Regression to Predict Car Prices

Rishi Mohan
Brandon Hom
David Ramirez

MSDS 601 Fall 2023
Prof. Shan Wang

INTRODUCTION AND EDA

Our group chose to use a dataset of vehicles for sale in India. The dataset includes some attributes about the vehicle as follows:

- Vehicle name - includes manufacturer, model, etc.
- Price - originally in Indian rupees
- Odometer - originally in kilometers
- Fuel type - categorical, includes petrol, diesel, electric, LPG, CNG
- Transmission type - categorical, includes manual and automatic
- Ownership history - categorical, either new or *N* previous owners
- Manufacture year - year of vehicle manufacture
- Engine displacement - measured in cubic centimeters
- Seats - number of seats in the vehicle

This dataset was retrieved from Kaggle and was originally scraped from a car listings website. The dataset has 5512 entries before any tailoring or cleaning was performed.

We chose to extract the manufacturer from each vehicle's name as an additional potential predictor, and we also converted the price (originally in Indian rupees, or INR to US dollars). This conversion is not strictly necessary, but helps make the results more interpretable from our perspective and also handles the difference in standard magnitudes used in measurements (e.g. lakh = 100,000, and crore = 10,000,000). We also extracted the numerical portion of Seats to treat it as a numerical predictor.

The goal of this analysis is to identify and select a model to predict the vehicle's price based on the other information we have about its attributes. To perform this, we will follow the multiple linear regression process, evaluate our model assumptions, perform model diagnosis checks, and select a final model that we find the most appropriate to match our assumptions. We start with some brief exploratory data analysis to observe some potential issues with the data and distributions.

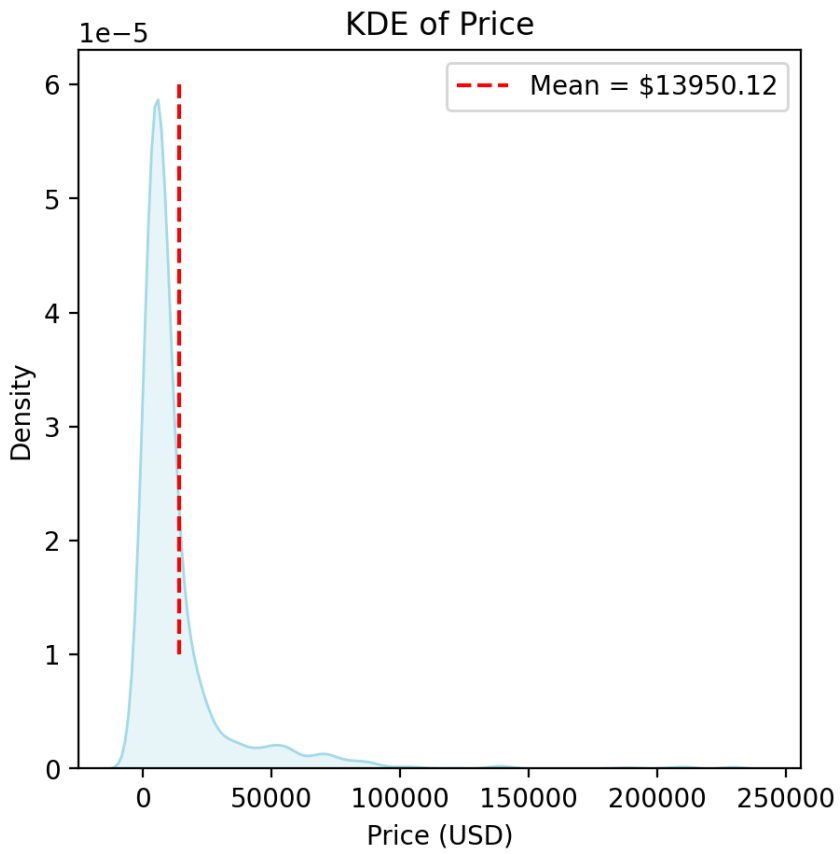

Figure 1: Kernel Density Estimate plot of Price, after conversion to USD

The KDE plot of our price variable clearly shows a strong left skew to our distribution. We likely have a couple vehicles with extremely high prices that are causing this bias, so we can consider tailoring the dataset or splitting the dataset to consider more valuable 'luxury' vehicles and more typical 'workhorse' or 'family'-oriented vehicles.

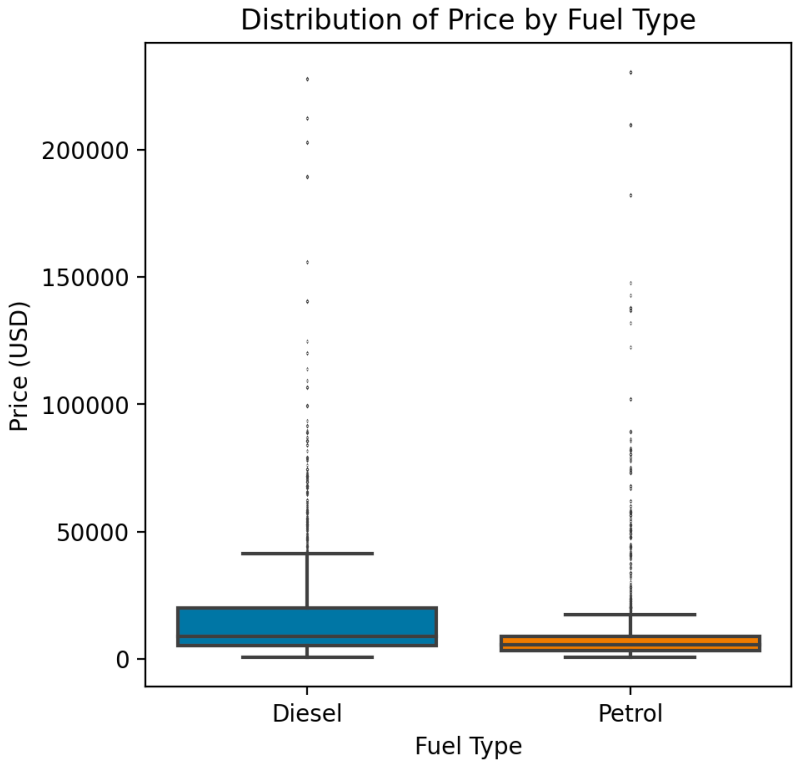

Figure 2: Distribution of Price by Fuel Type

The boxplots of price by fuel type show that there is a wider distribution of prices for diesel-powered vehicles, and that the median prices for diesel and petrol are somewhat different. There are also a substantial number of outliers for both diesel and petrol datatypes. We omitted the other fuel types in this dataset from this plot (CNG, LPG, and electric) since there are a negligible number of observations relative to diesel and petrol.

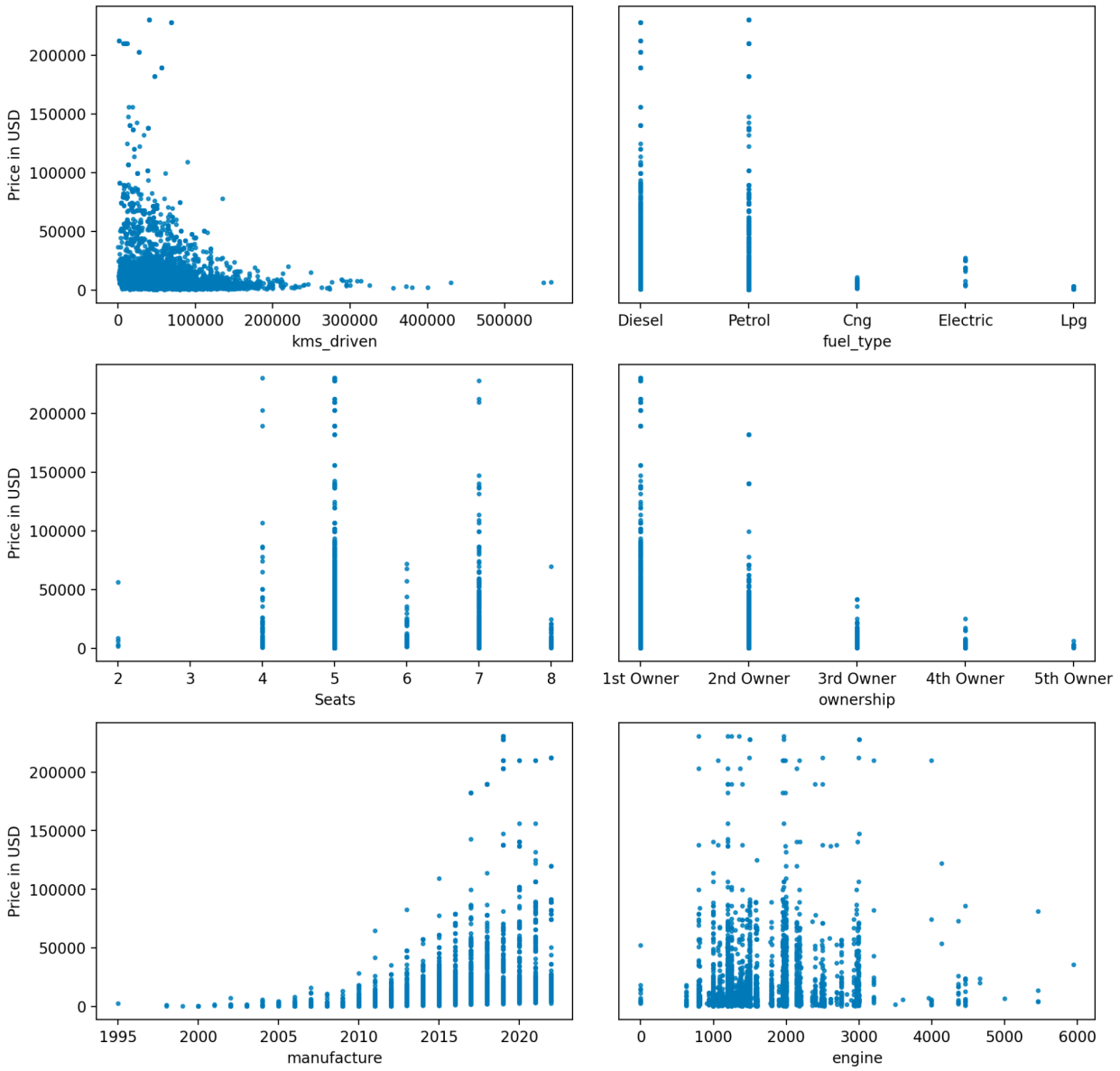

Figure 3: Potential predictors vs. Price in USD

We also plot the relationships between some of our potential predictors and the response variable of interest (price) to observe the extent of their potential relationships. We see a basic relationship in several cases, but nothing extremely compelling. We also observe that some variables may be contributing similar information, such as manufacture year and odometer reading.

REGRESSION ANALYSIS

We first regressed the price (USD) against fuel type, transmission, ownership, manufacture, engine and seats. We then performed model diagnostics to check for any violation of assumptions and influential points. From the results of the diagnostics we then decide to perform a log transformation on our price and fit the model again. The results are then given in the next section.

MODEL DIAGNOSIS

Multicollinearity

We checked multicollinearity by two different methods. First we plot a heatmap of the correlation matrix (Figure 4), and then we calculate the VIF factor for each predictor.

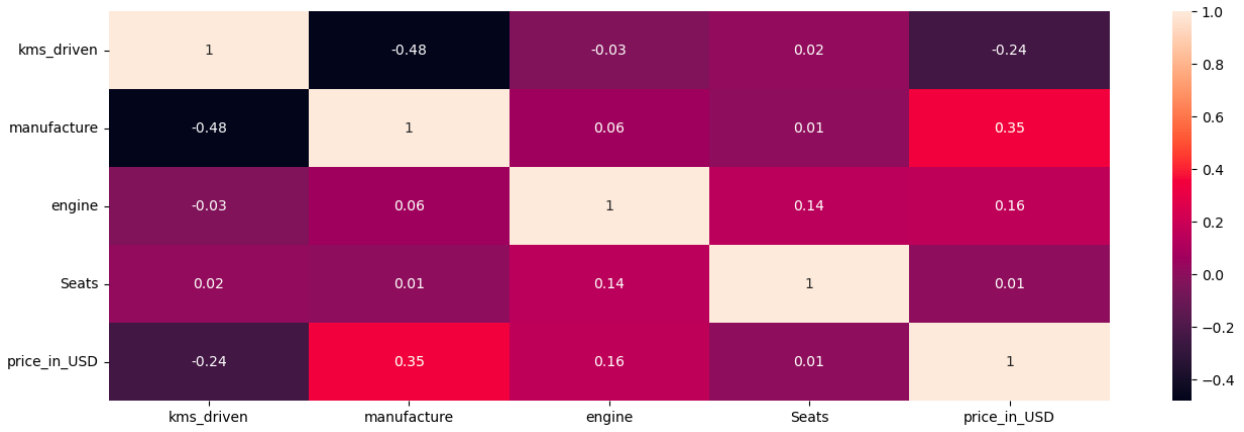

Figure 4: Heatmap of correlation matrix

We identified a strong correlation between kms_driven and manufacture predictors. We decided to remove kms_driven to eliminate the multicollinearity.

	VIF Factor	features
0	421224.169097	Intercept
1	17.759448	C(fuel_type)[T.Diesel]
2	1.186813	C(fuel_type)[T.Electric]
3	1.358954	C(fuel_type)[T.Lpg]
4	17.760131	C(fuel_type)[T.Petrol]
5	1.156486	C(transmission)[T.Manual]
6	1.158506	C(ownership)[T.2nd Owner]
7	1.145112	C(ownership)[T.3rd Owner]
8	1.052488	C(ownership)[T.4th Owner]
9	1.012981	C(ownership)[T.5th Owner]
10	1.504608	kms_driven
11	1.588007	manufacture
12	1.058015	engine
13	1.024558	Seats

Figure 5: VIF Factor results

VIF (Figure 5) showed multicollinearity on fuel_type [T.Diesel] and fuel_type [T.Petrol]. Furthermore, we had fitted a logistic regression model on the fuel types to see if any of the predictors were correlated with it based on the Wald test. As seen from the output below, there were significant predictors for petroleum, which indicated a multicollinearity problem. It is because those categories are the most common, then it seems to be extremely related. After analyzing this issue we decided to keep fuel_type on the model.

Dep. Variable:	is_petrol	**No. Observations:**	5505
Model:	GLM	**Df Residuals:**	5496
Model Family:	Binomial	**Df Model:**	8
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3384.0
Date:	Wed, 11 Oct 2023	**Deviance:**	6768.0
Time:	13:34:32	**Pearson chi2:**	4.57e+05
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.1403
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	316.0912	20.668	15.294	0.000	275.582	356.600
ownership[T.2nd Owner]	-0.4184	0.074	-5.681	0.000	-0.563	-0.274
ownership[T.3rd Owner]	-0.1539	0.127	-1.208	0.227	-0.404	0.096
ownership[T.4th Owner]	0.1431	0.251	0.571	0.568	-0.348	0.634
ownership[T.5th Owner]	0.2617	0.648	0.404	0.686	-1.007	1.531
engine	-0.0003	5.15e-05	-6.057	0.000	-0.000	-0.000
Seats	-0.1382	0.041	-3.393	0.001	-0.218	-0.058
manufacture	-0.1553	0.010	-15.179	0.000	-0.175	-0.135
kms_driven	-2.522e-05	1.09e-06	-23.207	0.000	-2.74e-05	-2.31e-05

Figure 6: First model results

Influential points

We used two different methods to identify and eliminate influential points. First we used an external studentized error method and then the Cook's distance method. We identified 82 points present on the results of both methods. Considering that we don't have access to the original source of the dataset, It was assumed that those points were real outliers (with fake information), therefore those 82 points were removed.

Heteroscedasticity

We plot residuals Vs fitted values (Figure 7) and performed a Breusch-Pagan test to evaluate the potential heteroscedasticity present in the model. The test finds a p-value below a

reasonable α=0.05, leading us to conclude significant evidence of heteroscedasticity. In order to eliminate heteroscedasticity we perform natural-log and Box-Cox transformations on the dependent variable to reduce the variance; however, none of those transformations eliminated heteroscedasticity (Figures 8 and 9). Considering that our t-test results would not be reliable because of the variability on error variance, we calculated model parameters using HC0 robust standard errors.

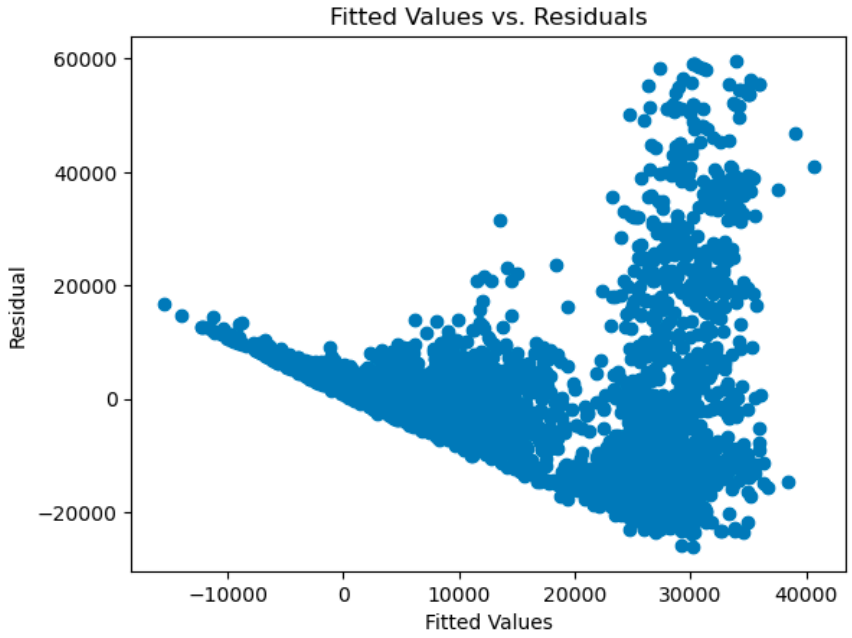

Figure 7: Residuals (Before Transformations)

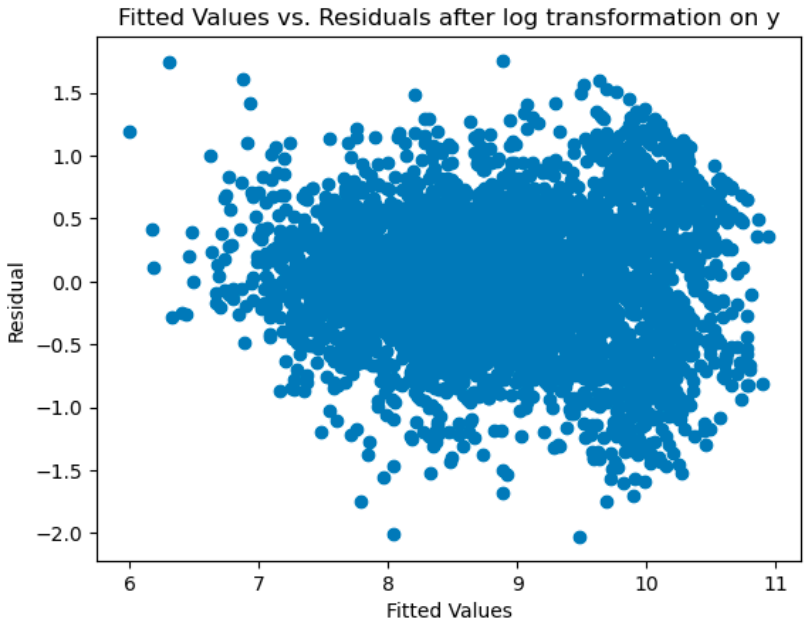

Figure 8: Residuals (After Log Transformation)

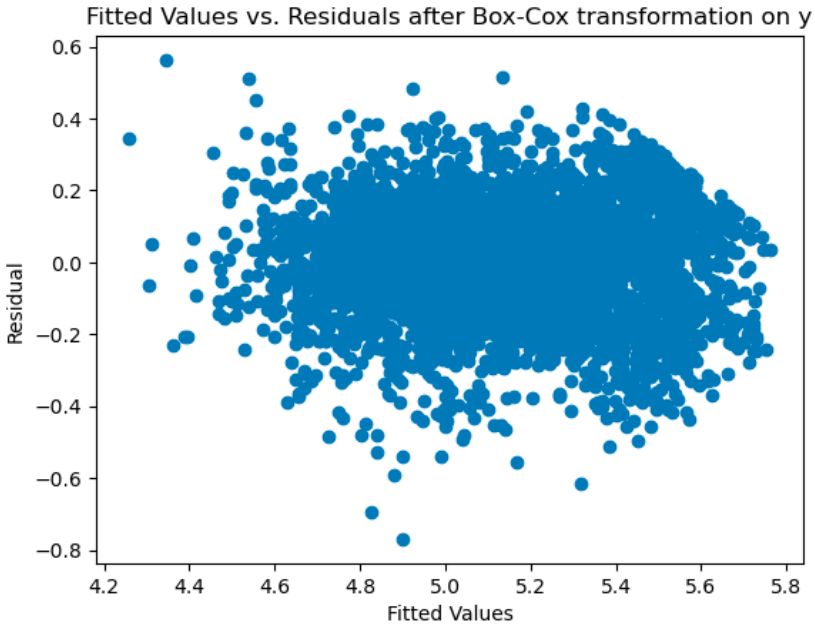

Figure 9: Residuals (After Box-Cox Transformation)

Normality

We generated QQ plots and performed the Jarque-Bera test (Figure 10). The test finds a p-value below a reasonable $\alpha = .05$, leading us to conclude that the skewness and excess kurtosis of the distribution of the residuals is significantly different from 0; the distribution of the residuals is not normal. In order to mitigate this somewhat we perform natural-log and Box-Cox transformations on the dependent variable; however, despite the improvement transformations produced, neither of those transformations completely eliminated non-normality (Figures 11-14).

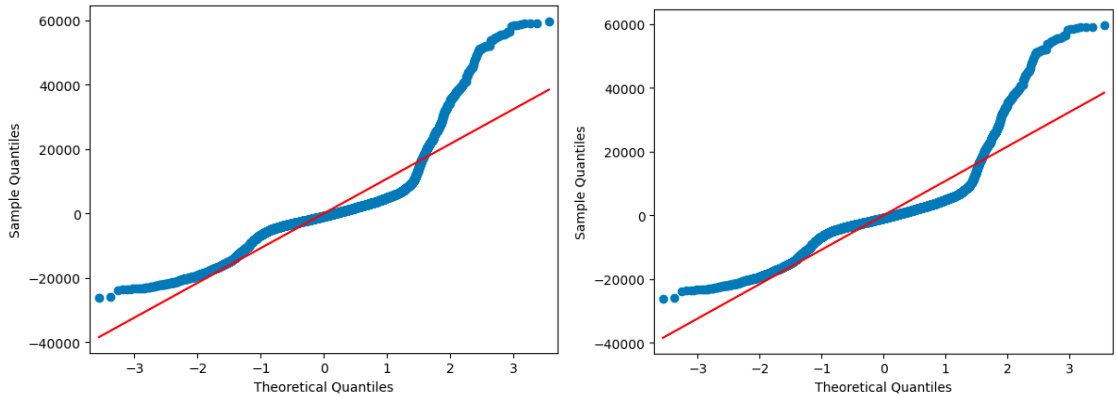

Figure 10: Q-Q Plots (Before Transformation)

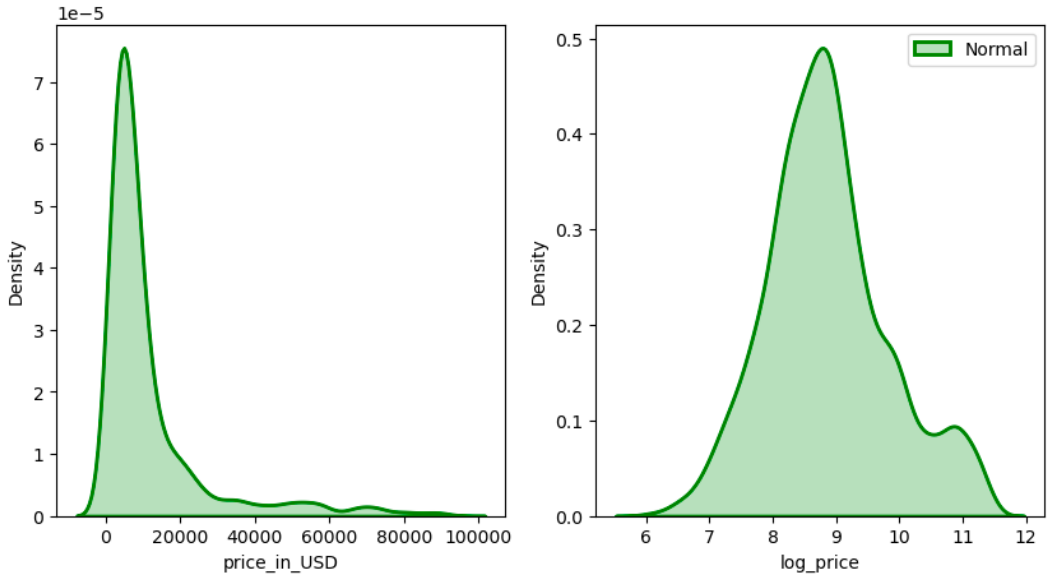

Figure 11: Dependent variable distribution (Before and after Log Transformation)

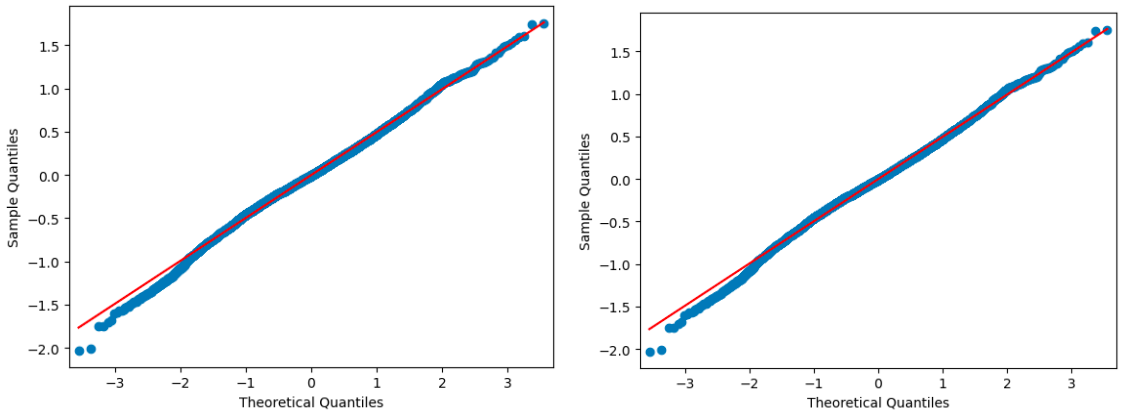

Figure 12: Q-Q Plots (After Log Transformation)

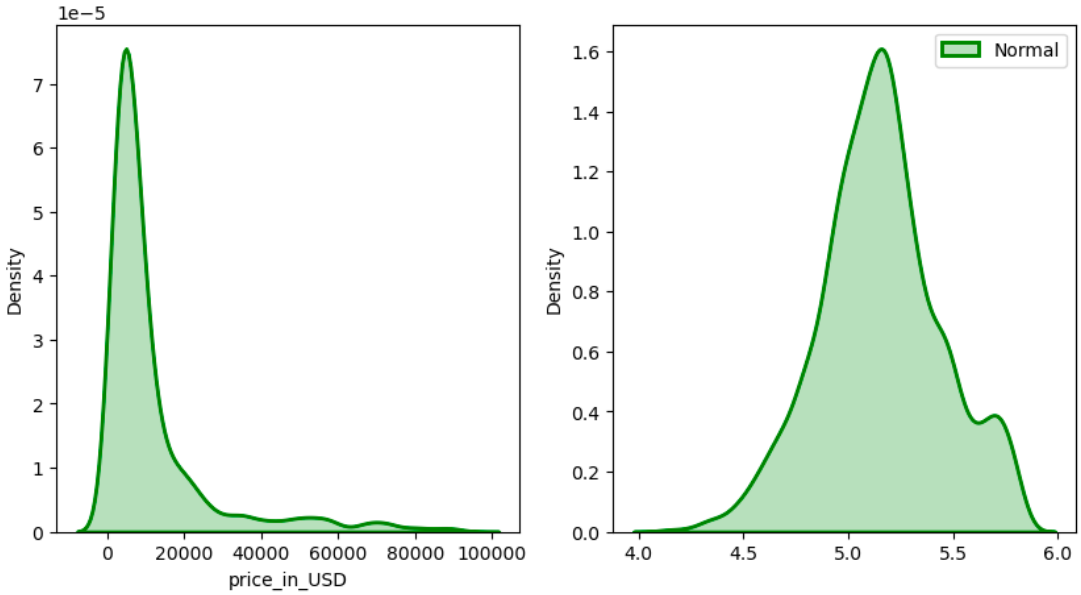

Figure 13: Dependent variable distribution (Before and after Box-Cox Transformation)

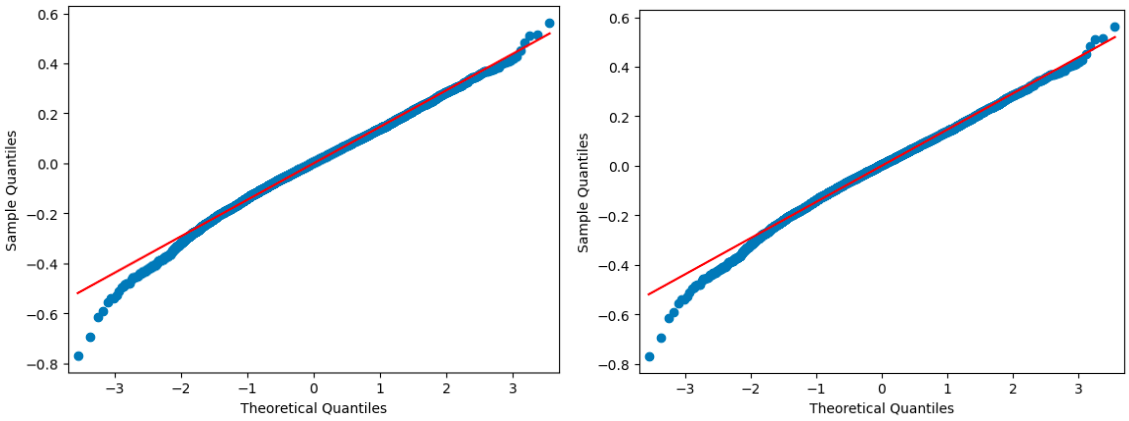

Figure 14: Q-Q Plots (After Box-Cox Transformation)

Non–linearity between predictors and dependent variable

We generated pairwise scatter plots between price and predictors. Linearity appeared present in most of the predictors, but Seats appeared to be less of a linear relationship and more bell-shaped; centered around 5.

MODEL SELECTION PROCESS

In order to obtain the best possible linear model we perform an OLSE modeling iterative process. The process consisted on 4 iterations:

1. Price ~ manufacture + Seats + engine + fuel_type + transmission + ownership (OLSE with all the registers)
2. Price ~ manufacture + Seats + engine + fuel_type + transmission + ownership (OLSE without influential points)
3. ln(Price) ~ manufacture + Seats + engine + fuel_type + transmission + ownership (OLSE without influential points and a natural-log transformation on Price)
4. Box-Cox(Price) ~ manufacture + Seats + engine + fuel_type + transmission + ownership (OLSE without influential points and a Box-Cox transformation on Price)

However, considering that none of those models pass the heteroscedasticity tests, we found it necessary to perform a 5th regression using HC0 Robust Standard Errors.

5. Box-Cox(Price) ~ manufacture + Seats + engine + fuel_type + transmission + ownership (Without influential points and a Box-Cox transformation on Price, using Robust Standard Errors)

Box-Cox transformation on Price was chosen for this final iteration because it was the one with better performance on the first 4 iterations.

	0	1	2	3
0	Model:	OLS	Adj. R-squared:	0.746
1	Dependent Variable:	box_price	AIC:	-5457.8274
2	Date:	2023-10-11 15:07	BIC:	-5372.0482
3	No. Observations:	5423	Log-Likelihood:	2741.9
4	Df Model:	12	F-statistic:	1150.
5	Df Residuals:	5410	Prob (F-statistic):	0.00
6	R-squared:	0.746	Scale:	0.021350

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	-75.112	1.279	-58.724	0.000	-77.620	-72.605
C(fuel_type)[T.Diesel]	0.177	0.012	15.218	0.000	0.154	0.200
C(fuel_type)[T.Electric]	-0.113	0.055	-2.054	0.040	-0.220	-0.005
C(fuel_type)[T.Lpg]	-0.077	0.027	-2.889	0.004	-0.129	-0.025
C(fuel_type)[T.Petrol]	0.038	0.012	3.280	0.001	0.015	0.060
C(transmission)[T.Manual]	-0.290	0.005	-52.850	0.000	-0.301	-0.279
C(ownership)[T.2nd Owner]	-0.017	0.005	-3.366	0.001	-0.027	-0.007
C(ownership)[T.3rd Owner]	-0.059	0.008	-6.955	0.000	-0.075	-0.042
C(ownership)[T.4th Owner]	-0.106	0.019	-5.660	0.000	-0.142	-0.069
C(ownership)[T.5th Owner]	-0.062	0.044	-1.418	0.156	-0.149	0.024
manufacture	0.040	0.001	62.840	0.000	0.039	0.041
Seats	0.006	0.003	2.010	0.045	0.000	0.011
engine	0.000	0.000	6.187	0.000	0.000	0.000

Figure 15: Robust standard error results

The resulting model was checked using ANOVA (Typ=1) with robust parameters, concluding that the best model was the full model:

box_price ~ manufacture + Seats + engine + fuel_type + transmission + ownership

	df	sum_sq	mean_sq	F	PR(>F)
C(fuel_type)	4.0	50.149762	12.537440	587.227884	0.000000e+00
C(transmission)	1.0	150.386164	150.386164	7043.778095	0.000000e+00
C(ownership)	4.0	36.598955	9.149739	428.554918	2.875462e-321
manufacture	1.0	101.568153	101.568153	4757.243029	0.000000e+00
Seats	1.0	0.182762	0.182762	8.560210	3.450187e-03
engine	1.0	0.899324	0.899324	42.122475	9.337615e-11
Residual	5410.0	115.504653	0.021350	NaN	NaN

Figure 16: ANOVA results

KEY FINDINGS AND RESULTS

The best model to describe cars price is:

box_price ~ manufacture + Seats + engine + fuel_type + transmission + ownership

Te equation that describes box_price is

$$
\begin{aligned}
\text{Box\_price} = {} & \beta_0 \\
& + \beta_1 \text{ manufacture} \\
& + \beta_2 \text{ Seats} \\
& + \beta_3 \text{ engine} \\
& + \beta_4 \text{ fuel\_type [T.Diesel]} \\
& + \beta_5 \text{ fuel\_type [T.Electric]} \\
& + \beta_6 \text{ fuel\_type [T.Lgp]} \\
& + \beta_7 \text{ fuel\_type [T.Petrol]} \\
& + \beta_8 \text{ transmission [T.Manual]} \\
& + \beta_9 \text{ ownership [T.2nd Owner]} \\
& + \beta_{10} \text{ ownership [T.3rd Owner]} \\
& + \beta_{11} \text{ ownership [T.4th Owner]} \\
& + \beta_{12} \text{ ownership [T.5th Owner]}
\end{aligned}
$$

Where:

$$box\_price = (Price^{-0.13491936524519552} - 1) \div (-0.13491936524519552)$$

B_0 = -75.11211

B_1 = 0.03985

B_2 = 0.00574

B_3 = 0.00002

B_4 = 0.17727

B_5 = -0.11258

B_6 = -0.07678

B_7 = 0.03779

B_8 = -0.28986

B_9 = -0.01724

B_{10} = -0.05869

B_{11} = -0.10555

B_{12} = -0.06245

Important considerations:

1. The model was obtained performing a Box-Cox transformation on the price with λ = -0.13491936524519552
2. Despite the Box-Cox transformation, the residuals have significant evidence of heteroscedasticity and the dependent variable has significant evidence of non-normality.

Other possible predictors

Kms_driven was removed as a predictor because it was highly correlated with the manufacture year predictor.

SUMMARY

After regressing price against fuel type, transmission, ownership, seats, manufacture and engine some of the assumptions were violated. Even with a natural log transformation on price, heteroskedasticity and normality were still violated, so in the end we had to use the robust standard errors. From figure 16, we can see that all of the features that we regressed are actually significant as most of the values in the PR(>F) column are all close to zero. This means that they all have a significant impact on price. From figure 15, we can see that electric and liquified petroleum gas decrease the price compared to the reference level of compressed natural gas, while diesel and petroleum increase the price. If the car has a manual transmission, the price will also decrease. For ownership, we can see that all of the ownership levels cause a decrease in price compared to the reference of being a first owner. Finally for our numerical variables of seats, engine and manufacture, a one unit increase leads to an increase in price. Overall, all of

our features that we picked of fuel type transmission, ownership, seats, manufacture and engine are all significantly associated with price.

POTENTIAL PROBLEMS

One of the major problems of this dataset is that it consists of cars from different populations. Since we have a mixing of populations, the assumptions of linear regression are more likely to be violated, such as constant variance. Despite the Box-Cox transformation, the residuals have significant evidence of heteroscedasticity and the dependent variable has significant evidence of non-normality, signifying that our t-tests are not reliable.