CS 301
Project 1

In this project, you will use a dataset of your choice from the provided open data portal and build a regression model to demonstrate your technical and analytical skills. Upon project completion, you are required to upload:

- A 5-minute recorded presentation with accompanying slides, summarizing key aspects and findings.
- The data used in a csv file format. If your data file is so large that Canvas will not accept it, you may submit a link to a shared file (Google Drive or equivalent).
- A Jupyter notebook showing all of your work.

The project consists of the following tasks:

Task 1: Data Collection and Preprocessing

- Use the following data portal to collect a dataset that has a minimum of seven features, including at least one categorical and one numerical feature:
  NYC Open Data Portal - https://opendata.cityofnewyork.us/
- Clean and preprocess the data to ensure accuracy and consistency. Elaborate on each preprocessing step to provide a detailed understanding of the process. Justify your choice of preprocessing techniques.

Task 2: Data Visualization

- Implement a range of data visualization methods to unveil inherent patterns and correlations in the dataset.
- Explain why the selected visualizations are appropriate for your analysis.

Task 3: Regression Analysis:

- Define a target variable in your dataset and apply a multiple regression model with at least one categorical and one numerical feature to analyze complex relationships.
- Interpret the model outcome in terms of learned parameters. You can make a prediction and explain it using the parameters learned by your model.
- Evaluate your model with at least two evaluation metrics and explain the result in detail.

Communication of Findings:

- Combine the insights gained from data visualization and regression analysis results to provide a holistic understanding of the dataset.
- Create compelling visualizations and presentation materials to effectively communicate results to both technical and non-technical audiences.

- Develop a comprehensive report summarizing key findings, insights, and the interplay between data visualization and regression analysis.