Applied Statistical Modelling

MAIN ASSIGNMENT

Sethuram Ramalinga Reddy
17311993

Yelp Dataset

Introduction:

The yelp dataset is obtained from the company called yelp that provides crowd sourced reviews for the local business and also helps them how to respond to their reviews that would result in improving their business. This dataset is divided into six sections and they give detailed information on Business, User, Reviews, Check-in, Tip and Photos. The Business dataset gives details regarding the type of business, name of the business, address, reviews, business attributes etc. and this dataset is used primarily for the analysis because it overall idea about the business and also the response given by the users to that business.

Data Examination:

From the dataset, we could find that the data is not clean, and it cannot be used straight away for the analysis. So, the pre-processing of data is necessary to proceed with the investigation.

Data Pre-processing:

The data for analysis is obtained from the main "Business.json" file. The data is filtered by selection of the following criteria's. They are:

- The City should be "Toronto".
- The line of Business should be "Restaurants".
- That Business should be currently running, and this is ensured by selecting "is_open == 1".
- The neighborhood and the other fields shouldn't be empty.

The resulted file from the previous process is in the form of JSON and this is imported into the R terminal using "JSONlite" package.

Missing Values:

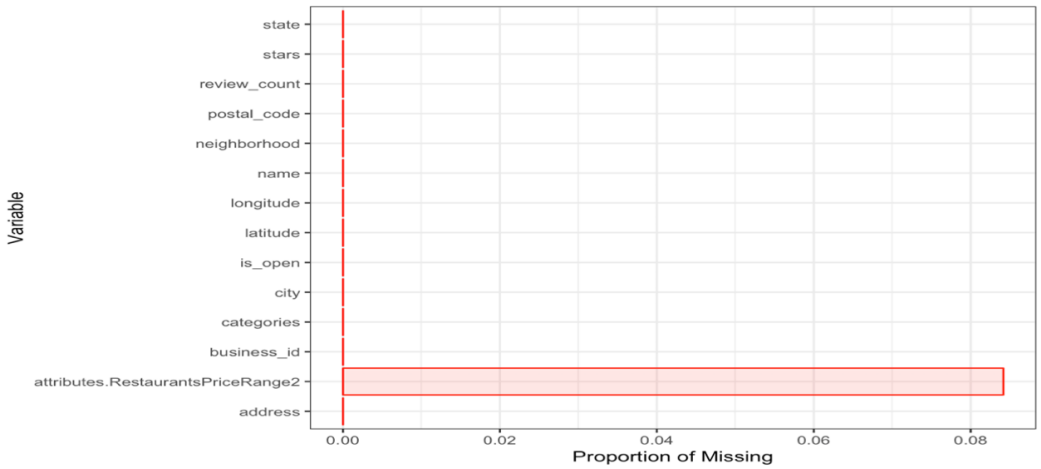


Figure 1 Proportion of Missing Values

As you can see from the plot, this dataset is cleaned, and there are no missing values in most of the columns except the "ResturantPriceRange2" column. This column has got missing value percentage of only 8%, and this doesn't affect our analysis since the percentage is very low. If need, missing value imputations can be done which is nothing but imputing the most repeated value to the missing cases.

Data Exploration:

To have proper visualization, all the neighborhoods are replaced with unique values, and their relationship with the stars given by the user is shown below.

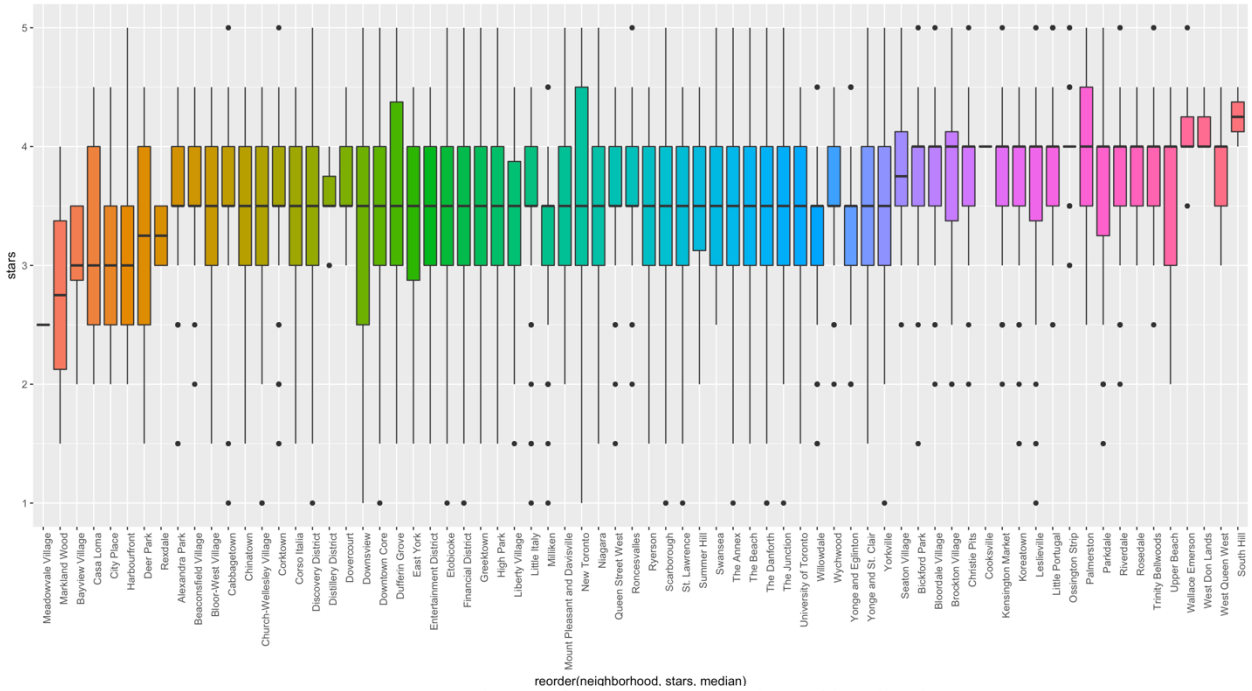


Figure 2 Distribution of Star Ratings in each Neighbourhood

In the above plot, we can see the distribution of star ratings for every neighbourhood. In some of the neighbourhood's the distribution is skewed either left or right which indicates the presence of a large number of outliers on either side and results in introducing bias on the mean ratings given by the users for that neighbourhood. In most of the neighbourhood's, the star rating has got high variability, and for some of them, the distribution is very small due to the presence of less number of ratings given by the user which might also indicate the presence of less number of restaurants in that neighbourhood. We can also see the median ratings of the user increases as we go along the neighbourhoods. From this plot, we cannot find the number of restaurants in each neighbourhood. The above plot also indicates that the average median for about 60% of the neighbourhoods is 3.5 which indicates that in Toronto most of the restaurant business has got good ratings and makes it a good place to start the restaurant business. This analysis also paves the way for finding the reason behind these good ratings in the Toronto for the restaurant business.

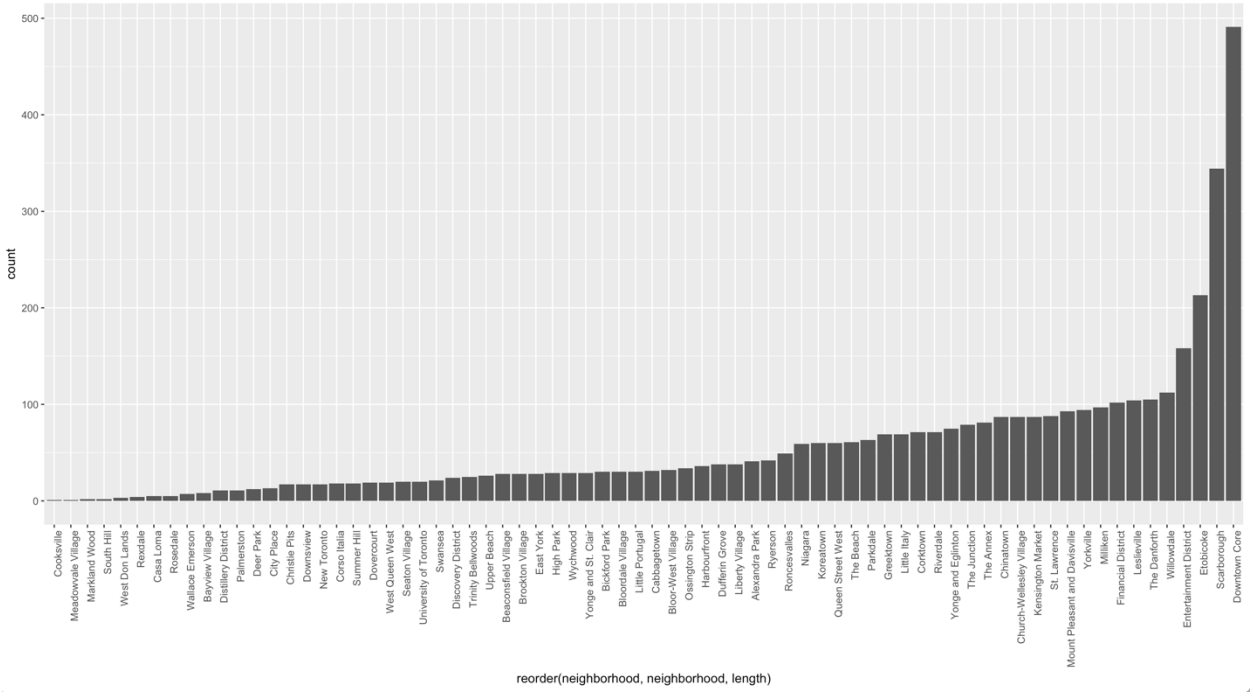


Figure 3 Number of Restaurants in each Neighbourhood

This plot gives the count of the number of restaurants for each neighborhood. From this we can see that some of the neighborhood has very less number of restaurants and some of them has got very large number of restaurants. We can remove the neighborhoods which has got very few restaurants since this doesn't contribute much to the analysis.

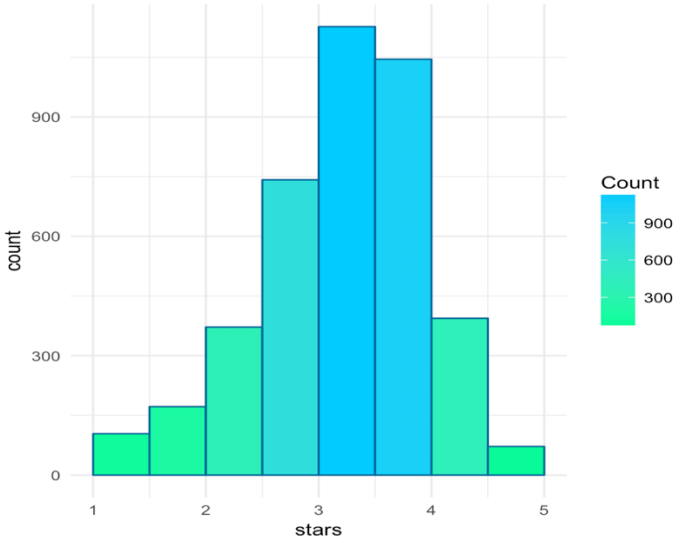


Figure 4 Distribution of Star Ratings

In this plot, we can see the distribution of star ratings of the restaurant in for all the neighbourhoods.

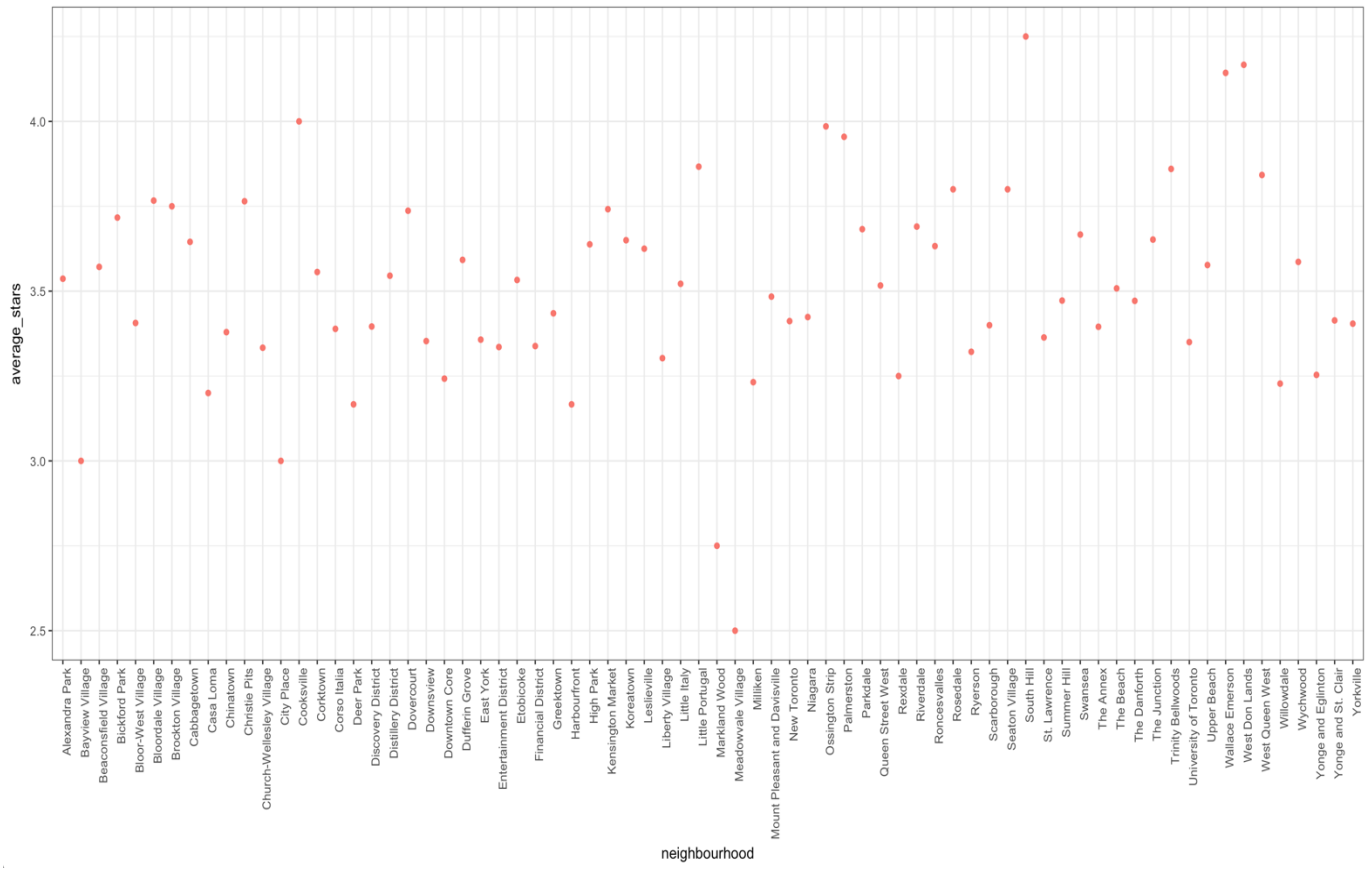


Figure 5 Average Star Ratings in each Neighbourhood

In the above plot we can see that the average star rating in each neighborhood and most of the ratings are between 3 and 3.5. This plot shows that the star rating is not influenced by the number of restaurants. For E.g. if we consider the Downtown Center and Scarborough, the number of restaurants is very high when compared to the other neighborhoods, but the star rating is less such as 3.25 and 3.3 respectively. We can also infer that most of the neighborhood has the average star rating between 3.25 and 3.75 and this indicates the overall restaurant business in the Toronto is good. But this result derived might be subjected to variation because of the variability present in the samples. For E.g. the user rating given for that restaurant will vary for different days. So, we have to incorporate this difference in order to build a proper model.

Question-1:

To compare the ratings of the different neighbourhood, the Hierarchical Model is considered. As stated earlier the variability between the observed data is incorporated into the model, and this model has got hierarchical levels to estimate the hyperparameters of the posterior distribution. This model is used when the observed data is present in several different units. A pictorial representation of these model is given below:

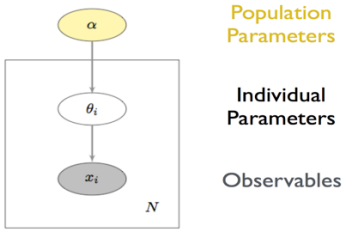


Figure 6 Hierarchical Model

From the above diagram, we can see clearly how each observed variable as got their parameters and the overall population parameters. The Bayes Equation for this model is given as below:

$$p(\alpha, \theta|x) \quad \propto \quad p(x|\theta, \alpha)p(\theta|\alpha) \quad p(\alpha)$$

Posterior likelihood Prior

In total, the hierarchical model quantifies the uncertainty in the hyperparameters and also deals with the probabilistic relation between the theory and the observations. This model always brings down the individual data estimates to the population mean and reduces the overall Root Mean Squared Error. This hierarchical structure for yelp business dataset is represented graphically as shown below:

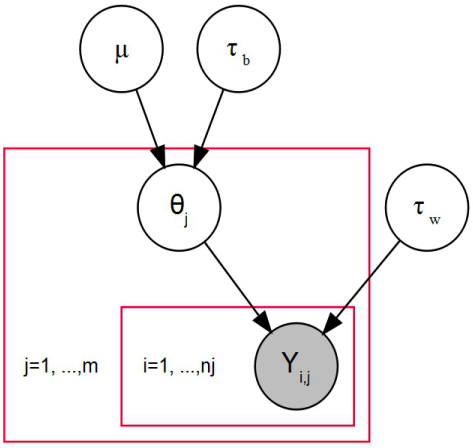


Figure 7 Bayesian Heierarchical Model

For the yelp business dataset, we have to find the ratings of different neighborhood and for this purpose we construct a hierarchical model with j = 1, . . ., m neighborhoods such that the average mean within the group be $\theta_j \sim N(\mu, 1/\tau_b)$ where μ is the overall mean and τ_b is the precision between the neighborhoods and also with i = 1, . . . , n_j : $y_{i,j} \sim N(\theta_j, 1/\tau_w)$. where i is

the number of ratings for each neighborhood and τ_w is the precision within the neighborhood. The total variability has been divided into two components. They are given below:

- **variability between the neighborhoods:** This is represented by the parameter $1/\tau_b$ which indicates the variance between the mean ratings given by the user for all the neighborhoods and also with the overall mean.
- **variability within the neighborhoods:** This is represented by the parameter $1/\tau_w$ which indicates the variance between the user ratings in each neighborhood.

Assumptions for model:

- We should make sure that the observations are exchangeable which means the ordering of each rating doesn't have an impact on the analysis.
- An exchangeable property should be established even at the group level.
- We assume a common precision τ_w within the groups.
- We assume only the mean performance of the group was normally distributed and nor the individuals.

The difference between the mean ratings of neighbourhoods can be estimated using a technique called Gibbs Sampling. This sampling technique is based on Markov Chain Monte Carlo algorithm, and this draws samples from a probability distribution when each sample is dependent on the previous sample. This method is used to convert a multi-dimensional problem to a lower dimensional problem by dividing the entire vectors in the multi-dimensional space into smaller sub vectors where each sub-vector is derived from its distribution based on the current sub-vector.

As discussed earlier to compare the means of star ratings between different neighbourhoods, the Gibbs sampler is used. The following prior distribution is assumed for the hyperparameters that are presented in the hierarchical model as shown above.

$$\mu \sim N(\mu_n, 1/\tau_n)$$
$$\tau_b \sim \text{Gamma}(\eta_0, \nu_0)$$
$$\tau_w \sim \text{Gamma}(a_0, b_0)$$

This distribution is chosen because the posterior distribution is also the same as their prior since these are conjugate priors. For, E.g. the posterior for the normal and gamma distribution will also be a normal and gamma distribution respectively. Since we know that the data y is independent of the overall mean and also the precision between the groups which makes the sampling easy. For the hyperparameters, weak informative priors are chosen which gives equal probability to all the values in the distribution to be sampled. Now Markov chain Monte Carlo algorithm is initiated with the chosen hyperparameters, and the new values for the population mean (μ) from the normal distribution, precision (τb & τw) from the gamma distribution is sampled from each iteration of the chain.

If we assume that most of the restaurants in all neighbourhoods of the Toronto have a star rating of 3.2 on an average from the previous analysis which means all the restaurants are good and the precision between them can be lower, so it is assumed to be a gamma distribution with low rate and shape parameters. The precision within the groups should be higher when compared to

the precision between the groups so this is also assumed to be gamma distribution with relatively high shape and rate parameters, then the results after the Gibbs sampling is obtained as follows:

```
> apply(fit2$params, 2, mean)
       mu     tau_w     tau_b
3.513018 1.899505 4.859433
> apply(fit2$params, 2, sd)
        mu       tau_w       tau_b
0.05943704 0.04227196 0.87373494
> mean(1/sqrt(fit2$params[, 3]))
[1] 0.4592797
> sd(1/sqrt(fit2$params[, 3]))
[1] 0.04257837
```

We can see from the calibrated results that the population mean is over 3.5 rating and the variance between the group is less compared to variance within the groups. The average ratings in each neighbourhood are shown below:

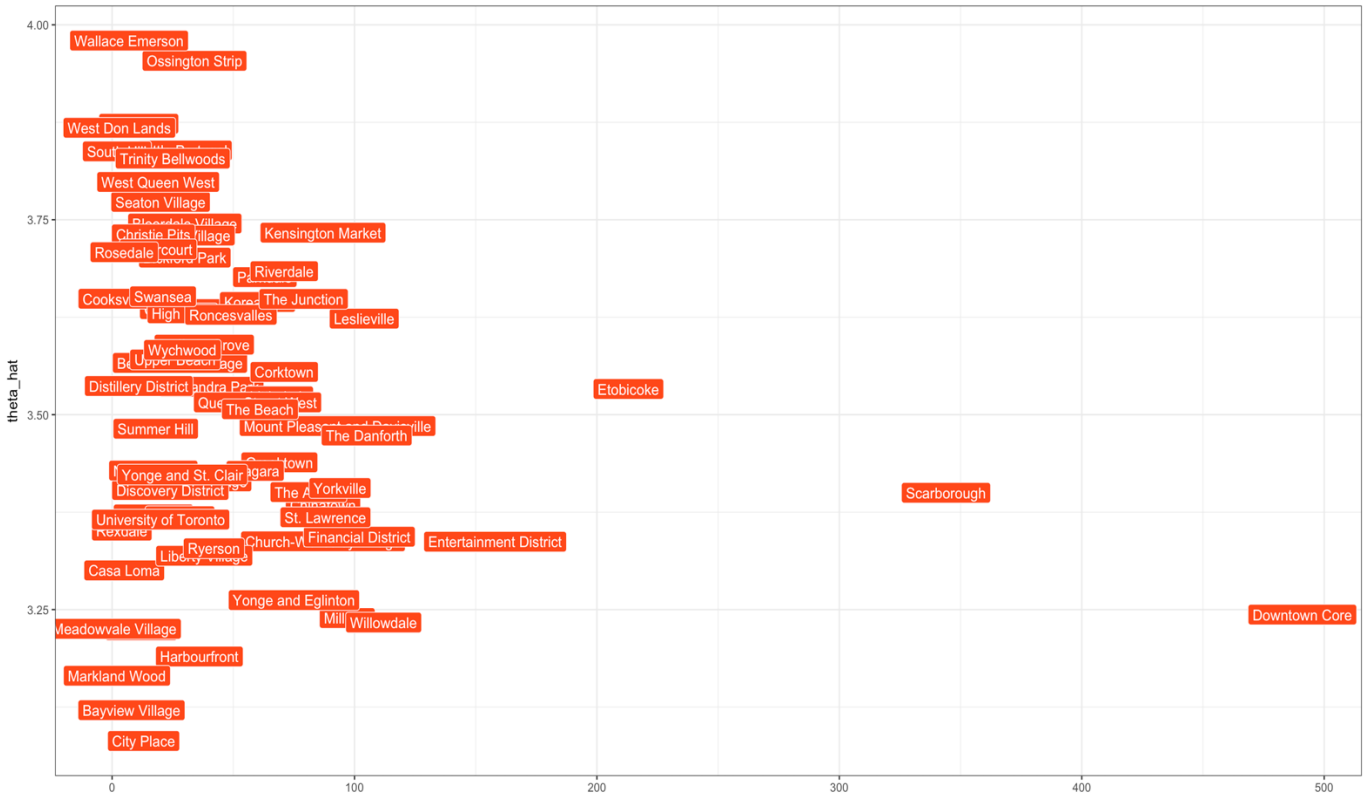


Figure 8 Average Star Ratings for each Neighbourhood VS Number of Restaurants

From the above plot, we can see that the neighbourhood is having 50 restaurants have an average rating between 3 and 4 and the distribution is also dense. The mean ratings are also high for these number of restaurants. When the number of restaurants is between 50 and 100, the average rating reduces between 3.25 and 3.75. Then as the number of restaurants increases the average rating also gets decreased. This is evident from the neighbourhood Downtown Core and Scarborough. From this, we can also infer that number of restaurants in a neighbourhood doesn't play a major factor in analyzing the neighbourhood rating.

Validating the sampler Performance:

The start of the chain has to be known exactly so that we can eliminate the variables selected during the burn-in period which doesn't require the condition we specified and also to make sure thinning has happened which means highly correlated successive samples are removed. This validation can be established by using the MCMC pack and also using raftery.diag method.

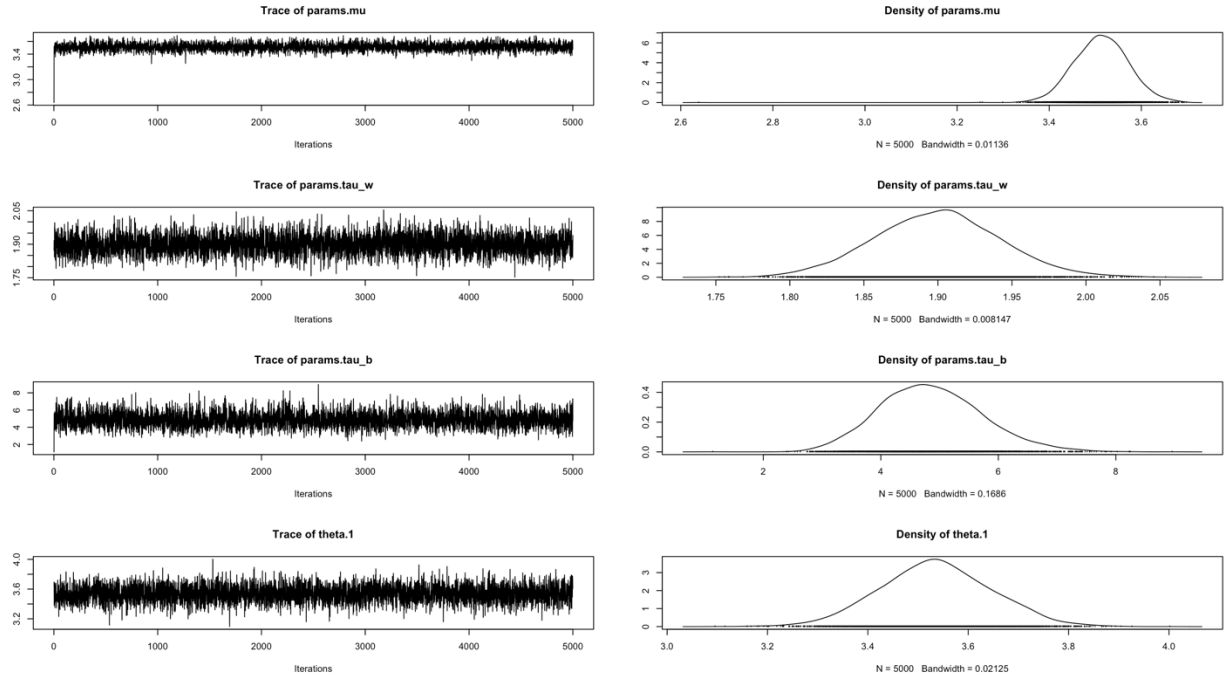


Figure 9 Trace plots

In the above plot, we could see the traces of the parameter mu, tau_w, tau_b and also the mean average of the first neighbourhood. We could also see their Density respectively. The densities of precision parameters (tau_w & tau_m), overall mean and also the mean average of the first neighbourhood follows a Normal distribution. The results from the rafter run diagnostic package are shown below:

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

            Burn-in  Total  Lower bound  Dependence
            (M)      (N)    (Nmin)       factor (I)
params.mu   2        3964   3746         1.060
params.tau_w 2       3561   3746         0.951
params.tau_b 2       3653   3746         0.975
theta.1     2        3680   3746         0.982
theta.2     2        3680   3746         0.982
theta.3     2        3620   3746         0.966
theta.4     2        3680   3746         0.982
theta.5     2        3869   3746         1.030
```

From the above result we could see that the number of burn-in iteration to be avoided is very less (It's two iterations). We can also verify that the thinning has happened since the total number of samples taken is less than the number of iterations, and also it states the minimum number of samples required to derive samples avoiding the correlation. The number of samples required increases in there is the positive autocorrelation. If we analyze the Dependency factor which indicates the auto correlation is all less that assures the good choice of hyper parameters.

Conclusion:

The ratings of the neighbourhoods are densely populated between 3.25 and 3.8. There are some neighbourhoods such as Wallace Emerson and Ossington Strip whose ratings are almost close to 4 and these are superior to others. If we consider the mean rating of 3.5 then these neighbourhood are superior by 0.5

Question-2:

To predict the restaurant rating the business and review datasets are combined in such a way that the reviews for that particular business are obtained. The reviews given by each user is quantified by using sentiment analysis, and the sentiment score and magnitude are calculated. This sentiment analysis is done using Google Cloud API which generates the score and magnitude. Here score means sentiment polarity of the review whether it is positive or negative review and magnitude indicates the intensity of the sentiment. This sentiment analysis on the reviews given by the user is the critical factor that influences the ratings of the restaurant in the business context. Now we build a model to find the relationship between the ratings obtained and the reviews given by the user for that business.

The model selected is the Multiple Linear regression assuming the dependent variable ratings(stars) to be a quantitative variable ranging from 0 to 5, and the restaurant rating is given by the average of all the individual user ratings.

Multiple linear regression is a particular case of linear regression model where we consider more than one independent variables for finding its relationship with the dependent variable by assuming that there is no correlation between the noise generated from each variable. The general form is given below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i.$$

The Regression model in Bayesian approach is depicted as follows:

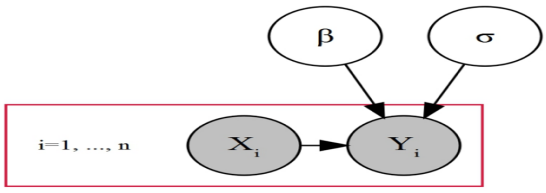


Figure 10 Bayesian Linear Model

The independent variables that are considered for building the model to predict ratings are given below:

1. Ratings given by each user
2. Useful
3. Funny
4. Cool
5. Review Count
6. Sentiment Score
7. Sentiment Magnitude
8. Price Range
9. Neighborhood
10. Rating of that business

The other independent variables such as city, state, postal code, latitude, longitude etc. are not considered because they don't give any sense in terms of business perspective.

Feature Selection by using standard correlation plot:

Now we have to find the variables that has mutual correlation between them so that we can avoid the repetitive information. The correlation gives the strength of the relationship between the variables along with its direction of the relationship. For this purpose, we use Pearson's correlation to determine the linear relationship between independent variables and the result is viewed using correlation plot.

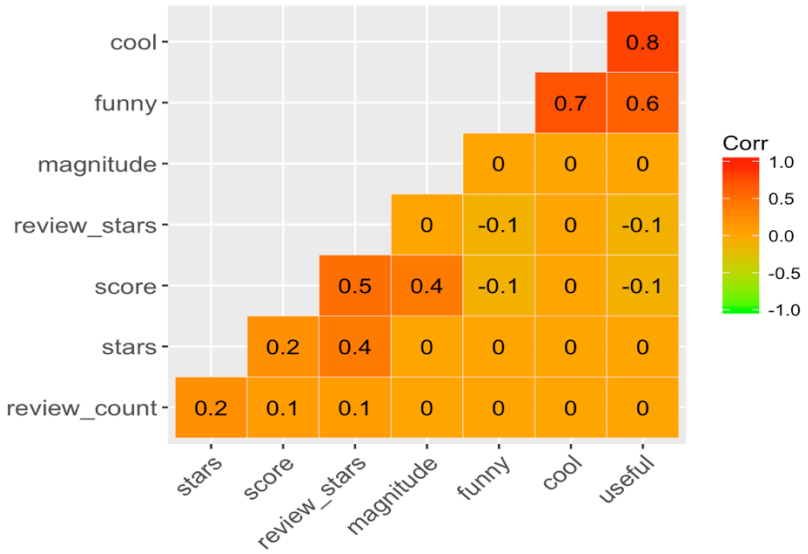


Figure 11 Correlation Plot

In the above figure color of the box represents the direction of the correlation (Red represents positive correlation, and Green represents negative correlation). The negative correlation means the independent variable has the linear relationship in the negative direction. The positive correlation means the independent variable has the linear relationship in the positive direction. The number represents the magnitude (Strength) of the correlation. If the quantity is 0.5

to 0.6, then it has the moderate correlation, and if the value is higher than 0.6, then it has a strong correlation.

Based on the correlation plot obtained above there is the high positive correlation between the cool vs useful and cool vs funny by taking the threshold for correlation as 0.7 which indicates the repetition of same information so that we can remove this variable. The variables considered for building the model based on the above analysis are given below:

1. Ratings given by each user
2. Useful
3. Funny
4. Review Count
5. Sentiment Score
6. Sentiment Magnitude
7. Price Range
8. Neighborhood
9. Rating of that business

Feature Selection by AIC and BIC:

The previous approach was based on correlation, and there is another approach for selecting the features by using a statistic called AIC (Akaike Information Criterion). This creates different models by removing each feature and estimating the AIC for that model. This is a statistic for estimating the performance of the model relative to the other models. It is used to do the tradeoff between the goodness of fit and simplicity of the model. Here goodness of fit is assessed through likelihood function. The model with low AIC statistic can be selected. To avoid overfitting to the model, we impose a penalty (k = log(n)), and this is referred to as BIC. The AIC approach is particularly used for prediction, and the BIC is used for understanding the data.

The linear model is created, and stepwise regression is implemented through step function and without giving any penalty. This will give the AIC static for each model. From the results, we choose the model with least AIC value. For the same model BIC (Bayesian Information Criterion) is determined by giving the penalty to the step function. As a result, overfitting is avoided.

In this scenario, AIC is chosen because we have to predict ratings of the restaurant and the AIC is the best suitable static for prediction. The features selected by the AIC model is given below:

```
> print(step_AIC)

Call:
lm(formula = stars ~ funny + magnitude + score + review_stars +
    useful + attributes.RestaurantsPriceRange2 + neighborhood +
    review_count, data = df_bus_rev_subset_model_selection)
```

We can see that the model chosen by both the correlation and AIC approach are same. This ensures that the correctness of the features selected for the building the linear model.

Now with the decided predictor variables, we build the linear regression model in the Bayesian way. This is represented in terms of probabilistic model as given below:

$$\mu_i = \alpha + \beta x_i$$

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

This means that the response variable follows a normal distribution with standard deviation σ and then mean μi which is a linear function of X parameterized by α and β. In this case, the optimal value is founded using the maximum likelihood estimation. Here the error term should be independent and identical normally distributed random variables (static noise).

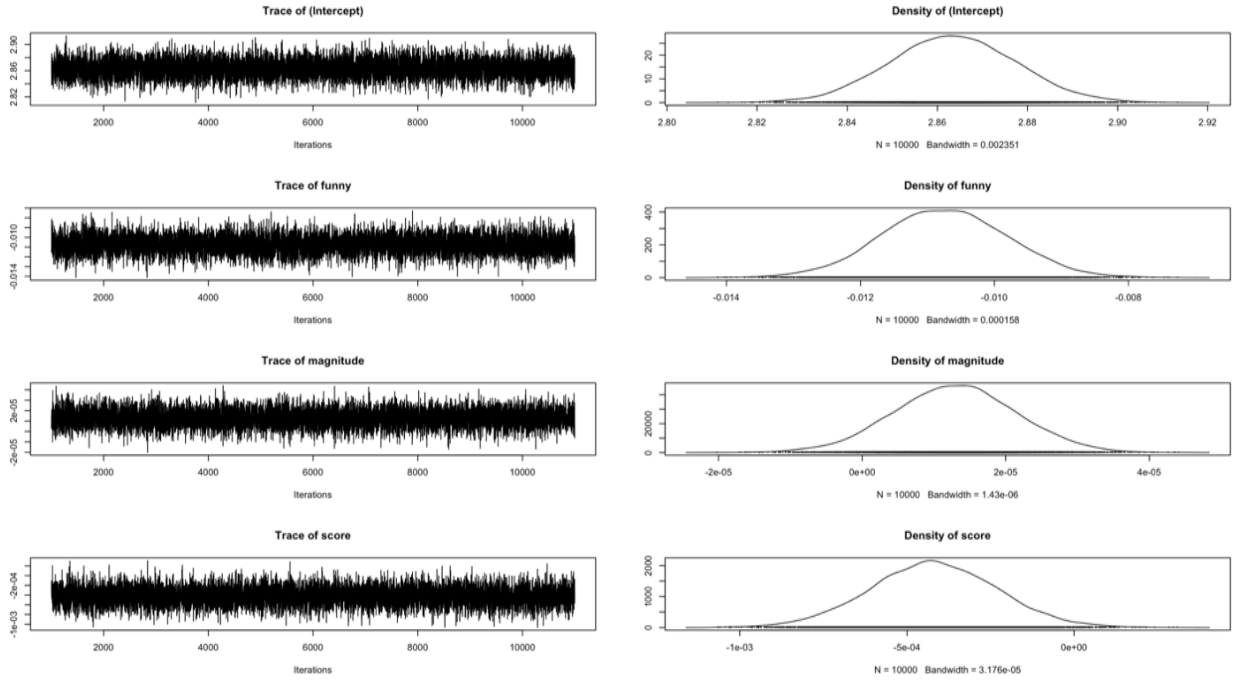
Now we use MCMCregress[1] function for creating the samples from the posterior distribution of the linear regression along with the Gaussian errors using Gibbs sampling which was explained earlier. The number of iterations after the burn-in period is set to 1000 and the result is obtained in the form of mcmc object. The summary of the derived model is given below:

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

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                                        Mean       SD  Naive SE Time-series SE
(Intercept)                         2.863e+00 1.400e-02 1.400e-04      1.400e-04
funny                              -1.075e-02 9.404e-04 9.404e-06      9.404e-06
magnitude                           1.291e-05 8.534e-06 8.534e-08      8.388e-08
score                              -4.211e-04 1.891e-04 1.891e-06      1.891e-06
review_stars                        1.476e-01 9.263e-04 9.263e-06      9.500e-06
useful                              8.477e-03 6.009e-04 6.009e-06      5.913e-06
attributes.RestaurantsPriceRange21.0 2.635e-01 1.119e-02 1.119e-04      1.151e-04
attributes.RestaurantsPriceRange22.0 1.512e-01 1.103e-02 1.103e-04      1.134e-04
attributes.RestaurantsPriceRange23.0 3.299e-01 1.141e-02 1.141e-04      1.141e-04
attributes.RestaurantsPriceRange24.0 3.486e-01 1.310e-02 1.310e-04      1.310e-04
neighborhood2                      -4.593e-01 2.482e-02 2.482e-04      2.482e-04
```

The obtained model is plotted as below:



[1] https://www.rdocumentation.org/packages/MCMCpack/versions/1.4-2/topics/MCMCregress

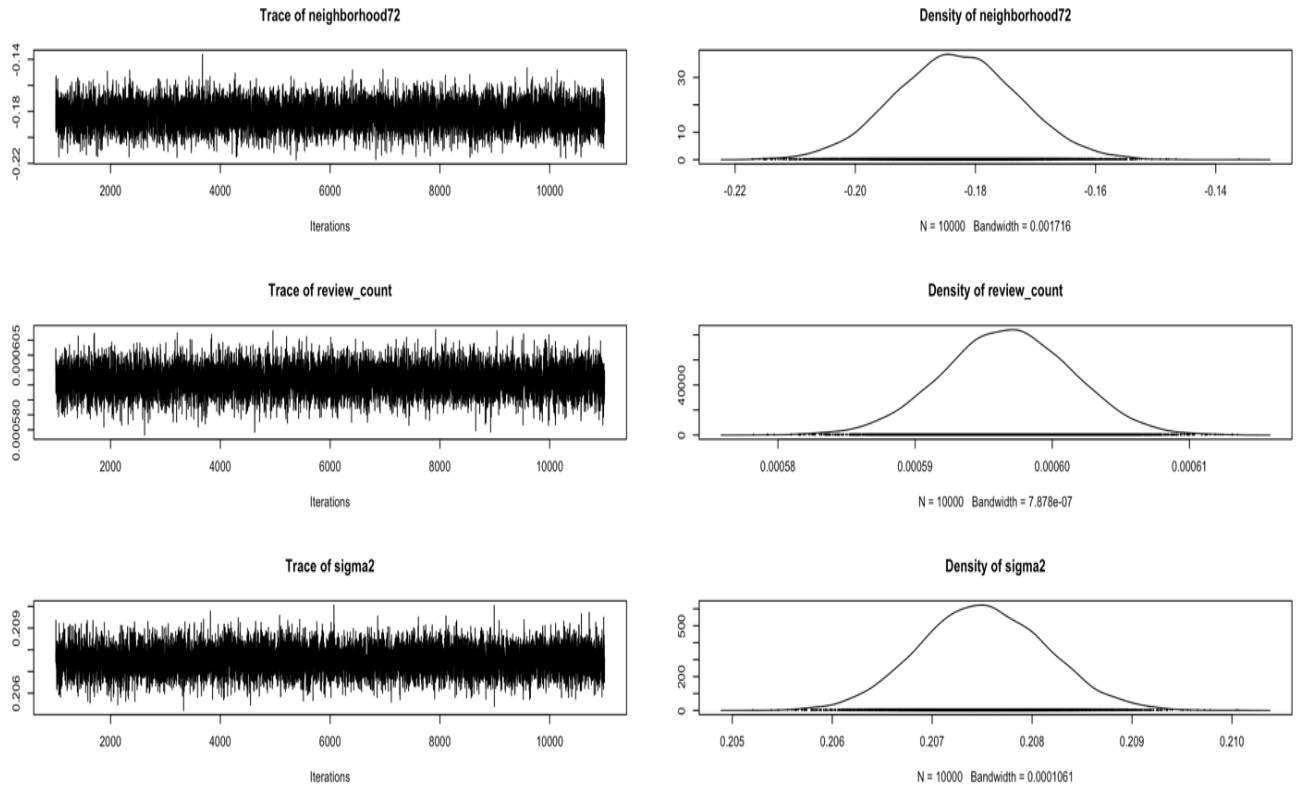


Figure 12 Trace plots for Linear Model

From the above plot, we can see that the static noise(sigma2) follows a normal distribution which indicates the correctness of the model. Now we find the mean of all the coefficients obtained from the model and use those to predict the ratings for the test dataset.

```
> beta_mean
                                    beta_mean
(Intercept)                      2.863408e+00
funny                           -1.074975e-02
magnitude                        1.290633e-05
score                           -4.210695e-04
review_stars                     1.475590e-01
useful                           8.477045e-03
attributes.RestaurantsPriceRange21.0  2.635349e-01
attributes.RestaurantsPriceRange22.0  1.511900e-01
attributes.RestaurantsPriceRange23.0  3.298865e-01
attributes.RestaurantsPriceRange24.0  3.485569e-01
neighborhood2                   -4.592964e-01
neighborhood3                   -1.709312e-02
neighborhood4                    6.814894e-03
neighborhood5                   -1.251913e-01
```

From the above coefficient values, we could see that the first value of the categorical values such as neighbourhood and attribute.RestaurantsPriceRange is missing. This is because when a dummy variable is created for every level in the factor, the resulting set of variables will be linearly dependent along with the intercept term. Therefore, one level is considered to be baseline, and other coefficients are calculated based on this baseline. As a result the coefficients of neighbourhood1 and attribute.RestaurantsPriceRange2.0 will be 0, and these variables are removed from the result. In the list of predictor variable that we have chosen contains two categorical variables neighbourhood and attribute.RestaurantsPriceRange, so the intercept is closer to the mean of intercept of both the selected categories. The resulted intercept is the standard baseline for all the variables. Now we have to do prediction using the test data which is created by removing the ratings (stars) from the dataset. The prediction is achieved by matrix multiplication of the obtained coefficients with the test dataset. But this process fails because of the two missing categories in the coefficients(beta_mean). This is addressed by introducing two column such as neighborhood1 and the attribute Restaurants PriceRange20.0 with value 0. Now the prediction ratings for the test dataset is obtained. The distribution of the predictions obtained is shown below:

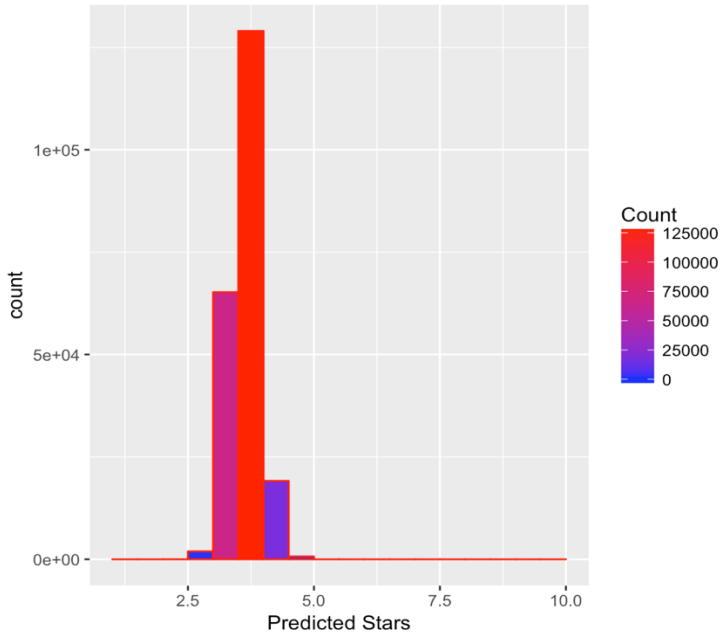


Figure 13 Distribution of Predicted Star Ratings

From the above histogram, we could see that most of the restaurant ratings are predicted between the range of 3 and 4. The performance of the model is analyzed using the Root Mean Square Error (RMSE) which is the difference between the predicted and observed value. In another word, it is the average of the distance of data points from the fitted line measured along the vertical line. The RMSE obtained is **0.45545**. This is small and indicates that the performance of the model is good.

Adding Quadratic Effect:

The quadratic terms are added to transform the linear model to curve without making them nonlinear. If by the addition of quadratic effect, the RMSE value decreases then it indicates that the dataset is curved, and the model is considered to be good fit. If there is increase in RMSE value then the dataset is linear or else if the RMSE value decreases, then the dataset is curved, and our model is not considered to be good fit for this dataset.

The model is built using the features that are used previously for the linear model and the quadratic terms chosen are sentiment score and magnitude since these are the most significant variables that influences the rating. If the sentiment score has the positive effect on the rating and the square of the sentiment score term shows more positive impact, then the square term is highly correlated than the untransformed variable, and this also indicates the presence of curvilinear nature within the dataset. Now we could see the number of predicted star ratings as below:

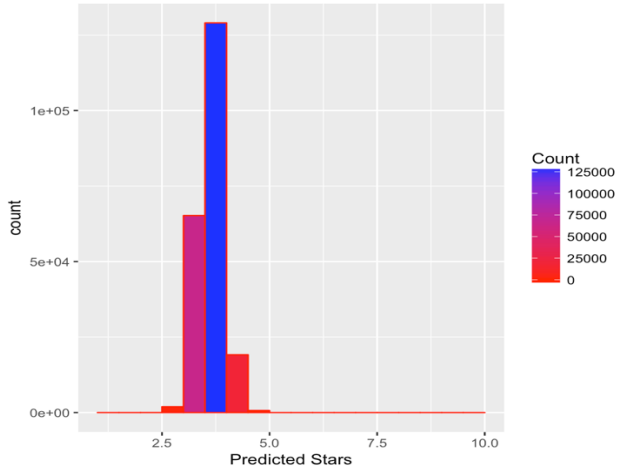


Figure 14 Distribution of Star Ratings with Quadratic Effect

From the above plot, we can see that number of ratings in the range 3.5 to 4 is higher than the model without quadratic effect and also the number of ratings higher than five is very less than the previous model. As a result, the RMSE value is **0.42**. Thus, we can infer that the data is curvilinear and by adding the quadratic terms the model fitness has been increased.

Conclusion:

The variables such as review count, price range, Ratings given by each user, useful, funny, sentiment score and magnitude, and the neighbourhood are most influential in predicting the restaurant ratings. We could see that the RMSE value for the model without the quadratic effect was **0.45** and this is pretty good for the model with less number of observations. When the quadratic effect was introduced on the sentiment score and magnitude variable by considering them as the most critical variable, the RMSE of the model got better and reduced to **0.42**. This indicates that the data was curvilinear data and our model is considered to be appropriate for this data.

Question-3:

The dataset considered for this case study is the "business_open_toronto.json" which has the variables Restaurant categories and neighbourhoods. Now we have to select the top 9 categories that occur in most of the restaurants. The steps to achieve this are given below:

- Split the categories in each business and find the number of the number of business/restaurants in each category.
- Sort the categories in decreasing order of the number of restaurants.
- Select the top nine categories.
- Check whether these categories exist in all the business and encode 1 if the categories are present for that respective category.
- Get all the details of the restaurants by joining the above result with the business dataset using business Id as the joining condition.

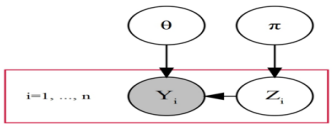
The top nine categories are given below:

```
> cat_names
[1] "Food"               "Nightlife"          "Bars"               "Sandwiches"
[5] "Breakfast & Brunch" "Chinese"            "Canadian (New)"     "Cafes"
[9] "Coffee & Tea"
```

To check the existence of an association between the neighbourhood and categories the Latent Class Analysis model is used.

Latent Class Analysis Model:

Latent class analysis model belongs to a domain of mixture models. The mixture model[2] is a probabilistic model to represent the existence of subpopulation within the overall population, and it is not necessary that the observed data should indicate to which subpopulation that an individual observation must belong. This represents the probability distribution of observations in the overall populations. This model is represented as below:



Here π indicates priors, $\theta = \theta_i \ldots \theta_k$ are the mean parameters and the Z is the latent group indicator

Figure 15 Mixture model

The Latent class analysis[3] is applied on multivariate categorical responses to carry out model-based clustering. In other words, we find the existence of similarities and groupings in the multivariable categorical data. This approach is similar to the Naïve Bayes algorithm in machine learning. This analysis includes the selection of variable step that is considered to be necessary for

[2] https://en.wikipedia.org/wiki/Mixture_model

[3] https://arxiv.org/pdf/1402.6928.pdf

this clustering approach. In this model, the observations that are similar are expected to be grouped based on the variables probability in each group. Finding the number of clusters is difficult to achieve, and this is done by using BIC approach.

The number of clusters are founded by building the latent models with clusters K =1...9 and finding their BIC values respectively. The BIC values obtained are given below:

```
> bic_values
[[1]]
[1] -25231.73

[[2]]
[1] -21797.32

[[3]]
[1] -20276.1

[[4]]
[1] -20171.36

[[5]]
[1] -20122.43

[[6]]
[1] -20195.07

[[7]]
[1] -20168.03

[[8]]
[1] -20206.89

[[9]]
[1] -20270.06
```

plot(x, y, ...)

According to the Kass and Raftery[4], the thumb rule for the difference in BIC is that a difference of less than two is not worth analyzing whereas a difference of ten is considered as strong evidence for the existence of patterns. So, based on the above values obtained there is the significant difference between the model with 1 group, two groups and three groups. Thus, the number of clusters chosen is 3 for building the model. The LCA model is obtained by passing the categories to the blca.em function with K=3 and the result obtained is shown below:

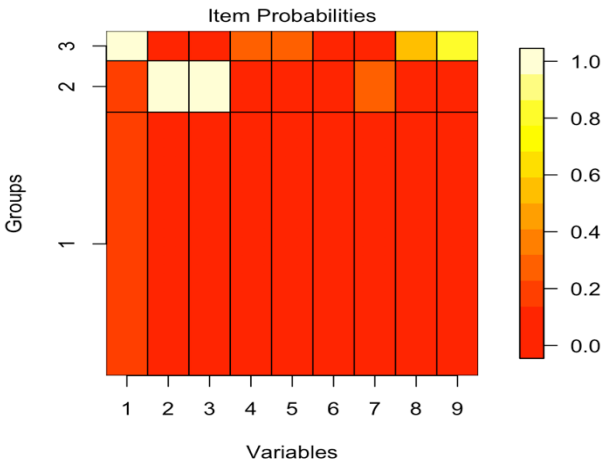


Figure 16 Cluster Representation

```
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 0.576        0.779

Membership Probabilities:

Group 1 Group 2 Group 3
  0.765   0.148   0.087
```

In the above plot, the item probabilities are assigned over a scale of 0 to 1. The Group Probability of Class 1 is very much higher when compared to the remaining groups. The group probabilities are influenced by the number of observations in that cluster. Three groups are taken into account for clustering since this produces the good result based on the BIC values. In Group 1 only the Food category is present with probability 0.2 and others are not present with probability

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2934856/ - R11

0. Similarly, we can find each variable (categories) dominance in each cluster. We can see that all the variables are present in the three groups except the variable six which is the category "Chinese" since the item probability is zero in all the classes so that this variable can be removed. In this way, the feature selection is carried out in the latent class analysis.

The algorithm convergence can also be seen for the LCA model that is constructed and this is given below:

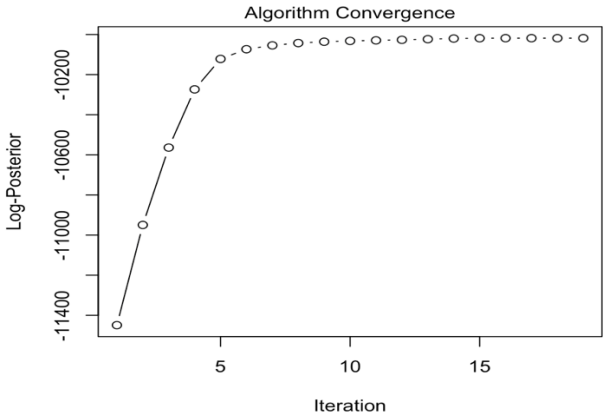


Figure 17 Convergence in each Model

From the above plot, we could see that the maximum posterior has reached a stationary value after ten iterations which means that the algorithm is converged.

Now we know the overall item probabilities in each cluster, and these clusters are considered as separate models. We use Z score function to find cluster probabilities for each data points. By this process, we can see to which cluster each restaurant belong and by what probabilities. Once the z score is known, we need to assign each restaurant to a particular cluster, and this is done by taking maximum likelihood estimation that gives the point estimate from previously calculated z score. Thus, we have grouped the restaurants into different clusters where each cluster has a definite pattern of categories in different probabailities.

To find the association between neighbourhood and categories, we need to separate the observations based on the neighbourhood and within each neighbourhood, we have to find the number of restaurants for different clusters. This is represented graphically as shown below:

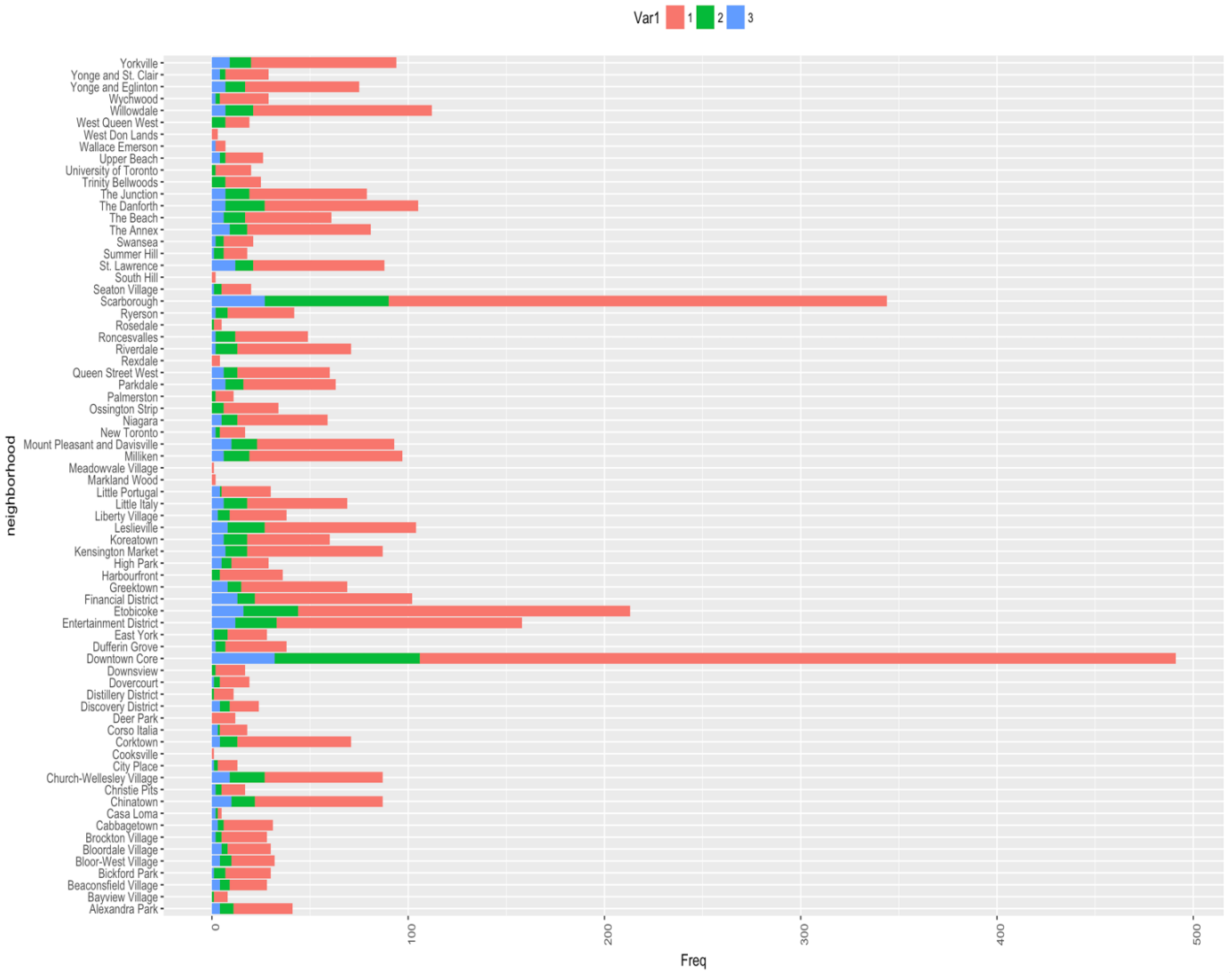


Figure 18 Association between Neighbourhood and Categories

In the above plot, we can see how the restaurants in each neighbourhood are grouped under different clusters. From the plot, we can infer that some neighbourhood contains the specified types of restaurant category more than the others.

Conclusion:

The analysis indicates that there is an association between the neighbourhood and restaurant categories. When compared to the other neighbourhoods, the Downtown Core and Scarborough has got the high number of restaurants that are more likely to have category food than the other eight categories which in turn represents the cluster 1. Similarly, we can also see that the for the same Downtown core and Scarborough the restaurants with categories in cluster 2 and cluster 3 is more than the rest of the neighbourhoods. Thus, we can infer that this neighbourhood has restaurants that are more likely to have these nine categories such as Food, Nightlife, Bars, Sandwiches, Breakfast and Brunch, Chinese, Canadian, Cafes, Coffee and Tea when compared to other neighbourhoods.