* Objective

For our project we will extract, transform and load several datasets to answer the following question; For films, video-games and the adaptations in between which came first and how successful where they?

An example is the terminator film series and terminator video games. Pandas .read\_html() and .read\_csv() will be done on the datasets to generated the Pandas dataframes.

From there our group will clean the data and generate processed versions of the dataframes which will be exported to .csv format and it will be imported into a SQL database through the use of PROGRESql on pgAdmin software.

Datasets will be taken from the following sources:

1. <https://data.world/popculture/imdb-5000-movie-dataset>

(generated by a third-party scrape of imdb website)

1. <https://www.kaggle.com/ashaheedq/video-games-sales-2019/>

(generated by a third-party scrape of vgchartz.com)

1. <https://en.wikipedia.org/wiki/List\_of\_films\_based\_on\_video\_games>
2. <https://en.wikipedia.org/wiki/List\_of\_video\_games\_based\_on\_films>

* Team
  + Daniel Cebula
  + William Lim
  + Rohan Chaudhari
* Objective:

To establish relationship between movies made into video games and vice-versa till now, using databases from Wikipedia, Kaggle and data world.

* Finding data :

1. Data Sources

As per project proposal we decided to source data from 3 resources mentioned below

* + **IMDB 5000 Movie Dataset**: (28 Variables for 5043 movies Spanning Across 100 years in 66 countries) by Chuan Sun [(@sundeepblue on Github)](https://github.com/sundeepblue/movie_rating_prediction) using scrapy:

<<https://data.world/popculture/imdb-5000-movie-dataset>>

* + **Video Games Sales 2019**: (There are 55,792 records in the dataset as of April 12th, 2019, scrape of vgchartz.com) by Abdulshaheed Alqunber.

<https://www.kaggle.com/ashaheedq/video-games-sales-2019/>

* + List of Films based on Video games and Video games based on films (by pandas .read\_html()): <https://en.wikipedia.org/wiki/List\_of\_films\_based\_on\_video\_games>

<https://en.wikipedia.org/wiki/List\_of\_video\_games\_based\_on\_films>

1. Data transformation:
2. Data Cleaning & processing:
   1. Extract:

Please look at the Clean\_ProgreSQL.ipynb jupyter notebook for detailed extraction details. Here is a summary only:

Video game and movie datasets from data.world and Kaggle were downloaded. They were read into pandas dataframes using .read\_csv(). The columns were dropped, renamed and any rows with selected null values were dropped. The dataframes were previewed in the jupyter notebook.

The films based on video-games and video-games based on films were extracted from Wikipedia using pandas .read\_html(). The columns were renamed, the [] links were removes as well as () references were removed as well. The extraction and transformations and loading was done in the Clean\_ProgreSQL.ipynb jupyter notebook.

Formats:

Data.word – CSV

Kaggle – CSV

Wikipedia – HTML

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

* 1. Transform:

1. All unwanted columns were dropped and dataset was stored according to common column nomenclature. All numeric values reformatted.
2. Challenge was to get dataset in the correct format, which was originally with many brackets and discrepancy. Using ‘Regex’ function those brackets were removed along with formatting of the data.
3. Data from Wikipedia came in three different files and hence they were cleaned and formatted thrice differently as mentioned and processed in Jupiter notebook file. At the end these datasets were appended together by keeping column terminology common, removing null values and unnecessary columns. Finally dataset was created for further processing comprising of HTML, data.world and Kaggle datasets.

Wikipedia – HTML to dataframe

Documentation – CSV format.

A picture containing screenshot

Description automatically generated

* 1. Load:

Once all the data tables were in the needed format, common relation was found between video games made till now in respective years along with movies matching on keywords, reflecting upon movies made into video games and vice-versa.

1. With PgAdmin, ‘ETL\_db’ was made, loading schemas based on ‘query.sql’ file.
2. In jupyter notebook, engine was created connecting to PgAdmin database.
3. Used pandas to load cleaned csv’s converted dataframe to database (ETL\_db) using to\_sql

