CS839 Data Science Spring 2018
<Group 9>
Ya-Chun Yang, yang364@wisc.edu
Han Wang, hwang729@wisc.edu
Yanghui Kang, ykang38@wisc.edu

Stage 4 Report - Integration and Analysis

1. INTRODUCTION

In the previous stage, a Naive Bayes matcher was built to match the restaurant data from Yelp and TripAdvisor. In this stage, we combined the two tables that contain matched entities, and performed several analysis on the integrated table to compare the distribution of ratings and reviews in the same restaurant from Yelp and TripAdvisor. We found that ratings and review count has positive correlations in most cases, though Yelp has generally lower rating and Tripadvisor has much less reviews.

2. DATA INTEGRATION

In the data integration step, we merged the tuple pairs from the previous stage classified by Naive Bayes matcher. The process is described as follows:

(1) We first collected the positive tuple pairs classified by Naive Bayes matcher
(2) We created two new features, [cat_loc] and [cat_food], to arrange and organize the numerous values in features [category_1] and [category_2] listed both in tables A and B. The former indicates the origin of food, and the later represents the food category. Next, we applied the value based on table A in the integrated process.
(3) For the feature [category], we first replaces some rare value with more common value. The replacing rules are: (a) Replacing 'ramen' with 'japanese'. (b) Replacing 'brazilian' and 'french' with 'other'. (c) Replacing 'indian' and other values that are not listed in our defined categories with 'asian other'. Next, we applied the value based on the feature [cat_food] in table A.
(4) For the feature [price] and the features related to open hours among the week, we used the value from table A if applicable. Otherwise, we filled in the value from table B.
(5) For the feature [rating], we averaged the values from table A and table B.
(6) For the feature [review_count], we summed the values from table A and table B.
(7) We renamed certain feature names to make them more readable. The final features and the description are described in the next section.

3. DATA DESCRIPTION

The integrated data table E contains 894 tuples, each corresponding to a restaurant in the Los Angeles area. There are 31 attributes in the table.

- Name [text]: name of restaurant
- City [text]: the city where the restaurant located
- Zipcode [5-digit]
- Address [text]
- Phone [10-digit]
- Category [text]: combined category based on origin country and food type
- Category_loc [text]: restaurant category based on country/location
- Category_food [text]: restaurant category based on food type
- Price [number of dollar signs]: price level of the restaurant (from Yelp)
- Rating [number]: Average rating based on Yelp and Tripadvisor
- Review_count [Number]: total count of review from Yelp and Tripadvisor
- Price_yelp [number of dollar signs]: price level from the Yelp dataset
- Rating_yelp [1 to 5]: rating from the Yelp dataset
- Review_count_yelp [number]: count of reviews from the Yelp dataset
- Price_trip [number of dollar signs]: price level from the Tripadvisor dataset
- Rating_trip [1 to 5]: rating from the Tripadvisor dataset
- Review_count_trip [number]: count of reviews from the Tripadvisor dataset
- A set of attributes for open hours (14 intotal)
    - hours_<day>_<open/close>: [HHMM] in 24 hour format
    - Examples:
        - hours_sun_open = 1100
        - hours_sun_close = 200

Below are the four example tuples in Table E.

Id	Name	City	Zipcode	Address	Phone
0	Hae Jang Chon Korean BBQ Restaurant	Los Angeles	90020	3821 W 6th St	2133898777
1	Kang Hodong Baekjeong	Los Angeles	90020	3465 W 6th St	2133849678
3	EMC Seafood and Raw Bar	Los Angeles	90020	3500 W 6th St	2133519988
4	Beer Belly	Los Angeles	90020	532 S Western Ave	2133872337

Category	Category_loc	Category_food	Price	Rating	Review_count

barbecue	korean	barbecue	2	4.25	4048
barbecue	korean	barbecue	2	4.5	3646
seafood	other	seafood	2	4.25	2963
american	american	other	2	4	2274

Price_yelp	Rating_yelp	Review_count_yelp	Price_trip	Rating_trip	Review_count_trip
2	4	3970	2.5	4.5	78
2	4.5	3533	2.5	4.5	113
2	4	2890	2.5	4.5	73
2	4	2165	2.5	4	109

hours_sun_open	hours_sun_close	hours_mon_open	hours_mon_close	hours_tue_open	hours_tue_close	hours_wed_open	hours_wed_close	hours_thu_open	hours_thu_close	hours_fri_open	hours_fri_close	hours_sat_open	hours_sat_close
1100	200	1100	200	1100	200	1100	200	1100	200	1100	200	1100	200
1130	130	1130	130	1130	130	1130	130	1130	130	1130	130	1130	130
1100	1400	1600	2400	1100	1400	1600	2400	1100	1400	1600	200	1100	1400
1700	2300	1700	2300	1130	2400	1130	2400	1130	100	1130	100	1130	2300

4. ANALYSIS & DISCUSSIONS

4.1 Analysis about the Restaurants

In our integrated table, we have three different attributes describing the type the type of restaurants. The "cateogry_loc" attribute categories restaurants by their region of origin, such as "american", "japanese", "mexican". The "category_food" attribute differentiate restaurants by the type of food they serve, such as "barbecue", "seafood", "pizza". The "category" attribute combines the above two categories, because none of them is able to fully describe all the types of restaurants. We analyzed the the distribution of the categories (Table 1), the result indicated that the category “American” is the most among Los Angeles.

Table 1. The distribution of categories among Los Angeles

Category	Count
American	139
Breakfast & Brunch	74
Cafe	67
Sushi	60
Seafood	59
Pizza	57
Mexican	54
Other	43
Fast Food	43
Barbecue	38
Japanese	34
Italian	31
Bars	30
burger	29
Mediterranean	27
Thai	27
Steakhouse	26
Chinese	20
Vietnamese	13
Asian Other	12
Korean	11

4.2 Comparison Between Websites

Based on the dataset, we used correlation, OLAP style analysis, and various data visualization tools to answer the question: how the ratings and reviews about the same restaurant differ from Yelp and TripAdvisor?

Comparison of restaurant ratings between Yelp and TripAdvisor

In this section, we focused on the measurement of rating differences from Yelp and TripAdvisor. We analyzed the difference from both all the restaurants and the restaurants in different categories.

First, we computed the count of restaurants within each pair of rating categories in Yelp and TripAdvisor. This is an OLAP style analysis, by rolling up the restaurant ID dimension.
We presented the results in a barplot (Figure 1), grouped by ratings in Yelp, and each bar shows the percentage of restaurants falling in a rating group of Tripadvisor. As shown in the figure 1, we found that the restaurant ratings of TripAdvisor is generally correlated with those of Yelp. The higher of restaurant rating in Yelp, the percentage of restaurant with high rating in TripAdvisor is also higher. For example, 5 rating restaurants in TripAdvisor are mostly rated 4.5 by Yelp. The majority of 4.5 rated restaurants in TripAdvisor is also rated 4.5 by Yelp. However, there is also some discrepancies. It is not hard to find out that TripAdvisor rating is generally higher than those of Yelp. There are also some interesting cases. The restaurants that are rated 5 in TripAdvisor is only rated 3.5 in Yelp. However, we could see that these restaurants were not reviewed by a lot of people in both websites, especially TripAdvisor, therefore, the ratings might be biased.

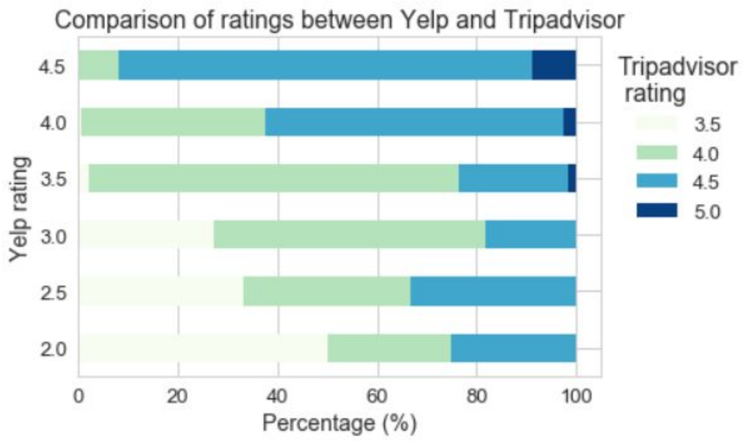


Figure 1. Comparison of ratings between Yelp and TripAdvisor.

Next, we compared the ratings from the two websites across different restaurant categories. The restaurant ratings are aggregated by mean/min/max within each category group as shown in Figure 2. This is an OLAP style analysis by rolling up the restaurant ID dimension to hierarchy of category. The plot shows the mean and range of ratings from each website grouped by restaurant categories. The y-axis shows the categories ranked by their average ratings in Yelp. From the visualization, it is clearly shown that ratings in TripAdvisor are generally much higher than those in Yelp, indicating that the users should be cautious to use ratings interchangeably or compare them in the two sources. Both websites rate mediterranean and asian other (Indian, Laos, Tibet, Nepal, etc.) restaurants the highest and gives bars the lowest rate. However, in the middle range, there is a substantial inconsistency between the ratings in the two websites. For example, Korean restaurants received high rate from Yelp but relatively low in TripAdvisor; fast food is not favored by Yelp users but are rated high in TripAdvisor.

The plot in Figure 2 also compares the range of ratings across various restaurant categories and across websites. Firstly, Yelp ratings have large variance than those of TripAdvisor, partly due to the larger review counts in Yelp. Secondly, some types of restaurants have generally high ratings in the two websites, such as Japanese and Barbecue, while others have ratings that vary a lot. For example, Mediterranean and Korean. This implies that if you are in Los Angeles and pick a random Japanese restaurant, it is more likely that the restaurant is pretty good. However, if one picks a random Korean restaurant, it is harder to predict how good it will taste.

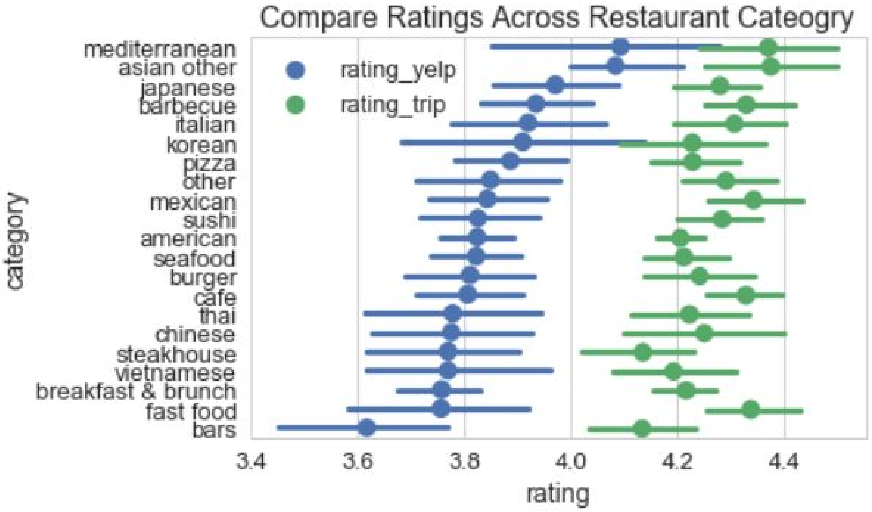


Figure 2. The comparison of ratings across restaurant categories

We also analyzed the ratings of restaurants from the two websites across different categories by country/region (Figure 3) and by food (Figure 4) using similar analysis. The result indicated that the best rated restaurant category (by region) from Yelp are Mediterranean, Indian, and Korean, while TripAdvisor users favored French, Mexican food (Figure 3). From the result in different categories by food (Figure 4), we found that both websites agree on the high ratings of ramen and barbecue. In addition, Yelp users likes pizza, while Tripadvisor users favors fast food, cafe, and sushi.

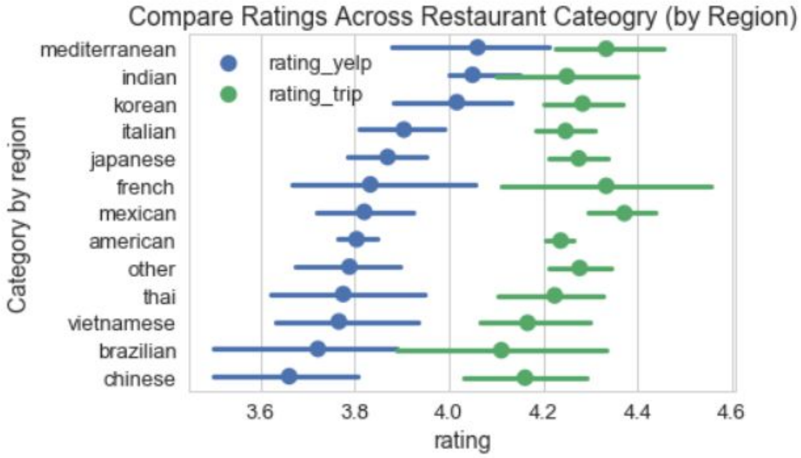


Figure 3. The comparison of ratings across restaurant categories by region

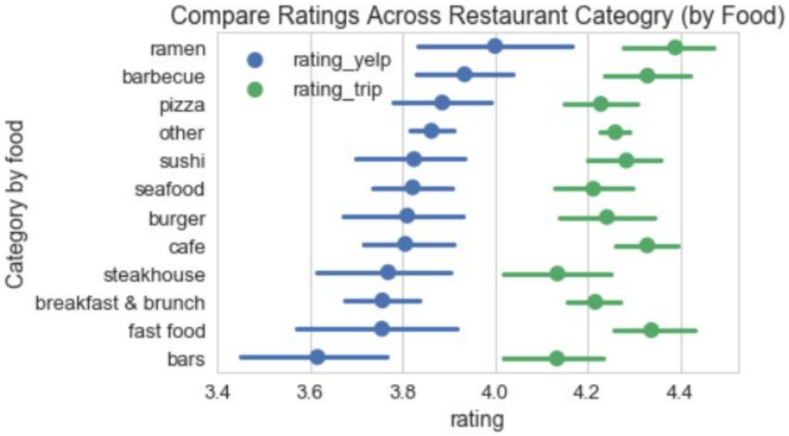


Figure 4. The comparison of ratings across restaurant categories by food

Correlation of restaurant ratings from Yelp and Tripadvisor

The above analysis shows that while Yelp and TripAdvisor agrees on the top and bottom rated restaurant categories, there are a lot of discrepancies in the middle range. Thus, we performed a correlation analysis based of the average rating of restaurant by category between the two websites. Note that, we did not perform the correlation analysis directly on instances of restaurants, because the ratings for each restaurant are non consecutive, meaning that they are categorical attribute rather than quantitative. The correlation coefficient between the ratings (aggregated by restaurant category) of Yelp and TripAdvisor is 0.68, indicating a positive correlation. The bubble plot shows the scatters between rating or the two restaurant and sized by the number of restaurants in each category (Figure 5). From the result, it is clear that there are still many types of restaurant with different ratings from the two websites. We thus presented two visualizations that identifies such restaurants by region (Figure 6) and by food (Figure 7). The result indicated that Mexican, French, Cafe, and Fast Food are relatively more favored by TripAdvisor users, while Yelp users prefers Steakhouse, Indian, and American foods more than those of Tripadvisor.

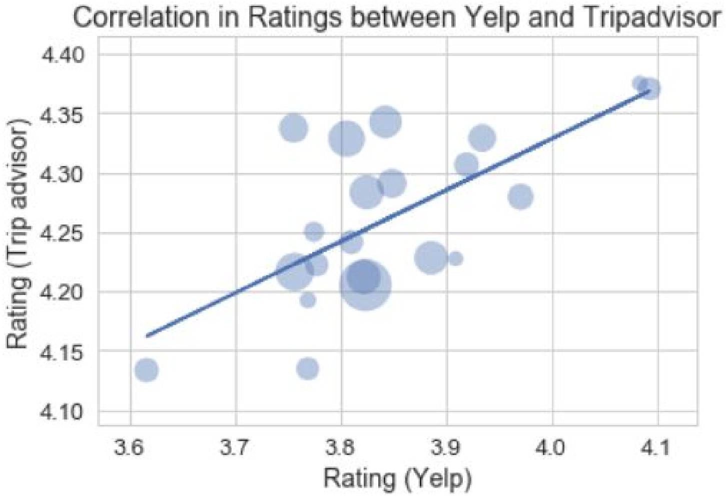


Figure 5. The correlation in ratings between Yelp and TripAdvisor

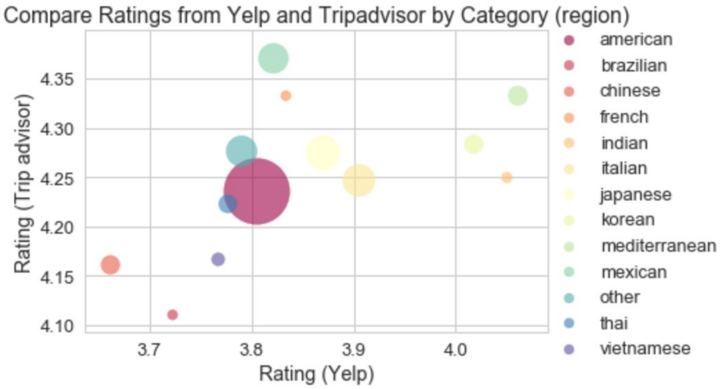


Figure 6. The correlation in ratings between Yelp and TripAdvisor by region

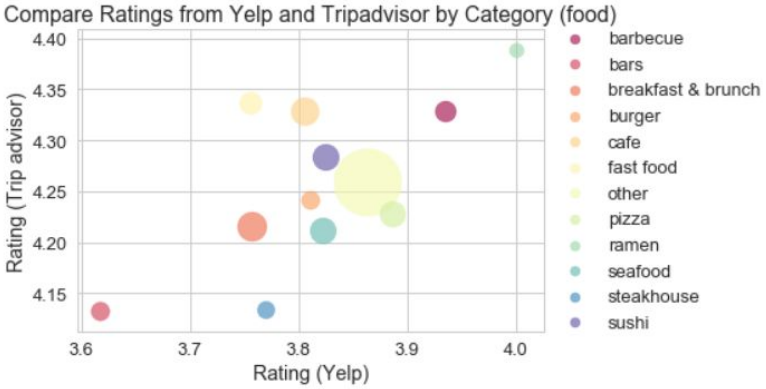


Figure 7. The correlation in ratings between Yelp and TripAdvisor by food

Analyze the correlation of review counts between Yelp and TripAdvisor

In this section, we examine the correlation of restaurant review counts between the two websites. Specifically, we try to answer the following questions: Do restaurants that got more reviews in Yelp also had more reviews in TripAdvisor? Does either website has more restaurant reviews in general?

We first analyzed correlations and distributions between the restaurant review count in the two website. As shown in Figure 8 which displayed a scatter plot of the restaurant review count and histogram from each website, we could conclude that Yelp generally has much more reviews compared to TripAdvisor. In addition, there is a positive correlation between the two website, but the correlation is not very high (r = 0.51). Furthermore, we performed a dice operation to look into restaurants that have the most reviews in both websites. There are three of them, Pink's Hog Dogs, The Griddle Cafe, and Philippe the Original (Table 2). These restaurants are not expensive and also have a good rating.

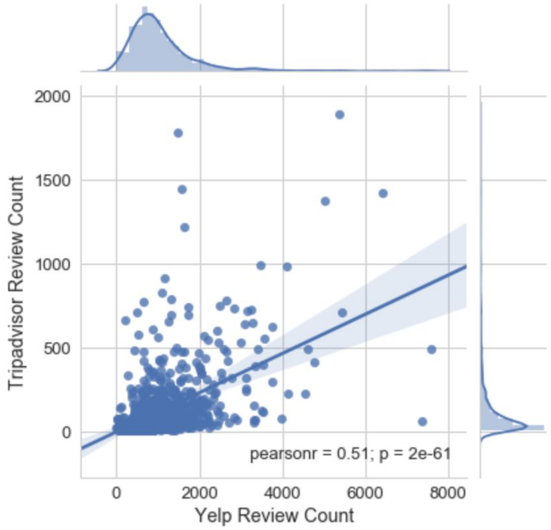
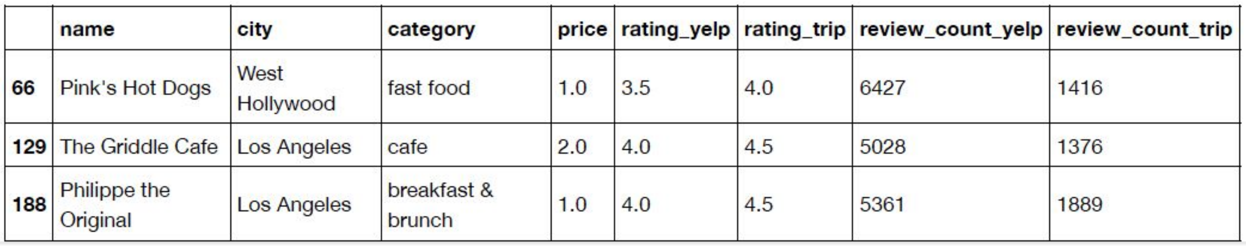


Figure 8. The correlation in review count between Yelp and TripAdvisor

Table 2. The list of restaurants with most reviews in both websites



	name	city	category	price	rating_yelp	rating_trip	review_count_yelp	review_count_trip
66	Pink's Hot Dogs	West Hollywood	fast food	1.0	3.5	4.0	6427	1416
129	The Griddle Cafe	Los Angeles	cafe	2.0	4.0	4.5	5028	1376
188	Philippe the Original	Los Angeles	breakfast & brunch	1.0	4.0	4.5	5361	1889

Relationship between review count and rating

In this section, we performed analysis to measure the relationship between restaurant rating and review count. Next, we determined how such relationship differs between the two websites. Our general assumption is: If a restaurant has high rating, it is likely to attract more customers who will provide their review.

We analyzed the distribution and the mean of review count for each rating category in both Yelp and TripAdvisor. As shown in Figure 9, the result indicated that the ratings range of Yelp (2 to 4.5) is much wider than that of TripAdvisor (3.5 to 5.0). Note that the y-axis is log-transformed to highlight the differences. Moreover, if we set aside the highest rating group in both websites, there is a general increasing trend of review count as with the increase of rating, which demonstrates our hypothesis that highly rated restaurants attracts more customers to post their reviews. It is also noteworthy that for Yelp, most of the reviews are concentrated in the high rating end (3.5 to 4.5), while for TripAdvisor, most of the reviews are given to 4 to 4.5 rated restaurants.

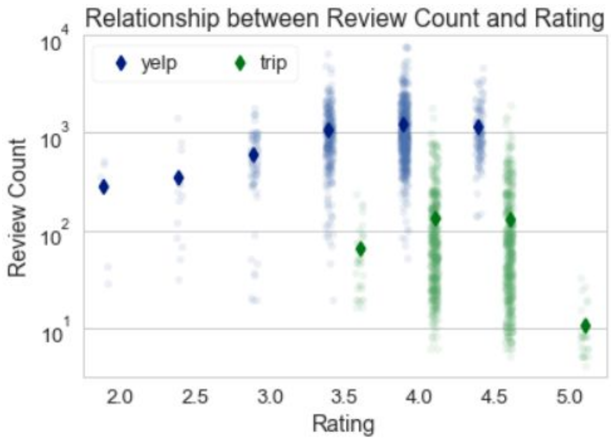


Figure 9. The relationship between review count and rating in Yelp and TripAdvisor

Relationship between review count and price level

Following the previous section, we next evaluated the correlation between review count and price level. We assume that fancy restaurants ($$$$, four dollar signs) might not have as much customers as mid-class ones ($$). In our analysis, there is clearly an increasing trend of review count with the increase of price level in both websites (Figure 10). In fact, this finding is not as we expected. It is interesting to note that a few very pricey restaurants have the most reviews in both websites. Therefore, we can suggest that the review count is associated with both price and rating from our analysis. We then presented the heatmap analysis to show the interaction between rating and price (Figure 11).

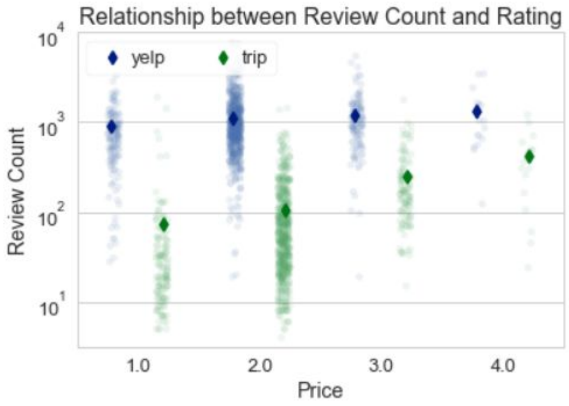


Figure 10. The relationship between review count and price in Yelp and TripAdvisor

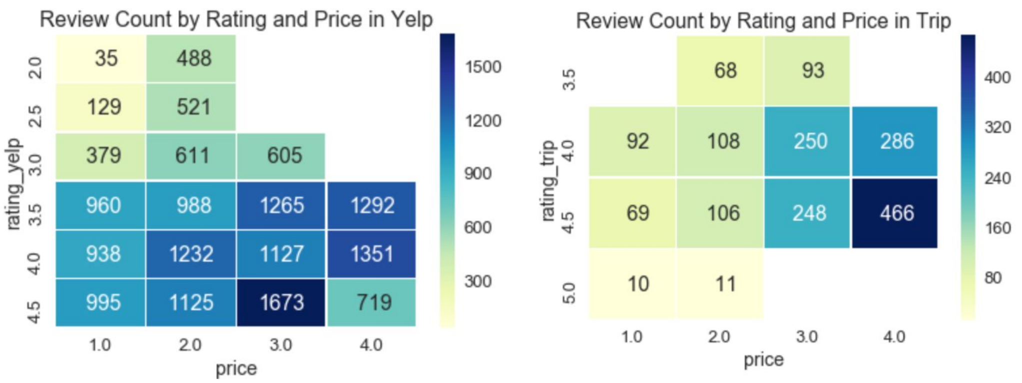


Figure 11. The heatmap analysis between rating and price in Yelp (left) and TripAdvisor (right)

Relationship between rating and price level

In this section, we examined the relationship between restaurant rating and its price level. We next determined whether such relationship differs between the two websites. Our general hypothesis is that there is a positive correlation between price and rating.

As shown in Figure 12, the highest ratings occurs in the highest price level in both websites, as we expected. However, the cheapest restaurants do not generally have low ratings. In TripAdvisor, the cheapest restaurants on average have the same ratings as the most expensive ones. In Yelp the one dollar sign ($) restaurants have an average rating of 3.8, which is similar to three dollar sign ($$$) restaurants. We also found that two dollar sign ($$) restaurants have the lowest rating in both websites.

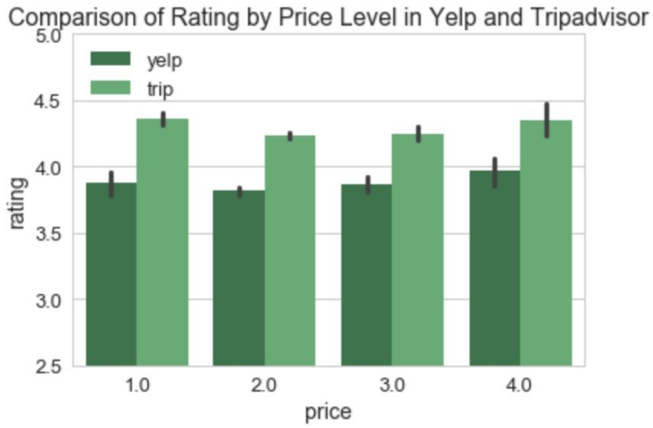
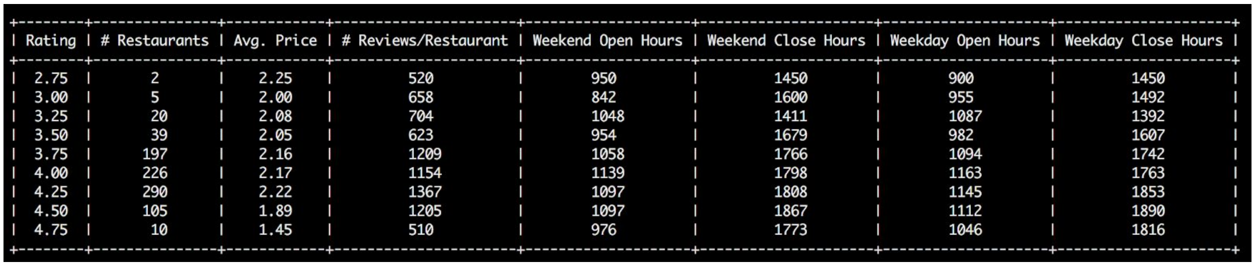


Figure 12. Comparison of rating by price level in Yelp and TripAdvisor

Relationship between hours, rating, price, and restaurant categories

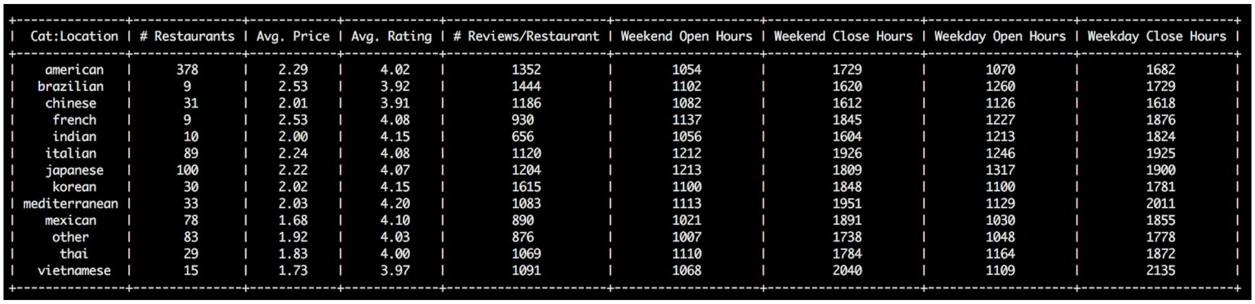
In this section, we performed OLAP analysis to examine the relationship between hours and another features. First, we investigated the relationship between rating and hours. We measured the average of weekend/weekday open/close hours by rolling up hours columns. The data indicated that the higher rating tend to open and close late in both weekend and weekday (Table 3). Next, we determined the relationship between restaurant categories and hours. The result indicated that vietnamese and mediterranean restaurant tend to close late (Table 4). Steakhouse restaurants set apparently higher price than other kind of restaurants (Table 5). Cafe surprisingly open the earliest (even earlier than breakfast & brunch) no matter in weekend or weekday (Table 5). Burger tend to close late during weekend, while steakhouse close late during weekday (Table 5).

Table 3. The relationship between rating and hours in Yelp and TripAdvisor



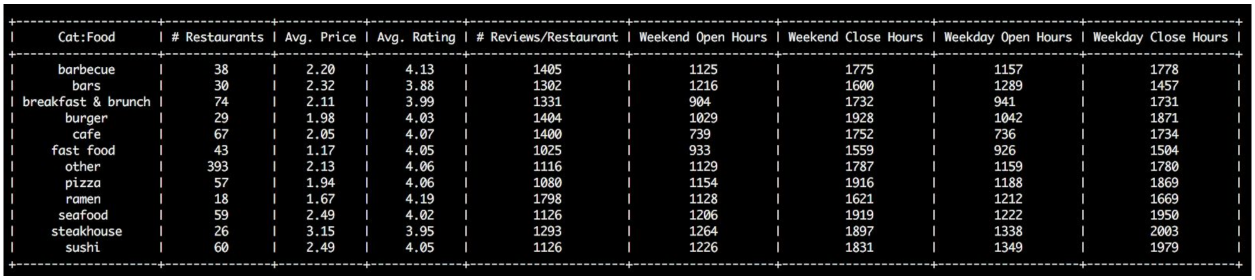
```
+--------+---------------+------------+----------------------+--------------------+---------------------+--------------------+---------------------+
| Rating | # Restaurants | Avg. Price | # Reviews/Restaurant | Weekend Open Hours | Weekend Close Hours | Weekday Open Hours | Weekday Close Hours |
+--------+---------------+------------+----------------------+--------------------+---------------------+--------------------+---------------------+
|  2.75  |       2       |    2.25    |         520          |        950         |        1450         |        900         |        1450         |
|  3.00  |       5       |    2.00    |         658          |        842         |        1600         |        955         |        1492         |
|  3.25  |      20       |    2.08    |         704          |        1048        |        1411         |        1087        |        1392         |
|  3.50  |      39       |    2.05    |         623          |        954         |        1679         |        982         |        1607         |
|  3.75  |      197      |    2.16    |         1209         |        1058        |        1766         |        1094        |        1742         |
|  4.00  |      226      |    2.17    |         1154         |        1139        |        1798         |        1163        |        1763         |
|  4.25  |      290      |    2.22    |         1367         |        1097        |        1808         |        1145        |        1853         |
|  4.50  |      105      |    1.89    |         1205         |        1097        |        1867         |        1112        |        1890         |
|  4.75  |      10       |    1.45    |         510          |        976         |        1773         |        1046        |        1816         |
+--------+---------------+------------+----------------------+--------------------+---------------------+--------------------+---------------------+
```

Table 4. The relationship between restaurant categories by location and hours in Yelp and TripAdvisor



```
+---------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
|  Cat:Location | # Restaurants | Avg. Price | Avg. Rating | # Reviews/Restaurant | Weekend Open Hours | Weekend Close Hours | Weekday Open Hours | Weekday Close Hours |
+---------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
|    american   |      378      |    2.29    |    4.02     |         1352         |        1054        |        1729         |        1070        |        1682         |
|   brazilian   |       9       |    2.53    |    3.92     |         1444         |        1102        |        1620         |        1260        |        1729         |
|    chinese    |      31       |    2.01    |    3.91     |         1186         |        1082        |        1612         |        1126        |        1618         |
|     french    |       9       |    2.53    |    4.08     |         930          |        1137        |        1845         |        1227        |        1876         |
|     indian    |      10       |    2.00    |    4.15     |         656          |        1056        |        1604         |        1213        |        1824         |
|    italian    |      89       |    2.24    |    4.08     |         1120         |        1212        |        1926         |        1246        |        1925         |
|    japanese   |      100      |    2.22    |    4.07     |         1204         |        1213        |        1809         |        1317        |        1900         |
|     korean    |      30       |    2.02    |    4.15     |         1615         |        1100        |        1848         |        1100        |        1781         |
| mediterranean |      33       |    2.03    |    4.20     |         1083         |        1113        |        1951         |        1129        |        2011         |
|    mexican    |      78       |    1.68    |    4.10     |         890          |        1021        |        1891         |        1030        |        1855         |
|     other     |      83       |    1.92    |    4.03     |         876          |        1007        |        1738         |        1048        |        1778         |
|      thai     |      29       |    1.83    |    4.00     |         1069         |        1110        |        1784         |        1164        |        1872         |
|   vietnamese  |      15       |    1.73    |    3.97     |         1091         |        1068        |        2040         |        1109        |        2135         |
+---------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
```

Table 5. The relationship between restaurant categories by food and hours in Yelp and TripAdvisor



```
+--------------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
|      Cat:Food      | # Restaurants | Avg. Price | Avg. Rating | # Reviews/Restaurant | Weekend Open Hours | Weekend Close Hours | Weekday Open Hours | Weekday Close Hours |
+--------------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
|      barbecue      |      38       |    2.20    |    4.13     |         1405         |        1125        |        1775         |        1157        |        1778         |
|        bars        |      30       |    2.32    |    3.88     |         1302         |        1216        |        1600         |        1289        |        1457         |
| breakfast & brunch |      74       |    2.11    |    3.99     |         1331         |        904         |        1732         |        941         |        1731         |
|       burger       |      29       |    1.98    |    4.03     |         1404         |        1029        |        1928         |        1042        |        1871         |
|        cafe        |      67       |    2.05    |    4.07     |         1400         |        739         |        1752         |        736         |        1734         |
|     fast food      |      43       |    1.17    |    4.05     |         1025         |        933         |        1559         |        926         |        1504         |
|       other        |      393      |    2.13    |    4.06     |         1116         |        1129        |        1787         |        1159        |        1780         |
|       pizza        |      57       |    1.94    |    4.06     |         1080         |        1154        |        1916         |        1188        |        1869         |
|       ramen        |      18       |    1.67    |    4.19     |         1798         |        1128        |        1621         |        1212        |        1669         |
|      seafood       |      59       |    2.49    |    4.02     |         1126         |        1206        |        1919         |        1222        |        1950         |
|     steakhouse     |      26       |    3.15    |    3.95     |         1293         |        1264        |        1897         |        1338        |        2003         |
|       sushi        |      60       |    2.49    |    4.05     |         1126         |        1226        |        1831         |        1349        |        1979         |
+--------------------+---------------+------------+-------------+----------------------+--------------------+---------------------+--------------------+---------------------+
```

5. LEARNINGS & CONCLUSIONS

From the above analysis, we derived several interesting insights in the restaurants in Los Angeles from Yelp and TripAdvisor, which are summarized as follows:

(1) We found that the the ratings in different restaurants have positive correlations in most cases in both websites, although the Yelp has lower ratings overall.

(2) The categories Mediterranean and Asian other (Indian, Laos, etc.) were rated highest by both Yelp and TripAdvisor users.

(3) Generally speaking, there is a positive relationship between review count and both rating and price level in both websites.

(4) We can also found the positive correlations between rating and price level.

In addition, we learn a lots from this project. First of all, we understand how to match two entities from different sources which is more complicated and more difficult than string matching. From the integrated table and the analysis, we obtained numerous findings. Secondly, we know how to view the data from various concepts, and how to dig into the data using different analysis, such as OLAP and classification. We also learn how to use Python package 'cube' to perform OLAP analysis. Furthermore, we have thorough grasps of the usage of py_entitymatching in the data matching, which is practical and interesting.

5. FUTURE PROPOSALS

From our analysis, we found the positive correlations between the ratings in both websites. We also analyzed the distributions of different restaurant categories. With these information in hand, we propose that we can further undertake classification tasks to predict the rating in either website. For example, we can try to answer the following questions:

(1) If we know the rating of a restaurant in Yelp, could we predict the approximate rating that would be rated by TripAdvisor users?

(2) If we know the rating of a restaurant in both Yelp and TripAdvisor, could we predict the price of the restaurant?

(3) If we know the rating across the restaurant categories, could we predict the data is from Yelp or TripAdvisor?

In addition, there are several features that we did not consider in our analysis, such as the spatial information and its relationship with restaurant type and ratings. The location of restaurant might have impact on the number and types of customer it serves as well as its ratings. With the location of a restaurant, the population composition, socio-economic status, the type of restaurant, price level, as well as information of other nearby restaurants, we might be able to predict the ratings and popularity for a given restaurant. Such information could be helpful to people who plan to open a new restaurant and would like to select a location for it.

Script for table integration

```
import py_entitymatching as em
import pandas as pd

#%% Matching tuples

# read tables
A = em.read_csv_metadata('A.csv', key='id')
B = em.read_csv_metadata('B.csv', key='id')

# Load blocked data pairs
C = em.read_csv_metadata('C.csv', key='_id',
                         ltable=A, rtable=B, fk_ltable='ltable_id', fk_rtable='rtable_id')

# Load the labeled data
G = em.read_csv_metadata('S_labeled.csv', key='_id',
                         ltable=A, rtable=B, fk_ltable='ltable_id', fk_rtable='rtable_id')

# Generate a set of features
F = em.get_features_for_matching(A.iloc[:, 1:8], B.iloc[:, 1:8],
validate_inferred_attr_types=False)

# Create a feature on the value of (price + rating), then compute Levenshtein
similarity
sim = em.get_sim_funs_for_matching()
tok = em.get_tokenizers_for_matching()
feature_string = """lev_sim(wspace(float(ltuple['price']) + float(ltuple['rating'])),
                     wspace(float(rtuple['price']) + float(rtuple['rating'])))"""
feature = em.get_feature_fn(feature_string, sim, tok)

# Add feature to F
```

```
em.add_feature(F, 'lev_ws_price+rating', feature)

# Convert the sample set into a set of feature vectors using F
H = em.extract_feature_vecs(G,
                            feature_table=F,
                            attrs_after='labe',
                            show_progress=False)

# impute missing values
H = em.impute_table(H,
                exclude_attrs=['_id', 'ltable_id', 'rtable_id', 'labe'],
                strategy='mean')

# Fit a Naive Bayes matcher
matcher = em.NBMatcher(name='NaiveBayes')
matcher.fit(table=H,
       exclude_attrs=['_id', 'ltable_id', 'rtable_id', 'labe'],
       target_attr='labe')

# Apply matcher to the whole dataset
Ht = em.extract_feature_vecs(C,
                            feature_table=F,
                            show_progress=False)
Ht = em.impute_table(Ht,
                exclude_attrs=['_id', 'ltable_id', 'rtable_id'],
                strategy='mean')
predictions = matcher.predict(table=Ht, exclude_attrs=['_id', 'ltable_id', 'rtable_id'],
              append=True, target_attr='predicted', inplace=False)

# save the final predictions
predictions.to_csv('P.csv', index = False)

#%% Merge columns to Table A and B
# get only the positive paris
```

```
predictions = pd.read_csv('P.csv')
P = predictions[predictions['predicted'] == 1]

# Add columns to E
E0 = P[['_id','ltable_id','rtable_id']]
E1 = pd.merge(E0, A, how = 'left', left_on = 'ltable_id', right_on = 'id')
E2 = pd.merge(E1, B, how = 'left', left_on = 'rtable_id', right_on = 'id',
              suffixes = ['_a','_b'])

# fill na: use hours of b to fill hours of a
E2.isnull().sum()
for i in range(14, 28, 1):
    E2.iloc[:,i] = E2.iloc[:,i].fillna(E2.iloc[:,i+25])
E2.isnull().sum()

# remove duplicated rows based on id of Table A
E2 = E2.drop_duplicates(subset = 'id_a')

# merge the category columns
# define two categories:
# category_loc: category by location
# category_food: category by food
cat_loc = ['american','japanese','mexican','italian','thai','chinese','korean',
           'vietnamese','mediterranean','french','indian','brazilian']
cat_food = ['barbecue','pizza','cafe','seafood','steakhouse','fast food','sushi',
            'ramen','breakfast & brunch','burger','bars']

E2['categories'] = E2[['category_1_b','category_2_b','category_1_a',
                       'category_2_a']].values.tolist()

def get_category_loc(row):
    """get the category based on ranked list of category_by_location
    """
    cat = 'other'
    for t in cat_loc:
        for item in row:
            if pd.isnull(item):
                continue
            if item.split()[0].lower() == 'american':
                return 'american'
            if item.lower() == 'pizza':
                return 'italian'
            if item.lower() == 'ramen':
                return 'japanese'
            if em.lev_dist(item.lower(), t) < 2:
                return t
    return cat

def get_category_food(row):
    """get the category based on a ranked list of food category
    """
    cat = 'other'
    for t in cat_food:
        for item in row:
            if pd.isnull(item):
                continue
            if item.split()[0].lower() == 'sushi':
                return 'sushi'
            if em.lev_dist(item.lower(), t) < 2:
                return t
    return cat

def add_asian_other(row):
    """Create a new category called "asian other"
    """
    if row['category'] == 'other':
        for cat in row['categories']:
            if pd.isnull(cat):
                continue
            for x in cat.split():
                if x.lower() == 'asian':
                    return 'asian other'
    return row['category']



E2['category_loc'] = E2['categories'].apply(get_category_loc)
E2['category_food'] = E2['categories'].apply(get_category_food)

# combine into a single category based primary on food, fill with location
E2['category'] = E2['category_food']
E2.loc[E2['category_food'] == 'other', 'category'] = E2.loc[E2['category_food'] ==
'other', 'category_loc']

# combine low count categories
E2['category'] = E2['category'].replace({'ramen':'japanese','brazilian':'other',
                         'french':'other','indian':'asian other'})
E2['category'] = E2.apply(add_asian_other, axis = 1)

Et = E2[['categories','category_loc','category_food','category']]
E2.category_loc.value_counts()
E2.category_food.value_counts()
E2.category.value_counts()

# merge rating, price, and review counts
E2['price'] = E2['price_a']
E2['price'] = E2['price'].fillna(E2['price_b']).round()
E2['rating'] = (E2['rating_a'] + E2['rating_b']) / 2
E2['review_count'] = E2['review_count_a'] + E2['review_count_b']

#%% Clean columns

# reorder columns
```

```
colindex = [0,3,28,4,29,33,9,7,10,56,54,55,57,58,59,11,12,13,36,37,38]
colindex.extend(range(14,28,1))
E = E2.iloc[:,colindex]

# rename some column names
colrelist = [5, 6, 7, 8]
colrelist.extend(range(21, 35, 1))
colnames = []
for i in range(E.shape[1]):
    if i in colrelist:
        coln = E.columns[i][0:-2]
    else:
        coln = E.columns[i]
    colnames.append(coln)

E.columns = colnames
E = E.rename(columns = {'price_a':'price_yelp','rating_a':'rating_yelp',
                        'review_count_a':'review_count_yelp',
                        'price_b':'price_trip','rating_b':'rating_trip',
                        'review_count_b':'review_count_trip',
                        'name_b':'name'})

E = E.drop(columns = ['_id','name_a'])
E.index.names = ['id']
E.to_csv('E_raw.csv')

Eclean = E.drop(columns = ['id_a','id_b'])
Eclean.to_csv('E.csv')
```