

Project 1 Overview

For Project 1, you will work with your group to find and analyze a dataset of your choice.

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



Finance

Exploratory data analysis is used by many individuals within the finance industry, including investment banking professionals, private equity analysts, lending analysts, financial administrators, and real estate professionals.

Exploratory data analysis is used for the following tasks in the financial sector:

- Identifying deals
- Analyzing private equity markets
- Researching arbitrage opportunities
- Evaluating liquidity
- Keeping up to date with finance and refinance trends

Project Examples

- **Equity Trading:** While working for a large equity-trading company, you're tasked with researching a client's portfolio. Your client wants to invest in telecom stocks and needs expert analysis to make the right decision. Using the [Nasdaq Data API](https://data.nasdaq.com/tools/api)  (<https://data.nasdaq.com/tools/api>), pull a year's worth of trading data for the major cell phone providers: AT&T, T-Mobile, and Verizon. Which stocks are trending upward? Which are trending downward? Based on the data, what would you recommend to your client?
- **New-Car Loan Analysis:** People have been financing higher car values over longer amounts of time. Explore what is driving this trend. Search for answers by using data collected from the [Federal Reserve Economic Data \(FRED\)](https://fred.stlouisfed.org/series/DTCTLVENANQ)  (<https://fred.stlouisfed.org/series/DTCTLVENANQ>). What other questions can you answer with the data? What do your results suggest about the time value of money? What about the impact of these loans as time goes on?



Healthcare

Exploratory data analysis is used by many individuals within the healthcare field, including clinical data analysts, pharmaceutical testers, healthcare-economics researchers, senior policy analysts, compliance operations analysts, and public health informatics scientists.

Exploratory data analysis is important for understanding the following healthcare considerations:

- Predicting and diagnosing illnesses
- Improving patient safety
- Reducing time to diagnosis
- Increasing our understanding of disease risks and causes
- Developing stronger prevention strategies

Project Examples

- Mental Health in Tech: People working in tech are often at their desks for extended amounts of time. Explore how this trend correlates with mental health. Examine the [data collected through surveys](https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey)  (<https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey>) and search for trends. Find out if there is a link between mental health and companies that offer wellness programs. What do the results show you about the state of mental health in tech? Can you suggest steps that companies can take to help their employees?
- Personal Fitness Analyst: Research whether working out helps a person become more active overall. Use data collected by the [Samsung Health application](https://www.kaggle.com/datasets/aroojanwarkhan/fitness-data-trends)  (<https://www.kaggle.com/datasets/aroojanwarkhan/fitness-data-trends>) to uncover relevant trends. What do the results tell you about individuals using this app? Have their lifestyles become more active? Less? Remained the same?




Custom

We've only specified healthcare and finance, but any industry can benefit from exploratory data analysis.

The following professionals also use data and can benefit from exploratory data analysis:


- Natural and environmental scientists
- Marketing professionals
- Information security analysts
- Business intelligence analysts

Project Examples

- Private Investigator: Use [aggregate crime data](https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i)  (<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>) from different police precincts in a city to uncover criminal activity patterns. Consider that [most crime in New York City takes place in the summer](https://www.nydailynews.com/new-york/nyc-crime/daily-news-analysis-reveals-crime-rankings-city-subway-system-article-1.1836918)  (<https://www.nydailynews.com/new-york/nyc-crime/daily-news-analysis-reveals-crime-rankings-city-subway-system-article-1.1836918>). Find out if you are able to uncover similar patterns in your city. What do your results suggest about how police should plan their patrols? What do your results suggest about how law enforcement resources should be distributed over the calendar year?
- Uber Rides and Weather: No one likes to walk in subzero temperatures or scorching heat. Do people use Uber more when the weather is uncomfortable? Using [Uber ride data from Kaggle](https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city)  (<https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city>) and data from a weather API, find out if people take Uber more during summer and winter months, and if there are relationships between daily temperature and ride frequency. What do the results tell you about surge-pricing strategies and commuter habits?

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
 - 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 in order 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.

Requirements

Completed Analysis Uploaded to GitHub (20 points)

- Final data analysis contains ample and complete information in README file (10 points)
- Final repository is acceptable for professional quality presentation (10 points)

Visualizations (20 points)

- 6–8 visualizations of data (at least two per question) (10 points)
- Clear and accurate labeling of images (5 points)
- Visualizations supported with ample and precise explanation (5 points)

Analysis and Conclusion (20 points)

- Write-up summarizes major findings and implications at a professional level (5 points)

- Each question in the project proposal is answered with precise descriptions and findings (5 points)
- Findings are strongly supported with numbers and visualizations (5 points)
- Each question response is supported with a well-discerned statistical analysis from lessons (e.g., aggregation, correlation, comparison, summary statistics, sentiment analysis, and time series analysis) (5 points)

Group Presentation (20 points)

- All group members spoke during the presentation (5 points)
- Group was well prepared (5 points)
- Presentation is relevant to material (5 points)
- Presentation maintains audience interest (5 points)

Slide Deck (20 points)

- Slides are visually clean and professional (5 points)
- Slides are relevant to material (5 points)
- Slides effectively demonstrate the project (5 points)
- Slides are clear and maintain audience interest (5 points)

This project will be evaluated against the requirements and assigned a grade according to the following table:

Grade	Points
A (+/-)	90+
B (+/-)	80–89
C (+/-)	70–79
D (+/-)	60–69
F (+/-)	< 60

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.

- 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.