Web Scraping Tata Car Details from Cars24.com

CONTENTS

1. Project Team E

Team Members

1   Team Lead

   B Deepakjeswi

2   Co-Lead

   Jaimon Jose

3   Team members

   Abinarthana Jayakumar, Digvijay Ingale, Veerta Rajput, Urmi
   Thakkar, Chandan R, Vishnu, Meda Srinivasa Rao

2. Introduction: Navigating the Used Car Market

The Indian used car market is vast, dynamic, and often opaque, making manual inventory and price analysis inefficient and time-consuming. With thousands of listings updated daily, extracting meaningful insights requires automated solutions. This project aims to extract structured, high-quality data on Tata cars from Cars24.com across Mumbai and fallback locations. We focus on key parameters that influence vehicle prices—kilometres driven, fuel type, transmission, and manufacturing year—transforming raw web data into actionable market intelligence.

3. Project Objectives

1

Systematic Data Extraction

· Extract and structure comprehensive Tata car listings data including price, kilometres driven, year of manufacture, fuel type, and transmission type from Cars24.com.

2

Technical Skill Development

· Build team proficiency in advanced techniques for scraping dynamic JavaScript-rendered content, handling pagination, and implementing robust data cleaning workflows.

4. Team Structure & Responsibilities

Core Data Acquisition Team

- Responsible for planning scraping strategy, writing Python scripts, executing data extraction, and performing data cleaning operations.

  - B Deepakjeswi

  - Jaimon Jose

  - Vishnu

  - Urmi

Reporting & Presentation Team

- Translating technical results into formal documentation, creating comprehensive reports, and developing stakeholder presentations.

  - Abinarthana Jayakumar

  - Digvijay Ingale

  - Veerta Rajput

  - Chandan R

  - Meda Srinivasa Rao

5. Technology Stack

Technology	Description
Python	Flexible and powerful programming language providing the foundation for automated web scraping, data processing, and workflow orchestration.
Pandas	Industry-standard library for data cleaning, transformation, and storage in structured DataFrames with powerful analysis capabilities.
Selenium WebDriver	Automates browser interactions and ensures complete loading of dynamic JavaScript-driven content before data extraction.
ChromeDriver Manager	Simplifies driver setup and version management for consistent browser automation across different development environments.
CSV Module	Exports cleaned data to a universally accessible CSV format compatible with Excel, databases, and analytics platforms.

6. Research & Planning Phase

Website Analysis

Research & Planning Phase

1

Pagination Logic

- Developed algorithms for handling multi-page navigation and fallback city scraping to maximize data coverage.

2

Ethical Considerations

- Incorporated controlled delays between requests to respect server capacity and avoid detection mechanisms.

- Used browser Developer Tools to inspect HTML DOM structure and identify CSS selectors for key data fields.

7. Data Extraction Workflow

Selenium automates the browser to navigate Cars24 pages, simulating real user behavior patterns to avoid detection.

Extract five critical attributes per listing: price, kilometres driven, manufacturing year, fuel type, and transmission type.

Robust exception management gracefully handles incomplete or missing data entries without disrupting the scraping process.

A 5-second delay per page ensures stable content loading and demonstrates respect for server resources.
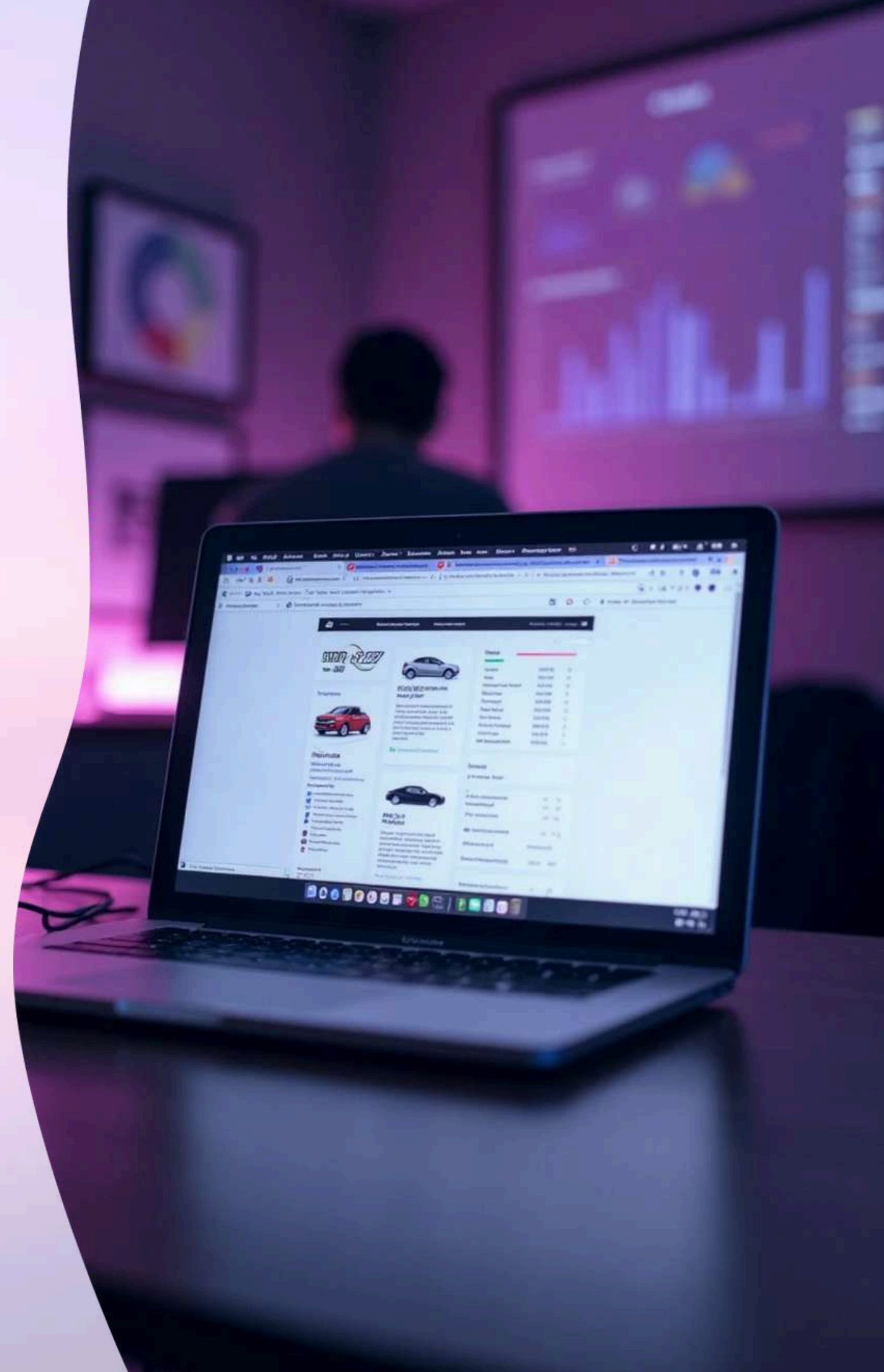
8. Data Cleaning & Feature Engineering

1 Data Sanitisation

• Removed non-numeric artifacts from price and kilometres fields, including currency symbols, commas, and unit labels.

2 Type Conversion

• Converted textual data to appropriate numeric types enabling quantitative analysis and statistical operations.

3 Feature Creation

• Engineered a vehicle age feature by calculating years elapsed from the manufacturing year to the current year.

4 Data Organisation

• Sorted and loaded data into Pandas DataFrame with consistent formatting and validated data types.

9. Data Analysis

Sample Dataset: Clean & Structured Output

The following sample demonstrates the clean, analysis-ready structure of our scraped data, ready for business intelligence and pricing analysis:

Car Name	Year	Kilometres	Fuel Type	Transmission	Price (₹)	Location
Tata Nexon XM	2020	38,000	Petrol	Manual	8,25,000	Mumbai
Tata Tiago XT	2019	45,500	Petrol	Manual	5,20,000	Mumbai
Tata Harrier XZ	2021	29,000	Diesel	Manual	15,90,000	Mumbai
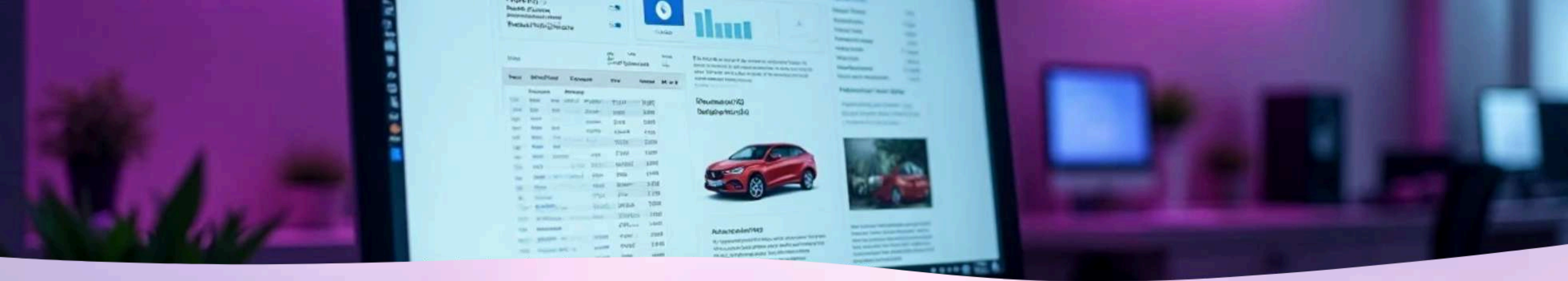
10. Key Takeaways & Impact

Automation Success

Successfully automated extraction of structured Tata car data from dynamic JavaScript websites, eliminating manual data collection.

Automated workflows included effective error handling that minimized the risk of failed data retrieval, ensuring continuous and dependable operations

Data Quality

Produced clean, analysis-ready datasets with consistent formatting enabling immediate business intelligence applications.

Maintaining standardized data structures across extraction cycles ensures that reporting and analytics processes remain accurate and comparably reliable over time

High data quality builds stakeholder confidence, empowering leaders to make informed decisions based on trustworthy analytics

Scalable Framework

Created a reusable scraping framework adaptable to other car brands, websites, and locations for future projects.

Integration with APIs and third-party tools allows for easy upgrades and connections to analytics platforms, supporting long-term business intelligence strategies

The framework's modular architecture supports seamless expansion, allowing it to accommodate higher data volumes and new data sources without performance bottlenecks

Thank You