A

MOOC based Seminar Report

On

Data Science with Python

lms.simplilearn.com

Submitted in partial fulfillment of the requirement Seminar for the I Semester

MCA

By

Lalit Singh

2351027

Under the Guidance of

Dr. Naveen Tiwari

Assistant professor

School of Computing

School of Computing

GRAPHIC ERA HILL UNIVERSITY BHIMTAL CAMPUS

SATTAL ROAD, P.O. BHOWALI

DISTRICT- NAINITAL-263136

2023 - 2024

Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

BHIMTAL CAMPUS

THIS IS TO CERTIFY THAT Mr. LALIT SNGH HAS SATISFACTORILY PRESENTED MOOC BASED SEMINAR. THE COURSE OF THE MOOC REGISTRATION "**Data Science with Python**"

IN PARTIAL FULLFILLMENT OF THE SEMINAR PRESENTATION REQUIREMENT IN I SEMESTER OF MCA DEGREE COURSE PRESCRIBED BY GRAPHIC ERA HILL UNIVERSITY, DEHRADUN BHIMTAL CAMPUS DURING THE YEAR 2023-2024.

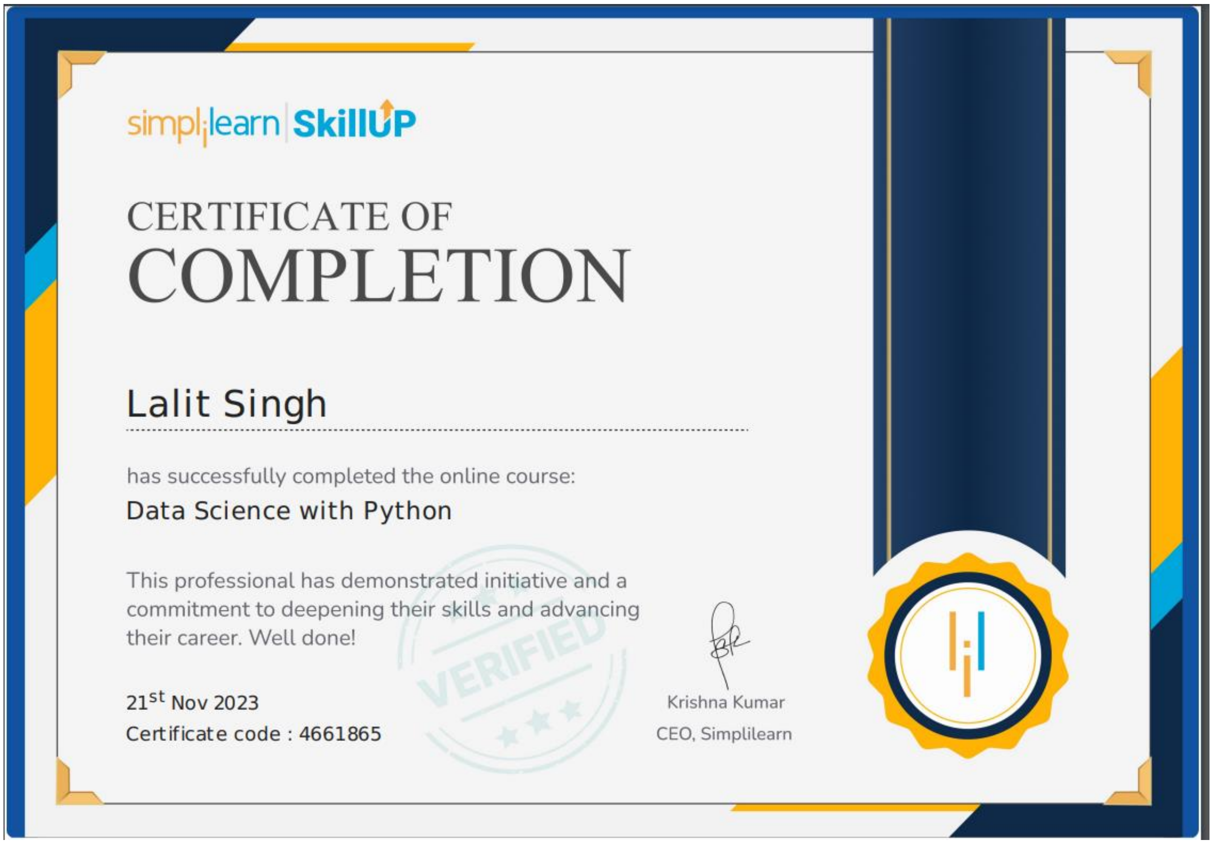
Class Coordinator	HOD
Dr. Naveen Tiwari	Dr. S.K. Budhani
Signature	Signature

Graphic Era HILL UNIVERSITY

Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

BHIMTAL CAMPUS

Copy of confirmation-Email of Course Completion Received



simplilearn | SkillUP

CERTIFICATE OF
COMPLETION

Lalit Singh

has successfully completed the online course:

Data Science with Python

This professional has demonstrated initiative and a commitment to deepening their skills and advancing their career. Well done!

VERIFIED

21st Nov 2023

Certificate code : 4661865

Krishna Kumar

CEO, Simplilearn

Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

BHIMTAL CAMPUS

Modules Attended

S. NO.	DAToE	Details of Modules Attended	PAGE NO.	Signature
1.	05/10/2023	Introduction | Demo Jupyter Lab Walk - Through	1	
2.	11/10/2023	Introduction to python	2	
3.	17/10/2023	Python libraries for data science	4	
4.	21/10/2023	Data Visualization libraries	6	
5.	27/10/2023	Statistics	8	
6.	2/11/2023	Data Wrangling	10	
7.	8/11/2023	Feature Engineering	13	
8.	14/11/2023	Exploratory Data Analysis	16	
9.	20/11/2023	Feature Selection	18	

ACKNOWLEDGEMENT

I Express my gratitude to Simplilearn and professors, for coming up with this interesting and thought provoking course, compiling the best contents and most important working towards changing the words, making the difference an d inspiring us to become the climate saver ourselves , knowledge that I gained will last a lifetime.

I Want to thanks lms.simplilearn.com for broadcasting the content and the very nice user interface.

Many thanks to my institution, Graphic Era Hill University, campus MOOC Coordinator – Dr. Naveen Tiwari for motivating us to look beyond the syllabus, to take up the opportunity and to learn as much as we can, as long as we can.

MODULE- 1

(Introduction | Demo Jupyter Lab Walk - Through)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 05/10/2023

WEEK OF EXECUTION: 1st Week

OBJECTIVE OF LEARNING:

- Provide a comprehensive introduction to the field of data science, emphasizing the role of Python as a primary programming language.
- Explore the significance of effective file structure design in the context of data science projects.
- Examine fundamental data structures essential for efficient data manipulation and analysis.
- Offer a practical, hands-on demonstration of working with Jupyter Lab, a popular interactive computing environment for data science.

CONTENT OF MODULE:

a. **Introduction to Data Science**
   a. Definition and key principles of data science.
   b. Overview of the data science lifecycle.
   c. Understanding the importance of data in decision-making.

b. **Jupyter Lab Walk-Through**
   a. Introduction to Jupyter Lab as an interactive computing environment.
   b. Hands-on demonstration of Jupyter Lab features for data analysis, visualization, and collaboration.

LEARNING OUTCOME:

- The understanding of the foundational concepts of data science.
- Recognize Python's pivotal role in data manipulation and analysis.
- Feel comfortable navigating and utilizing Jupyter Lab for interactive data exploration and analysis.

MODULE- 2

(Introduction to python)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 11/10/2023

WEEK OF EXECUTION: 2nd Week

OBJECTIVE OF LEARNING:

1. Provide a comprehensive introduction to the Python programming language, focusing on its relevance and applications in the field of data science.
2. Familiarize participants with the syntax, basic data types, and control structures in Python.
3. Highlight key features of Python that make it a popular choice for data science tasks.
4. Lay the foundation for further exploration of Python's capabilities in data manipulation and analysis.

CONTENT OF MODULE:

1. **Overview of Python:**
    a. Introduction to Python as a versatile and high-level programming language.
    b. Historical background and evolution of Python.
    c. Python's popularity and its role in various domains.
2. **Python Basics:**
    a. Syntax rules and structure in Python programming.
    b. Variables, data types, and basic operations.
    c. Understanding the concept of indentation in Python.
3. **Python for Data Science**
    a) Why Python is a preferred language for data science.
    b) Overview of key libraries for data manipulation and analysis (e.g., NumPy, Pandas).

LEARNING OUTCOME:

1. Have a solid understanding of Python's syntax and basic constructs.
2. Be familiar with Python's control structures and data types.
3. Understand the significance of Python in the context of data science

MODULE- 3

(Python libraries for data science)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 17/10/2023

WEEK OF EXECUTION: 3rd Week

OBJECTIVE OF LEARNING:

1. Explore key Python libraries that play a pivotal role in data science and analysis.
2. Understand the functionalities and applications of these libraries in real-world data manipulation tasks.
3. Provide hands-on experience with popular data science libraries to empower participants in their data analysis journey.

CONTENT OF MODULE:

1. **NumPy:**
   a) Introduction to NumPy as a fundamental library for numerical computing in Python.
   b) Understanding NumPy arrays and their operations.
   c) Practical applications of NumPy in data manipulation.
2. **Pandas:**
   a) Overview of Pandas as a powerful data manipulation library.
   b) Working with Pandas DataFrames for efficient data organization.
   c) Data cleaning, filtering, and transformation using Pandas.

3. **Hands-On Projects:**

   a) Practical exercises and projects leveraging the discussed libraries.

   b) Applying acquired knowledge to real-world data science scenarios.

   c) Building a foundation for further exploration and experimentation

<u>LEARNING OUTCOME:</u>

Upon completion of this module, participants should:

a) Have a comprehensive understanding of key Python libraries for data science.

b) Be proficient in using NumPy for numerical operations and Pandas for data manipulation.

MODULE- 4

(Data Visualization libraries)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 21/10/2023

WEEK OF EXECUTION: **4th Week**

OBJECTIVE OF LEARNING:

1. Explore essential Python libraries for data visualization to effectively communicate insights.
2. Understand the capabilities and applications of these libraries in creating compelling visual representations of data.
3. Provide hands-on experience with popular data visualization tools to enhance participants' data storytelling skills

CONTENT OF MODULE:

1. Matplotlib:

a) Introduction to Matplotlib as a versatile plotting library.

b) Creating basic plots, line charts, and scatter plots.

c) Customizing visualizations for better interpretation.

2. Seaborn:

a) Overview of Seaborn for statistical data visualization.

b) Utilizing Seaborn's specialized plots for enhanced visual insights.

c) Styling and theming options in Seaborn.

3. Plotly:

a) Introduction to Plotly for interactive and dynamic visualizations.

b) Building interactive dashboards and plots.

c) Embedding Plotly visualizations in Jupyter Notebooks.

LEARNING OUTCOME:

Upon completion of this module, participants should:

1. Have a comprehensive understanding of key Python libraries for data visualization.
2. Be proficient in creating static and interactive visualizations using Matplotlib, Seaborn, Plotly
3. Gain practical experience in visualizing diverse datasets for effective storytelling.
4. Acquire skills to choose the appropriate visualization tool based on data characteristics and analysis goals.
5. Develop the ability to customize visualizations to enhance clarity and impact.

MODULE- 5

(Statistics)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 27/10/2023

WEEK OF EXECUTION: 5th week

OBJECTIVE OF LEARNING:

1. Develop a foundational understanding of statistical concepts and methods crucial for data science.
2. Apply statistical techniques to analyze and interpret data effectively.
3. Familiarize participants with statistical tools in Python for practical implementation in data science projects.

CONTENT OF MODULE:

1. Introduction to Statistics:

a) Overview of basic statistical concepts.
b) Descriptive vs. inferential statistics.
c) Role of statistics in data-driven decision-making.

2. Descriptive Statistics:

a) Measures of central tendency (mean, median, mode).
b) Measures of dispersion (range, variance, standard deviation).
c) Visual representation of data through histograms and box plots.

3. Probability Distributions:

a) Understanding probability and its role in statistics.
b) Common probability distributions (normal, binomial, Poisson).
c) Probability density functions and cumulative distribution functions.

4. Statistical Inference:

a) Hypothesis testing and significance levels.

b) Types of errors and power of tests.

c) Confidence intervals for parameter estimation.

5. **Correlation and Regression:**

a) Exploring relationships between variables.

b) Pearson correlation coefficient and its interpretation.

c) Simple linear regression for predictive modeling.

LEARNING OUTCOME:

Upon completion of this module, participants should:

1. Have a solid foundation in basic statistical concepts and methods.
2. Be proficient in using statistical tools in Python for data analysis.
3. Understand the principles of hypothesis testing and statistical inference.
4. Apply correlation and regression analysis for variable relationships.
5. Interpret and communicate results from statistical analyses effectively.

MODULE- 6

(Data Wrangling)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 02/11/2023

WEEK OF EXECUTION: 6th week

OBJECTIVE OF LEARNING:

1. Gain proficiency in data wrangling techniques to prepare raw data for analysis.
2. Understand the importance of data cleaning, transformation, and aggregation in the data science process.
3. Learn to handle missing data, outliers, and inconsistencies for robust data analysis.

CONTENT OF MODULE:

1. Introduction to Data Wrangling:

a) Definition and significance of data wrangling in the data science lifecycle.

b) Challenges and common issues in raw datasets.

2. Data Cleaning:

a) Identifying and handling missing data.

b) Dealing with duplicate records.

c) Correcting inconsistent data entries.

3. Data Transformation:

a) Converting data types for compatibility.

b) Extracting and creating new features.

c) Scaling and normalizing numerical data.

4. Handling Outliers:

a) Identifying outliers and their impact on analysis.

b) Strategies for handling outliers (removing, transforming, or imputing).

5. Merging and Joining Data:

a) Combining datasets through merging and joining operations.

b) Types of joins (inner, outer, left, right) and their applications.

c) Handling overlapping column names.

6. Aggregating Data:

a) Grouping data for summary statistics.

b) Aggregating data using functions (mean, sum, count).

c) Pivot tables and cross-tabulations.

7. Handling Time Series Data:

a) Introduction to time series data and its characteristics.

b) Resampling and frequency conversion.

c) Time-based indexing and slicing.

8. Data Wrangling in Python:

a) Overview of Python libraries for data wrangling (Pandas, NumPy).

b) Practical examples of data wrangling using Python scripts.

c) Utilizing Jupyter Notebooks for interactive data wrangling.

9. Hands-On Projects:

a) Applying data wrangling techniques to real-world datasets.

b) Addressing specific challenges in raw data to prepare for analysis.

c) Documenting and sharing data wrangling workflows

LEARNING OUTCOME:

Upon completion of this module, participants should:

a) Be proficient in identifying and handling missing data, duplicates, and inconsistencies.

b) Understand various techniques for transforming and cleaning data.

c) Know how to handle outliers for robust data analysis.

d) Master the art of merging, joining, and aggregating data effectively.

e) Feel comfortable working with time series data.

f) Have practical experience with data wrangling in Python using Pandas and NumPy.

MODULE- 7

(**Feature Engineering**)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 08/11/2023

WEEK OF EXECUTION: 7th week

OBJECTIVE OF LEARNING:

1. Understand the significance of feature engineering in enhancing the performance of machine learning models.
2. Gain knowledge of techniques to create informative and relevant features from raw data.
3. Explore methods to handle categorical data, create interaction features, and address issues of dimensionality.

CONTENT OF MODULE:

1. Introduction to Feature Engineering:

a) Definition and importance of feature engineering in the machine learning pipeline.

b) Impact of quality features on model performance.

c) Overview of common feature engineering techniques.

2. Handling Categorical Data:

a) Strategies for encoding categorical variables (label encoding, one-hot encoding).

b) Dealing with high cardinality categorical features.

c) Feature hashing for large categorical datasets.

3. Creating Interaction Features:

a) Importance of interaction features in capturing complex relationships.

b) Techniques for creating interaction terms between variables.

c) Polynomial features and their application.

4. Handling Missing Data:

a) Strategies for handling missing values in features.

b) Imputation techniques (mean, median, forward-fill, backward-fill).

c) Impact of missing data on model performance.

5. Dimensionality Reduction:

a) Introduction to dimensionality reduction techniques.

b) Principal Component Analysis (PCA) for reducing feature space.

c) Singular Value Decomposition (SVD) and its applications.

6. Feature Scaling:

a) Importance of feature scaling in machine learning models.

b) Techniques for scaling features (min-max scaling, standardization).

c) Impact of feature scaling on different algorithms.

7. Feature Selection:

a) Identifying and removing irrelevant or redundant features.

b) Univariate and recursive feature selection methods.

c) Feature importance techniques.

8. Handling Time and Date Data:

a) Extracting information from time and date features.

b) Creating temporal features for machine learning models.

c) Handling time-based patterns in data.

9. Feature Engineering in Python:

a) Implementing feature engineering techniques using Python libraries (Pandas, Scikit-Learn).

b) Utilizing Jupyter Notebooks for interactive feature engineering.

10. Hands-On Projects:

a) Applying feature engineering techniques to enhance model performance.

b) Experimenting with different feature engineering strategies.

c) Documenting and sharing feature engineering workflows.

LEARNING OUTCOME:

Upon completion of this module, participants should:

a) Understand the importance of feature engineering in machine learning.

b) Be proficient in handling categorical data and creating interaction features.

c) Know how to deal with missing data and address issues of dimensionality.

d) Gain practical experience with feature scaling, selection, and extraction.

e) Be able to apply feature engineering techniques using Python.

f) Have hands-on experience with real-world feature engineering projects.

MODULE- 8

(Exploratory Data Analysis)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 14/11/2023

WEEK OF EXECUTION: 8th week

OBJECTIVE OF LEARNING:

1. Develop skills to explore and understand raw data before formal modeling.
2. Learn techniques for visualizing and summarizing data to uncover patterns and insights.
3. Gain proficiency in using exploratory data analysis (EDA) to inform subsequent steps in the data science process.

CONTENT OF MODULE:

1. **Introduction to Exploratory Data Analysis (EDA):**
    a. Definition and significance of EDA in the data science workflow.
    b. Role of EDA in hypothesis generation and data understanding.
    c. Overview of common EDA techniques.

2. **Data Summary Statistics:**
    a. Descriptive statistics for numerical and categorical data.
    b. Central tendency, dispersion, and distribution of data.
    c. Utilizing summary statistics to understand data characteristics.

3. **Data Distribution and Transformation:**
    a. Understanding data distribution shapes (normal, skewed, bimodal).
    b. Box plots and violin plots for visualizing distributions.
    c. Log and power transformations for non-normally distributed data.

4. **Data Visualization Tools in Python:**
    a. Introduction to Python libraries for data visualization (Matplotlib, Seaborn).
    b. Creating interactive visualizations with Plotly.
    c. Utilizing Jupyter Notebooks for dynamic EDA.

5. **Exploratory Data Analysis Projects:**
    a. Hands-on projects applying EDA techniques to real-world datasets.
    b. Documenting insights and patterns discovered during EDA.
    c. Communicating findings through visualizations and summary reports.

LEARNING OUTCOME:

Upon completion of this module, participants should:

a. Understand the role and significance of Exploratory Data Analysis in data science.

b. Be proficient in summarizing data using descriptive statistics.

c. Know how to perform univariate, bivariate, and multivariate analyses.

d. Gain skills in visualizing data distributions and transformations.

e. Be able to use Python libraries for effective data visualization.

f. Have hands-on experience with EDA projects, uncovering insights for subsequent data science tasks.

MODULE- 9

(Feature Selection)

NAME : Lalit Singh
COURSE: M.C.A
SEMESTER: 1st
ROLL NO.: 2351027
DATE: 20/11/2023

WEEK OF EXECUTION: 9th week

OBJECTIVE OF LEARNING:

1. Understand the importance of feature selection in optimizing model performance and reducing complexity.
2. Explore various techniques for selecting relevant features in a dataset.
3. Gain hands-on experience in implementing feature selection strategies using Python.

CONTENT OF MODULE:

1. Introduction to Feature Selection:

1. Definition and significance of feature selection in the machine learning pipeline.
2. Impact of irrelevant or redundant features on model performance.
3. Overview of common feature selection techniques.

2. Filter Methods:

1. Overview of filter methods for feature selection.
2. Statistical measures (e.g., correlation, chi-square) for ranking features.
3. Selecting top features based on statistical scores.

3. Wrapper Methods:

1. Introduction to wrapper methods for feature selection.
2. Recursive Feature Elimination (RFE) and Forward/Backward Selection.
3. Evaluating model performance during feature selection.

4. Embedded Methods:

1. Overview of embedded methods for feature selection.
2. Regularization techniques (e.g., Lasso, Ridge) and their impact on feature selection.
3. Feature importance from tree-based models (e.g., Random Forest).

5. Dimensionality Reduction vs. Feature Selection:

1. Understanding the difference between dimensionality reduction and feature selection.
2. Principal Component Analysis (PCA) as a dimensionality reduction technique.
3. When to choose feature selection over dimensionality reduction.

6.Sequential Feature Selection:

1. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS).
2. Dynamic programming approach to search optimal feature subsets.
3. Balancing computational complexity and model performance.

7. Advanced Feature Selection Techniques:

1. Overview of advanced techniques (e.g., Boruta algorithm, SHAP values).
2. Addressing challenges in high-dimensional datasets.
3. Ensemble methods for feature selection.

8. Feature Selection in Python:

1. Implementation of feature selection techniques using Python libraries (Scikit-Learn).
2. Utilizing Jupyter Notebooks for interactive feature selection workflows.

LEARNING OUTCOME:

Upon completion of this module, participants should:

1. Understand the importance of feature selection in enhancing model performance.
2. Be proficient in implementing filter, wrapper, and embedded methods for feature selection.
3. Differentiate between dimensionality reduction and feature selection.
4. Gain practical experience with advanced feature selection techniques.
5. Be able to apply feature selection using Python libraries for real-world datasets.
6. Have hands-on experience with feature selection projects, optimizing models for better interpretability and efficiency.