What Happened to Jane? — Answers

Data Computing

CVC 2015

Warm-ups

1. How many babies are represented?

Add up the `n` (count) over the names and years.

```
babynames %>%
  summarise(total = sum(n))
```

```
## Source: local data frame [1 x 1]
##
##       total
## 1 333417770
```

Note that `summarise()` clobbers all the variables in the input data table other than those used for grouping. (No variables were used for grouping here.)

2. How many babies are there in each year?

```
babynames %>%
  group_by(year) %>%
  summarise(total = sum(n))
```

```
## Source: local data frame [134 x 2]
##
##    year  total
## 1  1880 201484
## 2  1881 192700
## 3  1882 221537
## 4  1883 216952
## 5  1884 243468
## 6  1885 240856
## 7  1886 255320
## 8  1887 247396
## 9  1888 299481
## 10 1889 288952
## ..  ...    ...
```

With `year` made a grouping variable, a separate calculation is done for each year, and `year` appears in the output.

3. How many distinct names in each year?

```
babynames %>%
  group_by(year) %>%
  summarise(name_count = n_distinct(name))
```

```
## Source: local data frame [134 x 2]
##
##    year name_count
## 1  1880       1889
## 2  1881       1830
## 3  1882       2012
## 4  1883       1962
## 5  1884       2158
## 6  1885       2139
## 7  1886       2225
## 8  1887       2215
## 9  1888       2454
## 10 1889       2390
## ..  ...        ...
```

4. How many distinct names of each sex in each year?

```
babynames %>%
  group_by(year, sex) %>%
  summarise(name_count = n_distinct(name))
```

```
## Source: local data frame [268 x 3]
## Groups: year
##
##    year sex name_count
## 1  1880   F        942
## 2  1880   M       1058
## 3  1881   F        938
## 4  1881   M        997
## 5  1882   F       1028
## 6  1882   M       1099
## 7  1883   F       1054
## 8  1883   M       1030
## 9  1884   F       1172
## 10 1884   M       1125
## ..  ... ...        ...
```

Popularity of Jane and Mary

1. Track the yearly number of Janes and Marys over the years.

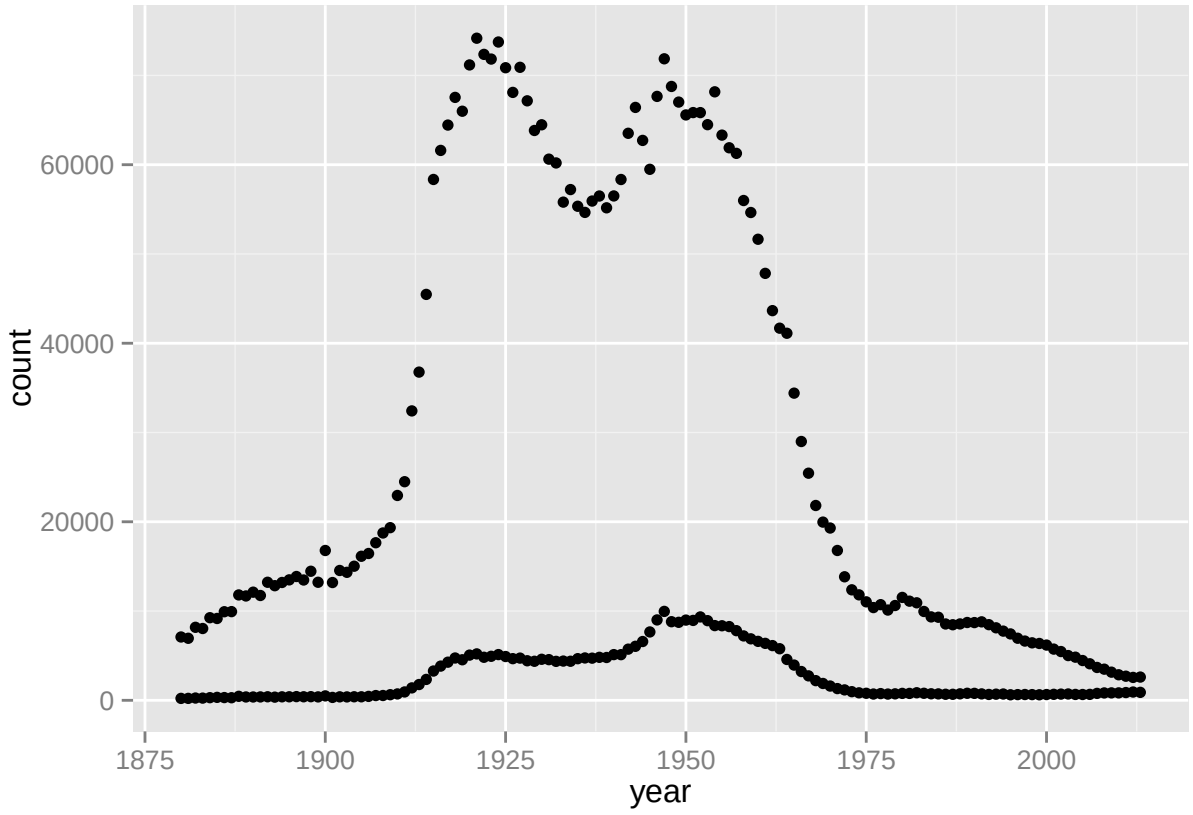
```
Result <-
  babynames %>%
  filter(name %in% c("Jane", "Mary")) %>%
```

```
  group_by(name, year) %>% # for each year
  summarise(count = sum(n))
```

2. Plot out the result of (1)

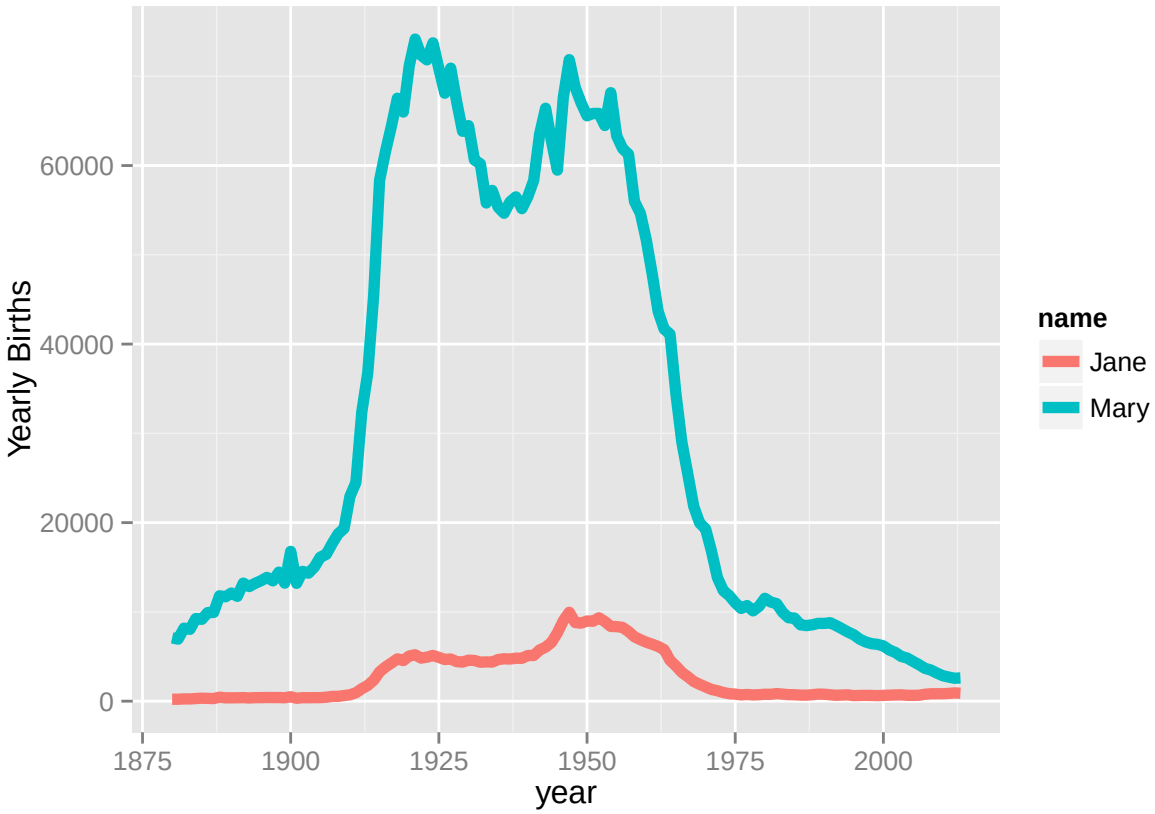
Put `year` on the x-axis and the count of each name on the y-axis.

```
ggplot(data=Result, aes(x = year, y = count)) +
  geom_point()
```



- *Map* the name (Mary or Jane) to the aesthetic of color. Remember that mapping to aesthetics is always done inside the `aes()` function.
- Instead of using dots as the glyph, use a line that connects consecutive values: `geom_line()`.
- Change the y-axis label to "Yearly Births": `+ ylab("Yearly Births")`
- *Set* the line thickness to `size=2`. Remember that "setting" refers to adjusting the value of an aesthetic to a constant. Thus, it's *outside* the `aes()` function.

```
ggplot(data=Result, aes(x = year, y = count)) +
  geom_line(aes(color = name), size=2) +
  ylab("Yearly Births")
```



3. Look at the *proportion* of births rather than the count

```
Result2 <-
  babynames %>%
  group_by(year) %>%
  mutate(total = sum(n)) %>%
  filter(name %in% c("Mary", "Jane")) %>%
  mutate(proportion = n / total)
```

- Why is `sex` a variable in `Result2`? Eliminate it, keeping just the girls.

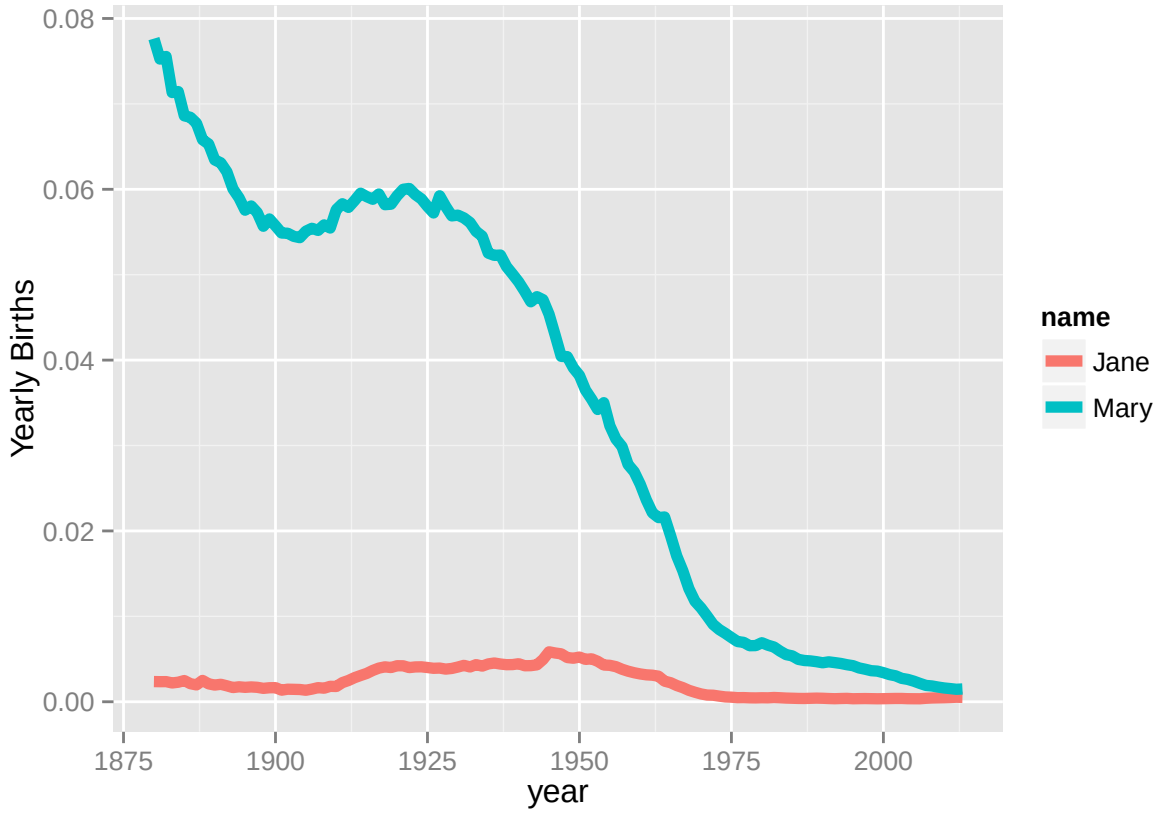
```
Result2 <-
  babynames %>%
  filter(sex == "F") %>%
  group_by(year) %>%
  mutate(total = sum(n)) %>%
  filter(name %in% c("Mary", "Jane")) %>%
  mutate(proportion = n / total)
```

- What happens if the `filter()` step is put *before* the `mutate()` step?

The `total` is just for Mary and Jane, ignoring all the other babies.

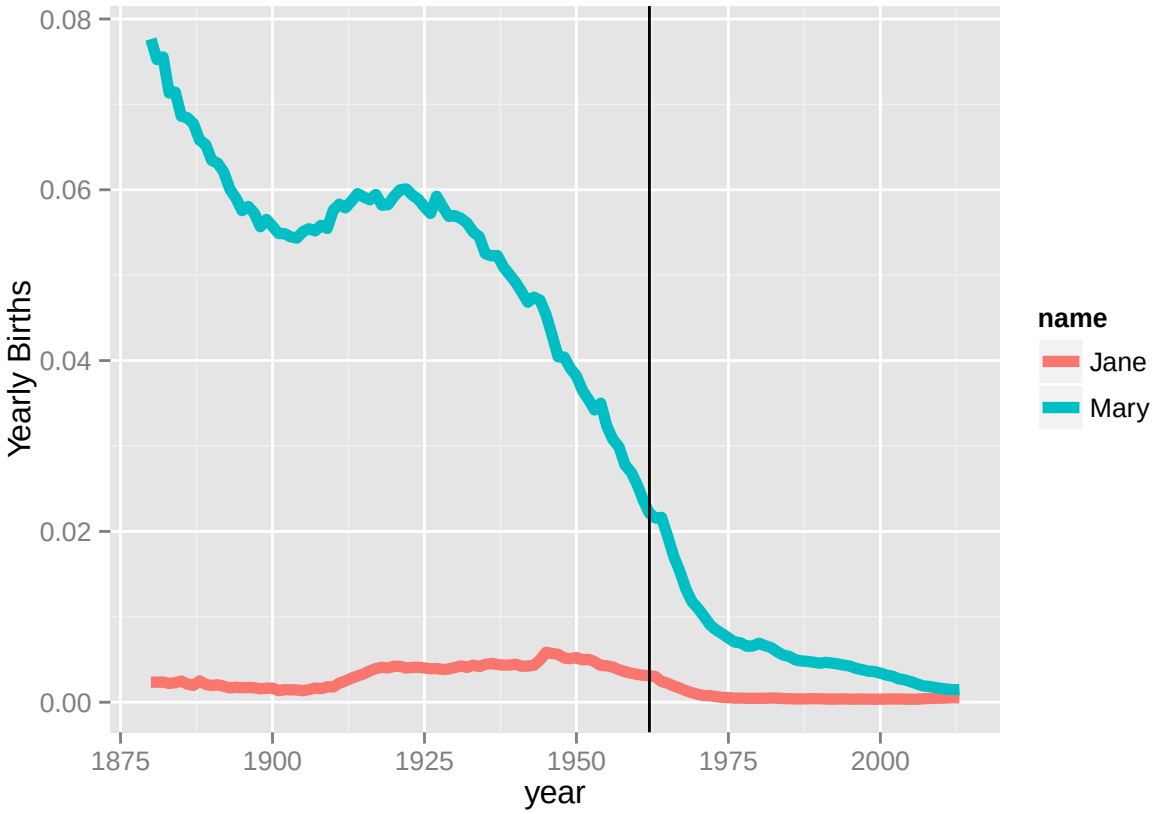
- Graph the results

```
ggplot(data=Result2, aes(x = year, y = proportion)) +
  geom_line(aes(color = name), size=2) +
  ylab("Yearly Births")
```



- Add a vertical line to mark a year in which something happened that might relate to the increase or decrease the popularity of the name. Example: The movie *Whatever Happened to Baby Jane* came out in 1962. The glyph is a vertical line: `geom_vline()`.

```
ggplot(data=Result2, aes(x = year, y = proportion)) +
  geom_line(aes(color = name), size=2) +
  ylab("Yearly Births") +
  geom_vline(x=1962)
```


0.08
0.06
0.04
0.02
0.00
Yearly Births
1875
1900
1925
1950
1975
2000
year
name
Jane
Mary