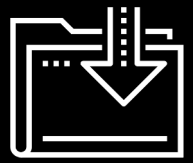
Projects & Collaboration with Git

Data Boot Camp

Lesson 7.1



© 2022 edX Boot Camps LLC. Confidential and Proprietary. All Rights Reserved.

Class Objectives

By the end of today's class, you will be able to:

Articulate the requirements for Project 1.

Draw and interpret diagrams of Git branching workflows.

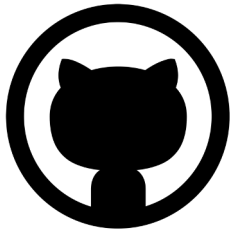
Create new branches with Git.

Push local branches to GitHub.

Time to divide into teams!

Instructor Demonstration

Introduction to Git



Git is a distributed version-control system for tracking changes in source code during software development.

Intro to Git

Working directory

Staging area

Local repository

Remote repository

git add

git commit

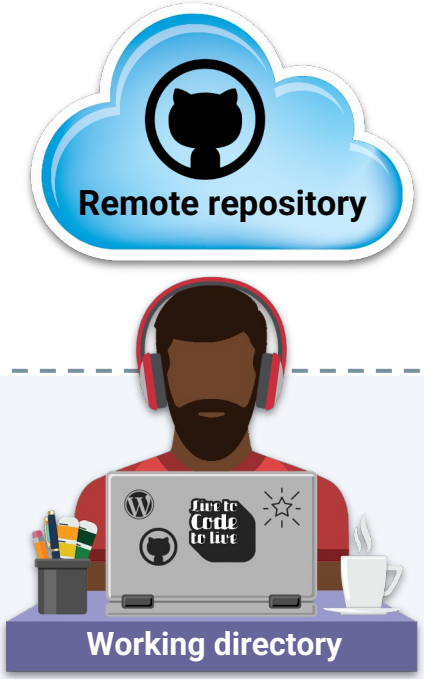
git push

git pull

git checkout

Git and Your Project

Scenario: Your group has been working with Uber's rider data, and you've decided to analyze the average age of the riders:
The root code for the project is called `main`.



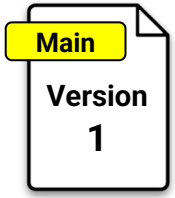
The **staging area** is where you edit the files that will be part of the next commit.

Staging area

Takes an existing GitHub repository and downloads it to the local computer, and links it to GitHub.

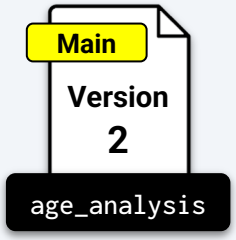
Git and Your Project

Git essentially allows us to write this code and save it with the name `age_analysis`.



The **staging area** is the where you edit the files that will be part of the next commit.

Staging area

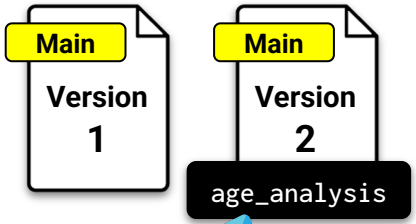


`git commit`

Your staged changes are saved once you commit.

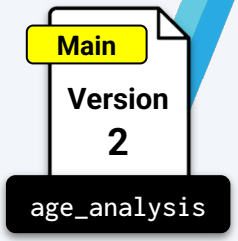
Git and Your Project

`age_analysis` is a branch that originates from the main branch. It contains updates that will be added to the main branch when it's ready to `merge`.



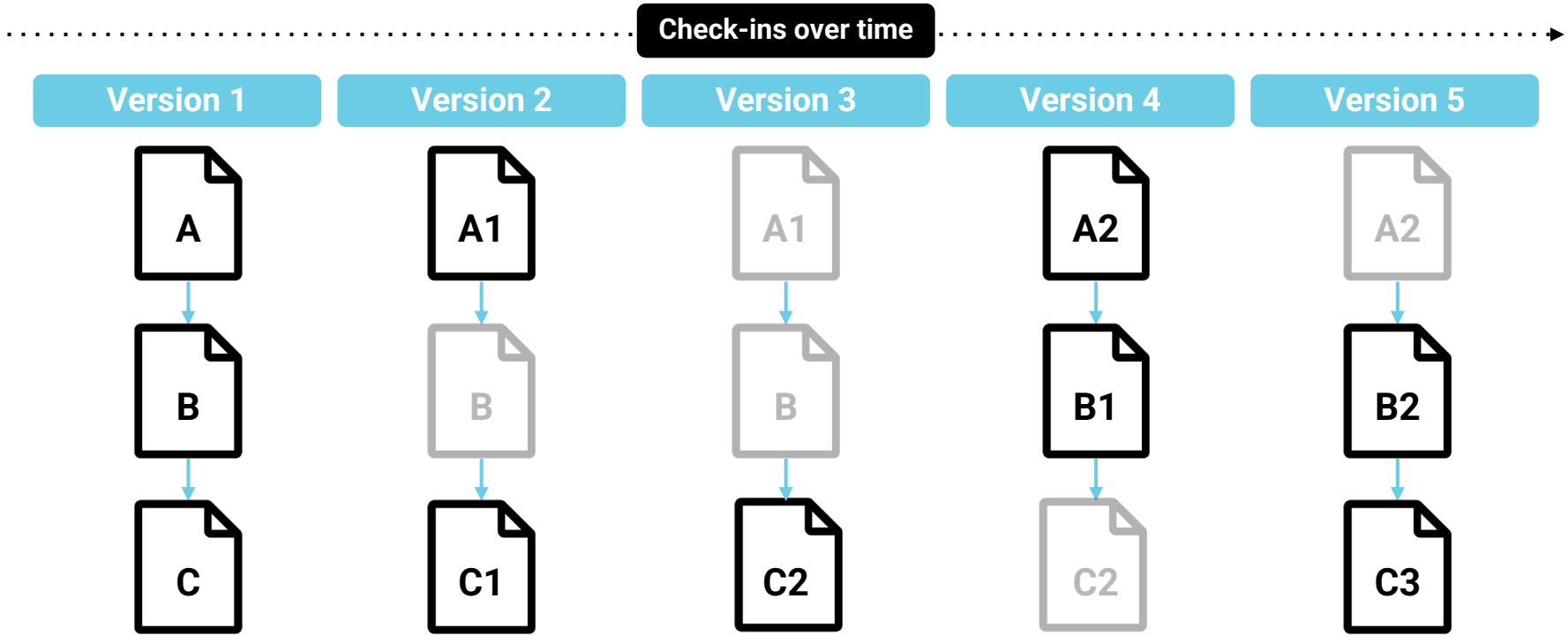
The **staging area** is the where you edit the files that will be part of the next commit.

Staging area



Working with GitHub: Snapshot Model

Version control is a system that allows you to manage changes over time.



Questions?

Activity: Creating a Project Repo

In this activity, one member of each group will set up a GitHub repository to use for the project.

Suggested Time:

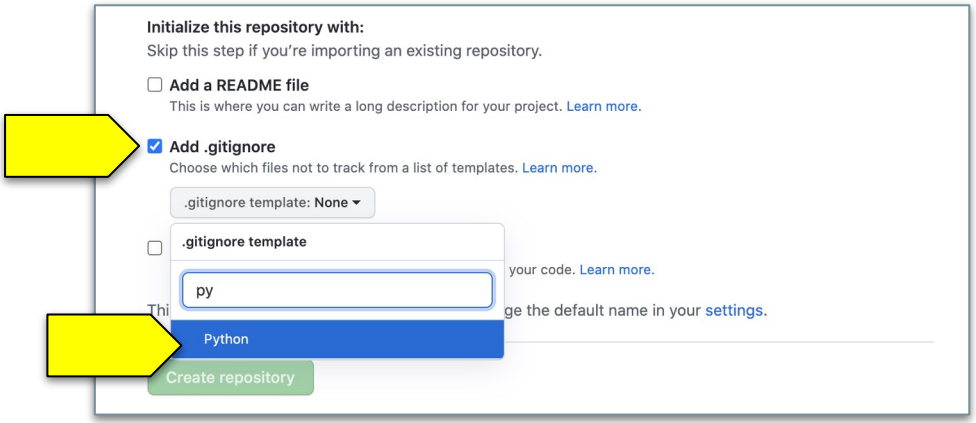
10 minutes

Activity Part 1: Creating a Project Repo

Instructions Part 1

Nominate a group member to create the project repository.
Go to GitHub, and perform the following steps:

- Click the plus sign **(+)** next to your profile picture.
- Click **New Repository**.
- Initialize with `.gitignore`.
- Choose **Python** in the gitignore dropdown.



Activity Part 2: Creating a Project Repo

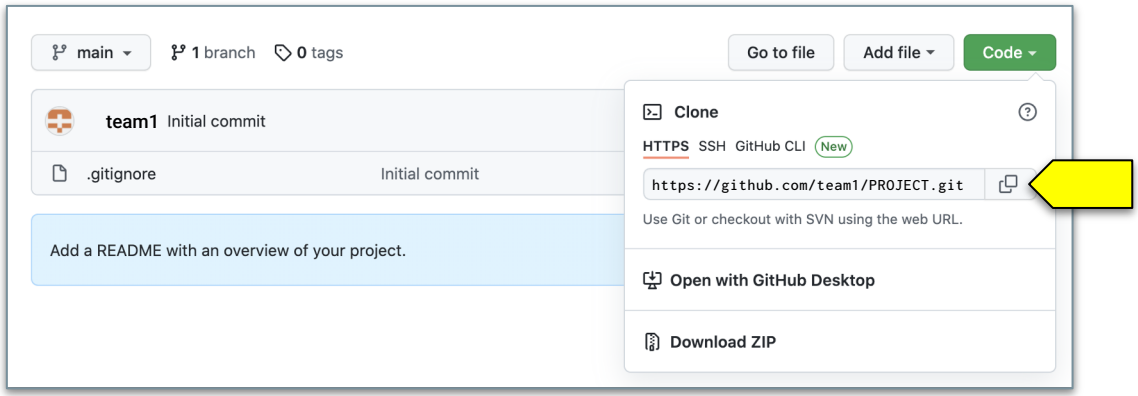
Instructions Part 2

On the repo page, click **Code** to open the **Clone** dropdown and click the clipboard to copy the URL.

Slack the remote URL to your team members.

Everyone on the team should now clone the repository to their local machines.

Open the terminal, and clone the repository with `git clone`, pasting the copied URL.



Activity Part 3: Adding Collaborators

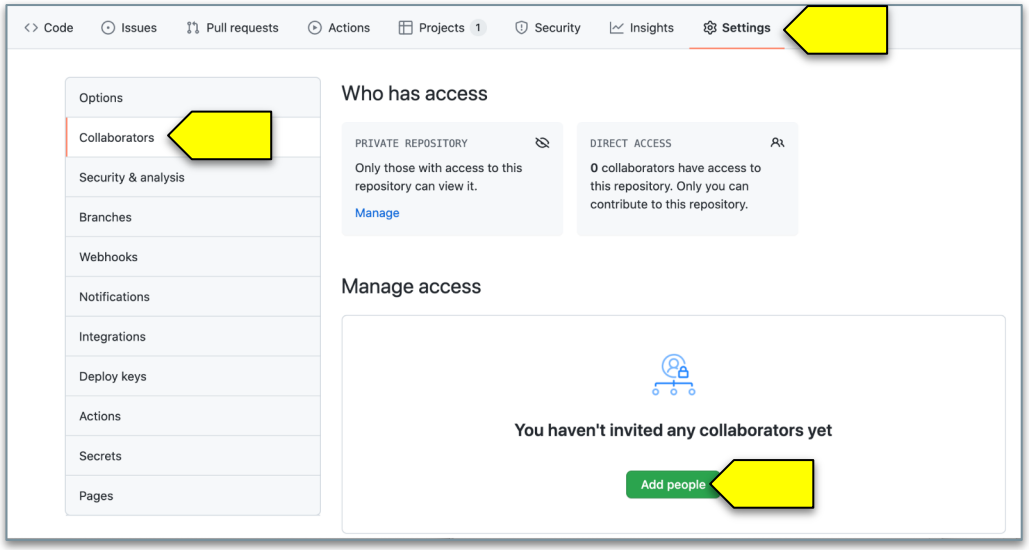
Instructions Part 3

In your Project Repo page, click the **Settings** tab.

Click **Collaborators** in the left-hand menu.

Click **Add people**.

Type in your collaborator's GitHub username, and add the user to the project.



Activity: Workflows

In this activity, you will review the concepts that we have learned so far.

Suggested Time:

5 minutes

Activity: Workflows

Instructions

For this diagramming exercise, either draw your solutions on paper or use the interface provided at Git Viz: https://peleke.github.io/git-viz/.

Check Slack for `Activities/02-Stu_Workflows/README.md`, or open the file directly from your student repo.

Time's Up! Let's Review.

Activity: Creating Branches

In this activity, we will review how to create a branch in GitHub.

Suggested Time:

10 minutes

Activity: Creating Branches

Instructions

To create a new, isolated development history, we must create branches.

```
git checkout -b data_analytics
```

Alternatively, we can create a branch and then switch to it as two separate steps, although this is less common.

```
git branch data_analytics
git checkout data_analytics
```

Activity: Pushing to GitHub

In this activity, we will review how to push a commit in GitHub.

Suggested Time:

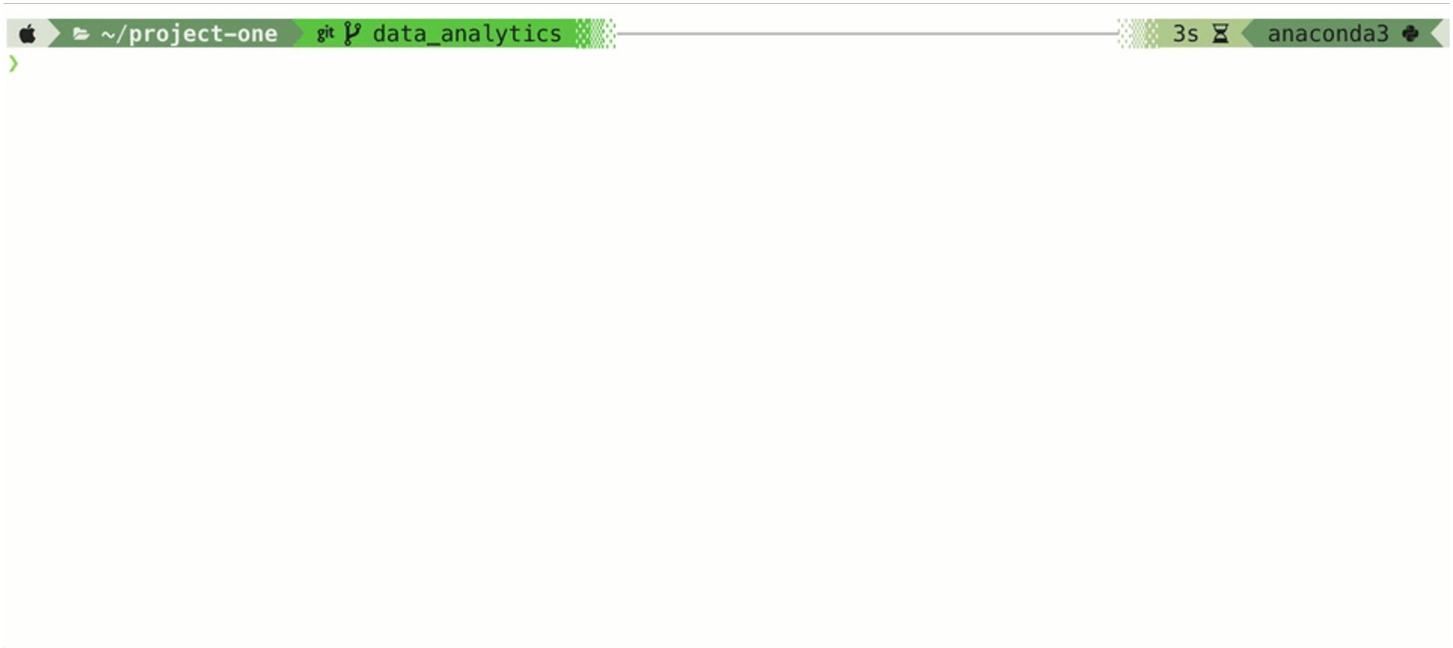
10 minutes

Activity: Pushing to GitHub

```
git push origin <branch_name>
```

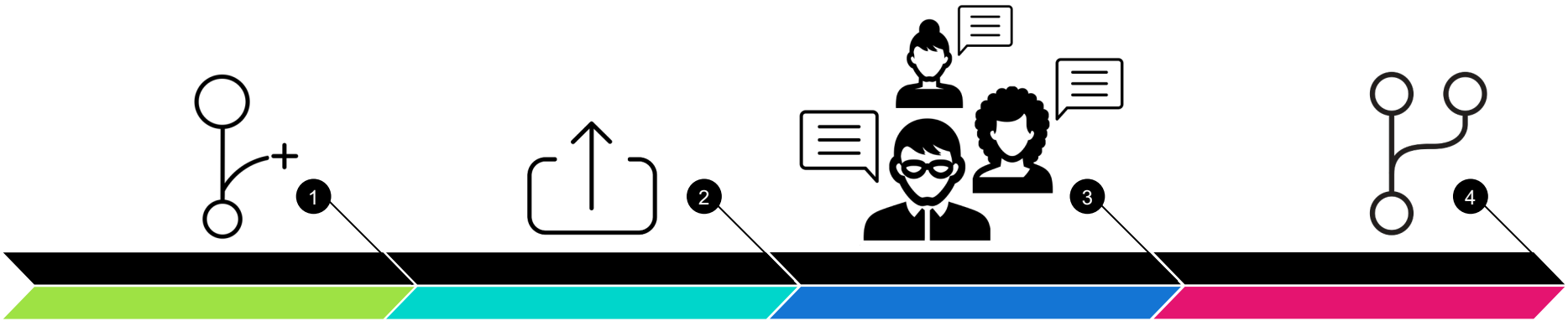
Activity: Pushing to GitHub

`git pull`



Recap

Recap of Git Workflow



Create a Branch:

Push up the branch to your remote GitHub repo. It should have created a remote branch with the same name.

Open a Pull Request:

A pull request is essentially a request to merge code from one branch into another branch on GitHub.

Discuss and Review:

This helps catch any mistakes you might have made and helps teammates understand the work you've done.

Merge and Deploy:

Merge changes into the main branch and deploy the code.

Recap of Git Workflow

`git clone`	Clones a git repository onto the local file system.
`git add`	Adds changed files to the queue of tracked files ready to be committed.
`git commit`	Adds tracked files as a bulk checkpoint ready to be pushed to the remote git repository.
`git push`	Uploads changed files from the local git repository to the remote git repository and updates the remote files.
`git pull`	Downloads changed files from the remote git repository to the local git repository and updates the local files.

Project Week (This Week)!

Project Week (This Week)!

Day 1	Form groups.
	Outline project ideas.
	Perform an initial data exploration.
	Begin research of datasets.
	Submit a project proposal for approval.
Day 2 & 3	Develop your project with your team.

Project Week (Next Week)!

Day 4	Develop your project with your team.
Day 5	Develop your project with your team and prepare your presentation.
Day 6	Present your project to the class.

Project Requirements

Presentation Requirements

You will prepare a formal 10-minute presentation that covers the following topics:

- Questions that you found interesting and what motivated you to answer them
- Where and how you found the data you used to answer these questions
- The data exploration and cleanup process (accompanied by your Jupyter notebook)
- The analysis process (accompanied by your Jupyter notebook)
- Your conclusions, including a numerical summary and visualizations of the summary
- The implications of your findings: what do your findings mean?

Development Requirements

Instructions	Use Pandas to clean and format your dataset or datasets.
	Create a Jupyter notebook **describing the data exploration and cleanup** process.
	Create a Jupyter notebook **illustrating the final data analysis**.
	Use Matplotlib to create 6 to 8 visualizations of your data (ideally, at least 2 visualizations per “question” that you ask your data).
	Save PNG images of your visualizations to distribute to the class and instructional team—and for inclusion in your presentation.
	Create a write-up summarizing your major findings. This should include a heading for each “question” that you asked your data as well as a short description of your findings and any relevant plots.
Bonus	Use at least one API—if you can find one with data pertinent to your primary research questions.

Suggested Data Sources

Suggestions for Data Sources

Feel free to ask the instructional staff for input, but our general advice is to stick to data sources that meet the following criteria:

- They are sufficiently large.
- They have a consistent format.
- They, ideally, contain more data than needed.
- They are well documented.

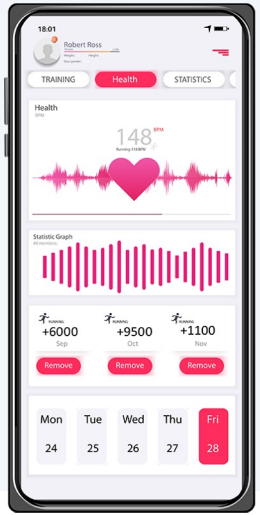
Choosing a Project Track

Choosing a Project Track

This project gives you the ability to focus your efforts within a specific industry. Here are the specializations:

Finance

Healthcare



Custom

EDA in Finance

Why is exploratory data analysis important in the financial sector?	• Identifying deals • Private equity • Arbitrage opportunities • Liquidity • Finance/refinance trends
Who would use this skill?	• Investment banking professionals • Private equity analysts • Lending analysts • Financial administrators • Real estate professional

Financial Analyst: Equity Trading

EDA in finance example:

01

While working for a large equity trading company, you're tasked with researching a client's portfolio. They want to invest in telecom stocks and need expert analysis to make the right decision.

02

Using the Nasdaq Data API, pull a year's worth of trading data for the major cell phone providers: AT&T, T-Mobile, and Verizon.

03

Which stocks are trending upward? Which are trending down? From these data, what recommendations would you make to your client?

Financial Analyst: New Car Loan Analysis

EDA in finance example:

01

People have been financing higher car values over longer amounts of time.

What is driving these decisions?

02

Search for answers using data collected from the Federal Reserve Economic Data (FRED).

What other questions can you answer with these data?

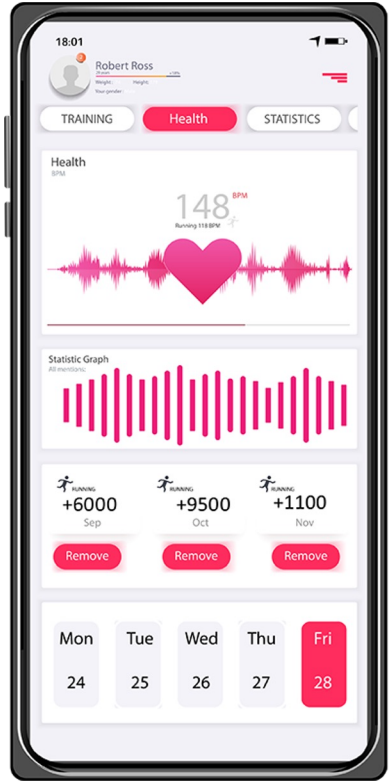
03

What do your results suggest about the time value of money?

What about the impact of these loans as time goes on?

EDA in Healthcare

Who in the healthcare industry does EDA?	• Clinical data analyst • Pharmaceutical testing • Healthcare economics researcher • Senior policy analyst • Compliance operations analyst • Public health informatics scientist
Why is it important?	• Predicting and diagnosing illnesses • Greater accuracy and impact • Improve patient safety • Improve diagnoses • Greater understanding of disease risks and causes • Greater prevention strategies



Mental Health in Tech

EDA in healthcare example:

01

People working in tech are often at their desks for extended amounts of time.

How does this correlate to one's mental health?

02

Examine the data collected through surveys and search for trends.

Find out if there is a link between mental health and companies that offer wellness programs.

03

What do the results show you about the prevalence of mental health in tech?

Can you suggest steps that companies can take to better aid their employees?

Personal Fitness Analyst

EDA in healthcare example:

01

Does working out help a person become more active overall?

An analytic team with Personal Fitness, Inc. has decided to tackle this very question.

02

Using data collected by the Samsung Health application, the team wants to uncover trends within the data.

03

What do the results tell you about individuals using this app?

Have their lifestyles become more active? Less? Remained the same?

EDA

Who else utilizes EDA? Everybody! We've only specified healthcare and finance as industry specifics, but any industry that uses data benefits from EDA.

Other professions that use it include:

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

Private Investigator

EDA example:

01

Use aggregate crime data from different police precincts in a city to uncover patterns in criminal activity.

02

Most crime in New York City takes place in the summer.

Can you uncover similar patterns in your city?

03

What do your results suggest about how police should plan their patrols?

What do your results suggest about how best to distribute law enforcement resources over the calendar year?

www.nydailynews.com

Uber Rides and Weather

EDA example:

01

No one likes to walk in subzero temperatures or scorching heat. Do people use Uber more when the weather is uncomfortable?

02

Using Uber ride data from Kaggle and data from a weather API, find out if people take Uber more during summer and winter, and if there are relationships between daily temperature and ride frequency.

03

What do the results tell you about surge-pricing strategies and commuter habits?

Today’s Focus

Today's Focus

By the end of today's class:

- Brainstorm possible project ideas.
- Begin data research.
- Write a description of the scope of your research.

Create a short 1-page project proposal that covers the following:

- Project Title
- Team Members
- Project Description/Outline
- Research Questions to Answer
- Datasets to Be Used
- Rough Breakdown of Tasks

Questions?