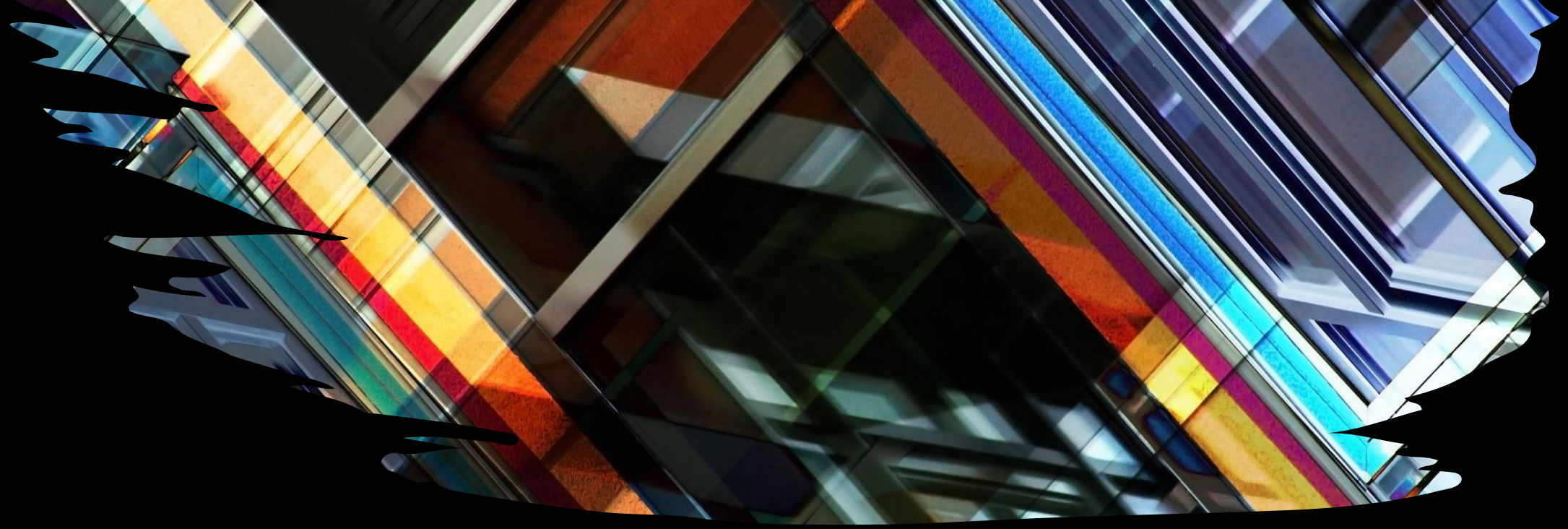


Predicting Used Car Prices

SPRINGBOARD CAPSTONE PROJECT

RYAN PAIK

The Problem

As a buyer, it is difficult to know what features of the vehicles and how they affect the price of a used vehicle. The focus is to predict the price of a used vehicle based on their features.

JUSTIN SULLIVAN/GETTY IMAGES

Who does this benefit?

Used Car Dealers

Individual Car buyers and sellers

Car Features Used to Predict Price

- MSRP
- Listed Price
- Year
- Condition
- Cylinders
- Fuel
- Odometer
- Transmission
- Drive
- Size
- Type

Data acquired from Kaggle

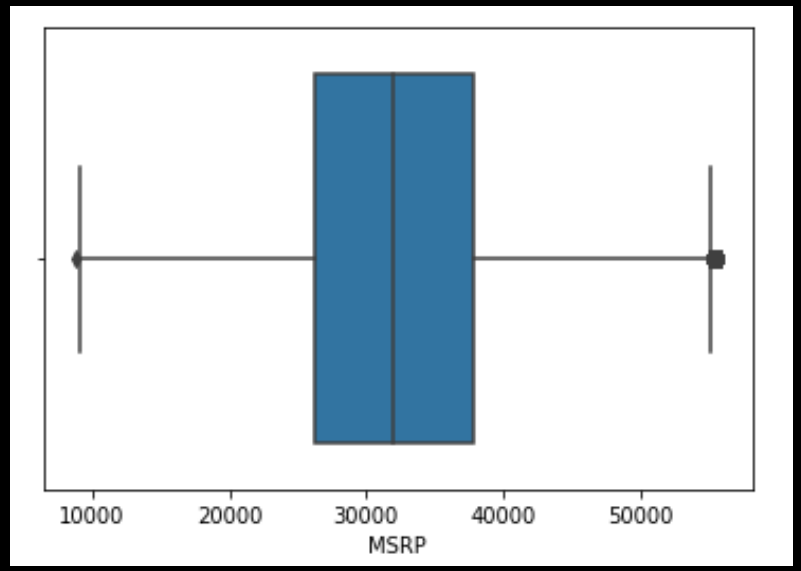
Exploratory Data Analysis

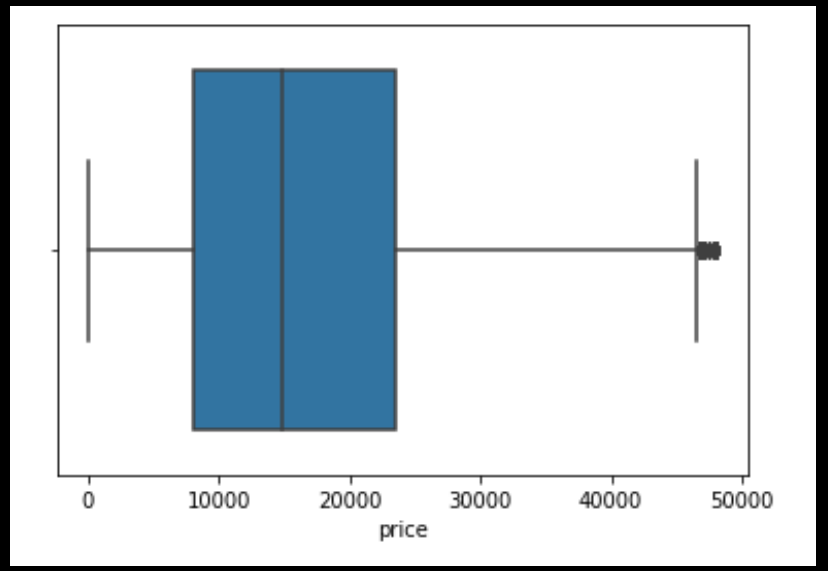
Merged Dataset

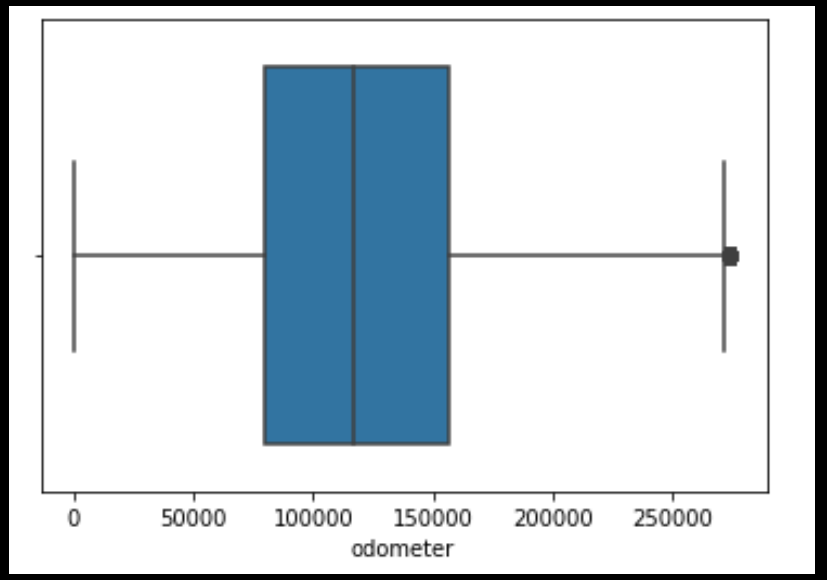
This dataset is the combination of 2 dataset: New car prices and used car prices

	Model	MSRP	price	year	condition	cylinders	fuel	odometer	transmission	drive	size	type
0	2019 acura rdx	40600	36250	2019	excellent	4 cylinders	gas	31250	automatic	4wd	mid-size	SUV
1	2019 acura rdx	45500	36250	2019	excellent	4 cylinders	gas	31250	automatic	4wd	mid-size	SUV
2	2019 acura rdx	43600	36250	2019	excellent	4 cylinders	gas	31250	automatic	4wd	mid-size	SUV
3	2019 acura rdx	37400	36250	2019	excellent	4 cylinders	gas	31250	automatic	4wd	mid-size	SUV
4	2019 acura rdx	42600	36250	2019	excellent	4 cylinders	gas	31250	automatic	4wd	mid-size	SUV
...	...	...	...	...	...	...	...	...	...	...	...	...
304065	2003 volvo xc90	35100	3500	2003	good	5 cylinders	gas	205000	automatic	4wd	full-size	SUV
304066	2003 volvo xc90	35100	2000	2003	fair	6 cylinders	gas	109000	automatic	4wd	mid-size	SUV
304067	1998 volvo s90	34300	2100	1998	good	6 cylinders	gas	144000	automatic	rwd	full-size	sedan
304068	1998 volvo s90	34300	2500	1998	good	6 cylinders	gas	160000	automatic	rwd	full-size	sedan
304069	1998 volvo s90	34300	3500	1998	good	6 cylinders	gas	170000	automatic	rwd	mid-size	sedan

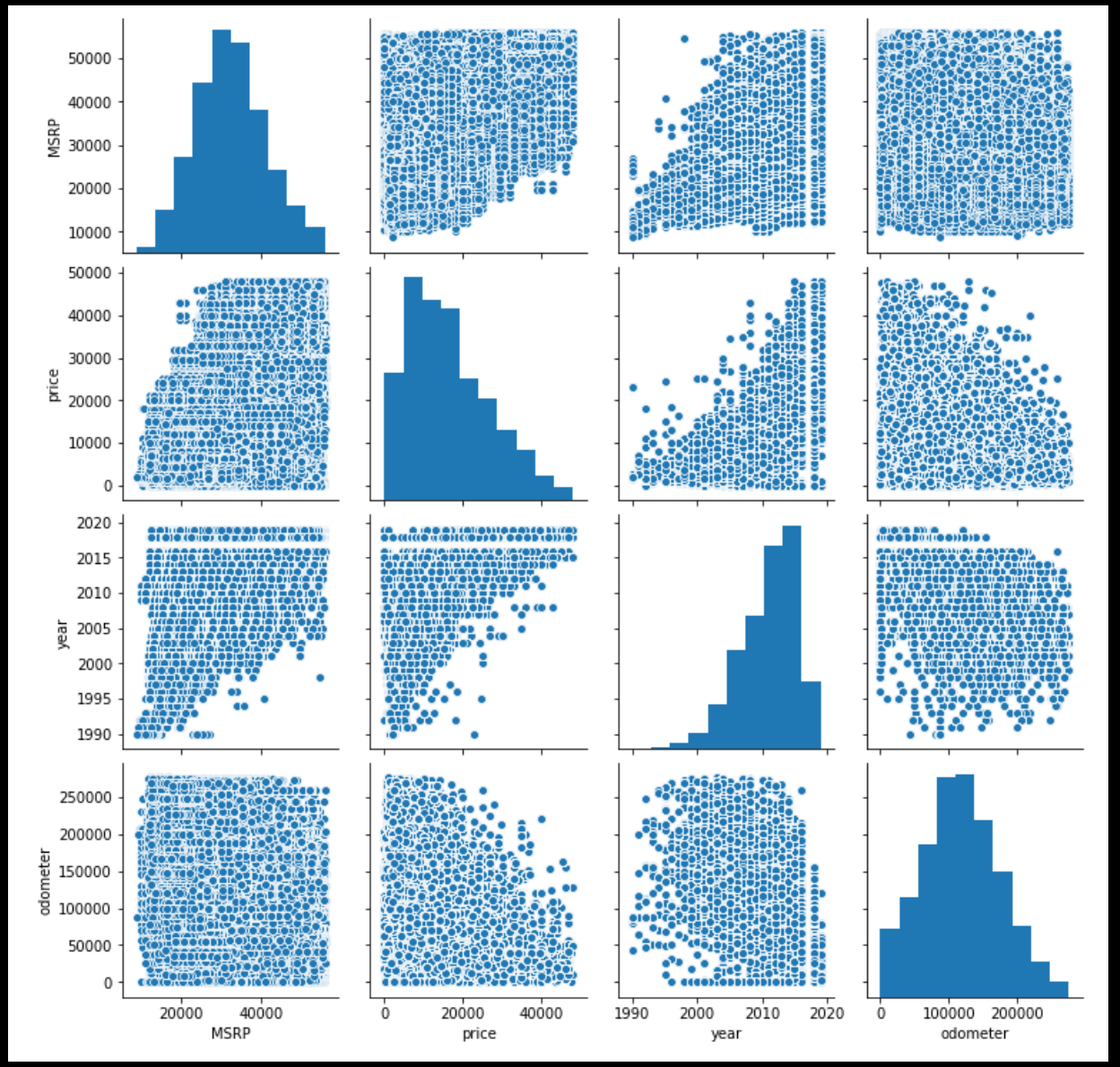
Outliers Removed





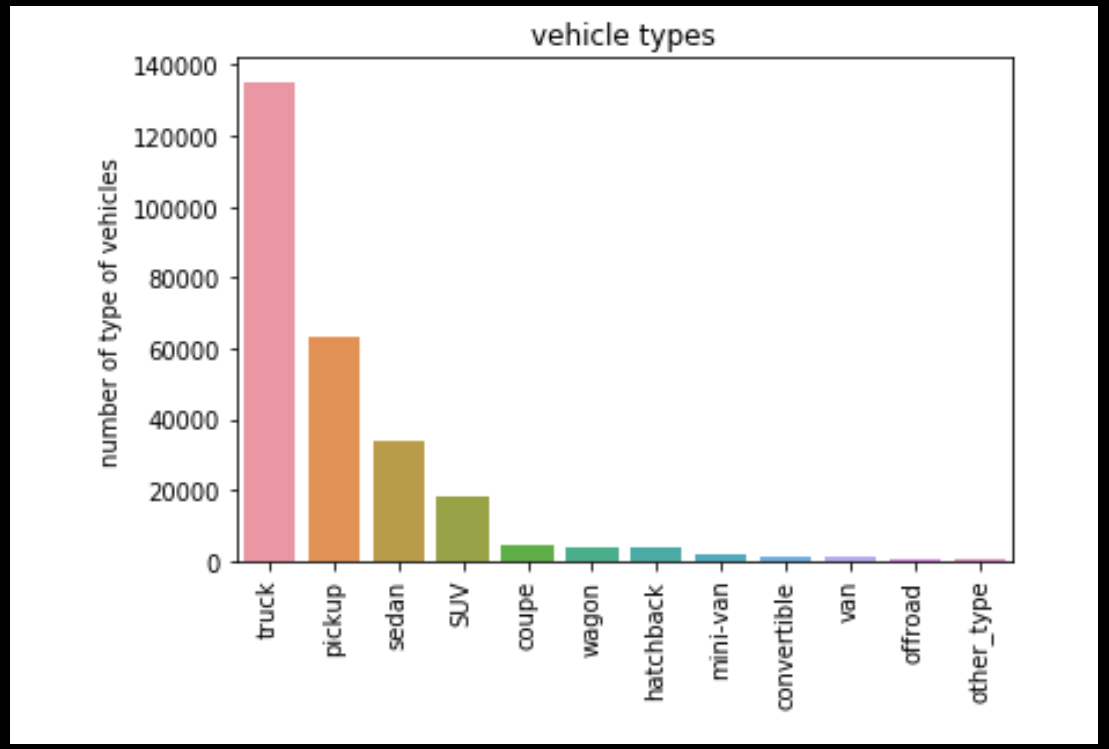


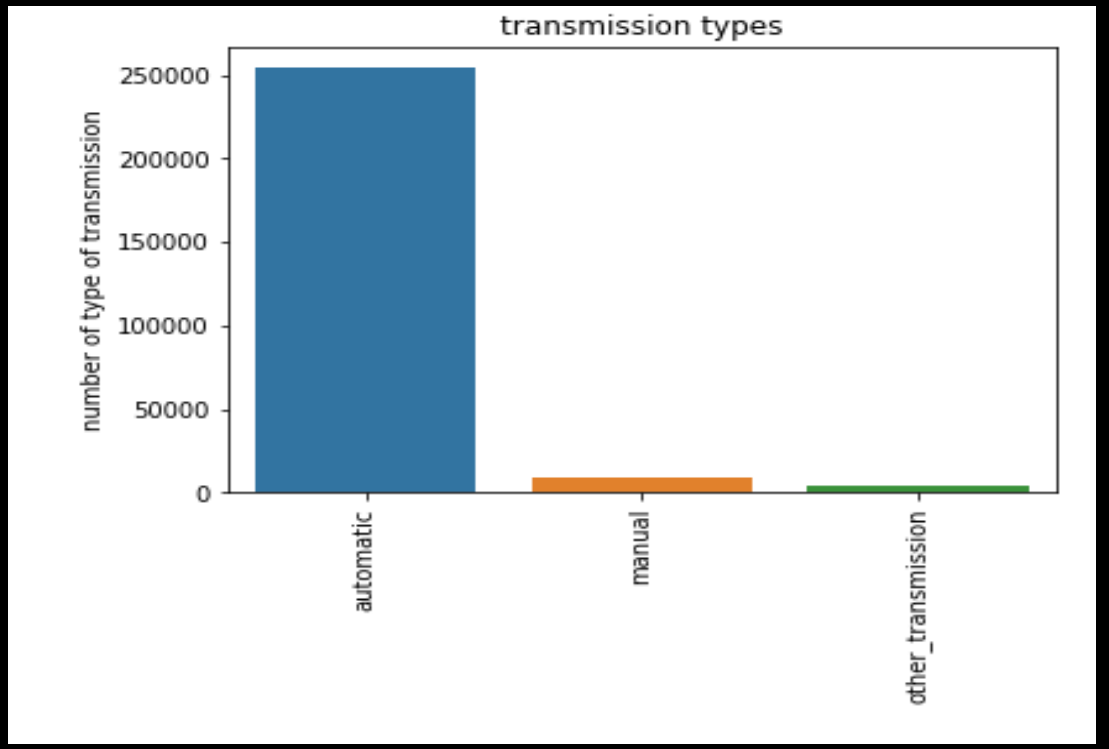
PairPlot

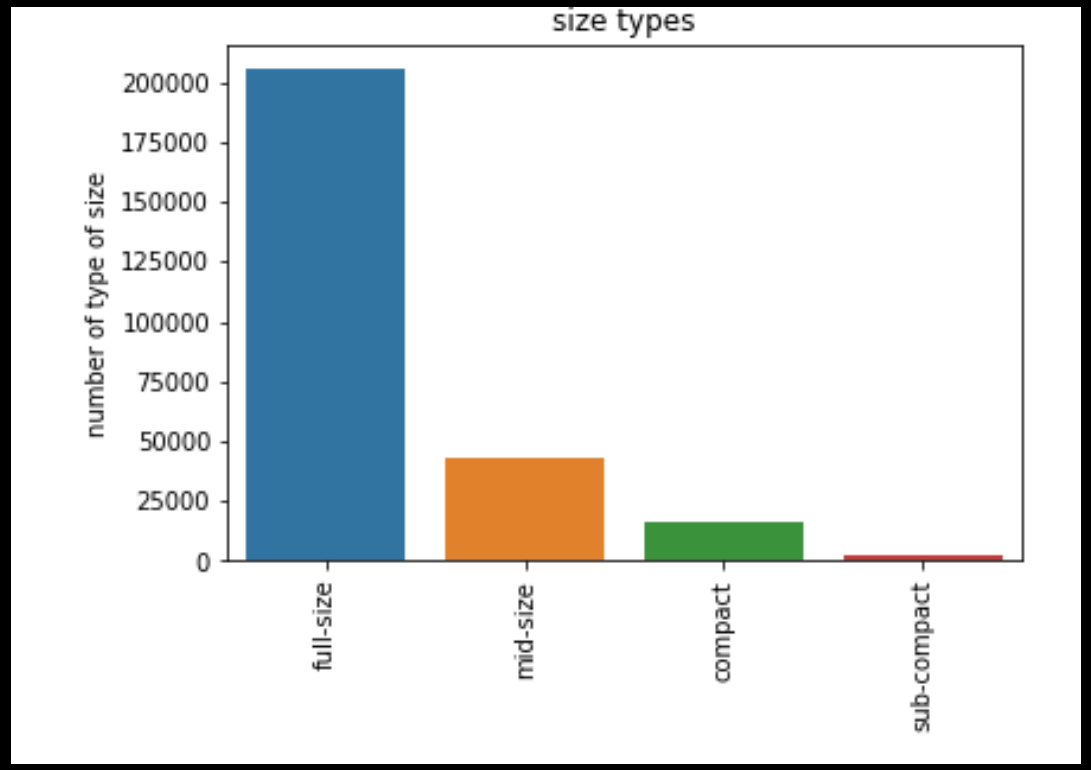

MSRP
price
year
odometer
50000
40000
30000
20000
10000
0
2020
2015
2010
2005
2000
1995
1990
250000
200000
150000
100000
50000
20000
40000
1990
2000
2010
2020
100000
200000

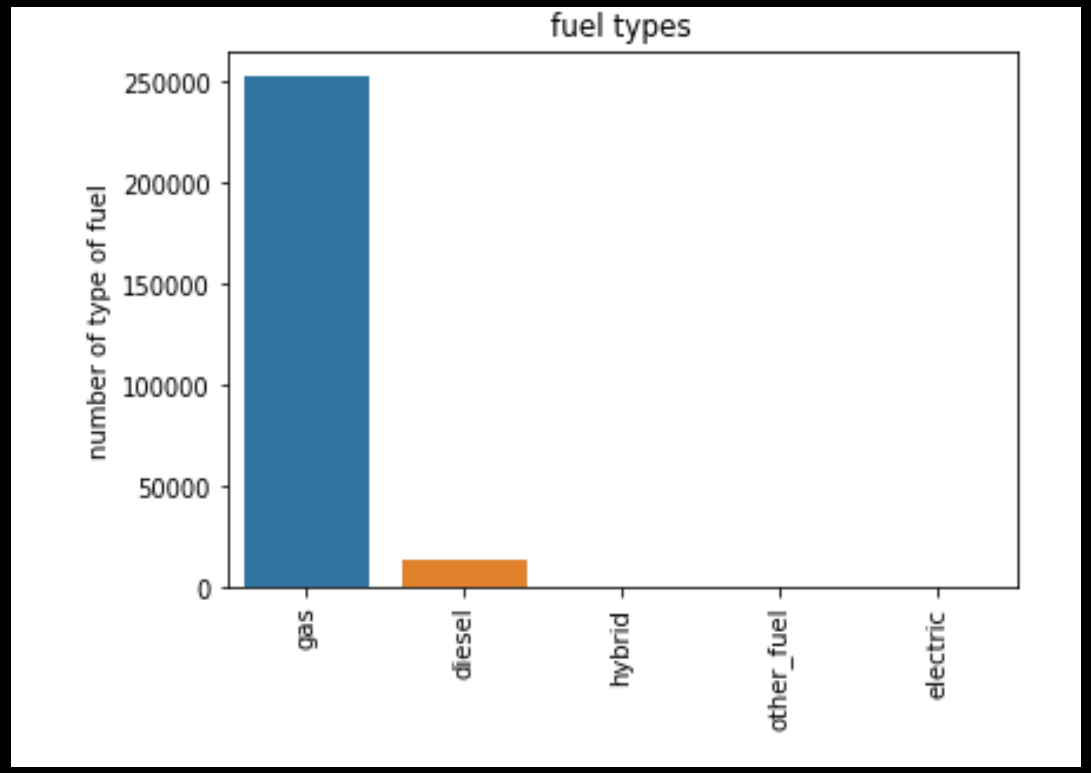
HeatMap

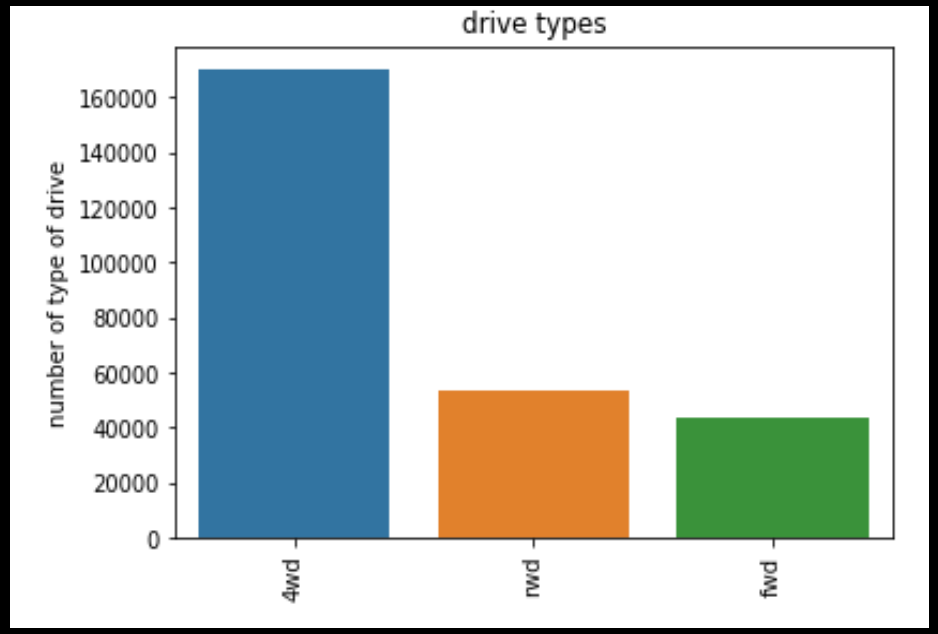

MSRP
price
year
odometer
1
0.54
0.47
-0.2
0.54
1
0.66
-0.5
0.47
0.66
1
-0.52
-0.2
-0.5
-0.52
1
MSRP
price
year
odometer
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

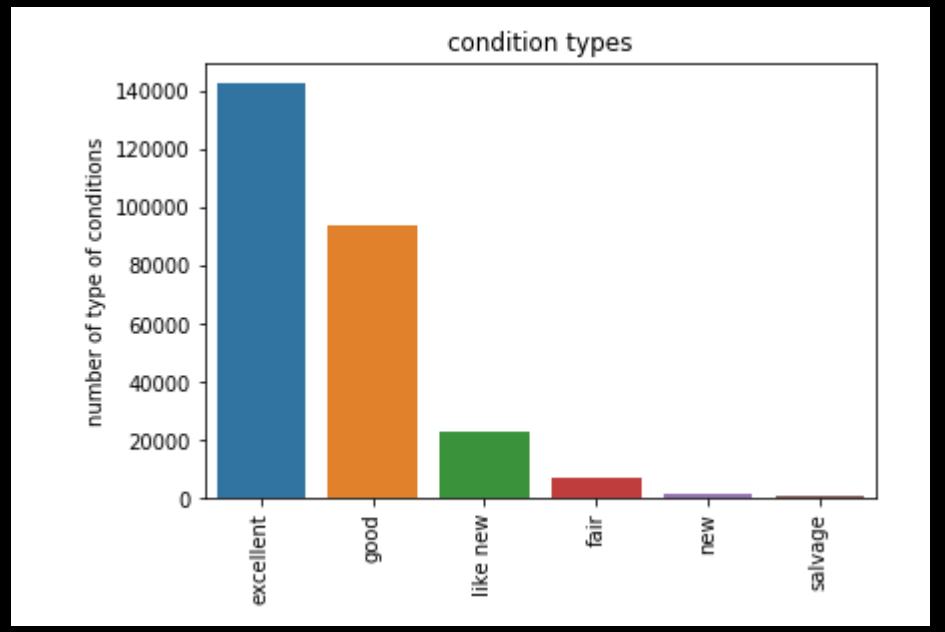

vehicle types
number of type of vehicles
140000
120000
100000
80000
60000
40000
20000
0
truck
pickup
sedan
SUV
coupe
wagon
hatchback
mini-van
convertible
van
offroad
other_type

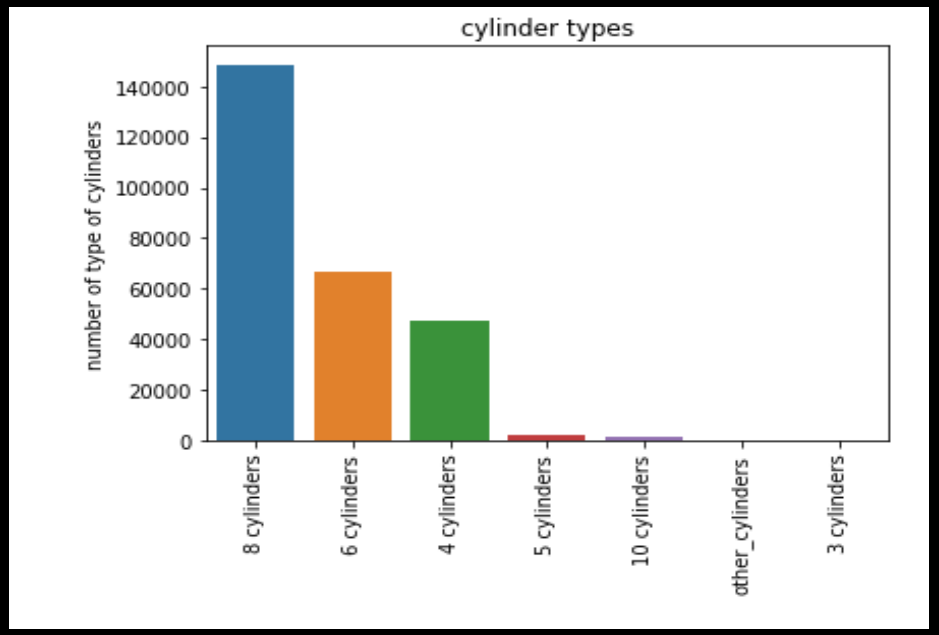

transmission types
number of type of transmission
250000
200000
150000
100000
50000
0
automatic
manual
other_transmission


size types
number of type of size
200000
175000
150000
125000
100000
75000
50000
25000
0
full-size
mid-size
compact
sub-compact


fuel types
number of type of fuel
250000
200000
150000
100000
50000
0
gas
diesel
hybrid
other_fuel
electric


drive types
number of type of drive
160000
140000
120000
100000
80000
60000
40000
20000
0
4wd
rwd
fwd


condition types
number of type of conditions
140000
120000
100000
80000
60000
40000
20000
0
excellent
good
like new
fair
new
salvage


cylinder types
number of type of cylinders
140000
120000
100000
80000
60000
40000
20000
0
8 cylinders
6 cylinders
4 cylinders
5 cylinders
10 cylinders
other_cylinders
3 cylinders

Machine Learning Models

Type: Supervised Learning – Regression

Linear Regression, Decision Tree,
Random Forest, Gradient Boosting

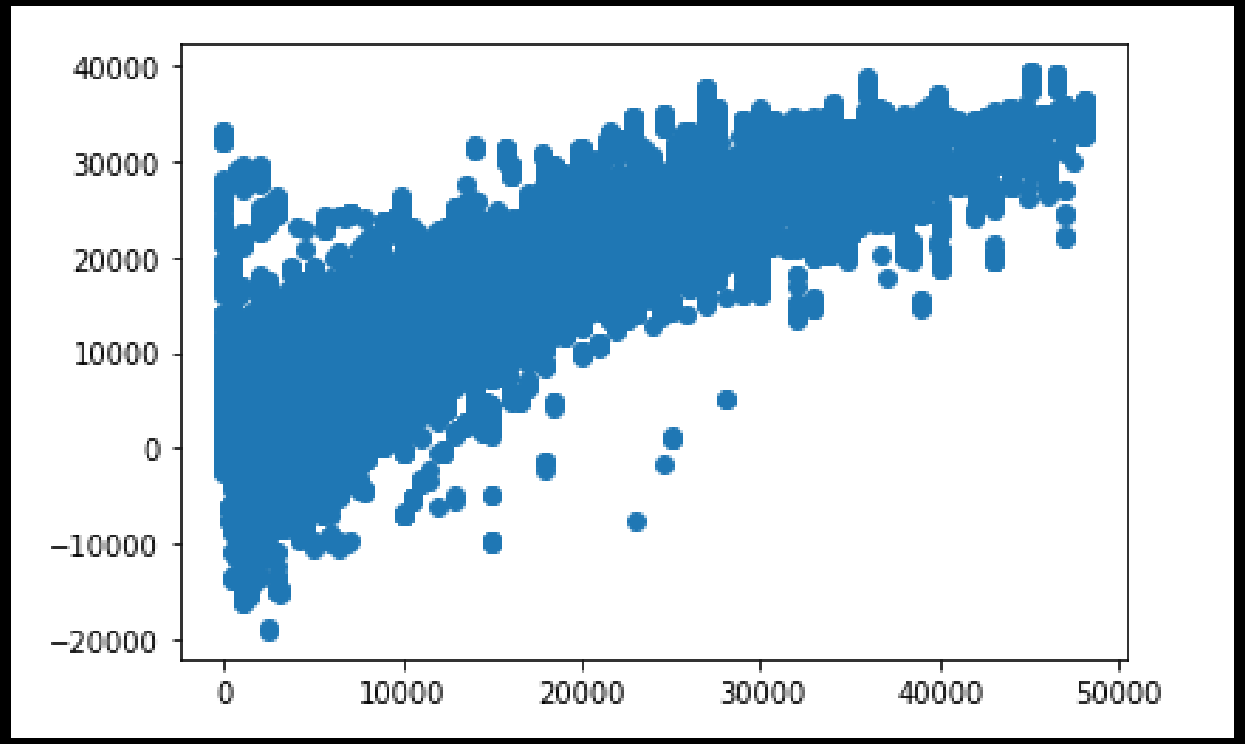
Standard Scaling

GridSearchCV

Linear Regression

Mean absolute error: 3910.62
Mean Squared Error: 27052164.51
Median Absolute Error: 3074.9
R2 score: 0.74

Average CV score: 0.74



Decision Tree

Mean absolute error: 89.03
Mean Squared Error: 740457.4
Median Absolute Error: 0.0
R2 score: 0.99

Average CV score: 0.99144

Hyperparameters:

Max depth: 10
Max features: 5

Train Score: 0.76
Test Score: 0.76

Average Validation Score: 0.728

Random Forest

Mean absolute error: 107.16
Mean Squared Error: 615529.52
Median Absolute Error: 0.0
R2 score: 0.99

Average CV score: 0.99297

Hyperparameters:

Max depth: 10
N estimators: 200

Train Score: 0.88
Test Score: 0.88

Average Validation Score: 0.884

Gradient Boosting

Mean absolute error: 3017.35
Mean Squared Error: 17363207.08
Median Absolute Error: 2272.67
R2 score: 0.84

Average CV score: 0.84

Hyperparameters:

Max depth: 10
N estimators: 300

Train Score: 0.99
Test Score: 0.99

Average Validation Score: 0.987

Conclusions

Model: Gradient Boosting

Future Improvements

- Improve vehicle feature grading system to be objective, rather than being subjective.
- Create a preferential input model for buyers