Data Wrangling

Objective: use available data to build recommendation models.

Dataset: There are two datasets used in this project. Both datasets are downloaded directly from https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data. The datasets contain content about Australian steam information. The first dataset ("australian_user_reviews.json") contains user information such as user id, user url, user review, etc. The second dataset ("steam_game.json") contains information on steam games such as item id, genres, specifications, publishers, etc.

Cleaning and Forming the Data: Because I do not need all the features in both data, I had to extract what is necessary. From "australian_user_reviews.json" dataset, after transforming the data into dataframe, I extracted user_id, recommend, item_id, and review to form a new dataframe. The new dataframe contains 4 columns and 59305 in total with no null values. For the second dataset "steam_games.json", I had to extract values of each key into lists and build a new dictionary due to complications of file's quality. Once I got the dictionary I turned it into a dataframe using pd.DataFrame. The second dataframe contains 13 columns and 32135 entries with many null values. I left joined the two datasets together on column item_id. The new dataset contains 16 columns and 59305 entries in total. To deal with null values, I used values from column tags to fill missing value in column genres. For the price column, I fill the empty cell with the mean price. For the early access column, empty cells are filled with "False". The missing value at the column title is filled with values from the column app name. I also dropped the column discount price because it only contains about 200 values. Finally, I dropped all rows that exist empty values that cannot be otherwise filled in. The clean dataset contains 15 columns and 49704 entries.

Saving the Data: The final dataframe is saved as a csv file named “data” in the directory for later analysis.