1. How did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).

In project 4, we trained a Random Forest Model. Apply this model to the table after blocking we got a set of predicted value. We treat the two restaurant be a match if the predicted value is 1 and mark two entries to be the same restaurant. After we find all matches in tuple pairs, we merge tables A and B by taking the union of them. During merging, we pick the entry in Yelp.csv if two entries appear to be a match.

2. Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.

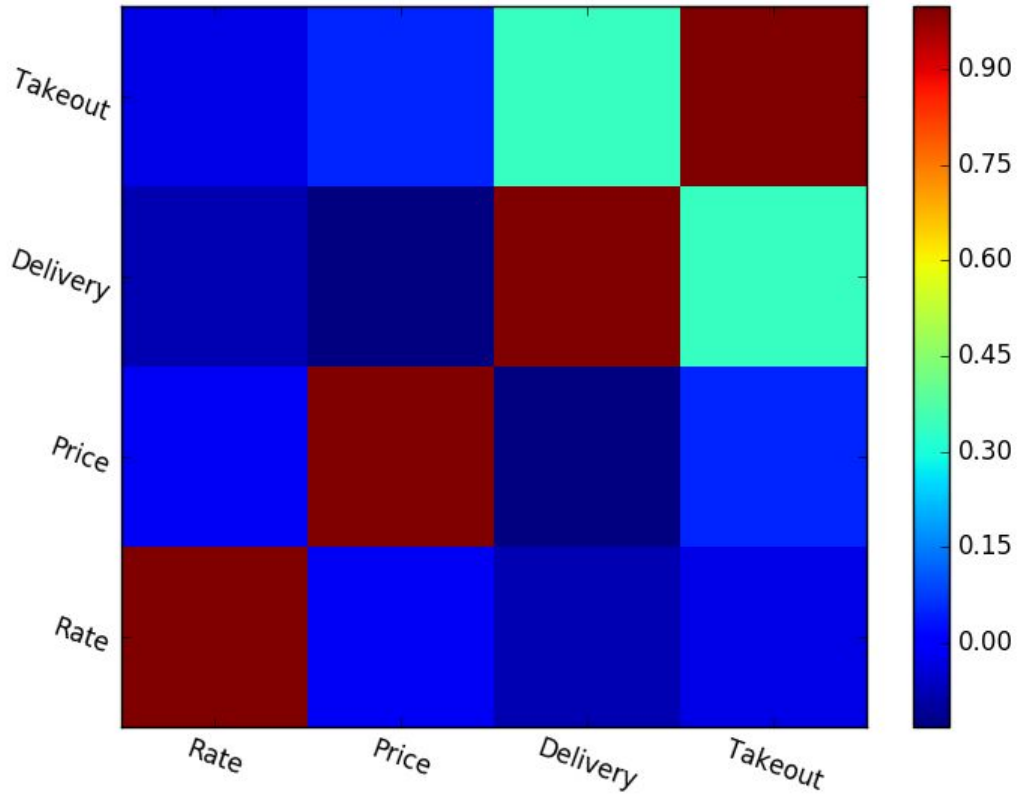
Schema: ID, Name, Phone, Rate, Price, Zipcode, State, City, Address, Delivery, Takeout

We have 5926 tuples in Table E
Sample tuples:
<60, Au Cheval, (312) 929-4580, 4.5, $$, 60607, IL, Chicago, 800 W Randolph S, No, No>
<605, Table, (773) 486-8525, 4, $$, 60647, IL, Chicago, 2728 W Armitage Ave, No, Yes>
<2681, The Social Study, (415) 292-7417, 4, $$, 94115, CA, San Francisco, 1795 Geary Blvd, No, No>
<511, Bongo Room, (773) 728-7900, 3.8, $$, 60640, IL, Chicago, 5022 N. Clark Street, No, Yes>

3. What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.

We want to explore the correlation between these four attributes: Rate, Price, Delivery and Takeout. Correlation is a broad class of statistical relationships involving dependence, though in common usage it most often refers to the extent to which two variables have a linear relationship with each other. First, we need to extract these four columns out, computing the correlation coefficient between each pair. The Pearson correlation can be calculated with numpy's corrcoef function.Then we use matplotlib to show the result in colorful grids.

4. Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).



5. What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?

We found that there is no strong correlation between these four attributes, especially for price and rate. Delivery and Takeout have some weak correlation, which intuitively makes sense. We think the major problem is that rate bases on a lot of other attributes that are hard to grade, like tastes, environment, services, etc. The given 3 aspects cannot fully predict the rate.

6. If you have more time, what would you propose you can do next?
If we have more time, we plan to grab more data and construct a larger schema so that we could do analysis on the major influence factor on restaurant rating. This might need more work on intelligently analyze all the comments, which is a precious source on a more complete view of each restaurant.