**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

**Second Semester 2022-2023**

**Programming Assignment-I**

**(EC-2 Regular)**

Course No. : **SE ZG568/SS ZG568**

Course Title : **Applied Machine Learning**

Weightage : **12%**

Duration : **March 20-30, 2023**

Course Instructor **: Dr Bharathi R**

***Assignment Objective: To analyze a given data set. Perform Exploratory Data Analysis. Suggest a regression model.***

|  |
| --- |
| **Dataset: student.csv** |
| The final deliverables of the Programming Assignment-I are   * + 1. a word file documenting all the findings of every stage     2. Python code in ipynb format   Save both files in a folder, zip and upload. |

# Tools and Techniques

Python libraries for data analysis. (**NumPy**,**SciPy**,**Matplotlib**,**Pandas**,**ScikitLearn**,**Statsmodels**,**Seaborn**,**Bokeh**,**Blaze**,**Scrapy**,**Req uests**,**BeautifulSoup**)

# Sample Exploratory Data Analysis Case Studies

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques> <http://ucanalytics.com/blogs/exploratory-data-analysis-retail-case-study-example-part-3/>

**Programming Assignment -I Guidelines**

These are the guidelines and questions that you are expected to answer. The student will have to analyze the data that he/she has been given and come up with meaningful insights for the given dataset. They have to decide a problem statement based on the dataset that they have received after they have performed descriptive statistics and data visualization. The steps that have to be taken are explained below.

1. **Descriptive Statistics**

Data given in the dataset has to be understood and every feature must be explained by the student. The datatypes present in the dataset must be found out. The measures of central tendency should be found and explained. Based on these values, there should be a few critical insights made that would then lead to their problem statement. ***Data cleaning should also be performed by suggesting appropriate techniques to handle missing data and outliers.***

|  |
| --- |
| ***Note:*** Exploratory Data Analysis (EDA) is used to tackle specific tasks such as:   * + 1. Spotting mistakes and missing data;     2. Mapping out the underlying structure of the data;     3. Identifying the most important variables;     4. Listing anomalies and outliers; |

1. **Data Visualization**

Data should be visualized using the various types of charts and graphs that the student has learnt. Utilizing these visualizations, there should be insights from every visualization that is submitted and they should help frame the problem statement that is intended to be solved.

1. **Framing problem statement**

The problem statement should be based on the numbers and visualizations that had been done so far. The problem statement should aim to establish a relation between variables. The problem statement should meet the following criteria

* Can be proven true or false
* Should be detailed and mention both the dependent and independent variable.

This has to be approved and written down

1. **Coming up with correlation analysis**

Correlation Analysis must be done to find out how the variables are related and how the regression model could be made.

1. **Regression Model**

A regression model should be established between the selected variables and should be used to predict the values of the dependent variable.

1. **Evaluating said regression model**

The regression model should be evaluated using the error scores and attempts should be made to increase the accuracy of the model and decrease the error of the model by using various variables.

# Presentation

* Presentation is key. Ensure that your notebook is capable of explaining your insights and visualizations by itself. Section your questions and emphasize your results. **Do not** hide your final result in a sea of code or debugging cells.

# Examples:

* + If your question is on data cleaning, highlight the rows which need to be cleaned and show the results of your data cleaning before and after it has been applied on those rows.
  + If your question asks you to prove a statement using visualizations, ensure that you actually have a concluding statement after your graphs. **Do not** leave the conclusion unstated after visualizing the data in your notebook.
* It is recommended to have short bullet points explaining what you have done before each task, especially for non-visualization tasks. This will help us understand your approach to the problem and can help with partial marks even if you are unable to solve the entire question.
* Prioritise interpretability over design. While it is encouraged to have visually appealing graphs, make sure that you do not lose interpretability of the data in the pursuit of aesthetic visualizations.

# Insights

* The last section of your report will have to be dedicated to an out of the box pursuit. If you think you have a better way of cleaning the dataset or visualizing a question, or if you believe that you have noticed an interesting insight that can be cleaned from the data, add them at the end of your notebook and elaborate why you think you’re right in your report or notebook and make sure you mention it in your recorded video. This carries weightage to your final scores.