

Portfolio 1: Analysis of a car sell record Dataset

Overview

This Jupyter Notebook focuses on cleaning and preprocessing a dataset of car sales. The dataset contains multiple features including car ownership, fuel type, and seller type, among others. The main objective of the notebook is to remove outliers and filter the data based on specific rules. In this Read Me file is the overview of the portfolio and the main findings.

Dataset Description

The dataset used in this notebook contains the following key columns:

1. **Owner:** Describes the ownership type, such as 'First Owner', 'Second Owner', etc.
2. **Fuel:** Specifies the fuel type, such as 'Diesel' or 'Petrol'.
3. **Seller Type:** Indicates whether the seller is an 'Individual' or a 'Dealer'.
4. **Other columns:** Additional information about the cars, such as price, model, etc.

Notebook Structure

1. Data Loading

The dataset is loaded and the initial structure is examined to ensure it is in a usable format.

2. Outlier Removal

Outliers are removed in a series of steps based on the following rules:

- Only cars with ownership of 'First Owner', 'Second Owner', or 'Third Owner' are retained.
- Only 'Diesel' and 'Petrol' fuel types are kept.
- Only sellers that are 'Individuals' or 'Dealers' are retained.

Each step prints the size of the dataset after applying the corresponding filter.

3. Final Cleaned Dataset

After all outliers are removed, the final cleaned dataset is displayed and is ready for further analysis or modeling.

Conclusion for Portfolio 1: Analysis of Car Sales Data

The portfolio effectively demonstrated the process of cleaning, preprocessing, and visualizing a car sales dataset. The key insights from the visualizations provide a deeper understanding of how various features impact the selling price of cars:

1. **Year vs Selling Price:** The analysis revealed that newer cars generally have higher selling prices, which aligns with the expected depreciation of cars over time. Older cars show a wider distribution in selling prices, indicating variability based on other factors such as condition and mileage.
2. **Seller Type vs Selling Price:** The plot comparing seller types showed that cars sold by dealers tend to have higher prices than those sold by individuals. This is likely due to added services such as warranties or repairs offered by dealers, making their cars more expensive.

Through the steps of data cleaning, including outlier removal and filtering by key attributes like fuel type and ownership, the dataset was refined for further analysis. These insights can be used for developing models to predict car prices or to make informed decisions in the used car market.

Usage

To run the notebook, open it in Jupyter Notebook or any compatible environment and execute the cells in sequence. The final output will be a cleaned dataset that has the following dimensions: `3657 rows and 8 columns`.

Files Included

- `48622613_Portfolio1.ipynb` - The Jupyter Notebook with all the code and analysis.
- `data/` - A folder containing the dataset `car_sell.csv`

lums`.

Dependencies

- Python 3.x
- Pandas
- Jupyter Notebook

Ensure all required packages are installed by running:

```
pip install pandas jupyter
```

;

