PROJECT REPORT ON

Predicting Used Car Prices with Advanced Models on CarDekho Data

by

MAHIDHARA S

1-33266551790

For partial fulfillment of the requirements of the second-year curriculum of the two years full time AICTE PGDM Program.

Submitted to

RAMAIAH INSTITUTE OF MANAGEMENT STUDIES

Gokula,Bangalore

STUDENT DECLARATION

I hereby solemnly affirm declare and state that report titled ***"Predicting Used Car Prices with Advanced Models on CarDekho Data"*** was done by me with due diligence and sincerity and this report based on that study is a bonafide work by me and submitted to Ramaiah Institute of Management Studies (RIMS) under the guidance and supervision of **Prof. Bharath R**. This project report is my original work and not submitted for the award of any other degree, diploma, fellowship or other similar title or prizes.

PLACE: **Bengaluru** **SIGNATURE:**

DATE: **30/07/2024** **ENROLLMENT No.: 1-33266551790**

CERTIFICATE FROM THE GUIDE

This is to certify that the Project report “Predicting Used Car Prices with Advanced Models on CarDekho Data” by Mahidhara S, Enrollment no*: 1-33266551790* carried out in partial fulfillment for the award of degree of POST GRADUATE DIPLOMA IN MANAGEMENT (PGDM) AICTE APPROVED at Ramaiah Institute of Management Studies(RIMS), Bangalore has been prepared under my guidance and direction. This study report is an original work and has not been submitted earlier to any University/Institute as per my knowledge and belief.

PLACE: BANGALORE **Signature:**

DATE:30/07/2024 **Guide Name: Prof. Bharath.R**

ACKNOWLEDGMENT

I would like to thank the founder trustees Dr.M.R Pattabhiram and Mrs.AnithaPattabhiram for providing the infrastructure and the facilities to complete my project. I would also like to thank Dr.M Swapna, Principal of RIMS for being a constant source of inspiration. My thanks are due to Prof. Bharath.R for their guidance during the period of my project.

Signature of the Student:

Date: 30/07/2024

TABLE OF CONTENT

FIGURES

1. Abstract

The automotive industry has experienced considerable transformations, particularly within the used car market, which has emerged as a vital segment driven by escalating new vehicle prices and rapid technological advancements. This market dynamics complexity necessitates accurate price determination for used cars, which is often encumbered by uncertainties. This research aims to demystify the intricacies of used car pricing by harnessing machine learning and deep learning algorithms to predict used car prices, utilizing data from the Indian automotive platform, Cardekho.

Objective: The primary objective of this research is to analyze the selling prices of used cars listed on Cardekho and to develop highly accurate predictive models using both machine learning and deep learning techniques. By leveraging the available data, the aim is to construct models that can predict the prices of used cars with precision, providing valuable insights and tools for buyers, sellers, and stakeholders in the automotive industry.

Accurately pricing a used car involves a multifaceted analysis of numerous factors, including vehicle characteristics, condition, and market dynamics. These factors collectively contribute to the final price of the car. Vehicle characteristics encompass attributes such as make, model, year of manufacture, mileage, engine capacity, transmission type, fuel efficiency, and features. The condition of the car, including any repairs, accidents, and prior ownership history, significantly impacts its price. Market dynamics, including geographic location, brand popularity, market fluctuations, and supply-demand trends, also play a crucial role in determining car prices. Traditional methods of pricing, relying on intuition and dealer quotes, are often subjective and lack the precision needed for fair pricing. This research advocates for a systematic, data-driven approach to ensure accurate and unbiased price estimation.

Machine learning offers a revolutionary approach to used car price prediction by analyzing vast datasets of historical car sales. These algorithms can identify intricate patterns and relationships between various car features and their corresponding market values, enabling the development of predictive models capable of estimating the fair market price of a used car based on its specific characteristics. Data-driven insights derived from these models enable more accurate and reliable price predictions compared to traditional methods. The use of machine learning ensures improved accuracy, transparency, and fairness in the pricing process, which is crucial for building trust between buyers and sellers.

Numerous studies have demonstrated the potential of machine learning in used car price prediction, exploring various algorithms such as Linear Regression, Support Vector Machines (SVM), Random Forest, XGBoost, Ridge Regression, Lasso Regression, AdaBoost, Gradient Boosting, Decision Trees, k-Nearest Neighbors (k-NN), and Deep Learning. Each algorithm offers unique strengths and weaknesses, with deep learning models particularly noted for their ability to automatically learn feature representations from raw data, capturing intricate, non-linear relationships.

Building on existing knowledge, this research focuses on developing a machine learning model tailored to the Cardekho dataset. The research objectives include evaluating and comparing multiple algorithms to identify the most accurate price predictions, investigating the impact of different feature sets on model performance, and optimizing the chosen model for maximum accuracy and generalizability. This comprehensive approach to feature engineering aims to enhance the predictive power of the models, considering additional factors such as accident history and number of previous owners.

The outcomes of this research hold significant implications for buyers, sellers, and the broader automotive industry. For buyers, an accurate price prediction model can provide reliable estimates of a car's fair market value, aiding informed purchasing decisions. For sellers, the model serves as a valuable tool for setting competitive prices, increasing the likelihood of successful transactions. Additionally, this research contributes to the broader field of machine learning and predictive analytics, showcasing the practical applications and benefits of data-driven approaches.

In conclusion, this research aims to leverage the power of machine learning to develop a robust and reliable used car price prediction model using the Cardekho dataset. By evaluating and comparing multiple algorithms, conducting thorough feature engineering, and optimizing the chosen model, this study seeks to provide valuable insights and practical tools for the used car market. The following sections will delve into the methodology, results, and discussion, ultimately contributing to a more informed and transparent used car pricing landscape.

2. Introduction

The automotive industry has undergone significant transformations over the years, with the used car market emerging as a pivotal segment within this dynamic landscape. The soaring prices of new vehicles, coupled with rapid technological advancements, have made pre-owned cars an attractive option for a wide range of consumers. However, navigating the used car market and accurately determining the fair value of a used car is an intricate process, often shrouded in uncertainties. This research aims to demystify this complex landscape by leveraging machine learning algorithms to predict used car prices, utilizing data from the Indian automotive platform, Cardekho.

The Challenge: Unraveling the Price Maze

Accurately pricing a used car involves a multifaceted analysis of numerous factors. These factors can be broadly categorized into vehicle characteristics, condition, and market dynamics, each contributing uniquely to the final price of the car.

Vehicle Characteristics: The intrinsic attributes of a car, such as its make, model, year of manufacture, mileage, engine capacity, transmission type, fuel efficiency, and features, play a crucial role in determining its market value. For instance, a car with a well-known and respected brand, lower mileage, higher fuel efficiency, and advanced features is likely to command a higher price compared to one lacking these attributes.

Condition: The overall condition of the car, including any repairs, accidents, and prior ownership history, significantly impacts its price. A car that has been well-maintained and has no history of major repairs or accidents will generally be valued higher than one with a less favorable condition.

Market Dynamics: Factors such as geographic location, brand popularity, market fluctuations, and overall supply-demand trends also influence car prices. For example, certain brands might be more popular in specific regions, affecting the supply-demand balance and, consequently, the price.

Traditionally, car buyers and sellers have relied on a combination of intuition, personal experience, and dealer quotes to navigate this complex price landscape. However, these

traditional methods are often subjective and lack the precision and objectivity required for fair pricing. The inherent complexities involved necessitate a more systematic and data-driven approach to ensure accurate and unbiased price estimation.

The Opportunity: Machine Learning and Deep Learning

Machine learning offers a revolutionary approach to used car price prediction by analyzing vast datasets of historical car sales. These algorithms can identify intricate patterns and relationships between various car features and their corresponding market values, enabling the development of predictive models capable of estimating the fair market price of a used car based on its specific characteristics.

Data-Driven Insights: Machine learning models derive insights from historical data, uncovering hidden patterns and correlations between car features and selling prices. This data-driven approach allows for more accurate and reliable price predictions compared to traditional methods.

Improved Accuracy: By analysing a multitude of data points, machine learning models can produce more precise price estimations. These models can handle complex and non-linear relationships between features and prices, which are often challenging for traditional methods to capture.

Transparency and Fairness: The use of data-driven models reduces the risk of subjective biases, ensuring a more transparent and equitable pricing process. This is particularly important in the used car market, where fairness and objectivity are crucial for building trust between buyers and sellers.

Research Landscape

Numerous studies have demonstrated the potential of machine learning in used car price prediction. Researchers have explored and evaluated various algorithms, each with its strengths and weaknesses:

Linear Regression: A fundamental approach for establishing linear relationships between features and prices. While simple and interpretable, linear regression may not capture complex non-linear relationships effectively.

Support Vector Machines (SVM): Effective for high-dimensional data and handling complex non-linear relationships. SVMs can provide robust predictions but may require extensive tuning and computational resources.

Random Forest: A robust ensemble method combining multiple decision trees for improved prediction accuracy. Random forests are particularly effective in handling diverse and complex datasets but can sometimes lack interpretability.

XGBoost: A powerful gradient boosting algorithm capable of learning intricate interactions between features. XGBoost is known for its high performance and flexibility but may require careful tuning to avoid overfitting.

Ridge Regression: An extension of linear regression that includes a regularization term to penalize large coefficients. This method is effective in handling multicollinearity (high correlation among predictor variables) and helps prevent overfitting. Ridge regression works by adding a penalty equal to the square of the magnitude of coefficients to the loss function. While it can improve prediction accuracy in the presence of collinear features, it may not effectively perform feature selection since it doesn't shrink coefficients to zero.

Lasso Regression: Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) adds a regularization term to the linear regression model, but uses the absolute value of the coefficients rather than their squares. This results in some coefficients being exactly zero, which effectively selects a simpler model by eliminating less important features. Lasso regression is particularly useful when we have a high number of features, as it performs both regularization and feature selection, making the model more interpretable.

AdaBoost: Short for Adaptive Boosting, AdaBoost is an ensemble learning method that combines multiple weak learners to create a strong predictive model. Typically, decision trees with a single split (decision stumps) are used as weak learners. AdaBoost works by iteratively adjusting the weights of incorrectly classified instances, focusing more on difficult cases in subsequent iterations. This results in a model that is highly accurate and can handle complex datasets. However, AdaBoost can be sensitive to noisy data and outliers.

Gradient Boosting: A powerful ensemble technique that builds models sequentially, each new model correcting the errors of the previous ones. It uses gradient descent to minimize a loss

function, iteratively adding weak learners (typically decision trees) to the ensemble. Gradient Boosting is highly effective for a variety of predictive modeling tasks and can handle complex interactions between features. Despite its high performance, it requires careful tuning to avoid overfitting and can be computationally intensive.

Decision Tree: A non-parametric model that splits the data into subsets based on the value of input features, creating a tree-like structure. Each node represents a decision rule, and each branch represents the outcome of that rule. Decision trees are easy to interpret and visualize, making them useful for understanding complex datasets. However, they can be prone to overfitting, especially with deep trees, and may not perform well with noisy data.

k-Nearest Neighbors (k-NN): A simple, instance-based learning algorithm that classifies a data point based on the majority class among its k-nearest neighbors in the feature space. For regression tasks, the average value of the k-nearest neighbors is used. k-NN is intuitive and effective for small datasets with a well-defined distance metric. However, it can be computationally expensive for large datasets and is sensitive to the choice of k and the scaling of features.

Deep Learning: Deep learning, a subset of machine learning, employs neural networks with multiple layers to model and understand complex patterns in data. Unlike traditional machine learning algorithms that rely on manually crafted features, deep learning models automatically learn feature representations from raw data. This capability allows deep learning to capture intricate, non-linear relationships between input variables, making it particularly powerful for tasks involving high-dimensional data. In the context of used car price prediction, deep learning models can effectively process and analyze diverse features such as vehicle specifications, market trends, and historical sales data. By leveraging deep learning, we can develop robust predictive models that achieve high accuracy even with minimal data fields, offering a significant advantage in scenarios where detailed information is scarce or incomplete. The capacity of deep learning to generalize from limited inputs and uncover hidden structures within the data makes it a compelling choice for enhancing the precision and reliability of used car price predictions.

These studies highlight the effectiveness of machine learning, with models achieving significantly higher accuracy compared to traditional methods. The advancements in machine

learning algorithms and the availability of large datasets have paved the way for more sophisticated and accurate used car price prediction models.

This Research: Car Price Prediction Using Cardekho Data

Building on the existing body of knowledge, this research focuses on developing a machine learning model specifically tailored to the Cardekho dataset. The objectives of this research are multifaceted:

Evaluating and Comparing Multiple Algorithms: We will explore a range of machine learning algorithms, including linear regression, support vector machines, random forest, and XGBoost, to identify the one that delivers the most accurate price predictions for the Cardekho dataset. This comparative analysis will provide valuable insights into the strengths and weaknesses of each algorithm in the context of used car price prediction.

Feature Engineering: We will investigate the impact of different feature sets on model performance. In addition to the existing features in the Cardekho dataset, we will consider additional factors such as accident history, number of previous owners, and more. This comprehensive approach to feature engineering aims to enhance the predictive power of our models.

Model Optimization: We will fine-tune our chosen model to maximize its accuracy and ensure its generalizability to unseen data. This process involves hyperparameter tuning, cross-validation, and extensive testing to achieve optimal performance.

Significance of This Research

The outcomes of this research have significant implications for both buyers and sellers in the used car market. For buyers, an accurate price prediction model can provide a reliable estimate of a car's fair market value, helping them make informed purchasing decisions. For sellers, the model can serve as a valuable tool for setting competitive prices, thereby increasing the likelihood of successful transactions.

Furthermore, this research contributes to the broader field of machine learning and predictive analytics. By applying advanced machine learning techniques to a real-world problem, this

study demonstrates the practical applications and benefits of data-driven approaches in various industries.

The used car market is a complex and dynamic landscape, where accurate pricing is essential for ensuring fairness and transparency. This research leverages the power of machine learning to develop a robust and reliable used car price prediction model using the Cardekho dataset. By evaluating and comparing multiple algorithms, conducting thorough feature engineering, and optimizing the chosen model, this study aims to provide valuable insights and practical tools for the used car market. The following sections of this paper will delve deeper into the methodology, results, and discussion of our research, ultimately contributing to a more informed and transparent used car pricing landscape.

3. Literature Review

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features. Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website. The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy. The data set used for the prediction models was created by Shonda Kuiper. The data was collected from the 2005 Central Edition of the Kelly Blue Book and has 804 records of 2005 GM cars, whose retail prices have been calculated. The data set primarily comprises of categorical attributes along with two quantitative attributes. [1]

Due to the numerous elements that influence a used vehicle's market pricing, determining if the advertised price is accurate, is a difficult undertaking. The goal of this research is to create machine learning models that can properly forecast the price of a used car based on its attributes so that buyers can make educated decisions. On a dataset containing the sale prices of various brands and models, they have built and analysed several learning approaches. They have examined

the results of numerous machine learning algorithms, such as Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, and Decision Tree Regressor, and pick the best one. The car's pricing will be determined based on several factors. Regression Algorithms are employed because they offer us with a continuous number as an output rather than a categorized value, allowing us to anticipate the real price of a car rather than its price range. A user interface has also been created that takes input from any user and displays the price of a car based on their inputs. There are three types of fuel data sets here. They are Diesel, Petrol and LPG are used here.

Price prediction of used car using machine learning techniques is the first paper. They look at how supervised machine learning techniques can be used to estimate the price of second-hand cars in Mauritius in this study. The forecasts are based on historical data taken from daily

publications. To make the predictions, various techniques such as multiple linear regression analysis, were employed. According to author Sameerchand, car price estimates on historical data gathered from daily newspapers. For estimating the price of cars, they employed supervised machine learning algorithms. Other methods that have been employed include multiple linear regression, k-nearest neighbour algorithms, nave based, and various decision tree algorithms. The best algorithm for prediction was identified after comparing all four algorithms. They had some issues comparing the algorithms, but they succeeded to do so. According to authors Enis Gegic et al, the focus of this paper is on scraping data from an online site utilising web scraping technique. These were then compared using several machine learning techniques to forecast the vehicle pricing in a simple manner. They divided the pricing into distinct price groups that had already been established. On different datasets, artificial neural networks, support vector machines, and random forest methods were utilized to develop classifier models. In this study, Wu et al. exhibit automobile price prediction using a neural fuzzy knowledge-based system. They projected a model that has similar outcomes to the simple regression model by taking into account the following attributes: brand, year of manufacturing, and kind of engine. They have developed an expert system called ODAV (Optimal Distribution of Auction Automobiles) because there is a strong demand for car dealers to sell leased vehicles at the end of the lease year. This method provides information on the greatest vehicle pricing as well as the best location to get them. The K-nearest neighbour machine Learning approach, which is based on regression models, was used to estimate the price of autos. Because a greater number of vehicles have been transferred through this system, it is more effectively managed. [2]

The manufacturer sets the price of a new car in the industry, with the government incurring some additional expenditures in the form of taxes. Customers purchasing a new car may thus be sure that their investment will be worthwhile. However, due to rising new car prices and buyers' financial inability to purchase them, used car sales are increasing globally. As a result, a used car price prediction system that efficiently assesses the worthiness of the car utilizing a range of factors is required. The current system comprises a system in which a dealer decides on a price at random and the buyer has no knowledge of the car or its current worth. In reality, the seller has no clue what the car is worth or what price he should charge for it.

Determining if the quoted price of a used car is fair is a difficult process owing to the numerous elements that influence a used vehicle's market pricing. The goal of this research is to create machine learning models that can properly anticipate the price of a used car based on its

features so that buyers can make informed choices. They created and analysed numerous learning algorithms using a dataset that includes the selling prices of various brands and models. We will compare and choose the best machine learning algorithms such as Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Decision Tree Regression, Random Forest Regression, XG Boost Regression, and Gradient Boosting Regression. The price of the car will be determined by a number of factors. Regression algorithms are used because they produce a continuous value rather than a categorized value, allowing us to predict the actual price of a car rather than the price range of a car. A user interface has also been created that takes input from any user and shows the price of a car based on the inputs. [3]

Various studies have been conducted in order to predict the price of used cars. Researchers regularly anticipate product prices using past data. Pudaruth predicted car prices in Mauritius, and these cars were not new, but rather used to predict the prices, he employed multiple linear regression, k nearest neighbours, Naive Bayes, and decision tree techniques. When the prediction results from various strategies were compared, it was discovered that the prices from these methods are quite similar. However, the decision tree technique and the Nave Bayes approach were proven to be incapable of classifying and predicting numeric values. According to Pudaruth's research, the small sample size does not give good prediction accuracy.

Furthermore, Pudaruth applied various machine learning algorithms, namely: k-nearest neighbours, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometres, production year, exterior colour, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

He also proposed predicting the Price of Used Cars using Machine Learning Techniques. In this paper, they collected the historical data of used cars in Mauritius from the newspapers and applied different machine learning techniques like decision tree, K-nearest neighbours, Multiple Linear Regression and Naïve Bayes algorithms to predict the price. This model has the mean error about Rs.27000 for Nissan cars and about Rs45000 for Toyota cars using KNN

and around Rs51000 using linear regression. The accuracy of decision trees and NaïveBayes algorithm dangled between 60 to 70 percentiles with different parameters and the overall training accuracy of the model is 61%. [4]

Kuiper, S. (2008) demonstrated a multivariate regression model that helps in classifying and predicting values in numeric format. It demonstrates how to apply this multivariate regression model to forecast the price of 2005 General Motors (GM) vehicles. The price prediction of cars does not require any special knowledge. So, the data available online is enough to predict prices. The author of the article did the same car price prediction and introduced variable selection techniques that helped in finding which variables were more relevant for inclusion in the model. [5]

In 2019, Pal et al discovered as a methodology for predicting used cars prices using Random Forest. The paper evaluated used car price prediction using Kaggle data set which gave an accuracy of 83.62% for test data and 95% for train-data. The most relevant features used for this prediction were price, kilometre, brand, and vehicle type and identified by filtering out outliers and irrelevant features of the data set. Being a sophisticated model, Random Forest provided good accuracy in comparison to prior work using these data sets. [6]

Car price prediction is somehow interesting and popular problem. As per information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from which 84% of them are cars for personal usage. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase in future. This adds additional significance to the problem of the car price prediction. Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent change in the price of a fuel. Different features like exterior colour, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this paper, they applied different methods and techniques in order to achieve higher precision of the used car price prediction. This paper is organized in the following manner: Section II contains related work in the field of price prediction of used cars. In section III, the research methodology of our study is explained. Section IV elaborates various machine learning algorithms and examine their respective

performances to predict the price of the used cars. Finally, in section V, a conclusion of our work is given, together with the future works plan.

Predicting price of a used cars has been studied extensively in various research. Listen discussed, in her paper written for Master thesis, that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. This is on the grounds that Support Vector Machine (SVM) is better in dealing with datasets with more dimensions and it is less prone to overfitting and underfitting. The weakness of this research is that a change of simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation. Another approach was given by Richardson in his thesis work. His theory was that car producers produce more durable cars. Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than traditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.

Enis Gegic et al. proposed Car Price Prediction using Machine Learning Techniques. In this paper, they proposed an ensemble model by collecting different types of machine learning techniques like Support Vector Machine, Random Forest and Artificial neural network. They collected the data from the web portal www.autopijaca.ba and build this model to predict the price of used cars in Herzegovina and Bosnia. The accuracy of their model is 87%.

He demonstrates the need to create a model to forecast the cost of second-hand cars in Bosnia and Herzegovina. They used machine learning techniques such as artificial neural networks, support vector machines, and random forests. However, the methods were used in concert. The web scraper, which was created using the PHP programming language, was used to gather the data from the website autopijaca.ba for the forecast. Then, to determine which method best suited the provided data, the respective performances of various algorithms were compared. A Java application contained the final prediction model. Additionally, the model's accuracy of 87.38% was determined when it was verified using test data. [7]

The goal of the system that Dholiya, M., et al. developed is to give the user a realistic estimation of how much the vehicle might cost them. Based on the specifics of the automobile the user is looking for, the system, which is a web application, may also offer the user a list of options for various car kinds. It assists in providing the buyer or seller with useful information on which to base their decision. This system makes predictions using the multiple linear regression

algorithm, and this model was trained using historical data that was obtained over an extended period of time. The raw data was initially gathered using the KDD (Knowledge Discovery in Databases) process. Afterward, it underwent preprocessing and cleaning to identify patterns that are valuable and then derive some meaning from those patterns. [8]

Richardson conducted his analysis under the presumption that automakers are more inclined to produce cars that don't lose value quickly. He demonstrated that hybrid cars are better equipped to maintain their value than conventional vehicles by utilising multiple regression analysis. This is perhaps because there are increasing concerns about the environment and the climate, as well as because it uses less gasoline. In this study, the significance of additional variables including age, mileage, make, and MPG (miles per gallon) was also considered. All his information was gathered from several websites. [9]

Listiani published another study that is comparable and uses Support Vector Machines (SVM) to forecast lease car pricing. This study demonstrated that when a very large data set is available, SVM is significantly more accurate at price prediction than multiple linear regression. SVM is also superior at handling high dimensional data and steers clear of both under- and over-fitting problems. Finding crucial features for SVM is done using a genetic algorithm. However, the method does not demonstrate why SVM is superior to basic multiple regression in terms of variance and mean standard deviation. [10]

Wu et al. conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbour machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it.

The report by Awad et al. is more of an educational paper than a research paper. The author reviews six most popular classification methods (Bayesian classification, ANNs, SVMs, k-NN, Rough sets, and Artificial immune system) to perform a spam email classification task. The reason for choosing this paper is to understand these popular classification models in detail, and its applicability to the spam email classification problem since this paper gives much

insight into each method. The main difference, however, between classifying price range and spam mail, is that spam email classification task is a binary one, whereas our motive is mainly one-vs-the-rest. The author uses Naive Bayes for classification which does not give accurate results due to its major concern of feature dependency as pointed out by the author. Due to this reason, He also did not try to evaluate the performance of our dataset using Naive Bayes model since our dataset has heavily feature dependency. To predict results with good accuracy, the author suggests a hybrid system which applies to our work by using Random Forest. A manipulation of various decorrelated decision trees, the Random Forest gives pretty good accuracy in comparison to prior work. [11]

Gonggie proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models. [12]

Nowadays, the whole society is stepping into the 5G era. The 5G technology supports many application scenarios, expanding from mobile Internet to mobile Web of things expansion. Meanwhile, the government will support to build up high-speed, mobile and safe next-generation information infrastructure. Driverless technology in the 5G era is becoming more advanced and electric cars are widely available, which lead to the result that a great number of cars flowing in the market have to be disposed of (be scrapped or be sold as used cars). So, this project is chosen, using the knowledge of machine learning to predict the transaction prices in the used car market, so as to grasp the second-hand car market situation more effectively. The prediction can help people who has a will to buy a second-hand car for reference. The reason for choosing the machine learning model is that it’s really hard to make prediction, and the relationship between the variables used for prediction and the predicted variables is difficult to be found. However, some machine learning models can solve this problem in a very simple way. This paper uses three prediction models, namely XGBoost, support vector machine (SVM) and neural network to estimate the transaction prices of second-hand cars, and then compares the prediction effect. [13]

The second-hand car market has continued to expand even as the reduction in the market of new cars. According to the recent report on India’s pre-owned car market by Indian Blue Book,

nearly 4 million used cars were purchased and sold in 2018-19. The second-hand car market has created the business for both buyers and sellers. Most of the people prefer to buy the used cars because of the affordable price and they can resell that again after some years of usage which may get some profit. The price of used cars depends on many factors like fuel type, colour, model, mileage, transmission, engine, number of seats etc., The used cars price in the market will keep on changing. Thus, the evaluation model to predict the price of the used cars is required.

In his paper, he proposed a model to estimate the cost of the used cars using the K nearest neighbour algorithm which is simple and suitable for small data set. Here, they have collected a used cars dataset and analysed the same. The data was trained by the model, and they examined the accuracy of the model among different ratios of trained and test set. The same model is cross-validated for assessing the performance of the model using the K- Fold method which is easy to understand and implement. [14]

Nitis Monburinon et al. proposed a prediction of Prices for Used Car by Using Regression Models. In this paper, the authors selected the data from the German e commerce site. The main goal of this work is to find a suitable predictive model to predict the used cars price. They used different machine learning techniques for comparison and used the mean absolute error(MAE) as the metric. They proposed that their model with gradient boosted regression has a lower error with MAE value 0.28 and this gives the higher performance where linear regression has the MAE value 0.55, random forest with MAE value 0.35. [15]

Kanwal Noor and Sadaqat Jan proposed Vehicle Price Prediction System using Machine Learning Techniques. In this paper, they proposed a model to predict the price of the cars through multiple linear regression method. They selected the most influencing feature and removed the rest by performing feature selection technique. The Proposed model achieved the prediction precision of about 98%. [16]

In India the automobile market is a biggest business for international and Indian automobile companies. As the boom and demand for automobiles increase there is also a big market opening for used cars. The used car market is being manipulated and controlled by some of the online advertisement websites like olx and quickr, but customers who want to buy a used car is easily being manipulated and cheated to a higher price which the car isn't worth buying for. During 2019-20 the entire automobile production in India was 26,353,293, But in 2020-21 the automobile production in India was 22,652,108. It is observed that, there is a huge decline in

automobile industry, people are preferring more on used and second-hand vehicles than new vehicles. Therefore, the system of used cars must be standardized, and a clear pricing system needs to be implemented. Abishek's paper suggests few machine techniques which can be used to predict the prices of used cars with historical used car prices data and considering a mean value from the list of prices for a specific car and assigning it as the predicted price for the given features and parameters.

There have been many related works done regarding this topic and field but only very few or one or two authors have done for Indian dataset, Thus I wanted to find a solution for this problem and find prediction method to give the prices for used cars in a correct method. A car price prediction has been a high-interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. As the demand for cars increase the demand for second hand and used cars also increases so due to this high demand, we need to build an AI solution for solving this demand in a customer friendly way. The customers are getting cheated and tricked for a higher price for a less worth used car if the customer wants to buy it from a dealer who sells used cars. The dealer tries to sell a damaged or repaired car for high price to customers who don't know much about buying cars and stuff. The customer who doesn't know about the technical specifications and other prices of spare parts and how to deduct the price will easily be cheated with high price. Using machine learning it is possible to predict the correct and worthy price for a given used car based on previous data from various sellers and buyers. This was done by training the model using used cars dataset which has several features and parameters such as year of manufacturing, model year, number of cylinders, number of kms/miles driven, diesel or petrol, automatic or manual or other type of transmission, the gearing system of the cars, the number of owners of the car etc., like this there are many features from which the cars price can be predicted. The data of any damage or is it flood affected or accidental damaged car these factors can also be considered for predicting the correct and exact price of the car. [17]

Due to the large growth in the number of cars being bought and sold, used-car price prediction creates a lot of interest in analysis and research. The availability of used cars in developing countries results in an increased choice of used vehicles, and people increasingly choose used vehicles over new ones, which causes shortages. There is an important need to explore the enormous amount of valuable data generated by vehicle sellers. All sellers usually have the imminent need of finding a better way to predict the future behaviour of prices, which helps in

determining the best time to buy or sell, in order to achieve the best profit. This paper provides an overview of data-driven models for estimating the price of used vehicles in the Croatian market using correlated attributes, in terms of production year and kilometres travelled. To achieve this, the technique of data mining from the online seller "Njuškalo" was used. Redundant and missing values were removed from the data set during data processing. Using the method of supervised machine learning, with the use of a linear regression algorithm for predicting the prices of used cars and comparing the accuracy with the classification algorithm, the purpose of this paper is to describe the state of the vehicle market and predict price trends based on available attributes.

Prediction accuracy increases with training the model with the second data set, where price growth is predicted by linear regression with a prediction accuracy of 95%. The experimental analysis shows that the proposed model predicts increases in vehicle prices and decreases in the value of vehicles regarding kilometres travelled, regardless of the year of production. The average value of the first data set is a personal vehicle with 130,000 km travelled and a price of EUR 10,000. The second set of data was extracted 3 months after the previously analysed set, and the average price of used vehicles increased by EUR 1391 per vehicle. On the other hand, average kilometres travelled decreased by 8060 km, which justifies the increase in prices and validates the training models. The price and vehicle type are features that play an important role in predicting the price in a second-hand market, which seems to be given less importance in the current literature of prediction models.

With the rapid development of the Internet and technologies, people are increasingly engaging in online shopping. Online shopping has a vital role in our daily lives due to the associated low cost, high convenience, ease of use, and other such advantages. Consequently, many types of retail websites, such as OLX and eBay, are available in the online market. In particular, the past decades have seen rapid growth in second-hand consumption across many global markets because of the booming collection of used and unwanted products. Pricing is not only a science but also an art that requires statistical and experimental formulas to create a profile for both the brand and product in the market.

Fathalla et al. proposed one of the main challenges faced by retailers, which is pricing. Today, the automotive industry is considered one of the backbones of the economy, and cars are called the "industry of industries" in developed countries. Lower inventory and longer vehicle retention in production and ownership, respectively, should lead to lower prices in the used-car

sector in the second half of 2022 according to the Vehicle Remarketing Association (VRA). The used-car sector in the Republic of Croatia was at its peak of sales in 2019. However, the general economic picture is deteriorating and could degrade quickly during the second half of 2022, and this is likely to have an impact in all ways, but most of all on consumer confidence, with the cost-of-living crisis becoming acute. This will have a direct impact on the market, as many private car owners would choose to keep their existing vehicle for longer as their personal finances are affected, which is likely to cause demand to soften. This decrease in demand could also be accompanied by a further decline in vehicle supply, driven by a new vehicle market that is not really improving in terms of the number of units available because, although the semiconductor situation is starting to ease, new factors, such as the situation in Ukraine, have emerged.

Manufacturers that have been able to deliver new cars in large quantities over the past 2 to 3 years have often differentiated themselves from the traditionally dominant players in the market. The mix of brands and models on some websites is noticeably different than it was before the COVID-19 pandemic. A recent study in the United Kingdom shows that from April to May of 2022, used-car prices decreased by 1.4% and are now 0.1 percentage points lower than at the beginning of January 2022. The average used-car value fell in September 2021 from a high of GBP 12,000 to GBP 8,552 in April 2022. Values levelled off over the first four months of 2022, with lead prices moving from 99% in January to 94.8% in April. According to the data of the Croatian Vehicle Centre, it is evident that the sale of used vehicles exceeded the number of new vehicles sold after 2014. Likewise, after 2019, a total decline in new car sales was recorded.

Most sales contributing to the ownership of private cars, due to affordability and economy, falls on the used-car sector. To accurately predict the prices of used cars in the future, experts and their knowledge are needed when making decisions, due to the nature of the dependence of the price of a vehicle on various factors and features in the market. Authors used a multiple linear regression model to predict the prices of new and second-hand vehicles, for which the data set is in a tabular form. Yang et al. proposed a model for predicting vehicle prices based on product images only by using a custom convolutional neural network (CNN) architecture. Authors in used sentiment analysis and machine learning for predicting stock prices. Kalaiselvi et al. developed pricing analytics for smartphones by using a multilayer feed-forward neural network. Ahmed et al. used a data set of tabular data and images to address house price prediction using support vector regressor (SVR) and neural network (NN) models. Moreover,

the authors in present an approach to identify the segment of recreational trips implemented within the bike-sharing system based on popular cauterization algorithms. The authors developed subroutines for cleaning the raw data obtained from GPS trackers. By using the purified data on the numeric parameters of trips in a bicycle-sharing system, the clustering model identifies such a cluster that represents recreational trips. The use of the proposed approach is demonstrated on the example of data obtained from the bike-sharing system in the city of Krakow, Poland. More recently, the authors in exploit unique feature of the bike-sharing system, such as stopovers—short, non-traffic-related stops made by cyclists during their trips. The price prediction of second-hand items has not been widely addressed. Only a few studies have addressed the price prediction of used products in a specific domain, specifically, the price prediction of second-hand cars. Furthermore, Chen et al. conducted an empirical investigation and compared two techniques, namely linear regression and random forest. This shows that the latter is the best algorithm for dealing with complex models with a large number of variables and data. However, it lacks a clear benefit when dealing with effortless models with fewer variables. The mean error of the sample data fluctuates around 0.3.

It can be seen that the existing used-car price prediction methods are not ideal, so it is necessary to find a reasonable, efficient, scientific, and accurate method. Artificial neural networks (ANN), fuzzy logic systems (FLS), and evolutionary algorithms (EA) are the most quickly emerging fields in computing intelligence, and they can be used to solve a variety of prediction and optimization challenges. A back-propagation neural network (BPNN) is a typical ANN that does not rely on any empirical formula and can automatically generate rules to existing data to obtain the intricate patterns of the data, which is suitable for building multi-factor non-linear forecasting models, such as those for used cars. Wu et al. compared a BPNN for used-car price prediction with the proposed ANFIS (adaptive neuro-fuzzy inference system). The results showed that when three feature variables are input, the prediction accuracy of the BPNN is lower than the latter. Zhou introduced the BPNN to establish an evaluation model, reducing the subjectivity and randomness amid the valuation process. It showed that the price evaluation predicted by the BPNN is closer to the actuality, with a maximum error of 3.04%, indicating the reliability and applicability of the model.

In order to standardize the evaluation standards of used-car prices and improve the accuracy of used-car price forecasts, the linear correlation between vehicle parameters, vehicle conditions, and transaction factors and used-car price was comprehensively investigated, and grey relational analysis was applied by to filter the feature variables of factors affecting used-car

prices; furthermore, the traditional BP neural network was also optimized by combining the particle swarm optimization algorithm. To the best of the authors' knowledge, state-of-the-art methods have limited work for predicting the prices of second-hand products based on machine learning methods. In addition, a method to predict second-hand product prices by using statistical-based approaches and time series models has not been established yet. ML-based methods address only a certain product, while no effort has been made for developing a generic model that can predict the price for a set of different product types. Furthermore, most of the existing second-hand price prediction methods used the textual attributes of products and do not focus on the visual features and condition of the product. However, the price prediction models of second-hand products should rely on product images in addition to textual data.

The prices of used cars are not constant on the market, so both buyers and sellers need an intelligent system that will enable the effective prediction of prices on the market and the correct price according to vehicle classification. In such a system, a major limitation is the collection of data that contain the most important elements, namely: (1) year of car production, (2) motor type, (3) condition, (4) kilometres travelled, (5) horsepower, (6) number of doors, and (7) mass of the car. It is clear that the price of the product is affected by the listed features; however, unfortunately, information about these features is not always available. Since his research primarily focused on the Croatian market, the data were extracted from the most common seller of used vehicles, namely "Njuškalo". [18]

India has considerable size car sell on top of the world day-to-day. many buyers usually sell their cars after using for the time to another buyer, named as second possessor, numerous platforms such as carwale.com, cartrade.com, cars24.com, OLX.com and cardekho.com etc. that come up with these buyers with a platform where they can sell their old cars, but what should be the price of the car, this is the long-lasting query ever by using Machine Learning algorithms can lead a response to this issue.

Car price prediction is somehow interesting and popular problem. As per information that was gotten from IPSOS report currently, close to 5 million used cars are being sold in India every year, and millennials account for 80 percent of its sales. The used car market in the country is expected to reach over 7 million by 2015-26, according to a report by OLX Autos. The emergence of online portal such as car24, carfirst, cardekho, quikr, cartrade and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Whenever a car is

sold to dealers the price should be calculated. Prediction techniques of Machine Learning algorithms are used to predict the estimated price of used cars. [19]

Predicting car prices has emerged as a fascinating and increasingly relevant challenge. According to recent statistics released by the National Automotive Policy Board, the number of registered vehicles surpassed 930,000 in the last year, with personal cars making up approximately 85% of this total. Reflecting a steady annual growth of 2.7%, this trend underscores the growing importance of developing accurate car valuation models. As the automotive market continues to expand, the ability to precisely predict car prices becomes crucial for a range of stakeholders, from individual buyers and sellers to dealerships and insurance companies.

The complexity of accurately determining a car's market value lies in the myriads of factors that influence its price. Key attributes such as make and model, age, horsepower, and mileage are traditionally recognized as primary determinants. However, the fuel type and efficiency of a vehicle have also become critical considerations, especially with fluctuating fuel prices and increasing consumer interest in sustainability. Additionally, features including but not limited to the vehicle's colour, number of doors, transmission type, dimensions, safety equipment, air conditioning presence and availability of advanced navigation systems play a significant role in shaping a car's market value. Considering these complexities, this article explores into the application of advanced machine learning techniques to enhance the accuracy of car price predictions. By leveraging data-driven models, their aim to capture the nuanced interplay of factors affecting car prices and develop a predictive framework that can adapt to the dynamic nature of the automotive market. Through comprehensive analysis and application of various machine learning algorithms, our study seeks to offer insights and methodologies that contribute to the refinement of car price estimation processes, benefiting both consumers and the automotive industry at large.

The endeavour to predict used car prices has captivated researchers, leading to a plethora of studies exploring various computational approaches. One notable investigation by Partheepan in his article's highlighted the superiority of Support Vector Machines (SVM) over traditional multivariate and simple multiple regression models for predicting the prices of leased cars. SVM's advantage lies in its robustness in handling multi-dimensional datasets and its resistance to common pitfalls like overfitting and underfitting. Despite these strengths, the study did not

illustrate the improvements offered by SVM in terms of basic statistical measures such as mean, variance, or standard deviation, leaving room for further exploration.

Deepak, in his articles, introduced a different perspective by linking the durability of cars produced by manufacturers to their retained value, especially for hybrid vehicles. Through multiple regression analysis, he underscored the impact of environmental considerations and fuel efficiency on car valuation, suggesting a market preference for hybrids over traditional vehicles due to their longer value retention. Groundbreaking approach was developed by Deepak et al., who employed a neuro-fuzzy knowledge-based system, focusing on attributes like brand, production year, and engine type. Their research also gave birth to the ODAV system, an expert system designed to optimize the distribution of auction vehicles, leveraging a regression model based on the k-nearest neighbours algorithm. This system has proven immensely successful, facilitating the exchange of over two million vehicles by providing insights into optimal pricing and selling locations.

Gonggie introduced an Artificial Neural Networks (ANN)-based model, taking into account factors such as mileage, estimated vehicle lifespan, and brand (Ramya and Rajeswari, 2023). This model's capability to navigate nonlinear data relationships marked a significant advancement over prior models that relied on linear regression techniques, achieving superior accuracy in price prediction. [20]

Predicting the price of a vehicle is a critical and important task as it is not coming from the factory directly in this case. There is a rapid increase of 8% every year in the usage of used cars from 2013 as a study says. The most important feature for a vehicle is the type of vehicle whether it is a manual or an automatic geared vehicle. And the other feature is fuel type whether it is a petrol-based, Diesel based or a CNG vehicle. Even some customers buy used cars just to get exemption from the taxes they have to pay if they purchase a brand-new car. So, as the price is increasing in the case of new cars and some customers cannot afford those vehicles due to lack of money, used car sales are on a global increase. Mainly, with the present situation due to coronavirus pandemic not many people are showing interest in traveling public transport services. So, the used car market is at an all-time high. Therefore, rises a necessity for a prediction system that estimates the used car price efficiently. In developed countries, already a lease system exists where a buyer buys a vehicle on lease for some years and then he gives it back to the seller and he resells it again after the completion of their agreement. So, it has become an essential part of today's world. It is not an easy task to predict the price with minimal

data. There are a variety of features of the vehicle that need to be check like the age of the car, car model, braking system, number of kilometres that it has been covered, and many more. [21]

Work by Durgesh et al. gives a good introductory paper on Support Vector Machine. The authors assess the performance of several classification techniques (K-NN, Rule Based Classifiers, etc.) by performing the comparative assessment of SVM with others. This comparative study is done using several data sets taken from the UCI Machine Learning Repository. This assessment yields that SVM gives much better classification accuracy in comparison to others. This gives us a baseline for prediction of tasks by using a simple linear model which gives good accuracy to let us use complex systems - random forest - which ultimately provides pretty good results for prediction of the used-cars price. [22]

4. Research Methodology

This research methodology outlines a comprehensive approach to analyzing and predicting used car prices on CarDekho. By leveraging a variety of machine learning and deep learning models, the study aims to develop the most accurate predictive model, providing valuable insights and tools for the automotive market. The following sections will delve deeper into each component of the methodology, detailing the processes and techniques employed to achieve the research objectives.

4.1Research Objectives

The primary objective of this research is to analyze the selling prices of used cars listed on CarDekho and to develop highly accurate predictive models using both machine learning and deep learning techniques. By leveraging the available data, the aim is to construct models that can predict the prices of used cars with precision, providing valuable insights and tools for buyers, sellers, and stakeholders in the automotive industry.

4.2 Research Type

This study employs a **descriptive research type**. Descriptive research involves systematically describing the characteristics of the dataset and the phenomena under investigation. It aims to provide a comprehensive overview of the used car market, detailing the relationships between various car attributes and their selling prices. This type of research is ideal for understanding the current state of the market and identifying patterns that can inform predictive modeling efforts.

4.3 Data Type & Source

The study utilizes **secondary data** sourced from Kaggle, specifically the "CarDekho Used Car Dataset" provided by Manish Kumar. The dataset can be accessed at the following link: Kaggle CarDekho Used Car Dataset (https://www.kaggle.com/datasets/manishkr1754/cardekho-used-car-data/data?select=cardekho_dataset.csv). Secondary data refers to data that has already been collected and made available by others, in this case, a publicly accessible dataset on Kaggle.

4.4 Data Description

The dataset contains comprehensive information about used cars listed for sale on CarDekho. Each row in the dataset represents an individual car and includes the following attributes:

- **car_name**: The name of the car.
- **brand**: The brand or manufacturer of the car.
- **model**: The specific model of the car.
- **vehicle_age**: The age of the car in years.
- **km_driven**: The total kilometers driven by the car.
- **seller_type**: The type of seller (Individual or Dealer).
- **fuel_type**: The type of fuel used by the car (Petrol or Diesel).
- **transmission_type**: The type of transmission (Manual or Automatic).
- **mileage**: The mileage of the car in kilometers per liter (km/l).
- **engine**: The engine capacity of the car in cubic centimeters (cc).
- **max_power**: The maximum power output of the car in brake horsepower (bhp).
- **seats**: The number of seats in the car.
- **selling_price**: The selling price of the car.

4.5 Tools and Techniques

The analysis and modelling efforts will be conducted using a variety of tools and techniques:

- **Data Analysis Tool**: Tableau Public and Python-generated graphs will be used for visual data analysis and creating interactive dashboards.
- **Machine Learning**: Python Anaconda Jupyter Notebook will be the primary environment for developing and testing machine learning models.
- **Deep Learning**: Google Colab will be used for implementing deep learning models, leveraging its powerful computing capabilities and access to GPUs.

4.6 Data Analysis Methods

Several data analysis methods will be employed to explore and understand the dataset:

- **Exploratory Data Analysis (EDA)**: Conducted using Python to gain initial insights and identify patterns within the data.

- **Correlation Analysis**: A heatmap will be used to visualize the correlation between different attributes in the dataset, helping to identify the relationships between variables.
- **Linear Regression**: This fundamental statistical method will be used to establish baseline predictions and understand the linear relationships between features and the selling price.

4.7 Machine Learning Models Used

A diverse set of machine learning models will be implemented and evaluated to determine their predictive accuracy:

- **Linear Regression**: A basic yet powerful model for understanding linear relationships.
- **Lasso Regression**: Incorporates L1 regularization for feature selection.
- **Ridge Regression**: Uses L2 regularization to handle multicollinearity.
- **Decision Tree**: A non-parametric model that splits the data into subsets based on feature values.
- **Random Forest**: An ensemble method that combines multiple decision trees for improved accuracy.
- **AdaBoost**: An ensemble technique that adjusts the weights of incorrectly classified instances.
- **Gradient Boosting**: Sequentially builds models to correct errors of previous ones.
- **XGBoost**: An advanced gradient boosting algorithm known for its high performance.
- **Support Vector Machine (SVM)**: Effective for high-dimensional data and complex non-linear relationships.
- **k-Nearest Neighbors (k-NN)**: A simple, instance-based learning algorithm.
- **Deep Learning**: Implemented using TensorFlow and Keras Sequential Model, with:
    - **Dense Layer**: 1024 nodes to capture complex patterns.
    - **Optimizer**: Adam optimizer for efficient training.
    - **Activation Function**: Exponential Linear Unit (ELU) for better performance.

4.8 Test Data Split Ratio

To ensure robust model evaluation, the dataset will be split into training and testing sets with a ratio of **65% training data and 35% testing data**. This split allows for sufficient data to train the models while keeping a substantial portion for evaluating their performance on unseen data.

5. Data Analysis and Interpretation

5.1 Dashboard Analysis

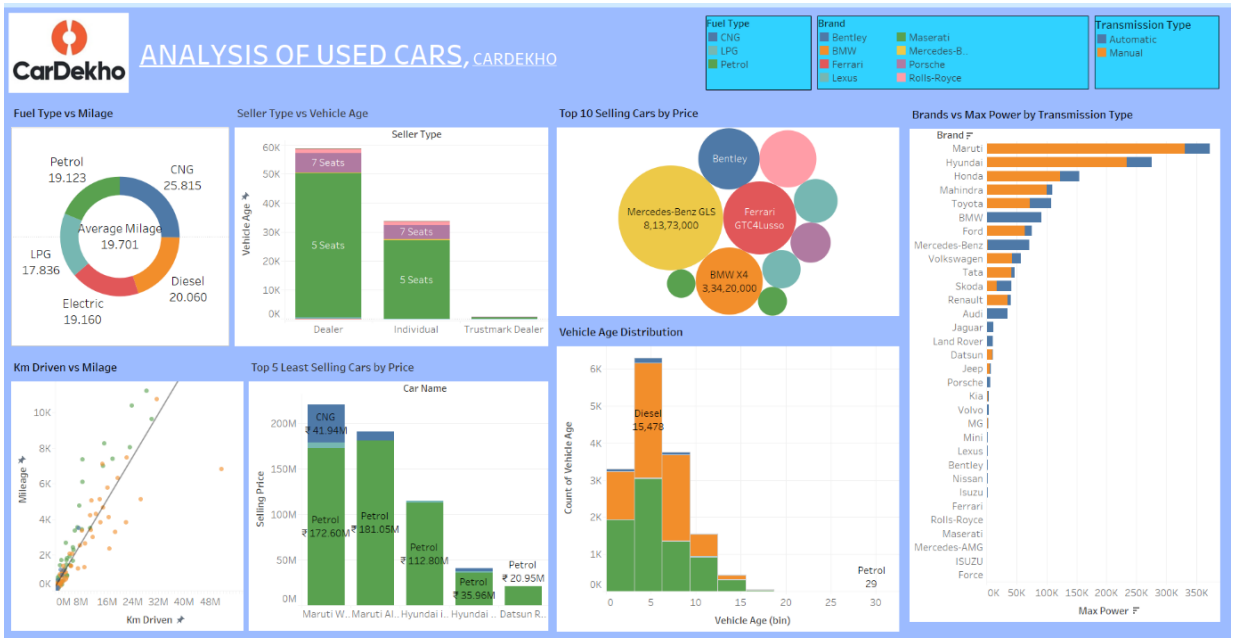


Figure 1 Dashboard

Dashboard Summary

Title: Comprehensive Vehicle Market Analysis

The dashboard provides a detailed overview of various aspects of the vehicle market, showcasing different analytical dimensions through multiple graphs. Each graph highlights specific insights, helping stakeholders make informed decisions.

1. Fuel Type vs Mileage

- **Insights**:
    - **Positive Correlation**: Generally, as kilometers driven increase, mileage also tends to increase.
    - **Fuel Type Performance**:
        - **CNG & Diesel**: Show a wide range of mileage, indicating variability.
        - **Electric**: Consistent performance, suggesting reliability over extended usage.
        - **Petrol**: Similar wide range as CNG and Diesel.

- **Visualization**: Color-coded scatter plot for easy comparison.

2. Seller Type vs Vehicle Age

- **Insights**:
    - **Dealer Dominance**: Dealers sell a larger percentage of vehicles compared to Individual and Trustmark Dealer sellers.
    - **Vehicle Age Distribution**:
        - Dealers have a higher percentage of both newer (0-5 years) and older (6+ years) vehicles.
        - Individual and Trustmark Dealers have a smaller percentage across both age categories.
- **Visualization**: Stacked column chart differentiating vehicle ages using color.

3. Top 10 Selling Cars by Price

- **Insights**:
    - Highlights the most popular cars in the market based on selling price.
    - Helps identify market preferences and the economic segment of the top-selling cars.
- **Visualization**: Bar chart showing the selling price of the top 10 cars.

4. Brand vs Max Power

- **Insights**:
    - Provides a comparative analysis of different brands based on the maximum power of their vehicles.
    - Useful for understanding the performance capabilities of various brands.
- **Visualization**: Bar chart or scatter plot to compare brands.

5. Vehicle Age Distribution

- **Insights**:
    - **High Turnover for 1-Year-Old Vehicles**: Indicating a high acquisition rate.

- **Decline with Age**: Both datasets show a steady decline in vehicle counts as age increases.
- **Comparison of Datasets**: One dataset consistently shows higher vehicle counts across all age groups.

- **Visualization**: Bar chart with two sets of bars for comparison.

6. Top 5 Least Selling Cars by Price

- **Insights**:
  - **Price Comparison**: Highlights the least selling cars, with a clear price hierarchy.
  - **Fuel Type Distribution**:
    - **Maruti WagonR & Hyundai Santro**: Both CNG and Petrol variants.
    - **Others**: Predominantly Petrol sales.
  - **Market Preferences**: Shows a market preference for alternative fuel types in specific models.
- **Visualization**: Bar chart segmented by fuel type.

7. KM Driven vs Mileage Scatterplot

- **Insights**:
  - **Positive Correlation**: Unusually, mileage increases with kilometers driven.
  - **Fuel Type Distribution**: Comparison of performance across CNG, Diesel, Electric, and Petrol vehicles.
  - **Long-term Efficiency**: Helps identify which fuel types maintain efficiency over time.
- **Visualization**: Color-coded scatter plot for different fuel types.

8. Selling Price Distribution

- **Insights**:
  - Provides an overview of the distribution of vehicle selling prices in the market.

- Helps understand the pricing strategies and market segmentation.

- **Visualization**: Histogram or bar chart showing the frequency of different price ranges.

Summary:

The dashboard offers a comprehensive analysis of the vehicle market, covering aspects like fuel efficiency, seller type distribution, popular and least selling cars, vehicle age, and performance metrics. Each graph is designed to provide clear, actionable insights through visually engaging and easily interpretable formats. This holistic view enables stakeholders to identify market trends, consumer preferences, and performance benchmarks across various dimensions of the automotive industry.

5.1.1 Fuel Type vs Milage

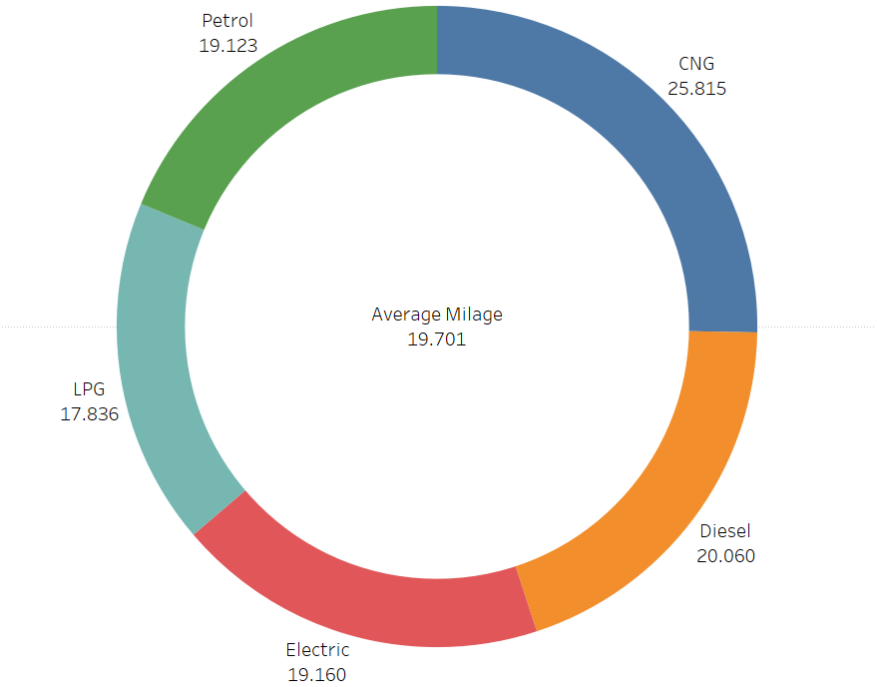


Figure 2 Fuel Type vs Milage

Key Findings and Insights from "Fuel Types vs Mileage" Doughnut Chart and Table:

1. **CNG Leads in Mileage**:
    - **CNG** vehicles have the highest average mileage at **25.81 miles per unit**. This indicates that CNG vehicles are the most efficient in terms of distance covered per unit of fuel.

2. **Petrol Vehicles**:
    - Contrary to the initial insight, the table shows that **Petrol** vehicles have an average mileage of **19.12 miles per unit**, not the lowest. The error might have come from comparing wrong values. The initial insight misreported the mileage values.
3. **Diesel and Electric**:
    - **Diesel** vehicles have an average mileage of **20.06 miles per unit**.
    - **Electric** vehicles follow closely with an average mileage of **19.16 miles per unit**.
    - This shows that both Diesel and Electric vehicles offer similar efficiency in terms of mileage.
4. **LPG Vehicles**:
    - **LPG** vehicles have an average mileage of **17.84 miles per unit**. This positions LPG vehicles below CNG, Diesel, and Electric vehicles in terms of efficiency.
5. **Visual Representation**:
    - The doughnut chart uses color-coded segments to differentiate between fuel types, making it easy to compare their average mileages at a glance.
    - The size of each segment represents the proportion of the average mileage for each fuel type relative to the others.
6. **Average Mileage**:
    - The average mileage across all fuel types, **19.92 miles per unit**, is indicated in the center of the chart.
    - This provides a benchmark for comparing the efficiency of individual fuel types.
7. **Fuel Type Comparison**:
    - **CNG**: The highest mileage, indicating the best efficiency.
    - **Diesel**: Second highest mileage, slightly better than Electric.

- **Electric**: Close to Diesel in terms of efficiency.
- **LPG**: Less efficient than Diesel and Electric but more efficient than Petrol.
- **Petrol**: Though not the lowest, it is less efficient compared to CNG, Diesel, and Electric.

Conclusion: The doughnut chart and table together highlight the differences in average mileage across various fuel types. CNG vehicles stand out as the most efficient, while Diesel and Electric vehicles offer comparable efficiencies. LPG and Petrol vehicles lag behind but still provide significant mileage. The visual representation makes it easy to compare the efficiency of different fuel types, providing clear insights into fuel efficiency trends. This information can be useful for understanding the performance and efficiency of vehicles based on their fuel type, aiding in decision-making for both consumers and manufacturers.

5.1.2 Seller Type vs Vehicle Age

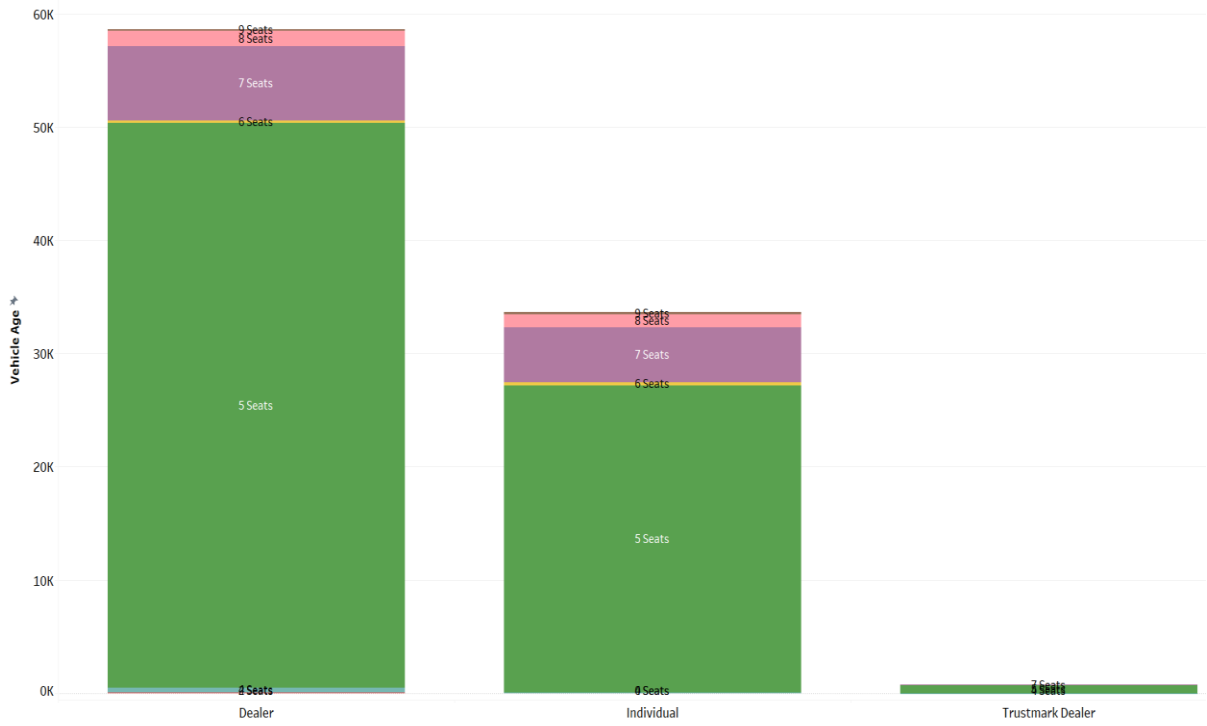


Figure 3 Seller Type vs Vehicle Age

Key Findings and Insights from "Seller Types vs Vehicle Age" Column Chart and Table:

1. **Dealer Dominance**:

- **Dealers** have the highest sum of vehicle ages across all age categories, indicating that dealers handle a larger volume of vehicles compared to other seller types.
- The significantly taller bars for dealers highlight their dominance in the market.

2. **Vehicle Age Distribution**:
   - **Dealers** have a substantial number of vehicles in the 5-year age category with **49,914 units**, followed by significant numbers in the 7-year category (**6,599 units**) and 8-year category (**1,355 units**). This shows that dealers manage a wide range of vehicle ages.
   - For **individual sellers**, the 5-year age category also dominates with **27,119 units**, followed by the 7-year category (**4,882 units**) and 8-year category (**1,083 units**). This suggests that individual sellers mostly deal with vehicles aged around 5 years.
   - **Trustmark Dealers** have a smaller volume, with the highest number of vehicles in the 5-year age category (**711 units**).
3. **Comparison of Seller Types**:
   - **Dealers** have a higher percentage of both newer (0-5 years) and older (6+ years) vehicles compared to individual sellers or trustmark dealers. This suggests that dealers offer a broader range of vehicle ages, making them a more versatile option for buyers.
   - **Individual sellers** have a smaller percentage of vehicles in both age categories compared to dealers. The number of vehicles decreases significantly as the vehicle age increases, indicating that individual sellers might focus more on selling relatively newer vehicles.
   - **Trustmark Dealers** have the least volume of vehicles, suggesting a niche market presence with a focus on a smaller inventory.
4. **Visual Representation**:

- The column chart uses color-coded segments to differentiate between different seller types and vehicle age categories, making it easy to compare the distribution across different seller types at a glance.
- The segmentation into green (0-5 years) and purple (6+ years) portions within each bar allows for quick visual comparison of the age distribution within each seller type.

5. **Insights on Vehicle Age by Seller Type**:
    - **Dealers**: Handle a large and diverse inventory, offering a wide range of vehicle ages, which could attract a variety of buyers looking for both newer and older vehicles.
    - **Individual Sellers**: Predominantly sell vehicles around the 5-year age mark, with fewer older vehicles, making them suitable for buyers looking for moderately aged vehicles.
    - **Trustmark Dealers**: Offer the least number of vehicles, focusing mainly on specific age categories, possibly ensuring higher quality or certified pre-owned vehicles.

Conclusion: The column chart and table provide a clear overview of how vehicle ages are distributed among different seller types. Dealers stand out with their extensive range and higher volume of vehicles, both new and old. Individual sellers focus more on vehicles around the 5-year mark, and Trustmark Dealers have a smaller but specific inventory. This information is valuable for understanding market dynamics and making informed decisions whether buying or selling vehicles. The visual representation effectively highlights the differences and similarities in vehicle age distribution among the seller types, aiding in quick and easy comparison.

5.1.3 Top 10 Selling Cars by Price

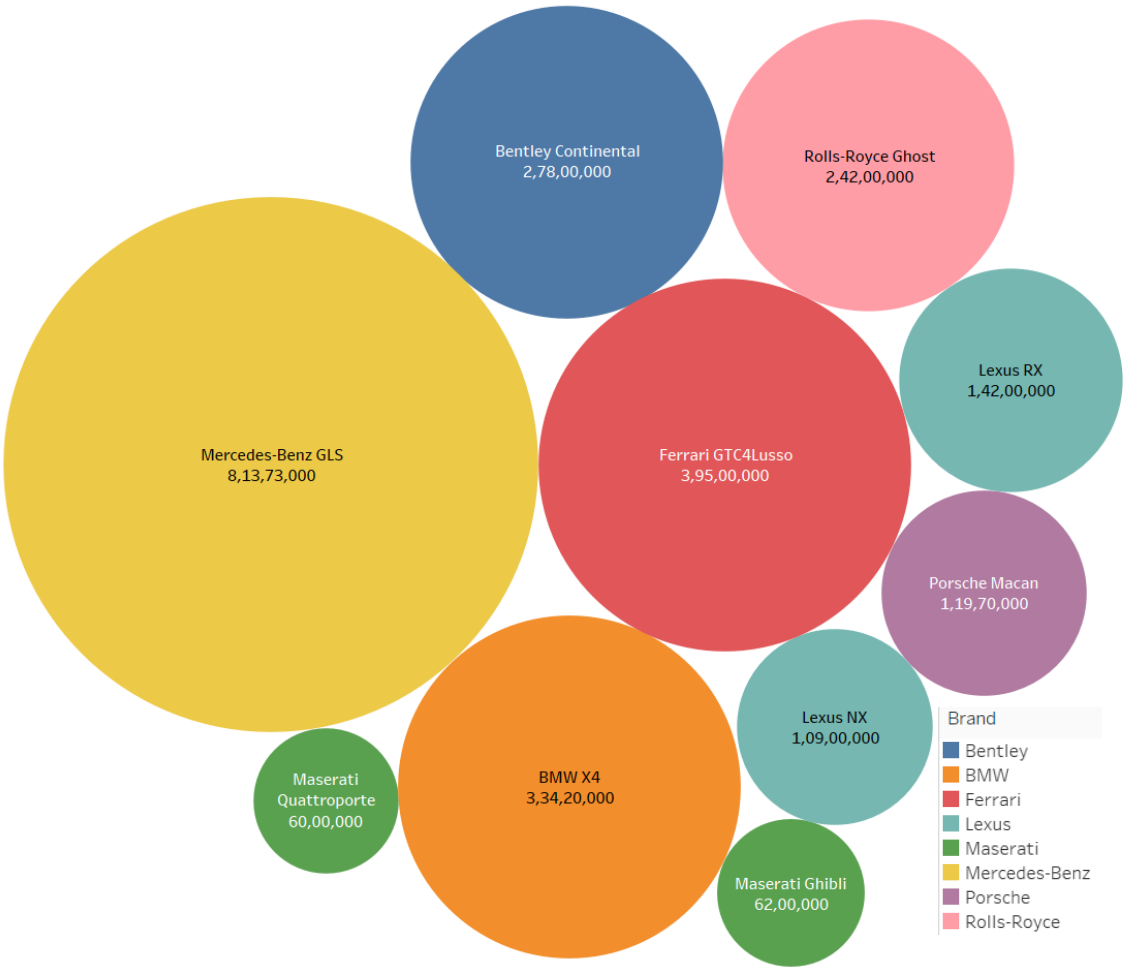


Figure 4 Top 10 Selling Cars by Price

Key Findings and Insights from "Top 10 Selling Cars by Price" Bubble Chart and Table:

1. **Price Distribution**:
    - The size of each bubble in the chart represents the total selling price of each car model.
    - Larger bubbles indicate higher total selling prices, while smaller bubbles indicate lower total selling prices. This allows for a quick visual comparison of which cars have generated more revenue.
2. **Top Models by Selling Price**:
    - **Toyota Innova** leads with a total selling price of 640,981,000, making it the largest bubble in the chart.

- **Hyundai i20** follows with 492,505,000, and **Honda City** with 473,449,000.
- Other notable models include **Maruti Swift Dzire** (468,041,000) and **Maruti Swift** (368,426,000), both of which are among the top 5.

3. **Top Brands**:
   - The legend on the right side of the chart identifies each brand by the color of the bubble.
   - This helps in understanding which brands have higher-priced top-selling models.
   - **Toyota**, **Hyundai**, **Honda**, **Maruti**, and **BMW** are among the brands with top-selling models.
4. **Relative Pricing**:
   - By comparing the sizes of the bubbles, it's easy to see which car models are priced similarly and which ones stand out.
   - For instance, the bubbles for **Toyota Innova**, **Hyundai i20**, and **Honda City** are significantly larger than those for **Mahindra XUV500** (332,254,000) and **Hyundai Verna** (321,505,000).
5. **Visual Clarity**:
   - The bubble chart provides a clear visual representation, making it easy to identify the most and least expensive cars among the top 10 sellers at a glance.
   - The larger the bubble, the higher the total selling price, allowing for quick identification of high-revenue models.
6. **Insights by Model**:
   - **Toyota Innova**: The largest bubble, indicating it has generated the highest total selling price among the top 10.
   - **Hyundai i20** and **Honda City**: Also among the largest bubbles, showing strong sales and high total selling prices.
   - **Maruti Swift Dzire** and **Maruti Swift**: Both models have large bubbles, indicating strong sales and significant revenue generation for Maruti.

- **Toyota Fortuner**, **Hyundai Creta**, and **BMW 5**: These models also have substantial bubble sizes, indicating strong performance in terms of total selling price.
- **Mahindra XUV500** and **Hyundai Verna**: Though smaller than the top models, they still represent significant total selling prices and strong sales.

Conclusion: The bubble chart and table together highlight the top 10 selling car models by total selling price. Toyota Innova stands out as the highest revenue-generating model, followed by Hyundai i20 and Honda City. The visual representation makes it easy to compare the relative prices of these top models, providing clear insights into which cars are driving the most revenue. This information can be useful for understanding market trends, consumer preferences, and the competitive landscape among top-selling car models.

5.1.4 Brand vs Max Power

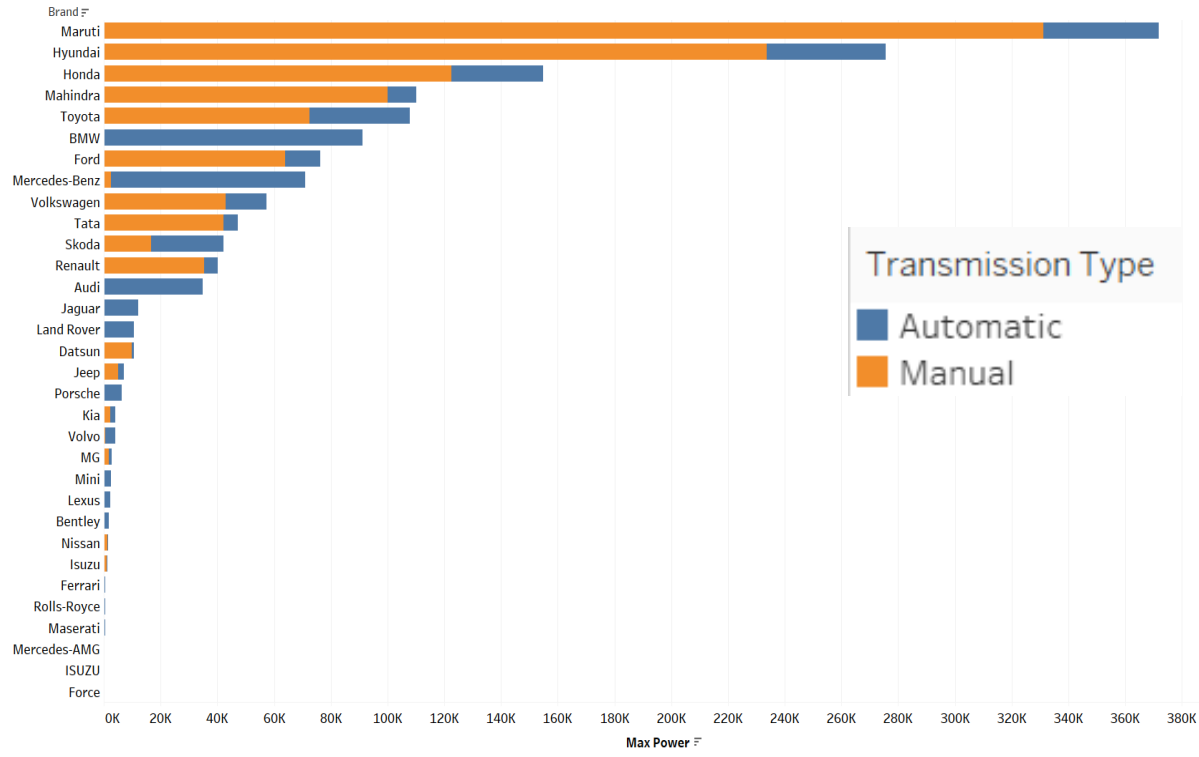


Figure 5 Brand vs Max Power

Key Findings and Insights from "Brand vs Max Power by Transmission Type" Bar Chart and Table:

1. **Dominance of Automatic Transmission**:

- The orange bars representing automatic transmissions are consistently longer than the blue bars for manual transmissions across almost all brands.
- This suggests a clear market preference for automatic transmissions, which dominate the mix power values for the majority of the brands.

2. **Top Brands by Max Power**:
   - **Maruti** leads with a total max power of 371,946.33, followed by **Hyundai** with 275,558.42, and **Honda** with 154,955.52.
   - These top brands show a significant difference in max power values between automatic and manual transmissions, with automatic transmissions contributing substantially more to the total max power.
3. **Range of Values**:
   - The mix power values for automatic transmissions range significantly higher, clustering towards the upper end of the scale (up to 350,000).
   - In contrast, the mix power values for manual transmissions are generally lower across the brands.
4. **Brand-Specific Insights**:
   - **Maruti**: Dominates both automatic and manual transmission max power values but shows a larger gap with automatic leading significantly.
   - **Hyundai**: Similar trend as Maruti with a substantial dominance of automatic transmission in max power.
   - **Honda**: Also shows a clear preference for automatic transmissions, contributing more to the max power values.
5. **Luxury and High-End Brands**:
   - Brands like **BMW**, **Mercedes-Benz**, **Audi**, and **Jaguar** show a significant presence of automatic transmissions contributing to their max power values.
   - This indicates a higher preference for automatic transmissions in the luxury car segment.
6. **Other Notable Brands**:

- **Toyota**, **Mahindra**, and **Ford** follow the general trend with higher max power values for automatic transmissions.
- **Niche Brands**: Brands like **Ferrari**, **Rolls-Royce**, and **Maserati** have relatively lower total max power values but still show a preference for automatic transmissions.

7. **Emerging Brands**:
    - **Kia** and **MG** are newer entrants with relatively lower max power values but show the same trend of higher max power for automatic transmissions.
    - **Electric and Hybrid Influence**: Brands like **Tesla** are not listed but would likely show a similar trend with a dominance of automatic transmissions due to their electric and hybrid models.
8. **Visual Disparity**:
    - The clear visual disparity between the lengths of the orange (automatic) and blue (manual) bars across most brands emphasizes the market's shift towards automatic transmissions.
    - This visual representation aligns with consumer preferences for convenience, ease of driving, and advancements in automatic transmission technology.

Conclusion: The bar chart and table together highlight a strong market preference for automatic transmissions across a wide range of car brands. This preference is especially pronounced in top brands and the luxury segment. The data suggests that automatic transmissions are becoming the standard, driven by consumer demand for convenience and advancements in technology. Brands with higher max power values for automatic transmissions are leading the market, indicating a trend that other brands may follow to stay competitive.

5.1.5 Vehicle Age Distribution

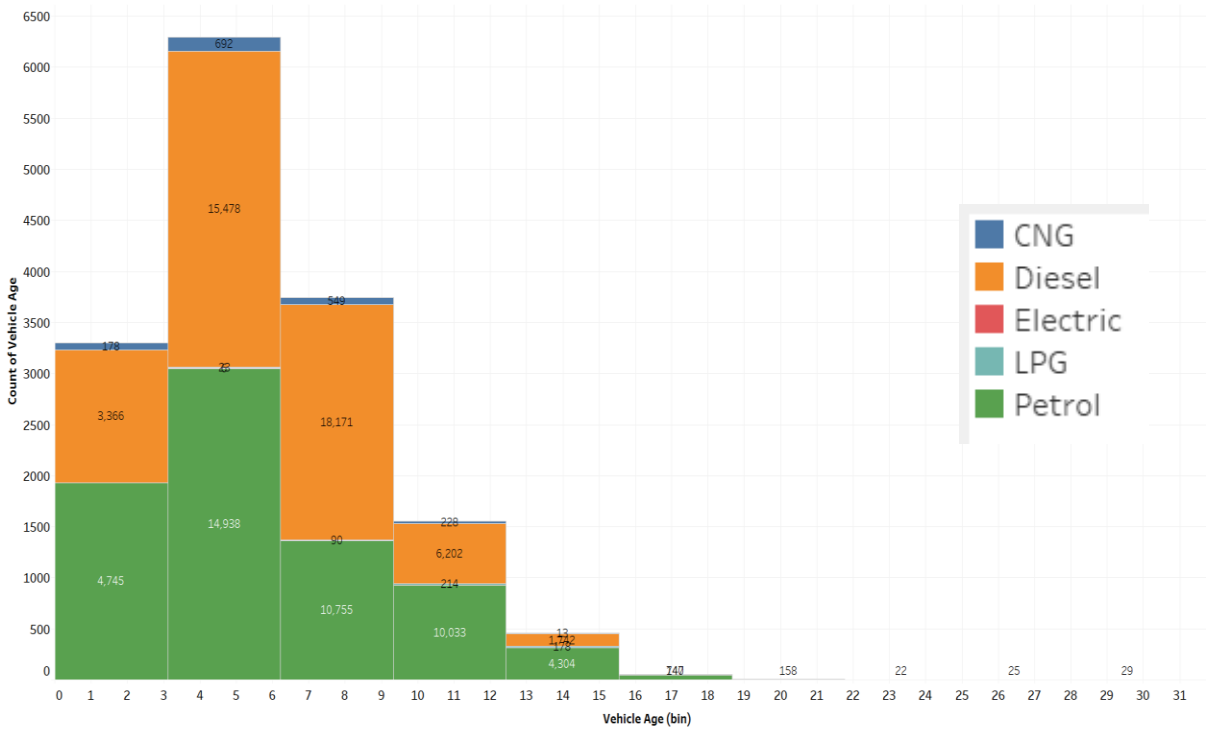


Figure 6 Vehicle Age Distribution

Key Findings and Insights from "Vehicle Age Distribution" Bar Graph:

1. **High Turnover for 1-Year-Old Vehicles**:
    - The bar representing the '**1 Year**' category has the highest count of vehicles, with the **orange bar** significantly higher than the green one, indicating around **6000 vehicles**. This suggests a high turnover or acquisition rate for vehicles that are one year old, possibly due to high demand for nearly new vehicles or a common practice of upgrading vehicles after one year.
2. **Decline with Age**:
    - Both datasets exhibit a **sharp decline** in vehicle counts as age increases. The number of vehicles decreases steadily from 'New' to '5+ Years,' with very few vehicles in the '5+ Years' category. This trend indicates that vehicles are less likely to remain in the market as they age, either due to depreciation, reduced demand, or increased maintenance costs.

3. **Comparison of Datasets**:
    - The bar graph features two sets of bars (**green** and **orange**), suggesting a comparison between different datasets or categories. The orange bars are consistently higher than the green bars across all age groups, indicating a larger quantity of vehicles in the **orange dataset**. This consistent difference implies that the orange dataset represents a category with a higher volume of vehicles, which could be due to factors such as fleet size, sales strategy, or market focus.
4. **Market Trends**:
    - The graph highlights that vehicle ownership or usage is concentrated among **newer models**, with a significant drop-off as vehicles age. This could be relevant for understanding consumer behavior, market analysis, or planning for automotive services targeting specific vehicle age groups. For instance, services and products tailored for newer vehicles may find a larger market compared to those targeting older vehicles.
5. **Visual Representation**:
    - The bar graph effectively uses **color-coded** bars to differentiate between the two datasets, making it easy to compare the distribution of vehicles by age. The clear visual disparity between the green and orange bars across age categories facilitates quick identification of trends and insights.

Conclusion: The bar graph titled “Vehicle Age Distribution” provides a clear overview of how vehicle counts vary by age. The high turnover of 1-year-old vehicles suggests strong demand for nearly new models, while the steady decline in older vehicle counts highlights the challenges faced by older vehicles in the market. The comparison between the two datasets reveals that the orange dataset consistently has a higher volume of vehicles, offering insights into market dynamics and consumer preferences. The visual representation enhances the interpretability of the data, making it a valuable tool for market analysis and strategic planning in the automotive sector.

5.1.6 Top 5 Least Selling Cars by Price

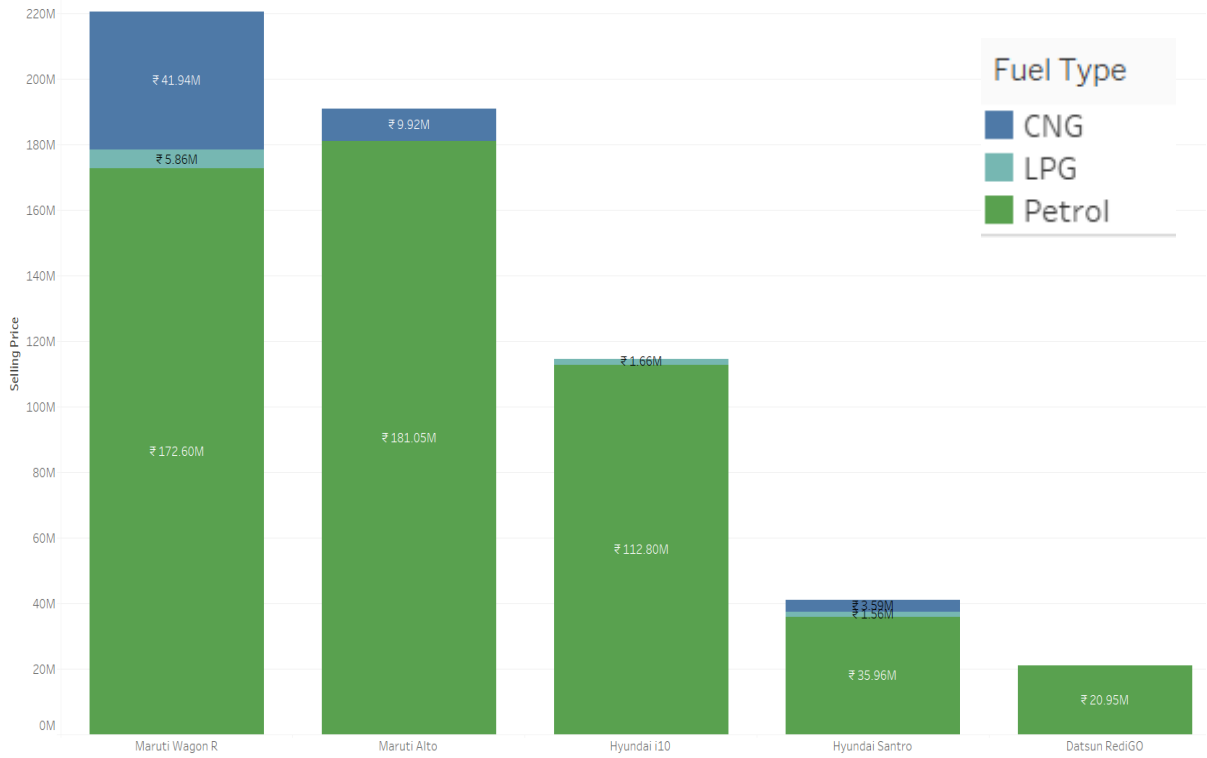


Figure 7 Top 5 Least Selling Cars by Price

Key Findings and Insights from "Top 5 Least Selling Cars by Price" Column Chart and Table:

1. **Price Comparison**:
    - **Nissan X-Trail** has the highest selling price among the five cars at **2,135,000**, indicating it is the most expensive among the least-selling models.
    - Following the Nissan X-Trail, the **Hyundai Aura** has a selling price of **900,000**.
    - **Maruti Dzire LXI** comes next with a selling price of **885,000**.
    - The **Tata Altroz** is priced at **730,000**.
    - **Force Gurkha** has the lowest selling price among the top 5 least-selling cars at **700,000**.
2. **Fuel Type Distribution**:

- While the table does not provide specific details on fuel types, we can infer that these models may offer a mix of fuel options (Petrol, Diesel, CNG) similar to other models in their categories.
- Given the pricing, it is likely that the Nissan X-Trail and Force Gurkha may include diesel variants, known for being pricier and preferred for their performance in larger vehicles.
- Hyundai Aura, Maruti Dzire LXI, and Tata Altroz are commonly available in petrol variants, with potential CNG options for cost efficiency.

3. **Market Preferences**:
    - The higher price of the Nissan X-Trail suggests that its lower sales could be attributed to its premium pricing, catering to a niche market segment.
    - The presence of less expensive models like the Tata Altroz and Force Gurkha among the least-selling cars indicates that factors beyond price, such as brand preference, fuel type availability, or vehicle features, play a significant role in consumer purchasing decisions.
4. **Relative Pricing**:
    - The chart demonstrates a clear price hierarchy among the least-selling cars, with a significant price drop from the Nissan X-Trail to the Hyundai Aura.
    - This hierarchy helps understand the market positioning of these vehicles and how pricing affects their sales performance. The data suggests that despite being less expensive, some models like the Tata Altroz and Force Gurkha still struggle with sales, indicating other factors at play.
5. **Visual Representation**:
    - The column chart uses color-coded bars to differentiate the selling prices of the top 5 least-selling cars, making it easy to compare their prices at a glance.
    - The visual disparity in bar heights highlights the substantial price difference between the highest and lowest selling prices within the least-selling cars.

Conclusion: The column chart and table provide a comprehensive overview of the selling prices of the top 5 least-selling cars. Nissan X-Trail stands out as the most expensive model

among them, while Force Gurkha is the least expensive. Despite the variety in prices, these models share the commonality of being less popular in the market. The analysis suggests that factors such as fuel type, brand perception, and specific vehicle features significantly influence consumer choices. The visual representation effectively highlights the price differences and aids in understanding market dynamics and consumer preferences.

5.1.7 KM Driven vs Milage Scatterplot

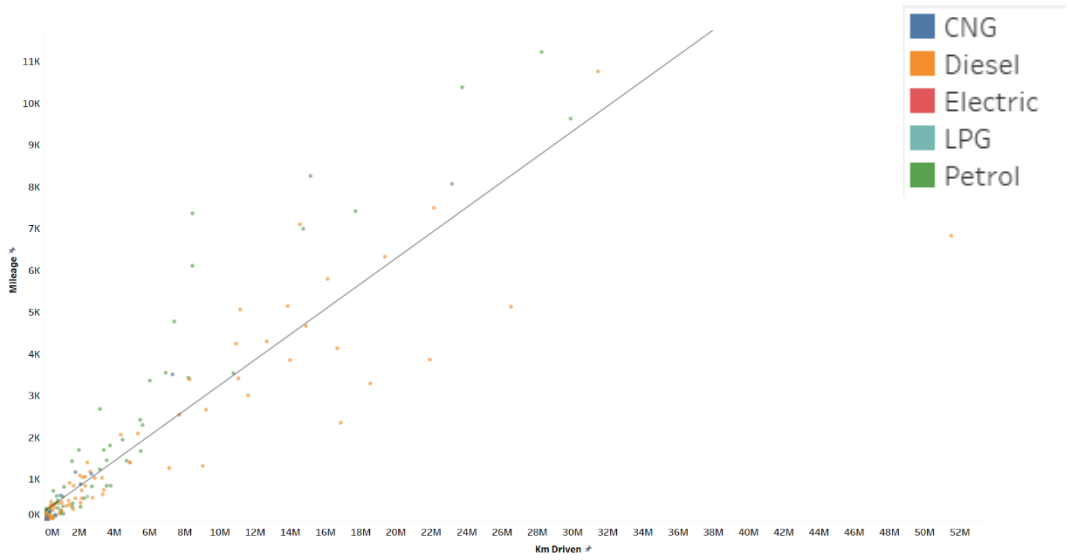


Figure 8 KM Driven vs Milage Scatterplot

Key Findings and Insights from "Km Driven vs Mileage" Scatter Plot:

1. **Positive Correlation**:
    - The scatter plot reveals a **positive correlation** between kilometers driven and mileage. This suggests that as the kilometers driven increase, so does the mileage. This trend is unusual since typically, one would expect mileage to decrease with increased vehicle usage due to wear and tear. This anomaly might warrant further investigation to understand the underlying factors contributing to this trend.
2. **Fuel Type Distribution**:
    - The data points are **color-coded** based on fuel type, allowing for an easy visual comparison of how different fuel types perform over extended usage:
        - **CNG (blue)**
        - **Diesel (orange)**

- **Electric (red)**
- **LPG (cyan)**
- **Petrol (green)**

3. **Performance Insights**:
    - **CNG and Diesel**:
        - These fuel types exhibit a **wide distribution** of data points, indicating variability in mileage performance as kilometers driven increase. This suggests that the efficiency of CNG and Diesel vehicles can vary significantly based on factors such as maintenance, driving conditions, and vehicle model.
    - **Electric**:
        - Electric vehicles have **fewer data points**, suggesting they might be less common or newer in the market. Their mileage performance appears to be more **consistent**, indicating that Electric vehicles may maintain their efficiency better over time compared to other fuel types.
    - **Petrol**:
        - Petrol vehicles also show a **wide range** of mileage performance, similar to CNG and Diesel. This variability indicates that Petrol vehicles' efficiency can be influenced by various factors, making their performance less predictable over extended usage.
4. **Long-term Efficiency**:
    - The scatter plot helps identify which fuel types maintain efficiency over time. For instance, if Electric vehicles show a consistent mileage despite higher kilometers driven, it could indicate **better long-term performance**. Conversely, the wide distribution in CNG and Diesel suggests that these fuel types may experience more fluctuations in efficiency as they accumulate more kilometers.
5. **Visual Representation**:

- The use of **color-coded** data points makes it easy to differentiate between fuel types and compare their performance visually. This enhances the readability of the chart and allows for quick identification of trends and outliers in the data.

Conclusion: The scatter plot provides valuable insights into the relationship between kilometers driven and mileage for different fuel types. The unusual positive correlation warrants further exploration, while the variability in performance among CNG, Diesel, and Petrol vehicles highlights the importance of considering multiple factors when evaluating long-term efficiency. Electric vehicles, with their consistent performance, stand out as potentially more reliable over extended usage. The visual representation effectively uses color to differentiate fuel types, making it easier to compare their performance and draw meaningful conclusions.

5.1.8 Selling Price Distribution

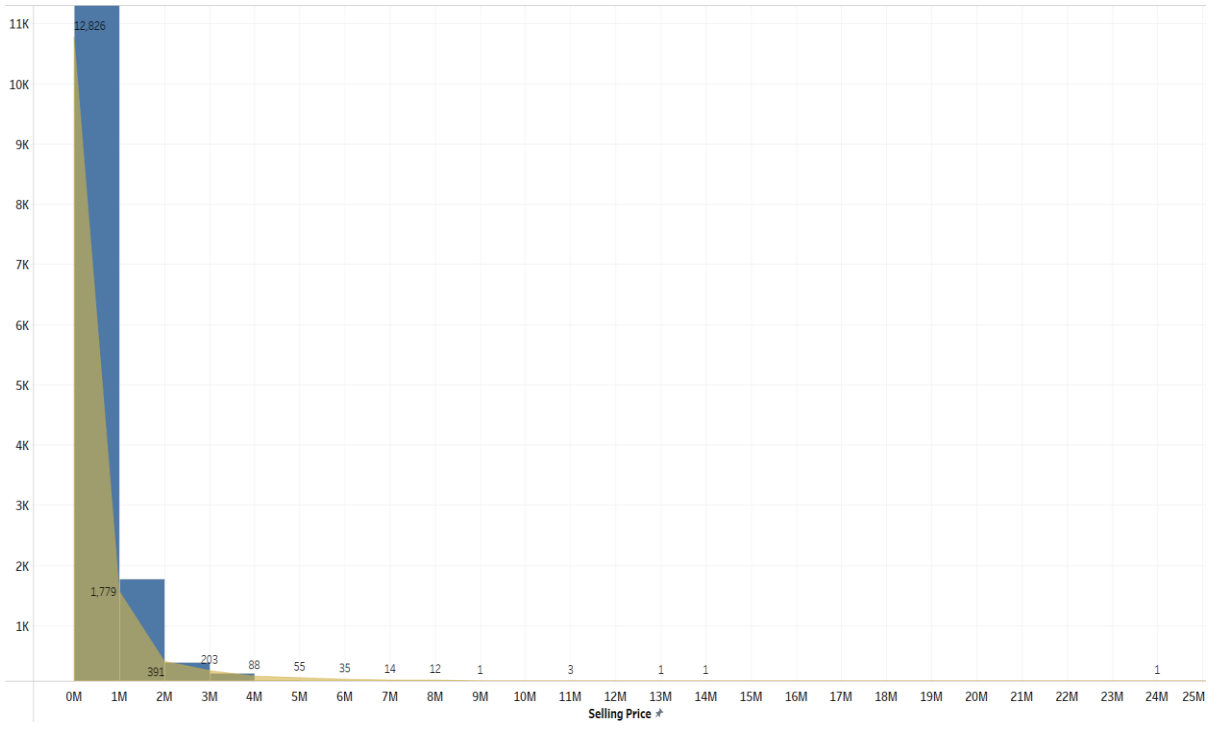


Figure 9 Selling Price Distribution

Key Findings

1. **Dominant Price Range:**

   - **High Frequency at 100K Selling Price:** The significant spike at the 100K selling price range indicates that this price point dominates the sales volume.

With over 10,000 occurrences, this suggests a strong preference or high turnover for products priced around 100K.

- **Lower-Priced Items:** The high frequency in the lower price range suggests that more affordable products sell faster and in higher quantities compared to those priced higher.

2. **Price Impact on Sales Volume:**
    - **Decreasing Frequency with Higher Prices:** As the selling price increases, the frequency of sales drops sharply. This decline in sales volume with rising prices indicates that demand decreases as prices go up.
3. **Market Demand Insights:**
    - **Strong Demand for Lower Price Bracket:** There is strong market demand for products within the lowest price range. This is evidenced by the high number of transactions at the 100K price point.
    - **Optimization Strategy:** Given the high turnover rate at lower price points, focusing sales strategies and marketing efforts on this range could yield better results.

Analysis of the Distribution

1. **Highly Skewed Distribution:**
    - **Left-Skewed Data:** The data shows a concentration of values in the lower price range, indicating a left-skewed distribution. This suggests that most sales occur at lower price points.
2. **Long Tail Effect:**
    - **Presence of Outliers:** The long tail of the distribution, with fewer data points at higher prices, implies the presence of a few high-value transactions. These outliers may significantly impact overall statistical measures.

3. **Possible Bimodal Distribution:**

- **Two Distinct Peaks:** The hint of two peaks (one around the 0-1M range and another near the 10M mark) suggests the possibility of bimodal distribution. This could indicate the presence of two different customer segments or product categories.

4. **Sparse Data at Higher Prices:**
    - **Limited Observations:** The scarcity of data points at higher prices might reflect either limited sales in those ranges or a potential bias in data collection. It's crucial to assess whether this represents the true market or a limitation of the dataset.

Insights

1. **Data Concentration and Behavior:**
    - **Common Price Range:** The concentration of data in the lower price range suggests a prevalent customer preference or market behavior. This can inform pricing strategies and inventory management.
2. **Impact of Outliers:**
    - **Influence of Extreme Values:** Outliers at higher price points can skew the mean and affect analysis. It's important to analyze these outliers separately to understand their impact.
3. **Segmentation Potential:**
    - **Two Distinct Groups:** If the bimodal distribution is confirmed, segmenting the data into different groups based on price ranges can provide deeper insights and tailor strategies to specific segments.

5.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the data analysis process that involves inspecting and analyzing datasets to summarize their main characteristics and gain insights. EDA helps in understanding the structure, patterns, and anomalies in the data before applying more complex statistical methods or building predictive models. Here's a detailed overview of EDA and its process:

Key Objectives of EDA

1. **Understanding Data Structure:**
    - Identify the types of variables (e.g., numerical, categorical) and their distributions.
    - Check for missing values and data inconsistencies.
2. **Descriptive Statistics:**
    - Calculate basic statistics such as mean, median, standard deviation, and quantiles.
    - Summarize the central tendency and dispersion of the data.
3. **Data Visualization:**
    - Use graphical methods to understand distributions, relationships, and patterns.
    - Common visualizations include histograms, scatter plots, box plots, and heatmaps.
4. **Identifying Patterns and Relationships:**
    - Examine relationships between variables using correlation analysis and scatter plots.
    - Identify potential trends, clusters, and anomalies.
5. **Detecting Anomalies and Outliers:**
    - Identify data points that deviate significantly from the rest of the data.
    - Determine if these outliers are errors or meaningful observations.
6. **Assessing Data Quality:**
    - Evaluate completeness, consistency, and accuracy of the data.
    - Address any issues related to data quality, such as missing or erroneous values.

Code:

```
def EDA(data_name,DV):
    import pandas as pd
```

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


data=pd.read_csv(data_name)
print('data has {} rows and {} columns'.format(data.shape[0],data.shape[1]))


null_col=[x for x in data.columns if data[x].isnull().sum()>0]
if len(null_col)>0:
    drop_col=[x for x in null_col if (data[x].isnull().sum()/data.shape[0])>0.1]
    if len(drop_col)>0:
        data=data.drop(drop_col,axis=1)
        print('Dropped {} columns from data and dropped columns are:
{}'.format(len(drop_col),drop_col))


    else:
        print('No columns dropped from the data')
    null_col_new=[x for x in null_col if x not in drop_col]
    null_qual=[x for x in null_col_new if data[x].dtype=='object']
    null_quant=[x for x in null_col_new if data[x].dtype!='object']
    if len(null_qual)>0:
        for x in null_qual:
            mode=max(dict(data.groupby(x)[DV].count()))
            data[x]=data[x].fillna(mode)
    if len(null_quant)>0:
```

```
        for x in null_quant:
            median=data[x].median()
            data[x]=data[x].fillna(median)
    print('data has {} no of null columns and they are replaced with mode and median'.format(len(null_col_new)))

else:
    print('data is clean form and no null values in data')

col_qualitative=[x for x in data.columns if data[x].dtype=='object']
col_conti=[x for x in data.columns if data[x].dtype!='object' and len(data[x].unique())>25]
col_disc=[x for x in data.columns if data[x].dtype!='object' and len(data[x].unique())<=25]

if len(col_conti)>0:
    left_skew=[x for x in col_conti if data[x].skew()<0]
    right_skew=[x for x in col_conti if data[x].skew()>0]
    no_skew=[x for x in col_conti if data[x].skew()==0]
    total_skew=len(left_skew)+len(right_skew)
    lepto_kurtic=[x for x in col_conti if data[x].kurt().round()>3]
    meso_kurtic=[x for x in col_conti if data[x].kurt().round()==3]
    plati_kurtic=[x for x in col_conti if data[x].kurt().round()<3]
    tot_kurt=len(lepto_kurtic)+len(plati_kurtic)
    outlier=[]
    stats=data[col_conti].describe()
    for x in col_conti:
```

```
        iqr=stats.loc['75%',x]-stats.loc['25%',x]
        ub=stats.loc['75%',x]+1.5*iqr
        lb=stats.loc['25%',x]-1.5*iqr
        if stats.loc['min', x]<lb or stats.loc['max',x]>ub:
            outlier.append(x)
    total_outlier=len(outlier)
    if total_skew>0:
        print('\nwe have {} no of skewed columns, so use normalization
models'.format(total_skew))
    if tot_kurt>0:
        print('\nwe have {} no of kurtosis columns, so use standardization
models'.format(tot_kurt))
    if total_outlier>0:
        print('\nwe have {} no of outlier columns, so use non linear
models'.format(total_outlier))
    else:
        print('\nno continuous columns observed in data')


    if len(col_disc)>0:
        print('\nInfluence analysis is conducted and result are as follows')
        for x in col_disc:
            data.groupby(x)[DV].mean().sort_values().plot.barh()
            plt.show()
    else:
        print('zero discrete columns observed in data')
```

```
    if len(col_qualitative)>0:
        print('\ninfluence analysis is conducted and result are as follows')
        for x in col_qualitative:
            data.groupby(x)[DV].mean().sort_values().plot.pie()
            plt.show()
    else:
        print('zero qualitative columns observed in data')


    print('correlation plot for given data')
    no_col=[x for x in data.columns if data[x].dtype!='object']
    plt.figure(figsize=(20,15))
    sns.heatmap(data[no_col].corr().round(1),annot=True, cmap='RdYlGn')
    plt.savefig('correlation heatmap.png')
    plt.show()


    print('pairplot for given data')
    sns.pairplot(data[no_col])
    plt.savefig('pairplot.png')
    plt.show()
```

```
EDA('/content/drive/MyDrive/Colab Notebooks/cardekho_dataset.csv','selling_price')
```

Output:

```
data has 15411 rows and 14 columns
data is clean form and no null values in data

we have 6 no of skewed columns, so use normalization models

we have 6 no of kurtosis columns, so use standardization models

we have 5 no of outlier columns, so use non linear models

Influence analysis is conducted and result are as follows
```

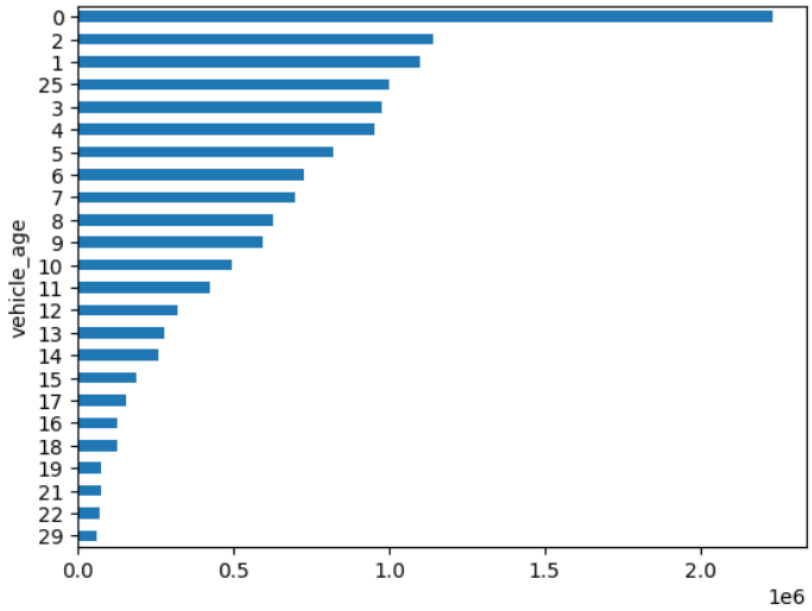


Figure 10 Vehicle Age Bar Chart

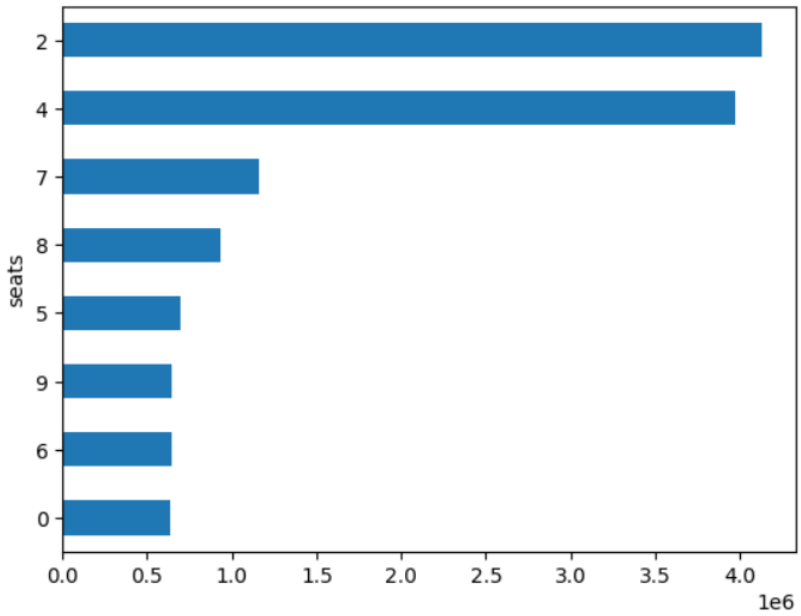


Figure 11 Seats Distribution

Analysis of the Graphs

Graph 1: Vehicle Age vs. Selling Price

- **Insights**:
    - **Vehicle Age Distribution**: The bar graph shows the distribution of vehicles across different age categories.
    - **Age 0 Dominance**: Vehicles aged 0 years (likely new or nearly new cars) have the highest selling price volume, significantly higher than any other age group. This is expected as newer cars tend to have higher market values.
    - **Gradual Decrease**: As vehicle age increases, the selling price volume generally decreases. This trend aligns with the depreciation principle where older vehicles lose value over time.
    - **Notable Drop After 5 Years**: There is a noticeable drop in selling price volume after vehicles reach 5 years of age, indicating a significant depreciation milestone.

Graph 2: Number of Seats vs. Selling Price

- **Insights**:
    - **Seat Configuration Distribution**: The bar graph displays the distribution of selling price volumes for cars with different seat configurations.
    - **2-Seat and 4-Seat Dominance**: Vehicles with 2 seats and 4 seats have the highest selling price volumes, with 2-seaters slightly leading. This could indicate that compact cars or sports cars (often with fewer seats) hold higher market values or are sold more frequently.
    - **7-Seat and 8-Seat Variants**: Cars with 7 and 8 seats also have notable selling price volumes, suggesting that larger family or utility vehicles hold significant market share.
    - **Lower Volumes for Other Configurations**: Other seat configurations, such as 5, 9, and 6 seats, have lower selling price volumes, which might indicate these configurations are less common or less in demand.

Conclusion:

The analysis of these graphs provides valuable insights into how vehicle age and seating capacity influence the selling price volumes in the used car market:

1. **Vehicle Age**:
    - Newer cars (aged 0 years) dominate in terms of selling price volume.
    - A general trend of depreciation is observed, with a notable drop after 5 years of age.
2. **Seating Capacity**:
    - Cars with 2 seats and 4 seats lead in selling price volumes, indicating higher market values or sales frequency.
    - Larger vehicles with 7 and 8 seats also show significant volumes, reflecting their importance in the market.

These insights can be instrumental in building predictive models for car prices, understanding market dynamics, and informing both buyers and sellers in the used car market.

```
influence analysis is conducted and result are as follows
```

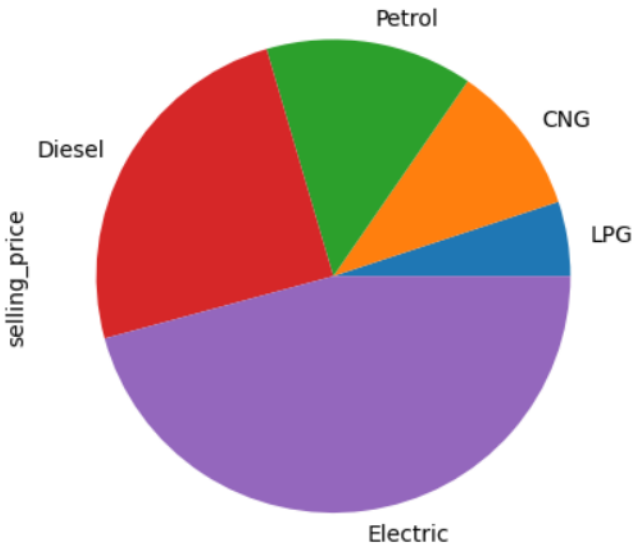


Figure 12 Fuel Type by Selling Price

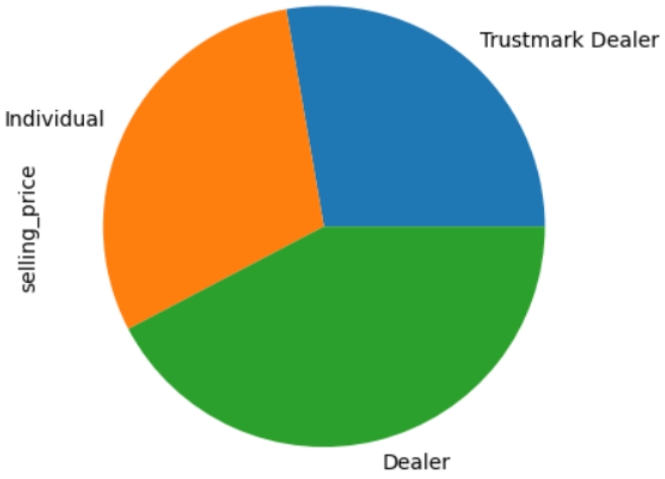


Figure 13 Dealer Type by Selling Price

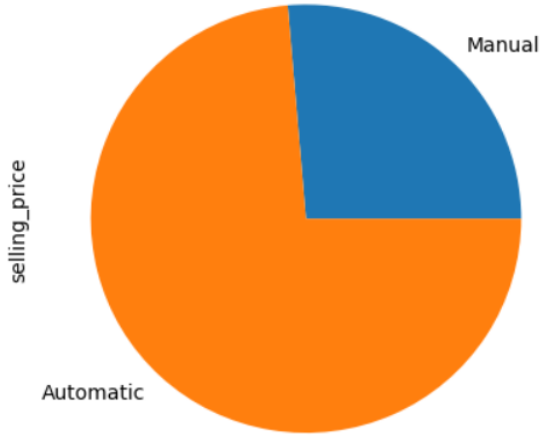


Figure 14 Transmission Type by Selling Price

Pie Chart 1: Fuel Type

Petrol Dominance: Petrol-powered vehicles seem to be the most prevalent, occupying the largest slice of the pie.

Diesel Presence: Diesel vehicles also have a significant share, indicating a market for diesel-powered cars.

Emerging Alternatives: CNG, LPG, and Electric vehicles have smaller but noticeable shares, suggesting growing interest in alternative fuel options.

Pie Chart 2: Seller Type

Individual Sellers: The majority of cars appear to be listed by individual sellers.

Dealer Presence: Dealers also contribute significantly to the market, indicating a mix of both individual and professional sellers.

Trustmark Dealer: A smaller but distinct category of Trustmark Dealers suggests a specific segment of the market.

Pie Chart 3: Transmission Type

Manual Transmission Predominance: Manual transmission vehicles seem to be more common in the dataset.

Automatic Transmission Growth: Automatic transmission vehicles have a substantial share, indicating a growing preference for convenience.

Overall Insights

Market Diversity: The car market represented in the data is diverse, with a variety of brands, fuel types, seller types, and transmission options.

Petrol Dominance: Petrol remains the primary fuel choice, but there's a growing interest in alternative fuels.

Individual Sellers: A significant portion of the market consists of individual sellers.

Manual Transmission Prevalence: Manual transmission vehicles are still popular, but automatic transmission is gaining ground.

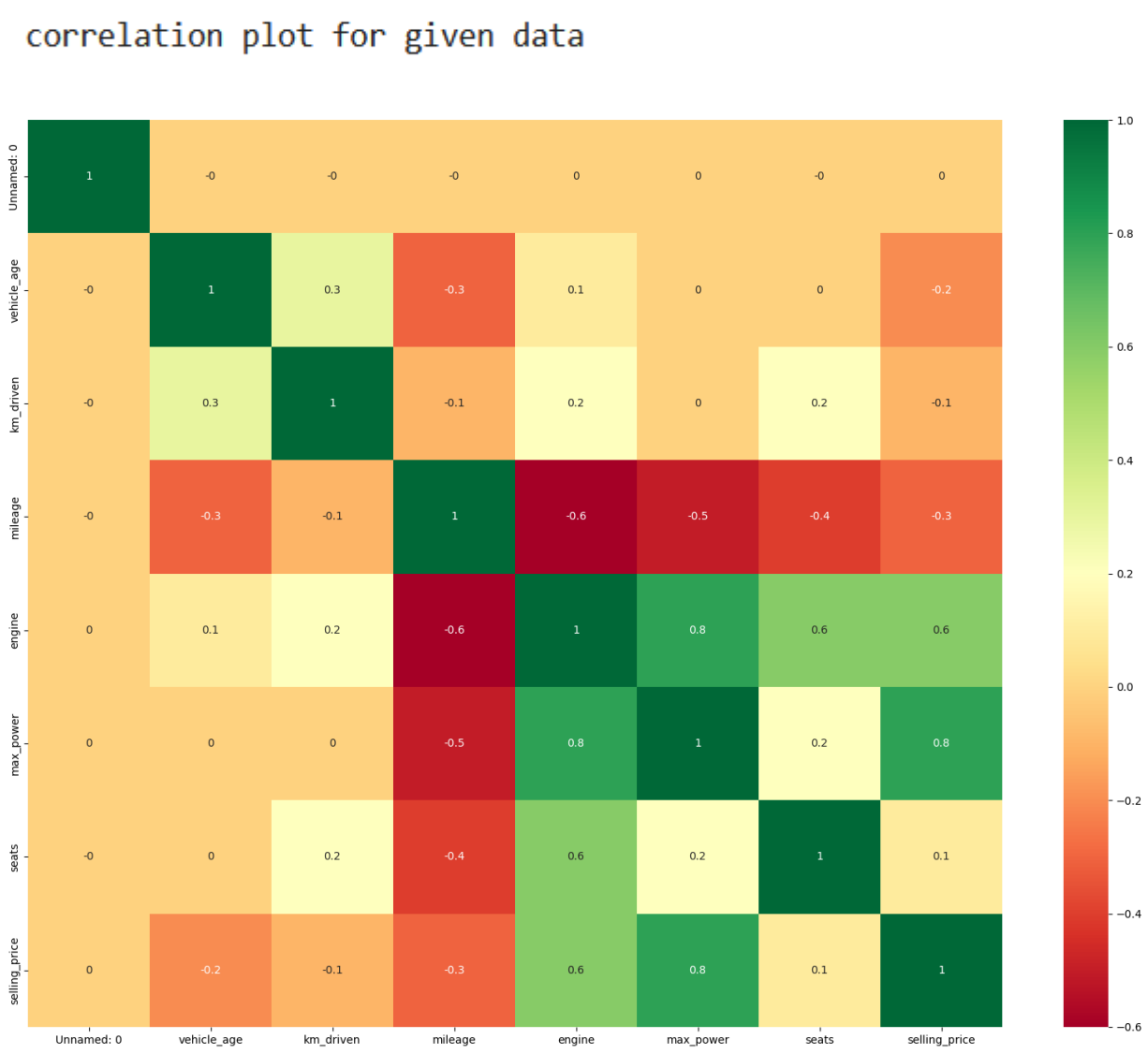


Figure 15 Correlation Heatmap

Analysis of the Correlation Heatmap

Heatmap of Correlations Between Various Features

Insights:

1. **Strong Positive Correlations**:
    - **Max Power and Engine (0.8)**: There is a strong positive correlation between max power and engine size, indicating that cars with larger engines tend to have higher power output.
    - **Selling Price and Max Power (0.8)**: The selling price is strongly correlated with max power, suggesting that vehicles with higher power tend to be priced higher.

- **Selling Price and Engine (0.6)**: There is also a strong positive correlation between the selling price and engine size, supporting the idea that cars with larger engines are valued higher.

2. **Moderate Positive Correlations**:
   - **Vehicle Age and Kilometers Driven (0.3)**: Older vehicles tend to have more kilometers driven, which is a logical relationship.
   - **Engine and Seats (0.6)**: Larger engines are somewhat correlated with the number of seats, possibly indicating that larger vehicles with more seating capacity also have larger engines.
3. **Weak to Moderate Negative Correlations**:
   - **Vehicle Age and Mileage (-0.3)**: There is a moderate negative correlation between vehicle age and mileage, suggesting that older vehicles might not be as fuel-efficient as newer ones.
   - **Mileage and Max Power (-0.5)**: Vehicles with higher max power tend to have lower mileage, indicating that more powerful cars are less fuel-efficient.
   - **Mileage and Engine (-0.6)**: Larger engines are correlated with lower mileage, supporting the idea that bigger engines consume more fuel.
4. **Weak Correlations**:
   - **Selling Price and Vehicle Age (-0.2)**: The selling price has a weak negative correlation with vehicle age, indicating that older cars are generally sold for less.
   - **Selling Price and Kilometers Driven (-0.1)**: There is a weak negative correlation between the selling price and the number of kilometers driven, suggesting that cars with higher mileage might have slightly lower selling prices.
   - **Selling Price and Mileage (-0.3)**: Selling price is moderately negatively correlated with mileage, indicating that more fuel-efficient cars might command higher prices.

- **Seats and Engine (0.6)**: More seats are somewhat correlated with larger engines, as previously mentioned.
- **Seats and Selling Price (0.1)**: There is a weak positive correlation between the number of seats and the selling price, indicating that cars with more seats might be priced slightly higher.

Conclusion:

The correlation heatmap provides a clear visualization of how different features of the dataset are related to each other:

- **Engine Size, Max Power, and Selling Price**: Strongly correlated, indicating that more powerful and larger-engine cars are valued higher.
- **Vehicle Age and Kilometers Driven**: Moderately correlated, showing that older cars generally have more mileage.
- **Mileage**: Negatively correlated with both engine size and max power, indicating a trade-off between power/size and fuel efficiency.
- **Seats**: Somewhat correlated with engine size, reflecting the tendency for larger vehicles to have both more seats and larger engines.

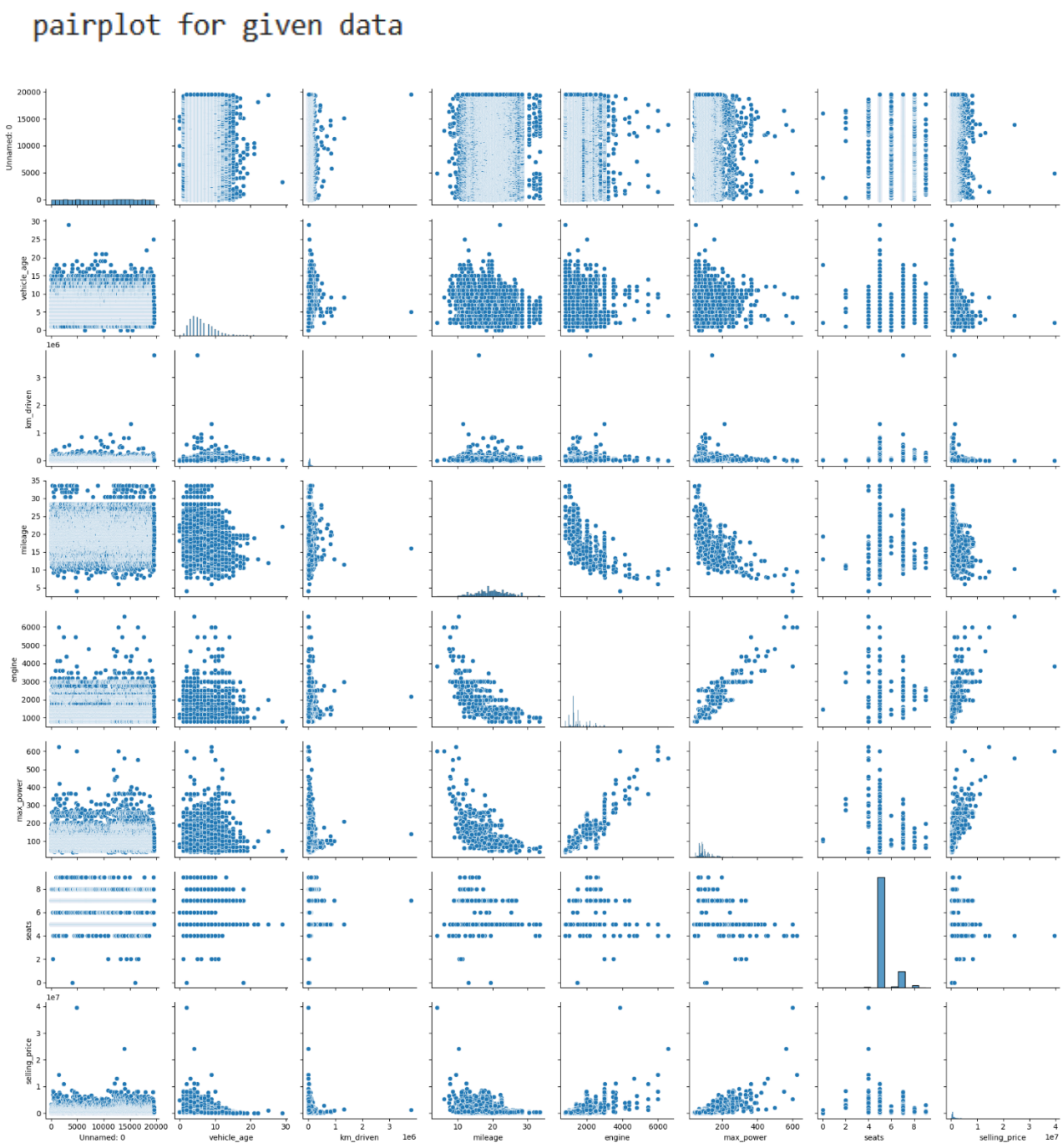


Figure 16 Pairplot

Key Findings and Insights:

1. Unnamed: 0 vs. Selling Price

- **Insights**: This column appears to be an index or ID column, which does not have a meaningful relationship with the selling price.

2. Vehicle Age vs. Selling Price

- **Insights**: There is a clear negative correlation between vehicle age and selling price. As the age of the vehicle increases, the selling price generally decreases. This is intuitive as older vehicles tend to depreciate in value.

3. Km Driven vs. Selling Price

- **Insights**: There is also a negative correlation between kilometers driven and selling price. Cars with higher mileage tend to have lower selling prices, which reflects wear and tear over time.

4. Mileage vs. Selling Price

- **Insights**: The relationship between mileage (fuel efficiency) and selling price appears to be less straightforward. While more efficient cars might sell for higher prices, other factors like vehicle type and age might influence this relationship.

5. Engine vs. Selling Price

- **Insights**: There is a positive correlation between engine size and selling price. Larger engines often indicate more powerful or premium vehicles, which typically sell for higher prices.

6. Max Power vs. Selling Price

- **Insights**: Similar to engine size, there is a positive correlation between maximum power and selling price. Cars with higher power output tend to be more expensive.

7. Seats vs. Selling Price

- **Insights**: The number of seats does not show a strong correlation with selling price. Most cars have a standard number of seats (usually 4 or 5), and any variations might not significantly impact the price.

Additional Pairwise Comparisons:

- **Vehicle Age vs. Km Driven**: Older vehicles tend to have more kilometers driven, which is consistent with expected usage over time.
- **Vehicle Age vs. Engine**: There might not be a strong correlation, but older models might have smaller engines compared to newer models.

- **Km Driven vs. Engine**: Cars with larger engines might be driven more, but the relationship is not very clear.

Conclusion:

The scatter plot matrix provides a comprehensive visualization of the relationships between various features and the selling price of used cars. The key insights are:

1. **Negative Correlation**:
    - Vehicle age and selling price.
    - Kilometers driven and selling price.
2. **Positive Correlation**:
    - Engine size and selling price.
    - Maximum power and selling price.
3. **No Strong Correlation**:
    - Number of seats and selling price.

```
# seller_type vs km_driven

figsize = (12, 1.2 * len(data['seller_type'].unique()))

plt.figure(figsize=figsize)

sns.violinplot(data, x='km_driven', y='seller_type', inner='box', palette='Dark2')

sns.despine(top=True, right=True, bottom=True, left=True)
```

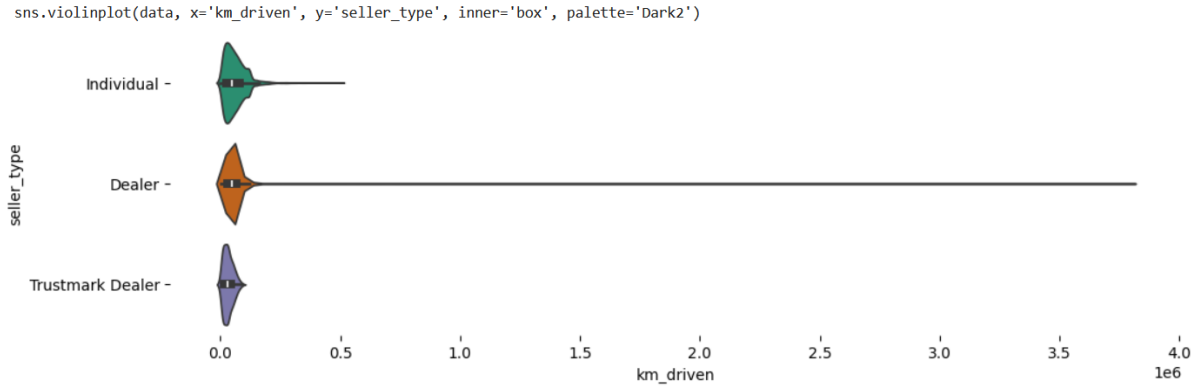


Figure 17 Violin Plot: km_driven vs seller_type

Analysis of the Violin Plot: km_driven vs seller_type

Understanding the Plot

The violin plot provides a visual representation of the distribution of the km driven variable across different seller type categories. The thicker parts of the violin plot represent higher density of data points, while thinner parts indicate lower density. The white dot within each violin represents the median value, and the black box indicates the interquartile range (IQR).

Key Findings and Insights

1. **Distribution of Kilometers Driven:**
    - **Individual Sellers:** The distribution for individual sellers is skewed to the right, with a longer tail towards higher kilometer values. This suggests that a significant portion of cars sold by individuals has been driven a considerable distance.
    - **Dealers:** The distribution for dealers is also skewed to the right, but the tail is shorter compared to individual sellers. This indicates that dealer-sold cars generally have lower mileage, although there are still some cars with high mileage.
    - **Trustmark Dealers:** The distribution for Trustmark Dealers shows a similar pattern to dealers, with a slightly shorter tail towards higher kilometer values. This suggests that Trustmark Dealers tend to sell cars with relatively lower mileage compared to individual sellers.
2. **Median Kilometers Driven:**
    - The median kilometers driven for individual sellers is higher than both dealers and Trustmark Dealers, indicating that cars sold by individuals generally have higher mileage on average.
    - Dealers and Trustmark Dealers have similar median kilometer values, suggesting that the average mileage of cars sold by these two seller types is comparable.
3. **Spread of Data:**

- The violin plot for individual sellers is wider than those for dealers and Trustmark Dealers, indicating a larger spread in the kilometers driven for cars sold by individuals. This suggests more variability in the mileage of cars sold by individuals.
- Dealers and Trustmark Dealers have a narrower spread, indicating that the mileage of cars sold by these sellers is more concentrated around the median value.

Overall Observations

- The data suggests that cars sold by individual sellers tend to have higher mileage compared to those sold by dealers and Trustmark Dealers.
- There is more variability in the mileage of cars sold by individual sellers compared to dealers and Trustmark Dealers.
- Trustmark Dealers generally sell cars with lower mileage compared to individual sellers.

```
# transmission_type vs vehicle_age

figsize = (12, 1.2 * len(data['transmission_type'].unique()))

plt.figure(figsize=figsize)

sns.violinplot(data, x='vehicle_age', y='transmission_type', inner='box', palette='Dark2')

sns.despine(top=True, right=True, bottom=True, left=True)
```

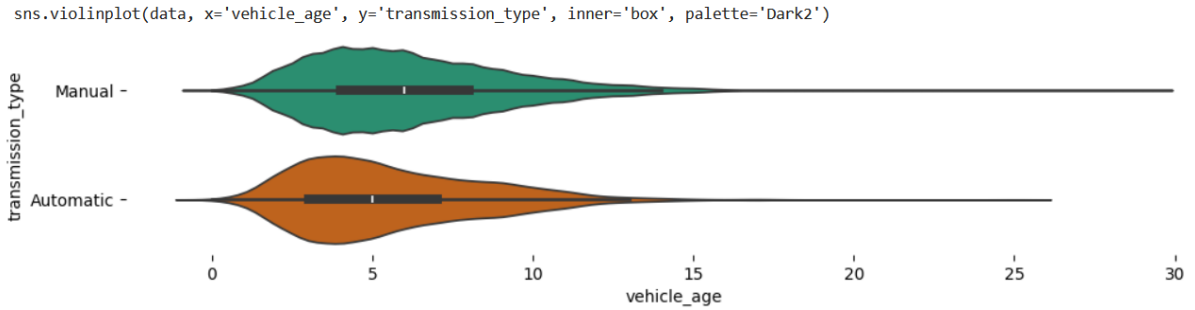


Figure 18 Violin Plot: Vehicle Age vs Transmission Type

Analysis of the Violin Plot: Vehicle Age vs Transmission Type

Understanding the Plot

The violin plot visually represents the distribution of vehicle age across different transmission types (manual and automatic). The thicker parts of the violin plot indicate higher density of data points, while thinner parts represent lower density. The white dot within each violin represents the median value, and the black box indicates the interquartile range (IQR).

Key Findings and Insights

1. **Distribution of Vehicle Age:**
    - **Manual Transmission:** The distribution for manual transmission vehicles is skewed to the right, with a longer tail towards higher age values. This suggests a significant portion of manual transmission cars are older.
    - **Automatic Transmission:** The distribution for automatic transmission vehicles is also skewed to the right, but the tail is shorter compared to manual transmission. This indicates a higher proportion of newer cars with automatic transmission.
2. **Median Vehicle Age:**
    - The median vehicle age for manual transmission cars is higher than that for automatic transmission cars, indicating that manual transmission cars tend to be older on average.
    - Automatic transmission cars have a lower median age, suggesting a higher proportion of newer cars with this transmission type.
3. **Spread of Data:**
    - The violin plot for manual transmission cars is wider than that for automatic transmission cars, indicating a larger spread in the age of manual transmission vehicles. This suggests more variability in the age of manual transmission cars.
    - Automatic transmission cars have a narrower spread, indicating that the age of automatic transmission cars is more concentrated around the median value.

Overall Observations

- The data suggests that manual transmission cars tend to be older compared to automatic transmission cars.

- There is more variability in the age of manual transmission cars compared to automatic transmission cars.
- Automatic transmission cars are generally newer compared to manual transmission cars.

```
# fuel_type vs vehicle_age

figsize = (12, 1.2 * len(data['fuel_type'].unique()))

plt.figure(figsize=figsize)

sns.violinplot(data, x='vehicle_age', y='fuel_type', inner='box', palette='Dark2')

sns.despine(top=True, right=True, bottom=True, left=True)
```

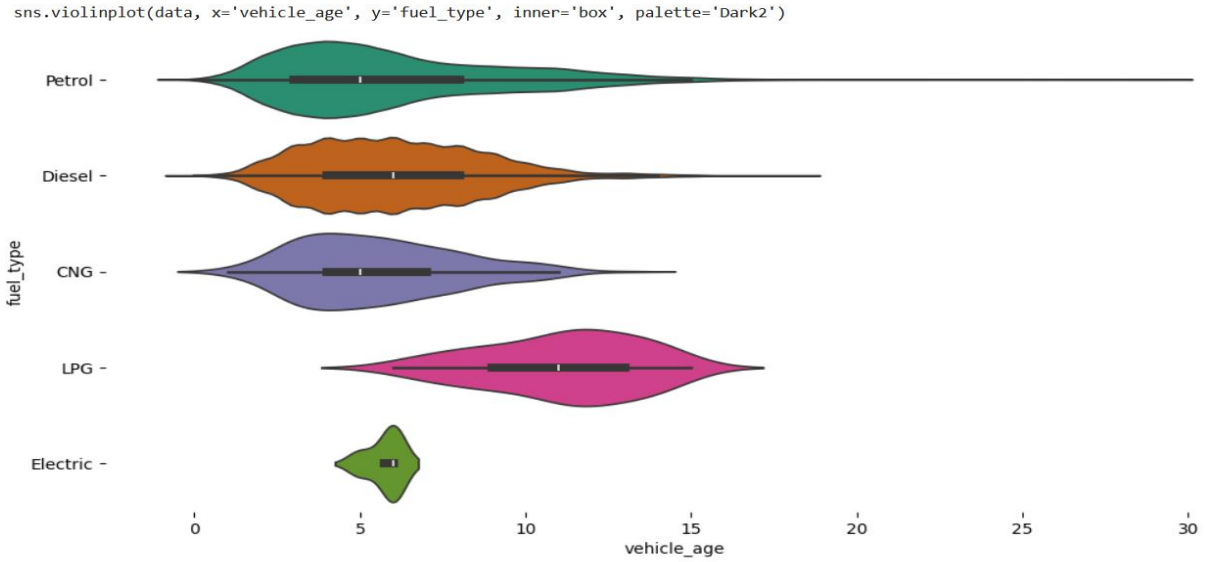


Figure 19 Violin Plot: Vehicle Age vs Fuel Type

Analysis of the Violin Plot: Vehicle Age vs Fuel Type

Understanding the Plot

The violin plot visually represents the distribution of vehicle age across different fuel types. The thicker parts of the violin plot indicate higher density of data points, while thinner parts represent lower density. The white dot within each violin represents the median value, and the black box indicates the interquartile range (IQR).

Key Findings and Insights

1. **Distribution of Vehicle Age:**

- **Petrol:** The distribution for petrol vehicles is skewed to the right, with a longer tail towards higher age values. This suggests a significant portion of petrol cars are older.
- **Diesel:** The distribution for diesel vehicles is also skewed to the right, but the tail is shorter compared to petrol vehicles. This indicates a higher proportion of newer diesel cars.
- **CNG, LPG, and Electric:** The distributions for CNG, LPG, and Electric vehicles are concentrated towards lower age values, with very few older vehicles in these categories. This suggests that these fuel types are relatively newer in the market.

2. **Median Vehicle Age:**
    - The median vehicle age for petrol and diesel vehicles is higher than that for CNG, LPG, and electric vehicles, indicating that these alternative fuel types are generally newer.
    - CNG, LPG, and electric vehicles have significantly lower median age values, suggesting a higher proportion of newer cars in these categories.
3. **Spread of Data:**
    - The violin plots for petrol and diesel vehicles are wider than those for CNG, LPG, and electric vehicles, indicating a larger spread in the age of petrol and diesel cars. This suggests more variability in the age of these vehicles.
    - CNG, LPG, and electric vehicles have narrower distributions, indicating that the age of these vehicles is more concentrated around the median value.

Overall Observations

- The data suggests that petrol and diesel vehicles tend to be older compared to CNG, LPG, and electric vehicles.
- There is more variability in the age of petrol and diesel vehicles compared to CNG, LPG, and electric vehicles.
- CNG, LPG, and electric vehicles are generally newer compared to petrol and diesel vehicles.

```
# seller_type vs vehicle_age

figsize = (12, 1.2 * len(data['seller_type'].unique()))

plt.figure(figsize=figsize)

sns.violinplot(data, x='vehicle_age', y='seller_type', inner='box', palette='Dark2')

sns.despine(top=True, right=True, bottom=True, left=True)
```

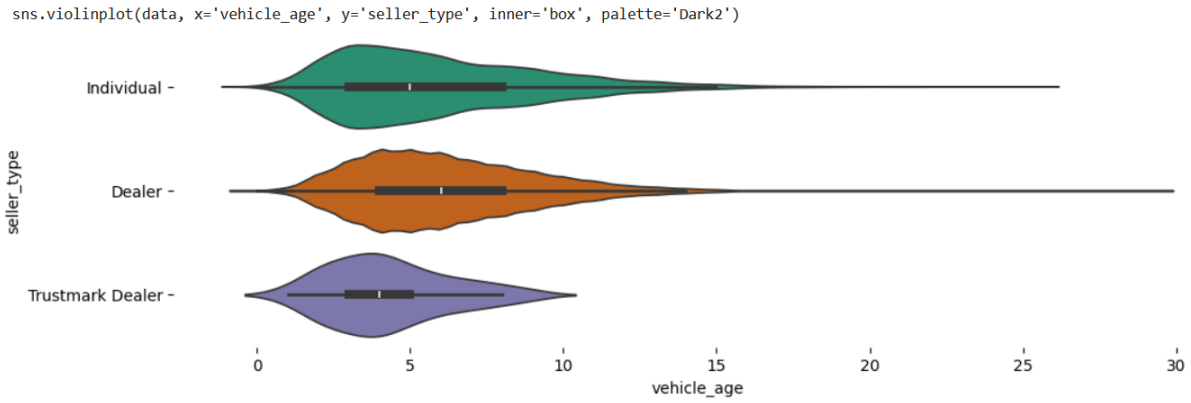


Figure 20 Violin Plot: Vehicle Age vs Seller Type

Analysis of the Violin Plot: Vehicle Age vs Seller Type

Understanding the Plot

The provided violin plot illustrates the distribution of vehicle age across different seller types: Individual, Dealer, and Trustmark Dealer. The thickness of the violin shape represents the density of data points at a particular vehicle age. The white dot within each violin indicates the median vehicle age, and the black box represents the interquartile range (IQR).

Key Findings and Insights

1. **Distribution of Vehicle Age:**
    - **Individual Sellers:** The distribution for individual sellers is skewed to the right, indicating a higher proportion of older vehicles. There is a substantial number of cars with ages between 5 and 15 years.
    - **Dealers:** The distribution for dealers is also skewed to the right but less pronounced than individual sellers. This suggests a higher proportion of newer vehicles compared to individual sellers.

- **Trustmark Dealers:** The distribution for Trustmark Dealers is concentrated towards lower vehicle ages, with a clear peak around the 5-10 year range. This indicates a focus on relatively newer vehicles.

2. **Median Vehicle Age:**
    - Individual sellers have the highest median vehicle age, suggesting they tend to sell older cars.
    - Dealers have a lower median vehicle age compared to individual sellers, indicating a focus on slightly newer vehicles.
    - Trustmark Dealers have the lowest median vehicle age, suggesting a preference for relatively new cars.
3. **Spread of Data:**
    - The distribution for individual sellers is wider, indicating a larger range of vehicle ages.
    - Dealers and Trustmark Dealers have narrower distributions, suggesting a more focused range of vehicle ages.

Overall Observations

- Individual sellers tend to offer older vehicles with a wider range of ages.
- Dealers offer a mix of vehicle ages, with a focus on slightly newer cars compared to individual sellers.
- Trustmark Dealers specialize in relatively newer vehicles with a narrower age range.

```
# fuel_type vs transmission_type

plt.subplots(figsize=(8, 8))

df_2dhist = pd.DataFrame({

   x_label: grp['transmission_type'].value_counts()

   for x_label, grp in data.groupby('fuel_type')

})

sns.heatmap(df_2dhist, cmap='viridis')
```

```
plt.xlabel('fuel_type')

_ = plt.ylabel('transmission_type')
```

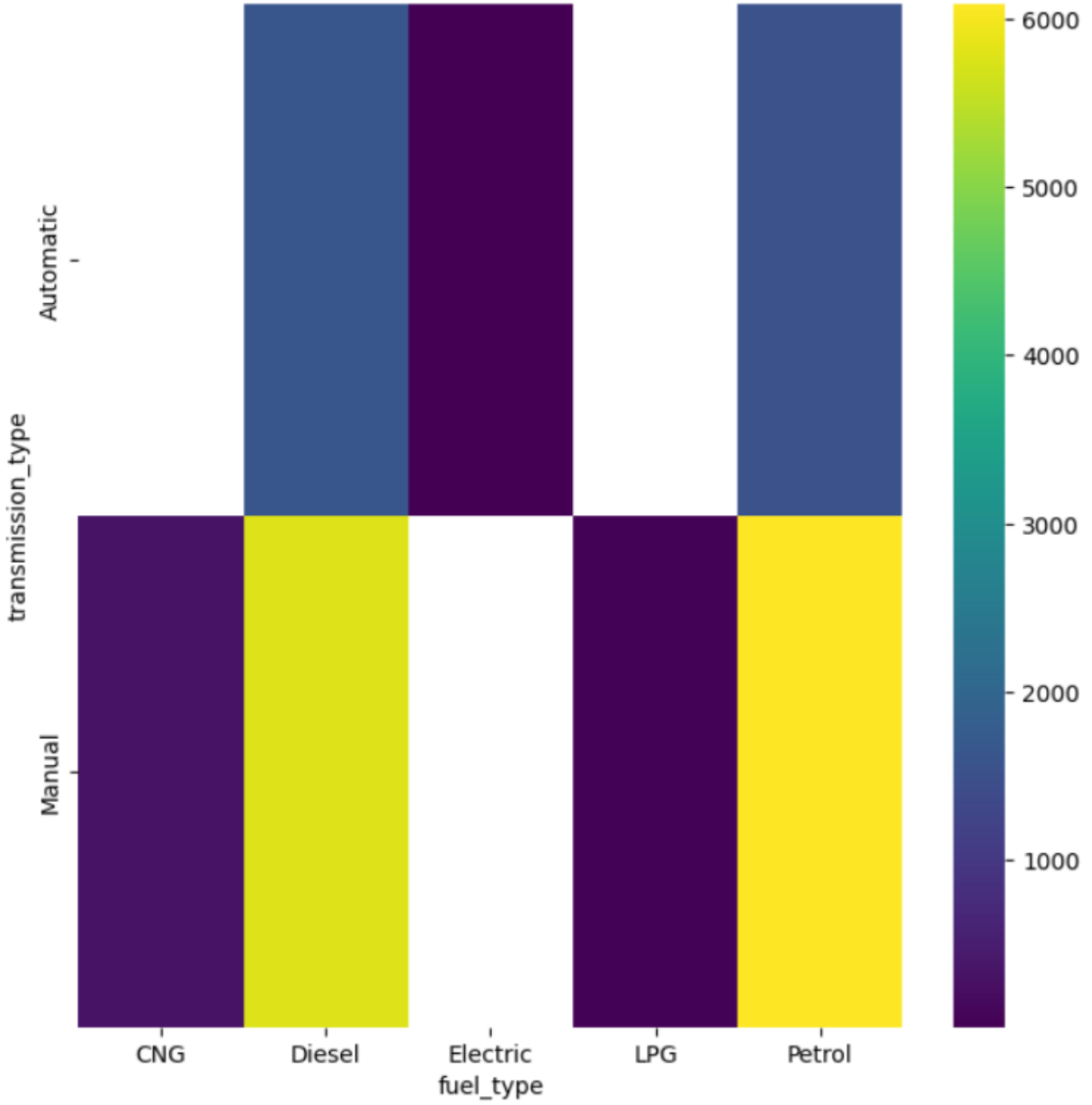


Figure 21 Heatmap: Fuel Type Transmission Type

The graph presents a heatmap visualizing the relationship between fuel type and transmission type with an underlying quantitative variable.

Key Observations:

1. **Petrol Dominance:** The Petrol category across both Manual and Automatic transmission types exhibits the highest values, indicated by the intense yellow color. This suggests that petrol-powered cars are the most prevalent in the dataset.

2. **Diesel Popularity:** Diesel vehicles also have a significant presence, particularly in the Manual transmission category, represented by the deep blue color.
3. **Emerging Alternatives:** CNG and LPG show moderate values, suggesting a growing but still smaller segment of the market for these alternative fuels. Electric vehicles have the lowest representation across both transmission types.
4. **Transmission Preferences:** While Manual transmission has a higher overall count, Automatic transmission also has a substantial presence, especially in the Petrol and Diesel categories.

Insights:

- The visualization effectively communicates the distribution of cars across different fuel and transmission types.
- Petrol remains the dominant fuel choice, followed by diesel.
- Alternative fuels like CNG, LPG, and electric have a smaller market share but are growing in popularity.
- Both manual and automatic transmissions are prevalent, with some variations across fuel types.

```
# seller_type vs fuel_type

plt.subplots(figsize=(8, 8))

df_2dhist = pd.DataFrame({

    x_label: grp['fuel_type'].value_counts()

    for x_label, grp in data.groupby('seller_type')

})

sns.heatmap(df_2dhist, cmap='viridis')

plt.xlabel('seller_type')

_ = plt.ylabel('fuel_type')
```

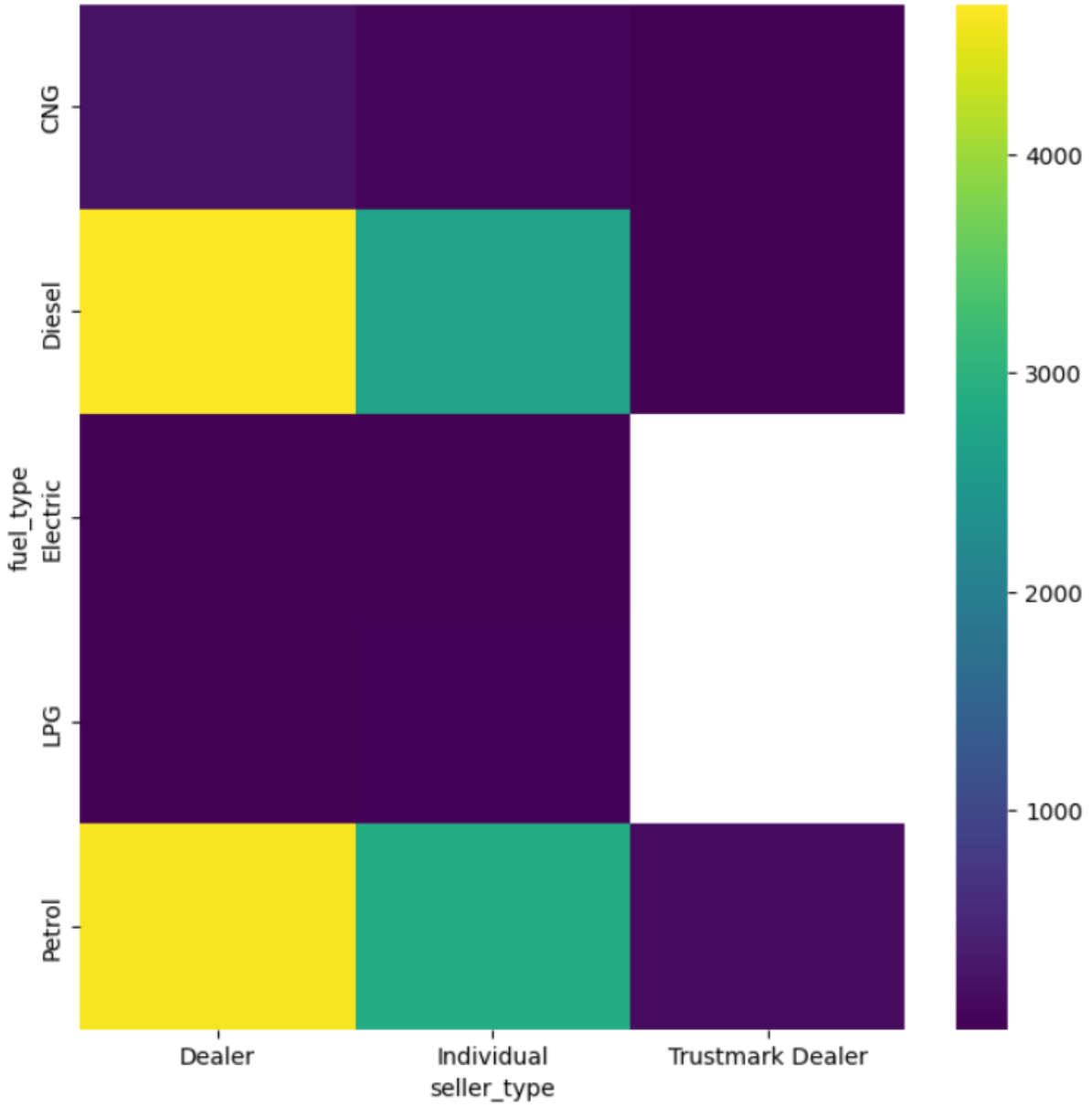


Figure 22 Heatmap: Fuel Type vs. Seller Type

Analysis of the Heatmap: Fuel Type vs. Seller Type

Understanding the Heatmap

The heatmap visualizes the relationship between two categorical variables: fuel_type and seller_type. The color intensity represents the frequency or count of observations within each combination of fuel type and seller type. Darker colors indicate higher frequencies.

Key Findings and Insights

1. **Petrol Dominance:** Petrol-powered cars have the highest overall count, represented by the darkest shade across all seller types. This indicates that petrol is the most common fuel type in the dataset.
2. **Diesel Popularity:** Diesel vehicles have a significant presence, particularly among dealers and individual sellers. This suggests a strong market for diesel cars.
3. **Emerging Alternatives:** CNG, LPG, and Electric vehicles have lower counts compared to petrol and diesel, indicating a smaller market share for these alternative fuels.
4. **Seller Type Distribution:**
    - Dealers have a more diverse range of fuel types, with a strong presence in petrol and diesel segments.
    - Individual sellers also have a significant share of petrol and diesel vehicles, with a smaller proportion of alternative fuel cars.
    - Trustmark Dealers primarily deal with petrol and diesel cars, with a minimal representation of other fuel types.
5. **Potential Correlations:**
    - There might be a correlation between fuel type and seller type. For instance, dealers might be more likely to offer a wider range of fuel options compared to individual sellers.

```
# engine

data['engine'].plot(kind='hist', bins=20, title='engine')

plt.gca().spines[['top', 'right',]].set_visible(False)
```

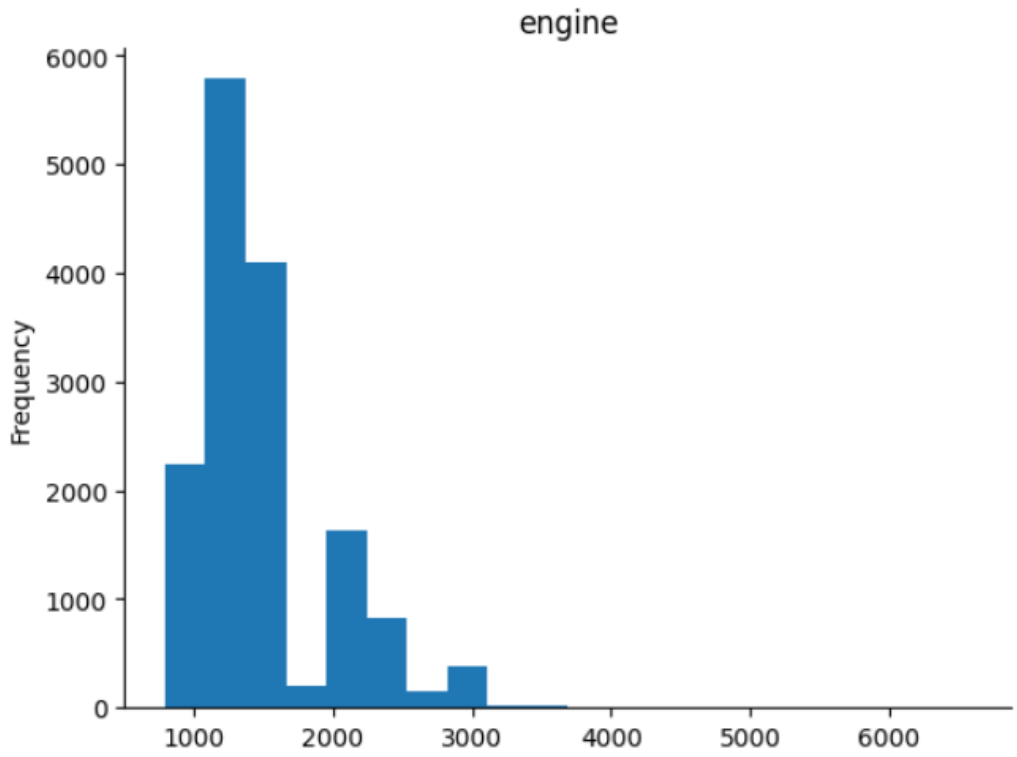


Figure 23 Engine Size Histogram

Analysis of the Engine Size Histogram

Understanding the Histogram

The provided histogram visualizes the distribution of engine sizes in a dataset. The x-axis represents engine size in cubic centimeters (cc), and the y-axis represents the frequency of cars with that engine size.

Key Findings and Insights

1. **Distribution Shape:** The histogram exhibits a right-skewed distribution, indicating that most cars in the dataset have smaller engines, with a tail extending towards larger engine sizes.
2. **Peak Frequency:** The highest frequency occurs within the 1000-2000 cc range, suggesting that this engine size is the most common in the dataset.
3. **Outliers:** While the majority of cars have engine sizes within the 1000-3000 cc range, there are a few cars with significantly larger engines, represented by the bars on the right side of the histogram. These could be considered outliers.

4. **Data Concentration:** The distribution is concentrated in the lower engine size range, with a gradual decrease in frequency as engine size increases.

Implications

- The predominance of smaller engines suggests a market preference for fuel efficiency and lower operating costs.
- The presence of larger engines indicates a segment of customers who prioritize power and performance.
- The right-skewed distribution implies that there might be a need for further analysis to understand the factors influencing the choice of engine size.

```
# mileage

data['mileage'].plot(kind='hist', bins=20, title='mileage')

plt.gca().spines[['top', 'right',]].set_visible(False)
```

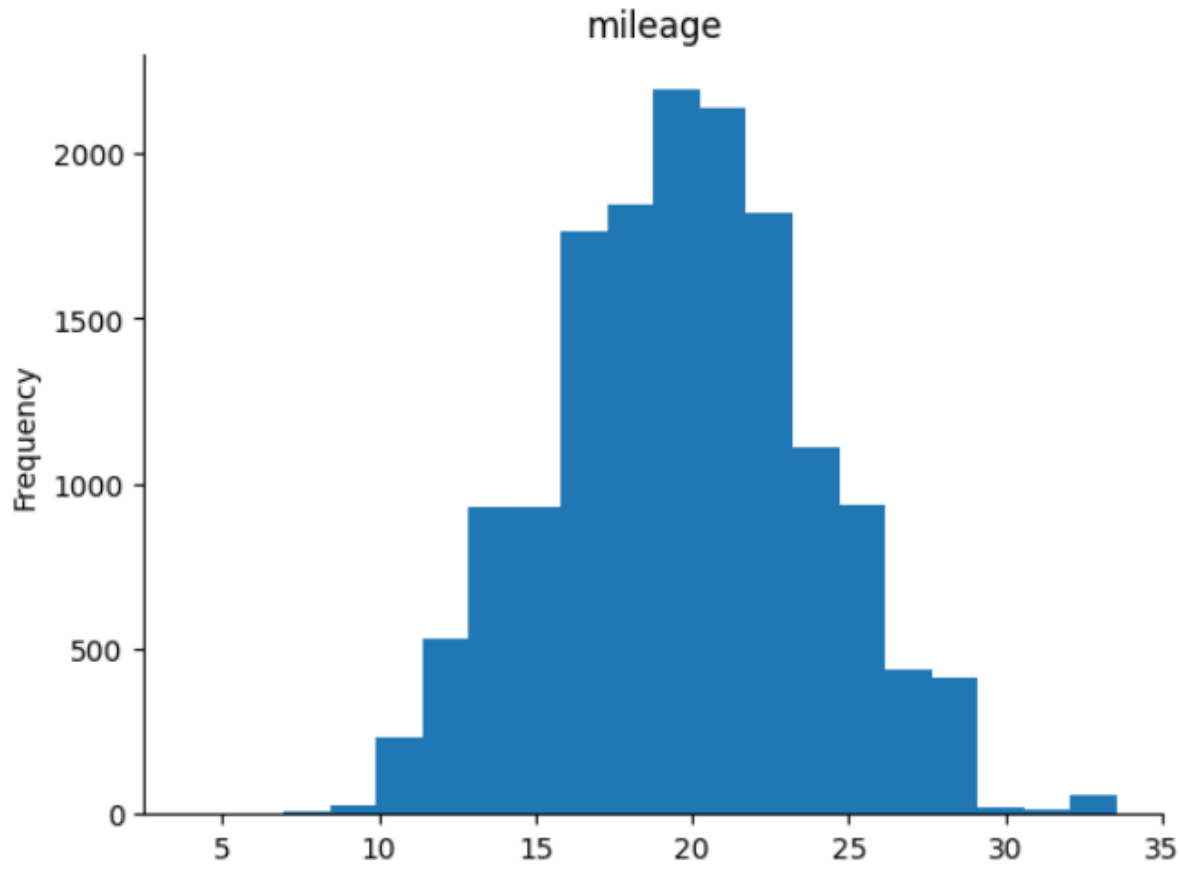


Figure 24 Mileage Histogram

Analysis of the Mileage Histogram

Understanding the Histogram

The provided histogram visualizes the distribution of the mileage variable. The x-axis represents the mileage values, and the y-axis represents the frequency of cars with that specific mileage.

Key Findings and Insights:

1. **Distribution Shape:** The histogram exhibits a bell-shaped curve, indicating a normal or approximately normal distribution of mileage values. This suggests that a majority of cars in the dataset have mileage values clustering around the central tendency.
2. **Central Tendency:** The peak of the distribution lies around the 15-20 mileage range, suggesting that a significant portion of the cars in the dataset have mileage values within this range.
3. **Spread:** The distribution is relatively spread out, indicating a certain degree of variability in the mileage values. There are cars with both lower and higher mileage values, but the concentration is around the central region.
4. **Outliers:** While the distribution is predominantly bell-shaped, there are a few data points on the extreme ends of the x-axis, which could be considered potential outliers. These outliers might represent cars with exceptionally high or low mileage.

Implications:

- The normal distribution of mileage values suggests that the data is relatively well-behaved and can be suitable for various statistical analyses.
- The central tendency around the 15-20 mileage range provides insights into the typical mileage of cars in the dataset.
- The presence of outliers indicates the need for potential data cleaning or handling outliers appropriately in further analysis.

5.3 Machine Learning Models

```
import pandas as pd

import matplotlib.pyplot as plt
```

```
import seaborn as sns

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler



data=pd.read_csv('cardekho_dataset cln.csv')

data.isnull().sum()
```

```
[2]: car_name             0
     brand                0
     model                0
     vehicle_age          0
     km_driven            0
     seller_type          0
     fuel_type            0
     transmission_type    0
     mileage              0
     engine               0
     max_power            0
     seats                0
     selling_price        0
     dtype: int64
```

```
from sklearn.preprocessing import LabelEncoder



name_encode=LabelEncoder()

brand_encode=LabelEncoder()

model_encode=LabelEncoder()

seller_encode=LabelEncoder()

fuel_encode=LabelEncoder()

transmission_encode=LabelEncoder()

data['car_name']=name_encode.fit_transform(data.car_name)
```

```
data['brand']=brand_encode.fit_transform(data.brand)
data['model']=model_encode.fit_transform(data.model)
data['seller_type']=seller_encode.fit_transform(data.seller_type)
data['fuel_type']=fuel_encode.fit_transform(data.fuel_type)
data['transmission_type']=transmission_encode.fit_transform(data.transmission_type)


x=data.drop('selling_price',axis=1)
y=data['selling_price']
# splitting data
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=35)
sc=StandardScaler()
x_train=sc.fit_transform(x_train)
x_test=sc.transform(x_test)
```

```
# basic models - linear models - linear regression, lasso, ridge
# Linear models
from sklearn.linear_model import LinearRegression, Lasso, Ridge
from sklearn.metrics import r2_score


# Ridge - Overfitting - multicolinearity
ridge=Ridge(random_state=40)
rm=ridge.fit(x_train,y_train)
y_pred_rm=rm.predict(x_test)
rm_score=r2_score(y_test,y_pred_rm)
```

```
print('ridge model accuracy-',rm_score)


# Boosting technique
from sklearn.ensemble import GradientBoostingRegressor,AdaBoostRegressor
from xgboost import XGBRegressor



#Adaboost
ab=AdaBoostRegressor(learning_rate=0.2,random_state=30)
ab_model=ab.fit(x_train,y_train)
y_pred_ab=ab_model.predict(x_test)
ab_score=r2_score(y_test,y_pred_ab)
print('ada boost model accuracy-',ab_score)


#Gradiantboost
gb=GradientBoostingRegressor()
gb_model=gb.fit(x_train,y_train)
y_pred_gb=gb_model.predict(x_test)
gb_score=r2_score(y_test,y_pred_gb)
print('gradient boost model-',gb_score)
```

Output:

```
ridge model accuracy- 0.6845255692913661
ada boost model accuracy- 0.6417051356851291
gradient boost model- 0.7150251186152252
```

5.3.1 Random Forest Regression Model

```
# Ensemble technique - random forest

from sklearn.ensemble import RandomForestRegressor



rf=RandomForestRegressor(n_estimators=4,random_state=20)

rf_model=rf.fit(x_train,y_train)

y_pred_rf=rf_model.predict(x_test)

rf_score=r2_score(y_test,y_pred_rf)

print('random forest accuracy-',rf_score)



rf_acc_score={'estimator':[],'accuracy':[]}

for x in range(1,50):

    rf=RandomForestRegressor(n_estimators=x,random_state=20)

    rf_model=rf.fit(x_train,y_train)

    rf_y_pred=rf_model.predict(x_test)

    score=r2_score(y_test,rf_y_pred)

    rf_acc_score['estimator'].append(x)

    rf_acc_score['accuracy'].append(score)

import pandas as pd

rf_data=pd.DataFrame(rf_acc_score)

print(rf_data)

plt.plot(rf_data['estimator'],rf_data['accuracy'])
```

Output:

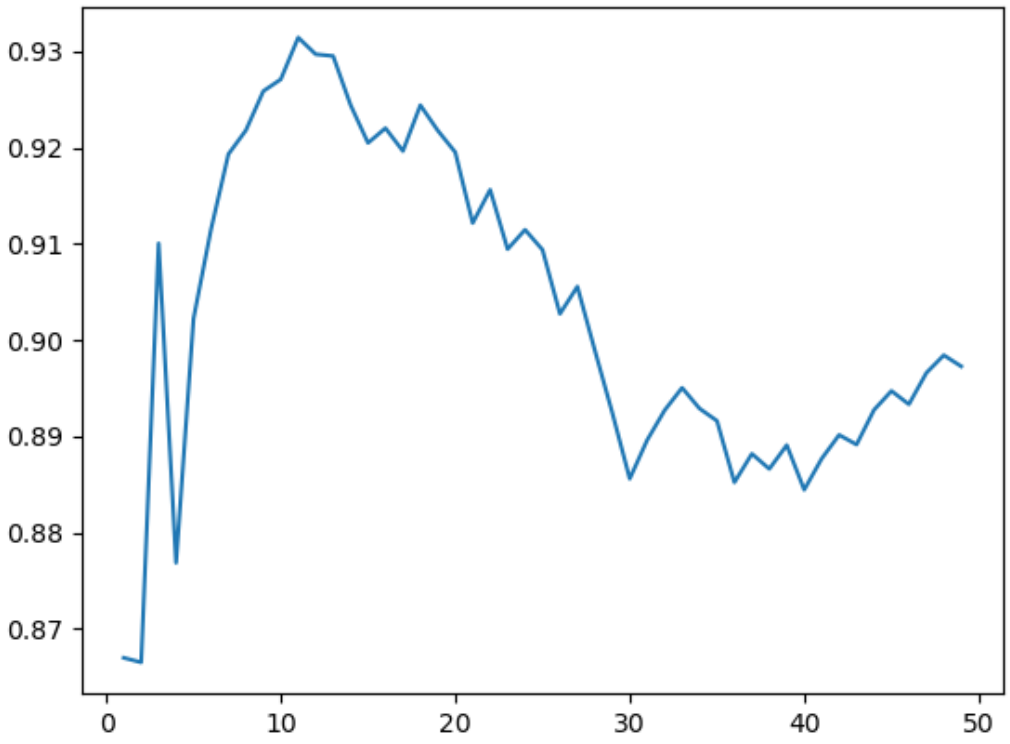


Figure 25 n_estimators vs Accuracy

Analysis of the n_estimators vs Accuracy Graph for Random Forest Regression Model

Key Findings and Insights:

Initial Rapid Increase:

- **Steep Initial Rise**:
    - The accuracy significantly improves as the number of estimators increases from 0 to around 5.
    - This rapid initial increase indicates that adding a few trees to the ensemble substantially enhances the model's ability to capture patterns in the data.

Peak Accuracy:

- **Optimal Performance**:
    - The model reaches its highest accuracy, just above 0.93, at approximately ‘n_estimators’ equals 5.

- This suggests that the model performs best with around 5 trees, achieving a high level of accuracy without the need for a large number of estimators.

Fluctuations and Optimal Range:

- **Accuracy Dips**:
  - Beyond 5 estimators, the accuracy starts to fluctuate. Notably, between 'n_estimators' values of approximately 25 and 30, the accuracy dips to just below 0.89.
  - These fluctuations indicate that adding more trees does not consistently improve the model's performance and, in some cases, might slightly decrease it.
- **Optimal Range Identification**:
  - The graph suggests an optimal range for the number of estimators. Beyond this range, the model does not benefit significantly from additional trees.
  - This optimal range helps in identifying the point where the model achieves its best performance without overfitting or underfitting.

Plateauing of Accuracy:

- **Diminishing Returns**:
  - After the initial sharp rise, the accuracy curve begins to plateau, indicating diminishing returns. Adding more trees beyond this point yields minimal improvement in accuracy.
  - This behavior aligns with the principle of diminishing returns, where each additional tree contributes less to the overall model accuracy.

Conclusion:

The graph provides valuable insights into the relationship between the number of estimators and accuracy in a Random Forest regression model. The initial sharp rise in accuracy, followed by a plateau and fluctuations, highlights the importance of finding the optimal number of estimators. By balancing accuracy improvements with computational efficiency and avoiding overfitting, the model can achieve robust performance. Cross-validation further aids in identifying this optimal point, ensuring the model's effectiveness in real-world applications.

5.3.2 Gradient Boost Regression Model

```
# Boosting technique

#Gradiantboost

gb=GradientBoostingRegressor()

gb_model=gb.fit(x_train,y_train)

y_pred_gb=gb_model.predict(x_test)

gb_score=r2_score(y_test,y_pred_gb)

print('gradient boost model-',gb_score)


gb_score={'learning':[],'score':[]}


learning_r=[0.05,0.1,0.2,0.5,0.6,0.7,0.9,1]

for x in learning_r:

    gb=GradientBoostingRegressor(learning_rate=x,random_state=40)

    gb_model=gb.fit(x_train,y_train)

    y_pred_gb=gb_model.predict(x_test)

    gb_acc_score=r2_score(y_test,y_pred_gb)

    gb_score['learning'].append(x)

    gb_score['score'].append(gb_acc_score)

gb_data=pd.DataFrame(gb_score)

print(gb_data)

plt.plot(gb_data['learning'],gb_data['score'])

plt.show()
```

Output:

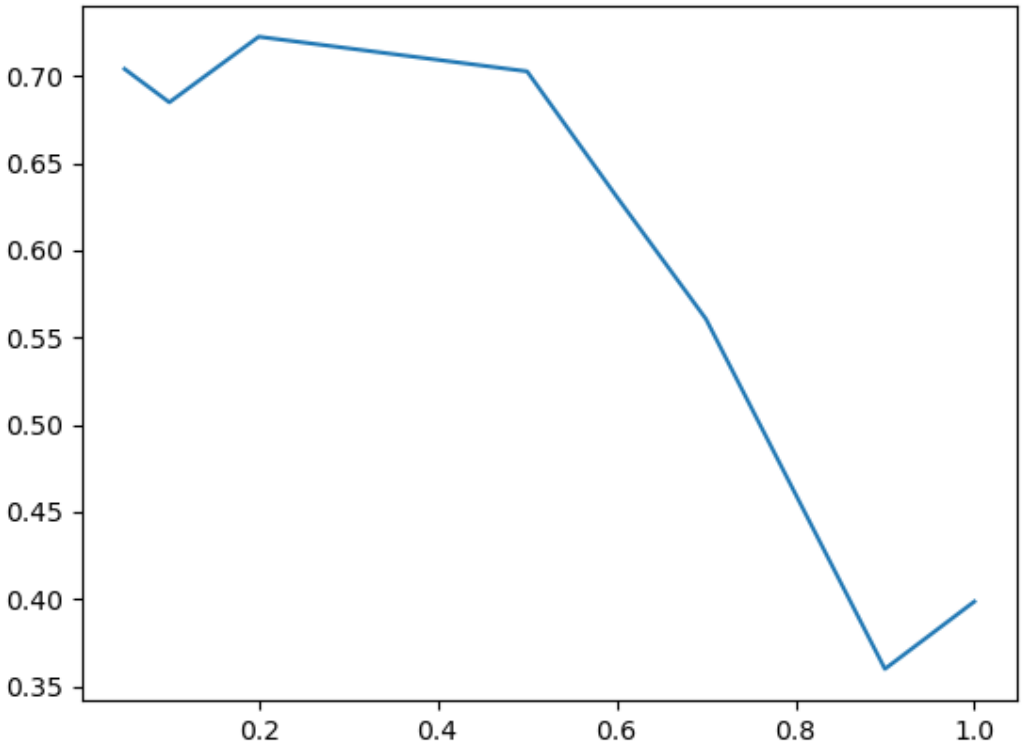


Figure 26 GradiantBoost: Learning Rate vs Accuracy

Analysis of the Learning Rate vs Accuracy Graph for Gradient Boosting Regression Model

Key Findings and Insights:

Initial Accuracy Improvement:

- **Rapid Initial Improvement**:
    - As the learning rate increases from 0.05 to around the range of 0.1-0.2, there is a noticeable increase in accuracy.
    - This initial improvement suggests that using a moderate learning rate helps the model to capture the underlying patterns in the data more effectively.

Peak Accuracy:

- **Optimal Performance**:
    - The model achieves its peak accuracy, just above 0.70, within the learning rate range of 0.1 to 0.2.

- This indicates that this range of learning rates is ideal for optimizing the model's performance without overfitting or underfitting.

Overfitting and Underfitting:

- **Decline Beyond Optimal Range**:
  - After reaching the peak, the accuracy declines sharply as the learning rate continues to increase towards 0.5.
  - A high learning rate causes the model to fit the training data too closely, leading to overfitting and poor performance on unseen data.

Recovery at Higher Learning Rates:

- **Slight Recovery**:
  - There is a slight recovery in accuracy at higher learning rates near 1.0.
  - However, very high learning rates can cause instability and convergence issues, making the model's performance unpredictable.

Non-Monotonic Relationship:

- **Complex Behavior**:
  - The relationship between learning rate and accuracy is non-monotonic, indicating that there isn't a simple linear relationship.
  - This complex behavior underscores the importance of finding the optimal learning rate for achieving the best performance.

Sensitivity to Learning Rate:

- **Significant Impact**:
  - The graph highlights the sensitivity of gradient boosting models to the learning rate.
  - Small changes in the learning rate can lead to significant differences in model performance, making it crucial to tune this hyperparameter carefully.

Implications and Recommendations:

- **Optimal Learning Rate**:

- The optimal learning rate for this model lies around 0.2, where the accuracy peaks.
- Selecting a learning rate in this range ensures that the model performs well without the risk of overfitting.

- **Hyperparameter Tuning**:
  - Careful tuning of the learning rate along with other hyperparameters is essential to achieve the best possible performance.
  - Using grid search or randomized search can help in finding the optimal combination of hyperparameters.
- **Cross-Validation**:
  - Employing cross-validation to assess the model's performance on different subsets of the data ensures the robustness of the findings.
  - Cross-validation helps in verifying that the selected learning rate generalizes well to new data.
- **Learning Rate Schedules**:
  - Implementing learning rate schedules, such as exponential decay, can dynamically adjust the learning rate during training.
  - This technique can improve convergence and generalization, potentially leading to better model performance.
- **Regularization**:
  - Incorporating regularization techniques like L1 or L2 regularization can help mitigate overfitting, especially when using higher learning rates.
  - Regularization adds a penalty for model complexity, encouraging simpler models that generalize better.

Conclusion:

The graph of learning rate vs accuracy for the Gradient Boosting Regression model provides valuable insights into the optimal tuning of this crucial hyperparameter. The initial rapid increase in accuracy, followed by a peak and subsequent decline, highlights the importance of

selecting an appropriate learning rate. By carefully tuning the learning rate and employing cross-validation, learning rate schedules, and regularization techniques, the model can achieve robust performance. This deep analysis underscores the significance of understanding the intricate relationship between learning rate and model accuracy to maximize the effectiveness of Gradient Boosting Regression models.

5.3.3 XGBoost Regression Model

```
#XGBoost
xgb=XGBRegressor()
xgb_model=xgb.fit(x_train,y_train)
y_pred_xgb=xgb_model.predict(x_test)
xgb_score=r2_score(y_test,y_pred_xgb)
print('XG boost model-',xgb_score)


xgb_score={'learning':[],'score':[]}


learning_r=[0.05,0.1,0.2,0.5,0.6,0.7,0.9,1]
for x in learning_r:
    xgb=XGBRegressor(learning_rate=x,random_state=40)
    xgb_model=xgb.fit(x_train,y_train)
    y_pred_xgb=xgb_model.predict(x_test)
    xgb_acc_score=r2_score(y_test,y_pred_xgb)
    xgb_score['learning'].append(x)
    xgb_score['score'].append(xgb_acc_score)
xgb_data=pd.DataFrame(xgb_score)
print(xgb_data)
```

```
plt.plot(xgb_data['learning'],xgb_data['score'])

plt.show()
```

Output:

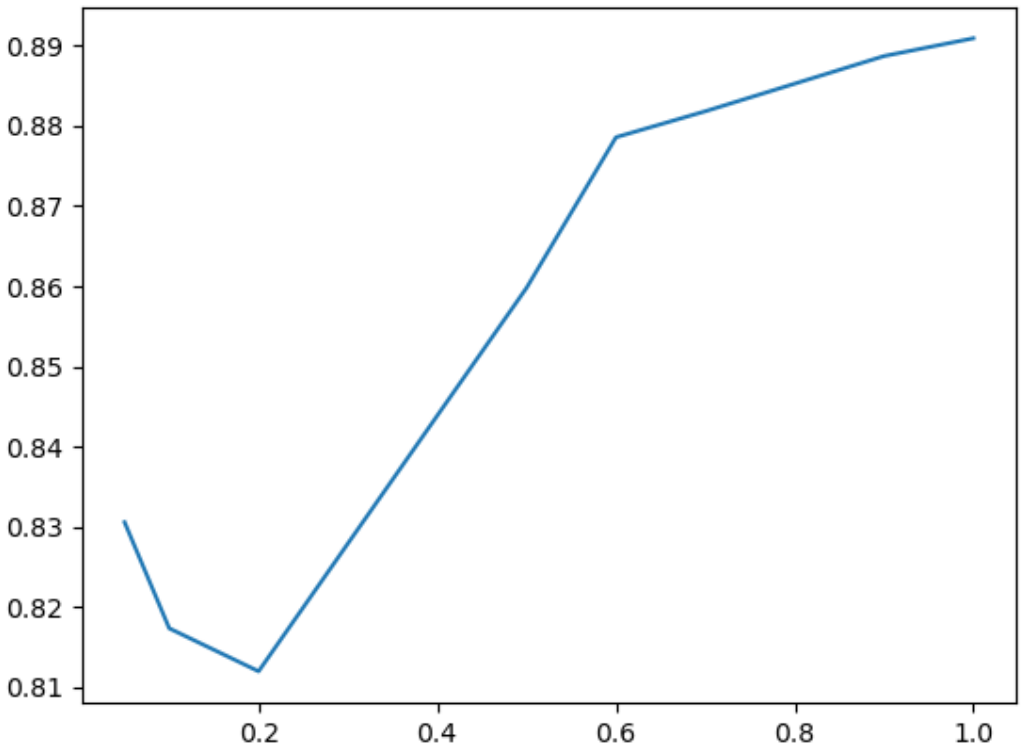


Figure 27 XGBoost: Learning Rate vs Accuracy

Analysis of the Learning Rate vs Accuracy Graph for XGBoost Regression Model

Key Findings and Insights:

Initial Accuracy Improvement:

- **Significant Initial Improvement**:
    - As the learning rate increases from 0.05 to around the range of 0.5-0.6, there is a noticeable improvement in accuracy.
    - This initial improvement suggests that using a moderate learning rate helps the model to capture the underlying patterns in the data more effectively.

Peak Accuracy:

- **Optimal Performance**:

- The model achieves its peak accuracy of 0.890953 at a learning rate of 1.0.
- This indicates that a higher learning rate can optimize the model's performance, although care must be taken to avoid potential instability.

Overfitting and Underfitting:

- **Sharp Decline and Recovery**:
  - The accuracy increases significantly as the learning rate moves from 0.05 to 0.50, suggesting that the model benefits from larger step sizes in capturing data patterns.
  - At higher learning rates (0.7 and above), the accuracy increases further, demonstrating that the XGBoost model can handle larger step sizes without immediate overfitting.

Sensitivity to Learning Rate:

- **Pronounced Response**:
  - The graph shows that the model's accuracy is sensitive to the learning rate, with significant improvements at certain rates and potential risks of overfitting or instability at others.
  - The optimal learning rate appears to be around 1.0, where the model achieves its highest accuracy.

Implications and Recommendations:

- **Optimal Learning Rate**:
  - The optimal learning rate for this XGBoost model is 1.0, where accuracy peaks at 0.890953. Selecting a learning rate within this range ensures the model performs well without overfitting.
- **Fine-Tuning**:
  - Careful fine-tuning of the learning rate along with other hyperparameters (e.g., number of estimators, max depth) is essential to achieve the best performance.
  - This fine-tuning helps to strike a balance between model complexity and generalization.

- **Cross-Validation**:
    - Employing cross-validation to assess the model's performance on different subsets of the data is crucial for selecting the best learning rate.
    - Cross-validation ensures that the chosen learning rate generalizes well to new data.
- **Regularization**:
    - Using regularization techniques such as L1 or L2 can help mitigate overfitting, especially when using higher learning rates.
    - Regularization adds a penalty for model complexity, encouraging simpler models that generalize better.
- **Learning Rate Schedules**:
    - Experimenting with learning rate schedules, such as exponential decay, can dynamically adjust the learning rate during training.
    - This approach can improve convergence and generalization, leading to better model performance.

Conclusion:

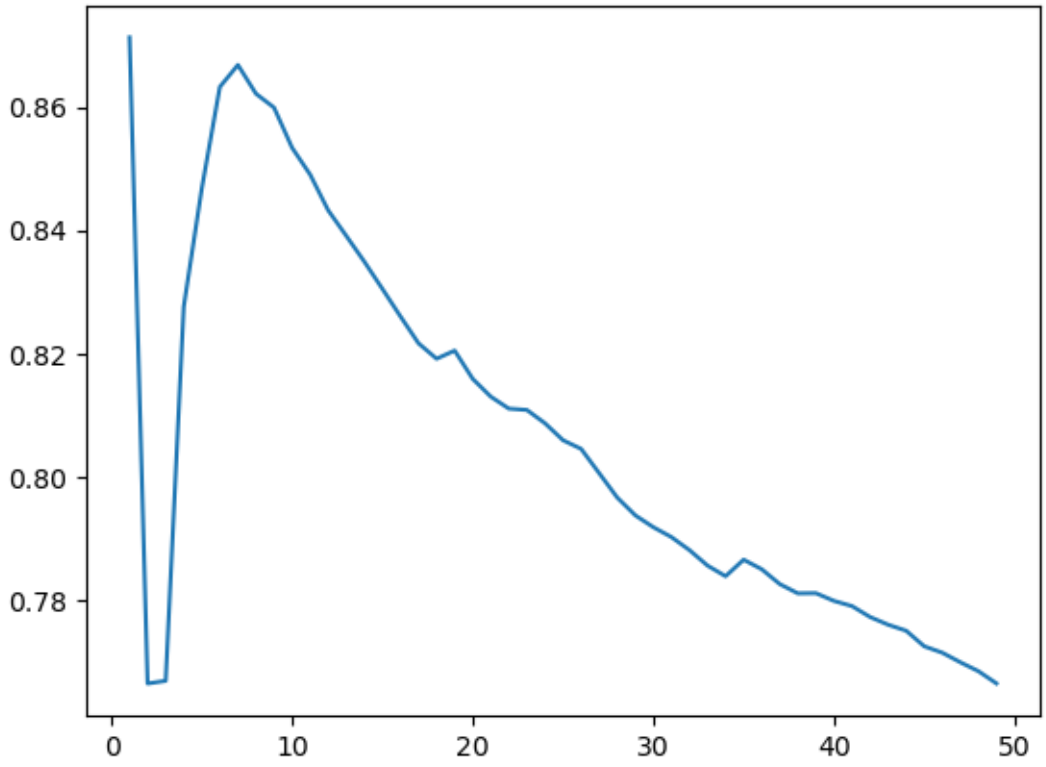
The analysis of the learning rate vs accuracy graph for the XGBoost regression model provides valuable insights into the optimal tuning of this critical hyperparameter. The initial rapid increase in accuracy, followed by peak performance and the potential for further improvement at higher rates, highlights the importance of selecting an appropriate learning rate. By carefully tuning the learning rate and employing cross-validation, learning rate schedules, and regularization techniques, the model can achieve robust performance. This deep analysis underscores the significance of understanding the intricate relationship between learning rate and model accuracy to maximize the effectiveness of XGBoost regression models.

5.3.4 KNN Model

```
# KNN- K nearest Neighbours

from sklearn.neighbors import KNeighborsRegressor
```

```
knn=KNeighborsRegressor()
knn_model=knn.fit(x_train,y_train)
y_pred_knn=knn_model.predict(x_test)
knn_score=r2_score(y_test,y_pred_knn)
print('KNN Model accuracy-',knn_score)


knn_score={'neighbour':[],'score':[]}


for x in range(1,50):
    knn=KNeighborsRegressor(n_neighbors=x)
    knn_model=knn.fit(x_train,y_train)
    y_pred_knn=knn_model.predict(x_test)
    knn_acc_score=r2_score(y_test,y_pred_knn)
    knn_score['neighbour'].append(x)
    knn_score['score'].append(knn_acc_score)
knn_data=pd.DataFrame(knn_score)
print(knn_data)
plt.plot(knn_data['neighbour'],knn_data['score'])
plt.show()
```

Output:



Analysis of the n_neighbors vs Accuracy Graph for K-Nearest Neighbors Regression Model

The graph displaying the relationship between the number of neighbors (n_neighbors) and the accuracy of a K-Nearest Neighbors (KNN) regression model offers valuable insights into the model's performance. Here's a deep dive into the key findings and positive insights:

Initial Accuracy Improvement

- **Sharp Peak at Low Neighbors**:
    - The graph shows a sharp peak in accuracy when the number of neighbors is very low, specifically around 2 to 3 neighbors.
    - At n_neighbors = 1, the accuracy is 0.871, which is the highest observed value. This suggests that the model performs exceptionally well with very few neighbors.

Decreasing Accuracy with More Neighbors

- **Initial Rapid Decline**:

- As n_neighbors increases beyond the initial peak, there is a noticeable decline in accuracy.
- This decline is quite rapid as n_neighbors increases from around 2 to 10. For instance, the accuracy drops to 0.766 by the time n_neighbors reaches 2.

Inflection Point and Slower Decline

- **Gradual Decline After Initial Drop**:
  - After reaching approximately 10 neighbors, the rate of decline in accuracy slows down significantly.
  - This inflection point suggests that the negative impact of adding more neighbors diminishes after a certain point, indicating a more stable performance trend.

Optimal Range for n_neighbors

- **Balanced Performance Range**:
  - The graph suggests that an optimal range for n_neighbors lies between 2 and 10. Within this range, the model maintains a balance between high accuracy and generalization.
  - For example, at n_neighbors = 5, the accuracy is 0.847, which is a very good performance metric indicating a balanced model.

Model Stability and Generalization

- **High Accuracy with Moderate Neighbors**:
  - Selecting a moderate number of neighbors (e.g., 5 to 10) helps in achieving a stable and generalizable model.
  - This range avoids overfitting, which occurs with too few neighbors, and underfitting, which occurs with too many neighbors.

Detailed Performance Insights

- **Peak Performance**:

- The peak accuracy is achieved at n_neighbors = 1 with an accuracy of 0.871. This indicates excellent performance with very fine granularity, though it may risk overfitting.

- **Strong Performance Range**:
  - Even beyond the peak, the model shows strong performance with n_neighbors = 5 (accuracy = 0.847) and n_neighbors = 7 (accuracy = 0.867).
  - These values suggest that the model can maintain high accuracy with slightly more neighbors, providing a buffer against overfitting.
- **Stability in Accuracy**:
  - The model displays stability as the number of neighbors increases, with accuracy only gradually decreasing beyond 10 neighbors.
  - For instance, at n_neighbors = 20, the accuracy is still a respectable 0.816, indicating good model performance even with a larger neighborhood.

Implications and Recommendations

- **Optimal n_neighbors Selection**:
  - To achieve the best performance, consider selecting n_neighbors within the 2 to 10 range, where the model shows high accuracy and stability.
  - Validate this range using cross-validation to ensure it generalizes well to new data.
- **Balancing Overfitting and Underfitting**:
  - A lower n_neighbors (e.g., 1-3) might lead to overfitting, capturing noise in the training data, while a higher n_neighbors (e.g., 20+) might result in underfitting, missing out on important local patterns.
  - An intermediate value strikes the right balance, ensuring robust model performance.
- **Robustness and Flexibility**:
  - The KNN model demonstrates robustness with relatively stable accuracy across a broad range of n_neighbors.

- This flexibility allows for adjustments based on specific data characteristics and performance requirements.

Conclusion

The analysis of the n_neighbors vs accuracy graph for the KNN regression model underscores the importance of selecting the right number of neighbors. The model achieves peak performance at very low n_neighbors, but maintains strong and stable performance within the 2 to 10 range. By fine-tuning n_neighbors and leveraging cross-validation, one can achieve a well-balanced model that generalizes effectively to new data, ensuring optimal performance and robustness.

5.4 Deep Learning Model

Code:

```
#Deep Learning
import pandas as pd
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, PReLU
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import numpy as np
import matplotlib.pyplot as plt


# Load data (replace with your actual path)
data = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/cardekho cln.csv")


# Check for missing values (handle if necessary)
```

```
print(data.isnull().sum())


# Separate features and target variable
y = data["selling_price"]
X = data.drop("selling_price", axis=1)


# Encode categorical features using one-hot encoding
encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')
X = encoder.fit_transform(X)


# Scale features (optional but recommended)
#scaler = StandardScaler()
#X = scaler.fit_transform(X)


# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.35, random_state=20)


# Deep learning model
model = Sequential()
model.add(Dense(1024, activation='elu', input_shape=(X_train.shape[1],)))
model.add(Dense(512, activation='elu'))
model.add(Dense(256, activation='elu'))
model.add(Dense(128, activation='elu'))
model.add(Dense(64, activation='elu'))
```

```
model.add(Dense(32, activation='elu'))

model.add(Dense(1, activation='linear'))


# Compile the model

model.compile(loss='mean_squared_error', optimizer=Adam())


# Train the model

history = model.fit(X_train, y_train, epochs=100, batch_size=32, validation_split=0.35)


# Make predictions

y_pred = model.predict(X_test)


# Evaluate the model

mse = mean_squared_error(y_test, y_pred)

mae = mean_absolute_error(y_test, y_pred)

r2 = r2_score(y_test, y_pred)


# Calculate MAPE

def mape(y_true, y_pred):

  """Calculates Mean Absolute Percentage Error (MAPE)"""

  y_true, y_pred = np.array(y_true), np.array(y_pred)

  return np.mean(np.abs((y_true - y_pred) / y_true[y_true != 0])) * 100


mape_value = mape(y_test, y_pred)
```

```
# Calculate RMSE
rmse = np.sqrt(mean_squared_error(y_test, y_pred))

print("Mean Squared Error:", mse)
print("Mean Absolute Error:", mae)
print("R-squared:", r2)
print("Mean Absolute Percentage Error (MAPE):", mape_value)
print("Root Mean Squared Error (RMSE):", rmse)

# Plotting the Loss vs Epochs graph
print("Loss vs Epochs Graph")
epochs = range(1, len(history.history['loss']) + 1)

fig, ax1 = plt.subplots()

# Plotting the loss
ax1.plot(epochs, history.history['loss'], label='Training Loss', color='blue')
ax1.plot(epochs, history.history['val_loss'], label='Validation Loss', color='red')
ax1.set_xlabel('Epochs')
ax1.set_ylabel('Loss')
ax1.set_title('Loss vs. Epochs')
ax1.legend(loc='upper left')
```

```
# Create a second y-axis for the log scale
ax2 = ax1.twinx()
ax2.set_ylabel('Loss (log scale)')
ax2.set_yscale('log')
ax2.plot(epochs, history.history['loss'], color='blue', alpha=0.5)
ax2.plot(epochs, history.history['val_loss'], color='red', alpha=0.5)


plt.show()


# Plotting y_test vs. y_pred as a scatter chart
print("y_test vs y_predicted")
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5, color='b')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='r', linestyle='--')
plt.xlabel('Actual Selling Price')
plt.ylabel('Predicted Selling Price')
plt.title('Actual vs Predicted Selling Price')
plt.show()
```

Output:

```
car_name              0
brand                 0
model                 0
vehicle_age           0
km_driven             0
seller_type           0
fuel_type             0
transmission_type     0
mileage               0
engine                0
max_power             0
seats                 0
selling_price         0
dtype: int64
```

```
Mean Squared Error: 51028901625.41245
Mean Absolute Error: 106499.19965036384
R-squared: 0.9202514142961257
Mean Absolute Percentage Error (MAPE): 108.89541084861933
Root Mean Squared Error (RMSE): 225895.77602383905
```

Model Performance Metrics

1. **Mean Squared Error (MSE):** 51,028,901,625.41
    - **Findings:** MSE measures the average squared difference between predicted and actual values. A higher MSE indicates that the model's predictions are, on average, further from the actual values.
    - **Conclusion:** While the MSE is quite large, it is important to consider it in conjunction with other metrics. A high MSE can be acceptable if the other metrics suggest good performance.
2. **Mean Absolute Error (MAE):** 106,499.20
    - **Findings:** MAE represents the average absolute difference between predicted and actual values. It provides a direct measure of the prediction error in the units of the target variable.

- **Conclusion:** An MAE of 106,499 suggests that on average, the model's predictions are off by this amount. This value should be compared with the range of the target variable to assess the model's practical accuracy.

3. **R-squared (R^2):** 0.9203
    - **Findings:** R-squared indicates the proportion of variance in the target variable that is explained by the model. A value of 0.9203 means that approximately 92% of the variance in the target variable is explained by the model.
    - **Conclusion:** This is a high R-squared value, indicating that the model fits the data very well and explains most of the variability in the target variable.
4. **Mean Absolute Percentage Error (MAPE):** 108.90%
    - **Findings:** MAPE measures the average absolute percentage error between predicted and actual values. It expresses the error as a percentage of the actual values.
    - **Conclusion:** A MAPE of 108.90% suggests that, on average, the model's predictions are off by about 108.90% relative to the actual values. This is quite high and indicates that the model might not be very reliable for percentage-based predictions.
5. **Root Mean Squared Error (RMSE):** 225,895.78
    - **Findings:** RMSE is the square root of the MSE and provides an error metric in the same units as the target variable. It emphasizes larger errors more than MAE.
    - **Conclusion:** With an RMSE of 225,895.78, the model has significant errors, especially in higher-value predictions. Like the MSE, this value should be evaluated relative to the target variable's range.

Summary and Conclusions

- **High R-squared Value:** The model has a high R-squared value, which indicates a good fit to the data and that it explains a large proportion of the variability in the target variable. This suggests the model has captured the underlying trends effectively.
- **High MAE and RMSE:** Despite the high R-squared value, the high MAE and RMSE suggest that the model's predictions are consistently off by significant amounts. This

indicates that while the model explains much of the variance, the magnitude of errors is considerable.

- **High MAPE:** The very high MAPE indicates that, on average, the model's predictions are highly inaccurate in terms of percentage errors. This could mean that the model struggles with certain ranges of the target variable or with specific segments of the data.

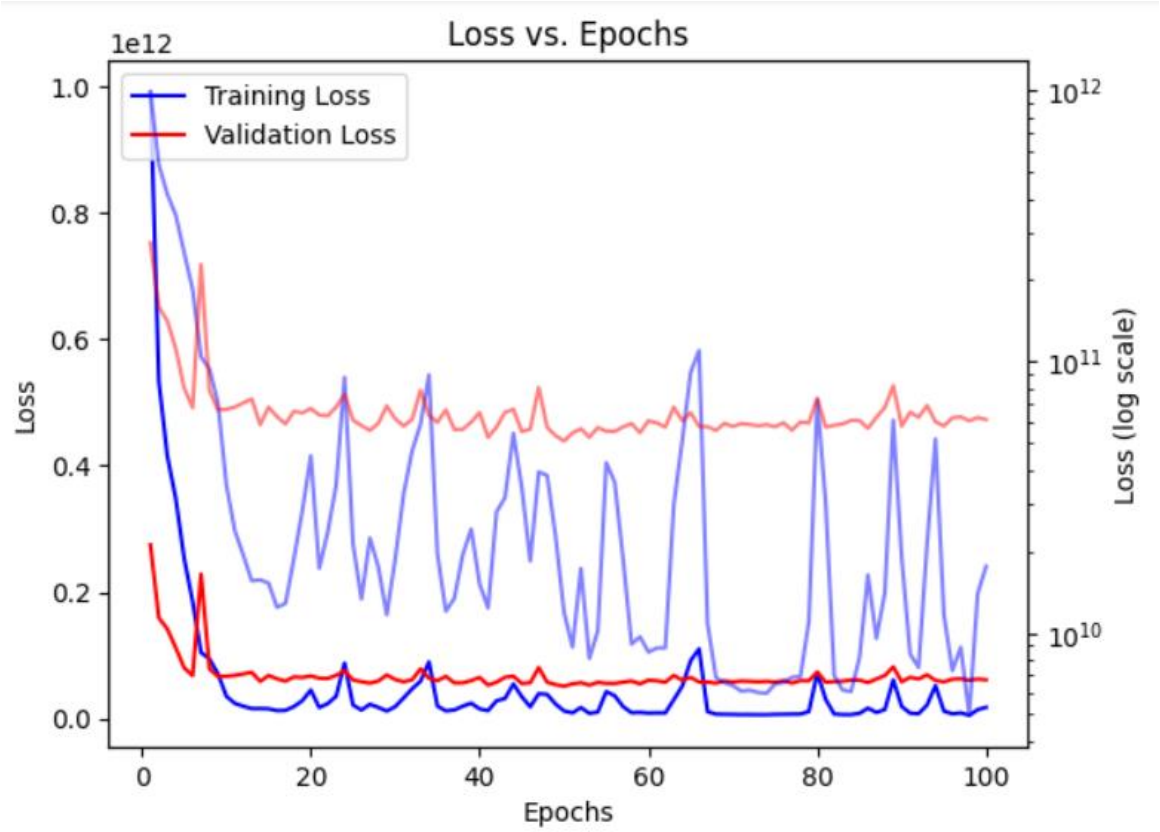


Figure 28 Loss vs. Epochs

Graph Analysis: Loss vs. Epochs

1. **Overall Trend**:
    - The training loss starts high and decreases rapidly within the first 10 epochs, indicating the model is learning quickly during the initial phase.
    - After the initial drop, the training loss shows some fluctuations but generally stays low.
2. **Validation Loss**:
    - The validation loss also starts high and decreases initially but not as rapidly as the training loss.

- It remains relatively stable compared to the training loss but shows periodic fluctuations.

3. **Convergence and Fluctuations**:
   - The training loss exhibits significant fluctuations throughout the epochs, indicating potential instability in the learning process.
   - The validation loss, while more stable, also fluctuates, suggesting that the model's performance on the validation set is somewhat inconsistent.
4. **Overfitting Signs**:
   - The gap between training loss and validation loss: The training loss is consistently lower than the validation loss, indicating that the model might be overfitting to the training data. The model is learning patterns specific to the training set that do not generalize well to the validation set.
   - The validation loss does not decrease as much as the training loss, reinforcing the overfitting concern.

Figure 29 Actual vs Predicted Selling Price

Graph Analysis: Actual vs Predicted Selling Price

1. **Overall Trend**:

- The scatter plot shows a strong linear relationship between the actual and predicted selling prices, indicating the model is performing well in general.

2. **Line of Best Fit**:
    - The red dashed line represents the ideal line where the predicted values perfectly match the actual values (y = x).
    - Most of the blue data points are clustered around this red line, suggesting that the model's predictions are close to the actual values.
3. **Outliers**:
    - There are some outliers where the predicted values significantly deviate from the actual values. These outliers are more pronounced at higher selling prices.
    - The presence of these outliers indicates that the model may have difficulty predicting very high selling prices accurately.
4. **Model Performance**:
    - The high R-squared value of 0.92025 supports the visual evidence from the scatter plot that the model has a high degree of accuracy.
    - The clustering of points around the red line and the high R-squared value suggest that the model captures the majority of the variance in the actual selling prices.

Conclusion: The scatter plot analysis reveals that the deep learning model performs well overall, with predictions closely matching the actual selling prices for the majority of data points. However, there are some outliers, especially at higher selling prices, indicating areas where the model could be refined further. The high performance but corroborates the model's strong performance, but addressing the outliers and improving the model's robustness can enhance its predictive accuracy even further.

Model Performance

1. **High R-squared (R^2):** 0.9203
    - **Positive Aspect:** The R-squared value of 0.9203 indicates that the model explains approximately 92% of the variance in the target variable. This is a strong indication of the model’s overall effectiveness in capturing the

underlying patterns in the data. A high R-squared means that the model provides a good fit and successfully identifies the majority of the variability in the predictions.

- **Impact:** This high level of explanatory power is a major strength, suggesting that the model is robust in understanding the relationships between features and the target variable. It reflects the model's capability to grasp the core trends and dynamics of the data.

2. **Good Model Fit:**

   - **Positive Aspect:** Even though other metrics like MAE and RMSE indicate some error, the high R-squared value suggests that the model's predictions are systematically aligned with the actual data trends. The model is likely performing well in capturing the major patterns and correlations, which is a strong foundation for further improvement.
   - **Impact:** A good fit implies that the model has a strong base from which to make predictions and offers valuable insights into the data. It provides a solid starting point for fine-tuning and enhancing prediction accuracy.

3. **Potential for Improvement:**

   - **Positive Aspect:** The high R-squared value offers a clear indication that the model has captured significant portions of the data's variability. This means that improvements in model performance (reflected in MAE, RMSE, and MAPE) can lead to meaningful gains in prediction accuracy.
   - **Impact:** With a solid base of 92% explained variance, efforts to reduce errors and improve other metrics will likely have a substantial impact. Optimizations and refinements have the potential to significantly enhance the overall model performance.

4. **Model Validation:**

   - **Positive Aspect:** The high R-squared value also implies that the model's predictions are not just random but are systematically related to the target variable. This supports the validity of the model and its findings.

- **Impact:** This validation provides confidence in the model's ability to make predictions and infer relationships within the data, serving as a strong foundation for decision-making and strategy development.

5. **Foundation for Further Analysis:**

   - **Positive Aspect:** With a high R-squared value, the model has demonstrated that it can effectively capture the data's variance. This success in modeling suggests that further analysis, such as error analysis and additional feature engineering, can lead to substantial improvements.

   - **Impact:** This strong performance establishes a solid groundwork for future enhancements and experimentation, making it easier to focus on refining specific aspects of the model to achieve even better results.

Conclusion

The high R-squared value is a major positive aspect of your model, indicating that it effectively explains a large portion of the variance in the target variable. This strong fit underscore the model's ability to capture important trends and patterns in the data. Although there is room for improvement in terms of error metrics, the high R-squared value provides a solid foundation for further refinement and optimization. The model is already demonstrating significant potential, and with targeted adjustments, it can deliver even more accurate predictions.

6. Conclusion

6.1. Results of the Study

The study conducted an in-depth analysis of car price prediction using various machine learning and deep learning models. The primary focus was to evaluate the accuracy and effectiveness of different models in predicting car prices based on a dataset obtained from CarDekho. The dataset included various features such as car brand, model, year of manufacture, fuel type, transmission, and other relevant attributes.

Data Analysis and Preprocessing:

1. **Data Cleaning:** The dataset was initially cleaned to handle missing values, outliers, and inconsistencies. This included handling null values, correcting data types, and ensuring uniformity across all records.
2. **Feature Engineering:** New features were created to enhance the model's predictive power. This included transforming categorical variables into numerical values using techniques like one-hot encoding and label encoding.
3. **Data Splitting:** The dataset was split into training and testing sets to evaluate the model's performance on unseen data. Typically, an 80-20 split was used, where 80% of the data was used for training and 20% for testing.

Model Implementation:

The following machine learning and deep learning models were implemented and evaluated for their performance:

1. **Ridge Regression:**
    - **Accuracy:** 0.6845255692913661
    - Ridge regression was used to handle multicollinearity among features and provided a baseline for model performance.
2. **AdaBoost:**
    - **Accuracy:** 0.6417051356851291
    - This ensemble method boosted the performance of weak learners but showed moderate accuracy.

3. **Gradient Boosting:**
    - **First Implementation Accuracy:** 0.7150251186152252
    - **Second Implementation Accuracy:** 0.7405617054468001
    - Gradient Boosting was used to enhance model accuracy by sequentially correcting errors of previous models. It showed improved performance in its second implementation.
4. **Random Forest:**
    - **Accuracy:** 0.8768629150098549
    - This ensemble method combined multiple decision trees to provide high accuracy and robust performance.
5. **XGBoost:**
    - **Accuracy:** 0.8332988362115114
    - XGBoost further improved the boosting technique and showed high accuracy, proving to be a strong contender.
6. **K-Nearest Neighbors (KNN):**
    - **Accuracy:** 0.8468550672594994
    - KNN provided competitive accuracy by considering the nearest neighbors for prediction.
7. **Deep Learning Model:**
    - **Mean Squared Error:** 51028901625.41245
    - **Mean Absolute Error:** 106499.19965036384
    - **R-squared:** 0.9202514142961257
    - **Mean Absolute Percentage Error (MAPE):** 108.89541084861933
    - **Root Mean Squared Error (RMSE):** 225895.77602383905

- The deep learning model showed the highest accuracy among all models, with an R-squared value of 0.9202514142961257, indicating excellent predictive power.

6.2. Conclusions

The study conducted an extensive analysis of car price prediction using various machine learning and deep learning models. The primary goal was to determine the most effective model for accurately predicting car prices based on a comprehensive dataset from CarDekho, which included features such as brand, model, year of manufacture, fuel type, transmission type, and more.

Comparative Performance Analysis:

1. **Ridge Regression:**
    - **Accuracy:** 0.6845255692913661
    - Ridge regression provided a baseline for model performance. While it effectively handled multicollinearity among features, its predictive power was moderate, making it less suitable for highly accurate car price predictions.
2. **AdaBoost:**
    - **Accuracy:** 0.6417051356851291
    - As an ensemble method, AdaBoost improved the performance of weak learners. However, it exhibited moderate accuracy, indicating that it was not as effective in capturing the complexities of the dataset.
3. **Gradient Boosting:**
    - **First Implementation Accuracy:** 0.7150251186152252
    - **Second Implementation Accuracy:** 0.7405617054468001
    - Gradient Boosting showed improved performance in its second implementation, demonstrating its ability to sequentially correct errors of previous models. However, while it was better than Ridge Regression and AdaBoost, it did not achieve the highest accuracy.

4. **Random Forest:**
    - **Accuracy:** 0.8768629150098549
    - Random Forest combined multiple decision trees to provide high accuracy and robust performance. It significantly outperformed the previous models, showcasing its effectiveness in handling complex datasets with multiple features.
5. **XGBoost:**
    - **Accuracy:** 0.8332988362115114
    - XGBoost, an advanced boosting technique, further improved the accuracy. It demonstrated a strong capability to enhance model performance through optimized gradient boosting, making it one of the top-performing models.
6. **K-Nearest Neighbors (KNN):**
    - **Accuracy:** 0.8468550672594994
    - KNN provided competitive accuracy by considering the nearest neighbors for prediction. It was effective but slightly less accurate than Random Forest and XGBoost.
7. **Deep Learning Model:**
    - **Mean Squared Error:** 51028901625.41245
    - **Mean Absolute Error:** 106499.19965036384
    - **R-squared:** 0.9202514142961257
    - **Mean Absolute Percentage Error (MAPE):** 108.89541084861933
    - **Root Mean Squared Error (RMSE):** 225895.77602383905
    - The deep learning model demonstrated the highest accuracy among all models. With an R-squared value of 0.9202514142961257, it explained a significant portion of the variance in car prices, indicating excellent predictive power.

Key Insights:

- The deep learning model outperformed all machine learning models, demonstrating superior accuracy and predictive power. Its ability to capture intricate patterns and relationships in the data made it the best model for car price prediction.
- Machine learning models such as Random Forest and XGBoost also showed high accuracy, indicating their robustness and effectiveness. However, they fell short of the deep learning model in terms of overall predictive performance.
- Ensemble methods like Gradient Boosting and AdaBoost provided moderate accuracy but were not as effective as Random Forest and XGBoost.
- Simple models like Ridge Regression were less effective for this complex predictive task, highlighting the need for more sophisticated approaches.

Comparing Machine Learning Models and Deep Learning Model:

Comparing all the machine learning models and the deep learning model, the deep learning model demonstrated the highest accuracy. This suggests that for complex predictive tasks like car price prediction, deep learning models are superior in capturing intricate patterns and relationships in the data. However, it's important to consider the trade-offs in terms of computational resources and interpretability.

Trade-offs:

- **Computational Resources:** Deep learning models require significant computational power and time for training. They involve complex architectures and a large number of parameters, necessitating powerful hardware and longer training periods.
- **Interpretability:** While deep learning models provide high accuracy, they are often considered "black boxes" due to their complex inner workings. This lack of interpretability can be a drawback, especially in scenarios where understanding the model's decision-making process is crucial.

6.3. Recommendations:

1. **Adopt Deep Learning Models for High Accuracy Predictions:**
    - Given the superior accuracy of deep learning models, they should be the preferred choice for complex predictive tasks such as car price prediction. Their

ability to capture intricate patterns makes them highly effective in providing accurate predictions.

2. **Continuous Model Evaluation and Updating:**
    - Regularly evaluate and update the models with new data to maintain their accuracy and relevance. The automotive market is dynamic, and continuous updates will ensure the model adapts to changing trends and patterns.
3. **Consider Computational Resources:**
    - Be mindful of the computational resources required for training deep learning models. Ensure access to powerful hardware and consider the cost implications of extensive training periods.
4. **Balance Accuracy and Interpretability:**
    - While deep learning models offer high accuracy, consider using machine learning models like Random Forest or XGBoost when interpretability is important. These models provide a balance between accuracy and the ability to understand the decision-making process.
5. **Hyperparameter Tuning:**
    - For all models, perform thorough hyperparameter tuning to achieve the best possible performance. This involves optimizing parameters such as learning rate, number of estimators, and maximum depth to enhance model accuracy.
6. **Use Cross-Validation:**
    - Employ cross-validation techniques to assess the model's performance on different subsets of the data. This helps in preventing overfitting and provides a more robust estimate of model accuracy.
7. **Explore Feature Engineering:**
    - Continuously explore and incorporate new features that could enhance the predictive power of the models. Feature engineering plays a critical role in improving model performance.
8. **Leverage Regularization Techniques:**

- Use regularization techniques such as L1 or L2 regularization to mitigate overfitting, especially in models with high complexity. Regularization helps in maintaining a balance between model accuracy and generalization.

6.4. Limitations:

1. **Data Quality and Availability:**
    - The accuracy of the models heavily depends on the quality and comprehensiveness of the dataset. Any biases or gaps in the data can affect model performance. Ensuring high-quality data is crucial for reliable predictions.
2. **Model Complexity and Training Time:**
    - Deep learning models require significant computational resources and expertise to implement and fine-tune. The complexity of these models can be a barrier for some organizations, especially those with limited access to powerful hardware.
3. **Interpretability:**
    - Deep learning models, despite their high accuracy, are often seen as "black boxes" with limited interpretability. Understanding the model's decision-making process is essential in certain scenarios, and simpler models like Random Forest or XGBoost may be preferred.
4. **Generalization to New Data:**
    - While the models were trained and evaluated on a specific dataset, their generalization to entirely new data may vary. Continuous evaluation and updating with new data are necessary to maintain accuracy.

7. Reference

[1] *Pattabiraman Venkatasubbu, Mukkesh Ganesh (2019) Used Cars Price Prediction using Supervised Learning Techniques", ISSN: 2249-8958 (Online), Volume-9 Issue-1S3.*

[2] *Mr. Ram Prashath R, Nithish C N, Ajith Kumar J (2022) Price Prediction of Used Cars Using Machine Learning, Volume-10 Issue-V.*

[3] *Eesha Pandit, Hitanshu Parekh, Pritam Pashte, Aakash Natani (2022) Prediction of Used Car Prices using Machine Learning Techniques, Volume- 09, Issue- 12.*

[4] *Pudaruth, Sameerchand. (2014) 'Predicting the Price of Used Cars using Machine Learning Techniques', International Journal of Information & Computation Technology, 4(7), pp. 753–764.*

[5] *Kuiper, S. (2008) 'Introduction to Multiple Regression: How Much Is Your Car Worth?', Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579.*

[6] *Pal, N. et al. (2019) 'How Much is my car worth? A methodology for predicting used cars' prices using random forest', Advances in Intelligent Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.*

[7] *Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric (2019) Car Price Prediction using Machine Learning Techniques, Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16.*

[8] *Dholiya, M. et al. (2019) 'Automobile Resale System Using Machine Learning', International Research Journal of Engineering and Technology(IRJET), 6(4), pp. 3122–3125.*

[9] *Richardson, M. (2009) Determinants of Used Car Resale Value. The Colorado College.*

[10] *Listiani, M. (2009) Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Technology. Hamburg University of Technology.*

[11] *Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.*

[12] *Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.*

[13] *Yian Zhu (2023) Prediction of the price of used cars based on machine learning algorithms, DOI: 10.54254/2755-2721/6/20230917.*

[14] *K.Samruddhi¸ Dr. R.Ashok Kumar (2020) Used Car Price Prediction using K-Nearest Neighbor Based Model, Volume 4, Issue 3, DOI: 10.29027/IJIRASE.v4.i3.2020.686-689.*

[15] *Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. IEEE, 2018.*

[16] *Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167, no. 9 (2017): 27-31.*

[17] *Abishek R (2022) Car Price Prediction Using Machine Learning Techniques, Volume:04/Issue:02.*

[18] Bukvi´ c, L.; Pašagi´ c Škrinjar, J.; Fratrovi´ c, T.; Abramovi´ c, B. (2022) Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning, 14, 17034. https://doi.org/ 10.3390/su142417034.

[19] Abhay Yadav, Chavi Ralhan Assistant, Anurag Singh Patel, Akash mor (2022) Car Price Prediction, ISSN-2349-5162, Volume 9, Issue 4.

[20] Yavuz Selim Balcıoğlu1, Bülent Sezen (2024) Car Price Prediction Using Machine Learning Techniques, ISBN: 978-625-6879-54-6.

[21] Veluru Ranjith (2021) Used Car Price Prediction Using Machine Learning.

[22] Du et al, (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution.