Get Ready for Baby

Kendrick Cole, Summer Gerry, Natalie Delworth, David Cabatingan (in order by strangeness of name, descending)

CSCI1951a Data Science - Spring 2019

BROWN

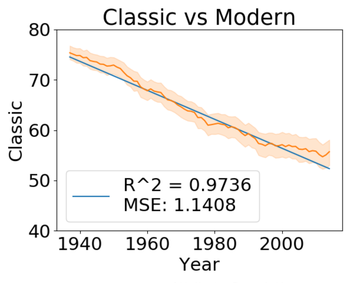
Data Sources and Gathering

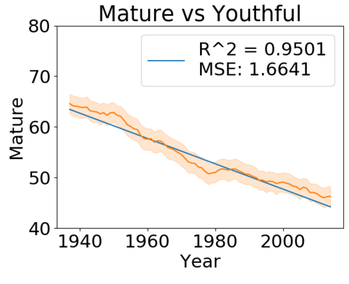
One of our data sets came from the Social Security Administration, and we scraped the other data set from BehindTheName.com, a website that crowd-sources name attribute ratings.

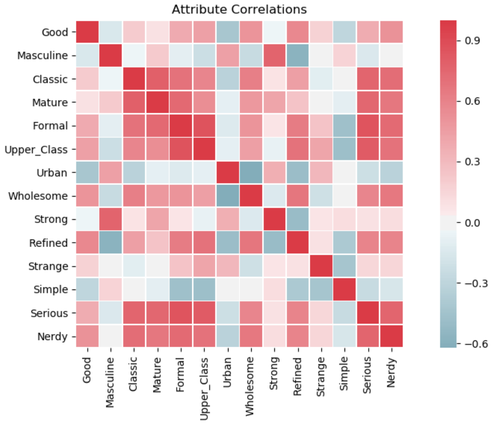
One of our biggest technical challenges came in scraping the website for data, since we had over 90,000 names from the SSA dataset to scrape for, and each web request took more than half a second to complete. We ended up using node.js to make the web requests concurrently to scrape in under an hour. We also had to deal with many edge cases, as there were names with accents or names with different pages for different roots, and the website had multiple ways to say a name did not exist.

Name Sentiment Analysis

Hypothesis: After seeing the behindthename.com data, we felt that it seemed biased. (This is possibly because all the names are being rated now so its only current perceptions of names, but we can't test for the cause). We did linear regression for the average attribute values for each year and found high correlation values of $r^2 = 0.9501$ and $r^2 = 0.9736$ for Mature vs. Youthful and Classic vs. Modern, respectively.







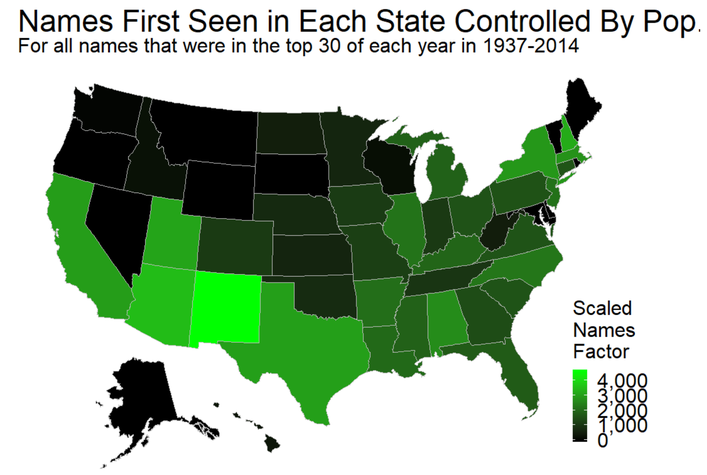
We created the correlation plot to the left and found that some of the attributes had high correlation values, indicating that the data was lower in dimensionality than the 14 attributes would suggest.

Name Origination Analysis

Hypothesis: New baby names originate uniformly from different states after controlling for population.

To control for population, we weighted each new name by the ratio of babies born in that state that year to the number of babies born that year, i.e. for state S, where b_y is number of babies born in year y and b_y^S is number of babies born in year y in state S,

$$S_{\text{score}} = \sum_{y=1937}^{2014} (\#\text{ of new names in } y \text{ in } S)\frac{b_y}{b_y^S}$$



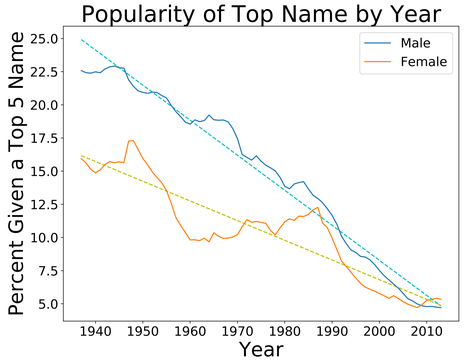
We ran a χ^2 test for GOF on the distribution of where baby names originated after normalization with H_0 as the uniform distribution, and had a significant p-value of $p = 2.2e-16$, so we can conclude that the distribution of where baby names originate is significantly different from uniform, after controlling for population of states.

Name Trend Analysis

We wanted to analyze the similarity in popularity over time between names and look for significant groupings. We computed the cross-correlation of the normalized popularity of each name with each other name to create a similarity matrix. We found the two most similar female names to be **Florence** and **Mildred**, and the two most similar male names to be **George** and **Arthur**. We then used t-distributed Stochastic Neighbor Embedding on the similarity matrix and found that the names formed one continuous cluster which was mostly ordered by the year a given name was most popular.

Name Trendiness

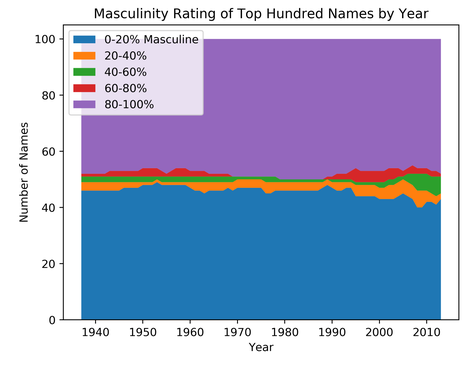
Hypothesis: the most popular names held a constant proportion of the babies born each year.



We ran linear regression and had an r^2 coefficient of 0.9720 for male names and 0.8284 for female names, so we can conclude that the popularity of the top 5 names over time is not uniform. The top 5 names comprise a smaller proportion of the population now.

Bloopers

We wanted to know whether gender neutral names have become more common over time. We tried a couple of strategies, including creating a sediment graph using the masculinity rating. There appears to be a small increase in the mid-masculinity range, but we were not able to find significant results.



Acknowledgements

Professor Ellie Pavlick, TA Anna Nakai, and the rest of the CS1951A TAs!

Sources

- Social Security Administration Name Dataset downloaded from www.kaggle.com/kaggle/us-baby-names
- Name Sentiment Data scraped from www.behindthename.com
- Poster adapted from Jacobs Landscape Poster LaTeX Template Version 1.1 (14/06/14) via http://www.LaTeXTemplates.com