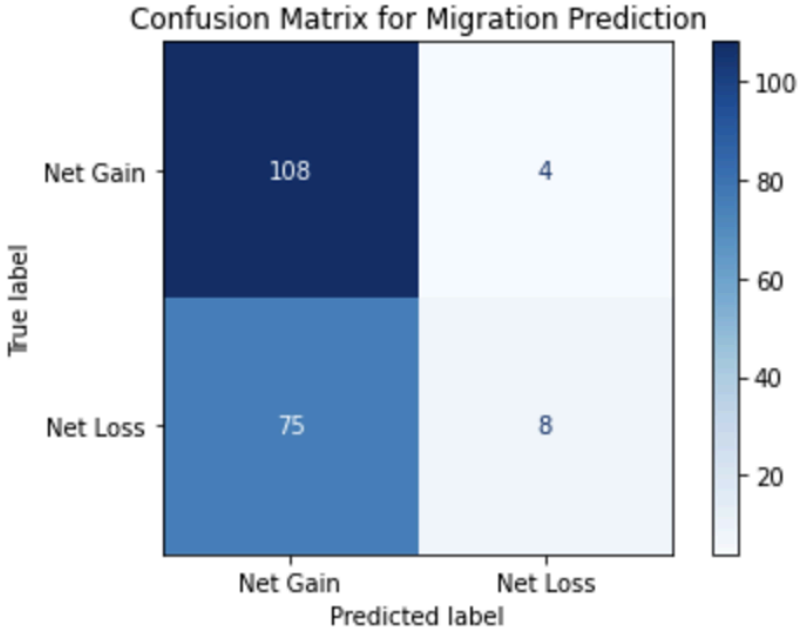



***Figure 18: Confusion Matrix for Migration Prediction***

### 5.3.3 Feature Importance and Insights

The classification model identified INflow, OUTflow, TotalPop, HousingCostAvg, and median_family_income as the most influential predictors of the migration class. These top predictors align with prior findings emphasizing the importance of migration flow and economic stability when determining population shifts. Additionally, feature importance visualization reaffirmed the critical importance of these variables with inflow and outflow contributing the most to the models decision making process.

Partial dependency plots (PDP) for affordability and migration classification models (**Figures 19 and 20**) provided further insights by visualizing how changes in HousingCostAvg and TotalLivingCost influence migration classifications. Moreover, these plots revealed non-linear relationships indicating that with increasing housing cost often correlating with a shift towards Net Loss classifications. These insights are highly valuable for understanding the model's internal decision making and validating its predictions against the known patterns in migration patterns.

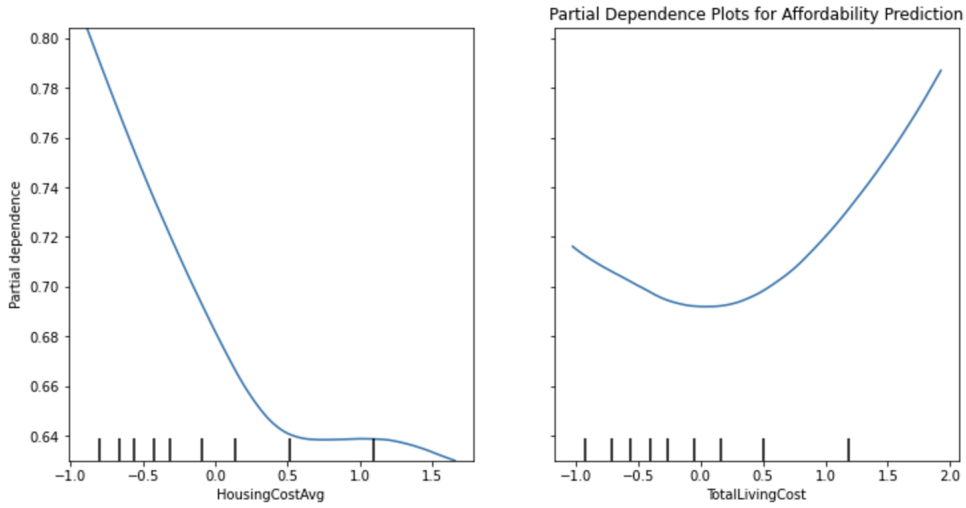



***Figure 19: PDP for Affordability Prediction (Left: HousingCostAvg, Right: TotalLivingCost)***

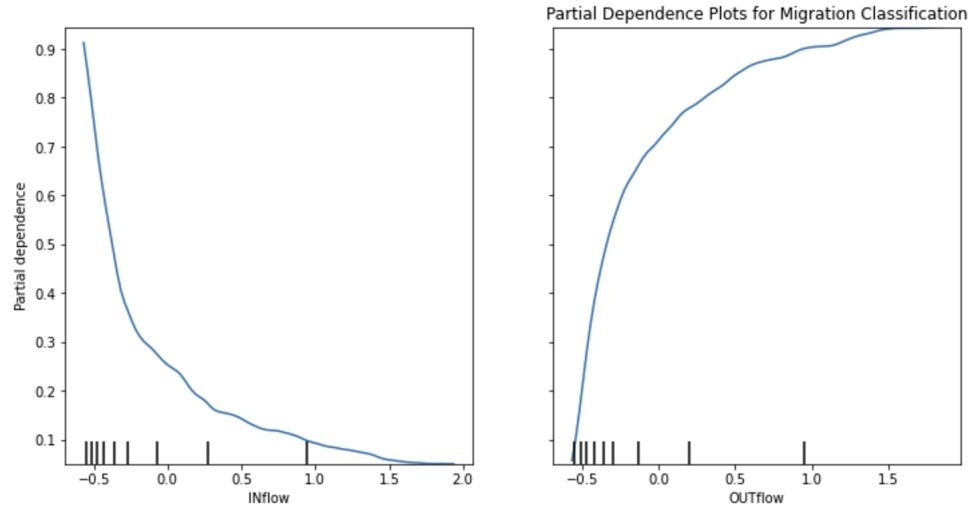



***Figure 20: PDP for Migration Classification (Left: INflow, Right, OUTflow)***

### 5.3.4 Cross Validation and Model Robustness

Cross validation serves as a critical step to ensure the robustness and generalizability of the neural network models. For regression, the cross validation MAE was 0.098, suggesting that the model consistently minimizes the prediction errors across the data folds even though the previously stated $R^2$ of -14.24 indicates issues with generalization. The discrepancy highlights the importance of balancing the cross validation performance with test set evaluation to avoid any overfitting of the data.

For classification, the cross validation accuracy of 89.2% demonstrated strong generalizability which highlighted the model's capability to identify migration patterns across the data. The high cross validation accuracy suggests that the classification model is able to capture the key factors influencing migration trends. However, the discrepancy between cross validation and test set accuracy (59%) raises questions about any data imbalances that could have affected test performance. Cross validation metrics provide a solid quantitative foundation to assess the model's reliability. This method also promotes model tuning by identifying optimal configurations that balance predictive accuracy and generalizability.

### 5.3.5 Summary of Findings

This experiment highlights the utility of neural networks in modeling affordability scores and migration patterns and highlights both the areas of strength and in need of improvement. While the regression model struggled with generalization, it was able to identify key affordability predictors and demonstrated interpretability through partial dependency plots. The classification model displayed signs of class imbalances however, it also showed strong cross validation results and provided actionable insights into migration trends.

The figures included in this experiment complemented the findings by offering sufficient visual evidence of model performance. These results emphasized the potential for neural networks when addressing complex predictive modeling, providing valuable insights for decision making related to housing and migration.

## 5.4 Interactive Tools for Analysis

To enhance the accessibility and usability of the predictive models developed in each experiment, interactive tools were implemented using Gradio. These tools provided a favorable interface for predicting affordability scores and classifying migration patterns, allowing users to engage with the models by providing structured inputs. By incorporating these tools, the project bridges the gap between complex data analysis and user friendly applications. This allows for users to experiment with various predefined scenarios within the affordability data.

### 5.4.1 Interactive Tool Design

These interactive tools were designed with two primary functions: predict affordability scores and classification for migration. Both functions allow users to input feature values either individually or in the form of structured arrays.

For affordability prediction, the regression model is used to analyze relationships between variables such as HousingCostAvg, TotalLivingCost, and median_family_income. With these tools, users can observe how adjustments made to these inputs impact the predicted affordability score. Additionally, the results are displayed in real time, offering immediate feedback.

Migration Classification relies on a random forest classifier which categorizes counties into migration such as Net Gain, Net Loss, and Neutral. The user specifies the input values including INflow, OUTflow, and HousingCostAvg which the tool processes to provide the predicted migration category.

Gradio allows for an efficient method of producing predictions as a user only has to modify the sliders in order to get an output of what their locale's affordability score or migration class may be. **Figure 21 and 22** displays a showcase of the Gradio GUI to allow a range of inputs to be cast and then made into a prediction without knowing any of the framework behind the models, hosted on HuggingFace.

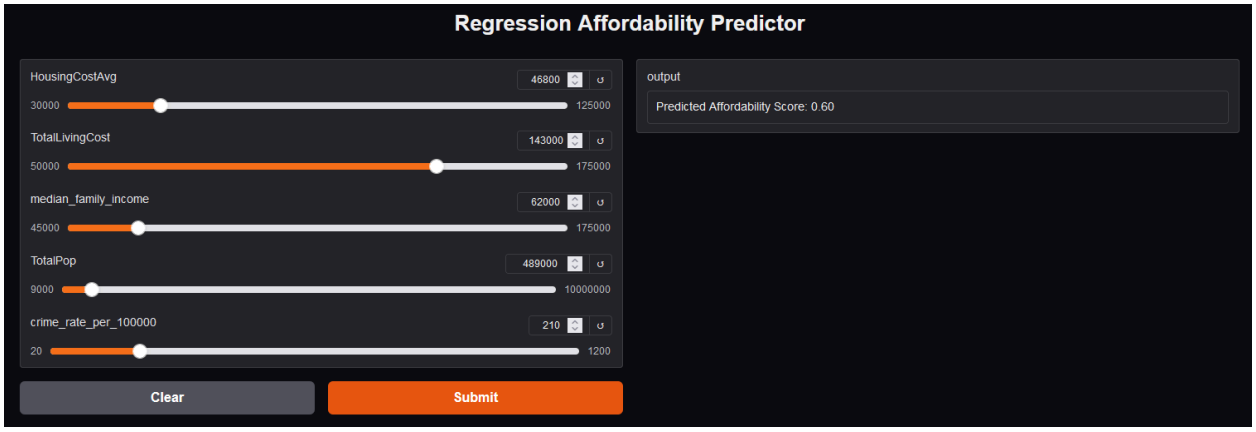



***Figure 21: Example of Affordability Prediction Using Gradio***

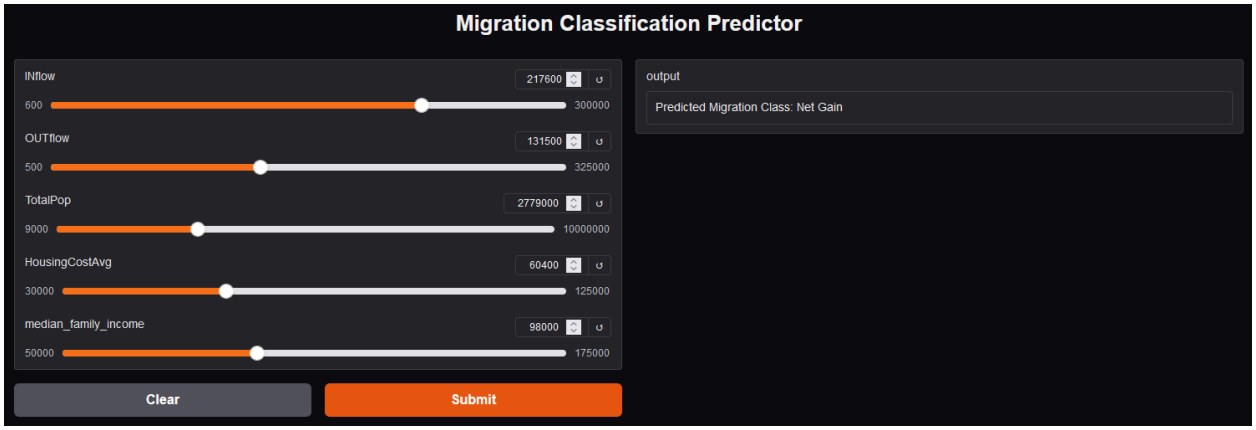



***Figure 22: Example of Affordability Prediction Using Gradio***

The interactive tools produced the following outputs during testing:

**Affordability Prediction:**

With the hypothetical parameters provided above, the predicted affordability score was 0.6, demonstrating the model's ability to generate user friendly and interpretable results, based on user provided data.

**Migration Classification:**

The classifier correctly categorized an example test scenario as Net Gain, emphasizing the robustness of the trained random forest model in analyzing migration trends.

### 5.4.3 Conclusion and Insights

The inclusion of interactive tools added significant value to the project by facilitating user engagement with the predictive models. Affordability prediction and migration classification tools offered accurate outputs while the feature importance visualization provided transparency into the visualization process. By simplifying access to machine learning models, each tool displayed their potential to guide decision making related to affordability and migration in various areas. Additionally, they allowed users to explore predictive trends interactively, transforming complex analyses into actionable insights.

# 6 Summary and Concluding Remarks

## 6.1 Overview of Findings

This project effectively combined various machine learning techniques to explore the intricate relationships between housing affordability and migration patterns across US counties. Three experiments were conducted to address distinct but also interconnected aspects of the data. The first experiment used clustering techniques to group the counties by affordability metrics, providing interpretable insights into affordability trends. The second experiment used predictive modeling to predict migration trends, revealing key socioeconomic factors that guide population shifts. The third experiment incorporated neural networks to refine affordability predictions and migration classifications, emphasizing their use in capturing complex and non-linear relationships. Additionally, the use of interactive tools with Gradio made the analysis more accessible and user friendly, allowing users to engage with the affordability and migration predictive models.

## 6.2 Key Implications

The findings of this study provide actionable insights for a variety of individuals and communities at large. Individuals and families can use these findings to determine where to live depending on their own socioeconomic and demographic situations. Policymakers can utilize the results to prioritize affordable housing initiatives in the counties with lower affordability scores. Urban planners can use the migration classifications to identify regions at risk of population decreases and allocate specific resources accordingly. In all, the results of this public can help a variety of individuals and stakeholders to make informed decisions on topics relating to affordability and migration.

Furthermore, the identified significant predictors such as HousingCostAvg and TotalLivingCost highlight the areas where interventions could most effectively influence affordability and migration patterns. Additionally, the integration of interactive tools further strengthens the accessibility of these models allowing individuals to test hypothetical scenarios and make data informed decisions.

### 6.3 Limitations and Future Directions

Despite the robustness of the methodologies used, this study does have some limitations. The datasets used to create the main affordability dataset left some counties unrepresented. This could potentially skew and impact results. Additionally, challenges in neural network models including discrepancies between test and cross validation metrics highlight the need for further model refinement. Addressing class imbalances in classification models and exploring alternative model architectures could enhance the predictive accuracy. Expanding the dataset to include in-depth temporal data could also provide a more dynamic view of affordability and migration patterns over time.

Future research could integrate additional variables to deepen the analysis. Additionally, expanding interactive tools to include geospatial visualizations or scenario planning capabilities would also further enhance their utilization in decision making.

### 6.4 Concluding Remarks

This project demonstrated the power of combining machine learning models with interactive tools to address critical societal challenges. By analyzing data on housing affordability and migration patterns, the project not only advances our understanding of these complex features and how they relate, but also provides practical tools for users to make informed decisions. The methodologies and insights presented lay a solid foundation for continued exploration and refinement. Through both innovation and collaboration, various approaches can contribute to more equitable and sustainable communities.

Furthermore, the integration of interactive tools bridges the gap between technical analysis and usability, enhancing the practicality and impact of the findings. These tools not only allow users to visualize data trends, but also provide the flexibility for exploration of different scenarios and tailor insights to various users. This fusion of data driven insights with user accessibility ensures that the knowledge generated is both actionable and inclusive.

# 7 References and Appendix

## 7.1 References

1. Asaniczka. “US Cost of Living Dataset (1877 Counties).” *Kaggle*, 17 Feb. 2024, www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties.
2. Johnson. “United States crime rates by county” *Kaggle,* 2016, www.kaggle.com/datasets/mikejohnsonjr/united-states-crime-rates-by-county/data
3. MuonNeutrino. “US Census Demographic Data.” *Kaggle*, 3 Mar. 2019, www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv.
4. U.S. Census Bureau. “County-to-County Migration Flows: 2016-2020 ACS”, 2020, www.census.gov/topics/population/migration/guidance/county-to-county-migration-flows.html
5. “Housing Data.” *Zillow*, 11 Nov. 2024, www.zillow.com/research/data/.

## 7.2 Appendix

A. Software and Tools

The following tools and software were used for analysis and modeling

1. **Python:** The programming language used for analysis and modeling
2. **Jupyter Notebooks:** Used locally for executing Python scripts
3. **Google Colab:** A cloud based environment for executing Python scripts
4. **Scikit-learn:** Used for implementing various machine learning models
5. **Matplotlib and Seaborn:** Libraries used for creating visualizations such as scatterplots, bar graphs, and heatmaps
6. **Gradio:** An interactive tool that facilitates the creation of an interface for testing model predictions and visualizations
7. **Hugging Face**: A platform to upload, share, and demonstrate Python-based machine learning models
8. **Pandas and NumPy:** Libraries used for data manipulation and numerical computations
9. **StandardScaler (from Scikit-learn):** Applied to normalize data prior to modeling

B. Experiment Scripts

This project was composed of several Python scripts, each corresponding to different experiments and tasks. The full Python scripts can be found in the GitHub repository linked on the Title Page.

1. **DataPreprocessing.ipynb:** Merges and reshapes the data and creates an affordability score using appropriate variables
2. **ExploratoryDataAnalysis.ipynb:** Contains exploratory data analysis including histograms, correlation matrices, and scatter plots for a deeper understanding of data distributions and relationships
3. **Experiment1_DataClustering.ipynb:** Implements K-means and hierarchical clustering to analyze patterns in affordability data
4. **Experiment2_PredictiveModeling.ipynb:** Focus on regression models and classification models using random forests classifiers and related machine learning techniques
5. **Experiment3_NeuralNetworkModels.ipynb:** Deploys neural networks for predictive modeling, emphasizing affordability and migration trends
6. **InteractiveTools.ipynb:** Integrates interactive tools such as Gradio for dynamic testing of predictive models

C. Data

The affordability.csv dataset forms the foundation of this project. It contains columns relating to housing costs, migration patterns, and demographics for US counties. Key variables include:

- **AffordabiltyScore:** The calculated metric summarizing affordability
- **HousingCostAvg, TotalLivingCost, median_family_income:** Economic indicators
- **INflow, OUTflow, NET in:** Migration statistics reflecting population movements

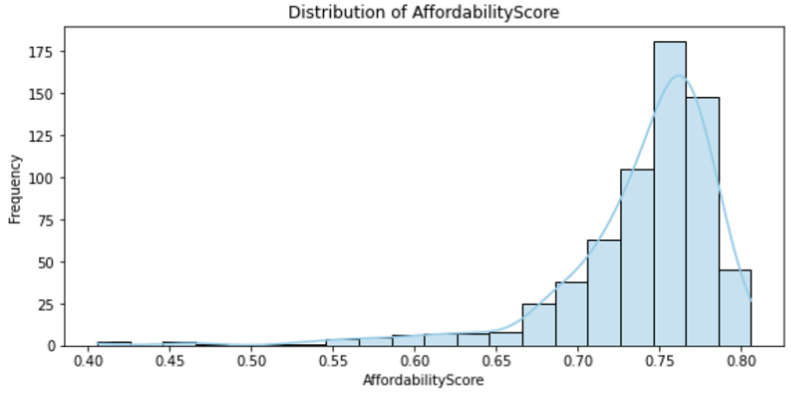
D. Demo

The files needed by the Hugging Face platform in order to build and run the model.

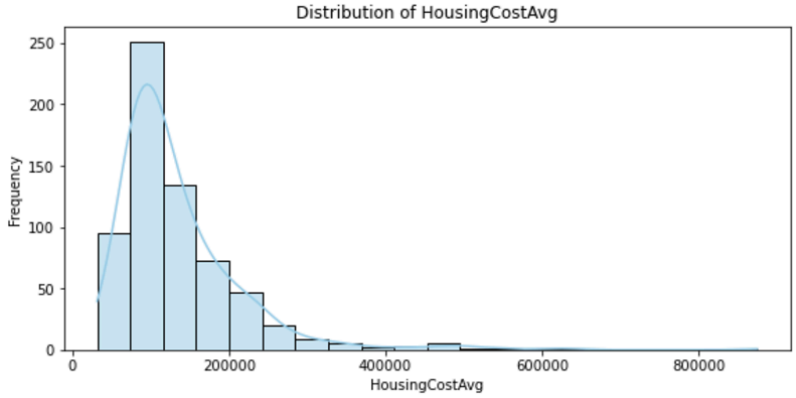
- **app.py:** The source code for the model to build
- **requirements.txt:** A list of libraries Hugging Face must install to run the model

E. Visualizations and Additional Figures and Tables

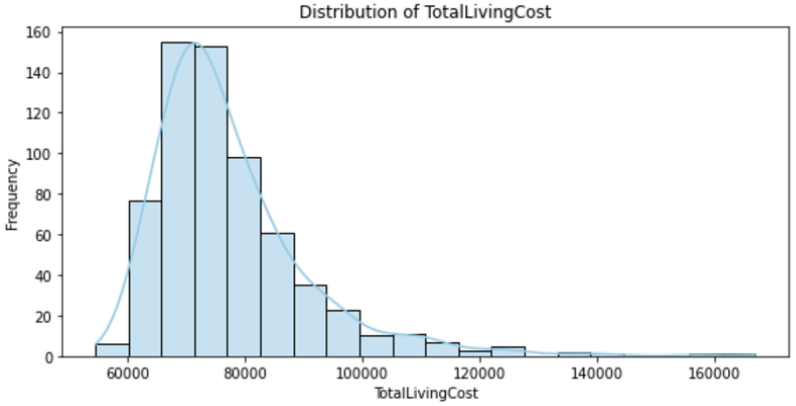
A variety of figures were used to support the findings and enhance the interoperability. These visualizations are referenced throughout the report and are available in the Python scripts. Additional figures and tables from the exploratory data analysis are displayed below for better statistical interpretation of the variables.



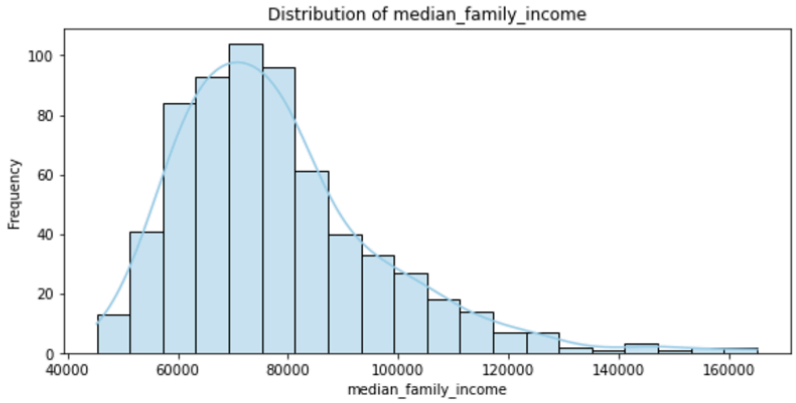

1. ***Figure 23: Distribution of AffordabilityScore (Appendix)***



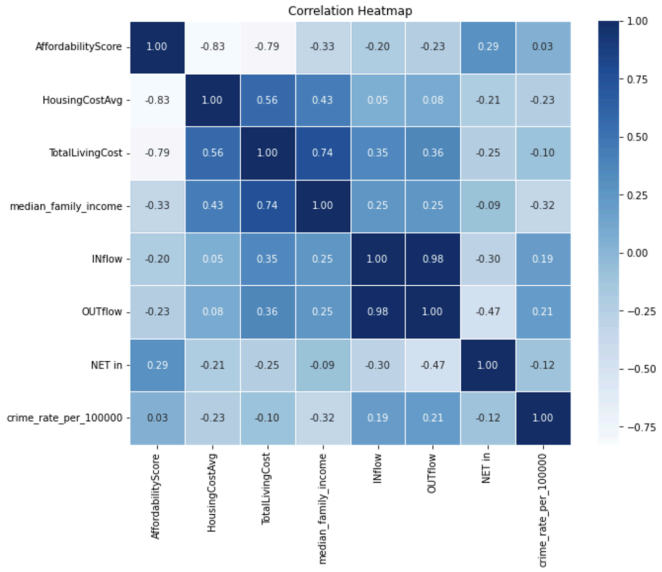

2. ***Figure 24: Distribution of HousingCostAvg (Appendix)***



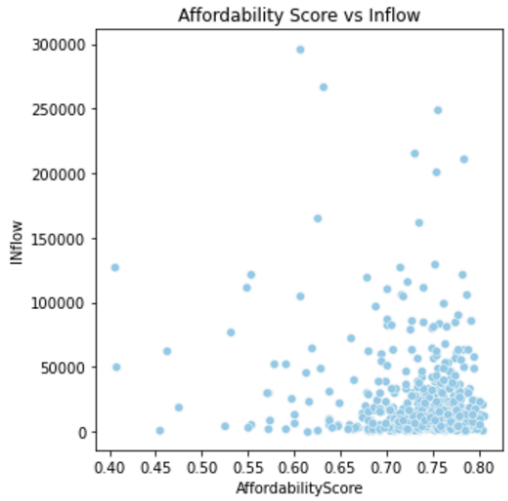

3. ***Figure 25: Distribution of TotalLivingCost (Appendix)***



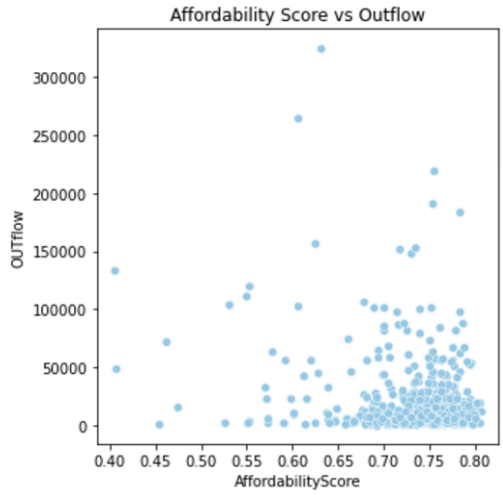

4. ***Figure 26: Distribution of median_family_income (Appendix)***



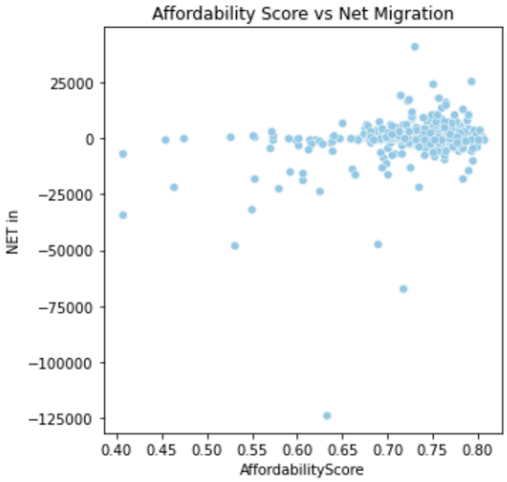

5. ***Figure 27: Correlation Heat Map for Key Variables (Appendix)***



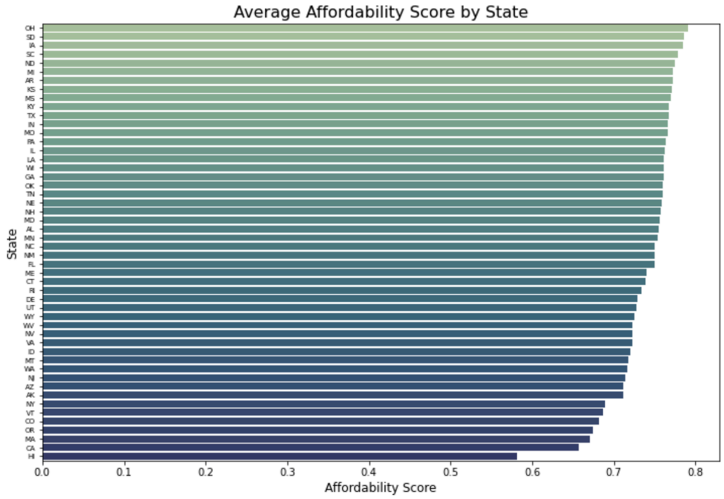

6. ***Figure 28: Relationship Between Affordability Score and Inflow (Appendix)***




7. ***Figure 29: Relationship Between Affordability Score and Outflow (Appendix)***




8. ***Figure 30: Relationship Between Affordability Score and Net In (Appendix)***




9. ***Figure 31: Average Affordability Score by State (Appendix)***

Top 10 Most Affordable Counties:

| | RegionName | StateName | AffordabilityScore |
|---|---|---|---|
| 2138 | Medina County | OH | 0.806156 |
| 1938 | Linn County | IA | 0.804568 |
| 1986 | Los Alamos County | NM | 0.804386 |
| 224 | Barton County | KS | 0.803034 |
| 1976 | Lorain County | OH | 0.802483 |
| 3738 | Wood County | OH | 0.801509 |
| 3711 | Williams County | ND | 0.800294 |
| 20 | Aiken County | SC | 0.800075 |
| 2648 | Pennington County | MN | 0.799958 |
| 3444 | Trumbull County | OH | 0.798211 |

***10. Figure 32: Top 10 Most Affordable Counties (Appendix)***

Top 10 Least Affordable Counties:

| | RegionName | StateName | AffordabilityScore |
|---|---|---|---|
| 2409 | New York County | NY | 0.405689 |
| 3049 | San Mateo County | CA | 0.407055 |
| 2711 | Pitkin County | CO | 0.454124 |
| 3018 | San Francisco County | CA | 0.462243 |
| 2087 | Marin County | CA | 0.474364 |
| 1695 | Kauai County | HI | 0.525276 |
| 2808 | Queens County | NY | 0.531052 |
| 3091 | Santa Clara County | CA | 0.548653 |
| 2995 | San Benito County | CA | 0.550185 |
| 1055 | Eagle County | CO | 0.552165 |

***11. Figure 33: Bottom 10 Least Affordable Counties (Appendix)***