STAT495 (Advanced Data Analysis) HW#7

Tam Tran-The

November 10, 2016

Exercise 12.10

Based on data from 2012 only, looking at flights that are 750 miles or shorter, and assuming that transportation to the airport is not an issue, would you rather fly out of JFK, LaGuardia (LGA), or Newark (EWR)? Why or why not? The intended audience is a non-technical manager who flies regularly. Be sure to include an executive summary (of no more than 2 pages including one graphical display) of your conclusion, followed by a technical appendix of your analyses that led to your decision.

SOLUTION:

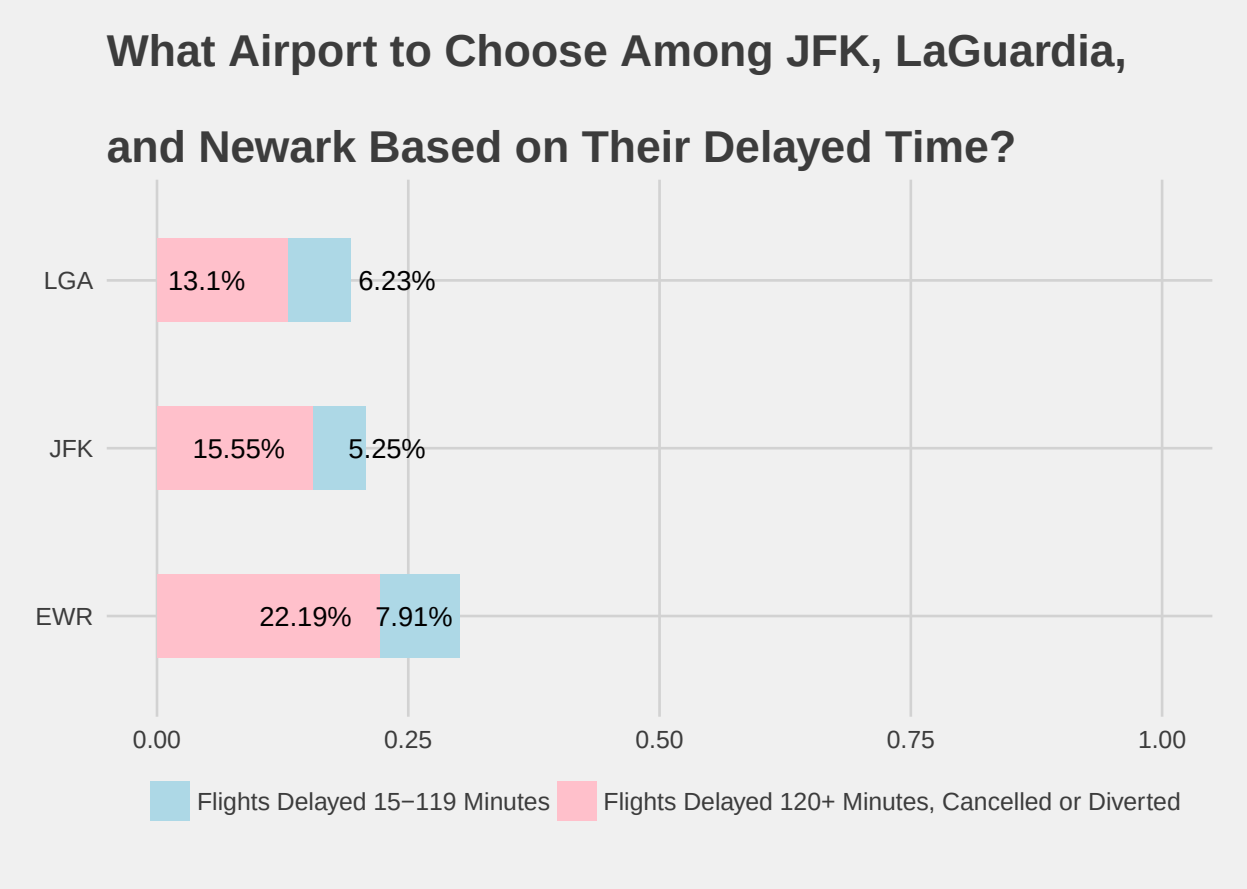
EXECUTIVE SUMMARY

This analysis takes an overview of how three major airports of New York, JFK, LaGuardia (LGA), and Newark (EWR) performed in 2012, according to the `flights` data set, which includes information of more than 169 million US flights from 1987 to 2016. The study only covers flights that were 750 miles or shorter, which were mostly domestic flights.

Assuming that transportation to the airport is not an issue, we determine how well an airport performs by looking at its proportion of delayed flights (including both short and long delays) and the total aggregate number of minutes of delay time in 2012 for each airport.

In this analysis, a flight is considered delayed when it arrived 15 or more minutes than the schedule. We consider delays that were less than 119 minutes as short, and longer than 120 minutes as long. We account for cancellations and diverted flights by grouping them into the same category as long delay since these kinds of experience usually cost travelers several hours. For computational purpose, we assume that each cancellation or diverted flight is equivalent to 240 minutes of delayed time. To have a holistic perspective, we also take into account early arrivals and subtract minutes of early arrival from the total aggregate number of minutes of delay time.

First, we would like to see how likely it is for an airport to have a delayed flight. From our analysis, we see that LaGuardia has the lowest proportion of flights being delayed (0.1933), followed by JFK with 0.2080, and Newark with 0.3010. In other words, whereas 80.67% of flights from LaGuardia arrived on time, only 69.9% of flights from Newark did. Performance ranking of the airports remains the same when we consider their proportions of short-delayed flights. Specifically, LaGuardia has the smallest proportion of short-delayed flights, whereas Newark has the highest. However, there is a change in order when we look at their proportions of long delay. Although Newark is still the worst, LaGuardia and JFK have switched their places.



We would also like to see the length of delay time an airport has accumulated over 2012. JFK appears to have the shortest length of delay time with 253,789 net minutes late, followed by LaGuardia with 443,472 net minutes late. Again, Newark comes last in the list for costing its fliers 1,200,721 net minutes.

	origin	distance	year	minutesLate	minutesEarly	netMinutesLate
1	JFK	228	2012	514109.00	-260320.00	253789.00
2	LGA	502	2012	966864.00	-523392.00	443472.00
3	EWR	529	2012	1652548.00	-451827.00	1200721.00

Overall, based on this analysis, Newark is not favorable under any circumstances. If we consider the probability of a flight being delayed, LaGuardia seems like the best choice among the three. On the other hand, if we take into account how long a flyer has to wait due to a flight delay, JFK seems more preferable.

TECHNICAL APPENDIX

```
db <- src_scidb("airlines")
flights <- tbl(db, "flights")
carriers <- tbl(db, "carriers")
airports <- tbl(db, "airports")
```

```
query1 <- "
SELECT origin, distance, year,
    sum(if(arr_delay > 15 and arr_delay <= 119, 1, 0)) / sum(1) as shortDelayProp,
    sum(if(arr_delay >= 120 or cancelled = 1 or diverted = 1, 1, 0)) / sum(1) as longDelayProp,
    (sum(if(arr_delay > 15 and arr_delay <= 119, 1, 0)) +
    sum(if(arr_delay >= 120 or cancelled = 1 or diverted = 1, 1, 0))) / sum(1) as delayProp,
    1 - (sum(if(arr_delay > 15 and arr_delay <= 119, 1, 0)) +
        sum(if(arr_delay >= 120 or cancelled = 1 or diverted = 1, 1, 0))) / sum(1) as ontimeProp
FROM flights
WHERE distance <= 750
AND origin IN ('JFK', 'LGA', 'EWR')
AND year = 2012
GROUP BY origin
ORDER BY delayProp"
ds1 <- DBI::dbGetQuery(db$con, query1)
```

```
query2 <- "
SELECT origin, distance, year,
    sum(if(cancelled = 1 OR diverted = 1, 240,
        if(arr_delay > 15, arr_delay, 0))) as minutesLate,
    sum(if(arr_delay < 0, arr_delay, 0)) as minutesEarly,
    sum(if(cancelled = 1 OR diverted = 1, 240, if(arr_delay > 15, arr_delay, 0))) +
        sum(if(arr_delay < 0, arr_delay, 0)) as netMinutesLate
FROM flights
WHERE distance <= 750
AND origin IN ('JFK', 'LGA', 'EWR')
AND year = 2012
GROUP BY origin
ORDER BY netMinutesLate"
ds2 <- DBI::dbGetQuery(db$con, query2)
```

```
ds1_tidy <- ds1 %>%
  select(origin, shortDelayProp, longDelayProp, delayProp) %>%
  tidyr::gather(key="delay_type", value="prop", -origin, -delayProp)
delay_chart <- ggplot(data=ds1_tidy,
                      aes(x=origin, y=prop)) +
  geom_bar(stat="identity", aes(fill=delay_type, width=0.5)) +
  scale_fill_manual(name=NULL, values = c("light blue", "pink"),
                    labels=c("Flights Delayed 15-119 Minutes",
                             "Flights Delayed 120+ Minutes, Cancelled or Diverted")) +
  scale_y_continuous(limits=c(0,1)) +
  coord_flip() +
  labs(title="What Airport to Choose Among JFK, LaGuardia, and Newark
         \nBased on Their Delayed Time?") +
  ggthemes::theme_fivethirtyeight() +
  geom_text(data=filter(ds1_tidy, delay_type=="shortDelayProp"),
            aes(label=paste0(round(prop*100, 2), "%")), hjust="left") +
  geom_text(data=filter(ds1_tidy, delay_type=="longDelayProp"),
            aes(label=paste0(round(prop*100, 2), "%")), hjust="right", nudge_y=0.05)
```

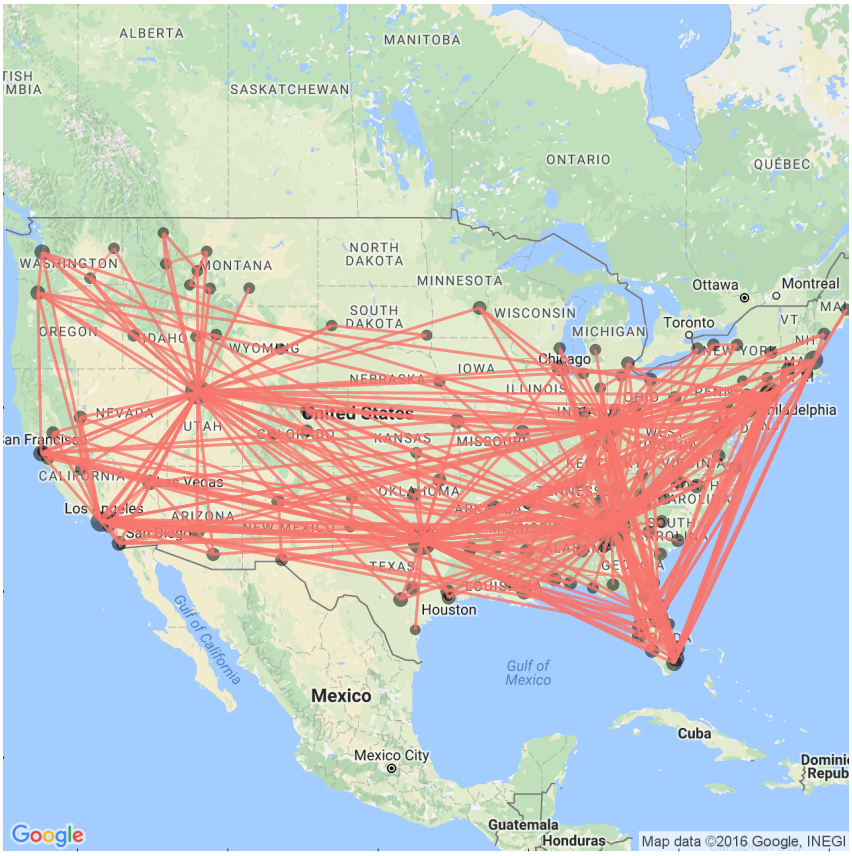
Exercise 14.9

Compare the airline route map for Delta Airlines in 2013 to the same map for Delta in 2003 and 1993. Discuss the history of Delta's use of hub airports. Reflect on the more general westward expansion of air travel in the US.

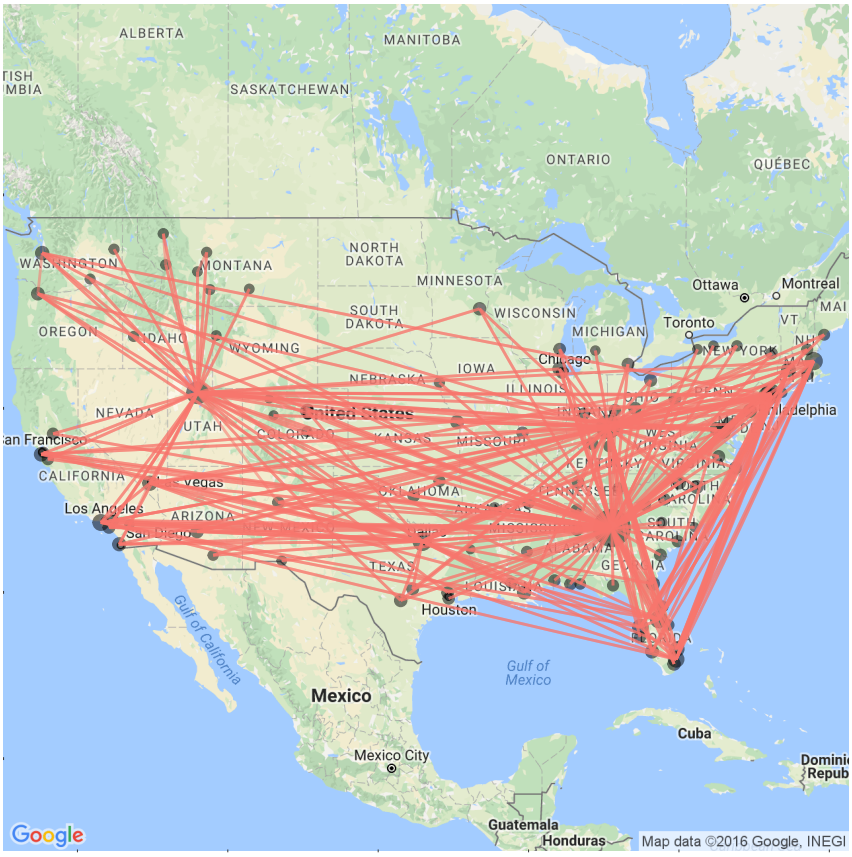
SOLUTION:

EXECUTIVE SUMMARY

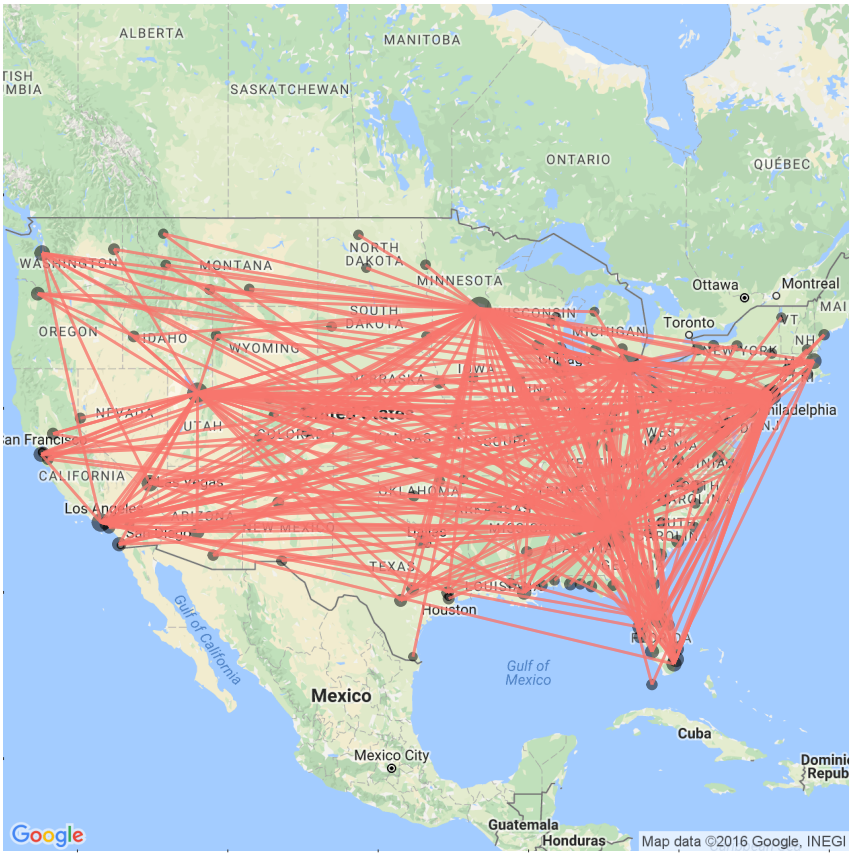
Airline route map for Delta Airlines in 1993



Airline route map for Delta Airlines in 2003



Airline route map for Delta Airlines in 2013



From sizes of the circles on the maps, we see that Delta's airport hubs have mostly remained the same over years. The biggest hub is located in Atlanta City, Georgia. The next biggest ones are located in Detroit (Michigan), Minneapolis (Minnesota), Salt Lake City (Utah), and Philadelphia (Pennsylvania). These hubs have disproportionately large number of flights routed to them compared to other cities. Some noticeable differences are that: first, Dallas (Texas) used to be a hub of Delta with significant number of flights routed through, but is no longer anymore as observed from 2003 and 2013 maps; second, Minneapolis (Minnesota) was not a hub in 1993 and 2003 but has become significantly more busy in 2013; third, since Minneapolis (Minnesota) has become a hub in 2013, it has shared a lot of flights that used to routed through Salt Lake City (Utah).

The increasing density of the line segments over years implies a growing number of flights in the US. Looking at 1993 and 2003 maps, we see that area on the right side of the map is much redder than the other side, which suggests flights were disproportionately concentrated in the east. In addition, flights from eastern hubs were mostly routed to other cities in the east, whereas flights from western hubs were mainly routed to cities around them; the density in the middle of the maps are very light which implies there were not a lot of flights connecting eastern and western cities. However, in 2013, not only the density of flights in each area increased significantly but also the density of flights connecting two regions rose considerably. This indicates that air travel in the US has greatly expanded to the west in 2013.

TECHNICAL APPENDIX

```
create_map <- function(my_carrier, my_year) {
  destinations <- flights %>%
    filter(year == my_year, carrier == my_carrier) %>%
    left_join(airports, by = c("dest" = "faa")) %>%
    group_by(dest) %>%
    summarize(N = n(), lon = max(lon), lat = max(lat),
              # note use of MySQL syntax instead of dplyr
              name = min(CONCAT("(", dest, ") ",
                REPLACE(name, " Airport", "")))) %>%
    collect() %>%
    na.omit()
  segments <- flights %>%
    filter(year == my_year, carrier == my_carrier) %>%
    group_by(origin, dest) %>%
    summarize(N = n()) %>%
    left_join(airports, by = c("origin" = "faa")) %>%
    left_join(airports, by = c("dest" = "faa")) %>%
    collect() %>%
    na.omit()
  route_map <- qmap("junction city, kansas", zoom = 4, maptype = "roadmap") +
    geom_point(data = destinations, alpha = 0.5,
               aes(x = lon, y = lat, size = N)) +
    scale_size() +
    theme_map() +
    geom_segment(data=segments, aes(x=lon.x, y=lat.x, xend=lon.y, yend=lat.y,
                                    alpha=0.7, colour="red")) +
    guides(colour=FALSE, alpha=FALSE)
  return(route_map)
}
```