Board Game Geek Database Clustering

# **Neil Opitz Summer 2021 https://github.com/FarmFresh-code/DSC680**

# Which Domain?

The data that I will be using for DSC680 Project #1 is data from the Board Game Geek database.

References:

https://medium.com/@galen.ballew/board-games-meet-machine-learning-34026870f8d5

Using SciKit-Learn for analysis of board game reviews

https://www.youtube.com/watch?v=I0WB0rRjO48

Using a linear regression algorithm for board game ratings analysis

https://rentzb.github.io/post/cluster-analysis/

Utilizing categorical variables in a clustering algorithm

# Which Data?

Dataset link: https://www.kaggle.com/seanthemalloy/board-game-geek-database

The dataset for this project was downloaded from the Kaggle website. The data is comprised of information about specific attributes of the game as well as user ratings for the game. In the absence of a codebook for this dataset, I have explained the contents below:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| name0 | The title of the game |
| average/stdev/bayesaverage | User ratings for the game |
| usersrated | Number of ratings for the game |
| weight/numweights | Physical weight of the game |
| minplayers/maxplayers | Number of players that the game can accommodate |
| playingtime/minplaytime/maxplaytime | Standard duration of game (range) |
| owned | Number of users that own game |
| wishing | Number of users that added game to their wishlist |
| wanting | Number of users that want the game |
| trading | Number of users that own the game and are willing to trade it |
| numcomponents | Number of different components in the game |
| yearpublished | Year the game was first published |
| abstractrank/ accessoryrank/ boardgamerank/ childrensrank/ customizablerank/ f amilyrank/ partyrank/ strategyrank/ thematicrank/ warrank | Ranking for game within categories |
| name1-9 | Alternate names for games (usually foreign versions) |
| designer0-9 | Name of game designer(s) |
| artist0-9 | Name of game artist(s) |
| subdomain0-2 | Alternate game category |
| category0-9 | Descriptive game categories |
| mechanic0-9 | Mechanics used in game |
| family0-9 | Descriptive family of game |
| publisher0-9 | Game publisher(s) |
| honor0-9 | Awards the game has won |
| expansion0-9 | Published expansions for primary game |
| version0-9 | Published version |
| implementation0-9 | Alternate versions derived from game |
| accessory0-9 | Accessories for game such as promotional items |
| compilation0-9 | Compilation of games that game belongs to |
| commerceweblink | Website where game is sold |
| description | Brief description of game |
| thumbnail | Link to small picture of game |

# Research Questions? Benefits? Why analyze these data?

I will start with exploratory data analysis and make decisions on how to deal with outliers and missing data. I will then explore the distributions for the primary variables that I am planning to use in the analysis. The dataset is very large at approximately 100,000 records, so I should have good representation for categorical variables.

Some questions that I have regarding this dataset include:

What are the attributes common to games that are in the higher performing clusters?

Do the highest rated and most popular game clusters have a higher proportion of the top game designers?

Are there clusters that have a higher proportion of games that were funded on Kickstarter?

Can a cluster of game ‘flops’ be identified with this method?

# What Method?

My plan is to use a clustering algorithm to identify clusters of games based on ratings and popularity then compare attributes of the games among clusters to identify if there are attributes that are more common to specific clusters. I am planning to scale the ratings and number of users data in order to obtain a visual that more clearly communicates the relationships and clusters in the dataset. I think that using cluster analysis will provide an interesting way of identifying groups of games from among the thousands in the database that have similar attributes within the group but have some attribute that varies from other clusters and to identify what attributes are driving cluster differentiation.

# Potential Issues?

It is possible that the data may not be distributed in a way that allows for clusters to be easily differentiated visually. I am hoping that after scaling the data and developing a scatterplot that the data is dispersed across the plot with some areas containing larger concentrations of data points (some clusters visible prior to applying the clustering algorithm.

Another challenge I may have is that the results may not be very interesting. It is possible that the clusters will all have the same attributes and the only thing differentiating them is the ratings. Since the rating data is basically driven by subjective assessment, there may not be any interesting relationships in the attribute data based on ratings.

# Concluding Remarks

I am an avid board game player and have a great interest in data about board games. I just learned this week that the database for the Board Game Geek website was available in Kaggle. Board Game Geek is the most popular website for information about board games and I have been visiting the site for more than a decade and have also contributed data to the website. I am using a clustering algorithm with board game attributes and ratings with the goal of finding games that are statistically similar to each other from an attribute and user perspective.