ETL Project #1

Emily Boling

Book reviews and revenue

EXTRACT

For this project I extracted data from two open source data sites, data.world and Kaggle. I wanted to obtain two different sources of data on book reviews from different sites and pull in book sales. My original question I was trying to answer was whether high average book reviews equated to more books sold or higher profit. I found it quite hard to find free data on book sales. I did find a source on data.world that had been obtained from NovelRank.com before Amazon bought and shut the site down. It is a limited set and its dated since the site was shutdown in mid 2018. My second piece of data came from Kaggle, it contains average book review ratings from goodreads.com and information on the books themselves.

TRANSFORM

The data I extracted was in CSV format, I loaded them into pandas. I had to read the goodreads data in with encoding because there were some strange characters within. I then removed a few columns I was not interested in from each data set. I also renamed a lot of columns to be more descriptive once joined. I merged the datasets with a right join on the book titles. I dropped the na’s and renamed more of the columns, I also added an index. I re-formatted the columns pertaining to sales.

LOAD

Once I was happy with my dataframe I used mysql.connector to connect to my mysql account. I created tables to house different columns of my original dataframe. I created three tables using title as a primary key: amazon, goodreads, book\_info. I used create\_engine and sqlite to load data into the database. I also exported my dataframe to a csv file for record keeping. I chose a relational database for this data because I was merging similar fields and I wanted to continue to work with mysql.

Just for fun I plotted the gross sales verses the average rating to see if they were correlated, there does not seem to be a strong correlation between sales and ratings for this data set: