DESCRIPTION : On-time data for all flights that departed NYC in 2013.

DATASET: File “flights.csv” - 336,776 rows, 19 columns
The table on the right side has the metadata.

BUSINESS CONTEXT : A data science team is working on an important project related to airports and optimization. You are a new team member and was hired as junior data scientist. Each team member has received a task and **your task is** to check if you can say that Delta Airlines (DL) flights are delayed more than United Airlines (UA) flights ? The file (data source) is already available to start your analysis. To solve this problem you will need to use statistical concepts and python programming language that are the foundation of modern machine learning development.

COLUMN NAME	TYPE	DESCRIPTION
year	year	Year of departure.
month	integer	Month of departure.
day	integer	Day of departure.
dep_time	integer	Actual departure times (format HHMM or HMM), local tz.
sched_dep_time	integer	Scheduled departure times (format HHMM or HMM), local tz.
dep_delay	integer	Departure delays, in minutes. Negative times represent early departures/arrivals.
arr_time	integer	Actual arrival times (format HHMM or HMM), local tz.
sched_arr_time	integer	Scheduled arrival times (format HHMM or HMM), local tz.
arr_delay	integer	Arrival delays, in minutes. Negative times represent early departures/arrivals.
carrier	string	Two letter carrier abbreviation.
flight	integer	Flight number.
tailnum	string	Plane tail number.
origin	string	Origin of flight.
dest	string	Destination of flight.
air_time	integer	Amount of time spent in the air, in minutes.
distance	integer	Distance between airports, in miles.
hour	integer	Hour of scheduled departure.
minute	integer	Minute of scheduled departure.
time_hour	datetimestamp	Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

Task #1: Create a Jupyter Notebook to present your analysis and perform an exploratory data analysis **(EDA)** on **flights.csv file.**

Task #2: Create at least 5 different views (reports) to the team using the imported dataset. The reports should be composed of a description, charts and tables based on some selected columns. These reports can give some insights to the team about the airport operation and potential optimizations.

Task #3: Create a new dataset named (**pop_data**) with flight data from airlines UA (United Airlines) and DL (Delta Airlines). The data set must contain only two columns, company name and delayed arrival flights. The data must be extracted from dataset flights.csv to build the pop_data dataset. The dataset should be limited to no more than 20,000 rows per airlines.

Task #4 Create two new datasets (“**dl” and “ua”**) of 1000 observations each from the “**pop_data”** dataset only with data from the **DL (Delta Airlines)** for “**dl”** and only data from **UA (United Airlines)** for “**ua”** . Tip: Include a column called sample_id populated with number 1 for the first sample and 2 for the second sample

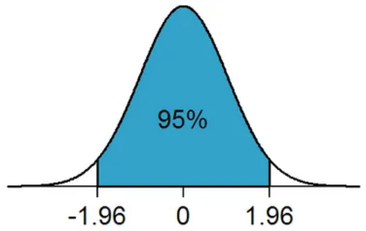
Task #5 : Create a new dataset (**samples**) containing the data of the 2 samples created in the previous item to be used in future analysis.

Task #6 : For each sample (“**dl” and “ua”**) calculate de standard error and the mean. Standard error can be calculated using this expression :

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample
σ = sample standard deviation
n = number of samples

Task #7 : For each mean calculated before we need to define the “confidence intervals” in this case 95% confidence interval. It means calculate lower and upper values.



Display :

LowerValue “UA” - MeanValeu “UA” - UpperValues “UA”

LowerValue “DA” - MeanValue “DA” - UpperValues “DA”

Task #8 : After these set of previous calculations it was requested that you take the **T-TEST** concept and apply it on your 2 groups of data (“means”) to let your team knows if Delta Airlines (DL) flights are delayed more than United Airlines (UA) flights ?

Note: Regarding T-TEST you can make a quick research about it on the web and find ways to use it considering these 2 samples.

Lab1 : After complete your Lab, please submit your Jupyter notebook with all codes, charts and descriptions that you have used to solve this important team task. The grade will be higher depending on how rich is your analysis regarding the original dataset and how many #tasks were successful delivered.