John Shutler

Declan Molony

Data 301

Professor Bodwin

***The Influence of Taxes on State Migration***

1. ***Intro***

US states govern themselves on a number of issues, such as, tax rates. We wanted to know, *to what extent do state tax rates have an impact on inter-state migration?* For example, is there a relationship between a state having relatively high tax rates and people leaving the state?

An event dubbed “[The Great California Exodus](https://www.wsj.com/video/series/journal-editorial-report/wsj-opinion-the-great-migration-out-of-california/AD5A0538-8173-450D-8893-93C5B02580DD)” piqued our interest because it describes the hundreds of thousands of Californians leaving the state each year. One of the reasons cited for leaving the state is high taxes. So which tax rates account for the majority of state revenue?

By far, state income and sales tax rates account for most of the state’s revenue. A third category we considered was property tax, but property tax was mostly a source of revenue for local governments (ie: counties) rather than state governments. We are going to use historical income and sales tax data from the last 15 years to predict the state migration patterns going into and out of states.

1. ***Data Collection and Cleaning***
   * ***Income Tax Brackets Data***

Income tax brackets were web scraped from <https://www.tax-brackets.org/> using Python’s Requests and BeautifulSoup libraries. To find a particular state and year, the format was the same for the url. For example, North Carolina tax brackets for 2015 could be found at <https://www.tax-brackets.org/northcarolinataxtable/2015>. In total, we web scraped 765 website url’s (51\* 15, we included the District of Columbia #DC should be a state). There was an issue with a few states and years. All of South Dakota was missing from <https://www.tax-brackets.org/>, but fortunately it is a no income tax state, and the data was easily entered. There were five other state and year combinations (e.g. California 2009) that the website was missing. We had to go find that state and year’s income tax data from their respective state government websites.

Cleaning the income tax data was relatively simple. It involved removing non-numeric characters to convert a few variables to a numeric type (ex: “$14,500.00+” to 14500).

* + ***Sales Tax Data***

We were able to find a multi-tabbed excel file [here](https://www.taxpolicycenter.org/statistics/state-individual-income-tax-rates-2000-2020) that contained the sales tax rates for states by year. Additionally, it was missing the years of 2005 and 2009. In order to find those years, we had to look at government websites for [2005](https://govinfo.library.unt.edu/taxreformpanel/meetings/pdf/Salestax_04182005.pdf) and [2009](https://taxfoundation.org/updated-state-and-local-option-sales-tax/).

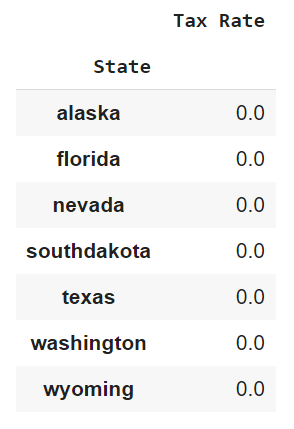
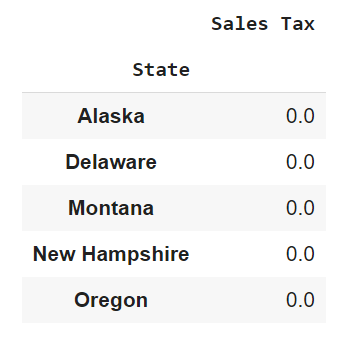
A great attention to detail was given while cleaning the data because it was formatted inconsistently from year to year. Several helper functions were created in aiding the structure of the dataframes we constructed due to inconsistent merged cells, lack of structure, and other details like superscripts in the excel file.

* + ***Migration Data***

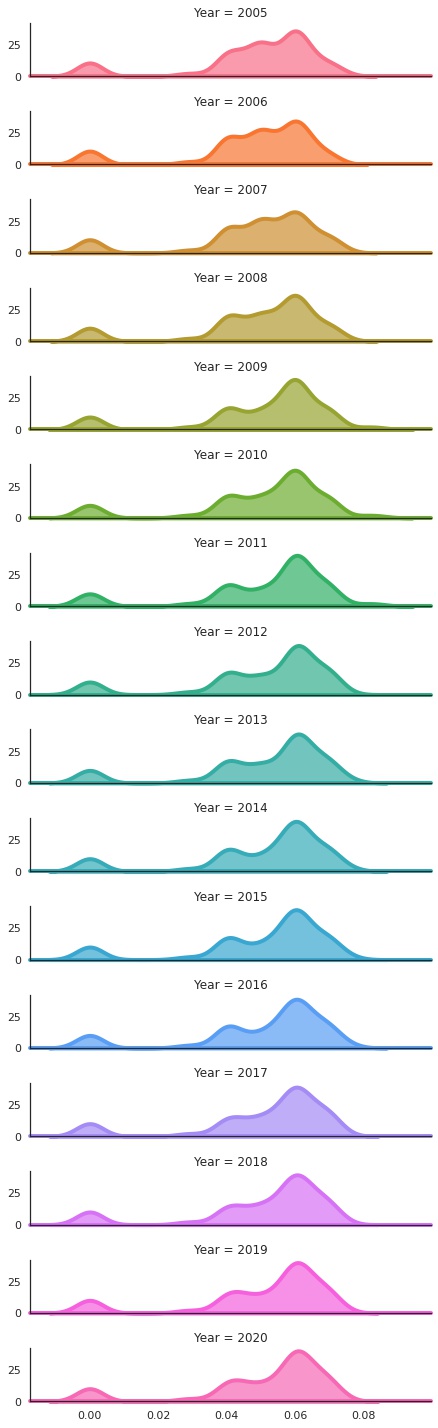
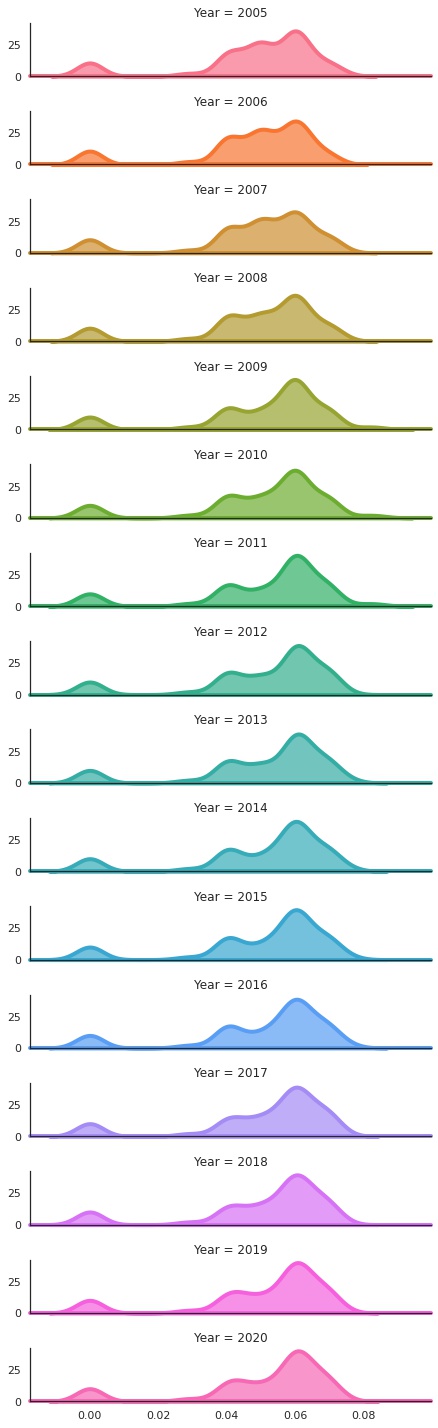
In order to get state by state migration data, we went to the United States Census Bureau where we found this [link](https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html). This site has access to state by state migration data between 2005-2019 contained in an excel file.

The data was formatted poorly. It had merged cells for columns and for readability purposes, it repeated the index column every 11th column. We had to remove all repeated indices, extrapolate the data in the merged cells to the column we cared about, remove all of the extraneous filler row the US government felt necessary to add to its files. Furthermore, the formatting was slightly different for each year, so we had to find the few common patterns that were similar across all years to make sure we could clean all excel files into a common format. After that we combined the data frames for each year into one main dataframe with the years included as an extra column.

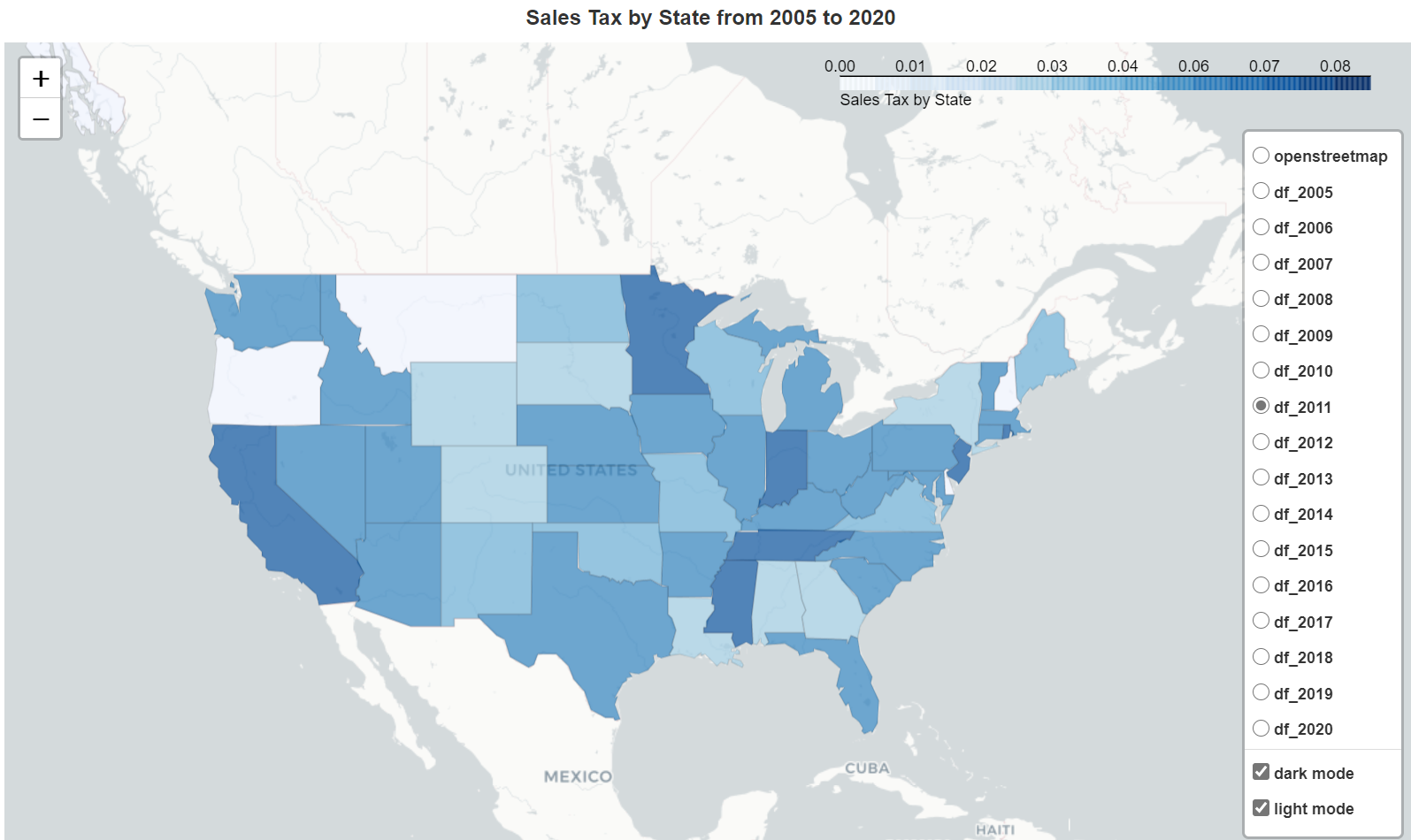
1. ***Exploratory Analysis***



These two tables show the states that do *not* have sales or income tax, respectively. Alaska is the one state that has neither, which made it difficult to predict migration patterns for.

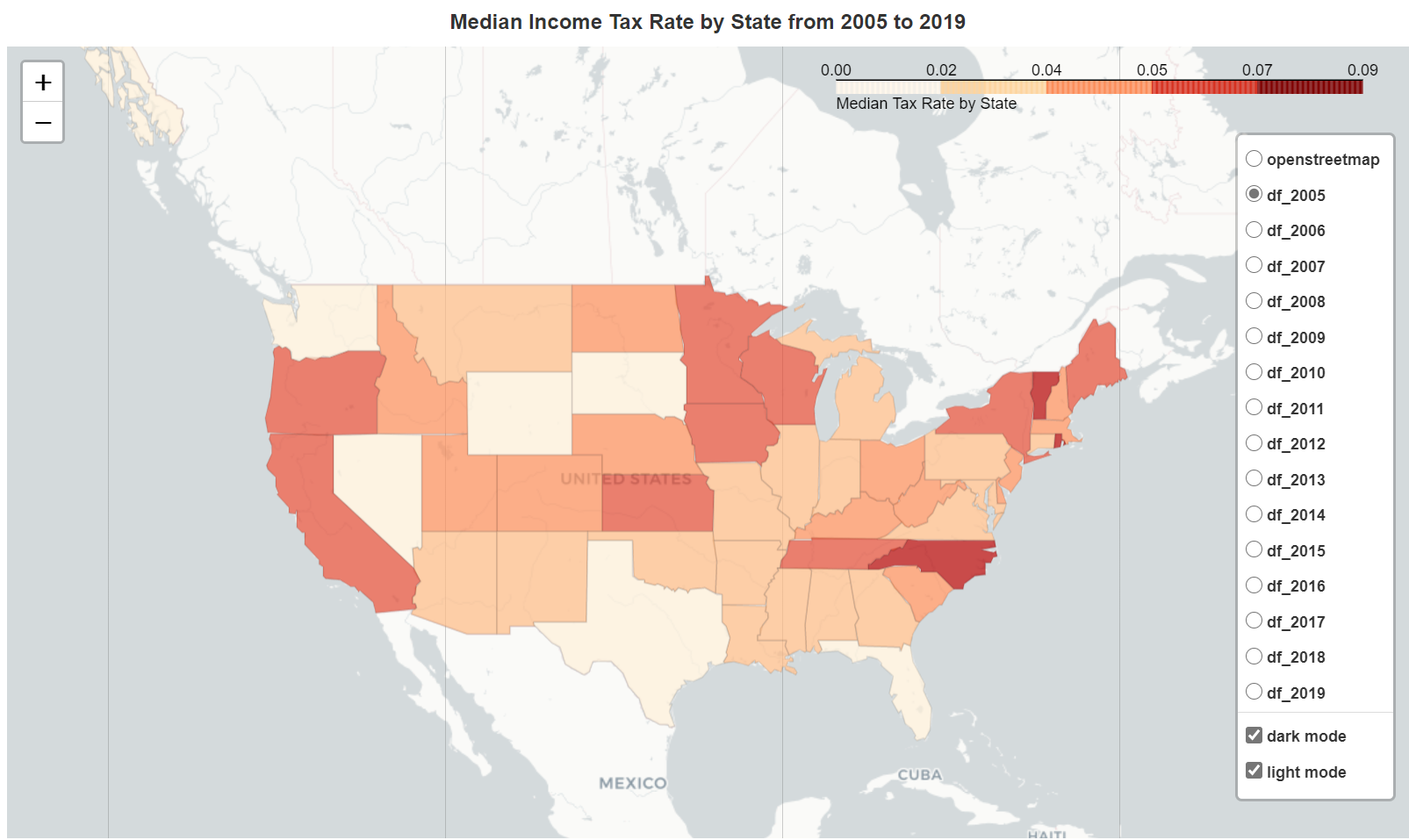


This facet graph shows that in the 2000's, sales tax has bimodal distribution. Going into the 2010's it slowly transitions to trimodal distribution. The small 'bump' on the left of each facet represents the group of states with no Sales Tax.



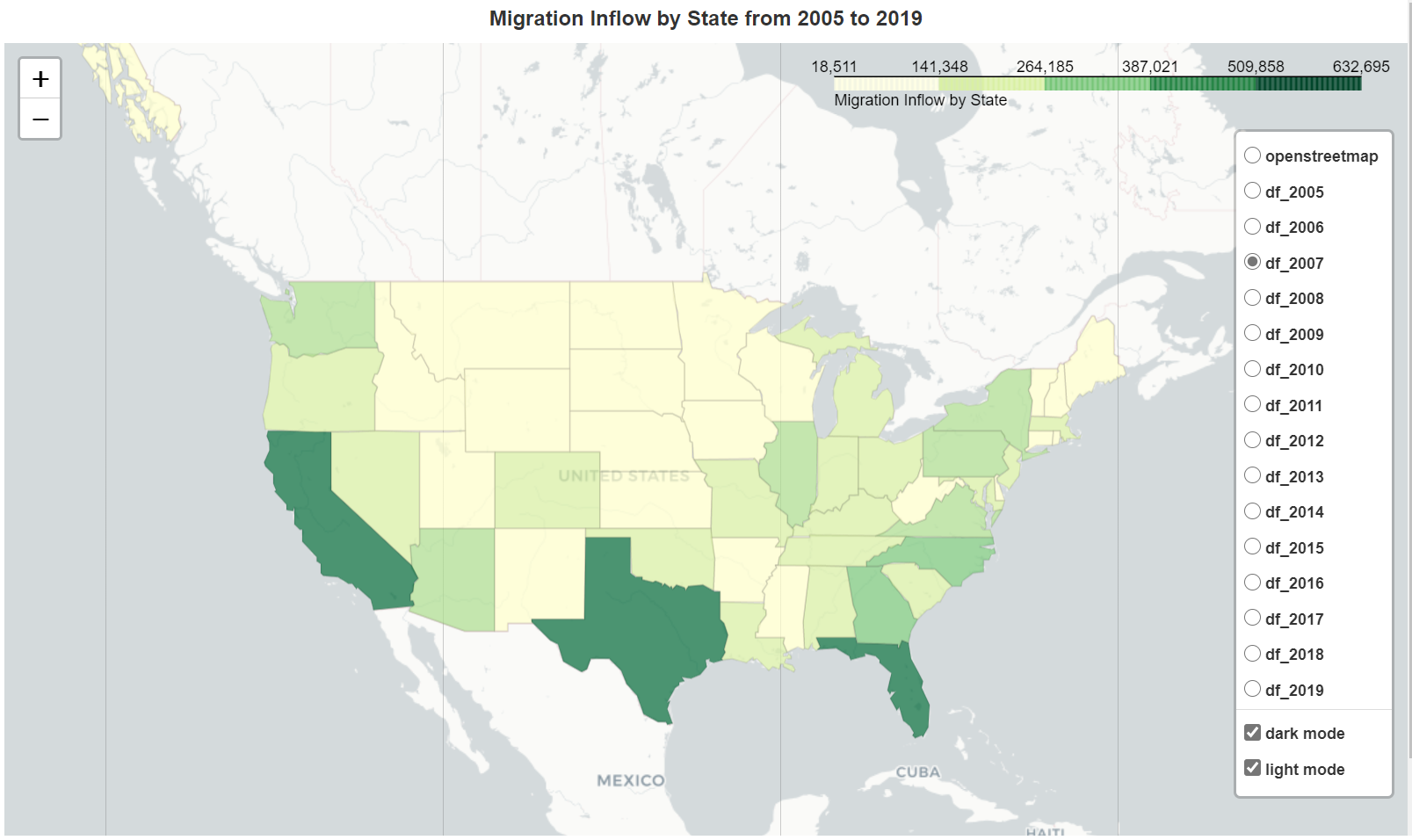
This choropleth illustrates how the Sales Tax changes in each state by year. It is interactive and in the “Data Exploration” google collab, one can filter by year.

There are two interesting observations from this graph. (1) Sales taxes from 2009-2011 seem to be particularly low compared to the rest of the years. This could be a response by state governments to the Great Recession during those years. (2) Sales tax tends to not vary greatly each year. It's mostly constant across each state, with increases or decreases every couple years.

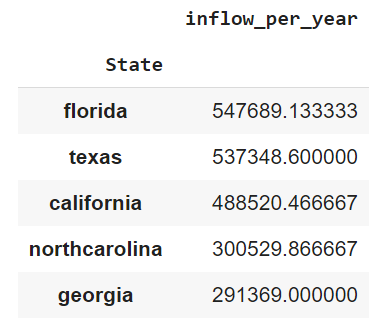


This choropleth illustrates how the median Income Tax changes in each state by year. It is interactive and in the “Data Exploration” google collab, one can filter by year.

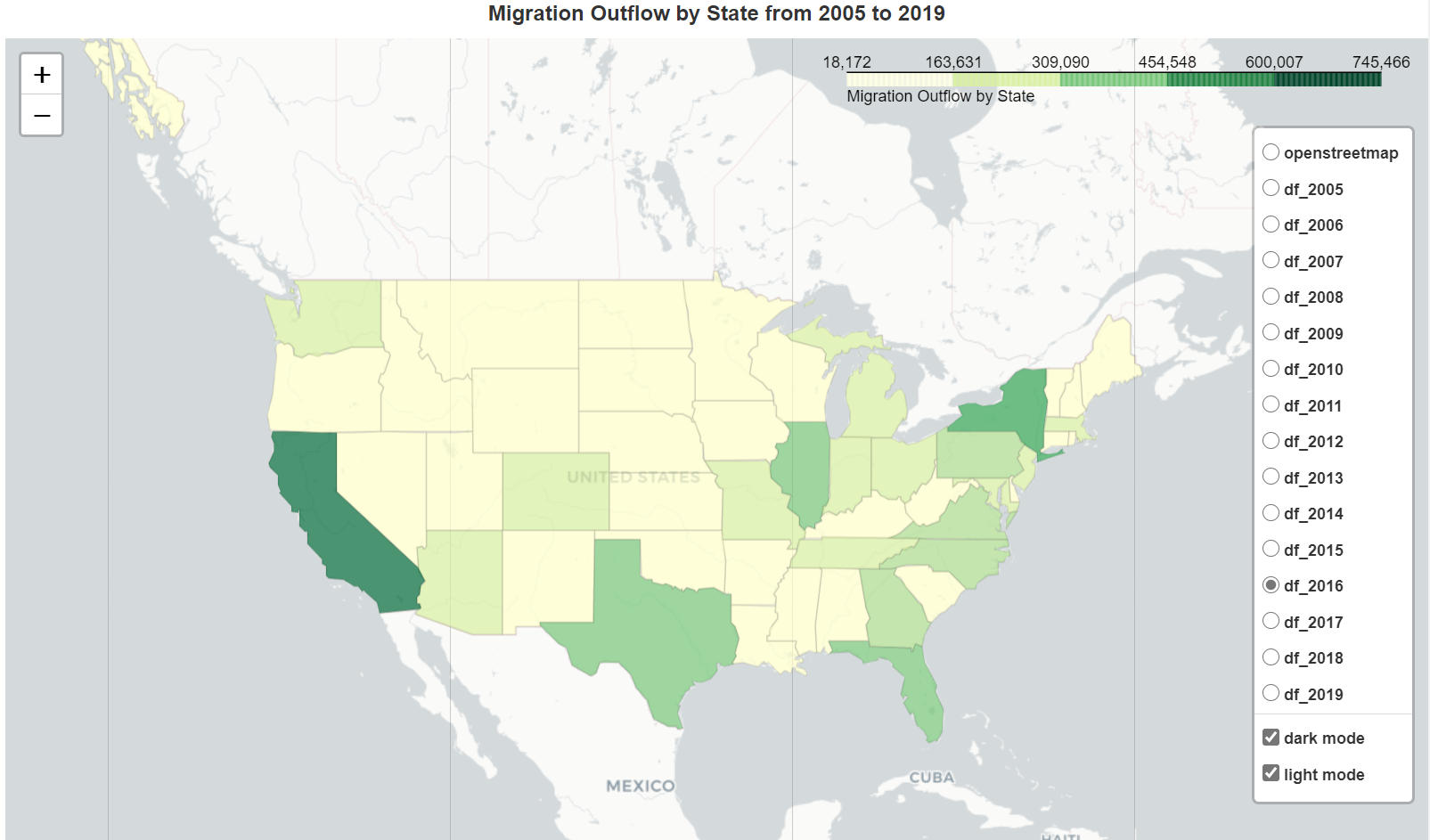
The observations based on this graph is that Income Tax median rates vary a great deal more than the Sales Tax Rates.



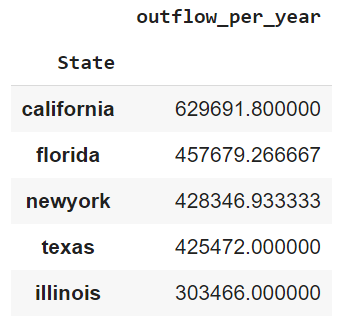
This choropleth illustrates Migration Inflow of each state by year. It is interactive and in the “Data Exploration” google collab, one can filter by year.



If you look at any year, the three states by far with the most migration inflow are California, Texas, and Florida. The table confirms those are in the top five States regarding mean inflow migration per year, along with North Carolina and Georgia.

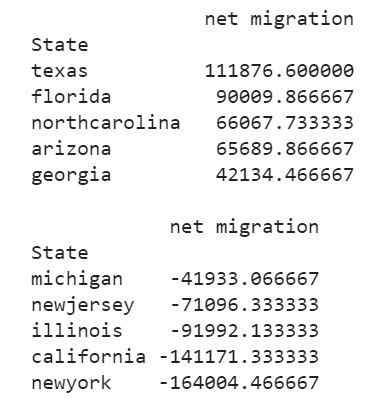


This choropleth illustrates Migration Outflow of each state by year. It is interactive and in the “Data Exploration” google collab, one can filter by year.



In terms of mean outflow migration per year, California, Florida, and Texas are still in the top 5. But the other states people are leaving are New York and Illinois.

Now hold on, if California, Texas, and Florida are in the top 3 for both inflow and outflow migration, what does that really mean? It means we need to take a look at net migration data, by state.



Here we see the top 5 states, on average, with the most net migration: Texas, Florida, North Carolina, Arizona, and Georgia. These states all have net positive migration. Compare those states to the bottom 5, which all have net negative migration. Those states are Michigan, New Jersey, Illinois, California, and New York.

There is a huge disparity between the top and bottom 5. About 111,000 people moved to Texas, on average, over the last 15 years. Meanwhile about 164,000 people moved out of New York, on average, over the same time period.

1. ***Predicting Migration Outflow and Inflow using Linear Regression***

Our variables for each model

* + Intercept
  + Minimum State Income Tax Rate:
  + Maximum State Income Tax Rate
  + Median State Income Tax Rate
  + State Sales Tax

1. Overview of our Models

Before we start, we want to specifically mention why we did not include dummy variables corresponding to each state in our model. This is because just knowing the state tells us the inflow and outflow by itself. In preliminary tests, we found that having only the state dummy variables gives us R^2 of .98, which is extremely high. This is because the state dummy encapsulates specific things about the state, its weather, its housing prices, and of course its tax rates. Meaning however much the tax of a state influences an individual to move/leave there is baked into that one dummy variable.

Knowing this we ran 2 main categories of regressions to predict the outflow and inflow of migration per state

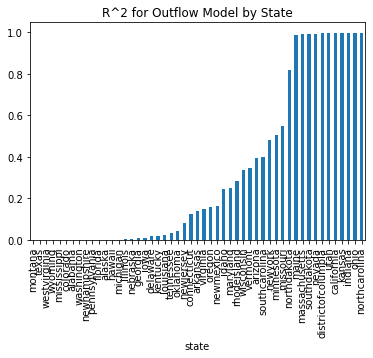
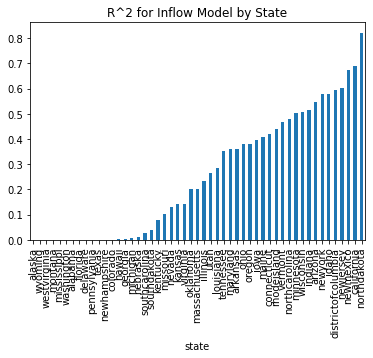
1. Predicting inflow/outflow using all of the state migration/tax data
2. Predicting inflow/outflow one state at a time, meaning we trained one model per state using only that specific states migration and tax data for prediction

For the first model, we got some interesting results depending on whether or not we added an intercept to the model

|  |  |  |
| --- | --- | --- |
| Model | Inflow | Outflow |
| R^2 (without intercept) | 0.602 | 0.609 |
| R^2 (with intercept) | 0.119 | 0.116 |

This shows us that the model has FAR more predictive power when there is no intercept given. However, we thought the interpretation of our coefficients seemed kind of silly if we didn’t include it, a state with no state taxes will never attract or lose it’s citizens? Seems silly to us at least, so we chose to keep the intercept going forward despite it weakening the model’s predictive power.

For the 2nd Model we tested its accuracy by the average R^2 model across all states models so the, we also only tested it with the intercept. Here are some graphs of the predictive power of each model by state:



We found that states that don’t have any taxes, like Texas, had no ability to predict what the rates were, resulting in R^2 values of 0. This is due to us trying to predict the migration of these states, using taxes that don’t exist. Conversely, states with higher tax rates, are easier to predict migrations using tax rates, for example, California has an R^2 of .99 when predicting how many people will leave, and an R^2 of .688 to predict how many people will come to the state in a given year.

B. Models Using All Migration and Tax Data

The first model we created was trying to predict the inflow of a generic state using all of the state’s tax data across all years collected (2005-2019). The output is below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | inflow\_per\_year | **R-squared:** | 0.119 |
| **Model:** | OLS | **Adj. R-squared:** | 0.114 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 1.06e+05 | 1.48e+04 | 7.177 | 0.000 | 7.7e+04 | 1.35e+05 |
| **min\_rate** | 7.844e+05 | 4.83e+05 | 1.624 | 0.105 | -1.64e+05 | 1.73e+06 |
| **max\_rate** | 1.326e+06 | 5.85e+05 | 2.268 | 0.024 | 1.78e+05 | 2.47e+06 |
| **median\_rate** | -3.27e+06 | 9.57e+05 | -3.416 | 0.001 | -5.15e+06 | -1.39e+06 |
| **salestax** | 1.656e+06 | 2.22e+05 | 7.441 | 0.000 | 1.22e+06 | 2.09e+06 |

As you can see, the R^2 on models with the intercept are not high; however, the P-values for all of the coefficients except for the minimum income tax rate are all below .05, showing evidence that they are important when predicting the inflow of a given state per year.

Interpretation of coefficients:

1. Intercept: If a state has no taxes, we predict that there will be 100,000 people trying to enter the state
2. Minimum Income Tax Rate: If a state increases it’s tax rate for the lowest tax bracket by 100%, then we expect 700,000 new people to enter the state)
3. Maximum Income Tax Rate: If a state increases it’s tax rate by 100% for the highest tax bracket by 100%, then we predict 1,326,000 people to enter the state
4. Median Income Tax Rate: If the income tax rate increases by 100% for the median tax bracket, then we predict for 3,270,000 to leave the state.
5. Sales Tax: If the sales tax rate increases by 1% for the state, then we predict for 1,656,000 people to enter the state.

The second model is the same as the first but attempting to predict the migration out of the state rather than into the state. The output is bellow:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | outflow\_per\_year | **R-squared:** | 0.116 |
| **Model:** | OLS | **Adj. R-squared:** | 0.111 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 6.747e+04 | 1.51e+04 | 4.478 | 0.000 | 3.79e+04 | 9.71e+04 |
| **min\_rate** | 1.487e+06 | 4.93e+05 | 3.017 | 0.003 | 5.19e+05 | 2.45e+06 |
| **max\_rate** | 2.912e+06 | 5.97e+05 | 4.880 | 0.000 | 1.74e+06 | 4.08e+06 |
| **median\_rate** | -4.977e+06 | 9.77e+05 | -5.094 | 0.000 | -6.9e+06 | -3.06e+06 |
| **salestax** | 1.761e+06 | 2.27e+05 | 7.758 | 0.000 | 1.32e+06 | 2.21e+06 |

From the output above we can see that R^2 is very similar to that of the inflow model, but the main difference between the two is that all the P-values for all coefficients for this model are below .05, meaning they have statistical significance in this model

Interpretation of coefficients:

1. Intercept: If a state has no taxes, we predict that there will be 60,000 people to leave the state
2. Minimum Income Tax Rate: If a state increases it’s tax rate for the lowest tax bracket by 100%, then we expect 1.487 million new people to leave the state
3. Maximum Income Tax Rate: If a state increases it’s tax rate for the highest tax bracket by 100%, then we predict 2.9 million people to enter the state.
4. Median Income Tax Rate: If the income tax rate increases by 100% for the median tax bracket, then we predict for 4.9 million people to enter the state.
5. Sales Tax: If the sales tax rate increases by 100% for the state, then we predict for 1.76 people to leave the state.

D. Conclusion about these models

The interpretation of the coefficient of these 2 models contradict each other. If you pay close attention, in the first model, we are predicting inflow, and the second we are prediction outflow. However, the signs on all of the coefficients are exactly the same for both models, meaning the interpretations of the two models directly contradict each other. For example, if the median tax rate increases by 100%, then the inflow model predicts that the state will LOSE over 3 million people, but in the exact same scenario, the outflow model predicts the state will GAIN nearly 5 million people. This shows a serious weakness in our model. If we were to try and improve it we would look for more confounding variables that explain some of the movement between states; examples being weather, housing prices, and employment rates. Adding confounding variables would allow our model to further isolate what the main causes of state by state migration are so we can have a more definitive interpretation of our tax coefficients.

1. ***Conclusion***

“In this world nothing is certain but death and taxes” - Benjamin Franklin

No one likes paying taxes, that is tautology if ever there was one, but we wanted to find out to what lengths people would go to not pay their taxes. Through our data exploration we found that states with the largest net migrations out of the state are California and New York, with the highest tax rates; furthermore, we found that Texas, a state with no income tax, has the highest net migration into the state. These simple statistics were evidence that tax rates play a big role in people choosing to move to or out of a given state.

We had hoped that we could use linear regression to find further evidence of this phenomenon, but our model was simply too weak to make any meaningful conclusions on the matter. We believe that we would need to account for other reasons people might choose to move to or from a state: a state’s weather, housing prices, and employment rates being some examples. And if we were to add these to our model, we think we would find more definitive evidence, one way or the other, as to the role taxes play in people’s decision to move states.

We hope you enjoyed reading our report as much as we had making it!