Layering Graphics

CMSC 205, Winter 2017

April 3, 2017

Motivating example

Today, we hope to recreate the below graphic that appeared in the FiveThirtyEight article "How to Tell Someone's Age When All You Know Is Her Name".

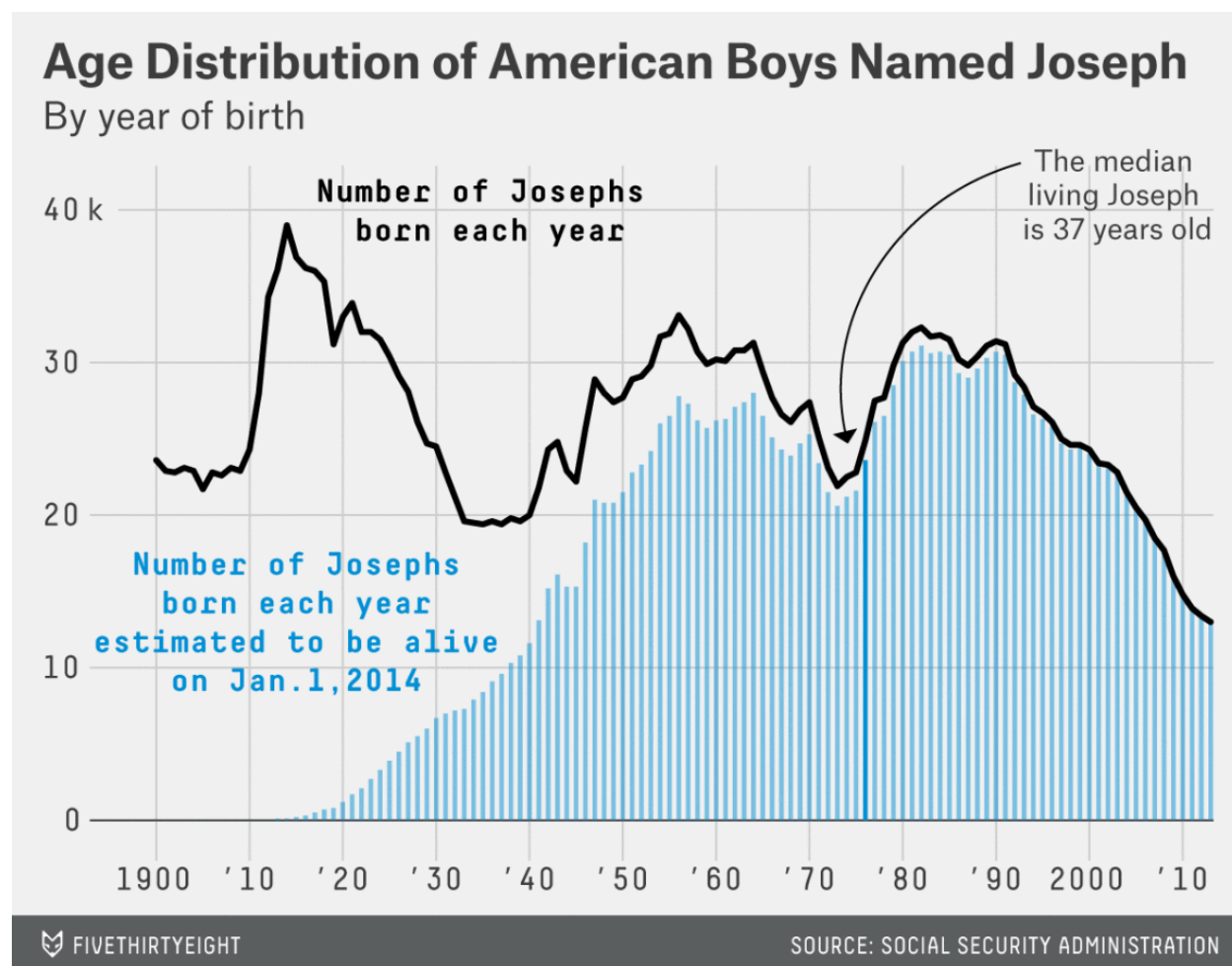The data we will use today are found on GitHub.

Figure 1:

Step 1: Load the tidyverse and the data

```
library(tidyverse)
library(mdsr)
```

```
babynames <- make_babynames_dist()
head(babynames)
```

```
## # A tibble: 6 × 9
##    year  sex      name     n       prop alive_prob count_thousands
##   <dbl> <chr>    <chr> <int>      <dbl>      <dbl>           <dbl>
## 1  1900    F      Mary 16707 0.05257295          0          16.707
## 2  1900    F     Helen  6343 0.01995991          0           6.343
## 3  1900    F      Anna  6114 0.01923930          0           6.114
## 4  1900    F  Margaret  5306 0.01669672          0           5.306
## 5  1900    F      Ruth  4765 0.01499432          0           4.765
## 6  1900    F Elizabeth  4096 0.01288914          0           4.096
## # ... with 2 more variables: age_today <dbl>, est_alive_today <dbl>
```
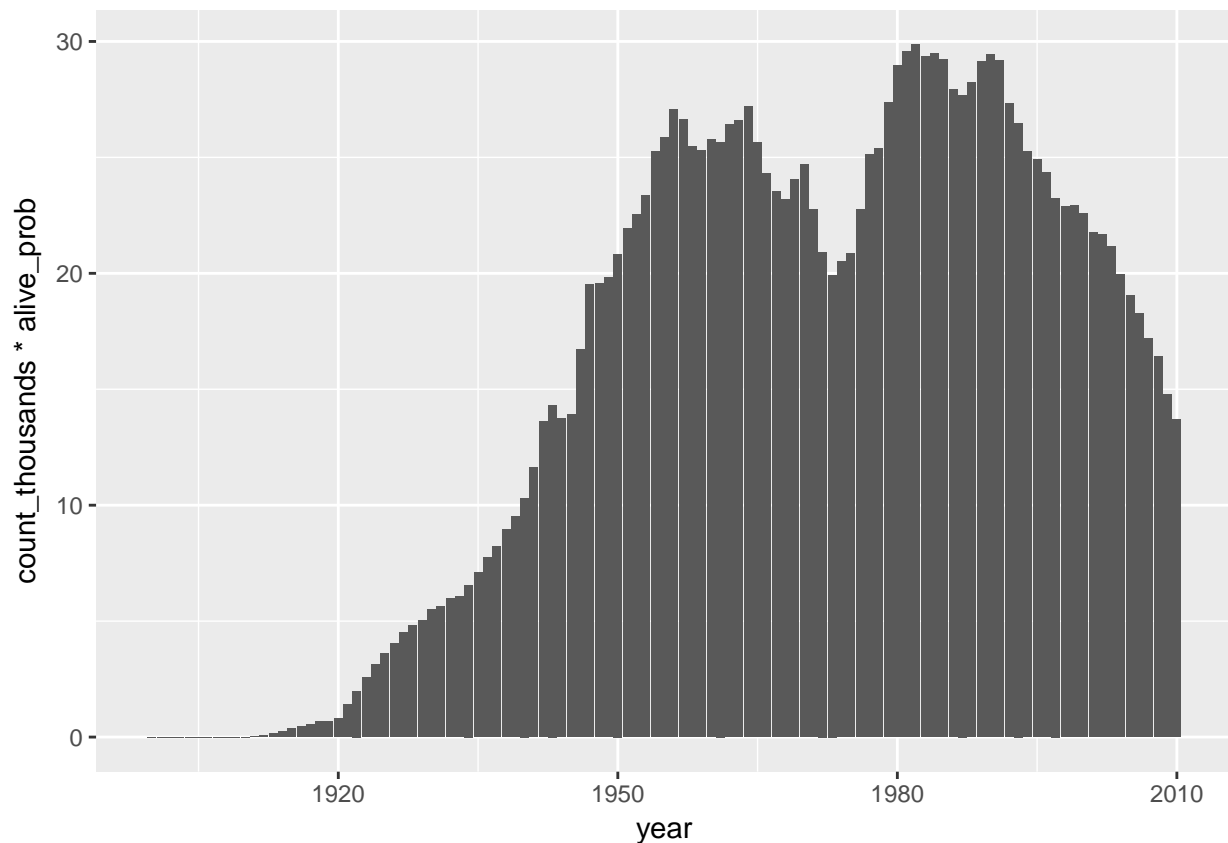
Step 2: Filter out the Josephs

```
joseph <- filter(babynames, name == "Joseph" & sex == "M")
```
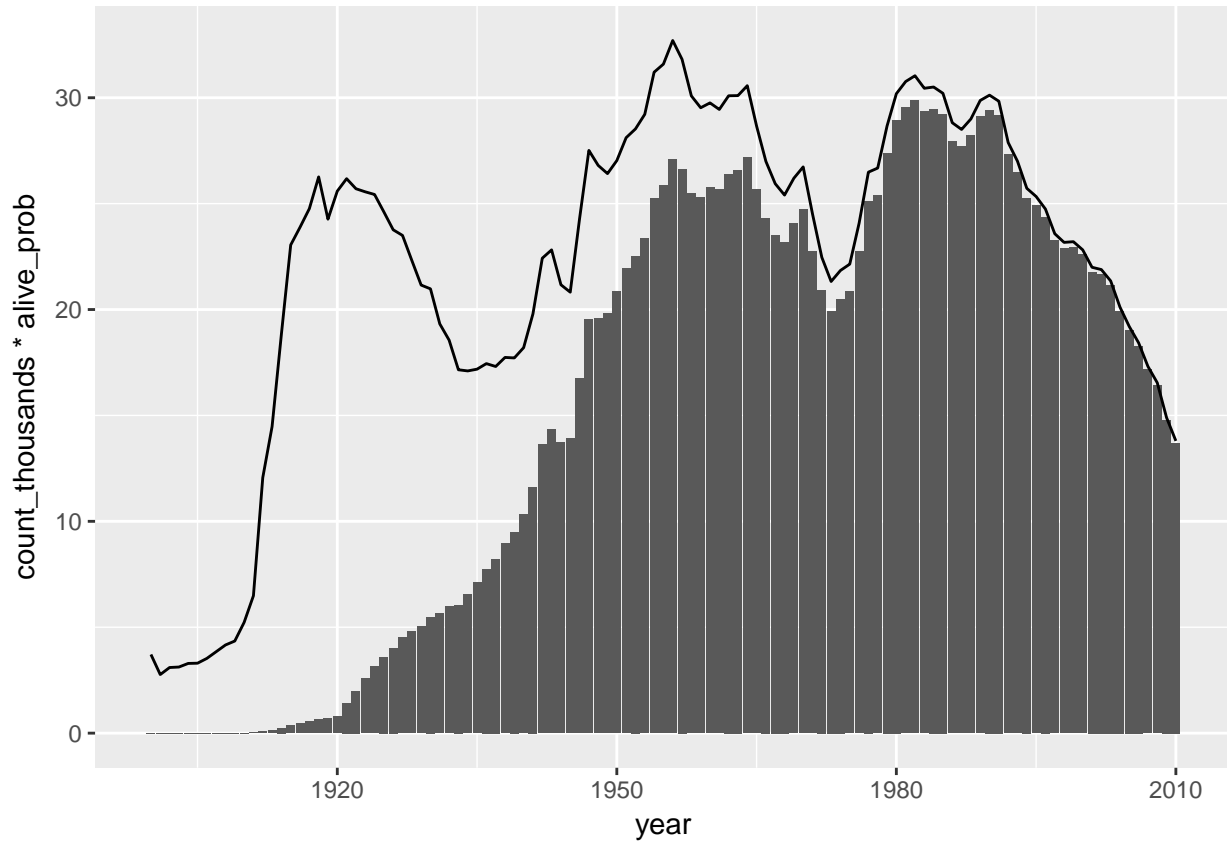
Step 3: Draw a bar chart on top for Josephs who are alive

```
name_plot <-
  ggplot(joseph, aes(x = year)) +
  geom_bar(stat = "identity", aes(y = count_thousands * alive_prob))
name_plot
```
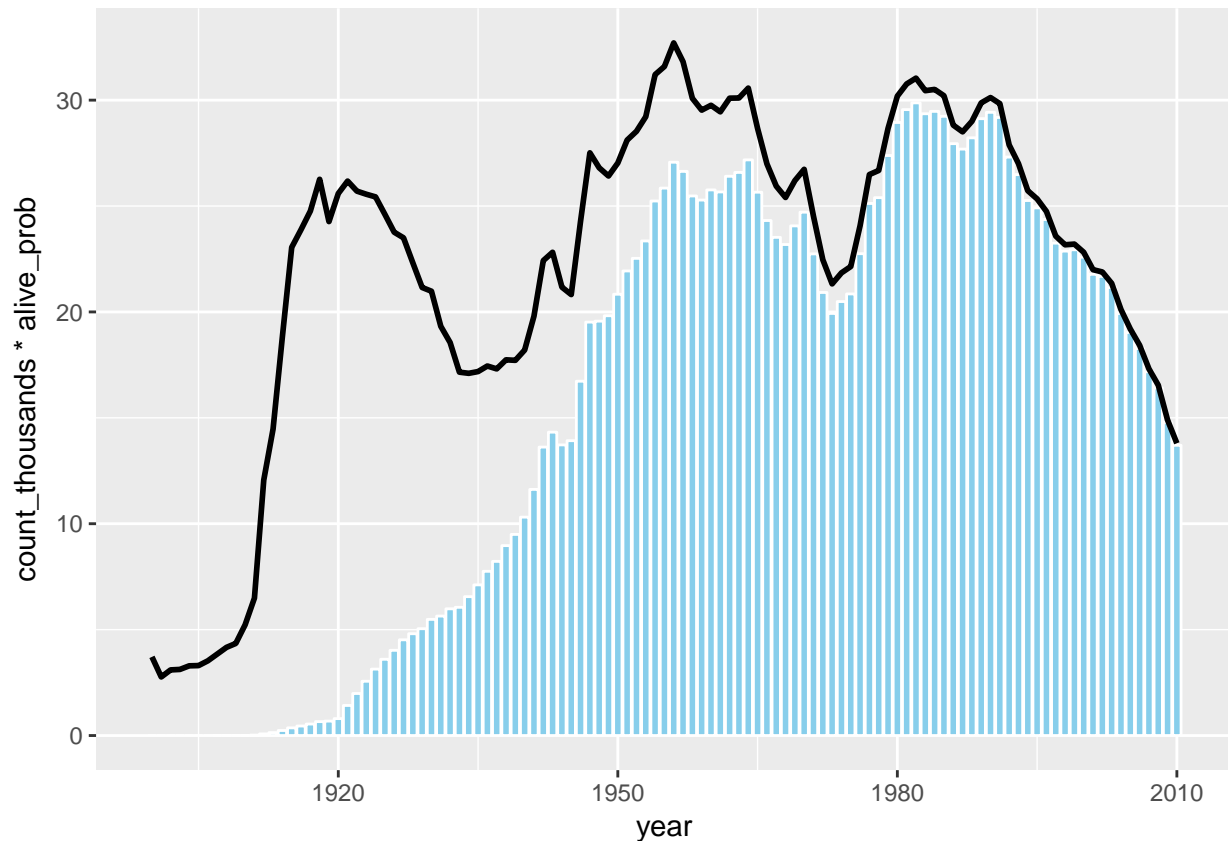
Step 4: Draw the line plot for all Josephs

```
name_plot <-
  name_plot +
  geom_line(aes(y = count_thousands))
name_plot
```


Step 5: Tweaking the colors and line width

```
name_plot <-
  ggplot(joseph, aes(x = year)) +
  geom_bar(stat = "identity", aes(x = year, y = count_thousands * alive_prob), fill = "skyblue", color =
  geom_line(aes(y = count_thousands), size = 1)
name_plot
```

Step 6: Highlighting the median age

```r
library(Hmisc) # to access wtd.quantile() function
```

```
## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
median_year <- wtd.quantile(x = joseph$year, weights = joseph$est_alive_today, probs = 0.5)
median_year
```
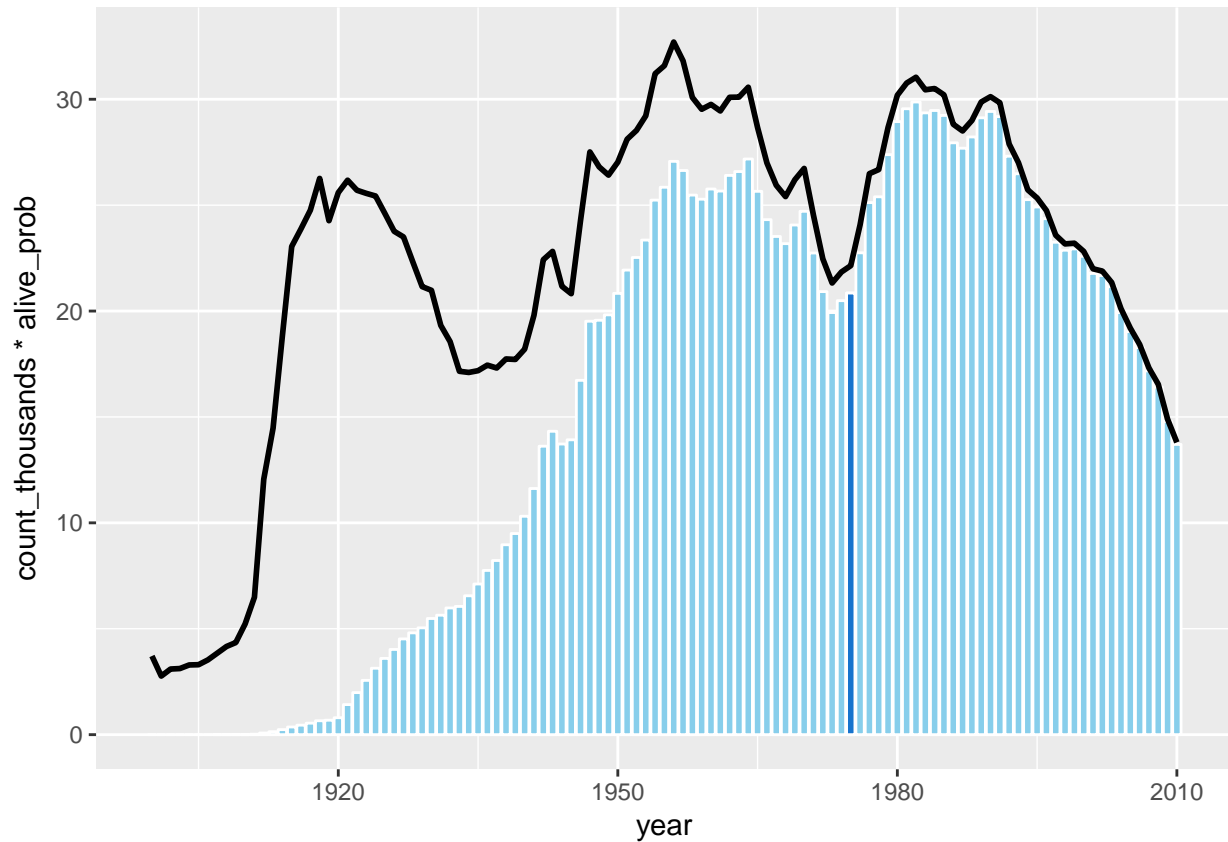
```
##  50%
## 1975
```

```r
median_year_alive <- filter(joseph, year == median_year)

name_plot <-
  name_plot +
```

```
    geom_bar(stat = "identity", color = "white", fill = "dodgerblue3", aes(y = ifelse(year == median_yea
name_plot
```
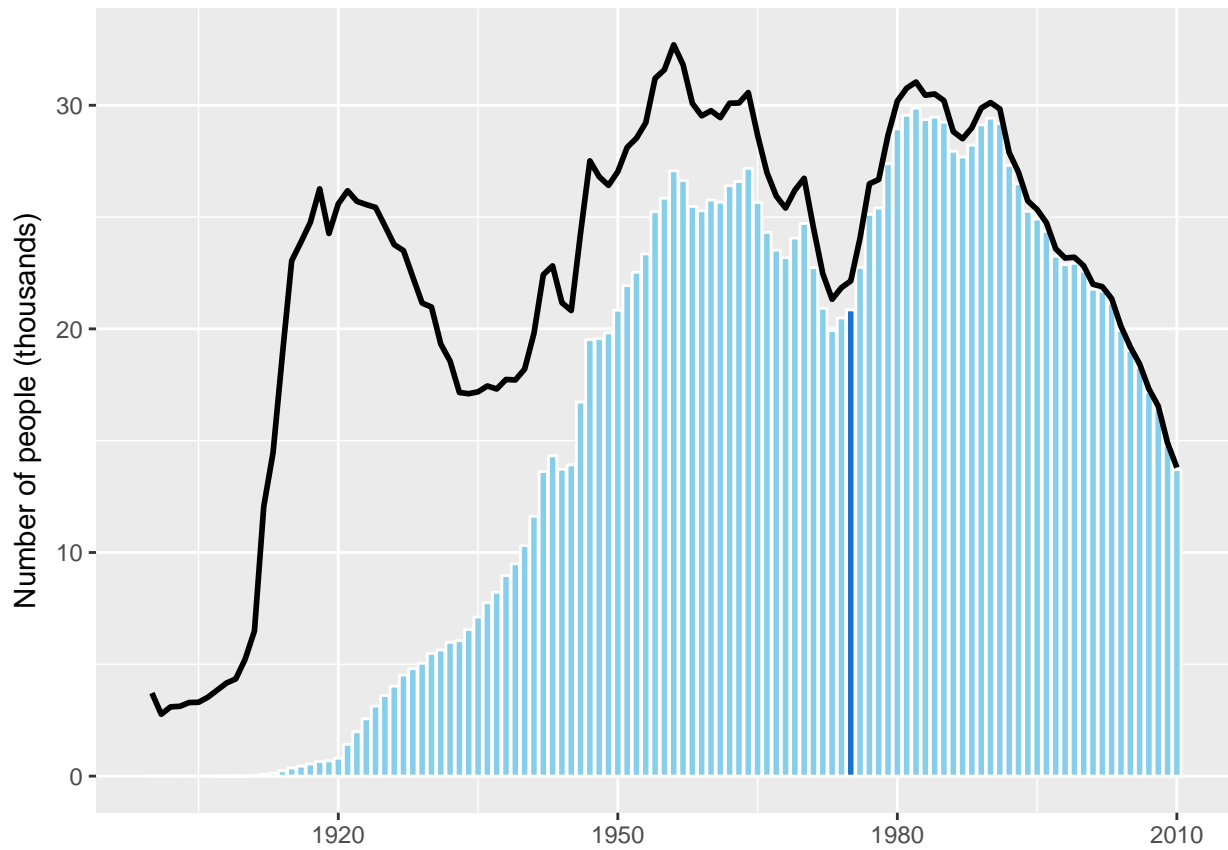
Warning: Removed 110 rows containing missing values (position_stack).

Step 7: Fixing the axis labels

```
name_plot <-
  name_plot +
  labs(x = NULL, y = "Number of people (thousands)")
name_plot
```
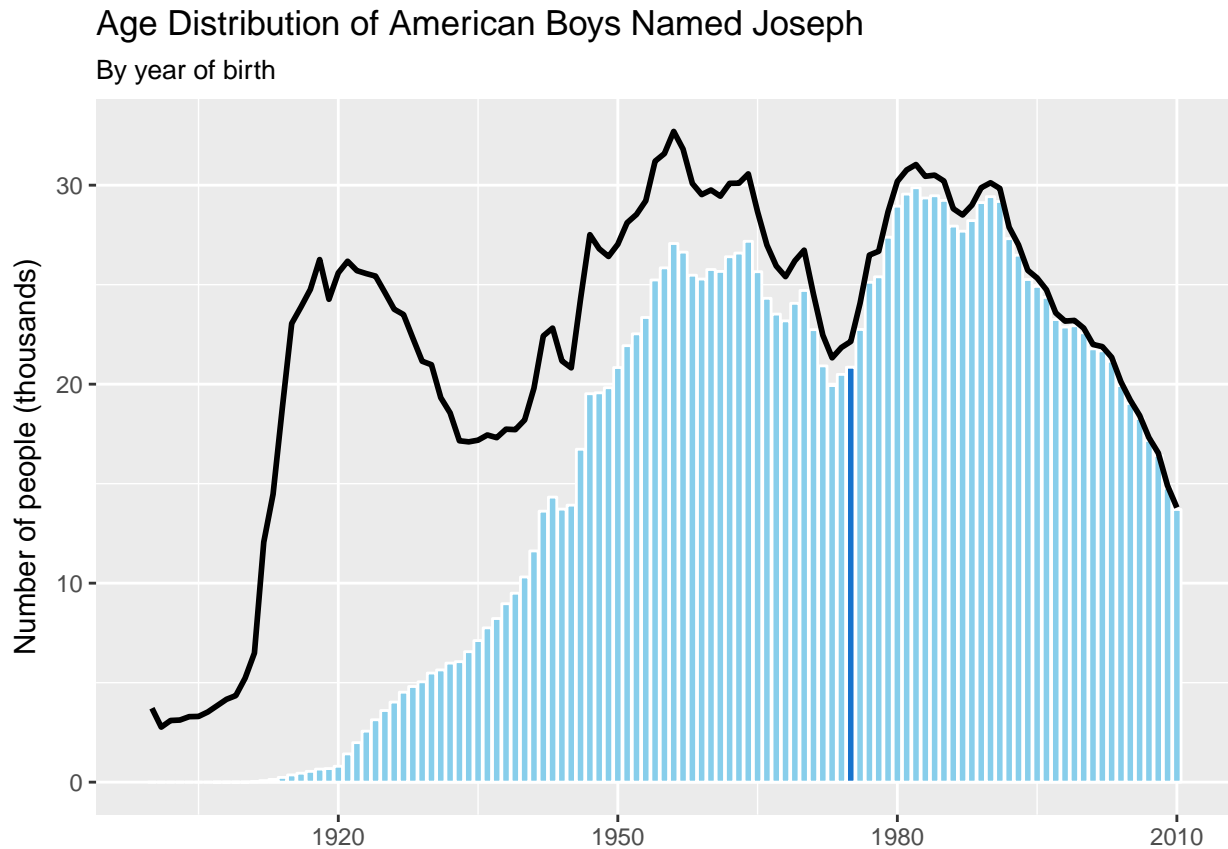
Warning: Removed 110 rows containing missing values (position_stack).

Step 8: Adding a title

```
name_plot <-
  name_plot +
  labs(title = "Age Distribution of American Boys Named Joseph",
       subtitle = "By year of birth")
name_plot
```
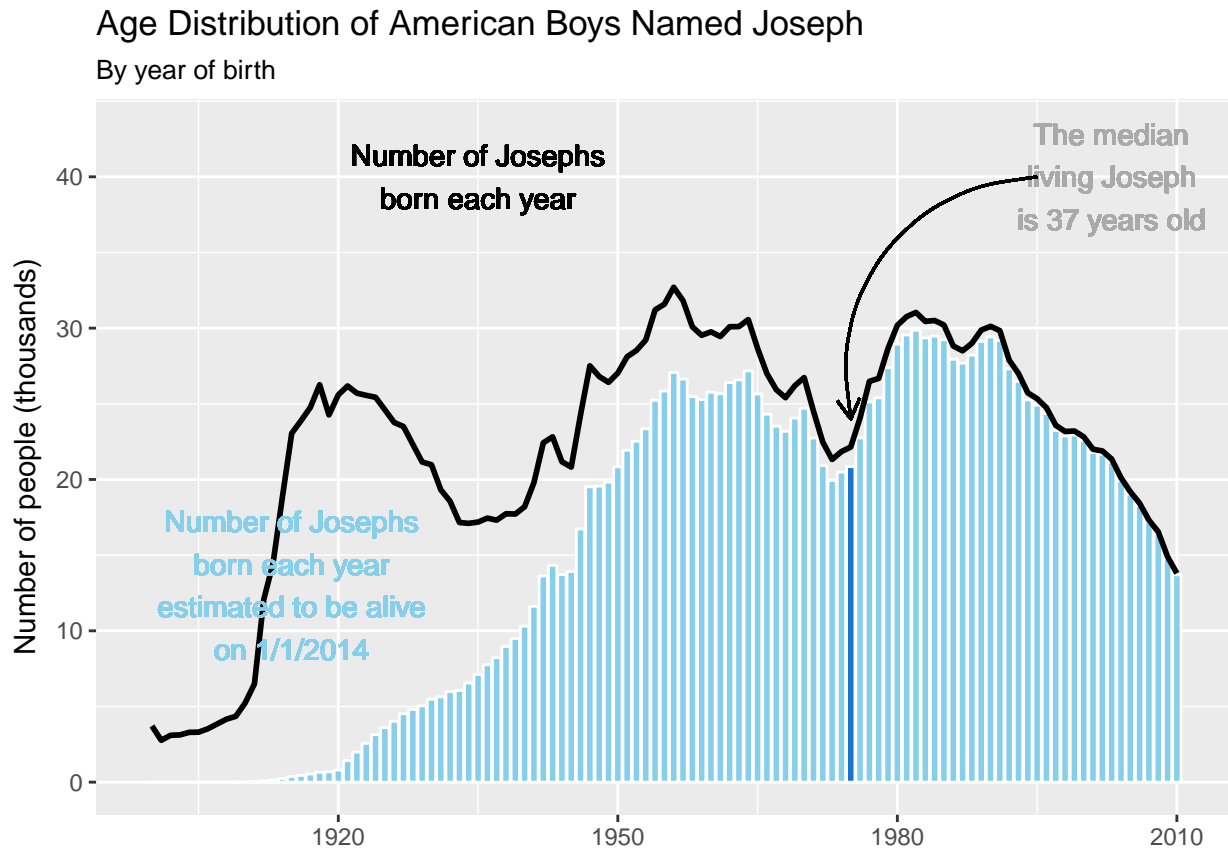
```
## Warning: Removed 110 rows containing missing values (position_stack).
```

Age Distribution of American Boys Named Joseph
By year of birth

Step 9: Adding text annotations

```
name_plot <-
  name_plot +
  ylim(0, 43) +
  geom_text(x = 1935, y = 40, label = "Number of Josephs\nborn each year") +
  geom_text(x = 1915, y = 13, label = "Number of Josephs\nborn each year\nestimated to be alive\non 1/1,
  geom_text(x = 2003, y = 40, label = "The median\nliving Joseph\nis 37 years old", color = "darkgray")
  geom_curve(x = 1995, xend = 1975, y = 40, yend = 24,
             arrow = arrow(length = unit(0.3, "cm")),
             curvature = 0.5)
name_plot
```

```
## Warning: Removed 110 rows containing missing values (position_stack).
```

Age Distribution of American Boys Named Joseph
By year of birth

Step 10: Changing the theme

When creating publication-quality graphics, we often want to change the **theme** of the plot. That is, we might not want a gray background, or we may which to change the tick marks on the axis, or some other characteristic. This type of change is a change to the theme. Today, we'll use the `ggthemes` package to quickly make a change to a different theme.

```r
library(ggthemes) # contains numerous different themes
```

You can find a gallery of the available themes here

```r
name_plot + theme_minimal()
```

```
## Warning: Removed 110 rows containing missing values (position_stack).
```

Age Distribution of American Boys Named Joseph

By year of birth