Project Objectives

This is an intensive project-based course. It enables students to perform data analysis using Python programming. Students should select their dataset from a free source and conduct a methodical data analysis using Machine Learning algorithms. The project's primary objectives are:

- Generate a data summary using descriptive analysis
- Create a sample and visualize sample data using graphs/charts and remove the unwanted outliers;
- Investigate the correlation between the variables;
- Perform hypothesis testing if you have any assumptions about your dataset;
- Perform data preprocessing prior to building a data model;
- Create and optimize the regression model for the selected dataset in order to predict the values; and
- Develop and Optimize the classification model for the selected dataset in order to predict the values.
- Analyze the data for patterns or groups based on clustering and optimize the model in order to obtain the desired output.
- Consolidate learning through the Anaconda Python for Data Science certification and reflect critically on your data science journey, connecting conceptual knowledge with applied project work.

Project Description

You are assigned to work on the data analysis for chosen dataset. The list of datasets is available in a Kaggle data source(https://www.kaggle.com/datasets). The project carries 40% of your coursework marks. You are required to work in a team of maximum TWO (2) members. It is important that you need to collaborate in working on the project within your team. The collaboration between the team members will be recorded, tracked, and monitored.

For CLO1, CLO2, and CLO3 – Regression, same dataset should be used

If required, then CLO3 – classification and Cluster, different dataset can be used.

New Requirement (Individual Component - 10 Marks): Anaconda Python for Data Science Professional Certificate (via Coursera) Each student is expected to complete the Anaconda Python for Data Science Professional Certificate course on LinkedIn Learning and submit a screenshot of the completion certificate. The program includes multiple courses (e.g., Intro to Data Science, Statistics Foundations, Learning Python, Python Data Analysis) and students must complete and submit all.
Students are required to:

- Submit individual course completion certificates for each course in the certification path sequentially.

- Submit the final Professional Certificate
- Submit individual reflection summarizing your learning experience, key takeaways, and how the certification supported your understanding of the project work (this reflection is mandatory)
- Answer a viva question drawn from any of the covered course content during the individual oral defense.
- This component carries 10 marks under the individual evaluation rubric.

Project Tasks/Questions

CLO	Deliverable Learning Outcomes	Marks
1	Define the purpose of data analysis for the chosen dataset	2
	Identify and Justify the type of programming used for data analysis	2
	Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset	3
	Identify and Justify the independent and dependent variables for the chosen dataset.	3
	Will you do the sampling? Identify and justify the type of sampling to be used for the chosen dataset	
	Total	**10**
2	1. Justify why you want to perform the descriptive analysis for the chosen dataset.	1
	2. Create a script to develop a Python function for descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics.	1
	3. Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created].	1
	4. Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created].	1
	5. Create a detailed descriptive statistics report about the dependent variable of the chosen dataset.	1
	6. Visualize the dependent variable by the Graph/Chart of the following using Python Program: a. Scatter plot b. Box Plot c. Histogram d. Heat Map Hint: Use Matplot or Ski-learn library	3
	7. Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable. Note: If you have more than one independent variable, then choose any one of the independent variables.	1
	8. Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test.	1
	Total	**10**
3	9. Build, Train, Develop and Evaluate using Simple Regression for chosen dataset.	5

	10. Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression	5
	11. Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree.	17
	12. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.	9
	13. Predict the dependent variable by using best-fit classifier.	1
	14. Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset.	8
	15. Explain the strategy for improving the system after viewing the cluster diagram.	2
	Total	**42**
4	16. Create a new repo for project in Git Hub	3
	17. Upload all the project files created for CLO1,CLO2 and CLO3 to the Git Hub repo	4
	18. Configure Git with GitHub	5
	19. Clone Git hub repo to Git	4
	20. Pull any file from Git Hub repo to Git	5
	21. Modify the pulled file and push the modified file to Git Hub	5
	Total	**26**

Please link each question/task to its corresponding CLO's and assign marks according to the CAP. Please note that a task might address many CLOs.

Project Deliverables

Project Report (50%)

1. **Deliverable 1**: A complete report about the purpose of data analysis, programming language chosen for data analysis, types of machine language algorithm to be analysed and the list of variables chosen for analysis [CLO1]
2. **Deliverable 2**: A detailed report about the summary of the data, sampling, graphs/charts to analyze the data, relationship between variables, evaluating assumptions using hypothesis testing, predicting the variables using the regression model [CLO 2,3]
3. **Deliverable 3**: A comprehensive description about the data model created using classification and clustering algorithm of machine learning. It should involve the narrative about the data model is optimize to predict the variables and bow the best fit model has been chosen. [CLO 2,3]
4. **Deliverable 4**: Complete narration about data versioning using Git. [CLO 4]
5. **Written Communication:** Complete report with specified format and structure [12 points].

LinkedIn Certificate Report

This section is dedicated to the submission of individual certificates for each course completed under the 'Anaconda Python for Data Science Professional Certificate' path on LinkedIn Learning. Students must ensure that all certificates are clearly labeled and arranged in the order of course completion.

Please include the following in this section:

1. Screenshot of course completion: 'Introduction to Data Science'
2. Screenshot of course completion: 'Statistics Foundations 1: The Basics'
3. Screenshot of course completion: 'Statistics Foundations 3: Using Data Sets'
4. Screenshot of course completion: 'Learning Python (2021)'
5. Screenshot of course completion: 'Python Data Analysis (2020)'
6. Final 'Anaconda Python for Data Science Professional Certificate' from LinkedIn Learning
7. Individual reflections on the learning experience (optional)

Project Oral (50%)

6. **Oral Communication**: Each student will be assessed in the form of individual oral defense with PowerPoint presentation. [All CLOs] [10 Marks]
7. **Follow-up Questions and Discussion** [All CLOs] [20 Marks]
8. **[Collaboration]** [10 Marks].
9. **Anaconda Python for Data Science Professional Certificate (via Linkedin Learning** submitted individually**), individual reflection & Viva Question** from certification [10 Marks].

Note: For oral presentations, a slide should be dedicated for each student to present their collaboration and lesson learnt. This will allow each student to showcase their individual contributions and reflect on the overall group experience. It also provides a structured format for sharing key takeaways and insights gained from working together.

Points 1 to 4 are a part of group grading [50%] and points 5,6, 7 and 9 contribute to individual grading.