**Car Price Analysis and Modeling Project Documentation**

**Step 1: Understanding the Dataset**

Overview:

* We have a dataset with 10,000 rows and 10 columns.
* The dataset is clean right from the start – there are no missing values!

Data Types:

* Categorical Variables:
  + Brand
  + Model
  + Fuel Type
  + Transmission
* Numerical Variables:
  + Year (manufacturing year)
  + Engine Size
  + Mileage
  + Doors
  + Owner Count
  + Price

**Step 2: Data Preparation**

Loading the Data:

* We read in the dataset using pandas.
* We display the first few rows to quickly inspect our data.

Data Cleaning:

* We verify that all entries are complete and there are no missing values.
* We also check for any formatting issues and potential outliers, which will be useful later on.

**Step 3: Exploratory Data Analysis (EDA)**

**3.1 Summary Statistics and Distributions**

* Year:  
  + The cars range between 2000 and 2023, with a mean around 2011. This suggests that many of the cars might be older, with a gradual shift towards newer models.
* Engine Size:  
  + Ranges from 1.0 to 5.0, with a mean of approximately 3.0. It's intriguing to see if larger engines relate to higher prices.
* Mileage:  
  + Values spread from 25 to 300,000. A high standard deviation in mileage (around 86,322) indicates there might be some outliers.
* Doors and Owner Count:  
  + Doors typically vary between 2 to 5, and the number of previous owners ranges from 1 to 5.
* Price:  
  + Prices range from roughly 2,000 to 18,000, with a mean of about 8,853 and a very similar median (≈ 8,858), suggesting a roughly symmetric distribution.

**3.2 Boxplot Analysis**

Mileage Boxplot:

* Range: 0 to 300,000 miles.
* Quartiles:
  + Q1 (25th percentile): around 100,000 miles
  + Median (50th): approximately 150,000 miles
  + Q3 (75th): about 200,000 miles
* Observation:
  + Most cars are used (100,000–200,000 miles), but we do have some very low and very high mileage entries—these outliers might be special cases or require further investigation.

Price Boxplot:

* Range: 2,000 to 18,000.
* Quartiles:
  + Q1: about 8,000
  + Median: 10,000
  + Q3: 12,000
* Observation:
  + While most cars are priced between 8,000 and 12,000, the extremes (2,000 and 18,000) suggest unique cases worth a closer look.

**3.3 Correlation Analysis**

For Numerical Variables:

* Strong Positive Correlation:
  + Year (≈0.66): Newer cars tend to be priced higher.
* Strong Negative Correlation:
  + Mileage (≈–0.55): More miles mean lower car prices.
* Moderate Correlation:
  + Engine Size (≈0.36) shows a moderate correlation with price.
* Minimal Correlation:
  + Owner Count and Doors – almost no linear relationship with Price.

For Categorical Variables:

* We calculate the correlation ratio (η) to understand the influence of these factors:  
  + Brand (η ≈ 0.04): Explains only about 4% of price variability.
  + Model (η ≈ 0.05): Also explains very little.
  + Fuel Type (η ≈ 0.26): Moderately associated with price.
  + Transmission (η ≈ 0.25): Also shows a moderate relationship.

**Step 4: Modeling**

**4.1 Selecting a Model**

* We’ll start with a Linear Regression model, popular for its simplicity and interpretability.
* Linear regression assumes a linear relationship between our features and the target (Price), making it a great baseline model.

**4.2 Building the Model**

* Features Used:
  + Engine Size, Mileage, Fuel Type, and Transmission
* Preprocessing:
  + We use one-hot encoding for the categorical variables (Fuel Type and Transmission) so they can be included in the regression.
* Workflow:
  + Split the dataset into training (80%) and testing (20%) sets.
  + Construct a pipeline that takes care of pre-processing and model building in one go.

**4.3 Evaluating the Model**

* Metrics:  
  + We evaluate using the Mean Squared Error (MSE).
  + For our model, an MSE of around 4,258,241.90 was calculated, and the Root Mean Squared Error (RMSE) came out to roughly 2,063.6.
* Interpretation:
  + The RMSE tells us, on average, our predictions are off by about 2,063.6 units (for example, dollars).
  + Depending on the scale of prices, this error might be acceptable or could suggest areas for further refinement.

**Step 5: Conclusion**

* Key Findings:
  + The year of the car and mileage are the most influential factors—newer cars have higher prices, and higher mileage tends to lower the price.
  + Other variables, like engine size, fuel type, and transmission, also play roles though to a lesser extent.
* Market Implications:
  + The typical car in this dataset is priced between 8,000 and 12,000, with our analysis suggesting a good baseline for further market studies.
* Future Improvements:
  + Further fine-tuning could involve exploring more advanced models, additional feature engineering, or incorporating other variables that might explain price variations better.
* Statistical Insights:
  + Our analysis includes a solid confidence interval and a t-test that highlights how electric cars differ significantly in price compared to petrol cars.

**References**

* National Institute of Standards and Technology. (n.d.). *NIST/SEMATECH e-Handbook of Statistical Methods.* Retrieved from [NIST Handbook](https://www.itl.nist.gov/div898/handbook/)
* Wikipedia contributors. (n.d.). *Correlation ratio.* In *Wikipedia, The Free Encyclopedia.* Retrieved from [Wikipedia: Correlation Ratio](https://en.wikipedia.org/wiki/Correlation_ratio)
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* New York, NY: Springer.





















