

IPDS Project (50%)

Project Scope

The project is designed to empower you in gathering and presenting intriguing information. Its inherent flexibility facilitates the development of creative endeavors tailored to **your interests** and aspirations.

Necessary Tasks

Your project must have the following utilised in some way:

- It must be an “end-to-end” system for data gathering/analysis and some visualisation.
- You gather data from multiple sources (details in next section).
- The data will need to be checked for errors/cleaned as necessary.
- Merge the multiple data into a database, all queries done in the workbook. Use SQLite database and include this file in your submission.
- Query the database to extract information into Dataframes.
- Plot the information, do some small analysis (maybe a basic Hypothesis test) on the data. Even simple things like means/medians etc.
- Perform some statistical analysis on the data. The best submissions will incorporate learning from your other modules. Are there trends in the data that could identify?
- Write an additional Document describing your work

Data Sources

When I say it must be from multiple sources, the sources can be

- Take an interesting dataset from <https://ourworldindata.org>, Kaggle, DataWorld
- Augment the data with some web-scraping that can be combined with your chosen dataset.
- Take information from an accessible API to build your dataset/augment the chosen one.

Whichever you use, incorporate at least one web-scraping into it.

Example 1 One of the datasets could be from Spotify about some artists. You could scrape some additional information from Wikipedia about the artists. You could get further information about an artist’s songs (such as the danceability of a track) from Spotify’s API. The end result could be a database somebody could query and some visualisations - How does the progression of danceability go through an album? Does an artist start with more danceable songs? Is it all killer no filler? Can you make predictions based on danceability vs another property of a song?

Example 2 Covid-19 is an obvious area of interest for many people in the world. You could build a system that reads in data from multiple sources, gathering it together into one database. There are lots of websites/news organisations visualising the data - you can build some of these visualisations using matplotlib. <https://covid19api.com/> is an API with data. What about more local information to Ireland or a different country of your choice? Compare different countries? John Hopkins have a Coronavirus GitHub repository

Document

These are the headings you should fill in. It does not need to be long, no more than 4 pages and written in any text editor but submit it as a PDF. It should be written in an academic

style, i.e. formal, without code blocks, numbered headings, and referencing data sources/papers correctly.

- Project Scope
- Description of the data
- Database
- Analysis/Visualization
- Conclusion
- References