💎🚗 SmartCar Price Predictor

Yousef Barakat
Machine Learning Engineer
Data Analysis | Predictive Modeling

In the used car market, pricing is often inconsistent and unclear.
This project builds a complete machine learning system that analyzes market patterns and predicts car prices with high accuracy.

❓ Why do two cars with similar specifications have completely different prices?

🤔Is it brand?

🤔Mileage?

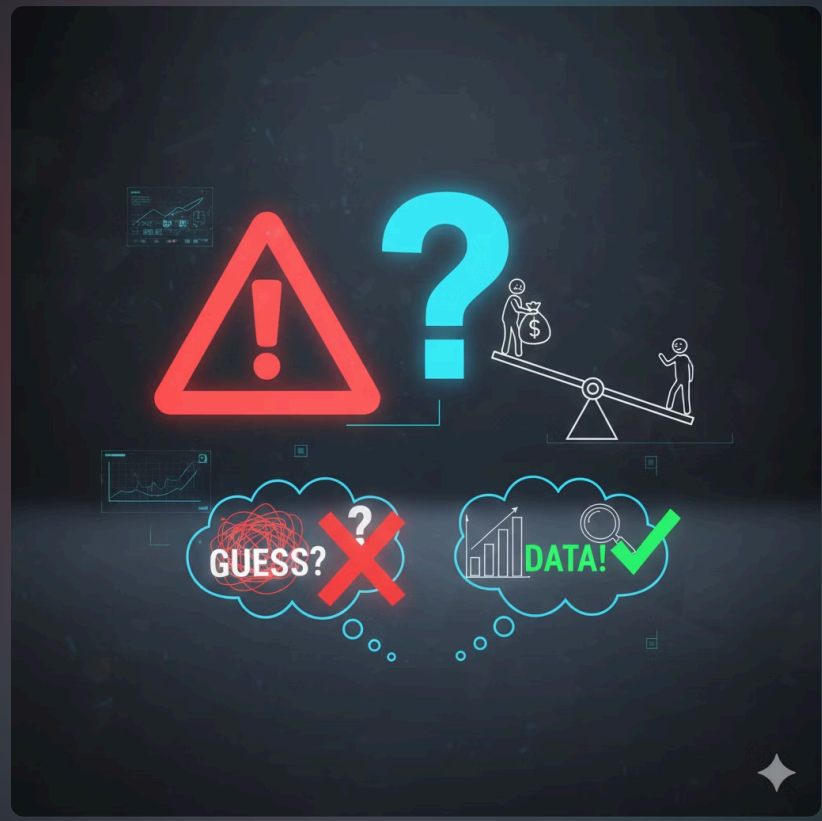
🤔Engine power?

🤔Market demand?

🤔Or hidden patterns we cannot see?

🚨 Problem Statement

The used car market lacks a transparent and reliable pricing system.
Buyers often overpay, and sellers struggle to determine the fair market value of their vehicles.
Pricing decisions are usually based on guesswork rather than data.



📌 Project Overview

> This project delivers a complete end-to-end machine learning solution for analyzing and predicting used car prices.

🔹 The system includes:

- 📊 Market Data Analysis
- 🧠 Predictive Machine Learning Model
- ⚙️ Hyperparameter Optimization
- 📈 Interactive Dashboard
- 🚀 Real-Time Price Prediction App

DATASET STRUCTURE

Understanding Our Car Data

Our dataset comprises critical attributes that directly influence a car's selling price, providing a rich foundation for our predictive model. Each column offers a unique insight into the vehicle's characteristics and market position.

- **Name:** Full car model name
- **Year:** Manufacturing year
- **Selling Price:** Target variable
- **KM Driven:** Total kilometers on odometer
- **Fuel:** Type of fuel used
- **Seller Type:** Individual, Dealer, Trustmark Dealer
- **Transmission:** Manual or Automatic
- **Owner:** Number of previous owners
- **Mileage:** Fuel efficiency (kmpl/km/kg)
- **Engine:** Engine displacement (CC)
- **Max Power:** Engine's maximum power (bhp)
- **Seats:** Number of seating capacity
- **Torque:** Engine's rotational force (dropped due to high missing values)
- **Brand:** Extracted car brand

Project Phases: A Detailed Checklist

Our journey to predict car prices is structured into distinct, logical phases, each building upon the last to ensure robust and accurate results.

1

Environment & Data Loading

Setting up the workspace and bringing in the raw data.

- Importing Libraries
- CSV Data Ingestion
- Initial Data Inspection

2

Data Cleaning

Refining the dataset to ensure quality and consistency.

- Unit Stripping (Regex)
- Missing Value Imputation
- Duplicate Removal

3

Exploratory Data Analysis (EDA)

Uncovering patterns and insights from the cleaned data.

- Univariate Distribution
- Bivariate Analysis
- Correlation Heatmaps
- Top Brands Analysis

4

Feature Engineering

Transforming and creating new features to enhance model performance.

- Car Age Calculation
- Ordinal & One-Hot Encoding
- Target & Features Splitting
- Feature Scaling

5

Machine Learning Modeling

Building, optimizing, and selecting the best predictive model.

- Multiple Model Testing
- Hyperparameter Tuning (GridSearch)
- Best Model Selection (Random Forest)
- Model & Scaler Export

6

Deployment & Application

Bringing the model to life with interactive tools for real-time predictions.

- Interactive Dashboard (Plotly)
- Streamlit Web Interface
- Real-time Prediction Engine

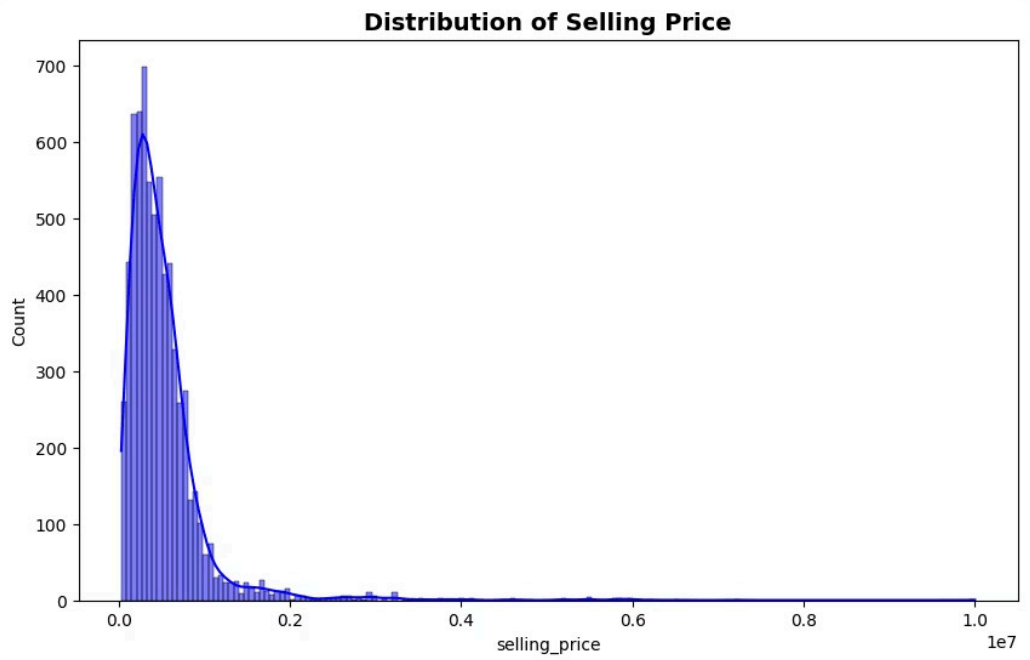
🧹 Data Cleaning & Preparation

Goal: Transforming messy raw data into a ML-ready structured format.

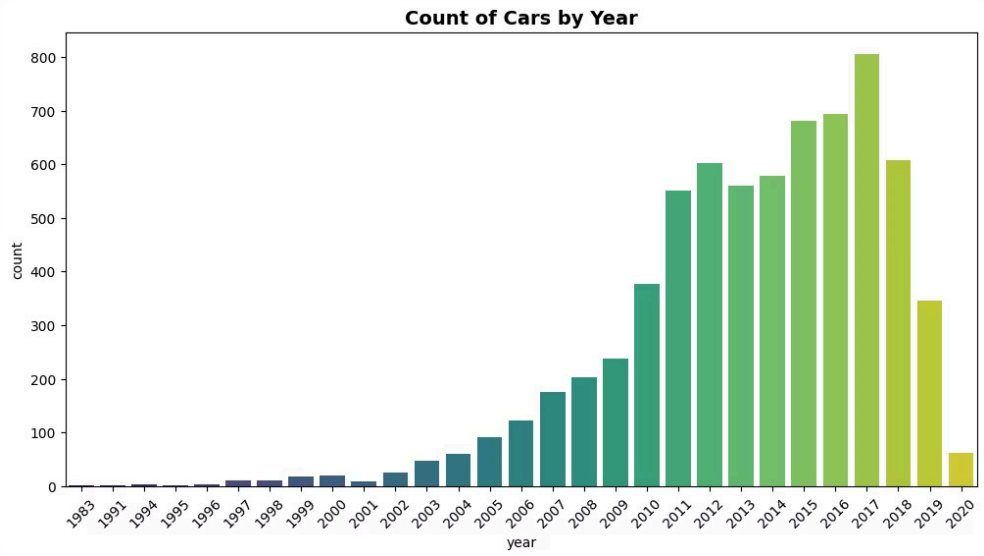
- ✂️ **String Manipulation**
  - **Problem:** Units like (kmpl, CC, bhp) in numeric columns.
  - **Action:** Used **Regex** to strip text and convert to `float`.
  - *Example: "1248 CC" → 1248.0*
- 🩹 **Missing Values**
  - **Problem:** Nulls in performance metrics.
  - **Action:** Applied **Median Imputation** for robust filling.
- 🚫 **Duplicates & Logic Filtering**
  - **Action:** Removed redundant entries to prevent **Overfitting**.
  - **Sanity Check:** Filtered out unrealistic data (e.g., 0 mileage, >14 seats).
- 🏷️ **Feature Extraction**
  - **Action:** Extracted **Brand Name** from the full car string to simplify categories and identify market leaders.

Key Data Insights (EDA) 📊

After analyzing the data, these five visualizations provided the most critical insights into the car market:

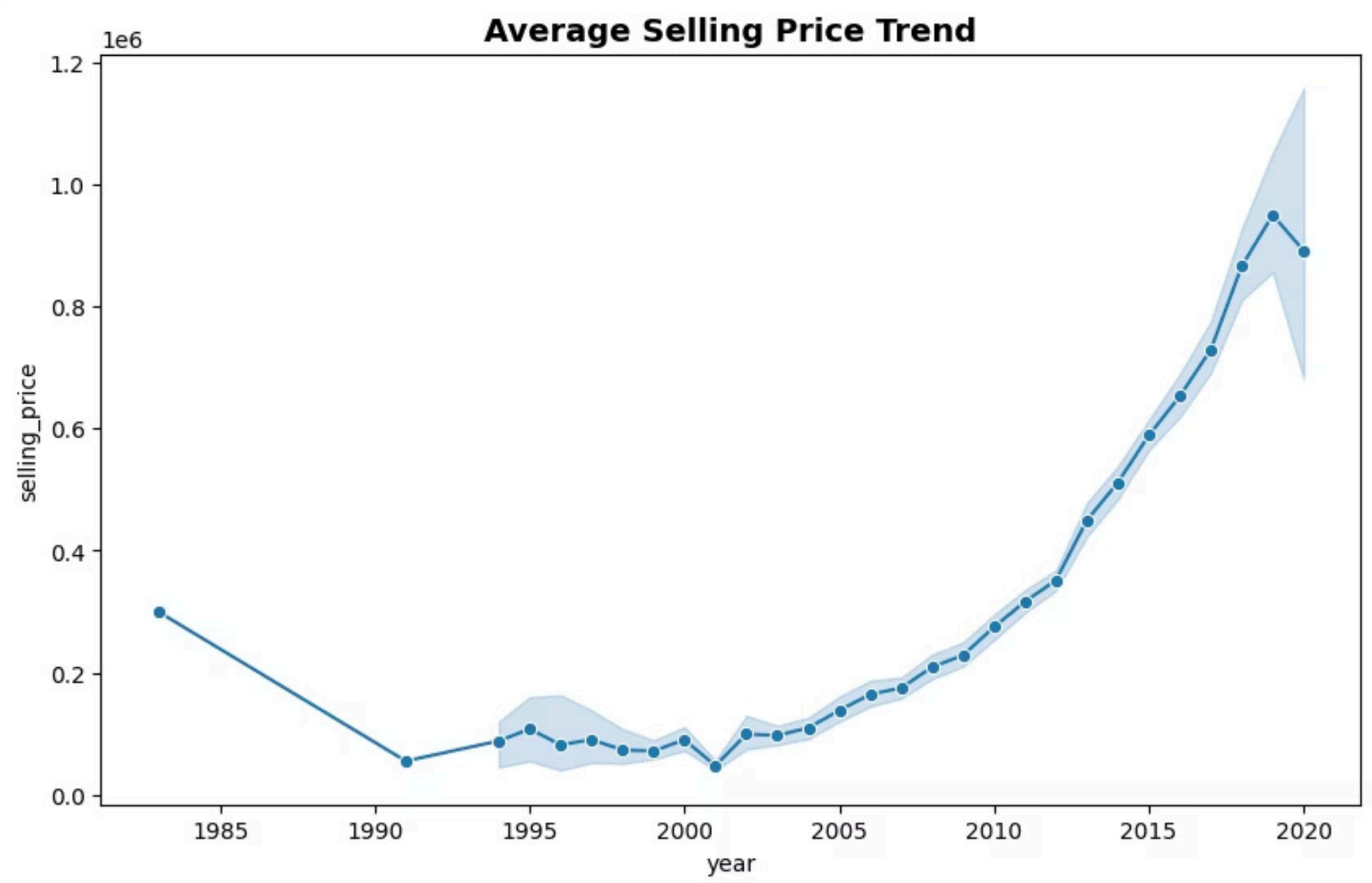


💰 Distribution of Selling Price

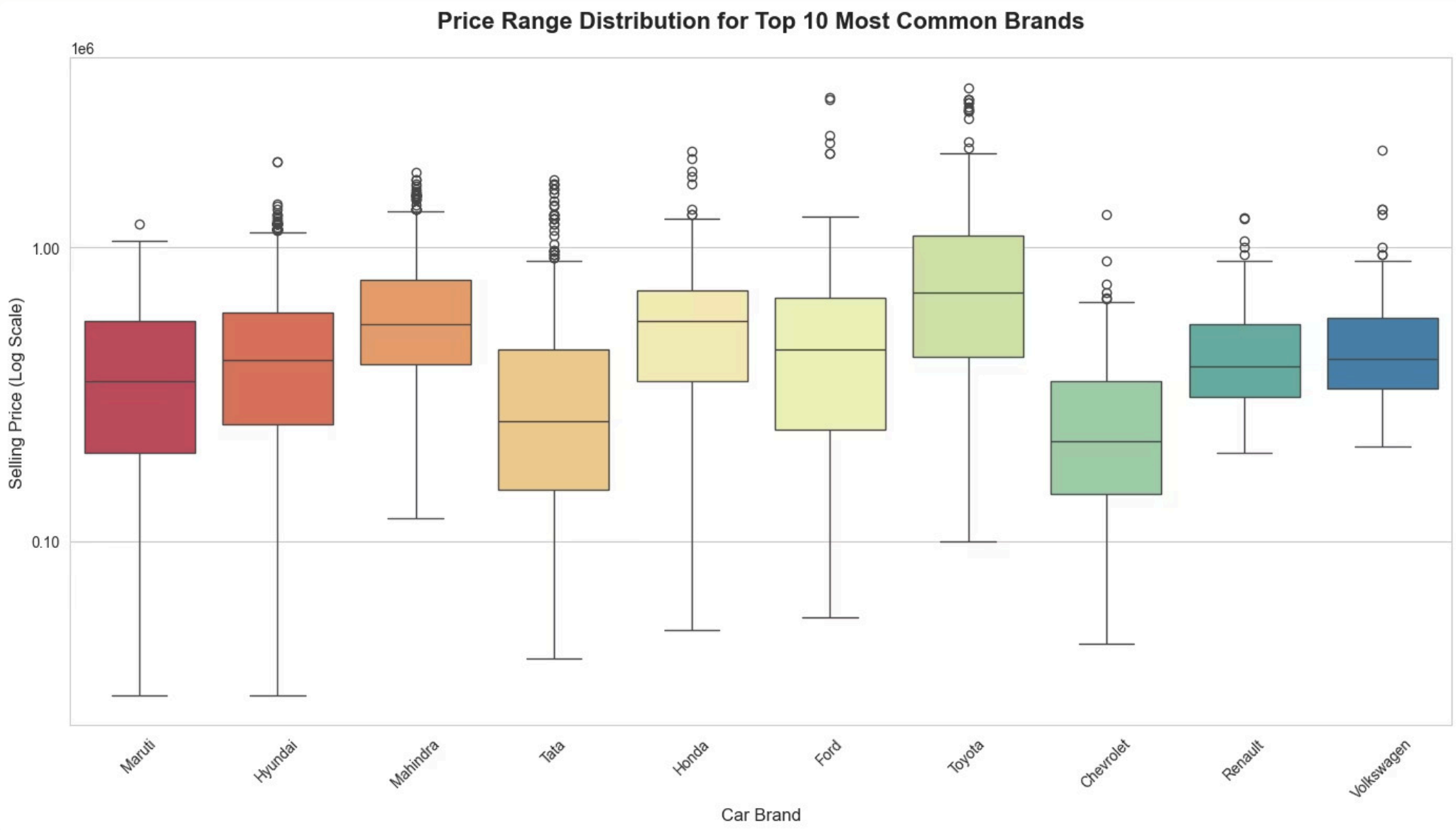


📅 Count of Cars by Year

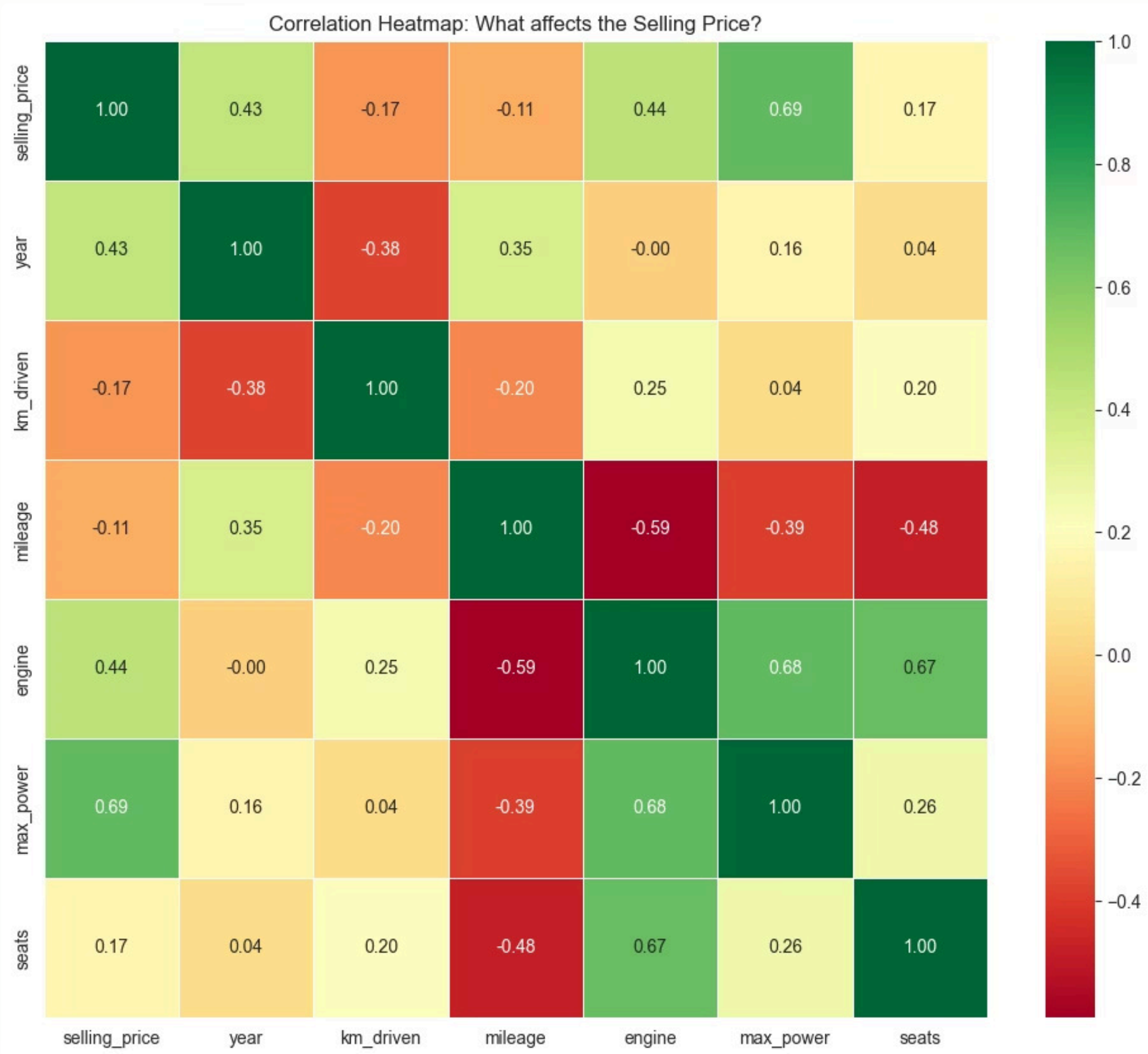
📈 Average Selling Price Trend



🏆 Price Range for Top 10 Brands



🔥 Correlation Heatmap: What affects Price?



⚙️ Feature Engineering

Goal: Optimizing data representation to boost model accuracy and predictive power.

- ⏳ **Smart Feature Creation: "Car Age"**
  - **Action:** Calculated `Current Year - Model Year`.
  - **Value:** Linear models capture **Depreciation** (الاستهلاك) much better through "Age" than raw "Year" timestamps.
- 🔢 **Categorical Encoding Strategies**
  - **Ordinal Encoding:** Used for `owner`. Preserves the logical rank (e.g., First Owner > Second Owner).
  - **One-Hot Encoding (OHE):** Used for `fuel`, `seller_type`, & `transmission`. Prevents the model from assuming a false mathematical order between categories.
- 🎯 **Data Splitting (80/20 Rule)**
  - **Action:** 80% Training (Learning patterns) | 20% Testing (Evaluating performance).
- ⚖️ **Feature Scaling (StandardScaler)**
  - **Action:** Applied **Standardization** to ensure all features (like KM vs. Seats) are on the same scale, preventing large numbers from dominating the model.

MODEL BUILDING & RESULTS

Selecting the Best Predictive Model

We evaluated several regression models to find the most accurate predictor for car selling prices, leveraging GridSearchCV for optimal hyperparameter tuning.

1

Linear Regression

A baseline model to establish initial performance benchmarks.

2

Lasso Regression

Introduces regularization for feature selection and preventing overfitting.

3

Decision Tree Regressor

A non-linear model capable of capturing complex relationships within the data.

4

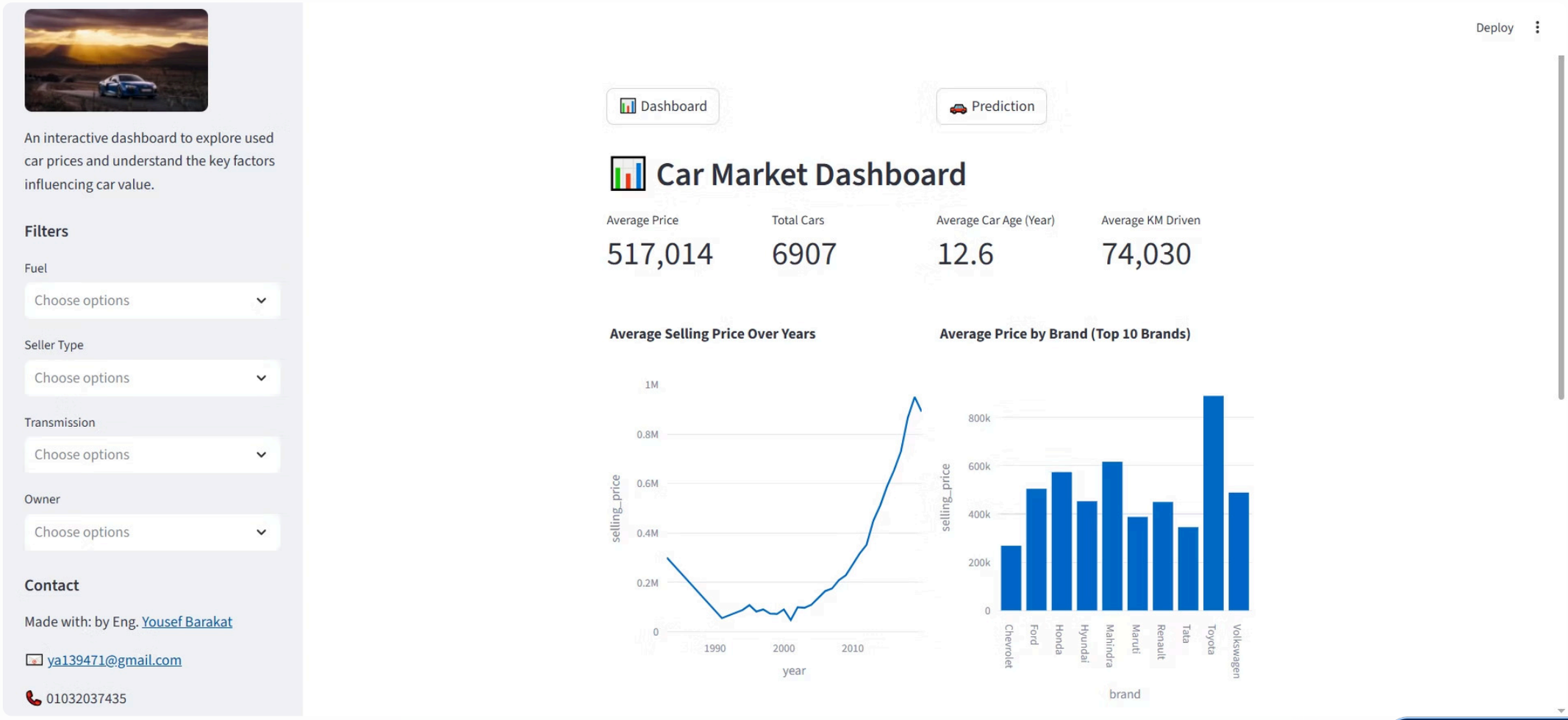
Random Forest Regressor

An ensemble method, often providing high accuracy and robustness. This model achieved the **best performance** with an R^2 score of **~85%**.

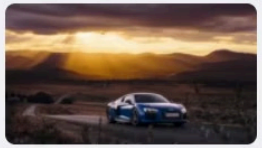
6. Deployment & Live Application 🚀

The final stage was bringing the model to life by building an interactive web application, allowing users to get real-time price estimates.

🖥️ Interactive Web Interface (Streamlit)



Real-Time Prediction Engine



Enter the car specifications to get an estimated market price based on our trained Machine Learning model.

The prediction is powered by an optimized Random Forest model trained on real used car market data.

Model Info

- Model: Random Forest Regressor
- Tuned with GridSearchCV
- Evaluation Metric: R^2 Score
- Preprocessing: StandardScaler + Pipeline

About the Developer

Developed by Eng. Yousef Barakat

ya139471@gmail.com

01032037435

This prediction is for estimation purposes only and may vary from actual market prices.

Deploy

CAR PRICE PREDICTION

Enter the car details to predict its selling price

Model Year

2014

KM Driven

0

0.00

Engine (CC)

500

Max Power (bhp)

0.00

Seats

2

Fuel Type

Petrol

Seller Type

Individual

Transmission

Manual

Owner

First Owner

Predict Price