Introduction:

Yelp (Yelp 2018) is an American Multinational organization that encourages rating and review arrangement of the neighbourhood organizations. Yelp.com and Yelp Application is exceptionally acclaimed and broadly utilized by clients all around the world, in this manner it has a tremendous measure of information of the neighbourhood organizations as well as of users, their reviews, checkins and relation between clients to other client and to the organizations that make it an informationrich ocean prepared to be dug and investigated.

The Yelp dataset is the blend of different datasets: Business dataset, Users dataset, Reviews dataset check-in dataset. These datasets contain a gigantic measure of data which can break down the patterns and examples to answer different queries like Which class of business has most elevated ratings? Users with most supportive reviews? Factors affecting the evaluations of the business Locations with most elevated reviews? and some more.

Yelp dataset has a gigantic accumulation of organizations having a place with different classifications, users all around the world and a great number of reviews, in this way examining such a big dataset isn't plausible, that is why this analysis is focussed around a subset that contains only the restaurants situated in city of Toronto. This analysis is mainly focussed on answering three main hypotheses:

- Examination of evaluations of various neighbourhoods, state which neighbourhood is predominant than the other.
- Components affecting the evaluations of the restaurants ratings.
- Setting up a relationship amongst neighbourhoods and classifications. Are there any areas that contain certain kind of restaurant categories?

All the hypotheses specified in the above section are analysed by utilizing various statistical modelling techniques in the below sections.

Exploratory Data-Analysis:

The dataset utilized for the whole analysis is the combination of Business and the Review Yelp datasets. Both the json files are merged into one Rds format containing the data related to the city of Toronto. Merging is done based on the business_id field present in both the datasets. To make the data manageable flatten is used on the dataset. The missing values are imputed which were present in the price range field. Categories, a complex variable contains entries for restaurants of different lengths which turns into a tider format by isolating this variable and forming an indicator matrix which denotes the categories each restaurant belongs to, for the most 9 relevant categories present in the data.

Further to gain the insight of the data both univariate and multivariate techniques are used for the analysis. Variables are analysed independently as well as to look out the dependency they are analysed with each other. Also, variables are analysed with respect to the target variable “stars”. Some sentiment analysis is also applied on the reviews to look out for finding the variable importance and to explore the affecting factors on the target variable.

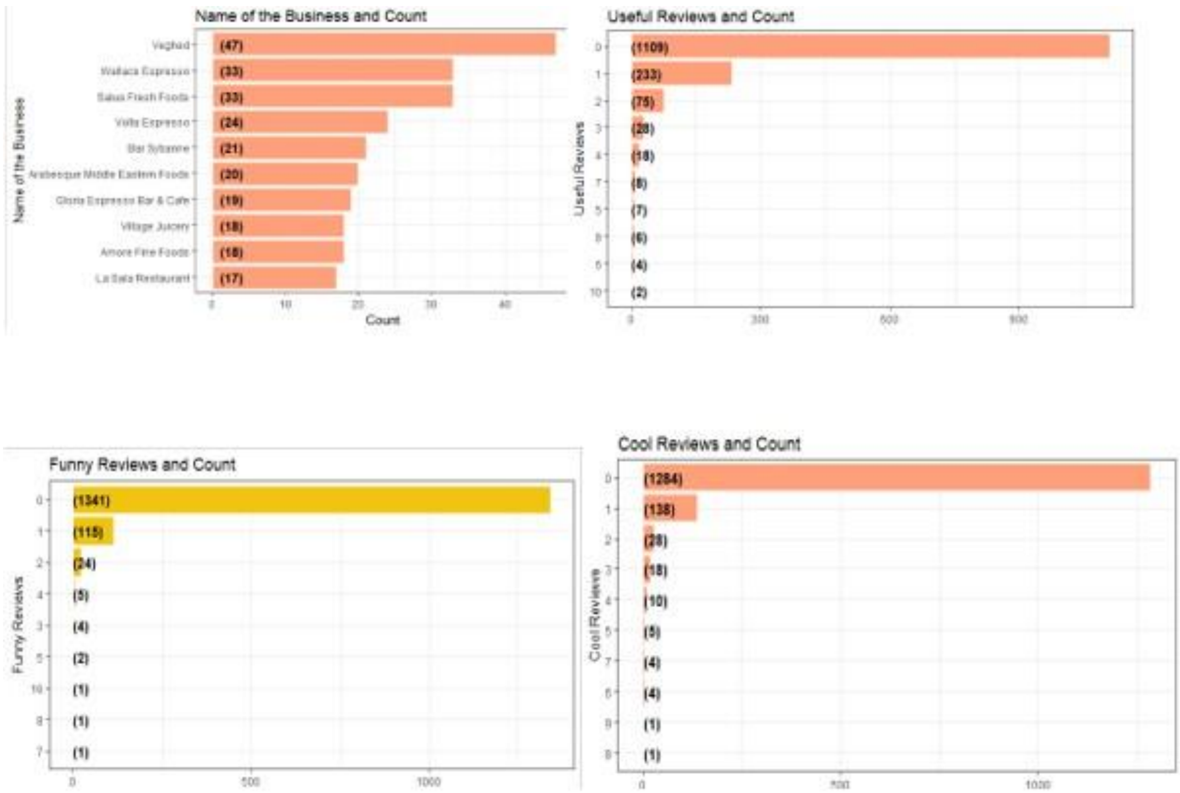
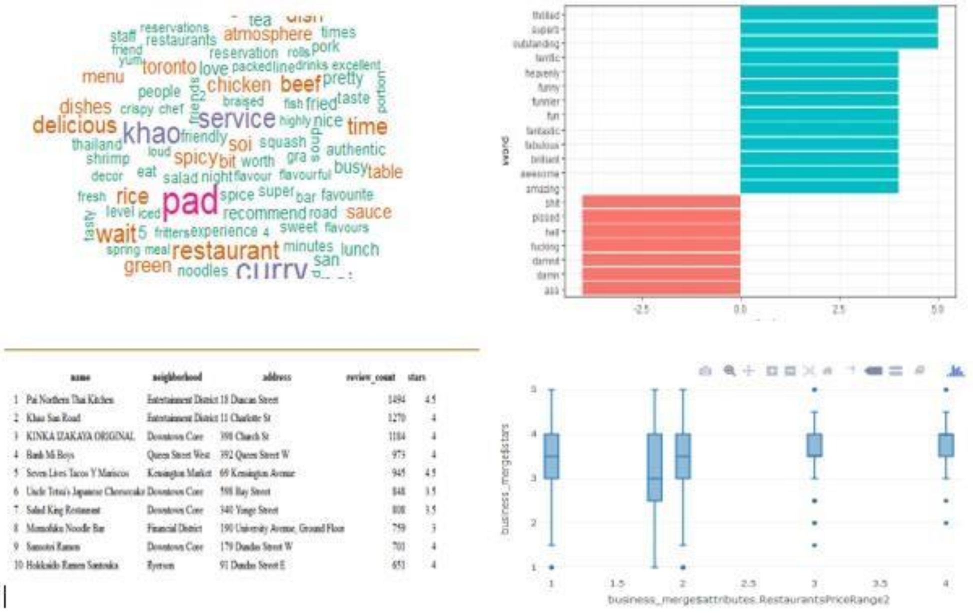


Figure 1 Figure represents the top businesses in Toronto and some review counts of different reviews



	name	neighborhood	address	review_count	stars
1	Pai Northern Thai Kitchen	Entertainment District	18 Duncan Street	1494	4.5
2	Khao San Road	Entertainment District	11 Charlotte St	1270	4
3	KINKA IZAKAYA ORIGINAL	Downtown Core	398 Church St	1184	4
4	Banh Mi Boys	Queen Street West	392 Queen Street W	973	4
5	Seven Lives Tacos Y Mariscos	Kensington Market	69 Kensington Avenue	945	4.5
6	Uncle Tetsu's Japanese Cheesecake	Downtown Core	598 Bay Street	848	3.5
7	Salad King Restaurant	Downtown Core	340 Yonge Street	808	3.5
8	Momofuku Noodle Bar	Financial District	190 University Avenue, Ground Floor	759	3
9	Sansotei Ramen	Downtown Core	179 Dundas Street W	701	4
10	Hokkaido Ramen Santouka	Ryerson	91 Dundas Street E	653	4

Figure 2 Figure represents the common words used in reviews and the plot between the price range and stars

Question 1: Compare the ratings of different neighbourhoods. Are any neighbourhoods clearly superior to others? If so, by how much?

For Finding the rating of different neighbourhoods the Hierarchical model approach is used. Before closer look onto the model, first the data is analysed. To get the thought of the data related to neighbourhood, a plot is made between the aggregated neighbourhood and the average stars corresponding to each neighbourhood.

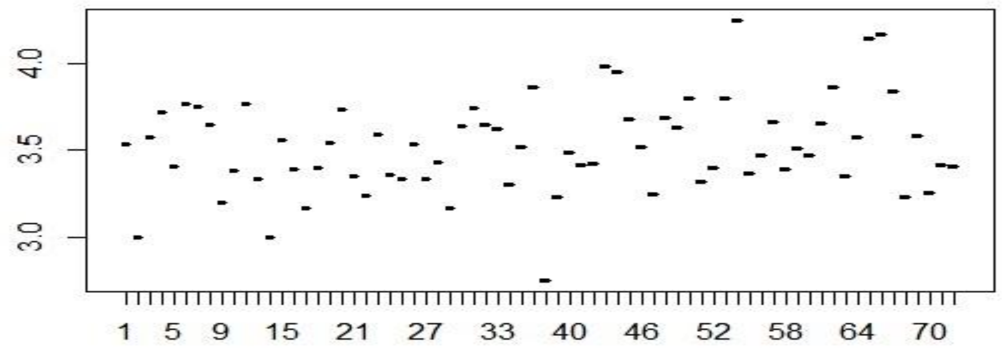


Figure 3 figure represents the stars corresponding to the neighbourhoods

The next plot is between the mean of stars and the corresponding neighbourhood.

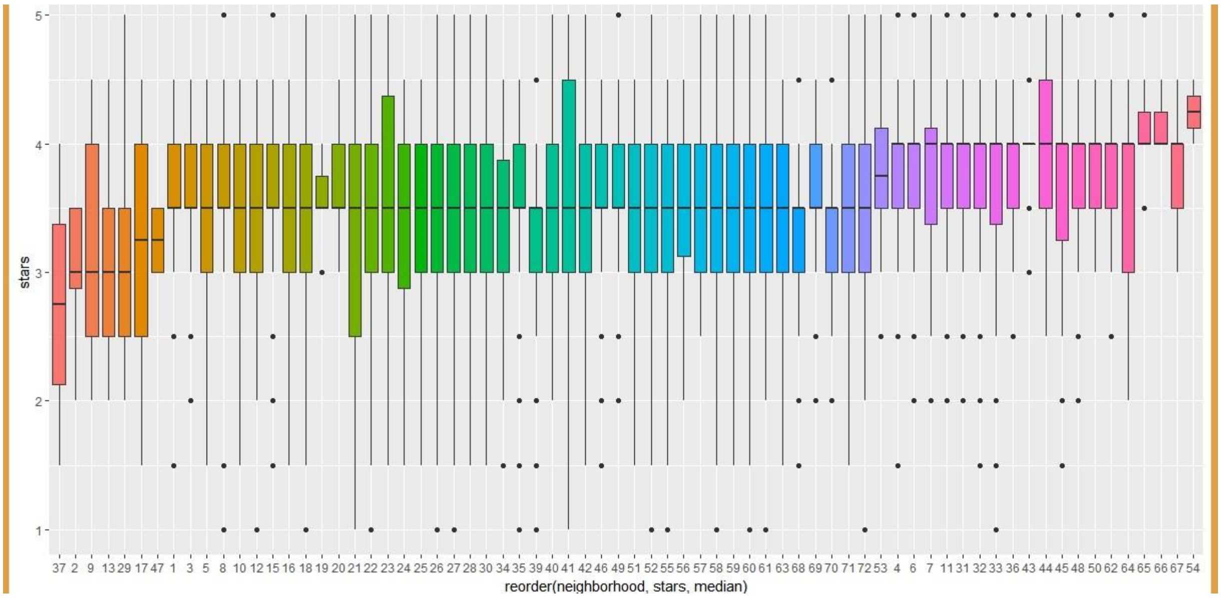


Figure 4 figure represents the mean ratings of the neighbourhoods

From the above plot it is clearly seen that data is skewed towards the higher ratings as most as the neighbourhoods are having more than 3.5 stars while very few neighbourhoods are having less than 2.5 stars.

Now To look onto the count of all ratings a plot is made

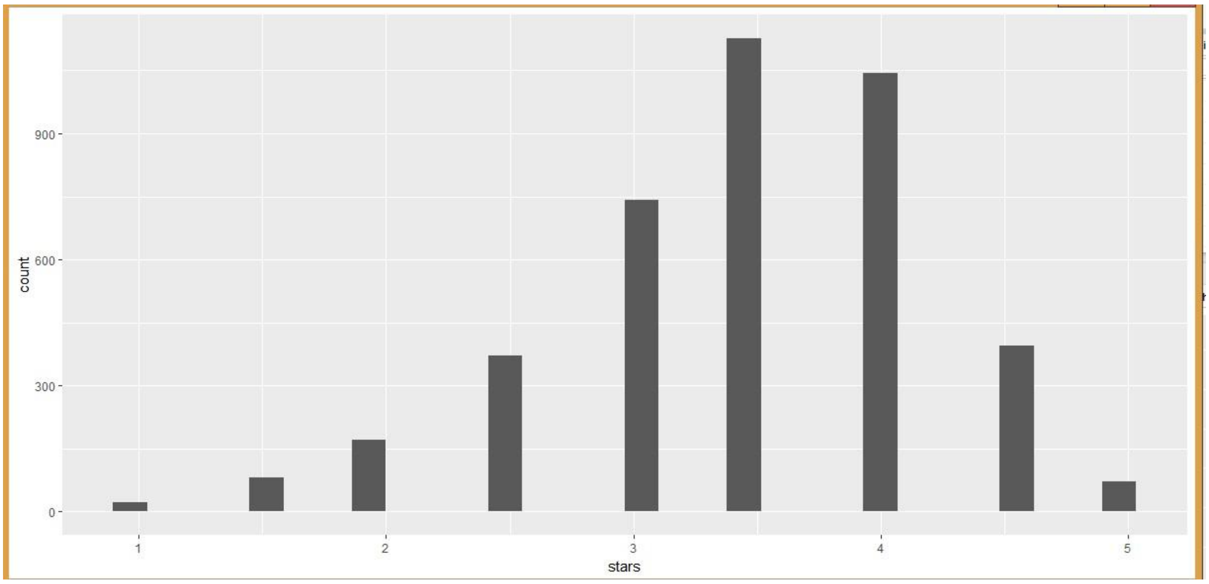


Figure 5 figure represents the number of counts of each star

From the above plot it can be seen that the highest number of reviews are for the 3.5 stars and negligible amount of reviews are present for the 1 and 5 stars which again making the data skewed which may cause the higher sample mean for the neighbourhoods with 3, 3.5 and 4 stars.

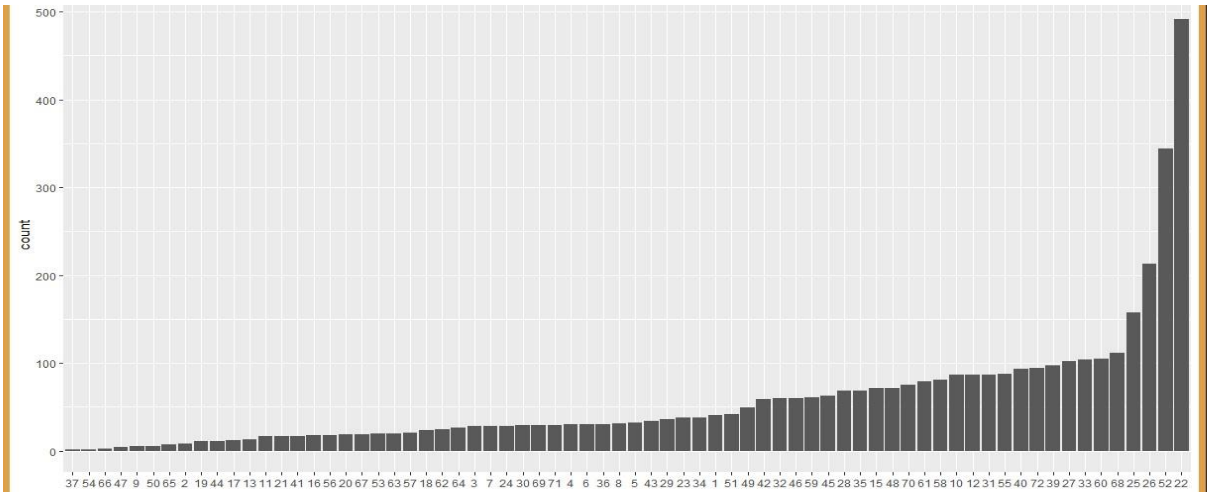


Figure 6 figure represents the count of reviews of the neighbourhoods

The above graph depicts the number of reviews of each neighbourhood. Downtown core is having highest number of reviews while Markland Wood and South Hill are having very less reviews.

Statistical Analysis

Gibbs sampler is utilised to model the difference between the mean stars of the neighbourhood. compare_m_gibbs function is defined in which the key quantities are:

- μ, the overall mean across the neighbourhoods
- τ_b, the precision between neighbourhoods;
- τ_w, the precision within neighbourhoods;

- θ_m, the mean awarded to neighbourhood m.

For building the model, different assumptions are used:

- The data sample y_1 = Neighbourhoods(y_1 ,…,y_{72}), and y_2 = stars. For both the samples Normal distributions are considered, $y_1 \sim N(\theta_1, 1/\tau)$ and $y_2 \sim N(\theta_2, 1/\tau)$ for all values respectively, and for the precision gamma distribution is taken, $\tau \sim$ Gamma (a_0, b_0). to compare the difference between the population means $d = \theta_1 - \theta_2$.

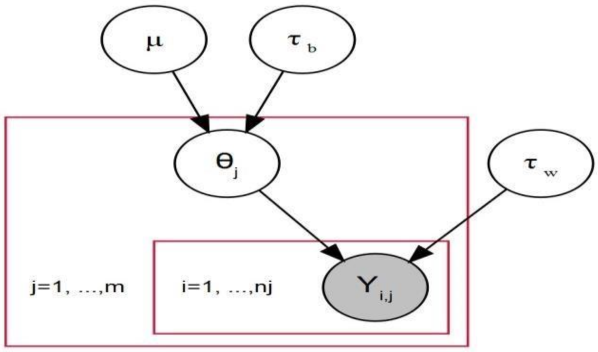


Figure 7 figure represents the model structure

- Gibbs sampler is utilized and the posterior is examined utilizing conditional posterior distributions for all the model parameters.
- The model is restrictively independent, which implies all values within a group are assumed to be autonomous.
- Groups are assumed to be independent at group level.
- τ_w is the precision within the group across all groups.
- the mean performance of the groups is normally distributed, and we assume that conditional on group membership, individual observations are normally distributed. The latter assumption is stronger than the first; again, both assumptions can be examined.

Model Analysis:

- The hyperparameter values best fulfilling the model are picked after experimentation. The default parameters did not do the trick, and the new parameters are put into the Gibbs work which keeps running for 3000, 5000 and 10000 iterations:

a_0 = 1.9 b_0 = 1 $etau_0$ = 0.5 t_0 = 5

mu_0 = 3.5 $gamma_0$ = 1.25

- On changing the values of prior many times, the priors don't strongly affect the posterior. The data assumes a more critical part.
- Even on the number of different iterations, no change can be seen in the values.
- On changing values of prior multiple times, we see that the priors do not have a strong effect on the posterior. The data plays a more important role.

Output:

	Mu	tau_w	SD_w	tau_b	SD_b

Mean	3.5165	1.6676	0.7642	4.8320	0.4603
SD	0.0583	0.0962	0.0272	0.8507	0.0411

From the table, the inferences are following:

- The mean of all neighbourhood is 3.51, which is like the average of all restaurants stars. tau_w tells how variable individual data points are conditional on belonging to group, and tau_b tells how variable the group means are from each other, and from the overall mean μ. The mean standard deviation between groups is 0.4603. From this, it can be concluded that the values will lie between $3.5165 \pm (0.4603)^2$.
- The variance within the neighbourhoods is higher than the variance between the neighbourhoods. This signifies the restaurants present within a neighbourhood are more varied, which makes sense because a neighbourhood will not have similar kind of restaurants.
- The standard deviation of means is 0.0583, which shows small amount of variance between sample means, a lot of variance is not recorder in the standard deviation within and between groups.

Next plot contains the values of theta, which holds the group mean rating for a neighbourhood. The theta parameter is present within the output of the Gibbs sampling function.

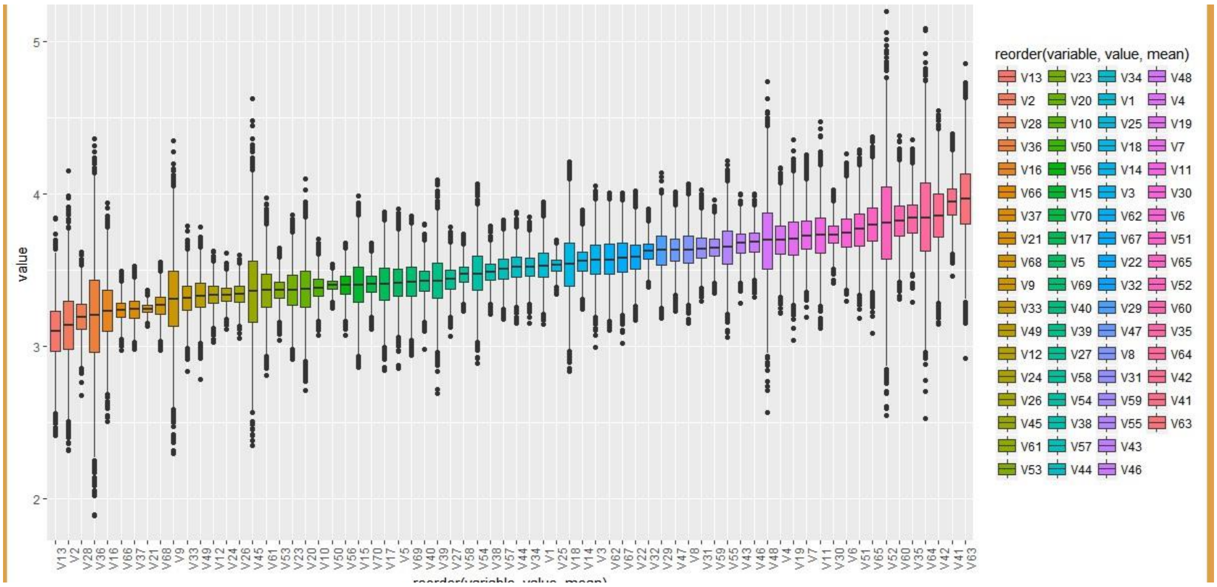


Figure 8 figure represents the mean ratings for each neighbourhood

Above graph is an important one, as it clusters neighbourhoods on their closeness to each other. Neighbourhoods with closest population means are shown by the same colour, and as the brightness decreases, the strength of relation also decreases. The graph represents two kinds of clusters, one in which neighbourhoods are significantly different from others, and other one in which clusters contain similar neighbourhoods, and also their difference from other clusters.

Now to check the significance among neighbourhoods, two random neighbourhoods are taken from two clusters. Neighbourhood 28 & 52 (Harbourfront and South Hill) are chosen.

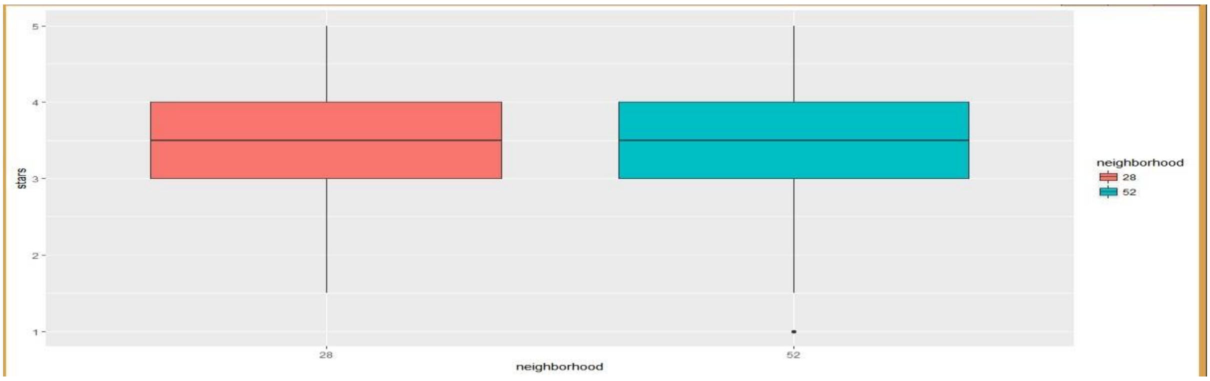


Figure 9 figure represents the comparison of two neighbourhoods

Now From the above plots it is very difficult to check the difference among the two plots. So, the violin plots are plotted to check the difference between two.

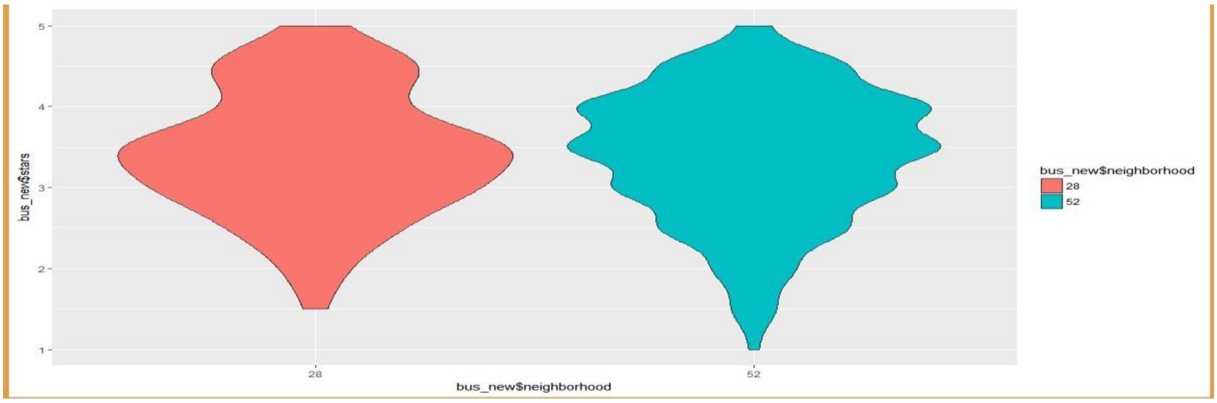


Figure 10 figure represents the comparison of two neighbourhoods

With the above graph for the same two neighbourhoods the difference can be seen among the distributions of the stars which was not evaluated in the previous box plot.

Now let’s look at other two neighbourhoods Harbourfront and Ossington Strip which are present in least and top neighbours list respectively according to the fitted model.

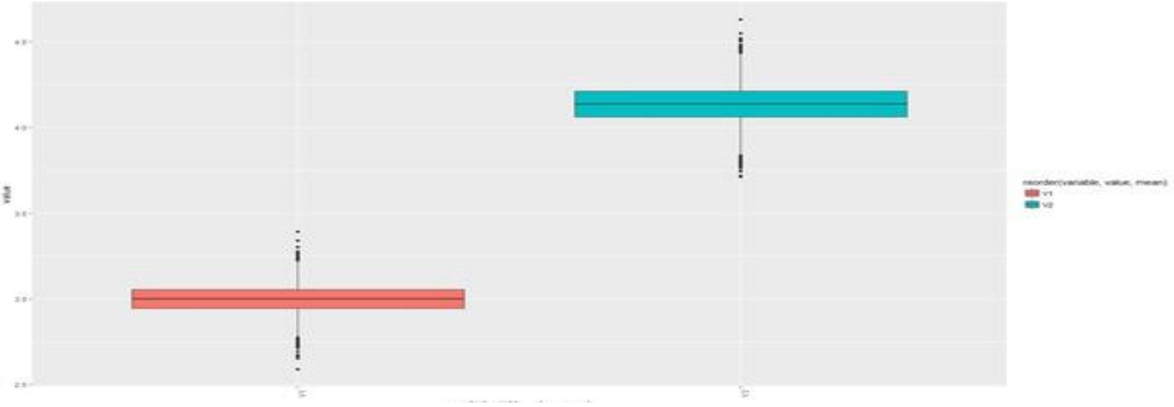


Figure 11 figure represents the comparison of two neighbourhoods

A decent approach is to explicitly demonstrate this distinction. Model the stars of neighbourhoods from every neighbourhood as:

- $\theta_1 = \mu + d$ and $\theta_2 = \mu - d$,
- $y_{i,1} = \mu + d + \epsilon$ i,1 <- for Ossington Strip
- $Y_{j,2} = \mu - d + \epsilon$ j,2 <- for Harbourfront

A hypothesis is taken, as a sample of size 10,000 is taken from both using the r-norm function, and mean function is utilised to get the probability that a randomly selected restaurant from neighbourhood Harbourfront have a better performance than a restaurant from neighbourhood Ossington Strip.

On computing, results show that the probability of randomly selected restaurant from neighbourhood 41 is having a better performance than a restaurant from neighbourhood 28 is 0.975.

The variance between groups value is 0.28, which show that both the groups have a significant variation.

2. What variables are most influential at predicting restaurant rating? How accurate are these predictions?

Keeping in mind the end goal to discover which factors influence the rating of a restaurant the most, we investigate the variables that affects to the stars. The merged dataset of business and review is used for this task. The different set of variables are considered in the analysis. Top 9 categories of the restaurants are taken during analysis.

- food quality
- restaurant has nightlife
- restaurant has a bar
- restaurant has Sandwiches
- restaurant has Breakfast/Brunch
- restaurant serves Chinese food
- restaurant serves new Canadian food
- restaurant serves Coffee/Tea
- restaurant café

Also, the sentiments related factors are taken into consideration in the analysis

- count of useful ratings of a restaurant
- count of cool ratings of a restaurant
- count of funny ratings of a restaurant
- The review counts of a restaurant
- Price range of the restaurants

Model Description:

Assumptions for the Bayesian Regression model (Adrian David Jennifer 1987):

- all covariates X = X1, . . . , Xp are informative.
- specific structure is most suitable, when any formulation ϕ(X) could be suitable.

MCMC regression or the Bayesian Regression is utilised for finding the variable importance on the target variable. MCMC package uses coda to perform posterior summarization and to evaluate the performance of the sample. Bayesian Regression model is given by:

$Y_i = X_{iT} + \varepsilon_i$

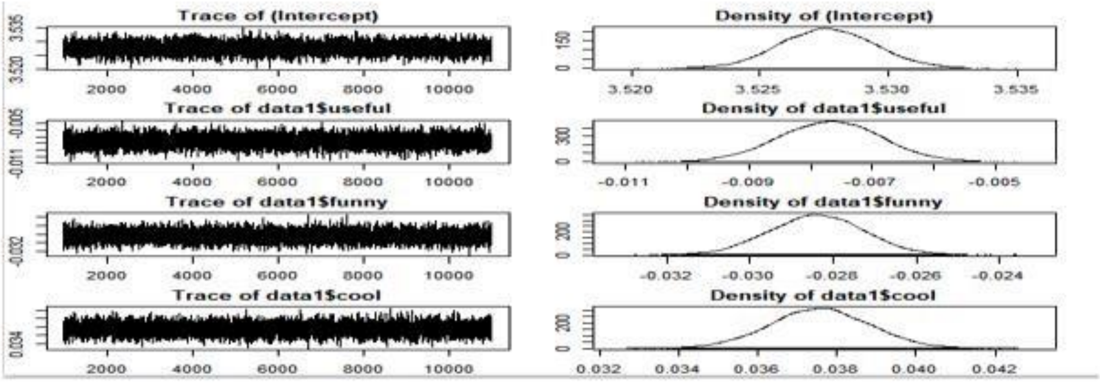
Where X_i is the predictor variables, and ε_i is the normal distributed random variables.
The model gives the posterior density sample from a linear regression model with Gaussian errors and it makes use of Gibbs sampling with a multivariate Gaussian prior on the beta vector, and an inverse Gamma prior on the conditional error variance.

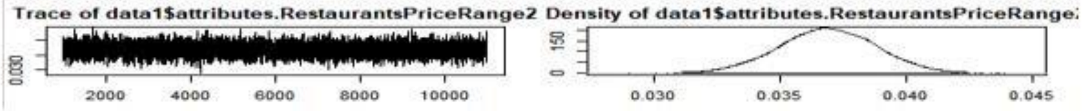
Bayesian regression model is fit to find the average stars of a restaurant based on the variables. Gibbs sampling is used to get a sample from the posterior distribution. Model is fitted and the following coefficient values are obtained:

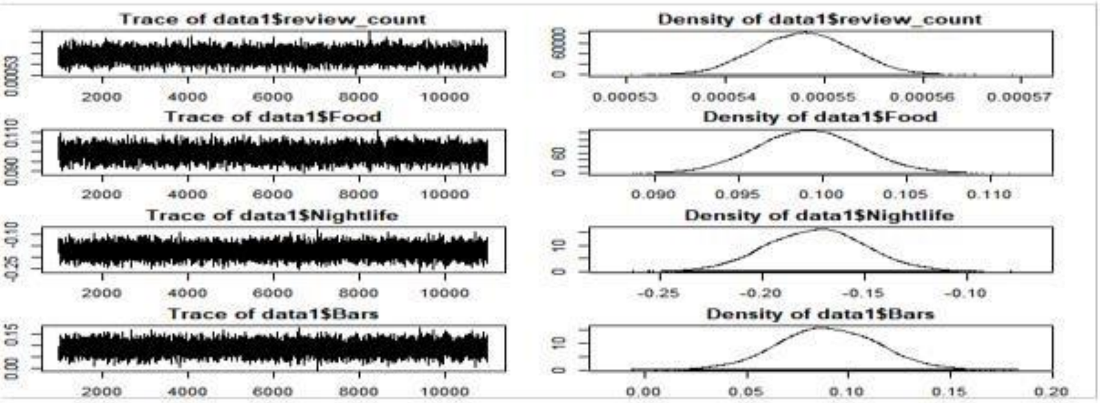
Useful	Cool	Funny	Price Range	Review Count	Food	Bars	Nightlife	Sandwiches	Breakfast/Brunch	Chinese	Canadian	Café	Tea/Coffee
-0.0084	0.038	-0.03	0.436	0.005	0.1055	0.0874	-0.1775	0.1346	-0.0336	-0.1899	-0.028	0.1299	0.077

The Mean Squared Error value obtained for the model is 0.512, so it can be stated that the model is accurate in predicting the ratings.

As per the results obtained from the MCMC regression coefficients, the most important variables affecting the rating of restaurants are Price Range, Café, Food, Chinese and Nightlife. Within sentiments related variables cool review counts is the most significant. For each variable Trace and density plots are plotted:







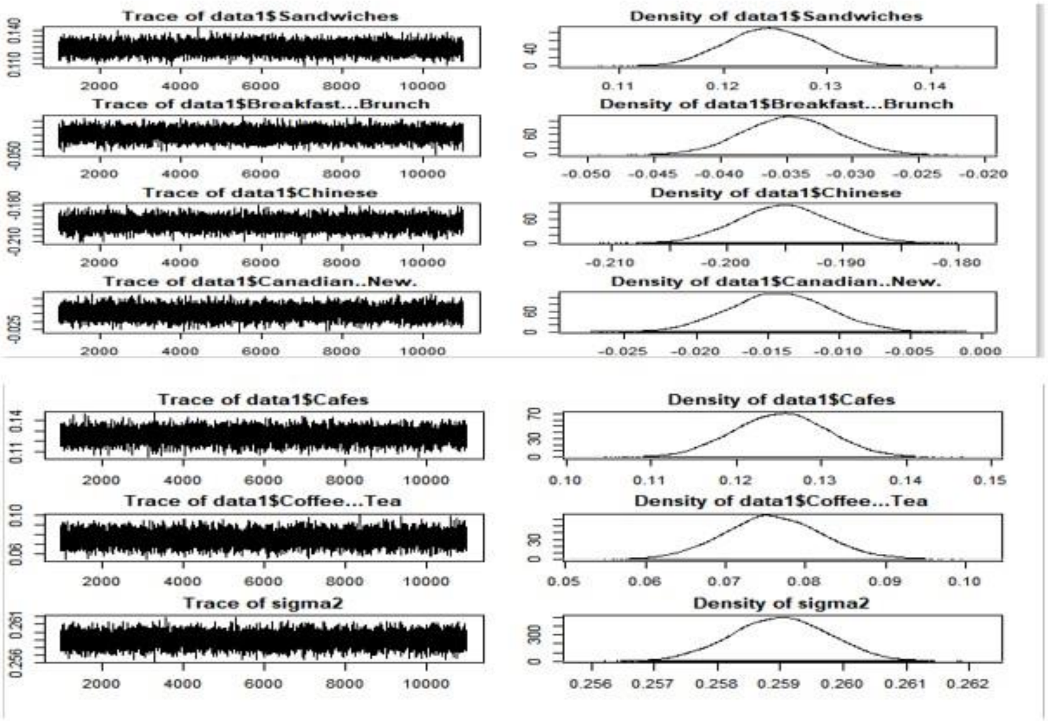


Figure 12 figure represents the Trace plots and Density plots for each variable used in the model fit

In the above figure two types of graphs are present, first one is the trace plot between the iterations and the sampled values within in the chain. Trace plots are indicating good sampling for the respective data. Second is the density plots which depicts the beta-coefficient values corresponding to the different variables.

MCMC regression Model fit:

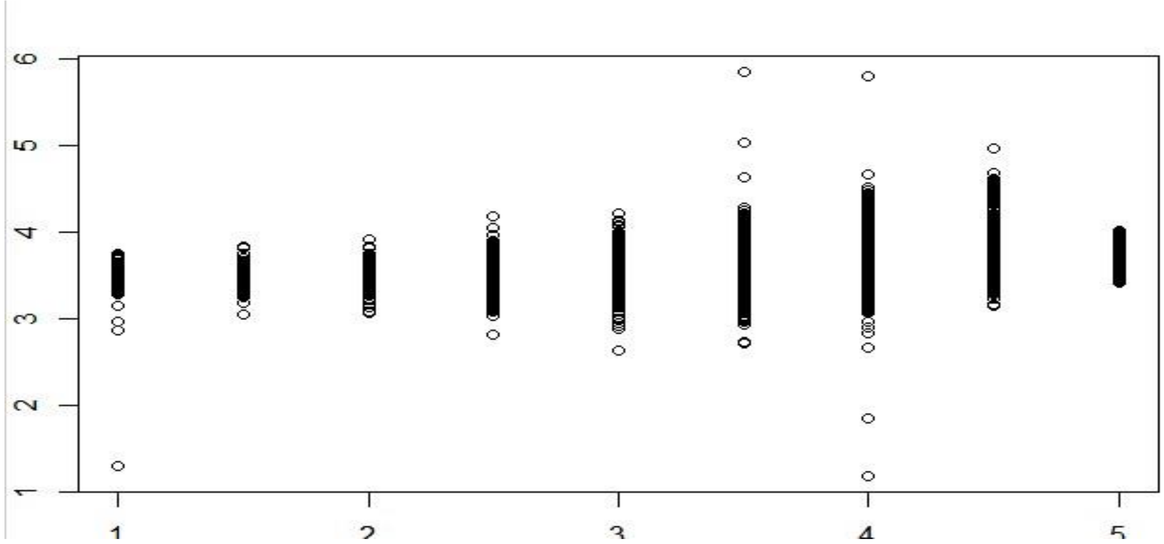


Figure 13 figure represents the MCMC model fit

Above graph depicts the model fit which is quite a decent fit as for the ratings 3, 3.5, 4 and 4.5 model is predicting well, while it does not predict well for other ratings may be due to following reasons:

- The data is skewed towards higher ratings, as lesser data points are present for lower ratings.
- There may be some latent variables present due to which output is deviating.

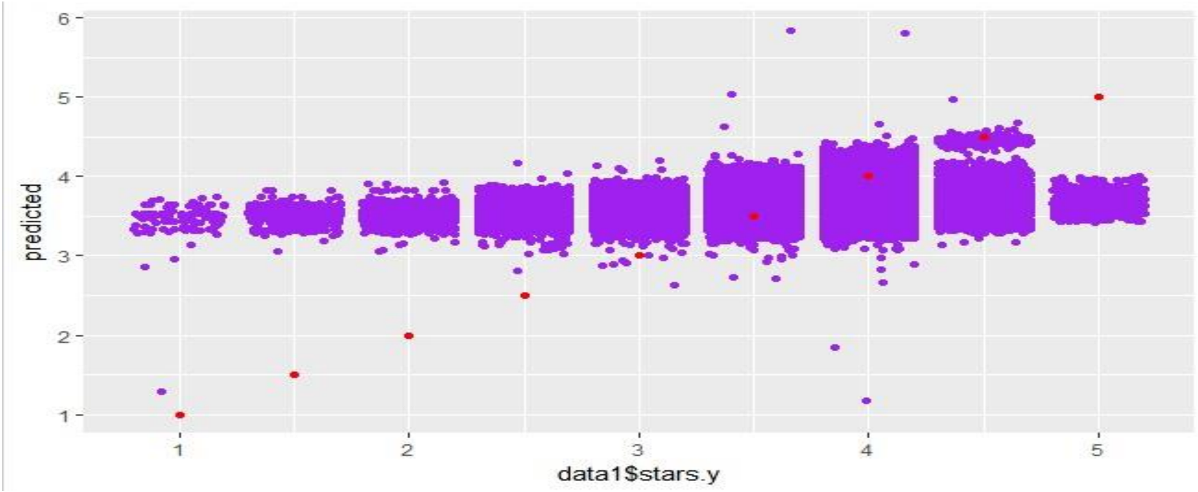


Figure 14 figure represents the MCMC fit with the outliers

Other Models can be considered for such set of problem. To check the results Random Forest Model is used which predicts better results than the MCMC regression which can be seen in the below graph. As this model was good with the handling of outliers. Bootstrap and cross-Validation approaches might be the reason for the better prediction of the Random Forest Model.

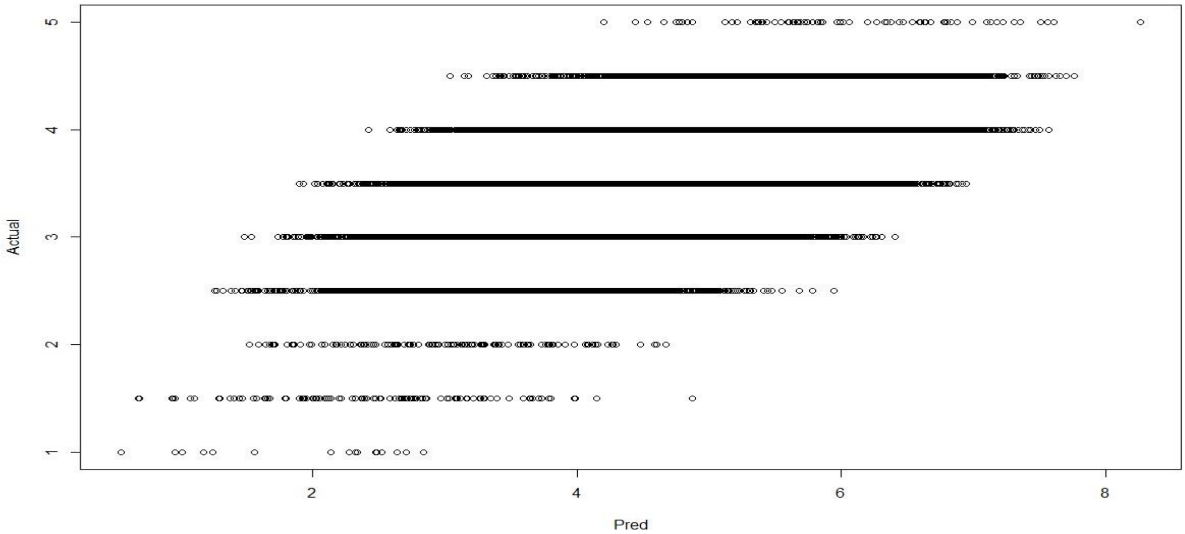


Figure 15 figure represents the Random Forest model fit **Model**

Evidence:

Akaike information criterion and Bayesian information criterion are utilised broadly in statistics to estimate the relative model quality. Both the criterions compare multiple models within the set, and estimate the quality of the models relatively.

AIC and BIC estimators are used to find the optimal model, with the significant variables. Step function is used by AIC. For BIC, a log function adds within the step function of AIC corresponds to the number of rows in the data frame. The lower the score of AIC, better will be the model.

The AIC and BIC functions are applied the Linear regression model. On execution of the steps function to calculate the AIC and BIC scores, the optimal model and with the important variables is obtained in the second run, with both AIC and BIC scores "-316930". Though this was time consuming criterion. For improving the performance "is_open" field is removed by the criterion in the second run. With the variables obtained by this method are then now used for MCMC regression.

Next, model selection is performed for the MCMC regression model. In the MCMC regress fit function an additional parameter is included, the marginal likelihood with the argument "Chib95". It has provided a sample from the posterior distribution of a binary model with numerous points. Next, the B_0 parameter is added which includes the prior precision for the Beta value.

Model selection has been performed for 3 model fits for comparing the marginal likelihood of each:

- Model fit1: In this model fit, MCMC regression model runs including all the variables. The marginal likelihood value for this fit is -169478.4.
- Model fit2: For this fit important variable obtained with the last AIC fit are used. The marginal likelihood value for this fit is -160696
- Model fit3: In this fit some variables are removed from previous fit to check for the change. The marginal likelihood value declines in this case -167591.3.

Model fit1	Model fit2	Model fit3
-169478.4	-160696	-167591.3

Conclusion:

In this model, the higher the marginal likelihood score, better the model will be. With this, it can be concluded that Model fit2 is the optimal model. Also, it reinforces from the output of the variables obtained from AIC fit which further proves that the set of variables which are stated above are the

3. Is there any association between neighbourhood and restaurant categories? Can you identify neighbourhoods that are more likely to contain certain types of restaurant category than others?

Latent Class Analysis (LCA) (Arthur White 2014)is a model based clustering method in which factors are classified into mutually exclusive groups, which are known as latent classes. The classification is based upon the factors interaction with each other, and the relationship between the output and the combination. LCA helps in detecting the unobservable subgroups within a problem.

Assumptions for LCA

- The variables in each class are independent of each other.
- The data fed into the model is categorical data, and not continuous data.
- LCA does not have any assumptions with respect to the data being linear, non-linear or normally distributed.

Conformation of Assumptions:

- The variables fed into the model are analysed and found to be categorical.
- For checking independence of variables "chisq" function is used to performs a Chi-squared Test.

Model Description:

For finding association between neighbourhoods and restaurant categories, the data is processed further for better analysis. The Business dataset is utilised for this analysis. Dataset is reduced and only two fields are considered during the analysis, namely the neighbourhood names and the categories present with in them. Due to the size of the dataset, it is not feasible to work with the entire data. So, categories are looked out first. Only the top most 12 popular categories are selected for the model. After that neighbourhoods are analysed and 20 neighbourhoods are picked, which contains the highest number of categories. The sample of data is being used for the analysis as the clustering for all the categories in plots was not a good way to interpret the results due to the large dataset. In this manner, sample data well represents the population data without any bias and skewness of the data.

The Bayesian Latent Class Analysis package is used to perform latent class analysis in R. Binary Data is required to find out the number of hidden classes. As Latent class analysis is done on the categorical data, dataset is changed and all the categories are encoded as 1 and 0, where 1 is representing category is present while 0 is coded for the absence of the category corresponding to the neighbourhoods. The bcla.em makes use of an expectation-maximization algorithm to find the estimates.

LCA can also be performed using Gibbs sampling. For determining the optimal value, the tuning parameters chosen by looking at the trace plots. However, analysis took too long on running the blca.gibbs function with 10000 iterations.

The number of restaurant categories taken are 12, the effectiveness is checked by taking 4,5,6 and 8 groups. The BIC value of the groups is taken as a measure for comparison of the models. fit_num$BIC is utilised to find the BIC value for each group. The higher the BIC score, better the model.

- 4 groups: -17137.09 • 5 groups: -17088.64 • 6
groups: -17111.59
- 8 groups: -17207.35

It can be drawn that the 5-groups model has the best BIC value. The expectation-maximization approach is an iterative one, and provides a way to check if the algorithm has converged sufficiently. Also, there is no further improvement has been recorded and the most recent iterations does not improve the convergence at all.

Encoding for the categories:

Food	Nightlife	Bars	Sandwiches	Breakfast	Chinese	Canadian	Cafes	Coffee	Pizza	Fast Food	Italian
1	2	3	4	5	6	7	8	9	10	11	12

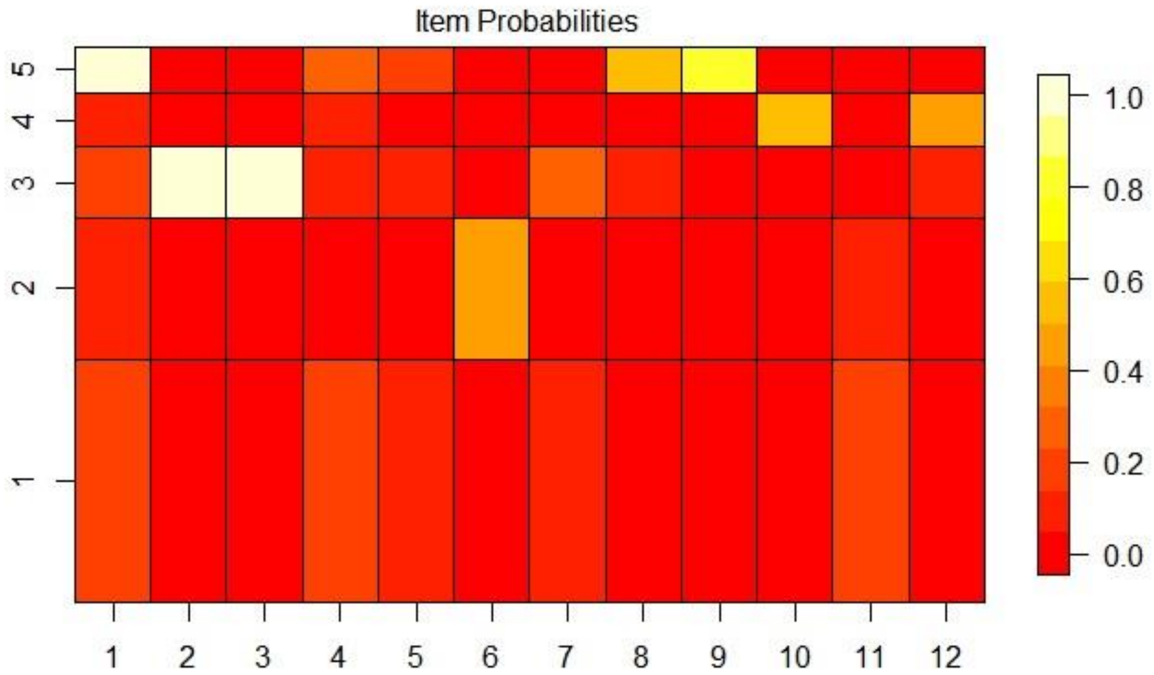


Figure 16 figure represents the groups division corresponding to the categories

In the above plot, the probability of each category falling within a group can be seen clearly. Group 5 has a high probability of containing category 1, and a medium probability of containing categories 8 & 9. Likewise, group 3 has a high probability of containing category 2 & 3, and a medium probability of containing category 7 and 1. Moreover, Group 4 is having moderate probability of containing categor ies 10 & 12. Group 1 contain neighbourhoods with low probabilities of categories 1, 4 and 11.

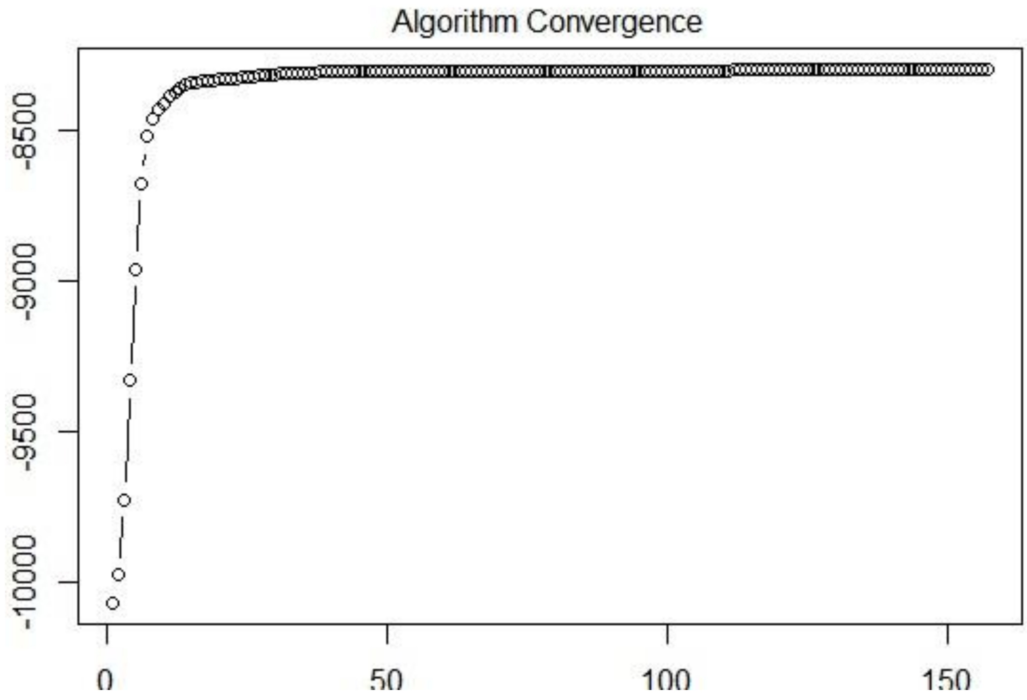


Figure 17 The sequence of lower bound as the algorithm proceeds towards convergence.

Number of restaurants in each group:

1	2	3	4	5
1590	282	340	220	207

Next, mapping of the neighbourhoods are made to the 5 groups.

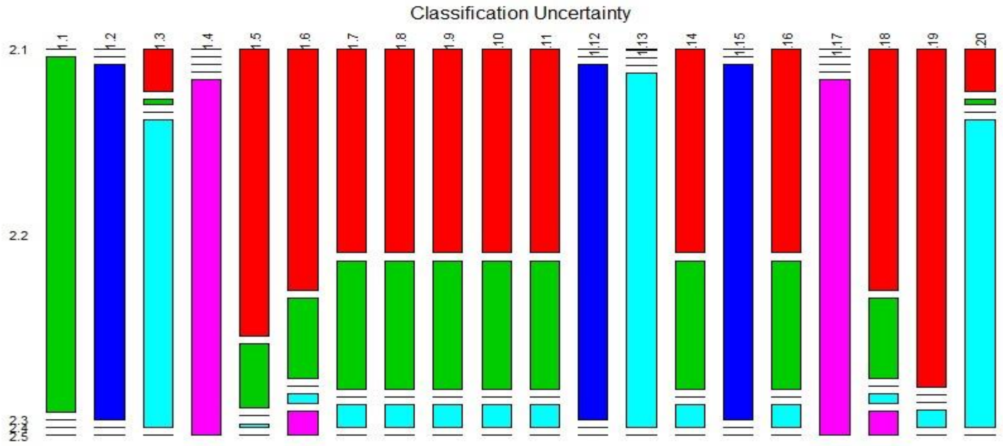


Figure 18 figure represents mosaic plot for the classification uncertainty

From inspection of the above figure, it can be drawn that around half of the points in the dataset are clustered with high levels of certainty, while the other half are still quite well distinguished.

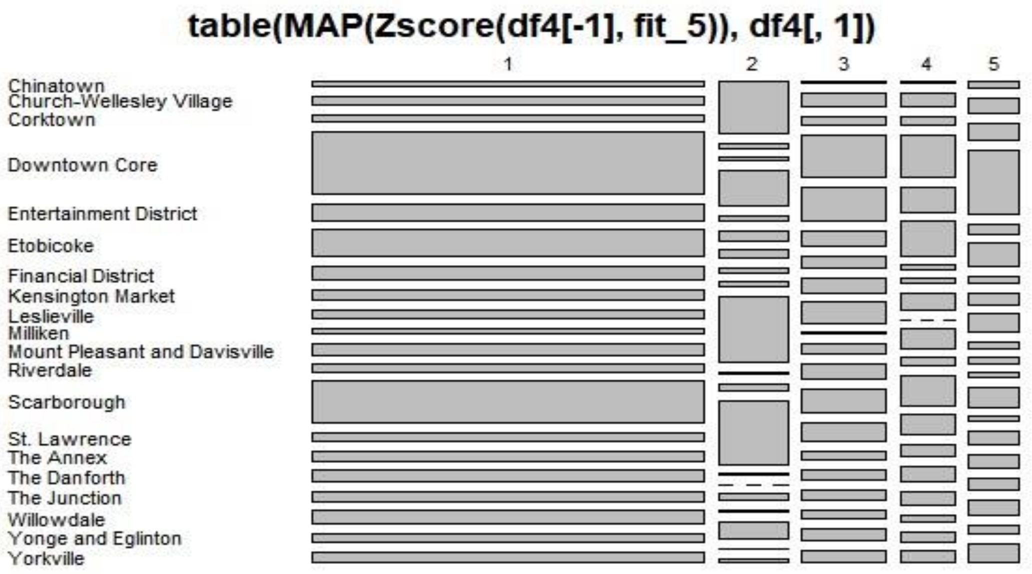


Figure 19 figure represents the mosaic plot of the neighbourhood belongs to the group number

The plot above represents the probability of the neighbourhood falling in a group (width of the box), and the size of each group (length of the box) i.e. classification Uncertainty. After combining the three plots, the following inference are made:

Group	Neighborhoods	Categories Probability
Group 1	Downtown Core	Food: Medium Sandwiches: Medium Fast Food: Medium
	Yonge and Eglington	
	Entertainment District	
	Etobicoke	
	Willow dale	
Group 2	Chinatown	Food: Medium Chinese: High Breakfast:Medium
	Downtown Core	
	Milliken	
	Scarborough	
	Willow dale	
Group 3	Downtown Core	Nightlife: High Bars: High Canadian: Medium Food:Medium
	Entertainment District	
	Church-Wellesley Village	
	Riverdale	
	Scarborough	
	St. Lawrence	
	The Annex	
Group 4	Etobicoke	Pizza: High Italian: Medium
	Downtown Core	
	CorkTown	
	RiverDale	
	Scarborough	
	Yorkville	
Group 5	Downtown Core	Food: High

	Etobicoke	Sandwiches: Medium Café: High Coffee: High
	Scarborough	
	The Junction	
	WillowDale	

Conclusion:

The following inference are made on the association between neighbourhood and categories:

- Neighbourhoods in group 1 contain restaurants of Food and Fast Food such as sandwiches.
- Neighbourhoods in group 2 contain restaurants of Food and Chinese cuisine. neighbourhoods are also popular for Breakfast.
- Neighbourhoods in group 3 contain restaurants of Food with Bars and Night Life. Also, Canadian restaurants are present in these neighbourhoods.
- Neighbourhoods in group 4 contain restaurants of Italian Cuisines and Pizza restaurants.
- Neighbourhoods in group 5 contain restaurants of Food, Sandwiches, along with café and Coffee restaurants.

Downtown Core and Scarborough are the most popular neighbourhoods Which are having the most variety of restaurants, and that might be a reason for being so popular. Also, these neighbourhoods contains distinct categories of restaurants which again makes them popular.

References

Arthur White, Thomas Brendan Murphy. 2014. "*BayesLCA: An R package for Bayesian latent class analysis.*" 28.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* 92, no. 437 (1997): 179-191.

2018. *Yelp.* 01 11. Accessed 04 02, 2018. https://www.yelp.com.

https://www.scss.tcd.ie/~arwhite/Teaching/CS7DS3/?C=M;O=D

Appendix:

Table contains the neighbourhood names corresponding to their ID's

neighbourhood_ID	neighbourhood name
1	Alexandra Park
2	Bayview Village
3	Beaconsfield Village
4	Bickford Park
5	Bloor-West Village
6	Bloordale Village
7	Brockton Village
8	Cabbagetown
9	Casa Loma
10	Chinatown
11	Christie Pits
12	Church-Wellesley Village
13	City Place
14	Cooksville
15	Corktown
16	Corso Italia
17	Deer Park
18	Discovery District
19	Distillery District
20	Dovercourt
21	Downsview
22	Downtown Core
23	Dufferin Grove
24	East York
25	Entertainment District

26	Etobicoke
27	Financial District
28	Greektown
29	Harbourfront
30	High Park
31	Kensington Market
32	Koreatown
33	Leslieville
34	Liberty Village
35	Little Italy
36	Little Portugal
37	Markland Wood
38	Meadowvale Village
39	Milliken
40	Mount Pleasant and Davisville
41	New Toronto
42	Niagara
43	Ossington Strip
44	Palmerston
45	Parkdale
46	Queen Street West
47	Rexdale
48	Riverdale
49	Roncesvalles
50	Rosedale
51	Ryerson
52	Scarborough
53	Seaton Village
54	South Hill
55	St. Lawrence
56	Summer Hill
57	Swansea
58	The Annex
59	The Beach
60	The Danforth
61	The Junction
62	Trinity Bellwoods
63	University of Toronto
64	Upper Beach
65	Wallace Emerson
66	West Don Lands
67	West Queen West
68	Willowdale

69	Wychwood
70	Yonge and Eglinton
71	Yonge and St. Clair
72	Yorkville