Machine Learning
Assignment 1 - Data Preprocessing

Vikas Thammanna Gowda

Due: 02/23/2025

Instructions

Collaboration Policy

This is an individual assignment. You may discuss concepts, problem formulations, and approaches to solving the problems, but you must write your own code and explanations.

- You are **not allowed** to share solutions, source code, or exact approaches.
- Any external sources (books, online resources, discussions, etc.) you refer to must be cited in your write-up.
- You do not need to cite course lecture notes, textbooks, or materials provided as part of the course.

Assignment Structure

Your submission consists of two parts:

Coding Component (Submit as `<your_name>_PA01.ipynb`)

- Use Markdown cells in Jupyter Notebook to add each question before solving it.
- Write clean, readable, and well-commented code.
- Define functions for repetitive tasks instead of redundant code.
- Ensure all visualizations are clear, properly labeled, and provide meaningful insights.
- Use at least two different types of visualizations for each data exploration question (e.g., histogram and box plot).

Report Write-up (Submit as `<your_name>_PA01.pdf`)

- Add each question to your write-up before answering them.
- The report should mirror the coding component and provide interpretations of results.
- Use Times New Roman, size 14 for questions, size 12 for answers.
- Ensure the document is justified and structured.
- Include properly labeled figures and tables, centered with captions.
- All the plots must be complete, be of the same size, and be centered with a figure number and a figure name.
- Clearly explain decisions regarding missing data handling, feature selection, scaling, and outlier removal.

Dataset Description

Attribute	**Description**
`id`	The unique car identifier
`region`	Area location of the car
`price`	The price of the car in USD
`year`	Year of manufacture
`manufacturer`	Company name of the manufacturer
`model`	The model of the car
`condition`	Condition of the car
`cylinders`	Total number of cylinders
`fuel`	Type of the fuel used
`odometer`	Odometer reading
`title_status`	Status of the title of the car
`transmission`	Type of transmission used in the car
`VIN`	Vehicle Identification Number
`drive`	Drive train used
`size`	Size of the car
`type`	Type of the car
`paint_color`	Color of the car
`county`	The county where the car is located
`state`	The state where the car is located
`lat`	Latitude
`long`	Longitude
`posting_date`	Date when the car was posted

Table 1: Description of Car Dataset Attributes

Assignment Questions: Coding vs. Write-Up

Each question involves both a coding component (implementation) and a write-up component (interpretation).

1. Handling Column-wise Missing Values

Coding:

- Compute the percentage of missing values in each column.
- Drop columns with more than 51% missing values.
- Display the percentage of missing values before and after dropping columns.

Write-up:

- Add a table showing the columns and their corresponding missing value percentages that you are dropping.

2. Feature Selection

Coding:

- Retain relevant features.
  Note: These features must be retained at a minimum - **year, manufacturer, condition, cylinders, fuel, odometer, title_status, type, and price**.
- Drop irrelevant columns.

Write-up:

- Explain the rationale for keeping and dropping specific columns.

3. Cleaning Up the Dataset

Coding:

- Drop rows with any remaining missing values.
- Display missing value counts before and after dropping rows.

Write-up:

- Did data loss affect the dataset significantly?

4. Grouping Manufacturers

Coding:

- Group manufacturers under their parent companies using the provided 14-company mapping. `https://www.visualcapitalist.com/14-companies-control-entire-auto-industry/`
- Categorize unlisted manufacturers as "Others."
- Display value counts before and after grouping.

Write-up:

- Explain how generalizing manufacturers improves data quality.
- Did the grouping simplify the dataset meaningfully?

5. Converting Data Types

Coding:

- Convert price, cylinders, and odometer to float.
- Plot distributions for each feature.

Write-up:

- Describe challenges in type conversion.
- Explain trends observed in distributions.

6. Assigning Ratings for Title and Condition

Coding:

- Assign ratings on a scale of 0.1 to 1.0 for title_status and condition.
- Convert both to float types.
- Plot distributions.
- Compute correlation between title_status and condition.

Write-up:

- Justify rating scale choices.
- Discuss correlation results.

7. Identifying and Treating Outliers

Coding:

- Detect outliers in numerical features.
- Use visualizations (e.g., box plots) to identify outliers.
- Apply appropriate outlier treatment.

Write-up:

- Explain why outliers were treated in a specific way for each of the numerical features.
- Compare distributions before and after treatment.

8. Feature Scaling

Coding:

- Apply different feature scaling methods.
- Visualize distributions before and after scaling.

Write-up:

- Justify which scaling method is best for each numerical feature.
- How did scaling affect feature distributions?

Grading Rubric (100 Points)

Category	**Criteria**	**Points**
Code Quality	Code is well-structured, commented, and readable.	10
Visualization Quality	Plots are clear, labeled, and informative.	10
Handling Column-wise Missing Values	Correctly identifies and drops necessary columns.	5
Feature Selection	Justifies removal of irrelevant features.	15
Data Cleaning	Drops missing rows effectively.	5
Manufacturer Grouping	Generalizes manufacturer names correctly.	10
Data Type Conversion	Converts price, cylinders, and odometer.	10
Rating System	Assigns appropriate ratings, checks correlation.	15
Outlier Treatment	Correctly detects and treats outliers.	15
Feature Scaling	Applies appropriate scaling method per feature.	15
Total	**Final Score**	**100**

Table 2: Grading Rubric

Final Submission Checklist

- Jupyter Notebook (`.ipynb`)
- PDF Write-up (`.pdf`)
- All visualizations included
- Formatted report with proper justifications
- Sources cited where applicable