Biostat 625 Final Project - Post 9/11 Flight Delay

https://github.com/chelleonis/flight_delay

Ralph Jiang, Xuelin Gu, Allen Li

December 18, 2019

Introduction

Delayed flights are a common occurance in the airline industry, with 25 million flights being delayed (for at least 15 minutes) for 20 years of data. The issue of delayed flights is seemingly unpredictable with so many factors preceding a successful flight. For the customer then, it is of increased importance to be able to anticipate what may cause a delay in their flight. To try to shed some light on the issue, we analyzed a large dataset from the post 9/11 era (2001-2008), given by the bureau of transportation statistics. The cuases of

Methods

Data Processing and Storage

Our dataset contains, at its base, 29 variables and 130 million entries of flight data, containing variables such as flight distance, arrival delay, and calendar month. We trimmed the 12 gigabyte dataset into our years of interest, reducing its size to around 4 GB (?).

We then combined weather data corresponding with the data at the 58 largest airports, including variables such as tempearture, rain (categorical), and wind speed.

biostat cluster using unix script - pre-prep and combining the separate data years (cp in code a bit later), but rbind in the cluster worked, merging the datafiles together to

read separate files in the dataset folder and conbine their rows

```
##read separate files in the dataset folder and conbine their rows
dataFiles = list.files(pattern = "*.csv") %>%
  lapply(read.csv, stringsAsFactors=F) %>%
  bind_rows
write.csv(dataFiles,file='out.csv')
```

Data cleaning and preprocessing

There are 72 variables in the table. Firstly, we deleted the cancelled and diverted flights which may have different situation with other common delayed flights. Secondly,to make the analysis more efficiently, we deleted some of the variables not included in the following analysis and make the table smaller for several reasons: 1 providing little information or providing information contained in other variables; 2 not included in this project objectives; 3 high ratio missing data. Third, we add covariate "season" based on "Month" value.

```
list_file = data.table::fread("out.csv")

dat = filter(list_file, Cancelled==0&Diverted==0)
dat = subset(dat, -c(V1))
dat = select(dat,-c(key2, key, DepTime, `STN---.x`, YEARMODA.x, `STN---.y`, YEARMODA.y,     #reason 1
                    CRSDepTime, ArrTime, CRSArrTime, ActualElapsedTime, CRSElapsedTime, #reason 1
                    Cancelled, CancellationCode, Diverted, #reason 2
                  CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, STP.x, STP.y,
#create season covariate
dat = mutate(dat, season = NA)
dat[which(dat$Month==1|dat$Month==2|dat$Month==3),]$season="winter"
dat[which(dat$Month==4|dat$Month==5|dat$Month==6),]$season="spring"
dat[which(dat$Month==7|dat$Month==8|dat$Month==9),]$season="summer"
dat[which(dat$Month==10|dat$Month==10|dat$Month==11),]$season="fall"
##
```

Descriptive statistics of variables.

```
##character variables:UniqueCarrier, FlightNum, TailNum, Origin, Dest, Fog.x, Rain.x, Snow.x, Hail.x, T
                    #Fog.y, Rain.y, Snow.y, Hail.y, Thunder.y, Tornado.y,season.
##numeric variables:Year, Month, DayofMonth, DayOfWeek, Distance, TaxiIn, TaxiOut, TEMP.x, DEWP.x, SLP.
                   #WDSP.x, MXSPD.x, MAX.x, MIN.x, PRCP.x, SNDP.x, TEMP.y, DEWP.y, SLP.y, VISIB.y,
                   #WDSP.y, MXSPD.y, MAX.y, MIN.y, PRCP.y, SNDP.y
##obtain descriptive statistic for numeric covaraites
select(dat, c(Year, Month, DayofMonth, DayOfWeek, Distance, TaxiIn, TaxiOut, TEMP.x, DEWP.x, SLP.x, VIS
              WDSP.x, MXSPD.x, MAX.x, MIN.x, PRCP.x, SNDP.x, TEMP.y, DEWP.y, SLP.y, VISIB.y,
              WDSP.y, MXSPD.y, MAX.y, MIN.y, PRCP.y, SNDP.y)) %>% summary()
select(dat, c(Year, Month, DayofMonth, DayOfWeek, Distance, TaxiIn, TaxiOut, TEMP.x, DEWP.x, SLP.x, VIS
              WDSP.x, MXSPD.x, MAX.x, MIN.x, PRCP.x, SNDP.x, TEMP.y, DEWP.y, SLP.y, VISIB.y,
              WDSP.y, MXSPD.y, MAX.y, MIN.y, PRCP.y, SNDP.y)) %>% var()
##obtain frequency tables for categorical covaraites
select(dat, UniqueCarrier) %>% table()
select(dat, FlightNum) %>% table() ## too many categories
select(dat, TailNum) %>% table() ## too many categories
select(dat, Origin) %>% table()
select(dat, Dest) %>% table()
select(dat, Fog.x) %>% table()
select(dat, Rain.x) %>% table()
select(dat, Snow.x) %>% table()
select(dat, Hail.x) %>% table()
select(dat, Thunder.x) %>% table()
select(dat, Tornado.x) %>% table()
select(dat, Fog.y) %>% table()
select(dat, Rain.y) %>% table()
select(dat, Snow.y) %>% table()
select(dat, Hail.y) %>% table()
select(dat, Thunder.y) %>% table()
select(dat, Tornado.y) %>% table()
select(dat, season) %>% table()
##obtain correlation coefficients of numeric covariates using complete data
select(dat, c(Year, Month, DayofMonth, DayOfWeek, Distance, TaxiIn, TaxiOut, TEMP.x, DEWP.x, SLP.x, VIS
              WDSP.x, MXSPD.x, MAX.x, MIN.x, PRCP.x, SNDP.x, TEMP.y, DEWP.y, SLP.y, VISIB.y,
              WDSP.y, MXSPD.y, MAX.y, MIN.y, PRCP.y, SNDP.y)) %>% cor(,use = "complete.obs")
##There is no highly correlations between pairwise covariates, except for MIN.x, MAX.x, MIN.y, MAX.y.
```

Linear regression based on descriptive result

```r
dat$UniqueCarrier = as.factor(dat$UniqueCarrier)
dat$Origin = as.factor(dat$Origin)
dat$Dest = as.factor(dat$Dest)
dat$season = as.factor(dat$season)
dat$Year = as.factor(dat$Year)
dat$Month = as.factor(dat$Month)
dat$DayOfWeek = as.factor(dat$DayOfWeek)
dat$Fog.x = as.factor(dat$Fog.x)
dat$Rain.x = as.factor(dat$Rain.x)
dat$Snow.x = as.factor(dat$Snow.x)
dat$Hail.x = as.factor(dat$Hail.x)
dat$Thunder.x = as.factor(dat$Thunder.x)
dat$Tornado.x = as.factor(dat$Tornado.x)
dat$Fog.y = as.factor(dat$Fog.y)
dat$Rain.y = as.factor(dat$Rain.y)
dat$Snow.y = as.factor(dat$Snow.y)
dat$Hail.y = as.factor(dat$Hail.y)
dat$Thunder.y = as.factor(dat$Thunder.y)
dat$Tornado.y = as.factor(dat$Tornado.y)
lr_result = biglm.big.matrix (DepDelay ~ Year + Month + DayofMonth + DayOfWeek + Distance + TaxiIn + Ta
                                DEWP.x + SLP.x + VISIB.x + WDSP.x + MXSPD.x + PRCP.x + SNDP.x + TEMP.y
                                SLP.y + VISIB.y + WDSP.y + MXSPD.y + PRCP.y + SNDP.y + UniqueCarrier +
                                Fog.x + Rain.x + Snow.x + Hail.x + Thunder.x + Tornado.x + Fog.y + Rain
                                Hail.y + Thunder.y + Tornado.y + season, data = dat )
```

bigmemory package (simplified to all number type columns, no need to use ff)
troubles with loading 9 GB file on a intel i7-6600U, 2.40 GHz, with 8GB memory had to restort to cluster commputing

(even bigmemory package would crash)

data was loaded in a variety of manners, for the parts we could break down by year, either read.table or read.csv were used.

```r
#fpath <- file.path(path,"flight_weather_cleaned.csv")
#tic("fread 6gb data import")
#flight_data <- data.table::fread(fpath)
#toc()
```

for the cleaned data, fread 6gb data import: 180.45 sec elapsed, approximately 3 minutes to import the data.

32 million observations

```r
> #fpath <- file.path(path,"flight_weather_cleaned.csv")
> #tic("fread 6gb data import")
> #flight_data <- data.table::fread(fpath)
> #toc()
>
> tic( .... [TRUNCATED]

> x_test <- read.csv("D:/bios625data/flight_weather_cleaned.csv")

> toc()
#test for read.csv: 2654.62 sec elapsed
```

each data table is around 1 gigabyte in size

Analytics

biganalytics package lme4 - Error: cannot allocate vector of size 256.0 Mb lol. . . .

Use a linear regression on Delay ~ Weather Age Year Linear miexed model on Delay ~ Weather Age + random Year

Visualization

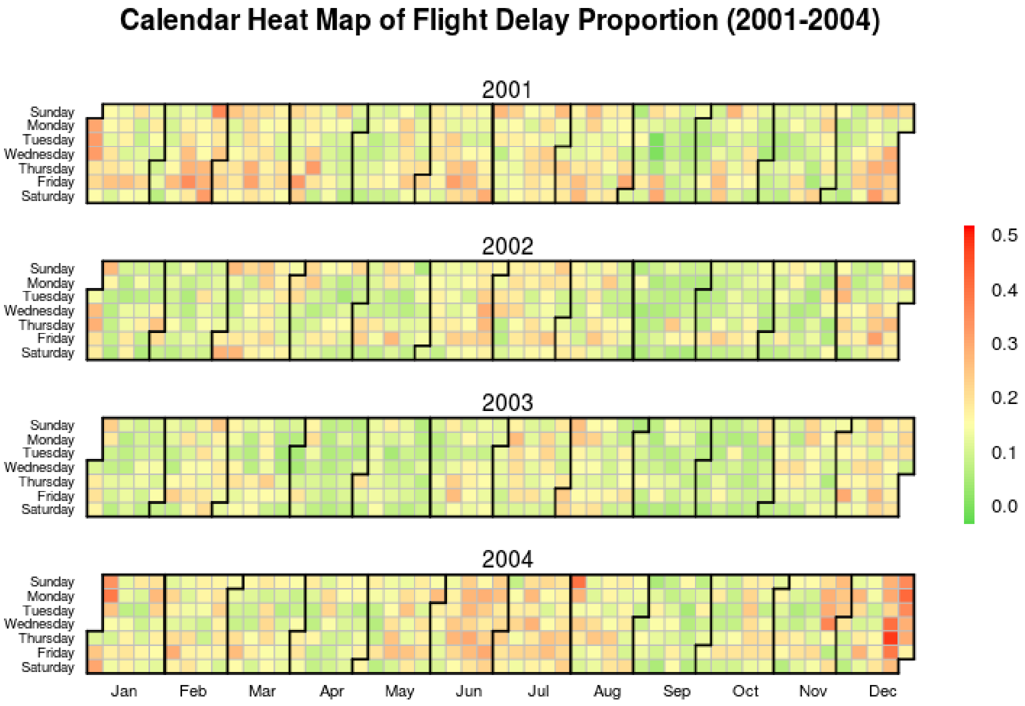
tidyverse family

Results

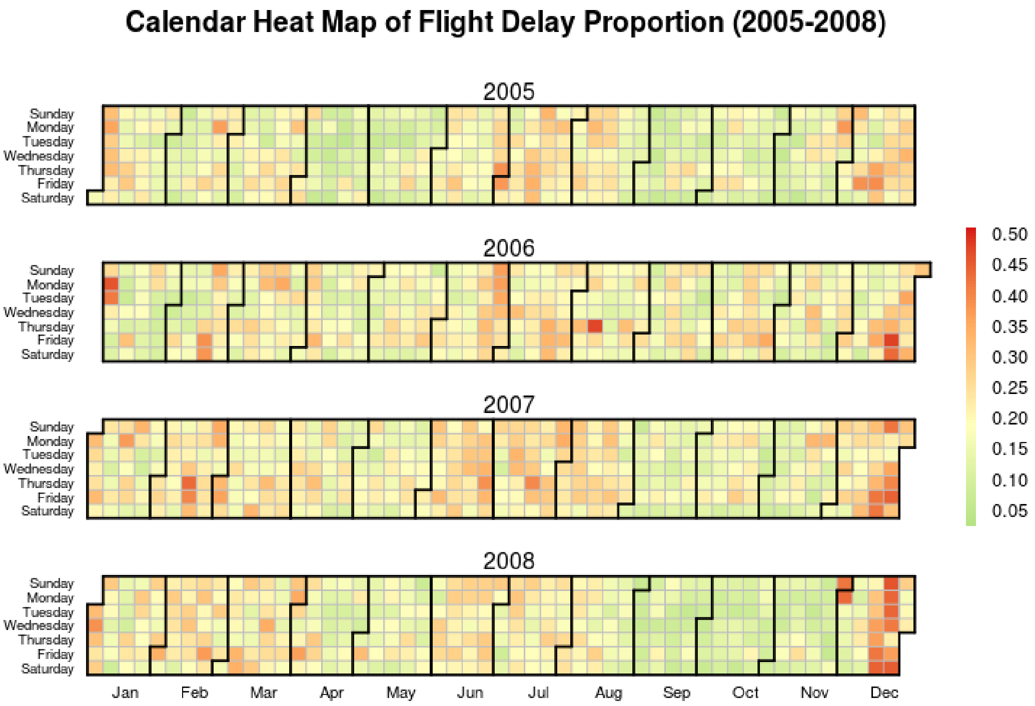
put linear regression results here

put lmm results here (if we get them)

put weather results here



analysis of this figure goes here



analysis of this figure goes here

Conclusion and Further Work

Flight delays may be indicated by the following significant factors:

Our work was comprehensive, but not exhaustive, with the following topics of interest for future investigation * Pre- 9/11 era and comparison + * Full time dataset is huge, would require different forms of data storage + having computation troubles as seen above, may need different forms + not limited to cluster computng * Investigation of different models * Dealing with missing data on a large scale: * Cross-validation