CS-5630 / CS-6630 Project - YelpHelper - Project Feedback

Sateesh Tata - u0942976
Vairavan Sivaraman - u0942570

We have discussed our project idea with Frederico Rocha, Anil Kumar Ravindra Mallapur and Anil Kumar Konasale Krishna who are working on the project, Brazil - Credit Operations in public sector and the Human Development Index. The following are some of the positive and negative points,

Positive:

- The data is very diverse, lot of possible ways to integrate and categorize the data by filtering and sorting.
- Good idea regarding the page layout and the visualization selection.
- Map integration can be helpful for selection and driving the interactivity between different visualizations.

Negative:

- The data is too large. The size of one of the JSON is around 1.6GB which requires on our part to do a tremendous amount of data processing and clean up.
- The file size is crucial as it might take longer on the server to load the data while hosting it. They suggested to either break down the files into smaller parts or structure them optimally so that we could load the data quicker.
- The number of categories of Businesses is just too high which may lead to a clutter when trying to sort/visualize all of them.
- In one of the bar charts we planned, we were pointed out that if the selection of businesses are too large, it may just scale down so small that representing a stacked bar chart would be difficult.
- There should be a more optimal way of reducing the cluttering of a lot of data by either categorizing the businesses or providing multiple filter options.

From the discussion we have actually considered changing the way we are processing our data. We started data processing in multiple steps. We filtered the contents of the JSON files for the data columns that wouldn't be used in the visualization to reduce the size. We were able to reduce the size of the large JSON file from around 1.6 GB to 240 MB. And also we are planning to reduce the data from a single JSON to multiple smaller files so that the data could be loaded quickly with minimal processing during the run time for the visualizations. Also, we see the problem with implementing the bar graphs for such a large data. So we

decided to add a brush on top of the data. This would help one to be able to view the subsections of the data without having to deal with the clutter. This method also helps to better view the stacked bars for different number of ratings for the same business upon zooming.

These changes along with the proper filtering of the data we believe would essentially help to speedup the data loading the faster responsiveness of the visualizations. We are planning to order the data in sections and subsections for every filter parameter so that the filtering would be done faster.