# Lasso and Ridge Regression



**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Prajay B. Urkude Batch ID: 16092021**

**Topic: Lasso and Ridge Regression**

**Problem Statements:**

1. Officeworks is a leading retail store in Australia, with numerous outlets around the country. The manager would like to improve the customer experience by providing them online predictive prices for their laptops if they want to sell them. To improve this experience the manager would like us to build a model which is sustainable and accurate enough. Apply Lasso and Ridge Regression model on the dataset and predict the price, given other attributes. Tabulate R squared, RMSE, and correlation values.





: **Business Objective:**

To improve the customer experience by providing them online predictive prices for their laptops if they want to sell them.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| price | Price of the computer | Quantitative, Ratio | Relevant |
| speed | Speed of the computer i,e, processor | Quantitative, Ratio | Relevant |
| hd | hd resolution | Quantitative, Ratio | Relevant |
| ram | Ram of the computer | Quantitative, Ratio | Relevant |
| screen | Screen size of the computer | Quantitative, Ratio | Relevant |
| cd | Is cd can be fit or not | Qualitative, nominal | Relevant |
| multi | Is computer multipurpose or not | Qualitative, nominal | Relevant |
| premium | Is computer premium or not | Qualitative, nominal | Relevant |
| ads | Alternate data stream | Quantitative, Ratio | Relevant |
| trend | Is computer trending or not | Quantitative, Nominal | Relevant |

Steps For the multilinear Regression by Lasso and Ridge Regression:

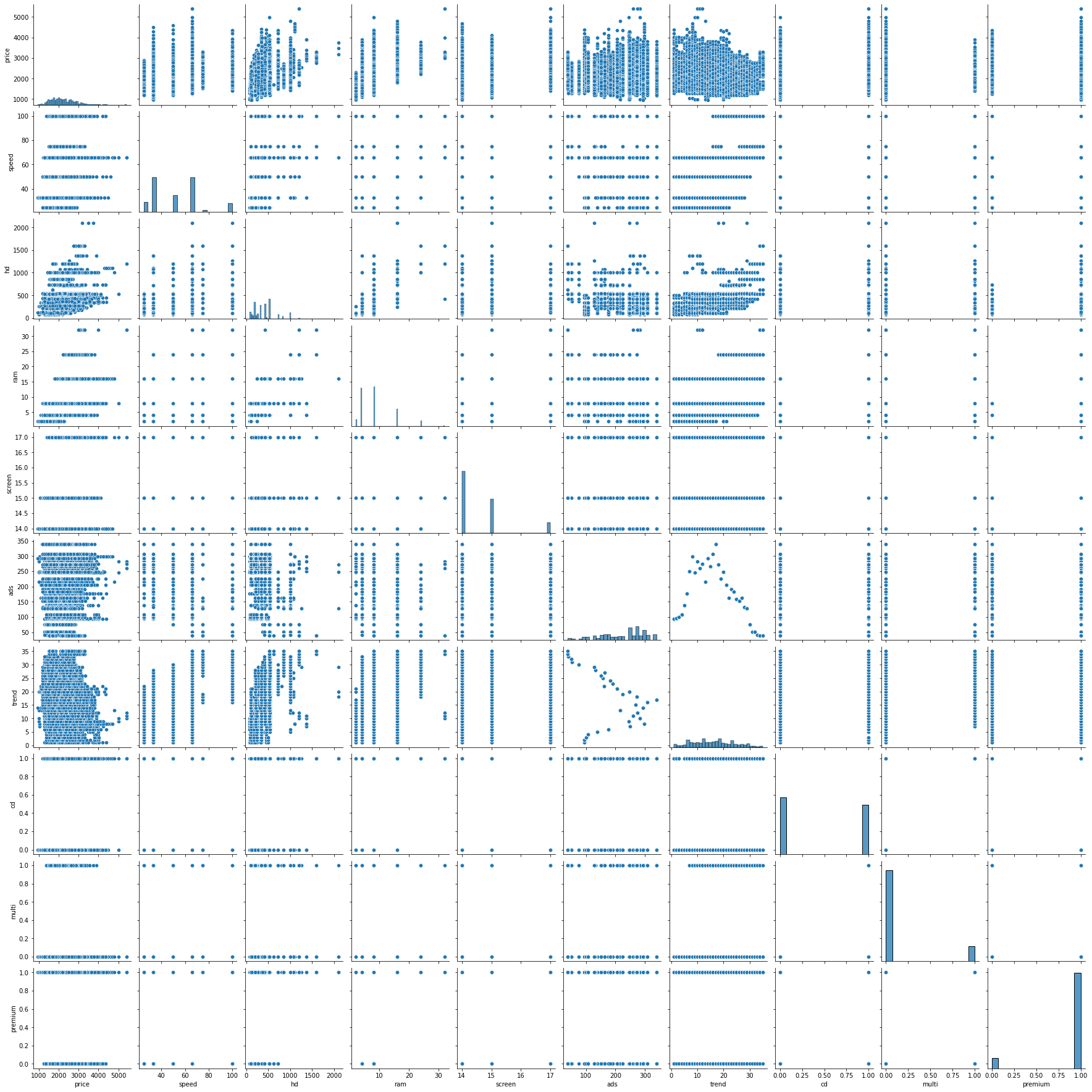
* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics is available for each estimator.

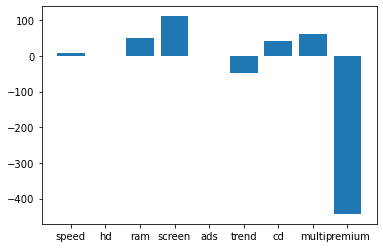
* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Find the correlation between the variables and by using the corr() function and plot the pairplot.
* 
* The above graph is the pair plot which shows the correlation matrix from which we can see the correlation between different variables.

# From the table we can say that ram and hd are moderately correlated

# ads and the trend are correlated with each other in the negative direction.

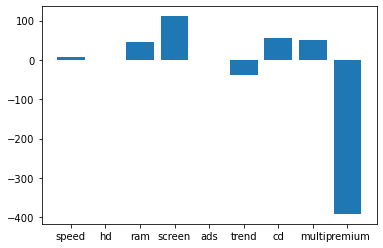
* Preparing multilinear regression model by considering all the variables and view the summary.
* From the summary we have to see the R2 value is 0.776 and adjusted R2 value is 0.775 which is less than 0.85.
* Calculate the RMSE value which is found out to be 275.
* To increase the accuracy of the model and to reduce the error we use Lasso and Ridge regression method.

**Lasso Regression Method:**

* Import the Lasso function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the the Lasso function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.
* From the above plot we can see that the coefficient for hd and ads become zero and other coefficients value are also adjusted according the correlation of the variable with the output variable and take only values which has the some coefficient value and avoid the variable which has the coefficient zero (feature selection)
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.77 and calculate the RMSE score which is coming out 278.
* We can repeat the process till we get the better accuracy by changing the alpha value.

**Ridge Regression Method:**

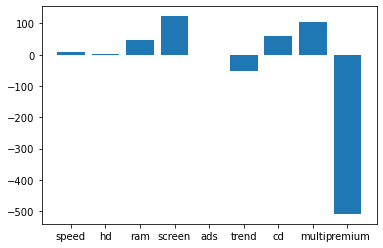
* Import the Ridge function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the Ridge function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.



* From the above plot we can see that the coefficient for hd and ad is nearly equal to become zero but not zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.75 and calculate the RMSE score which is coming out 290.
* We can repeat the process till we get the better accuracy by changing the alpha value.

**ElasticNet Regression Method:**

* Elastic net regression methid is the combination of L1 and L2 Regularization method
* Import the ElasticNet function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the ElasticNet function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.



* From the above plot we can see that the coefficient for hd is nearly equal to become zero and ads column became and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.77 and calculate the RMSE score which is coming out 275.
* We can repeat the process till we get the better accuracy by changing the alpha value.

From the above 3 regression model if we compare three methods the we found that elasticnet method gives the good result for same value of alpha = 0.15.

* For better accuracy of the model, we do the Hyperparameter tuning by using the GridSearchCV function of model\_selection package of sklearn.
* Hyperparameters tuning to be done for all the three methods of regression i.e. Lasso, Ridge and ElasticNet by taking the different alpha values as parameters and cross validation value =5 and find out the best alpha value for the best accuracy and predict the result.f
* Out of three methods after hyperparameter tuning we get the good result for Lasso regression method i.e. R2 = 0.77 ans RMSE value = 275 at best alpha value = 15

1. An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like to have an automated model which can predict the price of the car once the user inputs the required factors. Help the business achieve the objective by applying Lasso and Ridge Regression on it. Please use the below columns for the analysis: Price, Age\_08\_04, KM, HP, cc, Doors, Gears, Quarterly\_Tax, Weight.



**Business Objective:**

To improve its customer base and their experience by providing them an easy way to buy and sell cars.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Id | Price of the computer | Quantitative, Nominal | Irelevant |
| Model | Model of the car | Qualitative, nominal | Irelevant |
| Price | Offer Price in euros | Quantitative, Ratio | Relevant |
| Age\_08\_04 | Age in months as in August 2004 | Quantitative, Ratio | Relevant |
| Mfg\_Month | Manufacturing month (1-12) | Quantitative, Nominal | Irelevant |
| Mfg\_Year | Manufacturing Year | Qualitative, nominal | Irelevant |
| KM | Accumulated Kilometers on odometer | Quantitative, Ratio | Relevant |
| Fuel\_Type | Fuel Type (Petrol, Diesel, CNG) | Qualitative, nominal | Relevant |
| HP | Horse Power | Quantitative, Ratio | Relevant |
| Met\_Color | Metallic Color? (Yes=1, No=0) | Quantitative, Nominal | Irelevant |
| Color | Color (Blue, Red, Grey, Silver, Black, etc.) |  | Irelevant |
| Automatic | Automatic ( (Yes=1, No=0) |  | Irelevant |
| Cc | Cylinder Volume in cubic centimeters | Quantitative, Ratio | Irelevant |
| Doors | Number of doors | Quantitative, Ratio | Relevant |
| Cylinders | - Number of cylinders | Quantitative, Ratio | Relevant |
| Gears | Number of gear positions | Quantitative, Ratio | Relevant |
| Quarterly\_Tax | Quarterly road tax in euros | Quantitative, Ratio | Relevant |
| Weight | Weight in Kilograms | Quantitative, Ratio | Irelevant |
| Mfr\_Guarantee | Within Manufacturer's Guarantee period (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| BOVAG\_Guarantee | BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Guarantee\_Period | Guarantee period in months | Qualitative,Nominal | Irelevant |
| ABS | Anti-Lock Brake System (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airbag\_1 | Driver\_Airbag (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airbag\_2 | Passenger Airbag (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airco | Airconditioning (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Automatic\_airco | Automatic Airconditioning (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Boardcomputer | Boardcomputer (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| CD\_Player | CD Player (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Central\_Lock | Central Lock (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Powered\_Windows | Powered Windows (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Power\_Steering | Power Steering (Yes=1, No=0 | Qualitative,Nominal | Irelevant |
| Radio | Radio (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Mistlamps | Mistlamps (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Sport\_Model | Sports3 Model (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Backseat\_Divider | Backseat Divider (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Metallic\_Rim | Metallic Rim (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Radio\_cassette | Radio Cassette (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Tow\_Bar | Tow Bar (Yes=1, No=0) | Qualitative,Nominal | Irelevant |

Steps For the multilinear Regression by using Lasso ans Ridge Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

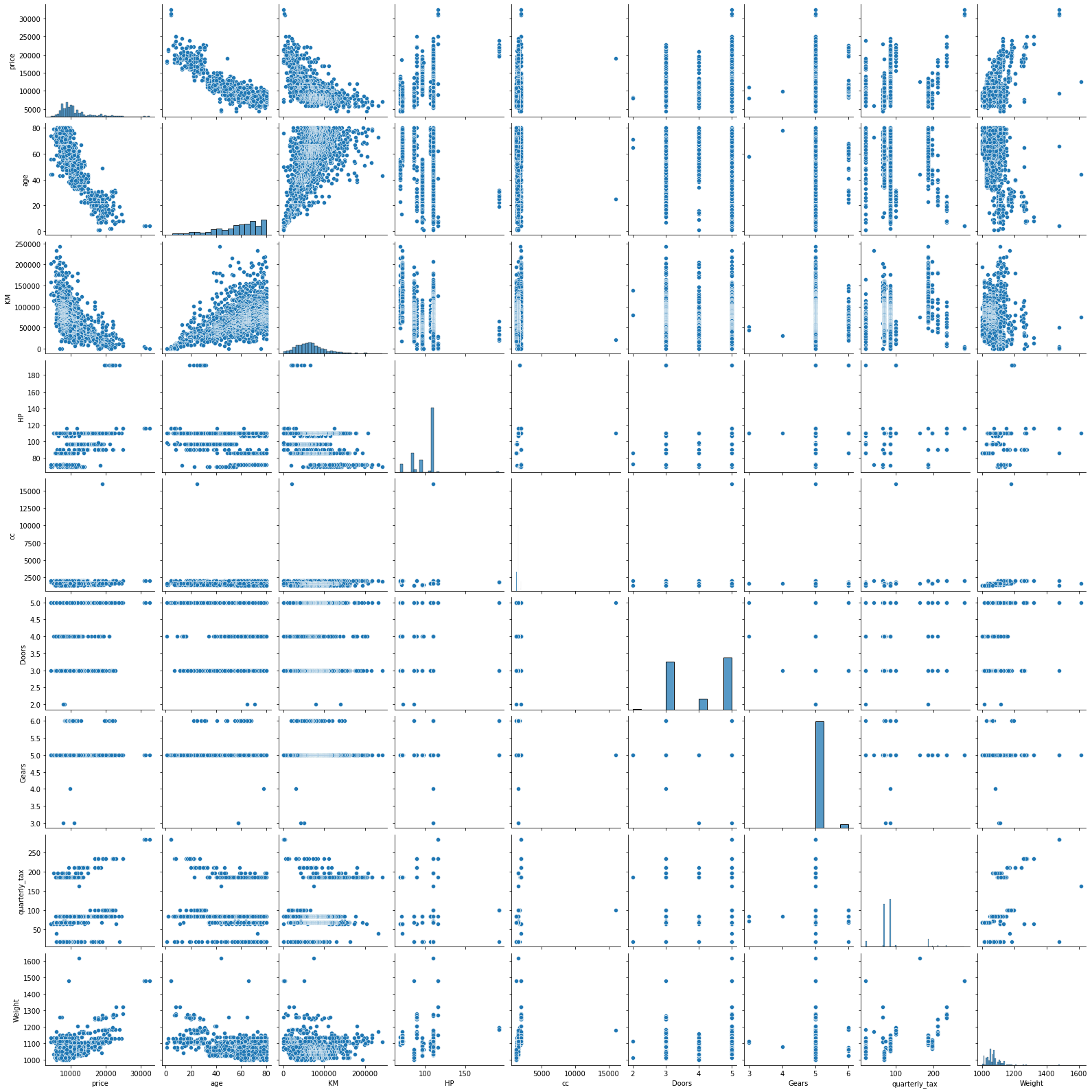
Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Find the correlation between the variables and by using the corr() function and plot the pairplot.



* The above graph is the pair plot which shows the correlation matrix from which we can see the correlation between different variables.

# From the pair plot we can say that age is strongly correlated with price

# KM, Weight is moderately correlated with price

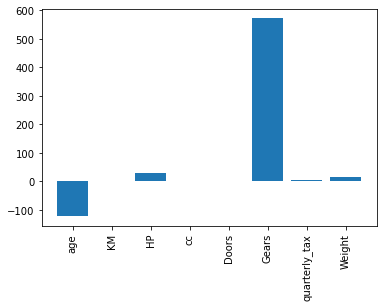
# Age and Km have some collinearity

# cc and qtax has some collinearity and cc and weight has moderate collinearity

# qtax and weight has moderate collinearity

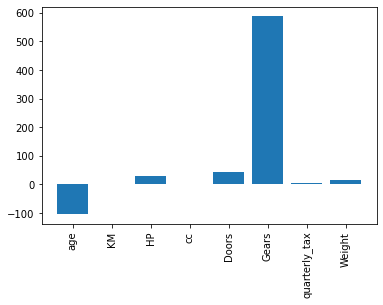
* Preparing multilinear regression model by considering all the variables and view the summary.
* From the summary we have to see the R2 value is 0.864 and adjusted R2 value is 0.863 which is greater than 0.85.
* Calculate the RMSE value which is found out to be 1338.
* To increase the accuracy of the model and to reduce the error we use Lasso and Ridge regression method.

**Lasso Regression Method:**

* Import the Lasso function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the the Lasso function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.
* From the above plot we can see that the coefficient for cc, KM, Doors became zero and other coefficients value are also adjusted according the correlation of the variable with the output variable and take only values which has some coefficient value and avoid the variable which has the coefficient zero (feature selection)
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.86 and calculate the RMSE score which is coming out 1338.
* We can repeat the process till we get the better accuracy by changing the alpha value.

**Ridge Regression Method:**

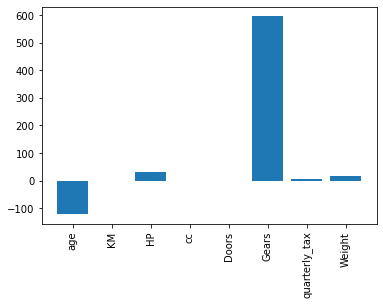
* Import the Ridge function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the Ridge function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.



* From the above plot we can see that the coefficient for KM and cc is nearly become zero and quarterly\_tax becomes nearly equal to zero but not zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.86 and calculate the RMSE score which is coming out 1371.
* We can repeat the process till we get the better accuracy by changing the alpha value.

**ElasticNet Regression Method:**

* Elastic net regression methid is the combination of L1 and L2 Regularization method
* Import the ElasticNet function by taking the alpha value = 0 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the ElasticNet function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.



* From the above plot we can see that the coefficient for KM, cc, doors becomes zero and quarterly\_tax and weight column became nearly equal to zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.86 and calculate the RMSE score which is coming out 1338.
* We can repeat the process till we get the better accuracy by changing the alpha value.

From the above 3 regression model if we compare three methods the we found that elastic net method gives the good result for same value of alpha = 0.

* For better accuracy of the model, we do the Hyperparameter tuning by using the GridSearchCV function of model\_selection package of sklearn.
* Hyperparameters tuning to be done for all the three methods of regression i.e. Lasso, Ridge and ElasticNet by taking the different alpha values as parameters and cross validation value =5 and find out the best alpha value for the best accuracy and predict the result.f
* Out of three methods after hyperparameter tuning we get the good result for Ridge regression method i.e. R2 = 0.86 and RMSE value = 1338 at best alpha value = 20

1. Data of various countries and the factors affecting their life expectancy has been recorded over the past few decades. An analytics firm would like to know how it varies country wise and what factors are influential. Use your skills to analyze the data and build a Lasso and Ridge Regression model and summarize the output. Snapshot of the dataset is given below.

A screenshot of a cell phone

Description automatically generated

Ans:- Business objective:

To find out the life expectancy based on the various factors affecting it.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Feature** | **Description** | **Type** | **Relevance** |
| Country | Country Observed. | Qualitative, Nominal | Irrelevant |
| Year | Year Observed. | Quantitative, Nominal | Irrelevant |
| Status | Developed or Developing status. | Qualitative, Nominal | Relevant |
| Life\_expectancy | Life Expectancy in age. | Quantitative, Ratio |  |
| Adult\_Mortality | Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population). | Quantitative, Ratio | Relevant |
| infant\_deaths | Number of Infant Deaths per 1000 population. | Quantitative, Ratio | Relevant |
| Alcohol | Alcohol recorded per capita (15+) consumption (in litres of pure alcohol). | Quantitative, Ratio | Relevant |
| percentage\_expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%). | Quantitative, Ratio | Relevant |
| Hepatitis\_B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%). | Quantitative, Ratio | Relevant |
| Measles | Number of reported Measles cases per 1000 population. | Quantitative, Ratio | Relevant |
| BMI | Average Body Mass Index of entire population. | Quantitative, Ratio | Relevant |
| under\_five\_deaths | Number of under-five deaths per 1000 population. | Quantitative, Ratio | Relevant |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%). | Quantitative, Ratio | Relevant |
| Total\_expenditure | General government expenditure on health as a percentage of total government expenditure (%). | Quantitative, Ratio | Relevant |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%). | Quantitative, Ratio | Relevant |
| HIV\_AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years). | Quantitative, Ratio | Relevant |
| GDP | Gross Domestic Product per capita (in USD). | Quantitative, Ratio | Relevant |
| Population | + Population - Population of the country. | Quantitative, Ratio | Relevant |
| thinness | Prevalence of thinness among children and adolescents for Age 10 to 19 (%). | Quantitative, Ratio | Relevant |
| thinness\_yr | Prevalence of thinness among children for Age 5 to 9(%). | Quantitative, Ratio | Relevant |
| Income\_composition | Human Development Index in terms of income composition of resources (index ranging from 0 to 1). | Quantitative, Ratio | Relevant |
| Schooling | Number of years of Schooling(years) . | Quantitative, Ratio | Relevant |

Steps For the multilinear Regression by using Lasso and Ridge Regression Method:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

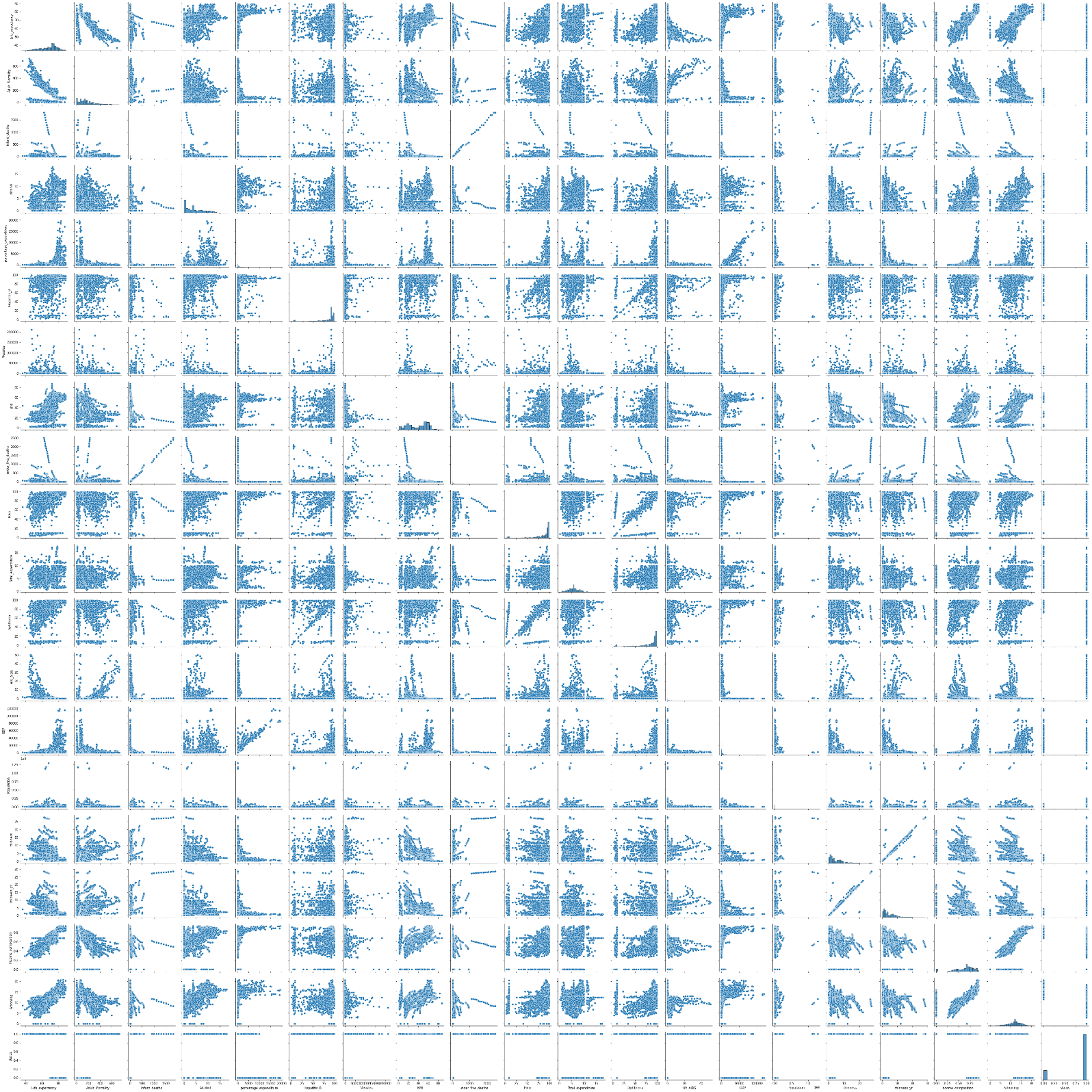
Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Find the correlation between the variables and by using the corr() function and plot the pairplot.



* The above graph is the pair plot which shows the correlation matrix from which we can see the correlation between different variables.

# From the pairplot and correlation table we can say that schooling is more correlated with the life expectancy

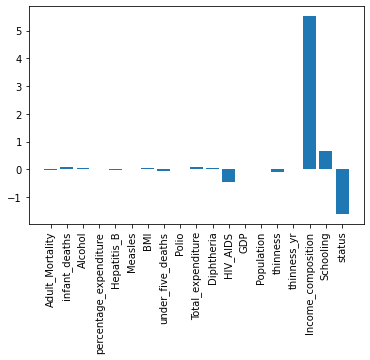
# infants death and under5 death are more correlated with each other

# GDP and percentage expenditure are more correlated with each other

# Income composition and life expectancy are correlated with each other.

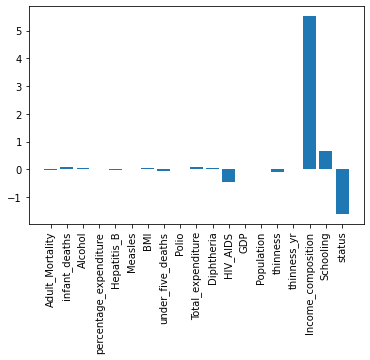
* Preparing multilinear regression model by considering all the variables and view the summary.
* From the summary we have to see the R2 value is 0.820 and adjusted R2 value is 0.819 which is lesser than 0.85.
* Calculate the RMSE value which is found out to be 4.0
* To increase the accuracy of the model and to reduce the error we use Lasso and Ridge regression method.

**Lasso Regression Method:**

* Import the Lasso function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the the Lasso function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.
* From the above plot we can see that the coefficient for percentage expenditure, measles, GDP, thinness\_yr became zero and other coefficients value are also adjusted according the correlation of the variable with the output variable and take only values which has some coefficient value and avoid the variable which has the coefficient zero (feature selection)
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.82 and calculate the RMSE score which is coming out 4
* We can repeat the process till we get the better accuracy by changing the alpha value.

**Ridge Regression Method:**

* Import the Ridge function by taking the alpha value = 0.15 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the Ridge function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.

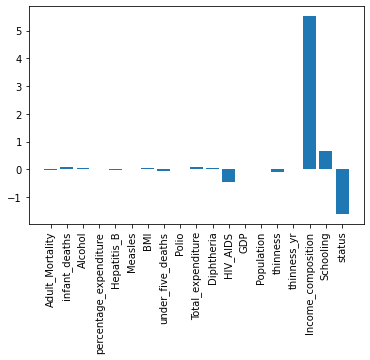


* From the above plot we can see that the for percentage expenditure, measles, GDP, thinness\_yr became zero and most of the variables becomes nearly equal to zero but not zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.82 and calculate the RMSE score which is coming out 4
* We can repeat the process till we get the better accuracy by changing the alpha value.

**ElasticNet Regression Method:**

* Elastic net regression methid is the combination of L1 and L2 Regularization method
* Import the ElasticNet function by taking the alpha value = 0 (alpha can be between 0 to 1) from the linear\_model package of sklearn.
* Define the ElasticNet function and fit the model on the data.
* Calculate the coefficient and intercept value and plot the bar graph between the coefficient value and the columns.
* From the above plot we can see that the coefficient for KM, cc, doors becomes zero and quarterly\_tax and weight column became nearly equal to zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.86 and calculate the RMSE score which is coming out 1338.
* We can repeat the process till we get the better accuracy by changing the alpha value.

From the above 3 regression model if we compare three methods the we found that elastic net method gives the good result for same value of alpha = 0.

*  for percentage expenditure, measles, GDP, population, thinness\_yr became zero and most of the variables becomes nearly equal to zero but not zero and coefficients value are also adjusted according the correlation of the variable with the output variable.
* Do the prediction on the data and find out the Adjusted R2 score which is come out 0.82 and calculate the RMSE score which is coming out 4
* We can repeat the process till we get the better accuracy by changing the alpha value.

From the above 3 regression model if we compare three methods the we found that elastic net method gives the good result for same value of alpha = 0.

* For better accuracy of the model, we do the Hyperparameter tuning by using the GridSearchCV function of model\_selection package of sklearn.
* Hyperparameters tuning to be done for all the three methods of regression i.e. Lasso, Ridge and ElasticNet by taking the different alpha values as parameters and cross validation value =5 and find out the best alpha value for the best accuracy and predict the result.f
* Out of three methods after hyperparameter tuning we get the good result for Lasso regression method i.e. R2 = 0.82 ans RMSE value = 4 at best alpha value =0.001