Statistics 650 - Project

Nurahmet Maimaitiyiming

Wednesday October 13, 2019

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggmap)

## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.

library(skimr)

##   
## Attaching package: 'skimr'

## The following object is masked from 'package:stats':  
##   
## filter

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(textreadr)  
library(nycflights13)

## Question1.

**Answer:**

I have used the data of months Jan-May 2018 as below.

flight\_bay201801 <- read\_csv(file="./data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv")

## Warning: Missing column names filled in: 'X110' [110]

## Warning: 31561 parsing failures.  
## row col expected actual file  
## 1861 DivReachedDest 1/0/T/F/TRUE/FALSE 1.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv'  
## 1861 DivActualElapsedTime 1/0/T/F/TRUE/FALSE 284.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv'  
## 1861 DivArrDelay 1/0/T/F/TRUE/FALSE 94.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv'  
## 1861 DivDistance 1/0/T/F/TRUE/FALSE 0.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv'  
## 1861 Div1Airport 1/0/T/F/TRUE/FALSE ABE './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_1.csv'  
## .... .................... .................. ...... ................................................................................  
## See problems(...) for more details.

flight\_bay201802 <- read\_csv(file="./data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv")

## Warning: Missing column names filled in: 'X110' [110]

## Warning: 124 parsing failures.  
## row col expected actual file  
## 51172 Div2Airport 1/0/T/F/TRUE/FALSE LAS './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv'  
## 51172 Div2AirportID 1/0/T/F/TRUE/FALSE 12889 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv'  
## 51172 Div2AirportSeqID 1/0/T/F/TRUE/FALSE 1288903 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv'  
## 51172 Div2WheelsOn 1/0/T/F/TRUE/FALSE 2345 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv'  
## 51172 Div2TotalGTime 1/0/T/F/TRUE/FALSE 4.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_2.csv'  
## ..... ................ .................. ....... ................................................................................  
## See problems(...) for more details.

flight\_bay201803 <- read\_csv(file="./data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv")

## Warning: Missing column names filled in: 'X110' [110]

## Warning: 84 parsing failures.  
## row col expected actual file  
## 30765 Div2Airport 1/0/T/F/TRUE/FALSE PIE './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv'  
## 30765 Div2AirportID 1/0/T/F/TRUE/FALSE 14112 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv'  
## 30765 Div2AirportSeqID 1/0/T/F/TRUE/FALSE 1411206 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv'  
## 30765 Div2WheelsOn 1/0/T/F/TRUE/FALSE 2031 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv'  
## 30765 Div2TotalGTime 1/0/T/F/TRUE/FALSE 4.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_3.csv'  
## ..... ................ .................. ....... ................................................................................  
## See problems(...) for more details.

flight\_bay201804 <- read\_csv(file="./data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv")

## Warning: Missing column names filled in: 'X110' [110]

## Warning: 6744 parsing failures.  
## row col expected actual file  
## 1391 DivReachedDest 1/0/T/F/TRUE/FALSE 1.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv'  
## 1391 DivActualElapsedTime 1/0/T/F/TRUE/FALSE 333.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv'  
## 1391 DivArrDelay 1/0/T/F/TRUE/FALSE 253.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv'  
## 1391 DivDistance 1/0/T/F/TRUE/FALSE 0.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv'  
## 1391 Div1WheelsOff 1/0/T/F/TRUE/FALSE 2243 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_4.csv'  
## .... .................... .................. ...... ................................................................................  
## See problems(...) for more details.

flight\_bay201805 <- read\_csv(file="./data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv")

## Warning: Missing column names filled in: 'X110' [110]

## Warning: 78 parsing failures.  
## row col expected actual file  
## 29552 Div2Airport 1/0/T/F/TRUE/FALSE LAS './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv'  
## 29552 Div2AirportID 1/0/T/F/TRUE/FALSE 12889 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv'  
## 29552 Div2AirportSeqID 1/0/T/F/TRUE/FALSE 1288903 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv'  
## 29552 Div2WheelsOn 1/0/T/F/TRUE/FALSE 1629 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv'  
## 29552 Div2TotalGTime 1/0/T/F/TRUE/FALSE 10.00 './data/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_(1987\_present)\_2018\_5.csv'  
## ..... ................ .................. ....... ................................................................................  
## See problems(...) for more details.

bay\_flight01<- flight\_bay201801%>% filter(Origin=="SFO" | Origin=="SJC"| Origin=="OAK")%>%select(c(Year:Reporting\_Airline,Tail\_Number,Origin:OriginState,Dest:DestState,CRSDepTime:DepDel15,CRSArrTime:ArrDel15,AirTime,Distance))  
bay\_flight02<- flight\_bay201802%>% filter(Origin=="SFO" |Origin=="SJC"|Origin=="OAK")%>%select(c(Year:Reporting\_Airline,Tail\_Number,Origin:OriginState,Dest:DestState,CRSDepTime:DepDel15,CRSArrTime:ArrDel15,AirTime,Distance))  
bay\_flight03<- flight\_bay201803%>% filter(Origin=="SFO" | Origin=="SJC"|Origin=="OAK")%>%select(c(Year:Reporting\_Airline,Tail\_Number,Origin:OriginState,Dest:DestState,CRSDepTime:DepDel15,CRSArrTime:ArrDel15,AirTime,Distance))  
bay\_flight04<- flight\_bay201804%>% filter(Origin=="SFO" | Origin=="SJC"|Origin=="OAK")%>%select(c(Year:Reporting\_Airline,Tail\_Number,Origin:OriginState,Dest:DestState,CRSDepTime:DepDel15,CRSArrTime:ArrDel15,AirTime,Distance))  
bay\_flight05<- flight\_bay201805%>% filter(Origin=="SFO" | Origin=="SJC"|Origin=="OAK")%>%select(c(Year:Reporting\_Airline,Tail\_Number,Origin:OriginState,Dest:DestState,CRSDepTime:DepDel15,CRSArrTime:ArrDel15,AirTime,Distance))

## Question2.

**Answer:**

There were 22606 flights in Jan 2018 flewed from the three airports in bay area.

dim(bay\_flight01)

## [1] 22606 26

sfoflights18<- bind\_rows(bay\_flight01,bay\_flight02,bay\_flight03,bay\_flight04,bay\_flight05)  
dim(sfoflights18)

## [1] 113058 26

head(sfoflights18)

## # A tibble: 6 x 26  
## Year Quarter Month DayofMonth DayOfWeek FlightDate Reporting\_Airli…  
## <dbl> <dbl> <dbl> <dbl> <dbl> <date> <chr>   
## 1 2018 1 1 27 6 2018-01-27 UA   
## 2 2018 1 1 27 6 2018-01-27 UA   
## 3 2018 1 1 27 6 2018-01-27 UA   
## 4 2018 1 1 27 6 2018-01-27 UA   
## 5 2018 1 1 27 6 2018-01-27 UA   
## 6 2018 1 1 27 6 2018-01-27 UA   
## # … with 19 more variables: Tail\_Number <chr>, Origin <chr>,  
## # OriginCityName <chr>, OriginState <chr>, Dest <chr>,  
## # DestCityName <chr>, DestState <chr>, CRSDepTime <chr>, DepTime <chr>,  
## # DepDelay <dbl>, DepDelayMinutes <dbl>, DepDel15 <dbl>,  
## # CRSArrTime <chr>, ArrTime <chr>, ArrDelay <dbl>,  
## # ArrDelayMinutes <dbl>, ArrDel15 <dbl>, AirTime <dbl>, Distance <dbl>

## Question3.

**Answer:**

nycflights13\_names <- tribble(  
 ~columns, ~description,  
 'year', 'Date of departure. - year',  
 'month', 'Date of departure. - month',  
 'day', 'Date of departure.',  
 'dep\_time', 'Actual departure and arrival times (format HHMM or HMM), local tz.',  
 'arr\_time', 'Actual departure and arrival times (format HHMM or HMM), local tz.',  
 'sched\_dep\_time', 'Scheduled departure and arrival times (format HHMM or HMM), local tz.',  
 'sched\_arr\_time', 'Scheduled departure and arrival times (format HHMM or HMM), local tz.',  
 'dep\_delay', 'Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.',  
 'arr\_delay', 'Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.',  
 'carrier', 'Two letter carrier abbreviation. See airlines to get name.',  
 'flight', 'Flight number.',  
 'tailnum', 'Plane tail number. See planes for additional metadata.',  
 'origin', 'Origin and destination. See airports for additional metadata.',  
 'dest', 'Origin and destination. See airports for additional metadata.',  
 'air\_time', 'Amount of time spent in the air, in minutes.',  
 'distance', 'Distance between airports, in miles.',  
 'hour', 'Time of scheduled departure broken into hour and minutes. - hour',  
 'minute', 'Time of scheduled departure broken into hour and minutes. minute',  
 'time\_hour', 'Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.'  
)  
  
kable(nycflights13\_names)

|  |  |
| --- | --- |
| columns | description |
| year | Date of departure. - year |
| month | Date of departure. - month |
| day | Date of departure. |
| dep\_time | Actual departure and arrival times (format HHMM or HMM), local tz. |
| arr\_time | Actual departure and arrival times (format HHMM or HMM), local tz. |
| sched\_dep\_time | Scheduled departure and arrival times (format HHMM or HMM), local tz. |
| sched\_arr\_time | Scheduled departure and arrival times (format HHMM or HMM), local tz. |
| dep\_delay | Departure and arrival delays, in minutes. Negative times represent early departures/arrivals. |
| arr\_delay | Departure and arrival delays, in minutes. Negative times represent early departures/arrivals. |
| carrier | Two letter carrier abbreviation. See airlines to get name. |
| flight | Flight number. |
| tailnum | Plane tail number. See planes for additional metadata. |
| origin | Origin and destination. See airports for additional metadata. |
| dest | Origin and destination. See airports for additional metadata. |
| air\_time | Amount of time spent in the air, in minutes. |
| distance | Distance between airports, in miles. |
| hour | Time of scheduled departure broken into hour and minutes. - hour |
| minute | Time of scheduled departure broken into hour and minutes. minute |
| time\_hour | Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data. |

X <- read\_html("readme.html")  
  
head(X)

## [1] "BACKGROUND"   
## [2] "The data contained in the compressed file has been extracted from the \n\tReporting Carrier On-Time Performance (1987-present) data table of the \"On-Time\" database from the TranStats data library. \n\tThe time period is indicated in the name of the compressed file; \n\tfor example, XXX\_XXXXX\_2001\_1 contains data of the first month of the year 2001."  
## [3] "RECORD LAYOUT"   
## [4] "Below are fields in the order that they appear on the records:"   
## [5] "Year"   
## [6] "Year"

# Remove the top four lines.  
  
X <- X[-c(1,2,3,4,5)]  
  
head(X)

## [1] "Year" "Quarter" "Quarter (1-4)" "Month"   
## [5] "Month" "DayofMonth"

# Now create a matrix that contains the variable names in the first column and the description in the second column.  
  
Y <- matrix(X[1:218], ncol = 2, byrow = TRUE)  
  
head(Y)

## [,1] [,2]   
## [1,] "Year" "Quarter"   
## [2,] "Quarter (1-4)" "Month"   
## [3,] "Month" "DayofMonth"   
## [4,] "Day of Month" "DayOfWeek"   
## [5,] "Day of Week" "FlightDate"   
## [6,] "Flight Date (yyyymmdd)" "Reporting\_Airline"

on\_time\_names <- tibble(columns = Y[,1], description = Y[,1])  
on\_time\_names

## # A tibble: 109 x 2  
## columns description   
## <chr> <chr>   
## 1 Year Year   
## 2 Quarter (1-4) Quarter (1-4)   
## 3 Month Month   
## 4 Day of Month Day of Month   
## 5 Day of Week Day of Week   
## 6 Flight Date (yyyymmdd) Flight Date (yyyymmdd)   
## 7 Unique Carrier Code. When the same… Unique Carrier Code. When the same …  
## 8 An identification number assigned … An identification number assigned b…  
## 9 Code assigned by IATA and commonly… Code assigned by IATA and commonly …  
## 10 Tail Number Tail Number   
## # … with 99 more rows

on\_time\_names <- on\_time\_names %>% mutate(columns = tolower(columns))  
head(on\_time\_names)

## # A tibble: 6 x 2  
## columns description   
## <chr> <chr>   
## 1 year Year   
## 2 quarter (1-4) Quarter (1-4)   
## 3 month Month   
## 4 day of month Day of Month   
## 5 day of week Day of Week   
## 6 flight date (yyyymmdd) Flight Date (yyyymmdd)

## Question4.

**Answer:**

Below is the table with the new variables available in flightsfo18 dataset.

sfoflights18\_names <- tribble(  
 ~columns, ~description,  
 'Quarter', 'Quarter (1-4)',  
 'DayOfWeek', 'Day of Week',  
 'OriginCityName', 'Origin Airport, City Name',  
 'OriginStat', 'Origin Airport, State Code',  
 'DepDelayMinutes','Difference in minutes between scheduled and actual departure time. Early departures set to 0.',  
 'DestCityName', 'Destination Airport, City Name',  
 'DestState', 'estination Airport, State Code.',  
 'DepDel15', 'Departure Delay Indicator, 15 Minutes or More (1=Yes).',  
 'ArrDelayMinutes', 'Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.',  
 'ArrDel15', 'Arrival Delay Indicator, 15 Minutes or More (1=Yes'  
 )  
  
kable(sfoflights18\_names)

|  |  |
| --- | --- |
| columns | description |
| Quarter | Quarter (1-4) |
| DayOfWeek | Day of Week |
| OriginCityName | Origin Airport, City Name |
| OriginStat | Origin Airport, State Code |
| DepDelayMinutes | Difference in minutes between scheduled and actual departure time. Early departures set to 0. |
| DestCityName | Destination Airport, City Name |
| DestState | estination Airport, State Code. |
| DepDel15 | Departure Delay Indicator, 15 Minutes or More (1=Yes). |
| ArrDelayMinutes | Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0. |
| ArrDel15 | Arrival Delay Indicator, 15 Minutes or More (1=Yes |
| ##Question5(4.2). |  |

**Answer:**

In the data for Jan-May 2018, March had the highes cancelled flights. Feb had the lowest cancelled flights. Maybe because of the rainy weather in 2018 from my memory.

sfoflights18%>%select(Month,ArrDelay)%>% group\_by(Month)%>%skim()

## Skim summary statistics  
## n obs: 113058   
## n variables: 2   
## group variables: Month   
##   
## ── Variable type:numeric ───────────────────────────────────────────────────────────────────────────  
## Month variable missing complete n mean sd p0 p25 p50 p75 p100  
## 1 ArrDelay 391 22215 22606 1 35.23 -62 -15 -7 5 1110  
## 2 ArrDelay 184 19958 20142 -0.47 35.01 -69 -15 -7 5 1216  
## 3 ArrDelay 560 22495 23055 5.83 40.97 -66 -14 -5 11 956  
## 4 ArrDelay 304 22786 23090 2.87 40.95 -68 -14 -6 7 1284  
## 5 ArrDelay 357 23808 24165 9.24 41.51 -57 -11 -2 14 1157  
## hist  
## ▇▁▁▁▁▁▁▁  
## ▇▁▁▁▁▁▁▁  
## ▇▁▁▁▁▁▁▁  
## ▇▁▁▁▁▁▁▁  
## ▇▁▁▁▁▁▁▁

## Question5(4.3).

**Answer:**

Flight N633VA travelled the most.

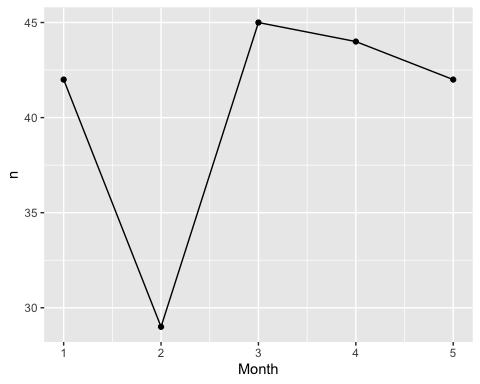
sfoflights18 %>%select(Tail\_Number) %>%group\_by(Tail\_Number)%>%count(count=n()) %>%arrange(desc(n))

## # A tibble: 3,592 x 3  
## # Groups: Tail\_Number [3,592]  
## Tail\_Number count n  
## <chr> <int> <int>  
## 1 N633VA 202 202  
## 2 N630VA 187 187  
## 3 N847VA 184 184  
## 4 N285VA 182 182  
## 5 N835VA 181 181  
## 6 N629VA 179 179  
## 7 N642VA 177 177  
## 8 N281VA 175 175  
## 9 N622VA 173 173  
## 10 N627VA 173 173  
## # … with 3,582 more rows

sfoflights18 %>%select(Year, Month, Tail\_Number) %>%filter(Tail\_Number=="N633VA") %>%group\_by(Year,Month) %>%summarise(count=n())

## # A tibble: 5 x 3  
## # Groups: Year [1]  
## Year Month count  
## <dbl> <dbl> <int>  
## 1 2018 1 42  
## 2 2018 2 29  
## 3 2018 3 45  
## 4 2018 4 44  
## 5 2018 5 42

sfoflights18 %>%filter(Tail\_Number=="N633VA") %>%group\_by(Month)%>%tally()%>%ggplot(aes(x= Month, y=n))+geom\_point()+geom\_line()



## Question5(4.4).

**Answer:** 3,322 planes flew from SFO are included in planes table.

planes\_1<- planes%>%select(tailnum,year)%>%group\_by(year)%>%arrange((year))  
planes\_1

## # A tibble: 3,322 x 2  
## # Groups: year [47]  
## tailnum year  
## <chr> <int>  
## 1 N381AA 1956  
## 2 N201AA 1959  
## 3 N567AA 1959  
## 4 N378AA 1963  
## 5 N575AA 1963  
## 6 N14629 1965  
## 7 N615AA 1967  
## 8 N425AA 1968  
## 9 N383AA 1972  
## 10 N364AA 1973  
## # … with 3,312 more rows

planes2 <-planes %>%rename( year\_m =year )  
planes2%>% left\_join(sfoflights18, by= c("tailnum"="Tail\_Number"))%>%group\_by(tailnum)%>% count(n=n())

## # A tibble: 3,322 x 3  
## # Groups: tailnum [3,322]  
## tailnum n nn  
## <chr> <int> <int>  
## 1 N10156 1 1  
## 2 N102UW 1 1  
## 3 N103US 1 1  
## 4 N104UW 1 1  
## 5 N10575 1 1  
## 6 N105UW 1 1  
## 7 N107US 1 1  
## 8 N108UW 1 1  
## 9 N109UW 1 1  
## 10 N110UW 1 1  
## # … with 3,312 more rows