Stage V Report

Chang Guo (cguo42@wisc.edu)
Yuncong Hao(hyuncong@wisc.edu)
Qun Zou(qzou5@wisc.edu)

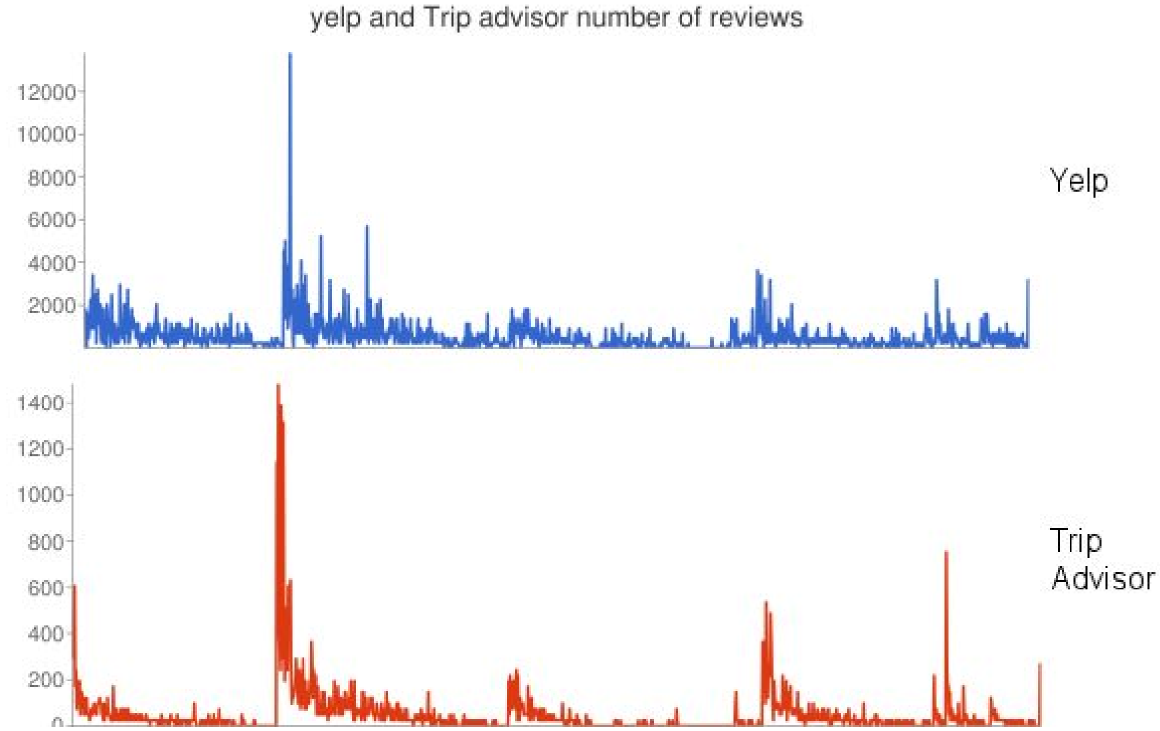
- **Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.**

There are 1360 tuples in Table E.

ID	Shop Name	Postal Code	Phone Number	Street Address	Yelp Price	Yelp Number of Reviews	Trip-Advisor Star	Trip-Advisor Price	Trip-Advisor Number Of Reviews
1	sugarfish by sushi nozawa	90017	2136273000	600 w 7th st	$$$	1776	4.5	$$-$$$	280
2	rock sugar pan asian kitchen	90067	3105529988	10250 santa monica blvd,	$$	1725,	4.5,	$$ - $$$,	336
3	katsuya hollywood	90028	3233912757	6300 hollywood blvd	$$$	1535	4.5,	$$$$	400
4	pagoda bar at yama-shiro hollywood	90068	3234665125	1999 n sycamore ave	$$$	14	4	$$$$	618

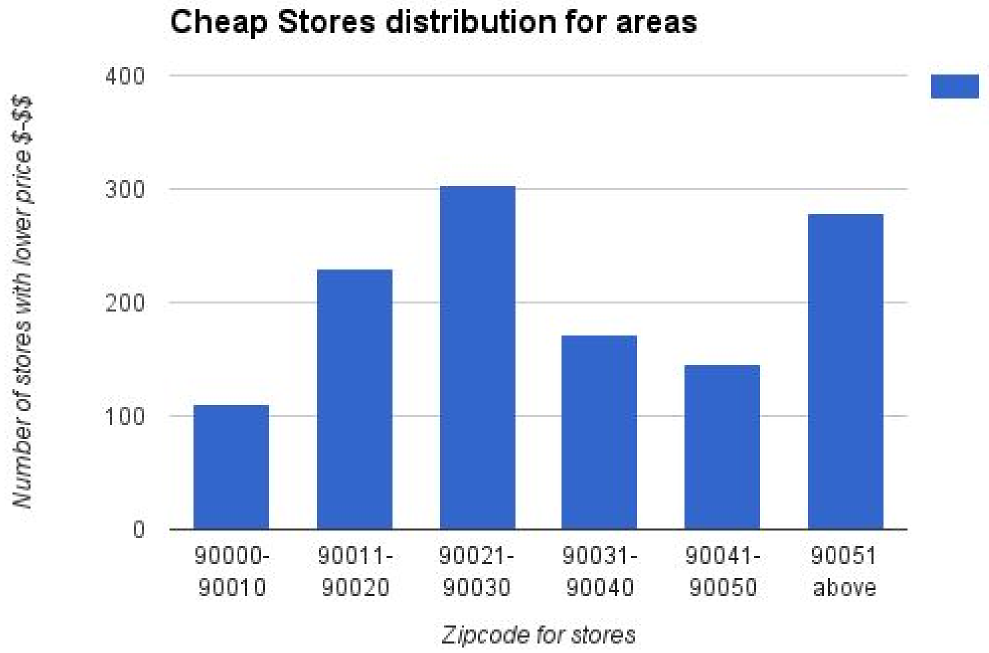
- **What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.**

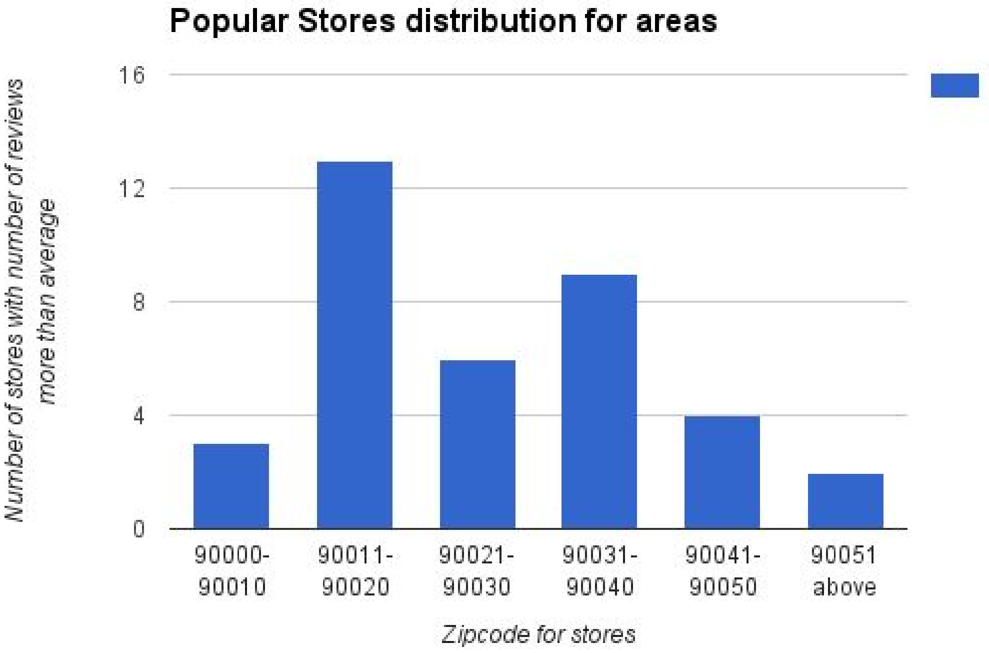
1. We want to check if there is any correlation relationship on the number of reviews and stars we got from the two different websites. We calculated the correlation value between the two websites on number of reviews = 0.6147.



The figure above shows the number of reviews we get from two websites, and the x-axis value is the ID number we assigned to all of the stores. From the picture we can see that the number of reviews from the two websites are identical which met with our expectation. And we also found the anomaly store with peak value of number of reviews from two websites, “philippe the original”, which can be selected as the top choice for restaurant recommendation in Los Angeles.

2. We want to find which area in Los Angeles has cheapest but also most popular restaurants. We used a OLAP-Style exploration. First we sort our matched pairs based on postal code and number of reviews. Then to find the cheapest restaurants, we filtered out all the restaurants with price more than two “$” sign. Then we counted the qualified restaurant number in the near area and make histogram figure on that.
To find the most popular restaurants in Los Angeles, we calculated the normalized value of number of reviews. We filtered out all the restaurants with reviews less than the average value, and counted the number of left restaurants in nearby area and make another histogram as below.





From the two figures above we can find that to find both cheap but also popular restaurants, the most efficient area we will recommend is with postal code between 90011 - 90020.

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

We didn't do classification analysis for our matching data and don't have accuracy numbers for shown.

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

1. We found that comparing to TripAdvisor, Yelp has more (around 2~3 times) number of reviews and ratings. So if you are looking for a better coverage on people's feedback of the restaurants in Los Angeles, it's better to check for Yelp.
2. The area in Los Angeles with the zip code from 90010 to 90040 has the most both cheap but still popular restaurants. We would recommend this area to search for good food for students who value the food quality as well as the price.
3. The popularity and rating for the same restaurants on both websites are identical, so we can highly trust the information on the two websites.

Problems:

Since Yelp has more user feedback (rating, reviews) comparing to TripAdvisor, there's ~100 matching tuples in our table E has missing value for either ratings, number of reviews or price from TripAdvisor website,, which made the analysis of comparisons between two websites difficult.

- **If you have more time, what would you propose you can do next?**

We would want zoom in more on the data of number of reviews and ratings, to see detect the restaurants with very different values on the two websites (ex. On one website it has a huge amount of reviews while on the other it only has a small number of reviews). By doing this we would want to analysis if there's any group of users would prefer one website over the other one, for example, if we detected most of the pairs with large differences in rating/reviews happened to be all Japanese restaurants, we may can conclude the website with more reviews have a larger user group on Japanese restaurants lovers. And also, they may want to release more Japanese food/culture related recommendations for their target users.