**GitHub Template**

**Goals of the project**

To curate a detailed data set containing all of Metallica’s live performances to date. Data to be enhanced by two other data streams in order to further enrich it and make it more useful/interesting for analytical purposes further down the line.

**Why you created the project**

Despite being one of the most successful heavy metal bands of all time (link), any readily available data for Metallica is limited.

There are a handful of existing (and unofficial) CSV files, although these focus on album and track data and not live performances.

https://www.kaggle.com/conklindata/metallica-song-catalog/version/1

The most thorough (and accurate) data available for live performances can be found on metallica.com, however it is only currently available in html form and they haven’t made it available for fans to download in a more manipulative state – which is where this project comes in.

So the benefits to this project are three fold:

* I get to create a useful and possibly unique dataset that doesn’t exist online anywhere
* I get to add value to the Metallica community!
* I get to practice some key data science skills:
  + Acquiring data from ‘hard-to-find’ locations (i.e web scraping)
  + I get to work with multiple data sets and engineer new features. These include integrating live performance data with:
    - Album data
    - Geocoded data

**Methodology & Approach**

The best approach will involve some kind of web scraping. I will use Selenium for the web scraping stage and Pandas to manage data manipulation and analysis. In addition to scraping from Metallica.com I also would like to integrate further streams of data unavailable on Metallica.com - specifically latitude and longitude data for each gig (we can make some interesting map visuals with this) and additional album level data (which we can use to work out live plays by album).

The diagram below summarises the data streams I intend to obtain and how I intend them to work with one another.

Once I’ve scraped, cleaned and integrated all these data streams into a final .csv file, I can then begin to look at analyzing and visualizing the data.

The project in its current state splits into three stages and for ease of reading I’ve split this across three notebooks

* 1.1\_Data\_Acquisition\_Primary\_Data
* 1.2\_Data\_Acquisition\_Supporting\_Data
* 2\_Data Cleaning\_Feature Generation

I intend to continuously build on this project as time goes on so will add new notebooks as and when I have time to add to the project

**How to install the project**

Feel free to fork and clone this project! You’ll need a github account and a python environment to run the code, Jupyter notbook might make sense seeing as that’s what I wrote in.

**Some key libraries are required in order to run this:**

Selenium

Chromedriver / Webdriver

TQDM Notebook (not essential, but helps to see progress)

Pandas

Numpy

Googlemaps( note - if you want to use Google’s API then you will need to add your API key in the script where I’ve made a note)

**How to run the project**

**Observations and improvements**

**Link to your Jupyter Notebook**