STAT 585: Lab Week 6

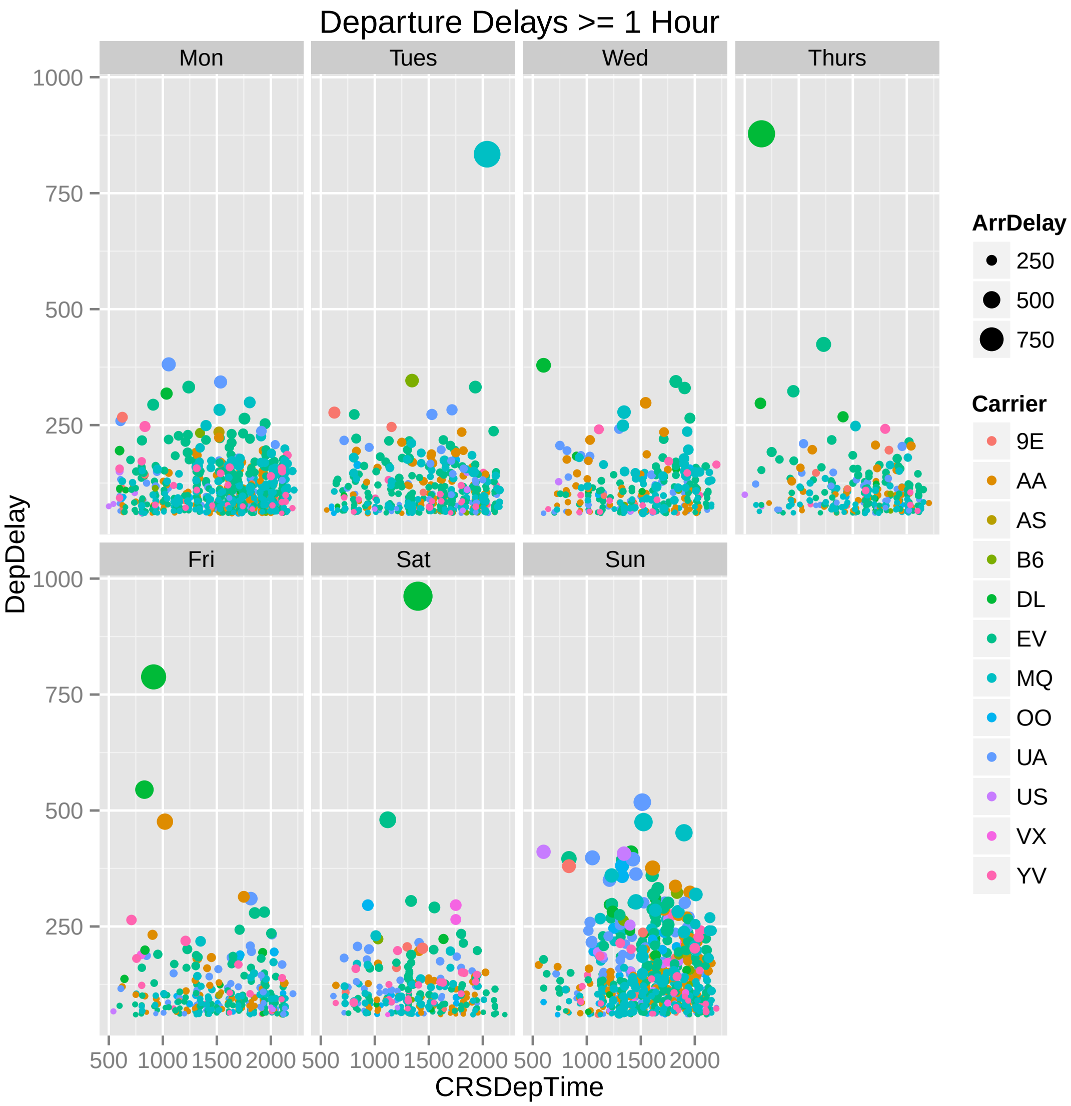
2.21.14 Connor Jennings, Yet Nguyen, Samantha Tyner

1 Delays

Is there a day of the week/time of day effect on departure or arrival delays? Provide a graphical overview of the situation. (Is there enough data to assess a carrier effect?)

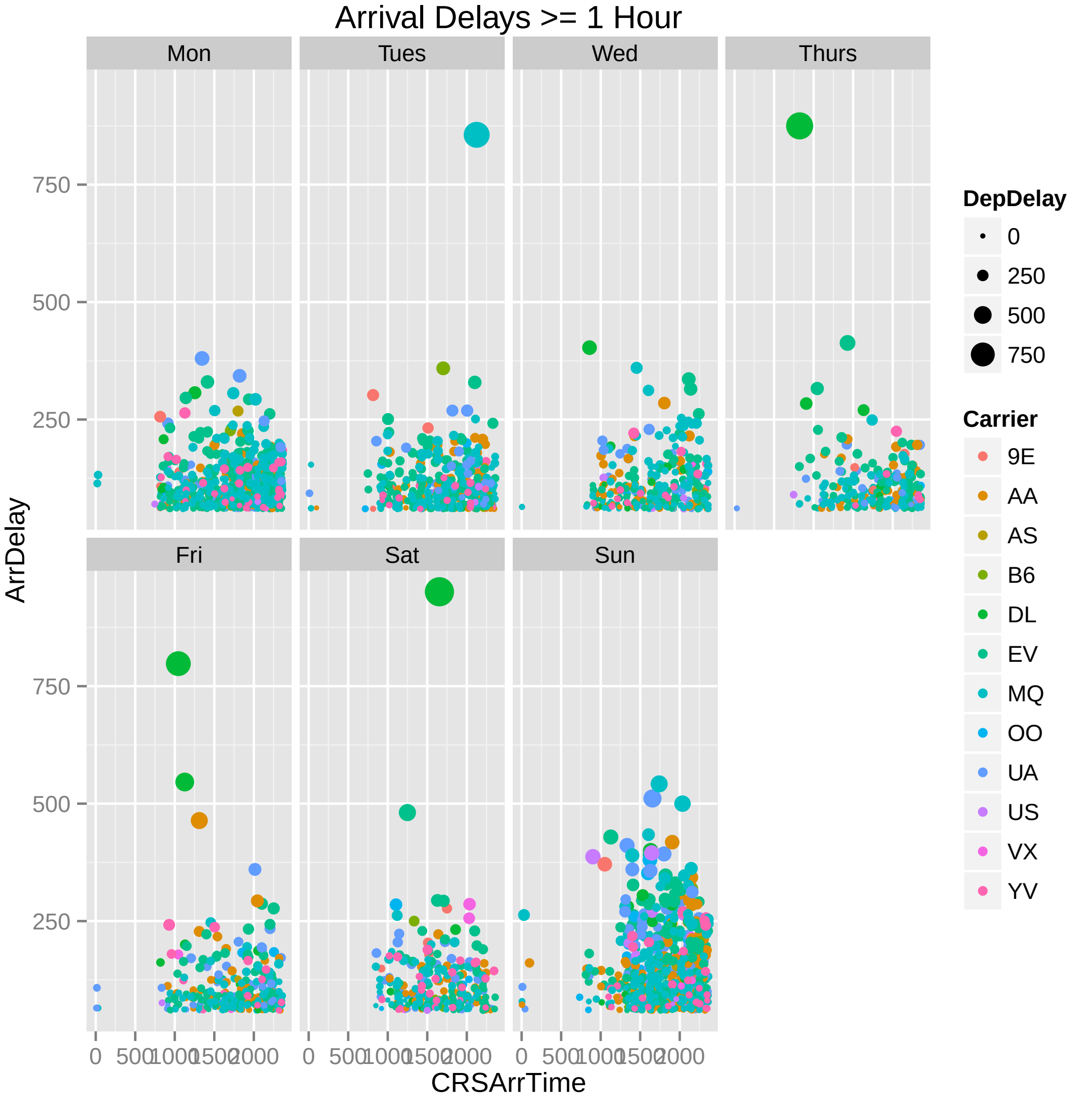
```
library(RSQLite)
library(dplyr)
library(ggplot2)
library(lubridate)
my_db <- src_sqlite("flights")
# read data into R after downloading and unzipping
# flights_nov2013<-read.csv('~/Desktop/585/nov2013flights.csv',header=T)
# flights_dec2013<-read.csv('~/Desktop/585/dec2013flights.csv',header=T)

# writing to db nov_flights_sqlite<-copy_to(my_db,flights_nov2013,
# 'flights', temporary=F) co <- dbConnect(dbDriver('SQLite'), 'flights')
# dbWriteTable(co, value=flights_dec2013, name='flights', append=TRUE,
# row.names=FALSE) dbGetQuery(co, 'select Month, count(*) from flights group
# by Month') create data table
flights <- tbl(my_db, "flights")

# making sure we got all the data flights %.% summarise(n=n()) = 1020035.
# Good!

ohare <- flights %.% filter(Origin == "ORD") %.% select(DayOfWeek, FlightDate,
    Carrier, Origin, Dest, DepDelay, ArrDelay, Cancelled, CancellationCode,
    CRSDepTime, DepTime, CRSArrTime, Distance)

ohare <- collect(ohare)
head(ohare$DayOfWeek)

## [1] 7 7 7 1 1 1

# Figure out labeling for graphs
ohare$FlightDate <- as.Date(ohare$FlightDate)
head(wday(ohare$FlightDate, label = T))

## [1] Sun Sun Sun Mon Mon Mon
## Levels: Sun < Mon < Tues < Wed < Thurs < Fri < Sat

# So 7 = Sunday
ohare$DayOfWeek <- as.factor(ohare$DayOfWeek)
levels(ohare$DayOfWeek) <- c("Mon", "Tues", "Wed", "Thurs", "Fri", "Sat", "Sun")

ohare_depdelays <- filter(ohare, DepDelay >= 60)
ohare_arrdelays <- filter(ohare, ArrDelay >= 60)
```

```r
# Plot Departure Delays & Arrival delays colored by Carrier
qplot(data = ohare_depdelays, y = DepDelay, x = CRSDepTime, colour = Carrier,
    size = ArrDelay, na.rm = T, main = "Departure Delays >= 1 Hour") + facet_wrap(~DayOfWeek,
    nrow = 2)
```



The above plot shows the scheduled departure time on the x axis and the departure delay on the y axis for all flights leaving O'Hare airport that were delayed for one hour or more. We chose to plot only delays (and arrivals in the next plot) of one hour or more because we decided that for us, a delay of 1 hour becomes a serious inconvenience as opposed to a mild annoyance. The size of the point represents the corresponding arrival delay of the flight, and the color represents the airline. The airlines with the most and largest departure delays are Delta (DL), Sky West (OO), and American Airlines (AA). The plots show that on average, there are more departure delays in the evening, which is evident by the increased number of points as the time increases on the x axis across all days. Addtionally, the most departure delays greater than 1 hour occur on Sunday and Monday. We think this could simply be due to there being more flights those days, or because more people flying those days are returning from a vacation to make it back to work the next day, so they aren't as experienced as the people flying on a Tuesday or Wednesday. Finally, note that there are no departure delays before 5am on any of the days, which makes sense because there are probably not many flights at that time, so fewer conflicts occur.

```
qplot(data = ohare_arrdelays, y = ArrDelay, x = CRSArrTime, colour = Carrier,
    size = DepDelay, na.rm = T, main = "Arrival Delays >= 1 Hour") + facet_wrap(~DayOfWeek,
    nrow = 2)
```



The conclusions drawn from the arrival delay plots are pretty much the same as the ones from the departure delay plots. The one noticeable difference here is that there are arrival delays before 5am because they are carried over from the late night departure delays from the previous day.

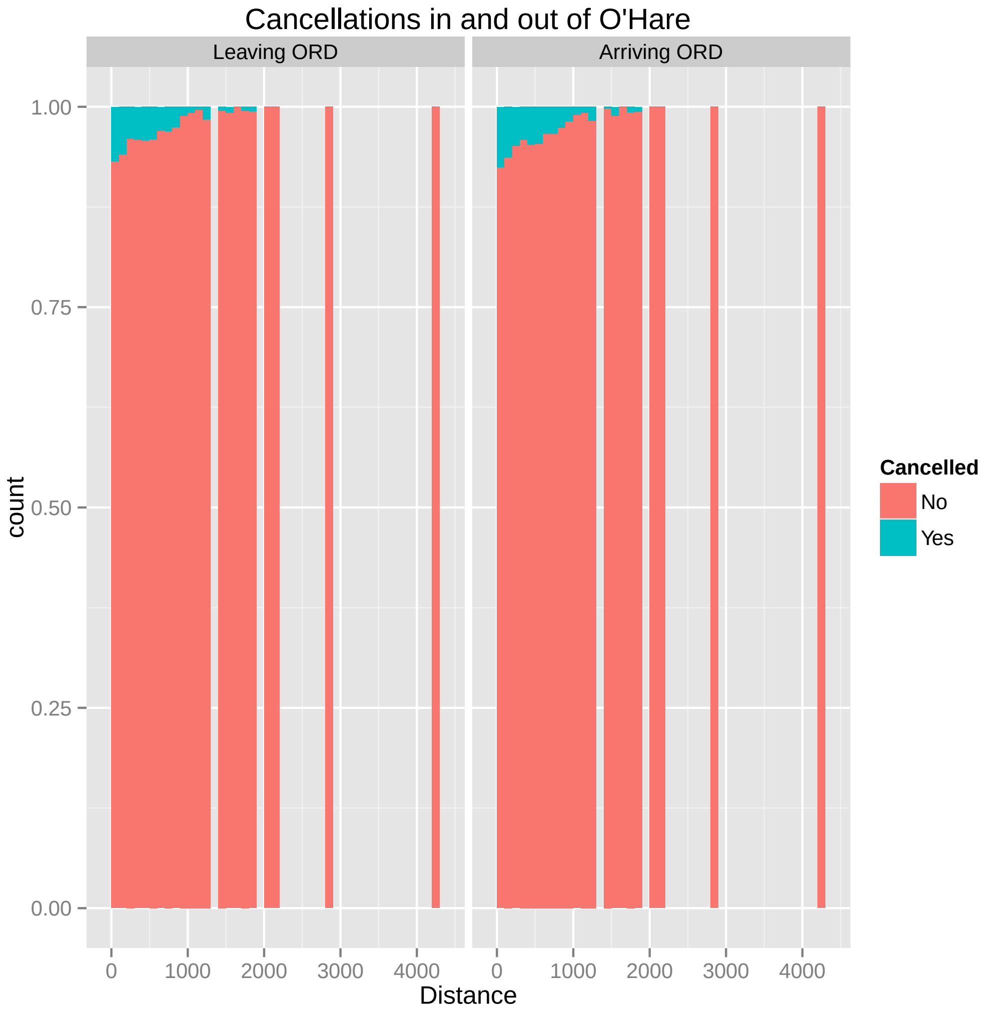
2 Cancellations

Investigate cancellations. For example, are flights to or from the airport cancelled more often? Smaller or bigger machines? (look at the flight distances) Again, provide graphical summaries.

```
my_db <- src_sqlite("flights")
flights <- tbl(my_db, "flights")
ohare_all <- flights %.% filter(Dest == "ORD" | Origin == "ORD") %.% select(DayOfWeek,
    FlightDate, Carrier, Origin, Dest, DepDelay, ArrDelay, Cancelled, CancellationCode,
    CRSDepTime, DepTime, CRSArrTime, Distance) %.% mutate(arrival = (Dest ==
    "ORD"))

ohare_all <- collect(ohare_all)
# arrival column is True when the destination is ORD, False otherwise
select(ohare_all, Origin, Dest, arrival)

## Source: local data frame [94,302 x 3]
##
##    Origin Dest arrival
## 1     ATL  ORD       1
## 2     ATL  ORD       1
## 3     ORD  ATL       0
## 4     ATL  ORD       1
## 5     ORD  ATL       0
## 6     ATL  ORD       1
## 7     ORD  ATL       0
## 8     ATL  ORD       1
## 9     DTW  ORD       1
## 10    DTW  ORD       1
## ..    ...  ...     ...

ohare_all$arrival <- as.factor(ohare_all$arrival)
levels(ohare_all$arrival) <- c("Leaving ORD", "Arriving ORD")
ohare_all$Cancelled <- as.factor(ohare_all$Cancelled)
levels(ohare_all$Cancelled) <- c("No", "Yes")
```

In the plot below, the distance travelled is on the x axis and the proportion of cancelled flights is shaded in blue (the smaller portions). The proportion of cancelled flights are about the same for flights arriving at O'Hare and flights departing O'Hare, but there do appear to be slightly more cancellations for arriving flights. We think this is due to the fact that since O'Hare is one of the largest airports in the country, it has more practice handling difficult situations, and thus less likey to cancel than say much smaller airports sending planes to O'Hare. The other trend we noticed in the plot below is that the proportion of cancellations decreases pretty steadily as the distance increases. We assume this happens because there are more flights per day going to closer locations, so cancelling really isn't a big problem when you can catch the next flight to that location in an hour or two. But for the really long distance flights (2000 miles or more), there are no cancellations, because even a several hour delay is less inconvenient than cancelling the flight entirely. Plus, the long distance flights are much less frequent overall.

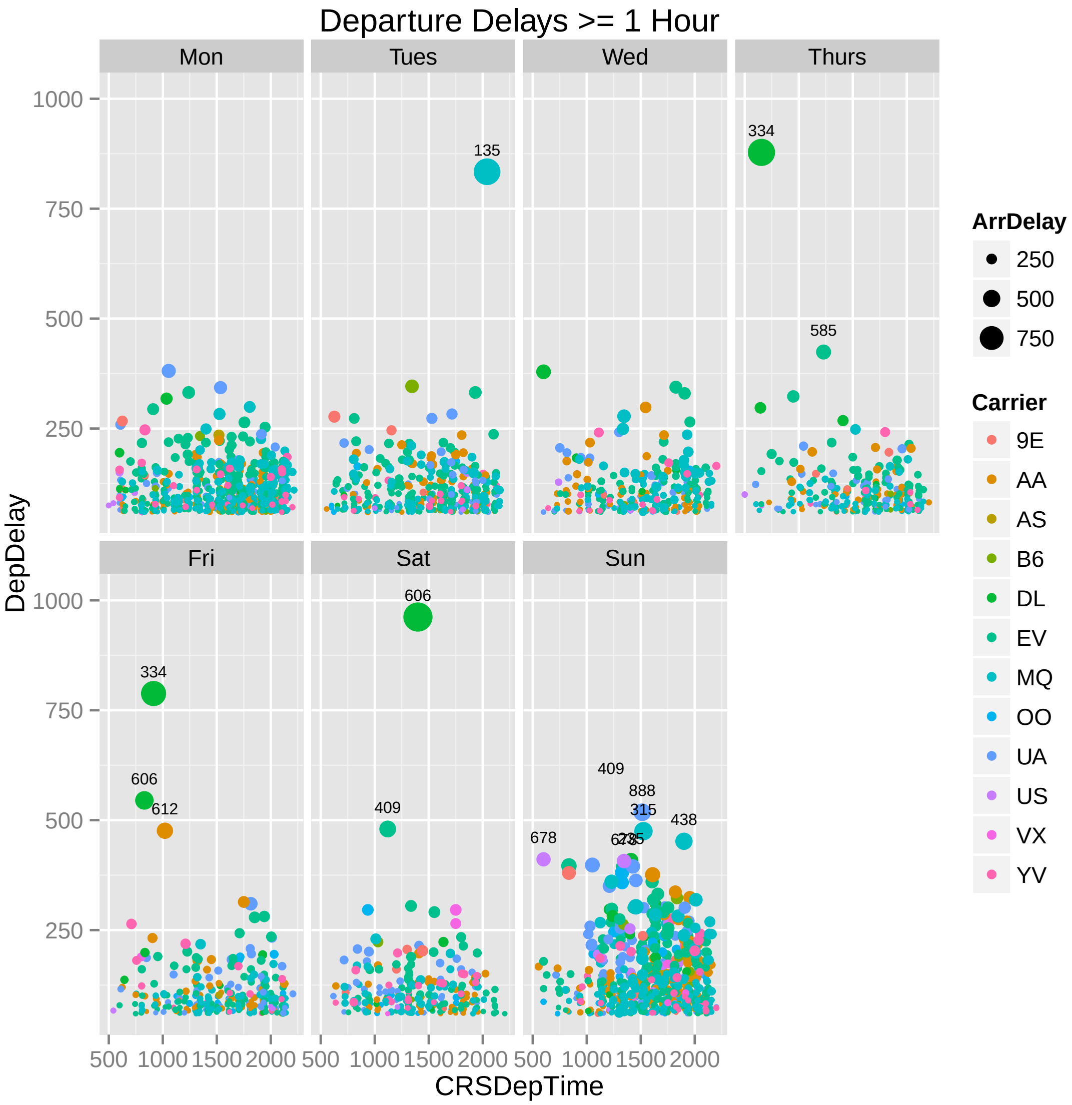
```
qplot(data = ohare_all, Distance, geom = "histogram", fill = Cancelled, position = "fill",
    binwidth = 100, main = "Cancellations in and out of O'Hare") + facet_wrap(~arrival)
```



3 Combining the Two

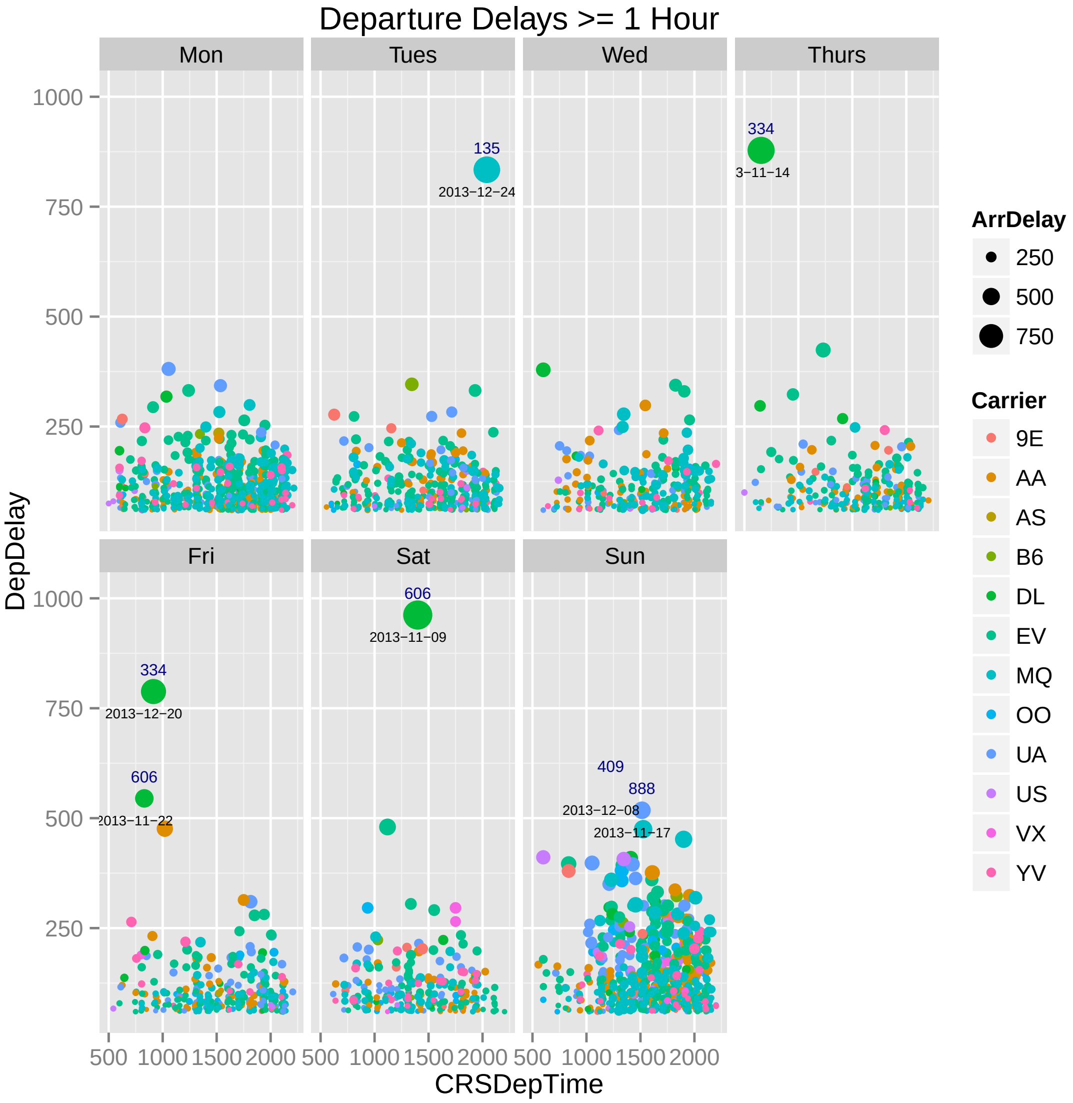
Finally, we looked at the disances travelled by the flights with the largest departure times because the cancellations plot caused us to think that maybe the flights with the largest delays were also the longer distance flights that were delayed a lot instead of being cancelled. However, this turned out not to be the case. In fact, one of the largest delays (over 750 minutes/12.5 hours) was a flight of only 135 miles.

```r
qplot(data = ohare_depdelays, y = DepDelay, x = CRSDepTime, colour = Carrier,
    size = ArrDelay, na.rm = T, main = "Departure Delays >= 1 Hour") + facet_wrap(~DayOfWeek,
    nrow = 2) + geom_text(data = subset(ohare_depdelays, DepDelay > 400), aes(label = Distance,
    x = CRSDepTime, y = DepDelay + 50), color = I("black"), size = I(2.5))
```



We wondered why this short flight was delayed for over 12 hours instead of just being cancelled. So, we looked at the dates of the most delayed flights, and we noticed why it wasn't cancelled right away:

```
qplot(data = ohare_depdelays, y = DepDelay, x = CRSDepTime, colour = Carrier,
    size = ArrDelay, na.rm = T, main = "Departure Delays >= 1 Hour") + facet_wrap(~DayOfWeek,
    nrow = 2) + geom_text(data = subset(ohare_depdelays, DepDelay > 500), aes(label = FlightDate,
    x = CRSDepTime - 90, y = DepDelay - 50), color = I("black"), size = I(2)) +
    geom_text(data = subset(ohare_depdelays, DepDelay > 500), aes(label = Distance,
        x = CRSDepTime, y = DepDelay + 50), color = I("navy"), size = I(2.5))
```



It was Christmas Eve! So, the short flight probably wasn't cancelled because there was no flight the next day, and people wanted to get to their families before Christmas.