Group03: Car Price Prediction

Programming For Data Science
Lecturer: CHAN Sophal
July 15 2023

3rd year Engineer's Degree in Data Science
Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

Group members:

SEAN VENGNGY	ID: e20201133
SABORN MENGHORNG	ID: e20200983
VEN VANNUTH	ID: e20201651
SOK SOPHEAK	ID: e20200668
THOU CHANMAKARA	ID: e20200227
SENG VATHANAK	ID: e20200468

Contents

Abstract

In this report, we try to predict the cars price based on some features such as brand, model, year, body types. Etc. After scraping data from website, we did some data cleaning and preprocessing, conducting EDA, building models such as Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, then we did compare those model ‘s performance. Finally, we will discuss about the problems we have so far and future study about new solutions and ideas.

1 Introduction

The automobile market in Cambodia has been growing rapidly in recent years, with an increasing number of car buyers seeking to purchase new and used cars. As a result, there has been a growing interest in determining the factors that influence car prices in Cambodia. In this study, we aim to explore the relationship between various car features and their impact on car prices in the Cambodian market. To achieve this goal, we collected data on car prices and features from Khmer24, one of the largest online marketplaces for cars in Cambodia. We scraped the data and compiled a dataset containing information on various car features such as brand, model, year, and transmission type, as well as their corresponding prices. In this report, we will be predicting the prices of used and new cars. We will be building various Machine Learning models with different architectures. In the end, we will see how machine learning models perform in comparison to each other. Lastly, we will discuss some various problem during performing regression analysis, such as outlier, model assumptions, then we will t

2 Metholoogy

2.1 Data set

Our data set get from **Khmer-24 website** by doing web scraping. We decided to choose **16 variable** such as Car Aid, Category, posted, Car Makes, Car Model, Year, Tax Type, Condition, Body type, Fuel, Transition, Color, Tittle, Price, and Link. Our target Variable is Price. We also get 17873 records.

- Ad ID: The unique identifier of cars listed on the website Khmer 24.
- Category: The type of cars indicated for sale on the website.

- Posted: The date when the car post was made on the website.
- Car Makes: The brand of cars.
- Car Model: The model of cars.
- Year: The year that the car was manufactured.
- Tax Type: Types of tax (Tax Paper or Plate Number).
- Condition: The condition of the car (New or Used).
- Body Type: The car body type (e.g., SUV, sports car).
- Fuel: The type of fuel used with the car.
- Transmission: The system that transfers power from the engine (manual or automatic).
- Fuel: The color of the car.
- Link: The link to the car information on the website.
- Title Description of Car: A description or title of the car listing.
- Price: The car price (our target variable).

  Target: Price

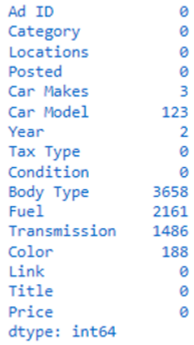
  Predictors: year (lifespan), brand, model, condition, body type, location, color, fuel, transmission, tax type.

We choose 4 Regression Model to predict price such as Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression.

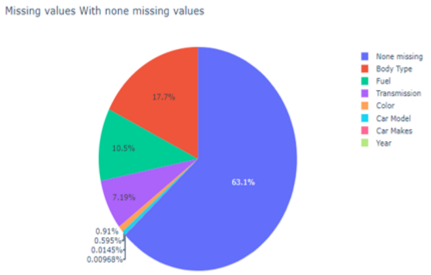
2.2 Data cleaning and preprocessing

1. Performs Exploratory Data Analysis (EDA) to understand the data and identify potential issues

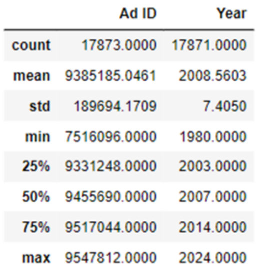
   From the above output we can say that there are 17873 rows/records and 16 columns Features present in our dataset. Then we check missing value.



```
Ad ID            0
Category         0
Locations        0
Posted           0
Car Makes        3
Car Model      123
Year             2
Tax Type         0
Condition        0
Body Type     3658
Fuel          2161
Transmission  1486
Color          188
Link             0
Title            0
Price            0
dtype: int64
```

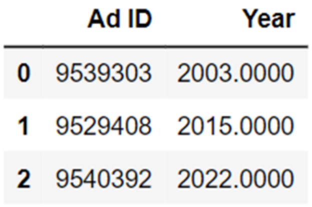


describe of our dataset.



	Ad ID	Year
count	17873.0000	17871.0000
mean	9385185.0461	2008.5603
std	189694.1709	7.4050
min	7516096.0000	1980.0000
25%	9331248.0000	2003.0000
50%	9455690.0000	2007.0000
75%	9517044.0000	2014.0000
max	9547812.0000	2024.0000

We have categorical Features 14 columns. Such as Category, location, posted, Car Makes, Car Model , Tax Type, Condition, Body Type, Fuel, Transmission, Color, Link, Title, Price. Numerical Features.



	Ad ID	Year
0	9539303	2003.0000
1	9529408	2015.0000
2	9540392	2022.0000

2. Provides descriptive statistics Uses appropriate visualizations Identifies and addresses any anomalies or outliers

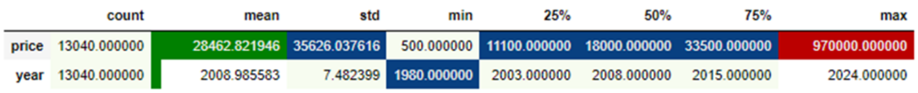
   Descriptive statistic table.

	year	price
count	13040.0000	13040.0000
mean	2008.9856	28462.8219
std	7.4824	35626.0376
min	1980.0000	500.0000
25%	2003.0000	11100.0000
50%	2008.0000	18000.0000
75%	2015.0000	33500.0000
max	2024.0000	970000.0000

Handling missing value technique can be removing or imputing. We choose the fastest is to remove missing value since they are not much compare to whole dataset. Handling outliers, there many techniques to deal with outliers such as removing or imputing it with mean, mode and median. In our project, we have try to keep outliers, However the result is not satisfied enough. The data seem to be non-linear and scatter too much. Therefore we also delete the outlier

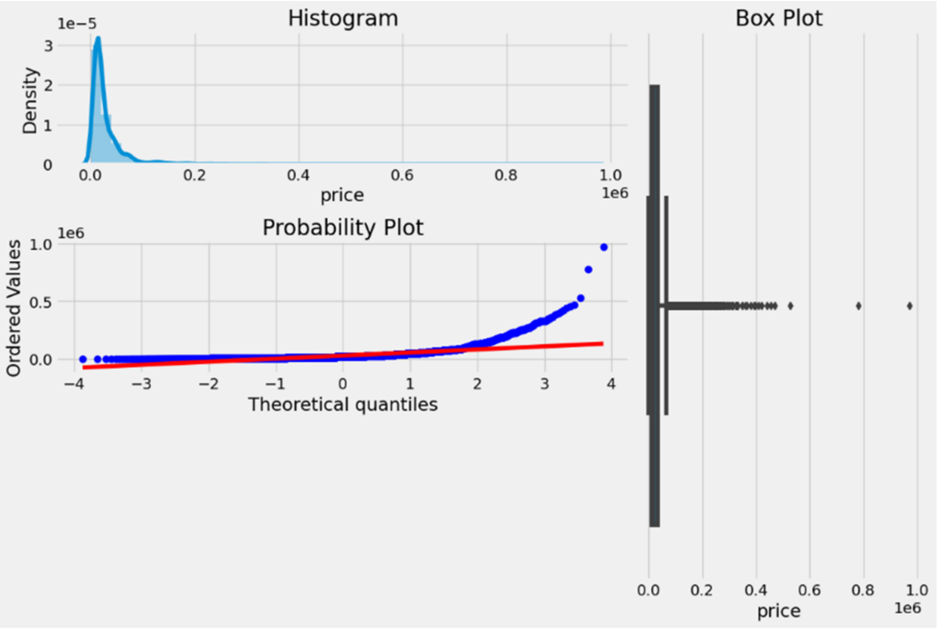
3. Handles missing data appropriately

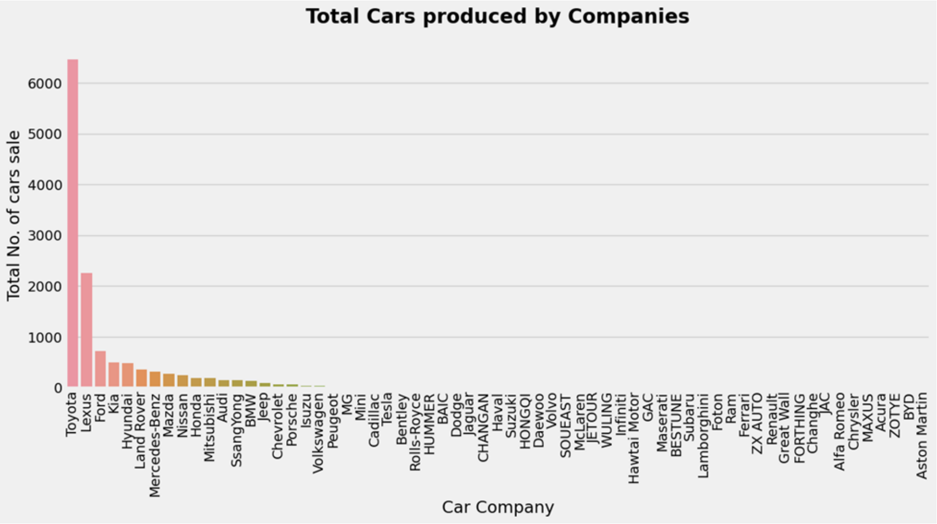
Overview of Data

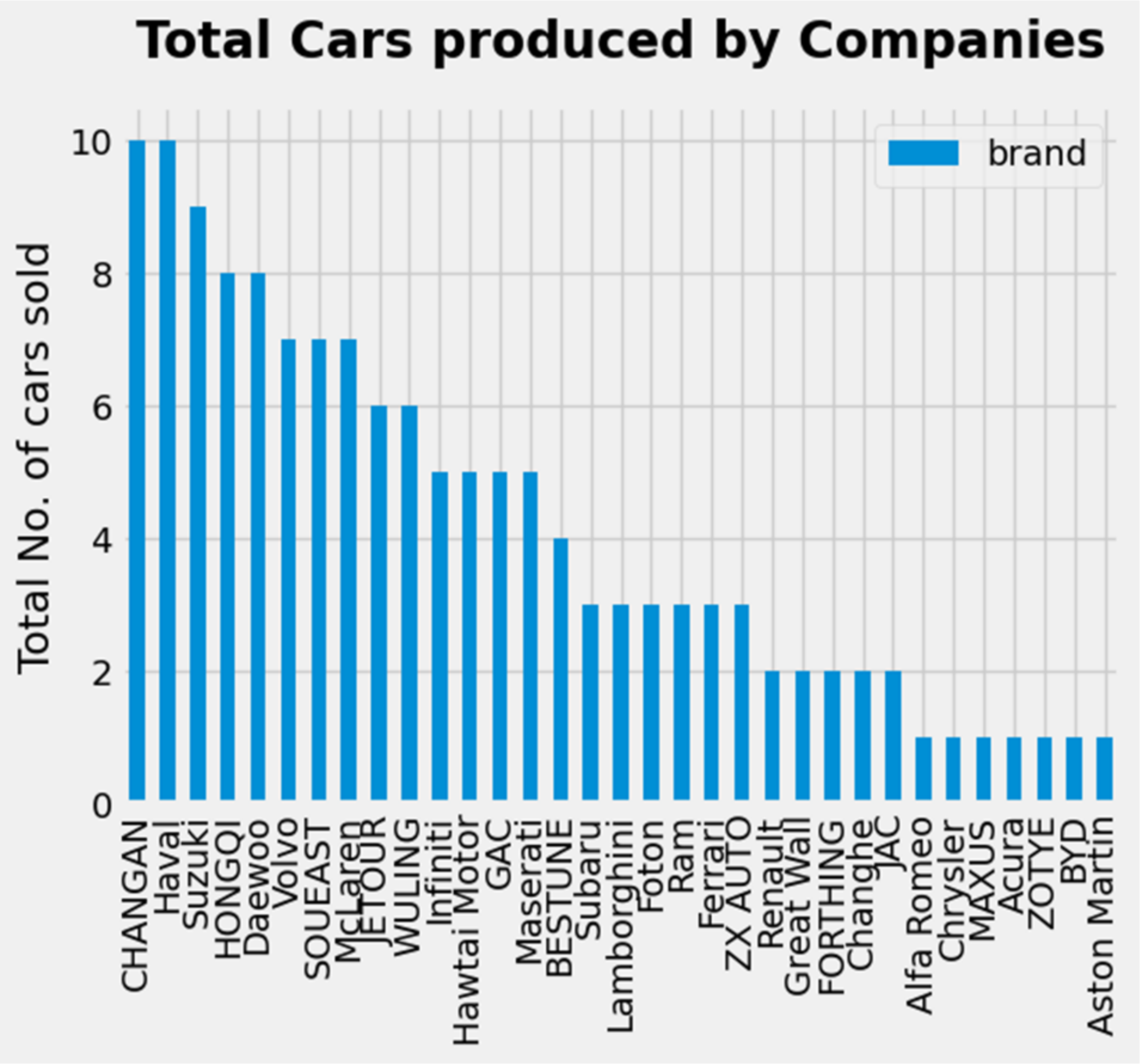


	count	mean	std	min	25%	50%	75%	max
price	13040.000000	28462.821946	35626.037616	500.000000	11100.000000	18000.000000	33500.000000	970000.000000
year	13040.000000	2008.985583	7.482399	1980.000000	2003.000000	2008.000000	2015.000000	2024.000000

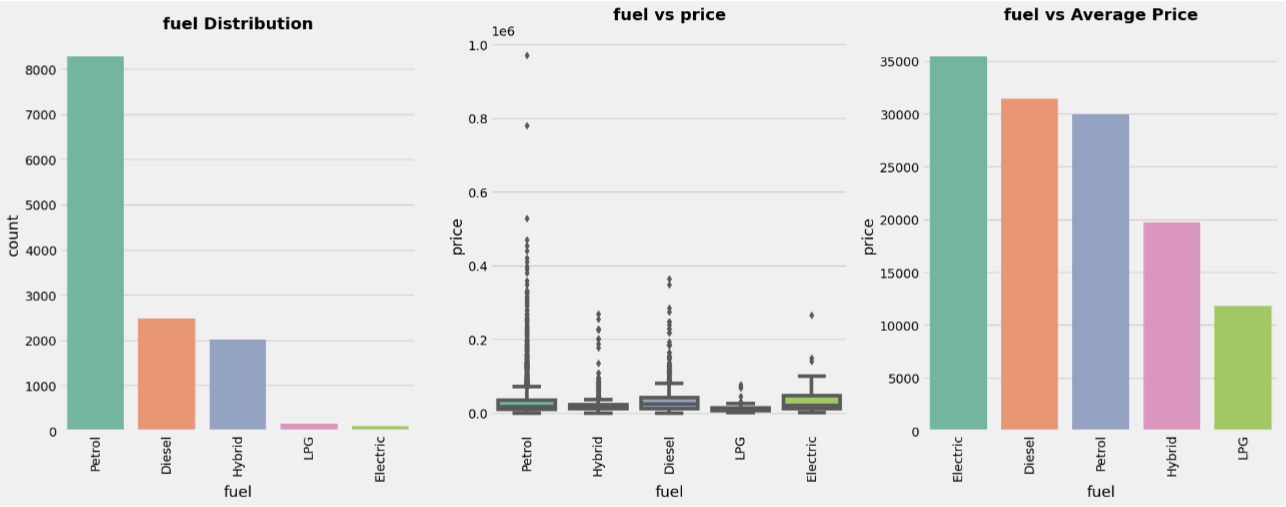
Target Distribution



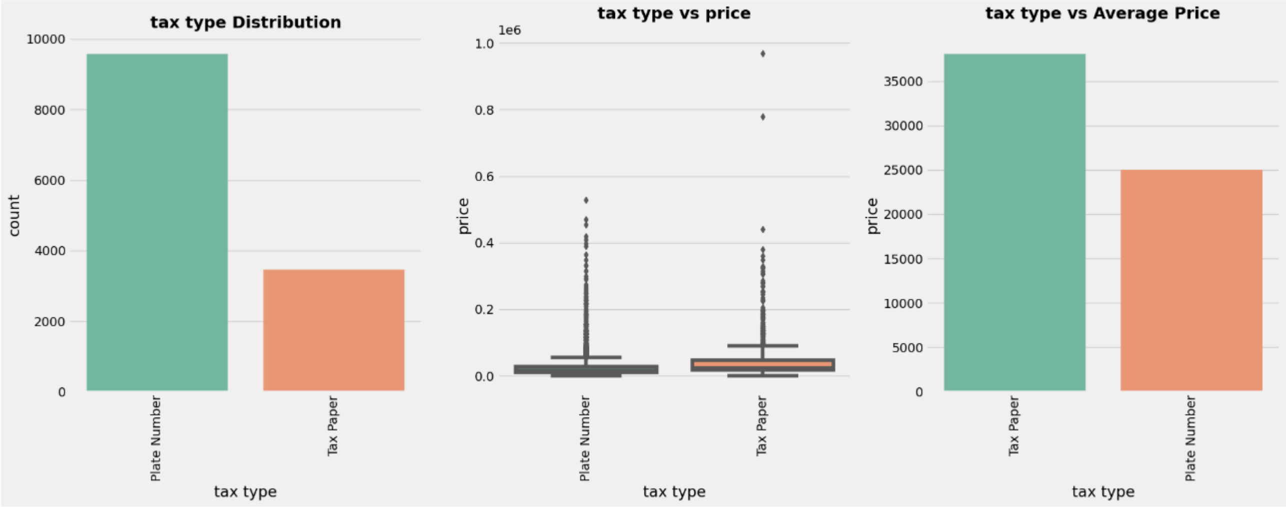

Total Cars produced by Companies
Total No. of cars sale
0
1000
2000
3000
4000
5000
6000
Car Company
Toyota
Lexus
Ford
Kia
Hyundai
Land Rover
Mercedes-Benz
Mazda
Nissan
Honda
Mitsubishi
Audi
SsangYong
BMW
Jeep
Chevrolet
Porsche
Isuzu
Volkswagen
Peugeot
MG
Mini
Cadillac
Tesla
Bentley
Rolls-Royce
HUMMER
BAIC
Dodge
Jaguar
CHANGAN
Haval
Suzuki
HONGQI
Daewoo
Volvo
SOUEAST
McLaren
JETOUR
WULING
Infiniti
Hawtai Motor
GAC
Maserati
BESTUNE
Subaru
Lamborghini
Foton
Ram
Ferrari
ZX AUTO
Renault
Great Wall
FORTHING
Changhe
JAC
Alfa Romeo
Chrysler
MAXUS
Acura
ZOTYE
BYD
Aston Martin



- *Toyota* company has sold the highest number of cars.
- So we can say that *Toyota* is kind of customers' most favored company.
- *MAXUS*, *Chrysler*, *Alfa Romeo*, *Acura*, *ZOTYE*, *BYD*, and *Aston Martin* are having very low data-points. So we can't make any inference of the least sales car companies.

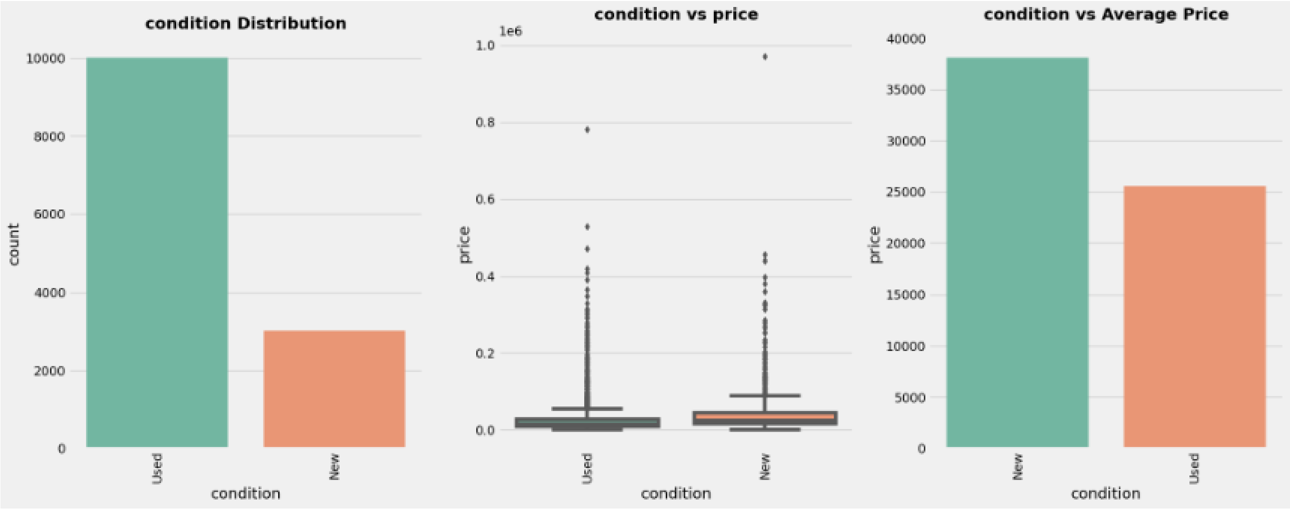


- We can clearly make an insight that Cars having petrol fuel systme is mostly sold.
- From the second plot we can make insight that Petrol Fuel System cars are availabe within every price range.
- From the third plot we can make an insight the **Average price of petrol fuel type cars are less than diesl fuel type

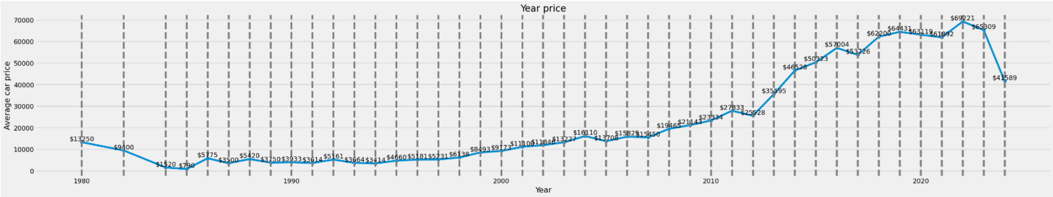


- Cars Plate Number are mostly sale when compared with cars Tax Paper.
- Cars Tax Paper are more expansive than Cars Plate Number

POutliers present in both



- The website has **cars used** more than **cars new**.
- **Cars new** have a higher average price than **cars used**.



The average prices seem to have increase from year 2012 to 2022.

4. Applies appropriate encoding techniques for categorical variables

5. Feature Engineering

   for feature engineering, to include categorical variable our models, we need to encoded them. there are many encoding method for categorical variables such as label encoding, one-hot encoding, for our dataset we choose target encoding since the unique values of each categorical variables are so many. Moreover, we create new column such as brand average price by grouping those brand and compute their mean. we also label those brand average price into 3 categories which are budget, mid-range, luxury price.

6. Explains and applies any necessary transformations or scaling to the data

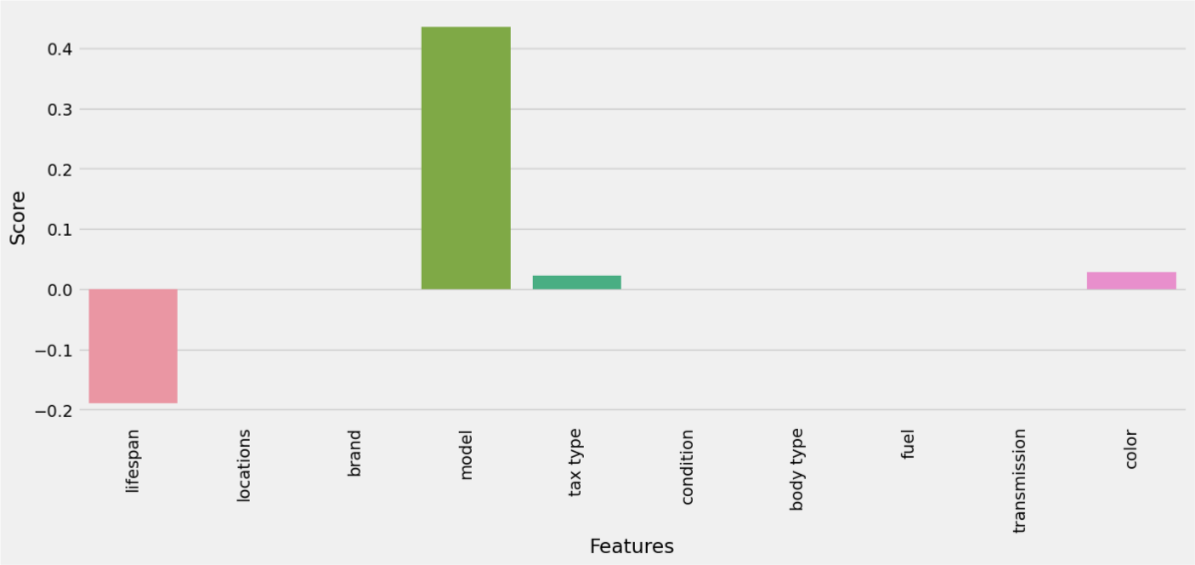
   for feature scaling we applied min-max scaling since our dataset full of encoded categorical data. so applying min max scaling would be suitable.

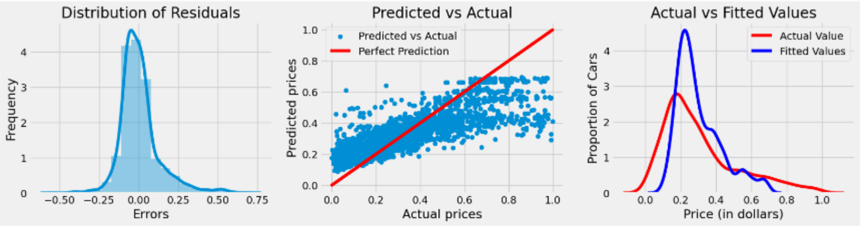
2.3 Regression Analysis

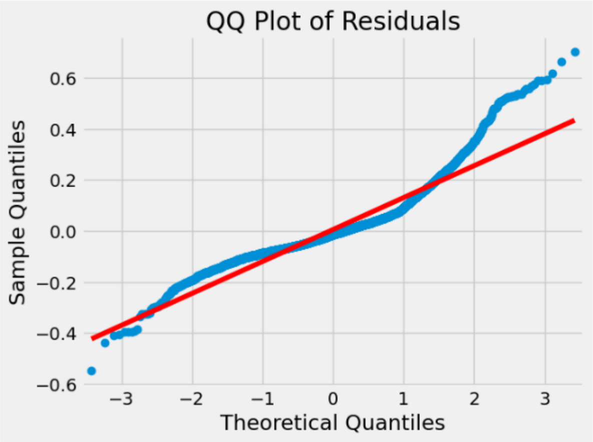
1. Build model

   we conducting 6 model for these dataset. those models are Multiple regression model, Ridge regression model, Lasso regression model, Polynomial regression, Ridge polynomial regression and Lasso polynomial regression. We had split our dataset for training and testing by 75

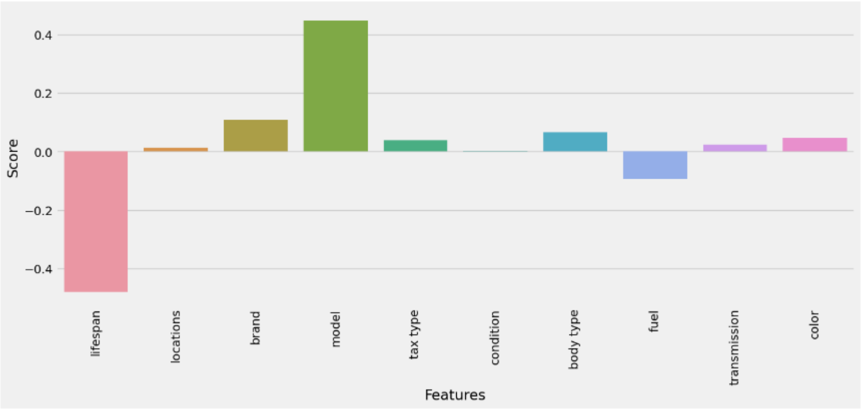
   - Linear Regression

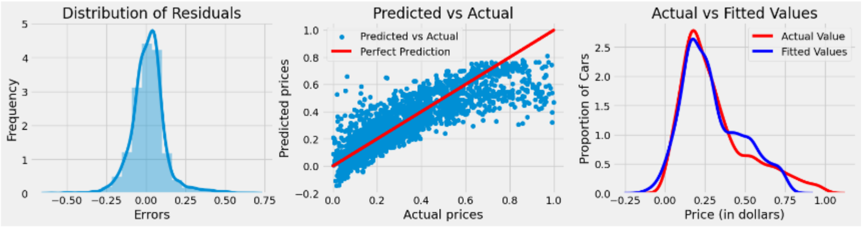


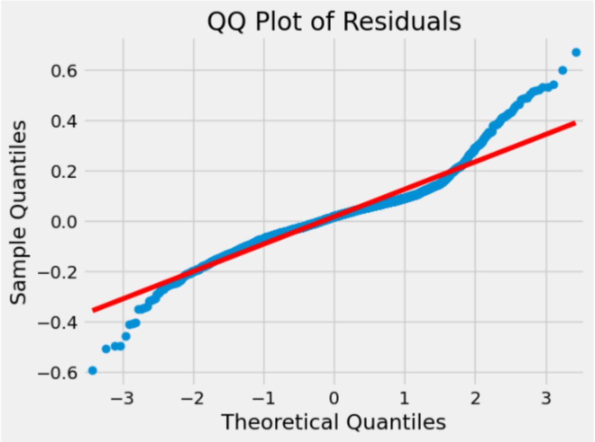




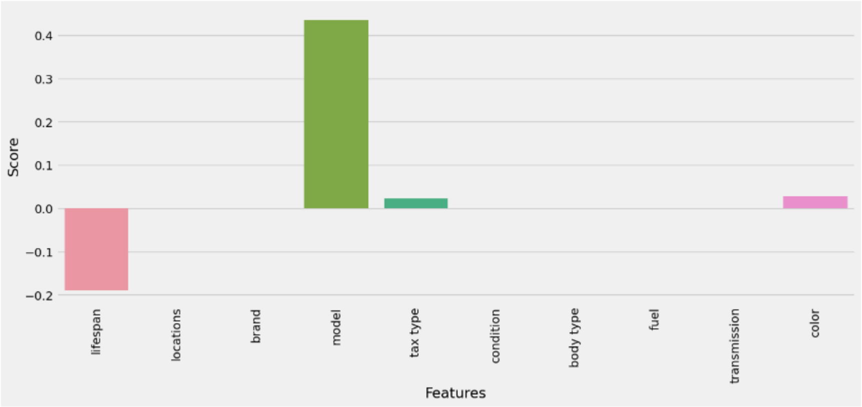
- Ridge Regression Model

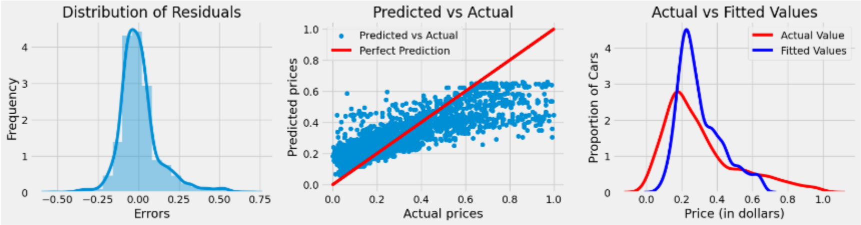


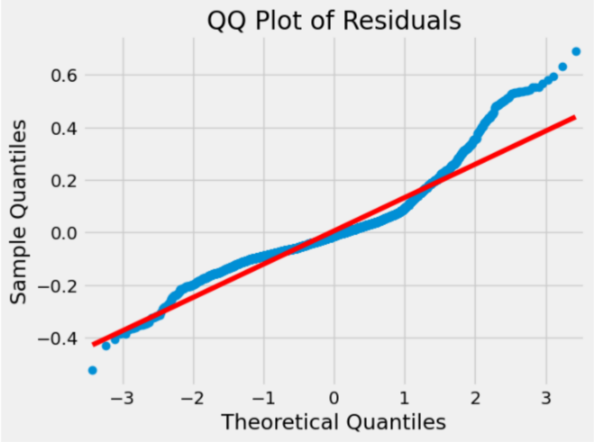




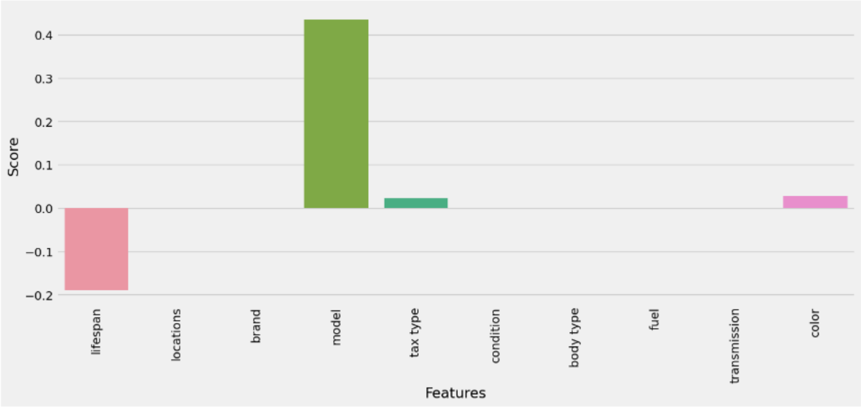
- Lasso Regressor Model

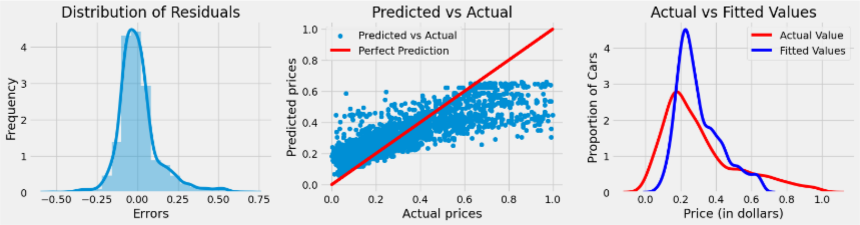


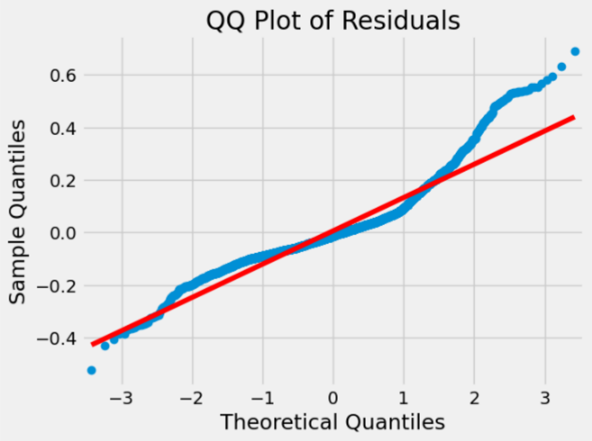




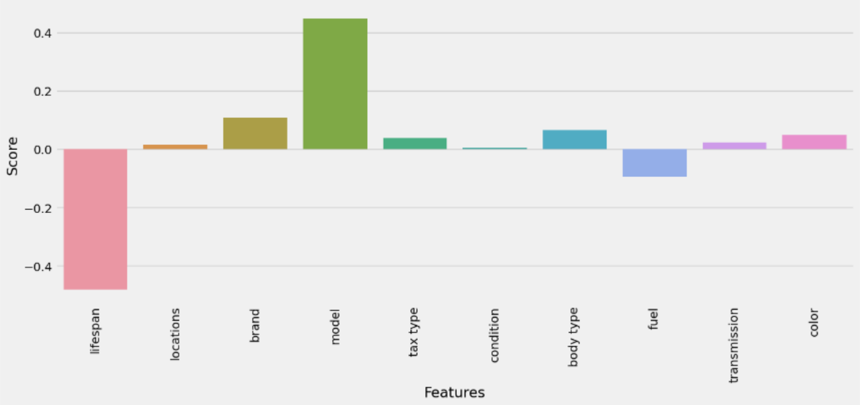
- Polynormail Regression Model

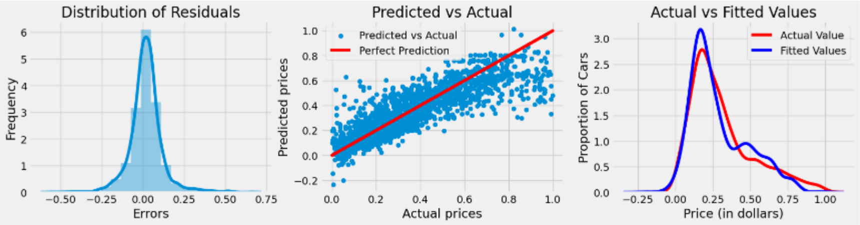


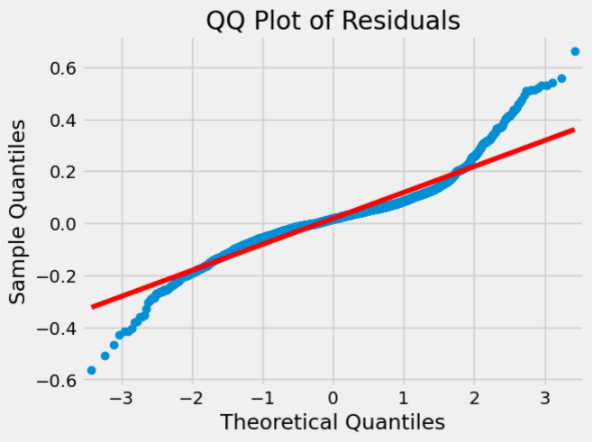




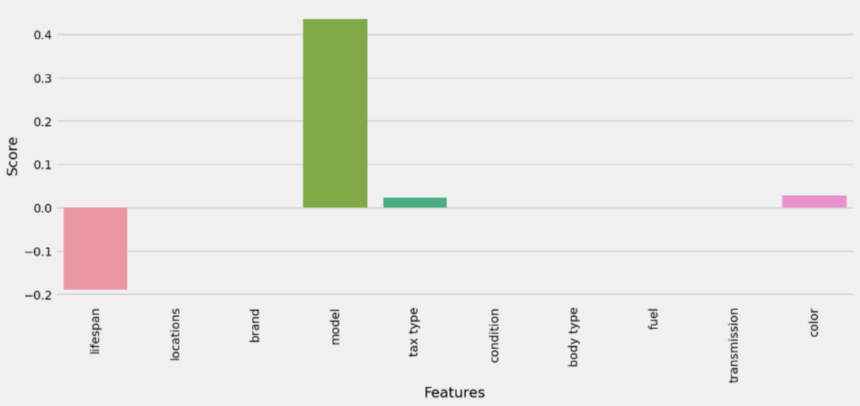
- Ridge Polynormail

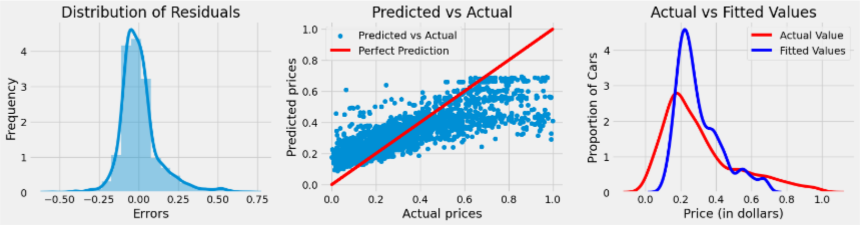


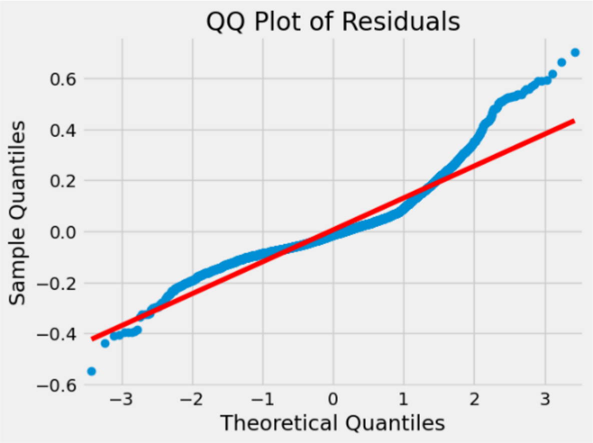




- Lasso Polynormail



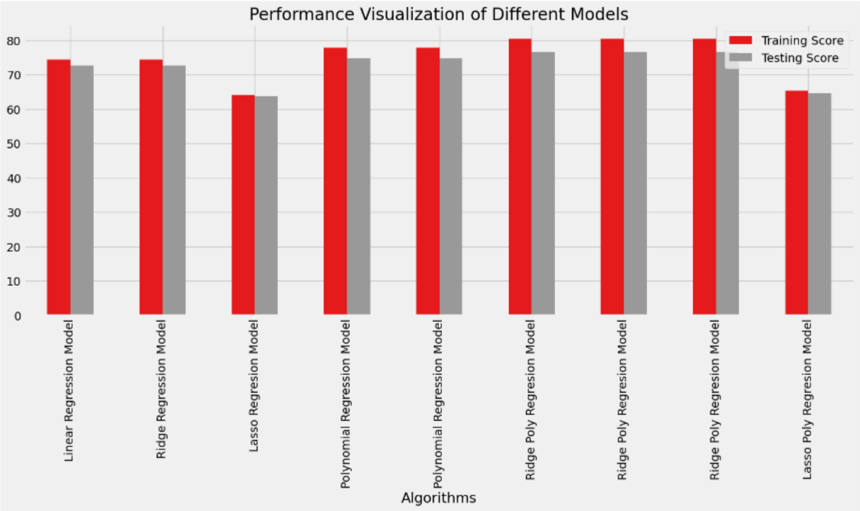




	Model	R^2 Train	R^2 Test	MSE Test
0	Linear Reg	74.40	72.60	0.012
1	Ridge	74.40	72.60	0.012
2	Lasso	64.10	63.70	0.016
3	Polynomial	79.10	71.09	0.012
4	Ridge Poly	80.45	76.70	0.010
5	Lasso Poly	65.27	64.61	0.015

3 Result

3.1 Model Comparison



Highest performance was given by the Ridge Polynomial Model, achieving around 80.The models are good enough to predict the car prices, explaining the variance of data up to 80, and the model is significant.

4 Discussion

1. Problem and Future Study:

   Problem and ideas for future study Since most assumptions are almost violated, so this issue may happened because of the dataset is not large enough, therefore we should study more about missing value and outlier instead of dropping them. Moreover we need to keep collecting up to date data too. If not we will explore and learn more about other method such effective sampling. Lastly, we want to try applying unsupervised learning like clustering and regression for our car dataset.

5 Conclusion

- First, we did the Basic Understanding of Data.
- Then we performed Data Cleaning to make the raw data more usable for analysis.
- Then we performed Exploratory Data Analysis to generate insights from the data.

- Then we performed Data Preprocessing to make the data suitable for model training and testing.
- Then we trained our model using different Machine Learning Algorithms.
- In the end, we achieved 80% accuracy with the Ridge Polynomial Regression model. So we can use this model for predicting the price of a car in the future.

Appendix - Python code

1. Sklearn Library For Linear Regression
2. Train Test Split in Sklearn Library
3. Feature selection using REF
4. P-values Feature Selection