جامعة بيرزيت

BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department
Machine Learning and Data Science - ENCS5341
Assignment #1
Submission deadline: 30/10/2024

This assignment may be completed by a group of up to two students.

The objective of this assignment is to work with a real-world dataset, focusing on data preprocessing, conducting exploratory data analysis (EDA), and effectively communicating your insights.

Dataset Overview:

Source and Description of the Dataset:

The dataset used for this assignment is titled "Electric Vehicle Population Data" and can be found on Data.gov: https://catalog.data.gov/dataset/electric-vehicle-population-data

Provided by the State of Washington, this dataset displays information about battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) currently registered through the Washington State Department of Licensing. Data is separated into 17 different columns, showing each vehicle's VIN, county and city of registration, make and model, electric type and electric range. Vehicle model years range from 2013 to the current year, with metadata being routinely updated by the Washington government.

Requirements:

Provide answers to the following questions as possible as you can. Provide a brief description, including the number of examples, number and type of features, and context.

Data Cleaning and Feature Engineering:

1. **Document Missing Values:** Check for missing values and document their frequency and distribution across features.
2. **Missing Value Strategies:** If missing values are present, apply multiple strategies (e.g., mean/median imputation, dropping rows) and compare their impact on the analysis.
3. **Feature Encoding:** Encode categorical features (e.g., Make, Model) using techniques like one-hot encoding.

4. **Normalization:** Normalize numerical features if necessary for chosen analysis methods.

Exploratory Data Analysis:

5. **Descriptive Statistics:** Calculate summary statistics (mean, median, standard deviation) for numerical features.
6. **Spatial Distribution:** Visualize the spatial distribution of EVs across locations (e.g., maps).
7. **Model Popularity:** Analyze the popularity of different EV models (categorical data) and identify any trends.
8. Investigate the relationship between every pair of numeric features. Are there any correlations? Explain the results.

Visualization:

9. **Data Exploration Visualizations:** Create various visualizations (e.g., histograms, scatter plots, boxplots) to explore the relationships between features.
10. **Comparative Visualization:** Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts.

Additional Analysis:

11. **Temporal Analysis (Optional):** If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.

Submission:

A- A comprehensive report (**5-8 pages**) that describes the dataset and summarizes and discusses all the results and findings as required above.

B- Your code (python code) in either .py format or a Jupyter Notebook with both the code and visualizations.

C- Please compress your files, including both the code and the report, into a single zip file and submit it to the ritaj before the deadline. The file name should follow this format: "LastName_ID_Student1_LastName_ID_Student2.ZIP".

D- Late submissions will be accepted up to 3 days after the deadline, with a 10% deduction for each day delayed.

Hint:

You can use the following python libraries: Pandas, NumPy, Matplotlib, Seaborn