Analysis of Airline Data In 1997 And 2004

By Jaemin Kim, Adam Lightner, Jonah Toch, Chengzhuang Zheng
Group 5 - STAT 480: Data Science Foundations

Table of Contents:

"Statistics is the science of variation"

— Douglas M. Bates

Introduction

In this project, we are tasked with analyzing the *airlines* dataset from the US department of Transportation Bureau of Transportation Statistics. Specifically, we are looking at data from 1997 and 2004. Our goal is to analyze and provide meaningful insights into trends within the available dataset using tools such as R, Hive and Python.

The *airlines* dataset has 29 variables and 12,541,113 observations for 1997 and 2004. It also contains descriptive information such as departure/arrival time, airtime, different types of delays, cancellations and other information about the specific airplane and airport for that flight. (Note: All mentions of airtime, departure delays and arrival delays are measured in minutes.) A full list of these variables is given at the end of the introduction section. We were also given three additional datasets that have information about the airports, airline carriers and planes mentioned in the original *airlines* dataset. Using Hive, we were able to combine these datasets into one large table and then query out the appropriate columns and rows as needed. For all the analyses, null values were excluded.

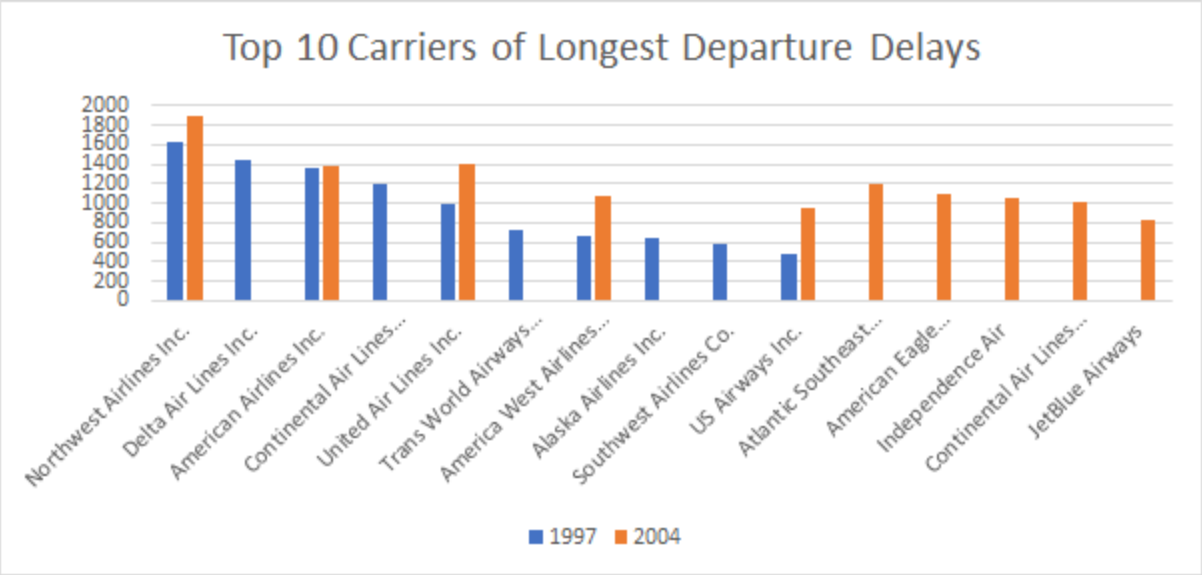
Airline Data Overview

Numerical Variables	Categorical Variables	cont.
DepTime	Year	CarrierDelay
CRSDepTime	Month	WeatherDelay
ArrTime	DayofMonth	NASDelay
CRSArrTime	DayOfWeek	SecurityDelay
ActualElapsedTime	UniqueCarrier	LateAircraftDelay
CRSElapsedTime	FlightNum	
AirTime	TailNum	
ArrDelay	Origin	
DepDelay	Dest	
Distance	Cancelled	
TaxiIn	CancellationCode	
TaxiOut	Diverted	

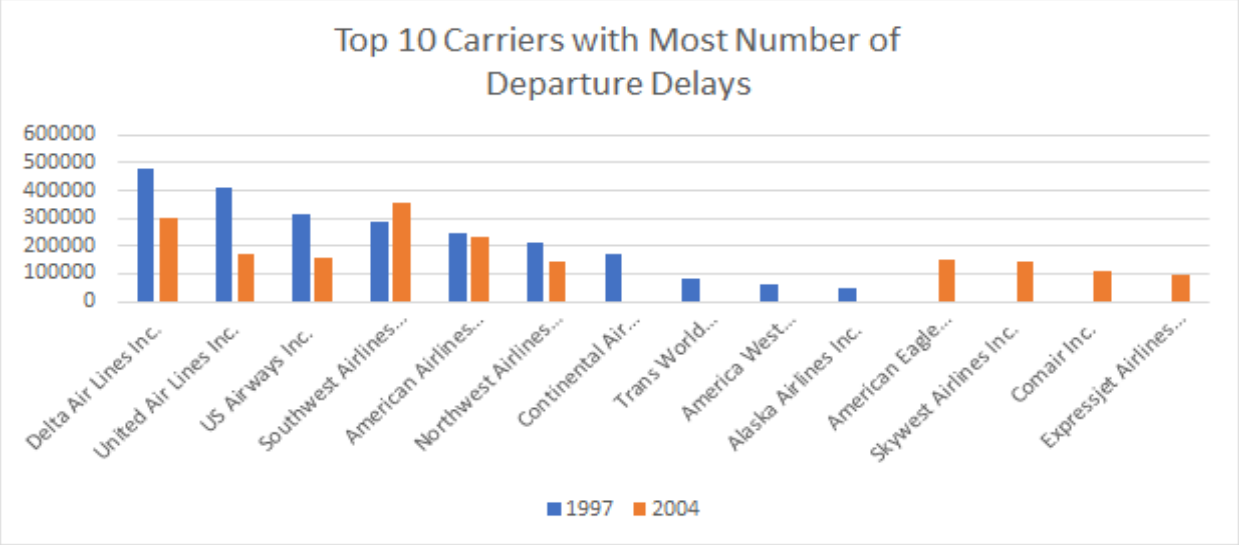
Section I: Airline Analysis

In this section, we looked into the *airlines* dataset and *carriers* dataset to see if there is any relationship between flight delays and airline carriers in 1997 and 2004. In order to use these two separate datasets in this analysis, we joined the two datasets using Hive with *UniqueCarrier* and

Code as key values and created a new dataset containing all variables from both datasets. Then, we executed Hive queries that selected the top 10 airlines with the longest delays as well as the most number of delays for 1997 and 2004 as seen in the charts below.



From the bar chart displaying the top 10 airlines with the longest delays, we can see which airlines marked the maximum delays in 1997 and 2004. Northwest Airlines Inc. had the longest delay in 1997 with around 1600 minutes delay, which is about 26.6 hours. It then also ranked number one for the longest delay in 2004 with over 1800 minutes delay, which is about 30 hours. Delta Airlines, American Airlines, Continental Air Lines, United Airlines, and other airlines made the top 10 list as well with varying delays. Airlines such as Cornair Inc., Atlantic Southeast Airlines, and American Eagle Airlines made the top 10 list in 2004 with relatively long delays compared to the other records in 1997. Looking at the graph, we were able to see that the overall delays in 2004 were higher than those in 1997.

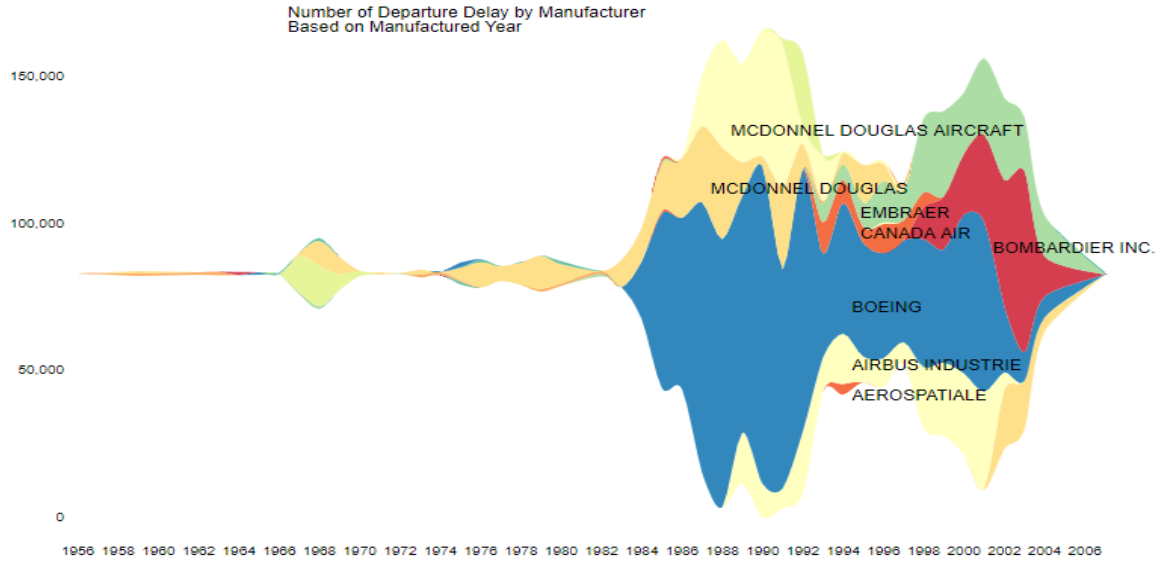


By comparing the number of delays for each airline in two different years, we were able to see some trends among airlines. For departure delays, Delta Air Lines Inc seemed to have the most delays, followed by United Air Lines, US Airways, Southwest Airlines, American Airlines, and so on. When comparing results between 1997 and 2004, we can see that generally the number of delays decreased significantly for most of the carriers.

However, interestingly, the overall departure delays of all airlines in 2004 actually increased almost double the number of that in 1997. We executed a Hive query to select departure delays that are greater than 0, and the outputs for those in 1997 and 2004 were 898,222 and 1,663,078. From these facts, we were able to infer that while airlines improved their delay rates over time, because of much higher flight rates and increased number of planes and airports, the overall number of delays increased nonetheless.

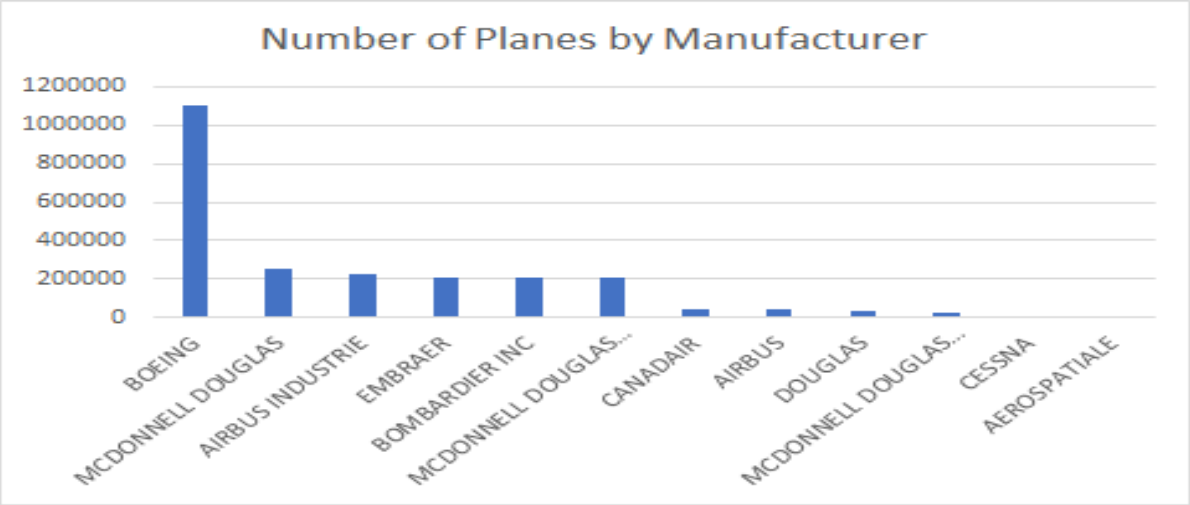
Section II: Analysis of Plane Type.

After looking into the relationship between flight delays and airlines, we were curious about exactly what causes the differences in delays among airlines. For example, why does Delta Airlines had more delays than Alaska Airlines? There were several reasons we could come up with, and here in Section 2, we took a closer look into how plane types can affect the delays using the given *plane-data* dataset. This dataset contains variables such as tail number, type, manufacturer, model, status, aircraft type, engine type, and year it was manufactured. Just like we did in Section 1, we merged the dataset with our original airlines dataset to have all necessary information in one piece. Among 9 different variables in the dataset, we mainly focused on the *manufacturer*, *engine_type*, *aircraft_type*, and *planeyear*, which is the year aircraft was manufactured.

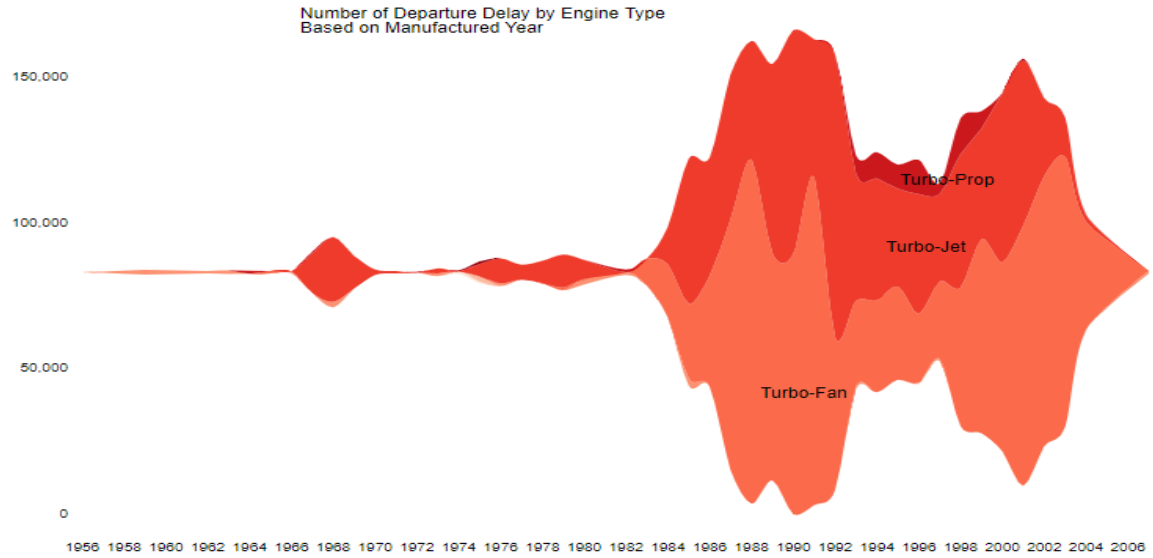


This stream graph depicts number of departure delay by manufacturer based on *planeyear*. Initially looking at this graph, we can see that Boeing company is dominant out of all

manufacturers such as McDonnel Douglas, Embraer, Bombardier, or Airbus. For example, planes manufactured by Boeing in 1990 had the most departure delays compared to other planes crafted in the same year. It is easy to overlook this data and conclude that planes built by Boeing has most problems that cause flight delays, but that is not true.

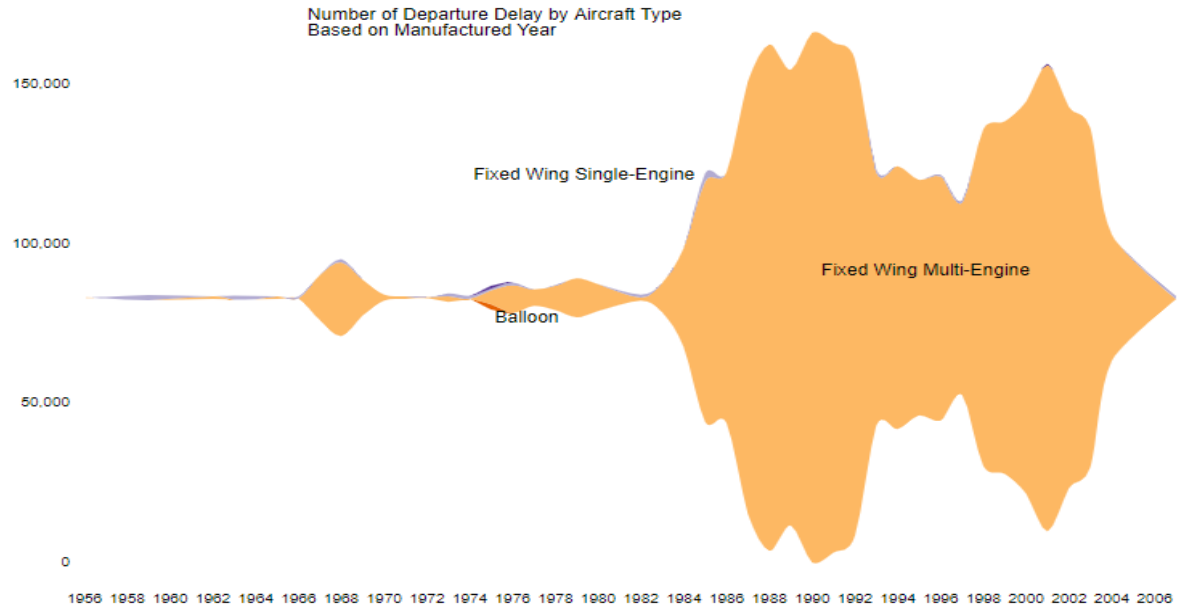


When looking at the whole population of planes grouped by manufacturer in the above bar graph, we end up realizing Boeing is an aircraft manufacturer that crafts more planes than all the other manufacturers combined. Because there are more number of flights by Boeing aircrafts, inevitably there has to be more delays from them. In fact, looking at above stream graph, airplanes recently manufactured by Boeing in early 2000s has almost the same delay rate as other manufacturers. Having plane quantities of almost 5 times the other manufacturers and still recording similar number of delays as others reflects how much Boeing has improved their manufacturing process and plane performance.



Above, we looked into the engine type of planes, instead of manufacturer, to see if there is any trends we could detect from differing engine types. Turbo-jet is a type of engines that was first

equipped onto aircrafts that uses turbine for power. Later, turbo-fans were introduced as a solution to turbo-jets inefficiency. This change of trend in engine type is displayed on our stream graph, where you mostly see turbo-jets until around 1982, and then turbo-fan begins to take up greater portion of the whole quantity after 1980s. In the early time period, majority of the delays were caused by planes with turbojet engines. However, as of 1980, with the increase in population of turbo-fan-equipped planes, delays by planes with turbo-fan dramatically increased. Turbo-prop engine type is also visible on the graph, but it is used for smaller and slower plane types with propellers thus have small number of delays compared to the other two engine types.



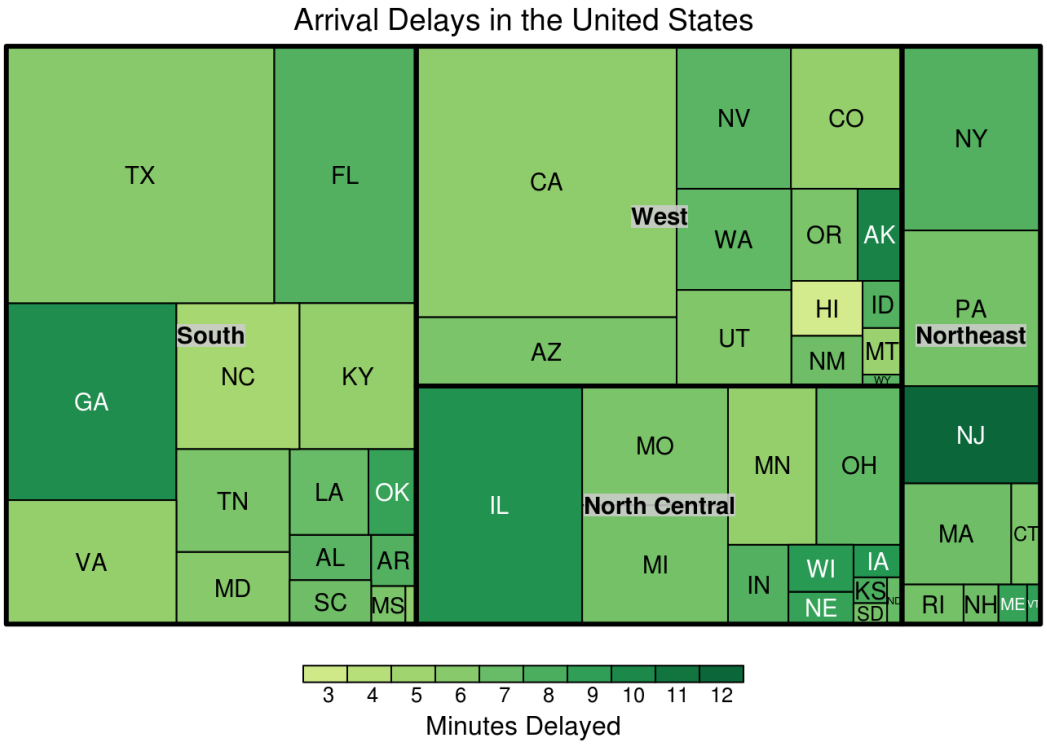
Lastly, we looked at the aircraft type for all manufactured years. As seen on the graph above, Fixed wing multi-engine aircraft represents almost all aircraft types from 1950s to early 2000s, with little glimpses of balloon and fixed wing single-engine aircrafts. Looking at this graph, we can conclude that from the very beginning of aircraft manufacturing, fixed wing multi-engine aircraft was the main manufacturing design, which has steadily proven its efficiency, safety, and overall performance over decades.

Section III: Analysis of Delays by Region

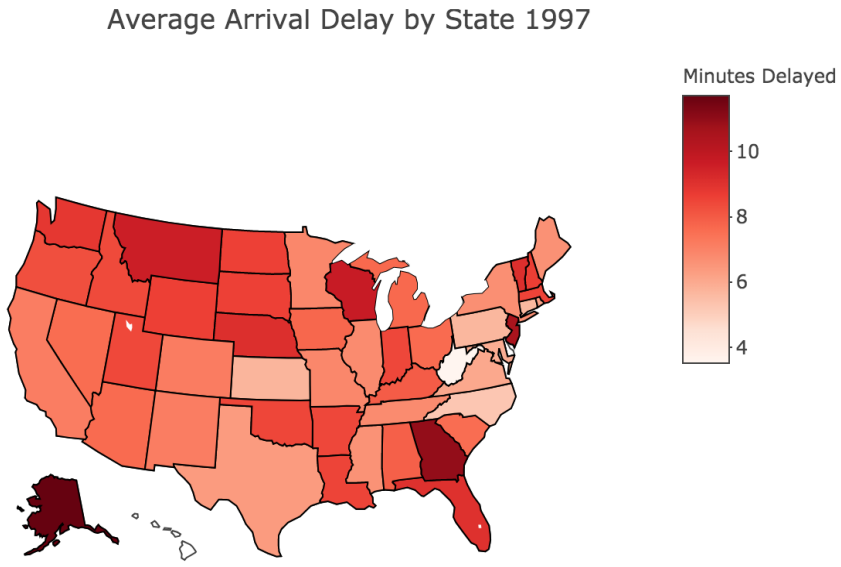
Location could be a key variable when looking at trends in flight patterns for numerous reasons such as weather, population and popularity of the origin and destinations of the flight. With population being so spread out in certain regions and very dense in others, we were interested in analyzing the flight traffic data by region, state and airport. Our end goal is to show how the underlying trends for these flight paths changed over the years 1997 and 2004. For the heatmaps below, destination is used to classify the flights into states and arrival delay is the metric.

To get an idea of the data, we first take a simple look at flight trends by region for both the years combined. Below is a treemap that visualizes the the number of flights by size and average arrival delay for each state by color. We can see that the South had the largest number of total

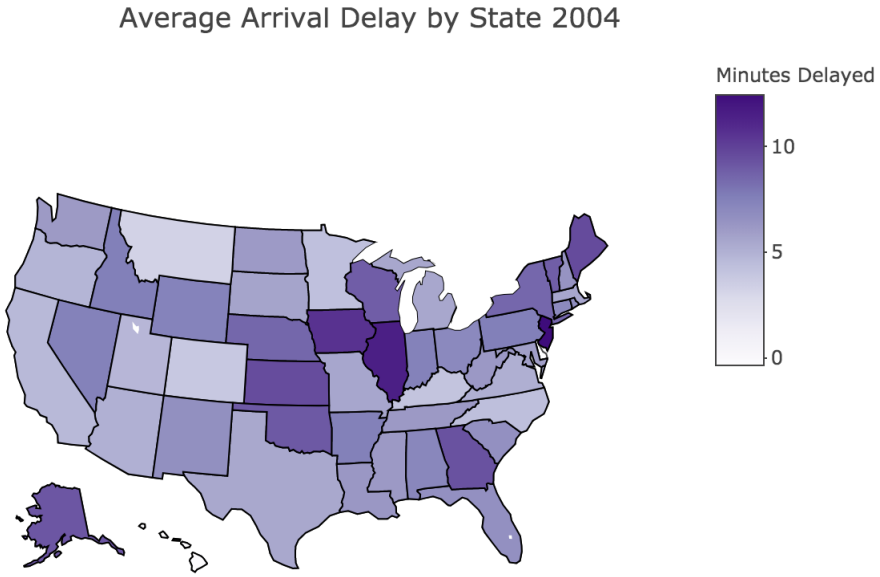
flights followed by the West, North Central and Northeast, respectively. Although the Northeast had the lowest number of total flights, it looks to have the darkest color which indicates that flights that are heading to that region have the largest arrival delay. This is not surprising given the fact that this region is so densely populated. New Jersey appears to have the largest average arrival delay of around 11-12 minutes followed by states like Illinois, Georgia and Arkansas that have a 9-10 minute delay.



Next, we plotted the same information by year to see if there are any differences between 1997 and 2004. To visualize this a little more clearly, the heat map is placed on top of a U.S. map. For 1997, Hawaii was excluded due to its extreme value for arrival delay. From the graph, it appears that in 1997, flying to the Southeast, North Central and parts of the Northeast resulted in high arrival delays. The Southwest and Midwest have low arrival delays compared to the other regions. States like Alaska, Georgia and New Jersey have some of the highest arrival delays at 11.71, 10.94, and 10.5 minutes, respectively. West Virginia had the lowest arrival delay in 1997 of only 3.52 minutes.



Next, move to 2004. The figure below shows that the Midwest has the largest arrival delays in this year. Illinois saw an increase in arrival delay from 6.7 minutes all the way up to 11.5 minutes. Other states in this region such as Iowa, Kansas and Oklahoma also saw large increases in delays too. States in the Northeast had an increase in delay. New Jersey had the highest arrival delay in 2004 at 12.43 minutes and Maine saw an increase from 6.6 minutes to 9.6 minutes. Also, Montana had a significant decrease in arrival delay from 9.6 minutes down to 3.3 minutes. In general, states in the West and Southeast look to have smaller arrival delays than the other regions.



When looking at individual flight paths, one interesting trend that we came across was how flights to the more popular airports in the United States decreased from 1997 to 2004. The tables below use Los Angeles International Airport, or LAX, as an example. It is shown that the five most popular flights from other airports within California to LAX fell significantly between the years. Flights from SFO to LAX fell nearly 47%.

Top Flights To LAX 1997

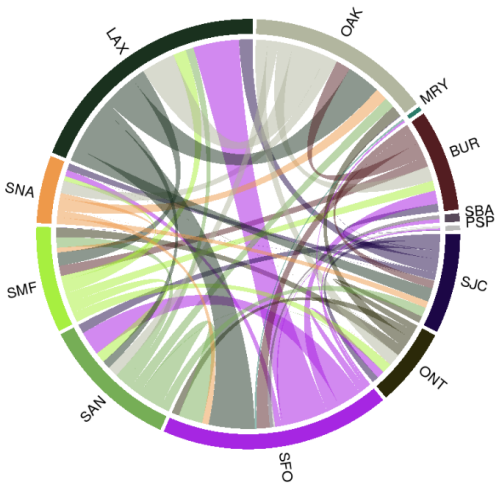
Origin	Dest	Flights
SFO	LAX	17217
LAS	LAX	16860
PHX	LAX	15490
ORD	LAX	12743
OAK	LAX	12686

Top Flights To LAX 2004

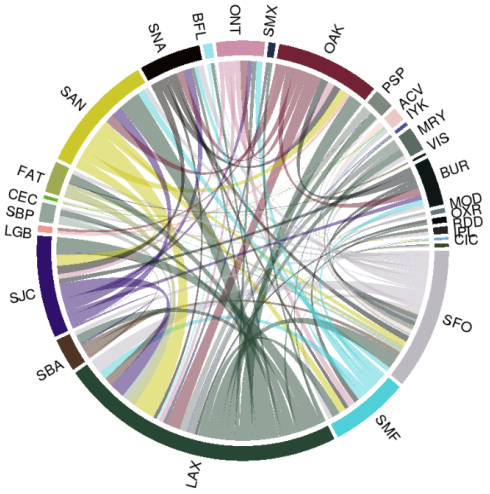
Origin	Dest	NUM
LAS	LAX	12487
PHX	LAX	10422
ORD	LAX	10134
OAK	LAX	9258
SFO	LAX	9191

To explain this, two chord diagrams where created showing in-state flight paths for California. It is clear that there are almost double the amount of airports in 2004 than in 1997. We can infer that the increase of smaller local airports have resulted in a decrease in the proportion of flights to LAX and other larger California airports.

1997 Flight Patterns



2004 Flight Patterns



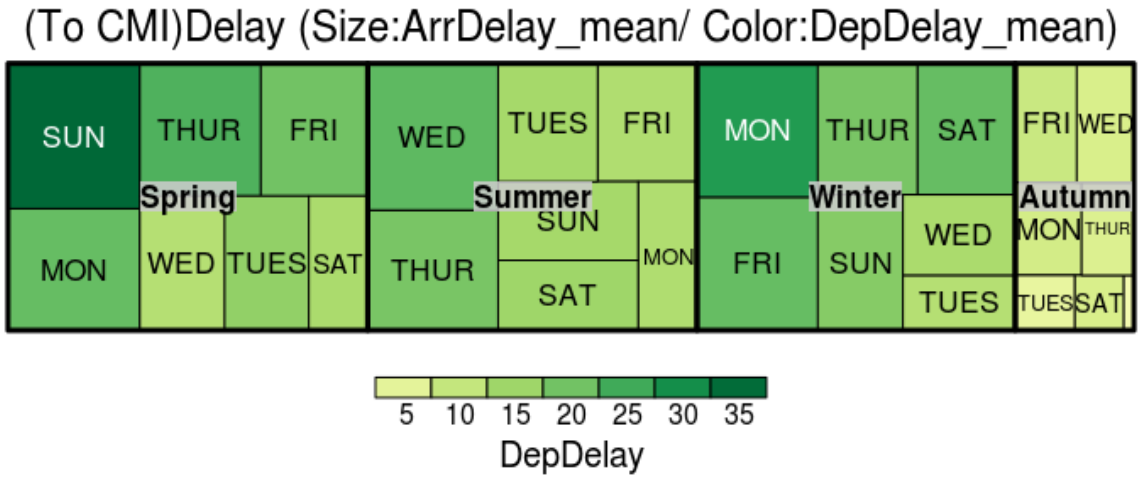
Section IV: Analysis of Delays and Number of Flights by Day, Season, and Year

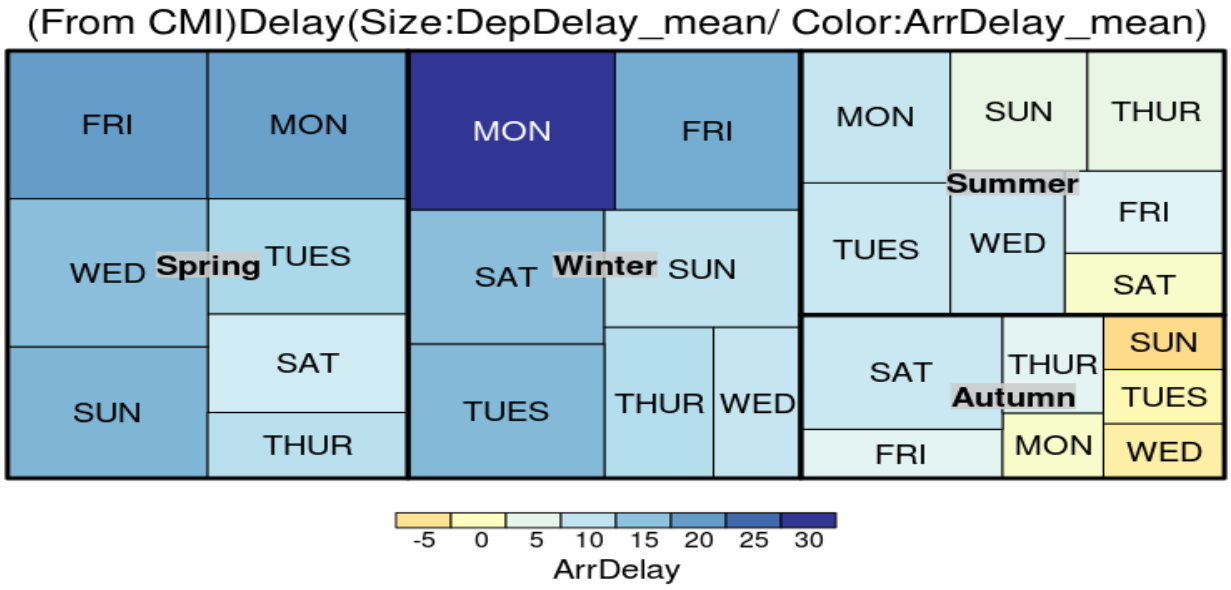
To get an idea of how these analyses are relevant to us in particular, we took a look at two airports that are in Illinois, Willard Airport at the University of Illinois(CMI) and the Chicago O'Hare International Airport(ORD). In the first part of this section, we explore the delays and number of flights of ORD and CMI by different time intervals (day of week, month, season and year) and by regions (other airports and states).

In the second part, we focus on the overall tendency of delays by year, season, day of week in 1997 and 2004. We visualized the mean and percentiles of delays (mainly focus on departure delays, because we found similar trend in arriving delays). We also offered some suggestions to the passengers about which season and which day of week might have higher probability of longer delays, and provide one possible reason to do with the increase in delays.

In this section we used Hive and Python for data preprocessing. We selected the data with specific *Dest* and *Origin* and group by specific columns like *year*, *month, dayofweek, state* and we also created new column *season*.

Specifically, for the first figure below, because there are negative average delays for some days of week, we converted the original data to ten intervals with equal length. This is used to create the size of the squares.



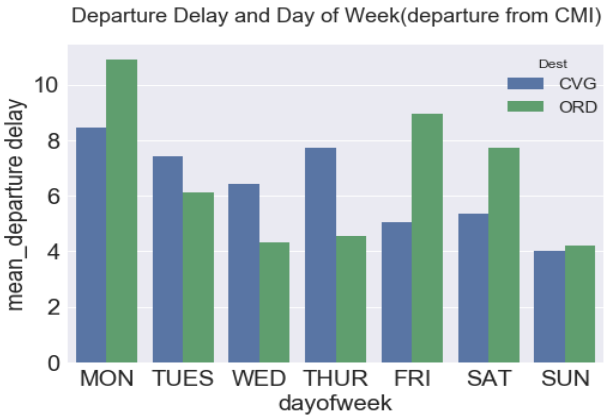


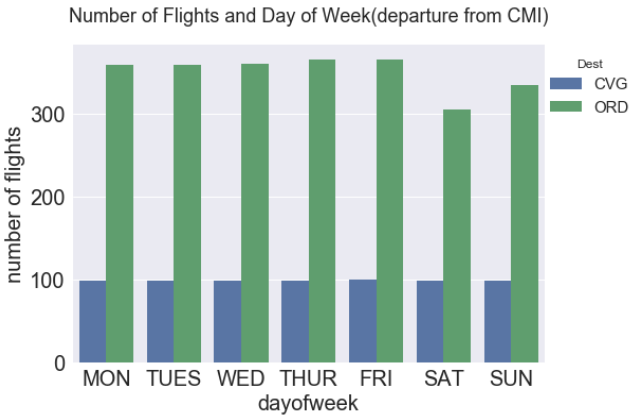
Willard Airport at the University of Illinois(CMI) is the closest airport to UIUC and in 2004, it only had flights to and from the Chicago O'Hare International Airport (ORD) and the Cincinnati/Northern Kentucky International (CVG) airports. The number of flights departing to and from CMI were both 3,140 in 2004, which was around 8.7 flights per day. Since CMI only has data for 2004, we were most interested in the delays for the four seasons and days of the week and how those delays relate with ORD and CVG.

The graphs above show the average delays(with known delays) for each season and day of week for flights to and from the CMI airport. The first graph has the arrival delays for each day of the week increase the size of the squares, while the departure delay darkens the shade. The second graph is the opposite, where the size of the graph is the departure delay and the shade is darkened by arrival delay. We included both as together they show the difference in the delays to and from CMI.

Now interpreting the data, we see for flights to CMI, Spring, Summer and Winter all had visibly longer arrival and departure delays than Autumn. Sundays in the Spring had an extremely high average departure delay of around 35 minutes. For the most part, flights with longer departure delay also had longer arrival delay. For flights from CMI, Spring and Winter had longer arrival and departure delay than the other two seasons. Fridays in Spring had higher average departure delay at around 16.33 minutes.

Now, we move onto analyzing how delays differ between the two airports CMI flies to (CVG and ORD).

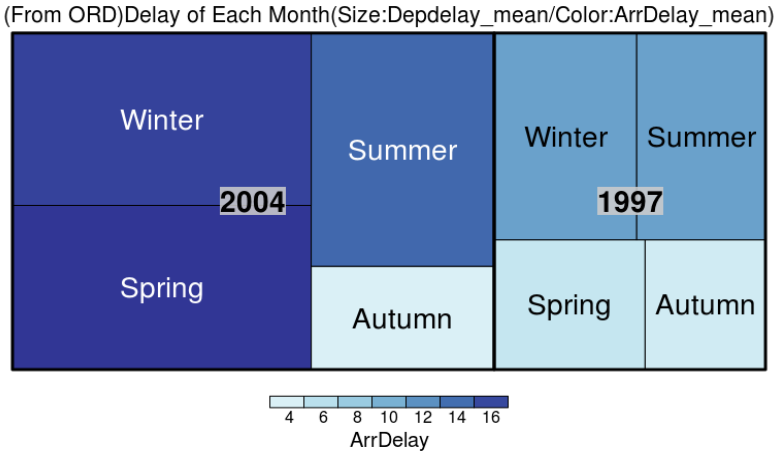


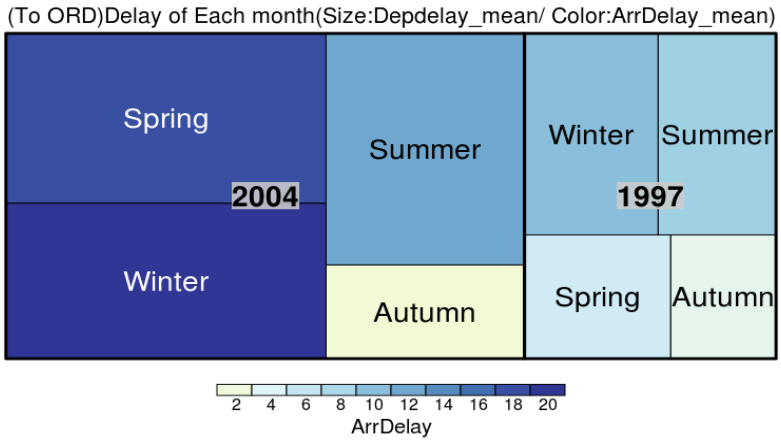


The graph on the top-left shows us the mean departure delay (with known delays) for each day of the week for the flights traveling from CMI to CVG and ORD. (The flights arriving at CMI share the same trend of the graph above) We notice for Tuesdays, Wednesdays and Thursdays ORD had higher delays, while CVG had higher delays for the other four days. For flights to or from CVG passengers would on average encounter lower delays during the weekend. For ORD, passengers were delayed the longest on Mondays and Fridays.

Looking at the graph on the top-right, we see there is about three times as many flights from CMI to ORD, making it a much more popular destination.

Next, we will explore the flights to and from ORD, one of the biggest airports in the world. We do this by visualizing the departure and arrival delays for flights to and from O'Hare International Airport.

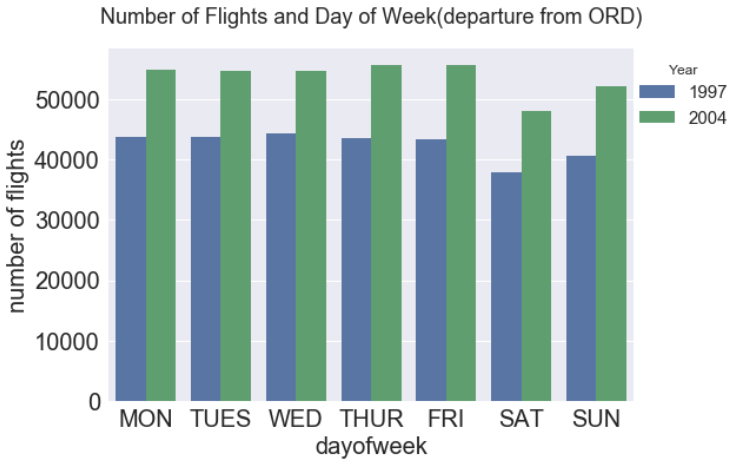




For the flight departures from ORD in Spring, Winter, and Summer in both 1997 and 2004, shown in the first tree diagram, Autumn showed the lowest departure and arrival delays, and Winter the highest. In the second diagram, for flights arriving at ORD, we see a similar delay pattern. This pattern is consistent with our previous analysis. One noticeable difference we found between the two diagrams was the average arrival delays are much higher in 2004 than that in 1997.

Overall, we can infer that there tends to be more severe delays in Winter and Spring than in Summer and Autumn. Also, both arrival and departure delays in 2004 were larger compared to that in 1997. From the above results, weather seems to be a potential reason causing the difference in delays in each season.

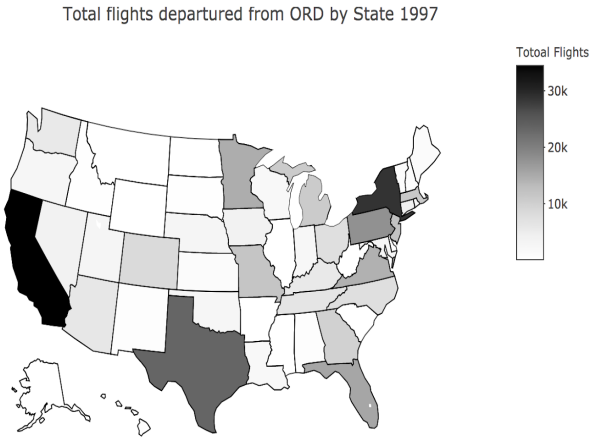
Next, we look at the number of flights leaving from ORD. As expected, we found that the flights to and from ORD increased in 2004. Also visualized, is the the number of flights for each day of week.



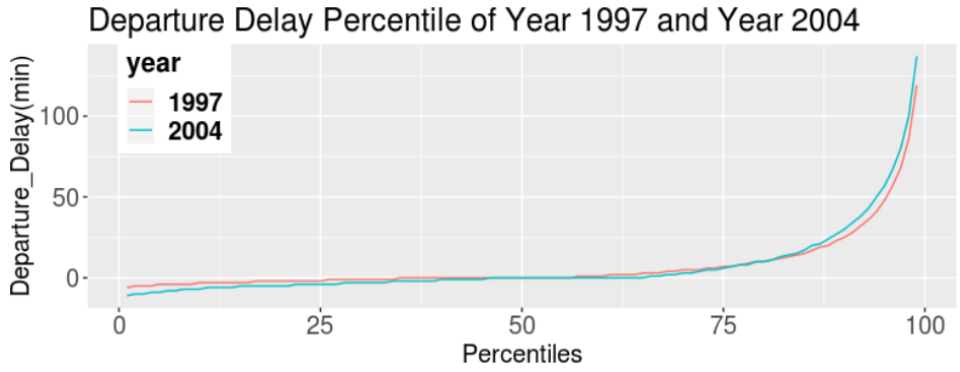
In 1997, the total number of flights was around 40,000 during weekends and 45,000 during weekdays (yearly). In 2004, the number of flights grew to 50,000 on weekends and 55,000 on weekdays (yearly).

The total number of flights departing or arriving at ORD in 1997 is around 297,500 and the total for 2004 is 376,000, which is a 26% increase. The average departure delay for flights from ORD for 1997 is 9.8 minutes and for 2004 it is 13.8 minutes, which is a 41% increase. A potential reason for this may be the expansion of ORD airport could not catch up with the speed of increasing flights, in turn leading to higher delays in 2004 compared to 1997.

Lastly, we wanted to see if there were any significant changes in destination for flights leaving ORD.

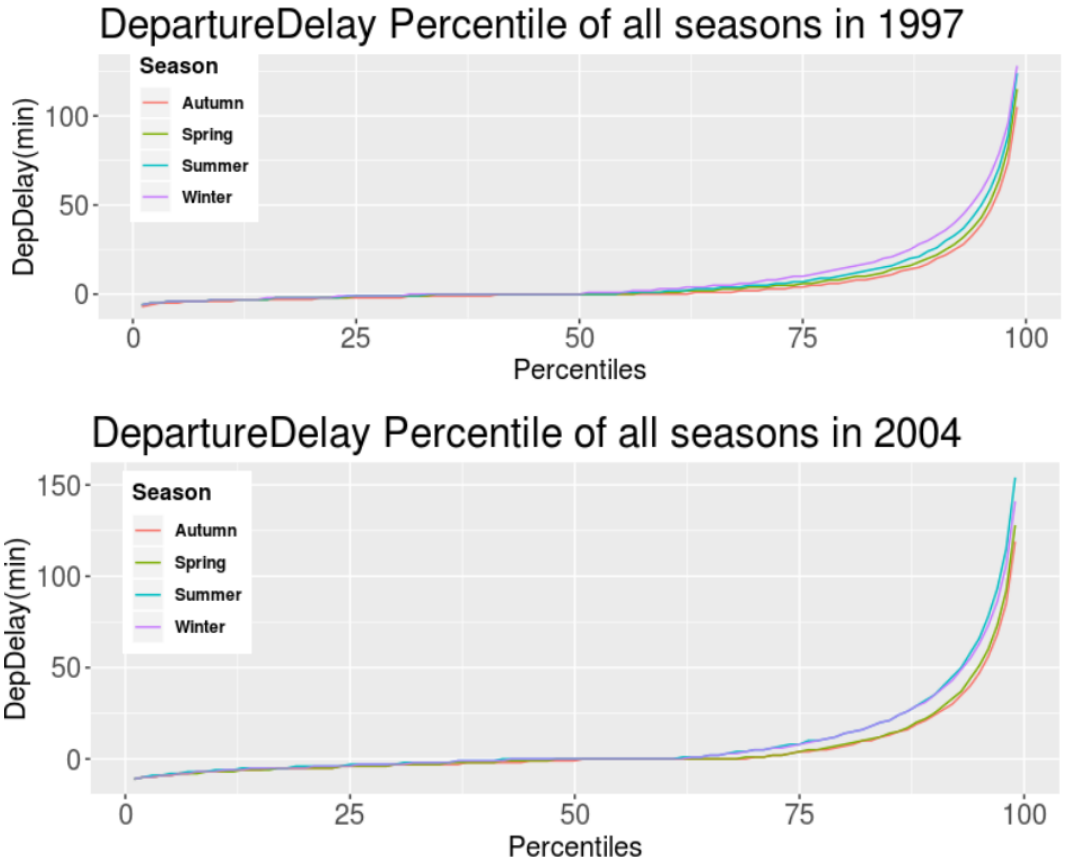


The graphs for the two years are nearly identical. Besides a few a few states in the Northeast where flights between ORD and their airports increased, the trends are the same. California, New York, and Texas are the main destinations for flights leaving ORD. With the exception of California, it seems that the majority of flights head East.



This graph visualizes departure delay quantiles (with known delays) for the years 1997 and 2004. The biggest difference between the two years are in the right tails. The delay median for 1997 is slightly below zero, while for 2004 it is slightly above zero. The higher 80-99 percentile for 2004 implies the probability of encountering extremely long flight delays increase in 2004. The 90th departure delay percentile for 1997 is 25 minutes and for 2004 it is 30 minutes. The 90th percentile for arrival delay in 1997 is 30 minutes and it is 35 minutes in 2004.

Our next analysis, focuses on looking at the percentiles for departure delay based on season.



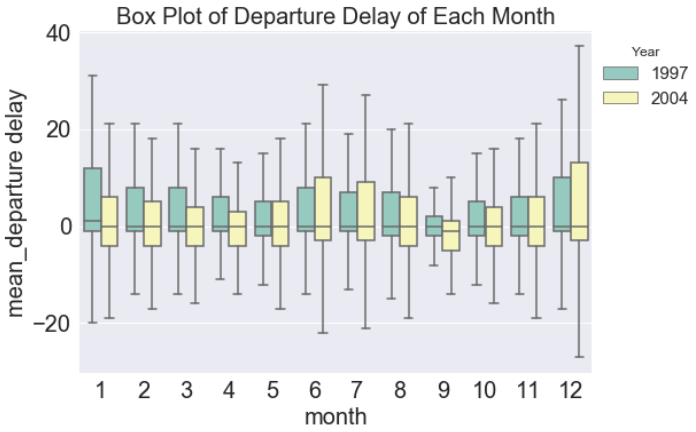
The first graph shows departure delay percentiles for each season in 1997, and the second is for 2004. The 1st to 50th departure delay percentiles in 1997 was similar to those in 2004. The median for delay was around zero minutes for all seasons in both years, which meant around 50% passengers would not run into any flight delay. For the 50th to 99th percentiles, the value for Winter was the highest, followed by the Summer season. In 2004, the percentiles for Summer increased and were extremely close to those for Winter season.

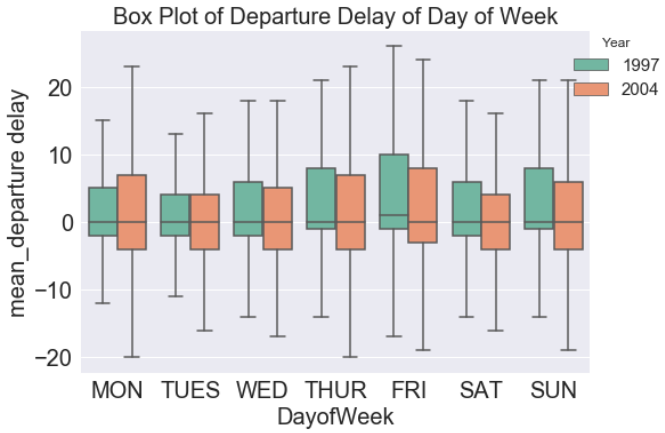
Overall, a passenger was more likely to wait longer for departure in Winter and Summer in these two years.

For example, we listed the 85th and 90th percentile for each season of these two years to show the percentiles as we approach the right tail.

	Year	1997				2004			
	Season	Winter	Summer	Spring	Autumn	Winter	Summer	Spring	Autumn
Percentile (minute)	85th	21	16	14	11	21	21	14	13
	90th	33	26	22	20	35	35	25	24

Moving on to the analysis of departure delays for the different months. The first graph (bottom-left) shows the the departure delay by month for 1997 and 2004. The median for all months is close to zero, except in January of 1997 where it is a bit higher. The interquartile ranges for the Summer and Winter months seem to be much larger than those of Spring and Fall, indicating that delays are more variable during these times.



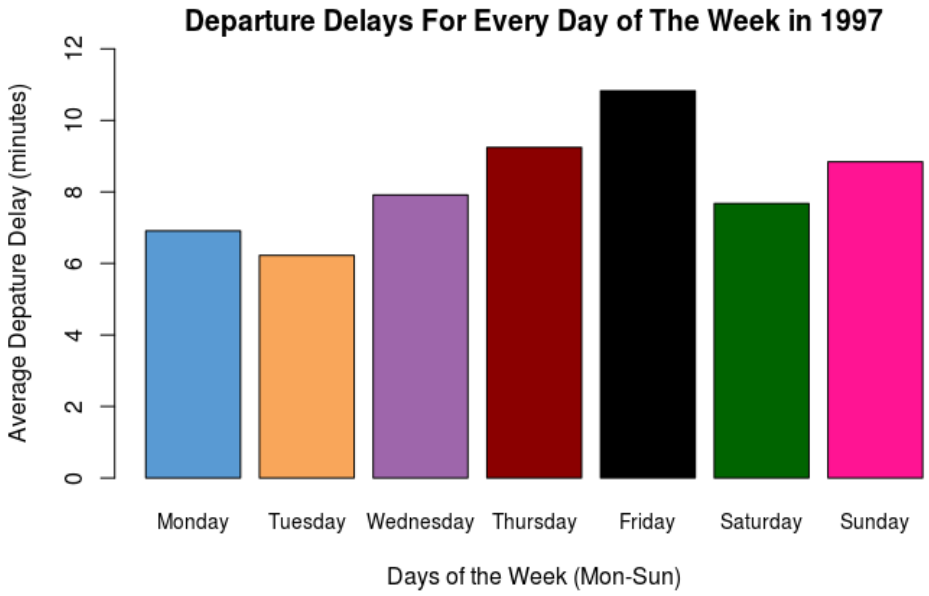


The second graph (top-right) explains the relationship between average departure delay percentiles and day of week. When we looked into the daily impact of these two years, the median departure delay were all around zero minutes, which meant half of the passengers would not have delay problems.

In 1997, flights on Friday were more likely to be delayed than other days. In year 2004, the departure delay on Thursday, Friday and Monday were more severe. In the table below, we show the number of flights by day of the week. We can see that the delay is positively correlated with the number of flights for each particular day.

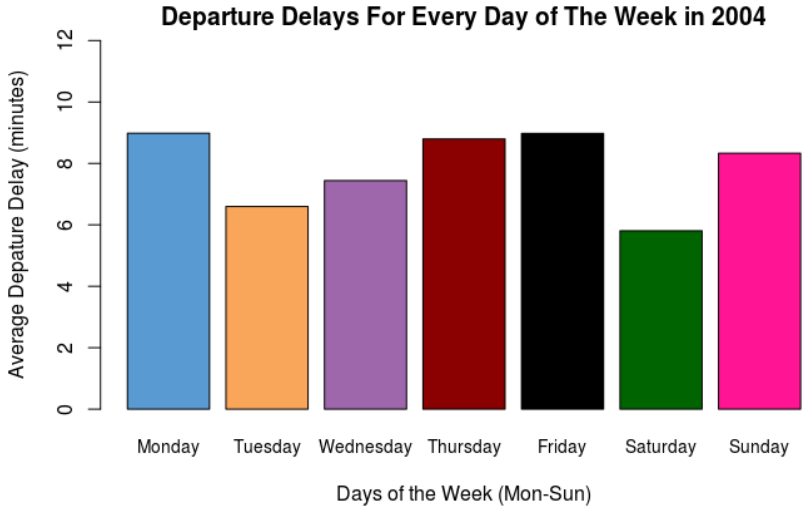
Number of flights	MON	TUES	WED	THUR	FRI	SAT	SUN
Year1997	790298	791617	802130	785731	786342	706198	749527
Year2004	1044508	1033863	1036521	1060245	1061447	903807	988879

Our next analysis is focused on the mean departure delay rather than the median. Visualized below, is the departure delay for each day of the week in 1997. This analysis should be interpreted as the difference between when a flight departs compared to when it is expected to. The mean values are higher than the median ones shown above, because a flight is much more likely to be heavily delayed than it is to leave very early.



1997 Mean Value Departure Delay	
Monday	6.91
Tuesday	6.23
Wednesday	7.91
Thursday	9.24
Friday	10.83
Saturday	7.68
Sunday	8.84

Friday had the longest average delay, being about 4 minute longer on average than the day with the lowest average delays, Tuesday.

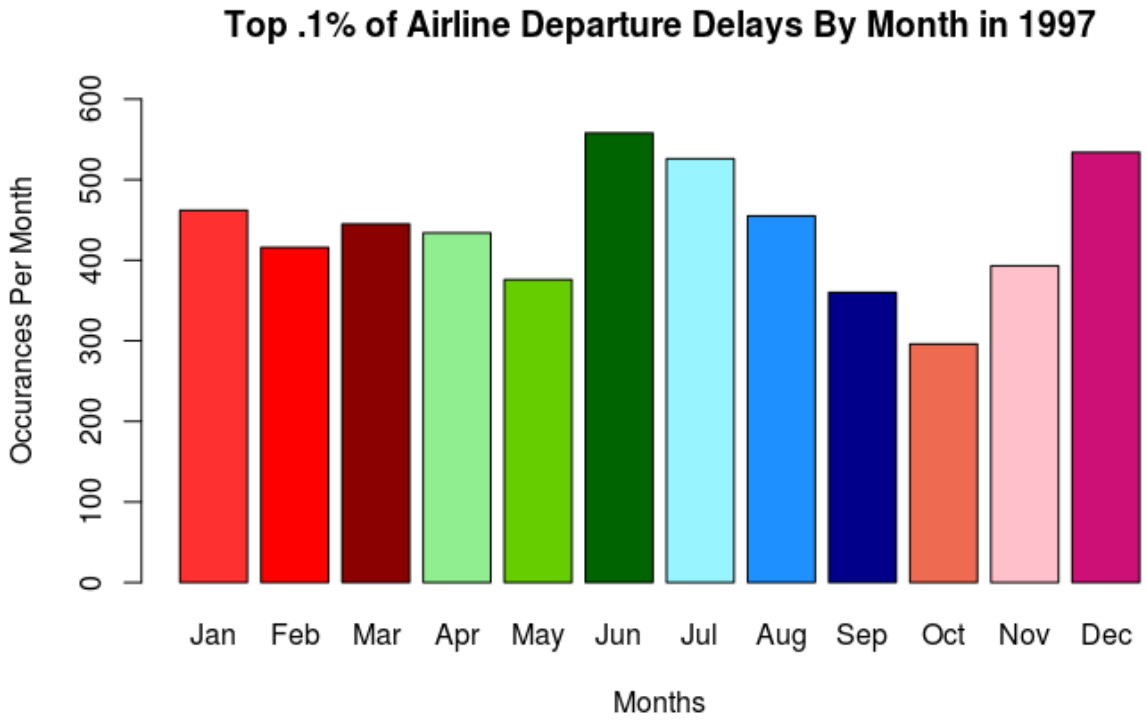


	1997 Rank	2004 Rank
Monday	6th	1st
Tuesday	7th	6th
Wednesday	4th	5th
Thursday	2nd	3rd
Friday	1st	2nd
Saturday	5th	7th
Sunday	3rd	4th

Looking at the 2004 day by day data, there is now much less of a difference from the longest and shortest delay. Surprisingly, Monday went from on average almost 4 minutes less of a delay than Friday, to just barely more than it. To the right of the 2004 bargraph, there is a table showing the difference in ranking for each day of the week (1st meaning highest delays, 7th meaning least). Besides Monday making a huge jump, for the most part the ranking of each each day of the week stayed the same.

While there is certainly a difference in mean delays for each day of the week for both 1997 and 2004, it is hard to determine whether or not the results would mean much to someone who was looking to decide which day of the week to fly on. Friday may consistently be one of the most delayed days of the week per year, but on average it is only by a few minutes, which most likely is not a deal breaker.

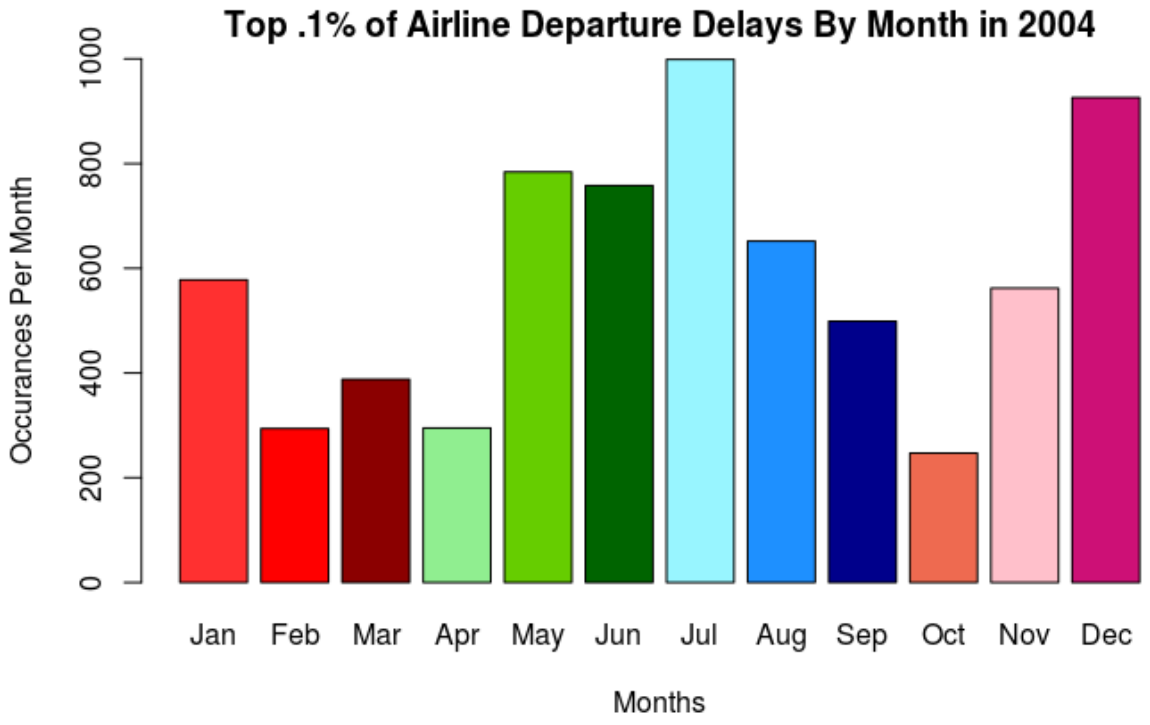
Earlier, we visualized the departure delays by season. From that, we were able to see on average, Winter and the beginning of Spring had the longest average delays. The graph below works as a focus the top .1% of the delays from that plot. Looking at this, we can better understand the very extremes of the data.



The first interesting thing that stands out is June and July are among the months with the highest delays (these departure delays are almost four hours long). This is interesting because Summer as a whole has a much lower average departure delay than Winter, but as shown here in 1997 it also has the highest extreme departure delays. This leads to the conclusion Summer on average has

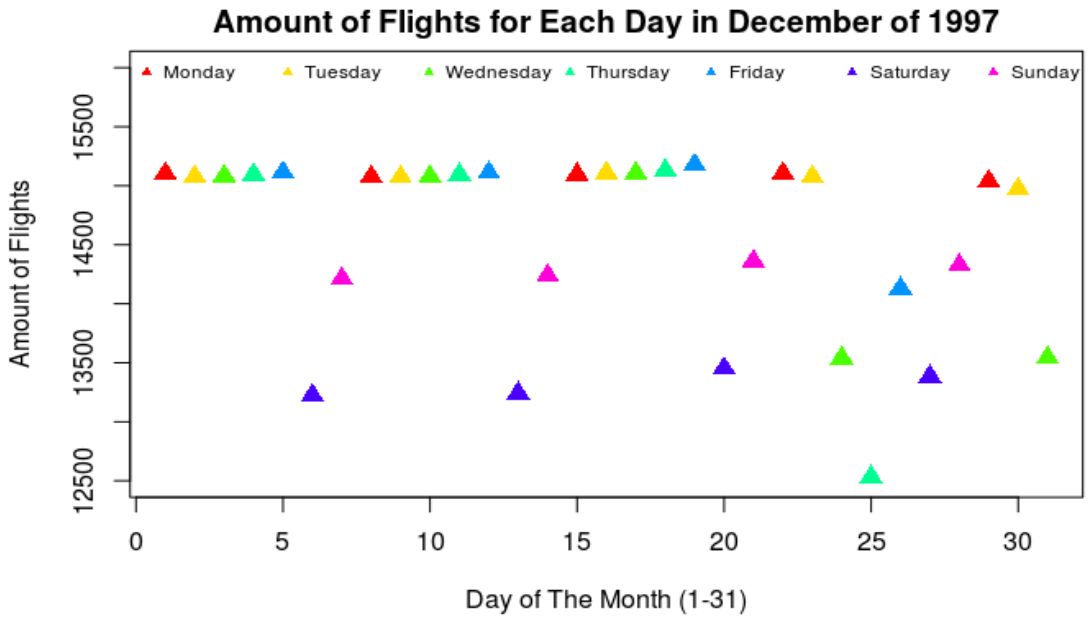
lower departure delays than Winter, but approximately the same amount of extreme departure delays.

2004 shows similar results, with the most standout months being in the Winter and during the Summer. The main difference, although consistent with other 1997 to 2004 comparisons, is the increased variability between the highest and lowest months. I attribute this fact to higher sample sizes for 2004, which allows the differences between the months to become more proclaimed.



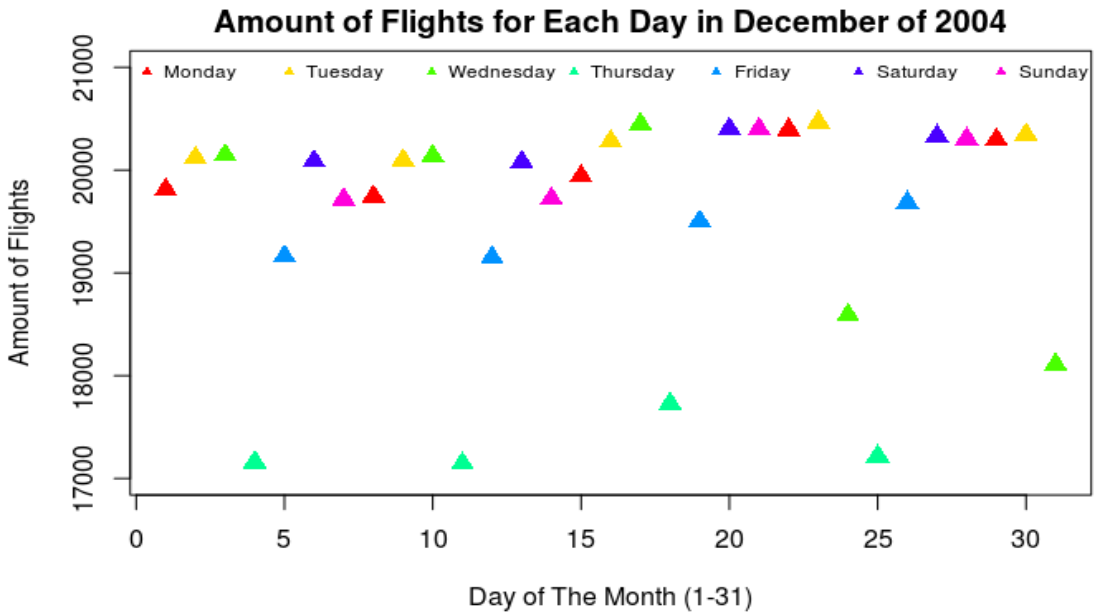
To summarize, if someone was fearful of an extremely long departure delay and was flying in 1997 or 2004, they are better off flying in Spring and Autumn.

Section V: Analysis of Flights by Holiday



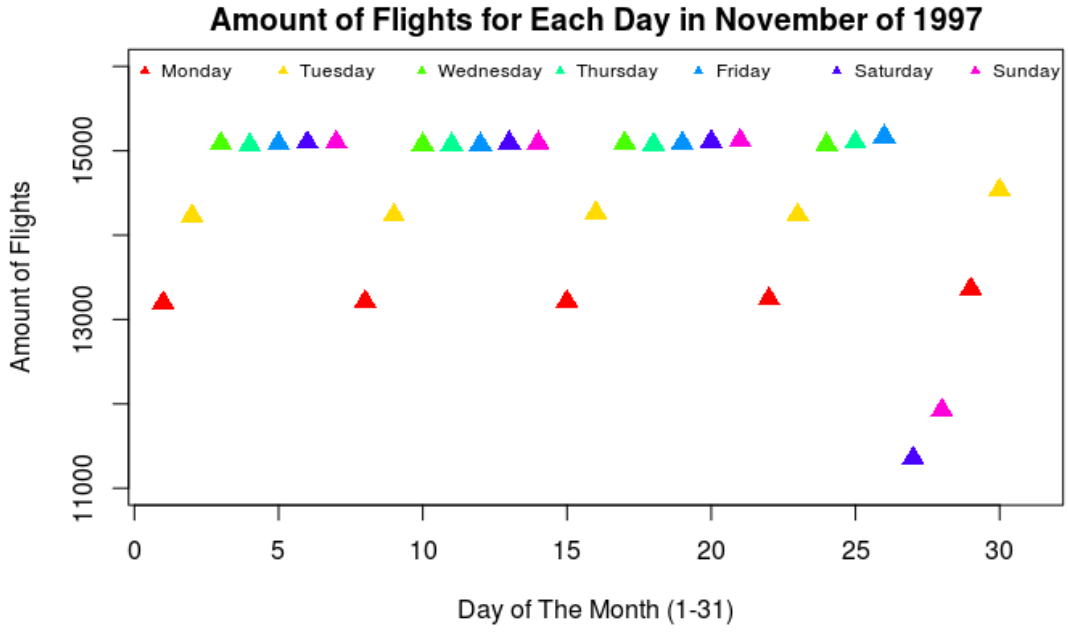
The plot above shows the amount of flights for each day of December 1997. The first three weeks of the month are relatively the same. Monday through Friday have about the same amount of flights, with Sunday having roughly 5% less flights than Mon-Fri, and Saturday having 10%. This trend changes on Christmas Eve, with the holiday having 10% less flights than the earlier Wednesdays in the month, and the following day on Christmas the amount of flights drops 17% compared to the first three Thursdays. The quantity of flights start to rise back to regular numbers just to drop again for New Years Eve, with a similar flight count to Christmas Eve.

Now looking at the results from 2004, there is a clear difference in the trends for each day of the week compared to 1997. There is substantially more flights on average compared to 1997, which better allows each days true proportion to show in addition to the fact the trends between the years are different.

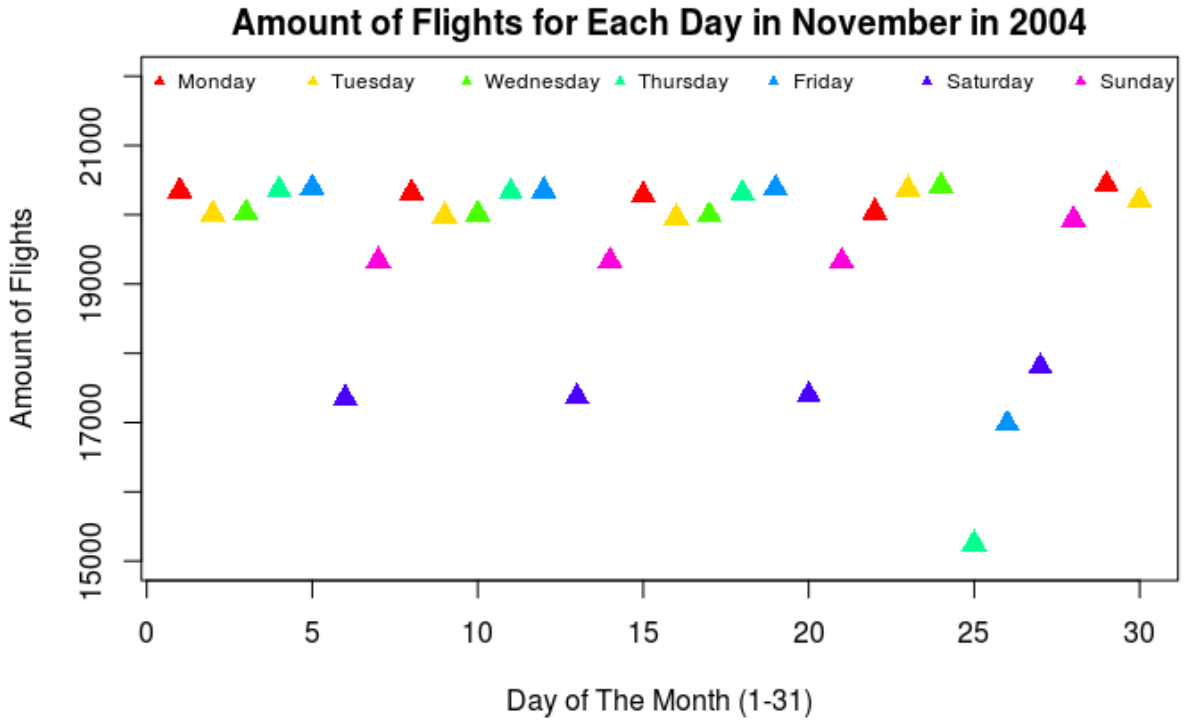


2004 has similar results in the way Christmas Eve and New Years Eve both have about 10% less flights than the same day of the week for the first three weeks of the month. What differs the most is Thursday, which has very low amount of flights throughout the month. This makes Christmas not stand out in particular, and leaves the interesting question on why Thursdays in December of 2004 had so many less flights relatively.

Looking at another month, November, we see data which is mostly uniform between again. The plot shows a clear drop in amount of flights on November 26th and 27th, the weekend following Thanksgiving.

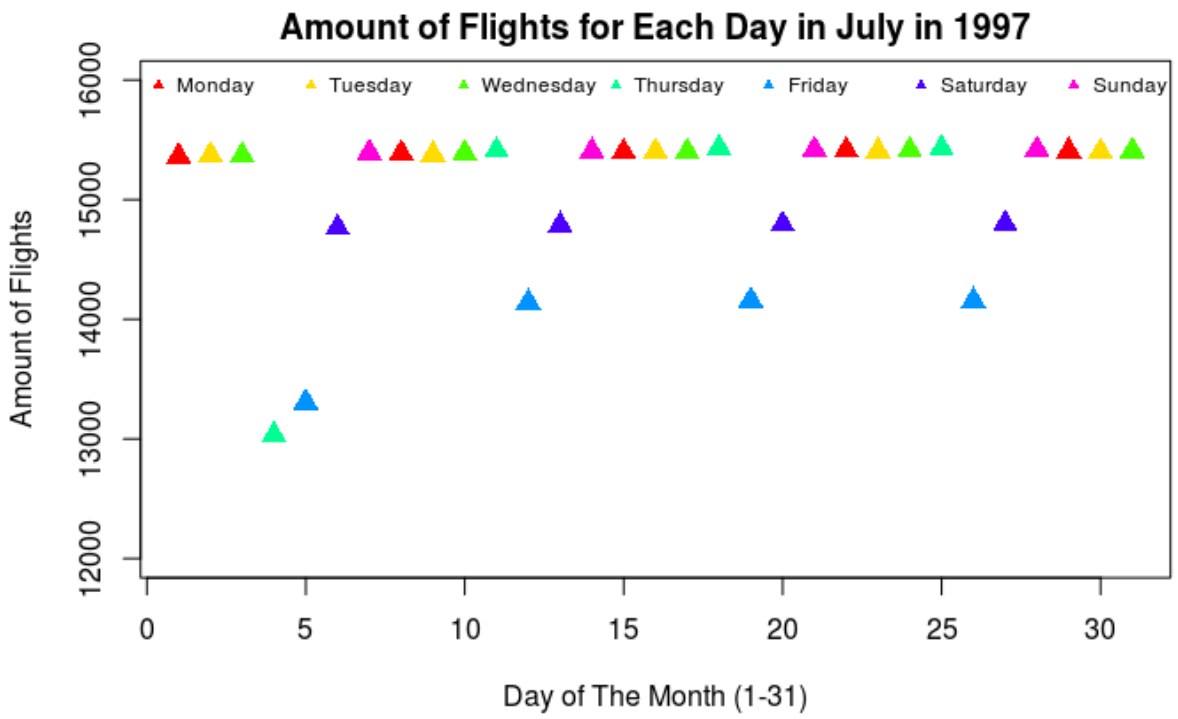


While we cannot definitively declare why these days have significantly less flights, it is certainly clear being near a holiday is a factor. Next, we will compare it to November of 2004, this time Thanksgiving was on the 25th.

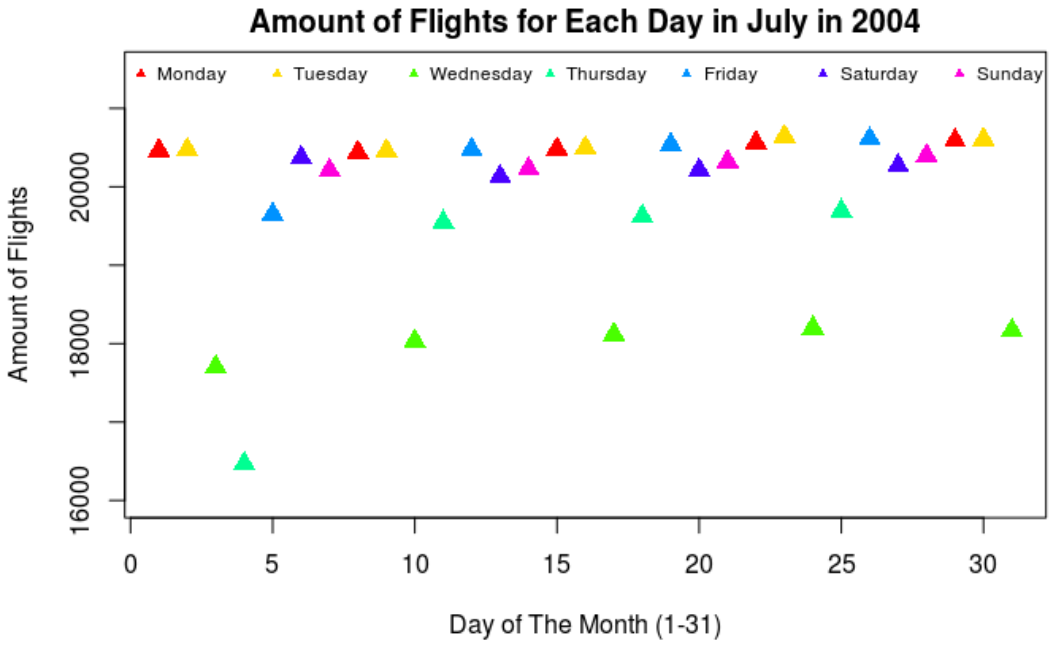


The 2004 data once again has much more variability between the 1997 data. Much like the 1997 data, the holiday of Thanksgiving is a good indication of a drop in flights, with Thanksgiving and the day following having the least amount of flights in the month by far. Interestingly, in 1997 only the weekend following Thanksgiving had visually less flights, while in 2004 Thanksgiving also showed much fewer flights flights than the Thursdays of the same month.

July of 1997 reinforces the conclusion of holidays having significantly less flights, with both July 4th and the following day having significantly less flights than the rest of the month.



July of 2004 is very similar, and just like the previous plots there is increased variability between the individual days of the weeks.



Conclusion:

In conclusion, we found many similarities and differences in flight trends between the years 1997 and 2004. Section I showed that delay rates slightly decreased over time even though the total number of delayed flights increased significantly. We found that overall delay times were fairly similar but changed significantly when inspected by region. It was also concluded that an increase in the number of local airports may have caused a decrease in number of total flights into larger international airports. In following sections, we showed that seasonal flight trends stayed the same, but delays by day of the week changed. It also revealed the overall tendency of flights delays. It is more likely to experience extreme long delays in Winter and Summer seasons. Looking into the flight connection between CMI and other two airports, and ORD with other states, we found that the delay patterns were consistent with the general tendencies. And the increasing number of flights might be the reason for higher delays in 2004 than that in 1997. Generally, as distance increases the number of flights decreased except for the large airports like airports in CA and NY. Lastly, Section V showed us the affect holidays have on the amount of flights.We learned holidays such as Christmas, the Fourth of July, and Thanksgiving were all good indicators of a lower amount of flights.

References:

1. Stat 480 - Data Science Foundation lecture slides and code.
2. Data sources:
    - 2.1. Airlines data: The Airlines Dataset comes from the US Department of Transportation's Bureau of Transportation Statistics (BTS) ,http://www.transtats.bts.gov/OT_Delay/NewAirportList.asp?xpage=OT_DelayCause1.asp&flag=undefined
    - 2.2. Airports data: The Airport data set comes from the Open Flights initiative,http://openflights.org/data.html
    - 2.3. Carriers data: The Carriers Dataset comes from the US Department of Transportation Bureau of Transportation Statistics (BTS),http://www.transtats.bts.gov/Download_Lookup.asp?Lookup=L_UNIQUE_CARRIERS
    - 2.4. Plane data: The Plane Data Dataset comes from the FAA registration database,http://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/

Contributions:

Jaemin - focused on the analysis of airlines(carriers) and plane types in Section 1 and 2. performed data manipulation in Hive to merge multiple datasets, remove NaN values, and extract only the information I need using multiple queries. Teamwise, Everyone in the team was attentive to team meetings and worked hard to complete sections each of us were responsible for.

Jonah - Developed the outline at the beginning on what topics we should focus on. Also, created the mean departure delays analysis in section IV, the analysis on the top .1% of the data, and the entirety of section V.

Chengzhuang - Help combine the dataset; Section IV, explored the delays and number of flights to do with ORD and CMI different time intervals (day of week, month, season and year) and by regions(other airports and states); Concentrated on the overall tendency of delays by year, season, day of week between years.

Adam - Was responsible for helping combine all data in Hive, writing the Introduction and Section III - Analysis of Delays buy region. Helped with U.S. map visualizations used in other sections.