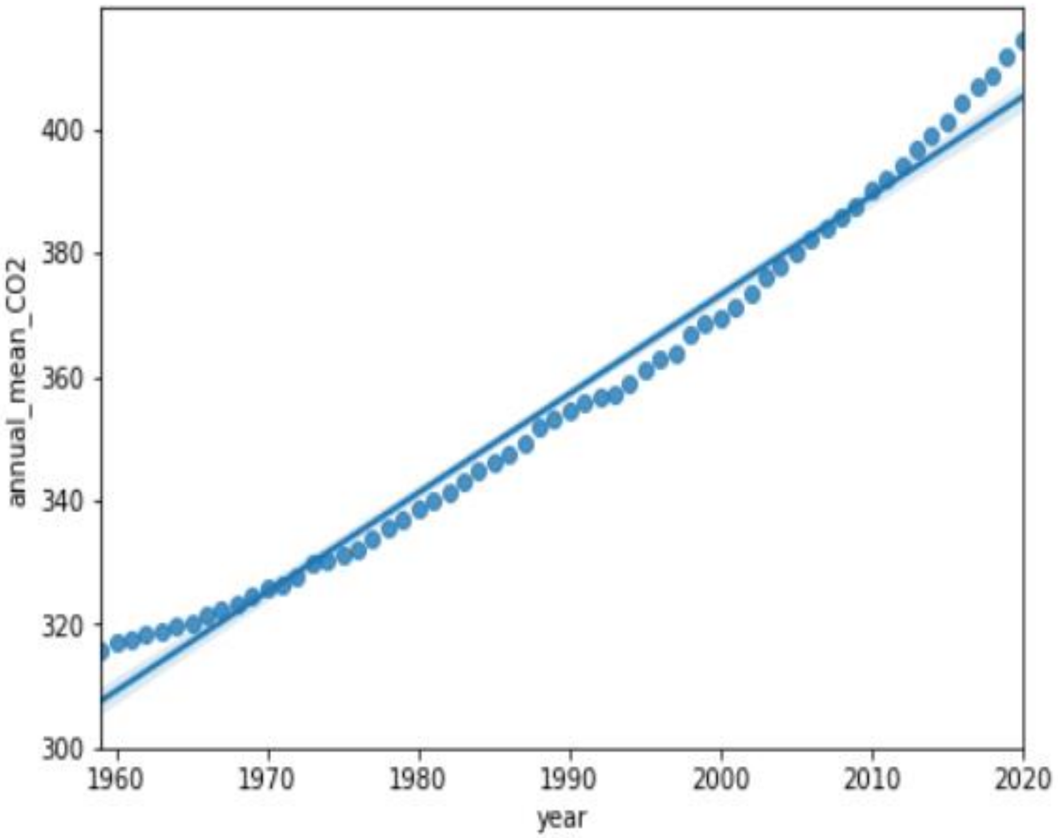
Assignment 10 (50 pts): Building Models: Regression

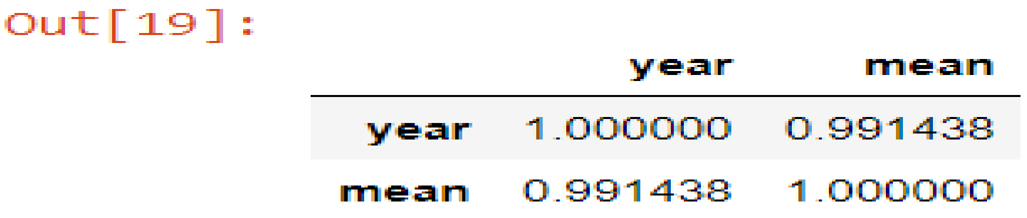
1. **Using the co2_annmean_mlo.txt file, create a csv file for analysis (any method).**

a. Copied the data into a text file CO2.txt, and read it using python
b. Converted the raw data into a list of data, and then into a Data frame
c. Cleaned the Data frame and exported it into CO2.csv file

2. **Create and interpret a linear model showing the relationship between year and annual mean CO_2 (using R or Python).**

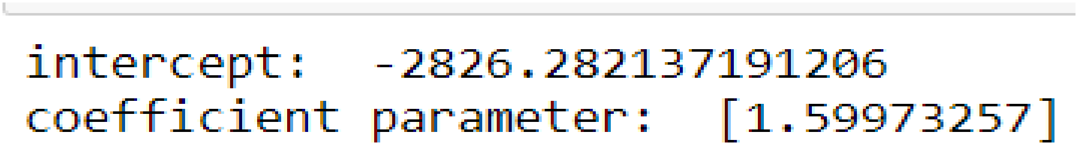


The correlation coefficient (R) between year and annual mean CO2 is almost equal to 1, i.e., 0.991438. This shows that there is a strong positive linear relationship between these two variables.


Out[19]:

	year	mean
year	1.000000	0.991438
mean	0.991438	1.000000

Interpretation of the Linear Model:



```
intercept:  -2826.282137191206
coefficient parameter:  [1.59973257]
```

Here the slope is 1.59973257. The linear equation is as follows (y (hat*) = b0+b1x)

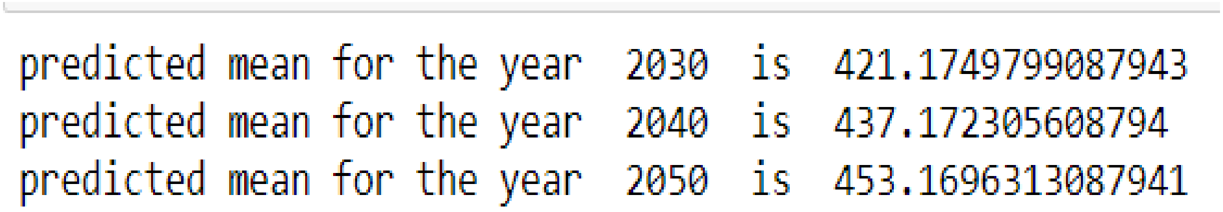
annual_mean_CO2 (hat*) = (-2826.282137191206) + ((1.59973257)*(year))

From the above linear equation, we can interpret that for every 1 unit increase in year, the expected value of annual_mean_CO2 increases by 1.59973257 times.

*predicted/expected value

3. **Using your model, estimate and interpret the predicted mean CO_2 for the next three decades (2030, 2040, 2050).**

The expected annual_mean_CO2 for years 2030, 2040, 2050 is 421.175, 437.173 and 453.170 respectively.



```
predicted mean for the year  2030  is  421.1749799087943
predicted mean for the year  2040  is  437.172305608794
predicted mean for the year  2050  is  453.1696313087941
```

We know that from the above liner equation, for every 1 unit increase in year, the expected value of annual_mean_CO2 increases by 1.6 times. So for every 10 units increase in year, the expected value of annual_mean_CO2 increases by 16 times which can be clearly seen from the above output. (437.17 – 421.17 = 16) / (453.17 – 437.17 = 16)

4. **Explain how your model's accuracy can be improved.**

   Ways to improve model's accuracy:

   1. Building many model's with different combination of independent variables and comparing them can help us arrive at a model with more accuracy

2. Checking if there are any outliers and treat them because these may have effect on the predicted values (exploratory data analysis)
3. We can judge our model by evaluating some factors like R square, coefficient values, p values
4. Checking if the residuals are nearly normal with a histogram and also if there is constant variability
5. Check if there is any structure in the data as its not accounted for best least squared line
6. Compare model's by transforming the variables like taking logarithmic values or normalizing or standardizing the data
7. Creating new variables out of the existing variables can also help to some extent

References:

Abhirami Sankar. Jigsaw Academy. https://www.jigsawacademy.com/5-super-tips-to-improve-your-linear-regression-models/

Sandeep Ram. Towardsdatascience.com. https://towardsdatascience.com/statistics-supporting-linear-models-bfc24fb9781f