**Technical Report**

For Project 2, we worked as a team to complete the ETL process on a dataset.

**Extract**

**Transform**

**&**

**Load**

After considering multiple topic areas and datasets our group decided on the topic of “gender representation within modern day video games across multiple platforms”.

**The Extract**

We found and used three datasets that focused on gender representation within modern day video games, sourced from Kaggle.com & diamondlobby.com\*. These datasets were games data, character data & sexualisation info all in the form of CSV files.

\*credit diamondlobby.com for allowing free access to the public to their datasets

**Group Communication**

To be able to complete this project effectively we ensure we had communication within our group. We put in place a few mediums firstly being opening a slack channel just with our group members in it.

Graphical user interface, text, application, chat or text message

Description automatically generated

Secondly we also created a WhatsApp group again just with our group members in it which we discuss the project during working hours.

Graphical user interface, text, application, chat or text message

Description automatically generated Icon

Description automatically generated

We also booked in extra session where we work together and on our respective parts using Team as the platform.

Graphical user interface, text, application, chat or text message

Description automatically generated

**The Project**

We created workload and separated then between the group and set deadline for review so that we had a realistic timeline to complete the project.

* + Project ideation
  + Data fetching
  + Data analysis
  + Cleaning & Transforming
  + Data loading into final database
  + Creating documentation (Technical report and Project Visuals)

After gaining all the information needed about the project we managed to break it down into manageable work chunks but also having the availability of the team to work on our work targets which we identified and is show in the process map on our git repository.

Process Map - [Project 2 Process Map](https://github.com/kass173/Project-2/blob/main/Process%20Map/Colorful%20Process%20Prjt%202.png)

**The Transform**

We had three datasets (638 character data across 65 modern day popular games)

* We joined the data set based on Characters.ID = Sexulixation.ID & Game.Game ID = Character.Game.
* The types of functions we used as part of our transform element to format the data to be able to load into our database (SQL) of the process is we cleaned select columns to ensure the dataframe in our jupyter notebook is similar to our tables within SQL over the 3 datasets and data tables. We cleaned up column names to either simplify names or renaming columns to shorter or more relevant names.
* We had to make sure that the numeric columns had no null values and fill them with zeros and string columns with nan so that the data could conform to our SQL table parameters.
* We had to change the numeric data column to date to fit sequel table requirements
* Make sure the data type in pandas agreed to the sequel datatypes form the jupyter dataframe and then connection strings
* We also sat and tested our individual work area’s together as we split the data sets between the group and once it worked we aggregated the codes into a master file.

**The Load**

The final jupyter file is Final\_jupytercode.ipynb and the sequel file final\_sequeldatabase.

The Schemata that is used in the final production of our database is saved into our group Git Hub Repository at along with the ERD with the relationships between our datasets and how we joined them together;

[Database Schema](https://github.com/kass173/Project-2/blob/main/Gen_Rep_%20Games.sql)

[Entity Relationship Diagram](https://github.com/kass173/Project-2/blob/main/ERD_Project_2.PNG)

**Final Database**

The final database consisted of 3 tables (Game, Character & Sexulisation Data). We had to ensure the primary keys used for the tables were unique.

Table 1, Games and Table 2, Characters were joined on 1 column (game.game\_id = character.game) .

Game\_id, is the primary key in table Game. In the Character table game column game character is where the datasets join hence functions as a foreign key.

Within the Characters Table we joined to the Sexualization table we linked on ID\_charecter column which is also a primary key.

game.game\_id = character.game

character.Id = sexualization.id

character.sexualization\_total = sexualization.sexualization\_total

Table1: Games, Primary Key: game\_id

Table2: Characters, Primary Key: Game; Foreign Key: Game

Table 3: Sexualization, Foreign Key: id

[Final SQL Database](https://github.com/kass173/Project-2/blob/main/final_sequeldatabase.sql)

The reason this topic was chose was due to our common interested in video games and being a majority female group when we discussed this dataset we were interested in finding if games still harboured patriarchal influences.

The data that we looked at in 2022 seems to suggest we clearly don’t have equality between male and female representation in video games yet. But, the industry is undoubtedly making improvements albeit slowly in introducing a large number of female main charecters.