Correlation Discovery on Yelp & Zomato Restaurants in the Madison Area

Lokananda Dhage, Mary Feng, Varun Naik
CS 838 Project Stage 5

1. Statistics on Table E (E.csv)

The schema for Table E contains 19 columns: yelp_id, zomato_id, neighborhood, city, zipcode, stars, review_count, is_open, attributes, categories, hours, cuisines, average_cost_for_two, price_range, user_rating, name, address, latitude, and longitude. There are 911 tuples in table E, which is E.csv. Here are five sample tuples from table E:

1) RJNAeNA-209sctUO0dmwuA,17502451,Capitol,Madison,53703.0,4.0,1236,1,"[u'Alcohol: full_bar', u""Ambience: {'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'divey': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': True}"", u'BikeParking: True', u'BusinessAcceptsCreditCards: True', u""BusinessParking: {'garage': False, 'street': True, 'validated': False, 'lot': False, 'valet': False}"", u'Caters: False', u'GoodForKids: True', u""GoodForMeal: {'dessert': False, 'latenight': False, 'lunch': True, 'dinner': True, 'breakfast': False, 'brunch': False}"", u'HasTV: True', u'NoiseLevel: loud', u'OutdoorSeating: True', u'RestaurantsAttire: casual', u'RestaurantsCounterService: True', u'RestaurantsDelivery: False', u'RestaurantsGoodForGroups: True', u'RestaurantsPriceRange2: 2', u'RestaurantsReservations: False', u'RestaurantsTableService: True', u'RestaurantsTakeOut: True', u'WheelchairAccessible: False', u'WiFi: free']","[u'American (Traditional)', u'Restaurants', u'German', u'American (New)', u'Steakhouses', u'Bars', u'Breakfast & Brunch', u'Salad', u'Nightlife']","[u'Monday 7:30-22:30', u'Tuesday 7:30-0:0', u'Wednesday 7:30-1:0', u'Thursday 7:30-1:0', u'Friday 7:30-1:30', u'Saturday 9:0-1:30', u'Sunday 9:0-22:30']","American, Breakfast, Burger",25,2,"OrderedDict([(u'aggregate_rating', u'4.8'), (u'rating_text', u'Excellent'), (u'rating_color', u'3F7E00'), (u'votes', u'1610')])",The Old Fashioned,23 N Pinckney St,43.0762316238,-89.3836456086
2) 6zZDTZ4ZZEYfN268iPz0uQ,17503679,Capitol,Madison,53703.0,4.5,38,1,"[u'Alcohol: full_bar', u'CoatCheck: True', u'NoiseLevel: loud', u'RestaurantsPriceRange2: 3', u'RestaurantsTableService: True', u'RestaurantsTakeOut: True', u'BusinessAcceptsCreditCards: True', u'GoodForKids: False', u'RestaurantsGoodForGroups: True', u'RestaurantsReservations: True', u""Ambience: {'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'divey': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': False}"", u'RestaurantsDelivery: False', u'BikeParking: True', u""BusinessParking: {'garage': False, 'street': False, 'validated': False, 'lot': False, 'valet': False}"", u'GoodForDancing: False', u""GoodForMeal: {'dessert': False, 'latenight': False, 'lunch': False, 'dinner': False, 'breakfast': False, 'brunch': False}"", u'HappyHour: True', u""Music: {'dj': False, 'background_music': True, 'no_music': False, 'karaoke': False, 'live': False,

'video': False, 'jukebox': False}"", u'OutdoorSeating: False', u'Smoking: no', u'HasTV: False', u'RestaurantsAttire: dressy']","[u'Bars', u'Sushi Bars', u'Restaurants', u'Nightlife', u'Lounges']","[u'Monday 11:30-14:30', u'Monday 16:30-22:0', u'Tuesday 11:30-14:30', u'Tuesday 16:30-22:0', u'Wednesday 11:30-14:30', u'Wednesday 16:30-22:0', u'Thursday 11:30-14:30', u'Thursday 16:30-22:0', u'Friday 16:30-22:0', u'Saturday 16:30-22:0', u'Sunday 16:30-22:0']","Asian, Japanese, Sushi",25,2,"OrderedDict([(u'aggregate_rating', u'4.2'), (u'rating_text', u'Very Good'), (u'rating_color', u'5BA829'), (u'votes', u'209')])",Red,"316 West Washington Ave, Ste 100",43.0738205,-89.384915

3) 0ETtpZSd6f7q_-o-7gutPg,17503369,Capitol,Madison,53703.0,4.5,264,1,"[u'Alcohol: beer_and_wine', u""Ambience: {'romantic': False, 'intimate': False, 'classy': False, 'hipster': True, 'divey': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': False}"", u'BikeParking: True', u'BusinessAcceptsCreditCards: True', u""BusinessParking: {'garage': False, 'street': True, 'validated': False, 'lot': False, 'valet': False}"", u'ByAppointmentOnly: False', u'Caters: False', u'GoodForKids: True', u""GoodForMeal: {'dessert': False, 'latenight': False, 'lunch': False, 'dinner': False, 'breakfast': True, 'brunch': True}"", u'HasTV: False', u'NoiseLevel: average', u'OutdoorSeating: True', u'RestaurantsAttire: casual', u'RestaurantsDelivery: False', u'RestaurantsGoodForGroups: False', u'RestaurantsPriceRange2: 1', u'RestaurantsReservations: False', u'RestaurantsTableService: False', u'RestaurantsTakeOut: True', u'WheelchairAccessible: True', u'WiFi: free']","[u'Coffee & Tea', u'Restaurants', u'Creperies', u'Food']","[u'Monday 6:30-18:30', u'Tuesday 6:30-18:30', u'Wednesday 6:30-18:30', u'Thursday 6:30-18:30', u'Friday 6:30-18:30', u'Saturday 6:30-18:30', u'Sunday 6:30-18:30']","Breakfast, Coffee and Tea",10,1,"OrderedDict([(u'aggregate_rating', u'3.8'), (u'rating_text', u'Good'), (u'rating_color', u'9ACD32'), (u'votes', u'114')])",Bradbury's,127 N Hamilton St.,43.0769705,-89.3838785
4) kz5UANVmPxplIcoLnZBNhg,17503613,McClellan Park,Madison,53718.0,3.5,88,1,"[u'Alcohol: full_bar', u""Ambience: {'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'divey': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': True}"", u""BestNights: {'monday': False, 'tuesday': False, 'friday': True, 'wednesday': False, 'thursday': False, 'sunday': True, 'saturday': True}"", u'BikeParking: True', u'BusinessAcceptsBitcoin: False', u'BusinessAcceptsCreditCards: True', u""BusinessParking: {'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}"", u'Caters: True', u'CoatCheck: False', u'DogsAllowed: False', u'GoodForDancing: False', u'GoodForKids: True', u""GoodForMeal: {'dessert': False, 'latenight': False, 'lunch': True, 'dinner': True, 'breakfast': False, 'brunch': True}"", u'HappyHour: True', u'HasTV: True', u""Music: {'dj': False, 'background_music': False, 'no_music': False, 'karaoke': False, 'live': False, 'video': False, 'jukebox': False}"", u'NoiseLevel: average', u'OutdoorSeating: True', u'RestaurantsAttire: casual', u'RestaurantsDelivery: False',

u'RestaurantsGoodForGroups: True', u'RestaurantsPriceRange2: 2', u'RestaurantsReservations: True', u'RestaurantsTableService: True', u'RestaurantsTakeOut: True', u'Smoking: no', u'WheelchairAccessible: True', u'WiFi: free']","[u'Pubs', u'Nightlife', u'Breweries', u'Bars', u'Gift Shops', u'Gastropubs', u'Food', u'Shopping', u'Flowers & Gifts', u'Restaurants']","[u'Monday 11:0-2:0', u'Tuesday 11:0-2:0', u'Wednesday 11:0-2:0', u'Thursday 11:0-2:0', u'Friday 11:0-2:30', u'Saturday 11:0-2:30', u'Sunday 10:0-2:0']",American,25,2,"OrderedDict([(u'aggregate_rating', u'3.6'), (u'rating_text', u'Good'), (u'rating_color', u'9ACD32'), (u'votes', u'61')])",The Great Dane Pub & Brewing Co.,876 Jupiter Drive,43.08536,-89.28035

5) ubEa6XiMt6gOJ9xsUkbpEw,17503345,Capitol,Madison,53703.0,4.5,201,1,"[u'Alcohol: full_bar', u""Ambience: {'romantic': False, 'intimate': False, 'classy': True, 'hipster': False, 'divey': False, 'touristy': False, 'trendy': False, 'upscale': True, 'casual': False}"", u'BYOB: False', u'BYOBCorkage: no', u'BikeParking: True', u'BusinessAcceptsCreditCards: True', u""BusinessParking: {'garage': False, 'street': True, 'validated': False, 'lot': False, 'valet': False}"", u'Caters: False', u'DogsAllowed: False', u'GoodForKids: False', u""GoodForMeal: {'dessert': False, 'latenight': False, 'lunch': False, 'dinner': True, 'breakfast': False, 'brunch': False}"", u'HasTV: False', u'NoiseLevel: quiet', u'OutdoorSeating: False', u'RestaurantsAttire: dressy', u'RestaurantsCounterService: False', u'RestaurantsDelivery: False', u'RestaurantsGoodForGroups: True', u'RestaurantsPriceRange2: 4', u'RestaurantsReservations: True', u'RestaurantsTableService: True', u'RestaurantsTakeOut: False', u'WheelchairAccessible: True', u'WiFi: no']","[u'Restaurants', u'American (New)']","[u'Monday 17:30-0:0', u'Tuesday 17:30-0:0', u'Wednesday 17:30-0:0', u'Thursday 17:30-0:0', u'Friday 17:30-0:0', u'Saturday 17:0-0:0']","American, European, French",70,4,"OrderedDict([(u'aggregate_rating', u'4.2'), (u'rating_text', u'Very Good'), (u'rating_color', u'5BA829'), (u'votes', u'334')])",L'Etoile Restaurant,1 South Pinckney Street,43.0755377281,-89.3827955168

2. <u>Data Analysis Task: Correlation Discovery with Association Rule Mining</u>

The data analysis task deemed most valuable and interesting was correlation discovery using association rule mining. We were interested in finding which characteristics of restaurants often occurred together. For this task, we thought of each restaurant, which is a tuple in E.csv (the integrated table from the previous project stage), as a transaction. The items in the transaction were characteristics of the restaurant.

After consulting with Professor Doan, we first looked at values of an individual column before combining values of two columns to form rules. For example, we first looked at the “cuisines” column, which is comprised of cuisine types such as “American”, “Sandwiches”, “Chinese”, and “Burger”. Each restaurant is a transaction with its cuisine values as its items. For instance, a restaurant with values “Asian”, “Chinese”, and

“Vegetarian” for “cuisines” would have “Asian”, “Chinese”, and “Vegetarian” as items in its transaction. The individual attributes considered were “Ambiences” (from the “attributes” column), “categories”, “cuisines”, and “GoodForMeal” (from the “attributes” column). Although we had more columns in the dataset, some did not make sense for this task in the project. Ambience describes the atmosphere of the restaurant with values like “romantic”, “casual”, and “hipster”. Categories is similar to cuisines, with values like “Bar”, “Nightlife”, and “Breweries”. Cuisines is cuisine types, as discussed earlier. Finally, GoodForMeal indicates whether the restaurant is good for lunch, dinner, breakfast, etc.

After evaluating columns individually, we combined two columns for association rule mining. Since the “stars” column was the rating of a restaurant (on a scale from 1.0 to 5.0, with half point increments), we were interested in combining it with individual columns examined previously in order to explore which characteristics of a restaurant correlated with higher or lower ratings. For example, to explore which cuisine types correlated to what star ratings, again each restaurant is considered a transaction as before, with items for the transaction consisting of its cuisine types and the star rating. For instance, if a restaurant had the values “American”, “Burger”, and “Pizza” for “cuisine” and value “3.0” for star rating, the items for this transaction/restaurant would be “American”, “Burger”, “Pizza”, and “3.0”. The combined characteristics we considered were ambiences & stars, categories & stars, cuisines & stars, cuisines & average cost for two, ambience & average cost for two, and GoodForMeal & average cost for two.

Next we joined E.csv with yelp_reviews_user_id_business_id.csv which contains the details of business_ids for which different users have written reviews. The resulting join table now has the details of users along with the restaurants for which the users have written reviews and hence helps us in mining association rules in the context of reviewers on Yelp. To mine such rules we can perform a group_by operation based on ‘user_id’ column on the join table. In the resulting table we can view each user id as a transaction and view the various other other columns as potential itemsets.

This helps us in mining rules such as “User who pens reviews for X type of restaurant also pens reviews for Y type of restaurant”. And since users generally write reviews about the restaurants they visit, we can generalize this to identify rules about the kind of restaurants user visits. In other words, this helps us in identifying the general taste/preferences of the user. i.e, we can identify rules such as “User who visits Mediterranean restaurants visits Indian restaurants”

To perform association rule mining, we used the Python package Orange. We created the corresponding .basket files for each set of rules to explore.

3. <u>Accuracy Numbers Obtained</u>

This is not applicable to this stage of our project.

4. <u>Insights from Association Rule Mining</u>

Some interesting rules found are listed below. Note the support for some of these rules may be low. This is due to the fact that some restaurants are missing values for certain attributes or may have only one value, and occurrences of certain values may not occur too often in general in this dataset. For instance, there are only 20 Thai restaurants in the Madison area, while there are many more American restaurants.

Association rules obtained from E.csv

a. Ambiences
   The rules below illustrate that restaurants with intimate ambiences are often romantic, and vice-versa. Similarly, upscale restaurants are often considered classy.
   i. intimate -> romantic with confidence 0.4167, support 0.0093
   ii. romantic -> intimate with confidence 0.3571, support 0.0093
   iii. upscale -> classy with confidence 0.8000, support 0.0075

b. Categories
   The below rules show that Bars are always accompanied with Nightlife, and that restaurants with Bars and Nightlife as categories are often considered traditional American restaurants.
   i. Bars -> Nightlife with confidence 1.0000, support 0.2173
   ii. Bars, Nightlife -> American_(Traditional) with confidence 0.5556, support 0.1207

c. Cuisines
   These rules demonstrate that Japanese restaurants are often correlated with sushi, restaurants often serve both seafood and steak, diners often serve breakfast, restaurants with desserts often serve sandwiches or pizza, many restaurants serving coffee and tea often serve sandwiches, and American restaurants serving bar food often have burgers.
   i. Japanese -> Sushi with confidence 0.6000, support 0.0167
   ii. Seafood -> Steak with confidence 0.5200, support 0.0145
   iii. Diner -> Breakfast with confidence 0.6364, support 0.0156
   iv. Desserts -> Sandwich with confidence 0.4545, support 0.0112
   v. Desserts -> Pizza with confidence 0.0.4091, support 0.0100
   vi. Coffee_and_Tea -> Sandwich with confidence 0.4054, support 0.0167
   vii. Bar_Food, American -> Burger with confidence 0.4211, support 0.0179

d. GoodForMeal
   Places that are good for breakfast and lunch are often good for brunch, and places that are good for dessert are often good for lunch.
   i. breakfast, lunch -> brunch with confidence 0.5517, support 0.0266
   ii. dessert -> lunch with confidence 0.4211, support 0.0133

e. Ambiences & Star Rating (from 1.0 to 5.0 in increments of 0.5)
   Restaurants with hipster, intimate, and romantic ambiences are often rated highly

at 4.0, while restaurants with classy and trendy atmospheres are often rated well at 3.5.

i. hipster -> 4.0 with confidence 0.5238, support 0.0121
ii. intimate -> 4.0 with confidence 0.6667, support 0.0088
iii. classy -> 3.5 with confidence 0.3750, support 0.0099
iv. trendy -> 3.5 with confidence 0.5000, support 0.0165
v. romantic -> 4.0 with confidence 0.7143, support 0.0110

f. Categories & Star Rating (from 1.0 to 5.0 in increments of 0.5)

There were no rules here we deemed of value.

g. Cuisines & Star Rating (from 1.0 to 5.0 in increments of 0.5)

Coffee and tea places, along with breakfast places, are often rated highly. Vietnamese and Chinese restaurants are also rated well.

i. Coffee_and_Tea -> 4.0 with confidence 0.5135, support 0.0209
ii. Vietnamese -> 3.5 with confidence 0.4000, support 0.0022
iii. Vietnamese -> 4.0 with confidence 0.6000, support 0.0033
iv. Chinese -> 3.5 with confidence 0.4510, support 0.0252
v. Breakfast -> 4.0 with confidence 0.4265, support 0.0318

h. Ambiences & Average cost for two

Restaurants with casual, divey, and hipster atmospheres are often less costly with average costs of $10. Meanwhile, restaurants with more intimate vibes are typically more expensive, and places with romantic and/or classy atmospheres are generally more expensive.

i. hipster -> 10 with confidence 0.5714, support 0.0132
ii. intimate -> 25 with confidence 0.5000, support 0.0066
iii. classy -> 40 with confidence 0.4167, support 0.0110
iv. romantic -> 40 with confidence 0.5714, support 0.0088
v. divey -> 10 with confidence 0.667, support 0.0285
vi. casual -> 10 with confidence 0.5640, support 0.2613

i. Cuisines & Average cost for two

Cheaper cuisine types include Mexican, coffee & tea, chinese, sandwiches, fast food, burgers, and diners, with average cost often around $10. Steak is often more expensive, which makes sense.

i. Steak -> 70 with confidence 0.3226 and support 0.0110
ii. Mexican -> 10 with confidence 0.6912, support 0.0516
iii. Coffee_and_Tea -> 10 with confidence 0.9189, support 0.0373
iv. Chinese -> 10 with confidence 0.7059, support 0.0395
v. Sandwich -> 10 with confidence 0.7651, support 0.1251
vi. Fast_Food -> 10 with confidence 0.9636, support 0.1164
vii. Burger -> 10 with confidence 0.7174, support 0.0724

viii. Diner -> 10 with confidence 0.8182, support 0.0198

Association rules obtained on the table obtained after joining E.csv with yelp_reviews_user_id_business_id.csv i.e, Rules mined in the context of reviewers to understand the correlation in the preferences of Restaurant goers.

a. Categories of restaurants that user writes reviews for.
   These rules show that users who pen reviews for American_(New) Restaurants also write reviews for Bars with high confidence and users who write reviews for American_(Traditional) Bars also write reviews for Restaurants Nightlife, so on.
   i. American_(New), Restaurants -> Bars with confidence 0.8126 and support 0.2169
   ii. American_(Traditional), Bars -> Restaurants, Nightlife with confidence 0.8052 and support 0.1564
   iii. Pubs -> Restaurants with confidence 0.9110 and support 0.1635

b. Cuisines of restaurants that user writes reviews for.
   These rules show that users who are interested in American cuisine are generally interested in Bar_Food and Burger
   i. Burger -> American with confidence 0.7865 and support 0.2077
   ii. Bar_Food -> American with confidence 0.7208 and support 0.1736

c. Ambiances of restaurants that users visits
   These rules show that users who visit restaurants with divey ambiance also visit restaurants with casual ambiance and users who visit restaurants with hipster ambiance also try out trendy restaurants
   i. divey -> casual with confidence 0.7960 and support 0.068
   ii. hipster -> trendy with confidence 0.4580 and support 0.0540

Another possible usage scenario for these results would be to infer certain descriptions for a restaurant, given other descriptions that are already known. For example, on Yelp, users can specify the cuisine of a restaurant, allowing other users to see that restaurant in search results for that cuisine. Search results would be better if a search for a particular cuisine also displays restaurants with related cuisines. Since a restaurant with the cuisine “Japanese” is likely to have the cuisine “Sushi”, it makes sense for a search for “Japanese” cuisine to also display restaurants with the cuisine “Sushi”, even if these restaurants are not labeled with the cuisine “Japanese”. However, before adding such inferences to the search functionality, it is important to remove spurious correlations. For example, the rule “Desserts -> Pizza” is likely a spurious correlation. We describe ways to eliminate spurious correlations in the “Future Work” section below.

From our results, we do not believe that combining multiple columns of data (such as cuisines and ratings) and mining association rules is a good way to perform classification. Firstly, we noticed that certain cuisines could correlate with multiple ratings, such as “Vietnamese” above. Secondly, our association rules only use a few columns with a small number of discrete values. A more conventional machine learning algorithm would use multiple columns to create more powerful features.

5. Future Work

To determine whether a given association rule actually represents a causal relationship, we would need to obtain another dataset and perform a statistical Z-test. For example, we consider the rule “Japanese -> Sushi” with confidence 0.600. The null hypothesis states that at least 60% of restaurants with the cuisine “Japanese” also have the cuisine “Sushi”. Using the new dataset, we determine the ratio of restaurants with the cuisine “Japanese” to restaurants with both cuisines “Japanese” and “Sushi”. If the test has less than a predetermined p-value, then we reject the null hypothesis and conclude that less than 60% of restaurants with the cuisine “Japanese” also have the cuisine “Sushi”.

As discussed earlier, it is not very effective to use association rules to predict the rating of a restaurant. In future work, we would use a multi-class machine learning algorithm, such as a neural networks, to predict the rating of a restaurant. Such an algorithm would use multiple columns of the dataset to create features.