# MOOC Seminar Report

*on*

*Data Analysis with Python*

*****(CSE 5th Semester MOOC Seminar ) 2023-2024*

Submitted to: Submitted by:

Mr. Samir Rana Mr. Satyam Parashar

(CC-CSE-F-5th -Sem) Roll. No.: 2119138

Guided by: CSE-F-5th -Sem

Mr. Samir Rana Session: 2023-2024

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA HILL UNIVERSITY, DEHRADUN**

# CERTIFICATE

*(from Internal Co-ordinator of MOOC i.e. Class Coordinator)*

#### Certified that Mr. Satyam Parashar (Roll No.- 2119138) has Completed the MOOC Seminar on the topic “**Data Analysis With Python”** from Coursera under the guidance of Mr. Samir Rana for fulfillment of CSE 5th Semester MOOC Seminar in Graphic Era Hill University, Dehradun. Students have Completed this Course to the best of my knowledge.

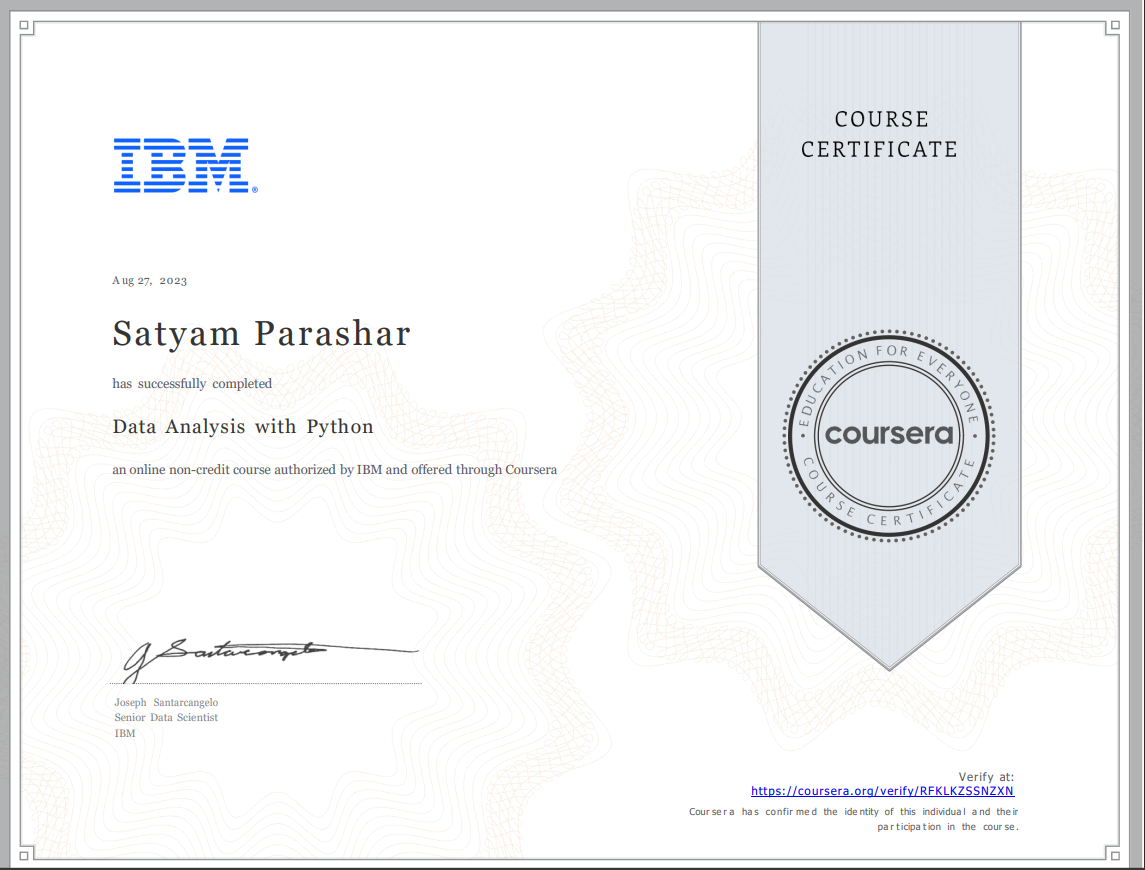
Date:

(Mr. Samir Rana) **Class Coordinator CC-CSE-F-5th - Sem**

CSE Department

GEHU, Dehradun

# CERTIFICATE

****

Refer: https://coursera.org/verify/RFKLKZSSNZXN

**ACKNOWLEDGMENT**

I want to thank particularly my Supervisor Mr. Samir Rana for his patience, support, and encouragement throughout the completion of this Course.

Last but not least I am greatly indebted to all other persons who directly or indirectly helped me during this course.

**Mr. Satyam Parashar**

**Roll No.- 2119138 CSE-F-5th -Sem**

**Session: 2023-2024 GEHU, Dehradun**

## TABLE OF CONTENTS

**PAGE NO**

1. **Importing Datasets (Module-1 of Course): 5**

* Introduction to Data Analysis with Python.
* Understanding various file formats (CSV, Excel, etc.).
* Using Pandas library for data manipulation.
* Reading and importing datasets into Pandas DataFrame.
* Exploring basic DataFrame operations and attributes.

1. **Data Wrangling (Module-2 of Course): 6**

* Data cleaning and preprocessing.
* Handling missing values and outliers.
* Data transformation techniques.
* Merging and joining datasets.
* Introduction to groupby and aggregation in Pandas.

1. **Extemporary Data Analysis (Module-3 of Course): 7**

* Exploratory Data Analysis (EDA) techniques.
* Basic statistical analysis using Pandas and NumPy.
* Data visualization using Matplotlib and Seaborn.
* Insight generation through data exploration.
* Understanding the distribution of data.

1. **Model Development (Module-4 of Course): 8**

* Introduction to machine learning.
* Preparing data for machine learning models.
* Splitting data into training and testing sets.
* Building and training machine learning models (e.g., regression, classification).
* Understanding the concept of features and labels.

1. **Model Evaluation (Module-5 of Course): 9**

* Evaluating model performance using various metrics.
* Cross-validation techniques.
* Fine-tuning and optimizing machine learning models.
* Hyperparameter tuning.
* Understanding overfitting and underfitting.

1. **Final Assignments: 10**

* Applying the knowledge gained in previous modules to real-world scenarios.
* Solving practical problems using data analysis and machine learning.
* Developing and evaluating models on specific datasets.
* Presenting findings and insights from the analysis.
* Completing comprehensive assignments that integrate all learned concepts.

**REFERENCE 11**

**Importing Datasets (Module-1 of Course):**

Introduction to Data Analysis with Python:

* Overview of the importance of data analysis.
* Introduction to Python as a powerful tool for data analysis.
* Understanding the role of Python libraries in data analysis.
* Understanding Various File Formats (CSV, Excel, etc.):
* Explanation of common data file formats like CSV, Excel, JSON, etc.
* How to identify and work with different file formats.
* Importance of choosing the right format for specific tasks.
* Using Pandas Library for Data Manipulation:

Introduction to the Pandas library.

* Overview of Pandas data structures: Series and DataFrame.
* Basic operations for data manipulation using Pandas.
* Reading and Importing Datasets into Pandas DataFrame:
* Hands-on experience on reading data from various sources into Pandas.
* Using Pandas functions to load data into DataFrames.
* Handling different parameters during the data importing process.

Exploring Basic DataFrame Operations and Attributes:

* Understanding basic DataFrame operations like indexing, slicing, and filtering.
* Overview of DataFrame attributes and methods.
* Practical exercises to reinforce DataFrame manipulation skills.

**Data Wrangling (Module-2 of Course):**

Data Cleaning and Preprocessing:

* Identifying and handling missing data.
* Dealing with duplicate values.
* Techniques for cleaning and preparing data for analysis.

Handling Missing Values and Outliers:

* Strategies for detecting and handling missing values.
* Methods to identify and manage outliers in the data.
* Importance of data integrity in the analysis process.

Data Transformation Techniques:

* Techniques for transforming data, including normalization and scaling.
* Feature engineering to create meaningful variables.
* Preparing data for machine learning algorithms.

Merging and Joining Datasets:

* Understanding merging and joining concepts.
* Techniques for combining data from different sources.
* Practical examples of merging datasets.

Introduction to Groupby and Aggregation in Pandas:

* Exploring the groupby operation in Pandas.
* Aggregating data based on specific criteria.
* Understanding and applying aggregation functions.

**Extemporary Data Analysis (Module-3 of Course):**

Exploratory Data Analysis (EDA) Techniques:

* Importance of EDA in understanding data.
* Univariate and bivariate analysis.
* Visualizing relationships and patterns in the data.

Basic Statistical Analysis Using Pandas and NumPy:

* Descriptive statistics using Pandas and NumPy.
* Understanding measures of central tendency and dispersion.
* Utilizing statistical functions for data analysis.

Data Visualization Using Matplotlib and Seaborn:

* Introduction to data visualization libraries (Matplotlib and Seaborn).
* Creating various types of plots for effective data communication.
* Enhancing visualizations for better interpretation.

Insight Generation Through Data Exploration:

* Extracting meaningful insights from data.
* Formulating hypotheses based on data exploration.
* Iterative exploration to refine analysis.

Understanding the Distribution of Data:

* Exploring the distribution of variables.
* Identifying and interpreting skewness and kurtosis.
* Visualizing data distributions using histograms and other plots.

**Model Development (Module-4 of Course):**

Introduction to Machine Learning:

* Overview of machine learning concepts.
* Understanding supervised and unsupervised learning.
* Real-world applications of machine learning.

Preparing Data for Machine Learning Models:

* Feature selection and extraction.
* Addressing imbalanced datasets.
* Preprocessing data for specific machine learning algorithms.

Splitting Data into Training and Testing Sets:

* Importance of training and testing data.
* Techniques for splitting data to avoid overfitting.
* Ensuring model generalization to unseen data.

Building and Training Machine Learning Models:

* Implementing regression and classification models.
* Choosing appropriate algorithms for specific tasks.
* Fine-tuning models for better performance.

Understanding the Concept of Features and Labels:

* Defining features and labels in a dataset.
* Importance of selecting relevant features.
* The role of labels in supervised learning.

**Model Evaluation (Module-5 of Course):**

Evaluating Model Performance Using Various Metrics:

* Metrics for regression (e.g., Mean Squared Error) and classification (e.g., accuracy, precision, recall).
* Interpreting model evaluation metrics.
* Balancing competing metrics based on the problem context.

Cross-validation Techniques:

* Understanding the concept of cross-validation.
* Implementing k-fold cross-validation for model assessment.
* Addressing issues related to overfitting and underfitting.

Fine-tuning and Optimizing Machine Learning Models:

* Techniques for hyperparameter tuning.
* Strategies for optimizing model parameters.
* Balancing model complexity and performance.

Hyperparameter Tuning:

* Identifying and tuning hyperparameters for better model performance.
* Utilizing grid search and randomized search techniques.
* Practical considerations in hyperparameter tuning.

Understanding Overfitting and Underfitting:

* Recognizing overfitting and underfitting in machine learning models.
* Techniques for mitigating overfitting.
* Balancing model complexity and generalization.

**Final Assignments:**

Applying the Knowledge Gained in Previous Modules to Real-world Scenarios:

* Solving practical problems using data analysis and machine learning techniques.
* Applying the learned concepts to diverse datasets.

Solving Practical Problems Using Data Analysis and Machine Learning:

* Developing solutions to real-world problems.
* Integrating data analysis and machine learning to address specific challenges.

Developing and Evaluating Models on Specific Datasets:

* Building machine learning models on provided datasets.
* Evaluating model performance and making recommendations.

Presenting Findings and Insights from the Analysis:

* Communicating results effectively through visualizations and summaries.
* Presenting insights and recommendations based on data analysis.

Completing Comprehensive Assignments that Integrate All Learned Concepts:

Integrating knowledge from importing datasets to model development and evaluation.

Showcasing proficiency in data analysis and machine learning.

**References**

**For Python and Pandas:**

* Official Python Documentation
* Pandas Documentation

**For Data Visualization:**

* Matplotlib Documentation
* Seaborn Documentation

**For NumPy:**

* NumPy Documentation

**For Machine Learning:**

* Scikit-learn Documentation
* Towards Data Science - Machine Learning

**For Exploratory Data Analysis (EDA):**

* DataCamp - Exploratory Data Analysis in Python
* Towards Data Science - Exploratory Data Analysis

**For Data Wrangling:**

* Data Wrangling with Pandas Cheat Sheet
* Real Python - Pythonic Data Cleaning With Pandas and NumPy

**For Model Evaluation and Hyperparameter Tuning:**

* Scikit-learn Model Evaluation
* Hyperparameter Tuning in Machine Learning

**For Assignments and Real-world Applications:**

Kaggle Datasets

UCI Machine Learning Repository

<https://docs.python.org/>

and the Pandas documentation at

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

.Similarly, you can find Matplotlib documentation at

<https://matplotlib.org/stable/contents>.html and Seaborn documentation at <https://seaborn.pydata.org/>

The same approach can be applied to other libraries and topics mentioned in the responses.