# Final Report
# Popular Game Mechanics - Capstone 3

## Original Hypothesis

The original business use case asserted that it would be possible to determine which game mechanics are most popular and lead to the most profitable games based on customer reviews.

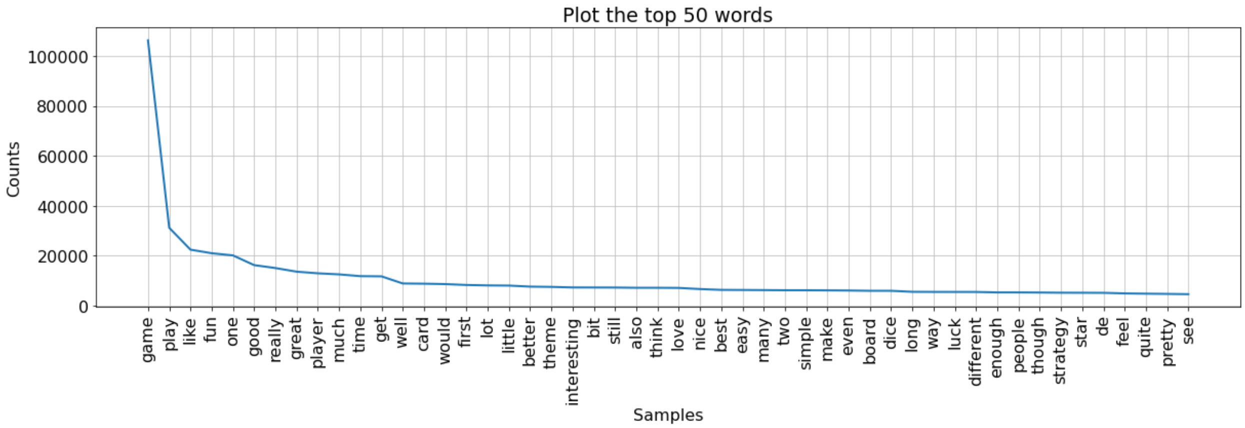
## Data Wrangling

Data was obtained from boardgamegeek.com by way of kaggle.com. The data set was posted by Jesse Van Elteren at [https://www.kaggle.com/datasets/jvanelteren/boardgamegeek-reviews](https://www.kaggle.com/datasets/jvanelteren/boardgamegeek-reviews) and captures the reviews of 350,000 unique users as of January 2022. The data included only games that had received more than 30 reviews, which brought the total of reviews to just under 19 million.

The dataset was unwieldy to work with locally and via Google CoLab, since the dataset was greater than one gigabyte. The data was exceptionally clean, so the main focus of data wrangling was stripping all unnecessary data to make the dataset easier to work with. The URL, thumbnail, user name, and users rated columns were removed. All reviews with number-only reviews and no text were removed, since it was the comment text under analysis. That took the dataset from 18,964,810 reviews to 3,368,619. Before running preprocessing, the data had to be further cut down to only analyze 20% of the data, and after non-English reviews were removed, the data was further cut down to analyze 25% of that set. This brought the number of reviews to 137,364. The mean of ratings remained at 6.88, even with the reduction in the size of the data set, and the 25%, 50% and 75% quartiles remained at 6, 7, and 8 star ratings, respectively. The standard deviation between the dataset of 3.3 million rows was the same as that of the reduced dataset, at 1.8.

## Exploratory Data Analysis

Exploratory data analysis was particularly useful for determining stop words during preprocessing. Words like “game” and “play” are used so commonly throughout the reviews that they do not lend much usefulness. Words like “interesting,” “fun,” and “like” would have been useful in the case of a sentiment analysis, but do not lend any information to the game mechanic analysis.





The most useful information to come out of the EDA was the results of the bigram and trigram association measures. When seeded with words like “mechanic,” “favorite,” and “theme,” these yielded the kind of game mechanic information that would be useful. Phrases like “drafting mechanic,” “auction mechanic,” “traitor mechanic,” “favorite worker placement,” “favorite area control,” “favorite easy learn,” “theme easy teach,” “deck building theme,” and “theme tile laying” indicated that the reviews had the data at the level of detail desired.

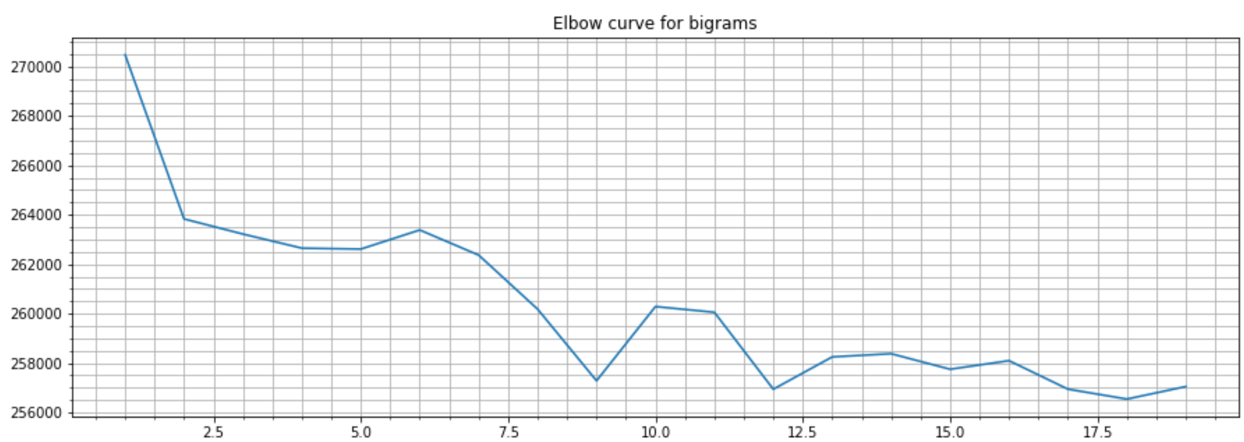
## Model Selection

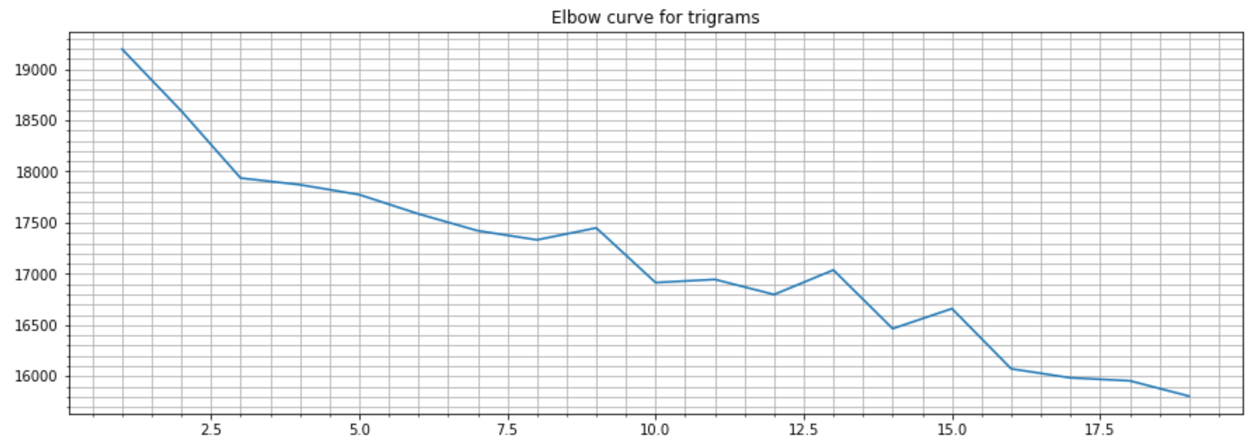
Three NLP models were used to examine the review test: KMeans, NMF, and Latent Dirichlet Allocation. Before running the text corpus through the models, all non-English reviews were stripped out. The data was split into bigrams and trigrams to yield more contextual meaning. Emoticons were also removed along with reviews that were digits only. The corpus was then split into test and training sets and run through a Count Vectorizer and a TF-IDF Vectorizer.

### KMeans

After running the corpus through KMeans the first time, the clustering of bigrams and trigrams that resulted showed clearly that more words needed to be added to the stop words being filtered out with the vectorizers. The decision was also made to drop the 'max_df' parameter for the vectorizers from 95% to 90%. Two of the three clusters were otherwise meaningless. For instance, one of the clusters consisted of reviews that contained bigrams with variations on "star star, half star, no star." There were a surprising number of these, which were meaningless without a much larger context. One assumes that they were part of larger phrases like "I gave this game no star for bad directions" or "one star for the high quality playing tokens." However, the utility of these longer phrases dropped out with just bigrams of "half star."

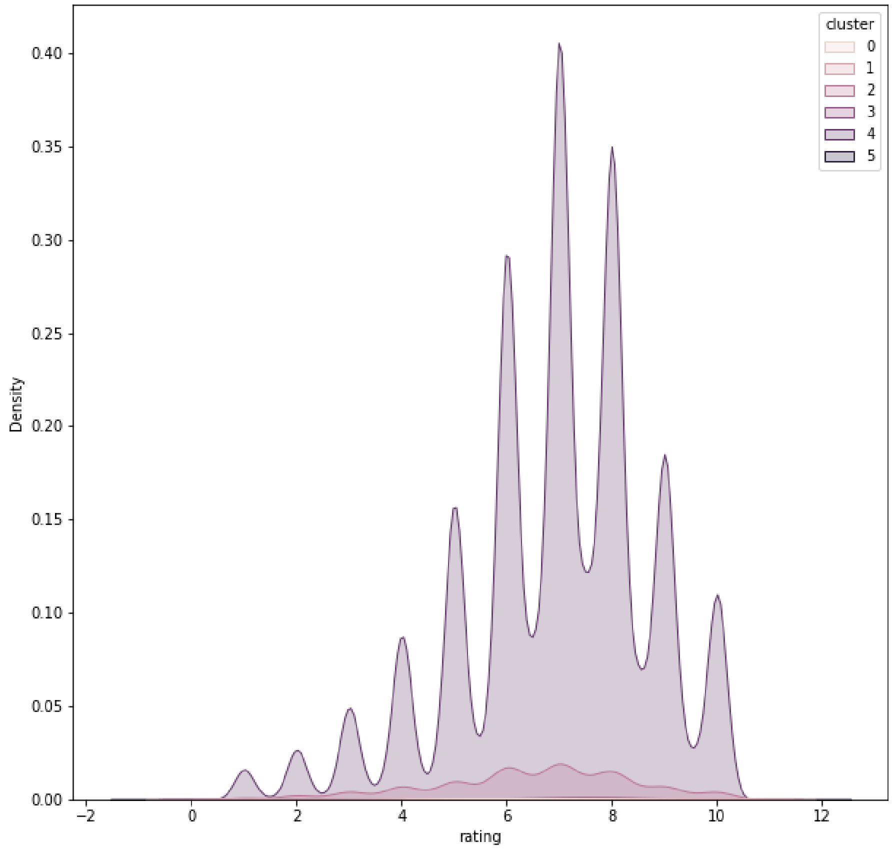
Once the stop words set was increased and the max_df percentage was dropped, the elbow curves determining the ideal number of clusters for KMeans lacked the ideal smooth elbow curve.






Running PCA to reduce the components did not yield promising results. When run without the components specified, the algorithm eventually crashed due to a memory overrun. Once the 'n_components' parameter was limited to two, the explained variance ratio showed why open-ended component numbers failed to run. The explained variance ratio numbers for the first two components were 0.025 and 0.012 once the stop words were refined.

Once the cluster IDs had been generated, dataframes were created comparing the distribution of ratings by the cluster IDs. Plotting these demonstrated that the clusters overlap almost completely and that one cluster had an overwhelming share of the ratings. The preferred outcome, showing specific topic clusters assigned to each rating, were not present at all. If that had existed, it would have allowed us to state that specific topics were associated with specific ratings.




At this point, it was determined that KMeans was not an ideal modeling approach to this corpus without significantly altering what data was being examined or running a sentiment analysis first. Since this would negate the hypothesis, the focus shifted to the topic modeling tools.

**NMF (Non-negative matrix factorization)**

The data from the Count Vectorizer and the TF-IDF Vectorizer were run through NMF for both bigrams and trigrams. From the corpus generated from the Count Vectorizer, several of the topics focus on words like: "don't know think mind care feel people don't enjoy players don't understand" and "bad great lot quite interesting theme isn't long time" which are of very limited utility. Some useful topic clusters were generated, though, with topic groupings like "worker placement resource management great worker victory points solid worker placement resource engine building player interaction," "area control player interaction deck building push luck set collection" and "easy teach non-gamers learn new players quick easy replay value." Results were very similar when run against the TF-IDF corpus, with the most useful topic clusters being "worker placement player interaction solid worker resource management dice rolling engine building deck building resource placement" and "area control set collection deck building hand management tile laying card drafting."

**LDA (Latent dirichlet allocation)**

Like NMF above, the data from the Count Vectorizer and the TF-IDF Vectorizer were run through LDA for both bigrams and trigrams. LDA yielded the most useful topic clusters, with the least amount of noise. For the Count Vectorizer corpus, topic clusters included "deck building, looking forward, player interaction, trick taking, push luck," "worker placement, simple rules non-gamers, great theme, simple years ago," and "area control, worker placement, player counts, paths victory." These topic clusters all came from the bigrams dataset, while the trigrams dataset had less meaningful clusters like "don't know, don't think, don't mind, cards, year old son, didn't think."

Results with LDA on the TF-IDF Vectorizer corpus were similar to those from the Count Vectorizer. While the trigram dataset had topics of limited utility as above, the bigrams dataset yielded the most semantically meaningful clusters: "trick taking, push luck, deck building, right group," "worker placement, non-gamers, great theme, simple rules, house rules, best number players," "area control, great family thing, hand management, roll write." Interestingly, in the analysis of the bigrams, the names of two specific games were popular enough to show up in the topic grouping: Puerto Rico and Ticket to Ride. This affirms the common wisdom that these two games are often considered good entry games into the hobby.

## Findings

LDA proved to be the most useful model. Because the data didn't cluster well to the rating numbers, a sentiment analysis would be required before running all the modeling detailed above to determine the sentiment of reviews before correlating them with the rating number.

Based on the LDA topic modeling, however, the game mechanics and aspects that receive the most focus, good or bad, are:

- Game mechanics:
    - Deck building
    - Trick taking
    - Push your luck
    - Worker placement
    - Area control
    - Set collection
    - Hand management
- Other important game elements:
    - Player interaction
    - Easy learn / easy teach

- Right group
- Player count
- Family game

## Recommendations and Future Research

**Data suggestions:**

Including a field in the data for language and another field for tracking “official” game mechanics present in each game would make the data more useful and easier to analyze. If fields were included for game mechanics, then those mechanics could be correlated with number ratings, and the subjective issues with the ratings could be avoided entirely.

**Future analysis suggestions:**

KMeans might be a more useful algorithm if sentiment analysis was run before clustering via rating numbers. Further revision of stop words and converting all contractions would need to be performed first, since a lot of reviewers used words like “didn’t” or “doesn’t”.

It would be worthwhile to obtain space on AWS to have larger memory and processing power for the dataset. More than half the time spent on analysis was on reducing the size of the dataset, waiting for algorithms to run, and recovering after memory overrun crashes.