For Assignment 1, I’ve attached the following folders-

1. Given Spreadsheets
2. Created Spreadsheets
3. Script
4. Graphs
5. My Details
6. Outputs

* “**Given Spreadsheets”** folder contains the scimago-excel files that we were supposed to download, and operate on this data, one for the journals, one for the conferences.
* For the journals, 80% of the data (title and H-indices) is taken and Impact Factor is calculated using the below mathematical concept-

**Impact Factor=Total Cites(3yrs)/Citable Docs(3yrs)**

***Note-*** *A PDF was attached for ECE students (a Reuters report) containing Impact factors of various journals in 2017, but after converting it into a csv file format, many errors creeped in due to spacing errors, line skip errors, junk values, issues with the delimiter shortening my data set from 601 to 287. So I’ve used the aforementioned mathematical formula for the sake of data- accuracy.*

* Based on impact factors of 80% of data, ‘a’ and ‘b’ were calculated by-

**a=r\*(sigma(y)/sigma(x))**

**b=mean(y)-a\*mean(x)**

where r= correlation coefficient

sigma(y)= standard deviation of Impact factor

sigma(x)= standard deviation of H-Index

* Based on a, b, we get the regression line-

**y=a\*x+ b**

according to my data set, a= 0.03509, b= 0.55064 and correlation coefficient=0.60633 (also printed in the script to the terminal).

* For the rest 20% data, we use the regression line obtained in the above result to find out Impact Factor of this fraction of data, and then we found out the “mean square error” of the new Impact factor values and original impact factor values (for 20% of the data set).

**Impact factor=y=a\*x+ b**

**Mean square error= (x\_new- x\_old)^2/number of values**

**Root Mean square error= (mean square error)^0.5**

according to my data set, Mean square error value= 0.50899, Root Mean square error value= 0.71343 (also printed in the script to the terminal).

* I created a new csv file called **“JOURNAL\_final.csv”** containing the required columns of Title, H-Index and Impact factor for all journals. This includes the 80% calculated data and 20% predicted data. For clarity, I created a csv file called **“EDITED1.csv”** that contains 80% of calculated data, and for the rest 20% data(towards the end), I’ve originally kept all Impact Factor values as -10, and then later I modify this data to obtain predicted values.
* For the conferences, I used the above regression line in the same way to obtain Impact Factor values for all conferences. I wrote all this data to a csv file called **“CONFERENCE\_final.csv”** which contains Title, H- Index and Impact Factor for all conferences.
* I’ve also attached screenshots showing graphically the regression line in the “**Graphs”** folder, containing two graphs- one for journals, one for conferences.
* “**My Details**” contains my name, roll number, email id and other details.
* “**Output**” folder has all values obtained from this assignment including a, b, correlation coefficient and errors.