Table of Contents

Introduction

This report will analyze the attributes of the Used Cars dataset:

The goal of this report is to gather useful statistics and information regarding the cars and predict the best selling price for it based on a number of predictors. We analyzed the dataset Vehicle dataset from cardekho | Kaggle published on the website Kaggle.

After the initial data pre-processing we are using a linear regression model to fit the observations and give us an insight on the important predictors that are contributing to a better price estimation. We are also validating the model assumptions and validating the predictor variables and observations to remove any discrepancies that may affect our model and add bias to our predictions.

Here is a snapshot of how the data looks like:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner

Research Questions

- What are the relevant characteristics of used cars that can influence our choice of target variable i.e selling price.
- Check the relationship between selling price and rest of the other predictors.
- Check different iterations of the model and compare the accuracy of the intermediate model to the finalized fitted model.

Description of your dataset

The dataset consists of information about used cars listed on cardekho.com. The dataset has a total of 4340 observations. It has 8 columns, each columns consisting of information about specific car features like:

Car_Name gives information about the car brand and model

Year gives information of when the brand new car was purchased

Selling_price is the price at which car is being sold. This will be the target variable for further prediction of price.

Km_driven number of kilometers the car has been driven so far.

Fuel gives information of the fuel type of car (CNG , petrol,diesel etc).

Seller_type gives us information of whether the seller is an individual or a dealer.

Transmission gives information if the car is automatic and manual.

Owner number of previous owner of the car

Exploratory Data Analysis

We also convert the year to a new column Year_used which we calculated from the year when the car was manufactured - yearPurchased.

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	Year_used
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner	14
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner	14
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner	9
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner	4
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner	7

The data was already clean and pre processed. There were no missing value in the dataset

```
name             0
year             0
selling_price    0
km_driven        0
fuel             0
seller_type      0
transmission     0
owner            0
Year_used        0
```

Next we had a look at the distribution of categorical variables to see if there is any bias in them in terms of one particular record level and modify them accordingly. Here we have combined the categorical variable levels with fewer number of levels into others category.

In the fuel predictor we convert CNG, LPG, Electric to other and in car owner type we convert Third Owner, Fourth & Above Owner, Test Drive Car to others.

```
Diesel      2153
Petrol      2123
CNG           40
LPG           23
Electric       1
Name: fuel, dtype: int64
```

```
Diesel    2153
Petrol    2123
other       64
Name: new_fuel, dtype: int64
```

```
First Owner             2832
Second Owner            1106
Third Owner              304
Fourth & Above Owner      81
Test Drive Car            17
Name: owner, dtype: int64
```

```
First Owner     2832
Second Owner    1106
other            402
Name: new_owner, dtype: int64
```

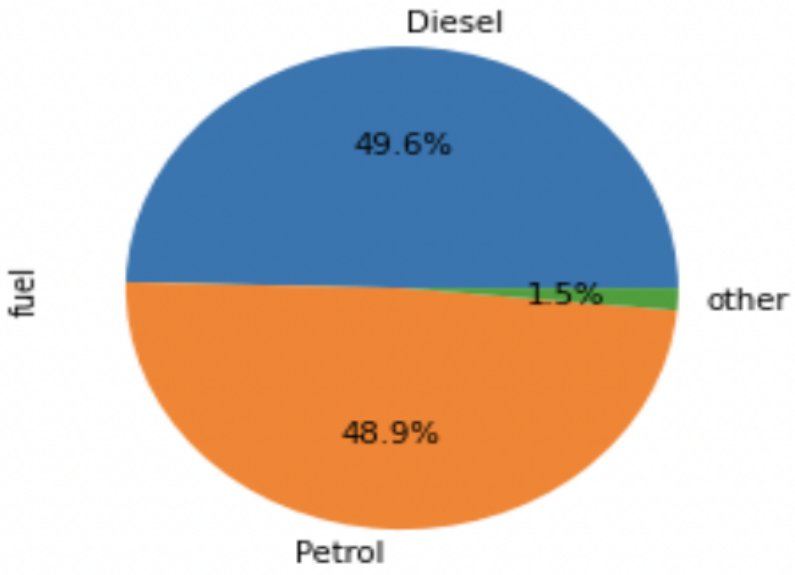


Figure: Pie Chart for fuel types

From the above pie chart we get a look at how many cars use what type of fuel as per our dataset.

From the below plot **Figure: Selling Price vs Year**, it shows that with the year increase, the price of the car also increases. The trend in the plot shows the linear increase in Y. So the increase in year causes an exponential increase in price.

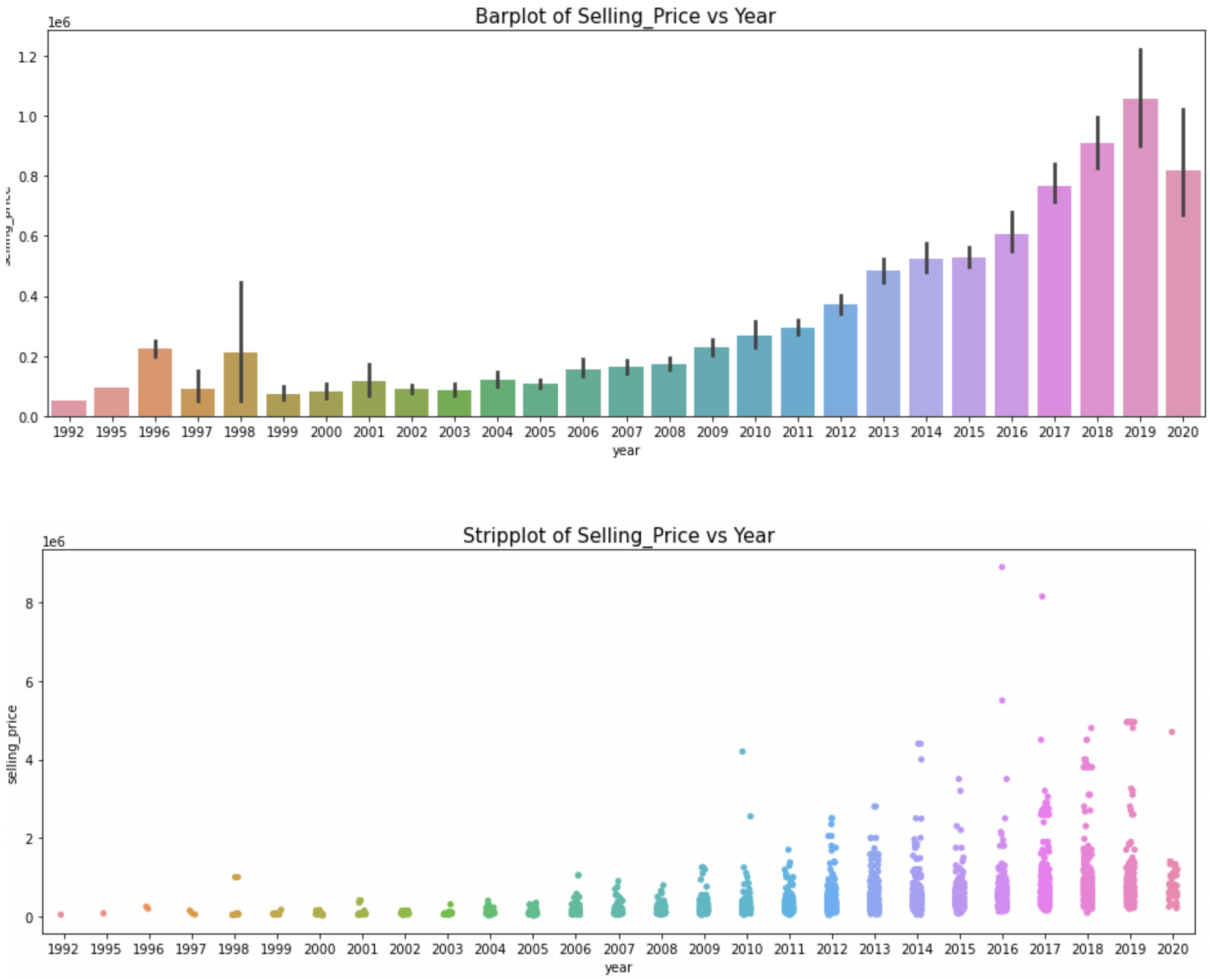


Figure: Selling Price vs Year

From plot **Figure: Average Car Price Vs Transmission,** we see that automatic cars have more average price as compared to manual cars.

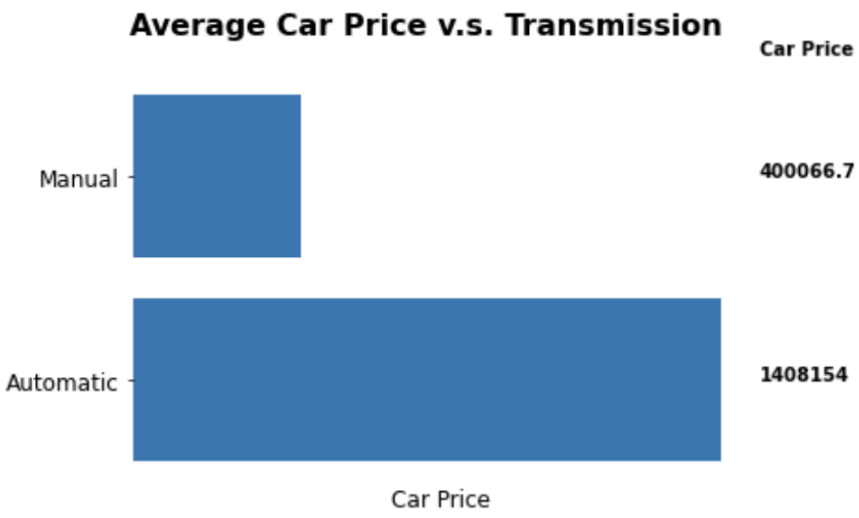


Figure: Average Car Price Vs Transmission

Methods

After doing the initial data understanding, EDA and visualization we decided to do a multivariate linear regression on our data set to predict the selling price for the used cars. We have 4 categorical variables: fuel, seller_type, owner and transmission and three numerical variables years_used, km_driven ,selling_price. Below is the initial model summary for our model.

OLS Regression Results

Dep. Variable:	selling_price	**R-squared:**	0.458
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	407.3
Date:	Fri, 15 Oct 2021	**Prob (F-statistic):**	0.00
Time:	12:15:04	**Log-Likelihood:**	-62411.
No. Observations:	4340	**AIC:**	1.248e+05
Df Residuals:	4330	**BIC:**	1.249e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.835e+06	2.45e+04	74.813	0.000	1.79e+06	1.88e+06
transmission[T.Manual]	-8.667e+05	2.2e+04	-39.457	0.000	-9.1e+05	-8.24e+05
seller_type[T.Individual]	-6.995e+04	1.63e+04	-4.289	0.000	-1.02e+05	-3.8e+04
seller_type[T.Trustmark Dealer]	1.642e+05	4.44e+04	3.696	0.000	7.71e+04	2.51e+05
owner[T.Second Owner]	-3.881e+04	1.66e+04	-2.338	0.019	-7.14e+04	-6262.069
owner[T.other]	-1.965e+04	2.47e+04	-0.795	0.427	-6.81e+04	2.88e+04
fuel[T.Petrol]	-2.901e+05	1.42e+04	-20.390	0.000	-3.18e+05	-2.62e+05
fuel[T.other]	-2.783e+05	5.42e+04	-5.133	0.000	-3.85e+05	-1.72e+05
km_driven	-0.9774	0.168	-5.819	0.000	-1.307	-0.648
year_used	-3.571e+04	1893.797	-18.855	0.000	-3.94e+04	-3.2e+04

Omnibus:	4368.779	**Durbin-Watson:**	1.935
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	502605.375
Skew:	4.659	**Prob(JB):**	0.00
Kurtosis:	54.890	**Cond. No.**	6.79e+05

We will be following the approach 1 method of model selection which is using t-test for individual coefficient and ANOVA(type=1 and type=1) to choose initial candidates with significant predictors. We will finally use the R-squared values to choose a final one. We will finally be checking model assumptions and doing model diagnosis to improve the model if there is any problem .

Model diagnosis

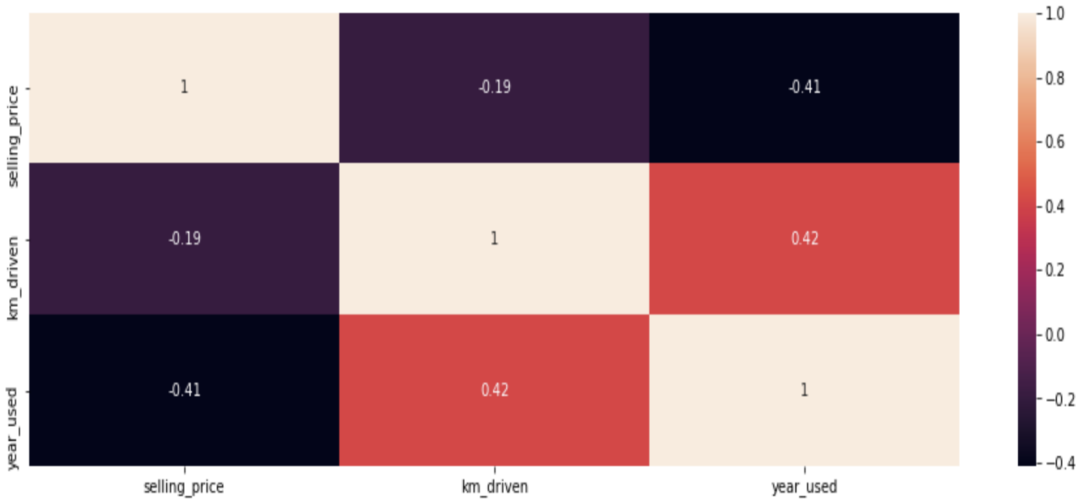
Multicollinearity:

After fitting the model with all the predictors of interest we checked the variance inflation factor(VIF) between all predictors. The result shows that there is no serious Multicollinearity between predictors as the VIF values for all are below 2. So we can proceed with our interpretation without dropping or combining any variables.

We used the following classification for VIF:
1 < VIF <= 4 : Light Multicollinearity
4 < VIF <= 10 : Moderate Multicollinearity
VIF >= 10 : Serious Multicollinearity

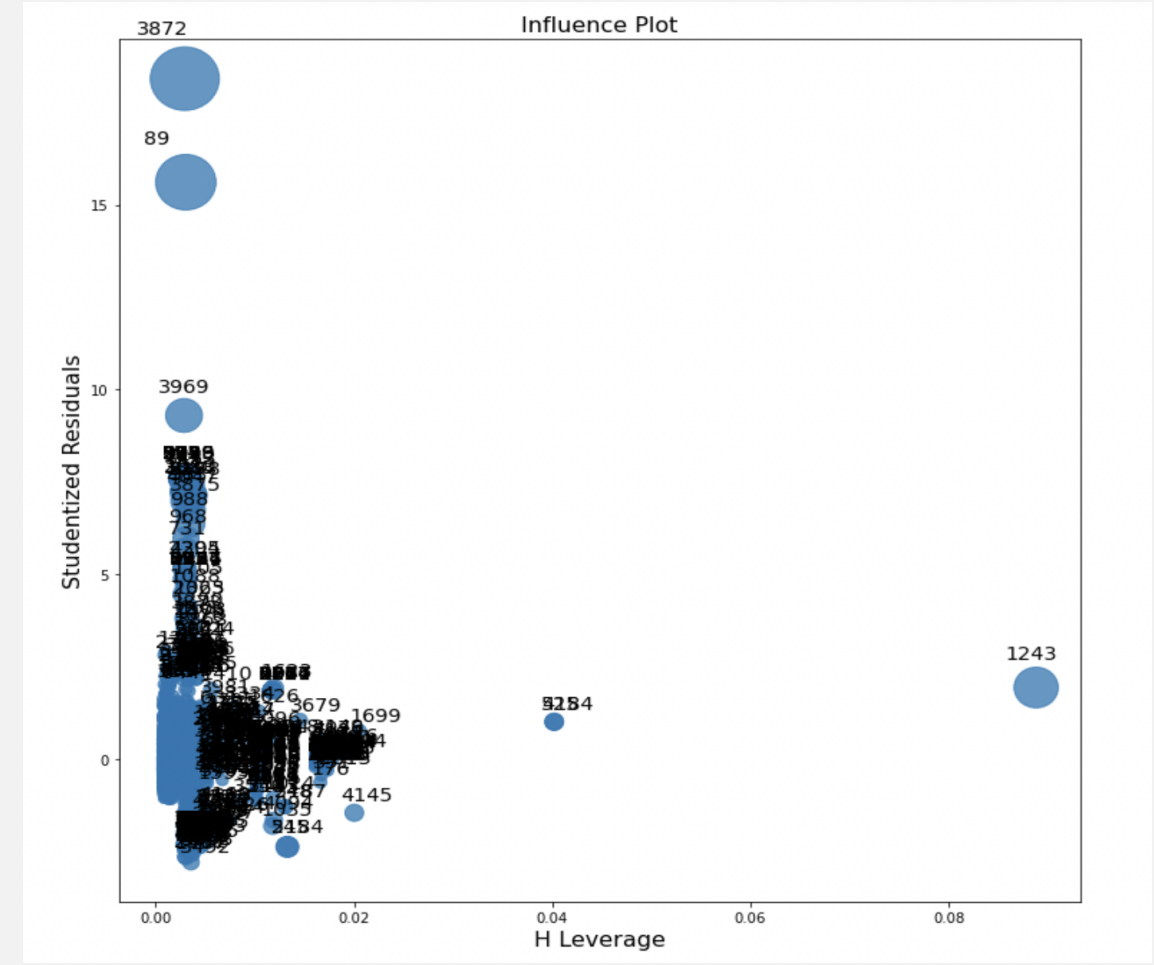
	VIF Factor	features
0	14.369745	Intercept
1	1.067157	transmission[T.Manual]
2	1.199645	seller_type[T.Individual]
3	1.081988	seller_type[T.Trustmark Dealer]
4	1.250392	owner[T.Second Owner]
5	1.227482	owner[T.other]
6	1.208988	fuel[T.Petrol]
7	1.020325	fuel[T.other]
8	1.466386	km_driven
9	1.522386	year_used

Also there is some correlation between year_used and km. However, since the VIF doesn't show multilinearity, we will still use these predictors in the model.



Influential Points:

1. To detect influential points in our data set we plotted the external studentized residuals against the leverage .



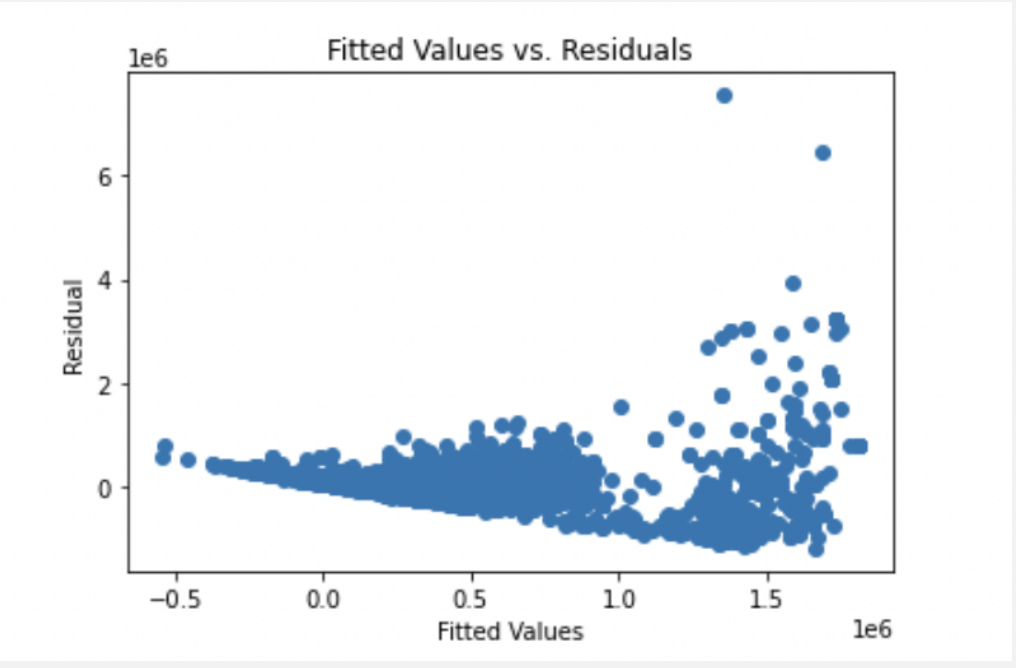
The plot shows the potential influential points as labelled . Cook's distance value determines the size of the points on the plot. Using the external studentized residual test with cutoff t-critical value at 5% significance and the Cook's distance test cutoff set at 4/n i.e 0.00092 we identified more than 100 influential points for our data set.

2. We Deleted those influential points that are recognized by both methods and obtained a dataset without influential points (data_woinf)

Later, we will report the result both with and without the influential points

Heteroscedasticity:

1. Drew the Fitted Values vs. Residuals Plot for the model with complete dataset .

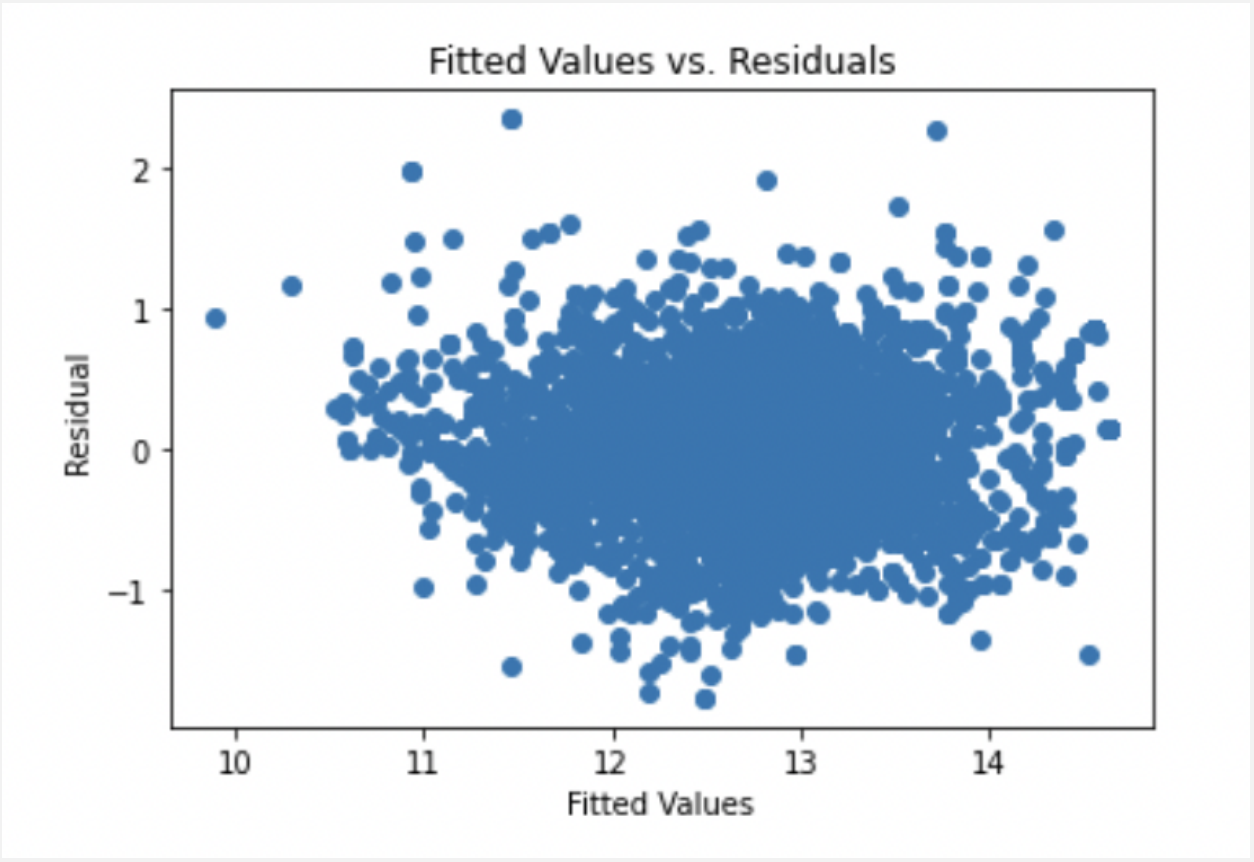


From this initial residual plot we see that there is a slight downward trend . The variance is not constant, being very narrow in the start and more spread at the end. The residuals are approximately around the mean 0 with deviations at the start and end. The slight downward trend hints at some correlation between the residual terms thus not being entirely independent.

We perform the (Breusch-Pagan) BP test. We get a high p value which allows us to reach the conclusion that Heteroscedasticity exists in the model.

2. To fix the issue a bit we performed log transformation on y and checked the heteroscedasticity again. We plot the residual plot and do the BP test again

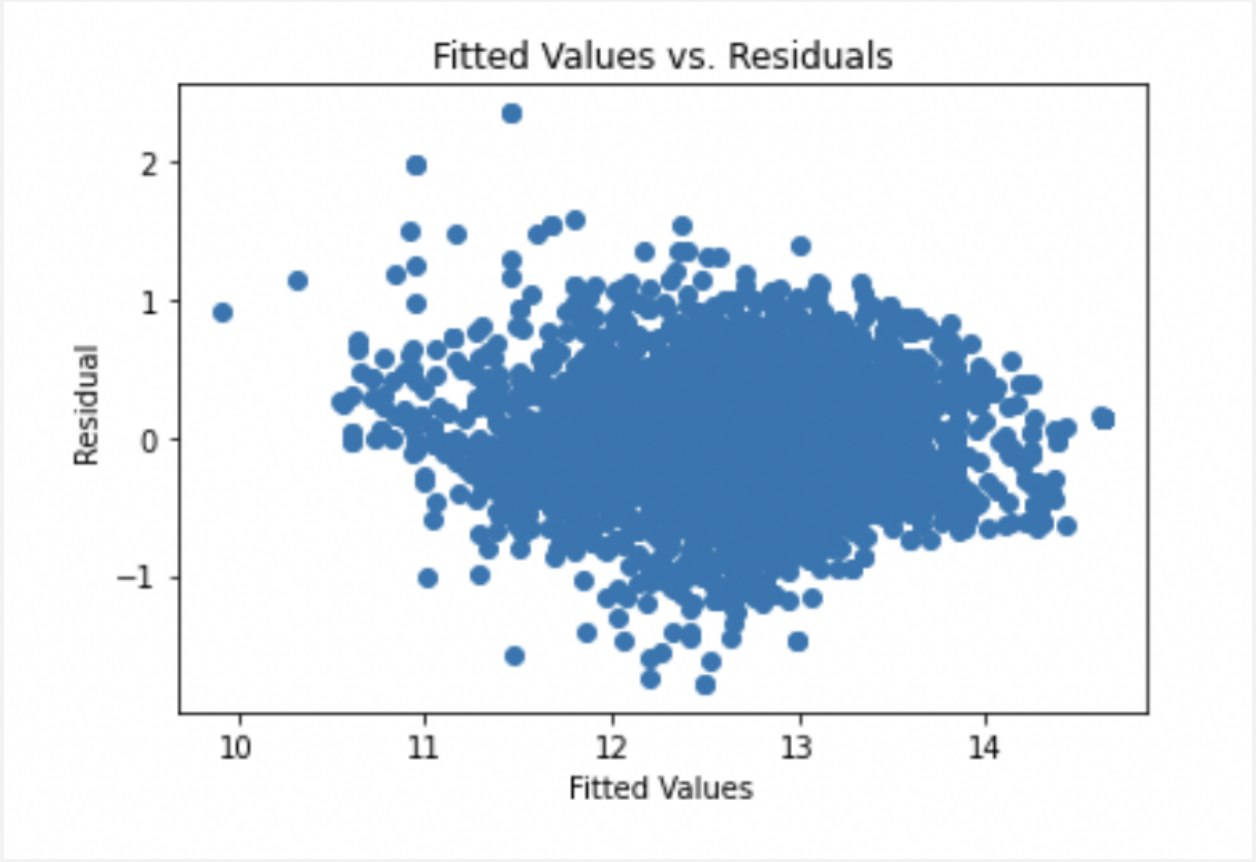
```
{'LM Statistic': 256.59732968168, 'LM-Test p-value': 4.033176530762076e-50}
```



From the above plot we see that the model assumption violation has been improved a lot. We now see the residual points approximately even around 0 and the spread looks constant at most places throughout the plot. Also the residual points are now completely random with no specific trend , thereby approving the assumption of independent and uncorrelated terms. We do the Breusch-Pagan (BP) test again and see that heteroscedasticity still exists but has improved a lot.

Iteration 2 (No influential points): We now checked the heteroscedasticity again, plotting both the residuals plot and doing the BP test for the model without the influential points. From the results we see that heteroscedasticity still exists but improved after deleting the influential points.

```
{'LM Statistic': 132.74890670864497, 'LM-Test p-value': 3.2244447299703192e-24}
```

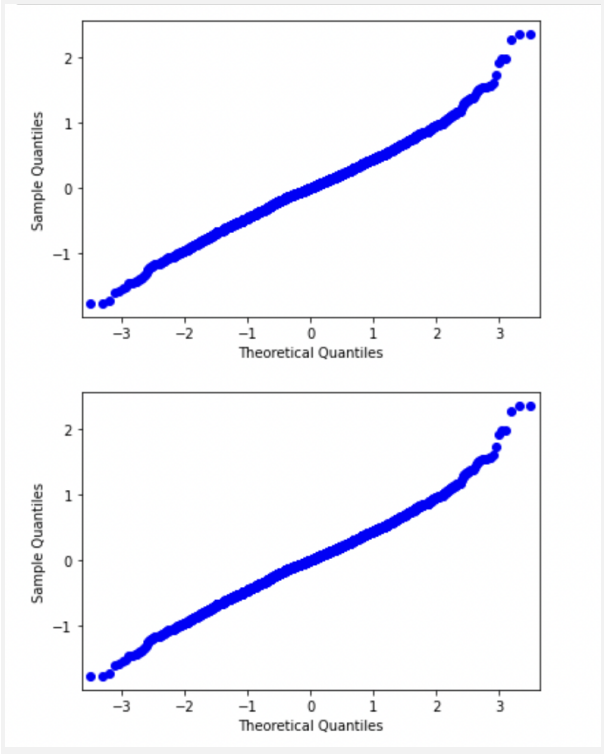
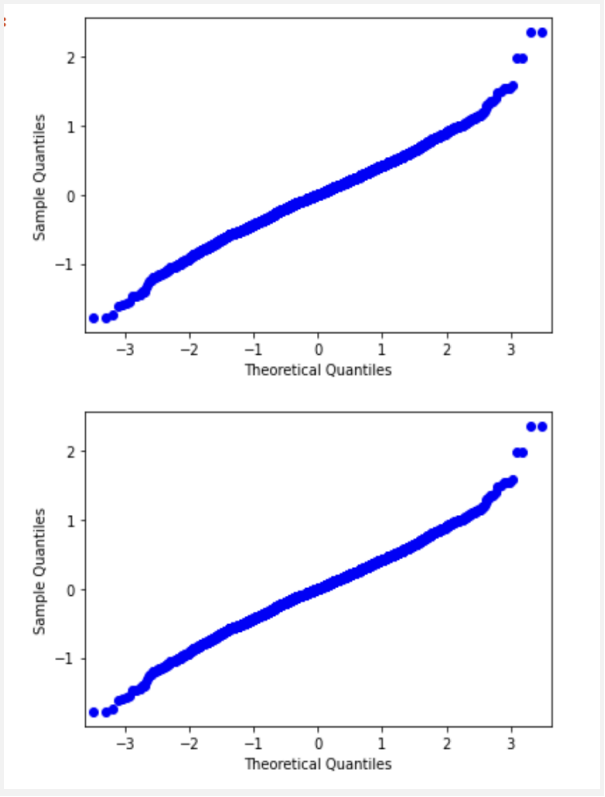


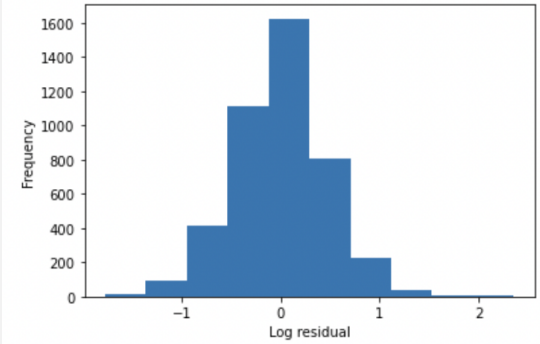
There isn't any dramatic difference between the plots with and without the influential points. We will keep doing rest of the diagnosis and report the analysis.

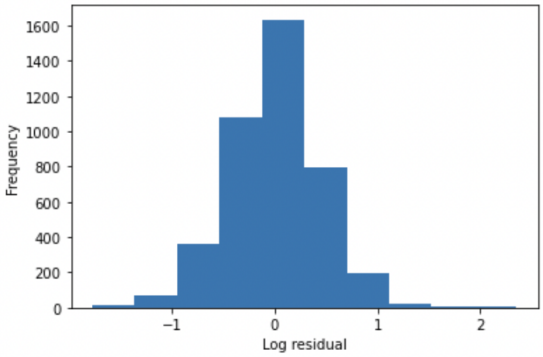
Normality:

We drew the QQ plot ,histogram of residuals and observed p-value of the JB test to check for normality assumption of the residuals for the complete dataset first and then for the model without the influential data points.

Figure : shows the change in the QQ plot and residual histogram for models with and without the influential points(iteration2)

Original dataset	After Removing influential points
QQ Plot	
	
Residual plot	





Jarque-Bera Test

Durbin-Watson:	1.856
Jarque-Bera (JB):	147.268
Prob(JB):	1.05e-32
Cond. No.	6.79e+05

Durbin-Watson:	1.895
Jarque-Bera (JB):	146.228
Prob(JB):	1.77e-32
Cond. No.	6.74e+05

Initially from the QQ plot we see that the line is almost linear at 45 degrees with being slightly curved in the end. The histogram showing the distribution of the residuals shows a shape similar to normal distribution. But when we see the Jarque-Bera test we see that normality was violated both with and without influential points.

Model selection

We performed ANOVA test (type 1) on our model and here are the results we got:

ANOVA Table with influential points:

	df	sum_sq	mean_sq	F	PR(>F)
transmission	1.0	522.595847	522.595847	2355.598272	0.000000e+00
seller_type	2.0	143.762823	71.881411	324.005117	7.421998e-132
owner	2.0	231.983966	115.991983	522.833307	4.076690e-204
fuel	2.0	406.302721	203.151360	915.703783	0.000000e+00
km_driven	1.0	137.596840	137.596840	620.217098	4.664632e-128
year_used	1.0	653.381102	653.381102	2945.112186	0.000000e+00
Residual	4330.0	960.622209	0.221853	NaN	NaN

ANOVA Table without influential points:

	df	sum_sq	mean_sq	F	PR(>F)
transmission	1.0	284.888233	284.888233	1441.508062	3.984851e-271
seller_type	2.0	151.063999	75.531999	382.184919	4.130354e-153
owner	2.0	220.937858	110.468929	558.962547	9.409443e-216
fuel	2.0	355.722763	177.861382	899.962112	0.000000e+00
km_driven	1.0	125.569073	125.569073	635.367875	1.345251e-130
year_used	1.0	639.404716	639.404716	3235.328619	0.000000e+00
Residual	4163.0	822.742338	0.197632	NaN	NaN

From the table we can see that the p values of all the F tests are smaller than 0.05 both with and without influential points. That means that we do not need to remove predictors from our model. All the predictors(transmission, seller_type, owner, fuel, km_driven, year_used) should be included in the model.

And we would get the same result if we use single value t tests to decide which predictors should be kept in the model.

Here is the summary table for the single value t tests with influential points:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.8013	0.027	546.120	0.000	14.748	14.854
transmission[T.Manual]	-0.8001	0.024	-32.958	0.000	-0.848	-0.752
seller_type[T.Individual]	-0.1634	0.018	-9.064	0.000	-0.199	-0.128
seller_type[T.Trustmark Dealer]	0.3024	0.049	6.161	0.000	0.206	0.399
owner[T.Second Owner]	-0.0417	0.018	-2.273	0.023	-0.078	-0.006
owner[T.other]	-0.1030	0.027	-3.769	0.000	-0.157	-0.049
fuel[T.Petrol]	-0.5103	0.016	-32.448	0.000	-0.541	-0.479
fuel[T.other]	-0.6007	0.060	-10.025	0.000	-0.718	-0.483
km_driven	-4.175e-07	1.86e-07	-2.249	0.025	-7.81e-07	-5.35e-08
year_used	-0.1136	0.002	-54.269	0.000	-0.118	-0.109

Here is the summary table for the single value t tests without influential points:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.7714	0.031	475.513	0.000	14.711	14.832
transmission[T.Manual]	-0.7912	0.028	-28.514	0.000	-0.846	-0.737
seller_type[T.Individual]	-0.1701	0.018	-9.701	0.000	-0.204	-0.136
seller_type[T.Trustmark Dealer]	0.3210	0.047	6.828	0.000	0.229	0.413
owner[T.Second Owner]	-0.0455	0.018	-2.596	0.009	-0.080	-0.011
owner[T.other]	-0.1006	0.026	-3.885	0.000	-0.151	-0.050
fuel[T.Petrol]	-0.4743	0.015	-31.300	0.000	-0.504	-0.445
fuel[T.other]	-0.5809	0.057	-10.268	0.000	-0.692	-0.470
km_driven	-3.313e-07	1.77e-07	-1.875	0.061	-6.78e-07	1.52e-08
year_used	-0.1134	0.002	-56.880	0.000	-0.117	-0.109

From the tables, we can see that the p values of all the t tests are smaller than 0.05 both with and without influential points, which suggests that we should keep all the predictors in our model.

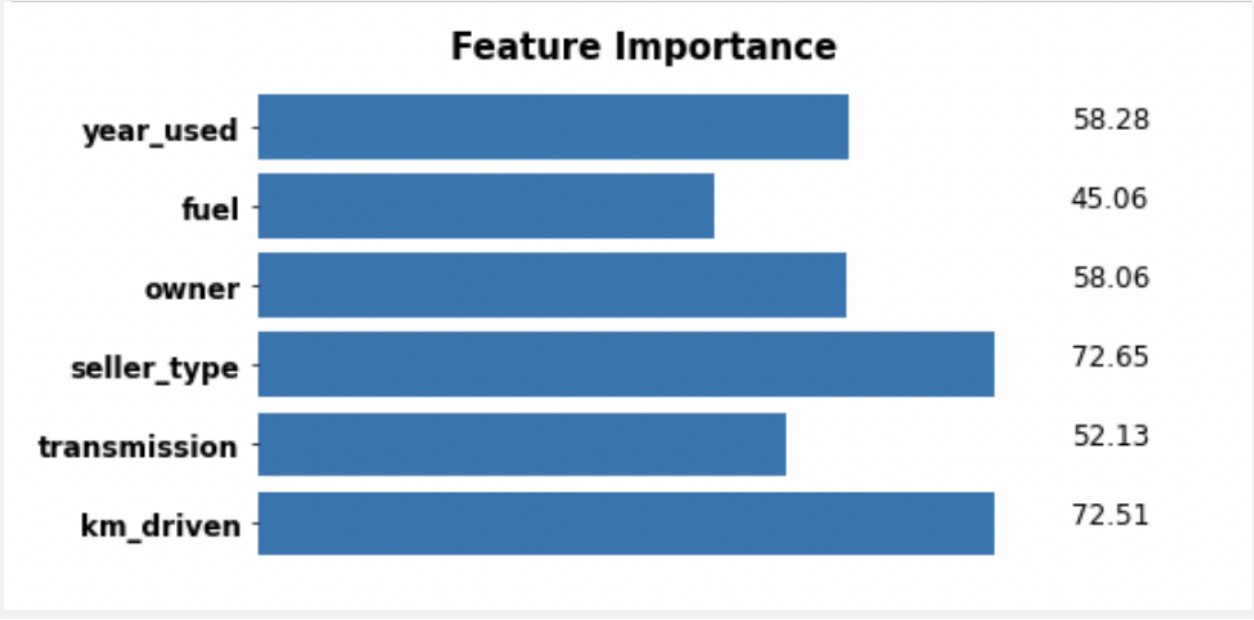
Final model

The best model for selling price in relation to its variables as per our interpretation is:
`log(selling_price) ~ km_driven + transmission + seller_type + owner + fuel + year_used`

According to the feature importance that we ran we see that all the predictors have a good impact on predicting the selling price of the used cars.

Extra: Feature importance

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Using the feature importance method from sklearn package we plot the importance value for each of our predictors. We can see below the importance level of each predictor for our selected model and we see that all of our selected predictors in the final model are a good estimator of our target variable selling price.



Potential Problems

Data: The data had very few data points for few of the levels in the categorical variables like fuel and owner. For our estimation we combined the levels with few data points into a common level called others. But if the data set had an approximately even number of data sets for each level we can get a better level wise estimation .

Analysis:
In our final model, we can still see heteroscedasticity which was not fixed after basic transformation and slight normality assumption is violated. The overall model is a good model with an average r squared value.

Future Analysis

By looking at our final model, we would like to see what other factors impact the selling price of the used cars. Since our model's r square value is average we would want to get a better dataset with more influential predictors to get a better model which gives a very accurate prediction of the actual selling price. We would also want to remove existing problems in the model.

We can apply feature engineering and other machine learning models like Random Forest to improve the model performance.

Conclusion

This project is aimed to build a linear regression model which can predict the best selling price using a bunch of predictors we select and construct from the dataset.

To realize this goal, we first performed a basic exploratory data analysis: We detected the missing values, reconstructed some categorical variables by combining some levels, and also visualized the relationship between predictors and response variables.

After that, We fit an initial model using all predictors we had here. And checked the multicollinearity of the model. At an initial glance, the model worked well with an adjusted R squared value 0.457. So, our next step was performing the model diagnosis to check the effectiveness of the model and also trying to decide the predictors we decided to remain in the model.

Our model diagnosis included checking heteroscedasticity, checking influential points, and checking normality. To solve the heteroscedasticity within our model, we did log transformation on y. We also detected some influential points, and reported the results both with and without influential points.

Finally, we fitted the model again after all the adjustments we did before. Checking the p value of all the predictors we have, We concluded all of them were significant to the model and we should keep all of them. Our final model is log_price ~ km_driven + transmission + seller_type + owner + fuel + year_used.