Stat 650 Project

YAN YIN

September 29, 2018

Stat 650 Section 1: 12:00pm - 1:40pm

# Answers:

## Part 1

### 1.Find the US Government website where Airline On-Time Performance Data can be downloaded. What website is this and how can you download the data? Download the data for the available months in 2018 for the Bay Area. Can you do this? If not, what can you download?

Answer:  
Airline On-Time Performance Data was downloaded on website. (<https://www.transtats.bts.gov/DL_SelectFields.asp>) When I get to the website, I select some variables and filter geography to “California”, filter year to “2018”, and filter Period to every month. And then, I clicked download button, zip files for every month are downloaded in my data folder in my R project. After that, I unzipped all the files and got data as .csv files for individual months. I checked all the files and found the records are from January to July. No records showed for August and after.

### 2.Once you have your data downloaded, develop your code for the first month of data. The last step will be to merge the data and perform an overall analysis for 2018. Extract the flights that departed from the Bay Area. Include all flights from department from San Francisco, Oakland, and San Jose. How many flight were there in January 2018.

Answer:  
After I read the data, I checked them with glimpse() function. I found there is a variable named X46 full of missing values which are not downloaded on my purpose. So I deleted the variable using the select() function. Then I filter all flights departed from San Francisco, Oakland, and San Jose.

sfoflight18Jan <- CAflightsJan18 %>%  
select(-X46) %>%  
filter(ORIGIN\_CITY\_NAME %in% c(“San Francisco, CA”, “Oakland, CA”, “San Jose, CA”))

After I cleaned data for January to July, I combined them together.

sfoflights18 <- bind\_rows(sfoflight18Jan, sfoflight18Feb, sfoflight18Mar, sfoflight18Apr, sfoflight18May, sfoflight18Jun, sfoflight18Jul)

I use glimpse() function to check sfoflight18Jan and sfoflights18. The results showed that the total flights of sfoflights18 are 163,056. And there are 22,606 flights in January 2018.

### 3.Compare the variables that are available in the full dataset with the variables in the nycflgihts13 data set. Make a table of the variables that are in both datasets, with a description of each variable. Hint: In RStudio see Help > RMarkdown Quick Reference > Tables.Report the intersection of the variables.

Answer:

|  |  |  |
| --- | --- | --- |
| ncyflights13 variables | sfoflights18 variabls | description |
| year | YEAR | year of flight |
| month | MONTH | month of flight |
| day | DAY\_OF\_MONTH | day of month |
| dep\_time | DEP\_TIME | Actual Departure Time (local time: hhmm) |
| sched\_dep\_time | CRS\_DEP\_TIME | schedualed/CRS Departure Time (local time: hhmm) |
| dep\_delay | DEP\_DELAY | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| arr\_time | ARR\_TIME | Actual Arrival Time (local time: hhmm) |
| sched\_arr\_time | CRS\_ARR\_TIME | schedualed/CRS Arrival Time (local time: hhmm) |
| arr\_delay | ARR\_DELAY | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| carrier | OP\_UNIQUE\_CARRIER | Unique Carrier Code. |
| flight | OP\_CARRIER\_FL\_NUM | Flight number |
| tailnum | TAIL\_NUM | Tail number |
| origin | ORIGIN | Origin Airport |
| dest | DEST | Destination Airport |
| air\_time | AIR\_TIME | Flight Time, in Minutes |
| distance | DISTANCE | Distance between airports (miles) |

### 4.What new variables do you now also have. Make a table of the variables that are in the new dataset, with a description of each variable. (Ok, this is kind of long. Make a table for 10 other variables you consider important.)

Answer:

|  |  |
| --- | --- |
| sfoflights18 variables | description |
| ORIGIN\_CITY\_NAME | Origin Airport, City Name |
| ORIGIN\_STATE\_ABR | Origin Airport, State Code |
| ORIGIN\_STATE\_NM | Origin Airport, State Name |
| DEST\_CITY\_NAME | Destination Airport, City Name |
| DEST\_STATE\_ABR | Destination Airport, State Code |
| DEST\_STATE\_NM | Destination Airport, State Name |
| DEP\_DELAY\_GROUP | Departure Delay intervals, every (15 minutes from <-15 to >180) |
| ARR\_DELAY\_GROUP | Arrival Delay intervals, every (15-minutes from <-15 to >180) |
| CANCELLED | Cancelled Flight Indicator (1=Yes) |
| CANCELLATION\_CODE | Specifies The Reason For Cancellation |

### 5.Answer Exercises 4.2, 4.3, 4.4 (you can only answer the second part of 4.4) on page 89 of the book, changing nycflights13 to sfoflights18. Answer all of the questions for the SF Bay Area in 2018.

**Question 4.2 What month had the highest proportion of cancelled fights? What month had the lowest? Interpret any seasonal patterns.**

Answer:  
March had the highest proportion of cancelled flights of 0.0225 while February had the lowest proportion of cancelled flights of 0.00715 .

I use the code to analyze the proportion of cancelled flights for each months.

sfoflights18 %>%  
group\_by(MONTH)%>%  
summarize(n=n(), canl\_num = sum(CANCELLED), canl\_prop = canl\_num / n) %>%  
arrange(canl\_prop)

And then I visualized the data using the code

ggplot(aes(x=MONTH, y= canl\_prop))+  
geom\_point()+geom\_line()

**Question 4.3 What plane (specifed by the tailnum variable) traveled the most times from Bay Area airports in 2018? Plot the number of trips per week over the year.**

Answer:

sfoflights18 %>%  
group\_by(TAIL\_NUM) %>%  
summarize(n=n()) %>%  
arrange(desc(n))

It shows that plane “N633VA” traveled 260 times which is the most times from Bay Area airports in 2018. I use the code below to plot the number of trips per week over the year:

sfoflights18 %>%  
filter(TAIL\_NUM == “N633VA”) %>%  
mutate(Week\_Num = week(FL\_DATE)) %>%  
group\_by(Week\_Num) %>%  
summarise(n=n()) %>%  
ggplot(aes(x= Week\_Num, y=n)) +  
geom\_point() +  
geom\_line() +  
geom\_smooth() +  
ggtitle(“N633VA”)

**Question 4.4 What is the oldest plane (specifed by the tailnum variable) that flew from Bay Area airports in 2018? How many airplanes that flew from Bay Area are included in the planes table?**

Answer:

planes %>%  
mutate(age = year) %>%  
select(age, tailnum) %>%  
right\_join(sfoflights18, by=c(“tailnum”=“TAIL\_NUM”)) %>%  
group\_by(age, tailnum) %>%  
summarize(n=n()) %>%  
arrange(age)

I just use plane data from nycflights13 package. And find the oldest plane that flew from Bay Area airports in 2018 are “N502UA” and “N505UA” of year 1989.

sfoflights18 %>%  
select(TAIL\_NUM, YEAR) %>%  
right\_join(planes, by=c(“TAIL\_NUM”=“tailnum”)) %>%  
select(TAIL\_NUM, YEAR) %>%  
filter(!is.na(YEAR)) %>%  
skim(TAIL\_NUM)

‘YEAR’ variable don’t have missing values in sfoflights18 data. After right-join to planes data, it shows that the complete value of TAIL\_NUM is 91693 which means 91693 records of sfoflights18 data can find matched data in planes. And result shows n\_unique of TAIL\_NUM is 1904. So in sfoflights18 data, there are 1904 TAIL\_NUM groups can find in planes data.

## Part 2

### 1.Find a data source for weather data. Download the temperature and other weather data for the Bay Area in 2018.

Answer:  
ASOS download from [Iowa Environmental Mesonet](https://mesonet.agron.iastate.edu/request/download.phtml)

select network: California ASOS  
Select station: [SFO] SAN FRANCISCO  
[OAK] OAKLAND  
[SJC] SAN JOSE INTL A  
Select From Available Data: All Available  
Specific Data Range: Start Date: 2018, January, 1  
End Date: 2018, October, 6  
Timezone of Observation Times: Corordinated Universal Time(UTC)  
Download Options: Data Format: Comma Delimited(NO DEBUG headers)  
include Latitude + Longitude? Yes

The selections are set as above and import the data into sfoweather18.

### 2.Try to find the same variables in the nycflights 13 weather data.

Answer:

|  |  |  |  |
| --- | --- | --- | --- |
| nycflights13 weather variables | description | sfoweather18 variables | description |
| origin | Weather station. Named origin to faciliate merging withflights() data | station | three or four character site identifier |
| temp | Temperature in F | tmpf | Air Temperature in Fahrenheit, typically @ 2 meters |
| dewp | dewpoint in F | dwpf | Dew Point Temperature in Fahrenheit, typically @ 2 meters |
| humid | Relative humidity | relh | Relative Humidity in % |
| wind\_dir | Wind direction (in degrees) | drct | Wind Direction in degrees from north |
| wind\_speed | wind speed in mph | sknt | Wind Speed in knots |
| wind\_gust | gust speed in mph | gust | Wind Gust in knots |
| precip | Precipitation, in inches | p01i | One hour precipitation in inches |
| pressure | Sea level pressure in millibars | mslp | Sea Level Pressure in millibar |
| visib | Visibility in miles | vsby | Visibility in miles |
| time\_hour | Date and hour of the recording as a POSIXct date | valid | timestamp of the observation |

### 3.Merge your new weather data with the same measures in the nycflights13 flights data from your sfoflights18 data.

Answer:  
Before I make merge, I changed the ‘valid’ variable into ‘dttm’ type and all the other ‘fct’ type variables into ‘character’ type.

sfoweather18Hour <- sfoweather18 %>%  
select(1:14) %>%  
mutate(wea\_date = date(valid), wea\_hour = hour(valid)) %>%  
group\_by(station, wea\_date, wea\_hour) %>%  
summarize( lon = mean(lon, na.rm =TRUE), lat = mean(lat, na.rm = TRUE), tmpf = mean(tmpf, na.rm = TRUE), dwpf = mean(dwpf, na.rm = TRUE), relh = mean(relh, na.rm = TRUE), drct = mean(drct, na.rm = TRUE), sknt = mean(sknt, na.rm = TRUE), p01i = mean(p01i, na.rm = TRUE), alti = mean(alti, na.rm = TRUE), mslp = mean(mslp, na.rm = TRUE), vsby = mean(vsby, na.rm = TRUE), gust = mean(gust, na.rm = TRUE) )

This new data set combined all the rows within 1 hour in sfoweather18.

And then join it with sfoflight18.

sfo\_join <- sfoflights18 %>%  
mutate(DEP\_hour = as.integer(substr(DEP\_TIME, 0, nchar(DEP\_TIME)-2))) %>%  
left\_join(sfoweather18Hour, by=c(“ORIGIN” = “station”, “DEP\_hour” = “wea\_hour”, “FL\_DATE” = “wea\_date”))

### 4.Questions to answer:Answer Exercises 4.6, and 4.7 for the SF Bay Area 2018 data.

**Question 4.6 Use the sfoflights18 and the weather table to answer the following questions: What is the distribution of temperature in July, 2018? 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?**

Answer:

sfoweather18 %>%  
filter(month(valid) == 7) %>%  
skim(tmpf, sknt)

sfoweather18 %>%  
filter(month(valid)==7) %>%  
ggplot(aes(x=tmpf))+  
geom\_bar()

Temperature(tmpf) in July, 2018 ranges from 52 to 93.9 with mean of 65.18 and sd of 7.43.

sfoweather18 %>%  
filter(month(valid)==7) %>%  
ggplot(aes(x=sknt))+  
geom\_bar()

Wind speed variable(sknt) in July, 2018 ranges from 0 to 27 with mean of 9.1 and sd of 4.92. After plot bar chart of sknt, we would identify outliers to be greater than 25.

Scatterplot was made to show the relationship between dewpoint and humidity, and the relationship between precipitation and visibility. There are no obvious linear relationship between them.

sfoweather18 %>%  
ggplot(aes(x=dwpf, y=relh))+  
geom\_point()

sfoweather18 %>%  
ggplot(aes(x=p01i, y=vsby))+  
geom\_point()

**Question 4.7 Use the sfoflights18 and the weather table to answer the following questions: On how many days was there precipitation in the Bay area in 2018? Were there differences in the mean visibility (visib) based on the day of the week and/or month of the year?**

Answer:

sfoweather18 %>%  
mutate(DATE=as.Date(valid, format=“%m%d%Y”)) %>%  
filter(p01i>0)%>%  
group\_by(DATE)%>%  
summarize(n=n())

The result shows a tibble of 43 rows. There are 43 days with precipitation in the Bay area in 2018.

sfoweather18 %>%  
mutate(Week\_Day = weekdays(valid)) %>%  
group\_by(Week\_Day) %>%  
summarize(visib\_mean = mean(vsby, na.rm=TRUE))

The mean of vsby(visibility) based on the day of the week is slightly different around 9.5 Week\_Day visib\_mean  
   
1 Friday 9.47  
2 Monday 9.47  
3 Saturday 9.70  
4 Sunday 9.68  
5 Thursday 9.32  
6 Tuesday 9.50  
7 Wednesday 9.46

sfoweather18 %>%  
mutate(Mon = month(valid))%>%  
group\_by(Mon)%>%  
summarize(visib\_mean = mean(vsby, na.rm=TRUE)) %>%  
arrange(visib\_mean)

Shows that January has the worst visibility of 8.33 and October has the best visibility of 9.95.

# Code and Explanation:

## Part 1

### 1.

None.

### 2.

library(tidyverse)  
library(nycflights13)  
library(skimr)  
library(lubridate)

Read .csv files

CAflightsJan18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_1.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsFeb18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_2.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsMar18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_3.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsApr18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_4.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsMay18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_5.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsJun18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_6.csv")

## Warning: Missing column names filled in: 'X46' [46]

CAflightsJul18 <-  
 read\_csv(file="./data/1028854783\_T\_ONTIME\_REPORTING\_7.csv")

## Warning: Missing column names filled in: 'X46' [46]

glimpse(CAflightsJan18)

## Observations: 107,006  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N477UA", "N16217", "N37255", "N424UA", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 368, 366, 361, 360, 358, 346, 341, 340, 33...  
## $ ORIGIN\_AIRPORT\_ID <int> 14747, 12892, 12889, 12892, 14771, 14831, ...  
## $ ORIGIN <chr> "SEA", "LAX", "LAS", "LAX", "SFO", "SJC", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Seattle, WA", "Los Angeles, CA", "Las Veg...  
## $ ORIGIN\_STATE\_ABR <chr> "WA", "CA", "NV", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "Washington", "California", "Nevada", "Cal...  
## $ ORIGIN\_WAC <int> 93, 91, 85, 91, 91, 91, 91, 38, 82, 91, 85...  
## $ DEST <chr> "SFO", "ORD", "SFO", "IAD", "LAS", "ORD", ...  
## $ DEST\_CITY\_NAME <chr> "San Francisco, CA", "Chicago, IL", "San F...  
## $ DEST\_STATE\_ABR <chr> "CA", "IL", "CA", "VA", "NV", "IL", "TX", ...  
## $ DEST\_STATE\_NM <chr> "California", "Illinois", "California", "V...  
## $ DEST\_WAC <int> 91, 41, 91, 38, 85, 41, 74, 91, 91, 85, 91...  
## $ CRS\_DEP\_TIME <chr> "0618", "0650", "1520", "0810", "1042", "1...  
## $ DEP\_TIME <chr> "0614", "0641", "1515", "0804", "1033", "1...  
## $ DEP\_DELAY <dbl> -4, -9, -5, -6, -9, -18, -7, -8, -4, -8, -...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -1, -1, -1, -1, -2, -1, -1, -1, -1, -1...  
## $ CRS\_ARR\_TIME <chr> "0831", "1250", "1703", "1602", "1221", "2...  
## $ ARR\_TIME <chr> "0813", "1242", "1641", "1548", "1208", "1...  
## $ ARR\_DELAY <dbl> -18, -8, -22, -14, -13, -23, -22, -31, -20...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -2, -1, -2, -1, -1, -2, -2, -2, -2, -2, -2...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 133, 240, 103, 292, 99, 247, 190, 372, 138...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 119, 241, 86, 284, 95, 242, 175, 349, 122,...  
## $ AIR\_TIME <dbl> 98, 212, 68, 256, 62, 218, 153, 332, 106, ...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 679, 1744, 414, 2288, 414, 1829, 1303, 241...  
## $ DISTANCE\_GROUP <int> 3, 7, 2, 10, 2, 8, 6, 10, 4, 1, 1, 11, 8, ...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsFeb18)

## Observations: 96,007  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_WEEK <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...  
## $ FL\_DATE <date> 2018-02-01, 2018-02-01, 2018-02-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N216FR", "N949FR", "N203FR", "N203FR", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 408, 670, 663, 668, 1747, 1746, 1923, 1924...  
## $ ORIGIN\_AIRPORT\_ID <int> 12892, 14771, 11292, 14771, 13204, 14679, ...  
## $ ORIGIN <chr> "LAX", "SFO", "DEN", "SFO", "MCO", "SAN", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Los Angeles, CA", "San Francisco, CA", "D...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CO", "CA", "FL", "CA", "TX", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "Colorado", "C...  
## $ ORIGIN\_WAC <int> 91, 91, 82, 91, 33, 91, 74, 91, 34, 91, 82...  
## $ DEST <chr> "DEN", "DEN", "SFO", "DEN", "SAN", "MCO", ...  
## $ DEST\_CITY\_NAME <chr> "Denver, CO", "Denver, CO", "San Francisco...  
## $ DEST\_STATE\_ABR <chr> "CO", "CO", "CA", "CO", "CA", "FL", "CA", ...  
## $ DEST\_STATE\_NM <chr> "Colorado", "Colorado", "California", "Col...  
## $ DEST\_WAC <int> 82, 82, 91, 82, 91, 33, 91, 74, 91, 82, 91...  
## $ CRS\_DEP\_TIME <chr> "0059", "0059", "1605", "1859", "1835", "2...  
## $ DEP\_TIME <chr> "0051", "0053", "1710", "1941", "1825", "2...  
## $ DEP\_DELAY <dbl> -8, -6, 65, 42, -10, -1, 0, -8, -5, -3, -6...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 65, 42, 0, 0, 0, 0, 0, 0, 0, 29, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -1, 4, 2, -1, -1, 0, -1, -1, -1, -1, 1...  
## $ CRS\_ARR\_TIME <chr> "0419", "0430", "1759", "2233", "2105", "0...  
## $ ARR\_TIME <chr> "0355", "0418", "1855", "2302", "2059", "0...  
## $ ARR\_DELAY <dbl> -24, -12, 56, 29, -6, -30, -19, -4, -26, 8...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 56, 29, 0, 0, 0, 0, 0, 8, 11, 19, 0,...  
## $ ARR\_DEL15 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -2, -1, 3, 1, -1, -2, -2, -1, -2, 0, 0, 1,...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 140, 151, 174, 154, 330, 295, 203, 175, 31...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 124, 145, 165, 141, 334, 266, 184, 179, 29...  
## $ AIR\_TIME <dbl> 108, 123, 137, 116, 317, 235, 161, 155, 27...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 862, 967, 967, 967, 2149, 2149, 1196, 1196...  
## $ DISTANCE\_GROUP <int> 4, 4, 4, 4, 9, 9, 5, 5, 8, 4, 4, 8, 4, 9, ...  
## $ CARRIER\_DELAY <dbl> NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, NA, ...  
## $ WEATHER\_DELAY <dbl> NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, NA, ...  
## $ NAS\_DELAY <dbl> NA, NA, 56, 0, NA, NA, NA, NA, NA, NA, NA,...  
## $ SECURITY\_DELAY <dbl> NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, NA, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, 0, 29, NA, NA, NA, NA, NA, NA, NA,...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsMar18)

## Observations: 110,768  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_WEEK <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...  
## $ FL\_DATE <date> 2018-03-01, 2018-03-01, 2018-03-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N206FR", "N213FR", "N213FR", "N221FR", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 409, 265, 266, 1631, 1630, 1479, 404, 371,...  
## $ ORIGIN\_AIRPORT\_ID <int> 11292, 11292, 14908, 13204, 12892, 10397, ...  
## $ ORIGIN <chr> "DEN", "DEN", "SNA", "MCO", "LAX", "ATL", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Denver, CO", "Denver, CO", "Santa Ana, CA...  
## $ ORIGIN\_STATE\_ABR <chr> "CO", "CO", "CA", "FL", "CA", "GA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "Colorado", "Colorado", "California", "Flo...  
## $ ORIGIN\_WAC <int> 82, 82, 91, 33, 91, 34, 91, 82, 91, 82, 91...  
## $ DEST <chr> "LAX", "SNA", "DEN", "LAX", "MCO", "LAX", ...  
## $ DEST\_CITY\_NAME <chr> "Los Angeles, CA", "Santa Ana, CA", "Denve...  
## $ DEST\_STATE\_ABR <chr> "CA", "CA", "CO", "CA", "FL", "CA", "CO", ...  
## $ DEST\_STATE\_NM <chr> "California", "California", "Colorado", "C...  
## $ DEST\_WAC <int> 91, 91, 82, 91, 33, 91, 82, 91, 82, 91, 34...  
## $ CRS\_DEP\_TIME <chr> "2215", "1705", "1920", "1715", "2055", "0...  
## $ DEP\_TIME <chr> "2248", "1711", "1920", "1735", "2207", "0...  
## $ DEP\_DELAY <dbl> 33, 6, 0, 20, 72, -1, 10, 80, 80, 75, 63, ...  
## $ DEP\_DELAY\_NEW <dbl> 33, 6, 0, 20, 72, 0, 10, 80, 80, 75, 63, 0...  
## $ DEP\_DEL15 <dbl> 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, ...  
## $ DEP\_DELAY\_GROUP <int> 2, 0, 0, 1, 4, -1, 0, 5, 5, 5, 4, -1, 0, 3...  
## $ CRS\_ARR\_TIME <chr> "2350", "1830", "2239", "2000", "0438", "0...  
## $ ARR\_TIME <chr> "0017", "1837", "2217", "2027", "0522", "1...  
## $ ARR\_DELAY <dbl> 27, 7, -22, 27, 44, 19, 14, 96, 55, 70, 49...  
## $ ARR\_DELAY\_NEW <dbl> 27, 7, 0, 27, 44, 19, 14, 96, 55, 70, 49, ...  
## $ ARR\_DEL15 <dbl> 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, ...  
## $ ARR\_DELAY\_GROUP <int> 1, 0, -2, 1, 2, 1, 0, 6, 3, 4, 3, 0, -2, 3...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 155, 145, 139, 345, 283, 316, 144, 143, 14...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 149, 146, 117, 352, 255, 336, 148, 159, 11...  
## $ AIR\_TIME <dbl> 128, 132, 105, 329, 236, 298, 110, 142, 99...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 862, 846, 846, 2218, 2218, 1947, 862, 819,...  
## $ DISTANCE\_GROUP <int> 4, 4, 4, 9, 9, 8, 4, 4, 4, 4, 8, 9, 9, 5, ...  
## $ CARRIER\_DELAY <dbl> 0, NA, NA, 20, 27, 0, NA, 10, 0, 70, 0, NA...  
## $ WEATHER\_DELAY <dbl> 0, NA, NA, 0, 0, 0, NA, 0, 0, 0, 0, NA, NA...  
## $ NAS\_DELAY <dbl> 27, NA, NA, 7, 0, 19, NA, 16, 0, 0, 0, NA,...  
## $ SECURITY\_DELAY <dbl> 0, NA, NA, 0, 0, 0, NA, 0, 0, 0, 0, NA, NA...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> 0, NA, NA, 0, 17, 0, NA, 70, 55, 0, 49, NA...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsApr18)

## Observations: 110,435  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_WEEK <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...  
## $ FL\_DATE <date> 2018-04-01, 2018-04-01, 2018-04-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N201FR", "N202FR", "N202FR", "N203FR", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 409, 1749, 1750, 1479, 404, 1701, 1702, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 11292, 11109, 14679, 10397, 12892, 10423, ...  
## $ ORIGIN <chr> "DEN", "COS", "SAN", "ATL", "LAX", "AUS", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Denver, CO", "Colorado Springs, CO", "San...  
## $ ORIGIN\_STATE\_ABR <chr> "CO", "CO", "CA", "GA", "CA", "TX", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "Colorado", "Colorado", "California", "Geo...  
## $ ORIGIN\_WAC <int> 82, 82, 91, 34, 91, 74, 91, 82, 91, 74, 91...  
## $ DEST <chr> "LAX", "SAN", "COS", "LAX", "DEN", "SAN", ...  
## $ DEST\_CITY\_NAME <chr> "Los Angeles, CA", "San Diego, CA", "Color...  
## $ DEST\_STATE\_ABR <chr> "CA", "CA", "CO", "CA", "CO", "CA", "TX", ...  
## $ DEST\_STATE\_NM <chr> "California", "California", "Colorado", "C...  
## $ DEST\_WAC <int> 91, 91, 82, 91, 82, 91, 74, 91, 82, 91, 74...  
## $ CRS\_DEP\_TIME <chr> "2215", "0935", "1150", "0730", "1036", "1...  
## $ DEP\_TIME <chr> "2257", "0927", "1141", "0728", "1031", "1...  
## $ DEP\_DELAY <dbl> 42, -8, -9, -2, -5, -2, -10, 5, -1, -6, -1...  
## $ DEP\_DELAY\_NEW <dbl> 42, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 1, 0,...  
## $ DEP\_DEL15 <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> 2, -1, -1, -1, -1, -1, -1, 0, -1, -1, -1, ...  
## $ CRS\_ARR\_TIME <chr> "2350", "1100", "1507", "0946", "1400", "1...  
## $ ARR\_TIME <chr> "0015", "1042", "1449", "0934", "1345", "1...  
## $ ARR\_DELAY <dbl> 25, -18, -18, -12, -15, -17, -16, 0, 5, -1...  
## $ ARR\_DELAY\_NEW <dbl> 25, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0,...  
## $ ARR\_DEL15 <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> 1, -2, -2, -1, -1, -2, -2, 0, 0, -1, -2, 0...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 155, 145, 137, 316, 144, 190, 160, 145, 13...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 138, 135, 128, 306, 134, 175, 154, 140, 14...  
## $ AIR\_TIME <dbl> 113, 121, 107, 273, 110, 158, 131, 124, 11...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 862, 816, 816, 1947, 862, 1164, 1164, 846,...  
## $ DISTANCE\_GROUP <int> 4, 4, 4, 8, 4, 5, 5, 4, 4, 5, 5, 4, 8, 4, ...  
## $ CARRIER\_DELAY <dbl> 14, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ WEATHER\_DELAY <dbl> 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ NAS\_DELAY <dbl> 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ SECURITY\_DELAY <dbl> 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> 11, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsMay18)

## Observations: 114,850  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_WEEK <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ FL\_DATE <date> 2018-05-01, 2018-05-01, 2018-05-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "AS", "AS", "AS", "AS", "AS", "AS", "AS", ...  
## $ TAIL\_NUM <chr> "N520AS", "N531AS", "N238AK", "N556AS", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 5, 6, 74, 86, 113, 150, 160, 165, 174, 181...  
## $ ORIGIN\_AIRPORT\_ID <int> 11278, 12892, 14747, 13891, 14057, 10299, ...  
## $ ORIGIN <chr> "DCA", "LAX", "SEA", "ONT", "PDX", "ANC", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Washington, DC", "Los Angeles, CA", "Seat...  
## $ ORIGIN\_STATE\_ABR <chr> "VA", "CA", "WA", "CA", "OR", "AK", "HI", ...  
## $ ORIGIN\_STATE\_NM <chr> "Virginia", "California", "Washington", "C...  
## $ ORIGIN\_WAC <int> 38, 91, 93, 91, 92, 1, 2, 91, 93, 91, 93, ...  
## $ DEST <chr> "LAX", "DCA", "SAN", "PDX", "ONT", "LAX", ...  
## $ DEST\_CITY\_NAME <chr> "Los Angeles, CA", "Washington, DC", "San ...  
## $ DEST\_STATE\_ABR <chr> "CA", "VA", "CA", "OR", "CA", "CA", "CA", ...  
## $ DEST\_STATE\_NM <chr> "California", "Virginia", "California", "O...  
## $ DEST\_WAC <int> 91, 38, 91, 92, 91, 91, 91, 93, 91, 93, 91...  
## $ CRS\_DEP\_TIME <chr> "0900", "1245", "1530", "2025", "1700", "2...  
## $ DEP\_TIME <chr> "0855", "1240", "1524", "2009", "1718", "2...  
## $ DEP\_DELAY <dbl> -5, -5, -6, -16, 18, -6, -13, -4, -10, -7,...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 18, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -1, -1, -2, 1, -1, -1, -1, -1, -1, -1,...  
## $ CRS\_ARR\_TIME <chr> "1203", "2055", "1810", "2244", "1915", "0...  
## $ ARR\_TIME <chr> "1211", "2028", "1757", "2228", "1921", "0...  
## $ ARR\_DELAY <dbl> 8, -27, -13, -16, 6, -35, -17, 17, -20, 8,...  
## $ ARR\_DELAY\_NEW <dbl> 8, 0, 0, 0, 6, 0, 0, 17, 0, 8, 0, 7, 0, 9,...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> 0, -2, -1, -2, 0, -2, -2, 1, -2, 0, -1, 0,...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 363, 310, 160, 139, 135, 317, 304, 161, 15...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 376, 288, 153, 139, 123, 288, 300, 182, 14...  
## $ AIR\_TIME <dbl> 328, 266, 129, 124, 107, 262, 285, 162, 13...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 2311, 2311, 1050, 838, 838, 2345, 2466, 93...  
## $ DISTANCE\_GROUP <int> 10, 10, 5, 4, 4, 10, 10, 4, 5, 3, 5, 11, 1...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA,...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 17, NA, NA, NA...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA,...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsJun18)

## Observations: 116,637  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ DAY\_OF\_MONTH <int> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20...  
## $ DAY\_OF\_WEEK <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ FL\_DATE <date> 2018-06-20, 2018-06-20, 2018-06-20, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "AS", "AS", "AS", "AS", "AS", "AS", "AS", ...  
## $ TAIL\_NUM <chr> "N569AS", "N517AS", "N581AS", "N464AS", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 881, 892, 895, 949, 982, 1001, 1002, 1005,...  
## $ ORIGIN\_AIRPORT\_ID <int> 12892, 12173, 14679, 14679, 14747, 14771, ...  
## $ ORIGIN <chr> "LAX", "HNL", "SAN", "SAN", "SEA", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Los Angeles, CA", "Honolulu, HI", "San Di...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "HI", "CA", "CA", "WA", "CA", "VA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "Hawaii", "California", "Cal...  
## $ ORIGIN\_WAC <int> 91, 2, 91, 91, 93, 91, 38, 91, 38, 91, 22,...  
## $ DEST <chr> "SEA", "SAN", "HNL", "SEA", "OAK", "DCA", ...  
## $ DEST\_CITY\_NAME <chr> "Seattle, WA", "San Diego, CA", "Honolulu,...  
## $ DEST\_STATE\_ABR <chr> "WA", "CA", "HI", "WA", "CA", "VA", "CA", ...  
## $ DEST\_STATE\_NM <chr> "Washington", "California", "Hawaii", "Was...  
## $ DEST\_WAC <int> 93, 91, 2, 93, 91, 38, 91, 38, 91, 93, 91,...  
## $ CRS\_DEP\_TIME <chr> "1151", "1448", "0720", "1938", "2120", "0...  
## $ DEP\_TIME <chr> "1157", "1521", "0718", "1943", "2119", "0...  
## $ DEP\_DELAY <dbl> 6, 33, -2, 5, -1, 12, 7, 2, -3, 27, -3, 4,...  
## $ DEP\_DELAY\_NEW <dbl> 6, 33, 0, 5, 0, 12, 7, 2, 0, 27, 0, 4, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> 0, 2, -1, 0, -1, 0, 0, 0, -1, 1, -1, 0, -1...  
## $ CRS\_ARR\_TIME <chr> "1439", "2324", "1029", "2229", "2327", "1...  
## $ ARR\_TIME <chr> "1433", "0014", "0959", "2232", "2319", "1...  
## $ ARR\_DELAY <dbl> -6, 50, -30, 3, -8, -5, 98, 24, -25, 38, -...  
## $ ARR\_DELAY\_NEW <dbl> 0, 50, 0, 3, 0, 0, 98, 24, 0, 38, 0, 0, 0,...  
## $ ARR\_DEL15 <dbl> 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, 3, -2, 0, -1, -1, 6, 1, -2, 2, -1, -2,...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 168, 336, 369, 171, 127, 315, 361, 319, 35...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 156, 353, 341, 169, 120, 298, 452, 341, 33...  
## $ AIR\_TIME <dbl> 131, 319, 318, 144, 100, 276, 341, 310, 29...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 954, 2614, 2614, 1050, 672, 2442, 2442, 23...  
## $ DISTANCE\_GROUP <int> 4, 11, 11, 5, 3, 10, 10, 10, 10, 4, 11, 11...  
## $ CARRIER\_DELAY <dbl> NA, 7, NA, NA, NA, NA, 7, 0, NA, 10, NA, N...  
## $ WEATHER\_DELAY <dbl> NA, 0, NA, NA, NA, NA, 0, 0, NA, 0, NA, NA...  
## $ NAS\_DELAY <dbl> NA, 17, NA, NA, NA, NA, 91, 24, NA, 11, NA...  
## $ SECURITY\_DELAY <dbl> NA, 0, NA, NA, NA, NA, 0, 0, NA, 0, NA, NA...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, 26, NA, NA, NA, NA, 0, 0, NA, 17, NA, ...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

glimpse(CAflightsJul18)

## Observations: 121,402  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ MONTH <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...  
## $ DAY\_OF\_MONTH <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...  
## $ DAY\_OF\_WEEK <int> 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, ...  
## $ FL\_DATE <date> 2018-07-01, 2018-07-02, 2018-07-03, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "B6", "B6", "B6", "B6", "B6", "B6", "B6", ...  
## $ TAIL\_NUM <chr> "N509JB", "N516JB", "N712JB", "N705JB", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14...  
## $ ORIGIN\_AIRPORT\_ID <int> 12954, 12954, 12954, 12954, 12954, 12954, ...  
## $ ORIGIN <chr> "LGB", "LGB", "LGB", "LGB", "LGB", "LGB", ...  
## $ ORIGIN\_CITY\_NAME <chr> "Long Beach, CA", "Long Beach, CA", "Long ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "JFK", "JFK", "JFK", "JFK", "JFK", "JFK", ...  
## $ DEST\_CITY\_NAME <chr> "New York, NY", "New York, NY", "New York,...  
## $ DEST\_STATE\_ABR <chr> "NY", "NY", "NY", "NY", "NY", "NY", "NY", ...  
## $ DEST\_STATE\_NM <chr> "New York", "New York", "New York", "New Y...  
## $ DEST\_WAC <int> 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22...  
## $ CRS\_DEP\_TIME <chr> "2046", "2046", "2046", "2046", "2046", "2...  
## $ DEP\_TIME <chr> "2042", "2036", "2040", "2045", "2045", "2...  
## $ DEP\_DELAY <dbl> -4, -10, -6, -1, -1, 29, 22, 27, 0, 11, 6,...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 29, 22, 27, 0, 11, 6, 0, 0,...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -1, -1, -1, -1, 1, 1, 1, 0, 0, 0, -1, ...  
## $ CRS\_ARR\_TIME <chr> "0505", "0505", "0505", "0505", "0505", "0...  
## $ ARR\_TIME <chr> "0502", "0502", "0508", "0506", "0517", "0...  
## $ ARR\_DELAY <dbl> -3, -3, 3, 1, 12, 35, 52, 61, 14, 8, 4, 10...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 3, 1, 12, 35, 52, 61, 14, 8, 4, 10, ...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -1, 0, 0, 0, 2, 3, 4, 0, 0, 0, 0, -1, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 319, 319, 319, 319, 319, 319, 319, 319, 31...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 320, 326, 328, 321, 332, 325, 349, 353, 33...  
## $ AIR\_TIME <dbl> 301, 309, 304, 302, 305, 305, 325, 328, 31...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 2465, 2465, 2465, 2465, 2465, 2465, 2465, ...  
## $ DISTANCE\_GROUP <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, 19, 22, 19, NA, NA, NA...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, 0, 0, 0, NA, NA, NA, N...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, 6, 30, 34, NA, NA, NA,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, 0, 0, 0, NA, NA, NA, N...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, 10, 0, 8, NA, NA, NA, ...  
## $ X46 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

sfoflight18Jan <- CAflightsJan18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Jan)

## Observations: 22,606  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...

sfoflight18Feb <- CAflightsFeb18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Feb)

## Observations: 20,142  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, ...  
## $ DAY\_OF\_WEEK <int> 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-02-01, 2018-02-01, 2018-02-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N949FR", "N203FR", "N941FR", "N953FR", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 670, 668, 1740, 596, 1532, 1124, 658, 1096...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14771, 14771, 14831, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SFO", "SFO", "SJC", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Francisco, CA", ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "DEN", "DEN", "MCO", "DEN", "ATL", "LAS", ...  
## $ DEST\_CITY\_NAME <chr> "Denver, CO", "Denver, CO", "Orlando, FL",...  
## $ DEST\_STATE\_ABR <chr> "CO", "CO", "FL", "CO", "GA", "NV", "CO", ...  
## $ DEST\_STATE\_NM <chr> "Colorado", "Colorado", "Florida", "Colora...  
## $ DEST\_WAC <int> 82, 82, 33, 82, 34, 85, 82, 85, 82, 82, 33...  
## $ CRS\_DEP\_TIME <chr> "0059", "1859", "2225", "1200", "1031", "2...  
## $ DEP\_TIME <chr> "0053", "1941", "2245", "1208", "1025", "2...  
## $ DEP\_DELAY <dbl> -6, 42, 20, 8, -6, -4, -5, -9, 38, -5, 0, ...  
## $ DEP\_DELAY\_NEW <dbl> 0, 42, 20, 8, 0, 0, 0, 0, 38, 0, 0, 0, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, 2, 1, 0, -1, -1, -1, -1, 2, -1, 0, -1,...  
## $ CRS\_ARR\_TIME <chr> "0430", "2233", "0651", "1539", "1816", "2...  
## $ ARR\_TIME <chr> "0418", "2302", "0643", "1530", "1748", "2...  
## $ ARR\_DELAY <dbl> -12, 29, -8, -9, -28, -10, -13, -12, 34, -...  
## $ ARR\_DELAY\_NEW <dbl> 0, 29, 0, 0, 0, 0, 0, 0, 34, 0, 0, 0, 0, 0...  
## $ ARR\_DEL15 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, 1, -1, -1, -2, -1, -1, -1, 2, -2, -1, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 151, 154, 326, 159, 285, 98, 165, 88, 151,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 145, 141, 298, 142, 263, 92, 157, 85, 147,...  
## $ AIR\_TIME <dbl> 123, 116, 261, 122, 242, 67, 113, 63, 117,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 967, 967, 2446, 948, 2139, 414, 967, 386, ...  
## $ DISTANCE\_GROUP <int> 4, 4, 10, 4, 9, 2, 4, 2, 4, 4, 10, 9, 2, 4...  
## $ CARRIER\_DELAY <dbl> NA, 0, NA, NA, NA, NA, NA, NA, 10, NA, NA,...  
## $ WEATHER\_DELAY <dbl> NA, 0, NA, NA, NA, NA, NA, NA, 0, NA, NA, ...  
## $ NAS\_DELAY <dbl> NA, 0, NA, NA, NA, NA, NA, NA, 0, NA, NA, ...  
## $ SECURITY\_DELAY <dbl> NA, 0, NA, NA, NA, NA, NA, NA, 0, NA, NA, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, 29, NA, NA, NA, NA, NA, NA, 24, NA, NA...

sfoflight18Mar <- CAflightsMar18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Mar)

## Observations: 23,055  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, ...  
## $ DAY\_OF\_WEEK <int> 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-03-01, 2018-03-01, 2018-03-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N939FR", NA, "N941FR", "N953FR", "N236FR"...  
## $ OP\_CARRIER\_FL\_NUM <int> 596, 668, 670, 1740, 658, 1096, 1532, 1124...  
## $ ORIGIN\_AIRPORT\_ID <int> 14831, 14771, 14771, 14771, 14771, 14831, ...  
## $ ORIGIN <chr> "SJC", "SFO", "SFO", "SFO", "SFO", "SJC", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Jose, CA", "San Francisco, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "DEN", "DEN", "DEN", "MCO", "DEN", "LAS", ...  
## $ DEST\_CITY\_NAME <chr> "Denver, CO", "Denver, CO", "Denver, CO", ...  
## $ DEST\_STATE\_ABR <chr> "CO", "CO", "CO", "FL", "CO", "NV", "GA", ...  
## $ DEST\_STATE\_NM <chr> "Colorado", "Colorado", "Colorado", "Flori...  
## $ DEST\_WAC <int> 82, 82, 82, 33, 82, 85, 34, 85, 82, 82, 33...  
## $ CRS\_DEP\_TIME <chr> "1154", "1904", "0055", "2245", "1148", "1...  
## $ DEP\_TIME <chr> "1159", NA, "0134", "0241", "1223", "1740"...  
## $ DEP\_DELAY <dbl> 5, NA, 39, 236, 35, 110, 16, 76, 34, 11, -...  
## $ DEP\_DELAY\_NEW <dbl> 5, NA, 39, 236, 35, 110, 16, 76, 34, 11, 0...  
## $ DEP\_DEL15 <dbl> 0, NA, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1,...  
## $ DEP\_DELAY\_GROUP <int> 0, NA, 2, 12, 2, 7, 1, 5, 2, 0, -1, 7, -2,...  
## $ CRS\_ARR\_TIME <chr> "1529", "2246", "0436", "0703", "1529", "1...  
## $ ARR\_TIME <chr> "1513", NA, "0451", "1018", "1626", "1859"...  
## $ ARR\_DELAY <dbl> -16, NA, 15, 195, 57, 108, -4, 63, -1, -11...  
## $ ARR\_DELAY\_NEW <dbl> 0, NA, 15, 195, 57, 108, 0, 63, 0, 0, 0, 9...  
## $ ARR\_DEL15 <dbl> 0, NA, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...  
## $ ARR\_DELAY\_GROUP <int> -2, NA, 1, 12, 3, 7, -1, 4, -1, -1, -1, 6,...  
## $ CANCELLED <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, "B", NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ CRS\_ELAPSED\_TIME <dbl> 155, 162, 161, 318, 161, 81, 287, 98, 161,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 134, NA, 137, 277, 183, 79, 267, 85, 126, ...  
## $ AIR\_TIME <dbl> 114, NA, 112, 256, 112, 61, 230, 62, 101, ...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 948, 967, 967, 2446, 967, 386, 2139, 414, ...  
## $ DISTANCE\_GROUP <int> 4, 4, 4, 10, 4, 2, 9, 2, 4, 4, 10, 9, 4, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, 0, 0, 17, 0, NA, 0, NA, NA, NA, 0,...  
## $ WEATHER\_DELAY <dbl> NA, NA, 0, 0, 0, 0, NA, 0, NA, NA, NA, 0, ...  
## $ NAS\_DELAY <dbl> NA, NA, 0, 0, 25, 0, NA, 0, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, 0, 0, 0, 0, NA, 0, NA, NA, NA, 0, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, 15, 195, 15, 108, NA, 63, NA, NA, ...

sfoflight18Apr <- CAflightsApr18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Apr)

## Observations: 23,090  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, ...  
## $ DAY\_OF\_WEEK <int> 7, 7, 7, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, ...  
## $ FL\_DATE <date> 2018-04-01, 2018-04-01, 2018-04-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "F9", "F9", "F9", "F9", "F9", "F9", "F9", ...  
## $ TAIL\_NUM <chr> "N923FR", "N938FR", "N941FR", "N211FR", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 670, 668, 1740, 658, 598, 1096, 1124, 1532...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14771, 14771, 14771, 14831, 14831, ...  
## $ ORIGIN <chr> "SFO", "SFO", "SFO", "SFO", "SJC", "SJC", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Francisco, CA", ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "DEN", "DEN", "MCO", "DEN", "DEN", "LAS", ...  
## $ DEST\_CITY\_NAME <chr> "Denver, CO", "Denver, CO", "Orlando, FL",...  
## $ DEST\_STATE\_ABR <chr> "CO", "CO", "FL", "CO", "CO", "NV", "NV", ...  
## $ DEST\_STATE\_NM <chr> "Colorado", "Colorado", "Florida", "Colora...  
## $ DEST\_WAC <int> 82, 82, 33, 82, 82, 85, 85, 34, 82, 33, 82...  
## $ CRS\_DEP\_TIME <chr> "0055", "1858", "2235", "1148", "1854", "1...  
## $ DEP\_TIME <chr> "0107", "1851", "2220", "1145", "1847", "1...  
## $ DEP\_DELAY <dbl> 12, -7, -15, -3, -7, -2, 14, 18, -14, -10,...  
## $ DEP\_DELAY\_NEW <dbl> 12, 0, 0, 0, 0, 0, 14, 18, 0, 0, 7, 31, 18...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, ...  
## $ DEP\_DELAY\_GROUP <int> 0, -1, -1, -1, -1, -1, 0, 1, -1, -1, 0, 2,...  
## $ CRS\_ARR\_TIME <chr> "0426", "2240", "0653", "1529", "2229", "1...  
## $ ARR\_TIME <chr> "0428", "2236", "0635", "1511", "2224", "1...  
## $ ARR\_DELAY <dbl> 2, -4, -18, -18, -5, -7, 12, -7, -22, -34,...  
## $ ARR\_DELAY\_NEW <dbl> 2, 0, 0, 0, 0, 0, 12, 0, 0, 0, 0, 49, 0, 0...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> 0, -1, -2, -2, -1, -1, 0, -1, -2, -2, -1, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 151, 162, 318, 161, 155, 81, 98, 287, 151,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 141, 165, 315, 146, 157, 76, 96, 262, 143,...  
## $ AIR\_TIME <dbl> 114, 121, 268, 117, 114, 58, 58, 234, 116,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 967, 967, 2446, 967, 948, 386, 414, 2139, ...  
## $ DISTANCE\_GROUP <int> 4, 4, 10, 4, 4, 2, 2, 9, 4, 10, 4, 10, 4, ...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

sfoflight18May <- CAflightsMay18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18May)

## Observations: 24,165  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...  
## $ DAY\_OF\_MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_WEEK <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ FL\_DATE <date> 2018-05-01, 2018-05-01, 2018-05-01, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "AS", "AS", "AS", "AS", "AS", "AS", "AS", ...  
## $ TAIL\_NUM <chr> "N506AS", "N527AS", "N236AK", "N468AS", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 303, 309, 312, 319, 321, 327, 337, 353, 36...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14831, 14831, 14831, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SJC", "SJC", "SJC", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "SEA", "PDX", "SEA", "SEA", "SEA", "SEA", ...  
## $ DEST\_CITY\_NAME <chr> "Seattle, WA", "Portland, OR", "Seattle, W...  
## $ DEST\_STATE\_ABR <chr> "WA", "OR", "WA", "WA", "WA", "WA", "OR", ...  
## $ DEST\_STATE\_NM <chr> "Washington", "Oregon", "Washington", "Was...  
## $ DEST\_WAC <int> 93, 92, 93, 93, 93, 93, 92, 93, 92, 93, 91...  
## $ CRS\_DEP\_TIME <chr> "2137", "1700", "0645", "2030", "0630", "0...  
## $ DEP\_TIME <chr> "2130", "1645", "0642", "2027", "0627", "0...  
## $ DEP\_DELAY <dbl> -7, -15, -3, -3, -3, -8, -9, -14, -14, -3,...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1...  
## $ CRS\_ARR\_TIME <chr> "2335", "1841", "0900", "2241", "0846", "1...  
## $ ARR\_TIME <chr> "2338", "1846", "0921", "2244", "0859", "1...  
## $ ARR\_DELAY <dbl> 3, 5, 21, 3, 13, 20, 4, -2, -10, 12, -12, ...  
## $ ARR\_DELAY\_NEW <dbl> 3, 5, 21, 3, 13, 20, 4, 0, 0, 12, 0, 9, 0,...  
## $ ARR\_DEL15 <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> 0, 0, 1, 0, 0, 1, 0, -1, -1, 0, -1, 0, -1,...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 118, 101, 135, 131, 136, 128, 101, 121, 10...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 128, 121, 159, 137, 152, 156, 114, 133, 10...  
## $ AIR\_TIME <dbl> 103, 103, 120, 113, 129, 126, 92, 111, 89,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 679, 569, 679, 696, 696, 696, 550, 672, 55...  
## $ DISTANCE\_GROUP <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, ...  
## $ CARRIER\_DELAY <dbl> NA, NA, 0, NA, NA, 0, NA, NA, NA, NA, NA, ...  
## $ WEATHER\_DELAY <dbl> NA, NA, 0, NA, NA, 0, NA, NA, NA, NA, NA, ...  
## $ NAS\_DELAY <dbl> NA, NA, 21, NA, NA, 20, NA, NA, NA, NA, NA...  
## $ SECURITY\_DELAY <dbl> NA, NA, 0, NA, NA, 0, NA, NA, NA, NA, NA, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, 0, NA, NA, 0, NA, NA, NA, NA, NA, ...

sfoflight18Jun <- CAflightsJun18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Jun)

## Observations: 24,526  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ MONTH <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ DAY\_OF\_MONTH <int> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20...  
## $ DAY\_OF\_WEEK <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ FL\_DATE <date> 2018-06-20, 2018-06-20, 2018-06-20, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "AS", "AS", "AS", "AS", "AS", "AS", "AS", ...  
## $ TAIL\_NUM <chr> "N926VA", "N627VA", "N640VA", "N847VA", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 1001, 1012, 1022, 1024, 1026, 1034, 1047, ...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14771, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Francisco, CA", ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "DCA", "JFK", "JFK", "JFK", "JFK", "JFK", ...  
## $ DEST\_CITY\_NAME <chr> "Washington, DC", "New York, NY", "New Yor...  
## $ DEST\_STATE\_ABR <chr> "VA", "NY", "NY", "NY", "NY", "NY", "HI", ...  
## $ DEST\_STATE\_NM <chr> "Virginia", "New York", "New York", "New Y...  
## $ DEST\_WAC <int> 38, 22, 22, 22, 22, 22, 2, 2, 36, 38, 38, ...  
## $ CRS\_DEP\_TIME <chr> "0800", "0700", "0909", "1345", "1554", "2...  
## $ DEP\_TIME <chr> "0812", "0704", "0906", "1340", "1609", "2...  
## $ DEP\_DELAY <dbl> 12, 4, -3, -5, 15, 14, 5, 44, 2, 0, 73, -2...  
## $ DEP\_DELAY\_NEW <dbl> 12, 4, 0, 0, 15, 14, 5, 44, 2, 0, 73, 0, 3...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, ...  
## $ DEP\_DELAY\_GROUP <int> 0, 0, -1, -1, 1, 0, 0, 2, 0, 0, 4, -1, 2, ...  
## $ CRS\_ARR\_TIME <chr> "1615", "1544", "1753", "2228", "0040", "0...  
## $ ARR\_TIME <chr> "1610", "1521", "1748", "2220", "0038", "0...  
## $ ARR\_DELAY <dbl> -5, -23, -5, -8, -2, -3, -3, 39, -23, 10, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 39, 0, 10, 80, 1, 29,...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, 2, -2, 0, 5, 0...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 315, 344, 344, 343, 346, 348, 326, 332, 31...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 298, 317, 342, 340, 329, 331, 318, 327, 28...  
## $ AIR\_TIME <dbl> 276, 299, 311, 299, 304, 293, 279, 294, 26...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 2442, 2586, 2586, 2586, 2586, 2586, 2338, ...  
## $ DISTANCE\_GROUP <int> 10, 11, 11, 11, 11, 11, 10, 10, 10, 10, 10...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 73,...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 0, ...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 39, NA, NA, 7,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 0, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 0, ...

sfoflight18Jul <- CAflightsJul18 %>%  
 select(-X46) %>%  
 filter(ORIGIN\_CITY\_NAME %in% c("San Francisco, CA", "Oakland, CA", "San Jose, CA"))

glimpse(sfoflight18Jul)

## Observations: 25,472  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...  
## $ MONTH <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...  
## $ DAY\_OF\_MONTH <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...  
## $ DAY\_OF\_WEEK <int> 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, ...  
## $ FL\_DATE <date> 2018-07-01, 2018-07-02, 2018-07-03, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "B6", "B6", "B6", "B6", "B6", "B6", "B6", ...  
## $ TAIL\_NUM <chr> "N985JB", "N987JB", "N988JB", "N923JB", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14771, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Francisco, CA", ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "JFK", "JFK", "JFK", "JFK", "JFK", "JFK", ...  
## $ DEST\_CITY\_NAME <chr> "New York, NY", "New York, NY", "New York,...  
## $ DEST\_STATE\_ABR <chr> "NY", "NY", "NY", "NY", "NY", "NY", "NY", ...  
## $ DEST\_STATE\_NM <chr> "New York", "New York", "New York", "New Y...  
## $ DEST\_WAC <int> 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22...  
## $ CRS\_DEP\_TIME <chr> "0630", "0630", "0630", "0630", "0630", "0...  
## $ DEP\_TIME <chr> "0618", "0650", "0625", "0623", "0721", "0...  
## $ DEP\_DELAY <dbl> -12, 20, -5, -7, 51, -4, -5, 30, 0, -3, -9...  
## $ DEP\_DELAY\_NEW <dbl> 0, 20, 0, 0, 51, 0, 0, 30, 0, 0, 0, 0, 0, ...  
## $ DEP\_DEL15 <dbl> 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, 1, -1, -1, 3, -1, -1, 2, 0, -1, -1, -1...  
## $ CRS\_ARR\_TIME <chr> "1509", "1509", "1509", "1509", "1509", "1...  
## $ ARR\_TIME <chr> "1459", "1509", "1456", "1537", "1624", "1...  
## $ ARR\_DELAY <dbl> -10, 0, -13, 28, 75, -15, -21, 7, -13, 8, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 28, 75, 0, 0, 7, 0, 8, 0, 0, 0, 0...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, 0, -1, 1, 5, -1, -2, 0, -1, 0, -1, -1,...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 339, 339, 339, 339, 339, 339, 339, 339, 33...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 341, 319, 331, 374, 363, 328, 323, 316, 32...  
## $ AIR\_TIME <dbl> 302, 298, 304, 326, 311, 300, 307, 302, 30...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 2586, 2586, 2586, 2586, 2586, 2586, 2586, ...  
## $ DISTANCE\_GROUP <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, 0, 51, NA, NA, NA, NA, NA, NA,...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, ...  
## $ NAS\_DELAY <dbl> NA, NA, NA, 28, 24, NA, NA, NA, NA, NA, NA...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, ...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, ...

sfoflights18 <- bind\_rows(sfoflight18Jan, sfoflight18Feb, sfoflight18Mar, sfoflight18Apr, sfoflight18May, sfoflight18Jun, sfoflight18Jul)

glimpse(sfoflights18)

## Observations: 163,056  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...

### 3.

library(nycflights13)

glimpse(flights)

## Observations: 336,776  
## Variables: 19  
## $ year <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...  
## $ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ dep\_time <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...  
## $ sched\_dep\_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...  
## $ dep\_delay <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...  
## $ arr\_time <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...  
## $ sched\_arr\_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...  
## $ arr\_delay <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...  
## $ carrier <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...  
## $ flight <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...  
## $ tailnum <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...  
## $ origin <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...  
## $ dest <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...  
## $ air\_time <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...  
## $ distance <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...  
## $ hour <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...  
## $ minute <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ time\_hour <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...

glimpse(planes)

## Observations: 3,322  
## Variables: 9  
## $ tailnum <chr> "N10156", "N102UW", "N103US", "N104UW", "N10575",...  
## $ year <int> 2004, 1998, 1999, 1999, 2002, 1999, 1999, 1999, 1...  
## $ type <chr> "Fixed wing multi engine", "Fixed wing multi engi...  
## $ manufacturer <chr> "EMBRAER", "AIRBUS INDUSTRIE", "AIRBUS INDUSTRIE"...  
## $ model <chr> "EMB-145XR", "A320-214", "A320-214", "A320-214", ...  
## $ engines <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...  
## $ seats <int> 55, 182, 182, 182, 55, 182, 182, 182, 182, 182, 5...  
## $ speed <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ engine <chr> "Turbo-fan", "Turbo-fan", "Turbo-fan", "Turbo-fan...

### 4.

None.

### 5.

**Question 4.2 What month had the highest proportion of cancelled fights? What month had the lowest? Interpret any seasonal patterns.**

Check the cancellation indicator variable “CANCELLED” in the dataset. There is no missing values in this variable.

sfoflights18 %>%   
 group\_by(CANCELLED) %>%  
 summarize(n=n())

## # A tibble: 2 x 2  
## CANCELLED n  
## <dbl> <int>  
## 1 0 160975  
## 2 1 2081

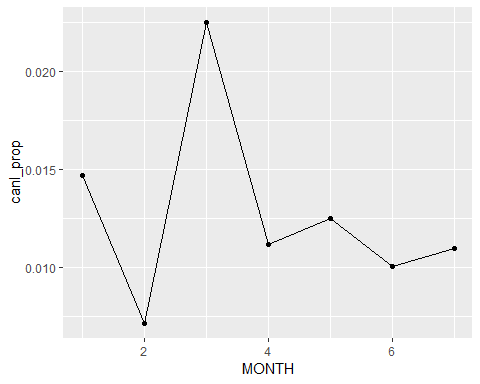
Analyze the proportion of cancelled flights for each months.

sfoflights18 %>%  
 group\_by(MONTH)%>%  
 summarize(n=n(), canl\_num = sum(CANCELLED), canl\_prop = canl\_num / n) %>%  
 arrange(canl\_prop)

## # A tibble: 7 x 4  
## MONTH n canl\_num canl\_prop  
## <int> <int> <dbl> <dbl>  
## 1 2 20142 144 0.00715  
## 2 6 24526 246 0.0100   
## 3 7 25472 280 0.0110   
## 4 4 23090 258 0.0112   
## 5 5 24165 302 0.0125   
## 6 1 22606 332 0.0147   
## 7 3 23055 519 0.0225

Visualize the propotion of cancelled flights for each months.

sfoflights18 %>%  
 group\_by(MONTH)%>%  
 summarize(n=n(),   
 canl\_num = sum(CANCELLED),   
 canl\_prop = canl\_num / n) %>%  
 arrange(canl\_prop) %>%  
 ggplot(aes(x=MONTH, y= canl\_prop))+  
 geom\_point()+geom\_line()



**Question 4.3 What plane (specifed by the tailnum variable) traveled the most times from Bay Area airports in 2018? Plot the number of trips per week over the year.**

Find the plane traveled the most times from Bay Area airports in 2018.

sfoflights18 %>%  
 group\_by(TAIL\_NUM) %>%  
 summarize(n=n()) %>%  
 arrange(desc(n))

## # A tibble: 3,724 x 2  
## TAIL\_NUM n  
## <chr> <int>  
## 1 N633VA 260  
## 2 N281VA 248  
## 3 N630VA 248  
## 4 N629VA 246  
## 5 N284VA 245  
## 6 N285VA 245  
## 7 N286VA 245  
## 8 N640VA 245  
## 9 N847VA 236  
## 10 N835VA 233  
## # ... with 3,714 more rows

Get the week variable.

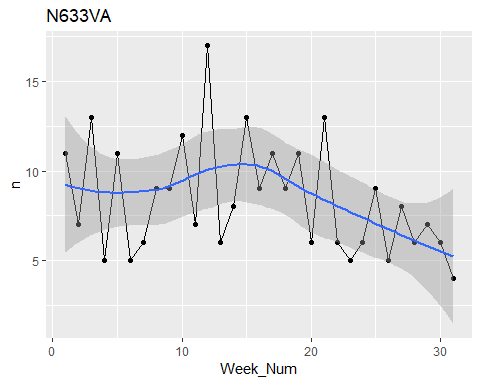
library(lubridate)  
  
sfoflights18 %>%  
 mutate(week = week(FL\_DATE)) %>%  
 glimpse()

## Observations: 163,056  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ week <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...

Visualize the flights of “N633VA” in weeks.

sfoflights18 %>%  
 filter(TAIL\_NUM == "N633VA") %>%  
 mutate(Week\_Num = week(FL\_DATE)) %>%  
 group\_by(Week\_Num) %>%  
 summarise(n=n()) %>%  
 ggplot(aes(x= Week\_Num, y=n)) +   
 geom\_point() +  
 geom\_line() +  
 geom\_smooth() +  
 ggtitle("N633VA")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

 **Question 4.4 What is the oldest plane (specifed by the tailnum variable) that flew from Bay Area airports in 2018? How many airplanes that flew from Bay Area are included in the planes table?**

Find the oldest plane. (I just use plane data from nycflights13 package.)

planes %>%  
 mutate(age = year) %>%  
 select(age, tailnum) %>%  
 right\_join(sfoflights18, by=c("tailnum"="TAIL\_NUM")) %>%  
 group\_by(age, tailnum) %>%  
 summarize(n=n()) %>%  
 arrange(age)

## # A tibble: 3,724 x 3  
## # Groups: age [26]  
## age tailnum n  
## <int> <chr> <int>  
## 1 1989 N502UA 103  
## 2 1989 N505UA 139  
## 3 1990 N152DL 6  
## 4 1990 N171DN 4  
## 5 1990 N172DN 5  
## 6 1990 N174DN 5  
## 7 1990 N175DN 1  
## 8 1990 N176DN 1  
## 9 1990 N510UA 127  
## 10 1990 N512UA 141  
## # ... with 3,714 more rows

Count airplanes that flew from Bay Area and included in the planes table.

planes

## # A tibble: 3,322 x 9  
## tailnum year type manufacturer model engines seats speed engine  
## <chr> <int> <chr> <chr> <chr> <int> <int> <int> <chr>   
## 1 N10156 2004 Fixed wi~ EMBRAER EMB-1~ 2 55 NA Turbo~  
## 2 N102UW 1998 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 3 N103US 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 4 N104UW 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 5 N10575 2002 Fixed wi~ EMBRAER EMB-1~ 2 55 NA Turbo~  
## 6 N105UW 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 7 N107US 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 8 N108UW 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 9 N109UW 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## 10 N110UW 1999 Fixed wi~ AIRBUS INDUS~ A320-~ 2 182 NA Turbo~  
## # ... with 3,312 more rows

sfoflights18 %>%  
 glimpse

## Observations: 163,056  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...

sfoflights18 %>%  
 skim(YEAR)

## Skim summary statistics  
## n obs: 163056   
## n variables: 45   
##   
## -- Variable type:integer -------------------------------------------------------  
## variable missing complete n mean sd p0 p25 p50 p75 p100  
## YEAR 0 163056 163056 2018 0 2018 2018 2018 2018 2018  
## hist  
## <U+2581><U+2581><U+2581><U+2587><U+2581><U+2581><U+2581><U+2581>

sfoflights18 %>%  
 select(TAIL\_NUM, YEAR) %>%  
 right\_join(planes, by=c("TAIL\_NUM"="tailnum")) %>%  
 select(TAIL\_NUM, YEAR) %>%  
 filter(!is.na(YEAR)) %>%  
 skim(TAIL\_NUM)

## Skim summary statistics  
## n obs: 91693   
## n variables: 2   
##   
## -- Variable type:character -----------------------------------------------------  
## variable missing complete n min max empty n\_unique  
## TAIL\_NUM 0 91693 91693 5 6 0 1904

## Part 2.

### 1.Find a data source for weather data. Download the temperature and other weather data for the Bay Area in 2018.

?weather

## starting httpd help server ... done

weather

## # A tibble: 26,115 x 15  
## origin year month day hour temp dewp humid wind\_dir wind\_speed  
## <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 EWR 2013 1 1 1 39.0 26.1 59.4 270 10.4   
## 2 EWR 2013 1 1 2 39.0 27.0 61.6 250 8.06  
## 3 EWR 2013 1 1 3 39.0 28.0 64.4 240 11.5   
## 4 EWR 2013 1 1 4 39.9 28.0 62.2 250 12.7   
## 5 EWR 2013 1 1 5 39.0 28.0 64.4 260 12.7   
## 6 EWR 2013 1 1 6 37.9 28.0 67.2 240 11.5   
## 7 EWR 2013 1 1 7 39.0 28.0 64.4 240 15.0   
## 8 EWR 2013 1 1 8 39.9 28.0 62.2 250 10.4   
## 9 EWR 2013 1 1 9 39.9 28.0 62.2 260 15.0   
## 10 EWR 2013 1 1 10 41 28.0 59.6 260 13.8   
## # ... with 26,105 more rows, and 5 more variables: wind\_gust <dbl>,  
## # precip <dbl>, pressure <dbl>, visib <dbl>, time\_hour <dttm>

glimpse(weather)

## Observations: 26,115  
## Variables: 15  
## $ origin <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "E...  
## $ year <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...  
## $ month <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ hour <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, ...  
## $ temp <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39...  
## $ dewp <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28...  
## $ humid <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62...  
## $ wind\_dir <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 2...  
## $ wind\_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11...  
## $ wind\_gust <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ precip <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ pressure <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 101...  
## $ visib <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,...  
## $ time\_hour <dttm> 2013-01-01 01:00:00, 2013-01-01 02:00:00, 2013-01-...

sfoweather18 <- read.csv("https://mesonet.agron.iastate.edu/cgi-bin/request/asos.py?station=SFO&station=OAK&station=SJC&data=all&year1=2018&month1=1&day1=1&year2=2018&month2=10&day2=6&tz=Etc%2FUTC&format=onlycomma&latlon=yes&direct=no&report\_type=1&report\_type=2",header = TRUE, na.strings = "M")  
glimpse(sfoweather18)

## Observations: 259,167  
## Variables: 24  
## $ station <fct> SFO, SJC, OAK, SJC, SFO, OAK, SFO, OAK, SJC, SJC, OAK,...  
## $ valid <fct> 2018-01-01 00:00, 2018-01-01 00:00, 2018-01-01 00:00, ...  
## $ lon <dbl> -122.3749, -121.9244, -122.2207, -121.9244, -122.3749,...  
## $ lat <dbl> 37.6190, 37.3594, 37.7213, 37.3594, 37.6190, 37.7213, ...  
## $ tmpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ dwpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ relh <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ drct <dbl> 50, 310, 330, 300, 70, 310, 60, 320, 300, 310, 330, 60...  
## $ sknt <dbl> 3, 6, 5, 6, 3, 3, 3, 6, 6, 4, 5, 3, 4, 5, 0, 5, 4, 0, ...  
## $ p01i <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ alti <dbl> 30.08, 30.08, 30.09, 30.08, 30.08, 30.09, 30.08, 30.09...  
## $ mslp <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vsby <dbl> 6, 4, 10, 4, 6, 10, 6, 10, 4, 4, 10, 6, 4, 10, 7, 4, 1...  
## $ gust <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc1 <fct> CLR, FEW, CLR, FEW, CLR, CLR, CLR, CLR, FEW, FEW, CLR,...  
## $ skyc2 <fct> NA, SCT, NA, SCT, NA, NA, NA, NA, SCT, SCT, NA, NA, SC...  
## $ skyc3 <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc4 <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl1 <dbl> NA, 100, NA, 100, NA, NA, NA, NA, 100, 100, NA, NA, 10...  
## $ skyl2 <dbl> NA, 12000, NA, 12000, NA, NA, NA, NA, 12000, 12000, NA...  
## $ skyl3 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl4 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ wxcodes <fct> HZ, HZ, NA, HZ, HZ, NA, HZ, NA, HZ, HZ, NA, HZ, HZ, NA...  
## $ metar <fct> KSFO 010000Z AUTO 05003KT 6SM HZ CLR 13/09 A3008 RMK T...

### 2.Try to find the same variables in the nycflights 13 weather data.

None.

### 3.Merge your new weather data with the same measures in the nycflights13 flights data from your sfoflights18 data.

sfoweather18 <- sfoweather18 %>%   
 mutate(  
 station = as.character(station),   
 valid = as\_datetime(valid),   
 skyc1 = as.character(skyc1),  
 skyc2 = as.character(skyc2),  
 skyc3 = as.character(skyc3),  
 skyc4 = as.character(skyc4),  
 wxcodes = as.character(wxcodes),  
 metar = as.character(metar)) %>%  
 glimpse()

## Observations: 259,167  
## Variables: 24  
## $ station <chr> "SFO", "SJC", "OAK", "SJC", "SFO", "OAK", "SFO", "OAK"...  
## $ valid <dttm> 2018-01-01 00:00:00, 2018-01-01 00:00:00, 2018-01-01 ...  
## $ lon <dbl> -122.3749, -121.9244, -122.2207, -121.9244, -122.3749,...  
## $ lat <dbl> 37.6190, 37.3594, 37.7213, 37.3594, 37.6190, 37.7213, ...  
## $ tmpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ dwpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ relh <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ drct <dbl> 50, 310, 330, 300, 70, 310, 60, 320, 300, 310, 330, 60...  
## $ sknt <dbl> 3, 6, 5, 6, 3, 3, 3, 6, 6, 4, 5, 3, 4, 5, 0, 5, 4, 0, ...  
## $ p01i <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ alti <dbl> 30.08, 30.08, 30.09, 30.08, 30.08, 30.09, 30.08, 30.09...  
## $ mslp <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vsby <dbl> 6, 4, 10, 4, 6, 10, 6, 10, 4, 4, 10, 6, 4, 10, 7, 4, 1...  
## $ gust <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc1 <chr> "CLR", "FEW", "CLR", "FEW", "CLR", "CLR", "CLR", "CLR"...  
## $ skyc2 <chr> NA, "SCT", NA, "SCT", NA, NA, NA, NA, "SCT", "SCT", NA...  
## $ skyc3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl1 <dbl> NA, 100, NA, 100, NA, NA, NA, NA, 100, 100, NA, NA, 10...  
## $ skyl2 <dbl> NA, 12000, NA, 12000, NA, NA, NA, NA, 12000, 12000, NA...  
## $ skyl3 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl4 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ wxcodes <chr> "HZ", "HZ", NA, "HZ", "HZ", NA, "HZ", NA, "HZ", "HZ", ...  
## $ metar <chr> "KSFO 010000Z AUTO 05003KT 6SM HZ CLR 13/09 A3008 RMK ...

glimpse(sfoweather18)

## Observations: 259,167  
## Variables: 24  
## $ station <chr> "SFO", "SJC", "OAK", "SJC", "SFO", "OAK", "SFO", "OAK"...  
## $ valid <dttm> 2018-01-01 00:00:00, 2018-01-01 00:00:00, 2018-01-01 ...  
## $ lon <dbl> -122.3749, -121.9244, -122.2207, -121.9244, -122.3749,...  
## $ lat <dbl> 37.6190, 37.3594, 37.7213, 37.3594, 37.6190, 37.7213, ...  
## $ tmpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ dwpf <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ relh <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ drct <dbl> 50, 310, 330, 300, 70, 310, 60, 320, 300, 310, 330, 60...  
## $ sknt <dbl> 3, 6, 5, 6, 3, 3, 3, 6, 6, 4, 5, 3, 4, 5, 0, 5, 4, 0, ...  
## $ p01i <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ alti <dbl> 30.08, 30.08, 30.09, 30.08, 30.08, 30.09, 30.08, 30.09...  
## $ mslp <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vsby <dbl> 6, 4, 10, 4, 6, 10, 6, 10, 4, 4, 10, 6, 4, 10, 7, 4, 1...  
## $ gust <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc1 <chr> "CLR", "FEW", "CLR", "FEW", "CLR", "CLR", "CLR", "CLR"...  
## $ skyc2 <chr> NA, "SCT", NA, "SCT", NA, NA, NA, NA, "SCT", "SCT", NA...  
## $ skyc3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyc4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl1 <dbl> NA, 100, NA, 100, NA, NA, NA, NA, 100, 100, NA, NA, 10...  
## $ skyl2 <dbl> NA, 12000, NA, 12000, NA, NA, NA, NA, 12000, 12000, NA...  
## $ skyl3 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ skyl4 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ wxcodes <chr> "HZ", "HZ", NA, "HZ", "HZ", NA, "HZ", NA, "HZ", "HZ", ...  
## $ metar <chr> "KSFO 010000Z AUTO 05003KT 6SM HZ CLR 13/09 A3008 RMK ...

glimpse(sfoflights18)

## Observations: 163,056  
## Variables: 45  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...

sfoflights18 %>%  
 mutate(DEP\_hour = as.integer(substr(DEP\_TIME, 0, nchar(DEP\_TIME)-2))) %>%  
 glimpse()

## Observations: 163,056  
## Variables: 46  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ DEP\_hour <int> 10, 13, 11, 22, 13, 10, 13, 11, 14, 22, 15...

sfoweather18Hour <- sfoweather18 %>%   
 select(1:14) %>%  
 mutate(wea\_date = date(valid), wea\_hour = hour(valid)) %>%  
 group\_by(station, wea\_date, wea\_hour) %>%  
 summarize(  
 lon = mean(lon, na.rm =TRUE),  
 lat = mean(lat, na.rm = TRUE),  
 tmpf = mean(tmpf, na.rm = TRUE),  
 dwpf = mean(dwpf, na.rm = TRUE),  
 relh = mean(relh, na.rm = TRUE),  
 drct = mean(drct, na.rm = TRUE),  
 sknt = mean(sknt, na.rm = TRUE),  
 p01i = mean(p01i, na.rm = TRUE),  
 alti = mean(alti, na.rm = TRUE),  
 mslp = mean(mslp, na.rm = TRUE),  
 vsby = mean(vsby, na.rm = TRUE),  
 gust = mean(gust, na.rm = TRUE)  
 )  
sfoweather18Hour %>% glimpse()

## Observations: 19,999  
## Variables: 15  
## $ station <chr> "OAK", "OAK", "OAK", "OAK", "OAK", "OAK", "OAK", "OAK...  
## $ wea\_date <date> 2018-01-01, 2018-01-01, 2018-01-01, 2018-01-01, 2018...  
## $ wea\_hour <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...  
## $ lon <dbl> -122.2207, -122.2207, -122.2207, -122.2207, -122.2207...  
## $ lat <dbl> 37.7213, 37.7213, 37.7213, 37.7213, 37.7213, 37.7213,...  
## $ tmpf <dbl> 55.9, 54.0, 52.0, 52.0, 48.9, 48.9, 48.0, 48.9, 48.0,...  
## $ dwpf <dbl> 48.9, 50.0, 50.0, 50.0, 48.0, 48.0, 48.0, 48.0, 46.9,...  
## $ relh <dbl> 77.28, 86.29, 92.86, 92.86, 96.69, 96.69, 100.00, 96....  
## $ drct <dbl> 308.46154, 313.07692, 350.76923, 293.84615, 88.46154,...  
## $ sknt <dbl> 4.4615385, 5.0000000, 3.7692308, 3.2307692, 2.4615385...  
## $ p01i <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ alti <dbl> 30.09308, 30.10538, 30.12538, 30.13692, 30.14846, 30....  
## $ mslp <dbl> 1019.1, 1019.6, 1020.4, 1020.6, 1021.1, 1021.4, 1022....  
## $ vsby <dbl> 10.000000, 10.000000, 10.000000, 10.000000, 9.307692,...  
## $ gust <dbl> NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN...

sfo\_join <- sfoflights18 %>%  
 mutate(DEP\_hour = as.integer(substr(DEP\_TIME, 0, nchar(DEP\_TIME)-2))) %>%  
 left\_join(sfoweather18Hour, by=c("ORIGIN" = "station", "DEP\_hour" = "wea\_hour", "FL\_DATE" = "wea\_date"))  
  
sfo\_join %>%  
 glimpse()

## Observations: 163,056  
## Variables: 58  
## $ YEAR <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, ...  
## $ QUARTER <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ MONTH <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DAY\_OF\_MONTH <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...  
## $ DAY\_OF\_WEEK <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...  
## $ FL\_DATE <date> 2018-01-27, 2018-01-27, 2018-01-27, 2018-...  
## $ OP\_UNIQUE\_CARRIER <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", ...  
## $ TAIL\_NUM <chr> "N36469", "N835UA", "N38268", "N14214", "N...  
## $ OP\_CARRIER\_FL\_NUM <int> 358, 346, 317, 313, 300, 294, 292, 285, 26...  
## $ ORIGIN\_AIRPORT\_ID <int> 14771, 14831, 14771, 14771, 14771, 14771, ...  
## $ ORIGIN <chr> "SFO", "SJC", "SFO", "SFO", "SFO", "SFO", ...  
## $ ORIGIN\_CITY\_NAME <chr> "San Francisco, CA", "San Jose, CA", "San ...  
## $ ORIGIN\_STATE\_ABR <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", ...  
## $ ORIGIN\_STATE\_NM <chr> "California", "California", "California", ...  
## $ ORIGIN\_WAC <int> 91, 91, 91, 91, 91, 91, 91, 91, 91, 91, 91...  
## $ DEST <chr> "LAS", "ORD", "IND", "RNO", "HNL", "DFW", ...  
## $ DEST\_CITY\_NAME <chr> "Las Vegas, NV", "Chicago, IL", "Indianapo...  
## $ DEST\_STATE\_ABR <chr> "NV", "IL", "IN", "NV", "HI", "TX", "FL", ...  
## $ DEST\_STATE\_NM <chr> "Nevada", "Illinois", "Indiana", "Nevada",...  
## $ DEST\_WAC <int> 85, 41, 42, 85, 2, 74, 33, 74, 23, 91, 13,...  
## $ CRS\_DEP\_TIME <chr> "1042", "1400", "1200", "2254", "1320", "1...  
## $ DEP\_TIME <chr> "1033", "1342", "1146", "2248", "1315", "1...  
## $ DEP\_DELAY <dbl> -9, -18, -14, -6, -5, -8, -5, -1, 8, -8, 1...  
## $ DEP\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 121, 0, 0, 0...  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ DEP\_DELAY\_GROUP <int> -1, -2, -1, -1, -1, -1, -1, -1, 0, -1, 8, ...  
## $ CRS\_ARR\_TIME <chr> "1221", "2007", "1913", "2359", "1659", "1...  
## $ ARR\_TIME <chr> "1208", "1944", "1851", "2341", "1711", "1...  
## $ ARR\_DELAY <dbl> -13, -23, -22, -18, 12, -22, -10, -6, -2, ...  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 71, 0, 0, 3...  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...  
## $ ARR\_DELAY\_GROUP <int> -1, -2, -2, -2, 0, -2, -1, -1, -1, -2, 4, ...  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ CRS\_ELAPSED\_TIME <dbl> 99, 247, 253, 65, 339, 215, 312, 236, 320,...  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 95, 242, 245, 53, 356, 201, 307, 231, 310,...  
## $ AIR\_TIME <dbl> 62, 218, 221, 32, 331, 179, 278, 197, 287,...  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ DISTANCE <dbl> 414, 1829, 1943, 192, 2398, 1464, 2446, 16...  
## $ DISTANCE\_GROUP <int> 2, 8, 8, 1, 10, 6, 10, 7, 11, 2, 11, 10, 2...  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 71...  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0,...  
## $ DEP\_hour <int> 10, 13, 11, 22, 13, 10, 13, 11, 14, 22, 15...  
## $ lon <dbl> -122.3749, -121.9244, -122.3749, -122.3749...  
## $ lat <dbl> 37.6190, 37.3594, 37.6190, 37.6190, 37.619...  
## $ tmpf <dbl> 45.0, 41.0, 43.0, 59.0, 42.1, 45.0, 42.1, ...  
## $ dwpf <dbl> 39.0, 37.9, 37.9, 44.1, 37.9, 39.0, 37.9, ...  
## $ relh <dbl> 79.36, 88.62, 82.04, 57.66, 84.93, 79.36, ...  
## $ drct <dbl> 75.38462, 0.00000, 185.38462, 323.84615, 1...  
## $ sknt <dbl> 0.9230769, 0.0000000, 2.4615385, 5.1538462...  
## $ p01i <dbl> 0.00000000, 0.00000000, 0.00000000, 0.0000...  
## $ alti <dbl> 30.39923, 30.40923, 30.39846, 30.39308, 30...  
## $ mslp <dbl> 1029.3, 1029.9, 1029.2, 1028.8, 1029.9, 10...  
## $ vsby <dbl> 10.000000, 10.000000, 10.000000, 10.000000...  
## $ gust <dbl> NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, Na...

### 4.Questions to answer:Answer Exercises 4.6, and 4.7 for the SF Bay Area 2018 data.

\*\* Question 4.6 Use the sfoflights18 and the weather table to answer the following questions: What is the distribution of temperature in July, 2018? 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?\*\*

sfoweather18 %>%   
 filter(month(valid) == 7) %>%  
 skim(tmpf, sknt)

## Skim summary statistics  
## n obs: 28925   
## n variables: 24   
##   
## -- Variable type:numeric -------------------------------------------------------  
## variable missing complete n mean sd p0 p25 p50 p75 p100 hist  
## sknt 79 28846 28925 9.1 4.92 0 5 9 12 27 <U+2583><U+2586><U+2587><U+2586><U+2583><U+2582><U+2581><U+2581>  
## tmpf 26542 2383 28925 65.18 7.43 52 59 64 69.1 93.9 <U+2583><U+2587><U+2587><U+2583><U+2582><U+2582><U+2581><U+2581>

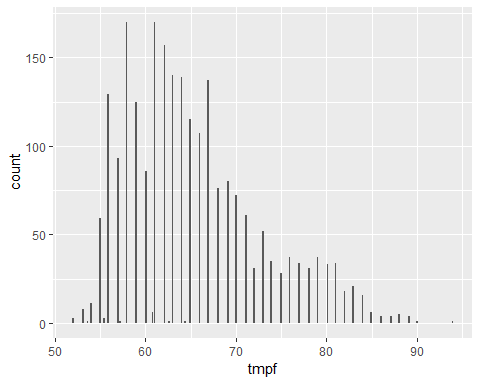
Temperature(tmpf) ranges from 52 to 93.9 with mean of 65.18 and sd of 7.43.

Wind speed variable(sknt) ranges from 0 to 27 with mean of 9.1 and sd of 4.92.After plot bar chart of sknt, we would identify outliers to be greater than 25.

Distribution of temperature in July, 2018

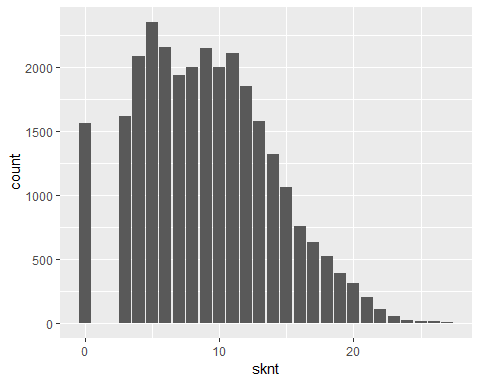
sfoweather18 %>%  
 filter(month(valid)==7) %>%  
 ggplot(aes(x=tmpf))+  
 geom\_bar()

## Warning: Removed 26542 rows containing non-finite values (stat\_count).

 Distribution of wind speed in July, 2018.

sfoweather18 %>%  
 filter(month(valid)==7) %>%  
 ggplot(aes(x=sknt))+  
 geom\_bar()

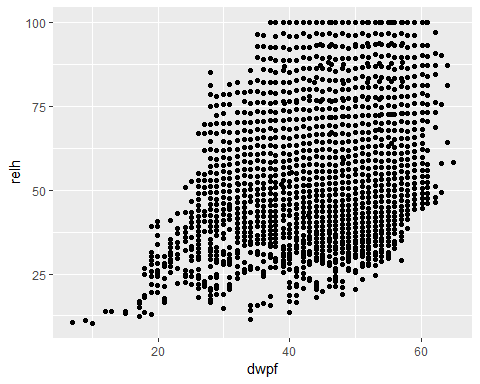
## Warning: Removed 79 rows containing non-finite values (stat\_count).



Relationship of dewpoint and relative humidity.

sfoweather18 %>%  
 ggplot(aes(x=dwpf, y=relh))+  
 geom\_point()

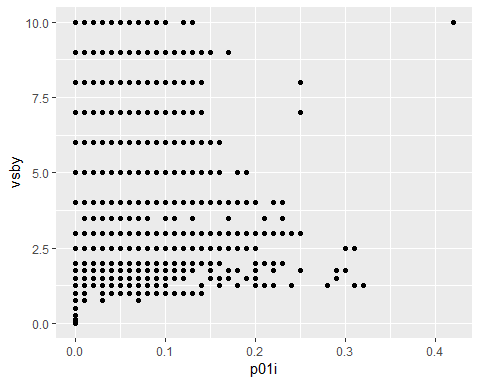
## Warning: Removed 237653 rows containing missing values (geom\_point).



Relationship of precipitaion and visibility.

sfoweather18 %>%   
 ggplot(aes(x=p01i, y=vsby))+  
 geom\_point()

## Warning: Removed 232772 rows containing missing values (geom\_point).



**Question 4.7 Use the sfoflights18 and the weather table to answer the following questions: On how many days was there precipitation in the Bay area in 2018? Were there differences in the mean visibility (visib) based on the day of the week and/or month of the year?**

sfoweather18 %>%  
 mutate(DATE=as.Date(valid, format="%m%d%Y")) %>%  
 filter(p01i>0)%>%  
 group\_by(DATE)%>%  
 summarize(n=n())

## # A tibble: 43 x 2  
## DATE n  
## <date> <int>  
## 1 2018-01-03 96  
## 2 2018-01-04 152  
## 3 2018-01-05 76  
## 4 2018-01-06 148  
## 5 2018-01-08 568  
## 6 2018-01-09 507  
## 7 2018-01-18 50  
## 8 2018-01-19 113  
## 9 2018-01-22 193  
## 10 2018-01-24 46  
## # ... with 33 more rows

sfoweather18 %>%  
 mutate(Week\_Day = weekdays(valid)) %>%  
 group\_by(Week\_Day) %>%  
 summarize(visib\_mean = mean(vsby, na.rm=TRUE))

## # A tibble: 7 x 2  
## Week\_Day visib\_mean  
## <chr> <dbl>  
## 1 Friday 9.47  
## 2 Monday 9.47  
## 3 Saturday 9.70  
## 4 Sunday 9.68  
## 5 Thursday 9.32  
## 6 Tuesday 9.50  
## 7 Wednesday 9.46

sfoweather18 %>%  
 mutate(Mon = month(valid))%>%  
 group\_by(Mon)%>%  
 summarize(visib\_mean = mean(vsby, na.rm=TRUE)) %>%  
 arrange(visib\_mean)

## # A tibble: 10 x 2  
## Mon visib\_mean  
## <dbl> <dbl>  
## 1 1 8.33  
## 2 3 9.42  
## 3 8 9.46  
## 4 9 9.60  
## 5 4 9.62  
## 6 5 9.69  
## 7 6 9.77  
## 8 7 9.86  
## 9 2 9.88  
## 10 10 9.95