# ST2195 PROGRAMMING FOR DATA SCIENCE

**Final Project Student N°2200580234**

12 DE ABRIL DE 2022

**Table of Contents:**

**Introduction: Hating Freedom**

The first time I read the coursework project description I was overwhelmed. I didn't know where to start, how to approach the questions. I felt that what was expected of me wasn't clear. I took my concerns to our professor. She told me that the whole idea of the project was exactly that. Questions where open on purpose and that we should expect this in our future jobs as data scientists. I felt frustrated. Where should I start? I went over and over the questions without knowing what to do. Then I thought, how can I work this project if I don't even know which data I'm working on? That took me to a series of (I hope fortunate) decisions. The structure of the project is separated by questions and stating clearly assumptions and decisions I made. The first three big decisions where left on the introduction because they are general for all questions. At the end of each question you can find a short conclusion for each topic.

**Decision N° 1: My data**. At first I wanted to work with information from several years. When I downloaded the data and tried to run simple operations I realized this would be impossible. I had to sacrifice a lot of data for performance. I chose going with the minimum two years because even with just too years my computation power was not enough. The years I chose where 2005 and 2006 because I was interested in working with years after 9/11 because safety regulations are similar to the ones used now a days and that could influence the delays. Now that I had two years I needed to sample. For my sample I only wanted two things. For it to be random and to have a manageable size. I chose this to be 30.000 rows. I made the randome sample in a separated R script in order to make it possible to work with the same samples in both R and Python, then saved it to a CSV file and imported that in both codes as my working data. Once I got my sample my first approach was to explore the data. I found a lot of columns that where not used in the whole project but at first I didn't know for certain which ones I was going to need and which I wasn't so I decided to keep them all and work with a specific data frame for each question.

**Decision N° 2: What is a Delay?** Before even approaching the first question I had a decision to make that would be crucial for the results of all my analysis. This would be the answer to the question "What is a Delay?" There are several possible and completely valid answers to this question but I needed to decide for one. I made some research on the matter and decided I would only consider delays on arrivals. This was not an arbitrary decision. The main reason for only considering delays in arrivals was that I wanted to make my whole project and analysis from a passenger point of view. My first hypothesis was that what really complicates a passenger is not the delay in departure (which might be annoying) but the delay in arrival. Once this was sorted out I still had to make a decision on what a delay is. My second hypothesis was that delays will be considered those that are larger than 15 minutes. I found 15 minutes to be a reasonable amount of delay that a passenger would accept as normal and anything larger than 15 minutes as something that might actually make a passenger loose a connection flight or complicate his trip. My third hypothesis was that I wouldn't consider negative values (arriving before time) as delays. This was probably the decision that got me thinking the most. Arriving before time might be as well prejudicial so I took another look at the data and found that only 19 flights in my sample had arrived more than 45 minutes earlier so I decided to ignore delays with negative values.

**Decision N°3: To Cancel or nor to Cancel.** My last decision regarding the raw data was what to do with cancelled flights. Although cancelled flights might be the worst thing for a passenger I don't

have any information regarding what happened to the passengers after cancellation. I found that on my sample only 536 flights had been cancelled so I decided to drop cancelled flights from my data. This would help me all along the way since cancelled flight had no data for arrival, delay and many other variables and didn't actually affect my work.

**Question 1: When is the best time of day, day of the week, and time of year to fly to minimise delays?"**

**The early bird catches de worm.**

I selected just the relevant columns to answer this question. This is how the head of my data frame looked like:

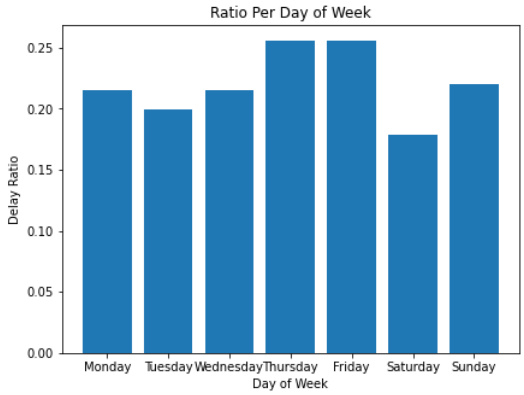
| | Year | Month | DayofMonth | DayOfWeek | ArrDelay | DepTime |
|---|---|---|---|---|---|---|
| 0 | 2006 | 11 | 6 | 1 | -26.0 | 926.0 |
| 1 | 2005 | 12 | 26 | 1 | 16.0 | 741.0 |
| 2 | 2006 | 7 | 20 | 4 | 107.0 | 2022.0 |
| 3 | 2005 | 4 | 12 | 2 | 7.0 | 1527.0 |
| 4 | 2005 | 10 | 14 | 5 | -12.0 | 1921.0 |

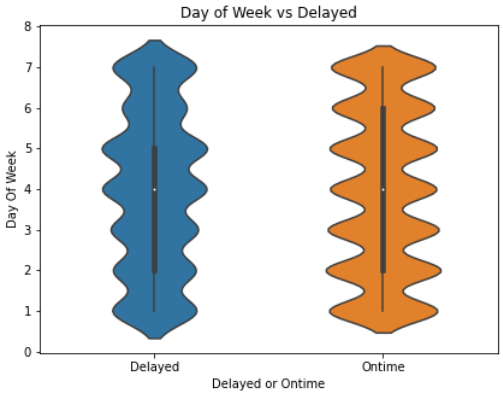
**Decision N°4: NAs.** The data presented 62 null registers for Arrival Delay. I looked for any information regarding their scheduled arrival and actual arrival but found that this where also null and because of the small amount of rows I decided to drop them.

**Decision N°5: Quantifying delays.** I knew that probably for each part of this question there would be a weekday, month or time with more flights than others so if delays where random these days/months/hours had more chance of having delays. So if I only took the amount of delays per class this would bias my answer. I decided to make a ratio (amount of delays divided total flights) per class. This is how it looked for each day of the week:

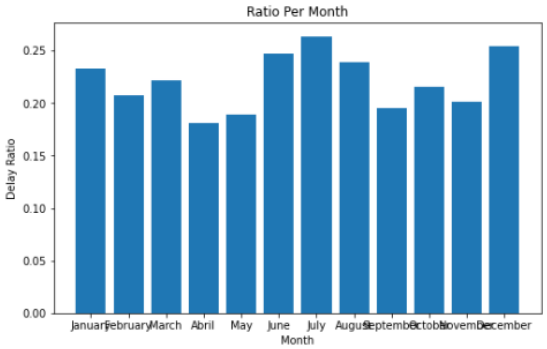
| | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Ratio | 0.215278 | 0.199726 | 0.215001 | 0.25555 | 0.255888 | 0.17844 | 0.220635 |

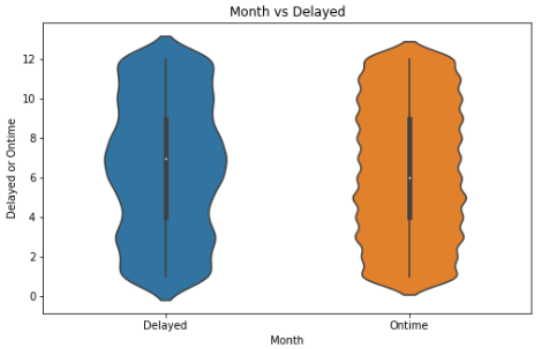
**Decision N°6: Presentation.** With this information I decided to plot two graphs per part of the question. They may be a bit redundant but they are two ways of easily viewing the results and violin plots are not that descriptive in R so I felt both plots would leave no doubts. One showed the ratio and the other one, a violin plot that showed the density of flights for whether they are on time or delayed per class.



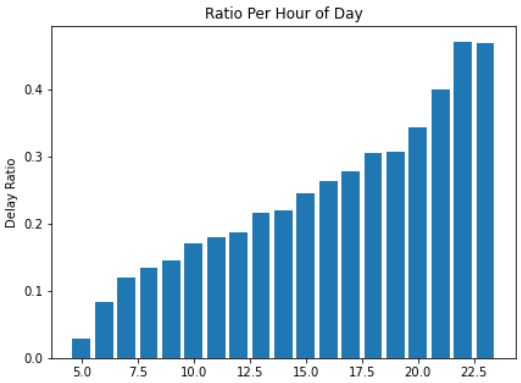



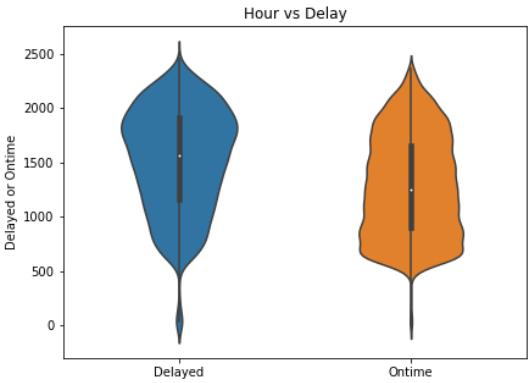

The same procedure was made for the time of year. The results of the plots where the following:




Lastly the same procedure was repeated for the best time of day to fly. I decided to use the scheduled time of departure for two reasons. Firstly when someone books a flight they takes more into account the departure time than the arrival and secondly when booking a flight you don't know the actual time of departure. I found that between 0 and 5 hours there were few flights so I decided to drop values for those hours so I would have a tidier plot, while maintaining this data in Hour vs Delay plot in order to be able to visualize this difference.



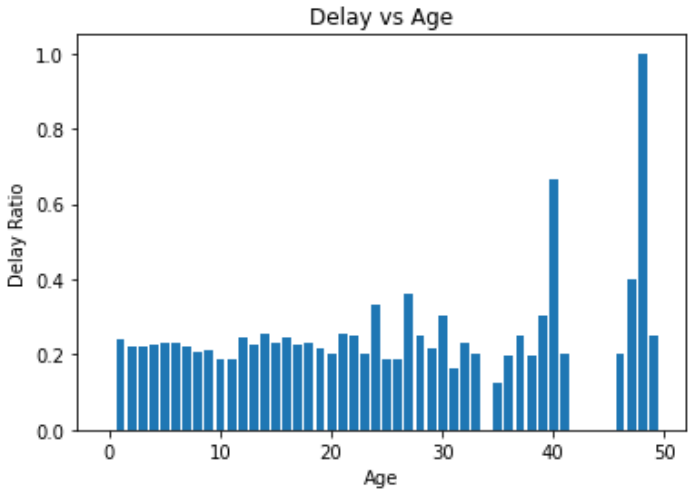

The results, in my opinion, where very clear for the three classes. In the first case Saturdays presented as a clear best choice in week day. The monthly results where the tightest. April would be the best month with May being near. In the hour graphs it's clear that the earlier you fly, the better. So the best day to fly would be a Saturday in April in the early morning.

**Question 2: Do older planes suffer more delays?**

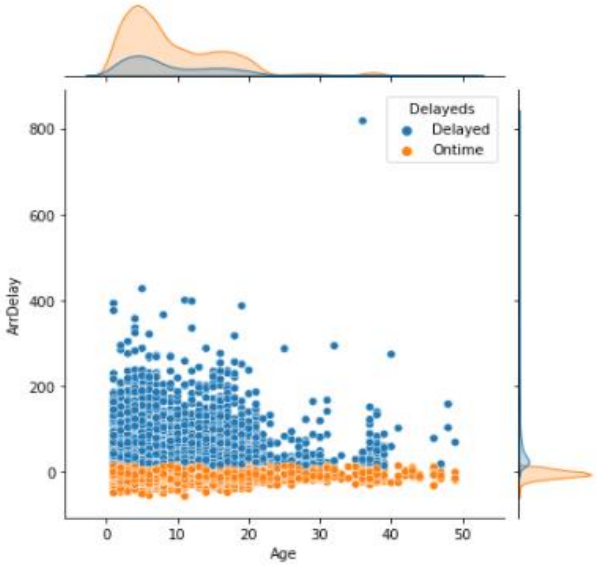
**Age matters not**

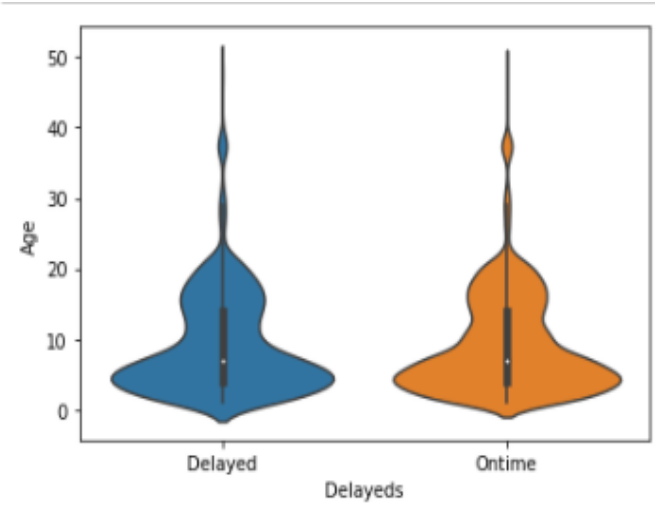
**Decision N°7: NAs strike back.** To approach this question I had to use another csv file with the planes information and merged it with mi flights data frame. My biggest problem with this new data frame was the amount of null registries for the planes information. Almost one third of my data frame had null values after the merge. I could try to substitute NAs with random values or with mean, median, etc. but because it represented a huge part of my variable of study and there was no way I could infer a year of manufacturing from the other data available I preferred to drop all NAs working with a smaller data frame. In my opinion imputation would only bias my result, it wasn't that I was imputing information on any variable, it was on the actual variable of study and although I was interested on a mean the disparity of classes could change my answer. When wrangling my data I found some impossible values. I had a plane that had age 2006 years and another that had -2 years. This outliers where clearly wrong and where dropped.

Here the ratio approach was not good enough. Although you could see that older planes might have more delay, the density of the classes was very different. Older planes had fewer flights than younger ones. This resulted in ages having a 1 ratio for age 48 and 0 ratio for age 42 for example. No good conclusions could be taken from this.




A closer look into the distribution and density of classes had to me made in order to answer this question. Here a scatter plot with its marginal densities and a violin plot where distribution of age is very clear:



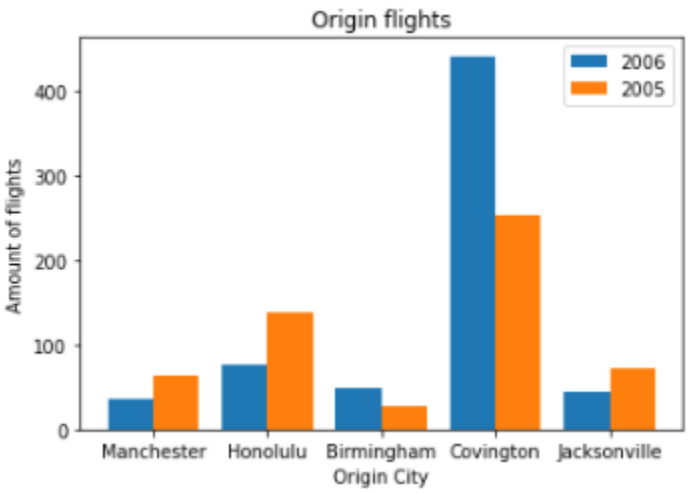



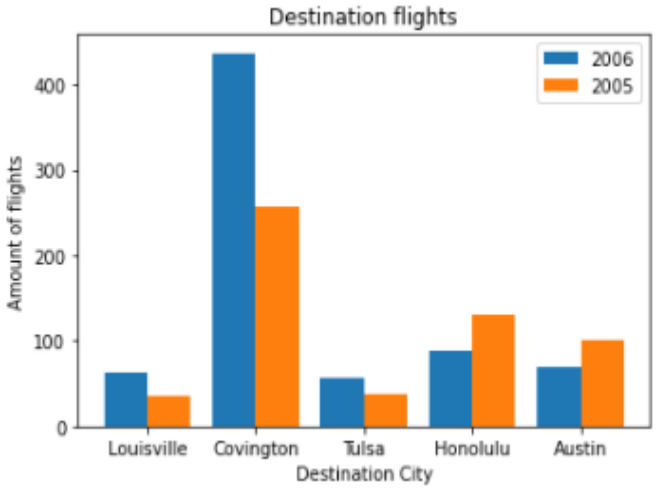

In both plots it's very clear that older planes don't travel much so data collected for these planes is probably not representative. One might infer that the reason older planes don't travel more is that they break more easily and cause delays but such conclusions cannot be made only considering the data provided.

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

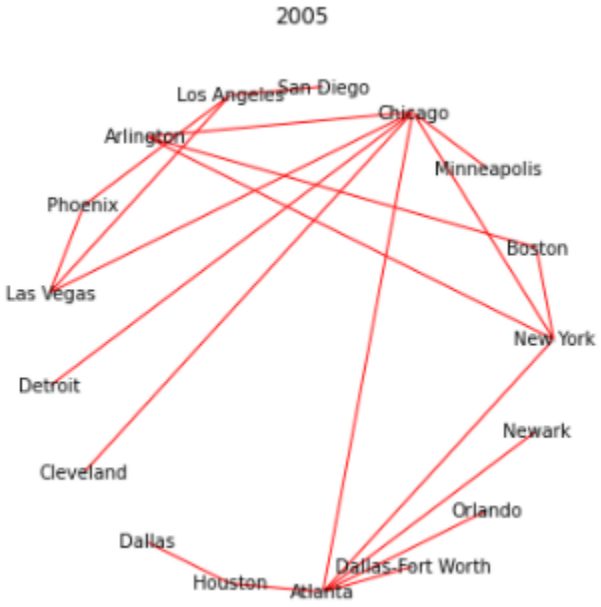
**Decision N°8: People what?** In the database provided there is no information on the amount of people in each flight, so there is not much to do in order to estimate the amount of people flying between different locations. Here I decided to use the amount of flights between different locations as my estimator of the amount of people, considering all flights as equally full. Another csv had to be imported for this questions with the airports information. This had to be joined twice with my flights data frame in order to get the airport and city for both origin and destination. Here once again null registries where few and so I decided to drop the almost 30 missing values. All the analysis was focused on the city not just on the airport so my data was grouped by the city of origin/destination.

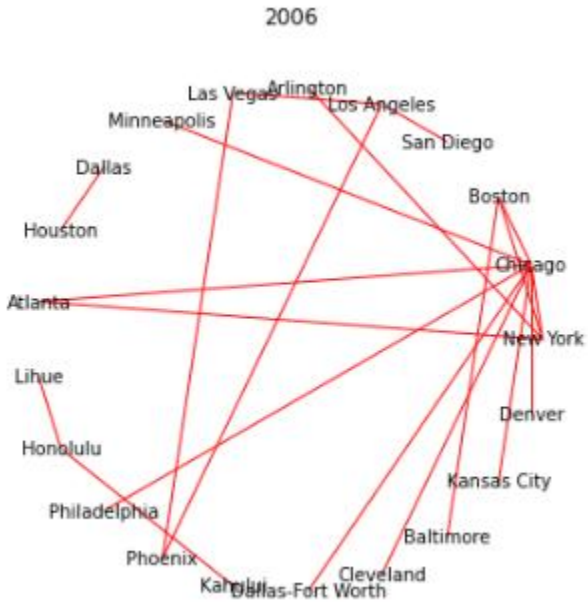
My first approach was to analyze the top origins and destinations that had changed the most between 2005 and 2006 as a percentage of the total flights of the two years. From this I selected the top 5 cities with a bigger percentage difference in absolute value and plotted the data for both 2005 and 2006 in the same graph. Here is the result:




I also wanted to see how the top routes changed between this two years so I decided to do a network graph for each year. The following resulted in:




As we can see on the network graphs there are changes in patterns of flights between the two years. Some new nodes got incorporated between de top routes while other where less common on the following year. Some cities like Covington had huge differences from one year to the next.

Comparing both years in the network graph also leads to think of a more interconnected world with more connections between cities in 2006 than in 2005.

**Question 4: Can you detect cascading failures as delays in one airport create delays in others?**

**Waterproof?**

**Decision N°9: All for one and one for all.** For this question I worked with the same data as the previous questions but analyzing airports instead of cities. In order to detect weather a delay on one airport creates a delay in another airport you have to be able to track a delayed flight and see what impact it had in another airport. To do this in a general way isn´t possible with the given tools so I decided to do this for the airport with most delays in one day. I grouped the information by year, month, day and airport of destination to see the airport that had most delays on one day and which day it was. I found that Atlanta airport received 6 flights with delays on the 19/10/2006. I traced to see where those flights came from.

| | Year | Month | DayofMonth | DepTime | ArrDelay | Origin | Dest | FlightNum | LateAircraftDelay | Delayed |
|---|---|---|---|---|---|---|---|---|---|---|
| **2970** | 2006 | 10 | 19 | 707.0 | 21.0 | MIA | ATL | 1653 | 0 | 1 |
| **6734** | 2006 | 10 | 19 | 1402.0 | 79.0 | SRQ | ATL | 4840 | 0 | 1 |
| **8899** | 2006 | 10 | 19 | 2223.0 | 66.0 | IAH | ATL | 2787 | 66 | 1 |
| **11004** | 2006 | 10 | 19 | 1154.0 | 35.0 | DHN | ATL | 4659 | 0 | 1 |
| **19215** | 2006 | 10 | 19 | 2215.0 | 222.0 | JAX | ATL | 1442 | 73 | 1 |
| **25661** | 2006 | 10 | 19 | 915.0 | 48.0 | SEA | ATL | 706 | 0 | 1 |

There were 6 flights arriving to Atlanta that day and the 6 flights arrived late. From the data we can see that from those six flights two had previous aircraft delays. Also if I look for the most flights with delays from an airport but change the filter from Destination to Origin. This was the result:

| | Year | Month | DayofMonth | DepTime | ArrDelay | Origin | Dest | FlightNum | LateAircraftDelay | Delayed |
|---|---|---|---|---|---|---|---|---|---|---|
| **786** | 2006 | 10 | 17 | 1815.0 | 26.0 | ATL | RIC | 4205 | 0 | 1 |
| **3143** | 2006 | 10 | 17 | 1711.0 | 25.0 | ATL | XNA | 4327 | 0 | 1 |
| **5022** | 2006 | 10 | 17 | 1510.0 | 15.0 | ATL | JAX | 973 | 0 | 1 |
| **8805** | 2006 | 10 | 17 | 1658.0 | 56.0 | ATL | MLB | 1197 | 52 | 1 |
| **20914** | 2006 | 10 | 17 | 1434.0 | 49.0 | ATL | CLE | 4391 | 0 | 1 |
| **25955** | 2006 | 10 | 17 | 2241.0 | 52.0 | ATL | PBI | 415 | 9 | 1 |

I don´t think that it is a coincidence that the day with most delays for both origin and destination airports are both Atlanta and both 2 days apart from each other, but I tried to trace back each flight and if it had an impact on the airport it arrived to and couldn't find any evidence. For instance the only destination that appears in both data frames is JAX. And JAX didn't have any flights at all on the 17th and just one flight on the 18$^{th}$ and that flight didn't have any delay so I cannot be sure if the flight on the 19th to Atlanta had any relationship with the flight leaving

Atlanta on the 17th. This differences probably might be generated because I had a sample of the data so it was impossible to trace back all flights.

**Decision N°10: Past is everything**. Getting to a dead end here I decided to try to look for a more general case rather than two specific cases. I went on to see which delayed flights had a previous aircraft delay (delay related to the aircraft arriving late from the flight before). Here I found that around 33% of the flights with a delay had a previous aircraft delay. Another very interesting fact was that the flights that had a previous aircraft delay bigger than 15 minutes where, in all of the cases, delayed. So if an aircraft arrived late from another flight this implies that the aircraft will be late for its scheduled flight, which might indicate clearly that cascade delays are happening.

Here working with a sample I believe really made it hard to try to detect and trace specific cases but I believe that the data found linked to previous aircraft is a great indicator that cascade delays are happening.

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

**I know you are late**

**Decision N° 11: Model.** In this case I chose two different models. Since what I wanted to predict was whether a plane was delayed or not I had to use two models that worked on predicting a categorical variable. To do so I chose Logistic Regression and Gradient Boosting. I based my code in the code we had available in the course material and the one provided by our class professor.

**Decision N°12: Variable**. My first step was to choose the variables which I wanted to consider for this model. I made several tries before deciding what the best approach would be. I wanted the model to be useful. For instance if I included the arrival delay variable the model would get a 99% accuracy which was amazing but made no sense because the arrival delay is what we use to measure whether the plane has a delay or not. I decided I wanted to use information that was available before the plane departed and if possible at the moment the passenger would purchase the ticket. Most of the variables chosen complied with this with two exceptions. The tail number of the plane and whether the plane arrived late from a previous flight. These two exceptions where made because both could be answered before the actual departure of the flight.

The variables chosen where: year, month, day of week, schedule time of departure, carrier, tail number, origin, destination, distance, aircraft delay.

In the case of departure time I used scheduled departure time for the same reason, this had an impact in model accuracy but I didn't want to use the actual time of departure because this wouldn't be known until the plane actually departed.

I imputed missing data using mean and median for continuous variables and most frequent and constant for categorical variables. I also split my data in a ratio of 80% for training and 20% for testing.

The results for the models where very similar.

For Logistic Regression we have:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.99 | 0.92 | 4689 |
| 1 | 0.95 | 0.44 | 0.60 | 1311 |
| accuracy | | | 0.87 | 6000 |
| macro avg | 0.91 | 0.72 | 0.76 | 6000 |
| weighted avg | 0.88 | 0.87 | 0.85 | 6000 |

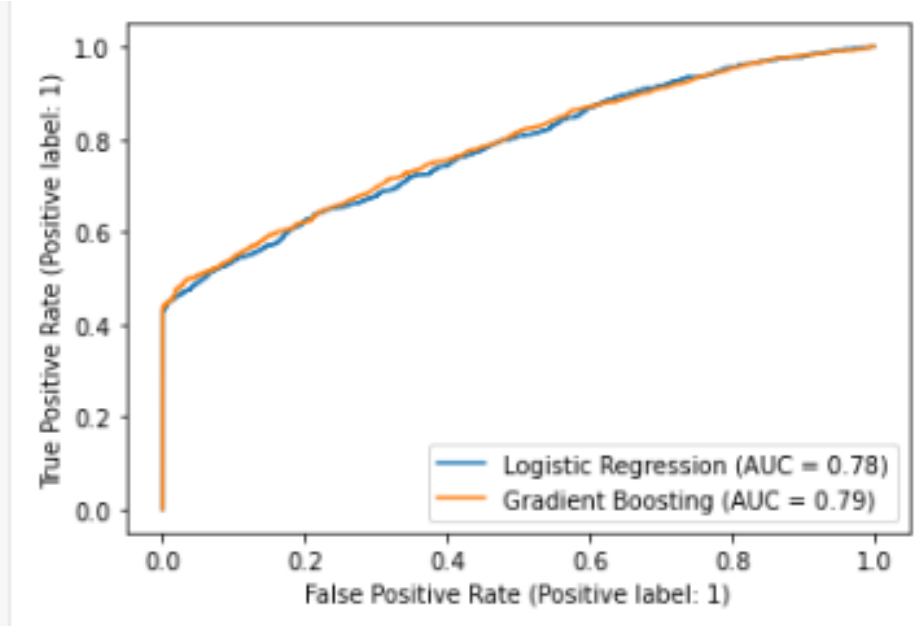
For Gradient Boosting we have:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.99 | 0.93 | 4634 |
| 1 | 0.95 | 0.44 | 0.60 | 1247 |
| accuracy | | | 0.88 | 5881 |
| macro avg | 0.91 | 0.72 | 0.77 | 5881 |
| weighted avg | 0.89 | 0.88 | 0.86 | 5881 |

With very similar results a slight improvement in precision but the same values for the other variables. Although our overall accuracy for both models was pretty good the model is not that good for what we want it: predicting delays. We were able to correctly detect only 44% of the flights that had a delay, whilst we were able to detect 99% of the flights that where not delayed. Also 95% of the flights predicted as delayed where delays so our precision for detecting delays is quite good.

These differences might arise because the classes are very unbalanced. For example in the train data the delayed values represented only 22% of the data. By balancing the classes we could get better results. Another tool we can use is tuning the hyper parameters which were left as default. This tuning could have an impact on the results of the modelling. Both techniques are beyond the scope of the course.

We also have the ROC curve and AUC score:



Where we also can see that both models have very similar results with slighter better results from gradient boosting model. In the case of the Gradient boosting model we have a 0.79 score that is the probability of correctly classifying a positive and a negative case.

Summing up I would say that the results of our model aren´t great but I think it's still quite useful. With very little information, that most can be available at the moment when the passenger purchases the ticket, we can predict almost for sure near half of the delays and we will miss the other half. I think it's still pretty good for a passenger with the given data to predict whether its flight might have a delay.

In the case of the R code I couldn't work with the whole data frame for performance issues. A simple test would take me up to five hours if it didn't crash. I decided to leave a representative code with a sample of 500 rows that works and (obviously) has very poor performance and focus my analysis on python which had greater performance and could run the whole code in minutes.

**Conclusion: End of Journey**

In my first report as an aspirant data scientist I learnt a lot about what might come up later in my career. At first I struggled with the openness of the questions but then I started to understand that not always assignments and tasks might be so direct and that we have to be prepared. I enjoyed thinking and trying to find the best solution for each problem that arose. I also enjoyed that every decision taken had a reason and nothing was decided by chance. For example before dropping a row I tried to see what options I had and which would be best decision for the project. I was surprised that I didn't use almost any technique to fill missing values (something I thought I would be doing all around the project), this didn't come from laziness but because I think that doing so would impact negatively on the results of the analysis.

Now that the project has come to an end I realized that decisions made at the beginning might affect results all the way to the end, and practice can save a lot of time. Several times I had to go to the beginning of my code to change how I wrangled data. I now finished this assignment with the feeling that in the path of a data scientist a lot of decisions have to be taken all the time, and they need to be sustained and coherent all around the project and I think I achieved that.

Regarding the actual coursework I tried to leave some conclusion at the end of each chapter to answer the question. Working with a sample might not be the best but given the performance issues I had I think some good conclusions still can be made.

In question one, we found that Saturdays mornings in April are the best days to fly in order to avoid delays. In question two, we had a big imbalance of classes that stains our conclusion with older planes having very few flights and in general having bigger delays. In question three, we saw some changes in patterns in people traveling from one year to the next and a bigger connection between cities. In question four, the sample didn't help much but we still could detect some evidence of cascade delays. In question five, we were able to make an acceptable model with very few variables that all where from information available before the departure of the plane. All in all, although I would have loved to work with the complete dataset I still think that good conclusions can be made from this sample.