To attempt this assignment, you need to identify a dataset through Kaggle (<https://www.kaggle.com/>) or GitHub (<https://github.com/>) portal or any other public/private repository. If you are not already registered in these portals for downloading datasets, you will need to register.

You can choose a dataset from a domain of your interest, such as sports, education, energy and environment, healthcare, financial stock market, agriculture, and so on. The dataset you choose should have structured (e.g. financial data etc.) and/or unstructured data (e.g., social media posts or images etc..) and can have some valuable insights drawn from it. You will need to share a publicly accessible source/link of the dataset you chose as part of the output of the assignment. In case you choose a private dataset ensure you have met all the licensing requirements before sharing the dataset.

**Note**: The dataset selected by you should have at least 1000 data points (records).

For the dataset selected by you, provide answers to Question 1 and 2:

**Question 1**

a) Explain the business or social problem you are addressing with the chosen dataset

b) Identify two *predictive machine learning insights* which you can draw from the dataset. Provide the steps of machine learning that you would undertake for implementing one of the two chosen insights.

**Question 2**

a) Implement a predictive model using scikit learn in Python for the chosen dataset above (an ipynb file, preferably on google colab needs to be shared). Clearly identify the features, the target variables and their types, and the choice of the algorithms used in your model.

b) Extend your model to make the prediction using an alternate algorithm. Comment on the relative performance, documenting the choice of you evaluation metrics and comparing the relative performances.

**Question 3**

a) In the Hands-On exercise in Module 10, Lesson 2, Videos/Topics 2 and 3, we train a sentiment analysis model (ANN Classification using TensorFlow.ipynb) using Neural Network. Extend the code in ANN Classification using TensorFlow.ipynb to implement the sentiment analysis using the hugging face transformer pipeline for sentiment analysis on the train.tsv file.

b) Comment on the accuracy achieved using the hugging face transformer pipeline in comparison to the Countvectorizer and Neural Network based TensorFlow model.

c) Explain the Transformers architecture and how it has contributed to the growth of Generative AI applications using Large Language Models (LLM). Mention a few emerging applications of LLM’s in your area of work to provide specific examples. You can include latest developments like multi modal applications and not necessarily restrict to the coverage in the course alone.