Analysis of Restaurant Business in Toronto

CS7DS3 - MAIN ASSIGNMENT

Sankalp Gupta

April 16, 2018

Abstract

Yelp contains the user, review and other business data of various restaurants in a city. In this assignment we focus on the analysis of restaurant business in the city of Toronto. The variables are responsible for predicting restaurant ratings are analyzed and the ratings of different neighborhoods are compared using Markov chain Monte Carlo methods, Gibbs Sampling and Hierarchical linear models. The association between neighborhoods and different restaurant categories are also defined and the neighbourhoods that are more likely to contain certain restaurant categories are identified.

1 Introduction

Yelp is an open platform for students and researchers to perform statistical data analysis on various datasets. The following dataset discusses on the restaurant business in the city of Toronto. Analyzing the dataset help restaurants optimizing their operations and strategize their business decisions with the help of insights obtained through the research. Different neighborhoods have different types of restaurants and hence attract specific users. The users generally rate the restaurants based on their own scale which differs from person to person. It is difficult to have a clear consensus on the true rating of a restaurant. The number of reviews should be sufficient enough to have a better picture of the actual rating of a restaurant. It is also important to know the most important factors influencing the ratings of a restaurant.

1.1 Data

Considering only the open restaurants in the city of Toronto, we use business_open_Toronto.json file along with the review.json file. The business file has 4028 observations of 14 variables. The review file was merged with the business file on the basis of common key business id. The variables of the business and the review file are as mentioned in the below table.

Table 1: business_open_Toronto.json Fields

Restaurant Attributes		
business_id	name	neighborhood
address	city	state
postal code	latitude	longitude
stars	review_count	is_open
categories	PriceRange	

Table 2: review.json Fields

Review Attributes		
business_id	user_id	review_id
stars	date	text
useful	funny	cool

The business dataset contains the restaurant details, the name, the location, the neighbourhood, the quality of the restaurant, the type of restaurant and the number of reviews it has received whereas the review dataset contains the details of each and every review given by the users for the restaurant. The below table gives a basic idea of how many restaurants every neighborhood has and the number of reviews received. It can be clearly seen that though the dataset is huge enough, it is not much balanced. Neighborhoods like MeadowvaleVillage, Markland Wood, Cooksville has very less restaurants in their neighborhood and also very less reviews. Hence, it is important to look at the distribution of the data and pre-process the data before performing analysis.

Table 3: Distribution of Neighborhood data

Sr No	Neighborhood	No of Restaurants	Total Number of Reviews
1	Alexandra Park	41	3101
2	Bayview Village	8	385
3	Beaconsfield Village	28	2050
4	Bickford Park	30	1588
5	Bloor-West Village	32	1138
6	Bloordale Village	30	1020
7	Brockton Village	28	830
8	Cabbagetown	31	1768
9	Casa Loma	5	120
10	Chinatown	87	7487
11	Christie Pits	17	1151
12	Church-Wellesley Village	87	5813
13	City Place	13	679
14	Cooksville	1	5
15	Corktown	71	4234
16	Corso Italia	18	449
17	Deer Park	12	439
18	Discovery District	24	601
19	Distillery District	11	2081
20	Dovercourt	19	852
21	Downsview	17	329
22	Downtown Core	491	32063
23	Dufferin Grove	38	1266
24	East York	28	579
25	Entertainment District	158	15687
26	Etobicoke	213	5152
27	Financial District	102	5850
28	Greektown	69	3646
29	Harbourfront	36	2435
30	High Park	29	737
31	Kensington Market	87	7159
32	Koreatown	60	4040
33	Leslieville	104	5330
34	Liberty Village	38	2554
35	Little Italy	69	4774
36	Little Portugal	30	1724
37	Markland Wood	2	11
38	Meadowvale Village	1	6
39	Milliken	97	3867
40	Mount Pleasant and Davisvill	93	3121
41	New Toronto	17	322
42	Niagara	59	3221
43	Ossington Strip	34	3629
44	Palmerston	11	521
45	Parkdale	63	3299

1.2 Pre-Processing Data

Figure 1 shows the stars distribution of the restaurants. We can see that the distribution is skewing towards left and there are more number of positive reviews (greater than 3.5 stars). One can also see that the common tendency of not giving the extreme ratings of 1 and 5 as compared to others, can be observed in the bar graph.

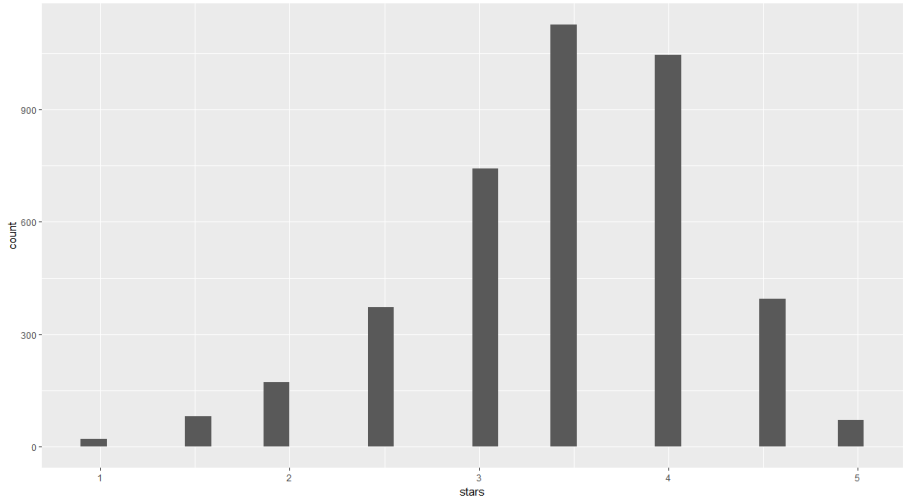


Figure 1: Rating Distribution

Its important to check the number of reviews for each restaurant and their associated rating because when very few people rate a restaurant it leads to biased decisions. On analyzing the data, it shows that 83.33% (i.e 60 of 72) of the restaurants having 5 star ratings and around 50% (224 out of 468) restaurants having ratings greater than 4.5 have total number of reviews less than 15. Due to this, the average rating of these restaurant increases and also that of a neighborhood thereby leading to false conclusions. For example, Cooksville neighborhood has only restaurant with only 5 number of reviews. Due to this the average rating of Cooksville neighborhood is very high i.e. 4. Although this may not be true given the less number of reviews and also less number of restaurants in that neighborhood. Overall, there are 1580 restaurants among the 4208 having less than equal to 15 total number of reviews.

Table 4: Some of the Restaurants with Review Count less than 15

	name	neighborhood	review_count	stars
15	Uncle Mikey's	Brockton Village	11	5.0
52	Noble Coffee	The Junction	6	5.0
91	Shalom Ethopian Restaurant	Cabbagetown	7	5.0
161	Wraps On The Go	Yorkville	4	5.0
171	McDonald's	Financial District	6	5.0
226	Royale's Luncheonette	Little Portugal	5	5.0
265	Village Pizza	Trinity Bellwoods	5	5.0
307	Cock-A-Doodle-Doo	Christie Pits	12	5.0
374	Starbucks	Discovery District	3	5.0
389	Shawarma Q	Yonge and St. Clair	3	5.0
448	Bebo's Authentic Grill	Etobicoke	4	5.0
559	Brasileirissimo Steak House	Little Portugal	3	5.0
607	Sweet Cocoa	Ossington Strip	7	5.0
778	Galaxy T&T	Bloordale Village	3	5.0
897	Patricia's Cake Creations	Etobicoke	10	5.0

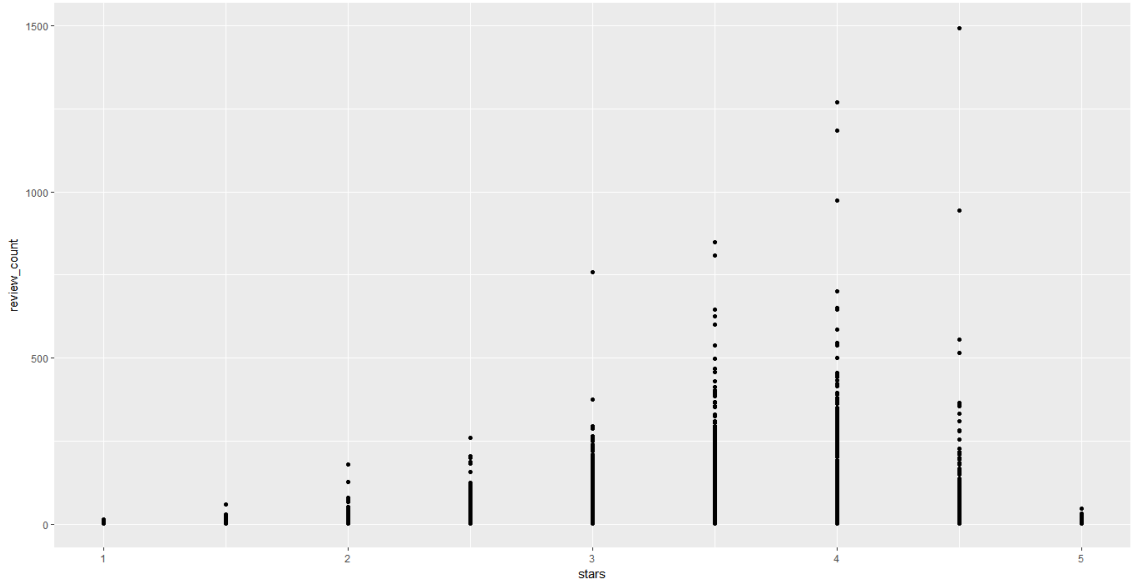


Figure 2: Review Count vs Rating

Therefore it is important to clean the data and remove all the restaurants with less than 15 reviews though the neighborhood might have less number of restaurants. Now we can analyze the ratings of different neighborhoods and identify the neighborhoods which are superior as compared to others. No Missing values are observed in the neighborhood and the stars variables as well.

COMPARISON OF RATINGS of DIFFERENT NEIGHBORHOODS

Neighborhood variable is converted into factors from character datatype so that it can be statistically analyzed. The distribution of the ratings for every neighborhood is shown through a boxplot in Figure 3. The neighborhoods are sorted with respect to the median ratings of the neighborhood.

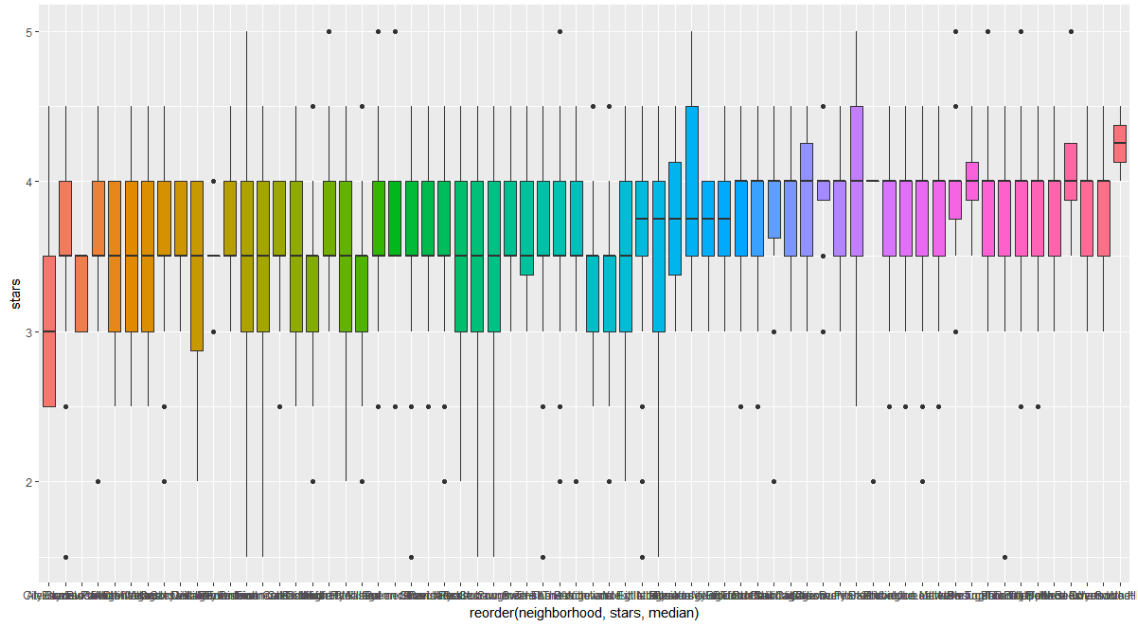


Figure 3: Boxplot of Ratings of Neighborhoods

We cannot compare and come to a conclusion of superior neighborhoods through a box plot. Gibbs sampling will be used to model the difference between the mean ratings of neighborhoods.

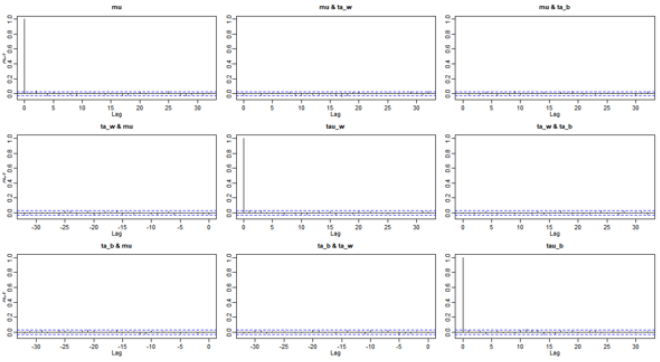
A function compare_m_gibbs having the following features

1. mu, the overall mean across neighborhoods
2. taob, the precision (inverse variability) between neighborhoods
3. taow, the precision (inverse variability) within neighborhoods
4. thetam, the mean rating awarded to neighborhood m.

Table 5: Hyperparameters Used

Hyperparameters							Output		
a0	b0	eta0	t0	mu0	gamma0	maximeter	mu	tau_w	tau_b
1.5	1	0.5	5	3.5	1.5	5000	3.651473	1221.9836	4.956068
0.5	66	0.5	66	3.5	0.04	5000	3.653045	18.4937	0.488908
3	3.5	0.5	3.5	3.5	0.04	5000	3.652364	349.6548	6.658669
2.5	3	0.5	3	3.5	1	5000	3.651385	407.4945	7.45624

Initial mu0 is kept at 3.5 as the distribution of rating has a central tendency near to 3.5 and we want our sampling to be centered around 3.5. The hyper-parameters were varied and the best results having mean near to 3.5 are with a0 = 2.5, b0 = 3, eta0 = 0.5, t0 = 3, mu0 = 3.5, gamma0 = 1. The results are shown in the above Table 5. However, the hyper-parameters have a very low effect on the final result as they are very weakly informed. It is majorly due to the skewness of the distribution. Even on changing the iterations, there is no major change in the trace output and on the final result. The sampling output appears to be performing very well and the same is diagnosed using the raftery.diag method. This diagnostic method helps in estimating how much burn in and thinning is required. The raftery.diag is a run length control diagnostic based on a criterion of accuracy of estimation of the quantile q. The results are good which are specified in Figure 4. An error message is shown indicating the minimum length of pilot run, if the number of iterations in data is taken very low.



```
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

       Burn-in  Total Lower bound  Dependence
       (M)      (N)   (Nmin)       factor (I)
mu     1        3712  3746         0.991
tau_w  2        3741  3746         0.999
tau_b  2        3837  3746         1.020
```

Figure 4: Raftery.diag Diagnostic

Gibbs model returns two objects param and theta. Theta stores the group mean values of all the neighborhoods whereas the param stores the above mentioned parameters. Plotting the theta object against the relative sample size can help us distinguish the superior neighborhoods.

Table 6: Mean Ratings of Neighborhoods

Sr No	Neighborhood	No of Restaurants	Total Number of Reviews	Avg Rating
50	South Hill	2	238	4.25
61	Wallace Emerson	4	156	4.125
20	Downsview	5	244	4
38	New Toronto	4	221	4
49	Seaton Village	12	495	4
24	**Entertainment District**	122	**15425**	3.401639
64	Willowdale	62	5355	3.395161
21	Downtown Core	290	30589	3.374138
66	**Yonge and Eglinton**	49	**3892**	3.367347
33	Liberty Village	31	2495	3.33871
16	Deer Park	8	414	3.3125
36	**Milliken**	46	**3484**	3.304348
2	Bayview Village	5	356	3.3
28	Harbourfront	19	2312	3.263158
13	City Place	11	669	3.090909

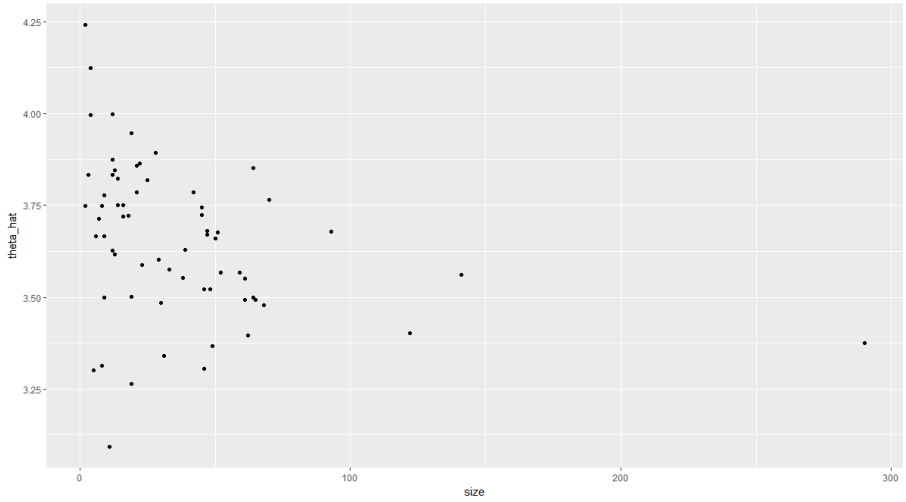


Figure 5: Mean Ratings vs Size

From Table 6, it is evident **South Hill** is the best performing neighborhood and superior to others with **mean rating of 4.25**. Wallace Emerson, Downsview and New Toronto are the next set of neighborhoods with a very good rating of 4. It is also evident from the Figure 5 that even though the number of restaurants are less, the neighborhoods Bayview Village, HarbourFront and the City Place are the worst performing restaurants as compared to SouthHill, Wallace Emerson or any other neighborhoods. The plot and the table clearly depicts the comparison of ratings between the neighborhoods and the superior neighborhoods are identified.

2 Variables influential in predicting Restaurant Ratings

2.1 Data Handling

The task is to identify the variables which are influencing the target variable the restaurant rating. The business.json file is merged with the review dataset to consider the reviews of each and every user for a particular restaurant. Even the business attributes like whether the restaurant has delivery options, accepts credit cards or not, noise level, has TV or not, attire, waiter service, reservations etc are considered and merged with the dataset to get a better understanding of the variables responsible for influencing the restaurant rating.

Although, the data was filtered considering only the restaurants category in the city of Toronto, there were lot of missing values present. The distribution of the same is found using the VIM package which can be seen in the below table.

Table 7: Missing Values Distribution

Missings in variables:	
Variable	Count
attributes.BusinessAcceptsCreditCards	22137
attributes.HasTV	8182
attributes.NoiseLevel	10884
attributes.RestaurantsAttire	9961
attributes.RestaurantsReservations	7076
attributes.RestaurantsTableService	13354
attributes.RestaurantsDelivery	8128
attributes.CoatCheck	179527
attributes.Smoking	179848

It can be seen that the attributes Coat-check and Smoking have more than 80% of their data missing (179527 out of 216006) which is more than the threshold and hence are not considered and removed from the dataset. The other attributes with missing data were imputed with the median and considered for further analysis. Also, the features which are not relevant in influencing the restaurant ratings such as business id, review id, user id, business name, city, state, postal code, location details are removed.

Also, the features with different levels were converted into factors. For using MCMC regressor model, it is necessary to convert the factors into numeric variables i.e dummy variables. Dummy variables for factors were created through One-hot encoding method using dummies package.

CORRELATION MATRIX

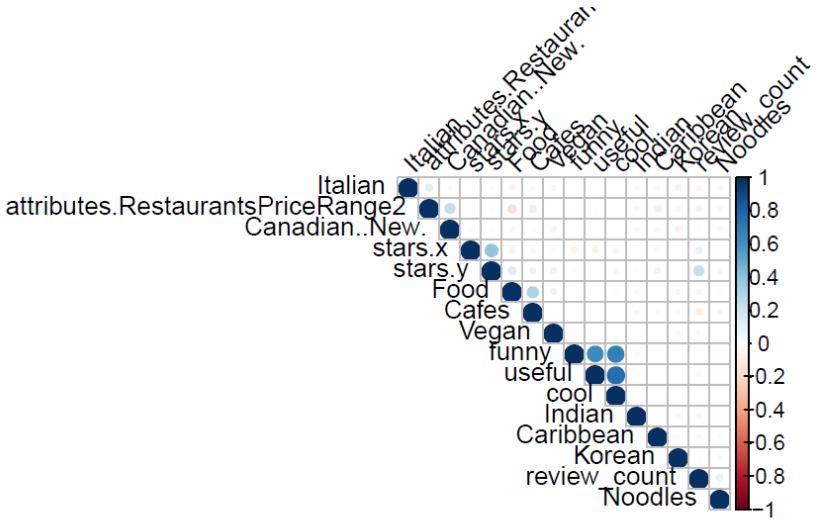


Figure 6: Correlation Matrix

For feature selection, it is necessary to remove highly correlated features to avoid col-linearity to find key features that affect overall restaurant rating given by star.y. The features with negligible effect on ratings can also be removed. In this way, a clear relationship between ratings and

the features is established. From the matrix shown in Figure 6, stars.x i.e. the rating of each review according to the review text is one of the important drivers of overall restaurant rating. Review_count is the next feature which exhibit strong correlations with stars.y. The restaurant categories show mild correlation with the stars. It can be seen that the Useful, funny and cool are highly correlated with each other and one can be avoided to avoid col-linearity.Feature selection is used to reduce the size of dataset with marginal decrease in performance. The reduction in size helps in reducing computational time and should be used if necessary.

2.2 Model Building

SIMPLE LINEAR REGRESSION MODEL

Initially, a simple linear regression model is developed considering all the variables using the lm() function. It is a very useful method for predicting a quantitative response. The stars.y is the target variable and the rest are the dependent variables. The necessary variables are converted into factors before fitting the model. A part of the summary of the output is shown below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.74597 -0.24267  0.01693  0.27248  2.42664
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.26E+00  9.53E-03  341.992  < 2e-16 ***
stars.x                          1.26E-01  7.69E-04  164.194  < 2e-16 ***
useful                           4.69E-03  6.98E-04    6.727  1.74E-11 ***
funny                           -7.77E-03  9.71E-04   -8.002  1.23E-15 ***
cool                             2.45E-03  1.09E-03    2.257  0.024032 *
neighborhoodBickford Park        3.46E-02  1.38E-02    2.513  1.20E-02 *
neighborhoodDownsview            2.65E-01  2.49E-02   10.649  < 2e-16 ***
neighborhoodNew Toronto          2.99E-01  2.53E-02   11.825  < 2e-16 ***
neighborhoodSouth Hill           4.39E-01  2.99E-02   14.681  < 2e-16 ***
neighborhoodUpper Beach          1.83E-01  1.64E-02   11.167  < 2e-16 ***
neighborhoodWallace Emerson      3.57E-01  3.40E-02   10.511  < 2e-16 ***
neighborhoodWest Don Lands       5.85E-01  5.94E-02    9.847  < 2e-16 ***
neighborhoodWest Queen West      4.55E-02  1.53E-02    2.964  3.03E-03 **
review_count                     7.32E-04  5.47E-06  133.968  < 2e-16 ***
attributes.RestaurantsPriceRange23 1.45E-01 4.85E-03  29.879  < 2e-16 ***
Food                             9.45E-02  3.90E-03   24.236  < 2e-16 ***
Bars                             9.15E-02  2.19E-02    4.181  2.90E-05 ***
Sandwiches                       1.06E-01  4.13E-03   25.643  < 2e-16 ***
Chinese                         -1.31E-01  4.71E-03  -27.816  < 2e-16 ***
Canadian..New.                   6.48E-02  3.88E-03   16.72   < 2e-16 ***
Cafes                            1.31E-01  5.23E-03   25.052  < 2e-16 ***
Coffee...Tea                     1.96E-02  5.62E-03    3.485  0.000492 ***
Pizza                           -2.71E-02  4.80E-03   -5.639  1.71E-08 ***
Fast.Food                       -1.24E-01  5.57E-03  -22.269  < 2e-16 ***
Italian                          1.95E-02  4.15E-03    4.704  2.55E-06 ***
Burgers                         -1.29E-01  4.58E-03  -28.151  < 2e-16 ***
Korean                           6.51E-02  5.24E-03   12.411  < 2e-16 ***
Indian                           5.56E-02  6.69E-03    8.314  < 2e-16 ***
Salad                            1.75E-01  7.05E-03   24.774  < 2e-16 ***
Vegan                            1.44E-01  7.18E-03   20.061  < 2e-16 ***
Cocktail.Bars                    1.22E-01  6.59E-03   18.497  < 2e-16 ***
Caribbean                        1.77E-01  8.31E-03   21.301  < 2e-16 ***
Noodles                          2.47E-01  7.61E-03   32.524  < 2e-16 ***
Latin.American                   1.40E-01  9.16E-03   15.311  < 2e-16 ***
Food.Delivery.Services           1.36E-01  1.08E-02   12.585  < 2e-16 ***
Fish...Chips                     2.23E-01  1.26E-02   17.692  < 2e-16 ***
Venues...Event.Spaces            8.93E-02  1.40E-02    6.382  1.76E-10 ***
attributes.BusinessAcceptsCreditCards2 -2.80E-02 3.86E-03 -7.248 4.25E-13 ***
attributes.HasTV2               -7.85E-02  2.31E-03  -33.99   < 2e-16 ***
attributes.NoiseLevel3           7.54E-02  3.39E-03   22.258  < 2e-16 ***
attributes.NoiseLevel4          -3.88E-01  7.17E-03  -54.14   < 2e-16 ***
attributes.RestaurantsAttire2    1.73E-02  6.15E-03    2.814  0.004901 **
attributes.RestaurantsReservations2 5.62E-04 2.76E-03  0.204  0.838542
attributes.RestaurantsTableService2 -1.42E-01 3.46E-03 -41.115 < 2e-16 ***
attributes.RestaurantsDelivery2 -4.31E-02  2.56E-03  -16.795  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4259 on 215843 degrees of freedom
Multiple R-squared: 0.3678,  Adjusted R-square
d: 0.3674
F-statistic: 775.3 on 162 and 215843 DF,  p-valu
: < 2.2e-16
Multiple R-squared:  0.3678,    Adjusted R-squared:  0.3674
F-statistic: 775.3 on 162 and 215843 DF,  p-value: < 2.2e-16
```

Figure 7: Simple Linear Regression Output Summary

On understanding the output, the summary of the variables t and p values are mentioned. Residuals is considered to be the error i.e. the difference between the predicted and the actual

values. Since, it can be seen that the residuals have a symmetric distribution around the mean, the model is considered to a good fit. Estimates denotes the slope i.e. the effect of the features on the target variable. The coefficient t-value signifies how much SDś our coefficient estimate is away from zero. Higher the distance, means the null hypothesis can be rejected and there is a strong relationship between the variables. From the summary, it can be seen that stars.x which represent the rating given by the user for each review on the basis of text and reviews_count have a strong relationship with the target variable stars.y having t-values greater than 100. Also the neighborhoods like SouthHill, Downsview, Upper Beach and Restaurant Categories like Food, Cafes, Noodles, Caribbean, Salad and Cocktail bars have a impact on the ratings. Also, the restaurant attributes like PriceRange, Noise Levels have a good t-value output. All the features have the p-value near to zero which indicates that the null hypothesis can be rejected and there is a relationship between the stars.y and the dependent features. Residual Standard Error is calculated to be 0.4259 which is the measure of the quality of a linear regression fit. The degrees of freedom 215843 is the number of data points that were taken into the estimation of the parameters used after taking into account the parameters.

MCMC LINEAR REGRESSOR

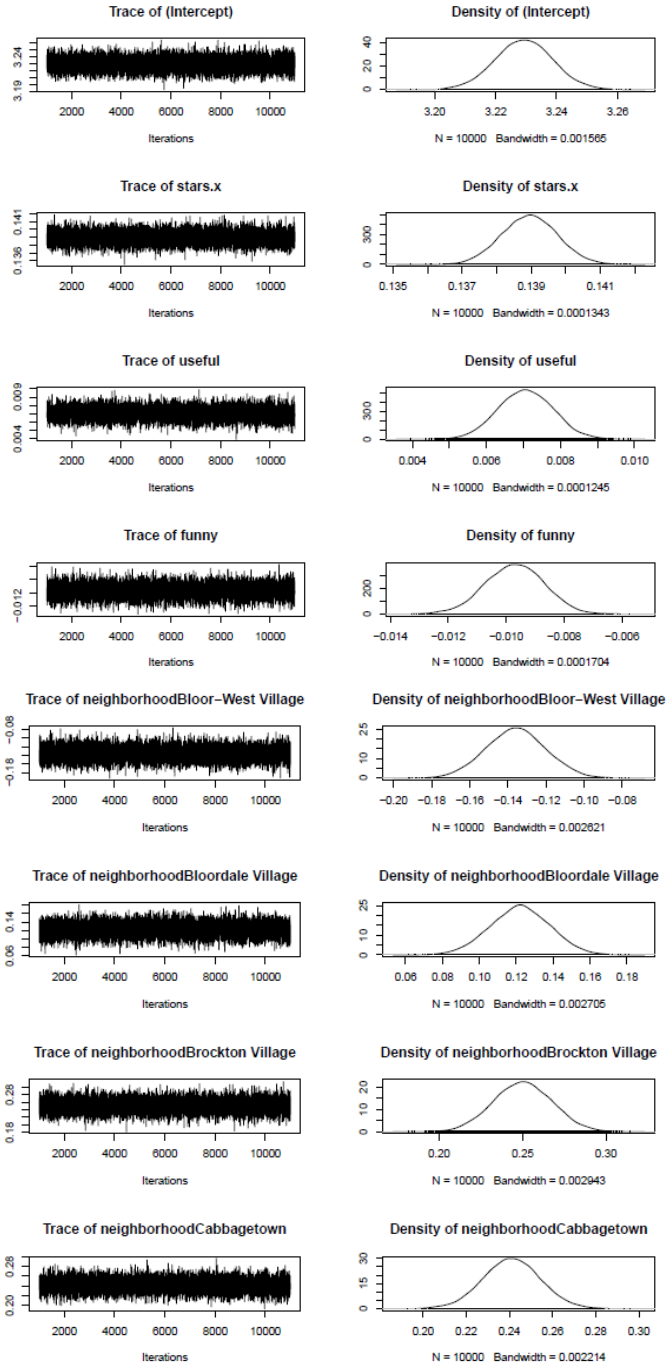
A Bayesian linear model is fitted using the MCMC regress package. The model to be fitted is given as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$b \sim \mathcal{N}_K(\mathbf{b}_0, \mathbf{B}_0^{-1}) \quad \sigma^2 \sim \mathcal{IG}(c_0/2, d_0/2)$$

The posterior distribution of the model is simulated using the MCMCregress package. The output from the MCMCregress function has been stored in an object called fit. Generally, a non-informative prior is used for the coefficient parameters by default in MCMCregress. It is better to consider weakly informative priors. The prior for b is given using the b0 and B0 hyper-parameters, where b0 is the prior mean and B0 is the prior precision of b. The prior for variance is controlled by the hyper-parameters c0 and d0. We use a weakly informative prior (c0 = 2 and d0 = 1). It is important to diagnose whether the Markov chain has explored the parameter space efficiently.

Using the plot method, the trace and density graphs diagnostics can be done to check whether the chain has converged to the stationary distribution. The trace and density plots are shown in the below figure. Trace plots analyzes the mixing of the chain and density are nothing but the histograms of the sample. The best trace plot is said to look like a hairy caterpillar avoiding the chain to be in a steady state for consecutive steps.


Trace of (Intercept)
Density of (Intercept)
Iterations
N = 10000 Bandwidth = 0.001565
Trace of stars.x
Density of stars.x
Iterations
N = 10000 Bandwidth = 0.0001343
Trace of useful
Density of useful
Iterations
N = 10000 Bandwidth = 0.0001245
Trace of funny
Density of funny
Iterations
N = 10000 Bandwidth = 0.0001704
Trace of neighborhoodBloor−West Village
Density of neighborhoodBloor−West Village
Iterations
N = 10000 Bandwidth = 0.002621
Trace of neighborhoodBloordale Village
Density of neighborhoodBloordale Village
Iterations
N = 10000 Bandwidth = 0.002705
Trace of neighborhoodBrockton Village
Density of neighborhoodBrockton Village
Iterations
N = 10000 Bandwidth = 0.002943
Trace of neighborhoodCabbagetown
Density of neighborhoodCabbagetown
Iterations
N = 10000 Bandwidth = 0.002214

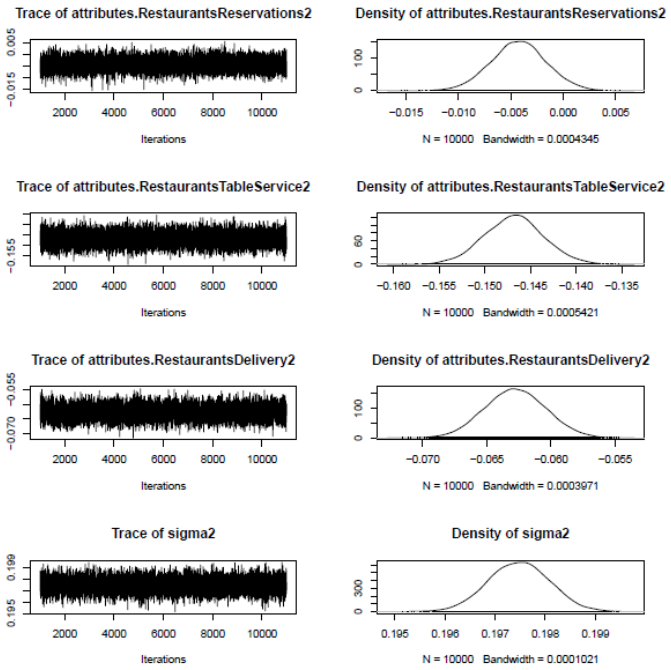


Figure 8: Trace and Density Plots for few Variables

Through the summary() function, we can check the properties of the fit. The default values of the iterations and other parameters are mentioned below:

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

It also describes the empirical mean and standard deviation for each variable , plus the standard error of the mean. The mean of all the columns is stored in the beta_means object. beta_means signifies the weight of each variable responsible or influencing the prediction of the rating. The results obtained are mentioned in the below Table 8. Clearly, it can be seen that the variables like stars.x (ratings according to the review text), Price-Range of 4th level, Noise level 3 have a higher weight as compared to other variables. If the categories of the restaurant is considered, restaurants categories with Noodles, Caribbean, Vegan ,Indian and Cafes have higher importance in influencing the rating of the restaurant. On the other side, attributes like absence of TV and very loud music or Noise Level 4 have a negative impact on the ratings. Other variables like cool, funny and whether credit cards are accepted or not are not of much importance to the final prediction of the ratings.

Table 8: Beta Mean Values

beta_mean	
(Intercept)	3.166066112
final1$stars.x	0.152394903
final1$useful	0.010487767
final1$cool	-0.002060817
final1$funny	-0.008345546
final1$review_count	0.000550639
final1$attributes.RestaurantsPriceRange22	-0.009497896
final1$attributes.RestaurantsPriceRange23	0.128198163
final1$attributes.RestaurantsPriceRange24	0.143626818
final1$attributes.BusinessAcceptsCreditCards2	-0.04013117
final1$attributes.HasTV2	-0.105156426
final1$attributes.NoiseLevel2	-0.108028114
final1$attributes.NoiseLevel3	0.119181799
final1$attributes.NoiseLevel4	-0.333463123
final1$attributes.RestaurantsAttire2	-0.051993741
final1$attributes.RestaurantsReservations2	-0.045082807
final1$attributes.RestaurantsTableService2	-0.102523784
final1$attributes.RestaurantsDelivery2	-0.055955334
final1$Food	0.092630105
final1$Canadian..New.	-0.001997944
final1$Indian	0.123163047
final1$Cafes	0.135030791
final1$Italian	0.033525132
final1$Korean	0.0394778
final1$Vegan	0.205613784
final1$Caribbean	0.256399754
final1$Noodles	0.092942468

The predicted values for the model is obtained by creating a dummy matrix and multiplying it with the beta_means which have the weights of all the variables. A dummy variable for intercept is added in the dummy matrix. The resultant column values obtained are the predicted ratings of the restaurant and the same are compared to the observed values. The performance of the model is calculated statistically using the metric Mean Squared Error (MSE) and visually by plotting the relation between the predicted and actual values.

Mean Squared Error = 0.25789 using MCMC Regression

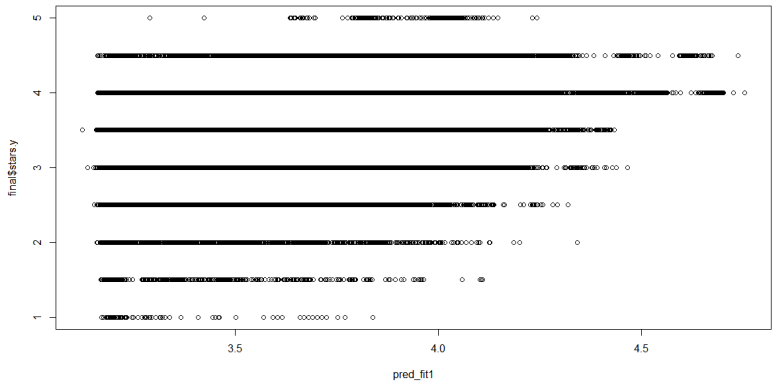


Figure 9: Predicted v/s Actual Ratings using MCMC

2.3 CONCLUSION

On analyzing both the linear regression model and MCMC regress, we can see that the MCMC regressor model performs better than the simple linear regression model. Both the models, predict similar variables which are most influential in predicting restaurant ratings. It can be concluded that the following features from our data have significant impact on a restaurant's rating: Review Ratings based on text (Stars.x), Noise Levels in the restaurant, Price Range of the restaurant. Also, the restaurant categories and Neighborhoods in which the restaurants are present do have a influence on the ratings provided by the user.

Table 9: Model Performance Comparison

Model	Simple Linear Regression	MCMC Regressor
Mean Squared Error	0.4259	0.25789

Table 10: Most Influential Variables

Most Influential Variables		
Restaurant Attributes	**Restaurant Categories**	**NeighborHoods**
Stars.x (Review Text Ratings)	Cafes	SouthHill
Noise Levels	Food	Upper Beach
Price Range	Caribbean	New Toronto
	Vegan	
	Noodles	

3 Association between Neighborhoods and Restaurant Categories

3.1 Data Handling

The business_open_Toronto.json dataset is used to find the association between the neighborhoods and the restaurant categories. In the dataset every restaurant contains a vector of categories which represents the different food categories the restaurant serves. The task is to segregate the restaurant categories into separate columns and identify whether a relation exists between them and whether a neighborhood is more likely to have certain types of restaurant categories. It can be seen that many of the restaurants in Toronto provide specific and niche category types such as Ethiopian, Tex-Mex, Nicaraguan etc. Therefore, it is necessary to clean the dataset and consider only those restaurant categories which are most common.

The nine most relevant categories are extracted from the list of categories using cbind() and rbind() functions. An indicator matrix denoting the categories that each restaurant belongs to is created using the %in% function. When there is match to the left operand of the %in% function, it returns a logical vector in binary format. The list of the top 9 most relevant categories obtained are mentioned in the below table. The dataset is merged with the existing dataset using the merge function and only the relevant categories and neighborhood features is considered for further analysis.

Table 11: List of 9 Most Relevant Categories

Restaurant Categories		
Food	Sandwiches	Canadian (New)
Nightlife	Breakfast & Brunch	Cafes
Bars	Chinese	Coffee & Tea

3.2 Model Building

Latent Class Analysis

As the task is to categorize different restaurant categories(observations) into different types of neighborhoods(latent classes), Latent Class Analysis model fits well in this problem. Latent Class Analysis (LCA) is a statistical technique for identifying hidden class groups among subjects using categorical or continuous observed variables. LCA is a method in which the groups are identified and created from latent subgroups, which are generally observations from multivariate categorical data. These models are also known as the finite mixture models.

The Bayesian Latent Class Analysis (BayesLCA) package is used to perform the LCA Analysis. There are three different methods blca.em,blca.gibbs and blca.vb to perform the blca analysis. The blca.em method finds the maximum a posteriori (MAP) estimates of the parameters using expectation-maximization algorithm.

It is very important to appropriate select the number of latent classes. As the BIC cross-validation method penalizes the likelihood criteria and heavily penalizes complex models, it is preferred. For eg, if BIC refers to 4 latent class and AIC refers to 6 latent class model, it is better to consider models with 4,5 or 6 groups. BIC is more efficient in selecting few classes rather than identifying all the latent classes.

example: **fit3 <- blca.em(categories, 3, restarts = 20)** # As more local nodes exist for models greater than 2, it is preferred to use more restarts.

The default number of restarts is set at 5. The model is fitted by blca.em with groups 3,4,5 and 7 to find the optimum group number and then the same are compared with each other using the Z-score and MAP functions. The BIC of all the models are tabulated and the group assignments of every fit is also shown in the below figure.

Table 12: Fit 3 Model Summary

Item Probabilities:

	Food	Nightlife	Bars	Sandwiches	Breakfast...	Brunch	Chinese	Canadian..	New.	Cafes	Coffee...	Tea
Group 1	0.144	0	0.000	0.093		0.070	0.105		0.055	0.030		0.000
Group 2	0.185	1	0.977	0.054		0.114	0.023		0.273	0.059		0.054
Group 3	1.000	0	0.000	0.226		0.220	0.021		0.020	0.577		0.781

Table 13: Fit 4 Model Summary

Item Probabilities:

	Food	Nightlife	Bars	Sandwiches	Breakfast...	Brunch	Chinese	Canadian..	New.	Cafes	Coffee...	Tea
Group 1	0.143	0	0.000	0.093		0.070	0.105		0.055	0.030		0.000
Group 2	0.136	1	0.975	0.046		0.102	0.023		0.282	0.036		0.000
Group 3	1.000	0	0.000	0.226		0.220	0.021		0.020	0.576		0.777
Group 4	1.000	1	1.000	0.185		0.316	0.031		0.123	0.438		0.963

Table 14: Fit 5 Model Summary

	Food	Nightlife	Bars	Sandwiches	Breakfast...	Brunch	Chinese	Canadian..	New.	Cafes	Coffee...	Tea
Group 1	0.169	0	0.000	0.153		0.110	0.006		0.086	0.045		0.000
Group 2	0.110	0	0.000	0.002		0.007	0.259		0.007	0.009		0.000
Group 3	0.137	1	0.975	0.046		0.102	0.023		0.282	0.036		0.000
Group 4	1.000	0	0.000	0.223		0.219	0.023		0.019	0.582		0.798
Group 5	1.000	1	1.000	0.186		0.316	0.031		0.123	0.438		0.971

Table 15: Fit 7 Model Summary

Item Probabilities:

	Food	Nightlife	Bars	Sandwiches	Breakfast...Brunch	Chinese	Canadian..New.	Cafes	Coffee...Tea
Group 1	0.141	0.000	0.000	0.003	0.017	0.137	0.023	0.019	0.000
Group 2	0.137	1.000	0.979	0.046	0.101	0.023	0.283	0.036	0.000
Group 3	0.180	0.005	0.000	0.097	0.331	0.000	0.240	0.089	0.000
Group 4	0.178	0.000	0.000	0.894	0.081	0.000	0.000	0.018	0.000
Group 5	1.000	0.000	0.000	0.026	0.069	0.004	0.000	0.899	0.755
Group 6	1.000	0.000	0.000	0.496	0.424	0.048	0.047	0.213	0.948
Group 7	1.000	1.000	1.000	0.186	0.316	0.031	0.123	0.438	0.974

Table 16: BIC Comparison Between Models

No of Groups	BIC
fit3	-20276.11
fit4	-20171.34
fit5	-20122.27
fit7	-20144.93

BIC score of all the models are compared and it can be seen that fit4 BIC value (-20171.34) is greater than the fit3 value (-20276.11). Also, from the above model summaries the group assignments of fit4 is better than the group assignments of 3. Also,the group assignments for fit4, fit5 and fit7 are similar. Hence, the four group model is preferred over other models and used for further analysis.

The probability of having a particular restaurant category is shown for each latent classes or groups in the model fit output. For example, consider the attributes Nightlife, Bars and Food for the model fit4 summary . The probability of restaurant category being "Nightlife" is 0% for the first group, 100% for the second group, again 0% for the third group and 100% for the fourth group. This pattern could be confirmed with restaurant category "Bars" having 97.5% for the second group, and 100% for the fourth group indicating direct correlation between the two categories of the restaurant. However, the difference between the second and the fourth group can be easily observed through the probability of category "Food" and "Coffee..Tea". The Group 4 has very high probability of having "Food" and "Coffee..Tea" categories whereas the same is very low in the Group 2. Same is the difference between the Group 1 and Group 3, Group 3 has a high probability of serving "Food" and "Coffee..Tea" whereas Group 1 has a very low probability of the said categories. Therefore, the groups or the latent classes can be differentiated through the conditional probabilities of the observations.

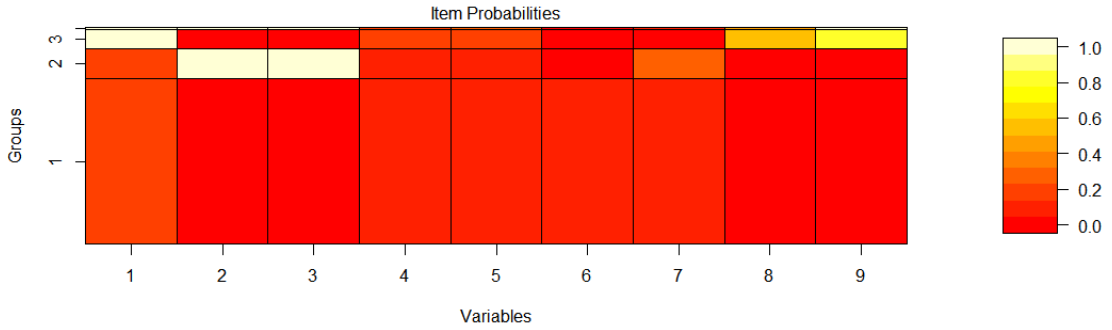


Figure 10: Item Probabilities

The following plot can be plotted using the following code
plot(fit4, which = 1).

It can be seen in the plot that large number of items belong to the Group 1 though the Group 1 has very low probabilities for all the features (observations). It is necessary to verify the convergence of the algorithm as blca.em is an iterative maximization algorithm. In Figure 11, after 6 iterations, it can be seen that the algorithm in very small margins.

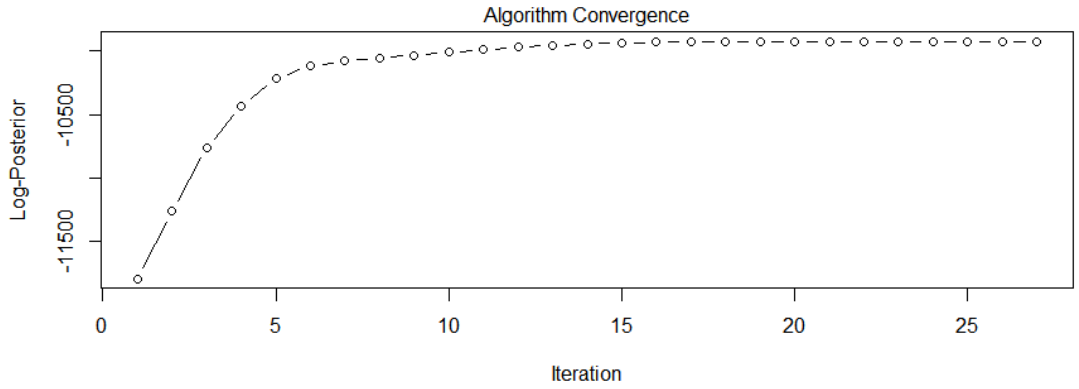


Figure 11: Algorithm Convergence

3.3 CONCLUSION

Table 17: Z score of Restaurant Categories

Neighbor-Hoods	Alexandra Park	Bayview Village	Beaconsfield Village	Downtown Core	Financial District	Kensington Market	Koreatown	Little Portugal	Milliken	Mount Pleasant and Davisville	Riverdale	Scarborough	St. Lawrence	Yonge and St. Clair	Yorkville
1	31	6	19	381	79	71	43	24	77	70	57	255	67	22	73
2	5	2	5	72	10	9	13	1	13	12	11	56	8	3	13
3	4	0	3	33	13	7	4	5	7	11	1	28	12	4	8
4	1	0	1	5	0	0	0	0	0	0	2	5	1	0	**0**

We can be certain that a relationship does exist between the neighborhoods and the restaurant categories. This can be verified by mapping each data point to a class based on the maximum posterio and mapping the Z scores of each restaurant categories with the Neighborhood attributes. For example, neighborhood Scarborough has a higher value in Group2 i.e 56 than that to Group 3 i.e.28. On mapping it with the probabilities obtained in the fit4, we can see that Scarborough neighborhood is more likely to have better NightLife and Bars but no food, whereas Little Portugal neighborhood is more likely to have Food and Coffee..Tea restaurant categories. It has low probability of 22% having " Sandwich","Breakfast..Brunch" restaurant categories.Similarly, we can check the association of rest of the neighborhoods and restaurant categories through conditional probability using Latent Class Analysis Model.

4 LIMITATIONS

1. It was a big dataset with many rows and columns. Therefore , the dataset had lot of noise.
2. There were lot of Missing Data for most of the Attributes
3. MCMCregressor package doesn't support factor variables and considers only numeric variables. Due to this, some of the attributes like PriceRange and Noise Levels were converted into numeric though being factor variables.
4. Analysis required high computational speed and resources. It took lot of computational time to build the models. e.g. Gibbs Sampling.
5. There were very less reviews for most of the restaurants. The dataset was not evenly balanced, which might lead to poor predictions.
6. Gibbs Sampling is a time consuming method and doesn't perform well or throws an exception with variables having few values or similar values.

5 REFERENCES

https://cran.r-project.org/web/packages/BayesLCA/BayesLCA.pdf
https://cseweb.ucsd.edu/ jmcauley/cse255/reports/fa15/017.pdf
https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a078.pdf
http://cs229.stanford.edu/proj2016spr/report/062.pdf
https://sites.lsa.umich.edu/admart/wp-content/uploads/sites/127/2014/08/jstatSoft11a.pdf
https://stats.idre.ucla.edu/mplus/dae/latent-class-analysis/