ETL Project  
Team Members: Jacob Lingwall and Jie Fan

Data Sources:

Kaggle Board Game Data Set: <https://www.kaggle.com/gabrio/board-games-dataset> (sqlite)  
St. Louis Pieces Game Inventory: <http://www.stlpieces.com/our-games/> (csv)

Data Cleanup & Analysis:

Looking through the St. Louis Pieces inventory webpage, their games table is a Google spreadsheet. The data was copied and pasted into Excel and saved as csv file. The Kaggle Board Game data set was downloaded as a sqlite database.

We used sqlalchemy to determine the tables and columns within the database. There were 5 tables in the dataset and the relevant table to our project was BoardGames. We initially removed the expansion games. The table has 81 columns and mostly were statistical analysis done by the creator. We used pandas to select only the columns needed and created two data frames, game\_details and game\_attributes.

The St. Louis Pieces inventory dataset did not have any unique identifying number to match with the game details and game attributes. We chose the game\_id, found in the Kaggle dataset, as our primary key. We imported the St. Louis Pieces csv and a newly created csv containing game id and name into MySQL via table import wizard. An inner join was performed on the two tables to create a new table where the St. Louis Pieces game inventory will have a corresponding game id. Inner join was used so that there would not be any nulls in the id column.

Its important to note that since this is an exercise we did not scrape all ninety thousand plus games from the Board Game Geek site. We captured prices for roughly 4000 games (zero to many per game) in the second hand market between December 20th and 21st.