Spring 2025 Data Science Bootcamp Syllabus

Welcome to the NYU Tandon Spring 2025 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

All sessions are virtual, and students who register will receive calendar invites with unique Zoom links. Students must log in using their NYU credentials for us to keep track of individual participation during the Boot camp.

Please email datasciencebootcamp@nyu.edu if you have any questions or comments.

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)

February 26
March 5, 12, 19
April 2, 9, 16, 23, 30

Section 1: Zoom Link

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

February 28
March 7, 14, 21
April 4, 11, 18, 25
May 2

Section 2: Zoom Link

Please fill out this teaching section feedback form each week

Bootcamp Topics & Timeline

Week 1 (02/26 & 02/28) - Introduction to Bootcamp and Python Fundamentals

Week 2 (03/05 & 03/07) - Numpy and Pandas

Week 3 (03/12 & 03/14) - Exploratory Data Analysis and Data Visualization

Week 4 (03/19 & 03/21) - SQL Fundamentals

Week 5 (04/02 & 04/04) - Mid Program Project Presentations

Week 6 (04/09 & 04/11) - Machine Learning

Week 7 (04/16 & 04/18) - Machine Learning

Week 8 (04/23 & 04/25) - Machine Learning

Week 9 (04/30 & 05/02) - 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.

2. SQL -
   1. Example SQL questions
   2. Harder SQL Questions
   3. SQL Practice

3. Python Libraries -
   1. Numpy
   2. Pandas
   3. Matplotlib
   4. Git

4. Neural Network Frameworks-
    1. Tensorflow
    2. Pytorch

5. Kaggle 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.

6. Another good resource to brush up your ML basics and Data Science skills.

7. Interview Preparation : Company based 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 to share your work and results each week by Wednesday/Friday at 11:59PM!

Industry Sessions

As an added bonus to topics covered in teaching sessions, the Spring 2025 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: 03/10, 1-2PM, Virtual
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

The Data Science Hiring Process, 04/14, 1-2PM, Virtual
In this presentation, students will hear from a data science professional about the ins and outs of the recruiting process for data science roles, what companies look for in potential data scientists, navigating technical interviews related to data science, and how to showcase their results to help land their next great job or internship

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-Spring-2025

Project

Students selected for the Spring 2025 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 4-6 students to complete the project and create their groups by **Friday, March 7 @ 11:59PM**.

Project groups will conduct their mid program presentations with industry professionals for 10 minutes (presentation and project judge Q&A) on either April 02 or April 04 (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 April 30 or May 02 (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

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:

- **Dataset Exploration:**
    - 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 hypotheses 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

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:

- **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 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 hypotheses 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

3. Movie Recommendation System

Dataset: TMDB Movie Dataset

Recommendation systems have become an integral part of how we discover and enjoy content, from movies and music to products and services. They help personalize the user experience by suggesting items that match individual preferences. A movie recommendation system, for example, identifies patterns in what people watch and uses that information to suggest films they're likely to enjoy.

Problem Statement:

The goal is to design and implement a movie recommendation system that suggests movies to users based on their viewing history, preferences, and the attributes of movies. This project focuses on exploring different recommendation approaches, including demographic filtering, content-based filtering, and collaborative filtering, to create a robust and effective recommendation engine.

Tasks:

- **Dataset Exploration:**
    - Gain familiarity with the TMDB 5000 Movie Dataset by analyzing its structure, key features, and the relationships between movies, genres, and user interactions.
    - Perform data cleaning to address missing values, inconsistencies, and duplicates.
- **Feature Analysis:**
    - Examine movie metadata such as genres, directors, cast, and descriptions to understand their role in influencing user preferences.
    - Explore user interaction metrics (e.g., ratings, watch counts) to identify patterns in movie popularity and user preferences.
- **Recommendation Model Development:**
    - Demographic Filtering**:** Implement a baseline system that recommends popular and highly rated movies to all users.
    - Content-Based Filtering: Develop a recommendation model that uses movie metadata (e.g., genres, actors, directors) to suggest similar movies based on a user's viewing history.
    - Collaborative Filtering: Build a model that predicts user preferences by analyzing interactions and matching users with similar tastes.
- **Evaluation of Recommendations:**
    - Measure the effectiveness of the recommendation system using metrics such as precision, recall, F1 score, and mean average precision (MAP).
    - If possible, collect user feedback or simulate user studies to evaluate the relevance of the recommendations.
- **Visualization of Results:**
    - Create visualizations to depict relationships between user preferences, movie features, and recommendation performance.
    - Showcase trends in movie popularity, genre preferences, and user interaction patterns.
- **Integration with User Interface (Optional):**
    - Develop a simple web or mobile interface to display recommended movies and allow users to provide feedback.
    - Include features for users to explore movies by genres, popularity, or personalized recommendations.

Suggested Timeline :

- ➔ **Week 1:** Project Introduction and Dataset Exploration
    - ◆ Introduction to recommendation systems and the TMDB 5000 dataset.
    - ◆ Initial dataset acquisition and exploration.
- ➔ **Week 2:** Feature Exploration
    - ◆ Data manipulation using libraries like Pandas and Numpy.
    - ◆ Exploration of movie metadata and user interaction features.
- ➔ **Weeks 3-4:** Recommendation Model Development
    - ◆ Implement demographic, content-based, and collaborative filtering models.
    - ◆ Train and optimize initial recommendation models.
- ➔ **Week 5:** Mid-Program Presentation
    - ◆ Present progress on dataset exploration, feature analysis, and initial models.
    - ◆ Gather feedback for further refinements.
- ➔ **Week 6:** Model Evaluation
    - ◆ Evaluate recommendation models using quantitative metrics.
    - ◆ Refine models based on evaluation results.
- ➔ **Week 7:** Visualization and Insights
    - ◆ Create visualizations to highlight key findings and system performance.
    - ◆ Summarize insights gained from the analysis.
- ➔ **Week 8:** Integration with User Interface (Optional)
    - ◆ Develop a user-friendly interface to showcase the recommendations.
    - ◆ Allow users to interact with and provide feedback on the system.
- ➔ **Week 9:** Final Presentation
    - ◆ Present the methodology, results, and insights from the project.
    - ◆ Discuss future improvements and applications of the recommendation system.

Outcomes:

This project will provide hands-on experience with building recommendation systems, enhancing understanding of machine learning, feature engineering, and data visualization techniques. The developed system will serve as a foundation for future improvements and real-world applications.

Digital Credential/Program Completion Requirements

Students will be issued a verifiable digital credential badge via Accredible (shareable on LinkedIn and other platforms) provided that they meet the following participation criteria:

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

    - Attendance will be monitored based on the duration logged on during each teaching section.
    - 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** (4-6 members) and communicate the group members' names and the prompt chosen to the instructor by **Friday, March 7 @ 11:59PM**.

- Groups must conduct their **mid-program project presentation** for 10 minutes on **April 2** or **April 4** (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 **April 30** or **May 2** (whichever date corresponds with your teaching section).

- It is required that all group participants be present during the presentation; it is up to the group to determine who the presenters are.

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: 11am - 3pm
Friday: 11am - 2pm

Slack Channel Link

Join Slack Channel

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.