Accuracy of SSA's USA Names Open Dataset

Google Data Analytics Professional Certificate Capstone Project

Tomas H Orihuela Jr

Last Updated June 14, 2021

Topics

1. Purpose of Case Study
2. Source and Subset Used in Study
3. The Process
4. Data Visuals
5. Conclusion

Propose of Case Study

Capstone Project

- Needed a case study to complete the Google Data Analytics Professional Certificate!
- Show the use of BigQuery, SQL, spreadsheets, pivot tables, and RStudio
- Select a open dataset that would be fun to play with!

Questions

- Could there be errors in the older data in the open dataset?
- Determine if an open dataset of first names in the US is accurate using myself as proof of accuracy!
- For accuracy testing, my SSN card shows Tomas. I was born in 1966 in the State of Louisiana.

Source and Subset

Source

- US Social Security Administration (SSA)
- Open Dataset: usa_names_1910_current (available in Google's BigQuery)
- Dataset last modified Sep 4, 2020.

Subset

- Based on my first name: Tomas
- Use different versions that included Thomas, Thom, Tom, Tomas, Tommie, and Tommy.
- Other name versions may existed but were not included in this study
- Years included: 2010 to 2020
- Genders included: male and female

The Process

BigQuery

I used BigQuery since it provides easy access to many open datasets. I created a SQL Query to obtain the data subset and exported it to a CSV file for cleaning purposes. CSV file included over 16,000 rows of data

MicroSoft Excel

I ssed Excel for the data cleaning since I used Google Sheets throughout the online course. I cleaned the data and found the data was clean as provided. I created filters and formulas to check the data. I could have used conditional formatting in place of some formulas but I didn't. I create two pivot tables for use in verification of results and plots created using RStudio. I could have done the data cleaning using RStudio but I wanted to play with pivot tables.

RStudio

Primarily used RStudio Cloud Free but I ran out of hours since I also used it throughout the course. I finished up the case study using the desktop application. I created a script using the tidyverse, ggplot2, and viridis packages. I spent the most time learning to create presentable plots.

Google Slides

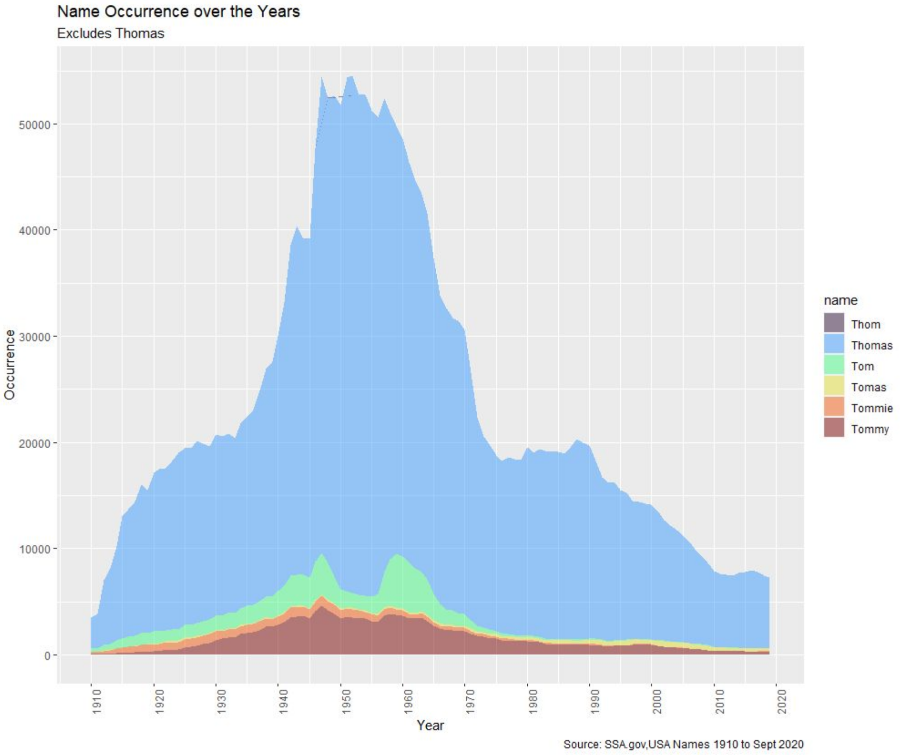
Created this gorgeous presentation in it!

Data Visuals

Trends

Thomas dominates!

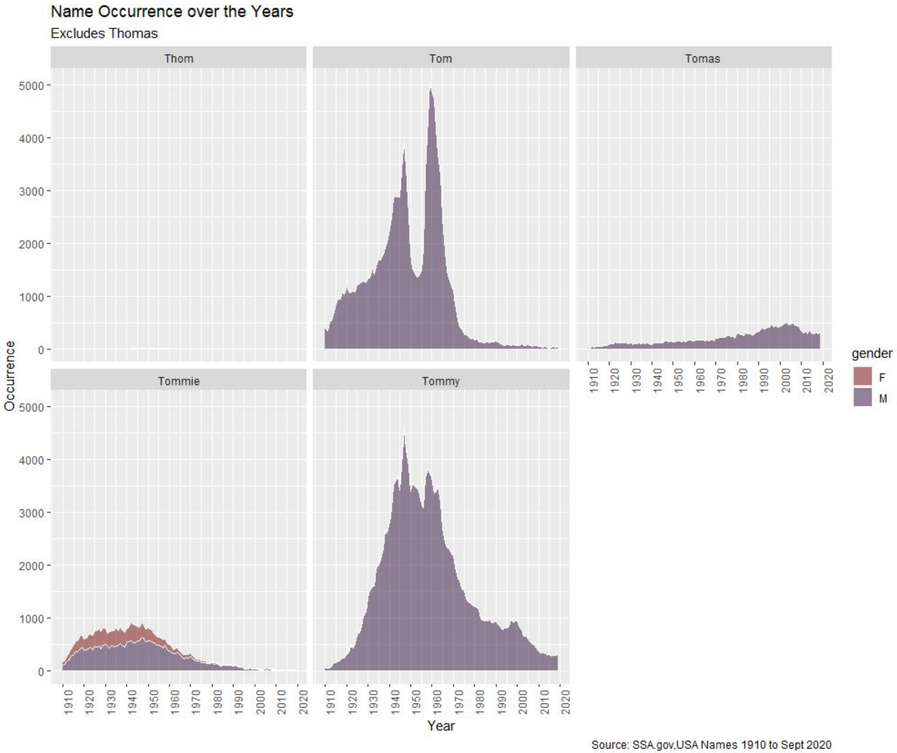
- Most names were most popular from 1930 to 1970
- Popularity of Tomas was highest from 1990 to 2010
- Dataset based on Thomas, Thom, Tom, Tomas, Tommie, and Tommy



Trends and Gender

Let’s Exclude Thomas!

- Thom was only used 49 times ever! More information coming.
- Tomas is trending upward in more recent years while all others are trending downward! Go Tomas!
- Females used four of the 6 names. More information provided later.
- Most common female name is Tommie.



Name Occurrence over the Years

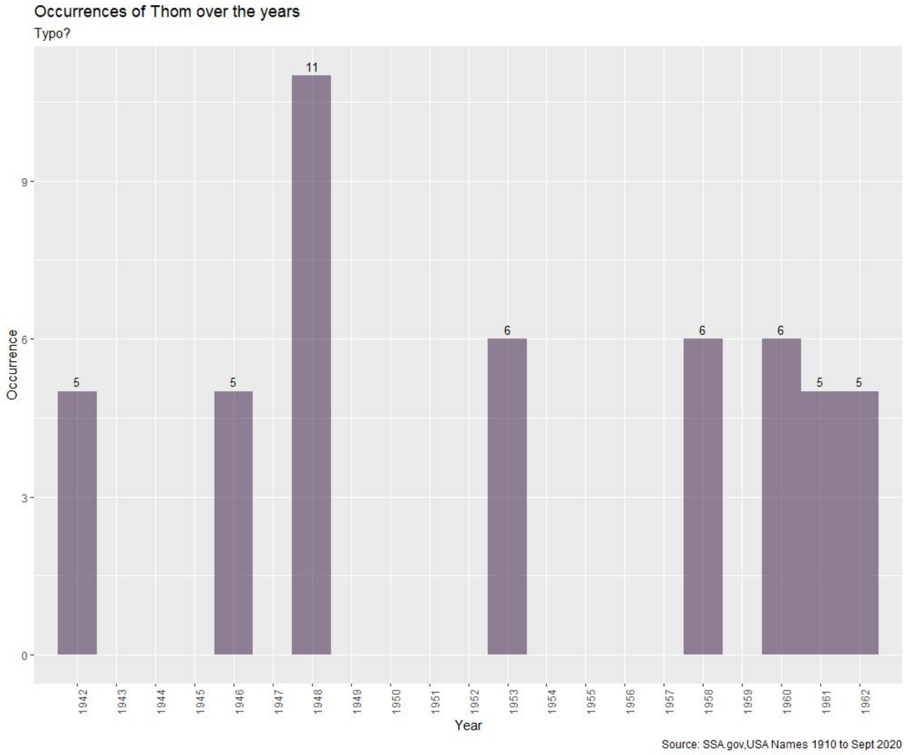
Excludes Thomas

Source: SSA.gov,USA Names 1910 to Sept 2020

What about Thom?

Thom needs attention too!

- Thom is only recorded 49 times from 1942 to 1962
- Were these typos?
- Should they be Thomas?

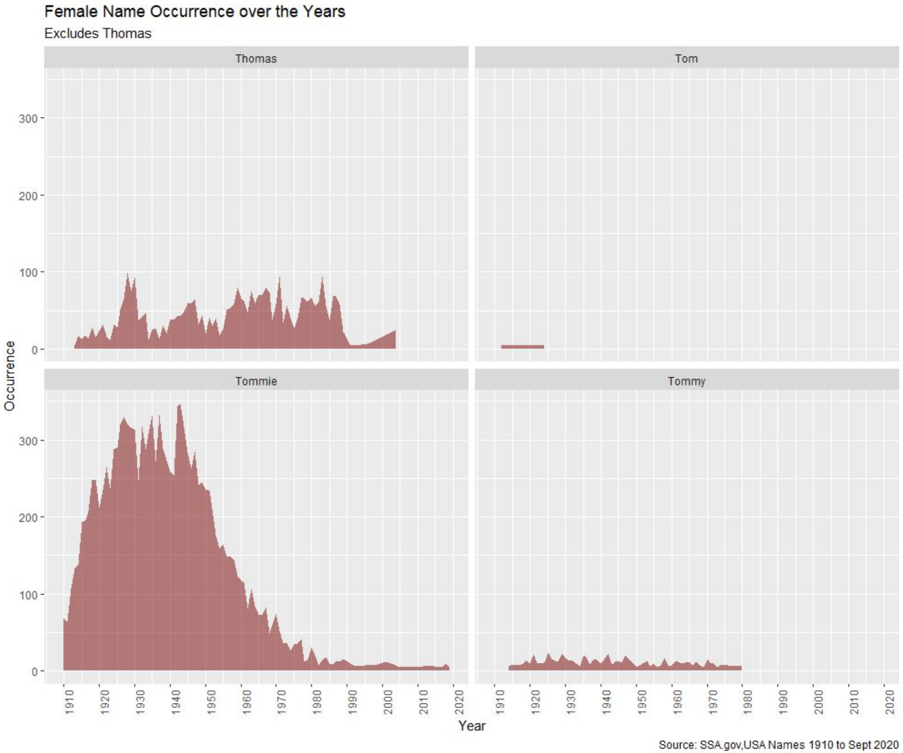


Source: SSA.gov,USA Names 1910 to Sept 2020

Female Names

Tommie is the most common and still used today!

- Were many of the Thomas' a typo on gender? Graph shows an abrupt ending at 2004.
- Were some of the Tom's typos for Tommie?
- Were some Tommy's typos for Tommie?



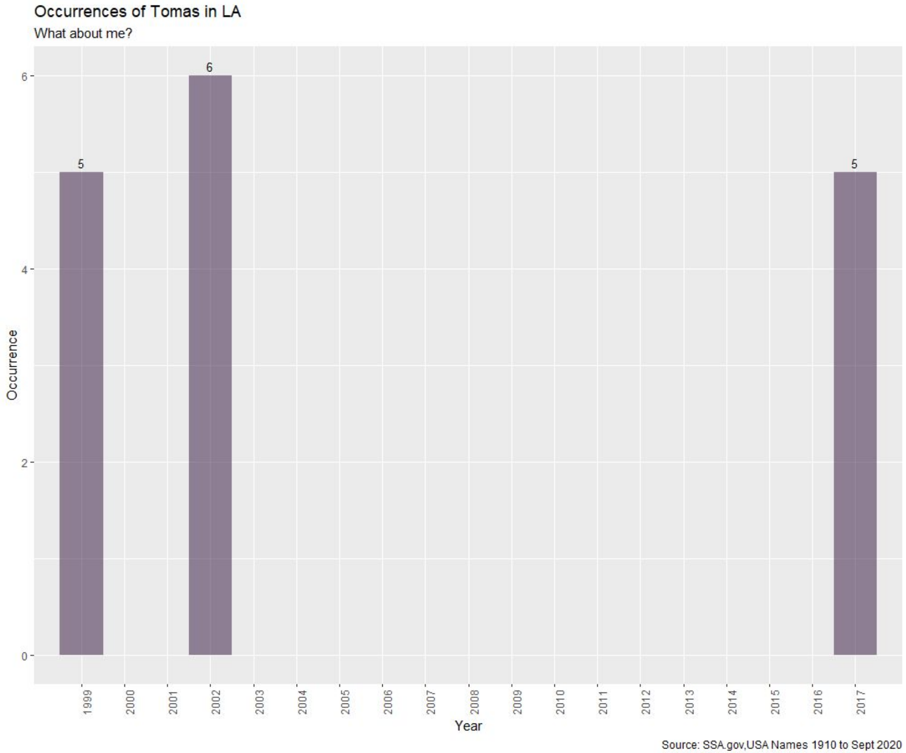
Source: SSA.gov,USA Names 1910 to Sept 2020

What about me?

Was I included?

Apparently I don't count!

- Name: Tomas
- Year: 1966
- State: Louisiana
- No data for 1966 for the state of Louisiana.
- So, was I included as Thomas or just ignored?



Conclusions

1. The SSA's open dataset of US Names from 1910 to present may not be accurate.
2. When is anything that the US Government tracks accurate? :)
3. Typographical errors during input of names are possibly included. Error in entering gender data may also be present.
4. IRS definitely has me in their system since they want their taxes!
5. Now, there is a new big question! Will I get my social security checks when I retire being that I am getting ever closer to retiring?