Ian Laurain

Sean Wang

6350.002

**Final Project – Phase 1**

**Team Members:**

Ian Laurain, Sean Wang

**Topic:**

Kaggle Contest: Finding naming trends from babies born in the US

**Data:**

Data is provided by Kaggle. Two datasets: NationalNames.csv & StateNames.csv

National Names

Number of instances: 1,825,433

Number of Attributes: 4 (5 including instance number)

Id, Name, Year, Gender, Count

Size: 44.4 MB

State Names

Number of Instances: 5,647,426

Number of Attributes: 5 (6 including instance number)

Id, Name, Year, Gender, State, Count

Size: 154.7 MB

\*\*\*\*DATA SCREENSHOTS SHOWN BELOW\*\*\*\*

**Preliminary Solution and Workflow:**

We plan to use Apache-Spark (pyspark, Ipython notebook, along with Pandas & possibly some other Python libraries and modules) to perform our data analysis. This will involve parsing the data from a csv. We plan to use Pandas to do this. Pandas is a Python library which allows for the tabular analysis of large datasets. It can be used alongside Apache-Spark (also the newer, Apache-Arrow) and Hadoop, and allows for an easy, columnar-like analysis of large datasets. Ipython notebook will be the primary tool we use to work on our data, instead of conventional IDE’s or text editors. Ipython is basically an IDE that can be run in your browsers on localhost, or another server. It provides sandboxed Python interpreters for step-by-step procedural implementations of Python programs. It is heavily utilized in the scientific, research, and data science communities.

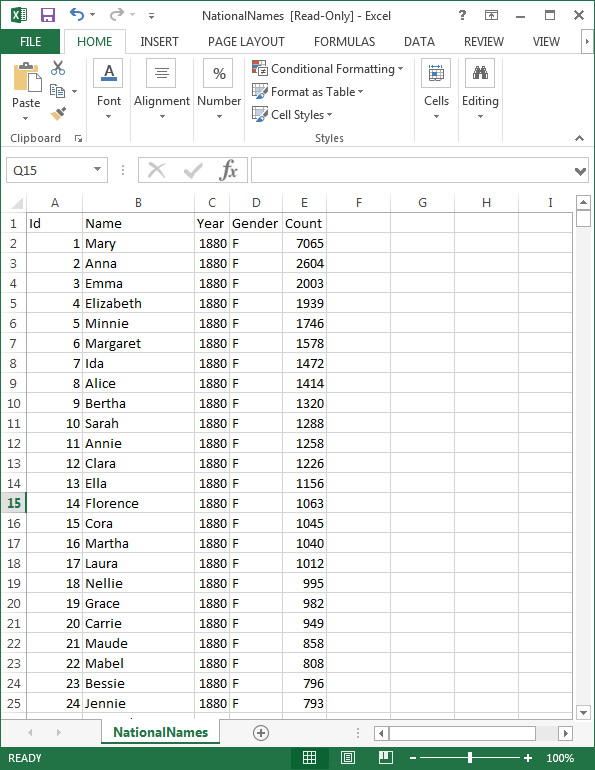
We plan to do some generally exploratory data analysis, to see if anything pops out at us, besides the hypotheses we have outlined below. Basically the project can be broken down into the following steps: research, EDA, algorithm design, data parsing, implementation, visualization, and reporting. We plan to do these steps in concert, on our own machines (if possible to avoid server overload), and with a strong emphasis on collaboration. We also hope to create well designed and interesting data visualizations. We plan to use the Seaborn Python library for these designs. Tableau might also be utilized, if a good solution is found.

**Hypothesis / What we want to accomplish:**

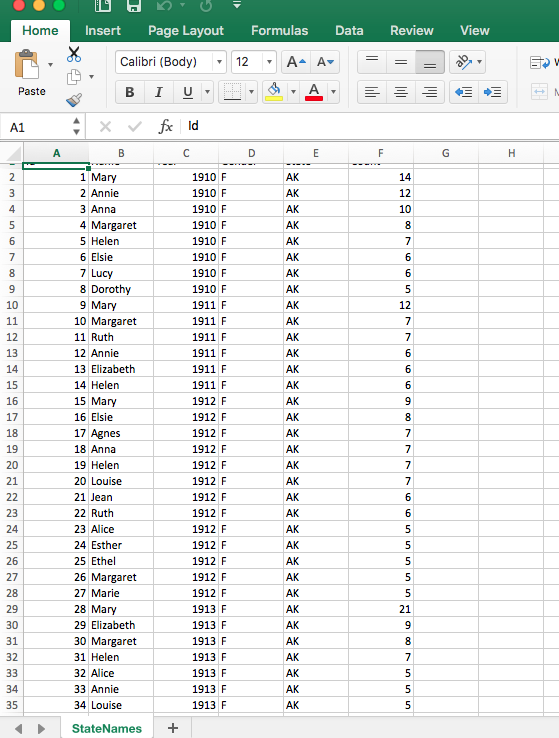
We hope to, besides finding general trends in baby name popularity, to see if we can determine when names seem to become popularity, as well as fall out of favor. We want to see if state trends affect national trends, or vice versa, and relate this to our desire to predict when names will fall out of favor. Although this step sounds simple, we think that predicting when a name will fall out of favor will be quite challenging. Although we won’t be able to determine every factor that leads to a name becoming less popular, we think we will be able to discern some factors that affect when and why a name becomes less popular. We hope to be able to predict when names will fall out of favor, and when names become classics and seems to have staying power. To do this, we might examine if a name is all of the sudden, out of nowhere, popular, and if names like these seems to flame out. By examining such behavior throughout the history covered by the dataset, we hope to determine, with some degree of accuracy, when a name becomes less common. Another trend we might analyze could be, is when certain names switch gender. However, the main emphasis will be on performing the prediction described above.

**Roles:**

We plan on working together in person for this project as this will get us both exposed to all parts on developing our Big Data skills. Both members will play heavy roles in the development side. Ian Laurain plans to handle setting up meetings, coding sessions, and setting up environments and software tools. Sean Wang plans to lead and collaborate on mathematical analysis and algorithm design, as well as project research. Both members plan to



NationalNames.csv



StateNames.csv