**Task# 1**

*Write a python3.7+ program which would run locally on the laptop and perform the following:*

*#1 Generation phase:*

*Program should generate N=~5000 JSON files on disk in /tmp/flights/%MM-YY%-%origin\_city%-flights.json or similar folder structure where each file is a JSON array of random size M = [50 – 100] of randomly generated flights data between cities. Total set of cities is K=[100-200]. Flight record is an object containing  {date, origin\_city, destination\_city, flight\_duration\_secs, # of passengers of board}. Some records, with probability L=[0.5%-0.1%] should have NULL in any of the flight record properties.*

*#2 Analysis & Cleaning phase:*

*Program should process those files in the most optimal way and produce the following result:*

*- #count of total records processed, #count of dirty records and total run duration.*

*- AVG and P95 (95th percentile) of flight duration for Top 25 destination cities.*

*- Assuming cities had originally 0 passengers, find two cities with MAX passengers arrived and left.*

**Task# 2:**

Dataset and note book : (please refer the link for the dataset and model)  
<https://www.kaggle.com/code/subhamchauhan1100/u-s-electricityprices-data-analysis-randomforest>  
  
**Now the task is following:**

**Instructions:**

* **Model Pipeline:** Build a basic CI/CD pipeline High-Level Flow Diagram
* **Containerization:** Containerize (e.g., Docker) the model and its dependencies for portability in your local

Please also define what is YAML file and why it is an essential component in automating CI/CD pipeline.