Homework 1

MIKE DIZON

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

#Store flights dataframe as variable for easy reference  
flights <- flights  
  
# Had an arrival delay of two or more hours (> 120 minutes)  
delayed\_arrival\_flights <- flights %>%   
 filter(arr\_delay > 120)  
  
# Flew to Houston (IAH or HOU)  
houston\_flights <- flights %>%   
 filter(dest == "IAH" | dest == "HOU")  
  
# Were operated by United (`UA`), American (`AA`), or Delta (`DL`)  
UA\_AA\_DL\_flights <- flights %>%   
 filter(carrier == "UA" | carrier == "AA" | carrier == "DL")  
  
# Departed in summer (July, August, and September)  
summer\_flights <- flights %>%   
 filter(month %in% c(7,8,9))  
   
# Arrived more than two hours late, but didn't leave late  
late\_arrival\_flights <- flights %>%   
 filter(arr\_delay > 120 & dep\_delay <= 0)  
  
# Were delayed by at least an hour, but made up over 30 minutes in flight  
early\_arrival\_flights <- flights %>%   
 filter(dep\_delay >= 60 & arr\_delay - dep\_delay >= 30)

## Problem 2: What months had the highest and lowest proportion of cancelled flights? Interpret any seasonal patterns.

# What months had the highest and lowest % of cancelled flights?  
  
#Calculate the total number of flights per month  
total\_monthly\_flights <- flights %>%   
#Group the flights dataframe by month  
 group\_by(month) %>%   
#Count the number of rows in each group  
 summarize(n())  
  
#Calculate the number of cancellations per month   
monthly\_cancellations <- flights %>%  
#Filter for departure times that have NA (i.e., cancelled flights)  
 filter(is.na(dep\_time)) %>%  
#Group the cancelled flights by month  
 group\_by(month) %>%   
#Count the number of cancelled flights by month  
 summarize(cancelled = sum(is.na(dep\_time))) %>%  
#Calculate the proportion of cancelled flights over the total monthly flights  
 mutate(prop\_cancelled = cancelled/total\_monthly\_flights)

**Answer:** February had the highest number of cancelled flights as a proportion of total monthly flights at 5.05% while October had the lowest at 0.09%. The proportion of cancelled flights are expectedly higher during the start and end of the year, as well as over the summer season as these are the times where there are relatively more flights, and hence have higher chances of cancellations.

## 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.

#Store planes dataframe as variable for easy reference  
planes <- planes  
  
#Filter flights with non-missing tail number and NYC origins  
flights\_filtered <- flights %>%  
 filter(!is.na(tailnum), origin %in% c("JFK", "EWR", "LGA"))  
  
#Compute number of flights for each tailnum and left join with planes  
flights\_by\_tailnum <- flights\_filtered %>%   
#Group flights by tail number  
 group\_by(tailnum) %>%   
#Count number of rows  
 summarize(total=n()) %>%   
#Arrange in descending order to identify tail number with most number of flights  
 arrange(desc(total)) %>%  
#Left join resulting table with planes dataframe  
 left\_join(planes, by = "tailnum")  
  
#Filter planes with more than 50 seats and select the one with the most flights  
plane\_with\_most\_flights <- flights\_by\_tailnum %>%  
#Filter data to get only those with seats grater than 50  
 filter(seats>50) %>%   
#Identify the tail number with the most flights  
 top\_n(1,total) %>%   
 select(tailnum)  
  
#Get the destinations for the selected plane during 2013  
destinations <- flights\_filtered %>%   
#Filter the tail number with the most flights  
 filter(tailnum == plane\_with\_most\_flights$tailnum) %>%   
#Filter the year to 2013  
 filter(year == 2013) %>%  
#Select the destinations where the plane flew  
 select(dest)  
  
#Compute a table of frequencies of destinations  
table(destinations)

dest  
BOS LAX MCO MIA SFO SJU   
 1 313 1 25 52 1

**Answer:** The plane that travelled most from New York City airports in 2013 had a tail number of N725MQ.

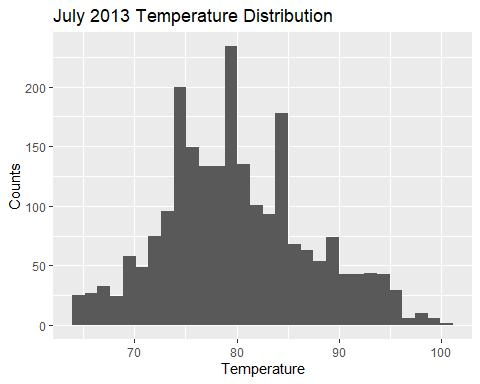
N328AA was the plane that had more than 50 seats. The destinations that it flew to in 2013 were BOS (1), LAX(313), MCO(1), MIA(25), SFO(52), and SJU(1).

## 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`?

#Store weather dataframe as variable for easy reference  
weather <- weather  
  
#Filter for data only for the month of July  
weather %>%   
 filter(month == 7) %>%   
   
#Use ggplot() to plot the data  
 ggplot() +  
#Use temperature as the x-axis  
 aes(x = temp) +  
#Visualize data using histogram to see the distribution  
 geom\_histogram() +  
#Include labels for the resulting diagram  
 labs(x = "Temperature", y = "Counts", title = "July 2013 Temperature Distribution")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

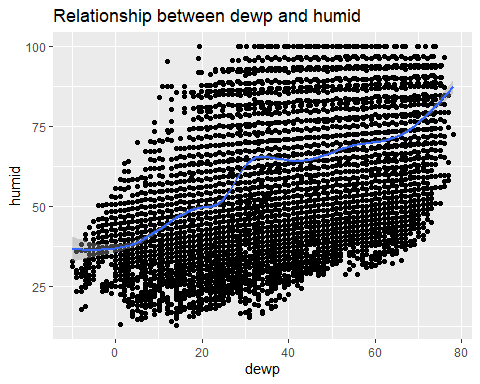


#Identify the quantile of wind\_speed and remove NA values  
quartile <- quantile(weather$wind\_speed, na.rm = TRUE)  
#Identify interquartile ranges as difference of quartile 4 and 2  
IQR <- as.numeric(quartile[4]) - as.numeric(quartile[2])  
#Identify outliers based on those who have wind speed values that are 3 IQR above third quartile or below the first quartile  
outliers <- weather%>%  
 filter((wind\_speed >= as.numeric(quartile[4]) + 3\*IQR) | (wind\_speed <= as.numeric(quartile[2]) - 3\*IQR)) %>%  
#Arrange filtered data to identify outlier  
 arrange(desc(wind\_speed))  
  
#Plot the data to identify relationship between dewp and humid  
weather %>%   
 ggplot() +  
#Use dewp as x-axis and humid as y-axis  
 aes(x = dewp, y = humid) +  
#Use scatter plot to visualize data points  
 geom\_point() +  
#Add fitted line to the plot to identify relationship  
 geom\_smooth() +  
#Include labels in the plot  
 labs(title = "Relationship between dewp and humid")

`geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 1 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 1 rows containing missing values (`geom\_point()`).



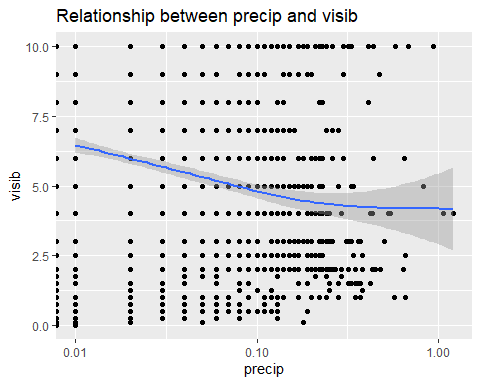
#Plot the data to identify relationship between precip and visib  
weather %>%   
 ggplot() +  
#Use precip as x-axis and visib as y-axis  
 aes(x = precip, y = visib) +  
#Use scatter plot to visualize data points  
 geom\_point() +  
#Add fitted line to the plot to identify relationship  
 geom\_smooth() +  
#Scale data to see trend clearly  
 scale\_x\_continuous(trans = "log10") +  
#Include labels in the plot  
 labs(title = "Relationship between precip and visib")

Warning: Transformation introduced infinite values in continuous x-axis

Warning: Transformation introduced infinite values in continuous x-axis

`geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 24366 rows containing non-finite values (`stat\_smooth()`).



**Answer:** We can observe that the distribution of the temperature for July 2013 is skewed to the right.

There also exists a positive relationship between dewp and humid and a negative relationship betwen precip and visib.

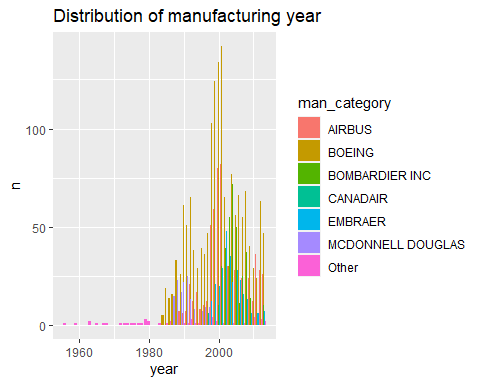
## Problem 5: Use the flights and planes tables to answer the following questions:

- 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.)

#Identify number of planes with missing date of manufacture  
missing\_dates <- planes %>%   
#Calculate the total number of planes  
 summarize(sum(is.na(year)))  
  
#Identify the five most common manufacturers  
top\_5\_manufacturers <- planes %>%   
#Count the number of manufacturers and sort number in descending order  
 count(manufacturer, sort = TRUE) %>%   
#Extract the top five values  
 top\_n(5)

Selecting by n

#Check whether the distribution of manufacturer changed over time  
planes1<- planes %>%   
#Count the combination of manufacturer and year  
 count(manufacturer, year) %>%   
#Add a new column to identify category of manufacturers between those with more than five and those less than five  
 mutate(man\_category = case\_when(n>=5 ~ manufacturer, TRUE ~ "Other")) %>%   
#Remove those with NA values  
 filter(!is.na(year))  
  
#Recode the names in planes1 dataframe to avoid duplicate values using case\_when  
planes1$man\_category <- case\_when(planes1$man\_category == "AIRBUS INDUSTRIE" ~ "AIRBUS", planes1$man\_category %in% c("MCDONNELL DOUGLAS AIRCRAFT CO", "MCDONNELL DOUGLAS CORPORATION") ~ "MCDONNELL DOUGLAS",  
 planes1$man\_category == "CANADAIR LTD" ~ "CANADAIR LTD",  
 TRUE ~ planes1$man\_category  
)  
  
#Use ggplot() to plot planes1 dataframe  
planes1 %>%   
 ggplot() +  
#Use year as x-axis and the number of manufacturers as y-axis  
 aes(x = year, y = n) +  
#Plot using stacked columns to see shape of distribution and color based on manufacturer category and also reduce the width of each column  
 geom\_col(aes(fill = man\_category), position = "dodge") +  
#Include labels in the plot  
 labs(title = "Distribution of manufacturing year")



**Answer:** A total of 70 planes have missing numbers of manufacture.

The top five most common manufacturers are BOEING (1630), AIRBUS INDUSTRIE (400), BOMBARDIER INC (368), AIRBUS (336), and EMBRAER (299).

The distribution of manufacturers have significantly changed over time. Prior to the 1980s, the manufacturing of planes was dominated by small manufacturers. However, by 1990s, the industry has gradually progressed with an increasing number of manufacturers. During this period up to 2000s, we can see that the industry has been dominated by Boeing although its market share has been slightly decreasing with other competitors such as Airbus starting to gain market share.

## 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?

#Identify the oldest plane from NYC airports in 2013  
oldest\_plane <- planes %>%   
#Right join the planes dataframe with the flights dataframe by tail number  
 right\_join(flights, by = "tailnum") %>%   
#Arrange data in ascending order by year  
 arrange(year.x) %>%   
#Select the tail number and year columns only  
 select(tailnum, year.x) %>%   
#Extract the top row from the dataframe to see oldest plane  
 head(1)  
  
#Identify the number of airplanes that flew from NYC that are included in the planes table  
planes\_nyc <- flights %>%   
#Inner join the planes and flights dataframe by tail number  
 inner\_join(planes, by = "tailnum") %>%   
#Count the number of planes using their tail numbers   
 summarize(n = n\_distinct(tailnum))

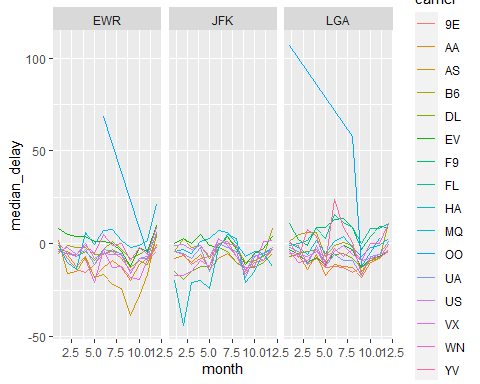
**Answer:** The oldest plane that flew from New York City airports in 2013 was the N381AA, which was manufactured in 1956.

The number of airplanes that flew from New York City that are included in the planes table is 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.

#Identify median arrival delay on a monthly basis for each airport  
median\_arrival\_delay <- flights %>%   
#Select carrier, origin, month, and arr\_delay columns from flights dataframe  
 select(carrier, origin, month, arr\_delay) %>%  
#Group data by carrier, origin, and month  
 group\_by(carrier, origin, month) %>%  
#Calculate median of arrival delay  
 summarize(median\_delay = median(arr\_delay, na.rm = TRUE), .groups = "drop")  
  
#Use ggplot() to plot median arrival delay  
median\_arrival\_delay %>%   
ggplot() +  
#Use month as x-axis, median delay as y-axis, and color by carrier  
 aes(x = month, y = median\_delay, color = carrier) +  
#Plot the data using a line chart  
 geom\_line() +  
#Obtain subplots for each airport  
 facet\_wrap(~ origin)



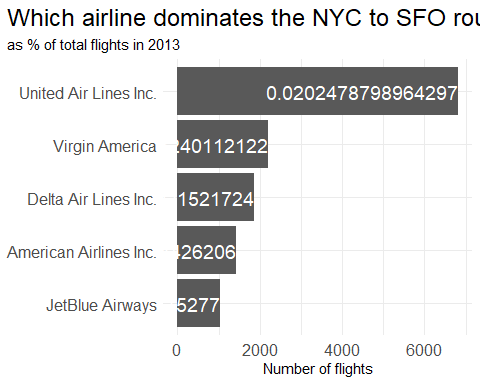
## Problem 8: Let’s take a closer look at what carriers service the route to San Francisco International (SFO). Join the flights and airlines tables and count which airlines flew the most to SFO. Produce a new dataframe, fly\_into\_sfo that contains three variables: the name of the airline, e.g., United 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.

#Left join airlines dataframe with flights dataframe and store into variable fly\_into\_sfo  
fly\_into\_sfo <- flights %>%   
 left\_join(airlines, by = "carrier") %>%   
#Filter the San Francisco destination  
 filter(dest == "SFO") %>%  
#Group data by name of airline  
 group\_by(name) %>%   
#Count the number of flights to SFO per airline and calculate the corresponding proportion of total flights   
 summarize(count = n(), percent = n() / nrow(flights), .groups = "drop")

**Answer:** It can be observed that of all the flights with San Francisco as destination, United Airlines had the most at 6819 while JetBlue Airways had the least at 1035.

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))

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.

**Answer:** I would start by creating a new dataframe that would summarize the cancellations by month, carrier, and airport origin. I would then use the geom\_col() under the ggplot() function to do a bar graph for the resulting dataset. Finally, I would integrate the facet\_wrap() function to organize the data into subplots organized according to airport origin and carrier.

## 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 |

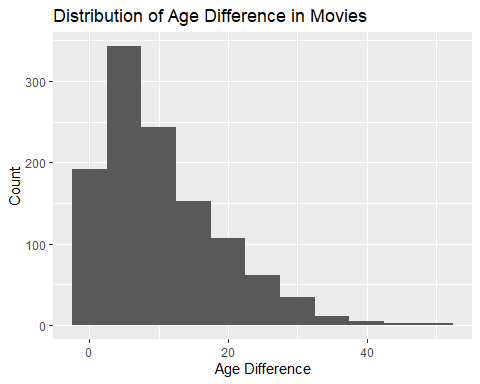
age\_gaps <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-02-14/age\_gaps.csv')

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?

#Use ggplot() to identify the shape of distribution of age difference  
age\_gaps %>%   
ggplot() +  
#Use the age difference as x-axis  
 aes(x = age\_difference) +  
#Plot data using histogram and adjust binwidth   
 geom\_histogram(binwidth = 5) +  
#Include labels in the plot  
 labs(x = "Age Difference", y = "Count",   
 title = "Distribution of Age Difference in Movies")



#Compute for the mean of the age difference  
mean(age\_gaps$age\_difference)

[1] 10.42424

#Compute for the median of the age difference  
median(age\_gaps$age\_difference)

[1] 8

**Answer:** The distribution of the age difference is skewed to the right, which means that most values are within the 0-20 range.

The mean age difference is 10.4 while the median age difference is 8.

* 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?

#Mutate age\_gaps dataframe  
age\_gaps <- age\_gaps %>%  
#Add half\_plus\_seven column which computes lower bound for the age difference  
 mutate(half\_plus\_seven = ((actor\_1\_age) / 2) + 7,  
#Add age\_max column which computes upper bound for the age difference  
 age\_max = ((actor\_1\_age) - 7) \* 2,  
#Add acceptable column which compares whether the age difference is within the established lower and upper bounds   
 acceptable = age\_difference >= half\_plus\_seven & age\_difference <= age\_max)  
  
# calculate the proportion of acceptable age differences  
prop\_acceptable <- mean(age\_gaps$acceptable)  
  
# calculate the frequency of acceptable age differences  
freq\_acceptable <- sum(age\_gaps$acceptable)

**Answer:** The half-your-age-plus-seven rule applies in 0.87% of cases (10 out of 1155). In other words, majority of cases do not conform to the rule in terms of age difference.

* Which movie has the greatest number of love interests?

#Group data by the name of movie  
most\_love\_interests <- age\_gaps %>%  
 group\_by(movie\_name) %>%  
#Compute for the sum of couple numbers per movie  
 summarise(num\_love\_interests = sum(couple\_number)) %>%  
#Arrange values in descending order  
 arrange(desc(num\_love\_interests)) %>%  
#Extract the top row  
 head(1)

**Answer:** Love Actually was the movie with the greatest number of love interests at 28.

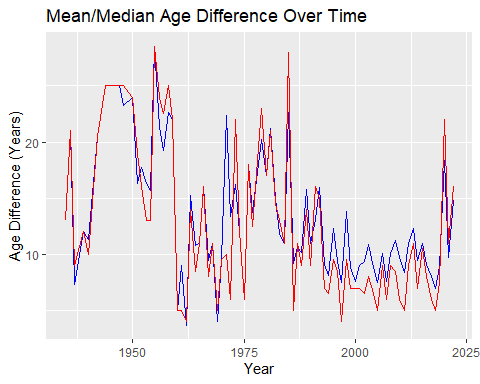
* Which actors/ actresses have the greatest number of love interests in this dataset?

#Group data by name of actors  
love\_interests <- age\_gaps %>%  
 group\_by(actor\_1\_name) %>%  
#Compute for distinct number of love interest for each specific actor  
 summarise(num\_love\_interests = n\_distinct(actor\_2\_name)) %>%  
#Arrange values in descending order  
 arrange(desc(num\_love\_interests)) %>%   
#Extract the top 5 actors  
 head(5)

**Answer:** The top five actors with the greatest number of love interests are Keanu Reeves (23), Adam Sandler (17), Roger Moore (16), Sean Connery (15), and Harrison Ford (13).

* Is the mean/median age difference staying constant over the years (1935 - 2022)?

#Group data by release year  
age\_diff\_by\_year <- age\_gaps %>%  
 group\_by(release\_year) %>%  
#Calculate the mean and meadian age difference by release year  
 summarise(mean\_age\_diff = mean(age\_difference), median\_age\_diff = median(age\_difference))  
  
#Use ggplot() to plot the mean and median age difference  
age\_diff\_by\_year %>%   
ggplot() +  
#Use release year as x-axis and mean age difference as y-axis  
 aes(x = release\_year, y = mean\_age\_diff) +  
#Plot data using line chart with blue color  
 geom\_line(color = "blue") +  
#Use line chart to also plot the median age difference with red as color  
 geom\_line(aes(y = median\_age\_diff), color = "red") +  
#Include labels in the plot  
 labs(x = "Year", y = "Age Difference (Years)", title = "Mean/Median Age Difference Over Time")



**Answer:** The mean and median of the age difference do not stay constant over time. It has peaked during the 1950s and slightly fell off during the 1960s-70s. However, it started to slightly increase again and peak by 1985. Around 2000s, the mean and median have once again decreased, but are starting to increase again in recent times.

* How frequently does Hollywood depict same-gender love interests?

#Filter for data with the same gender for both characters  
same\_gender <- age\_gaps %>%   
 filter(character\_1\_gender == character\_2\_gender) %>%  
#Count the values  
 count()

**Answer:** From 1935-2022, there have only been 23 instances wherein there were 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: N/A
* Approximately how much time did you spend on this problem set: 1 day
* What, if anything, gave you the most trouble: Using ggplots() for sophisticated graphs