Update 1:

The Yelp dataset is huge and it took me quite sometime to setup a streaming pipeline to succesfully filter it down to more manageable proportions. For this project I am looking just looking 'restaurants'(type of business) in 'Edinburgh'(one of the cities).

Merchandizing labels:

I identified attributes/details/information that a customer might specifically look for while deciding on a restaurant. I call these Merchandizing Labels. It doesn't matter what is the value of the attribute, just if it's available. For example, 'Noise Level' - it doesn't matter if it is 5,6 or other value. If a value is provided I say it is attributed.

The idea here is that the more a business is attributed the more likely it is that customers/users/reviewers interact with it online. This is because more often than not they find the detail that helps them make decisions(both positive and negative).

So, the hypothesis here is better merchandizing label attributed restaurants have more reviews/tips/stars etc. The second and third plots kind of agrees with this.

Operating Hours:

Although I encode operating hours as a Merchandizing Label, it has a very significant value compared to other merchandizing labels. I say a restaurant is attributed in Operating Hours when it mentions in operating hours for atleast 5 days of the week.

Location:

As seen in the distribution of review counts is super long tail. When I split out the long tail from the rest and plot a choropleth for merchandizing label counts, it nicely ties back into my earlier hypothesis.

The first map plot is for restaurants with less than 12 reviews and second map plot is restaurants for more than 12 reviews.

Customer(Network Design):

This is still WIP but I have identified customers who reviewed these restaurants and how many of them are elite reviewers. All I need to do is to build a network between business and customers and observe the interaction of elite/non-elite customers and more/less merchandized restaurants.

Original Post:

For my final project, I’ll be looking at the Yelp dataset. As one of the most famous and rich datasets for big data analytics, it provides me endless opportunities to mine the data to answer specific questions. It provides geo-location of each business, operating hour of business and customer and business interaction data. Ideally, I would like to exploit at least two of spatial, temporal and network visual design principles on this project.

The overarching question I would like to answer is ‘Does well merchandized businesses tend to get higher ratings and higher reviews i.e better sales?’. Breaking this down to sub-questions and deliverables –

What constitute a business to be well merchandized? (Multi-variate visual design)  
Does review counts solely depend on location? (Geo-spatial design)  
Do elite customers gravitate towards rating/reviewing only these well merchandized businesses? (Network Design)

Considering the size of the dataset, I will mostly concentrate on one particular state or city depending on some quick data sparsity analyses.

I will continually keep updating/changing/improving this post as I make progress.