Course Name: Fundamentals of Programming & Data Science	**Course Code:** CMPE-112L
Assignment Type: Lab	**Dated:** 22nd April 2024
Semester: 2nd	**Session:** 2023
Lab/Project/Assignment #: 11	**CLOs to be covered: CLO 4**
Lab Title: Introduction to Data Science	**Teacher Name:** Engr. Afeef Obaid

Lab Evaluation:

CLO 4	Display well-commented code					
Levels (Marks)	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
(10)						
					Total	**/10**

Rubrics for Current Lab Evaluation

Scale	Marks	Level	Rubric
Excellent	**9-10**	L1	Submitted all lab tasks, BONUS task, have good understanding.
Very Good	**7-8**	L2	Submitted the lab tasks but have good understanding
Good	**5-6**	L3	Submitted the lab tasks but have weak understanding.
Basic	**3-4**	L4	Submitted the lab tasks but have no understanding.
Barely Acceptable	**1-2**	L5	Submitted only one lab task.
Not Acceptable	**0**	L6	Did not attempt

LAB No. 11

Lab Goals/Objectives:

By reading this manual, students will be able:

- To Become familiar with the Data Sciences
- To learn about Data Collection
- To learn about Data Cleaning and Preprocessing in Python
- To learn about Exploratory Data Analysis (EDA) in Python

Equipment Required: Computer system with Pycharm IDE and python package installed on it

Introduction to Data Science

Data science is an interdisciplinary field that combines scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It involves various processes, techniques, and tools to analyze, interpret, and derive actionable information from data.

Components of Data Science

Data Collection: Data science starts with gathering relevant data from various sources, such as databases, APIs, websites, sensors, or any other data repositories. This data can be in structured format (organized in tables with fixed fields) or unstructured format (text, images, audio, video, etc.).

Data Cleaning and Preprocessing: Raw data often contains errors, missing values, outliers, or inconsistencies. Data cleaning involves removing or correcting these issues to ensure the quality and reliability of the data. Preprocessing includes transforming and standardizing the data, dealing with missing values, encoding categorical variables, and normalizing numeric features.

Exploratory Data Analysis (EDA): EDA is the process of examining and visualizing the data to gain insights and understand its characteristics. It involves summary statistics, data visualization, and correlation analysis to identify patterns, trends, outliers, and relationships within the data. EDA helps in formulating hypotheses and selecting appropriate techniques for further analysis.

Feature Engineering: Feature engineering involves creating new features or transforming existing features to enhance the predictive power of the data. It includes techniques like feature extraction, feature selection, dimensionality reduction, and creating derived features using domain knowledge.

Modeling and Machine Learning: This stage involves applying statistical models, algorithms, and machine learning techniques to the prepared data for prediction, classification, clustering, or other tasks. It includes selecting the appropriate model, training it on the data, and evaluating its performance using various metrics. Machine learning techniques commonly used in data science include linear regression, logistic regression, decision trees, random forests, support vector machines, neural networks, and more.

Model Evaluation and Validation: Once the model is trained, it needs to be evaluated to assess its performance and generalization capabilities. This involves splitting the data into training and testing sets, cross-validation, and using evaluation metrics like accuracy, precision, recall, F1-score, or area under the curve (AUC). Model validation ensures that the model is reliable and performs well on unseen data.

Deployment and Productionization: Once the model is deemed satisfactory, it can be deployed into a production environment where it can generate predictions or insights on new, real-time data. This stage involves integrating the model into a larger system, setting up appropriate APIs or interfaces for data input/output, and ensuring scalability, robustness, and security.

Monitoring and Maintenance: After deployment, the performance of the model needs to be continuously monitored to detect any degradation or drift in its accuracy. Regular updates, retraining, and fine-tuning may be required to maintain the model's effectiveness as the underlying data changes over time.

Communication and Visualization: Data scientists play a crucial role in communicating the findings and insights derived from the data to stakeholders, such as business executives, policymakers, or researchers. This often involves creating visualizations, reports, or interactive dashboards to effectively convey complex information in a clear and understandable manner.

Ethical Considerations: Data science also includes ethical considerations regarding privacy, data security, fairness, and bias. Data scientists need to ensure that the data they use is obtained and processed in an ethical manner, and that the resulting models and insights do not perpetuate discrimination or harm any particular group.

Data science is a constantly evolving field that requires a combination of domain knowledge, programming skills (e.g., Python), statistical knowledge, and an understanding of machine learning algorithms. It finds applications in various domains like finance, healthcare, marketing,

In this Lab our main focus is Data collection, Data Cleaning, Preprocessing and Exploratory data Analysis (EDA)

Libraries Used

There are a lot of libraries for data science in Python but the details of libraries used in Data Cleaning, Preprocessing and for Exploratory data Analysis (EDA) are given below.

Pandas: Pandas is a widely used library for data manipulation and analysis. It provides data structures like DataFrames, which allow for easy handling and manipulation of tabular data. Pandas offers functions to load, clean, filter, aggregate, and transform data, making it a fundamental library for EDA.

NumPy: NumPy is a fundamental library for scientific computing in Python. It provides powerful tools for working with multi-dimensional arrays, along with a collection of mathematical functions. NumPy is often used in EDA for numerical operations, array manipulation, and handling missing values.

Matplotlib: Matplotlib is a versatile library for creating static, animated, and interactive visualizations in Python. It offers a wide range of plotting functions and customization options, allowing for the creation of various types of charts, histograms, scatter plots, box plots, and more. Matplotlib is often used to visualize distributions, correlations, and patterns in the data.

Seaborn: Seaborn is a high-level statistical visualization library that works in conjunction with Matplotlib. It provides a simplified interface for creating attractive and informative statistical graphics. Seaborn offers functions for visualizing distributions, categorical data, regression models, and more. It simplifies the process of creating complex plots with fewer lines of code.

Lab Tasks

- Take any dataset from Kaggle and Perform Data Cleaning, Data Preprocessing, Exploratory Data Analysis (EDA) and Feature Engineering on it.