Roman Shaikh
Student number: 18300989

CS7DS3-Applied Statistical Modelling
Main-Assignment

Contents

Introduction:

Yelp was founded in 2004 in the US to help people find great local businesses like restaurants, dentists, hair stylists, and mechanics. Yelp works by people reviewing the local businesses and helping others. Yelp users have over 177 million of reviewers by the end of Q4, 2018.

The dataset provided by yelp covers a huge spectrum of information that yelp has on various business areas like user reviews, business information, and user details. Yelp has made this data set publicly available for analysis and study purposes. It is provided in the JSON format with three types of file. With this enormous data available a few interesting analyses can be done to answer questions like what are the upcoming businesses.? What business are most people moving to over the years? etc.

This huge dataset can prove a bit challenging to analyze. Therefore we focus only on a specific subset of the dataset which contains data for restaurants in Toronto, Canada. We will focus our attention in to answer the following questions in this report.

- Scrutinizing the reviews of various neighborhoods and evaluating the ratings of categories across multiple neighborhoods.
- Predictions of factors using the restaurant datasets.
- Classification of neighborhoods based on common factors, clustering based on specific factors.

We analyze all the above questions, with the help of statistical modeling techniques.

Question 1:

Comparing the ratings of currently open Indian restaurants in the neighborhoods of Scarborough and Etobicoke. Which neighborhood is best for this kind of food? How much better? Compare the ratings of (open) restaurants across multiple different neighborhoods in the city. Are any neighborhoods clearly superior to others? If so, by how much.?

Processing the Data:

We first load the JSON files into R for manipulation. Data cleaning is performed after the whole dataset is loaded into R. As required by the problem, we only require the data for two areas Scarborough and Etobicoke. Looking at the data we identify the following columns for processing.

- Business_id
- Name
- Neighborhood
- Address
- State
- City
- Postal_code
- Latitude
- Longitude
- Stars
- Review_Count
- Is_open

The highlighted columns are selected based on criteria that they will help us identify the Indian food category. Also, the data for open restaurants are only considered. For simplicity, the values for Etobicoke and Scarborough are encoded as 1 and 2 respectively.

Analysis:

We start by creating the boxplot for the data that we have. Fig. 1 shows the output for the boxplot.

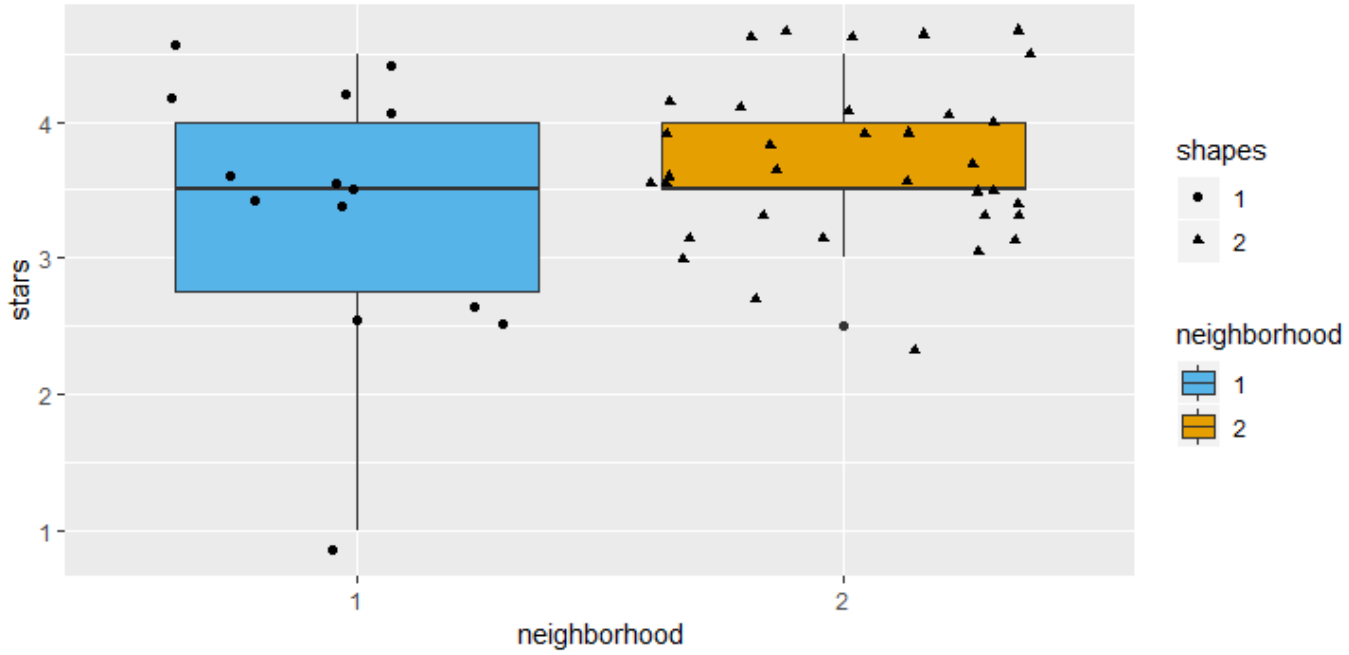

Fig. 1. Box Plot

As we observe from the box plot, neighborhood 1 is represented by blue color and neighborhood 2 is shown by yellow color. The First quartile for area 2 is overlapping with the medium of data. Area 2 data is concentrated around 4.0 to 3.5 while in neighborhood 1 data is spread across 2.5 to 4.

We follow this by calculating the mean rating for both the neighborhoods, which comes up to be 3.36 for area 1 and 3.68 for area 2. The median for area 1 and area 2 comes out to be same i.e. around 3.50. The standard deviation for each of the ratings in the neighborhood 1 and 2 which is about 0.949 and 0.562.

From the problem we assume the null hypothesis that the mean is equal amongst the two data samples, for testing this this assumption we conduct a t-test. The p-value comes out to be 0.154 which suggests that it is 15.4% likely that our null hypothesis is true. This high percentage indicates the acceptance of our null hypothesis that the difference between the ratings for neighborhoods is close to none i.e. zero. Therefore we accept the null hypothesis.

We use Gibbs sampling to model the difference between the mean of the ratings in the neighborhood. Gibbs sampling model is represented as shown in fig. 2

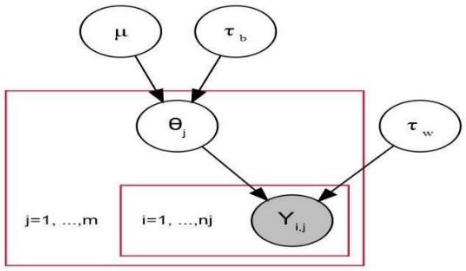

Fig. 2 :Model graph

In which, μ is defined as mean across of total data.

τ is defined as precision across total data.

Γ is defined as variance across all the data.

For fitting the prior parameter in Gibbs sampling, we take our best guess as follows. We have a rating system from 0-5, so a mean would be our best guess for μ would be 2.5 . We assume that $\mu + 2\sigma$ covers the all the data, we take the value of σ to be 1.25. And we know that variance = σ^2 and $\tau = \frac{1}{\sigma^2}$. By using these values, we can approximate the values of a_0 and b_0. This value are calculated with the formula $\frac{a_0}{b_0} = \mu$ and $\frac{a_0}{b_0^2} = \sigma^2$. This initial guess will eventually converge at their original values from data after the burn in period of the sampler. (The practice of removing an initial portion in a MC sample to reduce the effect of primary values on posterior inference).

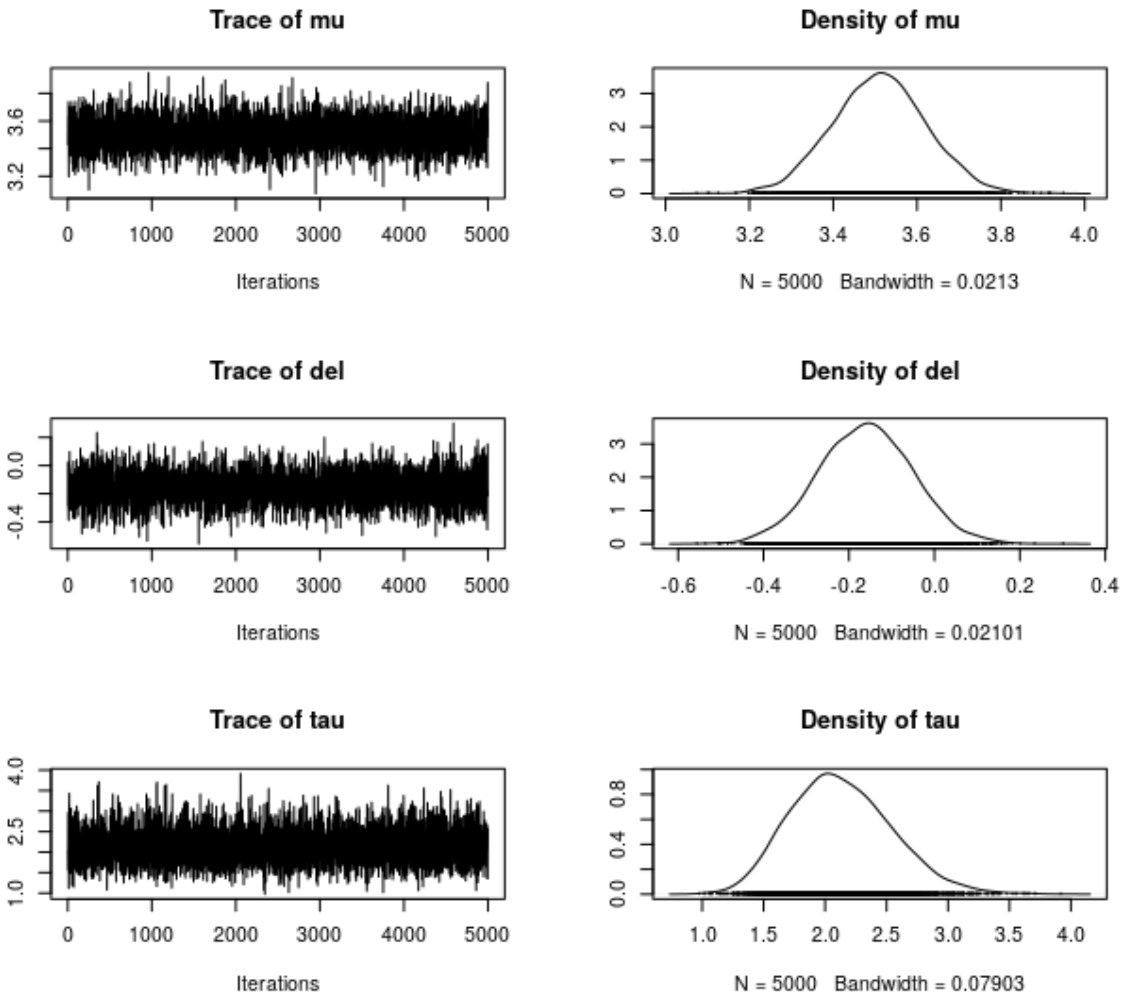

Fig. 3 : Trace and Density plot

Because the prior values can possibly create a high correlation in upcoming values. In practice, we discard this prior value out of the model. To avoid this correlation in prior values we toss values at even intervals. This process of throwing values at regular intervals is known as thinning.

In Fig.3 we can observe the values for $\tau, \delta \; and \; \mu$. All this value follows a normal distribution.

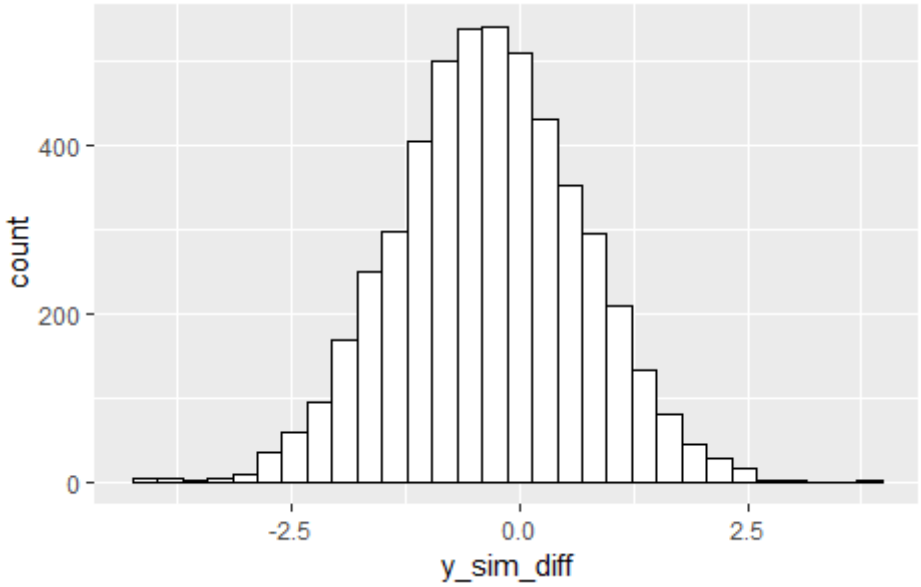

Fig. 4: Distribution for simulated data

In Fig. 4 above we clearly see the posterior values follow a normal distribution.

We again generate 5000 more samples for each neighborhood using Gibbs samplings posterior parameters. These samples help us in checking whether the mean of one area is better than other or no. Each sample of one area is checked against each sample of other neighborhood. This generates a probability which tells us what area is better than other and by what factor.

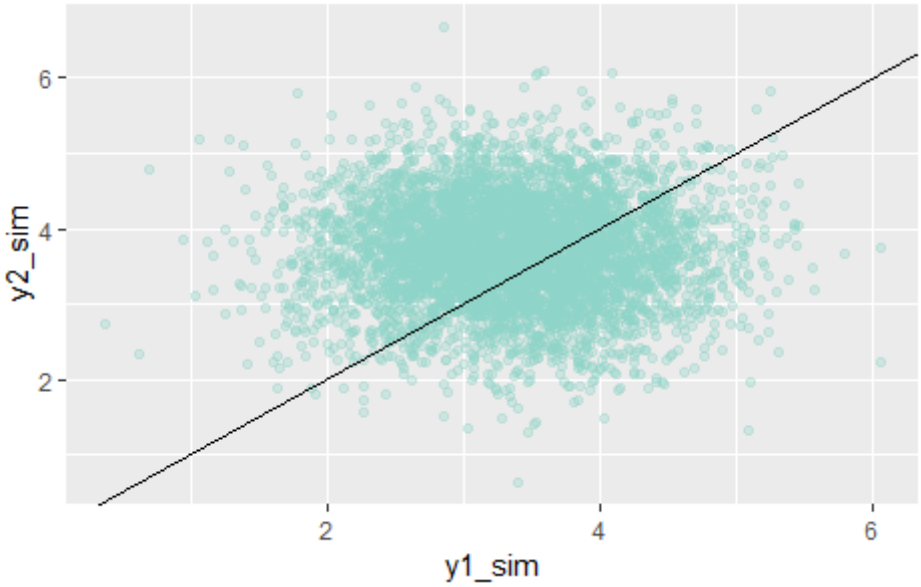

Fig. 5: scatter plot for simulated sample

Fig.5 shows the scatter plot of the two simulations that were generated. Conclusively the probability that 2nd simulation is better than 1st simulation is **0.626** as calculated by piecewise comparison.

To answer the second part of the question, we generate a new data frame with information about all the neighborhood (barring the records for which the neighborhood is not known and the records for which the

neighborhood as an only single business). We observer that total exclusive neighborhood in the dataset comes to be 72.

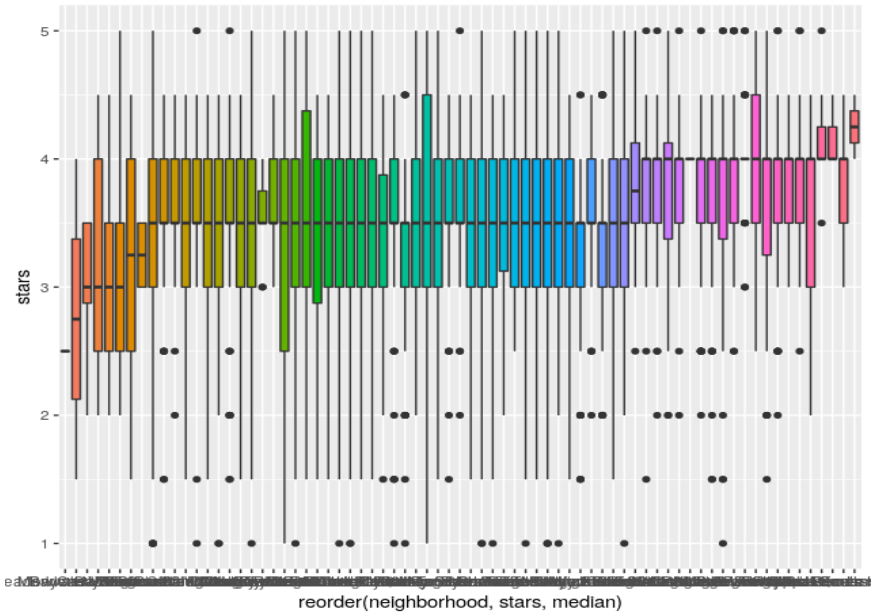

Fig. 6: Box plot for mean rating

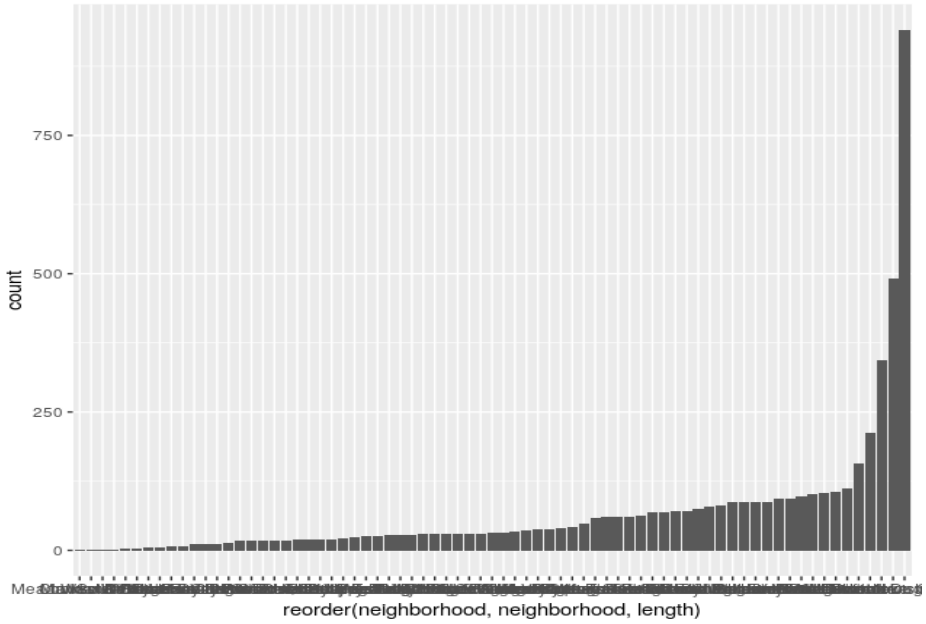

Fig. 7: Review count or each neighborhood.

Fig. 6 shows the data distribution assembled by a distinct neighborhood. It compares the ratings between neighborhoods. We clearly see the outliers in the graph, each box in the graph represents a unique neighborhood. Fig. 7 shows the count of businesses relevant to all the different neighborhood.

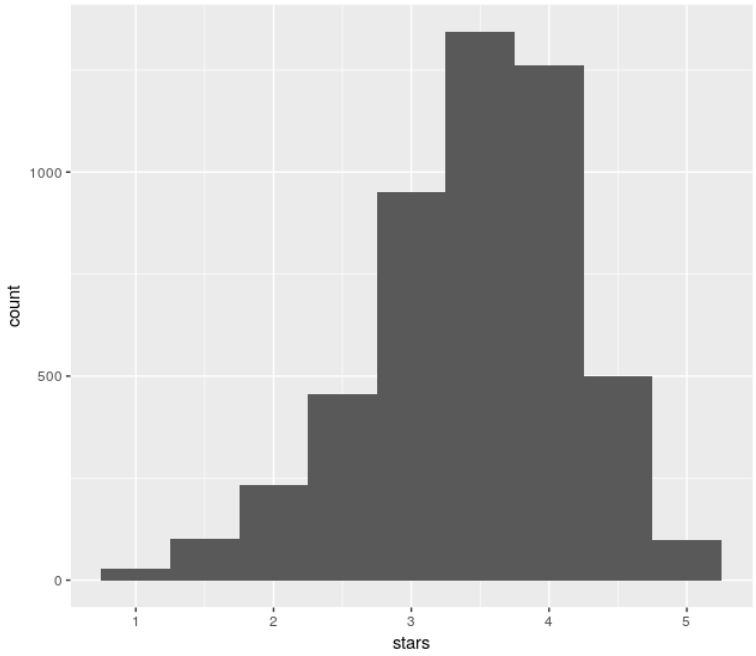

Fig. 8 Distribution of stars

Fig. 8 above shows the distribution of ratings of restaurants with the frequency of occurrence.

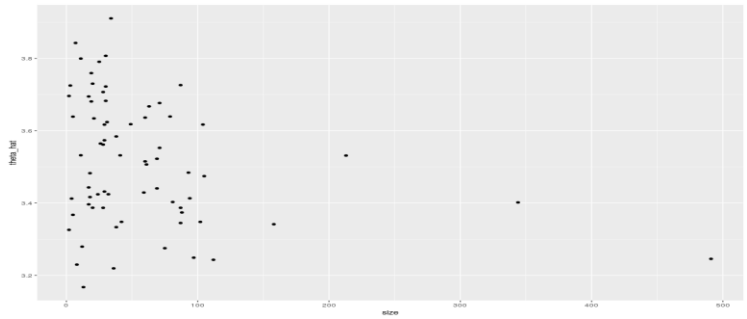

Fig. 9: $\hat{\theta}$ for each neighborhood

The above Fig. 9 shows the mean of all the theta values for a neighborhood. This is done using Gibbs sampling again for each neighborhood and each row. We do this to compare different neighborhood and discover if it is better in terms of ratings.

Conclusion:

From all of the above analysis, we can conclusively say that Scarborough neighborhood is better by 31% to the Etobicoke neighborhood. Calculated by the probability of occurrence of Scarborough minus that of Etobicoke i.e. 0.6266 – 0.3734 =31%. Thus, concluding that the Indian restaurants in Scarborough neighborhood have better ratings.

However, from graphs 8, we can clearly see that south hill have better ratings. However, we see that the number of reviews in the area are little (Fig. 9).

Thus, we cannot conclusively say which neighborhood is better. As $\hat{\theta}$ suggests that best restaurants are fa unction of review counts and star rating both.

Question 2:

What are the factors most strongly associated with restaurants being closed? How accurately can you predict when a restaurant in the dataset will be closed?

We can also consider this as a problem to find variables which contribute highest for predicting the **is_open** attribute in the dataset. Is_open is a binary attribute which identifies if the restraint is open or close.

Processing Data:

First of all, we identify the variables that will have no effect on the prediction of our target variable. They are listed below:

- Business ID – unique key
- Address – Address of the place
- Postal Code – postal code of the area
- Name – the name of the place
- City – city of the area
- Latitude – latitude of the place
- Longitude – longitude of the place
- State – state of the place.

All of the above variables (except business_id and Name) identify the neighborhood of the place, and for our analysis, we have kept the neighborhood variable hence these are not required for the analysis.

We have the dataset from two files one is for information on businesses and other is for information about the reviews.

Business data – most of the fields in this data have some relation with our target variable.

Review data – to capture most of the information about reviews, we add one more field in the data for the usefulness of each review (<= 4 people voted the review as useful). This choice was made after constricting the below graph in Fig. 10 which suggests that the review with at least votes have enough data samples for our analysis.

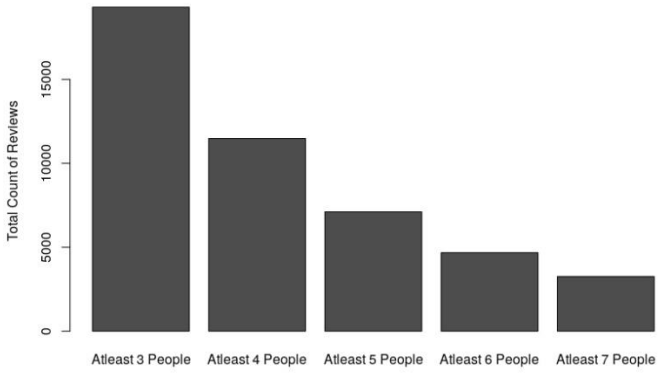

Fig. 10: people found the review useful count

Next, we flatten the attributes columns, which in turn creates multiple dummy columns with value 0 where that particular column is false and 1 otherwise. This gives us a total of 94 columns, from which we discard the once

who more than 75% values are missing. This is a purely intuitive choice which ensures that only columns with less than 25% missing values are used. This leaves us with a total of 38 columns for analysis.

Now we also flatten the categories column in a similar fashion. This leads to column count becoming 359. We then check the top 10 most popular categories in the column. As seen below in Fig. 11 the most popular category is a restaurant, which is irrelevant since we assuming every data point in the dataset a restaurant in some sense. So, we discard the category restaurant.

	cat_total	Freq
1	Restaurants	7051
2	Food	1455
3	Nightlife	936
4	Bars	903
5	***Canadian (New)***	***627***
6	Sandwiches	605
7	Breakfast & Brunch	601
8	Italian	559
9	Chinese	548
10	***Cafes***	***537***

Fig. 11: Word frequency

Modeling:

We create multiple models to predict our target variable **is_open**, which ensures we get maximum accuracy.

The first model we create is using logistic regression. The imputation approach we take is to drop all the rows with value as NA, which reduces our number of observations from 7051 to 3593 i.e. almost 51%. We then feed this data to a logistic regression model. We achieve a model accuracy of 79% with this approach. Accuracy is calculated by using the confusion matrix.

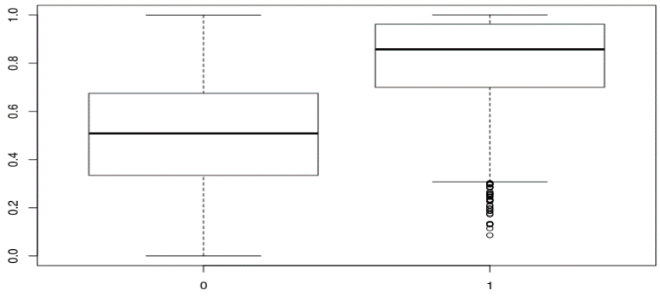

Fig. 12: NA values dropped regression model

The above Fig. 12 shows the box-plot of our logistic regression model with all the values of NA dropped. The plot shows the difference between calculated values of 0 and 1, the difference in the distribution of the two boxes shows the difference in predicted values for 0 and 1. Also, some outliers are visible from the box-plot.

For our next model, we choose a linear regression model. However we do not predict the target variable, instead is used to calculate the p-value, which in turn gives us the dependency of other input variables. All the significant

columns (i.e. for a p-value < 0.05) are selected and then fed into a new model. This new model is a Logistic regression model again. This time we have the column count as 106 significant columns. Therefor there is a considerable reduction (around 73%) in the number of columns used to create model previously. With the accuracy of 76%, this is model gives considerably fewer input columns.

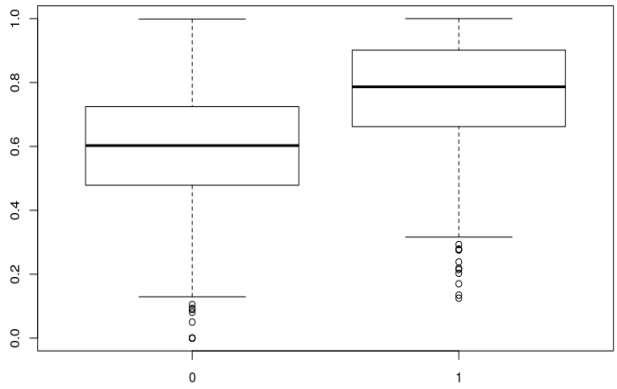

Fig.13: NA values dropped and feature selected regression model

The above Fig.13 shows the box plot of the previously discussed logistic regression model. With feature selection, the boxplot suggests the difference between the prediction value of both level 0 and 1.

The third model we implement is logistic regression. however, instead of discarding records with NA values, we mean impute them (i.e. replace the NA values with the overall mean of the variable). For the factor column, we use mode values to impute the data. After feeding this data to the logistic regression model we calculate the accuracy the model delivers. The accuracy comes to be around 75%.

For our fourth and last model, we user lasso logistic regression model. Comparing previous models we can see that the highest accuracy was given by discarding the NA values of the data. Therefor we choose the same this time.

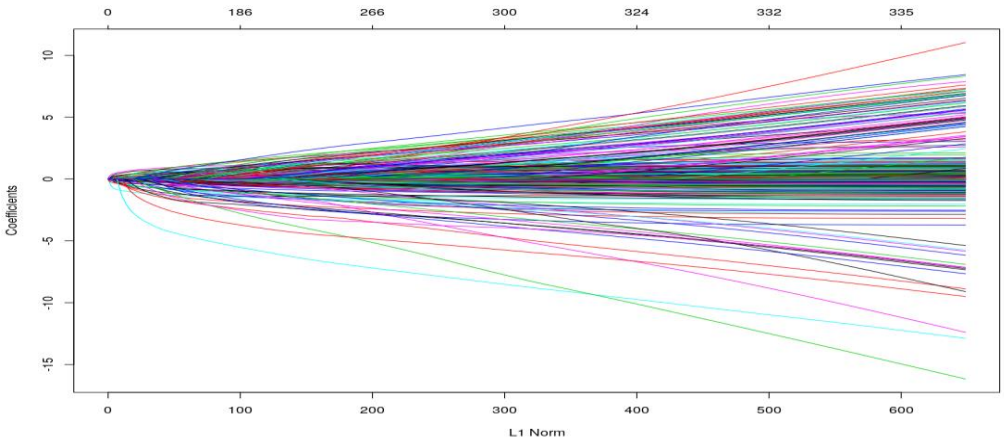

Fig. 14: Lasso regression

The above Fig. 14 shows lasso regression, it can be seen from the graph that no. of columns is too high i.e 359 the plot becomes difficult to read. Above figure shows the coefficient path of a fitted “glmnet” objects.

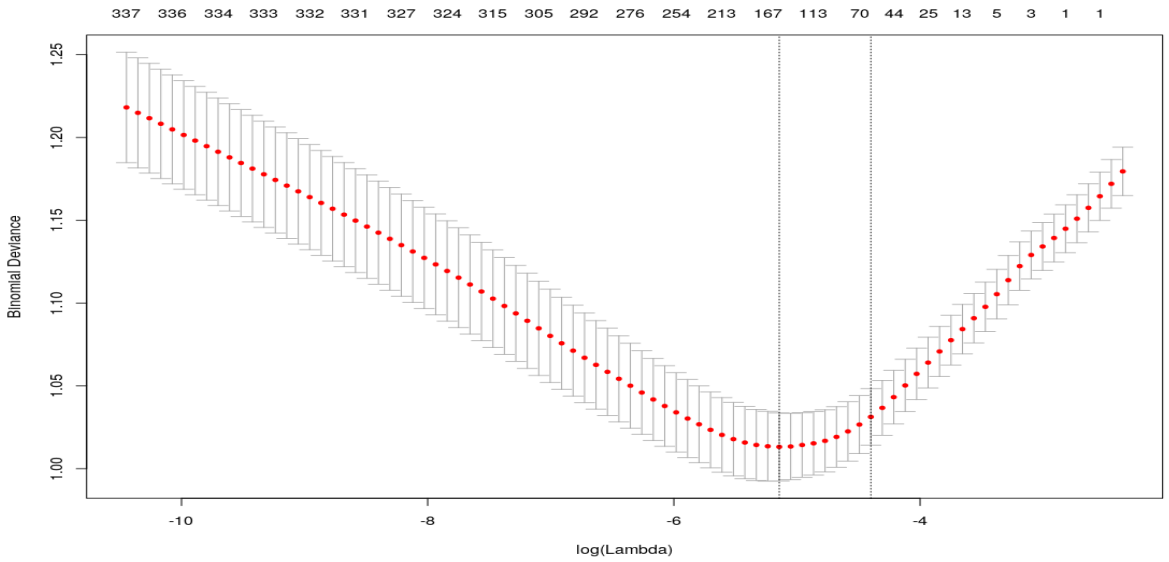

Fig. 15: Cross validation

The above Fig. 15 shows the plot for K-fold cross-validation in lasso regression. We can see the two values of lambda λ for which the error of the model is lowest. These two values are used to minimize the error function and also to avoid the overfitting problem.

Conclusion:

	Model with NA Dropped(without feature selection)	Model with NA Dropped(with feature selection)	Model with mean imputation	Model with Lasso Regression
1	0.797940439743947	0.76148065683273	0.75492837895334	0.778458112997495

Fig. 16: Model comparison

The above table compares all the accuracy values achieved by four of our models. As we see from the table the maximum accuracy was given by discarding the NA columns and not specifying the significant columns. And the lease accuracy was achieved by the model in which we used mean imputation.

Therefor we conclude that the columns most tightly connected to the output variable is_open are the once with p-value is most significant(<0.01) :

- neighborhoodKensington Market
- neighborhoodPalmerston
- neighborhoodQueen Street West
- neighborhoodWest Queen West
- review_count
- attributes.NoiseLevelquiet

- attributes.RestaurantsTableServiceTRUE
- attributes.BusinessParking.streetTRUE
- attributes.GoodForMeal.lunchTRUE
- attributes.GoodForMeal.breakfastTRUE
- Italian
- Fast Food
- Sushi Bars
- Lounges
- Portuguese
- Delicatessen
- Local Flavor

Thus we can say that by using all of the above columns we can predict if the restaurant is closed or not with 79% of the accuracy.

Question 3:

Classification of neighborhoods based on common factors, clustering based on specific factors.

Restaurants are organized based on neighborhoods. Visualizing the data, we see that there are a lot of neighborhoods in the datasets. We utilize the lat/lng of each restaurant to see and organize all the restaurants.

Data handling:

The **business.json** is first filtered to find the data only for Toronto. We use this data to find an association between latitude and longitude. Since these are the only two columns in the data used for the association, we run a clustering model on it to try to cluster the neighborhood into multiple parts.

Model building:

We use a model-based approach for clustering. The R package *mclust* uses finite normal mixture modeling for clustering, classification, and density estimation [1]. It provides a function from the EM algorithm for normal mixture model with a variety of covariance structures. Bayesian Information Criterion (BIC) in wide-ranging approaches for clustering. The higher score of BIC gives us comparatively better score. Once we get the highest BIC value, we finalize the clusters. We will try multiple BIC values to get the best results.

Execution:

First, we find the largest BIC value according to the model building. We try different models for the same. As we can see from the fig.17 the result for parameter G from 0:10. Additionally, we also try parameter G values from 11:20 as visible in fig.18

We see from both figure's the VVV model is the best model in terms of the highest likelihood, however, it needs a larger number of parameters. To select the optimal cluster, we consider 13 as our optimal cluster value as the values are at a peak on 13 and decrease thereafter.

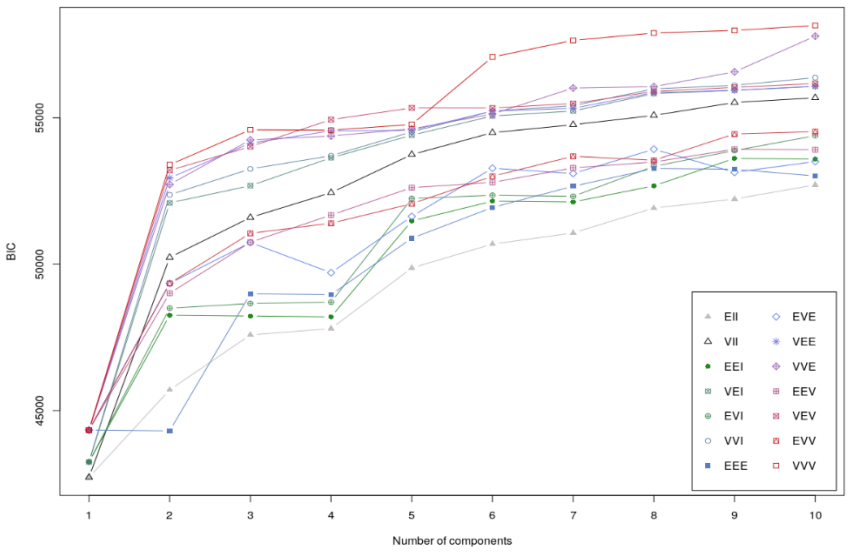

Fig. 17: mclust plot for G = 1:10

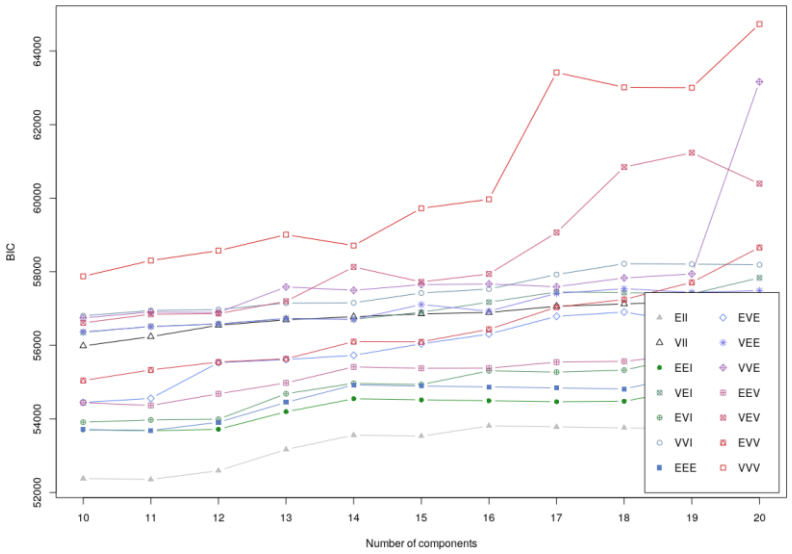

Fig. 18: mclust plot with G = 11:20

As seen from the below scatter plot, without any clustering, the data (lat/lng)looks scattered

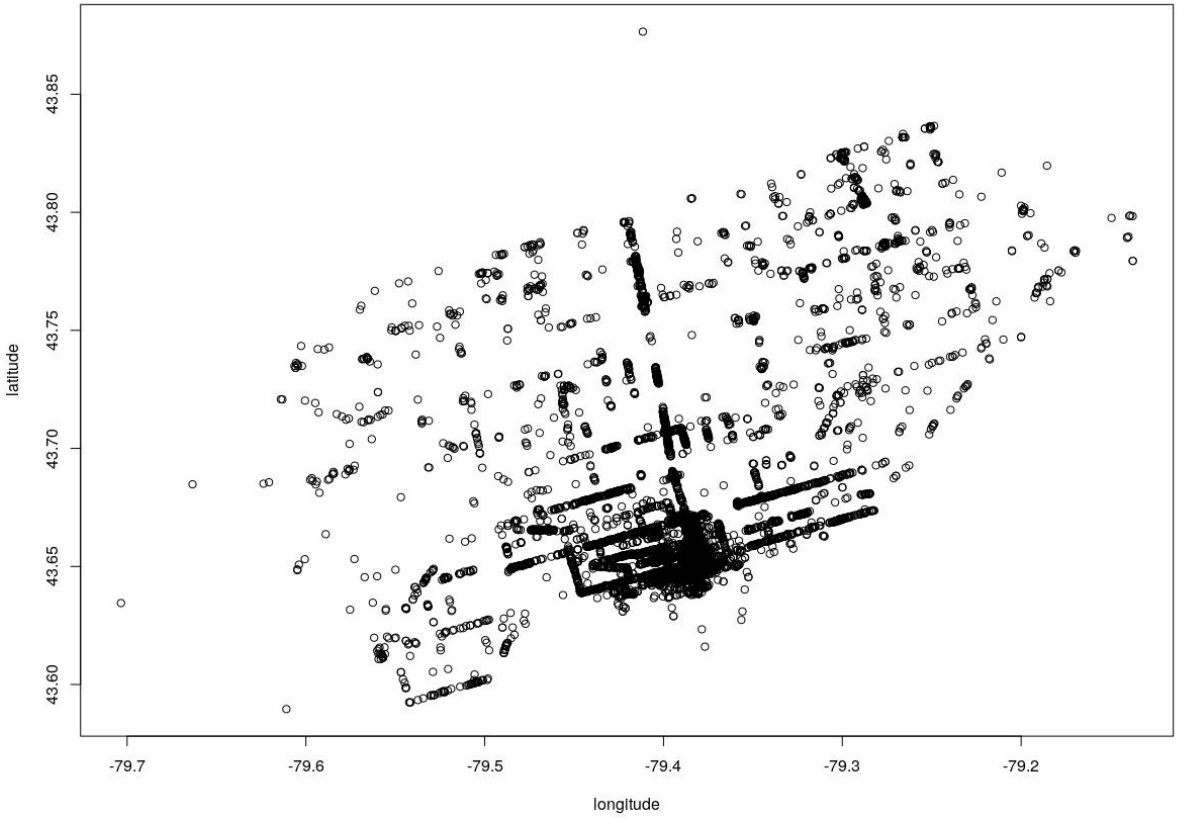

Fig. 19: uncertainty plot

Now that we have finalized the cluster value to 13, we divide the data (lat/lng) into 13 clusters. Fig.20 shows the plot for all the 13 clusters.

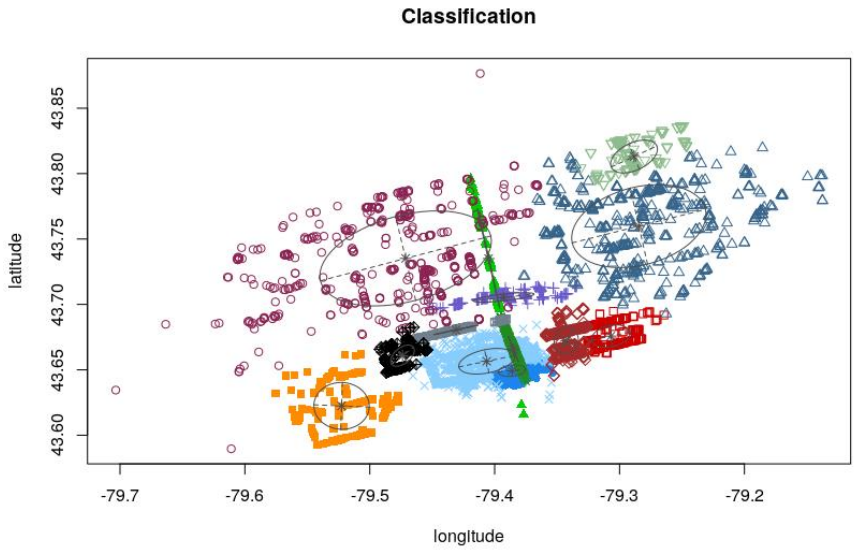

Fig. 20: Clustered neighborhood

Can you find any interesting associations with other elements of the data using this clustering?

By comparing the mean and standard deviation we can better identify the association between different clusters. As from the previous cluster data we start analyzing the association between the below variables.

- Star mean rating
- Review Count rating

Comparing Star ratings:

Below figure fig. 21 depicts the relationship between the star rating and the clusters. We clearly see from barograph that there is no relation between both variables. Also, the standard deviation of the data doesn't give any conclusive result.

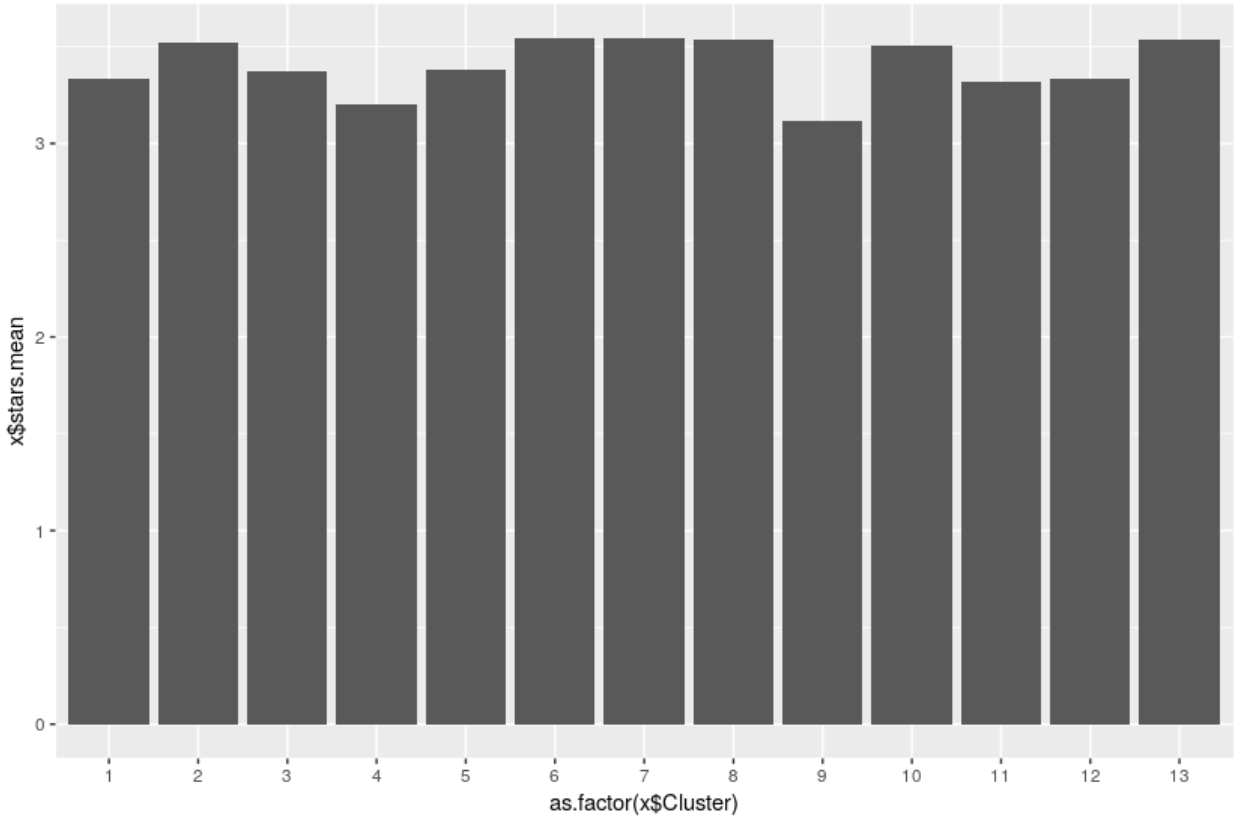

Fig. 21: Mean of star ratings in clusters

Comparing review counts:

Now, we compare the mean of review count rating against the newly created clusters. We see from Fig. 22 that cluster 3 has the greatest number of reviews.

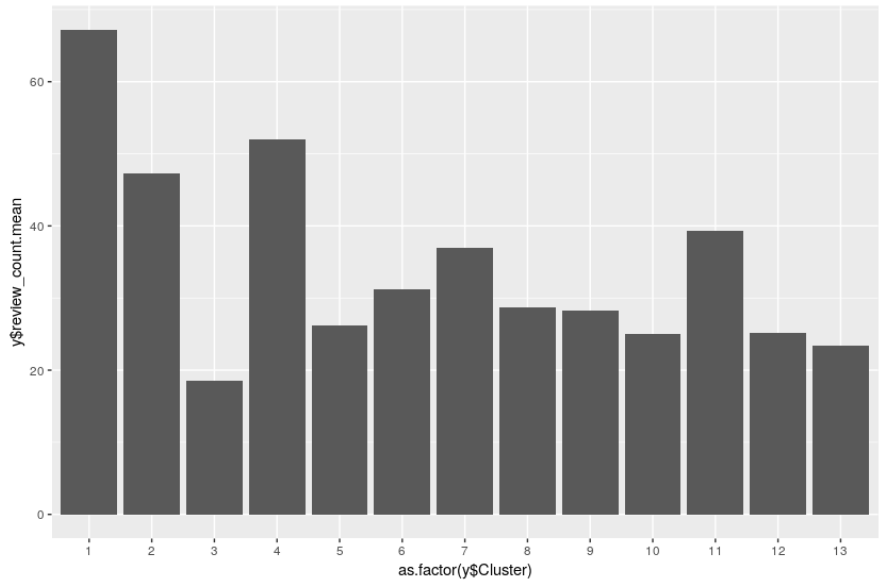

Fig. 22: Mean of review count in clusters

We assume the hypothesis that the cluster with the highest number of reviews will have the highest number of restaurants comparative to other clusters of restaurants.

We test this hypothesis by plotting the rows of reviews with the cluster count. As seen from fig.23 below cluster 3 seems to have the highest number of review count.

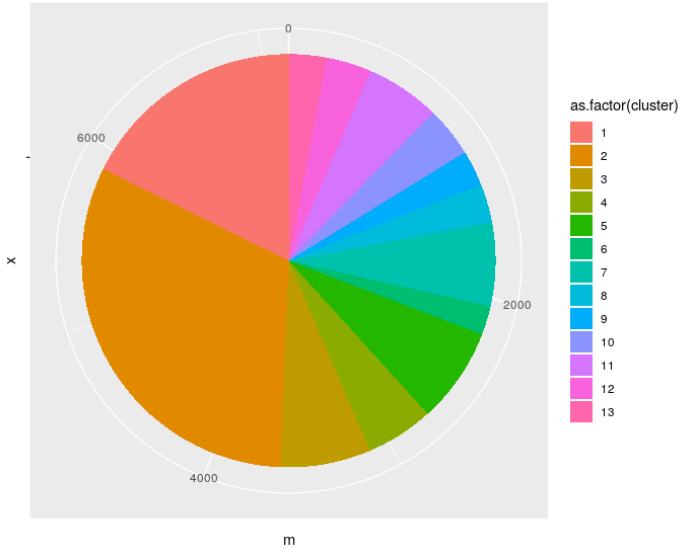

Fig. 23: Size of each cluster.

This proves our hypothesis wrong, as most of the restaurants fall in cluster 1 (ref Fig. 22).

Conclusion:

We have used the mclust package in R to divide the data for the restaurants in 13 clusters based on the lat/lng location information.

Next, we tried to find some kind of association between the clusters and review count or star ratings. Based on our analysis there was no suitable association found between them.

We found that there may be some association between the clusters and the review count, however, this can be due to the size of the cluster i.e. high mean review count because the cluster is larger.