Table of Contents:

1. Business Context

Business Context:

The client company is a car dealership, that manages a diverse inventory of vehicles spanning multiple makes and models. The dealership maintains data on different vehicle characteristics.

The dealership needs me to analyse the inventory data to uncover insights into customer preferences, customer segregation and pricing trends. They want to understand which car makes are in high demand, and how the vehicle characteristics are influencing their price, and how inventory can be optimised to serve budget-conscious, mid-range and premium buyers.

Additionally, they require me to build a predictive model that can correctly predict the price of a vehicle based on the characteristics of the vehicle which will help in data-driven pricing strategies and inventory management. This analysis will help the dealership improve sales efficiency, align marketing strategies with customer demand and maximise profitability across all market segments.

2. Data Exploration and Preprocessing

Data Exploration:

1. The data has mainly 3 data types (float, object and int).
2. The data has 11914 rows and 16 columns.
3. The data has missing values in some columns. There are 3 missing values in Engine Fuel Type, 69 missing values in Engine HP, 30 in Engine Cylinders, 6 in number of doors and 3742 in market category.
4. The categorical features in the dataframe are Make, Model, Engine Fuel Type, Transmission Type, Driven Wheels, Market Category, Vehicle Size and Vehicle Style and the numerical features are Year, Engine HP, Engine Cylinders, Number of Doors, highway MPG, city MPG, Popularity and Price.
5. The most highly correlated features with price are engine horsepower, year, engine cylinder and number of doors. These features can be really useful while predicting the price of a car based on its characteristics and also for exploratory data analysis.
6. There are a lot of outliers in the different columns in the dataset. The column Year has 661 outliers, Engine HP has 509 outliers, Engine Cylinders has 357 outliers, highway MPG has 192 outliers, city MPG has 316 outliers, Popularity has 881 outliers and Price has 996 outliers.
7. There are some rows with transmission type as unknown.

Data Preprocessing:

1. The numerical missing values are filled with the median of the numerical columns, and the categorical missing values are filled with mode of the categorical columns.

2. The outliers have been identified using the IQR method and have been filtered out of the dataset, by only selecting the data which was less than the upper bound and greater than the lower bound.

3. The records which had Transmission Type as unknown have been filtered out.

3. Explanatory Data Analysis and Model Building

1. **Which car makes represent the premium section of the market and what does this reveal about their market positioning?**

Ans: The analysis shows that car makes like Maserati, Tesla, Alfa Romeo, Lotus, Cadillac, BMW, Genesis, Land Rover, Infiniti and Lincoln represent the premium section of the listings. All these car makes have a higher average price than the other car makes. So, these car makes represents the premium section of the market and falls into market categories such as exotic, high performance, luxury, performance and crossover. This can help with inventory management, targeted marketing and pricing strategies especially targeted towards premium vehicle buyers.

2. **How does engine fuel type influence the average price of vehicles and what insights can it provide for pricing strategy and marketing?**

Ans: From this analysis it can be seen that vehicles with engine fuel types such as flex-fuel (premium unleaded required/E85), flex-fuel (premium unleaded recommended/E85), premium unleaded (required), diesel, flex-fuel (unleaded/natural gas), premium unleaded (recommended) and electric have a high average price and are associated with market categories like flexible fuel, diesel, hatchback, performance and luxury. Vehicles with engine fuel types such as flex-fuel (unleaded/E85), natural gas and regular unleaded have a medium to low average price and are associated with market categories like flexible fuel and crossover. This information will help for targeted marketing strategies for different customer groups based on their budget and preferences.

3. **How does transmission type relate to vehicle pricing and listing frequency, and what does this reveal about buyer preferences across different budget segments?**

Ans: The analysis shows that cars with automatic transmission types have the most listings and have a mid-range average price showing general buyer preference. Manual transmission types are less common but are the most affordable with the lowest price range. Cars with automated manual are the next most common in the listings, which also has a mid-range average price. the rarest are the cars with direct drive which have a high average price. This analysis will help in targeted marketing for different groups of buyers with different budgets and preferences.

4. **How does horsepower relate to vehicle pricing, and which car makes have the highest average horsepower and which buyer segments can they be marketed to?**

Ans: The high correlation between engine horsepower (Engine HP) and price shows that the greater the horsepower higher will be the price. The analysis shows that luxury car makes such as Genesis, Maserati, Cadillac, Infiniti, BMW, Lincoln, Mercedes-Benz, Lexus and Hummer have the highest average engine horsepower and also have premium pricing and belong to the luxury and performance market category. These insights suggest that high horsepower vehicles are in the premium segment and marketing campaigns are more likely to be effective when targeting buyers interested in luxury and performance.

5. **What car makes are the most popular and how do they rank against each other?**

Ans: From the analysis it can be observed that the top 10 most popular car makes among customers are Ford, BMW, Audi, Honda, Toyota, Nissan, Dodge, Kia, Porsche and Cadillac. The 10 least popular car makes are Pontiac, Acura, Infiniti, Buick, Hummer, Alfa Romeo, Scion, Lincoln, Oldsmobile and Genesis. This analysis helps to identify the most popular mainstream and luxury car makes and least popular and less favoured car makes among customers, which would help in targeted marketing of mainstream and niche segments effectively.

6. **How does the model year of a vehicle influence its price?**

Ans: The dealership's inventory shows a clear relationship between model price and year. Older vehicles (1990-2000) have a low average price under $5000, and with very limited variation. The vehicles starting from the year 2001, the average jumps above $20,000 and continue to rise steadily, surpassing $35,000 for models after 2015. Newer vehicles not only have higher average prices but also exhibit wider price distributions, thus having both budget-friendly and premium options. Overall, the dealership can serve budget conscious buyers, mid-range shoppers and premium customers by offering diverse vehicles across model years and price levels.

7. **Create a machine learning model pipeline to predict the prices of cars based on their characteristics**

Ans: A pipeline has been built with a very high r2 score of 86.2%, which shows this model can very accurately predict prices of cars based on their characteristics. The model can be applied to future vehicle listings, allowing the dealership to estimate the prices for new vehicles as they arrive. The pipeline automates preprocessing and model building, ensuring consistent, efficient and data driven pricing decisions over time.

4. Discussion

Overall Strengths:

1. Provides a detailed understanding of the dealership's inventory, linking vehicle characteristics such as horsepower, transmission, fuel type, model year and make to pricing and customer preferences.
2. Identifies market segregation clearly, distinguishing premium, mid-range and budget vehicles.
3. Highlights patterns in popular makes and market segments which can directly inform inventory management and marketing strategies.
4. The machine learning model will be able to predict the price of a vehicle based on their characteristics.

Overall limitations:

1. The data is limited to the dealership's current inventory, so results may not reflect broader market trends or competitor offerings.
2. The popularity of certain car makes may change over time, which is not taken into account in the EDA.

Business Implications and Recommendations:

1. Premium makes attract buyers seeking luxury, high performance and exotic vehicles. These vehicles should be marketed towards buyers who are interested in premium and high-performance cars.
2. Vehicle characteristics like fuel type, transmission type, horsepower and model year significantly influence pricing and appeal to different market segments. Targeted marketing can leverage these characteristics match customer's preferences and budgets.
3. The inventory of available cars varies across a wide range of model years, with varying prices. Vehicle models before 2001 fall into the very low-price segment best for buyers with a very limited budget and vehicles after 2001 are distributed across budget, mid-range and premium buyers, giving flexibility to meet the needs of different buyers.
4. Popularity analysis of the cars reveals the makes of cars which resonate the most and least with customers. The dealership can prioritize promotional efforts and marketing of the most popular car makes to buyers across all price segments and the least popular car makes can be promoted to niche segments.

5. The dealership can use the machine learning pipeline to predict the price of new listings based on the characteristics of the vehicle.

5. Links:

Kaggle Link for dataset: https://www.kaggle.com/datasets/CooperUnion/cardataset

Google Drive Link for Video:
https://drive.google.com/file/d/1xTOU8G_eFoD1BScOqprlwxTgY6PXixGU/view?usp=sharing

GitHub Link: https://github.com/anuronmitra2001/AI-and-Applications-M504D-Project_Anuron-Mitra