Analyzing Readability of Yelp Reviews with Respect to Business Data

Project Goals:

We sought to create an interactive visualization where the user could explore the relationship between the readability of Yelp reviews, and the attributes associated with those reviews (e.g., numbers of stars, price, city, etc.). We define readability by a number of commonly used metrics, which tend to use a combination of text metrics such as average word length and average sentence length.

Some task categories we expect to user to be able to do efficiently are:

- Compare the distribution of a readability metric within a certain category (e.g., price)
- Filter the dataset based on city and business types
- Change the variables plotted on the axes
- Change the plot type

Below are two examples of tasks that should be easily achieved with our visualization.

Example 1:
A user would like to look at automotive shops in Las Vegas only, and would like to see how the price of the automotive shop affects the readability of the reviews.

Example 2:
A user would like to look at restaurants, and would like to see how the readability of reviews varies among the cities included in the dataset.

Dataset / Libraries Used:

We used data from the Yelp dataset challenged found here:
https://www.yelp.com/dataset_challenge

We used the 4.1 million reviews contained in the dataset, along with the data on the business associated with each review.

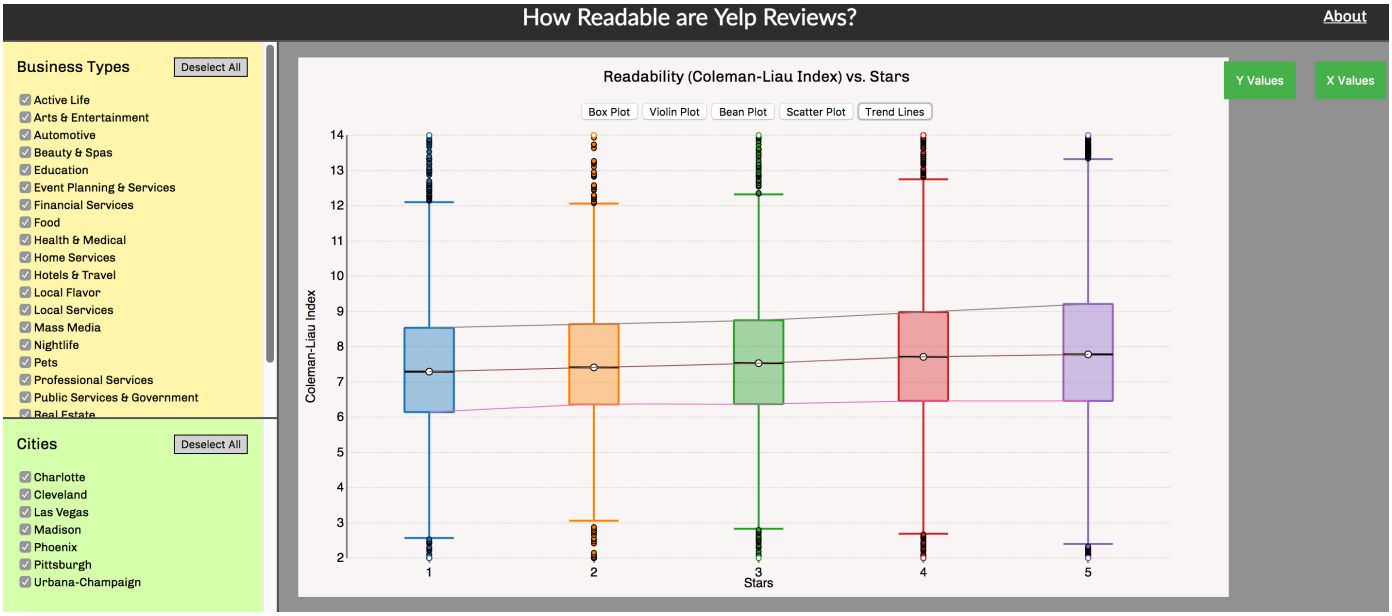
To analyze the readability of the review texts, we used the textstat Python package found here:
https://pypi.python.org/pypi/textstat/0.1.4

To produce the plots, we used a d3 library found here:
http://bl.ocks.org/asielen/92929960988a8935d907e39e60ea8417

An important source of inspiration for the project, which analyzed the readability of Tweets, is found here:
https://arxiv.org/abs/1401.6058

Description:

A screenshot of our visualization is shown below.



The main focus of our visualization are the plots shown at the center of the page. These plots allow the user to visualize the distribution of the data selected in a variety of ways by using the buttons at the top of the graph to switch plot type.

The two panels on the left of the page allows the user to filter the dataset, and the dropdown boxes on the right allow the user to change the variable plotted on the axes.

Our visualization has four main levels of interactivity:

- Filtering the dataset by “Business Type” and “City”
- Changing the readability metric plotted on the y-axis (e.g., length, Flesch reading ease)
- Changing the attribute plotted on the x-axis (stars, price, or city)
- Changing the type of plotting method used (e.g., bar plot, violin plot, scatter plot)

Encoding:

The encoding of the data varies on the type of plot being used. For all the plots, the position on the x-axis is used to separate groups of categorical data, while the y-position is used to represent the readability metric.

Box plot:
The box represent the region between Q1 and Q3, with the line in the middle being the median. The points outside of the box are outliers, defined as outside 3*IQR.

Violin / Bean plot:
The violin and bean plots use kernel density estimates to show the distribution of the data. We provide a bandwidth slider to allow the user to adjust the bandwidth to accurately fit the data.

Scatter plot:
The scatter plot uses the y-position to represent the readability metric, where each review is encoded as a circle.

Trend line:
The trend lines correspond to Q1, median, and Q3 for the metric plotted on the vertical axis.

Reflection:

Overall, we believe that our project successfully achieves the goals we had. While there are more features we would have liked to implement, given the time allotted we feel that our project satisfies the requirements.

Some feedback we received in the critique was that we should bin our data in order to reduce the number of points being plotted. We were not able to successfully achieve this, but instead we took a random sample of 30,000 reviews from our 4.1 million review dataset, which should be a representative sample.

Some future features we would have liked to implement are:

- viewing the review text on mouseover for the scatter plot
- make a slider for filtering review data by date
- add an option for plotting businesses on the x-axis

We feel our questions were answered well by the visualization, but our results were not very significant, although this is a result in itself.

This project was very informative in terms of learning CSS, HTML, and d3. We also learned a lot about the best ways to create elements that afford interactivity, such as sliders, buttons and checkboxes, and how to implement those elements. The biggest challenge was processing and plotting our large dataset, about 5 gigabytes in size.