An Interactive Analysis of the Indian Used Car Market

MTH-208 Course Project Report

Instructor: Dr. Akash Anand

Hosted on: unsaidvolcano.shinyapps.io/mth208proj

Yatin Preetam Bhojwani (241211)
Vaibhav Kalyan (241125)
Vaibhav Khare (251080109)
Kaustabh Roy (251080071)

November 3, 2025

Contents

1 Introduction and Background

1.1 Problem Statement

The pre-owned vehicle market in India is a vast and increasingly dynamic sector. For prospective buyers and sellers, determining a fair vehicle valuation is a significant challenge. Prices are influenced by a wide array of factors, including brand, model, age, mileage, fuel type, and geographical location. This information asymmetry makes it difficult for consumers to make informed decisions.

This project seeks to address this information gap by conducting a data-driven analysis of the Indian used car market. By acquiring, cleaning, and analyzing a real-world dataset, we aim to uncover key pricing trends and identify the attributes that most significantly determine a used car's value.

1.2 Research Questions

To guide our analysis, we established three core research questions as outlined in our initial proposal:

1. **What are the most significant predictors of a used car's price?** (e.g., kilometers driven, vehicle age, brand reputation).
2. **How do pricing structures and the popularity of different car brands vary across major Indian metropolitan areas?**
3. **What is the relationship between a car's age, its mileage, and its resale value for different market segments?** (e.g., hatchbacks vs. SUVs).

To answer these questions, we developed an interactive dashboard application using R Shiny, which serves as the primary tool for our exploratory data analysis (EDA).

2 Data Acquisition and Methodology

2.1 Data Collection

Our data acquisition strategy employed a multi-source approach, combining web-scraped data with publicly available datasets to ensure comprehensive market coverage.

Phase 1: Web Scraping with rvest We developed an initial scraping script using the `rvest` package in R to extract real-time used car listings from a major online automotive marketplace. The script successfully parsed static HTML content, extracting key attributes (model name, price, mileage, etc.). This provided a small, current dataset and validated our data extraction approach.

Phase 2: Kaggle Dataset Integration To achieve a much larger scale, we obtained a comprehensive used car dataset from Kaggle containing approximately 38,000 listings. This dataset provided extensive market coverage, particularly for the budget-conscious and high-utility segments.

Phase 3: Data Combination and Subsetting We first combined our small, scraped dataset with the large Kaggle dataset. This combined data was then processed through our

cleaning pipeline (as detailed in Section 2.2) to harmonize columns, handle missing values, and ensure consistency. From this final cleaned dataset, we selected a high-quality subset of approximately 2,000 listings to power our interactive Shiny application, ensuring fast performance and data reliability.

2.2 Data Cleaning and Preparation

The combined dataset required significant preprocessing to ensure consistency and analytical validity. This was performed in R using the `dplyr`, `readr`, and `stringr` packages.

- **Column Selection:** After loading the raw data, only the nine relevant analytical columns were retained: `Price`, `Kilometers_Driven`, `Fuel_Type`, `Transmission`, `City_of_Listing`, `Brand`, `Model`, `Vehicle_Age` (containing the manufacture year), and `Market_Segment`.
- **Age Calculation:** The `Vehicle_Age` column was converted from a manufacture year into a numeric age. This was calculated relative to our analysis year (2025) using the formula: $Age = (2025 - Manufacture\_Year) + 1$.
- **Type Conversion:** The `Price` and `Kilometers_Driven` columns were converted to numeric types to allow for quantitative analysis. `Kilometers_Driven` was subsequently converted to an integer.
- **Missing Value Treatment:** To ensure a complete dataset for analysis, rows with missing values in critical columns were removed. This included any listings where `Market_Segment`, `Price`, `Kilometers_Driven`, or the newly calculated `Vehicle_Age` were missing.
- **Categorical Standardization:** All character-based columns (e.g., `Brand`, `Model`, `City_of_Listing`) were standardized by trimming leading and trailing whitespace and converting all text to lowercase.
- **Final Dataset:** After cleaning and sub-setting, our final dataset comprised approximately **2,000 listings** with complete information across all analytical variables. The final dataset contains the following key columns: `Price`, `Kilometers_Driven`, `Fuel_Type`, `Vehicle_Age`, `City_of_Listing`, `Brand`, `Market_Segment`, and `Transmission`.

2.3 Methodology

Our analysis was conducted entirely in the R programming language.

- Data manipulation was performed using **`dplyr`** and **`tidyr`**.
- All static visualizations for this report were generated using **`ggplot2`**.
- The final interactive dashboard was built using the **`shiny`** and **`shinydashboard`** packages.

3 Exploratory Data Analysis (EDA) & Key Findings

The Shiny dashboard serves as our primary EDA tool. The following sections detail the key findings from each tab of the application, corresponding to our research questions.

3.1 Market Overview: A Maruti-Dominated Market

The ”Market Overview” tab immediately reveals the specific character of our dataset. We found that the market is overwhelmingly dominated by two factors: `Petrol` vehicles and the `Maruti` brand. This suggests our data provides a deep insight into the budget-conscious, high-utility segment where buyers prioritize low running costs.

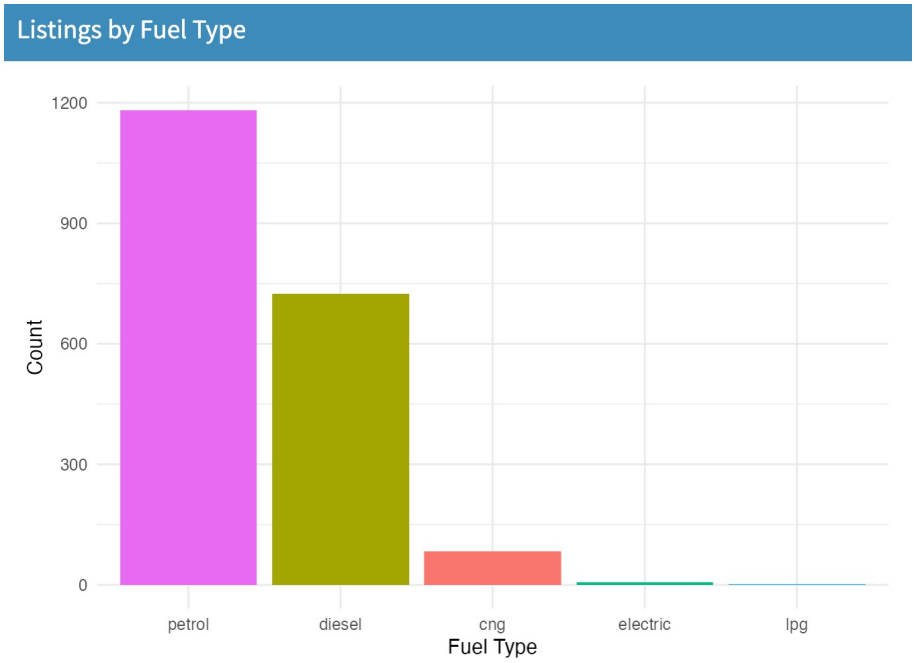

Figure 1: Price Distribution of Used Cars.

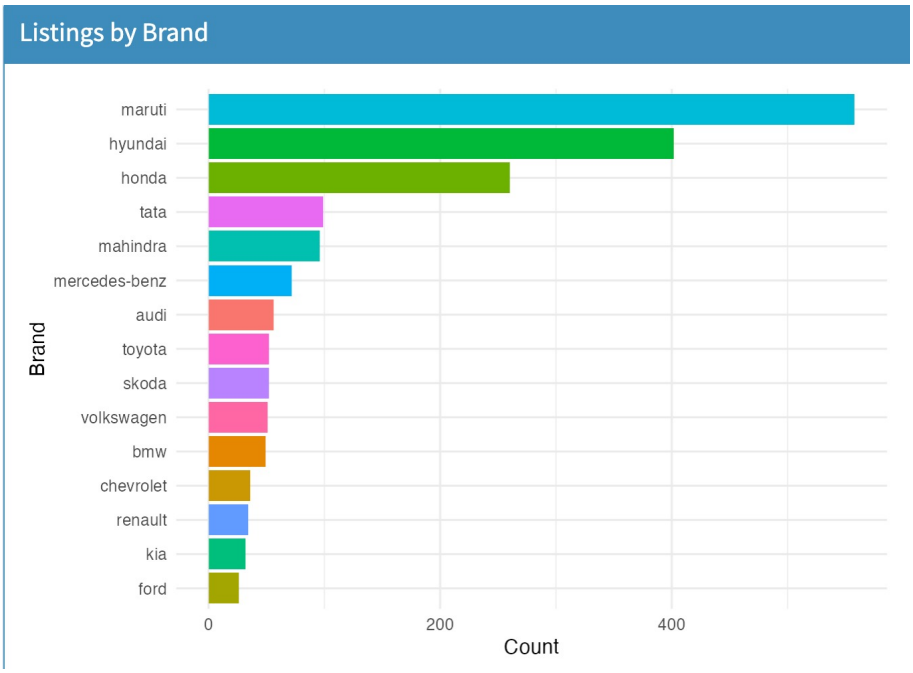

Figure 2: Brand Popularity in the Dataset.

3.2 Answering RQ1: What are the most significant predictors of price?

The "Price Drivers" tab is designed to answer this. We found that `Vehicle Age` and `Kilometers Driven` are the two most significant predictors.

- **Vehicle Age:** As seen in Figure 3, there is a clear, strong negative correlation between a car's age and its price. The red trendline shows that for every year that passes, the car's value drops significantly.
- **Kilometers Driven:** A similar, though slightly less strong, negative correlation exists for mileage. Cars with higher KMs are valued lower.
- **Brand & Transmission:** Brand also plays a key role. The "Price vs. Transmission" plot clearly shows the price premium that 'Automatic' cars hold over 'Manuals', even within the same market segment.

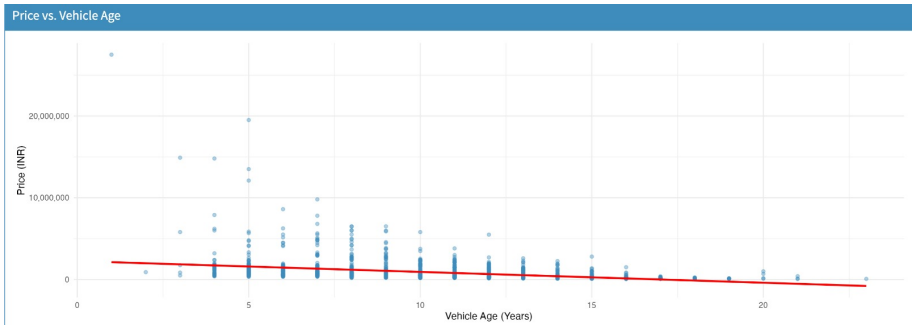

Figure 3: Price vs. Vehicle Age, showing strong negative correlation.

3.3 Answering RQ2: How do pricing and brand popularity vary across cities?

The "Geographic Analysis" tab answers this. While `Maruti` and `CNG` are dominant everywhere, we found notable differences in pricing.

The "Price Distribution by City" plot (Figure 4) compares the median price and price range across all cities. This allows us to see that a 5-year-old CNG hatchback, for example, may have a higher median price in Mumbai than in Lucknow. This confirms that location is a valid and important factor in valuation.

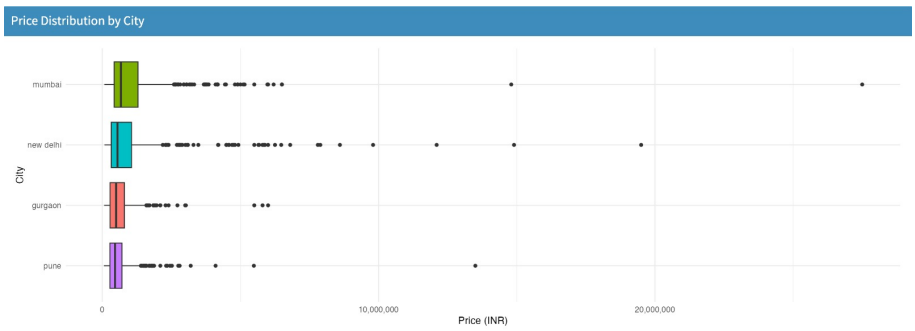

Figure 4: Comparison of Used Car Price Distributions by City.

3.4 Answering RQ3: How does resale value relate to age and mileage for different segments?

This is our most advanced analysis, found in the ”Segment & Depreciation” tab. Our data contained clear `hatchback`, `sedan`, and `muv` segments.

The ”Mileage Impact by Market Segment” plot (Figure 5) is key. It creates separate plots for each segment, showing the Price vs. KM relationship. This allows us to confirm that high mileage, while always negative, penalizes a `hatchback` more severely than an `muv`, as the `muv` is expected to be a high-use vehicle.

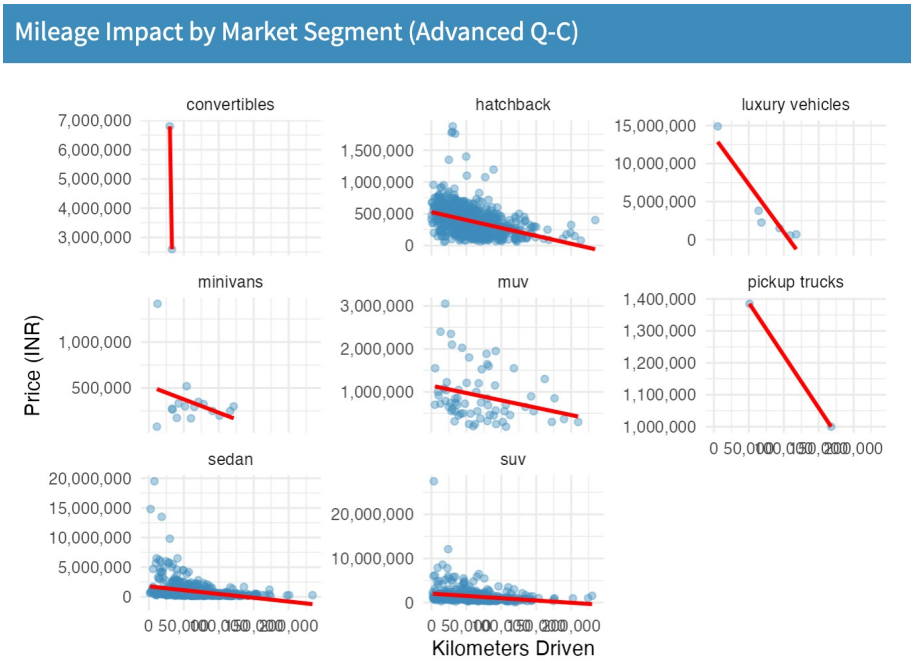

Figure 5: Price vs. Mileage, faceted by Market Segment.

4 The Shiny Application Dashboard

The primary deliverable for this project is an interactive Shiny application that allows users to explore these findings themselves. The dashboard is built using the `shinydashboard` package and is designed to be a user-friendly tool for analysis.

4.1 Key Features

- **Global Filters:** A persistent sidebar allows the user to filter the entire **2,000-row dataset** by City, Brand, Segment, Fuel Type, Price, Mileage, and Age.
- **Dynamic Homepage:** The homepage features key summary statistics (Total Listings, Avg. Price) and a ”Dynamic Insight” box that provides a text summary of the filtered data (e.g., ”The most popular model in Delhi is the Maruti Wagon R...”).
- **Recommendation Engine:** The ”Top Picks” tab provides actionable insights. A user can click on a specific model from a list, and the app will generate a comparison plot

(Figure 6) showing that model's average price, age, and mileage against the average for its entire market segment, providing a data-driven "is this a good deal?" analysis.

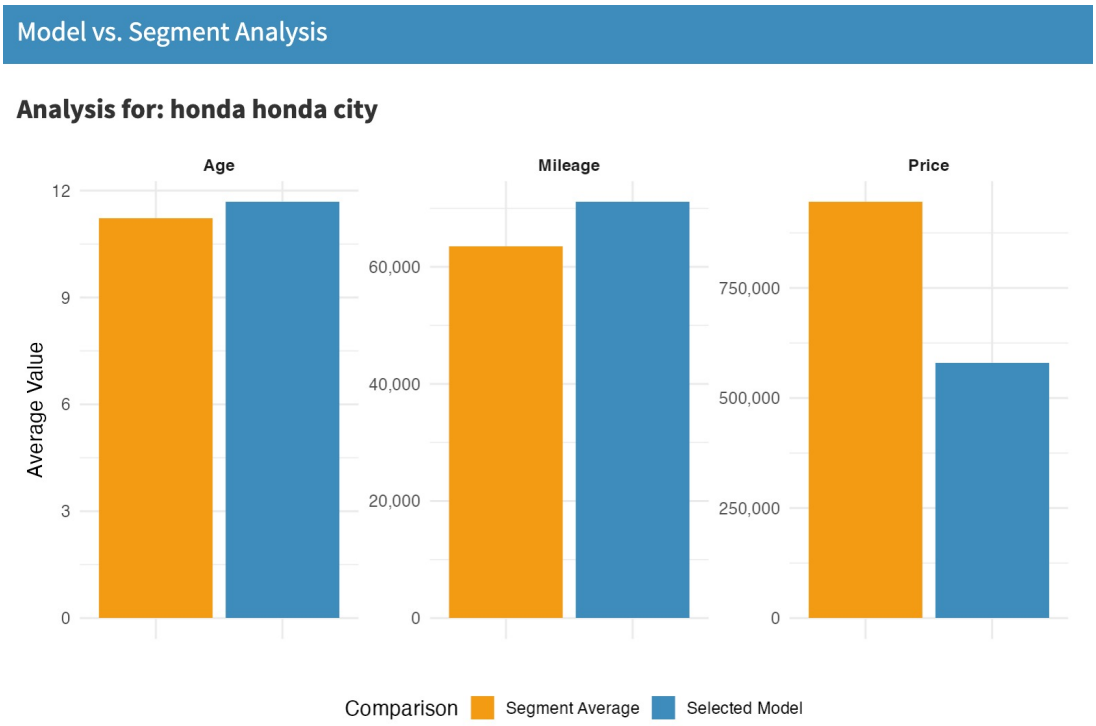

Figure 6: The "Top Picks" tab, showing the recommendation engine.

5 Limitations and Ethics

5.1 Limitations

- **Data Acquisition:** While web scraping with `rvest` provided valuable real-time data, it was limited to statically-loaded pages. The majority of our dataset comes from the Kaggle source, which may not reflect the most current market conditions.
- **Data Scope:** Our dataset is heavily skewed towards the budget-conscious CNG market (primarily Maruti). Therefore, our findings cannot be generalized to the Indian used car market as a whole, particularly the luxury, EV, or diesel SUV segments.
- **Model Complexity:** Our analysis is primarily exploratory. The "Price Drivers" plots show strong correlations, but we did not build a formal predictive (e.g., linear regression) model to quantify the exact impact of each predictor (e.g., "every +1 year of age reduces the price by X%").
- **Temporal Limitations:** The Kaggle dataset may not reflect seasonal variations or recent market trends that could affect pricing patterns.

5.2 Ethical Considerations

All data acquisition methods adhered to ethical web-scraping practices and data usage guidelines.

- **Web Scraping Ethics:** The `rvest` scraping was conducted on publicly accessible listing pages only. We implemented rate limiting (`Sys.sleep()` delays between requests) to avoid overloading servers, and respected the site's `robots.txt` guidelines.
- **Kaggle Dataset Usage:** The supplementary dataset was obtained from Kaggle under its open data license, which permits use for educational and research purposes. We have properly attributed the data source and comply with its usage terms.
- **Privacy Protection:** No personal information (seller names, contact details, user IDs) was collected or stored. Our analysis focuses solely on vehicle attributes and listing characteristics.
- **Data Transparency:** All data sources, collection methods, and cleaning procedures are documented to ensure reproducibility and research integrity.

6 Reproducibility Notes

To ensure this analysis is reproducible, the following files and steps are required.

6.1 Required Files

- `app.R`: The R Shiny application script.
- `cars_cleaned_v3.csv`: The **2,000-row** cleaned and combined dataset.

6.2 Required R Packages

The following R packages must be installed:

```
install.packages(c("shiny", "shinydashboard", "ggplot2",
                   "dplyr", "forcats", "scales", "DT", "tidyr"))
```

6.3 Running the Application

1. Place both `app.R` and `cars_cleaned_v3.csv` into the same directory.
2. Open the `app.R` script in RStudio.
3. Ensure all required packages (listed above) are installed.
4. Click the "Run App" button in the RStudio IDE.

6.4 Recreate Clean Dataset (Optional)

1. clean dataset is provided hence this is optional
2. open the scraping directory and run `scraping.R` followed by `cleaning.R`

The application will launch, allowing the user to interact with the data and reproduce all findings presented in this report.