Part A — Descriptive Analysis

1. Description of the data:

* + The data files are separated by US state (and D.C.). Within each file, the tabular data is represented as comma separated values (“csv”) with five columns and without headers. Each line represents the two-letter state abbreviation, a single letter for male or female (‘M’ or ‘F’), year of birth, first name (between 2 and 15 characters), and frequency of the name within the particular state and year.
  + Per the readme file included with the data, the data does not include names with fewer than five occurrences within a year and state. This means that the data does not represent all births in the US and does not represent all unique names. Because there could be many unique names occurring fewer than five times in a year and state, there is a lot of potentially missing data.
  + Based on the Social Security website providing the data, clerical errors exist in the data. For example, the sex may be misclassified, the name could be recorded as ‘Babyboy’, ‘Baby’, ‘Babygirl’, ‘Female’, various date, month, or day of week names not typically used as names of people, etc. The occurrences of these issues are relatively small.

Part B — Predicting Delays