Car Price Prediction Using Data Analysis of Quickr Data and Machine Learning

Akash Banik (M25MAC001), Ali Asad Quasim (M25MAC002)

Abstract

This report presents a thorough data analysis performed on a used car dataset to support the development of a machine learning model for predicting car prices. The focus is on understanding data quality, distributions, relationships between features, and deriving insights useful for predictive modeling.

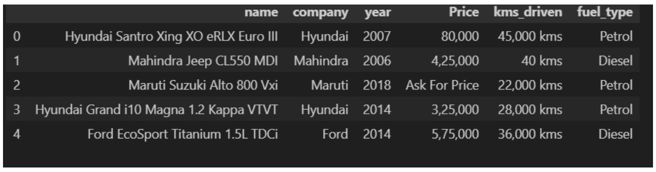
1 Introduction

Car price prediction is a valuable application of machine learning, where the goal is to estimate the market value of used cars based on multiple attributes such as brand, model, year of manufacture, fuel type, and kilometers driven. This project utilizes a dataset containing various features of used cars along with their prices. This data was made available by data scraping from Quickr website of used cars. A robust exploratory data analysis (EDA) provides the foundation for building accurate predictive models by revealing patterns, data issues, and relationships.

2 Data Understanding and Cleaning

2.1 Dataset Description

The scraped dataset from **Quickr** contains the following main columns/features:



	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing XO eRLX Euro III	Hyundai	2007	80,000	45,000 kms	Petrol
1	Mahindra Jeep CL550 MDI	Mahindra	2006	4,25,000	40 kms	Diesel
2	Maruti Suzuki Alto 800 Vxi	Maruti	2018	Ask For Price	22,000 kms	Petrol
3	Hyundai Grand i10 Magna 1.2 Kappa VTVT	Hyundai	2014	3,25,000	28,000 kms	Petrol
4	Ford EcoSport Titanium 1.5L TDCi	Ford	2014	5,75,000	36,000 kms	Diesel

Figure 1: Sample view of the dataset

- **Car Name**: Brand and model name of the car.
- **Year**: Year of purchase/manufacture of the car.
- **Kilometers Driven**: Total kilometers the car has been driven.
- **Fuel Type**: Type of fuel used (Petrol, Diesel, CNG, etc.).
- **Seller Type**: Whether it is a dealer or an individual seller.
- **Price**: The selling price of the car (target variable).

2.2 Data Cleaning

Ensuring data reliability is a necessary step before performing meaningful analysis or machine learning. The raw dataset exhibited issues such as formatting inconsistencies, missing values, and extreme outliers. The cleaning procedure consisted of the following:

- **Removal of redundant fields**: Export-generated or index-like columns that carried no analytical value were dropped.
- **Handling missing values**: Rows with missing or zero *Price* values were removed since the target variable must be valid. Numerical fields with few missing entries were imputed using the median, while missing categorical values were filled using the mode.
- **Cleaning numeric fields stored as text**: Features such as *Price* and *Kilometers Driven* initially contained non-numeric characters (e.g., “km”, “kms”, “₹”). These were cleaned by removing the unwanted characters and converting the results into proper numerical types. Invalid values, such as negative mileage or unrealistic car ages, were filtered out.
- **Standardizing categorical variables**: Categories such as Fuel Type, Seller Type, and Transmission contained inconsistent spellings and formatting (e.g., different casings or variants of brand names). These were standardized to ensure consistent grouping during analysis.
- **Duplicate removal**: Identical listings containing the same brand, model, year, kilometers driven, and price were removed to avoid over-representation of specific samples.
- **Outlier detection**: Outliers in numerical features such as *Price* and *Kilometers Driven* were removed using the Interquartile Range (IQR) method. Values lying outside the range

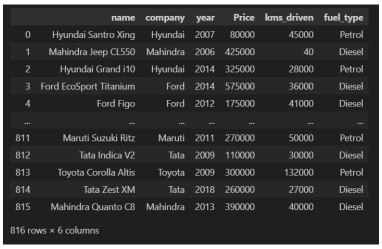
  $$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

  were treated as extreme and excluded to reduce skewness and improve model stability.

After completing these steps, the dataset became consistent and ready for exploratory analysis and predictive modeling.

Table 1: Summary Statistics for Price and Kilometers Driven

Statistic	Price	kms_driven
Mean	325692.59	41425.237
Median	284999.0	40000.0
Skewness	0.864037	0.357494
Kurtosis	0.27225	0.022601
IQR	281999.0	28000.0
Mode	250000	45000



	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	Ford Figo	Ford	2012	175000	41000	Diesel
...	...	...	...	...	...	...
811	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
812	Tata Indica V2	Tata	2009	110000	30000	Diesel
813	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
814	Tata Zest XM	Tata	2018	260000	27000	Diesel
815	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

816 rows × 6 columns

Figure 2: Sample view of the cleaned dataset

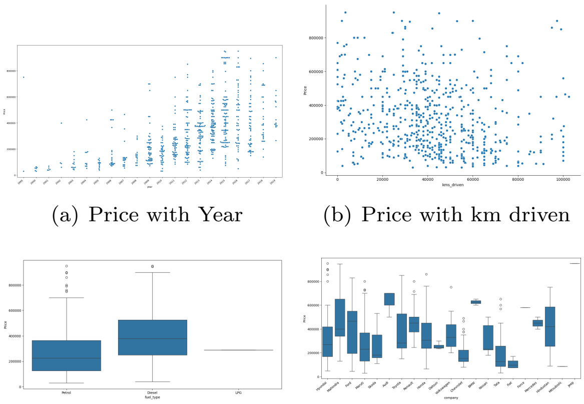
3 Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics

Summary statistics such as mean, median, standard deviation, minimum, and maximum values provide insights into the distribution and variability of the features, especially the target variable *Price*.

3.2 Categorical Features Distribution

Pie charts or bar plots are useful to visualize the frequency of categorical variables such as *Fuel Type*, *Seller Type*, and *Transmission*. This helps understand which categories dominate the dataset.



(a) Price with Year

(b) Price with km driven

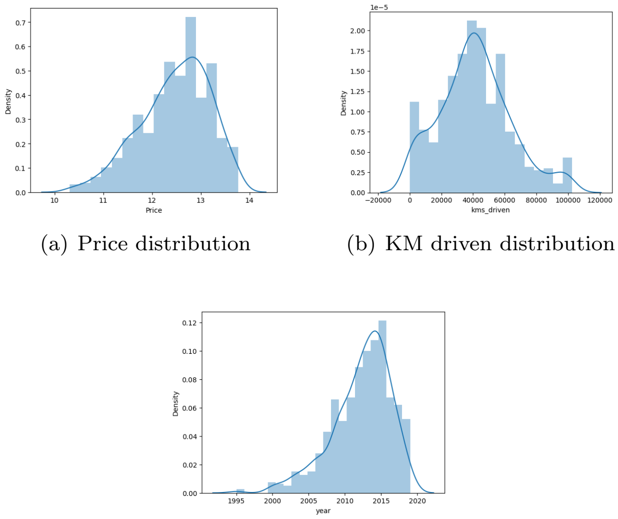
(c) Price with fuel type

(d) Price with company

Figure 3: Comparison plots of price, kilometers driven, fuel type and years driven

3.3 Numerical Features Distribution

Histograms and box plots are employed for visualizing the spread of continuous variables like *Kilometers Driven* and *Price*. For example, prices typically show right-skewed distributions with some very expensive cars.



(a) Price distribution

(b) KM driven distribution

(c) Years driven distribution

Figure 4: Distribution plots of price, kilometers driven, and years driven

3.4 Correlation Analysis

Correlation matrices between numerical features reveal relationships that can influence the price prediction. Typically, features like car age and kilometers driven are negatively correlated with price, while brand reputation might show positive effects.

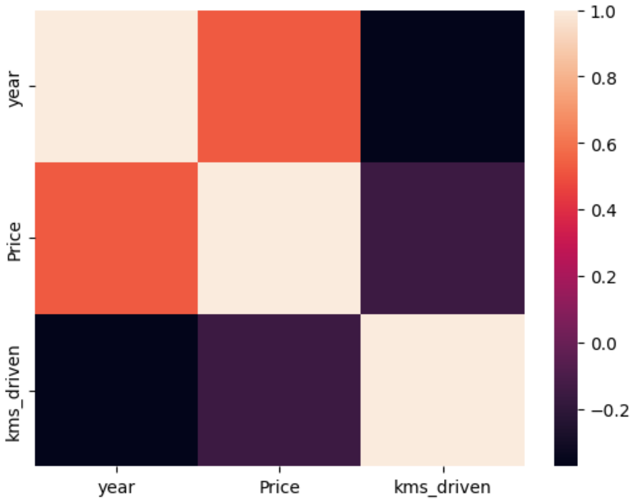


Figure 5: Correlation heatmap of numerical attributes.

3.5 Feature Relationships

Scatter plots of price versus key features help reveal the nature of dependencies. For example, plotting price against year or kilometers driven uncovers trends useful for model selection.

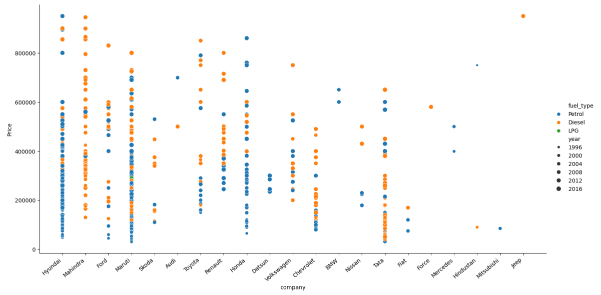


Figure 6: Price variation with car age, fuel type and kms driven

3.6 Model Building and final prediction

This project focuses on developing a predictive model to estimate the selling price of used cars using features such as **year of manufacture**, **kilometers driven**, **company**, and **fuel type**. The original dataset contained inconsistencies like missing values, non-numeric entries in the price and kilometers driven columns, and categorical variables with many unique values. To address this, extensive data preprocessing was applied. Non-numeric price entries such as "Ask For Price" were removed, price and kilometers driven fields were cleaned and converted to numeric types, and missing or invalid values were filtered out. Categorical variables were normalized by limiting car names to the first few tokens and encoding companies and fuel types via **OneHotEncoding** for effective model ingestion.

A machine learning pipeline was constructed combining the encoding of categorical features and inclusion of numerical predictors. The pipeline employed a **linear regression** model to learn the relationship between input features and car prices. The cleaned dataset was split into training and testing sets to evaluate model performance objectively. The resulting model achieved a high **$\mathbf{R^2}$ score** of approximately **0.92** on the test set, indicating that the model explains 92% of price variance. This result underscores the importance of rigorous data cleaning and feature engineering. Future work may involve exploring regularized linear models or ensemble methods to capture nonlinear patterns and further enhance predictive accuracy.

4 Summary and Insights

The exploratory data analysis indicates the following key points relevant for model building:

- **Strong influence of car age and kilometers driven**: Older cars and those driven longer generally have lower prices.
- **Fuel type impact**: Diesel cars might have higher prices on average compared to petrol or CNG.

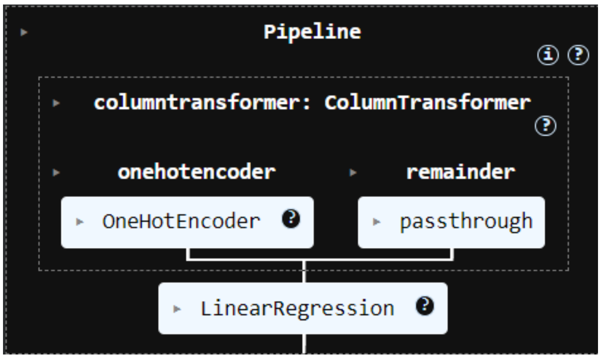


Figure 7: Model Pipeline

- **Categorical variables importance**: Transmission type, seller type, and ownership history play a noticeable role in pricing.
- **Presence of outliers**: Extreme values in price and kilometers should be treated carefully either by transformation or removal.
- **Data consistency**: Proper cleaning and encoding of categorical variables are critical before feeding data into models.

These observations ensure a more informed approach to feature engineering, selection, and model tuning for accurate price prediction.

5 References

- Machine Learning Basics: Géron, Aurélien. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow." 2nd Edition, 2019.
- Exploratory Data Analysis Techniques: Tukey, John W. "Exploratory Data Analysis." 1977.

Appendix

- **Project Repository:** `https://github.com/aliasad20/Car-Price-Predictor`
- **Dataset Link:** `https://drive.google.com/file/d/1zrlMEQupTlejy28RVYx9_LOzBMXAHrGF/view?usp=sharing`