The Impact of Socioeconomic Status on Yelp Usage and Reviews

by Dustin Howell, Olga Romanova, Martha Poole, and Dileep Ciddi

Part 1. Introduction

This paper will utilize data from Yelp, Zillow, and other socioeconomic data sources to investigate the correlation between socioeconomic status and Yelp review quantity, quality, composition, variation. We will also look at the relationship between home values and characteristics of the businesses and neighborhoods themselves, including attributes such as business density and category. We seek to establish expected trends such as wealthier areas attracting more reviews, as well as seek out more surprising revelations that will shine some light on how the proliferation of social media may impact societal development and vice versa.

Part 2. Abstract

Users of social media and review sites such as Yelp skew towards the wealthier end of the population distribution. Our work will attempt to see if this fact is meaningful to how we interact with such services on a daily basis. Insights gleaned from the analysis of the data will help us interpret how we should utilize the information gleaned from using these sites and whether or not such information should be discounted or utilized differently.

Part 3. Related Work

Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud

The paper investigates a business's incentives to engage in review fraud using data from yelp.com. Overall, roughly 16% of reviews are identified as fake and are subsequently filtered. It is also found that a restaurant's "offline" reputation is a determinant of its decision to seek positive fake reviews. Legitimate reviews are unimodal with a sharp peak at 4 stars(on a scale of 5). By contrast, the distribution of fake reviews is bimodal with spikes at 1 star and 5 stars. Estimate the characteristics of filtered reviews by using a linear probability model. It is also observed that low ratings increase incentives for positive review fraud, and high ratings decrease them.

Poster: Towards Detecting Yelp Review Campaigns

The goal of this paper is to detect review campaigns, concerted efforts to bias public opinion. The notion of user ratings is used to identify low quality reviews. Expertise coefficient of a user for a review is defined as the number of reviews written by the user to the number of active users by the user. Timeline for a venue is defined to be a set of tuples (user and time) of the reviews in chronological succession. Spikes in the evolution of positive reviews are studied. Box plots are used to identify outliers. The amplitude of the spikes in positive reviews has a long-tail distribution. Corresponding parameter values for the distributions are calculated.

Narrative Framing of Consumer Sentiment in Online Restaurant Reviews

This paper attempts to provide psychological insights through sentiment analysis of Yelp reviews. The study evaluates differences in word choices for positive and negative reviews. Language choice for positive reviews tended to include metaphors for sensual pleasure and addiction. They also tended to include more sophisticated word choices, especially when the venue being reviewed was more expensive.

Negative reviews most often resembled language that described trauma. Descriptions of bad service or a negative experience were similar to how one might describe a tragedy or death of a loved one. The study deduced that this reflected the reviewers attempt to reconcile the experience through the act of writing the review.

Inequality and Web Search Trends

This project analyzes the correlation between google search terms and regional socioeconomic status. Regions were weighted on a relative spectrum, as "easier" to live in versus "harder." The criteria for this spectrum included income, life expectancy, and education. In particular, the focus was on terms that correlated highly for one group relative to the inverse correlation for the other group.

Search terms most prevalent in easiest places to live were related to technology, exercise equipment, and foreign travel. Search terms that correlated most highly with the hardest places to live included topics relating to medical conditions, guns, and religion. The findings were characterized as a "glimpse" into the divergent concerns that accompany economic and social inequality.

New Retail Capital and Neighborhood Change: Boutiques and Gentrification in New York City

The study investigates the effects of gentrification in Williamsburg and Harlem through the changes in types of retail and food establishments in these neighborhoods. The effect of displacing local retail stores with upscale, but non chain businesses is referred to as 'boutiquing'. The new upscale stores cater to a more affluent clientele and can "reinforce a sense of the neighborhood's cultural distinction." Chain-stores can have negative effects such as disrupting a sense of community based on patronage of small, local businesses. Furthermore, the introduction of privatized spaces for more affluent newcomers often comes with more strict policing against the long term residents who choose to congregate in public spaces. Boutiquing benefits some residents while increasing economic inequality between the poor and well-off since the arrival of boutiques designates an area as 'safe' and leads to rises in rent. Harlem benefited from several government agencies establishing the Upper Manhattan Empowerment Zone to reinvest in the neighborhood. As a result, the type of boutiquing that occurred in Harlem "seems to reflect a specific kind of state-sponsored development: oriented toward the new middle class and tourists." Williamsburg experienced 'market-led gentrification' after decades of 'market disinvestment' without the aid of government intervention or zoning permits. In contrast to Harlem, most government intervention in Williamsburg occurred after gentrification began. Retail

entrepreneurs are drawn to neighborhoods partially because they belong to the new population and make the new residents 'feel more comfortable.' The analysis concludes that with both types of gentrification, the first wave is accomplished by the "first pioneers" who open individual retail establishments.

Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database

The study researched the effects of star ratings on restaurant customer flows. Since Yelp rounds to the nearest half star, two restaurants with similar ratings can vary by .5 star overall. The effects of the star ratings on restaurant customer flows are greatest for restaurants that have little information available outside of Yelp. For a reservation at 7pm, "moving from 3 to 3.5 stars reduces the likelihood of availability from about 90% to 70%. A fourth star reduces the likelihood of availability further to 45%, and that possibility drops to 20% at 4.5 stars". For "unfamiliar" restaurants with less than 500 reviews, an extra half star reduces reservation availability by 20 to 30 percentage points. If the restaurant does not have other recognition outside of Yelp, the effect is more pronounced. The study surveyed wait times for walk in customers as well which correlated with the reservation results. In conclusion, small variations in Yelp ratings had a significant impact on customer flows.

In Chicago, Food Inspectors Are Guided by Big Data – by Mohana Ravindranath (Washington Post)

The task of keeping Chicago's 15,000+ restaurants in a clean, healthy condition falls on only 32 food inspectors. Given this enormous task, Chicago is engaging big data analytics to guide their efforts. Currently inspection scheduling is guided by past performance on prior inspections, with violators incurring more frequent inspections. However, they are adopting an analytical approach that layers data on conditions such as weather and construction over data on past health code violations. The model takes in data from various publicly available sources, such as records of building and sanitation code violations, demographic characteristics of the local populace, restaurant density and characteristics such as which establishments have a liquor license. It goes back about 10 years over 13 variables and looks at which of these variables correlated most highly with health code violations. It turns out the variable most strongly correlated is fluctuations in weather, which is stronger than the restaurants location or history of past violations. The system also mines twitter posts to find clues about food poisoning, and while these tweets are not directly used in the analytic, they are used to encourage those affected to file formal reports, which are integrated into the analytic, which is a great example of using data mining techniques to enrich their primary data set. The system has boosted critical health code violation detection by 4%, which though it sounds modest, is a substantial boon to public health.

New York City restaurants; my search for the dirty truth – by Frances Angulo

In this article, the author mines the NYC open data related to food inspections to see if there are any correlations between cuisine type and health code violations. The data was joined together such that all

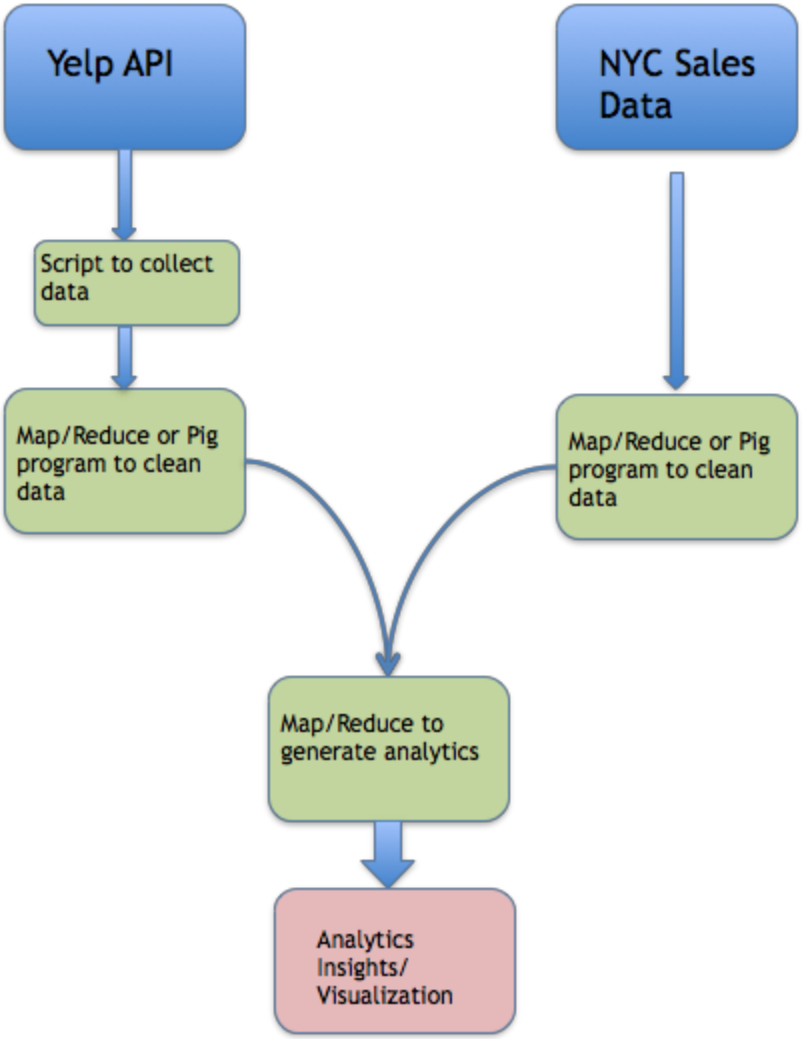
health code grades were aggregated and normalized using basic averages (to prevent the relative prevalence of certain cuisines from biasing the results). The author found no meaningful correlation between cuisine type and health-code grades assigned by the NYC health code authorities. He also found very little contribution/correlation by zip code. This is a good warning because it shows how pre-determined bias can affect expectation and guide wrong-headed efforts.

Part 4. References

1. *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud* – Michael Luca, Georgios Zervas
2. Rethinking Big Data to Give Consumers More Control – James Wilson
3. nEmesis: Which Restaurants Should You Avoid Today? - Sadilek, Brennan, Kautz, and Silenzio
4. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness – Harrison, et al
5. New Retail Capital and Neighborhood Change: Boutiques and Gentrification in New York City – Sharon Zukin
6. Learning From the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database
7. In Chicago, Food Inspectors Are Guided by Big Data – Mohana Ravindranath
8. New York City Restaurants; My Search for the Dirty Truth – Frances Angulo
9. Narrative Framing of Consumer Sentiment in Online Restaurant Reviews - Jurafsky, Chahuneau, Routledge, Smith
10. Optimal Circle of Friends Depends on Socioeconomic Conditions– Oishi, Kesebir
11. Inequality and Web Search Trends - Leonhardt
12. Demographics of Yelp

Part 5. Design

The architecture of the project is outlined in the figure below. The two main Data sources are from Yelp reviews and NYC sales. For the former we use the developer tools and API provided by Yelp to get information that is relevant to our analysis. Next stage involves cleaning the data, replacing missing values and formatting it so that it is in a format that is understandable by the map-reduce program in the next stage. The output of the map-reduce is fed as input to the next stage of mapreduce which uses both the data sources and generates analytics as desired and as mentioned in the project objectives. Visualization stage involves generating plots and charts that depict the analytics in a presentable format.


Yelp API
NYC Sales Data
Script to collect data
Map/Reduce or Pig program to clean data
Map/Reduce or Pig program to clean data
Map/Reduce to generate analytics
Analytics Insights/ Visualization