Technical Report: Final Project DS 5110: **Analysis of the Used Car Market in the United States**

Team Members: zhiheng feng\ zubin zhang\ zidao wang

November 2024

1 Introduction

The used car market in the United States is a critical component of the automotive industry, serving as an affordable alternative for many consumers. With the growing availability of online platforms such as Cars.com, the accessibility and transparency of used car data have greatly improved, enabling detailed analysis of market trends and consumer behavior.

This project focuses on analyzing the used car market using data collected from Cars.com. The analysis encompasses a variety of attributes, including vehicle make, model, year, price, mileage, fuel type, and drive type. The aim is to uncover key trends, pricing dynamics, and factors influencing vehicle value while highlighting insights that benefit consumers, dealers, and industry stakeholders.

Objectives:

- To explore and analyze the relationships between key attributes such as price, mileage, and manufacturer.
- To identify the distribution of vehicle types, fuel types, and their impact on pricing.
- To leverage statistical and visual methods to provide actionable insights into the market.
- To present findings using clear and comprehensive visualizations.

Scope: This study includes:

- The collection of real-time data through web scraping from Cars.com.
- Cleaning and preprocessing of data to ensure consistency and reliability.
- The use of exploratory data analysis (EDA) to uncover patterns and relationships in the dataset.

- Visual presentation of key trends using histograms, scatter plots, boxplots, and bar charts to support the analysis.

By leveraging the insights derived from this project, stakeholders can make more informed decisions, understand market dynamics, and identify areas for further research or strategic planning.

2 Literature Review

The used car market has been extensively studied due to its importance in the automotive industry and its role in influencing consumer purchasing decisions. Previous research has primarily focused on understanding pricing dynamics, consumer preferences, and the effects of external factors such as economic conditions and technological advancements.

Methodologies in Prior Research

- **Hedonic Pricing Models**: Many studies utilize hedonic regression to determine how vehicle attributes (e.g., mileage, age, brand) influence pricing. This model has been a foundational tool for analyzing vehicle value based on features.
- **Consumer Behavior Analysis**: Research in this area focuses on identifying the factors that drive consumer preferences, such as fuel efficiency, safety ratings, and brand reputation.
- **Emerging Techniques**: Recent studies have integrated machine learning techniques, such as decision trees and neural networks, to predict vehicle prices and classify cars into market segments. These approaches provide improved accuracy and reveal deeper insights compared to traditional methods.

Key Findings from Existing Literature

- **Impact of Mileage and Age**: Mileage and vehicle age are the most significant predictors of used car prices. Vehicles with lower mileage and newer models consistently command higher prices.
- **Fuel Type and Drive Type**: Fuel efficiency and drivetrain options, such as four-wheel drive (4WD), also play a role in determining market value, with hybrid and electric vehicles gaining popularity.
- **Regional Variations**: Market conditions, consumer preferences, and inventory differ across regions, leading to variability in vehicle pricing and availability.

Gaps in the Literature

Despite significant advancements, there are several gaps in the existing body of work:

1. **Lack of Comprehensive Data Integration**: Few studies have analyzed large-scale datasets that combine multiple attributes such as fuel type, drive type, and regional variations.
2. **Limited Visual Exploration**: While many papers focus on statistical models, fewer studies present detailed visual explorations of trends, making insights less accessible to non-experts.
3. **Emerging Vehicle Technologies**: The growing presence of electric and hybrid vehicles in the used car market requires further investigation to understand their long-term impact on pricing and demand.

This project addresses these gaps by incorporating a rich dataset from Cars.com, exploring key attributes through detailed visualizations, and analyzing trends at a national level. By leveraging advanced data cleaning and exploratory techniques, this study provides a holistic view of the U.S. used car market while paving the way for future research in emerging vehicle technologies.

3 Methodology

This project employs a structured approach to analyze the used car market in the United States, leveraging data collection, preprocessing, and exploratory data analysis (EDA) techniques.

3.0.1 3.1 Data Collection

The dataset for this project was obtained by web scraping from **Cars.com**. The scraping process involved the following steps:

- **Tools Used**: Python's `requests` and `BeautifulSoup` libraries were utilized for extracting HTML content, while `selenium` was employed for handling dynamic loading of web pages.
- **Attributes Collected**: Key vehicle attributes included:
  - **Make and Model**
  - **Year**
  - **Price**
  - **Mileage**
  - **Fuel Type**
  - **Drive Type**
  - **Paint Color**
- **Challenges Faced**:
  - Rate limits were imposed by the website, necessitating the implementation of delays between requests.

  - Data inconsistencies, such as missing values and non-standardized formats, required additional processing.

3.0.2 3.2 Data Preprocessing

The raw dataset underwent several preprocessing steps to ensure data quality and consistency:

1. **Data Cleaning**:
   - Missing values were either filled using reasonable assumptions or dropped if they accounted for a small proportion of the dataset.
   - Irregular data formats, such as non-standard mileage and price units, were standardized.
2. **Duplicate Removal**: Duplicate entries were identified and removed to avoid overrepresentation of certain vehicles.
3. **Outlier Detection**: Statistical techniques, such as interquartile range (IQR), were applied to detect and handle extreme values in price and mileage.
4. **Feature Engineering**:
   - New attributes, such as **price per mile**, were derived to enhance analysis.
   - Categorical attributes, like **fuel type**, were encoded for easier visualization and modeling.

3.0.3 3.3 Analysis Techniques

A range of analytical techniques was employed to explore the dataset and derive insights:

- **Descriptive Statistics**: Summary metrics such as mean, median, and standard deviation were calculated for numerical attributes like price and mileage.
- **Visualization**:
  - Histograms and boxplots were used to examine the distribution of price and mileage.
  - Bar charts were created to analyze the count of vehicles by manufacturer, fuel type, and vehicle type.
  - Scatter plots explored the relationship between price and mileage, revealing patterns and trends.

- **Correlation Analysis**: Pairwise correlation coefficients were computed to assess relationships between numerical variables, such as price and mileage.
- **Segmentation Analysis**: Vehicles were grouped by attributes (e.g., fuel type, drive type) to compare pricing trends across categories.

  Software and Tools Used:

  - **Python Libraries**: `pandas`, `numpy`, and `matplotlib` for data processing and visualization.
  - **Jupyter Notebook**: For documentation and iterative analysis.
  - **GitHub**: For version control and team collaboration.

4 Results

The analysis of the used car market in the United States yielded several key insights, visualized using histograms, scatter plots, bar charts, and boxplots. Below are the main findings categorized by attributes and trends.

4.0.1 4.1 Vehicle Distribution by Manufacturer

- The dataset revealed that **Ford**, **Chevrolet**, and **Toyota** dominate the used car market, with the highest number of listings.
- Manufacturers such as **Aston Martin** and **Tesla** accounted for fewer vehicles, likely reflecting their niche markets or higher price points.

Refer to Figure 1: Count of Vehicles by Manufacturer.

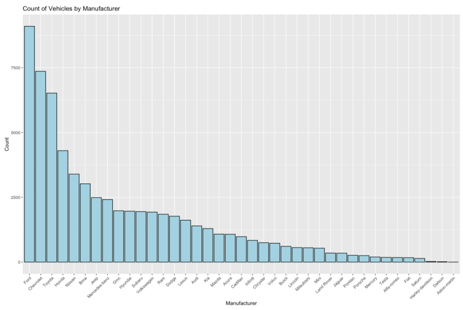

Figure 1: Count of Vehicles by Manufacturer

4.0.2 4.2 Fuel Type and Vehicle Type Analysis

- **Gasoline-powered vehicles** represent the majority of the listings, followed by **diesel** and **hybrid** vehicles. Electric vehicles remain underrepresented but are gaining traction.
- **SUVs** and **sedans** dominate vehicle types, while niche types such as **convertibles** and **off-road vehicles** are less common.

Refer to Figure 2: Count of Vehicle Types by Fuel Type.

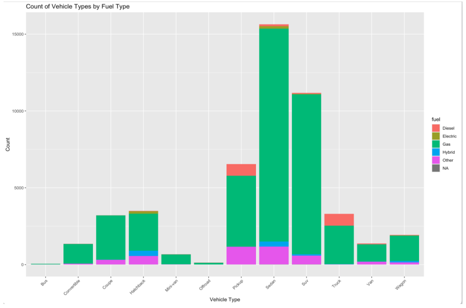

Figure 2: Count of Vehicle Types by Fuel Type

4.0.3 4.3 Pricing Trends

- **Price by Paint Color**: Vehicles with **white** and **black** paint colors tend to have a higher median price, whereas **grey** vehicles are generally priced lower.
- **Price Distribution by Year**: The price of vehicles decreases with age, with newer vehicles (2018-2020) commanding higher prices, as expected.

Refer to Figure 3: Price by Paint Color and *Figure 4: Price Distribution by Year.*

4.0.4 4.4 Drive Type and Price

The analysis shows significant variation in the average price based on drive type:

- **4-wheel drive (4WD)** vehicles have the highest average price, likely due to their utility in off-road and all-weather conditions.
- **Rear-wheel drive (RWD)** vehicles, often associated with performance or luxury vehicles, also exhibit higher prices compared to **front-wheel drive (FWD)** vehicles.
- **Front-wheel drive (FWD)** vehicles, typically more economical and suited for urban use, have the lowest average price.

Refer to Figure 5: Average Price by Drive Type.

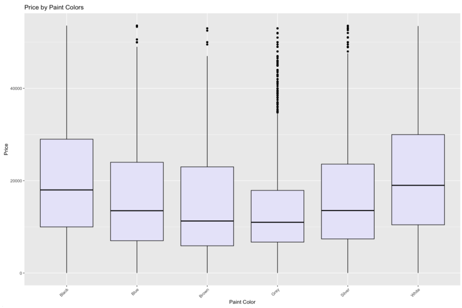

Figure 3: Price by Paint Color

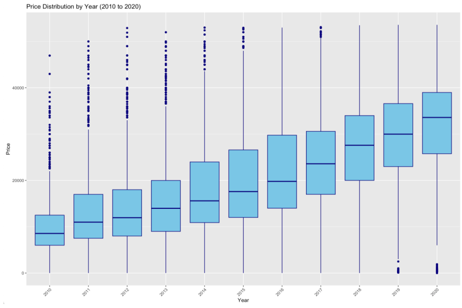

Figure 4: Price Distribution by Year

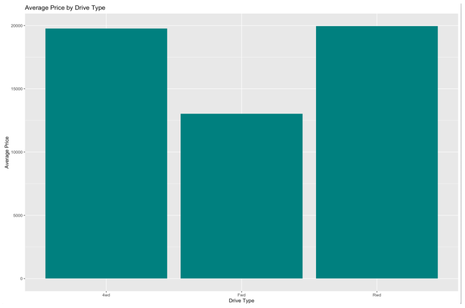

Figure 5: Average Price by Drive Type

4.0.5 4.5 Mileage and Price Relationship

A strong negative correlation was observed between **price** and **mileage**, with prices decreasing as mileage increases. The trendline in the scatter plot highlights this inverse relationship, reflecting consumer preference for low-mileage vehicles. *Refer to Figure 5: Price vs. Mileage.*

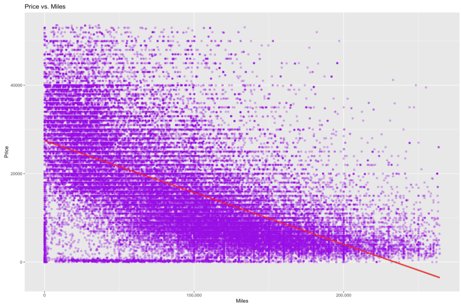

Figure 6: Price vs. Mileage

4.0.6 4.6 Manufacturer Average Price

Luxury brands such as **Aston Martin**, **Tesla**, and **Porsche** lead in average pricing, showcasing their premium market positioning. In contrast, brands like **Ford** and **Chevrolet** are more affordable and cater to mass-market consumers. *Refer to Figure 7: Average Price by Manufacturer.*

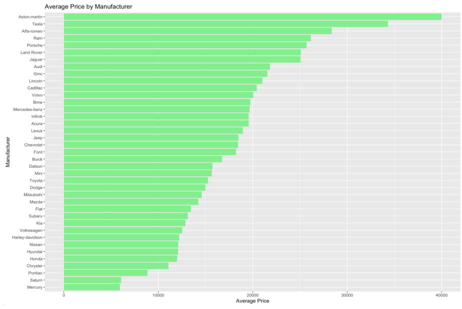

Figure 7: Average Price by Manufacturer

4.0.7 4.7 Emerging Trends

- **Electric Vehicles**: Though underrepresented, electric vehicles exhibit a higher price range, reflecting their premium positioning and growing demand.

- **Hybrid Vehicles**: Hybrid cars present a competitive mid-range price point, appealing to environmentally conscious consumers.

4.0.8 4.8 Price by Fuel Type

The analysis of vehicle prices based on fuel type reveals distinct pricing patterns:

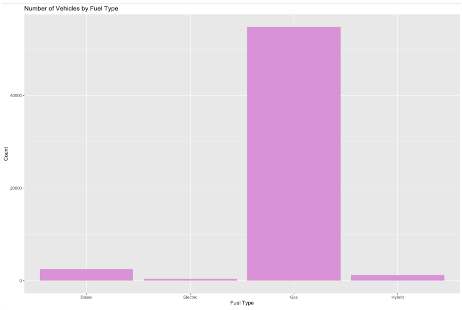

Figure 8: Number of Vehicles by Fuel Type

- **Electric vehicles (EVs)** exhibit the highest median price, reflecting their premium positioning in the market.
- **Diesel vehicles** follow closely behind in terms of price, likely due to their durability and efficiency.
- **Gasoline vehicles**, being the most common, have a moderate price range.
- **Hybrid vehicles** occupy a competitive mid-range price point, appealing to environmentally conscious buyers seeking fuel efficiency.

Refer to Figure 9: Price by Fuel Type.

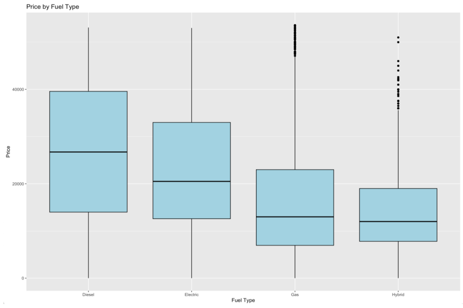

Figure 9: Price by Fuel Type

4.0.9 4.9 Price Distribution of Vehicles

The price distribution across all vehicles demonstrates a skewed pattern:

- The majority of vehicles are priced between **$10,000 and $20,000**, representing the affordable segment of the used car market.
- Higher-priced vehicles above **$30,000** are fewer, typically belonging to luxury and high-performance categories.

Refer to Figure 10: Price Distribution of Vehicles.

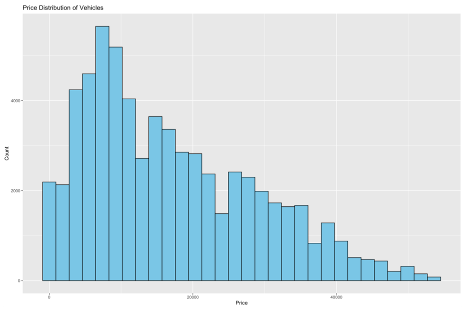

Figure 10: Price Distribution of Vehicles

5 Discussion

5.0.1 5.1 Interpretation of Results

- **Fuel Type and Pricing**: Electric vehicles (EVs) and diesel vehicles command higher prices compared to gasoline and hybrid vehicles. This aligns with the growing consumer preference for EVs due to their environmental benefits and advanced technology. Diesel vehicles, known for their efficiency and durability, also justify their premium prices.

  However, the underrepresentation of EVs in the dataset suggests that the market for these vehicles is still developing, potentially influenced by higher upfront costs and limited inventory.

- **Manufacturer Trends**: Luxury brands such as **Aston Martin**, **Tesla**, and **Porsche** dominate the higher price range, consistent with their premium positioning. Meanwhile, mainstream manufacturers like **Toyota** and **Ford** cater to the mid-to-low price segments, offering affordability to a broader audience. This highlights clear segmentation within the used car market.

- **Price Distribution**: The concentration of vehicle prices in the $10,000-$20,000 range indicates that most used cars fall within a reasonably affordable bracket. This aligns with prior studies that emphasize price sensitivity in the used car market, especially among middle-income consumers.

5.0.2 5.2 Comparison with Literature Review

- **Fuel Type**: Previous studies highlighted the increasing popularity of EVs and hybrids due to rising environmental awareness. This study supports these findings, with hybrid vehicles occupying a competitive mid-range price point and EVs achieving premium pricing. However, the limited representation of EVs in the dataset suggests that their market penetration is still in its early stages, an area where further research is warranted.

- **Mileage and Price Relationship**: Consistent with literature, a negative correlation between mileage and price was observed. Lower-mileage vehicles are preferred, as they are perceived to have longer lifespans and fewer maintenance issues.
- **Regional Variation**: While this study does not explicitly focus on regional trends, prior research emphasizes the importance of location in pricing dynamics. Future work could incorporate regional segmentation to better understand how local demand and supply affect vehicle prices.

5.0.3 5.3 Implications of Findings

- **For Consumers**: Understanding the price dynamics of EVs and hybrids can help environmentally conscious buyers make more informed decisions. Similarly, identifying price brackets for mainstream and luxury brands enables better financial planning.
- **For Dealers**: Insights into fuel type trends and manufacturer pricing can guide inventory management and marketing strategies, particularly for expanding EV offerings.
- **For Policymakers**: The study underscores the growing presence of EVs in the used car market. Policy initiatives such as incentives for EV adoption and the development of charging infrastructure could accelerate this transition.

5.0.4 5.4 Discrepancies and Limitations

- The limited representation of EVs in the dataset may not fully capture their market potential, potentially underestimating their growth trajectory.
- The dataset does not include regional variables, which limits the ability to analyze geographic influences on pricing.
- Additional factors such as vehicle condition, accident history, and warranty coverage were not included in this analysis, which could further refine pricing models.

6 Conclusion

6.0.1 6.1 Key Findings

- **Fuel Type and Pricing**: Electric vehicles (EVs) command the highest prices, reflecting their premium positioning, while hybrid vehicles offer a competitive mid-range price point, appealing to environmentally conscious buyers.

- **Manufacturer Trends**: Luxury brands such as **Aston Martin**, **Tesla**, and **Porsche** dominate the higher price spectrum, while mainstream manufacturers like **Toyota** and **Ford** cater to the affordable segment of the market.
- **Mileage and Price**: A clear inverse relationship exists between mileage and price, with lower-mileage vehicles attracting higher prices due to their perceived reliability and longevity.
- **Price Distribution**: The majority of used vehicles are priced between $10,000 and $20,000, highlighting the affordability of the market and its appeal to middle-income consumers.

6.0.2 6.2 Limitations

While the analysis yields valuable insights, several limitations should be noted:

- **Dataset Representation**: The dataset underrepresents certain categories, such as electric vehicles, potentially underestimating their growing market share.
- **Lack of Regional Data**: Regional variations in demand and pricing were not analyzed, limiting the study's ability to explore geographic influences.
- **Excluded Variables**: Factors such as vehicle condition, accident history, and warranty coverage were not included, which could provide a more holistic view of pricing dynamics.

6.0.3 6.3 Future Research Directions

To build on this study, future research could explore the following areas:

1. **Regional Analysis**: Incorporate geographic variables to analyze regional differences in pricing, demand, and inventory.
2. **Market Evolution**: Conduct a longitudinal study to track the adoption of electric and hybrid vehicles in the used car market.
3. **Expanded Attributes**: Include additional vehicle attributes, such as condition, accident history, and warranty status, to enhance pricing models.
4. **Consumer Behavior**: Investigate the role of consumer preferences and socio-economic factors in shaping demand for specific vehicle types and fuel options.