# Homework 2 Question 1

*Bejan Sadeghian*

*August 17, 2015*

## Task

To do some exploratory data analysis on the provided ABIA dataset from 2008 and graph a, mostly, self explanatory story.

## My Method

What I wanted to do was find, for each carrier, if there was any extra delay time when leaving Austin Bergstrom (AUS) that seemed to have been induced by the plane's late arrival into AUS. In other words, I wanted to see if certain carriers handled less time between arrival and departure at AUS better than others.

What I did to accomplish this is I first took the dataset and found only flights that came into and left out of Austin Bergstrom (AUS) in the same day and matched them by plane tail ID. I then created a scatter plot showing, by carrier, the arrival into Austin delay time vs the delay time out of Austin.

Below is the piece of code that was the "data wrangling"

```
library(ggplot2)
library(plyr)

rawdata = read.csv('ABIA.csv')
#Question I want to answer...
#What is the impact of a plane's arrival time into Austin on its departure time?
#Aux: What is the average TAT for a carrier?
#This will of course require only flights that leave in the same day as their arrival.

flights = rawdata
#Remove any rows where the tail number is blank or NA
flights$TailNum[flights$TailNum == ''] = NA
flights = flights[complete.cases(flights$TailNum),]

#number sequentially the grouped tailnumbers
flights_seq = ddply(flights, .(TailNum), mutate, id = seq_along(TailNum))

flights_seq = flights_seq[c('TailNum','id','Month','DayofMonth','DepTime','UniqueCarrie
r','ArrDelay','DepDelay','Origin','Dest')]

flights_seq$MonthDay = paste(flights_seq$TailNum, flights_seq$Month, flights_seq$DayofMon
th, sep='_')

flights_distilled = subset(flights_seq,duplicated(flights_seq$MonthDay) | duplicated(flig
hts_seq$MonthDay, fromLast = TRUE))

flights_intoAUS = subset(flights_distilled, Origin!='AUS')
flights_outofAUS = subset(flights_distilled, Origin=='AUS')

#Merge the two together to allow for comparison of plane arrival and then plane departur
e.
Merged_flights = merge(flights_intoAUS, flights_outofAUS, by='MonthDay', type='full', al
l=TRUE, suffixes=c('_Arrive','_Depart'))
Merged_flights$ID_diff = Merged_flights$id_Depart - Merged_flights$id_Arrive

#Removing extra rows (flights that did not directly follow the arrival flight)
Merged_flights = subset(Merged_flights, ID_diff == 1)

#Select out only columns I care about (the merge resulted in a lot of extra columns)  -re
using the variable flights here
flights = Merged_flights[c('MonthDay','TailNum_Arrive','Month_Arrive','DayofMonth_Arriv
e','UniqueCarrier_Arrive','ArrDelay_Arrive','DepDelay_Arrive','Origin_Arrive','DepDelay_D
epart','Dest_Depart')]
```

# Result

As you can see from the graph below, the carrier “US” appears to have been unaffected by incoming delay time, this could be due to the limited data and over 100 minutes there may eventually be a increase but for the data from 2008 it was relatively flat. What I think this is really telling us is that their

time between flights is large enough to where it wont affect their departure time if the flight is up to 100 minutes late.

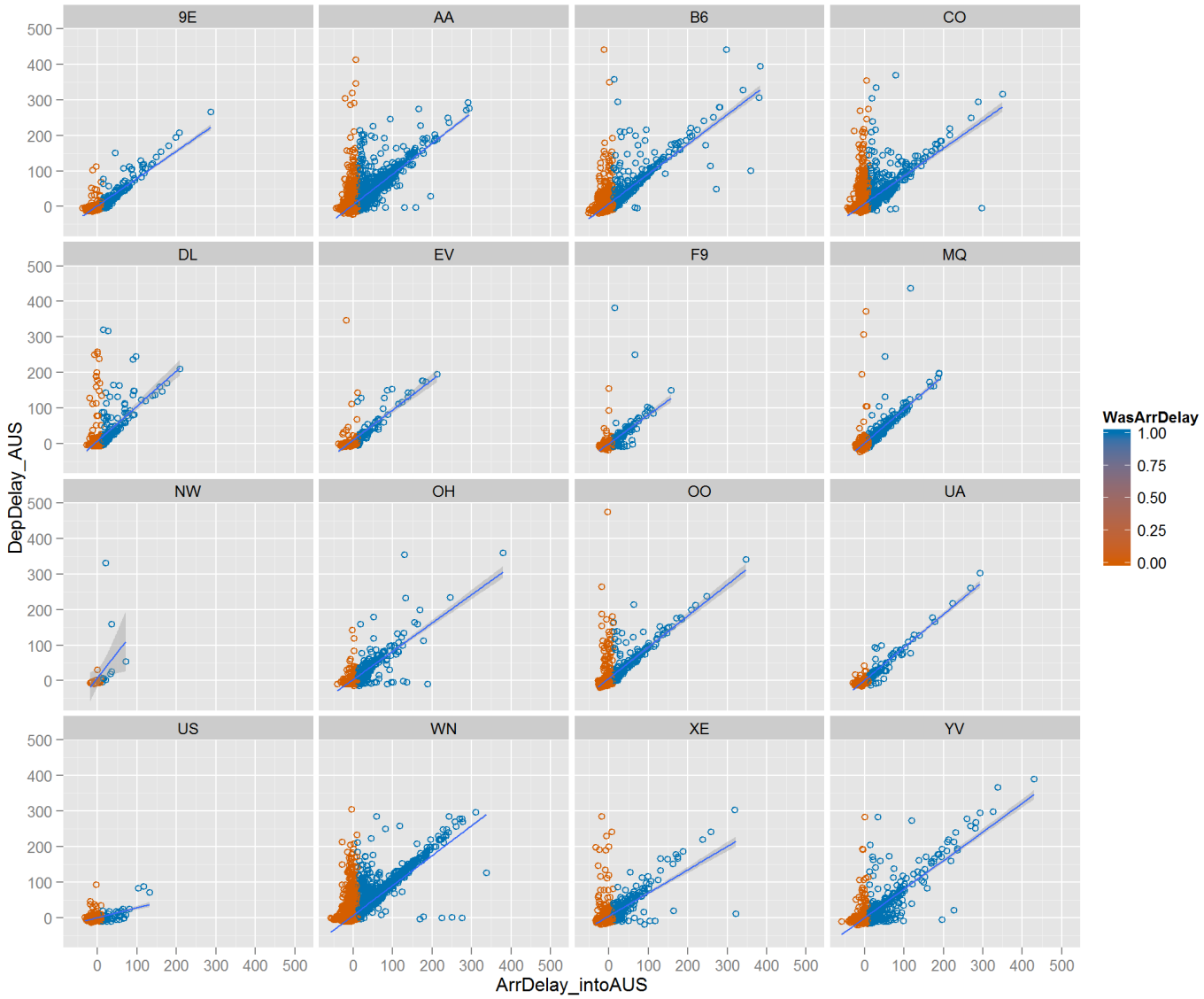
“AA”, “WN”, and a few other carriers have high departure delay times regardless of arrival delay.

You can see below that each airline has a linear reaction to the arrival delay time. The slope indicates how well the carrier react to this incoming delay time. Most airlines are about the same in this regard, “CO” might be slight better.

Finally, something to note is most bigger airlines dont seem to experience any impact to their departure time until the incoming flight is delayed by ~40 minutes.

```
library(ggplot2)
library(plyr)
library(RCurl)

data <- getURL("https://raw.githubusercontent.com/jgscott/STA380/master/data/ABIA.csv",
               ssl.verifypeer=0L, followlocation=1L)
rawdata = read.csv(text=data)

#rawdata = read.csv('ABIA.csv')
#Question I want to answer...
#What is the impact of a plane's arrival time into Austin on its departure time?
#Aux: What is the average TAT for a carrier?
#This will of course require only flights that leave in the same day as their arrival.

flights = rawdata
#Remove any rows where the tail number is blank or NA
flights$TailNum[flights$TailNum == ''] = NA
flights = flights[complete.cases(flights$TailNum),]

#number sequentially the grouped tailnumbers
flights_seq = ddply(flights, .(TailNum), mutate, id = seq_along(TailNum))

flights_seq = flights_seq[c('TailNum','id','Month','DayofMonth','DepTime','UniqueCarrie
r','ArrDelay','DepDelay','Origin','Dest')]

flights_seq$MonthDay = paste(flights_seq$TailNum, flights_seq$Month, flights_seq$DayofMon
th, sep='_')

flights_distilled = subset(flights_seq,duplicated(flights_seq$MonthDay) | duplicated(flig
hts_seq$MonthDay, fromLast = TRUE))

flights_intoAUS = subset(flights_distilled, Origin!='AUS')
flights_outofAUS = subset(flights_distilled, Origin=='AUS')

#Merge the two together to allow for comparison of plane arrival and then plane departur
e.
Merged_flights = merge(flights_intoAUS, flights_outofAUS, by='MonthDay', type='full', al
l=TRUE, suffixes=c('_Arrive','_Depart'))
Merged_flights$ID_diff = Merged_flights$id_Depart - Merged_flights$id_Arrive
```

```
#Removing extra rows (flights that did not directly follow the arrival flight)
Merged_flights = subset(Merged_flights, ID_diff == 1)

#Select out only columns I care about (the merge resulted in a lot of extra columns)  -re
using the variable flights here
flights = Merged_flights[c('MonthDay','TailNum_Arrive','Month_Arrive','DayofMonth_Arriv
e','UniqueCarrier_Arrive','ArrDelay_Arrive','DepDelay_Arrive','Origin_Arrive','DepDelay_D
epart','Dest_Depart')]

#Rename the columns to make them more managable
flights = rename(flights,c('MonthDay'='Unique_ID','TailNum_Arrive'='TailNum','Month_Arriv
e'='Month','DayofMonth_Arrive'='DayofMonth','UniqueCarrier_Arrive'='UniqueCarrier','ArrDe
lay_Arrive'='ArrDelay_intoAUS','DepDelay_Arrive'='DepDelay_origin','Origin_Arrive'='Origi
n','DepDelay_Depart'='DepDelay_AUS'))

#If the delay of arrival into austin was >10 min mark as 1, else 0
flights$WasArrDelay = ifelse(flights$ArrDelay_intoAUS > 10, 1, 0)

#Calculate the Difference between Departure Delay from AUS and Arrival Delay into AUS
flights$Dep_Arr = flights$DepDelay_AUS - flights$ArrDelay_intoAUS

#Begin plotting
ggplot(flights, aes(x=ArrDelay_intoAUS, y=DepDelay_AUS)) + geom_point(shape = 1, aes(colo
ur=WasArrDelay)) + facet_wrap( ~ UniqueCarrier) +scale_colour_gradientn(colours=c("#D55E0
0","#0072B2")) + stat_smooth(method="lm", geom="smooth", se=TRUE)
```




```
#The following three plots were mostly unfruitful, only the preceeding plot gave useful information in regards to its facet's impact to delay time.
#ggplot(flights, aes(x=ArrDelay_intoAUS, y=DepDelay_AUS)) + geom_point(shape = 1, aes(colour=WasArrDelay)) + facet_wrap( ~ Dest_Depart) +scale_colour_gradientn(colours=c("#D55E00","#0072B2")) + stat_smooth(method="lm", geom="smooth", se=TRUE)
#ggplot(flights, aes(x=ArrDelay_intoAUS, y=DepDelay_AUS)) + geom_point(shape = 1, aes(colour=WasArrDelay)) + facet_wrap( ~ Month) +scale_colour_gradientn(colours=c("#D55E00","#0072B2")) + stat_smooth(method="lm", geom="smooth", se=TRUE)
#ggplot(flights, aes(x=ArrDelay_intoAUS, y=DepDelay_AUS)) + geom_point(shape = 1, aes(colour=WasArrDelay)) + facet_wrap( ~ DayofMonth) +scale_colour_gradientn(colours=c("#D55E00","#0072B2")) + stat_smooth(method="lm", geom="smooth", se=TRUE)
```