

CDS6214

Data Science Fundamentals

Project (40%)

Group Number: G10 I

Title: Transportation

|  |  |  |
| --- | --- | --- |
| **ID** | **Name** | **Contribution (in %)** |
| 242UC24511 | Chan Jia Jun | 33 |
| 242UC24513 | Chen Yong Qi | 30 |
| 242UC24515 | Tong Jian Hao | 37 |
|  | **Total** | **100** |

Note: Please ensure all group members agree to the submitted contribution (in %).

|  |  |
| --- | --- |
| **YouTube Link** |  |
| **Dataset Link** |  |

**Table of Contents**

**Introduction and Objectives 2**

**Dataset Description 3**

**Exploratory Questions and Answers 4**

**Modeling and Evaluation 10**

**Discussion and Insights 10**

**Conclusion 11**

**References 12**

# Introduction and Objectives

Problem Statement  
In the dynamic transportation industry, used car buyers and sellers are often faced with uncertainty around vehicle value and risk factors. Pricing disparities and lack of transparency in history (service, accident, ownership) make informed decisions difficult.  
  
Objectives and Scope  
This project aims to:  
- Identify patterns in pricing related to service history.  
- Analyze trends in engine size over time to predict buyer preferences.  
- Determine if certain colors correlate with accident reports.  
- Examine ownership patterns among economy vs premium cars.  
  
Scope: Limited to structured analysis of a used car dataset with over 1,000 entries and 10+ attributes. Analysis is exploratory and interpretive, not predictive.  
  
Background  
Used cars occupy a significant position in the automotive market in developing countries such as Malaysia, the United States, and China. Factors such as accident history, engine displacement, and number of ownerships directly affect the perceived vehicle reliability and price. As data-driven decision making becomes crucial in this field, actionable insights are crucial for dealers, insurers, and buyers.

# Dataset Description

The dataset contains 10,000 entries of used cars with 13 features including make year, mileage, engine capacity, fuel type, ownership history, color, transmission, accidents reported, insurance validity, service history, and brand. All data entries were pre-cleaned to remove duplicates and missing values. Categorical variables such as fuel type and brand were label-encoded for modeling.

Target variable: price\_usd (continuous).

# Exploratory Questions and Answers

**Q1: Does higher car price correlate with a full service history?**

A blue rectangular bars with white text

AI-generated content may be incorrect.

Figure 1: Avg Price by Service History

Chart Observation

Figure 1 compares the average price (USD) of used cars based on their service history: Partial vs Full.

* Vehicles with a full service history show a slightly higher average price than those with only partial service history.
* Although the price gap is not large, the upward trend is noticeable.

Information Understanding

* Trust Factor: A thorough service history indicates appropriate upkeep, which can boost buyer confidence and lower perceived risk.
* Price Premium: Customers may be prepared to pay more for verified maintenance if the price is even slightly raised.
* Market Strategy: In order to defend higher prices, auto dealers may highlight complete service records as a selling point.

Conclusion

Figure 1 shows that service history has a value-added role in the used car market, and cars with full service records usually sell for a slightly higher price. Even if the difference is small, this can be a selling point for the car.

**Q2: How can manufacturers or resellers adapt to the trend of engine sizes in recent years to better match consumer demand?**

A graph with blue lines and numbers

AI-generated content may be incorrect.

Figure 2: Avg Engine Capacity by Year

Chart observations

* Figure 2 shows the average engine capacity (in CC) of cars by year of manufacture from 1995 to 2023.
* Between 1995 and 2015, engine displacement varied widely, generally remaining above 2300 CC.
* However, starting around 2016, the situation changed, and engine displacement began to decline, and is now mostly around 2200 CC or even lower.

Information Understanding

* Drivers are more focused on fuel efficiency and reducing running costs. Large engines tend to consume more fuel, so smaller engines are more attractive (especially against the backdrop of rising oil prices).
* Environmental concerns and government emission regulations have also prompted manufacturers to turn to cleaner, smaller engines.

Conclusion

The continued reduction in engine capacity reflects a general pursuit of efficiency and sustainability. Whether making or selling cars, adapting to this trend is key to staying competitive and meeting the real needs of today's buyers.

**Q3: Do certain car colors correlate with higher accident reports?**

A bar graph with blue and white stripes

AI-generated content may be incorrect.

Figure 3: Avg Accidents by Colour

Chart observations

* Figure 3 shows the average number of accidents reported for vehicles of different colors. The graph compares common colors such as blue, gray, black, white, red, and silver.
* The differences are small, with average accident rates close to 0.5 for most colors.
* However, blue, gray, and black vehicles have slightly higher average accident rates, while red and silver vehicles have the lowest accident rates by a small margin.

Information Understanding

* Visibility may be a factor, as vehicles with brighter colors (e.g., red) are easier to see in low light and bad weather, while darker colors (e.g., black) are harder to see.
* Another factor is driver behavior. Certain colors may appeal to different types of drivers or be more common in specific driving environments (e.g., rural vs. urban).

Conclusion

While there is no clear correlation between car color and accident rate, this analysis reveals a subtle trend: cars with darker colors (e.g., blue, gray, and black) tend to have slightly higher accident rates. It’s a small but intriguing detail that provides some information.

**Q4: What are the ownership trends among economy vs high-end car brands?**

A graph of a number of people

AI-generated content may be incorrect.

Figure 4: Avg Owner Count by year & Category

Chart observations

* Figure 4 shows the average number of previous owners for economy and premium cars from 1995 to 2023.
* Economy cars (e.g. Toyota, Honda) generally have slightly more average owners than premium brands (e.g. BMW, Mercedes-Benz).
* This difference has remained consistent over time, with the average number of owners for economy cars being around 3, while the average number of owners for premium cars has mostly remained around 2.8.

Information Understanding

* Economy cars are resold more frequently, due to their affordability, reliability, and continued demand in the used market.
* Premium cars are more expensive to maintain, so they may stay in the hands of the first or second owner longer.
* Newer models in both categories have slightly fewer owners, suggesting that they have higher retention rates or shorter resale cycles.

Conclusion

The data confirms the hypothesis that older economy cars generally have more previous owners than premium cars. This reflects their role as reliable, affordable options that have been through more hands, which is an important consideration for buyers and sellers in the used car market.

# Modeling and Evaluation

A Random Forest Regressor was trained using features such as brand, mileage, engine size, owner count, fuel type, and color. The dataset was split into 80% training and 20% testing. Categorical features were label-encoded. Model performance was evaluated using two metrics:

* Mean Squared Error (MSE): 1,486,082.94
* R² Score: 0.8132

The R² score of 0.8132 indicates strong explanatory power, meaning over 81% of the price variance can be explained by the input features.

# Discussion and Insights

**Finding Summary**

* Maintenance History: Verified vehicles are more expensive, an indicator of value for buyers.
* Engine Trends: Manufacturers may need to prioritize compact, eco-friendly engines.
* Color Bias: Subtle correlations suggest that insurers may be basing risk analysis on color.
* Ownership: Economy models have higher ownership churn, a consideration for resale value.

**Limitations**

* The dataset may be biased toward certain brands or regions.
* The color-accident correlation may involve confounding factors (e.g., usage pattern, city driving).

**Practical Recommendations**

* Dealers should highlight verified maintenance history in product listings.
* Manufacturers should evaluate the trend toward smaller engines after 2015.
* Buyers may consider accident-prone colors when assessing risk.
* Insurers can evaluate long-term ownership trends to adjust policies.

# Conclusion

This project has highlighted the practical value of data science in finding meaningful insights within the used car market. By concentrating on four clear questions, we explored connections between vehicle price, service history, engine size trends, accident records, and ownership patterns. Each analysis uncovered subtle yet important trends that reflect consumer preferences, market behaviors, and opportunities for resellers and manufacturers.

For example, we discovered that cars with a full service history generally sell for slightly higher prices, showing that verified maintenance can build buyer trust and support higher value. The analysis of engine capacity over time showed a growing shift toward smaller, more efficient engines, indicating that both manufacturers and resellers need to respond to changing environmental and economic pressures. Similarly, while exploring accident reports by color was not conclusive, it suggested that visibility and driving patterns might affect risk profiles. Lastly, comparing ownership trends between economy and high-end cars reinforced the idea that affordability drives quicker turnover, which is an important insight for vehicle lifecycle planning and pricing strategies.

Overall, these findings emphasize how structured data, when examined systematically, can aid in making better decisions across the automotive resale landscape. From selecting inventory and pricing to marketing strategies and assessing consumer risk, data-driven insights lead to greater profitability, efficiency, and customer satisfaction. This project ultimately shows the value of using data analytics not just to understand the market but also to influence it.

# References

American Psychological Association. (2020). \*Publication manual of the American Psychological Association\* (7th ed.).

Scikit-learn Developers. (2024). \*Scikit-learn documentation\*. https://scikit-learn.org/

XGBoost Contributors. (2024). \*XGBoost documentation\*. https://xgboost.readthedocs.io/

Pandas Documentation. (2024). \*Pandas User Guide\*. https://pandas.pydata.org/docs/

Seaborn Documentation. (2024). \*Seaborn visualization library\*. https://seaborn.pydata.org/