MASTERCLASS FOR BUSINESS ANALYTICS

BUSM-131

USED CAR LISTING PRICE PREDICTION

SNEHAL SRIVASTAVA

TEAM-13

220316567

Table of Contents

I. ABSTRACT

This study focuses on developing a tool using machine learning to predict the listing price of used cars as per historical data. It aims to solve the ambiguity that sellers and buyers face while selling or buying a used car at a certain listed price. The dataset used for this purpose was published on Craigslist.org and shows the listing price of used cars listed over the years. It gives us information about the used cars up for sale and their prominent features like paint colour, manufacturer, number of cylinders, etc. After performing several ML models, XGBoost has given us key results with highest accuracy for prediction amongst all. This study included intensive exploratory data analysis and heavy data cleaning.

II. INTRODUCTION

In the UK, second-hand cars are selling more than new cars, with a 3.7% increase in used car sales in the first quarter of 2021, compared to a 12.4% decrease in new car sales over the same period (Society of Motor Manufacturers and Traders, 2021). The popularity is increasing due to economic uncertainty, changing preferences, and advancements in technology. Second-hand cars are more affordable than new cars, and they are becoming more reliable thanks to technological advancements. As these factors continue to shape the market, it is likely that the popularity of second-hand cars will continue to grow.

Machine learning (ML) methods have been utilized in many studies to build models that can accurately predict the price of a used car based on various factors such as age, mileage, condition, and features. One such study conducted by Xu et al. (2020) used a deep neural network model to predict the listing price of used cars in China. They found that their model outperformed traditional linear regression models, achieving a prediction accuracy of 85.5%.

Another study by Shahriar et al. (2020) compared various ML models for predicting the price of used cars in the US market. They found that random forest and gradient boosting models outperformed other models such as linear regression and neural networks.

A third study by Joshi et al. (2021) used a hybrid model that combined deep learning and gradient boosting to predict the listing price of used cars in the Indian market. The authors found that their model achieved a prediction accuracy of 90%, outperforming other traditional models.

Lastly, Huang et al. (2021) developed a hybrid model that combines gradient boosting decision tree and convolutional neural network algorithms to predict used car prices. They used a dataset of over 150,000 used car listings and achieved an average error rate of 0.05%.

Overall, these studies demonstrate the effectiveness of ML methods in predicting the listing price of used cars. The models developed by these researchers can help inform pricing decisions for used car sellers and provide valuable insights for buyers.

Inspired by the various studies and research carried out, our project aimed to use ML models to predict listing prices of used cars. Craigslist has the most used cars for sale in the world. However, finding them all at the same time is extremely difficult. Our dataset aids in the collection of aggregate data for all used cars in the USA.
This report elaborates on the steps involved in developing an ML model for prediction with a significant level of accuracy. It includes an elaborative exploratory data analysis, pre-processing data, building pipelines, testing ML models and hyper tuning to get a robust framework.

III. BUSINESS PROBLEM

The used car market has been expanding rapidly in recent years, with increased buyers and sellers turning to online marketplaces to make their transactions. While these marketplaces have their advantages, they also have their limitations, which can lead to problems for both sellers and buyers.
One major limitation of marketplaces for used cars is the lack of transparency in the pricing process. Sellers often struggle to determine the fair market value of their vehicles, which can lead to overpricing or under-pricing. This

can result in the vehicle sitting on the market for an extended period or being sold for less than its worth. Buyers, on the other hand, may have difficulty in finding the right vehicle for their budget, or may end up paying too much for a vehicle that is not worth the price (Cox Automotive, 2021).
Sellers also face problems in terms of presenting their vehicles in the best light. They may not have the resources or expertise to make necessary repairs and maintenance, or to provide detailed information about the vehicle's history. Buyers may therefore be hesitant to make a purchase, leading to prolonged negotiations and ultimately, a loss of potential sales (iSeeCars.com).
Buyers, on the other hand, face a distinct set of challenges. They may not have the knowledge or expertise to evaluate the condition of a vehicle properly, which can lead to costly repairs and maintenance down the road. Additionally, they may not have access to a wide range of options or may have difficulty in finding a vehicle that meets their specific needs and budget (Autolist.com).

To address these challenges, a valuation tool to predict the listing price for used cars can be incredibly beneficial. Such a tool can provide sellers with an accurate and objective assessment of their vehicle's worth, based on a range of factors such as mileage, age, condition, and market trends. This can help sellers determine the right price for their vehicle and avoid under-pricing or overpricing.

Similarly, buyers can use such tools to evaluate the fairness of a seller's asking price and make more informed purchasing decisions. They can also access a wider range of options and filter their search results based on their specific needs and budget (CarGurus.com).

Overall, while marketplaces for used cars have their limitations, the use of valuation tools can help to mitigate some of these challenges and ensure a more transparent and efficient buying and selling process.

Companies Solving Similar Problems:

In recent years, many companies have used ML to predict listing prices for used cars. They have developed algorithms that analyse various data points, including historical transaction data, market trends, and consumer behaviour, to predict the fair market value of used cars. The results of these efforts have been promising.

Here are some examples of companies that use machine learning to predict the value of used cars:

- CarGurus (2019) uses a proprietary algorithm to analyse millions of data points for prediction of fair market value of used cars. In a study, they found that their ML model predicted fair market value with an average accuracy of within 2.4% of the final transaction price.
- A study by Google (2020) states that Edmunds uses machine learning to analyse market trends and consumer behaviour to predict the value of used cars. The company’s ML model predicted the value of used cars with an accuracy rate of 90%.
- Vroom (2021), an online used car retailer, uses ML to analyse numerous factors, including market demand, condition, and mileage for their prediction. Their ML model has achieved an accuracy rate of over 95% when predicting the value of used cars.
- Microsoft’s study in 2019 stated that Autotrader, another online used car retailer, uses ML to analyse historical transaction data, market trends, and consumer behaviour to predict the value of used cars. Their ML model predicted the value of used cars with an accuracy rate of 90%.

Overall, the use of ML to predict the value of used cars is a promising innovative technology that has the potential to benefit both sellers and buyers.

IV. DATA, EDA, AND METHODS

The dataset scraped from Craigslist has 426880 rows and 26 features, out of which 14 are categorical, 7 are numerical and 5 text datatypes. The data lasts from the year 1900 till 2022.

Exploratory data analysis (EDA), which frequently makes use of data visualization techniques, is used to examine and study data sets and summarize their key properties.
It can aid in better understanding data patterns, spotting outliers or unusual events, and spotting intriguing relationships between the variables.

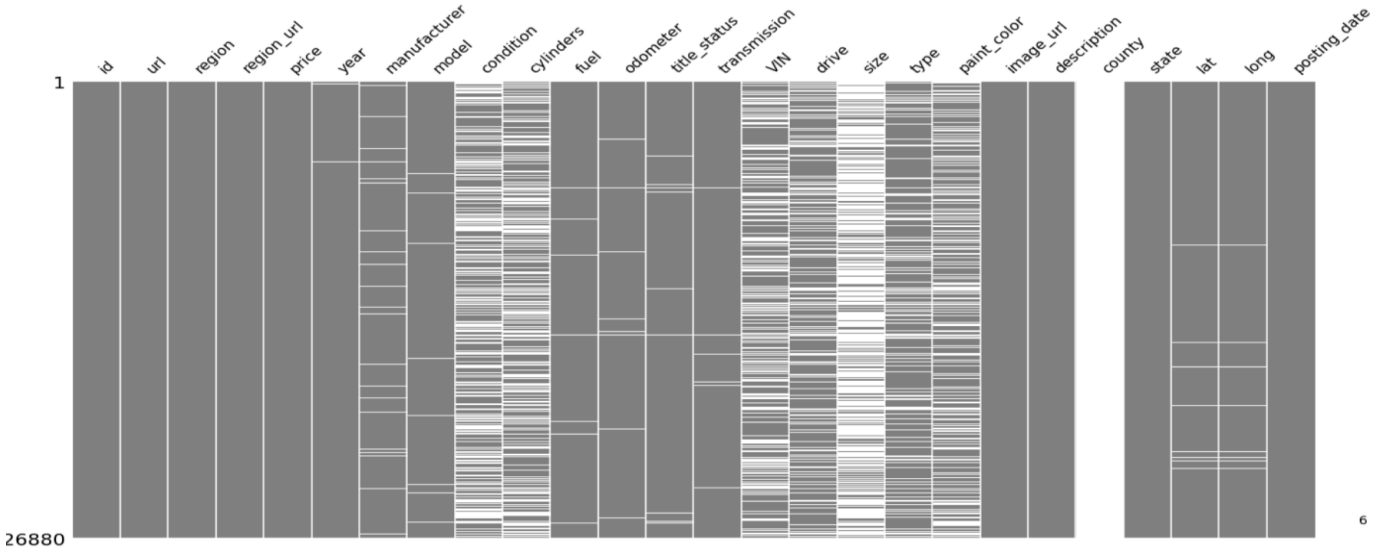

Figure 1

The missingno library is used to visualize missing values in the dataset as a matrix.
After plotting the raw data in terms of various features of the cars, we produced this heat map (Figure 1). This visual shows the missing values in the data for each of the columns. We can see here that that column county has no data, and the column size has a lot of missing values. The data is also irregular in the columns condition, cylinder, VIN, drive, type, and paint colour. The following represents the percentage of null values present in the columns:
Condition ~ 37%
Cylinders ~ 40%
Drive ~ 30%
Type ~ 21 %
Paint Color ~ 29%

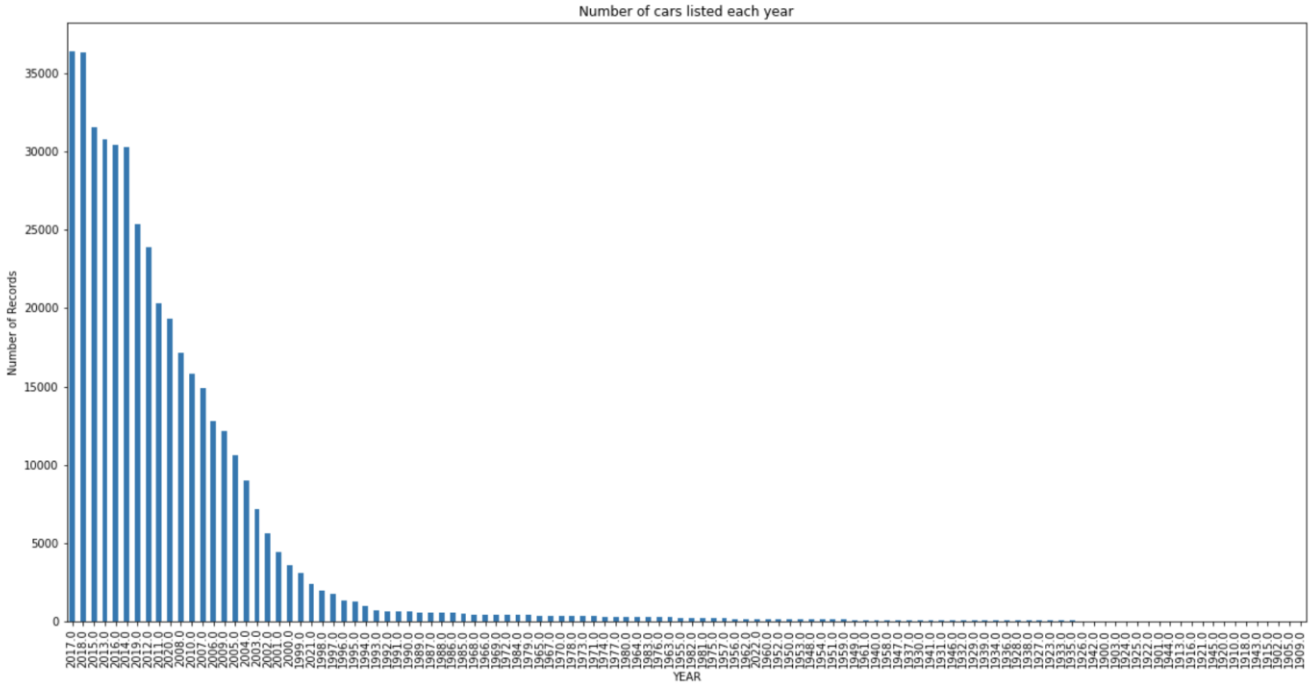

Figure 2

This visual (Figure 2) shows the number of records in each year, starting from the year 1900 till the year 2022. It depicts the number of used cars listed for sale each year. We see that there is negligible to no data available for listed cars during the initial years. The used cars listed from the year 1990 onwards, make for a considerable data for our analysis. This trend might be seen due to the growing popularity of the buying and selling of used cars only in recent years whereas earlier people preferred buying new cars instead of used ones.
The sweetviz library is used to generate detailed reports for the dataset, including descriptive statistics and visualizations.

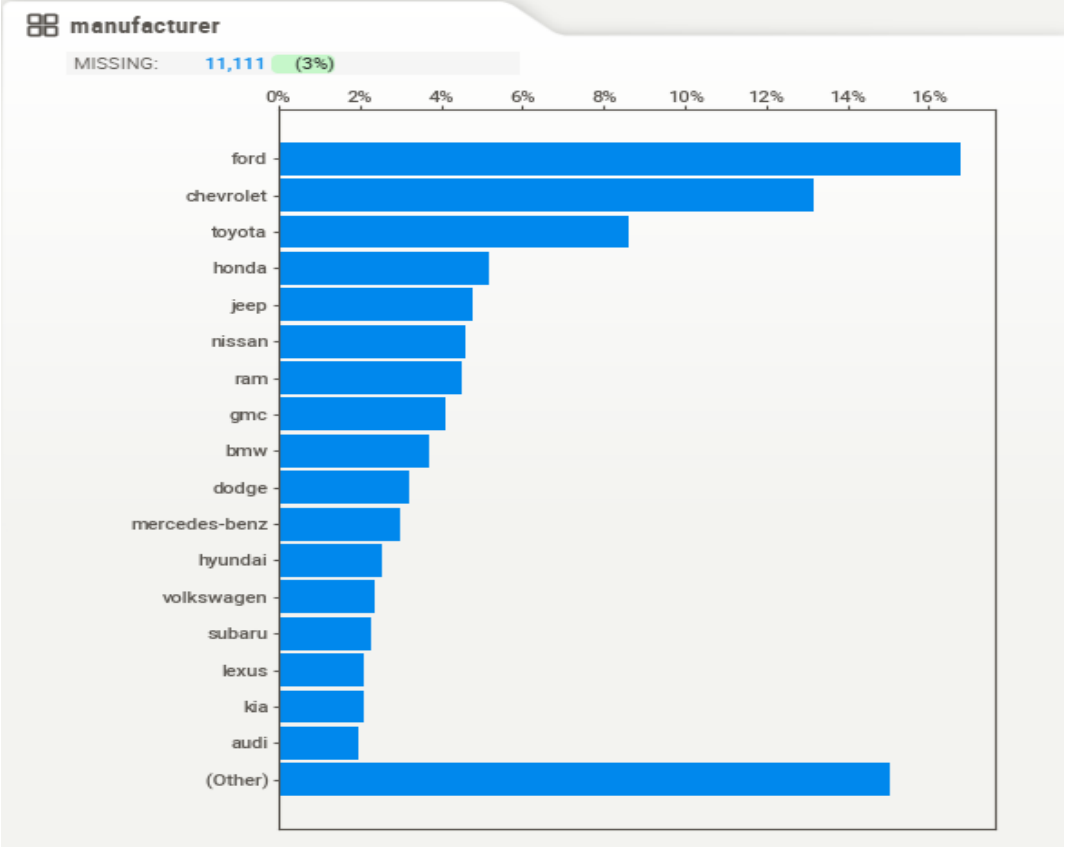

Figure 3

Figure 3 represents the number of cars of different manufacturers listed for resale. From the data, we can see that most cars are manufactured by Ford. The second most common manufacturer in our dataset is Chevrolet. Looking at this data we can assume about the target audience.

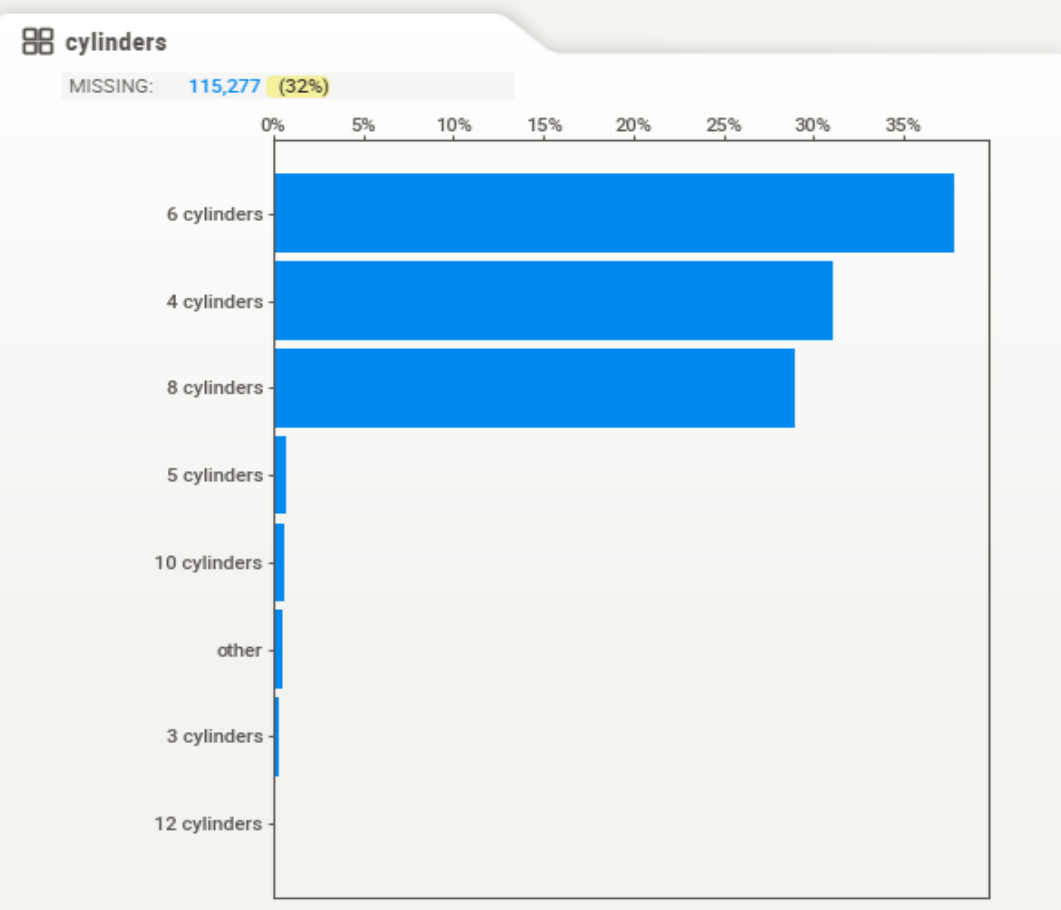

Figure 4

This graph in Figure 4 shows the number of cylinders in different cars listed for resale. From this graph we can see that most cars have 6 cylinders and hardly any car has 12 cylinders. This is a major feature that the customers consider while buying a car.

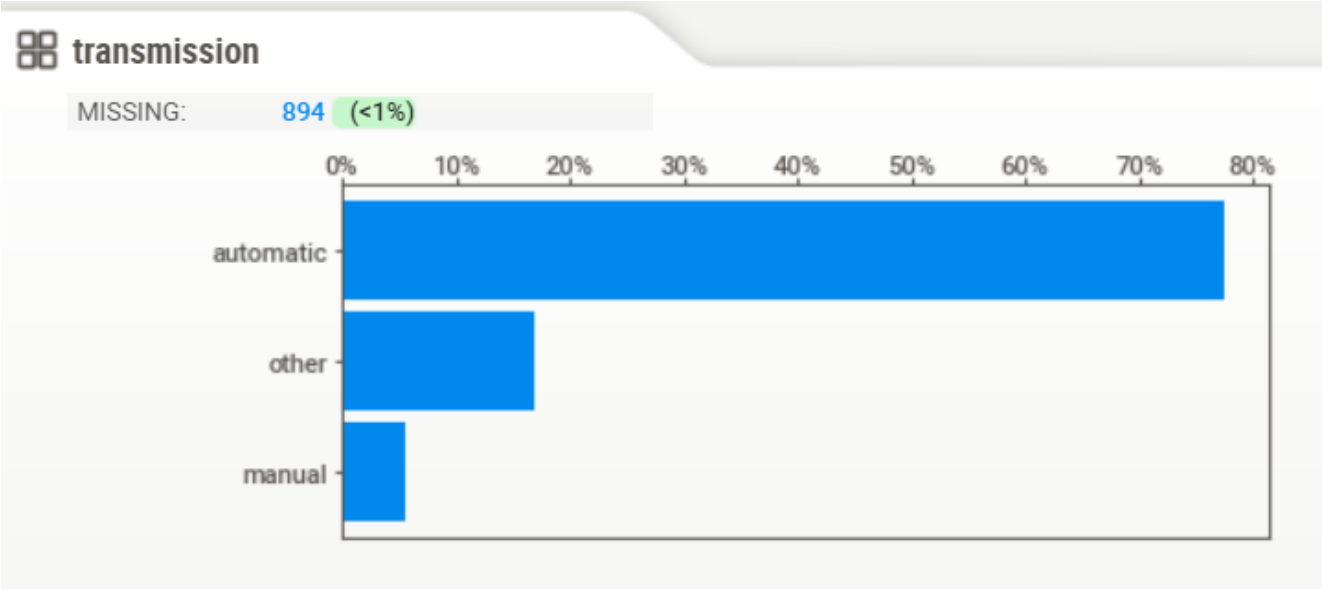

Figure 5

Another important feature of a car is the transmission. This graph in Figure 5 shows the percentage of cars that

have automatic, manual, and other transmission. A major chunk of cars that are listed show automatic transmission.

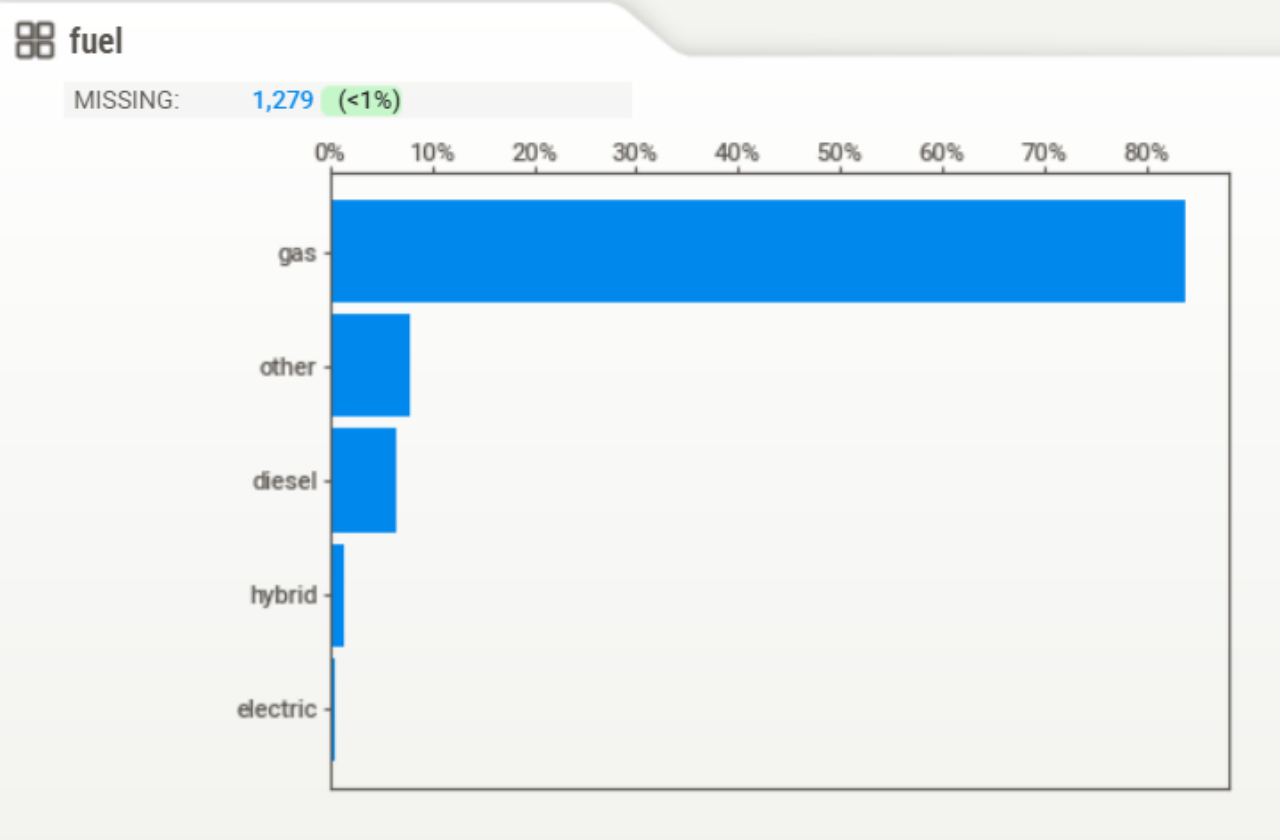

Figure 6

Fuel prices highly impact the car consumers and Figure 6 shows the type of fuel each car uses. As it can be seen, 84% of the cars listed use gas and only 6% cars use diesel. We can see that there are a very low number of electric cars that have been listed. This can be due to the poor popularity of electric cars during the years 1990-2010.

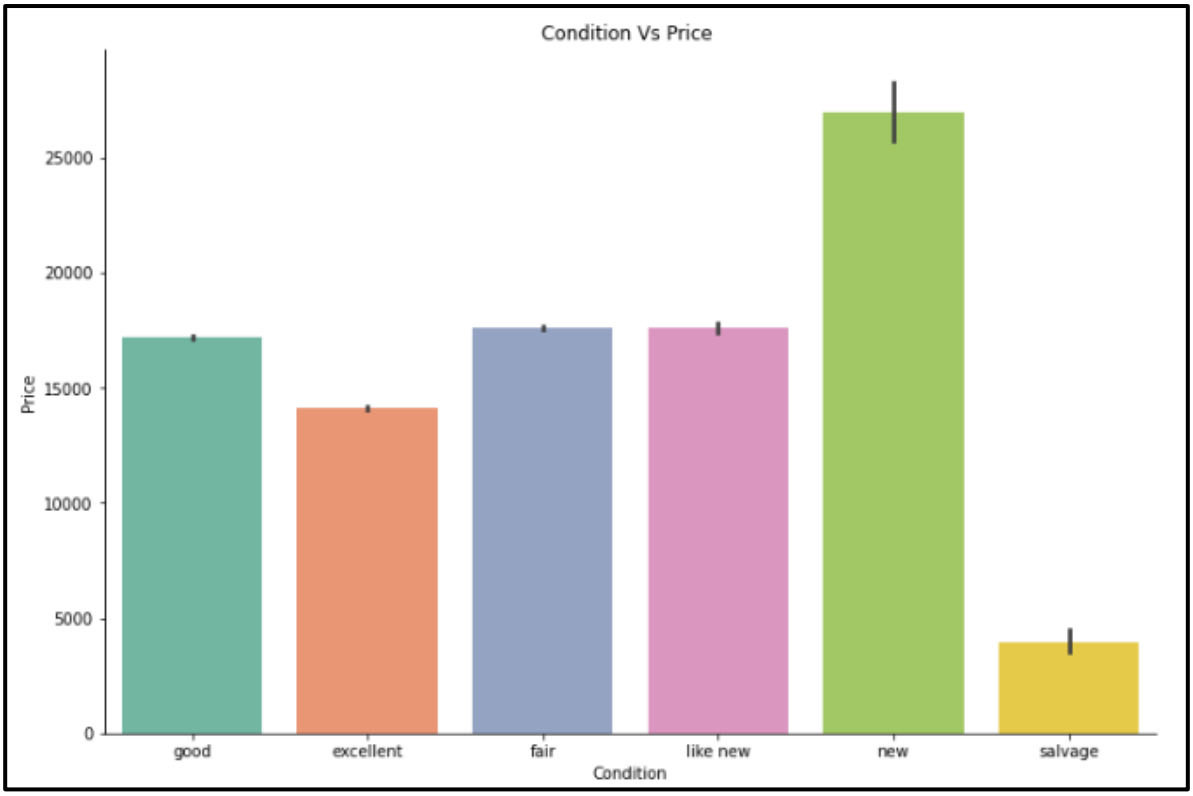

Figure 7

Figure 7 is a bar graph which shows the relationship between the condition of the car and the amount of variance Price has in regards with it. We can see that the lowest prices are for the salvaged cars whereas the highest prices are for the new cars, as expected. We also observed how cars in excellent condition are priced lower than cars in good and fair conditions.

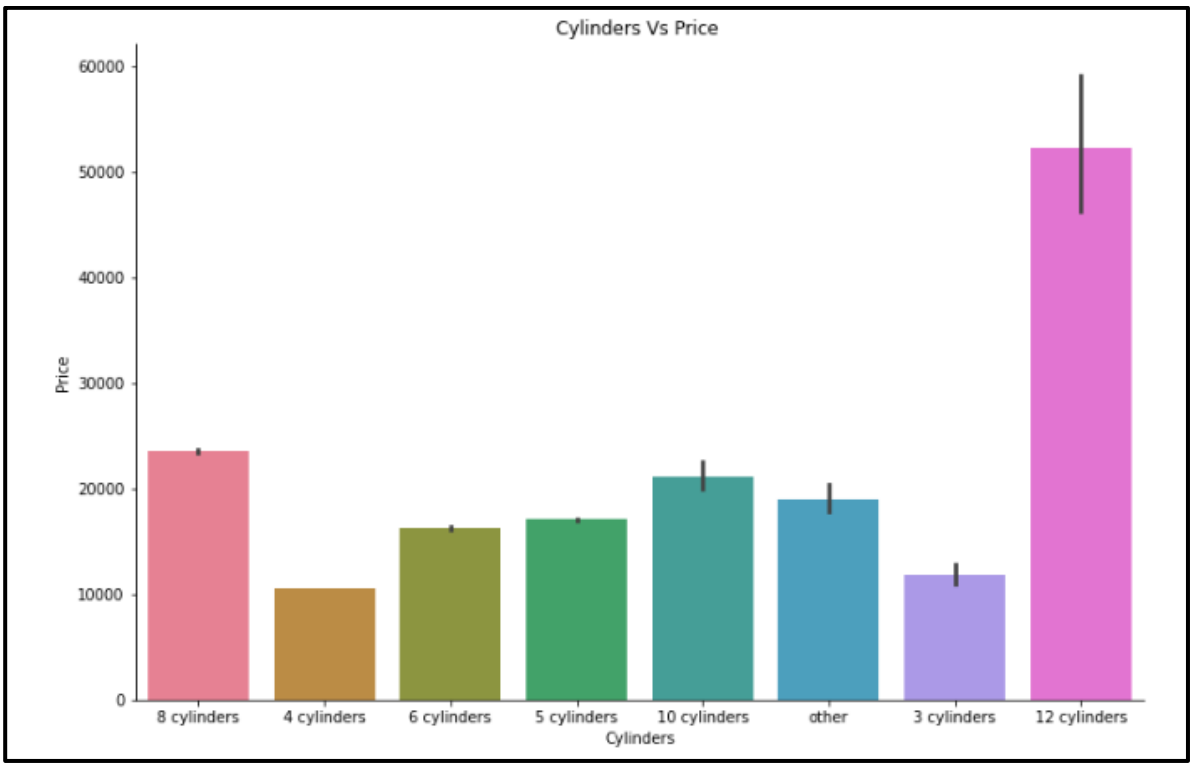

Figure 8

Figure 8 shows the relationship between the number of Cylinders the car has and the amount of variance Price has in regards with it. Even though Figure 4 showed that there are very few cars with 12 cylinders, these cars seem to be the most expensive ones.

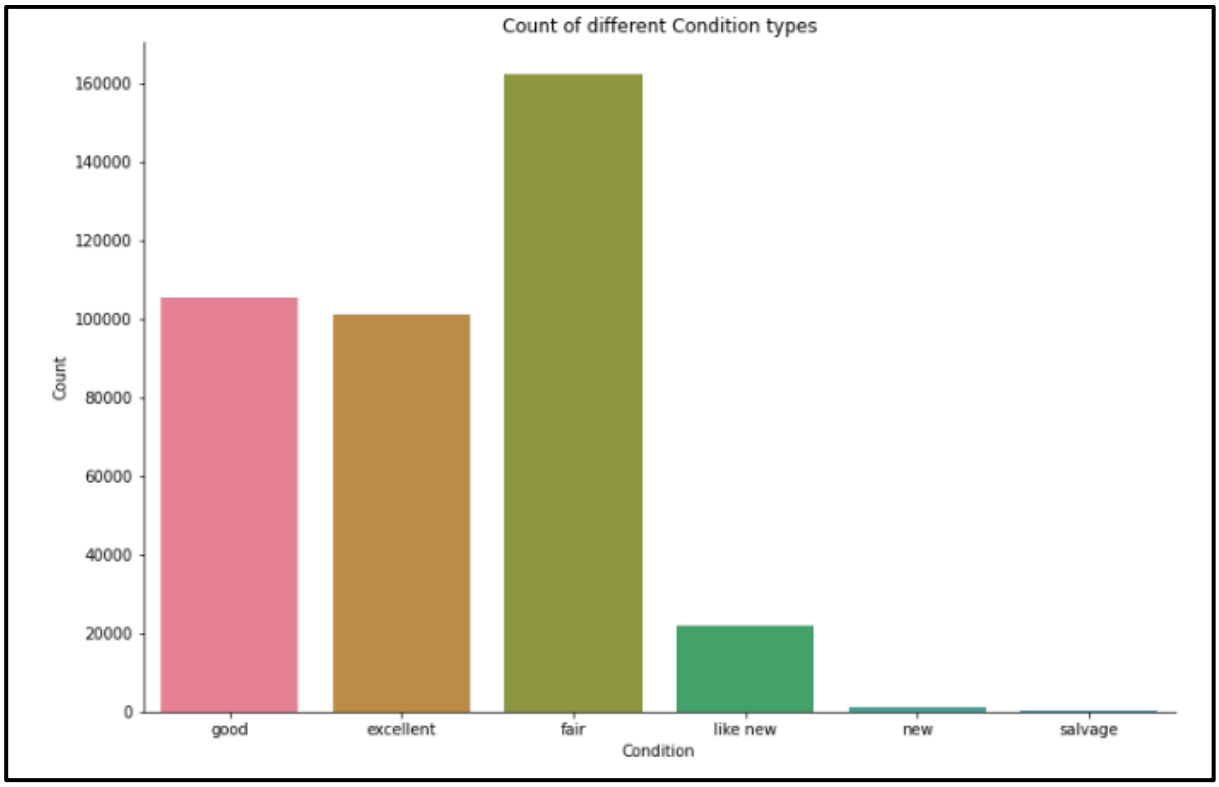

Figure 9

Figure 9 shows the count of different cars based on their conditions. Majority of the cars were in good and fair condition, which seems fair assuming they have been used for some years.

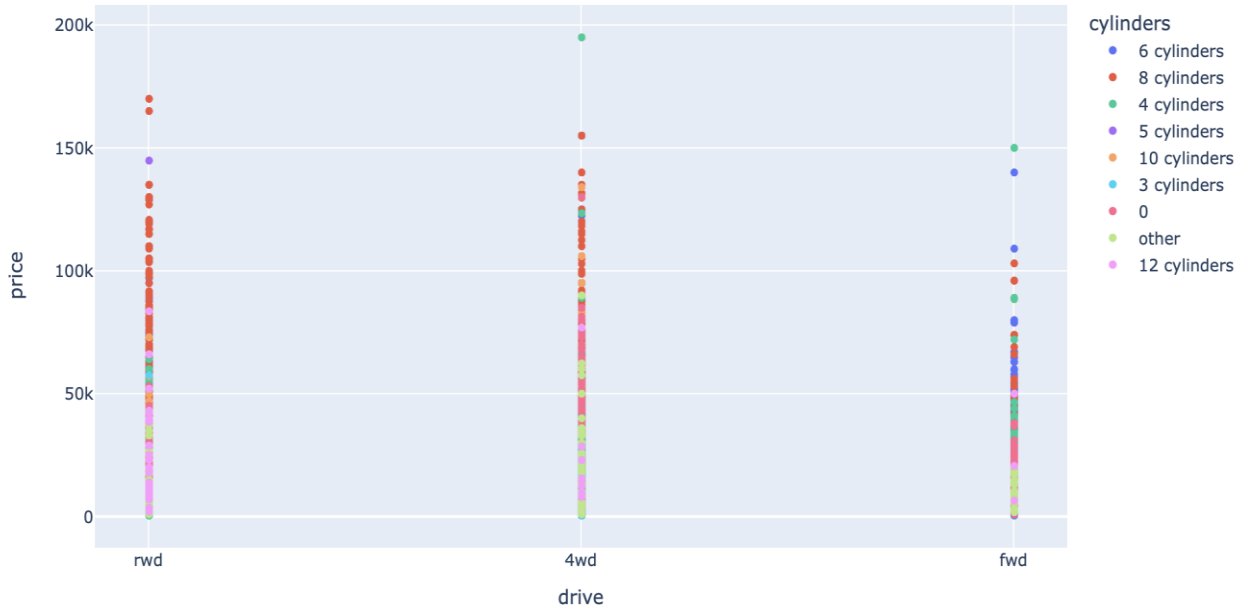

Figure 10

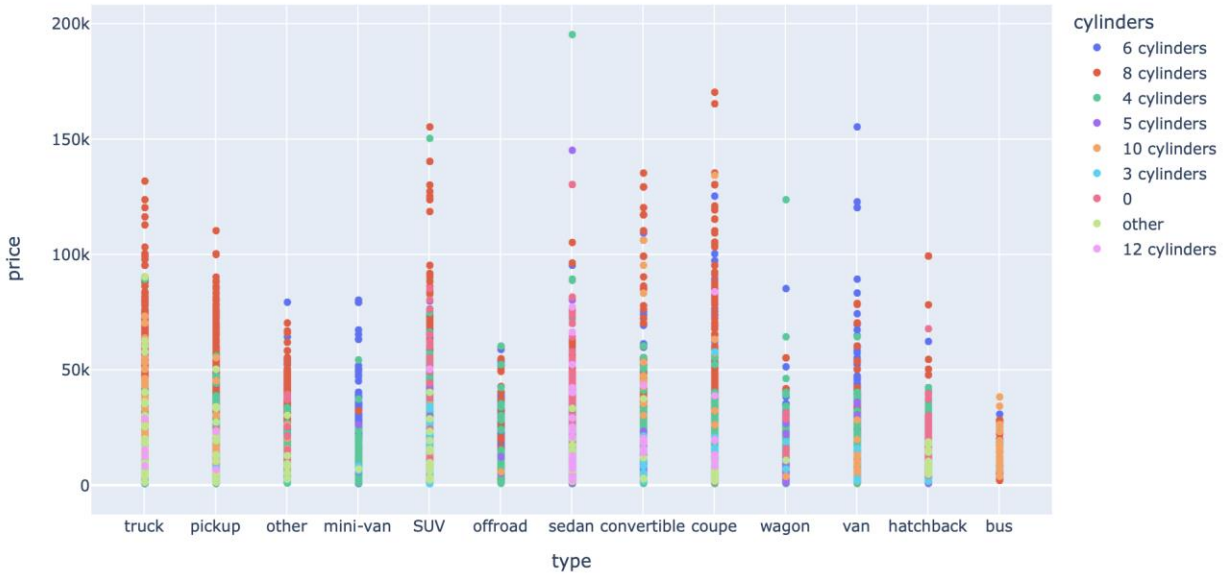

Figure 11

From figures 10 and 11, we can say that Rwd and 4wd drive cars are more expensive than fwd. It also shows that vehicle type and drive are highly correlated with Price. Heavy vehicles like SUV, truck tend to have a greater number of cylinders (>=8) and more expensive.

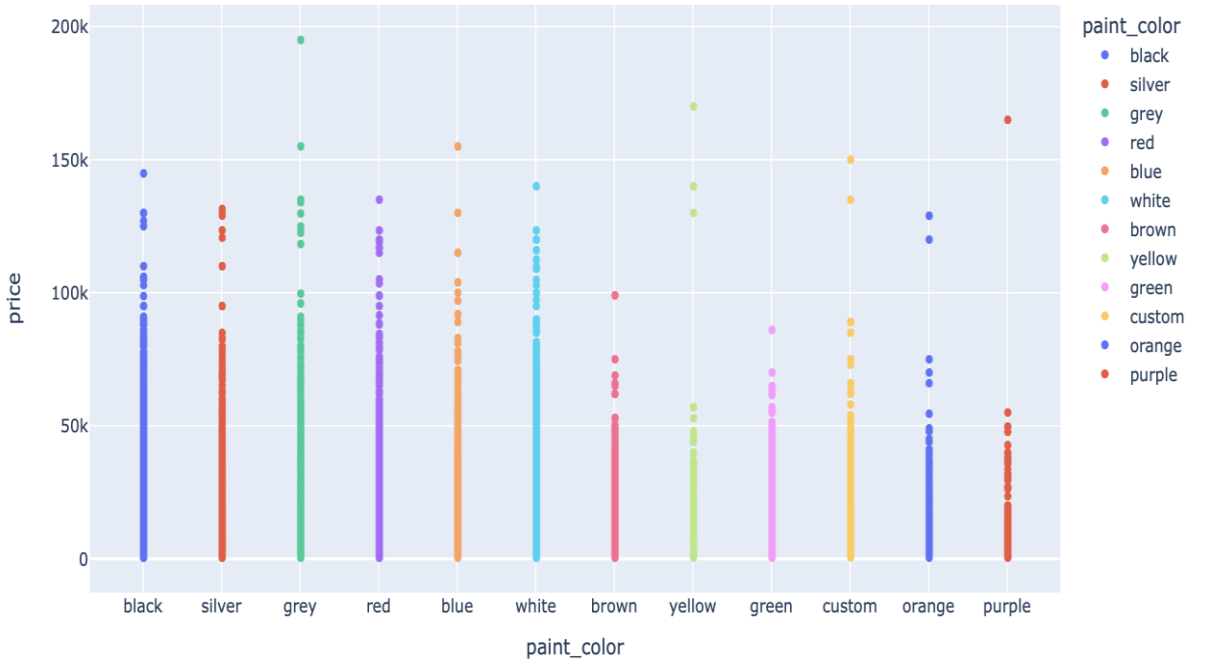

Figure 12

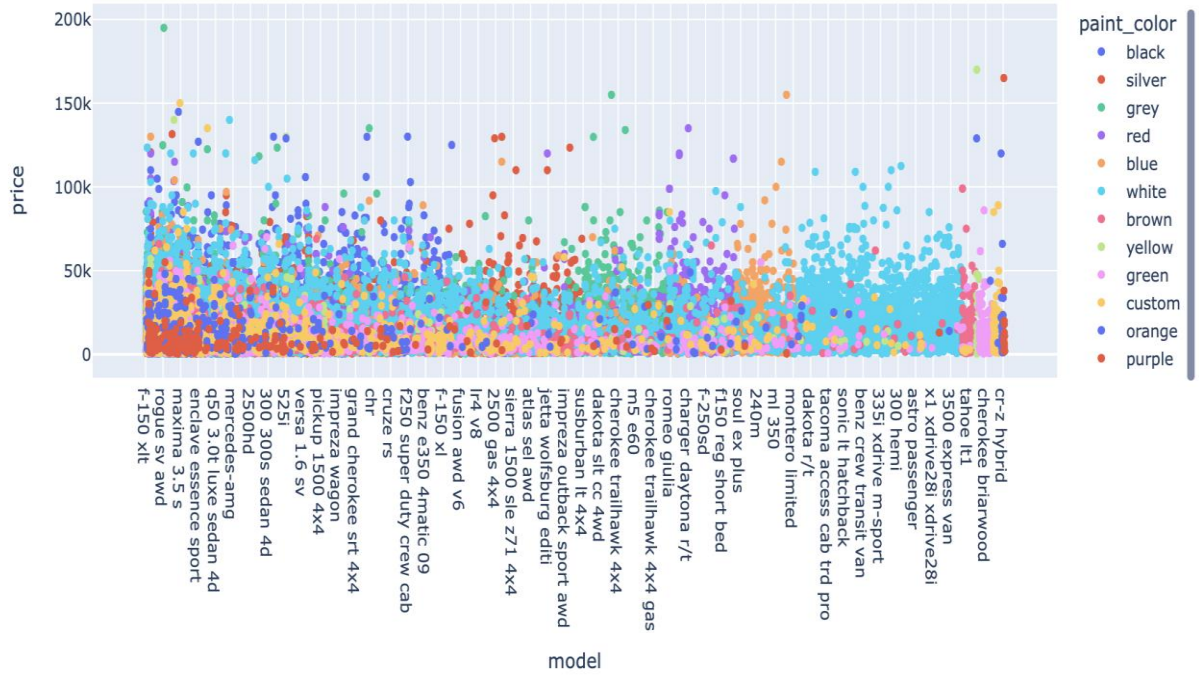

Figure 13

Figures 12 and 13 show that paint color affects car prices. Certain colors like Black, White, Grey are high in demand.

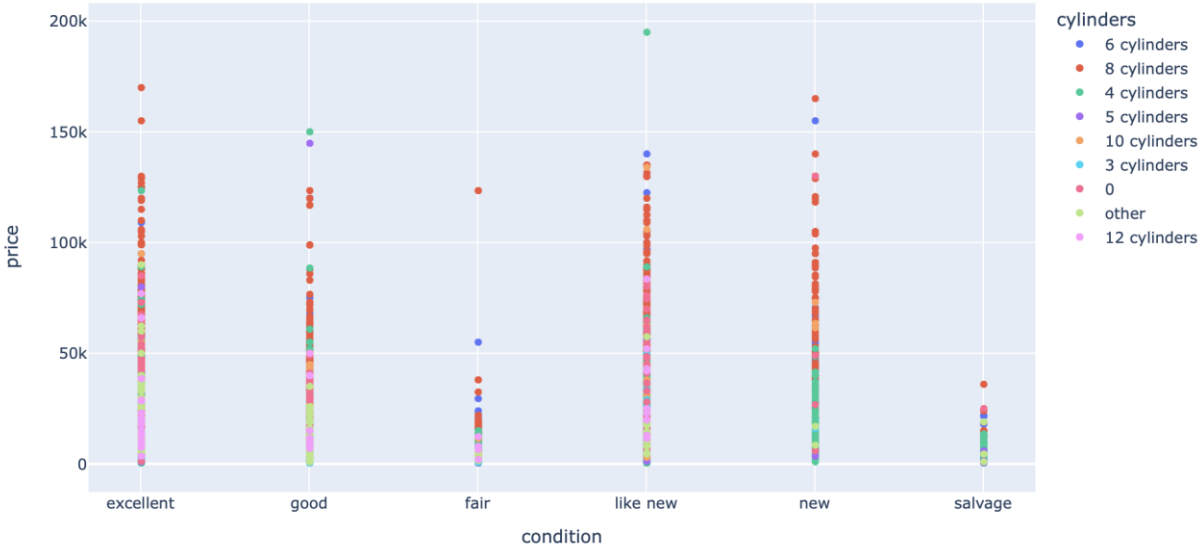

Figure 14

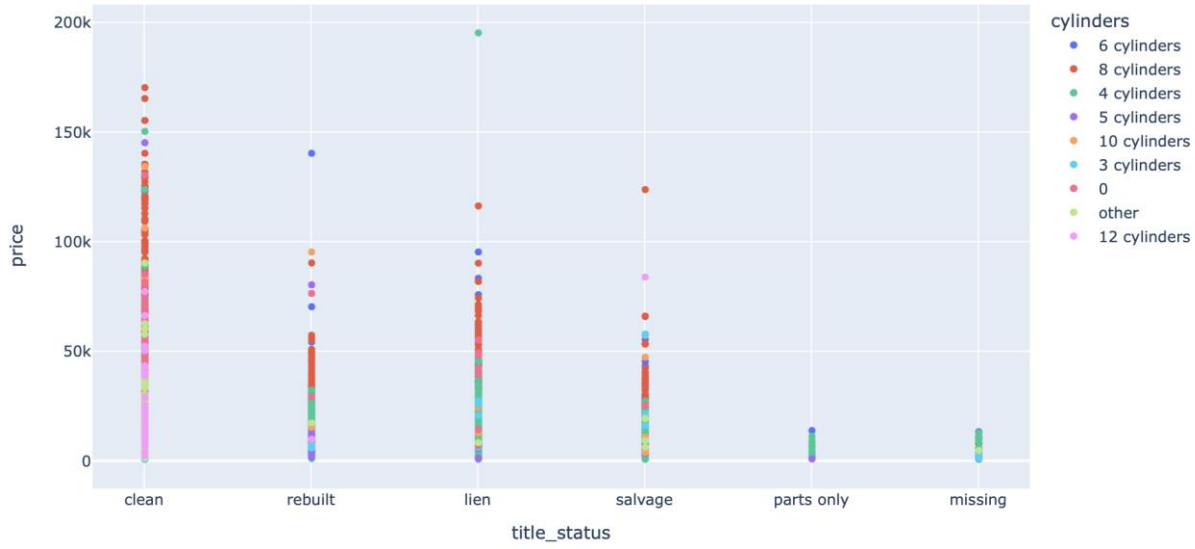

Figure 15

Figures 14 and 15 display that the car's condition and title status affect car prices. Cars with excellent/new/like new conditions are higher in price and cars in 'clean' category are more expensive.

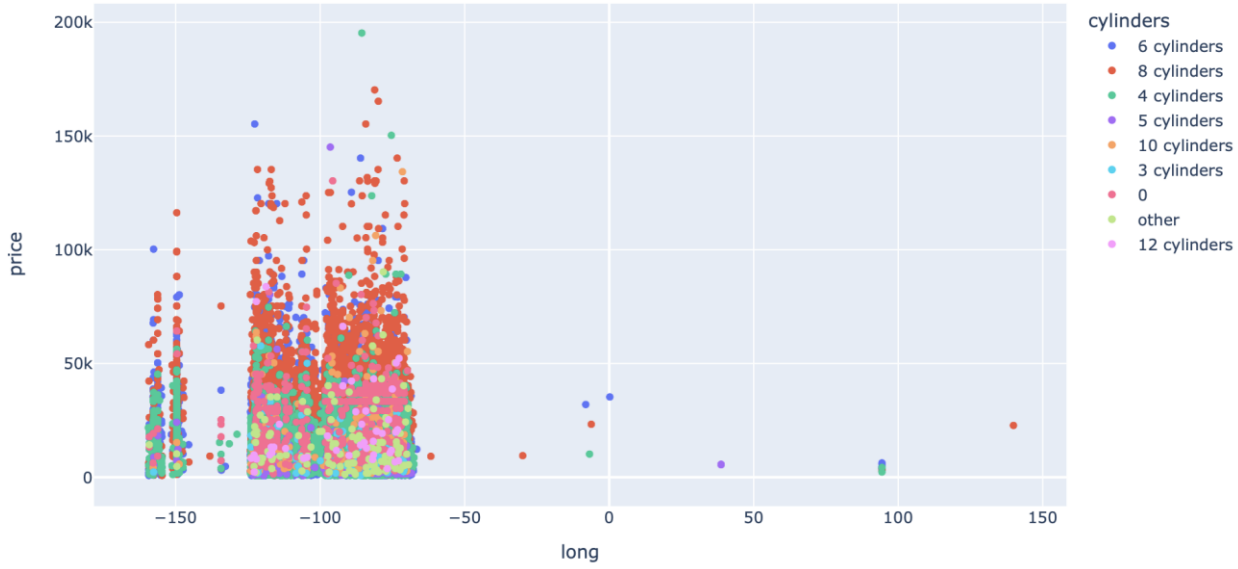

Figure 16

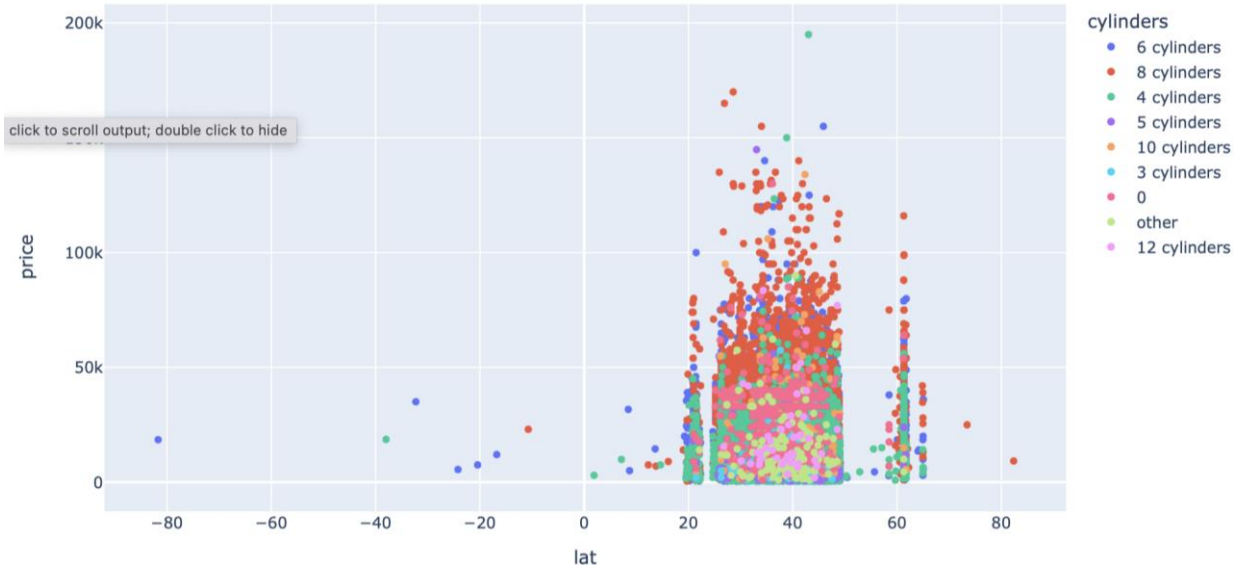

Figure 17

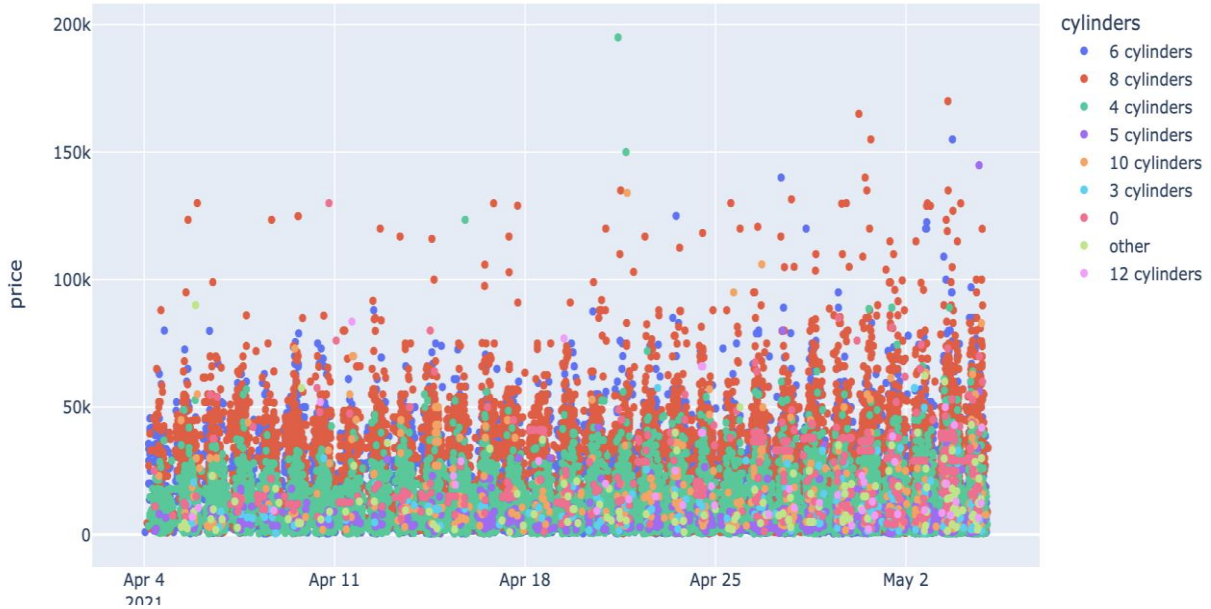

Figure 18

Figures 16, 17 and 18 show how lat(Latitude), long(Longitude) and posting date don't seem to be correlated with price.

After performing EDA on our dataset, we can expect below features may affect car prices:

- Year
- Manufacturer
- Model
- Condition
- Cylinders
- Fuel
- Odometer
- Title Status
- Transmission
- Drive
- Type
- Paint Color

Pre-Processing Data

Data preprocessing, which is a crucial phase in the data mining process, can be defined as the altering or dropping of data before usage to ensure or increase performance. It describes any type of processing performed on raw data to prepare it for another data processing procedure. Preprocessing data can increase the accuracy and quality of a dataset, making it more dependable by removing missing or inconsistent data values brought on by human or computer mistakes. It ensures consistency in data.

a. Data Cleaning

b. Handling outliers

c. Missing Values Relationship

d. Information from "Description"

1. **Data Cleaning**

Data cleaning is the process of identifying and correcting inaccurate, incomplete, irrelevant, or duplicated data in a dataset. This is a crucial step in data analysis, as it ensures that the data is accurate and reliable.

As per the heat map in Figure 1, we observed high missing values in the columns county and size. While the column county had no data, the column size had 67% missing values. Since this data is not enough to make any firm predictions, we dropped these two columns from our dataset.
We also dropped 'URL', 'Region_URL','Image_URL' ,'VIN' as they are not significant for price and the URL's given in this dataset do not work.
We also found data in the column price which had values less than $500. Looking at Figure 3, this does not seem realistic and counts for dirty data. Hence, we dropped rows in the column price with values less than $500.
Considering Figure 2, as we noticed, there was very negligible data in the initial years, hence we decided to take the data from the year 1990 onwards. In addition, we also decided to drop rows where the null values are present for more than 17 columns.
Furthermore, we calculated the percentage of null values in each column and decided to drop the rows with null values for columns which had more than 10% of null values present.

After the data cleaning performed in previous steps, we found that the null value percentage for column Type had dropped below our assumed threshold (10%). Hence, we dropped all NaN values from the column Type. In addition, for car type as "Electric", we found 1110 records where "Number of cylinders" is NaN. We have imputed these values with 0. The cylinders for electric cars are imputed with 0 using the np.where method. The cylinders for electric cars with non-zero values are corrected to 0 using the np.where method.

2. **Handling Outliers**

Outliers are data points that are significantly different from other data points in a dataset. They are values that are located far away from the other values in the dataset, and they can have a big impact on the overall statistical analysis and conclusions drawn from the data. It is important to identify outliers in a dataset because they can distort the statistical analysis and

lead to incorrect conclusions. One common method for identifying outliers is to use a box plot or a scatter plot to visually examine the distribution of the data.

Figure 19

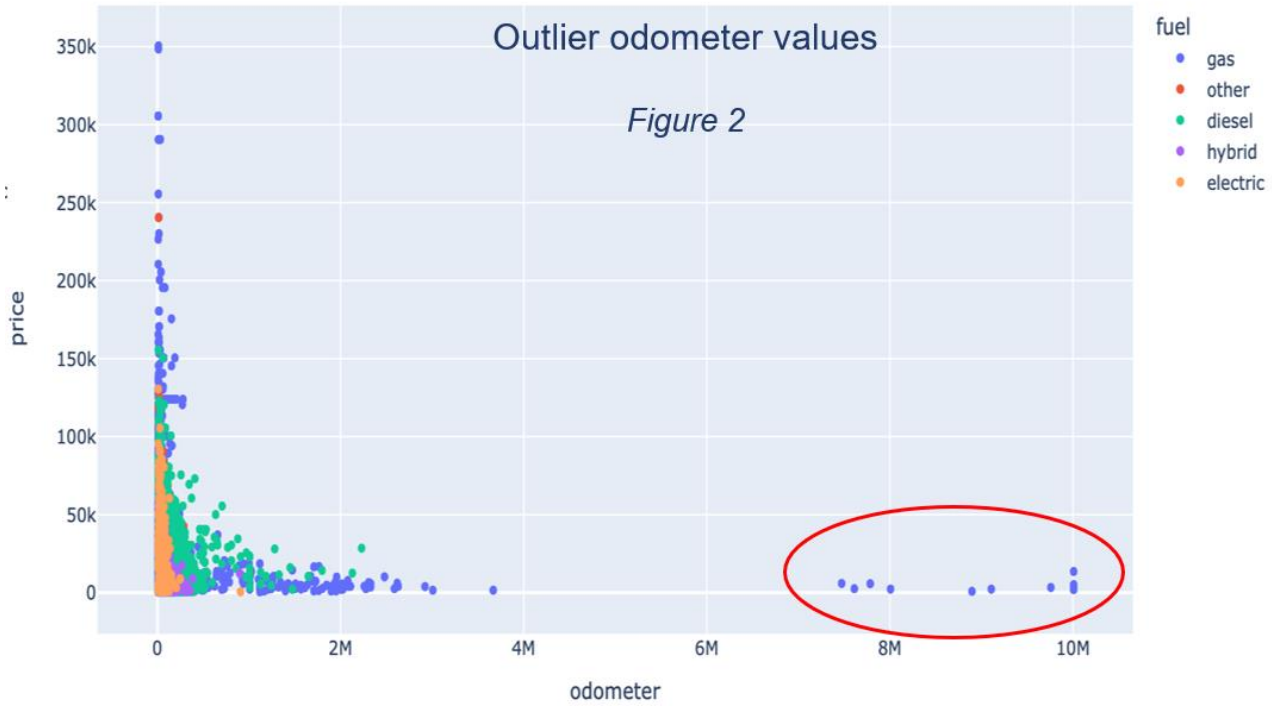

Figure 20

We identified some outliers in our data. In figure 19, the encircled data points are the price outliers during several years. In figure 20, the encircled data points are the outlier odometer values with respect to the prices of used vehicles. To enhance the data, we remove the odometer values > 5,000,000. We will drop these outliers for better visualization.

Figure 21

In figure 21, after removing the outliers, we can see that till around the year 2000, the market was dominated by gas and diesel cars, however few hybrid cars started being listed around 2002 and a major rise in electric cars can be seen from around 2012. It can also be deduced that gas cars have been among the higher priced cars among all.

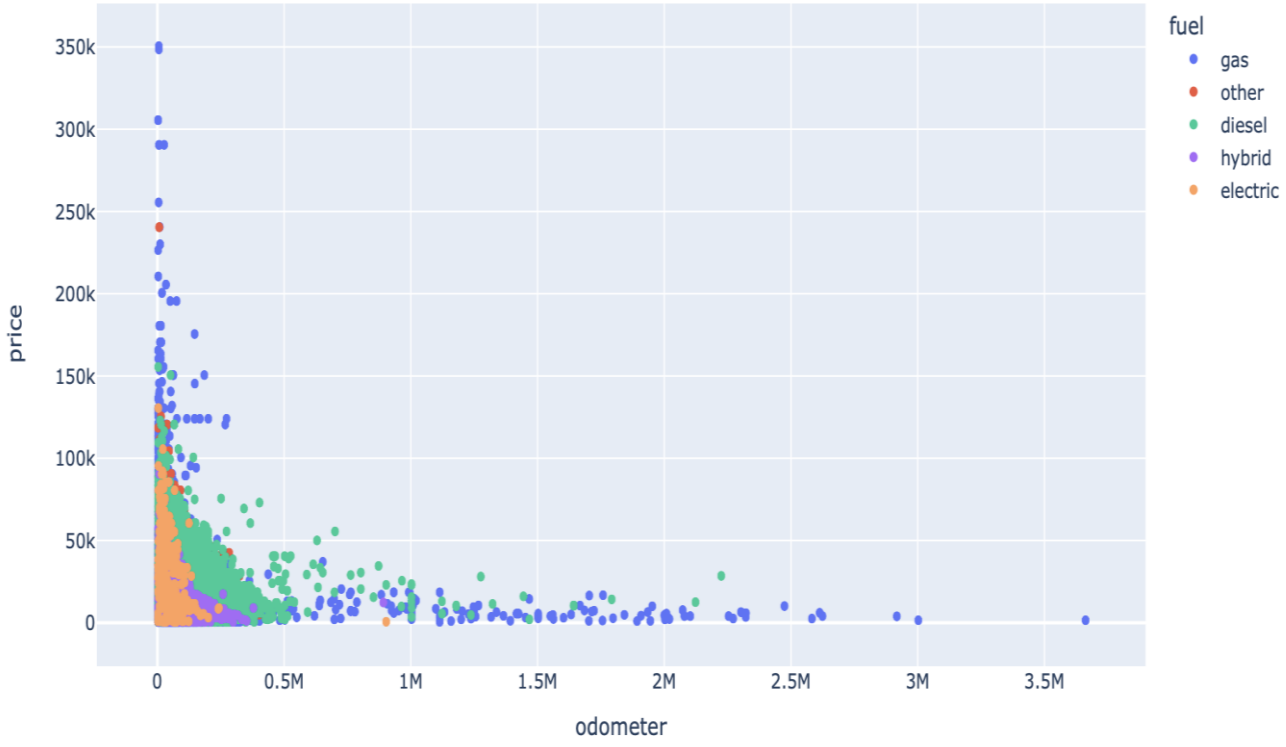

Figure 22

In figure 2, we can see that electric cars listed for sale are newer than diesel cars which have a higher odometer reading. We can also see how prices for newer gas cars are higher and the gas cars with more odometer value have lower price. This also shows that most used cars with more odometer values are fueled by gas.

3. **Missing Values Relationship**

So far, we had done data cleaning based on assumed threshold (drop NaN's with <10%) and drilled down to the following columns which still have a significant percentage of null values to deal with:
Condition ~ 29 %
Cylinders ~ 32 %
Drive ~ 19 %
Paint Color ~ 15%

We have also tried to analyse if there is any pattern in missingness of data. We used some algorithms and methods to investigate this. First, we used DBSCAN but we could not continue with it as it is primarily used for 'float' values. Then we did K-prototype clustering and ran a chi-squared test, results for which are below.

K-Prototype Clustering:
K-Prototype clustering is used for clustering analysis of categorical variables. Here we converted the whole data into True and False values. The data is substituted with True if the value is missing and False if the value is not missing.

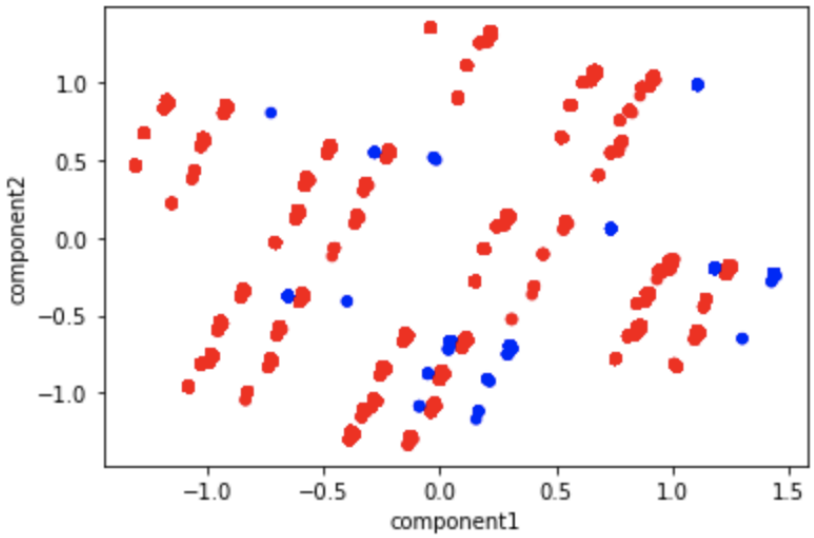

Figure 23

Chi-Square Test:
The chi-square test is a test that helps to determine whether there is a significant association

between two categorical variables in a dataset.

	Condition	Cylinders	Drive	Paint_color
Condition		5879	17940	4946
Cylinders			315	2478
Drive				2181
Paint_color				

Figure 24

After running these various algorithms and tests, we could not find any pattern or relationship in the missingness of data.

4. **Information from "Description"**

The following are some snippets from the descriptions of various cars given by the sellers. Our next challenge is to extract information of paint_color and drive from this column. The highlighted text represents the type of information that we are looking for and can be useful for our study.

```
'2011 Ford F-150 2WD XL PickUp Truck Regular
Cab Long Bed 3.7 Liter V-6 Engine  164,000
Original Miles Cold A/C Excellent Condition
Runs Great  Needs Tires Clear Title
$8950.00 CASH Marietta, Georgia   show
contact info       Vin# 1FTMF1CM6BKD55331'
```

```
'2006 Jeep Grand Cherokee Laredo FOR SALE BY OWNER. This vehicle has some miles on it, but runs well and I need
a car. Had the transmission replaced roughly 1 1/2 years ago.  Mechanically and Electronically sound, although
the check engine light has stayed on since my uncle sold it to me after his shop repaired the transmission.  I
replaced the alternator roughly 1 year ago and fixed another problem that arose from it regarding a cable
connected to the alternator. I have done a good job of maintaining it and keeping plenty of oil in the vehicle.
- 3.7 Liter / V6 / 5-speed automatic engine - 2WD automatic  - Black exterior with leather grey seats and grey
paneling (very clean) - 17 - 22 MPG - 281000 miles  - 4 power windows / 2 power seats - Non-smoker vehicle - New
tires placed on the vehicle roughly 9 months ago   The negatives: - Dent on passenger side rear door (as shown)
- Scratch on driver side back quarter panel (as shown) - Driver side inside rear door handle is down in the door
(as shown) - Check engine light and Tire Pressure lights are on although nothing is wrong here   Contact me via
phone at: 2 zero five - 5 six 7 - four 3 8 zero'
```

However, the issue is the randomness of the information present in this column. We tried to use this information to impute missing values but since it does not have any definite pattern, it ends up imputing junk values, as can be seen in the snippet below.

```
'MagneticOEM', 'Ice', 'AblazeOEM', 'Smoke', 'Sandalwood',
'MaroonOEM', 'GRAYMPG', 'Iridium', 'Shadow', 'Atlas', 'Cocoa',
'PREDAWN', 'Metal', 'Light', 'Medium', 'GrayDoors', 'Bordeaux',
'Mesa', 'River', 'Bronze', 'Oak', 'Austin', 'Med', 'BURGUN',
'ObsidianInterior', 'DARK', 'Oxford', 'BURGAN', 'Tuxedo',
'Monterey', 'Blade', 'Driftwood', 'Gunmetal', 'GrayFuel',
'TANInterior', 'Quartz', 'Space', 'MEDIUM', 'Alabaster', 'Taffeta',
'Premium', 'Arctic', 'GrayDescription', 'CHARCOALInterior',
'MAROONInterior', 'Cappuccino', 'Galaxy', 'Crystal', 'Coppertino',
'Smokey', 'Star', 'Ford', 'Pepperdust', 'Sunset', 'Storm',
'Pyrite', 'Neutral', 'Fine', 'Imperial', 'Anthracite', 'Steelmist',
'CopperOEM', 'Saharan', 'Umbria', 'Meteor', 'Quicksilver',
'Dorado', 'Ginger', 'Triathlon', 'GrayIf', 'Metallic',
'Smokestone', 'BeigeOEM', 'Warm', 'True', 'Blonde', 'designo',
'Caspian', 'Havana', 'Iron', 'Ibiza', 'Kona', 'MagneticInterior',
'Mojave', 'Delmonico', 'Cassis', 'Mountain', 'GrayYou', 'GRAYYou',
'Sandstorm', 'tan', 'Aegean', 'Super', 'Stone', 'Winterberry',
'Onyx', 'CEMENT', 'SIVER', 'GrayClearpathpro', 'GRAYClearpathpro',
'BURG', 'BurgundyOEM', 'Panthera', 'WHITE', 'GrayEquipmentVehicle',
'BlackEquipmentVehicle', 'Mahogany', 'BROWN', 'San', 'Mosaic',
'Special', 'HEATED', 'ONE', 'SAFE', 'SunsetOEM', 'SPEED',
'GrayVehicle', 'Triple', 'PlatinumOEM', 'Brownstone', 'NA', 'RED',
'Redline', 'TorredOEM', 'Bikini', 'CrushOEM', 'Golden',
'InfernoOEM', 'BLACKInterior', 'Kalapana', 'SNOW', 'WHITEInterior',
'Century', 'BLUEInterior', 'GUN', 'PLATINUM', 'GREENInterior',
'AVALANCHEInterior', 'SUMMIT', 'TUXEDO', 'OXFORD', 'SPICY',
'MAXIMUM', 'DEEP', 'TITANIUM', 'SHADOW', 'ROYAL', 'BurgundyDoors',
'BRIGHT', 'NO', 'SILVERInterior', 'ONYX', 'Copperhead', 'Bohai',
'Cement', 'Liquid', 'Moondust', 'Barcelona'], dtype=object)
```

Test-Train-Validation Split

For the train and test stage of this study, we have used 3-way split to generate a validation set. The train_test_split function from sklearn.model_selection library is used to divide the dataset into training, validation, and test sets. First the data is split into train and test sets and then the train set is split into train and validation sets.

The train set is used to train the machine learning model. The validation set is used to tune the ML model. The test set is used to evaluate the ML model.

The test_size parameter specifies the size of the test set. In this case, it is set to 0.2, which means that 20% of the data will be used for testing. The random_state parameter is set to 42 to ensure that the same random split is generated each time the code is run.

Further, a pipeline is defined to perform one-hot encoding on the categorical columns. One-hot encoding is a technique that is used to convert categorical data into numerical data. The fit_transform method of the pipeline object is called on the training set. The ColumnTransformer is used to select the categorical columns and apply the one-hot encoding to them. The ‘remainder’ parameter is set to 'passthrough' to pass through any columns that are not transformed

Machine Learning Models

Simple Linear Regression:

Simple linear regression is a statistical method used to model the relationship between two variables, where one variable (the independent variable) is used to predict the other variable (the dependent variable). It assumes that there is a linear relationship between the two variables, meaning that changes in the independent variable are associated with changes in the dependent variable that can be expressed by a straight line.

In simple linear regression, the goal is to find the equation of the straight line that best describes the relationship between the two variables. This equation can be used to predict the value of the dependent variable for any given value of the independent variable.

Decision Tree Method:

A Decision Tree regressor is a machine learning algorithm that is used to model and predict continuous numerical values based on input features. It works by recursively partitioning the data based on the values of the input features, until a terminal node is reached where a prediction is made. The decision tree is constructed by selecting the input feature that provides the best split of the data based on some criterion. The predicted value for a new input is the average or median value of the target variable in the leaf node. Decision tree regression is easy to interpret and can handle non-linear relationships but can be prone to overfitting.

Random Forest Regressor:

Random Forest regression is a machine learning algorithm that combines multiple decision trees to improve the prediction accuracy and reduce overfitting. It works by randomly selecting subsets of the input features and the training data for each tree, and then aggregating the predictions of all the trees to obtain the final prediction. This approach reduces the variance and increases the robustness of the model, by reducing the impact of individual trees and the noise in the data.

XGBoost (Extreme Gradient Boosting):

XGBoost is a machine learning algorithm that uses gradient boosting to train a boosted ensemble of decision trees. It works by sequentially adding decision trees to the ensemble, where each tree is trained to correct the errors of the previous trees. XGBoost uses a novel regularization technique called "gradient-based regularization", which penalizes the complexity of the model based on the gradient of the loss function, and a distributed computing framework that allows for parallelization and scalability.

Machine Learning Libraries

The following are some libraries used to perform a variety of functions in our analysis procedures:

1. Pandas: Pandas is a Python library used for data manipulation and analysis. It is commonly used in data pre-processing and cleaning tasks in machine learning workflows.
2. Seaborn and Matplotlib: These libraries are used for data visualization tasks, which are an important part of exploratory data analysis in machine learning workflows.
3. Missingno: This library provides a visual representation of missing data in a dataset, which can be useful in identifying patterns and trends in missing data.
4. Plotly and Sweetviz: These libraries provide additional data visualization capabilities, including interactive visualizations and detailed statistical summaries.
5. KPrototypes: This library is used for clustering categorical and numerical data, which can be useful in exploratory data analysis and unsupervised learning tasks.
6. Scipy: This library provides a wide range of statistical functions, including hypothesis testing, probability distributions, and descriptive statistics. These functions are commonly used in machine learning tasks such as feature selection and model evaluation.
7. NumPy is a library for scientific computing with Python. It contains a high-performance multidimensional array object and tools for working with arrays. NumPy is the foundation of many other Python libraries for scientific computing, such as SciPy and Matplotlib.

8. Scikit-Learn is a library for machine learning. It provides a wide variety of machine learning algorithms, including support vector machines, decision trees, and random forests. Scikit-Learn is easy to use and has a large and active community of users.

V. ANALYSIS AND RESULTS

After performing EDA and pre-processing the data, we applied several machine learning models to predict the listing price of the used cars.
We first ran a Linear Regression on our data. The linear regression model is performed twice. The first time it is performed on the original target variable, and the second time on the log-transformed target variable. It is done so because using different target variables can affect the accuracy and interpretability of the model's predictions. However, even after performing it the second time, the accuracy level remained the same (0.7114).
We got the following results:

➢ LINEAR REGRESSION

- Original Target Variable

  R-squared value= 0.7114

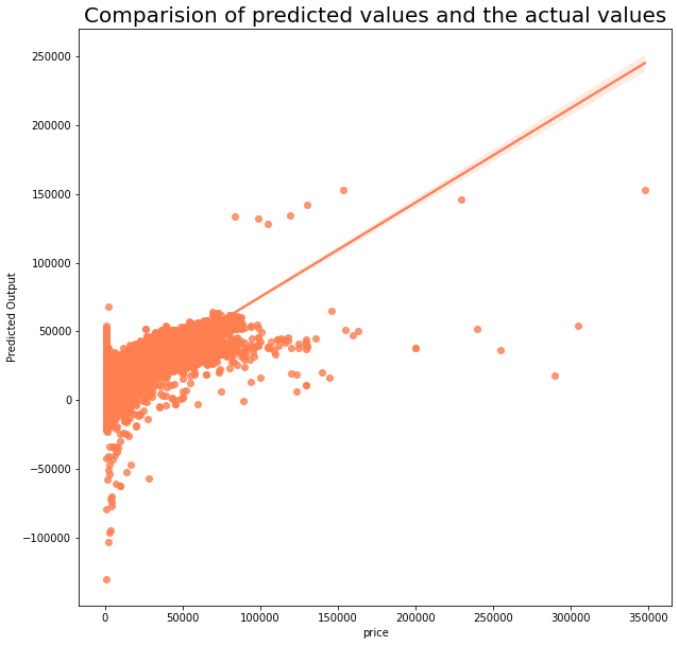

Figure 25

- Log-Transformed Target Variable

  R-squared value= 0.7114

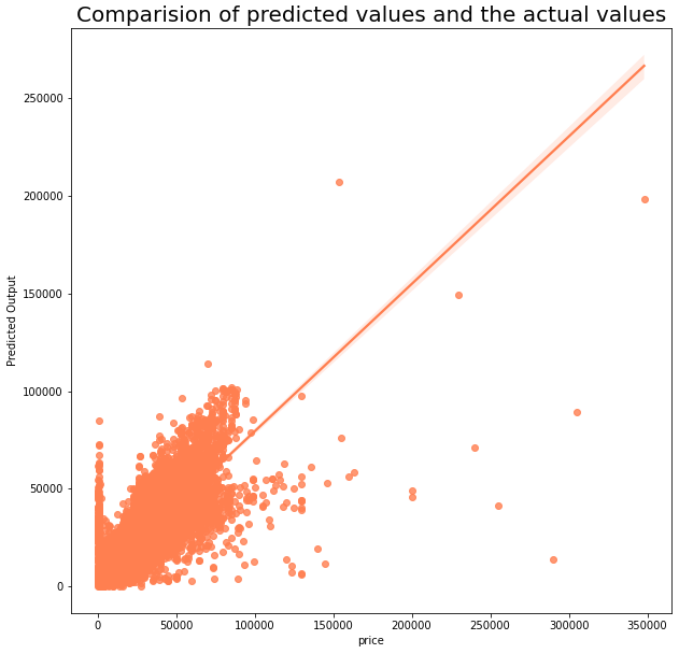

Figure 26

Since the accuracy level isn't too high, we decided to change our ML model to Decision Tree Regression.
It is also performed twice. The first time it is performed on the original target variable and then on the log-transformed target variable.
The following are the results:

- **DECISION TREE REGRESSOR**
  - Original Target Variable

    Training Set: $R^2 = 0.9999$, RMSE= 124.85
    Validation Set: $R^2 = 0.8253$, RMSE= 6124.77

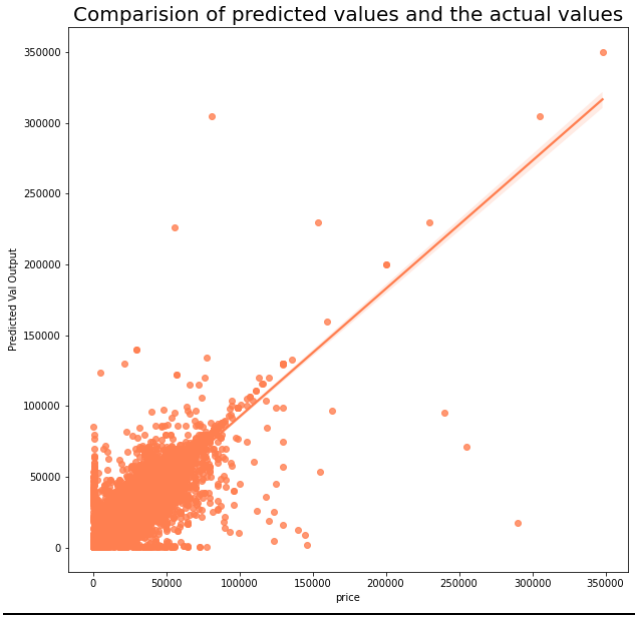

Figure 27

- Log-Transformed Target Variable

  Training Set: R^2 = 0.9999, RMSE= 132.21
  Validation Set: R^2 = 0.8214, RMSE= 6186.80

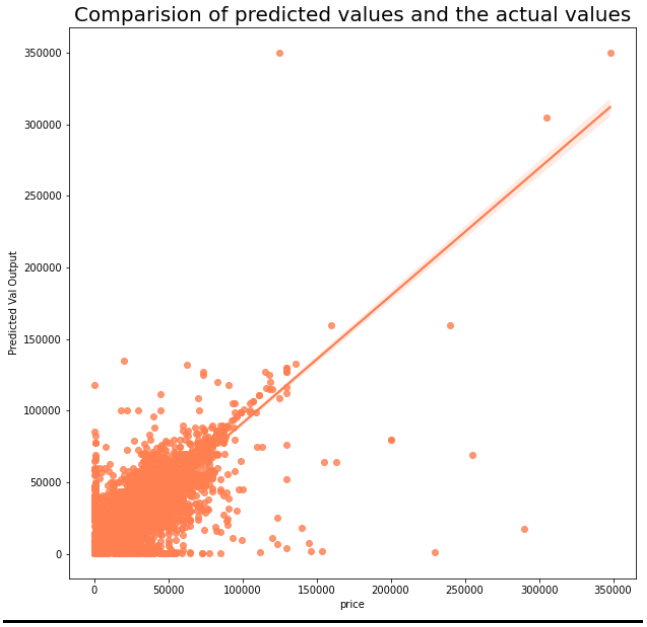

Figure 28

The Decision Tree Regressor model is overfitting the data, as the training R^2 score is very high (close to 1) but the validation R^2 score is much lower (0.82). This suggests that the model is fitting too closely to the noise in the training data and is not generalizing well to new data. Hence, we decided to move on to Random Forest Regressor, which is performed on original target variable and log-transformed target variable. The code trains a random forest regression model with 500 trees and a maximum depth of 10, using the training data
The following are the results:

➢ **RANDOM FOREST REGRESSOR**

- Original Target Variable

  Training Set: R^2 = 0.8195, RMSE= 6170.1
  Validation Set: R^2 = 0.7874, RMSE= 6756.7

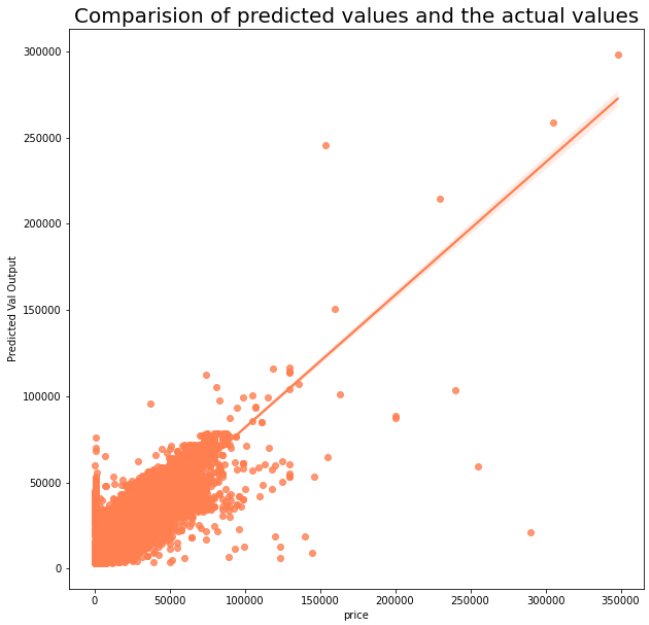

Figure 29

- Log-Transformed Target Variable

  Training Set: R^2 = 0.7624, RMSE= 7080.5
  Validation Set: R^2 = 0.7301, RMSE= 7612.4

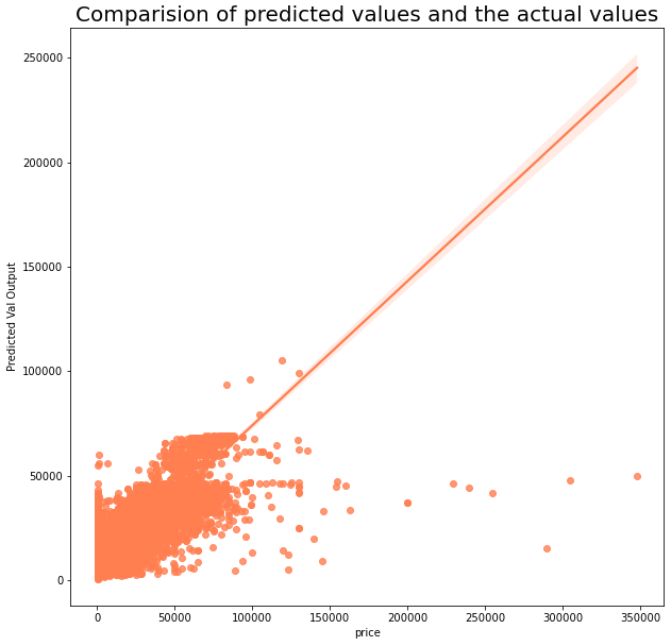

Figure 30

These scores suggest that the model performs worse on both the training and validation datasets after transforming the target variable, indicating that the log transformation may not be helpful in this case. This model, too, does not deal well with the outliers and we would like to test another model for a higher accuracy level. Hence, we move on to XGBoost Model to check for its performance.
The following are the results:

➢ **XGBOOST MODEL**

- Original Target Variable

Training Set: R^2 = 0.895, RMSE= 4706.42
Validation Set: R^2 = 0.857, RMSE= 5530.85

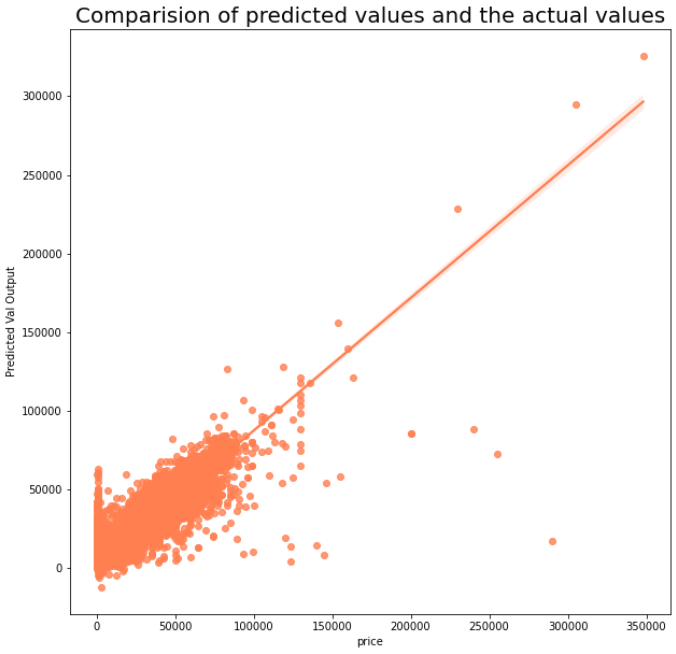

Figure 31

- Log-Transformed Target Variable

Training Set: R^2 = 0.864, RMSE= 5342.72
Validation Set: R^2 = 0.830, RMSE= 6033.07

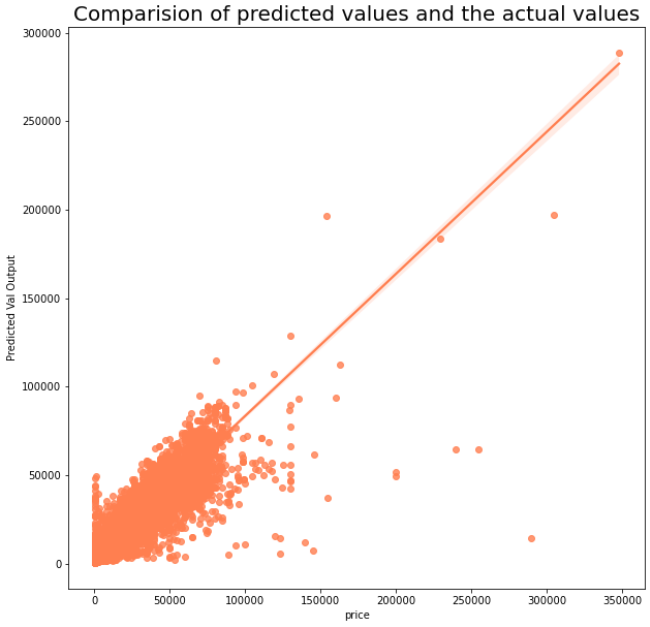

Figure 32

Since this model gave us the best accuracy as compared to other models, we stuck with this model. For further enhancement, hyperparameter tuning was performed using randomized search. The params dictionary defines a set of hyperparameters to tune, including max_depth, learning_rate, n_estimators, gamma, reg_alpha, and reg_lambda. The RandomizedSearchCV function is used to perform a randomized search over these hyperparameters. The following table mentions the results of each hyperparameter used.

HYPERPARAMETERS	RMSE VALUE		R-SQUARED VALUE		CHANGES/OBSERVATIONS
	Training	Validation	Training	Validation	
Default	*4706.41*	*5530.85*	*0.895*	*0.8575*	
Hyperparameter 1	1213.57	4669.72	0.993	0.8984	Overfitting observed
Hyperparameter 2	3541.79	4626.80	0.9405	0.9003	Regularization parameters, reduced learning rate, max depth reduced Overfitting observed
Hyperparameter 3	3016.61	4825.56	0.9568	0.8915	Reduced max_depth
Hyperparameter 4	3016.61	4825.56	0.9568	0.8915	Increased reg_alpha
Hyperparameter 5	3881.22	5144.07	0.9286	0.8767	Reduced n_estimators
Hyperparameter 6	3541.79	5065.57	0.9405	0.8805	Increased max_depth
Hyperparameter 7	3361.03	4952.56	0.9464	0.8857	Reduced n_estimator to 150
Hyperparameter 8	4794.86	5514.34	0.8910	0.8583	Reduced max_depth and min_child_weight is set to 5
Hyperparameter 9	4689.57	5478.55	0.8957	0.8602	Learning rate increased
Hyperparameter 10	4011.23	5122.58	0.9237	0.8777	Increased max_depth

After tuning our model with several hyperparameters, we got the results as shown in Table 1. From the table, we can learn that Hyperparameter 4 has performed better than the rest in terms of fitting the data

Table 1

well and showing a better RMSE and R^2 values. Hyperparameters 1 and 2 have better scores but they seem to be overfitting the data and not generalizing well. Hence, we will move forward with Hyperparameter 4 which gives us a an RMSE value of 4825.56 and R^2 value as 0.8915.

VI. DISCUSSION AND CONCLUSIONS

In this study, we started with a dataset that had loads of dirty data. Upon performing EDA, we found out that the dataset had lots of missing data and outliers. We also saw columns which were not relevant for the listing price prediction. We then performed pre-processing of the data which included dropping some columns and rows, handling outliers and trying to find a pattern in the missingness of data. We also tried to deduce data from the description, but it wasn't helpful. Further, we performed four ML models, results for which are summarized in Table 2 below. We also made use of one-hot pipeline. As per first level prediction, XGBoost performed with the maximum accuracy as compared to the other three models. To enhance this, we then used hyperparameters which finally gave us an accuracy of 89.15%.

ACHINE LEARNING MODEL	RMSE VALUE	R^2 VALUE
Linear Regression	6284.26	0.7114
Decision Tree Regressor	6124.77	0.8253
andom Forest Regressor	6756.71	0.7874
XGBoost	5530.85	0.8575

Table 2

With the accuracy attained by this ML model, we can develop this into a valuation tool which can be present as a toolbar on websites. On the other hand, we can also develop it into a valuation app which would require the user to mention some details and features about the car they want to check the price for. With these developments, we can pitch our services to the platforms which facilitate the buying and selling of used cars. This can be beneficial for the platform, buyers and sellers altogether.

For the platform, it can enhance the user experience and boost up engagement on the online platforms. It can optimize the operations as deciding on an acceptable listing price with the seller can be very time consuming. It creates a consistent and standardized valuation which reduces ambiguity.

For the sellers, it considers the historical data and hence gives a suggested listing price for the car based on the market forces of supply and demand. This tool suggests listing price based on all the features of the car, which the sellers may not consider in the usual, providing with an optimal valuation. It also saves a lot of time and efforts since it will only need one click for the seller to get the best listing price.

As for the buyers of used cars, it will be a big relief for them to know the actual worth of the cars they're

about to purchase without the risk of fraud or fear of paying way beyond the actual worth of the car. This tool will increase the transparency in pricing and reduce fraudulent activities which will be beneficial for the social good.

To inculcate this tool, the organisations will need to adopt the ML model for efficient ways of valuation which can be a challenge for those who rely on their own methods of valuation or those who are not tech-savvy and do it manually. For this, our valuation tool needs to have the simplest user interface which is as easy to operate as a regular mobile phone. This tool will smoothen out a major chunk of the process for selling and buying of used cars.

VII. LIMITATIONS & IMPROVEMENTS

Some limitations were also observed during this study. One of the major ones is that the dataset provided had a huge portion of dirty data and null values. Hence, a lot of time was spent on data cleaning. There were too many outliers present and this problem was not solved completely but minimized to some extent. Another major limitation is that the data doesn't have some important and relevant features like mpg, number of doors, etc. Also, some imputation methods may not be accurate. Further on, we could have tried using polynomial regression but we performed models based on the assumption that there is only one dependent variable.

To make more developments in our study, a dataset which contains the sales price of used cars instead of list price would make a great change to our application. Listing price is not the same as selling price, since the seller can decide to sell it at a price lower or higher than the list price. Hence, sales data would help us make our application more accurate by predicting the sales price instead of just the list price.

VIII. REFERENCES

Autolist.com. "10 Challenges of Buying a Used Car and How to Overcome Them."

CarGurus.com. "The Pros and Cons of Selling Your Car Online."

Cox Automotive. (2021). The Used Car Market Report 2021.

Dataset for the project: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

CarGurus. (2019). Predicting the Value of Used Cars: A Machine Learning Approach. Retrieved from https://static1.cargurus.com/gfx/cg_media/whitepapers/predicting-the-value-of-used-cars.pdf

Google. (2020). Edmunds Case Study: Machine Learning Helps Improve Car Buying. Retrieved from https://services.google.com/fh/files/blogs/edmunds_case_study.pdf

Huang, J., Hu, Z., Zou, J., & Tang, Y. (2021). A Hybrid Model of Gradient Boosting Decision Tree and Convolutional Neural Network for Used Car Price Prediction. IEEE Access, 9, 39435-39445.

iSeeCars.com. "What Are the Challenges of Selling a Used Car Online?"

Joshi, A., Kumar, A., & Bhattacharya, S. (2021). A hybrid model for used car price prediction in Indian market. International Journal of Intelligent Systems and Applications, 13(3), 10-18.

Microsoft. (2019). Autotrader Drives Sales with Machine Learning. Retrieved from https://customers.microsoft.com/en-us/story/autotrader-retail-azure-machine-learning

NerdWallet.com. "How to Price Your Used Car to Sell."

Society of Motor Manufacturers and Traders (2021) 'UK new car market falls -12.4% in Q1 2021 as showrooms remain closed', available at: https://www.smmt.co.uk/2021/04/uk-new-car-market-falls-12-4-in-q1-2021-as-showrooms-remain-closed/

Shahriar, N., Islam, M. R., Azad, M. A., & Ahmed, T. (2020). Comparison of machine learning techniques for used car price prediction. arXiv preprint arXiv:2002.06228.

U.S. News & World Report. "The Best Used Car Websites for 2021."

Vroom. (2021). Vroom Announces Fourth Quarter and Full Year 2020 Financial Results. Retrieved from https://investors.vroom.com/news-releases/news-release-details/vroom-announces-fourth-quarter-and-full-year-2020-financial

Xu, Y., Li, X., Xu, M., Chen, K., & Wang, G. (2020). Prediction of used car price based on deep neural network. Neurocomputing, 396, 152-160.