# *ASSIGNMENT 5*

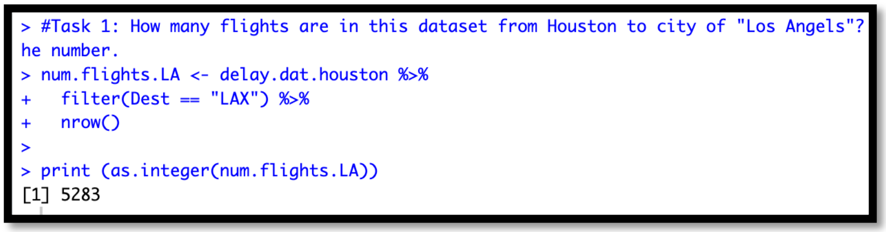
GSI Intro to Big Data and Data Mining

*The University of Texas at Austin*

*Zhaowen Fan*
*Rafael Ignacio Gonzalez Chong*

## Table of Contents

### Task 1: How many flights are in this dataset from Houston to city of "Los Angeles"? Print the number.

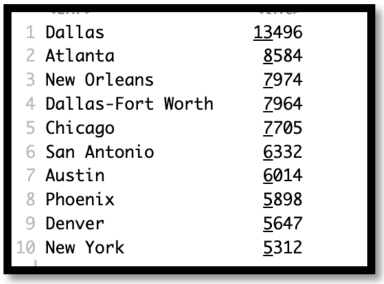


```
> #Task 1: How many flights are in this dataset from Houston to city of "Los Angels"?
he number.
> num.flights.LA <- delay.dat.houston %>%
+   filter(Dest == "LAX") %>%
+   nrow()
>
> print (as.integer(num.flights.LA))
[1] 5283
```

Fig. 1 Flights from Houston to LAX.

There are 5283 flights from Houston to Los Angeles.

### Task 2:  Which top 10 destination cities have the greatest number of flights?

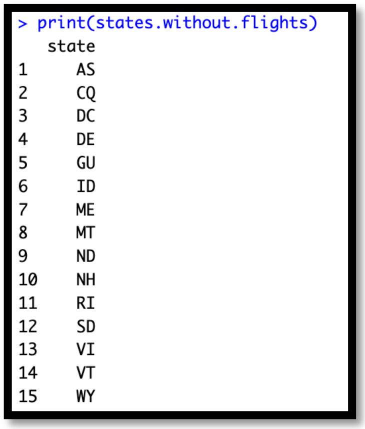


```
 1 Dallas               13496
 2 Atlanta               8584
 3 New Orleans           7974
 4 Dallas-Fort Worth     7964
 5 Chicago               7705
 6 San Antonio           6332
 7 Austin                6014
 8 Phoenix               5898
 9 Denver                5647
10 New York              5312
```

Fig. 2 Top 10 cities with the greatest number of flights.

The table represents the city with the number of flights.

### Task 3:  Which states have no flights?

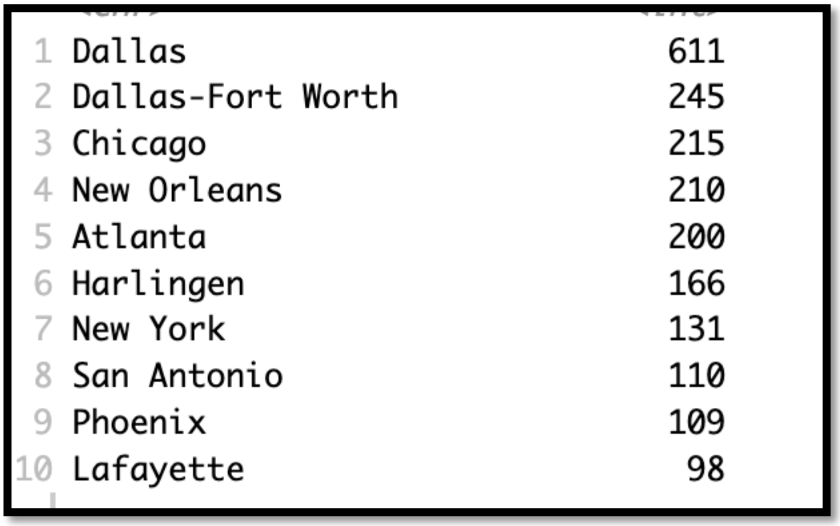


```
> print(states.without.flights)
   state
1     AS
2     CQ
3     DC
4     DE
5     GU
6     ID
7     ME
8     MT
9     ND
10    NH
11    RI
12    SD
13    VI
14    VT
15    WY
```

Fig. 3 States with no flights.

There are 15 states with no flights.

### Task 4: Which top 10 destination cities have the most cancelations of flights?

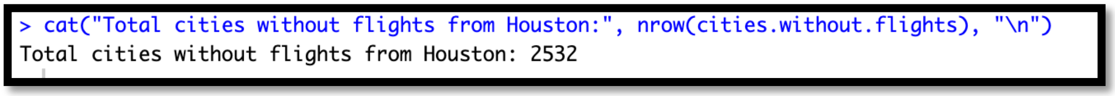


```
 1 Dallas                        611
 2 Dallas-Fort Worth             245
 3 Chicago                       215
 4 New Orleans                   210
 5 Atlanta                       200
 6 Harlingen                     166
 7 New York                      131
 8 San Antonio                   110
 9 Phoenix                       109
10 Lafayette                      98
```

Fig. 4 Cities with most cancelations of flights.

These 10 cities are the ones with the greatest number of canceled flights in the US.

### Task 5: Is there any Cities without a Flight from Houston?

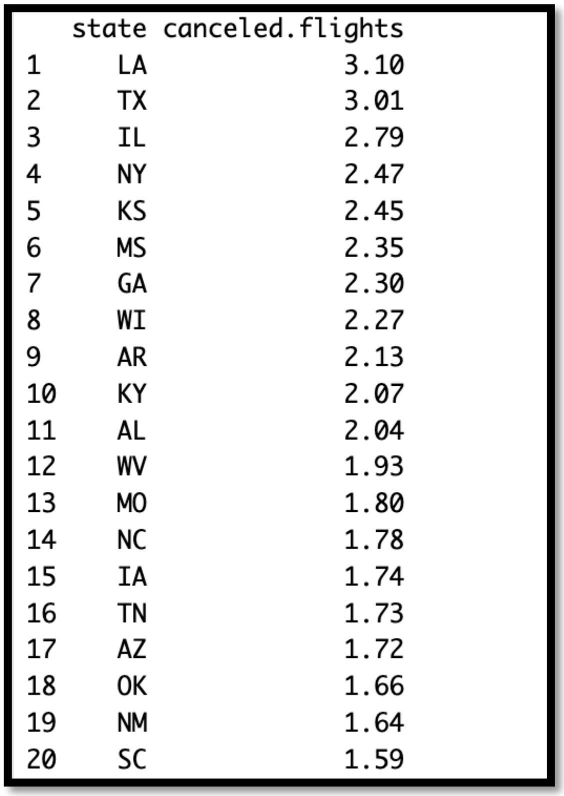


```
> cat("Total cities without flights from Houston:", nrow(cities.without.flights), "\n")
Total cities without flights from Houston: 2532
```

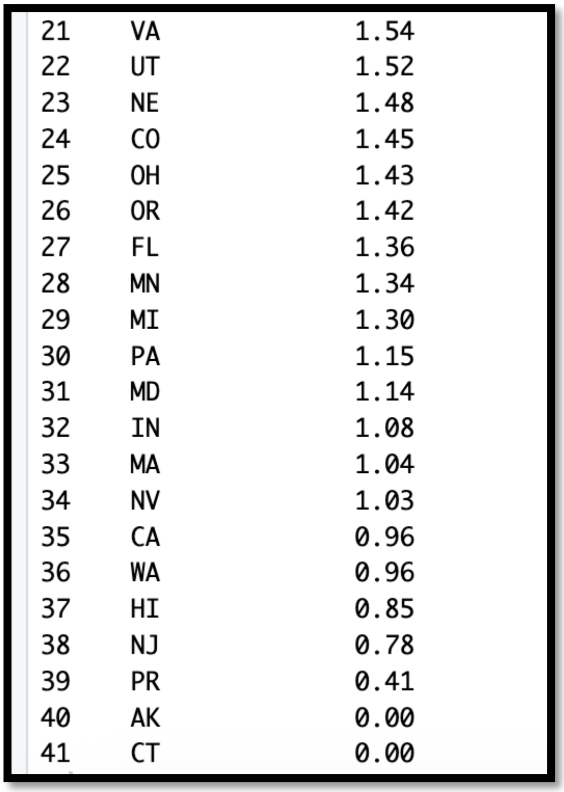
Fig. 5 Cities without flights from Houston.

There are 2532 flights that not departure from Houston.

### Task 6: What is the ratio of flights canceled for each state?



```
   state canceled.flights
1     LA             3.10
2     TX             3.01
3     IL             2.79
4     NY             2.47
5     KS             2.45
6     MS             2.35
7     GA             2.30
8     WI             2.27
9     AR             2.13
10    KY             2.07
11    AL             2.04
12    WV             1.93
13    MO             1.80
14    NC             1.78
15    IA             1.74
16    TN             1.73
17    AZ             1.72
18    OK             1.66
19    NM             1.64
20    SC             1.59
```



```
21    VA             1.54
22    UT             1.52
23    NE             1.48
24    CO             1.45
25    OH             1.43
26    OR             1.42
27    FL             1.36
28    MN             1.34
29    MI             1.30
30    PA             1.15
31    MD             1.14
32    IN             1.08
33    MA             1.04
34    NV             1.03
35    CA             0.96
36    WA             0.96
37    HI             0.85
38    NJ             0.78
39    PR             0.41
40    AK             0.00
41    CT             0.00
```

Fig. 5 and 6 Ratios of flights canceled with its percentage.

### Appendices (Code)

```
#ASSIGNMENT 5

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong


DelayDataLocation <- "https://raw.githubusercontent.com/kiat/R-Examples/master/Datasets/airline/HoustonAirline.csv"

delay.dat.houston <- read.csv(DelayDataLocation,

                         header=TRUE,

                         stringsAsFactors = FALSE)
```

```
airportDataLocation <- "https://raw.githubusercontent.com/kiat/R-Examples/master/Datasets/airline/airports.csv"
airports <- read.csv(airportDataLocation,
                     header=TRUE,
                     stringsAsFactors = FALSE)
```

```
#Task 1: How many flights are in this dataset from Houston to city of "Los Angels"? Print the number.
num.flights.LA <- delay.dat.houston %>%
  filter(Dest == "LAX") %>%
  nrow()

print (as.integer(num.flights.LA))
```

```
#Task 2:  Which top-10 destination cities have the greatest number of flights?
delay.dat.houston %>%
  left_join(airports, by = c("Dest" = "iata")) %>%
  group_by(city) %>%
  summarise(
    NFlights = n()
  ) %>%
arrange(desc(NFlights)) %>%
  slice_head(n=10)
```

```
#Task 3:  Which states have no flights?
all.us.states <- airports %>%
  select(iata, state) %>%
  distinct()

states.with.flights <- delay.dat.houston %>%
  left_join(all.us.states, by = c("Dest" = "iata")) %>%
  distinct(state)

states.without.flights <- all.us.states %>%
  distinct(state) %>%
  filter(!state %in% states.with.flights$state) %>%
  arrange(state)

print(states.without.flights)



#Task 3 (changed question):To which states do we have direct flights?
states.with.direct.flights <- delay.dat.houston %>%
  left_join(airports, by = c("Dest" = "iata")) %>%
  distinct(state) %>%
  arrange(state) %>%
  mutate(Number = row_number()) %>%
  select(Number, State = state)

print(states.with.direct.flights)
```

```
#Task 4: Which top 10 destination cities have the most cancelations of flights?
delay.dat.houston %>%
  filter(Cancelled == 1) %>%
  left_join(airports, by = c("Dest" = "iata")) %>%
  group_by(city) %>%
  summarise(CancelledFlights = n()) %>%
  arrange(desc(CancelledFlights)) %>%
  slice_head(n = 10)


#Task 5: Is there any Cities without a Flight from Houston?
iata.with.flights <- unique(delay.dat.houston$Dest)

all.us.cities <- airports %>%
  filter(nchar(iata) == 3, iata != "") %>%
  select(city, iata) %>%
  distinct()

all.us.cities <- all.us.cities %>%
  mutate(has.flight = iata %in% iata.with.flights)

cities.without.flights <- all.us.cities %>%
  group_by(city) %>%
  summarise(any.flight = any(has.flight)) %>%
  filter(!any.flight) %>%
  arrange(city)
```

```
cat("Total cities without flights from Houston:", nrow(cities.without.flights), "\n")


#Task 6: What is the ratio of flights canceled for each state?
flights.with.states <- delay.dat.houston %>%
  left_join(airports, by = c("Dest" = "iata"))

cancel.percentage.by.state <- flights.with.states %>%
  group_by(state) %>%
  summarise(
    total.flights = n(),
    canceled.flights = sum(Cancelled == 1)
  ) %>%
  filter(!is.na(state) & total.flights > 0) %>%
  transmute(
    state,
    canceled.flights = round(100 * canceled.flights / total.flights, 2)
  ) %>%
  arrange(desc(canceled.flights))

print(as.data.frame(cancel.percentage.by.state))
```