Homerwork 1

Igor Kuznetsov

2023-05-14

[[1]]  
[1] TRUE  
  
[[2]]  
[1] TRUE  
  
[[3]]  
[1] TRUE  
  
[[4]]  
[1] TRUE

# 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

IK: The task is not clear in terms these conditions are AND or OR. I assume that these are separate conditions. Otherwise, it would not make sense.

# Had an arrival delay of two or more hours (> 120 minutes)  
flights %>% filter(arr\_delay >120, na.rm=TRUE)

# A tibble: 10,034 × 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 10,024 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"), na.rm=TRUE)

# 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`)  
flights %>% filter(carrier %in% c("UA","AA","DL"), na.rm=TRUE)

# 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)  
flights %>% filter(month %in% c(7,8,9), na.rm=TRUE)

# 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  
flights %>% filter(arr\_delay >120 & dep\_delay<=0, na.rm=TRUE)

# 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  
flights %>% filter(arr\_delay <=30 & dep\_delay>=60, na.rm=TRUE)

# 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

IK: The lowest proportion of cancelled flights is November. The weather is still good, no snow. not too cold and not too hot. The largest fraction of cancelations fall on February. The main reason is probably weather, and snow in particular.

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

# What months had the highest and lowest % of cancelled flights?  
flights %>%   
 filter(is.na(dep\_time)) %>% count(month, sort = TRUE) %>% mutate(prop = n/sum(n)) %>% filter(prop %in% c(max(prop),min(prop))) %>% arrange(prop)

# A tibble: 2 × 3  
 month n prop  
 <int> <int> <dbl>  
1 11 233 0.0282  
2 2 1261 0.153

## 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).

IK: 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.

# counting number of flights by the plain  
flights1 <- flights %>%   
 filter(!is.na(dep\_time)) %>% count(tailnum, sort = TRUE)   
  
# printing the result  
flights1

# A tibble: 4,037 × 2  
 tailnum n  
 <chr> <int>  
 1 N725MQ 546  
 2 N722MQ 487  
 3 N723MQ 480  
 4 N711MQ 467  
 5 N713MQ 455  
 6 N258JB 422  
 7 N353JB 403  
 8 N298JB 402  
 9 N351JB 392  
10 N328AA 389  
# ℹ 4,027 more rows

# joining the data with the plain info  
planes1 <- left\_join(x=flights1,y=planes,by ='tailnum' )  
  
# Filtering the data for the condition of at least 50 seats and max number of flights. We need to use filter twice here to 1) first select the planes with seats >50 and from the selected list we filter the plain that had the most flights (it is possible that the plane with the largest number of flights overall had <=seats)  
planes1<- planes1 %>% filter(seats>50) %>% filter(n==max(n)) %>% select('tailnum' )  
  
# Extracting the unique combinations of plains and destinations  
flights2 <- flights %>%   
 filter(!is.na(dep\_time)) %>% group\_by(tailnum,dest) %>% select (tailnum,dest)   
  
# creating a list of destinations for the select plain  
planes2<- left\_join(x=planes1,y=flights2,by ='tailnum' )   
  
# printing the results  
planes2

# A tibble: 389 × 2  
 tailnum dest   
 <chr> <chr>  
 1 N328AA LAX   
 2 N328AA LAX   
 3 N328AA LAX   
 4 N328AA LAX   
 5 N328AA LAX   
 6 N328AA LAX   
 7 N328AA LAX   
 8 N328AA LAX   
 9 N328AA LAX   
10 N328AA LAX   
# ℹ 379 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:

IK Note: The question is not clear. What does it mean by “distribution”? What specific statistics one needs to put together.

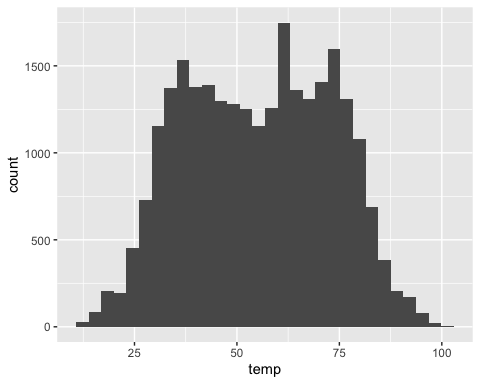
- What is the distribution of temperature (`temp`) in July 2013? Identify any important outliers in terms of the `wind\_speed` variable.  
- What is the relationship between `dewp` and `humid`?  
- What is the relationship between `precip` and `visib`?

IK: There is a positive relationship between dewp and humid but there does not seem to be a relationship between precip and visib

# plotting temp disctribution  
ggplot(weather, aes(x=temp))+geom\_histogram()

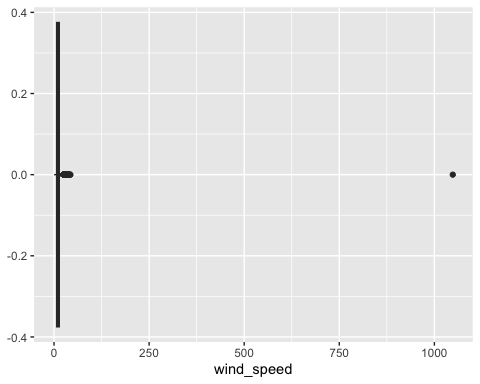
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

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



# box plot of wind speed to see outliers  
ggplot(weather, aes(x=wind\_speed))+geom\_boxplot()

Warning: Removed 4 rows containing non-finite values (`stat\_boxplot()`).

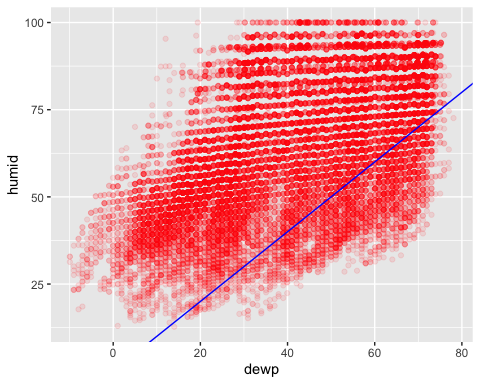


# selecting the top 1% in terms of wind spped and sorting in decending order  
weather %>% mutate(wind\_speed\_perc=percent\_rank(wind\_speed)\*100) %>% filter(wind\_speed\_perc>99) %>% arrange(desc(wind\_speed\_perc))

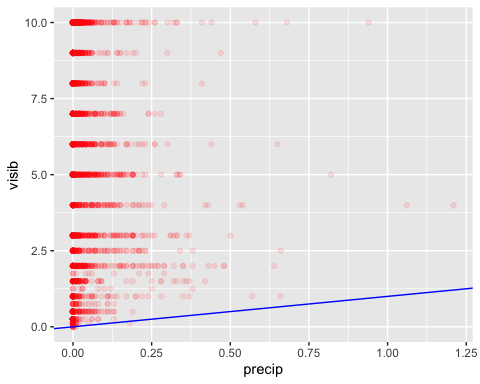
# A tibble: 219 × 16  
 origin year month day hour temp dewp humid wind\_dir wind\_speed  
 <chr> <int> <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 EWR 2013 2 12 3 39.0 27.0 61.6 260 1048.   
 2 EWR 2013 1 31 6 57.2 53.6 87.7 270 42.6  
 3 JFK 2013 1 31 4 53.6 53.1 100 200 42.6  
 4 EWR 2013 1 31 4 60.8 59 93.8 230 40.3  
 5 LGA 2013 1 31 4 59 55.4 93.7 230 40.3  
 6 EWR 2013 1 31 8 46.0 30.0 53.3 270 39.1  
 7 JFK 2013 3 6 14 41 28.9 61.9 50 38.0  
 8 JFK 2013 1 31 3 53.1 52.0 100 180 36.8  
 9 JFK 2013 1 31 7 51.8 46.4 81.7 270 36.8  
10 JFK 2013 11 24 10 28.0 -0.04 29.2 310 36.8  
# ℹ 209 more rows  
# ℹ 6 more variables: wind\_gust <dbl>, precip <dbl>, pressure <dbl>,  
# visib <dbl>, time\_hour <dttm>, wind\_speed\_perc <dbl>

# scatter and the fitted line for `dewp` and `humid`  
ggplot(weather, aes(x=dewp, y= humid), na.rm=TRUE)+geom\_point(alpha = 1/10,color = "red")+geom\_abline(color = "blue")

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



# scatter and the fitted line for `precip` and `visib`  
ggplot(weather, aes(x=precip, y= visib), na.rm=TRUE)+geom\_point(alpha = 1/10,color = "red")+geom\_abline(color = "blue")



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

- How many planes have a missing date of manufacture?  
IK: There are 70 such obs.  
- 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.)

IK: There does not seem to be large changes in terms of plane manufacturers, only the number of Bombardier planes have increased.

# counting the number of obs with missing year of manuf in planes data  
planes %>% filter(is.na(year)) %>% count()

# A tibble: 1 × 1  
 n  
 <int>  
1 70

# five most common manufacturers  
planes %>% group\_by(manufacturer) %>% summarise(plane\_count = n()) %>% arrange(desc(plane\_count)) %>% top\_n(5)

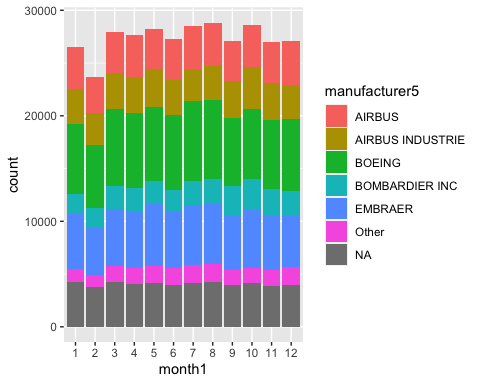
Selecting by plane\_count

# A tibble: 5 × 2  
 manufacturer plane\_count  
 <chr> <int>  
1 BOEING 1630  
2 AIRBUS INDUSTRIE 400  
3 BOMBARDIER INC 368  
4 AIRBUS 336  
5 EMBRAER 299

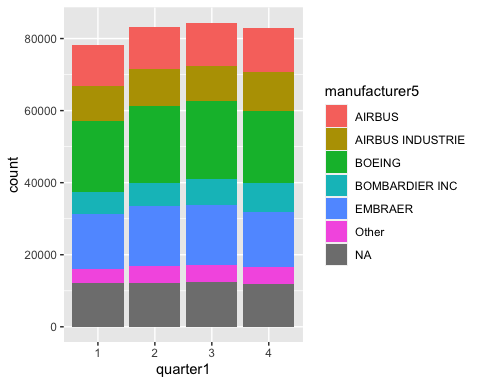
# - Has the distribution of manufacturer changed over time as reflected by the airplanes flying from NYC in 2013? (Hint: you may need to use case\_when() to recode the manufacturer name and collapse rare vendors into a category called Other.)  
  
  
# five most common manufacturers  
top\_5\_manuf <- planes %>% group\_by(manufacturer) %>% summarise(plane\_count = n()) %>% arrange(desc(plane\_count)) %>% top\_n(5)

Selecting by plane\_count

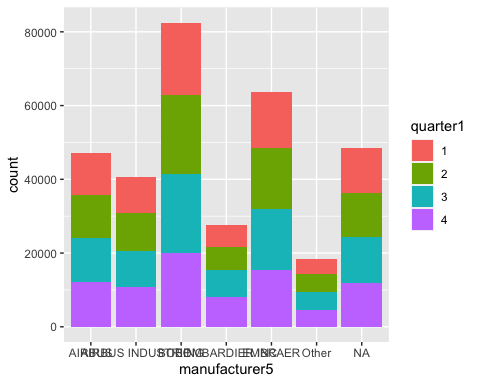
planes <-planes %>% mutate(manufacturer5 = case\_when(  
 manufacturer %in% top\_5\_manuf$manufacturer ~ manufacturer,  
 .default = "Other"  
))  
  
# joining the data with the plain info  
flights\_planes <- left\_join(x=flights,y=planes,by ='tailnum' )  
flights\_planes <- flights\_planes %>% filter(!is.na(dep\_time)) %>% arrange(month)  
  
# create date variables  
flights\_planes$month1 <- as.factor(flights\_planes$month)  
flights\_planes <-flights\_planes %>% mutate(quarter = case\_when(  
 month %in% c(1,2,3) ~ 1,   
 month %in% c(4,5,6) ~ 2,  
 month %in% c(7,8,9) ~ 3,  
 .default = 4  
))  
  
flights\_planes$quarter1 <- as.factor(flights\_planes$quarter)  
  
# Plotting distribution of fligts by manufacturer over time (by month)  
ggplot(flights\_planes, aes(x=month1, fill=manufacturer5)) + geom\_bar(just = 0.5)



# Plotting distribution of fligts by manufacturer over time (by quarter1)  
ggplot(flights\_planes, aes(x=quarter1, fill=manufacturer5)) + geom\_bar(just = 0.5)



# Plotting distribution of fligts by quarter over time (by manufacturer5)  
ggplot(flights\_planes, aes(x=manufacturer5, fill=quarter1)) + geom\_bar(just = 0.5)



## 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?   
There are planes that were made in 1956  
- How many airplanes that flew from New York City are included in the planes table?  
There are 3316 such planes.

# finding the oldest airplane  
planes1 <- planes %>% filter(!is.na(year))  
flights1 <- flights %>% filter(!is.na(dep\_time)) %>% select(tailnum)  
flights\_planes <- left\_join(x=flights1,y=planes1,by ='tailnum' )  
flights\_planes %>% arrange(year) %>% select(tailnum,year)

# A tibble: 328,521 × 2  
 tailnum year  
 <chr> <int>  
 1 N381AA 1956  
 2 N381AA 1956  
 3 N381AA 1956  
 4 N381AA 1956  
 5 N381AA 1956  
 6 N381AA 1956  
 7 N381AA 1956  
 8 N381AA 1956  
 9 N381AA 1956  
10 N381AA 1956  
# ℹ 328,511 more rows

# counting airplanes that flew from New York City are included in the planes table  
# Note: here we are dropping cancelled flights but we keep planes with no manufacturing date  
flights2 <- flights1 %>% group\_by(tailnum) %>% count()  
flights2 <- left\_join(x=flights2,y=planes,by ='tailnum' )  
flights2 %>% mutate(present = case\_when(is.na(type) ~ "No", .default = "Yes")) %>% group\_by(present) %>% count()

# A tibble: 2 × 2  
# Groups: present [2]  
 present n  
 <chr> <int>  
1 No 721  
2 Yes 3316

## 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.

IK: it is actually hard to come with a plot that would depict 4 dimensions nicely in this case, even with a grid unless one goes 3d.

# median arrival delay on a month-by-month basis in each airport?  
flights %>% filter(!is.na(dep\_time)) %>% group\_by(month,dest) %>% summarise(arr\_delay\_median = median(arr\_delay)) %>% arrange(month,arr\_delay\_median)

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

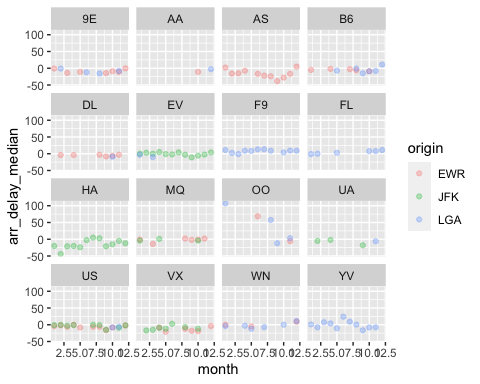
# A tibble: 1,112 × 3  
# Groups: month [12]  
 month dest arr\_delay\_median  
 <int> <chr> <dbl>  
 1 1 MYR -18  
 2 1 PSP -18  
 3 1 MTJ -16  
 4 1 OAK -14  
 5 1 BHM -11  
 6 1 BDL -10  
 7 1 HNL -10  
 8 1 MIA -9  
 9 1 PSE -9  
10 1 STT -9  
# ℹ 1,102 more rows

# For each airline, plot the median arrival delay for each month and origin airport.  
flights1 <- flights %>% filter(!is.na(dep\_time)) %>% group\_by(carrier,month,origin) %>% summarise(arr\_delay\_median = median(arr\_delay))

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

ggplot(flights1,aes(x =month , y = arr\_delay\_median, color = origin)) +geom\_point(alpha = 0.3) +facet\_wrap(~ carrier)

Warning: Removed 241 rows containing missing values (`geom\_point()`).



## 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.

# counting the flights by carier to SFO (dropping canceled ones)  
flights1 <- flights %>% filter(!is.na(dep\_time))  
flights1 <- left\_join(x=flights1,y=airlines,by ='carrier' )  
flights1 %>% filter(dest == "SFO") %>% group\_by(name) %>% summarise(count\_flight =n()) %>%   
arrange(desc(count\_flight))

# A tibble: 5 × 2  
 name count\_flight  
 <chr> <int>  
1 United Air Lines Inc. 6762  
2 Virgin America 2187  
3 Delta Air Lines Inc. 1850  
4 American Airlines Inc. 1402  
5 JetBlue Airways 1029

# the same by alternative way of doing it  
flights1 %>% filter(dest == "SFO") %>% count(name, sort = TRUE) %>%  
 mutate(prop = n/sum(n))

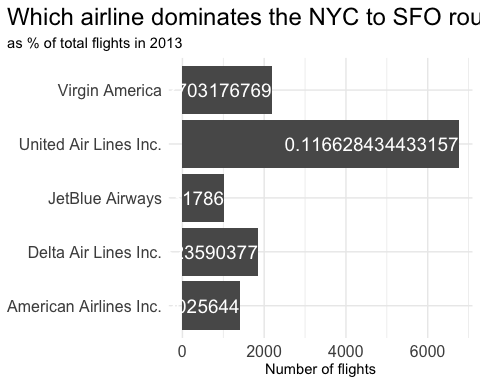
# A tibble: 5 × 3  
 name n prop  
 <chr> <int> <dbl>  
1 United Air Lines Inc. 6762 0.511   
2 Virgin America 2187 0.165   
3 Delta Air Lines Inc. 1850 0.140   
4 American Airlines Inc. 1402 0.106   
5 JetBlue Airways 1029 0.0778

# creating the dataset fly\_into\_sfo (dropping canceled flight)  
flights\_by\_carier <- flights1 %>% count(name, sort = TRUE)   
flights\_by\_carier <- rename(flights\_by\_carier, n\_total = n)  
  
fly\_into\_sfo <- flights1 %>% filter(dest == "SFO") %>% group\_by(name) %>% count(dest, sort = TRUE)   
   
fly\_into\_sfo <- left\_join(x=fly\_into\_sfo,y=flights\_by\_carier,by ='name' )  
  
fly\_into\_sfo <- fly\_into\_sfo %>% mutate(percent = n/n\_total ) %>% rename( count = n) %>% select(name,count, percent )  
fly\_into\_sfo

# A tibble: 5 × 3  
# Groups: name [5]  
 name count percent  
 <chr> <int> <dbl>  
1 United Air Lines Inc. 6762 0.117   
2 Virgin America 2187 0.426   
3 Delta Air Lines Inc. 1850 0.0387  
4 American Airlines Inc. 1402 0.0437  
5 JetBlue Airways 1029 0.0190

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.

IK: Steps are the following: 1. Filter for two origin airports 2. Filter for 5 carriers 3. Do bar plot of cancellations dataset by month with ggplot 4. Add facet\_wrap by the 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?
* 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:

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: No One
* Approximately how much time did you spend on this problem set: 4 hours
* What, if anything, gave you the most trouble: Git Hub but the instructions were super clear.

**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.