Homerwork 1

Ignacio Landerreche

5/16/23

# 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)  
flights %>%   
 filter(arr\_time>120)

# A tibble: 319,950 x 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 1 544 545 -1 1004 1022  
 5 2013 1 1 554 600 -6 812 837  
 6 2013 1 1 554 558 -4 740 728  
 7 2013 1 1 555 600 -5 913 854  
 8 2013 1 1 557 600 -3 709 723  
 9 2013 1 1 557 600 -3 838 846  
10 2013 1 1 558 600 -2 753 745  
# i 319,940 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)  
flights %>%   
 filter(dest %in% c("IAH","HOU"))

# A tibble: 9,313 x 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`)  
flights %>%   
 filter(carrier %in% c("UA","AA","DL"))

# A tibble: 139,504 x 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)  
flights %>%   
 filter(month %in% c(7,8,9))

# A tibble: 86,326 x 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  
flights %>%   
 filter(arr\_delay>120 & dep\_delay==0)

# A tibble: 3 x 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 10 7 1350 1350 0 1736 1526  
2 2013 5 23 1810 1810 0 2208 2000  
3 2013 7 1 905 905 0 1443 1223  
# 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>

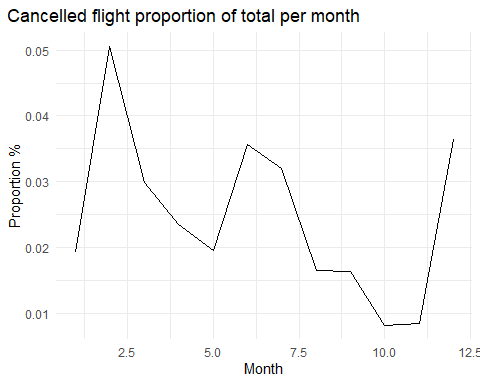
# Were delayed by at least an hour, but made up over 30 minutes in flight  
flights %>%   
 filter(arr\_delay>=-30 & arr\_delay<0 & dep\_delay>=60)

# A tibble: 3 x 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 2 26 1000 900 60 1513 1540  
2 2013 4 5 932 831 61 1149 1151  
3 2013 7 11 2018 1915 63 2210 2211  
# 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>

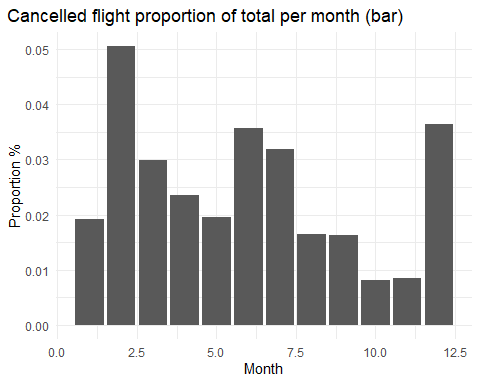
## Problem 2: What months had the highest and lowest proportion of cancelled flights? Interpret any seasonal patterns. To determine if a flight was cancelled use the following code

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

# What months had the highest and lowest % of cancelled flights?  
cancelled\_flights<-flights %>%   
 group\_by(month) %>% #groupping by month  
 summarise(total\_flights\_m=n(),cancelled=sum(is.na(dep\_time)),proportion\_can=cancelled/sum(total\_flights\_m)) #calculating total number of flights, the cancelled number of flights and the proportion  
  
  
#Creating a plot to see the month seasonality  
ggplot(cancelled\_flights,aes(x=month,y=proportion\_can))+geom\_line()+theme\_minimal()+labs(  
 title = "Cancelled flight proportion of total per month",  
 x= "Month",  
 y = "Proportion %")+ theme(plot.title.position = "plot")



ggplot(cancelled\_flights,aes(x=month,y=proportion\_can))+geom\_col()+theme\_minimal()+  
labs(  
 title = "Cancelled flight proportion of total per month (bar)",  
 x= "Month",  
 y = "Proportion %")+ theme(plot.title.position = "plot")

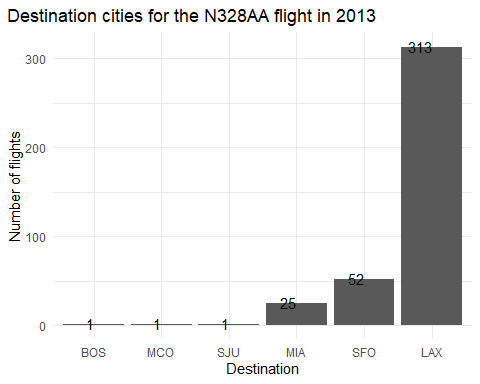


One can see from the above figure that there are three peaks: February, June and December. One could infer that June and December are high seasonality because of school vacations and another assumption would be that flights are cancelled because of cold weather and ergo Jan-March are months with a high proportion of cancelled flights. However without a deeper analysis one can’t conclude much from this.

## 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.

problem3<-flights %>%   
 filter(year==2013) %>%   
 left\_join(planes,by="tailnum") %>% #Left joining so we can have the number of seats by each flight   
 filter(seats>50)%>% #filtering by flights > 50 seats  
 group\_by(tailnum)%>% #groupping by flight number and counting it  
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
#Creating second part of the problem  
planemoreflown<-flights %>%  
 filter(year==2013) %>%   
 left\_join(planes,by="tailnum")%>%  
 filter(tailnum=="N328AA")%>% #Having only the flight >50 seats that flew the most  
 count(dest,sort=TRUE) %>%   
 mutate(dest=fct\_reorder(dest,n)) #arranging them for the graph  
  
#Creating the plot to show which were the cities where this plane flew the most  
ggplot(planemoreflown,aes(x=dest,y=n))+geom\_bar(stat="identity")+geom\_text(  
 aes(label = n, y = n + 1),  
 colour = "black",  
 size = 4,  
 hjust = 1  
 ) +  
 theme\_minimal()+labs(  
 title = "Destination cities for the N328AA flight in 2013",  
 x= "Destination",  
 y = "Number of flights")+ theme(plot.title.position = "plot")

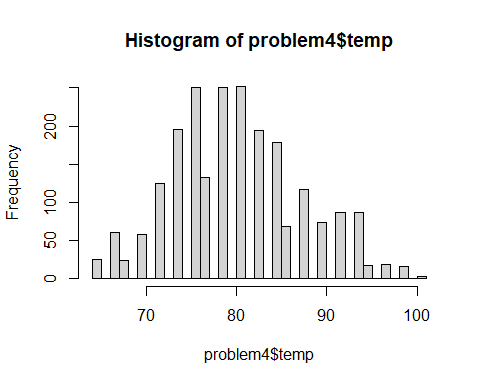


Based on the above, we know that *N328AA* was the plane with more than 50 seats with greatest number of flights from NYC ariports. The city the flight N3288AA flights the most is from NY to Los Angeles with 313 flights in 2013.

## 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`?

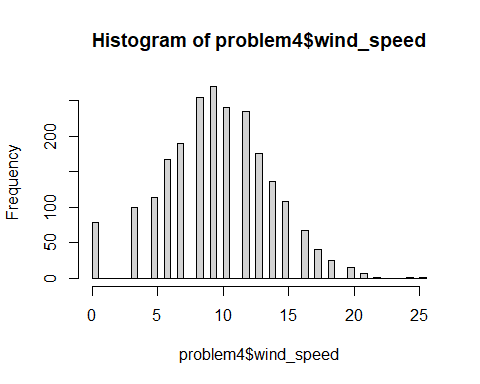
##Working on the data base  
problem4<-weather %>%   
 filter(year==2013 & month==7) %>%   
 arrange(wind\_speed)  
  
##Analyzing temperature  
hist(problem4$temp,breaks=50) #Distribution of temperature in July, on a bare eye it is difficult to see which distribution it might look like but it seems like it is a tri-modal distribution



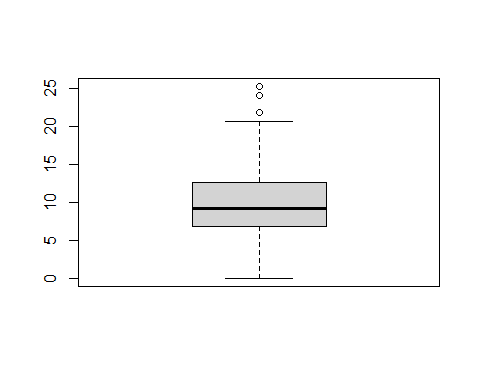
summary(problem4$temp) #basic stats from the sample of the temperature

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 64.04 75.02 78.98 80.07 84.20 100.04

##Analyzing wind\_speed  
hist(problem4$wind\_speed,breaks=50) #distribution of wind\_speed, looks more normal-like



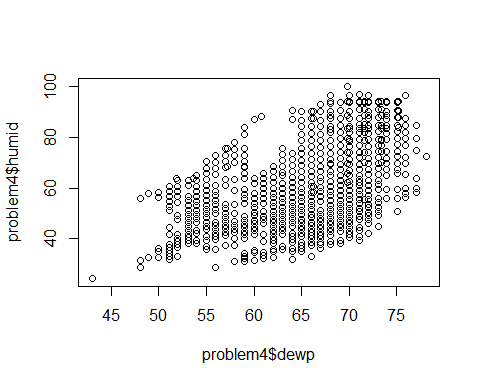
boxplot(problem4$wind\_speed) #with the boxplot we can identify the IQ range and outliers (specifically 3)



tail(problem4$wind\_speed,10)# these might be the outliers from the data distribution, everything above 20, basing everything also with the boxplot

[1] 20.71404 20.71404 20.71404 20.71404 20.71404 21.86482 24.16638 25.31716  
 [9] NA NA

##Relationship between dewp, and humid  
  
plot(problem4$dewp,problem4$humid) #there is no clear linear trend nor any quadratic trend based on the plot.



cor(problem4$dewp,problem4$humid) #based on the plot and the correlation, we see it is positive

[1] 0.5354781

plot(problem4$visib,problem4$precip) #looks like there is a relation between the two variables. Additionally, it looks like they are both discrete variables.

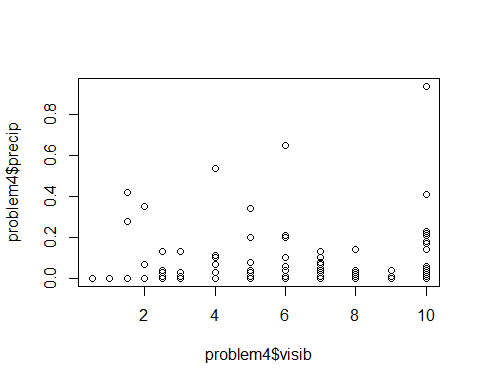
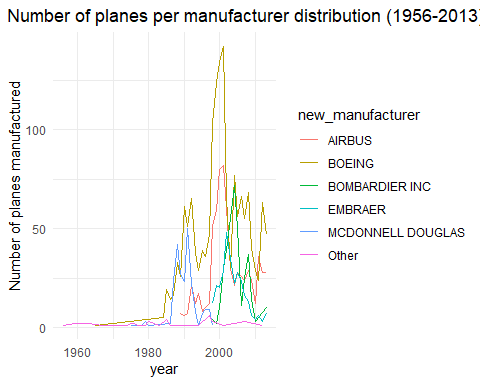


tabla1<-table(problem4$visib,problem4$precip) #based on this table, we see the relationship between these variables, but no further information can be gotten from this

## 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.)

#Creating the database with manufacturer as #NA  
missing\_manufacturer<- planes %>%   
 filter(is.na(manufacturer))  
view(missing\_manufacturer)  
  
#Creating database groupped by manufacturer, counting them and arranging them  
manufacturer<-planes %>%   
 group\_by(manufacturer) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
# It seems that there are couple of manufacturers that are the same but written differently, I will manually change them.  
total\_planes<- manufacturer%>% summarise(sum(count))  
#Changing the duplicate names  
  
manufacturer<-manufacturer %>%   
 mutate(manufacturer=case\_when(manufacturer=="AIRBUS INDUSTRIE" ~ "AIRBUS",  
 manufacturer %in% c("MCDONNELL DOUGLAS AIRCRAFT CO","MCDONNELL DOUGLAS AIRCRAFT CO","MCDONNELL DOUGLAS CORPORATION")~"MCDONNELL DOUGLAS",  
 .default=manufacturer  
   
 )) %>% group\_by(manufacturer) %>%   
 summarise(count=sum(count)) %>%   
 arrange(desc(count))  
  
#Validation:  
#validation<- manufacturer %>% summarise(sum(count))-total\_planes  
  
topmanufacturer<-head(manufacturer$manufacturer,5)   
  
#Creating a new database with new classification of manufacturer only mentioning the first 5 and then having "Other"  
planes\_modified<-planes %>%   
 mutate(manufacturer=case\_when(manufacturer=="AIRBUS INDUSTRIE" ~ "AIRBUS",  
 manufacturer %in% c("MCDONNELL DOUGLAS AIRCRAFT CO","MCDONNELL DOUGLAS AIRCRAFT CO","MCDONNELL DOUGLAS CORPORATION")~"MCDONNELL DOUGLAS",  
 .default=manufacturer  
   
 )) %>%   
 mutate(new\_manufacturer=case\_when(  
 manufacturer %in% topmanufacturer ~ manufacturer,  
 .default="Other"  
 ))  
  
  
#Groupping the newly created data base  
groupped\_manufacturer<- planes\_modified %>%   
 filter(!is.na(year))%>%   
 group\_by(new\_manufacturer,year)%>%   
 count(new\_manufacturer,year)  
  
#Plotting by year the new manufacturers and count  
ggplot(groupped\_manufacturer,aes(x=year,y=n,group=new\_manufacturer))+geom\_line(aes(color=new\_manufacturer))+theme\_minimal()+labs(  
 title = "Number of planes per manufacturer distribution (1956-2013)",  
 x= "year",  
 y = "Number of planes manufactured")+ theme(plot.title.position = "plot")



* Based on the above, there are no flights with missing manufacturer
* These are the top 5 manufacturers: “BOEING” ,“AIRBUS” , “BOMBARDIER INC” , “EMBRAER” “MCDONNELL DOUGLAS”
* Distribution has changed drastically since beginning, see above plot.

## 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?

problem6<-flights %>%   
 filter(year==2013) %>%   
 left\_join(planes,by="tailnum") %>%   
 arrange(year.y)  
  
problem6$tailnum[1]

[1] "N381AA"

problem6$year.y[1]

[1] 1956

problem6$dest[1]

[1] "SFO"

problem6$origin[1]

[1] "JFK"

#Anti\_joining the planes table to the flights  
problem6\_2<-planes %>%   
 anti\_join(flights,by="tailnum")  
view(problem6\_2)

* N381AA is the oldest plane built in 1956 and travelled from JFK to SFO
* Since the database is empty after doing the “anti\_join” and the “flights” database had only flights from NY, one may conclude that all of the planes in the “planes” table came from New York City’s airports

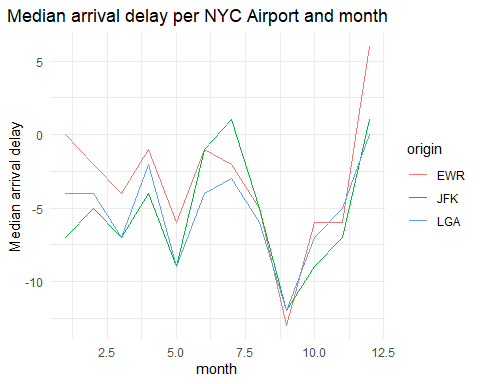
## Problem 7: Use the nycflights13 to answer the following questions:

- 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.

#Creating the data set groupped by month, origin and median delay  
problem7<-flights %>%   
 filter(year==2013) %>%   
 group\_by(month,origin)%>%   
 summarise(median\_delay=median(arr\_delay,na.rm=TRUE))

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

ggplot(problem7,aes(x=month,y=median\_delay,group=origin))+geom\_line(aes(color=origin))+theme\_minimal()+labs(  
 title = "Median arrival delay per NYC Airport and month",  
 x= "month",  
 y = "Median arrival delay")+ theme(plot.title.position = "plot")

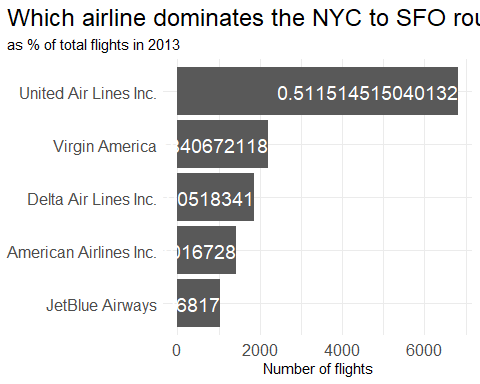


Creatting the plot with the above information- it seems that August is the month with less delays and december with most delays. The graph is very volatile within the months

## 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.

fly\_into\_sfo<- flights %>%  
 filter(dest=="SFO") %>%   
 left\_join(airlines,by="carrier")%>%   
 count(name, sort=TRUE)%>%   
 mutate(percent = n/sum(n),count=n)

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))



What I would do is the following:

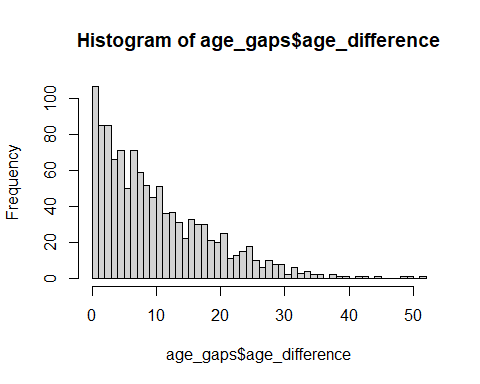
1. Filter only cancelled flights
2. Filter only EWR and JFK
3. Filter only the top 5 biggest airlines
4. Group by (1) and (2)
5. Graph it or use the facet graph

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

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.

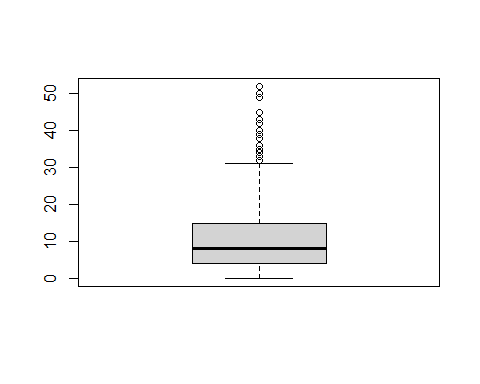
##Distribution of "Age\_difference"  
hist(age\_gaps$age\_difference,breaks=50)



summary(age\_gaps$age\_difference)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.00 4.00 8.00 10.42 15.00 52.00

boxplot(age\_gaps$age\_difference)



* It is a very right-skewed distribution, meaning that most of the age differences are low
* Additionally we can see the boxplot, IQ Range and the many outliers that exist with this difference
* One can see that the mean age difference is around 10 years, and half of the actors have at least 8 years of difference. The Maximum is 52 years which looks very like an outlier, as the 3rd quartile is just 15

##Rule of half plus seven  
  
age\_gaps\_new<-age\_gaps %>%   
 mutate(rule=case\_when(  
 (actor\_2\_age>(actor\_1\_age/2)+7&(actor\_1\_age-7)\*2>actor\_2\_age)~TRUE,  
 .default=FALSE  
   
 ))   
#Creating a table by groupping and calculating relative frequency  
test<-age\_gaps\_new%>% group\_by(rule) %>%   
 summarise(count=n()) %>%   
 mutate(freq=count/sum(count))  
  
  
print(test)

# A tibble: 2 x 3  
 rule count freq  
 <lgl> <int> <dbl>  
1 FALSE 360 0.312  
2 TRUE 795 0.688

* There are 795 actors who follow the rule (69% approximately)

##Movie with most love interests  
movie\_love<-age\_gaps %>%  
 group\_by(movie\_name) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
head(movie\_love,5)

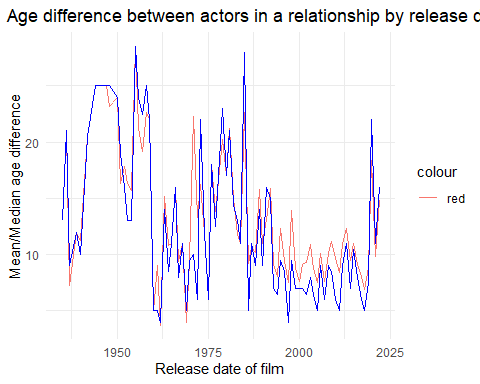
# A tibble: 5 x 2  
 movie\_name count  
 <chr> <int>  
1 Love Actually 7  
2 The Family Stone 6  
3 A View to a Kill 5  
4 He's Just Not That Into You 5  
5 Mona Lisa Smile 5

## Actors with most love interests  
#Creating a table only with actor 1 and actor 2  
actor1<-age\_gaps %>% select(actor=actor\_1\_name)  
actor2<-age\_gaps %>% select(actor=actor\_2\_name)  
#binding both tables into a single one  
list\_actors<-bind\_rows(actor1,actor2)  
  
#Groupping by name and arranging it descending order  
final\_list\_actors <- list\_actors %>% group\_by(actor) %>% summarise(count=n()) %>% arrange(desc(count))  
  
head(final\_list\_actors,5)

# A tibble: 5 x 2  
 actor count  
 <chr> <int>  
1 Keanu Reeves 27  
2 Adam Sandler 20  
3 Leonardo DiCaprio 17  
4 Roger Moore 17  
5 Sean Connery 17

* Movie with most love interests is called “Love actually” with 7 love interests
* Keanu Reeves is the actor that has had more love interests (27 in total)

##Mean/Median age by year  
mean\_median<-age\_gaps %>%   
 group\_by(release\_year) %>%   
 summarise(mean=round(mean(age\_difference),1),median=round(median(age\_difference),1))  
  
ggplot(mean\_median,aes(x=release\_year))+geom\_line(aes(y=mean,color="red"))+geom\_line(aes(y=median),color="blue")+theme\_minimal()+labs(  
 title = "Age difference between actors in a relationship by release date of the film",  
 x= "Release date of film",  
 y = "Mean/Median age difference")+ theme(plot.title.position = "plot")



* Conclusion: the mean/median ages have not been static throughout the years

#LGBTQ romances  
age\_gaps\_new<-age\_gaps\_new %>%   
 mutate(orientation=case\_when(  
 (character\_1\_gender=="man" & character\_1\_gender==character\_2\_gender)~"gay",  
 (character\_1\_gender=="woman" & character\_1\_gender==character\_2\_gender)~"lesbian",  
 .default="heterosexual"  
 ))  
  
summary\_orientation<-age\_gaps\_new %>%   
 group\_by(orientation) %>%   
 summarise(count=n()) %>%   
 mutate(frequency=count/sum(count))  
  
view(summary\_orientation)

* **Heterosexual couples are: 98% of total, gay relationships are 1.04% of total and lesbian are 0.952% of total**

# Details

* Who did you collaborate with: Only me
* Approximately how much time did you spend on this problem set: did not count, but at least 4 hours
* What, if anything, gave you the most trouble: **groupping by 2 variables and ggplot**

# 