Report for A4

Manthan Thakar, Vineet Trivedi and Sharad Boni

October 13, 2017

```
#install.packages("tidyr")
#install.packages("igraph")
```

```
require("ggplot2")
require("dplyr")
require("igraph")
library(plyr)
library(tidyr)
library(ggplot2)
RESOURCE_DIR = "resources/"
library(maps)
library(dplyr)
```

MapReduce Jobs

There are Jobs running in 3 phases:

1. **Data Cleaning Job:** This job is responsible for cleaning data and testing data for sanity check conditions. The output of this job generates a `SequenceFile` which is used by all subsequent jobs.
2. **Delay Jobs:** This jobs calculates delays for airlines and airports per month, per year, this job uses Clean Data generated by the first job.
3. **Activity Jobs**: This job calculates the most active airports and airlines and sorts them in descending order.

Execution Environment

AWS Cluster Configurations

The Amazon Map Reduce cluster was setup using Amazon EMR.

Instance Type: m3.xlarge

Hadoop Distribution: Amazon 2.7.3 (EMR 5.8.0)

Memory: 15GB

Storage: 2 x 40 GB SSD

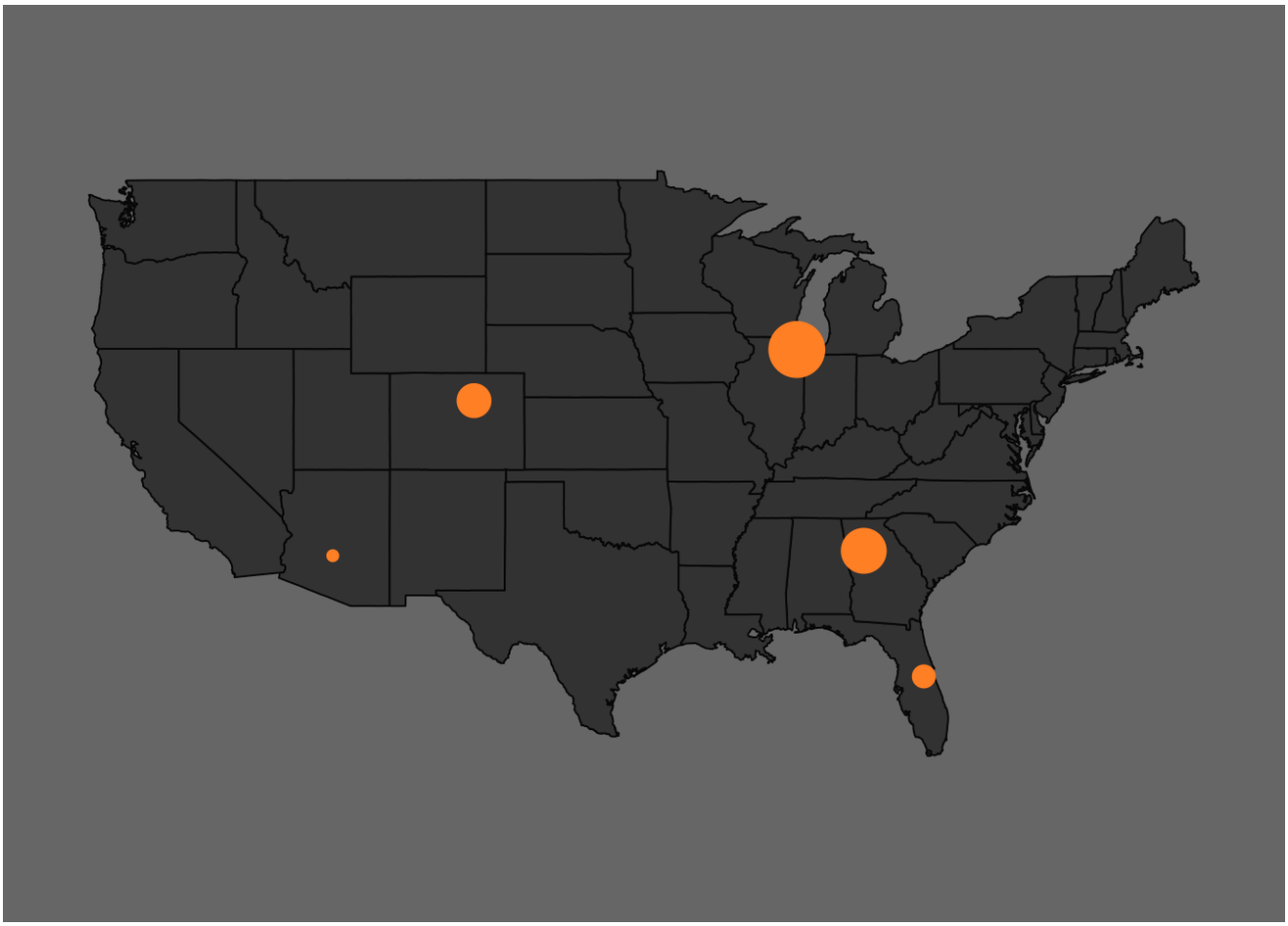
vCPU: 4

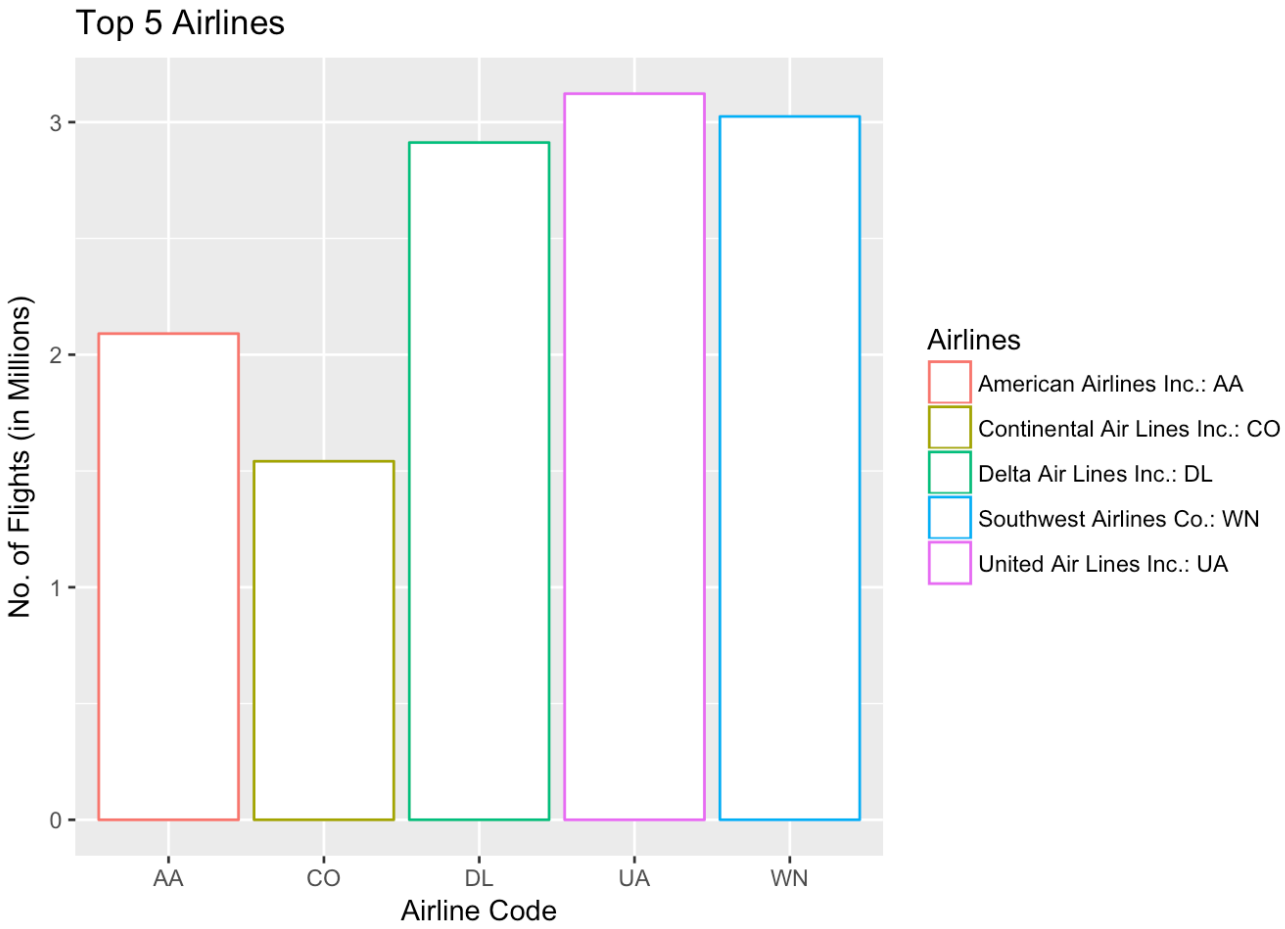
On a 4-node m3.xlarge cluster, it takes about **19 minutes** to run all three jobs, with Data Cleaning job

taking **15 minutes**.

Most Active Airports and Airlines

The following graphs show the top 5 most active airlines and airports extracted from the data.





The graph above is for Top 5 Active Airlines. The graph above is plotted for Airline Flights in millions (y-axis) vs Airline Code (x-axis). A legend has been provided to map the Airline Code with the name of the Airline. As is evident from the graph above: United Airlines (UA) is the most active airline with just over 3 million flights. Southwest Airlines (WN) and Delta Airlines (DL) are also around 3 million flights ranking 2nd and 3rd respecitvely. American Airlines (AA) ranks 4th with just over 2 million flights. Continental Airlines (CO) is the least active airline with around 1.5 million flights.

Airline Delays

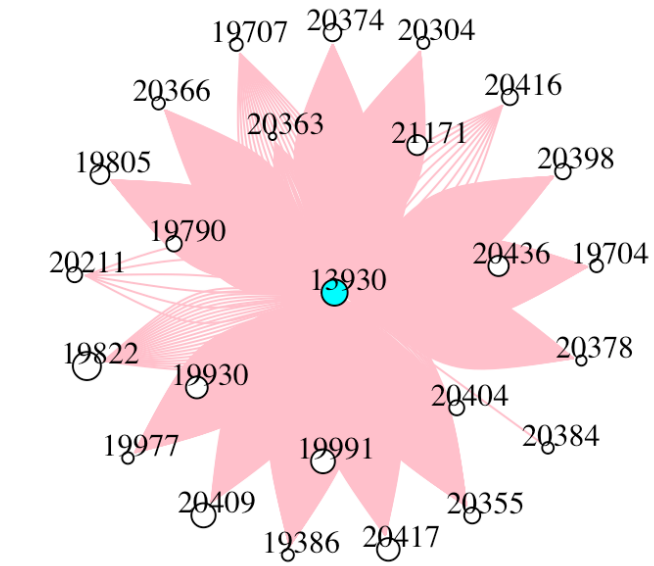
The following graph shows airline delays with respect to the top airport. The size of the vertices are proportional to the delay of that airline going to the airport. The top airport is depicted as a blue vertex. The airlines are depicted as white vertices.

```
read_files = function(dirname, col_names=c()) {
  files <- dir(paste(RESOURCE_DIR, dirname, sep=""), recursive=TRUE, full.names=T
RUE)
  tables <- lapply(files, read.csv)
  tables <- lapply(tables, setNames, nm = col_names)
  do.call(rbind, tables)
}
col_names <- c("airlineID", "airportID", "month", "year", "mDelay")
airline_delay = read_files("airline-delay-data", col_names)
airport_delay = read_files("airport-delay-data", c("airportID", "month", "year",
"mDelay"))
```

```
most_active_airports = read.table(paste(RESOURCE_DIR, "most-active-airport/part-r
-00000", sep=""))
most_active_airlines = read.table(paste(RESOURCE_DIR, "most-active-airlines/part-
r-00000", sep=""))

names(most_active_airlines) = c("activity", "airlineID")
names(most_active_airports) = c("activity", "airportID")

top_5_dest = most_active_airports[1, ]
top_5_airlines = most_active_airlines[1:5, ]

airline_id = read.csv("resources/Airline_id.csv", header=T)
filtered = filter(airline_delay, airline_delay$airportID %in% c(top_5_dest$airpor
tID))
dest_graph = graph.data.frame(filtered, directed=T)
V(dest_graph)$color<-ifelse(V(dest_graph)$name %in% top_5_dest$airportID, 'cyan',
ifelse(V(dest_graph)$name %in% top_5_airlines, 'blue', 'white'))
E(dest_graph)$color = "pink"
par(0, 0, 1, 0)
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

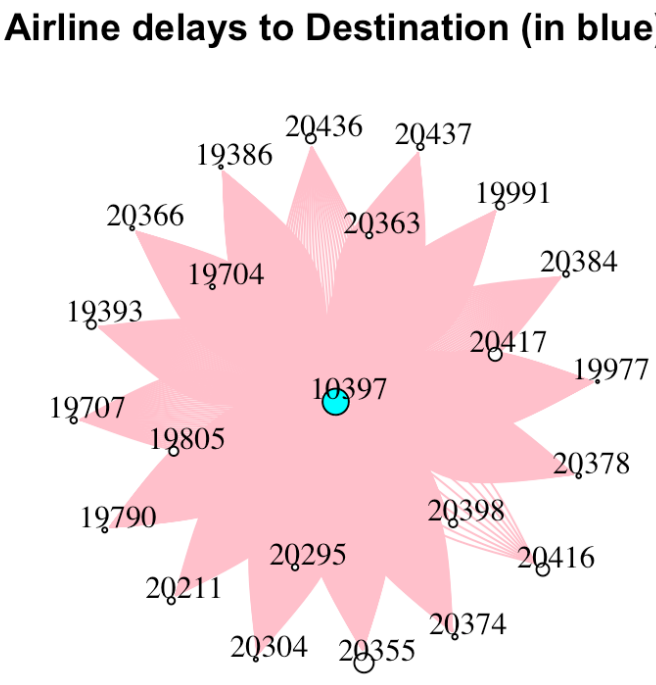
```
V(dest_graph)$size<-ifelse(V(dest_graph)$name %in% top_5_dest, 10, E(dest_graph)$
mDelay * 20)
plot(dest_graph, layout=layout.kamada.kawai,
vertex.label.dist=1,
vertex.label.color='black',
vertex.label.font = 0.5,
vertex.label=V(dest_graph)$name,
main="Airline delays to Destination (in blue)")
```




```
top_5_dest = most_active_airports[2, ]
top_5_airlines = most_active_airlines[1:5, ]

airline_id = read.csv("resources/Airline_id.csv", header=T)
filtered = filter(airline_delay, airline_delay$airportID %in% c(top_5_dest$airpor
tID))
dest_graph = graph.data.frame(filtered, directed=T)
V(dest_graph)$color<-ifelse(V(dest_graph)$name %in% top_5_dest$airportID, 'cyan',
ifelse(V(dest_graph)$name %in% top_5_airlines, 'blue', 'white'))
E(dest_graph)$color = "pink"
par(0, 0, 1, 0)
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

```
V(dest_graph)$size<-ifelse(V(dest_graph)$name %in% top_5_dest, 10, E(dest_graph)$
mDelay * 20)
plot(dest_graph, layout=layout.kamada.kawai,
vertex.label.dist=1,
vertex.label.color='black',
vertex.label.font = 0.5,
vertex.label=V(dest_graph)$name,
main="Airline delays to Destination (in blue)")
```



As we can see in the figure above, airlineID 19822 - Piedmont has more delays in flights when going to

Chicago Airport (13930).

Mean delays for top 5 destinations across months

Mean delays for top 5 destinations are displayed across months for all years in the graph below.

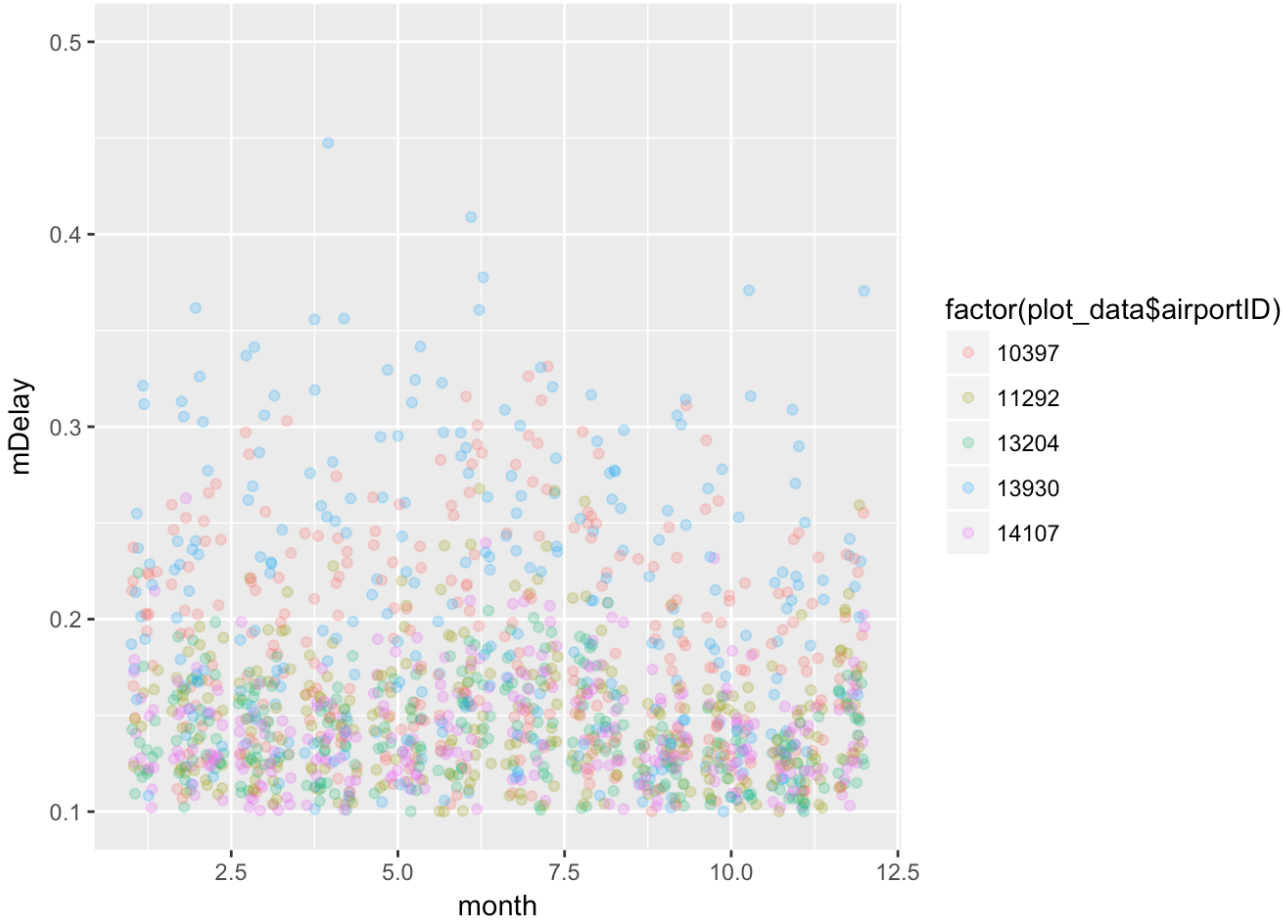
Here, we can see that airport ID `13930` (chicago airport), has more delays lying above 0.3 mean delay. This airport also happens to be the most active airport in the dataset.

```
top_5_dest = most_active_airports[1:5, ]
plot_data = airport_delay %>%
  filter(airportID %in% top_5_dest$airportID) %>%
    group_by(month,airportID)

b1<-ggplot(plot_data, aes(month, mDelay)) + xlim(1, 12) + ylim(0.1, 0.5) +
  geom_jitter(alpha=I(1/4), aes(col = factor(plot_data$airportID)))
  labs(x="Month of year",y="Mean Delay",title="Flight Delays to Top 5 Destination
s", col="Airport IDs")
```

```
## $x
## [1] "Month of year"
##
## $y
## [1] "Mean Delay"
##
## $title
## [1] "Flight Delays to Top 5 Destinations"
##
## $colour
## [1] "Airport IDs"
##
## attr(,"class")
## [1] "labels"
```

```
b1
```

```
## Warning: Removed 262 rows containing missing values (geom_point).
```



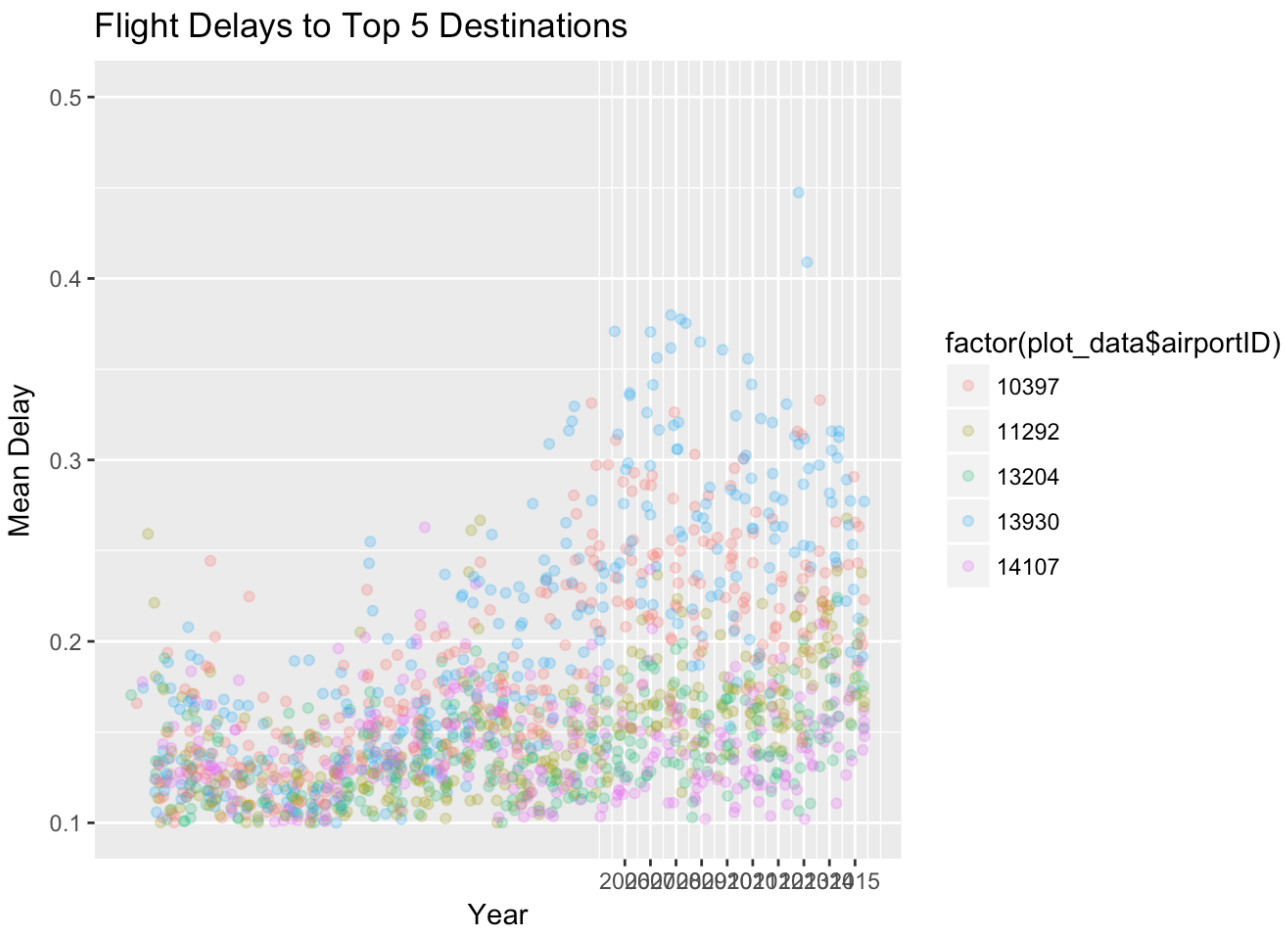
Mean delays for destinations across years

The graph below shows mean delows for destinations across years. Interestingly, the delays for top 5 most active airports have increased after the year 2006. Moreover, in 2015 there they mean delay has increased significantly for `13930` which seems to have the highest mean delay in all other years.

Interestingly, the two datapoints for year 2013 are sort of the outliers. This could be because of sequestration-related furloughs being in effect during that year in Chicago.

```
top_5_dest = most_active_airports[1:5, ]
plot_data = airport_delay %>%
  filter(airportID %in% top_5_dest$airportID) %>%
    group_by(year,airportID)

b1<-ggplot(plot_data, aes(year, mDelay)) + ylim(0.1, 0.5) +
  geom_jitter(alpha=I(1/4), aes(col = factor(plot_data$airportID))) +
  scale_x_continuous(breaks=seq(2006, 2015)) +
  labs(x="Year",y="Mean Delay",title="Flight Delays to Top 5 Destinations")
b1
```

```
## Warning: Removed 122 rows containing missing values (geom_point).
```


Flight Delays to Top 5 Destinations
Mean Delay
0.5
0.4
0.3
0.2
0.1
Year
factor(plot_data$airportID)
10397
11292
13204
13930
14107