Data Science – Technical Inteview

Data Science – Technical Inteview
Select a dataset you like and do the following activities:

1. Describe the dataset and tell us why you find it interesting (include a link or file of the dataset).

   In this dataset we have information about flight delays, it is interesting to learn more about this, because it is an unavoidable event and it plays an important role in both profit and loss of airlines.

   In the data, we have 29 columns and could well be used to predict the flight delay at the destination airport.

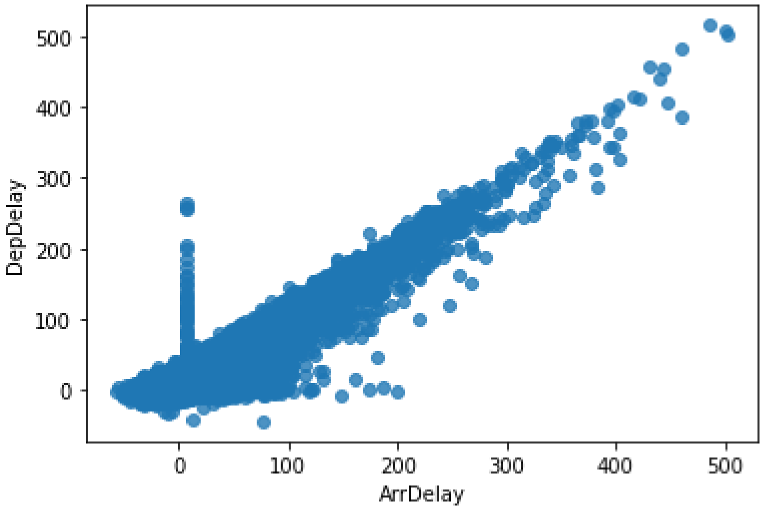
`Year`	Year of the Flight Trip
`Month`	Month of the Flight Trip
`DayofMonth`	Day of the Flight Trip
`DayOfWeek`	Day of week of the Flight Trip
`DepTime`	Actual Departure Time
`CRSDepTime`	Planned Departure Time
`ArrTime`	Actual Arrival Time
`CRSArrTime`	Planned arrival time
`UniqueCarrier`	A carrier code is a four-character unique identifier that is assigned by the CBSA to identify a carrier
`FlightNum`	Flight Identifier
`TailNum`	Aircraft Identifier
`ActualElapsedTime`	AIR_TIME+TAXI_IN+TAXI_OUT
`CRSElapsedTime`	Planned time amount needed for the flight trip
`AirTime`	The time duration between wheels_off and wheels_on time
`ArrDelay`	Total Delay on Arrival in minutes
`DepDelay`	Total Delay on Departure in minutes
`Origin`	Starting Airport

`Dest`	Destination Airport
`Distance`	Distance between two airports
`TaxiIn`	The time duration elapsed between wheels-on and gate arrival at the destination airport
`TaxiOut`	The time duration elapsed between departure from the origin airport gate and wheels off
`Cancelled`	Flight Cancelled (1 = cancelled)
`CancellationCode`	Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
`Diverted`	Aircraft landed on airport that out of schedule
`CarrierDelay`	Delay caused by the airline in minutes
`WeatherDelay`	Delay caused by weather
`NASDelay`	Delay caused by air system
`SecurityDelay`	Delay caused by security
`LateAircraftDelay`	Delay caused by aircraft

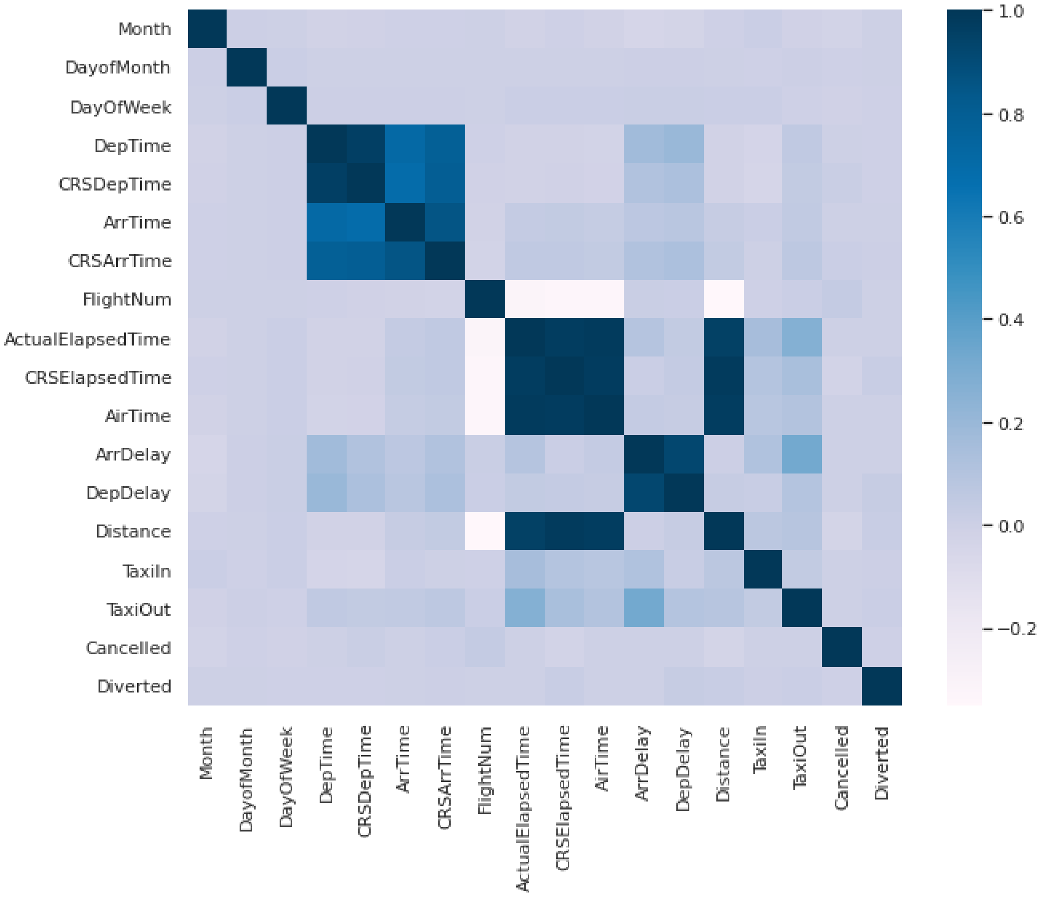
2. Create a correlation plot and write down a few key observations.

A scatterplot shows the relationship between two quantitative variables measured for the same individuals.

We can visualize the relationship between two variables with a scatter plot. For example, it is evident that the greater the delay of a flight in its departure `DeepDelay`*, the greater the delay in its arrival* `ArrDelay` *and the graph shows it with a strong positive correlation.*

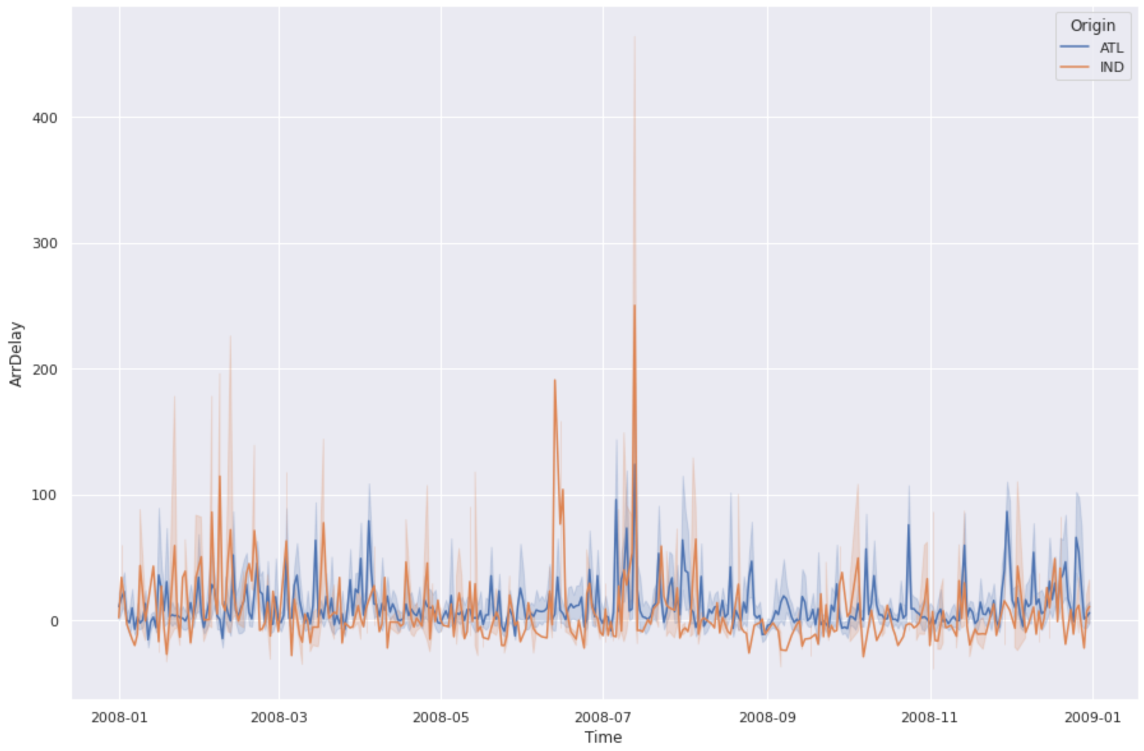


If we want to know the correlation of all the variables, we can use a heat map. In the figure, we notice that variables such as `DepTime`, `CRSDepTime`, `ArrTime`, and `CRSArrTime` have a strong positive correlation with each other. On the other hand, the variables `ActualElapsedTime`, `CRSElapsedTime`, `AirTime` and `Distance` have a negative correlation with the `FlightNum` variable, although it would be necessary to analyze if this really makes sense.



3. Provide 1 visualization about the information you found the most interesting.

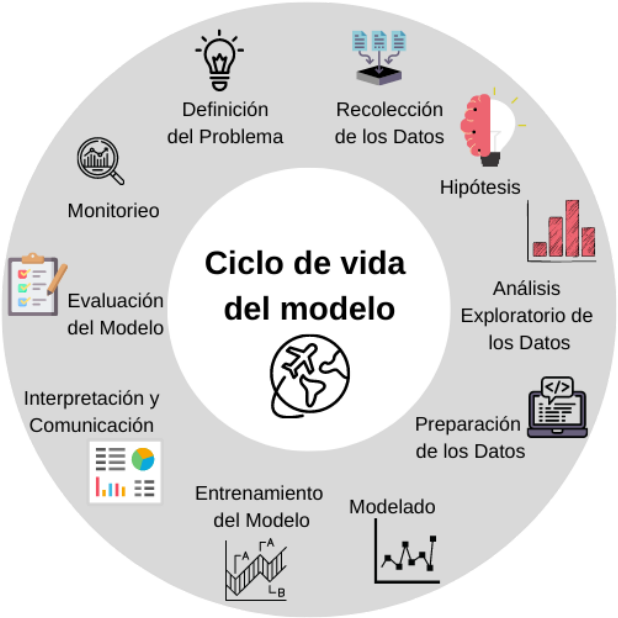
   One of the most interesting graphs that we can make is using time series. In the following image we are going to show a graph where we have the time on the 'x' axis, the delays on the 'y' axis and a filter of two origin airports.



4. What kind of machine learning algorithms can you use with this dataset?
    a. *Linear Regression*
    b. *Logistic Regression*
    c. *Decision Tree*
    d. *Random Forest*
    e. *Naive Bayes*
    f. *K Means*
5. Select one of the algorithms from the previous question and tell us which variables you
    would use to build it.

    Applying the random forest algorithm, we could try to estimate the `ArrDelay`, that is, the arrival delay of a flight.

    I would choose the variables:
    `Distance`, `AirTime`, `TaxiIn`, `TaxiOut`, `DepDelay`.
6. Make a flowchart for the life cycle of the model you proposed.

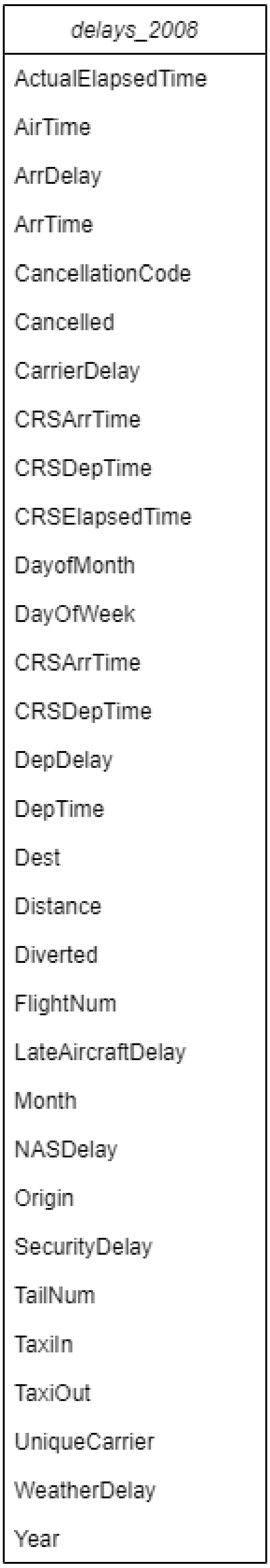


7. What frameworks and libraries would you use to develop this project? Can you estimate how long it would take to deliver a first version?

   For the development of this project, libraries will be necessary according to the selected algorithm. The main aspects would be treated with

   - *visualization: matplolib, seaborn*
   - *data manipulation: pandas, numpy*
   - *modeling: sklearn, scipy*

8. (Bonus) Suppose you manage to obtain a data feed similar to your dataset. Design a
   relational schema that will be used to store the received data.



delays_2008
ActualElapsedTime
AirTime
ArrDelay
ArrTime
CancellationCode
Cancelled
CarrierDelay
CRSArrTime
CRSDepTime
CRSElapsedTime
DayofMonth
DayOfWeek
CRSArrTime
CRSDepTime
DepDelay
DepTime
Dest
Distance
Diverted
FlightNum
LateAircraftDelay
Month
NASDelay
Origin
SecurityDelay
TailNum
TaxiIn
TaxiOut
UniqueCarrier
WeatherDelay
Year