

## Project 4 Overview

---

For Project 4, you will work with your group to solve, analyze, or visualize a problem using machine learning (ML) with the other technologies we've learned. Here are the specific requirements:

1. Find a problem worth solving, analyzing, or visualizing.
2. Use machine learning (ML) with the technologies we've learned.
3. You must use Scikit-learn and/or another machine learning library.
4. Your project must be powered by a dataset with at least 100 records.
5. You must use at least two of the following:
  - Python Pandas
  - Python Matplotlib
  - HTML/CSS/Bootstrap
  - JavaScript Plotly
  - JavaScript Leaflet
  - SQL Database
  - MongoDB Database
  - Google Cloud SQL
  - Amazon AWS
  - Tableau

For this project, you can focus your efforts within a specific industry, as detailed in the following examples.

### Finance

- Create an algorithm that analyzes credit scores and predicts consumer personal-loan eligibility.
- Using natural language processing, create a chatbot to perform simple tasks and help users find information.
- Train an algorithm to analyze consumer spending and predict trends.
- Train an image classifier to assess property value, which could then be used to calculate insurance quotes.

### Healthcare


- Train an algorithm to recognize disease symptoms and predict if a patient is at risk.
- Train an image classifier to recognize anomalies, such as suspicious vs healthy areas of skin.
- Using natural language processing, create a chatbot that will help connect patients with doctors.
- Create an algorithm to analyze patient history and predict the likelihood of inherited illness.

## Custom

We've only specified healthcare and finance, but any industry can benefit from machine learning. Consider preparing a 15-minute data deep dive or infrastructure review that shows machine learning in the context of what we've already learned.


- Create a front-end interface that maps to an API to "smarten" the algorithm.
- Perform a deep dive on existing data using machine learning.
- Create a visualization that continues to learn where clusters lie based on ML (use Leaflet or Plotly to change the visualization).
- Create an idea using mock data, and simulate how machine learning might be used.
- Create an analysis of existing data to make a prediction, classification, or regression.

## Considerations

- **Remember to closely monitor any AWS resources that you choose to use.** It's crucial that you clean up and stop, or shut down any AWS resources to avoid accruing additional costs.
  - [AWS Billing Guide](https://static.bc-edx.com/data/dl-1-2/m23/supplemental/AWS_check_billing.pdf)  ([https://static.bc-edx.com/data/dl-1-2/m23/supplemental/AWS\\_check\\_billing.pdf](https://static.bc-edx.com/data/dl-1-2/m23/supplemental/AWS_check_billing.pdf))

## Working with Your Group

When working on an online group project, it's crucial to meet with your group and communicate regularly. Plan for significant collaboration time outside of class. The following tips can help you make the most of your time:

- Decide how you're going to communicate with your group members when you begin. Create a Slack channel, exchange phone numbers, and ensure that the group knows each group member's available working hours.
- Set up an agile project by using [GitHub Projects](https://docs.github.com/en/free-pro-team@latest/github/managing-your-work-on-github/managing-project-boards)  (<https://docs.github.com/en/free-pro-team@latest/github/managing-your-work-on-github/managing-project-boards>) so that your group can track tasks.
- Create internal milestones to ensure that your group is on track. Set due dates for these milestones so that you have a timeline for completing the project. Some of these milestones might include:
  - Project ideation
  - Data fetching/API integration
  - Data analysis
  - Building the ML model
  - Testing

- Creating documentation
- Creating the presentation

Since this is a two-week project, make sure that you have done at least half of your project by the end of the first week to stay on track.

Although you will divide the work among the group members, it's essential to collaborate and communicate while working on different parts of the project. Be sure to check in with your teammates regularly and offer support.

## Support and Resources

Your instructional team will provide support during classes and office hours. You will also have access to learning assistants and tutors to help you with topics as needed. Make sure to take advantage of these resources as you collaborate with your group on this first project.

## Project Guidelines

The following project guidelines focus on teamwork, your project proposal, data sources, and data cleanup and analysis.

## Collaborating with Your Team

Remember that these projects are a group effort. The experience of close collaboration will create better project outcomes and help you in your future careers. Specifically, you'll learn collaborative workflows that will enable you to approach and solve complex problems. Working in groups allows you to work smart and dream big. Take advantage!

## Project Proposal

Before you start writing any code, your group should outline the scope and purpose of your project. This will help provide direction and safeguard against **scope creep** (the tendency for projects to become more complex after work begins).

The proposal is essentially a brief summary of your interests and intent. Be sure to include the following details:








- The kind of data you'd like to work with and the field you're interested in (finance, healthcare surveys, etc.)
- The questions you'll ask of the data
- Possible source for the data

Use the following example for guidance:

The aim of our project is to uncover patterns in credit card fraud. We'll examine relationships between transaction types and location, purchase prices and times of day, purchase trends over the course of a year, and other related relationships derived from the data.

## Finding Data

Once your group has written a proposal, it's time to start searching for data. We recommend the following curated sources of high-quality data:

- [data.world](https://www.data.world)  (<https://www.data.world>)
- [Kaggle](https://www.kaggle.com)  (<https://www.kaggle.com>)
- [Data.gov](https://www.data.gov)  (<https://www.data.gov>)
- [Awesome Public Datasets](https://github.com/awesomedata/awesome-public-datasets)  (<https://github.com/awesomedata/awesome-public-datasets>)
- [Public-APIs](https://github.com/n0shake/Public-APIs)  (<https://github.com/n0shake/Public-APIs>)
- [Awesome API](https://github.com/Kikobeats/awesome-api)  (<https://github.com/Kikobeats/awesome-api>)
- [Medium API List](https://benjamin-libor.medium.com/a-curated-collection-of-over-150-apis-to-build-great-products-fdcfa0f361bc)  (<https://benjamin-libor.medium.com/a-curated-collection-of-over-150-apis-to-build-great-products-fdcfa0f361bc>)

### IMPORTANT

Whenever you use a dataset or create a new dataset based on other sources (such as existing datasets or information scraped from websites), make sure to use the following guidelines:

1. Check for copyright protections, and make sure that the way you plan to use this dataset is within the bounds of fair use.
2. Document how you intend to use this dataset now and in the future. Find any licenses or terms of use associated with the dataset, and review them to confirm that your intended use is in compliance.
3. Investigate how the dataset was collected. Identify any indicators that the data was obtained from a source that the compilers were not authorized to access.

You'll likely have to adjust your project plan as you explore the available data. That's okay! This is all part of the process. Just make sure that everyone in the group is aligned on the project's goals as you make changes.

Make sure that your datasets are not too large for your personal computer. Big datasets are difficult to manage locally, so consider using data subsets or different datasets altogether.

## Data Cleanup and Analysis

Now that you've picked your data, it's time to tackle development and analysis. This is where the fun starts!

The analysis process can be broken into two broad phases: (1) exploration and cleanup, and (2) analysis.

As you've learned, you'll need to explore, clean, and reformat your data before you can begin answering your research questions. We recommend keeping track of these exploration and cleanup steps in a dedicated Jupyter notebook to keep you organized and make it easier to present your work later.

After you've cleaned your data and are ready to start crunching numbers, you should track your work in a Jupyter notebook dedicated specifically to analysis. We recommend focusing your analysis on multiple techniques, such as aggregation, correlation, comparison, summary statistics, sentiment analysis, and time-series analysis. Don't forget to include plots during both the exploration and analysis phases. Creating plots along the way can reveal insights and

interesting trends in the data that you might not notice if you wait until you're preparing for your presentation. Presentation requirements will be further explained in the next module.

## Presentation Day

It's crucial that you find time to rehearse before presentation day.

On the day of your presentation, each member of your group is required to submit the URL of your GitHub repository for grading.

### NOTE

Projects are requirements for graduation. While you are allowed to miss up to two Challenge assignments and still earn your certificate, projects cannot be skipped.

© 2025 edX Boot Camps LLC