Assignment 1 Data Wrangling

Yaniv Bronshtein

1/31/2021

Import the libraries

```
library(babynames)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

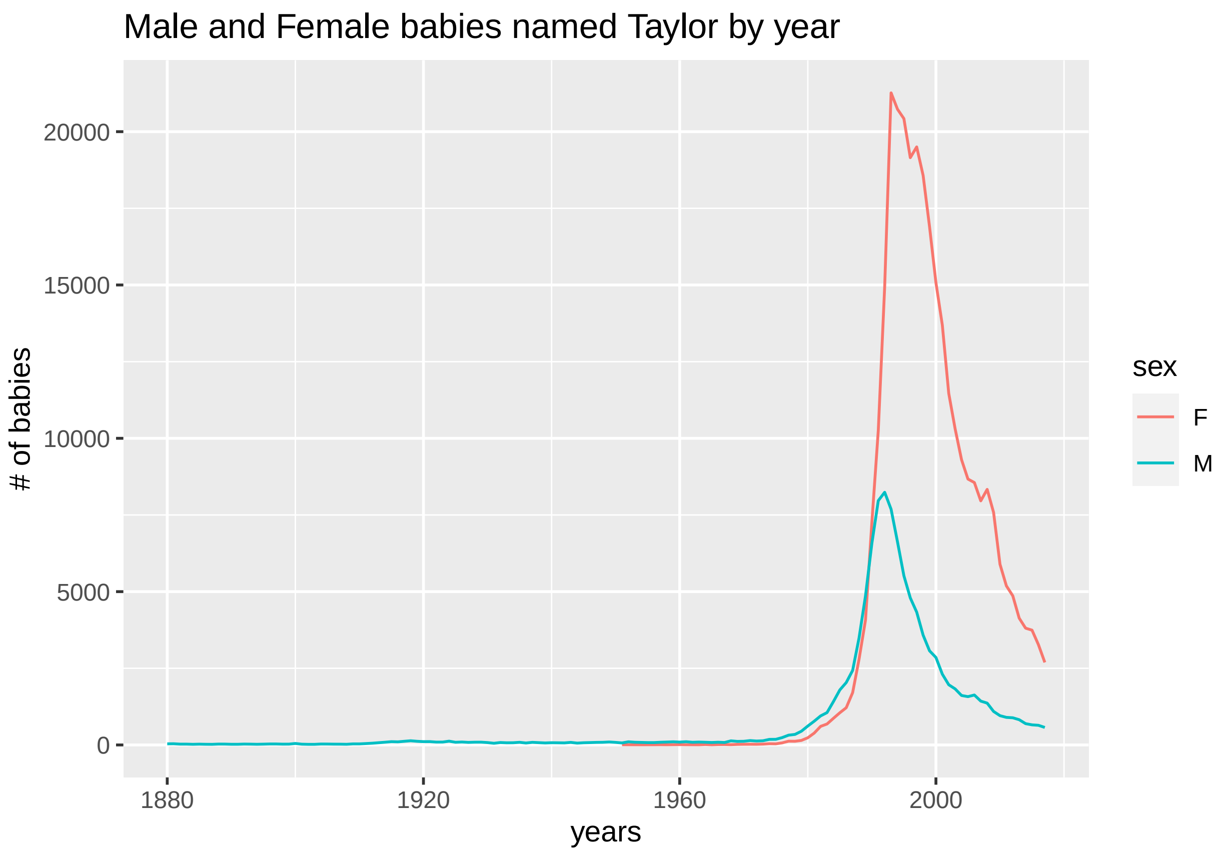
```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

Plot the number of male and female babies named Taylor *by year*

Filter the babynames dataset for the name Taylor

Create a ggplot using the newly filtered data where the x-axis is the year, y-axis is the count of the number of babies, and the color of each line is defined by the sex variable



Answer the following questions, showing plots to substantiate your answers:

1. Is a 16 year old named Quinn more likely to be a boy or a girl?

Initialize the variables for the scale

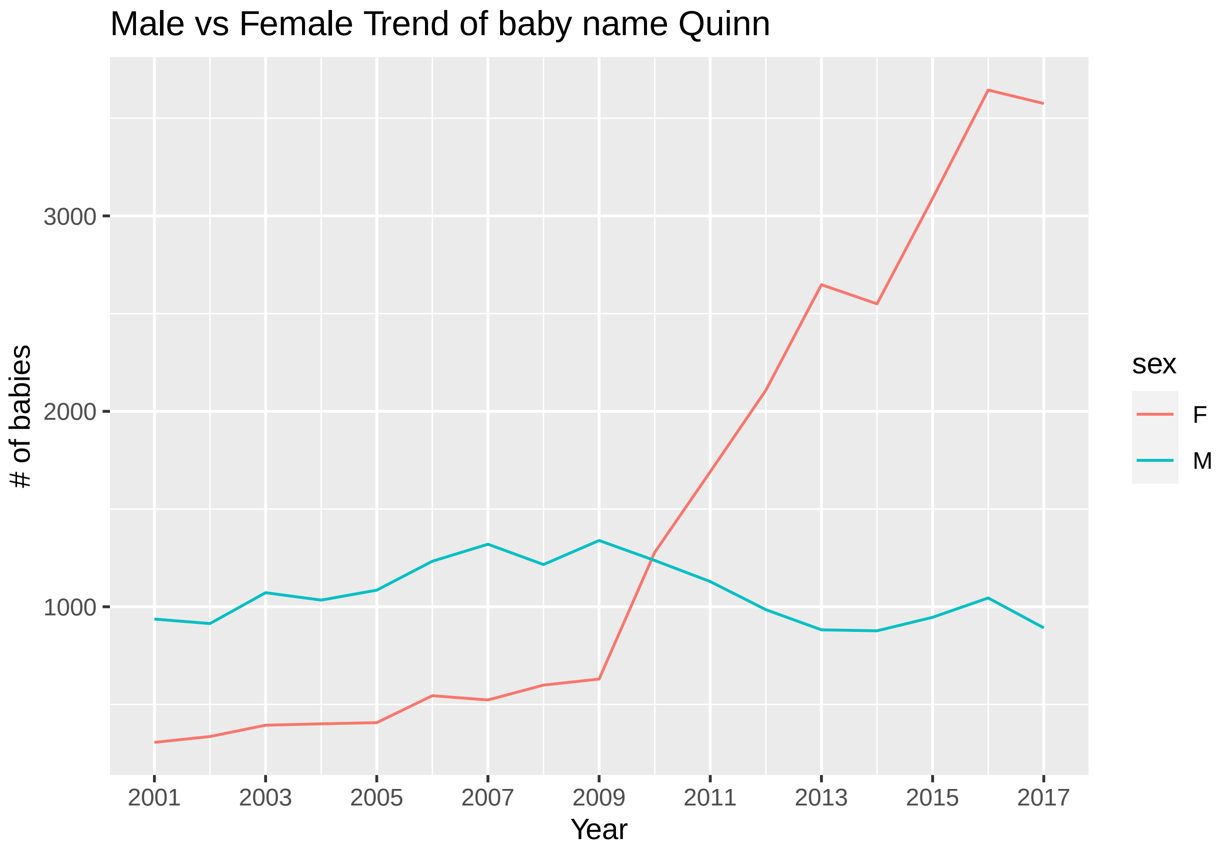
```
max_year <- max(babynames$year)
year_16 <- max_year - 16
```

Filter the babynames data set for the name Quinn, and the year is 2001 onward

```
baby_Quinn = filter(babynames, name=="Quinn", year>= year_16)
```

Create a ggplot using the newly filtered data where the x-axis is the year, y-axis is the count of the number of babies, and the color of each line is defined by the sex variable. The scale is set at 2 year increments starting 2001 and ending 2017

```
ggplot(data = baby_Quinn) +
  geom_line(mapping=aes(x=year, y=n, color=sex)) +
  labs(title = "Male vs Female Trend of baby name Quinn",
       x = "Year",
      y = "# of babies") +
  scale_x_continuous(breaks = seq(year_16, max_year, by = 2))
```



According to the plot above, a 16 year old named Quinn was born in 2001 since the data set ends in 2017. Even if one were to consider the current year to be 2021(which would mean Quinn was born in 2005,Quinn was still mostly a boy's name.)

2. Is a 2 year old named Quinn more likely to be a boy or a girl?

According to the plot above, a 2 year named Quinn would have been born in 2015. During this time, Quinn was overwhelmingly a girl's name. Since the data ends in 2017, it is impossible to consider the current year to be 2021 with a baby born in 2019.

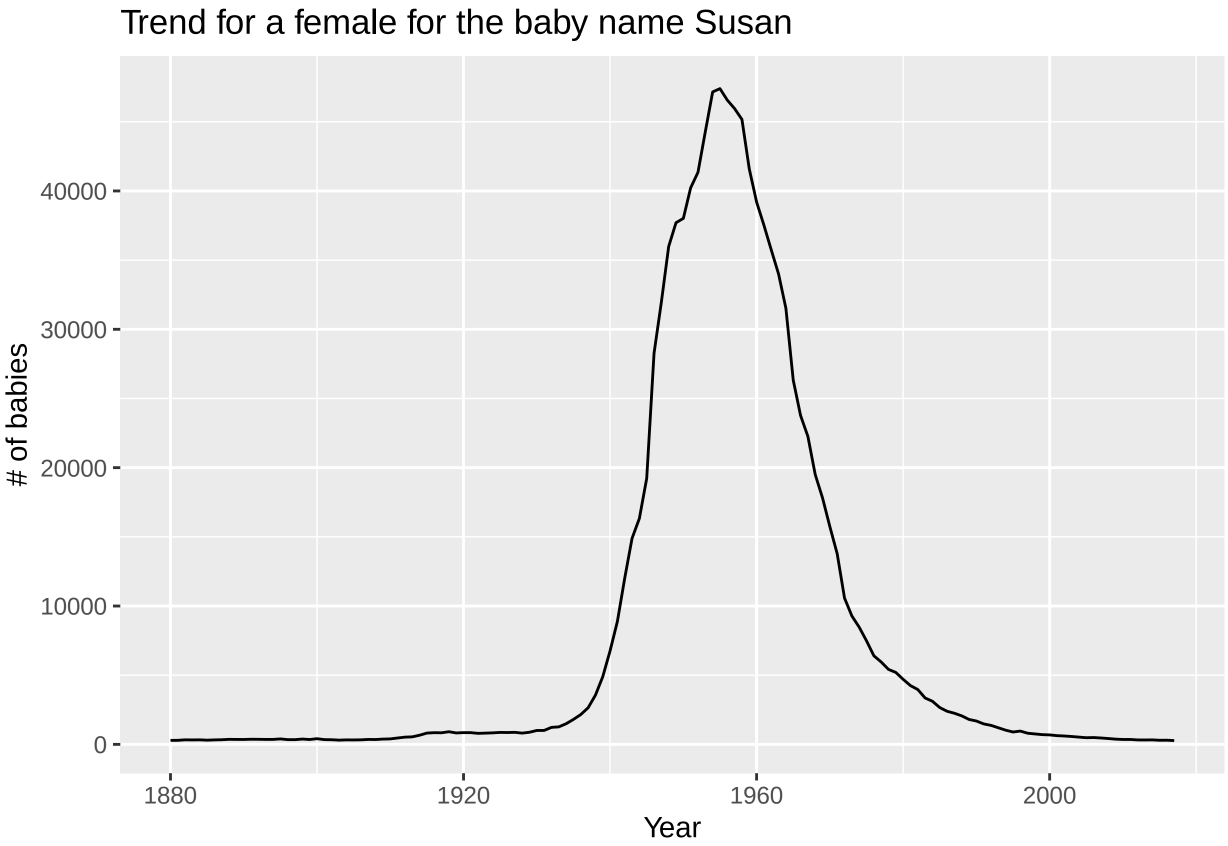
3. What is your best guess as to how old a woman named Susan is?

Filter the babynames dataset for the name Taylor

```
susan_female <- filter(babynames, name=="Susan", sex == "F")
```

Create a ggplot using the newly filtered data where the x-axis is the year, y-axis is the count of the number of babies

```
ggplot(data = susan_female) +
  geom_line(mapping=aes(x=year, y=n)) +
  labs(title = "Trend for a female for the baby name Susan",
       x = "Year",
       y = "# of babies")
```



According to the plot above, the popularity of the baby name Susan can be approximated by the normal distribution with a peak popularity between 1950 and 1960. If one takes 2017 to be the final year, that would put our Susan at between 57 and 67. I believe the question to be ambiguous as even in 1980 there are roughly 5000 Susans born. For all we know, there was an attack on the Susans born between 1950 and 1960 and the only Susans left were born in 1980 making them 37 years old.