Assignment 4: Swetha Adike, Venu Goud Raparti, Srujan Namburu

**Team Assignment – Toyota Corolla dataset**

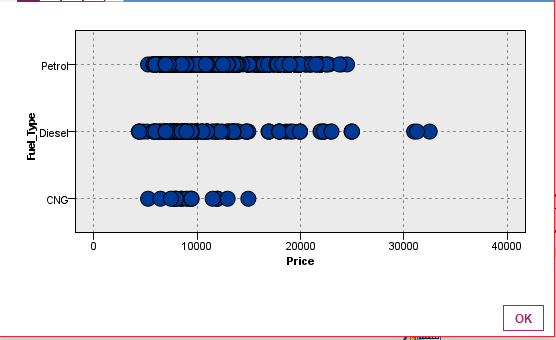
A Toyota dealership offers its clients the option to trade-in their used Corollas when they purchase a new car. To determine the trade-in price, the dealership has decided to collect data on all previous trade-ins. The dataset contains the characteristics of the traded Corollas and the price the dealership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

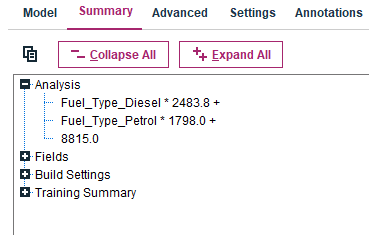
**Assignment**

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.

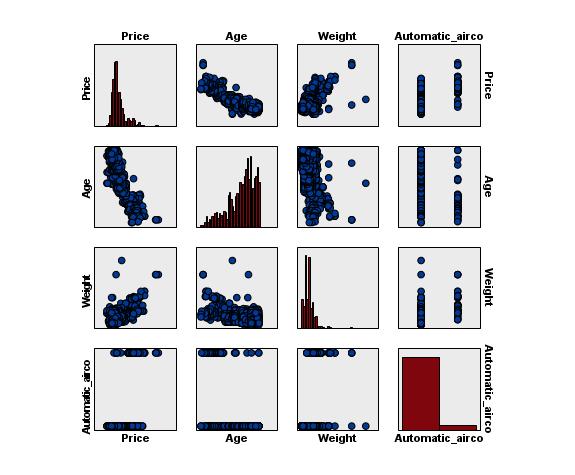


Since fuel is a discrete value, it needs to be restructured. Flag values are allotted after restructuring fuel type to petrol, diesel and CNG. Considering CNG as a reference value the price vary with petrol and diesel fuel types as the equation: 8815+fuel type diesel\*2483.8+fuel type petrol\*1798



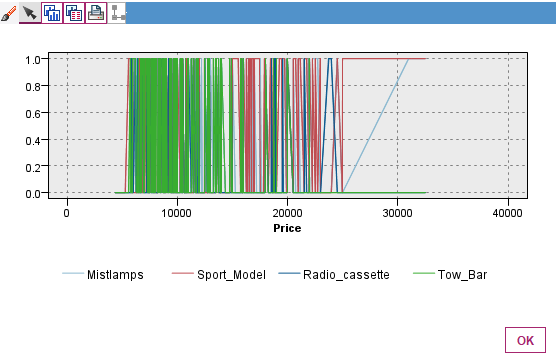
1. Explore the relationship between price and each of the continuous predictors. Does there seem to be a linear relationship? (Note: if you plan to use a scatterplot matrix, do not plot more than 4 or 5 variables together, as the plot consumes plenty of memory; create more than one scatterplot to visualize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

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



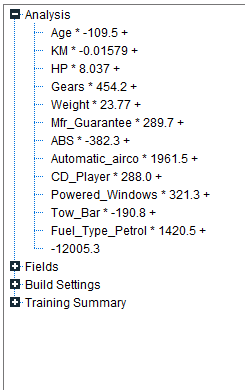
The picture depicts that there is a linear relation between price and age, price and weight mostly.

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

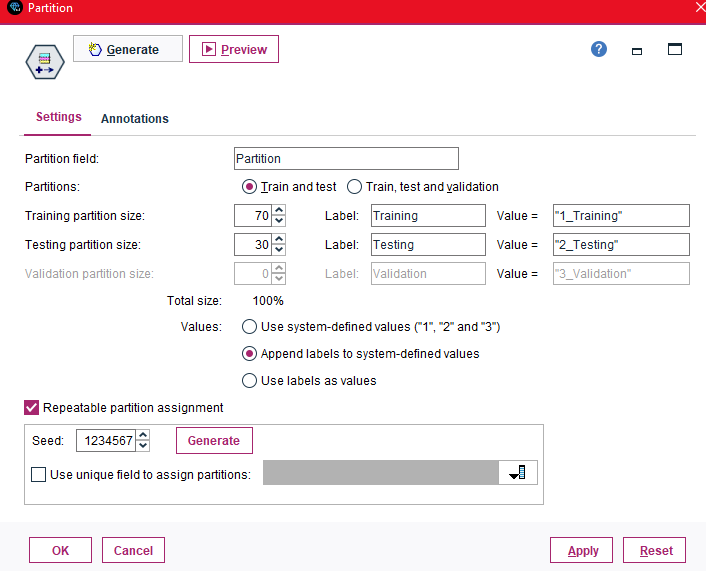


Individual relation depiction is not clear here, so considering the linear regression equation,

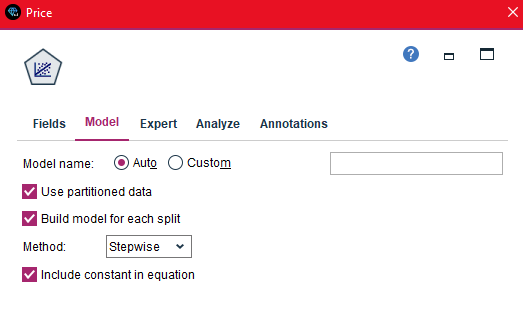
Price= Fuel type\*1420.5+ Age\*-109.5+KM\*-0.01579+HP\*8.037+gears\*454.2+weight\*23.77+Mfr\_guarantee\*289.7+ABS\*-382.3+Automatic\_airco\*1961.5+CD\_Player\*288+Powered\_Window\*321.3+Tow\_Bar\*-190.8-12005.3.



1. To fit a predictive model for price of used cars:
2. Partition the dataset into training and testing data sets.

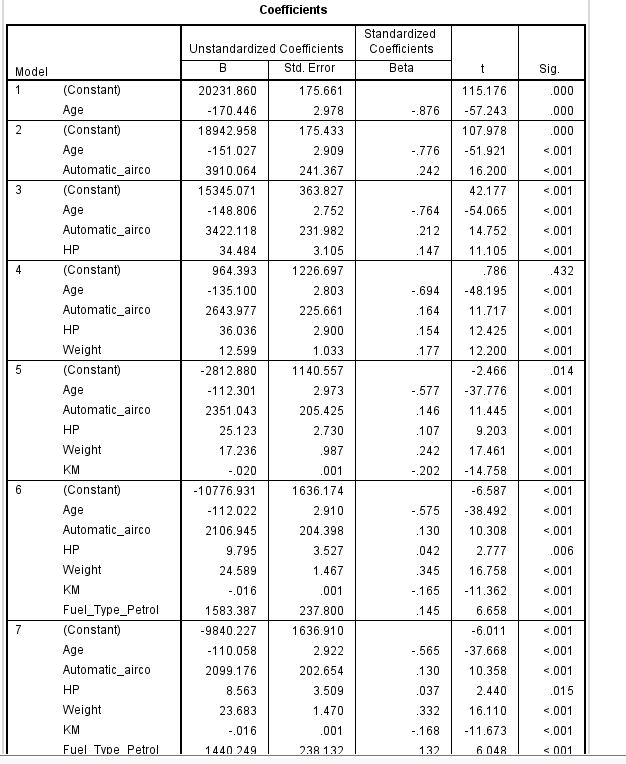


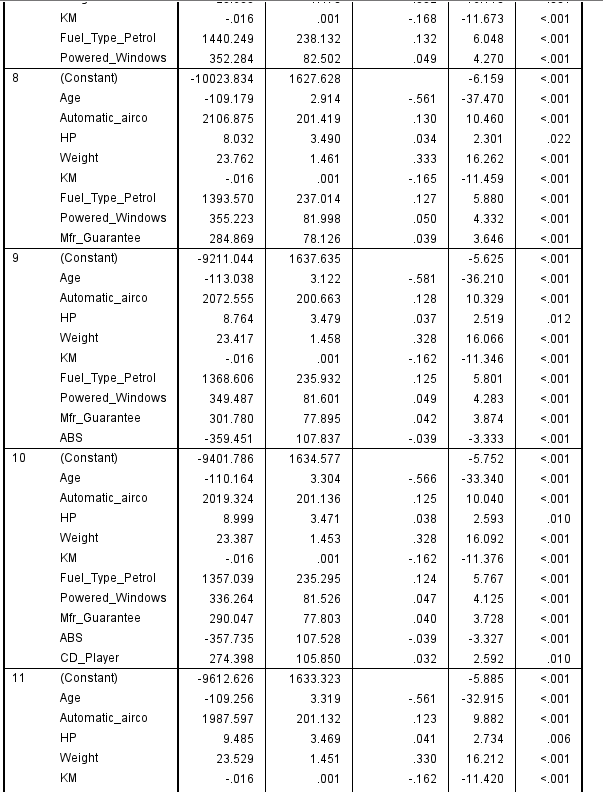
1. Use stepwise regression to reduce the number of predictors using the training data

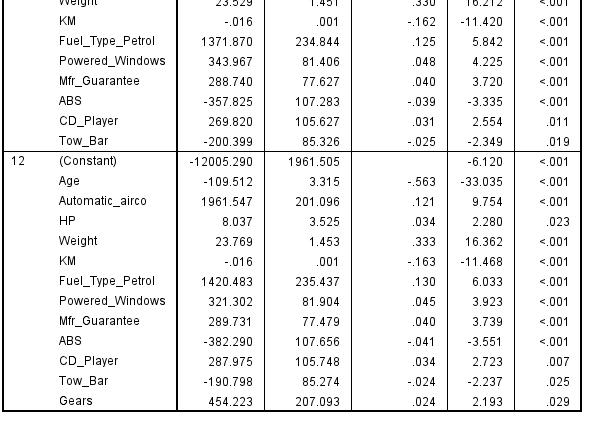


1. Report the model parameters: regression coefficient estimates with their standard errors, goodness of fit metrics (R-squared, adjusted R-squared), standard error of the estimate (s), t-test values (scores and p-values), F-test values (F-score and p-value)

The model parameters are as follows





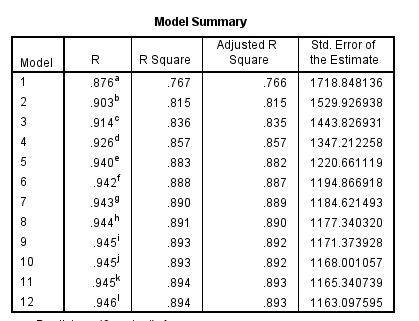


From the above table, H0 can be rejected for T values as they are mostly at smaller confidence level

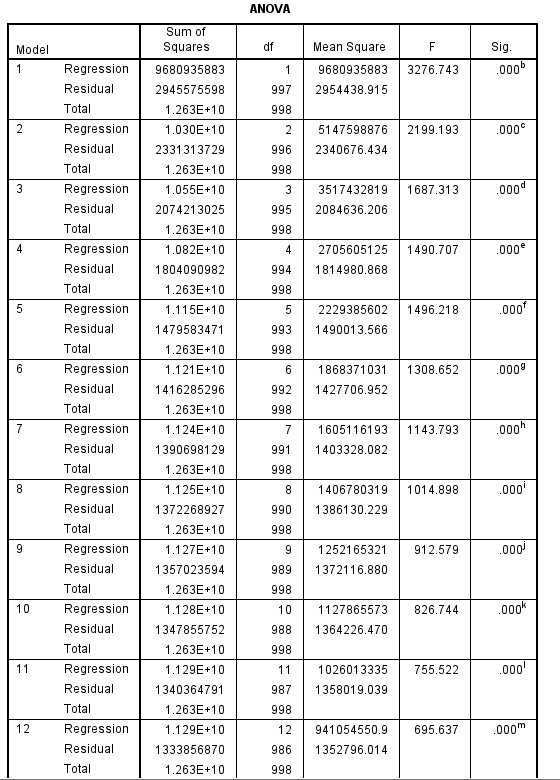
Goodness of fit and standard error of estimates

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

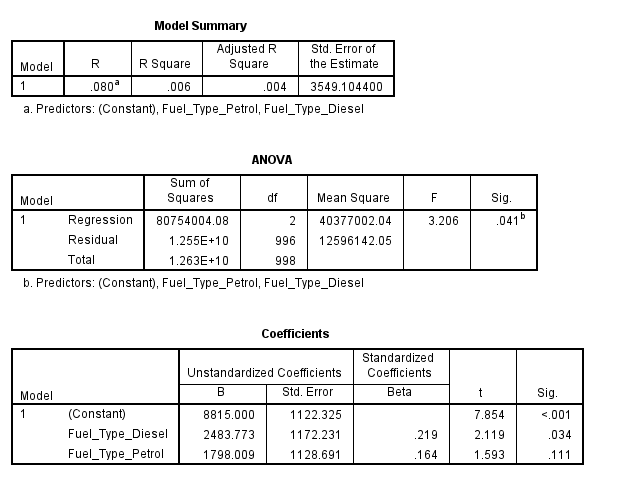
The Radj<R2 in all the parameters, hence omitting may be considered.



Standard error of estimates



Considering the relation between price and fuel type,

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

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

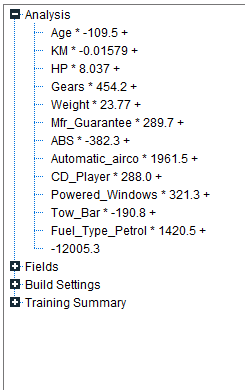
For diesel fuel type confidence level, the regression coefficient not zero is 1-0.034=0.966 96.6%

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

1. Write and explain the regression equation.



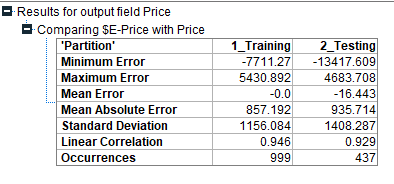
The regression equation Price= Fuel type\*1420.5+ Age\*-109.5+KM\*-0.01579+HP\*8.037+gears\*454.2+weight\*23.77+Mfr\_guarantee\*289.7+ABS\*-382.3+Automatic\_airco\*1961.5+CD\_Player\*288+Powered\_Window\*321.3+Tow\_Bar\*-190.8-12005.3 gives below evaluations

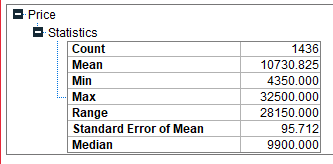
It denotes that a unitary increase in the respective parameters produces an increase or decrease based on + or – of the parametric slopes when remaining are kept constant. That is, with a unitary increase in fuel type, an increase of 1420.5 is observed when remaining parameters are kept constant.

Negative regression coefficients indicates negative relationships.

1. 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

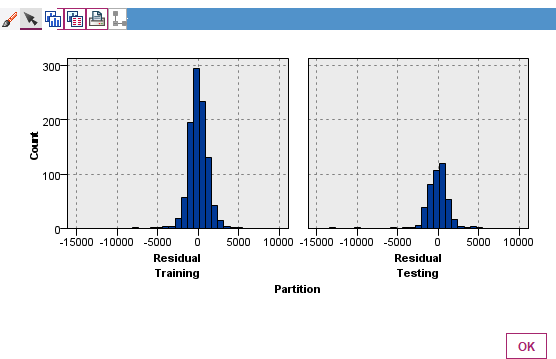




On performance evaluation, we got the above.

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%

1. 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.