A picture containing icon

Description automatically generated

| *Student Full Name* | **Yuri Braga** |
| --- | --- |
| *Student Number* | **sba24328** |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 2 |
| *Assessment Due Date* | *15th December 2024 (23:59 IST)* |
| *Date of Submission* | Thursday, 02nd January 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

### 

[**1. Strategic Overview of the Business Problem 2**](#_8t4oeo4hrubj)

[**2. Project Plan 3**](#_pootlg35fkz6)

[**3. Business understanding 5**](#_77gf4uiu0tt)

[Stakeholders 5](#_6oix3vogrize)

[**4. Data Understanding 6**](#_uyxzz1p4unl7)

[Data Details 7](#_39jxzhfb9cle)

[Summary Statistics 9](#_j9ezn9izgbz5)

[Analysis of the Categorical Features 11](#_7w6cuajrla40)

[**5. Data preparation 14**](#_b4m5ayemxawl)

[Data engineering 14](#_5fz7y7nmufzi)

[Conversion to Numeric Values 14](#_f80eubm5buut)

[Handling of Missing Values 14](#_st8g3258c7r6)

[Distributions of Numerical Features 15](#_748na86ytrei)

[Correlation Analysis Report 19](#_7y9b45p790qj)

[Dependency Relationships Between Predictors 20](#_o21lxiv5uw4d)

[Multicollinearity 21](#_7qa8cbjtm9iq)

[Why One-Hot Encoding? 23](#_jgd0r5rem3eu)

[Feature analysis in predicting vehicle prices 24](#_6w4xk4z7f17t)

[Summary of the Outlier Capping Process 25](#_ag98efen2c4d)

[Summary of Capping Ranges and Their Impacts 26](#_8ni8n4ra48yw)

[Derived Feature Engineering Report 27](#_cwqf2y2v6x65)

[Feature Transformation and Scaling Report 28](#_qygsph2ewuff)

[**6. Findings & Recommendations 29**](#_tzfa5ub3g0tu)

[Findings 29](#_9b3itwkvchvy)

[Recommendations 29](#_pglp7fdu21et)

[**References 30**](#_ujkpc2jwmb8z)

[Github: https://github.com/bragayuri/strategic-thinking-ca-1 32](#_8uwq18l15utl)

### 

### **1. Strategic Overview of the Business Problem**

The used car market is expected to reach $460 billion by 2029, registering a CAGR of over 10% over the forecast period. And with demand expected to continue growing throughout 2025, accompanied by rising new car prices, still a reflection of the COVID pandemic that has stalled the auto industry, there is a certain inclination among consumers to buy used cars, strengthening the entire sector. The growth of digital platforms for buying and selling used cars has further strengthened consumer confidence in buying used vehicles.

While the industry keeps growing, more dealerships are joining Saas solution platform to process online orders and transactions. It is noticed that there is a need for a tool that can be used to price a used car.

For instance when a customer offers their own car in exchange as part of the payment, it is difficult for a seller to understand whether the price that was sent or asked for the used car is compatible with a list price for that model.

**Solution**

**Price prediction tool in the used car market**

This capstone project aims ***to deliver an accurate price prediction tool in the used car market***. Once implemented, this tool would be available specially for dealerships to handle vehicle pricing. The tool can offer several benefits as such:

**Environment and Sustainability**

Accurately priced used cars can help with environmental sustainability as a longer vehicle lifespan reduces the need for new car production which involves higher emissions and uses of natural resources.(ConsumerAffairs, 2024)

**Customer Satisfaction**

When a customer submits a price for their used car or is looking to purchase one, it is important that the price is fair to both parties (buyers and sellers). The price assigned needs to be in line with the market, ensuring that the consumer feels confident in their decision.(International Journal of Research Publication and Reviews, 2021)

**Market Efficiency**

Appropriate pricing helps the market remain efficient and stable. It helps to maintain stable levels without excessive prices causing a distortion of the market with unfair buying and selling conditions.(Economics Discussion, n.d.)

**Increase profitability**

Since accurate price prediction allows for more effective management of inventory control. By applying real market values ​​to vehicles, dealers can improve their inventory levels and work on turnover strategies, thus increasing overall profitability.(Sulaiman, Mustapha and Shareef, 2022)

### **2. Project Plan**

This project aims to cover the following deliverables:

* A deployed model that predicts used vehicle price;
* A web-component so dealerships can integrate the tool to their website.
* A backend to communicate with the LLM model and the front-end

This capstone project will use the Agile philosophy and Scrum methodology.

Developed in the early 1990s, Scrum is an Agile framework that helps to generate value through its adaptive solutions for complex problems.(Schwaber and Sutherland, 2020).

**Scrum Framework**

When applying Scrum on a project, the load of work is divided into Sprints, which are fixed-duration iterations and it typically lasts two weeks. Scrum involves different roles and process:

* A product owner requires the work for a complex problem creating a product backlog.
* The scrum team turns a selected part of this into an increment of value during the sprint.
* Stakeholders along with the scrum team will review the results and if necessary adjust the next sprint.
* The process should repeat until the goal is accomplished.

Implementation

Timeline and Sprints

This capstone project aims to have two-week Sprints. On each sprint here is the list of deliverables:

**Sprints**

| 1 | 2 | 3 | 4 | 5 | 6 |
| --- | --- | --- | --- | --- | --- |
| Data Acquisition and EDA | Data cleaning and preprocessing | Exploratory Data Analysis and Initial Modeling | Advanced Modeling and Initial Results | Model Refinement and Validation | Backend development and Integration |

| 7 | 8 | 9 | 10 | 11 |  |
| --- | --- | --- | --- | --- | --- |
| Front-end design and development | Integration Front-end, Back-end, LLM | Deployment | Testing | Finish Documentation |  |

**Sprint 1**

Goal: Acquire necessary vehicle datasets and conduct initial exploratory data analysis to get information about the data and its patterns.

**Sprint 2**

Goal: Data cleaning and processing. Important to remove missing values or incorrect data. Perform the necessary data transformations to prepare data for model training.

**Sprint 3**

Goal: EDA to refine hypotheses and initial predictive models built.

**Sprint 4**

Goal: Develop more complex models. Implement analytics techniques to improve prediction accuracy.

**Sprint 5**

Goal: Refine models by conducting validation to ensure model reliability.

**Sprint 6**

Goal: Develop a backend service in Javascript and Node.js to integrate with the LLM service.

**Sprint 7**

Goal: Design and Develop a front-end UI using javascript and VueJs to integrate with the backend. It should gather vehicle data so the tool can predict its value.

**Sprint 8**

Goal: Integrate end to end service flow, front-end with backend service and the LLM service.

**Sprint 9**

Goal: Deploy the tool using Amazon S3 buckets.

**Sprint 10**

Goal: Apply regression testing to verify the results.

**Sprint 11**

Goal: Create a detailed documentation with the EDA, report and findings.

### 3. Business understanding

### 

This capstone project would directly impact decision-making and improve profitability.

Key impacts this tool would bring:

**Accurate pricing**

Undervaluing or overprice a car can disrupt the transaction and bring user dissatisfaction. This tool will enhance pricing accuracy which is essential to improve customer trust and increase transaction volumes.

**Operational Efficiency**

Once a trade-in is received, Dealerships can assess those and set fair sales prices quickly reducing the time vehicles spend in inventory reducing holding costs.

**Higher Profitability**

When the price asked for a vehicle is closely aligned with the market conditions and car valuations, dealerships will optimize their profit margins.

**Liability**

On digital marketplaces, providing a pricing tool that offers AI can help to stardize car prices which leads to market transparency. This leads to greater trust among end users, attracting even more traffic to the platform.(Information Age, 2022)

**Single point data for negotiation**

Both ends, buyers and sellers can use data from the same tool to inform their negotiations, which leads to a more balanced and fairer transaction based on current market tendencies.

#### **Stakeholders**

The price prediction tool will benefit the following stakeholders:

**Digital selling platforms**

Digital marketplaces that enable users to buy or sell their used cars can integrate this service to deliver instant listings pricing.

**Dealerships selling used cars**

Being the primary users of the tool, Dealerships often need to provide instant evaluation over listings on a daily basis. This process might happen multiple times along the day and require precision with an accurate result. This predicting tool can complement or replace any already in place pricing methodology with the goal to maximize profitability and improve efficiency.

**Individual sellers**

Private users that want to trade-in their used vehicle can use this tool to get instant evaluation before a trade-in is submitted. This brings a fair negotiation with potential buyers with a better understanding of their vehicle value.

### **4. Data Understanding**

**Data Source**

The data set used during this study is the “Car details v3.csv” extracted on [car-dataset](https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho) which is subject to the Open Database License (ODbL). In accordance with such license the users are allowed to distribute, alter and utilize the database but observing the licensing conditions for attribution requirements and akin sharing constraints. This dataset is very relevant for the study as it contains adequate and organized information on used vehicles, which is indispensable in the creation of a potent price estimation tool.

**Licensing details**

The dataset is made available under the Open Database License (ODbL) v1.0 by the Open Knowledge Foundation.

This license makes provisions for the right to:

* Apply the database for any purposes, whether commercial or noncommercial.
* Change and develop new works from this database.
* Replicate or pass the database or its modified works on to third parties under the same licensing regime.

**The conditions include**

* Attribution: credit must be given to the original author of the database for each use or derivative work.
* Share Alike: ODbL licensing that is attributed on any changes made or any derivative database created must be sustained on that same ODbL licensing.

**Compliance Plan**

Concerning the licensing restrictions, this project will: Adhere to the licensing restrictions, and apply appropriate strategies. Proper acknowledgment of origin of sources will be made by ensuring that relevant details such as names of datasets and links to their licensing terms are provided in the documentation of the project.

The dataset contains the history of used car sales collected from various online selling websites. These sites gather and store lots of data pertaining to sales and listings making it a comprehensive data set of used cars market in all the territories over a wide time frame and includes many details.

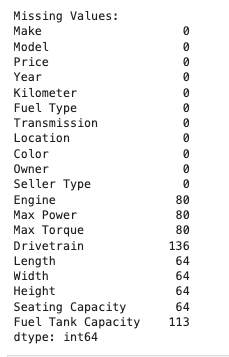
#### Data Details

**Features description**

In the given dataset, there are 20 features encapsulating the attributes of used cars. A brief narration of those features is given below:

| Feature | Description |
| --- | --- |
| Make | Manufacturer’s name and well-known trademark (e.g, Toyota & Honda). |
| Model | The name of the model (for example, Civic, Swift Dzire). |
| Price | The selling price of the car (the variable to explain). |
| Year | The year when the car was manufactured; it is useful in determining the age of a vehicle. |
| Kilometer | Total distance covered in kilometers which shows the usage of the car. |
| Fuel Type | The kind of fuel utilized (Petrol, diesel, LPG). |
| Transmission | The kind of transmission employed (Manual, Automatic). |
| Location | The state or city in which the vehicle is placed for sale. |
| Color | Exterior color of the car. |
| Owner | The number of owners e.g. First-owner, Second-owner etc. |
| Seller Type | It specifies whether the seller is a private individual or a dealer. |
| Engine | Engine capacity in CC which gives strength to power and efficiency. |
| Max Power | The Engine Power maximum output in BHP. |
| Max Torque | The maximum torque generated, as measured in Nm or Kgm. |
| Drivetrain | The vehicle striving factor e.g. Front-Wheel Drive. |
| Length | Length of the car in millimeters. |
| Width | The width measure of the vehicle in millimetres. |
| Height | The height measure of the vehicle in millimetres. |

**A Detailed Review and Description of the Data**

****

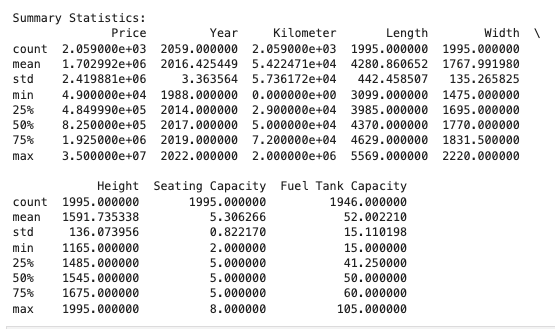
Missing values on the following attributes:

* **Engine, Max Power, Max Torque**: Each of these has 80 missing values – about 3.9 per cent in the data set.
* **Drivetrain**: Absent in 136 rows, which is near 6.6% of the dataset as a whole.
* **Length, Width, Height, Seating Capacity**: Each of these has 64 missing records. This is about 3.1 percent of the total observations.
* **Fuel Tank Capacity**: Absent in 113 rows hence about 5.5 % of the dataset.

**Interpretation**

Although a great portion of the values are complete and accounted for, the most likely dry areas in the data set such as **Engine, Max Power** and **Max Torque** could prove to spoil the quality of the models which will be trying to predict. There will be some imputation techniques that might be necessary to cope with these voids.

#### Summary Statistics



First, it would be best to analyze and see the statistical aspects of the continuous attribute features below.

**Price (Target Variable):**

* Range: Prices from 49,000 and about 35 million with most cars priced at around 1.7 million and an average price of around 825,000.

Observations: This range is large because the dataset is made up of both budget and luxury cars. The high standard deviation also means there is great variability in the price of the cars.

**Year (Manufacturing Year):**

* Range: The vehicles range from being manufactured in 1988 to 2022, with the midpoint being the year of 2017.

Observations: As seen from 75% being 2019, the dataset consists of newer vehicles.

**Kilometer (Driven) Range:** These ranges are from zero km (i.e. ‘almost new cars’) through to two million km, whereby 50,000 km is the mode.

Observations: The overall wide range in mileage provides for all possible usage scenarios which will be vital in evaluation of wear.

**Physical Dimensions (Length, Width, Height):**

* Values observed for length ranges from 3099 mm to 5569 mm with the mean being 4280 mm.
* Values observed for width ranges from 1475 mm to 2220 mm with the mean being 1768 mm.
* Values observed for height ranges from 1165 mm to 1995 mm with the mean being 1592 mm.

Observations: These sizes tell the kind of vehicles present in the dataset such as compact, sedan, SUV and so on.

**Seating Capacity Range**

* Minimum - 2 seats, maximum - 8 seats and (5) being the median as most vehicles have five seats.

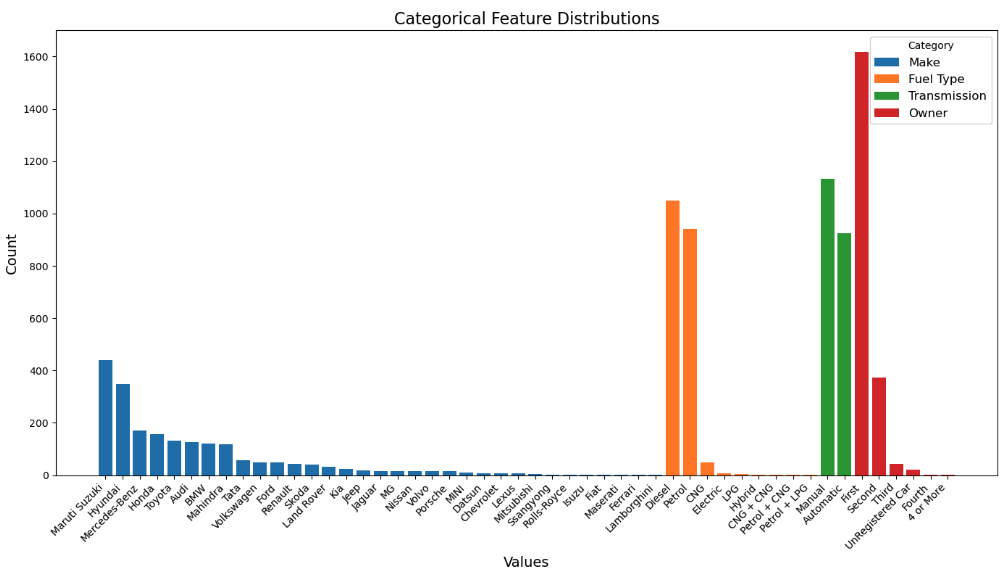
Observations: The dataset predominantly consists of medium segment passenger normal use cars.

**Fuel Tank Capacity Range**

* Minimum - 15 liters, maximum - 105 litters with a mean of about 52 liters.

Observations: It is possible that bigger tanks may be associated with SUVs or some powerful vehicles.

#### Analysis of the Categorical Features



This data set includes various information and that presents the various attributes as well as the variation of the vehicles in the set. So here is an overview of the distinct features and their proportions:

| Category | Detail | Count |
| --- | --- | --- |
| Make | Number of Makes | 34 |
| Make | Maruti Suzuki | 440 |
| Make | Hyundai | 349 |
| Make | Mercedes-Benz | 171 |
| Make | Honda | 158 |
| Make | Toyota | 132 |
| Make | Luxury Brands (e.g., Audi, BMW, Porsche) | Present (low volumes) |
| Make | Exotic Brands (e.g., Lamborghini, Ferrari, Maserati) | Present (1-2 each) |
| Fuel Type | Number of Fuel Types | 9 |
| Fuel Type | Diesel | 1049 |
| Fuel Type | Petrol | 942 |
| Fuel Type | CNG | 50 |
| Fuel Type | Electric | 7 |
| Fuel Type | LPG | 5 |
| Fuel Type | Hybrid | 3 |
| Transmission | Number of Transmission Types | 2 |
| Transmission | Manual | 1133 (55%) |
| Transmission | Automatic | 926 (45%) |
| Owner | Number of Ownership Categories | 6 |
| Owner | First Owner | 1619 (79%) |
| Owner | Second Owner | 373 (18%) |
| Owner | Third Owner | 42 (2%) |
| Owner | Unrecorded Buy | 21 (1%) |
| Owner | Fourth Owner | 3 (<1%) |
| Owner | 4+ Owners | 1 (<1%) |

**Observations:**

Those who drive diesel and petrol cars form the majority and when the two types are put together, they account for over 95% of the sample population.

CNG, Electric and Hybrid alternatively fuelled vehicles are only hardly available in numbers and this is probably an indication to their availability in the market or acceptance in the particular region.

Manual transmission, though close, is more than automatic transmission, portraying a general liking for the manual vehicles which could be a result of its cost or more of a preference in the area of usage.

Most of the cars are first owner cars that are 79%, and these are the cars expected to be sold at a relatively high price.

However second owner cars are relatively tiny in percentage but third owner and up seems rather uncommon which suggests depletion of value in the market.

**Key Insights Market Developments:**

Market Trends: Dataset covers a wide market spectrum dominated by popular brands like Maruti Suzuki and Hyundai, as well as luxury and other specialized brands.

Diesel and Petrol cars are still the two most common types of vehicles on the market in accordance with usual patterns in the market.

**Transmission and Ownership:**

The market differences between manual and automatic transmissions highlight consumer preferences for certain driving styles. In addition to this, vehicles that had only one single owner are highly valued by buyers, as they are assumed to be more reliable and well-maintained.(Quest Journals, 2021)

**Data Diversity**

Data diversity is essential to develop a robust and unbiased model as it enables the model to learn from different examples, improving its capability to generalize to new and unseen data. A study on data diversity conducted by Gong et al. (2019) demonstrates how diversity contributes to improved machine learning performance. Grong highlights that a diverse training provides more information for the model which leads to better generalization. Although the dataset demonstrates strong dominance of the popular vehicle types and fuel types, it also encompasses rare categories (hybrid, luxury brands, etc.) as well which increases diversity.

### 

### 

### 5. Data preparation

#### Data engineering

Creation of Age Column

A study conducted by Milunovich et al. (2023), showed that car age is a key attribute in creating second-hand vehicle price predictions. By converting the Year column to Age and removing the original feature, it is possible to improve the data clarity and prevent redundancy, which removes noise from the model. This will simplify the analysis by providing a vehicle’s age measure which is a vital factor in predictive models.

#### Conversion to Numeric Values

The conversion of Engine and Max power features from a categorical to a numerical value is crucial as machine learning models utilize numerical data. Models require numeric inputs when computing data and such preprocessing enforces that those features can be standardized, scaled and used in mathematical tasks. It also eliminates noise improving the quality of the analysis and modeling.(Geron, 2019).

#### Handling of Missing Values

Missing values handling is essential as it can introduce noise bias which leads to model inaccuracy.

Completing or modeling missing data enforces that models can utilise a complete and consistent dataset resulting in a more reliable prediction.(Géron, 2019)

**Before Handling:** Missing values were found for such features as Engine, Max Power, Max Torque, Drivetrain and some others.

**Imputation Applied:** Numerical features such as Engine and Max Power which had no values were replaced with the respective median of the particular characteristic.

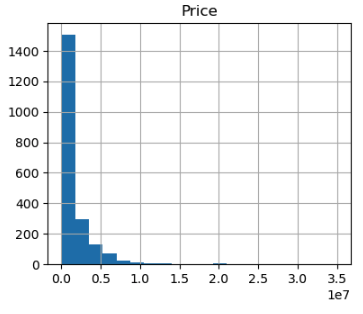
**After Handling:** Missing values in Engine and Max Power have been entirely solved, whereas the other columns such Max Torque, Drivetrain and dimensions (length, width) have been resolved but still require additional attention.

**Remaining Missing Data:** Furthermore, some columns such as Max Torque (80), Drivetrain (136), Fuel Tank Capacity (113) have lost values which can be addressed need targeted handling.

#### Distributions of Numerical Features

The histograms depict the frequency distribution of the numerical features: Price, Kilometer, and Age. The following are the critical insights for every feature:

1. Price (Target Variable)



Most of the prices are focused at the bottom of the range making it a right-tailed price variable frequency distribution.

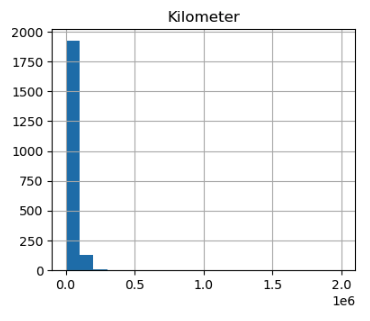
There’s a long tail because there aren’t many cars that are much more expensive than 5,000,000; only high range expensive cars exist.

**Insights:**

The distribution is telling us that there are cars that are overpriced and this can skew the performance of predictive models unless approaches such as log transformations are deployed.

This distribution depicts that there are more budget and middle range cars in the market compared to the high-end cars.

2. Kilometer (Driven)



Kilometer, like Price, is a positively skewed distribution. Most of the values fall under lower mileage.

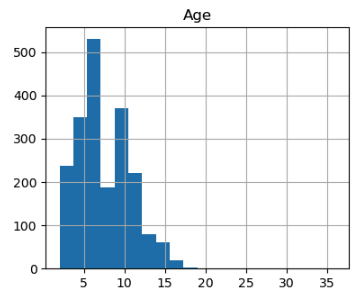
Most of the cars would have a mileage count lesser than a 1000, while a very few would have counted mileage of 1000 or more.

Insights:

This is the trend that can be anticipated because cars with higher mileages are infrequent and are not likely to be resold for much.

The profile of moderate mileages would seem to align with the distribution of used cars with varied conditions in the population.

3. Age (Derived from Year)



The Age feature is usually vertically symmetrical with the most common range for vehicles being 5 years and 10 years older.

The dataset consists of limited quantities of young (up to 2 years) and old (20 ages and above) cars.

**Insights:**

The graphs peak at 5 to 7 years which can be explained by that being the age when vehicle owners sell their cars.

Older cars exist in large numbers but most of them are not retraded and therefore are not actively traded in the market.

**Final Observations**

Outlier distribution in Price and Kilometer:

Both variables are right skewed hence suggesting that the market doesn't lack cars which are overpriced or over-miler but such cars are in low supply, thus a number of cheap and low mileage cars are available.

Age as a Generalization and easy to control variable:

The ages for the cars on sale are mostly centered on the 5 – 10 year range areas which is expected as it shows how long banned owners wait on average before they put their car for resale after purchase. Age is also expected to have a greater impact on the vehicle and its price.

**Possible Improvements**

For features that have a skewed distribution like Price and Kilometer, there exists ways of altering these features for example normalizing the feature or using some transformations such as logarithmic scale which is essential to stabilize variance and achieve a better distribution enhancing model performance.

Whenever high prices or high mileage, among other anomalies, are present in the data, the outliers should be checked in order to avoid biases during analysis or modeling. (Hubert and Van der Veeken, 2008)

#### Correlation Analysis Report

#### 

The correlation matrix is used in order to examine the relationship between all the numerical features in the dataset.(Géron, 2019).Each cell displays the correlation value which varies between -1 (a perfect inverse correlation) and 1 (a perfect correlation).

Here are the major takeaways from the study:

1. Price (Target Variable)

**Strong positive relationship**

**Max Power (0.78)**: More powered muscle cars tend to fetch higher prices and this is logical since high performance cars can be sold at a premium.

**Engine (0.61)**: Higher engine volume almost always propels the price upward since these vehicles are likely to embed enhanced performance or luxury features.

**Oil Tank (0.58):** Cars with higher oil tanks, which are mainly SUVs or bigger automobiles, usually come at a higher price.

**Length and Width (0.56 each):** Bigger sedans or SUVs attract relatively higher prices.

**Moderate negative correlation**

**Age (-0.31):** As cars become older, the likely price of a vehicle reduces because of appreciation of its value.

**Weak or Insignificant correlations:**

**Kilometer (-0.15):** This weak relationship indicates that purchase price may be affected by the factor but no specific firm or high consistency is expected from the entire dataset.

**The height (0.08) and Seating capacity ( -0.04):** What is noted as a very negative correlation with price has a very limited effect on the vehicle’s valuation.

#### Dependency Relationships Between Predictors

**Max Power and Engine (0.87):** It means that the relationship is very strong with a positive number which suggests that the greater the vehicle engine size the greater the vehicle power output. This is expected since engine specifications are rated according to performance.

**Length and width (0.81):** These features are highly correlated, which reflects with the above that the greater the size of the vehicles with the larger dimensions the greater the dimensions.

**Engine and fuel tank capacity (0.80):** Fuel tanks over insulations’ for greater engines usually come so as most sundry weight is needed by those vehicles for range or performing

**Age and kilometer (0.30):** A moderate positive relationship means that the older vehicles are driven more in general, but not in a strong manner.

#### Multicollinearity

Max Power, Engine, Length, Width, Fuel Tank Capacity, among other features, have so much correlation and thus cannot stand uniquely. This multicollinearity will probably have and endanger model effectiveness and interpretability because they are essential features that don’t provide extra information.

**Key insights**

Price predictors. It only occurs that max power, engine, fuel tank capacity, length, width of the vehicle are strong predictors of price that triggers other factors negotiable with price. Thus priority is the focus area of the modeling.

**Depreciation Impact**

Age of cars shows a negative correlation with Price and can be assumed as a case of depreciation.

**Multicollinearity Considerations**

A notable diagonal relationship can be seen across Engine, Max Power, Length and Width. This may imply that the feature space is a bit redundant. It may be worth looking into dimensionality reduction techniques like PCA or even feature selection.

**Unexpected Insights**

Kilometer (mileage) appears to bear minimal effect on Price, furthering the argument that it isn’t a critical factor when valuing vehicles in this database. Further exploration is required to substantiate this claim.

**Feature Reduction Report**

The purpose of the removal of the columns was to enhance simplicity of the dataset while retaining only the features that are most likely to impact outcome the most.

**Removed features:**

- **Location**: Tend to offer minimal or inconsistent impact on the car prices.

- **Colour**: Because colour is a more subjective factor of the aesthetics, it would not overly impact value.

- **Drivetrain**: Where a significant share of this variable is missing, the variable is likely to lack unique predictive power.

- **Height, Length, Width**: These variables discussed all received reciprocal influences with respect to the dimensions such as Engine and Max Power and therefore, were deemed as redundant variables.

- **Fuel Tank Capacity**: Like the preceding variables, Fuel Tank capacity exhibited high correlation coefficients towards Engine and Max Power thus having little considerable value in modeling.

**Reasoning for Dropping Columns**

**Multicollinearity**: The removal of dimensions (car dimensions in this context) which are highly interrelated reduces redundancy whilst improving the efficiency of the model.

**Irrelevance:** Such features like Colour and Location are assumed to have low influences in predicting the prices of vehicles of the models in question.

**Missing Data:** Columns such as the Drivetrain, on the other hand, have very bad missing values that affect the model performance.

**Categorical Feature Encoding Report**

The categorical features Make, Fuel Type, Transmission and Owner were converted to numerical values using One-Hot Encoding.

The first category of each feature was omitted in order to reduce the risk of redundancy on the encoded data known as the dummy variable trap.

#### Why One-Hot Encoding?

Such categorical features cannot directly be applied in the machine learning models which have numerical features as their inputs.

The One-Hot Encoding method allows each of the categories to be represented in separate binary columns thus making the dataset compatible with the algorithms while retaining the categorical features.

There are additional columns representing each unique category in the encoded features; these are in columns where one category of the feature has been dropped.

Example:

Within Fuel Type, these categories of Diesel, CNG, etc., each are converted as types Fuel Type with the suffix of Diesel, CNG respectively retaining the baseline category with none.

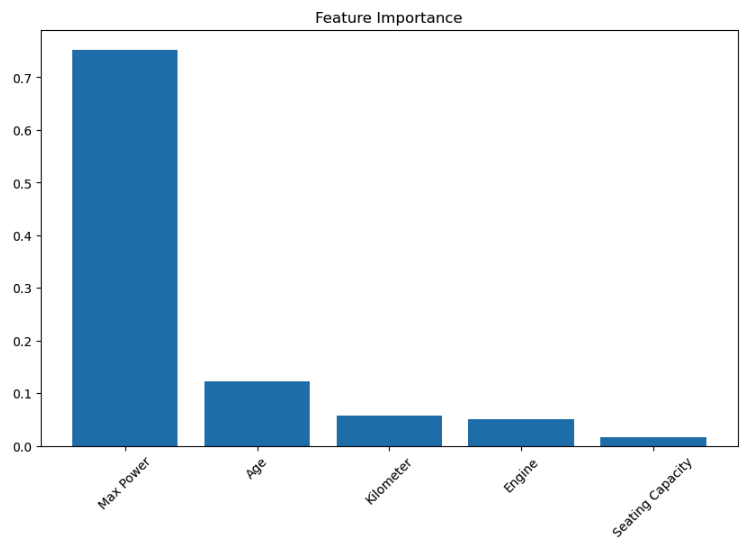
Outcome

The categorical data has now been fully prepped for the predictive modeling loop.

Although the number of columns in the dataset has risen due to the addition of binary columns, this increase guarantees that all categorical data is well accounted for.

This step makes sure that the dataset is prepared in a form for numerical computations which will aid in improving the performance along with the accuracy of the model.

#### Feature analysis in predicting vehicle prices

****

**Findings**

**Max Power:** It is the most important feature with the importance score of 0.7 and above. Cars that provide higher power output are more expensive, which shows the features that customers are willing to have in a vehicle.

**Age:** This is also significant but to a moderate level in this case, as it indicates the pattern of depreciation- in other words newer cars are fresher and of higher value compared to those that are older.

**Kilometers traveled:** This number is the least significant which might mean that mileage is not as important as age and performance in determining the prices of cars.

**Max Power:** This does have an effect, as expected, but is not as strong as the Max Power feature.

**Seating Capacity:** This is also of little value since all vehicles have standard seating arrangements.

**Relevance**

Max Power and Age are the primary determinants of price in the pricing models and the results match with previously collected data and the market (Jain et al., 2021; Nanda et al., 2020).

The reduced mileage is consistent with the other data as there are signs that the majority of buyers are more interested in the performance and overall condition of the vehicle than how it has been used.

Other features that have a very low contribution (like, Seating Capacity) can be removed and a more useful model can be built.

**Conclusion**

**Max Power** and **Age** are the core predictors of price. These insights and predictions are expected and consistent with the trends in the market which guarantees that the models are relevant in regard to the automotive market.

#### Summary of the Outlier Capping Process

Outlier capping was carried out on four important variables in order to improve the homogeneity of the dataset and limit the influence that outlying observations may exert over the model. Attainment of such a control level over exceptional cases of a specific feature area’s values is advantageous as it helps to realize the desired effect in the distribution of that feature case’s data in most cases. (Raschka & Mirjalili, 2022)

#### Summary of Capping Ranges and Their Impacts

| Category | Capping Range | Description | Impact |
| --- | --- | --- | --- |
| Cost | 160,000 – 10,817,640 | Cutting extreme high prices of luxury vehicles to reduce very high outliers. | Shifts focus towards mid-range and premium vehicles, offsetting the absence of luxury pricing. |
| Distance Traveled | 2,459.78 – 145,882 | Removed extremely high and low mileage values that are rare or erroneous. | Normalizes vehicle usage distribution, improving mileage-focused modeling accuracy. |
| Engine Size | 796 – 3,198 | Applied upper caps to engine sizes offered by performance or exotic cars. | Focuses on conventional engine sizes, enhancing the evaluation of common vehicle models. |
| Power Rating of Engine | 47 – 335 | Addressed extreme power outputs from high-performance vehicles to avoid upper-end outliers. | Balances performance and standard cars, reducing bias in price forecasts and maintaining data integrity. |

#### Derived Feature Engineering Report

This report outlines the engineering of new features aimed at giving more discretized information to the dataset for better predictive capabilities. The analysis for the newly created features is as follows.

| Feature | Definition | Key Statistics | Description | Impact |
| --- | --- | --- | --- | --- |
| Mileage Intensity | The ratio of kilometers driven to the age of a vehicle (Kilometer / (Age + 1)). | Range:  409.96 - 20,840.28  Median: 5750  Mean: 6113.48 | Indicates the average kilometers driven per year, representing the intensity of vehicle usage over its service life. | Quantifies the depreciation of vehicles due to age and usage, improving the model's handling of vehicle conditions. |
| CC Price | The price of the vehicle divided by its engine capacity (Price / Engine). | Range:  64.15 - 5441.47  Median: 564.30  Mean: 856.15 | Reflects the cost attributed per cubic centimeter of engine capacity. High values may indicate luxury or efficient vehicles. | Positions vehicles within a targeted pricing structure based on engine size, enhancing performance pricing insights. |

**Data Quality**

No Missing Values: Both features contain complete records after the infinities and this measure is carried out to fill the NaNs with the median.

Statistical Balance: The calculated statistics reveal that both features are dark in distribution yet the level of conducive skewness is not excessive.

**Importance**

Enhanced Predictive Power: Mileage Intensity and Price per cc are both critical parameters within the valuation models uniquely used to impact the price of the vehicle.

Feature Relevance:These features are relevant in increasing the understanding of the relationship between pricing on mileage and engine size with the trending automotive market trends (Raschka & Mirjalili, 2022).

#### Feature Transformation and Scaling Report

To improve the performance of other machine learning models and reduce the volume of skewness in the dataset, log transformations and Min–Max scaling were employed. This step ensures model performance enhancement along with robustness against performance variability.

| Transformation | Applied To | Purpose | Outcome |
| --- | --- | --- | --- |
| Log Transformation | Price, Kilometer, Engine, Max Power | Handles highly skewed data distributions, ensuring log-transformed variables are normally distributed (Han et al., 2022). | Created new columns (e.g., Price, Kilometer), with skewed features now better suited for algorithms sensitive to normality. |
| Min-Max Scaling | All numerical columns | Normalizes feature distributions to a range of 0 to 1, maintaining value proportions. | Standardized all numerical features, ensuring no single feature (e.g., Price) dominates algorithms sensitive to scale differences. |

**Importance**

Lending flexibility to the model, log transformation renders many potential advantages as indicated below:

* Aids in the reduction of extreme value bias.
* Assists in understanding Price, Engine features in a much better manner (Han et al., 2022).
* There are also unique advantages of the Min-Max scaling.
* Proves to be important for neural networks and distance-based models such as k-NN scale which is very important for the learning process.
* Increases the improvement in convergence speed for gradient based optimizers.

### 6. Findings & Recommendations

#### Findings

1. Top price predictors are **Age and Max Power**. The stronger the engine power, the higher the price will be; That also means the older age represents a lower price.
2. The actual condition and how well a vehicle performs tend to overshadow the total number of kilometers and how it has been driven, therefore, it is safe to state that **Mileage** is not that much of an issue.
3. For some models it might be necessary to use dimension-reduction techniques (such as PCA) because large inter-feature relationships (between **Engine** and **Max Power**) might prevail.
4. There are a variety of fuel types available, but **Diesel** and **Petrol** still dominate the market. Hybrid, electric and CNG are of less significance.

#### Recommendations

* **Model Focus**: In any predictive approach engine power and age must be prioritized.
* **Outlier Review**: Revisit some “exotic” cars such as Ferrari, etc. to ensure valuable data is not discarded.
* **Sprints and Planning:** This project covered 3 sprints (1,2,3) of two weeks. The data preparation will require more time than it was planned so an additional sprint might be required.
* **Deployment Strategy**: A research on app deployment platforms such as Amazon S3 is needed.

### 

### **References**

**Used car market and forecast**

* Mordor Intelligence (n.d.). Used Car Market - Growth, Trends, COVID-19 Impact, and Forecasts (2023-2028). [online] Available at:<https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024> [Accessed 02 Nov. 2024].

**Programming for Machine Learning**

* Raschka, S. & Mirjalili, V. (2022). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. 4th ed. Birmingham: Packt Publishing.

**Customer Satisfaction**

* International Journal of Research Publication and Reviews (2021). *A Study on Consumer Behaviour towards Used Cars (within the Special Reference to Coimbatore City)*. [online] Available at:<https://ijrpr.com/uploads/V4ISSUE10/IJRPR18035.pdf> [Accessed 04 Nov. 2024].

**Environment and Sustainability**

* ConsumerAffairs (2024). Electric vehicle prices continue their free fall. [online] Available at:<https://www.consumeraffairs.com/news/electric-vehicle-prices-continue-their-free-fall-031524.html> [Accessed 10 Nov. 2024].

**Increase profitability**

* Sulaiman, S., Mustapha, A. and Shareef, M. (2022). Price Prediction and Classification of Used-Vehicles Using Machine Learning Techniques. Sustainability, [online] 14(24), p.17034. Available at:<https://www.mdpi.com/2071-1050/14/24/17034> [Accessed 10 Nov. 2024].

**Usage of AI in sales**

* Information Age (2022). AI increasing transparency in used car sales. [online] Available at:<https://www.information-age.com/ai-increasing-transparency-in-used-car-sales-19811/> [Accessed 28 Nov. 2024].

**Importance of pricing in Marketing**

* Economics Discussion (n.d.). Importance of Pricing: 7 Importance of Pricing in Marketing. [online] Available at: https://www.economicsdiscussion.net/marketing-management/pricing/importance-of-pricing/31838 [Accessed 28 Nov. 2024].
* Quest Journals (2021). Comparative Analysis of Consumers’ Perception of Different Transmission Technologies. [online] Available at:<https://www.questjournals.org/jrbm/papers/vol11-issue7/1107127137.pdf> [Accessed 02 Dec. 2024].

**One-hot encoding**

* Williams, C.K.I. (2024). Naive Bayes Classifiers and One-hot Encoding of Categorical Variables. [online] Available at:<https://arxiv.org/abs/2404.18190> [Accessed 03 Dec. 2024].

**Outlier detection for skewed data**

* Hubert, M. and Van der Veeken, S. (2008). *Outlier detection for skewed data*. *Journal of Chemometrics*, 22(3-4), pp.235-246.

**Data Engineering**

* Raj, P., Varma, G. and Kumar, M. (2021). Data Engineering for Machine Learning: A Roadmap. [online] Available at:<https://arxiv.org/pdf/2102.11447> [Accessed 08 Dec. 2024].
* Milunovich, G., Wu, L., & Zhao, Y. (2023). Forecasting the Prices of Used Cars: A Comparative Analysis of Supervised Learning Algorithms. [online] Available at:<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4334750> [Accessed 08 Dec. 2024].

**Data diversity**

* Gong, B., Li, W., Grauman, K., and Sha, F., (2019). Diversity in machine learning. arXiv preprint, arXiv:1807.01477. Available at:<https://arxiv.org/pdf/1807.01477> [Accessed 08 Dec. 2024].
* Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. Sebastopol: O'Reilly Media, p.851. ISBN: 9781492032649. Available at: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ [Accessed 28 Dec. 2024].

#### Github: https://github.com/bragayuri/strategic-thinking-ca-1