STA242 - H01

Hugh Crockford

January 23, 2013

Contents

1 Late Flights - by airline

1.1 Q: What airports are the worst for on time performance?

Examining which airports are worst for on time performance may be useful to see if there are any airports or areas which are consistently poor performers.

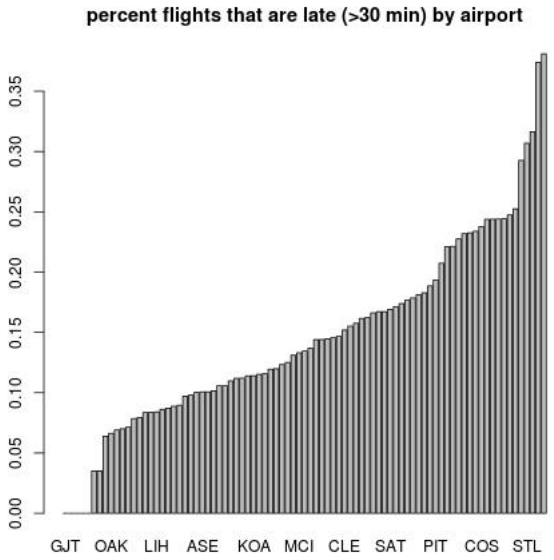

Figure 1: Percent flights that are late (more than 30 min) by airport.

1.2 Q: Is there a relationship between airport location and on time performance?

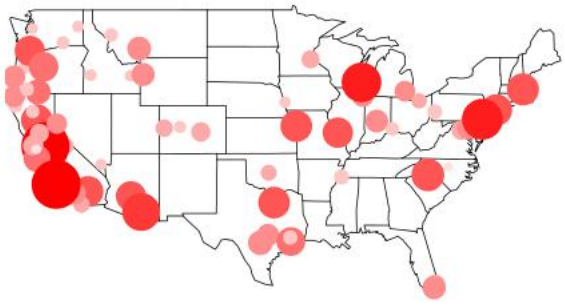

Figure 2: Map showing location of airport

The above figure shows the regional hubs(Ohare, LAX etc) are the worst for on time performance, which is to be expected as they are most likely to suffer the effects of concertina like accumulation of delays.

1.3 Q: Which airlines are the worst for on time performance?

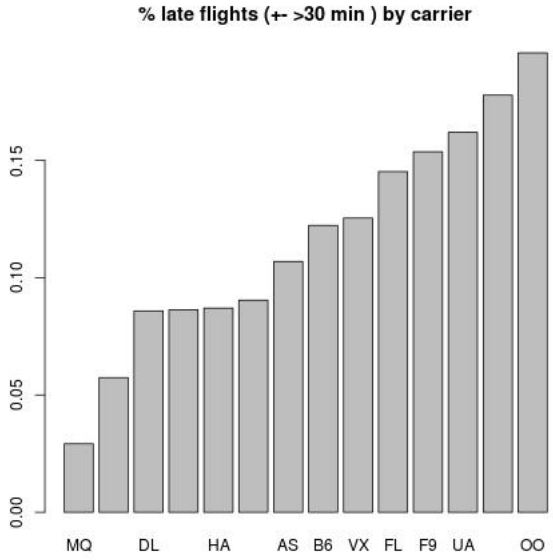

Figure 3: Late flights (+- ¿30 min) by carrier

The Above figure shows a ranking of airlines by on time percentage. An investigation of the effect of lateness threshold on on time performance follows in section 3.

2 Late flights - by time of year

2.1 Q: Is there a pattern of delays throughout the year?

An investigation into the seasonal pattern of delays, and the reason for the delays follows. The figure above does not display any strong seasonal ten-

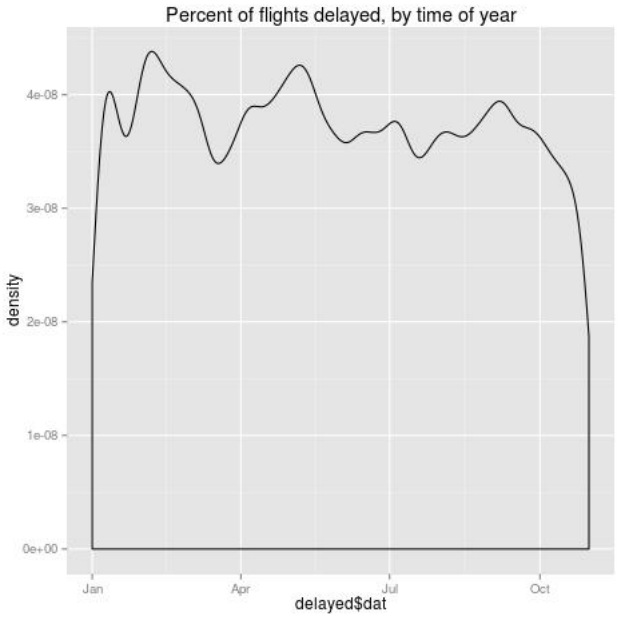

Figure 4: Percent of flight delayed, by time of year

dancies, although this analysis may be getting confounded by delay reason, i.e there actually is an increase in weather delays in winter, but there is a corresponding increase in other reasons for delays during summer which offsets this.

2.2 Q: Is there more weather delays in winter?

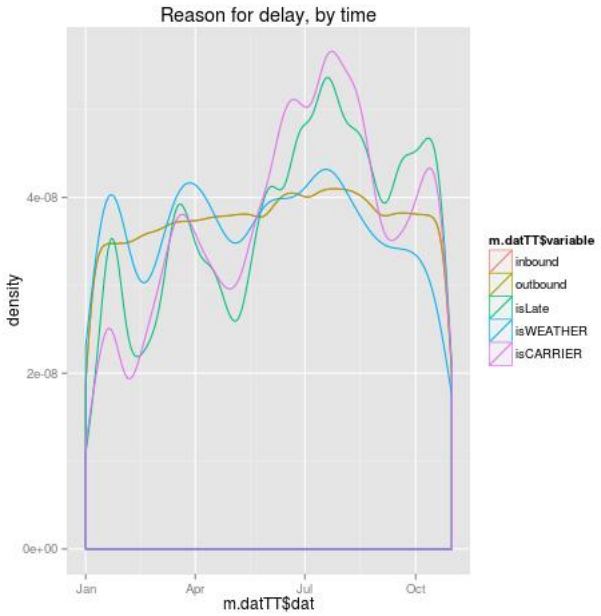

Figure 5: Patterns of delay reason - more weather delays in winter?

Increasing the resolution of delay reason and subsetting for data encoded reasons(inbound, outbound, late plane, weather, carrier) revealed there actually was an increase in weather delays during winter, as well as carrier delay (maybe some carrier delays are being miscalssified by carrier or data collecter as carrier delays when they are infact due to weather?)

3 Late flights - effect of cutoff value.

3.1 Q: Does late time cutoff affect percent planes late

The relationship between cutoff for flight delays and percent delayed by various groups was examined.
This investigation may reveal some airlines/airports which are often late by only a little and others that are rarely late, but when they are planes miss their scheduled departure/arrival time by a large margin.
The above figure shows the expected relationship between late time cutoff

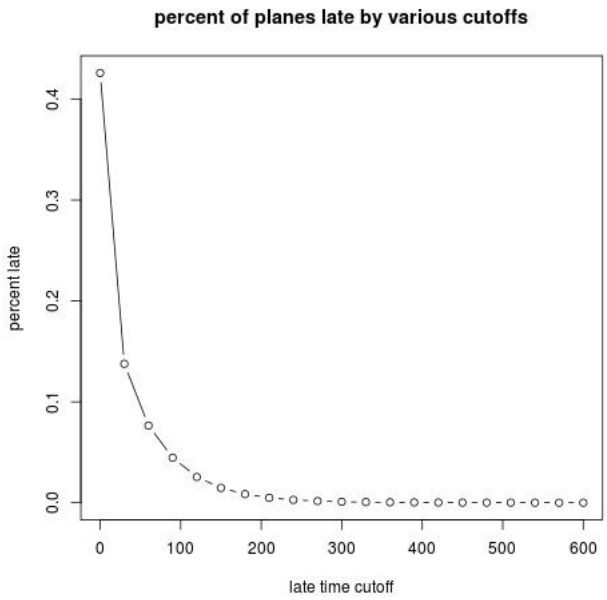

Figure 6: Percent of planes late by various cutoffs

(both arriving and departing), and the percent of late planes, i.e as the late time cutoff is increased, there are less planes classified as late.

3.2 Q: Do different airlines/airports perform better under different late cutoff points

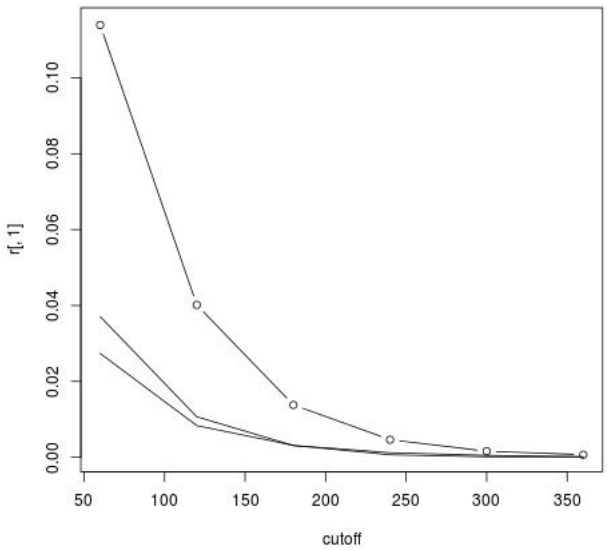

Figure 7: Plot of lateness by carrier, for different late cutoffs.

The above figure shows there is a relationship between different airlines and the effects of lateness cutoff.
(I was unable to plot this to my satisfaction, please see code)
The below figures shows there is a relationship between the destination and origin airport, and the effects of lateness cutoff

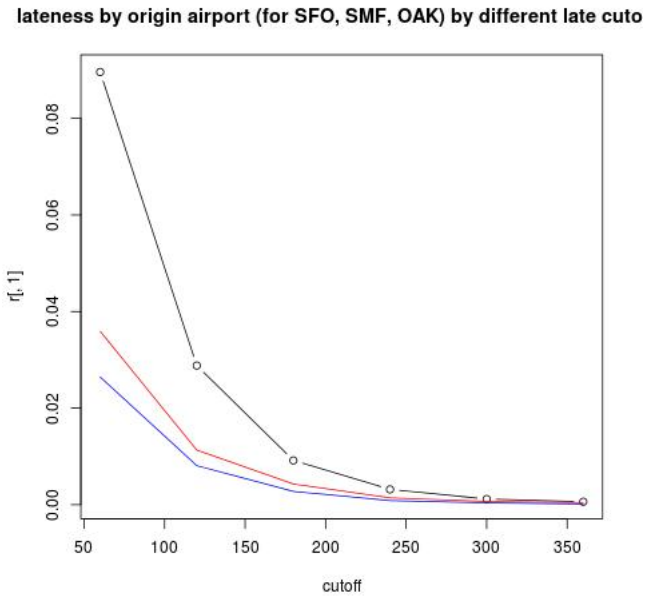

Figure 8: Plot of lateness by origin airport, for different late cutoffs.

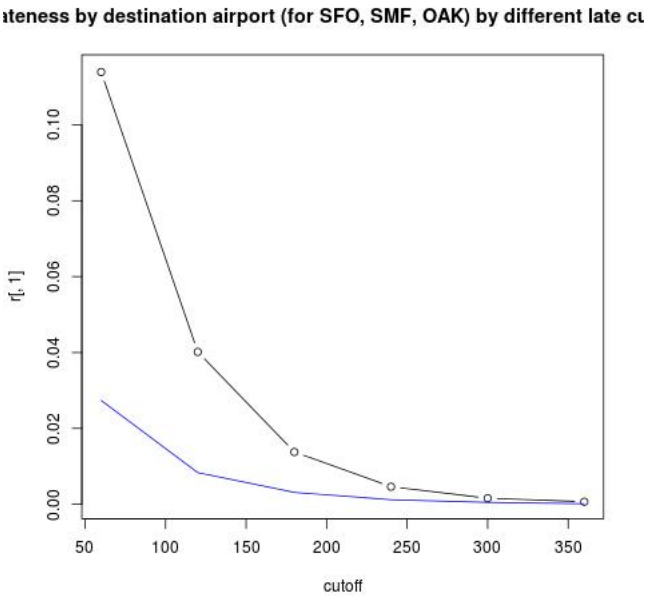

Figure 9: Plot of lateness by destination airport, for different late cutoffs.

4 CODE

```
# STA 242 HW01
# 20130109 - hughcrockford

# things to try out:
## melt/cast
## plyr - aaply - parralle
## ggplot2 + - lattice?
## sqldf? data.table?

#########################
# Questions
#########################

# delay by carrier
# delay by quarter - season?

library(lattice)
library(ggplot2)
library(reshape2)


# load(url('http://eeyore.ucdavis.edu/stat242/data/BayAreaDelays.rda'))
save("BayAreaDelays.rda")


load("BayAreaDelays.rda")
trim = BayAreaDelays[,c(2,6,9,10,15,24,30,31:32,41:43,48:52,55,57:61)]

####################
# use of airports
####################

# which airports are the worst for on time performance?

table(trim$ORIGIN)
```

```
table(trim$DEST)
trim = transform(trim, inbound =  trim$DEST %in% c("SFO","OAK","SMF"), outbound =
#  inbound vs outbound.
latelimit = 30                          #
trim$isLate = (trim$inbound == TRUE & trim$ARR_DELAY > latelimit) | (trim$outbound
llate = tapply(trim$isLate, trim$DEST , sum,na.rm = TRUE) / tapply(trim$isLate, tr


jpeg("lateCar.jpg")
lateCar = tapply(trim$isLate, trim$CARRIER , sum,na.rm = TRUE) / tapply(trim$isLat
barplot(lateCar[order(lateCar)],main = "% late flights (+- >30 min ) by carrier")
dev.off()

jpeg("pcl_airport.jpg")
barplot(late[order(late)],main = "percent flights that are late (>30 min) by airpo
dev.off()

# by map - color scale to red of lateness??
library(maps)
library(maptools)
library(RColorBrewer)
ports = readShapeSpatial("./airports/airports.shp")
map("state")
air = ports@data$LOCID
colScale = colorRampPalette(brewer.pal(3,"Reds"))
lt = (late / max(late))*10
points(ports[air %in% rownames(late),],pch = 20,col = colScale[rank(late)],cex = 1

# playt with color ramp

col = colorRamp(colors = c("white","red"))
points(ports[air %in% rownames(late),],pch = 20,col = col(late),cex = 2)

jpeg("latemap.jpg")
map("state",main = "Late flights by airport, size and color = % late")
col = colorRampPalette(c("white","red"))
points(ports[air %in% rownames(late),],pch = 20,col = col(10)[cut(late,breaks = 10
dev.off()
```

```
# size by rank lateness

###############
# how late are most of the late planes"
###################

late = function(late) {
isLate = (trim$inbound == TRUE & trim$ARR_DELAY > late) | (trim$outbound == TRUE &
lp = sum(isLate,na.rm = TRUE)/length(isLate)
return(lp)
}

llist = seq(0,600,30)
lateness = sapply(llist,late)
jpeg("lateness.jpg")
plot(llist,lateness,type = "b", main = "percent of planes late by various cutoffs'
dev.off()


late2 = function(varb,late) {
trim$isLate = (trim$inbound == TRUE & trim$ARR_DELAY > late) | (trim$outbound == T
late = tapply(trim$isLate, trim[[varb]] , sum,na.rm = TRUE) / tapply(trim$isLate,
return(late)
}

l = sapply(c("CARRIER","DEST","ORIGIN"),function(x) mapply(late2,x,seq(60,360,60))
ld =lapply(l,data.frame)
lapply(ld,function(i) names(i) = seq(60,360,60)) # didnt work?
n = seq(60,360,60)
names(ld$CARRIER) = n
names(ld$DEST) = n
names(ld$ORIGIN) = n

# need a better way of plotting this, but catagorical makes hard- 3d barplot?

jpeg("ltcarrier.jpg")
y=ld$CARRIER
```

```
y = t(y)
# sapply(1:14,function(i) lines(1:6,y[,i])) # couldnt get to work, lattice? ggplot
cutoff=n
plot(cutoff, y[,1],type = "b",main = "lateness by carrier with different late cutc
lines(cutoff,y[,2])
lines(cutoff,y[,3])
lines(cutoff,y[,4])
lines(cutoff,y[,5])
lines(cutoff,y[,6])
lines(cutoff,y[,7])
lines(cutoff,y[,8])
dev.off()

jpeg("ltdest.jpg")
y=ld$DEST
i=t(y)
r=i[,c("SFO","SMF","OAK")]
cutoff=n
plot(cutoff, r[,1],type = "b",main = "lateness by destination airport (for SFO, SM
llines(cutoff,r[,2],col="red")
lines(cutoff,r[,3],col = "blue")
dev.off()

jpeg("ltorig.jpg")
y=ld$ORIGIN
i=t(y)
r=i[,c("SFO","SMF","OAK")]
cutoff=n
plot(cutoff, r[,1],type = "b", main = "lateness by origin airport (for SFO, SMF, C
lines(cutoff,r[,2],col = "red")
lines(cutoff,r[,3], col = "blue")
dev.off()

###################33
# delay by carrier
#####################

levels(trim$CARRIER)
```

```
summary(trim$DEP_DELAY)
boxplot(trim$DEP_DELAY ~ trim$CARRIER,main = "boxplot showing level of dely by air
# prob most useful

#delay vary by time of year, some carriers better than others?
qplot(delayed$dat, color=delayed$CARRIER, geom = "density")

########################
# delay by time of year
########################3

#delay vary by time of year,?
del = transform( trim , isLate = trim$ARR_DELAY < 0 )
delayed = subset(del, del$isLate)
delayed$dat = strptime(delayed$FL_DATE,format = "%Y-%m-%d")
jpeg("delaydate.jpg")
qplot(delayed$dat,geom="density",main = "Percent of flights delayed, by time of ye
dev.off()

# pattens in delay reason, more weather delay in winter??
reasons = transform( trim, isWEATHER = trim$WEATHER_DELAY > 0, isCARRIER = trim$CA
dat = reasons[,c(2,3,24:28)]
m.datT = melt(dat,id = 1:2)
m.datTT = m.dat[m.dat$value == TRUE,] # for some reason didnt work??
m.datTT = subset(m.datT,m.datT$value)
m.datTT$dat = strptime(m.datTT$FL_DATE,format = "%Y-%m-%d")
qplot(m.datTT$dat, geom = "density")

jpeg("weather.jpg")
qplot(m.datTT$dat, color=m.datTT$variable, geom = "density",main = "Reason for del
dev.off()



# delay explained by recent flights - to what number?
late = as.numeric(del$isLate)
plot(late[200:500],type = "l")
```

```
# whichc airline takes loongest to taxi over same route - pref treatment by air tr
taxi = BayAreaDelays[,c(6,9,15,24,31,37:40,42)]
names(taxi)

# arriving flights

trim = transform(trim, inbound =  trim$DEST %in% c("SFO","OAK","SMF"), outbound =
#  inbound vs outbound.


# add in weather data - Rcurl fill form to source data?
## do like thi - cumulative days of shit weather, wind above threshold, test diff
```