**Bill Yerkes**

**CS5542 Big Data Apps and Analytics**

**In Class Programming –9**

**22nd October 2020**

**Submit ICP Feedback in Class. :** [**Lnik to Feed back Form**](https://docs.google.com/forms/d/e/1FAIpQLSesllFh5_STnj7RbHyQainRG_2EIKw1csp8ObP5FWjpVnGVOg/viewform)

**Variational Autoencoders:**

**Create a linear regression model in python using any dataset of your choice. For this model you can also create your own data. Find the best fit line in the data and calculate SSE (sum of square error) or MSE (Mean square error) , Y intercept, and Slope for the relationship in data. Explain your findings and understanding of these terms in detail in the report.**

ICP Requirements:

1. Successfully executing the code with linear regression model and calculating following:
   1. SSE or MSE
   2. Y intercept
   3. Slope

(75 points)

1. Detail explanation of each in report (5 points)
2. overall code quality (10 points)
3. Pdf Report quality, video explanation (10 points)

Submission Guidelines:

Same as previous ICPs.

**ICP Report:**

**What I learned in the ICP:**

I learned another method of consuming / linking data in CoLab. I learned how to compute the Mean Square Error, Slope, and Y intercept for linear regression by hand and by using the Regression Model object.

Looking at the data and the results of the linear regression, the data does suggest that the SAT store can predict how a student will do in college. However, it is not a strong, compact correlation. There are students who get a total score of over 1400 (SAT only had Math and Verbal at this time), and they had a GPA of around 2. There are students who had a total score of under 1000, and their GPA was around 4. So there is a good amount of error, variance, in the prediction, but in general the SAT does give a general indication on how a student will do in College, based upon this given data set.

**Description of what task I was performing:**

I chose to use a data set of SAT (Standardize College entrance tests) and GPA scores as a model to predict how well a student will due in college, based up their GPA in college.

The first part of the lab, computed the values by hand, and the second part used the LinearRegression object to produce the answers.

The first item that need to be computed was the slope.

x is the SAT total score (sat\_sum)

y is the GPA (fy\_gpa)

x\_mean is the mean of x

y\_mean is the mean of y

n is the number of elements

Sxy = sum(x \* y) – n \* x\_mean \* y\_mean

Syy = sum(x \* x) – n \* x\_mean \* x\_mean

**Slope = Sxy / Sxx**

Now that we have the slope, we can compute the y intercept.

**y\_intercept = y\_mean – slope \* x\_mean**

Now that we have the y\_intercept we can compute the Mean Squared Error.

y\_pred = slope \* x + y\_intercept

error = y – y\_pred

se = sum(error^2)

**Mean Squared Error = se / n (mse = se/n)**

Using the LinearRegression object, I just had to plug in the values and the object did all of the math for me.

Started with calling the fit method and passing in the x and y values:

x = x.reshape(-1,1)

Reshaping x, so it can be used in the object.

regressionModel = LinearRegression()

regressionModel.fit(x,y)

Next we compute the expected value of y:

y\_predicted – regressionModel.predict(x)

Next we can use the object to give us the Mean Squared Error:

mse = mean\_squared\_error(y,y\_predicted)

Finally we can use the object to give the Slope and Y intercept:

slope is regressionModel.coef\_[0]

and

Y intercept is regressionModel.intercept\_

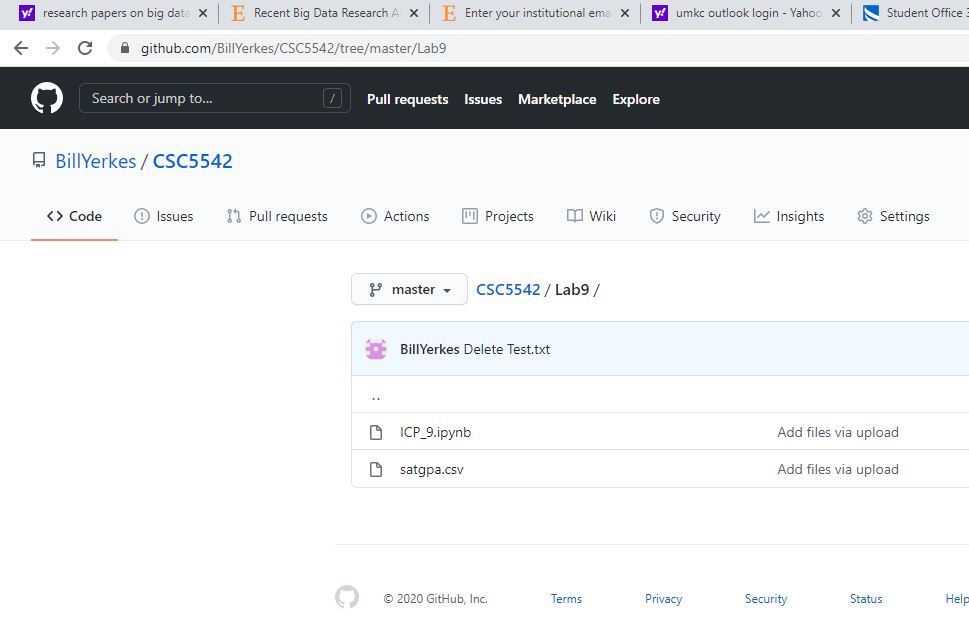
**Challenges I faced:**

The biggest challenge I faced was fining a data set to work with.

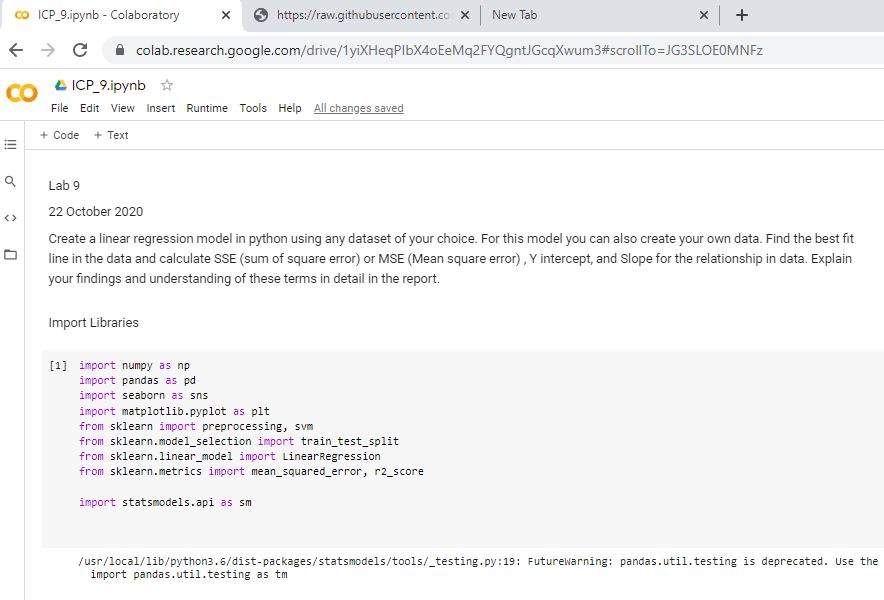
**Screen Shots**

[GitHub:](https://github.com/BillYerkes/CSC5542/tree/master/Lab9)

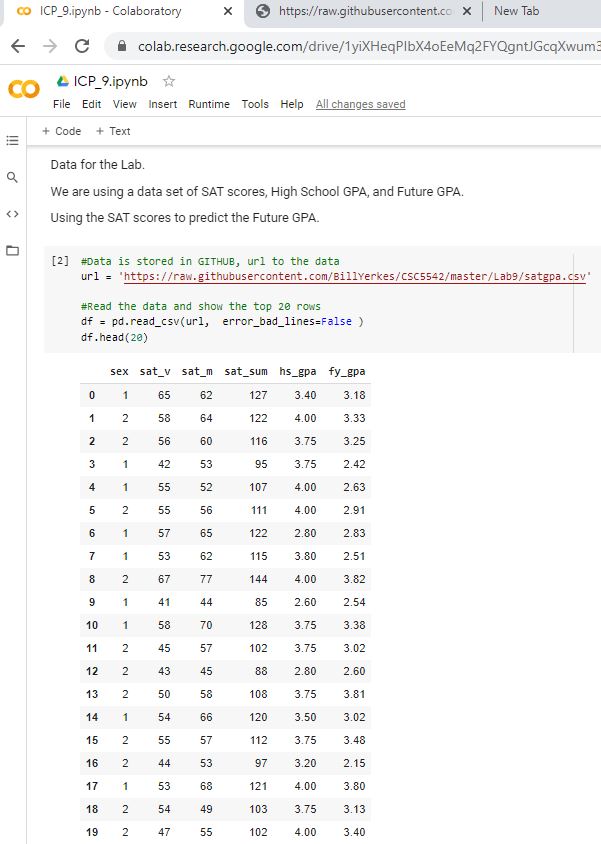
Image of GitHub repository

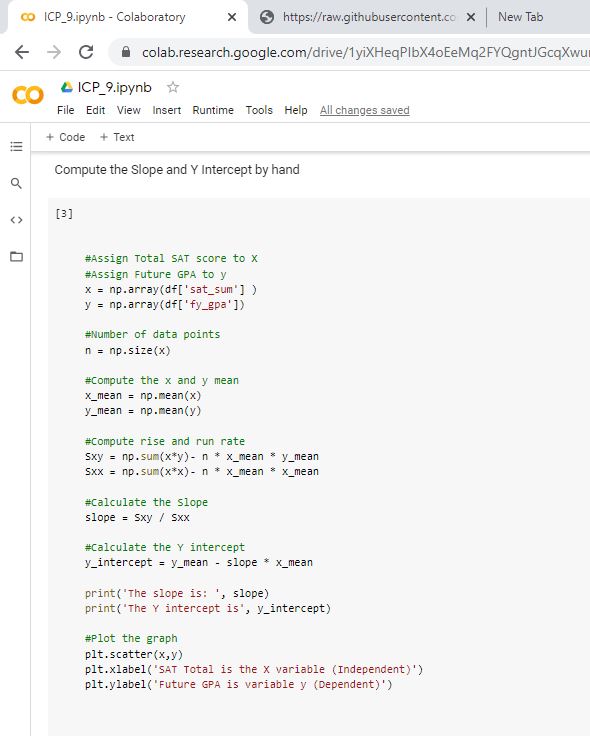


**Lab 9, Libraries declaration:**

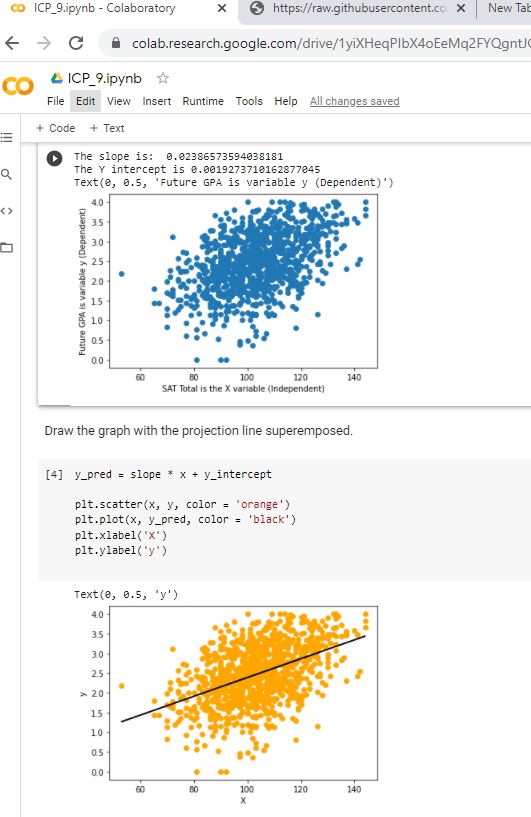


**Data for the Lab. Using SAT / GPA data. Determining if GPA can be predicted based upon SAT Score.**

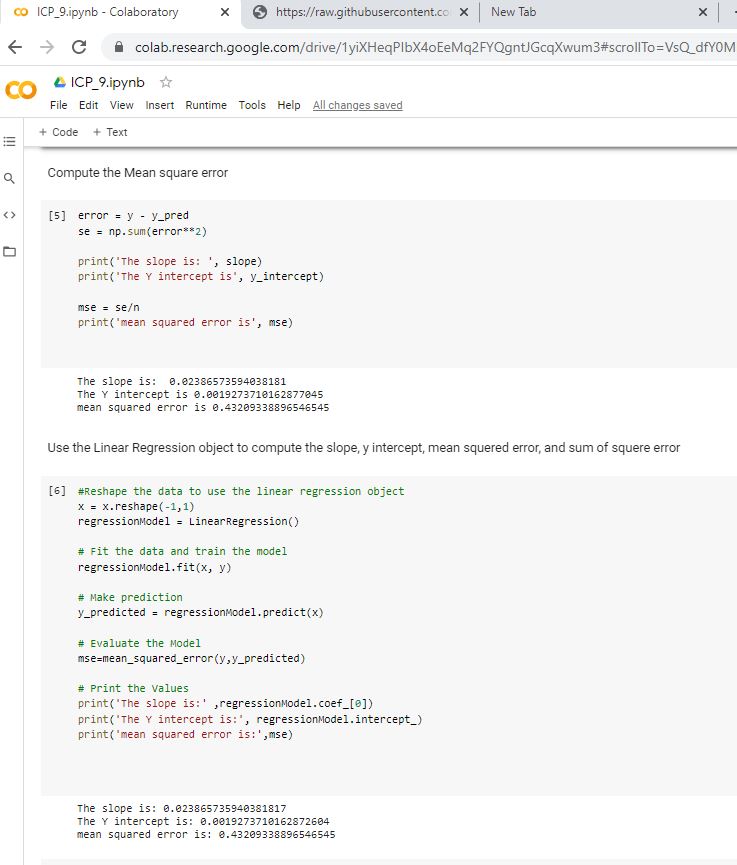


**Computing the Slope and Y Intercept by hand**

**Plotting the Data:**



**Compute Mean Square Error and use Linear Regression Object to produce the same results:**



[**Video Link**](https://youtu.be/LvB6Km4e8e4)

**Any in site about the data or the ICP in general**

I enjoyed the lab, going to try and figure out on my own how to add more dimensions to the model and see if that improve the prediction or not.