



REGRESSION TREATMENT CAR PRICE PREDICTION

IBM Machine Learning Professional Certificate:
Course 2 : Supervised Machine Learning : Regression

By Zidan Qurosey Sabilla



CONTENT INDEX



- Dataset Description
- Objective of Analysis

- Exploratory Data Analysis
- Feature Engineering

- Applying various Regression Models
- ML Analysis and Evaluation

DATA DESCRIPTION

ABOUT DATA

The Car Price Prediction dataset, available on Kaggle, is a comprehensive collection of automotive information tailored for predictive modeling. It encompasses a diverse range of features, including specifications like engine size, horsepower, fuel type, and more.

With data sourced from various car models, manufacturers, and regions, this dataset offers a rich and varied landscape for exploring the factors influencing car prices.

Whether you're a data scientist, machine learning enthusiast, or automotive analyst, this dataset provides an exciting opportunity to delve into the intricacies of the automotive market and develop predictive models that can accurately forecast car prices based on key attributes.



Car Price Prediction Multiple Linear Regression

Predicting the Prices of cars using RFE and VIF

 [kaggle.com](https://www.kaggle.com)

Data Source : <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/>

Notebook : <https://github.com/zidangrs/IBM-Machine-Learning-Course/blob/master/2-Supervised-Machine-Learning-Regression/Project-2.ipynb>



DATA OVERVIEW

	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	carheight
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3

- **Car_ID** : Unique ID of each observation (Integer)
- **Symboling** : Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. (Categorical)
- **carCompany** : Name of car company (Categorical)
- **fueltype** : Car fuel type i.e. gas or diesel (Categorical)
- **aspiration** : Aspiration used in a car (Categorical)
- **doornumber** : Number of doors in a car (Categorical)
- **carbody** : body of car (Categorical)
- **drivewheel** : type of drive wheel (Categorical)
- **enginelocation** : Location of car engine (Categorical)
- **wheelbase** : Wheelbase of car (Numeric)
- **carlength** : Length of car (Numeric)
- **carwidth** : Width of car (Numeric)
- **carheight** : Height of car (Numeric)

DATA OVERVIEW

curbweight	enginetype	cylindernumber	enginesize	fuelsystem	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
2548	dohc	four	130	mpfi	3.47	2.68	9.0	111	5000	21	27	13495.0
2548	dohc	four	130	mpfi	3.47	2.68	9.0	111	5000	21	27	16500.0
2823	ohcv	six	152	mpfi	2.68	3.47	9.0	154	5000	19	26	16500.0
2337	ohc	four	109	mpfi	3.19	3.40	10.0	102	5500	24	30	13950.0
2824	ohc	five	136	mpfi	3.19	3.40	8.0	115	5500	18	22	17450.0

- **curbweight** : The weight of a car without occupants or baggage (Numeric)
- **enginetype** : Type of engine (Categorical)
- **cylindernumber** : Cylinder placed in the car (Categorical)
- **enginesize** : Size of car (Numeric)
- **fuelsystem** : Fuel system of car (Categorical)
- **boreratio** : Boreratio of car (Numeric)
- **stroke** : Stroke or volume inside the engine
- **compressionratio** : Compression ratio of car (Numeric)
- **horsepower** : Horsepower (Numeric)
- **peakrpm** : Car peak rpm (Numeric)
- **citympg** : Mileage in city (numeric)
- **highwaympg** : Mileage on highway (Numeric)
- **price** : Price of car (Numeric **Dependent Variable**)

DATA OVERVIEW

	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	enginetype	cylindernumber	fuelsystem
count	205	205	205	205	205	205	205	205	205	205	205
unique	6	147	2	2	2	5	3	2	7	7	8
top	0	toyota corona	gas	std	four	sedan	fwd	front	ohc	four	mpfi
freq	67	6	185	168	115	96	120	202	148	159	94

	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
count	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00	205.00
mean	98.76	174.05	65.91	53.72	2555.57	126.91	3.33	3.26	10.14	104.12	5125.12	25.22	30.75	13276.71
std	6.02	12.34	2.15	2.44	520.68	41.64	0.27	0.31	3.97	39.54	476.99	6.54	6.89	7988.85
min	86.60	141.10	60.30	47.80	1488.00	61.00	2.54	2.07	7.00	48.00	4150.00	13.00	16.00	5118.00
25%	94.50	166.30	64.10	52.00	2145.00	97.00	3.15	3.11	8.60	70.00	4800.00	19.00	25.00	7788.00
50%	97.00	173.20	65.50	54.10	2414.00	120.00	3.31	3.29	9.00	95.00	5200.00	24.00	30.00	10295.00
75%	102.40	183.10	66.90	55.50	2935.00	141.00	3.58	3.41	9.40	116.00	5500.00	30.00	34.00	16503.00
max	120.90	208.10	72.30	59.80	4066.00	326.00	3.94	4.17	23.00	288.00	6600.00	49.00	54.00	45400.00

DATA OVERVIEW

```
car_ID      0
symboling   0
CarName     0
fueltype    0
aspiration  0
doornumber  0
carbody     0
drivewheel  0
enginelocation  0
wheelbase   0
carlength   0
carwidth    0
carheight   0
curbweight  0
enginetype  0
cylindernumber  0
enginesize  0
fuelsystem  0
boreratio   0
stroke      0
compressionratio  0
horsepower  0
peakrpm     0
citympg     0
highwaympg  0
price       0
dtype: int64
```

The data has 11 categorical variables and 14 numerical variables, and it appears the data doesn't have missing values on it. So it's great to go!

Objective of the Analysis

1

EXPLORATORY DATA ANALYSIS

Through Exploratory Data Analysis (EDA), We will employ Scatter Plots and correlation heatmaps to visualize numerical data relationships, analyze data distribution, and plot categorical data using visualizations like bar charts. This succinct approach aims to swiftly uncover patterns and insights within our dataset, guiding subsequent analytical steps effectively.

2

FEATURE ENGINEERING

In our Feature Engineering objective, we focus on aligning the dataset with regression assumptions. Applying Log, Reciprocal, and Box-Cox transformations to the dependent variable ensures a normal distribution, while OneHotEncoder swiftly converts categorical variables to binary, meeting the prerequisites for effective regression modeling.

3

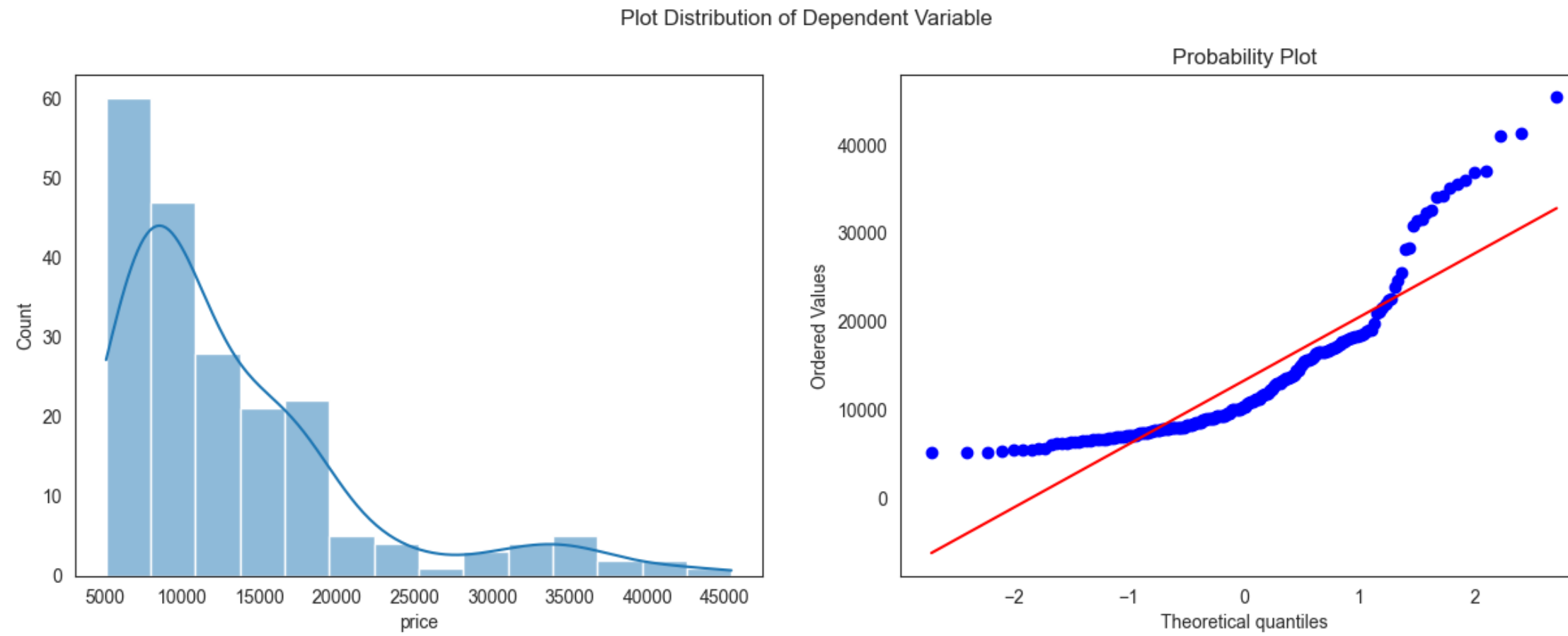
MODELLING & EVALUATION

In Modelling & Evaluation, we'll utilize Baseline Linear Regression, Lasso, Ridge, and Elastic Net on the dataset, and incorporate polynomial features. This succinct approach aims to identify the best model for accurate predictions, streamlining our focus on optimal predictive performance.



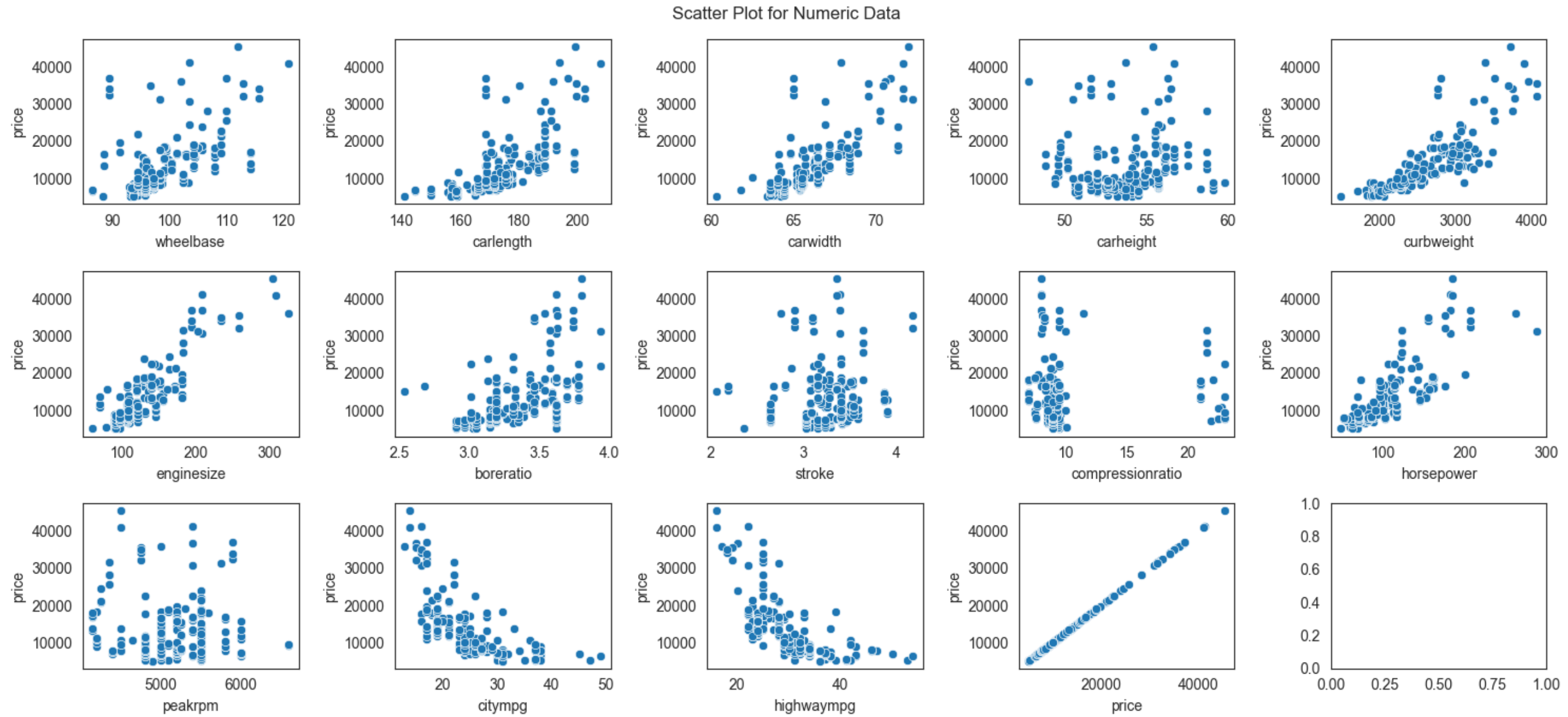
EXPLORATORY DATA ANALYSIS

NORMALITY ASUMPTION



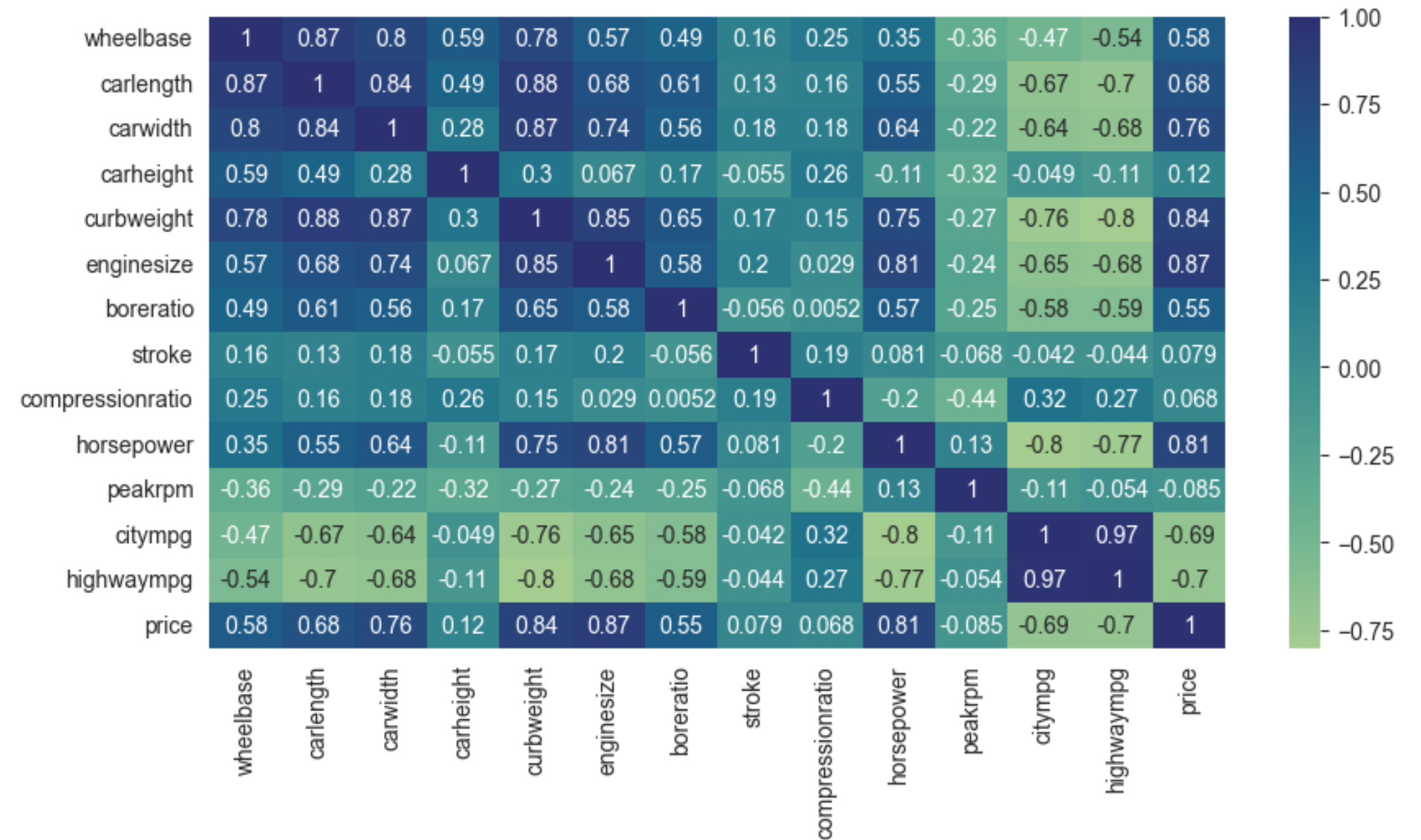
The P-value (Kolmogorov-Smirnov Test) is 0.000 with a significance value of 0.05, thus rejecting the null hypothesis where the null hypothesis means the data is normally distributed. Next, we will apply Log, reciprocal, and BoxCox transformation for changing the distribution of dependent variable.

VISUALIZE NUMERIC DATA

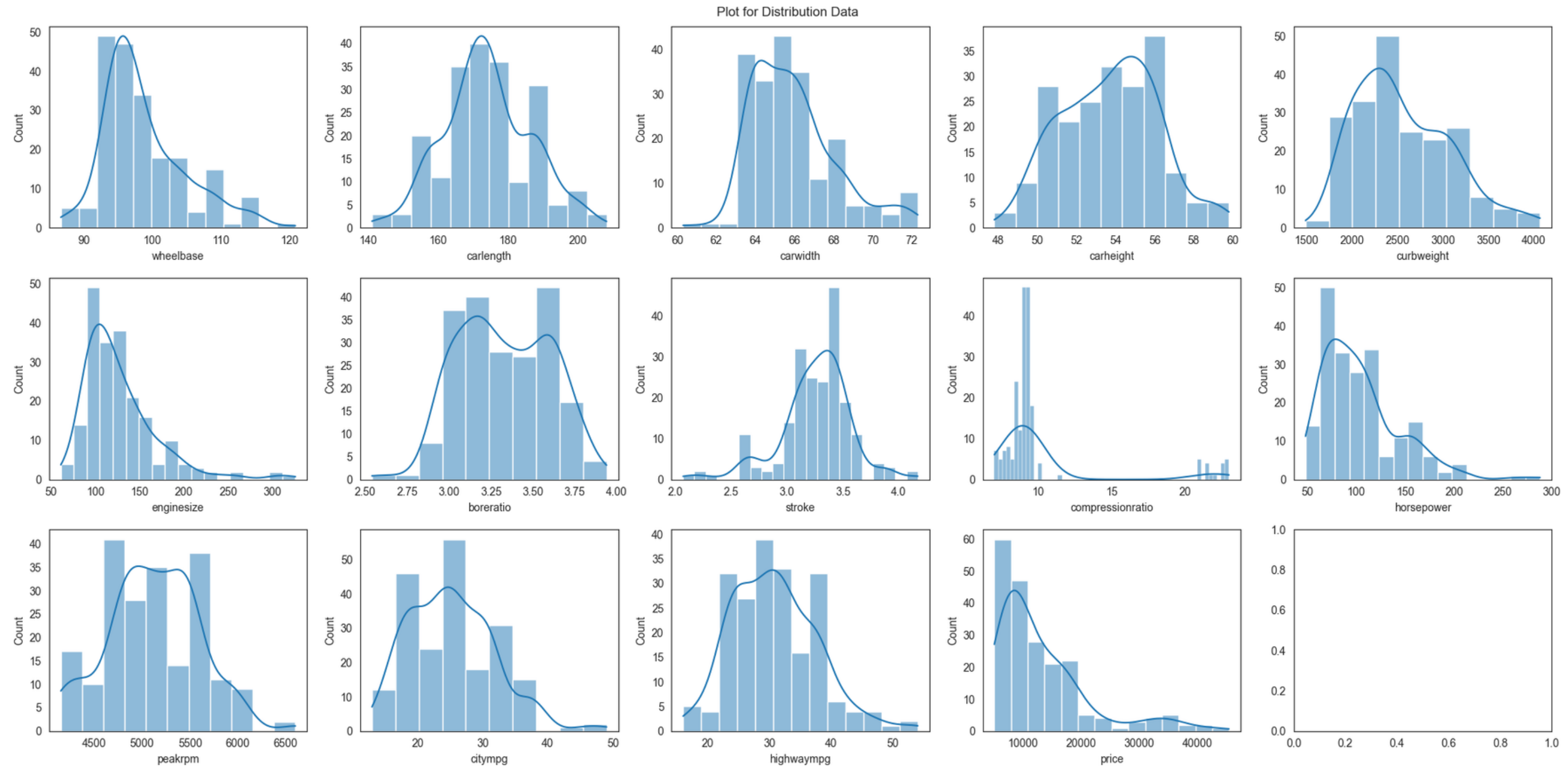


VISUALIZE NUMERIC DATA

It can be seen that some variables has a high correlation with the dependent variable, but the variable **peakrpm**, **compressionratio**, and **stroke** have very low correlation.

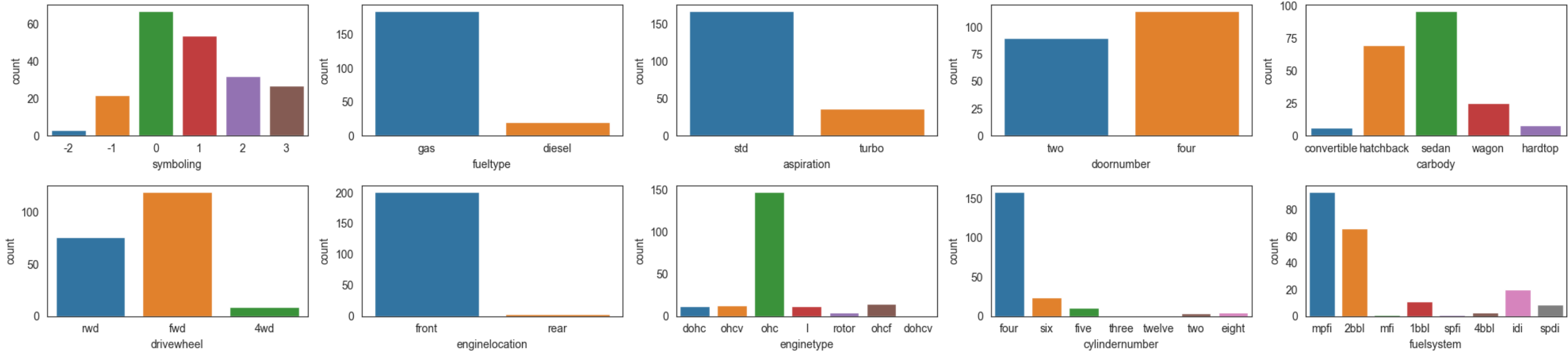


VISUALIZE NUMERIC DATA



VISUALIZE CATEGORIC DATA


Plot for Categorical Data



FEATURE ENGINEERING

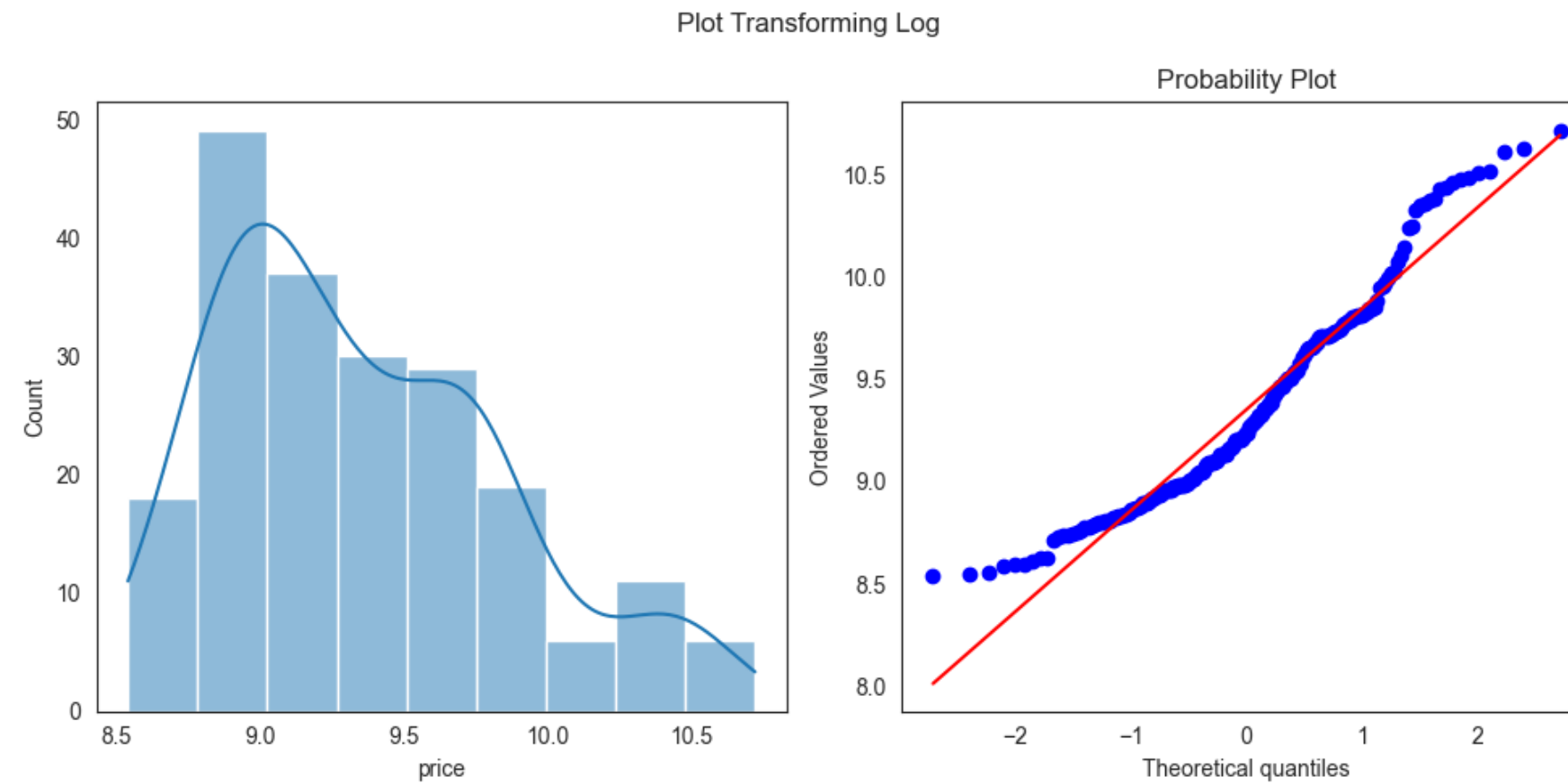
CONVERT CATEGORIC DATA

```
array(['symboling', 'fueltype', 'aspiration', 'doornumber', 'carbody',  
      'drivewheel', 'enginelocation', 'enginetype', 'cylindernumber',  
      'fuelsystem'], dtype=object)
```



```
array(['symboling_-2', 'symboling_-1', 'symboling_0', 'symboling_1',  
      'symboling_2', 'symboling_3', 'fueltype_diesel', 'fueltype_gas',  
      'aspiration_std', 'aspiration_turbo', 'doornumber_four',  
      'doornumber_two', 'carbody_convertible', 'carbody_hardtop',  
      'carbody_hatchback', 'carbody_sedan', 'carbody_wagon',  
      'drivewheel_4wd', 'drivewheel_fwd', 'drivewheel_rwd',  
      'enginelocation_front', 'enginelocation_rear', 'enginetype_dohc',  
      'enginetype_dohcv', 'enginetype_l', 'enginetype_ohc',  
      'enginetype_ohcf', 'enginetype_ohcv', 'enginetype_rotor',  
      'cylindernumber_eight', 'cylindernumber_five',  
      'cylindernumber_four', 'cylindernumber_six',  
      'cylindernumber_three', 'cylindernumber_twelve',  
      'cylindernumber_two', 'fuelsystem_1bbl', 'fuelsystem_2bbl',  
      'fuelsystem_4bbl', 'fuelsystem_idi', 'fuelsystem_mfi',  
      'fuelsystem_mphi', 'fuelsystem_spdi', 'fuelsystem_spfi'],  
      dtype=object)
```

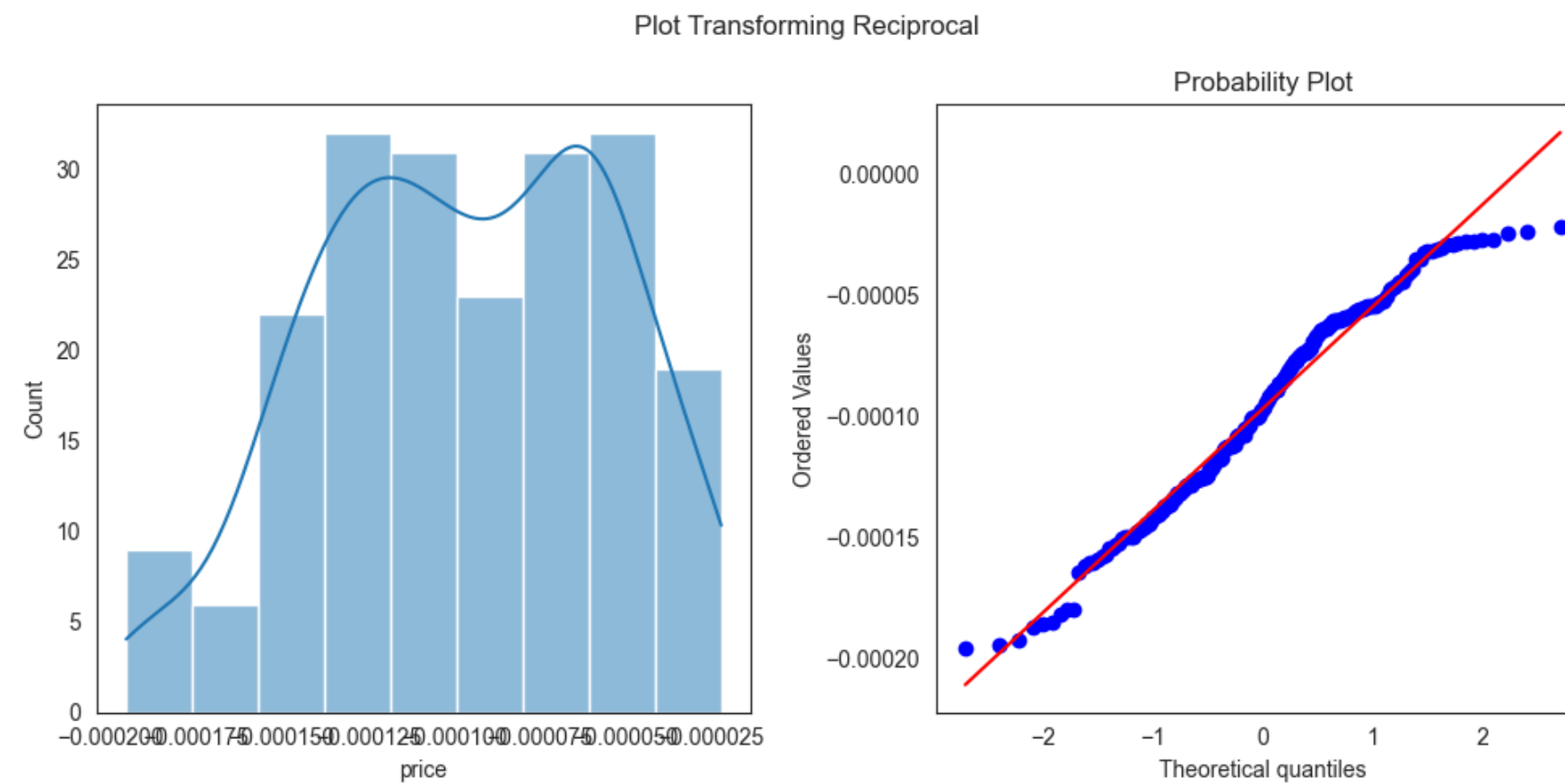
LOG TRANSFORMATION



Kolmogorov-Smirnov Test

P-Value = 0.000

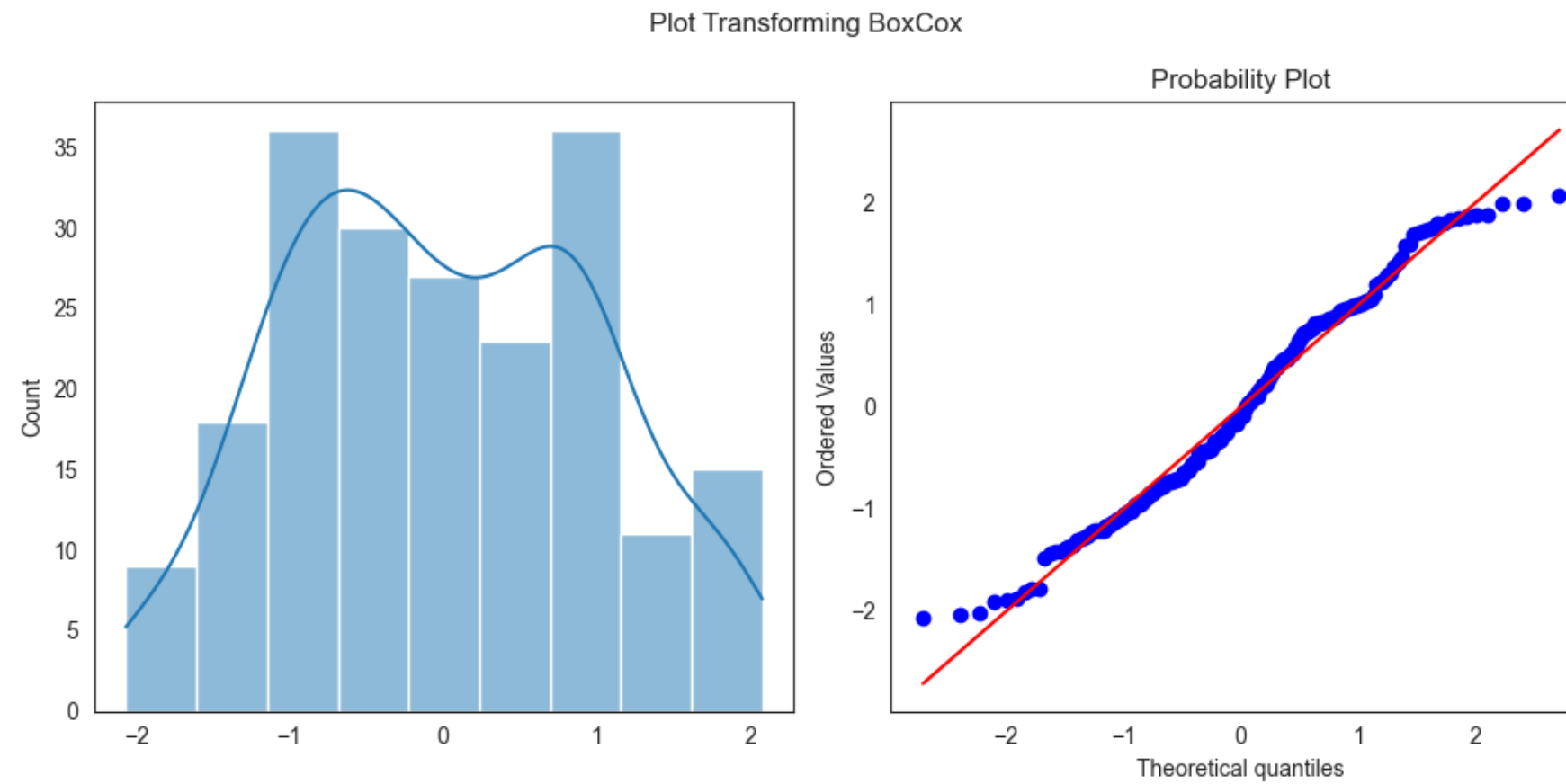
RECIPROCAL TRANSFORMATION



Kolmogorov-Smirnov Test

P-Value = 6.438497839806477e-48

BOX-COX TRANSFORMATION

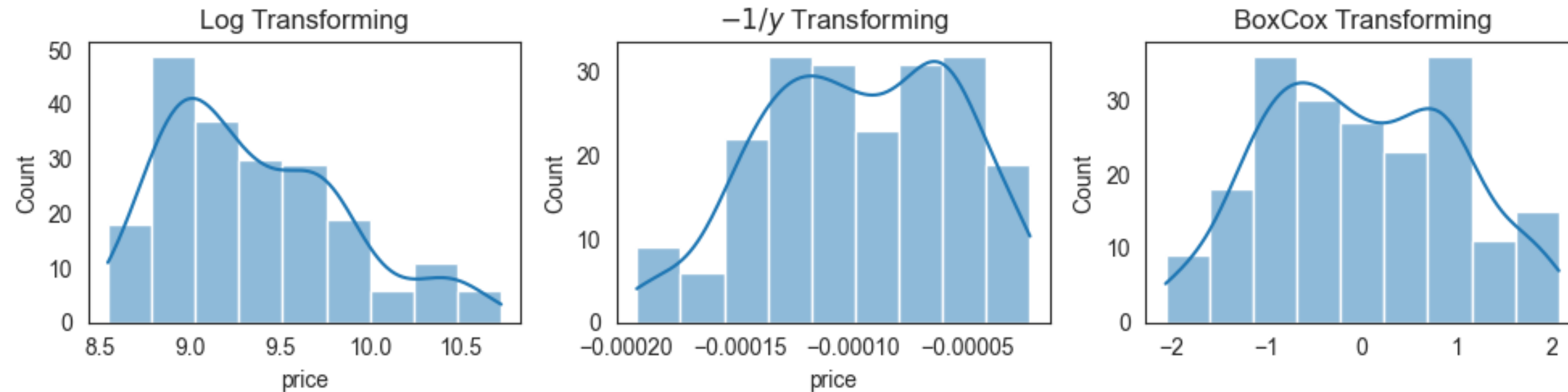


Kolmogorov-Smirnov Test

P-Value = 0.302

DETERMINING NORMALITY

Result of Transforming Dependent Variable



Only the Box-Cox Transformation that the P-Value is not rejecting the null hypothesis.
So with the Box-Cox Transformation, the dependent variable are normally distributed.

	Method	P-Value
0	Log	0.000
1	-1/y	0.000
2	BoxCox	0.302

MODELLING & EVALUATION

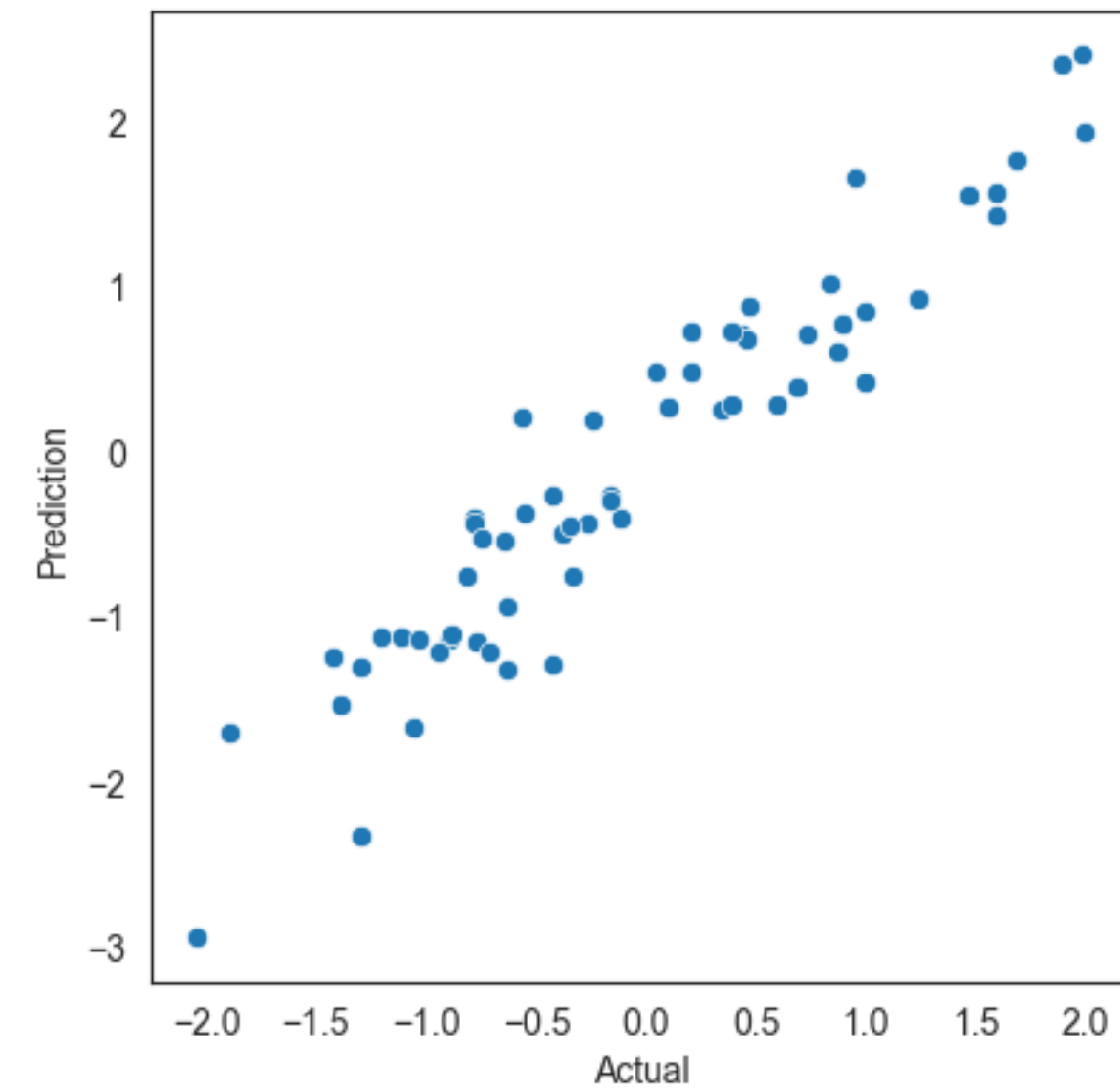
MODEL 1

Baseline Linear Regression

$R^2 = 0.927$

CV RMSE = [0.38105016 0.37823469 0.35377234 0.38250705]

CV RMSE Mean = 0.3738910617512583



RMSE Testing : 0.36848104390970576

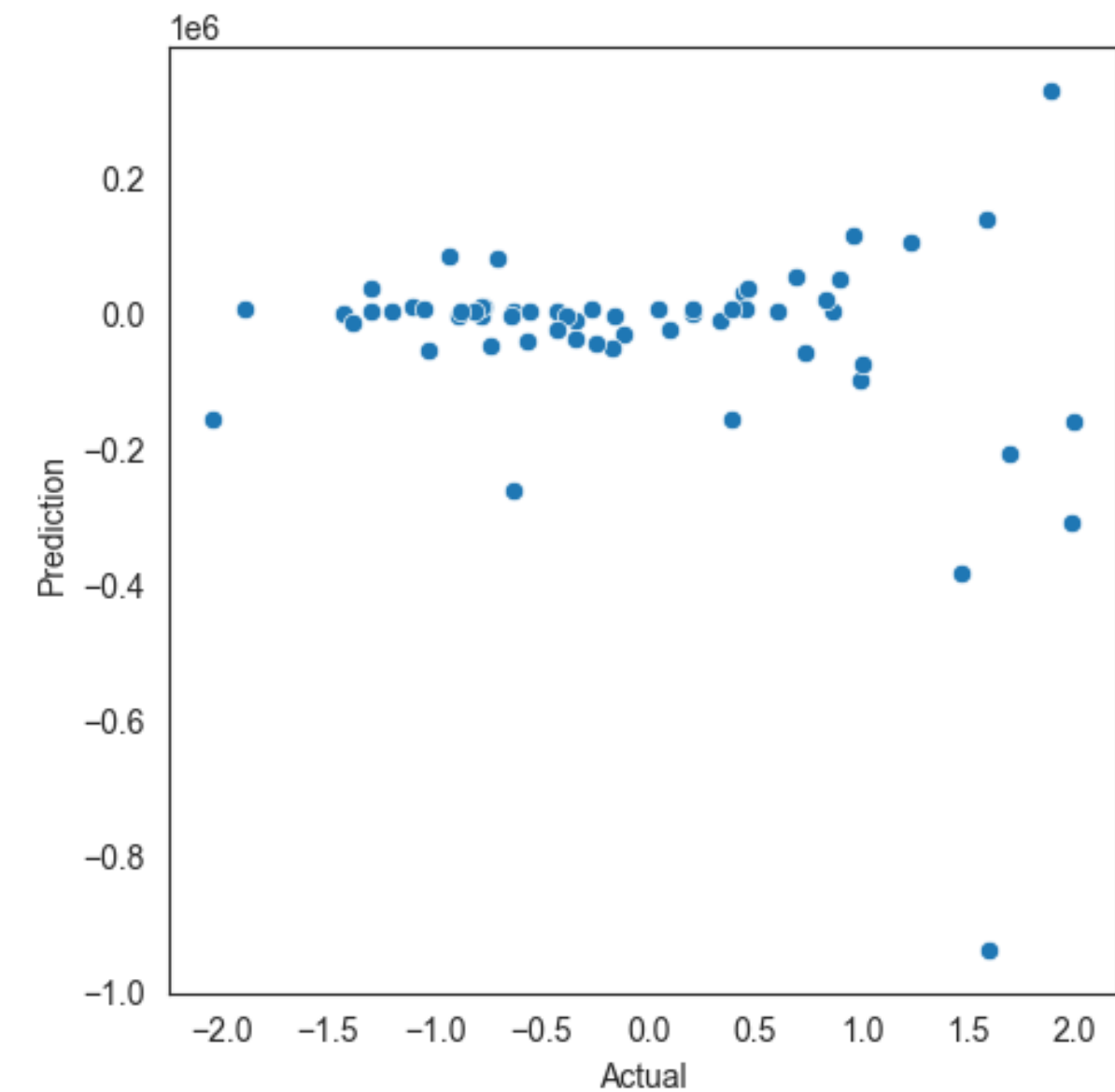
MODEL 2

Polynomial Regression (degree=3)

$R^2 = 0.996$

CV RMSE = [9.695148e+02 1.82351007e+05 5.77679331e+01
4.36480230e+02]

CV RMSE Mean = 45953.69240560229



RMSE Testing : 156560.93660299678

MODEL 3

Lasso Regression

Parameter

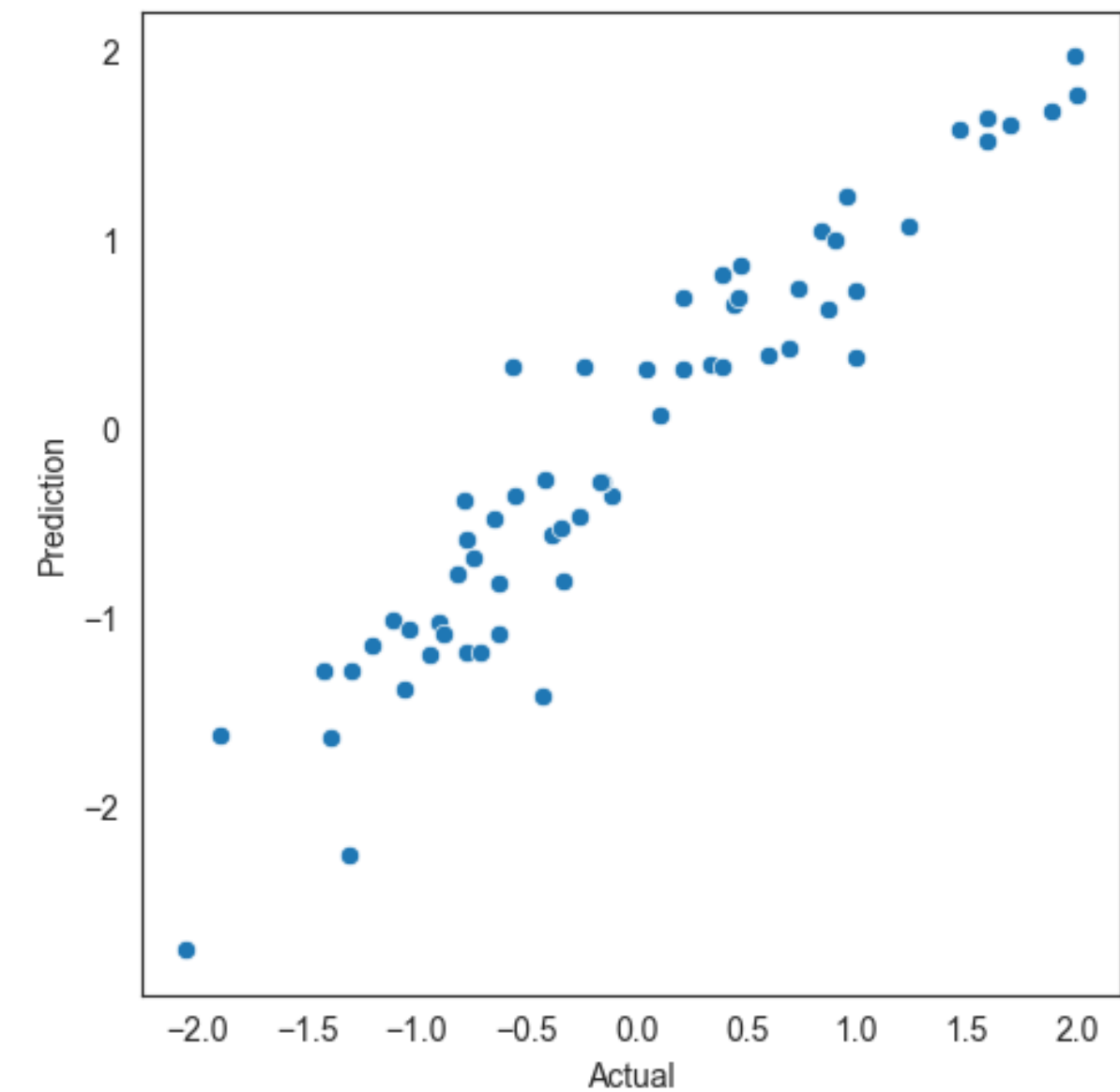
Alpha = 0.0036696850832181275

Metrics

R2 = 0.919

CV RMSE = [0.34494462 0.35785994 0.34471434 0.33327068]

CV RMSE Mean = 0.34519739268144656



RMSE Testing : 0.33398862769025334

MODEL 4

Lasso Polynomial Regression (degree = 3)

Parameter

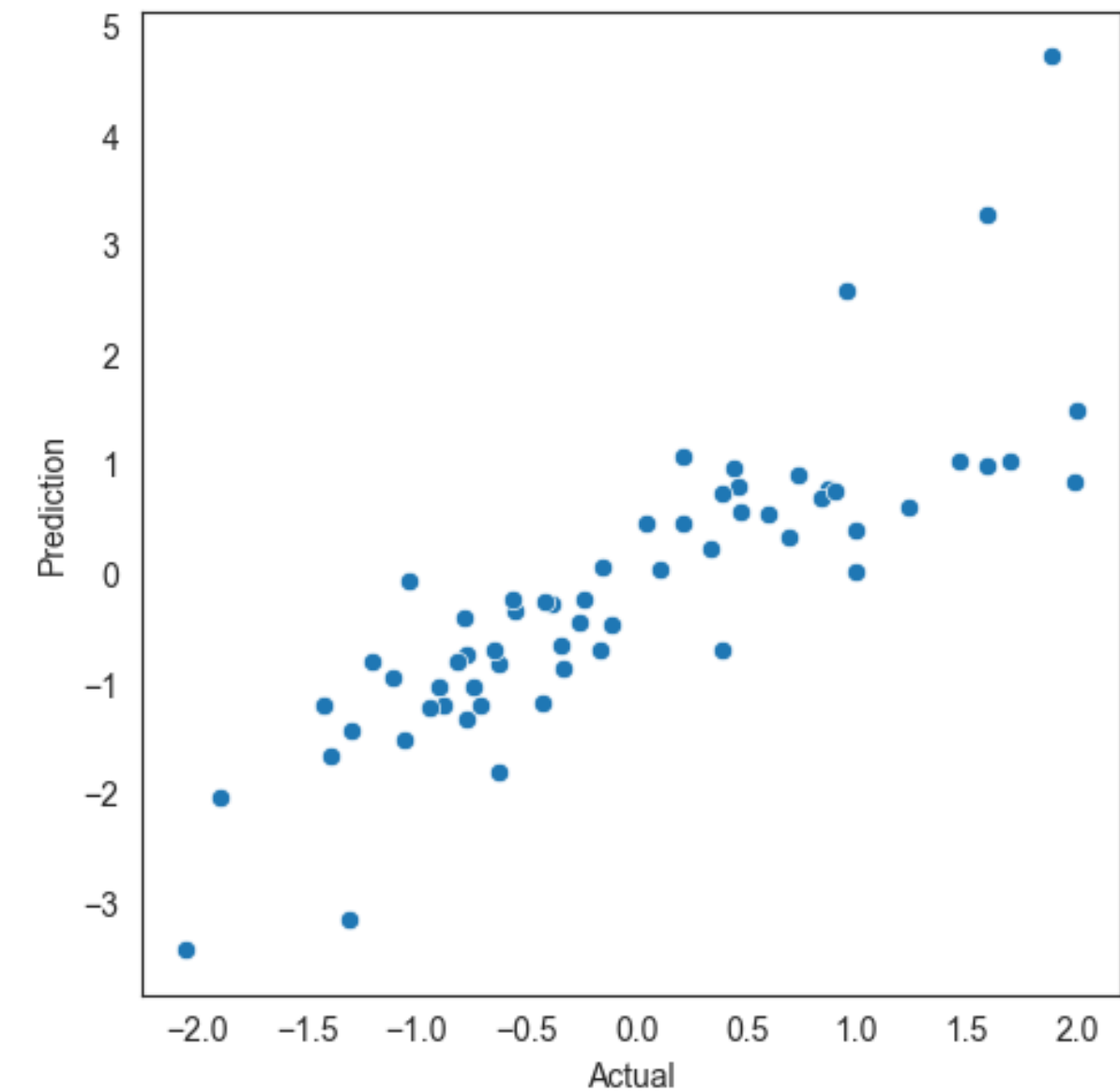
Alpha = 46.31611723780748

Metrics

R2 = 0.971

CV RMSE = [0.76234581 0.43527632 0.4671957 0.87320495]

CV RMSE Mean = 0.6345056948880022



RMSE Testing : 0.7115543163954181

MODEL 5

Ridge Regression

Parameter

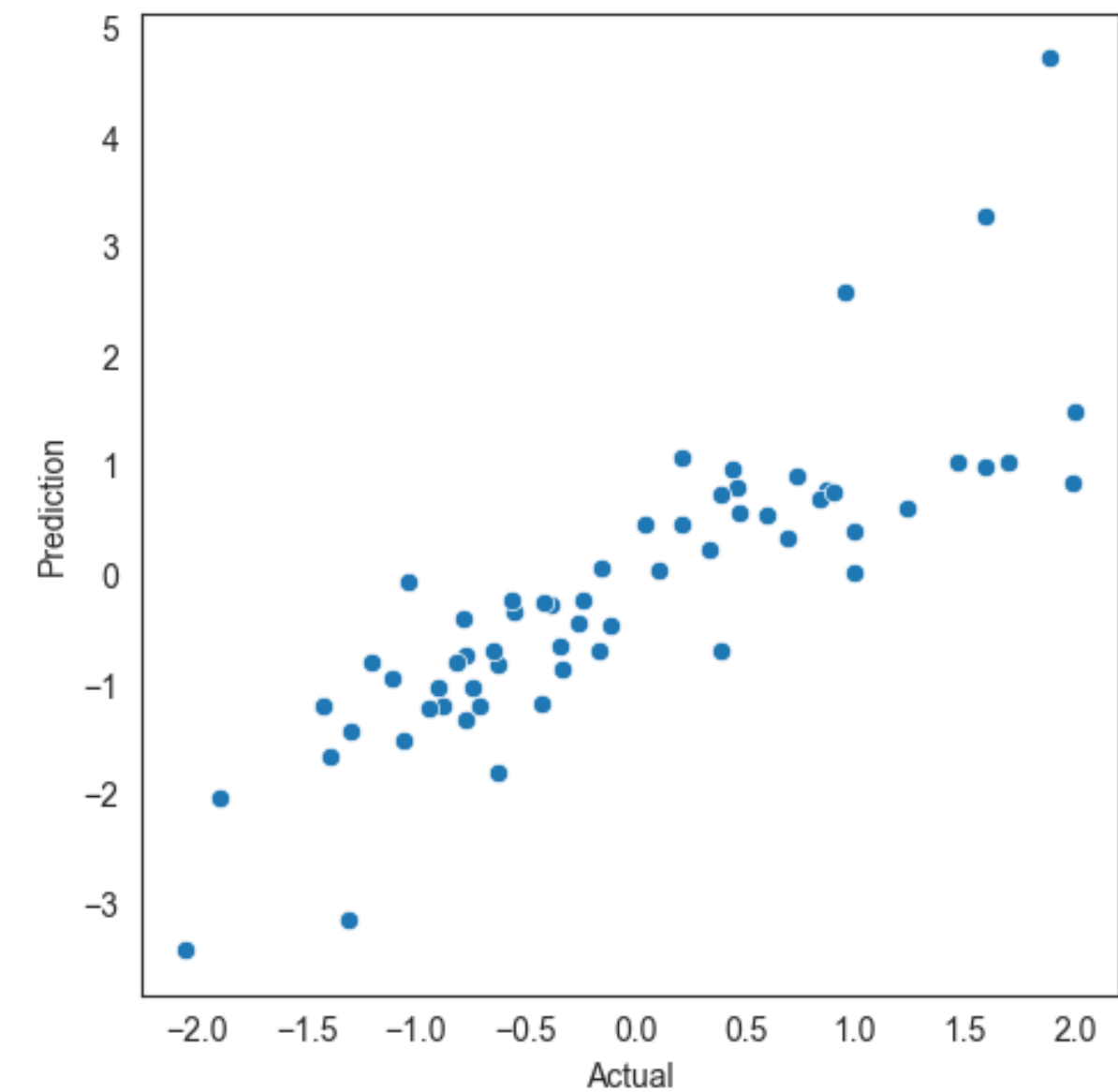
Alpha = 1.8985883261921552

Metrics

R2 = 0.924

CV RMSE = [0.35033307 0.35718678 0.3496061 0.35019249]

CV RMSE Mean = 0.35182961010279323



RMSE Testing : 0.33974494971093816

MODEL 6

Ridge Polynomial Regression (degree = 3)

Parameter

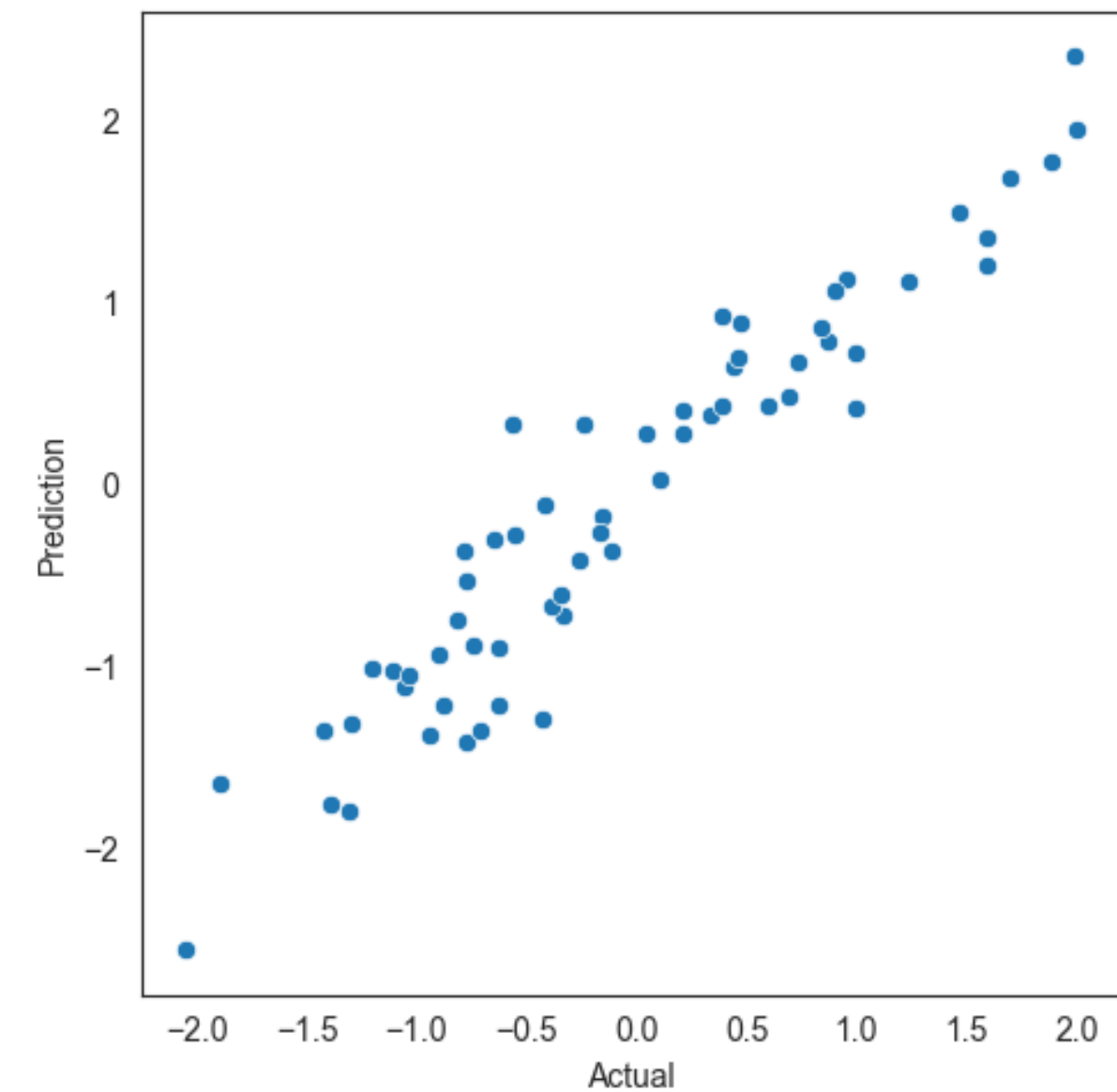
Alpha = 1.8995240732827625

Metrics

R2 = 0.914

CV RMSE = [0.32800631 0.35200752 0.36296785 0.40833523]

CV RMSE Mean = 0.36282922832581854



RMSE Testing : 0.32773604480047663

MODEL 7

Elastic Net Regression

Parameter

Alpha = 0.007972184720447488

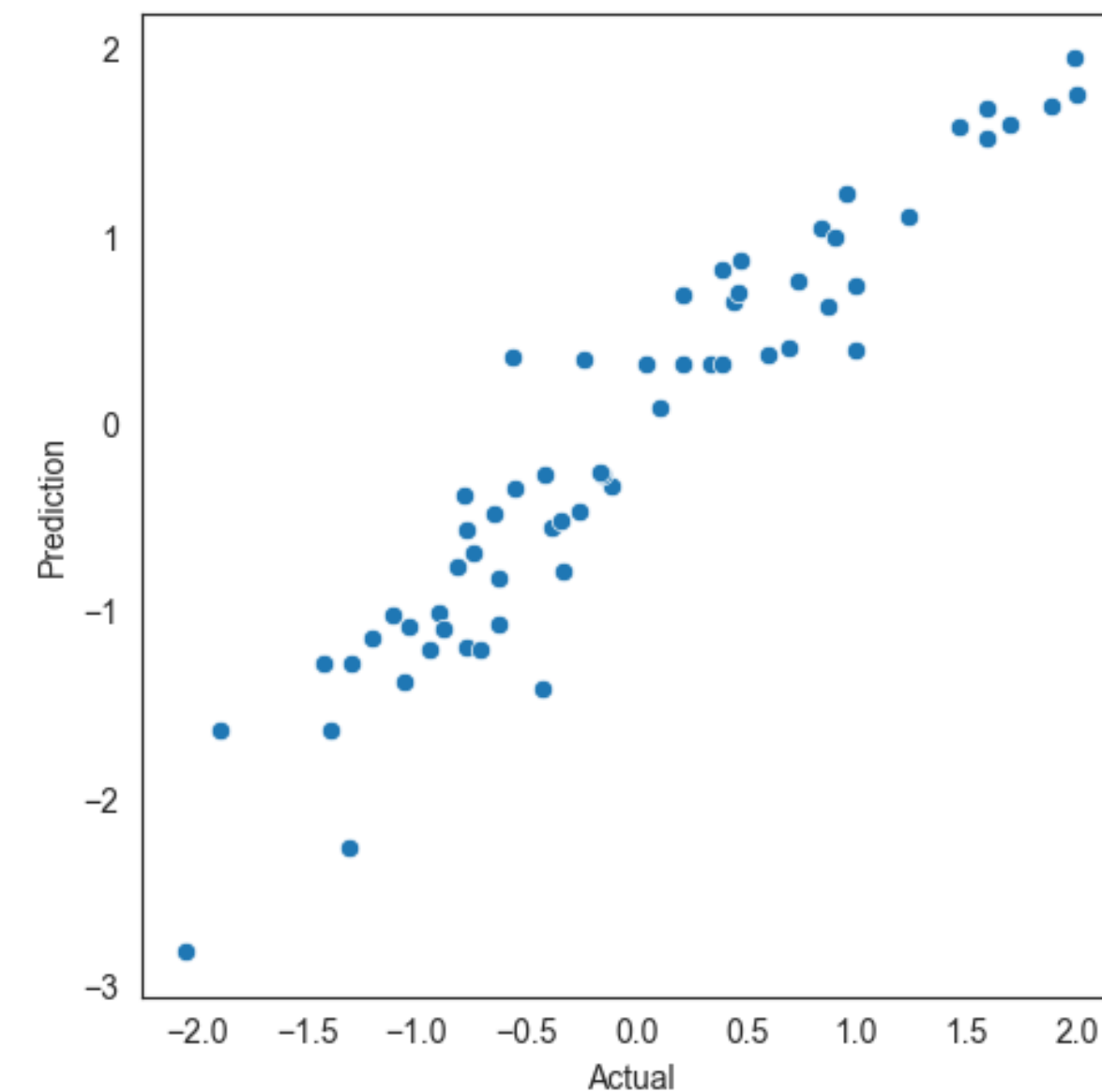
L1 Ratio = 0.316711512080725

Metrics

R2 = 0.920

CV RMSE = [0.34684147 0.35535837 0.3456632 0.33798389]

CV RMSE Mean = 0.34646173340606884



RMSE Testing : 0.3374714706015728

MODEL 8

Elastic Net Polynomial Regression (degree=2)

Parameter

Alpha = 54.964619082268285

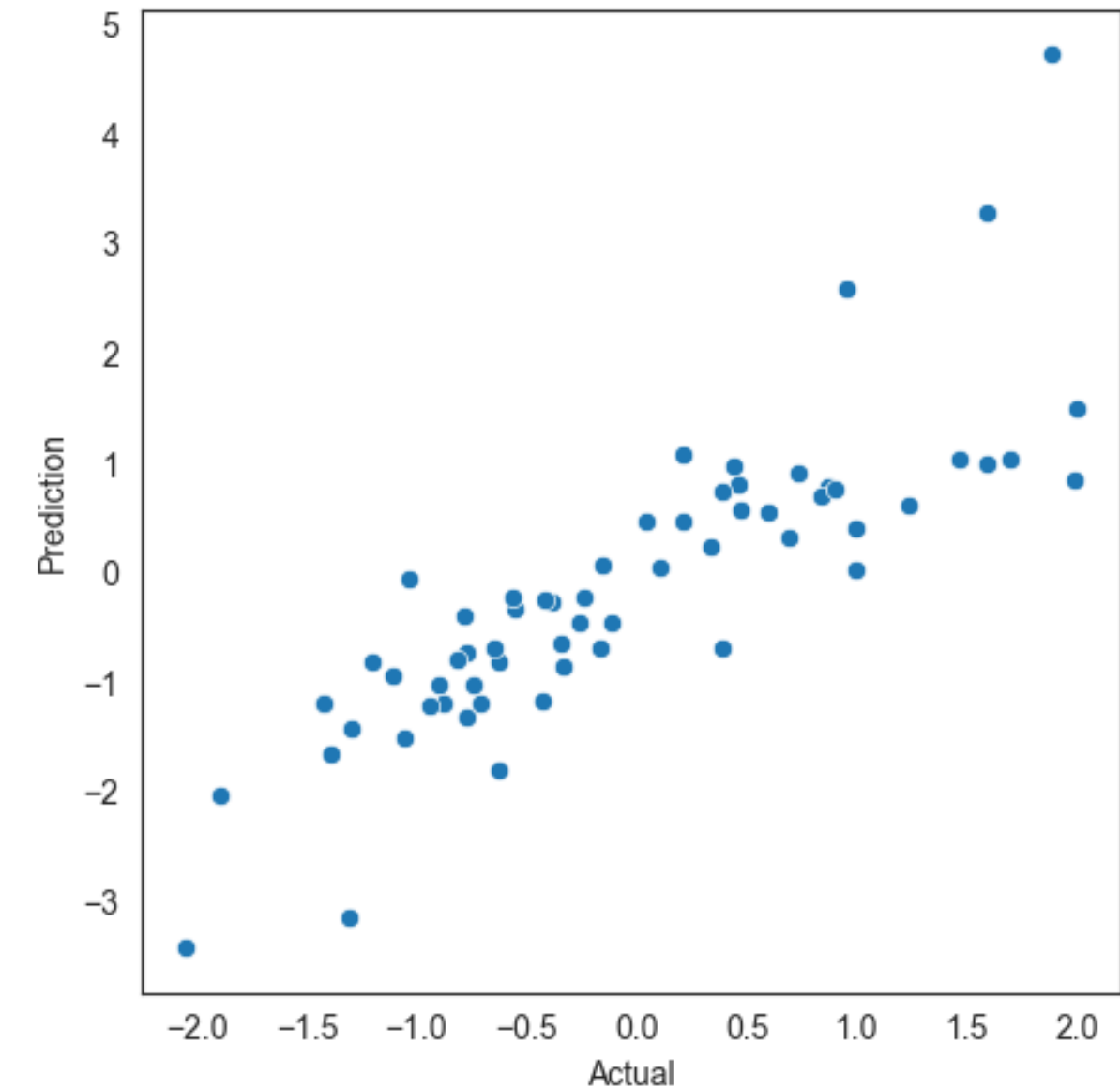
L1 Ratio = 0.8418943037846904

Metrics

R2 = 0.971

CV RMSE = [0.76238423 0.43533113 0.46720449 0.87357962]

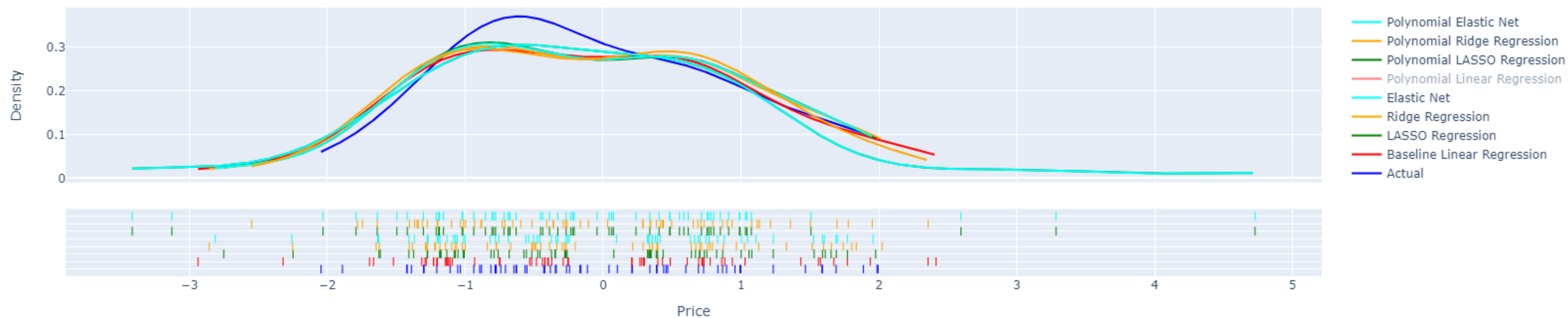
CV RMSE Mean = 0.634624867458729



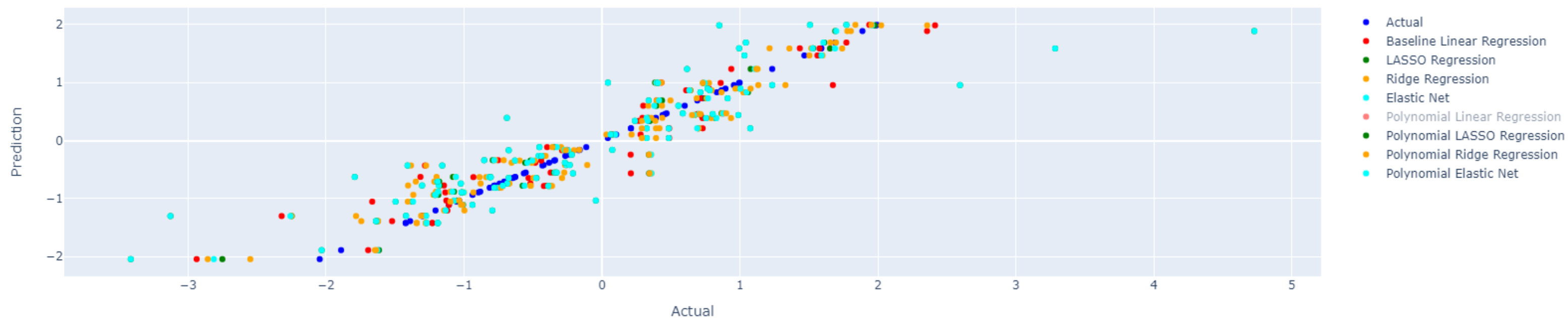
RMSE Testing : 0.7116338544495328

MODEL ANALYSIS

Distribution Plot



Scatter Plot



MODEL ANALYSIS

As indicated in the table and preceding analysis, it is evident that the **Polynomial Linear Regression model exhibits signs of overfitting**. This is evident from the **significantly high scores in both CV RMSE Mean and RMSE Testing**. However, despite these drawbacks, it boasts the **highest R2 among all models**, indicating a strong fit to the data.

In terms of model selection, the **Lasso Regression** method emerges as the most favorable, with the lowest CV RMSE Mean. Conversely, when considering the lowest RMSE Testing, the **Polynomial Ridge Regression** model stands out.

Upon thorough evaluation, it becomes apparent that the **Lasso Regression model outperforms others**, as it consistently demonstrates the lowest CV RMSE Mean and RMSE Testing. Therefore, **the Lasso Regression model is identified as the best-performing model for accurate predictions**.

	Method	CV RMSE Mean	RMSE Testing Data	R^2
0	Baseline Linear Regression	0.373891	0.368481	0.927310
1	Polynomial Linear Regression	45953.692406	156560.936603	0.995745
2	LASSO Regression	0.345197	0.333989	0.919076
3	Polynomial LASSO Regression	0.634506	0.711554	0.971162
4	Ridge Regression	0.351830	0.339745	0.924333
5	Polynomial Ridge Regression	0.362829	0.327736	0.913829
6	Elastic Net	0.346462	0.337471	0.919830
7	Polynomial Elastic Net	0.634625	0.711634	0.971167

Sorted:

	Method	CV RMSE Mean	RMSE Testing Data	R^2
2	LASSO Regression	0.345197	0.333989	0.919076
6	Elastic Net	0.346462	0.337471	0.919830
4	Ridge Regression	0.351830	0.339745	0.924333
5	Polynomial Ridge Regression	0.362829	0.327736	0.913829
0	Baseline Linear Regression	0.373891	0.368481	0.927310
3	Polynomial LASSO Regression	0.634506	0.711554	0.971162
7	Polynomial Elastic Net	0.634625	0.711634	0.971167
1	Polynomial Linear Regression	45953.692406	156560.936603	0.995745

CONCLUSION

In conclusion, the comprehensive analysis of various regression models sheds light on their strengths and limitations. Despite the overfitting concerns observed in the Polynomial Linear Regression model, its remarkable R^2 value indicates a robust fit to the data. However, for optimal predictive performance, the Lasso Regression model emerges as the most favorable choice, consistently demonstrating the lowest CV RMSE Mean and RMSE Testing across the evaluated models.

While the Lasso Regression model proves effective for accurate predictions, it is essential to consider the complexity of advanced algorithms like XGBoost, LGBM, and Random Forest. While these algorithms may offer improved predictive capabilities, they introduce complexity that may not always be warranted. Therefore, for statistical analyses where interpretability and simplicity are valued, it is prudent to adhere to linear regression. Specifically, using linear regression with adherence to key assumptions such as normality of residuals and addressing issues like autocorrelation ensures a reliable foundation. Additionally, conducting thorough diagnostic tests, including Overall and Partial tests, contributes to the robustness of the analysis. Striking a balance between predictive power and interpretability is crucial, tailoring the choice of models based on the specific objectives and complexity considerations.



THANK YOU!

IBM Machine Learning Professional Certificate:
Course 2 : Supervised Machine Learning : Regression



By Zidan Qurosey Sabilla