## Assignment 6: Multilinear and tree regression model

**Objective:**

The objective of this assignment is to practice data exploration and apply regression techniques, including Multilinear Regression and Tree Regression, to make predictions. Consider the code that we developed in class as a guide.

---

## Instructions:

### 1. Data Exploration and manipulation (2 points)

1. **Dataset Overview**:
    - Describe the dataset:
    - Identify the target variable.
2. **Check for Missing Values**:
    - Identify and report any null values in the dataset.
    - Handle missing values appropriately (if present).
3. **Convert Categorical Variables**:
    - Identify categorical columns.
    - Convert them into dummy variables using pandas.
4. **Correlation Analysis**:
    - Plot a correlation matrix to visualize relationships between variables.
    - Include annotations in the correlation matrix for clarity.
5. **Additional Visualization**:
    - Create at least one additional plot using seaborn to explore the data (e.g., scatter plot, box plot, or pair plot).
    - Interpret the plot in a few sentences.

---

### 2. Multilinear Regression Model (2 points)

1. **Preprocessing**:
    - Standardize the numerical variables using `StandardScaler` from scikit-learn.
2. **Model Creation**:
    - Train a multilinear regression model on the dataset.
3. **Model Summary**:
    - Use `statsmodels` to print the regression summary
4. **Make Predictions**:
    - Use the model to calculate predictions on the test data.
5. **Evaluate Model Performance**:

    - Calculate the following metrics for the model:
        - Mean Absolute Error (MAE)
        - Root Mean Squared Error (RMSE)
6. **Fine-Tuning**:
    - If applicable, identify statistically significant features (p-value < 0.05).
    - Create a new regression model using only significant features and evaluate its performance.

---

**3. Tree Regression Model (2 points)**

1. **Model Creation**:
    - Train a regression tree model on the dataset using scikit-learn.
2. **Feature Importance**:
    - Identify and report the most important features from the tree model.
3. **Make Predictions**:
    - Use the model to calculate predictions on the test data.
4. **Evaluate Model Performance**:
    - Calculate the following metrics for the model:
        - Mean Absolute Error (MAE)
        - Root Mean Squared Error (RMSE)
5. **Fine-Tuning**:
    - If applicable, refine the tree regression model using the most important features and evaluate its performance.

---

**4. Extra Mile (Optional 1 point)**

1. Create a **Random Forest Regression Model**:
    - Train a random forest model using the dataset.
    - Evaluate its performance using MAE and RMSE.
    - Compare the results with the previous models.

---

## Submission Requirements:

1. Upload your code to your GitHub, send the link of your GitHub repository and record a video between 5 and 10 minutes long explaining your code and conclusions. Attach the video as a single file or upload it to YouTube or Vimeo. If you prefer, you can set it to be accessible only to those with the link. It is not necessary for you to appear in the video, but it is appreciated if you do. Please note that I will not accept files shared via

WeTransfer, as they expire too quickly. Additionally, I will not accept the audio and video as separate files.This assignment will count as two assignments and is the last one.

## Deadline: November 24 23:59