Capstone Project

Analysing Naming Trends using Python

By

Aniket Tidke

aniket.tidke.1997@gmail.com

A project Report submitted

In Advanced Certification Data science

And Artificial Intelligence by CCE, IIT Madras

Table of Contents:

- Problem Statement
- Project Objective
- Data Description
- Data Pre-processing Steps and Inspiration
- Choosing the Algorithm for the Project
- Motivation and Reasons for Choosing the Algorithm
- Assumptions
- Model Evaluation and Techniques
- Inferences from the Same
- Future Possibilities of the Project
- Conclusion
- References

Problem Statement:

The problem is to extract and analyse a zipped format dataset containing information about babies born in different years. We have to analyse the naming trends using Python.

The objective is to visualize the number of male and female babies born in a particular year and **identify popular baby names**. To achieve this, we will use Python and necessary libraries for data extraction, manipulation, and visualization.

Project Objective:

The main objective of this project is to extract, analyse, and visualize a zipped dataset containing information about babies born in different years containing their names.

Specifically, we aim to:

1. Extract the dataset from the zipped file and load it into a Pandas data frame.
2. Visualize the number of male and female babies born in a particular year using graphs and charts. Plot trend of it.
3. Identify the most popular and commonly used baby names based on the frequency of occurrence in the dataset.
4. Analysing baby names of top 100 birth count.
5. Analyse trends and patterns in the data, such as changes in the frequency of certain names over time or variations in gender distribution across years.
6. Use the insights gained from the analysis to make informed decisions and recommendations, such as suggestions for baby names or insights into societal changes in attitudes towards gender and naming conventions.

Data Description:

The dataset which is popular baby names is provided by the Social Security Administration (SSA) of the United States.

The Popular Baby Names dataset contains the names of babies born in the United States, along with their gender and the number of babies given that name in each year since 1880 till year 2021. The data is based on applications for Social Security cards, which are often filled out shortly after a baby is born.

For each year of birth YYYY after 1879, they have created a comma-delimited file called yobYYYY.txt. Each record in the individual annual files has the format "name,sex,number," where name is 2 to 15 characters, sex is M (male) or F (female) and "number" is the number of occurrences of the name. Each file is sorted first on sex and then on number of occurrences in descending order. When there is a tie on the number of occurrences, names are listed in alphabetical order. This sorting makes it easy to determine a name's rank. The first record for each sex has rank 1, the second record for each sex has rank 2, and so forth.

All this .txt file stored in a zip file called ‘names.zip’. Data set contain four columns Name, Sex, Babies, Year having object, object, int64, int64 data types respectively.

Data Pre-processing Steps And Inspiration:

The dataset may require cleaning and pre-processing before it can be analysed, such as removing duplicates, dealing with missing data, and standardizing the format of the data. However, once cleaned, it can provide valuable insights into naming trends and patterns in the United States over the past century.

The pre-processing of the data included the following steps:

1. Data has to download form the website by following steps:
   a. Go to https://www.ssa.gov/oact/babynames/limits.html
   b. Click on 'National data'
   c. Get the zipped file

2. Open new python file and import important libraries like numpy, pandas, matplotlib.pyplot , ZipFile and BytesIO.
3. Since data given in Zipped format use ZipFile to unzip all the text file into folder .
4. After this we will use for loop to read all files from range 1880 to 2022.
5. The range function is used to create a list of integers from 1880 to 2021. This creates a list of years that we want to read in the dataset for loop iterates over each year in the list of years. "pd.read_csv()" function is used to read in the dataset for each year.
6. The filename for each dataset follows the format "yob{year}.txt", where {year} is the year we want to read. The names parameter is used to specify the column names for the dataset. The column names for the dataset are "Name", "Sex", and "Babies", where Name is name of baby, Sex is gender that is male or female and Babies gives us number of babies names for a particular year.
7. for loop appends the year to the dataset using the Year column.
8. As years will contain list of different years data so concatenating data into a dataframe.
9. Top 5 and bottom 5 rows are checked using "df.head()" and "df.tail()".
10. Duplicates check performed.
11. Dataframe info checked to understand the data type which is int or object.
12. Check shape of data.
13. Data contain 2052781 rows and 4 columns.

Choosing the Algorithm for the Project:

1. Data is analysed using pivot table.
2. Numpy and pandas libraries are used to extract useful data from data set.
3. Matplotlib library is used to visualise data.
4. WordCloud library is also use for visualisation.
5. Plotted Barchart, Linechart, wordclound for visualisation.

Assumptions:

Based on the problem statement and the provided dataset, here are some assumptions that we can make for this project:

1. The dataset provided by the SSA (Social Security Administration) is accurate and comprehensive for the years and all regions covered.

2. The dataset provides accurate counts of the number of babies born and their corresponding names and genders.

3. The dataset covers all regions of the United States and is representative of the entire population.

4. The dataset only includes names that were given to at least five babies in a given year to protect the privacy of individuals and families.

5. The dataset does not include any special characters or diacritical marks in the names.

6. The popularity of names is based solely on the number of babies given that name in a given year, and does not take into account any other factors such as cultural or historical significance.

7. It's important to keep in mind these assumptions and their potential limitations as we perform our analysis and draw conclusions from the data.

Model Evaluation and Technique:

As this project involves analysing and visualizing data, we don't have a specific model to evaluate. Instead, we will be using a variety of techniques to explore and analyse the dataset. Some of the techniques we can use for this project include:

By code:

Here I have used “pd.options.display.max_rows=100” to display 100 rows of dataframe. Code sets the maximum number of rows to display to 100 and then sorts the dataframe by the number of babies in descending order and resets the index. Finally, it displays the first 100 rows of the sorted dataframe. This allows us to see the most popular baby names in the dataset.

Note that we can't see the entire dataset using this code because the head function only displays the first 100 rows. If we want to see the entire dataset, we can remove the head function.

Data Visualization:

Created visualizations such as line charts, bar charts and tables to explore the data and identify trends and patterns.

Following are some examples of data visualization:

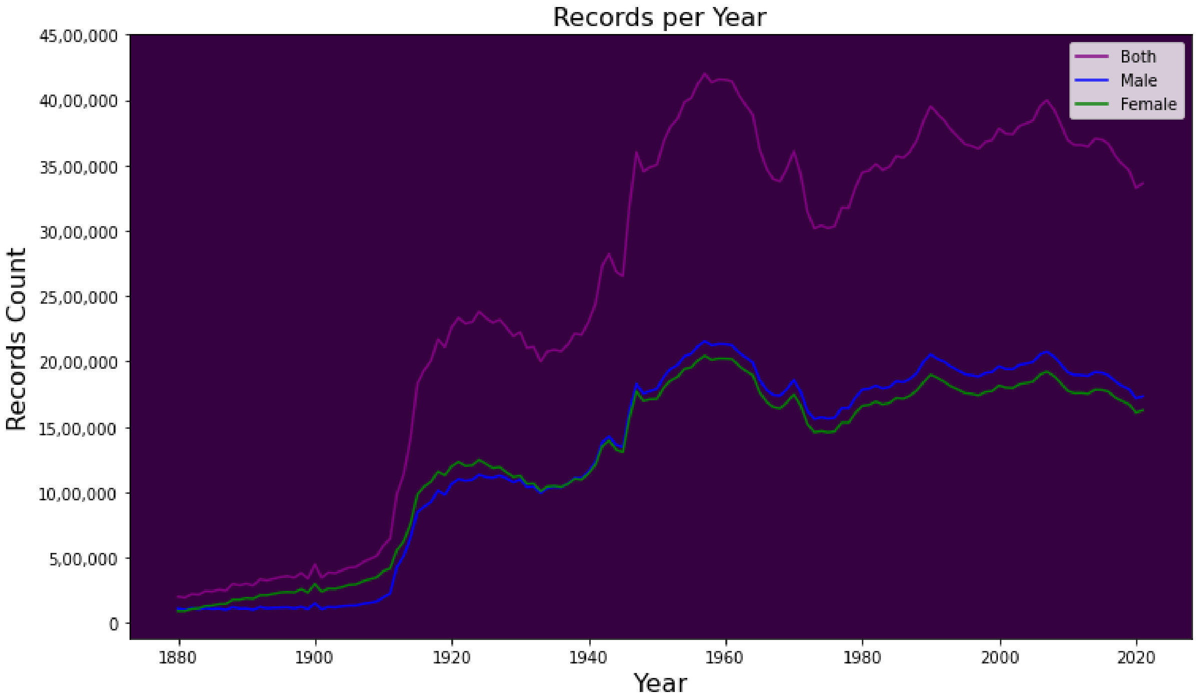

Figure-1: Line Chart showing birth counts of Male, Female and Both combined year wise from 1880 to 2021

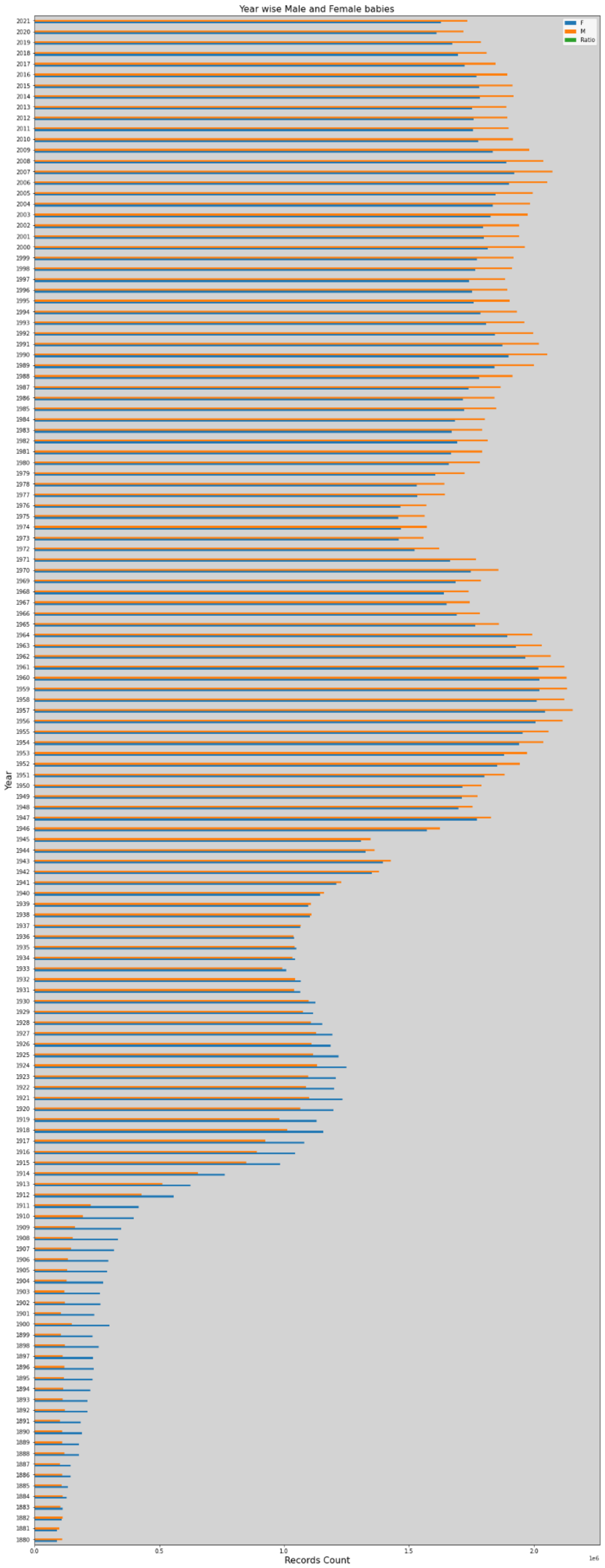

Figure-2: Bar Chart showing birth counts of Male, Female year wise from 1880 to 2021

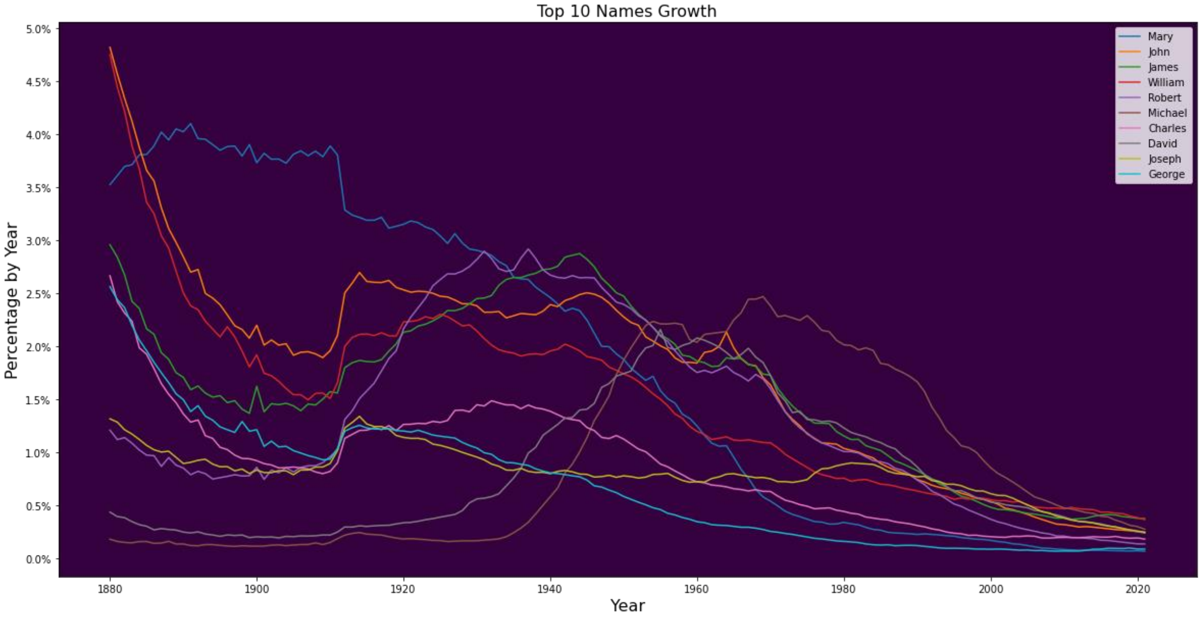

Figure-3: Line Chart showing % trend of using top 10 names from year 1880 to 2021

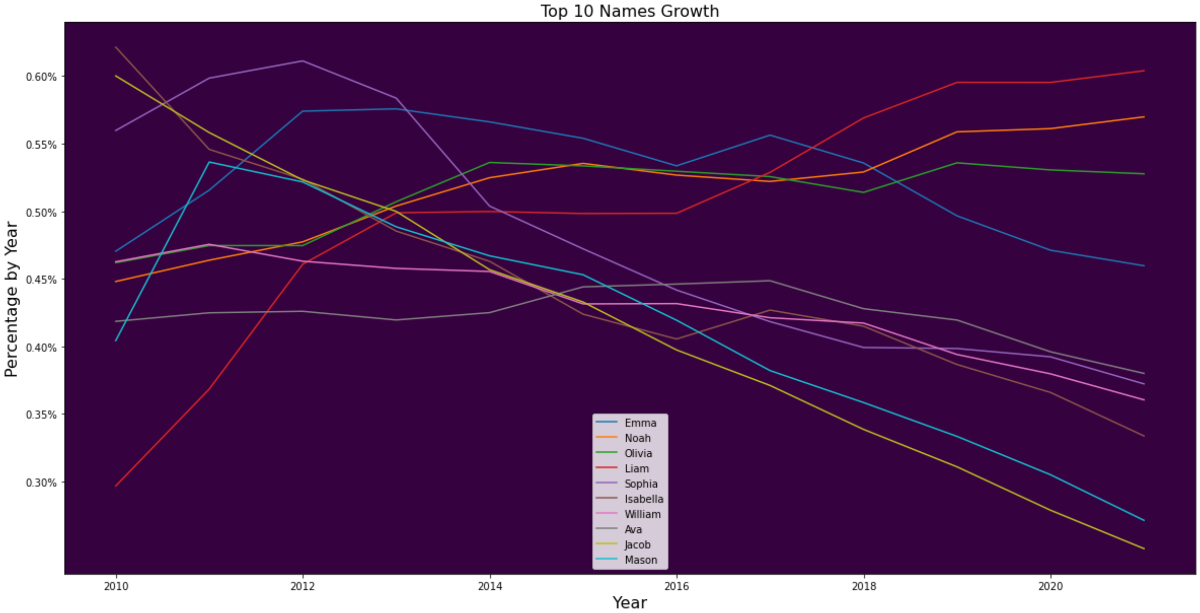

Figure-4: Line Chart showing % trend of using top 10 names from year 2010 to 2021

Inferences from the Project:

Based on the project objective and the analysis of the dataset, we can draw the following inferences:

- The number of babies born in the United States has generally increased over time, with some fluctuations from year to year.

- Data Set contain 1,01,338 unique names and 43,093 are unique male names, 69,527 are unique female names and 11,282 are unique gender-neutral names.

- Top 10 male names used from 1880 to 2021:

Name	Count
James	5202714
John	5150510
Robert	4834094
Michael	4392696
William	4156142
David	3646903
Joseph	2639396
Richard	2571082
Charles	2411608
Thomas	2331794

- Top 10 female names used from 1880 to 2021:

Name	Count
Mary	4132497
Elizabeth	1661030
Patricia	1572795
Jennifer	1469379
Linda	1453755
Barbara	1435386
Margaret	1255686
Susan	1122518
Dorothy	1109423
Sarah	1087196

- Top 10 gender neutral names used from 1880 to 2021:

Name	Female	Male	Total Count
James	23595	5202714	5226309
John	21723	5150510	5172233
Robert	20105	4834094	4854199
Michael	21811	4392696	4414507
William	16003	4156142	4172145
Mary	4132497	15172	4147669
David	12936	3646903	3659839
Joseph	10687	2639396	2650083
Richard	9528	2571082	2580610
Charles	12436	2411608	2424044

- Out of 141 years, total 87 years has more number of male babies birth than female babies and total 54 years has more number of female babies birth than male babies.

- Some stats on birth count-
  1. Highest male babies birth count in particular year is 21,56,314 in year 1957.
  2. Lowest male babies birth count in particular year is 1,00,737 in year 1881.
  3. Highest female babies birth count in particular year is 20,44,615 in year 1957.
  4. Lowest male babies birth count in particular year is 90,994 in year 1880.
  5. Year 1957 has highest number of babies born which is 42,00,929.
  6. Year 1881 has lowest number of babies born which is 1,92,690

- The most popular baby names for boys and girls have changed over time, with some names remaining popular for many years and others only being popular for a short period of time.

- Some names are more popular for boys and others are more popular for girls, although there is some overlap in the names that are used for both genders.

- The dataset is limited to names that were given to at least five babies in a given year, which means that it does not include all possible names and may not accurately represent the entire population.

- The dataset does not include any special characters or diacritical marks in the names, which may limit the accuracy of the analysis for names that are commonly spelled with such characters.

- Overall, this analysis provides insights into the trends and patterns in baby names in the United States over the past century, and highlights the cultural and societal influences that shape the popularity of certain names.

Future Possibilities:

There are several future possibilities for this project, including:

- Expanding the dataset: The dataset used in this project includes data from 1880 to 2021, but it could be expanded to include more recent years as well as data from other countries.

- Analysing name popularity by region: The dataset could be further analysed to identify regional trends in name popularity and variations in popular names by state or city.

- Identifying patterns in naming trends: Using natural language processing techniques, it could be possible to identify patterns in the types of names that are popular at different times or in different regions.

- Identifying correlations with cultural and social trends: The dataset could be further analysed to identify correlations between popular names and cultural or social trends, such as the popularity of names associated with certain celebrities or events.

- Predicting future naming trends: By analysing historical trends and identifying patterns in name popularity, it may be possible to predict future naming trends and anticipate changes in popular names.

- Developing a web application: A web application could be developed to allow users to explore the dataset and visualize name popularity trends over time. Users could also search for specific names and see how their popularity has changed over time.

- Limitations:

    a. Data set is taken for name which is been used more than 5 time for privacy purpose, so data is not complete. There are some names which are left out.
    b. States and region is not available for this data set, so that by using this we can find naming trend in different regions.

Conclusion:

The analysis of the popular baby names dataset provides insights into the trends and patterns of baby names in the United States over the past century. The dataset shows that the popularity of baby names has changed over time and that the most popular names for boys and girls have varied throughout the years. The analysis also highlights the influence of cultural and societal trends on baby name popularity.

Although the dataset has some limitations, it provides a valuable resource for understanding historical trends in baby names and identifying patterns in name popularity. There are also many future possibilities for further analysis of the dataset and for developing applications that allow users to explore and visualize name popularity trends.

Overall, this project demonstrates the value of data analysis for understanding cultural and societal trends, and highlights the importance of using data to inform decision-making and policy development.

The popular baby names given below:

Name	**Total Count**
James	5226309
John	5172233
Robert	4854199
Michael	4414507
William	4172145
Mary	4147669
David	3659839
Joseph	2650083
Richard	2580610
Charles	2424044

References:

1. https://pandas.pydata.org/docs/user_guide/index.html
2. https://numpy.org/doc/
3. https://matplotlib.org/stable/tutorials/introductory/pyplot.html
4. https://www.labmanager.com/news/glimmers-of-evolution-in-naming-babies-choosing-a-dog-28196
5. Popular Baby Names (ssa.gov)
6. https://pypi.org/project/wordcloud/