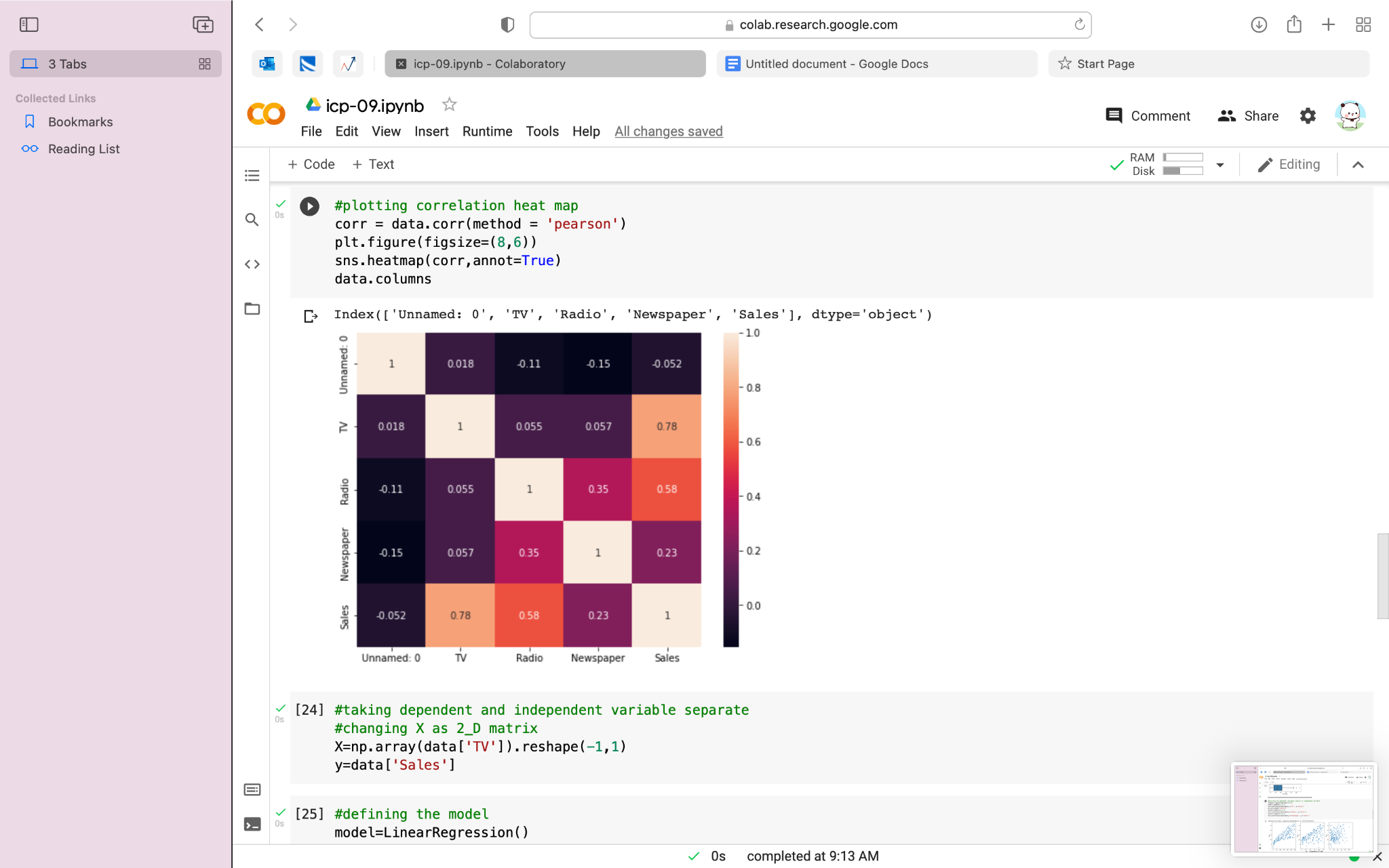
**WIKI REPORT ICP-09**

**A)** I have learned how to create a linear regression model in python using a given dataset. And also learned how to find a best fit line in the data and how to calculate SSE (sum of square error), MSE(mean square error) and y-intercept and slope for the relationship in data.

**B)**The task that I was performing was finding the best fit line in the data and calculating SSE, MSE, Y intercept and slope for the relationship in data.

First I have imported all the required libraries like pandas, numpy for mathematical purposes from google colab import files for uploading a file from local disk to google colab. For visualization I have imported seaborn and matplotlib for plotting graphs and imported Machine Learning libraries like LinearRegression and MeanSqaureError to perform the given task.

In the Dataset there are 4 columns, First 3 columns are budgets spent for different marketing types for advertising and last column is sales. Sales here are dependent variable and the rest of the columns are independent variables. Here I have to predict sales in this scenario using a linear regression model. Then I have checked for null values and then performed outlier Analysis and I have observed that there is no considerable outliers data. Then plotted the dependent variable (Sales) and independent variable (TV,Radio,Newspaper). From the plot i have observed that column ‘TV’ has more of a linear relationship with column ‘Sales’, So i’m going to find and plot correlation between Sales and the rest of the columns like tv, radio, newspaper. Plotted the correlation heat map using the ‘**Pearson**’ method. The Pearson map tells about the correlation between the sales and rest of the columns. In this map if we get -1 then it is negatively correlated , if it is +1 it is positively correlated. If it is 0 then there is no correlation.

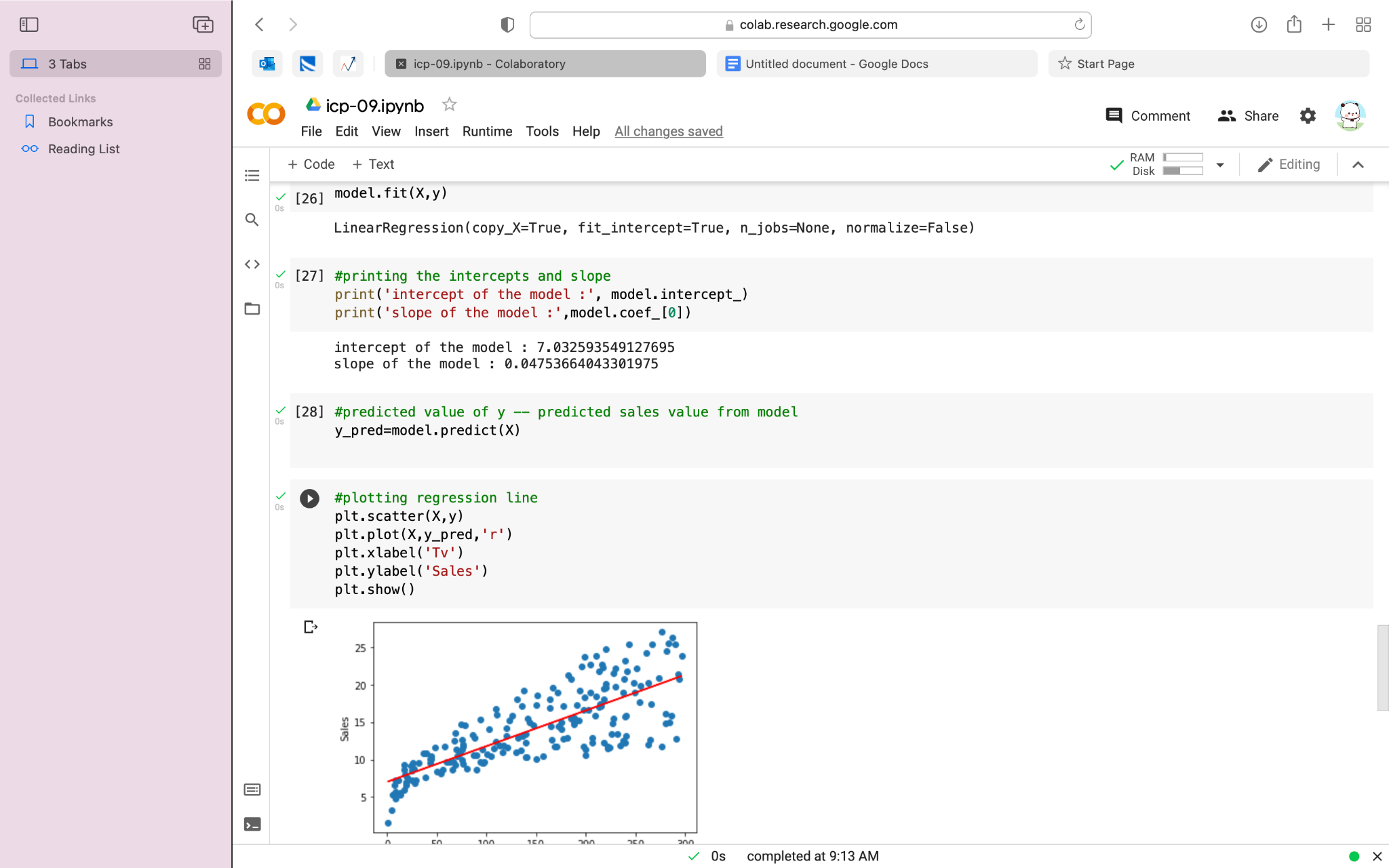


From the above method I have observed that there is a correlation of 0.9 between Tv and sales. So these two columns are highly positively correlated. Radio and sales have correlation of 0.35and Newspaper and sales have correlation of 0.16. It means change in the tv column has a major positive impact on Sales columns.

Then plotted a linear model using TV as a feature and Sales as output. Defined the model using **LinearRegression()** and fitted the model with the data using **fit()**. Then found the y intercept and slope using **model.intercept\_ and model.coef\_[0].**

**So we can write linear model as:**

**sales=0.055546477 x+6.97488229891(y=mx+b)**



Then predicted the value of y using x and plotted the regression line .

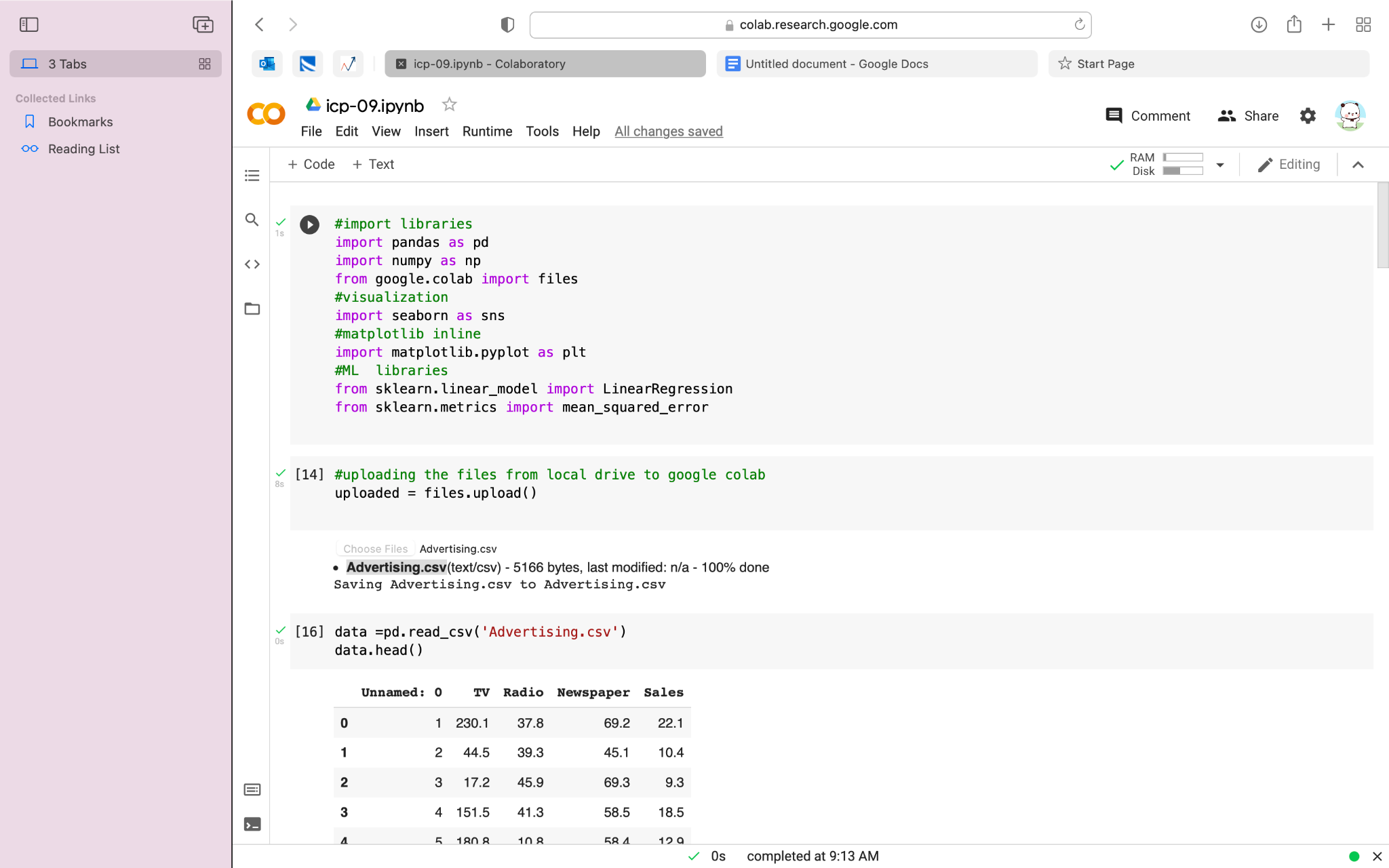
From the plot we can see there is a positive slope and positive intercept .

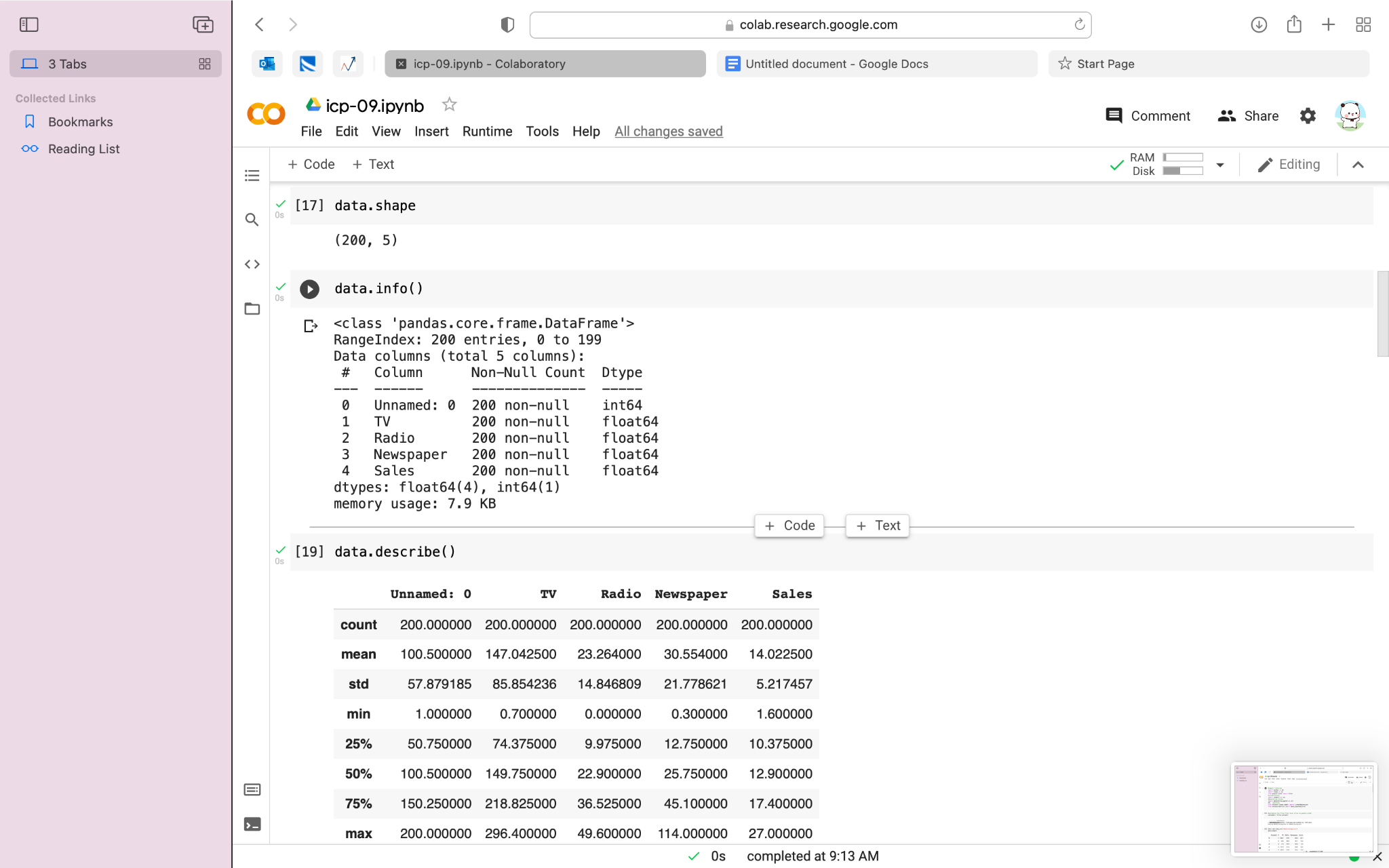
Then Calculated the R square using model.score(x,y) and got 81.22% of the variance in sales is explained by tv.

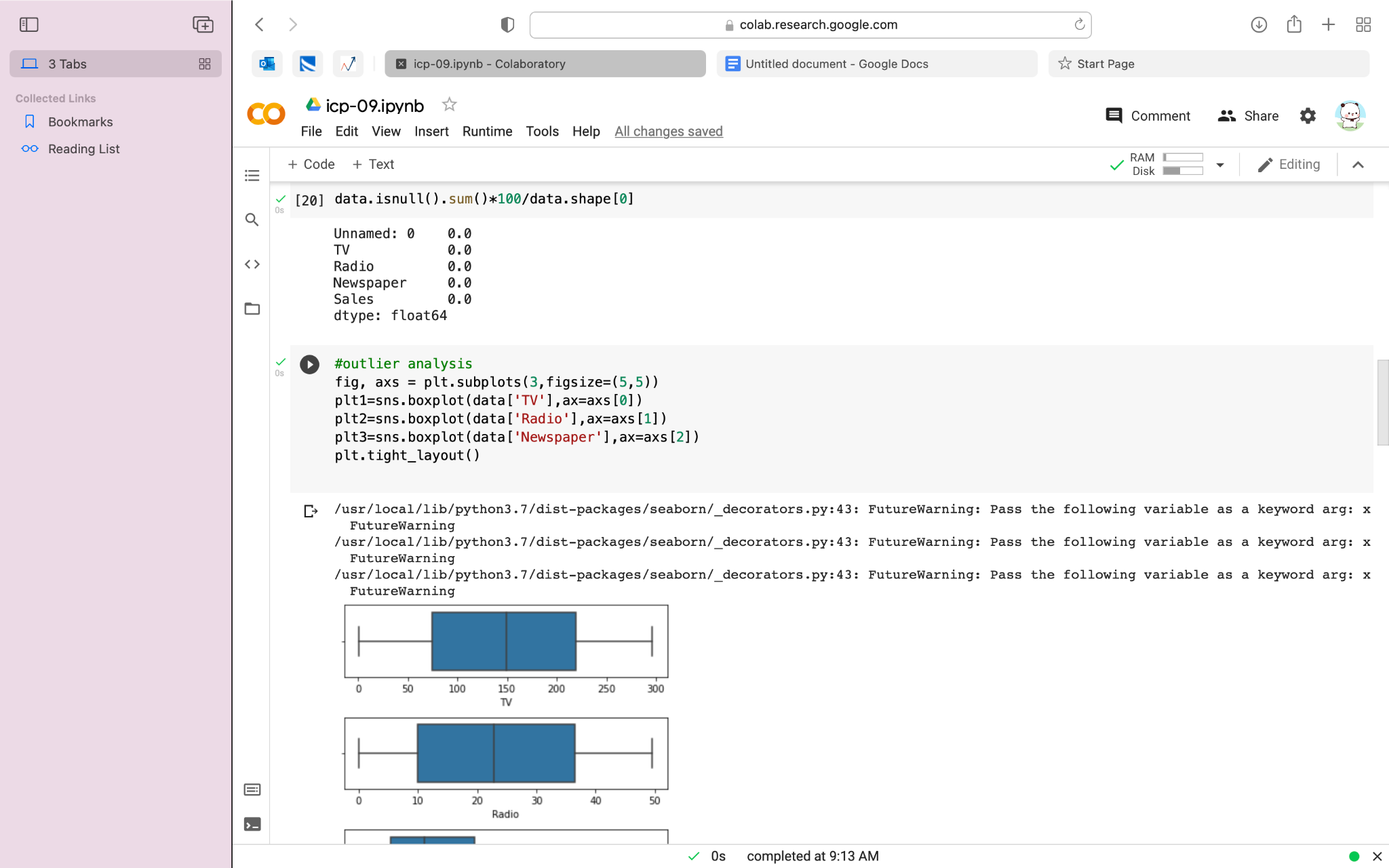
Then Calculated MSE (Mean Square Error ) using the library(MeanSqaureError).

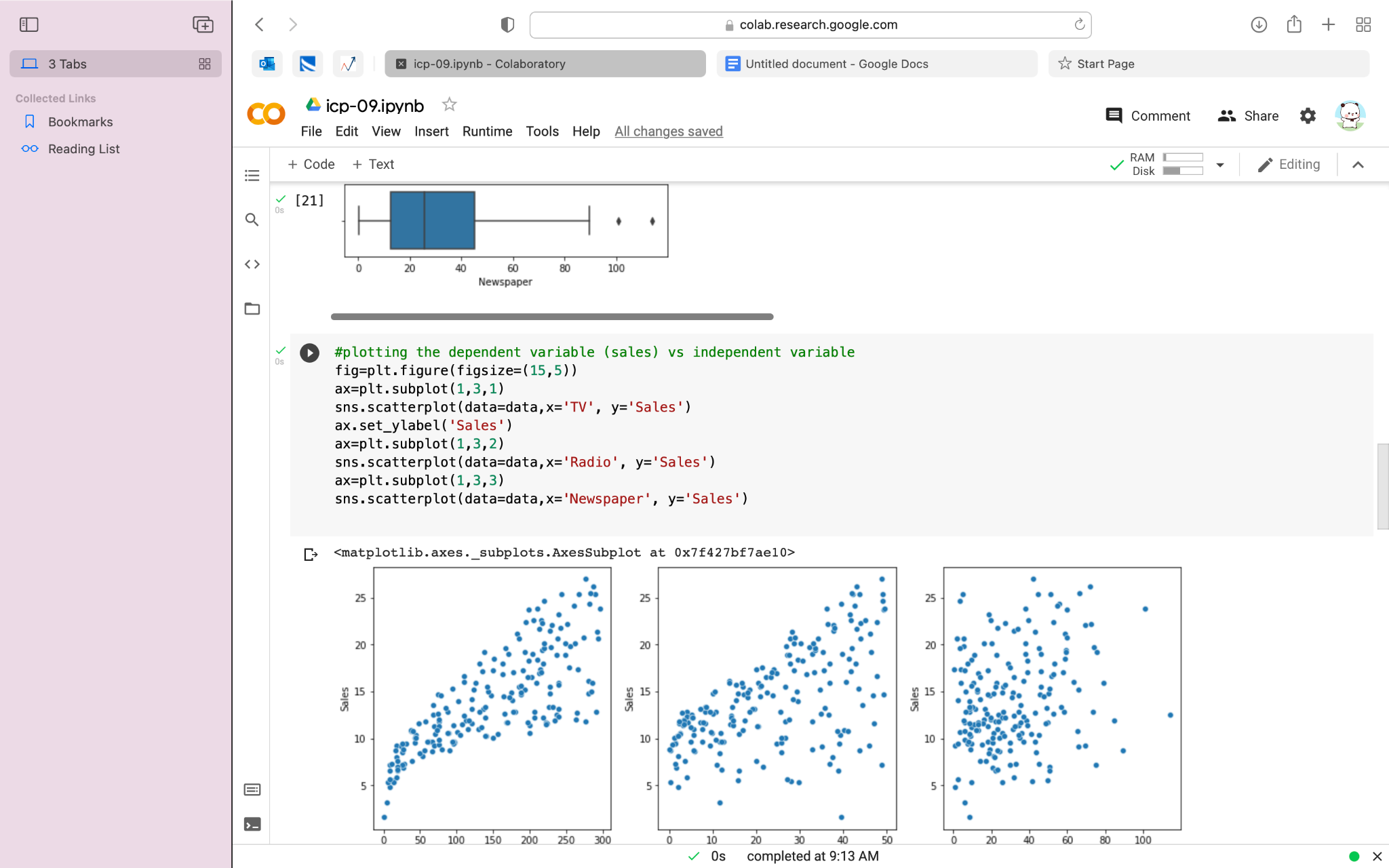
**C)**Challenges that I have faced in selecting the dataset to perform linear regression.

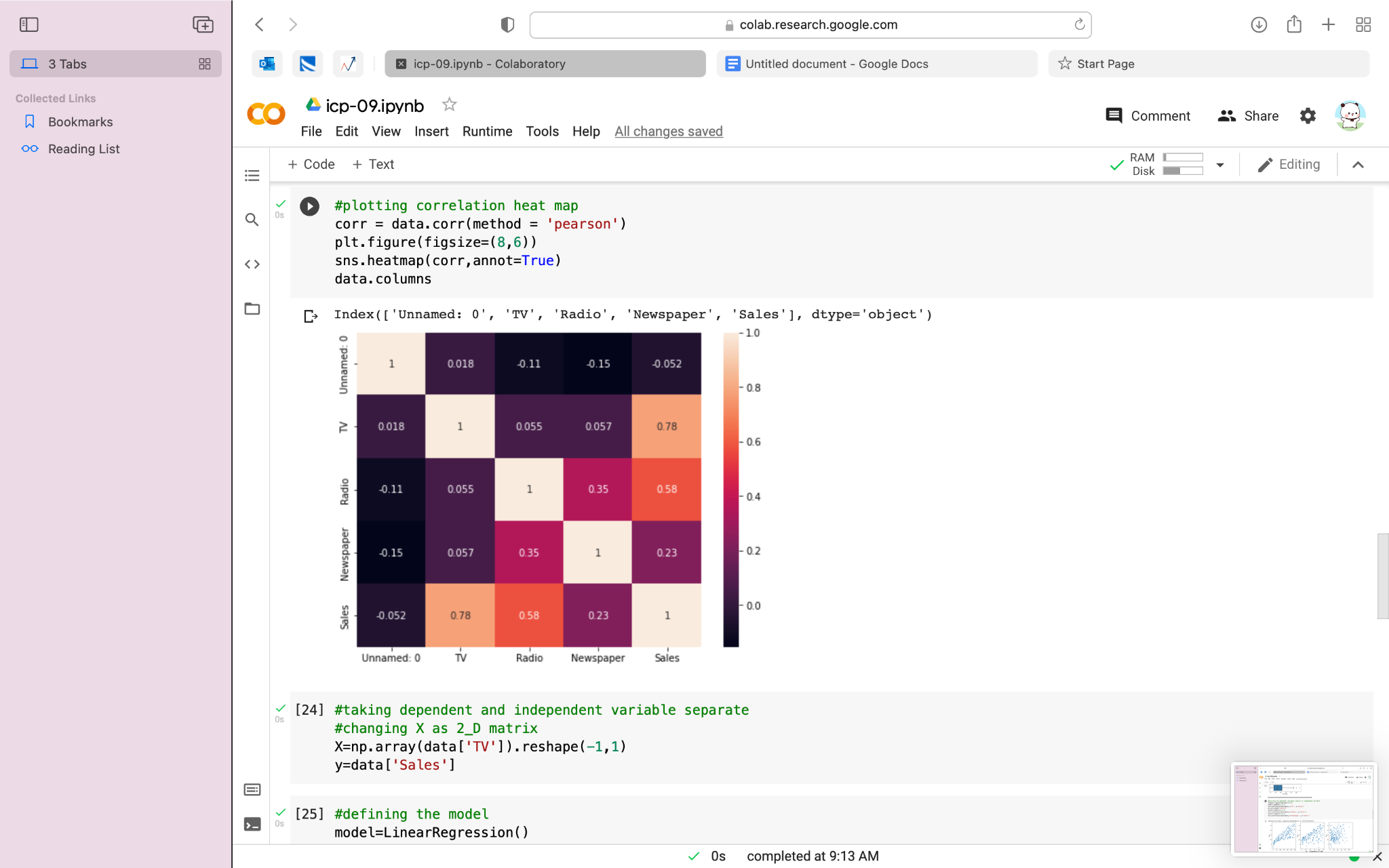
**D)Screenshots:**

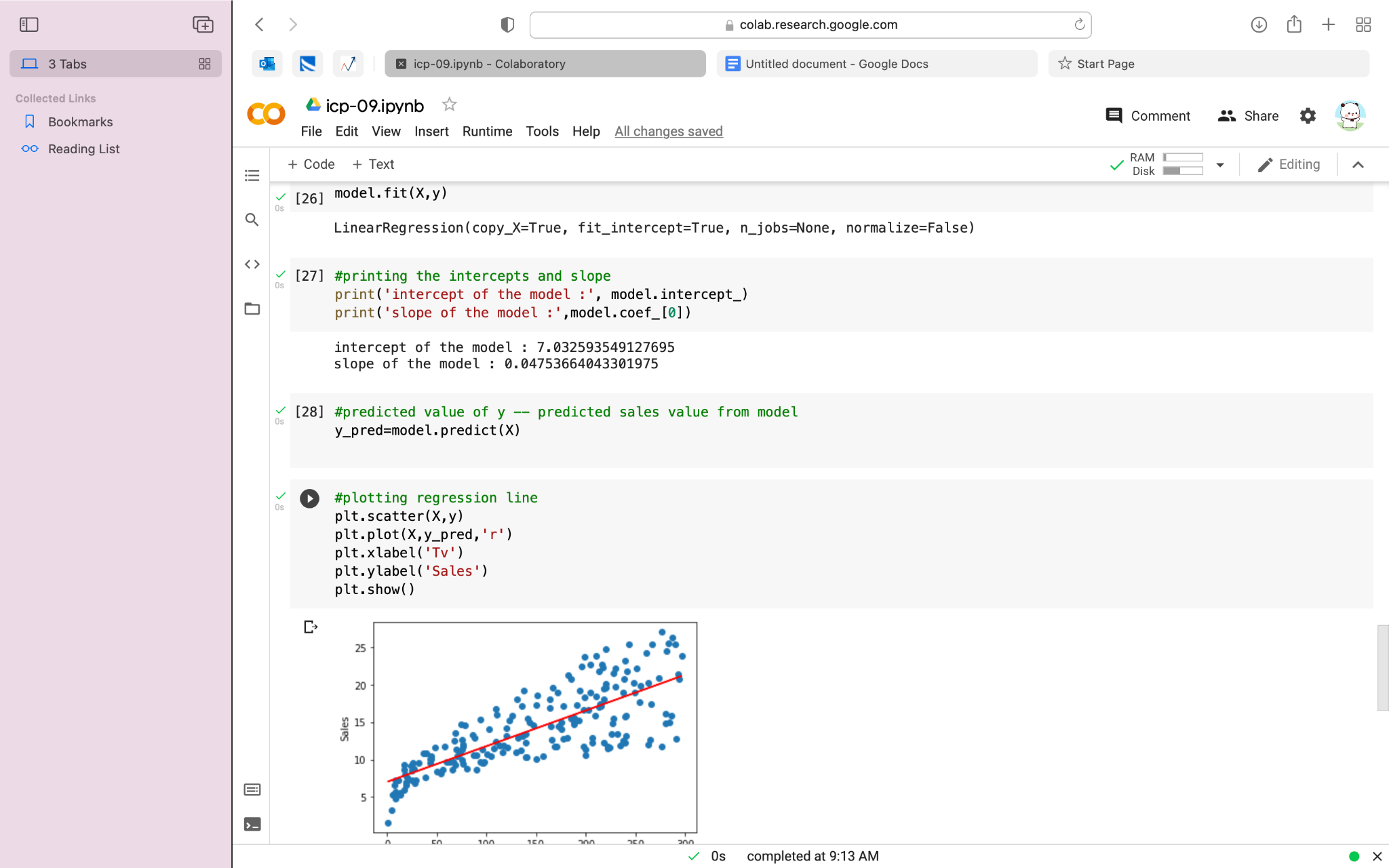


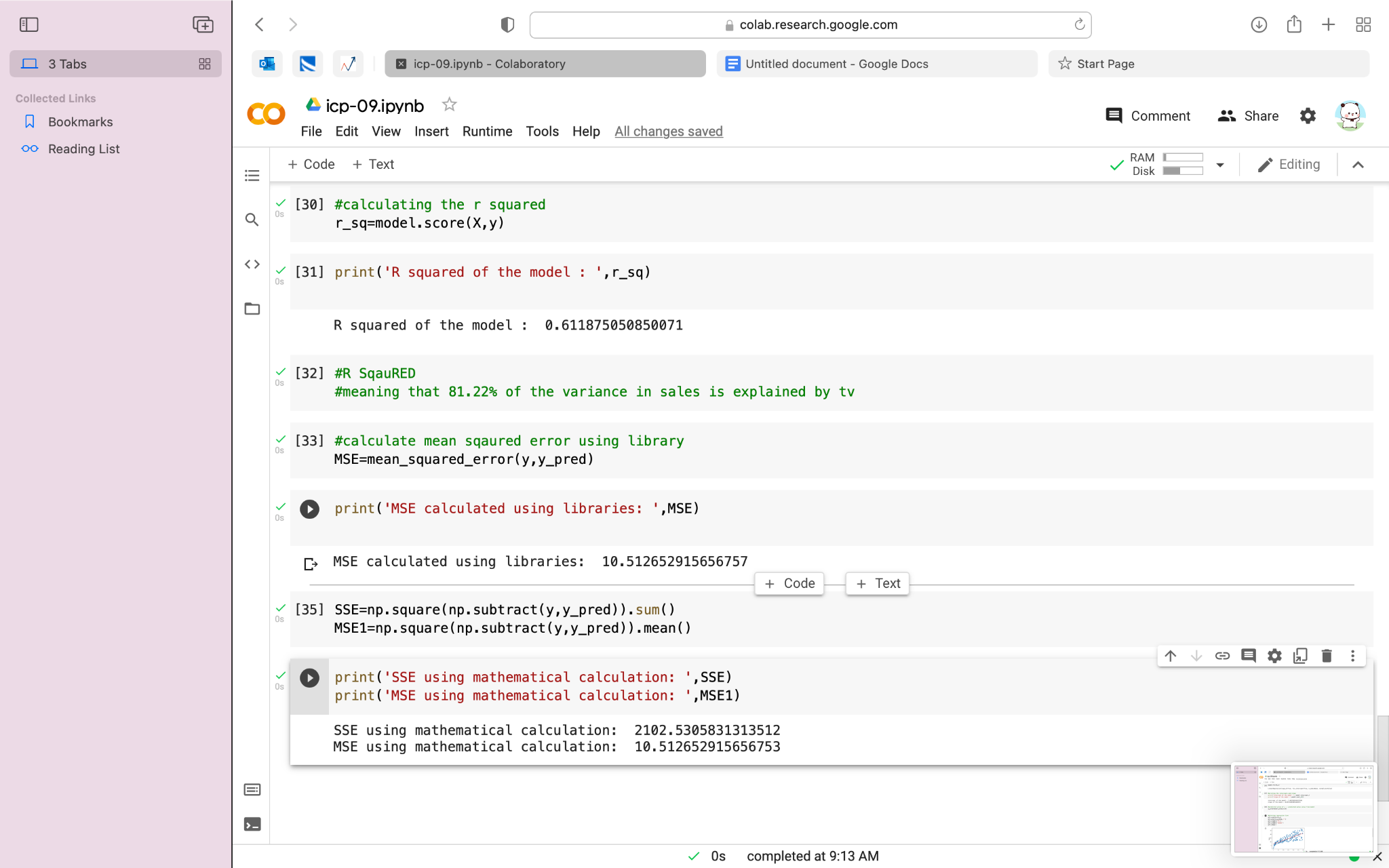












**F)Videolink:**

[**https://github.com/saidurga-kanuganti/icp09/blob/main/Mon%20Nov%2001%202021%2010\_21\_44.webm**](https://github.com/saidurga-kanuganti/icp09/blob/main/Mon%20Nov%2001%202021%2010_21_44.webm)

**G)**The ICP is about performing linear regression on a given dataset and calculating MSE AND SSE ie.. mean square error and sum of square error and finding y intercepts and slope for relationship in data.