Steam Sentiment Analysis and Gameplay Element Correlation

Unsupervised Learners
Group Leader: Paul Rosa
Group Members: Max Baskin, Hudson Langham
DS 4002
February 10th, 2025

Hypotheses:

- The sentiment scores for Steam games with a majority-positive rating will be above the neutral score (-0.05 to 0.05), and below the neutral score for games with a majority-negative rating.
- Positive sentiments towards game components such as story, visuals, gameplay, and more will positively correlate with ratings.

Research Questions

- How do review sentiments as analyzed by our models relate to Steam ratings?
- How do positive sentiments towards individual game components affect overall game ratings?

Model Approach

- We will preprocess the Steam review data by cleaning and preparing the text for analysis.
- We will then use a pre-trained sentiment analysis model, such as BERT or RoBERTa, fine-tuned on our dataset to classify the sentiment of review texts.
- To identify gameplay elements, we will attempt to implement NER, an NLP, in our model to help identify what elements make up a game.
- We will use statistical methods like logistic regression to analyze the correlation between identified gameplay elements, sentiment scores, and game recommendations.
- Visualize our findings using techniques such as heatmaps or network graphs to illustrate the relationships between these factors.

Executive Summary

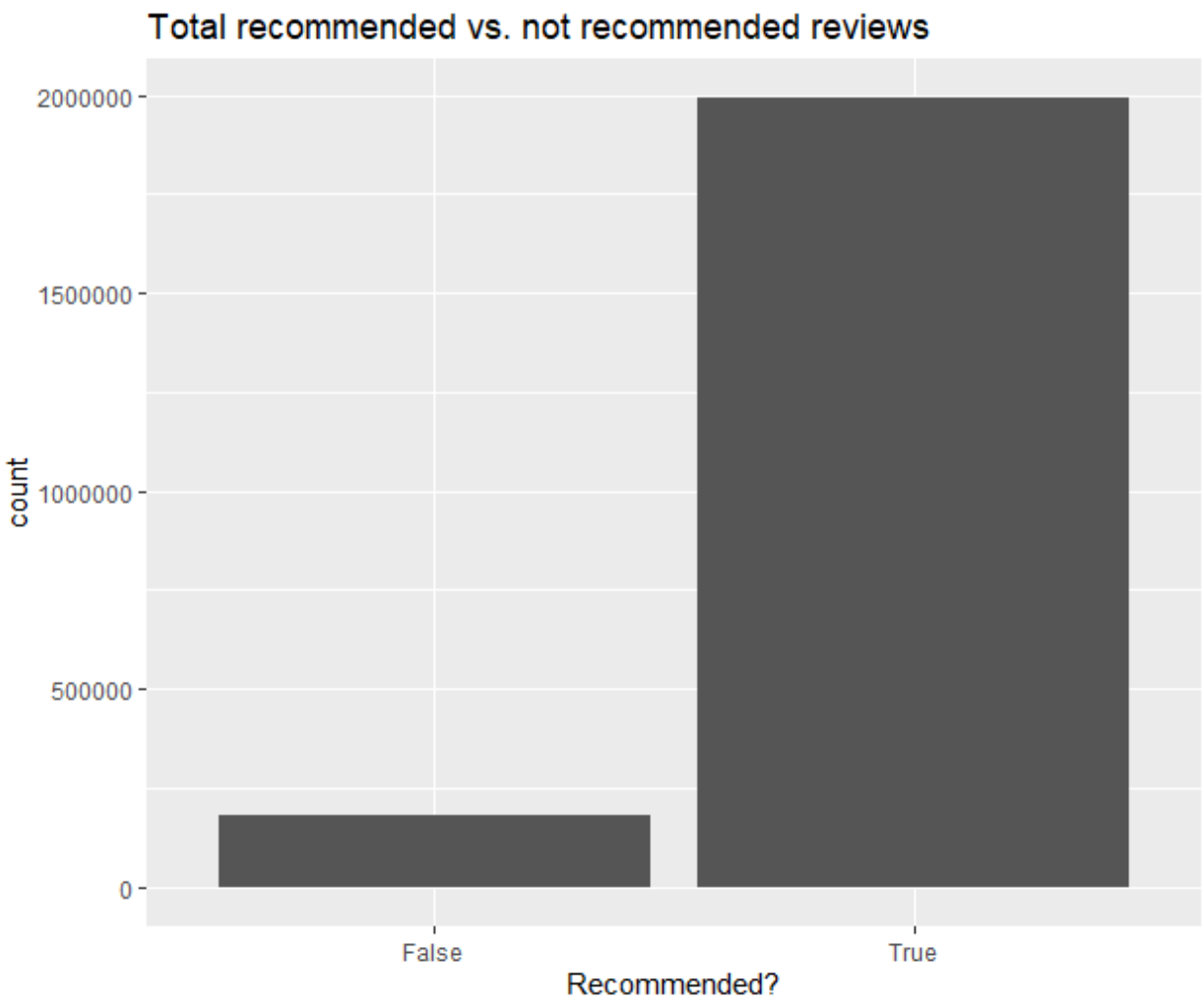
This document outlines our data set and analysis plan for analyzing Steam review data to determine the relation between review sentiment and review ratings, as well as review sentiment towards specific gameplay elements and review ratings.

Data Set Establishment Details

- **Goal**: Our dataset will be gathered from a Steam Review Dataset on Kaggle and preprocessed to suit our needs.
- **Summary of Established Data Set**: The dataset contains approximately 21 million reviews from games across the steam platform. We will likely use a subset of the data for processing efficiency.
- **Data Dictionary:**

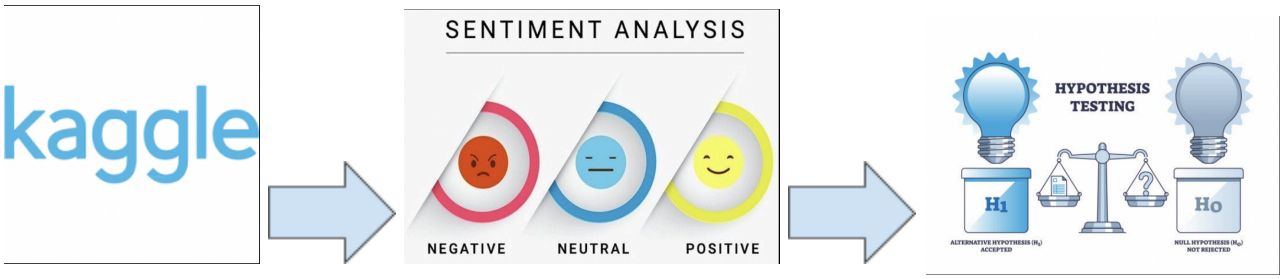
Column	Description	Value
review_id	Unique identifier for each review	*xxxxx*
game_id	Unique identifier for each game	*xxxxx*
game_name	Name of the game	Counter-Strike: Global Offen...
review_text	Full text of the review	"This game was so cool I love CS:GO thank god they made it free its the best"'
review_score	Binary score of (0 for not recommended, 1 for recommended)	0, 1
review_votes	Number of users who found the review helpful	*x*
review_language	The language the review was in	*English, Spanish, **

- **Data is on Kaggle**: https://www.kaggle.com/datasets/najzeko/steam-reviews-2021
- **Questions Explored in EDA:**
    - What is the distribution of review scores?
    - What are the average sentiments for each game?
    - What is the distribution of the length of comments?
- **Refinement of hypothesis, research model/plan:** Not yet since we haven't had the chance to look through all the data and clean anything that isn't eligible nor decide on our approach with languages. Research model still stays the same for now and may be tweaked a bit after our sentiment data is finished.
- **Current Unknows:** How playtime may be a factor since it's not a column in our data set.
- **Goal, put another way:** Build a model that can extract, analyze, and relate game elements to review score on the Steam platform, providing insights into what drives player satisfaction.



- As taken from reviews labeled as English-language.

Analysis Plan

- **Goal**: Use steam data from Kaggle and analyze the sentiment
- **Graphic:**
- 

- Get data on ratings and sentiment from the kaggle dataset
- First analyze the sentiment using BERT
- Hypothesis Testing:
    - Majority-positive rated Steam games:

        - H0: Mean sentiment score for positively rated games is less than or equal to the neutral score
        - H1: Mean sentiment score for positively rated games is greater than the neutral score
    - Majority-negative rated Steam games:
        - H0: Mean sentiment score for negatively rated games is greater than or equal to the neutral score
        - H1: Mean sentiment score for negatively rated games is less than the neutral score
    - Split into groups one being Majority-negative rated games and the other being Majority-Positive rated games
    - Perform t-test for each group against the neutral score
- Next analyze the sentiments of gameplay components effect on ratings
    - Get data on ratings, and game component sentiment from kaggle dataset
    - Apply NER to identify key gameplay elements within the reviews
    - Define a custom set of game related labels for the NER model (e.g. Character, Ability, Weapon, Location)
    - Use a BERT sentiment analysis model to evaluate the sentiment associated with each identified gameplay component
    - Hypothesis testing:
        - H0: There is no significant relationship between positive sentiment towards game components and positive overall game ratings
        - H1: Positive sentiments towards game components are significantly associated with positive overall game ratings
    - Perform t-test against the neutral score
- **Paragraphs:**
    - Step 1: Data Collection
        - The dataset with approximately 2,100,000 english reviews then cleaned and analyzed for sentiment analysis
    - Step 2: Sentiment Analysis
        - Each review is analyzed using a BERT sentiment analysis model. From this positive, negative and neutral sentiment scores for each review will be appended to the dataset.
        - The game component sentiment scores will also be appended to the dataset, after which the dataset will be complete
    - Step 3: Hypothesis Testing
        - Our hypothesis will be tested using the final dataset

- Quantifiable Goal: All of the approximately 2,100,00 english reviews in the dataset will be analyzed.

References

https://www.kaggle.com/datasets/najzeko/steam-reviews-2021