ST2195 Programming for Data Science

Coursework Project

Analysis of US flights

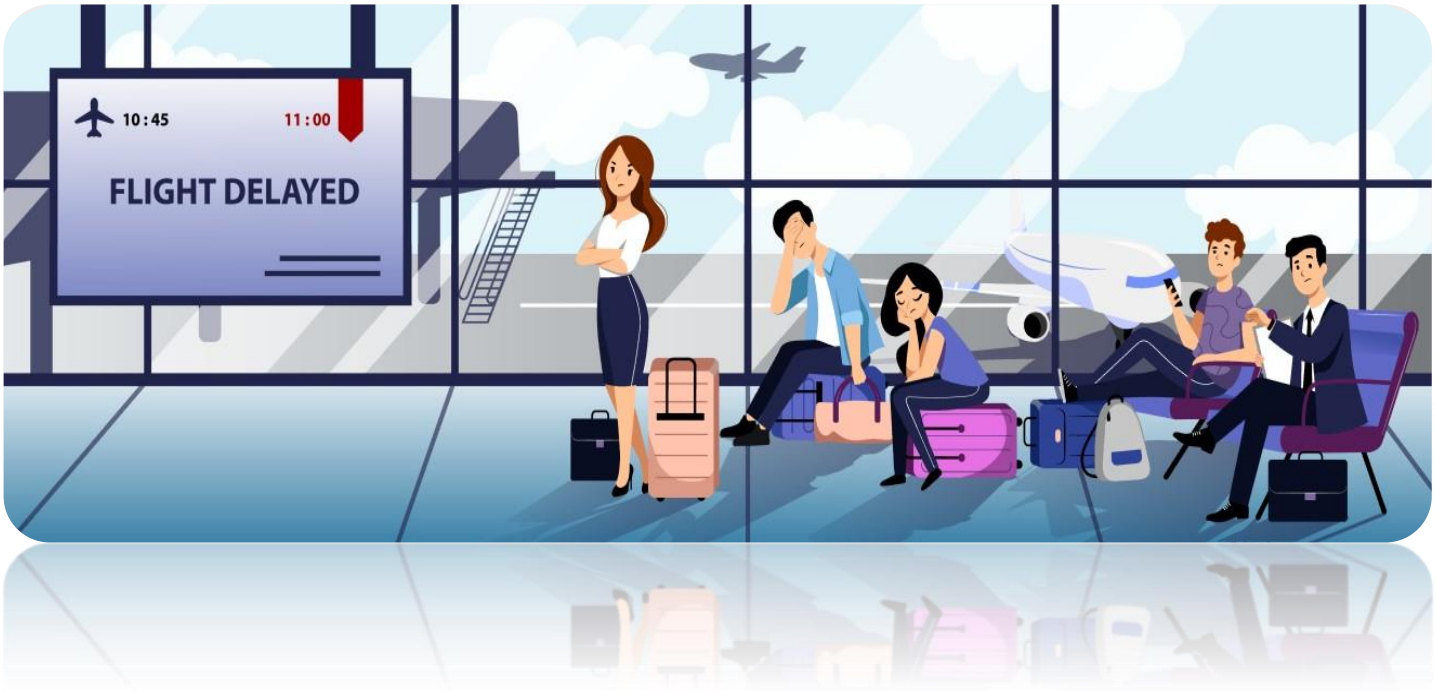


Table of content

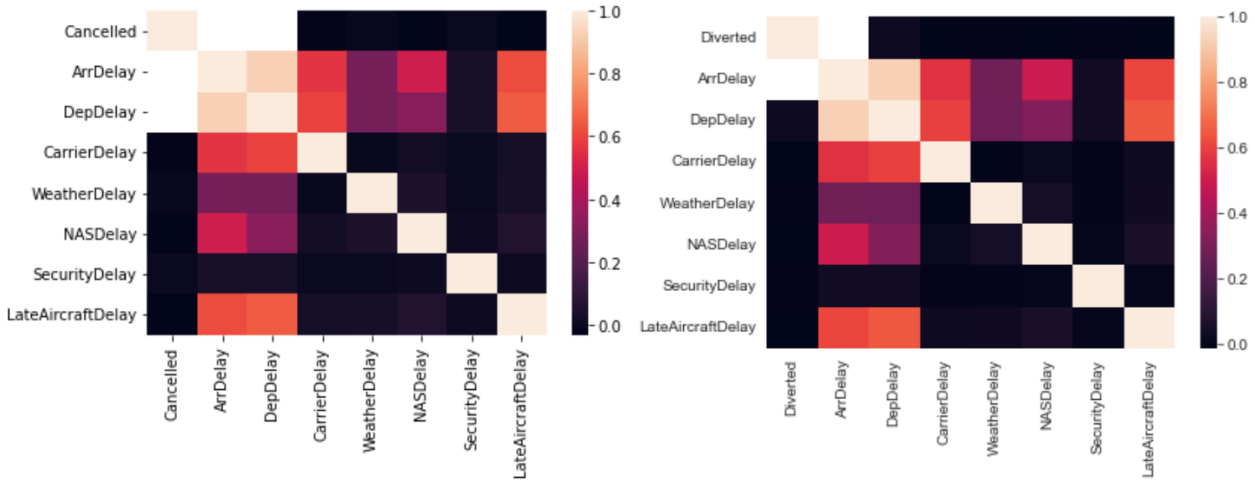
1. Introduction

In this report I am going to study American flights from year 1997 to 2008, I will mainly focus on two consecutive years (2006 and 2007) for the most part of the report. The dataset and supplementary variables are collected from the Harvard Dataverse.[1] The report include answers to question 1-5 in both R and Python. For a deeper insight I strongly recommend looking at the RMarkdown and Jupyter notebook files, this is especially true for the models in Question 5.

2. Data reading and cleaning

First, we read the data for year 2006, 2007, and 2008. We also read the supplementary variables such as airport, carrier, and airplane data. When comparing the 2006 and 2007 datasets against 2008, it is clear that 2008 only contains datapoints up to the 4th months, so to avoid any potential bias we won't use the data from that year.

There is a high amount of missing value for both the datasets, the most prominent are departure/arrival delay and cancelation/diverted related columns. We don't address the missing values by simply removing them, because we want a better insight into the variables and their connection, also I want to avoid loss of any useful datapoints. Instead we'll try to see if there is a rational explanation for some of these, i.e. if a flight is cancelled, then a departure delay entry won't be possible. Meaning that if a flight is cancelled then there can't be a departure delay registered in that specific row, nor a reason for the delay like weather, security, etc. related to departure delay. In fact, the number of missing values in the departure and arrival columns are exactly equal to the when there is a cancelation in both years. Hence, we set each dataset to only contain non-cancelled flight. There are still 16.186 missing values in the arrival delay column for 2006 and 17.179 missing values for year 2007. Again we'll try to see if there is a logical explanation, e.g. if non-cancelled flight were related to departure delay, then the remainder of the missing values in arrival delay could be related to diverted flights. Below is two heatmaps, the left describes the relationship between cancelled flights and delays, the right heatmap does the same for diverted flight and delays.



As we might have suspected, both heatmaps suggest a very strong correlation. So I only select data with non-cancelled and non-diverted flights for each year.

After the missing values are addressed, we sample the 2007 dataset to match the number of flights in 2006 (7.003802) by doing so we avoid bias towards 2007. Then we merge to datasets together, the combined dataset contains over 14 million flights. Next, we create a 'delay' column which adds departure and arrival delay together into a total delay column (minutes). We also merge the dataframe with the carrier dataset to get the full names of the carriers. The last thing we do before EDA, is creating useful date and time columns with the data entries from the month, day of month, and departure time columns.

3. EDA

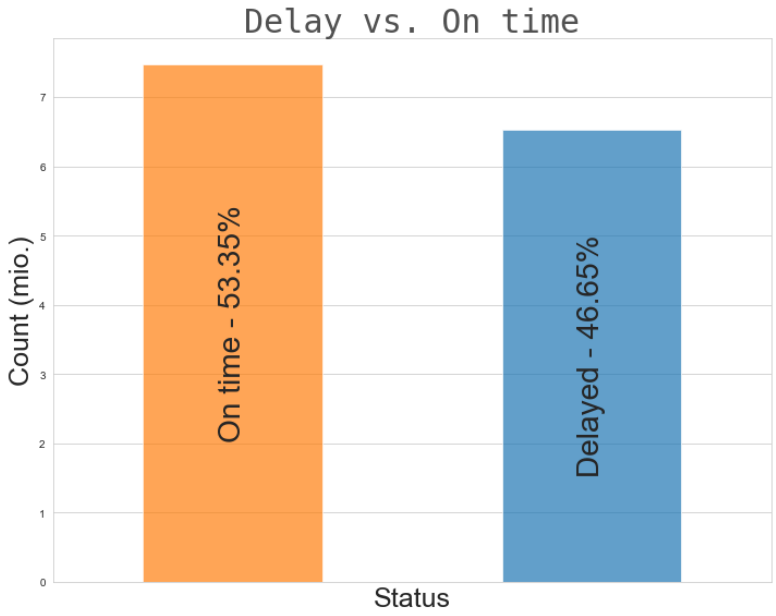
In this section we will try to gain some valuable insight into the dataset, we are especially interest in the delay and all related information hereof. Below is a table with the statistics for delay and all delay related columns. The maximum arrival and departure delay in both is around 43h, and the mean is approx. 10 minutes for both. Security delays have the lowest mean and max. Weather delays also have a low mean delay, but the max is a little higher so when they occur, they can potentially last longer than security related delays. Late aircraft and carrier delay seems to be the most unpredictable delays.

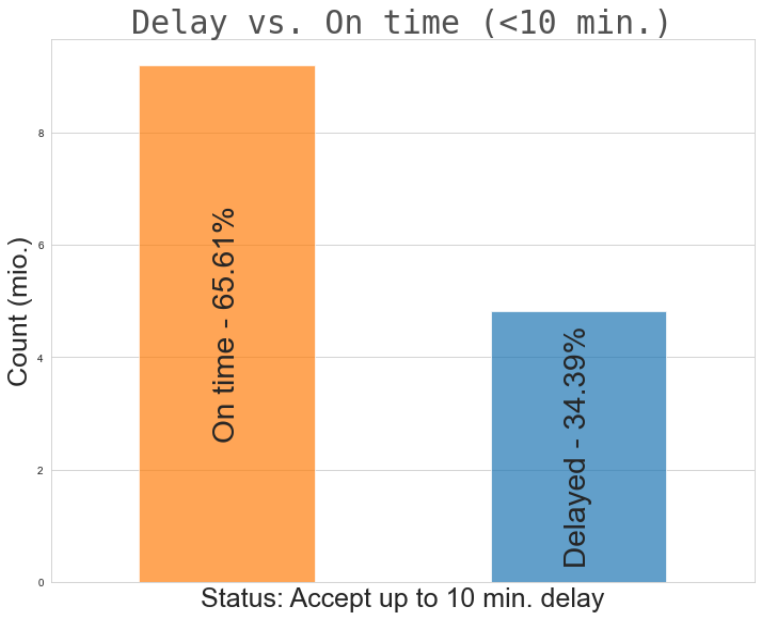
	count	mean	std	min	25%	50%	75%	max
Arrival delay	14007604.0	9.44092765615	37.9851044114	-592.0	-9.0	-1.0	13.0	2598.0
Departure Delay	14007604.0	10.7127158221	34.7757936178	-1200.0	-4.0	0.0	10.0	2601.0
Carrier Delay	14007604.0	3.7124069898	20.0757984464	0.0	0.0	0.0	0.0	2580.0
Weather Delay	14007604.0	0.741176078364	9.17072270731	0.0	0.0	0.0	0.0	1429.0
NAS Delay	14007604.0	3.76740476101	16.0751211202	0.0	0.0	0.0	0.0	1392.0
Security Delay	14007604.0	0.027943251394	1.19842947292	0.0	0.0	0.0	0.0	382.0
Late Aircraft Delay	14007604.0	4.91507034322	20.8009429358	0.0	0.0	0.0	0.0	1366.0
Total Delay	14007604.0	20.1536434782	71.3996103799	-1196.0	-12.0	-1.0	22.0	5199.0

Both arrival, departure, and carrier delay have a high max observation. in the table below I explore the data a bit further and we can see that high max is due to an extreme outlier related to the same flight which assumingly got postponed multiple times.

Carrier	TailNum	DepDelay	ArrDelay
Northwest Ai	N329NW	2601.0	2598.0

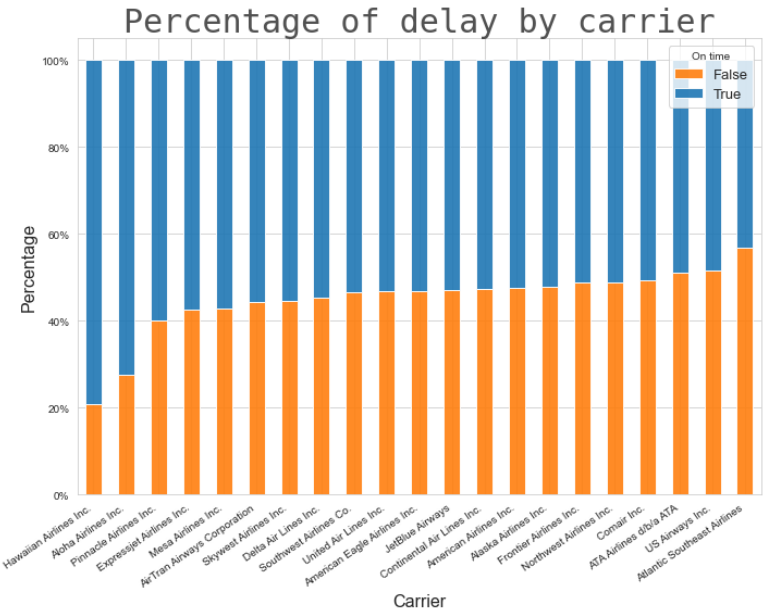
Now let's gain some knowledge about the total delays. In the plot below we are looking at the delay percentage for all flights in the set. It gives further insight on the number and percentage of delayed flights. The total number of delays is 6534776 (approx. 47%). On the plot on the right side, we assume that passengers are okay with a little delay and are willing to accept up to 10 minutes delay, then setting our threshold to max. 10 min we get 4671457 delays (approx. 34%)

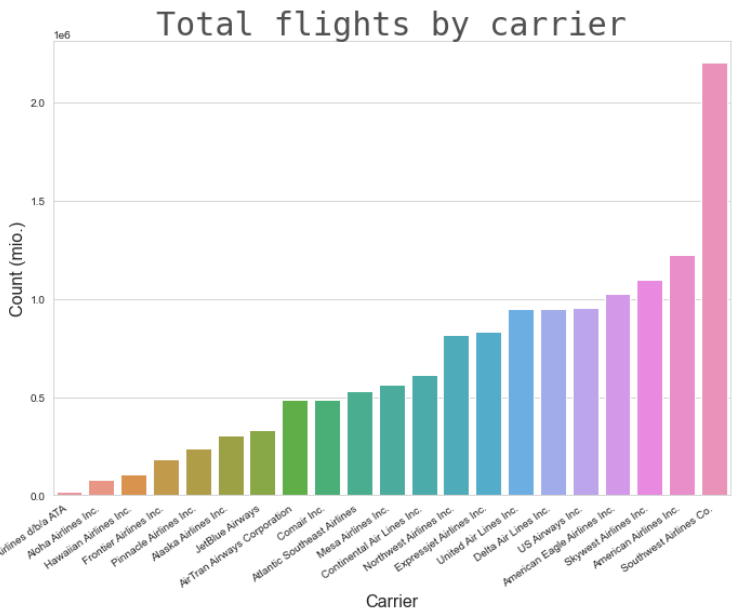




In the left plot above we can see that almost half of all the flights are actually experiencing some kind of delay. Later in this analysis we want to explore how to best avoid or minimise the risk of experiencing a delay from the passenger's perspective, so it seems logical to want to explore which carriers to trust and which to avoid at any cost. In the python notebook we also explored which airports to avoid but given that passengers probably won't travel across the country to select the airport with lowest delays, it is not that valuable here.

The left plot below shows the percentage of delays by carrier, here we can see that most carriers experience very similar percentage of delays. Atlantic Southeast stands out with the worst delay percentage, Hawaiian and Aloha Airlines have the lowest percentage of delays, but that doesn't mean we should travel with those two at all costs, for example bigger carriers would almost certainly have more flights and destinations. In the second plot where we can see that Hawaiian and Aloha Airline also have very few flights, and presumably also only travel to few locations. The carrier with the highest number of flights is Southwest Airlines with almost the double the flights of all the other big competitors.





Given the similarities in delay percentage I would recommend traveling with Atlantic, as they have the highest no. of flights and a delay percentage better or similar to its closest competitors. I would advise against traveling with Atlantic Southeast as they have the highest delay percentage despite having relatively few flights. Further delay exploration graphs can be found in the attached Python notebook.

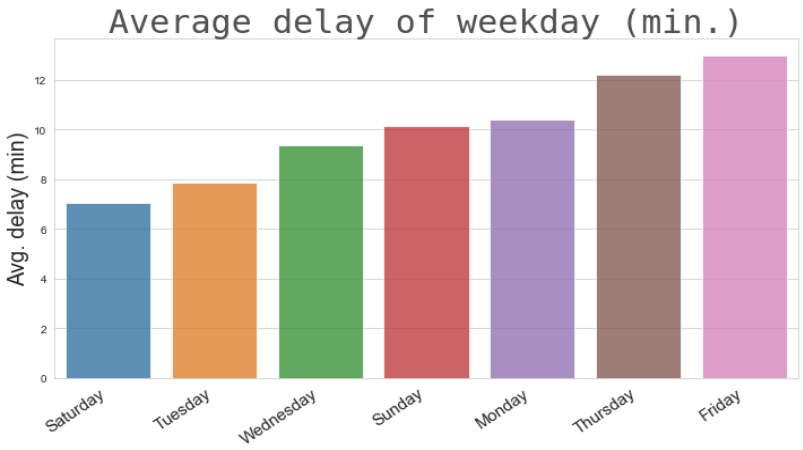
4. Questions

Q.1: When is the best time of day, day of the week, and time of year to fly to minimise delays?

1. When is the day of the week to fly to minimise delays?

Before analysing what day of the weeks to fly that minimise delays, we assume that some the passengers are interested in the average delay time per day, and that some are more interested in the total number of delays by weekday. The graph and table below show the mean delay in minutes by weekday.

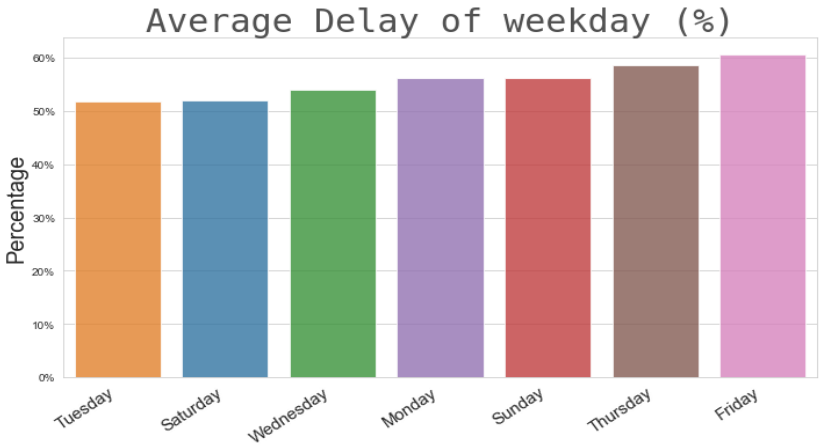
Day of Week	Avg. Departure Delay	Avg. Arrival Delay	Avg. delay	Total avg. delay
Saturday	8.609168	5.500266	7.054717	14.109433
Tuesday	8.517488	7.228652	7.873070	15.746140
Wednesday	9.756744	8.971638	9.364191	18.728382
Sunday	11.064771	9.235061	10.149916	20.299832
Monday	11.095161	9.691710	10.393436	20.786871
Thursday	12.304571	12.134668	12.219620	24.439239
Friday	13.300742	12.703370	13.002056	26.004112



Saturdays is the day of the week with lowest average delay, with around only 7 minutes. Tuesday has the second's lowest average delay at around 8 minutes. In the opposite end we find Thursdays and Fridays. Fridays has almost the double delay time on average than Saturday's (approx. 13 minutes.)

Before concluding that the passenger should always look for flights on Saturday or Tuesdays, lets have a look at the total number of flights, delays, and the percentage of the flights that are delayed by weekday. The table and graphs below give us a valuable insight in that regard, again Saturday stands out as the day with fewest total delays and second fewest percentage of flights that are delayed. Tuesdays has the lowest percentage of delayed flights and second lowest number of delays. In the high end, again we see Thursday and Fridays were around 60% of all flights are delayed.

Day of Week	Flights	Flights(%)	Delays	Delays (%)
Tuesday	2023777	14.447703	1047246	0.517471
Saturday	1767905	12.621038	918094	0.519312
Wednesday	2042684	14.582680	1104987	0.540949
Monday	2075083	14.813975	1164797	0.561325
Sunday	1972786	14.083679	1107784	0.561533
Thursday	2058307	14.694212	1206897	0.586354
Friday	2067062	14.756714	1255420	0.607345

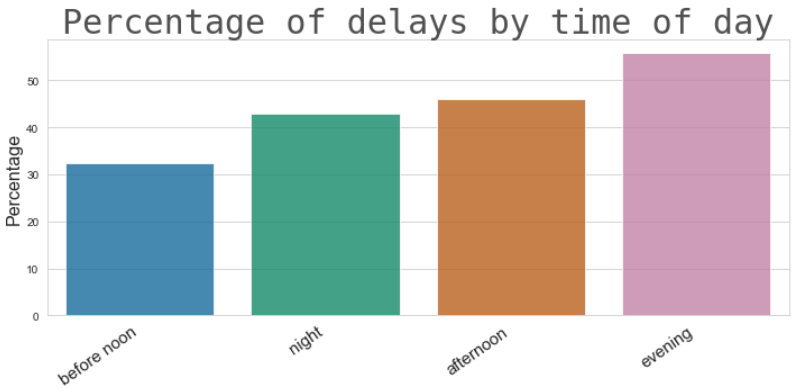


So, we can conclude that to minimise delays, the best day to travel is either Tuesday or Saturdays. Both days have similarly percentage of delay occurrences. If a delay is to occur, Saturdays offer slightly lower avg. delay time. But, on Tuesdays there is perhaps more personal on duty to handle unexpected delays due to the higher number of total flights. Chosen one of the two days to conclude as the best would be up to the passenger's preference. I would choose the day with the lowest ticket price (in all fairness that could also be difficult, as both Tuesday and Saturday are considered among the cheapest days to fly[2], but that a question for another report.)

2. When is the best time of day to minimise delays?

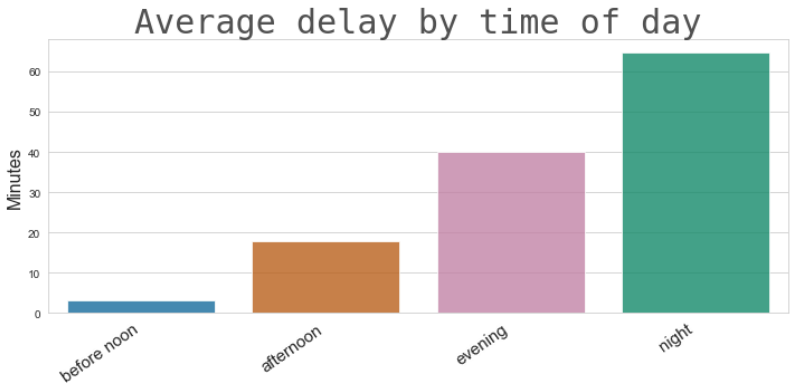
Before answering this question, we make the assumption that most people have a preferred time of day instead of one specific hour for their travel plans. Therefore, dividing the day into 6-hour bins: night, before noon, afternoon, and evening. In the first table and graph below we can see that the before noon suffer the lowest percentage of delays, only 32% delayed flight which is significant below the total we saw in the EDA section and much lower than delays occurring in the evenings where 55% of all flights are delayed. We can also see that there are very few flights in the night-time so potential less queue, and it also has the second lowest delay percentage — so seemingly traveling at night also seems like a very good choice. But lets also explore the mean delay in minutes before making any conclusions.

Time of day	Flights	Delays	Delayed (%)
before noon	9247472	2985532	32.284845
night	706817	303498	42.938696
afternoon	10442055	4813189	46.094270
evening	7611520	4248785	55.820454



Below on the left is a table with departure, arrival, and total delay in minutes on average and on the right, we have a graph of average delay in minutes by time of the day. Flights before noon have the lowest departure, arrival, and total average delay in minutes. Flights in the night suffer the most delay arrival and total delay in minutes.

Time of day	Avg. Departure Delay	Avg. Arrival Delay	Total avg. Delays
before noon	3.224788	-0.028797	3.195991
afternoon	11.101738	6.702518	17.804255
evening	23.040655	17.104592	40.145247
night	10.594667	54.089681	64.684348



The answer to the question depends on how you define the best way to minimise delays, if the passenger is interested in the lowest average minutes, then traveling before noon is best. If the passenger is more risk averse, then traveling at night is best, but if a delay occurs at night, then the probability for it to be longer than before noon is very high. So in the end it depends on how risk adverse the passenger is – I would choose to travel before noon.

3. When is the best time of year to fly to minimise delays?

For this question, I have chosen to divide the time of year into months initially, because dividing into seasons (spring, summer, autumn, winter) can possibly introduce some bias, e.g. if January and February have a low number of delays, it could mask the busy Christmas holiday season of December, and similarly for May and the summer holidays.

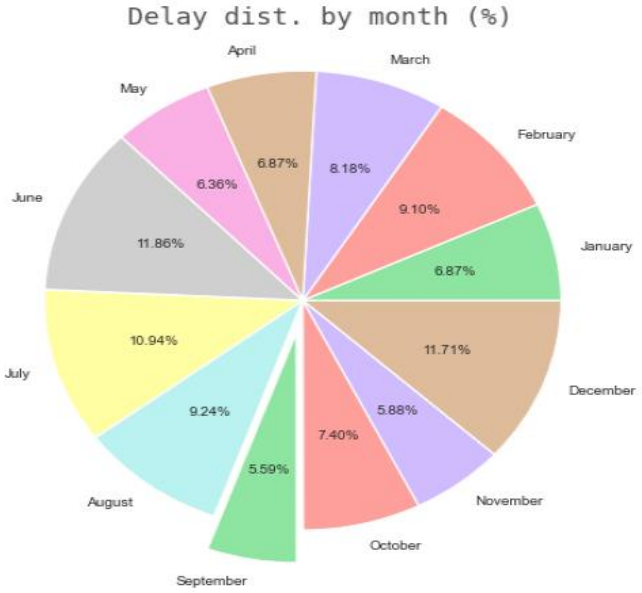
Month	Departure delay	Arrival Delay	Total Avg. Delay
September	7.440537	6.080626	13.521163
November	8.179276	6.032137	14.211414
May	8.415244	6.971385	15.386629
January	9.168771	7.420203	16.588974
April	9.116255	7.474179	16.590434
October	9.261392	8.592198	17.853589
March	10.759757	9.006834	19.766591
February	11.553899	10.447546	22.001445
August	11.699477	10.638009	22.337486
July	13.720379	12.711885	26.432265
December	14.619105	13.664733	28.283838
June	14.509099	14.112587	28.621686

Month	Flights	Flights(%)	Delays	Delayed (%)
September	1144554	8.170948	565474	49.405620
May	1194869	8.530145	623761	52.203296
November	1150995	8.216930	603368	52.421427
January	1152288	8.226161	608992	52.850676
April	1157432	8.262884	615026	53.137117
October	1197728	8.550556	658692	54.995124
March	1194570	8.528011	665851	55.739806
August	1232369	8.797857	701480	56.921263
February	1037214	7.404650	597747	57.630055
July	1217922	8.694720	729458	59.893655
June	1173535	8.377842	713324	60.784212
December	1154128	8.239296	722052	62.562558

In the tables above, we can see what we might have suspected: the Christmas and summer holiday months are experiencing the highest average delay time (min.) and also the highest percentage of delayed flights. In the piechart below, we can see the percentage distribution of delay by months. The months with fewest average delay (min.) and percentage of delay is September followed by November and May.

Season	Flights	Flights(%)	Delays	Delayed (%)
Autumn	3493277	24.938433	1827534	52.274057
Spring	3546871	25.321040	1904638	53.693406
Summer	3623826	25.870420	2144262	59.199710
Winter	3343630	23.870107	1928791	57.681096

Season	Dep. delay (min)	Arr. delay (min)	Total Avg. Delay
Autumn	24.881204	20.704961	45.586166
Spring	28.291256	23.452398	51.743654
Summer	39.928955	37.462481	77.391436
Winter	35.341775	31.532482	66.874257



If the passenger instead wants to know the best season of the year that minimises delays, we can divide the year into seasons. From the tables above to the left of the pie chart, we can see that the best seasonal time of the year to travel is autumn which has a total avg. delay of approx. 46 min and the worst season is summer with around 77 minutes total delay on average.

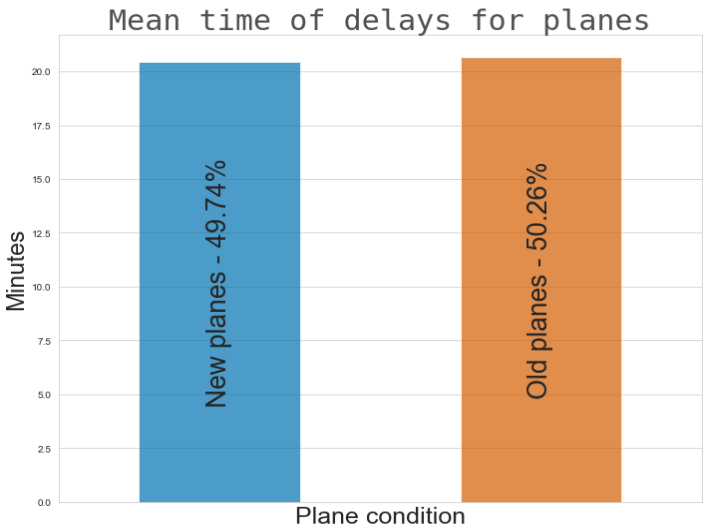
Q.2: Do older planes suffer more delays?

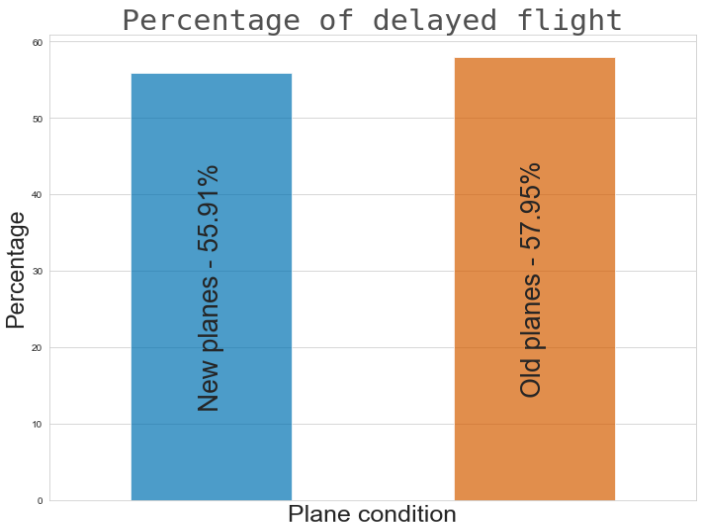
First, we must define what an old plane is: We make the assumption that a plane manufactured over 20 years ago is considered old.[3] The latest datapoints in our set is from 2007, so we define an old plane to be before 1987. From the table below we can

see that old and new planes suffer similar delay time in minutes on average, new planes only have an 0.84% less average minutes delay in comparison. We can also see that older planes have slightly higher percentage of flights that are delayed by 2%.

Plane Condition	Flights	Flights(%)	Delays	Delayed (%)	Avg. Departure Delay	Avg. Arrival Delay	Total avg. delay
new	10100202	89.122774	5646938	55.909159	10.792953	9.655673	20.448626
old	1232894	10.878885	714373	57.942775	10.501595	10.139588	20.641182

Graph showing the comparisons for old vs. new planes:





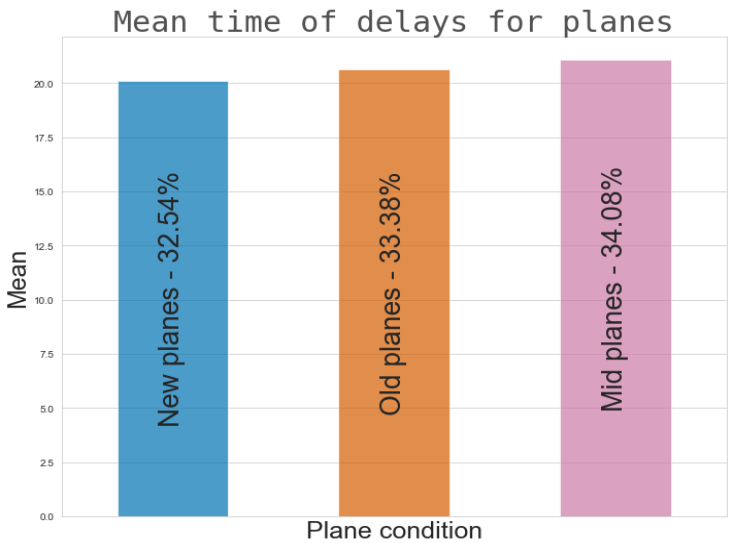
The difference seems small that it would presumably be reasonable to say older planes do not suffer significantly more delay time. To conclude that old planes to not suffer more delay with confidence I did $\chi 2$ test to decide if there if a relationship exist between the manufacturing year and delays (test for independence): H_0: Old planes does not suffer more delay vs. Old planes suffer more delay, we rejected the null hypothesis at 5% and 10% significance level, and conclude that there is very strong evidence suggesting that old planes do not suffer significantly more delay[3].

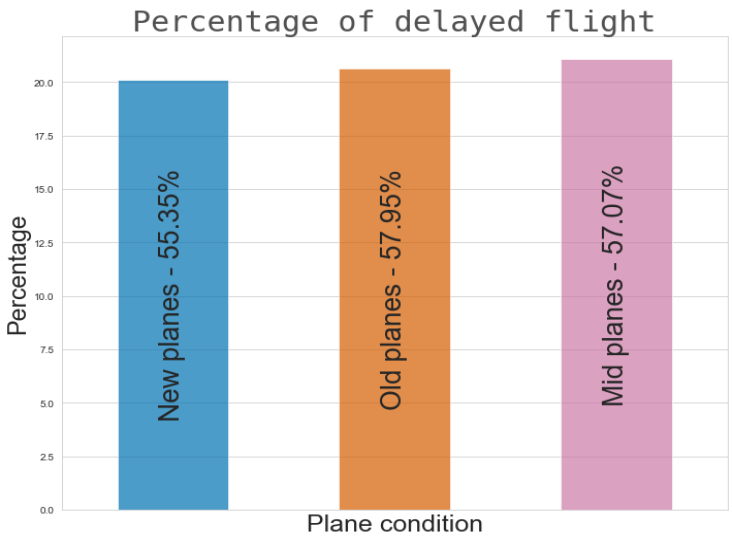
We therefore conclude that old planes no not suffer any significantly extended delays over newer planes.

I did a fascinating observation when changing the threshold my definition of an old plane.

Plane Condition	Flights	Flights(%)	Delays	Delayed (%)	Avg. Departure Delay	Avg. Arrival Delay	Total avg. delay
1956-1987(old)	1232894	10.879007	714373	57.942775	10.501595	10.139588	20.641182
1987-1997(mid)	3278748	28.931539	1871475	57.078952	11.059478	10.027815	21.087293
1997-2007(new)	6821454	60.192233	3775463	55.346895	10.664847	9.476803	20.141650

Graph showing the comparisons for old, mid, and new airplanes :



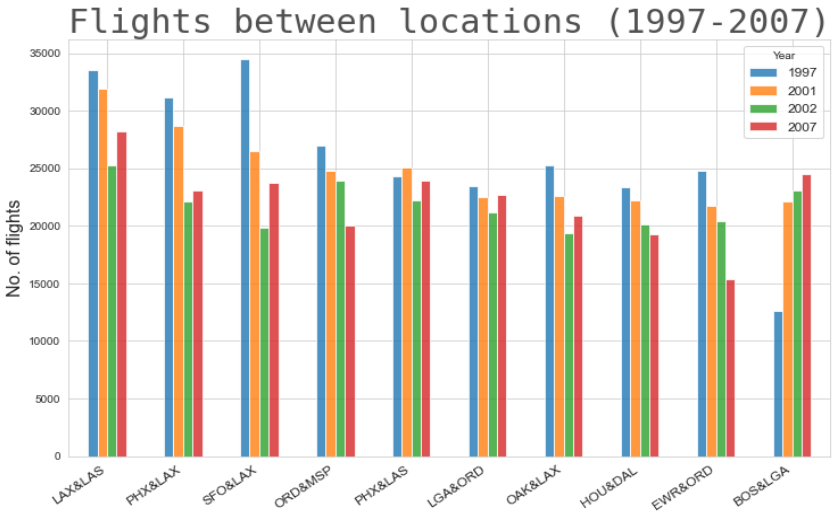


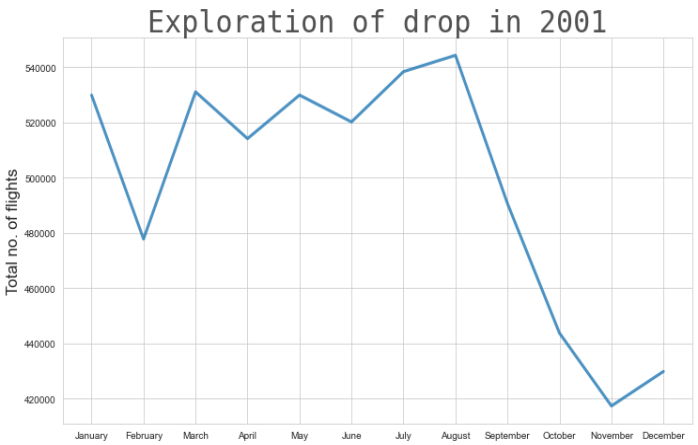
Airplanes tend to suffer the fewest mean delay in minutes when new, then increase around the 10-year mark and decrease again when passing the 20-year mark.

Q.3: How does the number of people flying between different locations change over time?

I have chosen to extend the timeframe for this question to 11 years (1997-2007), by doing so we gain a more valuable insight in the changes over time. I have also decided to focus on the 20 busiest destination connections (in- and outbound flights) to reflect the population best possible. Looking at my table and graph for total flights between locations below, we see that the number of people flying in the last 11 year have fallen for 9 out of the 10 connections. Opposite of all other connections, flights between Boston Logan and LaGuardia airport in New York has actually increased during the period. Flights from San Francisco to Los Angeles International Airport has decreased most, the lowest drop is between Phoenix and Las Vegas. Generally speaking there has been a drop in flights in the period. As the available variables don't offer any insight on the people onboard of a given flight, we assume a fixed average number of passengers.

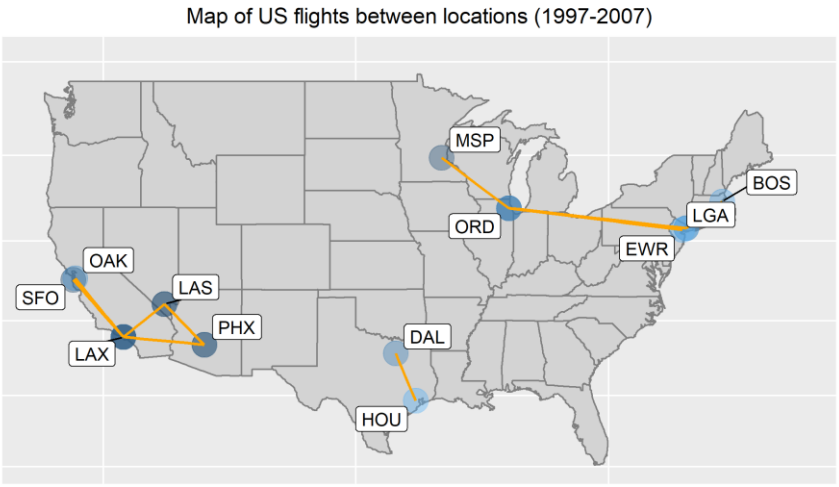
Route	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
LAX&LAS	33527	32605	30311	34644	31953	25259	22439	24981	25055	26669	28200
PHX&LAX	31120	28765	30775	33693	28664	22093	21016	20891	21639	23429	23059
SFO&LAX	34479	32058	32376	31988	26504	19879	17425	18240	16930	19897	23751
ORD&MSP	27006	26445	26497	25972	24780	23914	24631	23671	19541	20012	20061
PHX&LAS	24300	24576	25014	26252	25086	22234	20564	21571	21569	23294	23959
LGA&ORD	23449	21490	22401	22407	22474	21205	21775	22573	22944	22944	22670
OAK&LAX	25311	22335	21748	23805	22624	19398	19177	18489	19311	19001	20857
HOU&DAL	23393	22967	22902	22977	22232	20119	19794	19825	19071	19091	19289
EWR&ORD	24812	25654	25398	24010	21789	20404	18752	18444	16128	15705	15364
BOS&LGA	12612	11082	11302	17299	22172	23076	25230	26535	26347	24768	24485

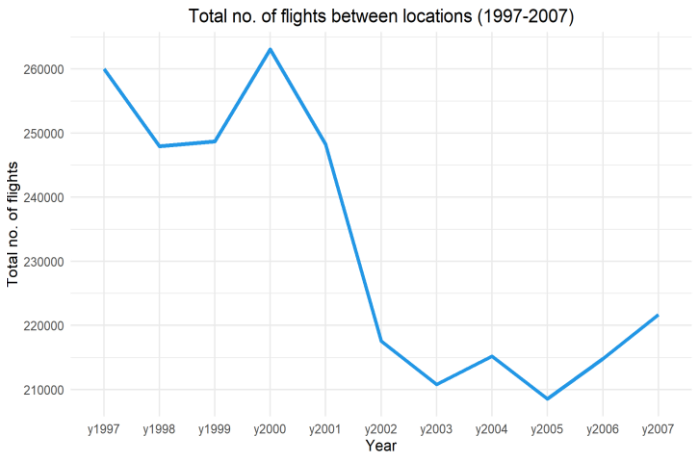




Furthermore, there seems to be a noticeable drop around year 2001 that affect flights going forwards. In the graph below, we can see that the drop is very significant and occurring around the 9/11 attack.

The map below is a graph of the connection between locations over time, the airport points darkness indicates no. of flights from destination, the darker the more flights. The graph to the right of the map, is a line plot of all flights over the 11 years, here we can see the rising trend after 2005.





We can conclude that number of people flying between different locations changes a lot over time, in 9 out of 10 connections they fall over time, which is especially true due to the 9/11 terror attack and the impact it has on the industry ever since. Most connections (7 out of 10) have experienced an increase in people flying between location again after 9/11. The Boston and New York connection saw a high increase in flights from 1997 to 2001 but has since normalised again.

Q.4: Can you detect cascading failures as delays in one airport create delays in others?

Because the question is somewhat ambiguous, I have focused on answering on the behalf of an airline carrier focusing on three different potential scenarios. In section 1. We look at a single airplane that create cascading delay failures for itself.

Section 2. We identify cascading delay failures of directly following flights on the same route from Las Vegas (A) to Los Angeles (B) and eventually ending in LAS.

And finally, section 3. Here I identify delays in Airport A (LAS) that creates further delay in airport B (LAX) and then for airport C (All other airports).

1. We can identify 568 cascading delay failures for the airplane with tail number 'N485HA' and 487 for airplane with tail number 'N477HA'. The failure involves the plane has a departure delay (airport A), then also a departure delay (airport B) and in the end an arrival delay for the planes next destination (airport C)

TailNum	Flights	Cascading Failures	Casc. Failure (%)
N485HA	7893	568	7.196250
N477HA	7696	485	6.301975

For the plane with tail number N485HA: A delay in airport A leads to approx. 60%, if there a delay in both airport A & B occur then around 71% of the flights experience cascading delay failures. For the plane with tail number N477HA, there is a similar pattern if both airports A & B are delayed, although a little lower. There is approx. occurring a cascading failure in 7% of all flights for airplane N485HA and in 6% for the plane with tail number N477HA.

2. For this section we look at cascading delay failures for the same route. The data is collected by sampling, so small differences is expected between runs. The flight is approx. one hour, so we are only interested in delays in airport B occurring around an hour later. We can detect cascading delay failures for the same route if a flight has departure delay in Las Vegas (airport A), that also creates departure delay in Los Angeles (airport B), which in return results in a arrival delay in Las Vegas again (airport C):

Airport A	Airport B	Airport C	Flights	Cascading failures	Casc. Failure (%)
LAS	LAX	LAS	26033	3685	14.155111

On the route Los Angeles to Las Vegas, there is a total of 26033 flights, we can detect 3685 cascading delay failures. There is occurring a cascading delay failure in 14.16% of the flights on the route.

3. Let's also have a look at how a delay between Las Vegas and Los Angeles spirals to delay in other airports from thereon: A departure delay in the origin airport of Las Vegas (airport A), creates a departure delay in destination airport of Los Angeles (airport B), which then finally causes an arrival delay in the next airport (airport C)

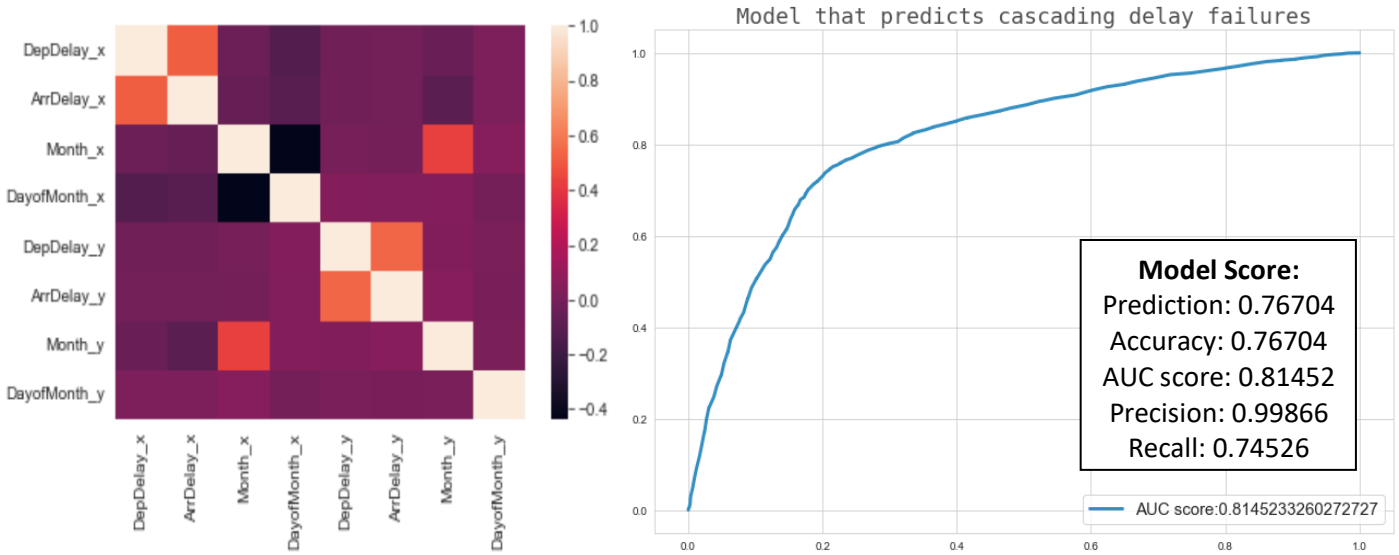
Airport A	Airport B	Airport C	Flights	Cascading failures	Casc. Failure (%)
LAS	LAX	All other airports	455728	12903	2.831294

From the table, we can see that we can detect 12903 cascading delay failures. A delay in Las Vegas (airport A) create cascading delays for approx. 2.83% of all flights.

Alternative approach for section 2 and 3: If we assumed the time range be 3 hours instead of account for the flight time from Las Vegas to Los Angeles, the no. of cascading delay failures detected would have been higher. But due to the fact that arrival delay for airport C is also specified in the same row of airport B (column: 'ArrDelay_x') in the merged dataset, I believe it to be more precise focusing on the flight duration instead.

We see a lot higher cascading failure percentage for the same plane and on the same route, this might be because the same planes, staff, gate, and terminals are used on the same route which have a higher impact on following flights. Hence, if a Las Vegas flight creates a delay in the Los Angeles gate, it might have a low impact on delays in a completely different terminal and gate of the airport (i.e. a New York flight in the opposite end of airport and with entirely different carrier, plane, staff, etc.)

In the Python notebook I also built a model that predicts cascading delay failures. The hyperparameter is present in the correlation map.



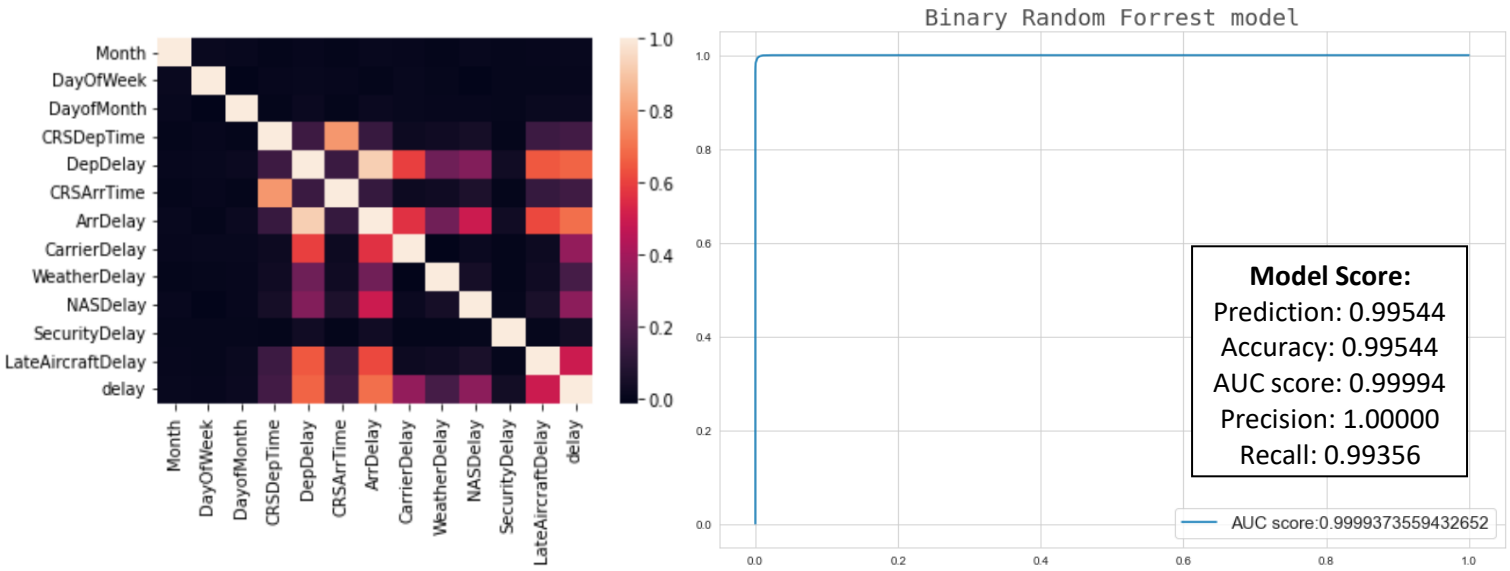
Q.5: Use the available variables to construct a model that predicts delays.

I built a Gradient Boost, Random Forest, and a Decision Tree classifier. All three models do a good job at predicting delays, both Random Forest and Decision Tree classifiers do exceptionally well. Here I focus on the Random Forrest models (the other models can be found in the python notebook.) The models are built on the year 2006 and 2007 data.

For the multiclass models: I created a total delay column ('delay') which is also the target column; where 2: >45 min. delay, 1: < 45 min. delay, and 0: on time or arriving before.

For the binary models: the total delay column ('delay') was set to 1: delayed or 0: no delay.

Due to computational limitations I took a sample of 250.000 from combined dataset. Indicators are created from the origin and destination columns and the data was split into a training and testing set (80/20%)

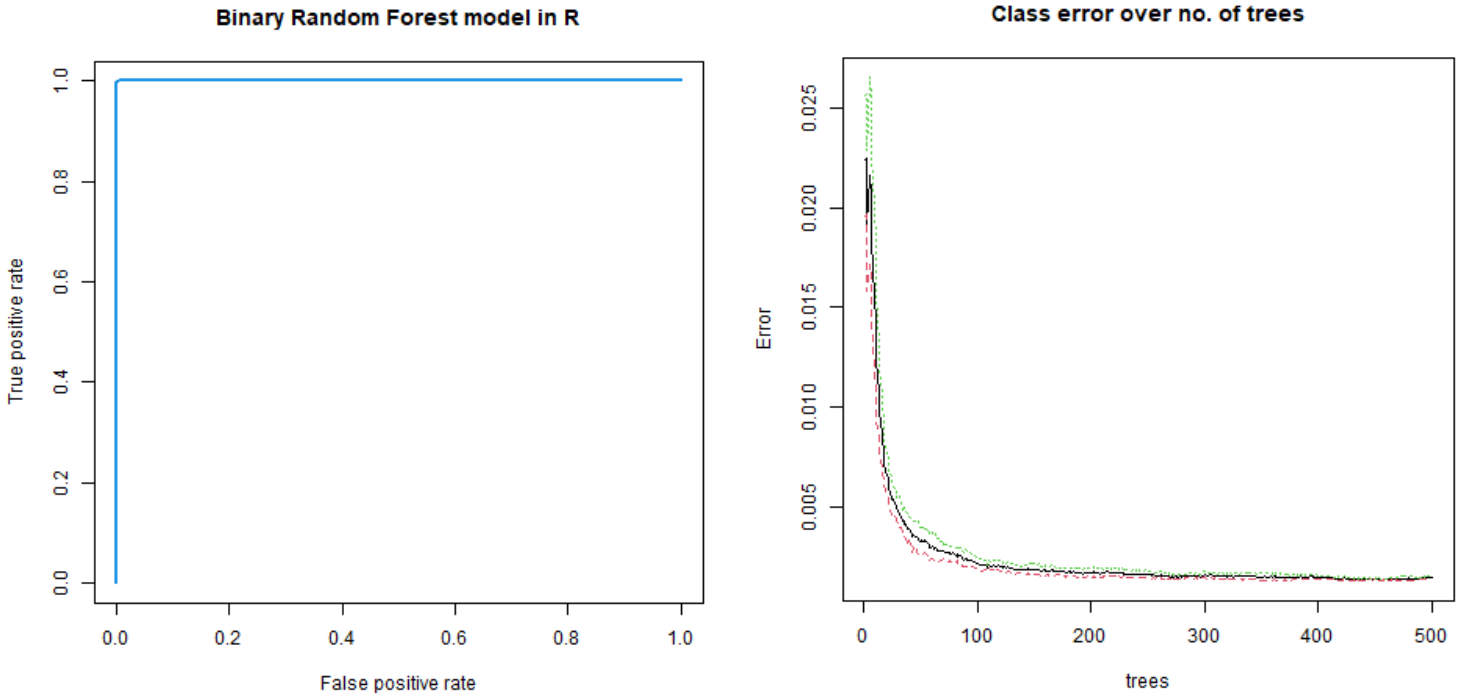


The correlation map shows the hyperparameter used to build the models. The model that performs best is a Random Forest model with multiclass classification. The multiclass is split into 3 classes: 0: on time or before, 1: less than 45 minutes late, and 2: more than 45 minutes. I also built a Random Forest model with binary classification that performed well.

Multiclass Random Forrest Score:
Prediction: 0.99606
Accuracy: 0.99606
AUC score: 0.99746
Precision: 1.00000
Recall: 0.99606

Multiclass Decision Tree Score:
Prediction: 0.99940
Accuracy: 0.99940
AUC score: 0.98413
Precision: 1.00000
Recall: 0.99940

I also built a Binary Random Forest model in R: the model performs similarly, with an accuracy of 0.99898 and a OOB estimate of error rate of 0.12%. Below I have plotted the performance of the model.



To get a better insight on how the results is derived in general, I encourage to take a look at the included notebook files. All answers in R and Python are very similar and the small difference seems to be due to data sampling.

5. References

Image credit: https://www.shutterstock.com/image-vector/flight-delay-cancel-concept-vector-flat-1498542281

1.Data: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

2: Price reference for best day to fly conclusion: https://www.farecompare.com/travel-advice/tips-from-air-travel-insiders/

3: Airplane age reference: https://www.paramountbusinessjets.com/faq/age-of-aircraft-safety-factor.html

4: Guide used to perform a Chi-squared test: https://stackoverflow.com/questions/64669448/understanding-scipy-stats-chisquare