Can I Get There?  
Exploratory Data Analysis of

2013 NYC Flight Data with R

By: Ben Schmidt, James Rullo, Kshitij Bhat,

Srilatha Lakka, Sam SR

DATS6101: Introduction to Data Science

Prof. Edwin Lo

Project 1

George Washington University

**Introduction:**

Air travel is a massive industry, one which impacts the life of almost every American. As an industry, it has a GDP of $846.3 billion within the US alone. (Airlines for America Homepage, 2018) Growth of the industry has been meteoric in recent years, with a 40.7% increase in the number of flights between 2002 and 2007. (Peterson, Neels, Barczi, & Graham, 2013) In addition to the thousands of jobs it creates, approximately 45% of Americans flew in 2015 (Reed, 2016) and delays cost a collective $32.9 billion yearly, with fully half of that being passed to the customers (Peterson, Neels, Barczi, & Graham, 2013). Due to the way airlines manage airplanes, delays propagate down the schedule. Planes are reused on a tight schedule, so the delay of one flight often delays the next flight using the same plane. Additionally, delay on a single plane can lead to delays for others due to constraints on runways and gates (Martinez, 2012).

To properly understand the cascading nature of flight delays, it is important to understand the process of use for a plane. Most flights are scheduled with what is called a schedule buffer, a built-in amount of time designed to allow for delays. In most cases when a flight arrives “early”, it is simply because it left on time and took the expected amount of time for its trip. The schedule buffer is meant to account for a leaving delay, and is designed to keep customers happy by making them feel they have arrived on time or even early. (Michael Ball, 2010) When a flight is delayed beyond this buffer, it can lead to delays down the line and may also be the result of earlier delays. Many of these delays are unavoidable, such as weather delays. No amount of infrastructure change given current aviation technology will make it acceptably safe to fly a passenger liner in a snowstorm. However, many can be changed with changes to airports or planning procedure. Flying each plane fewer times in the day would reduce load on the system, but it would require more ground space in airports. Likewise, air congestion could be reduced by increasing airport size and number of gates and runways. (Michael Ball, 2010) This, however, would require extensive infrastructure investment, and presumable could result in greater delays during the course of the work.

To address these concerns, we studied the nycflights13 dataset. This is a collection collected of all flights leaving from New York City area airports, including LaGuardia, JFK, and Newark International Airport. These three airports receive a huge amount of traffic every day, leading to a very dense collection of data that can be used for analysis and predictive purposes. This particular dataset reflects the entirety of flights to and from NYC in 2013. It comes built into the R program as “flights” and can be readily accessible through this software program.

In considering this dataset, the following variables were of interest: year, month, day, dep\_time, sched\_dep\_time, dep\_delay, arr\_time, sched\_arr\_time, arr\_delay, carrier, flight, tailnum, origin, dest, air\_time, distance, hour, minute, and time\_hour. Of interest for our purposes, we utilized month (which shows the month of the year the flight occurred), the dep\_delay (the departure delay from leaving New York City), the arr\_delay (the delay of arriving at the destination), and carrier (the airline that operated the flight). The codes for the airlines can be found in Figure 1. By examining these variables, were are focused solely on carrier patterns and which airlines have the largest problems when it comes to delays. We will be assuming that weather and other non-operational factors will be impacting each carrier equally and will not be considered in this particular study. Further due to the issues of major inconvenience of a flight delay and the possibility of missing a connecting flight, only flights delayed greater than 20 minutes are included in this study.

Figure 1: Airline Acronym

|  |  |  |  |
| --- | --- | --- | --- |
| Abbreviation | Airline | Abbreviation | Airline |
| 9E | Endeavor Air Inc. | F9 | Frontier Airlines Inc. |
| AA | American Airlines Inc. | FL | AirTran Airways Corporation |
| AS | Alaska Airlines Inc. | HA | Hawaiian Airlines Inc. |
| B6 | JetBlue Airways | MQ | Envoy Air |
| DL | Delta Air Lines Inc. | OO | SkyWest Airlines Inc. |
| EV | ExpressJet Airlines Inc. | UA | United Air Lines Inc. |
| WN | Southwest Airlines Co. | US | US Airways Inc. |
| YV | Mesa Airlines Inc. | VX | Virgin America |

**SMART Question Development**

There are several SMART questions that can be developed from this dataset and explored through Exploratory Data Analysis. The first question is: Which carrier has the most delays and the worst delays? This question is specific and will identify carriers with problems with delays. This question is measurable, as we can use percentages and and the time ranges of the percentage of each flight by carrier. This can also be measured using a linear regression and an anova test. This question is achievable, as the data collects both these carrier identity and the delay information. The question is relevant because identifying delays is critical to making flying easier and more profitable. The question is timely, because the data is well formatted and will enable quick analysis.

The second question is: Are there flights which often depart earlier than the scheduled time? If yes, do they have early arrival too?. This question is very similar to the previous question and can be measured using a t-test. There are two other questions of interest: Does departure delay can be used to predict the arrival delay? and Do particular carriers have expected delays more often? Analysis can be used to inform the passengers to avoid flights of such carriers in case of urgent travel.

**Summary of Data**

This dataset follows a largely normal structure. The well ordered nature of this dataset will allow us to focus on the issues related to the carriers and their delays. In a brief overview, Figure 2, we can see the percent of flights by each carrier. There is also surprising number of delayed flights in the summer and winter months, perhaps due to higher number of flights. This can be seen in Figure 3.

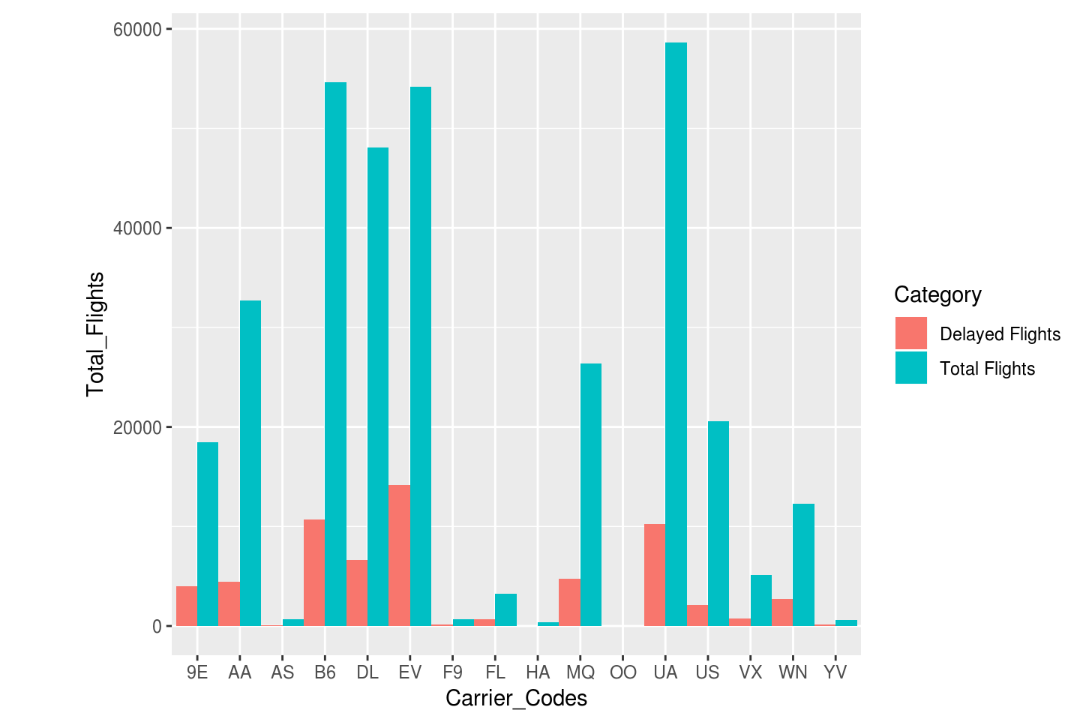
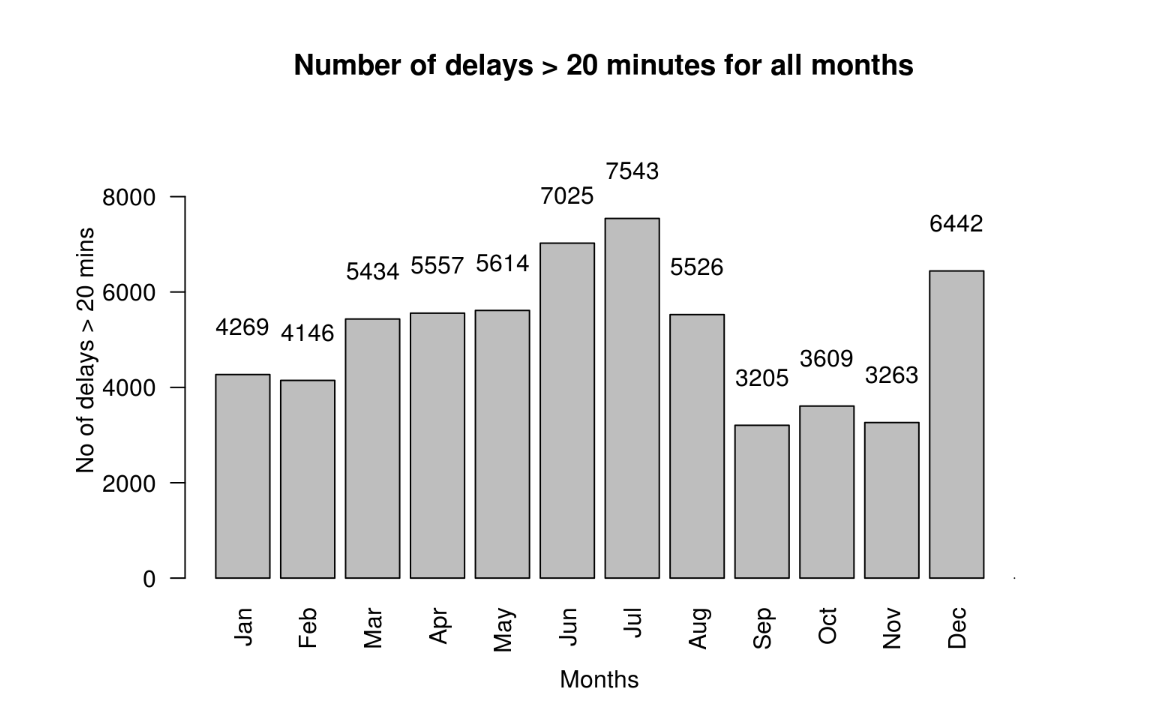
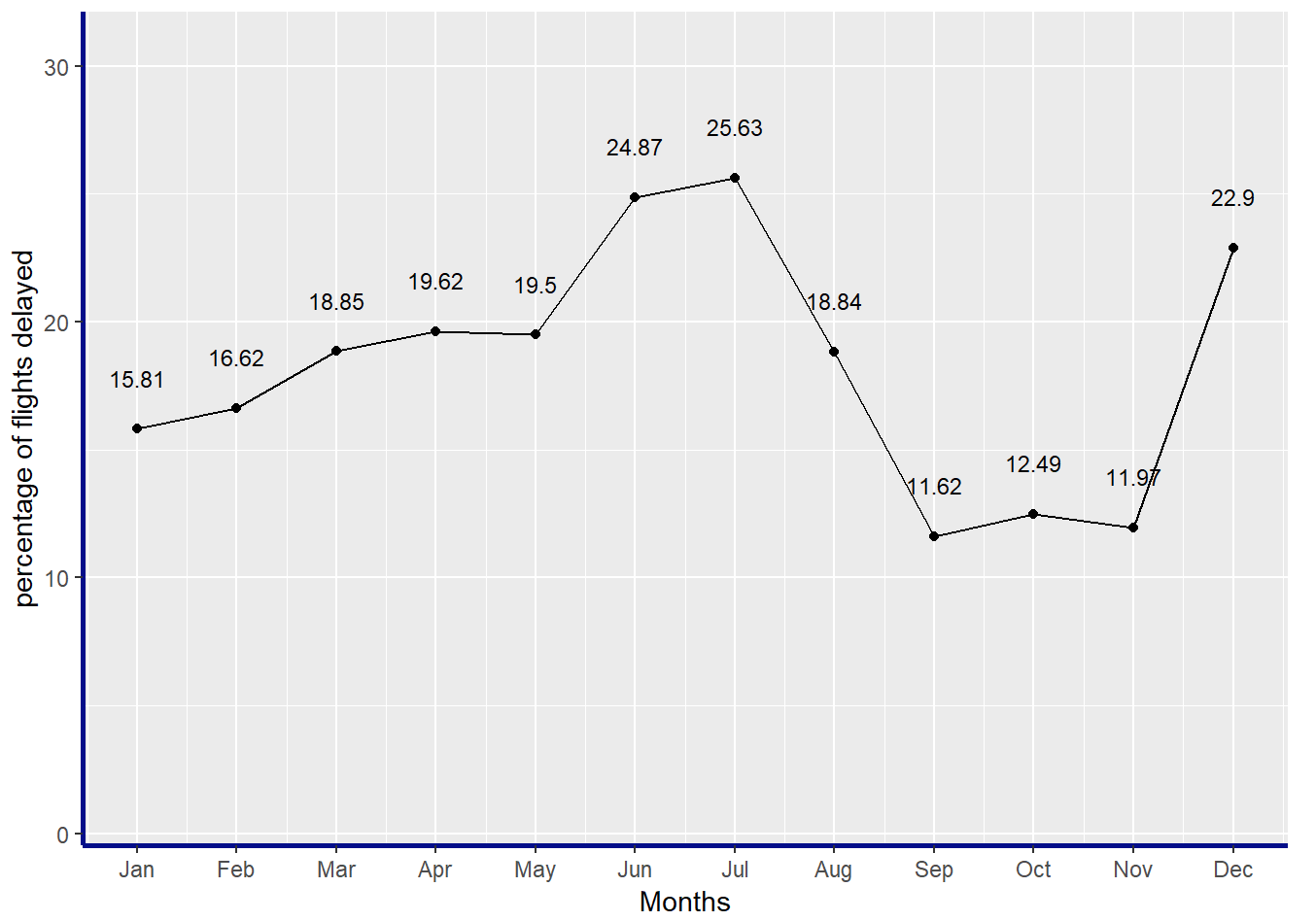
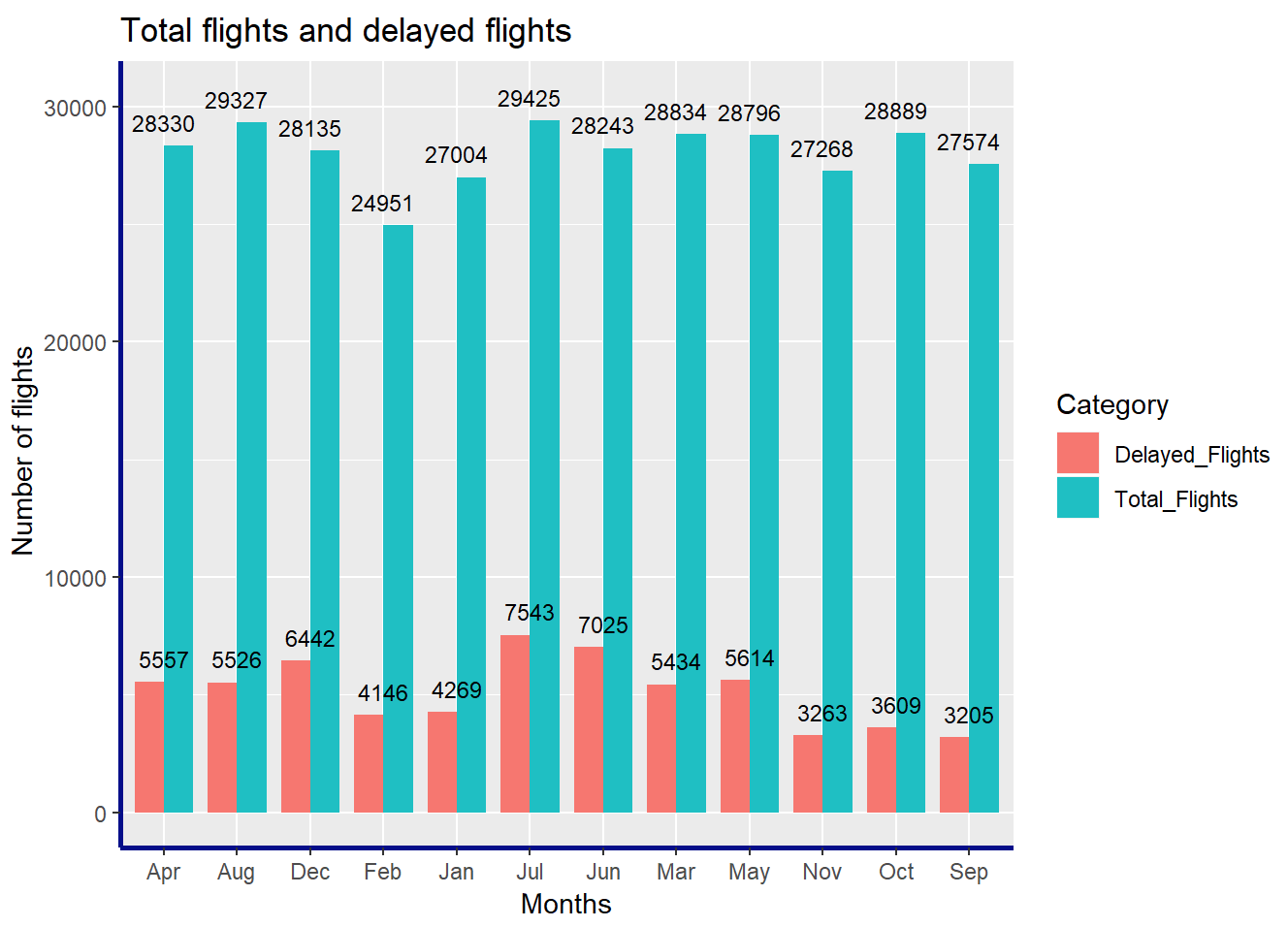
Figure 2:Total flights and delays by different carriers

Figure 3: Number of delays greater than 20 minutes in each month

Consideration of Figure 3 shows that correlating delay to carrier will provide important information helpful in assessing the reliability of various airlines, which will in turn give consumers a tool for choosing a carrier. Likewise, time of day will provide information for customers to make a choice as to when they fly to try and avoid delays. Time of year has a number of confounding factors; weather changes based on the time of year and has an unpredictable effect on delay, but there are also peak travel times that lead to delays. For instance, November and December correlate to snow storms, but also holidays, and this dataset will not be able to identify what the reasons are behind delays.

Figure 4 and Figure 5 show the percentage of flights being delayed in each month, and Comparison of total flights and delayed flights each month respectively. We can see that there is significant rise in the percentage of flights being delayed from the beginning of the year till July, and then a drastic fall till November,and then again gain in the percentages and numbers during december.

Figure 4: Percentage of flights delayed each month

Figure 5: Comparing total flights with delayed flights

**Linear Models: Correlation Between Delays and Carrier**

We performed a linear model to determine if there was any correlation between the arrival delay of flights with the other variables in the data frame. When we ran a model of 3 variables, flight number, distance, and air time, we had a low coefficient of determination at .02275 as our R2. However, when adding a fourth variable of departure delay as well, the coefficient of determination was much higher at .9156. This demonstrates that the second linear model had a much stronger fit and that there is likely a correlation between the arrival delay and departure delay, which makes sense, given that a late flight would arrive late.

**Histogram / Barplot Comparisons Between Carriers**

The following graphs detail the delayed flights per carrier, as a percentage of total flights by that carrier. Each graph shows a different subset of the flights that were delayed.

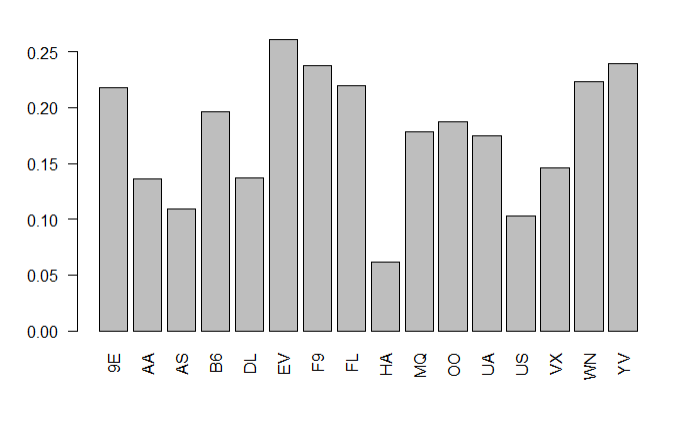


Figure 6: This graph shows, for each carrier, what percentage of their flights in the dataset departed more than 20 minutes late.

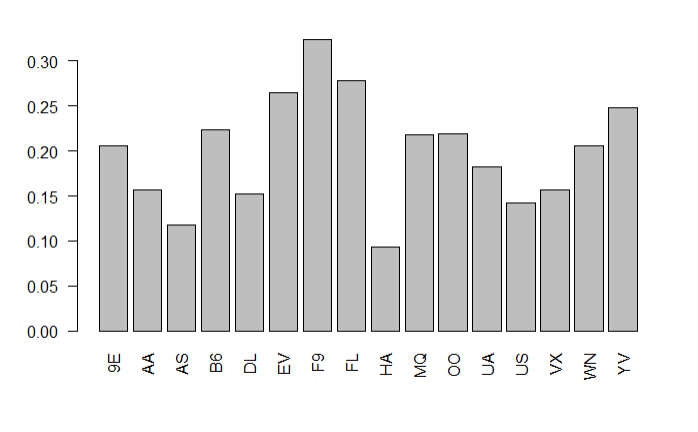
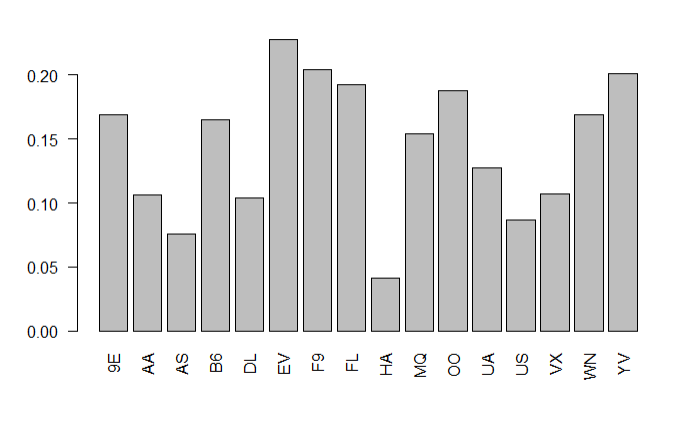


Figure 7: This graph shows, for each carrier, what percentage of their flights in the dataset arrive more than 20 minutes late.

Figure 8: This graph shows, for each carrier, what percentage of their flights in the dataset *both* departed late *and* arrived late, each by more than 20 minutes.

These three graphs are meant to illustrate that EV, F9, FL, and YV rank among the worst airlines for flight delays, but that every airline except HA has at least 10% of their flights either arriving late or departing late. Over all carriers, there is an average of 17.67% of flights that have departure delay, 19.92% of flights that have arrival delay, and 14.49% of all flights across all carriers that have delays on both ends of the trip. We can thus conclude that carrier has a demonstrable effect on delays, though further analysis would be needed to determine how much of an affect the carriers have on the delay itself.

The following two graphs are each a 3D histogram, plotting carrier on the y-axis, departure delay time (for Figure 7), and arrival delay time (for Figure 8) in minutes on the x-axis, and frequency of departure delay time on the z-axis. This graph demonstrates just how much each carrier actually delays their flights, with US, UA, OO, and DL being some of the most egregious carriers for long delays, with the large tails you see going out in the plot. DL has a large number of short delays (21-50 minutes or so), whereas other carriers like VX, MQ, HA, and FL barely show up on the chart. We know from the previous chart that they all have around 10% of flights as delayed, but this can be attributed to both lower volume of flights and fewer delays.

Interestingly, US, UA, and OO all have significantly more short term (21-100) arrival flight delays than they did for departure delays. The overall shape of the tails however, as delay time increases, is about the same between the arrival delay 3D histogram and the departure delay 3D histogram.

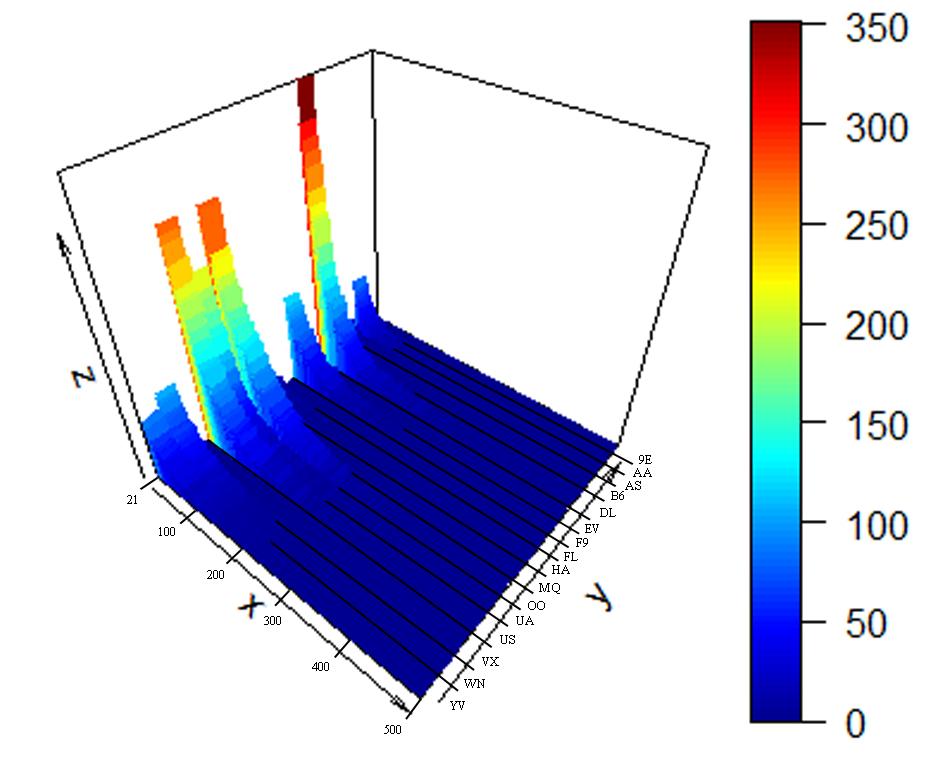
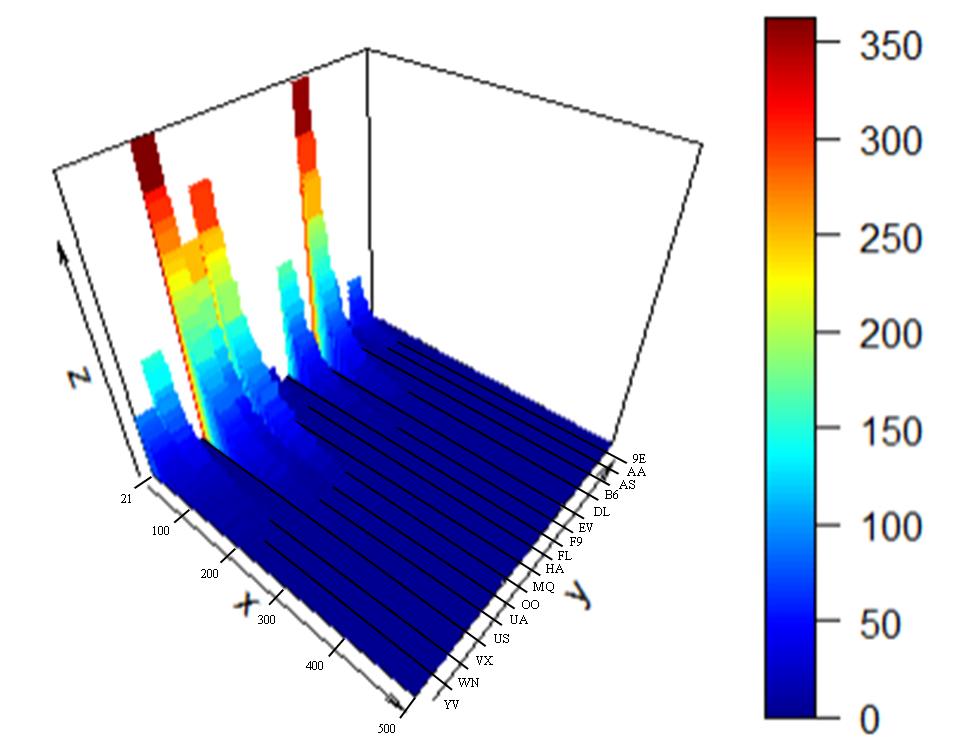
Figure 7: 3D Histogram of Departure Delay Frequency by Carrier (in Minutes)

Figure 8: 3D Histogram of Arrival Delay Frequency by Carrier (in Minutes)

**Conclusion**

Based on the descriptive analysis that was presented here, we found that major delays are happened during the month of June/July and December.We can see that there is a strong correlation between departure delay and arrival delay. Over all carriers, there is an average of 17.67% of flights that have departure delay, 19.92% of flights that have arrival delay, and 14.49% of all flights across all carriers that have delays on both ends of the trip. We can thus conclude that departure delay is one of the positive influencing factors on arrival delay. Our prediction accuracy could potentially improve if we include other strong influencing factors such as weather conditions and air traffic.

Further, we are able to note that there is variation between delays greater than 20 minutes and the airline providing the service. With strong correlation between the carrier and delay, we can confidently say that the airline can have a major impact on the number of flights and extremes of the delays of the individual flights. Of these, Delta Airlines stands out as particularly likely to have lengthy delays and frequent delays. Further examination into the destinations of these flights is expected to revel problematic routes, which may explain some airlines delays.

As flying is a an expensive and costly process, fully understanding the inefficiencies caused by flight delays is critical to understanding this industry. The analysis studied here suggests a strong sense of correlation between various aspects impacting delays. These include carriers and month of the year. The reasons for this are multi-dimensional and requires further datasets to study the reasons behind this. Fully understanding the nature an airline delay is in the vested financial interest of passengers and airlines. The exploratory data analysis described in this paper presents several trends for future exploration and research into the airline industry.

**Areas for Further Research**

Though the exploratory data analysis presented here does allow us to answer the SMART questions examined, it does not provide us with a complete picture. To further understand flight delays and the overall travel experience for anyone taking domestic and international air travel, we will need to incorporate two datasets. The first will include weather, which can support the correlation that inclement weather is the cause of delays in December and January. The second will include details on the individual flights and causes of delays and customer reviews of each airlines. This second dataset will support hypotheses about the carriers and the experience they provide their customers. By incorporating these data sources, we will be able to have a complete view of what causes flight delays and when those delays are most costly.

**References**

1. Airlines for America Homepage. (2018). Retrieved from Airlines for America: <http://airlines.org/data/>
2. Martinez, V. (2012). Flight Delay Prediction. ETHzurich.
3. Michael Ball, C. B. (2010). Total Delay Impact Study. Nextor.
4. Peterson, E. B., Neels, K., Barczi, N., & Graham, T. (2013). The Economic Cost of Airline Flight Delay. Journal of Transport Economics and Policy, 107-121.
5. Reed, D. (2016, 4 14). Americans Love To Complain About Flying, But Probably Less Than You Think. Retrieved from Forbes: https://www.forbes.com/sites/danielreed/2016/04/14/americans-love-to-fly-they-also-complain-about-it-a-lot-but-probably-less-than-you-think/#3faf22196423