(Marsh, 2016)

Which Baby Names Are Popular in USA?

DATA MANIPULATION FINAL PROJECT REPORT

Rohita Tikoo | SI601 | 10-18-2016

Motivation

When people choose a name, they don't realize how much they are influenced by the world around them. It is also interesting to note how many people in the world have the same name. Sometimes it seems like a certain name is more common in a certain state or a country.

My project topic is inspired by the above fact and I would analyze the data sets (which contain records from the year 1879 to 2014) of US Baby names to answer the below questions:

- Which is the most popular name at the national level in the recent years (2004 to 2014)?
- Which is the most popular name at the Michigan state level in the recent years (2004-2014)?

Data Sources

The data sources used for this project were taken from the publicly available US Baby Names data sets on the website of data.gov (Administration, 2016) and Kaggle (Kaggle, 2016).The URLs for these data sets are mentioned in the references section of this report.

The two data sets used for this project belong to the different formats as mentioned below:

1. 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 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 the NationalNames table are Id, Name, Year, Gender 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.

2. The second data set is the CSV file named “MichiganStateNames_all.csv” which contains the records of baby names in Michigan queried from the StateNames database table. I used a python script to extract the records for the state of Michigan.

This CSV file has a size of 4779 KB and contains records for the time period of 1879 to 2014.
The fields present in "MichiganStateNames_all.csv" are Id, Name, Year, Gender, State and Count.

Data Manipulation Methods

The data sets downloaded from the sources were already clean and did not contain any missing rows. I decided to look at records from the date range of 2004 to 2014. The process of combining these two data sources was relatively straightforward. In the first part of the code I wrote a SQL query (used sqlite3 package) to fetch all the baby girl records from NationalNames table where the year would be greater than 2003 and ordered by descending value of 'Count'. In this way I got all the records from date range of 2004 to 2014 for girls at the national level. The first value in the result returned by the above SQL query would have the name of the most popular baby girl name. Using the pandas package I wrote the SQL query to fetch all the records for most popular baby girl. This result was written to the file named "NationalNames_popular.csv".

Similarly, I queried the records to fetch all baby boy names from the NationalNames table where the year would be greater than 2003 and ordered by descending value of 'Count'. The first value in the result would give me the name of the most popular baby boy name from 2004 to 2014 at the national level. Using the pandas package I wrote the SQL query to fetch all the records for most popular baby boy. This result was written to the file named "NationalNames_popular.csv".

For the second part of the code, I used the python CSV package to read all the records present in the data set "MichiganStateNames_all.csv". Next, I stored these records in a list of tuples called 'data' that contains the count, name, gender and year. Then 'data' was sorted as per the value of count. Records that had year greater than 2003 and gender as male were stored in a list called 'newdata_male' and similarly when gender was female the records were stored in a list called 'newdata_female'. From the above two lists we get the names of the most popular boy and girl at the Michigan state level.

For the third part of the code, we fetch the records for the most popular boy and girl names at the Michigan state level and store them in two different lists called pop_male_mi and pop_female_mi. Using the python CSV package the lists pop_male_mi and pop_female_mi are written to a csv file called "test.csv".

For the last part of the code the two data files "NationalNames.csv" and "test.csv" are merged into one file named "Nat_state_popularnames.csv". This merged csv file is further used to plot a pivot chart for analysis of the national level and Michigan state level baby names.

The size of the NationalNames data set was huge and considering all records from the date range from 1879 to 2014 increased the execution of my program time to up to ten minutes. To decrease the execution time. I decided to consider the records belonging to the date range of 2004 to 2014 only and this was also the time range that would reflect the most recent trend in names. Another challenge faced was while merging the selected records of most popular names at state level with national level, for this I ended up merging the two resulting csv files using the concat function of pandas package in python.

Analysis and Visualization

To answer the question of "Which is the most popular name at the national level in the recent years (2004 to 2014)?" I sorted the list of records in a descending order of count in the NationalNames table for boys and girls respectively. The resulting records gave me the answer to this question: the most popular name at the National level for the time range of 2004 to 2014 was "Jacob" for boys and "Emily" for girls.

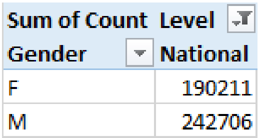
Similarly, to answer the second question "Which is the most popular name at the Michigan state level in the recent years (2004-2014)?" I read the csv file "MichiganStateNames_all.csv" and fetch the records for the most popular boy and girl names at the Michigan state level and store them in two different lists called pop_male_mi and pop_female_mi. The resulting records gave me the answer: the most popular baby names at the Michigan state level for the time range of 2004 to 2014 was "Jacob" for boys and "Emma" for girls.

The most interesting insight that I got from this analysis was that for the time range of 2004 to 2014 the most popular baby boy name at the national level (all over USA) was same as the most popular baby boy name at the Michigan state level i.e. the name "Jacob".

After merging the resulting data records from the two data sources I stored it in a csv and further modified it a bit manually to add the level (stored in a new csv called Merged_Pop_name_national_state.csv) then plotted a pivot chart which would show the sum of count by gender and level (state or national) from the year 2004 to 2014.

A PivotTable report is an interactive way to quickly summarize large amounts of data. The PivotChart report helps to visualize that summary data in a PivotTable report, and to easily see comparisons, patterns, and trends.

Set the filters of the PivotTable to National level, it shows us the trend of the sum of count for the most popular baby boy and girl names in USA (Figure 1).



Sum of Count	Level
Gender	National
F	190211
M	242706

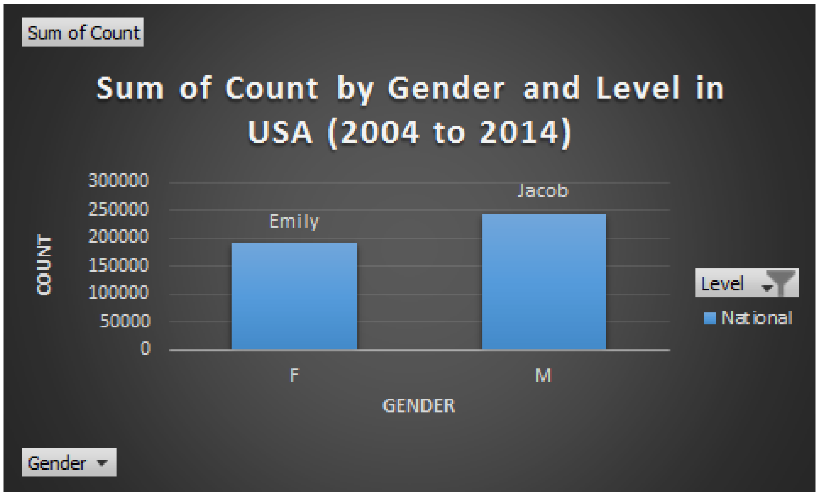
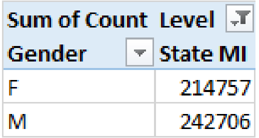


Figure 1

Set the filters of the PivotTable to Michigan State level, it shows us the trend of the sum of count for the most popular baby boy and girl names in Michigan (Figure 2).



Sum of Count	Level
Gender	State MI
F	214757
M	242706

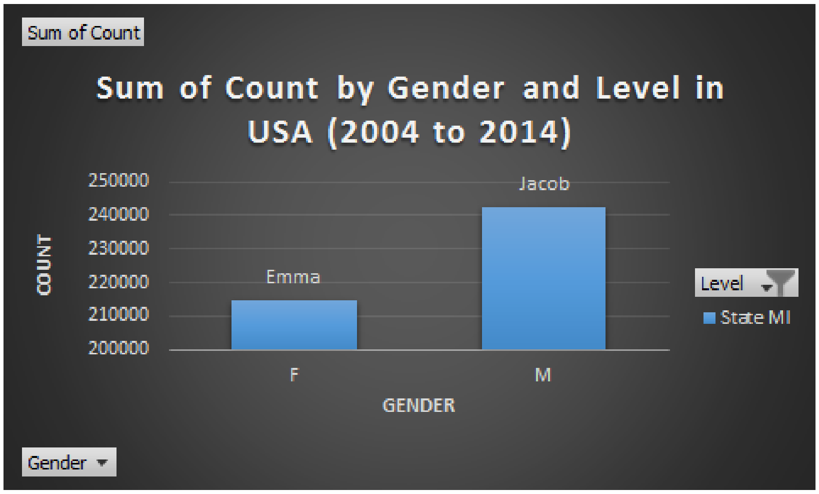
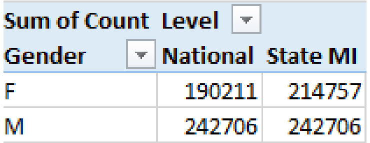


Figure 2

Set the filters of the PivotTable to both National and Michigan State level, it shows us the trend of the sum of count for the most popular baby boy and girl names in USA and Michigan (Figure 3).



Sum of Count	Level ▾	
Gender ▾	National	State MI
F	190211	214757
M	242706	242706

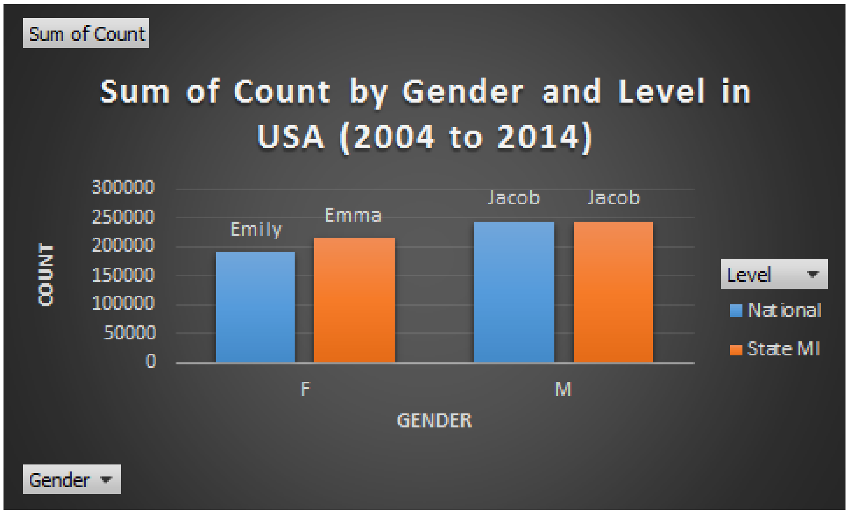


Figure 3

This project was very interesting, but there a few things that I would have done differently. I used only the recent year records from 2004 to 2014, eventhough the data sets had the records from 1879 to 2014. I also limited the scope of my project to national level names and Michigan state names but I could have also compared the trend of baby names in different states and also the trend of names that were popular in the 80's with the names in the 90's.

References

Administration, S. S. (2016, September 28). *Baby Names from Social Security Card Applications-National Level Data*. Retrieved from DATA.GOV : https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data

Kaggle. (2016, September 01). *US Baby Names*. Retrieved from Kaggle: https://www.kaggle.com/kaggle/us-baby-names

Marsh, C. (2016, January 05). *HomePregnancyBaby namesHow to name your baby...without the stress!* Retrieved from Today's Parent: http://www.todaysparent.com/pregnancy/baby-names/how-to-name-your-baby-without-the-stress/