Identifying Fast Food Restaurant In The Same Location by Zip Code

Kelvin Tiongson
kelvinjtiongson@lewisu.edu
DATA-51000-001, Summer 2020
Data Mining and Analytics
Lewis University

I. Introduction

The goal of this analysis was to find an association between restaurants and their respective location. The analysis takes a look at restaurants in a specific zip code. The data that was used was downloaded as a CSV from an open data community known as "data.world" [1]. We will be able to discover which restaurants that are frequently or infrequently placed close to each other.

The future sections of this report describe the dataset, methodology, results along with a discussion, and a conclusion. Section II contains a description of the dataset. Section II also contains a frequency distribution of the restaurants, descriptive statistics of the frequency of restaurants, and a box plot of the frequency of restaurants in the dataset. The methodology for analysis is presented in section III. For the association rule mining method, I chose support, confidence, and lift thresholds that best matched the problem. There were multiple rules that were created from the association, so I chose the top ten rules by their measures of support, confidence, and top five rules by highest and lowest lift. In section IV, the reports and results of the analysis are discussed. Finally, section V provides conclusions about the association rules that were found for certain restaurants and their location.

II. Data Description

Table I describes the dataset used in the analysis. The dataset contained eleven columns. The goal of the analysis was to look at fast food restaurants in certain areas by their postal code. From Table I, the attributes used in the analysis were the name of the restaurant and the postal code of the restaurant's location. None of the attributes used were normalized when looking for association rules between restaurants. Some restaurants had to be renamed because the name of the restaurant had different spelling. For example, there were different variations of the spelling of "McDonald's" such as "Mcdonalds", "McDonalds", and "Mcdonald's". This was the case for several of the big chain restaurants like KFC, Popeye's, and Wendy's just to name a few.

TABLE I. Fast Food Restaurants

Attribute	Type	Example Value	Description
ADDRESS	Nominal (string)	"324 Main St."	Address
CITY	Nominal (string)	"Oklahoma City"	City
COUNTRY	Nominal (string)	"US"	Country
KEYS	Numeric (real)	"us/ny/massena/324mainst/-1161002137"	Datafini business record identifier
LATITUDE	Numeric (real)	44.9213	Latitude
LONGITUDE	Numeric (real)	-74.89021	Longitude
NAME	Nominal (string)	"Wendy's"	Name of restaurant
POSTAL CODE	Numeric (real)	13662	Zip code
PROVINCE	Nominal (string)	"IL"	State
WEBSITE	Nominal (string)	https://www.wendys.com	URL to the specific restaurant or to the company of the restaurant

Renaming the restaurants to have the same name rather than different spellings of their names allowed me to create better frequencies of the restaurants. From Figure I, there were a few restaurants that separated themselves from most of the restaurants. To get a better look at the separation, I displayed the top fifteen restaurants in Table II. McDonald's had the most restaurants in the

dataset at 1365. The top five restaurants in the dataset were McDonald's, Burger King, Taco Bell, Wendy's, and Arby's. Table II shows descriptive statistics of the restaurants. The standard deviation from Table II shows that the values are far from the mean. This is evident when comparing the minimum, maximum, and quartile values of the frequencies of the restaurants. Figure II supplements the differences between the quartiles with mean, median, Q1, and Q3 at the lower end of the plot. To the right of the box plot, we have the higher frequencies of restaurants with McDonald's being the restaurant with the farthest point.

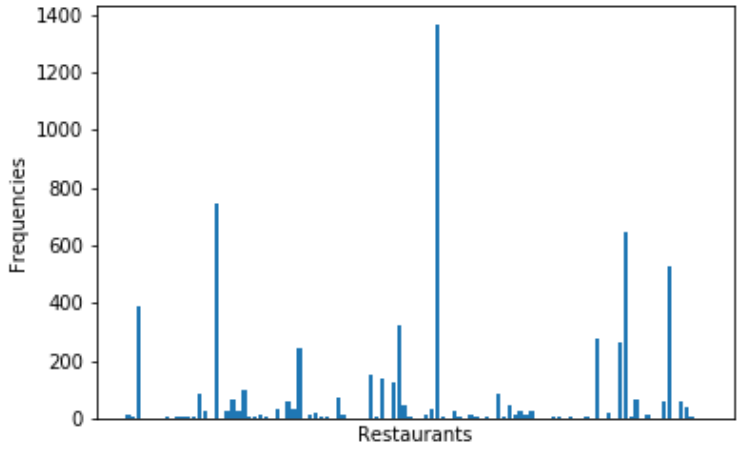

Fig. 1. Frequency distribution of restaurants

TABLE II. FREQUENCY OF THE TOP FIFTEEN RESTAURANTS

Name	Frequency
McDonald's	1365
Burger King	748
Taco Bell	645
Wendy's	529
Arby's	392
KFC	322
Sonic	275
Subway	266
Domino's Pizza	243
Hardee's	154
Jack in the Box	137
Jimmy John's	123
Chick-fil-A	102
Bojangles' Famous Chicken 'n Biscuits	86
Pizza Hut	86

TABLE III. DESCRIPTIVE STATISTICS OF THE FREQUENCIES OF RESTAURANTS

	Restaurants
Count	6627
Mean	63.11
Standard Deviation	177.77
Minimum	2
25%	3
50%	7
75%	31
Maximum	1365

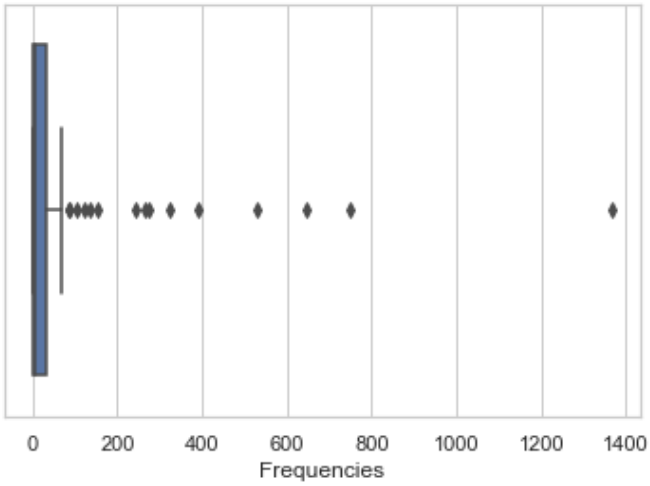

Fig. 2. Box plot of frequency distribution of restaurants

III. METHODOLOGY

In order to successfully identify association rules from the dataset, I reshaped and created a new data frame of the dataset where the data represented a transactional dataset. Each row represented the zip codes and the columns were the restaurants of the dataset. If a restaurant existed in the zip code, its value was the count of the restaurant in the given area.

When deciding a minimal support threshold, I originally wanted to find the rules that appeared in half of the locations. This would give me a starting point of restaurants that were very common to find together. So, I initially set to find rules with a minimum support threshold of fifty percent. When deciding on a minimal confidence threshold, I wanted to see rules that did not have restriction based on the confidence threshold. So, I decided to choose the lowest minimum confidence threshold at one percent. This provided me with no rules. As a result, I continuously moved the minimum support threshold down. Finally, a threshold of one percent helped identify four hundred and thirty-eight rules. The top ten rules by support are shown in Table IV. The top ten rules by confidence are shown in Table V. From the associated rules, the lift shows how an item in the consequent is likely or unlikely to be in the same basket based on the item in the antecedent. A lift value greater than 1 means that an item in the consequent column is likely to be in the basket if an item in the antecedent column is in the basket, while a value less than 1 means that an item in the consequent column is unlikely to be in the basket if an item in the antecedent column is in the basket [2]. Table VI shows the associated rules of the top five items with the highest lift. Table VII shows the associated rules of the top five items with the lowest lift.

TABLE IV. TOP TEN RULES BY SUPPORT

Support	Confidence	Lift	Antecedent	Consequent
0.121	0.290	1.116	McDonald's =< 1.5	Taco Bell =< 1.5
0.121	0.465	1.116	Taco Bell =< 1.5	McDonald's =< 1.5
0.102	0.245	0.875	McDonald's =< 1.5	Burger King =< 1.5
0.102	0.365	0.875	Burger King =< 1.5	McDonald's =< 1.5
0.099	0.238	1.104	McDonald's =< 1.5	Wendy's =< 1.5
0.099	0.461	1.104	Wendy's =< 1.5	McDonald's =< 1.5
0.084	0.322	1.492	Taco Bell =< 1.5	Wendy's =< 1.5
0.084	0.388	1.492	Wendy's =< 1.5	Taco Bell =< 1.5
0.075	0.180	1.030	McDonald's =< 1.5	Arby's =< 1.5
0.075	0.430	1.030	Arby's =< 1.5	McDonald's =< 1.5

TABLE V. TOP TEN RULES BY CONFIDENCE

Confidence	Support	Lift	Antecedent	Consequent
0.857	0.011	3.295	Arby's =< 1.5, KFC =< 1.5, Wendy's =< 1.5	Taco Bell =< 1.5
0.786	0.010	3.020	Arby's =< 1.5, McDonald's = 1.5-2.5, Wendy's =< 1.5	Taco Bell =< 1.5
0.722	0.012	2.776	KFC =< 1.5, McDonald's =< 1.5, Wendy's =< 1.5	Taco Bell =< 1.5
0.710	0.010	3.290	Arby's =< 1.5, McDonald's = 1.5-2.5, Taco Bell =< 1.5	Wendy's =< 1.5
0.706	0.011	3.272	Arby's =< 1.5, KFC =< 1.5, Taco Bell =< 1.5	Wendy's =< 1.5
0.688	0.010	3.928	McDonald's=< 1.5, Taco Bell =< 1.5, Wendy's =< 1.5	Arby's =< 1.5
0.660	0.015	2.535	Arby's =< 1.5, McDonald's = 1.5-2.5	Taco Bell =< 1.5
0.654	0.016	2.513	Arby's =< 1.5, KFC =< 1.5	Taco Bell =< 1.5
0.643	0.013	2.471	Arby's =< 1.5, Burger King =< 1.5, Wendy's =< 1.5	Taco Bell =< 1.5
0.639	0.037	2.457	Arby's =< 1.5, Wendy's =< 1.5	Taco Bell =< 1.5

TABLE VI. TOP FIVE RULES BY HIGHEST LIFT

Lift	Support	Confidence	Antecedent	Consequent
5.690	0.010	0.328	McDonald's = 1.5-2.5, Taco Bell =< 1.5	Arby's =< 1.5, Wendy's =< 1.5
5.690	0.010	0.180	Arby's =< 1.5, Wendy's =< 1.5	McDonald's = 1.5-2.5, Taco Bell =< 1.5
5.591	0.010	0.124	Taco Bell =< 1.5, Wendy's =< 1.5	Arby's =< 1.5, McDonald's = 1.5-2.5
5.591	0.010	0.468	Arby's =< 1.5, McDonald's = 1.5-2.5	Taco Bell =< 1.5, Wendy's =< 1.5
5.568	0.010	0.379	McDonald's = 1.5-2.5, Wendy's =< 1.5	Arby's =< 1.5, Taco Bell =< 1.5

TABLE VII. TOP FIVE RULES BY LOWEST LIFT

Lift	Support	Confidence	Antecedent	Consequent
0.579	0.013	0.061	Wendy's =< 1.5	Subway =< 1.5
0.579	0.013	0.125	Subway =< 1.5	Wendy's =< 1.5
0.669	0.018	0.174	Subway =< 1.5	Taco Bell =< 1.5
0.669	0.018	0.071	Taco Bell =< 1.5	Subway =< 1.5
0.703	0.013	0.293	Chick-fil-A =< 1.5	McDonald's =< 1.5

IV. RESULTS AND DISCUSSION

From Table IV, most of the rules have a below average confidence. Their lift scores imply that there is barely an association between those rules. From Table V, most of the rules lead to a consequent of a Taco Bell. The lift scores of Table V are all above one by more than one point. So, their lift scores imply that it is likely that we will find the restaurants in the consequent column if there are specific restaurants in their respective antecedent column. Table VI shows us the likely consequent restaurants of the antecedent restaurants. The confidence in this table is lower than average and are right at the minimal support threshold. So, it may not be likely that the consequent restaurants are in the same location as the antecedent restaurants. Table VII shows us the unlikely consequent restaurants of their antecedent restaurants. The confidence in this table is lower than Table VI. So, it is highly unlikely that the restaurants in the consequent column are in the same location as the restaurants in the antecedent column.

From Table IV, McDonald's appears in eight out of the top ten rules. From the top ten rules of support, we can say that if there is at least one McDonald's in an area, then there should be a Taco Bell, Burger King, Wendy's, and Arby's. However, the lift scores of each rule are not too far from one. So, there is barely an association between a location having at least one McDonald's and at least one of Taco Bell, Burger King, Wendy's or Arby's. Taco Bell and Wendy's have the highest lift in this table. So, if there is a Taco Bell, we can associate a Wendy's being in the same place. From Table V, most of these rules have Arby's, Taco Bell, Wendy's, and McDonald's. Most of the consequent columns in Table V associate a Taco Bell in the area. From Table VI, the restaurants are ether Arby's, McDonald's, Wendy's, or Taco Bell. Since these columns have a high lift score, better confidence score than Table VII, and passes the minimal support threshold, we can associate that a combination of a pair of the mentioned restaurants will lead to an association of the other remaining pair of the mentioned restaurants. From Table VII, these rules have the lowest lift, are close to the minimum support threshold, and have a very low confidence score. We can expect to not find a Wendy's with a Subway, Subway with a Taco Bell, and a Chick-fil-A with a McDonald's in the same location.

V. CONCLUSIONS

I graphed the frequency distribution of the restaurants and displayed the top fifteen restaurants. I computed the descriptive frequencies of the restaurants and supplemented the frequencies with a box and whisker plot. From the figures and table, there were several restaurants that stood out within the data. McDonald's dominates the amount of residence of fast food restaurants. McDonald's and the other top fifteen restaurants from Table II are among the largest fast food restaurant chains in the world [3]. In order to obtain the association rules of the restaurants with their respective location by zip code, I had to reshape the data to create a transactional dataset. The rows of the new data frame were the zip code locations. Its columns were the count of restaurants in their location. With a minimal support threshold and minimal confidence threshold of one percent, I was able to identify four hundred and thirty-eight rules. I displayed the top ten rules by support, top ten rules by confidence, top five items with the highest lift, and the top five items with the lowest lift. From these rules, we can expect to find most places with a McDonald's. However, when trying to associate two restaurants within an area, we can associate a Taco Bell with a Wendy's. We can also associate a combination of Arby's, Taco Bell, McDonald's, Wendy's together. The fast food restaurants that we can expect to be infrequently placed together are Subway with a Wendy's or Taco Bell and a Chick-fil-A with a McDonald's.

REFERENCES

[1] https://data.world/datafiniti/fast-food-restaurants-across-america/workspace/file?filename=FastFoodRestaurants.csv
[2] https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html
[3] https://www.qsrmagazine.com/content/ranking-top-50-fast-food-chains-america