HW #4: Data Analysis of Building Energy Benchmarking Data

Course: **DATA 601**
Instructor: **Syed Tauhid Ullah Shah**

Objective:
The purpose of this HW is to perform an in-depth analysis of the City of Calgary's Building Energy Benchmarking dataset. Students will use Python, **Regular Expressions (Regex)**, *Pandas (for minimal tabular operations),* NumPy, and Matplotlib to preprocess, analyze, and visualize the data. The assignment requires cleaning the dataset, **extracting relevant data using Regex**, performing aggregations, detecting outliers, and conducting exploratory visualizations.

Submission Format:

- Jupyter Notebook (.ipynb) containing all code, analysis, and visualizations.
- A structured report (.pdf or .md) summarizing key findings, trends, and insights.
- Submit the GitHub repository link in the D2L report along with the Jupyter Notebook.
- Clearly document in the README.md how **Regex** was used for data extraction and cleaning.
- The repository must include:
    - A well-organized folder structure.
    - A README.md with clear explanations of methodology, challenges faced, and insights gained.
    - Proper version control with meaningful commit messages.

Part 1: Data Cleaning and Preprocessing

1.1 Load and Inspect the Dataset

- Load the dataset and display its shape, column names, and data types.
- Identify and list the number of missing values in each column.

1.2 Handling Missing Data

- Drop columns with more than 40% missing values.
- For numerical columns, fill missing values with the median of their respective column.
- For categorical columns, fill missing values with the mode of their respective column.

1.3 Extracting and Cleaning Data Using Regex

- **Use Regex only to:**
    - Extract numeric values from text-based numeric columns (e.g., Property GFA, Energy Use, Emissions).
    - Standardize Postal Codes to follow the Canadian format (A1A 1A1).

- Clean and extract meaningful text from Property Names and Addresses.
- Ensure extracted values are properly converted to numerical types for analysis.

Part 2: Exploratory Data Analysis (EDA) and Aggregations

2.1 Statistical Summary

- Generate summary statistics for numerical features using extracted data.
- Identify and explain key observations (e.g., outliers, mean vs. median differences).

2.2 Aggregations

- Compute the average Energy Use Intensity (EUI) by Property Type.
- Compute the total Greenhouse Gas (GHG) emissions by year.
- Identify the top 5 properties with the highest total energy consumption.

2.3 Detecting Outliers Using Regex and IQR

- **Use Regex only to:**
  - Identify values that do not conform to expected numeric formats.
  - Remove or correct incorrectly formatted numeric values.
- Apply the Interquartile Range (IQR) method to detect outliers in Total GHG Emissions (Metric Tons CO2e).
- Replace outliers with the median value for that property type.

Part 3: Data Visualization

3.1 Time-Series Visualization

- Plot the yearly trend of average Site Energy Use Intensity (EUI).
- Highlight any significant increases or decreases in energy usage.

3.2 Comparative Bar Charts

- Create a bar chart showing the top 10 buildings with the highest GHG emissions.
- Annotate the bar chart with emission values.

3.3 Heatmap Visualization

- Create a heatmap of energy usage intensity (Site EUI (GJ/m^2)) across different property types.

Part 4: Further Analysis

4.1 Correlation Analysis

- Compute and visualize the correlation matrix between energy consumption, emissions, and building size.
- Identify any strong correlations and explain their implications.

4.2 Hypothesis Testing

- Conduct a t-test (**t-test** is used to compare the means of two groups to determine if they are significantly different from each other. More at Student's t-test - Wikipedia) comparing the average Energy Star Score between two different property types (e.g., Offices vs. Residential buildings).
- Interpret the results and discuss statistical significance.

Part 5: Reporting and Insights

5.1 Summary Report

- Write a structured report (300-500 words) covering:
  - Key trends in energy consumption and efficiency.
  - Seasonal and property type variations.
  - Recommendations for improving energy efficiency and reducing emissions.
- Include supporting visualizations with clear titles, labels, and legends.
- Submit the GitHub repository link in the report on D2L along with the Jupyter Notebook.
- Highlight in the report where Regex was used for data cleaning and extraction.

General Considerations

- **Missing Required Parts**: Deduction for missing or incomplete sections.
- **Incorrect Calculations**: Deduction for incorrect or misinterpreted calculations.
- **Poor Code Readability**: Deduction for code that lacks comments, is difficult to follow, or is poorly structured.
- **Incomplete or Unclear Report**: Deduction for reports lacking depth, clarity, or proper formatting.
- **Improper GitHub Usage**: Deduction for repositories that are not structured properly, lack a README.md, or have non-meaningful commits.

Additional Tasks

- Analyze the relationship between building age and energy efficiency.
- Use Regex only to clean and standardize text-based data such as property names and addresses.
- Generate a dashboard-style visualization combining multiple Matplotlib plots for an interactive overview.
- Ensure the GitHub repository follows best practices, including an organized folder structure, detailed README.md, and version-controlled commits with meaningful messages.

Good luck!