Discovering County Trends with Machine Learning

Team 9 - Kim Di Camillo, Liz Giancola, Ben Merrill

Human migration, the movement of people with the intention of resettling, impacts many dynamics of society including city planning applications, international trade, spread of infectious diseases, conservation planning, and public policy development. California is the most populous state in the US, and the third-largest by area, containing the Greater Los Angeles area, which comprises the second largest urban area in the country, and the San Francisco Bay Area, the fifth largest urban area in the country.[1]

In this paper, we explore some of the social, economic, and geographical characteristics of California counties to determine their influence on population inflow and outflow. We aggregate data for the state at the county-level from a variety of sources and create both supervised and unsupervised methods to explore county trends. We conclude that there are clear geographic indicators of different trends among counties in the state and discover correlations between fields that are indicative of population changes.

Datasets and Cleaning

For each of the following datasets, we transformed the data into 58 rows (one for each county) for each year from 2014 through 2019. This data was used for both supervised and unsupervised approaches.
IRS Tax Migration Data | Excel | 348 rows | Inflow and outflow
CA County Geographic Boundaries | Shapefile | 58 rows total | geography plus land and water area
American Community Survey (ACS) Data | API pull | 348 rows | demographic data
Federal Emergency Management Agency Data | csv | 2000 rows | different types of disasters
California School Expenditures Data | csv | 348 rows | average daily attendance expense
California Voter Registration Data | csv | 348 rows | voter registration by party
California Weather Data | csv | 88334 rows | average daily, min, and max temperature and precipitation
Bureau of Economic Analysis (BEA) Data | csv | 348 rows | job, income and gross domestic product data
California Crime Data | xlsx | 348 rows | major crime category counts
California Coast Data | html | 58 rows | flag 1 or 0

We aggregated datasets from various government agencies and public sources and utilized them in exploratory data analysis, supervised learning, and unsupervised learning. After cleaning the data and creating appropriate fields, we were able to merge each of these datasets on unique county codes, state codes, and years. Our final file, ca_counties_full_dataset, contains all fields shown in the appendix for California and has 342 rows and 82 columns. For more detail on the variables, see Appendix A.

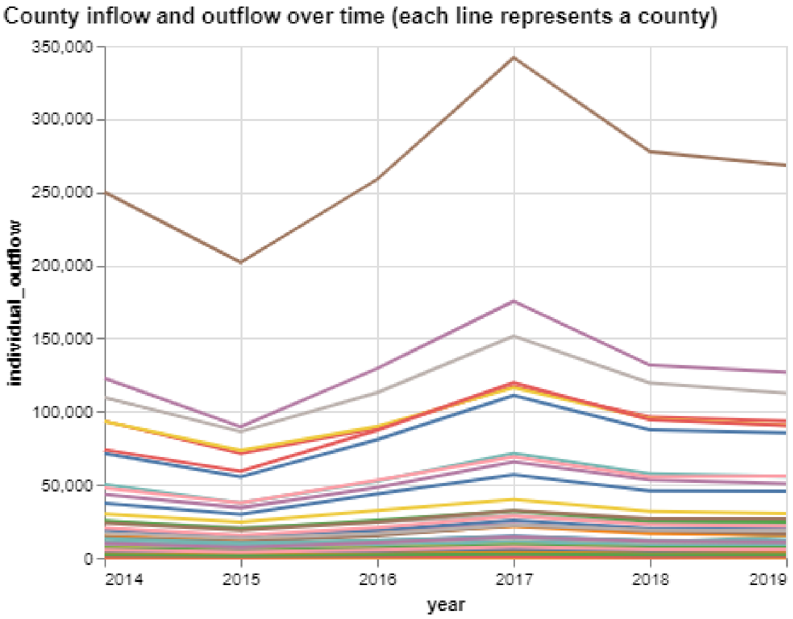
Exploratory Data Analysis

We were interested in understanding the relationship between inflow and outflow and how that varies both by county and by year. To that end, we performed EDA to investigate the change in inflow and outflow from 2014 through 2019 in all 58 CA counties. In the following graph, each line represents the change in inflow for one county over time. For the very similar graph of outflow, see Appendix B. We see that inflow remains pretty consistent throughout the period of this study. There are two notable systematic differences. For 2015, all counties with inflow over 20,000, have a decrease in both inflow and outflow. For 2017, we see all counties with an inflow greater than 20,000 have an increase in both inflow and outflow. This systematic difference has been documented by researchers at the University of Minnesota who caution against using this data after the IRS took over managing it from the US Census, noting anomalies in the reporting of this data.[2] It is unlikely that nearly all counties in California experienced a decrease in a given year. Instead, the method for determining the pool of possible tax records to include

[1] https://en.wikipedia.org/wiki/California
[2] DeWaard, Jack, et al.

in this data likely changed from year to year. For example, the decision to include taxes that were filed late could have been included in 2017 but not 2015. We discuss potential implications for the supervised learning model in that section.

See Appendix B for more exploratory data analysis and a graph of outflow with the same relationship.

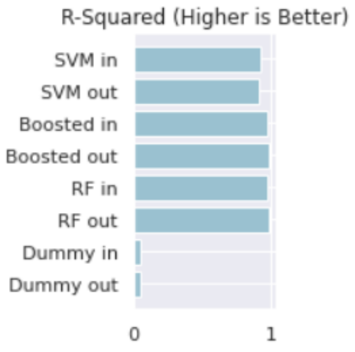


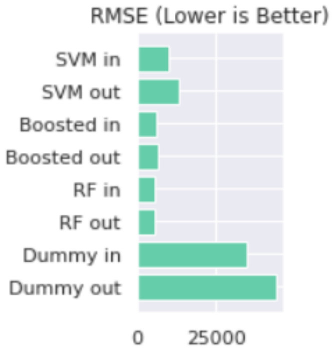
Supervised Learning Approaches

For supervised learning we used three different models to predict inflow and outflow of people in California counties. We then evaluated the most effective model to understand the reason for population changes across counties in California. The hope is that the accurate prediction of these models could be useful for counties to better understand the potential changes that may arise in the future.

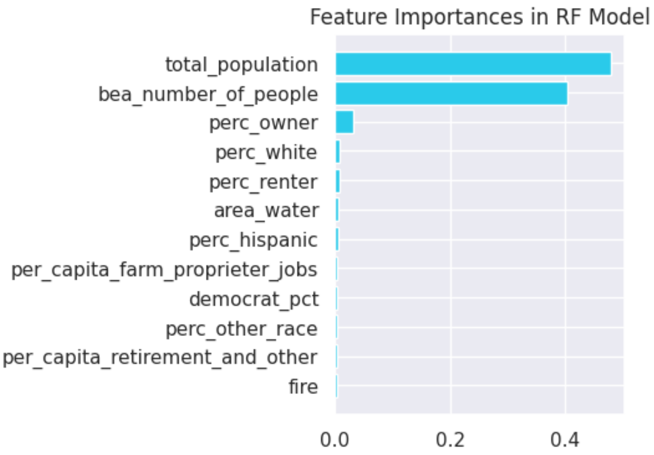
Our approach to this prediction problem started with the creation of a training and testing set. We split our target variables of inflow and outflow from the rest of the data and dropped fields irrelevant to the model, such as county or state code. As we are dealing with time-series data, the model was built without allowing any of the training data to hold future information. We also wanted the model to use current data to predict future population inflow and outflow (i.e. all 2014 fields are paired with 2015 inflow/outflow and considered 2014.) Thus, we used 2014-2017 as the training set and 2018 as the testing set. Finally, we used a standard scaler to fit each field in our training set fields on a normalized bell curve.

After normalizing our data, we built three models to predict our target variables of inflow and outflow. We chose Random Forest Regressor, Gradient Boosted Regressor, and Support Vector Machines as our models to fit the data, as well as a dummy regressor for a baseline. We decided that these three models give a more holistic approach to regression with different model assumptions. Random Forest is a powerfully simple multifaceted algorithm that can detect easily interpretable trends in the data. Gradient Boosted Regression utilized the power of a weak learner. Support Vector Machines use a more classification-based approach to regression. With models selected, we used grid search cross validation to detect the best hyperparameters for each model. We then looked at model accuracy. Random Forest had an R^2 value of 0.98, Gradient Boosting had 0.97, and Support Vector Machines had 0.91. Among these three models, we saw that Random Forest had not only the best performance with R^2 , but also the lowest root mean squared error.

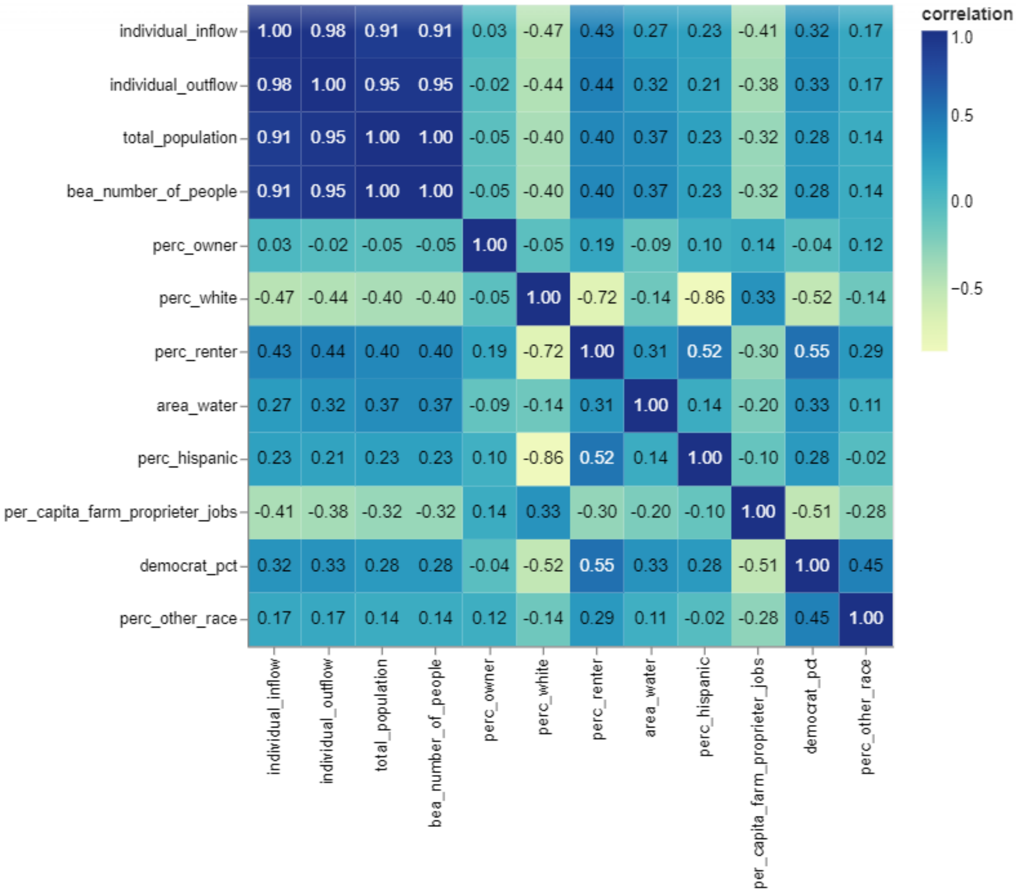




Next, we wanted to better understand why the performance of the Random Forest Regressor was better than the other models. We pulled the most important features, and they are displayed in the figure to the right. The top two fields deal with population growth and are unsurprising as a higher population would naturally increase the number of people traveling between counties. Interestingly, percentage measures of ownership and renting, race (white, hispanic, and other races), and political views also affect the model, as well as retirement, water in an area, and farming. Uncovering these ranked importances helped to better understand which of the many fields in the dataset were indicative of a higher inflow and outflow of a county.



Although these measures are useful, they don't give any indication as to how change in a given field might change population diffusion in a county. To solve this problem, we built a correlation matrix of the most important features in the random forest regressor.

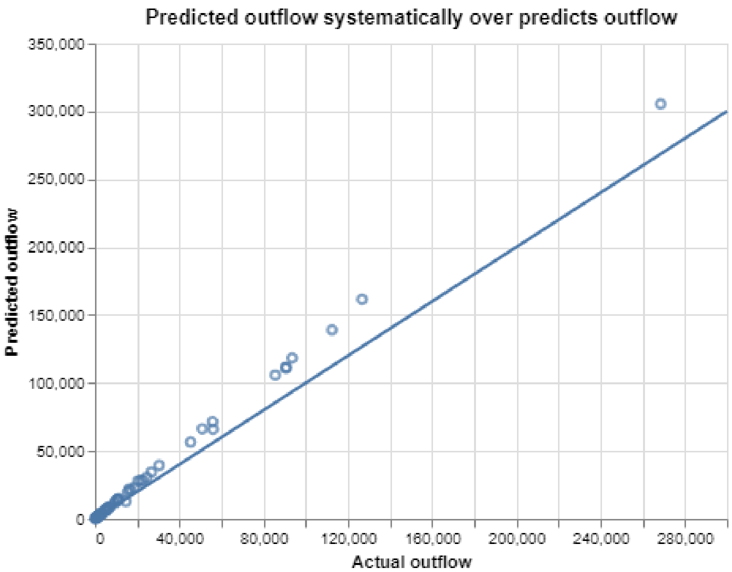


In the figure above, we see the correlation of the ten most important variables in our random forest regression. We see that population inflow, outflow, and total population are all highly correlated, hence the blue square in the top left corner of the matrix. It looks like a higher white population and more farm jobs per capita indicate a lower inflow and outflow in a county, meaning that counties with a higher proportion of white people and farmers may imply less transience. On the other side of the spectrum, high rentership, a higher democratic percentage, more lakes and water on the land, and higher hispanic and other race populations imply higher inflow and outflow. Likely these are outlining the highly populated counties such as Orange County, Los Angeles County, and San Francisco County, where there are more people. Also, high ownership percentage is not indicative of inflow and outflow at all.

There are also some other high correlations that are worth mentioning. We see that a high proportion of white people in a county is correlated negatively with population size, hispanic populations, rentership rates, and democratic political views. Likely, the counties with higher white population are rural, indicated by lower population and more farm jobs, are less transient, have less renters, and less hispanic and other race people. We also see a negative correlation between democrats and farm jobs. Alternatively, we see a high correlation between rentership rates and hispanic population, indicating that a higher proportion of hispanic households in a county may imply that more of the population is renting. It is also clear that a democratic political view is highly correlated with rentership.

We were curious if we would be able to achieve similarly strong predictive results with a supervised learning model that used only the IRS data inflow and outflow. To create such a model, we transformed the IRS data so that each row contained county-level IRS inflow and outflow for each year from 2014 through 2019. The training data used the year 2018 inflow/outflow as the target, and input variables were the inflow and outflow for the four previous years. The test set used 2019 inflow/outflow as the target. Because the outflow and inflow model are so similar, we will only analyze the outflow model. Using a linear regression model, we achieved an R^2 value of 0.94 and a mean squared error of 10,468 for the outflow model. The coefficient of determination means that 94% of the variation in the actual value can be explained by the predicted outflow value and the mean squared error means that the average squared distance between the actual and predicted outflow is 10,468 people. Notably, we achieved an R^2 value of 0.999 on the training data, indicating that our model may have overfit to the training data.

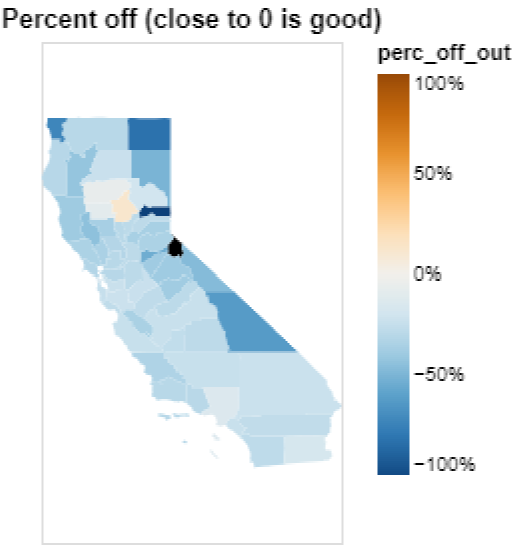
Though this model does not have the highest R^2 value, because it is so simple, we thought it would be more likely to be deployed. As such, we decided to further investigate the accuracy of the model.



Despite the high R^2 value for the linear model, the model has some significant weaknesses, including high bias. If our model was perfect, the actual value and the predicted value would be the same. Perfect predictions are represented by the blue line in this model. As you can see, our model systematically overpredicts outflow. Our predictions are higher than the actual value for all but one county. If we inspect the difference between actual and predicted values, we find that the counties with the largest outflow have the largest residuals (LA, San Diego, and Orange). This is not particularly useful.

Instead, we can inspect the percent difference between actual and predicted outflow, calculated as (actual - predicted)/actual. In the map below, we see that we overestimate the outflow for all but one county (Butte). We predicted that Butte County would have an outflow of 12,833, but its actual outflow was 14,647. The actual value was 12% higher than the prediction. Butte county (in orange on the map), had a significant increase in outflow between 2018 and 2019 because of the wildfires that destroyed over

10,000 homes in that county.[3] Our model appears to do a good job predicting this county, but because this county is an anomaly, this model should actually do a poor job of predicting this. This makes us wonder whether the issue with the model is connected to previously discussed inconsistencies of the IRS data or a change between 2018 and 2019 that our model didn't account for.

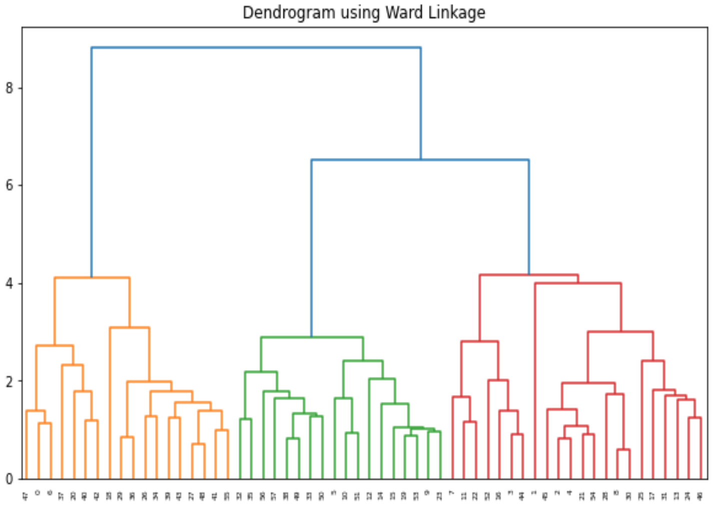


The other notable error is Alpine county (in dark blue in the inland elbow of CA). For this county, the actual outflow was 96, and our prediction was 255. The difference was 316% off, which is about 200% larger than the county with the next highest error. This huge percent difference is largely because the county outflow was so small. Using percent difference over values counties with small outflows and undervalues those with large outflows.

Unsupervised Learning Approaches

We used unsupervised learning approaches on 2018 data to look at our counties to try to understand which are similar, and what features are most influential in any similarities. In order to do this, we utilized multiple methods, including K-Means clustering, agglomerative clustering, and DBSCAN clustering. With 72 numeric scaled features, we realized it would be prudent to run a Principal Component Analysis on our data. After doing so, we re-ran the K-Means clustering to compare the results.

Our first clustering analysis was completed on data using a MinMax Scaler, which scaled all of our data in the range [0,1]. We first ran K-Means clustering on a range of 2-18 clusters and calculated the Davies-Bouldin Index and Calinski-Harabasz Index to assess the best cluster size. Additionally, we used the inertia_attribute and the elbow method to find where the sum of the squared error started decreasing in a linear fashion. We tried several types of initializations for K-Means, and neither showed a strong "elbow". Our next step was to perform a silhouette analysis to study the separation distance between the resulting clusters. Our best result of 5 clusters had an average silhouette coefficient of about 0.21. This low average value, plus the fact that none of our clusters scored above 0.5 indicated that our clusters were not very well separated. This led us to the conclusion that maybe K-Means on our raw data was not the best choice for us.

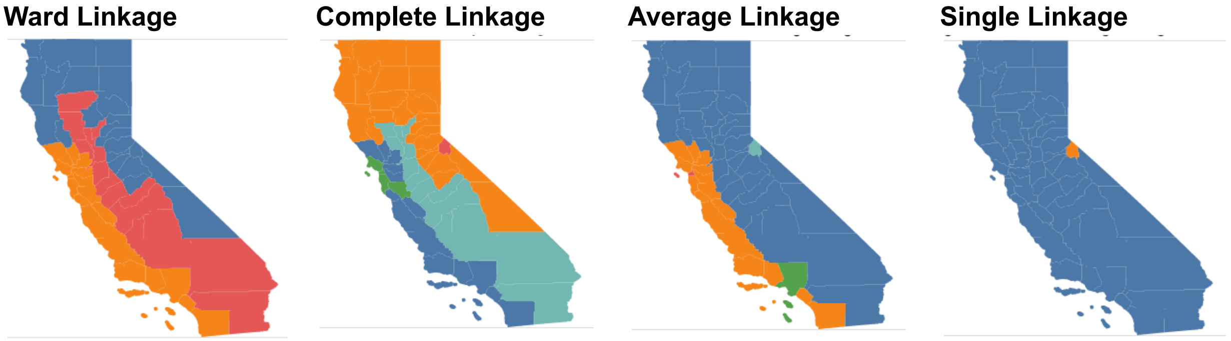


At this point, we decided to try agglomerative clustering as it uses a bottom up approach instead of the top-down approach of K-Means. For agglomerative clustering, we ran 4 types of linkage models: single for minimum distance between clusters, complete for maximum distance between clusters, average distance between clusters, and Ward's Method, which calculates the difference between the total within cluster

[3] Chabria, Anita

sum of squares for the two clusters separately and the within-cluster sum of of squares if the clusters were merged. The dendrogram showing resultant clusters using Ward Linkage is shown above.

Running the other 3 methods resulted in quite different results. Here are the clusters presented on a map of California for our 4 methods:



Ward Linkage **Complete Linkage** **Average Linkage** **Single Linkage**

In the Ward linkage map, we see three clusters that represent three distinct areas of California, the Central/Southern coastal area, the Central Valley and deserts, and the Northern coast and mountains. The complete linkage map is similar, but adds an additional cluster for the Bay Area and some surrounding counties, as well as an additional cluster for Alpine County. In fact Alpine County is its own cluster in 3 out of our 4 results. Alpine County is primarily composed of three national forests and about 96% of land is owned by the federal government[4]. Its tiny population of 1,146 in 2018 is less than half the size of the next most populous county of Sierra. The average linkage map is interesting in that it groups most of the state together, but separates out San Francisco County, Los Angeles County, Alpine County, and then the rest of the Central/Southern Coast.

The next type of clustering we tried was DBSCAN, which works well with datasets that have more complex cluster shapes. We ran DBSCAN using various values for the eps parameter which is the maximum distance between two samples for one to be considered as in the neighborhood of the other. The two best scenarios were 4 clusters with 10 outliers and 3 clusters with 7 outliers. In both situations, DBSCAN generally made one large cluster with a few smaller ones, and then multiple outliers, as seen in Appendix D. It did not seem like a good choice for learning more about California's counties.

At this point, we decided to perform some feature reduction by using PCA, with the idea that maybe all of the features were impacting the ability to create useful clusters. We created a PCA Explained Variance chart and found that our first 2 principal components gave us an explained variance ratio of 0.49. Here are the features that provide the most variance:

Principal Component 1		**Principal Component 2**	
coastal_flag	0.419181	perc_65_over	0.305379
median_home_value	0.234739	perc_white	0.280679
median_income	0.226722	perc_hispanic	0.267592
democrat_pct	0.224378	perc_poverty	0.236784
median_rent	0.221641	per_capita_num_vehicle_theft	0.224469
american_independent_pct	0.213186	registered_pct	0.216534
republican_pct	0.211032	total_precip_amt	0.210019
educational_attainment	0.209305	per_capita_nonfarm_proprieter_jobs	0.201308
perc_asian	0.194298	max_temp	0.192144
per_capita_personal_income	0.177468	avg_temp	0.190186

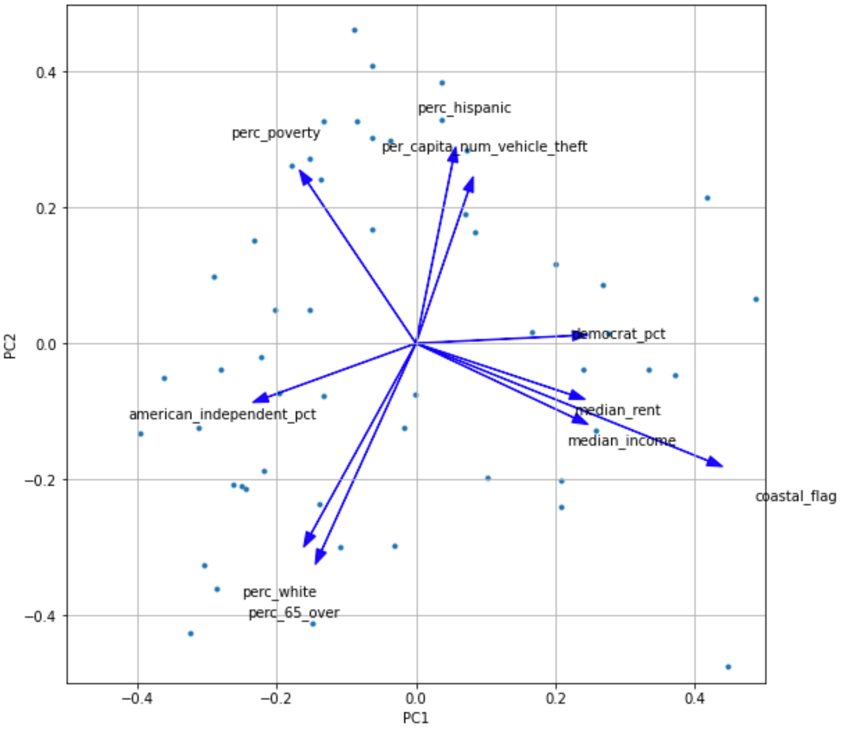
[4] https://en.wikipedia.org/wiki/Alpine_County,_California

Interestingly, the analysis showed that 14 of the 15 FEMA variables had a variance of 0 in both principal components. The only FEMA variable with any impact was “fire”, with a tiny impact on PC-2, of 0.007.

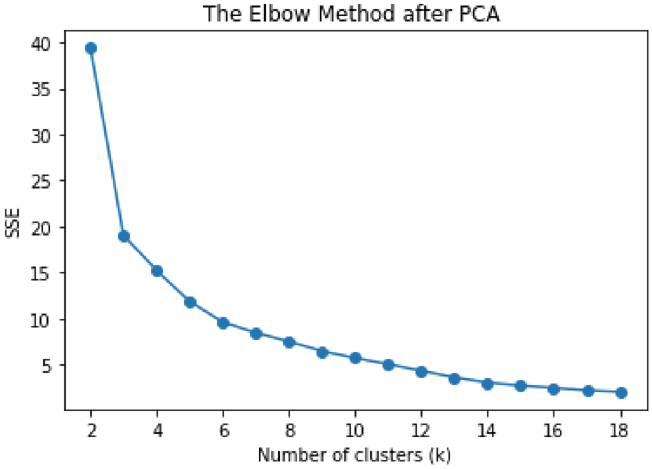
The biplot on the right shows the vectors for 10 of the top differentiating features across the first two principal components, along with the data markers for our counties. Here we can see multiple groups of variables that are likely highly correlated (likely because our first two principal components do not make up 80% of the variation in the data):

- perc_white and perc_65_over
- median_rent, median_income, and coastal_flag
- per_capita_num_vehicle_theft and perc_hispanic

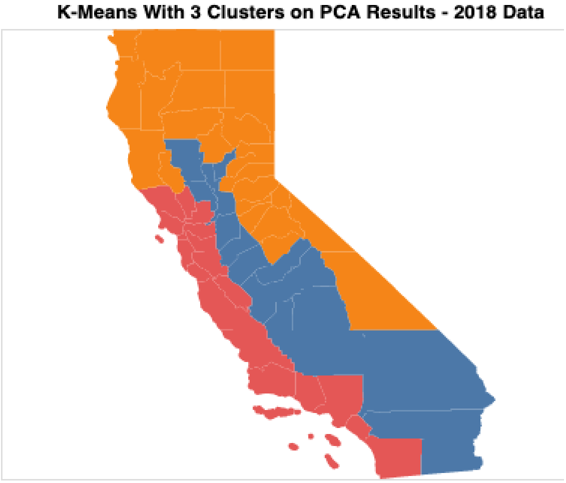
We can also see that some features are negatively correlated, such as perc_65_over and per_capita_num_vehicle_theft, which are orthogonal to each other. It is also clear that democrat_pct is highly correlated with principal component 1.

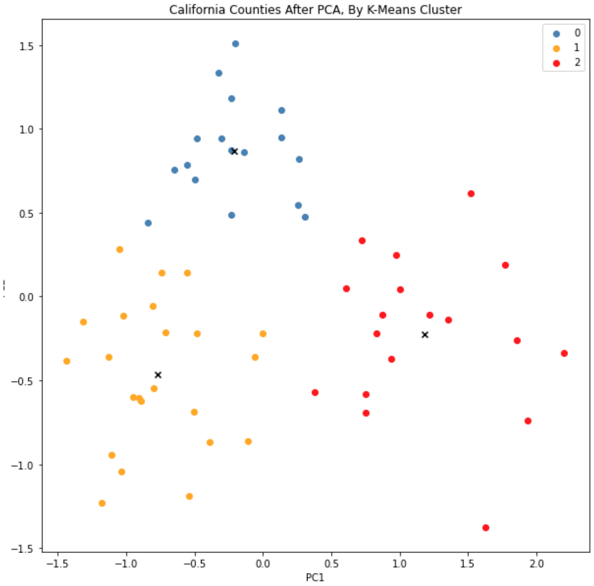


After performing feature reduction using PCA, we decided to try K-Means again to see if our results were any more conclusive. The elbow method after PCA proved to be much more pronounced, clearly showing an elbow at 3 clusters.



The resulting clusters for K-Means with 3 clusters after PCA are shown in the two maps below, one by geography, and one by principal component values. The “x” represents the cluster center:



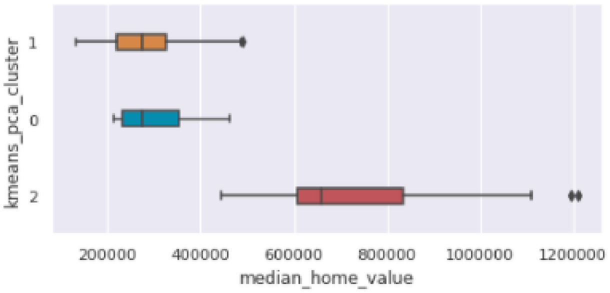


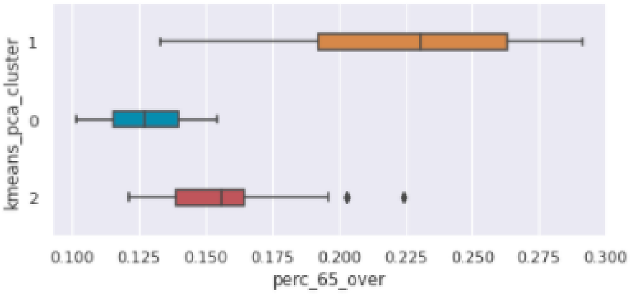
Looking at our map, we notice that it is nearly identical to our Ward linkage agglomerative clustering result. The only difference is that the Central Valley cluster in the Ward result includes the county of Tehama, and this result includes Tehama in the northern coast and mountains cluster.

By viewing the scatter plot of CA counties color coded by cluster and the feature vectors in the biplot on the previous page, we'd predict that cluster 1, the northern region (orange), has a higher percentage of 65+ and white residents. We'd also predict what we'll call the Central Valley (0 in blue), has a higher percent of poverty, hispanic population, and higher vehicle theft. Finally, for the coastal region (red), we'd expect a higher percentage of democrats, and higher median income and rent. We'd also expect a higher coastal flag, which we can visually verify via the map. With a basic knowledge of CA geography, none of these findings are surprising and the following table, containing mean values for each cluster, confirms this. With more time, we’d consult a statistics expert to determine the statistical significance of these findings.

	Over 65	White	Hispanic	Poverty	# Veh. Thefts per Capita	Coastal Flag	Median Home Value	Median Income	Democrat	Median Rent
0	12.8%	35.5%	49.3%	17.5%	0.0040	0%	294,312	57,718	36.3%	1,054
1	22.6%	74.2%	14.9%	14.2%	0.0022	13%	287,322	55,549	31.1%	1,001
2	16.0%	44.2%	33.3%	9.7%	0.0032	100%	750,772	92,003	46.7%	1,773

We also examined the distribution of the second feature in PC-1 (median home value) and the first feature in PC-2 (percent over 65). We see that nearly all counties in the coastal cities cluster (red) have higher median home value and that in the northern region cluster (orange), 75% of counties have a larger percent elderly population than nearly all of the counties in the other two clusters.

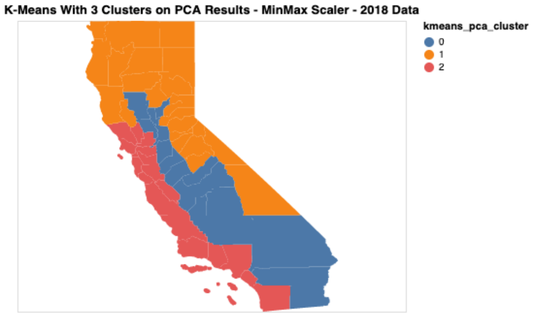


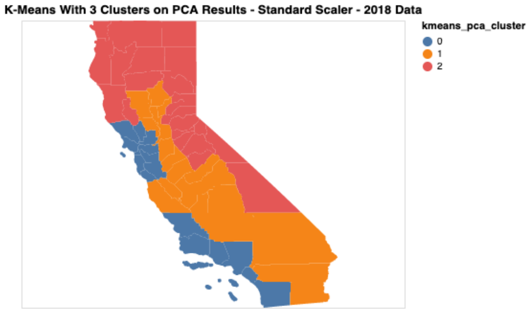


At some point in our analysis, we realized that we were performing supervised learning with the standard scaler and unsupervised learning with the MinMax scaler. To that end, we reran the entire unsupervised learning process again with standard scaling. We were extremely surprised, and concerned, to see how different the results were. This really drove home the importance of data preparation and how impactful it can be in data science. Some of the notable differences are:

- Using MinMax scaling led to clusters where all counties in a cluster were adjacent to each other. With standard scaling several of the cluster results had counties that were not adjacent.
- The top 10 list of differentiating features for the first principal component of PCA shared 8 values. However, the standard scaler did not have coastal_flag (the top differentiating feature in MinMax) and per_capita_personal_income, and MinMax did not have perc_renter and min_temp.
- The second principal component of PCA overlapped by 7 of the 10 most differentiating features.

The following shows how scaling impacted K-Means clustering after PCA:


K-Means With 3 Clusters on PCA Results - MinMax Scaler - 2018 Data
kmeans_pca_cluster
● 0
● 1
● 2


K-Means With 3 Clusters on PCA Results - Standard Scaler - 2018 Data
kmeans_pca_cluster
● 0
● 1
● 2

Discussion and Conclusions

In the supervised section, we looked at the power of regressors built on using all 70+ variables as inputs and the power of simple models. We learned that of the models chosen, Random Forest Regressor was the most accurate. This may be because the model is easily able to pick predefined fields to split on, such as population, that are indicative of population diffusion among counties. Using the results, we were surprised to see such interesting correlations between different demographic groups, population sizes, and political views among different counties. Many of the correlations seen in the data are popular topics in the news at the moment. After all of our efforts to wrangle and prepare the data, it was also illuminating to see how well our simple linear regression model performed in comparison.

Generally, the unsupervised learning for this project showed us that choice of model and data preparation can greatly change the results of the analysis. Even within agglomerative clustering, we saw four quite different results based on the type of linkage chosen. It is important to have domain knowledge to help understand how to select between different outcomes, because simply using the algorithms does not help you to know which one is providing the best result. We also learned that dimension reduction using PCA can be very helpful when working with varied features in unsupervised learning tasks. More specifically, we learned that similar California counties are located near each other, forming distinct geographic areas that are representative of some of the regions of the state. This was surprising, because we removed latitude and longitude from the dataset. We were also surprised how impactful the choice of a scaling tool was to the results.

Some ideas we had to extend our analysis on this subject further are:

- Build a model on each of the three clusters found in the unsupervised learning section to discover differences between clusters using a supervised approach.
- Use grid search cross validation on other models to see if we could improve model accuracy. We could also extend our dataset nationwide or expand the number of years shown.

- Train another linear regression model with input as year t-1 outflow/inflow, and target as year t. That way, we could have 2015-2018 inflow/outflow as target variables in our training set and still have 2019 as a test set. This would likely reduce the bias that we saw with this model. We could also use 2018 as a validation set so that we could complete more model tuning without the risk of data leakage. If we had 2018 data as a validation set, we could tune variables in the same way that we tune model parameters, adding some variables from our large variable data set to the array of inputs and analyzing how they impact model performance.
- Explore the network of migratory patterns of people from county-to-county. In the news, we hear stories about mass migration out of California to places with lower tax rates, including Idaho and Texas. It would be interesting to explore where people who leave California counties go and where people are from who move to California. It would be interesting to see if we could use 2018 county-to-county migration data to predict the 2019 county-to-county migration patterns and to explore where those predictions fail.
- Explore the results of the correlation analysis we ran on our 3 clusters after PCA (see Appendix C) to understand more about each cluster.
- Use results of unsupervised learning to help understand migration trends.
- Perform unsupervised analysis on different years of data, looking for changes over time
- Rework our usage of FEMA data. Instead of using dummy variables, obtain some type of numeric measure, such as risk of disaster, to allow these features to provide an impact.

Ethical Issues - Supervised Learning

The IRS data systematically underrepresents low income people who are not required to file taxes. As a result, predictions made from this data systematically disregard a vulnerable and important segment of our population. If this data were used for city planning, because those people are left out of the data, using this to make decisions could result in furthering the aggregation of resources toward those with power and money. To address this, we'd need to consider alternate data sources. The ACS provides county-level data, but for smaller counties, because the population is under 65,000, they use five-years worth of survey data to produce data about that population. As a result, for smaller counties, comparison of year-over-year migration estimates is not always advisable.

Ethical Issues - Unsupervised Learning

We saw a surprising correlation in the data, specifically that rural counties, indicated by lower population and more farm jobs, are less transient, have less renters, and less hispanic and other race people. This correlation is surprising because the Central Valley is largely farm country, where historically much of the manual labor has been completed by migrant workers who are sometimes undocumented. These workers are likely to be missed in most of our data sources. The percentage of hispanic people was influential in our unsupervised learning results, based on our PCA analysis. To get a true understanding of the people who live and work in California, and definitely before making any policy decisions based on migration, it would be useful to speak with domain experts and actual workers to understand this population better. Potentially, we might even be able to find a data source on this population.

Statement of Work

Kim sourced the data and performed data manipulation for the California crime, weather, voter registration, coastal, economic, and school expenditure data sets. She also performed the unsupervised learning project work. Liz provided a bulk of the research during the problem formulation stage. She sourced the data and performed data manipulation for the ACS, geographic, and FEMA data. She performed exploratory data analysis to look for inconsistencies, identify initial correlations and identify the best way to approach the inflow/outflow comparison. Liz also performed a simplified supervised learning algorithm on inflow/outflow numbers alone, to compare against the results with the complete dataset.

Ben provided data manipulation on the IRS dataset. He set up the data for the supervised learning section, selected models, evaluated model performance, and analyzed fields of significance.

All three team members consolidated information and wrote the project report.

Appendix A: Data Sources

IRS Tax Migration Data

- Unique Fields: individual_inflow, individual_outflow
- Description: Measurement of the number of personal exemptions claimed, approximating the quantity of individuals moving in and out of counties on a yearly basis.

CA County Shapefile - County level geographic data

American Community Survey (ACS) Data

- Unique Fields: total_population, median_income, median_rent, median_home_value, total_housing_units, educational_attainment, av_commute_time, perc_poverty, perc_white, perc_black, perc_american_indian, perc_asian, perc_hawaiian, perc_other_race, perc_hispanic, perc_65_over, perc_enrolled_undergrad, perc_owner, perc_renter, perc_vacant, perc_unemployed
- Description: demographic information from the Census Bureau aggregated by county and year

Federal Emergency Management Agency (FEMA) Data

- Unique Fields: chemical, coastal_storm, dam_levee_break, earthquake, fire, flood, hurricane, mud_landslide, severe_ice_storm, severe_storm, snow, tornado, toxic_substances, typhoon, volcano
- Description: Created a boolean variable to represent whether a specific natural disaster occured in a given county each year.

California School Expenditures Data

- Unique Fields: avg_daily_attendance_expense
- Description: This field measures the average yearly cost per student per county. It divides the total expenditures for the year by the “average daily attendance” which is the total count of student attendance days divided by the total days of instruction.

California Weather Data

- Unique Fields: avg_temp, min_temp, max_temp, total_precipitation
- Description: Weather data per county per year.

California Voter Registration Data

- Unique Fields: registered_pct, democrat_pct, republican_pct, american_independent_pct, green_pct, liberterian_pct, peace_and_freedom_pct, no_party_pct, other_pct
- Description: This source provides voter registration data for registered voters during election years. 2014 data was also used for 2015, 2016 data was also used for 2017, and 2018 data was also used for 2019. Note that the files retrieved were one page pdf files that were copied into text files and then regex was used to isolate the columns in the data.

Bureau of Economic Analysis (BEA) Data

- Unique Fields: bea_number_of_people, per_capita_personal_income, per_capita_retirement_and_other, per_capita_unemployment_ins_comp, per_capita_curr_dollar_real_gdp, per_capita_num_jobs, per_capita_farm_proprieter_jobs, per_capita_nonfarm_proprieter_jobs
- Description: This source contains gross domestic product data, job data, and income data for counties by year. All states were available, but we are only using California for this analysis

California Crime Data

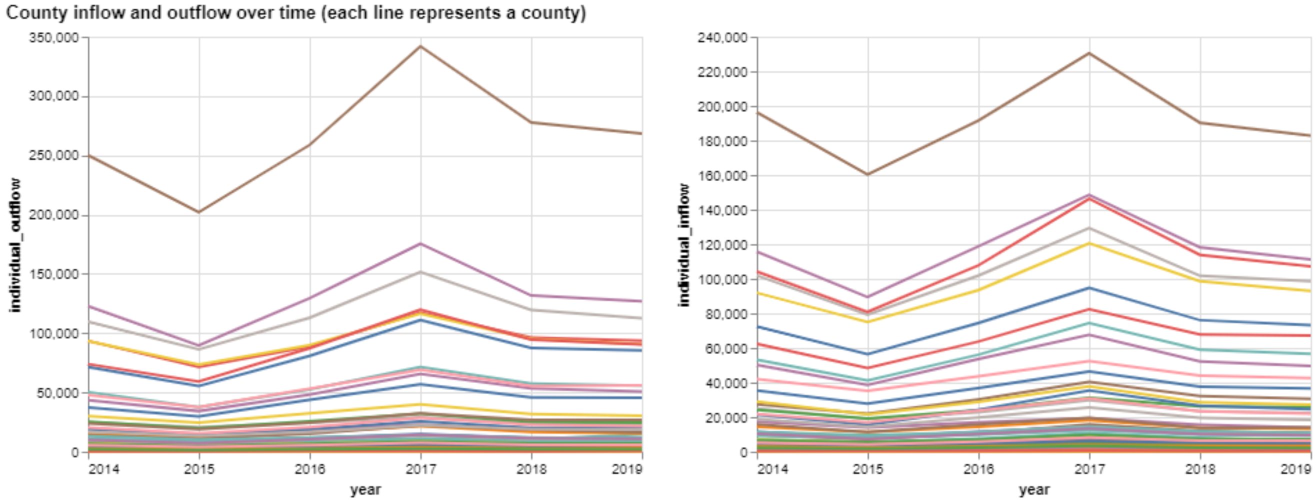
- Unique Fields: num_violent_crimes, num_homicide, num_rape, num_robbery, num_agg_assault, num_property_crimes, num_burglary, num_vehicle_theft, num_larceny_theft, num_arson
- Description: Major crime category counts by county by year. These numbers were converted to per capita values using the BEA number of people variable.

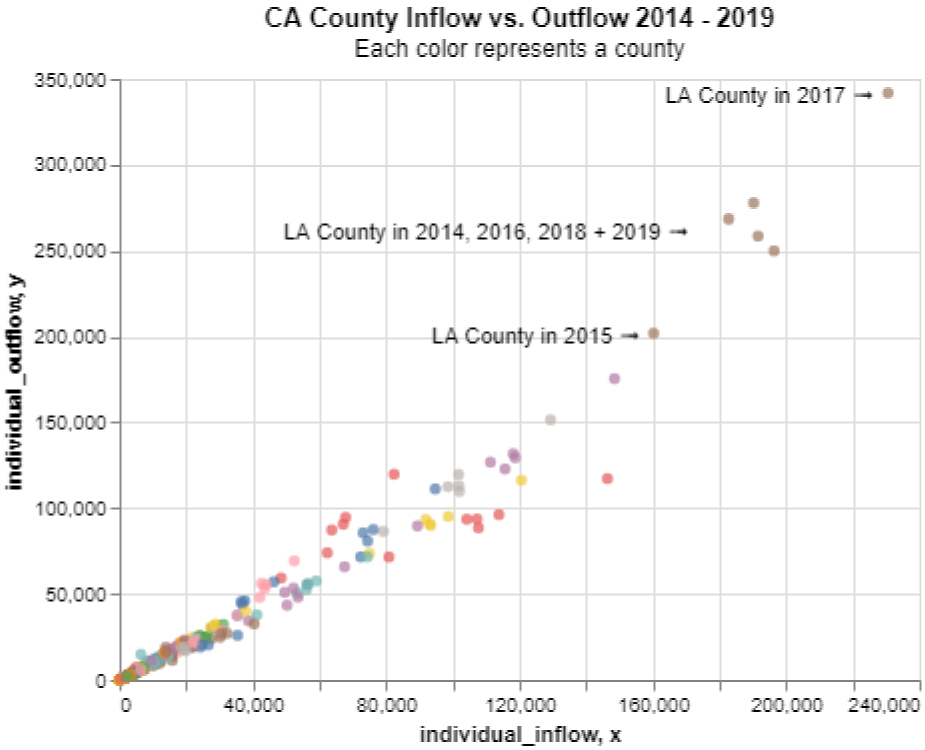
California Coast Data

- Unique Fields: coastal_flag
- Description: This source was created manually using the linked Wikipedia article. It provides an indicator of whether each California county is coastal or not.

Appendix B: Exploratory Data Analysis

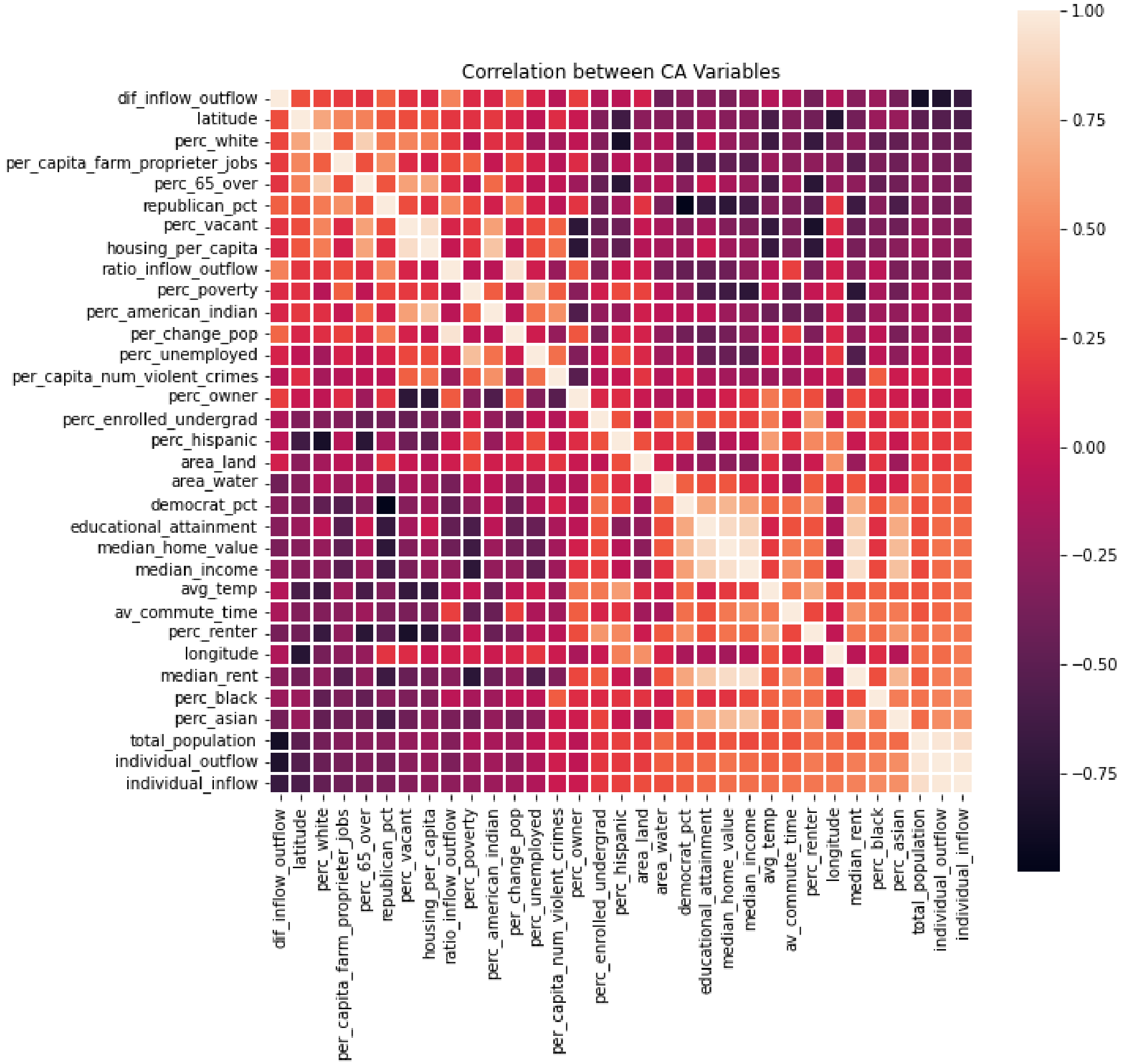
Here are the side-by-side inflow and outflow graphs over time. We can see the similarity between these two graphs across all counties.



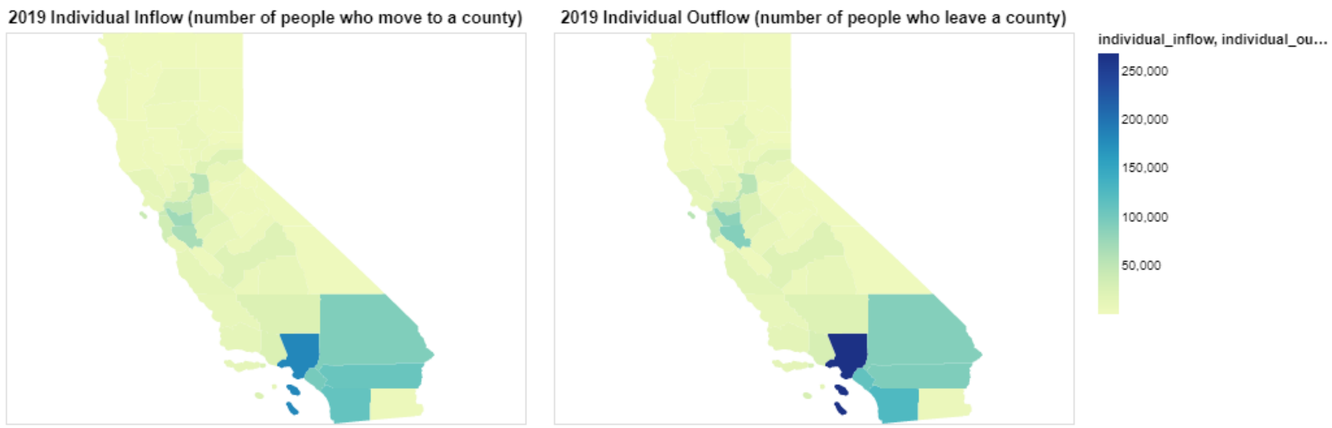


In the graph to the right, we see that there is a strong positive correlation between inflow and outflow for all counties across years. When we turn our attention to color, which represents counties, we also see that, generally, a county's inflow and outflow are most similar to that same county's inflow and outflow for other years. For example, LA county's migration data, in brown, is most similar to its own migration patterns for different years. This led us to hypothesize that the next year's inflow or outflow can be accurately predicted using only data about a county's inflow and outflow in earlier years. Though we have over 70 variables about California counties, we anticipate that we could create similarly accurate models without using such a large amount of data in the supervised learning task.

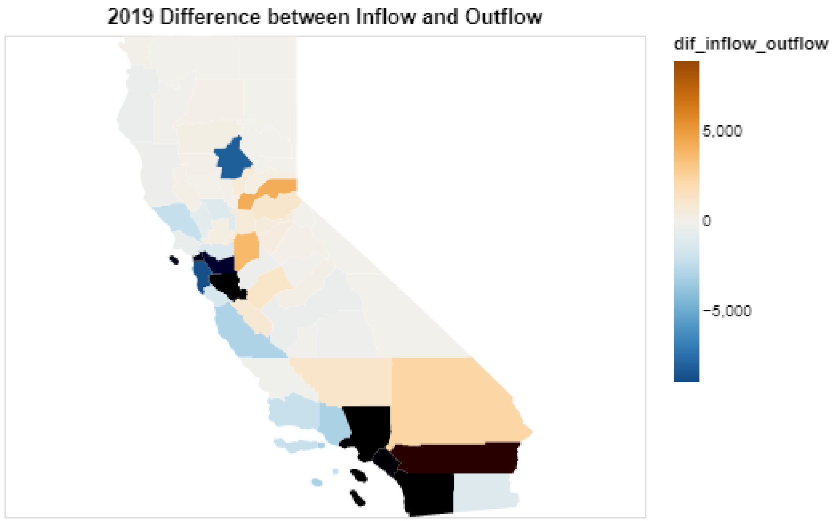
Since there were over 70 variables about CA counties, creating a heatmap containing the correlation between all of the variables would be impossible. Instead, we chose a subset to represent the important features in our data. The variables in the map are sorted from most negative correlation to most positive correlation with the variable inflow. Here we see that the difference between inflow and outflow, latitude, percent white, farm proprietor jobs per capita, percent over 65, and percent republican, and percent of housing units that are vacant are all strongly negatively correlated with both inflow and outflows. We also see that inflow and outflow are strongly positively correlated with each other, and the total population. They have a more moderate positive correlation with many variables, including percent Asian, percent black, and median rent. Counter to our expectations, the percent enrolled in college only has a weak positive correlation with inflow and outflow.



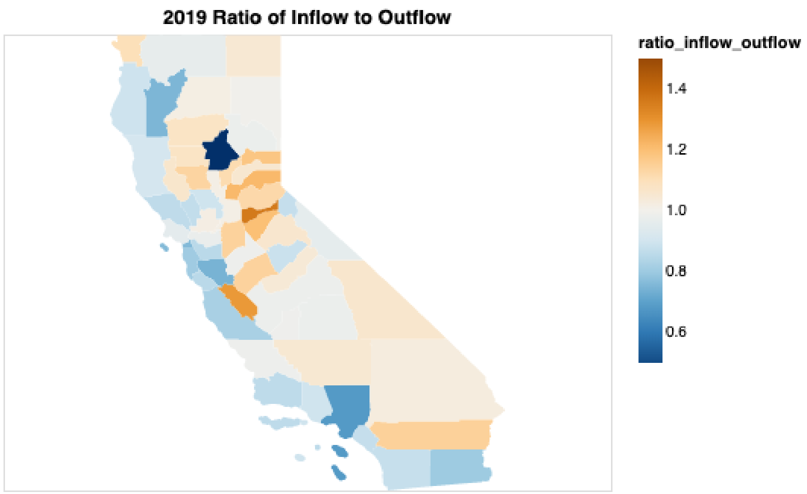
There are many ways to compare counties and their relative inflow and outflow. We could take the ratio of inflow to outflow, the net inflow (inflow minus outflow), or the percent change in population, (inflow - outflow)/population. Each method will yield a different ranking for cities with the highest or lowest migration. This is an investigation of the variation and similarity between these methods. If we want larger counties to hold a larger weight in our model than smaller counties, using raw numbers is most logical (inflow, outflow, or inflow - outflow). If instead, we want all counties to hold equal weight in the model, normalized versions of these metrics are more logical (percent change in population or ratio of inflow to outflow). Here is a look at the similarities and differences between the methods.


2019 Individual Inflow (number of people who move to a county)
2019 Individual Outflow (number of people who leave a county)
individual_inflow, individual_ou...
250,000
200,000
150,000
100,000
50,000

When we view inflow and outflow geographically, we see that the majority of CA counties have individual inflow and outflow below 40,000. This map displays patterns that could also be detected if we color-coded counties by population. The high-population counties, including those in the bay area and southern California, have higher inflow and outflow than smaller ones.

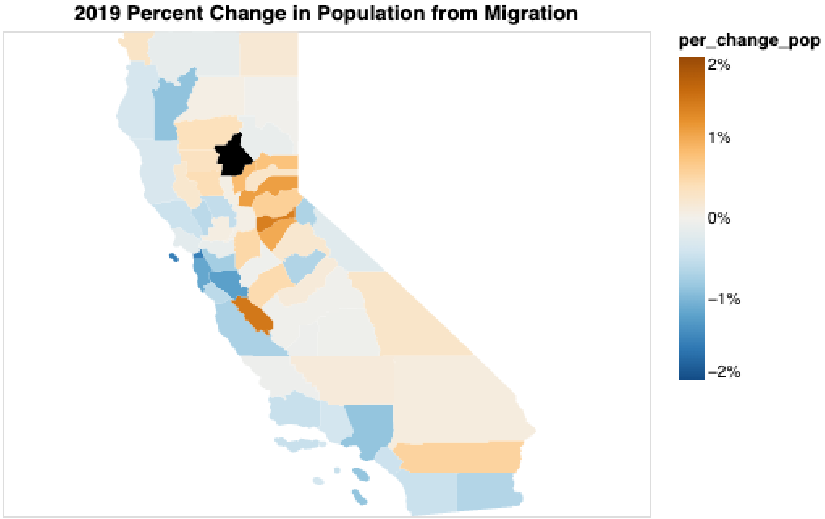


If we instead look at the difference between inflow and outflow, more interesting patterns emerge. Notably, coastal counties have net negative migration with very large negative differences for those in the Bay area, LA, and other coastal southern California counties. We also notice that the majority of inland counties have very small positive net migration, with those inland of LA and the Bay Area having larger net migration. The one exception to this is Butte County, which is inland and north of San Francisco.



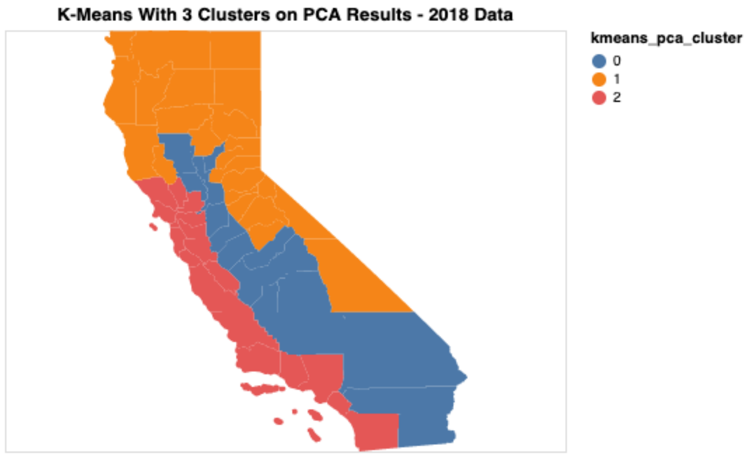
If we turn our attention to the ratio of inflow to outflow, we again notice that Butte county stands out with the lowest ratio of outflow to inflow (0.44). This means that 44 people move to this county for every 100 that leave. Between 2018 and 2019, there was a 3.7% decrease in the population because of migration, leaving it as the county with the sharpest decrease in population, far outpacing the next closest county of San Francisco, which had a 1.5% decrease in population. This is largely due to the 2018 wildfire that hit

Paradise, a town within the county. This may also account for the large increase in population in some surrounding counties including Yuba (0.8% increase), and Placer (1% increase). The correlation between the ratio of inflow to outflow and the percent change in population is 0.96. Since they are so similar, the map of the percent change of population looks nearly identical. For our supervised learning task, we decided to predict the actual outflow and inflow, thus weighting the larger counties higher in our model. We imagined that if a CA treasurer were to use this data, they would want to more accurately capture changes in larger counties.

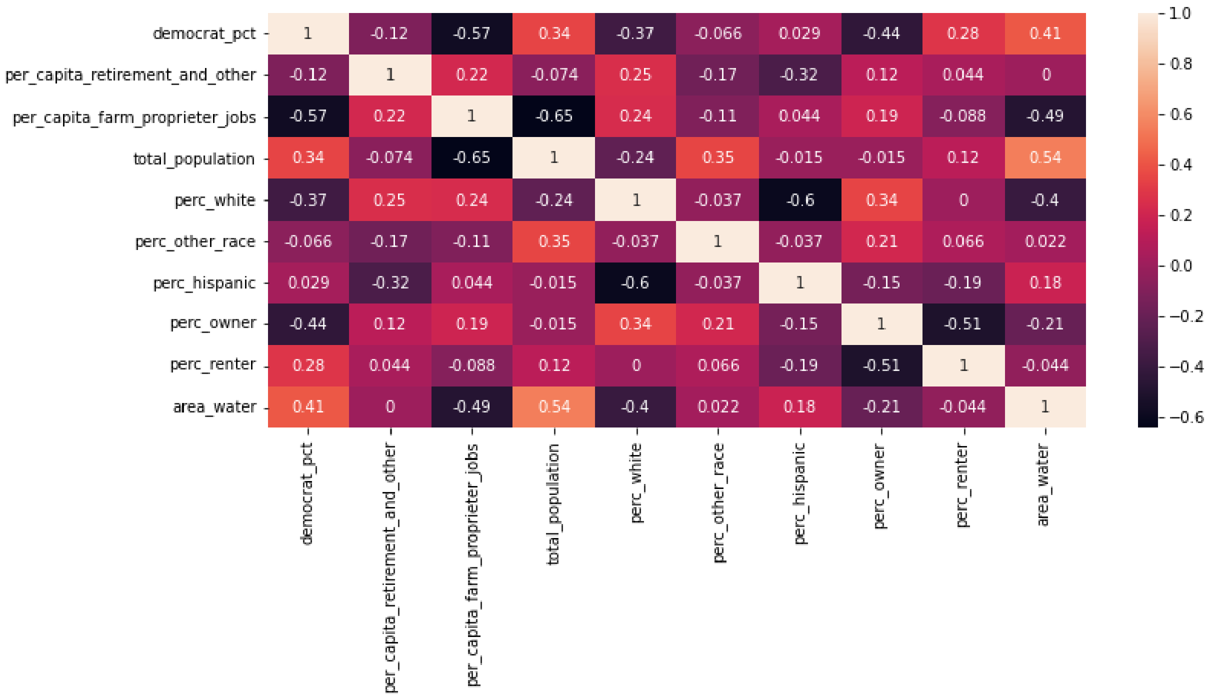


Appendix C: Correlation within Clusters

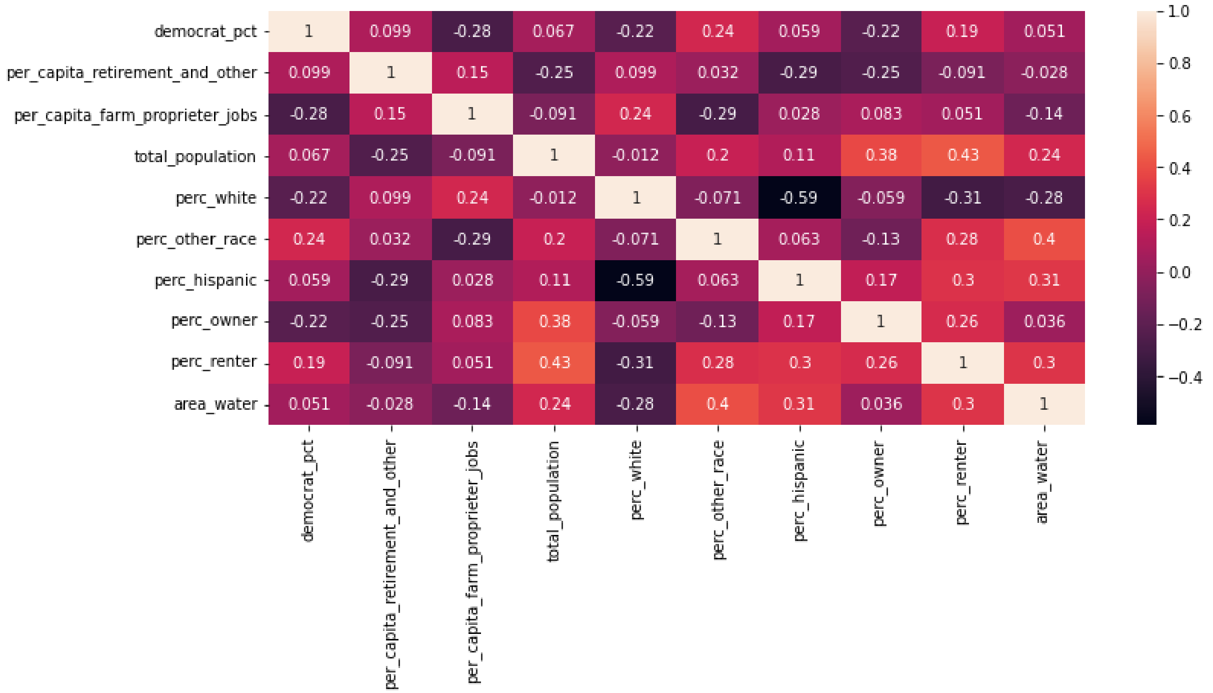
We ran a correlation analysis for each cluster identified in the K-Means after PCA clustering, using the variables identified as most influential in supervised learning:



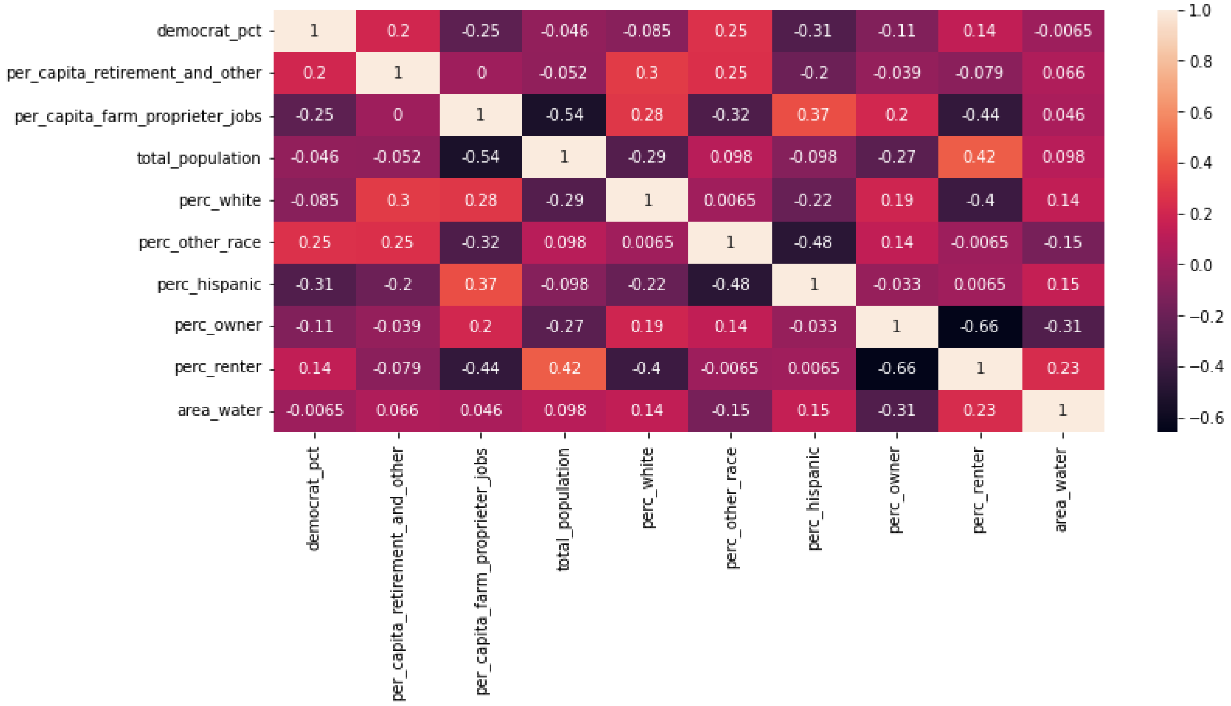
Cluster 0



Cluster 1

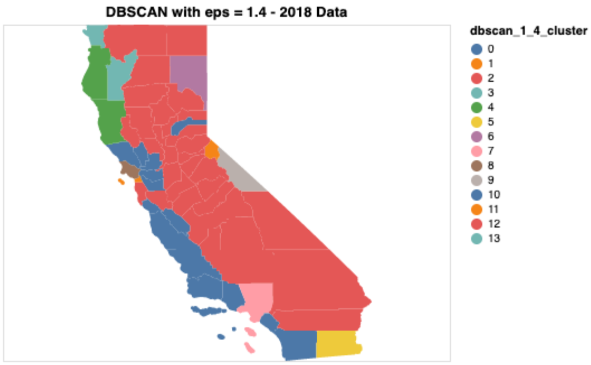

democrat_pct 1 0.099 -0.28 0.067 -0.22 0.24 0.059 -0.22 0.19 0.051
per_capita_retirement_and_other 0.099 1 0.15 -0.25 0.099 0.032 -0.29 -0.25 -0.091 -0.028
per_capita_farm_proprieter_jobs -0.28 0.15 1 -0.091 0.24 -0.29 0.028 0.083 0.051 -0.14
total_population 0.067 -0.25 -0.091 1 -0.012 0.2 0.11 0.38 0.43 0.24
perc_white -0.22 0.099 0.24 -0.012 1 -0.071 -0.59 -0.059 -0.31 -0.28
perc_other_race 0.24 0.032 -0.29 0.2 -0.071 1 0.063 -0.13 0.28 0.4
perc_hispanic 0.059 -0.29 0.028 0.11 -0.59 0.063 1 0.17 0.3 0.31
perc_owner -0.22 -0.25 0.083 0.38 -0.059 -0.13 0.17 1 0.26 0.036
perc_renter 0.19 -0.091 0.051 0.43 -0.31 0.28 0.3 0.26 1 0.3
area_water 0.051 -0.028 -0.14 0.24 -0.28 0.4 0.31 0.036 0.3 1
democrat_pct
per_capita_retirement_and_other
per_capita_farm_proprieter_jobs
total_population
perc_white
perc_other_race
perc_hispanic
perc_owner
perc_renter
area_water
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

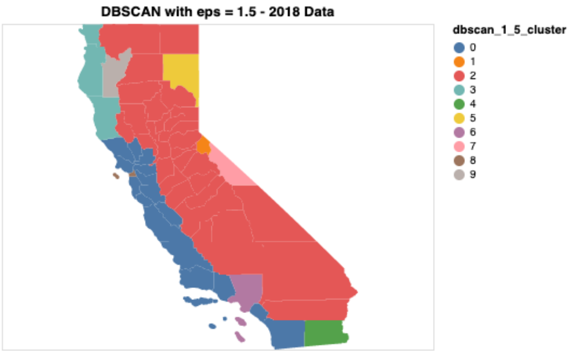
Cluster 2


democrat_pct 1 0.2 -0.25 -0.046 -0.085 0.25 -0.31 -0.11 0.14 -0.0065
per_capita_retirement_and_other 0.2 1 0 -0.052 0.3 0.25 -0.2 -0.039 -0.079 0.066
per_capita_farm_proprieter_jobs -0.25 0 1 -0.54 0.28 -0.32 0.37 0.2 -0.44 0.046
total_population -0.046 -0.052 -0.54 1 -0.29 0.098 -0.098 -0.27 0.42 0.098
perc_white -0.085 0.3 0.28 -0.29 1 0.0065 -0.22 0.19 -0.4 0.14
perc_other_race 0.25 0.25 -0.32 0.098 0.0065 1 -0.48 0.14 -0.0065 -0.15
perc_hispanic -0.31 -0.2 0.37 -0.098 -0.22 -0.48 1 -0.033 0.0065 0.15
perc_owner -0.11 -0.039 0.2 -0.27 0.19 0.14 -0.033 1 -0.66 -0.31
perc_renter 0.14 -0.079 -0.44 0.42 -0.4 -0.0065 0.0065 -0.66 1 0.23
area_water -0.0065 0.066 0.046 0.098 0.14 -0.15 0.15 -0.31 0.23 1
democrat_pct
per_capita_retirement_and_other
per_capita_farm_proprieter_jobs
total_population
perc_white
perc_other_race
perc_hispanic
perc_owner
perc_renter
area_water
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

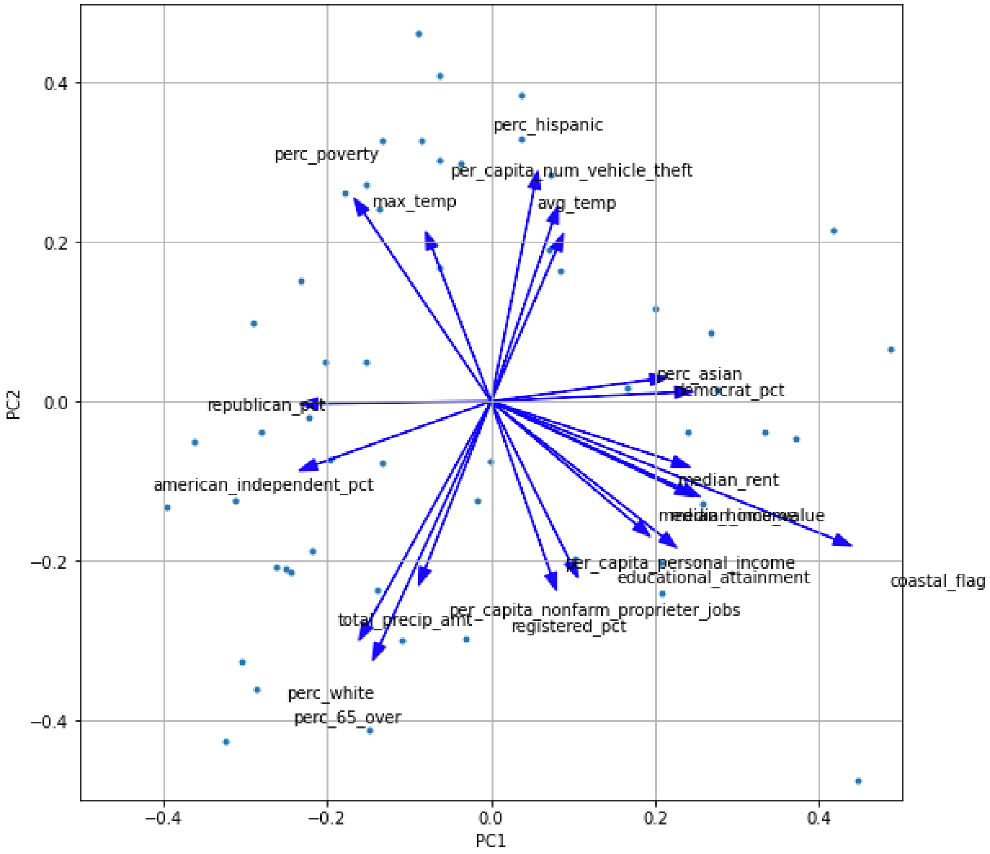
Appendix D: Additional Unsupervised Charts

DBSCAN Cluster Results





Biplot Containing Top 10 Features From Principal Components 1 & 2



Work Cited

“Alpine County, California.” *Wikipedia*, https://en.wikipedia.org/wiki/Alpine_County,_California. Accessed 27 January 2022.

“California.” Wikipedia, https://en.wikipedia.org/wiki/California. Accessed 28 January 2022.

Chabria, Anita. “Paradise, devastated by 2018 Camp fire is, threatened again.” *Los Angeles Times*, 9 September 2020, https://www.latimes.com/california/story/2020-09-09/paradise-devastated-by-californias-deadliest-fire-again-threatened-by-new-blazes. Accessed 28 January 2022.

DeWaard, Jack, et al. “User Beware: Concerning Findings from the Post 2011–2012 U.S. Internal Revenue Service Migration Data.” *Population Research and Policy Review*, 2021, https://doi.org/10.1007/**s11113-021-09663-6.**