*The University of Texas at Austin*

*Zhaowen Fan*

*Rafael Ignacio Gonzalez Chong*

*Assignment 5*

GSI Intro to Big Data and Data Mining

Table of Contents

[Task 1: How many flights are in this dataset from Houston to city of "Los Angeles"? Print the number. 2](#_Toc204156202)

[Task 2:  Which top-10 destination cities have the greatest number of flights? 2](#_Toc204156203)

[Task 3:  Which states have no flights? 2](#_Toc204156204)

[Task 4: Which top 10 destination cities have the most cancelations of flights? 3](#_Toc204156205)

[Task 5: Is there any Cities without a Flight from Houston? 3](#_Toc204156206)

[Task 6: What is the ratio of flights canceled for each state? 4](#_Toc204156207)

[Appendices (Code) 4](#_Toc204156208)

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

A close-up of a data set

AI-generated content may be incorrect.

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?**

A screenshot of a computer

AI-generated content may be incorrect.

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?**

A screenshot of a computer

AI-generated content may be incorrect.

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?**

A screenshot of a computer

AI-generated content may be incorrect.

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?**

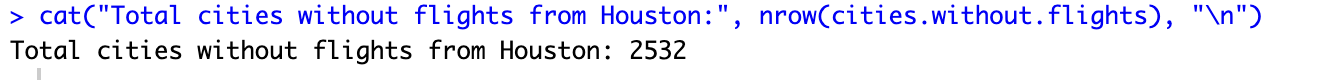


Fig. 5 Cities without flights from Houston.

There are 2532 flights that not departure from Houston.

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

A screenshot of a computer

AI-generated content may be incorrect.A table of numbers with letters and numbers

AI-generated content may be incorrect.

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))