Linear Regression Project

Lingraj S Vannur
Madhav Ponnudurai
Rushil Manglik

Outlines

1. Problem Statement and Data Description
2. Exploratory Data Analysis
3. Multicollinearity and VIF
4. Influential points
5. Statistical tests
6. Residual Plots (Model Assumptions).
7. Final model and Contribution
8. Conclusion

Problem Statement

Our motivation is to select the best model using linear regression, which could assist in predicting the car price depending on the various features a car brand is offering. This could also help us understand why some car brands are very expensive while most others are in the affordable range.

Dataset Description

1. **Target column:** Price
2. **Number of features:** 25
3. **Number of observations:** 205

Variable description:

a) **15 numerical variables** (symboling, normalized-losses, wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg).

b) **10 categorical variables** (fuel-type, aspiration, num-of-doors, body-style, drive-wheels, engine-location, engine-type, num-of-cylinders, fuel-system, make)

Research questions

To get the best model we need to answer the following questions:

1. How many and what predictors are **correlated**?
2. Are there any **outliers/influential points** that are severely affecting the model?
3. Are there any **insignificant predictors**?
4. Are any **model assumptions** being violated?
5. Finally checking model performance.

Summary of methods

1. Exploratory Data Analysis
2. Checking for multicollinearity using VIF
3. External studentised t-tests and Cook’s distance
4. Breusch-Pagan test for heteroscedasticity and Jarque-Bera test for normality
5. T-test and ANOVA(typ=1)
6. Adj-R^2 and MSE

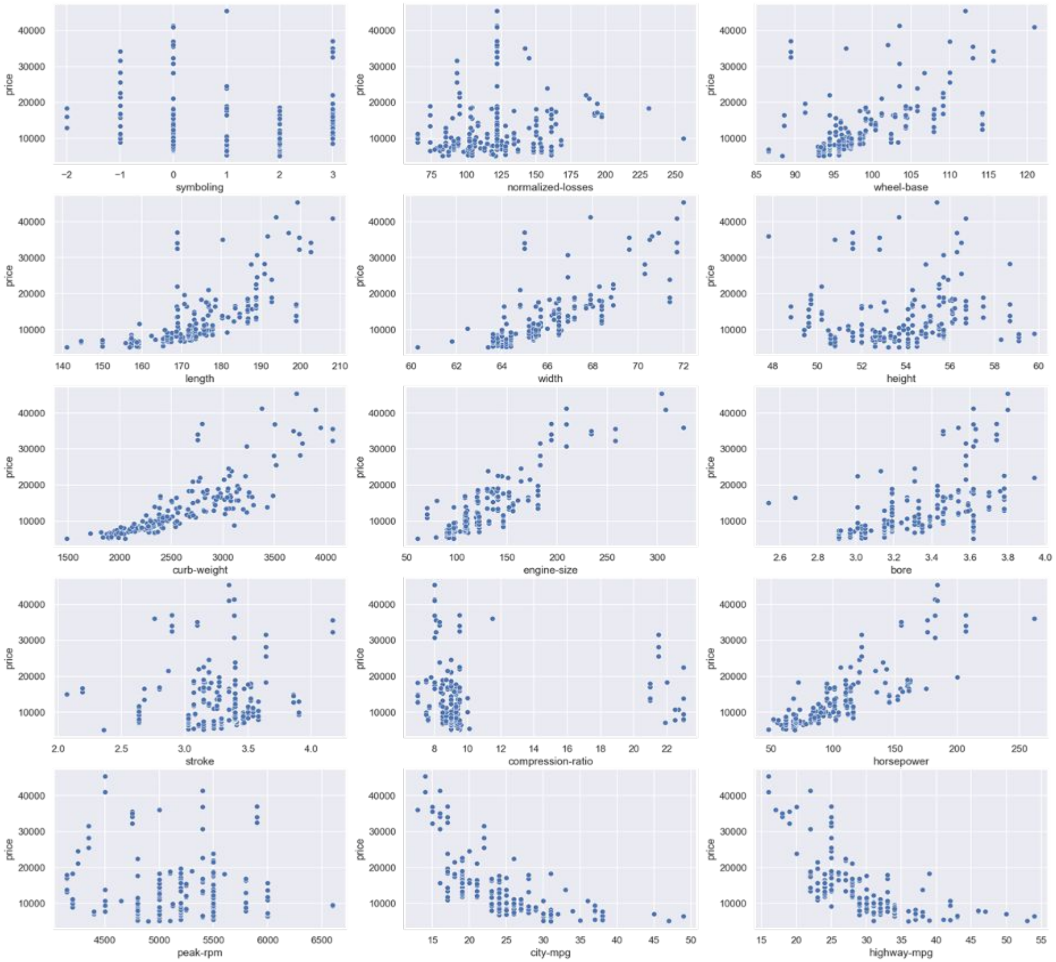
Exploratory Data Analysis

1. There were no null values but **few cells have '?'** which had to be replaced in columns like normalized-losses, num-of-doors, bore, stroke, horsepower, and peak-rpm variables.

2. The **null values were replaced by the mean value** of the respective columns. For num-of-doors null value was filled based on domain knowledge.

3. For target variable ‘price’, the unknown values were saved as test_data to check model accuracy.

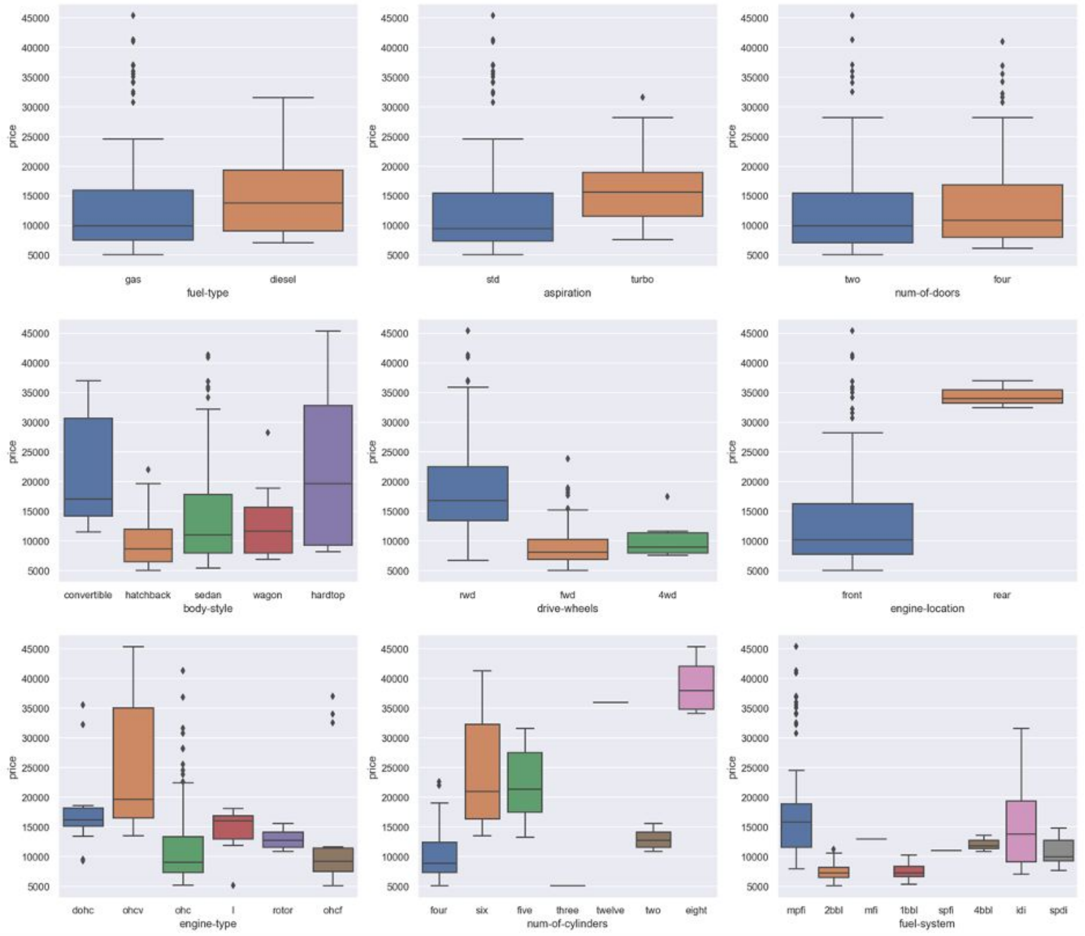
4. For numerical variables we plot scatterplots to understand their relationship with price.



- Columns - compression ratio, height, symboling, normalised-losses, stroke and peak-rpm attributes **has very less relation or no relation with price.**

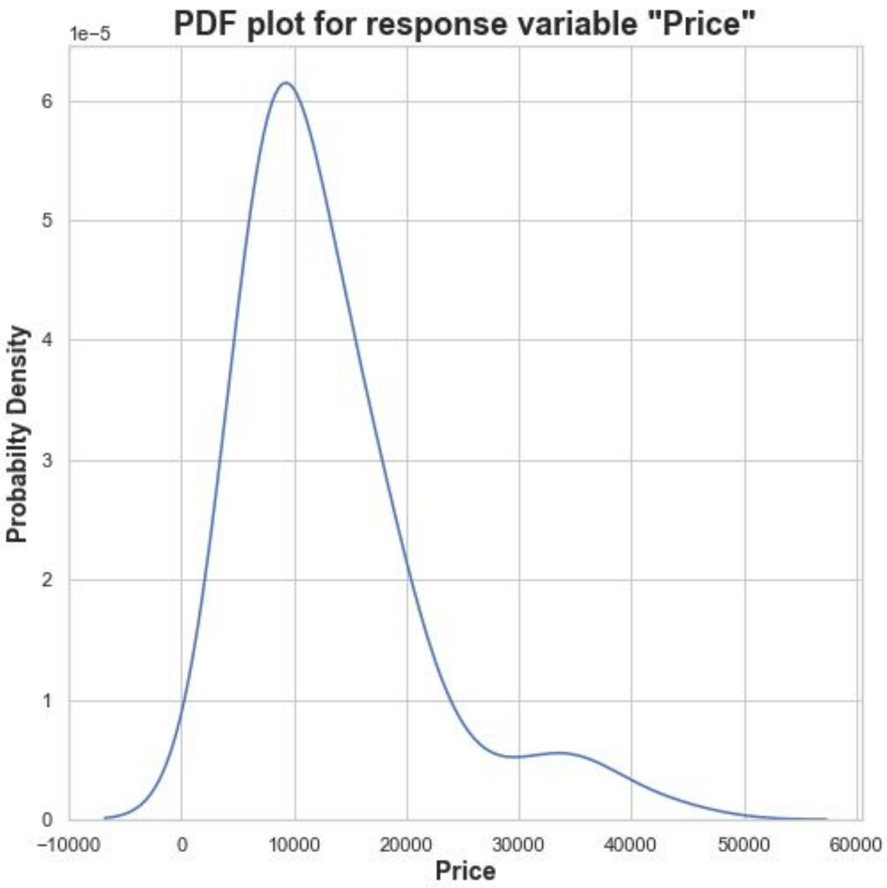
- Other columns have a **positive or negative relationship with price.**

5. For categorical variables we plotted boxplots to understand their relationship vs price.



- Rear wheel drive, rear engine location, eight number of cylinders **make the vehicles costlier.**
- **Multiple outliers** can be identified for each case.

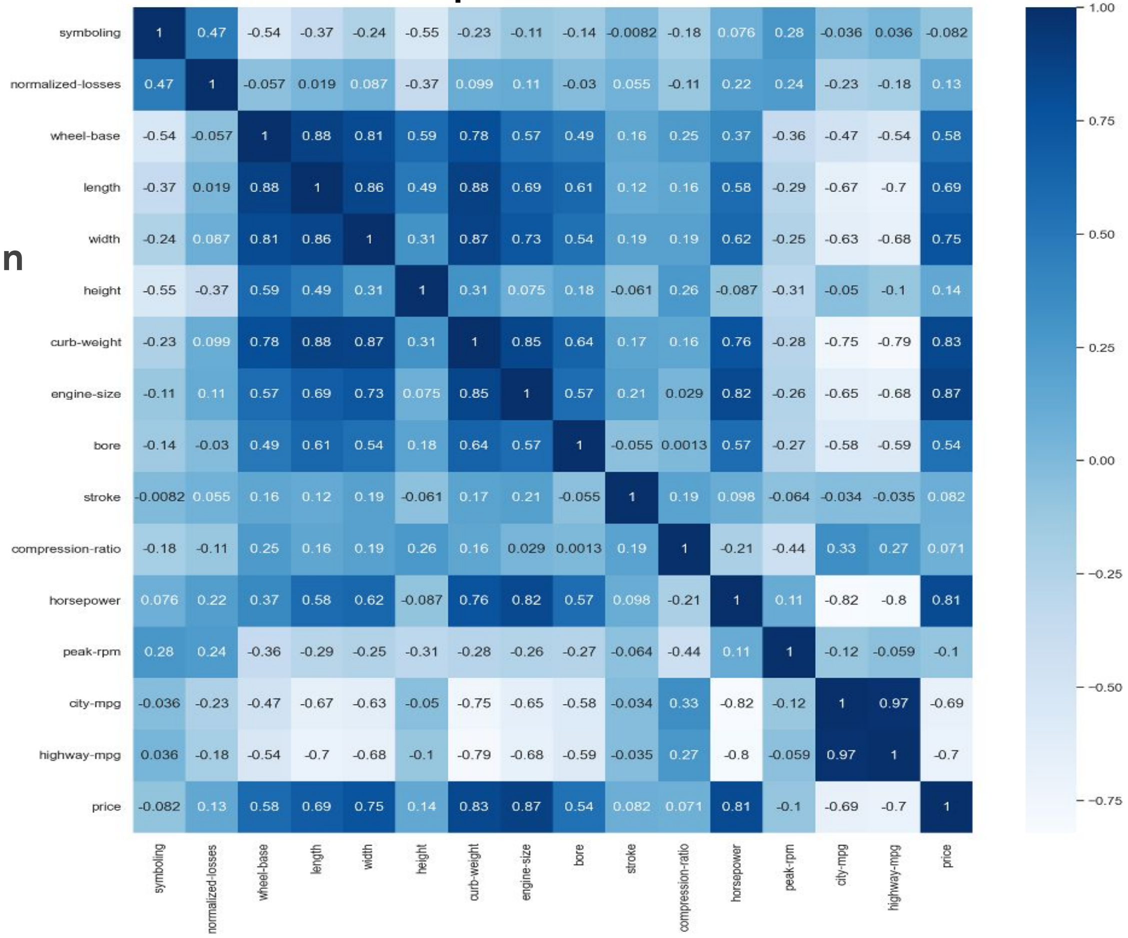
Target variable: “Price” PDF plot



Multicollinearity

We combined the length, width and height columns to create a new column "volume".

Heatmap for numerical variables



Variance Inflation Factor

1. After encoding the categorical features we get 72 features in our data set.

2. As we have only **22 columns out of 72 columns with VIF<10**, this isn't a reliable method to filter out predictors.

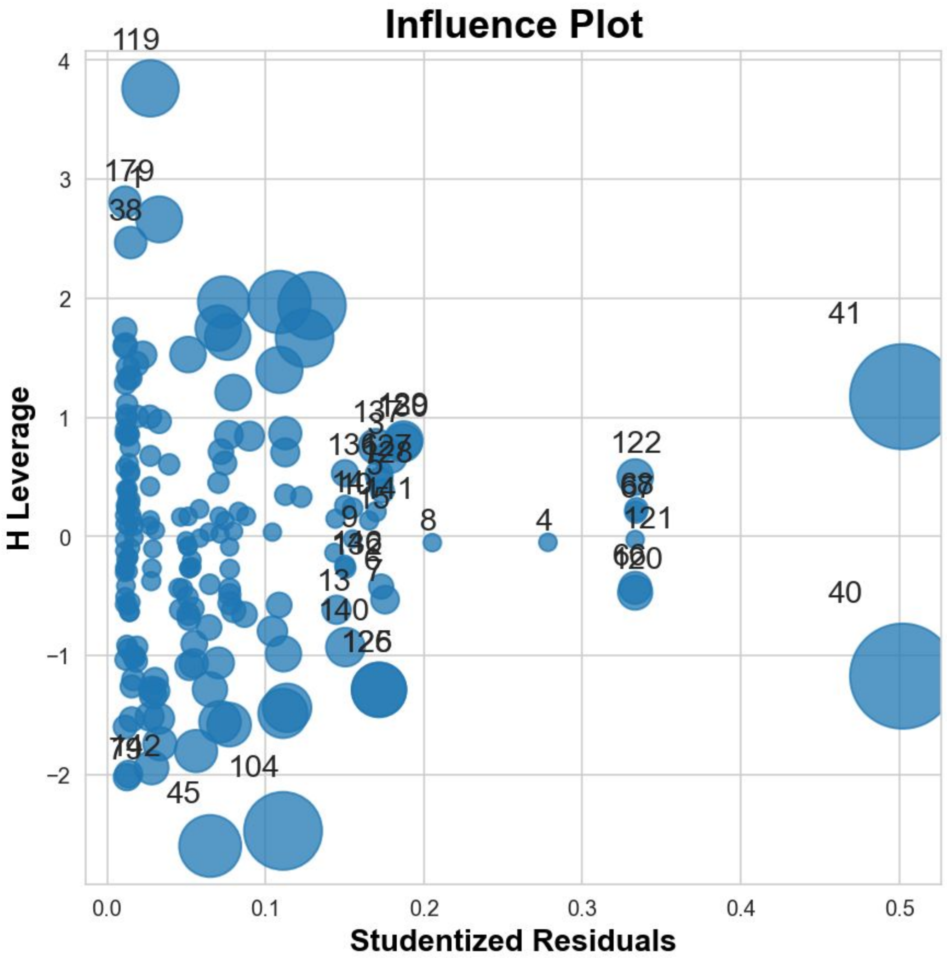
3. Figure beside shows the features which had VIF<10.

	VIF Factor	features
0	0.000000	Intercept
3	5.148115	make[T.chevrolet]
6	4.012004	make[T.isuzu]
7	7.706195	make[T.jaguar]
10	2.938450	make[T.mercury]
14	8.352246	make[T.plymouth]
16	3.468340	make[T.renault]
17	9.790689	make[T.saab]
23	6.483269	aspiration[T.turbo]
24	4.202776	num_of_doors[T.two]
25	3.572518	body_style[T.hardtop]
35	6.169138	engine_type[T.ohcv]
41	8.732602	num_of_cylinders[T.twelve]
44	6.311034	fuel_system[T.4bbl]
46	2.330516	fuel_system[T.mfi]
48	9.671580	fuel_system[T.spdi]
49	3.084567	fuel_system[T.spfi]
50	7.138638	symboling
51	3.864447	normalized_losses
55	8.596078	height
59	6.435681	stroke
62	6.367247	peak_rpm

Influential Points

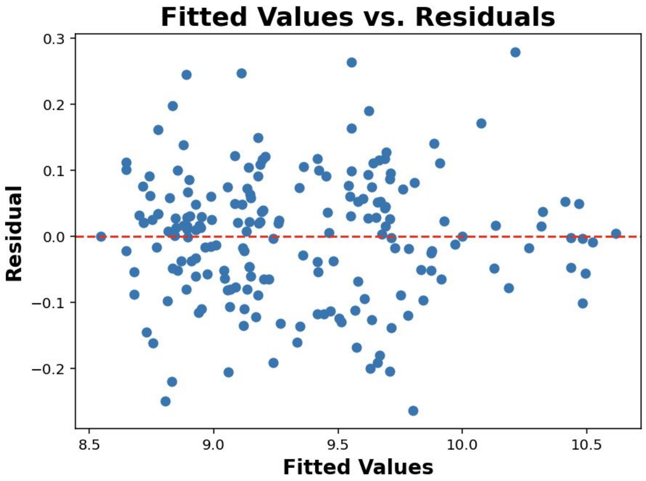
Observations

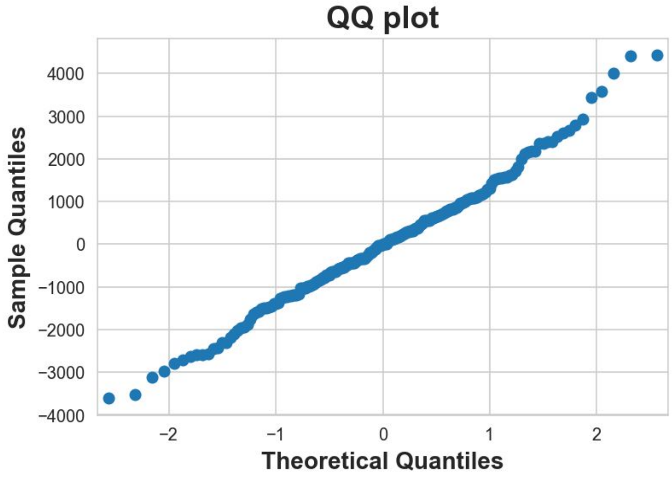
1. We remove the points which we get common in both **Cook's distance** and **studentized residual test results**.
2. **7 records are removed as outliers** (indexes: 40, 41, 119….).
3. Now we are **left with 194 observations.**



Model Assumptions

1. **Heteroscedasticity:**
   a. Initially found that heteroscedasticity existed.
   b. So, **applied log transformation** on the Price variable and performed **Breusch Pegan Test:** *LM-Test p-value:* 0.15 >0.05
   c. Hence, **resolved the issue of heteroscedasticity.**
2. **Normality:** From **QQ plot** we can observe the linear plot representing and performing **Jarque-Bera test:** *(0.406>alpha)*, normality assumption is not violated.
3. **E[ei] = 0**. We can observe that average mean of residuals is 0.





T-test and ANOVA(typ=1) Results

1. **Once the outliers and multicollinearity** is **handled,** model assumptions were agreed upon, We ran the the t test to select the significant features in the model.
2. To verify we are not missing any significance in the features, **we run ANOVA test type 1**.
3. Finally the table beside shows **few features selected** after doing t test and anova type 1.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.7496	0.196	39.467	0.000	7.362	8.137
make_audi	0.3043	0.050	6.141	0.000	0.207	0.402
make_honda	0.0853	0.033	2.557	0.011	0.019	0.151
body_style_convertible	0.1550	0.049	3.191	0.002	0.059	0.251
make_saab	0.1344	0.049	2.732	0.007	0.037	0.231
curb_weight	0.0006	4.54e-05	14.309	0.000	0.001	0.001
make_bmw	0.4168	0.046	9.142	0.000	0.327	0.507
aspiration_turbo	0.0678	0.026	2.632	0.009	0.017	0.119
num_of_cylinders_two	0.1672	0.071	2.360	0.019	0.027	0.307
engine_location_rear	0.3865	0.125	3.093	0.002	0.140	0.633
num_of_cylinders_three	0.4153	0.124	3.347	0.001	0.170	0.660
make_isuzu	-0.2380	0.081	-2.942	0.004	-0.398	-0.078
fuel_system_mpfi	0.0779	0.024	3.213	0.002	0.030	0.126
make_porsche	0.4222	0.110	3.841	0.000	0.205	0.639
city_mpg	-0.0075	0.004	-1.992	0.048	-0.015	-6.98e-05
fuel_system_idi	0.1308	0.048	2.730	0.007	0.036	0.225
num_of_cylinders_eight	0.2034	0.071	2.864	0.005	0.063	0.344

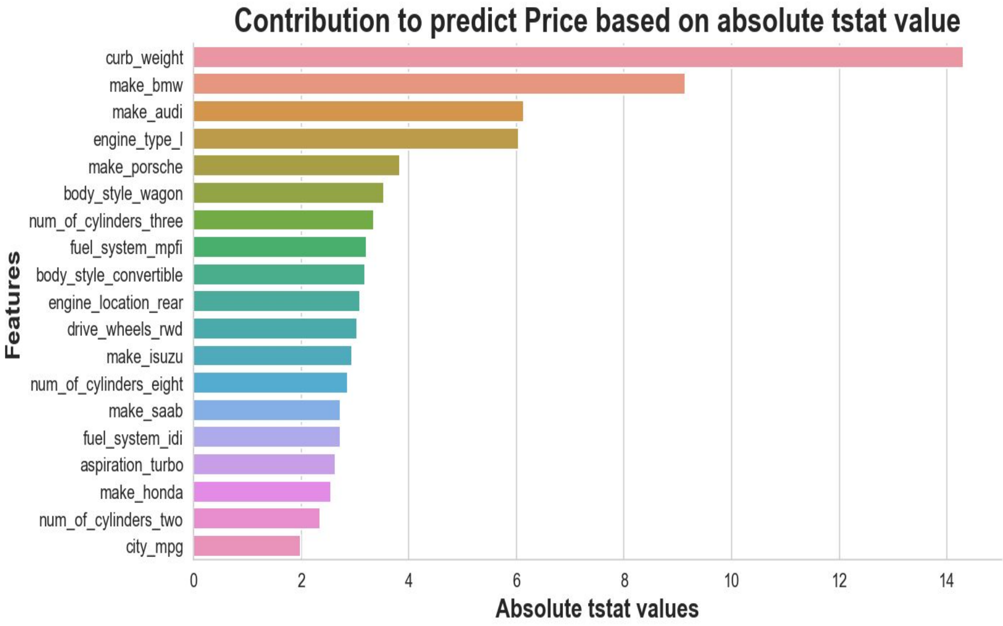
Final model summary

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.956
Model:                            OLS   Adj. R-squared:                  0.951
Method:                 Least Squares   F-statistic:                     198.7
Date:                Sat, 15 Oct 2022   Prob (F-statistic):          2.96e-107
Time:                        16:40:01   Log-Likelihood:                 168.50
No. Observations:                 194   AIC:                            -297.0
Df Residuals:                     174   BIC:                            -231.7
Df Model:                          19
Covariance Type:            nonrobust
```

Final model selected achieves:

1. Adj.R2 = 95.1%
2. MSE: 0.0114

From the right graph, we can observe that ***car maker, curb weight, engine type, mpg*** are some major features contributing in **predicting the car price.**

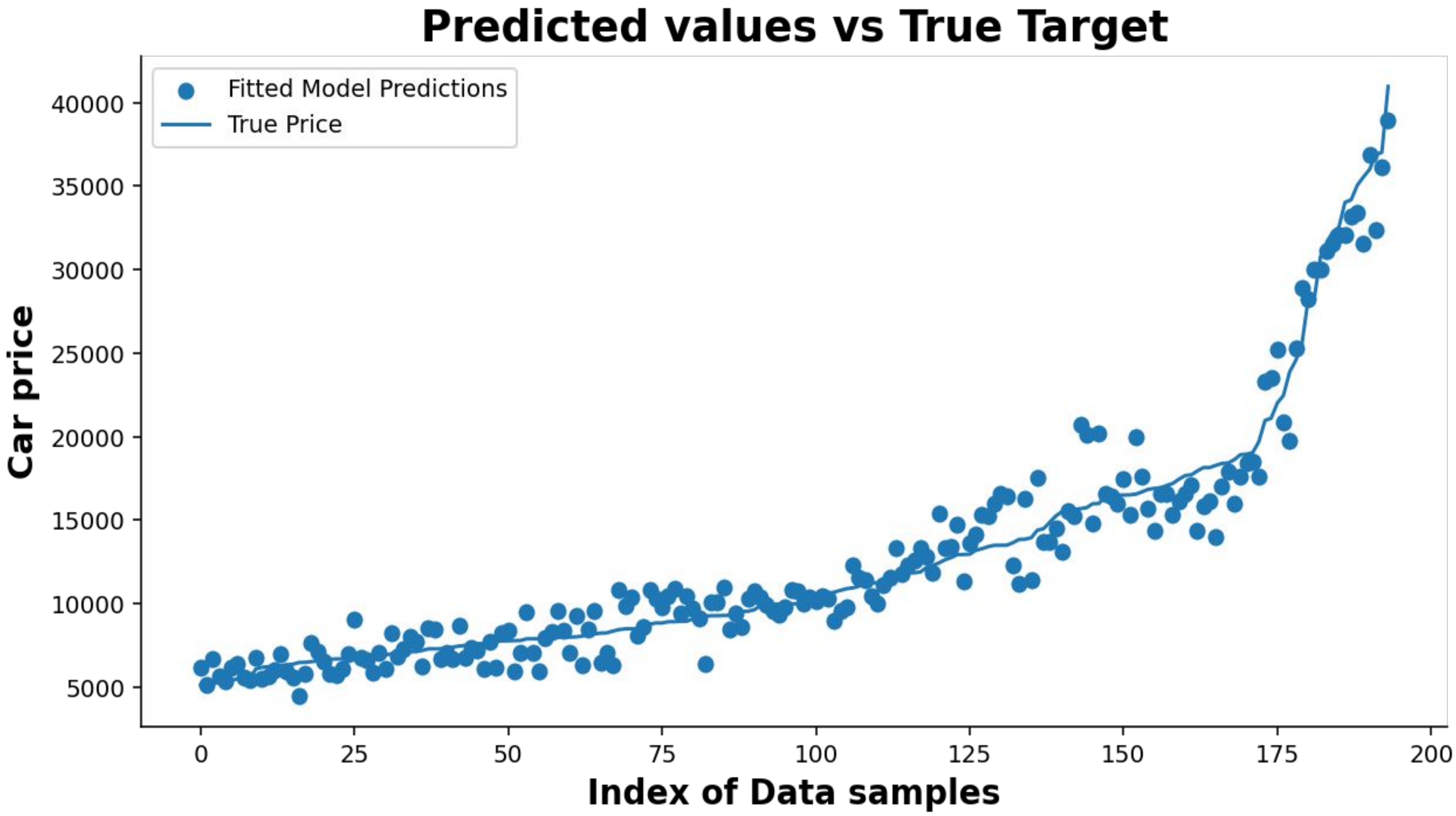


Final model in equation form:

Price~

engine_size+make_toyota+body_style_convertible+make_plymouth+'make_porsche+make_peugot+engine_type_rotor+make_bmw+engine_type_ohcv+curb_weight+stroke+num_of_cylinders_three+wheel_base+bore+make_dodge+horsepower+make_mercedes_benz+make_mitsubishi.

Model Performance Graph



Conclusion

- We observed that the data was too messy with many variables and null values.
- Once the data was cleaned we **performed statistical tests** and **resolved the problems** like:
    - Multicollinearity
    - Influential points
    - Statistical tests
    - Model assumptions and mitigated the violations.
- We came to know that the **car price depends on** the brand(car maker), engine type, mileage, body style whether it's convertible, fuel system etc..
- Models like decision tree or xgboost can perform better.