ST2195 Programming for Data Science

Name: Thilini Madushani Herath

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

Approach: The problem is asking to find the best "time" to fly to minimize delay (Arrival delay) in three-time cohorts. To decide the best time of the day, four sections (Morning 5:00 AM-12:00PM, Afternoon 12:00 PM-5:00 PM, Evening 5:00 PM-9:00 PM, and Night 9:00 PM-5:00 AM) have been considered. For the day of the week, the obvious Monday-Sunday has been considered. However, the problem has not specified the time cohort for the year so monthly analysis has been considered for "year". I first focus my analysis on the largest cohort of time which is "months" then narrow the focus to weeks and finally time of day.

Data preparation: My plan is the analyse data yearly so, I do not create a database consisting of all the data at this step but for question 2. For each year's flight data frame, first I removed canceled and diverted flight observations. And for each of three cases, I considered relevant time frequency at a time (Month, Day of week, Time).

There are two critical facts to consider when comparing delays. One is the percentage of flights that get delayed in each session and the second average delay time. Hence two similar analysis has been conducted focusing on those. To check the consistency of the results, 5 years (2003-2007) delay data have been considered for the analysis.

Solution: Monthly: There is a periodic pattern in the percentage of flights and delay time over months. And this pattern occurs in all 5 years 2003-2005. Hence based on the percentage and average delay time April or September are ideal months to minimize delays.

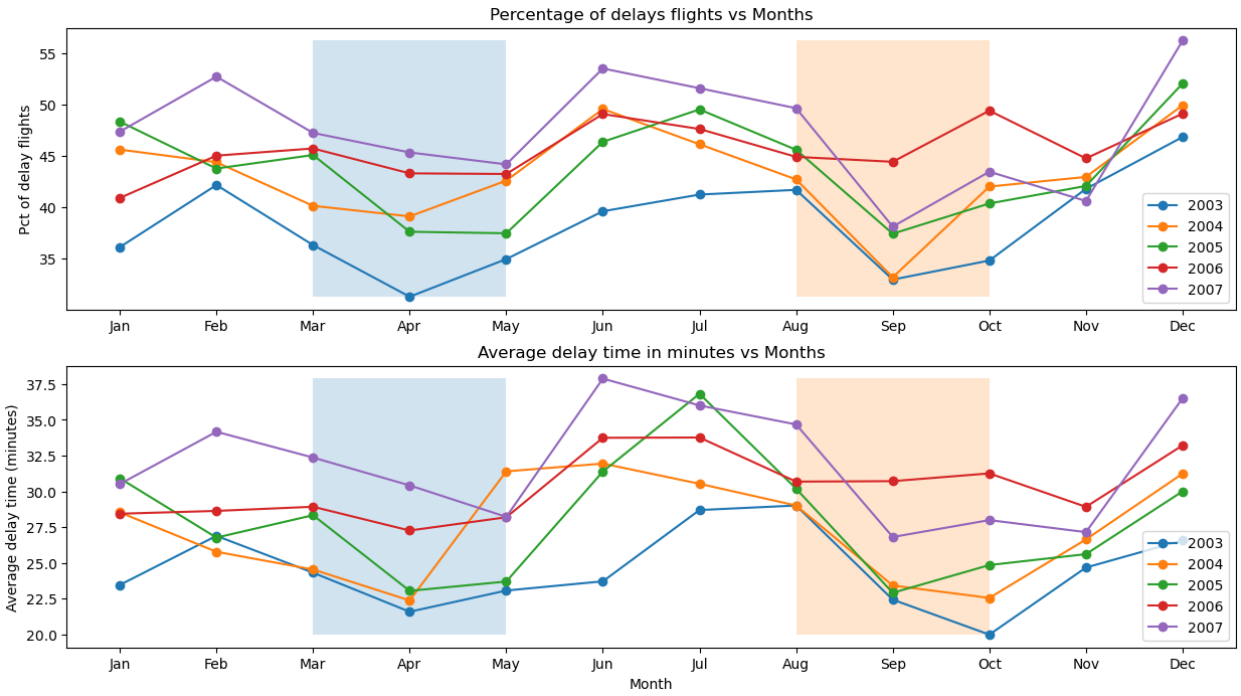


Figure 1.

Day of week: This analysis suggests that Saturday is ideal for flying. And the result is consistent for all five years.

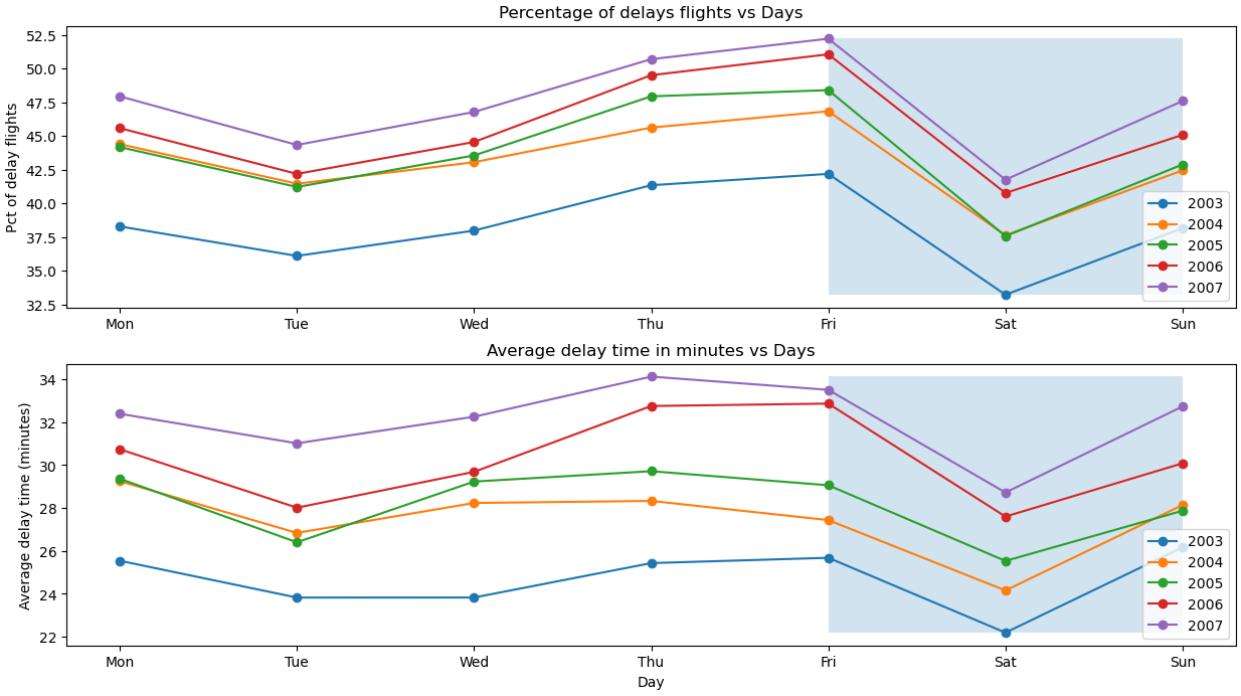


Figure 2.

Day of time: This analysis suggests that the Morning (5:00 AM-12:00 PM) is ideal for flying. And the percentage and average delay time increase over the rest of the day. The result is consistent for all five years.

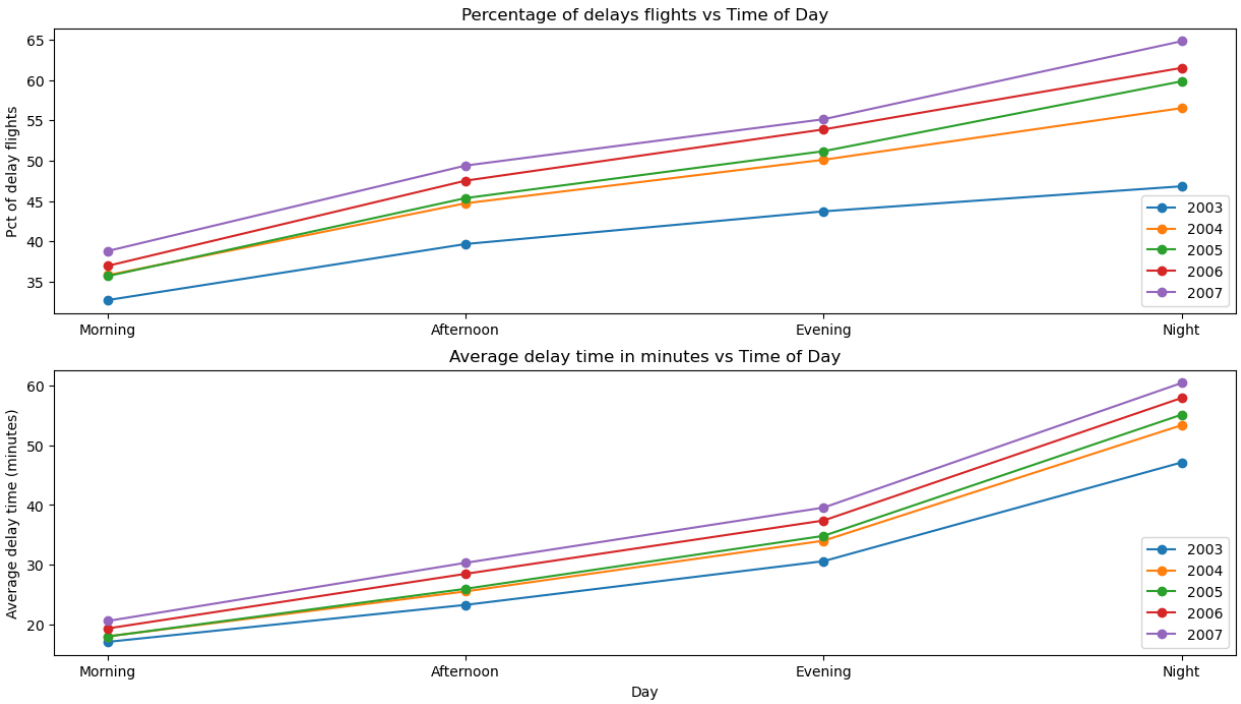


Figure 3.

2. Do older planes suffer more delays?

Approach: Two years (2006-2007) worth of data through the database which was created based on flight and plane data have been considered for this analysis. First, “issue_date” of planes and flight dates were used to compute the age in years of planes at the time of each journey. The maximum age of a plane is found to be 32 years approximately. The histogram below shows the count of flights with ages. Older planes make fewer journeys compared to new planes. The approximate threshold to classify new from old planes is between 10 and 32. The 75th percentile of age is 9.1 years hence it is considered to be the Threshold. Complete age percentiles are as follows.

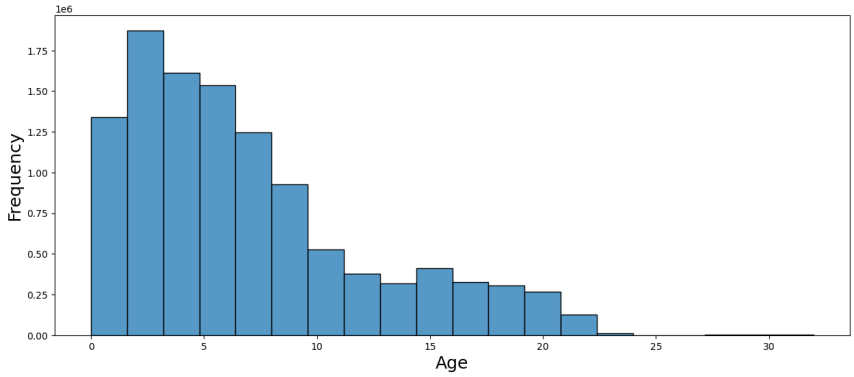


Figure 4.

Percentile	Age
25%	2.861054e+00
50%	5.579740e+00
75%	9.308693e+00
Max	3.197536e+01

Data preparation: To prepare data in order to answer this question, I needed to create a database that consists of both flights and plane information. I joined these two tables with the unique identifier Tailnum and consider two years (2006 and 2007) worth of data. After removing all the null values, I created a new feature, the Age of planes at the time of the flight.
Solution: No! With the decided threshold of 9.1 years. The percentage of delayed journeys made by new planes is 47.02% and for old planes, it is 45.80%. There is no significant difference between the average delay time between new and old planes.
However, this does not infer the population of planes. A careful statistical analysis (t-test) is needed to make a final conclusion.

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

Approach: The question has not specified what it refers to as locations. The obvious approach to the problem is to consider flights between pairs of airports and check if there is any trend that exists over time. The challenge in approach is that there are nearly 1200 pairs of airports. To narrow the scope of the analysis, the 20 pairs of airports with the highest number of journeys between them are considered for the analysis. Further to check the trend (if exists) five years (2003-2007) worth of data is considered.

Data preparation: For each year's flight data frame, first I removed canceled and diverted flight observations and considered only the “Origin” and “Dest” columns which contain airports departed and arrived and this information has been used to compute the number of journeys between them. I included a new column called “comb” by combining Origin and Destinations and created two functions (unicom2 and existing_unicom) that compute variations of each airport pair and extract all existing pairs from data frames.

Solution: Figure 5 represents the distribution of the number of journeys among all pairs of airports separated by years. Just by looking at this figure, we can see that there is a slight increment in total flights in 2007 compared to 2006 and so on. Figure 8 can be used to investigate this further. In Figure 8, the x-axis (pairs) have been sorted by the number of flights in 2003 between the top 20 pairs. Clearly, there are some increasing trends in some pairs (JFK_HPN, LAS_COS) and decreasing trends in some (CLE_IAN, PHL_MCN), and inconsistency among others. My observation is that there does not an exact trend between all airports.

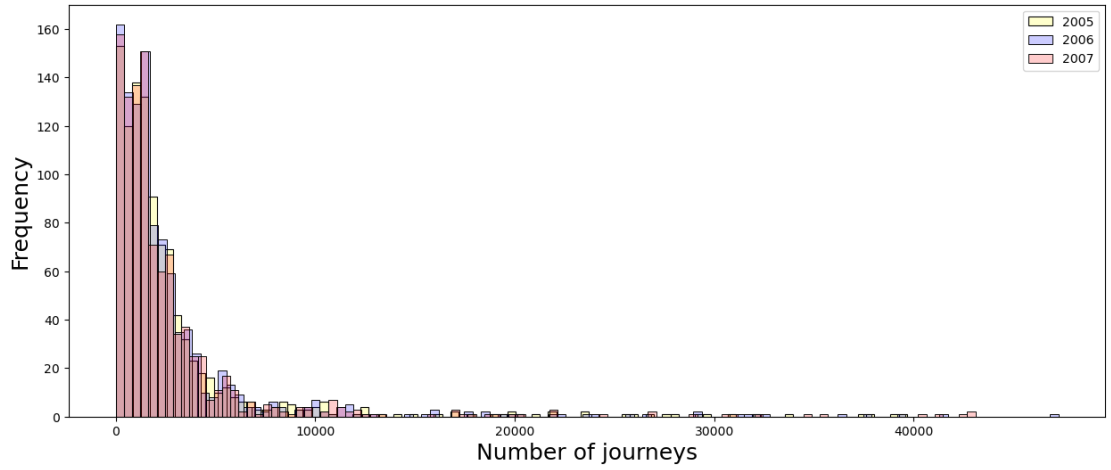


Figure 5.

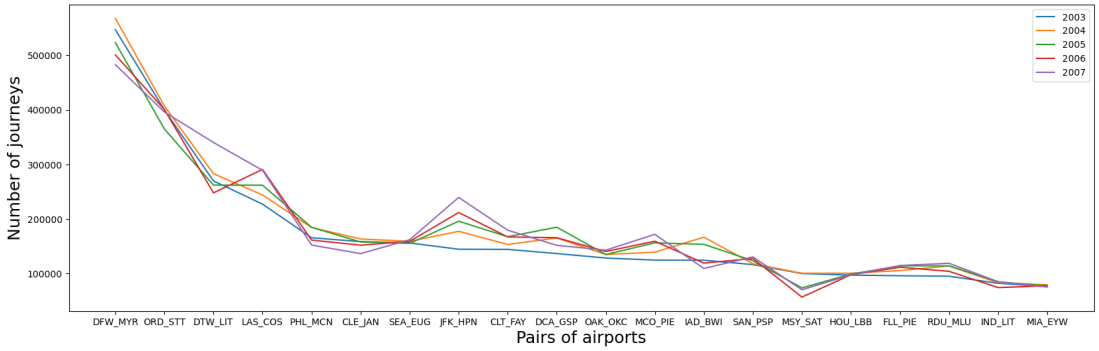


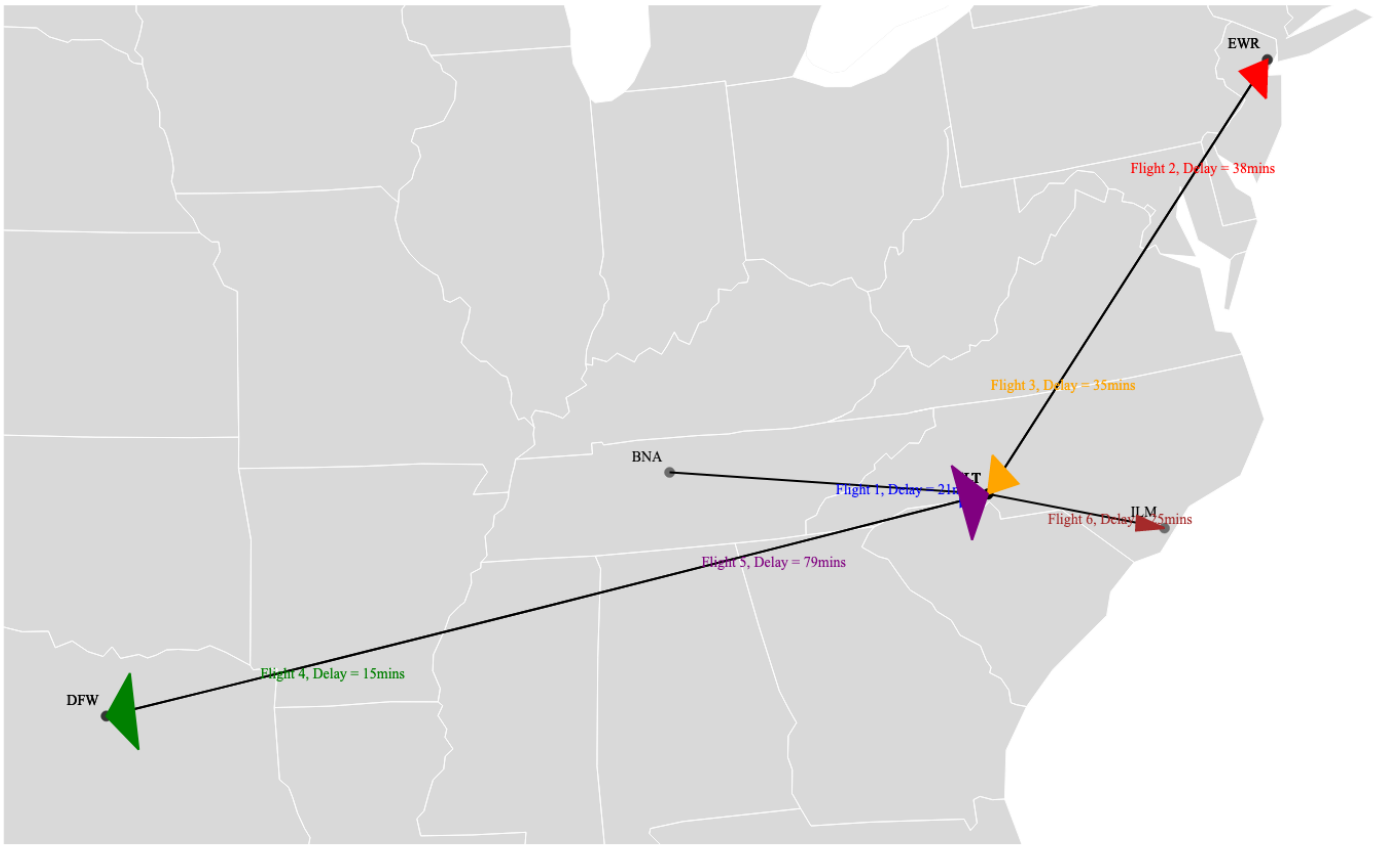
Figure 6.

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

Approach: There are multiple approaches to solving this problem. My approach is to detect the cascading failures due to the delay (Arrival delay) of a given plane. In order to detect this series of delays, the first delay of the given plane should NOT be due to a delay of another plane. As there are categories of delays caused in the flight data frames, I look for the first delay of a given flight (sorted by day and time). And check if this delay of this plane follows more delays in its subsequent journeys.

Data preparation: Again I got the help of the database to select data joined with flights and planes in the year 2006. Created a new column “fly_date” which is the DateTime data type of the flight date. And selected observations only with some delays based on the category of delays. Selected one tailnumber (a plane) and ordered by date, time, and minutes of delay. Check if the first delay is NOT due to “LateAircraftDelay”, if this is true then use this data to detect the series of delays. A picture of the prepared data for plane tailnumber, N514AU is shown below.

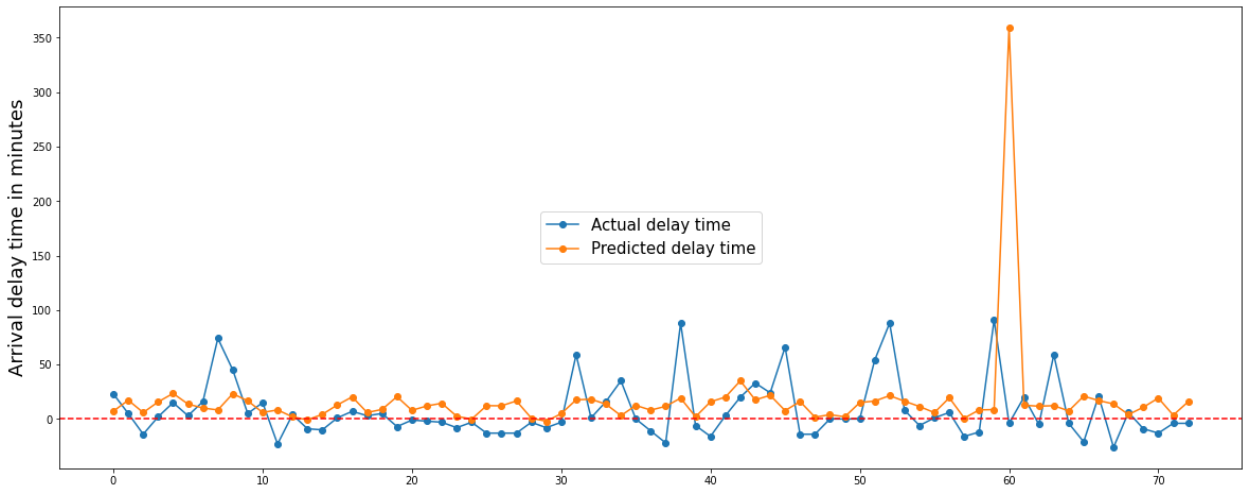
CRSArrTime	CRSDepTime	TailNum	ArrDelay	DepDelay	Origin	Dest	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	fly_date
835	620	N514AU	21	-8	BNA	CLT	0	0	21	0	0	2006-08-06
1111	920	N514AU	38	29	CLT	EWR	13	0	9	0	16	2006-08-06
1329	1145	N514AU	35	45	EWR	CLT	9	0	0	0	26	2006-08-06
1558	1420	N514AU	15	37	CLT	DFW	5	0	0	0	10	2006-08-06
2010	1640	N514AU	79	14	DFW	CLT	11	0	65	0	3	2006-08-06
2226	2130	N514AU	25	33	CLT	ILM	0	0	0	0	25	2006-08-06

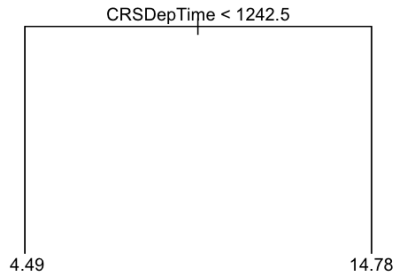


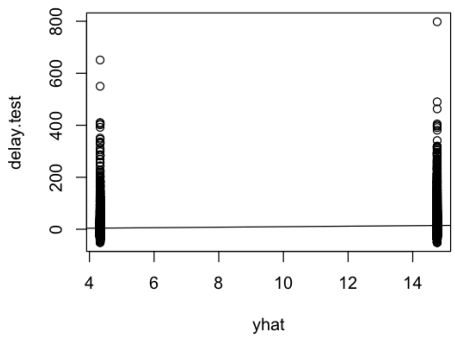
This shows that Origins match with destinations. And I used the library Plotly to illustrate this scenario.
Solution: Flight 1 from BNA to CLT made a delay of 21 minutes due to NAS delay. The next flight (Flight 2 in the Figure) from CLT to EWR made a 38 minutes delay which consists of 16 minutes of Late Aircraft Delay and this series of delays propagates throughout the journeys. The full figure included is in the solution files (delay_series.html).

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

Approach: For this problem, my plan is to use the ML model, GradientBoostingRegression (GB regression) model from the Sklearn package in Python, and a basic tree base model from the "Tree" package in R to predict arrival delay. In question 2, we concluded that the age of the planes does not affect the delay time. Hence I assumed that the delay is only due to factors such as distance, airports, and time.
Data preparation and solution: Again, with the help of the database, I join flights, planes, and airport data. Mapped all the relevant categorical factors such as month and day. And encoded all the categorial variables into numerical variables and trained a GB regression with max_depth = 2 and n_estimators = 200. I had to limit the amount of data I used for training as full data collection requires a large amount of computational power. Also further it was separated into two groups training and testing. GB regression has an MSE of 2486.08. Tree-based model developed in R has an MSE of 1269.







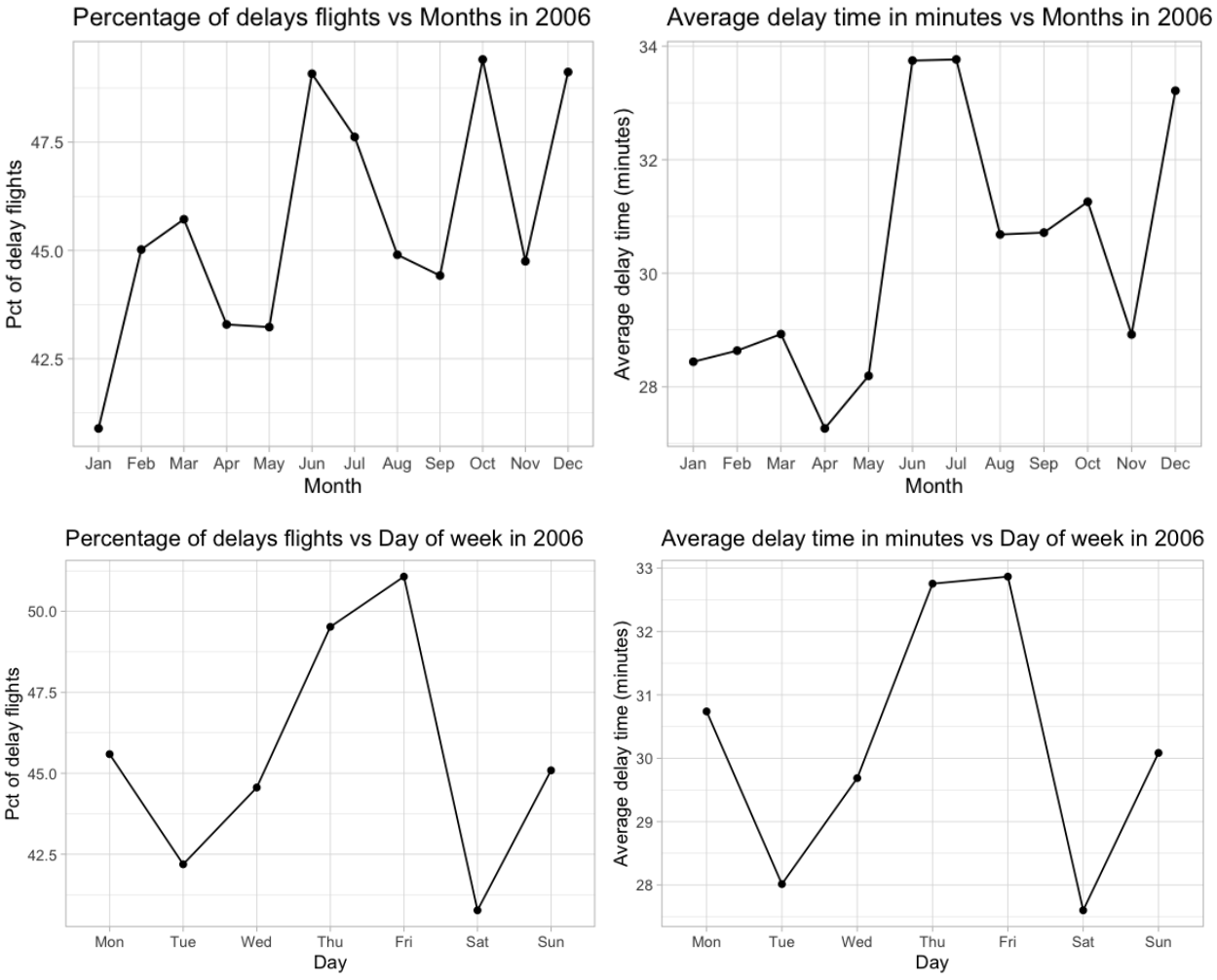
Basic tree model from R.

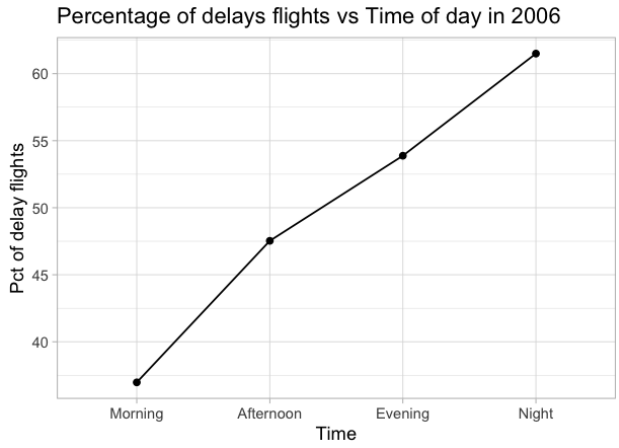
Reference used for building models:

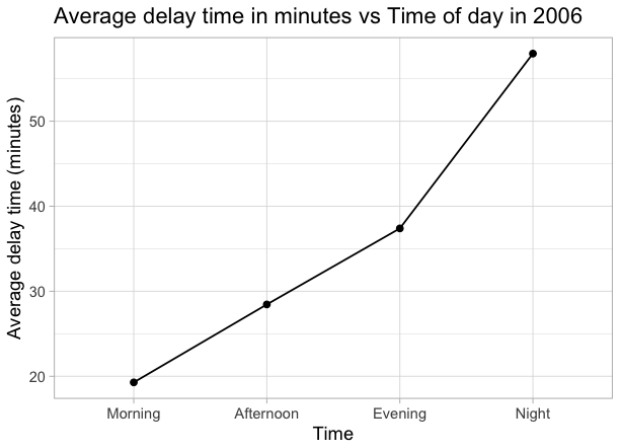
- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed; by Aurélien Géron.
- An Introduction to Statistical Learning, 2nd ed; by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Below are the corresponding R outputs for each problem:

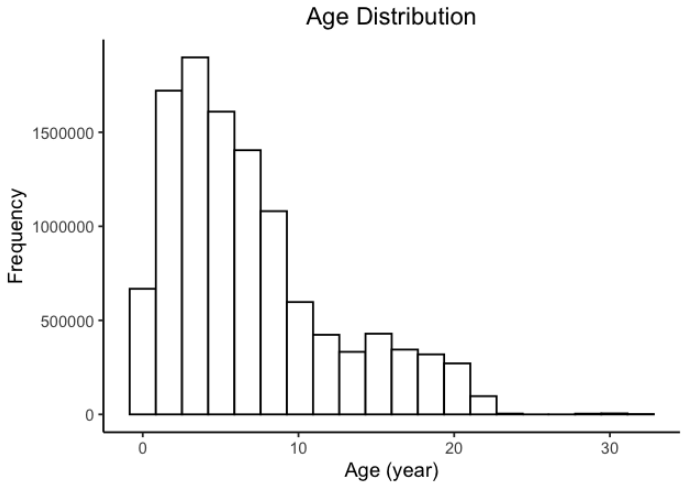
Quesition1:



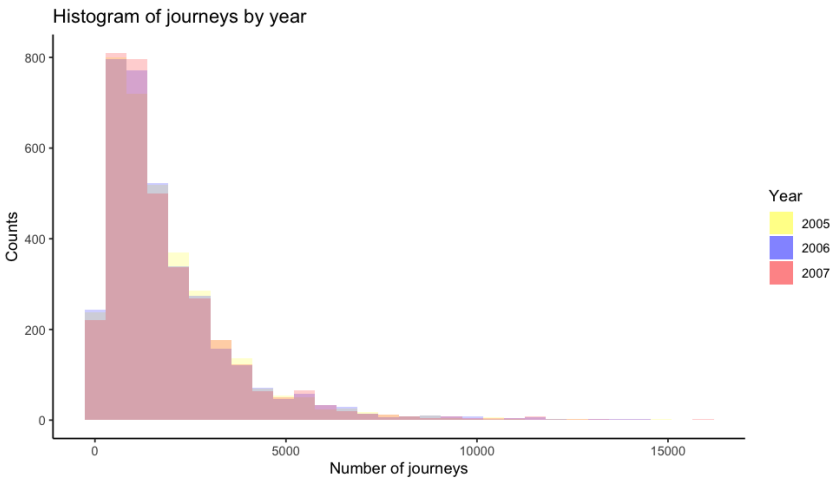


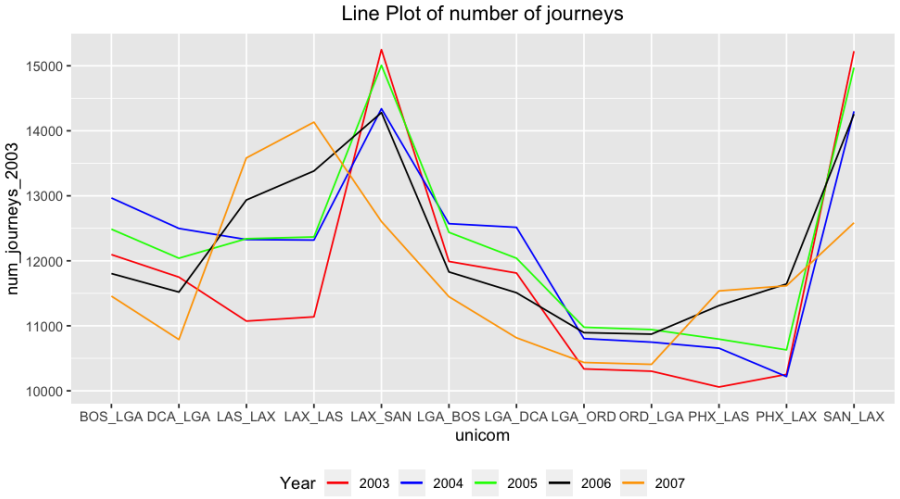


Question 2:

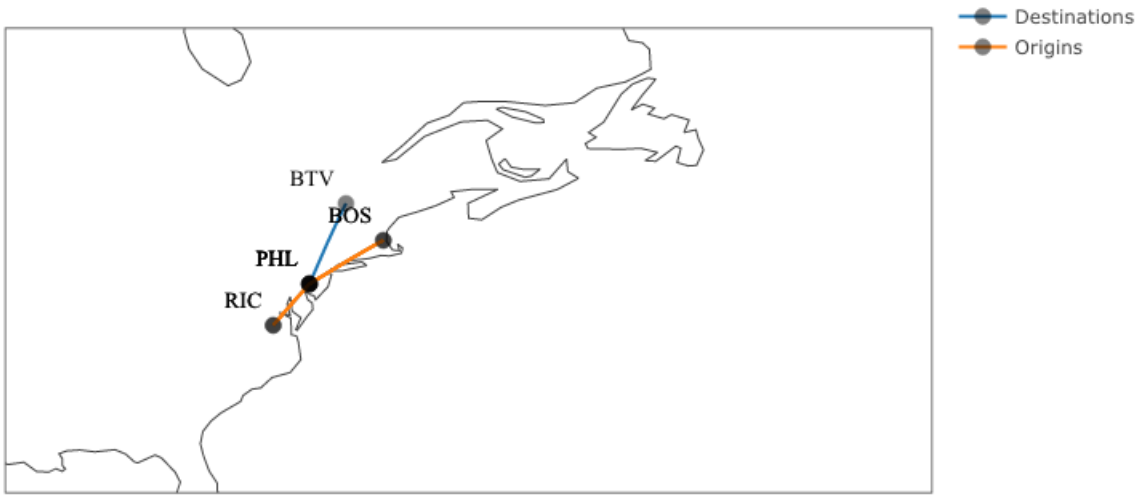


Question 3:




Line Plot of number of journeys
num_journeys_2003
15000
14000
13000
12000
11000
10000
BOS_LGA DCA_LGA LAS_LAX LAX_LAS LAX_SAN LGA_BOS LGA_DCA LGA_ORD ORD_LGA PHX_LAS PHX_LAX SAN_LAX
unicom
Year
2003
2004
2005
2006
2007

Question 4:


Destinations
Origins
BTV
BOS
PHL
RIC