Team Assignment – Toyota Corolla dataset

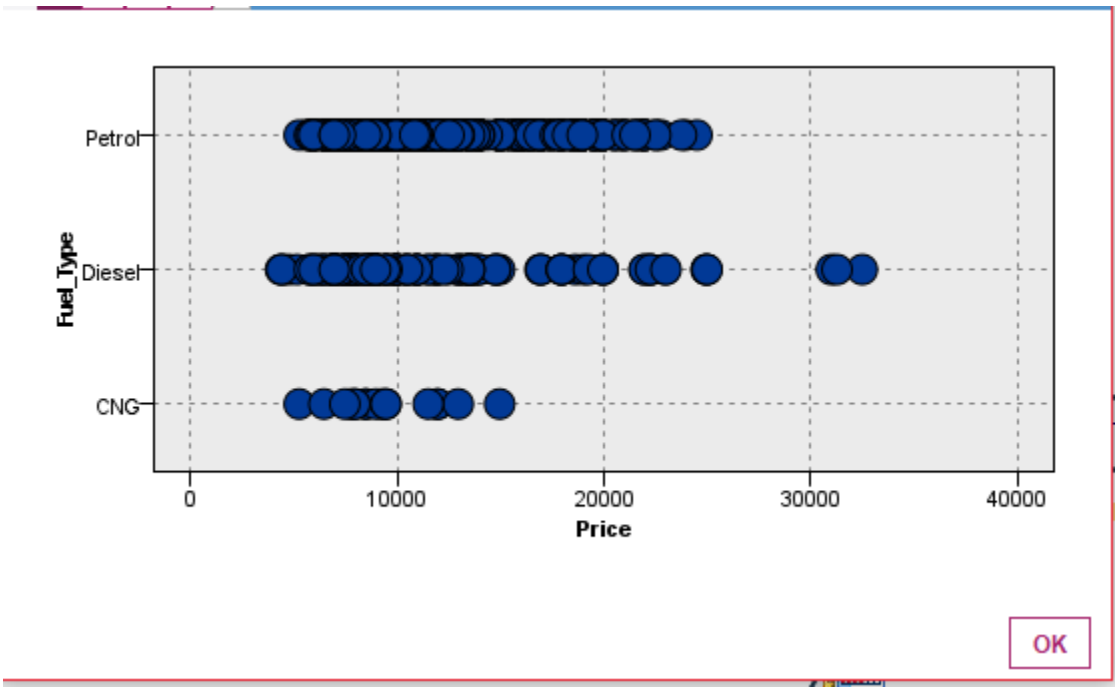
A Toyota dealership offers its clients the option to trade-in their used Corollas when they purchase a new car. To determine the trade-in price, the dealership has decided to collect data on all previous trade-ins. The dataset contains the characteristics of the traded Corollas and the price the dealership paid for them. The file ToyotaCorolla.xlsx contains info on 1436 purchases of used Corollas.

Before getting into the assessment, analyzing the data by data audit node, we found some outliers in few fields but as they do not seem out of the data, we considered the given data as it is.

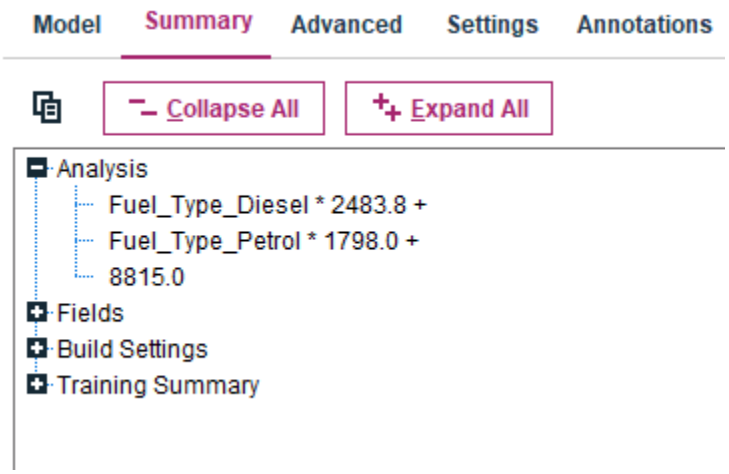
Assignment

1. Explore the price in relation to the fuel type.

On the data, a graph is plotted between price and the fuel type. The graph showed that, compared to petrol and diesel, CNG has low effect on the price of Corollas.

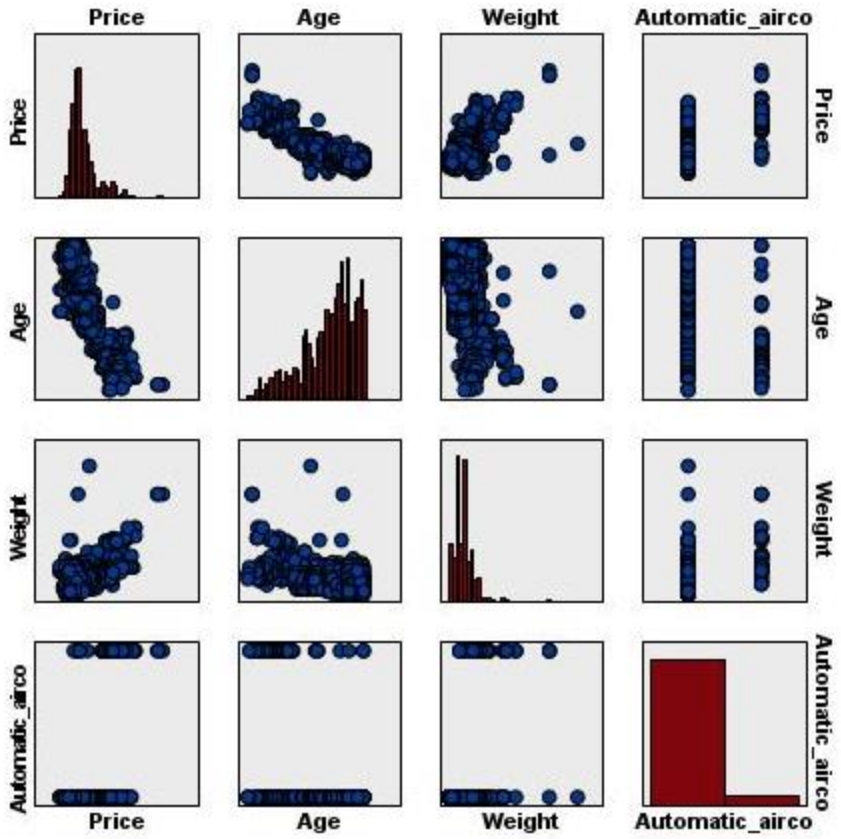


Since fuel is a discrete value, it needs to be restructured. Flag values are allotted after restructuring fuel type to petrol, diesel and CNG. Considering CNG as a reference value the price vary with petrol and diesel fuel types as the equation: 8815+fuel type diesel*2483.8+fuel type petrol*1798



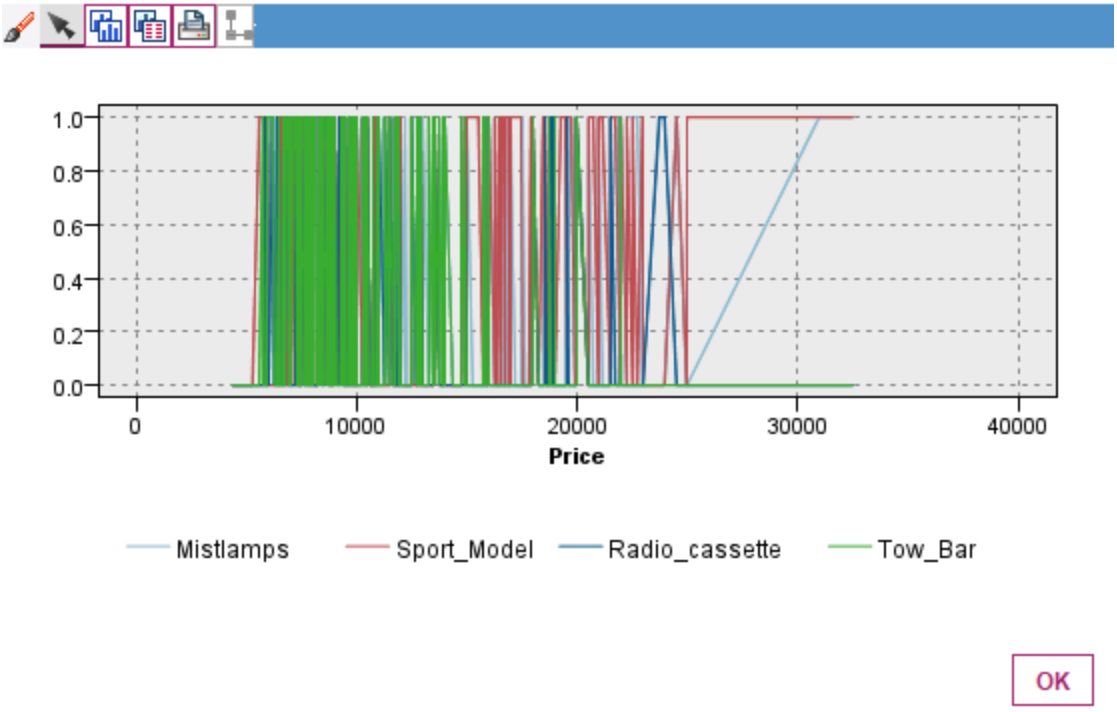
2. Explore the relationship between price and each of the continuous predictors. Does there seem to be a linear relationship? (Note: if you plan to use a scatterplot matrix, do not plot more than 4 or 5 variables together, as the plot consumes plenty of memory; create more than one scatterplot to visualize the variables)

On the given data, a scatter plot matrix is plotted between price, age, weight and automatic airco.



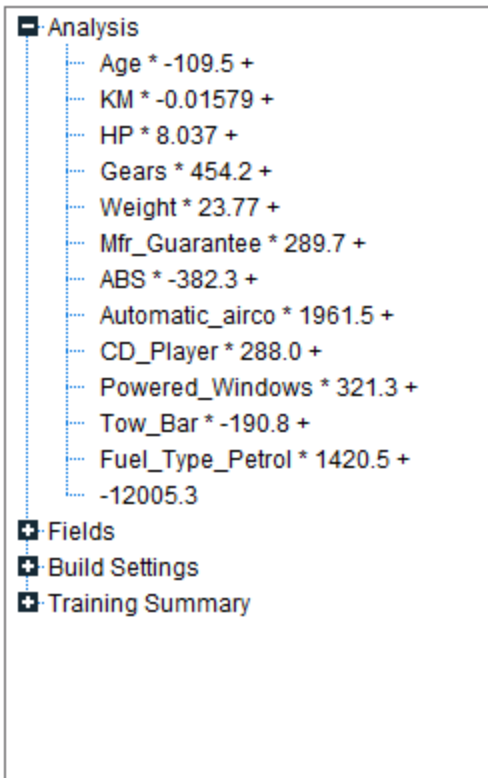
The picture depicts that there is a linear relation between price and age, price and weight mostly.

We tried to produce more scatter plots but we got only blank pictures. So we tried with multiplot node



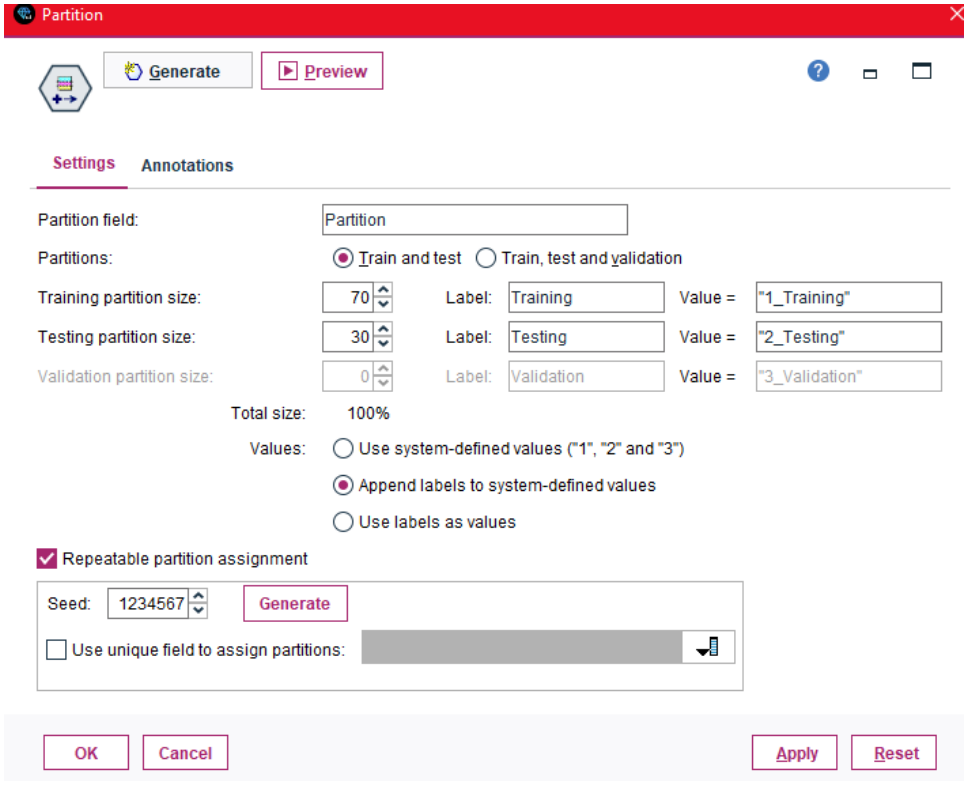
Individual relation depiction is not clear here, so considering the linear regression equation,

Price= Fuel type*1420.5+ Age*-109.5+KM*-0.01579+HP*8.037+gears*454.2+weight*23.77+Mfr_guarantee*289.7+ABS*-382.3+Automatic_airco*1961.5+CD_Player*288+Powered_Window*321.3+Tow_Bar*-190.8-12005.3.

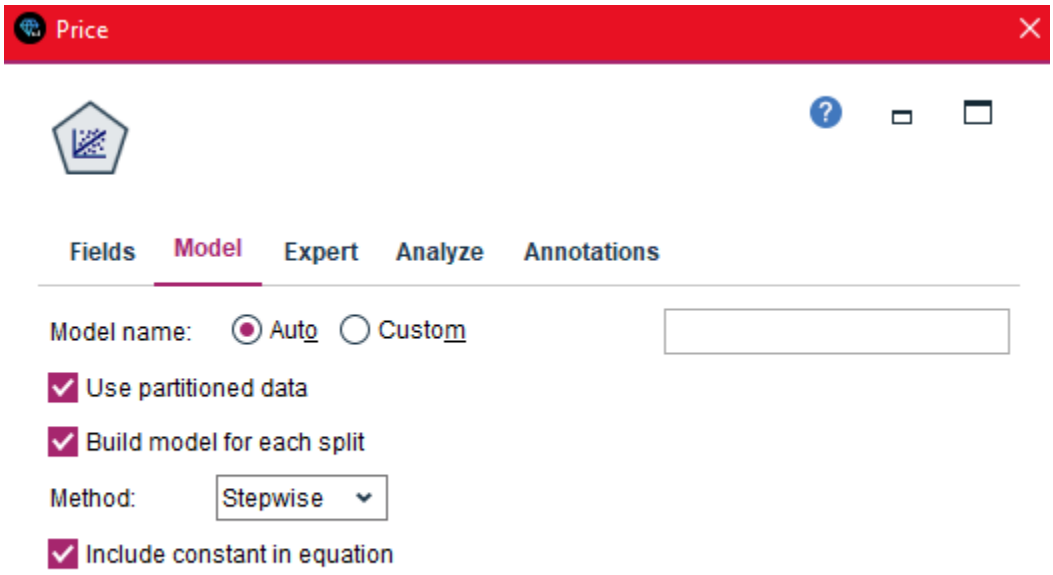


3. To fit a predictive model for price of used cars:

i. Partition the dataset into training and testing data sets.



ii. Use stepwise regression to reduce the number of predictors using the training data



iii. Report the model parameters: regression coefficient estimates with their standard errors, goodness of fit metrics (R-squared, adjusted R-squared), standard error of the estimate (s), t-test values (scores and p-values), F-test values (F-score and p-value)

The model parameters are as follows

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20231.860	175.661		115.176	.000
	Age	-170.446	2.978	-.876	-57.243	.000
2	(Constant)	18942.958	175.433		107.978	.000
	Age	-151.027	2.909	-.776	-51.921	<.001
	Automatic_airco	3910.064	241.367	.242	16.200	<.001
3	(Constant)	15345.071	363.827		42.177	<.001
	Age	-148.806	2.752	-.764	-54.065	<.001
	Automatic_airco	3422.118	231.982	.212	14.752	<.001
	HP	34.484	3.105	.147	11.105	<.001
4	(Constant)	964.393	1226.697		.786	.432
	Age	-135.100	2.803	-.694	-48.195	<.001
	Automatic_airco	2643.977	225.661	.164	11.717	<.001
	HP	36.036	2.900	.154	12.425	<.001
	Weight	12.599	1.033	.177	12.200	<.001
5	(Constant)	-2812.880	1140.557		-2.466	.014
	Age	-112.301	2.973	-.577	-37.776	<.001
	Automatic_airco	2351.043	205.425	.146	11.445	<.001
	HP	25.123	2.730	.107	9.203	<.001
	Weight	17.236	.987	.242	17.461	<.001
	KM	-.020	.001	-.202	-14.758	<.001
6	(Constant)	-10776.931	1636.174		-6.587	<.001
	Age	-112.022	2.910	-.575	-38.492	<.001
	Automatic_airco	2106.945	204.398	.130	10.308	<.001
	HP	9.795	3.527	.042	2.777	.006
	Weight	24.589	1.467	.345	16.758	<.001
	KM	-.016	.001	-.165	-11.362	<.001
	Fuel_Type_Petrol	1583.387	237.800	.145	6.658	<.001
7	(Constant)	-9840.227	1636.910		-6.011	<.001
	Age	-110.058	2.922	-.565	-37.668	<.001
	Automatic_airco	2099.176	202.654	.130	10.358	<.001
	HP	8.563	3.509	.037	2.440	.015
	Weight	23.683	1.470	.332	16.110	<.001
	KM	-.016	.001	-.168	-11.673	<.001
	Fuel_Type_Petrol	1440.249	238.132	.132	6.048	<.001

	KM	-.016	.001	-.168	-11.673	<.001
	Fuel_Type_Petrol	1440.249	238.132	.132	6.048	<.001
	Powered_Windows	352.284	82.502	.049	4.270	<.001
8	(Constant)	-10023.834	1627.628		-6.159	<.001
	Age	-109.179	2.914	-.561	-37.470	<.001
	Automatic_airco	2106.875	201.419	.130	10.460	<.001
	HP	8.032	3.490	.034	2.301	.022
	Weight	23.762	1.461	.333	16.262	<.001
	KM	-.016	.001	-.165	-11.459	<.001
	Fuel_Type_Petrol	1393.570	237.014	.127	5.880	<.001
	Powered_Windows	355.223	81.998	.050	4.332	<.001
	Mfr_Guarantee	284.869	78.126	.039	3.646	<.001
9	(Constant)	-9211.044	1637.635		-5.625	<.001
	Age	-113.038	3.122	-.581	-36.210	<.001
	Automatic_airco	2072.555	200.663	.128	10.329	<.001
	HP	8.764	3.479	.037	2.519	.012
	Weight	23.417	1.458	.328	16.066	<.001
	KM	-.016	.001	-.162	-11.346	<.001
	Fuel_Type_Petrol	1368.606	235.932	.125	5.801	<.001
	Powered_Windows	349.487	81.601	.049	4.283	<.001
	Mfr_Guarantee	301.780	77.895	.042	3.874	<.001
	ABS	-359.451	107.837	-.039	-3.333	<.001
10	(Constant)	-9401.786	1634.577		-5.752	<.001
	Age	-110.164	3.304	-.566	-33.340	<.001
	Automatic_airco	2019.324	201.136	.125	10.040	<.001
	HP	8.999	3.471	.038	2.593	.010
	Weight	23.387	1.453	.328	16.092	<.001
	KM	-.016	.001	-.162	-11.376	<.001
	Fuel_Type_Petrol	1357.039	235.295	.124	5.767	<.001
	Powered_Windows	336.264	81.526	.047	4.125	<.001
	Mfr_Guarantee	290.047	77.803	.040	3.728	<.001
	ABS	-357.735	107.528	-.039	-3.327	<.001
	CD_Player	274.398	105.850	.032	2.592	.010
11	(Constant)	-9612.626	1633.323		-5.885	<.001
	Age	-109.256	3.319	-.561	-32.915	<.001
	Automatic_airco	1987.597	201.132	.123	9.882	<.001
	HP	9.485	3.469	.041	2.734	.006
	Weight	23.529	1.451	.330	16.212	<.001
	KM	-.016	.001	-.162	-11.420	<.001

	Weight	23.529	1.451	.330	16.212	<.001
	KM	-.016	.001	-.162	-11.420	<.001
	Fuel_Type_Petrol	1371.870	234.844	.125	5.842	<.001
	Powered_Windows	343.967	81.406	.048	4.225	<.001
	Mfr_Guarantee	288.740	77.627	.040	3.720	<.001
	ABS	-357.825	107.283	-.039	-3.335	<.001
	CD_Player	269.820	105.627	.031	2.554	.011
	Tow_Bar	-200.399	85.326	-.025	-2.349	.019
12	(Constant)	-12005.290	1961.505		-6.120	<.001
	Age	-109.512	3.315	-.563	-33.035	<.001
	Automatic_airco	1961.547	201.096	.121	9.754	<.001
	HP	8.037	3.525	.034	2.280	.023
	Weight	23.769	1.453	.333	16.362	<.001
	KM	-.016	.001	-.163	-11.468	<.001
	Fuel_Type_Petrol	1420.483	235.437	.130	6.033	<.001
	Powered_Windows	321.302	81.904	.045	3.923	<.001
	Mfr_Guarantee	289.731	77.479	.040	3.739	<.001
	ABS	-382.290	107.656	-.041	-3.551	<.001
	CD_Player	287.975	105.748	.034	2.723	.007
	Tow_Bar	-190.798	85.274	-.024	-2.237	.025
	Gears	454.223	207.093	.024	2.193	.029

From the above table, H0 can be rejected for T values as they are mostly at smaller confidence level

Goodness of fit and standard error of estimates

R2 increased by adding additional predictors. It implies the percentage of variability in price is accounted for linear relationships with other parameters

The Radj<R2 in all the parameters, hence omitting may be considered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.876[a]	.767	.766	1718.848136
2	.903[b]	.815	.815	1529.926938
3	.914[c]	.836	.835	1443.826931
4	.926[d]	.857	.857	1347.212258
5	.940[e]	.883	.882	1220.661119
6	.942[f]	.888	.887	1194.866918
7	.943[g]	.890	.889	1184.621493
8	.944[h]	.891	.890	1177.340320
9	.945[i]	.893	.892	1171.373928
10	.945[j]	.893	.892	1168.001057
11	.945[k]	.894	.893	1165.340739
12	.946[l]	.894	.893	1163.097595

Standard error of estimates

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9680935883	1	9680935883	3276.743	.000[b]
	Residual	2945575598	997	2954438.915		
	Total	1.263E+10	998			
2	Regression	1.030E+10	2	5147598876	2199.193	.000[c]
	Residual	2331313729	996	2340676.434		
	Total	1.263E+10	998			
3	Regression	1.055E+10	3	3517432819	1687.313	.000[d]
	Residual	2074213025	995	2084636.206		
	Total	1.263E+10	998			
4	Regression	1.082E+10	4	2705605125	1490.707	.000[e]
	Residual	1804090982	994	1814980.868		
	Total	1.263E+10	998			
5	Regression	1.115E+10	5	2229385602	1496.218	.000[f]
	Residual	1479583471	993	1490013.566		
	Total	1.263E+10	998			
6	Regression	1.121E+10	6	1868371031	1308.652	.000[g]
	Residual	1416285296	992	1427706.952		
	Total	1.263E+10	998			
7	Regression	1.124E+10	7	1605116193	1143.793	.000[h]
	Residual	1390698129	991	1403328.082		
	Total	1.263E+10	998			
8	Regression	1.125E+10	8	1406780319	1014.898	.000[i]
	Residual	1372268927	990	1386130.229		
	Total	1.263E+10	998			
9	Regression	1.127E+10	9	1252165321	912.579	.000[j]
	Residual	1357023594	989	1372116.880		
	Total	1.263E+10	998			
10	Regression	1.128E+10	10	1127865573	826.744	.000[k]
	Residual	1347855752	988	1364226.470		
	Total	1.263E+10	998			
11	Regression	1.129E+10	11	1026013335	755.522	.000[l]
	Residual	1340364791	987	1358019.039		
	Total	1.263E+10	998			
12	Regression	1.129E+10	12	941054550.9	695.637	.000[m]
	Residual	1333856870	986	1352796.014		
	Total	1.263E+10	998			

Considering the relation between price and fuel type,

R2:6% of variability in price is accounted for an increase(change) in fueltypes

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.080[a]	.006	.004	3549.104400

a. Predictors: (Constant), Fuel_Type_Petrol, Fuel_Type_Diesel

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	80754004.08	2	40377002.04	3.206	.041[b]
	Residual	1.255E+10	996	12596142.05		
	Total	1.263E+10	998			

b. Predictors: (Constant), Fuel_Type_Petrol, Fuel_Type_Diesel

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8815.000	1122.325		7.854	<.001
	Fuel_Type_Diesel	2483.773	1172.231	.219	2.119	.034
	Fuel_Type_Petrol	1798.009	1128.691	.164	1.593	.111

T test is for relation between price and specific predictor fuel type, in presence of other predictors

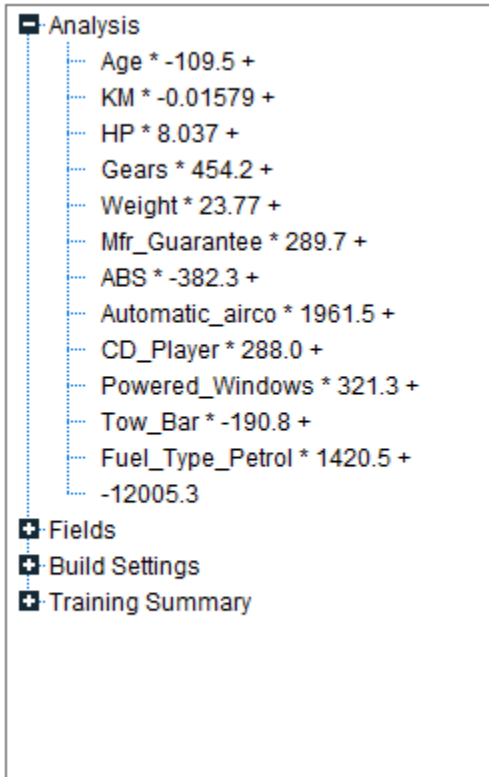
For diesel fuel type confidence level, the regression coefficient not zero is 1-0.034=0.966 96.6%

For petrol fuel type confidence level, 1-0.111=0.889, 88.9%

So null hypothesis is rejected as there is linear relation between price and fuel types at a confidence of 96.6% and 88.9%

For f test, MSR/MSE=3.2, F=3.206 from ANOVA table. degrees of freedom=999-2-1=996 >3.206, we cannot reject H0

iv. Write and explain the regression equation.



The regression equation Price= Fuel type*1420.5+ Age*-109.5+KM*-0.01579+HP*8.037+gears*454.2+weight*23.77+Mfr_guarantee*289.7+ABS*-382.3+Automatic_airco*1961.5+CD_Player*288+Powered_Window*321.3+Tow_Bar*-190.8-12005.3 gives below evaluations

It denotes that a unitary increase in the respective parameters produces an increase or decrease based on + or – of the parametric slopes when remaining are kept constant. That is, with a unitary increase in fuel type, an increase of 1420.5 is observed when remaining parameters are kept constant.

Negative regression coefficients indicates negative relationships.

4. Using the test subset, compute the predictive accuracy metrics (MAE, max, min errors, stddev of the predictive error). What is the typical predictive error that you can expect with this model?

The MAE, min, max errs and std dev are as follows

Results for output field Price

Comparing $E-Price with Price

'Partition'	1_Training	2_Testing
Minimum Error	-7711.27	-13417.609
Maximum Error	5430.892	4683.708
Mean Error	-0.0	-16.443
Mean Absolute Error	857.192	935.714
Standard Deviation	1156.084	1408.287
Linear Correlation	0.946	0.929
Occurrences	999	437

Price

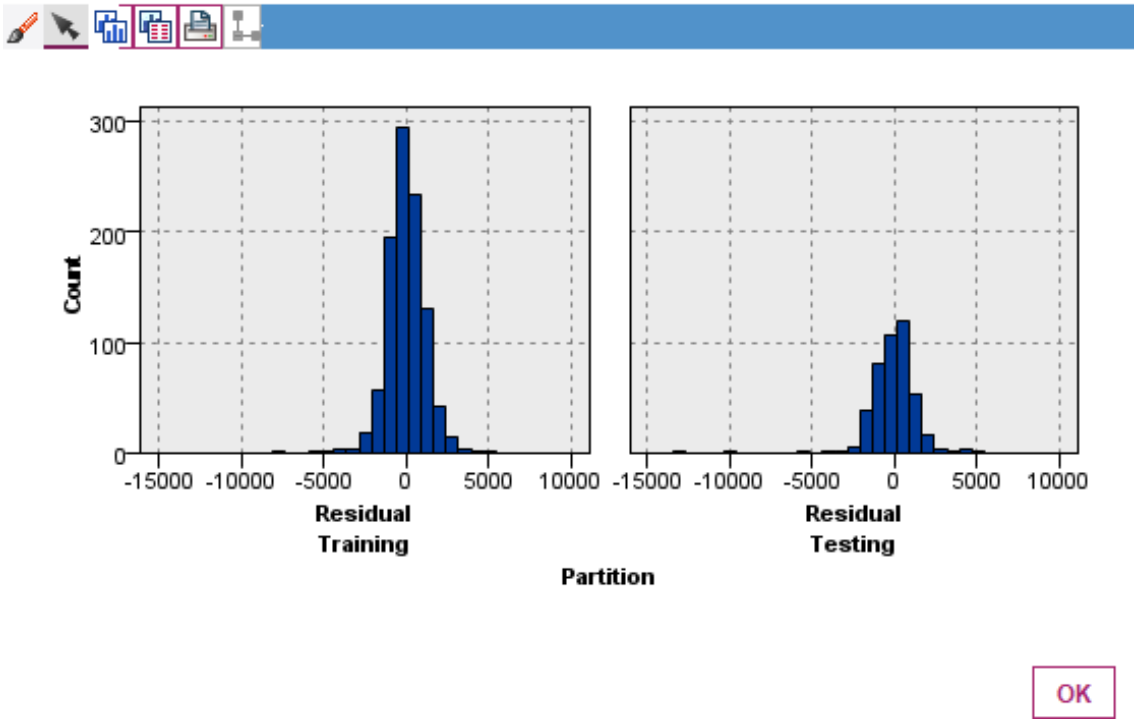
Statistics

Count	1436
Mean	10730.825
Min	4350.000
Max	32500.000
Range	28150.000
Standard Error of Mean	95.712
Median	9900.000

On performance evaluation, we got the above.

The MEA for training and testing vary a lot which means the model is overfitted. The typical error is very high, of approximately (MAE/Mean) 900/10000=9%

5. Create a histogram of the model residuals. Do they appear to follow a normal distribution? How does this affect the predictive performance of the model?



The histogram shows a normal distribution but there are few outliers and it is not skewed.