Fall
2023

Assignment 01

CSE 574 (SECTION: D)

NAZMUS SAQUIB

UBIT: NSAQUIB2

PERSON NUMBER: 50510460

Part 1

Ans to ques. 1

I have used 2 datasets here. One is “penguins”, another is “emissions_by_country”.

Penguins:

This dataset has 10 features and 344 rows. This dataset is about 3 species of penguins from 3 different islands. In the dataset, penguins’ physical features and a few behaviors are stated briefly in numbers.

Emissions:

This dataset has 13 features and 63104 rows. It consists of environmental, financial, and type of emissions data of more than 220 countries in a brief. There are about 319695 missing values in the dataset which is huge. To make this dataset with such a huge missing value I cleaned it thoroughly and modified it in a usable form. Among all emissions done by the countries mentioned in the dataset, oil based emission is the most with a mean of 39076.41 and a standard deviation of 819.50.

Ans to ques. 2

The data description methods I used are described below.

dropna(): I used it to drop rows containing empty values in particular columns.

value_counts(): I used it to count particular records in columns to identify the spelling mismatches.

replace(): Then with replace method, I fixed all the spelling mismatches with the same spelling.

corr(): With this I found out the co-relation between features to identify important ones.

factorize(): This one is used to convert the string values to numeric for better training.

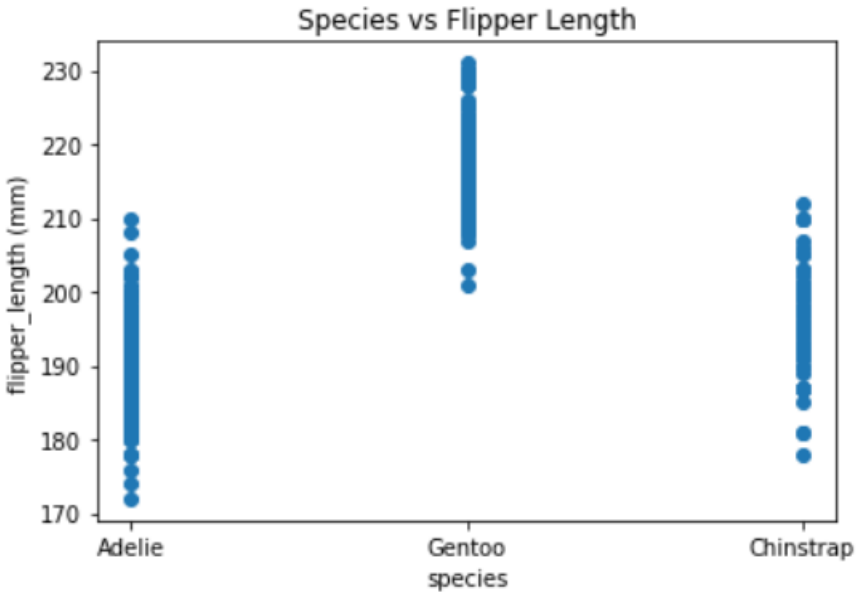
fillna(): This one is to replace NaN values in my directed way. There are plenty of NaN or empty records in the emission dataset. For performing the best analysis, I kept data of 1990-2020 because in this period, most countries have proper data. Still there are many empty cells with I either removed or replaced NaN using the “mean” of columns.

normalize(): I wrote this custom function to normalize data between 0 and 1.

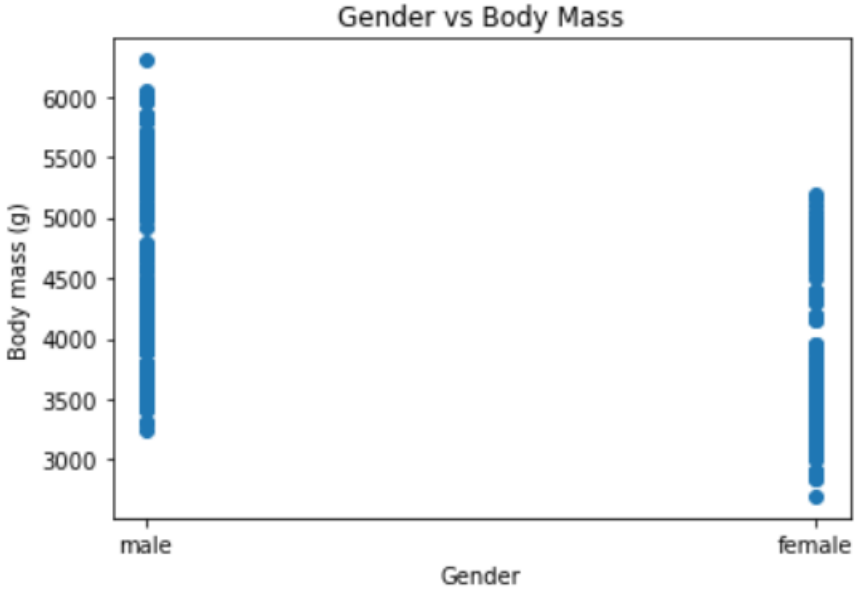
Ans to ques. 3

Penguins:

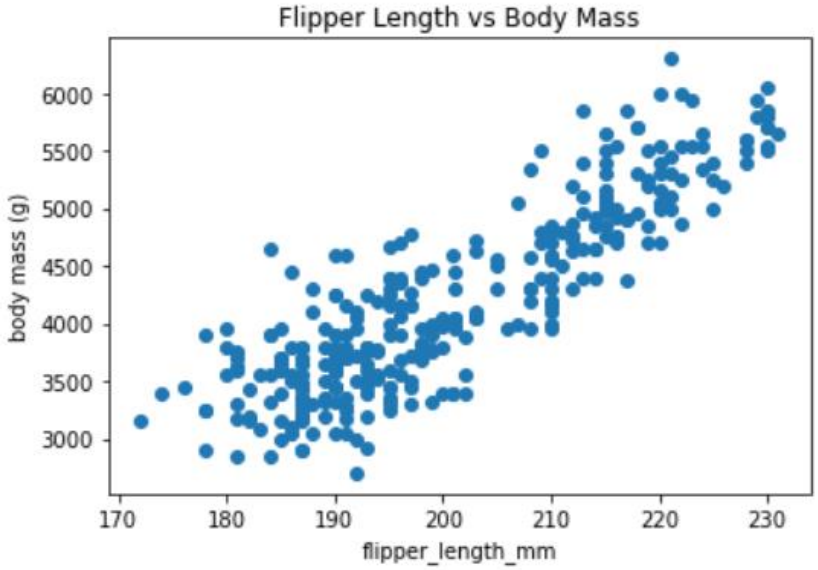
For Penguins dataset I plotted 5 graphs. All the 5 plots are given below.



The above plot shows that flipper length depends on the species. Flippers of Gentoo are usually bigger than the other ones.

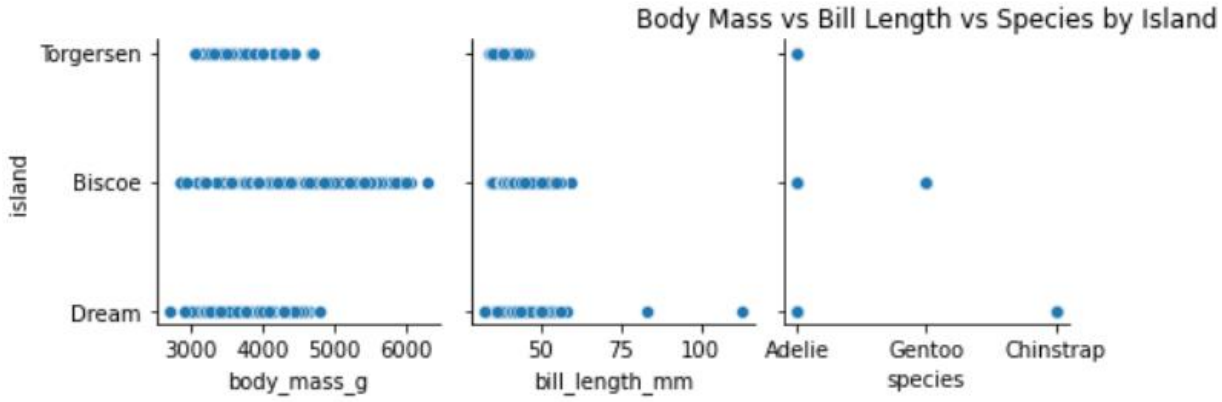


The above plot shows that body mass varies by gender. Male penguins are usually heavier than the female ones.



The above plot shows that body mass of penguins increase with their flipper size.

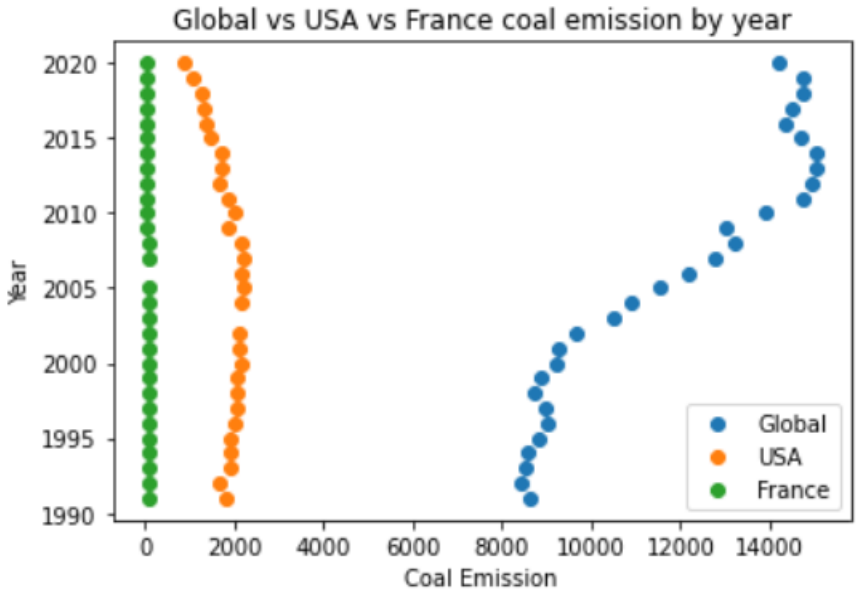
Above is a co-relation matrix which shows that flipper_length, body_mass, bill_length are the most important features to use for finding out insights from this dataset.



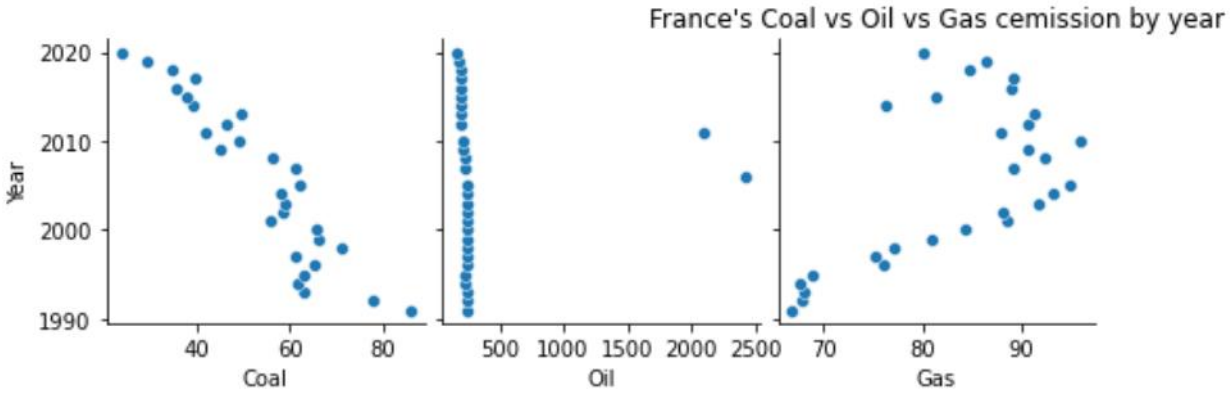
The above plot shows the variation of penguins' body mass, bill length, and species based on the islands they live on.

<u>Plots of emission dataset:</u>

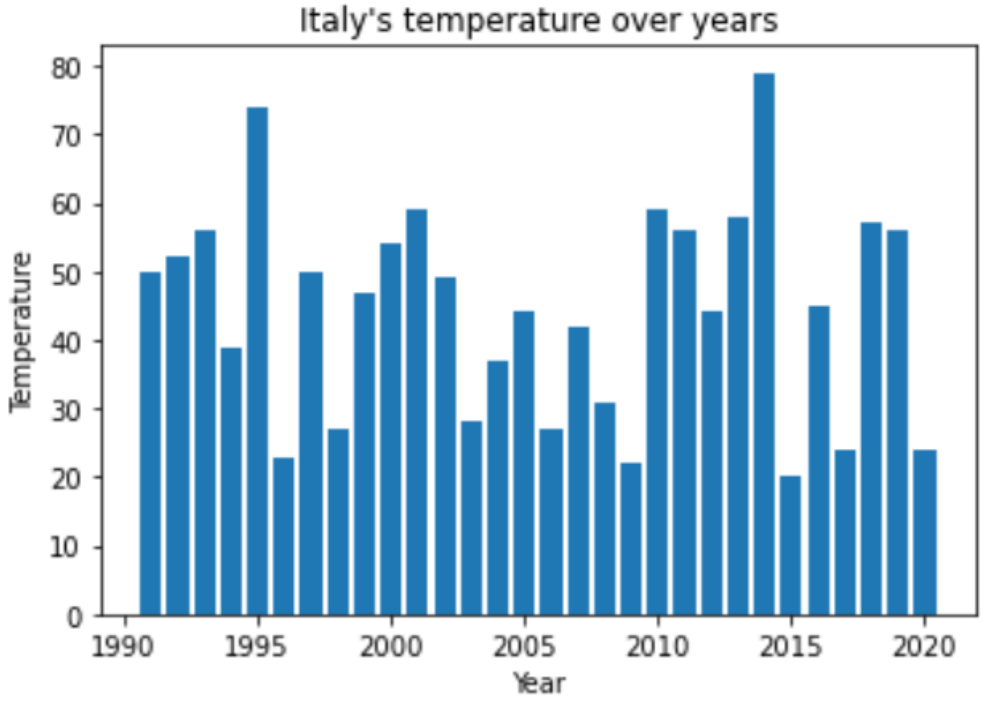
For emission dataset I plotted 5 graphs. All the 5 plots are given below.



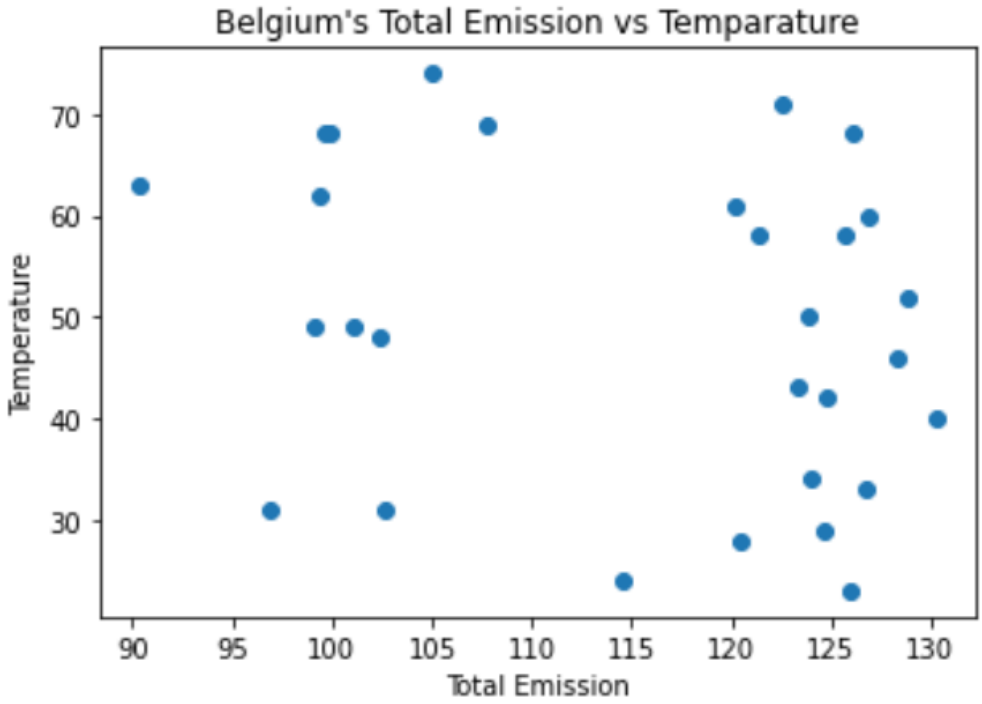
The above plot shows coal emissions of France, USA, and all the other countries combined.



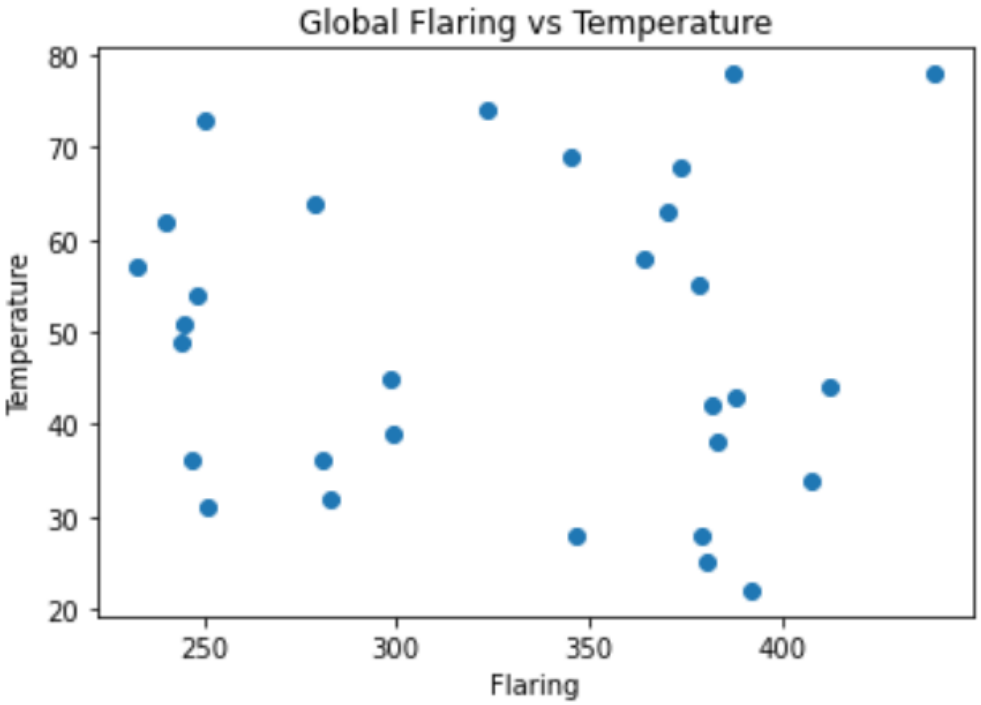
The above plot shows that oil takes major portion of the emission by France.



The above plot shows the temperature of Italy over years. The fluctuations did not change much over the last 30 years.



The above plot shows that Belgium's Total emission and temperature doesn't follow any synchronization. So, emissions don't affect their temperature much.



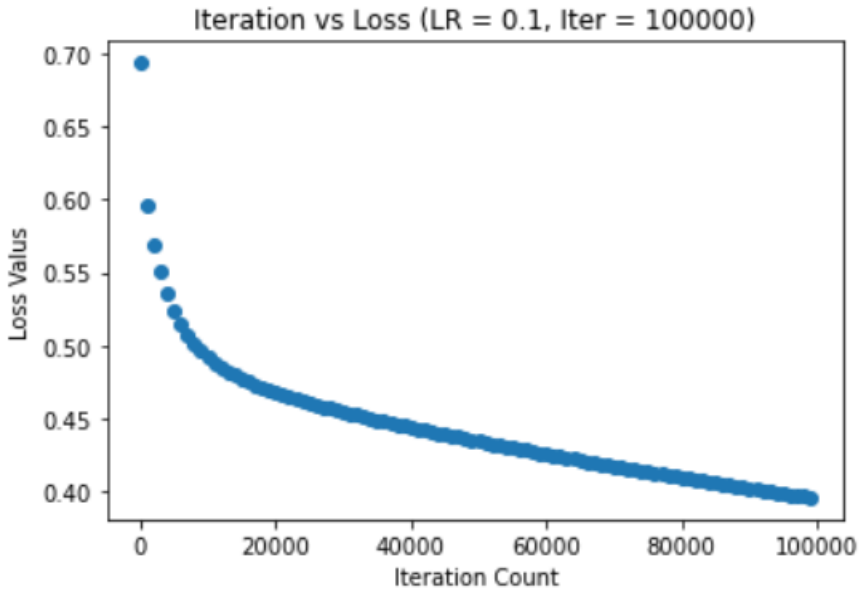
The above plot shows that global flaring mostly increases global temperature.

Part 2

Ans to ques. 1

Best Accuracy = 74.6%

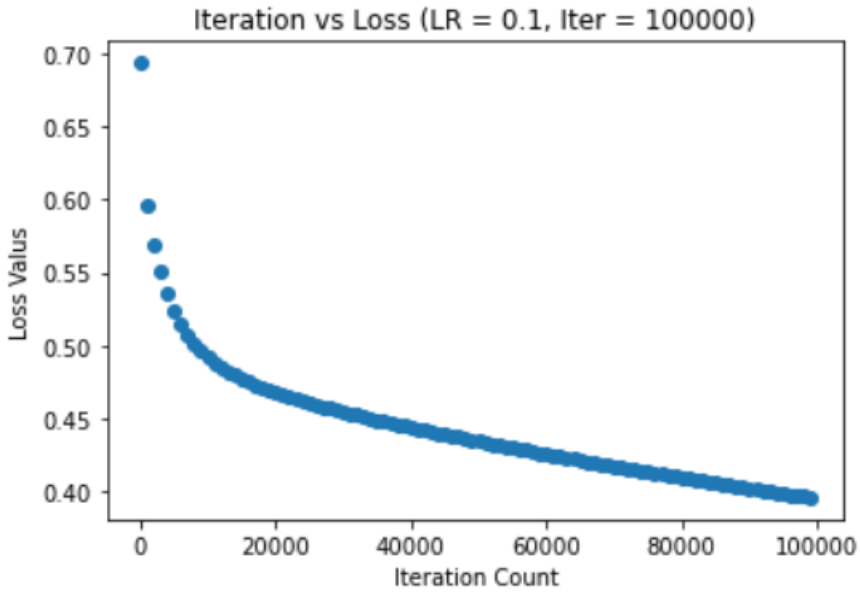
Ans to ques. 2



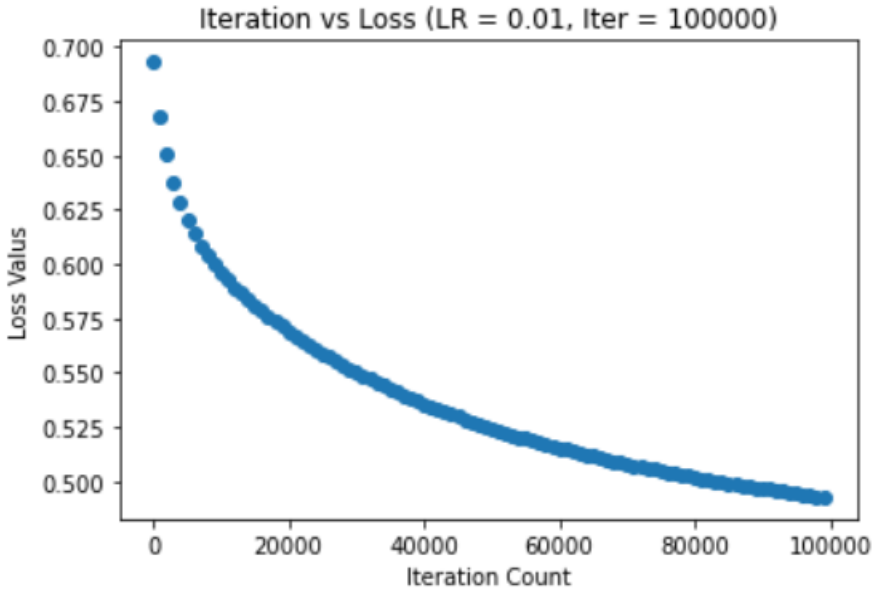
The graph shows that if we take learning rate of 0.1 with 100000 iterations, the loss value decreases too fast.

Ans to ques. 3

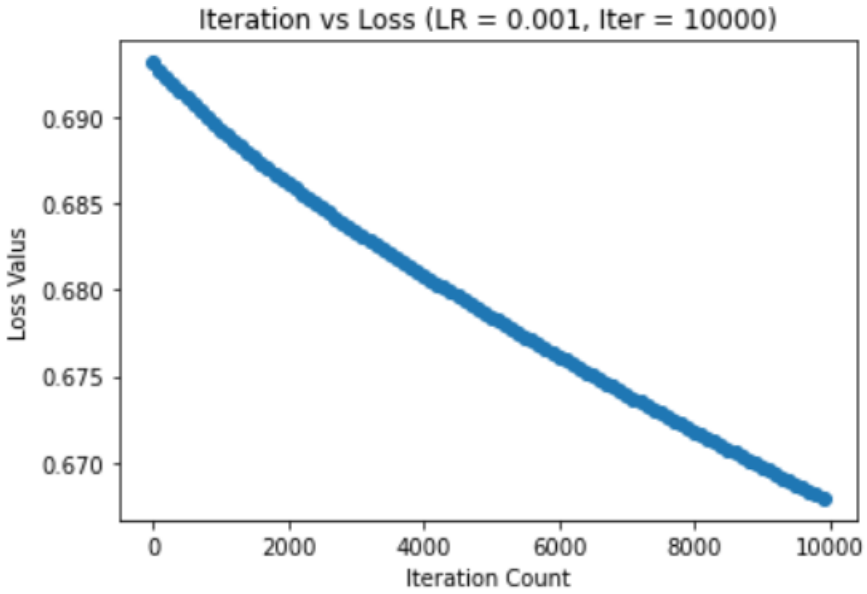
I tried 3 different set ups of iteration and learning rate and found different loss values over number of iterations. I plotted the graphs (snapshots attached) and discussed the impact and loss over the iterations below.



If we take learning rate of 0.1 with 100000 iterations, the loss value decreases too fast resulting in a model accuracy of 65.08%.



If we decrease the learning rate and make it 0.01 with the same 100000 iterations, the loss value decreases a bit slower than the previous one, resulting in a model accuracy of 57.14%. So, decreasing the learning rate is decreasing our accuracy for this case.



Finally, if I decrease the learning rate and make it 0.001, and at the same time decrease the iterations, the loss value decreases much straighter than the previous one, resulting in a model accuracy of 74.06% which is the highest as per my observation. So, decreasing the learning rate and number of iterations at the same time is increasing the accuracy for this case.

The learning rate affects the step size while optimizing. From the above visualization my understanding is that a model with a too fast learning rate may exceed the optimal parameters, whereas a model with too low learning rate may converge slowly or become trapped in poor solutions. Tuning the learning rate can have a significant influence on the training process and accuracy.

Ans to ques. 4

Benefits:

Regression can handle irrelevant features without affecting its performance considerably. The weight feature assigns small weights to uninformative features to get less impact from them. It can also handle enormous datasets effectively and does not require a lot of computer power.

Drawbacks:

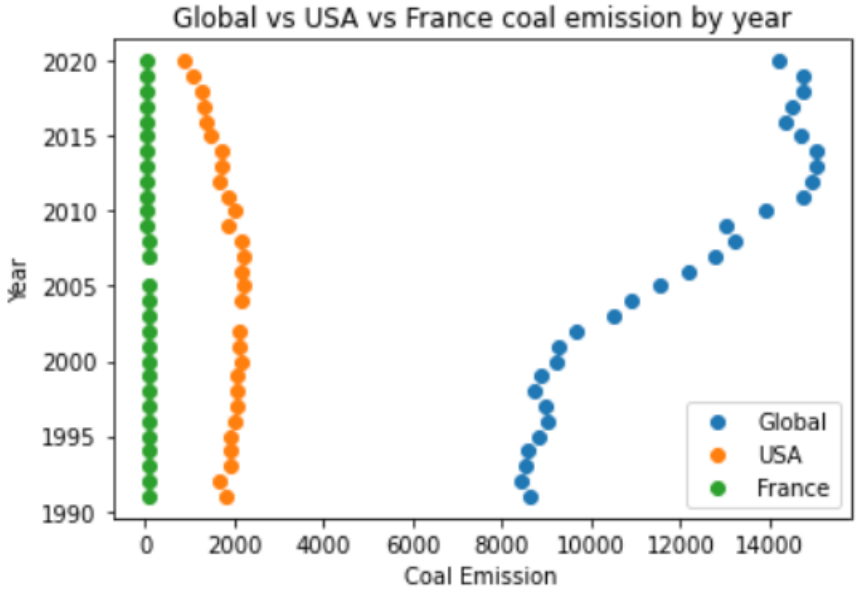
Logistic regression fails to predict a continuous outcome. In addition, if the sample size is too small, the conclusion may be inaccurate.

Part 3

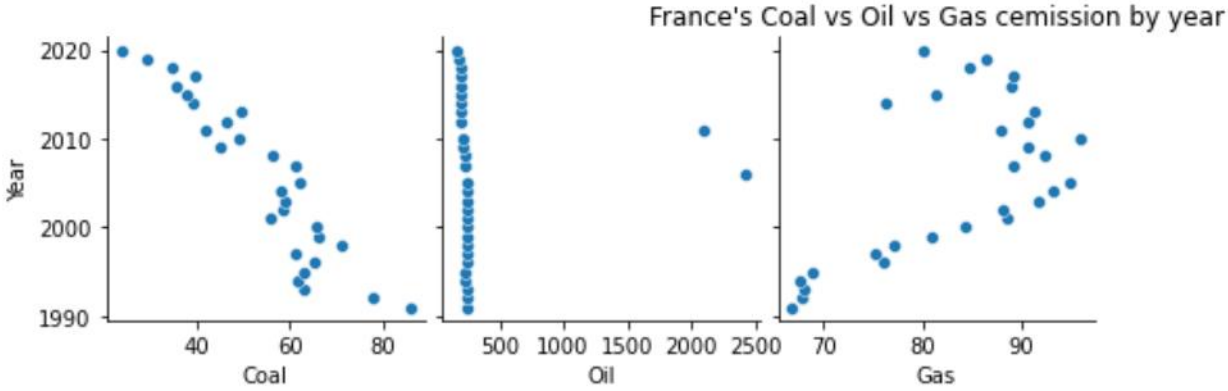
Ans to ques. 1

This dataset has 13 features and 63104 rows. It consists of environmental, financial, and type of emissions data of more than 220 countries in a brief. There are about 319695 missing values in the dataset which is huge. To make this dataset with such a huge missing value I cleaned it thoroughly and modified it in a usable form. Among all emissions done by the countries mentioned in the dataset, oil based emission is the most with a mean of 39076.41 and a standard deviation of 819.50.

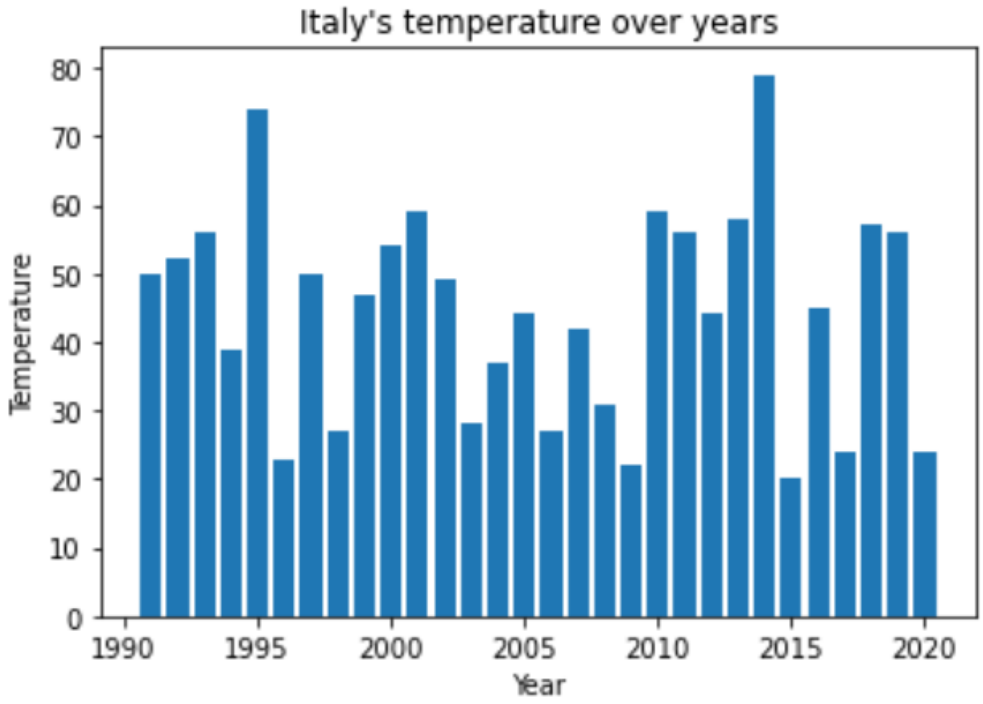
Five visualization plots of the dataset are given below.



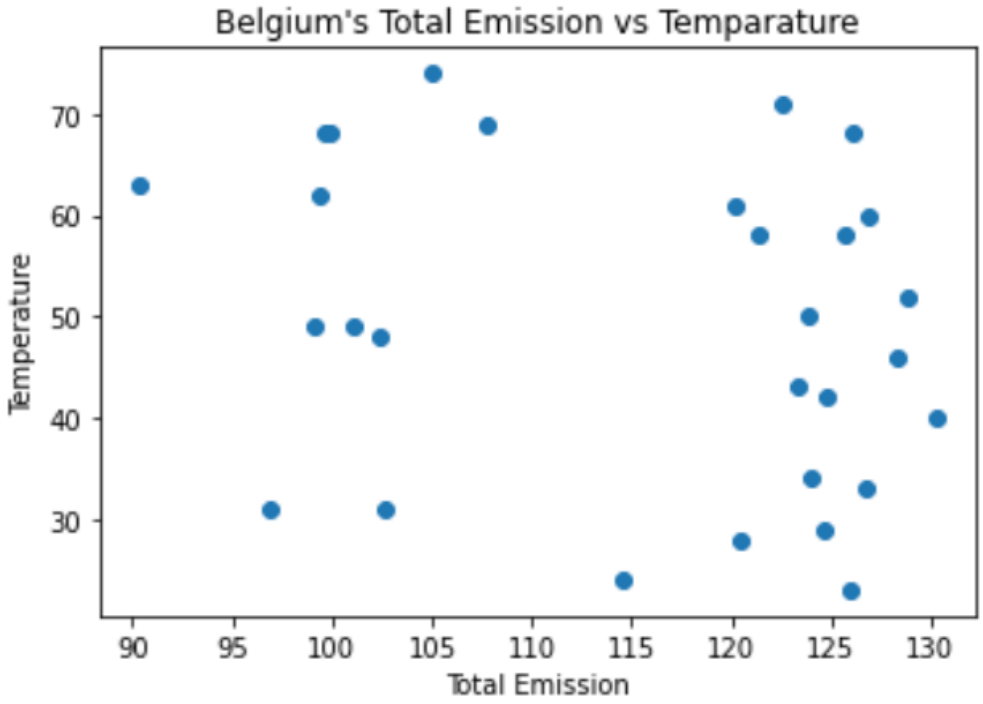
The above plot shows coal emissions of France, USA, and all the other countries combined.



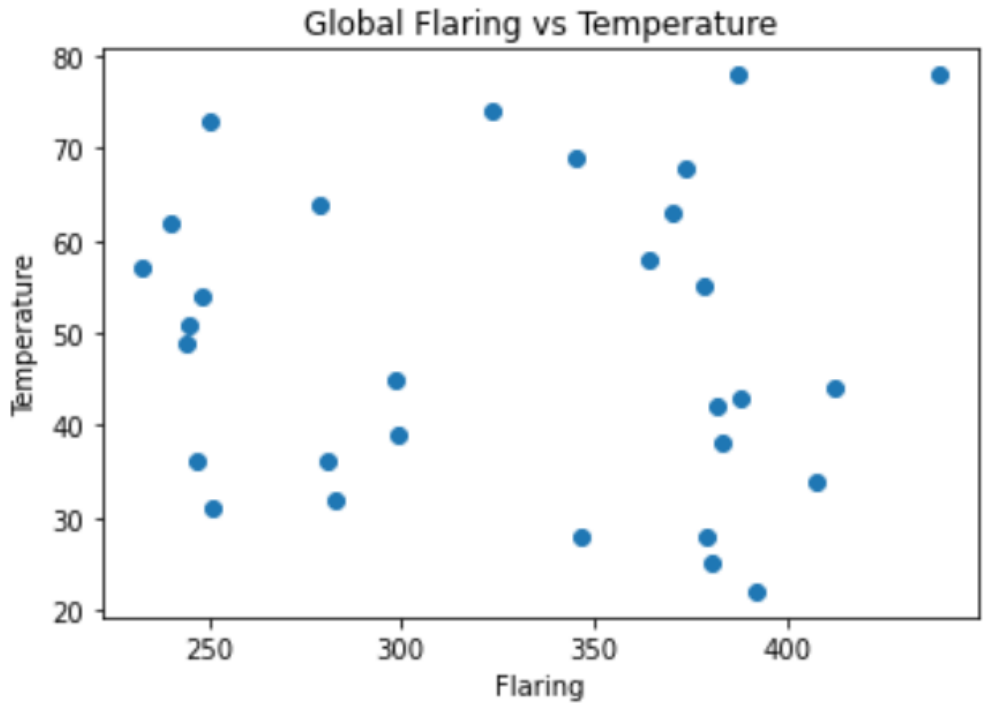
The above plot shows that oil takes major portion of the emission by France.



The above plot shows the temperature of Italy over years. The fluctuations did not change much over the last 30 years.



The above plot shows that Belgium's Total emission and temperature doesn't follow any synchronization. So, emissions don't affect their temperature much.



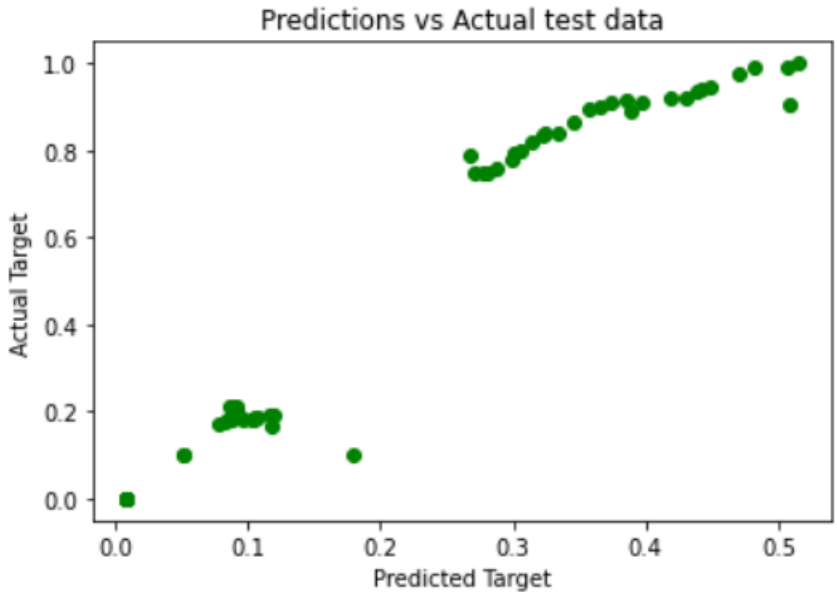
The above plot shows that global flaring mostly increases global temperature.

Ans to ques. 2

After finding out gradient descent and running loss function, I found,

Loss Value = 0.0009631996945221432

Ans to ques. 3



Above is the plot comparison of predicted target vs actual target after linear regression. It proves that our actual output has a linear relation with the predicted outputs.

Ans to ques. 4

Benefits and drawbacks of using OLS estimate for computing the weights are given below.

Benefits:

OLS is simple to apply and comprehend. It gives a straightforward closed-form method for estimating model coefficients, making it suitable for a wide variety of users. OLS estimators are efficient and have acceptable statistical features when the requirements of linear regression are satisfied.

Drawbacks:

OLS is vulnerable to data outliers. Extreme values can have an outsized impact on the model, resulting in coefficients that do not adequately represent the remainder of the data. Furthermore, while OLS does not need a particular distribution, it frequently assumes that errors are normally distributed.

Ans to ques. 5

Benefits and drawbacks of Linear Regression are given below.

Benefits:

Linear regression is simple and effective. The concept is intuitive and easy to explain which made it user friendly. Also. linear regression is computationally efficient, especially when the number of independent variables is minimal. A linear regression model may be trained and predicted quickly.

Drawbacks:

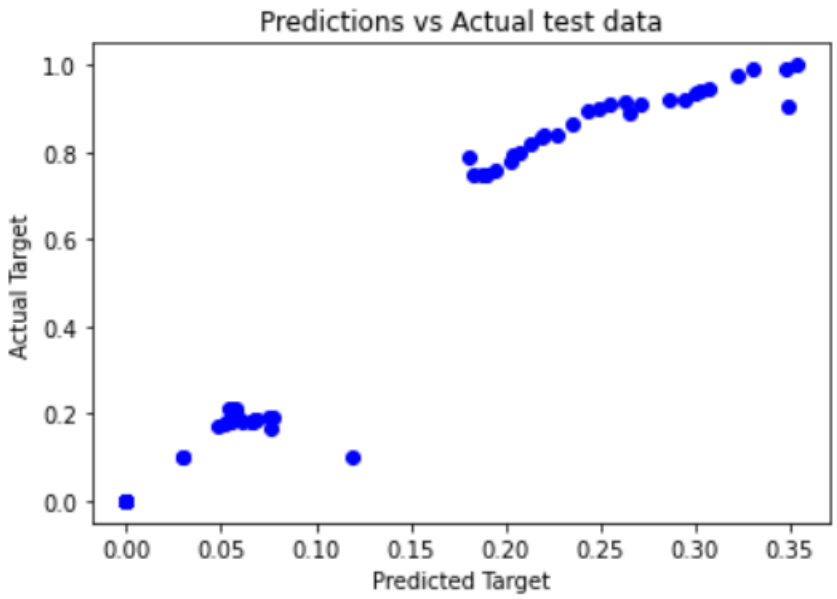
One drawback is linear regression presupposes a linear connection between independent and dependent variables. The model may provide erroneous results for non-linear relationships. It is also susceptible to outliers, which can have a major influence on the model's coefficients and predictions. Outliers must be detected and dealt with appropriately.

Part 4

Ans to ques. 1

Loss Value = 0.00032665

Ans to ques. 2



Above is the plot comparison of predicted target vs actual target after ridge regression. The linear relation between the actual target and the predicted outputs are a bit better than the linear regression shown before.

Ans to ques. 3

The ordinary least squares (OLS) loss function is used in linear regression. It seeks to reduce the sum of squared errors between projected and actual target values. Ridge regression, on the other hand, employs a modified loss function that incorporates the sum of squared weights as a penalty term in addition to the OLS loss.

The major reason for applying L2 regularization is to increase the model's generalization. It is useful when linear regression may overfit or when there are several correlated features. Ridge regression supports a smoother model with more evenly weighted features by include the L2 regularization term.

Ans to ques. 4

Benefits and drawbacks of Ridge Regression are given below.

Benefits:

Ridge regression fosters a smoother model by applying regularization. It aids the model's generalization to previously unknown data, making it more robust and stable. In ridge regression, the regularization term decreases the variance in the model's predictions. When you have little data, this is critical since it produces more reliable estimations.

Drawbacks:

Ridge regression is particularly useful when there is cause to suspect overfitting. If this concerns do not exist, ridge regression may not be as beneficial as regular linear regression.