**Final Research Paper**

Shawn Behrend

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Orenthio Goodwin

8 May 2022

**Contents**

[Abstract 3](#_Toc102565958)

[Objectives 4](#_Toc102565959)

[Overview of Study 5](#_Toc102565960)

[Research Questions and Hypotheses 5](#_Toc102565961)

[Literature Review 6](#_Toc102565962)

[Research Design 9](#_Toc102565963)

[Methodology 9](#_Toc102565964)

[Methods 10](#_Toc102565965)

[Limitations of the Dataset 11](#_Toc102565966)

[Ethical Considerations for the Dataset 12](#_Toc102565967)

[Findings 12](#_Toc102565968)

[Conclusion 18](#_Toc102565969)

[Recommendations 19](#_Toc102565970)

[References 20](#_Toc102565971)

[Appendix A 22](#_Toc102565972)

# Abstract

Choosing a first name for a baby can be one of the most fun, yet also most frustrating activities for parents-to-be. Many parents choose a pre-existing name for their child, while others choose to design and invent a new unique name for their child (Hahn & Bentley, 2003). The majority of chosen baby names in the United States are gender-specific, meaning they are used for a particular gender, either male or female. A small minority of chosen baby names are considered gender-neutral, androgynous or unisex, in that they can be used for either a male or a female (Seguin et al., 2021).

This project involves an analysis of trends in use of gender-neutral names in the United States over 14 decades, from 1880-2020, as well as uses statistical analysis via correlation tests to better understand these trends over time. The objective is to explore, analyze and gain understanding towards insights about gender-neutral baby names based on the data contained in the "Baby Names from Social Security Card Applications - National Data" dataset for years 1880-2020, published by the Social Security Administration (Nuessel, 2017; SSA, 2021).

**Final Research Paper**

Choosing a name for a new baby can be one of the most fun, yet also most frustrating activities for parents-to-be. Many parents choose a common or popular name, while others design and invent a new unique name for their child (Hahn & Bentley, 2003). A child's name expresses the child’s individuality and can also represent family traditions.

The majority of chosen baby names in the United States are gender-specific, meaning they are used for a particular gender, either male or female. A small minority of chosen baby names are considered gender-neutral, androgynous or unisex, in that they can be used for either a male or a female (Seguin et al., 2021). Additional background content about the history of use of gender-neutral names is contained in [Appendix A, Background](#_Background).

# Objectives

This project involves an analysis of trends by decade for gender-neutral names in the United States. The objective is to explore, analyze and gain understanding towards insights about gender-neutral baby names based on the data contained in the "Baby Names from Social Security Card Applications - National Data" dataset for years 1880-2020, published by the Social Security Administration (Nuessel, 2017; SSA, 2021).

The results of this project would be beneficial to organizations that want to gain knowledge regarding usage of gender-neutral names in the U.S. This knowledge could help businesses who develop personalized products, software designers who utilize user name information for customer service purposes, and company human resources departments in hiring of new employees and for employee training purposes.

# Overview of Study

This study uncovers trends in the use of gender-neutral names over 14 decades, from 1880-2020, as well as uses statistical analysis via correlation tests to better understand these trends over time. This subject is of interest to the author of this paper, as growing up with a gender-neutral name has had few advantages and many disadvantages. Learning about use of gender-neutral names over many decades, and how popularity of using them has increased or decreased was the inspiration for this study.

# Research Questions and Hypotheses

**Research Questions**

Research questions for this project, shown in Table 1, involve the quantity and usage of a set of unique gender-neutral names per decade used by both males and females, with the exact same spelling.

**Table 1**

*Research Questions for Gender-Neutral Baby Names Project*



As mentioned in Research Question 1 and Research Question 2, the data that was collected and analyzed consists of names for which Social Security applications were submitted, which are the same for both males and females within the fourteen decades shown here: 1880-1889, 1890-1899, 1900-1909, 1910-1919, 1920-1929, 1930-1939, 1940-1949, 1950-1959, 1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2009, and 2010-2020. Names are considered to be gender-neutral if they appear in both the male and female list and have identical spelling. The number of occurrences of the name for each gender was tabulated and the ratio of occurrences of gender-neutral name per the total number of names was calculated per gender per decade (SSA, 2021).

**Hypotheses**

The null and alternative hypotheses for Research Questions 1 and 2 for this project are shown in Table 2.

**Table 2**

*Hypotheses for Research Questions 1 & 2*



# Literature Review

A literature review was conducted to find sources with relevance to this research project. The Colorado State University Global online library, with key words of *baby names*, *unisex names*, *gender-neutral names*, and *androgynous names* was used*.* Topics of the relevant papers include the evolution of gender-neutral names, cultural change in naming of children due to random drift, the instability in the use of gender-neutral names, and its opposite view of the stability in the use of gender-neutral names.

**Evolution of Usage of Gender-Neutral Names**

Barry & Harper (1982) studied how the use of gender-neutral names has evolved through time. Their hypothesis was that all gender-neutral names start as a masculine name, evolve to a unisex name, then become a feminine name. Their belief was that parents would more often name their female child with a masculine name rather than name a male child with a feminine name due to cultural attitudes and stereotyping.

The reference sources utilized by Barry & Harper (1982) consisted of six baby name books from the 1930s/1940s and the 1960s/1970s. While the research showed that some classic masculine names became unisex names, it was rare and there was not a clear progression from masculine to unisex to feminine as was suspected per their hypothesis. One of the biggest finds from the research was that gender-neutral names made popular by famous people, like movie actors and musicians, were often used for children.

**Random Drift in Cultural Change**

Hahn & Bentley (2003) studied how cultural drift affects the naming of children. The term *drift* in this source refers to the way that "functionally neutral cultural elements come and go in society" (Hahn, M., & Bentley, R., 2003, p. S120, para. 1). They note that when naming children, parents introduce new names often, while other names fall into dis-use. This occurs in nearly every decade. The research determined that large increases or decreases in the frequency of usage of specific names are common because names are copied from others, such as through use of baby name books.

**Instability of Gender-Neutral Names**

Lieberson et al. (2000) wrote of the rise of the use of gender-neutral, or *androgynous*, names as a result of feminism and the drive towards achieving equality between men and women. This research involved a small subset of all baby names, based on data from one U.S. state and one race of children. The prediction for the research is similar to the evolution concept by Barry & Harper (1982). Lieberson et al. (2000) state that gender-neutral names are unstable because they start as masculine names, become androgynous, then end up being feminine names. The conclusion of the research revolves around the lifespan of particular gender-neutral names over time and how these names drop out of popular use more quickly than gender-specific names, which shows proof of their instability.

**Stability of Gender-Neutral Names**

Seguin et al. (2021) argue against the theory of instability of gender-neutral names, as was researched by Lieberson et al. (2000). Seguin et al. (2021) state that gender-neutral names are stable because popularity trends in names for boys and girls flow in tandem and that gender-neutral names remain in use longer than the individual use for female or male children. This research involved the Social Security applicant dataset from 1880-2016 and attempted to refute the claims from the Lieberson et al. (2000) paper using the larger dataset. This research claims to have been successful due to delving further into the use of gender-neutral names, such as the sound of the name. Name sounds, such as ending in a long "e" sound, could be a reason particular names became more feminine over time.

**Results of Literature Review**

The results of the literature review are that all four reference sources note similarities in their research into gender-neutral names. It is discussed that choice of baby names is highly dependent on popularity across the region and country, as well as use by famous people such as movie actors and musicians. Some of the references reviewed refer to one another: Barry & Harper (1982) is referred to in Lieberson et al. (2000); Seguin et al. (2021) refers back to Lieberson et al. (2000). All four references used different datasets for their name research, which resulted in differing conclusions in each case.

# Research Design

## Methodology

The dataset used for this project is quantitative in that it contains the names themselves and the quantity of occurrences of that name. The Social Security baby names dataset consists of one data file per Year-of-Birth (YOB) for the years 1880-2020. As each year passes, the baby names for that year are published and added to the dataset. Each YOB data file contains three columns in a Comma Separated Value (CSV) format, where female names are at the top, followed by male names. The top 1000 most popular names each for females and males are included in each YOB file. Two excerpts from the YOB data file for baby names from 1880 (*yob1880.txt)* are shown in [Appendix A, Figure A-1](#_Figure_A-1) (SSA, 2021).

Of note, each data file is ordered by highest occurrence of names. In the example YOB data file ([Appendix A, Figure A-1](#_Figure_A-1)), the name *Mary* had the highest number of Social Security applications for females with 7065 occurrences, meaning *Mary* was the #1 ranked name for females. The example YOB data file also shows an excerpt from the center of the *yob1880.txt* file showing the last 5 rows of female names plus the first ten rows of male names in the 1880 YOB file. In this portion, the name *John* had the highest number of Social Security applications for males with 9655 occurrences, meaning *John* was the #1 ranked name for males. Appendix A includes the complete [Type of Variables and Data Dictionary](#_Type_of_Variables_1) for the YOB files.

## Methods

#### Data Collection Methods

The dataset is publicly available and was downloaded from SSA (2021). The data for analysis was prepared by searching each YOB file for names that were used by both males and females with the exact same spelling. A complete YOB data file was used, so the data searched included the top 1000 names for males, and the top 1000 names for females for that year. In order to do this analysis per decade, the last year of the decade was used as the search set for the names that were used for both genders.

#### Data Analysis Methods

Tools for analysis in this project were R for data analytics and Microsoft Excel. R was used to ingest the raw name data, clean and organize the data for analysis, perform statistical calculations and produce plots as needed for relevant visual inspection (Davies, 2016). Microsoft Excel was used to ingest the resultant CSV data from the R data analysis for visual inspection in spreadsheet form, as well as to perform basic tasks such as alphabetizing and numerical ordering as needed for visual inspection.

The research method for this project was to analyze the SSA name dataset per year and decade using descriptive analytics to search for trends of gender-neutral name usage, as well as to visually show the trends in tables and plots. As an example, there were 91 gender-neutral names with exact same spelling for females and males found using this research method for the year 1959. This list of names was then used to search each individual YOB file from 1950-1959. The table containing the gender-neutral names for 1959 is shown in [Appendix A, Table A-2](#_Table_A-2_2).

Using a set of names per decade, as in the described example for 1959 ([Appendix A, Table A-2](#_Table_A-2_2)), allowed analysis of:

* the occurrences of males and females using this name versus total occurrences, per individual year
* the quantity of names used for either gender, versus total quantity of names

To prove or disprove the hypotheses for Research Questions 1 and 2, scatter plots were created to look for trends in the time-based data. The data tested was the occurrences of gender-neutral names versus the occurrences of all names for that gender. Using R, the Pearson Correlation test was also done per gender for all years in the dataset to evaluate association between the occurrences of gender-neutral names versus occurrences of all names, per gender (Davies, 2016).

## Limitations of the Dataset

Prior to the Tax Reform Act of 1986 (99th Congress, 1986), U.S citizens were required to apply to join the Social Security program. Nuessel (2017) notes that the names included in the Social Security dataset were gathered from Social Security card applications for those born after 1879. After the Tax Reform Act of 1986 was passed, all U.S.-born children's names from 1987 and beyond appeared in the data file for their year of birth. Nuessel (2017) provides two additional limitations about the Social Security names dataset that apply directly to this project:

* The dataset includes gender and spelling as was recorded on the Social Security application. No edits were done if incorrect.
* Names can be used for both genders, but will have differing ranks in popularity per gender.

## Ethical Considerations for the Dataset

The United Stated Department of Labor (US DoL)defines Personally Identifiable Information (PII) as "any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means" (US DoL, n.d., para. 1). Examples of PII that identify individuals directly are full name (first, middle and last), address, social security number, telephone number, and email address (US DoL, n.d.).

Howe & Elenberg (2020) state that computational analysis of datasets containing PII should only be used to advance knowledge and research by identifying trends and associations that were previously unknown. There are enormous ethical considerations that need to be taken into account when using datasets that include PII. The five principles of ethics laid out by Cote (2021), including Ownership, Transparency, Privacy, Intention and Outcome, are explained further in [Appendix A, Five Principles of Ethics](#_Five_Principles_of_1).

  For this project, there are no ethical concerns that need to be taken into account. The data fields in the SSA name dataset include only the individual's first name, their gender and their birth year. The dataset includes U.S. born applicants for the U.S. Social Security program, but no other geographic information is included. Additionally, care was taken by the SSA to include only names that appear five times or more, such that no individual could be directly identified and privacy is preserved (Nuessel, 2017; SSA, 2021).

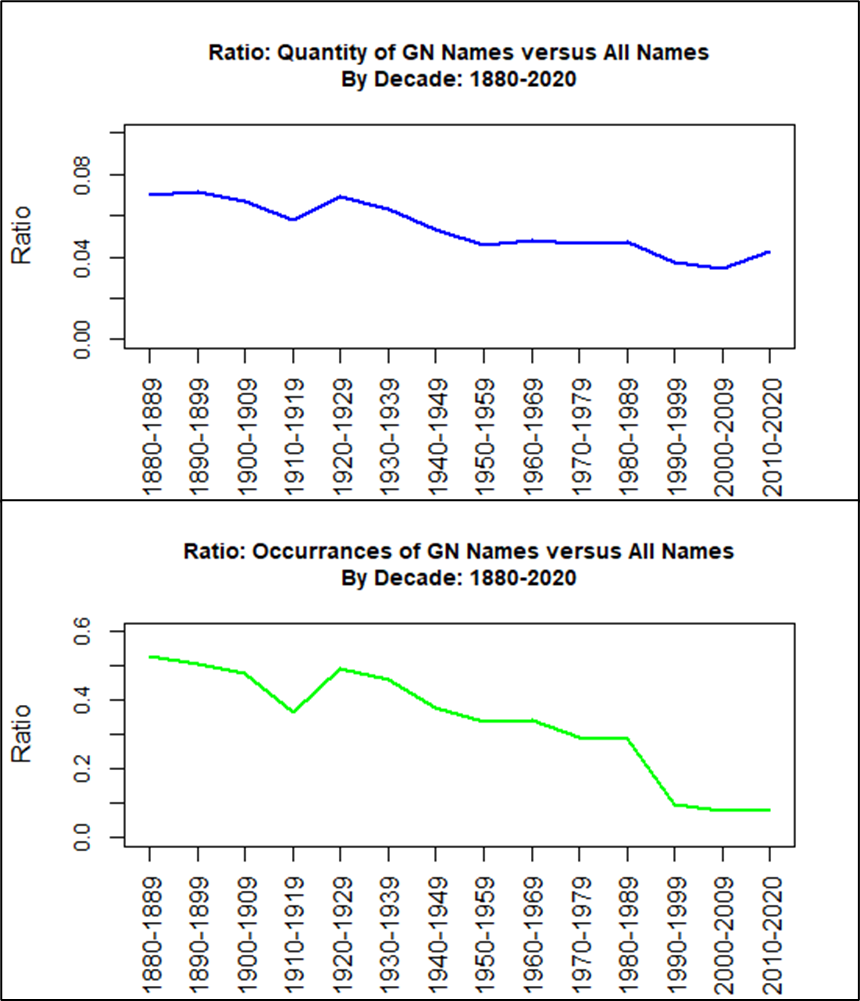
# Findings

**Findings: Research Question 1**

Research Question 1 asks about the quantity of unique gender-neutral names, with same spelling, used for both males and females. The quantity of gender-neutral names used per decade as well as the occurrences of gender-neutral names and the total occurrences of all names used per decade were compiled and are displayed in [Appendix A, Table A-3](#_Table_A-3_1). For ease of visualization of this data, the ratio of the quantity of gender-neutral names versus the total number of names per decade is shown on the upper portion of Figure 1. The ratio of the occurrences of gender-neutral names versus total occurrences of names is shown on the lower portion of Figure 1.

**Figure 1**

*Ratio Plots of Quantity of Gender-Neutral Names and Occurrences of Gender-Neutral Names*



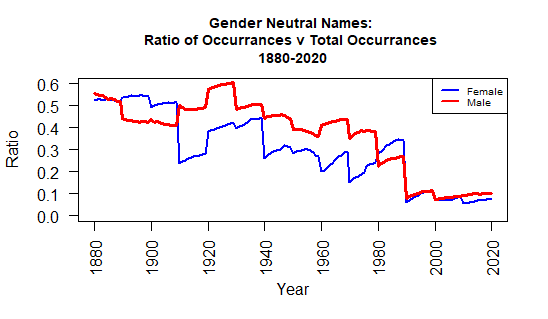
As is seen in the upper portion of Figure 1, the quantity of gender-neutral names used decreases from a ratio of 0.07, or 7%, in 1880 down to 0.043 or 4.3% in 2020. In looking at the lower portion of Figure 1, the ratio of occurrences of gender-neutral names being used goes from 0.53, which is 53%, in 1880 down to 0.084 or 8.4%, in 2020. Based on these visualizations it is clear that the null hypothesis for Research Question 1 can be rejected as there is a clear trend downward of gender-neutral names being used through the 140-year time period. This shows that as time has gone on, most parents prefer to name their children with distinct gender-specific names.

**Findings: Research Question 2**

Research Question 2 asks about the number of occurrences of unique gender-neutral names, with same spelling, used for both males and females. Preparation of the data towards analysis of the hypotheses for Research Question 2 included compilation of the sum of the number of occurrences of gender-neutral names per gender, versus the total occurrences of all names for that gender. Data for every year in the 140-year time period was compiled to develop a comprehensive picture of the trend. An example of the data compiled for the years of 1970-1979 is shown in [Appendix A, Table A-4](#_Table_A-4_1). The ratio of occurrences of gender-neutral names with the occurrences of all names was also calculated. A plot to visualize this ratio of occurrences per gender for all 140 years in the dataset is shown in Figure 2.

**Figure 2**

*Ratio of Occurrences of Gender-Neutral Names with Occurrences of All Names, per Gender*



As is seen in Figure 2, the ratio of occurrences varies per gender, with the female occurrences shown in blue and male occurrences shown in red. While the use of gender-neutral names was fairly even for both females and males in 1880, it is seen that the usage of gender-neutral names, which are considered to be more masculine, was more popular for males in the years and decades that follow than for females. It is also seen that there is a downward trend in use of gender-neutral names for either gender over the 140-year time period. This downward trend is a clear indicator that the null hypothesis for Research Question 2 can be rejected. The answer to Research Question 2 shows that as time has gone on, parents more rarely chose gender-neutral names for their children. As was stated previously, it appears that most parents prefer to name their children with distinct gender-specific names.

**Research Question 2: Correlation Tests**

As an additional method of analysis for Research Question 2, scatter plots were made and the Pearson correlation test was used to evaluate the association between occurrences of gender-neutral names and occurrences of all names, per gender. Figure 3 shows the scatter plots where the female occurrences are on the left portion and the male occurrences are on the right portion.

**Figure 3**

*Scatter Plots of Occurrences of Gender-Neutral Names versus All Names, per Gender*

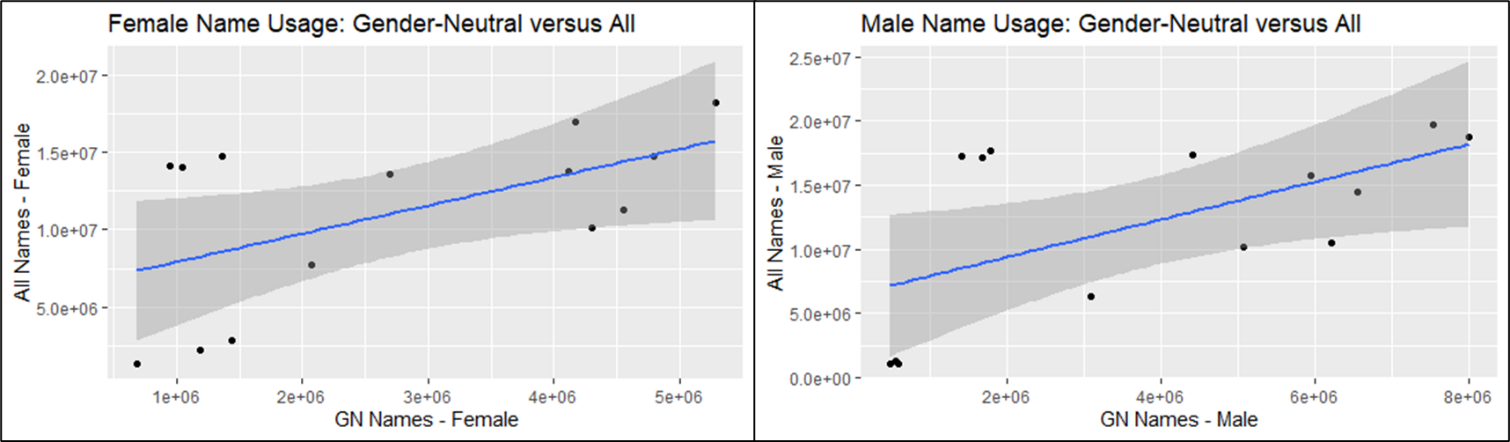


Table 3 shows the Pearson's correlation test results with female names in the left portion and male names in the right portion.

**Table 3**

*Pearson's Correlation Test Results*



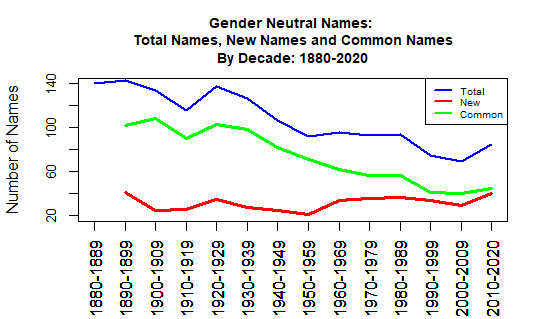
As can be seen in Table 3, the p-values are similar with females at 0.038 and males at 0.03. These p-values are very close to the significance level of 0.05. Therefore, the correlation of the occurrences of use of gender-neutral names versus all names are not statistically significant for either gender.

**Additional Analysis: Theory by Hahn & Bentley (2003)**

Based on the theory from Hahn & Bentley (2003) discussed in the Literature Review, which stated that new names are invented and others fall into dis-use with every decade, analysis was done beginning with the 1880-1889 decade to determine whether gender-neutral names in following decades were either in common with the previous decade, or were new for that decade. In some cases, a gender-neutral name was included in a decade, but then skipped one or more decades before appearing in the list again. If this happened, that particular name was considered common in some decades and new in a decade further on. This occurred with the name *Charlie*, which was considered common since 1889. The name falls into disuse as a gender-neutral name in the years from 1950-1999, but then reappears in 2000-2009. The name *Charlie* is then considered *New* as of 2009 when it reappears as a name used by both females and males. A table showing this is included in [Appendix A, Table A-5](#_Table_A-5). A plot of the quantity of gender-neutral names and their status of *New* or *Common* is shown in Figure 4 along with the total number of gender-neutral names used for the 14-decade time period.

**Figure 4**

*Names since 1889: Common and New per previous decade*

**

As shown in Figure 4, the total number of gender-neutral names is plotted in blue. Gender-neutral names that were in common with the previous decade are shown in green. This line trends downward as the decades went on. Gender-neutral names that were new in each decade are shown in red. Interestingly, the number of new gender-neutral names increases from 1950 to 2020. The number of new gender-neutral names nearly reaches the level of the gender-neutral names that were in common with the previous decade. It can be speculated that parents desired gender-neutral names for children, but did not want to use the same names that previous generations had used. They instead introduced new gender-neutral names. Based on analysis, these new gender-neutral names may have actually been brought forward from previous decades as in the case of the name *Charlie*.

### Conclusion

In conclusion, this study analyzed trends and correlation in usage of gender-neutral children's names over 14 decades, from 1880-2020. Research Question 1 involved the quantity of gender-neutral names being used overall. For this question, the null hypothesis was rejected, meaning that there was a difference in quantity of gender-neutral names being used per decade. The analysis showed there was a significant decrease in quantity of gender-neutral names being used from 1880-2020.

Research Question 2 involved the number of occurrences of gender-neutral names being used per gender versus all names. For this question, the null hypothesis was rejected, meaning that there was a difference in number of occurrences of gender-neutral names per decade, per gender. The analysis showed there was a significant decrease in occurrences of gender-neutral names being used from 1880-2020 for both males and females. Additionally, the correlation between usage of gender-neutral names versus all names was not statistically significant for either males or females.

Lastly, analysis was done on gender-neutral names being commonly used as decades go on versus new gender-neutral names being introduced. This analysis showed that while usage of many common gender-neutral names decreased and names fell into dis-use, there was an increase in new gender-neutral names being added over the last 50 years.

# Recommendations

Further analysis of trends in the use of gender-neutral names in the United States in the future will require addition of new YOB files as more are published when each year passes. It is recommended that future analysis be done at the end of each decade to enable gathering of the most useful set of data, which includes a full decade worth of name data.

### References

Associated Press. (2018, March 21). Why gender-neutral baby names are on the rise. *New York Post*. https://nypost.com/2018/03/21/why-gender-neutral-baby-names-are-on-the-rise

Barry, III, H. & Harper, A. (1982). Evolution of Unisex Names. *Names: A Journal of Onomastics, 30*(1), 15–22. https://doi.org/10.1179/nam.1982.30.1.15

Cote, C. (2021). 5 principles of data ethics for business. *Harvard Business School Online*. https://online.hbs.edu/blog/post/data-ethics

Davies, T. (2016). *The Book of R, A First Course in Programming and Statistics*. San Francisco, CA: No Starch Press.

Hahn, M., & Bentley, R. (2003). Drift as a mechanism for cultural change: An example from baby names. *Proceedings: Biological Sciences, 270*, S120–S123. https://doi.org/10.1098/rsbl.2003.0045

Howe III, E., & Elenberg, F. (2020). Ethical Challenges Posed by Big Data. *Innovations in Clinical Neuroscience, 17*(10-12), 24–30.

Lieberson, S., Dumais, S., & Baumann, S. (2000). The instability of androgynous names: The symbolic maintenance of gender boundaries. *American Journal of Sociology, 105*(5), 1249–1287.

Nuessel, F. (2017). A note on popular baby names on the Social Security website: An important onomastic resource*. Names: The Journal of the American Name Society, 65*(1), 45–50. https://doi.org/10.1080/00277738.2016.1197651

O'Leary, Z. (2021). *The essential guide to doing your research project*. SAGE Publications Ltd.

Seguin, C., Julien, C., & Zhang, Y. (2021). The stability of androgynous names: Dynamics of gendered naming practices in the United States 1880–2016. *Poetics, 85*. https://doi.org/10.1016/j.poetic.2020.101501

Social Security Administration (SSA). (2021). Baby names from Social Security card applications - National data. *Data.gov*. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data

United States Department of Labor (US DoL). (n.d.) Guidance on the Protection of Personal Identifiable Information. *Dol.gov*. https://www.dol.gov/general/ppii

# Appendix A

**Additional Content**

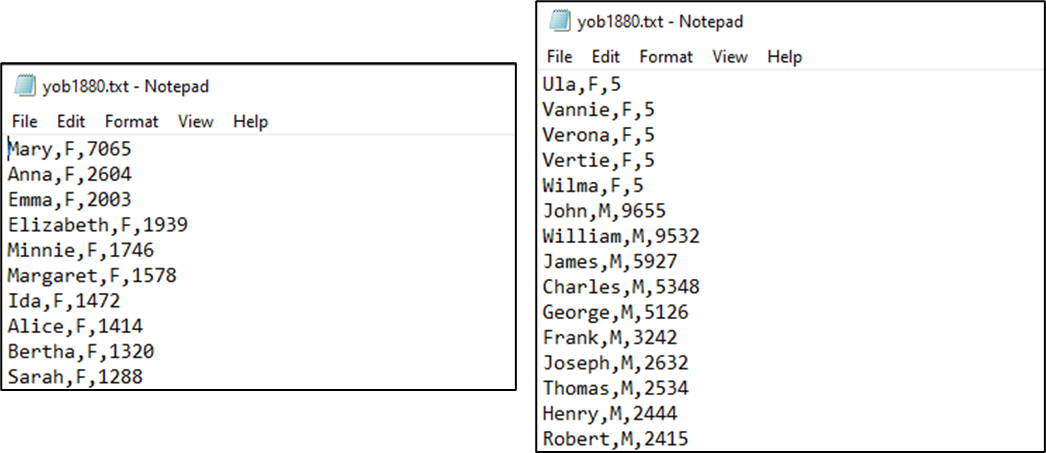
#### Background

A parent's decision to use a gender-neutral name for their baby can occur for various reasons. Some parents want to lessen the emphasis on gender differences in their child's world and prevent potential gender bias for them in later life. Other parents believe that choosing a more masculine name for daughters will lead to an impression of strength versus more traditional feminine sounding names (Lieberson et al., 2000). Even others will choose a gender-neutral name because they believe it would sound good for either gender of child, and it saves the time and potential frustration of having to choose both a male name and a female name.

While the individual names themselves change throughout time, the idea of using gender-neutral names for children has existed for centuries. From Lieberson et al. (2000), names of male saints were given to both sons and daughters in the 13th century in England. Conversely, in Europe, names of female saints were used for both females and males. Sometimes parents would even use a boy's name for a daughter to show their disappointment in not having a son (Lieberson et al., 2000). Gender-neutral naming is not accepted worldwide however. In modern times, some countries such as Portugal, Denmark, and Iceland forbid the use of unisex names by law. In Germany, local registrars can make the decision to go against the parent's choice of their child's name if they determine the name is too unusual and could negatively impact a child's life (AP, 2018).

#### Figure A-1

*Excerpts from yob1880.txt data file*

**

#### Type of Variables and Data Dictionary

The data file format is comma separated and consists of three columns. The data files being used do not have a header row that describes the labels for each column. Table A-1 shows the data dictionary for the SSA YOB data files. The data names in Table A-1, column 2, were assumed based on Nuessel (2017).

#### Table A-1

*Data Dictionary for the SSA YOB data files*



#### Table A-2

*Table of Gender-Neutral Names for 1959*



#### Five Principles of Ethics by Cote (2021)

1. Ownership: the individuals own their PII (Cote, 2021)
2. Transparency: the individuals are informed how their PII is collected, how and where it will be stored and how it will be used (Cote, 2021)
3. Privacy: If the individual gives consent of use of their PII, this data is not considered publicly available. It is still private and only for the uses disclosed to the individuals (Cote, 2021)
4. Intention: The collector of the PII needs to know why it is needed, what insights will be gained from it, and how this data will be used to make changes (Cote, 2021)
5. Outcome: It is on the collector of the PII to understand if their analysis will cause inadvertent harm to the individuals the PII was collected from (Cote, 2021)

#### Table A-3

*Quantities and Occurrences of Gender-Neutral Names per Decade*



#### Table A-4

*Occurrences of Gender-Neutral versus All Names for Males and Females from 1970-1979*



#### Table A-5

*Example: New and Common Name Determination for Name = Charlie*

