# Analysis of U.S. Airlines Cancellation and Denial of Borading Rate Using Data Visualization

*Joe Zou*

*2017/11/31*

## 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. So, it might occur to you that which airline has the best performance in case of cancellation and denied boarding?

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 performance of low-cancellation rate and involuntarily denied boarding 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 performance based on cancellation rate and involuntarily denied boarding rate?**

### 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). The website allows us to choose the variables we need before downloading it as a csv file. I chose variables that were most related to our topic, such as carrier, cancelled flight indicator, and causes of cancellation.

Because of the shocked incidence of United Airlines passenger being dragged from an overbooked flight, I also decided to take a closer look at the number of passengers who are denied boarding involuntarily for each airline. I also found the data, Passengers Denied Confirmed Space Report, from BTS. In order to align these two data sets, I chose to analyze the data of Passengers Denied Confirmed Space Report from the year of 2009 to 2016. Since the original data contains other information not related to our research, I first filtered out the information I need, which is the total number of passengers denied boarding involuntarily for the airlines we are interested. However, since the website also only allows users to download the data of one quarter of year each time, I combined all data together first in yearly basis. Then, I changed the number to numeric class, and added a new column showing the sum of passengers denied boarding involuntarily for each airline every year.

To get a larger data set, data from 2009 to 2017 is collected and combined using the following example of command lines.

```
data_n <- read.csv("/Users/Joezou/Desktop/Stat 133/data/n.csv")
completedata_n <- rbind(data_1,data_2,data_3,...,data_n)
```

```
#Overbooking data scraping for 2016 as an example
OB16_1q <- read.csv("/Users/Joezou/Desktop/Stat 133/data/2016_1q_0.csv") %>% select(CARRIER, X3)
OB16_2q <- read.csv("/Users/Joezou/Desktop/Stat 133/data/2016_2q_0.csv") %>% select(CARRIER, X3)
OB16_3q <- read.csv("/Users/Joezou/Desktop/Stat 133/data/2016_3q_0.csv") %>% select(CARRIER, X3)
OB16_4q <- read.csv("/Users/Joezou/Desktop/Stat 133/data/2016_4q_0.csv") %>% select(CARRIER, X3)
OB16_1 <- OB16_1q %>% left_join(OB16_2q,by = c(CARRIER = "CARRIER"))
OB16_2 <- OB16_1 %>% left_join(OB16_3q,by = c(CARRIER = "CARRIER"))
OB16 <- OB16_2 %>% left_join(OB16_4q,by = c(CARRIER = "CARRIER")) %>%
filter(CARRIER %in% c("American Airlines","Alaska Airlines","Delta Air Lines","Southwest Airlines","United Air Lin
es","JetBlue Airways"))
OB16['2016'] = apply(apply(apply(OB16[,-1],1,FUN = function(x) gsub(",","",x)),1,FUN = function(x) as.numeric(x)),
1,FUN = function(x) sum(x))
# combining the whole 8 years of overbooking data
OB_Total <- OB09 %>%
left_join(OB10,by = c(CARRIER = "CARRIER")) %>% left_join(OB11,by = c(CARRIER = "CARRIER")) %>% left_join(OB12,by
= c(CARRIER = "CARRIER")) %>% left_join(OB13,by = c(CARRIER = "CARRIER")) %>% left_join(OB14,by = c(CARRIER = "CAR
RIER")) %>% left_join(OB15,by = c(CARRIER = "CARRIER")) %>% left_join(OB16,by = c(CARRIER = "CARRIER")) %>%
select(CARRIER, `2009`, `2010`, `2011`, `2012`, `2013`, `2014`, `2015`, `2016`) %>%
filter(grepl("Alaska Airlines|American Airlines|Delta Air Lines|JetBlue Airways|Southwest Airlines|United Air Line
s",CARRIER))
```

### Overview of the analysis

Cancellation is an important aspect reflecting an airline's performance, and it can more severely negatively impact passenger satisfaction. I first compared the airlines' average overall cancellation rates over the years. Then, after again considering the controllable factors and uncontrollable factors, I decided to draw a pie chart for each airline to show the portion of cancellations occurred solely due to the airline's responsibility; and the larger this portion is, the worse the airline's cancellation status is.

Another aspect I focused on was the cases where passengers were involuntarily denied boarding. Because for most of the time, when someone is denied boarding, he/she will definitely not make to their destination on time. According to federal law, airlines have the rights to have their tickets overbooked and involuntarily deny boarding to anyone as long as they provide reasonable compensation to the passengers since the statistical chance of all passengers with a valid ticket actually showing up is less that 1/10,000. Just like the previous factor, I first compared the airlines' total number of passengers being denied using box plot. Next, I drew a regression line for each airline wanting to observe their trend over the years.

## Data Visualization and Analysis through Plots

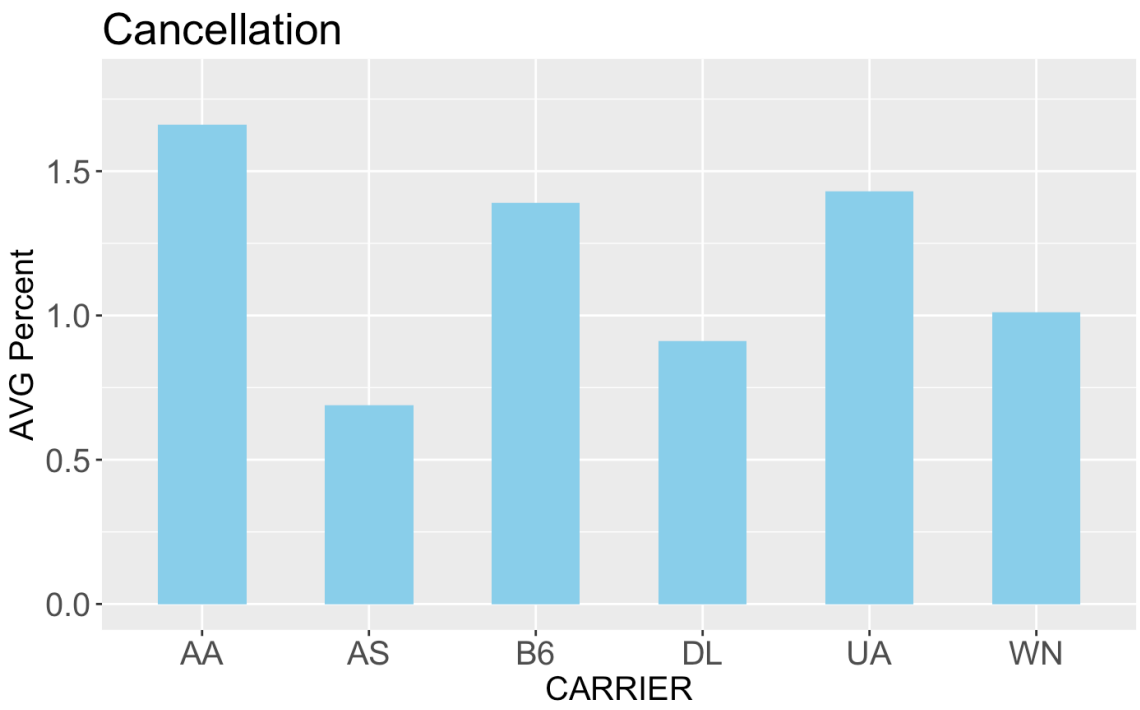
### Cancellation rate on a bar chart

To assess the aspect of flight cancellation, I calculated the average rate of cancellation for each airline of 8 years. I first selected the flights that

had a canceled value of 1 in our dataset for each corresponding year and got the total number of canceled flights. Then, I divided that value with the total number of flights in the corresponding year. Finally, I calculated the average percentage over the years by summing the separate percentages and divide by eight. Average percentage is a reasonable test statistic in this case due to the differentiating airlines' sizes. The total numbers of cancellation for each airline were incomparable because some airlines simply had more total number of flights, thus more cancellation. For instance, in 2009, American Airline had 9,987 cancellation, Alaska had 1,809, JetBlue had 1,924, Southwest had 8563, Delta had 5,793, and United had 6,907. By converting the total number of canceled flight into percentages, we eliminated the difference of size among the 6 airlines to better evaluate the airlines' on-time performances.

The average of approximately 1 percent may seem insignificant or trivial. However, considering that the major airlines have over 5,000 flights per day, even 0.1% counts. The numbers of cancellation adding up (e.g. the numbers as listed above) would be noteworthy and consequential for a customer while considering which airline to choose.

```
# example of data scraping for cancellation as well as their percentage.
new_n_cancel <- completedata_n %>% filter(CANCELLED == 1) %>%
group_by(UNIQUE_CARRIER) %>% summarise(total_cancel = n()) new_n_total <- completedata_n %>% group_by(UNIQUE_CARRI
ER) %>% summarise(total = n())
new_n_cancel_Percentage <- new_n_total %>% left_join(new_n_cancel) %>% mutate(cancel_percentage = (total_cancel/to
tal) * 100) head(new_n_cancel_Percentage)
# calculating the percentages
cancelpercentage <- rbind(new_09_cancel_Percentage, new_10_cancel_Percentage, new_11_cancel_Percentage, new_12_can
cel_Percentage, new_13_cancel_Percentage, new_14_cancel_Percentage, new_15_cancel_Percentage, new_16_cancel_Percen
tage) %>% filter(grepl("(AS|AA|WN|B6|DL|UA)", UNIQUE_CARRIER))
cancelpercent_AA<-cancelpercentage%>% filter(UNIQUE_CARRIER=="AA")
mean(cancelpercent_AA$cancel_percentage)
cancelpercent_AS<-cancelpercentage%>% filter(UNIQUE_CARRIER=="AS")
mean(cancelpercent_AS$cancel_percentage)
cancelpercent_B6<-cancelpercentage%>% filter(UNIQUE_CARRIER=="B6")
mean(cancelpercent_B6$cancel_percentage)
cancelpercent_WN<-cancelpercentage%>% filter(UNIQUE_CARRIER=="WN")
mean(cancelpercent_WN$cancel_percentage)
cancelpercent_DL<-cancelpercentage%>% filter(UNIQUE_CARRIER=="DL")
mean(cancelpercent_DL$cancel_percentage)
cancelpercent_UA<-cancelpercentage%>% filter(UNIQUE_CARRIER=="UA")
mean(cancelpercent_UA$cancel_percentage)
cancelpercent <- data.frame(c("AA","AS","B6","WN","DL","UA"), c(1.66, 0.69, 1.39, 1.01,
0.91, 1.43))
names(cancelpercent) <- c("CARRIER","AVG")
#Create bar graphs of cancel percentage for the six airline companies
cancelpercent %>%
ggplot(aes(x= CARRIER, y = AVG)) +geom_bar(stat="identity",fill="skyblue", width
=0.53)+ylim(0,1.8)+
labs(title="Cancellation", y = "AVG Percent") + theme(plot.title=element_text(size=20), axis.text=element_text(siz
e=15), axis.title = element_text(size=15))
```




As we can see from the graph, Alaska, Delta, and Southwest performed relatively well. Their average cancellation rates were completely below 1%. American Airlines had the worst performance, which was about 1.7%. JetBlue and United Airline had a slightly better performance, but still exceeding 1.25%.

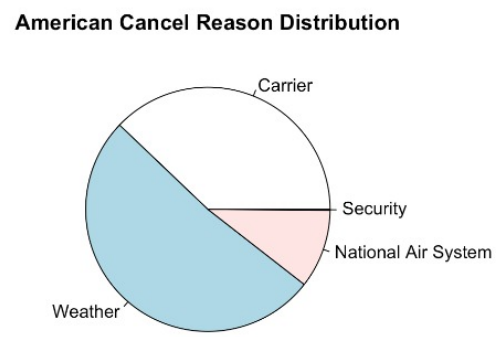
### Cancellation reasons on pie charts

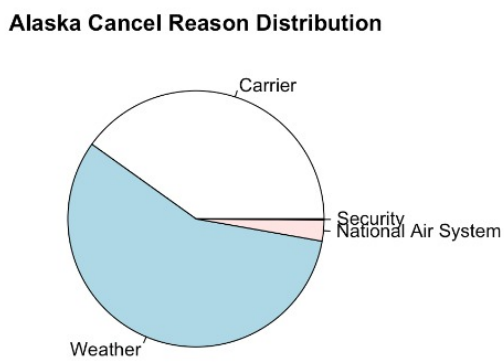
While the cancellation rate of each airline was telling us a lot, I wanted to further analyze the cancellation reason distribution in order to examine whether an airline's cancellation was reasonable or not. For some cases, maybe the airport locations - which affect an important factor of cancellation, weather - would influence the cancellation rate. So, it is important to analyze whether the cancellation was caused by the airline

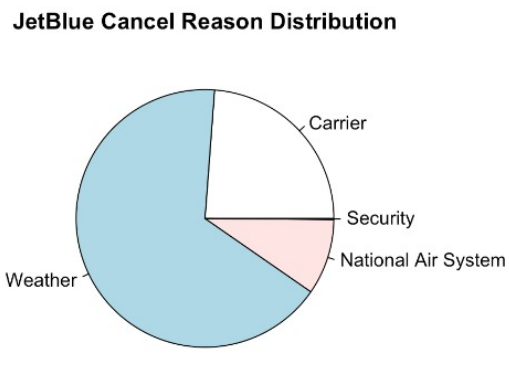
carrier.

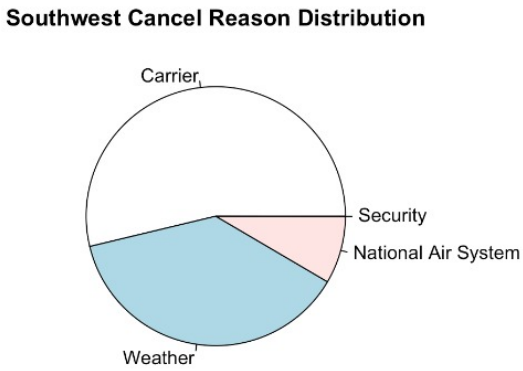
There are four types of cancellation in total - cancellation caused by carrier, security, weather, or National Air System (e.g. air traffic control). Cancellation caused by security, weather, or NAS is considered as justified, since factors like this cannot be controlled manually by the airline carriers themselves. It was nevertheless truth that the cancellation caused by carrier was the kind of cancellation we wanted to investigate. By grouping the airline cancellation reasons, we could calculate the percentage distribution of each cancellation reason.

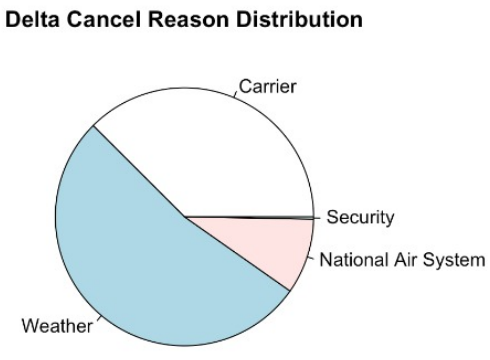
```
# Create pie chart respectively for airline companies, using AA as an example of calculation
AAcr09 <- completedata09 %>% filter(CANCELLED=="1", UNIQUE_CARRIER=="AA") %>%
summarise(totalcancel=n())
AAcr09_2 <- completedata09 %>% filter(CANCELLED=="1", UNIQUE_CARRIER=="AA") %>% group_by(CANCELLATION_CODE) %>% su
mmarise(cancel_reason=n()) %>% mutate(cancel_reason=cancel_reason/AAcr09$totalcancel*100)
AArc_A <- rbind(AAcr09_2,AAcr10_2,AAcr11_2, AAcr12_2, AAcr13_2, AAcr14_2, AAcr15_2, AAcr16_2) %>%
filter(CANCELLATION_CODE=="A")
mean(AArc_A$cancel_reason)
AArc_B <- rbind(AAcr09_2,AAcr10_2,AAcr11_2, AAcr12_2, AAcr13_2, AAcr14_2, AAcr15_2, AAcr16_2) %>%
filter(CANCELLATION_CODE=="B")
mean(AArc_B$cancel_reason)
AArc_C <- rbind(AAcr09_2,AAcr10_2,AAcr11_2, AAcr12_2, AAcr13_2, AAcr14_2, AAcr15_2, AAcr16_2) %>%
filter(CANCELLATION_CODE=="C")
mean(AArc_C$cancel_reason)
AArc_D <- rbind(AAcr09_2,AAcr10_2,AAcr11_2, AAcr12_2, AAcr13_2, AAcr14_2, AAcr15_2, AAcr16_2) %>%
filter(CANCELLATION_CODE=="D")
mean(AArc_D$cancel_reason)
#Pie chart
slices <- c(37.89,51.62, 10.44,0.11)
lbls <- c("Carrier","Weather","National Air System", "Security")
pie(slices, labels=lbls, main="American Cancel Reason Distribution")
```

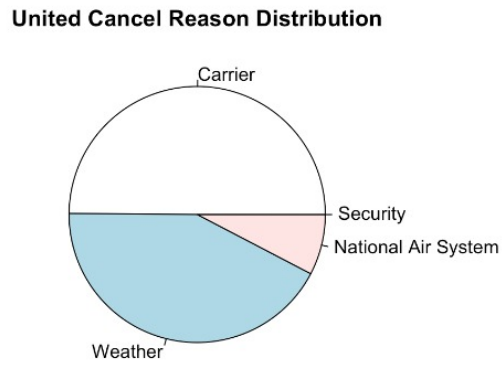



The white segment represents cancellation caused by the carriers themselves, and others are caused by uncontrollable factors. So, the larger the white slice is, the worse is the airline's performance. From the graph, we concluded that Southwest was the worst airline, considering approximately 54% of the cancellations are caused by Southwest itself. It was the only airline that had a carrier caused cancellation rate surpasses 50%. United Airline was also not very ideal with its 49% carrier caused cancellation. Alaska had a surprisingly high rate of 40%. However, Alaska had the lowest cancellation rate among all its opponents. The rest were keeping a relatively good record of about 30~37%.

Based on our analysis of the two graphs, we can see that Alaska was the best airline on the evaluating aspect of cancellation, with lowest

average cancellation rate and middle carrier caused cancellation. United Airline was the worst, with its both second worst cancellation rate and carrier caused cancellation rate. American Airline and Southwest were also not very suitable choices for airline customers who want their airline to be on time: one for its highest cancellation rate among all the 6 airlines, and one for its highest carrier caused cancellation rate.

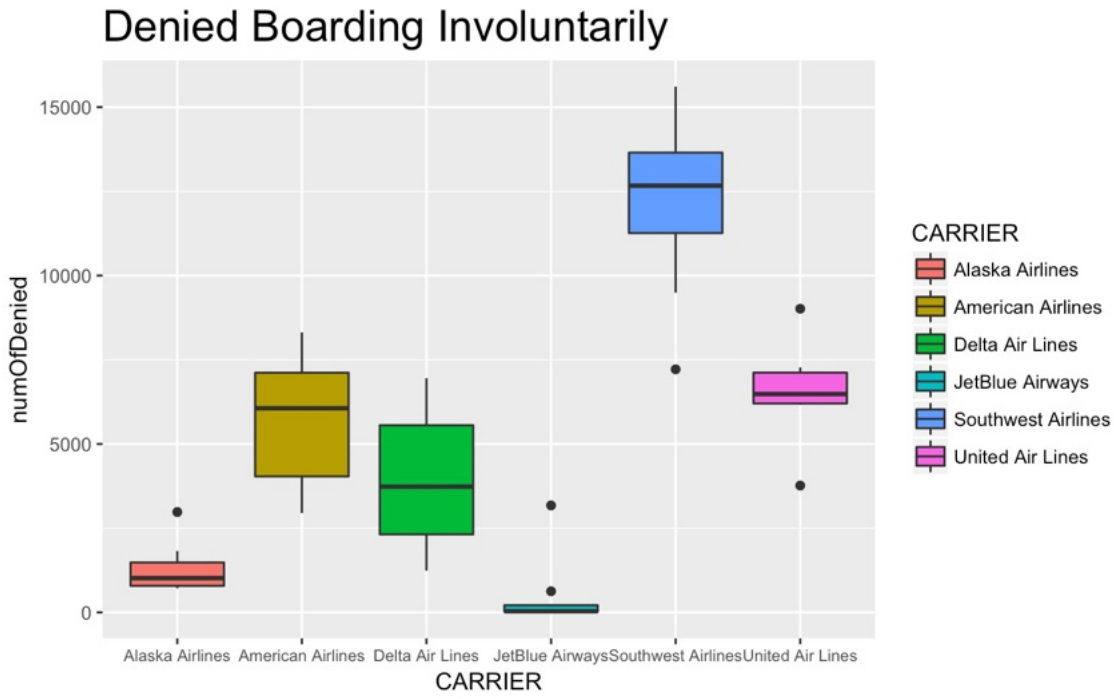
## Overbooking data on a boxplot

From the recent news of United Airlines passenger being forcibly removed from an overbooked flight, the public begins to get aware of the airlines' practice of overbooking and their solutions when facing overbooked flights. First of all, what is overbooking? Overbooking is when airlines sell more tickets than the available seats in the plane with the assumption that there will be a certain percentage of "no shows". The commercial reason behind overbooking is for airlines to ensure they fill their planes up as much as possible, and also to maximize their profits. However, while overbooking is common practice, the number of people who are denied boarding, both voluntarily and involuntarily, is in fact relatively tiny. When overbooking occurs, the first step for airlines is to offer some form of compensation for passengers. Some passengers may be willing to have the compensation and take the next arranged flights. However, if there are not enough volunteers, it is legal for the airline to deny some of their passengers boarding against their will for some type of compensation.

Nevertheless, it is the airlines' responsibility to ensure the least number of people are affected by their overbooking strategy. For our analysis, we will focus on passengers who are mostly affected by overbooking flights, which are the passengers denied boarding involuntarily. Our data contains the information of total number of passengers who are denied boarding involuntarily from the year of 2009 to 2016.

Based on such information, I created a box plot comparing all six airline carriers.

```
OB_Total_Narrow <- OB_Total %>%
gather(key = year, value = numOfDenied, `2009`, `2010`, `2011`,`2012`, `2013`, `2014`,
`2015`,`2016`)
OB_Total_Narrow %>%
ggplot(aes(x= CARRIER, y = numOfDenied)) + geom_boxplot(aes(group = CARRIER, fill = CARRIER))
+ labs(title="Denied Boarding Involuntarily") + theme(axis.text.x=element_text(size=7.5),
plot.title=element_text(size=20))
```




The box plot clearly indicates the median and range of the number of passengers who are denied boarding involuntarily during the past eight years. The plot shows that Southwest Airlines has the most number of passengers who are denied boarding involuntarily, with a median of 12,714 and a huge range of 8,392. Since our data is based on only eight consecutive years, its high median and big difference show that Southwest Airlines' forecast for no shows is not very accurate and fluctuates dramatically. In addition, even though Southwest Airlines has an outlier much lower than its averaged number, the outlier is as large as other airlines' average, or even largest, data point. Southwest Airlines' large number of passengers denied boarding involuntarily and the huge fluctuation of its numbers show that Southwest Airlines does not perform well in term of its overbooking strategy and forecast.

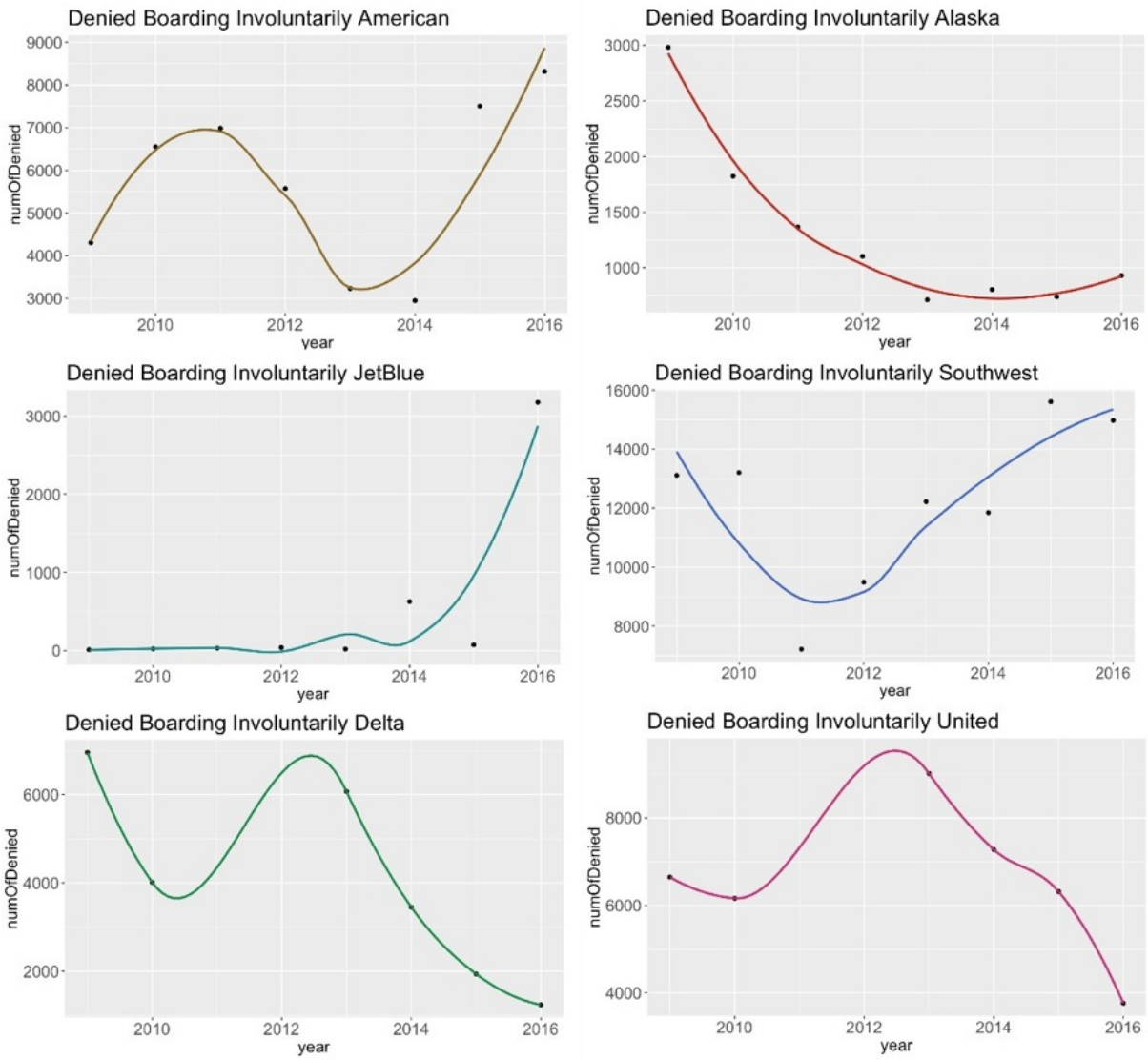
Compared to Southwest Airlines, Alaska Airlines and JetBlue show a quite different situation. Alaska Airlines is the second lowest airline of the number of people denied boarding involuntarily, and the range of its numbers is relatively small. Its data represents that Alaska Airlines has the ability to accurately and stably forecast no shows, and satisfy its flying passengers. Similar to Alaska Airlines, JetBlue also has small number of people who are denied boarding involuntarily. Even though it has extreme large outliers, compared to other airlines, these outliers can be seen as normal or even low point.

From the visual representation of the number of people who are denied boarding involuntarily, Southwest Airlines, instead of American Airlines, surprisingly, performs the worst. And Alaska Airlines and JetBlue do a good job satisfying their passengers on their overbooking strategy.

## Overbooking regression

I further developed graphs for each of the airline carrier respectively to show their trends over the past eight years and also forecast for the following years. With the providing data, I got regression line for each of the graph.

```
# regression plot using Alaska as an example
OB_regression <- OB_Total_Narrow %>% filter(CARRIER == "Alaska Airlines") %>%
  mutate(year = as.numeric(year)) %>% ggplot(aes(x = year, y = numOfDenied)) + geom_point() + geom_smooth(method =
'loess', color = "#CC0000", se = FALSE) + labs(title="Denied Boarding Involuntarily Alaska") +
  theme(plot.title = element_text(size = 20), axis.text = element_text(size = 15), axis.title = element_text(size
= 15))
OB_regression
```




Based on the direction of the regression lines, Alaska Airlines, Delta Airlines, and United Airlines all show trends of decreasing their number of people denied boarding involuntarily. Alaska Airlines, started with one of the few number of denying passengers, has continually put the effort of dragging down its number throughout the years, and had its lowest point of 714 in 2013. Delta Airlines and United Airlines started with similar numbers in 2009, 6956 and 6645 respectively, both got increased in the first few years, and drastically decreased their numbers in the last few years. Both airline carriers had much lower number in 2016 than they had in 2009. In contrast, American Airlines, JetBlue, and Southwest Airlines ended up with bigger number of passengers denied boarding involuntarily than they had in 2009, and their regression lines show they will continue to increase their numbers in following years. For both American Airlines and Southwest, even though their numbers decreased in early years, they have continually set new highs in recent days. JetBlue's number is the most drastically increasing one among the six airlines. Even though JetBlue has the lowest number of 9 in the year of 2009, it seemed to find out the profitability of overbooking and thus has increased its number of involuntarily denied boarding passengers rapidly, especially in the year of 2016. In 2015, JetBlue had its number of 73; in the year of 2016, however, its number climbed up to 3176, increasing by more than 4200%! Therefore, even though the first box plot shows that JetBlue has lower numbers historically, it has the trend of strengthening its overbooking strategy and getting larger number of people who will be denied boarding involuntarily.

In conclusion, based on these two sets of graphs, in terms of these airlines' overbooking strategy, Alaska Airlines performs the best, followed by JetBlue Airways, Delta Air Lines, United Airlines, American Airlines, and with Southwest Airlines performs the worst.

## Conclusion

In summary, based on the six airlines' flight cancellation rate and involutarily denied boarding rate, we found out that Alaska Airlines is the best, with Delta Airlines ranked the second. JetBlue Airline has moderate performances. However, Southwest, American, and United Airlines perform not as satisfactory as the rest. In fact, if you wish your flight is neither cancelled nor overbooked, you should not take a flight provided by any of these three.

# 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. Passengers Denied Confirmed Space Report
7. resource on how to draw box plot using gg plot