أكاديمية سدايا
SDAIA Academy

Predict Car Prices

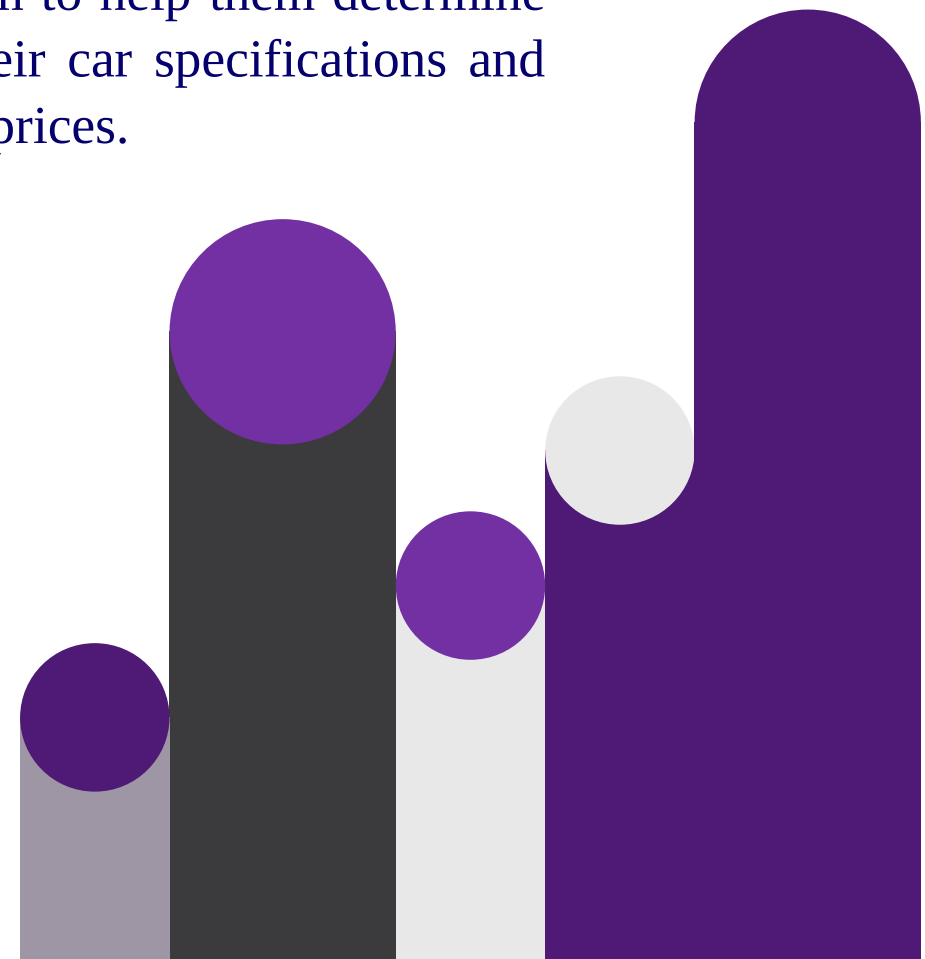
Scraping and Regression

Presented by:
Ghadah Alharbi and Rahaf Alyousef

24-10-2021

Business Objective

Most people have the problem of not knowing the price of a car when selling it. So, we will use data from cars.com to help them determine the expected price of their car based on their car specifications and build a Regression Algorithms to predict car prices.



Data Description

Using cars.com dataset. The focus is on Audi cars.
Using Python library Beautiful Soup to scrape the cars information.

	DESCRIPTION	DEALER	MILES_DRIVEN	RATE	NO_OF_REVIEWS	PRICE	MODEL	CAR_NAME
0	[2018, Audi A4 2.0T Premium Plus]	The Audi Exchange	43152.0	4.9	654	29587.0	2018	Audi A4 2.0T Premium Plus
1	[2016, Audi Q5 3.0T Premium Plus]	Fletcher Jones Audi	32666.0	4.8	1048	31881.0	2016	Audi Q5 3.0T Premium Plus
2	[2012, Audi S5 3.0 Prestige quattro]	Volkswagen of Downtown Chicago	28552.0	4.6	1442	29000.0	2012	Audi S5 3.0 Prestige quattro
3	[2021, Audi Q3 45 S line Premium]	Volkswagen of Downtown Chicago	16027.0	4.6	1442	39000.0	2021	Audi Q3 45 S line Premium
4	[2014, Audi Q5 2.0T Premium Plus]	Toyota of Lincoln Park	64698.0	4.2	216	20900.0	2014	Audi Q5 2.0T Premium Plus
...	...	...	...	...	...	...	...	...
791	[2014, Audi S4 3.0T Premium Plus]	Adam Auto Group	124582.0	4.6	1074	20885.0	2014	Audi S4 3.0T Premium Plus
792	[2013, Audi A5 2.0T Premium Plus]	Guaranteed Motor Cars	74437.0	4.7	3942	24900.0	2013	Audi A5 2.0T Premium Plus
793	[2011, Audi S5 4.2 Premium Plus quattro]	Coda Motors	122279.0	4.2	55	16990.0	2011	Audi S5 4.2 Premium Plus quattro
794	[2019, Audi Q7 55 Prestige]	Audi Morton Grove	48335.0	4.8	630	52999.0	2019	Audi Q7 55 Prestige
795	[2010, Audi Q7 3.6 Prestige]	ACL Sales & Leasing	119432.0	4.6	634	14995.0	2010	Audi Q7 3.6 Prestige

796 rows × 8 columns

Preprocessing

1

Check if there are any missing values.

2

Convert the datatype of MILES_DRIVEN, MODEL, and PRICE.

3

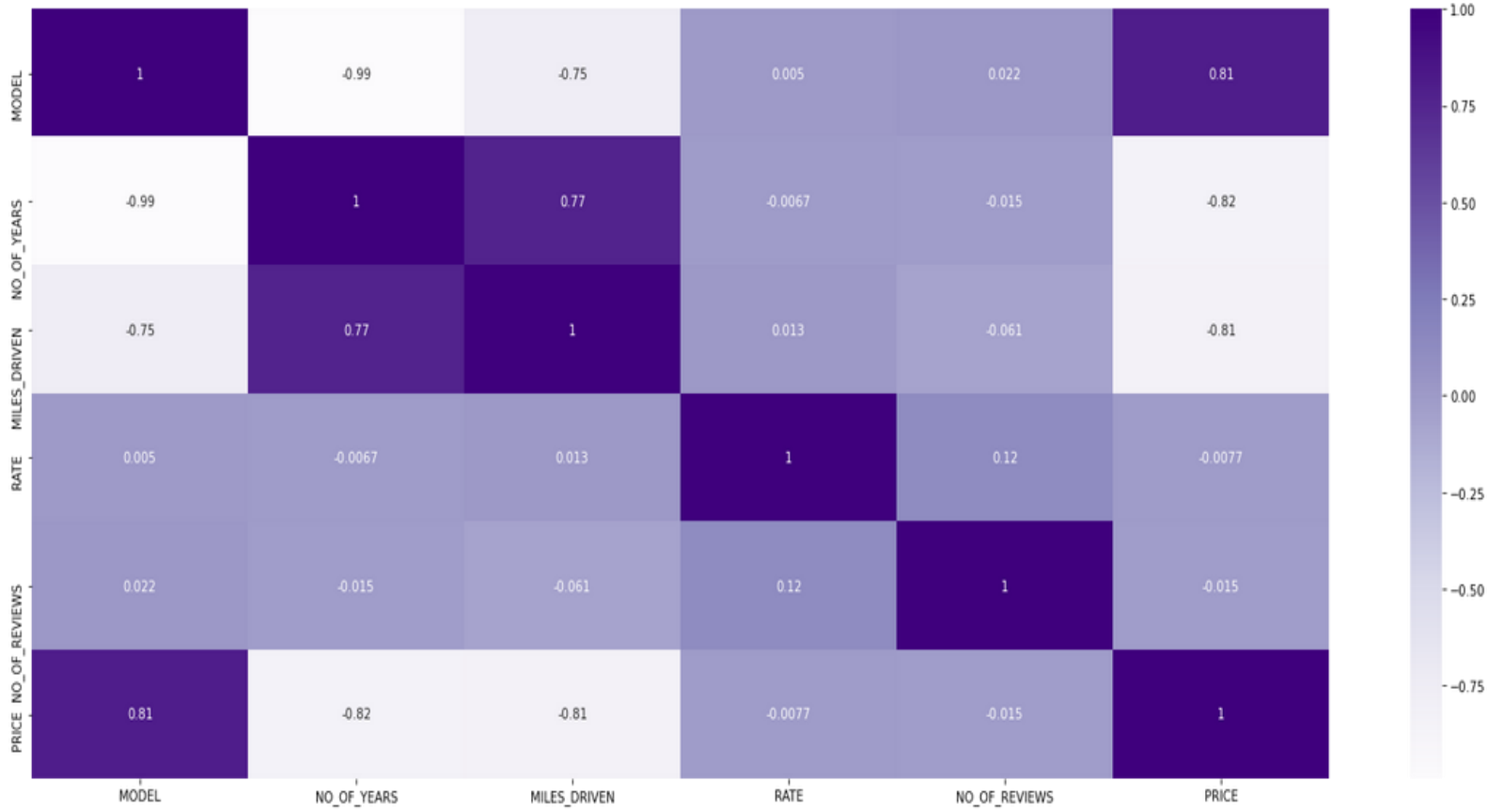
Check if there are any duplicate values.

4

Check if there are any outlier values.

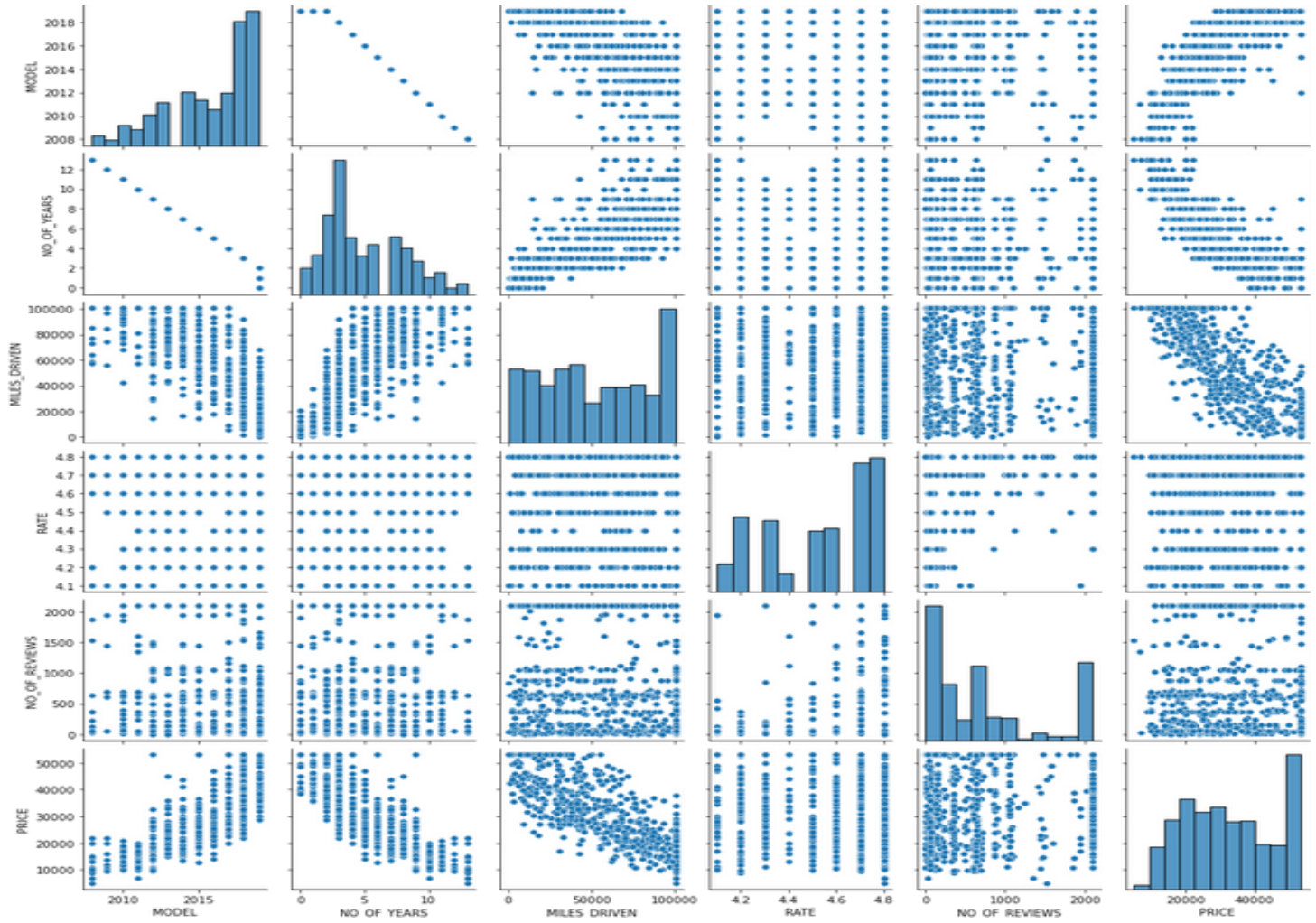
Visualizations

Heatmap Plot the correlations between the variables



Visualizations

Plot all the variable-to-variable relations as scatterplots



Feature Engineering

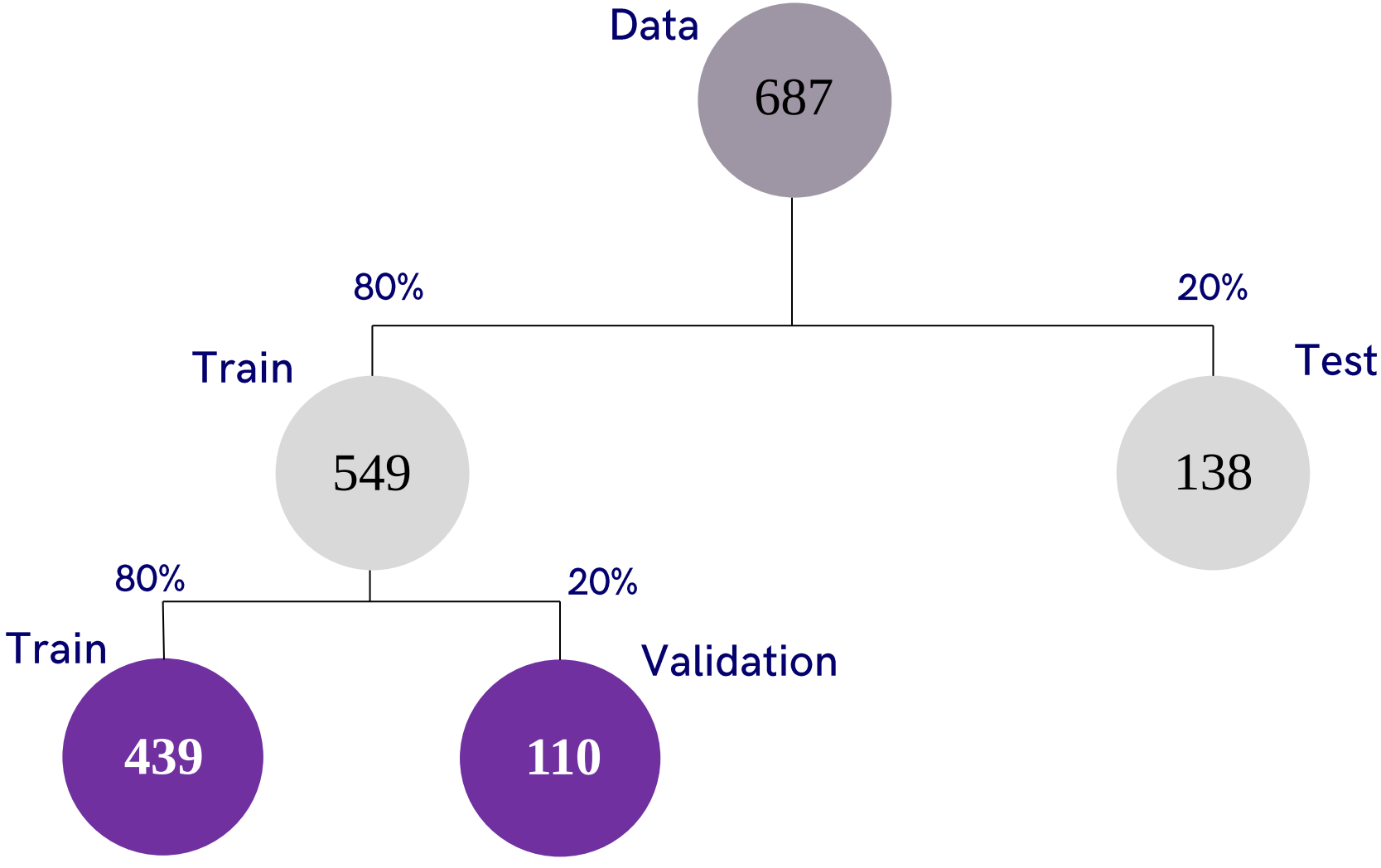
1

Creating Dummy Variables.

2

Create a new column (NO_OF_YEARS) from column MODEL.

Split Data


Data
687
80%
20%
Train
Test
549
138
80%
20%
Train
Validation
439
110

Regression Algorithms

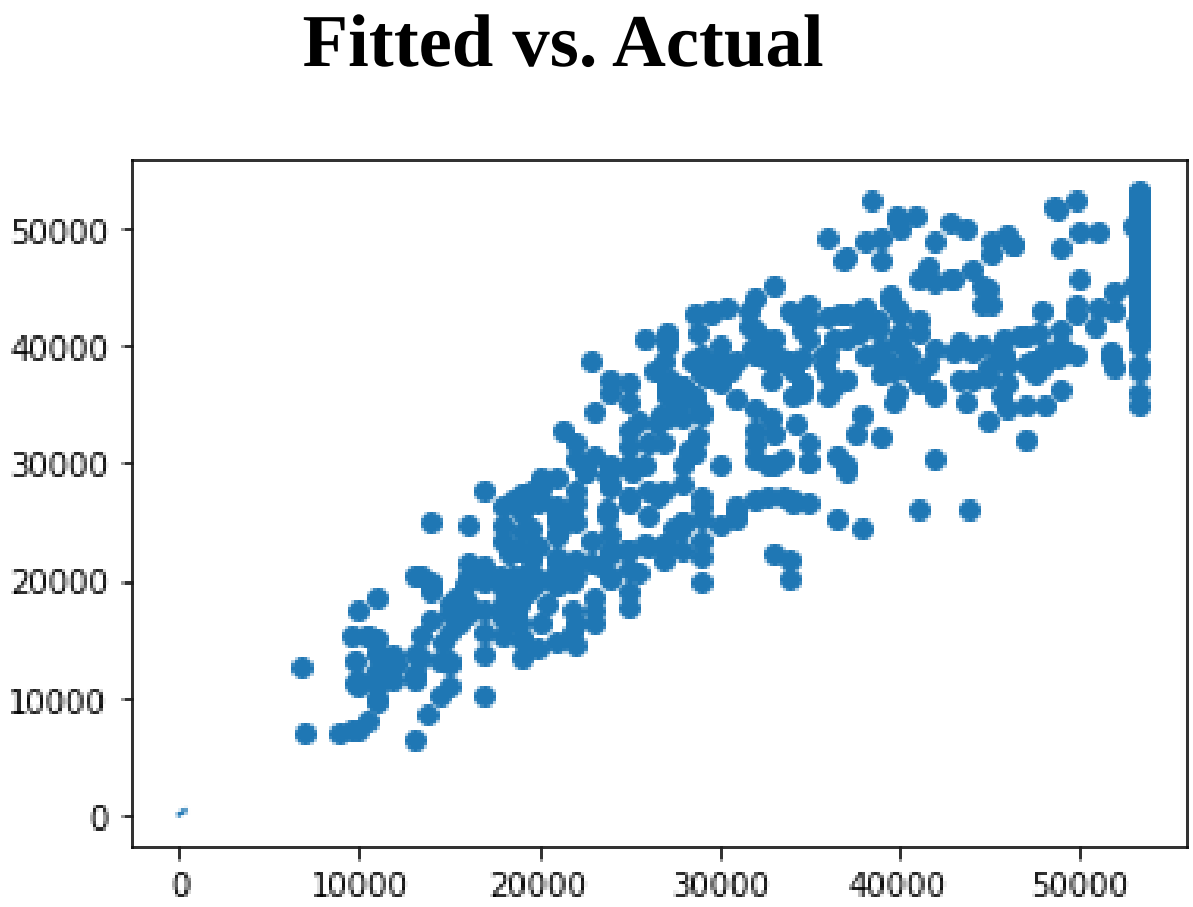
NO.	Regression Algorithms	Training Score	Validation Score
1	Simple Linear Regression	0.76114	0.74205
2	Polynomial	0.77348	0.73379
3	Ridge Regression	0.75584	**0.75785**
4	Lasso Regression	0.75584	0.75771
5	Cross Linear Regression	0.75584	0.75770

Best Model

Ridge Regression

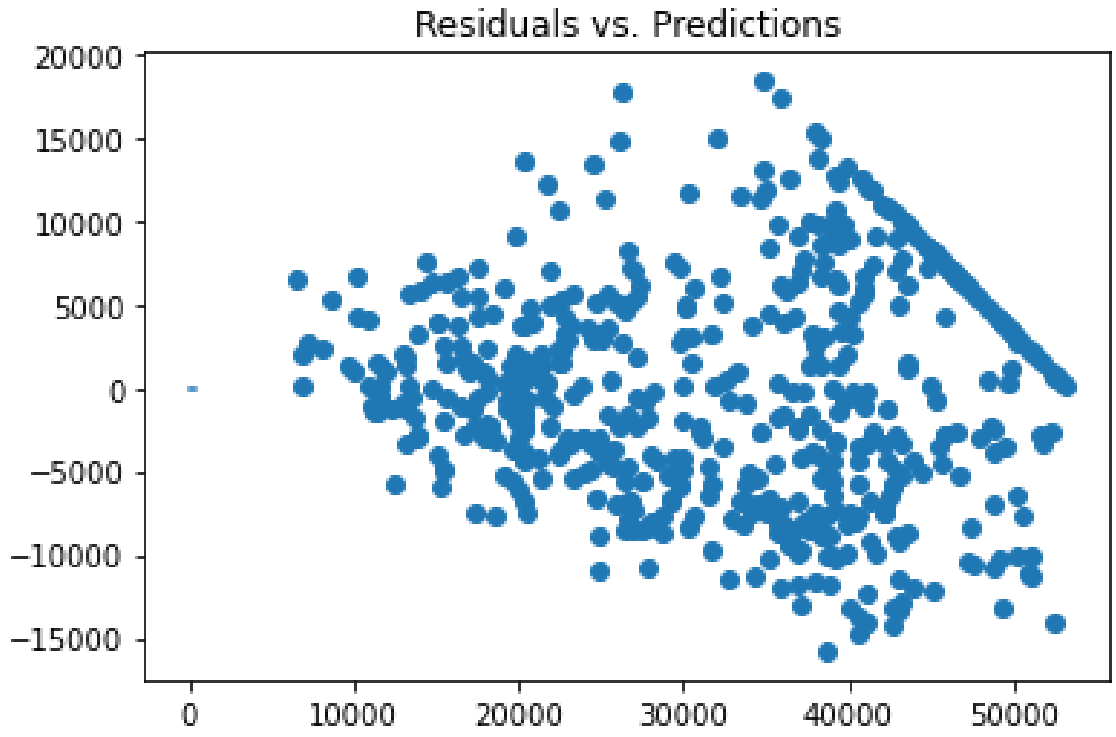
Training Score	Validation Score	Testing Score
0.75584	0.75785	0.77345

Evaluating Ridge Regression Model



Evaluating Ridge Regression Model

Plot Residuals vs. predicted



Conclusion

Ridge Regression has the best results

Evaluation (Validation)

SSE	RMSE	MAE	R2
38609913.3	6213.6	5175.9	0.729

Thank you for listening