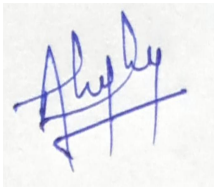
Project Summary

Batch details	PGPDSE-FT Gurgaon Oct22
Team members	Abhinav Tyagi, Aditya Kapur, Ankur Kumar, Jaswant Singh, Kritagya Kashyap, Lakshya Sharma, Muskaan Passi
Domain of Project	Predictive Analysis
Proposed project title	A Regression Study of Price Determinants of used cars in US Automobile markets
Group Number	5
Team Leader	Ankur Kumar
Mentor Name	Ms Vibha Santhanam

Date: 31/03/2023



Signature of the Mentor

Signature of the Team Leader

Table of Contents

1. Project Overview

For this project, we used the data set of used car listings in the USA, available on Kaggle, a popular platform for data-science competitions. The data set has over 4 lakh records and 26 features which contain information on various used cars in the USA, including their model, year of manufacture, condition and other relevant features such as engine type, transmission type, fuel type, and more.

The objective of this project is to build a machine-learning model that can accurately predict the price of a used car in the USA based on various features. The project consists of data preprocessing, feature selection, model selection, training and model evaluation.

2. Business problem statement

a) What would you achieve by this project?

The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, to make informed purchases. We will compare the performance of various machine learning algorithms and choose the best one out of them. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous values as output and not categorical values because of which it will be possible to predict the actual price of a car rather than the price range of a car.

b) How would this help the business or clients ?

These systems open new opportunities for businesses and individuals. As data scientists, we extract insights and behavioural patterns of fluctuating prices of used cars. Through a used car price prediction system we can help the buyers as well as the sellers to effectively determine the worthiness of the car using a variety of features.

c) What is the further scope of the project ?

Real-time data such as current market trends, seasonal variations, and demand and supply fluctuations could be integrated into the model to provide more accurate and up-to-date price predictions.

The project could collaborate with industry partners, such as car manufacturers or dealerships, to develop customized price prediction models that can be used for specific make, models and types of vehicles.

d) Limitations of the Project :

The accuracy of a used car price prediction model depends heavily on the availability and quality of data. Limited or incomplete data can result in inaccurate predictions and potentially harm the usability of the model. The used car market is highly dynamic and influenced by various factors, such as seasonality, changes in consumer preferences, and economic conditions. This makes it challenging to create a model that can accurately predict prices over a long period.

3. Topic Survey in brief

a) Problem understanding:

The main objective of this problem is to study the factors that affect the price of a vehicle in the second-hand market and predict its price. It is a complex problem that requires understanding the dynamics of the used car market. Some of the key factors that impact the pricing of used vehicles include the make and model of the car, its age and condition, mileage, location, and demand-supply dynamics in the market. Other factors such as the quarter's economic conditions, and the availability of financing options also play a role. To predict the price of a used car accurately, it is crucial to have a comprehensive understanding of these factors. Machine learning algorithms can be used to analyze large amounts of data and identify patterns that can help predict the price of a used vehicle.

b) Current solution to the problem:

The majority of the sales in the used car market still happen through offline channels. This is because of the consumer preference for the conventional mode of buying. Although, the online sales channel segment is witnessing significant growth over the years. The market has become more competitive due to the emergence of online tools designed for both buyers and sellers. Dealers are leveraging technically advanced tools that incorporate artificial intelligence and machine learning technologies to expand their customer base and network. By analyzing the data stored in their dealer management systems, AI applications
are optimizing marketing and sales strategies, resulting in an improved buying experience.

c) Proposed solution to the problem:

We will try to find a solution that can mimic or better the current statistical and machine learning techniques that go into predicting the price of used cars.

d) Reference to the problem (blogs, articles or startups in this domain):

Some startups in the US used car market :

1. Carvana
2. Vroom
3. Cars.com
4. The AutoTempest Blog
5. Own A Car Fresno Blog
6. The Shabana Motors Auto Blog
7. Car Time Auto Blog
8. We Buy All Cars Blog

4. Critical assessment of topic survey

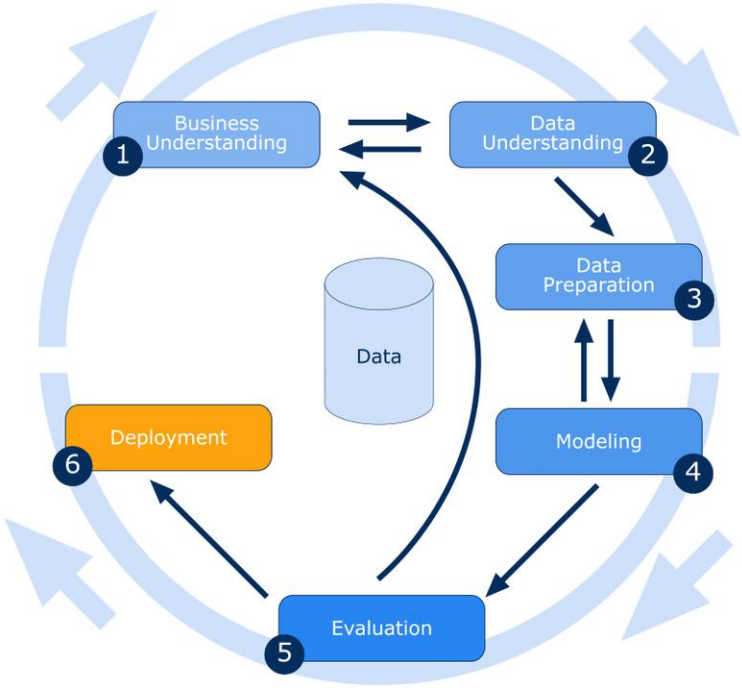
a) Find the key area and gaps identified in the topic survey where the project can add value to the customers and business

- It will improve clarity in the used car market by providing more accurate and reliable price estimates to buyers and sellers.
- It will reduce information imbalance between buyers and sellers, which can lead to more efficient and fair negotiations.
- It will be saving time and effort for buyers and sellers by providing a quick and easy way to estimate a car's value.
- It can help dealerships and car rental companies in the automotive industry to improve their profits by suggesting ways to set prices effectively. This involves analyzing market trends, customer behaviour and competition, and implementing strategies like value-based pricing and dynamic pricing to stay competitive and maximize revenue. Regularly monitoring and adjusting prices can also help improve profitability.

b) What key gaps are you trying to solve ?

- The project aims to improve the performance of the machine learning model by exploring different algorithms, feature engineering, and hyperparameter tuning.
- the project may aim to enhance the user experience of the web or mobile app by implementing a user-friendly interface and providing more accurate and detailed information on used car prices.
- Finally, the project may aim to expand the data set to include more diverse and relevant features that can improve the accuracy of price predictions.

5. Methodology to be followed



a) Business Understanding:

Determining the listing price of a used car is a challenging task, due to the many factors that affect a used vehicle's price in the market. The major focus of this project is to develop a machine-learning model that can accurately predict the price of a used car based on its features. This model will be useful to businesses involved in the retail of such vehicles. We will implement and evaluate various learning methods on the data set.

b) Data Understanding:

i. Variable Identification:

Independent: Id, URL, Region, Region URL, Year, Manufacturer, Model, Condition, Cylinders, Fuel, Odometer, Title Status, Transmission, VIN, Drive, Size, Type, Paint color, Image URL, Description, State, Lat, Long, Posting Date.

Target: Price

The Data set has 426880 rows and 26 columns

ii. Variable Information:

- **Price:** Price of the used cars (The estimated price of the car listed by the seller)
- **Id:** This column tells us about the unique ID for each unique category/row.
- **URL:** This column tells us about the unique URL for each listing with complete details.
- **Region:** Region of where the car is from.
- **Region URL:** URL of a specific region of the listing of the car.
- **Year:** In which year the car was purchased
- **Manufacturer:** This column consists of the Company names, the car belongs to.
- **Model:** Name of the car.
- **Condition:** The condition in which the car is kept.
- **Cylinders:** No. of cylinders present in the car.
- **Fuel:** The kind of fuel on which the car operates.
- **Odometer:** Number of miles travelled by the car's previous owner (Distance that car has travelled).
- **Title Status:** Current Status of vehicles.
- **Transmission:** Mode of power transfer in a vehicle (automatic/manual etc).
- **VIN:** Unique Vehicle identification number.
- **Drive:** Type of drive (rear-wheel drive, front-wheel drive, 4-wheel drive).
- **Size:** Segment of the car (6seater/4seater, etc).
- **Type:** Type of the vehicle.
- **Paint Color:** The colour of the car.
- **Image URL:** This column shares the links where we can see the images of the cars.
- **Description**: Detailed description of the condition of the car.
- **County:** This column has all values as Nan so we will drop this column.
- **State:** The area where the car is situated.
- **Lat:** Latitudinal position of the car.
- **Long:** Longitudinal Position of the car.
- **Posting Date:** Date of posting of the ad.

c) Data Preparation:

i. Data Filtering:

We started by checking the data-types of the the data set using *info()* method. We found that *posting_date* had object data-type, which was converted into *datetime* data-type.
Next we checked the *5-point summary* of our data using *describe* method.

	id	price	year	odometer	county	lat	long
count	4.268800e+05	4.268800e+05	425675.000000	4.224800e+05	0.0	420331.000000	420331.000000
mean	7.311487e+09	7.519903e+04	2011.235191	9.804333e+04	NaN	38.493940	-94.748599
std	4.473170e+06	1.218228e+07	9.452120	2.138815e+05	NaN	5.841533	18.365462
min	7.207408e+09	0.000000e+00	1900.000000	0.000000e+00	NaN	-84.122245	-159.827728
25%	7.308143e+09	5.900000e+03	2008.000000	3.770400e+04	NaN	34.601900	-111.939847
50%	7.312621e+09	1.395000e+04	2013.000000	8.554800e+04	NaN	39.150100	-88.432600
75%	7.315254e+09	2.648575e+04	2017.000000	1.335425e+05	NaN	42.398900	-80.832039
max	7.317101e+09	3.736929e+09	2022.000000	1.000000e+07	NaN	82.390818	173.885502

Figure 1, *5-point summary of data*

We found that minimum value of *price* and *odometer* variables is 0. Further on checking the null values in the data using *isnull* method and data using the *head* method, we made the following observations and inferences:

- **Id** - It has only unique values and no underlying pattern, so it doesn't give any useful information.
- **county** - It has 100% null values.
- **Lat, long** - They have the co-ordinates of the location of the car. We also get the broader location information from *region* and *state. Lat, Long* as such don't provide any useful information.
- **image_url, region_url** - They contain web-links, and as such don't provide any useful information.

Therefore we dropped the above mentioned 6 columns.

Next we checked and dropped the duplication in records using the *drop_duplicates* method. After this initial filtering, the size of our data set was 2,99,509 rows and 20 columns.

From earlier we know that our target variable *price* has 0 as minimum value. Using the *count* method we found that there are 20,070 records where *price* of vehicle is 0. We dropped these records from our data.We saved these records separately in a new data-frame *absurd*, to use them as unseen data for our model.

Using *boxplot* and *distplot* we try to visualize our target variable, *price.*

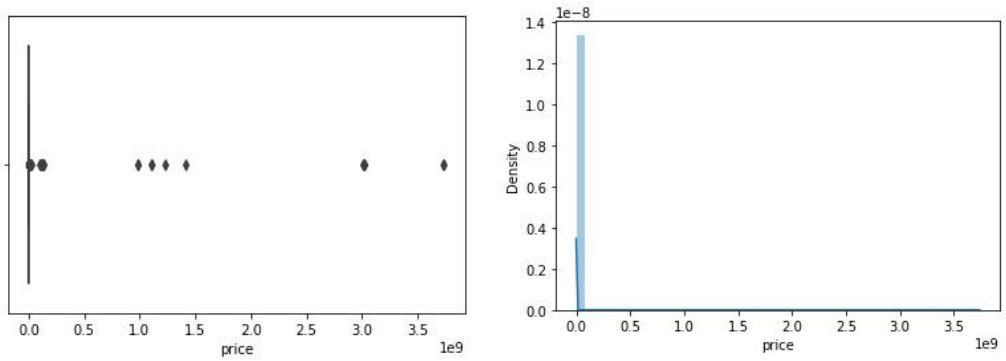


Figure 2, box-plot and dist-plot of the data set

The visualizations suggest that there is some noisy data in our data set.

Using domain knowledge, we conclude that price of even a used car below 500$ is absurd.There are 5589 records with price less than 500$. We add these records to the absurd data set.Using IQR method for outlier analysis, we find the upper limit of target variable to be 57,000$.

Subsequently we split our data set into car1 (data set without outliers) and car 2 (data set of outliers only). car1 has records with price greater than 500$ and less than 57,000$. car2 has records greater than 57,000$.

ii. Uni-variate Analysis:

Visualizing each feature to get a better understanding of data and its distribution

- Numerical Columns:

Price:

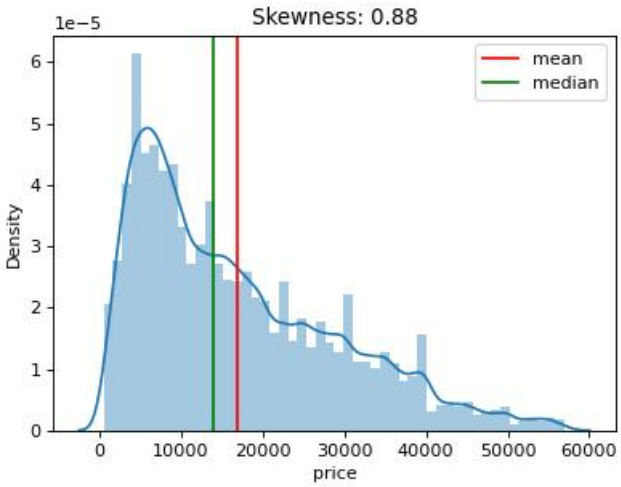


Figure 3, dist-plot of price (target variable)

Inference: The data is Right Skewed with the skewness of 0.88.

Odometer:

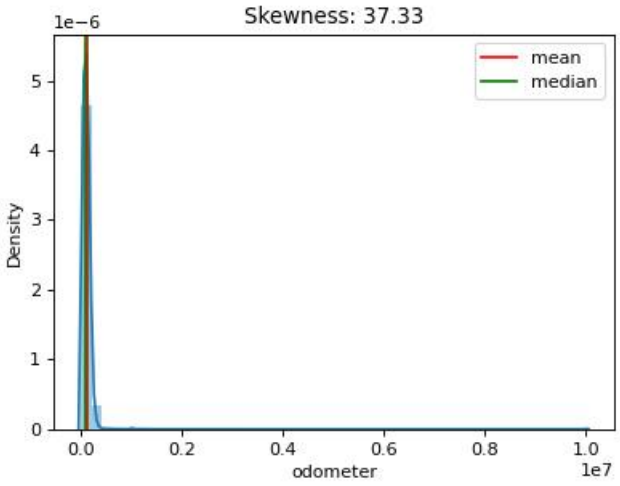


Figure 4, dist-plot of odometer

Inference: The data is Right skewed data with the skewness of 37.33

- Categorical Columns:

Regions:

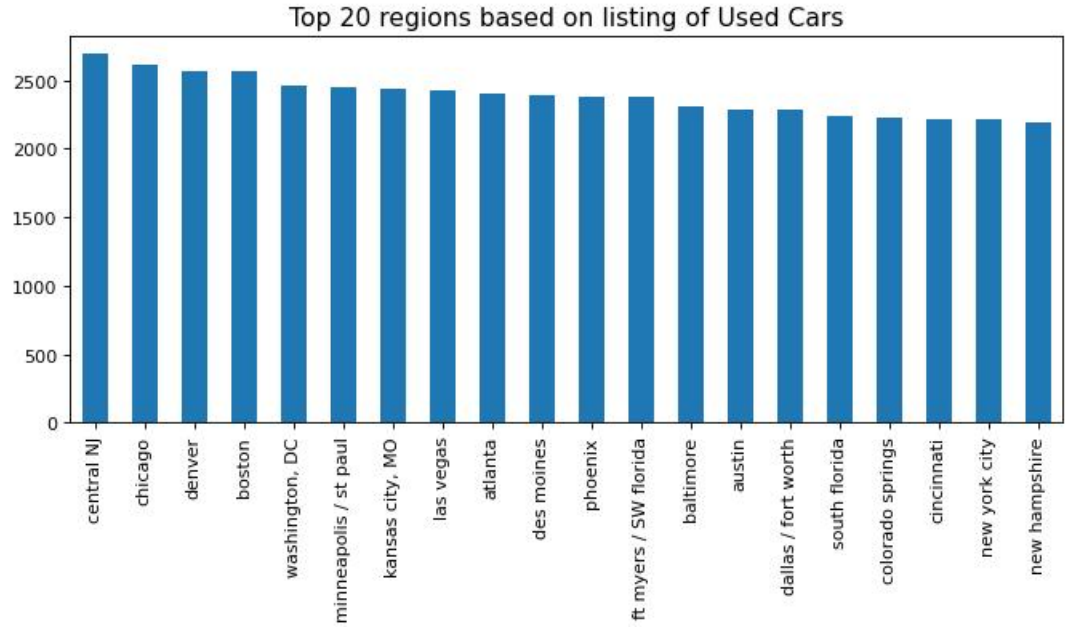


Figure 5, bar-plot for top 20 regions based on number of car listings

Inference: *Central Nj* have highest number of listed used cars.

Manufacturer:

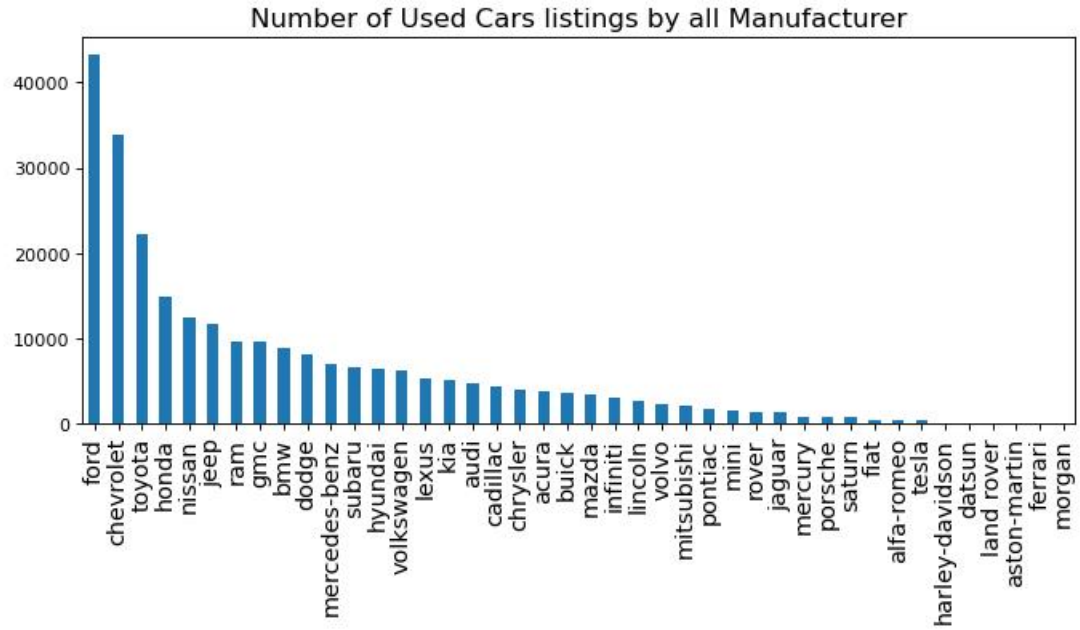


Figure 6, bar-plot for frequency of cars sold for each manufacturer

Inference: The top 3 most popular listed used car manufacturers are ford, chevrolet, and toyota while the least 3 are morgon, ferrari and aston-martin.

Model:

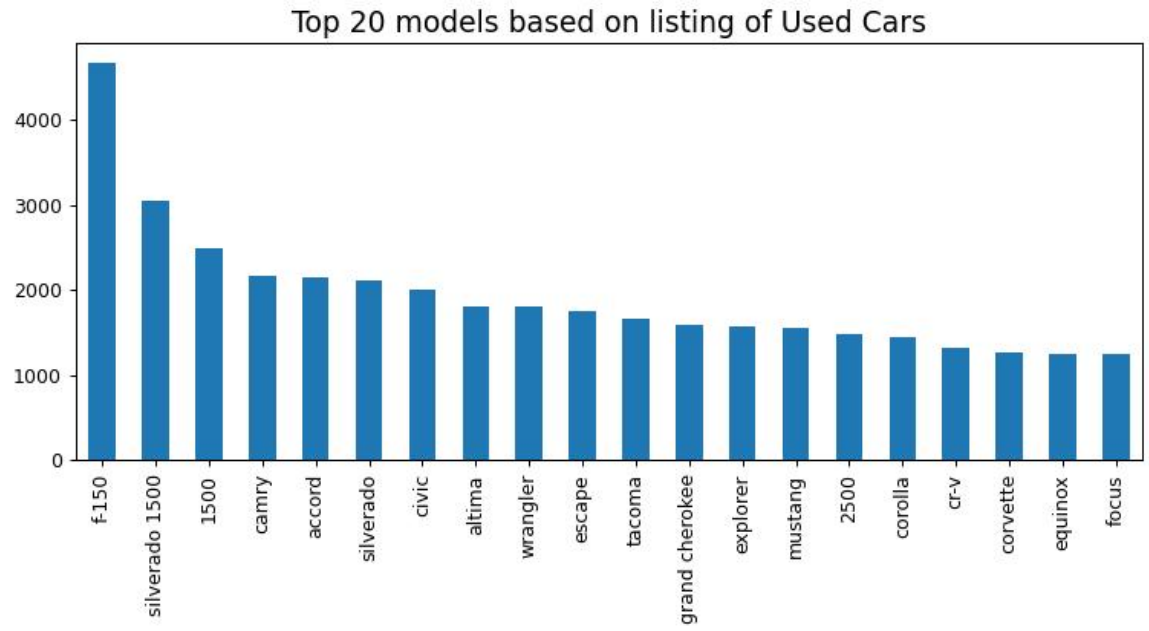


Figure 7, bar-plot for top 20 models based on number of car listings

Inference: f-150 model car is highest listed in the used car market which belongs to Ford.

Conditions:

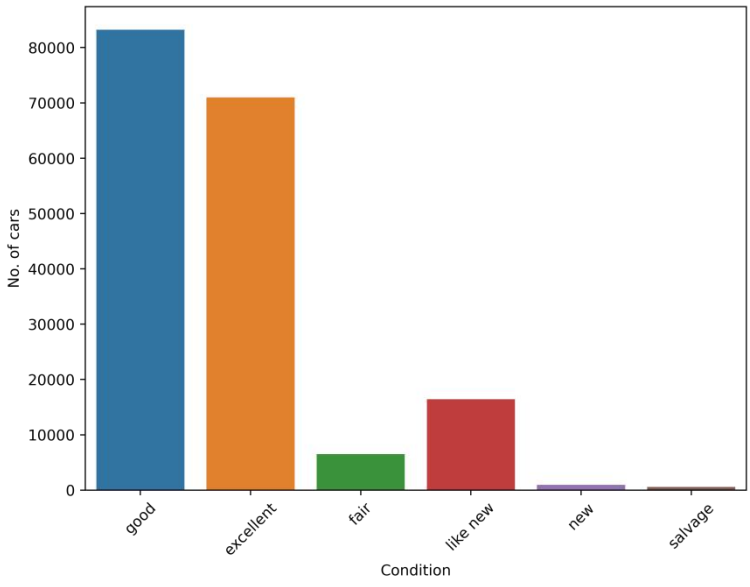


Figure 8, bar-plot for conditions of listed used car

Inference: Good condition car is most listed which contributed around 50% of the distribution followed by excellent condition which contributes around 38%.

Cylinders:

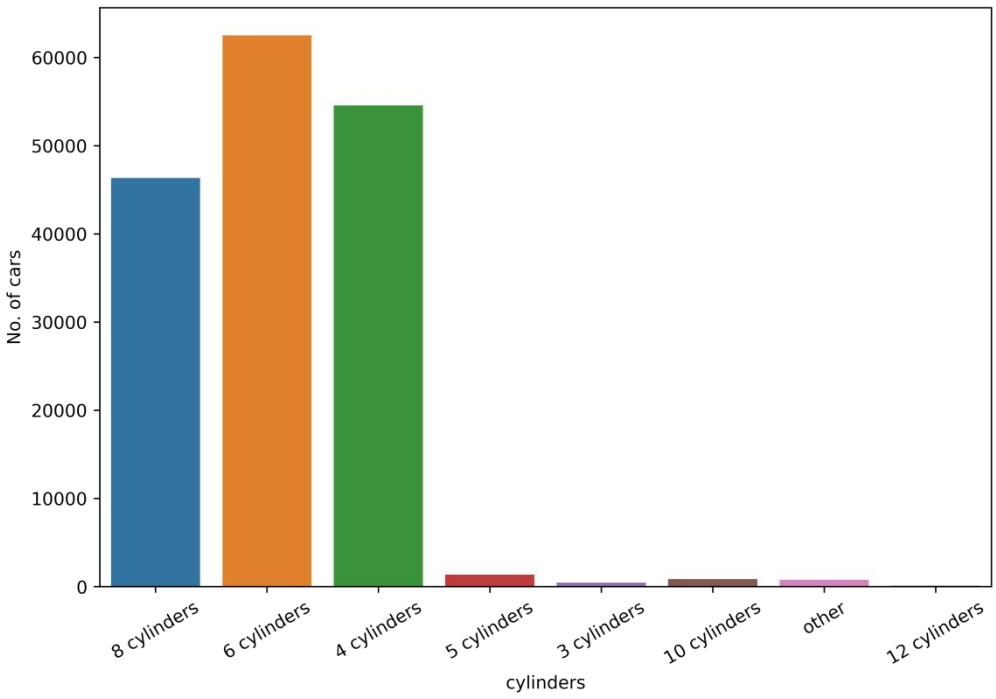


Figure 9, bar-plot for Number of Cylinders in listed car

Inference: Cars with 6, 4, and 8 cylinders are the most popular listed in used car market. This is expected because most cars are equipped with cylinders in that range.

Fuel:

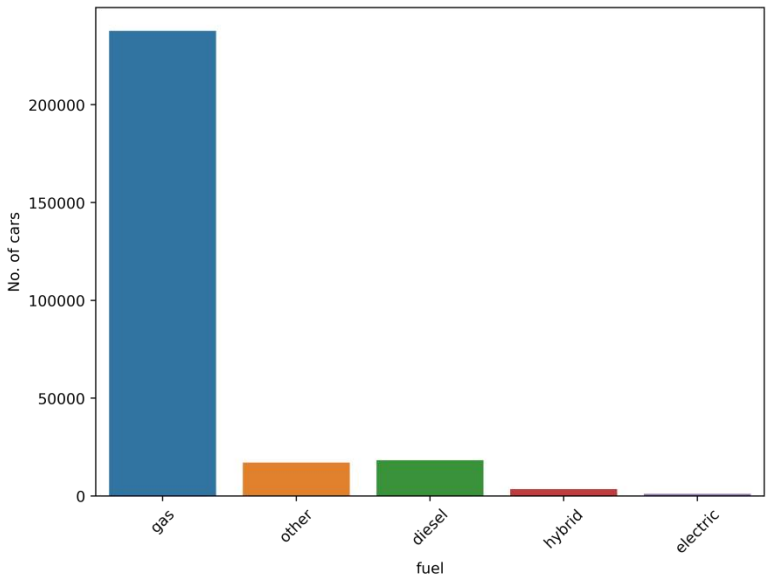


Figure 10, bar-plot for Number of Cylinders in listed car

Inference: Most of the cars whose price range is between 500 to 57,000 have fuel type gas. As expected, gas or petrol is the most common type of fuel for the listed cars. Very few of them are electric cars.

Title Status:

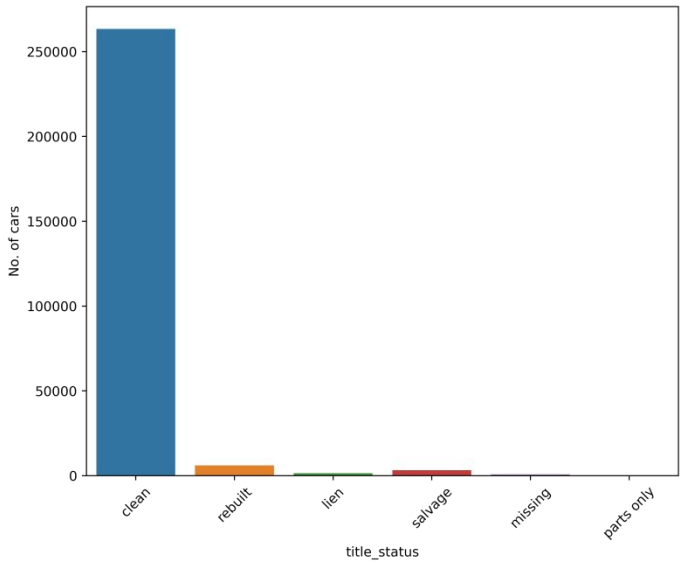


Figure 11, bar-plot for Distribution of Title Status in listed used car

Inference: Clean is the most occurring title status for the used car listings.

Transmission:

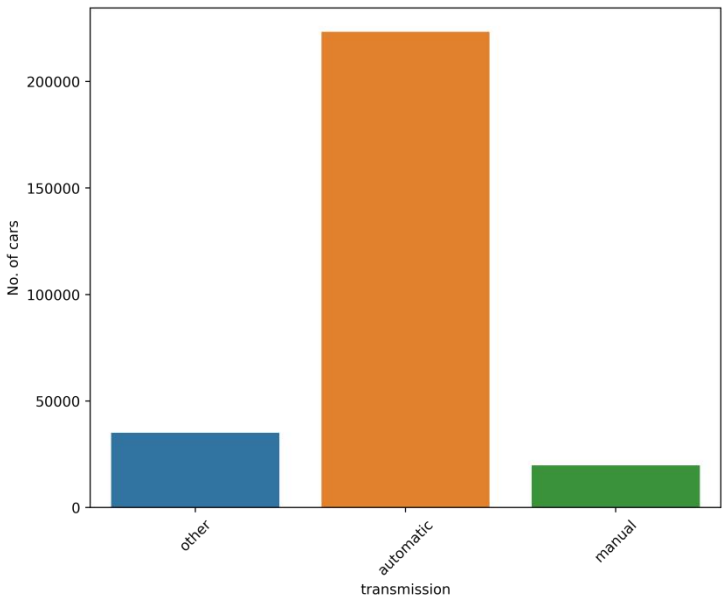


Figure 12, bar-plot for Type of Transmission in listed car

Inference: The most common car transmission type is automatic as expected.

Drive:

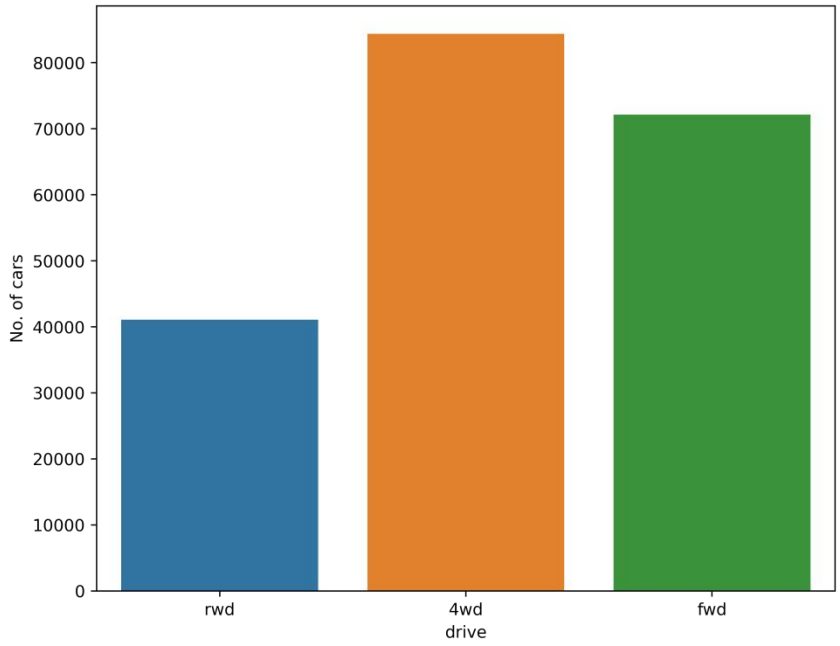


Figure 13, bar-plot for Type of Drive available in listed car

Inference: 4 wheel drive is the most commonly listed car drive followed by front wheel drive Rear wheel drive is the least in used car market 1.

Size:

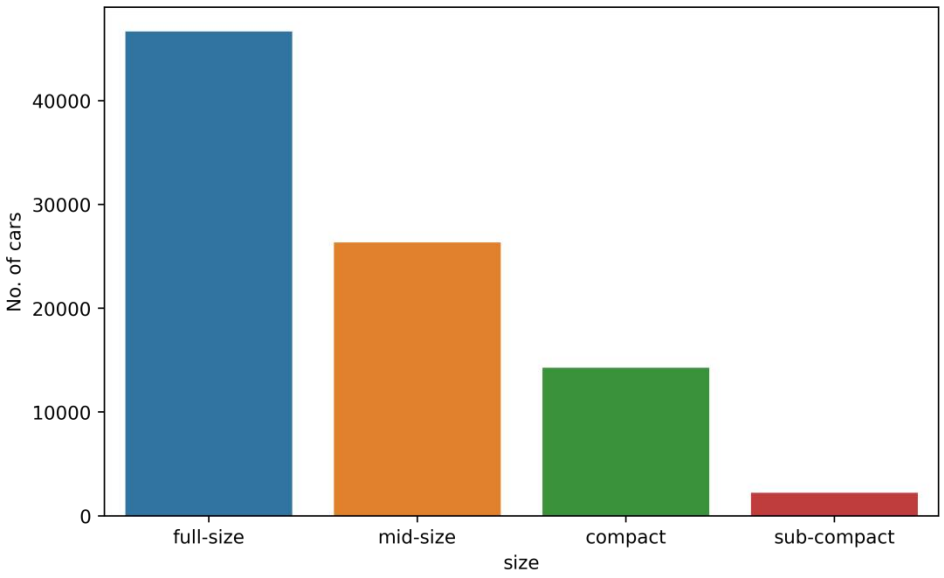


Figure 14, bar-plot for Size available in listed car

Inference: Full-size cars are most listed in used car market. Generally, sedan type cars are considered as full-sized cars in USA.

Type:

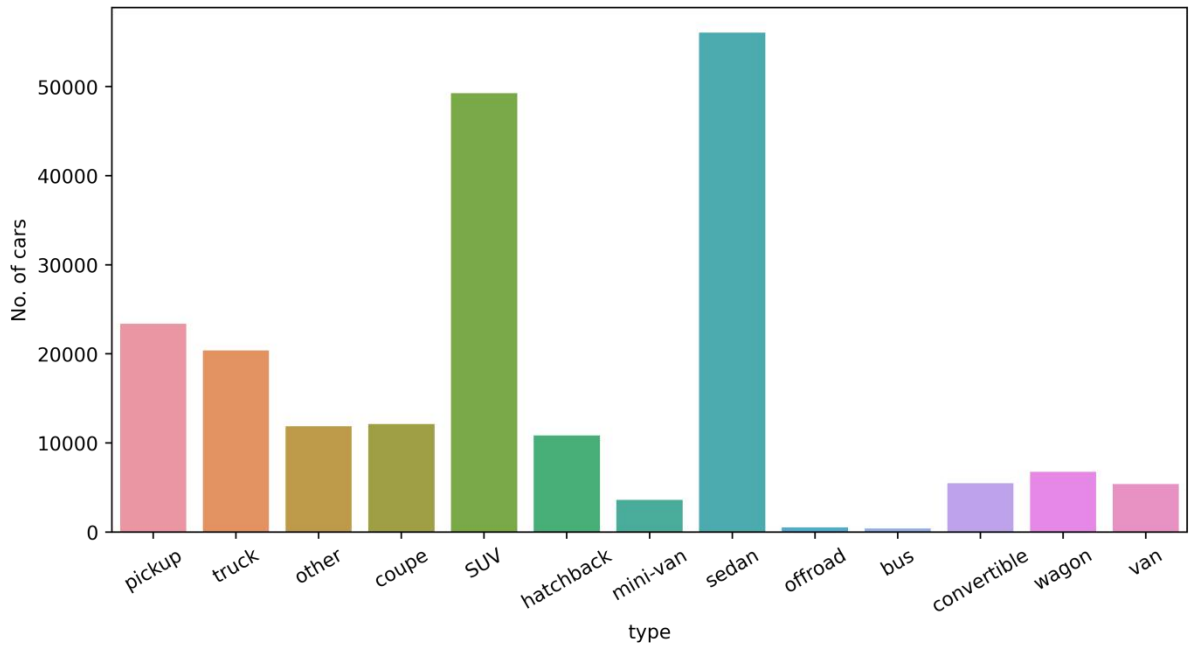


Figure 15, bar-plot for Type of listed car

Inference: Sedans, SUVs are the two most popular car listings followed by pickups and trucks.

Paint Color:

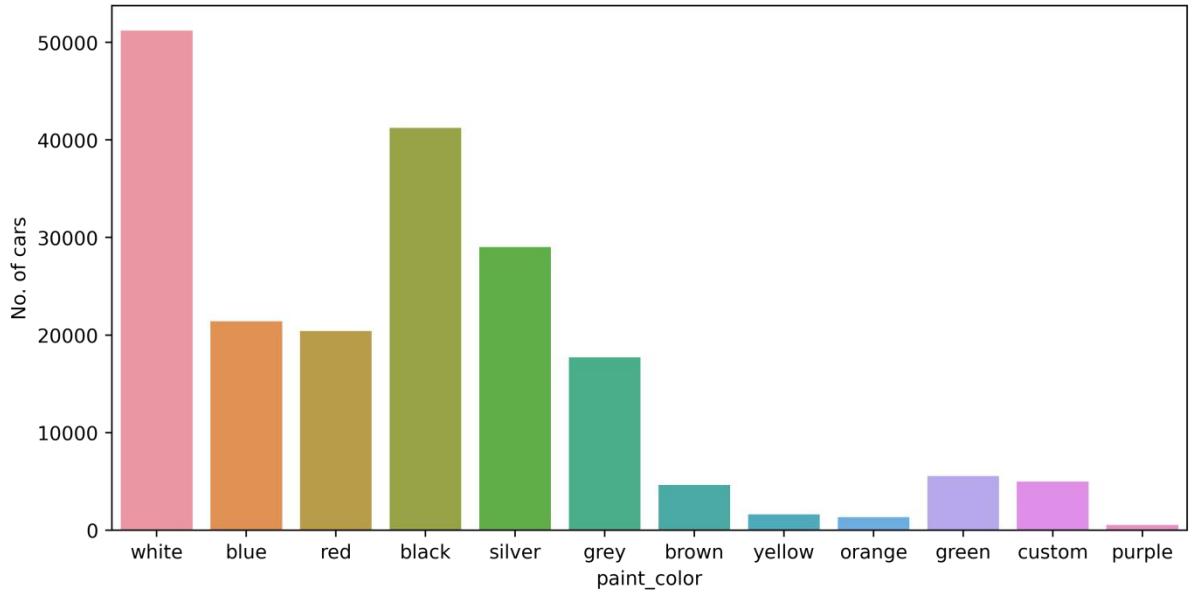


Figure 16, bar-plot for paint color in listed car

Inference: White and black are the two most popular colors of cars being listed, followed by silver.

State:

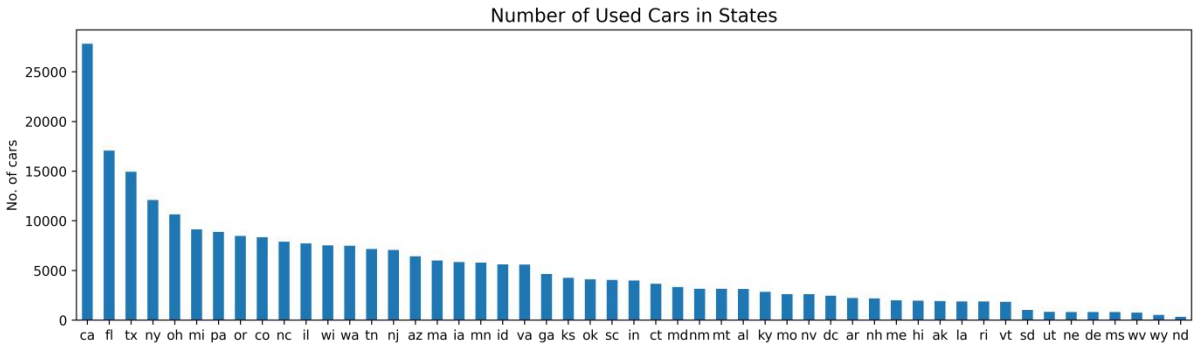


Figure 17, bar-plot for frequency of cars sold for each State

Inference: In used car market most of the cars are from California. Top 5 states are California, Florida, Texas, New York, Ohio in used car market.

iii. Bi-variate Analysis:

Visualizing each feature to get a better understanding of data and its distribution

Condition vs Price:

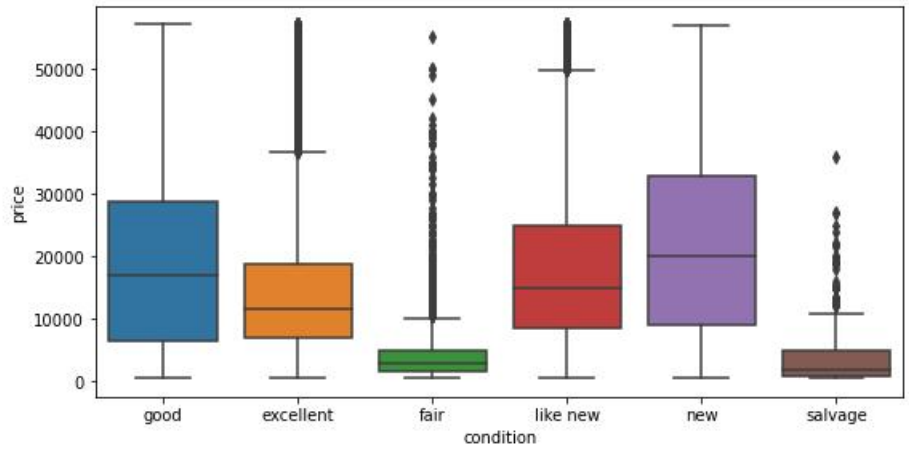


Figure 18, box-plot of condition vs price

Inference:

- Newly purchased cars have higher Average price.
- Salvaged cars have the least Average price.
- There are outliers present in Excellent, fair, like new and salvage condition categories.

Cylinder vs Price:

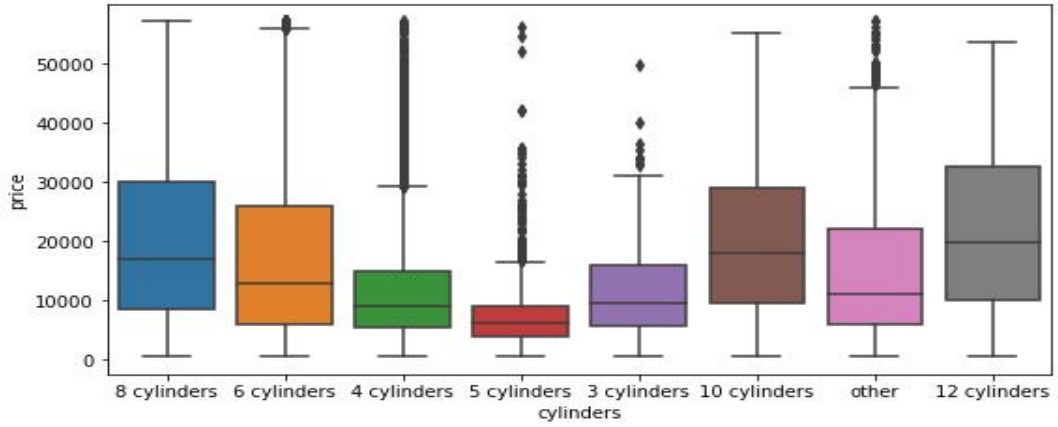


Figure 19, box-plot of cylinder vs price

Inference:

- Cars having 12 cylinders have maximum average price followed by 10 and 8 cylinders.
- There are outliers present in each of the cylinder categories except 8, 10 and 12.

Fuel vs Price:

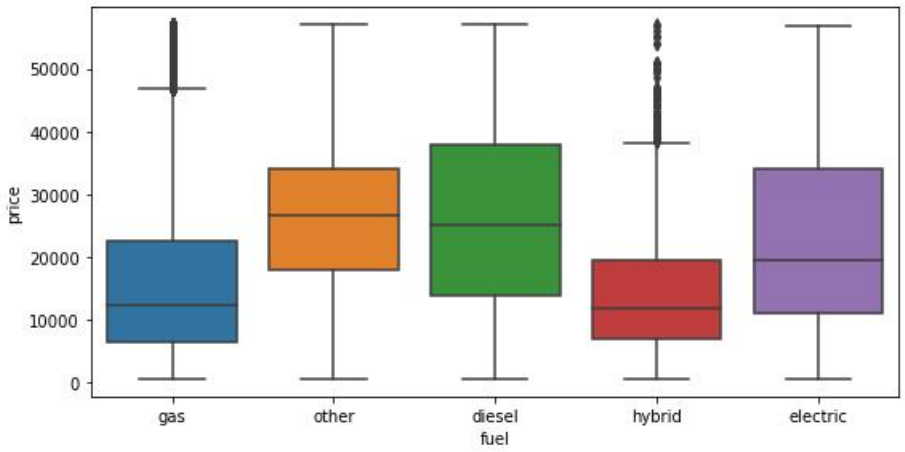


Figure 20, box-plot of fuel vs price

Inference:

- As we can see Average cost of 'other' category fuels is the highest and as all the major fuel types are already displayed, we can interpret that 'other' category fuels may contain exotic/costly fuels.
- 'Diesel' has 2nd highest Average price in fuel type.
- The fuel types with Least Average price are gas and hybrid.
- There are outliers present in gas and hybrid fuel categories.

Title Status vs Price:

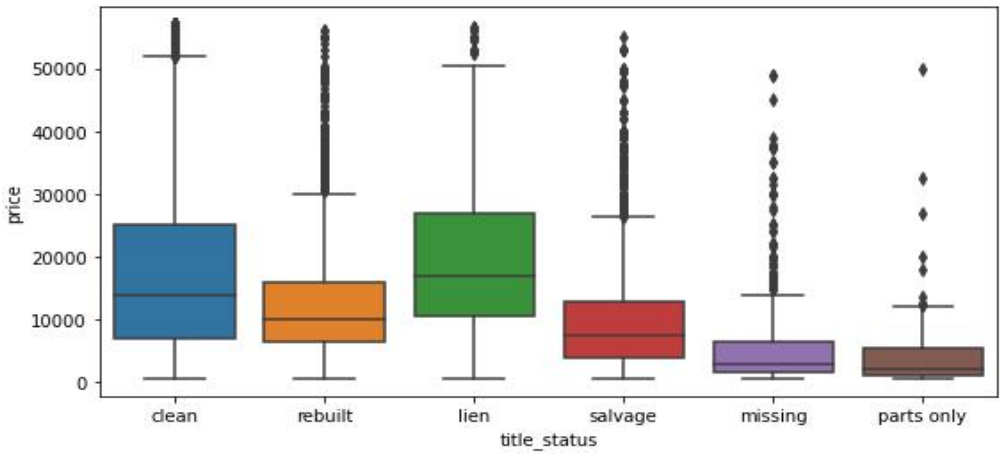


Figure 21, box-plot of title_status vs price

Inference:

- Cars with title_status 'lien' has the highest average price.
- There are outliers present in each of the title_status categories.

Transmission vs Price:

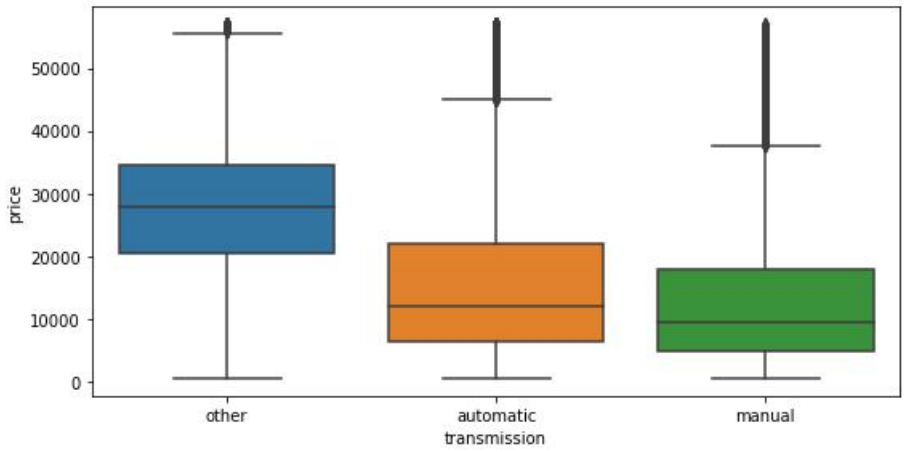


Figure 22, box-plot of transmission vs price

Inference:

- 'Automatic' has higher average price than 'manual' cars
- In transmission category 'other' has the highest average price.
- There are outliers present in each of the transmission categories.

Drive vs Price:

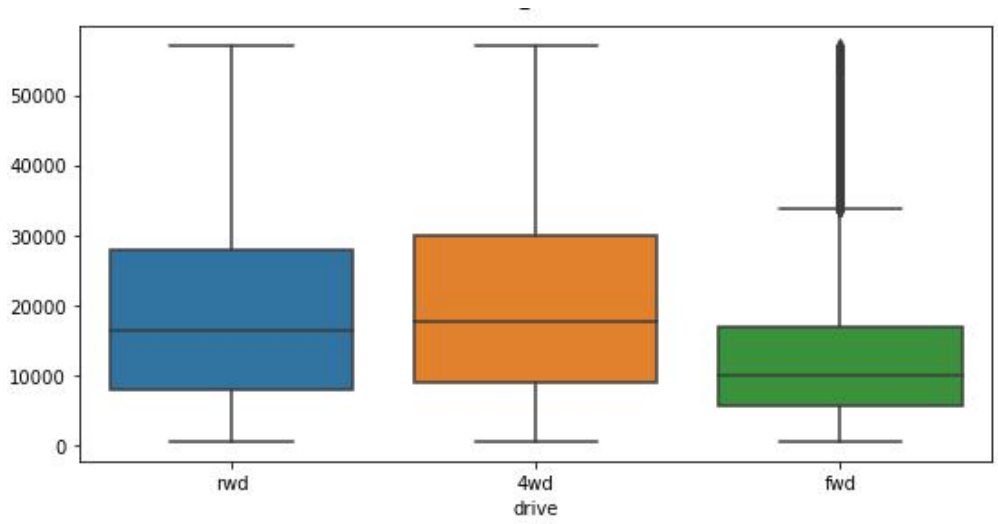


Figure 23, box-plot of drive vs price

Inference:

- 4wd (4 wheel drive) has the highest average price.
- From business understanding we can say that 4wd are commonly found in SUV's and pickup trucks which are usually on the higher price side.
- There are outliers present in fwd drive category.

Size vs Price:

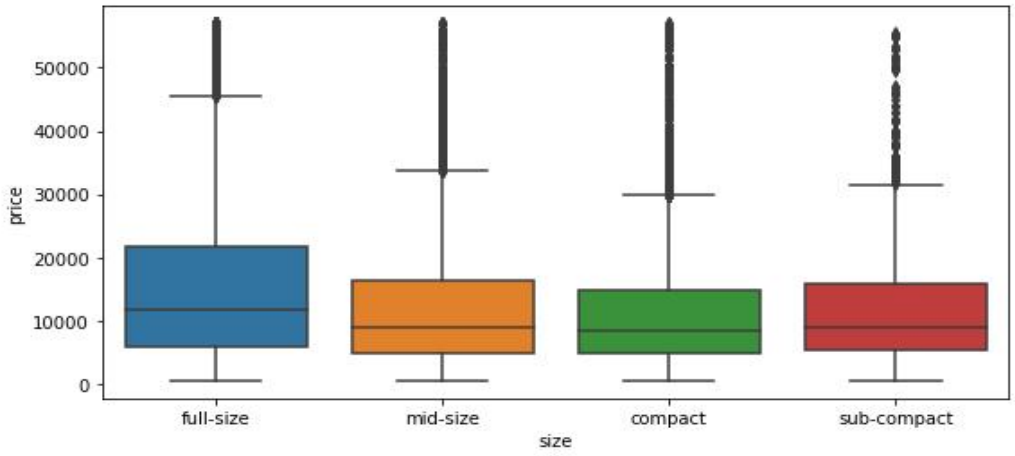


Figure 24, box-plot of size vs price

Inference:

- 'full-size' cars have the highest average price.
- There are outliers present in each of the size categories.

Type vs Price:

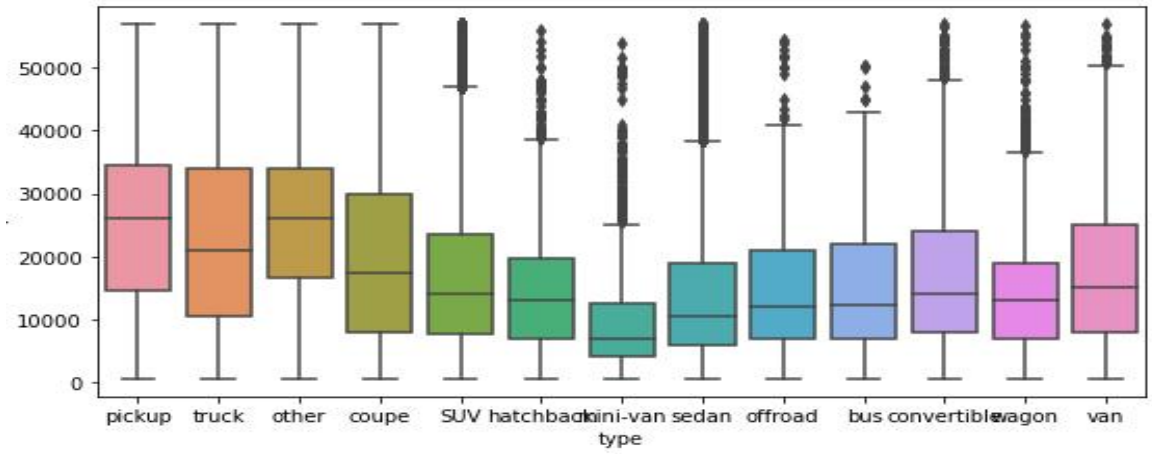


Figure 25, box-plot of type vs price

Inference:

- Cars type 'pickup' has the highest average price.
- Mini-van has the least average price.
- There are outliers present in each of the type categories except pickup, truck, coupe.

Paint Color vs Price:

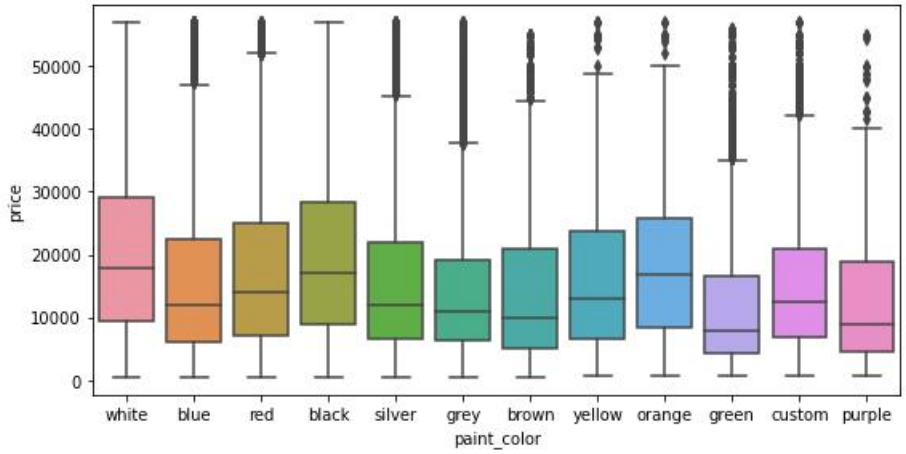


Figure 26, box-plot of paint color vs price

Inference:

- Cars with 'paint_color' white and black have the highest average price.
- There are outliers present in each of the paint_color categories except white and black.
- Cars with paint-color green has the least average price.

iv. **Missing value treatment:**

Variable name	Missing values	Treatment
Year	0.2822 %	• URL • VIN (using *vin* method in python) • Mode.
Manufacturer	4.13 %	• VIN (using *vin* method in python) • Model • URL • Rest binned as unknown
Condition	40.78 %	• Title Status • Forward fill
Fuel	0.70 %	• Mode values on basis of model and manufacturer
Cylinder	41.62 %	• Description • Model • Manufacturer • Electric cars binned as other
Odometer	1.03 %	• All values > 999999 set to 999999 • 0 and Null filled with mean w.r.t model • rest filled with median
Transmission	0.59 %	• Mode values on basis of model and manufacturer
Type	21.75 %	• Mode values on basis of model and manufacturer
Drive	30.58 %	• Mode values on basis of model, manufacturer and type
Size	71.76 %	• Mode values on basis of model, manufacturer and type
Paint Colour	30.50 %	• Mode values on basis of model, manufacturer and type

v. Feature Engineering:

Age:

It is a Feature which directly or indirectly affects the price of the used car in the market. From the Business Knowledge we came to conclusion that as the age of the car increases price of the car decrease.

This feature was extracted by taking difference between car purchased year and car posting year which can be extracted from posting date column.

State_Zone:

This can be used which zone the car is present in the USA market. We divided the state in 4 Zone which are Northeast, South, West and Midwest.

This Feature was extracted with the help of State column.

is_tax:

This feature helps us to tell to whether we have to pay tax on selling the used car. Through the research we came to conclusion that there are only 5 states in which we doesn't have to pay tax else we have to pay tax.

We used the State column to extract this feature.

vint_car_antique:

This feature can be used tell us which type of used car is listed whether it is a Vintage Car, Classic car, Antique car or modern car. The price of used Classic car, Vintage car or Antique car is higher comparison to modern car as they are used as collectable items.

We used the year column to extract this column if year is less than 1930 then it is a Vintage car, if it is less than 1977 then it is a Antique car, if it is less than 2002 it is a classic car else it is a modern car.

vi. Statistical Tests:

Significant Variables		
Features	**Test**	**P value**
Year	Mann-Whitney U	0.0
Odometer	Mann-Whitney U	0.0
Condition	Anova	0.000001
State	Anova	0.032514
Is tax	T-test independent	0.009575
Age	T-test independent	0.0005157
Vint_car	Anova	0.000009
Model	Due to high categories we can't run statistical test and from domain knowledge we keep this variable for modelling.	

Non significant variables		
Features	**Test**	**P value**
Region	Anova	1.0
Manufacturer	Anova	1.0
Cylinder	Anova	0.390706
Fuel	Anova	0.964539
Transmission	Anova	0.270099
Drive	Anova	0.057264
Size	Anova	0.208265
Type	Anova	0.388538
Paint colour	Anova	0.294331
State Zone	Anova	0.588757

vii. Encoding:

Categorical columns would be encoded into numerical value as the machine can only process numerical values.

S.No.	Encoding technique	Variable name
1	Label encoding	Region, Manufacturer, Type, Paint Color, State, State Zone
2	Ordinal encoding	Cylinders, Condition, Size
3	Target encoding	Model
4	One-hot encoding	Fuel, Transmission, Drive, Vint_car

d) Modeling:

After train test split, we applied Decision Tree as our base model.

i. Base Model:

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
Decision Tree	0.999604	0.696382	247.414093	6816.5272	40.538615

Decision Tree Regressor is over-fitting as its *Train R2 score* is 99.9% and *Test R2 score* is 69.6%.

ii. Comparison and Model Selection:

We applied ensemble techniques Ridge, Lasso, Elastic Net.

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
Ridge	0.551710	0.550717	8323.137291	8291.999291	65.752941
Lasso	0.551695	0.550725	8323.279008	8291.927331	65.720329
Elastic Net	0.420750	0.422053	9461.076248	9404.668752	62.097979

Ridge's performance is barely above average (50% to 95%) as its *Train R2 score* is 55.17% and Test R2 Score is 55%.

Lasso' performance is approximately same as Ridge as its *Train R2 score* is 55.16% and Test R2 Score is 55%.

Elastic Net is under-fitting model as it only gives 42% *Train R2 score* and 42.2% Test R2 Score.

After that we applied the Random Forest Model.

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
Random Forrest	0.978557	0.851884	1820.349500	4761.028578	43.168481

Random Forest Model gives us 97.8% Train R2 score and 85.1% Test R2 score. From R2 score we can conclude that it is also a over-fitting model like Decision Tree Regressor.

Now we apply Boosting algorithms like Ada-boosting Regressor, Gradient Boosting Regressor and XGBoost Regressor.

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
Ada-boosting	0.635130	0.635568	7508.906508	7468.054537	60.193155

Gradient Boosting	0.781405	0.783581	5812.027572	5755.016932	50.773069
XG Boosting	0.856421	0.837678	4710.345448	4984.118219	42.128192

AdaBoost model gives us better performance then all the above model as its Train R2 score is 63.5% and Test R2 score is 63.5%. We are looking for model which perform better than AdaBoost Model,

Gradient Boost Model gives us better performance then AdaBoost model as its Train R2 score is 78.1% and Test R2 score is 78.3%.

For more better performance we also try **XGBoost Regressor** and is gives us better performance the Gradient Boost Model. It gives us the Train R2 score 85.6% and Test R2 score is 83.7%.
We also try CatBoost Regressor for check it performs better than XGBoost

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
CatBoost	0.851534	0.841018	4789.844196	4932.564387	42.989694

CatBoost gives us 85.1% Train R2 score and 84.1% Test R2 score. On comparing the XGBoost and CatBoost we found that accuracy difference is less in CatBoost Model. And also RMSE difference is less in CatBoost Model.

So our best model is CatBoost Regressor.

iii. Hyper Parameter Tuning:

In order to improve the accuracy of our model we do parameter tuning.

Model name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
Tuned CatBoost	0.882055	0.851753	4269.212741	4763.129450	41.563495

The **Tuned CatBoost** gives us Best R2-score at *learning rate **0.3***. It gives us the Train R2 score is 88.2% and Test R2 score 85.1%.

So we finalized the Tuned CatBoost Model is our Final model.

Feature Significance:

Using Feature Importance method we visualize the significant features in our best fit model.

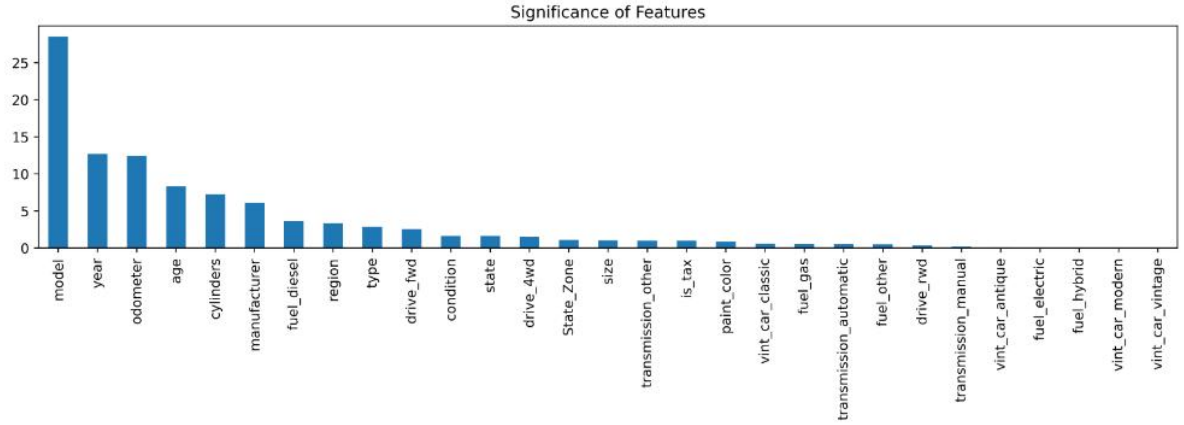


Figure 27, bar-plot showing significance of features

Conclusion:

1. Used cars with high Mileage are cheaper

2. Used cars with better Appearance are expensive.

3. Used cars which come from big manufacturer are cost more.

4. Used cars with pickup or truck or coupe or convertible type are cost more.

5. Used cars with white or black paint-color are expensive.

6. Used cars with more cylinders are expensive.

7. Used cars with higher vehicle age are cost less.

6. Intermediate milestones (based on project deadlines)

Week	Task	Meeting	Deliverable
Week 1	Synopsis	Group meetings Group meeting with Mentor	Synopsis submission (03 Feb)
Week 2	EDA + Feature engineering	Group meetings Group meeting with Mentor	Work Progress Report 1
Week 3	Feature and target variable analysis	Group meetings Group meeting with Mentor	Work Progress Report 2
Week 4	Basic model fitting	Group meetings Group meeting with Mentor	Interim Presentation and Report (23 Feb)
Week 5	Best fit model (fine-tuning the model)	Group Meetings Group Meeting with Mentor	Work Progress Report 3
Week 6	Prepare final report and presentation.	Group Meetings Group Meeting with Mentor	Final Report (31 Mar)
Final Presentation (1 Apr)			

7. References

- Reference documents of CRISP-DM
  - https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
  - https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
- data set Source: Used Cars Data set Vehicles listings from Craigslist.org
  Retrieved from Kaggle (Link)
- Article: India: Used Car Market - Growth, Trends, Covid-19 Impact, and Forecast
  Retrieved from https://www.mordorintelligence.com/industry-reports/india-used-car-market
- Journal: Irjet Journal: Used car price prediction
  Retrieved from https://www.academia.edu/51235585/Used_Car_Price_Prediction
- Machine Learning Workflow – Process stops
  Retrieved from https://www.gatevidyalay.com/tag/machine-learning-process-diagram/
- Enes Gocse, Predicting used car prices using machine learning techniques
  Retrieved from https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952
- Report: Used Car Market Size and Share Report 2022-2030
  Retrieved from https://www.grandviewresearch.com/industry-analysis/used-car-market

Notes For Project Team

The original owner of the data	Austin Reese
Data set information	Used Cars data set: Vehicles listings from Craigslist.org
Any past relevant articles using the data set	*Medium article:* Exploratory data analysis and model building of Craigslist used car data set
Link to data set	https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

****