BIG DATA: AIRLINE DELAYS

Daniel Antalan

CIS 320 Professor Thomas

03/02/16

Introduction

Part of our everyday lives is spent on the move whether it be walking, driving, or flying. When flying to a destination, we often think about whether we will depart and/or even arrive on time. This begs the question: how frequent do flights get delayed? And when the delay occurs, what are the major factors that have resulted in that particular delay. Despite a world well-connected through technology, flying has remained one of the most popular modes of transportation for various reasons, such as for business or vacation.

Hypotheses

Based on the questions mentioned, I have developed a few hypotheses I would like to try to prove. My first hypothesis is that I believe about 25% of flights are delayed according to my own flight history. I have been on 26 flights within the last two years and 6 of those 26 flights have been delayed. My next hypothesis is more specific. Taking the data regarding the flights that were delayed, I propose that the cause of delays for flights in the winter months (December to February) were the result of weather delays more than the summer months (June to August). Since the United States experiences the most unfavorable weather for aircraft during the winter months, it would make sense that the reason for flights being delayed at this timeframe would be due to the inclement weather. The winter months would also be a reasonable amount of data to utilize since approximately 23.3% of flights within the year compared to the 26.4% of the summer months.

Methods

To provide proof for the hypotheses, data from the Bureau of Transportation Statistics will be extracted, cleaned, organized, and tested for analysis. For simplicity, I took the data from flights

within the domestic United States and within the last year (December 2014-November 2015). November 2015 was the recent month the Bureau of Transportation Statistics had on file in the database. To simplify matters further, I only considered the flights that were delayed, not cancelled. Cancelled flights obscured the data for delays because of their missing information. Consequently, to clean the data for delays, I filtered out cancelled flights. With such a large volume of data, I extracted the data for every month of the period into Comma Separated Value files. I then compiled the data into IBM SPSS Statistics program and saved it as .sav file, giving me a file with all the consolidated information I would need to run various types of analysis. Using IBM SPSS Statistics even further, I ran the frequency distribution as shown below in **Figure 1**. By having the frequency distribution, I was able to find out which months experienced the most flights and which ones experienced the least. Using this spectrum, I was able to gather further insights about the data. As aforementioned, the summer months experienced the most, mainly due to a time period known for vacationing. Conversely, the winter months experienced the least amount of flights despite being also known for a time to vacation. The underlying reason could be that less flights are scheduled as a result of the poor weather conditions. Along with IBM SPSS Statistics, I also loaded the data through R to obtain descriptive statistics as shown in **Figure 2** and create a pie chart (**Figure 3**) based on the frequency distribution displayed in IBM SPSS Statistics. I also utilized R to find the rate at which flights were delayed relating to one of my hypotheses. In addition, I particularly focused on the weather delay variable in both the summer and winter months since it pertained to my other hypothesis. Lastly, I ran a two sample t-test through R to determine the significance of the means for the weather delays in the summer and winter.

Analysis

In terms of the first hypothesis (how often flights are delayed throughout a one year period), I was able to obtain the rate for both arrival delays and departure delays. For simplicity, whether it be classified as arrival or departure, I generally categorized it as delays. But to give an idea of how much of a difference the rates were, I ran the data through R (**Figure 4**). For departure delays, R calculated 18.52% whereas for arrival delays it was 18.85%. In general, from December 2014 to November 2015, about 19% of flights were delayed, which is substantially less compared to the rate of 25% I hypothesized based on my flight history.

The second and final hypothesis dealt specifically with the weather delays in the summer and winter months. I proposed that weather delays had a prominent occurrence in winter months due to the unfavorable conditions that happen during this period. Their rate was compared to the weather delays in the summer months since one would assume that summer weather would work in favor of aircraft. When calculating the rates, I again used R as shown in **Figure 5**. Initially, I used the total flights in the given seasonal period as my reference when finding the rate of weather delays. However, as seen in **Figure 6**, I used an R command to omit the missing data from flights that were not delayed in order to get the rate of weather delays that occur from the total amount of delays. This would give a better understanding of how often delays are caused by weather delays. For winter months, the rate was about 6.4% while for summer, it was around 6.0%. By simply comparing the rates, it exhibits that relatively speaking flights in the winter months experience slightly more weather delays than do the summer months. However, when taking into account the mean, or average, time for delays, weather delays in the winter took an average of 2.71 minutes whereas the ones in the summer took 2.77 minutes (**Figure 7**). Consequently, weather delays were much frequent in the winter but they took up a longer

amount of time in the summer. To make a more proper assessment of these two means, I decided to test it further and ran a two sample t-test in R (**Figure 8**). The result was a p-value of 0.1108 which clearly is significantly higher than the alpha of 0.05. Therefore, we fail to reject the null hypothesis at the 95% confidence interval that the difference between the mean of the summer months and the mean of the winter months is equal to zero. The result, moreover, is statistically nonsignificant.

Conclusion

After running the calculations and tests in R, my first hypothesis was rejected since the rates that were computed had large difference relative to the rate that I proposed. The calculated rate was approximately 19%, which is significantly less than the predicted 25%. On the other hand, my second hypothesis was supported through the calculating the rates as done for the first hypothesis. Specifically, it was the rate of delays that occurred due to weather delays in the summer and winter months. When considering the means of the two periods, the summer months, on average, had slightly longer delays than the winter months. But after running a t-test to verify, this became statistically nonsignificant. However, this does not mean that the null hypothesis is true; it only signifies that insufficient evidence exists to reject it.

Figure 1: Frequency Distribution of Flights

Frequencies

[DataSet1] E:\Airlines Delay rev.sav

Statistics

MONTH

N	Valid	5713691
	Missing	0

MONTH

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	457013	8.0	8.0	8.0
	2	407663	7.1	7.1	15.1
	3	492138	8.6	8.6	23.7
	4	479251	8.4	8.4	32.1
	5	489641	8.6	8.6	40.7
	6	492847	8.6	8.6	49.3
	7	514384	9.0	9.0	58.3
	8	503956	8.8	8.8	67.2
	9	462153	8.1	8.1	75.2
	10	482878	8.5	8.5	83.7
	11	462367	8.1	8.1	91.8
	12	469400	8.2	8.2	100.0
	Total	5713691	100.0	100.0	

Figure 2: Descriptive Statistics of Flight Data

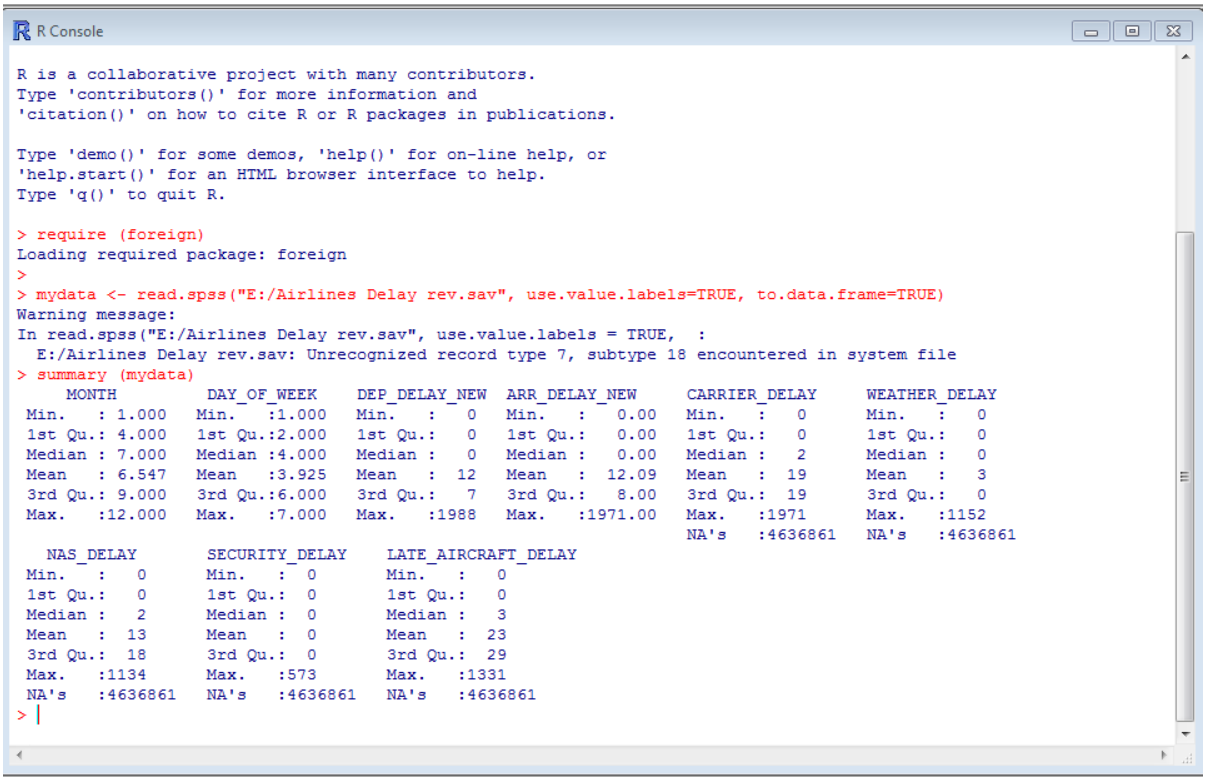
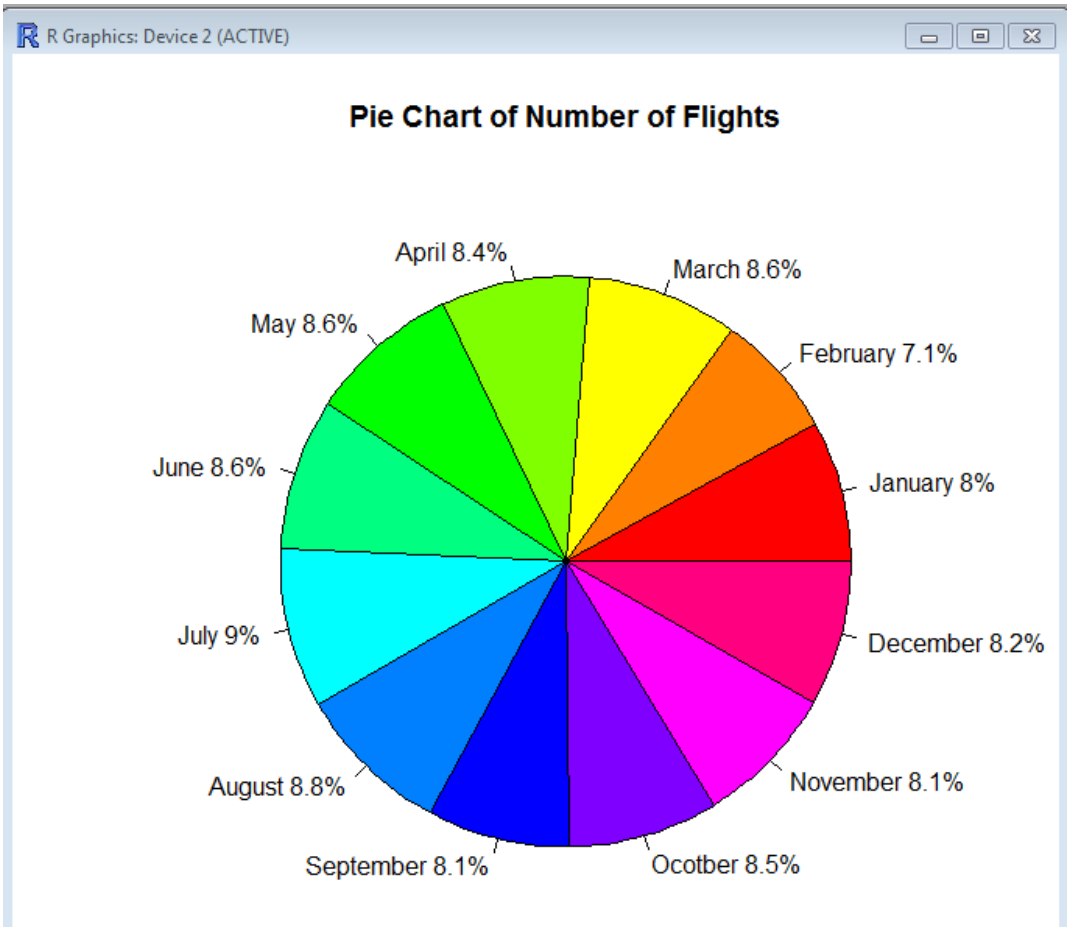


Figure 3: Pie Chart of Frequency Distribution



R Code of Figure 3

```
> slices <- c(8.0, 7.1, 8.6, 8.4, 8.6, 8.6, 9.0, 8.8, 8.1, 8.5, 8.1, 8.2)
> lbls <- c("January", "February", "March", "April", "May", "June", "July", "August", "September", "Ocotber", "November", "December")
> lbls <- paste(lbls, slices) # add percents to labels
> lbls <- paste(lbls,"%",sep="") # ad % to labels
> pie(slices,labels = lbls, col=rainbow(length(lbls)),
+    main="Pie Chart of Number of Flights")
>
```

Figure 4: Delay Rates of Departures and Arrivals

```
> DEP <- (mydata$DEP_DELAY_NEW)
> length(which(DEP >= 15))/length(which(DEP >= 0))
[1] 0.1852353
> ARR <- (mydata$ARR_DELAY_NEW)
> length(which(ARR >= 15))/length(which(ARR >= 0))
[1] 0.1884649
```

Figure 5: Weather Delay Rates for Flights in Winter and Summer Months

```
> length(which(winterdata$MONTH < 13))
[1] 1334076
> length(which(winterdata$WEATHER_DELAY > 0))
[1] 19229
> length(which(winterdata$WEATHER_DELAY > 0))/length(which(winterdata$MONTH < 13))
[1] 0.01441372
```

```
> length(which(summerdata$MONTH < 13))
[1] 1511187
> length(which(summerdata$WEATHER_DELAY > 0))
[1] 18925
> length(which(summerdata$WEATHER_DELAY > 0))/length(which(summerdata$MONTH < 13))
[1] 0.01252327
```

Figure 6: Weather Delay Rates for Delayed Flights in Winter and Summer Months

```
> winteromitdata <- na.omit(winterdata)
> length(which(winteromitdata$MONTH < 13))
[1] 301293
> length(which(winteromitdata$WEATHER_DELAY > 0))
[1] 19229
> length(which(winteromitdata$WEATHER_DELAY > 0))/length(which(winteromitdata$MONTH < 13
))
[1] 0.0638216
```

```
> summeromitdata <- na.omit(summerdata)
> length(which(summeromitdata$WEATHER_DELAY > 0))/length(which(summeromitdata$MONTH < 13
))
[1] 0.05960968
```

Figure 7: Mean and Standard Deviation of Weather Delays in Winter and Summer (in minutes)

```
> mean(winteromitdata$WEATHER_DELAY)
[1] 2.708715
```

```
> mean(summeromitdata$WEATHER_DELAY)
[1] 2.768532
```

```
> sd(summeromitdata$WEATHER_DELAY)
[1] 18.23873
> sd(winteromitdata$WEATHER_DELAY)
[1] 20.1541
```

Figure 8: Two Sample t-Test Comparing the Means in Figure 7

```
> t.test(summeromitdata$WEATHER_DELAY, winteromitdata$WEATHER_DELAY, alternative = "grea
ter")

        Welch Two Sample t-test

data:  summeromitdata$WEATHER_DELAY and winteromitdata$WEATHER_DELAY
t = 1.222, df = 604870, p-value = 0.1108
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.02069646         Inf
sample estimates:
mean of x mean of y
 2.768532  2.708715
```