

Spectrum Club,
College of Engineering and Technology,
Bhubaneswar
www.spectrumcet.com

Dear Intern,

Welcome to **Data Science and Machine Learning Task 1**. Hope you are now comfortable with the basic blocks of mathematical functions in python using **Numpy**, and plotting and visualizing your data using **Matplotlib**. This task's aim will be to introduce you and make you used to working with data, cleaning data, various types of plotting and visualization you can do with data, and how numpy and matplotlib libraries help you with the same.

Technology Stack to be used:

We will be mainly be working on importing and manipulating data, as well as cleaning the same. The libraries we will be using for this task include:-

- Numpy
- Matplotlib
- Pandas

Stage 1:


Tasks:





Billy has his exams coming up. He hasn't been performing well on maths throughout the year, but this time he has an idea. He has the data of his friends and how they have fared in the exam. But when he tries to feed the data to his computer, the computer denies it and says it only understands numeric values. Now it's up to you to help him.

1. With the given dataset, named "student-math", import the dataset using the library pandas, and create a dataframe of it.
2. As there are three grades given, create a new column named "final_grade", which shall be the sum of the respective three grades in the row.
3. Delete the three grades from the dataframe after step 2.
4. Replace all binary values with 1 and 0 in the dataframe. E.g. – For values having "Yes", replace them with 1, and for values having "No", replace them with 0. Do this for all columns having binary type of values.
5. Plot a scatter plot between the values of column "studytime" and your new column named "final_grade" that you created in step 2. Observe the relation between both the attributes. You can assign different colours/symbols for the points of a particular time step of "studytime" to observe it better.

- 
6. Plot a boxplot of the new column “final_grade” and the column “studytime” in the same graph, and observe the difference between them.

RESOURCES: -

The resources provided here include both documentation and Youtube tutorials for the above tasks. If any case of any confusion, do Google once and search around or ask in the forum, but do understand the concepts behind the functions properly. Remember, Stackoverflow is your best friend for programming.

➤ Pandas –

(Documentation)

<https://pandas.pydata.org/docs/>

(Video Tutorial)





<https://www.youtube.com/playlist?list=PLeo1K3hjS3uuASpe-1LjfG5f14Bnozjwy>

(video 1-11 should be enough, you can refer for learning more about the library)

The dataset has been provided with this zip file, and the attributes, and the type of data they contain(numeric, non-numeric/binary, nominal), etc. have been provided too.

Good luck!

LAST DATE OF SUBMISSION: 17th May-2020



WARM REGARDS,
SPECTRUM, CET-B

