USED CAR PRICE PREDICTION

Aditi Aserkar

Bhagyashri Raghunath Gadkari

Shilpa Nidhi Kirubanidhi

Group No:4

Problem definition:

We are using the used cars dataset from craigslist to predict the prices of the cars using machine learning models. We also identify important features of the car that influence its price in the market. This prediction will be of use to:

- Used car dealers: Dealers can benefit from this as they can better understand the important features of the car that influence the price, and what makes a car desirable.
- Online car pricing websites: This prediction may be used as a point of reference pricing websites and as a result, they can list a better price estimate.
- Individuals who are interested in selling/purchasing a car: People can better understand the market value of their used cars and fix a price accordingly.

Data set:

https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

This data from craigslist contains 26 fields.

Data cleaning:

Step 1. Drop the unnecessary columns:

Id: Field that contains an identifier for each record

url: URL of the car posting

region, region_url, lat, long: removed as we are using State field for location dimension

VIN: Vehicle identification number

image_url &description: details of the car (image and description)

size: More than 50% contain NULLS

county: Dummy field with 100% NULL values

Step 2: Drop the duplicate records.

Step 3: Filling NULL values of categorical columns:

For the categorical columns namely transmission, cylinders, title_status, fuel, paint_color, drive, manufacturer, type and model, we are using the forward fill (ffill) option to fill the NULL values.

The ffill() method replaces the NULL values with the value from the previous row (or previous column, if the axis parameter is set to 'columns').

Step 4: Filling NULL values for numerical columns:

Odometer:

To deal with NULLS in odometer field, we are first creating a new field namely AGE of a car. This is done by subtracting the date of advertisement posting (Posting_date) from the year field.

Age=Posting_Date-year

We then find the mean value of odometer reading based on the age and condition of the car. This value is used to fill the NULL values. Ex: If a car is old and in good condition, we use the mean value of all cars in the same age and condition.

Step 5: Removing outliers

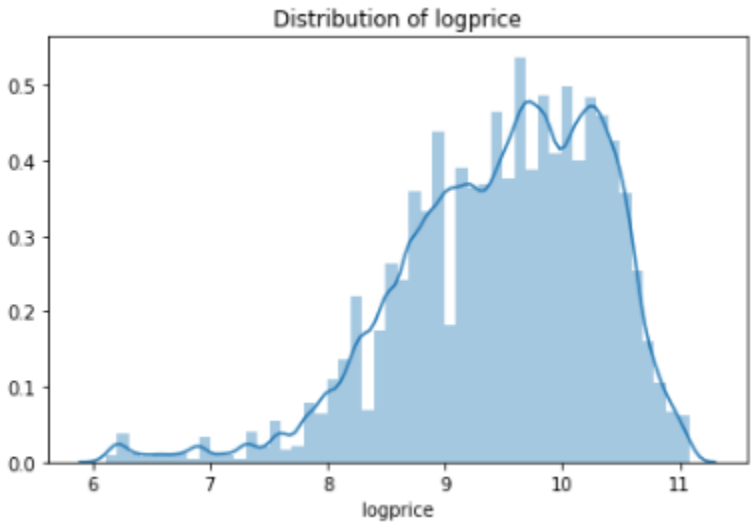
For the odometer and price fields, we are removing the outlier data by calculating quartiles. We then filter out the data that fall outside of the quartile limits.

Exploratory data analysis:

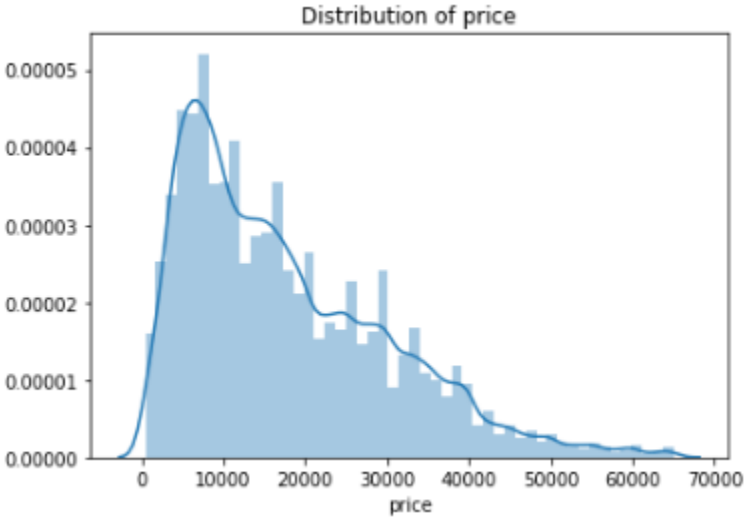
Data visualization is the representation of data and information in a graphical or visual format.

Here, after the data is cleaned, we do some visualization to clear the relation more clearly between different factors. Some of the visualizations are defined below:

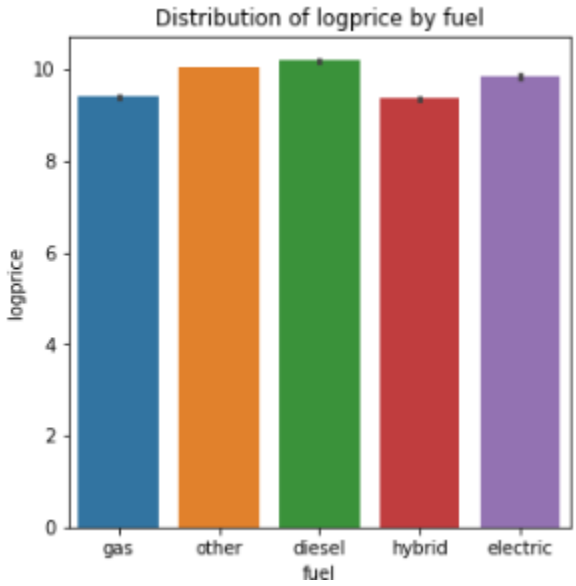
1. The distribution of logprice in the clean dataset:

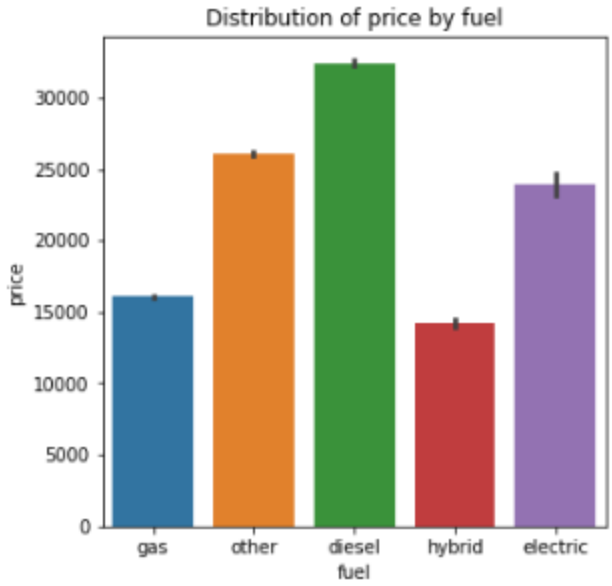


2. The distribution of the price in clean dataset:

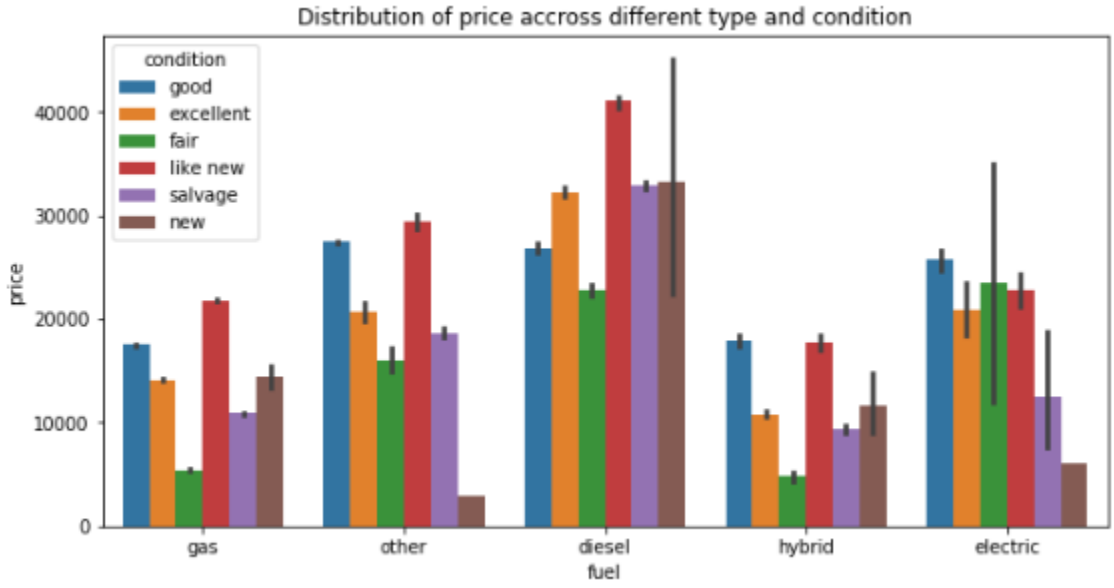


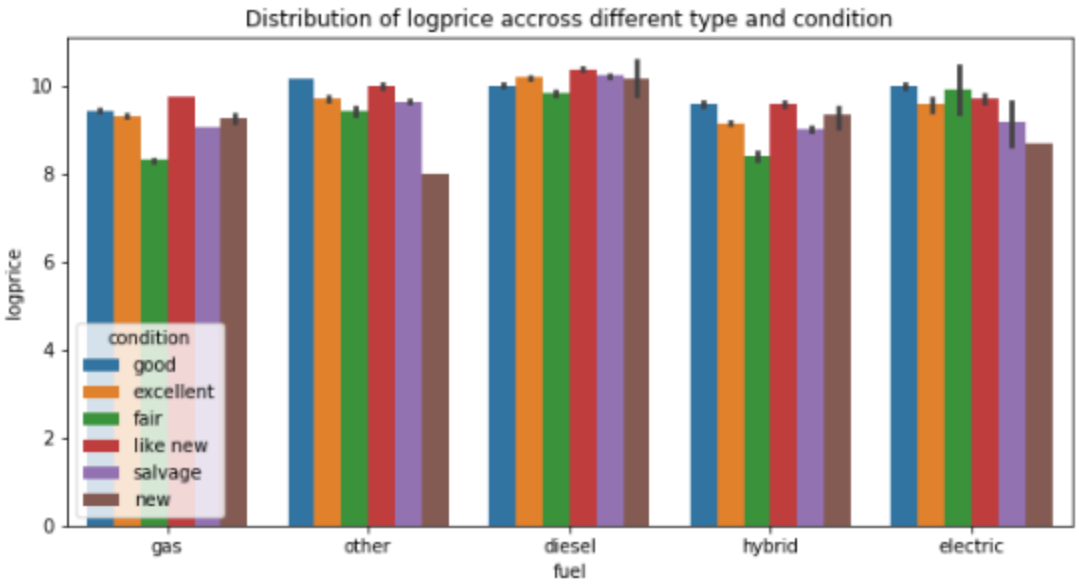
3. Distribution of price by fuel:



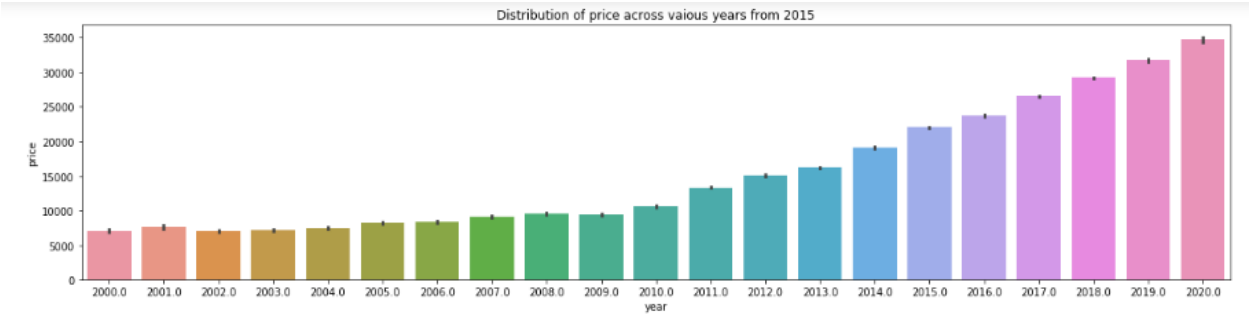


4. Distribution of price across different type and condition:

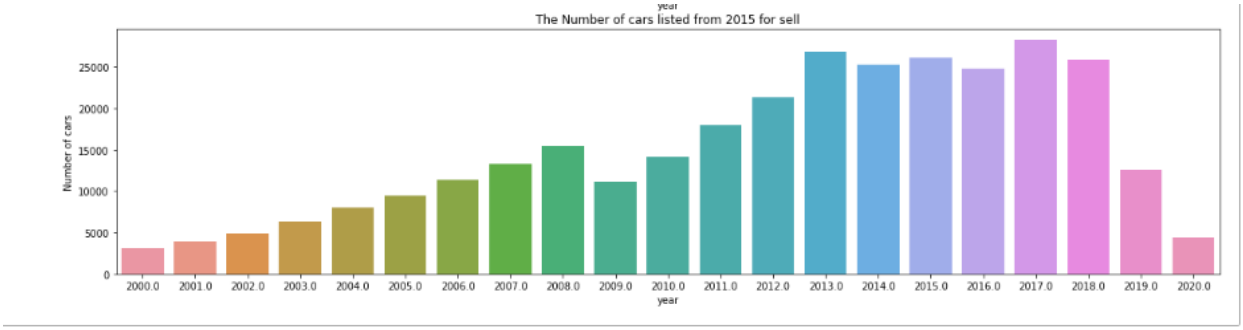




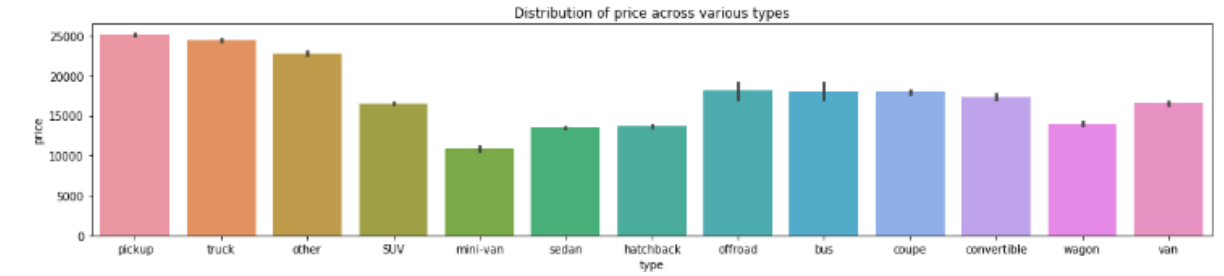
5. Distribution of price across various years from 2000

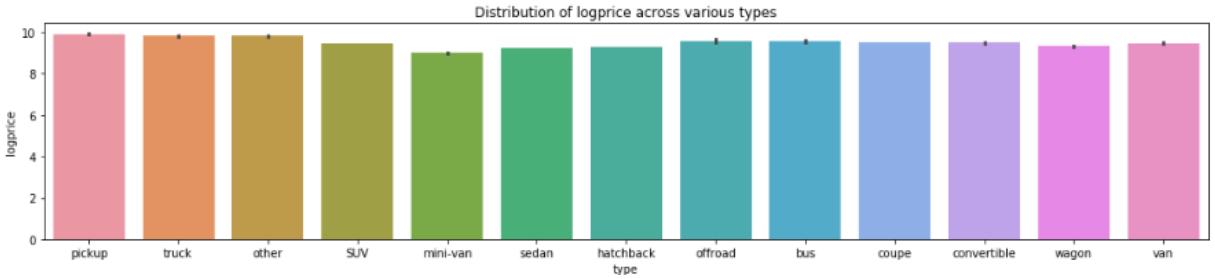


6. The number of cars listed for sell from 2000:

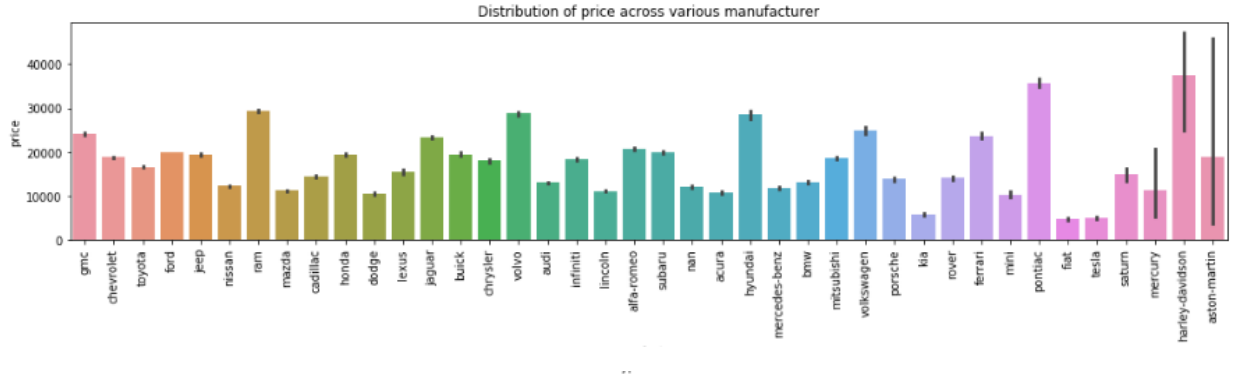


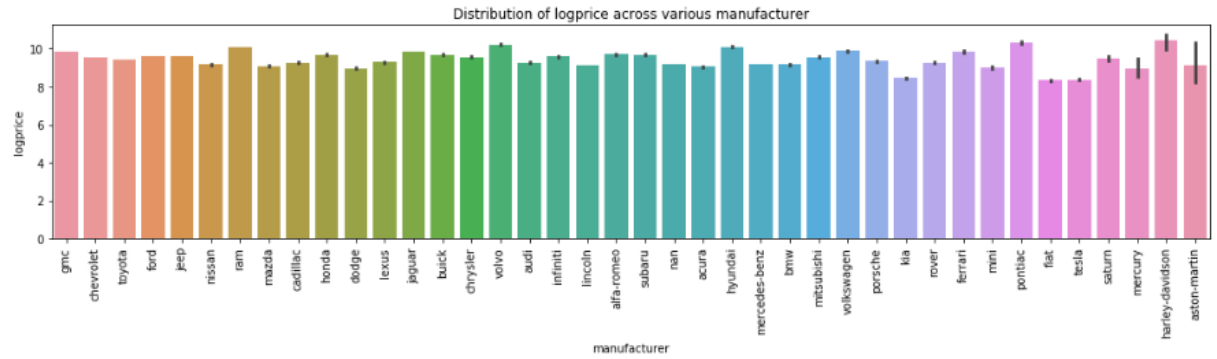
7. Distribution of the price across various types:



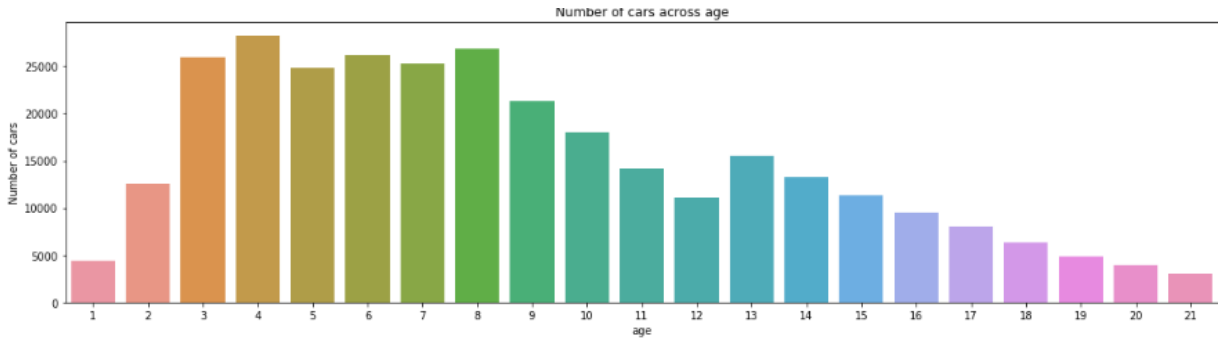


8. Distribution of price across various manufacturer:

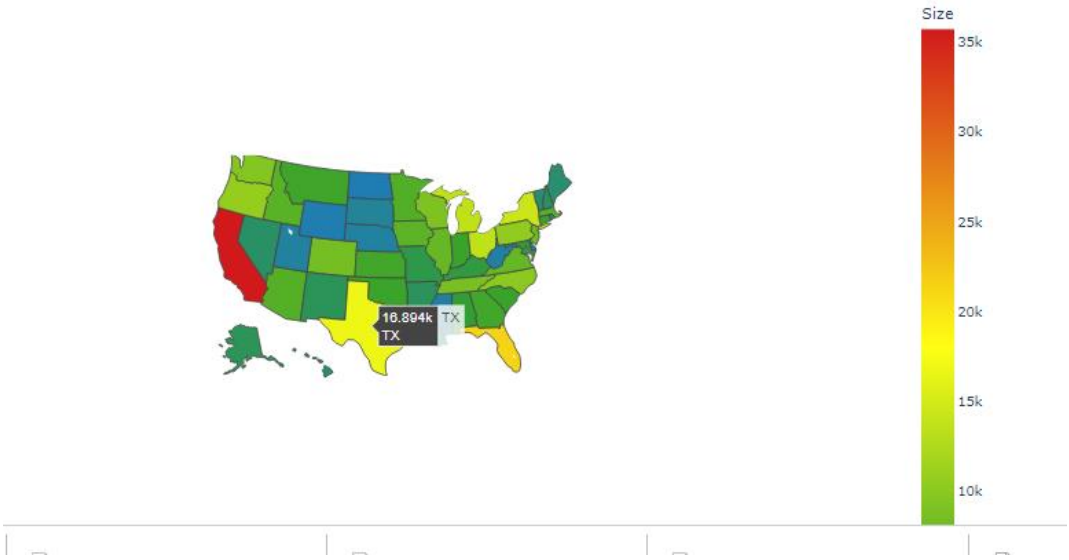




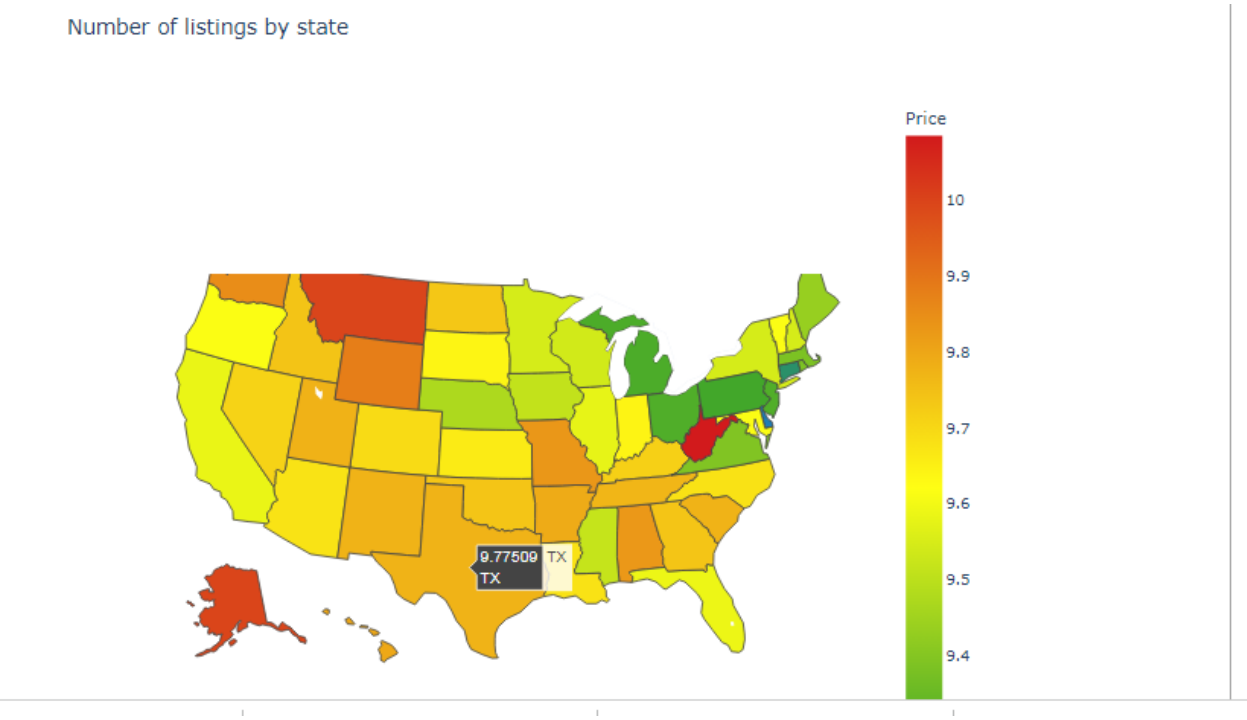
9. Number of cars across ages:



10. Median price of cars by state:



11. Number of listings of cars by state:



Data Modeling:

The below models were implemented in the analysis:

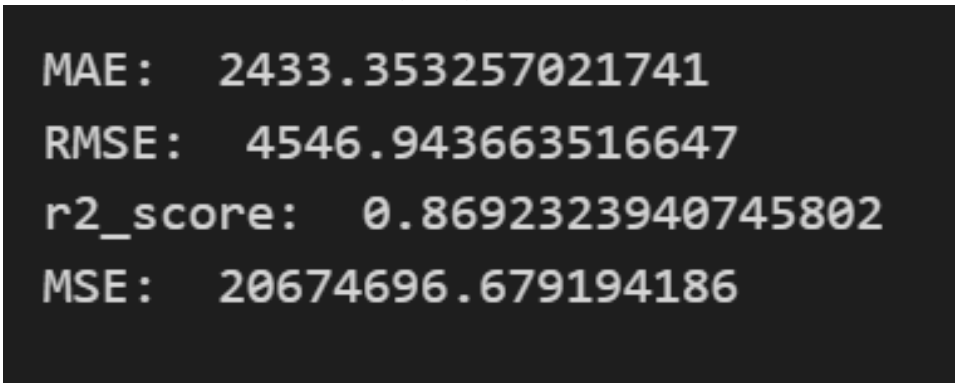
1. Linear Regression (Price)

```
MAE:  6552.971552764372
RMSE:  8802.493262912541
r2_score:  0.5081849047839497
MSE:  77483887.64362067
Coefficients:  [[ 5.51068539e+02 -2.52250748e+01  3.26878473e-02 -2.39219201e+02
   2.21599590e+03 -3.54511851e+03 -5.99981622e-02 -5.51068539e+02
  -1.33092103e+03  1.56927658e+03 -1.81727993e+03  1.14612642e+02
   5.41340127e+01 -5.21238050e+00]]
Intercept:  [-1082607.1773502]
```

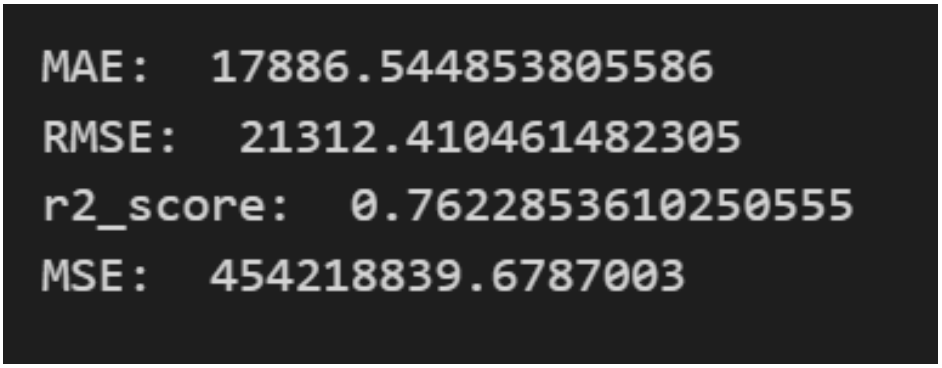
Logprice:

```
MAE:  0.41326842315369927
RMSE:  0.6068660891374787
r2_score:  0.49170277887026803
MSE:  0.3682864501450183
Coefficients:  [[ 3.98378258e-02 -1.30271295e-03  2.76536522e-06 -1.28197851e-02
   1.33571041e-01 -1.73517967e-01 -3.81925955e-06 -3.98378258e-02
  -8.89263978e-02  9.72257620e-02 -8.34746177e-02  3.42833137e-03
   4.21636027e-03 -1.22153364e-03]]
Intercept:  [-70.16383301]
```

2. RandomForest Generator Model (Price):



Logprice:



Possible Insights:

- As per our findings, Random forest model fits data and measures 86 % variance in the car price that is predictable from the car features
- In our analysis, year, odometer, age, model and fuel are the most impactful features.
- It is a challenging task as large number of features need to be considered and collected. In future we would like to collect more accurate data and use advanced machine learning algorithms for better prediction.