SI 618
Rohita Tikoo

SI 618 Project Proposal

Dataset:

The data sources used for this project were taken from the publicly available US Baby Names data sets on the website of Kaggle (https://www.kaggle.com/kaggle/us-baby-names). Sqlite database which contains the tables called NationalNames and StateNames.
This database has the size of 130.4MB. This database contains records for the time period of 1879 to 2014. The national data on the relative frequency of given names in the population of U.S. births where the individual has a Social Security Number. Each record has a name(character datatype) which is 2 to 15 characters, sex is M (male) or F (female) and count(numeric datatype) is the number of occurrences of the name.
The fields present in the NationalNames table are Id(numeric datatype), Name, Year(numeric), Gender (character datatype) and Count.
State-specific data on the relative frequency of given names in the population of U.S. births where the individual has a Social Security Number (Tabulated based on Social Security records as of March 8, 2015). Each record has a name which is 2 to 15 characters, sex is M (male) or F (female) and count is the number of occurrences of the name. The fields present in StateNames table are Id, Name, Year, Gender, State and Count.
This dataset was also used by me in my SI 601 final project, where my focus was mostly on demonstrating my data manipulation skills using Python and its various libraries. For the SI 618 project my focus would be more of diving into and exploring this dataset and analyzing it using R and its various libraries.

Exploratory Questions

1. What is the trend in the average length of baby names been from 1880 to 2000?
2. What is the impact of President names on baby names, while the Presidents were still in office?
3. Which old names (boys and girls) have not been used since 2000's?
4. Which were the most popular female names in each decade from 1880 till 2014? What is the trend like?

Proposed Methods:

Exploratory Data Analysis Method: For question 1,2, 3 and 4 I would be using the smoothing and trend finding in R. I would also use the subsetting, transforming, summarizing data slicing and dicing data with the dplyr package. Using data frames would be a powerful way to keep related variables together in a package.

Visualization Methods: I would use ggplot2 package and plot the appropriate visualizations for the above questions.
For 1 – use Scatter plot with layers
For 2- use Histogram
For 3.-use scatter plot
For 4 – scatter plots and bar chart.