DataDive Week 1

SheffDataSoc

Data Science Society

Session Outline

- What will the DataDive be?
- The DS Project Pipeline (CRISP-DM)
- How to form a good research question
- Using VSCode (& Other IDEs)
- How to collaborate using git and GitHub
- Details of the next sessions

What is the DataDive?

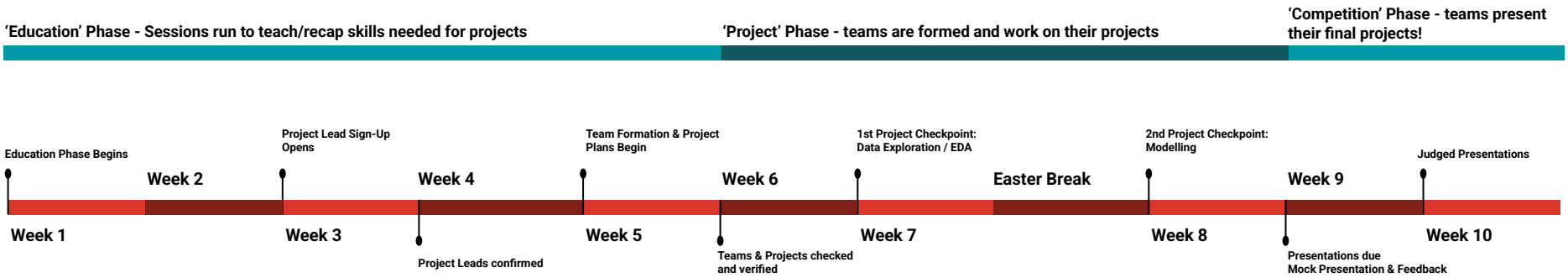
- Data science competition-based to showcase and develop essential data-driven skills
- Inspired by UIUC Illinois Data Science Club
- Example of a GitHub repo where teams will work together: GitHub
  - Organize how the team due fits
  - Should have a **README.md** to describe the data
- Example of presentation powerpoint to present in front of judges: Presentation
  - Shows the different stages of the data science pipeline
  - Findings when reaching to the project's research question/goal

The DataDive!

This is a great opportunity to meet new people, and work on a data science project of your choosing! You can mention your project on your CV and in interviews.

The final presentations of the projects will be judged in-person in Week 10 by people from industry, and the winning teams will receive prizes worth up to ~£200!

DataDive Timeline



Project Pipeline - CRISP-DM*

*the Cross-Industry Standard Process for Data Mining

This is one of the frameworks used in industry for the lifecycle of a Data Science project.

The DataDive will roughly follow this structure, with teams defining a problem, acquiring, exploring and preparing data, and building and evaluating a model.

*the deployment stage is not necessary for this competition

How To Form A Good Project & Research Question

A good project vision is one that all team members can understand and define.

The question/topic should cover:

- **What**: what data is the project analysing / what is the model doing?
- **Who**: who does the project benefit?
- **Why**: why is the project necessary?

It can be on (*almost*) anything you like, but should be something that has some kind of real-world benefit or impact for a certain group of people.

The projects are judged on how they address the overall question, and what insights are found.

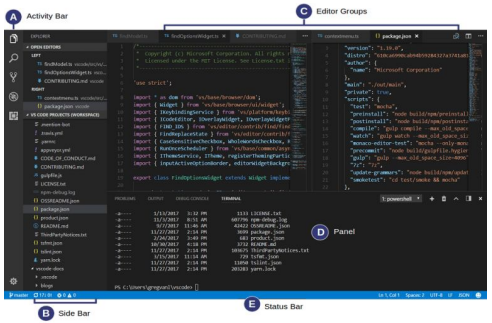
(Importantly, the project should be realistic to complete within the timeframe!)

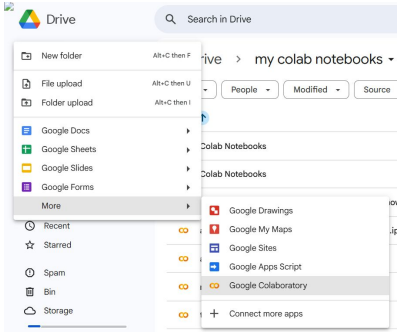
Working On The Project - Coding

As these are Data Science projects, you will likely want to use Python to work on them (though other languages like R can also be used).

There are a variety of platforms/IDEs (integrated development environments) you could work with, such as: VSCode, Spyder, Jupyter notebooks, Google Colab.

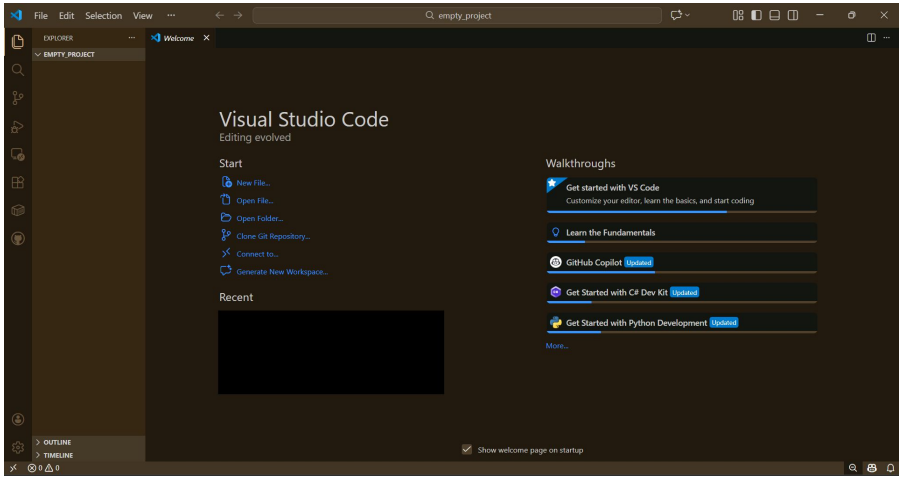
Your project lead may have more experience with some over others, and you may want to use different platforms for different tasks (eg VSCode for collaborative coding, but Colab for training models).



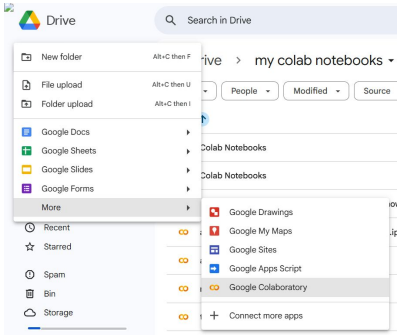


Working On The Project - VSCode Set-Up

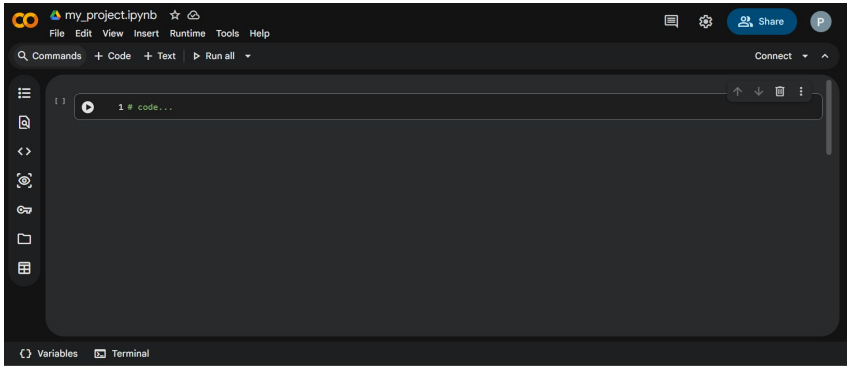
Useful links for installing and setting-up VSCode:
https://code.visualstudio.com/?wt.mc_id=vscom_downloads
https://code.visualstudio.com/docs/setup/setup-overview



Working On The Project - Other Ways To Run Notebooks

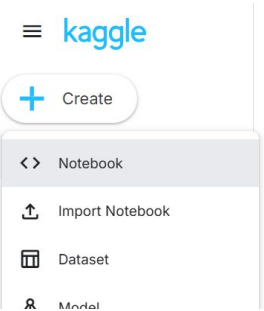


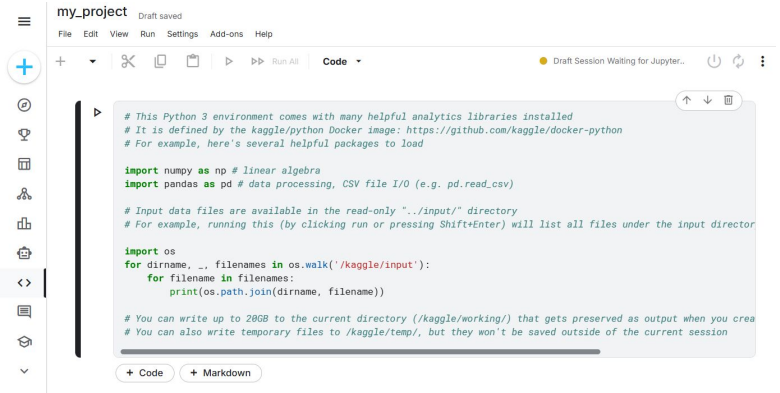
You can also create Google Colab python notebooks from within Google Drive, and work on and run them in the browser.



Similarly, you could use kaggle notebooks!

These options will be useful if you need more processing power (eg to train ML models).





Collaborating On The Project - Using GitHub

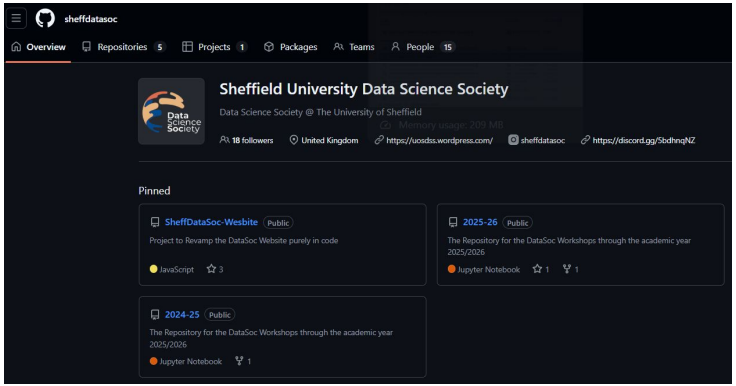
Once teams have formed (week 5-6 of the semester), a GitHub repository will be created by DataSoc and the project leads for each project.

GitHub is a very widely used platform for remotely storing and collaborating on code. It integrates well with IDEs like VSCode, but can also be used with git (see next slide) or via the website.

Signing up for a new personal account

1. Navigate to https://github.com/.
2. Click **Sign up**.
3. Alternatively, click on **Continue with Google** to sign up using social login.
4. Follow the prompts to create your personal account.

During sign up, you'll be asked to verify your email address. Without a verified email address, you won't be able to complete some basic GitHub tasks, such as creating a repository.



Collaborating On The Project - Using Git

Git is a version control system, integrated with GitHub. Teams can use git to save changes on their code, push this to the remote repository, pull updated code from other teammates from the remote repository, create branches for separate users/features, etc.

Git can be used via the command line.

How you install Git can depend on your operating system, see here for details: https://git-scm.com/book/en/v2/Getting-Started-Installing-Git
It may also be useful to get Git Bash alongside, if using Windows.

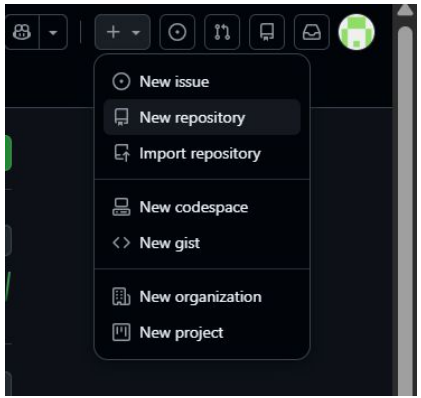
Some resources for using git commands:

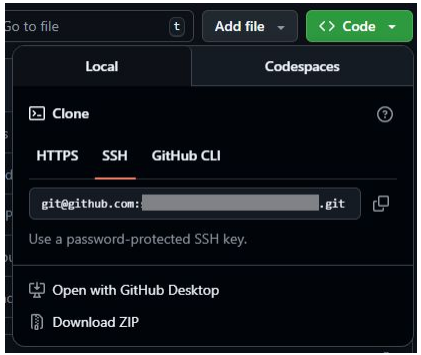
- https://education.github.com/git-cheat-sheet-education.pdf
- https://git-scm.com/cheat-sheet

Collaborating On The Project - Set-up Example

1. Once logged into GitHub, you can create a repository (folder for a project's files), and follow the instructions it gives you

2. Within the repository, you can then copy the link to be able to 'clone' it to your machine

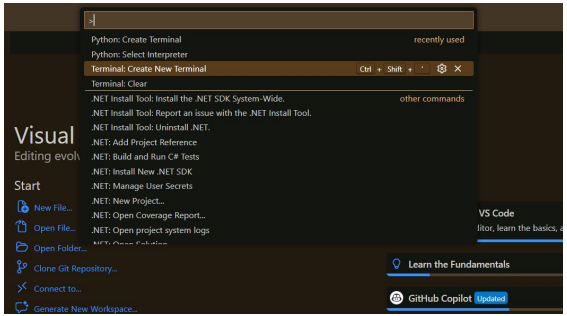


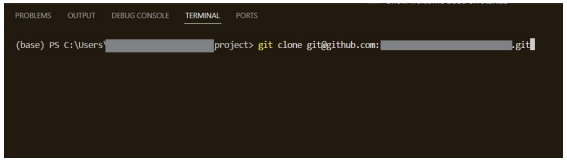


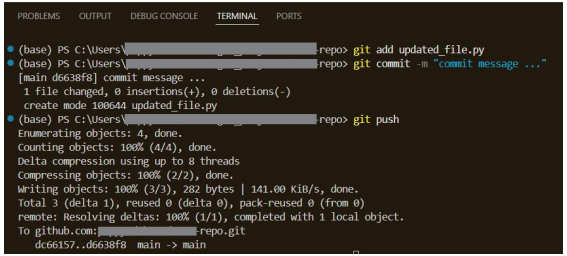
3. Within VSCode (or your chosen IDE), open a terminal

4. Clone the GitHub repository with **`git clone`**

5. Begin to add files and make changes! Use **`git add <file>`** to stage a change, **`git commit`** to commit it to your local repository, and **`git push`** to push the change to the remote repository on GitHub







What comes next?

Thank you for coming to Week 1! Hopefully you are excited to get involved in your own DataDive project.

These slides will be accessible after the session to re-visit, should you need it.

You could have a think about the sort of project you would like to work on, or who you might want to work with (but team-formation will come later on, so don't worry at this stage).

If you think you might want to be a *Project Lead*, applications for this will open in Week 3.

The remaining DataDive 'education' sessions will be →

Week	Session	Where/When
2	Data Understanding & Collection	**Diamond LT9** **Wednesdays** **4-6pm**
3	Data Cleaning & Pre-processing	
4	Creating ML Models	
5	Evaluating & Interpreting ML Results	