1. Exercise: Search and Summarize the tables (Table to insights)

- **Background**: A company sells IT software, hardware, and related business services. The company has asked you to provide quick, actionable insights and alerts from their machine usage logs.
  The company has provided you data to create this system

- **Type of data:** Structured text present in tabular format

- **Data details:**
  - Machine usage logs and corresponding insights (please note that the insights are generated using a GPT model and may not be 100% accurate). Our goal is to build models that can give summaries as close as possible to that of GPT.
  - Data dictionary of the columns available in the data
  - The data is randomly split into train and test. Please use the test data as unseen data in case you choose to train/ finetune an existing model. If you use a pre-trained LLM, you can use train data for prompt Engineering and test data for evaluation.
  - Link to data:Read the complete details and access data at: “Table insights” dataset in the datasets topic of this classroom

- **Expected Result:** A solution that can summarize tables and give actionable business insights.

- **Libraries to use:** huggingface pipelines, Langchain, Llamaindex and some other libraries can also be used

- **Table extraction packages :** Tabula, Camelot

- **Open source models**: Dolly, Falcon, Llama, T5, BART, Flan-T5 , Pegasus can be used.

2. Deliverables

- Do the assignment on google colab
- Upload colab notebook containing the codes
- Provide a report (txt or or doc or docx) file containing following details:

    - Section 1: Model details (i.e. model name, number of parameters, etc)
    - Section 2: Experiment results(i.e. Prompt used for generation of results, Five examples results, metrics like BLEU, Rouge and semantic match scores, etc)
    - Create the docs pointwise
- The colab notebook and the report should be uploaded in the folder named “Assignment 4” on github