Flights Lab

Kate Bailey

11/16/2021

Introduction

The three major New York City-area airports — JFK, LaGuardia, and Newark — are some of the busiest airports in the country. Each sees a similar number of flights (over 10,000) depart each year. But do those numbers stay equal once you start looking at individual carriers?

In this lab, I'll use a data set with every 2013 LGA, JFK, and EWR flight to look into **airline-specific** flight information, in an effort to figure out if major airlines tend to use certain NYC airports over others. To build on this analysis, I'll try to ascertain whether airlines tend to choose one airport as their NYC "hub" — that is, as a gateway to many different *destinations*.

Observations & Analysis

First step: importing the data set.

```
flight.data <- read.csv("nycflights.csv")
head(flight.data)
```

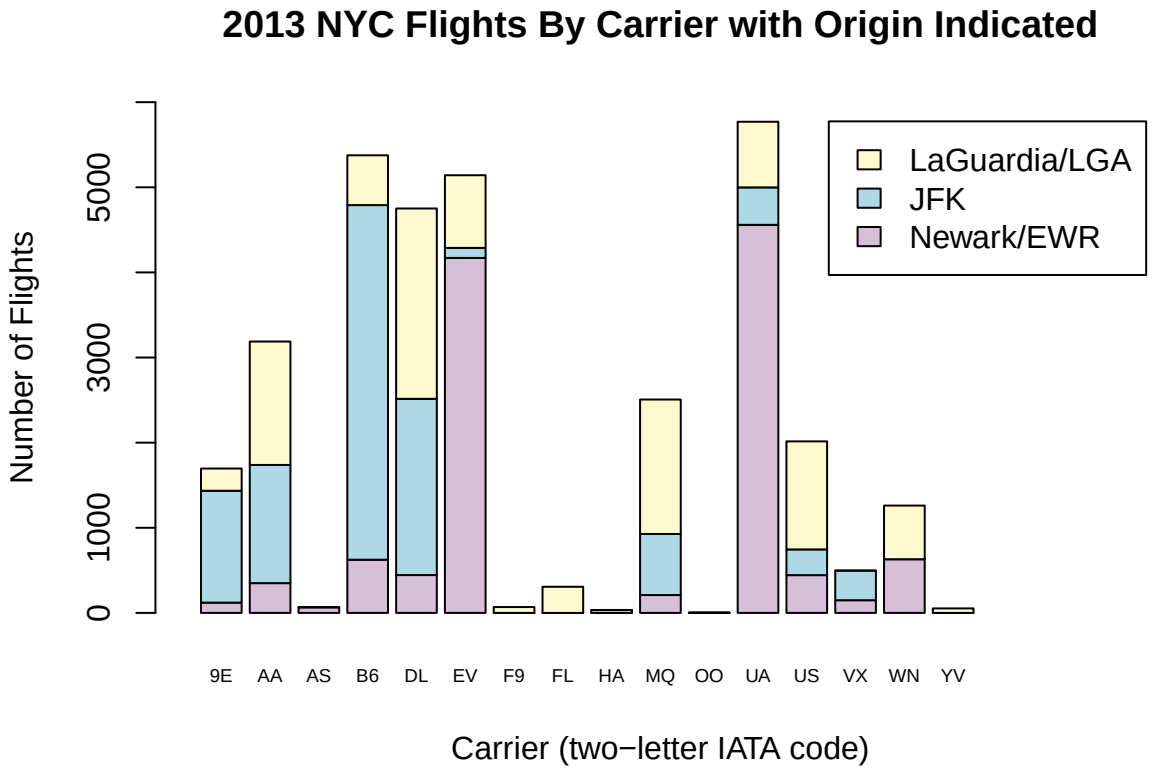
```
##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
## 1 2013     6  30      940        15     1216        -4      VX  N626VA    407
## 2 2013     5   7     1657        -3     2104        10      DL  N3760C    329
## 3 2013    12   8      859        -1     1238        11      DL  N712TW    422
## 4 2013     5  14     1841        -4     2122       -34      DL  N914DL   2391
## 5 2013     7  21     1102        -3     1230        -8      9E  N823AY   3652
## 6 2013     1   1     1817        -3     2008         3      AA  N3AXAA    353
##   origin dest air_time distance hour minute
## 1    JFK  LAX      313     2475    9     40
## 2    JFK  SJU      216     1598   16     57
## 3    JFK  LAX      376     2475    8     59
## 4    JFK  TPA      135     1005   18     41
## 5    LGA  ORF       50      296   11      2
## 6    LGA  ORD      138      733   18     17
```

Narrowing the search

Over 30,000 flights took off from LGA, JFK, and EWR in 2013. Here's how that breaks down by carrier:

```
barplot(table(flight.data$origin,flight.data$carrier),
        main = "2013 NYC Flights By Carrier with Origin Indicated",
        cex.names=0.6,
        ylim = c(0,6000),
        xlim = c(0,23),
        xlab = "Carrier (two-letter IATA code)",
```

```
        ylab = "Number of Flights",
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR","JFK","LaGuardia/LGA"))
```



There are a number of carriers using the three NYC airports, but that's about all we can get from this graph in ten seconds. There are no trends that stand out immediately from this graph, because there's just so much going on! So, let's narrow our search down to just the **three most active airlines: UA, B6, and EV.**

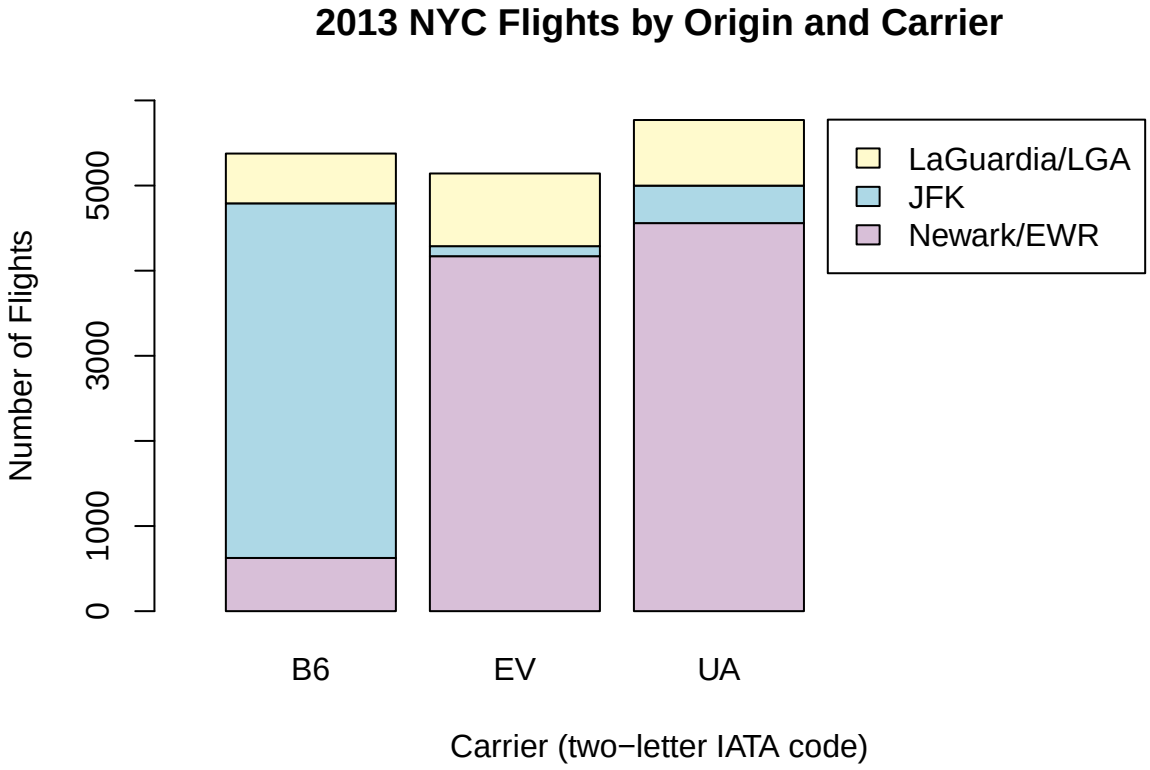
To do so, we can make a new data frame with condensed carriers ("cc.data"):

```
cc.data <- flight.data[flight.data$carrier == "UA"
                       |flight.data$carrier == "B6"
                       |flight.data$carrier == "EV",]
```

Part 1: Where did you come from?

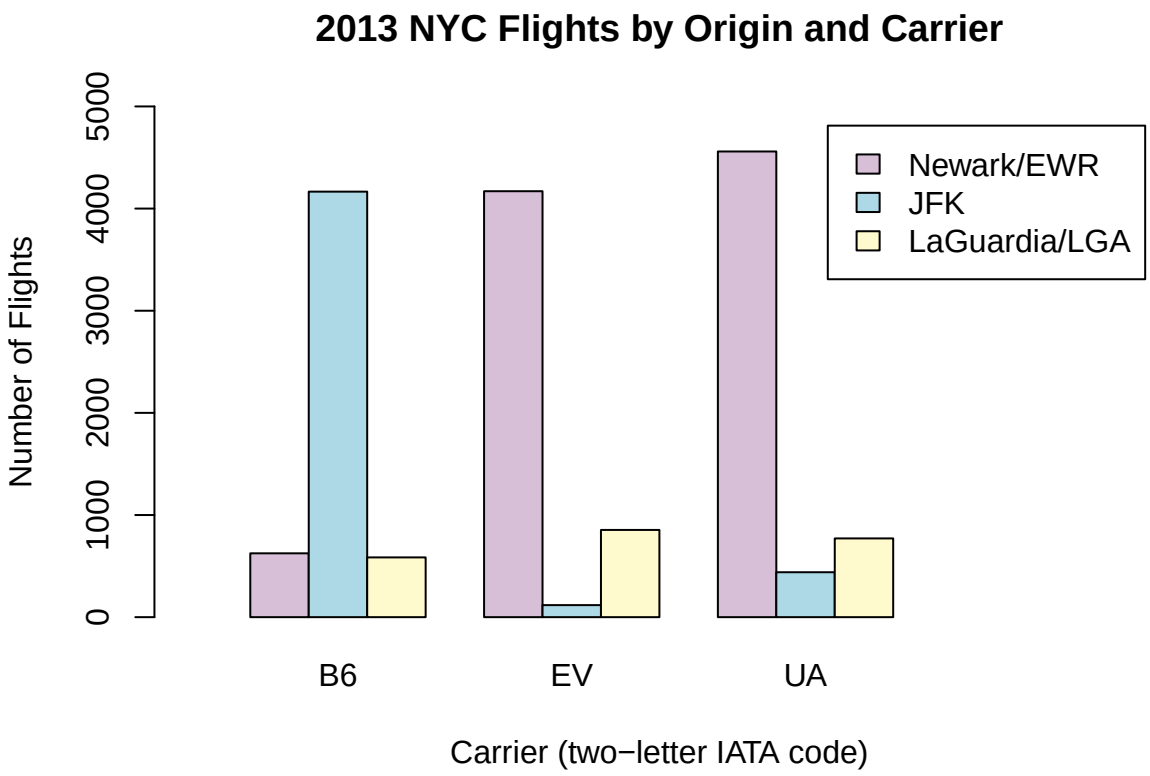
Now that we've narrowed down the carriers in the data set, we can look at how many flights depart on each airline, broken down by airport of origin.

```
barplot(table(cc.data$origin, cc.data$carrier),
        main = "2013 NYC Flights by Origin and Carrier",
        cex.names = 1,
        ylim = c(0,6000),
        xlim = c(0,5.5),
        xlab = "Carrier (two-letter IATA code)",
        ylab = "Number of Flights",
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR","JFK","LaGuardia/LGA")
        )
```



In this graph, we can start to see that each airline appears to have a major chunk of its flights departing from one airport (indicated by the block of yellow for B6 and the blocks of purple for EV and UA). This graph is also helpful in verifying that these airlines are comparable, as according to the graph they each send about 5,500 flights out of NYC airports each year. However, it'll be useful to look at this data in a side-by-side bar chart in order to see the trends more clearly.

```
barplot(table(cc.data$origin, cc.data$carrier),
        main = "2013 NYC Flights by Origin and Carrier",
        beside = TRUE,
        cex.names = 1,
        ylim = c(0,5000),
        xlim = c(0,16),
        xlab = "Carrier (two-letter IATA code)",
        ylab = "Number of Flights",
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR","JFK","LaGuardia/LGA")
        )
```



To me, this graph answers the question (Do certain airlines use certain airports more frequently?) far more clearly. **For each carrier, there is a dramatic spike for a particular airport.** This indicates that B6 is sending a much higher quantity of flights out of JFK than any other airport; similarly, EV and UA are sending very high numbers of flights out of Newark. In other words, these three airlines have each chosen — and, by virtue of how dramatic the difference is, seemingly chosen *deliberately* — between EWR, JFK, and LGA as the airport from which they're directing most of their NYC-area flights.

Part 2: Where did you go?

Next, I wanted to see if certain airlines also routed the **majority of their destinations** out of a particular airport. Theoretically, airlines could divide up destinations equally and/or "randomly" among the NYC airports, or they could choose one airport (imagine lots of lines going out of that one airport) — the latter is known as the "hub and spoke" model for air travel, but not all airlines use it. I wanted to see if I could figure out which airport, if any, each airline uses as a **hub**.

I had to do a little bit of coding to figure this out. In order to make charts that show the number of *destinations* on the y-axis (as opposed to the number of flights), I figured I would make a **new data frame** that contains **only one row for each carrier/origin/destination combination**. So, this code makes a new data frame ("unique.data") that contains only the first observation for each unique combination (e.g. the first instance of a UA flight from LGA to CLT, or a B6 flight from JFK to ATL).

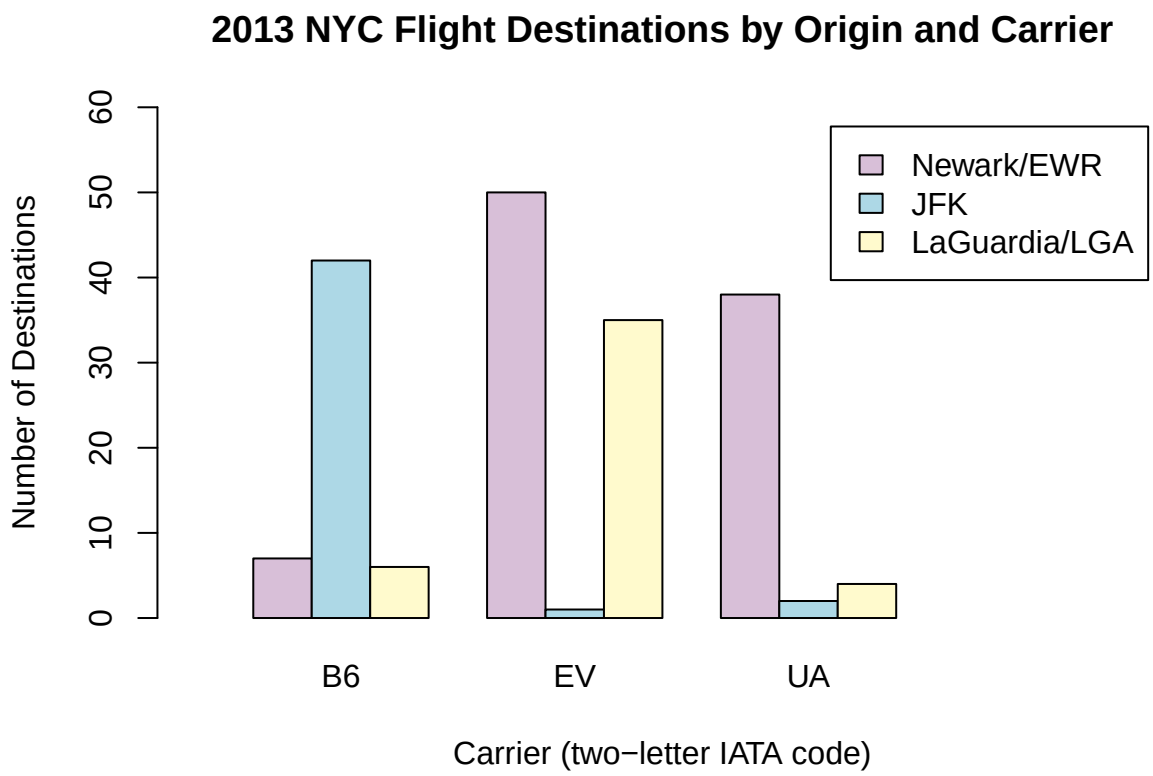
```
# Make data frame with just the carrier/origin/dest variables
new.data <- data.frame(cc.data$carrier,cc.data$origin,cc.data$dest)
# Use "unqiue()" to remove duplicates
unique.data <- unique(new.data)
# Note new variable names
head(unique.data)
```

```
##   cc.data.carrier cc.data.origin cc.data.dest
## 1              B6            JFK          IAD
## 2              EV            EWR          JAX
```

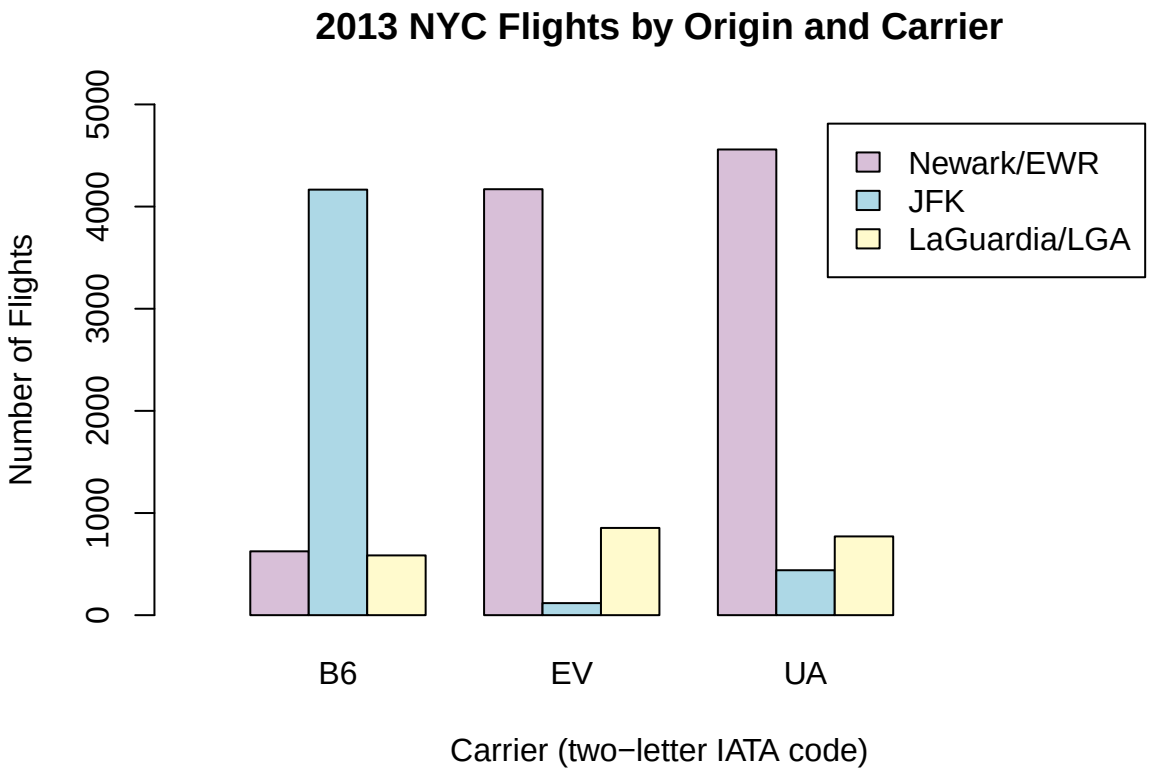
```
## 3              B6            JFK           ROC
## 4              B6            LGA           RSW
## 5              EV            EWR           DAY
## 6              B6            JFK           BTV
```

Here are two bar charts. The first shows the number of unique destinations each airline flies to, divided into airports of origin. The second is the same carriers by origin bar chart as before, for comparison.

```
barplot(table(unique.data$cc.data.origin, unique.data$cc.data.carrier),
        main = "2013 NYC Flight Destinations by Origin and Carrier",
        beside=TRUE,
        xlim = c(0,16),
        ylim = c(0,60),
        ylab = "Number of Destinations",
        xlab = "Carrier (two-letter IATA code)",
        cex.names = 1,
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR","JFK","LaGuardia/LGA")
        )
```




```
barplot(table(cc.data$origin, cc.data$carrier),
        main = "2013 NYC Flights by Origin and Carrier",
        beside = TRUE,
        cex.names = 1,
        ylim = c(0,5000),
        xlim = c(0,16),
        xlab = "Carrier (two-letter IATA code)",
        ylab = "Number of Flights",
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR","JFK","LaGuardia/LGA")
        )
```



The first chart shows that yes, **each airline has a very different number of destinations for each NYC airport of origin.** For example, at UA, you can fly to 40+ different places if you're flying out of Newark, but only a few if you're flying out of LaGuardia or JFK.

Additionally, interestingly but perhaps not completely surprisingly, **each airline's spikes in number of destinations overlap with the spikes in number of flights in genera**l (as shown in the second graph). B6, who mainly flew out of JFK in 2013, has a far greater variety of destinations for JFK. And EV and UA have greater variety in destinations for Newark, their main airport from part 1. It seems that the more flights an airline is routing out of a given NYC airport, the more destinations the airline will have for that NYC airport.

Based on the first graph, it seems reasonable to conclude that **UA's NYC-area hub is Newark**, while **B6's hub is JFK**, based on the fact that they have so many more destinations for those airports.

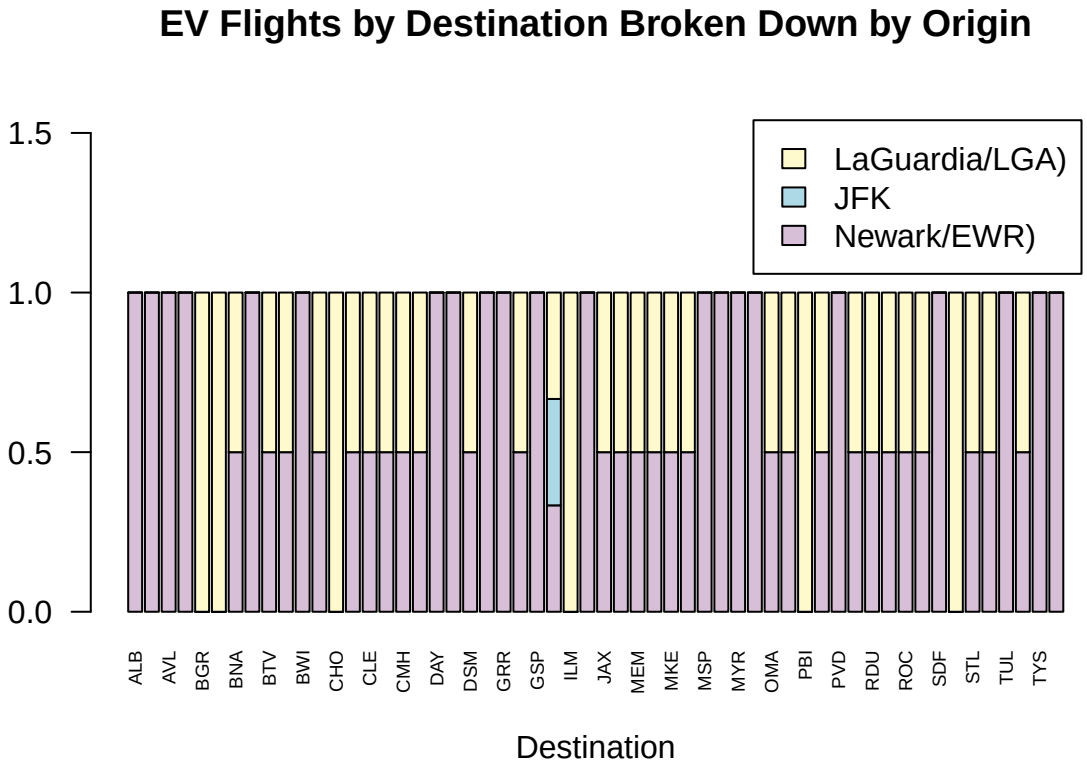
But, there **is** something unusual and surprising about the first graph! The airline EV has TWO spikes. Newark's follows the aforementioned trend, but LaGuardia's spike is befuddling at first glance. **Why does EV have ~35 destinations for LGA travelers** when they only send about 1,000 flights out of LGA per year?

Further search: The mystery of EV!

To answer the question about EV's LaGuardia flights, I decided to look at the individual destinations to see whether I could figure out a trend regarding which destinations were routed out of LGA and which ones out of EWR. I know this is not the prettiest/best way of looking at the data, but it works okay!

```
barplot(prop.table(table(unique.data$cc.data.origin[unique.data$cc.data.carrier=="EV"],
                         unique.data$cc.data.dest[unique.data$cc.data.carrier=="EV"]),2),
        main = "EV Flights by Destination Broken Down by Origin",
        cex.names=0.6,
        las=2,
        ylim=c(0,1.6),
        xlab = "Destination",
```

```
        col = c("thistle","lightblue","lemonchiffon"),
        legend.text = c("Newark/EWR)","JFK","LaGuardia/LGA)"),
    )
```



If a destination is purple all the way through, that is a EWR-only destination; if it's only yellow, it's LGA-only. Split yellow and purple means there are flights out of both EWR and LGA to that destination (and IAD — aka Washington, DC — has the distinction of being EV's only JFK destination).

There are only six LGA-only destinations, and many more "EWR *and* LGA" destinations than I expected (over 20!). **There doesn't seem to be much rhyme or reason to which destinations are out of which airports.** I can't quickly figure out what exactly the six LGA-only destinations have in common, or why EV would specifically choose to route flights to them from LGA.

One thing that helped me to understand this data came from Google: EV stands for ExpressJet. (FYI, UA is United and B6 is JetBlue). ExpressJet is a regional airline, and it uses **the "point-to-point" model** for air travel, which means that an airline will set up a bunch of flights between many airports rather than routing all travelers through a handful of hubs like UA/B6 seem to. So, maybe that partially explains the odd variety in destinations for LGA — **it doesn't have a NYC hub at all!** (But of course it opens up another question: why basically boycott JFK?)

Unfortunately, after playing around with the data for a bit, I don't think we can figure out exactly why EV has so many destinations for LGA based on this data set alone. Maybe there is **another factor** that's not included as a variable in this data set that would explain why EV chooses to send certain flights out of LGA and not EWR.

Conclusion

In this lab, I was able to guess that United/UA and JetBlue/B6 both have defined NYC-area hubs: Newark and JFK, respectively. This is based on the number of destinations that originate from each of the three airports. I also learned that each of the biggest three airlines have a preference, in terms of number of flights, for one of the three airports.

In a future lab, **I would want to look at more airlines** to see if this trend continues. It would be interesting to see if it's possible to use this data set to accurately guess whether or not an airline uses the "point to point" or "hub and spoke" model, and if they use the latter, which airport if any serves as their NYC area hub.