CS-5630 / CS-6630 Project - YelpHelper - Process Book

Sateesh Tata - u0942976
Vairavan Sivaraman - u0942570

Overview and Motivation: This project is basically a simple visual representation of the yelp sample data set provided by YELP for their Yelp Dataset Challenge competition for 2016. The dataset is really vast with a numerous possible attributes and factors that could be considered to analyze and visualize the data and trends. We were mainly motivated by the variety of factors that could be considered to actually visualize such a large data. Although this might not be a standard bread and butter dataset to directly throw a bunch of numbers and statistics to the user to plug them in and use in the visualizations, we believe that the richness and vastness of the data would provide us with a challenging task to analyze, understand the visualize the data that might not be pretty straight in the most obvious ways. Some of the main ideas that were suggested with the data by YELP in their official challenge are the following,

- **Location Mining and Urban Planning:** How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?

- **Seasonal Trends:** What about seasonal effects: Are HVAC contractors being reviewed just at onset of winter, and manicure salons at onset of summer? Are there more reviews for sports bars on major game days and if so, could you predict that?

- **Infer Categories:** Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food, and vice versa? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?

- **Natural Language Processing (NLP):** How well can you guess a review's rating from its text alone? What are the most common positive and negative words used in our

reviews? Are Yelpers a sarcastic bunch? And what kinds of correlations do you see between tips and reviews: could you extract tips from reviews?

- **Changepoints and Events:** Can you detect when things change suddenly (i.e. a business coming under new management)? Can you see when a city starts going nuts over cronuts?

- **Social Graph Mining:** Can you figure out who the trend setters are and who found the best waffle joint before waffles were cool? How much influence does my social circle have on my business choices and my ratings?

These problems might not all be standard visualization techniques or provide straightforward data to plug in but are definitely interesting in many ways. We would like to push our understanding of our concepts and visualization techniques to bring out ways to be able to better show such subtle changes and properties and relations between various factors and trends.

Related Work: The websites which have quite a lot of user base always come up with some really nice and innovative approaches to visualize the information on their websites. This has primarily inspired us to pick up this particular set of data and come up with better and more cleaner visualizations. So we picked the YELP dataset challenge so that we could understand the data and represent it in such a way so that an end user could clearly see the data representation and could understand how the businesses are doing in and around a particular area and also some user related data.

With this kind of data and many of the visualizations that were discussed from the class, we think we can better represent the data. For example, the geo location of the maps and also the bar and area graphs with brushes included are a good way to analyze large amounts of data location wise, over a period of time or over a big group. Also, we feel that this could be a good exercise to kind of implement the standard visualization techniques on some real world data sets which are not just tailored for visualization demonstrations but are actually the kind of data we expect from an actual application.

Questions: We have considered a lot of possible questions which we would like to answer as a part of our analysis and visualization of the data. The data we have is a wholesome representation of business data from the application's perspective of the various businesses in a lot of areas along with a lot of users who are interacting with these businesses and are rating them. We primarily can think of two kinds of analysis that could be drawn from the given dataset.

One approach is to show the users the trends of businesses in a particular area over a period of time or the overall ratings in every category. The data given is per business with a lot of factors to consider and analyse how good is a business doing based on user ratings and also how each business fares with a bunch of other businesses of the same category with respect to the overall ratings.

The other approach is to view the customer interactions with the businesses based on the overall ratings of every single customer and the number of reviews that are given. When both the business and customer interactions are taken into consideration, there are several other possible trends that could be visualized using the data.

Data: The data we have for the given task is massive. The following are the rough estimates of various parameters of data size we are going to deal with during the course of this project.

- **1.6M** reviews and **500K** tips by **366K** users for **61K** businesses.
- **481K** business attributes, e.g., hours, parking availability, ambience.
- Social network of **366K** users for a total of **2.9M** social edges.
- Aggregated check-ins over time for each of the **61K** businesses.

There are primarily 5 kinds of json files present in the entire dataset namely, *yelp_academic_dataset_business.json, yelp_academic_dataset_checkin.json, yelp_academic_dataset_review.json, yelp_academic_dataset_tip.json and yelp_academic_dataset_user.json.*

These JSONs are quite large and so we have decided to process the data and clean it up in multiple steps so that the data would finally be reduced to a small enough size and also

the format of the data would be in such a way that there would be not a lot of effort for the application to process and filter the data in the runtime. We have initially parsed the JSON objects and removed the set of features that we are not going to be using for our visualizations. This has drastically reduced the size of the files that are to be loaded (more details to be added on finalizing all of our data formats). In the second step, we have parsed the data from the reduced JSON files in order to change the format of the data so that it is readily available for the visualizations to use. This would reduce the effort and decrease the time for the data to be loaded in the visualizations and for faster updates. The data count is so huge that every simple modification takes us a few hours to convert using a standard python script.

Lastly, we are planning to organize the data from the formatted JSONs in such a way that the filtering of the data would be really fast so that the visualizations update with very less delay.

Exploratory Data Analysis: We initially had the following ideas for visualization in mind in order to represent the visual analysis of the Businesses in an area across a bunch of categories.

- *Map* - This visualization would plot the locations of all the businesses and also the filtered data based on the categories, regions and other factors on the map. This would help us visualize the trend of locations on certain kinds of businesses and also how well a particular category of businesses are doing in a certain area. This helps the user to quickly select and compare between multiple local businesses and view their ratings history and overall scores over a bunch of businesses in an area of interest. This task would be challenging because the sheer number of businesses that are covered as a part of YELP data are huge and could cause problems and cluttering in the data on the map if not filtered properly.
- *Bubble/Pie Diagram* - This visualization is to represent the percentage weightage of each category among the list of all businesses. This representation shows how dense is a particular category of a business over a certain region and count (and probably more data). This gives the user an overall understanding of how and area could influence any particular category of business and also how dense it is in general in terms of sheer number.

- *AreaChart* - This visualization helps us to analyse how the ratings for a certain business or a bunch of businesses have changed over a period of time. With the help of this the user can see how the ratings for that business have changed over a certain interval of time or till it went out of business. This we believe would help one to understand the general trends over a bunch of businesses in the same or different categories over time.
- BarGraph - This visualization helps to analyse the overall ratings of a business and also specifically the number of ratings across all the stars. This would help the user to see how well the business is doing not just based on the number of ratings in general but also based on the number of ratings per every single category. This graph would plot a lot of data and in order to be able to view the data clearly, we are implementing a brush on top of this so that the user gets to select a subsection of data based on his interest.
- *BipartiteGraph* - This visualization helps the user to better understand the overall data that is being represented by the visualizations. This represents a bipartite graph relation between the areas and different categories of businesses so that a user can easily check which area has what categories of businesses in majority and also which category of business is more well established in which area. This provides the user with better understanding and helps him better decide what kind of filters he needs to use over a certain area or for a particular category of business.

These are the primary ideas of various visualizations to represent the data we have to analyse the trends in local businesses from the YELP data. We may come up with more interesting and different ways of handling the data in time and also by considering other factors which we haven't so far from the huge dataset.

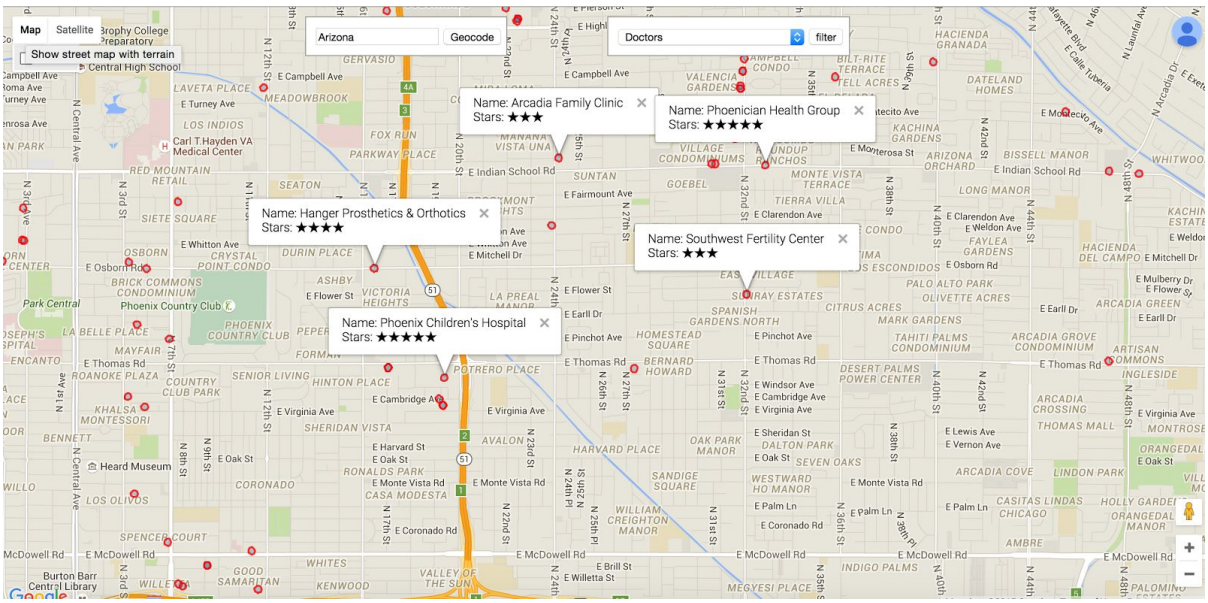
Design Evolution: (What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?) The list of above visualizations are a part of our initial phase to represent the data. We have encountered a lot of problems while trying to represent the data in these visual charts. So there are eventually a few places where we had to change our approach a little and compromise on certain things due to the limited time resources to work on such vast data. Since there is no straight method to

represent this data, there have been a few changes in the designs and representations as a whole and a few ideas that we stumbled upon later which we thought were more suitable for such data representation. More on this section to come up...

Implementation:(Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.) This section would be added soon once we have finalized all the designs and implemented the changes.

1) *Maps*:

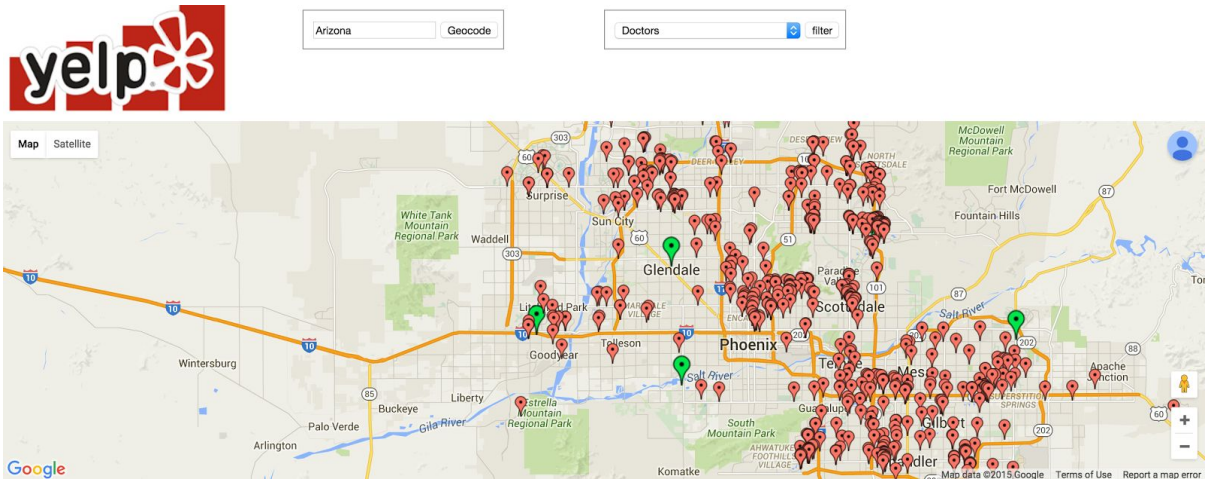
In this we plot the business on the google maps.We choose google maps because that shows us the street name which makes it easy to perceive. We plot the markers for each businesses using location details (latitude and longitude). Initially to plot businesses, we chose circles to plot on the maps, then we investigated various sites which uses similar models and ***we changed circles to markers***.On Click of markers or circles we initially had info window popping up the details of the business, then we found it messed the map as it was open everywhere until a user closes it. Due to this mess, we changed from **on click to on mouseover functions to display the business details.**



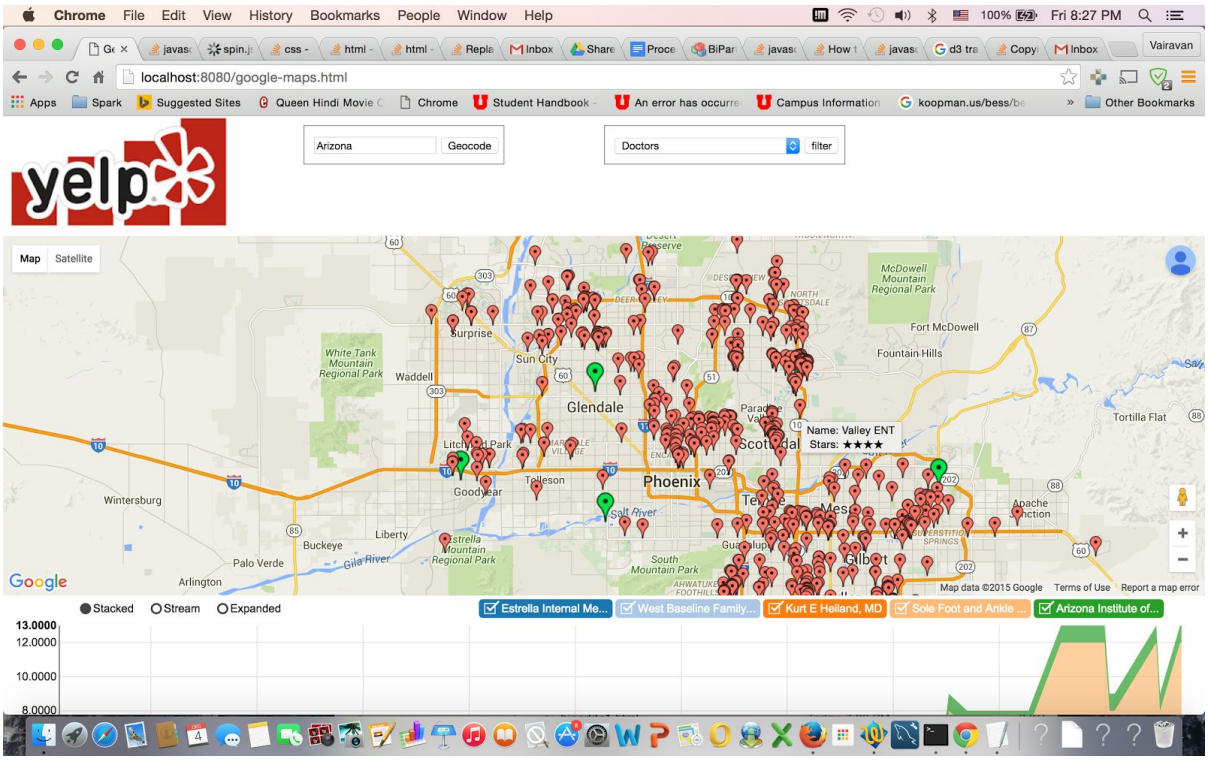
From the beginning, we decided to compare five business in each region to let the business owner who is running a business there, or a prospective business owner to analyse how well a business is going. If he is a business owner, he can see his reviews going down and may try to make to make it better or if he is a prospective business owner. He may analyze the popular business there,compare how well it is going and know the competition. **With this he will be able to choose the area where he could start a business where there are more chances of him becoming successful based on the statistics and also the concentration of businesses around that area over a period of time.**

The following are the various functionalities of markers:

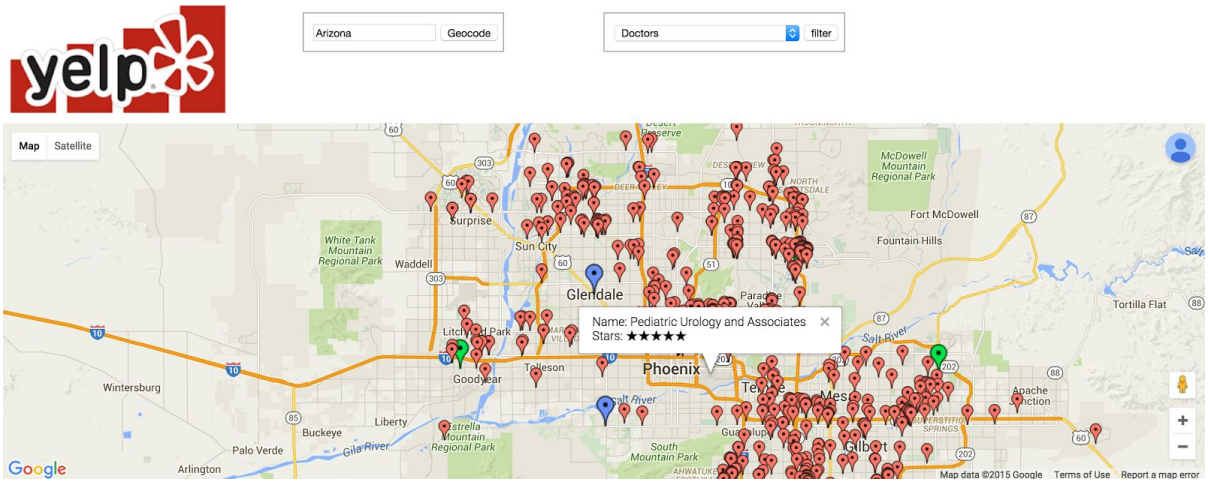
1. ***Red Markers:*** Those are markers which marks the business in that location. Just a normal marker.Initially if you select a business,there were no special notification in the map. Then the info windows are the notifications. Finally, we implemented green and blue markers.
2. ***Green Markers:*** If you click on a red marker, it will turn bigger and to green color to notify that you have selected that business for comparision. It will start to bounce out of happiness.
3. ***Blue Marker:*** if you click on a marker which is already green, it will turn to blue to let you know that you have already chosen this business for comparison in the past. If you again click the blue marker, it will turn to green and will get ready to get compared.
4. ***Info Window:*** Initially these were used to identify whether a business is selected for comparison or not. But, this made the graph clutter. Then to remove the clutter we throwed the info window out of design and we just made information to display on mouseover. Then we decided to display info window for last clicked business on map.
5. ***Geocode:*** This is a same as google maps location searcher. If you type some address, it will take you there. If you wish to look at different location, this is what you want.
6. ***Categories filter:*** This has all the categories that this dataset has, You could choose any category you want and it will display the businesses which are tagged under that category.



on Mouseover:



Blue markers and Info window:



Once you have chosen five businesses and try to choose an another business, on your info window it shows you have already chosen five business for comparison and It remains as red marker. You have click on the green marker to deselect it and then you have select new selections.

on selection of markers on map, it Interacts with:

1. Stacked Area Chart.
2. Stacked Bar Chart.

Map updates on:

1. Sunburst chart.
2. Stacked Bar Chart brush.

When there are a lot of businesses clustered the maps takes a lot of time to drop down the markers, So to indicate that we have incorporated spin.js to indicate whether the markers drop down is done or not.

2) Stacked Area Chart:

Using this area chart, you could see the reviews given by user in the each point of time sorted time wise. This might be useful to identify when the business started to go down and this might be helpful to identify the problem which led to poor ratings.

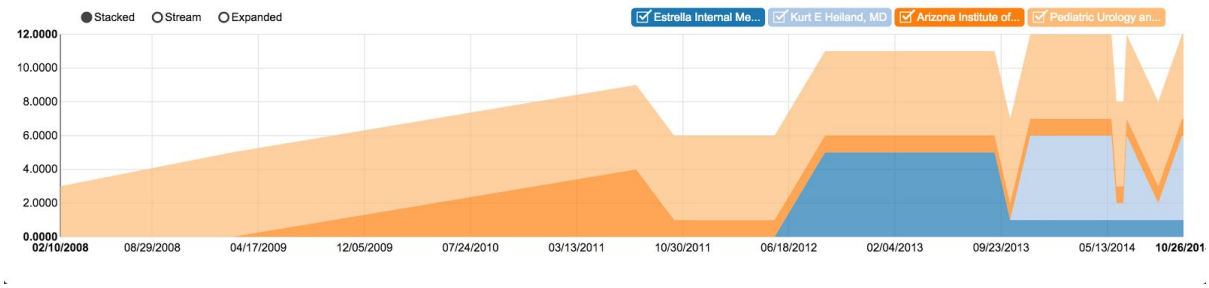
For this stacked area chart, we need ratings for all dates, but the rating for businesses is sparse for some and dense for others. We had to fill all dates with previous rated rating until a new rating was found. This was computationally expensive and might take some time to load as we do this task.Each businesses are given a color and legends will give you the relation between color and the name of the business. There are three different views:

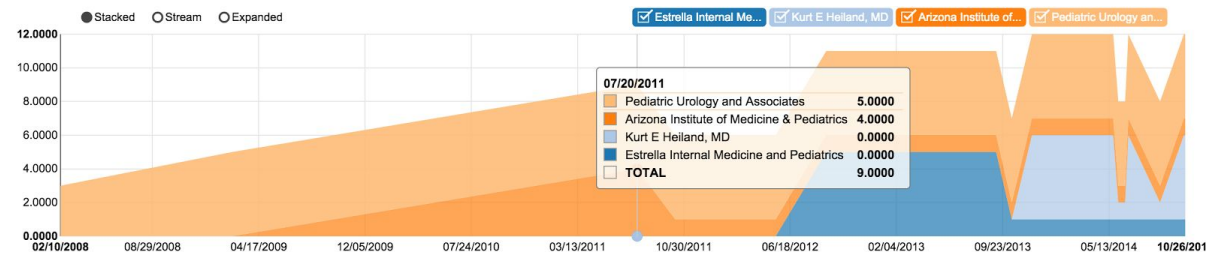
In stacked area chart, there are basically three views:

1. Stacked View(Absolute value)
2. Streamed View(Absolute value)
3. Expanded View(Relative value)

1.Stacked View:

Stacked area chart will give you the overall rating for the hotel. You cannot totally rely on stacked area chart. Since it is stacked, on just looking at the chart you cannot determine which is better, so we have given a on mouseover function to compare at each point of time.

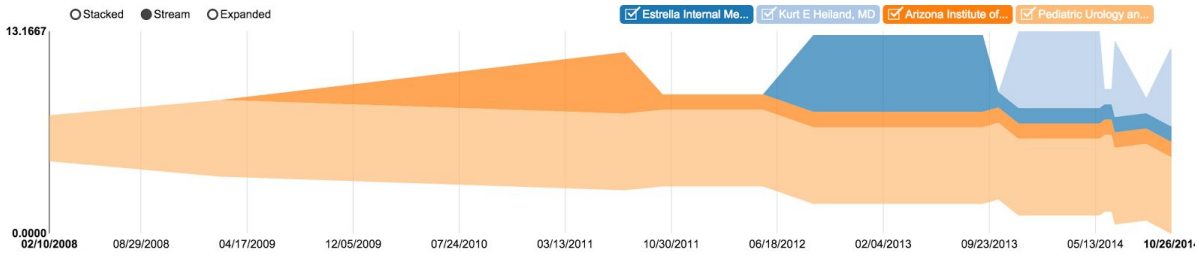


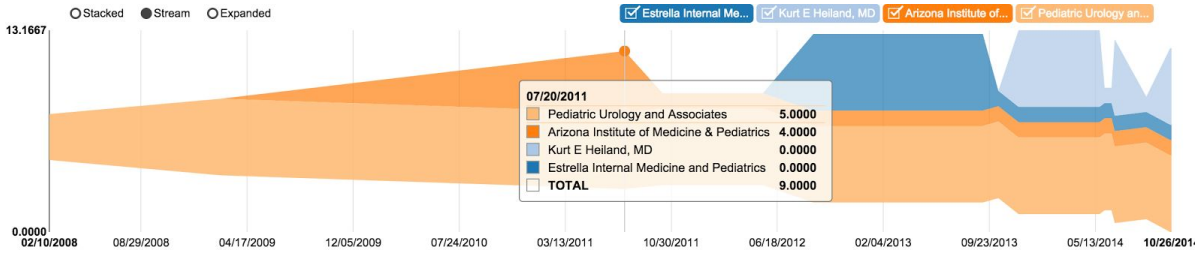


2.Streamed View:

Stream area charts are a generalization of stacked area graphs where the baseline is free. By shifting the baseline, it is possible to minimize the change in

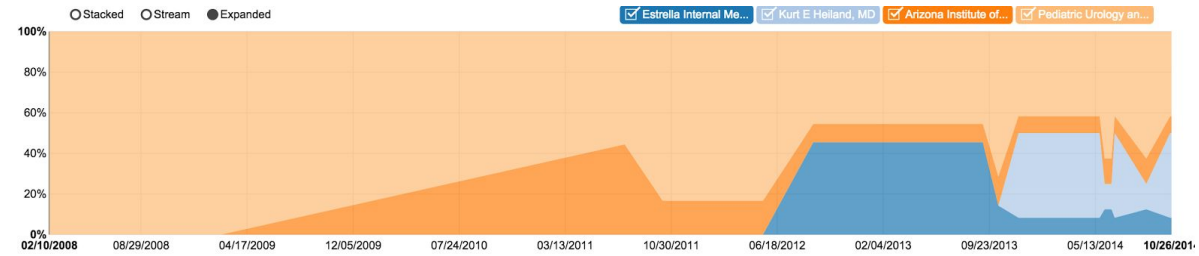
slope in individual series, thereby making it easier to perceive the thickness of any given layer across the data.

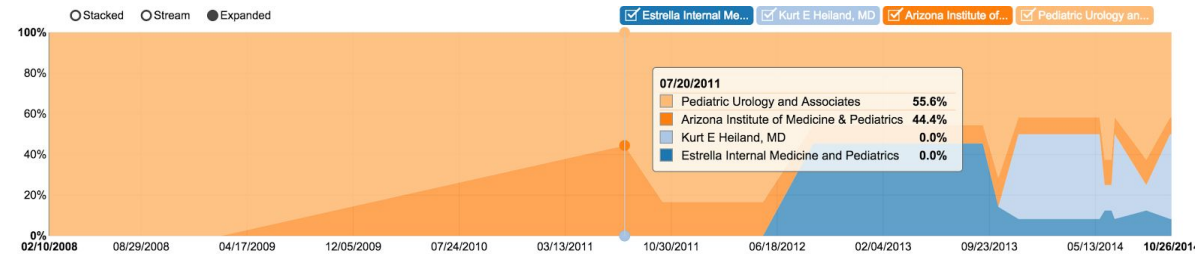




3.Expanded area charts:

Depending on percentage of ratings the area graph is expanded to the entire chart. This view with on mouseover option will give us clear view on which restaurant is better at a point of time.





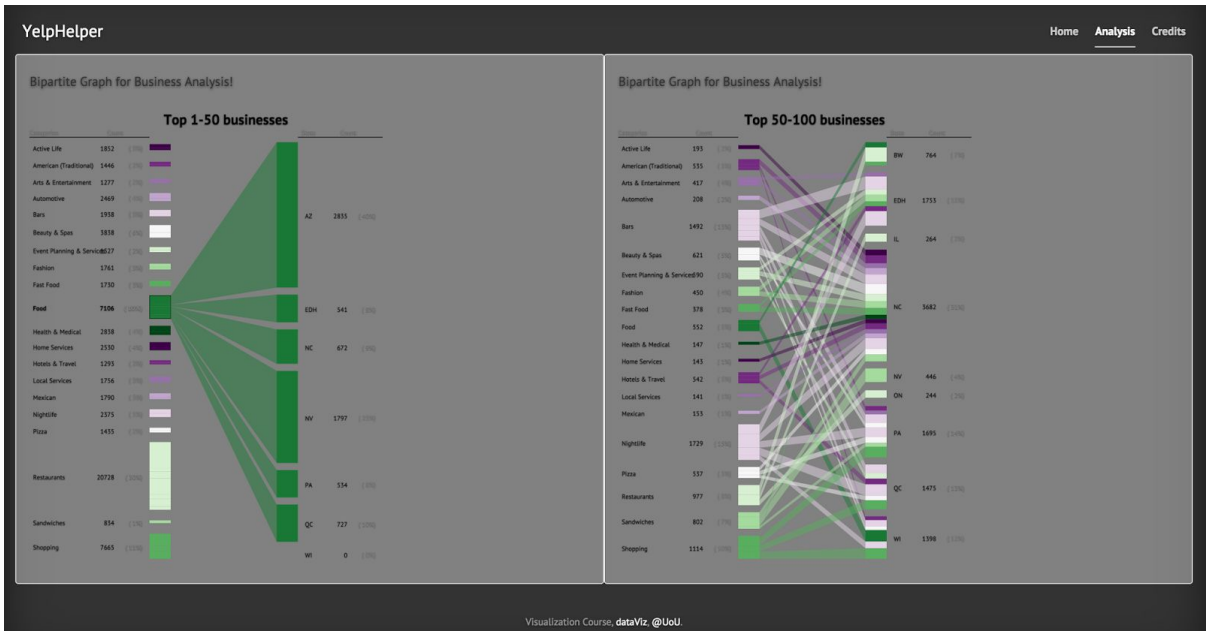
The difficult part of this implementation is formatting the data for this graph.

This graph interacts on click by changing its contents but not with other parts.This graph is intentionally made not to interact with other parts because then only the user can click on a business and can see that business performance. To give that space to user, we made it not to interact with others.

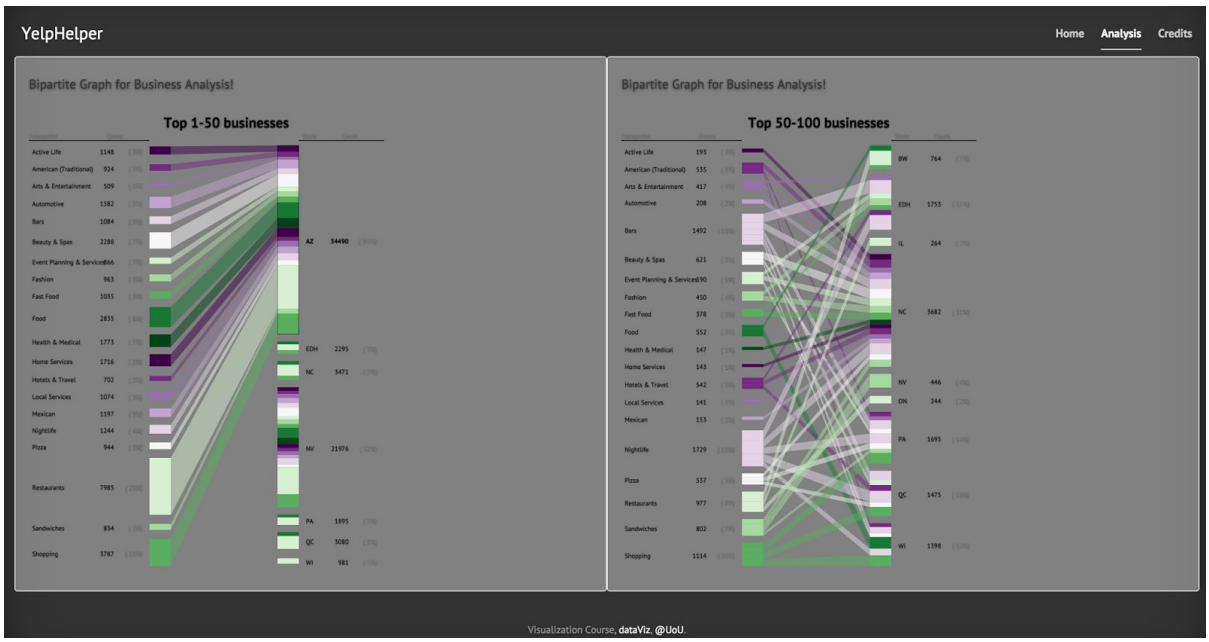
5) Bipartite Graph:

This can be used to analyze which business categories are popular in each area. For each tuple of categories and states, the number of businesses for each set is calculated and the top 100 business is displayed to show which are the popular business that are listed according to each area by yelp dataset. The Business analyzer can know the popular business on each region as listed by yelp.

On selecting restaurants from Top 1-50 we get:



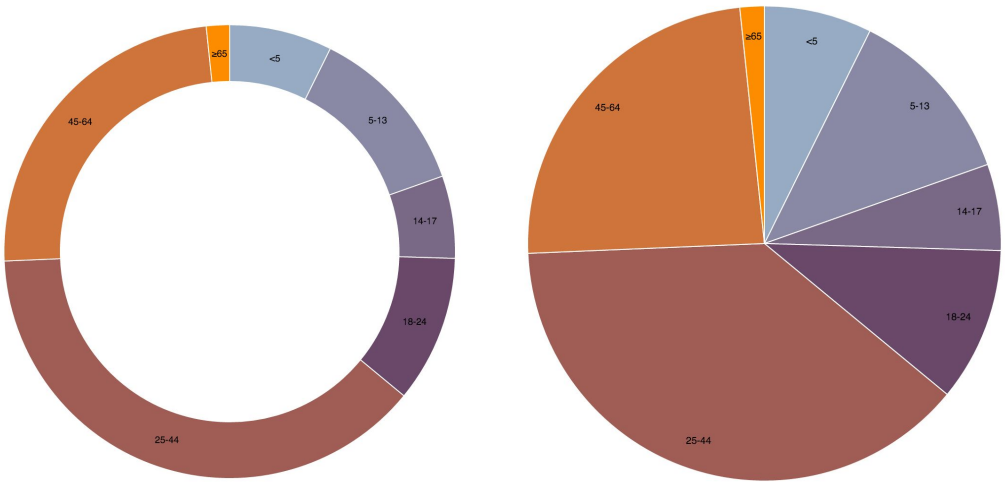
On Selecting AZ from Top 1-50 we get:



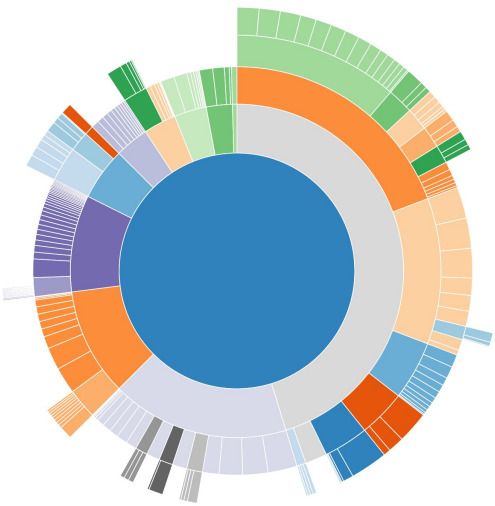
You could do the same thing for the bipartite graph on the right.

3) *SunBurst Graph*:

We wanted to find an appropriate visualization for representing the vast number of businesses and their various categories with respect to area. We initially planned on doing this using a simple pie chart or donut chart or a combination of these for representing hierarchies. The following are some of the examples which we had initially in mind,

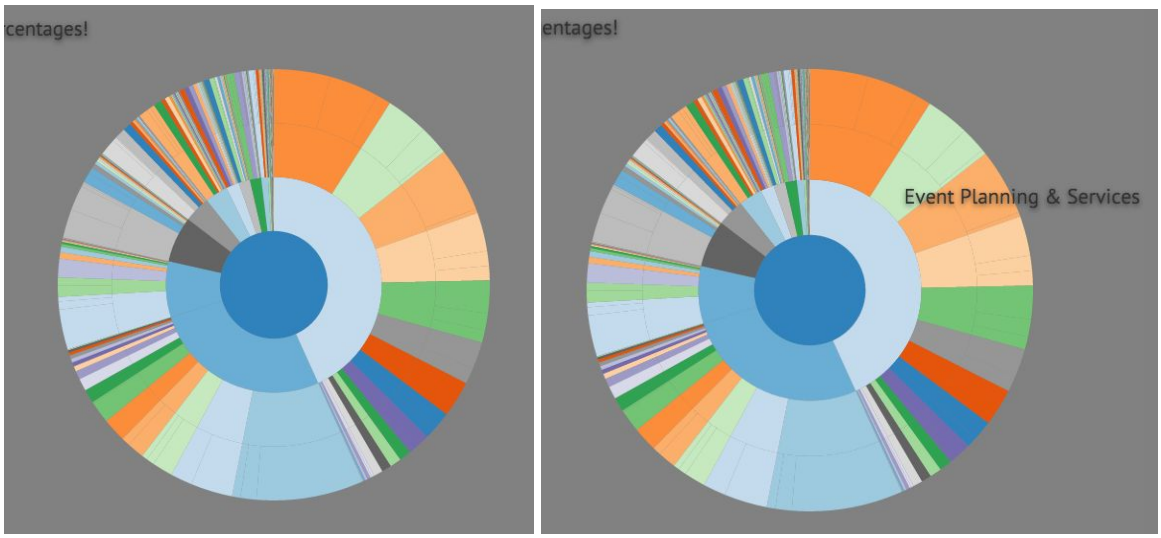


But these representations, either single or multiple donut/pie charts aren't very effective in representing the number of categories and also their hierarchy. In addition, these representations also cost a lot of screen real estate which is also a problem because we are trying to optimize the screen usage to fit in all the visualizations so that they could be more interactive. So we finally came across a really good way to represent this variety of data with hierarchies using a *SunBurst Graph*. The following is a standard sunburst graph,



The center of the sunburst represents the root node of the tree and the rest of the outer partitions represent various children of the root node and the sub partitions outside these partitions are the children of the children and so on. This representation of hierarchy coupled with event handling and selection of any particular section allows for a good representation of our complex data.

So, we formatted our data for all the businesses based on the field *categories* and *state* in their corresponding order such that the root node represents all the businesses in the data set and the first outer ring of sections represent the different states across which the data is spread. The next ring is the subsections of various types of major categories of businesses (for examples restaurants is a major category of business because it encompasses all the food joints, restaurants and other smaller places together) in that area. Then the further ring of subsections is the different categories of businesses under the major category and their corresponding counts. Once, we parsed this information from the given data in a tree like structure with every element having a *name* which is the name of the category and *children* that has a list of its children. The following is the way our sunburst graph looked. We then added a tooltip on mouseover for this graph so that we could easily identify the category name which we are about to select. The following are the images of the sunburst graph with and without the tooltip.



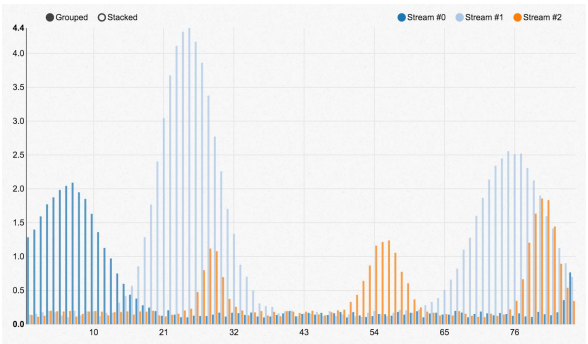
The only thing that could be improved here is the separation line thickness between different subsections with the same color. We tried multiple ways but sadly we couldn't get the desired output. This sunburst graph represents the overall generic hierarchy and categorization of the businesses we are dealing with. This graph is interactive and upon

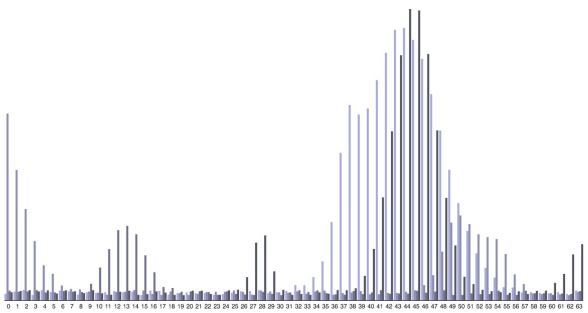
clicking on any particular selection, it would return the list of business IDs under this category and its subcategories.

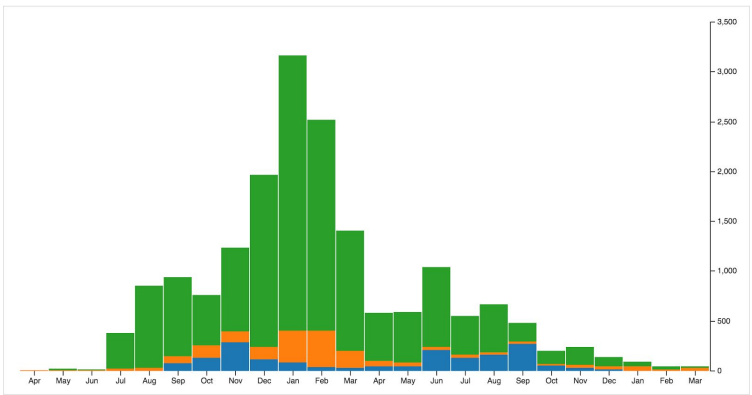
This visualization interacts with all the other graphs in the page, on selection, it plots the corresponding businesses on the map and also populates the vertical bar graph in the context and focus areas.

4) *Stacked Bar Graphs with Brush*:

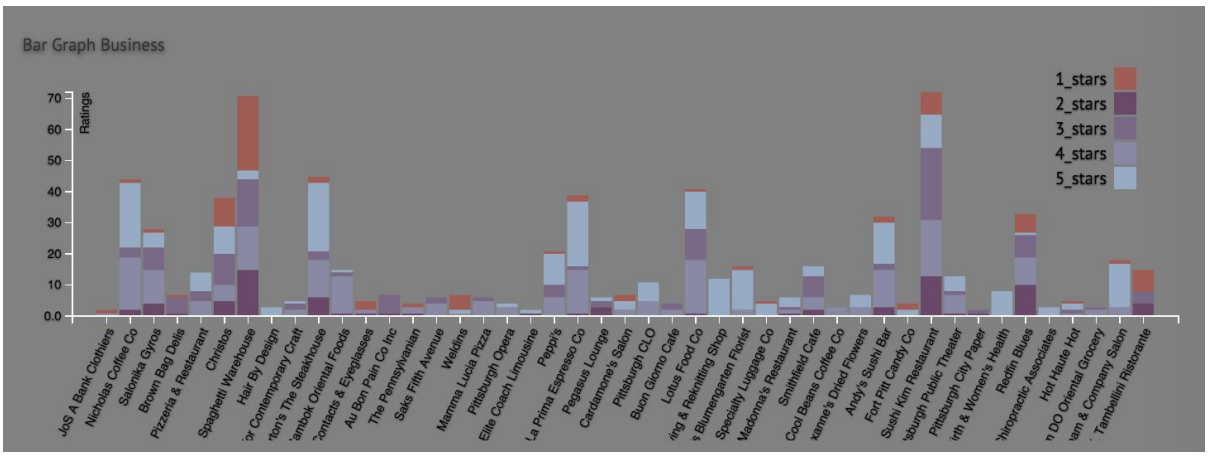
Once the user selects a subsection or category from the list of all the categories, we wanted to represent all the individual businesses from within this category to the user in a neat format showing him the total number of ratings the business has and also the information about the number of ratings with *5 stars*, *4 stars*, *3 stars*, *2 stars* and *1 star*. We wanted to do this using a stacked bar chart with the x-axis representing various businesses with their names on an ordinal scale and the y-axis representing the number of ratings the business has. The following are some of the examples of stacked bar charts we wanted to try out,

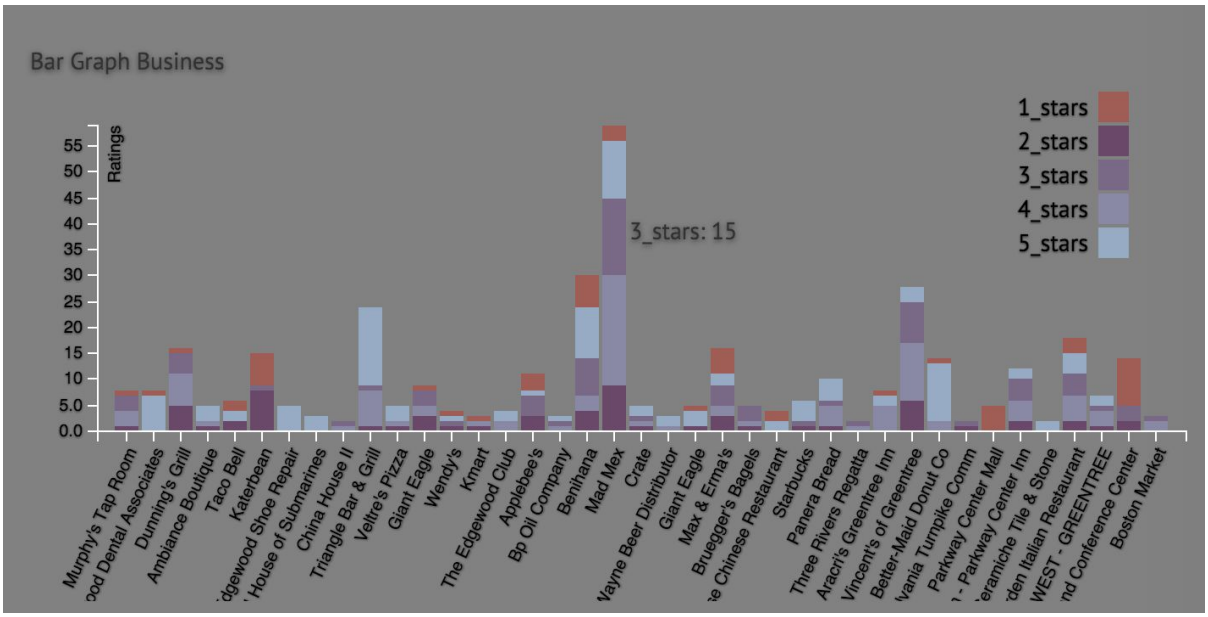




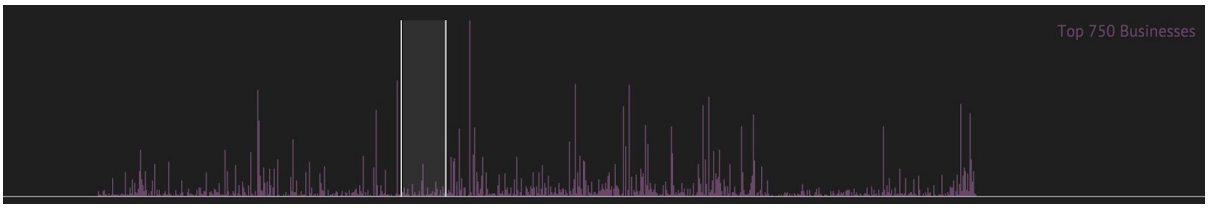


The first two examples have bars stacked horizontally which we thought was not a great way to represent huge selection of businesses as they would take up quite a lot of space while brushing and also would be difficult to scale. So we decided to go with the vertically stacked bar graphs as shown in the last image which represents various subsections of the bar vertically in a different color. After we passed the whole data to the visualization we have the following vertically stacked bar graphs with and without tooltips.





Since we are visualizing around 61k businesses, we have created a separate area for the brush context so that we could represent large number of businesses in a better way,



This scales and automatically populates the focus area with the selected extent of a brush with rectangles that could be hovered with a tooltip.

This visualization interacts with the map and also the focus area of the bar graphs upon brushing. This plots the selected areas across the map and pans and zooms the map layout while focussing on the individual businesses along with the sub section of ratings in the focus area of the bar graph.

Challenges with dataset:

1. The dataset which came in JSON format was not in proper JSON format, we had to change it to a proper format.
2. We had to clean the dataset to a large extent to bring down the dataset size. We wrote many python scripts to clean data at each point of time. We faced a hard time formatting each dataset.
3. Each dataset is very huge, with the largest being around 1.6 GB. This vast amount of information took quite a while to load in the web page due to all the data processing happening in the background. To overcome this we had implemented a spinner which is a loading sign to show exactly when the data is ready for visualization.
4. The map cannot plot markers more than a certain number. It gets unresponsive if we try to mark all businesses. So if you select businesses in sunburst it will show a alert window to choose a state. This also helps us to load the page quickly than spend a lot of time waiting for the data to be plotted.
5. The number of businesses are so huge that even the ordinal scale when calculating the width of the rectangle to be appended to the bar graph is almost going to zero beyond a certain point. This lead to almost no bars being plotted even on the extended context scale of the graph. To overcome this, we are only plotting the top 750 businesses under that particular sub selection using the sunburst or map

filters. This optimally plots the businesses and we could use the brush and also the map or even the sunburst to narrow down the search results.

Evaluation:

The following are some of the bullet points that we kind of figured out and think are important for understanding the dataset.

- We learned using this yelp dataset the kind of businesses that are well established for every region(for example restaurants in arizona).
- Arizona has the largest number of businesses in the yelp dataset.
- We can choose businesses to compare and we could see the number of ratings for each category (1star, 2star, 3star, 4star, 5star)using bar charts.
- The data is largely sparse when it comes to the reviews over a time period. If we try to plot the rating of the businesses over a period of time, there are only a few entries over a large stretch of time for every single business. This makes it harder to evaluate the business over area chart.
- We can choose businesses and compare their performance using the area charts.
- The bipartite and sunburst graphs will clearly show which business dominates a region.
- Using sunburst, you can easily filter data and more filters can be done using maps and brush in stacked bar chart to do the evaluation of businesses.

How our Visualization answers the problem:

Our problem is to compare businesses in that particular area using different view:

1. Sunburst chart: This is used to select from broader category of businesses.
2. Map chart: used to select five businesses to compare the businesses.
3. Stacked area chart: Used to compare different business rating with respect to time.
4. Bar chart: Used to list the star ratings for each business and number of ratings in each category.
5. Bipartite graph: this graph will list top 100 categories.