ETL Project

Joon Lee, Simon Eyob, Jamie Chamberlain

Due to the pandemic and lockdown going on, we decided to use databases that had to do with entertainment that you could enjoy at home. The first database was a database from Kaggle involving video games. It contained information about video game sales in the last 4 decades. Our second dataset also comes from Kaggle but contained information about Netflix movies and tv shows. This one contained show ids, title, and other detailed information about each show/movie on Netflix.

Netflix DB Cleanup, Transformation, Load:

* Imported the pandas and sqlalchemy libraries
* Converted the dataset from a csv to a pandas dataframe
* Cleaned up data using panda modules
  + We dropped columns containing null values and columns containing information we deemed unnecessary
  + We also split the dataset into 2 categories which are Movies and Tv Shows
  + Set the index to show id because they were unique values
* Used sqlalchemy to connect pandas to sql creating an engine
  + In pgadmin, we created 2 tables with the same columns and imported the data to the sql database tables.
* We then used pandas to check to see that the information imported correctly

Video Game DB Cleanup, Transformation, and Load:

* Imported the pandas and sqlalchemy libraries
* Converted the dataset from a csv to a pandas dataframe
* Cleaned up data using panda modules
  + Split the dataset into 4 different dataframes by years using the loc function
  + After splitting the datasets, we reset the index
  + Sql automatically set the columns names to lower case, but our dataframe columns had capital letters so we had to use a function that changed all letters in the columns to a lower case (syntax: ***[c.lower() for c in vg\_1980.columns]***
* Used sqlalchemy to connect pandas to sql creating an engine
  + In pgadmin, we created 4 tables with the same columns and imported the data to the sql database tables.
* We then used pandas to check to see that the information imported correctly
* We ended up with 4 tables by decade. We did this to make the information easier to access.