**US Baby Names Analysis**

Santosh Selvaraj

14th Feb’2019

**Background:**

The dataset provided contains information regarding the popularity of baby names in the United States (measured by Social Security card applications) organized by state, gender, and year.

**Tool :** Python (Jupyter Notebook)

**Data Preprocessing:**

Collated the multiple text files, by state, into one single dataset for performing the analysis using Python. Also performed basic quality checks and summarized the data as a part of understanding the dataset provided.

**Analysis:**

**Part 1:**

Analyzed the unisex baby names : "Jessie" and "Riley" and conducted statistical test to identify whether either of the names is significantly more associated with a particular gender than the other. Leveraged the Chi Square test to understand the significance of association between the two categorical variables : baby names(‘Jessie’/’Riley’) and gender(‘male’/’female’). The test yielded a very small p-value, close to 0, and hence lead to the rejection of the null hypothesis that there is no association between the two variables. This indicated a significant difference between the observed and expected values.

Post identifying a significant association between the two variables, the individual chi square statistic values were calculated for the two names and the genders to understand whether either of the names(Jessie/Riley) show a higher association with gender(male/female). The individual statistic values revealed that the name 'Riley' associated more strongly with both genders as compared to the name 'Jessie'. since the statistic values were higher for Riley.

**Part 2:**

Analyzed the baby names dataset and performed descriptive analysis to identify the 5 most common female and 5 most common male names between the time period: 1900-2000 and visualized their popularity trends over time.

Results for female names showed that although the identified names are the most popular female names overall, the trends show a diminishing popularity around the end of 20th century. The popularity of names such as 'Linda', 'Barbara' and 'Patricia' peaked around the 1950s and eventually reduced over time. Also other names such as 'Jennifer' picked up popularity only towards the end of the 20th century and was uncommon during the earlier days.

Results for male names showed a similar trend for all the popular names, with an increased popularity during the 1950s, much like some of the popular female names. Even though the popularity of male common names had dropped over time, the decline was not observed to be as sharp as in the case of the popular female names.

**Part 3**:

1. Analyzed the change of diversity in names over time owing to factors such as exposure, influences from different nations/cultures, education etc. The results showed a steep rise in the number of unique baby names for both male and female over time, indicating diversified choices of names as hypothesized
2. Analyzed the popularity of baby names based on the alphabets they begin with, and their trends over time. The analysis results showed that names starting with 'E','O','Q' are increasing greatly in popularity. Also that the more uncommon alphabets such as 'X','Y','Z','U' are surfacing more frequently over time as compared to the early 1900s
3. Analyzed the length of baby names to identify bias in selection of names The results of the analysis showed that baby names that have length between 5-10 letters are the most popular choices and remain popular throughout the chosen time period. Results also showed that the baby names with length greater than 10 letters are not preferred and see a continuing trend of low popularity throughout the time period, as expected, indicating people's inclination towards shorter names over long ones