Analysis of U.S. Airlines On-Time Performance Using Data Visualization

Joe Zou

2017/10/31

Introduction

An incident happened early April about United Airlines uncourtly dragging a man down their airplane after boarding at Chicago O'Hare International Airport has brought us attention to the United States' airline policies and the comparison between major airlines' performances. The man who refused to get off the plane is a doctor who needed to see his patients in the following morning, so it's crucial for him to arrive on-time. In general, do passengers care about airline's on-time performance even though they don't have an emergency? The answer is definitely yes. Studies show that cleaner planes, better in-flight services, improving on-time performances and bumping fewer passengers from their flights all have significant positive effects on passenger satisfaction and their choice of airlines. An impressive on-time performance ranking of an airline ensures their customers that they will manage their travels in a timely manner, reflecting their effectiveness.

In this post, I want to examine the question: **Which major airline in the U.S. has the best on-time performance based on delay rates?**

Data Collection and Analyzing Process

In this post, I analyzed data collected from the United States Bureau of Transportation. In order to answer the proposed question, I first picked the six largest U.S. airlines in terms of enplaned passengers, fleet size and number of destinations as potential candidates. They are American Airlines (AA), Delta Air Lines (DL), United Airlines (UA), Southwest Airlines (WN), JetBlue Airways (B6), and Alaska Airlines (AS).

Issues on delay

In order to answer the key question in as many aspects as I can, I first analyzed the airlines' performances in delay. According to definition, a flight delay is when an airline flight takes off or lands later than its scheduled time. Interestingly, most analyses available online examine the airlines' on-time arrival performances, and very few studies have focused on if they take off on-time, which is the other crucial part of the definition of delay. I chose to investigate this "other part," which is closely interconnected with arrival performance. I first used dot plot and regression line to compare the total delay percentages of the six airlines from 2009 to 2016 to analyze their overall delay performances, as well as look at the trend of each airline to see if there were improvements made throughout years.

However, only comparing the percentages was not enough, we also wanted to take into consideration of the airlines' total number of flights and total number of delays. A bubble chart done with "plotly" from R graphing library would exactly satisfy my need.

Data Visualization through Plots

Delay

It's apparent that passengers' most common source of frustration are flight delays. Delay is kind of the nature of air travel, and is a huge factor affecting the airline's on-time performance. Without going too deep to explore the possible reasons for delay, I first used regression line to compare the six airlines' overall delay performance and observe their individual trend from 2009 to 2016.

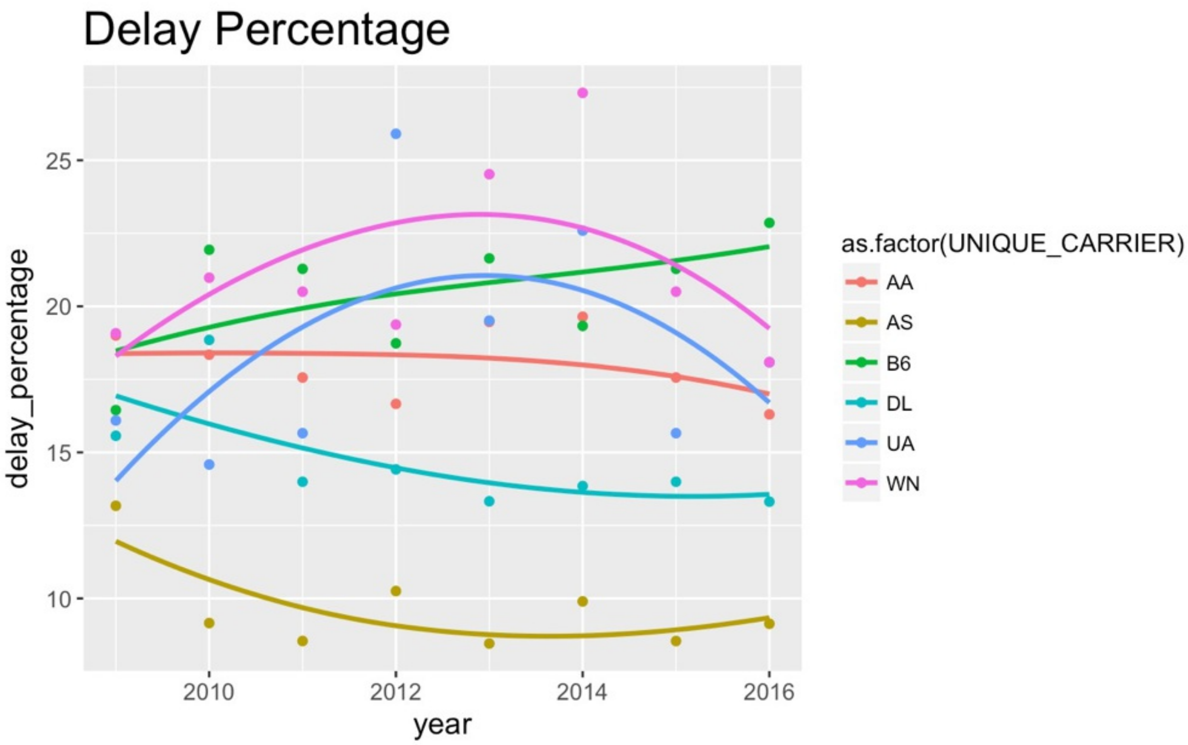
```
# This part contains all the data processing

# read in the csv file
data_n <- read.csv("data/n.csv")
completedata <- rbind(data1,data2,data3,...,data_n)

#For each year from 2009 to 2016, I filter to find data that are delayed. Then two tables, respectively for state
and for city, are grouped by airline company names and city/state names
nrow(completedata_n)
delay_n <- completedata_n %>% filter(DEP_DELAY > 0) new_n_state <- delay_n %>%
group_by(UNIQUE_CARRIER,ORIGIN_STATE_NM) %>%
summarise(total = n()) new_n_city <- delay_n %>%
group_by(UNIQUE_CARRIER, ORIGIN_CITY_NAME) %>%
summarise(total = n())

#Same data scraping for delay as well as their percentage.
new_n_delay <- delay_n %>% group_by(UNIQUE_CARRIER) %>% filter(DEP_DELAY>15) %>% summarise(total_delay = n()) new_
n_delay_Percentage <- new_n_total %>% left_join(new_n_delay) %>%
mutate(delay_percentage = (total_delay/total) * 100) head(new_n_delay_Percentage)
```

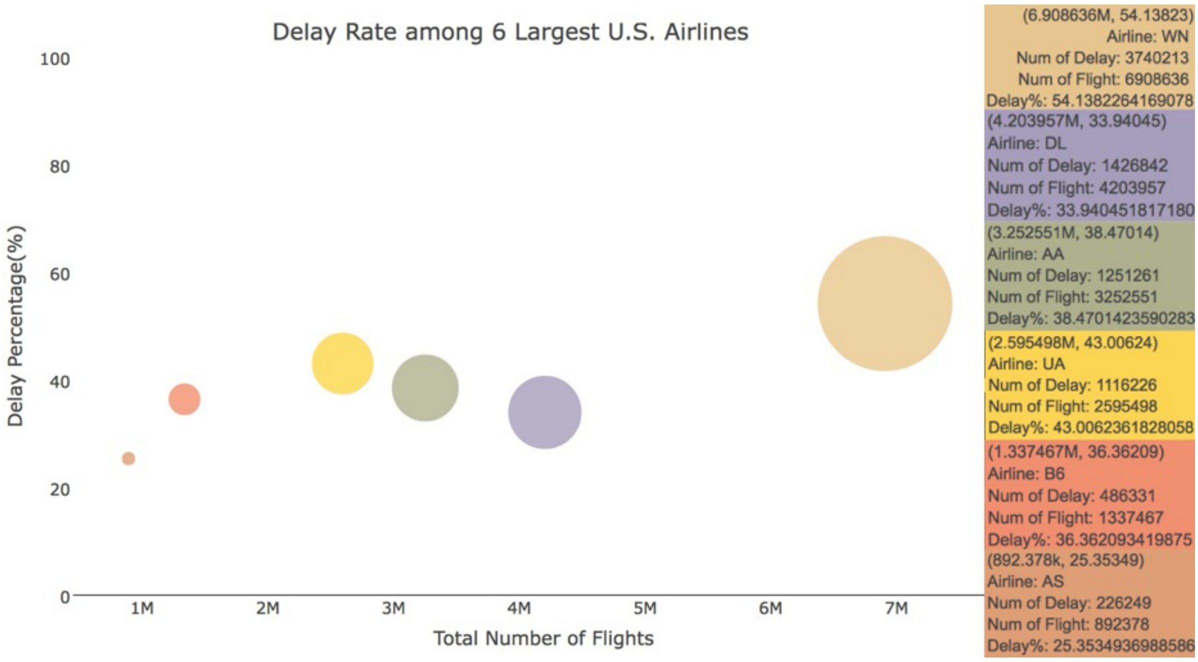
```
#Filter the six largest airlines in the U.S. and produces ggplot showing year as x and delay percentage as y, dist
inguished by airlines.
delay_percent <- rbind(delay_percent_09,delay_percent_10, delay_percent_11, delay_percent_12, delay_percent_13, de
lay_percent_14, delay_percent_15,delay_percent_16) %>% filter(grepl("(AS|AA|DL|B6|WN|UA)", UNIQUE_CARRIER))
delay_percent %>%
ggplot(aes(x= year, y = delay_percentage, col=UNIQUE_CARRIER))
+geom_point()+geom_smooth(method="loess", span=3, se=FALSE)+ labs(title="Delay Percentage")+ theme(plot.title=elem
ent_text(size=20),
axis.text=element_text(size=10), axis.title = element_text(size=13))+
scale_fill_discrete(name="CARRIER")
```



here’s the graph

It’s not hard to notice that Alaska Airlines has the best performance in this category. It has not only the lowest delay percentage in every single year (all below 12%), but also an overall decreasing trend. Delta’s delay percentage is on average 5% higher than Alaska’s each year, and has very similar trend with Alaska, resulting in the second best performance. When we look at the top of the graph, we saw that Southwest Airlines has on average a 21% delay rate, which is the highest, reflecting the worst performance. However, it began to improve after 2013. Starting from 2015, JetBlue has replaced Southwest to become the airline with the highest total delay percentage. Nevertheless, delay percentage can only reflect the portion of delayed flights over total flights, but I am also curious about the total number of delays and the total number of flights over the years to compare the overall population size being affected.

```r
#Create bubble chart to further show the delay percentage for each airlines by using plotly
library(plotly)
total_flight <- completedata %>% group_by(UNIQUE_CARRIER)%>% summarise(total = n())
delay_flight <- completedata %>% filter(DEP_DELAY > 0) %>% group_by(UNIQUE_CARRIER)%>% summarise(total_delay = n())
table_complete <- total_flight %>% left_join(delay_flight) %>% mutate(delay_percentage = (total_delay/total) * 100) table_analyzed <- table_complete %>% filter(UNIQUE_CARRIER == "AA"|UNIQUE_CARRIER == "AS"|UNIQUE_CARRIER == "B6"|UNIQUE_CARRIER == "WN"|UNIQUE_CARRIER == "DL"|UNIQUE_CARRIER == "UA")
p <- plot_ly(table_analyzed, x = ~total, y = ~delay_percentage, type = 'scatter', mode = 'markers', size = ~sqrt(total_delay /pi), color = ~UNIQUE_CARRIER, text = ~paste('Airline:', UNIQUE_CARRIER,'<br>Num of Delay:', total_delay,'<br>Num of Flight:', total, '<br>Delay%:', delay_percentage), marker = list(symbol = "circle",opacity = 0.8, sizemode = 'diameter')) %>% layout(title = 'Delay Rate among 6 Largest U.S. Airlines', xaxis = list(title = "Total Number of Flights", showgrid = FALSE), yaxis = list(title = "Delay Percentage(%)", showgrid =FALSE, range = c(0,100)), showlegend = FALSE)
p
```



(6.908636M, 54.13823)
Airline: WN
Num of Delay: 3740213
Num of Flight: 6908636
Delay%: 54.1382264169078

(4.203957M, 33.94045)
Airline: DL
Num of Delay: 1426842
Num of Flight: 4203957
Delay%: 33.940451817180

(3.252551M, 38.47014)
Airline: AA
Num of Delay: 1251261
Num of Flight: 3252551
Delay%: 38.4701423590283

(2.595498M, 43.00624)
Airline: UA
Num of Delay: 1116226
Num of Flight: 2595498
Delay%: 43.0062361828058

(1.337467M, 36.36209)
Airline: B6
Num of Delay: 486331
Num of Flight: 1337467
Delay%: 36.362093419875

(892.378k, 25.35349)
Airline: AS
Num of Delay: 226249
Num of Flight: 892378
Delay%: 25.3534936988586

Here’s another graph

I therefore drew a bubble chart using “plotly.” Instead of putting years on the horizontal axis to see the trend for each airline, this time I chose to put the total number of flights as our variable. Moreover, the size of the bubbles reflects the total number of delays from 2009 to 2016, the height of each represents their specific average delay percentage, and the colorful legend on the right side of the graph provides more detailed information with exact numbers. I observed from the bubble chart that from 2009 to 2016, Southwest has not only the highest delay percentages, but also the biggest total number of delays, affecting the largest passenger population size compared to other airlines. In contrast, Alaska Airlines has the lowest delay rate and the smallest total delay number, again proving to be the best.

Conclusion

In summary, if we solely take into consideration of the six airlines’ delay performances, we found out that Alaska Airlines is the best, with Delta Air Lines ranked the second. United Airlines and American Airlines have moderate performances. However, the two largest low-cost carriers in the U.S., Southwest and JetBlue, perform not as satisfactory as the rest.

References

1. Police violently drag man from United plane after airline reportedly overbooked flight
2. Study of customer satisification in airline industry
3. United States Bureau of Transportation
4. List of largest U.S. airlines
5. Flight cancellation and delay
6. R documentation on plotly
7. resource on how to draw smooth lines