STAT 480 Group Project

Airline Data Analysis: 1998 & 2002

Yutong Wang, Yihang Zhang, Brandon Nsiah-Ababio, Siddharth Ahuja

Goal of Project

The general goal of the project is to extract interesting information, trends and comparisons about flights in the United States during the years of 1998 and 2002.

To achieve this goal we will explore the following areas:

- Trends in cancellations of flights
- Trends in delayed arrival or departure of flights
- Trends in number of flights
- Trends in diversions of flights
- Trends of flights by region

Data Description

Setup for Project

Downloading the necessary data (done in terminal):

```
wget https://raw.githubusercontent.com/coatless/stat490uiuc/master/airlines/airlines_data.sh
chmod u+x airlines_data.sh
./airlines_data.sh 2002 2002
mv airlines.csv groupproject.csv
./airlines_data.sh 1998 1998
tail -n+2 airlines.csv >> groupproject.csv
```

Deleting columns with all or most data missing (done in terminal):

```
cut -d ',' -f 23,25,26,27,28,29 --complement groupproject.csv > groupprojectcl.csv
```

Creating the necessary tables (done in hive):

Creating table with flight data:

```
CREATE TABLE flights (Year INT, Month INT, DayofMonth INT,  DayOfWeek INT, DepTime INT,
CRSDepTime INT, ArrTime INT, CRSArrTime INT, UniqueCarrier STRING, FlightNum INT, TailNum STRING,
ActualElapsedTime INT, CRSElapsedTime INT, AirTime INT, ArrDelay INT, DepDelay INT, Origin STRING,
Dest STRING, Distance INT, TaxiIn INT, TaxiOut INT,Cancelled INT, Diverted INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH 'groupprojectcl.csv'
OVERWRITE INTO TABLE flights;

ALTER TABLE flights set tblproperties("skip.header.line.count"="1");
```

Creating table with airport data:

```
CREATE TABLE airports (iata STRING, airport STRING, city STRING, state STRING,
country STRING, lat DOUBLE, long DOUBLE)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH 'airports.csv'
OVERWRITE INTO TABLE airports;

ALTER TABLE airports set tblproperties("skip.header.line.count"="1");
```

Loading data into a big matrix (done in R):

```
#install.packages('biganalytics')
#install.packages('foreach')
library(biganalytics)
```

```
## Loading required package: bigmemory
```

```
## Loading required package: foreach
```

```
## Loading required package: biglm
```

```
## Loading required package: DBI
```

```
library(foreach)

## Function to apply to a CSV file to convert a list of columns to integer index values.
##
#convertCSVColumns <- function(file, collist){
#  fulldata<-read.csv(file)
#   for (i in collist) {
#     fulldata[,i]<-convertColumn(fulldata[,i])
#   }
#   write.csv(fulldata, file, row.names=FALSE)
# }
# # The following function is called by convertCSVColumns. It converts a single
# #column to integer indices.
# convertColumn <- function(values){
#   allvals<-as.character(values)
#   valslist<-sort(unique(allvals))
#   xx<-factor(allvals, valslist, labels=1:length(valslist))
#   rm(allvals)
#   rm(valslist)
#   gc()
#   as.numeric(levels(xx))[xx]
# }
#
# # Now use the function on the data.
# convertCSVColumns("groupprojectcl.csv", c(9,11,17,18))
#
# x <- read.big.matrix("groupprojectcl.csv", header = TRUE,
#                      backingfile = "gp.bin",
#                      descriptorfile = "gp.desc",
#                      type = "integer")

x <- attach.big.matrix('gp.desc')
```

Initial Observations

Missing airlines ‘American Eagle’

Cancellation Trends

```
year = split(1:nrow(x), x[, 'Year'])
cancelA = foreach(i = year, .combine = cbind) %do% {
  CancelledCount = sum(x[i, 'Cancelled'])
  Total = length(x[i, 'Cancelled'])
  list(CancelledCount = CancelledCount, Total = Total)
}
colnames(cancelA) = c('1998', '2002')
cancelA
```

```
##                1998    2002
## CancelledCount 144509  65143
## Total          5384721 5271359
```

```
cancelB = foreach(i = year, .combine = cbind) %do% {
  CancelledPercent = (sum(x[i, 'Cancelled'])/(length(x[i, 'Cancelled'])))*100
  list(CancelledPercent = CancelledPercent)
}
colnames(cancelB) = c('1998', '2002')
cancelB
```

```
##                   1998     2002
## CancelledPercent 2.683686 1.235791
```

```
# Plot with Cancelled Count and Percent on same bar graph, two y-axis?
```

by Month

1998

```
month1998 = split(1:sum(x[, 'Year'] == 1998), x[x[, 'Year'] == 1998, 'Month'])

month1998A = foreach(i = month1998, .combine=cbind) %do% {
  a = sum(x[i,'Cancelled'])
  b = length(x[i, 'Cancelled'])
  c = (a/b)*100
  list(CancelledCount1998 = a, Total1998 = b, CancelledPercent1998 = c)
}
colnames(month1998A) = c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'D
month1998A
```

```
##                       Jan      Feb      Mar      Apr      May      Jun
## CancelledCount1998   7414     4529     6128     4624     4598     7571
## Total1998            452001   412832   459703   442644   449293   446427
## CancelledPercent1998 1.640262 1.097056 1.333035 1.044632 1.023386 1.69591
##                       Jul      Aug     Sep       Oct      Nov      Dec
## CancelledCount1998   6012     5337    3900      4655     4473     9163
## Total1998            462429   465910  443901    457954   436528   455099
## CancelledPercent1998 1.300091 1.1455  0.8785743 1.016478 1.024677 2.013408
```

2002

```
month2002 = split(1:sum(x[, 'Year'] == 2002), x[x[, 'Year'] == 2002, 'Month'])

month2002A = foreach(i = month2002, .combine=cbind) %do% {
  a = sum(x[i,'Cancelled'])
  b = length(x[i, 'Cancelled'])
```

```
  c = (a/b)*100
  list(CancelledCount2002 = a, Total2002 = b, CancelledPercent2002 = c)
}
colnames(month2002A) = c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', '
month2002A
```

```
##                       Jan      Feb      Mar      Apr      May       Jun
## CancelledCount2002   7301     4323     6033     4513     4442      7666
## Total2002            436336   399535   447896   438141   450046    448333
## CancelledPercent2002 1.673252 1.082008 1.346964 1.030034 0.9870102 1.70989
##                       Jul      Aug      Sep      Oct      Nov      Dec
## CancelledCount2002   6260     5339     3686     4549     3675     7356
## Total2002            465573   466764   429996   446590   415024   427125
## CancelledPercent2002 1.34458  1.143833 0.8572173 1.018608 0.885491 1.722212
```

by Day

1998

```
day1998 = split(1:sum(x[, 'Year'] == 1998), x[x[, 'Year'] == 1998, 'DayOfWeek'])

day1998A = foreach(i = day1998, .combine=cbind) %do% {
  a = sum(x[i,'Cancelled'])
  b = length(x[i, 'Cancelled'])
  c = (a/b)*100
  list(CancelledCount1998 = a, Total1998 = b, CancelledPercent1998 = c)
}
colnames(day1998A) = c('Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat')
day1998A
```

```
##                       Sun      Mon      Tue      Wed      Thu      Fri
## CancelledCount1998   10163    10035    10124    10231    9701     8692
## Total1998            788003   789241   789076   796404   782299   694528
## CancelledPercent1998 1.289716 1.271475 1.28302  1.284649 1.240063 1.251497
##                       Sat
## CancelledCount1998   9458
## Total1998            745170
## CancelledPercent1998 1.269241
```

```
day2002 = split(1:sum(x[, 'Year'] == 2002), x[x[, 'Year'] == 2002, 'DayOfWeek'])

day2002A = foreach(i = day2002, .combine=cbind) %do% {
  a = sum(x[i,'Cancelled'])
  b = length(x[i, 'Cancelled'])
  c = (a/b)*100
  list(CancelledCount2002 = a, Total2002 = b, CancelledPercent2002 = c)
}
colnames(day2002A) = c('Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat')
day2002A
```

```
##                       Sun      Mon      Tue      Wed      Thu      Fri
## CancelledCount2002   9251     12139    11420    12427    8507     5967
## Total2002            774285   780556   769799   768973   771174   674222
## CancelledPercent2002 1.19478  1.555173 1.483504 1.616052 1.103123 0.8850201
##                       Sat
## CancelledCount2002   5432
```

```
## Total2002            732350
## CancelledPercent2002 0.7417219
```

Plot Month

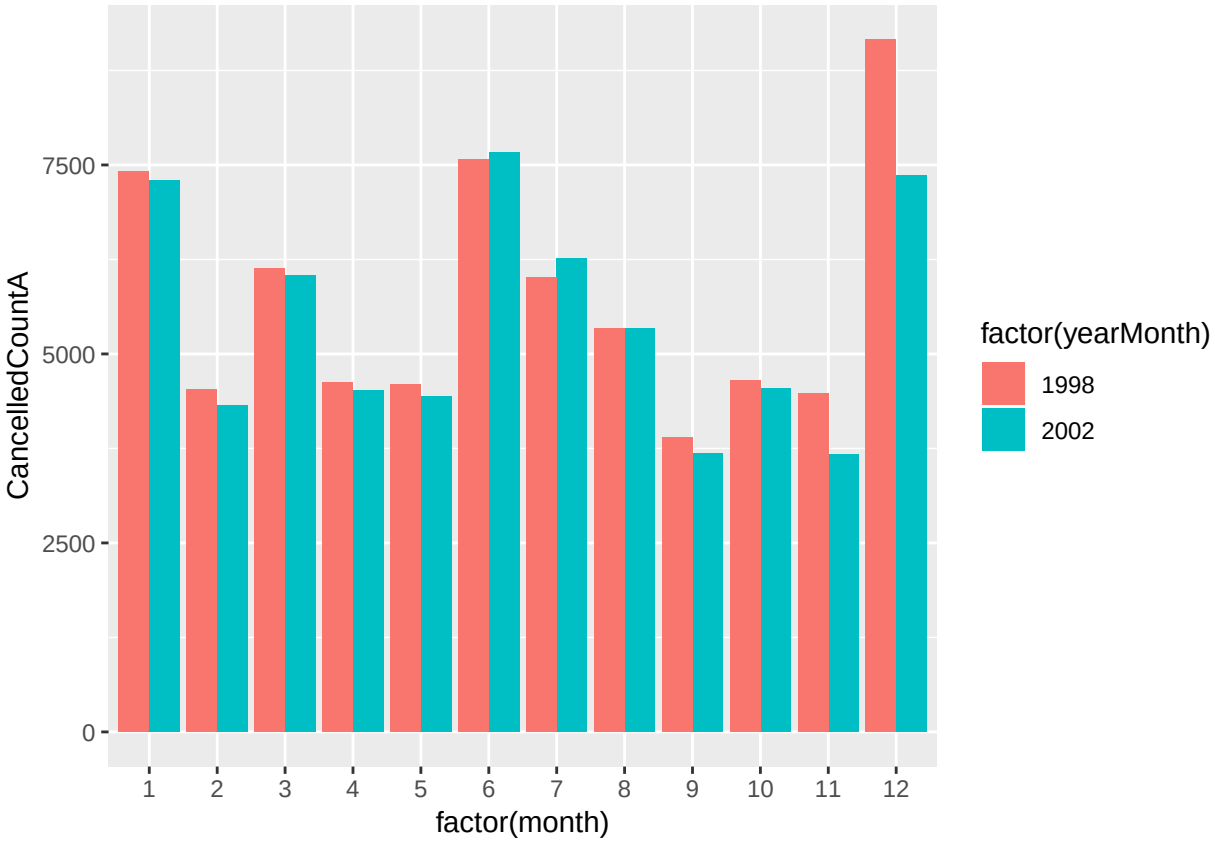
```
#install.packages('ggplot2')

library(ggplot2)
CancelledCountA = unlist(c(month1998A[1,], month2002A[1,]))
month = rep(seq(1:12), times = 2)
yearMonth = rep(c(1998, 2002), each = 12)

dfMonthCount = as.data.frame(t(rbind(yearMonth, month, CancelledCountA)))
dfMonthCount
```

```
##       yearMonth month CancelledCountA
## Jan        1998     1            7414
## Feb        1998     2            4529
## Mar        1998     3            6128
## Apr        1998     4            4624
## May        1998     5            4598
## Jun        1998     6            7571
## Jul        1998     7            6012
## Aug        1998     8            5337
## Sep        1998     9            3900
## Oct        1998    10            4655
## Nov        1998    11            4473
## Dec        1998    12            9163
## Jan.1      2002     1            7301
## Feb.1      2002     2            4323
## Mar.1      2002     3            6033
## Apr.1      2002     4            4513
## May.1      2002     5            4442
## Jun.1      2002     6            7666
## Jul.1      2002     7            6260
## Aug.1      2002     8            5339
## Sep.1      2002     9            3686
## Oct.1      2002    10            4549
## Nov.1      2002    11            3675
## Dec.1      2002    12            7356
```

```
ggplot(dfMonthCount, aes(x = factor(month), y = CancelledCountA, fill = factor(yearMonth))) + geom_bar(s
```


CancelledCountA
7500
5000
2500
0
1
2
3
4
5
6
7
8
9
10
11
12
factor(month)
factor(yearMonth)
1998
2002