**NAME: BISIRIYU AFEEZ TOLULOPE**

**STUDENT ID: 24834237**

**INTRODUCTION**

This report presents a comprehensive analysis of a vehicle dataset The primary objective is to uncover patterns, relationships, and actionable insights from the data using a structured data science pipeline. By examining key questions such as the best predictors of vehicle price and interesting groupings within the dataset, the analysis aims to inform data-driven decisions and enhance business strategies.

**DATASET OVERVIEW**

The dataset contains information about vehicles, including attributes such as mileage, price, year of registration, fuel type, body type, and vehicle condition. The dataset comprises 402,005 rows and 12 features, which include numerical, categorical, and Boolean data types. Key target variables include **price** (numerical) and **vehicle condition** (categorical).

**1.DATA UNDERSTANDING AND EXPLORATION**

**1.1 Feature Descriptions and Data Types:** The dataset comprises the following features:

* **Mileage**: Indicates the total distance a vehicle has travelled. Units are assumed to be in miles.
* **Public Reference**: A special number or reference code linked to the vehicle listing so that it may be found and retrieved by the public.
* **Reg Code:** The vehicle's registration code provides information about the vehicle's official registration status.
* **Price**: The selling price of the vehicle in local currency.
* **Year of Registration**: The year the vehicle was registered, ranging from 1999 to 2020.
* **Fuel Type**: Types of fuel used by vehicles, including Petrol, Diesel, Hybrid, and Electric.
* **Body Type**: Categories describing the vehicle structure, such as SUV, Hatchback, Convertible, and Saloon.
* **Vehicle Condition**: Binary classification indicating whether the vehicle is new or used.
* **Standard Colour:** The exterior colour of the car.
* **Standard Make:** The cars’ manufacturer or brand.
* **Standard Model:** This is the specific model or version of the car.
* **Vehicle Age**: Calculated as the difference between the max year (2020) and the year of registration.

**Initial Data Inspection**

**Data Shape and Missing Values**

* The dataset originally contained **402,005 rows and 12 columns**.
* Missing data was present in numerical features like **mileage** and **year of registration**, as well as categorical features like **fuel type** and **body type**.

**A screenshot of a computer program

Description automatically generated**

**1.2 ANALYSIS OF DISTRIBUTIONS**

The distributions for numerical and categorical variables are analysed below:

A screenshot of a computer

Description automatically generated

**1.2.1 Numerical Features**

* **Mileage**: Highly skewed distribution with a significant number of low-mileage vehicles. The 'mileage' column contains 401,878 entries, with an average mileage of around 37,743 miles. The minimum recorded mileage is 0, which could indicate an anomaly or data entry error. The median mileage is 28,629.5 miles, marking the midpoint of the distribution. The wide range of mileage values, extending up to 999,999, suggests the presence of outliers or potential errors in the dataset that warrant further review.
* **Price**: Exhibits a wide range with substantial outliers at higher price points. The 'price' column contains 402,005 entries, with an average value of approximately $17,342. The lowest recorded price in the dataset is $120, while the median price stands at $12,600, representing the midpoint of the distribution. The maximum price is $9,999,999, which may be an outlier or the result of a data entry error.

**1.2.2 Categorical Features:**

* **Standard Make**: By analysing the distribution of the most popular car makes in the dataset, it is evident that certain car brands dominate the market. BMW, Audi, Volkswagen, Vauxhall, and Mercedes-Benz are among the most frequently occurring makes in the dataset. The counts of each make reflect their market presence, with higher counts indicating greater popularity and dominance within the analysed data.

A graph of cars with names

Description automatically generated with medium confidence

* **Fuel Type:** Petrol vehicles dominate the dataset, followed by Diesel, Hybrid, and Electric vehicles. In Body Type, SUV and Hatchback are the most common body types

A screenshot of a computer

Description automatically generatedA graph with blue bars

Description automatically generated

-**Vehicle Condition:** The dataset reveals a significantly larger proportion of used cars compared to new ones. This indicates that the majority of vehicles in the dataset are pre-owned, potentially due to factors such as depreciation, affordability, or consumer buying preferences. The distribution is visualized in the following plot

A graph of a vehicle condition

Description automatically generated

**1.2 Analysis of Predictive Power of Features**

* **Correlation Analysis**: Mileage exhibited a strong negative correlation (-16) with Price. Vehicles with higher mileage tend to have lower prices.

A screenshot of a graph

Description automatically generated

Categorical variables showed clear distinctions in average prices. For instance, new vehicles had consistently higher prices than used ones.

* **Scatterplots**: Relationships between Mileage and Price showed a downward trend. Clustering within fuel categories suggested further potential for feature engineering.

A graph with blue dots

Description automatically generated

* Scatterplots confirmed significant relationships between predictors and the target variable. These guided feature engineering decisions to enhance the model's predictive capabilities.

**1.3 Data Processing for Exploration and Visualization**

* **Handling Missing Values**:

A screen shot of a computer program

Description automatically generated

Missing numerical values, such as Mileage and year of registration were imputed using the median to maintain the central tendency without being influenced by outliers. For categorical features, missing values were replaced with the most frequent category (mode) to preserve data integrity.

* **Target Encoding and Scaling**:

Target encoding was applied to categorical features, such as Fuel Type, to preserve their ordinal relationships and avoid dimensionality explosion. Standard Scaler was used to normalize numerical features, ensuring all features contributed equally to model training.

A graph of a graph of a graph

Description automatically generated with medium confidence

Pre- and post-transformation visualizations, such as scaled histograms and encoded categories, demonstrated the effectiveness of these preprocessing steps. The processed data was better suited for machine learning tasks, enabling the models to learn patterns more effectively.

**2. DATA PREPROCESSING**

* 1. **Data Cleaning**

-Missing values were handled by imputing **numerical features** with their median and **categorical features** with their most common values.

1. mileage filled with the median value.
2. year\_of\_registration filled with median.
3. reg\_code, standard colour, fuel\_type, and other missing values were replaced with "Mode".
4. Duplicates were checked and removed from the dataset.

A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated

1. Handling Outliers: Outliers in **mileage** and **price** were detected using the **IQR method**, which flagged values outside the range of 1.5 times the interquartile range as potential outliers. Instead of removing these values, they were capped to the nearest valid boundary to preserve data integrity. Boxplots before and after capping demonstrated a significant reduction in extreme values, improving the dataset's suitability for analysis while retaining meaningful variability.

A graph showing a mileage

Description automatically generatedA green rectangle with white text

Description automatically generated

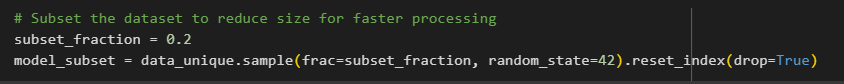
A graph with numbers and lines

Description automatically generated with medium confidenceA green graph with a line

Description automatically generated

1. Subsetting

A screen shot of a computer

Description automatically generated

The dataset was initially filtered to include vehicles registered between **2013 and 2025**, based on the distribution of the year\_of\_registration feature. This step ensured that outdated and erroneous entries, such as the registration year 999, were excluded, resulting in a cleaner and more relevant subset for analysis. The subset retained most valid data points, aligning with the interquartile range of key features like mileage and price. Since the first subset was large, we further sampled **20%** of this subset to reduce computational overhead while maintaining a representative dataset for analysis and modeling. This approach balanced efficiency and accuracy, ensuring the subset remained statistically significant for further exploration and machine learning tasks.

### **Analysis of Detecting and Removing Unrealistic Mileage for Used Cars**

A screenshot of a computer

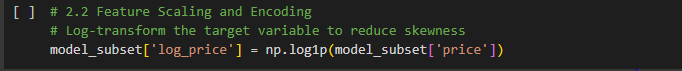
Description automatically generatedA computer screen shot of a computer code

Description automatically generated

During the data cleaning process, **342 rows** were identified where the vehicle\_condition was labeled as **'USED'**, but the mileage was recorded as **0**. This is unrealistic, as used vehicles are expected to have non-zero mileage. These inaccuracies likely stem from missing mileage values imputed as zero. To maintain data integrity and enhance the reliability of the predictive model, these rows were removed. The updated dataset now excludes these entries, ensuring a more accurate representation of used vehicles. This cleaning step reduces noise and biases in the dataset, thereby improving the quality of insights derived from subsequent analyses. By addressing this issue, we ensure that the predictive models are trained on realistic data, enhancing their ability to generalize to new, unseen cases. The cleaned dataset is now better suited for robust machine learning modeling.

**2.2 Feature Engineering and Transformations**

* **Log Transformation:**



Applied to Price, resulting in a normalized distribution that improved the linearity of relationships with other features.

* **Encoding and Scaling:**

Target encoding effectively reduced high cardinality in categorical features like Fuel Type, preserving their predictive potential.

A screen shot of a computer program

Description automatically generated

Scaling via Standard Scaler ensured numerical features like Mileage had consistent ranges, facilitating gradient-based learning.

A computer screen with white text

Description automatically generated

**Visual Evidence and Interpretation:**

A graph of a graph of a graph

Description automatically generated with medium confidence

* Pre- and post-transformation histograms highlighted the benefits of these preprocessing steps. The normalized and scaled data provided a strong foundation for modeling.

**3. Model Building**

**3.1 Algorithm Selection, Model Instantiation, and Configuration**

Three regression models were selected based on their capabilities to capture different types of relationships:

A screen shot of a computer program

Description automatically generated

Linear Regression served as a baseline for linear trends, Decision Tree captured non-linear feature interactions, and k-NN explored localized patterns. Linear Regression required no hyperparameter tuning, while Decision Tree and k-NN were configured with varying depths and n\_neighbors’ values, respectively, to optimize performance. Visual analyses highlighted each model's strengths: Decision Trees effectively segmented data, Linear Regression revealed feature impacts through coefficients, and k-NN demonstrated its ability to model localized relationships. This comprehensive approach ensured a robust exploration of the dataset.

**3.2 Grid Search, Model Ranking, and Selection**

* **Hyperparameter Tuning:**

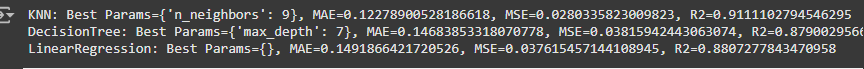
A computer code with numbers and symbols

Description automatically generated

For k-NN, tuning identified n\_neighbors = 9 as optimal. Decision Tree Regressor performed best with max\_depth = 7, balancing complexity and performance. Linear Regression required no hyperparameter tuning,

**Performance Metrics:**

A computer screen with white text

Description automatically generated

A screenshot of a data

Description automatically generated

**Summary Analysis: Model Results and Insights**

The evaluation of three regression models highlights key differences in performance and suitability:

**-k-Nearest Neighbors (KNN):** KNN emerged as the top-performing model, achieving the lowest MAE (0.1228) and MSE (0.0280), along with the highest R² (0.9111), indicating its ability to capture over 91% of the variance. Its precision and ability to minimize large errors make it the most suitable model for this task.

**-Decision Tree:** With an MAE of 0.1468 and an R² of 0.8790, the Decision Tree strikes a balance between interpretability and accuracy. While slightly less accurate than KNN, it remains a strong choice for capturing non-linear patterns and providing interpretable results.

**-Linear Regression:** Linear Regression, though simple and interpretable, showed the lowest performance with an MAE of 0.1492 and an R² of 0.8807. It is suitable when simplicity and ease of explanation are prioritized over predictive precision.

* Top of Form

### **4. Model Evaluation and Analysis**

#### **4.1 Coarse-Grained Evaluation**

**Metrics Summary:**

A screenshot of a graph

Description automatically generatedA screenshot of a computer screen

Description automatically generated

* Bottom of Form

Among the models tested, k-NN provides the best overall performance, with the highest R² (0.9111) and lowest error metrics (MAE = 0.1228, MSE = 0.0280). This indicates its superior ability to predict outcomes accurately and minimize errors. Linear Regression, while offering good interpretability, shows a lower R² (0.8807) and higher error metrics, especially in capturing non-linear relationships. Decision Tree performs similarly to Linear Regression in R² (0.8790) but exhibits a tendency to overfit small data segments, as seen in its marginally better error metrics (MAE = 0.1468, MSE = 0.0382). Visual analysis further supports k-NN’s stronger alignment with actual values, while Linear Regression and Decision Tree reveal performance limitations in specific regions of the data. Therefore, k-NN is the most reliable model for this task, with Linear Regression and Decision Tree being less effective due to their respective limitations

**.4.2 Feature Importance**

**-Decision tree**

A graph with a bar graph

Description automatically generated

In the Decision Tree model, the most important feature is standard\_model, with an importance score of approximately 0.7, indicating its significant influence on predictions. Other moderately important features include reg\_code and mileage, while features such as standard\_make, vehicle age, fuel\_type, and body\_type contribute minimally. The dominance of standard\_model suggests it plays a critical role, while low-importance features, such as fuel\_type and body\_type, may be redundant. To optimize the model, removing less relevant features (e.g., standard\_colour, vehicle condition) could enhance training speed without losing accuracy. Additionally, a deeper investigation into standard\_model may help assess whether its dominance is justifiable or overly relied upon.

**-Linear Regression**

A graph with blue and white bars

Description automatically generated

In the Linear Regression model, standard\_model and vehicle\_condition are the most influential features, with absolute coefficients of approximately 0.7 and 0.6, respectively, indicating strong alignment with the target variable. Features such as vehicle age, body\_type, and mileage have minimal coefficients, suggesting weaker contributions to the model. The high coefficients for standard\_model and vehicle\_condition imply these variables have a strong linear relationship with the target, while the smaller coefficients for other features may indicate low correlation or potential multicollinearity.

**-k-NN**

A graph with blue lines

Description automatically generated

In the KNN model, Vehicle Age and Mileage are the most influential features, reflecting their strong predictive power for the target variable, likely log\_price. Fuel Type, Body Type, and Vehicle Condition also have significant contributions. In contrast, features like Standard Color and Registration Code show minimal influence, suggesting they may have limited correlation with the target. The prominence of Vehicle Age and Mileage aligns with real-world expectations, as these factors directly affect vehicle value. This analysis highlights the potential to streamline the model by prioritizing impactful features and removing less relevant ones.

**4.3 Fine-Grained Evaluation**

**Error Analysis**

A detailed breakdown of the error metrics provides a clearer picture of model performance. Here's the analysis:

**Linear Regression**

A graph of error

Description automatically generated

The Linear Regression model shows a relatively high Mean Error of 0.149, indicating challenges in accurately capturing patterns in the data. The Median Error is slightly lower at 0.145, suggesting that most errors are close to this value, but outliers are likely inflating the mean. This discrepancy highlights the model's sensitivity to extreme values, which may be due to the linear nature of the model. Since Linear Regression assumes a linear relationship, this could be problematic if the actual relationships in the dataset are non-linear, and outliers may further impact model performance.

**Decision Tree** A graph of error

Description automatically generated

The Decision Tree model shows a Mean Error of 0.147, slightly better than Linear Regression but still not optimal. The Median Error of 0.143, with a smaller gap between the mean and median, suggests fewer extreme outliers compared to Linear Regression. Despite the model’s ability to capture nonlinear patterns, the high error values may be due to overfitting, especially with deeper trees. The small difference between the mean and median indicates moderate variance in the error distribution, suggesting that while the model performs better than Linear Regression, further tuning is needed to reduce error and improve generalization.

**k-Nearest Neighbors (k-NN)**

A graph of residual error

Description automatically generated

The k-NN model exhibits the lowest Mean Error (0.122) and Median Error (0.118) among the three models, indicating its effectiveness in capturing data patterns and making consistently accurate predictions. The small difference between the mean and median errors suggests minimal impact from outliers. k-NN's strength lies in its use of local neighbors for predictions, which makes it robust in handling datasets with nonlinear relationships. This localized approach helps the model maintain accuracy even when the data is complex, further supporting its superior performance compared to the other models.

A screenshot of a computer

Description automatically generated

**Accuracy:**

k-NN performs the best with the lowest mean and median errors, making it the most reliable model. Decision Tree follows closely but may suffer from slight overfitting. Linear Regression underperforms due to its inability to handle nonlinear relationships and outliers.

**Robustness to Outliers:**

k-NN and Decision Tree are less impacted by outliers, indicated by the small difference between their mean and median errors. Linear Regression is more sensitive to outliers.

**Conclusion:**

Best Model: k-NN is the most accurate and reliable model.

**4.4 Actual vs Predicted plot**

**-k\_NN**

A graph of blue dots

Description automatically generated

The scatter plot comparing Actual log\_price with Predicted log\_price for the k-NN model shows a strong alignment, indicating high predictive accuracy. The small spread around the line suggests minimal prediction error. k-NN effectively captures the relationships in the dataset, as demonstrated by the tight clustering along the diagonal, reflecting its strong generalization capability. Any slight deviations from the line may result from inherent data noise or limitations in the model's hyperparameters (e.g., n\_neighbors = 9).

**-Actual vs Predicted (Decision Tree)**

A graph with blue dots

Description automatically generated

The scatter plot comparing Actual target values with Predicted values for the Decision Tree model shows a general alignment with the diagonal, but there is noticeable spread and clustering. Horizontal bands are visible in some regions, a result of the Decision Tree's piecewise constant predictions. While the model captures the overall trend, it lacks the precision of k-NN. The horizontal bands suggest overfitting, where the tree splits too finely on features, leading to poor generalization in certain areas of the data. The spread suggests that the model's predictions are less consistent and likely influenced by noise in the data or suboptimal depth (max\_depth = 7).

-**Actual vs Predicted (Linear Regression):**

A diagram of a graph

Description automatically generated

The **Actual vs Predicted (Linear Regression)** plot demonstrates that the model captures the general trend of the data, with predictions closely aligning with actual values along the diagonal. However, noticeable dispersion in the lower and upper ranges of log\_price suggests that the model struggles to accurately predict extreme values. This limitation underscores Linear Regression's inability to fully account for non-linear relationships in the dataset, impacting its overall predictive performance.