Stage 5 Report

1. Statistics of Table E:
    - Schema of Table E

ID	Integer
Name	Text
Address	Text
City	Text
Zipcode	Number/Text
Latitude	Float value
Longitude	Float value
Review_count	Integer
Rating	Range 0-5
Zomato_id	Id to track lineage
Yelp_id	Id to track lineage
Wifi	Boolean 0/1
Researvations	Boolean 0/1
Parking	Boolean 0/1
Wheelchair Accessible	Boolean 0/1
Outdoor Seating	Boolean 0/1
Is_expensive	Boolean 0/1

    - Number of Tuples in E - 718
    - Examples

ID	480	1204	1327	1464	1937
name	The Cracked Egg	Sauce Pizza & Wine	Firebirds Wood Fired Grill	Bonfyre American Grille	Presti's Bakery & Caf
address	1000 N Green Valley Parkway #480$*$ Henderson$*$ NV 89074	2470 W Happy Valley Rd$*$ Phoenix$*$ AZ 85085	6801 Northlake Mall Drive$*$ Charlotte$*$ NC 28216	2601 West Beltline Highway$*$ Madison$*$ WI 53713	12101 Mayfield Rd$*$ Cleveland$*$ OH 44106
city	Henderson	Phoenix	Charlotte	Madison	Cleveland
zipcode	89074	85085	28216	53713	44106
latitude	36.02807582	33.71409205	35.351243	43.03476069	41.5088518
longitude	-115.0851223	-112.1124008	-80.85076	-89.42187057	-81.598275
review_count	714	374	533	754	779
rating	3.811764706	3.787433155	4.158724203	4.033819629	4.279332478
zomato_id	16981241	17030169	17147235	17503464	16962390
yelp_id	At2bqa8emnEr5WNIosi0ow	8J55FMsOXei4Xh1jHSpElw	qVVjbYR0LIfJuIIzgPMTuw	2YlUn3s132hNq5ueGeliJg	orrrhqRRUORIzUSxWTveKg
wifi	0	0	1	1	0
reservations	0	0	1	1	0
parking	1	1	1	1	1
wheelchairaccessible	1	1	1	1	1
outdoorseating	0	1	1	1	1

Is_expensive	0	0	1	1	0

2. Data Analysis Task
   A. We used columns rating, has_restaurant_reservations, parking, wheelchair accessible, outdoor seating, to predict the "is_expensive" of the restaurant. The is_expensive is a binary valued attribute with 0 representing "less than 30$" and 1 representing "more than 30$". We trained 5 models – Random Forests, Linear Regression, Logistic Regression, Decision Tree and SVM. Using these models, we are trying to predict if the given restaurant is expensive.

   B. We performed the roll-up OLAP operation using MS Excel on our "E.csv" file. The objective of this operation was to group the restaurants by city, calculate the average rating for each city and thus find out the top 5 cities with highest ratings. The results were following :

City	Count (# of Restaurants)	Avg Rating Scale : (1-5)	Avg Price Range Scale : (1-4)
Toronto	35	4.42	2.45
Pittsburgh	80	4.35	1.975
Cleveland	71	4.27	2.098
Charlotte	75	4.22	2.96
Las Vegas	95	4.12	2.715

NOTE : These are the top 5 cities out of the ten overlapping cities in our datasets as the operations have been performed on the merged table obtained after Stage #4 of our project. The ten overlapping cities were : Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland, Edinburgh, Montreal and Waterloo.

3. Numbers
   These are the results of 5 classifiers for cross-validation (for 2A) :

Classifier	Precision	Recall	F1
Random Forests	0.85	0.75	0.75
Linear Regression	0.89583	0.875	0.86818
Logisitic Regression	0.74063	0.62727	0.60616
Decision Trees	0.5	0.5	0.5
SVM	0.85	0.75	0.75

Based on the result of classifier, Linear Regression offered best Precision, Recall and F1.
We ran all the models on Test Set just to ascertain results of cross-validation.
As expected Linear Regression again offered best results.

Classifier	Precision	Recall	F1
Random Forests	0.65625	0.625	0.63095
Linear Regression	0.89583	0.875	0.86818
Logisitic Regression	0.73088	0.62712	0.60935
Decision Trees	0.62395	0.60169	0.6029
SVM	0.70635	0.65254	0.64816

From the Stage4 output, we had 718 records for data analysis. We used 500 records for training the models, 100 for tuning models and 118 to test the accuracy.

Linear Regression provided us precision of 89.5 % and recall of 87.5 %.

4. Conclusion:
   We trained ML model to predict if the restaurant is expensive (cost for one people exceeds the threshold). The threshold was defined by one of the two data sources as 30$. Currently, the model is predicting the Boolean valued attribute is_expensive with good precision and recall. Thus, we can infer that we can predict if the restaurant is expensive by using facilities offered by the restaurant.
   Challenges we faced:
   1. We found it difficult to increase precision as number of records were low. Also, most of the restaurants in our data had good ratings (3.8+ out of 5.0). We need to get more data with varying ratings (low, medium and high) to increase precision further.
   Also, our dataset offered restaurant of selected cities. We need more restaurants of a particular state or country to perform more analysis. We can detect preferred cuisine of the region.

5. Future Work
   We believe following extensions are feasible from our current progress:
   1. Reviews provided by customers can be used to figure out best dish served in the restaurant.
   2. Get more data from source to perform demographic analysis. Currently, data from limited cities is available. But, datasources (Zomato and Yelp) provide APIs to procure more data
   3. Some columns like cuisine types can be used to get better insights. But, it requires more work as cuisine types are enormous in the data.
   4. The Yelp also offers multi valued ranges as – (below 10$, 10-30$, 30-50$ and 60$+). We need to verify if current features work for predicting multi-valued price ranges.