Zewail City of Science and Technology University of Science and Technology CIE 427- fall 2021 

Flight Records Analysis

Asmaa Ismail 201600780

Anhar Hassan 201601171

Ahmed Elgabry 20170100

# DATASET:

The Reporting Carrier On-Time Performance [Dataset](https://developer.ibm.com/exchanges/data/all/airline/) contains information on approximately 200 million domestic US flights reported to the United States Bureau of Transportation Statistics. The dataset contains basic information about each flight (such as date, time, departure airport, arrival airport) and, if applicable, the amount of time the flight was delayed and information about the reason for the delay. We will use the dataset to answer specific questions.

# DATA COLLECTION AND CLEANING:

The full dataset file is 81 GB .csv compressed in tar.gz format. We worked on a 58 GB subset of the data from the years 1987 to 2007. We manually downloaded the data for 2020 and 2019 from the american bureau of transportations statistics [site](http://www.transtats.bts.gov) . The cleaning steps involved:

**1- Filtering out the columns we need:** the original dataset contains 128 columns out of which we only used 26 columns containing information about the source and destination airports, cities, and states, the year, month, and day, as well as the departure and arrival delay among other things. This step helped us reduce the data size to a more manageable 18 GB down from 58 GB.

**2-removing null-containing records:** this ended up removing all of the data from older years prior to 1995. The older records contained more null values than was acceptable for us; We suspect it’s because in earlier years there was less emphasis on building detailed databases and they did not record all of the flight information. It may also be that additional performance metrics were introduced post 1995 and were filled out with nulls in prior years records.

**3- extracting the cancellation data in a separate csv file:** cancellation records contain null values in most datafields, therefore they would be lost in the previous step. Consequently , we decided to extract the cancelled records into a separate csv to use them in relevant analysis questions.

**Our final clean data amounted to 11 GB + 6 GB of cancellation records. This is around 82 million flight records + 1.5 million cancelled flight records.**

# ANALYSIS QUESTIONS:

***Question one:* what is the best month, day of month/ of week to minimize delay? :**

To answer this requirements we do the following:

1- group by: month ,day of month ,day of week

2- aggregate the average of departure delay - and arrival delay

3- sort ascendingly by delay

**We observe that**

**⇒** The best month to travel in with minimum delays is September while the worst months are June and december.

***⇒***the best day to travel (minimum delay) is saturday

***Question two:* compare 2020 travel patterns vs prior years (effect of covid19 and lockdown)**

To reach this requirements we will calculate the

**1- travel frequency (number of flights) across years && across the months of 2020 and 2019**

* group by (year, month) should yield 15yrs(our dataset contains 15 year)\* 12 months
* count the rows (flights)
* aggregate the delays

**2- flight cancellation rate across years && across the months of 2019 and 2020**

* group by (year, month) should yield 15yrs\* 12 months
* count the rows (flights)
* aggregate the delays

**We observe that**

**⇒** 2020 is associated with by far the smallest delay values

***⇒*** the minimum number of flights dispatched as well as the highest number of flights cancelled was in 2020

⇒ the sharpest decrease in dispatched flights and the sharpest increase in flight cancellation happened in april of 2020

***Question Three:* compare the number of flights and average delays**

**before and after september 11, 2001 (the twin tower**

**attack).**

this is the exact same process as the previous requirement but we focus on the year 2001

**We observe that**

**⇒** The number of dispatched flights took a very sudden sharp fall on september 11, 12, and 13 with the minima being on sept 12 (the day after the twin tower attack).

***Question Four:* what are the delay patterns per airport - the airports associated with the highest departure delays?**

To reach this requirements we will

**1- group by origin airport ID, origin city, and state**

**2- aggregate the average departure delay**

**We observe that**

**⇒** Four Corners Regional Airport (FMN) has The maximum delay.

***Question Five: The change of the number of flights between different locations over time.***

We made two kinds of analysis here:

1- Grouping by the origin state, and year, then counting those groupings.

2- Concatenating origin-destination pairs, then grouping them together with the year, then counting the number of flights.

We observe that: When searching, the trend is somehow related to the state number of airports.

***Question six: the total distance travelled in each year:***

We simply group by the year and add up the distances.

**We observe that**

**⇒** more recent years tend to have higher distances with the travelling distance climaxing in 2019. However, only 2020 significantly breaks this trend by having the lowest distance of all (due to covid lockdown). Generally we expect the total travelled distance to increase in more recent years; since air travel is getting more affordable and the world experiences more globalization, more and more people are travelling by air. Not to mention that the technology and capacity of airports have gotten better as time goes on and more flights can be dispatched serving more people and covering higher distances.

***Question Seven: Average Delay Across Seasons.***

Done by mapping the months into seasons.

Then Calculating the Average Arrival Delay, and the average Departure Delay across the years.

**We observe that**

⇒summer has dominated the Departure delay since 1998, after it was the winter with the highest average dep delay, and Fall is always with the lowest average departure delay

⇒ Another insightful observation is how radically the number of flights reduced in 2002, especially in the Fall, which is the first anniversary of the 9 eleven attack.

***Question Eight:* what are the Most popular destinations across the years?**

To reach this requirements we will

**1- group by both year and destination city, count the number of flight records.**

**2- take a yearly window and find the city record associated with the maximum flight counts per year.**

**We observe that**

**⇒**the most popular destination has been overwhelmingly Chicago, Illinois since 1995.

***Question Nine:* what are the busiest airports across years?**

To reach this requirements we will

**1- group by both year and originairportid and count the flight records**

**2- group by both year and destairportid and count the flight records**

**3- join the two dfs on airport id and sum the two flights' count columns.**

**We observe that**

**⇒** The busiest airports across the years are O'Hare International Airport (ORD) in Chicago Illinois, and Hartsfield-Jackson International Airport (ATL) in Atlanta Georgia.

***Question Ten: the machine learning requirements:***

1- predicting departure delay based on time of year, month, and week, as well as source, destination and distance:

We used a regression model to estimate a predicted delay given the aforementioned features, however, we couldn’t achieve a decent prediction accuracy. We calculated the correlation between the delay and every other variable in the data and the found correlation coefficients to be all +-0.2 or less so we conclude that there isn’t really much of a linear correlation between them. We suspect it’s because the data is full of outlier values because each year, month or even week has its own unexpected events that may affect the delay in an unpredictable way. Bad weather for example, or any such small unpredictable factors can dramatically affect the delay, not to mention major events like 9/11 or the covid pandemic.

2- predicting flight cancellation based on the same features:

We tried a random forest classifier to predict flight cancellation patterns with mild success. We attribute the low accuracy of this model to the same factors that limited the other one.

# CHALLENGES AND TRIALS:

1- The biggest challenge we faced was data preprocessing and collection. The first obstacle was decompressing the whole 81 GB file in google colab and drive. The disk space of colab was very limited and would always crash after around 50 GBs were extracted. We tried to decompress it directly to google drive using colab, but it failed because colab saves the intermediate outputs to its own desk before moving them to drive and deleting the intermediates, so we were always limited by the desk space of colab. We couldn’t segment the compressed file or extract parts of it in any way because it was all one big csv compressed in a non-splittable format. We couldn’t segment the zipped file and extract each segment because it couldn’t be done with non-splittable compression unless we had access to the uncompressed form (which we didn’t). To our knowledge the only solution that remained was to download the data and decompress it in our own laptops and work locally (because we could upload that many gigabytes to drive). However, we really didn’t want to do that as it wouldn’t allow for collaborative working (everything will be in one pc) and it may overwhelm our system.

Our solution:

We worked on the 58 Gb bits that managed to get decompressed on colab and overlooked the rest. Unfortunately, this left out the more recent years data entirely. To meet the covid effect requirement, we had to go to the american bureau of transportations site and scrape 2020 data manually and add it to our dataset. There’s a consequence to this however, and that’s the gap between the last year in our original data 2007 and 2020. When we compare the 2020 data to the rest of the years, the large time gap may add some bias to the analysis, and it was generally better to compare 2020 to the years closely preceding it. We downloaded the 2019 data as well to mitigate this effect and we assume that the patterns will not change dramatically from older times to recent times and operate accordingly.

2- Another consequence of our truncated data is that the decompression failed mid 2007 so only about a third of 2007’s records were actually extracted. This would be a problem when we want to analyze the number of flights per year, because the would be a very sharp drop in 2007 and we may, if we’re not careful, interpret that as an insight of the data.

Our solution:

We simply exclude 2007 in analysis requirements that consider the number of flights per year.

3- The cascading requirement, this requirement was quite complicated and promising, we implemented it, however the very last step took really long time that we eventually had to give it up.

Our approach is to chose a specific airport, join its delay as a reference delay columns with all other delays in the same year, month, and day, then with a margin “we decided it to be 15 minutes” we accumulate all the delays of all other flights following the in-hand flight in time “with 15 minutes as mentioned”.

The very last step which was to sum only these delays is what took forever.

Our solution:

We substituted it for the distance requirement.

4- The machine learning tasks produced really bad accuracy:

As we mentioned before we assume this is because of the unpredictable factors that significantly affect delay and cancellation whose effect is hidden in the data but that we don’t have access to, like weather data for example, or political events. We conclude that this analysis requirement is not very meaningful given the features that we have available.