CAPSTONE PROJECT: PREDICTING USED CAR PRICES

Machine Learning Exercise

By Vivek Mistry

Introduction

- Overview of the problem
- Approach for the solution
- Key findings and insights
- Business recommendations

Problem and Objective:

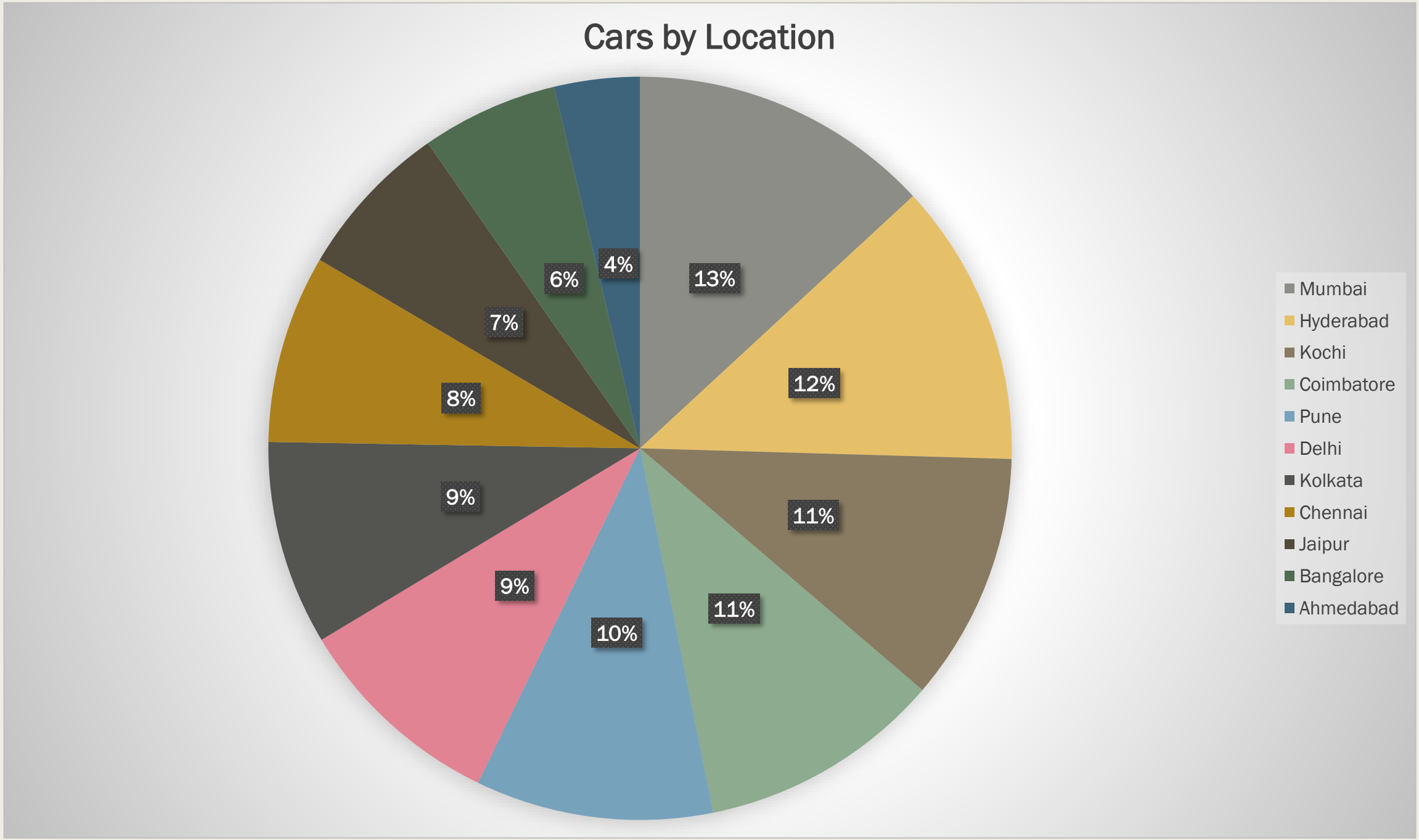
- The demand for used cars in the Indian market is huge these days. Because new car sales have slowed. Recently, the used car market has continued to grow in recent years and is now bigger than the new car market. Cars4u needs provide a more accurate and reliable way to estimate the value of a Used Car to price the cars accordingly.
- Come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing.

The dataset contains 7253 records with 14 columns of fields following attributes:

- **S.No.:** Serial Number
- **Name:** Name of the car which includes Brand name and Model name
- **Location:** The location in which the car is being sold or is available for purchase (Cities)
- **Year:** Manufacturing year of the car
- **Kilometers_driven:** The total kilometers driven in the car by the previous owner(s) in KM
- **Fuel_Type:** The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG)
- **Transmission:** The type of transmission used by the car (Automatic / Manual)
- **Owner:** Type of ownership
- **Mileage:** The standard mileage offered by the car company in KMPL or KM/KG
- **Engine:** The displacement volume of the engine in CC
- **Power:** The maximum power of the engine in BHP
- **Seats:** The number of seats in the car
- **New_Price:** The price of a new car of the same model in INR 100,000
- **Price:** The price of the used car in INR 100,000

Solution Design:

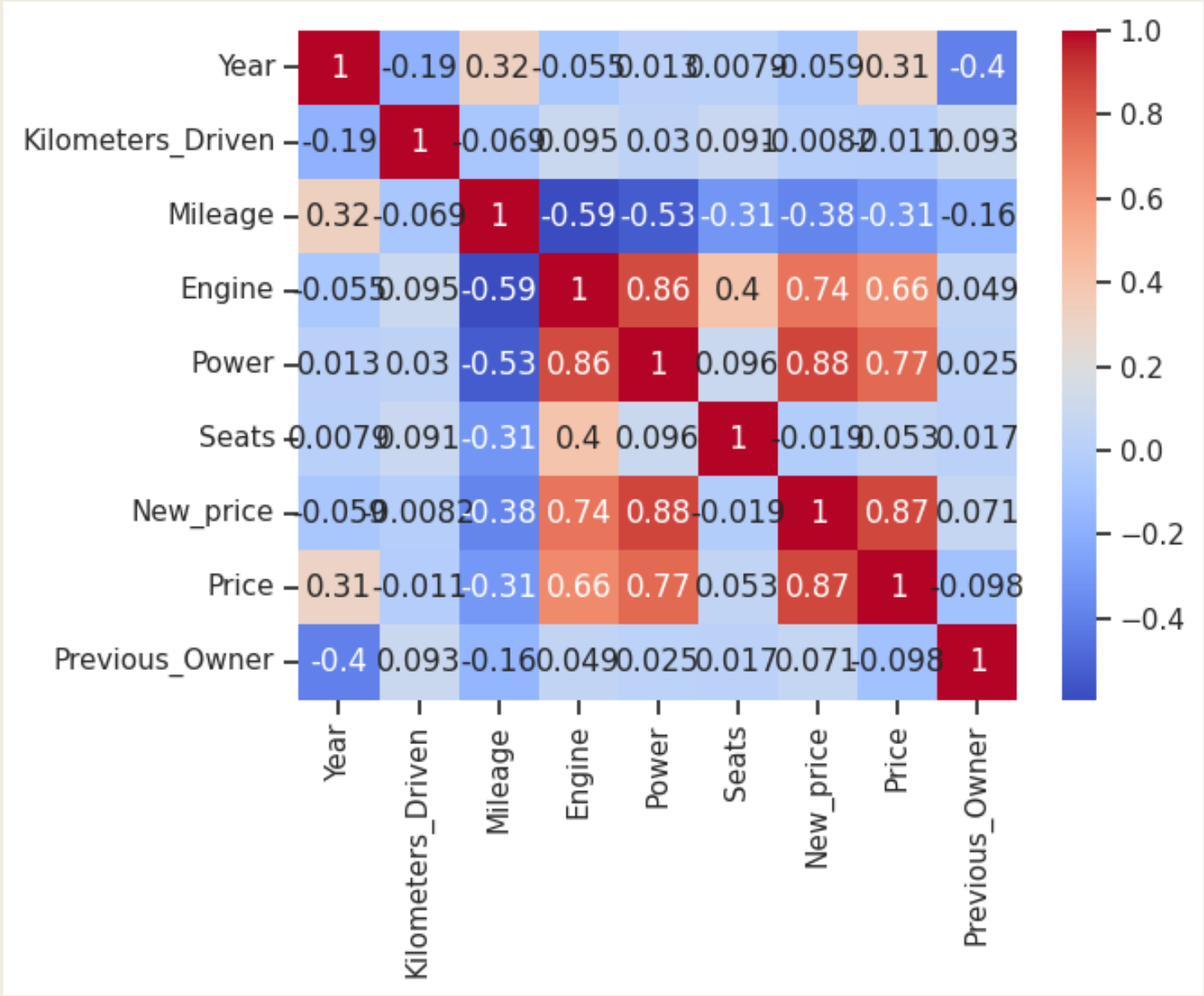
- Use regression models to help design a solution as
  - *Problem involves predicting a continuous target variable (car prices).*
  - *Regression models are well-suited for such tasks*
  - *Regression models can provide insights into the factors that affect car prices and help in pricing strategies.*
- Data Preprocessing Steps were used for handling missing values, outliers, inconsistencies, handling skewness and other steps.
- Data was also populated based on name of the car.
- Train various models with regression and identify which is the more performant model for more accurate predictions.

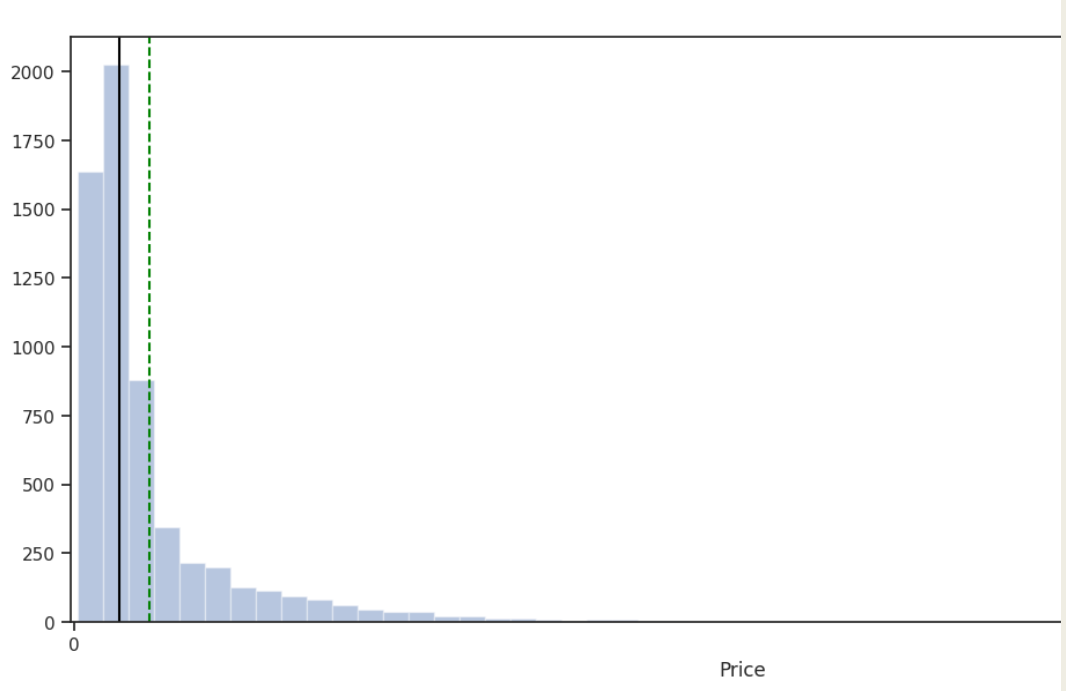

Cars by Location
13%
12%
11%
11%
10%
9%
9%
8%
7%
6%
4%
Mumbai
Hyderabad
Kochi
Coimbatore
Pune
Delhi
Kolkata
Chennai
Jaipur
Bangalore
Ahmedabad

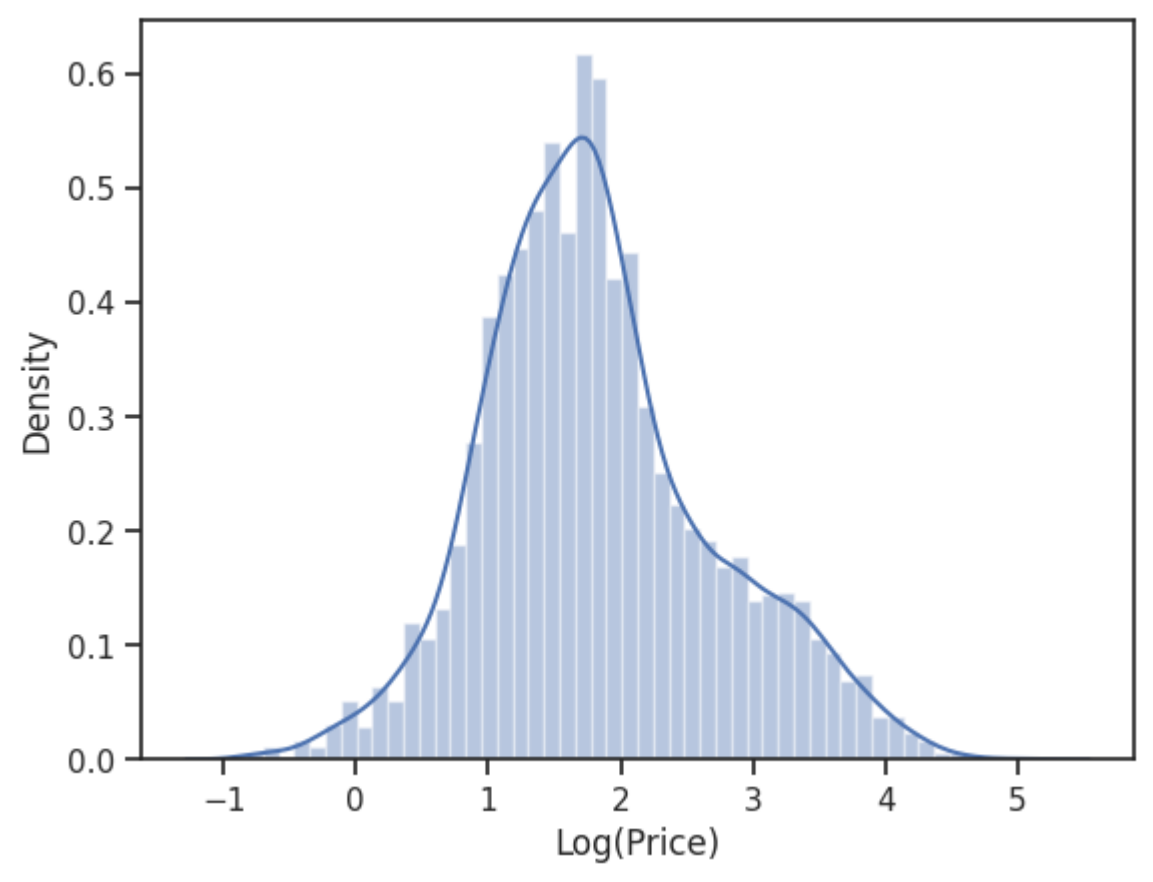
Clean Dataset Details

Metric	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	New_price	Price	Previous_Owner	Kilometers_Driven_Log	Price_Log
count	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00	6018.00
mean	2013.36	57668.05	18.14	1620.38	112.58	5.28	20.19	9.47	1.20	10.76	1.82
std	3.27	37878.78	4.58	600.66	53.66	0.80	22.79	11.17	0.46	0.71	0.87
min	1998.00	171.00	0.00	72.00	34.20	2.00	3.91	0.44	1.00	5.14	-0.82
25%	2011.00	34000.00	15.17	1198.00	74.00	5.00	8.80	3.50	1.00	10.43	1.25
50%	2014.00	53000.00	18.16	1493.00	92.85	5.00	11.67	5.64	1.00	10.88	1.73
75%	2016.00	73000.00	21.10	1984.00	138.10	5.00	16.95	9.95	1.00	11.20	2.30
max	2019.00	775000.00	33.54	5998.00	560.00	10.00	230.00	160.00	4.00	13.56	5.08

correlation matrix using a heatmap




2000
1750
1500
1250
1000
750
500
250
0
0
Price


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
−1
0
1
2
3
4
5
Log(Price)

Data Cleanup Insights

- Using the name of the car filled in data for seats, power, engine, new_price and price.
- Similar name for vehicles with similar power and seats when you look up internet
- Engine specs on names for example
  - *Maruti Swift 1.3 Vxi – 1.3 is 1300cc engine spec*
- Dropped data stayed null for price however used median value for remain colums.

Models investigated

- Linear Regression:
  - *Simple and interpretable model that captures linear relationships between features and car prices.*
- Ridge Regression:
  - *Regularized regression model that addresses multicollinearity and improves generalization.*
- Decision Tree:
  - *Non-linear model that captures complex relationships and interactions between features.*
- Random Forest:
  - *Ensemble of decision trees that provides robust predictions, handles non-linearity, and performs feature importance analysis.*
- Tuned Decision Tree:
  - *Fine-tuned version of the decision tree model to improve performance and mitigate overfitting.*
- Tuned Random Forest:
  - *Fine-tuned version of the random forest model to improve performance and mitigate overfitting.*

Model Performance:

Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
Linear Regression	0.85	0.87	4.26	4.09
Ridge Regression	0.85	0.86	4.32	4.10
Decision Tree	1.00	0.84	0.02	4.48
Random Forest	0.98	0.89	1.71	3.70
Tuned Decision Tree	0.95	0.87	2.53	4.06
Tuned Random Forest	0.96	0.88	2.13	3.79

Model Performance Cont.

- Random Forest performs relatively better among the chosen techniques, with the highest R-squared value on the test set.
- Decision Trees have the highest R-squared value on the training set but lower performance on the test set, indicating overfitting.
- Linear Regression and Ridge Regression have similar R-squared values on both the training and test sets, indicating consistent performance. However, their R-squared values are lower compared to the Decision Trees and Random Forest techniques.

Hyperparameter Tuning

- Hyperparameter tuning is an important step to optimize the performance of machine learning models.
- The tuned models showed improved performance compared to their default counterparts.
- R-squared values increased, indicating better fit to the data.
- RMSE values decreased, indicating reduced error in predicting car prices.
- Both Tuned Decision Tree and Tuned Random Forest improved

Insights

- Features Importance : "Power," "Year," and "New_price“
- Newer cars with higher power and a higher starting cost tend to have higher prices in the used car market.
- Engine and Mileage: Cars with larger engines and better mileage generally cost more
- The car's location ("Location") plays a significant role in determining its price.
- Brand reputation, perceived quality, and consumer preferences play a crucial role in determining used car prices.
- Random Forest model is better suited for predicting used car prices in this dataset

Final Proposed Model

- Random Forest model as the ultimate resolution based on the investigation and modeling fit and score

Random Forest model benefits

- the Random Forest model exhibits the highest R-squared values
- Consistent performance on training and test sets
- manage missing data and outliers, minimising the need for labor-intensive data preprocessing.
- Provide more precise forecasts, it uses the combined knowledge of several trees.

Business Recommendations

- Pricing Strategy: Consider feature like power, year and other to when setting prices
- Brand Positioning: Leverage the influence of brand on used car prices.
- Market Segmentation: Recognize the regional variations in used car prices and target marking accordingly. For example more ads at Kolkata, Hyderabad, Coimbatore, and Bangalore to due higher price from their region.
- Inventory Management: Manage the inventory of used cars based on acquiring vehicles with these desirable features

Conclusions

- power, year, new price, engine size, mileage, location, and brand reputation are significant drivers of used car prices
- The Random Forest regression model emerged as the best-performing model
- Limitations and Potential Risks with market trends, economic conditions, and individual car condition
- Incorporating real-time data and sentiment analysis from online platforms would help with more customer and market insights
- A streamlined real-time data collection with an online platform/service, can further refine the models' predictive capabilities.
- Further more converting the prediction service into a web service for better access web application