Clarence Li

ETL Project Final Report

For this project, I seek to analyze board game trends. Although board games (also known as tabletop games) have been around since the 1960s and 1970s (and probably even earlier!), there has been a resurface of the popularity of board games despite the rise of technology and video gaming among family and peers. The goal of this project is to analyze that rising trend in the 21st century.

To start, board game data has been extracted from BoardGameGeek’s data on their website via Kaggle (<https://www.kaggle.com/gabrio/board-games-dataset>) of which an SQLite file was obtained. BoardGameGeek is a popular website that provides descriptions, rankings, and reviews for many board games. Kickstarter data, a website known for crowdsourcing projects (especially board games), was also extracted from Kaggle (<https://www.kaggle.com/kemical/kickstarter-projects>) of which a csv file was obtained.

In order to transform the datasets from BoardGameGeek’s SQLite file, SQLAlchemy was utilized. An engine and connector were created to connect to the SQLite file, and an inspector was created to get a quick view of all related tables and table columns needed for analysis. Afterwards, the relevant SQLite table was converted into a pandas DataFrame, and cleaning functions such as df.iloc and df.drop were used to clean up the DataFrame. Df.info() was used to collect all the names of the columns to assist in dropping the unnecessary columns. Also, one thing to note was that the column names for the BoardGameGeek datasets had periods in the variable names, so every column had to be renamed by replacing “.” With “\_” to make python coding much easier.

In the Kickstarter CSV file dataset, the file was converted into a pandas DataFrame via pandas. Since Kickstarter is known for projects other than just board games (tabletop in this case), df.unique was utilized to only identify tabletop categories, especially the ‘successful’ projects since failed projects will not help with this analysis. Further cleansing was performed to drop unnecessary columns in the Kickstarter DataFrame.

Finally, both BoardGameGeek and Kickstarter DataFrames were combined via the pd.merge() function. In order to merge those two DataFrames, I had to make sure that the column names that describes each board game names were matching. So, the ‘name’ column from the Kickstarter DataFrame was changed to ‘details\_name’ to match that of the BoardGameGeek column name. It is there that allows both DataFrames to be merged. Once completed, the combined DataFrame was loaded into an excel file called ‘output.xlsx’.

There are many analyses one can perform given the provided combined DataFrame. Also, more data can be collected from other sources, such as extracting the ratings of board game items sold on Amazon.com using web-scraping techniques. Listed below are some examples of possible board game analyses:

1. Identifying the trend of when each board game was published to analyze the rising popularity and growth of the tabletop gaming genre.
   1. Via details\_yearpublished column from BoardGameGeek
   2. Via launched (or deadline) column from Kickstarter
2. Identify what board game category is most popular.
   1. Ex: stats\_family\_strategygames\_bayesaverage for 1-10 rating and stats\_family\_strategygames\_pos for BGG ranking in the strategy game category
3. Identify what type of board game do most people prefer.
   1. Via the suggested number of players criteria: ex: polls\_suggested\_numplayers\_3
   2. Via average time length of board game to play: details\_playingtime
   3. Via board game difficulty level which BGG gives 1-5 rankings from easy to hard: stats\_averageweight
4. Identify the popularity of new board games from Kickstarter projects.
   1. Via the number of backers and US dollars donated to each board game, ex: backers vs usd pledged