This will be a basis for analysis. I will keep the journal in a separate doc.

**Initial thoughts and motivation**

1. The dataset is very rich, encompassing 3 out of 4 major data types that one would typically find in ML applications:
   1. Tabular data (mixed data: numeric, string, dates)
   2. Text, where one would use NLP
   3. Images

The only data type missing would be sound or other time-varying biometric measurements (movements, etc.).

1. It describes two interacting ‘agents’
   1. Businesses
   2. Customers

This constitutes a possibly rare opportunity of analyzing the data more from a business point of view, i.e. how businesses fare as ‘measured’ by customers, but it can also reflect on the customers themselves, i.e. do customers have distinctive features, e.g. are they ‘spammers’, are they connected, etc. Unlike Uber data, I presume, there is feedback from customers to businesses but not vice-versa, i.e. businesses do not rate customers.

1. It is a real-world data set, with all its quirks. From the more mundane like missing data, misspellings, but also time-varying nature of reviews, etc.

**The Challenge Dataset:**

* **2.2M** reviews and **591K** tips by **552K** users for **77K** businesses
* **566K** business attributes, e.g., hours, parking availability, ambience.
* Social network of **552K** users for a total of **3.5M** social edges.
* Aggregated check-ins over time for each of the **77K** businesses
* **200,000** pictures from the included businesses

## The Challenge

Not only would we like to give you our data, we’d also like to announce the seventh round of the **Yelp Dataset Challenge**. We challenge you to use this data in an innovative way and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

**Cultural Trends:** By adding a diverse set of cities, we want participants to compare and contrast what makes a particular city different. For example, are people in international cities less concerned about driving in to a business, indicated by their lack of mention about parking? What cuisines are Yelpers raving about in these different countries? Do Americans tend to eat out late compared to the Germans and English? In which countries are Yelpers sticklers for service quality? In international cities such as Montreal, are French speakers reviewing places differently than English speakers?

I guess, generically: what do customers care about?

**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?

Very interesting question!

**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?

I assume businesses are categorized by type?

**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?

Very interesting

**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?

Very interesting

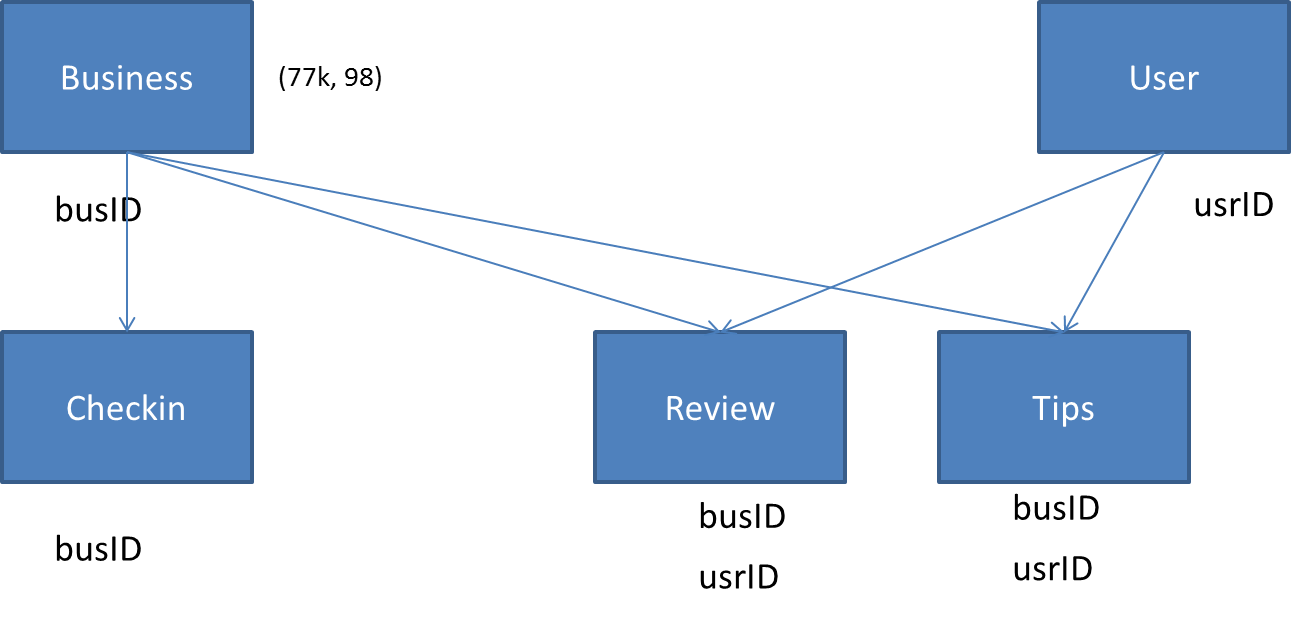
**Changepoints and Events:** Can you detect when things change suddenly (i.e. a business coming under new management)? Can you see when a city starts going nuts over cronuts?

**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?

**Data Structure**

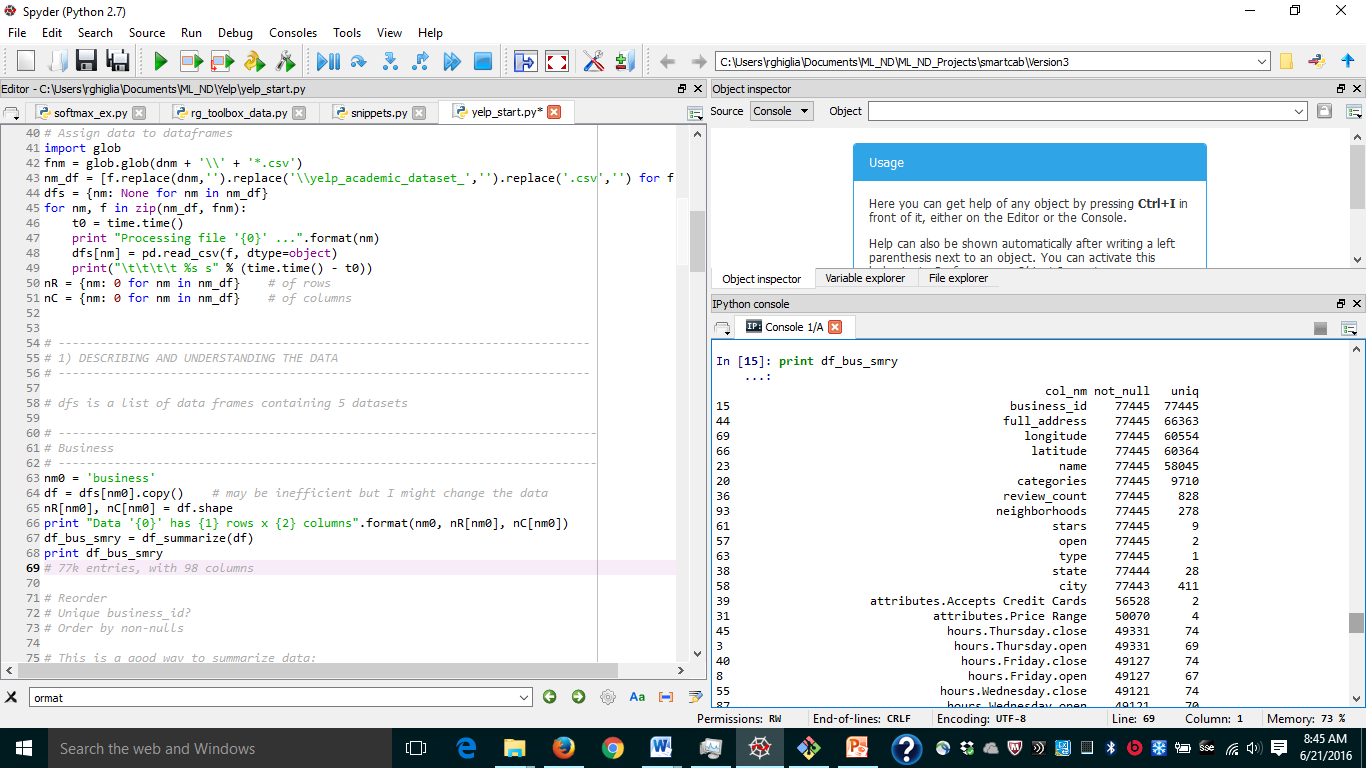
The data consists of 5 sets (Figure 1).

Figure 1



**Business**

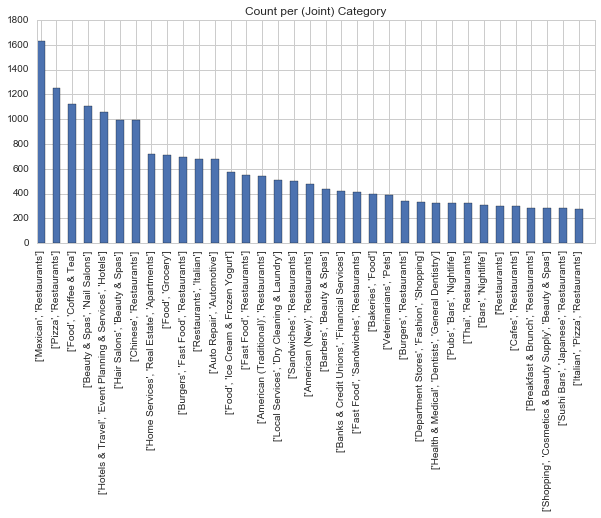
There are 77k unique businesses, although not all have distinct addresses, nor distinct locations, or distinct names.



The businesses are categorized. However, each category is really a combination of categories, e.g.:

['Hotels & Travel', 'Event Planning & Services', 'Hotels']. The number of businesses in each category decreases but we have to go to 8,935 categories before we cover 99% of the businesses.

Figure 2



We therefore separate out the subcategories.