# Assignment

The assignment has two sections:

1. Section 1
2. Section 2

Kindly ensure both the sections are completed before submission

The zip file has a folder “data”, the subfolder has two folders highlighting section\_one\_data and section\_two\_data.

You will find all the data files highlighted below in this folder itself

Please ensure your output folder should be zipped, and have all files that are highlighted below

# Section 1

Introduction:

The objective of this case study is to analyze the email metadata contained in the Excel file named "Emails\_CS.xlsx." The focus will be on cleaning and organizing the data to ensure its accuracy and usability. Once the data has been refined, a cleaned version of the Excel file will be shared with stakeholders. This cleaned dataset will enable your manager to identify specific dollar amounts, which are crucial for their data modeling efforts.

Use a Jupyter notebook to write your code.

***Please ensure the approach has been highlighted for each task in the markdown section***

***Kindly add comments to the code and ensure standard coding practices***

1. Conduct Exploratory Data Analysis (EDA) and perform the necessary data cleaning steps on the provided dataset. **(input data ref: use Emails\_CS.xlsx as input data)**
   1. Clearly explain each step/steps using markdowns.
   2. For each step, where needed, indicate the number of rows affected or changed as a result of your actions.
2. There are 5 excel files in the folder given (input data ref: **ExcelFiles).** Read the files and add the metadata in the excel Emails\_CS.xlsx.

**Note: Use a single approach to read the excel files.**

1. Identify and create flags for rows that contain specific dollar values within the dataset. **(input data ref: the data formulated from the above 2 tasks serves as an input to this)**

Extract the dollar values in a new column where found.

* 1. Create a markdown and include the total number of rows where dollar values were identified.
  2. You can do further analysis at this point keeping in mind the data needs to be easily usable by your manager.

Submission:

Save the Jupyter notebook in the following format: FirstName\_LastName.ipynb.

Save the output in csv / excel format and attach it with the jupyter notebook.

# Section 2

Introduction: section\_two\_data – this folder has pdf files – Extract data from these pdfs and perform the following tasks.

***Please ensure the approach has been highlighted for each task in the markdown section***

***Kindly add comments to the code and ensure standard coding practices***

1. Extract data (perform pdf extraction / mining) from the 6 pdfs in a unified file as master dataframe from each pdf
2. Extract bold lines present (page level), from each pdf in a master dataframe\_two
3. Extract underlined words (page level), from each pdf in a master dataframe\_three
4. Extract tables from pdf (page level), from each pdf in a master dataframe\_four

Data Files:

Pdf\_data – has pdf files

Submission:

Please submit your jupyter notebook, please create functions, add comments to your code. Explain the approach in markdowns.

Output: 4 Master dataframes as highlighted above, export them in csv / xlsx format