Baby names NY State 2007-2012 Analysis

Daniel Toran Mercade

January 2021

Baby Names NY 2007-2012 has many problems due to inconsistent data

Dataset pitfalls

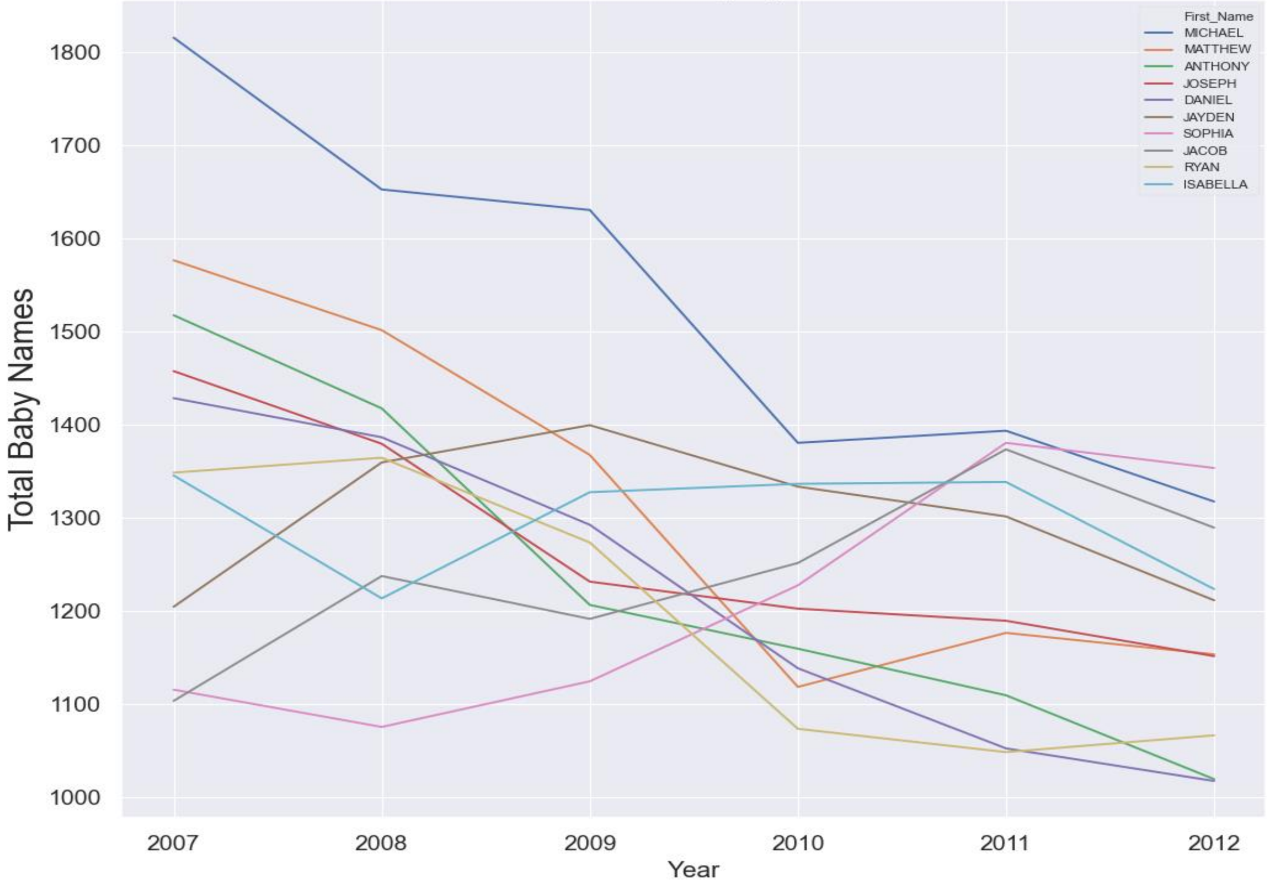
- **Less names per capita:** there are some counties that have much less names per capita than they should have and much less than the rest of the counties.
- **Sudden decrease in 2010:** those counties have also a weird sudden decrease in the baby names registration in the year 2010 (specially those with less baby names)
- **High proportion of male babies :** male babies are represented in around 60% whereas females have 40%. This trend is similar among years and counties

Assumptions

- Use the births_1000_inhabitants variable to exclude those that have less than 3 (KINGS has the highest with 8.87)
- Those states with the sudden decrease can be singled out by using the variable Decrease_pop_2010
- Try to find complementary data since this problem does not seem to come from a mislabeling in the data in names that can be used for both genders

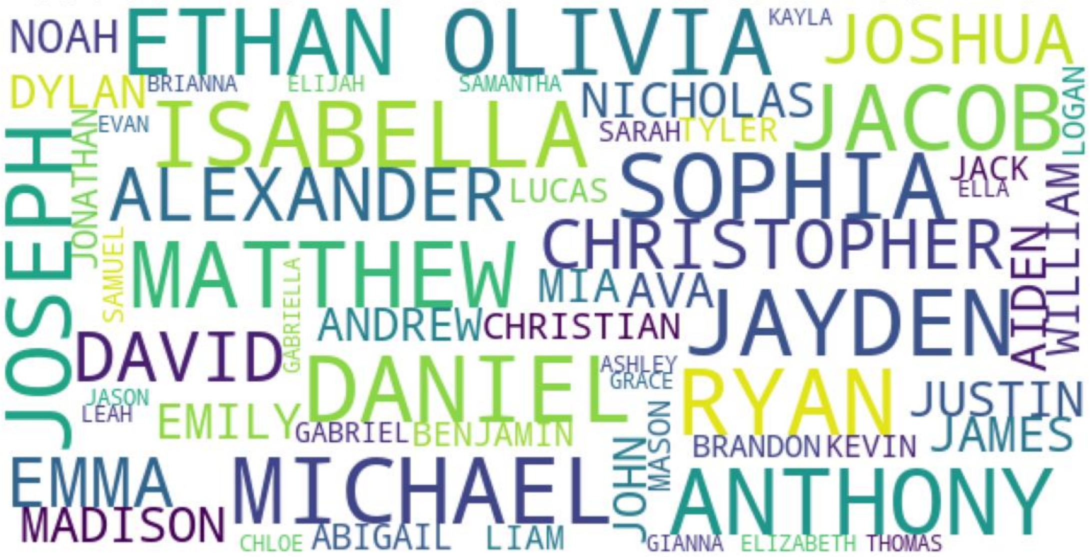
10 most popular names have a decrease in total baby names year after year

Evolution 10 most popular names



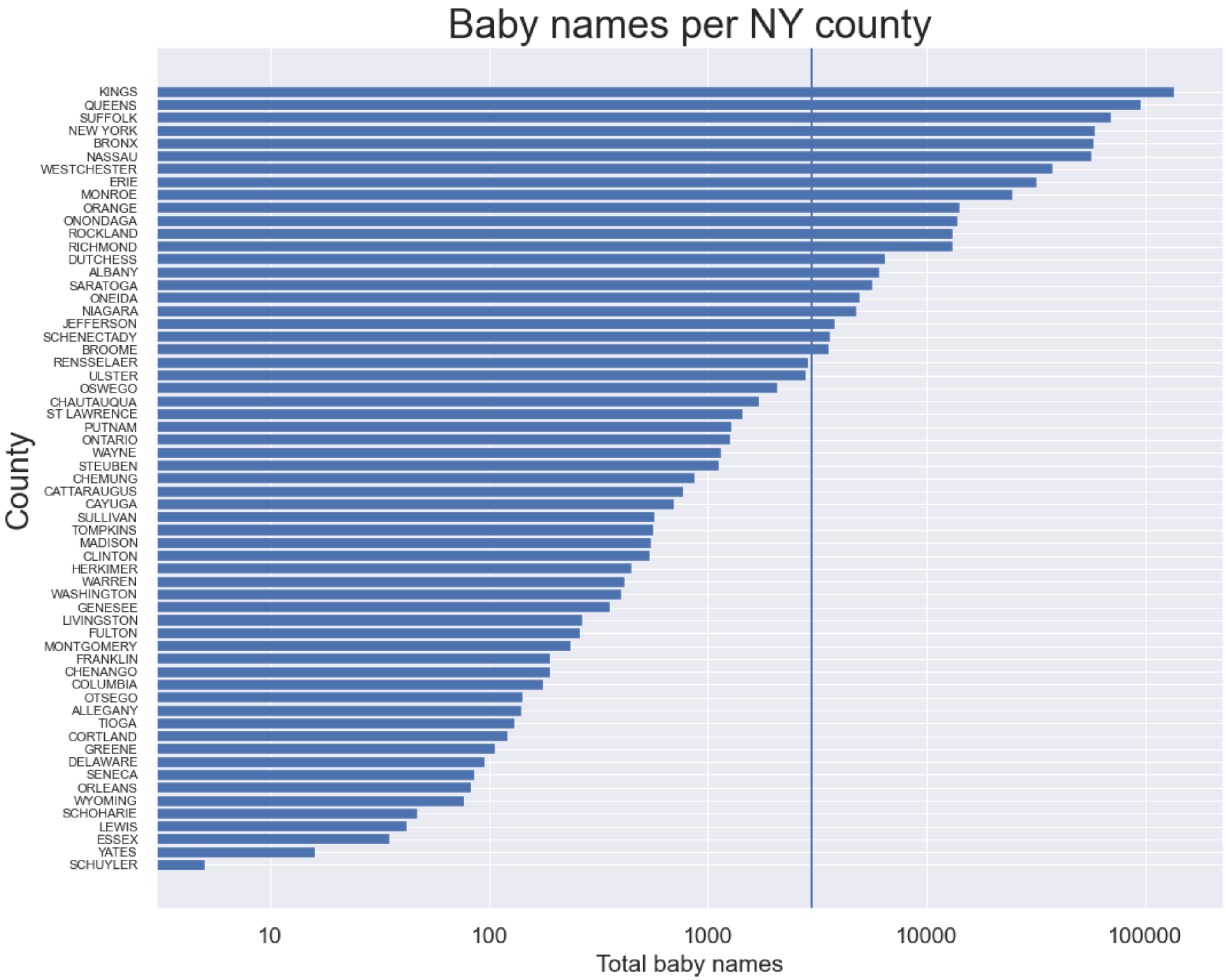
This could be done by the fact that the diversity of the names used tends to increase

Most common names in NY State 2007-2012



A lot of counties have less than 3000 total baby names registered

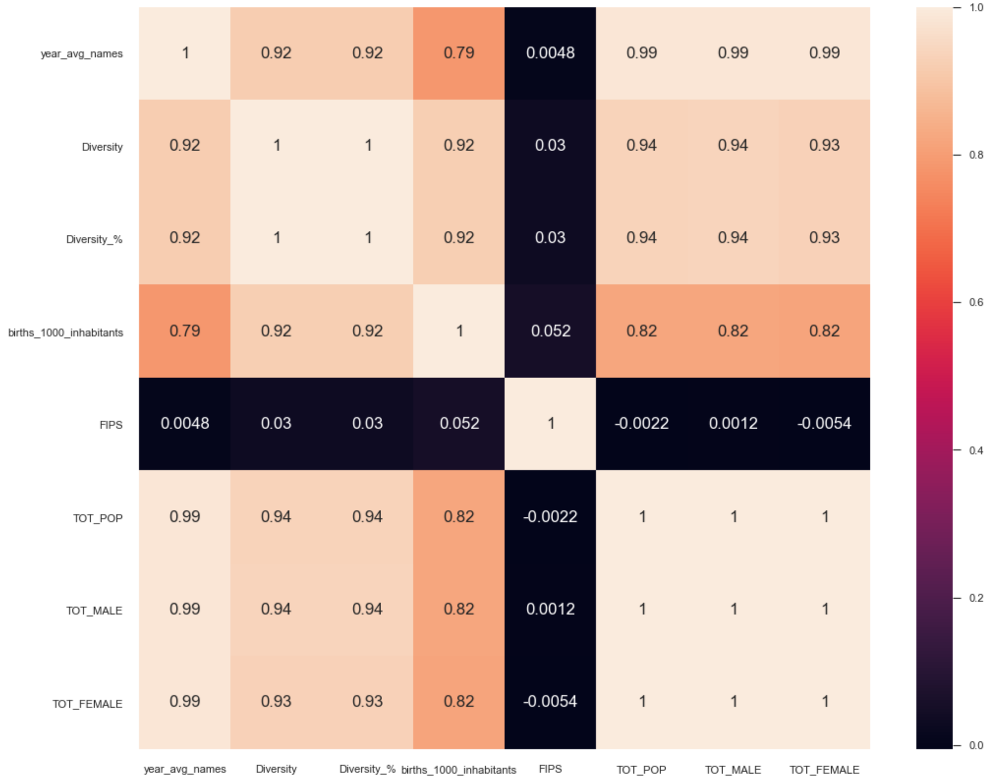
- Depending on the analysis to perform these counties could mislead the results
- Assess for each method whether to exclude these counties or not



Correlation analysis of the main aggregated variables

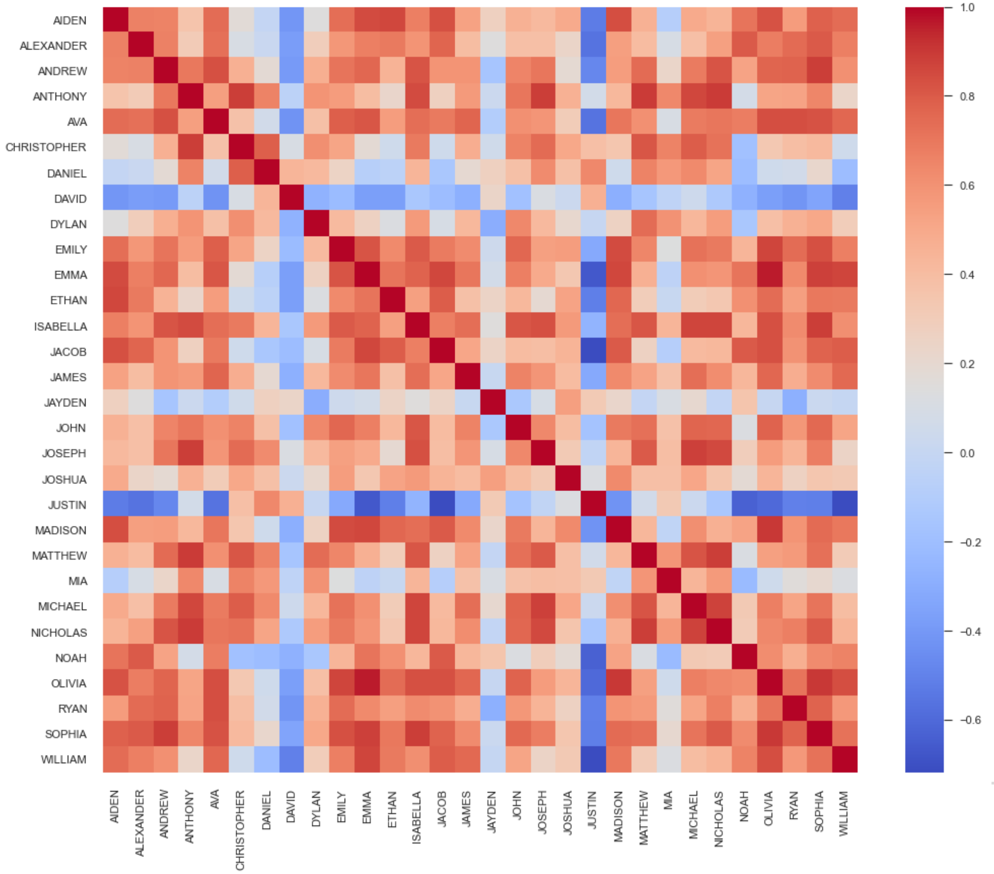
It can be seen how diversity is positively correlated with year_avg_names and births_1000_inhabitants.

Hence, those states with that have higher year_avg_names and births_1000_inhabitants will have more easily much more diversity in the unique names.



	year_avg_names	Diversity	Diversity_%	births_1000_inhabitants	FIPS	TOT_POP	TOT_MALE	TOT_FEMALE
year_avg_names	1	0.92	0.92	0.79	0.0048	0.99	0.99	0.99
Diversity	0.92	1	1	0.92	0.03	0.94	0.94	0.93
Diversity_%	0.92	1	1	0.92	0.03	0.94	0.94	0.93
births_1000_inhabitants	0.79	0.92	0.92	1	0.052	0.82	0.82	0.82
FIPS	0.0048	0.03	0.03	0.052	1	-0.0022	0.0012	-0.0054
TOT_POP	0.99	0.94	0.94	0.82	-0.0022	1	1	1
TOT_MALE	0.99	0.94	0.94	0.82	0.0012	1	1	1
TOT_FEMALE	0.99	0.93	0.93	0.82	-0.0054	1	1	1

Most names are positively correlated with each other with some of them like Daniel and David that are more common when the rest are less common



Which county may be more representative of the whole NY state in terms of name distribution?

Difference analysis to find the difference between the proportion of each names in the NY state with the proportion of each name in each county

Assumption

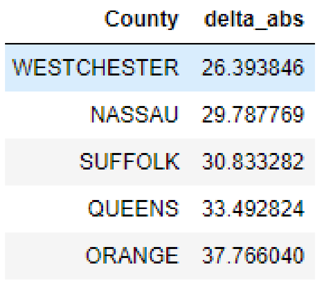
1. **All counties should be included since the faulty data will compensate**

Methodology

1. **Difference of the proportion**
2. **Distance and Variance Calculation**

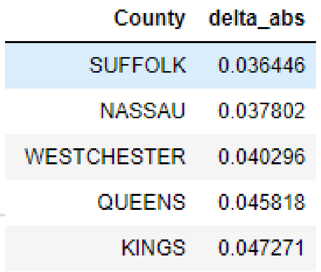
Results

1 Distance Result



	County	delta_abs
	WESTCHESTER	26.393846
	NASSAU	29.787769
	SUFFOLK	30.833282
	QUEENS	33.492824
	ORANGE	37.766040

2 Variance Result



	County	delta_abs
	SUFFOLK	0.036446
	NASSAU	0.037802
	WESTCHESTER	0.040296
	QUEENS	0.045818
	KINGS	0.047271

The counties that differ less from the proportion of names of NY state are **WESTCHESTER, NASSAU, SUFFOLK and QUEENS**.

They are, in that order, the ones closer to the NY distribution of names and also the ones with less dispersion, i.e., the ones that differ less from the real proportion on average.

Which counties can be grouped together in terms of the name distribution?

Techniques

1. **PCA**
2. **K-means**
3. **Hierarchical clustering**

Assumptions

1. **Exclude those that have less than 3 births_1000_inhabitants**
2. **Maintain counties even if they have huge decrease in population in 2010**
3. **Use the 200-400 more popular names in NY state**

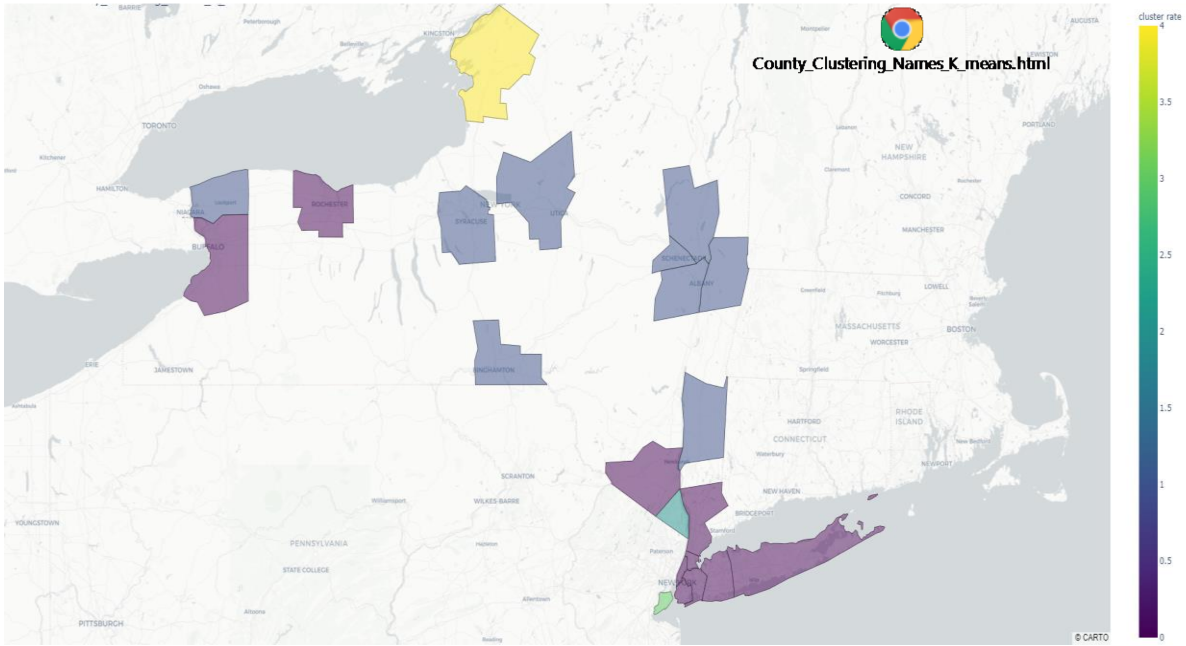
Methodology

1. Use the proportion of each name in each county as variables and the counties as observations to cluster them using clustering techniques
2. Since many variables are used (400 names %), Using PCA to reduce the number of variables is advised to reduce computational time
3. Cluster using K-means with the PC variables
4. Cluster using Hierarchical clustering with the PC variables
5. Compare the clusters obtained with both methods

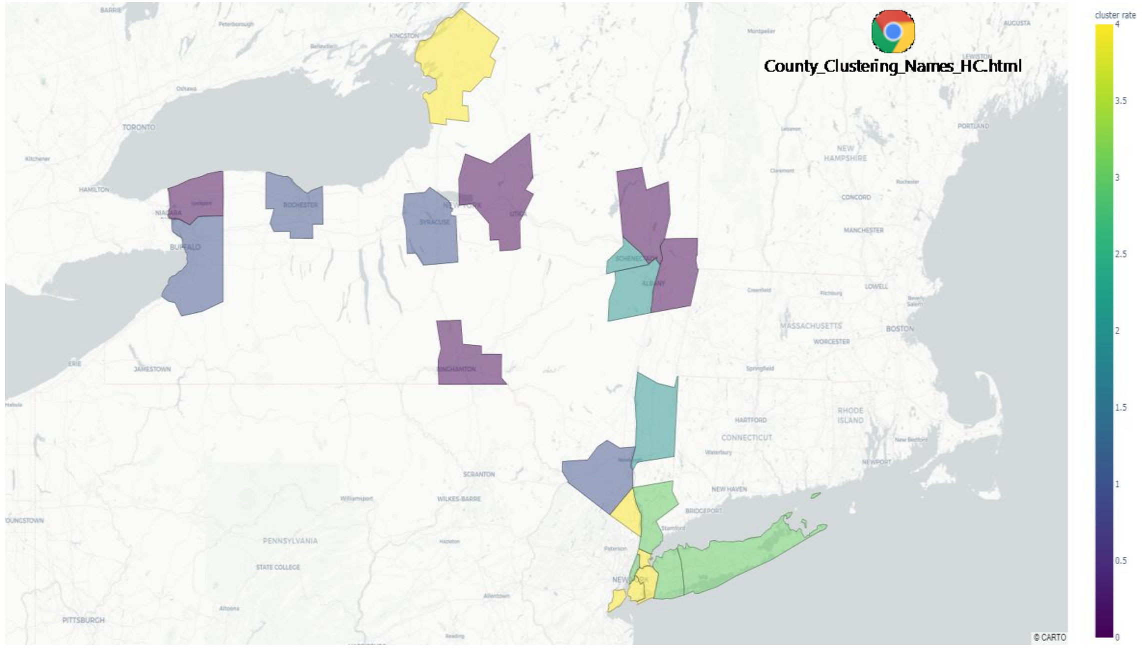
Note: Since the number of counties is still low (60) the analysis could be performed with all the names as variables, however, if the same analysis would be done with all the counties of the USA then this method would reduce a lot the computational time

There are clearly differences in name distribution between the counties of NY city and the inner counties of NY state

K-Means Clusters



HC Clusters

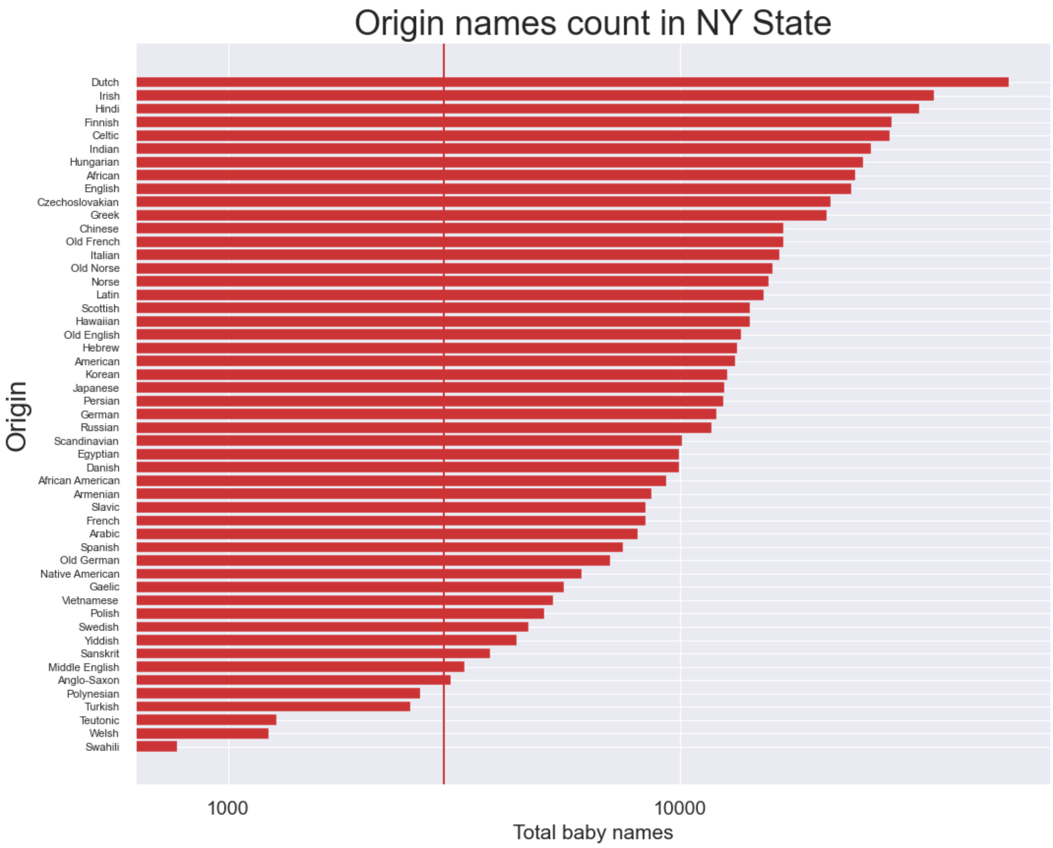


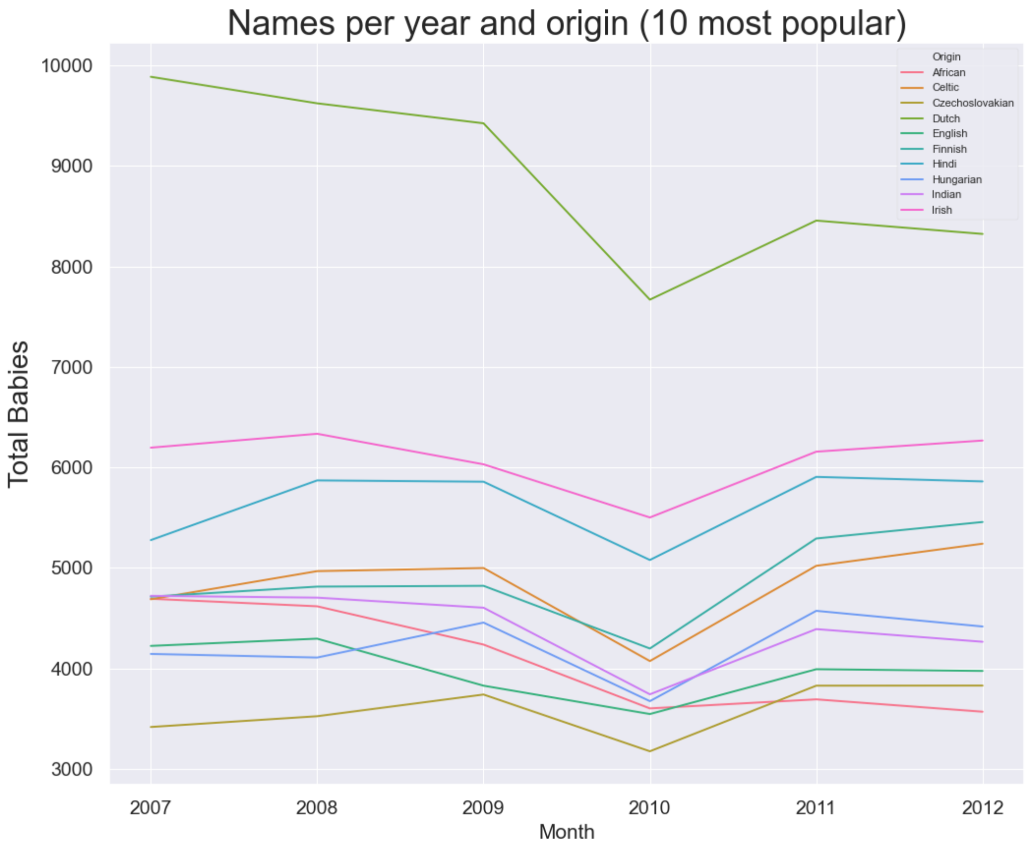
Both clustering techniques give similar results. NY city counties are grouped together. It is interesting to note that some inner counties have similar name distribution than the NY city counties

Note: Since the number of counties is still low (60) the analysis could be performed with all the names as variables, however, if the same analysis would be done with all the counties of the USA then this method would reduce a lot the computational time

Dutch is the origin of most of the name babies in the whole NY state between 2007-2012

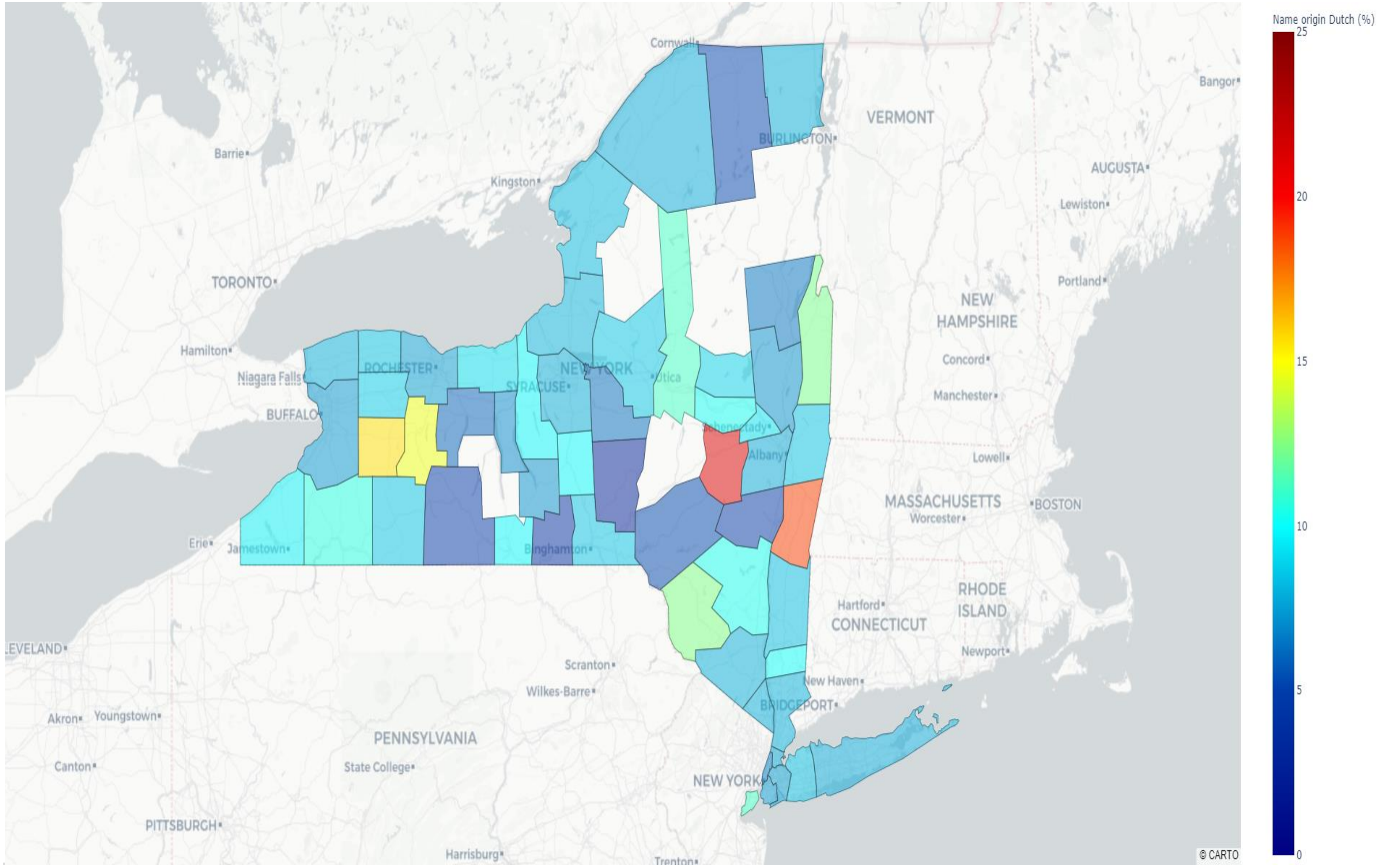
Most of the names maintain their level year after year. **The most popular name origin (Dutch) has a decrease from 10k in 2007 to 8 k in 2012**


Origin names count in NY State


Names per year and origin (10 most popular)

Distribution of Dutch origin names in the NY state per county

Dutch origin names Distribution



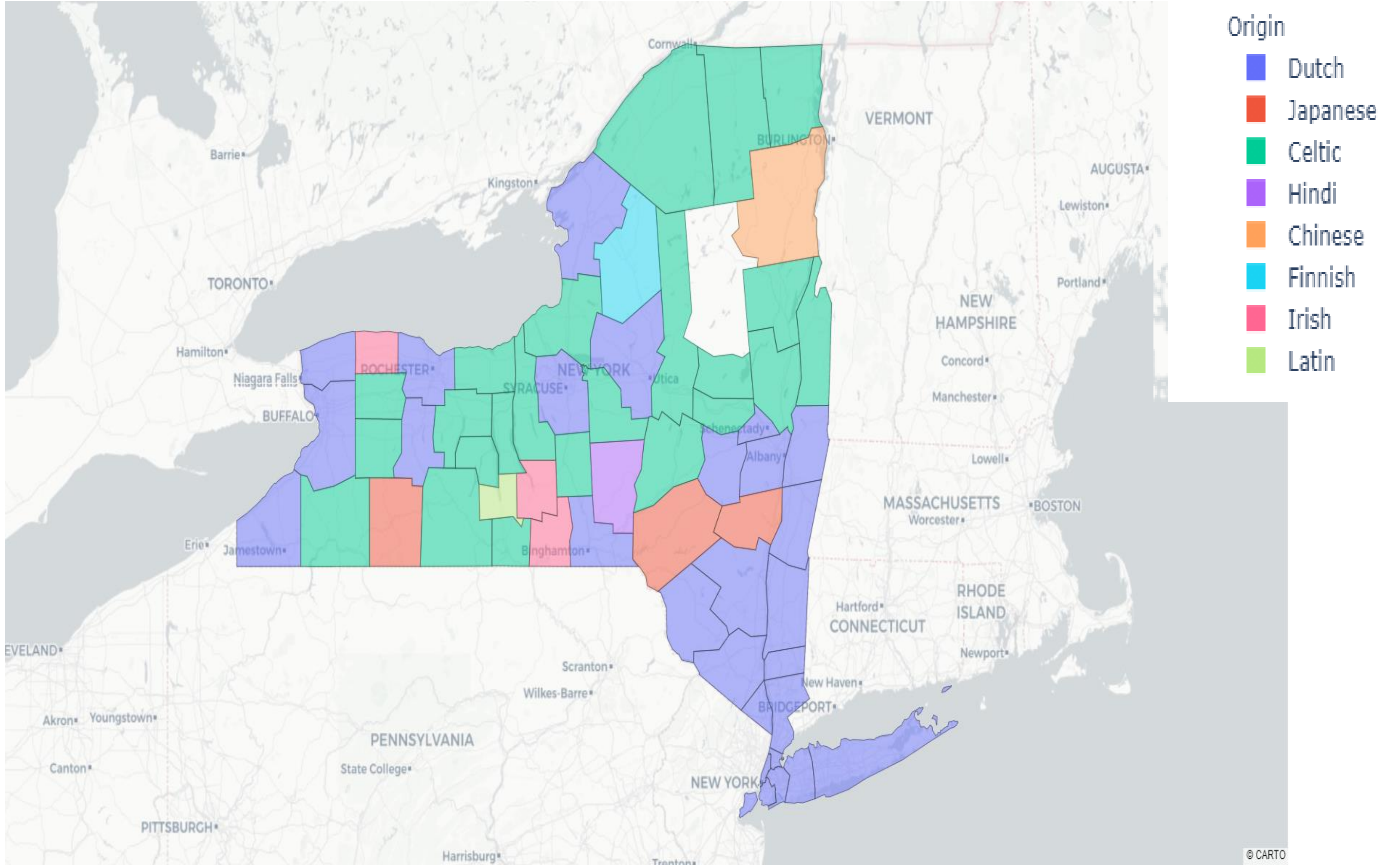
Names of Dutch origin have the highest proportion (21%) in the county of SCHOHARIE. It can be seen that its distribution its pretty uniform across the state.

There are some counties that do not have a single name of Dutch origin. This can be an indication of lack of the complete data for those years

Origin_name_Dutch_Distribution.html

Most common name origin by county in NY State

Dutch origin names Distribution



- Dutch names are found in counties near NY city, which makes sense since it was founded by Dutch settlers

- In the inner counties, the predominant name origin is Celtic,

- Some inner counties have Dutch as main name origin, which in fact are those that are in the same group as NY city counties resulted from the K-means clustering

Most common name origin by county.html

Next Steps and Recommendations

- Clustering could be done yearly to see if the same counties continue to be associated year after year

- The same clustering analysis could be generalized to compute the similarities between counties from different states that maybe have a similar race composition

- Group the name origin by continent to see which continents origin names are more representative

- The same analysis can be scaled to a dataset with the names and origins of babies from all across the US