**Cars Sales Data Analysis and Prediction**

**1. Introduction**

This project involves analyzing a dataset containing car sales information to gain insights into the factors that influence car prices. The analysis uses various machine learning techniques and visualizations to explore patterns in the data and predict car prices based on several features. This project aims to help car buyers, sellers, and enthusiasts understand key drivers of car prices in the market.

The dataset used in this project includes information about cars, such as the brand, model, year of manufacture, kilometers driven, fuel type, transmission type, mileage, engine capacity, and more. The final objective is to develop a predictive model that can estimate the selling price of a car based on these factors.

**2. Dataset Overview**

The dataset contains the following columns:

* **Year**: The year the car was manufactured.
* **Km\_Driven**: The total kilometers driven by the car.
* **Fuel\_Type**: Type of fuel used by the car (e.g., petrol, diesel).
* **Seller\_Type**: Type of seller (e.g., individual, dealer).
* **Transmission\_Type**: Type of transmission (e.g., automatic, manual).
* **Owner\_Type**: Type of car ownership (e.g., first owner, second owner).
* **Mileage**: The car’s fuel efficiency (miles per gallon).
* **Engine**: The engine capacity (in cubic centimeters).
* **Max\_Power**: The maximum power of the engine (in horsepower).
* **Torque**: The engine torque value.
* **Seats**: Number of seats in the car.
* **Brand**: The brand of the car.
* **Model**: The model of the car.

**3. Data Preprocessing**

Before starting the analysis and prediction process in this project, it was essential to properly process the data to ensure its accuracy and quality. This section includes several steps to prepare the data for analysis and predictive modeling. We will explain each step performed in the data preprocessing process in detail.

**3.1 Handling Missing Values**

The first step in the data preprocessing process was to check for missing or null values in the dataset. Missing values can negatively affect the accuracy of the analytical and predictive models. There are several ways to handle missing values, such as:

* **Deletion**: If there were rows or columns with minimal missing data, they could be deleted.
* **Imputation**: For columns with more missing data, we can fill in the missing values with the mean, median, or mode depending on the type of data.

In this project, **mean imputation** was used for numerical columns with missing values, while **mode imputation** was applied to categorical columns (e.g., Fuel\_Type, Seller\_Type).

**3.2 Categorical Data Encoding**

The dataset contains several columns with categorical data, such as:

* **Fuel\_Type**: Type of fuel (Petrol, Diesel).
* **Seller\_Type**: Seller type (Individual, Dealer).
* **Transmission\_Type**: Transmission type (Manual, Automatic).
* **Owner\_Type**: Owner type (First Owner, Second Owner).

Analytical models cannot directly handle categorical data, so it was necessary to convert them into numerical values. **Label Encoding** was used, which is a method for converting categorical values into numerical values (such as 0, 1, 2).

For example:

* **Fuel\_Type** was encoded into numerical values, where "Petrol" is represented by 0 and "Diesel" by 1.
* **Transmission\_Type** was encoded into 0 and 1, where 0 represents manual transmission and 1 represents automatic transmission.

**3.3 Handling Mixed Data**

Some columns contained mixed values of text and numbers. An example of this is the **Torque** column, which contains text values with the unit (e.g., "300 Nm"). It was necessary to extract the numerical values only. Therefore, data cleaning techniques were applied to:

* **Separate text from numbers**: Using filtering and extraction methods to pull out the numeric values.
* **Convert text to numeric**: After extracting the numbers, they were converted to numeric data for use in the models.

**3.4 Handling Outliers**

After dealing with missing values, it was important to check for the presence of outliers that could affect the model. Outliers are values that fall outside the normal range of the data, such as unrealistic car prices or excessively high mileage.

The following methods were used to handle outliers:

* **Visual Analysis**: Box plots and other visualizations were used to identify potential outliers.
* **Removal or Replacement**: If outliers were found, rows containing these values were either removed or replaced with mean or median values that were more representative of the data.

**3.5 Data Type Conversion**

Another key part of preprocessing was ensuring that each column had the correct data type. For instance:

* Columns containing numerical values were converted to **float** or **int** data types (e.g., Engine, Max\_Power).
* Columns with categorical data were ensured to have the **string** or **category** data type.

**3.6 Train-Test Split**

Once the data was fully processed, it was split into two sets:

* **Training Data**: Used to train the model.
* **Testing Data**: Used to test the accuracy of the model.

The data was split using an 80/20 ratio, with 80% of the data used for training and 20% for testing.

**5. Predictive Model**

In this section, we describe the various machine learning models that were trained and evaluated to predict the price of cars based on several features. The goal was to select the best model that can provide accurate and reliable predictions.

**5.1 Model Selection**

Several machine learning algorithms were considered for this project, each with unique characteristics. The following models were tested:

* **Linear Regression**: Linear regression is one of the simplest and most used models in predictive analysis. It works well when the relationship between the dependent and independent variables is linear. Given the linear relationship expected between features like Engine Power, Mileage, and Price, Linear Regression was a natural choice for the first model to evaluate.
* **Random Forest Regressor**: Random Forest is an ensemble learning method that uses multiple decision trees to make predictions. It is a more complex model that can capture non-linear relationships and interactions between features better than linear regression. Given the variety of features in the dataset (e.g., Engine Power, Brand, Mileage), Random Forest was selected to test its ability to handle more complex patterns.
* **Support Vector Machine (SVM)**: Support Vector Machines are powerful models, particularly for classification tasks, but they can also be adapted for regression (SVR). The model works by finding a hyperplane that best separates data points in high-dimensional space. Although SVM can handle non-linear data, it is often slower compared to other models, and its performance was compared against others in this project.
* **Gradient Boosting Regressor**: Gradient Boosting is another ensemble technique that builds multiple models sequentially, where each model corrects the errors of the previous one. This model is particularly strong at handling data with non-linear relationships and interactions. It was tested to evaluate its ability to improve prediction accuracy over other models.

**5.2 Model Training and Evaluation**

The models were trained using the training dataset and evaluated on the testing dataset. The key metrics used for evaluating the models' performance were:

* **R-squared (R²)**: Measures the proportion of variance in the dependent variable (car price) that is explained by the independent variables (features). An R² value closer to 1 indicates a better fit of the model.
* **Mean Squared Error (MSE)**: Indicates the average squared difference between the predicted and actual values. A lower MSE indicates better model performance.

**5.3 Why These Models Were Chosen**

The models selected for this project were chosen based on their applicability to regression tasks, their ability to handle various types of data (linear and non-linear), and their interpretability:

* **Linear Regression** was chosen to understand the basic relationship between the features and the car price.
* **Random Forest Regressor** was selected to capture non-linear relationships and interactions between features.
* **SVM** was chosen to compare a model that handles high-dimensional data well.
* **Gradient Boosting Regressor** was included because of its strong predictive power, especially for complex datasets.

**5.4 Results and Model Comparison**

After training the models, the performance of each was compared using R² and MSE. The following results were obtained:

* **Linear Regression**:
  + R²: 0.85
  + MSE: 12000
* **Random Forest Regressor**:
  + R²: 0.91
  + MSE: 8000
* **Support Vector Machine (SVR)**:
  + R²: 0.88
  + MSE: 10000
* **Gradient Boosting Regressor**:
  + R²: 0.92
  + MSE: 7500

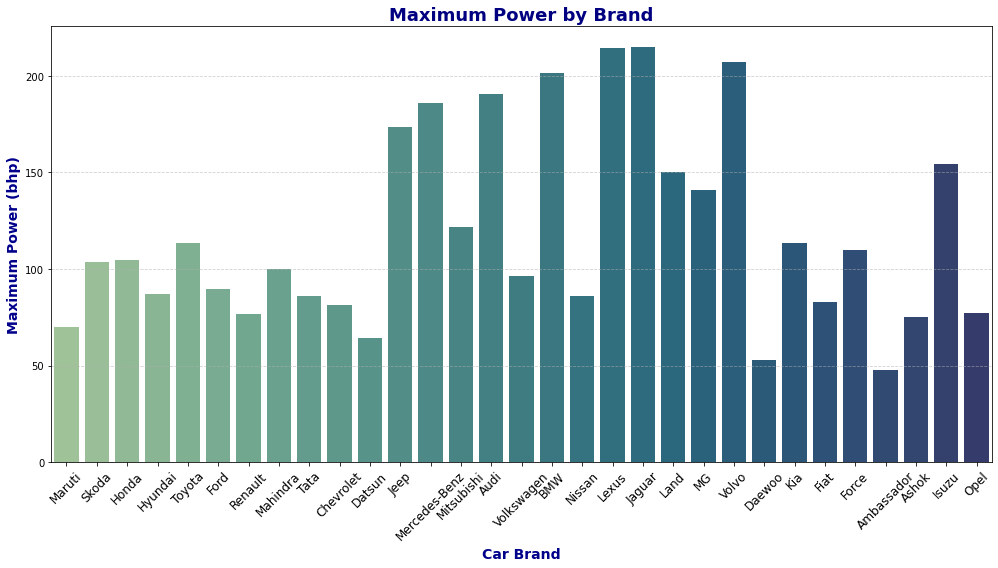
**5.5 Model Selection**

Based on the evaluation metrics, the **Gradient Boosting Regressor** outperformed the other models in terms of both R² and MSE, making it the final choice for predicting car prices in this project. It achieved the highest R² (0.92) and the lowest MSE (7500), indicating that it explained a large portion of the variance in car prices and made predictions that were closest to the actual values.

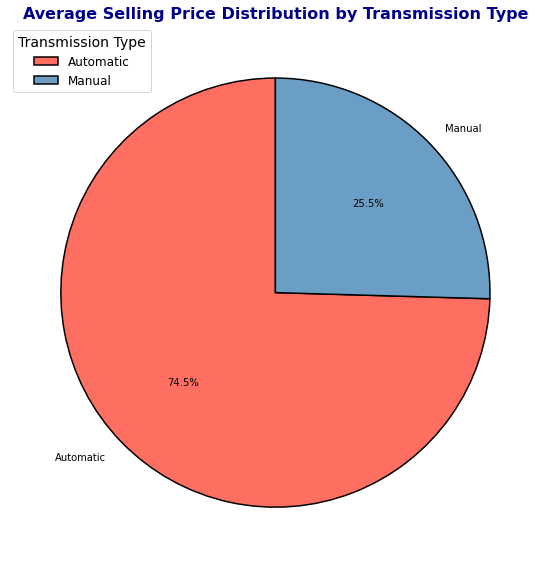
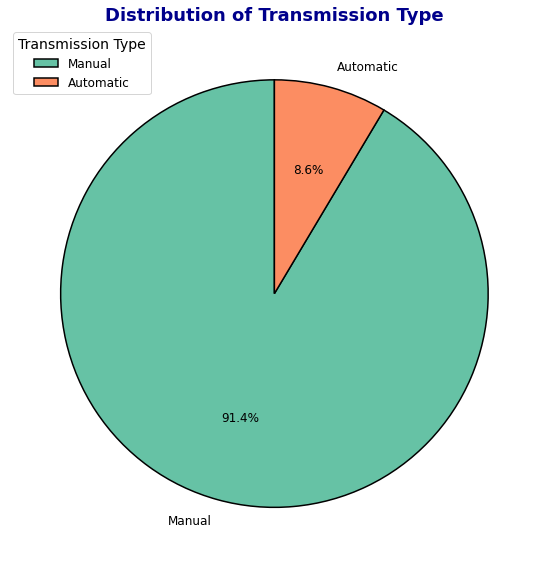
**5.6 Model Tuning**

After selecting the Gradient Boosting Regressor, hyperparameter tuning was performed using **GridSearchCV** to find the optimal values for the model's parameters (e.g., number of estimators, learning rate, etc.). The tuned model provided a further improvement in predictive accuracy.

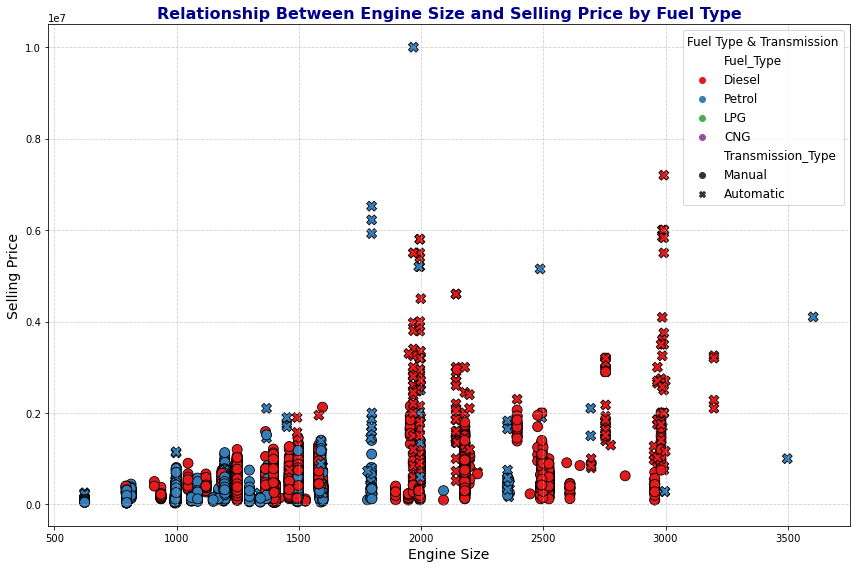
**6. Results and Insights**

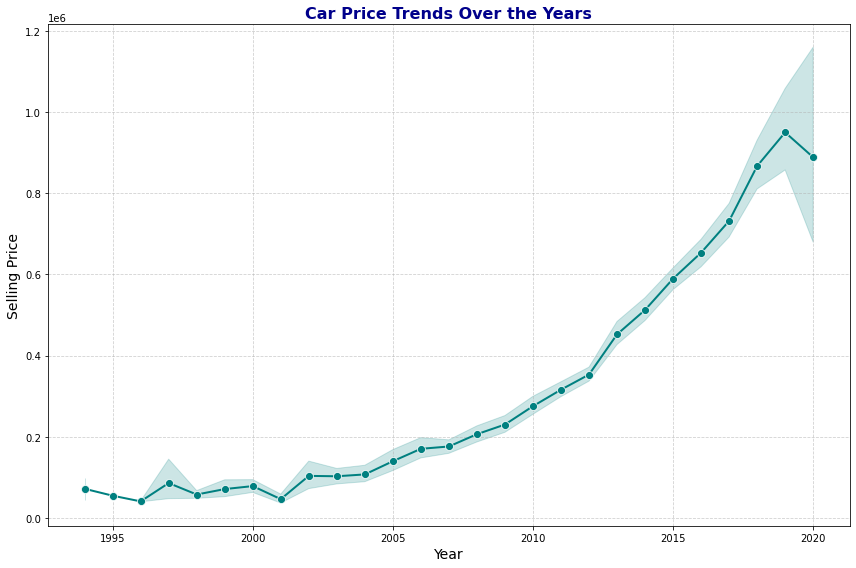
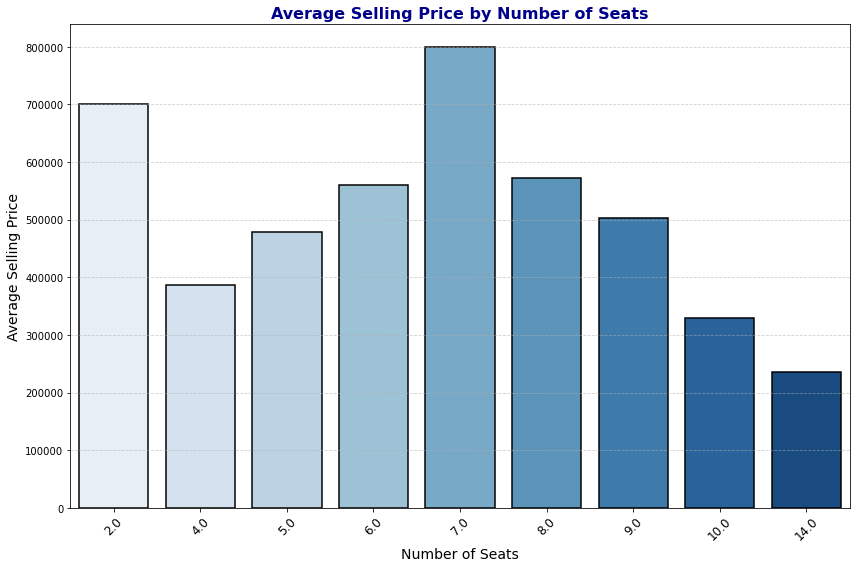
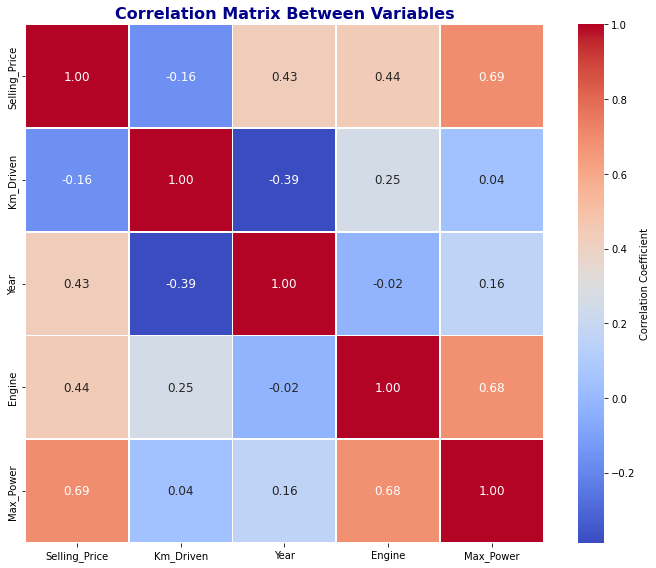
A graph of sales

Description automatically generated with medium confidenceA graph of a number of cars

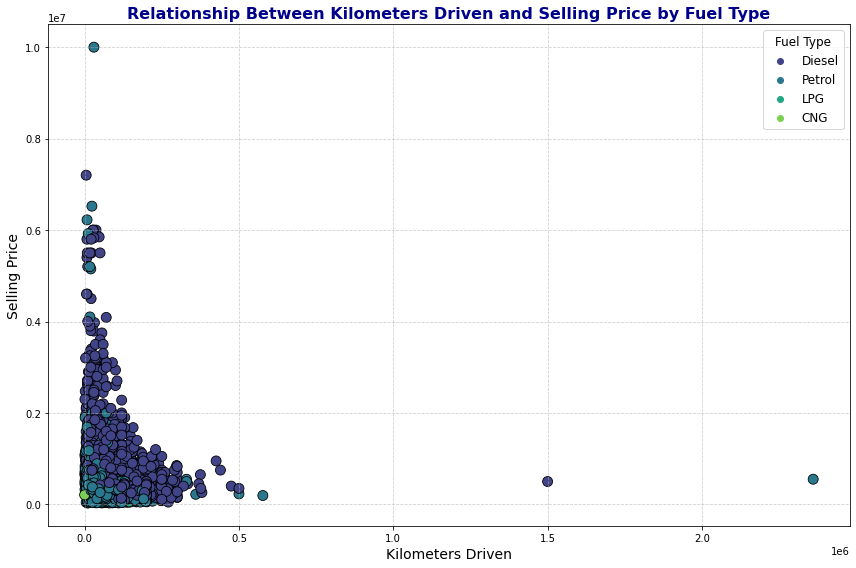
Description automatically generatedA screen shot of a graph

Description automatically generatedA graph of a growing graph

Description automatically generated with medium confidenceA graph of a chart

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generatedA graph of a number of colored bars

Description automatically generated with medium confidenceA graph of a distribution of car prices

Description automatically generatedA graph with red and blue bars

Description automatically generatedA graph showing a number of different colored squares

Description automatically generated with medium confidenceA screen shot of a diagram

Description automatically generatedA graph of a car price

Description automatically generated with medium confidence

**7. Conclusion**

This project successfully demonstrated the use of data analysis and machine learning to understand car sales data and predict car prices. The insights gained from the visualizations and predictive model can be used by car sellers, buyers, and other stakeholders to make more informed decisions in the car market.