Command to get data set in R

```
library(readr)
birthsData <-
read_csv(url("https://raw.githubusercontent.com/fivethirtyeight/data/refs/heads/master/births/US_births_2000-2014_SSA.csv"))
```

GREEN still has to be done
RED is completed
GRAY is finished code
YELLOW is written explanation portion for the Quarto document

1. Describe the provenance of your data. That is, where did you get the data, who collected the data, for what purpose, who/what make up the cases.

The data set we are using is from the github 538 data library, titled “births”. Each case represents the number of births on a specific day of the year. This data was contributed to by Jay Boice, Dhrumil Mehta, and Andrew Flowers, with data provided by the Social Security Administration to keep written records on US births.

US Births (there's a lot of data we can just focus on 2000)-
[data/births/US_births_2000-2014_SSA.csv at master · fivethirtyeight/data · GitHub]

	year	month	date_of_month	day_of_week	births
2	2000	1	1	6	9083
3	2000	1	2	7	8006
4	2000	1	3	1	11363
5	2000	1	4	2	13032
6	2000	1	5	3	12558
7	2000	1	6	4	12466
8	2000	1	7	5	12516
9	2000	1	8	6	8934
10	2000	1	9	7	7949
11	2000	1	10	1	11668
12	2000	1	11	2	12611
13	2000	1	12	3	12398
14	2000	1	13	4	11815
15	2000	1	14	5	12180

2. Explain how your data meet the FAIR and/or CARE Principles.

FAIR

Findable: It is sourced directly from the US Social Security Administration and contains a large breadth of fortified birth data from the years 2000-2014. It is titled “birth” in the github fivethirtyeight data library, which is clear, concise, and easy to find.

Accessible: the dataset is open, free, and public use. It can be found by anyone at the link data/births/US_births_2000-2014_SSA.csv at master · fivethirtyeight/data · GitHub.

Interoperable: The data is clear with all value meanings. It is stated that for days of the week, 1 represents Monday, and 7 represents Sunday, with the days in between assigned in increasing order. The months and years are clearly stated through their number representations (12 for the month of December, 2000 for the year Two-Thousand)

Reusable: This data is easy to manipulate and handle for many different uses. It is high quality and easy to be used in the future by anyone who downloads the data set off github.

3. Describe what attributes you'll focus your analysis on (mention if they are part of your data sets or if you created them out of your data sets).

We are focusing on the year 2000 in our project. To do this, we will simply use the R filter command for the year 2000. We will answer questions regarding the patterns found in birth dates across the whole year.

- What was the most common month to be born in 2000?

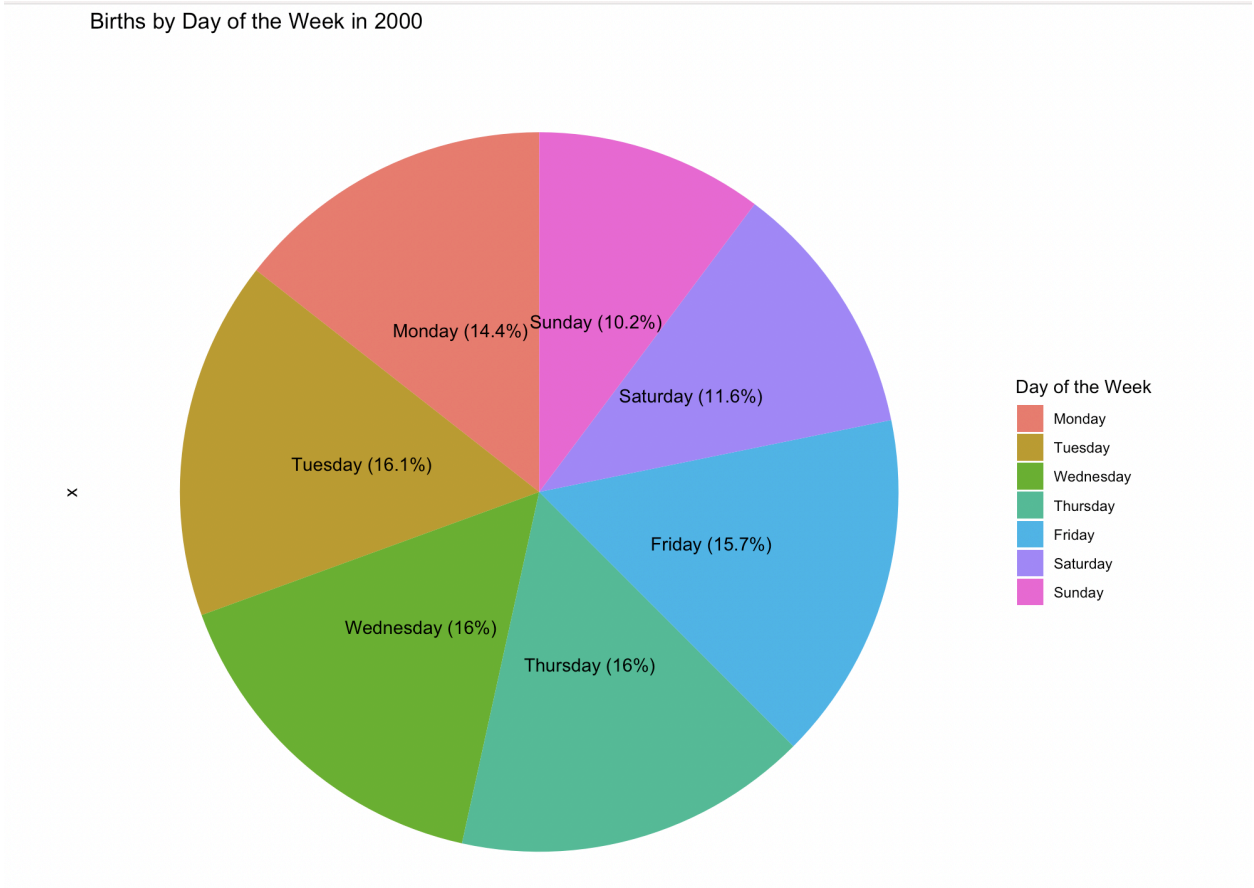
- Most common date of the month? (1-31)

-Compare 2000 to 2009, make line chart for each year to compare visual trends.
Find total births in 2000 and total births in 2009, compare the respective US population to find the birth rate: was the birth rate in 2009 smaller than 2000 due to the recession?

282.2 million (2000)
306.8 million (2009)

- What was the most common day of the week to be born in 2000?

1 = MONDAY
2= Tuesday
3= Wednesday
4= Thursday
5=Friday
6=Saturday
7 = SUNDAY



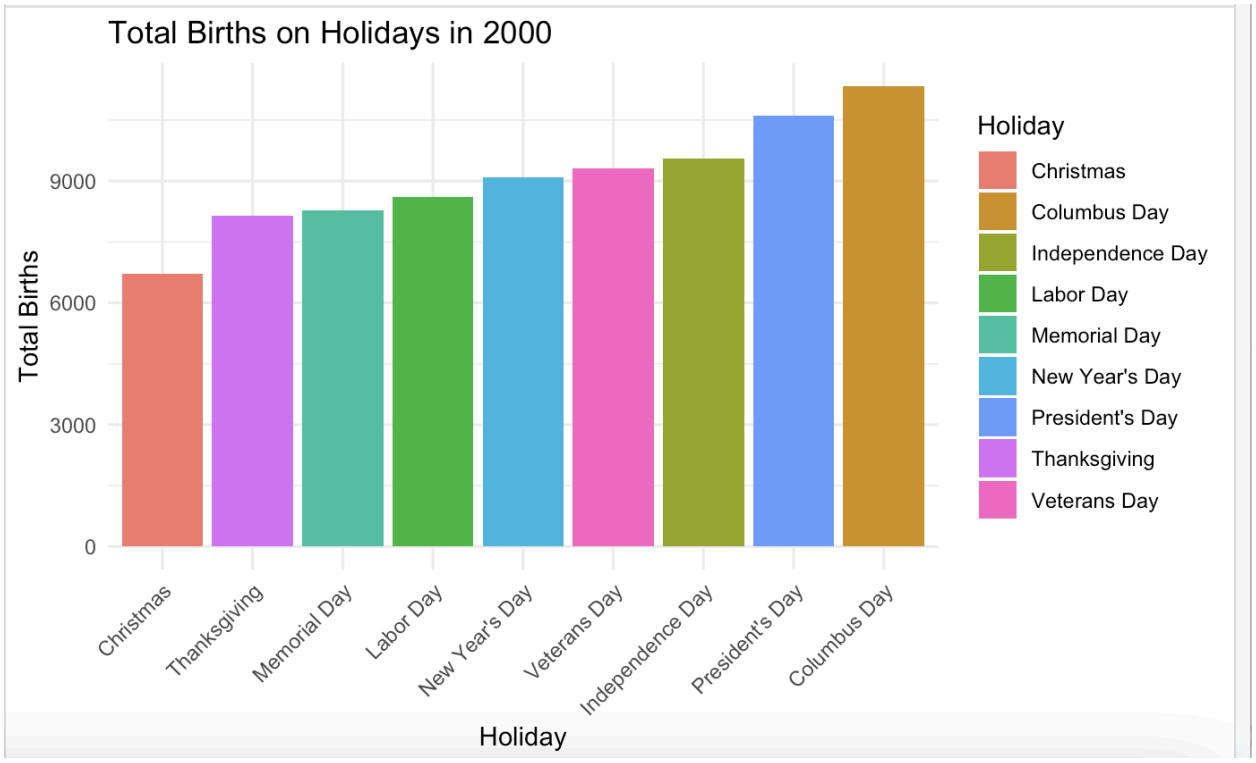

```
library(dplyr)
library(ggplot2)

births_2000 <- birthsData %>%
filter(year == 2000)

births_by_day <- births_2000 %>%
group_by(day_of_week) %>%
summarize(total_births = sum(births)) %>%
ungroup()
```

```
day_labels <- c("Monday","Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
births_by_day <- births_by_day %>%
mutate(day_label = factor(day_of_week, levels = 1:7, labels = day_labels))
daybirths_by_day <- births_by_day %>%
mutate(percentage = total_births / sum(total_births) * 100,
label =paste0(day_label, " (", round(percentage, 1), "%)"))
ggplot(births_by_day, aes(x = "", y =total_births, fill = day_label)) + geom_bar(stat = "identity",
width = 1) + coord_polar(theta = "y") + geom_text(aes(label = label),
position = position_stack(vjust = 0.5)) +
labs(title = "Births by Day of the Week in 2000",
fill = "Day of the Week") + theme_minimal() +
theme(axis.text = element_blank(),
axis.ticks = element_blank(),
panel.grid =element_blank())
```

While the data shows no dramatic bias towards a certain day of the week, there are subtle patterns revealed from the pie chart. Weekend days including Saturday and Sunday see slightly lower birth rates than Tuesday, Wednesday, and Thursday, which fall in the middle of the week. This could suggest that there is a slight preference for mothers and medical professionals to schedule non-emergency births on weekdays rather than weekends.

- What was the most common holiday to be born on? (calendar official holidays) Barchart to compare holidays and find the most avoided ones
OFFICIAL Bank Calendar Holidays in 2000 (No valentines or st patricks):

New Years Day 1/1/2000
President’s Day 2/21/00
Memorial Day 5/29/00
Independence Day 07/4/00
Labor Day 9/4/00
Columbus Day 10/9/00
Veterans Day 11/11/00
Thanksgiving 11/23/00
Christmas 12/25/00




```
library(ggplot2)
holidays <- data.frame(
  month = c(1, 2, 5, 7, 9, 10, 11, 11, 12),
  date_of_month = c(1, 21, 29, 4, 4, 9, 11, 23, 25),
  holiday_name = c("New Year's Day", "President's Day", "Memorial Day",
                   "Independence Day", "Labor Day", "Columbus Day",
                   "Veterans Day", "Thanksgiving", "Christmas")
)
  filter(year == 2000) %>%
  inner_join(holidays, by = c("month", "date_of_month"))

ggplot(filtered_data, aes(x = reorder(holiday_name, births), y = births, fill = holiday_name)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Births on Holidays in 2000",
       x = "Holiday", y = "Total Births") +
  scale_fill_discrete(name = "Holiday") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

This bar chart highlights an interesting trend: women are the least likely to give birth on two major US holidays, Thanksgiving and Christmas. In particular, Christmas stands out with a strikingly low 6719 births compared to a less relevant holiday, like Columbus Day, which saw 11343 births. Interestingly, the day after Christmas, December 26th, 2000, had a large spike

with 10395 births, a 54% increase. and the number of births on the day directly before Thanksgiving was 52% higher than Thanksgiving Day with 12420 births compared to 8144. The data suggests a deliberate effort from families and medical professionals to schedule births for the day before or after major holidays if possible.

- most common date to be born in 2000?

```
births_2000 <- birthsData %>%
  filter(year == 2000)
births_2000 <- births_2000 %>%
  mutate(date = as.Date(paste(year, month, date_of_month, sep = "-")))
births_by_date <- births_2000 %>%
  group_by(date) %>%
  summarize(total_births = sum(births)) %>%
  ungroup()
most_common_date <- births_by_date %>%
  filter(total_births == max(total_births))
print(most_common_date)
```

The most common date to be born in the year 2000 was November 21st with 13,991 births.