CS 838 (Spring 2017): Data Science Project Report - Stage 5 (Group 12)

Deepanker Aggarwal
deepanker@cs.wisc.edu

Saket Saurabh
ssaurabh@cs.wisc.edu

Vishnu Lokhande
lokhande@cs.wisc.edu

1 Objective

The objective of this stage was to perform data analysis to derive some insights on the integrated and cleaned Table E obtained at the end of Stage 4. We performed OLAP analysis and classification on our Table E and derived some interesting results.

2 Statistics of Table E

- The schema of table is as follows (name, address, zipcode, cuisine, price, violation_code, critical_flag, grade, median_household_income, median_real_estate_value, population_density, cost_of_living, population, neighborhood, borough).

- ### 2.1 Attribute definitions

  —- Fill in the details here of what each attribute means —-

- The number of tuples in the Table E are $5,560$.

- Some sample tuples from table E are as follows:

  1. (juniper,"237 w 35th st,",10001,bars american,$$,06C,Critical,A, 81671,650200,35350,157.4,21966,Chelsea and Clinton,Manhattan)
  2. (mission chinese food,"171 e broadway, ,",10002,desserts bars chinese,$$,10F,Not Critical,A,33218,535600,93461,168.9,82191, Lower East Side, Manhattan)
  3. (okinii,"216 thompson st, ,",10012,japanese bars,$$,10B,Not Critical,A, 86594,1000001,80873,164.7,26145, Greenwich Village and Soho,Manhattan)

4. (the fitz,"fitzpatrick manhattan, 687 lexington ave,",10022,bars american irish,$$,10B,Not Critical,A,109019,866100,67873,158.2,29618, Gramercy Park and Murray Hill,Manhattan)
5. (ny sweet spot cafe,"2376 coney island ave,",11223,middle eastern european,$$,06C,Critical,A,41328,613000,35968,167.4,74606,Southern Brooklyn,Brooklyn)

- Apart from the attribute 'cuisine', no other attribute in the table has missing values.

3 Data Analysis Tasks

We undertook the following categories of data analysis tasks on the Table E.

- OLAP Analysis
- Classification

We describe the process and the outcome of these tasks in the following sections.

4 Data Analysis I: OLAP

4.1 Questions that we wanted to answer:

Using OLAP analysis, we wanted to answer the following questions for our NYC Restaurant dataset as represented by Table E:

1. How many restaurants are located in each borough and neighborhood of New York City?
2. How many NYC restaurants are there for each kind of price type as reported at Yelp?
3. How many NYC restaurants are there for each kind of Yelp rating?
4. How many NYC restaurants are there for each grade of health violation?
5. How many NYC restaurants have received 'critical' health violation status?
6. Which borough and neighborhoods of New York City have restaurants with highest number of 'critical', grade 'A' health violations?
7. Which Yelp ratings correspond to the highest number of 'critical', grade 'A' health violations?
8. How many 'critical', grade 'A' health violations exist for even expensive restaurants in New York City?

4.2 OLAP Data Analysis Process

We had to do a couple of interesting things before we could perform OLAP analysis on our dataset as represented by Table E. We loaded the dataset in a Python package called 'cubes', which provides a ROLAP server backed by SQLite relational database. We introduced a concept hierarchy for area of restaurant in New York City- borough > neighborhood > zipcode. We defined a set of dimensions across which we could perform aggregation, roll-up, drill-down, slicing and dicing. The following subsections define our OLAP methodology in detail.

4.2.1 Introducing Concept Hierarchy for Area

The table E produced at the end of Stage 4 just had the zipcode information for each restaurant. Therefore, in this stage, we added two more new attributes: borough and neighborhood, and created a concept hierarchy for area. This borough and neighborhood data for each zipcode was obtained by web scraping this information from this url: `https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm` and afterwards, we joined this scraped data with the zipcode available in Table E.

The resulting concept hierarchy can be represented as: borough > neighborhood > zipcode. This implies that each zipcode belongs to a larger neighborhood, which in turn is a part of larger borough. In fact, New York City has 5 boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island) and 42 neighborhoods and 180 zipcodes.

4.2.2 Dimensions and Fact in our OLAP analysis

Based on the data in the Table E, we created these dimensions for our OLAP cube: area, price, rating, and critical_flag. The dimension 'area' represents a concept hierarchy as explained above. The restaurant data represents the fact table with the number of restaurants as one of the measures of the fact table.

4.2.3 Summary of Schema used in our OLAP analysis

1 summarizes the logical Star Schema of our OLAP model, as interpreted by the Python 'cubes' package.

4.2.4 A Note on using the Python 'cubes' package for OLAP

We used the Python 'cubes' package to perform our OLAP analysis. The 'cubes' package provides a default ROLAP server backed by a SQLite database that was sufficient for our given dataset size. The 'cubes' package takes as input a JSON file that describes the OLAP logical model. (We have attached this JSON file in the Appendix section of this report.) The logical model described by the JSON file is independent of the underlying physical model. In our case,

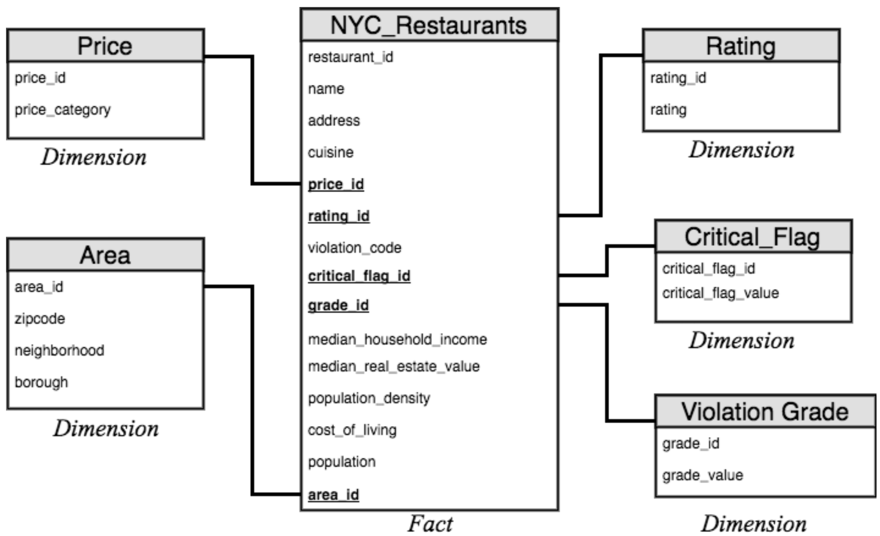

Figure 1: Logical Star Schema for the OLAP Model

the underlying physical model is defined by a single relational table in which we load the Table E. The 'cubes' package defines a nice set of abstractions on top of the logical model that offers standard operations like drill-down, slicing, dicing, etc. For example, a browser represents a view of the cube, a point cut represents a slice along a dimension and one can invoke the aggregate() method on the browser object to perform various kinds of drill-down.

Borough	Total Number of Restaurants for the given borough	Number of Restaurants with 'critical' violation	Number of Restaurants with 'critical' & Grade A violation
Bronx	519	281	230
Brooklyn	1846	977	821
Manhattan	1821	1007	839
Queens	1078	561	500
Staten Island	296	165	140
	Aggregation	*Slicing*	*Dicing*

Figure 2: Aggregation, Slicing and Dicing by borough and health violation type for NYC restaurants

4.3 Outcomes of OLAP Analysis

In this section, we report the results of various OLAP analysis that we performed.

Yelp Restaurant Rating	Total Number of Restaurants for the given rating	Number of Restaurants with 'critical' violation	Number of Restaurants with 'critical' & Grade A violation
1.0	43	21	19
1.5	85	46	40
2.0	168	82	73
2.5	379	211	175
3.0	780	414	333
3.5	1424	793	666
4.0	1825	979	836
4.5	726	378	329
5.0	130	67	59
	Aggregation	*Slicing*	*Dicing*

Figure 3: Aggregation, Slicing and Dicing by Yelp Rating and health violation type for NYC restaurants

Yelp Restaurant Price Category	Total Number of Restaurants for the given price category	Number of Restaurants with 'critical' violation	Number of Restaurants with 'critical' & Grade A violation
$	2280	1215	1047
$$	2930	1597	1331
$$$	300	155	132
$$$$	50	24	20
	Aggregation	*Slicing*	*Dicing*

Figure 4: Aggregation, Slicing and Dicing by Yelp Price Category and health violation type for NYC restaurants

- Roll-up on the number of restaurants in each NYC Borough:
  Figure 2 shows the roll-up on the number of restaurants in each New York Borough. Both, Brooklyn and Manhattan have the highest number of restaurants in New York.

- Drill-down on number of restaurants in each NYC Neighborhood:
  We performed drill-down on the number of restaurants in each neighborhood by going one level down in the concept hierarchy from borough to neighborhood. Due to paucity of space, we do not report the whole statistics for all the neighborhood- however, it is available in the OLAP Juypter notebook at our GitHub repository. From our analysis, we concluded that Northwest Brooklyn neighborhood has the highest number(442) of restaurants.

Health Violation Grade	Total Number of Restaurants for the given grade
A	4881
B	422
C	46
P	10
Z	201

Aggregation

Figure 5: Aggregation by Health Violation Grade for NYC restaurants

- Roll-up on number of restaurants with each type of rating:
  Figure 4 shows the roll-up on the number of restaurants for each type of rating. About 75% of restaurants in New York have a Yelp rating of more than 3.5.

- Roll-up on number of restaurants with each price category:
  Figure 4 shows the roll-up on the number of restaurants for each price category. More than 85% of restaurants in New York are cheap or moderately priced. Only 50 restaurants are ultra-expensive.

- Roll-up on number of restaurants with critical health violation type:
  By performing a roll-up on the 'critical' vs 'non-critical' flag, we found that 2991 restaurants had 'critical' health violation, while the remaining 2569 restaurants were marked 'non-critical'.

- Roll-up on number of restaurants with each health violation grade:
  There were 5 grades of health violation that were assigned to each restaurant with 'A' being the assigned for gravest case of violation and 'Z' being assigned for the most mild case of violation. A couple of restaurants were assigned 'P' grade, which meant that their evaluation was still pending. Figure 5 shows the number of restaurants for each type of health violation grade. Clearly, more than 86% of New York Restaurants have Grade 'A' health violation.

- Slice and dice on health violation by area:
  Figure 2 shows the results of slicing and dicing for various borough types by 'critical' flag and grade 'A' health violation. Almost every borough of New York City had more than 45% of restaurants that had grade 'A' health violation and were flagged 'critical'.

- Slice and dice on health violation by restaurant rating:
  Figure 3 shows the results of slicing and dicing for various restaurant

rating types by 'critical' flag and grade 'A' health violation. Most restaurants have 4.0 rating, however, those restaurants also saw one of the most number of health code violation.

- Slice and dice on health violation by restaurant price category:
  Figure 4 shows the results of slicing and dicing for various restaurant price category types by 'critical' flag and grade 'A' health violation. From the data, we can conclude that as the restaurant becomes more pricey, they have less number of health code violations.

5 Data Analysis II: Classification

The attributes from the table E which have been used to solve the classification questions are (zipcode, price, violation_code, critical_flag, grade, median_household_income, median_real_estate_value, population_density, cost_of_living, population, neighborhood, borough).

5.1 Questions that we wanted to answer:

1. Can we predict the price of the restaurant from the other attributes used for classification?
2. Can we predict the critical flag of the restaurant based on the other attributes used for classification?
3. Can we predict which borough the restaurant belongs to based on the other attributes used for classification?
4. Can we predict borough from other attributes?

5.2 Classification Process

5.3 Preprocessing

We analyzed the dataset for missing values and found none. As some attributes had string values, we transformed them to integer values and also analyzed how data is distributed in Figure 6. We find that price attribute is skewed for $ and $$, grade is also skewed. Before classifying the data, we also standardize it, such that mean of the data is 0 and variance is 1.

5.4 Prediction

For each question, we try below classifiers:-

- Logistic Regression
- Linear Discriminant Analysis

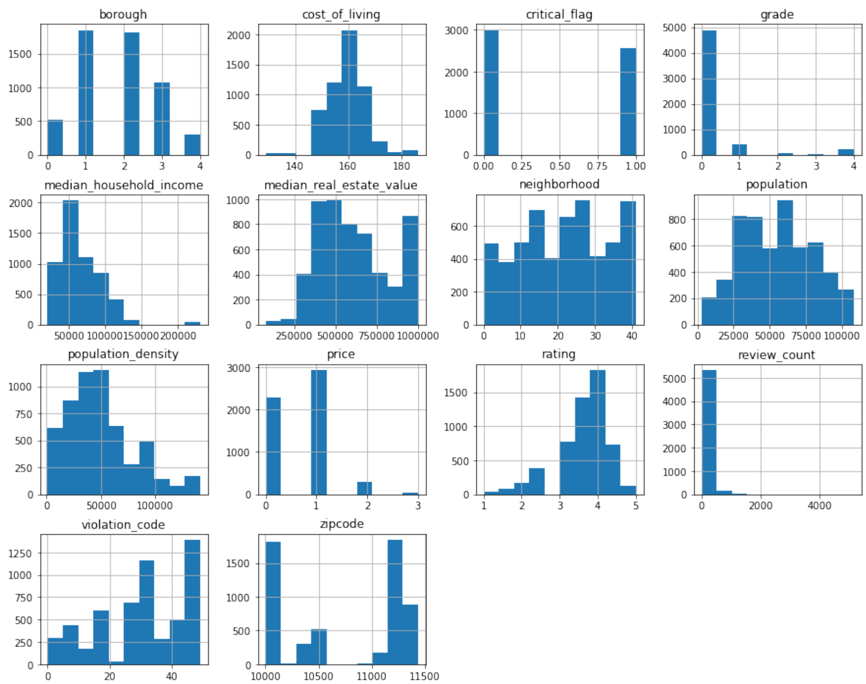

Figure 6: Dataset distribution of various attributes

- K-Nearest Neighbors
- Decision Tree (CART)
- Naïve Bayes
- SVM

We divide the dataset into two sets called Set I(train) of size 80% and Set J(test) of size 20%. We do 5-fold Cross Validation on Set I, then choose the best classifier on the basis of accuracy as all labels are equally important in the questions we are trying to answer. We use grid search to tune this classifier. Then we report Accuracy, Precision, Recall and F1-Scores for this tuned classifier on Set-J.

6 Classification Results

We report the results for each question below.

6.1 Question 1

Can we Predict the price of the restaurant from the other attributes used for classification?

The cross-validation accuracy measures for the five algorithms considered can be found in Table 1. The accuracies can also be visualized in the form of a box plot in Figure 7. It is observed that Logistic Regression performs the best among the other techniques. We then tune the parameters for Logistic Regression and managed to increase the train accuracy to 0.661(±0.010). We obtained 62.16% on test set. Precision, recall and F-1 scores for each label using tuned Logistic Regression($C = 10$) on test set are in Table2. Our classifier doesn't predict any price as expensive or highly expensive. We think this happens because data is highly skewed in favor of cheap and moderately priced restaurants.

Matcher	Accuracy(Mean)
Logistic Regression	65.46±2.4
Linear Discriminant Analysis	60.43±1.5
K-Nearest Neighbors	55.66±2.7
Decision Tree Classifier	57.73±1.2
Naïve Bayes	39.29±6.7
SVM	62.13±±2.2

Table 1: Question 1- Classifier to predict the price of the restaurant from the other attributes used for classification

Predicted Price	Precision(Mean)	Recall(Mean)	F1(Mean)	Support
Cheap($)	0.58	0.67	0.0.62	1813
Moderate($$)	0.65	0.66	0.66	2361
Expensive($$$)	0	0	0	233
Ultra Expensive($$$$)	0	0	0	41

Table 2: Question 1 - Precision, Recall and F1 score for Logistic Regression

We report the results for each question below.

6.2 Question 2

Can we predict the critical flag of the restaurant based on the other attributes used for classification?

The cross-validation accuracy measures for the five algorithms considered can be found in Table 3. The accuracies can also be visualized in the form of a box plot in Figure 8. It is observed that Linear Discriminant Analysis performs the best among the other techniques. We then tune the parameters for the same and managed to increase the train accuracy to 0.347(±0.015). We obtained 31.87%

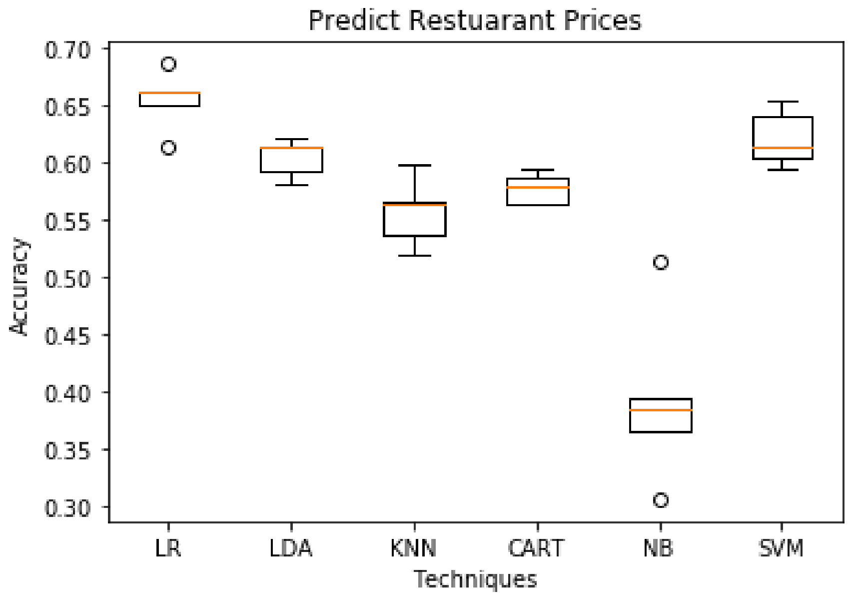

Figure 7: Question 1 - Box plot of accuracies of different classifiers

on test set. Precision, recall and F-1 scores for each label using tuned Linear Discriminant Analysis($'solver' :' eigen'$) on test set are in Table4.

Matcher	Accuracy(Mean)
Logistic Regression	34.08±2.2
Linear Discriminant Analysis	33.36±1.9
K-Nearest Neighbors	26.16±2.6
Decision Tree Classifier	23.55±3.1
Naïve Bayes	22.39±1.6
SVM	32.82±±2.1

Table 3: Question 2- Classifier to predict the critical flag of the restaurant based on the other attributes used for classification

7 Learnings and Conclusion

—- Fill in the details here —-

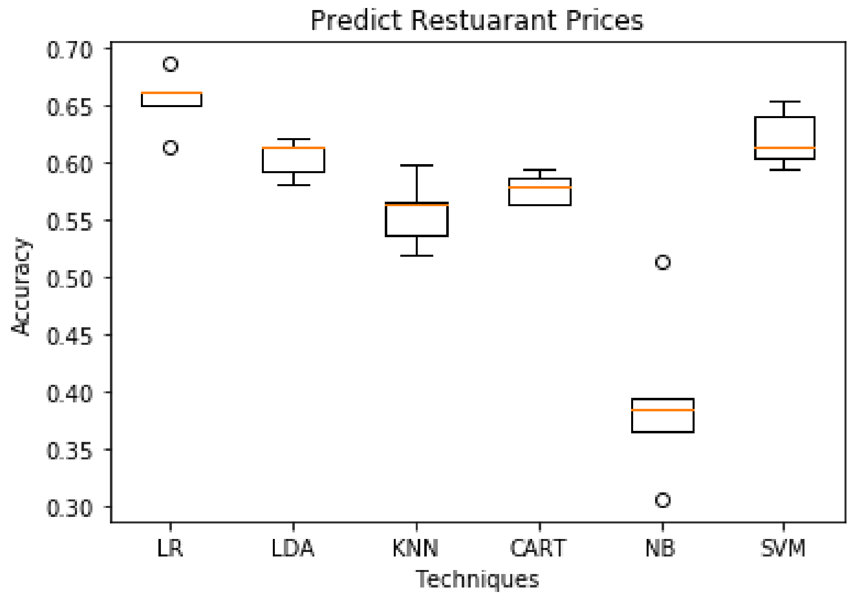

Figure 8: Question 2 - Box plot of accuracies of different classifiers

Predicted Price	Precision(Mean)	Recall(Mean)	F1(Mean)	Support
Cheap($)	0.58	0.67	0.0.62	1813
Moderate($$)	0.65	0.66	0.66	2361
Expensive($$$)	0	0	0	233
Ultra Expensive($$$$)	0	0	0	41

Table 4: Question 2 - Precision, Recall and F1 score for Logistic Regression

8 Future proposals

—- Fill in the details here —-

9 Appendix

```
###################################
#--- olap_model.json ---#
###################################
# This is the logical model for the OLAP cube that was fed into
# the Python 'cubes' package for analysis.

{
```

```
"dimensions": [
   {
     "name": "area",
     "levels": [
          {
              "name": "zipcode",
              "label": "ZipCode"
          },
          {
              "name": "neighborhood",
              "label": "Neighborhood"
          },
          {
              "name": "borough",
              "label": "Borough"
          }
       ],
     "hierarchies": [
          {
              "name": "area_hierarchy",
              "levels": ["borough", "neighborhood", "zipcode"]
          }
     ]
   },
   {
     "name": "price",
     "levels": [
          {
              "name": "price",
              "label": "Price"
          }
     ]
   },
   {
     "name": "rating",
     "levels": [
          {
              "name": "rating",
              "label": "Rating"
          }
     ]
   },
   {
     "name": "critical_flag",
     "levels": [
          {
              "name": "critical_flag",
              "label": "Critical Health Violation Flag"
          }
     ]
```

```
    },
    {
      "name": "grade",
      "levels": [
            {
                "name": "grade",
                "label": "Health Violation Grade"
            }
      ]
    }
],
"cubes": [
    {
        "name": "nyc_restaurants",
        "dimensions": ["area", "price", "rating", "critical_flag",
             "grade"],
        "measures": [
                {"name":"median_household_income", "label":"Household
                     Income"},
                {"name":"median_real_estate_value", "label":"Real
                     Estate Value"},
                {"name":"population_density", "label":"Population
                     Density"},
                {"name":"cost_of_living", "label":"Cost of Living"},
                {"name":"population", "label":"Population"}
              ],
        "aggregates": [
                {
                    "name": "household_income_avg",
                    "function": "avg",
                    "measure": "median_household_income"
                },
                {
                    "name": "real_estate_value_avg",
                    "function": "avg",
                    "measure": "median_real_estate_value"
                },
                {
                    "name": "cost_of_living_avg",
                    "function": "avg",
                    "measure": "cost_of_living"
                },
                {
                    "name": "population_avg",
                    "function": "avg",
                    "measure": "population"
                },
                {
                    "name": "num_restaurants",
                    "function": "count"
```

```
                }
            ],
          "mappings": {
                        "area.zipcode": "zipcode",
                        "area.neighborhood": "neighborhood",
                        "area.borough": "borough"
                       }
        }
    ]
}
```