Problem Set 04

WRITE YOUR NAME HERE

2021-09-18

Learning goals

- Using more data visualization techniques: changing colors and adding trend lines
- First data wrangling exercise

Setup

Load necessary packages:

```
library(ggplot2)
library(dplyr)
library(babynames)
```

Question 1: Honor code

For this problem set I worked with (please indicate even if with no one):

Question 2

In this exercise, you're going to recreate the figure from Practice Midterm I Question 4 (see `#midterms` channel in Slack), allowing us to visualize the degree to which the names "Casey" and "Riley" were used for babies of both sex male and female.

Part a)

Perform the data wrangling necessary to transform the `babynames` data frame included in the `babynames` package into a new data frame called `babynames_riley_casey` that will allow us to create the visualization.

Hint: I recommend you first draw on a piece of paper what the data frame should look like; that way you'll know what your target looks like and when you've hit it.

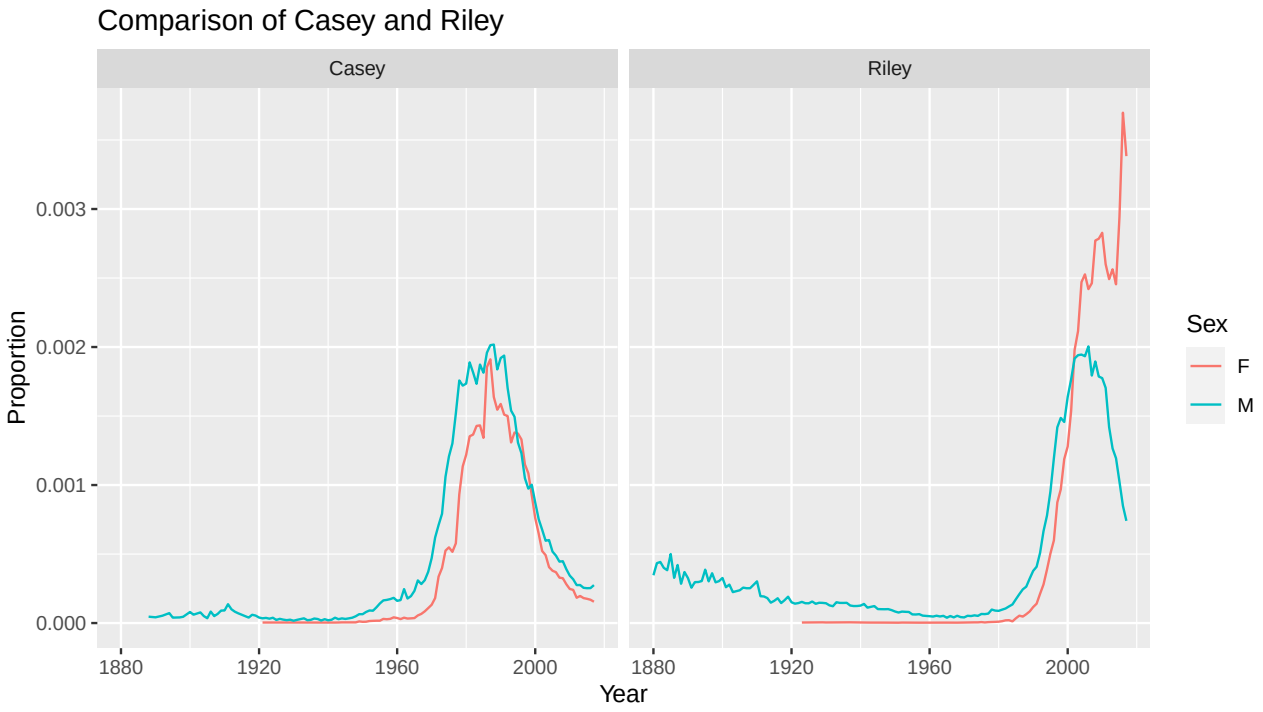
```
babynames_riley_casey <- babynames %>%
  filter(name == "Riley" | name == "Casey")
```

Part b)

Recreate the visualization from Practice Midterm I Question 4 *exactly* including the capitalization of all label text.

```
ggplot(babynames_riley_casey, aes(x=year, y=prop, col=sex)) +
  geom_line() +
```

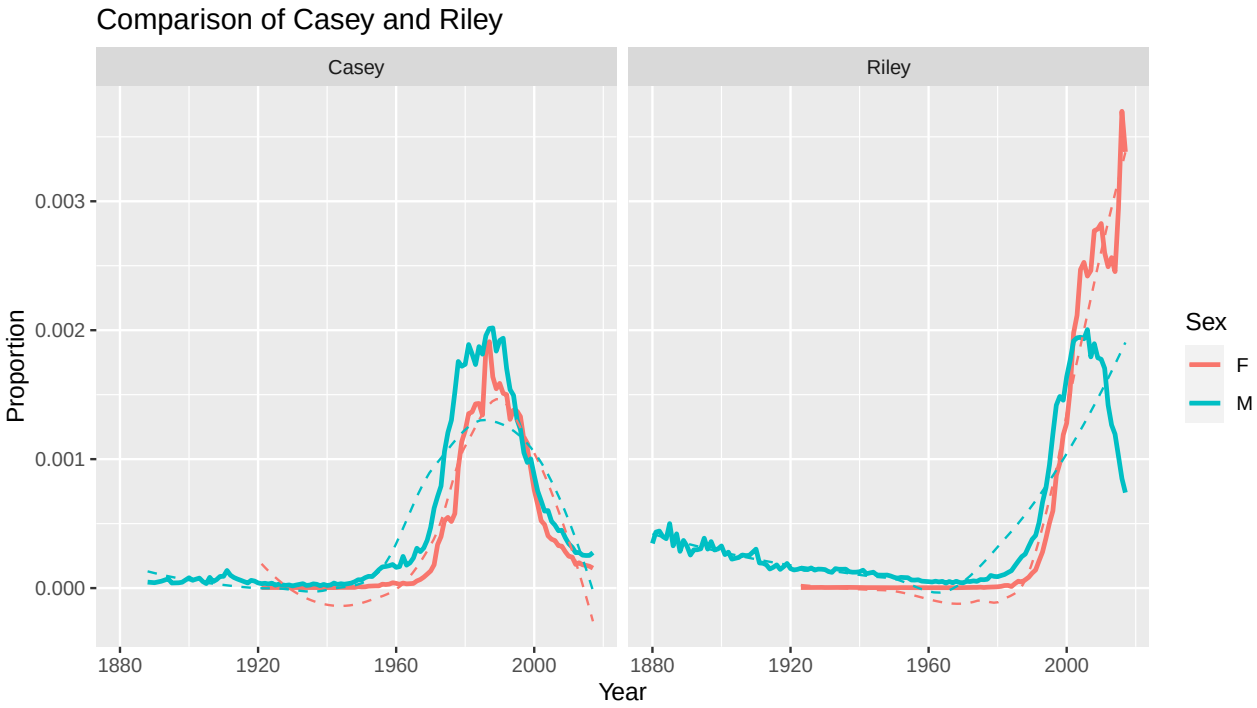
```
  facet_wrap(~name) +
  labs(x = "Year", y = "Proportion", color = "Sex", title = "Comparison of Casey and Riley")
```



Part c)

Once again, recreate the visualization from Practice Midterm I Question 4 *exactly*, however this time add appropriately chosen trend lines. For clarity's sake, do NOT include the standard error bars.

```
ggplot(babynames_riley_casey, aes(x=year, y=prop, col=sex)) +
  geom_line(size = 1) +
  facet_wrap(~name) +
  labs(x = "Year", y = "Proportion", color = "Sex", title = "Comparison of Casey and Riley") +
  geom_smooth(se = FALSE, size = 0.5, linetype = "dashed")
```



Bonus

In this exercise, you're going to recreate the figure from Practice Midterm I Question 4 (see `#midterms` channel in Slack). This time however, you're going to limit it to years 1960 and later:

Part a)

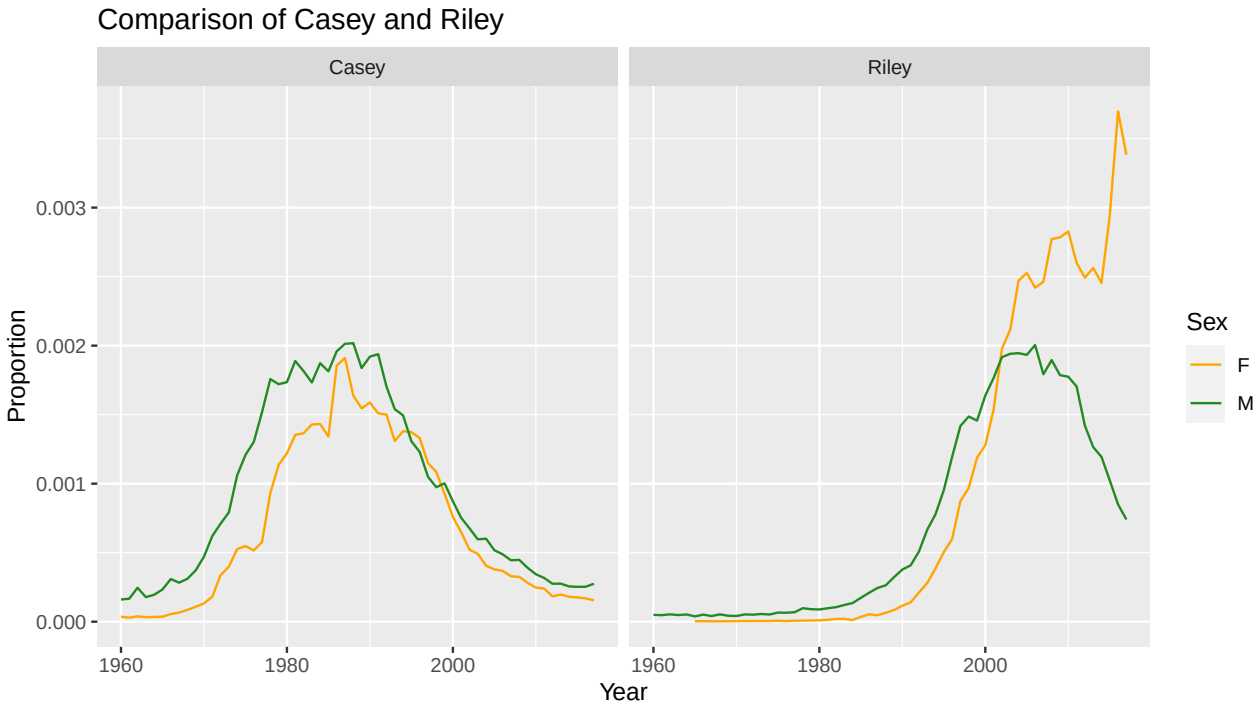
Perform the data wrangling necessary to transform the `babynames` data frame included in the `babynames` package into a new data frame `babynames_riley_casey_1960_later` that only has data for 1960 or later.

```
babynames_riley_casey_1960_later <- babynames %>%
  filter(name == "Riley" | name == "Casey") %>%
  filter(year >= 1960)
```

Part b)

Create the same version of the visualization as in Q2.b), but for the `babynames_riley_casey_1960_later` data frame and with `"forestgreen"` and `"orange"` lines for male and female respectively. This time, the x-axis should only be for years 1960 and later, as saved in the `babynames_riley_casey_1960_later` data frame.

```
ggplot(babynames_riley_casey_1960_later, aes(x=year, y=prop, col=sex)) +
  geom_line() +
  facet_wrap(~name) +
  labs(x = "Year", y = "Proportion", color = "Sex", title = "Comparison of Casey and Riley") +
  scale_color_manual(values = c("orange", "forestgreen"))
```


Comparison of Casey and Riley
Casey
Riley
Proportion
0.003
0.002
0.001
0.000
1960
1980
2000
Year
Sex
F
M