Diary Entry

Loy Yee Keen

2023-11-16

Week 9

(1) What is the topic that you have finalised? (Answer in 1 or 2 sentences).

The topic that I have finalised is “Names”, specifically “Is there a rise in the gender-neutral names given to babies born in America?”

(2) What are the data sources that you have curated so far? (Answer 1 or 2 sentences).

I have curated a dataset consisting of a list of names given to babies born in the US each year, the gender of the babies, and the count of each name per gender. The data sources span from the years 1880 to 2022. These data sources are extracted from the website of the United States Social Security Administration, an independent agency of the U.S. federal government.

Week 10

(1) What is the question that you are going to answer? (Answer: One sentence that ends with a question mark that could act like the title of your data story)

Is there a rise in the gender-neutral names given to babies born in America?

(2) Why is this an important question? (Answer: 3 sentences, each of which has some evidence, e.g., “According to the United Nations…” to justify why the question you have chosen is important)

The Council of Europe underscores the role of gender in shaping power dynamics and opportunities in society. The popularity of gender-neutral names reflects a broader shift toward inclusivity and the challenge of traditional gender roles.

Source: https://www.coe.int/en/web/gender-matters/exploring-gender-and-gender-identity#:~:text=Gender%20is%20of%20key%20importance,equality%20and%20freedom%20from%20discrimination (https://www.coe.int/en/web/gender-matters/exploring-gender-and-gender-identity#:~:text=Gender%20is%20of%20key%20importance,equality%20and%20freedom%20from%20discrimination

Additonally, gender-neutral names empower girls and women by challenging gender stereotypes. According to a New York Times article, some parents opt for these names to counter biases and promote strength for their daughters.

Source: https://nypost.com/2018/03/21/why-gender-neutral-baby-names-are-on-the-rise/ (https://nypost.com/2018/03/21/why-gender-neutral-baby-names-are-on-the-rise/)

This trend aligns with the United Nations’ Sustainable Development Goal 5, which aims to achieve gender equality and empower women and girls.

Source: https://sdgs.un.org/goals/goal5 (https://sdgs.un.org/goals/goal5)

(3) Which rows and columns of the dataset will be used to answer this question? (Answer: Actual names of the variables in the dataset that you plan to use).

I will use multiple datasets to answer the chosen question. All the datasets have the same format; each dataset represents a specific year spanning from 1882 to 2022.

In each dataset, there are 3 columns, corresponding to the name of the baby, the sex of the baby and the count of babies with that name (The original dataset did not define the names of the variables, so I will redefine the variables as Name, Sex and Count).

The number of rows, each corresponding to the observation for each name, differ for every year.

Every row and column of the datasets will be used to answer my chosen question as all of them are relevant in comparing the shifts in naming trends over the years.

I will use rbind to combine the datasets into a single dataset.

(4) Challenges and errors that you faced and how you overcame them.

I encountered difficulties when I read the files using read_csv because the datasets did not have column names. Consequently, the output assigned the first value in each cell of the respective columns as the column names. This approach was erroneous, as the data in the first row represented observations, not variables. To resolve this issue, I tried to look up the answer in the textbook reading (https://r4ds.hadley.nz/data-import (https://r4ds.hadley.nz/data-import)) provided in the Lecture 9 slides.

Week 11

(1) List the visualisations that you are going to use in your project

- I. Barplot of proportion of US babies with gender-neutral names by year:
    - Variables: y axis = Proportion of gender-neutral names; x-axis = Year.
    - Purpose: Investigate the trend over time, determining whether there is an increase in the use of gender-neutral names for babies in the US.

II. Table of gender-neutral names in 2022:

- Columns: Gender-neutral names in 2022, count of male babies, count of female babies, total count and the proportion overlap between the counts of both sexes.
- Purpose: Provides a more detailed visualisation of the specific names of which the trends would be plotted in III.

III. 5 line plots of the trend of gender-neutral name over time, grouped by gender:

- Variables: x-axis: Proportion of babies with names that at least 90% male-female overlap, y-axis: Year
- Purpose: Analyse the trends of these names, understanding whether these names have been consistently gender-neutral or have transitioned from a more gender-specific association over time. This could possibly answer the question of why there is a rise/fall in gender-neutral names over time.

(2) How do you plan to make it interactive?

- I. Features:
    - A slider for adjusting the number of bars in the barplot.
    - Radio buttons displaying the proportion of male and female babies with gender-neutral names.

II. Features:

- A button to highlight values in the table which have more than 90% male-female proportion overlap.
- numbericInput function to choose the number of rows to display in the table.

III. Features:

- SelectInput function to choose a name and display the corresponding plot.
- Forward (and backward) navigation buttons to show the plot for the next 50 years.
- A card providing an explanation of the plot, updating itself when the navigation buttons are pressed.
- Tab panel that displays all the 5 plots in a single view (using facet wrap).

(3) What concepts incorporated in your project were taught in the course and which ones were self-learnt?

```
## Warning: package 'readxl' was built under R version 4.3.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## # A tibble: 61 × 2
##    Topic                                                    Week
##    <chr>                                                    <chr>
##  1 library                                                  2
##  2 pull                                                     2
##  3 logical operators (==, &, <, >)                          2 & 4
##  4 ggplot                                                   2 & 7
##  5 as.integer                                               3
##  6 as.character                                             3
##  7 vector('list', length=)                                  3
##  8 list[['']]                                               3
##  9 c()                                                      3
## 10 read_csv                                                 3
## 11 $                                                        3
## 12 filter                                                   4
## 13 select                                                   4
## 14 mutate                                                   4
## 15 arrange(desc())                                          4
## 16 seq(from=, to=, by=)                                     4
## 17 slice                                                    4
## 18 : eg 1:5                                                 4
## 19 ifelse                                                   4
## 20 functions                                                5
## 21 paste0                                                   5
## 22 for loop                                                 6
## 23 aes(group=)                                              7
## 24 aes(colour=)                                             7
## 25 geom_col(fill=)                                          7
## 26 geom_col(alpha=)                                         7
## 27 guides()                                                 7
## 28 facet_wrap(~)                                            8
## 29 shiny                                                    8
## 30 sliderInput                                              8
## 31 unlist                                                   <NA>
## 32 abs                                                      <NA>
## 33 scales = 'free_y'                                        <NA>
## 34 geom_line                                                <NA>
## 35 read_csv(col_names=)                                     <NA>
## 36 merge                                                    <NA>
## 37 data.frame                                               <NA>
## 38 rbind                                                    <NA>
## 39 bind_rows                                                <NA>
## 40 do.call                                                  <NA>
## 41 css                                                      <NA>
## 42 is.null                                                  <NA>
## 43 reactiveVal                                              <NA>
## 44 observeEvent                                             <NA>
## 45 plotly                                                   <NA>
## 46 DT package                                               <NA>
## 47 scale_colour_manual                                      <NA>
## 48 guide_legend(title = NULL)                               <NA>
## 49 geom_col                                                 <NA>
## 50 as.double(digits=2)                                      <NA>
## 51 scale_x_continuous/scale_y_continuous (limits=, breaks=) <NA>
## 52 radioButtons                                             <NA>
## 53 numericInput                                             <NA>
## 54 actionButton                                             <NA>
## 55 div                                                      <NA>
## 56 tabs                                                     <NA>
## 57 switch                                                   <NA>
## 58 min                                                      <NA>
## 59 max                                                      <NA>
## 60 backticks                                                <NA>
## 61 sum                                                      <NA>
```

Explanations for some of the functions I used are as follows:

filter(): filter Sex == “M” and filter Sex =“F” from the original dataset and store in a new variable each

mutate(): create new columns in the dataframe called `Count of Female Babies` , `Count of Male Babies` , `Total Count` and `Proportion Overlap Between M & F`

arrange(desc()): sort dataframe in descending order of `Total Count`

seq(from=1882, to=2022, by=10): set the breaks of the bar plot to be in intervals of 10 from 1882 to 2022

ifelse: if radio buttons are clicked, display its corresponding plot in the output of distPlot1 highlight the names in the shiny output table if `Proportion Overlap Between M & F` is more than 0.9 plot the lines for the next 50 years and update the text if the forward button is pressed, reverse if backward button is pressed print the text corresponding to a particular name if that name is selected in the selectInput function on shiny

functions function to filter male and female names function to merge the dataframes for male and female into one dataframe called gender-neutral names (using **merge()** function) function to read dataset for each year (see paste0 function below)

paste0(): concatenate the strings: “yob”, year & “.txt” as they have no separators

```
read_year_data <- function(year) {
  data <- read_csv(paste0("yob", year, ".txt"), col_names = c("Name", "Sex", "Count"))
}
```

for loop: execute functions across multiple names and multiple years from 1882 to 2022

geom_col(): bar plot of the proportion of gender-neutral names for each year and another bar plot of the proportion of names occurring less than 50 times for each year (geom_col is used as y variable is proportion not count)

geom_col(fill=): fill the bars with different colors based on the male and female proportions

geom_col(alpha=): to make the plot translucent so the overlapping colours for male and female proportions can be compared more clearly

ggplot(data)+aes(group=Sex, color=Sex)+geom_line(): plot a line plot of proportion of babies with a particular gender-neutral name, with 2 separate lines of different colours for male and female babies.

facet_wrap(~ Name, scales = “free_y”): plot each name in a separate plot and show all the plots at the same time and allow the scales of the y-axis to vary between facets based on the data for each ‘Name’ otherwise some of the y-axes will be so large that the plot cannot be seen clearly

reactiveVal(): store the selected name in selectInput

observeEvent(): observe changes in this stored value (the selected name), and updates the plot/text whenever the user selects a new name or presses a button

switch In this code chunk,

```
selected_data <-
switch(input$proportion_per_sex,
       "Male" = filtered_proportion_m_df,
       "Female" = filtered_proportion_f_df
      )
```

the value of input$proportion_per_sex determines which alternative is selected. If it's "Male," selected_data will be assigned the value of filtered_proportion_m_df; if it's "Female," it will be assigned the value of filtered_proportion_f_df

rbind: combined_name_trend_df <- do.call(rbind, name_trend_list) Combine name_trend_list for all the years into a single dataframe, combined_name_trend_df

(4) Include the challenges and errors that you faced and how you overcame them.

The first error I faced was when I used a "for loop" that loops through each dataframe for the years 1882 to 2022, and merges the names that are given to both male and female babies. However, when I tried to access the output outside the loop, I was returned "Error: object 'gender_neutral_names' not found".

To overcome this error, I went back to lecture 6 about "for loops" and recalled that we had to pre-allocate space to store the output. I then corrected the code by adding

```
gender_neutral_names <- vector("list", length =1882:2022)
```

before the loop.

The next error I faced was when I tried to access a year from the gender_neutral_names list using gender_neutral_names("2022"), but was returned with Error in gender_neutral_names("2022") : could not find function "gender_neutral_names".

I then referred to lecture 3 and realised that we need to access columns in a list using gender_neutral_names[["2022"]] instead.

Another error I faced was when I tried to arrange the data in a column named "Total Count". However, when I printed the output, the data was not arranged.

To overcome this challenge, I went online to search for how to reference to column names that include spaces and found out that we need to use backticks around the column name.

Week 12

Challenges

For my first visualisation, I tried to plot a barplot of the proportion of gender-neutral names given to babies per year using

```
ggplot(gender_neutral_names) +
      aes(x = year, y = proportion) +
      geom_bar()
```

However, I was returned with "Error occurred in the 1st layer. Caused by error in `setup_params()` :! `stat_count()` must only have an x or y aesthetic".

To overcome this error and plot the barplot successfully, I copied this error into Google. The first result came from stackoverflow which said that I need to either use geom_col() or geom_bar(stat="identity"). This is because the default for geom_bar() is (stat=count) which counts the aggregate number of rows for each x value (year), but I am instead providing the y-values (proportion) for the barplot.

I also included radio buttons (none/male/female/both) that fill this gender-neutral barplot according to the proportion of male/female/both that are given gender-neutral names.

I initially tried to do this by binding these 3 dataframes together using rbind, and then write a code using ifelse, something like

```
if(input$proportion_per_sex=female) {
plot<-ggplot(total_proportion)+
aes(x=year, y=proportion, fill=female_proportion)
} else if(input$proportion_per_sex=male) {
plot<-ggplot(total_proportion)+
aes(x=year, y=proportion, fill=male_proportion)
} else {plot<-total_proportion)
+aes(x=year, y=proportion)
}
```

but this doesn’t work since female_proportion and male_proportion are separate columns in the combined dataframe and fill works by colouring binary factor variables within the same column

Instead, I googled if it was possible to layer multiple ggplots in the same plot using eg

```
ggplot() +
geom_col(data = total_proportion) + aes(x = year, y = proportion) +
geom_col(data = male_proportion) + aes(x = year, y = proportion) +
geom_col(data = female_proportion) + aes(x = year, y = proportion)
```

Apparently it was possible, so I used this method instead.

For my second visualisation, I had 2 navigation buttons in the side panel that display the line plot when pressed. For example, the original plot (without pressing any buttons) displays an empty plot with x-axis from 1882 to 2022. When the forward button is pressed once, it displays the line from 1882 to 1932 (50 years), then when it is pressed again, it displays from 1882 to 1982 (next 50 years). To accumulate the years, I tried to use “<<-” which I learnt during Week 5’s tutorial. My original code was interval_end <<- interval_end + 50. However, the plot is empty when I rendered it.

Thus, for this visualisation to work effectively, I created three functions, one to store the interval end, one to store the cumulative end, and another one to accumulate the years.

```
interval_end <- reactiveVal(1882)
cumulative_end <- reactiveVal(1882))
new_cumulative_end <- cumulative_end() + 50
```

I also faced another challenge for this plot. Even if a name only has records starting from a specific year that is not 1882 (for instance, 1960), the plot does not start the line from 1960 but instead extends the line back 1882, which would otherwise erroneously imply the name’s existence before it was actually established. Also, the colour aesthetic for the plot gets mixed up. For instance, if a name only has records for male babies in the 1960s, the plot displays a pink line to represent this trend. However, when there are records for both sexes in the later years, say 1980, the plot displays 2 lines, but now the male line is blue while the female line is pink. Hence, I included a code chunk to add two rows to the dataframe (one for each sex), setting the proportion to 0 when the year is 1882.

```
filtered_df <- filtered_df %>%
   bind_rows(
    data.frame(Name = selected_name, Year = 1882, Sex = "M", Proportion = 0),
   data.frame(Name = selected_name, Year = 1882, Sex = "F", Proportion = 0)
 )
```

Final Write-up

Theme of Datastory and its Importance

Focusing on the question of whether there is a growing prevalence of gender-neutral names in the U.S., this datastory explores names as indicators of evolving societal attitudes. If there is indeed a growing prevalence, it may signify a society embracing gender fluidity ((Kihm, n.d., as cited in Mustafa, 2023) and challenging traditional gender stereotypes (Mahdawi, 2016), holding implications for fostering inclusivity and advocating an egalitarian culture. By unravelling the trends in baby naming, we gain insights into the broader narrative of societal progress.

Curated Data Sources

To answer this question, 141 datasets spanning from 1882 to 2022 were compiled. These datasets not only cover a vast timeline but are also sourced from the official records of the Social Security Administration, a federal agency of the U.S. government. Thus, they serve as a comprehensive and reliable source for analysing the historical trends of names given to babies born in the U.S.

To streamline the analysis, the following criteria were established to define gender-neutral names The difference in counts between both sexes is less than 300 The count for either sex is less than double that of the opposite sex

Implementation of the Project

First Plot

In exploring the evolving prevalence of gender-neutral names and aiming to visually represent the proportion of babies given gender-neutral names over the years, I opted for a barplot due to its clarity in illustrating the distribution of data. However, given that proportion is represented on the y-axis, I used geom_col() instead of geom_bar(). Unlike geom_bar(), which counts the frequency of occurrences on the x-axis and plots them on the y-axis, geom_col() allows for the direct plotting of the pre-calculated values on the y-axis.

To elevate the user experience, I introduced a slider that allows users to adjust the number of bars displayed. Alongside this, I integrated radio buttons, utilising the switch function for seamless toggling between visualisations of proportions for male or female babies with gender-neutral names. These visualisations overlay the original plot, allowing users to compare the distribution of gender-neutral names among male and female babies.

To further enrich the visual representation, I introduced an additional radio button labeled “Both”, allowing users to concurrently observe the overlay of male and female proportions of babies with gender-neutral names on the original plot. However, as the switch function is designed for toggling between distinct visualisations, I drew upon the concepts introduced in Lectures 2 and 7 that ggplots are composed of layers, and self-experimented with the feasibility of using multiple + geom_col() functions to layer the plots of both male and female proportions.

Insights from First Plot

The barplot depicted a significant rise in the popularity of gender-neutral names over the last 140 years. Interestingly, the proportion of gender-neutral names in the late 1800s was quite high, but the trend declined, reaching its lowest point around the 1960s. It was not until the late 1900s that the trend experienced a resurgence and has since increased exponentially, more than doubling since the 1800s. This provides testament to the shift in preference for gender-neutral names over the years, which is perhaps reflective of the increased acceptance of gender fluidity, prompting parents to seek names that transcend traditional gender boundaries.

The plot also reveals a shift in the gender distribution of these names. Initially, during the late 1800s to early 1900s, gender-neutral names were more prevalent among male babies. However, the proportion has since become more spread out across both sexes, often more common among female babies, especially during the mid-late 1900s. This shift suggests that parents are increasingly open to the idea of giving their daughters gender-neutral names, challenging the traditional stereotypes that dictate how names should align with gender norms. This shift perhaps reflects a broader societal advancement toward a culture that prioritises gender equality.

Second Plot

To delve deeper into the historical associations of names currently considered gender-neutral and understand their evolution—whether they were consistently popular among both sexes or if there was a historical bias towards a particular sex, potentially more commonly associated with males—I studied the trends of the top 10 most popular gender-neutral names of 2022. However, to ensure these names are popular across both sexes, I only included those with a proportion overlap of more than 90% in the counts of both male and female babies. There were 5 such names out of the 10.

To effectively showcase these trends, I opted for a line plot for each of the names, grouped and coloured by sex to facilitate a clearer comparison between the sexes. A dropdown menu allows users to select a name, instantly displaying its corresponding plot. Each plot is accompanied by text explanations within a div element—a self-learned concept. I also incorporated navigation buttons that, when clicked, reveal subsequent 50-year intervals of the plot. The reactiveVal() function, which was also learned independently, accumulated the years when these buttons were clicked. Additionally, I introduced another tab displaying all the plots combined in a single view to facilitate a more effective comparison of the trends across the 5 names.

However, given the extensive time span of 140 years on the x-axis, there were certain parts that were hard to see, hence I explored options for zoom functionality. My research led me to discover that Plotly offers a solution for this purpose.

Insights from Second Plot

The plots provide some support for the hypothesis that names currently considered gender-neutral used to be more commonly associated among males. 3 out of the 5 names (Blake, Charlie, and Finley) were originally male-associated but have seen a recent surge in popularity among female babies. Blake and Charlie used to be highly popular among male babies, but this popularity has fallen drastically over the years. As of 2022, more female babies are named Charlie than male babies, and the same trend is also observed for Finley. On the contrary, none of the 5 names were predominantly female-associated in the past.

This trend could potentially be attributed to parents wanting to challenge traditional gender stereotypes, as proposed earlier. This could be due to potential advantages in the workplace for females with less explicitly feminine names, as they are seen as more competent (Mahdawi, 2016). In fact, 2 out of the 5 names mean “warrior”, projecting a sense of capability.

Conclusion:

The rise of gender-neutral names in the U.S. mirrors a broader societal shift towards inclusivity, challenging traditional gender boundaries. In exploring these trends, this datastory not only uncovered historical trends but also provided insights into the evolving associations of specific names. As society embraces more egalitarian values, the choice of gender-neutral names stands as a testament to this evolving cultural landscape.

References

Mustafa, T. (2023, September). The gender neutral baby names taking the UK by storm. Metro. https://metro.co.uk/2023/09/27/unisex-baby-names-rise-as-parents-want-gender-neutral-options-19544405/# (https://metro.co.uk/2023/09/27/unisex-baby-names-rise-as-parents-want-gender-neutral-options-19544405/#)

Mahdawi, A. (2016, September). Initial impressions: will hiding my name force gender equality in the workplace?. The Guardian. https://www.theguardian.com/lifeandstyle/2016/sep/29/gender-neutral-name-equality-work (https://www.theguardian.com/lifeandstyle/2016/sep/29/gender-neutral-name-equality-work)

Social Security Administration. (n.d.). National Data. (1882-2022). https://www.ssa.gov/oact/babynames/limits.html (https://www.ssa.gov/oact/babynames/limits.html)

Word Count: 1191