Yelp Helper

Yelp Dataset Visualization

Introduction

Goal

There is a yelp dataset, which comprises of information about business, users, reviews, tip and check in. Withthese information, we are trying to visualize the different business. With this the objective is to find the highest rated business among others in that specific region. With this representation we could easily find which business does great. It would be easy for users to filter the business and this will help to narrow down the choices.

Data Set:

- The data set consist of business of the following cities: Edinburgh(U.K), Karlsruhe(Germany), Montreal(Canada) , Waterloo(Canada), Pittsburgh(U.S), Charlotte(U.S), Urbana-Champaign (U.S), Phoenix(U.S), Las Vegas(U.S), Madison(U.S).
- There are five dataset : Business.json, ratings.json, user.json, tip.json, checkin.json

1. Business.Json: (BID, Name, Neighborhoods, City, State, Latitude, Longitude, stars, Review_count, hours (optional), Open(status), categories)
2. Rating.Json: (BID, UID, Stars, Date, Votes (optional), Text)
3. User.Json: (UID, Name, Review Count, Avg_stars ,Votes (Optional))
4. Checkin.Json(Optional) : (BID, Check in info)
5. Tips.Json (Optional): (Text, BID, UID, date, likes)

Cleaning Dataset:

The size of each datasets are Business.Json(55.4MB), Check in.Json(20.7MB), Review.Json(1.43 GB), Tip.Json(98.5MB), User.Json(166.2MB). When we analyzed the largest dataset we understood that the comment for each review is about 20 words each in average which contributed most to the large size of that dataset. We anyhow will not be useful text for our visualization, so we thought of filtering that out. At first we are going to start with visualizing with three datasets for which we have attached the figures. The three datasets that we are going to visualize are:

1. Business.Json
2. User.Json
3. Review.Json

Optional:

1. Checkin.Json
2. Tip.Json.

VISUALIZATION:

Color Schema: 5 Palate Colors based on ratings.

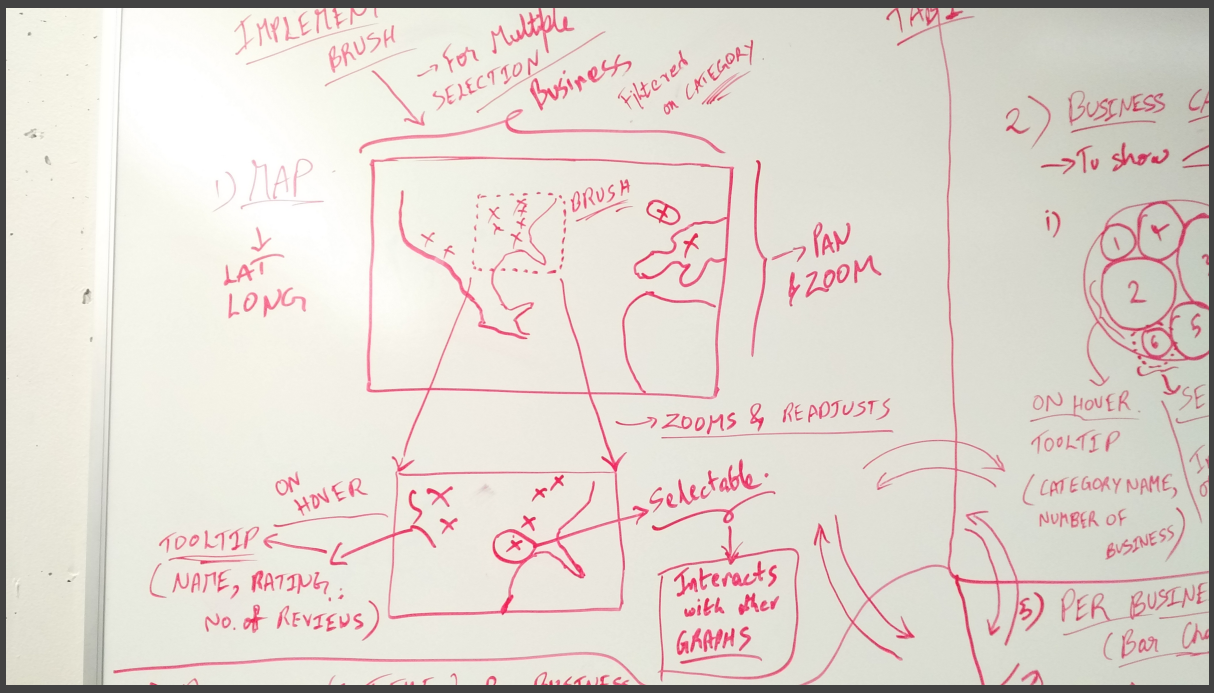
This schema is going to be uniform across all the graphs in the first tab. The first tab contains four different sections that interact with each other.

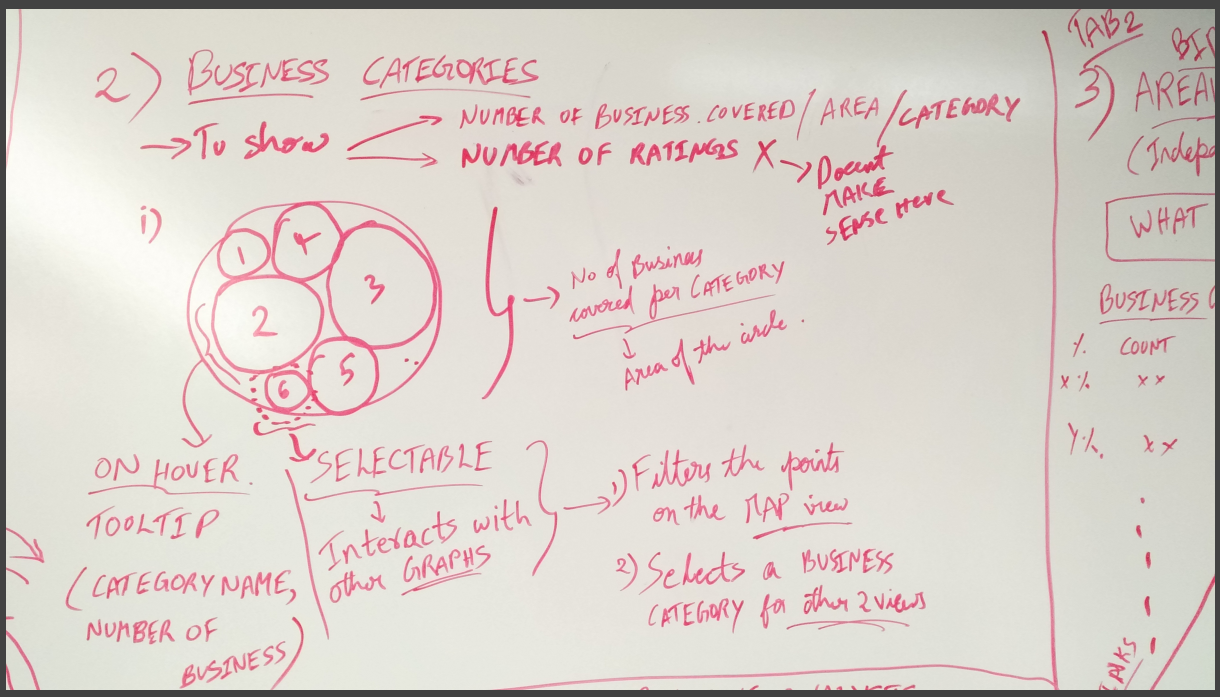
- **The Map:**
  This section plots the business establishments based on the location data on a map. We can pan and zoom on the map. Also, multiple points could be selected using a brush. The individual points could be selected and the details of the business would be displayed in a tool tip on hover.
- **Business Categories**:
  This section contains all the business categories represented as circles whose area denotes the number of establishments in that category. These are selectable and on selection are used as a filter.
- **Ratings vs Time:**
  This section plots the average ratings over a period of time for a business selected which is used to analyze how good the business did at some point in time. This could be represented using a bar chart, scatter plot or area graph depending on how we are going to handle multiple business selections.
- **Review Count and Rating Analysis:**
  This section plots the review counts per the number of business selected in an area using bar charts. The color schema remains standard.

The Second tab consists of a Bipartite Graph for analyzing the business in an area and vice versa. This is independent of the interactions from the first tab and visualizes the core data.

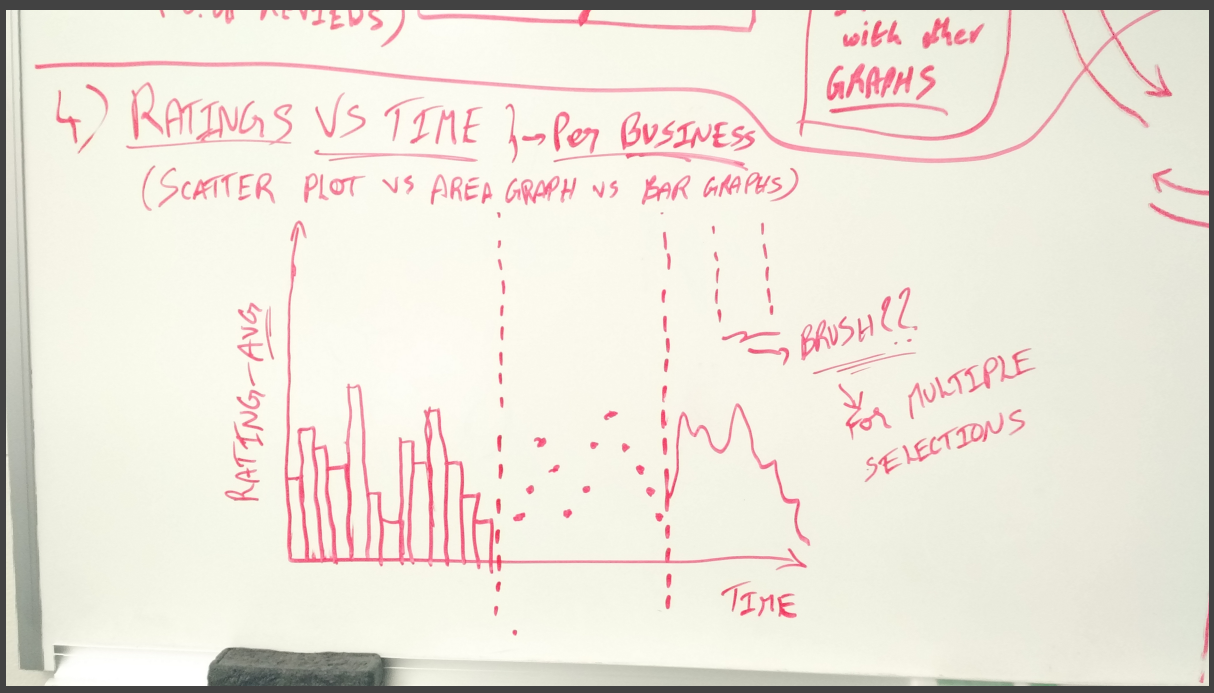
Filters: The data can be filtered based on Business Category, Ratings and the Area.

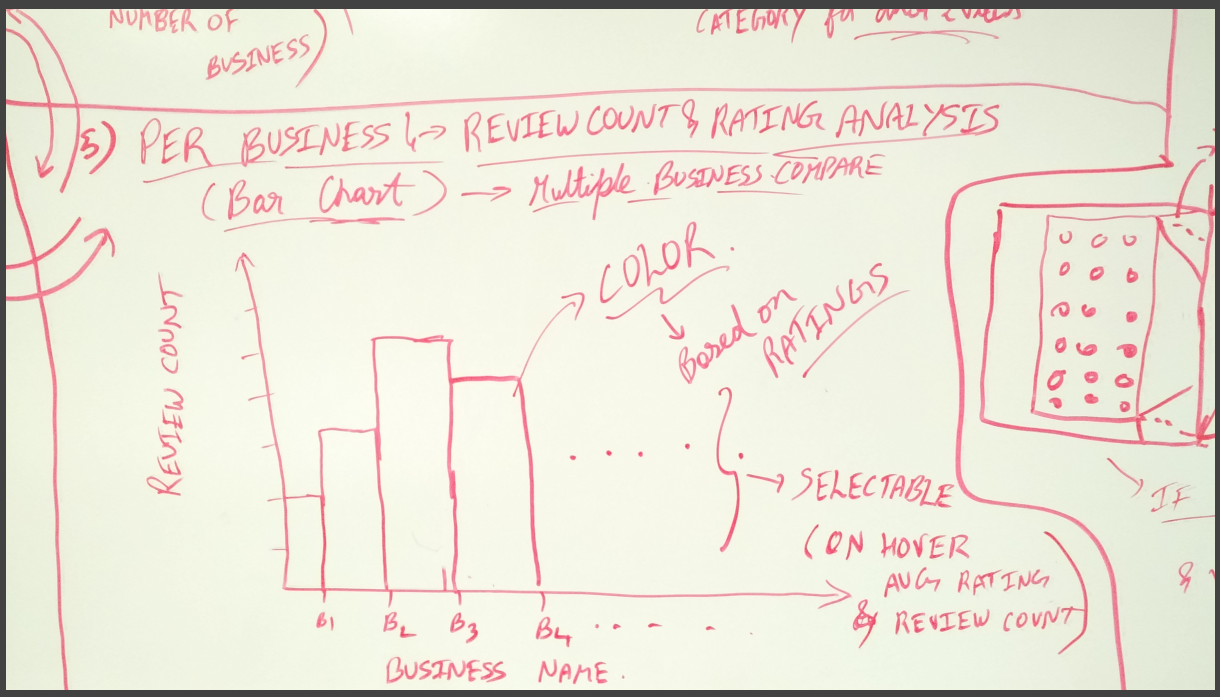
The Map & Business Categories:



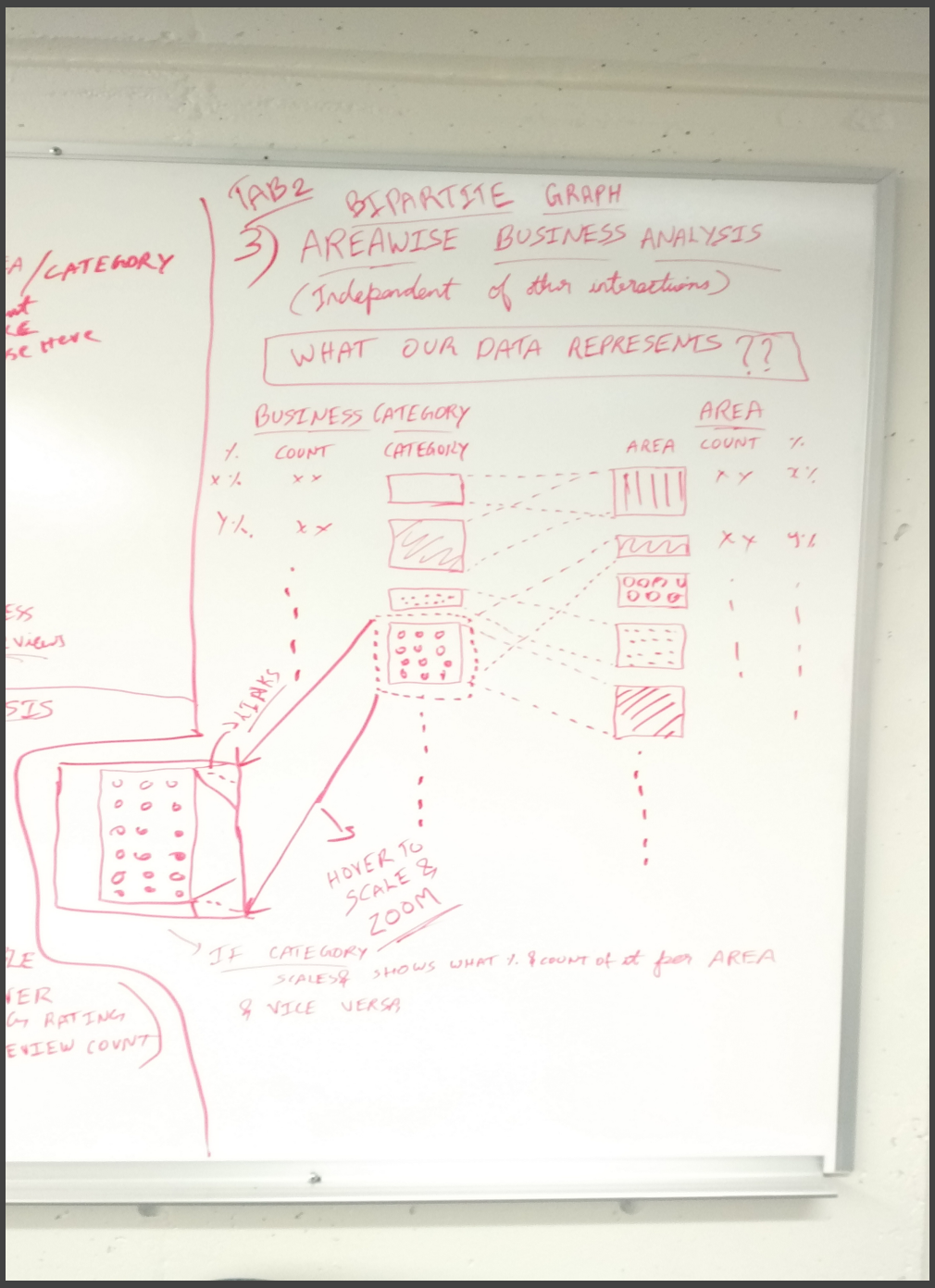


Rating vs Time & Review Count and rate analysis:





Area Wise Business Analysis:



Page Layout (1st Tab & 2nd Tab)