Practice / Getting to know R

Prachi Sardana

## Question1 - To load data csv file into a dataframe df.flights

### used stringAsFactors = False so that the character coloumns don’t convert into the factors.

df.flights <- read.csv(url("https://s3.us-east-2.amazonaws.com/artificium.us/datasets/FlightsWithAirlines.csv"),stringsAsFactors = FALSE)

## Question 2

### The structure of data consitute 18 observation of 14 Variables showing the summary of dataframe

## 'data.frame': 18 obs. of 14 variables:  
## $ year : int 2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...  
## $ month : int 9 2 1 6 1 5 5 10 6 8 ...  
## $ day : int 18 14 7 7 30 2 5 24 16 18 ...  
## $ dep\_hr : int 2 12 13 22 24 11 5 20 5 2 ...  
## $ dep\_min : int 38 25 56 56 2 2 50 23 31 38 ...  
## $ dep\_delay: int 0 13 9 7 14 49 13 50 30 20 ...  
## $ carrier : chr "UA" "AA" "AA" "B6" ...  
## $ airline : chr "United" "American Airlines" "American Airlines" "JetBlue" ...  
## $ country : chr "USA" "USA" "USA" "USA" ...  
## $ flight : int 1545 441 1141 725 461 1696 507 5708 411 1545 ...  
## $ equip : chr "B737-8" "B777" "A321" "A321" ...  
## $ tailnum : chr "N14228" "N24211" "N619AA" "N804JB" ...  
## $ origin : chr "EWR" "MIA" "JFK" "JFK" ...  
## $ dest : chr "IAH" "GRU" "MIA" "BQN" ...

## Question 3

### The head function to display first 4 rows

## year month day dep\_hr dep\_min dep\_delay carrier airline country  
## 1 2022 9 18 2 38 0 UA United USA  
## 2 2022 2 14 12 25 13 AA American Airlines USA  
## 3 2022 1 7 13 56 9 AA American Airlines USA  
## 4 2022 6 7 22 56 7 B6 JetBlue USA  
## flight equip tailnum origin dest  
## 1 1545 B737-8 N14228 EWR IAH  
## 2 441 B777 N24211 MIA GRU  
## 3 1141 A321 N619AA JFK MIA  
## 4 725 A321 N804JB JFK BQN

## Question 4

### Tail function to display last 5 rows of a dataframe

## year month day dep\_hr dep\_min dep\_delay carrier airline country  
## 14 2022 2 30 24 2 0 DL Delta Airlines USA  
## 15 2022 6 2 11 2 68 UA United USA  
## 16 2022 6 5 5 50 850 B6 JetBlue USA  
## 17 2022 11 24 20 23 20 NK Spirit Airways USA  
## 18 2022 7 16 5 31 18 LH Lufthansa Germany  
## flight equip tailnum origin dest  
## 14 461 B757-2 N668DN LGA ATL  
## 15 1696 B737-MAX N39463 EWR ORD  
## 16 507 A321 N516JB EWR FLL  
## 17 5708 A321 N829AS BOS PBI  
## 18 411 B747-4 N593JB CLT MUC

## Question 5

### carrier, flight, origin, and destination columns from the dataframe

### subseted the specific coloumns using a vector and selected the specific coloumns

specific\_col <- df.flights[,c("carrier","flight","origin","dest")]  
print(specific\_col)

## carrier flight origin dest  
## 1 UA 1545 EWR IAH  
## 2 AA 441 MIA GRU  
## 3 AA 1141 JFK MIA  
## 4 B6 725 JFK BQN  
## 5 DL 461 LGA ATL  
## 6 UA 1696 EWR ORD  
## 7 B6 507 EWR FLL  
## 8 NK 5708 BOS PBI  
## 9 LH 411 CLT MUC  
## 10 UA 1545 EWR IAH  
## 11 AA 441 MIA GRU  
## 12 AA 1141 JFK MIA  
## 13 B6 725 JFK BQN  
## 14 DL 461 LGA ATL  
## 15 UA 1696 EWR ORD  
## 16 B6 507 EWR FLL  
## 17 NK 5708 BOS PBI  
## 18 LH 411 CLT MUC

## Question 6

### To calculate the average mean departure delay and display the result in R using cat function

mean\_dep\_delay <- mean(df.flights$dep\_delay)  
  
cat("Average(mean) departure delay is",mean\_dep\_delay)

## Average(mean) departure delay is 69.66667

## Question 7

### Add a new column ‘tod’ to the data frame with a value of “am” or “pm” depending whether the departure time was AM (before 12 noon) or PM (on or after 12 noon). The time in the data file is in 24-hour format. Hint: Look up how to use the ifelse function. Use the dep\_hr column. Print the dataframe to ensure the new column is there and is correct, but only display the carrier, flight, dep\_hr, and the new tod columns.

# Created an empty dataframe tod  
df.flights$tod <- 0  
  
# used ifelse statement based on the syntax ifelse(test\_expression, x, y)  
# where test\_expression is the logical condition or vector, x is the value to return when condition = True and y is the value to return when condition = False  
df.flights$tod <- ifelse(df.flights$dep\_hr < 12, "am", "pm")  
  
specific\_col <- df.flights[,c("carrier","flight","dep\_hr","tod")]  
print(specific\_col)

## carrier flight dep\_hr tod  
## 1 UA 1545 2 am  
## 2 AA 441 12 pm  
## 3 AA 1141 13 pm  
## 4 B6 725 22 pm  
## 5 DL 461 24 pm  
## 6 UA 1696 11 am  
## 7 B6 507 5 am  
## 8 NK 5708 20 pm  
## 9 LH 411 5 am  
## 10 UA 1545 2 am  
## 11 AA 441 12 pm  
## 12 AA 1141 13 pm  
## 13 B6 725 22 pm  
## 14 DL 461 24 pm  
## 15 UA 1696 11 am  
## 16 B6 507 5 am  
## 17 NK 5708 20 pm  
## 18 LH 411 5 am

## Question 8

### For each flight, display the carrier, flight number, and the actual departure time (scheduled departure plus departure delay) for flights that were delayed. Display the time in the format hh:mm in 24-hour format, e.g., display 23:20 rather than 11:20 or 11:20PM.

# actual departure = scheduled departure time in min + the departure delay in min   
actual\_departure <- df.flights$dep\_min +df.flights$dep\_delay\*60  
  
# as.POSIXlt function covert the time into list of date-time components  
df.flights$act\_dep\_hh <- format(as.POSIXlt(actual\_departure, origin = "2023-09-01"), format = "%H:%M")  
  
select\_col <- df.flights[,c("carrier","flight","act\_dep\_hh")]  
print(select\_col)

## carrier flight act\_dep\_hh  
## 1 UA 1545 20:00  
## 2 AA 441 20:13  
## 3 AA 1141 20:09  
## 4 B6 725 20:07  
## 5 DL 461 20:14  
## 6 UA 1696 20:49  
## 7 B6 507 20:13  
## 8 NK 5708 20:50  
## 9 LH 411 20:30  
## 10 UA 1545 20:20  
## 11 AA 441 21:31  
## 12 AA 1141 20:00  
## 13 B6 725 20:02  
## 14 DL 461 20:00  
## 15 UA 1696 21:08  
## 16 B6 507 10:10  
## 17 NK 5708 20:20  
## 18 LH 411 20:18