# **Red wine Quality**

**Problem Statement**

**About Data**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

Data Description  
  
1 - fixed acidity  
  
2 - volatile acidity  
  
3 - citric acid  
  
4 - residual sugar  
  
5 - chlorides  
  
6 - free sulfur dioxide  
  
7 - total sulfur dioxide  
  
8 - density  
  
9 - pH  
  
10 - sulphates  
  
11 - alcohol  
  
12 - quality (score between 0 and 10)

## Table of Contents

1. [**Import Libraries**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#import_lib)
2. [**Set Options**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#set_options)
3. [**Read Data**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Read_Data)
4. [**Data Analysis and Preparation**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#data_preparation)

4.1 - [Understand the Data](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Data_Understanding)

* + - 4.1.1 - [Data Dimension](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Data_Shape)
    - 4.1.2 - [Data Types](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Data_Types)
    - 4.1.3 - [Summary Statistics](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Summary_Statistics)
    - 4.1.4 - [Missing Values](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Missing_Values)
    - 4.1.5 - [Correlation](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#correlation)
    - 4.1.8 - [Feature Engineering](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Feature_Engineering)
    - 4.1.9 - [Discover Outliers](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#outliers)
    - 4.1.10- [Recheck the Correlation](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Recheck_Correlation)

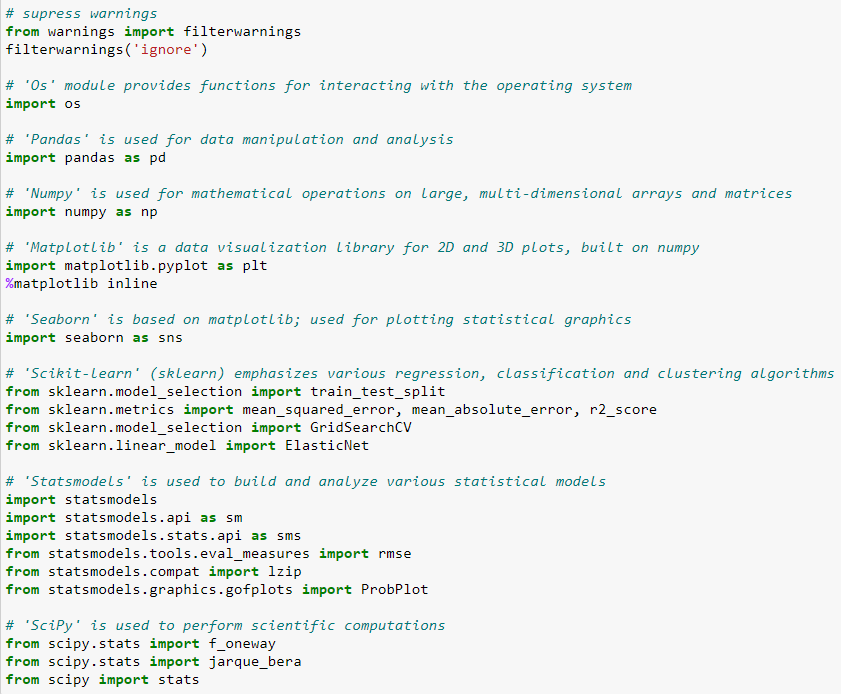
1. [**Linear Regression (OLS)**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#LinearRegression)
   * - [Multiple Linear Regression - Full Model](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#withLog)
   * - [Fine Tune Linear Regression Model (OLS)](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Finetuning)
2. **Feature Engineering**

* RFE( Recursive Feature Elimination
* RFE CV
* Forward Selection Process

1. [**Regularization (OLS)**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#regularization)
   * [Ridge Regression Model (OLS)](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Ridge_Regression)
   * [Lasso Regression Model (OLS)](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Lasso_Regression)
   * [Elastic Net Regression Model (OLS)](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#Elastic_Net)
2. [**Conclusion and Interpretation**](http://localhost:8888/notebooks/Project/Machine%20Learning/case%20study/SLR%20-%20Case%20Study%20-%20Insurance%20Claim%20-%20%5Bv1.6%20-%2013052020%5D.ipynb#rmse_and_r-squared)

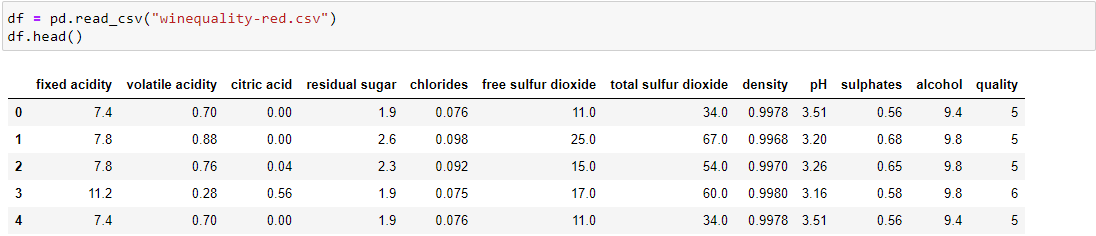
## **Import Libraries**

**Let us import the required libraries and functions.**

****

## **Read Data**

Read and display data to get insights from the data . Here we use pd.read\_csv() to desire data set. and print the head of the file



# Data Analysis and Preparation

### Data preparation is the process of cleaning and transforming raw data prior to building predictive models.

##### *Here we will analyse and prepare data to perform regression analysis:*

1. Check dimensions of the data frame in terms of rows and columns
2. Check the data types. Refer data definition to ensure your data types are correct
3. If data types are not as per business context, change the data types as per requirement
4. Study summary statistics
5. Check for missing values
6. Study correlation
7. Perform feature engineering
8. Detect outliers
9. Recheck the correlation

##### *Note: It is an art to explore data and one will need more and more practice to gain expertise in this area.*

#### Data Dimension

For data dimension we use df.shape() to get the rows and columns in the dataset.

#### Data Types

For Data types we use df.dtypes to get the respective data types of the columns.

### Summary Statistics

### It helps to get the statistical relationship between the numerical columns.

### We use df.describe() to get the desired output.

### 

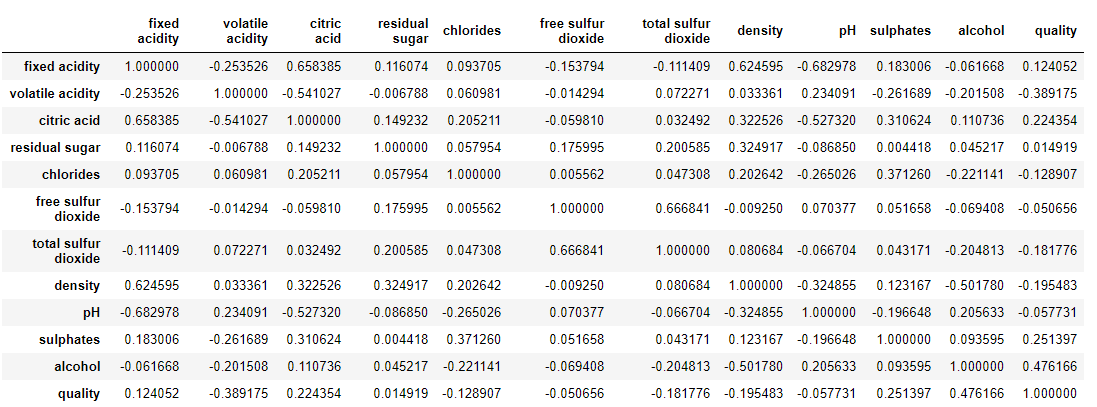
### Missing Value

### For missing value detection we use df.isnull().sum() function .

### It will give the missing value count of each columns.

## Checking the correlation

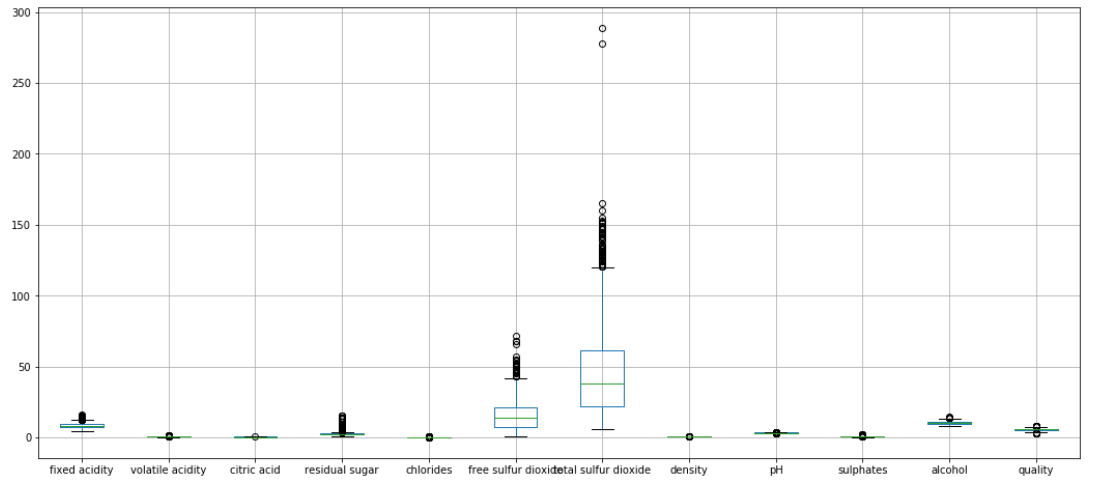
**df.corr()** is used to find the pairwise correlation of all columns in the data frame. Any **null** values are automatically excluded. For any non-numeric data type columns in the data frame it is ignored.



### 

## **Discover Outliers**

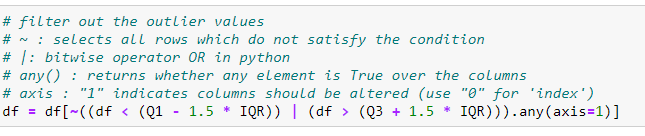
## Importance of detecting an outlier. An outlier is an observation that appears to deviate distinctly from other observations in the data. If the outliers are not removed, the model accuracy may decrease. Recollect that one of the assumptions of Linear Regression is there should be no outliers present in the data



### Using IQR Method

### The interquartile range rule is useful in detecting the presence of outliers. [Outliers](https://www.thoughtco.com/what-is-an-outlier-3126227) are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

### 



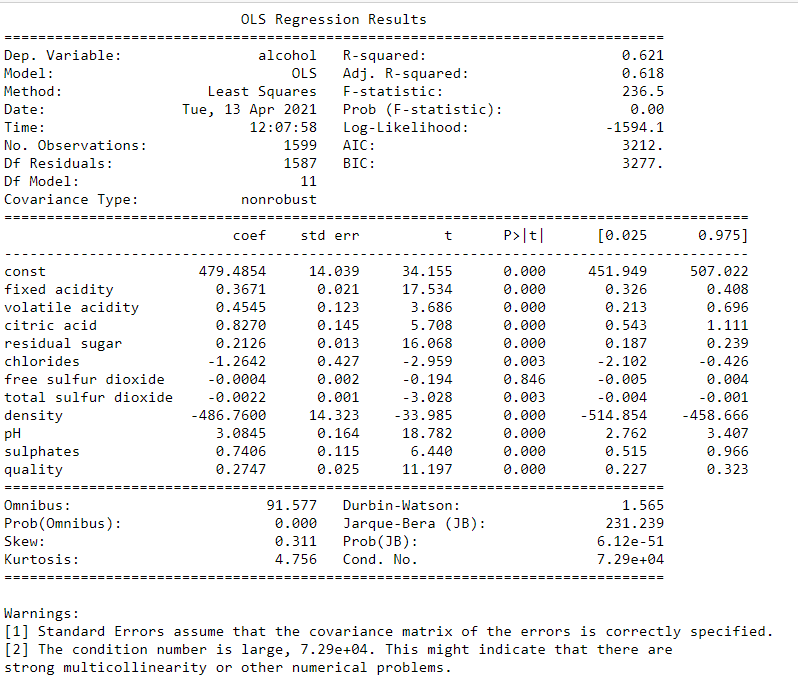
After the IQR test now the rows reduced to to 1177. Which is previously 1569.

## **Linear Regression (OLS)**

Linear Regression is the family of algorithms employed in supervised machine learning tasks Knowing that supervised ML tasks are normally divided into classification and regression, we can collocate Linear Regression algorithms in the latter category. It differs from classification because of the nature of the target variable: in classification, the target is a categorical value (‘yes/no’, ‘red/blue/green’, ‘spam/not spam’…); on the other hand, regression involves numerical, continuous values as target, hence the algorithm will be asked to predict a continuous number rather than a class or category. Namely, imagine you want to predict the price of a house based on some relative features: the output of your model will be the price, hence a continuous number.

Regression tasks can be divided into two main groups: those which use only one feature to predict the target, and those which use more than one features for that purpose. To give you an example, let’s consider the house task above: if you want to predict its price only based on its squared meters, you will fall into the first situation (one feature); if you are going to predict the price based on, let’s say, its squared meters, its position and the liveability of the surrounding environment, you are going to fall into the second situation (multiple features, in that case, three).

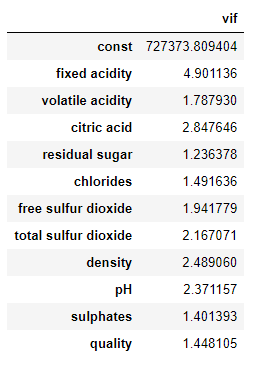
In the first scenario, the algorithm you are likely to employ will be the Simple Linear Regression, which is the one we are going to talk about in this article. On the other side, whenever you are facing more than one features able to explain the target variable, you are likely to employ a Multiple Linear Regression.



In the Above result r2 value is 62% we have to improve the value for a good fit model.

## **Variance Inflation Factor**

A variance inflation factor (VIF) detects [multicollinearity](https://www.statisticshowto.com/multicollinearity/)in [regression analysis](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/). Multicollinearity is when there’s [correlation](https://www.statisticshowto.com/probability-and-statistics/correlation-analysis/)between predictors (i.e. [independent variables](https://www.statisticshowto.com/independent-variable-definition/)) in a model; it’s presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

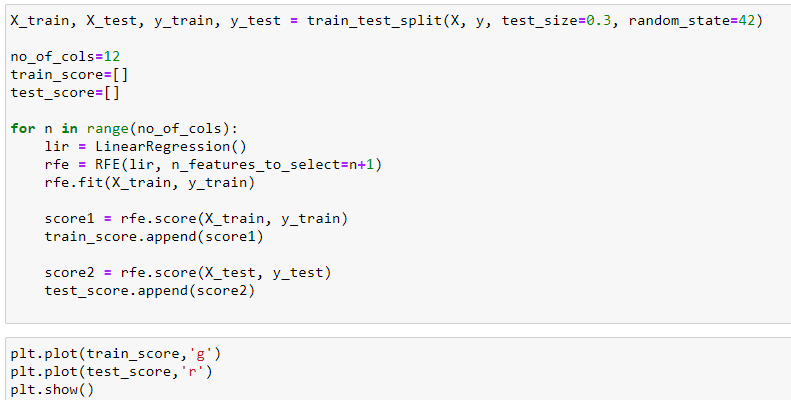


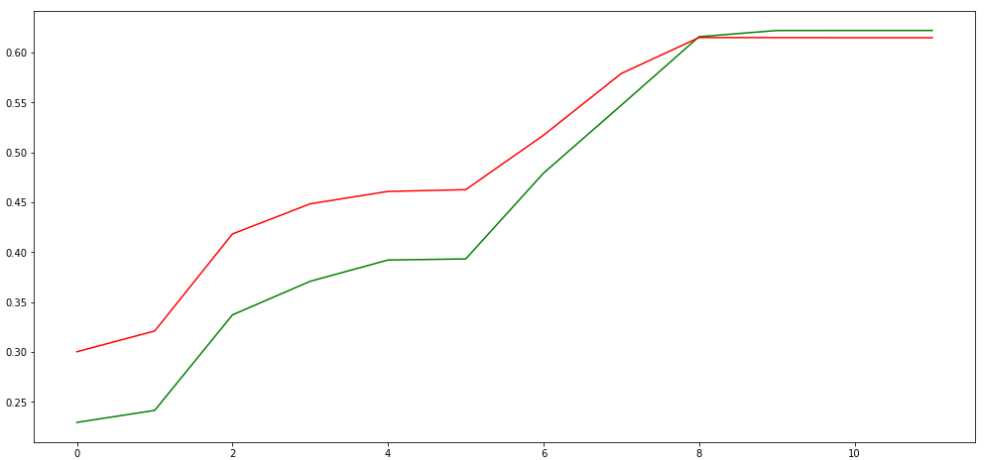
# **Feature Engineering**

### RFE -Recursive Feature Elimination

### RFE is a wrapper-type feature selection algorithm. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains.

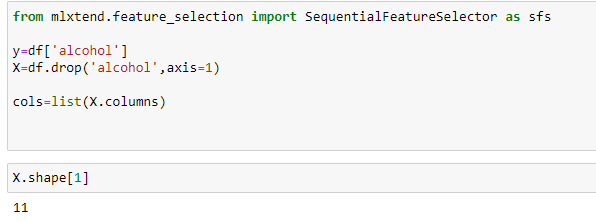
There are two important configuration options when using RFE: the choice in the number of features to select and the choice of the algorithm used to help choose features. Both of these hyperparameters can be explored, although the performance of the method is not strongly dependent on these hyperparameters being configured well.

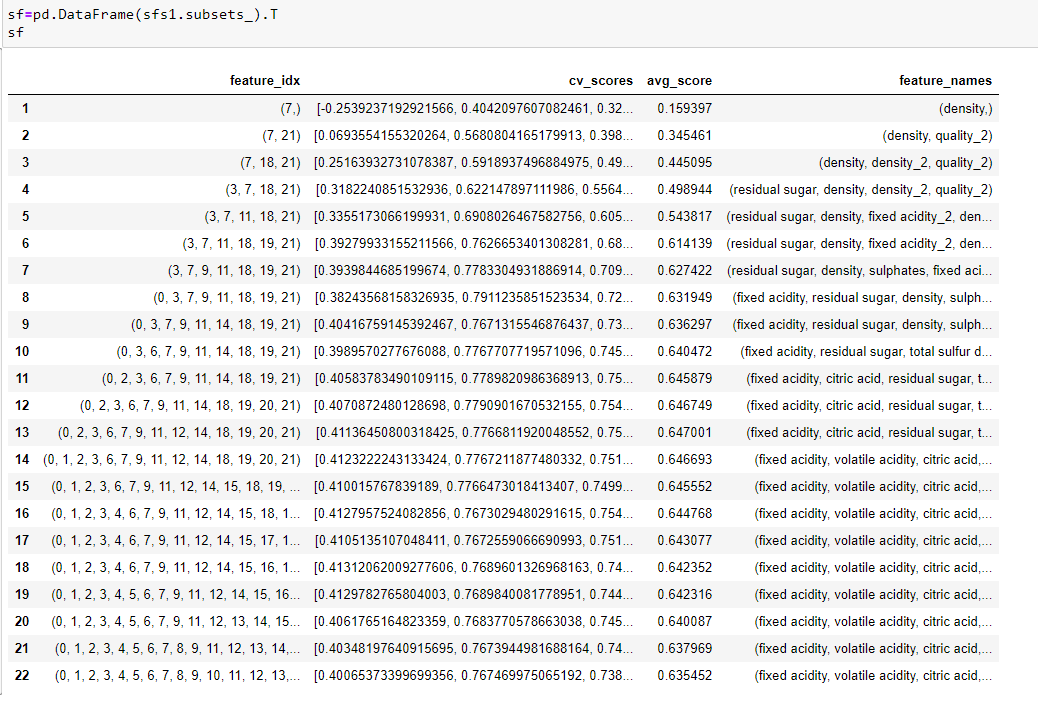


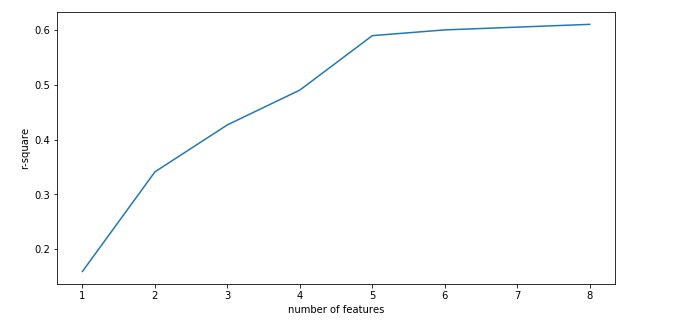


### Forward Selection Approaches

### Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model

****

****

****

**Regularization**

Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

**Lasso**

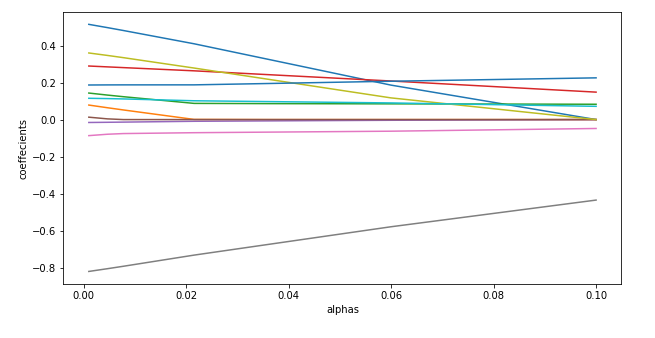
This regularization technique performs L1 regularization. Unlike Ridge Regression, it modifies the RSS by adding the penalty (shrinkage quantity) equivalent to the sum of the absolute value of coefficients.

Looking at the equation below, we can observe that similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition to this, it is quite capable of reducing the variability and improving the accuracy of linear regression models.

**Limitation of Lasso Regression:**

* If the number of predictors (p) is greater than the number of observations (n), Lasso will pick at most n predictors as non-zero, even if all predictors are relevant (or may be used in the test set). In such cases, Lasso sometimes really has to struggle with such types of data.
* If there are two or more highly collinear variables, then LASSO regression select one of them randomly which is not good for the interpretation of data.

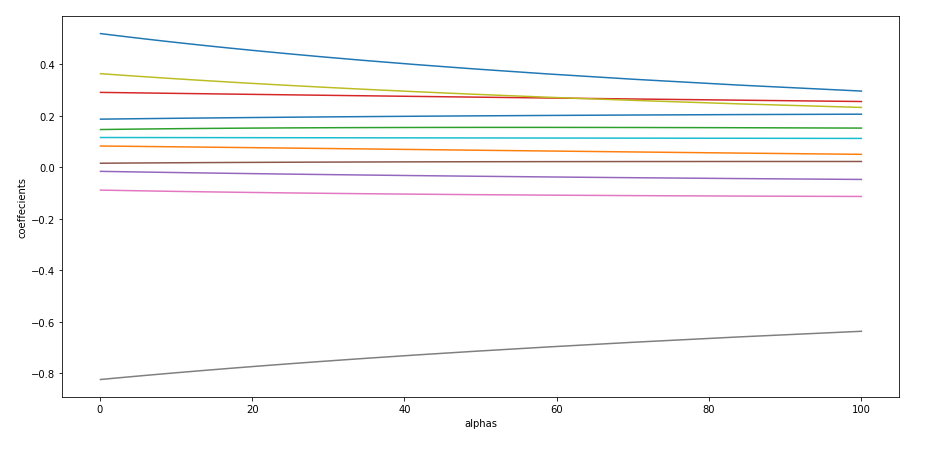
Lasso regression differs from ridge regression in a way that it uses absolute values within the penalty function, rather than that of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. The more penalty is applied, the more the estimates get shrunk towards absolute zero. This helps to variable selection out of given range of n variables.



# **Ridge Regression**

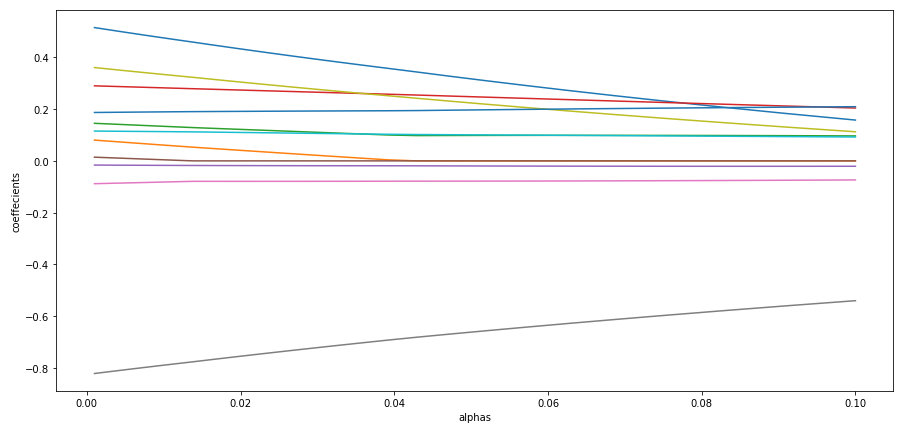
Ridge [regression](https://www.mygreatlearning.com/blog/what-is-regression/) is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

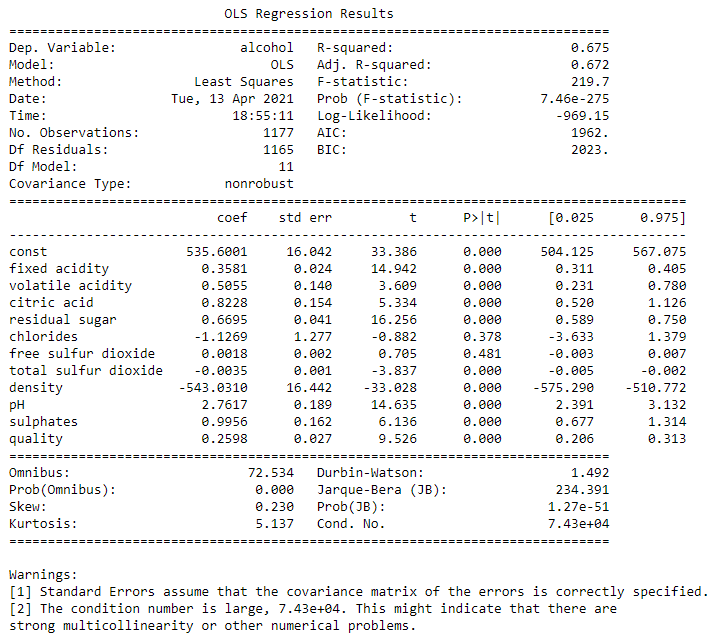


**Elastic Net**

The elastic net method overcomes the limitations of the [LASSO](https://en.wikipedia.org/wiki/Lasso_(statistics)) (least absolute shrinkage and selection operator) method which uses a penalty function based on Use of this penalty function has several limitations.[[1]](https://en.wikipedia.org/wiki/Elastic_net_regularization#cite_note-ZH-1) For example, in the "large *p*, small *n*" case (high-dimensional data with few examples), the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations,



**Rechecking the Linear model**

****

**Conclusion**

As we can see the r2 value of the OLS model is 0.621 and ad.r2 value is 0.618.later when we perform recursive feature elimination we get the value of r2 as 0.671 and ad.r2 as 0.668.

Where as after performing the fine tunning r2 value is 0.675 and ad.r2 value is 0.672, which indicates there is no much significance change post tunning of the data.

But r2 and ad.r2 value come to a very close value after fine tunning which indicates the model fits good post removal of insignificant variables.