STA 380 Exercise 2

Shreyas K.S.

Introduction

This exercise goes over 3 questions: exploratory data analysis at ABIA, author attribution for Reuters articles, and association rules in music.

Exploratory Data Analysis: Flights at ABIA

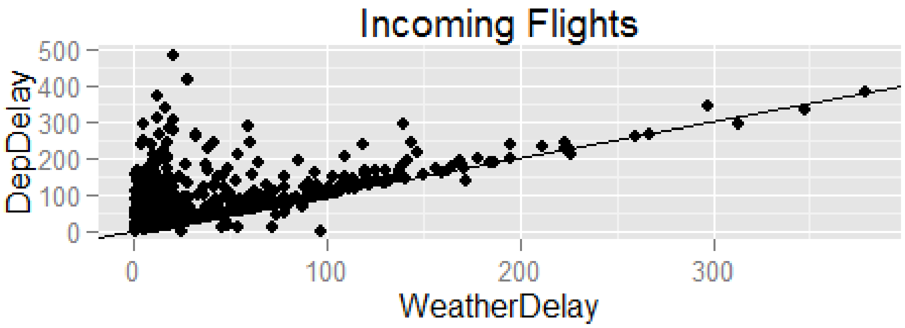
Austin Bergstrom International Airport (ABIA) is the only major domestic airport at Austin, Texas. The data set being explored here contains flights in and out of ABIA in 2008. First, lets start by loading and partitioning the data into incoming and outgoing flights from/to Austin:

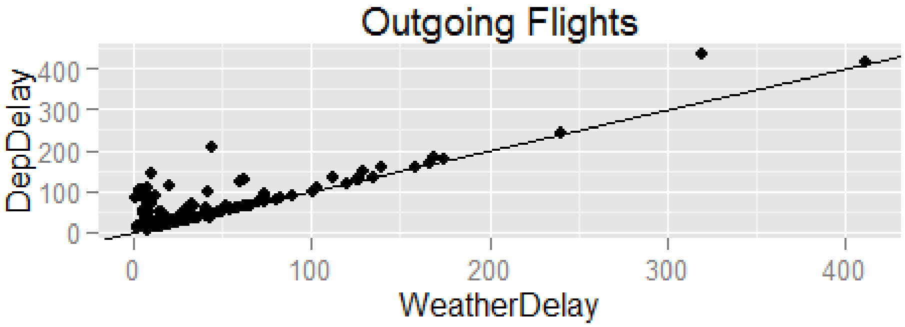
```
abia = read.csv("ABIA.csv", header=TRUE)
incoming <- subset(abia, Dest == "AUS")
outgoing <- subset(abia, Origin == "AUS")
```

Since Austin is known for its beautiful weather, I start by looking at how forecasted weather delay relates to actual delay in departure. Both these attributes correspond to the airport of origin, hence they can be compared. I start by cleaning up the data. I drop NA values in Weather Delays for incoming and outgoing flights.

```
in_weather <- subset(incoming, !is.na(incoming[,26]))
in_weather <- subset(in_weather, WeatherDelay>0)
out_weather <- subset(outgoing, !is.na(incoming[,26]))
out_weather <- subset(out_weather, WeatherDelay>0)
```

The following set of plots shows the difference in forecasted delay due to weather and actual delay in departure for incoming and outgoing flights:

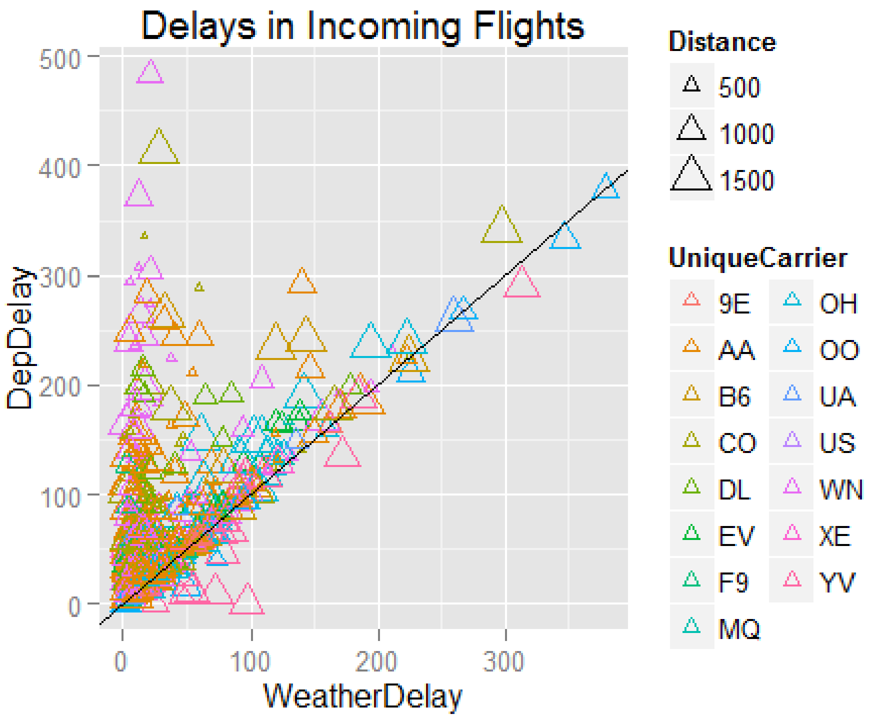




A line of slope 1 and origin 0 (Line formula:y=x) is included in the plots. Being along the line means forecasted weather delay was exactly equal to actual departure delay. A point below the line represents a flight taking off in less time than its predicted delay due to weather.

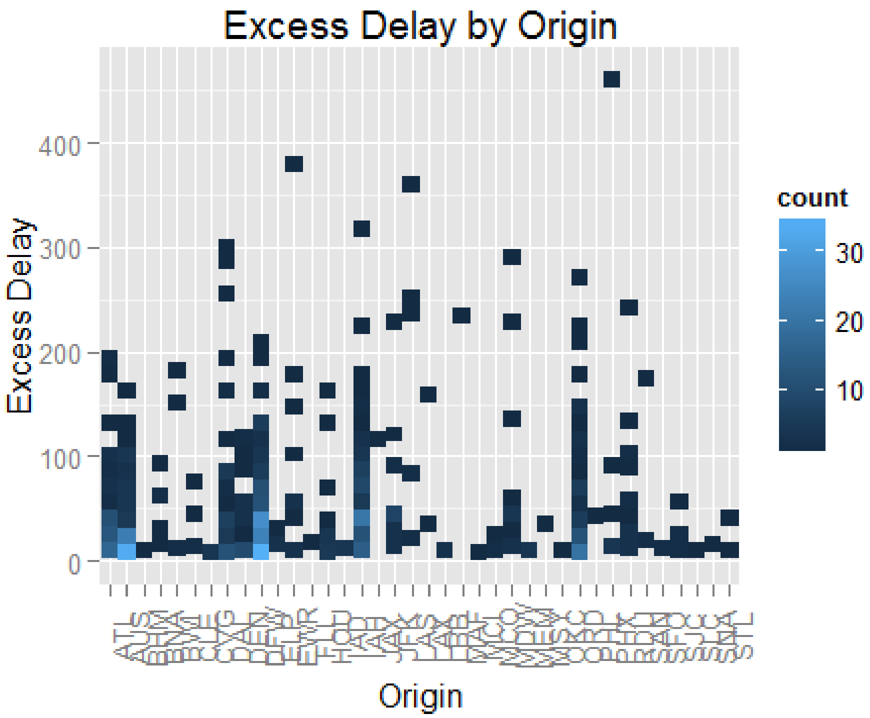
One extremely surprising observation is that for flights going out of Austin, every time there a weather delay is forecasted, the flight is delayed by the time of predicted weather delay or higher. This is helpful information for travelers out of Austin when it comes to planning their trip to the airport.

The differences between incoming and outgoing flights may be because of two reasons. An airport may generally be safer (less security delay), less busy (less overall aircrafts compared to other airports and less late aircraft delay), or have better airport staff (less carrier delay). Let's look closely at the airports with delays. Lets try to analyze this in more detail.



The above graph is similar to the visuals we saw earlier, but gives us more information. The color of each point is the carrier. We can't see many clear trends in carriers when it comes to excess delay. However, we can see that more long distance flights are involved when there is excess delay.

Lets call the difference between expected weather delay and actual departure delay excess delays. Excess delays can be represented by the vertical distance between a point and the line. Lets subset the data further to look into this. There are 612 flights which have been delayed due to weather AND other reasons. The following plot shows the excess delay (in minutes) against origin:



We can see that most of the points are towards the lower half of the plot, which means excess delays were generally less than 200 minutes. The lighter shades are seen in Austin (AUS), Dallas Fort Worth (DFW), Chicago O'Hare (ORD), George Bush Intercontinental Houston (IAH), Hartsfield Jackson Atlanta (ATL). These airports had most excess delays. This is not very surprising, since most of these airports are major international airports. Additionally, DFW, ORD, and IAH are hubs for either United or American Airlines, which are the two biggest airlines in the country. Delta uses ATL as a hub. A delay in one of these hub airports may cascade and go on to affect other flights, which is what we might be seeing here.

It is generally difficult to map out the exact reason for a cancellation or delay in airplanes. However, exploratory analysis of the data shows that in Austin, any predicted weather delay results in a delay of longer, and hub airports tend to have a higher frequency of delays.

Association Rule Mining with Groceries

The data contains rows with baskets of goods purchased by each customer. First, lets generate association rules with the apriori principle and inspect it. The support was set at 0.01, which means a combination of goods would have to be purchased 1% of overall baskets to be included in our analysis. The confidence is set at 0.6, and maximum length is 4, since I felt 60% is a reasonable confidence level for meaningful interpretation, and large baskets will not impact the analysis by much.

```
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##         0.6    0.1    1 none FALSE            TRUE    0.01      1      4
##  target   ext
##   rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## apriori - find association rules with the apriori algorithm
## version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 7011 transaction(s)] done [0.01s].
## sorting and recoding items ... [101 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4 done [0.01s].
## writing ... [6 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

269 rules have been generated. Lets subset the results of the rules in a few ways:

```
## NULL

## NULL

##   lhs                      rhs              support confidence     lift
## 1 {butter,
##    root vegetables}  => {whole milk} 0.01155327  0.6377953 2.012413
## 2 {butter,
##    yogurt}           => {whole milk} 0.01312224  0.6388889 2.015864
## 3 {other vegetables,
##    tropical fruit,
##    yogurt}           => {whole milk} 0.01069748  0.6250000 1.972041
## 4 {other vegetables,
##    root vegetables,
##    yogurt}           => {whole milk} 0.01098274  0.6111111 1.928218
```

We see that a lot of rules end up modeling combinations of items in baskets that also include dairy items such as yogurt and milk or different types of vegetables. We can see that people who buy vegetables and fruits in general tend to purchase root vegetables, and people who buy certain combinations of yogurt and fruit are likely to purchase whole milk.