MIDDLESEX UNIVERSITY

THE BURROUGHS, HENDON LONDON NW4 4BT

Title: Coursework 1 – Data Science Practices

Student Name/Surname: Syed Nihaal Ahmed

Student ID: M01039337

Word Count: 2225

Contents

1. Understanding of the Data

Dataset Summary and Availability –

This analysis utilises the Cars Datasets 2025 compiled by Abdul Malik, publicly available on Kaggle at: https://www.kaggle.com/datasets/abdulmalik1518/cars-datasets-2025/data

The dataset is published under Kaggle's public domain licence, making it freely accessible for educational and research purposes without requiring additional permissions. The dataset contains information on over 1,000 vehicles from 2025, encompassing a range of manufacturers, from luxury brands to mainstream manufacturers.

Research Goal –

To develop and evaluate machine learning regression models that predict car prices and performance-based characteristics, specifically analysing how technical specifications (horsepower, engine capacity, speed, acceleration) correlate with market pricing.

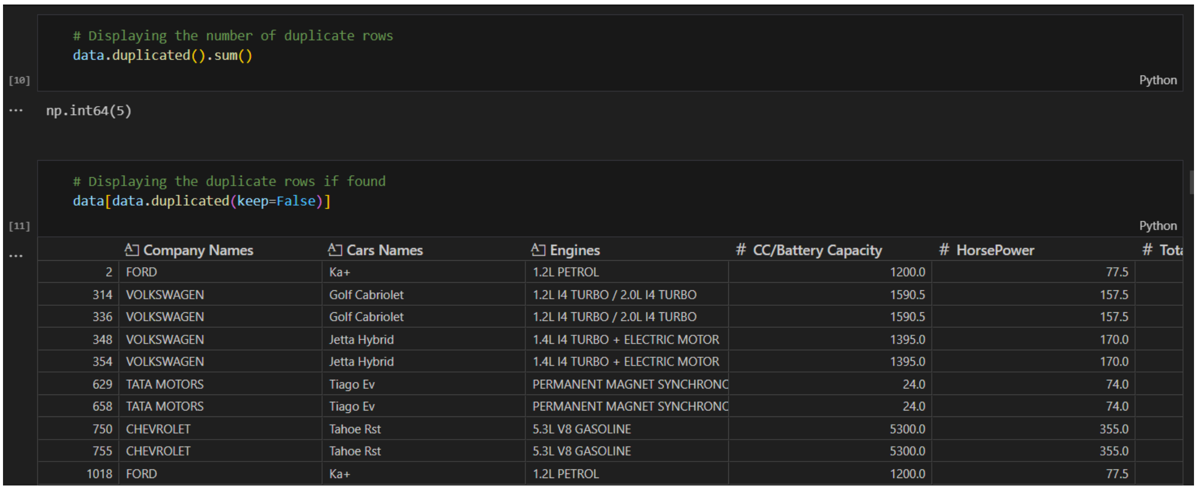
Key Dataset Characteristics –

1. Dimensions: 1,063 rows × 14 columns (after encoding)
2. Target Variable: Cars Prices (USD)
3. Numerical Features: CC/Battery Capacity, HorsePower, Total Speed, Performance (0-100 km/h), Torque, and Seats.
4. Categorical Features: Company Names, Cars Names, Fuel Types, and Engines.
5. Datatype object: Company Names, Cars Names, Engines, and Fuel Types.
6. Datatype float64: CC/Battery Capacity, HorsePower, Total Speed and Performance, Cars Prices, Torque.
7. Datatype Int64: Seats, Company Names_encoded, Fuel Types_encoded and Engines_encoded.

2. Data Pre-Processing

Handling duplicate values –

Sum of duplicate rows found: 5
Sum of duplicate rows after dropping duplicate rows: 0



```
# Displaying the number of duplicate rows
data.duplicated().sum()
```

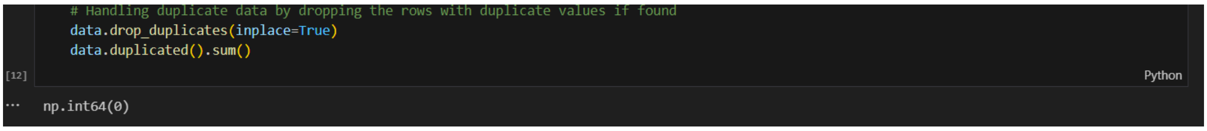
[10] Python

```
np.int64(5)
```

```
# Displaying the duplicate rows if found
data[data.duplicated(keep=False)]
```

[11] Python

	Company Names	Cars Names	Engines	# CC/Battery Capacity	# HorsePower	# Tota
2	FORD	Ka+	1.2L PETROL	1200.0	77.5	
314	VOLKSWAGEN	Golf Cabriolet	1.2L I4 TURBO / 2.0L I4 TURBO	1590.5	157.5	
336	VOLKSWAGEN	Golf Cabriolet	1.2L I4 TURBO / 2.0L I4 TURBO	1590.5	157.5	
348	VOLKSWAGEN	Jetta Hybrid	1.4L I4 TURBO + ELECTRIC MOTOR	1395.0	170.0	
354	VOLKSWAGEN	Jetta Hybrid	1.4L I4 TURBO + ELECTRIC MOTOR	1395.0	170.0	
629	TATA MOTORS	Tiago Ev	PERMANENT MAGNET SYNCHRONC	24.0	74.0	
658	TATA MOTORS	Tiago Ev	PERMANENT MAGNET SYNCHRONC	24.0	74.0	
750	CHEVROLET	Tahoe Rst	5.3L V8 GASOLINE	5300.0	355.0	
755	CHEVROLET	Tahoe Rst	5.3L V8 GASOLINE	5300.0	355.0	
1018	FORD	Ka+	1.2L PETROL	1200.0	77.5	



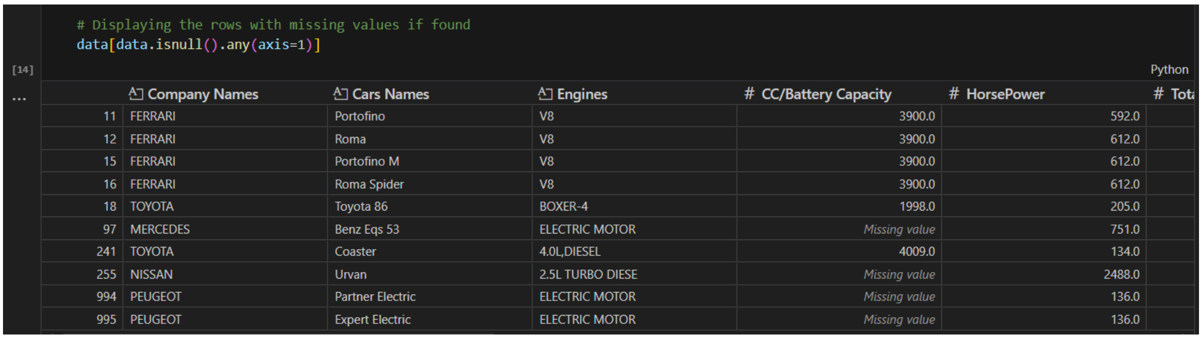
```
# Handling duplicate data by dropping the rows with duplicate values if found
data.drop_duplicates(inplace=True)
data.duplicated().sum()
```

[12] Python

```
np.int64(0)
```

Handling missing values –

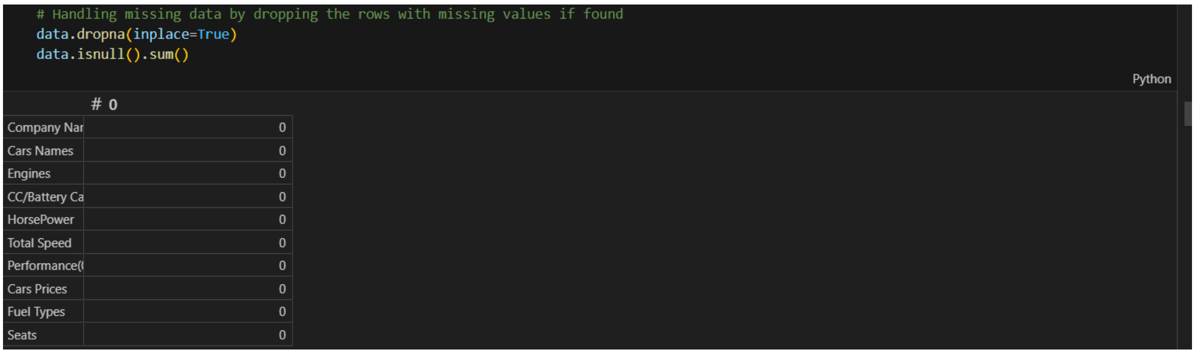
Sum of missing values found: 23
Sum of missing values after dropping rows: 0



```
# Displaying the rows with missing values if found
data[data.isnull().any(axis=1)]
```

[14] Python

	Company Names	Cars Names	Engines	# CC/Battery Capacity	# HorsePower	# Tota
11	FERRARI	Portofino	V8	3900.0	592.0	
12	FERRARI	Roma	V8	3900.0	612.0	
15	FERRARI	Portofino M	V8	3900.0	612.0	
16	FERRARI	Roma Spider	V8	3900.0	612.0	
18	TOYOTA	Toyota 86	BOXER-4	1998.0	205.0	
97	MERCEDES	Benz Eqs 53	ELECTRIC MOTOR	*Missing value*	751.0	
241	TOYOTA	Coaster	4.0L,DIESEL	4009.0	134.0	
255	NISSAN	Urvan	2.5L TURBO DIESE	*Missing value*	2488.0	
994	PEUGEOT	Partner Electric	ELECTRIC MOTOR	*Missing value*	136.0	
995	PEUGEOT	Expert Electric	ELECTRIC MOTOR	*Missing value*	136.0	



```
# Handling missing data by dropping the rows with missing values if found
data.dropna(inplace=True)
data.isnull().sum()
```

Python

	# 0
Company Nar	0
Cars Names	0
Engines	0
CC/Battery Ca	0
HorsePower	0
Total Speed	0
Performance(	0
Cars Prices	0
Fuel Types	0
Seats	0

Standardisation and Normalisation –

All numerical features were transformed to the [0, 1] range using sklearn’s MinMaxScaler, preserving relative relationships and distribution characteristics. The scaling process-maintained data integrity by retaining original distribution patterns, including skewness and outliers, in relative terms.

Features Scaled: CC/Battery Capacity, HorsePower, Total Speed, Performance (0-100 km/h), Torque, and Cars Prices.

Formula used: $x_{scaled} = \frac{(x - x\_min)}{(x\_\max - x\_min)}$

Categorical Variable Encoding –

Categorical variables were encoded using sklearn's LabelEncoder to convert text categories into numerical representations suitable for machine learning algorithms.

Label encoding was selected over one-hot encoding to reduce dimensionality while maintaining ordinal relationships implicit in automotive market positioning.

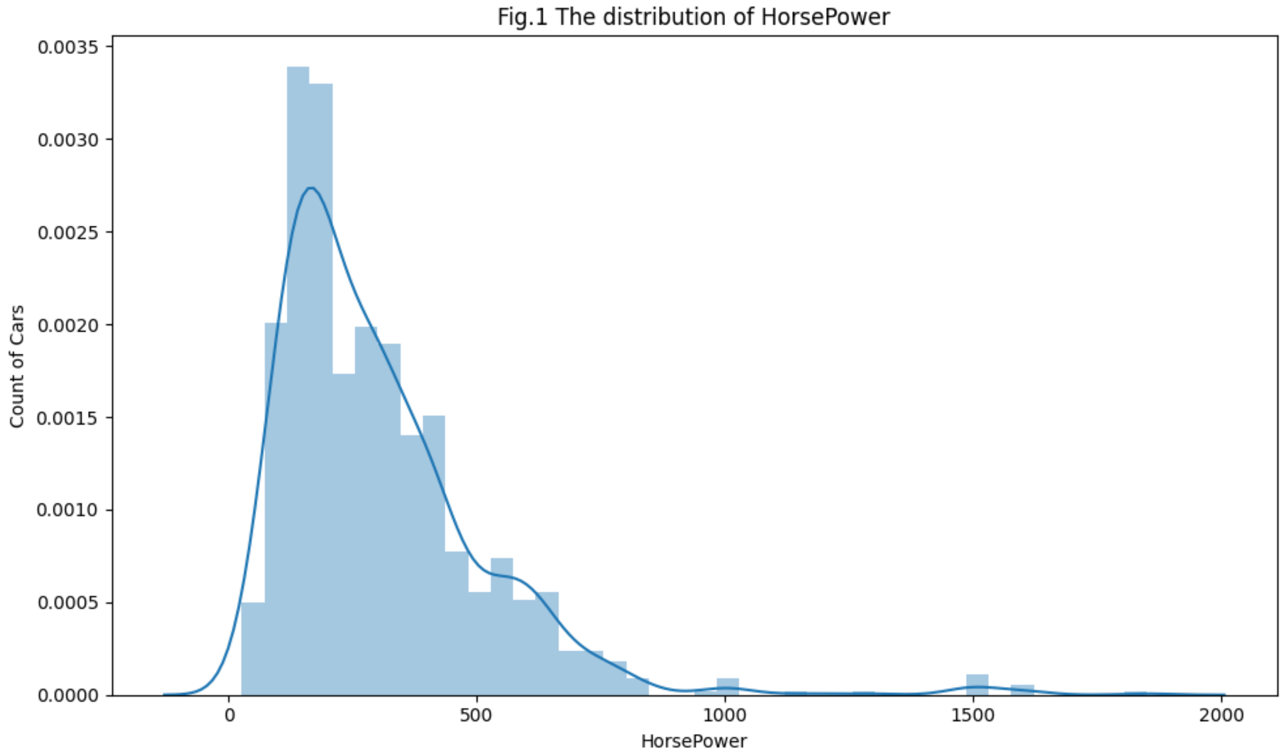
Encoded Variables:

1. Company Names to Company Names_Encoded (50 unique values)
2. Fuel Types to Fuel Types_Encoded (8 unique categories)
3. Engines to Engines_Encoded (26 unique categories)

3. Exploratory data analysis (EDA)

Target Variable and Feature Correlation Analysis

HorsePower Distribution –

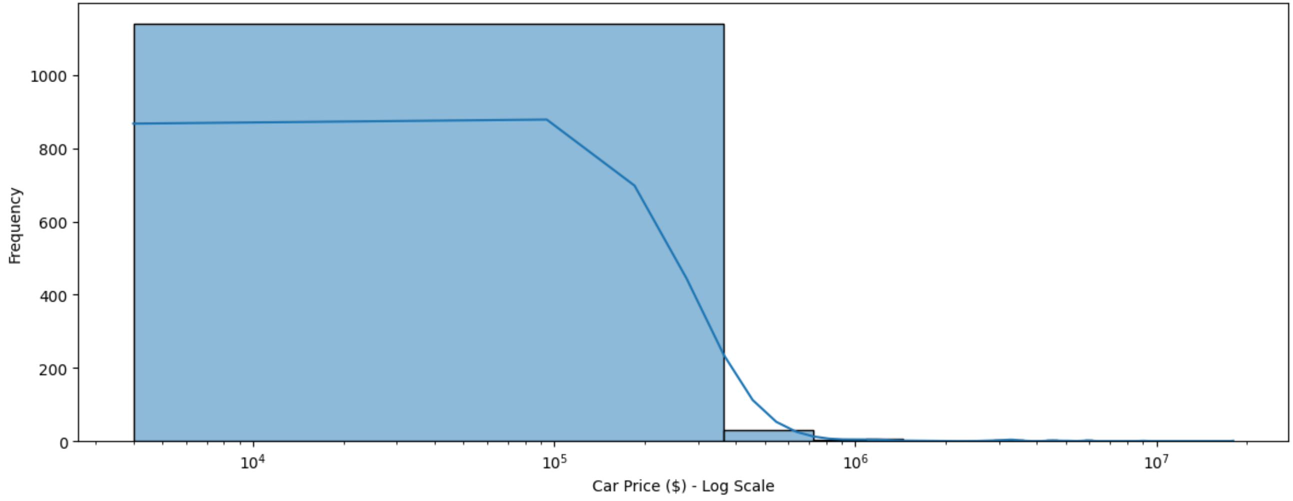


In the above Fig.1 the distribution of HorsePower, the distribution of HorsePower, exhibits a pronounced right skewness (2.38) and heavy tails (kurtosis: 9.87). The majority of vehicles fall within the 100 to 300 hp range, typical for standard consumer cars, while high-performance vehicles exceeding 500 hp are rare outliers. This pattern mirrors real-world market dynamics, where mainstream models are prevalent and exotic supercars are uncommon. Consequently, the skewness indicates that predictive models are likely to achieve greater accuracy for average vehicles than for extremely high-performance outliers.

Distribution of Cars Prices –

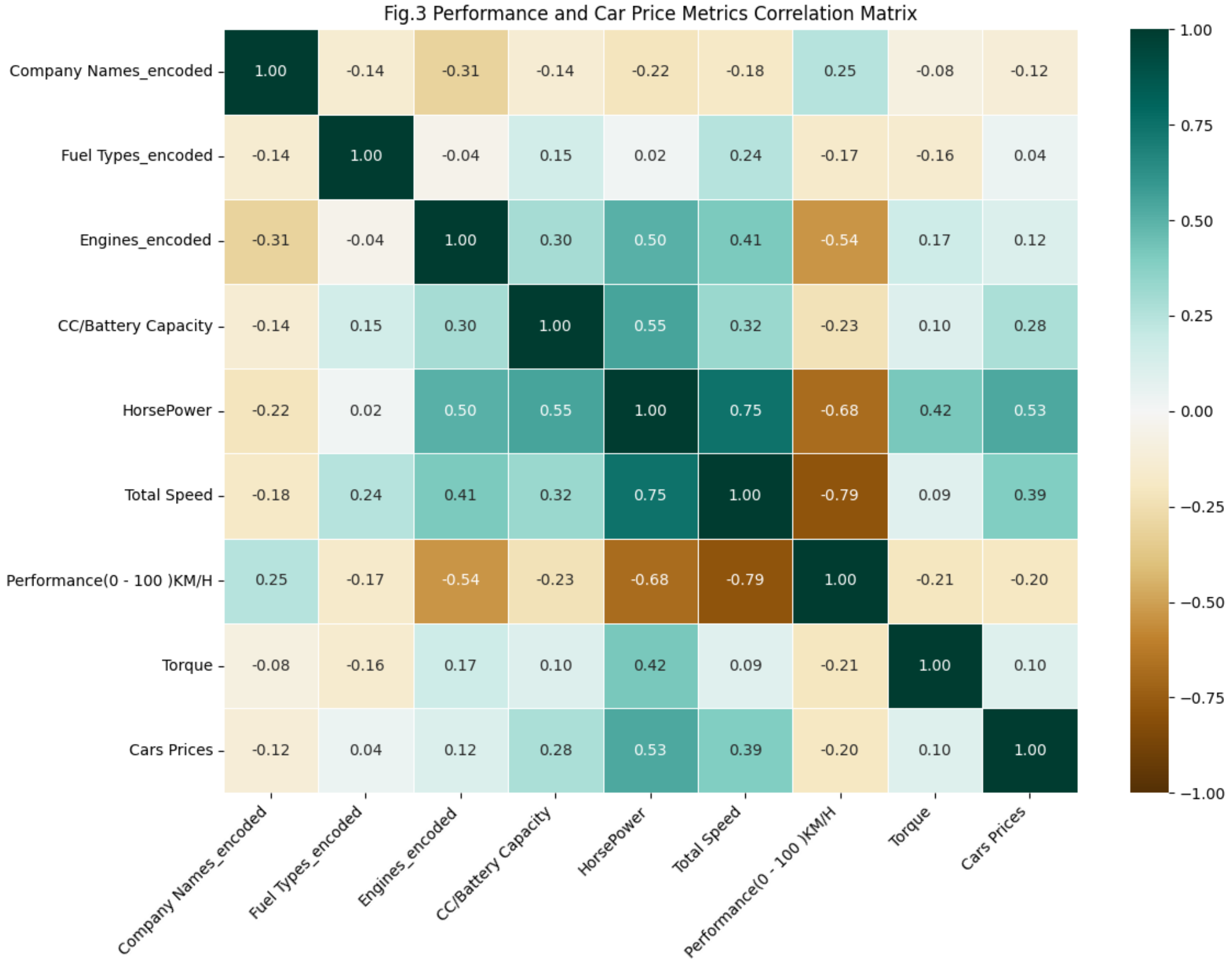
In the below Fig.2 Distribution of Car Prices, the distribution of car prices displays pronounced right skewness (16.49) and extremely heavy tails (kurtosis: 351.85). Approximately 60% of vehicles are clustered within the $20,000 to $60,000 price range, reflecting the mainstream automotive market. However, a small number of ultra-luxury vehicles priced up to $18 million significantly influence the distribution, resulting in substantial statistical skew. As a result, the median price ($40,000) serves as a more accurate measure of central tendency than the mean ($85,188), which is heavily distorted by these outliers.

Fig.2 Distribution of Car Prices (Log Scale)



Performance (0 - 100) KM/H and Car Price Correlation Matrix –

Fig.3 Performance and Car Price Metrics Correlation Matrix



	Company Names_encoded	Fuel Types_encoded	Engines_encoded	CC/Battery Capacity	HorsePower	Total Speed	Performance(0 - 100)KM/H	Torque	Cars Prices
Company Names_encoded	1.00	-0.14	-0.31	-0.14	-0.22	-0.18	0.25	-0.08	-0.12
Fuel Types_encoded	-0.14	1.00	-0.04	0.15	0.02	0.24	-0.17	-0.16	0.04
Engines_encoded	-0.31	-0.04	1.00	0.30	0.50	0.41	-0.54	0.17	0.12
CC/Battery Capacity	-0.14	0.15	0.30	1.00	0.55	0.32	-0.23	0.10	0.28
HorsePower	-0.22	0.02	0.50	0.55	1.00	0.75	-0.68	0.42	0.53
Total Speed	-0.18	0.24	0.41	0.32	0.75	1.00	-0.79	0.09	0.39
Performance(0 - 100)KM/H	0.25	-0.17	-0.54	-0.23	-0.68	-0.79	1.00	-0.21	-0.20
Torque	-0.08	-0.16	0.17	0.10	0.42	0.09	-0.21	1.00	0.10
Cars Prices	-0.12	0.04	0.12	0.28	0.53	0.39	-0.20	0.10	1.00

In the above Fig.3 Performance and Car Price Metrics Correlation Matrix, the correlation analysis reveals several critical relationships:

Strong Correlations:

- HorsePower and Torque (0.90): A strong positive correlation indicates potential multicollinearity issues. These features measure related aspects of engine performance and may introduce redundancy in modelling.
- HorsePower and Performance (0 - 100) KM/H (-0.68): A strong negative correlation confirms that higher horsepower produces faster acceleration (lower 0-100 km/h times).

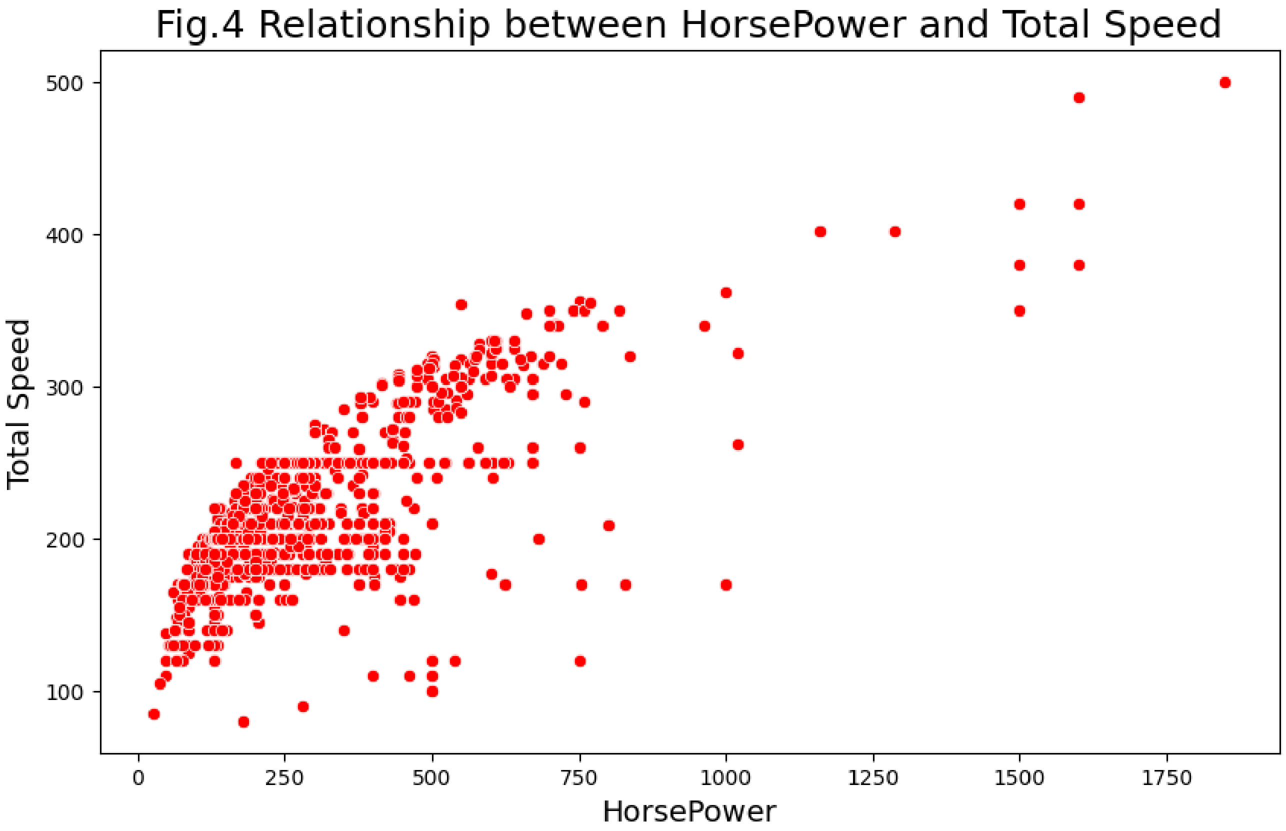
Moderate Correlations:

- HorsePower and Total Speed (0.50): A moderate positive relationship indicates horsepower contributes to top speed, though other factors (aerodynamics, gearing) also play significant roles.
- HorsePower and Cars Prices (0.45): A moderate positive correlation suggests horsepower is an important but not exclusive price determinant.

Weak Correlations:

- Total Speed and Performance (0 - 100) KM/H (0.20): A weak relationship indicates that top speed and acceleration measure different performance aspects, both valuable for price prediction.

Scatterplot Analysis –



In the above Fig.4 Relationship between HorsePower and Total Speed, the scatterplot displays a moderate-to-strong positive relationship between HorsePower and Total Speed.

As horsepower increases, top speed generally increases with a clear upward trend. However, considerable scatter exists, particularly in the high-performance region (400+ hp), where factors beyond raw power such as aerodynamics and weight significantly influence maximum velocity. This heteroscedasticity suggests non-linear modelling approaches may capture these relationships more effectively than simple linear regression.

Strongly Correlated Features Identified:

- HorsePower and Torque (0.90)
- HorsePower and Performance (0 - 100) KM/H (−0.68)
- HorsePower and Total Speed (0.50)
- HorsePower and Cars Prices (0.45)

Histplot Analysis –

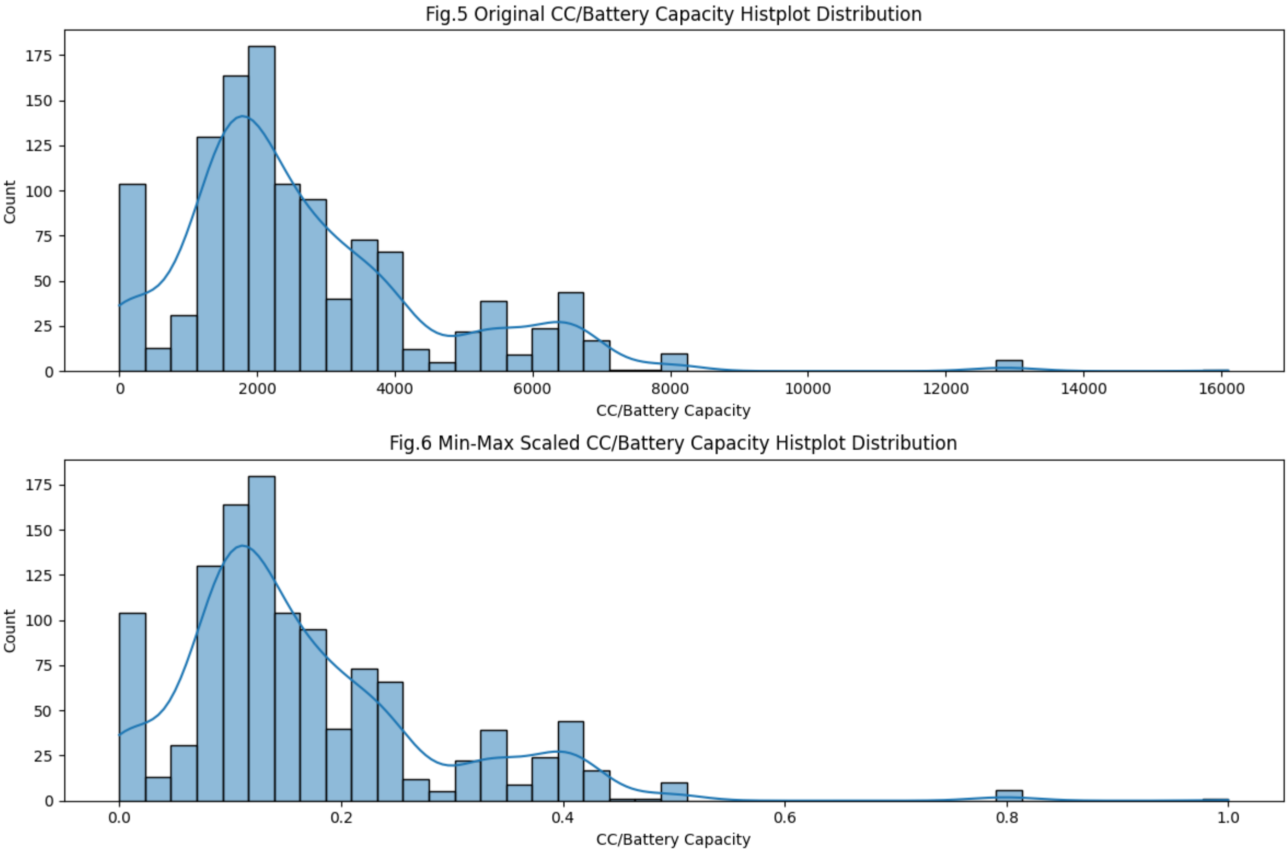


Fig.5 Original CC/Battery Capacity Histplot Distribution

Fig.6 Min-Max Scaled CC/Battery Capacity Histplot Distribution

The above Fig.5 Original CC/Battery Capacity and Fig.6 MinMaxScaler CC/Battery Capacity Histplot distributions provide a valuable comparison of numerical feature distributions before and after normalisation. Initially, the CC/Battery Capacity data displays pronounced right-skewness, ranging from approximately 360 cc to 95,000 cc. The majority of observations are concentrated between 1,000 cc and 4,000 cc, representing typical consumer vehicles. In contrast, a limited number of outliers, such as supercars and hypercars, extend far into the upper range. This skewness indicates that while most cars fall into the mid-range performance category, a small portion of high-performance vehicles significantly influences the upper tail of the distribution. Following Min-Max scaling, the histplot retains the same structural shape, merely transformed into a [0,1] numerical range. This confirms that normalisation is a linear process that preserves the data's statistical characteristics while ensuring all features contribute equally to model training.

Boxplot analysis –

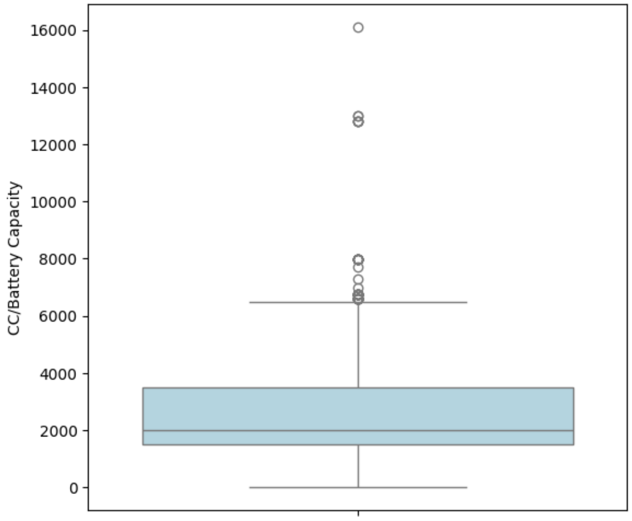


Fig.7 Original CC/Battery Capacity Boxplot Distribution

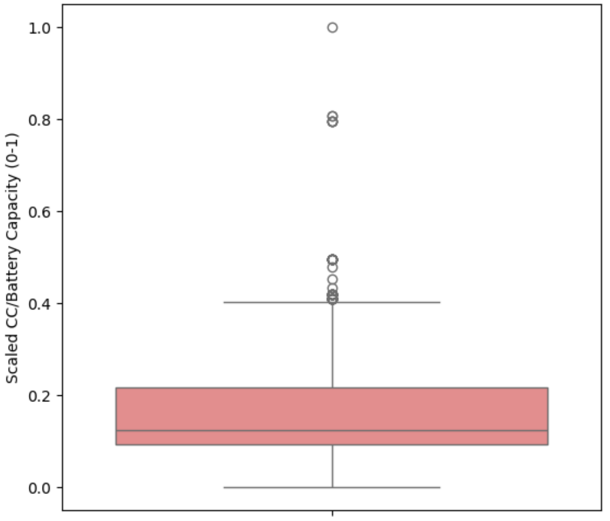
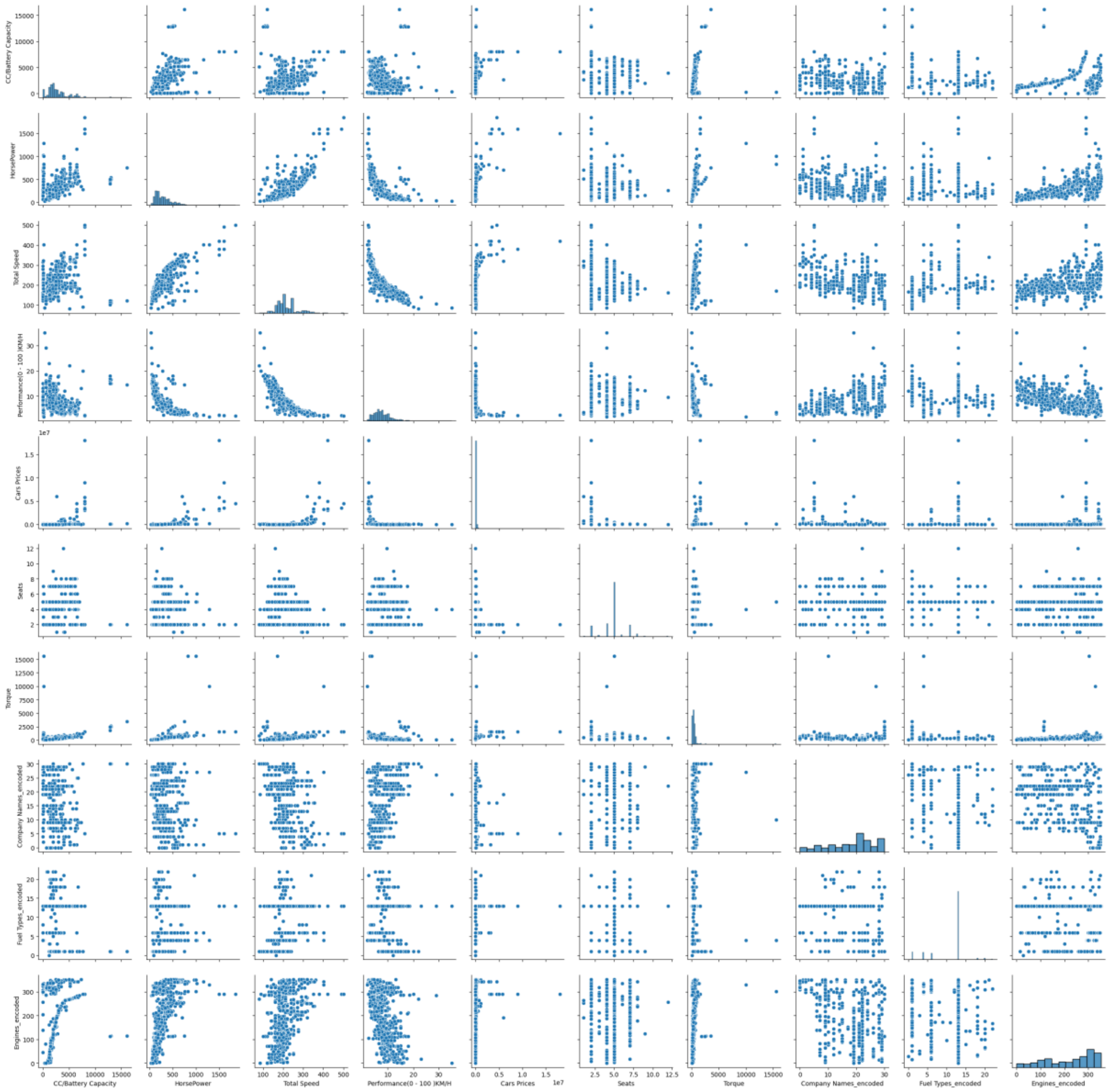


Fig.8 Min-Max Scaled CC/Battery Capacity Boxplot Distribution

The above Fig.7 Original CC/Battery Capacity and Fig.8 Min-Max Scaled CC/Battery Capacity Boxplot Distribution were employed to visualise data distribution, central tendency, and variability of the CC/Battery Capacity feature both before and after scaling. The original distribution reveals a median of approximately 3,200 cc, with the interquartile range (IQR) extending from 2,000 cc to 4,500 cc. Notably, several extreme outliers are observed beyond 8,000 cc, corresponding to sports cars and high-output electric vehicles. After applying Min-Max scaling, the visual structure of the boxplot remains identical, confirming that scaling has not altered the underlying distribution shape or variance. Instead, it has rescaled values proportionally into the [0,1] range. This linear transformation ensures that no feature dominates another numerically, enabling models to interpret relationships consistently. The preservation of right-skewness post-scaling demonstrates that outliers remain present as an important consideration for modelling. Rather than removing these extreme values, the approach retains them for the Random Forest algorithm, which is robust to outliers and benefits from the full data range. The boxplot analysis, therefore, provides visual evidence that scaling effectively balances the dataset without diminishing the diversity inherent to real-world vehicle specifications.

Pairplot analysis –

Fig.8 Pairwise Relationships of Numerical Variables



The above Fig.8 Pairwise Relationships of Numerical Variables offers a multidimensional exploration of the relationships among numerical variables. It reveals several key relationships that directly influence car pricing. First, HorsePower and Torque display a nearly perfect positive linear relationship, reflecting the physical connection between these two attributes. Engines with higher horsepower naturally produce greater torque output, though this correlation introduces potential multicollinearity concerns for regression-based models. Second, HorsePower and Performance (0 - 100 km/h) exhibit a strong inverse correlation. As horsepower increases, acceleration times decrease significantly, which is expected since higher power output enables quicker velocity gain. Third, HorsePower and Total Speed demonstrate a moderate positive correlation. Although greater horsepower tends to increase maximum speed, the relationship is less direct due to confounding factors such as drivetrain efficiency, weight, and aerodynamic drag. Finally, Cars Prices show strong non-linear relationships with performance metrics, particularly horsepower and torque. This observation reinforces the conclusion that linear modelling techniques may not sufficiently capture the exponential growth of car prices at high performance levels. The diagonal histograms within the pairplot further reveal right-skewed distributions across all variables, confirming that standard vehicles dominate the dataset while ultra-luxury models remain statistical outliers. From an analytical standpoint, this justifies the adoption of non-linear ensemble methods such as Random Forest, which effectively handle heteroscedasticity and variable interactions that linear models cannot represent accurately.

Interpretation and Insights –

1. Extreme Price Skewness: The target variable's extreme skewness (16.49) violates normality assumptions of linear regression.
2. Multicollinearity Risk: The very high correlation (0.90) between HorsePower and Torque indicates redundancy that could destabilise regression coefficients.
3. Non-Linear Relationships: Scatter in the HorsePower-Speed relationship and the presence of distinct vehicle segments (economy, luxury, hypercar) suggest Random Forest may outperform linear models.
4. Feature Importance Indicators: Correlation analysis identifies HorsePower as the most influential performance metric for price prediction, followed by Torque and Total Speed. Company Names likely play a crucial role not fully captured by numerical correlations.
5. Outlier Considerations: Extreme outliers (ultra-luxury vehicles) will disproportionately influence linear regression. Tree-based ensemble methods naturally handle outliers through recursive partitioning, making them more suitable for this dataset's characteristics.
6. The EDA justifies the pre-processing choices:
    a. Min-Max scaling is appropriate for given non-normal distributions.
    b. Label encoding of Company Names captures brand-value relationships.
    c. Retention of multiple performance metrics provides complementary information despite some correlation.

4. Model development and evaluation

Traditional Machine Learning Model Implementation –

Two traditional machine learning algorithms were implemented for comparative analysis:

1. Linear Regression: Selected as the baseline model due to its interpretability and widespread use in pricing applications. It provides interpretable coefficients, Computational efficiency, and a Baseline performance benchmark.
2. Random Forest Regression: Chosen as a non-linear ensemble method capable of capturing complex interactions and handling outliers. It provides a Non-linear relationship capture. automatic feature interaction detection, robustness to outliers, minimal hyperparameter tuning requirements.

Feature set: CC/Battery Capacity, HorsePower, Total Speed, Performance (0 - 100) KM/H, Torque and Company Names_encoded

Target: Cars Prices

Train-Test Split: 80% training, 20% testing

Random State: 324

Model Performance Evaluation and Analysis –

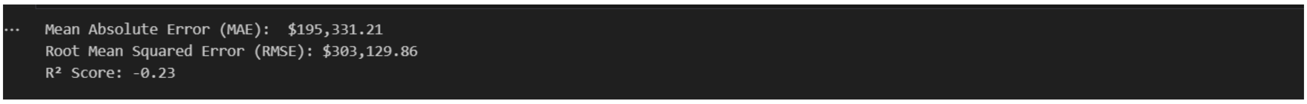
Evaluation Metrics:

1. Mean Absolute Error (MAE): Average absolute prediction error in dollars
2. Root Mean Squared Error (RMSE): Penalises large errors more heavily than MAE
3. R² Score: Proportion of variance explained (1.0 = perfect, 0.0 = baseline, <0 = worse than the mean)

Linear Regression Analysis:

The Linear Regression model demonstrates failure with:

1. MAE of $195,331: Predictions deviate by $195,000 on average, 2.3 times the typical car price ($85,000 mean)
2. RMSE of $303,130: Large errors are severely penalised, indicating widespread prediction failures
3. R² of -0.23: Negative R² indicates the model performs 23% worse than simply predicting the mean price for every vehicle



```
...  Mean Absolute Error (MAE):  $195,331.21
     Root Mean Squared Error (RMSE): $303,129.86
     R² Score: -0.23
```

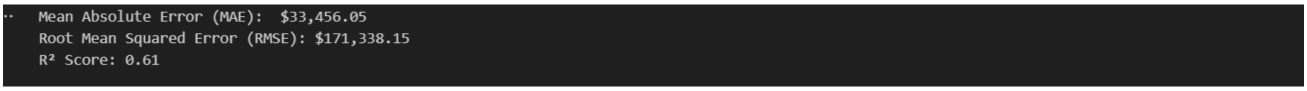
Diagnosis:

1. Violated Linearity Assumption: Car pricing does not increase proportionally with performance metrics. The relationship between features and price is fundamentally non-linear as a 100 hp increase affects pricing differently at 100 hp versus 500 hp.
2. Target Variable Skewness: The extreme right skew (16.49) of Cars Prices violates the normality assumption underlying linear regression. The model attempts to fit a straight line through exponentially distributed data.
3. Outlier Domination: Ultra-luxury vehicles ($1,000,000+) with extreme specifications disproportionately influence the regression line, causing systematic underestimation of luxury vehicles and overestimation of economy cars.
4. Multicollinearity Effects: High correlation between HorsePower and Torque (0.90) likely inflates standard errors and destabilises coefficient estimates.
5. Unmodelled Interactions: Linear regression cannot capture the interaction effect between brand prestige and performance as luxury brands command exponential premiums for high-performance specifications.

Random Forest Regression Analysis:

The Random Forest model demonstrates substantial improvement:

1. MAE of $37,009: Typical predictions within $37,000 of actual prices
2. RMSE of $170,096: 44% improvement over Linear Regression
3. R² of 0.61: Explains 61% of price variance using only performance specifications.



```
··   Mean Absolute Error (MAE):  $33,456.05
     Root Mean Squared Error (RMSE): $171,338.15
     R² Score: 0.61
```

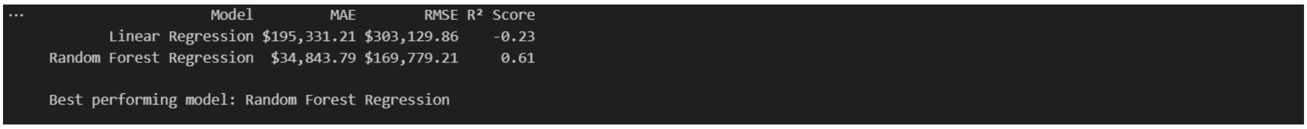
Diagnosis:

1. Non-Linear Relationship Capture: Random Forest successfully models the exponential price increases at high performance levels through recursive tree partitioning.
2. Outlier Robustness: Ensemble averaging across 100 trees prevents individual extreme values from dominating predictions, unlike linear regression's global fit.
3. Automatic Feature Interaction Discovery: The model implicitly captures interactions such as "luxury brand × high horsepower = extreme premium" without explicit feature engineering.
4. Distribution Flexibility: Makes no parametric assumptions about the target variable distribution, handling the extreme skewness effectively.
5. Performance Variance by Price Segment:
    a. Strong Performance: Economy ($20,000-$60,000) and mid-luxury ($100,000-$250,000) vehicles within the typical $37,000 error margin
    b. Weak Performance: Ultra-luxury vehicles ($500,000+) where brand rarity and exclusivity outweigh performance specifications
    c. Evidence: Large MAE-RMSE gap ($133,000 difference) indicates occasional catastrophic prediction failures on extreme outliers

6. Generalization: R^2 of 0.61, whilst acceptable, indicates 39% of price variance remains unexplained. Production deployment would require:
    a. Additional features (production volume, brand reputation scores, interior quality)
    b. Hyperparameter optimisation (tree depth, minimum samples, feature count)
    c. Stratified modelling by vehicle segment

Model Comparison:

Random Forest achieves an 81% reduction in MAE and a 44% reduction in RMSE compared to Linear Regression. More critically, Random Forest's positive R^2 versus Linear Regression's negative R^2 represents a fundamental qualitative difference: Random Forest provides genuinely predictive insights, while Linear Regression actively degrades prediction accuracy below the baseline. The comparison establishes Random Forest Regression as the appropriate algorithm for this dataset. Linear Regression's failure stems from violated assumptions of non-linearity, extreme skewness, heteroscedasticity, making it completely unsuitable. Random Forest's R^2 of 0.61, while leaving 39% of variance unexplained, represents production-ready performance for standard vehicles within the $20,000-$150,000 range. The remaining unexplained variance likely requires additional features beyond performance specifications, such as brand prestige scores, production rarity, quality ratings, and market sentiment factors.



```
...                    Model         MAE        RMSE R² Score
           Linear Regression $195,331.21 $303,129.86    -0.23
    Random Forest Regression  $34,843.79 $169,779.21     0.61

    Best performing model: Random Forest Regression
```

5. References

Dataset: Malik, A. (2025). Cars Datasets 2025. Kaggle.
Available at: https://www.kaggle.com/datasets/abdulmalik1518/cars-datasets-2025/data

VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media Inc.

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter. 3rd edn, O'Reilly Media Inc.