HW1_DataViz_Evans

Maryanne Evans

4/17/2022

```
#install.packages("babynames")
library(babynames)
```

```
## Warning: package 'babynames' was built under R version 4.1.3
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(colorspace)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

Looking at the data to determine what questions to ask:

```
?babynames
```

```
## starting httpd help server ... done
```

```
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

```
nrow(babynames)
```

```
## [1] 1924665
```

```
ncol(babynames)
```

```
## [1] 5
```

There is obviously an outrageous amount of data in this data frame, so I am inclined to immediately subset this considering my five variables before continuing on. . .

Subsetting and identifying the variables of interest in the creation of new dataframes is critical for this analysis.

```
newdata <- babynames

#look at the data frame
nrow(newdata)
```

```
## [1] 1924665
```

```
ncol(newdata)
```

```
## [1] 5
```

```
head(newdata)
```

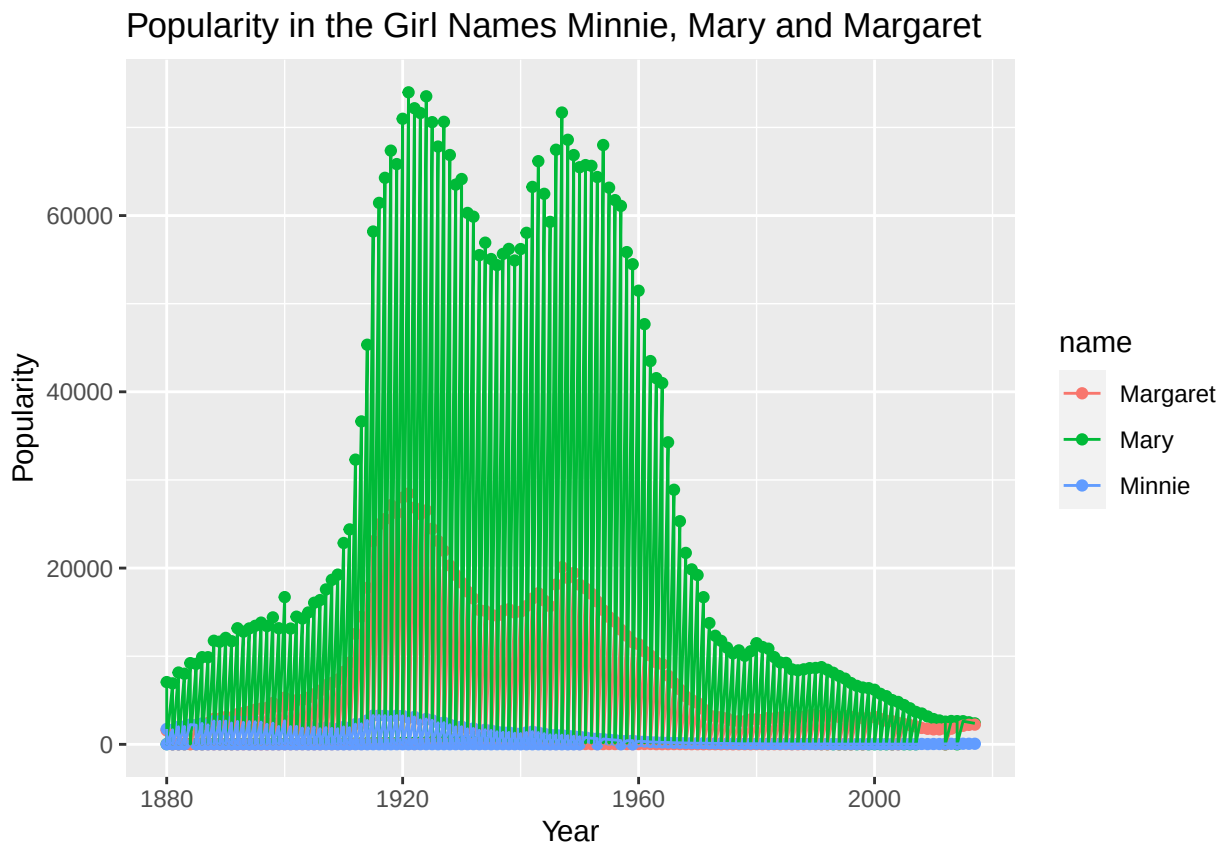
```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

```
#I am curious only about the trend over time for any given certain names (Mary, Minnie, Margaret), so I

marydf <- newdata[newdata$name == "Mary",]
margaretdf <- newdata[newdata$name == "Margaret",]
minniedf <- newdata[newdata$name == "Minnie",]
#combining
both_names <- rbind(marydf, margaretdf)
all_names <- rbind(both_names, minniedf)

#reframing the same question but with more interesting, gender-neutral names
#new data subset
alexdf <- newdata[newdata$name == "Alex",]
charliedf <- newdata[newdata$name == "Charlie",]
parkerdf <- newdata[newdata$name == "Parker",]
#combining
neutral_names <- rbind(alexdf, charliedf)
their_names <- rbind(neutral_names, parkerdf)

#if interested in seperating out F and M in columns
#true_names <- all_names %>%
#  pivot_wider(names_from = sex, values_from = sex)

#head(true_names)
```

Overall Question: Is there a relationship between the proportion of baby names and the gender of the babies born in the USA from years 1880 to 2017? In other words, does time change the outcome of popularity (frequency babies are named a certain name) for these names? And how is this trend reflected by gender?

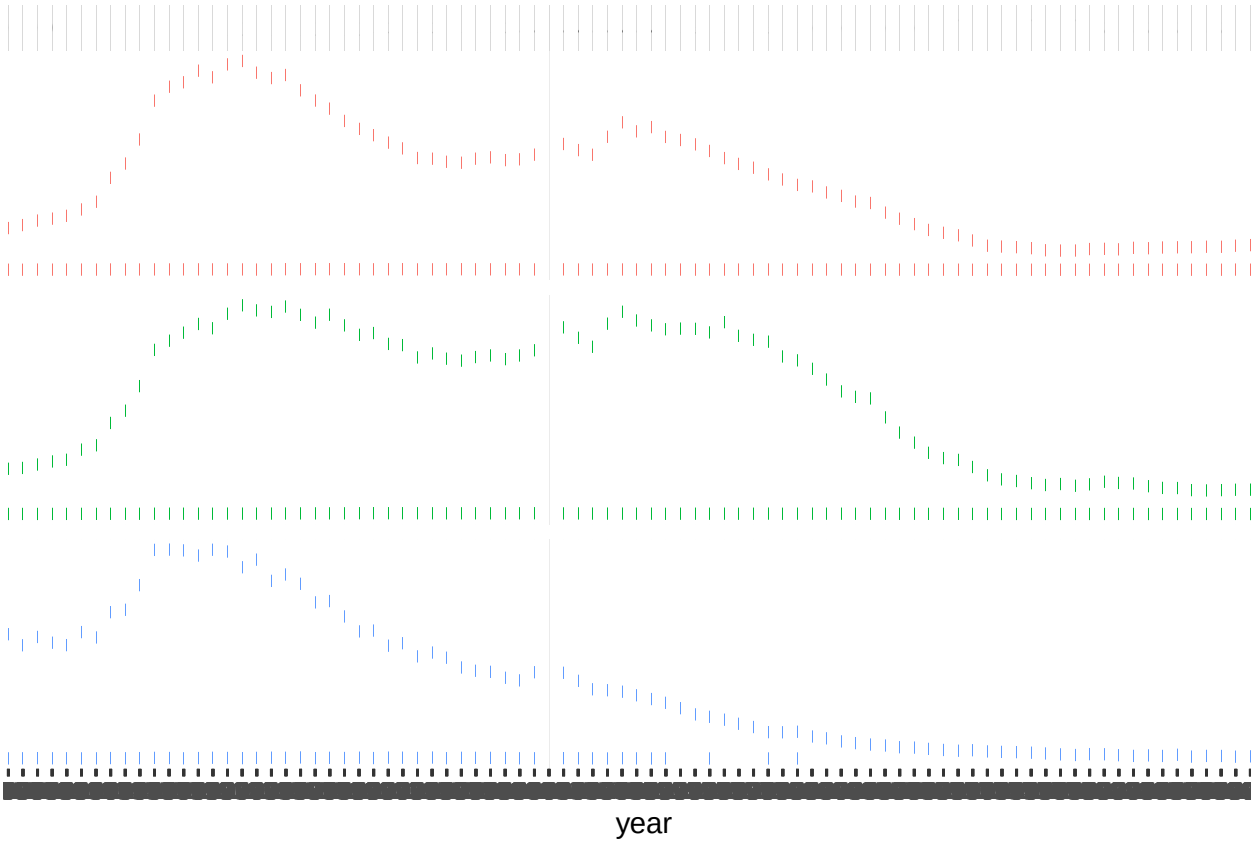
Plot 1 is the first attempt to visualize these data. The plot tries to simply group by name using color. But, unfortunately fails to distinguish between sex, leaving the visual chaotic, misleading and challenging to interpret. The story of these data are lost with this first attempt.

```
plot1 <- ggplot(data = all_names) +
  geom_point(mapping = aes(x = year, y = n, color = name)) +
  geom_line(mapping = aes(x = year, y = n, color = name)) +
  labs(title = "Popularity in the Girl Names Minnie, Mary and Margaret") +
  xlab("Year") +
  ylab ("Popularity")

plot1
```



Plot 2 attempts to separate out the variables using facet(). Though this function is useful, the visual seems busy and complicated. Per suggestion, using shape is an effective way to parse out the gender of the babies. while still viewing trends in the names if interest.

```
plot2 <- ggplot(data = all_names) +
  geom_point(mapping = aes(x = year, y = n, color = name)) +
  facet_grid(name ~ year, scales = "free") +
   theme(axis.title.y=element_blank()) +
    theme(legend.position="none")

plot2
```

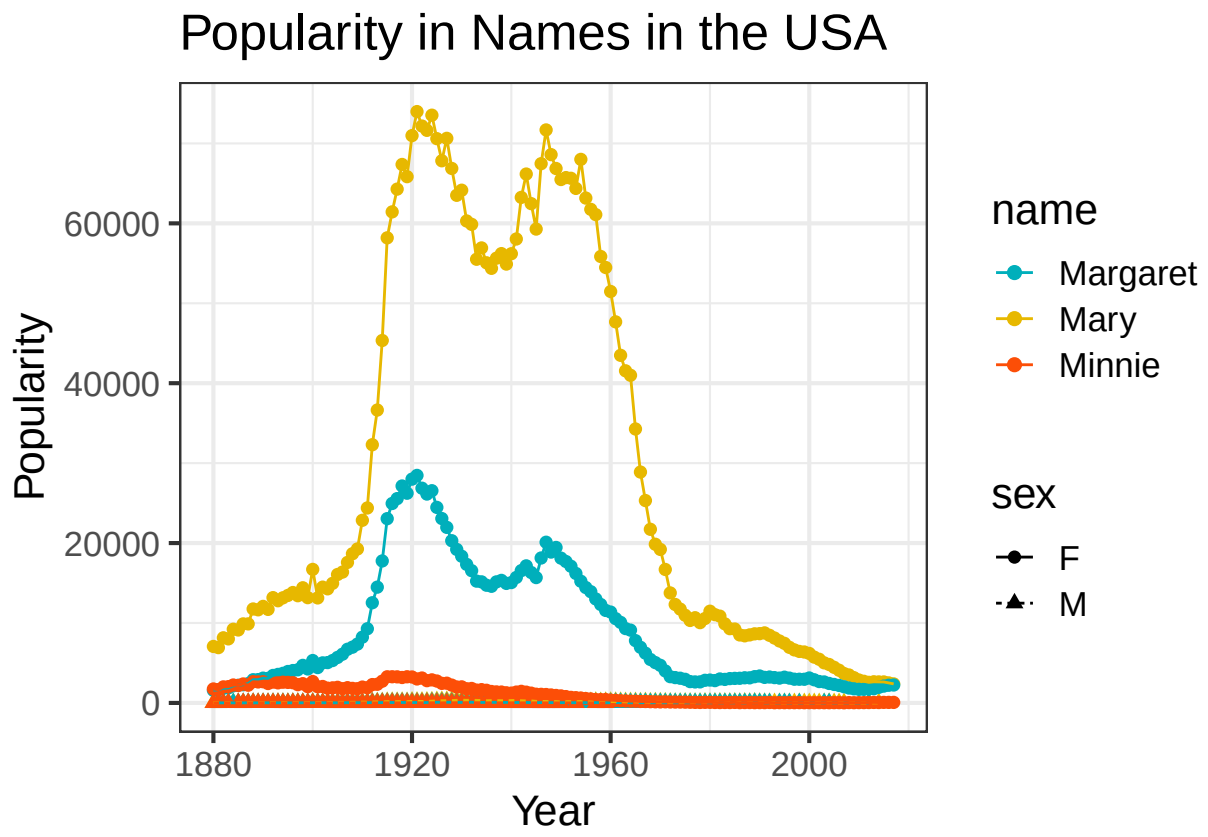


Plot 3 assigns shape to sex to distinguish between groupings, names the axes and includes a title, fonts are increased and a border implemented.

```
plot3 <- ggplot(data = all_names) +
  geom_point(mapping = aes(x = year, y = n, color = name, shape = sex, shape = sex), size = 2) +
  geom_line(mapping = aes(x = year, y = n, color = name,
                          linetype = sex)) +
 scale_colour_manual(values = c("#00AFBB", "#E7B800", "#FC4E07")) +
  theme_bw(base_size = 16) +
  labs(title = "Popularity in Names in the USA") +
  xlab ("Year") +
  ylab ("Popularity")
```

```
## Warning: Duplicated aesthetics after name standardisation: shape
```

```
plot3
```



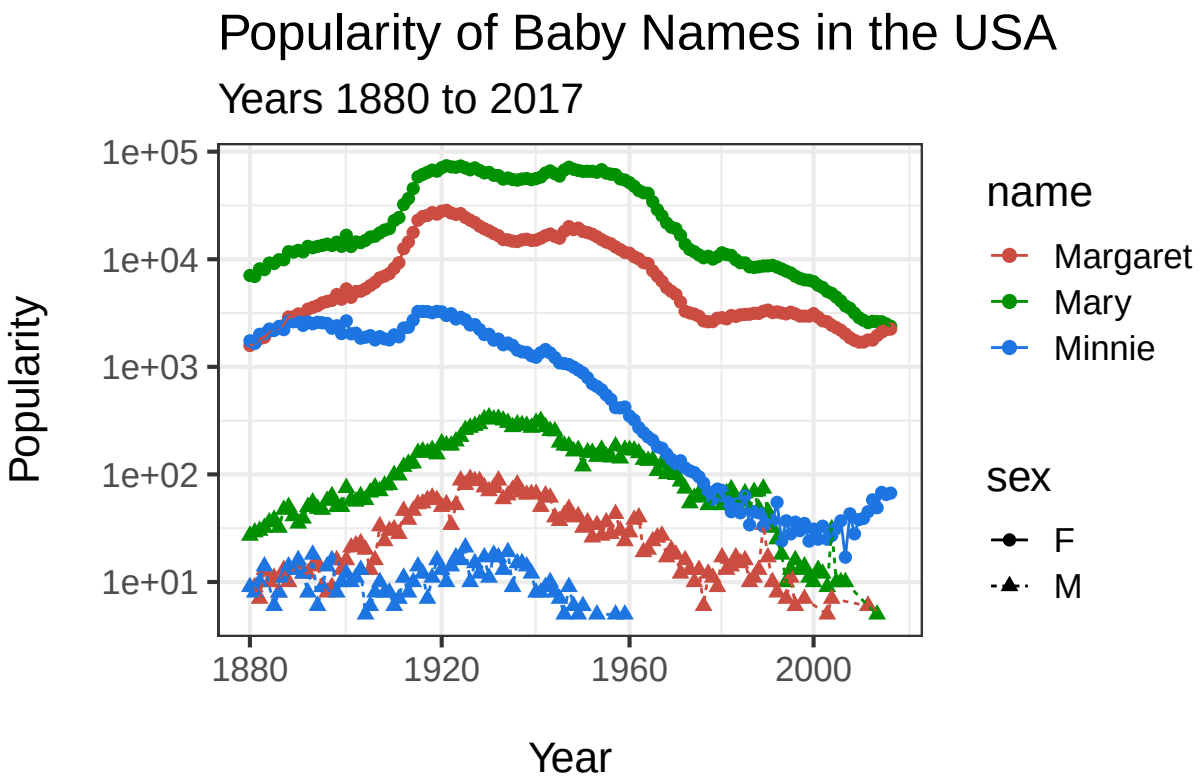
#plot 4 implmenting diffent colors ggplot2 package, and attempts to implement hue. It also applies a log transformation to improve these data trends visually.

```
plot4 <- ggplot(data = all_names) +
  geom_point(mapping = aes(x = year, y = n, color = name, shape = sex, shape = sex), size = 2) +
  geom_line(mapping = aes(x = year, y = n, color = name,
                          linetype = sex)) +
 scale_colour_manual(values = c ("#0072B2", "#009E73", "#CC79A7")) +
  scale_colour_hue(l=50) +
  theme_bw(base_size = 16) +
 theme(axis.title.y = element_text(margin = margin(t = 0, r = 20, b = 0, l = 0))) + theme(axis.title.x =
  scale_y_log10() +
  scale_x_log10() +
  labs(title = "Popularity of Baby Names in the USA", subtitle = "Years 1880 to 2017") +
  xlab ("Year") +
  ylab ("Popularity")
```

```
## Warning: Duplicated aesthetics after name standardisation: shape
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```

```
plot4
```

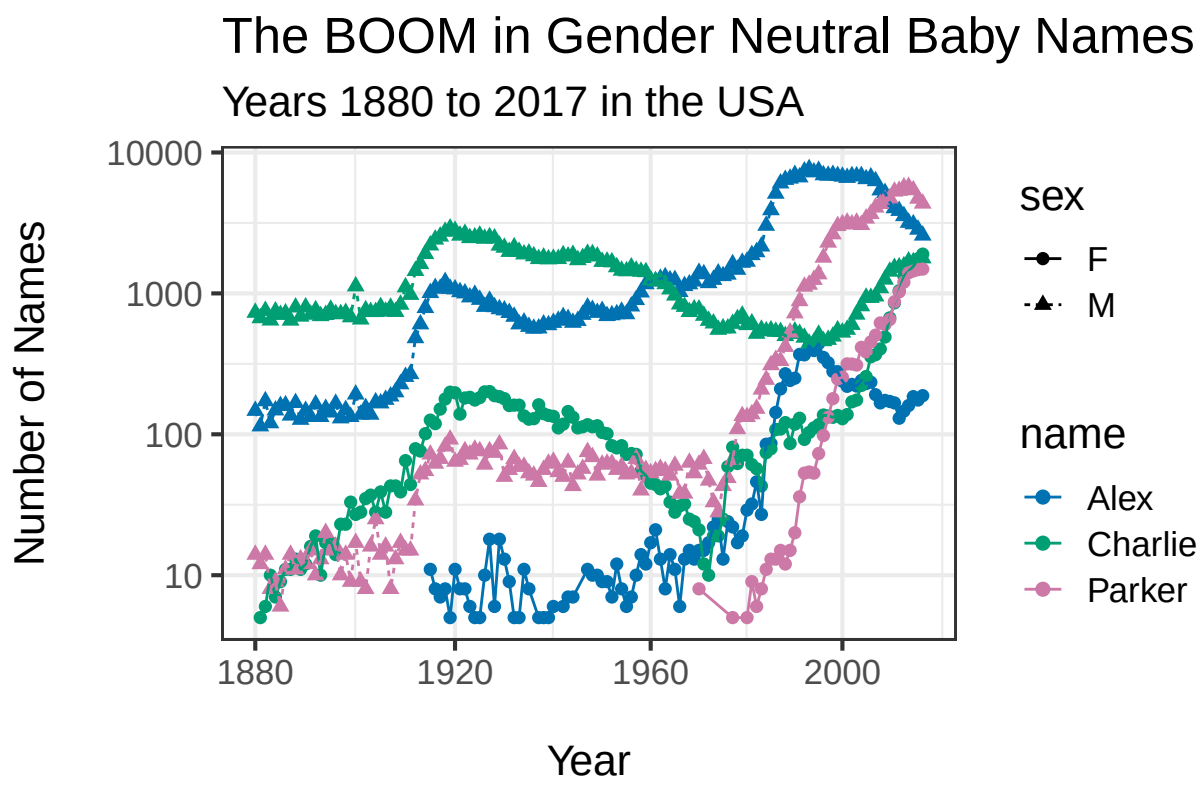


Plot 5 looks at more interesting names (Alex, Parker, Charlie), eliminates the attempts to implement luminescence, uses colorblind friendly palettes from ggplot2, adjusts for spaces in the axes, log transforms these data for better visual trends, groups by name and sex. The title is changed to be more descriptive of the data, and less of the actual axes.

```
plot5 <- ggplot(data = their_names) +
  geom_point(mapping = aes(x = year, y = n, color = name, shape = sex, shape = sex), size = 2) +
  geom_line(mapping = aes(x = year, y = n, color = name,
                          linetype = sex)) +
 scale_colour_manual(values = c ("#0072B2", "#009E73", "#CC79A7")) +
  theme_bw(base_size = 16) +
 theme(axis.title.y = element_text(margin = margin(t = 0, r = 20, b = 0, l = 0))) + theme(axis.title.x =
  scale_y_log10() +
  scale_x_log10() +
  labs(title = "The BOOM in Gender Neutral Baby Names", subtitle = "Years 1880 to 2017 in the USA") +
  xlab ("Year") +
  ylab ("Number of Names")
```

```
## Warning: Duplicated aesthetics after name standardisation: shape
```

```
plot5
```

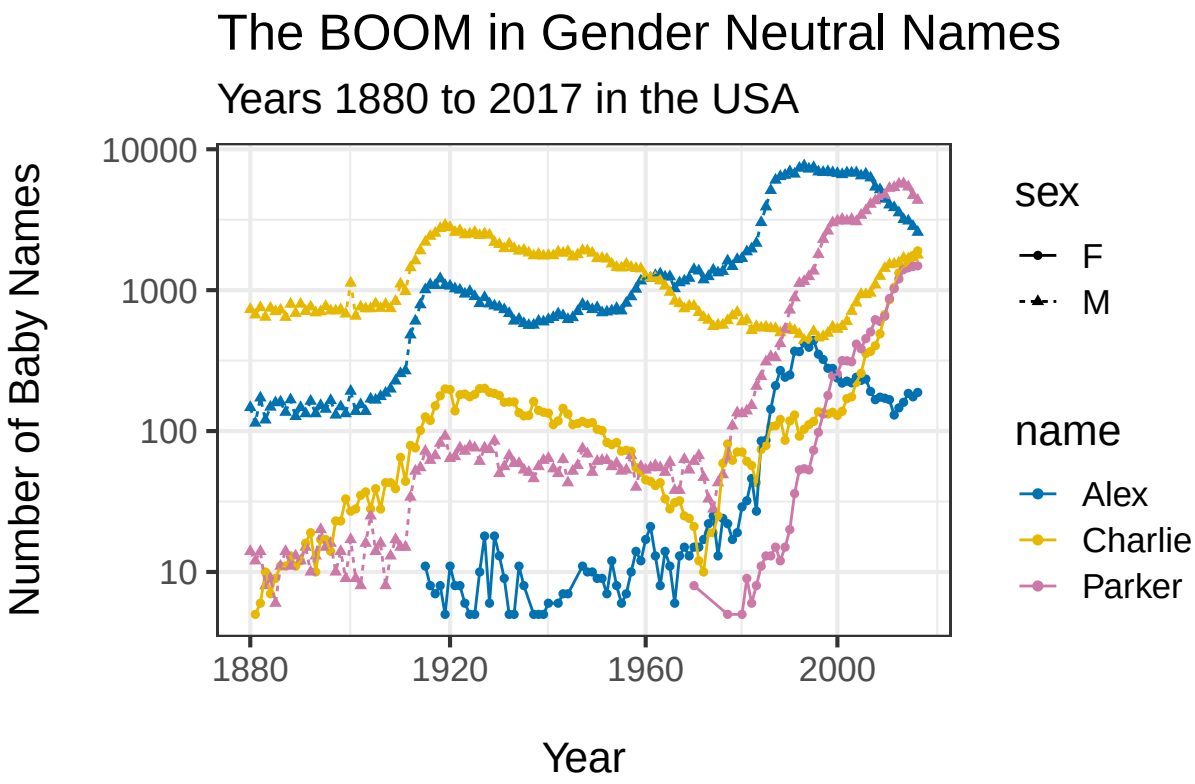


View color palettes offered with package colorspace.

```
hcl_palettes(plot = TRUE)
```

Plot 6 finalizes the color selection (which is evidently very challenging) for the final polished plot. Additionally, a couple of adjustments for titles, trying to enhance comprehension but eliminate "wordiness". After playing around with various colors from colorspace, I decided manual selection from a list of ggplot2 color blind palattes was best. To test the plot success, I viewed it using http://hclwizard.org:3000/cvdemulator/. It passed the test, so this is my final plot. I did not feel the need to adjust for black and white printing because the shapes distinguish the difference between sexes, and those trends are made evident without the need to view different colors.

```
plot6 <- ggplot(data = their_names) +
  geom_point(mapping = aes(x = year, y = n, color = name, shape = sex), size = 1.3) +
  geom_line(mapping = aes(x = year, y = n, color = name, linetype = sex)) +
    scale_colour_manual(values = c ("#0072B2", "#E7B800", "#CC79A7"))+
    theme_bw(base_size = 16) +
    theme(axis.title.y = element_text(margin = margin(t = 0, r = 20, b = 0, l = 0))) + theme(axis.title
  scale_y_log10() +
  scale_x_log10() +
  labs(title = "The BOOM in Gender Neutral Names", subtitle = "Years 1880 to 2017 in the USA") +
  xlab ("Year") +
  ylab ("Number of Baby Names")

plot6
```


The BOOM in Gender Neutral Names
Years 1880 to 2017 in the USA
Number of Baby Names
10000
1000
100
10
1880
1920
1960
2000
Year
sex
F
M
name
Alex
Charlie
Parker