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* | Friday, 29th November 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.

**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**

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.

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.

**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.

**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.

**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.

### **2. Project Plan**

This project aims to cover the following deliverables:

* A deployed model that predicts used vehicle price
* A simple user interface to interact with the model

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 | User interface, Documentation and Deployment |

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: Create a detailed documentation with the EDA, report and findings. Develop a simple user interface to interact with the model. Deploy the tool.

**Resources**

**This capstone will use Python for data cleaning and model implementation.**

**It will be delivered as a Jupyter Notebook. State libraries.**

### **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.**

**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 primary beneficiaries of the used car price prediction tool include:**

**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.**

### **3. Business Understanding**

* **Stakeholders: Identify who will benefit from this tool (e.g., individual sellers, used car dealerships, online marketplaces).**
* **Business Impact: Discuss how improving price prediction accuracy can impact decision-making and profitability in the used car market.**

### **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.

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:

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.

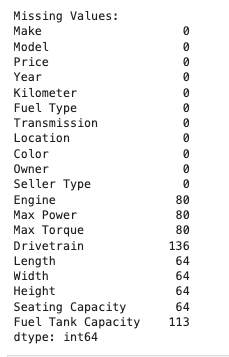
Width: The width measure of the automobile in millimetres.

Height: The height measure of the automobile in millimetres.

Seating Capacity: Number of seats which are available for passengers (e.g., 5 , 7).

Fuel Tank Capacity: Size of the fuel tank measured in liters.

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

****

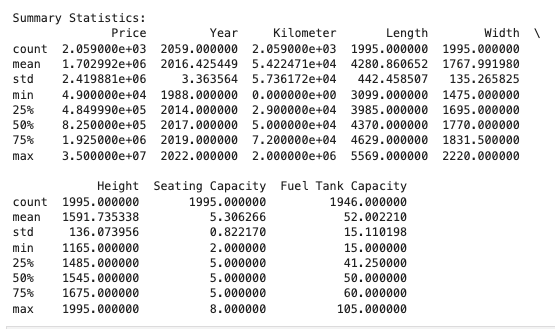
Missing values on the following attributes:

* ***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):**

The values observed for length ranges from 3099 mm to 5569 mm with the mean being 4280 mm.

The values observed for width ranges from 1475 mm to 2220 mm with the mean being 1768 mm.

The 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.

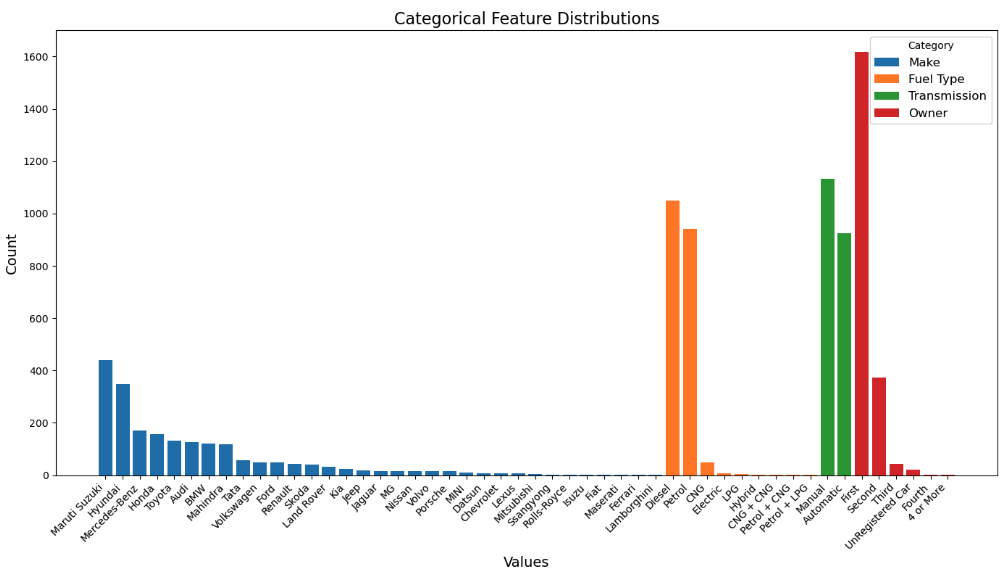
**Seating Capacity Range**: Minimum - 2 seats, maximum - 8 seats and (5) being the median as most vehicles have five seats.

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:**

Because manual and automatic transmission, the difference in market clearly illustrates the preferences for driving patterns of the consumers.

First owner cars clearly have a notable preference in the buyers as all the vehicles are seen to be on the positive side in this regard.

**Data Diversity:**

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.

**Data engineering**

Creation of Age Column

The Year column has turned out to be Age by determining the difference of the current calendar year from the car manufacturing year.

The original Year column was then deleted in order to prevent redundancy. WHY is important?

**Conversion to Numeric Values:**

The Engine and Max Power values which were earlier captured as strings with non-alphabetical characters were converted to numeric values as well.

This extracted numeric data guarantees that these features are suitable for modeling.Why is important? reference

**Handling of Missing Values:**

**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.

Categorical features must be replaced with their most common (mode) or otherwise appropriate.

**Why is it important**

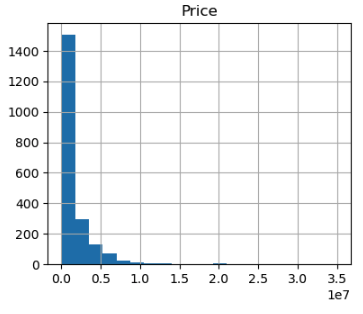
**After Handling:** Missing values in Engine and Max Power have been entirely solved, whereas the other columns such Max Torque, Drivetrain and dimensions (length, width) have been resolved but still require additional attention.

**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.

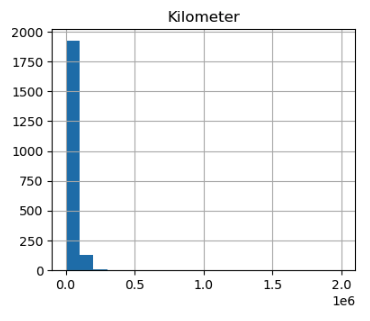
There’s a long tail because there aren’t many cars that are much more expensive than 5,000,000; only high range expensive cars exist.

**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.

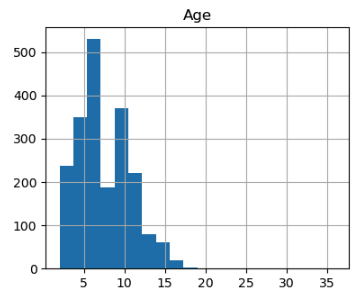
Most of the cars would have a mileage count lesser than a 1000, while a very few would have counted mileage of 1000 or more.

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.

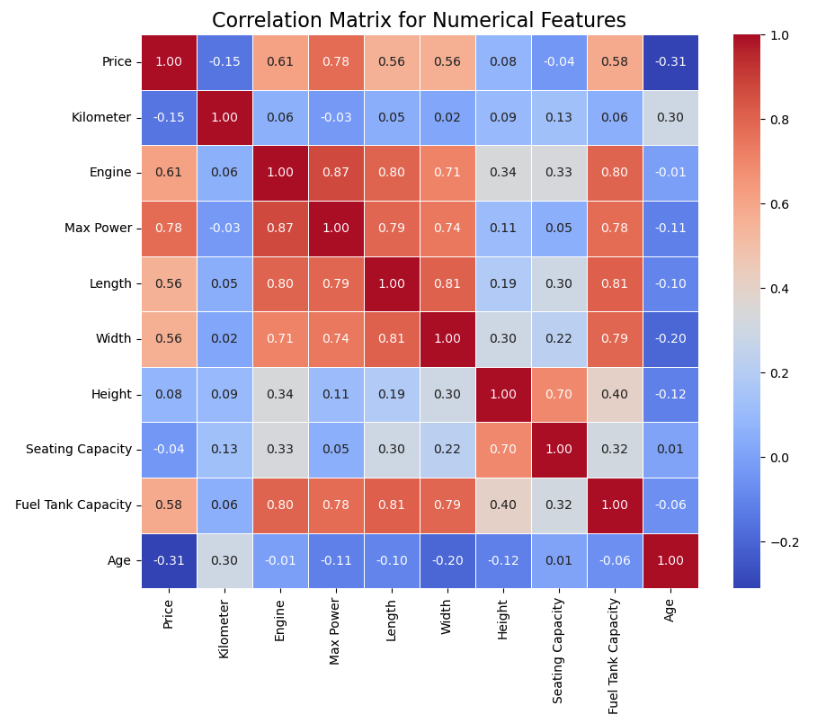
Age as a Generalization and easy to control variable:

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.Research use reference.

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. Reference

**Correlation Analysis Report**

The correlation matrix is used in order to examine the relationship between all the numerical features in the dataset. 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.

**2. 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.

**3. 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.

- **Color**: Because color 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.

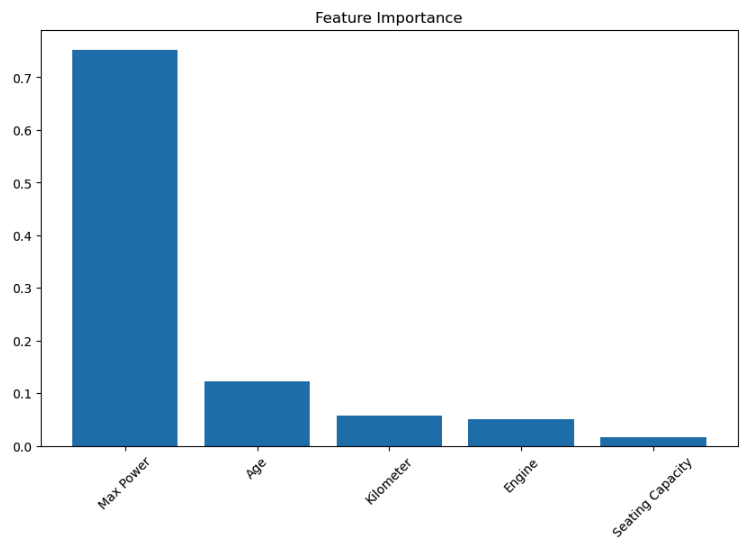
Outcome

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**

In order to enhance the homogeneity of the dataset and reduce the effect that exceptional cases may have on the quality of the model, outlier capping was done on four important variables. Achieving such control on exceptional values of a particular feature area is beneficial as it helps achieve appropriate balance in the distribution of the data of that feature across the majority of cases. (Raschka & Mirjalili, 2022)

#### **1. Cost**

* **Capping Range:** 160,000 to 10,817,640
* **Description:** Extreme luxury vehicle prices were capped to ensure there are no excessively high outliers.
* **Impact:** The approach assists in avoiding concentrating on luxury pricing by balancing the emphasis on mid-range and premium vehicles.

#### **2. Distance Traveled (in Kilometers)**

* **Capping Range:** 2,459.78 – 145,882
* **Description:** High and low values of mileages that are extreme in nature and are most likely infrequent and incorrect were applied.
* **Impact:** Utilizes realistic vehicle usage patterns, thus enhancing model performance with respect to mileage-focussed features in the context of integrated modelling.

#### **3. Engine Size**

* **Capping Range:** 796 to 3198
* **Description:** Engine capacities above a certain threshold, which are usually offered by performance or exotic automobiles, were capped.
* **Impact:** A decreased ability to focus on outlier-sized engines enhances common automobiles’ model explainability.

#### **4. Engine Power**

* **Capping Range:** 47 to 335
* **Description:** Capped extreme outliers with extreme wlper powers in the segment of high-performance vehicles.
* **Impact:** Helps maintain equilibrium between performance specific and standard cars thereby biasing the predicted prices even less.

### **The Need for Outlier Capping**

Outlier capping increases the usefulness of the machine learning datasets through the following ways.

1. **Creating More Robust Models:** Models are less sensitive to noise and extreme values which makes them perform well on general data.
2. **Reducing Skewness:** This helps to achieve a reasonable degree of normalcy in the distributions of data, which is fundamental for most machine learning algorithms.
3. **Retention of Variability:** Going to the extreme of removing, capping nevertheless concerns more data points thereby allowing the calibration of the entire data set.

This step ensures that the data set conforms to established standards for feature engineering which will further guarantee the validity and interpretability of the model to follow.

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.

1. Mileage Intensity

Definitions: The ratio of the number of kilometers driven to the age of a vehicle (Kilometer / (Age + 1)).

Key Statistics:

Range: 409.96 - 20840.28.

Median: 5750.

Mean: 6113.48.

Description:

This feature relates to the average kilometers driven per year and hence indicates the intensity of the usage of the vehicle over its service life.

Being in the high classes implies that its vehicle has indeed been driven much more than other units of the same age.

Impact. This feature helps in the quantification of the depreciation in the condition of the vehicles allowing the model to better accommodate depreciation due to the age of the vehicle having been used.

1. CC Price

Definition: It is the price of the vehicle divided by its engine capacity (Price / Engine).

Key Statistics:

Range: 64.15 to 5,441.47.

Median: 564.30.

Mean: 856.15.

Description:

Cost attributed per cc of the overall available engine capacity

High values mean the vehicles with a small engine price probably more than its value and are either luxurious or better efficient vehicles.

Impact: Aided in positioning the vehicles in terms of a targeted pricing structure engine size, which helps in the understanding of performance pricing considerably better.

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.

1. Log Transformation

Applied To: Price, Kilometer, Engine, Max Power.

Purpose:

Handling highly skewed data distributions especially log transformed variables which are normally distributed.

It makes sure that extreme values do not overpower the model learning (Han et al., 2022).

Outcome:

Created new columns such as Pricel, Kilometrel, etc. for the log-transformed values.

Skewed features are now better distributed for machine learning algorithms that are sensitive to normality and the presence of outliers.

1. Min-Max Scaling

Applied To: All numerical columns.

Purpose:

Normalizes the distribution of features to a standard range of zero to one while maintaining the proportions of the values.

Ensures that features with a larger range such as Price do not have the most impact on the algorithms such as SVM or k-NN.

Outcome:

Make sure all numerical features are standardized in the dataset for algorithms that are sensitive to scale difference.

Importance

Log Transformation Benefits:

Helps in lessening the adverse effects of extreme values.

Helps in interpreting Price, Engine features much more effectively (Han et al., 2022).

Min-Max Scaling Benefits:

Proves to be important for neural networks and distance-based models such as k-NN in which scale affects the learning very much.

Enhances the improvement in convergence speed for gradient based optimizers.

### **Reference**

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.

* **Initial Observations: Share any initial insights from a cursory look at the data.**

### **5. Data Preparation**

* **Cleaning Steps: Explain how you handled missing data, outliers, and errors.**
* **Feature Engineering: Describe any new features you created and why.**
* **Data Transformation: Document any transformations or scaling applied to the dataset.**

### **6. Exploratory Data Analysis**

* **Descriptive Statistics: Provide statistics that summarize the central tendency, dispersion, and shape of the dataset's distribution.**
* **Visualizations: Include charts and graphs to illustrate relationships between features and the target variable.**
* **Preliminary Findings: Highlight any significant trends or patterns you observed.**

### **7. Findings & Recommendations**

* **Model Performance: Discuss how the models performed, including metrics like RMSE or R².**
* **Insights: Share insights on what drives car prices based on your model’s findings.**
* **Recommendations: Offer suggestions for potential buyers, sellers, or platforms based on your findings.**
* **Future Work: Propose further areas of research or additional features to improve the model.**

**References**

Jain, R., Gupta, P., & Khare, N. (2021). Factors Affecting Resale price of Cars.

Nanda, P., Yadav, S., & Sharma, V. (2020). Trends in Depreciation of Valuation models of automotives markets.

### **8. Version Control and Repository**

* **GitHub Usage: Explain how you used GitHub for version control.**
* **Repository Link: Include a link to the GitHub repository containing your Jupyter Notebook and report.**

### **9. References**

* **Citation: Ensure all your sources are properly referenced using the Harvard referencing style. Include any datasets, articles, or books you utilized.**

### **Word Count and Submission**

* **Adherence to Guidelines: Keep the report within the 2,000-word limit, focusing on clarity and conciseness.**
* **Professional Presentation: Ensure the report is well-organized, properly formatted, and free of grammatical errors.**

### **Tools and Tips for Execution**

* **Jupyter Notebook: Use this for all your coding, analysis, and visualization tasks.**
* **Libraries: Utilize Pandas for data manipulation, Matplotlib/Seaborn for data visualization, and scikit-learn for modeling.**
* **Iterative Approach: Continuously refine your analysis based on the findings and feedback.**

**This structured approach will help you create a comprehensive and coherent report that meets the requirements of your assignment while providing valuable insights into the used car market.**

**Assessment Task:**

This assessment aims to evaluate your ability to apply project management methodology to develop and execute a capstone project. You will select a dataset, conduct exploratory data analysis, pre-process the data, implement at least one machine learning algorithm, and present your findings effectively through a comprehensive report. The capstone project will be based on a dataset of your choice from any domain, such as finance, marketing, or any other.

You will

submit a comprehensive report detailing the following;

* Strategic overview of the business problem Project plan
* Business understanding
* Data understanding
* Data preparation
* Machine learning implementation Findings
* Conclusions
* Any future recommendations

The report should be presented in a clear and concise manner, and it should demonstrate your ability to use a project management methodology. The project management methodology should enable you to prioritize tasks and monitor the progress of the capstone project. Additionally, the report should provide a background to the business problem and its importance from a strategic viewpoint, an overview of the project's timeline, milestones achieved, and any challenges faced during the implementation phase. It should highlight the key insights gained from analyzing the data and present any significant trends or patterns observed. Moreover, the report should address any limitations or constraints encountered during the project and propose potential solutions for future improvements.

This is a two-semester module, and the capstone project will continue into semester two. Students are advised to review and adhere to the submission requirements documented after the assessment task.

Further details of the assessment:

* a) **Continue to use the GitHub repo provided in CA 1, the Jupyter Notebook and report**    
  **Word document must be put into a GitHub repo for version control. The GitHub repo link will be added at the end of the report. There should be another 5 to 10 commits throughout the time worked on CA2.**
* **b) Exploratory data analysis of your dataset. Use descriptive statistics.**
* c) **Use at least one machine learning algorithm.**
* d) **Support your analysis with references and properly reference ALL sources that you**    
  **have used. WARNING – If you do not support your work, you will not receive a high mark!**
* e) WORD COUNT: 2,000 words. If your report is too short or long, you may *lose up to 10% of marks*!

**Assessment Requirements**

All assessment submissions must meet the following minimum requirements:

* ● Include a professional report paper in Word format ONLY of about 2,000 words.
* ● Code must be submitted as a Jupyter Notebook artefact.
* ● ZIP or RAR files will not be accepted. Files must be submitted separately.
* ● Be submitted by the deadline date specified or be subject to late submission penalties.
* ● Be submitted via Moodle upload.
* ● Use Harvard Referencing when citing third party material.
* ● Be the student’s own work.
* ● Include the CCT assessment cover page.  
     
  **Learning Outcomes:**    
  This assessment addresses the following module learning outcomes for this module:  
   1. Critically evaluate the relationship between information technology infrastructure and organisational competitive advantage.  
   2. Critically analyse and select open source and proprietary software with a view to developing IT  
   solutions for business and business-related IT problems.  
   3. Utilise tools of strategic business analysis to evaluate the current macro and micro business environment with a view to formulating future action plans.  
   4. Research emerging technologies and critically evaluate their impact on business and business information systems in general.  
   5. Understand the relationship between data gathering/utilisation and business intelligence and its impact on industry policy.

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

* **~~Context~~**~~: Start by describing the used car market and why accurate price prediction is critical.~~
* **~~Problem Statement~~**~~: Define the specific problem your project addresses—predicting the resale value of used cars.~~
* **~~Importance~~**~~: Explain why this is important for buyers, sellers, and businesses within the automotive industry.~~

### **~~2. Project Plan~~**

* **~~Goals~~**~~: List the primary goals of your project.~~
* **~~Timeline~~**~~: Outline the timeline of your project with key milestones and deadlines.~~
* **~~Resources~~**~~: Describe the tools and technologies (e.g., Python, Jupyter Notebook, specific libraries) you will use.~~

### **3. Business Understanding**

* **Stakeholders**: Identify who will benefit from this tool (e.g., individual sellers, used car dealerships, online marketplaces).
* **Business Impact**: Discuss how improving price prediction accuracy can impact decision-making and profitability in the used car market.

### **4. Data Understanding**

* **Data Source**: Describe where your data comes from and why it is suitable for this project.
* **Features Description**: Provide a detailed description of the dataset features (e.g., make, model, mileage, year, etc.).
* **Initial Observations**: Share any initial insights from a cursory look at the data.

### **5. Data Preparation**

* **Cleaning Steps**: Explain how you handled missing data, outliers, and errors.
* **Feature Engineering**: Describe any new features you created and why.
* **Data Transformation**: Document any transformations or scaling applied to the dataset.

### **6. Exploratory Data Analysis**

* **Descriptive Statistics**: Provide statistics that summarize the central tendency, dispersion, and shape of the dataset's distribution.
* **Visualizations**: Include charts and graphs to illustrate relationships between features and the target variable.
* **Preliminary Findings**: Highlight any significant trends or patterns you observed.

### **7. Findings & Recommendations**

* **Model Performance**: Discuss how the models performed, including metrics like RMSE or R².
* **Insights**: Share insights on what drives car prices based on your model’s findings.
* **Recommendations**: Offer suggestions for potential buyers, sellers, or platforms based on your findings.
* **Future Work**: Propose further areas of research or additional features to improve the model.

### **8. Version Control and Repository**

* **GitHub Usage**: Explain how you used GitHub for version control.
* **Repository Link**: Include a link to the GitHub repository containing your Jupyter Notebook and report.

### **9. References**

* **Citation**: Ensure all your sources are properly referenced using the Harvard referencing style. Include any datasets, articles, or books you utilized.

### **Word Count and Submission**

* **Adherence to Guidelines**: Keep the report within the 2,000-word limit, focusing on clarity and conciseness.
* **Professional Presentation**: Ensure the report is well-organized, properly formatted, and free of grammatical errors.

### **Tools and Tips for Execution**

* **Jupyter Notebook**: Use this for all your coding, analysis, and visualization tasks.
* **Libraries**: Utilize Pandas for data manipulation, Matplotlib/Seaborn for data visualization, and scikit-learn for modeling.
* **Iterative Approach**: Continuously refine your analysis based on the findings and feedback.