**Fall 2024 Data Science Bootcamp Syllabus**

Welcome to the NYU Tandon Fall 2024 Data Science Bootcamp! Over the next nine weeks, students will gain hands-on experience practicing data science concepts & questions in a supportive group environment to learn the fundamentals of answering technical interviewing questions. Here is more information about what to expect during the course of this bootcamp:

**Instructor Information:**

**Rohan Chopra**

datasciencebootcamp@nyu.edu

**Teaching Sections**

Two teaching sections are available and the times and dates are listed below. A different topic will be delivered each week. Participants must stay with their chosen teaching section.

**Section 1: Wednesdays (12:00PM-2:00PM EST)**

October 2, 9, 16, 23, 30

November 6, 13, 20

December 4

**Section 1 Zoom Link: https://nyu.zoom.us/meeting/register/tJUufuqopzojG9xP3Th19IEEmb6keFqK-xNQ**

**Section 2: Fridays (12:00PM-2:00PM EST)**

October 4, 11, 18, 25

November 1, 8, 15, 22

December 6

**Section 2 Zoom Link: https://nyu.zoom.us/meeting/register/tJ0udOygrTgiGdwwTer83hch2IiEiqaluTZg**

\***Please fill out this** [**teaching section feedback form**](https://forms.gle/JkCWnxAALKecxHcw9) **each week\***

**Bootcamp Topics & Timeline**

**Week 1 (10/2 & 10/4) -** Introduction to Bootcamp and Python Fundamentals

**Week 2 (10/9 & 10/11) -** Numpy and Pandas

**Week 3 (10/16 & 10/18) -** Exploratory Data Analysis and Data Visualization

**Week 4 (10/23 & 10/25) -** SQL Fundamentals

**Week 5 (10/30 & 11/1) -** Mid Program Project Presentations

**Week 6 (11/6 & 11/8)** - Machine Learning

**Week 7 (11/13 & 11/15)** - Machine Learning

**Week 8 (11/20 & 11/22)** - Machine Learning

**Week 9 (12/4 & 12/6)** - Final Project Presentations

**Pre-bootcamp resources**

The NYU Tandon Data Science bootcamp is open to all NYU Tandon students, no matter what your level of data science proficiency may be. Please consult the following resources to help you better understand technical interviews and practice in your free time during your participation in the bootcamp:

1. Python - <https://www.hackerrank.com/domains/python>

Note: We would recommend going through easy and medium questions. The objective should not be to finish all questions, but rather to explore as many topics as possible.

1. SQL -
   1. [Example SQL questions](https://www.nicksingh.com/posts/30-sql-and-database-design-questions-from-real-data-science-interviews)
   2. [Harder SQL Questions](https://towardsdatascience.com/5-common-sql-interview-problems-for-data-scientists-1bfa02d8bae6)
   3. [SQL Practice](https://www.w3schools.com/sql/)

1. Python Libraries -
   1. [Numpy](https://numpy.org/doc/stable/user/quickstart.html)
   2. [Pandas](https://pandas.pydata.org/docs/getting_started/intro_tutorials/)
   3. [Matplotlib](https://matplotlib.org/stable/tutorials/index.html)
   4. [Git](https://learngitbranching.js.org/)

1. Neural Network Frameworks-
   1. [Tensorflow](https://www.tensorflow.org/tutorials)
   2. [Pytorch](https://pytorch.org/docs/stable/index.html)

1. [Kaggle](https://www.kaggle.com/) is a good place to start your Data Science learning journey. You can find a lot of datasets and problem statements. Most of the time you will learn a lot from others' submissions.
2. [Another good resource](https://aman.ai/coursera-ml/) to brush up your ML basics and [Data Science](https://github.com/varunkashyapks/Books/blob/master/Data%20Science%20from%20Scratch.pdf) skills.
3. Interview Preparation : [Company based questions](https://datalemur.com/questions)

**How each session works**

The format of each session may vary depending on the specific topic being covered, but typically each session will comprise of the following -

1. **Lecture**: The instructor will present material on the topic for the session, using slides, jupyter notebooks and other visual aids to help explain concepts and examples.

2. **Hands-on exercises** - Each session will give participants an opportunity to apply the concepts and techniques learnt in the lecture through hands-on exercises. These may involve working with datasets, programming in python or using tools to analyze and visualize data.

3. **Q&A** - Participants are strongly encouraged to ask questions! The more questions you ask, the better your learning. We strongly believe that active participation from the audience will help make each session and the bootcamp a success. So ask away!

4. **Take-home exercises** - Please see below for more information regarding take-home exercises.

5. **Breaks-**Students will be given several breaks throughout each session.

Overall, the goal of each session is to provide a combination of theory and hands-on experience, with a focus on practical skills and techniques that participants can apply in real-world data science projects.

**Take-home assignments:**

Students will be given problems to work on at home each week related to the week's topic. Students may work individually to solve take home problems.

**Please use this** [**submission form**](https://docs.google.com/forms/d/e/1FAIpQLSeFTMoA8ydZHx1B6Cpgyr1dgXQnYBdP_5kEwlJcKMSK9_OwUQ/viewform?usp=sharing) **to share your work and results each week by Tuesdays at 11:59PM!**

**Industry Sessions**

As an added bonus to topics covered in teaching sessions, the Fall 2024 Data Science Bootcamp will feature 2 presentations from industry practitioners about presenting work as a software engineer and acing software engineering technical interviews. We will also have 1 panel discussion with recruiters about the hiring process for software engineers.

**Dates:**

[How To Present Your Work as a Data Scientist: 10/21, 1-2PM, Virtual](https://nyu.zoom.us/meeting/register/tJ0ucOyrrDosGdLFW_h09nz5kMZ2r7U7hoKV)

In this presentation, students will hear from a data science industry practitioner about tips and tricks for navigating professional data science projects, communicating technical concepts to non-technical audiences, and how to present their work effectively

[Acing Data Science Interviews, 11/4, 1-2PM, Virtual](https://nyu.zoom.us/meeting/register/tJAqc-iqrzwiGNNFQ9GdGxzkg6FHkPJHQ2yJ)

In this presentation, students will hear from a data science professional about how to navigate technical interviews related to data science, best practices for articulating their process, and how to showcase their results to help land their next great job or internship

[The Data Science Hiring Process, 11/11, 1-2PM, Virtual](https://nyu.zoom.us/meeting/register/tJYtdu-gqDsvE92N28tNMZc3TH6tyiKogGEL)

In this interactive panel discussion, students will have the opportunity to hear from technical recruiters and hiring professionals about the ins and outs of the recruiting process for data science roles, what companies look for in potential data scientists, and best practices to help them move forward in that process.

**Data Science Bootcamp GitHub**

Please review the following GitHub link for take-home assignments and other pertinent content throughout the course of the bootcamp.

<https://github.com/rohnnie/NYU-Data-Science-Bootcamp-Fall-2024>

**Project**

Students selected for the Fall 2024 Data Science Bootcamp will have the opportunity to work on a real-world project to apply the skills and knowledge gained during the bootcamp.

Projects are required as part of student participation in the Data Science bootcamp. Students may work with each other in groups of 4-5 students to complete the project and create their groups by Friday, October 11 @ 11:59PM.

Project groups will conduct their mid program presentations with industry professionals for 10 minutes (presentation and project judge Q&A) on either October 30 or November 1 (whichever date corresponds with your teaching section) in order to receive feedback and make revisions for final project presentations

Groups will conduct their final project presentations with industry professionals for 10 minutes (presentation and project judge Q&A) on December 4 or December 6 (whichever date corresponds with your teaching section). Prizes will be awarded to the winning group.

Students are encouraged to utilize weekly individual review time and participate in Slack discussions to communicate with the instructor about their project progress, ideas, or questions regularly.

**Project Judging Criteria**

Projects will be judged on the following criteria and prizes will be awarded to the top performing projects:

Using 1 (lowest)-3 (highest)

**Project Design**: How was the project designed and did students ask questions that helped move the project forward proactively?

● 1: Project not designed well; yields no tangible results or applicability

● 2: Project designed to move forward, but disjointed approach

● 3: Project designed to move forward in systematic way

**Addressing Stated Problem:** How well does the project solve the stated problem?

● 1: Results did not address stated problem at all

● 2: Results include missing details and inconclusive results

● 3: Results solves stated problem at face value

**Presentation and Demo:** How was the presentation?

● 1: Project presentation lacked any sort of depth and clarity, student(s) presenting did not clearly address issues from prompt

● 2: Project presentation included little detail and inconclusive results, difficult to understand

● 3: Project presentation addresses challenges stated in prompt clearly and concisely, easy for audience to follow and understand

**Project description(s): Please select one to work on:**

**1. Data Visualization with Air Quality Data**

Dataset: [Air Quality in India](https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data/data)

This project is intended to explore India's air pollution levels over the years using the provided dataset. The dataset represents a combined and cleaned version of the Historical Daily Ambient Air Quality Data .

Problem Statement:

The primary goal of this project is to analyze India's air pollution data and derive meaningful insights. Identify local trends in air quality, examine the correlation between air quality changes and shifts in environmental policies in India, and explore factors influencing air pollution levels.

Tasks:

● Explore the dataset to understand its structure and features.

● Identify key pollutants and their variations over the years.

● *Temporal Analysis:*

Analyze air quality trends over the years.

Identify any seasonal patterns or significant changes.

● *Regional Trends*:

Investigate regional variations in air quality.

● Explore differences in pollution levels between states and cities.

● Create visualizations to illustrate trends, patterns, and regional variations in air quality.

● Predictive Modeling: If feasible, consider building predictive models for air quality based on historical data.

● Evaluate model performance and explore its potential application.

Students are encouraged to draw connections between data-driven insights and potential policy implications. The project should foster a deeper understanding of the dynamics of air quality in India and its impact on public health and the environment.

Suggested Timeline :

➔ Week 1: Project introduction and Dataset Exploration

◆ Team formation (4-5 students).

◆ Introduction to the project, objectives, and Air quality dataset.

◆ Dataset acquisition and initial exploration.

➔ Week 2: Feature Exploration

◆ Data manipulation using Numpy and Pandas.

◆ Exploring the dataset.

➔ Week 3-4: EDA & Visualization

◆ Data manipulation using Numpy and Pandas.

◆ Exploring EDA and visualization techniques.

◆ Select relevant features and formulate the problem statement.

➔ Week 5: Mid-Program Presentation

◆ Present progress achieved till the EDA stage.

◆ Receiving feedback and suggestions for further analysis.

➔ Week 6: Initial Modeling

◆ Begin experimenting with different ML models

◆ Train Initial model and evaluate performance.

➔ Week 7: Hypothesis Testing

◆ Formulate hypothesis related to factors influencing air quality.

◆ Conduct hypothesis testing and statistical analysis.

➔ Week 8: Final Model Training and Interpretation

◆ Train final predictive models incorporating insights from previous analysis.

◆ Interpret results, summarize key insights

➔ Week 9: Final Presentation

◆ Present methodology, results and insights

**2. Sentiment Analysis of Social Media Content**

**Dataset**: [Social Media Sentiment Analysis Dataset](https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset/code)

This project aims to analyze user-generated content across various social media platforms to uncover sentiment trends and user behavior. The dataset offers a rich source of data, including text-based content, user sentiments, timestamps, hashtags, user engagement metrics (likes and retweets), and geographical information. By exploring this data, we can identify how emotions fluctuate over time, platform, and geography. We will also investigate the correlation between popular content and user engagement metrics.

Problem Statement:

The primary goal is to perform sentiment analysis, investigate temporal and geographical trends in user-generated content, and analyze platform-specific user behavior. The project will focus on identifying popular topics through hashtags, exploring engagement levels, and understanding regional differences in sentiment trends.

Tasks:

**Tasks**:

● **Dataset Exploration**:

○ Gain familiarity with the dataset by understanding its structure and key features such as sentiment, timestamps, and user engagement (likes and retweets).

● **Sentiment Analysis**:

○ Conduct sentiment analysis to classify the user-generated content into different categories such as surprise, excitement, admiration, etc.

○ Visualize the distribution of sentiments and examine the emotional landscape of social media platforms.

● **Temporal Analysis**:

○ Explore temporal patterns in user sentiment over time using the "Timestamp" column.

○ Identify recurring themes, seasonal variations, or any significant trends in the data.

● **User Engagement Insights**:

○ Analyze user engagement by studying the likes and retweets associated with posts.

○ Investigate how sentiment correlates with higher levels of user engagement.

● **Platform-Specific Analysis**:

○ Compare sentiment trends across various platforms using the "Platform" column.

○ Identify how emotions differ depending on the platform.

● **Hashtag and Topic Trends**:

○ Explore trending topics by analyzing the hashtags.

○ Investigate the relationship between hashtags and user engagement or sentiment.

● **Geographical Trends**:

○ Examine regional sentiment variations using the "Country" column.

○ Understand how social media content and sentiment differ across various regions.

● **Cross-Feature Analysis**:

○ Combine features (e.g., sentiment and hashtags, sentiment and platform) to uncover deeper insights about user behavior and content trends.

● **Predictive Modeling (Optional)**:

○ Explore the possibility of building predictive models to predict user engagement (likes/retweets) based on sentiment, hashtags, and platform.

○ Evaluate the performance of the model and explore its potential for predicting popular content.

Students are encouraged to draw connections between data-driven insights and potential policy implications. The project should foster a deeper understanding of the dynamics of air quality in India and its impact on public health and the environment.

Suggested Timeline :

➔ Week 1: Project introduction and Dataset Exploration

◆ Team formation (4-5 students).

◆ Introduction to the project, objectives, and Sentiment Analysis of Social Media dataset.

◆ Dataset acquisition and initial exploration.

➔ Week 2: Feature Exploration

◆ Data manipulation using Numpy and Pandas.

◆ Exploring the dataset.

➔ Week 3-4: EDA & Visualization

◆ Data manipulation using Numpy and Pandas.

◆ Exploring EDA and visualization techniques.

◆ Select relevant features and formulate the problem statement.

➔ Week 5: Mid-Program Presentation

◆ Present progress achieved till the EDA stage.

◆ Receiving feedback and suggestions for further analysis.

➔ Week 6: Initial Modeling

◆ Begin experimenting with different ML models

◆ Train Initial model and evaluate performance.

➔ Week 7: Hypothesis Testing

◆ Hypothesis testing and further data-driven analysis of factors influencing sentiment and engagement.

◆ Conduct hypothesis testing and statistical analysis.

➔ Week 8: Final Model Training and Interpretation

◆ Train final predictive models incorporating insights from previous analysis.

◆ Interpret results, summarize key insights

➔ Week 9: Final Presentation

◆ Present methodology, results and insights

**3. Music Recommendation System**

**Dataset:** [Spotify Dataset](https://www.kaggle.com/datasets/vatsalmavani/spotify-dataset)

This project aims to develop a music recommendation system that leverages user listening history and song characteristics from the Spotify dataset. The dataset includes detailed information about songs, such as audio features (tempo, danceability, energy, etc.), user playlists, user interactions, and other metadata. By analyzing this data, insights can be gained into user preferences and music trends, leading to personalized recommendations that enhance the listening experience.

Problem Statement:

The primary goal is to build a recommendation system that suggests songs to users based on their listening habits and preferences. The project will focus on understanding user behavior, exploring song characteristics, and applying machine learning techniques to create effective recommendations.

**Tasks**:

**Dataset Exploration:**

● Familiarize with the dataset by understanding its structure, key features, and relationships between songs, artists, and user interactions.

● Perform initial data cleaning to handle missing values and inconsistencies.

**Feature Analysis:**

● Analyze audio features such as tempo, energy, and danceability to understand their impact on user preferences.

● Explore user engagement metrics (e.g., play counts, likes) to identify popular songs and trends.

**Recommendation Model Development:**

● Implement collaborative filtering techniques to recommend songs based on user listening patterns.

● Explore content-based filtering methods using song characteristics to provide personalized recommendations.

**Evaluation of Recommendations:**

● Assess the effectiveness of the recommendation system using metrics such as precision, recall, and F1 score.

● Conduct user studies or surveys to gather feedback on the quality of recommendations.

**Visualization of Results:**

● Create visualizations to illustrate the relationships between audio features, user preferences, and song popularity.

● Present the results in a user-friendly format to showcase the effectiveness of the recommendation system.

**Integration with User Interface (Optional):**

● Develop a simple web or mobile interface to display recommended songs and allow users to provide feedback on recommendations.

● Incorporate features for users to explore new music based on their interests.

Students are encouraged to draw connections between data-driven insights and potential policy implications. The project should foster a deeper understanding of the dynamics of air quality in India and its impact on public health and the environment.

Suggested Timeline :

➔ Week 1: Project Introduction and Dataset Exploration

◆ Introduction to the project, objectives, and the Spotify dataset.

◆ Dataset acquisition and initial exploration.

➔ Week 2: Feature Exploration

◆ Data manipulation using Numpy and Pandas.

◆ In-depth exploration of key features and their relationships.

➔ Week 3-4: Recommendation Model Development

◆ Implement collaborative filtering and content-based filtering techniques.

◆ Begin training initial models.

➔ Week 5: Mid-Program Presentation

◆ Present progress achieved until the recommendation model development stage.

◆ Receive feedback and suggestions for further analysis.

➔ Week 6: Model Evaluation

◆ Evaluate the recommendation models using various metrics.

◆ Refine models based on evaluation results.

➔ Week 7: Visualization and Insights

◆ Create visualizations to illustrate key findings and recommendations.

◆ Summarize insights gained from data analysis.

➔ Week 8: Integration with User Interface (Optional)

◆ Develop a user interface to showcase the recommendation system.

◆ Allow users to explore and provide feedback on recommendations.

➔ Week 9: Final Presentation

◆ Present methodology, results, and insights.

◆ Discuss potential future enhancements and applications of the recommendation system.

# 

# Certificate of Completion Requirements

**Students will receive a certificate of completion if they meet the following participation criteria:**

● Attend at least 7 live sessions; attendance will be taken each session.

○ Attend at least 60 minutes of each teaching section to receive credit for attendance.

○ If a student must arrive late to a session or leave early, they must receive prior approval by emailing us at datasciencebootcamp@nyu.edu

○ Leaving a session early or coming late without prior approval will result in the student being marked absent.

● Submit weekly **take-home problem** solutions via submission form.

● Submit weekly **teaching section feedback form**.

● Students must **form project groups** (5 members) and communicate the group members' names and the prompt chosen to the instructor by **Friday, October 11 @ 11:59 PM**.

● Groups are highly encouraged to perform their **mid-program project presentation** for 10 minutes on **October 30** or **November 1** (whichever date corresponds with your teaching section) to receive feedback and make revisions for the final project presentations.

● Groups must conduct their **final project presentation** on **December 4 or December 6** (whichever date corresponds with your teaching section).

● All group members must speak and present during the project demonstrations.

**Slack Channel**

The NYU Tandon Data Science bootcamp has a dedicated Slack channel and students are highly encouraged to participate in discussions with the bootcamp instructor and each other in order to have questions related to weekly topics, optional take home assignments, and projects answered.

The instructor will monitor the Slack channel during the following periods:

Monday: 1pm - 3pm

Tuesday: 12pm - 4pm

Friday: 11am - 2pm

**Slack Channel Link https://nyudatascienc-f1l2595.slack.com/archives/C07JK4X79QF**

All sessions are virtual and students who register will receive calendar invites with unique Zoom links.

**One on One Technical Interviewing Coaching**

Instructors will be available for scheduled 30 minute coaching sessions to assist students with technical interviewing questions and practice. Please see the following Calendly link for availability:

calendly.com/rc4920-nyu

**Students must be logged in using their NYU credentials in order for us to keep track of individual participation during the course of the bootcamp.**

Please email datasciencebootcamp@nyu.edu if you have any pressing questions or comments and we look forward to your participation!