# Extract, Transform, Load

# Project Report: Books Library

As outlined in the proposal, I set out to create the database of an online bookstore containing information from the ‘Books’, ‘Ratings’ and ‘Users’ datasets. The aim was to merge the different datasets to create a comprehensive database of the customers and the books they bought.

# Extraction

The Extraction process was carried out by sourcing the datasets from Kaggle in the CSV format. The first challenge I had with these datasets was that they were semi-colon (“ ; “) separated values rather than the comma separated values I’ve become accustomed to. After several attempts at reading the datasets into Jupyter Notebook using Python, I reverted to Excel to make the conversion.

When satisfied with this stage of the cleaning process, I proceeded by loading the datasets into python by saving file paths as variables and then converting into Data Frames; at which stage I carried out further cleaning using Python.

# Transform

The transformation process in Jupyter Notebook involved checking the deleting columns, renaming, checking for Null values and deleting Null rows. This allowed me to clean up the date and rearrange the columns to make them more reader friendly.

# Load

The final Data Frame was then loaded into a SQL database by first creating the ‘records.db’ data base and initialising the ‘storedata’ table. An ‘Engine’ was created using SQLalchemy establishing a connection between Pandas and pgAdmin.

*“final\_df.to\_sql(…”* The three was used to load the dataframe into SQL and read using the pd.read\_sql command.