Project #1: EDA and Git Collaboration

Data Boot Camp

Lesson 7.1

Class Objectives

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

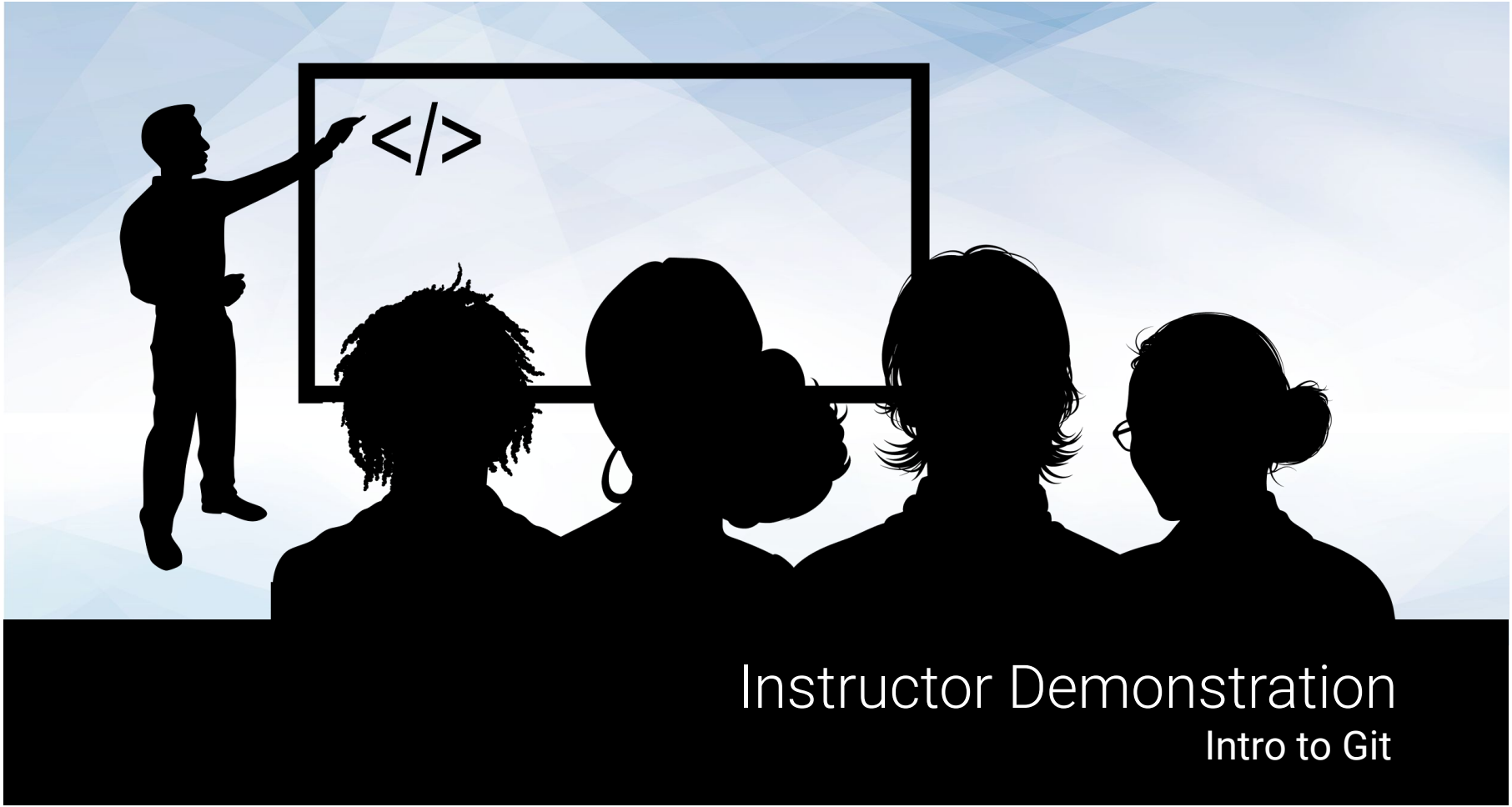
Articulate the requirements for Project 1.

Draw and interpret diagrams of Git branching workflows.

Create new branches with Git.

Push local branches to GitHub.


START



Instructor Demonstration

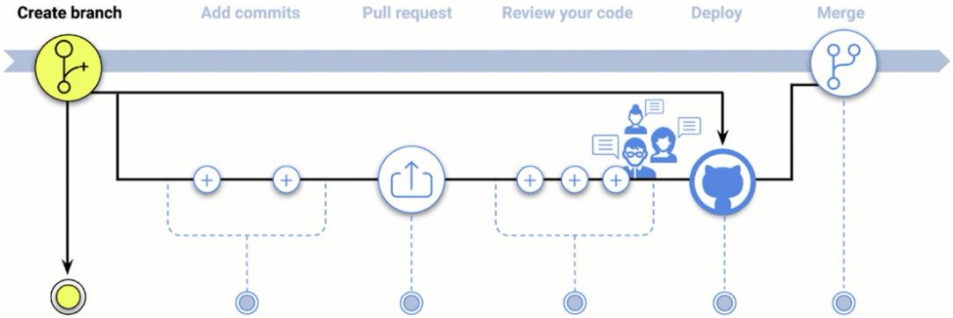
Intro to Git

Intro to Git

git --everything-is-local

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

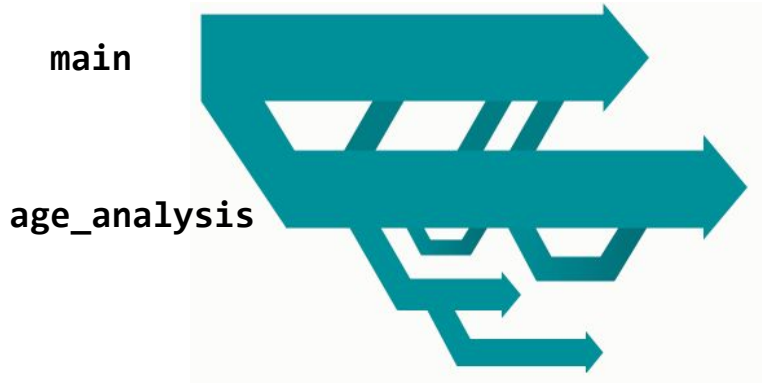
git commit -a



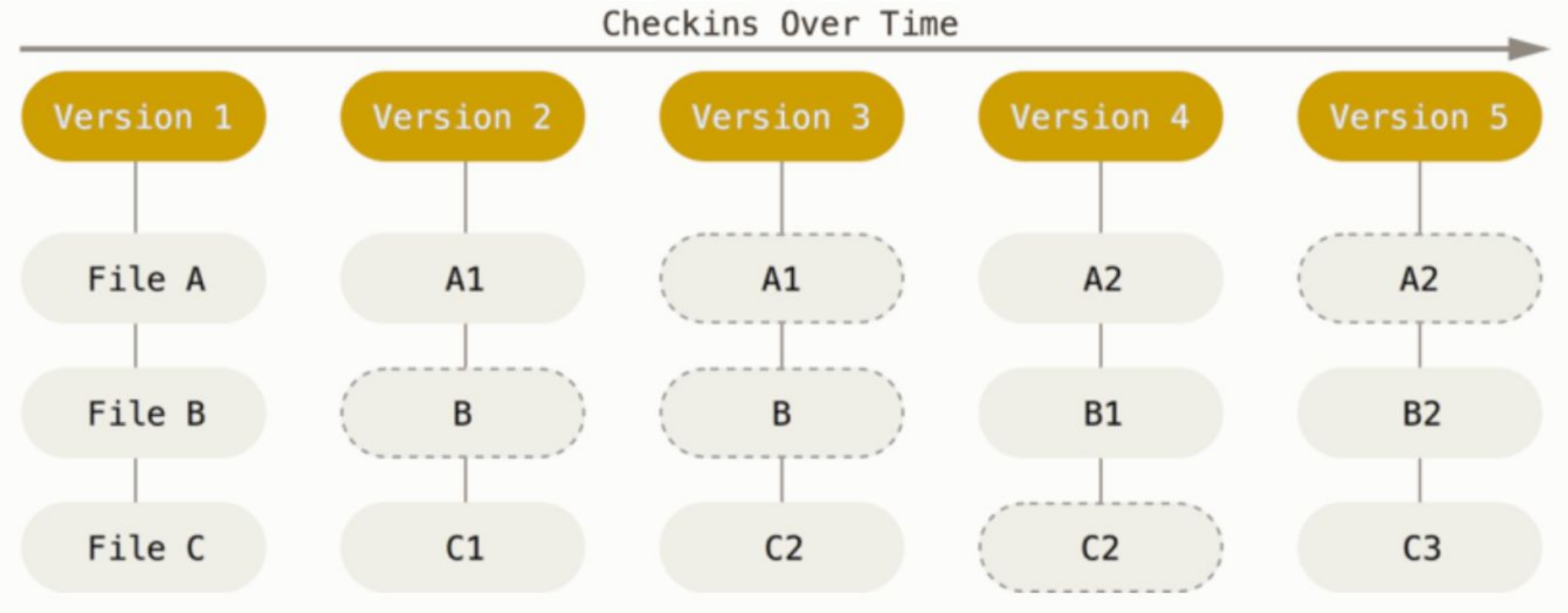
Intro to Git

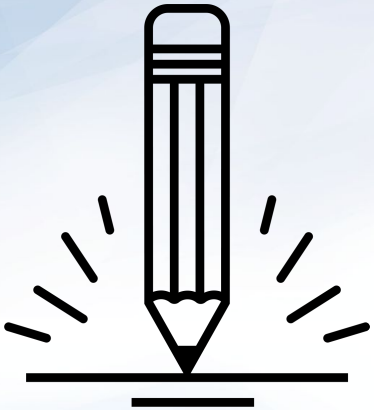
Git and your Project!

- **Scenario**: Your group has been working with Uber's rider data, and you decided to analyze the average age of the riders:
  - Git essentially allows us to write this code, and save it with the name: `age_analysis`.
  - The code in `age_analysis` differ from the root code.
  - The root code for the project is called `main`.
  - `age_analysis` is a branch originated from the `main` branch. It contains updates that will be added to the main branch when it's ready to merge.



Git's "Snapshot model"





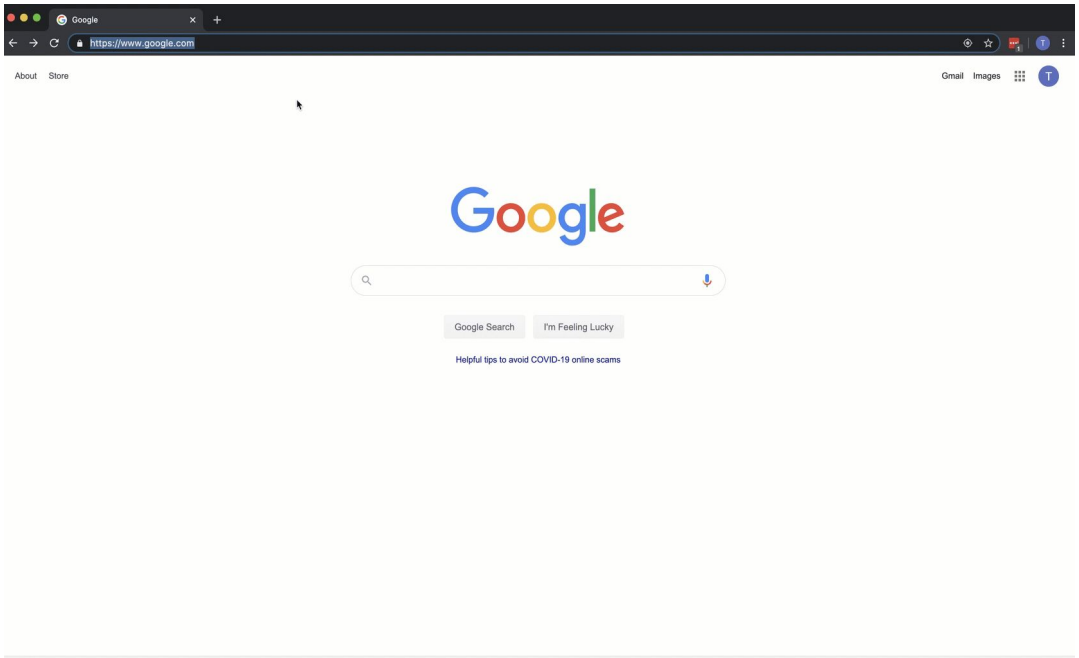
Activity: Creating a Project Repo

In this activity, everyone will set up a GitHub repository that we can use for our projects.

Suggested Time:
10 Minutes

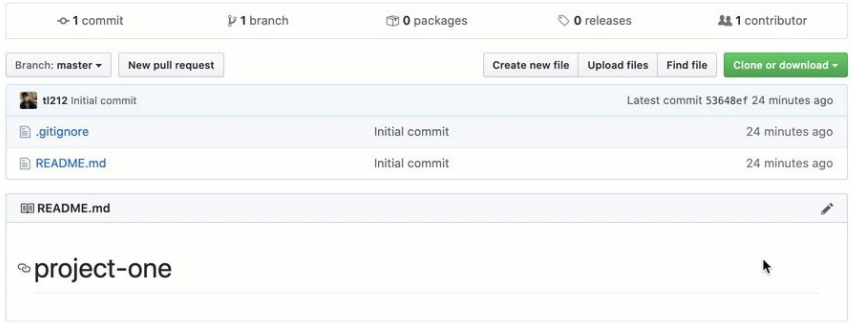
Activity: Creating a Project Repo

- Nominate one of your group member to create the Project Repository.
- Go to GitHub and do the following:
    - Click on the plus sign next to your profile picture.
    - Click on ‘New Repository’.
    - Initialize with `.gitignore`.
    - Choose `Python` in the gitignore dropdown.



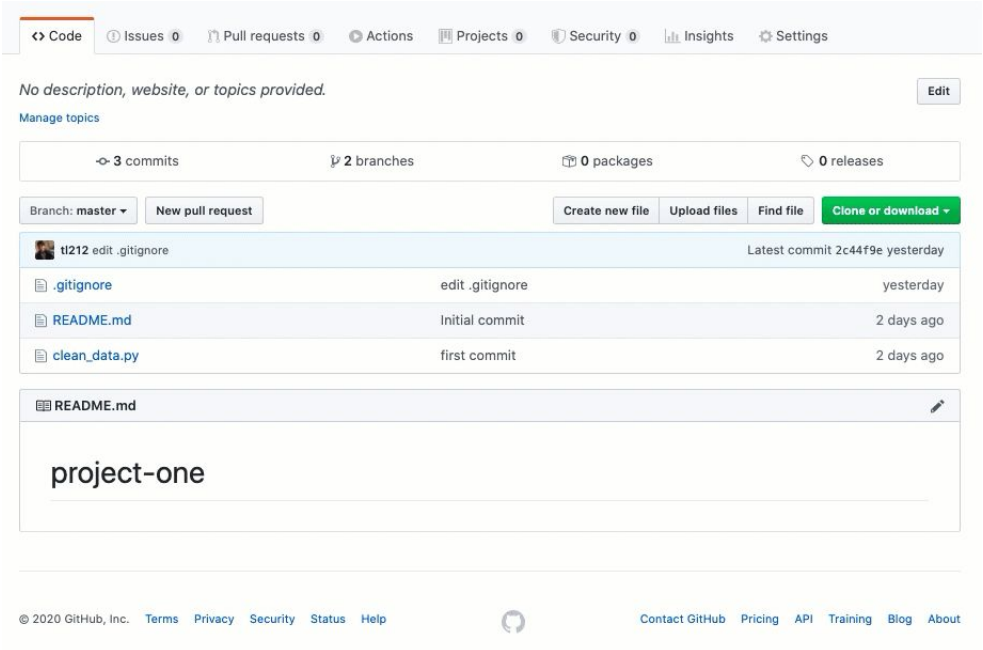
Activity: Creating a Project Repo

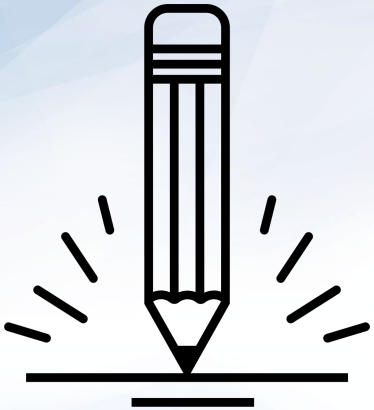
- Slack the remote URL to your team members.
- `git clone` the repository.
    - Once in the repo page click ‘Clone or download’ and copy.
    - Open Terminal and `git clone`



Activity: Creating a Project Repo

- In your Project Repo page click on the “Settings” tab.
    - Click on “Manage access” on the left panel.
    - Click on ‘Invite collaborator”.
    - Type in your college GitHub username and click add.





Activity: Workflows

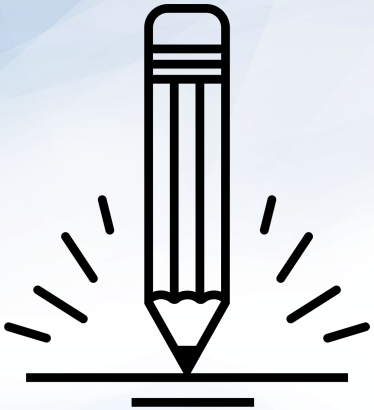
In this activity, you will take a few minutes to review the concepts we have learned.

Suggested Time:
5 Minutes

Activity: Workflows

- This is a diagramming exercise. You can either draw your solutions on paper, or use the interface provided at Git Viz.
- Check your slack for: Activities/02-Stu_Workflows/README.md or open directly from your students repo.

Time’s Up! Let’s Review.



Activity: Creating Branches

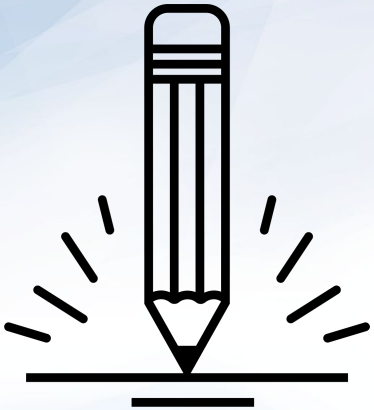
In this activity, everyone will review how to create a branch in GitHub.

Suggested Time:
10 Minutes

Activity: Creating Branches

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

- Alternatively, we can create a branch and then switch to it as two separate steps, though this is uncommon.



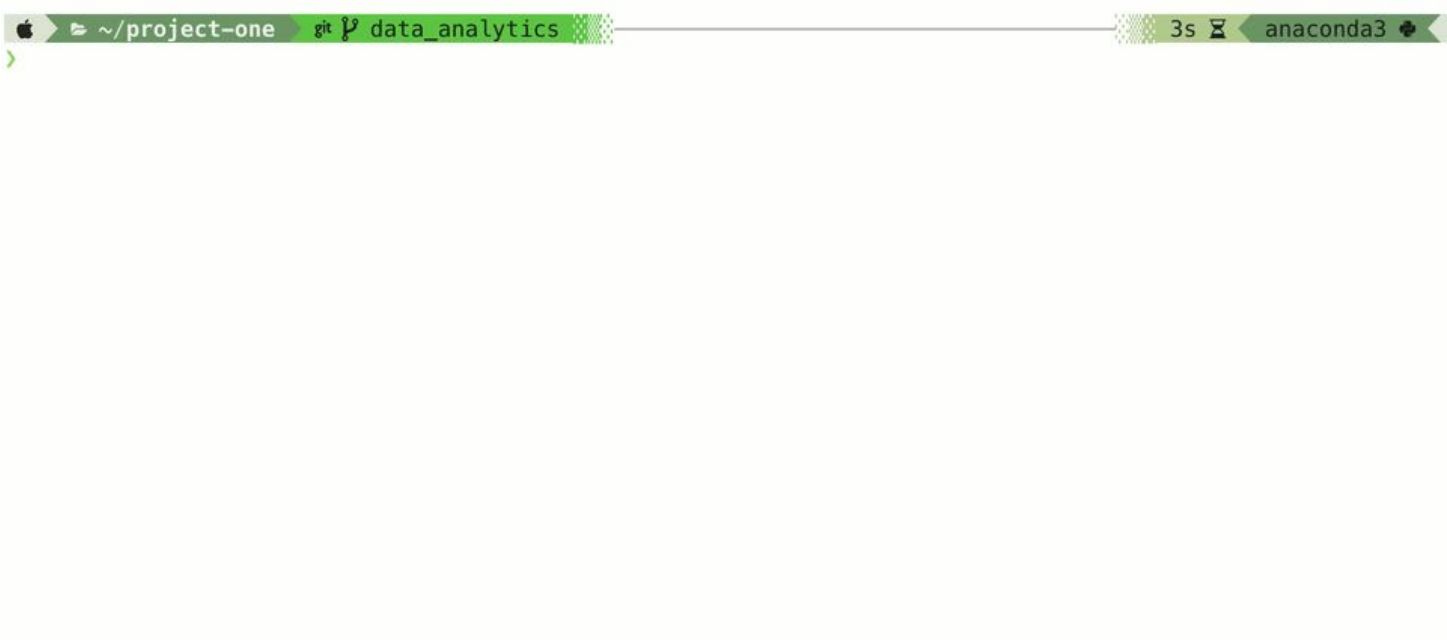
Activity: Pushing to GitHub

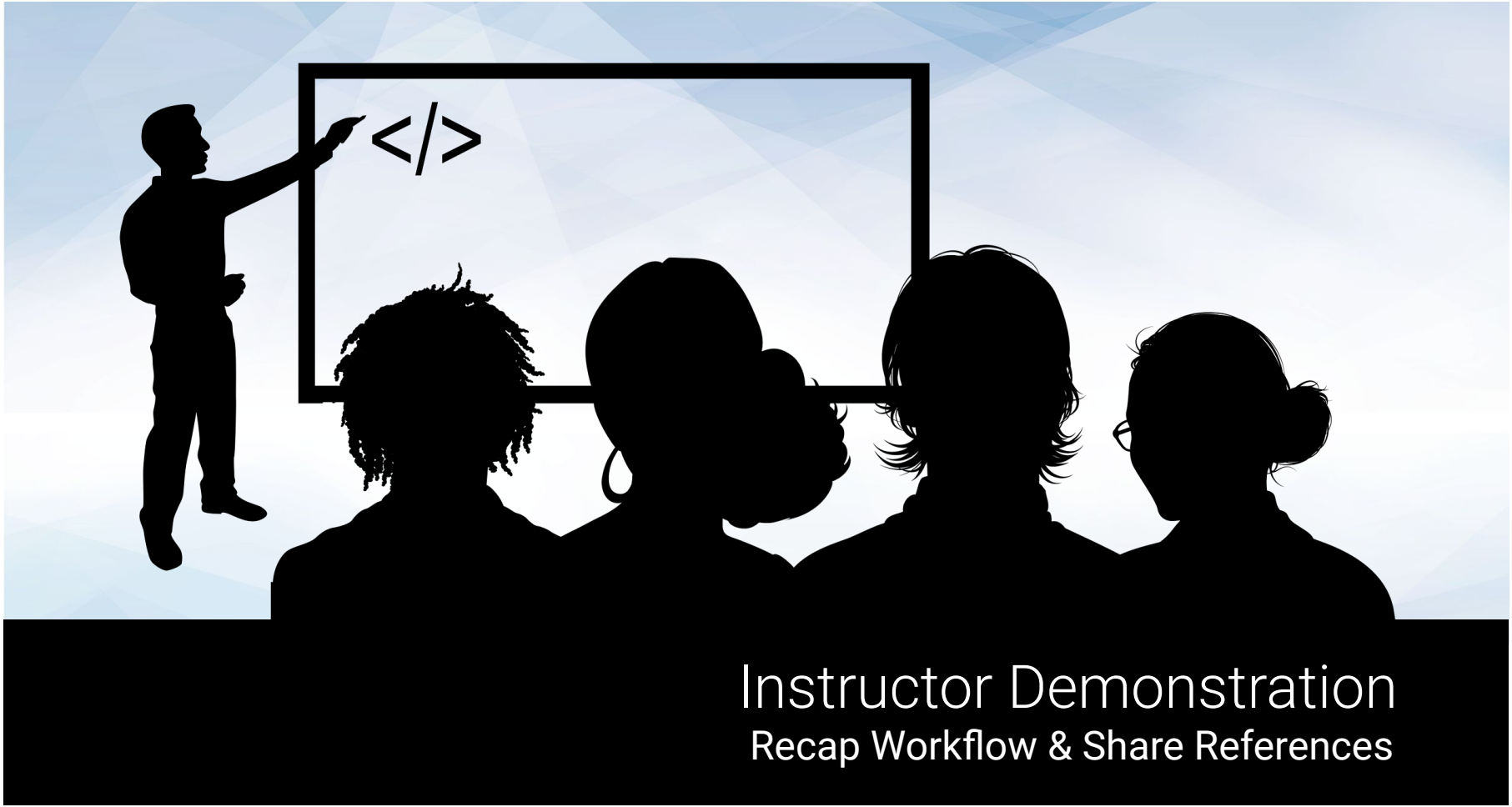
In this activity, everyone will review how to push a commit in GitHub.

Suggested Time:
10 Minutes

Activity: Pushing to GitHub

Activity: Pushing to GitHub

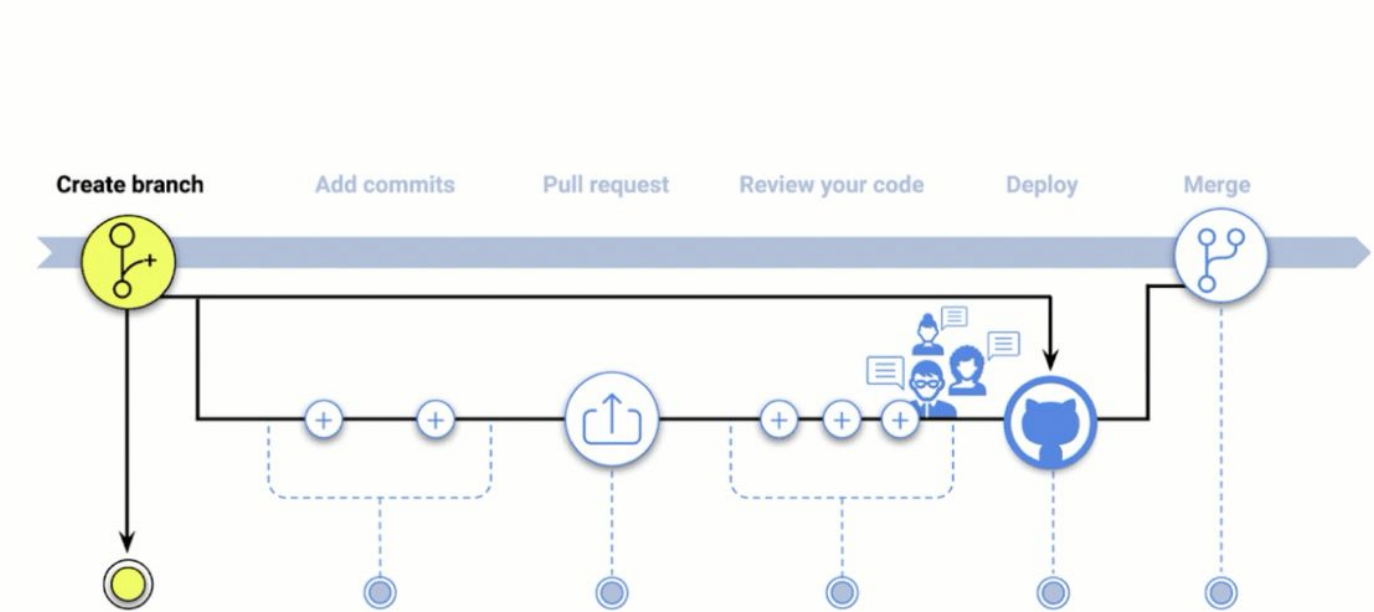




Instructor Demonstration

Recap Workflow & Share References

Recap Workflow & Share Preferences

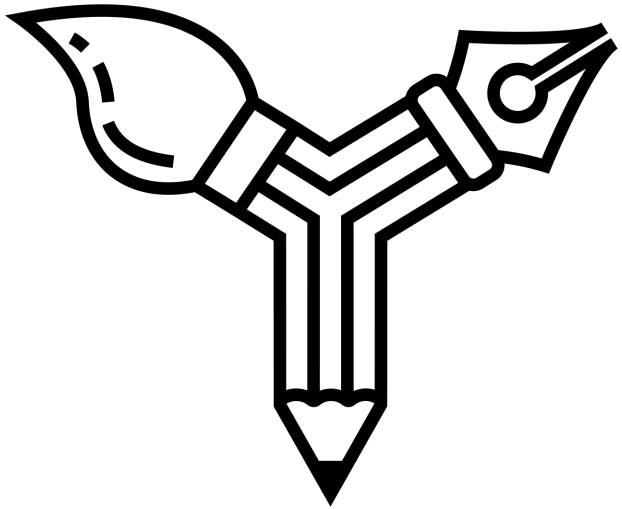


Project Week Overview

Project Week! (This Week)

Day 1:

- Form groups (3-5 members each)
- Outline project ideas
- Initial data exploration
- Begin research of datasets
- Submit project proposal for approval



Day 2:

- Hardcore development

Day 3:

- Hardcore development

Project Week! (Next Week)

Day 4:

Hardcore development

Day 5:

Hardcore development

Presentation prep

Day 6:

Presentations

Project 1: EDA and Git Collaboration

Development Requirements

You will also be responsible for:

- Using Pandas to clean and format your dataset(s).
- Creating a Jupyter Notebook describing the **data exploration and cleanup** process.
- Creating a Jupyter Notebook illustrating the **final data analysis**.
- Using Matplotlib to create a total of 6–8 visualizations of your data (ideally, at least two per ”question” you ask of your data).
- Saving PNG images of your visualizations to distribute to the class and instructional team, and for inclusion in your presentation.
- (Optional) Using at least one API if you can find an API with data pertinent to your primary research questions.
- Creating a write-up summarizing your major findings. This should include a heading for each “question” you asked of your data and a short description of your findings and any relevant plots.

Presentation Requirements

You will also be responsible for preparing a formal, 10-minute presentation that covers:

- Questions 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, which should include a numerical summary and visualizations of that summary
- The implications of your findings: What do your findings mean?

Project Rubric

Rubric at a Glance

Categories for grading

GitHub repository (20 points)

Visualizations (20 points)

Analysis and conclusion (20 points)

Group presentation (20 points)

Slide deck (20 points)

Suggested Data Sources

Suggestions for Data Sources

Feel free to ask us (the instructional staff) for input, but our general advice is to stick to data sources that:

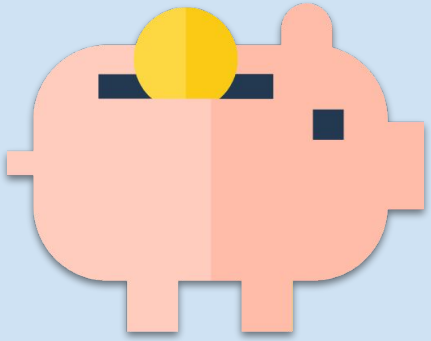
- Are sufficiently large.
- Have a consistent format.
- Ideally, contain more data than needed.
- Are well documented.

Choosing a Project Track

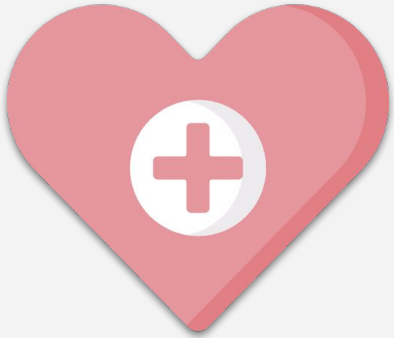
Choosing a Project Track

This project gives you the ability to focus your efforts within a specific industry.

Here are the specializations:



Finance



Healthcare

Custom

EDA in Finance

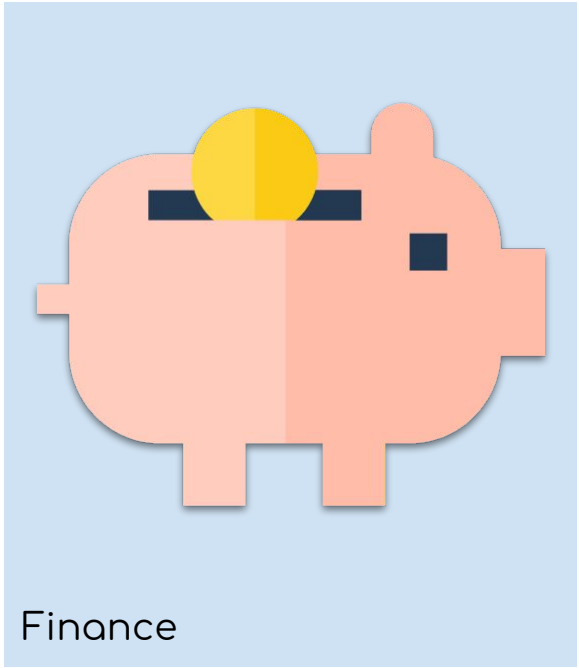
Why is exploratory data analysis important in the financial sector?

- 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 professionals

Let's look at some examples!



Financial Analyst: Equity Trading

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 Quandl 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?

https://www.quandl.com/tools/api

Financial Analyst: New Car Loan Analysis

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?

https://fred.stlouisfed.org/series/DTCTLVENANQ

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

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?

https://www.kaggle.com/osmi/mental-health-in-tech-survey

Personal Fitness Analyst

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?

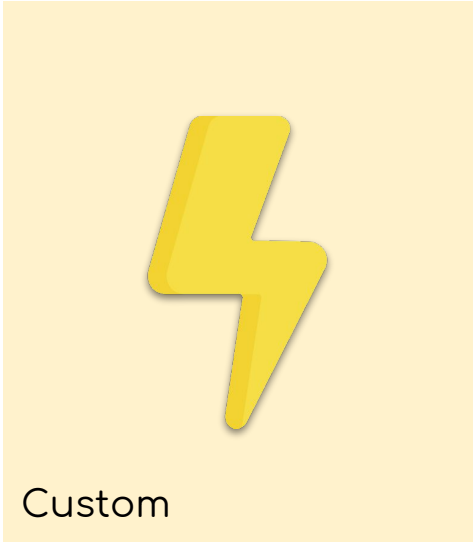
EDA

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

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?

Uber Rides and Weather

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

By the End of Today's Class...

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

Create a short, one-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

Time to divide into teams!

Questions?

Countdown timer

15:00

(with alarm)


TRADING PLATFORM
ETH USD LTC
190.34 +3.44 +1.88%
199.31 +0.31 +0.05% 19:55 Jun 01.06
month
SELL
BUY
Sales
UP 23%
DOWN 14%
Month
190.34
BTC USD


The End