Institute of Tecnology of Cambodia
Department of Applied Mathematics and Statistics

Car Price Prediction

Programing for Data Science

LECTURER: MR. CHAN SOPHAL

Lecture

Sophal CHAN

Team member

SOBON Menghorng

e20200983

SEAN VENGNGY

e20201133

SOK Sopheak

e20200668

VEN Vannuth

e20201651

THOU Chanmakara

e20200227

SENG Vathanak

e20200463

CONTENTS

Car Price Prediction

Goals 01
INTRODUCTION

Goals 02
EDA

Goals 03
MODELING

Goals 03
CONCLUSION

car price prediction

Data loading
-Basic Understanding
-data clearning

EDA
-overview data
-feature distribution
-correlation
-Featrure Engneering

Data preprocessing

Modeling
-linear Regression
-Ridge Regression
-Lassso Regression
-polynomail regression

1.introduction

Car Price Prediction:

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 make, model, year, and transmission type, as well as their corresponding prices.

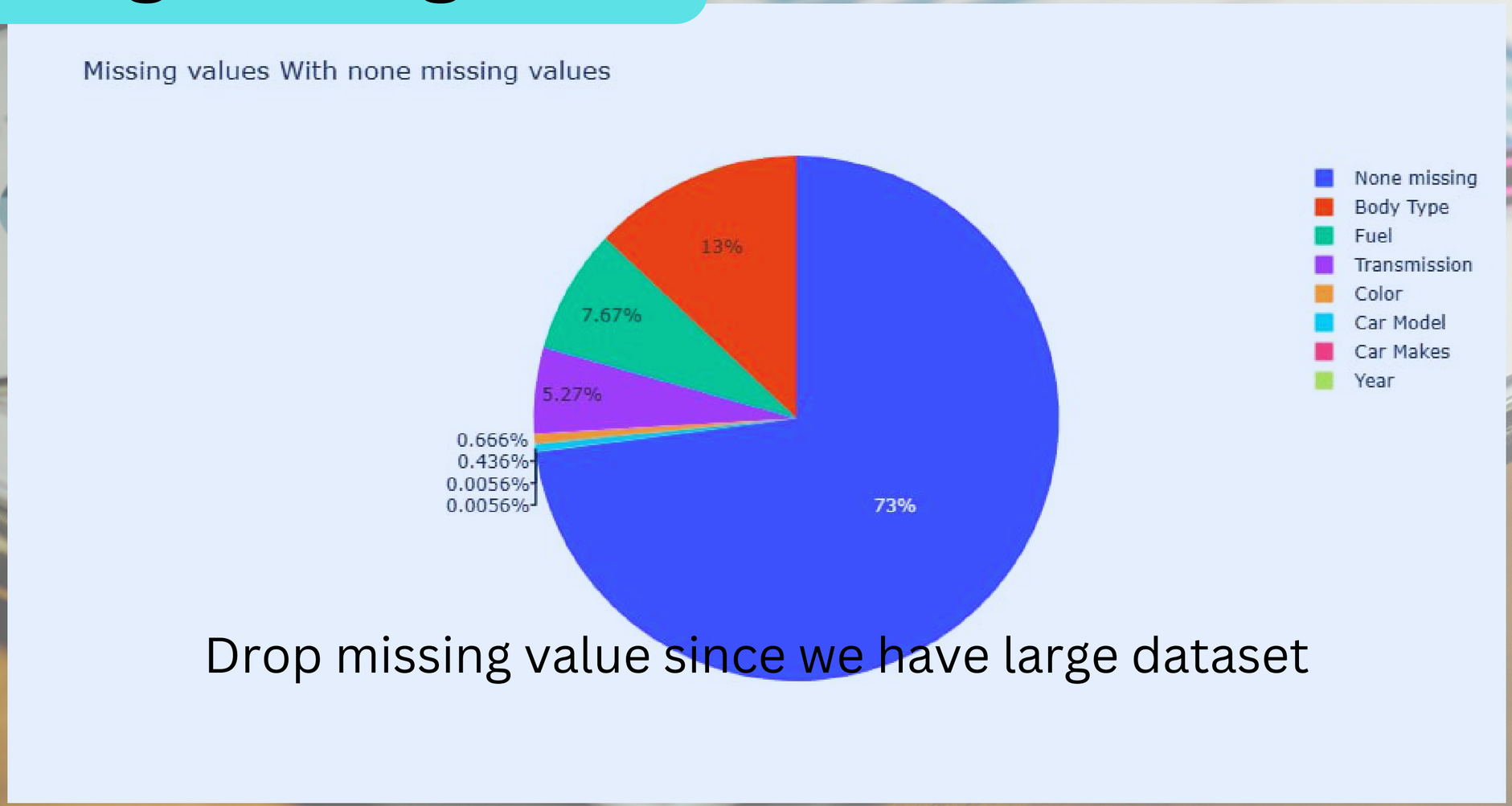
2. EDA

Variables descriptions:

- Ad ID: A unique identifier assigned to each car listed on the Khmer 24 website.
- Category: The type of car that is being advertised for sale on the website.
- Posted: The date when the car was posted for sale on the website.
- Car Makes: The brand or manufacturer of the car.
- Car Model: The specific model of the car.
- Year: The year that the car was manufactured.
- Tax Type: The type of tax associated with the car. There are two types of taxes.
- Condition: The condition of the car, which can be either new or used.
- Body Type: The type of car body, such as SUV, sports car, or other.
- Fuel: The type of fuel used by the car.
- Transmission: The type of transmission system that transfers power from the engine to the wheels.
- Color: The color of the car.
- Link: A link to the car information on the Khmer 24 website.
- Title: A description of the car.
- Price: The price of the car, which is the target variable in this context

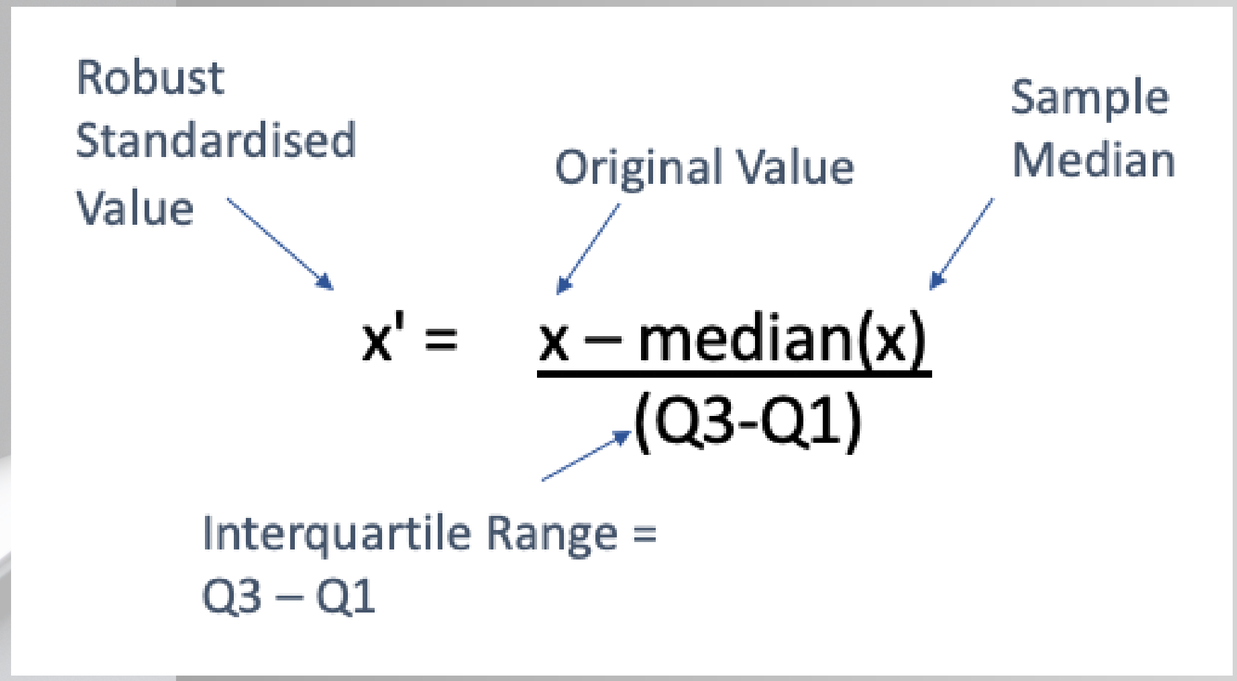
Our dataset was collecting from khmer24 website by using web scraping technique we obtain 16 columns and around 17 873 rows for oringinal dataset about car model price and other features such as year, color condition, brand, body types.etc.

Handling missing value



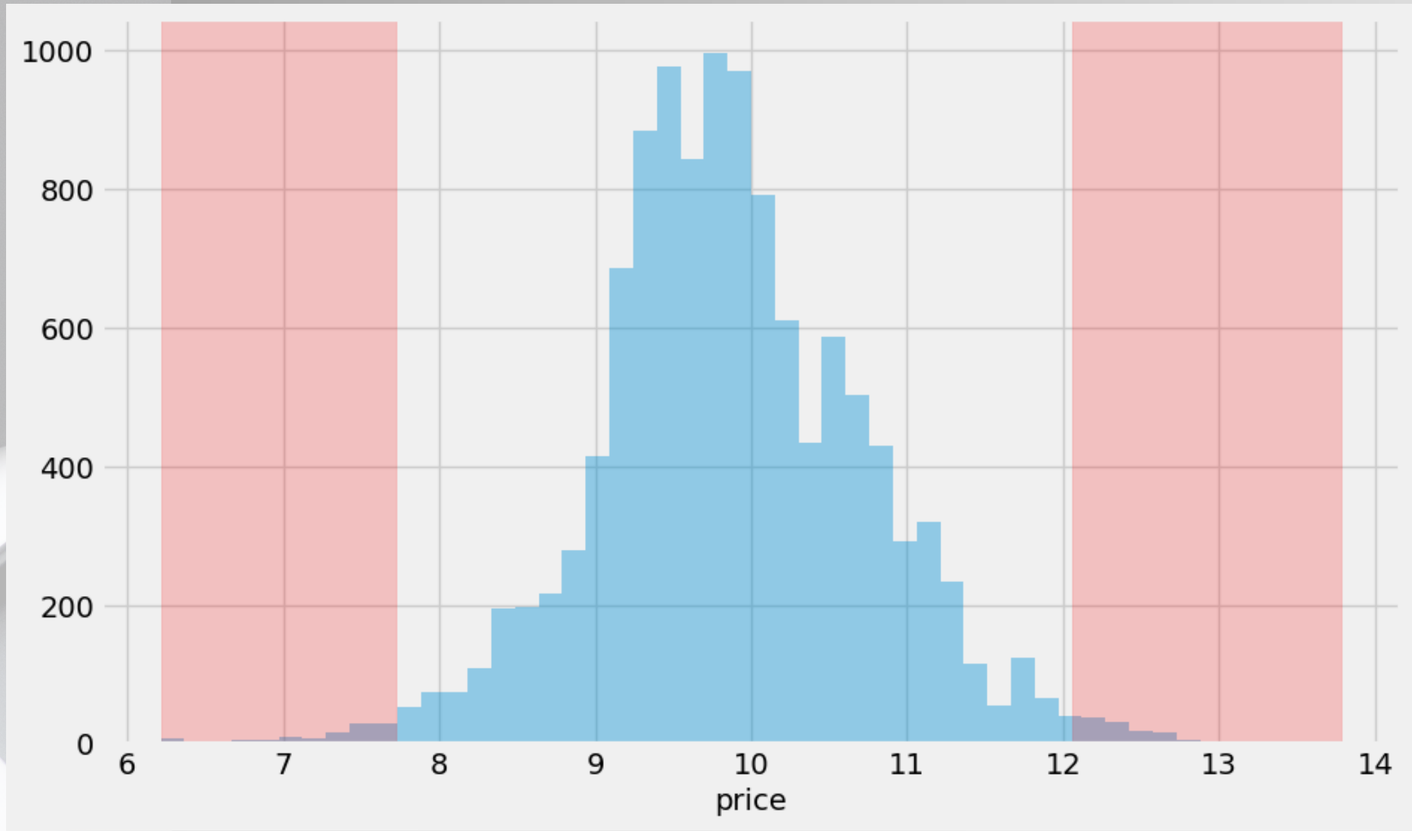
Drop missing value since we have large dataset

feature scalling



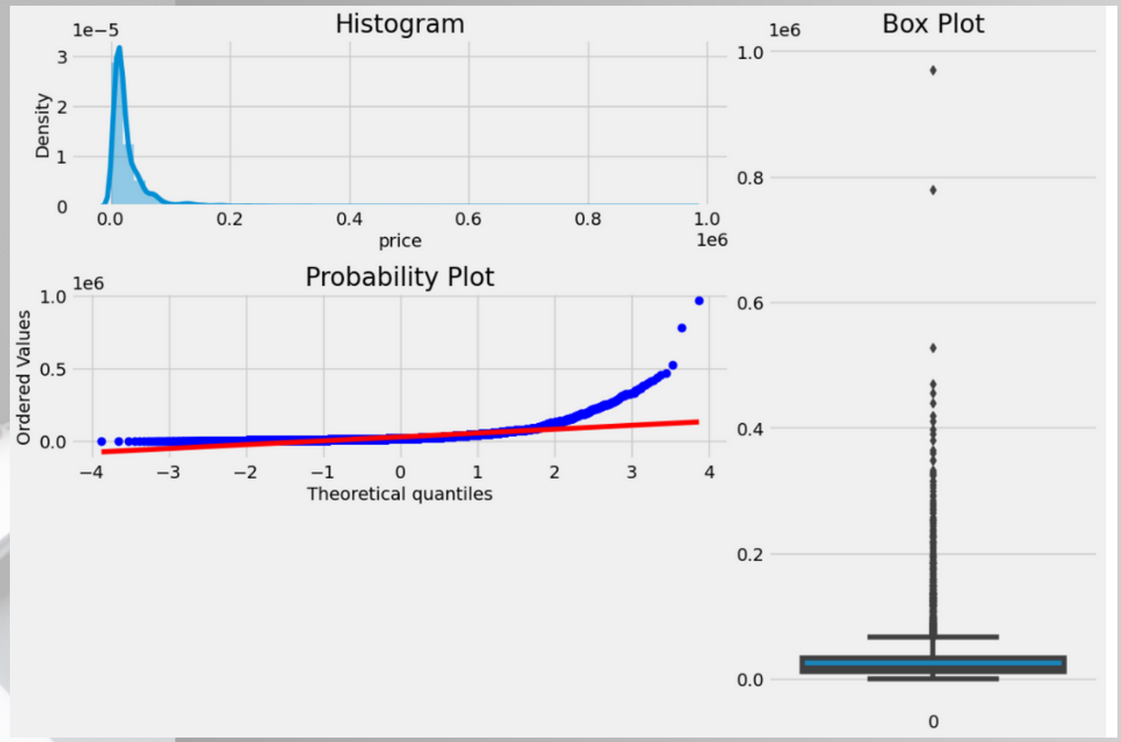
Applied the Robust scaling to all features.
RobustScaler is a technique that scales features to median and quantiles instead of mean and variance. It is used when the data has outliers, which can skew the mean and variance and make the scaling less effective.

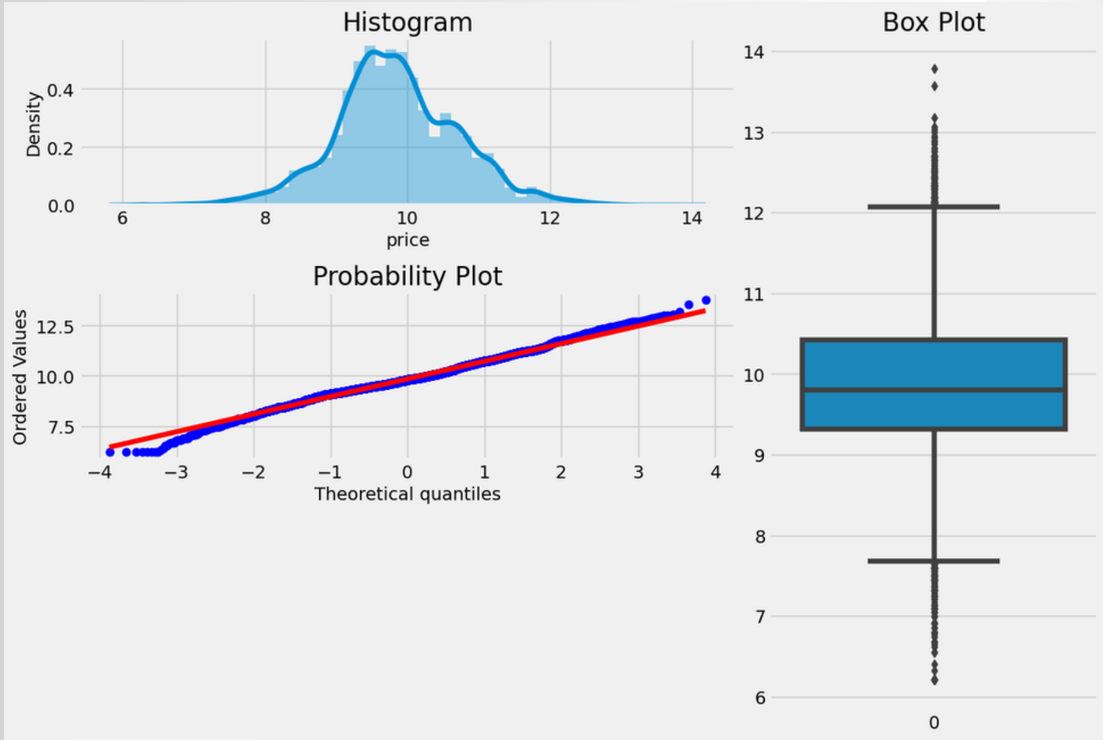
Outlier handling



outlier for price 170k above and 2.2k dropped

Log Transforming





OVERVIEW

	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

	count	unique	top	freq
model	13040	462	Prius	2096
brand	13040	63	Toyota	6462
color	13040	14	White	5524
body type	13040	9	SUV	5035
fuel	13040	5	Petrol	8280
tax type	13040	2	Plate Number	9577
condition	13040	2	Used	10017
transmission	13040	2	Auto	12426

label encoded variables

	1.0000	0.0000	-1.0000
0	Tax Paper	Plate Number	NaN
1	NaN	Used	New
2	Manual	Auto	NaN

Target encoded variables

	0.0000	-0.1049	0.2233	0.0491	-0.6178
0	Petrol	Hybrid	Diesel	Electric	LPG

	0.0808	-1.0000	0.0000	-0.7119	0.7331	0.4294	-0.8498	-0.1008
0	Pickup	Sedan	SUV	Hatchback	Sports	MPV (Minivan)	Station Wagon	Convertible

	-1.1173	0.0000	-1.0000	-0.3782	0.1759	-0.6616	-1.0221	-1.6240	-0.8842	-0.4109	-0.3962	-0.4040	-0.7922
0	Blue	White	Silver	Gray	Black	Brown	Gold	Green	Red	Yellow	Beige	Orange	Purple

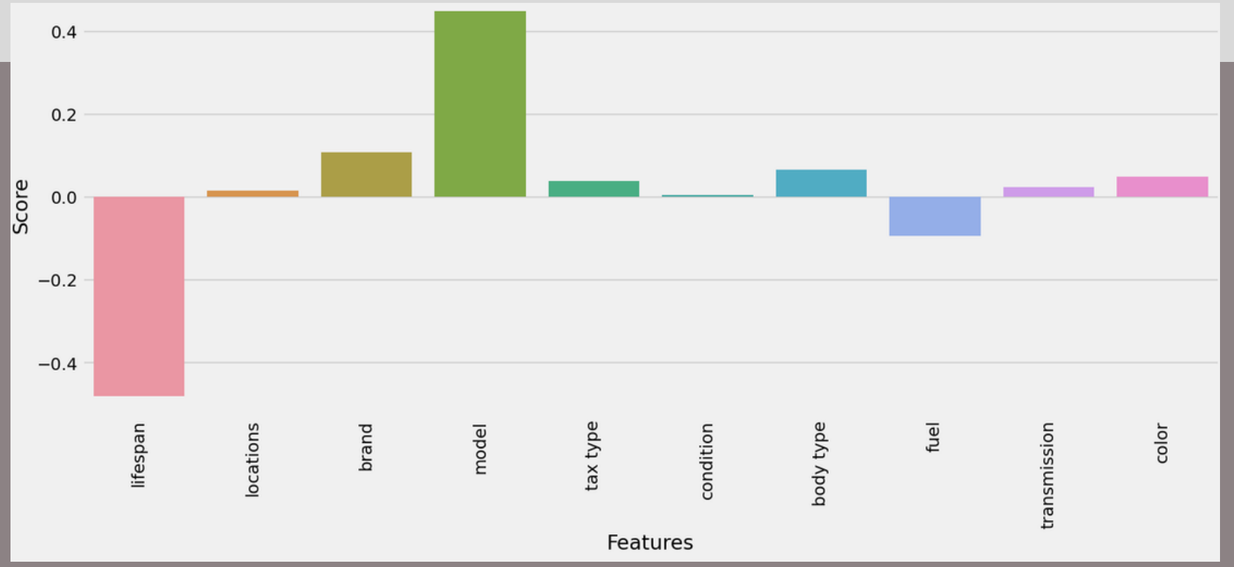
	0.0000	-0.5998	-0.4551	-0.4775	-0.5380	-0.5260	-0.3914	-0.5190	-0.0665	-0.6392	...	-0.5568	-0.0567	-0.2904	-0.3258	-0.2314	-0.1938	-0.1041	-0.2603	-0.1546	-0.2542
0	Phnom Penh	Kampong Cham	Prey Veng	Battambang	Kandal	Siem Reap	Preah Sihanouk	Banteay Meanchey	Ratanakiri	Takeo	...	Kampong Thom	Mondulkiri	Stung Treng	Kampong Chhnang	Svay Rieng	Koh Kong	Pailin	Kratie	Preah Vihear	Oddar Meanchey

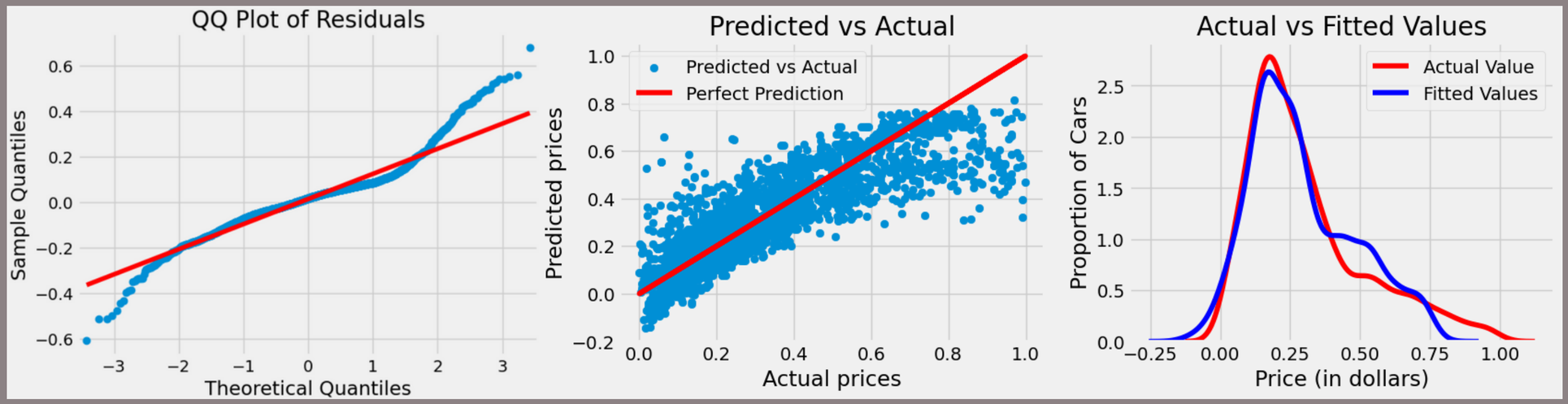
	0.6244	0.0000	1.0000	-1.1483	-0.5055	2.0289	0.8256	-0.0530	-1.7996	2.4927	...	0.5906	1.1231	0.9221	0.4820	0.5860	0.5809	0.5872	0.5423	0.2608	0.66
	HUMMER	Toyota	Lexus	Kia	Hyundai	Ford	MG	Nissan	Honda	Porsche	...	FORTHING	Ram	Acura	Subaru	SOUEAST	MAXUS	Renault	BYD	Great Wall	Chang

	0.5027	-0.1071	-0.5458	1.9428	-0.3895	2.0823	-1.4755	1.7629	0.2282	0.7980	...	0.4647	0.8108	0.5124	0.3814	0.1032	0.5295	0.5138	0.6225	0.6433	0.1031
0	H2 SUT	Prius	Camry	NX	RX300	RX350	Morning	Land Cruiser	Santa Fe	CT	...	Datsun/Nissan Z-car	Armada	408	Benteyga	Carrera GT	Poer-1	Galant	Celerio	Micra	GLB-Class

3. Model Building

Linear-Regression Model





Model Results

OLS Regression Results

Dep. Variable:	price	R-squared:	0.744
Model:	OLS	Adj. R-squared:	0.744
Method:	Least Squares	F-statistic:	2448.
Date:	Fri, 14 Jul 2023	Prob (F-statistic):	0.00
Time:	15:20:36	Log-Likelihood:	6135.7
No. Observations:	7577	AIC:	-1.225e+04
Df Residuals:	7567	BIC:	-1.218e+04
Df Model:	9		
Covariance Type:	nonrobust		

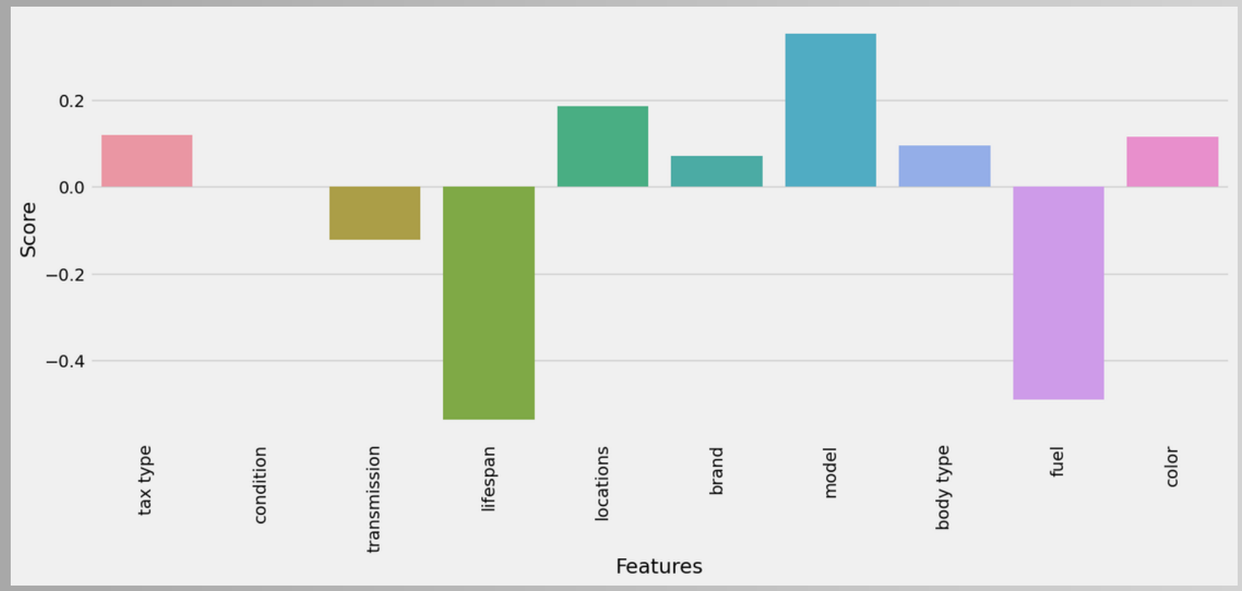
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2622	0.013	20.609	0.000	0.237	0.287
lifespan	-0.4827	0.010	-47.079	0.000	-0.503	-0.463
locations	0.0142	0.005	2.716	0.007	0.004	0.024
brand	0.1077	0.009	12.594	0.000	0.091	0.124
model	0.4480	0.008	58.923	0.000	0.433	0.463
tax type	0.0404	0.003	12.683	0.000	0.034	0.047
body type	0.0661	0.005	12.320	0.000	0.056	0.077
fuel	-0.0937	0.008	-11.903	0.000	-0.109	-0.078
transmission	0.0243	0.007	3.481	0.001	0.011	0.038
color	0.0478	0.004	12.570	0.000	0.040	0.055

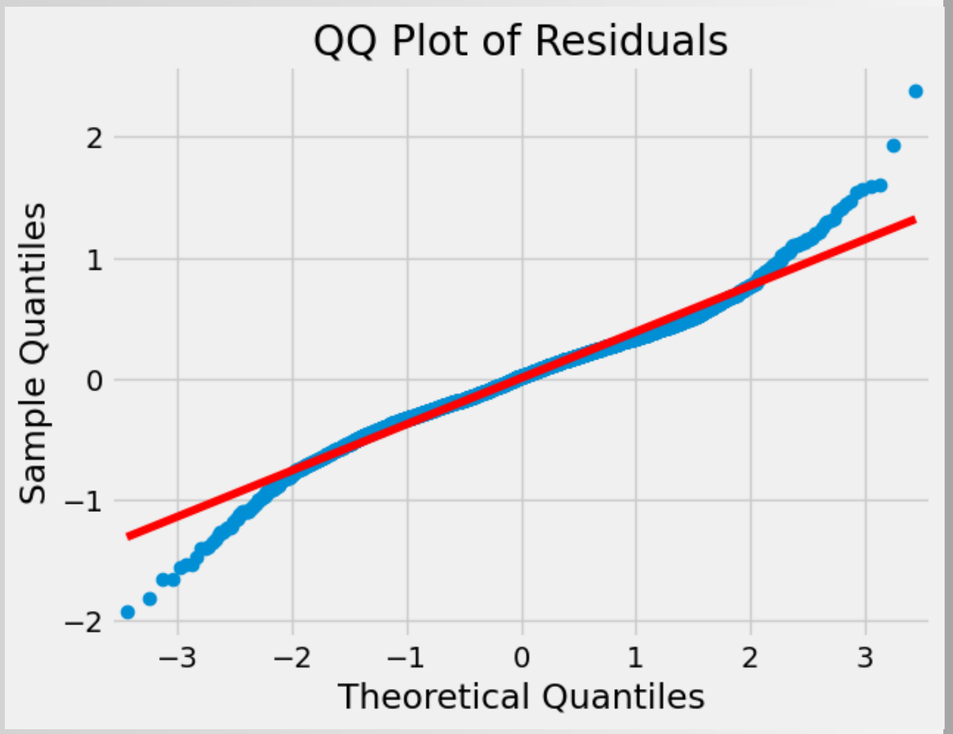
Omnibus:	992.561	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5549.506
Skew:	0.500	Prob(JB):	0.00
Kurtosis:	7.071	Cond. No.	26.6

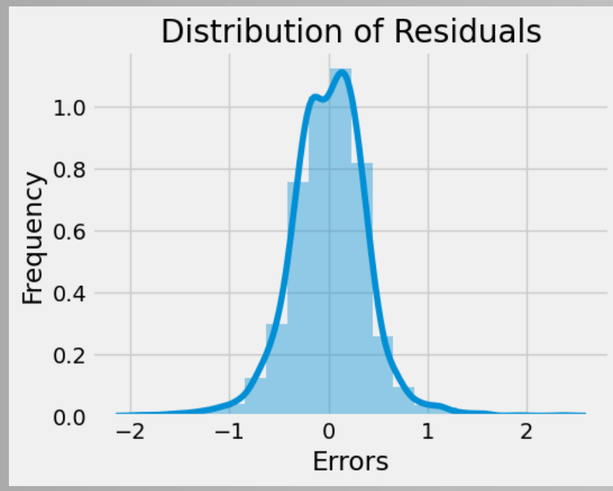
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie

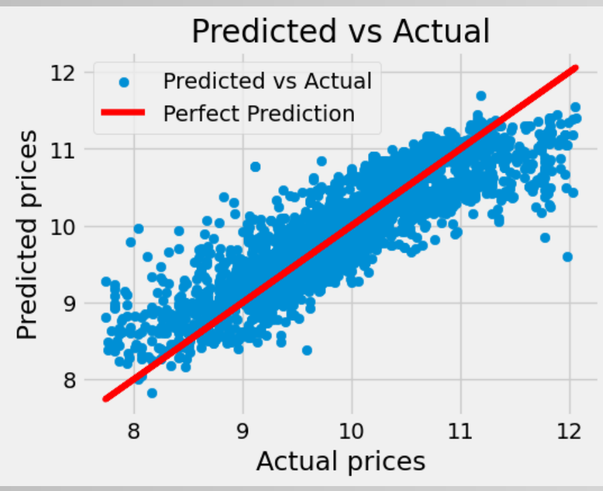
	Features	VIF
0	model	2.2200
1	lifespan	1.9400
2	brand	1.6700
3	tax type	1.4200
4	body type	1.4100
5	color	1.3400
6	fuel	1.2400
7	locations	1.1500
8	transmission	1.1100

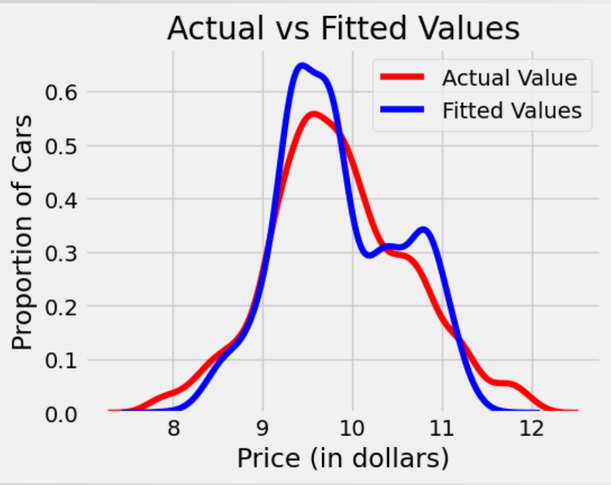
Ridge Regressor Model





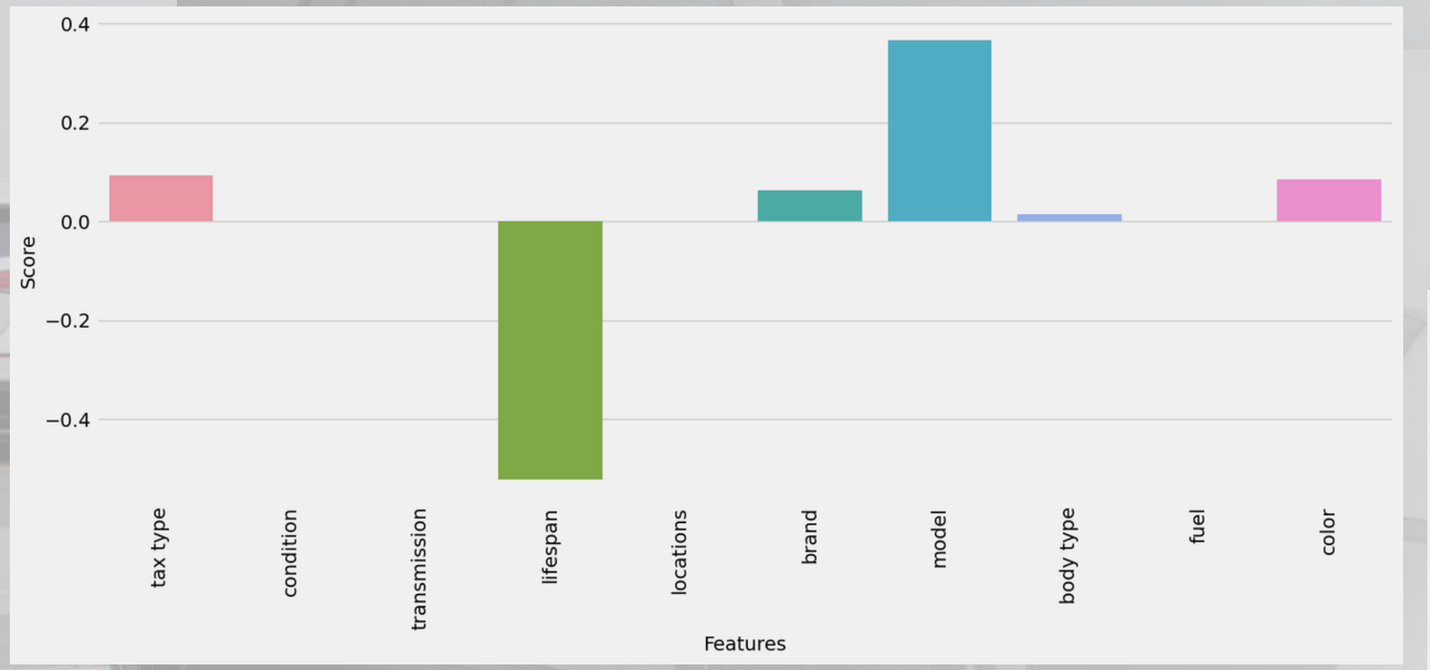


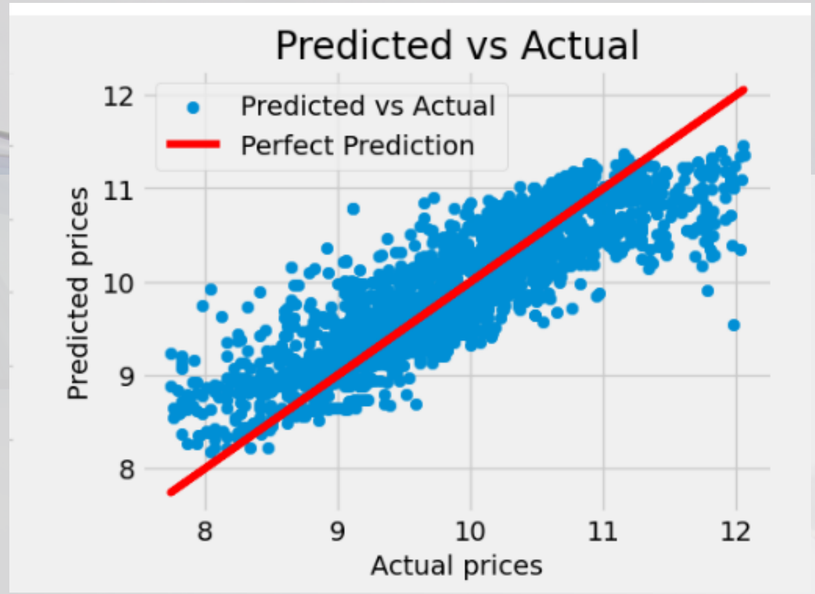


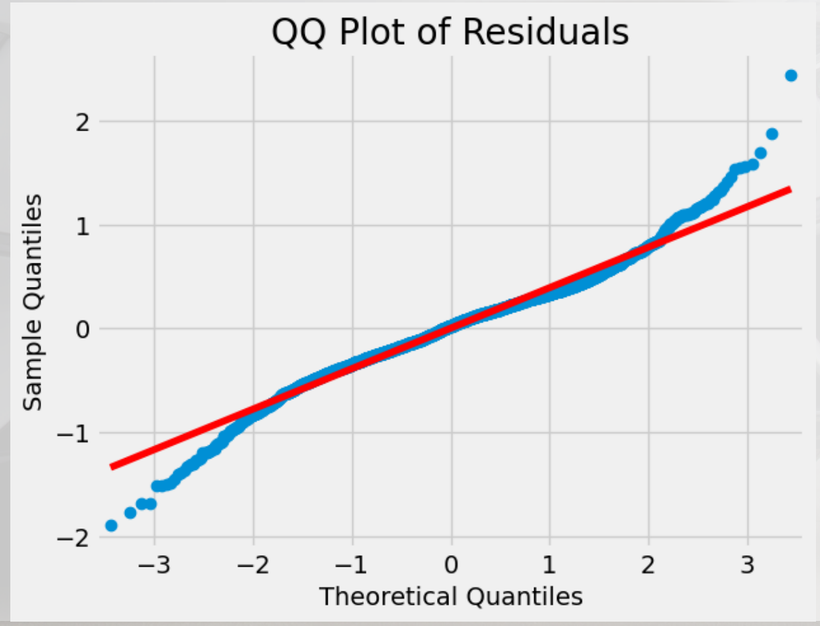


3. Model Building

Lasso Regression

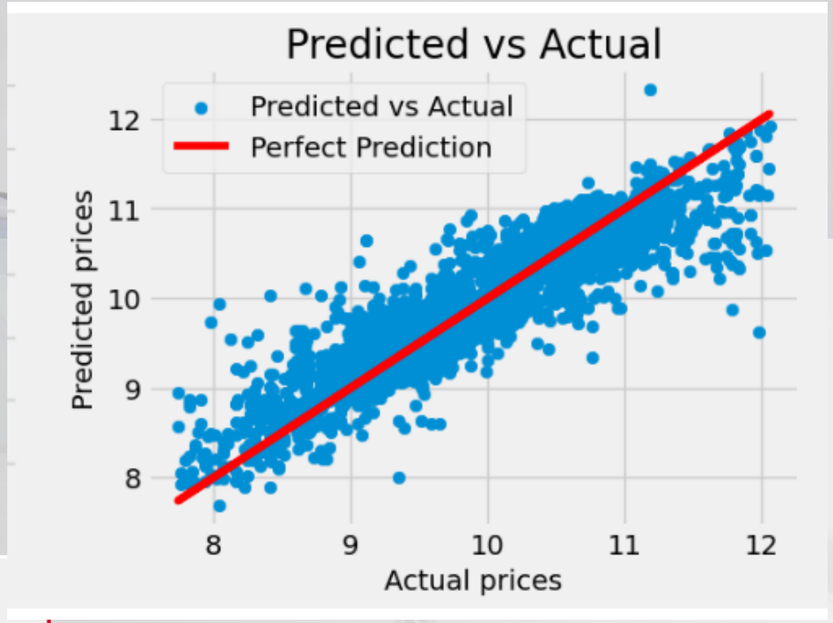


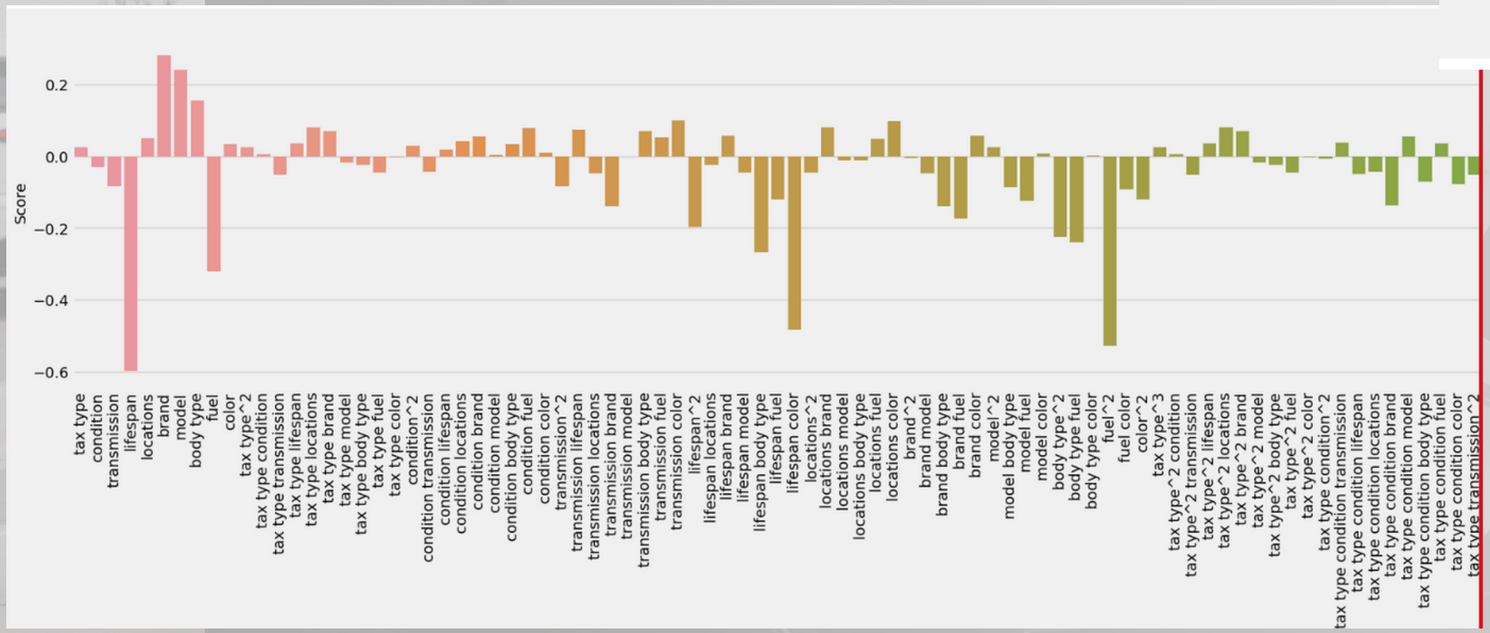


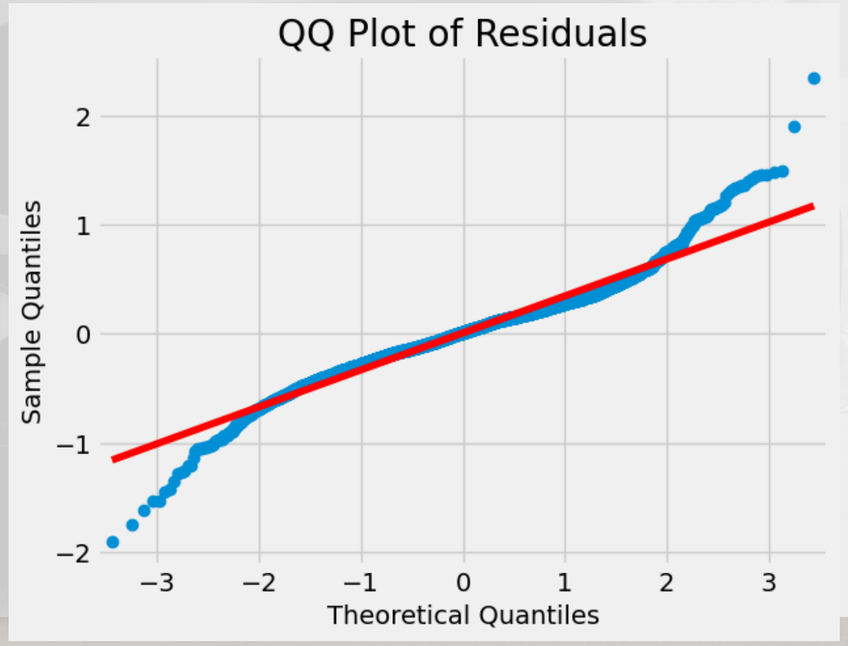


3. Model Building

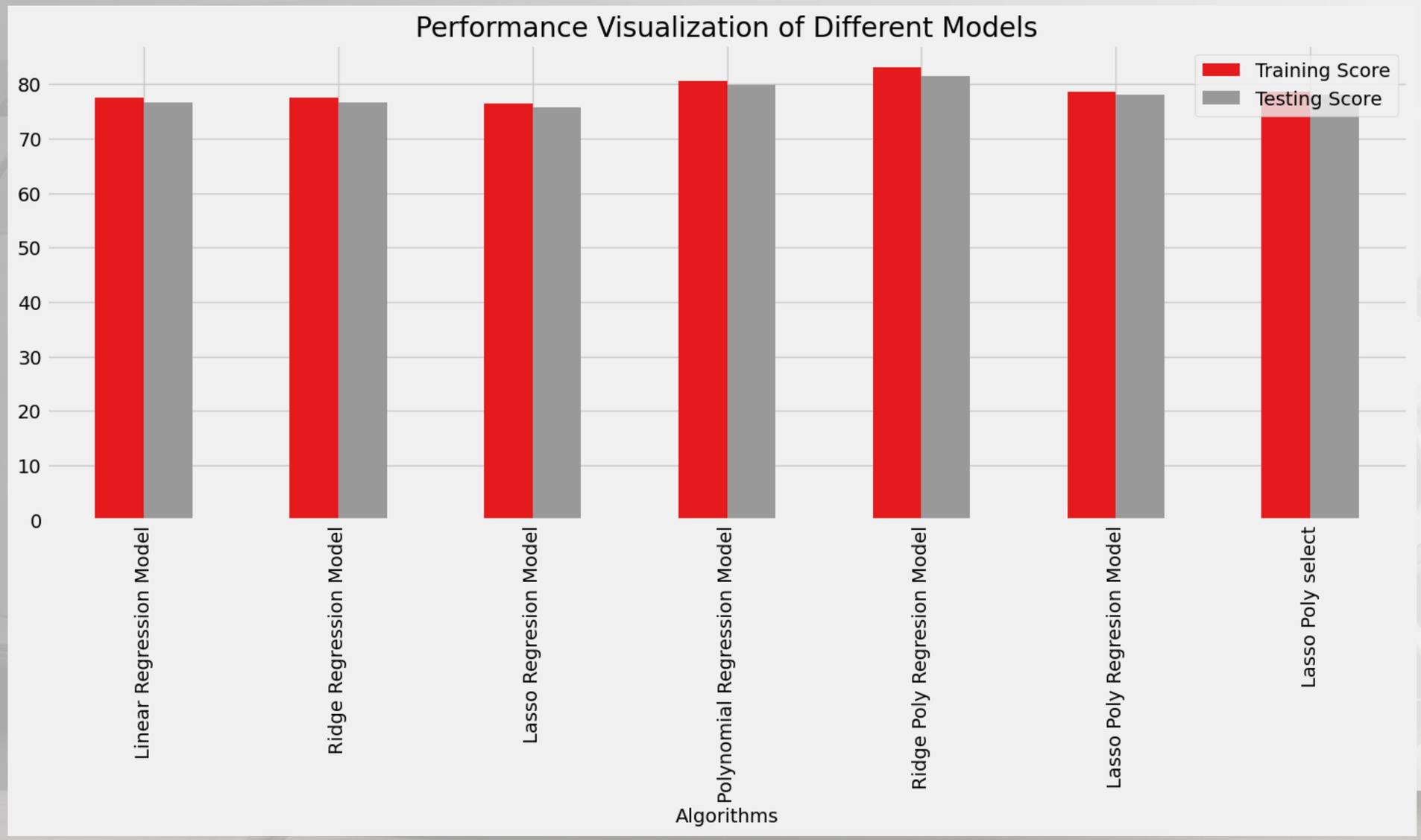
Polynomial Ridge Regression







Model Performance



Model Performance

	Training Score	Testing Score	Algorithms
0	77.5916	76.5947	Linear Regression Model
1	77.5913	76.5971	Ridge Regression Model
2	76.4847	75.6655	Lasso Regresion Model
3	80.5060	79.8510	Polynomial Regression Model
4	83.0927	81.5533	Ridge Poly Regresion Model
5	78.5731	78.1463	Lasso Poly Regresion Model
6	78.6269	78.1457	Lasso Poly select

TESTING UNSEEN DATA

	Actual_values	Predicted_values	Error_values
0	14000.0000	9454.1454	4545.8546
1	155000.0000	79603.4473	75396.5527
2	9800.0000	18918.3803	-9118.3803
3	7900.0000	17706.3923	-9806.3923
4	109100.0000	114046.6757	-4946.6757
5	12500.0000	14379.4799	-1879.4799
6	15300.0000	17321.5283	-2021.5283
7	8500.0000	11827.0773	-3327.0773
8	22800.0000	23874.0431	-1074.0431
9	18500.0000	26695.1865	-8195.1865
10	19600.0000	32883.4450	-13283.4450
11	13200.0000	18481.2970	-5281.2970
12	13200.0000	18481.2970	-5281.2970
13	79500.0000	68888.0117	10611.9883
14	13300.0000	11383.5157	1916.4843
15	3000.0000	7871.4943	-4871.4943
16	20500.0000	14818.4893	5681.5107
17	17500.0000	17195.4505	304.5495
18	17500.0000	17195.4505	304.5495
19	10000.0000	11669.0034	-1669.0034
20	18300.0000	16416.9636	1883.0364
21	8900.0000	8170.1501	729.8499
22	14900.0000	12843.8851	2056.1149
23	4500.0000	4468.5634	31.4366
24	11500.0000	13904.7939	-2404.7939
25	9200.0000	10967.7023	-1767.7023
26	18500.0000	17491.3313	1008.6687

4. conclusion

Car Price Prediction with Machine Learning.

Key-Points

- First we did the **Basic Understanding of Data**
- Then we performed **Data Cleaning** to make the raw data more useable while analysis.
- Then we performed **Exploratory Data Analysis** to generate insights from the data.
- Then we performed **Data Preprocessing** to make data suitable for model training & testing.
- Then we trainined our model using different Machine Learning Algorithms.
- In the end we came with **80% accuracy** which was given by **Ridge Polynormail Regression model**.So we can use this model for predicting price of a car in future.

GOT QUESTIONS?