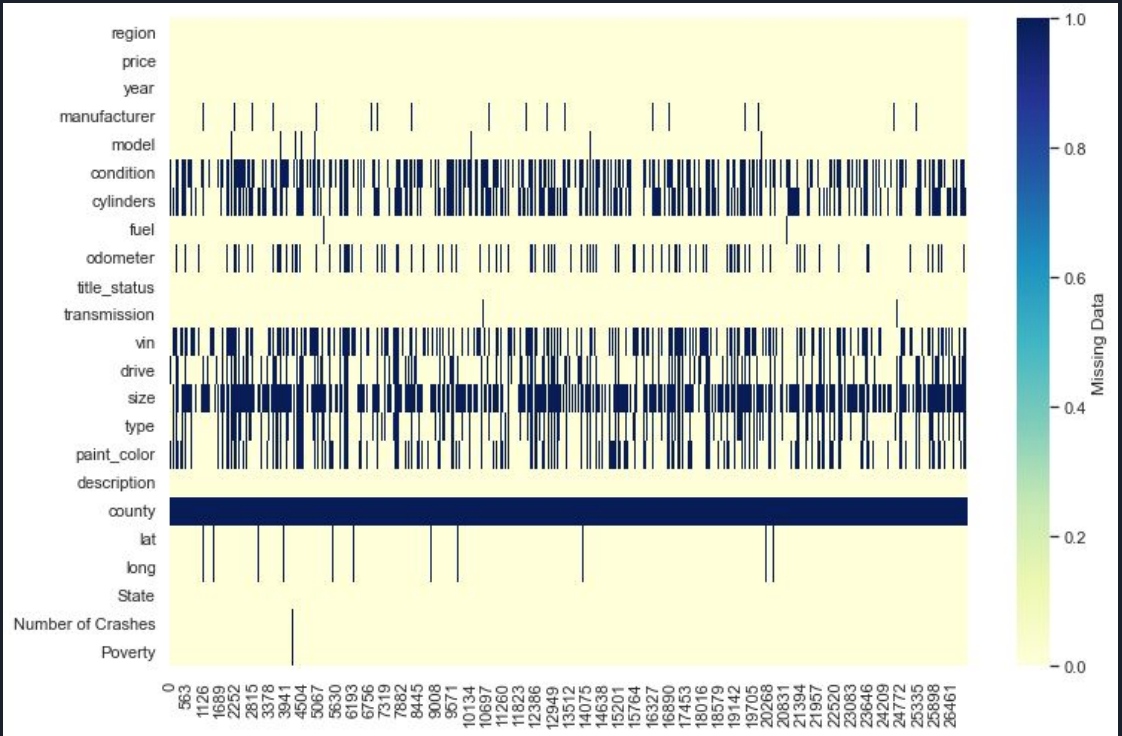
Car Price Prediction Multiple Linear Regression

Juan Felipe Latorre Gil
jflatorreg@unal.edu.co

Exploratory Data Analysis

- Null Data



- 23 variables
- 18 categorical variables
- 5 numeric variable

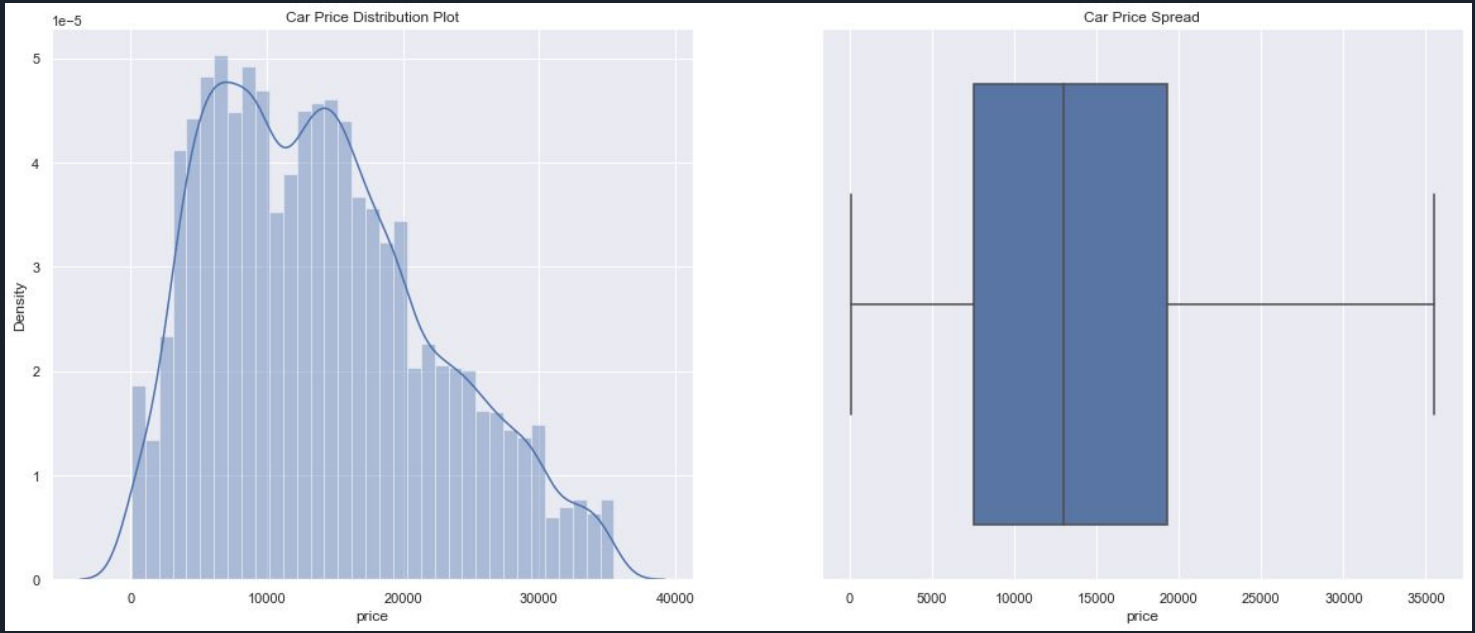
- vin and county are not statistically relevant
- size is the least represented variable

Exploratory Data Analysis

Numeric Data

Exploratory Data Analysis

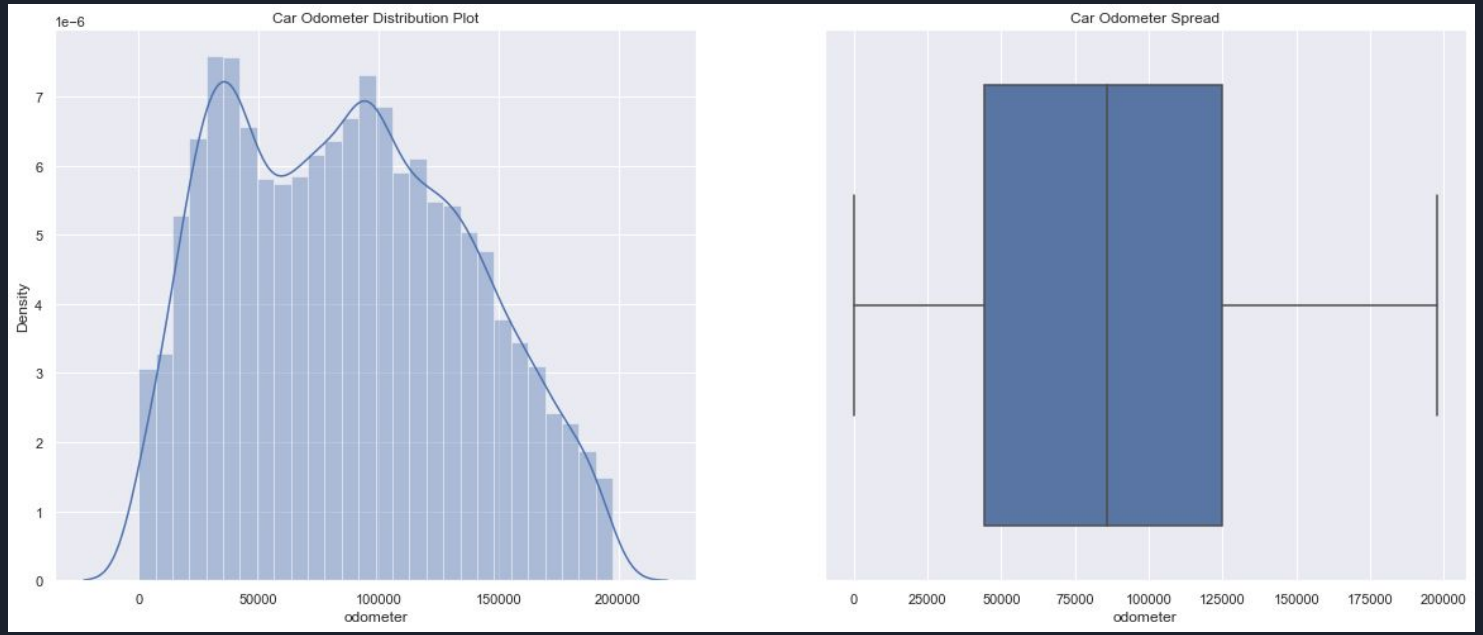
- Price - Response Variable



- Range less than the 95% percentile and greater than 50.

Exploratory Data Analysis

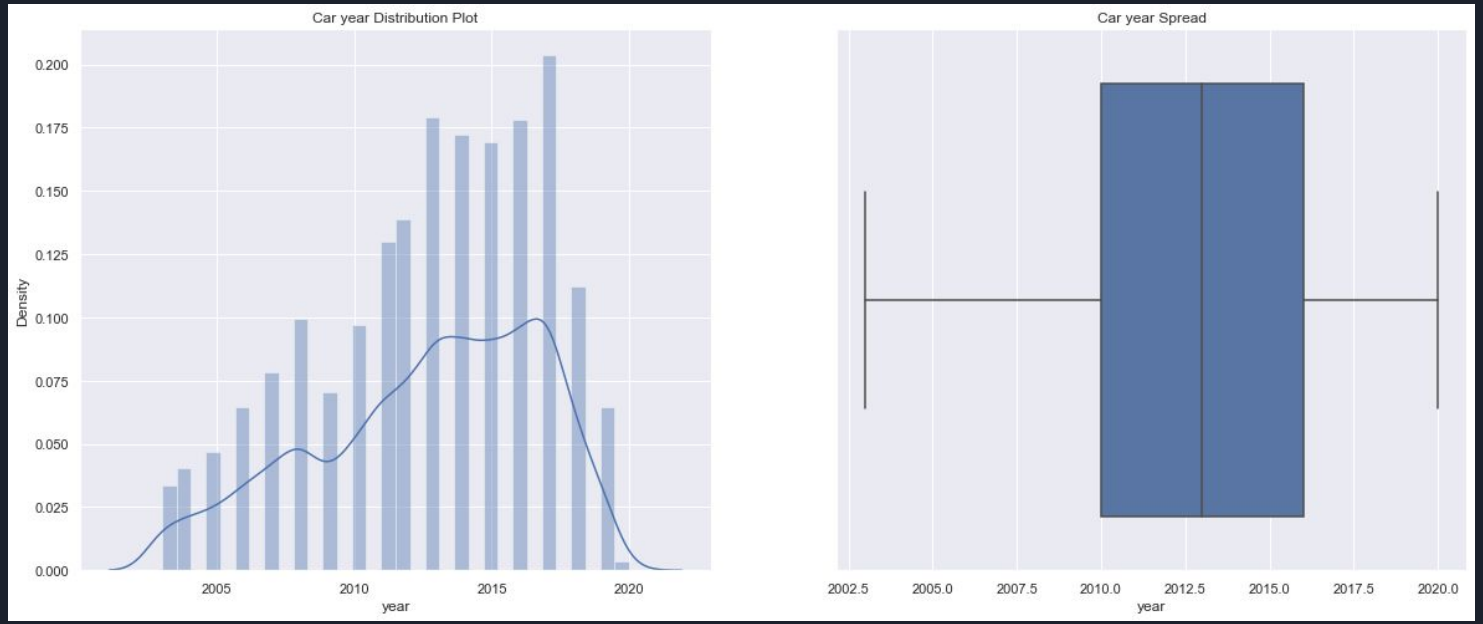
- Odometer



- Range less than the 95% percentile.

Exploratory Data Analysis

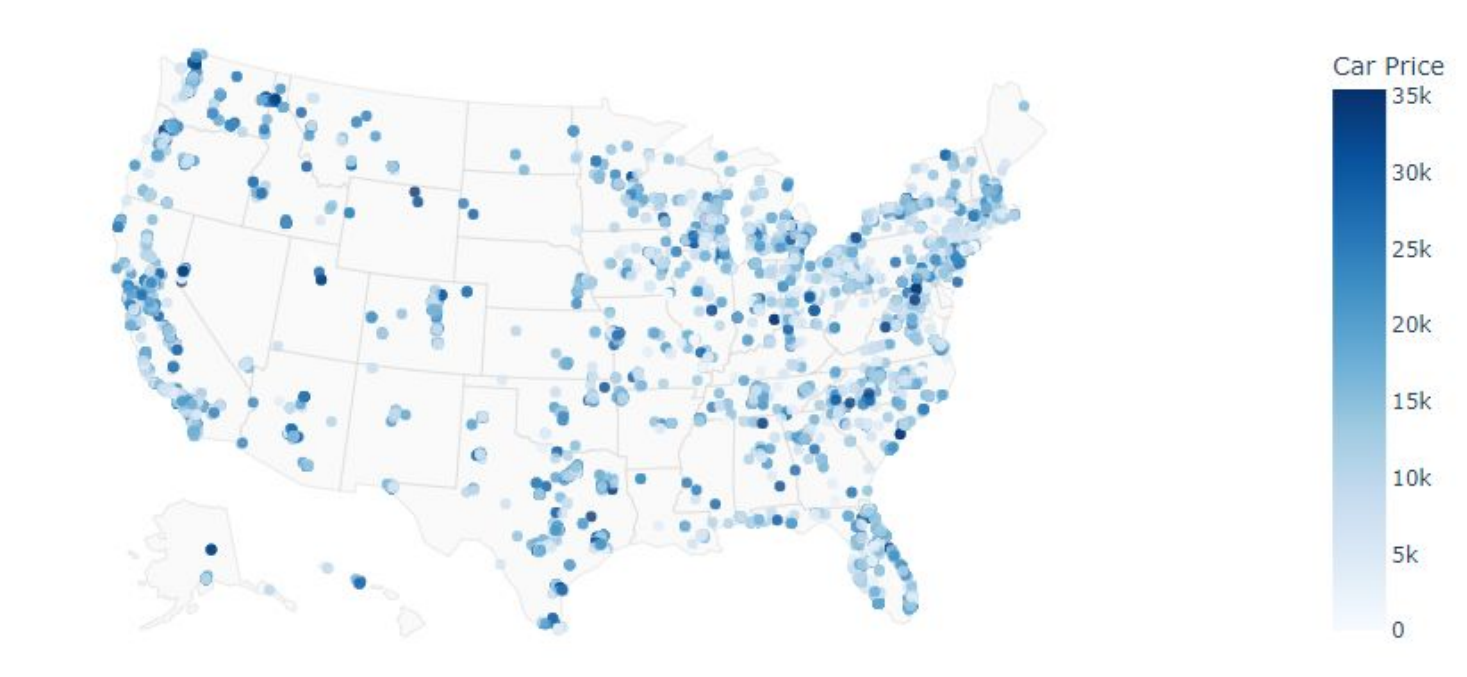
- Year



- Range greater than the 5% percentile.

Exploratory Data Analysis

- Latitude and Longitude

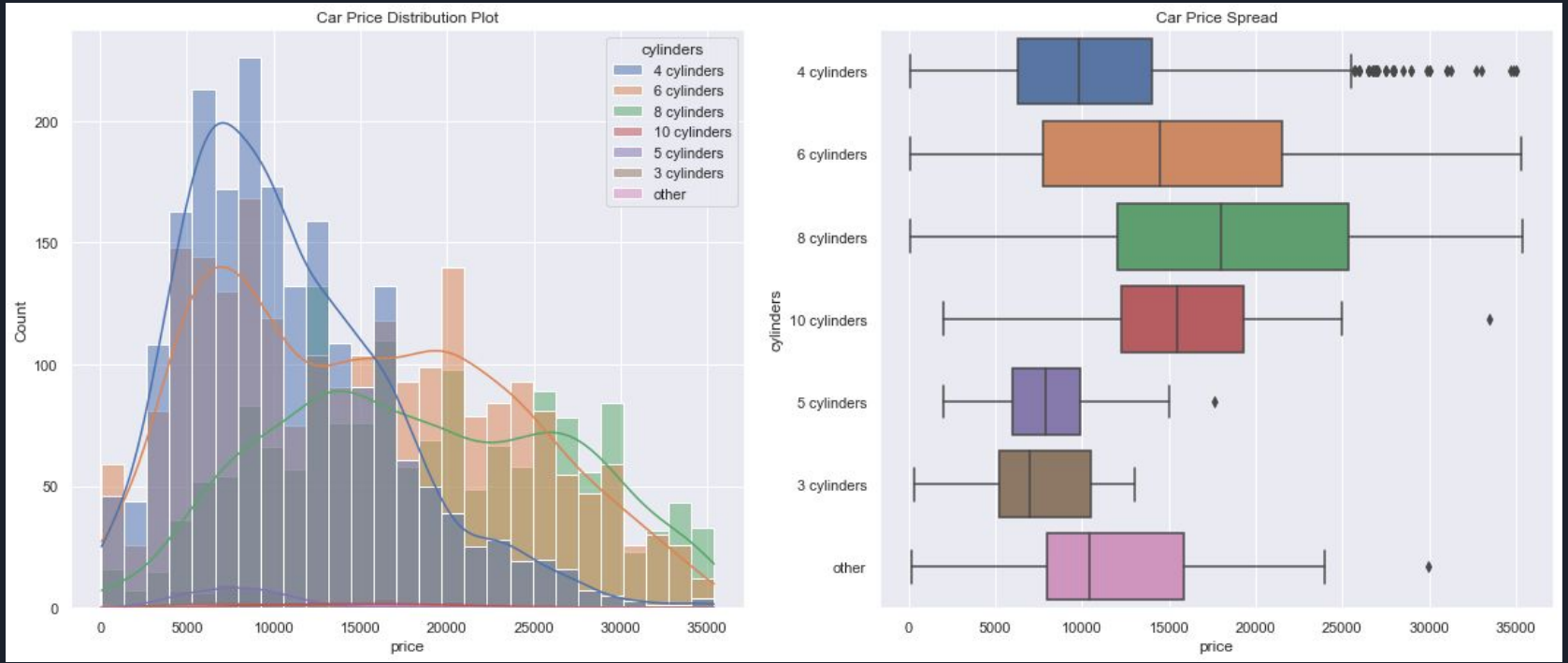


Exploratory Data Analysis

Categorical Data

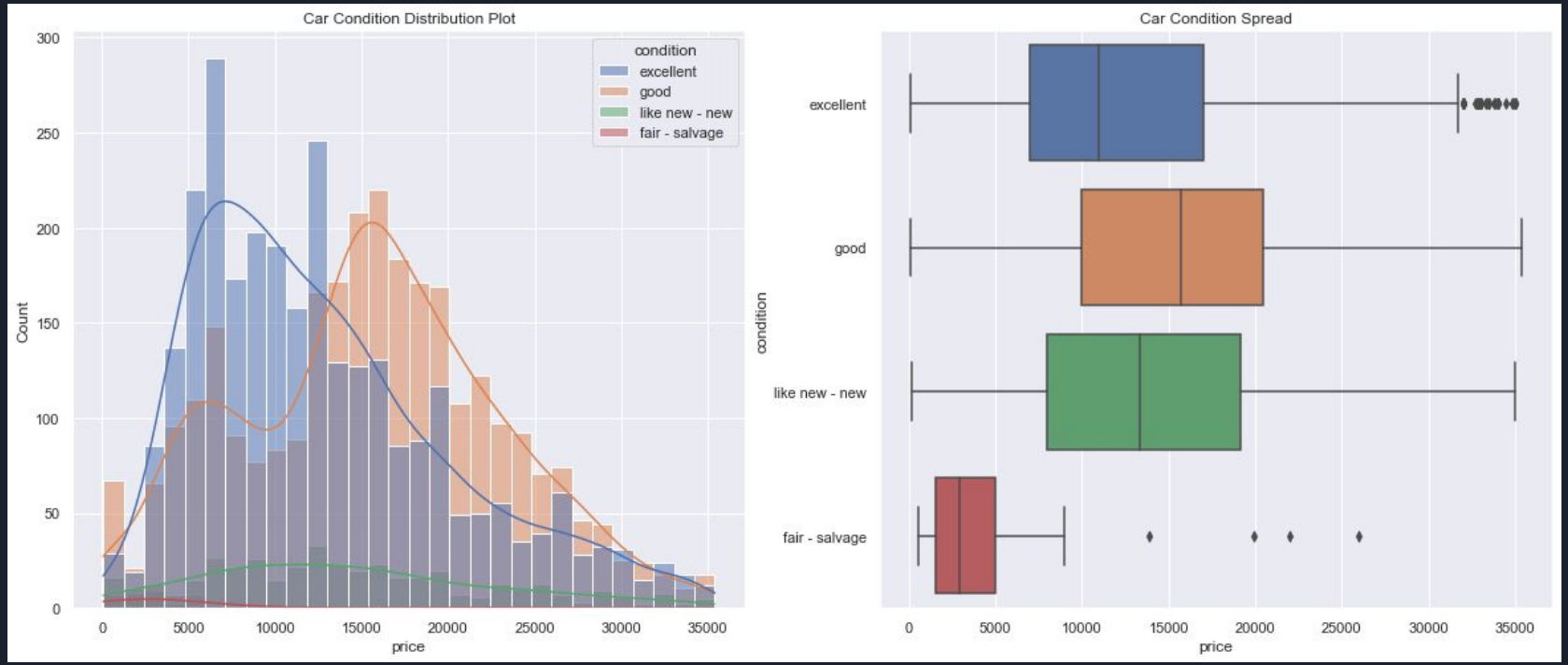
Exploratory Data Analysis

- Cylinders



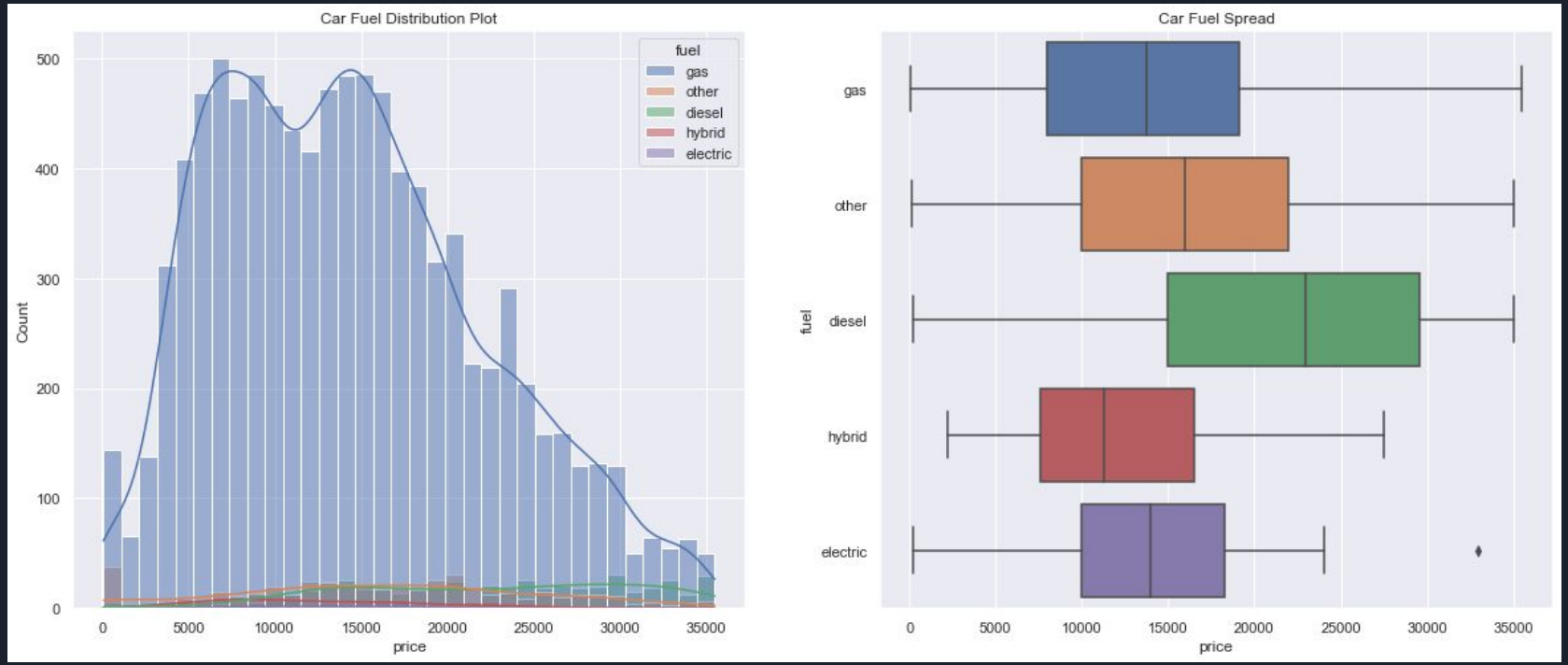
Exploratory Data Analysis

- Condition



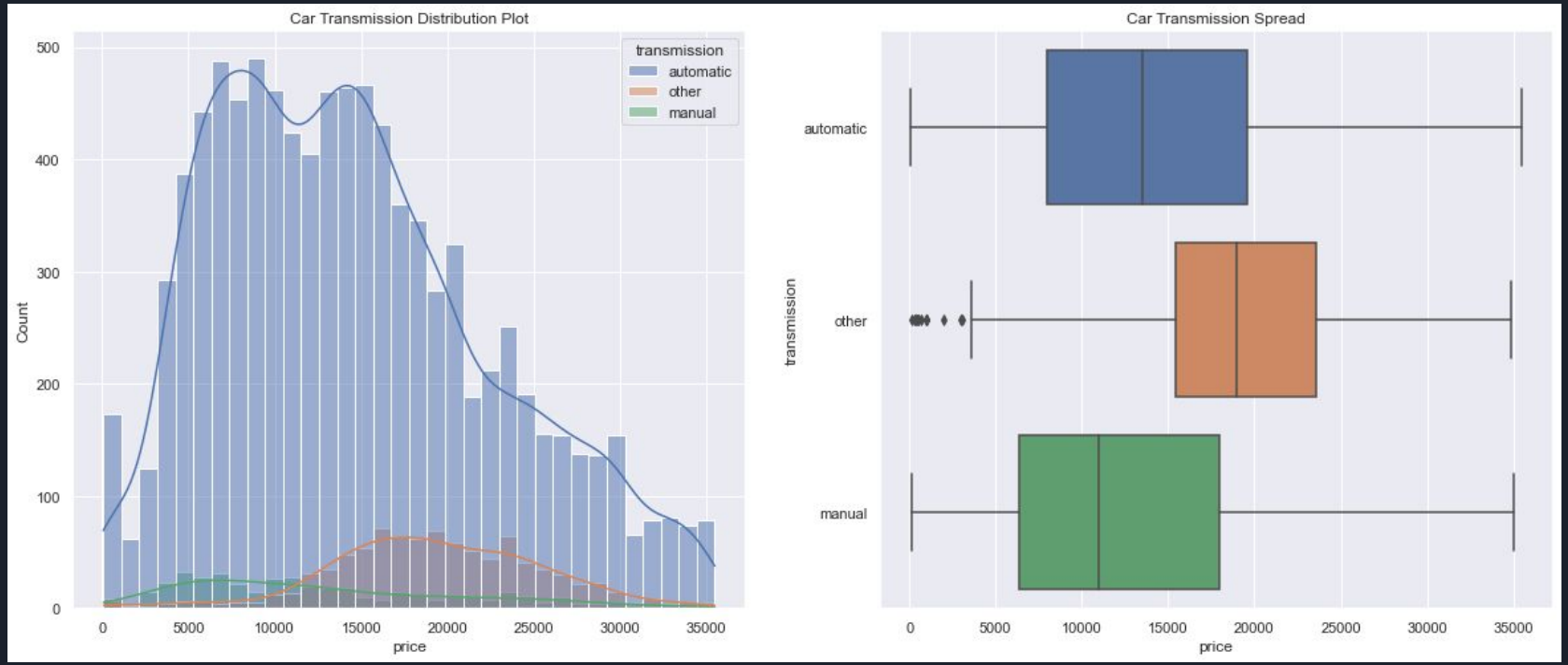
Exploratory Data Analysis

- Fuel



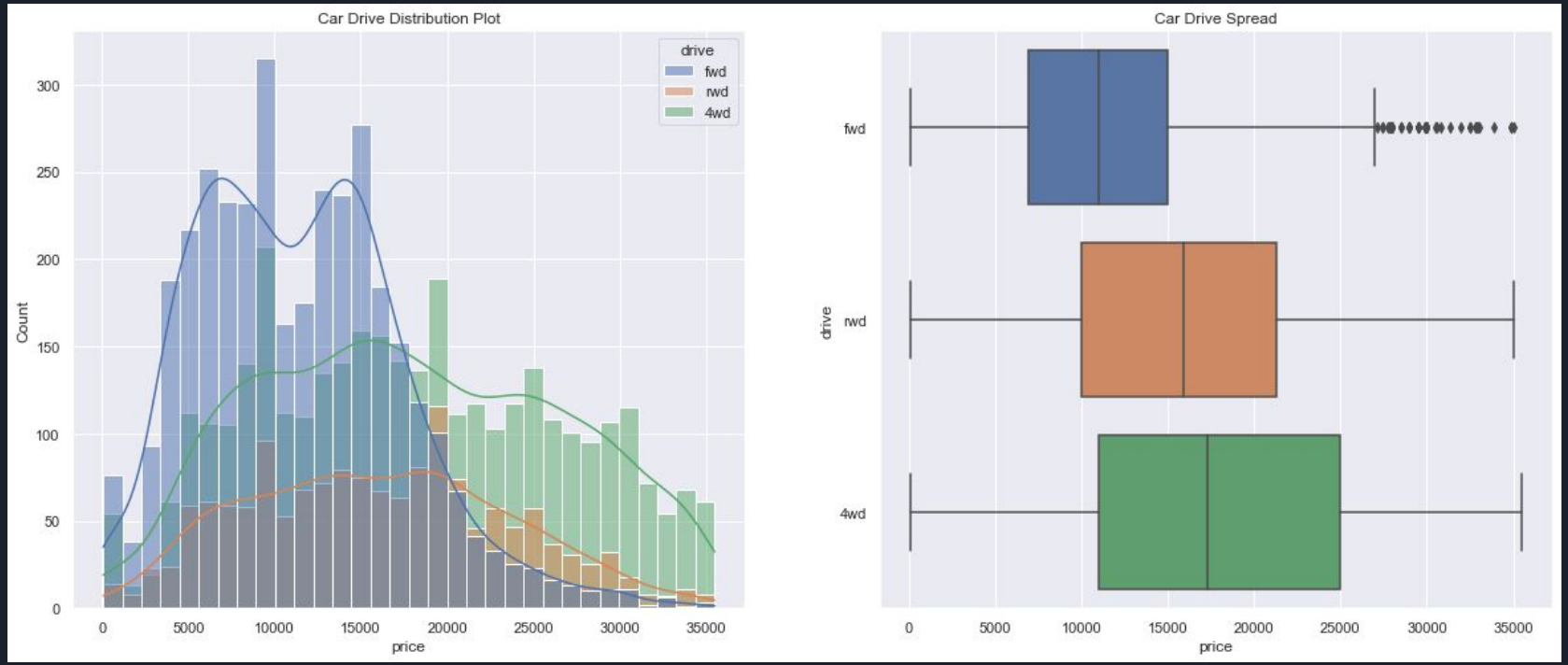
Exploratory Data Analysis

- Transmission



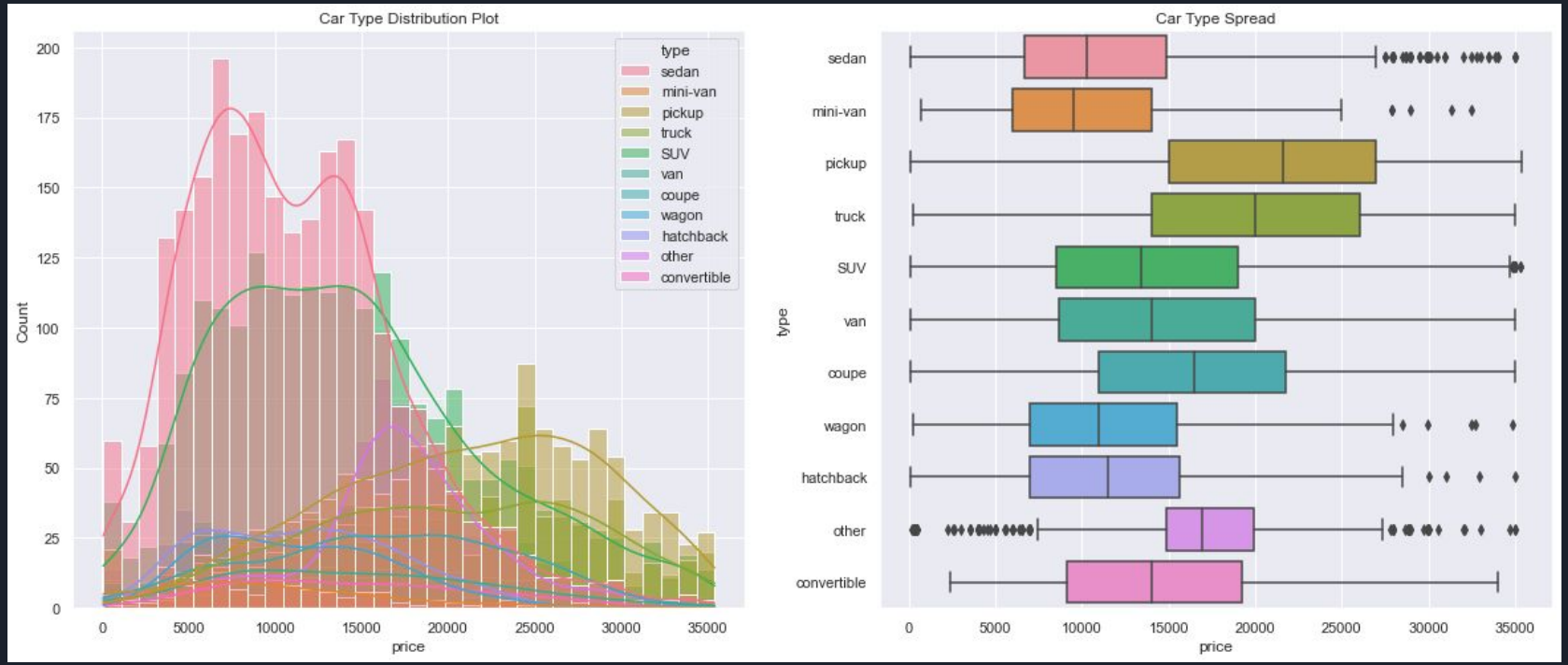
Exploratory Data Analysis

- Drive



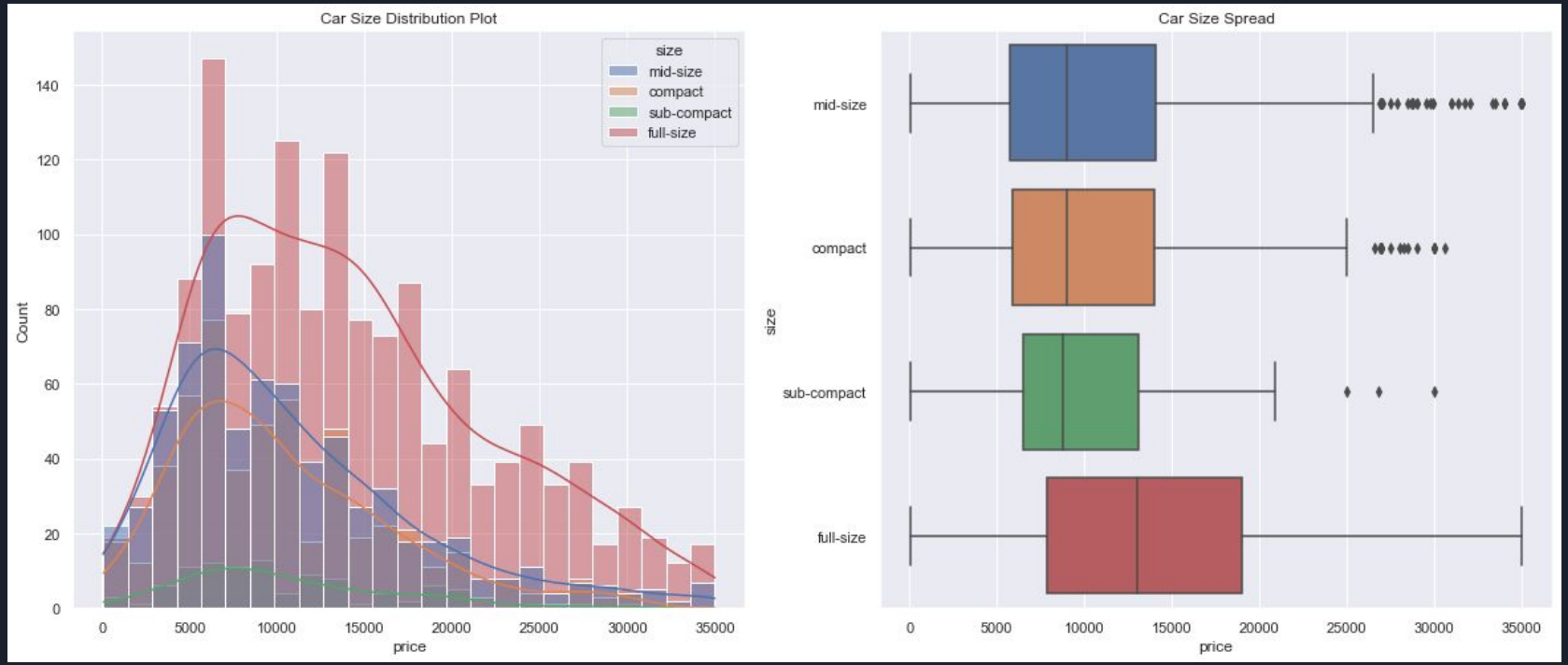
Exploratory Data Analysis

- Type



Exploratory Data Analysis

- Size



Exploratory Data Analysis

- Region has 356 different non-normalized values.
- Manufacturer has 37 different non-normalized values.
- Model has 37 different non-normalized values.
- Title Status has non-normalized values.
- Paint Color 12 different non-normalized values.

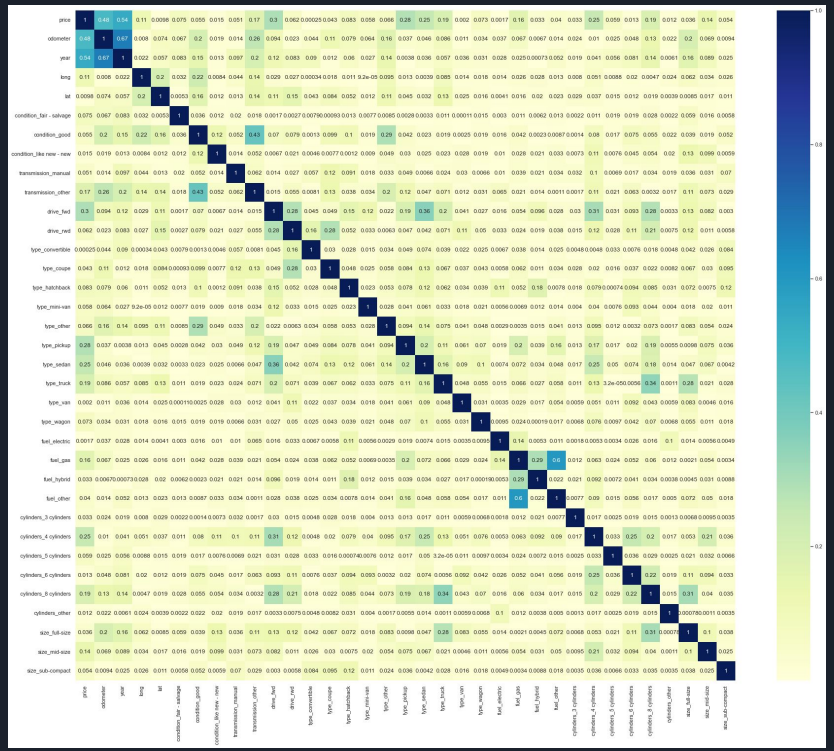
Model Variables:

- price
- odometer, year,
- long, lat,
- condition, transmission,
- drive, type,
- fuel, cylinders, size

11 explanatory variables

Exploratory Data Analysis

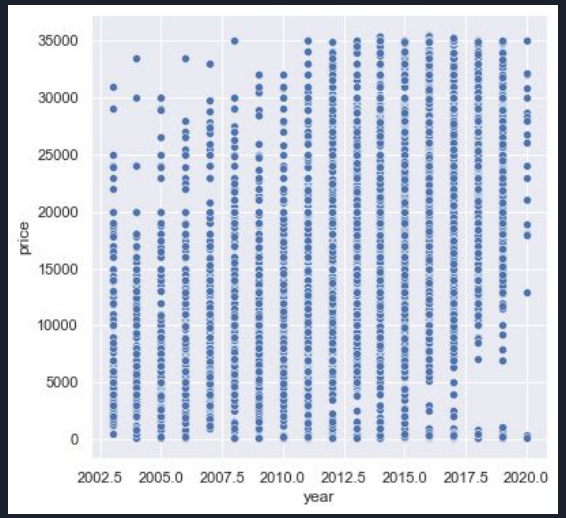
- Correlation

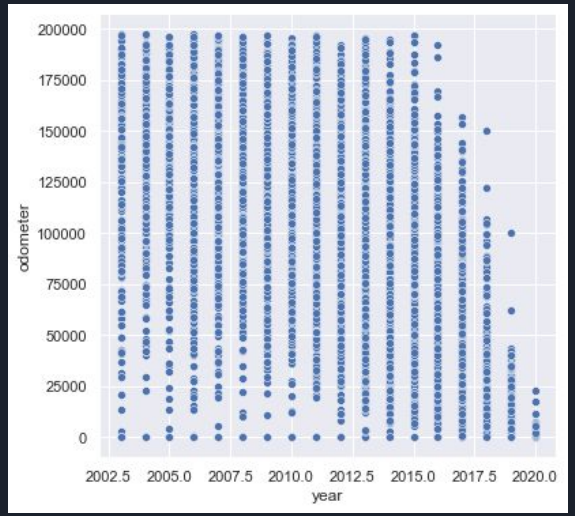


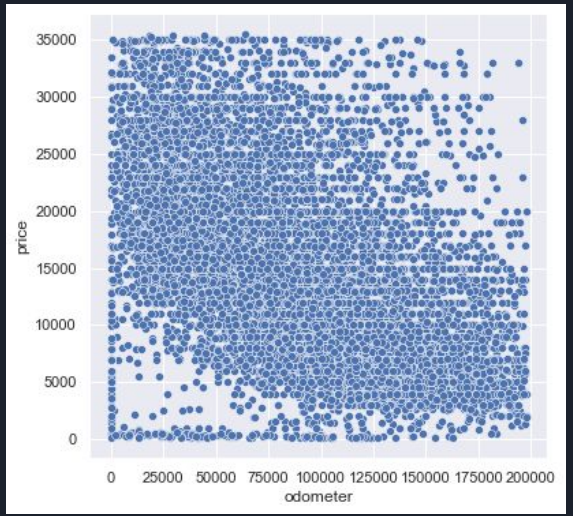
- Strong correlation between year and price, between odometer and price, also enrew year and odometer
- correlation between independent variables

Exploratory Data Analysis

- Correlation







Model Building

Results

Model Building

Results

Model Building

Linear regression

OLS Regression Results

Dep. Variable:	y	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.626
Method:	Least Squares	F-statistic:	473.3
Date:	Tue, 09 Aug 2022	Prob (F-statistic):	0.00
Time:	10:38:06	Log-Likelihood:	-95251.
No. Observations:	9612	AIC:	1.906e+05
Df Residuals:	9577	BIC:	1.908e+05
Df Model:	34		
Covariance Type:	nonrobust		

Model Building

Feature Selection

OLS Regression Results

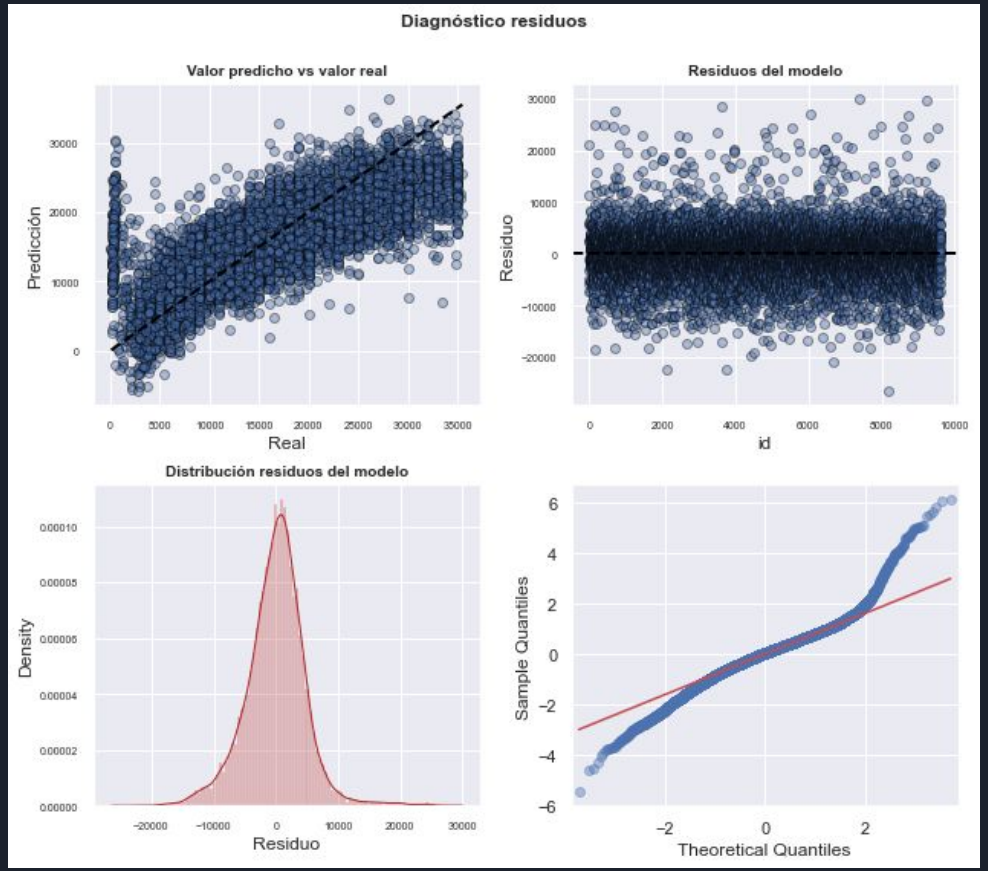
Dep. Variable:	y	R-squared:	0.626
Model:	OLS	Adj. R-squared:	0.625
Method:	Least Squares	F-statistic:	668.1
Date:	Tue, 09 Aug 2022	Prob (F-statistic):	0.00
Time:	13:22:15	Log-Likelihood:	-95265.
No. Observations:	9612	AIC:	1.906e+05
Df Residuals:	9587	BIC:	1.908e+05
Df Model:	24		
Covariance Type:	nonrobust		

- Bonferroni correction

B= 0.0014

Model Building

Performance



Model Building

Business Presentation

Losses per car: $-1,711.92
Profits per car: $1,893.85
P(Buying|PriceReal): 65%
Car Value Mean: $14,555.05
12%