NYC ⇒ Miami - Detecting Patterns in Flight Delays

Azam Bin Ahmad Bakir, Robert Kwasny, Emily Mitaro, Frank Portman

April 25, 2013

1 Introduction

It's rare to find someone who reacts positively, or even ambivalently, to a flight delay. Typically, people react with considerable frustration, making delays notable and likely to stand out when reflecting on flying experiences. People then like to look for reasons to explain why their flights are delayed, which has led society to create many "myths" about when flights are most likely to be delayed and for how long. The Wednesday before Thanksgiving is rumored to be the worst day of the year for flying, and delays and airport congestion are said to be more common before and after every major holiday. But it is a proven psychological phenomenon that people remember more negative experiences than positive experiences, making it likely that we will remember more flight delays than on-time flights. People's memories and opinions are very biased sources for determining when a flight is going to be delayed and for how long, yet society tends to trust the "myths" they have created. In this report, we test these "myths" of flight delays and determine whether or not they hold any merit.

The data set used in this report was downloaded from stat-computing.org, which compiled the data from the Research and Innovative Technology Administration (RITA) in the Bureau of Transportation Statistics (BTS). The BTS tracks the on-time performance of domestic flights operated by large carriers and has released public records dating fom 1987. Because these public records contain such a massive number of flight records, we built a sequel database consisting only of flights from the three major New York City airports - LaGuardia, John F. Kennedy, and Newark Airports. However, even this data consisted of over 6,000,000 data points, so we decided to focus on flights from New York to Miami to make the data more managable. New York to Miami is one of the most popular flight routes in the country, making it a relevant topic for an analysis of flight performance.

Specifically, we conduct a temporal analysis of several of the given flight variables, including day of the week, hour of departure, month of the year, and year itself; test the "airline hub theory" of departure delays; and develop a model to predict whether a given flight will be delayed en route from New York to Miami. In order to test these variables, we did some considerable subsetting of the data. Our subset includes flights departing from LGA, JFK, or EWR, and arriving at Miami International Airport (MIA). American Airlines and Continental fly to Miami considerably more often than any of the other airlines, with American totaling 115,436 flights and Continental totaling 42,778 flights in the Miami subset. Because the next most frequent airline was Trans World Airlines with 14,637 airlines, we decided to stick with American Airline and Continental for the temporal analysis and the testing of the "hub theory", though the predictive model does not account for this in the data set. Furthermore, we removed data points with a departure time or air time of "0" that do not make sense in the analysis. Ultimately, we find mixed evidence to back these "myths". Flights appear to be most delayed on average in the summer months and December, as well as in the evening and towards the weekend, but these averages differ by a total of under ten minutes. While the number of flights has not increased over the years since 1987, the length of delay has increased with time. Furthermore, there is no evidence that flights are more delayed when flying from a hub of the airline.

We have a strong suspicion of human error in this data set that could contribute to errors in our analysis. Throughout our exploration of the data, we noticed that the entries seem to be clustered around intervals of five and ten, which is very characteristic of human data entry. Additionally, inconsistincies in the dataset, such as the use of both "0" and "2400" as flight times or the entry of "0" for air time without entering the flight as cancelled, are also likely due to human error and provide suspicion that rest of the data set may also contain errors. These suspicions must be taken into consideration when reading this analysis.

2 Temporal Analysis

In this section we highlight and analyze several the patterns in on-time performance with respect to five time variables: year, month, day of week, and departure time.

2.1 Year

Year	Dep Delay	Freq(NYC-MIA)	Freq(Total)	Arr Delay	Elapsed Tim	CRS Elapsed
1995	6.98	8811	255004	6.43	176.71	177.85
1996	10.85	9102	259514	13.07	181.33	179.10
1997	8.65	8684	257936	9.02	182.38	182.01
1998	10.30	8323	257251	7.58	182.35	185.06
1999	14.83	8069	269181	13.69	183.11	184.26
2000	16.23	7996	274251	14.34	184.95	186.84
2001	13.39	7656	298449	8.67	182.73	187.45
2002	8.40	7682	246398	3.60	180.58	185.37
2003	9.48	7729	331649	9.13	183.28	183.63
2004	11.63	8090	377164	9.18	182.42	184.87
2005	12.53	8204	385118	12.24	183.60	183.91
2006	13.98	8177	396117	12.54	183.90	185.34
2007	20.54	7834	403378	19.92	186.66	187.28
2008	19.41	8137	376445	21.38	189.91	187.94

Table 1

As Table 1 shows, the average departure delay for flights between New York City and Miami has increased from 7 minutes in 1995 to almost 20 minutes in 2008. However, the growth in the average delay has not been linear. The delay time was at a local peak of 16 minutes in 2000, only to fall to below 10 minutes in 2001 and 2002.

A plausible cause of the decrease in the early 2000s is the large decrease in the number of flights. Even though there was little change in the frequency of flights between NYC and Miami, the three New York airports were the origin for over 50,000 fewer flights in 2002 than in 2001. The 9/11 terrorist attacks, and the reevaluation of the security measures they caused, are most obvious explanation. From 2002 to 2003 the number of flights originating in the three New York City airports grows by staggering 85,000. However, by this time the flight traffic was working smoothly and, as the table above shows, the increase in delays was negligible.

Figure 1 shows the ratio of the average arrival delay to the average departure delay. If the ratio is below 1, it means that the plane 'made up' some of the lost time during one, or all, of the three intermediate stages: taxi out, the flight itself, and taxi in.

However, according to table 1 and the left-hand plot of figure 2, no such thing takes place. Here the average time elapsed is the sum of the taxi out, taxi in, and air time. The CRS time elapsed is the predicted

total time of these three stages. The difference between the two is negligible during most years. Thus, the only remaining explanation of the trends in the figure 1 is that airlines intentionally overestimate the arrival time.

Figure 1: The ratio of the average arrival delay to the average departure delay for (red) flight between NYC and MIA and (blue) all flights departing from NYC

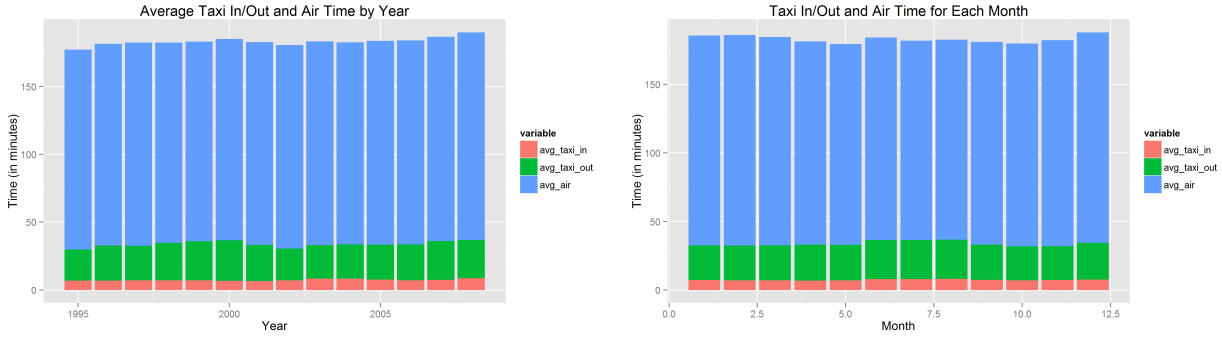

Figure 2: (Left) The average taxi in, taxi out, and air time of flight between NYC and MIA. (Right) The average taxi in, taxi out, and air time of flight between NYC and MIA for each month

2.2 Months

Table 2 shows that there is a good deal of variation in the arrival delays between months. The shortest delays are in October and November (roughly 6 minutes), while the longest are in June, July, August, and December (between 16 and 18 minutes). The difference is not explained by the total number of flights, as this does not change significantly throughout the data. The month with the most flights to Miami, March, averages 11-minute delays. However, when examining flights from New York airports to all destinations, we find that October has more flights than the other months, yet this is one of the best months to fly to Miami.

As the right hand plot of figure 2 shows, air time and taxi in time are fairly constant throughout the year. However, during the months with the longest arrival delay (June, July, August, and December) taxi out takes discenibly more than during other months.

Month	Dep Delay	Arr Delay	Ratio	Freq(NYC-MIA)	Freq(Total)
1	13.68	10.88	0.80	9703	524977
2	12.29	11.47	0.93	8897	484152
3	12.65	10.73	0.85	10023	538111
4	10.79	9.64	0.89	9669	519616
5	9.24	7.06	0.76	9743	527333
6	15.41	17.62	1.14	9348	521879
7	16.49	16.08	0.98	9987	541319
8	16.31	16.28	1.00	9917	547268
9	8.48	7.70	0.91	8695	508481
10	8.53	6.26	0.73	9175	550814
11	9.49	6.01	0.63	9484	526018
12	16.74	17.13	1.02	9853	541301

Table 2

We cannot deduce the reason for this pattern using this dataset. However, we speculate that, since June-August and December are times when many people take vacation, outside variables may lead to flight delays. For example, the planes may be always completely filled, more families with young children may travel, and inexperienced and infrequent travellers are likely to constitute a higher proportion of passangers than usual. Each of these variables could contribute to these changes in the data, but without information detailing these variables, we cannot draw conclusions.

Moreover, the ratio of the average arrival delay to the average departure delay is at or below 1 for all months except for June and December. It is consistent with our findings in the subsection 2.1, and it further shows that departure delay overstates the length of the arrival delay. This finding implies that, since departure delay and arrival delay are usually not equal, prospective passengers should pay more attention to the latter.

To a certain extent the patterns we found conform with expectations. Miami is a popular tourist destination with an excellent climate throughout the year. Thus, increased delays ought to be expected on fully-booked flights during summer months as well as during Holiday season. We speculate that the March is the month with the highest number of flights to Miami can be explained by the city's popularity as a spring break destination. However, we were surprised to discover that November, despite the surge of travel caused by the Thanksgiving, is among the best times of year to make the trip, indicating that the infamous "Wednesday before Thanksgiving" myth may not actually hold weight.

2.3 Weekly and Daily Flight Frequency

By analyzing a weekly profile of this flight route, we can uncover patterns of flight availabilities throughout the week and their distributions across time. Furthermore, weekly and daily trends of flight delays can be observed to allow for analysis on the relationship of delay time with flight frequency.

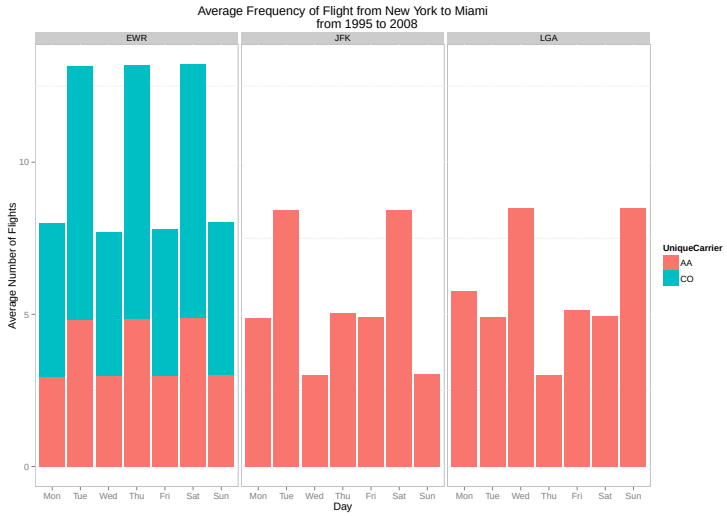

Figure 3: Daily flight frequency between New York and Miami faceted across major departure airports and major airlines.

Figure 3 shows the average daily frequency of flight from each major New York airport offered by major airlines operating on this particular route. Overall, more flights fly weekly from Newark on average than LaGuardia and JFK as both American Airlines and Continental fly from Newark while only American Airlines flies from the other two airports. The frequency of departing flights from Newark is highest on Tuesday, Thursday and Saturday with about 12 flights per day while the other days' frequencies average about 6 flight per day. JFK's flight frequency was highest on Tuesday and Saturday with approximately 6 flights per day, while LaGuardia's flight frequency is highest on Wednesday and Sunday with about 6 flights per day. It can be seen that AA flies this route most often as it flies from all three airports, while Continental only operates from Newark. However, the average frequency of AA flights was lowest at Newark compared to the other airports where it is operating, likely because JFK is a hub for AA and LaGuardia is a focus city.

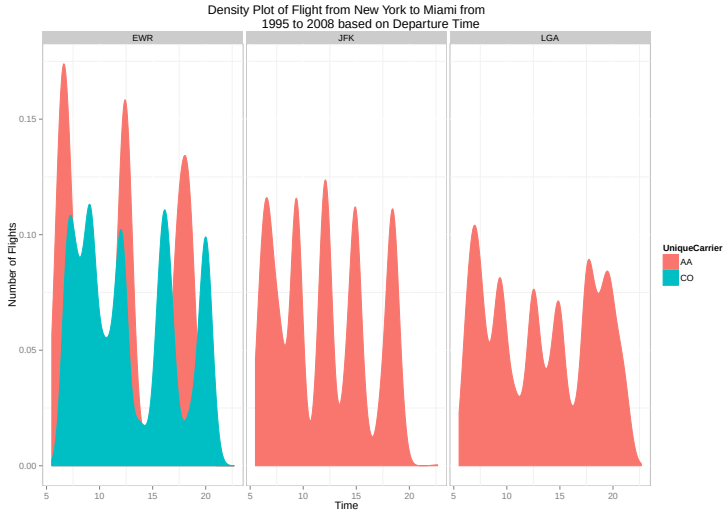

Figure 4: Flight density between New York and Miami faceted across major departure airports and major airlines at particular time of a day.

Flights occur at the highest density in the early morning for all the three airports. This may be due to the demand from business travelers, which require early arrival at their destinations. Also, it can be seen that AA operated flights from Newark mostly in the early morning with decreasing density throughout the afternoon and evening. The density of flights for Continental at Newark and AA at JFK seemed to be fairly constant at their departing times. The density of flights at LaGuardia was highest in the morning and evening and dipped in the afternoon. This trend may be caused by the demand of local New Yorkers business travellers, who are departing to Miami in the morning, and business travelers that are returning back to Miami at night. Furthermore, the smaller distance between LaGuardia Manhattan makes it attractive for business travelers.

2.4 Weekly and Daily Trend of Arrival and Departure Delay

We classify delay intervals of less than 5 minutes as on-time, 5 to 30 minutes as minor delay, 30 minutes to 2 hours as moderate delay, and more than two hours as major delay. In general, about 30-50% of flights experienced arrival delays daily. Most of delays are classified by our terms as minor delays, with moderate delays and major delays contributing in lower frequencies. Less than 5% of flights every day experience major arrival delays. The most arrival delays were seen on Thursdays and Fridays at all airports and by both airlines, while weekends showed lowest percentage of delay. There were also negligible differences between the two airlines, though Continental had a slightly higher percentage of delay than AA on Thursday and Friday. AA flights from Newark were most often on-time compared to AA flights from JFK and LaGuardia and Continental flights from Newark. This trend may be caused by the lower frequency of flights by AA at Newark in comparison to Continental and the other airports.

The daily trend showed that arrival delay mostly increased from morning to evening. This may be due to the accumulation of delay caused by other previous flights. AA arrival delay at Newark was highest in the early evening around 3:00pm, then drops for flights later in the evening. LaGuardia showed the smallest duration of delay, which may have been due to its smaller operation compared to the other two major international airports. Also, there seems to be correlation between flight density and time of day. Comparing figure 6 with flight density plot, it can be seen that the slight bump of increase in arrival delay can also be noticed in the figure at the time of high flight density. This may be related to more aerospace traffic, leading to delays in attempt to space out aircraft landings. Trends in departure delays were similar to trends in arrival delays. This is expected as a departure delay will likely cause arrival delay.

By looking at weekly and daily trends in flight delays, we see that arrival and departure delays are influenced by flight frequency at particular days and times. It was also found that arrival and departure delay were highest approaching the weekend. In addition to that, passengers are also more likely to experience a flight delay if they are taking the evening flight compared to the morning flight.

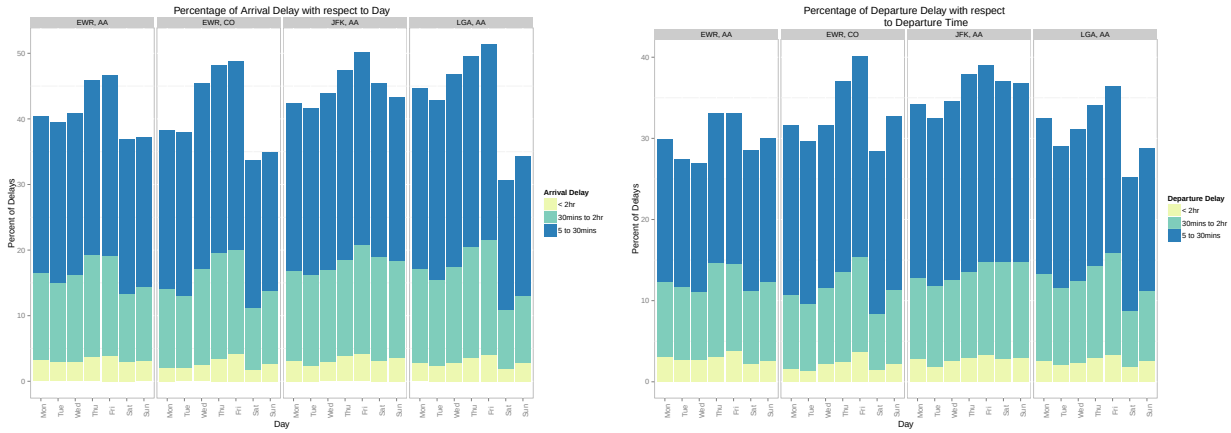

Figure 5: (Left) Daily average arrival delay faceted to airports and airlines. (Right) Daily average departure delay faceted to airports and airlines

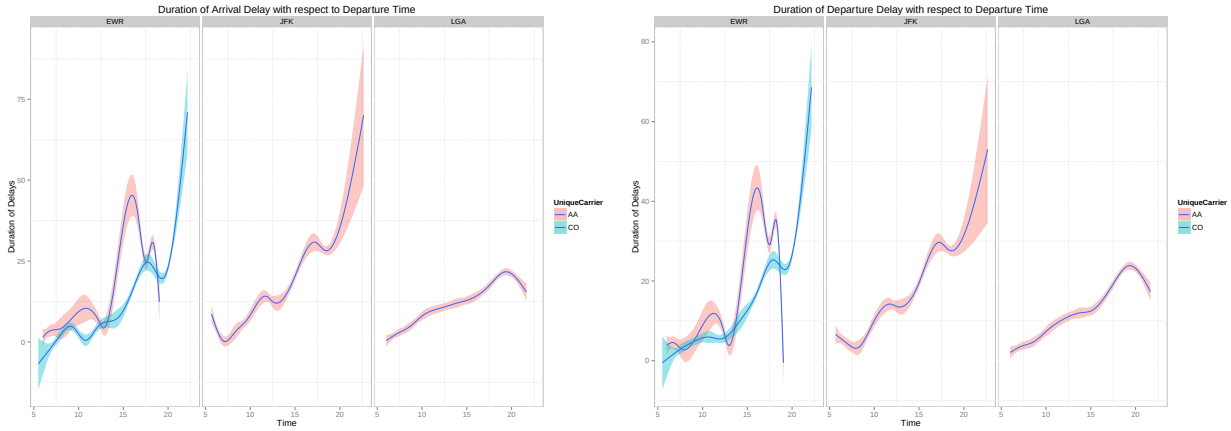

Figure 6: (Left) Duration of arrival delay over departure time faceted to airports and airlines. (Right) Duration of departure delay over departure time faceted to airports and airlines.

3 Airline Hubs - Does It Make a Difference?

3.1 Overview

We decided to focus this section on comparing three flight routes: American Airlines from JFK (hub), American Airlines from Newark (not a hub) and Continental from Newark (not a hub). Continental does not have any hubs in the New York area airports, and only flies to Miami from Newark. American Airlines, on the other hand, has a hub at JFK, a focus city at Laguardia, and flies from Newark as well. Since Laguardia is only relevant to American and is not a hub, we decided to exclude it from this section of the analysis. This subset allows us to compare the difference between hubs and non-hubs both within and between airlines. However, there are several variables that are not taken into account in doing so. We do not have information on airlines fares, passenger volume, or flight popularity and fullness, all of which could have significant impact on flight delays. Unless the planes are roughly the same size and price, these variables cannot be ignored and could be contributing factors for flight delays. Also, it significantly hinders the analysis that we

do not have information for a hub at Newark, a hub for Continental, and a non-hub at JFK. Results could be attributed to differences between Newark and JFK, or to differences between American and Continental, rather than a hub and a non-hub. Additionally, air congestion could vary between airports, as could baggage loading processes and flight crew policies, which could all lead to differences between airports but cannot be controlled for in this analysis.

3.2 The Data

	Hub	Origin	Carrier	Average Delay (mins)	Number of Flights
1	No	Newark	American Airlines	11.66	21,320
2	No	Newark	Continental	11.62	36,375
3	Yes	JFK	American Airlines	12.10	33,375

Table 3: Average Departure Delays of Hubs and Non-Hubs

Table 3 provides an overview of the data used in this section. The averages in this data set are calculated with given departure delays beginning in 1990, which is the earliest year at which all three routes began flying to Miami. The differences in average flight delays are negligible, averaging less than one minute, among all three flight routes. This indicates that there may not be significant differences in delay patterns even if an airline has a hub at an airport. Figure 7 shows a more detailed representation of average delays by each of the three routes.

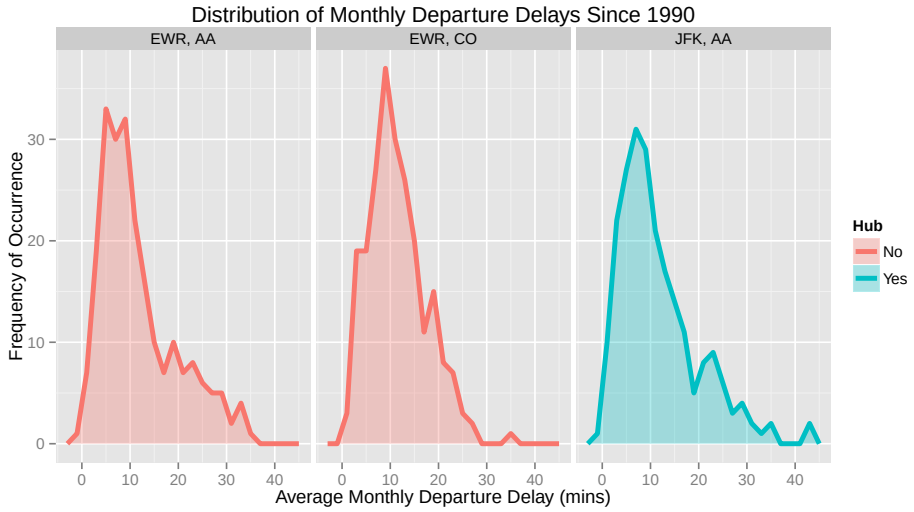

Figure 7: Each of the three routes shows very similar positively skewed distributions, with the majority departures averaging a delay of about 12 minutes. Continental seems slightly more concentrated to the left than both American Airlines routes, but it does not appear to be significant.

However, it is still possible that, when analyzed in terms of other variables, differences might arise. For example, it seemed likely that differences in length of delay could exist among these routes. It seemed possible that the proportion of exceptionally long delays would be roughly equivalent at hubs and non-hubs, but shorter delays could be more frequent at hubs due to the large number of flights coming and going each

day, leading to congestion and issues relating to baggage loading or crew. We decided to examine delay size further to see if these differences might arise.

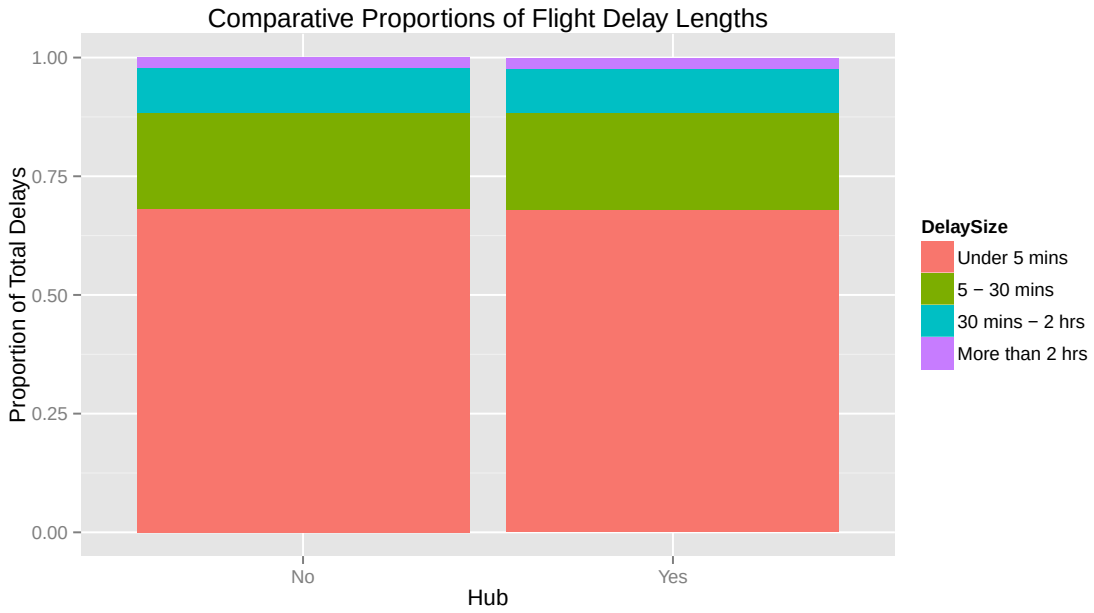

Figure 8

To create figure 8, we created a new variable "DelaySize" to group the individual departure delay times in a more meaningful way. We determined a delay of under five minutes to be essentially on-time for all practical purposes. We grouped the next level to be five to thirty minutes, as delays in this interval could be deemed as significant but not disruptive. A delay of thirty minutes to two hours would likely be considered disruptive, and over two hours is indisputably disruptive. We chose these intervals to reflect how a consumer would group intervals of flight delays to make the analysis as relevant to the consumer as possible. However, it is possible that differences would appear if broken down further. For this graph, we compared delays based on whether or not they occurred at an airport with a hub for the given airline. It should be noted that this graph does not account for the differences in airline, therefore lumping the data from both airlines at Newark into one. Also, the "Yes" hubs and "No" hub categories are from different airports, which needs to be taken into account. It's possible that other outside variables contributed to the pattern seen in the visual because the airport was not controlled for.

Very clearly, there are no visible differences in the number of flights falling into these intervals, regardless of whether the airport was a hub for the airline or not. Every interval is represented in the same proportions between the "Yes" hub and "No" hub categories. This indicates that the proportion of flights that are on time, slightly delayed, considerably delayed, and very significantly delayed from New York to Miami is the same whether the airport is or is not a hub for that airline. If such a conclusion were true, a passenger traveling from New York to Miami would be just as likely to experience a given flight delay at a hub for an airline as an airport that is not a hub for an airline. However, this does not mean that differences do not exist. The flight patterns in 1990 that are contributing to this visualization might be significantly different from the patterns in 2008, and flight patterns from the 90s arent relevant to someone flying in 2013. We therefore decided to examine this relationship further, also taking year into consideration.

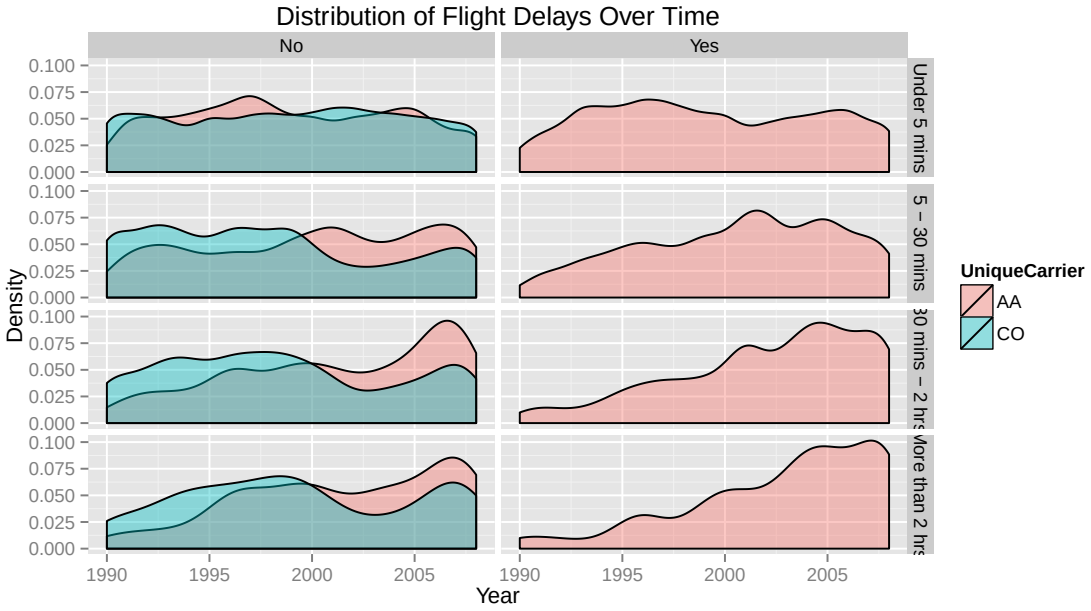

Figure 9

Figure 9 provides a breakdown of the distribution flight delays by year, and compares the data based on whether the flight departed from an airport where the airline has a hub. The graph also allows us to compare the differences between American Airlines and Continental at Newark alone, where neither airline has a hub. This makes it easier to understand which differences can be attributed to the fact that the airport is different and not a hub, and which differences can be attributed to the differences in airline. The distributions were created using a density curve that shows the patterns of delay over time. The data was then facet wrapped, colored, and made transparent to improve the effectiveness of the visualization.

At JFK, where American Airlines has a hub, we see very different distributions among the delay intervals over time. The proportion of very delayed flights (more than thirty minutes) is steadily increasing with each passing year, displaying a negatively skewed distribution concentrated around later years. However, flights that are not very delayed appear to be evenly distributed over the years. This indicates that while the proportion of on-time or short-delayed flights is not changing, the number of very delayed American Airlines flights to Miami from JFK is growing with each passing year. If this is true, then it is more likely that a flight would be largely delayed in lthe last 2000s than it was in the early 1990s.

The pattern is slightly different at Newark, where both American Airlines and Continental are displayed and neither has a hub. On American, we do see a similar upward trend in proportion of very delayed flights, though the rise is less steady and less extreme. Still, this indicates that the increase could be attributed to the carrier rather than the airport as a hub, because it is not seen on Continental. For Continental flights, although the on-time flights are evenly distributed over the years of this data set, the other delay intervals show significant declines in the distributions in the early 2000s, indicating that flights were actually less delayed than they had been in previous years. The drop specifically happens around 2002, which could be related to the September 11th attacks. The distributions then show increases in proportions again in the most recent years. It is interesting that Continental would decline around 2002 while American Airlines continues its increase, as the hijacked planes were American Airlines planes. Further analysis could explore why this might be the case.

Clearly, there are differences in these distributions and we see that the delay patterns are actually not

identical. However, it seems likely that these differences could be due to Airline carriers rather than the airport serving as a hub. Still, even for just American Airlines, the rise is most drastic when the airport has a hub at JFK, where the airline has a hub, than at Newark, where is does not. Figure 10 shows more clearly the average delay patterns over time, with AA increasing at both airports and Continental fluctuating.

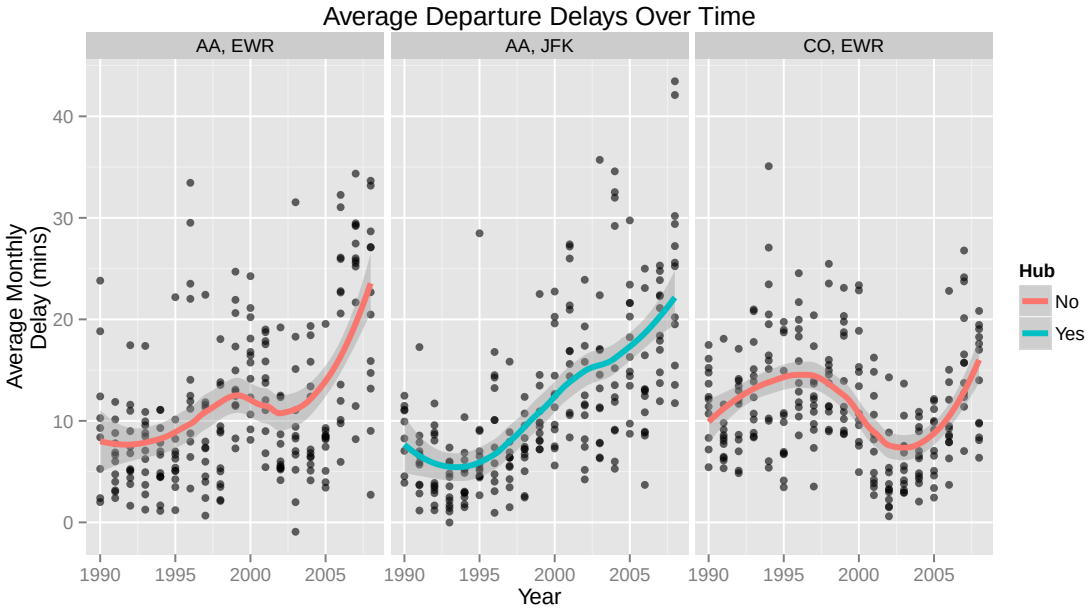

Figure 10: The most consistent increase is seen where American Airlines has a hub at JFK. Average delays for each month are plotted at each year and the smooth curve shows the general distribution. The AA curve at Newark resembles the AA curve at JFK more than the CO curve at Newark, indicating that the differences in delays are more likely attributed to carrier than hub or airport.

For all practical purposes, there do not seem to be significant differences in flight delays if an airport is a hub for an airline when flying from New York to Miami. The average flight delays are roughly equivalent and are distributed similarly among the three flight routes, which indicates that, on average, a consumer will experience the same likelihood of delay and delay length at both JFK and Newark on American and Continental. There is therefore no evidence to back the common complaint that flights get delayed more when the airport is a hub for a given airline. We should be skeptical that carrier may be playing a larger role than accounted for in other areas as well. Further analyses would require more detailed information, such as passenger volume, cargo volume, air fare, and flight fullness, to account for variables that could contribute to delays. This analysis has significant room for improvement. Neither airport has Continental as a hub, and differences between the American hub and non-hub could actually be attributed to differences in airports, which was not controlled for because JFK does not have a non-hub airline that flies to Miami.

4 Predictive Modelling

4.1 Introduction

Using our dataset, we sought to make a meaningful forecast of future flight information. Considering our working subset has over 100,000 datapoints, we should be able to recognize some underlying pattern within the data and use it to make predictions. More specifically, we decided to build a model around predicting the severity of departure delay. With this in mind, we created several restrictions for which data we would use. For example, none of the individual delay variables were considered when constructing the model - if these features are known, one already has a good idea about whether a flight will be delayed or not. The driving motivation behind our model is to give an individual some idea of how long their flight will be delayed using basic information that can only be known at the time of their purchase (usually weeks to months in advance).

4.2 Data Processing & Feature Selection

Before constructing models, some amount of data processing was necessary in order to build a stronger predictor. Firstly, the 'Scheduled Departure Time' variable was in an ambiguous format. We decided to convert this variable into one that represented how late in the day the flight was as a fraction of 1. In this way, we avoid oversaturating the weights of our models by inputting large hours. Also, we converted 'Year', 'Origin Airport', and 'Airline' into factors. These are known well in advance, so a robust prediction based off of these translates into highly valuable information for the individual. Finally, our working subset for the exploratory analysis was randomly split into an 80% training set and a 20% testing set.

We took two basic approaches in constructing our predictive models - regression and classification. For the regression technique, we were interested in predicting the actual departure delay in minutes. The classification approach was inspired by the results of the regression model which will be discussed later in this section. For that model, we split our departure delays into five different factors as follows in Table 4:

Table 4: Our Classifications

Delay	Classification
Less than -5 Minutes	Early
Between -5 and 5 Minutes	On Time
Between 5 and 30 Minutes	Late
More than 30 Minutes	Very Late

In this way we feel that we fully encapsulate all the possible delay scenarios fairly well. The 'On Time' grouping takes into consideration human error in recording results because it is very rare for a flight to be EXACTLY (precisely zero seconds of departure delay) on time. There is much evidence to suggest that the data is recorded by hand due to the structured intervals represented in some variables, so we allow a small buffer for that. In addition, knowing that your flight will be 'delayed' by three minutes isn't crucial information, so we feel our range is fair.

To select our features, we first glanced at the results of the exploratory analysis conducted in the beginning of the paper and balanced those results with information that an individual might have upon purchasing a flight ticket. Our primary analysis suggests differences and patterns between several of the variables. After deciding on the inputs as: Year, Day of Month, Month, Day of Week, Time (as proportion of full day), and Airline, we leave it to the sophisticated machine learning techniques to translate our findings into predictive power.

4.3 Models & Results

For several reasons, the first technique we turned to was a Random Forest implementation. Random Forests are fast, require almost no tuning, automatically rank importance of variables, and are immune to collinearity. The speed of the machine learning approach we took turned out to be a very important factor in deciding on how to proceed. Since we have 'infinite time' (for all technical purposes) to produce a model, we should theoretically be able to resort to a method of high complexity and computational intensity. However, we don't know the exactly model we want to build and our final product is the result of rapid prototyping and tuning in R. Thus, selecting a fast, yet powerful, algorithm is crucial since the dataset is quite large.

The regression results from the Random Forest left much to be desired, with a RMSE (Root Mean Squared Error) of 36.54 on the training set, and 42.02 on the test set. However, in this case it proved fruitful to actually go to the data with a visual approach. We noticed that while some of our predicted departure delays were far off in a minute sense, the 'magnitude' of the delay was preserved by our prediction. To preserve this effect, we reformulated our problem as a classification challenge.

Table 5: (Left) Train Confusion Matrix (Right) Test Confusion Matrix

	Early	Late	On Time	Very Late
Early	417	112	5997	144
Late	92	1785	14780	1444
On Time	449	1822	60102	1375
Very Late	79	1216	8755	2602

	Early	Late	On Time	Very Late
Early	690	10	18	15
Late	18	1995	102	94
On Time	1030	2192	15811	1464
Very Late	26	120	93	1615

Interestingly enough, our model seemed to perform better on the test set rather than the training one, with accuracies (predicting the exact category) of 79.51% and 64.15% respectively. Table 5 shows us the confusion matrix from the results of our Random Forest prediction on the two separate training sets. The horizontal table axis represents the predicted class, while the vertical one represents the actual classes. Confusion matrices are useful in that they may tell us how 'far off' our predictions are when dealing with a classification problem. In this case, if most of our wrong predictions are in neighbouring time slots, that would mean our model isn't extremely wrong even when it fails to predict the correct class. The confusion matrix of the training set depicts that our model is rather pessimistic (predicting longer delays). More specifically, it predicts many on time flights as late or very late. On the other hand, the test set displayed very good results. A huge portion of the predicted results were in the correct category, or in a neighbouring one. However, when wrong, the model leaned towards the sign of optimism (predicting shorter delays).

4.4 Conclusion & Improvements

Our Random Forest classification has fairly good accuracy for predicting departure delays of flights from NYC airports to Miami. Of course, one should be mindful of using this model for several reasons. For one, the model relies on very basic features and principles in giving its prediction. If the model states that a flight should be leaving 'Early', but it is blizzarding on the day of your flight, there's a good chance that it will be wrong. Therefore, considering the fact that many delays are a result of weather, security delays, and other unpredictable events, we feel that our model does a good job at weaving through those special cases to present an individual with an accurate benchmark prediction.

In addition, there are a number of things that can be improved on to possibly yield a stronger model:

1. Hardware and R-language limitations hindered our ability to test more computationally intensive models that may have performed better.

2. Using 'Year' as a predictor is questionable. For it to be an important predictor, one would have to know

a significant number of observations in that year in order to train the model properly. Therefore, the model isn't static and would have to be re-generated.

3. There are an endless number of ways to process and bin the data. Experimenting with several of these, especially for the time-related variables (such as Month) could improve the model.

5 Conclusion

Ultimately, our report shows that there are certain temporal factors that contribute to flight delays, however many of the common societal "myths" may not hold weight. In the temporal analysis section, we showed that the average departure delay has been steadily increasing; it went from 7 minutes in 1995 to 20 in 2008. We also showed that the length of the delay is only loosely related to either the number of flights from New York to Miami or the total number of all outgoing flights from New York. Further, we argued that the length of the departure delay is usually greater than the length of the arrival delay. The weekly trends show that flight arrival and departure delays are highest as the weekend approaches. Duration of flight delay was also higher in the evening flights compared to the morning flights due to the propagation of delay acquired throughout the day. The myth regarding airport hubs does not seem to hold weight. Though we do see increases in delays over time from JFK moreso than Newark, this cannot necessarily be attributed to the fact that JFK is a hub for American and Newark is not, and carrier may actually be a more likely explanation. Finally, using our dataset we constructed a Random Forest in order to predict the severity of departure delay from several variables in the dataset. Despite weather, security, and other unpredictable events that contribute significantly to delays, our model produces a relatively accurate prediction of a given flight delay.

Further analyses could easily build upon this report to produce a more broad representation of flight delay patterns. Given the limitations of our hardware, we could not conduct data analysis on additional flights, such as Miami to JFK, LaGuardia, and Newark, that may show interesting and relevant patterns. However, this would be another interesting topic for comparison, as would the addition of other flight routes or even international flight information. With data on passenger volume, airfare, cargo volume, and other variables, this analysis would also be more complete and conclusive. Still, we believe this report to be a good starting point for an analysis of flight delay patterns.

A Code

```
#######################
##### Stat 405    #####
##### Project 3   #####
##### Final Code  #####
#######################
library(RSQLite)
library(lubridate)
library(ggplot2)
library(plyr)
library(randomForest)
library(caret)
library(xtable)
library(stringr)
library(reshape2)




###
## Data Cleaning and Extracting
###

## Connect with our local SQL Database
ontime <- dbConnect("SQLite", dbname = "ontime.sqlite3")

## Write a function to pass queries to it
from_db <- function(sql) {
  dbGetQuery(ontime, sql)
}

## Extract the airports separately
jfk <- from_db("select * from ontime where Origin = 'JFK'")
ewr <- from_db("select * from ontime where Origin = 'EWR'")
lga <- from_db("select * from ontime where Origin = 'LGA'")

## Make full NYC Dataset
nyc <- rbind(rbind(jfk, ewr), lga)

## Save/Load easily because this subset is huge
save(nyc, file = "nyc.RData")
load("nyc.RData")

## We only care about flights to Miami
nyc.mia <- subset(nyc, Dest == "MIA")
```

```
## Clean the dataset to remove some NA's and flights that have 0 duration
nyc.mia.clean <- subset(nyc.mia, AirTime != 0)
nyc.mia.clean <- subset(nyc.mia.clean, DepDelay != "NA")
nyc.mia.clean <- subset(nyc.mia.clean, AirTime != "NA")




###
## Temporal Analysis
###

## Include only AA and CO airlines
nyc_mia2 <- subset(nyc.mia.clean, UniqueCarrier %in% c("AA", "CO"))

## Create summaries of different variables by year
mia_table <- ddply(nyc_mia2, "Year", summarise,
                   Avg_Dep_Delay = mean(DepDelay),
                   Avg_Arr_Delay = mean(ArrDelay),
                   Avg_Taxi_in = mean(TaxiIn),
                   Avg_Taxi_Out = mean(TaxiOut),
                   Avg_AirTime = mean(AirTime),
                   Ratio = Avg_Arr_Delay/Avg_Dep_Delay)

## Same, except for the whole NYC dataset, to compare
whole_table <- ddply(nyc, "Year", summarise,
                     Avg_Dep_Delay = mean(DepDelay),
                     Avg_Arr_Delay = mean(ArrDelay),
                     Avg_Taxi_in = mean(TaxiIn),
                     Avg_Taxi_Out = mean(TaxiOut),
                     Avg_AirTime = mean(AirTime),
                     Ratio = Avg_Arr_Delay/Avg_Dep_Delay)

## Extract from our dataframe
Ratio_mia <- mia_table[, 7]
Ratio_whole <- whole_table[, 7]
Ratio_whole <- Ratio_whole[9:22]

## Vector of years
Year <- c(1995:2008)

## Create a dataframe for our ratios (only MIA and whole)
Plot1 <- data.frame(Year, Ratio_mia, Ratio_whole)

## Melt by year
Plot1a <- melt(Plot1, id = c("Year"))
```

```
## Compare Delays across years
qplot(Year, value, data = Plot1a, color = variable, geom = "line",
      ylab = "Arrival Delay/Departure Delay",
      main = "Relationship between departure and arrival delay, 1995-2008")

ggsave(filename = "plot_1.png")

## Extract mean Taxi In, Taxi Out, Flight Time, by year
y <- ddply(nyc_mia2, "Year", summarise,
           avg_taxi_in = mean(TaxiIn),
           avg_taxi_out = mean(TaxiOut),
           avg_air = mean(AirTime))

## Melt these all so Year is a variable
my <- melt(y, id = c("Year"))

## Look at trend of Taxi Time and Air Time by year
qplot(Year, value, fill = variable, data = my, ylab = "Time (in minutes)",
      geom = "bar", stat = "identity",
      main = "Average Taxi In/Out and Air Time by Year")

ggsave(filename = "Plot_2.png")

## Extract average departure delay by year
mia_table2 <- ddply(nyc_mia2, "Year", summarise,
                    Avg_Dep_Delay = mean(DepDelay))

## Count number of flights from all of NYC by year
count_year_whole <- count(nyc, "Year")

## Do the same for flights to just Miami
count_year <- count(nyc_mia2, "Year")

## Combine all three of these dataframes
table <- data.frame(mia_table2, count_year, count_year_whole[9:22, ])

## Initialize
table$Year.1 <- NULL
table$Year.2 <- NULL

## Take Some averages by year into a dataframe
table2 <- ddply(nyc_mia2, "Year", summarise,
                avg_delay = mean(DepDelay),
                avg_delay_arr = mean(ArrDelay),
                avg_act_elapsed = mean(ActualElapsedTime),
                avg_crs_elapsed = mean(CRSElapsedTime))

## Join them for presentation
```

```
table3 <- join(table, table2)
xtable(table3)

## Looking at monthly averages of several variables
avg_month <- ddply(nyc_mia2, "Month", summarise,
                   avg_delay = mean(DepDelay),
                   avg_delay_arr = mean(ArrDelay),
                   ratio = avg_delay_arr/avg_delay,
                   avg_taxi_in = mean(TaxiIn),
                   avg_taxi_out = mean(TaxiOut),
                   avg_air = mean(AirTime),
                   avg_act_elapsed = mean(ActualElapsedTime))

## Count number of flights by month to Miami
count_month_mia <- count(nyc_mia2, "Month")

## Count number of flights by month from all of NYC
count_month_whole <- count(nyc, "Month")

## Create our table for presentation
table4 <- data.frame(avg_month[, 1:4], count_month_mia, count_month_whole)
table4$Month.1 <- NULL
table4$Month.2 <- NULL
xtable(table4)

## Take monthly averages of several variables
plot5 <- ddply(nyc_mia2, "Month", summarise,
               avg_taxi_in = mean(TaxiIn),
               avg_taxi_out = mean(TaxiOut),
               avg_air = mean(AirTime))

## Melt the dataframe so month is a variable
plot5a <- melt(plot5, id = c("Month"))

## Take a look at Taxi Time + Flight Time over months
qplot(Month, value, data = plot5a, stat = "identity", geom = "bar",
      fill = variable,
      ylab = "Time (in minutes)",
      main = "Taxi In/Out and Air Time for Each Month")

ggsave(filename = "Plot_5.png")

## Create slightly different subsets along the NYC to MIA Route
## We consider only AA and CO as carriers, but this time we exlude all flights
## less than 100 minutes flight time
nycmia <- subset(nyc.mia, nyc.mia$AirTime > 100)
nycmia2 <- subset(nycmia, nycmia$UniqueCarrier %in% c("AA", "CO"))
nycmia2 <- subset(nycmia2, CRSDepTime != 0)
```

```
## Convert numeric value to actual day label in DayOfWeek
levels <- 1:7
labels <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")
nycmia2$DayOfWeek <- ordered(nycmia2$DayOfWeek, levels = levels,
                             labels = labels)

## Make time a decimal for smooth plots
nycmia3 <- nycmia2
nycmia3$time <- hm(nycmia3$CRSDepTime)
nycmia3$time <- nycmia3$time@hour + nycmia3$time@minute / 60

## Construct a function to classify delays
delaydivide <- function(time){

  if(time >= 5 & time <= 30){
    counttime = "5 to 30mins"
  }
  else if(time > 30 & time <= 120){
    counttime = "30mins to 2hr"
  }
  else if(time > 120){
    counttime = "< 2hr"
  } else {
    counttime = "Ontime"
  }

  counttime
}

# Count number of days between all acquired days in dataset
firstday <- ymd("1995-01-01")
lastday <- ymd("2008-12-31")
nday <- (firstday %--% lastday) %/% days(1)
Mon <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Monday")
Tue <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Tuesday")
Wed <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Wednesday")
Thu <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Thursday")
Fri <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Friday")
Sat <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Saturday")
Sun <- sum(weekdays(seq(firstday, lastday, "days")) %in% "Sunday")

## Create dataset of number of days
day <- c(Mon, Tue, Wed, Thu, Fri, Sat, Sun)
DayOfWeek <- 1:7
day <- cbind(DayOfWeek, day)
day <- data.frame(day)
levels <- 1:7
```

```
labels <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")
day$DayOfWeek <- ordered(day$DayOfWeek, levels = levels,
                         labels = labels)

## Find number of flights per day
## Combine day dataset and dayfreq
dayfreq <- count(nycmia2, var = c("DayOfWeek", "Origin", "UniqueCarrier"),
                 wt_var = NULL)

dayfreq2 <- join(dayfreq, day, type = "right", match = "all")

## Calculate frequency of flights per day
dayfreq2 <- mutate(dayfreq2, AverageDailyFlight = freq / day)

## Plot frequency of flights per day
qplot(DayOfWeek, round(dayfreq2$AverageDailyFlight, 2), data = dayfreq2,
      fill = UniqueCarrier, geom = "bar") + facet_grid(.~Origin) +
      ylab("Average Number of Flights") + xlab("Day") +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ggtitle("Average Frequency of Flight from New York to Miami
              from 1995 to 2008")

ggsave(filename = "ave_day.pdf", width = 11.69, height = 8.27)

## Plot density of flights per time of day
qplot(time, data = nycmia3, fill = UniqueCarrier, geom = "density",
      color = UniqueCarrier) +   facet_wrap(~ Origin) +
      ylab("Number of Flights") + xlab("Time") +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ggtitle("Density Plot of Flight from New York to Miami from
              1995 to 2008 based on Departure Time")

ggsave(filename = "timedensity.pdf", width = 11.69, height = 8.27)

## Initialize an array
arrdelaydivide <- rep(NA, length(nycmia2$ArrDelay))

#classify delay according to time
for (i in 1:length(nycmia2$ArrDelay)){
  time = nycmia2$ArrDelay[i]
  arrdelaydivide[i] <- delaydivide(time)
}

## Combine new delay classification to dataset
nycmia4 <- cbind(nycmia2, arrdelaydivide)
arrdelaycount <- count(nycmia4,
                       var = c("DayOfWeek", "Origin",
                               "UniqueCarrier", "arrdelaydivide"))
```

```
## Find the percentage delayed according to our classification
totalflight <- count(nycmia2, var = c("DayOfWeek", "Origin", "UniqueCarrier"))

## Clean Column Names
names(totalflight) <- c("DayOfWeek", "Origin", "UniqueCarrier", "TotalFlight")

## Combine our data frames
arrdelaycount2 <- join(arrdelaycount, totalflight,
                       type = "right", match = "all")

## Add a percent delayed column and remove all ontime flights
arrdelaycount2 <- mutate(arrdelaycount2, percent = (freq / TotalFlight) * 100)
arrdelaycount2 <- subset(arrdelaycount2, arrdelaydivide != "Ontime")

#plot of Percent of arrival Delay per day
qplot(DayOfWeek, percent, data = arrdelaycount2,
      fill = arrdelaydivide, geom = "bar") +
      facet_wrap(Origin~  UniqueCarrier, nrow = 1, ncol = 4) +
      opts(axis.text.x = theme_text(angle = 90)) +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ylab("Percent of Delays") + xlab("Day") +
      scale_fill_brewer(name = "Arrival Delay", palette = "YlGnBu") +
      ggtitle("Percentage of Arrival Delay with respect to Day")

ggsave(filename = "daypercent.pdf", width = 11.69, height = 8.27)

## Plot actual delays per time of day
qplot(time, ArrDelay, data = nycmia3, geom = "smooth",
      fill = UniqueCarrier) + facet_wrap(~Origin) +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ylab("Duration of Delays") + xlab("Time") +
      ggtitle("Duration of Arrival Delay with respect to Departure Time")

ggsave(filename = "timearrdelay.pdf", width = 11.69, height = 8.27)

## Initialize an array
depdelaydivide <- rep(NA, length(nycmia2$DepDelay))

## Classify each delay according to duration
for (i in 1:length(nycmia2$DepDelay)){
    time = nycmia2$DepDelay[i]
    depdelaydivide[i] <- delaydivide(time)
}

## Join our classification to the dataset
nycmia5 <- cbind(nycmia2, depdelaydivide)
depdelaycount <- count(nycmia5,
```

```
                        var = c("DayOfWeek", "Origin",
                                "UniqueCarrier" , "depdelaydivide"))

## Combine our data frames
depdelaycount2 <- join(depdelaycount, totalflight, type = "right",
                        match = "all")

## Add a percent delayed column and remove all ontime flights
depdelaycount2 <- mutate(depdelaycount2, percent = (freq / TotalFlight) * 100)
depdelaycount2 <- subset(depdelaycount2, depdelaydivide != "Ontime")

## Plot the percentage of flights with departure delay by day
qplot(DayOfWeek, percent, data = depdelaycount2,
      fill = depdelaydivide, geom = "bar") +
      facet_wrap(Origin ~ UniqueCarrier, nrow = 1, ncol = 4) +
      opts(axis.text.x = theme_text(angle = 90)) +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ylab("Percent of Delays") + xlab("Day") +
      scale_fill_brewer(name = "Departure Delay", palette = "YlGnBu") +
      ggtitle("Percentage of Departure Delay with respect
              to Departure Time")

ggsave(filename = "daydepdelay.pdf", width = 11.69, height = 8.27)

## Plot the actual duration of departure delay over time of day
qplot(time, DepDelay, data = nycmia3, geom = "smooth",
      fill = UniqueCarrier) + facet_wrap(~ Origin) +
      opts(panel.background = theme_rect(fill = 'white', colour = 'grey')) +
      ylab("Duration of Delays") + xlab("Time") +
      ggtitle("Duration of Departure Delay with respect to Departure Time")

ggsave(filename = "timedepdelay.pdf", width = 11.69, height = 8.27)




###
## Airline Hubs
###

## Subset to inclue only American Airlines and Continental flights from JFK
## and Newark after 1990, removing empty rows and ensuring that there are
## always flights to Miami in the subset.
hubs <- subset(nyc.mia, (Origin == "EWR" | Origin == "JFK") &
               (UniqueCarrier == "AA" | UniqueCarrier == "CO") &
                Year >= 1990)
```

```
## Save a second copy just in case
hubs2 <- hubs

## Defines hubs to be flights from JFK (only includes AA) and not hub to be
## flights from Newark (CO and AA, neither has hub)
isHub <- hubs2$Origin == "JFK"
notHub <- hubs2$Origin == "EWR"

## Creates new variable "Hub", labels "Yes" if has a hub and "No" if does not
## have hub
hubs2$Hub[isHub] <- "Yes"
hubs2$Hub[notHub] <- "No"

## Split data by origin, airline carrier, and whether or not the airport
## is a hub. Summarise the data and calculate the mean departure delay
## for the data collapsed by these variables.
hubgeneral <- ddply(hubs2, .(Origin, UniqueCarrier, Hub), summarise,
                    avgDelay = mean(DepDelay))

xtable(hubgeneral, caption = "Average Departure Delays of Hubs and Non-Hubs")

## Create new variable DelaySize to group departure delays categorically
hubs3 <- hubs2
hubs3 <- mutate(hubs3, DelaySize = DepDelay)

## Define several conditions for delay classification
ontime <- hubs3$DepDelay < 5
minor <- hubs3$DepDelay >= 5 & hubs3$DepDelay < 30
med <- hubs3$DepDelay >= 30 & hubs3$DepDelay < 120
big <- hubs3$DepDelay >= 120

## Give labels to the new variable
hubs3$DelaySize[ontime] <- "Under 5 mins"
hubs3$DelaySize[minor] <- "5 - 30 mins"
hubs3$DelaySize[med] <- "30 mins - 2 hrs"
hubs3$DelaySize[big] <- "More than 2 hrs"

## Reorder levels by increasing delay
hubs3$DelaySize <- factor(hubs3$DelaySize, levels =
                          c("Under 5 mins", "5 - 30 mins",
                            "30 mins - 2 hrs", "More than 2 hrs"))

## Plot density curve by delay size and hub over time
ggplot(hubs3, aes(Year)) +
       geom_density(aes(fill = UniqueCarrier), alpha = 0.4) +
       facet_grid(DelaySize ~ Hub) + xlab("Year") + ylab("Density") +
       labs(title = "Distribution of Flight Delays Over Time")
```

```
ggsave("del_overtime.pdf")

## Plot comparative delay sizes by hub
ggplot(hubs3, aes(Hub)) + geom_bar(aes(fill = DelaySize), position = "fill") +
       xlab("Hub") + ylab("Proportion of Total Delays") +
       labs(title = "Comparative Proportions of Flight Delay Lengths")

ggsave("delaysize_prop.pdf")

## Split data by origin, airline, year, month, and whether or not the
## airport is a hub for the airline. Summarises the data and compute
## the average delay per month per year.
hubinfo <- ddply(hubs3, .(Origin, UniqueCarrier, Year, Month, Hub),
                 summarise, avgDelay = mean(DepDelay))

## Plot average monthly delays by hub and airline
ggplot(hubinfo, aes(avgDelay)) +
       geom_area(aes(y = ..count.., fill = Hub),
                 stat = "bin", binwidth = 2, alpha = I(.3)) +
       facet_wrap(Origin~UniqueCarrier) +
       geom_line(aes(y = ..count.., color = Hub),
                 stat = "bin", binwidth = 2, size = 1.5) +
       xlab("Average Monthly Departure Delay (mins)") +
       ylab("Frequency of Occurrence") +
       labs(title = "Distribution of Monthly Departure Delays Since 1990")

ggsave("avgdelay_dist.pdf")

## Plot the trend in monthly average delays by year
qplot(Year, avgDelay, data = hubinfo, xlab = "Year", ylab = "Average Monthly
      Delay (mins)",
      main = "Average Departure Delays Over Time", alpha = I(.6)) +
      geom_smooth(aes(group = 1, color = Hub), size = I(1.5)) +
      facet_wrap(UniqueCarrier ~ Origin) + xlim(1990, 2008)

ggsave("yearly_avg_delays.pdf")

## Collapse data across pre-existing variables to collapse different causes
## of delay

hubstidy <- melt(hubs3, c("Year", "Month", "DayofMonth", "DayOfWeek",
                          "DepTime", "CRSDepTime", "ArrTime", "CRSArrTime",
                          "UniqueCarrier", "FlightNum", "TailNum",
                          "ActualElapsedTime", "CRSElapsedTime", "AirTime",
                          "ArrDelay", "DepDelay", "Origin", "Dest",
                          "Distance", "TaxiIn", "TaxiOut", "Cancelled",
                          "CancellationCode", "Diverted", "Hub",
                          "DelaySize"))
```

```
## Rename the columns of "variable" as "DelayType" and "value" as "Number"

names(hubstidy) <- c("Year", "Month", "DayofMonth", "DayOfWeek", "DepTime",
                     "CRSDepTime", "ArrTime", "CRSArrTime", "UniqueCarrier",
                     "FlightNum", "TailNum", "ActualElapsedTime",
                     "CRSElapsedTime", "AirTime", "ArrDelay", "DepDelay",
                     "Origin", "Dest", "Distance", "TaxiIn", "TaxiOut",
                     "Cancelled", "CancellationCode", "Diverted", "Hub",
                     "DelaySize", "DelayType", "Number")

## Exclude delay causes with value 0 or NA, and excludes weather and
## security as sources of delay as they only account for a very small
## portion
hubstidy2 <- subset(hubstidy, Number != 0 &
                              Number != "NA" &
                              DelayType != "WeatherDelay" &
                              DelayType != "SecurityDelay")

## Include only significant delays
delays <- subset(hubstidy2, DelaySize != "Under 5 mins")

## Plot breakdown of delays by cause
qplot(DelaySize, data = delays, fill = Hub, position = "fill", alpha = I(.9),
      xlab = "Delay Size", ylab = "Proportion of Total Delays",
      main = "Comparative Causes for Delays by Length") +
      facet_grid(~DelayType) +
      theme(axis.text.x = element_text(angle = 30, color = "black")) +
      geom_hline(aes(yintercept = 0.50), color = "black", linetype = "dashed")

ggsave("delay_type.pdf")




###
## Predictive Modelling - Departure Delay
###

## Create a variable for how late in the day the flight was scheduled
nyc.mia.clean <- nyc.mia.clean[-(which(nyc.mia.clean$CRSDepTime == 0)), ]
nyc.mia.clean$time <- hm(nyc.mia.clean$CRSDepTime)
nyc.mia.clean$time <- nyc.mia.clean$time@hour + nyc.mia.clean$time@minute / 60
nyc.mia.clean$time <- nyc.mia.clean$time / 24

## Turn some variables into factors
nyc.mia.clean$UniqueCarrier <- factor(nyc.mia.clean$UniqueCarrier)
```

```
nyc.mia.clean$Year <- factor(nyc.mia.clean$Year)
nyc.mia.clean$Origin <- factor(nyc.mia.clean$Origin)

## Sample 10% of the data since the set is too large
nyc.mia.clean.train <- nyc.mia.clean[sample(nrow(nyc.mia.clean),
                                            nrow(nyc.mia.clean) * .1), ]

## Naive regression model
rf <- randomForest(DepDelay ~ DayOfWeek + DayofMonth + Month + time + Year +
                   UniqueCarrier, data = nyc.mia.clean.train,
                   na.action = na.omit, importance = TRUE)

## Turn it into a classification problem
nyc.mia.clean$delay <- NA

nyc.mia.clean$delay[which(nyc.mia.clean$DepDelay < -5)] <- "Early"

nyc.mia.clean$delay[which(nyc.mia.clean$DepDelay >= -5 &
                          nyc.mia.clean$DepDelay <= 5)] <- "On Time"

nyc.mia.clean$delay[which(nyc.mia.clean$DepDelay > 5 &
                          nyc.mia.clean$DepDelay <= 30)] <- "Late"

nyc.mia.clean$delay[which(nyc.mia.clean$DepDelay > 30)] <- "Very Late"

## Make it a factor for predictions
nyc.mia.clean$delay <- factor(nyc.mia.clean$delay)

## Clean rownames
rownames(nyc.mia.clean) <- NULL

## Create new test and training subsets
nyc.mia.clean.train <- nyc.mia.clean[sample(nrow(nyc.mia.clean),
                                            nrow(nyc.mia.clean) * .8), ]

nyc.mia.clean.test <- nyc.mia.clean[-as.numeric(rownames(nyc.mia.clean.train))
                                    , ]

## Train randomForest for classification
rf1 <- randomForest(delay ~ DayOfWeek + DayofMonth + Month + time + Year +
                    UniqueCarrier, data = nyc.mia.clean.train,
                    na.action = na.omit, importance = TRUE)

## Check how many are exactly accurate
length(which(rf1$predicted != nyc.mia.clean.train$delay))
## [1] 36265 aren't exactly correct
## 64.15% Accuracy
```

```
## Check how we perform on the test
nyc.mia.clean.test$pred <- predict(rf1, nyc.mia.clean.test)
length(which(nyc.mia.clean.test$pred != nyc.mia.clean.test$delay))
## 5182 aren't exactly correct
## Accuracy of 79.51

## Get confusion matrices
## Make factors to use confusionMatrix function
nyc.mia.clean.test$pred <- factor(nyc.mia.clean.test$pred)
cm <- confusionMatrix(nyc.mia.clean.test$pred, nyc.mia.clean.test$delay)

xtable(cm$table)
xtable(rf1$confusion)
```