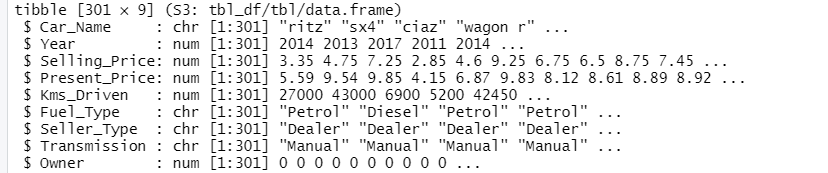
**LM LAB ASSIGNMENT**

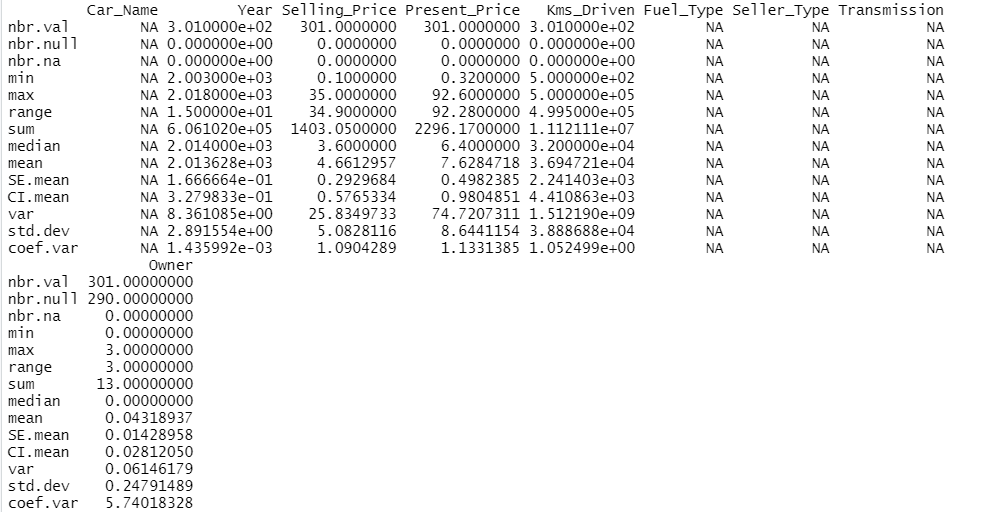
1. **Select parameters of your choice and apply Simple Linear regression and interpret the results.**

On given car data:

* The structure of the data and the data types are given below:

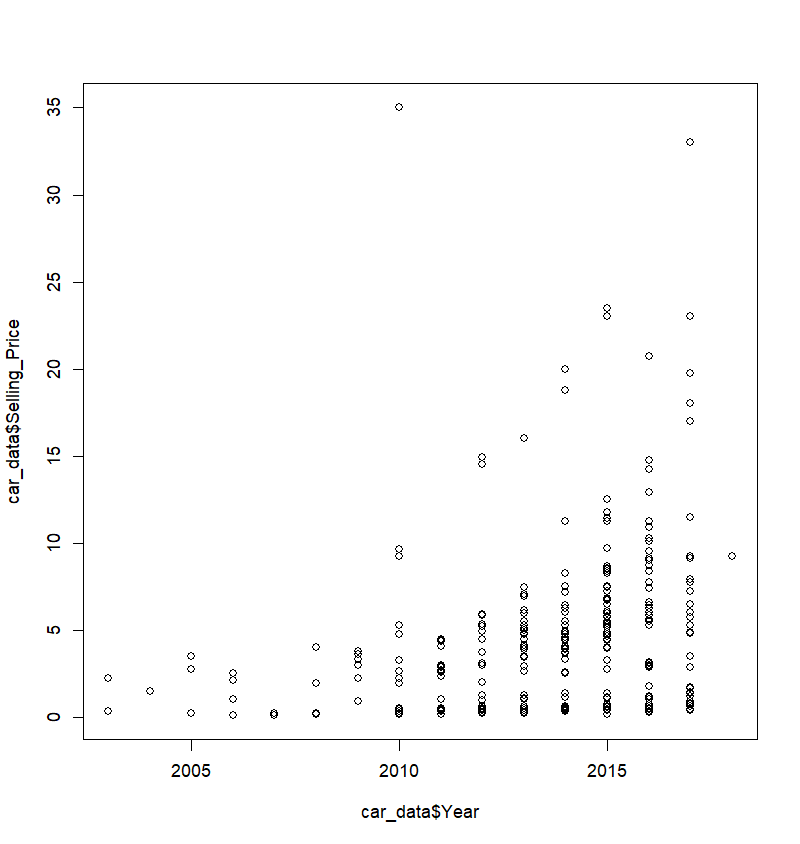


* The basic statistics are given as:

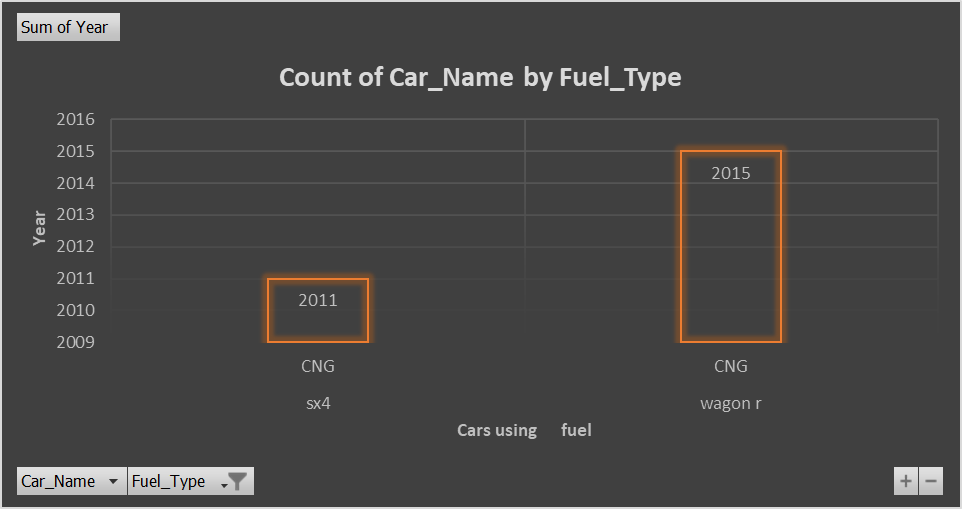


* The scatter plot of the selling price and year is given as:

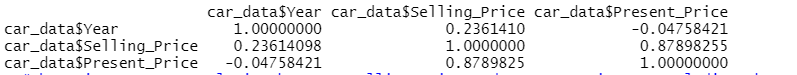
We see that the selling price increases as the year goes on increasing



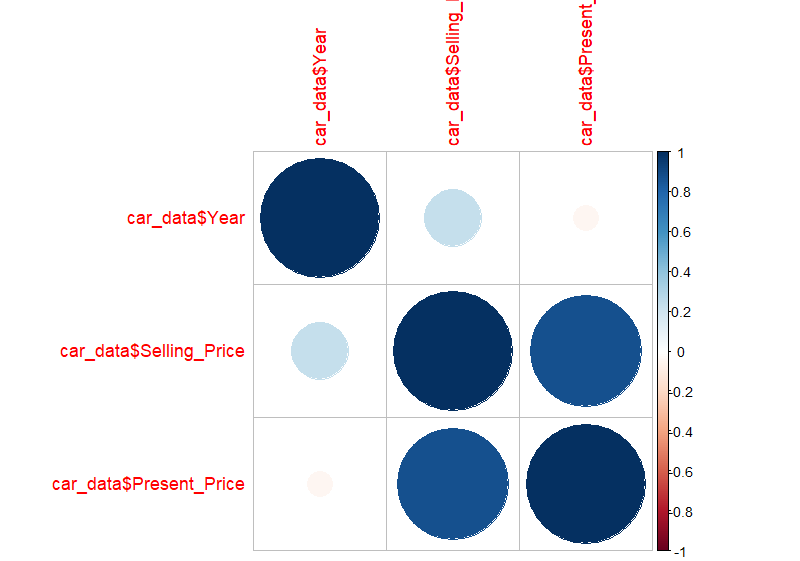
* The following bar plot shows that there are only 2 types of cars in the dataset which use CNG as the fuel type for the last 7 yrs. In other words, CNG gas is the least used fuel among the given three types of fuel and petrol is used at a high rate.



* The correlation matrix:

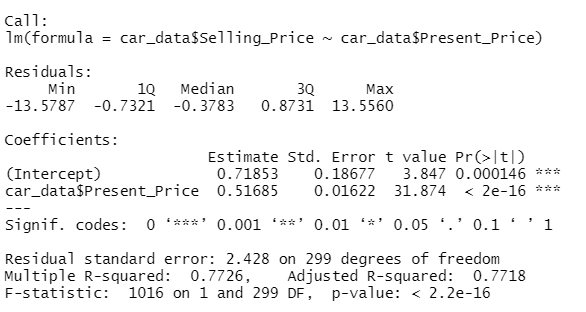


And the correlation plot,



We see that there is a strong correlation between the present price and the selling price. And very less correlation between the year and the present price.

* We fit a simple linear regression with the selling price as the dependent variable and the present price as the independent variable. The summary of the model is,

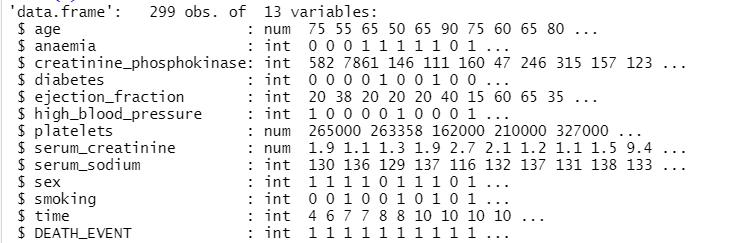


From Adjusted , we see that 77.18% of the variability in selling price is explained by the model.

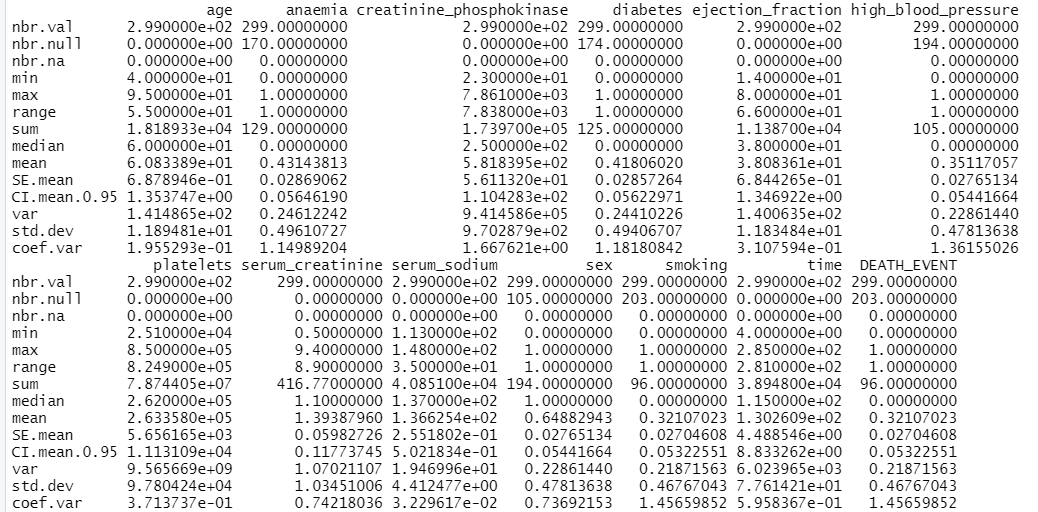
1. **Use Logistics regression using the attached data file. Choose parameters of interest. Interpret the results.**

# Data set is about heart failure clinical record

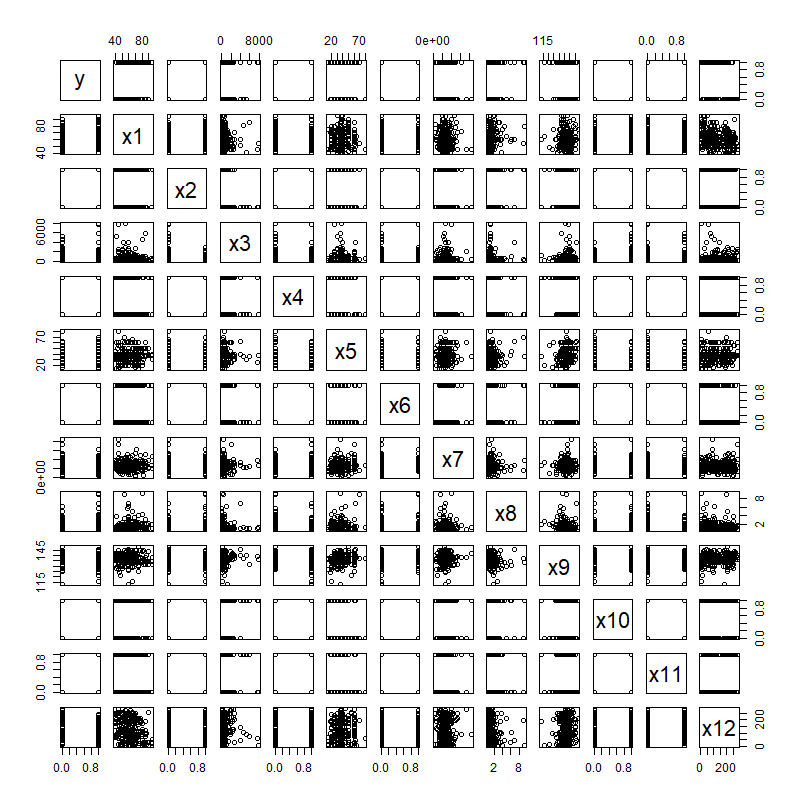
* The structure of the dataset is there are 299 medical records of heart failure patients with 13 features. Death event is the dependent variable.



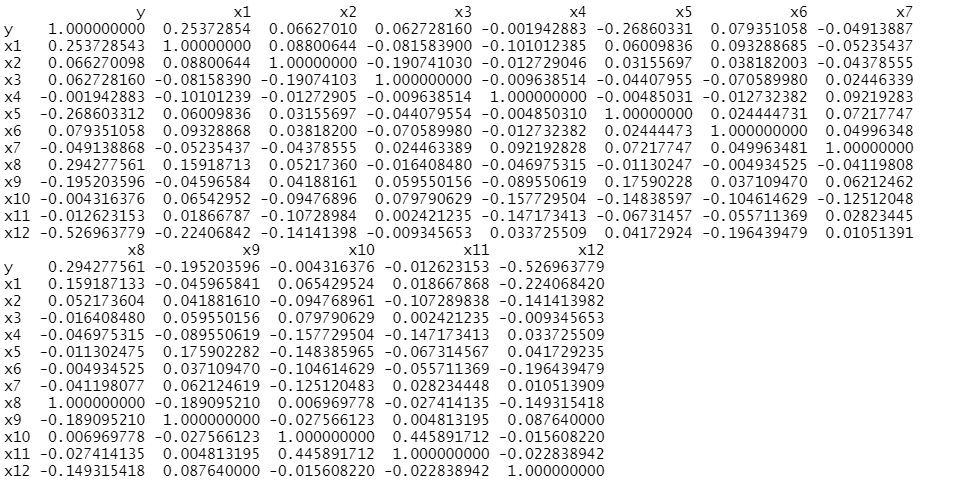
* The EDAs (or statistical description) of the dataset is,



This gives us the min, max values, mode, mean, standard deviations, and many more statistical quantities of each variable. For example, in Age, the minimum age is 40, the maximum age is 95 years, its standard deviation is 11.89481, etc.

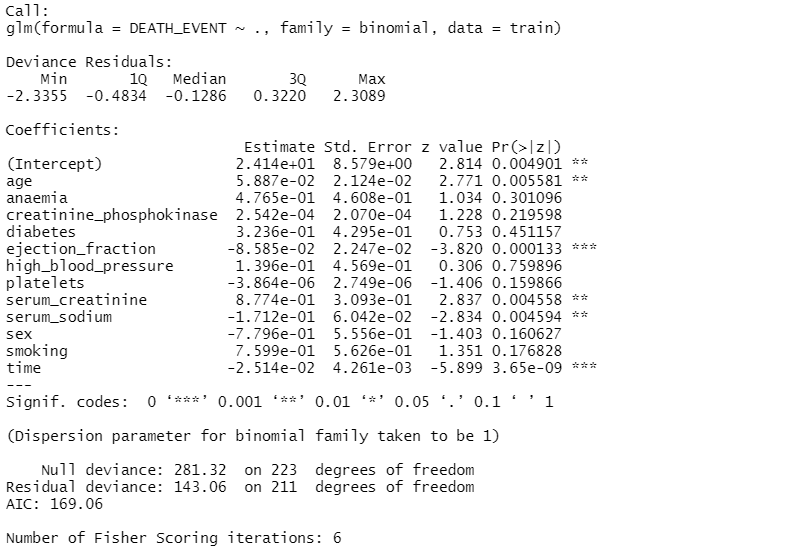
The covariance plot:

* The correlation matrix:



#There is a negative correlation between the time and death event & positive correlation between age & death event & serum\_creatinine and death event.

* Let X1, X2,...X12 be the 12 regressors and death event (Y) be the response variable. The data set does not have any null value.
* Fitting of Logistic Regression Model:



From the O/P we observe that, estimates of parameters involved in this model: Intercept:2.414e+01, S1=5.887e-02, S2=4.765e-01, S3=...,s12=-2.514e-02 .Standard errors of regression coefficient are given in 2nd column of coeff. table. The logistic regression model may be written as, pi(x)=e^( 2.414e+01 +5.887e-02x1+4.765e-01x2+2.542e-04x3+3.236e-01x4-8.585e-02x5+ 1.396e-01x6 -3.864e-06x7+8.774e-01x8-1.712e-01x9-7.796e-01x10+7.599e-01x11-2.514e-02x12)/(1+e^(2.414e+01 +5.887e-02x1+4.765e-01x2+2.542e-04x3+3.236e-01x4-8.585e-02x5+ 1.396e-01x6 -3.864e-06x7+8.774e-01x8-1.712e-01x9-7.796e-01x10+7.599e-01x11-2.514e-02x12))Beta1=5.887e-02 implies that e^(beta1)=1.0606455. It can be interpreted as , for every unit increase in age the chance of death event increase by 06.06%.

Null Deviance=281.32 Residual Deviance=143.06

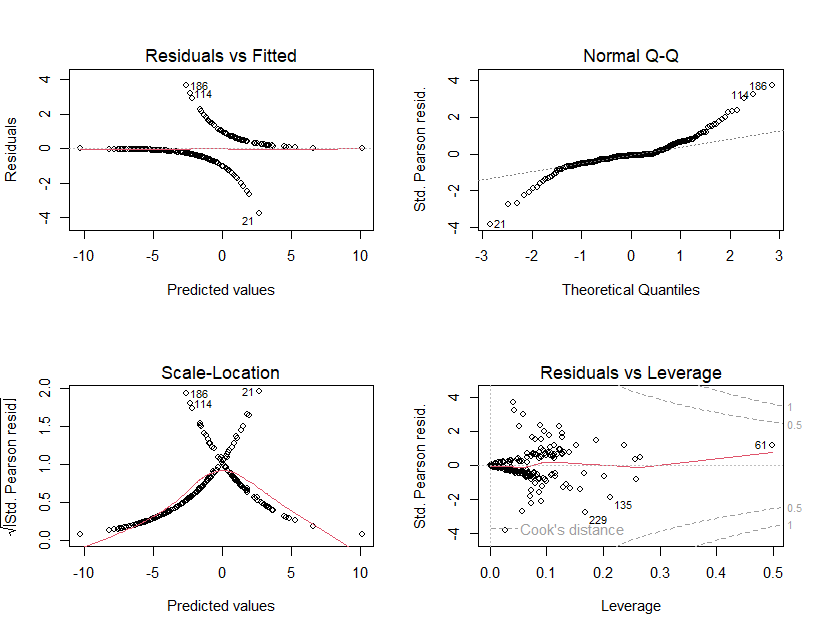
G=Hessian matrix follows chisq with df 1

G=Null Deviance- Residual Deviance=138.26

Chisq(12,0.05)=21.026

Since, G exceeds, H01 is rejected at 5% los. Hence, atleast 1 regression coeff. is significant. From pvalue in last column we observe that age, ejection\_fraction, serum\_creatinine, serum\_sodium, time are significant.

* Diagnostic plots



Interpretation from the plots:

Normal Q-Q plot: Flat extremes indicate that observations are from a distribution having tails thinner than the normal distribution. This shows a linear trend.

Residual Vs Leverage plot: There is no influential case. We can barely see Cook’s distance lines because all cases are well inside Cook’s distance lines.

Residual Vs Fitted: This shows a pattern that is not equally spread residuals around a horizontal line.

The accuracy of model id is 78%