Data Science 287 Final Project

What Makes a Good Business?

12/13/2017

Chen Ling

Computer Science Department

College of Engineering and Mathematics Science

University of Vermont

Introduction

To begin with, there are millions of gigabytes every day are generated by blogs, social websites, and web pages. Companies are gathering all of these data for understanding users and their passions and give these reports to the companies to adjust their plans. Therefore, understanding these data is really important for the development of a company. Besides, users can also get useful information from the data, such as the business information and all users' comments.

Yelp.com is a great example, where people can leave comments and give stars on business's location, decoration, service, etc. I intend to find commonplaces among high-rates businesses, including the type of business, location and the quality of reviews. Besides, like we discussed in the first few lectures, the star rating system has been considered harmful, and the Yelp.com just uses the star rating system. Therefore, I will discuss the rating system by comparing the aggregated rating and the actual rating. Last but not least, I intend to use text sentiment analysis to test the user's review in a specific state in order to get the polarity and subjectivity of the reviews of the high rated business on Yelp.

Dataset Example

The dataset I am using is from Yelp.com, which is a subset of businesses, reviews, and user data. I chose to focus on the following two JSON files in the dataset. All business registered in Yelp has its own business ID, and we can retrieve business information and the related reviews by the unique business ID. Since the dataset is 5GB and it would take a long time to read all the records, I

specifically chose the top 5 states with the largest number of business IDs, and they are AZ, NV, NC, OH, and PA.

Review.json file contains full review text data including the user_id that wrote the review and the business_id the review is written for.

```
{
   // string, 22 character business id, maps to business in business.json
   "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

   // integer, star rating
   "stars": 4,

   // string, date formatted YYYY-MM-DD
   "date": "2016-03-09",

   // string, the review itself
   "text": "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good.",
}
```

Business.json contains the business information, including the location and rating (star).

```
{
   // string, 22 character unique string business id
   "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

   // string, the city
   "city": "San Francisco",

   // string, 2 character state code, if applicable
   "state": "CA",

   // string, the postal code
   "postal_code": "94107",

   // float, star rating, rounded to half-stars
   "stars": 4.5,

   // an array of strings of business categories
   "categories": [
      "Mexican",
      "Burgers",
      "Gastropubs"
   ],
}
```

The following data is the joint dataset about Arizona's population and household annual income by zip code from U.S. Census Bureau website. I chose Arizona because this state has the most number of business IDs as well as the corresponding reviews. This dataset is sorted by the annual income, and we can get to know the wealthy level of a specific area in AZ. By combining with the Yelp dataset, we will have a clearer mind of the relationship between the business rating in a area and the wealthy level in this area.

11	*85749*	*"32.405505, -110.667429"*	*"Tucson, Arizona"*	*"18,267"*	*"$78,026.00 "*
12	*85310*	*"33.713148, -112.174366"*	*"Glendale, Arizona"*	*"22,916"*	*"$77,116.00 "*
13	*85254*	*"33.614359, -111.952292"*	*"Scottsdale, Arizona"*	*"49,904"*	*"$73,758.00 "*
14	*85296*	*"33.308990, -111.759970"*	*"Gilbert, Arizona"*	*"32,732"*	*"$72,019.00 "*

Result

Aggregated Data from Business.json

I intend to start with the aggregated business information. Firstly, I select 5 states in United States which has the largest number of reviews from Business.json.

```
business_location = []

for line in gzip.open("data/business.json.txt.gz", 'rt',
encoding='utf-8'):

    business_info = json.loads(line.strip())

    if business_info["state"] in list_states:

        business_location.append(business_info["state"])
```

After counting the number of business records, we can list top five states (AZ, NV, NC, OH and PA), and we can review data from these states. Figure 1 shows the number of business IDs in the five states. (final_cling.ipynb, In [6])

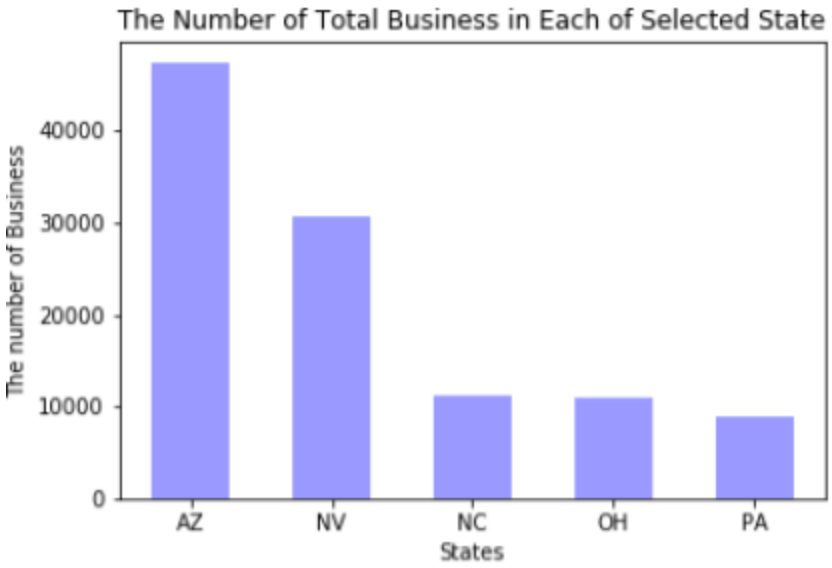

Figure 1

Then, we can retrive all the business ID from these five states. Based on the business ID, we can get these states' rating information.

```
for line in gzip.open("data/business.json.txt.gz", 'rt',
encoding='utf-8'):

   business_info = json.loads(line.strip())

   for k in range(len(selected_states)):

      if business_info["state"] == selected_states[k]:

        business_id_states[k].append(business_info["business_id"])
```

```
business_rates_states[k].append(business_info["stars"])
```

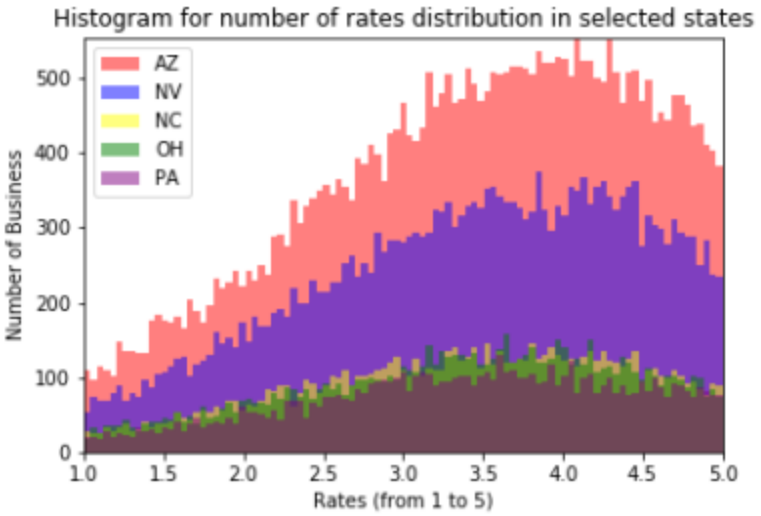

Figure 2

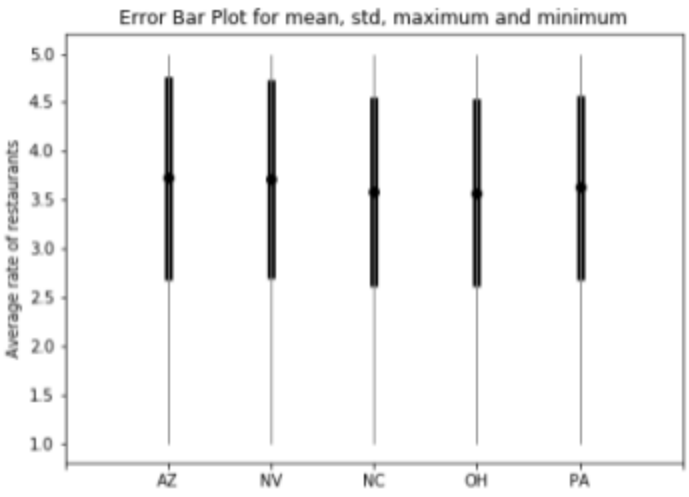

Figure 3

Figure 2 (final_cling.ipynb, In [10]) represents the histogram distribution of business average rating among five selected states. We can see the star ratings are from 1 to 5, and the average is approximately around 3.6 to 3.7.

Figure 3 (final_cling.ipynb, In [11]) is the Error Bar plot. The light line is the range of rating for each states (from 1 to 5), and the black think line for each state represents the standard deviation. Therefore, we can see the most of business's average rating are around from 2.6 to 4.6, which means people tend to leave rates around 3 and 4.

The next step is analyzing the high rated businesses' common grounds; for example, category. Figure 4 (final_cling.ipynb, In[20]) shows the most popular business in each states. In Arizona, the most popular business type is Home Service. (Figure 4 is in the figure folder since it's too large to put in the document). Figure 5 (final_cling.ipynb, In[22]) shows the most common and popular business category in the selected five states.

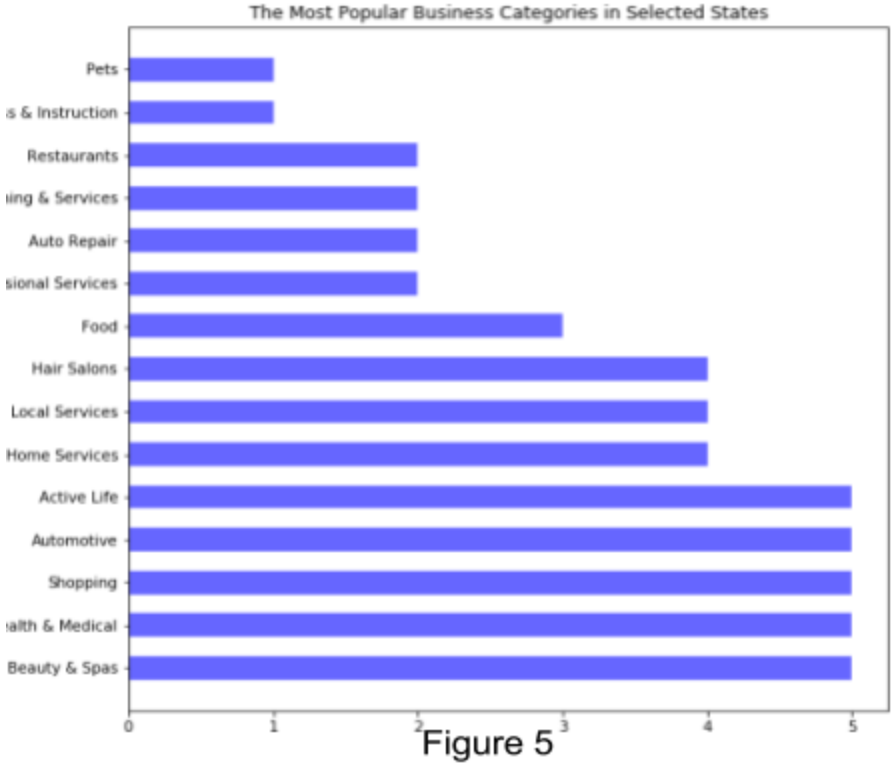

Figure 5

Next, I am interested in the location of high-rated business in the selected area, and I will use Arizona as an example, because Arizona has the most number of business records as well as reviews. The next block of code is used to read and save data from the joint dataset.

```
lines = []

with open("data/zipcode_wealthlevel.txt", 'r', encoding="utf-16")
as file:

    for l in file:

        lines.append(l.strip().split('\t'))


zipcode = []; cities = []; income = []

for i in range(len(lines)):

    zipcode.append(lines[i][1])

    cities.append(lines[i][3].split(",")[0].replace('"', ""))

    income.append(float(sub(r'[^\d.]', '',lines[i][5].replace('"',
"").replace(' ', ""))))
```

I generated the list of businesses with high average rating. Based on the business id, I am able to get the corresponding zip code from business.json. After sorting the zip code from the lowest income to highest income, I generated the following Figure. Figure 6 (final_cling.ipynb, In[30]).

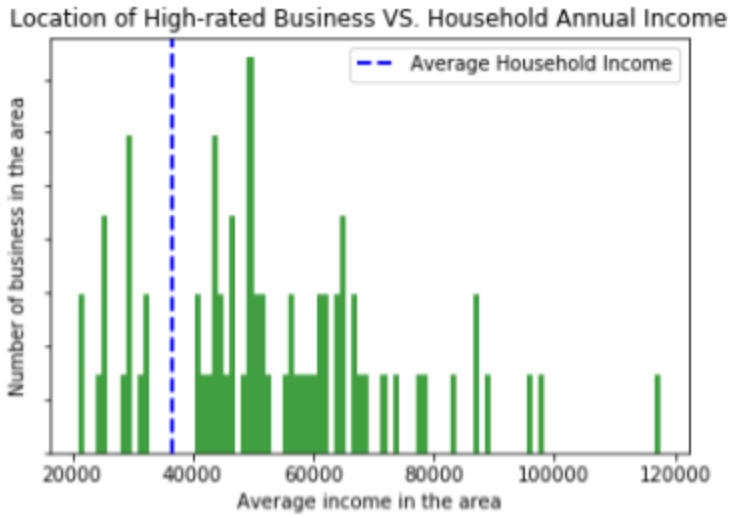

Figure 6

Actual Data from Review.json

In the next section, we will analysis the actual review from users. First thing is reading data from review.json.

```
actual_stars_AZ = []

reviews_AZ = []


for line in gzip.open("data/review.json.txt.gz", 'rt',
encoding='utf-8'):

    review_info = json.loads(line.strip())

    if review_info["business_id"] in business_id_states[0]:

        actual_stars_AZ.append(review_info["stars"])

        reviews_AZ.append(review_info["text"])
```

By plotting the actual review, I generated the Figure 7 (final_cling.ipynb, In[32]).

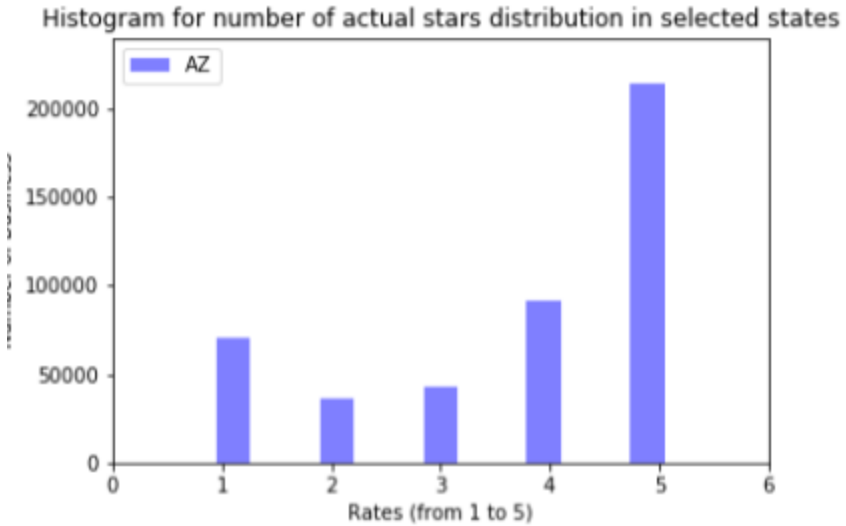

Figure 7

The difference of actual rate and average rate is apparent. However, some users tend to only leave stars without comments, and some people's comments are really subjective, so the rating may not be accurate. From the review itself, I want to compare the result of text sentiment analysis with the actual rating. Next, by using the Natural Language Processing Python library, I want to take a closer look at the sentiment analysis on actually reviews. Figure 8 (final_cling.ipynb, In[36]) is generated by the sentiment analysis result.

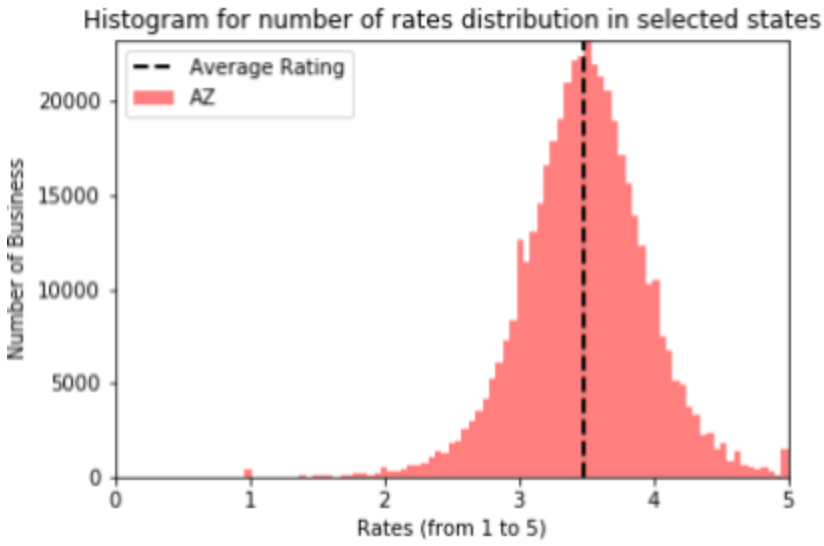

Figure 8

Then, I want to analysis the subjectivity of reviews. The following figures Figure 9 (final_cling.ipynb, In[37]) and Figure 10 (final_cling.ipynb, In[38]) are the subjectivity of Yelp reviews. From the figure we can see, the subjectivity scale is from 0 to 1 and the average is 0.5 (neutral). More than half of the reviews are subjective, which means people tend to give emotional and inaccurate grading.

Figure 9

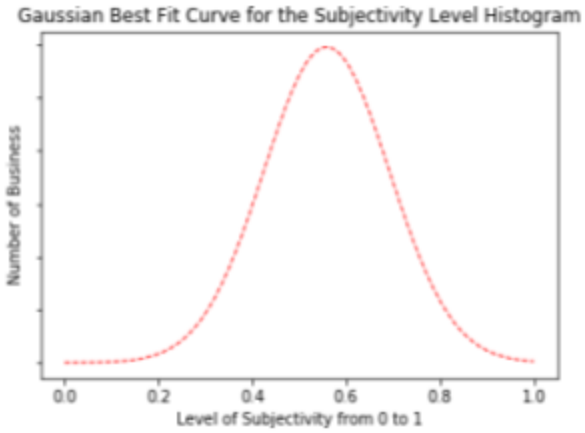

Figure 10

Conclusion & Discussion

Characteristics of High-rated Business

From the above plots, we can see in the selected states, residents tend to give high rates in the following categories: Active Life, Automotive, Shopping, Health & Medical and Beauty & Spa among the five selected states. Figure 5 reflects that people nowadays generally pay more attention to life quality because the five categories listed above are related to personal care (Health & Medical and Beauty & Spa) and living standard (Active Life, Automotive, Shopping). However, there are examples showing people from one state have a preference for certain business; for example, only people who live in North Carolina generally give Pets and Fitness related business high score.

Besides, based on the location of those high-rated businesses in Arizona, we can also find common places between those businesses. By calculating the average household income according to the census data, the average annual income in an area is 36611.15. From Figure 6, we can see that

most of high-rated business locations' annual income are above the average line, which means restaurants and other businesses in the wealthy area are more popular and people tend to give a high score for these businesses.

Summing up, high-rated businesses are more likely to exist in the relatively wealthy area; and high-rated businesses category reflects that people nowadays care more about personal care and living quality.

The Disadvantage of Star Rating System

From the comparison of Figure 2 and Figure 7, the difference is obvious. More people tend to give a polarized rating, which means people either give a full score if they really like the business or low score if they hate them. Therefore, it is safe to say that the result of star rating system is not accurate enough. Like we discussed in the first few lecture, a simple thumb up and thumb down would make more sense. In addition, I am interested in analyzing what is the sentiment analysis score for the reviews. Some people may leave the comment only, and some may give stars without comments, which makes it useful to check the difference between the result of sentiment analysis and the aggregated rating.

"*TextBlob is a Python library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.*" I used the well-trained classifier provided by TextBlob to test the polarity and subjectivity of the reviews in Arizona. Figure 8 shows the average score of rating by sentiment analysis is a bit less than the aggregated one, which can be tolerated. However, the average of the subjectivity score is not close to neutral (0.5), which means people

may leave emotional comments and inaccurate rating. Therefore, this further proves the star rating system is not really helpful.

Last but not least, the interesting thing is both polarity and subjectivity curves are close to the Gaussian function. Therefore, I plotted the Figure 10 as the best fit function to the histogram Figure 9.

$$f(x) = ae^{-(x-b)^2/2c^2}$$

Citation & Reference

Yelp Dataset Documentation

https://www.yelp.com/dataset/documentation/json

TextBlob Documentation

https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis

Gaussian Function

https://en.wikipedia.org/wiki/Gaussian_function