Jiali Zhou, jz2312@nyu.edu

Capital One Labs Coding Challenge

Jiali Zhou, jz2312@nyu.edu

Code Test Part 1: Model building on a synthetic dataset

This project is written in python, and the project interpreter is python 2.7.

The main function is main.py, and all the functions are included in process_model.py.

For processing the dataset, I use import_data function to import data, normalize_data function to deal with missing data, and normalize the data. For missing data, I calculate the median of the data from continuous data, and the most frequently occurring category to represent the categorical data.

For modeling, I use cross-validation to train the model and test on the validation set. I used L1 and L2 regularization to prevent overfitting. For L1 regulariztion, I use AkaShooting algorithm to train the model, and get the square loss at minimum about 12.3. For L2 regulariztion, I use stochastic grad descent to train the model, and get the square loss at mimimum about 13~14. So I choose L1 regularization for training and testing, and output the result in y_test.txt.

All the functions and algorithms are written on my own. And main.py and process_model.py are also included.

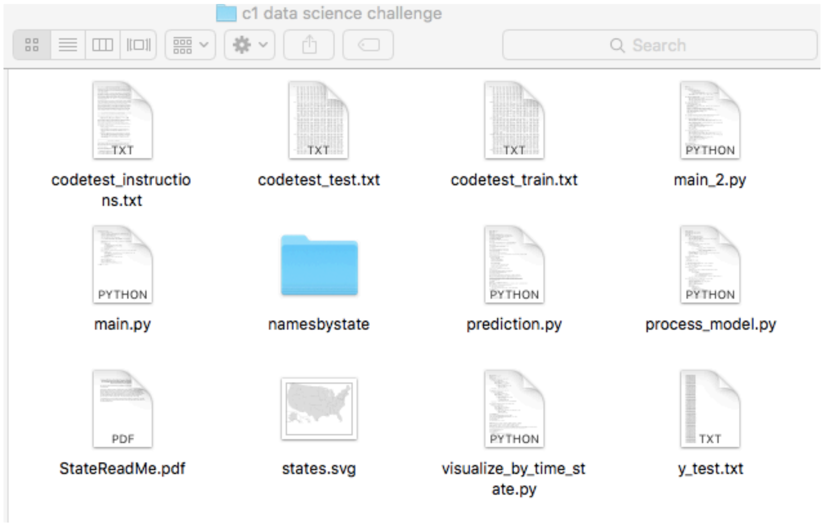
There are also lots of packages in python, like AdaboostRegressor, gradientRegressor which can be used to solve the problem.

Code Test Part 2: Baby Names!

This project is written in python, and the project interpreter is python 2.7.

The main function is main_2.py, and include all the functions to calculate the answers for the questions. For visualization, please run visualize_by_time_state.py, and include the blank states.svg and folder namebystate in the same folder.

Jiali Zhou, jz2312@nyu.edu



1. Please describe the format of the data files. Can you identify any limitations or distortions of the data?

Each record in a file has the format: 2-digit state code, sex (M = male or F = female), 4-digit year of birth (starting with 1910), the 2-15 character name, and the number of occurrences of the name. Fields are delimited with a comma.

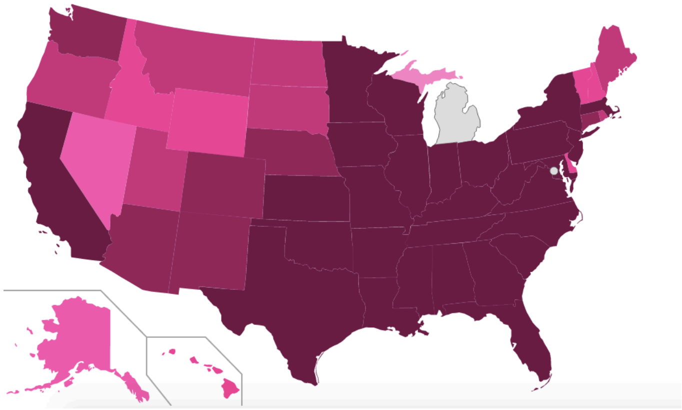
For limitations, list of names to those with are at least 5 occurrences. If a name has less than 5 occurrences for a year of birth in any state, the sum of the state counts for that year will be less than the national count

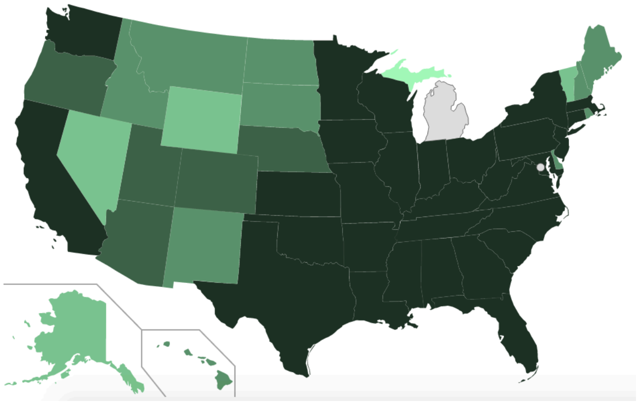
2. What is the most popular name of all time? (Of either gender.)

Running function most_popular_name, I find that the most popular name of all times of Female is Mary, and of Male is James.

For visualization, see below. I get the average counts from 1910-2014 in each state of name Mary and Male and plot in United States map. The darker color, the more counts in each state.

From the visualization, we can see that both names are very popular in more than half the states, mostly from the mid to east states.



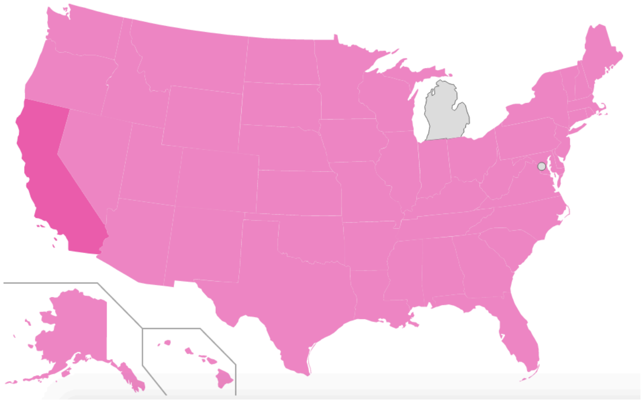


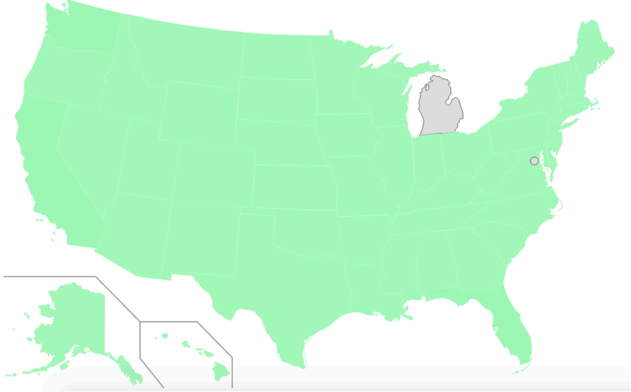
3. What is the most gender ambiguous name in 2013? 1945?

Running ambiguous_name function, we can find that the most ambiguous name in 2013 is Nikita, and in 1945 is Maxie.

Jiali Zhou, jz2312@nyu.edu

For visualization, Nikita 2013 popularity in Female and Male is as below. And for Maxie is almost the same.

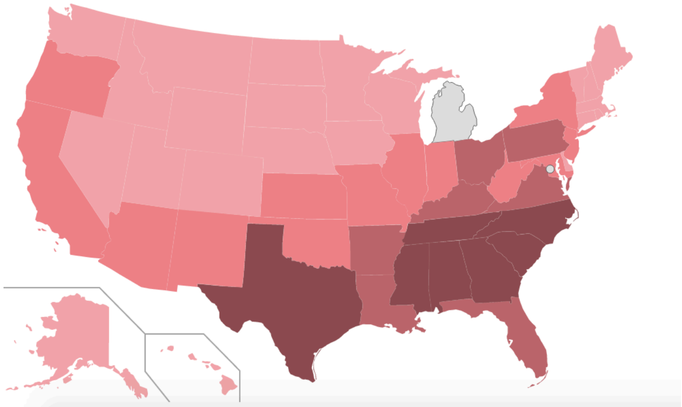


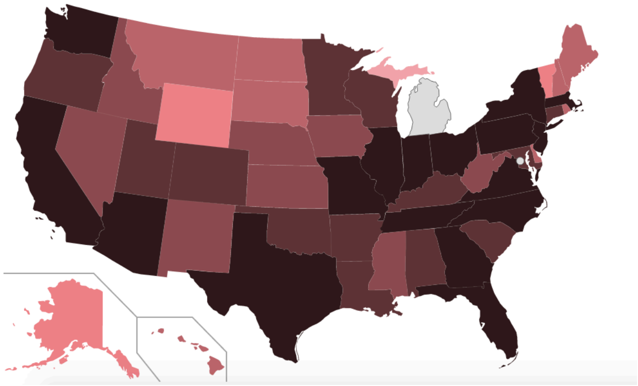


4. Of the names represented in the data, find the name that has had the largest percentage increase in popularity since 1980. Largest decrease?

Running find_largest_increase_or_decrease function, we can find that the name that has had the largest percentage increase in popularity since 1980 is Emma, and largest decrease is Jennifer.

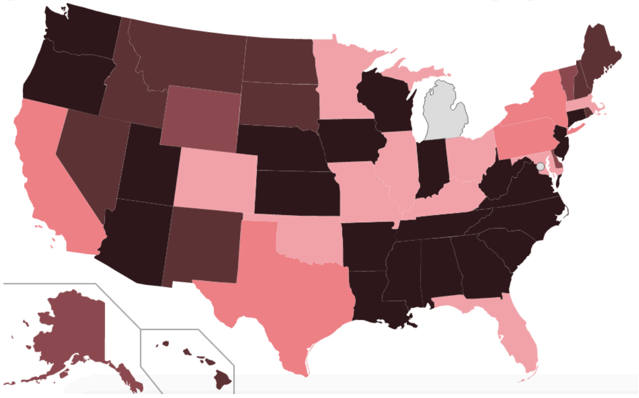
For visualization, the counts in each state for Emma in 1980 and 2014 are as below.

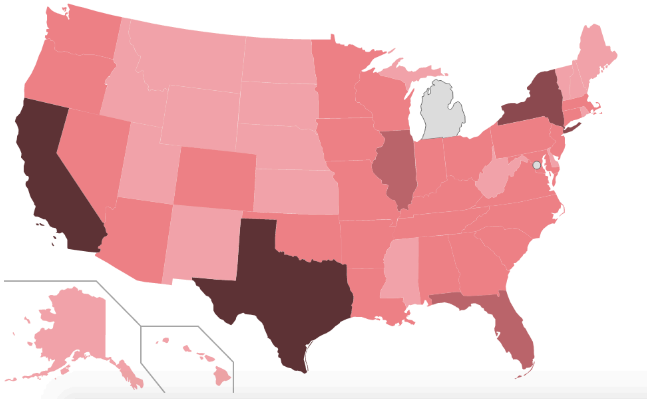




And for Jennifer in 1980 and 2014.

We can find an interesting phenomenon that for Jennifer, the popularity in each state is almost oposite in 1980 and 2014. The more popular in 1980, the less popular in 2014.



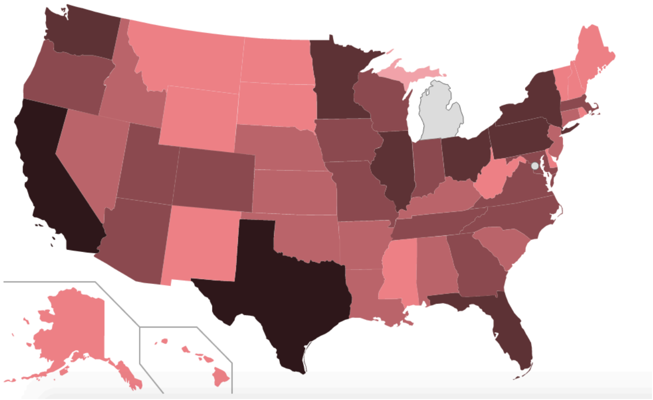


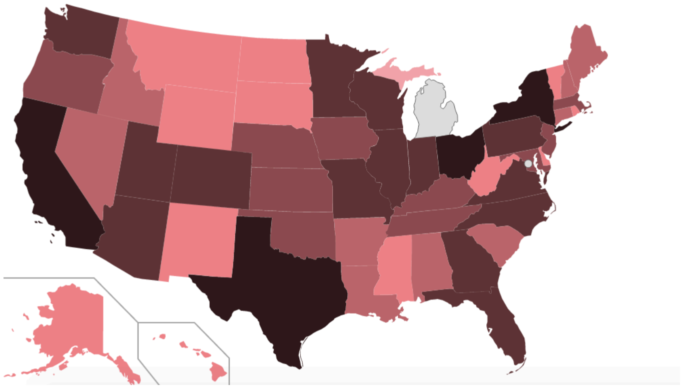
Jiali Zhou, jz2312@nyu.edu

5. Can you identify names that may have had an even larger increase or decrease in popularity?

Running prob_largest_increase_or_decrease function, we can find that the name that will have the largest percentage increase in popularity is Oliver, and will probably have the largest percentage decrease in popularity is Sophia.

For visualization, the popularity in 2013 and 2014 in each state of Oliver is as below. From this, we can find that the states which popularity increases most are UT, AZ, CO and northeast states.





For visualization, the popularity in 2013 and 2014 in each state of Sophia is as below. From this, we can find that the states which popularity decreases most are OR, NV, LA and OK.