Homerwork 1

Andres Felipe Montoya Torres

5/14/23

# Data Manipulation

## Problem 1: Use logical operators to find flights that:

- Had an arrival delay of two or more hours (\> 120 minutes)  
- Flew to Houston (IAH or HOU)  
- Were operated by United (`UA`), American (`AA`), or Delta (`DL`)  
- Departed in summer (July, August, and September)  
- Arrived more than two hours late, but didn't leave late  
- Were delayed by at least an hour, but made up over 30 minutes in flight

# Had an arrival delay of two or more hours (> 120 minutes)  
library(nycflights13)  
data("flights")  
filter (flights, arr\_delay >= 120)

# A tibble: 10,200 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 1 1 811 630 101 1047 830  
 2 2013 1 1 848 1835 853 1001 1950  
 3 2013 1 1 957 733 144 1056 853  
 4 2013 1 1 1114 900 134 1447 1222  
 5 2013 1 1 1505 1310 115 1638 1431  
 6 2013 1 1 1525 1340 105 1831 1626  
 7 2013 1 1 1549 1445 64 1912 1656  
 8 2013 1 1 1558 1359 119 1718 1515  
 9 2013 1 1 1732 1630 62 2028 1825  
10 2013 1 1 1803 1620 103 2008 1750  
# ℹ 10,190 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

# Flew to Houston (IAH or HOU)  
library(nycflights13)  
data("flights")  
filter (flights, dest == "IAH" | dest=="HOU")

# A tibble: 9,313 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 1 1 517 515 2 830 819  
 2 2013 1 1 533 529 4 850 830  
 3 2013 1 1 623 627 -4 933 932  
 4 2013 1 1 728 732 -4 1041 1038  
 5 2013 1 1 739 739 0 1104 1038  
 6 2013 1 1 908 908 0 1228 1219  
 7 2013 1 1 1028 1026 2 1350 1339  
 8 2013 1 1 1044 1045 -1 1352 1351  
 9 2013 1 1 1114 900 134 1447 1222  
10 2013 1 1 1205 1200 5 1503 1505  
# ℹ 9,303 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

# Were operated by United (`UA`), American (`AA`), or Delta (`DL`)  
library(nycflights13)  
data("flights")  
filter (flights, carrier == "UA" | carrier=="AA" | carrier=="DL")

# A tibble: 139,504 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 1 1 517 515 2 830 819  
 2 2013 1 1 533 529 4 850 830  
 3 2013 1 1 542 540 2 923 850  
 4 2013 1 1 554 600 -6 812 837  
 5 2013 1 1 554 558 -4 740 728  
 6 2013 1 1 558 600 -2 753 745  
 7 2013 1 1 558 600 -2 924 917  
 8 2013 1 1 558 600 -2 923 937  
 9 2013 1 1 559 600 -1 941 910  
10 2013 1 1 559 600 -1 854 902  
# ℹ 139,494 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

# Departed in summer (July, August, and September)  
library(nycflights13)  
data("flights")  
filter (flights, month == 7 | month==8 | month==9)

# A tibble: 86,326 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 7 1 1 2029 212 236 2359  
 2 2013 7 1 2 2359 3 344 344  
 3 2013 7 1 29 2245 104 151 1  
 4 2013 7 1 43 2130 193 322 14  
 5 2013 7 1 44 2150 174 300 100  
 6 2013 7 1 46 2051 235 304 2358  
 7 2013 7 1 48 2001 287 308 2305  
 8 2013 7 1 58 2155 183 335 43  
 9 2013 7 1 100 2146 194 327 30  
10 2013 7 1 100 2245 135 337 135  
# ℹ 86,316 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

# Arrived more than two hours late, but didn't leave late  
library(nycflights13)  
data("flights")  
filter (flights, arr\_delay >= 120 & dep\_delay <=0)

# A tibble: 29 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 1 27 1419 1420 -1 1754 1550  
 2 2013 10 7 1350 1350 0 1736 1526  
 3 2013 10 7 1357 1359 -2 1858 1654  
 4 2013 10 16 657 700 -3 1258 1056  
 5 2013 11 1 658 700 -2 1329 1015  
 6 2013 3 18 1844 1847 -3 39 2219  
 7 2013 4 17 1635 1640 -5 2049 1845  
 8 2013 4 18 558 600 -2 1149 850  
 9 2013 4 18 655 700 -5 1213 950  
10 2013 5 22 1827 1830 -3 2217 2010  
# ℹ 19 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

# Were delayed by at least an hour, but made up over 30 minutes in flight  
library(nycflights13)  
data("flights")  
filter (flights, dep\_delay >=60 & arr\_delay <=30)

# A tibble: 239 × 19  
 year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
 <int> <int> <int> <int> <int> <dbl> <int> <int>  
 1 2013 1 3 1850 1745 65 2148 2120  
 2 2013 1 3 1950 1845 65 2228 2227  
 3 2013 1 3 2015 1915 60 2135 2111  
 4 2013 1 6 1019 900 79 1558 1530  
 5 2013 1 7 1543 1430 73 1758 1735  
 6 2013 1 11 1020 920 60 1311 1245  
 7 2013 1 12 1706 1600 66 1949 1927  
 8 2013 1 12 1953 1845 68 2154 2137  
 9 2013 1 19 1456 1355 61 1636 1615  
10 2013 1 21 1531 1430 61 1843 1815  
# ℹ 229 more rows  
# ℹ 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
# tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
# hour <dbl>, minute <dbl>, time\_hour <dttm>

## Problem 2: What months had the highest and lowest proportion of cancelled flights? Interpret any seasonal patterns. To determine if a flight was cancelled use the following code

flights %>%   
 filter(is.na(dep\_time))g

# What months had the highest and lowest % of cancelled flights?  
library(nycflights13)  
data("flights")  
flights %>%  
filter(is.na(dep\_time)) %>%   
group\_by (month) %>%   
summarise(count = n()) %>%   
mutate(prop = count/sum(count))

# A tibble: 12 × 3  
 month count prop  
 <int> <int> <dbl>  
 1 1 521 0.0631  
 2 2 1261 0.153   
 3 3 861 0.104   
 4 4 668 0.0809  
 5 5 563 0.0682  
 6 6 1009 0.122   
 7 7 940 0.114   
 8 8 486 0.0589  
 9 9 452 0.0548  
10 10 236 0.0286  
11 11 233 0.0282  
12 12 1025 0.124

## Problem 3: What plane (specified by the tailnum variable) traveled the most times from New York City airports in 2013? Please left\_join() the resulting table with the table planes (also included in the nycflights13 package).

For the plane with the greatest number of flights and that had more than 50 seats, please create a table where it flew to during 2013.

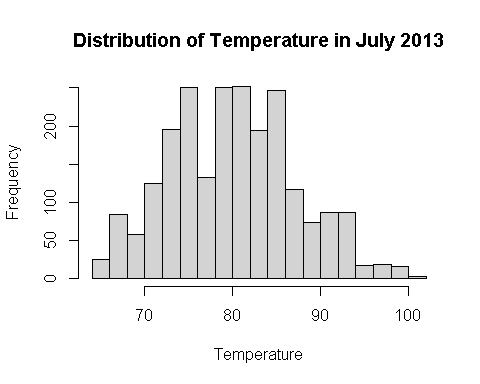
#Load data  
library(tidyverse)  
library(nycflights13)  
library(skimr)  
  
#create the object (table) that will show the planes by tailnum in descent order of number of times it flew to nyc  
Q3prime <- flights %>%   
 group\_by (tailnum) %>%   
 summarise(count = n()) %>%   
 arrange(desc(count))  
  
#join Q3prime (tailnum and number of flights to NYC) with planes  
Q3join <- left\_join(Q3prime, planes, by = "tailnum")  
   
# Filter flights with over 50 seats  
large\_planes <- Q3join %>% filter(seats > 50)  
  
# Join flights and large\_planes datasets  
flights\_with\_large\_planes <- flights %>% left\_join(large\_planes, by = "tailnum")  
  
# Filter flights with over 50 seats  
flights\_with\_large\_planes <- flights\_with\_large\_planes %>% filter(!is.na(seats))  
  
# Count number of flights for each plane  
num\_flights <- flights\_with\_large\_planes %>% group\_by(tailnum) %>% summarise(count = n())  
  
# Find the plane with the most flights  
most\_flights <- num\_flights %>% filter(count == max(count))  
  
# Join with flights to get flight details  
most\_flights\_with\_details <- most\_flights %>% left\_join(flights\_with\_large\_planes, by = "tailnum") %>% select(tailnum, flight)  
  
# Join with flights to get flight details  
most\_flights\_with\_details <- most\_flights %>% left\_join(flights\_with\_large\_planes, by = "tailnum")  
  
# Select columns to display  
most\_flights\_with\_details <- most\_flights\_with\_details %>%   
 select(tailnum, flight, origin, dest)  
  
# Display the flights with the most number of flights and over 50 seats  
most\_flights\_with\_details

# A tibble: 393 × 4  
 tailnum flight origin dest   
 <chr> <int> <chr> <chr>  
 1 N328AA 19 JFK LAX   
 2 N328AA 19 JFK LAX   
 3 N328AA 3 JFK LAX   
 4 N328AA 1 JFK LAX   
 5 N328AA 1 JFK LAX   
 6 N328AA 19 JFK LAX   
 7 N328AA 33 JFK LAX   
 8 N328AA 185 JFK LAX   
 9 N328AA 185 JFK LAX   
10 N328AA 181 JFK LAX   
# ℹ 383 more rows

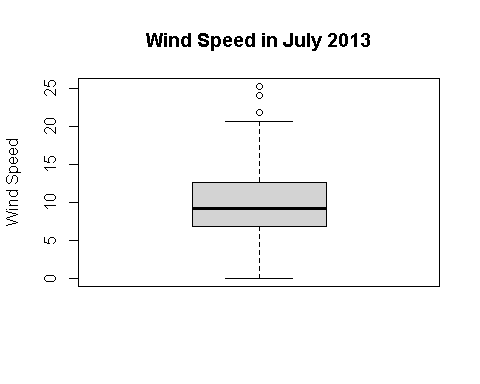
## Problem 4: The nycflights13 package includes a table (weather) that describes the weather during 2013. Use that table to answer the following questions:

- What is the distribution of temperature (`temp`) in July 2013? Identify any important outliers in terms of the `wind\_speed` variable.  
- What is the relationship between `dewp` and `humid`?  
- What is the relationship between `precip` and `visib`?

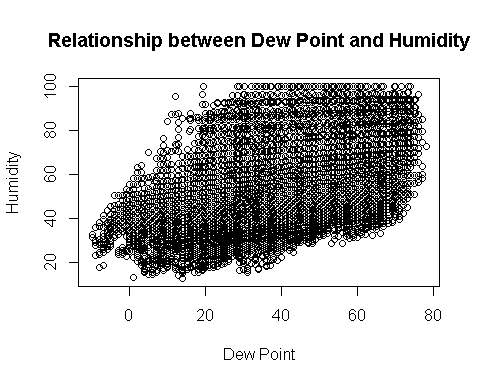
#Filter weather data to only include July 2013  
weather\_july <- filter(weather, year == 2013, month == 7)  
  
# Distribution of temperature in July 2013  
hist(weather\_july$temp, xlab = "Temperature", main = "Distribution of Temperature in July 2013")



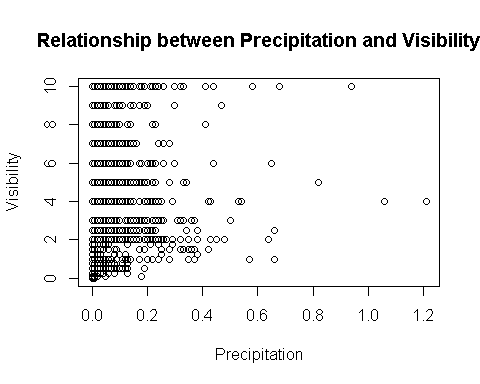
# Identify any important outliers in terms of wind\_speed  
boxplot(weather\_july$wind\_speed, main = "Wind Speed in July 2013", ylab = "Wind Speed")



# Scatterplot of dewp and humid  
plot(weather$dewp, weather$humid, xlab = "Dew Point", ylab = "Humidity", main = "Relationship between Dew Point and Humidity")



# Scatterplot of precip and visib  
plot(weather$precip, weather$visib, xlab = "Precipitation", ylab = "Visibility", main = "Relationship between Precipitation and Visibility")



## Problem 5: Use the flights and planes tables to answer the following questions:

- How many planes have a missing date of manufacture?  
- What are the five most common manufacturers?  
- Has the distribution of manufacturer changed over time as reflected by the airplanes flying from NYC in 2013? (Hint: you may need to use case\_when() to recode the manufacturer name and collapse rare vendors into a category called Other.)

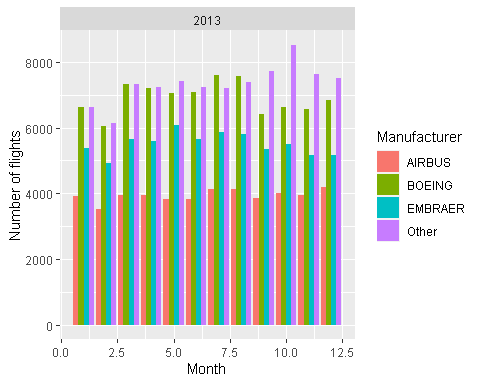
library(nycflights13)  
library(dplyr)  
library(ggplot2)  
# Count number of planes with a missing date of manufacture  
missing\_manufacture <- planes %>%   
 filter(is.na(year)) %>%   
 nrow()  
  
# Print number of planes with missing date of manufacture  
cat("Number of planes with missing date of manufacture:", missing\_manufacture, "\n")

Number of planes with missing date of manufacture: 70

# Count the number of planes by manufacturer and show top 5  
planes\_by\_manufacturer <- planes %>%   
 count(manufacturer, sort = TRUE) %>%   
 head(5)  
  
# Recode manufacturer names and collapse rare vendors into "Other"  
planes\_recode <- planes %>%  
 mutate(manufacturer = case\_when(  
 manufacturer %in% c("BOEING", "AIRBUS", "EMBRAER") ~ manufacturer,  
 TRUE ~ "Other"))  
  
# Join with flights data to get year of manufacture for each flight  
flights\_planes <- flights %>%  
 inner\_join(planes\_recode, by = "tailnum")  
  
# Group by month, manufacturer, and year  
flights\_by\_month\_manufacturer <- flights\_planes %>%  
 group\_by(month, year.x, manufacturer) %>%  
 summarise(n = n())

`summarise()` has grouped output by 'month', 'year.x'. You can override using  
the `.groups` argument.

# Plot the monthly distribution per manufacturer  
ggplot(flights\_by\_month\_manufacturer, aes(x = month, y = n, fill = manufacturer)) +  
 geom\_col(position = "dodge") +  
 facet\_wrap(~ year.x, ncol = 2) +  
 labs(x = "Month", y = "Number of flights", fill = "Manufacturer")



## Problem 6: Use the flights and planes tables to answer the following questions:

- What is the oldest plane (specified by the tailnum variable) that flew from New York City airports in 2013?  
- How many airplanes that flew from New York City are included in the planes table?

# Find the oldest plane  
oldest\_plane <- flights\_planes %>%  
 filter(!is.na(year.y)) %>%  
 arrange(year.y) %>%  
 slice(1) %>%  
 pull(tailnum)  
  
oldest\_plane

[1] "N381AA"

# count number of unique tailnums in resulting table  
num\_planes <- flights\_planes %>%  
 distinct(tailnum) %>%  
 nrow()  
  
cat("Number of airplanes that flew from New York City and are included in the planes table:", num\_planes, "\n")

Number of airplanes that flew from New York City and are included in the planes table: 3322

## Problem 7: Use the nycflights13 to answer the following questions:

- What is the median arrival delay on a month-by-month basis in each airport?  
- For each airline, plot the median arrival delay for each month and origin airport.

#Calculate de median by grouping by month and destination  
median\_arrival\_delay <- flights %>%  
 group\_by(month, dest) %>%  
 summarise(median\_arrival\_delay = median(arr\_delay, na.rm = TRUE))

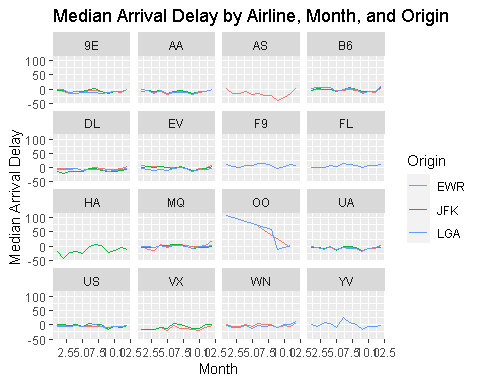
`summarise()` has grouped output by 'month'. You can override using the  
`.groups` argument.

#Now grouping by origin  
median\_arrival\_delay\_nyc <- flights %>%  
 group\_by(month, origin) %>%  
 summarise(median\_arrival\_delay = median(arr\_delay, na.rm = TRUE))

`summarise()` has grouped output by 'month'. You can override using the  
`.groups` argument.

#  
flights %>%  
 group\_by(month, origin, carrier) %>%  
 summarize(median\_arr\_delay = median(arr\_delay, na.rm = TRUE)) %>%  
 ggplot(aes(x = month, y = median\_arr\_delay, color = origin)) +  
 geom\_line() +  
 facet\_wrap(~ carrier, nrow = 5) +  
 labs(x = "Month", y = "Median Arrival Delay", color = "Origin", title = "Median Arrival Delay by Airline, Month, and Origin")

`summarise()` has grouped output by 'month', 'origin'. You can override using  
the `.groups` argument.

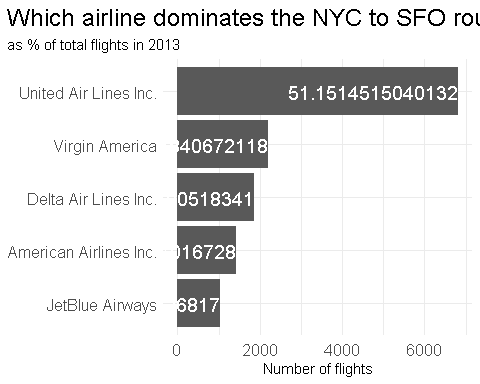


## Problem 8: Let’s take a closer look at what carriers service the route to San Francisco International (SFO). Join the flights and airlines tables and count which airlines flew the most to SFO. Produce a new dataframe, fly\_into\_sfo that contains three variables: the name of the airline, e.g., United Air Lines Inc. not UA, the count (number) of times it flew to SFO, and the percent of the trips that that particular airline flew to SFO.

# Join the flights and airlines tables  
fly\_sfo <- flights %>%  
 inner\_join(airlines, by = "carrier") %>%  
 filter(dest == "SFO")  
  
# Count the number of times each airline flew to SFO  
fly\_into\_sfo <- fly\_sfo %>%  
 group\_by(name) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 mutate(percent = count/sum(count)\*100)

And here is some bonus ggplot code to plot your dataframe

fly\_into\_sfo %>%   
   
 # sort 'name' of airline by the numbers it times to flew to SFO  
 mutate(name = fct\_reorder(name, count)) %>%   
   
 ggplot() +  
   
 aes(x = count,   
 y = name) +  
   
 # a simple bar/column plot  
 geom\_col() +  
   
 # add labels, so each bar shows the % of total flights   
 geom\_text(aes(label = percent),  
 hjust = 1,   
 colour = "white",   
 size = 5)+  
   
 # add labels to help our audience   
 labs(title="Which airline dominates the NYC to SFO route?",   
 subtitle = "as % of total flights in 2013",  
 x= "Number of flights",  
 y= NULL) +  
   
 theme\_minimal() +   
   
 # change the theme-- i just googled those , but you can use the ggThemeAssist add-in  
 # https://cran.r-project.org/web/packages/ggThemeAssist/index.html  
   
 theme(#  
 # so title is left-aligned  
 plot.title.position = "plot",  
   
 # text in axes appears larger   
 axis.text = element\_text(size=12),  
   
 # title text is bigger  
 plot.title = element\_text(size=18)  
 ) +  
  
 # add one final layer of NULL, so if you comment out any lines  
 # you never end up with a hanging `+` that awaits another ggplot layer  
 NULL



## Problem 9: Let’s take a look at cancellations of flights to SFO. We create a new dataframe cancellations as follows

cancellations <- flights %>%   
   
 # just filter for destination == 'SFO'  
 filter(dest == 'SFO') %>%   
   
 # a cancelled flight is one with no `dep\_time`   
 filter(is.na(dep\_time))  
  
#HOW TO PLOT: starting from the cancellations database, i would group the flights by carrier and month. Then to plot i would generate a grahp by carrier for each of the NYC airports that counts the number of cancellations (flights) by month.

I want you to think how we would organise our data manipulation to create the following plot. No need to write the code, just explain in words how you would go about it.



## Problem 10: On your own – Hollywood Age Gap

The website <https://hollywoodagegap.com> is a record of *THE AGE DIFFERENCE IN YEARS BETWEEN MOVIE LOVE INTERESTS*. This is an informational site showing the age gap between movie love interests and the data follows certain rules:

* The two (or more) actors play actual love interests (not just friends, coworkers, or some other non-romantic type of relationship)
* The youngest of the two actors is at least 17 years old
* No animated characters

The age gaps dataset includes “gender” columns, which always contain the values “man” or “woman”. These values appear to indicate how the characters in each film identify and some of these values do not match how the actor identifies. We apologize if any characters are misgendered in the data!

The following is a data dictionary of the variables used

| variable | class | description |
| --- | --- | --- |
| movie\_name | character | Name of the film |
| release\_year | integer | Release year |
| director | character | Director of the film |
| age\_difference | integer | Age difference between the characters in whole years |
| couple\_number | integer | An identifier for the couple in case multiple couples are listed for this film |
| actor\_1\_name | character | The name of the older actor in this couple |
| actor\_2\_name | character | The name of the younger actor in this couple |
| character\_1\_gender | character | The gender of the older character, as identified by the person who submitted the data for this couple |
| character\_2\_gender | character | The gender of the younger character, as identified by the person who submitted the data for this couple |
| actor\_1\_birthdate | date | The birthdate of the older member of the couple |
| actor\_2\_birthdate | date | The birthdate of the younger member of the couple |
| actor\_1\_age | integer | The age of the older actor when the film was released |
| actor\_2\_age | integer | The age of the younger actor when the film was released |

age\_gaps <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-02-14/age\_gaps.csv')

Rows: 1155 Columns: 13  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (6): movie\_name, director, actor\_1\_name, actor\_2\_name, character\_1\_gend...  
dbl (5): release\_year, age\_difference, couple\_number, actor\_1\_age, actor\_2\_age  
date (2): actor\_1\_birthdate, actor\_2\_birthdate  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

How would you explore this data set? Here are some ideas of tables/ graphs to help you with your analysis

* How is age\_difference distributed? What’s the ‘typical’ age\_difference in movies?
* The half plus seven\ rule. Large age disparities in relationships carry certain stigmas. One popular rule of thumb is the [half-your-age-plus-seven](https://en.wikipedia.org/wiki/Age_disparity_in_sexual_relationships#The_.22half-your-age-plus-seven.22_rule) rule. This rule states you should never date anyone under half your age plus seven, establishing a minimum boundary on whom one can date. In order for a dating relationship to be acceptable under this rule, your partner’s age must be:

$$\frac{\text{Your age}}{2} + 7 \\\\\\\< \text{Partner Age} \\\\\\\< (\text{Your age} - 7) \\\\\\\\* 2$$

How frequently does this rule apply in this dataset?

* Which movie has the greatest number of love interests?
* Which actors/ actresses have the greatest number of love interests in this dataset?
* Is the mean/median age difference staying constant over the years (1935 - 2022)?
* How frequently does Hollywood depict same-gender love interests?

# Deliverables

There is a lot of explanatory text, comments, etc. You do not need these, so delete them and produce a stand-alone document that you could share with someone. Render the edited and completed Quarto Markdown (qmd) file as a Word document (use the “Render” button at the top of the script editor window) and upload it to Canvas. You must be commiting and pushing tour changes to your own Github repo as you go along.

# Details

* Who did you collaborate with: TYPE NAMES HERE
* Approximately how much time did you spend on this problem set: ANSWER HERE
* What, if anything, gave you the most trouble: ANSWER HERE

**Please seek out help when you need it,** and remember the [15-minute rule](https://mam2022.netlify.app/syllabus/#the-15-minute-rule). You know enough R (and have enough examples of code from class and your readings) to be able to do this. If you get stuck, ask for help from others, post a question on Slack– and remember that I am here to help too!

As a true test to yourself, do you understand the code you submitted and are you able to explain it to someone else?

# Rubric

13/13: Problem set is 100% completed. Every question was attempted and answered, and most answers are correct. Code is well-documented (both self-documented and with additional comments as necessary). Used tidyverse, instead of base R. Graphs and tables are properly labelled. Analysis is clear and easy to follow, either because graphs are labeled clearly or you’ve written additional text to describe how you interpret the output. Multiple Github commits. Work is exceptional. I will not assign these often.

8/13: Problem set is 60–80% complete and most answers are correct. This is the expected level of performance. Solid effort. Hits all the elements. No clear mistakes. Easy to follow (both the code and the output). A few Github commits.

5/13: Problem set is less than 60% complete and/or most answers are incorrect. This indicates that you need to improve next time. I will hopefully not assign these often. Displays minimal effort. Doesn’t complete all components. Code is poorly written and not documented. Uses the same type of plot for each graph, or doesn’t use plots appropriate for the variables being analyzed. No Github commits.