IS4116 Business Intelligence Systems

Assignment: Data Analytics Process and Interpretation

W.D Kumudika | 20020597

Github-:https://github.com/DinithKumudika/IS4116---Business-Intelligence-Systems---Assingnment-2
Dataset-: https://www.kaggle.com/datasets/alikalwar/uae-used-car-prices-and-features-10k-listings

Dataset Domain

- automotive industry specifically focusing on the **used car market in the UAE**.

Dataset Description

- The dataset contains 10,000 used car listings from the UAE market with detailed features, pricings, and conditions.
- Dataset include realistic price variations based on UAE market conditions
- Dataset has 11 columns (features): Make, Model, Year, Price, Mileage, Body type, Cylinders, Transmission, Fuel type, Color, Location, Description

Business problem

- What are the key factors influencing used car prices in the UAE market, and build a model to predict the valuation of a used car based on its features?

Business challenges that can be addressed

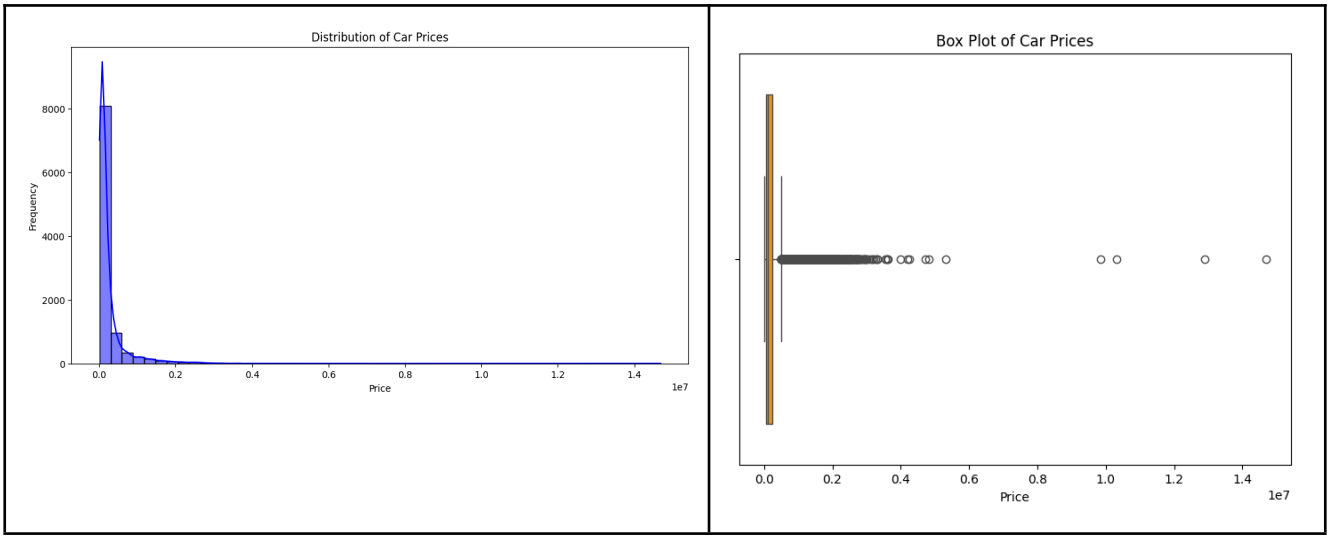
- Buyers want to ensure they are paying a fair price for a used car based on its features. Understanding these price drivers helps buyers make informed decisions.
- Sellers need to price their cars competitively to attract buyers while maximizing profit. Insights into pricing trends can help sellers set optimal prices.
- Dealers want optimized pricing strategies to stay competitive and understanding market trends helps dealerships stock in-demand cars.
- Customers need the best vehicle for the price they are willing to pay and customers can identify the best deals based on market trends.

Analytical Process

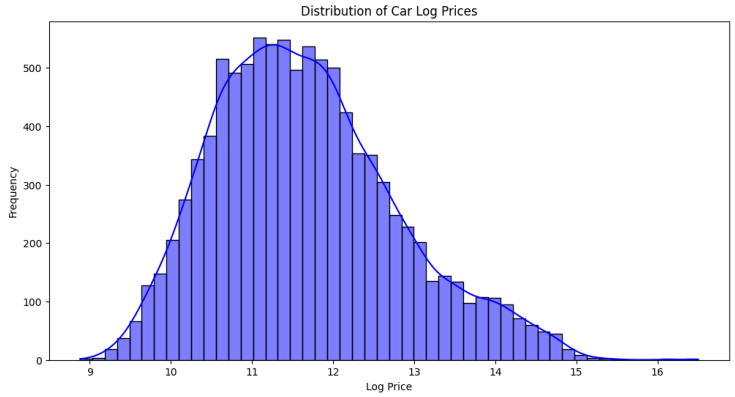
- Data Collection-: extracted dataset from Kaggle through kagglehub library.
- Data Processing-: preprocessed the dataset by handling missing values, removing duplicates, handling outliers (Price), Feature Engineering, encoding categorical features, standardizing numerical features and removing irrelevant columns (Location, Description)
- EDA-: visualize categorical and numerical features to identify their relationship with the price of vehicles.
- Statistical Analysis-: a linear regression model to predict the vehicle price.

Exploratory Data Analysis

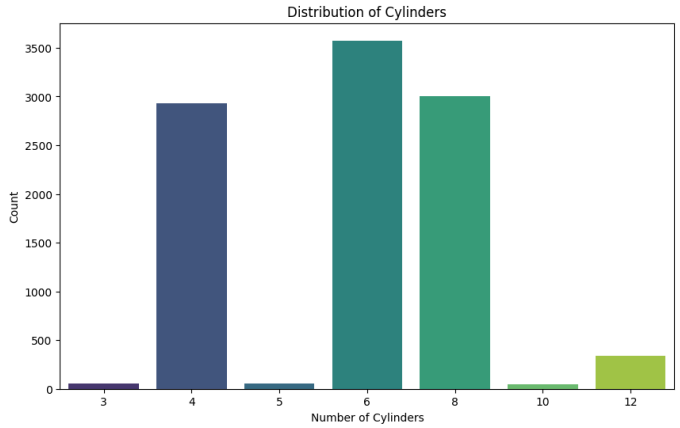
Distribution of Car Prices - histogram and box plot was used to analyze distribution of vehicle prices.



Since car price distribution contained many outliers log transformation was applied

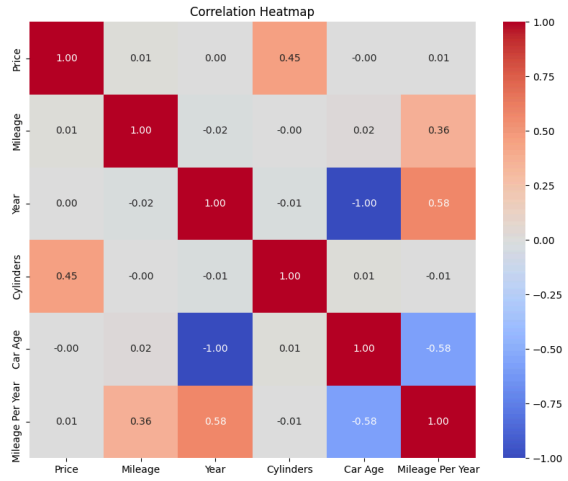


Distribution of Cylinders - since no of cylinders is generally a critical factors for determining car prices it was analyzed find no of cylinders in most of the vehicles



Correlation between numerical features - correlation matrix and heatmap was used to identify patterns and relationships between numerical features

	Price	Mileage	Year	Cylinders	Car Age	Mileage Per Year
Price	1.000000	0.010685	0.000188	0.446045	-0.000188	0.005784
Mileage	0.010685	1.000000	-0.017118	-0.002258	0.017118	0.360396
Year	0.000188	-0.017118	1.000000	-0.010376	-1.000000	0.576625
Cylinders	0.446045	-0.002258	-0.010376	1.000000	0.010376	-0.007067
Car Age	-0.000188	0.017118	-1.000000	0.010376	1.000000	-0.576625
Mileage Per Year	0.005784	0.360396	0.576625	-0.007067	-0.576625	1.000000



*** more data analysis steps and visualizations are available in data_analysis.ipynb**

Feature Engineering

- Z-score scaling was used to standardize numerical features (year, mileage, price, cylinders, car age and mileage per year)
- Used log transformation to to transform right-skewed car distribution into symmetrical to reduce the impact of outliers
- Used,
    - Label encoding encode transmission
    - One hot encoding to encode fuel type and body type
    - Target encoding to encode make, color and model

Regression Modeling

Linear regression modeling was used to predict car price based on transmission, fuel type, body type, make, model, color, year, mileage per year, cylinders, car age features as predictors.

Evaluation-:

MAE-: 102445.09118158213	MSE: 39631054027.52635	R^2: 0.8488379010874164