CIND820: Big Data Analytics Project

Literature Review, Data Description and Approach

Dealership Vehicle Dataset

Name: Aaron Hosein

Student Number: 500437092

Supervisor: Ceni Babaoglu, PhD

Date: June 23, 2025

GitHub Repository

https://github.com/Logic9397/Vehicle_Dataset_for_CIND820/tree/main

Introduction:

The automotive industry is continuing to evolve with a growing emphasis on consumer purchasing, pitting conventional gasoline vehicles with newer electrical counterparts. As vehicle options diversify, so does the complexity of determining pricing—particularly for new vehicles. This project adopts a machine learning approach to analyze, classify, and model vehicle price behavior, contributing to the theme of predictive analytics and pattern recognition in the automotive market.

Problem Statement and Research Objectives

The primary objective of this study is to investigate the relationship between vehicle features and their market value, using a dataset of new vehicle listings. The project aims to answer sample questions like the below listed:

Classification / Clustering: Can vehicles be automatically categorized into groups such as budget, mid-range, and luxury based on pricing and specifications?

Feature Contribution: What is the correlation between vehicle MSRP (Manufacturer's Suggested Retail Price) and other features such as engine type, make, drivetrain, fuel type, and trim? Which attributes most significantly influence price?

Distribution Analysis: What does the price distribution of new vehicles look like when grouped into $5,000 or $10,000 intervals?

Consumer Preferences: What are the most common exterior and interior color choices among new vehicles?

Predictive Modeling: Can we construct an accurate machine learning model to predict vehicle price based on the available features?

These questions aim to uncover actionable patterns in automotive pricing, enhance understanding of market segments, and develop a framework for predictive pricing in vehicle sales.

Literature Review:

Lapatta, N. T., & Husin, A. (2024). *Predicting potential car buyers using logistic regression algorithm*. Sistemasi: Jurnal Sistem Informasi, 13(3), 1147–1156. http://sistemasi.ftik.unisi.ac.id

Lapatta et al provide a fantastic source by analyzing a similar vehicle dataset using logistic regression. Previous attempts have used other modelling methods such as Naive Bayes and ARIMA, but they have focused on sales. Instead Lapatta et al appear to use logistic regression as a means to predict consumer behaviour. Their final logistic model incorporating factors such as income, age, gender and marital status reached 95% accuracy.

Kavitha, V., Sai, P. D., Revathi, S., & Reddy, P. V. K. (2019). Car buying decision using machine learning algorithms. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 7*(IV), 293–298. https://doi.org/10.22214/ijraset.2019.4494

Kavitha et al provide a good high level overview of various machine learning models used to predict car purchasing. The group tried using logistic regression, K-Nearest Neighbor and decision tree classification. Of the three, logistic regression performed poorly and decision tree classifier performed the best with a 96.5% accurate model.

Ledesma-Alonso, R., & Becerra-Nuñez, G. (2024). Electric or gasoline: A simple model to decide when buying a new vehicle. *Environmental Research Communications, 6*(2), 025015. https://doi.org/10.1088/2515-7620/ad2949

Ledesma-Alonso et al build a non-machine learning model using simple linear mathematics to predict breakeven points of car ownership and purchase price. They focus on four consumer driven behaviours including: driving habits, economic trends, duration of car ownership and vehicle depreciation. They conclude that electric vehicles are better for high mileage trips in congested areas, while gasoline vehicles are better for low mileage and light traffic areas.

Ong, A. K. S., Cordova, L. N. Z., Longanilla, F. A. B., Caprecho, N. L., Javier, R. A. V., Borres, R. D., & German, J. D. (2023). Purchasing intentions analysis of hybrid cars using random forest classifier and deep learning. *World Electric Vehicle Journal, 14*(8), 227. https://doi.org/10.3390/wevj14080227

Ong et al provide a resource for modelling Filipino car purchasing behaviour towards hybrid vehicles. They use a variety of machine learning models to gauge accuracy including: decision tree, random forest and deep learning neural network. Their findings concluded that deep learning neural networks provided the most accurate model at 96.60%, with random forest classifiers coming second at 94% accuracy. For CIND820, deep learning neural networks may be beyond the scope of the course, but random forest classifiers could be investigated.

Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). *Application of machine learning techniques to predict the price of pre-owned cars in Bangladesh*. Information, 12(12), 514. https://doi.org/10.3390/info12120514

Amik et al built a machine learning model to predict the price of pre-owned vehicles in Bangladesh. They covered a variety of machine learning models including: linear regression, LASSO regression, decision tree, random forest and XGBoost. Among the five machine learning methods tested, XGBoost generated the highest accuracy with a R-squared value of 91.32%. The second best model was a random forest model with an R-squared value of 90.14%. For CIND820, random forest appears to be performing well among peer-reviewed sources.

Al-Turjman, F., Hussain, A. A., Alturjman, S., & Altrjman, C. (2022). *Vehicle price classification and prediction using machine learning in the IoT smart manufacturing era*. Sustainability, 14(15), 9147. https://doi.org/10.3390/su14159147

Al-Turjman et al focus on testing machine learning models to predict vehicle prices based on the features of those vehicles. They test four machine learning models including: linear regression, support vector machine (SVM), decision tree and neural networks. Unfortunately, the models alone performed quite poorly and an “ensemble” or combination of machine learning models was required. Standalone, the models performed poorly with SVM performing the best at 49% accuracy. The group used a combination of models that ended up with a SVM ensemble model predicting with 90% accuracy.

Data Description:

I have selected a vehicle dataset consisting of new vehicles available for purchase. The dataset consists of 18 variables for each vehicle in the table below with a total 1002 data points.

Variable	Description	Data Type
name	Name of vehicle	string
description	Description given by dealership and highlighted features	string
make	Make / brand of vehicle	string
model	Model of vehicle	string
type	New vehicle or used vehicle (Dataset only contains new)	string
year	Model year of vehicle	integer
price	Dealership price of new vehicle	float
engine	Model of engine in vehicle	string
cylinders	Number of cylinders (if applicable)	float
fuel	Fuel source of vehicles (Gasoline, electric etc.)	string
mileage	Number of miles on the vehicle odometer	float
transmission	Vehicle transmission	string
trim	Vehicle trim level for model	string
body	Vehicle body type	string
doors	Number of doors on the vehicle	float
exterior	Exterior paint color of the vehicle	string
interior	Interior paint color of the vehicle	string
drivetrain	Drivetrain of the vehicle (All wheel drive, front wheel drive etc.)	string

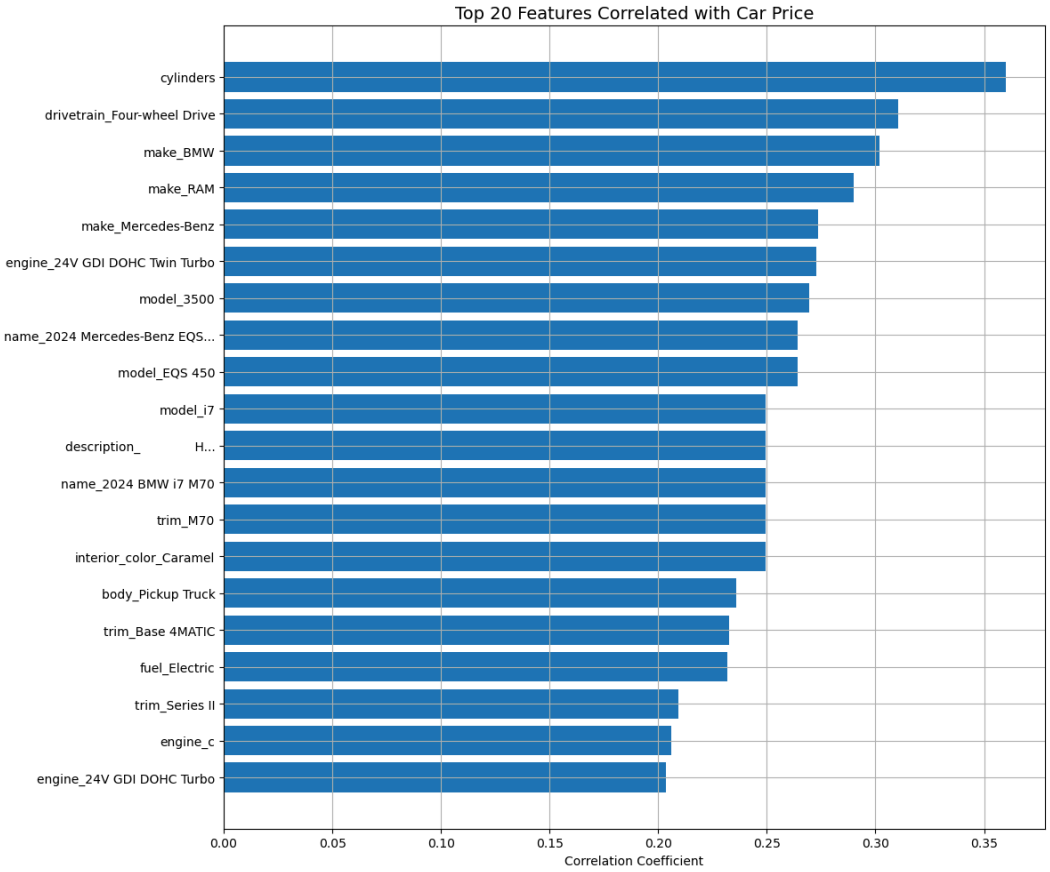
Abnormal & Missing Data:

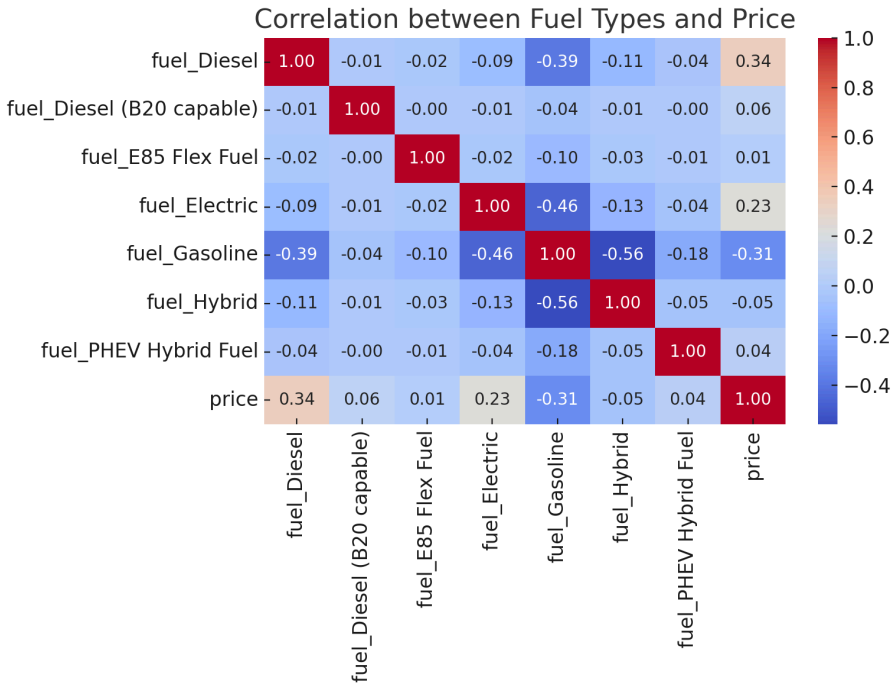
Initial examination of the raw dataset revealed several instances of missing data as outlined below with the approach to fix:

name (0)	N/A
description (56)	Substitute null values with a placeholder such as “No Description”
make (0)	N/A
model (0)	N/A
type (0)	N/A
year (0)	N/A
price (23)	Drop rows missing price, as price is the dependent variable for modelling
engine (2)	Substitute null values with a placeholder such as “Unknown Engine”
cylinders (105)	Substitute median value to limit outliers
fuel (7)	Substitute mode value as the most common fuel type
mileage (34)	Since mileage varies in this dataset, median limits skewing the data
transmission (2)	Substitute mode value as the most common transmission type
trim (1)	Substitute null values with a placeholder such as “Unknown Trim”
body (3)	Substitute mode value as the most common body type
doors (7)	Substitute mode value as the most common door configuration
exterior (5)	Substitute null values with a placeholder such as “Unknown”
interior (38)	Substitute null values with a placeholder such as “Unknown”

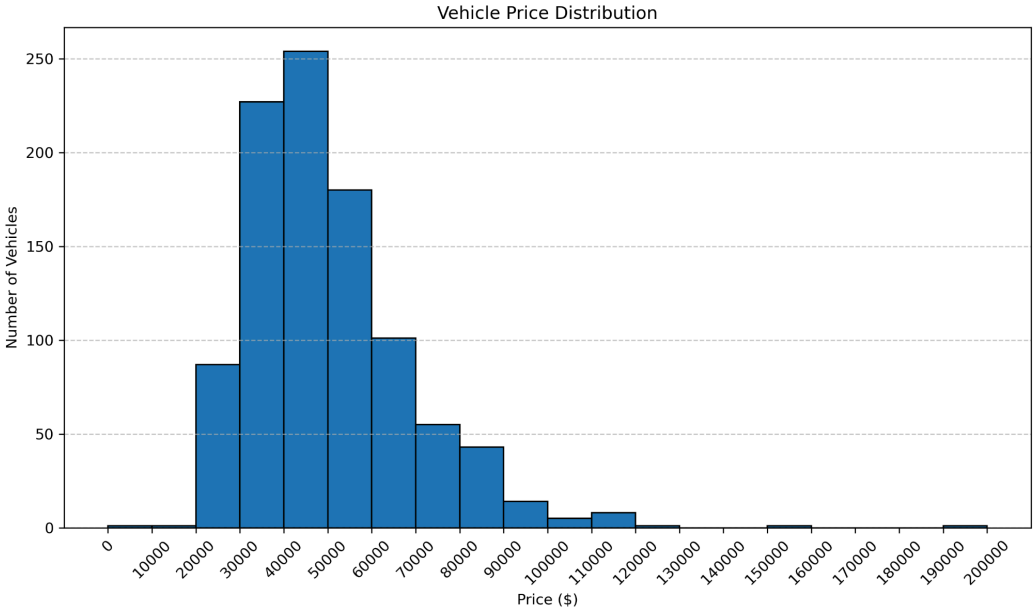
drivetrain (0)	N/A

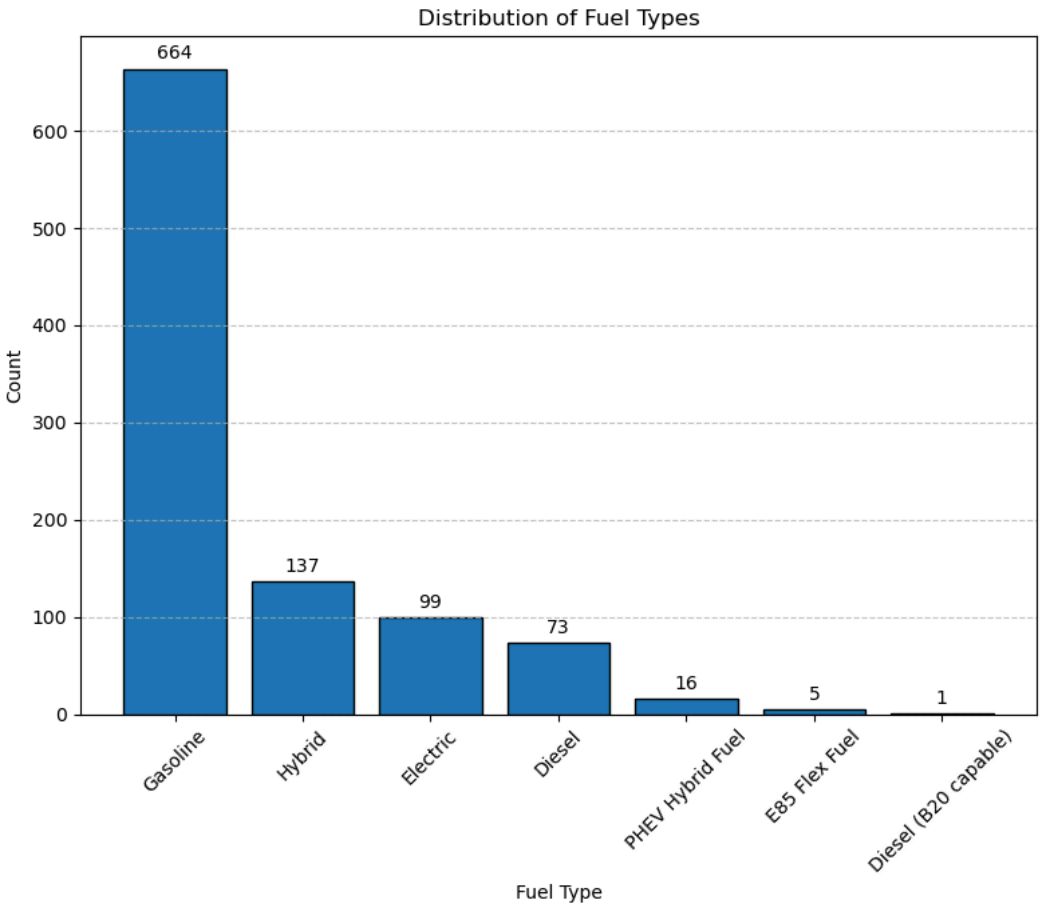
Correlation:

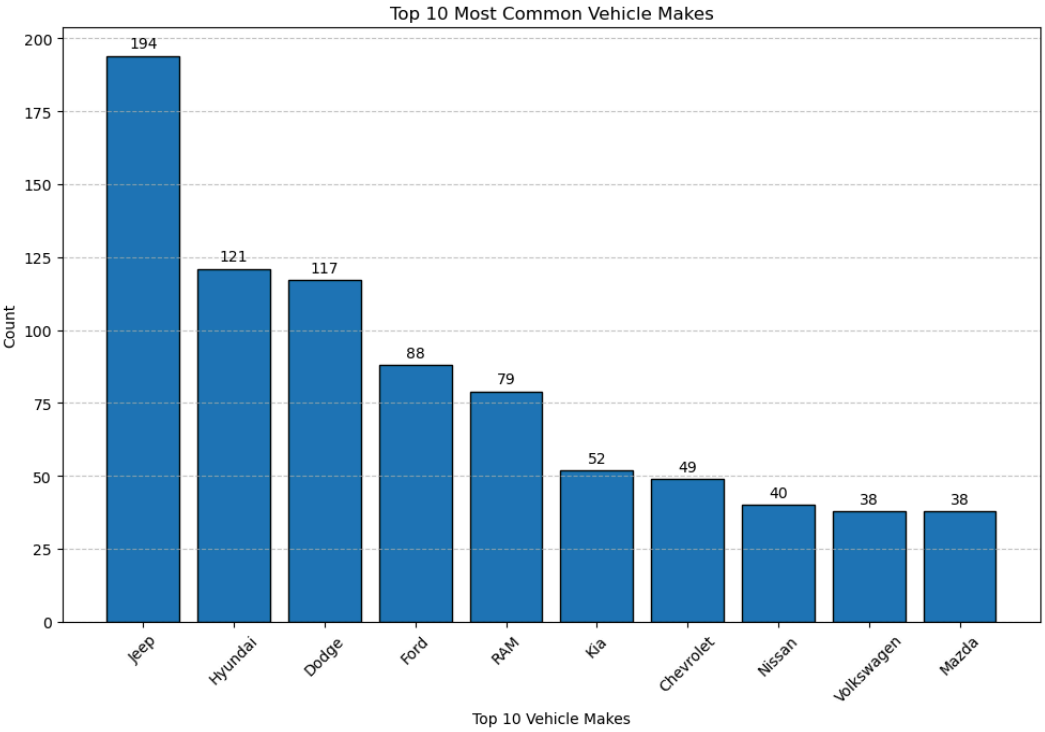


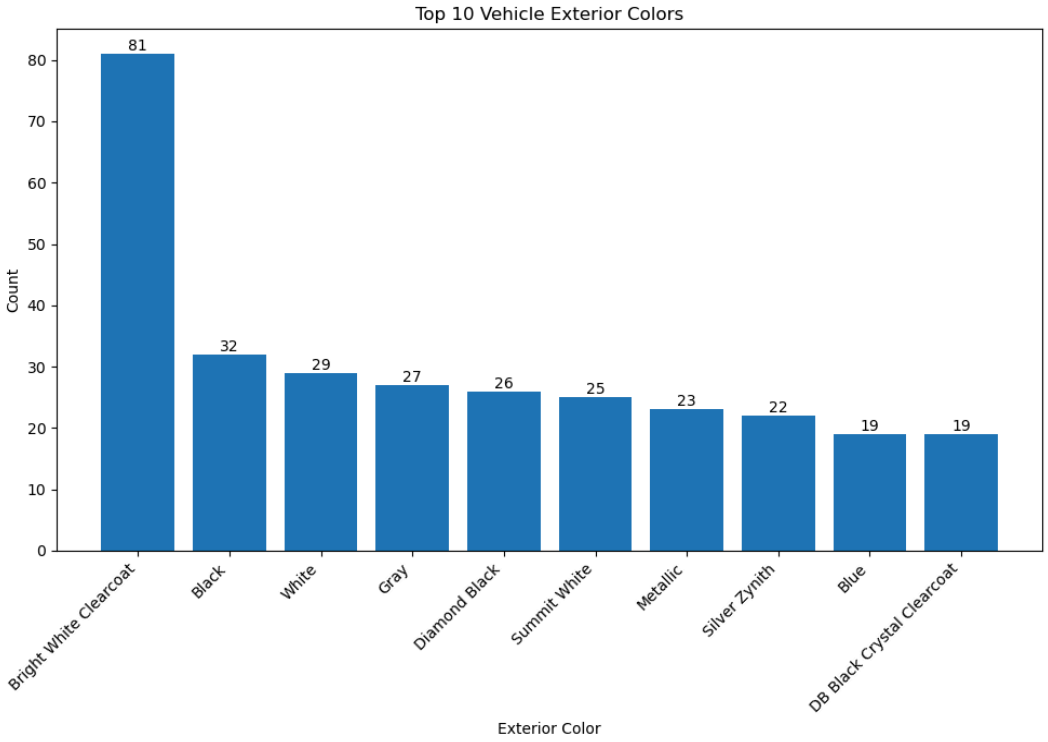


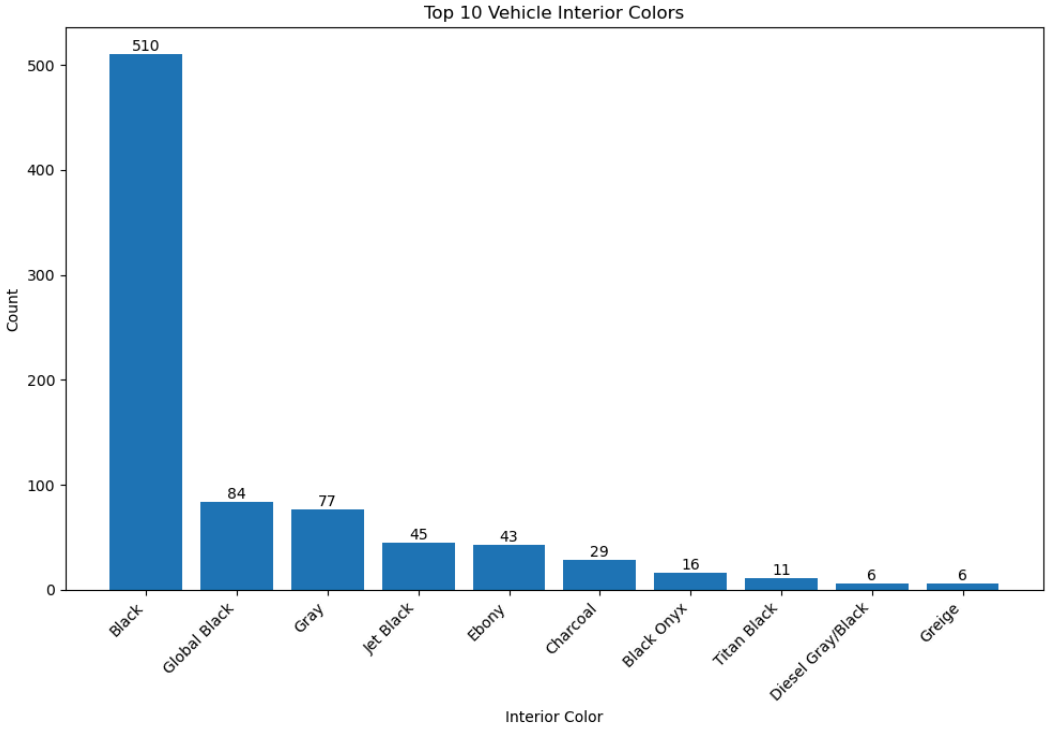
Other Data Visuals:



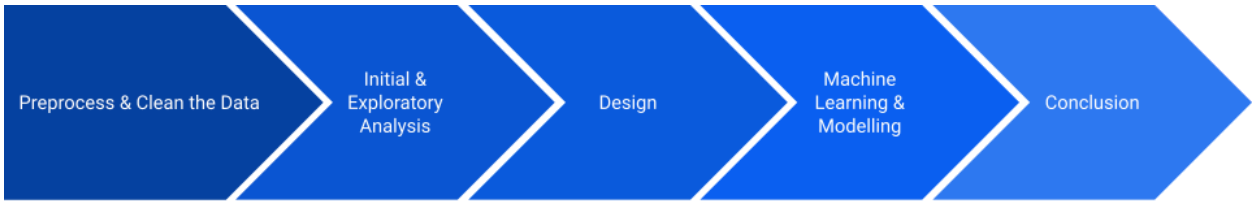

Distribution of Fuel Types
Count
Fuel Type
664
137
99
73
16
5
1
Gasoline
Hybrid
Electric
Diesel
PHEV Hybrid Fuel
E85 Flex Fuel
Diesel (B20 capable)


Top 10 Most Common Vehicle Makes
Count
Top 10 Vehicle Makes
194
121
117
88
79
52
49
40
38
38
Jeep
Hyundai
Dodge
Ford
RAM
Kia
Chevrolet
Nissan
Volkswagen
Mazda


Top 10 Vehicle Exterior Colors
Count
Exterior Color
Bright White Clearcoat
Black
White
Gray
Diamond Black
Summit White
Metallic
Silver Zynith
Blue
DB Black Crystal Clearcoat
81
32
29
27
26
25
23
22
19
19


Top 10 Vehicle Interior Colors
Count
Interior Color
Black
Global Black
Gray
Jet Black
Ebony
Charcoal
Black Onyx
Titan Black
Diesel Gray/Black
Greige
510
84
77
45
43
29
16
11
6
6

Approach:



Examine the data with summary statistics and descriptions.

Search for missing data and deal with numeric & categorical missing values.

Look for outlier data that may have an effect on numerical means, , median and modes.

Univariate analysis. Look to see how categorical data will be encoded for modelling, without including too many categorical selections.

Bivariate analysis. Look for correlations and scatter plots to identify potential modelling opportunities.

Multivariate analysis may be used for statistical analysis.

Clustering to identify potential pair groupings within the dataset. For example, budget friendly prices and luxury prices.

Normalizing the data for the machine learning model to provide accurate models.

Feature scaling as needed.

Splitting the data into training and test data sets (80-20).

Cross-validation to ensure the data is accurate for the model and that the model is processing the data cleanly.

Classification and regression can both be used to analyze the data. Namely, classification for grouping and regression for price prediction. Logistic regression and random forest classifiers to be focus.

R squared and goodness of fit for regression analysis.

Confusion matrix for classification model. Recall, precision and accuracy can also be evaluated for the classification model.

The process begins with preprocessing and cleaning the dataset by using summary statistics to understand its structure. Missing values are handled appropriately, and outliers are identified to reduce their influence on model accuracy. This is followed by exploratory analysis, including univariate and bivariate analysis to study feature distributions and relationships with the target variable, supported by visual tools like scatter plots and correlation matrices.

In the design phase, we explore patterns in the data, such as grouping vehicles by price range or type, and apply normalization and feature scaling as needed. This ensures the dataset is well-prepared for modeling, especially for algorithms like logistic regression that are sensitive to feature scales.

For modeling, we split the data into training and testing sets and apply cross-validation for evaluation. Both logistic regression and random forest models are implemented based on whether the target is categorical or continuous. Model performance is then assessed using metrics such as R-squared for regression or accuracy, precision, recall, and confusion matrix for classification, helping to select the most effective and interpretable model.

References:

Al-Turjman, F., Hussain, A. A., Alturjman, S., & Altrjman, C. (2022). *Vehicle price classification and prediction using machine learning in the IoT smart manufacturing era*. Sustainability, 14(15), 9147. https://doi.org/10.3390/su14159147

Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). *Application of machine learning techniques to predict the price of pre-owned cars in Bangladesh*. Information, 12(12), 514. https://doi.org/10.3390/info12120514

Beliveau, M., Rehberger, J., Rowell, J., & Xarras, A. (2010). *A study on hybrid cars: Environmental effects and consumer habits* (Undergraduate Interactive Qualifying Project, Worcester Polytechnic Institute). https://digital.wpi.edu/concern/student_works/1v53jx69g

Fujita, K. S., Yang, H.-C., Taylor, M., & Jackman, D. (2022). Green light on buying a car: How consumer decision-making interacts with environmental attributes in the new vehicle purchase process. *Transportation Research Record, 2676*(7), 743–762. https://doi.org/10.1177/03611981221082566

Jayaraman, K., Wong, W. Y., Seo, Y. W., & Joo, H. Y. (2015). Customers' reflections on the intention to purchase hybrid cars: An empirical study from Malaysia. *Problems and Perspectives in Management, 13*(2), 304–312.

Kavitha, V., Sai, P. D., Revathi, S., & Reddy, P. V. K. (2019). Car buying decision using machine learning algorithms. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 7*(IV), 293–298. https://doi.org/10.22214/ijraset.2019.4494

Lapatta, N. T., & Husin, A. (2024). *Predicting potential car buyers using logistic regression algorithm*. Sistemasi: Jurnal Sistem Informasi, 13(3), 1147–1156. http://sistemasi.ftik.unisi.ac.id

Ledesma-Alonso, R., & Becerra-Nuñez, G. (2024). Electric or gasoline: A simple model to decide when buying a new vehicle. *Environmental Research Communications, 6*(2), 025015. https://doi.org/10.1088/2515-7620/ad2949

Ong, A. K. S., Cordova, L. N. Z., Longanilla, F. A. B., Caprecho, N. L., Javier, R. A. V., Borres, R. D., & German, J. D. (2023). Purchasing intentions analysis of hybrid cars using random forest classifier and deep learning. *World Electric Vehicle Journal, 14*(8), 227. https://doi.org/10.3390/wevj14080227