# Migration Season: The Top 100 Most Populous Counties and Their Hidden Gems

Kerstin Fontus, Katha Korgaonkar, Joel Westmark
Group 144

## Problem Statement:

There have been significant shifts in the US population with the rise of remote work and increases in the cost of living. Before 2020, migration rates had been trending downwards, even in the 25-34 age range which historically made up the largest share of movers (1). After the 2020 Census, New York lost a congressional seat by a very slim margin, and was joined by California who had yet to lose one in the history of its statehood. Texas, Florida, and Colorado all gained seats, indicating a shifting preference away from historical hotspots and towards the West and the South. This is not to say that states like New York and California are not still popular destinations, but their slow decline in population indicates a change in migration patterns.

There have been many studies and surveys to examine the factors underlying these trends. We are interested in finding what we consider “hidden gems” among the various destinations for domestic migrants. There will always be people who want to move to population centers such as New York City, Chicago, Los Angeles, etc. We are looking to uncover which American counties have quietly become migration hubs amongst the top 100 most populous American counties. They may not have the same dramatic increases in population as the three aforementioned centers, but nevertheless have significant amounts of newcomers from all over the country, while potentially having a low attrition rate. We finally will discover “sister hidden gems” by grouping our original nominated gems into clusters based upon their features.

Through various methods and data sources, we hope to create a methodology to analyze and recommend counties that have enjoyed an influx of new residents, uncover the features that have influenced this change, and nominate “sister gems" that have similar characteristics that may have an increase in population soon. We will also try to provide insight into population growth in general.

## Data Source:

The datasets needed to complete our analysis include:

- County to County migration (2): The original plan was to study the migration rate between cities, but the United States Census Bureau has a complete dataset documenting county to county migration from 2016 to 2020. From there we can extrapolate what cities or metropolitan areas have become the most attractive for domestic migrants. After restructuring, cleaning and transforming the data, we have the bidirectional, net, and gross migration estimates between every US county that people moved between in this time frame.
- County metadata (3): After studying domestic migration and identifying the counties of interest, we also want to see if we can identify the features that make these areas “hidden gems”. Demographic features such as total population, unemployment, sex

ratios, housing price index (or housing price), etc. may allow us to have a more explicit methodology in recommending destinations. We had some significant issues obtaining this data, as the datapull tool on the census.gov quick facts website broke after we chose this topic. So we manually pulled the top 100 counties in terms of population.

## Methodology:

We decided to approach county migration by interpreting population movement as a weighted graph. First, we used spectral clustering to find unique migratory clusters. Specifically, we tried to find small, single county clusters in the 100 counties we examined. We subsetted these counties to find only counties with positive growth. This step was prone to how many eigenvalues we used and how many clusters we wanted to examine.

After identifying small migratory clusters or “hidden gems”, we used classification algorithms to identify the defining traits of these small clusters and to attempt to find a method to predict the gem or non-gem status of a county. Due to the overabundance of non-gem counties, we evaluated our results with F-1 score to better evaluate the ability of the models to find true positives.

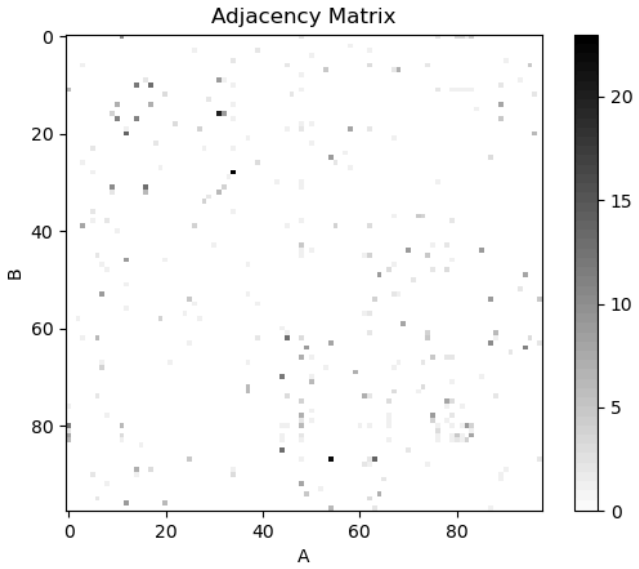
In order to relate these unique counties with other counties, we performed K-Means clustering of the 100 counties using the county metadata, which included demographics, resources, etc. We then identified the K-Means clusters that contained counties in our unique Spectral clusters. We define these special K-Means clusters as “sister hidden gems”. Meaning their migration patterns do not stand out, but they share the same features as the unique Spectral cluster counties.

Finally, we performed Elastic Net regression to identify causes for a high population percent increase using county metadata. This helps us learn the “why” of population increases and can add context to our sister hidden gem analysis. We also used a Markov Chain matrix of the top 100 counties to predict the future population (assuming births = deaths). Theoretically, we would reach a steady state population under regular Markov Chain assumptions, but we were only interested in near term projections, since long term projections using a simple markov chain would be very inaccurate. Using all of the methods above (some more attainable given the data than others), we reached some interesting conclusions.

### Spectral Clustering:

The migration dataset is composed of pairs of counties that had any exchange of residents. Each data point has a feature with the estimates of the number of people who moved from county A to county B, and vice versa. We also have the 2020 population for each county, which we used to calculate the percentage of the population of a given county to move to its new destination. There are 3,143 counties in the United States, and thus there are almost 5,000,000 ways to create a duo and look at the migration between them. We are only looking at the movements between the 100 most populous counties as a way to test our methodology.

After scaling the new population variables, we constructed an adjacency matrix $A=\{a_{ij}\}$ where entry *(i,j)* is the percentage of the population of county *i* to move to county *j*. This resulted in a very sparse matrix that had 100 distinct counties.




After constructing the degree matrix and the Laplacian, we found its 31 largest eigenvectors and grouped the rows of the resulting matrix into 31 clusters. This resulted in 30 clusters that each had a single county and one cluster with the remaining 68. The 30 counties with their own cluster were designated as potential hidden gems, as they were counties that had intake and output migration patterns that were distinct from all the others. We aggregated the migration dataset to find the gross and net migration as a percentage of its total population for each county. The 14 of the potential hidden gems that had a positive net migration (more people moved in than out of the county) were identified as the final hidden gems. These counties, their largest cities, and the metropolitan area they belong to are:

| **County** | **Largest City; Metropolitan area** |
|---|---|
| Broward County, Florida | Fort-Lauderdale, Florida; Miami |
| Cobb County, Georgia | Mableton, Georgia; Atlanta |
| Collin County, Texas | Plano, Texas; Dallas |
| Denton County, Texas | Denton, Texas; Dallas |
| DuPage County, Illinois | Aurora, Illinois; Chicago |
| Fort Bend County, Texas | Sugar Land, Texas; Houston |
| Gwinnett County, Georgia | Lawrenceville, Georgia; Atlanta |

| | |
|---|---|
| Norfolk County, Massachusetts | Quincy, Massachusetts; Boston |
| Pierce County Washington | Tacoma, Washington; Seattle & Olympia |
| Riverside County, California | Riverside California; Los Angeles County |
| San Bernardino County, California | San Bernardino, California; Los Angeles |
| San Mateo County, California | San Mateo, California; San Francisco |
| Snohomish County, Washington | Everett, Washington; Seattle |
| Tarrant County, Texas | Fort Worth, Texas; Dallas |

From 2016 to 2020, these are the counties that appear to be the most unique positive moving destinations for US residents. Our “hidden gems” are sometimes not so hidden, but they are unique migration centers. One could also argue that we should get the same results by looking at the 14 counties that had the highest net intake percentage, or the highest net intake by numbers. For both of these metrics, 7 of the top 14 counties are not included in our gem list. And both the gems and non-gems in those lists differ. The spectral clustering done here gives the best insight into which counties with a net positive gain in residents had distinct enough migration patterns to be considered a hidden gem.

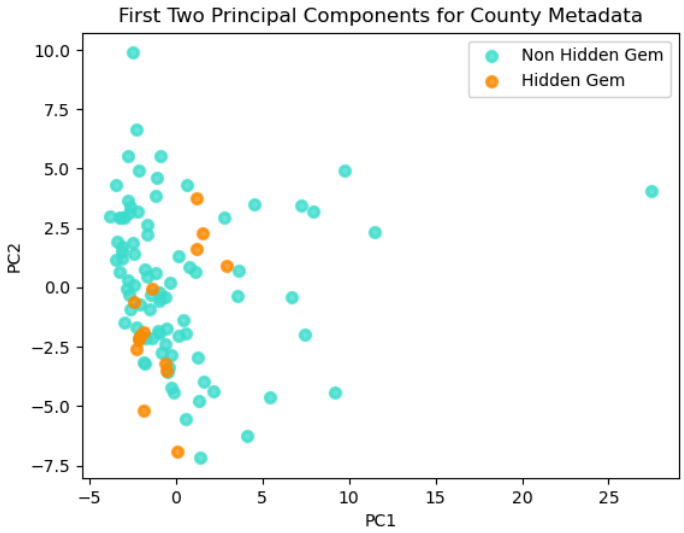
**Classification:**

After finalizing our list of hidden gems, we created an indicator variable, where 1 represents a hidden gem and 0 means otherwise. We then implemented multiple classification models to try and see if we could use our metadata to predict if a county could be considered a hidden gem. The models used were Gaussian Naive Bayes, Logistic Regression, K-Nearest-Neighbors with k=1 to 10 neighbors, linear SVM, RBF-kernel SVM, and Neural Networks. We also tried a One Class SVM model with the RBF kernel to see if an anomaly detection algorithm yielded better results.

The metadata dataset has 62 independent variables. We removed those having to do with population and population change, as we were trying to look for factors unrelated to those metrics that determine if a city is a hidden gem. We also removed variables involved with land area in 2010, as they were virtually identical to measures in 2020. This left us with 56 independent variables. The Pearson correlation coefficients between the metadata variables shows some high correlation, indicating that our models could benefit from dimension reduction.

One challenge with building these models comes from the prior distributions of the classes. There are 14 hidden gems among 100 counties, so there is a higher chance of False Negative predictions, as well as the accuracy being reliant on the number of correctly classified non-gems. The F-1 score shows a better evaluation of the models, since it balances the recall score with the accuracy.

Another challenge is the lack of clear separability of gems and non-gems. A visualization of the first two principal components of the entire dataset indicate that it may be difficult to separate the two classes.




We divided the 100 rows of data with a 70-30 train-test split. This split was chosen to ensure there would be enough gems in the test set.

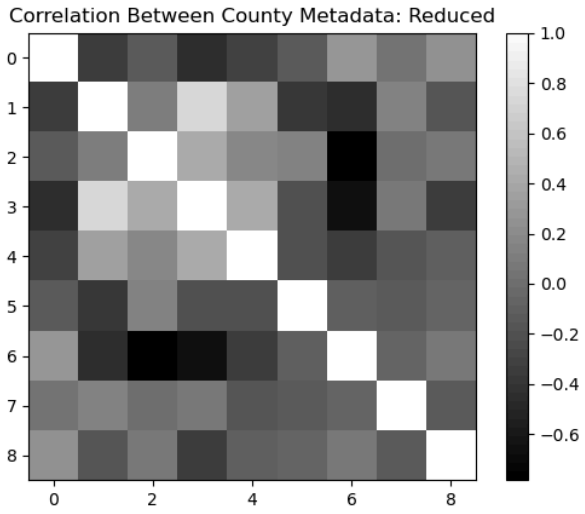
Each model was trained three times: once with the full dataset, once with the data transformed into the first two principal components, and once with a subset of features chosen by LASSO Logistic Regression. When building the models with the full data, the resulting accuracy and F-1 Scores were:

| Model | GNB | Log. Reg. | KNN: k=5 | LSVM | KSVM | NN | 1-C SVM |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.667 | 0.833 | 0.9 | 0.833 | 0.867 | 0.833 | 0.7 |
| F-1 | 0.167 | 0.444 | 0.400 | 0 | 0 | 0.286 | 0.816 |

The models built with the first two principal components had the following scores:

| Model | GNB | Log. Reg. | KNN: k=5 | LSVM | KSVM | NN | 1-C SVM |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.867 | 0.867 | 0.867 | 0.867 | 0.867 | 0.833 | 0.733 |
| F-1 | 0 | 0 | 0.0 | 0 | 0 | 0.286 | 0.84 |

The LASSO model was built with the same train-test split as the other models. The 9 features with non-zero coefficients were Persons under 18 years, percent; White alone, not Hispanic or Latino, percent; Households with a computer, percent, 2018-2022; High school graduate or higher, percent of persons age 25 years+, 2018-2022; Total retail sales per capita, 2017; Mean travel time to work (minutes), workers age 16 years+, 2018-2022; Persons in poverty, percent; Total employment, percent change, 2020-2021; and Land area in square miles, 2020. The resulting model had an accuracy of 0.867 and an F- score of 0.5. There are still some strong correlations, like the negative relationship between percentage of households with a computer and the percentage of persons in poverty, but there are less overall than with the full dataset.




The models built using only with these variables had the follow results:

| Model | GNB | Log. Reg. | KNN: k=5 | LSVM | KSVM | NN | 1-C SVM |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.9 | 0.933 | 0.933 | 0.933 | 0.933 | 0.933 | 0.7 |
| F-1 | 0 | 0 | 0.0 | 0 | 0 | 0 | 0.824 |

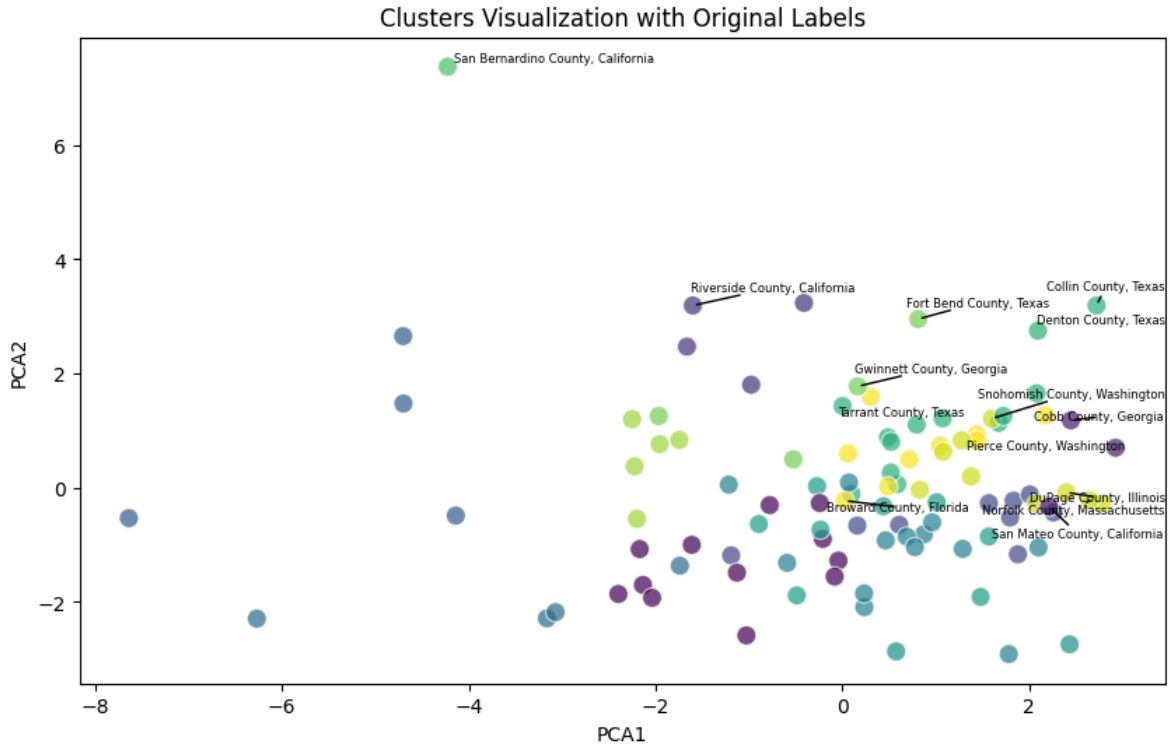
The low F-1 Scores across the various models indicate that they have poor predictive performance, especially in finding True Positives. The One Class SVM models performed much better. The accuracy was lower than for the other models but the higher F-1 score shows that it had better performance for both the gem and non-gem counties.

In conclusion, we are not able to use basic classification algorithms to decide if counties are gems based on the available metadata. There is likely too much overlap between the two classes and too many non-gems that it is too difficult to tell one class from the other. The fact

that anomaly detection has better performance does mean that there is a discernible difference between the two classes, but it is best to view the hidden gems as a deviance from the other counties instead of there being a large separation between them. Although our classification performance was not as successful as we had hoped, there is still more to glean from the county metadata. Next we will use K-Means to delve deeper into the features and investigate the relationships between hidden gems and other counties.

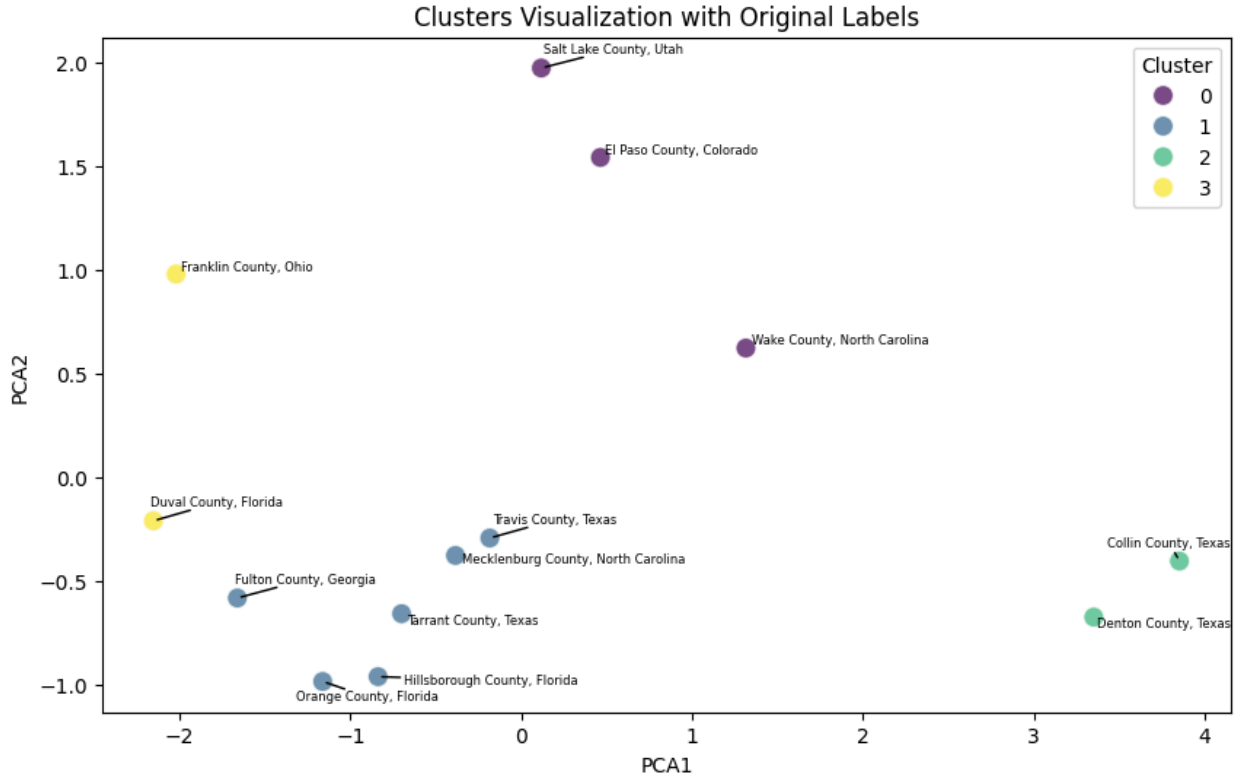
**K-Means Clustering:**

K-Means is susceptible to the curse of dimensionality, so we do not want to use all 56 features. Selecting the features highlighted through the LASSO classification model, we were able to cluster the counties. Features include demographic information, housing information, employment information, and more. Using a k value of 15, we were able to separate the top 100 populated counties into clusters. Below is a plot with the gem counties highlighted. Our clustering is done with 8 features, while visualization is done in the first 2 principal components. We are confident that 8 features is not too many dimensions for K-Means to work successfully. Please note that PCA and the following visualization lose roughly half of the variation in the data.




From here, we want to nominate “sister gem counties”. After highlighting migratory gem counties, we want to find other counties that have similar features. For example, if we look at cluster 1 where Cobb County and San Mateo County are assigned, we can see that King County is also grouped with them. Therefore, we can nominate this county as “sister gem.” Although it wasn’t highlighted in our original gem search, it seems to have similar features and could be a potential migratory gem in the future.

We can replicate this analysis for other gem cities as well, sub-clustering as needed if there are too many counties assigned to a certain cluster. For example, Collin County, Tarrant County, and Denton County are all in cluster 9, which has 13 counties total. We could label the other 10 counties as “sister gems”, but we want to be more precise. Therefore we sub-cluster group 9 and see where the hidden gems are assigned.




Sub-clustering cluster 9 with a k-value of 4 gave us this plot above. Looking at Tarrant County in cluster 1 we can nominate the counties it is grouped with as sister gems. There are no sister gems to be nominated for Collin County and Denton County since they are clustered to the side on their own. By sub clustering when needed, we are investigating deeper into the underlying structures between counties.

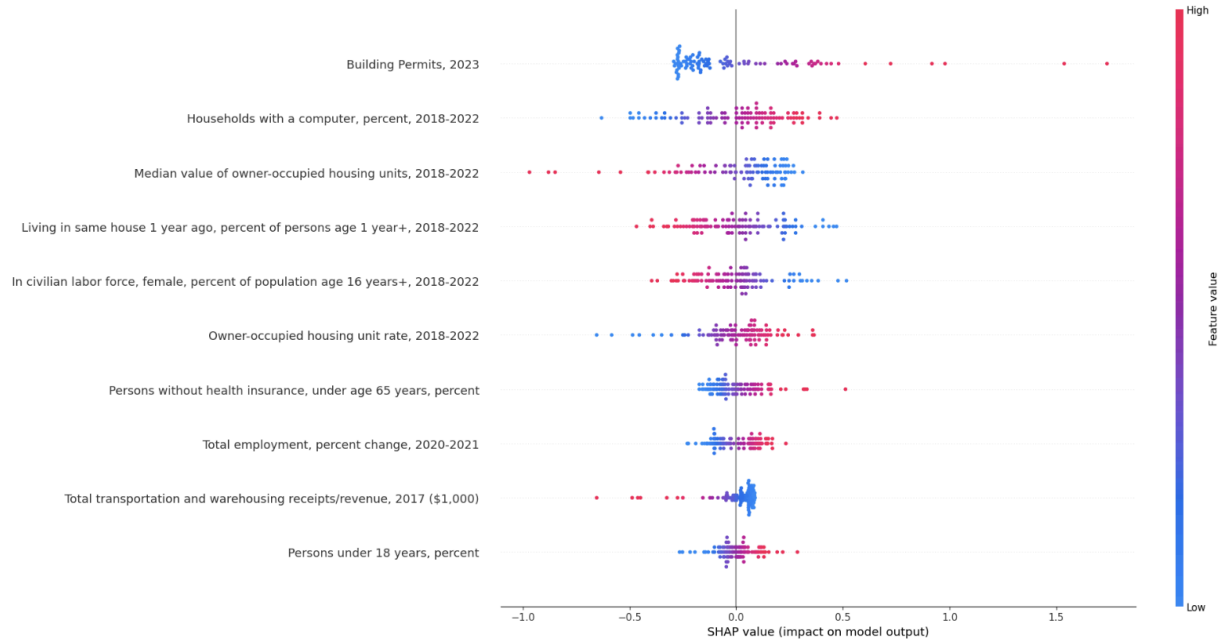
We can use the outcome of this clustering to uncover counties that have similar traits to hidden gems that have not yet been discovered. They act as a list of potential hidden gems where people might be moving to next in the coming years.

**Elastic Net Regression:**

We decided we needed to know more of the “why” for population percent increases. We believe that regression would produce a reasonable model, but we were data poor and feature rich. We had just over half as many features as we did data points. For this reason, we wanted to use LASSO regression for feature selection. However, as mentioned before we had many

columns that were strongly positively or negatively correlated. To handle this, we would need to use Ridge Regression. Regardless, we scaled and imputed our data for model stability.

Elastic Net uses aspects of both LASSO and Ridge, so we fit a cross validated elastic net model. We checked for many values of alpha and l1. Alpha helped us decide the ratio of Ridge vs LASSO, while l1 told us how strict to be, meaning how many features we wanted to eliminate. Our best model had **alpha=0.1 and l1_ratio=0.2**. Our model achieved an R2 score of 0.62. We believe this score indicates decent performance. The SHAP plot below shows our most important features in our model, sorted by importance.



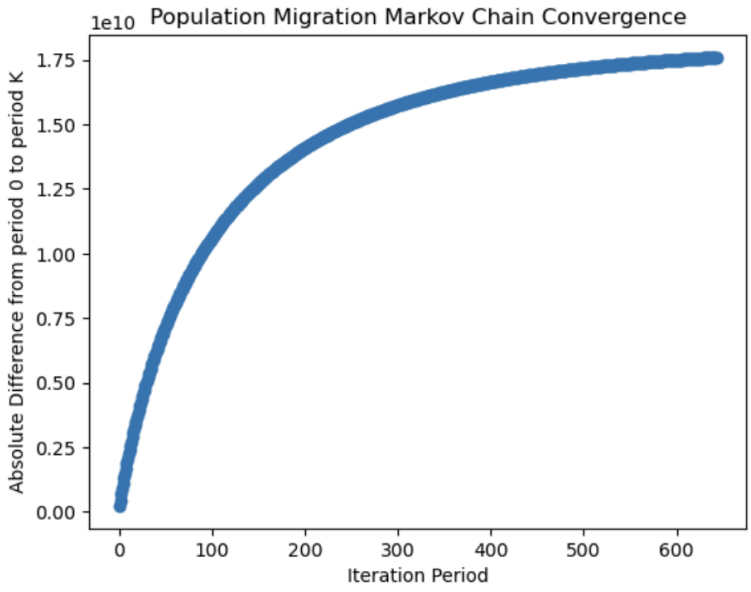

Counties that experience more building permits would tend to experience more economic growth, and therefore more population growth. Households with more computers may be more likely to participate in the digital economy. Similarly, cheap housing might be more attractive, where cities with high populations but high housing prices might experience a population market correction. Counties that have more housing mobility may be more likely to efficiently accommodate more people as well. This could signify that their housing market has more “wiggle room”. Other top features may be more correlated. For example, Persons under 18 and persons without health insurance might be correlated by the hypothesis that young people are more likely to skip insurance. One important note is that the current population is not a top predictor in this model.

**Markov Chain Population Prediction:**

Finally, we wanted to run a simple model to predict the population for all counties given comovement between counties. There are a number of ways to tackle this, but we decided to use Markov Chains. We created an adjacency matrix that contained the population for all counties, and the percent of county A that migrated to county B, and the percent of county B that migrated to county A. We then calculated the population at Period K using the population from period K-1 and the percent changes between counties at Period 0. This Markov Chain has a nice property that eventually, the populations will reach a "steady state". Some key assumptions for this:

- The percent movement from A to B and B to A are static, meaning they are the same at Period 0 through Period K. Note that the population movement is not static - if A increases in population, then the number of people moving from A to B will increase.
- Births = deaths at the county level.
- There is no migration to or from outside of these 100 counties.

To confirm that we will eventually reach a steady state, we ran 650 migration periods. The Y Axis is the sum of the absolute row-wise difference between Period 0 and Period K. As we can see, eventually the differences in populations will even out. So if nothing ever changed in migration patterns, the population of our counties would be easy to predict! The overall population across all counties never changes, but the population within counties does change. For example, one of our "sister hidden gems" Mecklenburg County, North Carolina would see a population increase of 2.3% in 10 years when only considering the top 100 counties. If we are correct in our hypothesis that counties more similar to "hidden gems" will experience more growth eventually, this would be a lower bound. We would also like to evaluate this model using all US counties instead of the top 100 in the future.



Unfortunately, none of our three assumptions hold in reality. But our Markov Chain approach is a strong way to estimate population changes, especially in the short term. We like that it includes the comovement between all 100 counties and we would use this as a baseline for predicting the population of our sister hidden gems. Given our K-Means analysis, we would envision the sister hidden gems to eventually exceed the estimates from this Markov model.

## Evaluation and Final Results:

Through our analyses we were able to successfully identify migratory hidden gems, discover the features that may be responsible for increasing migration, discover sister hidden gems that may be candidates for a high influx of migrants in the coming years, and predict future populations. By conducting spectral clustering on the migration data, we were able to identify the following 14 counties as hidden gems: Broward County, Florida, Cobb County, Georgia, Collin County, Texas, DuPage County, Illinois, Fort Bend County, Texas, DuPage County, Illinois, Gwinnett County, Georgia, Norfolk County, Massachusetts, Pierce County Washington, Riverside County, California, San Bernardino County, California, San Mateo County, California, Snohomish County, Washington, and Tarrant County, Texas. Once labeled as hidden gems, we were able to run a LASSO logistic regression model to do feature selection on our list of metadata features. With this reduced list, we applied K-Means clustering on the counties to identify sister hidden gems that have similar features to the original migratory gems. Finally with linear regression, we identified the key factors for population increase rates. We also created a markov model to estimate a baseline population growth.

Our models varied in performance. In particular, our classification models encountered issues, and outlier models had more success. Our Elastic Net regression had a fairly decent R2 of  0.62. The rest of our models are either unsupervised or we would need to wait for more data to successfully evaluate. Our PCA dimension reduction, which we primarily used for visualization, captured about 40% of our variance.

Given more data and more time, we believe we could fine-tune our approach to achieve better predictive and explanatory power, and even provide a prescriptive recommendation to people given their living preferences. For example, if you selected a city you like or metadata features most important to you, we could identify a list of counties you might also like, and make predictions on the future population.

We would also like to replicate this approach with different subsets of counties such as top 101-200 most populated counties, counties just on the west coast, and all counties in the US. Because we do spectral clustering to identify hidden gems we were bound by the number of counties we could use since the matrix size and runtime grow quickly with the addition of more counties. We also had unforeseen data collection issues as the census website went under partial construction while we were gathering data. It would also be interesting to include our list of hidden gems with counties of lesser populations to see which mid-sized counties have similar features.

The scope of this project allows for many different avenues of exploration. Whether it's focusing on subsets of features, or investigating different subsets of counties, there is a lot more insight to uncover and we would love to extend it given the time (and financial compensation).

## Division of Labor:

**Kerstin:**

- Project Proposal Drafting
- Spectral clustering
- Classification models

**Katha:**

- County metadata compilation
- Migration to metadata linkage
- K-means model
- Spectral clustering

**Joel:**

- Elastic Net Regression
- Markov Chain forecasting
- County to County Migration Data Cleaning

**Kerstin, Katha, & Joel:**

- Methodology Brainstorming
- Final Report Writing
- Peer Review of each other's work
- Emotional Support

## References:


1. Introduction to problem: https://www.thepolicycircle.org/minibrief/migration-between-states/
2. County to county migration dataset: https://www.census.gov/data/tables/2020/demo/geographic-mobility/county-to-county-migration-2016-2020.html
3. County metadata source https://www.census.gov/quickfacts/fact/table/tulsacountyoklahoma,jeffersoncountyalabama/PST045223
4. Markov Chain Inspiration: https://setosa.io/ev/eigenvectors-and-eigenvalues/