Six Week Summer Training Report

Data Science

Submitted by:-

Name: Anubhav Singh
Reg. no.: 12208856
Programme Name: Bachelors in Computer Applications (BCA)

School of Computer Application
Lovely Professional University
Phagwara, Punjab
(June – July 2024)

DECLARATION:

I hereby declare that I have completed my Six weeks Summer training at **Coursera** online platform from 10/06/2024 to 31/07/2024. I have declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements of training for the award of degree of **Bachelor Of Computer Application**, Lovely Professional University, Phagwara.

Anubhav Singh
12208856

Date: 31/07/2024

Acknowledgement

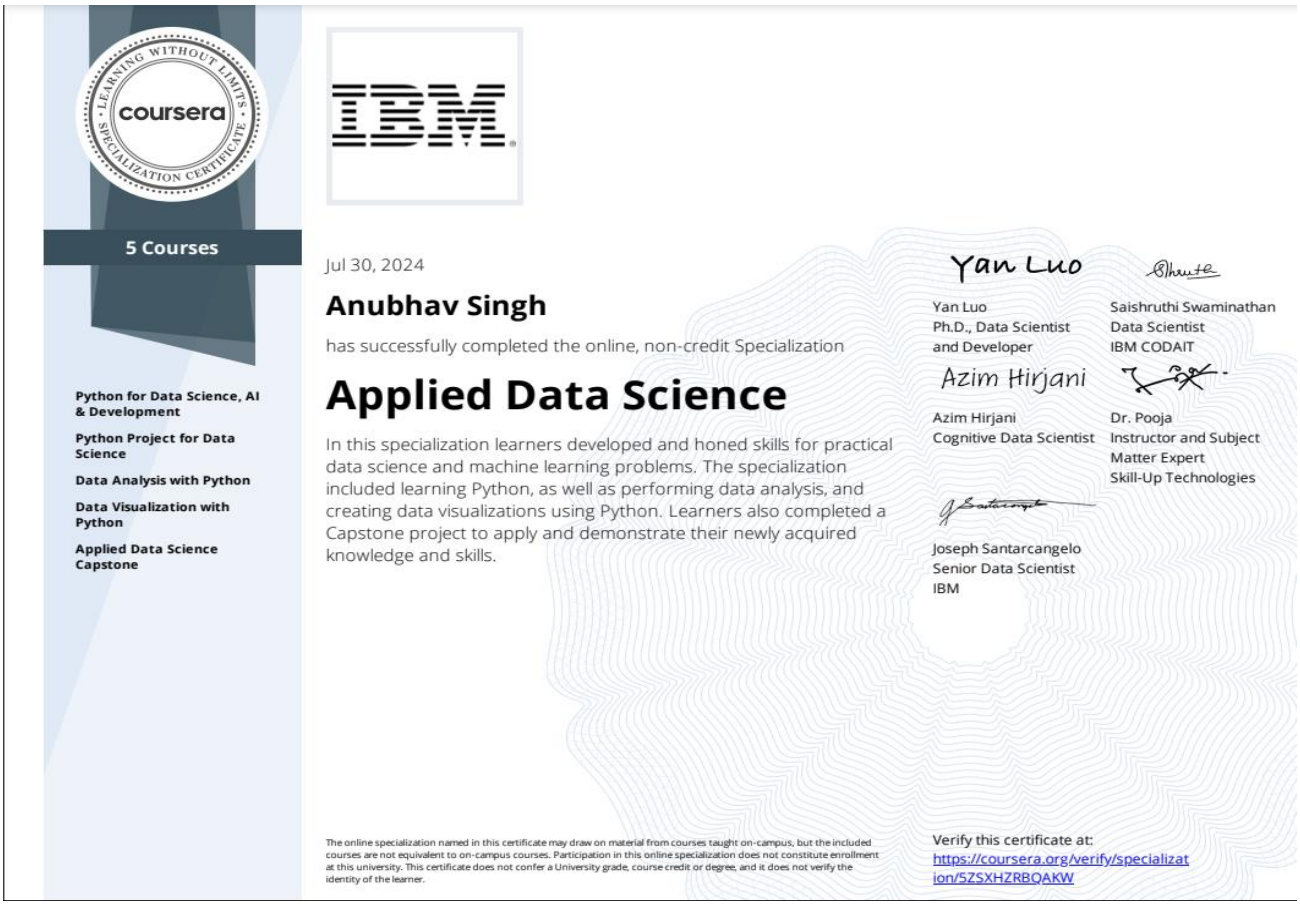
I would like to express my sincere gratitude to Coursera, for providing their invaluable guidance, comments and suggestions throughout the course. I would specially thank my training mentor for constantly motivating me to work harder and for encouragement and invaluable assistance.
Also, I would like to thanks to my parent for giving encouragement and support. Without their constant support, I might not be able to complete this training properly. I would also like to say thanks to my friends for being there for me when I needed their help.

Anubhav Singh
12208856

Date: 31/07/2024

Table Of Contents

Training Certificate



LEARNING WITHOUT LIMITS · coursera · SPECIALIZATION CERTIFICATE

5 Courses

- Python for Data Science, AI & Development
- Python Project for Data Science
- Data Analysis with Python
- Data Visualization with Python
- Applied Data Science Capstone

IBM

Jul 30, 2024

Anubhav Singh

has successfully completed the online, non-credit Specialization

Applied Data Science

In this specialization learners developed and honed skills for practical data science and machine learning problems. The specialization included learning Python, as well as performing data analysis, and creating data visualizations using Python. Learners also completed a Capstone project to apply and demonstrate their newly acquired knowledge and skills.

Yan Luo

Yan Luo
Ph.D., Data Scientist
and Developer

Saishruthi Swaminathan
Data Scientist
IBM CODAIT

Azim Hirjani

Azim Hirjani
Cognitive Data Scientist

Dr. Pooja
Instructor and Subject
Matter Expert
Skill-Up Technologies

Joseph Santarcangelo
Senior Data Scientist
IBM

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
https://coursera.org/verify/specialization/5ZSXHZRBQAKW

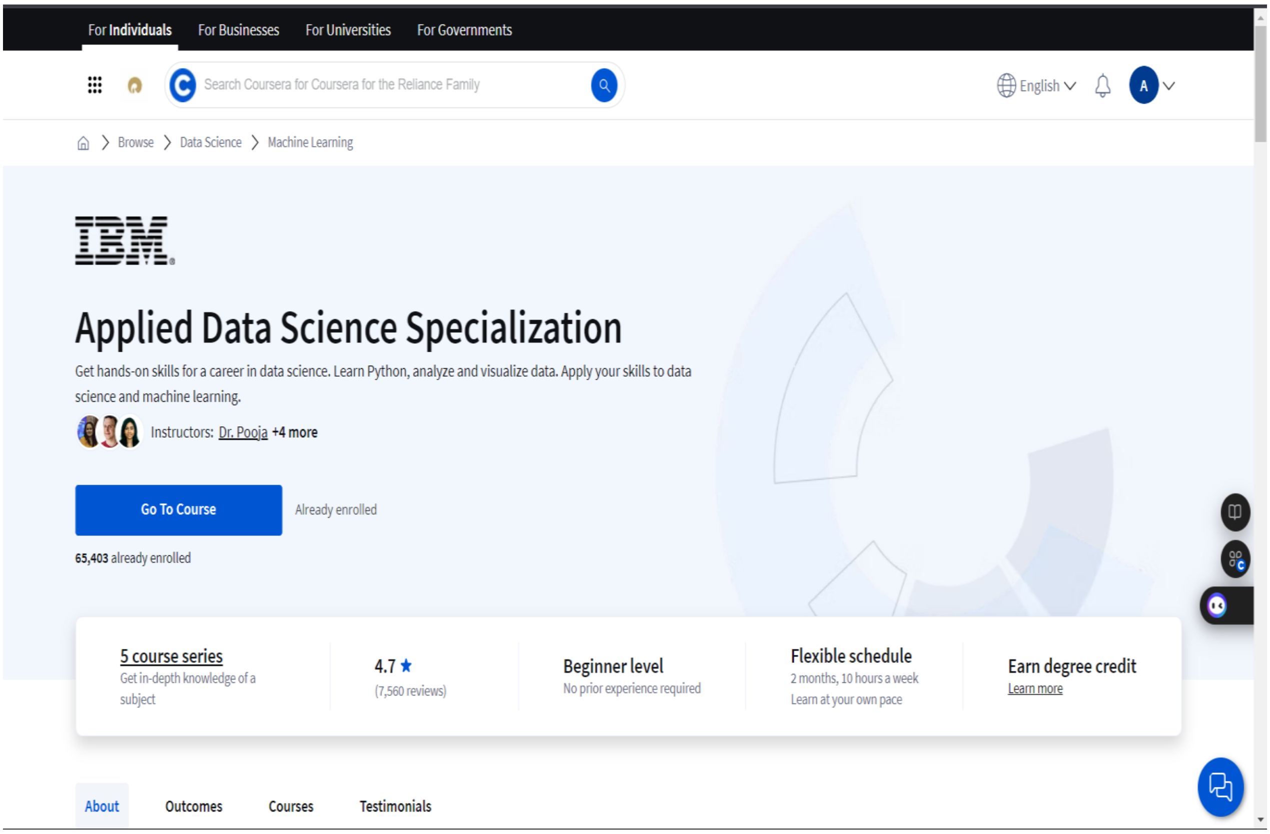
INTRODUCTION

During the summer vacation, I had the opportunity to enhance my skills in data science by enrolling in the IBM Applied Data Science course on Coursera. This comprehensive course provided me with a solid foundation in various aspects of data science, from the fundamentals of data analysis to the application of machine learning algorithms. The course was designed by industry experts from IBM, offering both theoretical knowledge and practical experience through hands-on labs and projects.

The curriculum covered a broad spectrum of topics, including Python programming, data visualization, statistical analysis, and the use of popular data science tools such as Jupyter Notebooks, Pandas, and Matplotlib. Additionally, the course introduced me to advanced concepts like machine learning, data modeling, and the deployment of data science models in a business context.

Through this course, I gained valuable insights into the data science workflow, from data collection and cleaning to exploratory data analysis and model building. The practical exercises allowed me to apply what I learned in real-world scenarios, reinforcing my understanding of the concepts.

This report aims to summarize the key learnings from the IBM Applied Data Science course, highlighting the skills acquired and the projects completed. It will also reflect on the overall experience and how this course has prepared me for future challenges in the field of data science.



Applied Data Science Specialization

Get hands-on skills for a career in data science. Learn Python, analyze and visualize data. Apply your skills to data science and machine learning.

Instructors: Dr. Pooja +4 more

65,403 already enrolled

5 course series — Get in-depth knowledge of a subject

4.7 ★ (7,560 reviews)

Beginner level — No prior experience required

Flexible schedule — 2 months, 10 hours a week. Learn at your own pace

Earn degree credit — Learn more

Chapter 1:
Python for Data Science, AI & Development

Python has emerged as the de facto language for data science, artificial intelligence (AI), and development due to its simplicity, versatility, and a vast ecosystem of libraries and tools. Let's delve deeper into why Python is so popular in these fields:

1. Readability and Simplicity:

- Clean syntax: Python's syntax is designed to be easy to read and understand, reducing the learning curve for beginners.
- English-like structure**:** Python uses English keywords and a natural language style, making it more intuitive.

2. Versatility:

- Data analysis: Python's libraries like NumPy, Pandas, and Matplotlib are powerful tools for data manipulation, analysis, and visualization.
- Machine learning: Scikit-learn, TensorFlow, and PyTorch provide comprehensive frameworks for building and training machine learning models.
- Deep learning: Frameworks like Keras and TensorFlow offer high-level APIs for designing and implementing deep neural networks.
- Web development: Django and Flask are popular Python web frameworks for building dynamic web applications.
- Automation: Python can be used to automate repetitive tasks and streamline workflows.

3. Large and Active Community:

- Extensive documentation and tutorials: The Python community has created a wealth of resources to help users learn and solve problems.
- Active forums and online communities: There are numerous online platforms where developers can ask questions, share knowledge, and collaborate.
- Third-party libraries: The Python Package Index (PyPI) offers a vast repository of libraries for various tasks.

4. Platform Independence:

- Cross-platform compatibility: Python code can run on Windows, macOS, Linux, and other operating systems.

5. Integration with Other Tools:

- Seamless integration: Python can easily be integrated with other programming languages and tools.

Key Libraries and Frameworks:

- **NumPy:** For numerical computations and array operations.

- **Pandas:** For data manipulation and analysis.
- **Matplotlib:** For creating visualizations.
- **Scikit-learn:** For machine learning algorithms
- **TensorFlow:** For deep learning and machine learning.
- **PyTorch:** Another popular deep learning framework.
- **Django:** A web framework for building complex web applications.
- **Flask:** A lightweight web framework for smaller-scale applications.

In conclusion, Python's combination of readability, versatility, and a strong community has made it an indispensable tool for data scientists, AI researchers, and developers. Whether you're exploring data analysis, building machine learning models, or developing web applications, Python offers a powerful and flexible platform to achieve your goals.

Skills Learnt:

1.Data Science:

Data science is an interdisciplinary field that combines statistics, computer science, and domain knowledge to extract insights from data. It involves techniques such as data collection, cleaning, analysis, and modeling.

Advantages:

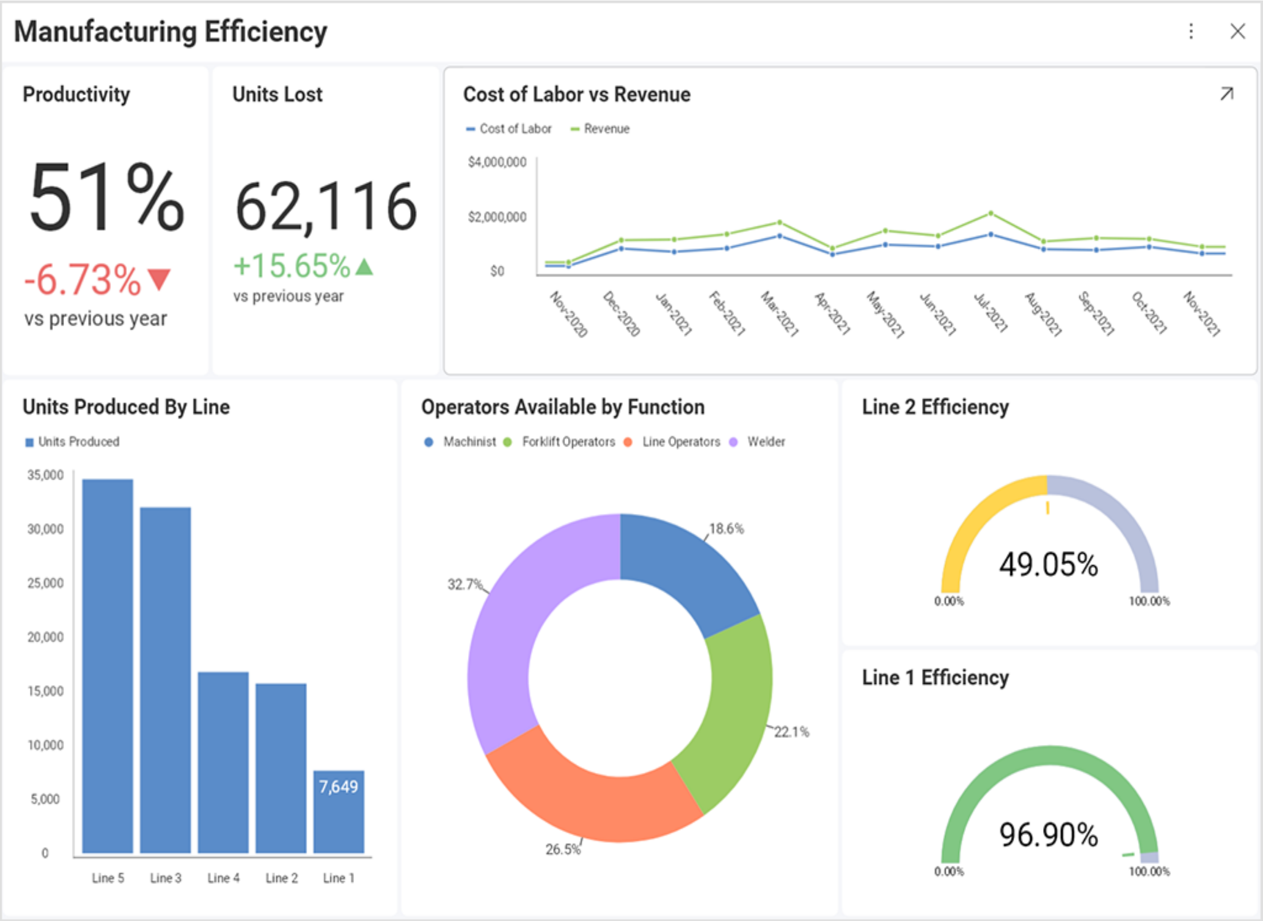
- Solves problems with data: Uses data to answer questions and find solutions.
- Makes businesses better: Helps companies improve their operations and make more money.

2. Data Analysis:

Data analysis is the process of examining data to discover patterns, trends, and insights. It involves techniques such as data cleaning, exploration, and statistical analysis.

Advantages:

- Uncovers hidden treasures: Finds interesting patterns and insights in your data.
- Makes smart decisions: Helps you make informed choices based on the data.



3. Python Programming:

Python is a high-level, general-purpose programming language widely used in data science. Its simplicity, versatility, and extensive libraries make it a popular choice for tasks like data analysis, machine learning, and web development.

Advantages:

- Easy to learn: A friendly language that's simple to understand.
- Lots of tools: Has many libraries to help you do data science tasks.
- Popular and helpful: Many people use it, so you can find lots of help and resources.

4 . NumPy:

NumPy is a Python library for numerical computing. It provides support for large arrays and matrices, along with a collection of mathematical functions.

Advantages:

- Does math really fast: Handles numbers and calculations efficiently.
- Forms the base: A foundation for many other data science tools.

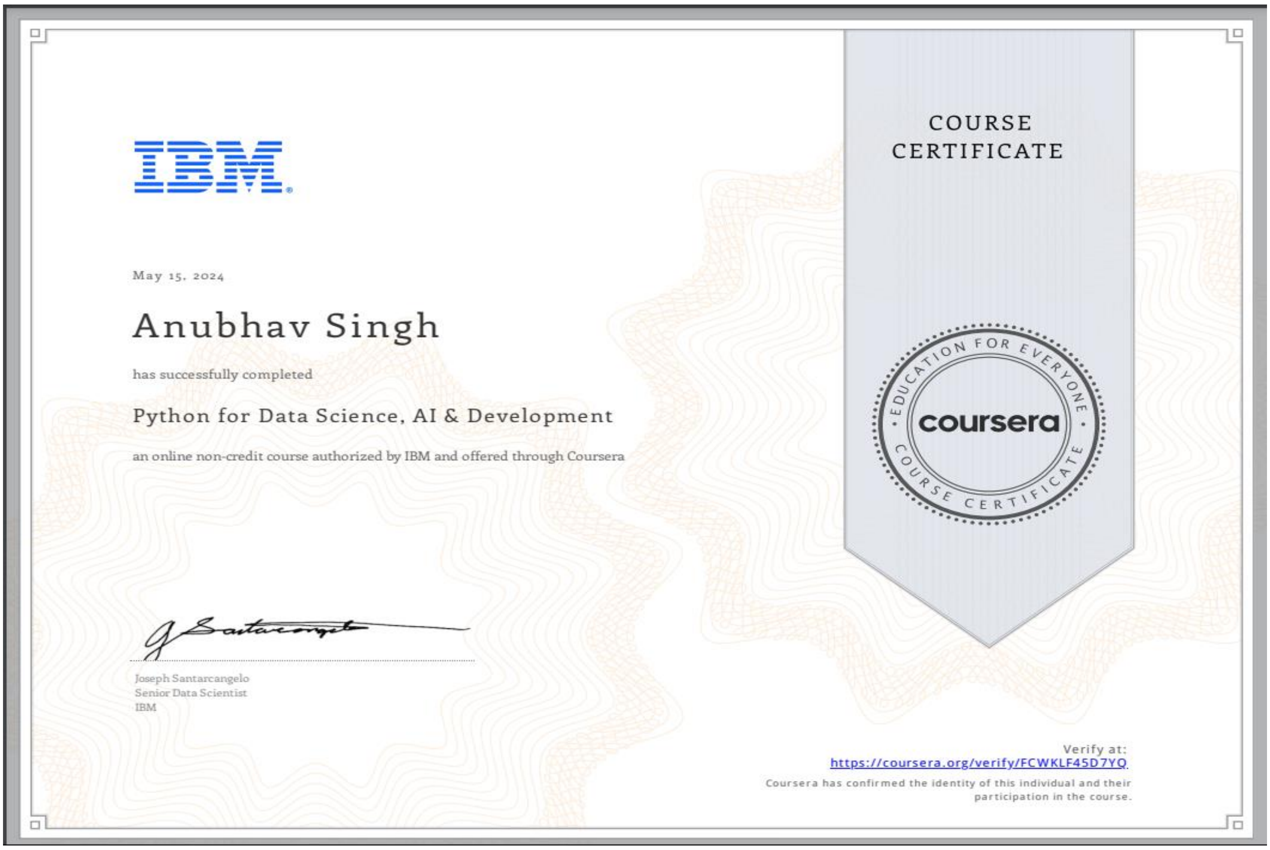
5. Pandas:

Pandas is a Python library for data manipulation and analysis. It provides data structures like Data Frames, which make it easy to work with structured data.

Advantages:

- Works with data like a pro: Makes it easy to manipulate and analyze data.
- Handles big data: Can handle large datasets without problems.

Certificate



IBM

May 15, 2024

Anubhav Singh

has successfully completed

Python for Data Science, AI & Development

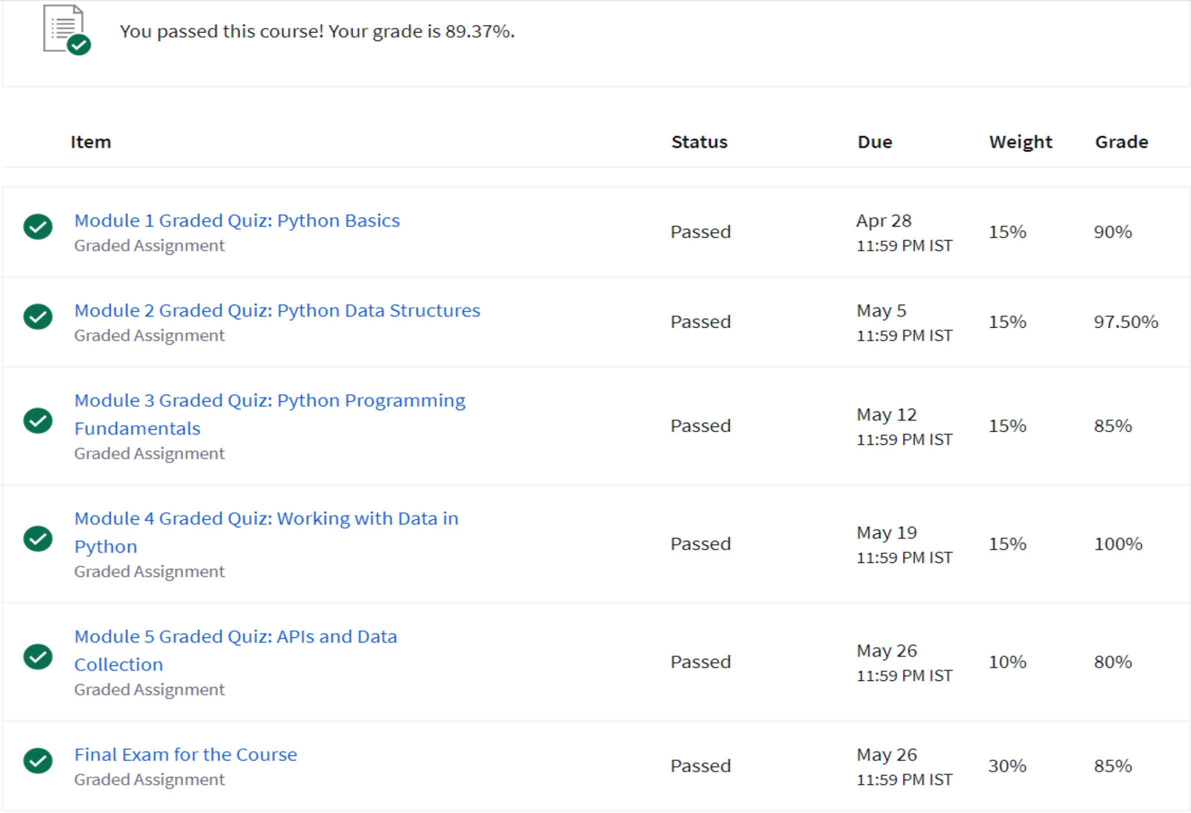
an online non-credit course authorized by IBM and offered through Coursera

Joseph Santarcangelo
Senior Data Scientist
IBM

COURSE CERTIFICATE

EDUCATION FOR EVERYONE
coursera
COURSE CERTIFICATE

Verify at:
https://coursera.org/verify/FCWKLF45D7YQ
Coursera has confirmed the identity of this individual and their participation in the course.



You passed this course! Your grade is 89.37%.

	Item	Status	Due	Weight	Grade
✓	Module 1 Graded Quiz: Python Basics Graded Assignment	Passed	Apr 28 11:59 PM IST	15%	90%
✓	Module 2 Graded Quiz: Python Data Structures Graded Assignment	Passed	May 5 11:59 PM IST	15%	97.50%
✓	Module 3 Graded Quiz: Python Programming Fundamentals Graded Assignment	Passed	May 12 11:59 PM IST	15%	85%
✓	Module 4 Graded Quiz: Working with Data in Python Graded Assignment	Passed	May 19 11:59 PM IST	15%	100%
✓	Module 5 Graded Quiz: APIs and Data Collection Graded Assignment	Passed	May 26 11:59 PM IST	10%	80%
✓	Final Exam for the Course Graded Assignment	Passed	May 26 11:59 PM IST	30%	85%

Chapter 2:
Python Project for Data Science

Project is crucial for solidifying your data science skills and gaining practical experience. Here are some project ideas that cover a range of difficulty levels and applications:

Beginner-Friendly Projects:

1. Predicting House Prices: This classic project involves building a regression model to predict house prices based on features like square footage, number of bedrooms, and location.
2. Customer Churn Prediction: Using a classification model, you can predict which customers are likely to churn based on their behavior and demographics.
3. Sentiment Analysis: Analyze social media data to determine the sentiment of people towards a particular topic or product.
4. Sales Forecasting: Build a time series model to predict future sales trends.

Intermediate-Level Projects:

1. Recommendation System: Create a recommendation system for movies, products, or articles using techniques like collaborative filtering or content-based filtering.
2. Image Classification: Train a convolutional neural network (CNN) to classify images into different categories.
3. Natural Language Processing (NLP) Tasks: Experiment with tasks like text summarization, machine translation, or question answering.
4. Fraud Detection: Develop a model to identify fraudulent transactions in financial data.

Advanced-Level Projects:

1. Time Series Forecasting with Deep Learning: Use deep learning models like LSTM or GRU to forecast complex time series data.
2. Computer Vision Applications: Explore projects like object detection, image segmentation, or facial recognition.
3. Reinforcement Learning: Train an agent to learn optimal actions in a given environment through trial and error.
4. Generative Adversarial Networks (GANs): Create realistic images or other data using GANs.

Tips for Choosing and Executing Projects:

- Start small and gradually increase complexity.
- Leverage existing datasets and libraries.
- Document your work and share your findings.
- Collaborate with others and learn from their experiences.
- Continuously update your skills and knowledge.

By working on these projects, i'll gain practical experience with data cleaning, feature engineering, model building, evaluation, and deployment**.** Remember, the most important thing is to have fun and learn something new along the way!

Skills Learnt:

1.Data Science:

Data science is an interdisciplinary field that combines statistics, computer science, and domain knowledge to extract insights from data. It involves techniques such as data collection, cleaning, analysis, and modeling.

Advantages:

- Solves problems with data: Uses data to answer questions and find solutions.
- Makes businesses better: Helps companies improve their operations and make more money.

2.Data Analysis:

Data analysis is the process of examining data to discover patterns, trends, and insights. It involves techniques such as data cleaning, exploration, and statistical analysis.

Advantages:

- Uncovers hidden treasures: Finds interesting patterns and insights in your data.
- Makes smart decisions**:** Helps you make informed choices based on the data.

3.Python Programming:

Python is a high-level, general-purpose programming language widely used in data science. Its simplicity, versatility, and extensive libraries make it a popular choice for tasks like data analysis, machine learning, and web development.

Advantages:

- Easy to learn: A friendly language that's simple to understand.
- Lots of tools: Has many libraries to help you do data science tasks.

Popular and helpful**:** Many people use it, so you can find lots of help and resources

4.Pandas:

Pandas is a Python library for data manipulation and analysis. It provides data structures like Data Frames, which make it easy to work with structured data.

Advantages:

- Works with data like a pro: Makes it easy to manipulate and analyze data.

- Handles big data: Can handle large datasets without problems.

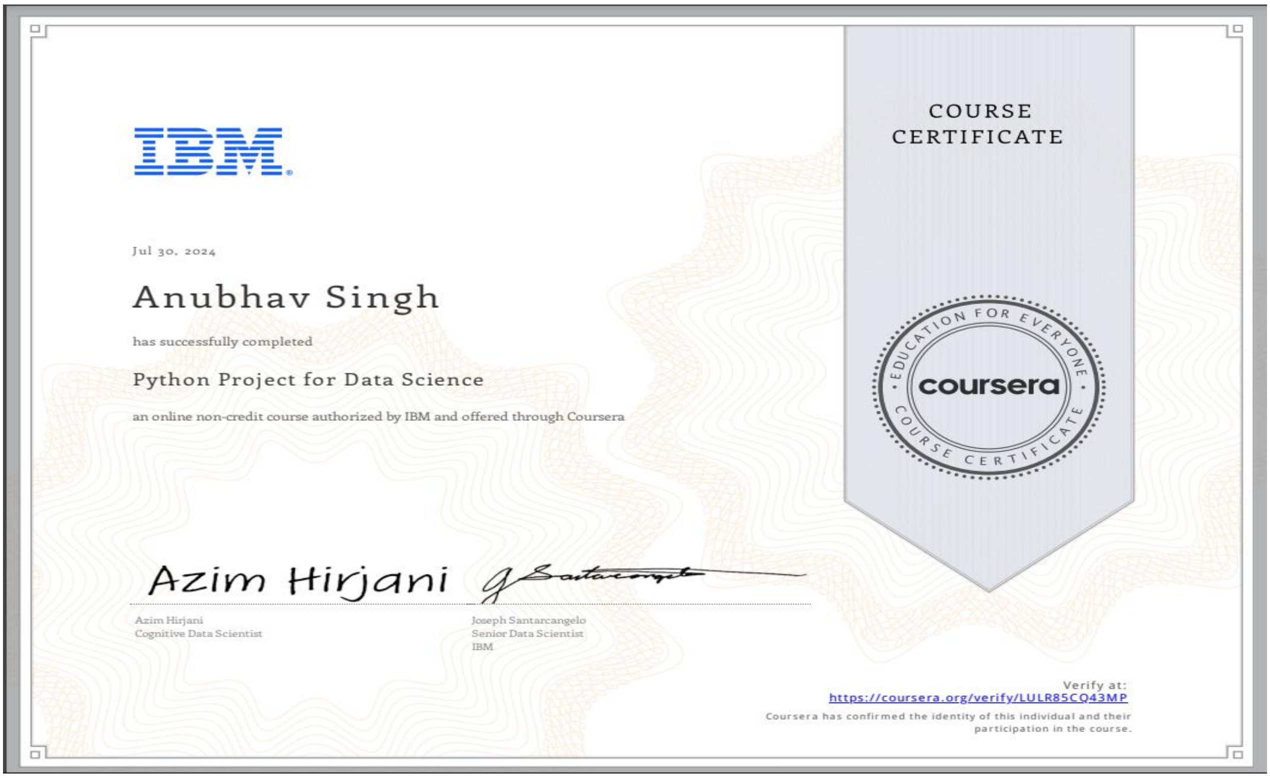
5.Jupyter Notebooks:

Jupyter Notebooks is an interactive environment for creating and sharing documents containing live code, equations, visualizations, and explanatory text. It is commonly used in data science for its interactive nature.

Advantages:

- Interactive playground: Lets you try things out and see the results immediately.
- Explains your work: Helps you document your code and explain what it does.

Certificate:



IBM

COURSE CERTIFICATE

Jul 30, 2024

Anubhav Singh

has successfully completed

Python Project for Data Science

an online non-credit course authorized by IBM and offered through Coursera

Azim Hirjani

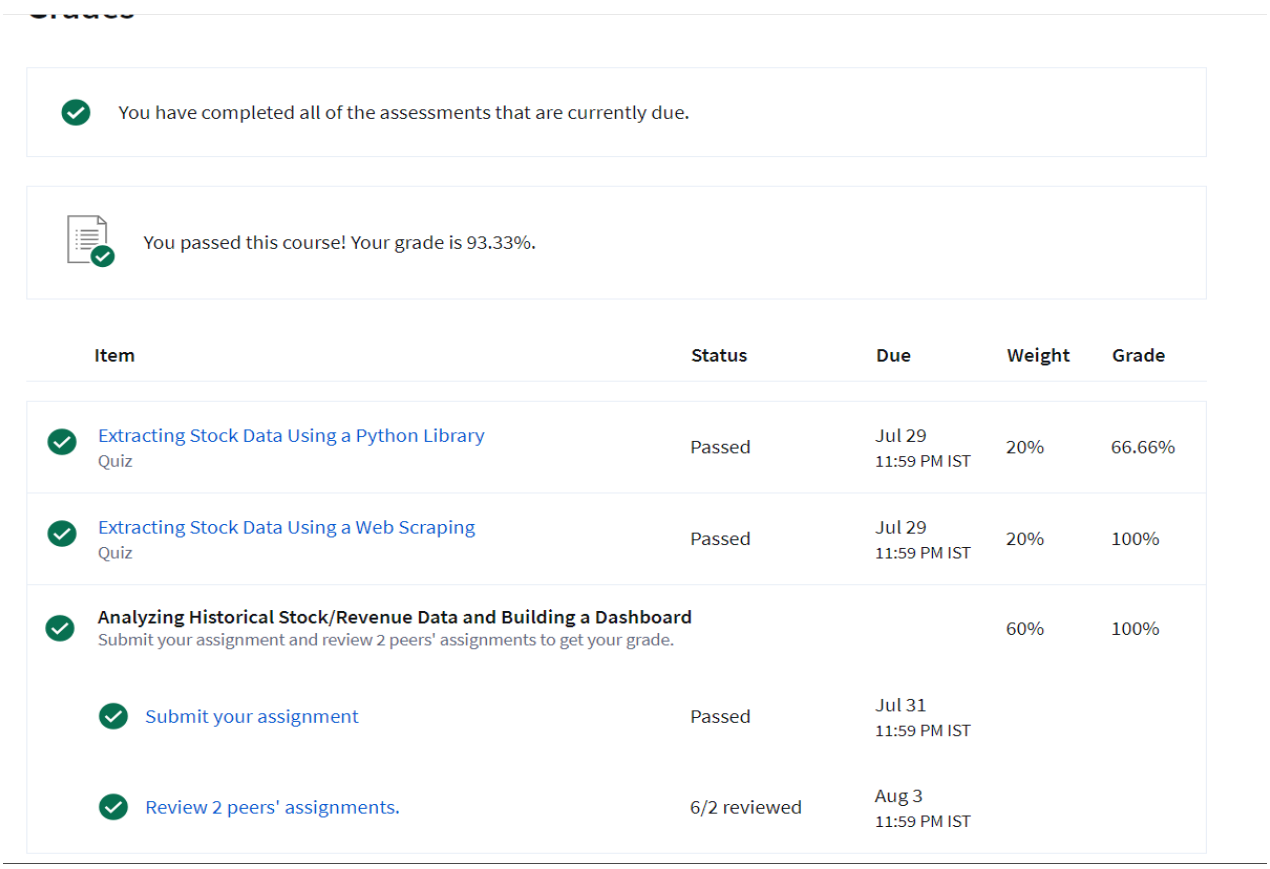
Azim Hirjani
Cognitive Data Scientist

Joseph Santarcangelo
Senior Data Scientist
IBM

EDUCATION FOR EVERYONE · coursera · COURSE CERTIFICATE

Verify at:
https://coursera.org/verify/LULR85CQ43MP

Coursera has confirmed the identity of this individual and their participation in the course.



✓ You have completed all of the assessments that are currently due.

You passed this course! Your grade is 93.33%.

	Item	Status	Due	Weight	Grade
✓	Extracting Stock Data Using a Python Library Quiz	Passed	Jul 29 11:59 PM IST	20%	66.66%
✓	Extracting Stock Data Using a Web Scraping Quiz	Passed	Jul 29 11:59 PM IST	20%	100%
✓	**Analyzing Historical Stock/Revenue Data and Building a Dashboard** Submit your assignment and review 2 peers' assignments to get your grade.			60%	100%
	✓ Submit your assignment	Passed	Jul 31 11:59 PM IST		
	✓ Review 2 peers' assignments.	6/2 reviewed	Aug 3 11:59 PM IST		

Chapter 3:
Data Analysis with Python

A Powerful Toolset
Python has become a go-to language for data analysis due to its simplicity, versatility, and a rich ecosystem of libraries. Let's explore some key Python libraries and their applications in data analysis:

Essential Libraries

- NumPy:
  - Provides efficient multi-dimensional arrays and matrices.
  - Offers mathematical functions for operations like linear algebra, Fourier transforms, and random number generation.
- Pandas:
  - Offers data structures like DataFrames and Series for data manipulation and analysis.
  - Provides functions for data cleaning, filtering, grouping, and aggregation.
- Matplotlib:
  - Creates static, animated, and interactive visualizations.
  - Supports various plot types like line plots, scatter plots, histograms, and bar charts.
- Seaborn:
  - Builds on Matplotlib and provides a higher-level interface for creating attractive statistical visualizations.

Additional Libraries for Specific Tasks

- Scikit-learn:
  - Provides algorithms for machine learning tasks like classification, regression, clustering, and dimensionality reduction.
- Statsmodels:
  - Offers statistical modeling tools for econometric and statistical analysis.
- NLTK (Natural Language Toolkit):
  - Provides tools for text processing, classification, and analysis.
- Plotly:
  - Creates interactive and customizable visualizations.

Common Data Analysis Tasks

- Data Cleaning and Preparation:
  - Handling missing values, outliers, and inconsistencies.
  - Transforming data into a suitable format for analysis.
- Exploratory Data Analysis (EDA):
  - Summarizing data characteristics.
  - Identifying patterns, trends, and relationships.
  - Visualizing data to gain insights.
- Statistical Analysis:
  - Calculating summary statistics (mean, median, mode, standard deviation).
  - Conducting hypothesis testing and statistical inference.
- Machine Learning:
  - Building and training models to make predictions or classifications.
  - Evaluating model performance.

- Data Visualization:
    - Creating informative and visually appealing charts and graphs.

Example: Analyzing Customer Churn Data

1. Load the data using Pandas.
2. Clean the data by handling missing values and outliers.
3. Explore the data using summary statistics and visualizations.
4. Feature engineering to create new features or transform existing ones.
5. Build a model (e.g., logistic regression) to predict customer churn.
6. Evaluate the model using metrics like accuracy, precision, recall, and F1-score.

Python's powerful libraries and ecosystem make it an ideal choice for data analysts and scientists.

Skills Learnt

1.Model Selection:

Model selection is the process of choosing the most appropriate machine learning algorithm for a given problem. It involves considering factors such as the nature of the data, the desired outcome, and the computational resources available.

Advantages:

- Finds the best tool for the job: Helps you pick the right machine learning algorithm to solve your problem.
- Avoids mistakes: Prevents you from using the wrong tool and getting bad results

2.Data Analysis:

Data analysis is the process of examining data to discover patterns, trends, and insights. It involves techniques such as data cleaning, exploration, and statistical analysis.

Advantages:

- Uncovers hidden treasures: Finds interesting patterns and insights in your data.

- Makes smart decisions**:** Helps you make informed choices based on the data.

3.Python Programming:

Python is a high-level, general-purpose programming language widely used in data science. Its simplicity, versatility, and extensive libraries make it a popular choice for tasks like data analysis, machine learning, and web development.

Advantages:

- Easy to learn: A friendly language that's simple to understand.

- Lots of tools: Has many libraries to help you do data science tasks.

Popular and helpful: Many people use it, so you can find lots of help and resources

4.Data Visualization:

Data visualization is the process of representing data graphically to make it easier to understand and interpret. It involves using charts, graphs, and other visual techniques to communicate insights.

Advantages:

- Makes data look cool: Turns numbers into pictures that are easy to understand.
- Tells a story**:** Helps you explain your findings in a clear and interesting way.

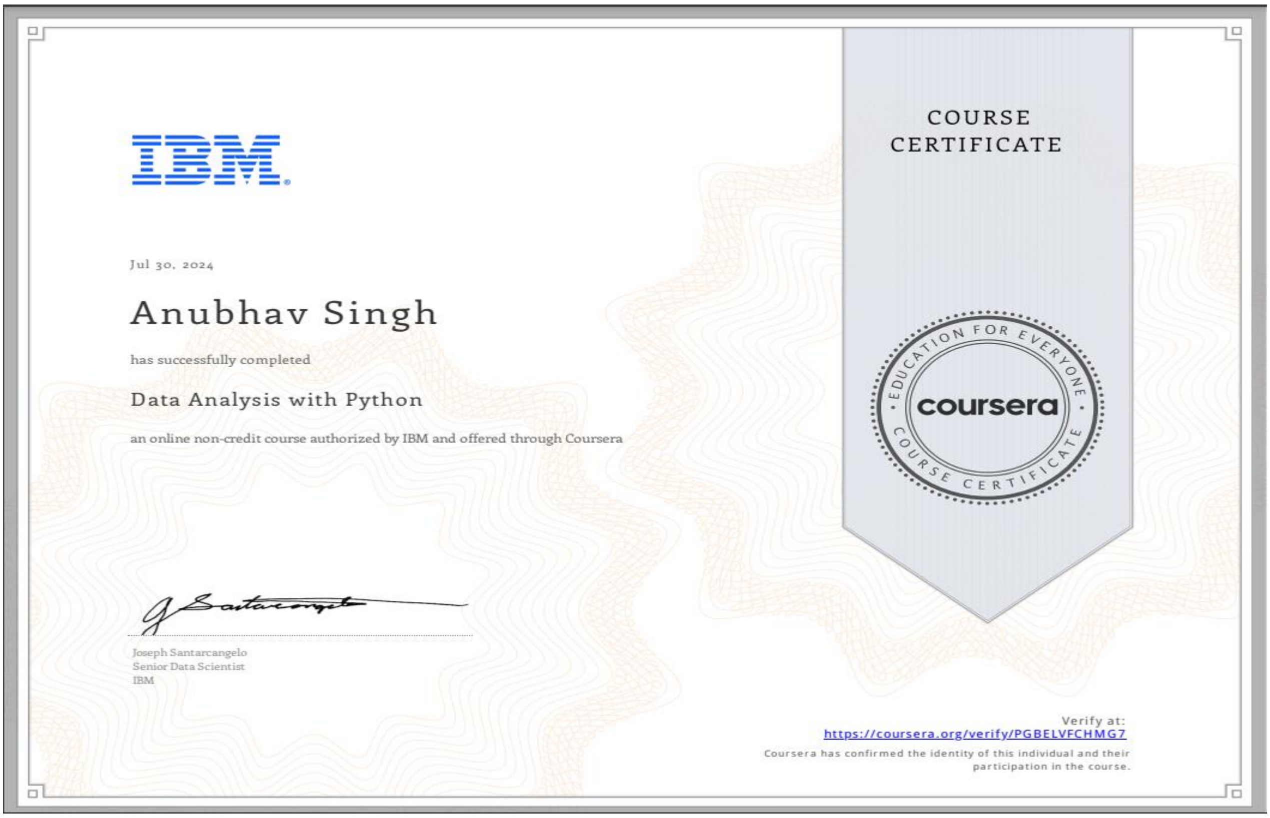
5.Predictive Modeling:

Predictive modeling is the process of building models that can predict future outcomes based on past data. It involves using machine learning algorithms to identify patterns in data and make predictions.

Advantages:

- Forecasts the future: Helps you guess what might happen next.
- Makes better plans**:** Guides your decisions by showing possible outcomes.

Certificate:



IBM

COURSE CERTIFICATE

Jul 30, 2024

Anubhav Singh

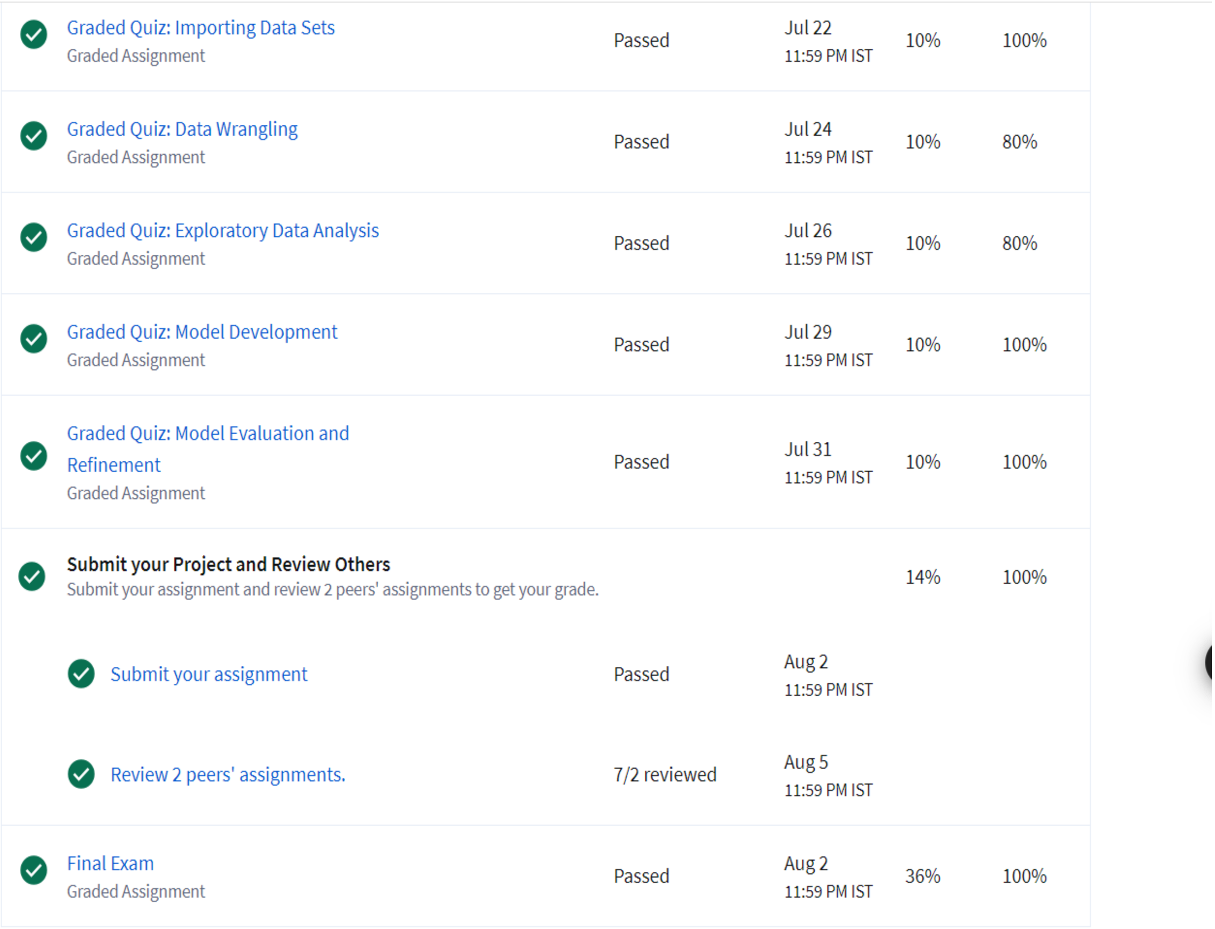
has successfully completed

Data Analysis with Python

an online non-credit course authorized by IBM and offered through Coursera

Joseph Santarcangelo
Senior Data Scientist
IBM

Verify at:
https://coursera.org/verify/PGBELVFCHMG7
Coursera has confirmed the identity of this individual and their participation in the course.



✓	Graded Quiz: Importing Data Sets Graded Assignment	Passed	Jul 22 11:59 PM IST	10%	100%
✓	Graded Quiz: Data Wrangling Graded Assignment	Passed	Jul 24 11:59 PM IST	10%	80%
✓	Graded Quiz: Exploratory Data Analysis Graded Assignment	Passed	Jul 26 11:59 PM IST	10%	80%
✓	Graded Quiz: Model Development Graded Assignment	Passed	Jul 29 11:59 PM IST	10%	100%
✓	Graded Quiz: Model Evaluation and Refinement Graded Assignment	Passed	Jul 31 11:59 PM IST	10%	100%
✓	**Submit your Project and Review Others** Submit your assignment and review 2 peers' assignments to get your grade.			14%	100%
✓	Submit your assignment	Passed	Aug 2 11:59 PM IST		
✓	Review 2 peers' assignments.	7/2 reviewed	Aug 5 11:59 PM IST		
✓	Final Exam Graded Assignment	Passed	Aug 2 11:59 PM IST	36%	100%

Chapter 4:
Data Visualization with Python

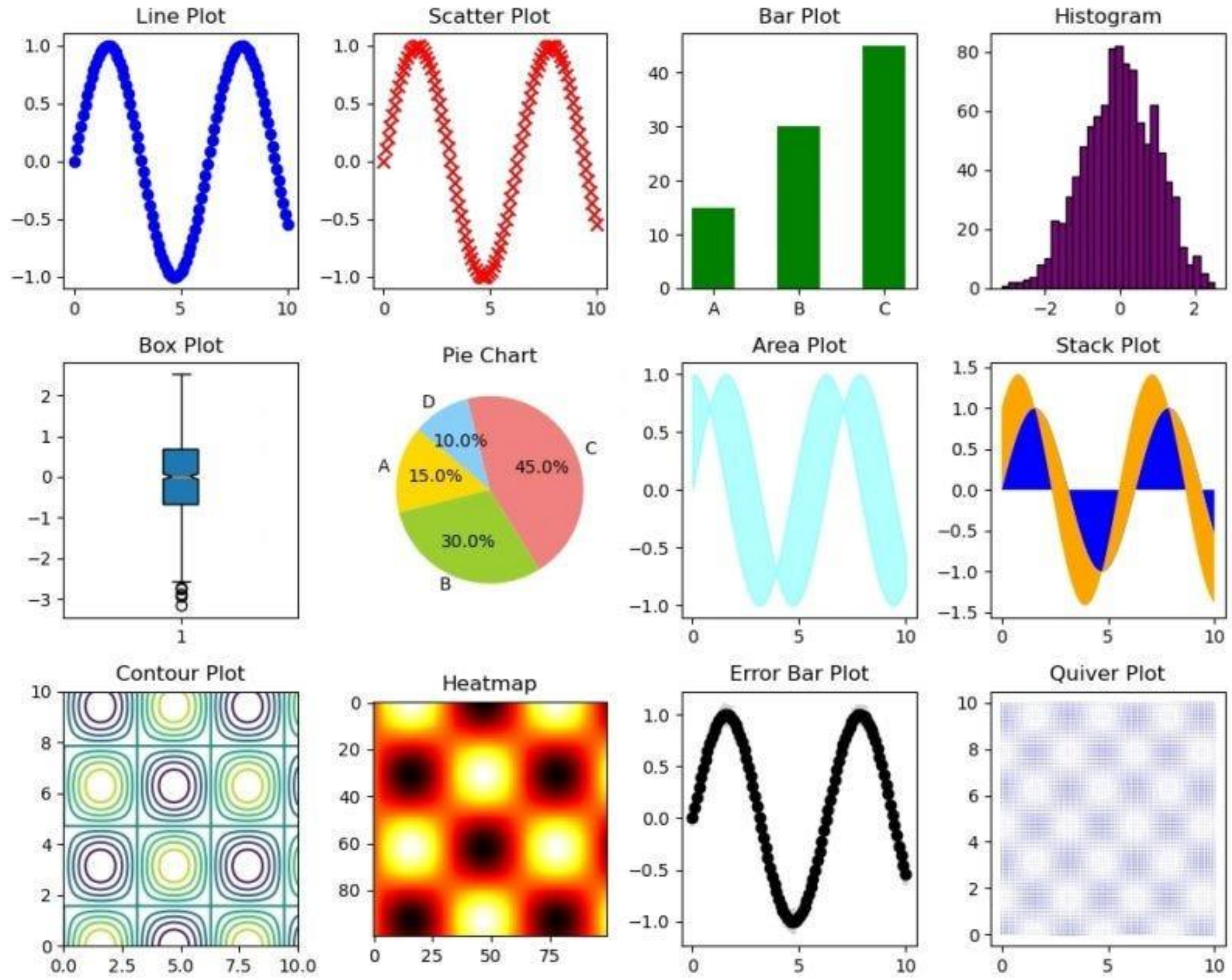
Bringing Data to Life
Python has become a powerful tool for data visualization, thanks to its rich ecosystem of libraries that offer a wide range of capabilities. Let's explore some of the most popular libraries and their applications:

Essential Libraries

- Matplotlib:
  - A fundamental plotting library, providing a wide range of plot types (line, scatter, bar, histogram, etc.).
  - Highly customizable and flexible for creating custom visualizations.
- Seaborn:
  - Built on top of Matplotlib, Seaborn offers a higher-level interface for creating visually appealing statistical graphics.
  - Provides themes, color palettes, and statistical plots specifically designed for data analysis.
- Plotly:
  - Creates interactive and customizable visualizations that can be embedded in web pages or dashboards.
  - Supports a wide range of plot types, including 3D plots and geographical maps.

Key Visualization Techniques

- Line Charts:
  - Show trends over time or relationships between variables.
  - Useful for visualizing time series data or comparing multiple variables.
- Bar Charts:
  - Display categorical data as bars.
  - Effective for comparing values across different categories.
- Scatter Plots:
  - Show the relationship between two numerical variables.
  - Useful for identifying patterns, correlations, and outliers.
- Histograms:
  - Show the distribution of a single numerical variable.
  - Helpful for understanding the frequency of different values.
- Box Plots:
  - Summarize the distribution of a numerical variable by showing the median, quartiles, and outliers.
  - Useful for comparing the distribution of data across different groups.
- Heatmaps:
  - Visualize two-dimensional data as a colored grid.
  - Effective for showing relationships between two variables or for clustering data.
- Geographical Plots:
  - Create maps to visualize data based on geographical locations.
  - Useful for analyzing spatial patterns and distributions.



Example: Visualizing Customer Churn Data

1. Load the data using Pandas.
2. Create a bar chart to visualize the distribution of customer churn.
3. Create a scatter plot to show the relationship between customer tenure and total spending.
4. Create a heatmap to visualize the correlation between different features.

Python's data visualization libraries offer a wide range of tools to help you explore, understand, and communicate your data findings effectively.

Skills Learnt

1.Python Programming:

Python is a high-level, general-purpose programming language widely used in data science. Its simplicity, versatility, and extensive libraries make it a popular choice for tasks like data analysis, machine learning, and web development.

Advantages:

- Easy to learn: A friendly language that's simple to understand.
- Lots of tools**:** Has many libraries to help you do data science tasks.
- Popular and helpful: Many people use it, so you can find lots of help and resources

2.Dashboards and Charts:

Dashboards and charts are visual tools used to present data in a concise and informative way. They can be used to monitor performance, track trends, and communicate insights to stakeholders.

Advantages:

- Shows data in a cool way: Makes data look interesting and easy to understand.
- Keeps track of things: Helps you monitor important information.

3.Dash:

Dash is a Python framework for building interactive web applications. It can be used to create custom dashboards and visualizations for data science projects.

Advantages:

- Builds fancy websites: Lets you create interactive websites that show data.
- Makes things easy: Simplifies the process of building web apps.

4.Data Visualization:

Data visualization is the process of representing data graphically to make it easier to understand and interpret. It involves using charts, graphs, and other visual techniques to communicate insights.

Advantages:

- Makes data look cool: Turns numbers into pictures that are easy to understand.
- Tells a story: Helps you explain your findings in a clear and interesting way

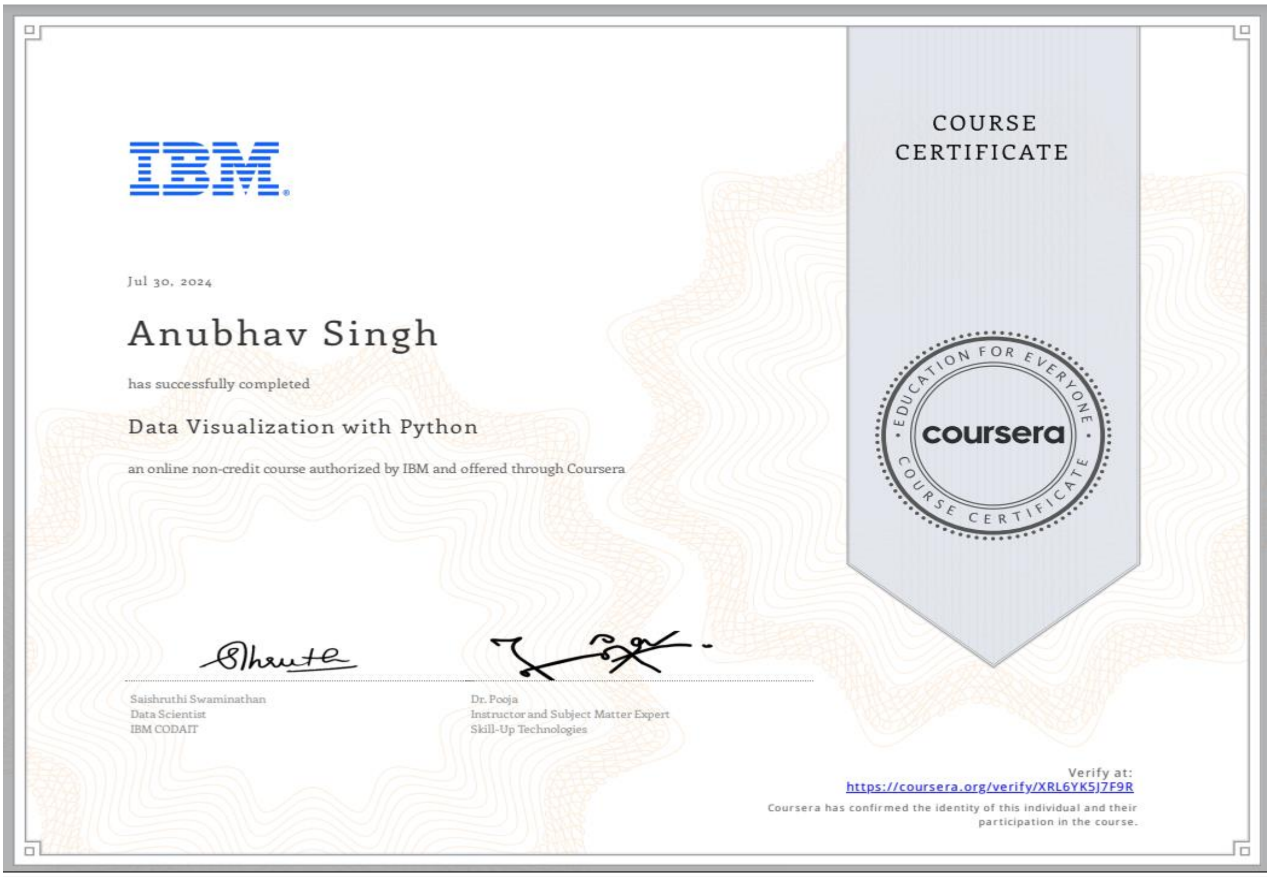
5.Matplotlib:

Matplotlib is a Python library for creating static, animated, and interactive visualizations. It is commonly used for data visualization tasks.

Advantages:

- Draws pictures: Helps you create graphs and charts to show your data.
- Customizable: Lets you make the pictures look exactly how you want.

Certificate:



COURSE CERTIFICATE

IBM

Jul 30, 2024

Anubhav Singh

has successfully completed

Data Visualization with Python

an online non-credit course authorized by IBM and offered through Coursera

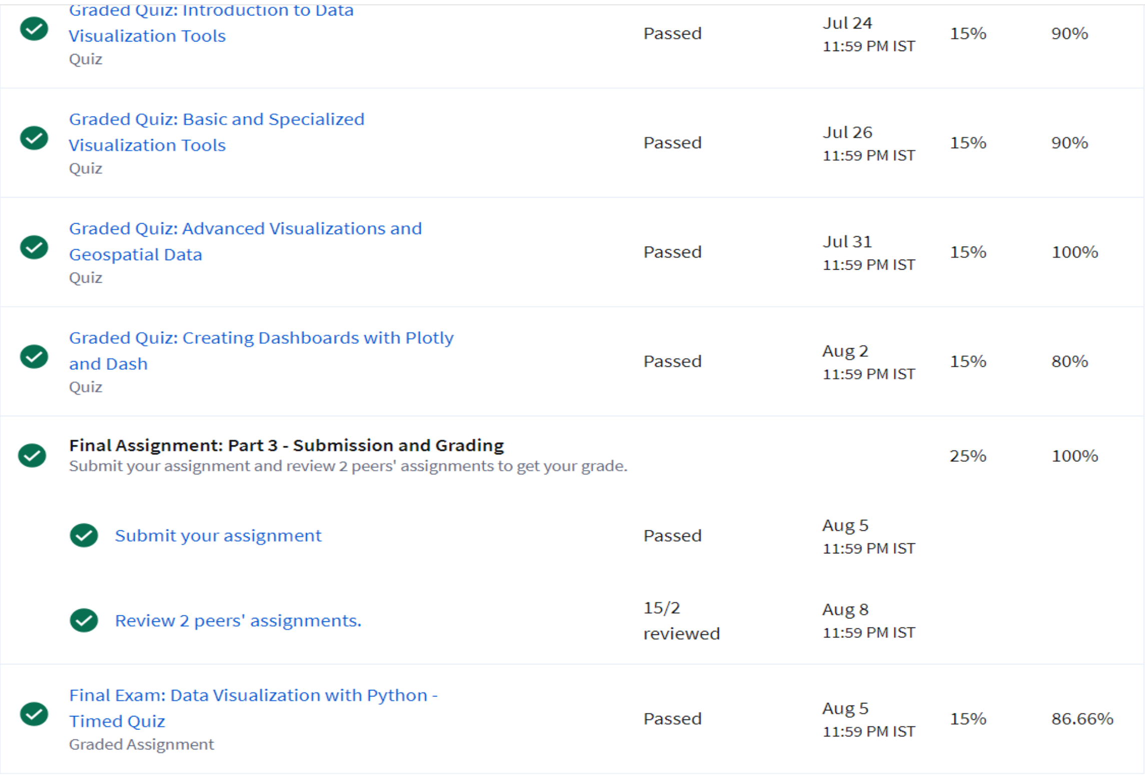
EDUCATION FOR EVERYONE · COURSE CERTIFICATE · coursera

Saishruthi Swaminathan
Data Scientist
IBM CODAIT

Dr. Pooja
Instructor and Subject Matter Expert
Skill-Up Technologies

Verify at:
https://coursera.org/verify/XRL6YK5J7F9R

Coursera has confirmed the identity of this individual and their participation in the course.



✓	Graded Quiz: Introduction to Data Visualization Tools Quiz	Passed	Jul 24 11:59 PM IST	15%	90%
✓	Graded Quiz: Basic and Specialized Visualization Tools Quiz	Passed	Jul 26 11:59 PM IST	15%	90%
✓	Graded Quiz: Advanced Visualizations and Geospatial Data Quiz	Passed	Jul 31 11:59 PM IST	15%	100%
✓	Graded Quiz: Creating Dashboards with Plotly and Dash Quiz	Passed	Aug 2 11:59 PM IST	15%	80%
✓	**Final Assignment: Part 3 - Submission and Grading** Submit your assignment and review 2 peers' assignments to get your grade.			25%	100%
✓	Submit your assignment	Passed	Aug 5 11:59 PM IST		
✓	Review 2 peers' assignments.	15/2 reviewed	Aug 8 11:59 PM IST		
✓	Final Exam: Data Visualization with Python - Timed Quiz Graded Assignment	Passed	Aug 5 11:59 PM IST	15%	86.66%

Chapter 5:
Applied Data Science Capstone

A Showcase of Skills

An applied data science capstone project is a culminating experience that allows you to demonstrate your mastery of data science concepts, tools, and techniques. It's a chance to apply your knowledge to a real-world problem and showcase your ability to:

- Identify and define a relevant problem.
- Gather and clean data.
- Perform exploratory data analysis.
- Build and evaluate predictive models.
- Communicate your findings effectively.

Here are some popular areas for applied data science capstone projects:

Healthcare

- Predicting patient outcomes: Develop models to predict patient outcomes based on medical records, demographics, and other factors.
- Disease outbreak detection: Use data to identify early signs of disease outbreaks and track their spread.
- Personalized medicine**:** Develop algorithms to recommend tailored treatment plans for individual patients.

Finance

- Fraud detection: Build models to detect fraudulent transactions in financial data.
- Risk assessment: Assess the risk of financial investments using data-driven techniques.
- Customer churn prediction: Predict which customers are likely to churn and take proactive steps to retain them.

Marketing

- Customer segmentation: Identify distinct groups of customers based on their behavior and characteristics.
- Recommendation systems: Build systems to recommend products or services to customers based on their preferences.
- Marketing campaign optimization: Optimize marketing campaigns for maximum ROI using data-driven insights.

E-commerce

- Product recommendation: Develop recommendation systems to suggest products to customers based on their purchase history and browsing behavior.
- Price optimization: Determine optimal pricing strategies for products using data analysis.
- Inventory management: Optimize inventory levels to minimize costs and avoid stockouts.

Social Media

- Sentiment analysis: Analyze social media data to understand public sentiment towards brands, products, or events.
- Community detection: Identify communities or groups of users within social networks.
- Fake news detection: Develop models to detect and prevent the spread of fake news.

When choosing a capstone project, consider the following factors:

- Your interests and passions: Choose a topic that genuinely interests you.
- The availability of data: Ensure that you have access to relevant and sufficient data.

- The complexity of the problem: Start with a manageable problem and gradually increase the complexity as you gain experience.
- The potential impact: Choose a project that has the potential to make a positive impact.

By completing a successful applied data science capstone project, you'll not only demonstrate your skills but also gain valuable experience that can help you land your dream job.

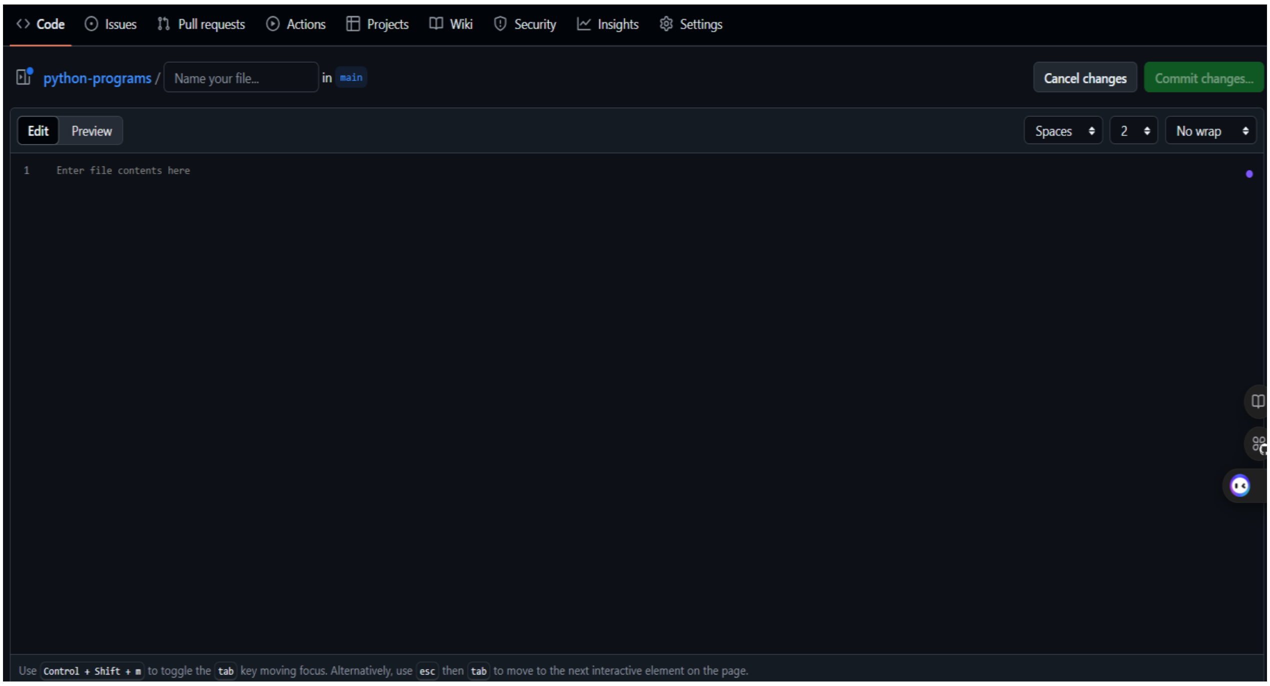
Skills Learnt

1.GitHub:

GitHub is a platform for version control and collaboration. It is commonly used by data scientists to manage their code, collaborate with others, and share their work.

Advantages:

- Stores your code: Keeps your work safe and organized.
- Collaborates with others**:** Lets you work with other people on projects.



```
1   Enter file contents here
```

Use `Control + Shift + m` to toggle the `tab` key moving focus. Alternatively, use `esc` then `tab` to move to the next interactive element on the page.

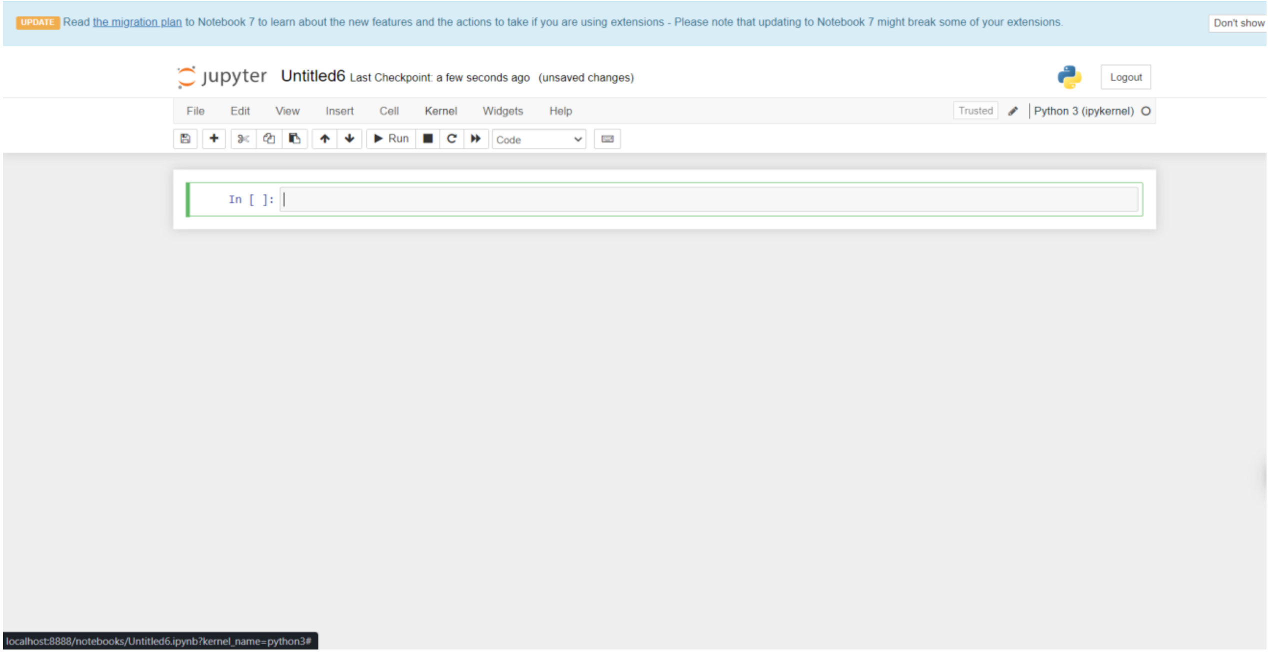
2.Jupyter Notebooks:

Jupyter Notebooks is an interactive environment for creating and sharing documents containing live code, equations, visualizations, and explanatory text. It is commonly used in data science for its interactive nature.

Advantages:

- Interactive playground: Lets you try things out and see the results immediately.

- Explains your work: Helps you document your code and explain what it does.



UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

jupyter Untitled6 Last Checkpoint: a few seconds ago (unsaved changes)

In [ ]:

localhost:8888/notebooks/Untitled6.ipynb?kernel_name=python3#

3.K-Means Clustering:

K-Means Clustering is an unsupervised machine learning algorithm used to group data points into clusters based on their similarity.

Advantages:

- Groups things together: Finds natural groups within your data.
- Used in many ways: Can be used for customer segmentation, image analysis, and more.

4.Methodology

Methodology refers to the systematic approach used to conduct research or analysis. In data science, it includes the steps involved in data collection, cleaning, analysis, and modeling.

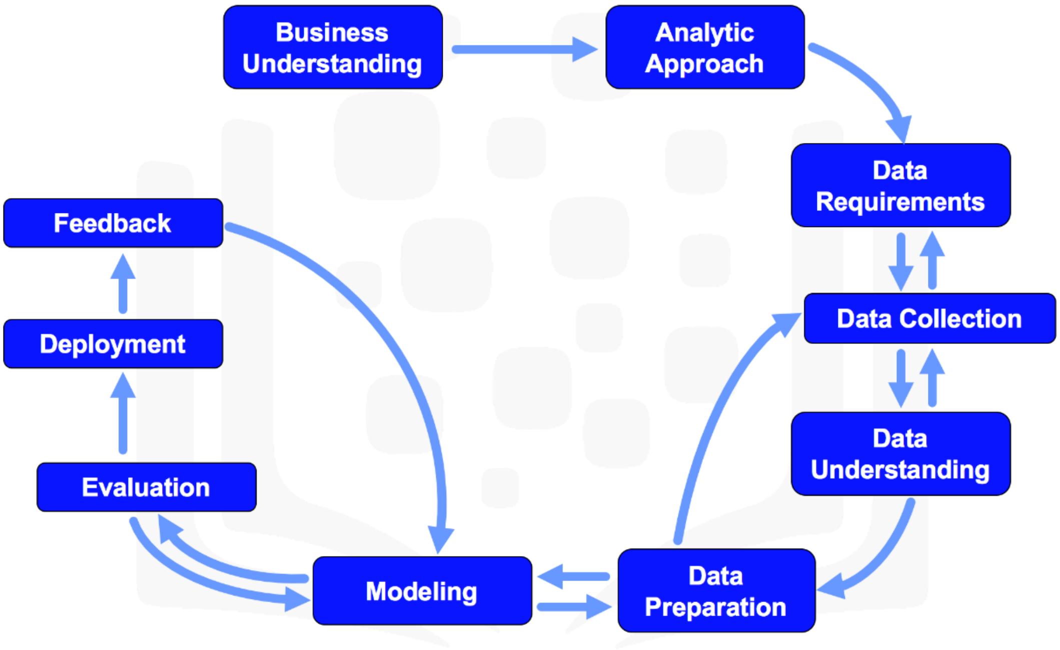
Advantages:

- Follows a plan: Provides a step-by-step guide for doing data science.
- Ensures quality: Helps you do things the right way and get reliable results.

5.Data Science Methodology

Data science methodology typically involves the following steps:

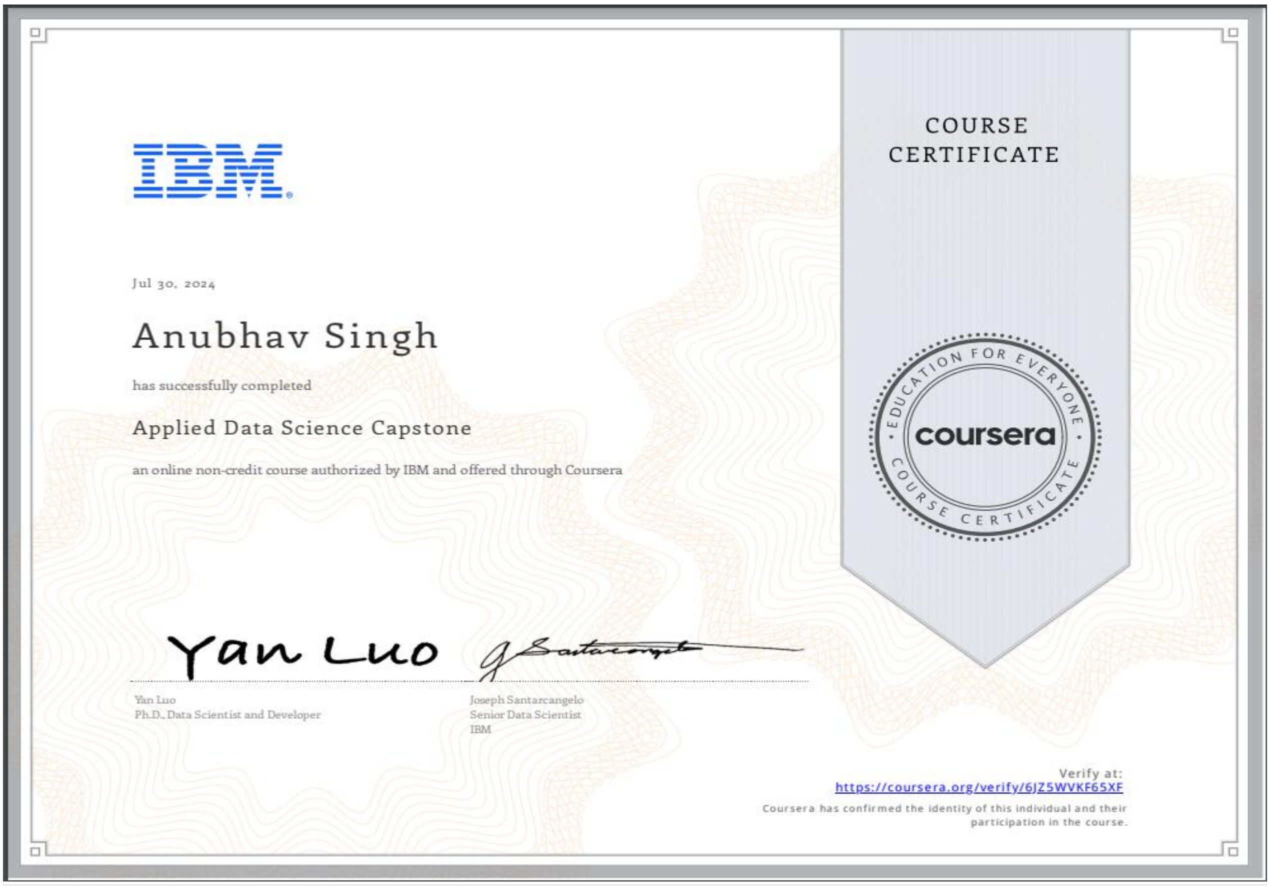
1. Problem Definition: Clearly define the problem to be solved.
2. Data Collection: Gather relevant data.
3. Data Preparation: Clean, preprocess, and explore the data.
4. Model Selection: Choose appropriate machine learning algorithms.
5. Model Training: Train the models on the data.
6. Evaluation: Evaluate the performance of the models.
7. Deployment: Deploy the best-performing model for use.



Advantages:

- The big picture: Outlines the whole process of data science from start to finish.
- Avoids mistakes: Helps you avoid common pitfalls and do things correctly.

Certificate:



IBM

COURSE CERTIFICATE

Jul 30, 2024

Anubhav Singh

has successfully completed

Applied Data Science Capstone

an online non-credit course authorized by IBM and offered through Coursera

EDUCATION FOR EVERYONE • COURSE CERTIFICATE

coursera

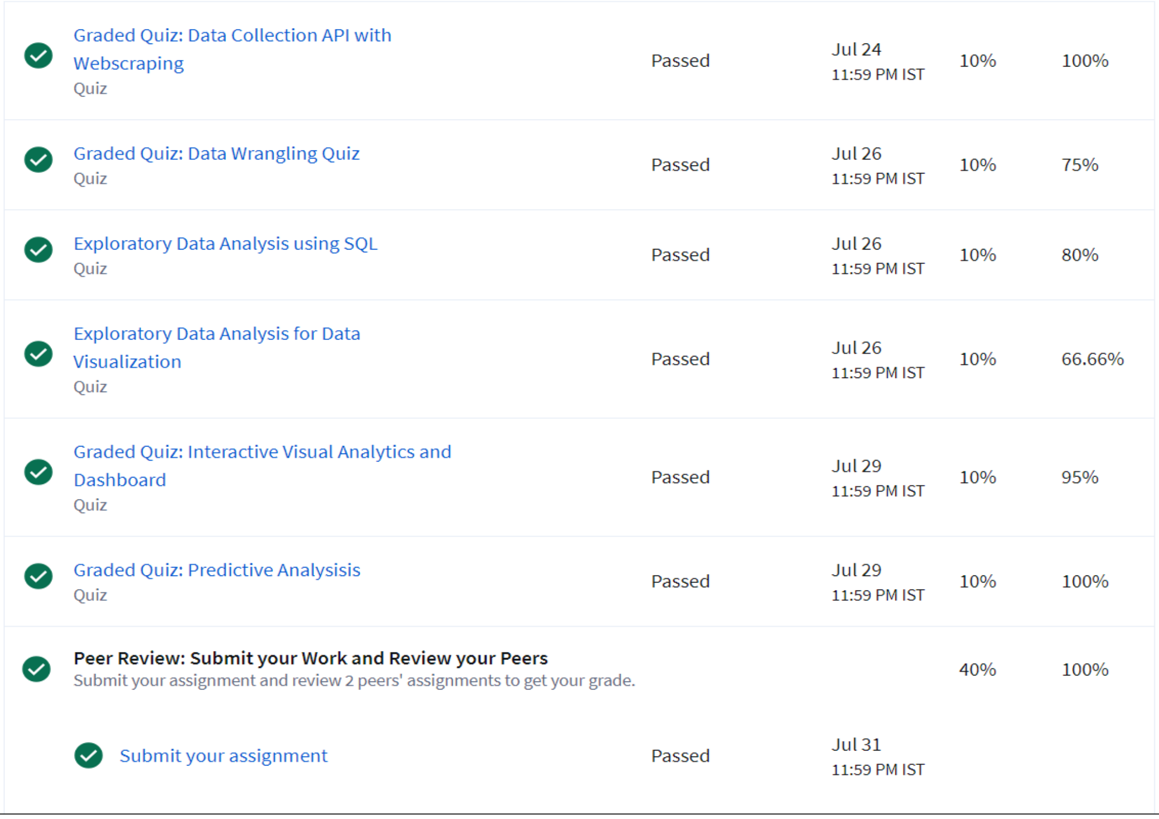
Yan Luo

Yan Luo
Ph.D., Data Scientist and Developer

Joseph Santarcangelo
Senior Data Scientist
IBM

Verify at:
https://coursera.org/verify/6JZ5WVKF65XF

Coursera has confirmed the identity of this individual and their participation in the course.



✓	Graded Quiz: Data Collection API with Webscraping Quiz	Passed	Jul 24 11:59 PM IST	10%	100%
✓	Graded Quiz: Data Wrangling Quiz Quiz	Passed	Jul 26 11:59 PM IST	10%	75%
✓	Exploratory Data Analysis using SQL Quiz	Passed	Jul 26 11:59 PM IST	10%	80%
✓	Exploratory Data Analysis for Data Visualization Quiz	Passed	Jul 26 11:59 PM IST	10%	66.66%
✓	Graded Quiz: Interactive Visual Analytics and Dashboard Quiz	Passed	Jul 29 11:59 PM IST	10%	95%
✓	Graded Quiz: Predictive Analysisis Quiz	Passed	Jul 29 11:59 PM IST	10%	100%
✓	**Peer Review: Submit your Work and Review your Peers** Submit your assignment and review 2 peers' assignments to get your grade.			40%	100%
	✓ Submit your assignment	Passed	Jul 31 11:59 PM IST		

Technology Learnt

In the IBM Applied Data Science course on Coursera, several key technologies and tools were used to help learners develop their data science skills. These include:

1. Model Selection
2. Data Analysis
3. Python Programming
4. Data Visualization
5. Predictive Modelling
6. Dashboards and Charts
7. dash
8. Matplotlib
9. Data Science
10. Pandas
11. Jupyter notebooks
12. Numpy
13. Github
14. K-Means Clustering
15. Methodology
16. Data Science Methodology

Learning Outcome:

The course equipped learners with a strong foundation in data science, including proficiency in Python, data manipulation techniques, data visualization, statistical analysis, and machine learning. I have successfully completed data science projects, demonstrating their ability to apply theoretical knowledge to real-world problems. They also gained experience with cloud-based tools, version control, and containerization. Additionally, the course developed their problem-solving, critical thinking, communication, and collaboration skills. Overall, learners were well-prepared to execute end-to-end data science projects and contribute to data-driven initiatives.

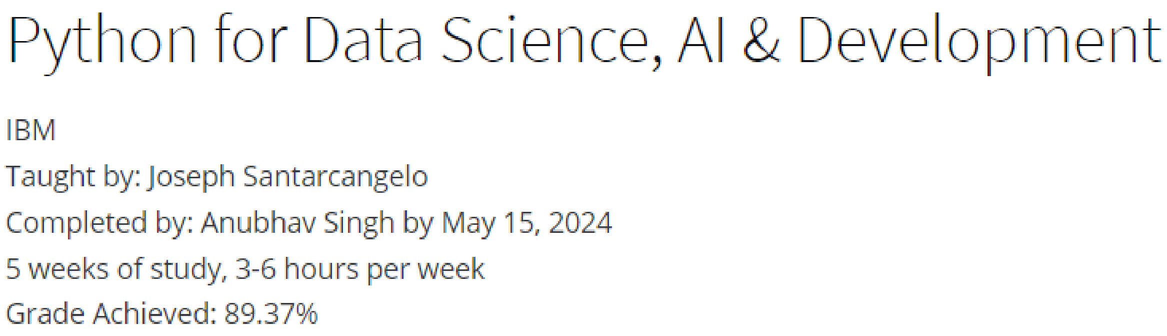
Throughout this project I have gained or I come to know the skills to

- Developed and honed skills for practical data science and machine learning problems: This indicates that I gained hands-on experience and expertise in applying data science techniques to real-world challenges.
- Learned Python: The specialization covered the fundamentals of Python programming, equipping learners with the necessary skills to work with data and build models.
- Performed data analysis: I practiced data analysis techniques, including cleaning, exploring, and preparing data for analysis.
- Created data visualizations using Python: I learned how to effectively communicate data insights through visualizations.
- Completed a Capstone project: The specialization culminated in a Capstone project, allowing me to apply my knowledge and skills to a comprehensive data science problem.

Overall, the specialization aimed to equip learners with the practical skills and knowledge required to excel in the field of data science and machine learning.

Grades and Faculty Taught by

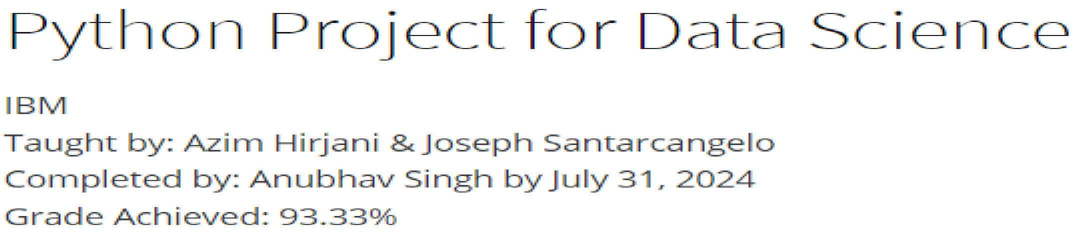
Chapter 1



Python for Data Science, AI & Development

IBM
Taught by: Joseph Santarcangelo
Completed by: Anubhav Singh by May 15, 2024
5 weeks of study, 3-6 hours per week
Grade Achieved: 89.37%

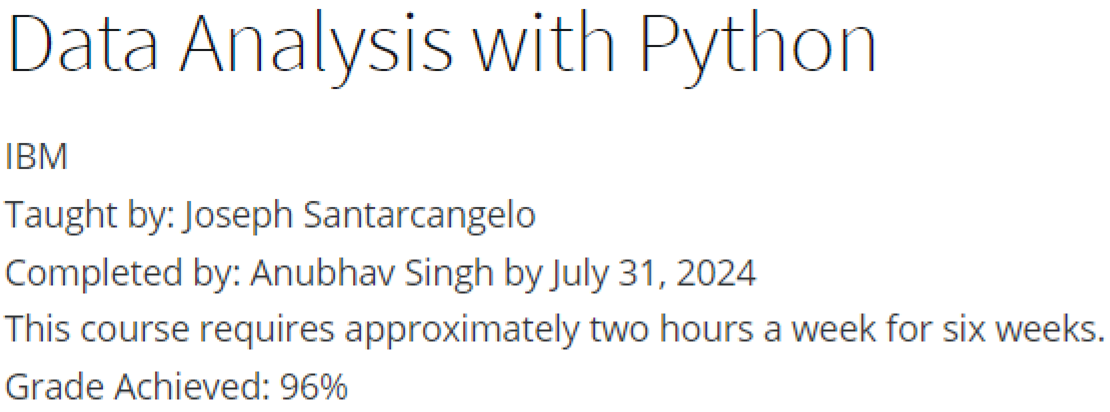
Chapter 2



Python Project for Data Science

IBM
Taught by: Azim Hirjani & Joseph Santarcangelo
Completed by: Anubhav Singh by July 31, 2024
Grade Achieved: 93.33%

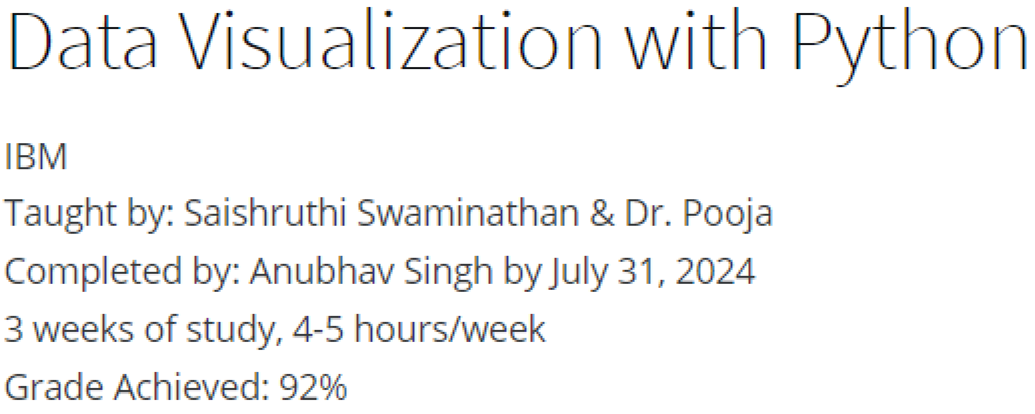
Chapter 3



Data Analysis with Python

IBM
Taught by: Joseph Santarcangelo
Completed by: Anubhav Singh by July 31, 2024
This course requires approximately two hours a week for six weeks.
Grade Achieved: 96%

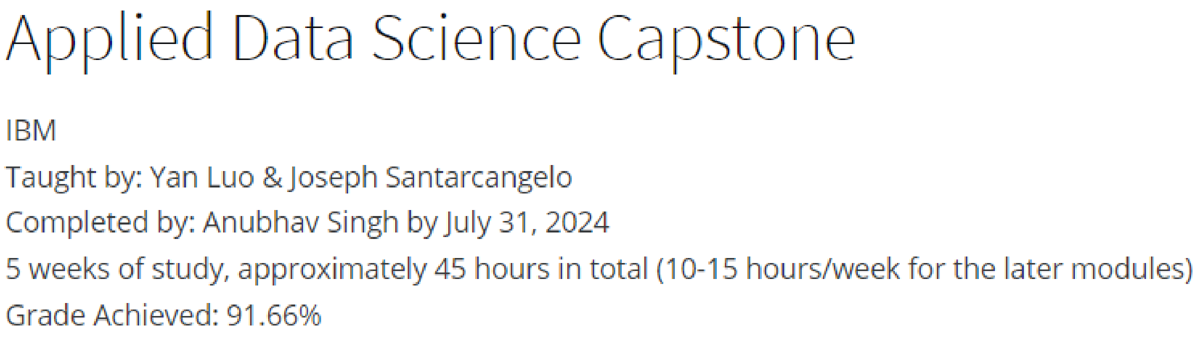
Chapter 4



Data Visualization with Python

IBM
Taught by: Saishruthi Swaminathan & Dr. Pooja
Completed by: Anubhav Singh by July 31, 2024
3 weeks of study, 4-5 hours/week
Grade Achieved: 92%

Chapter 5



Applied Data Science Capstone

IBM

Taught by: Yan Luo & Joseph Santarcangelo

Completed by: Anubhav Singh by July 31, 2024

5 weeks of study, approximately 45 hours in total (10-15 hours/week for the later modules)

Grade Achieved: 91.66%

Bibliography

1. IBM. (2024). *IBM Data Science Professional Certificate*. Retrieved from https://www.coursera.org/specializations/applied-data-science#courses
2. VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.
3. McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
5. Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90-95.
6. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). *Array Programming with NumPy*. Nature, 585(7825), 357-362.
7. Grus, J. (2019). *Data Science from Scratch: First Principles with Python* (2nd ed.). O'Reilly Media.
8. Seaborn Documentation. (2023). *Seaborn: Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org.
9. Docker Inc. (2023). *Docker Overview*. Retrieved from https://docs.docker.com/getstarted/overview/.