Fei Guo
Victor Li
Jonathan Lin
Xinyu Zheng

Preliminary Results

Preliminary analysis and summary statistics Produce tables and graphs that show exactly what data you have, and that contain summary statistics about the data. Questions to answer for each data source include:

How many unique observations to you have?

We have 3 websites, IGN, Gamespot, and Giantbomb that we've scraped and of those 3 websites, have some of the following unique observation counts:

Reviewer Counts:

Gamespot:	39
Giantbomb:	11
IGN:	49

This aligns with our knowledge of the companies as Gamespot and IGN are larger operations while Giantbomb is a smaller editorial team

Games Reviewed:

Gamespot:	387
Giantbomb:	639
IGN:	275

It is possible that the same game has been reviewed by all three sites. The unique observations are per site and may include reviews of a game that has been released on multiple platforms, which is significant.

Average Score out of 100:

Gamespot:	65.62
Giantbomb:	70.50

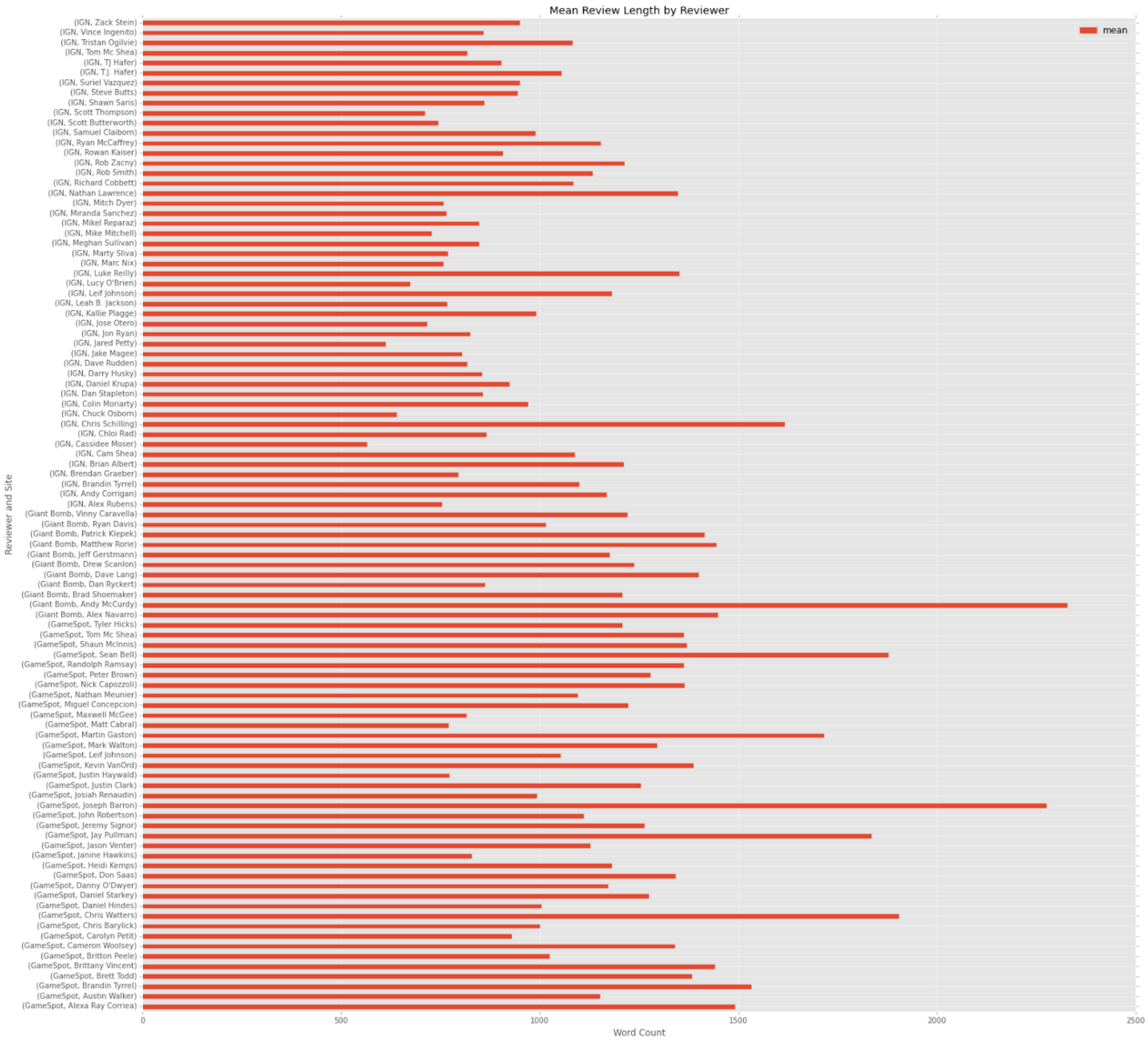
Score were converted to the 100 point scale because each site employs a different scale. We are still working to gather scores for IGN, which appear in a different section of the review.

What information/features/characteristics do you have for each observation?

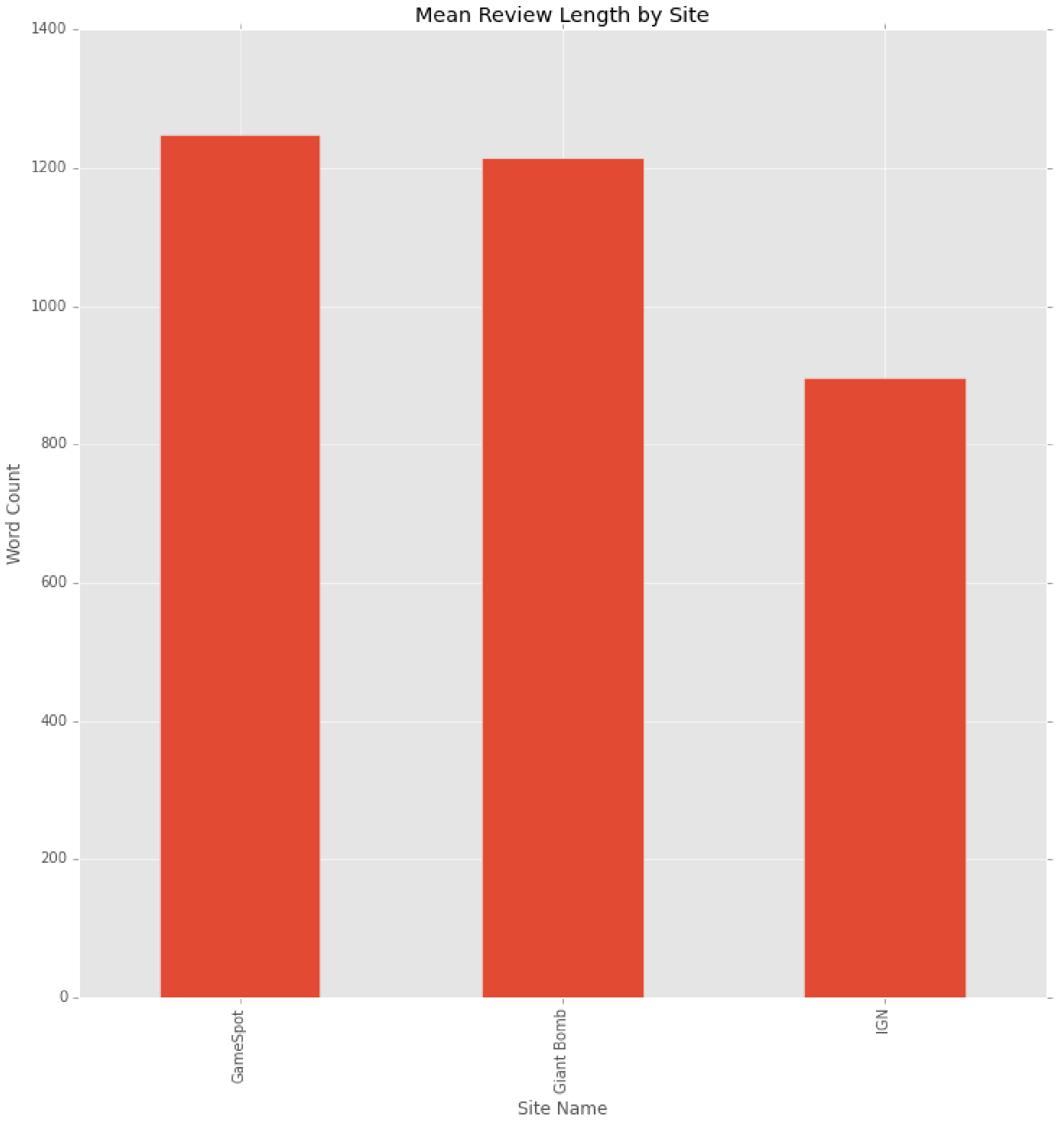
For each observation, we have the date of the review, the name of the game, a link to the review, what platform the game was released on, review text and word count, the site's score, that same score converted to 100, and the name of the website.

What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?

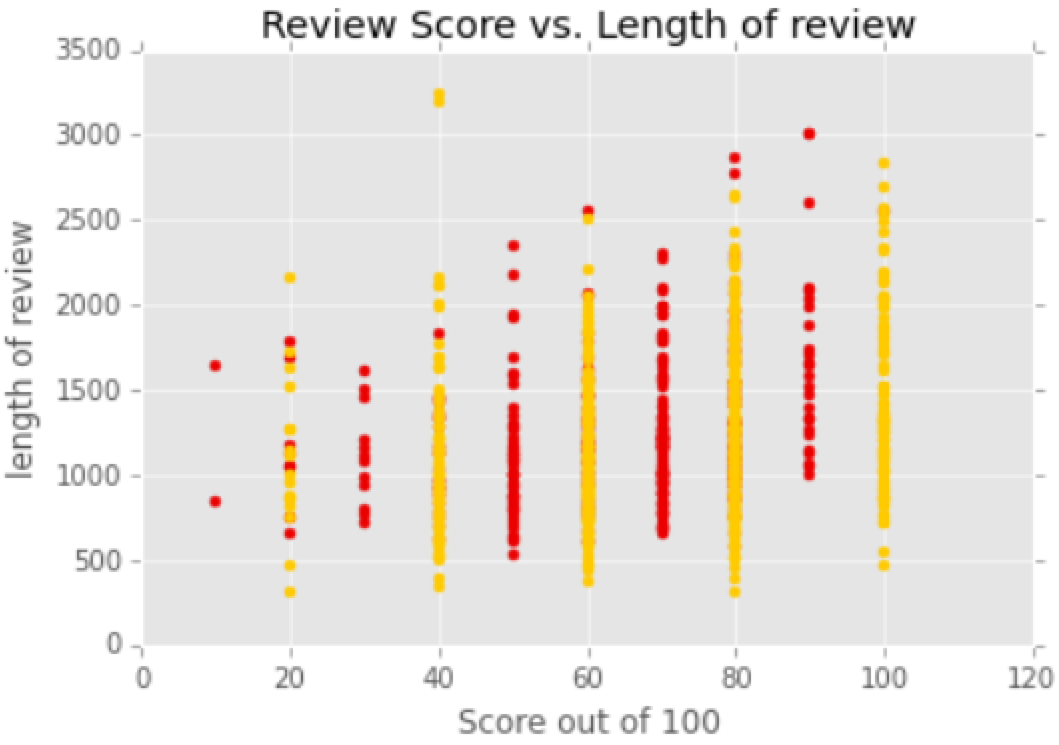
We computed the average review length of a text review for each of the reviewers of each site. The results are presented in a chart bel



An average reviewer from any of the 3 sites was around 900 words or so, with the most wordy reviewer logging in around 2,250. This may be skewed due to sample size for some of the reviewers. We also calculated the average review length for each of the sites with the results presented in the bar graph below.



We then compared the review length to the score a game received. The results are presented on this plot below.



This plot shows that longer reviewers don't necessarily translate out to higher scores.

What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)

Textual analysis could possibly reveal what kind of words and phrases are associated with positive and negative reviews. We will need to filter out games that have scored well and those that have scored poorly to create a random sample for that area of analysis.

Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.

- Perform LDA on the text of the reviews on a per site basis.
- Extract scores from IGN dataset into main dataset.
- Perform LDA on high and low rated review text reviews.
- Compute Jargon Distance for some of the reviewers within certain sites and against other reviewers from other sites.
- Produce more visualizations based on the results.