As discussed, this is an open ended research assignment where you get to use what you have learnt on an actual problem that has not been addressed so far.

This is what you have to do step by step –

1. Go to https://www.scimagojr.com/journalrank.php

2. CSE students should filter by Subject Area = Computer Science, Type = Journal or conference.

3. ECE students filter by Subject Area = Engineering, Category = Electrical and Electronics Engineering, Type = Journal or Conference.

4. Download the data.

5. The data will come in a messy .csv file, you have to write a script to extract the relevant data.

6. At the end of point 5, you will have a spreadsheet with the following categories only:

    For journals – Name of Journal, H-index, Impact Factor

    For conference – Name of Conference, H-index, Impact Factor (empty initially)

7. On the journals you have to find out the correlation coefficient between H-index and Impact Factor.

8. On the journals you have to fit a straight line (regression) between H-index (input) and Impact Factor (output). This you have to do by considering a subset of 80% data (called training data).

9. On the remaining 20% of data (test data), you have to use the learnt regression line to predict the Impact Factor, given the H-index. You have to report what is the error (mean squared error) between the actual Impact Factor and the Predicted Impact Factor.

10. Use the learnt regression line to predict the Impact Factor of the conferences (which was empty initially).

You have to submit the following documents –

1. Spreadsheet for journals and spreadsheet for conferences (including predicted Impact Factor).

2. All script files, such as cleaning data, computing correlation, computing regression coefficients, predicting etc.

3. Error between actual Impact Factor and predicted Impact Factor for journals.