# Business review of Toronto Restaurants

By: Saumya Bhatnagar
16338296

Abstract:
In data analytics or business analytics, modelling signifies defining, understanding, analyzing or predicting the organization of the data or business, by using software tools and mathematical techniques, basing assumptions on existing business processes or systems. When we generate data for the same, the methodologies that are used are termed as applied statistical modelling. In this report, we are analyzing the yelp dataset using applied statistical modelling for making inferences which will help Toronto restaurants to generate more business and provide better services.

## Introduction:

This analysis provides various strategies that can be directly applied on Toronto restaurants. The dataset is taken from yelp webpage. The key focus is on restaurant business. In order to make the analysis more applicable on the running businesses of Toronto, only the restaurants that are opened in Toronto are analyzed. The methodology can used applied on other cities as well.

Key problems addressed are: if restaurants in some neighborhoods tend to be more superior than others; the factors responsible higher restaurant ratings; the categories that are more likely to be in one neighborhood than other

## Computer Software and Hardware Architecture:

RAM: 16GB. The details are shown in Table 1

## Data Cleaning

The tables used for merging and cleaning criteria is shown in Table 2. For second problem statement, the three datasets were considered to get more factors in the factor-map.

## R Packages Used:

The same is shown in Table 3.

| Tasks | Tools | Method |
|---|---|---|
| Initial Exploration | Excel | Extracted initial 100 rows |
| Data Merging | R version 3.4.4 | Data merged: business.json, review.json, user.json |
| Sentiment Analysis | Google API | Review.json is uploaded for getting the score and magnitude of the review texts given by the customers |

Table 1: Tools and Techniques for various procedures

| Problem Statement | Datasets | Join key | Filter Criteria |
|---|---|---|---|
| if restaurants in some neighborhoods tend to be more superior than others | Open_business_Toronto.json | | "Business.json" -> City is Toronto, State is Ontario, Open business, Null values are removed |
| the factors responsible higher restaurant ratings | Business.json, review.json, user.json | business + review (on business_id)<br>review + user (on user_id) | |
| the categories that are more likely to be in one neighborhood than other | Open_business_Toronto.json | | |

Table 2: Tools and Techniques for various procedures

| S. No. | Package | Description | Version | R Dependence |
|---|---|---|---|---|
| 1 | Readr | Read Rectangular Text Data | 1.1.1 | ≥ 3.0.2 |
| 2 | jsonlite | A JSON Parser and Generator | 1.5 | > 3.4.3 |
| 3 | VIM | Visualization and Imputation of Missing Values | 4.7.0 | ≥ 3.1.0 |
| 4 | dplyr | Eases manipulation/workings on data | 0.7.4 | ≥ 3.1.2 |
| 5 | ggplot2 | Data Visualisations Using the Grammar of Graphics | 2.2.1 | ≥ 3.1 |
| 6 | MCMCpack | Markov Chain Monte Carlo (MCMC) Package Contains functions to perform Bayesian inference using posterior simulation for a number of statistical models | 1.4-2 | ≥ 2.10.0 |
| 7 | MASS | Support Functions and Datasets for Venables and Ripley's MASS | 7.3-49 | ≥ 3.1.0 |
| 8 | Coda | Output Analysis and Diagnostics for MCMC | 0.19-1 | ≥ 2.14.0 |
| 9 | Mclust | Model based clustering, parameter estimation via EM algorithm, BIC steps | | |
| 10 | BayesLCA | | | |
| 11 | MCMCglmm | | | |

Table 3: R Packages Used

# Problem statement 1: Compare the ratings of different neighborhoods. Are any neighborhoods clearly superior than others?

## Methodologies:

The question that we are trying to solve is:
Do differences exist between two or more groups on one DV?
Clearly, we need to compare the variation in the means of various variables. We consider: Factorial ANOVA and then modeling using Hierarchical models

<u>ANOVA:</u>

Anova is an omnibus to t-tests or Generalized t-test
The hypothesis formulation:

$$H_o: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

$H_a$ : At least one of the means is different from the others

Relation between F-value and t value: $F = t^2$
In this case we use Factorial Anova: multiple independent variables, one dependent variable.
WHY ANOVA?
Gives an exploratory data analysis
Effect of many variables at once
Generates ‘F’ statistics: allows testing of a nested sequence of models
organization of an additive data decomposition, sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
Analysis of a variety of experimental designs.
Handles experimental error
Reduces chances of Type 1 error
The more statistical tests run, the greater likelihood that the researcher will obtain seemingly significant effects due to chance alone. (ANOVA determines whether the amount of variance between the groups is greater than the variance within the groups)
ANOVA is computed with the three sums of squares
Total – Total Sum of Squares, a measure of all variations in the dependent variable, SST
Treatment (Between) – Sum of Squares Treatments (Between), SSC
Error (Within) – Sum of Squares of Errors; yields the variations within treatments (or columns), SSE

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$$

$$SSC = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x})^2$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

SST = SSC + SSE

$$MST = \frac{SST}{df(SST)} = \frac{SST}{N-1},$$

$$MSC = \frac{SSC}{df(SSC)} = \frac{SSC}{k-1}.$$

$$MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{N-k},$$

$$F = \frac{MSC}{MSE}.$$

$$F > F_{1-\alpha, k-1, N-k},$$

<u>HIERARCHICAL MODELING</u>: It is a Bayesian statistical modeling of hierarchical structure, where using prior parameters posteriors are estimated

# Model Understanding

## What is Hierarchical Modelling?

Hierarchical Modelling is used to compare the means of various population means

## What is Gibbs Sampler?

It is a MCMC algorithm for Bayesian inference. Gibbs sampler is a technique for generating random variables from a distribution indirectly, without having to calculate the density.
We make the following assumptions,
The initial x0~N(μ0, V0),
The covariance matrix Σ and Γ are known, and
Given F, the distribution xt is Gaussian.
The Gibbs sampling through m replications of the i iterations produces i iid k tuples,
Z1j(i): Zkj(i), j=1, 2, 3, …, m,
which the proposed density estimates for [Zs] having form

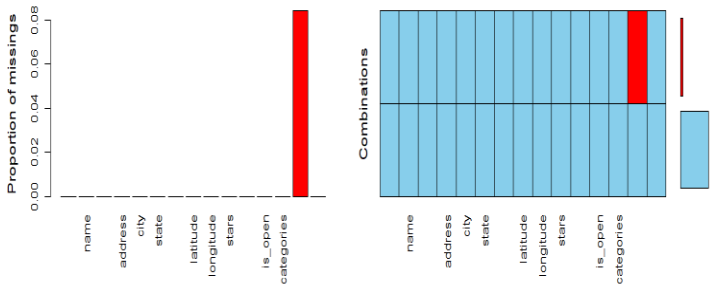
$$\left[\hat{Z}_s\right] = \frac{1}{m} \sum_{j=1}^{m} \left[ Z_s \middle| Z_r^{(j)}, r \neq s \right]$$

## Model Assumptions

- Conditional independence assumption:
  - individual observations are assumed to be independent.
  - Similarly, at group level, groups are assumed to be all exchangeable
- We assume a common within group variance/precision parameter τ across all groups.
- Normal distribution:
  - mean of the groups are normally distributed, and
  - conditional on group membership, individual observations are normally distributed.

# Results:

1) Checking for missing values, see Figure 1 and 2.
   a. Since, there is high proportion of missing values, we ignore the factor “RestaurantPriceRange2”
2) Exploration of data
   a. Looking at the inter-quartile range and the mean in the below graph, we conclude that the average ratings of various neighborhood are varied. See figure 3.
3) Consider review count: but we need to consider the review count for the various neighborhoods as well.
   a. Rearranging the neighborhoods shows that very few do have higher ratings than others. See Figure 4.
   b. Plotting the bar-graph for ratings vs review counts, we find Gaussian distribution
   c. Thus, we lack rigid conclusions. So, we use MCMC method- Gibbs sampler for sampling the data.
4) Selection of hyperparameters: Used dnorm and dgamma for selecting hyper-parameters
5) Gibbs Sampling: Figure 6 shows the trace and the density graph of the sampled values, for the hyperparameters and 72 neighbors. The high noise in the graph and the normal distribution of the density shows that the sampled values are fit to us.
6) Raftery diagnostics: Table 3 shows Raftery diagnostics. Taking:
   a. Quantile (q) = 0.025
   b. Accuracy (r) = +/- 0.005
   c. Probability (s) = 0.95
   d. Dependence factor I = (M+N)/Nmin, where M, N, Nmin are length of burn-in, Sample size and min sample size respectively
7) I > 5 => high autocorrelation => poor choice of hyperparameters. In raftery diagnostics, I < 5, the choice of hyperparameters is acceptable.



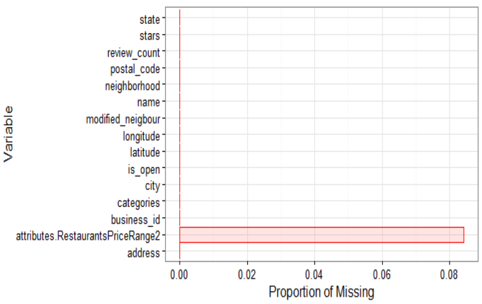




Figure 1: Aggregated Missing Values Visualization

Figure 2: Missing values by proportion

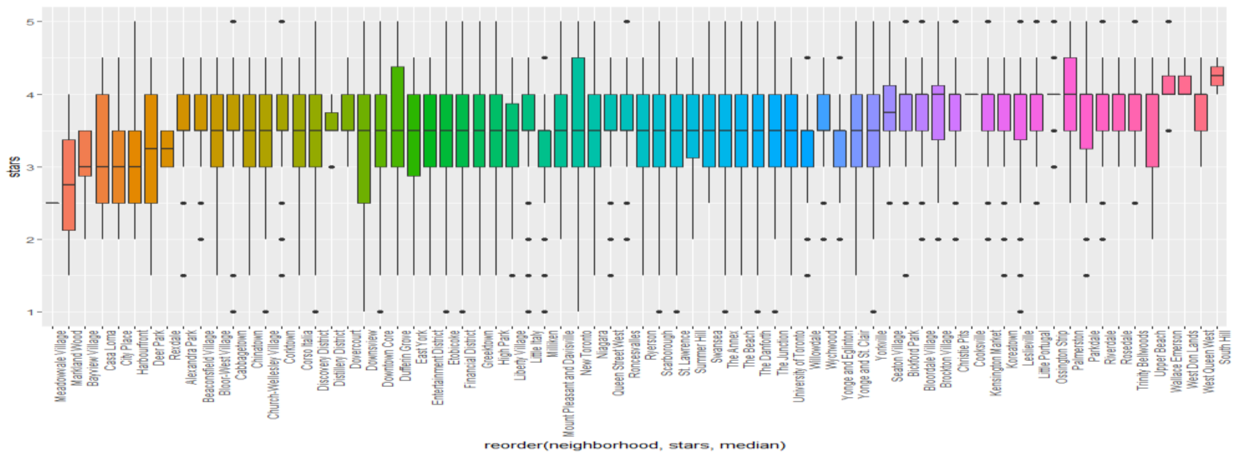



Figure 3: Data Exploration, shows interquartile and variation of ratings using boxplots

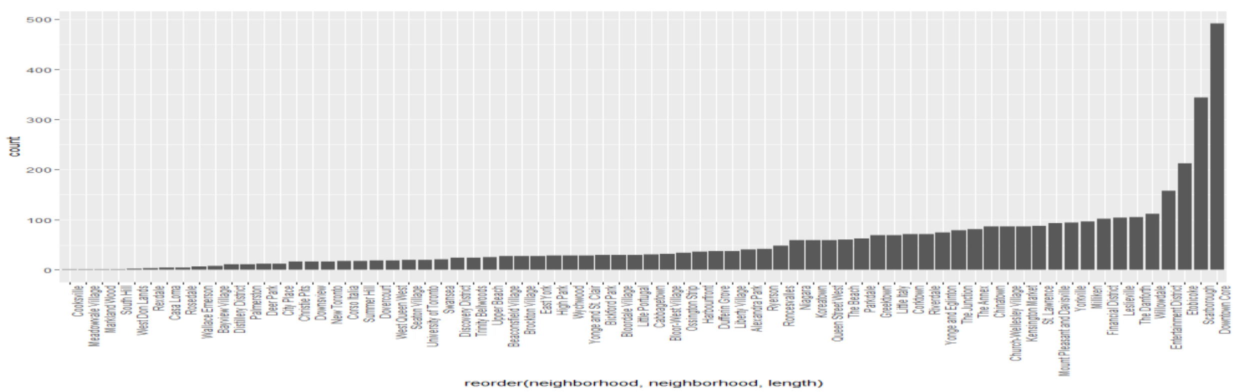



Figure 4: Bar chart to show the distribution of Number of ratings distribution

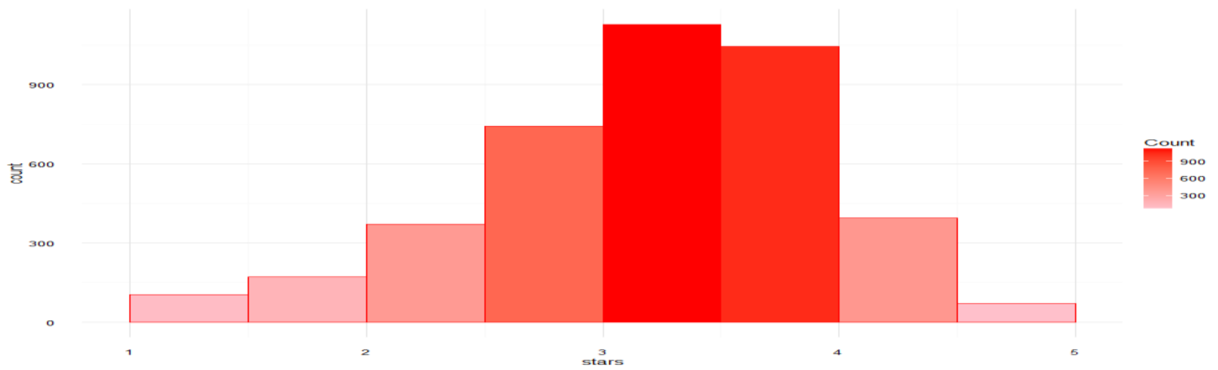



Figure 5: Bar graph of ratings distribution vs review counts

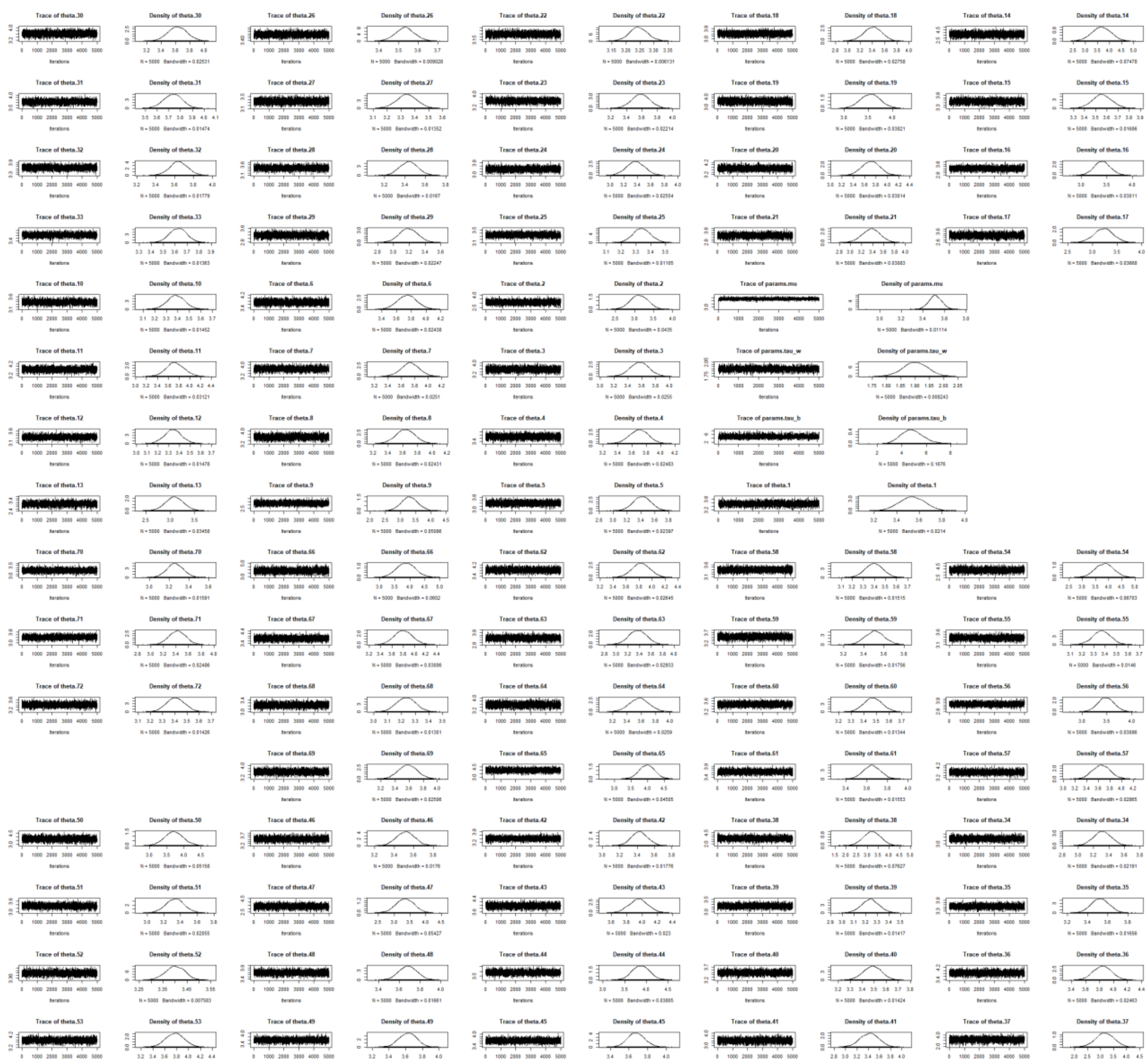


Figure 6: MCMC sampled mean of various parameters

| **Burn-in(M) Total(N) Lower bound(Nmin) Dependence factor (I)** | **(M) (N) (Nmin) factor (I)** |
|---|---|
| **(M) (N) (Nmin) factor (I)** | theta.35 2 3866 3746 1.030 |
| params.mu 3 4030 3746 1.080 | theta.36 2 3680 3746 0.982 |
| params.tau_w 2 3866 3746 1.030 | theta.37 2 3837 3746 1.020 |
| params.tau_b 3 4030 3746 1.080 | theta.38 2 3837 3746 1.020 |
| theta.1 2 3866 3746 1.030 | theta.39 2 3741 3746 0.999 |
| theta.2 2 3680 3746 0.982 | theta.40 2 3561 3746 0.951 |
| theta.3 2 3620 3746 0.966 | theta.41 2 3680 3746 0.982 |
| theta.4 2 3680 3746 0.982 | theta.42 2 3930 3746 1.050 |
| theta.5 2 3803 3746 1.020 | theta.43 2 3561 3746 0.951 |
| theta.6 2 3805 3746 1.020 | theta.44 2 3866 3746 1.030 |
| theta.7 2 3741 3746 0.999 | theta.45 2 3680 3746 0.982 |
| theta.8 2 3805 3746 1.020 | theta.46 2 3866 3746 1.030 |
| theta.9 2 3680 3746 0.982 | theta.47 2 3680 3746 0.982 |
| theta.10 2 3620 3746 0.966 | theta.48 2 3805 3746 1.020 |
| theta.11 2 3620 3746 0.966 | theta.49 2 3995 3746 1.070 |
| theta.12 2 3741 3746 0.999 | theta.50 2 3680 3746 0.982 |
| theta.13 1 3712 3746 0.991 | theta.51 2 3620 3746 0.966 |
| theta.14 2 3866 3746 1.030 | theta.52 2 3803 3746 1.020 |
| theta.15 2 3741 3746 0.999 | theta.53 2 3930 3746 1.050 |
| theta.16 2 3561 3746 0.951 | theta.54 2 3561 3746 0.951 |
| theta.17 2 3741 3746 0.999 | theta.55 2 3741 3746 0.999 |
| theta.18 2 3995 3746 1.070 | theta.56 2 3680 3746 0.982 |
| theta.19 2 3620 3746 0.966 | theta.57 2 3620 3746 0.966 |
| theta.20 2 3803 3746 1.020 | theta.58 2 3866 3746 1.030 |
| theta.21 2 3620 3746 0.966 | theta.59 2 3620 3746 0.966 |
| theta.22 2 3803 3746 1.020 | theta.60 2 3741 3746 0.999 |
| theta.23 2 3680 3746 0.982 | theta.61 2 3803 3746 1.020 |
| theta.24 2 3741 3746 0.999 | theta.62 2 3741 3746 0.999 |
| theta.25 2 3803 3746 1.020 | theta.63 2 3680 3746 0.982 |
| theta.26 3 4062 3746 1.080 | theta.64 2 3930 3746 1.050 |
| theta.27 2 3680 3746 0.982 | theta.65 2 3930 3746 1.050 |
| theta.28 2 3741 3746 0.999 | theta.66 2 3803 3746 1.020 |
| theta.29 2 3741 3746 0.999 | theta.67 2 3620 3746 0.966 |
| theta.30 2 3741 3746 0.999 | theta.68 2 3680 3746 0.982 |
| theta.31 2 3930 3746 1.050 | theta.69 2 3741 3746 0.999 |
| theta.32 2 3620 3746 0.966 | theta.70 2 3741 3746 0.999 |
| theta.33 2 3741 3746 0.999 | theta.71 2 3741 3746 0.999 |
| theta.34 2 3741 3746 0.999 | theta.72 2 3680 3746 0.982 |

Table 3: Raftery diagnostics

8) Comparing the means of the groups with
9) in and between for the parameters

Mean of μ = 3.513029

Mean of τ_within = 1.900632

Mean of τ_between = 4.847146

SD of μ = 0.05859155

SD of τ_within = 0.04280421

SD of τ_between = 0.86872781

Mean of √(1/ τ within) = 0.4598075

SD of √(1/ τ between) = 0.04231237

# Conclusions:

Plotting the mean values of the sampled ratings

- Inclusive of all data points, Figure 7
- Excluding the outliers, Figure 8
- Getting the neighborhoods, Figure 9

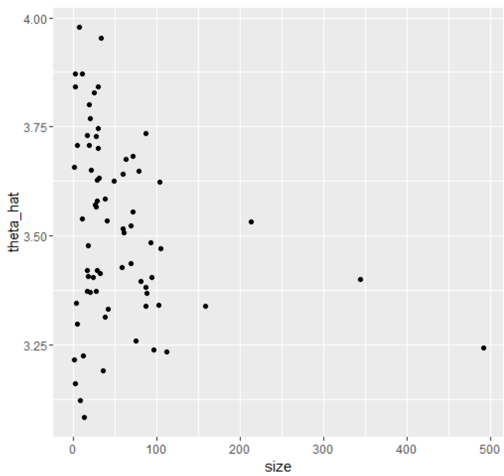



Figure 7: Plot of sampled ratings for neighborhoods

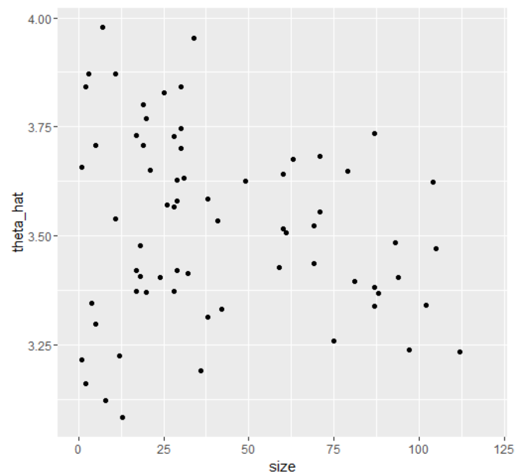



Figure 8: Plot of sampled ratings for neighborhood, excluding outliers

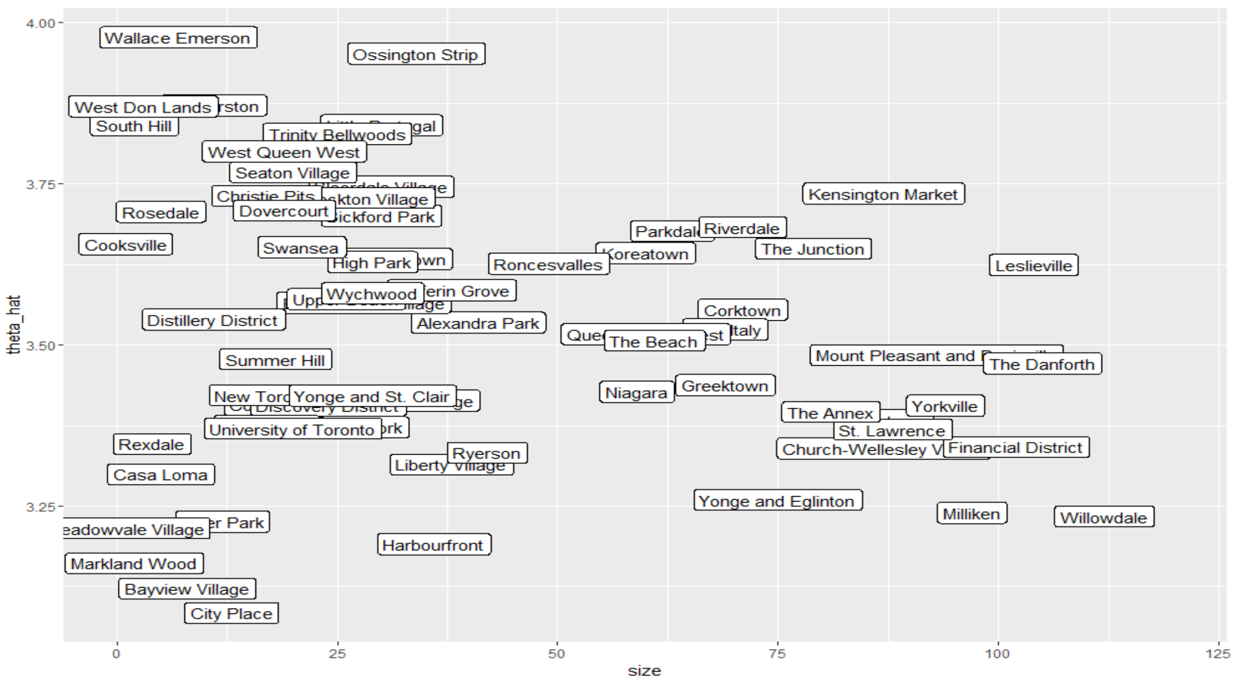



Figure 9:

## Problem Statement 2: What variables are most influential at predicting restaurant rating? How accurate are these predictions?

## Factor Map:

Considered the factors that might impact the restaurant rating, and
added magnitude and score from sentiment analysis.
See Table 4, Table 5, and Table 6

| Variable | Define | Explanation |
|---|---|---|
| address | Address of the restaurant | Accounted by neighborhood variable |
| Business_id | Business id of the restaurant | Accounted by neighborhood variable |
| City | City of restaurant | Filtered for Toronto |
| Hours_open | The opening time | Rating will matter if the restaurant is open, so ignored |
| Hours_closing | The closing time | Rating will matter if the restaurant is open, so ignored |
| Is_open | If the restautant is open | Same for all as restaurant should be open |
| Latitude | Gives location | Accounted by neighborhood variable, useful in data visualization |
| Longitude | Gives locations | Accounted by neighborhood variable, useful in data visualization |
| Business_name | The business name | Accounted by neighborhood variable |
| Neighborhood | Neighborhood of restaurant | Considered |
| Postal_Code | Postal code | Taken into account by neighborhood variable |
| Review_count | Number of reviews | Considered (Extensive variable) |
| Stars | review rating of restaurant | Predictor variable |
| State | State of the restaurant | Same for all |

Table 4: Factors from business file

| Variable | Define | Explanation |
|---|---|---|
| Business_id | Business id of the restaurant | Accounted by neighborhood variable, used for join |
| Review | Review given | Used in Sentiment analysis |
| Votes | Count of votes | Considered above, extensive |
| Date | Date of the review | Since action timeline is missing, this factor is not actionable |
| Review_ID | Review id | Used for join with business dataset |
| Stars | Rating | Rating per review, considered |
| Text | Review text | useful in data visualization |
| User_id | User giving review | Used to join with user dataset |
| Useful | Filter chosen by the customer | Considered |
| Funny | Filter chosen by the customer | Considered |
| Cool | Filter chosen by the customer | Considered |

Table 5: Factors from review file

| Variable | Define | Explanation |
|---|---|---|
| User id | Business id of the restaurant | used for join |
| Average_stars | Review given | Used in Sentiment analysis |
| Elite | The price range of user | Considered |
| Fans | User attribute - # fans | Irrelevant for ratings |
| Friends | # friends user has | Irrelevant for ratings |
| Name | Name of user | Categorical, low p-value |
| Review_Count | The no of reviews user gave | Low p-value |
| Type | The type of user | The factor is accounted in Elite variable |

Table 6: Factors from user file

# Model Understanding:

To predict the response variables for the predictor variable, regression is used to create and fit the model.

## What is Regression?

Regression analysis is a statistical technique to assess the relationship between an predictor variable and one or more response factors.

Typically, GLM models a conditional expectation of Y given X and is defined as:

$\eta = X\beta\eta = X\beta$

$g(E(Y|X)) = \eta * g(E(Y|X)) = \eta$

$E(Y|X) = \mu = g^{-1}(\eta)E(Y|X) = \mu = g^{-1}(\eta)$

$Y|X \sim f(\mu, \sigma^2)$ ,

where, E is the expected value,

g is the link function,

Y is the dependent variable,

X = {X1, X2, …, Xk} is the independent variable, and

f is a probability distribution of the exponential family (Y follows f conditionally on X).

Types of regression models, the families and link to be sued?

A generalized linear model (GLM) family comprise a link function as well as a mean-variance relationship. While GLM is generally with a log link function, the linear regression(LM) is a Gaussian GLM with identity link.

In R, the formulae primarily used for modelling: lm(linear regression) and glm(GLM)

The selection criteria is used in Table 7.

## Which regression model is to be used?

The choice of this conditional distribution depends on the data knowledge/assumptions on the relation between Y and X. If the outcome variable is count/rate, Poisson or negative binomial distribution, with a log link function is checked for fitting.

Here, the response variable here i.e. "Restaurant rating" is the rate (average of review ratings by the customers), it indicates Poisson model. That is a GLM with a log link function.

The response variable i.e. restaurant rating is an extensive variable, i.e. the values will change depending on the size of the system, which in our case is number of users. It is additive in nature. In case of multiple addtitive variables, the correlation between these variables become high. The more additive variables are the more repetitively one property is analyzed in a model.

Mathematically,

Assume,

a, b, c are three variables in a dataset, and a1, b2,c1 and a2, b2, c2 are the corresponding variables after dividing the dataset into two, division ratio be α

a is extensive in nature and c as the only extensive covariable

Initial Model: $a=\mu+\beta1c+\beta2b$

Model for part 1, after the division: $\alpha a1 = \alpha\mu+\alpha\beta1c+\alpha\beta2b = \alpha\mu+\beta1c1+\alpha\beta2b1$ … ($a1=\alpha a$, $c1=\alpha c$, $b1=b$).

Similar Model for part two.

Thus, the model choice for both the parts varies after the division

Consider using a log link function:

Initial Model: $a = \exp(\mu+\beta1c+\beta2b)$

Model for part 1, after the division:

$a1 = \exp(\log\alpha)\ \exp(\mu+\beta1c+\beta2b) = \exp(\log\alpha+\mu+\beta1c+\beta2b1)$

Here, the variable b is the differentiating factor, and is still in extensive format, making choice of model different.

Consider using log scale:

Initial Model: $a = \exp(\mu+\beta1c+\beta2b)$

Model for part 1, after the division:

$a1 = \exp(\log\alpha)\ \exp(\mu+\beta1\log c+\beta2b) = \exp(\log\alpha+\mu+\beta1\log c+\beta2b1) = \exp((1-\beta)\log\alpha+\mu+\beta1\log c1+\beta2b1)$

$= \exp(\mu'+\beta1\log c1+\beta2b1)$

which is our initial model assumption, with the difference in the intercept μ′

| Outcome variable | GLM Family | Link | Mean : variance |
|---|---|---|---|
| Continuous, unbounded | Normal/standard Gaussian | Identity | |
| Continuous, non-negative | Gamma or inverse gamma | | |
| Discrete, counts, rate | Poisson | Log | Identity |
| Discrete, counts, rate | Quassi-poisson or negative binomial | Log | If not identity |
| Count | Gamma | | Over-dispersion |
| Counts with multiple zero | Zero-inflated Poisson may be checked for fitting | | |
| Binary | Binomial or Logistic regression | | |
| Binary, with more than 2 categories | Multinomial regression | | |

Table 7: Model selection Criteria

<u>Steps to be used in modelling:</u>

1) Get scatter plot for understanding the data
2) Do correlation analysis to quantify the association between the variables
   - Get correlation coefficient (a.k.a. Pearson Product Moment correlation coefficient), denoted by $r = \frac{Cov(x,y)}{\sqrt{s_x^2 * s_y^2}}$ ,where $Cov(x,y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$ and $s_x^2$ and $s_y^2$ are the sample variances of x and y respectively and are defined as $s_x^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$ and $s_y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n-1}$
   - r ∈ (-1, +1) signifying direct or inverse relationship

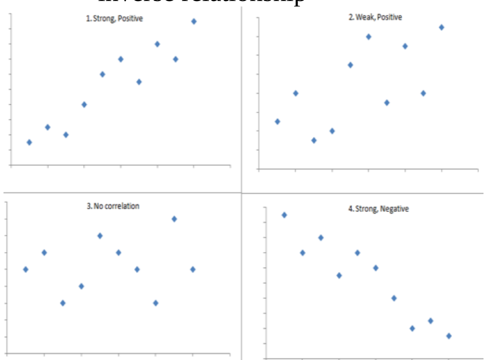



Figure 10: correlation and significance

3) Express covariables in intensive form.
   a. Use log scale for any covariable that is still extensive
4) Use Poisson regression
   a. Use log link
   b. Use log scale
5) For validating, use ANOVA (analysis of variance)

While using GLMs, the Pearson residuals are checked, for they help in understanding the mean-variance relationship, in case of multiple 0 values in the variable.

# Results:

1) Scatter plot: See figure 11
2) Correlation analysis: See figure 12

| | stars | funny | cool | useful | magnitude | score |
|---|---|---|---|---|---|---|
| Stars | 1.00 | -0.06 | 0.05 | -0.06 | -0.01 | 0.77 |
| Funny | -0.06 | 1.00 | 0.70 | 0.63 | 0.18 | -0.09 |
| Cool | 0.05 | 0.70 | 1.00 | 0.77 | 0.20 | 0.00 |
| Useful | -0.06 | 0.63 | 0.77 | 1.00 | 0.27 | -0.10 |
| magnitude | -0.01 | 0.18 | 0.20 | 0.27 | 1.00 | -0.05 |
| Score | 0.77 | -0.09 | 0.00 | -0.10 | -0.05 | 1.00 |

Table 8: Correlation of factors impacting rating

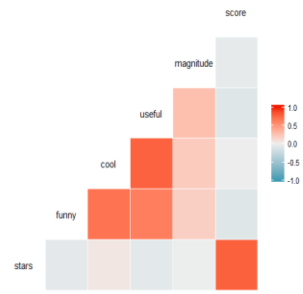



Figure 12: Correlation between various factors

3) Analysis of covariables to be used

| Variables | Extensive | Intensive |
|---|---|---|
| cool | yes | no |
| useful | yes | no |
| funny | yes | no |
| score | yes | no |
| magnitude | yes | no |
| neighborhood | no | yes |

Table 9: Analysis of covariables to be used

4) Model equation:
   Linear model,
   stars ∝ [$\beta_1$ × cool + $\beta_2$ × useful + $\beta_3$ × funny+ $\beta_4$ × score+ $\beta_5$ × magnitude+ $\beta_6$ × neighborhood]
   Generalized model,
   stars ∝ [$\beta_1$ × log(cool) + $\beta_2$ × log(useful) + $\beta_3$ × log(funny) + $\beta_4$ × log(score) + $\beta_5$ × log(magnitude) + $\beta_6$ × log(neighborhood)]

5) Model fitting:
   Conclusion from running the linear model:
   The values of AIC, SE and quartile range suggests that magnitude and score are to be excluded for the model fitting. The remaining factors, since are positive, a poisson model can be tried fitting
   See Table 10:12
   Figure 13:14

# Conclusion:

The model fitted into was poisson model. All the three variables: funny, cool and useful can be used for predicting restaurant rating.

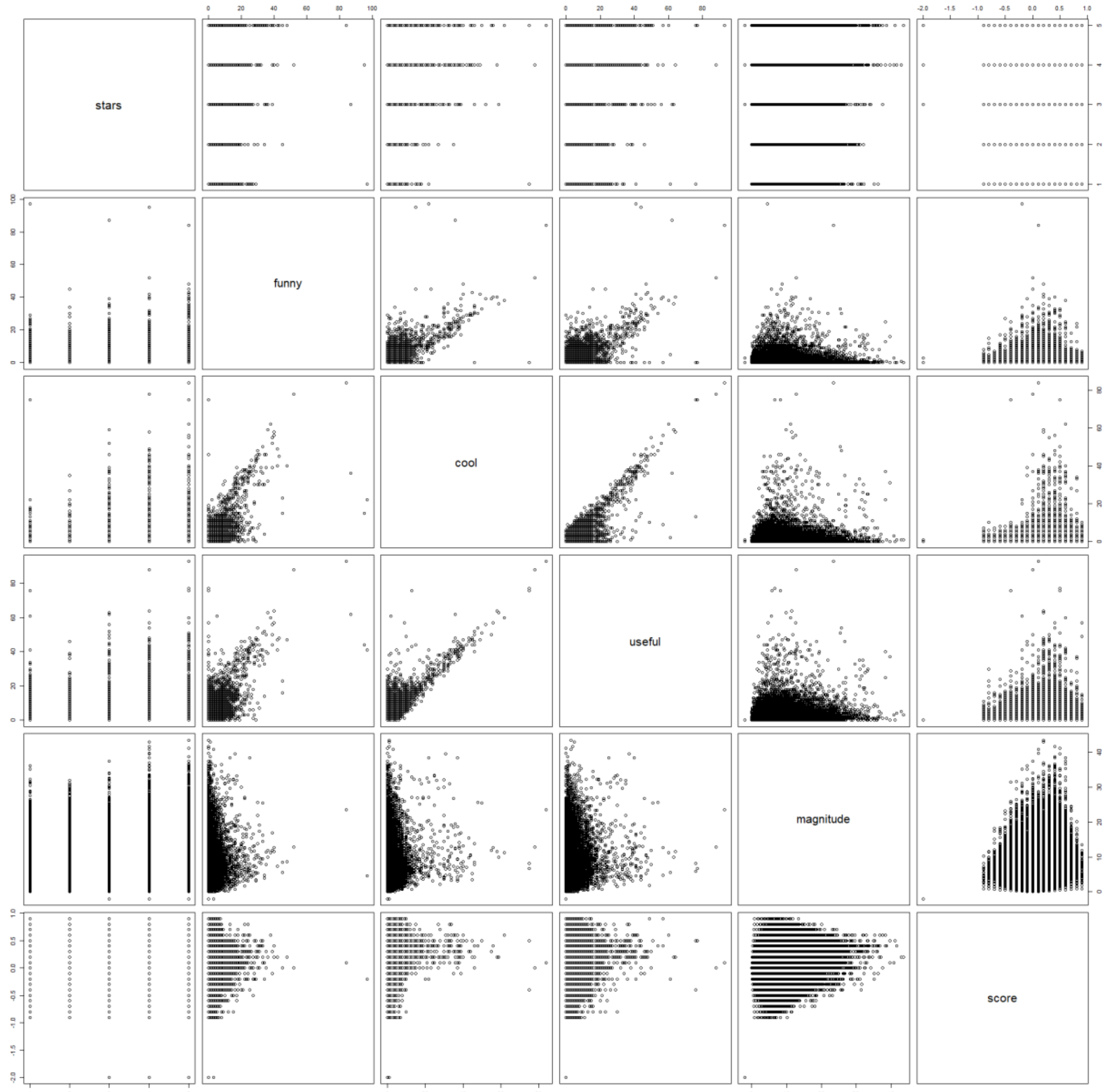



Figure 11: Scatterplots for the data

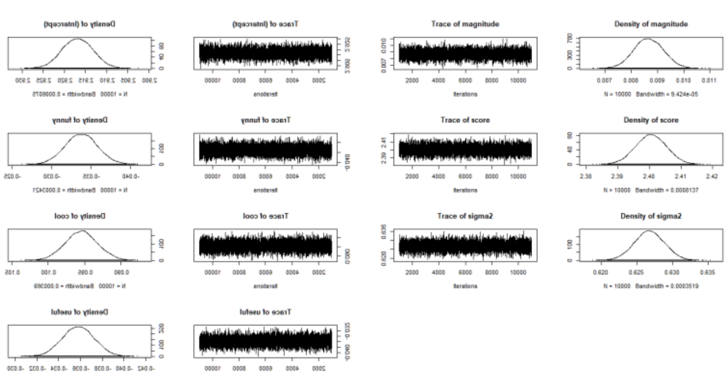



Figure 13: MCMC regression results using linear model

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.4649 -0.5426  0.0300  0.5657  6.9014 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.9169138  0.0036125  807.45   <2e-16 ***
funny       -0.0337988  0.0020384  -16.58   <2e-16 ***
cool         0.0954445  0.0022149   43.09   <2e-16 ***
useful      -0.0358046  0.0014473  -24.74   <2e-16 ***
magnitude    0.0086611  0.0005602   15.46   <2e-16 ***
score        2.4005062  0.0048340  496.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7917 on 177724 degrees of freedom (38276 observations
deleted due to missingness)
Multiple R-squared:  0.5974,   Adjusted R-squared:  0.5974 
F-statistic: 5.275e+04 on 5 and 177724 DF,  p-value: < 2.2e-16
```

Table 10: Linear Model Results

```
Start:  AIC=-83029.26
stars ~ funny + cool + useful + magnitude + score

            Df Sum of Sq    RSS    AIC
<none>                   111389 -83029
- magnitude  1       150 111539 -82792
- funny      1       172 111562 -82757
- useful     1       384 111773 -82420
- cool       1      1164 112553 -81184
- score      1    154560 265950  71643

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)  2.916905 0.0036158 3.616e-05      3.616e-05
funny       -0.033797 0.0020362 2.036e-05      2.036e-05
cool         0.095443 0.0022280 2.228e-05      2.264e-05
useful      -0.035805 0.0014591 1.459e-05      1.459e-05
magnitude    0.008664 0.0005609 5.609e-06      5.609e-06
score        2.400480 0.0048433 4.843e-05      4.843e-05
sigma2       0.626760 0.0021181 2.118e-05      2.118e-05

2. Quantiles for each variable:
                 2.5%       25%       50%       75%     97.5%
(Intercept)  2.909866  2.914427  2.916908  2.919399  2.923911
funny       -0.037815 -0.035160 -0.033786 -0.032428 -0.029818
cool         0.091040  0.093973  0.095435  0.096917  0.099754
useful      -0.038646 -0.036786 -0.035808 -0.034826 -0.032971
magnitude    0.007577  0.008283  0.008655  0.009041  0.009765
```

```
score        2.390883  2.397194  2.400470  2.403760  2.409952
sigma2       0.622607  0.625362  0.626755  0.628169  0.630898
```

Table 10: MCMC regression results

```
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-8.9303   -0.3753   0.1551    0.6424    5.1587

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.3077259  0.0012465 1049.12   <2e-16 ***
funny       -0.0378737  0.0012411  -30.52   <2e-16 ***
useful      -0.0394065  0.0009367  -42.07   <2e-16 ***
cool         0.0731543  0.0012223   59.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 107837  on 216005  degrees of freedom
Residual deviance: 104075  on 216002  degrees of freedom
AIC: 773039

Number of Fisher Scoring iterations: 4
```

Table 10: Poisson regression result

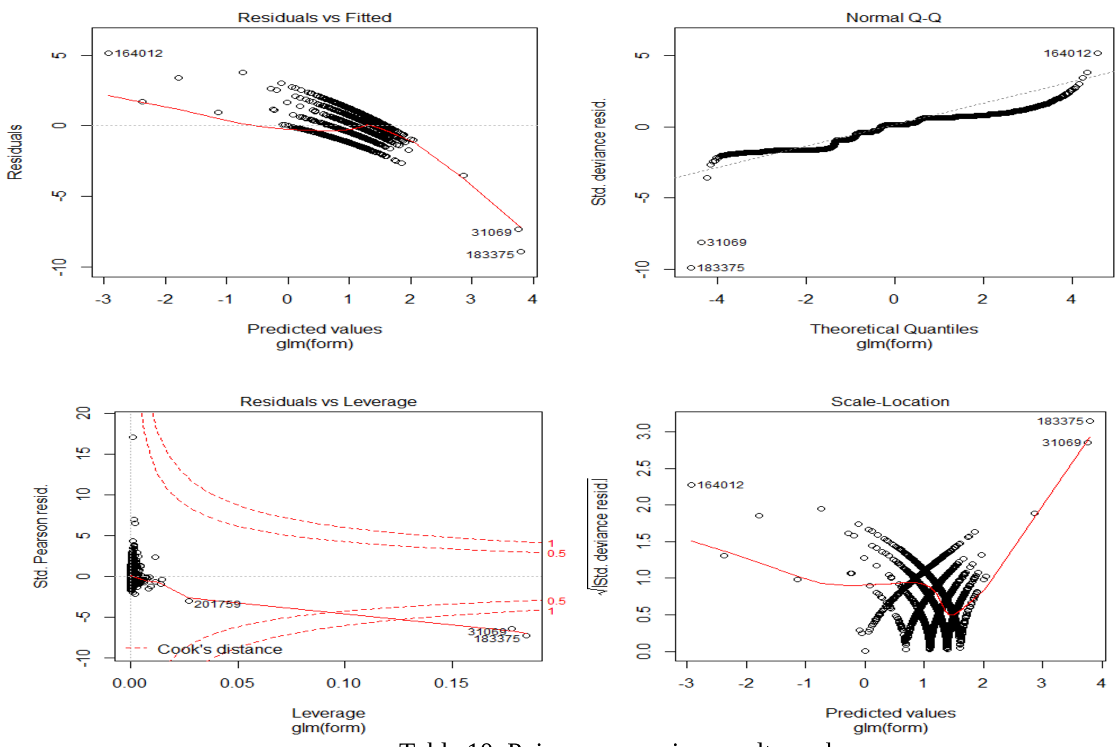



Table 10: Poisson regression result graphs

# Problem Statement 3: Is there any association between neighborhoods and restaurant categories? Can you identify neighborhoods that are more likely to contain certain types of restaurant category than others?

## Model Understanding:

The requirement of the problem is cluster analysis and extraction of a hidden variable.

Cluster analysis

Type of clustering:
Hard clustering: Clusters do not overlap
Soft clustering: Clusters overlap
Cluster analysis can be done by either of the following methods:

- K means clustering: Circular clusters are misleading, plus, in soft clustering, the results become non-informative
- Mixture models

Mixture Model

It is a sound clustering method where soft clustering is expected. Here, each cluster is a generative model (Gaussian or multinomial). Mixture models give more control as while modeling the number of clusters can be specified. The clustered data is obtained by assigning the data point to the cluster component with which it has highest estimated posterior probability.
Data is assumed to be continuous.

Latent Class Analysis

LCA is a modelling technique for observing categorical unobserved factors.

EM algorithm: Expectation Maximization algorithm

Assigns data to the clusters with given probability. EM algorithm provides the parameters of the probability distribution. It starts with multiple random distributions, and for each point gives the probability of it coming from those distributions. It iteratively adjusts the distributions to fit the points assigned to them. These iterations are run till convergence is achieved.

## Process:

1. Create a matrix distributing the category list in various variable columns, add the neighborhoods variable in it.
2. Define the value of G whiles using MCLUST
3. Plot the density and the uncertainty using MCLUST
4. Get BIC values for all the categories
5. Get various fit models for various groups
6. Get the convergence for these models

## Results:

- Various fit models at G levels can be seen in Figure 15
- The convergence of these models can be seen in Figure 16
- The high convergence with G $\geq$ 8 can be seen in the convergence plot.
- It's also worth re-starting the algorithm multiple times to see if better solutions are available.
- We can also perform LCA using Gibbs sampling, and will take longer to run.

## Conclusions:

Yes, there are neighbors that will tend to have specific categories

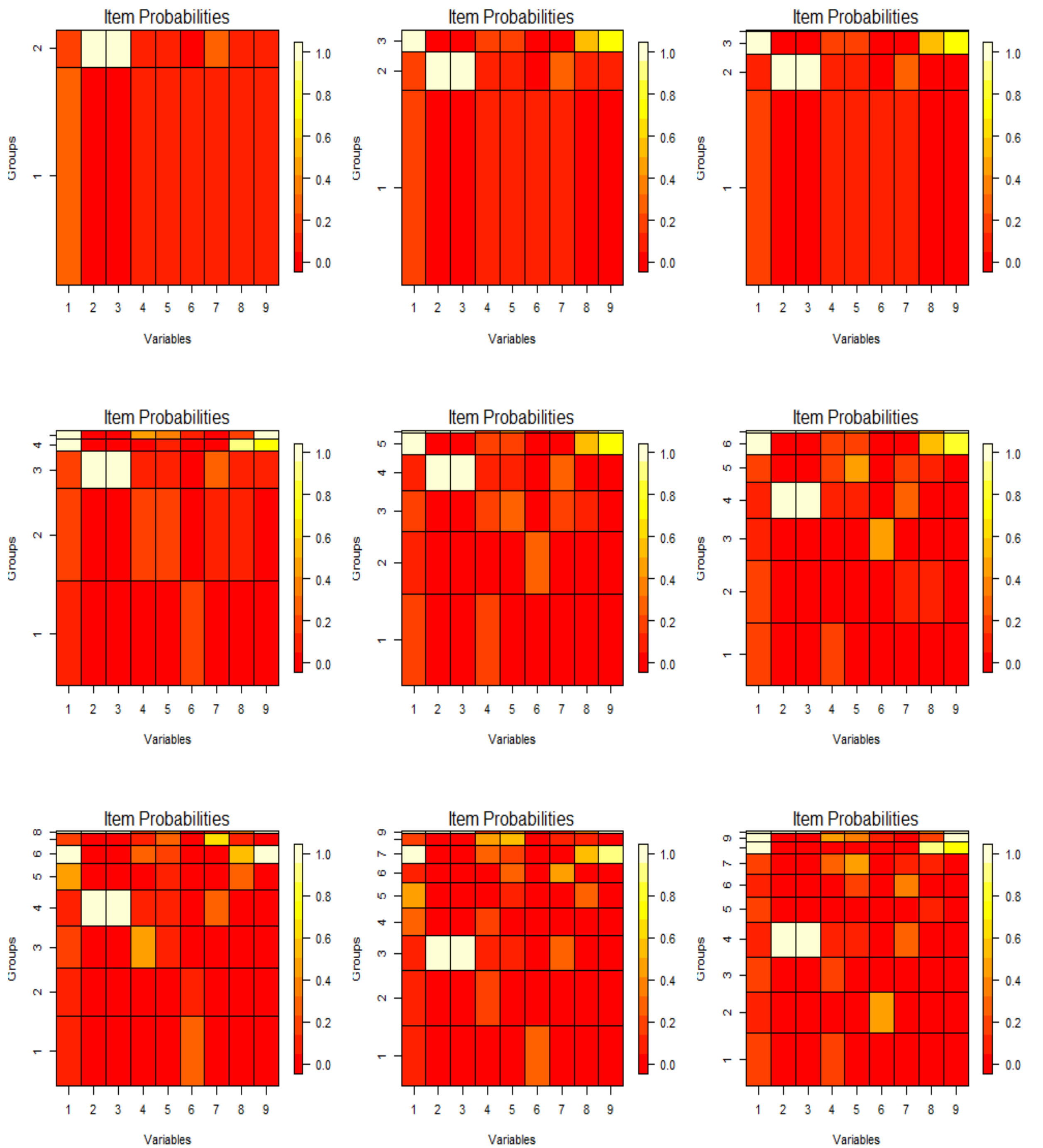



Figure 15: The fit models for various G values

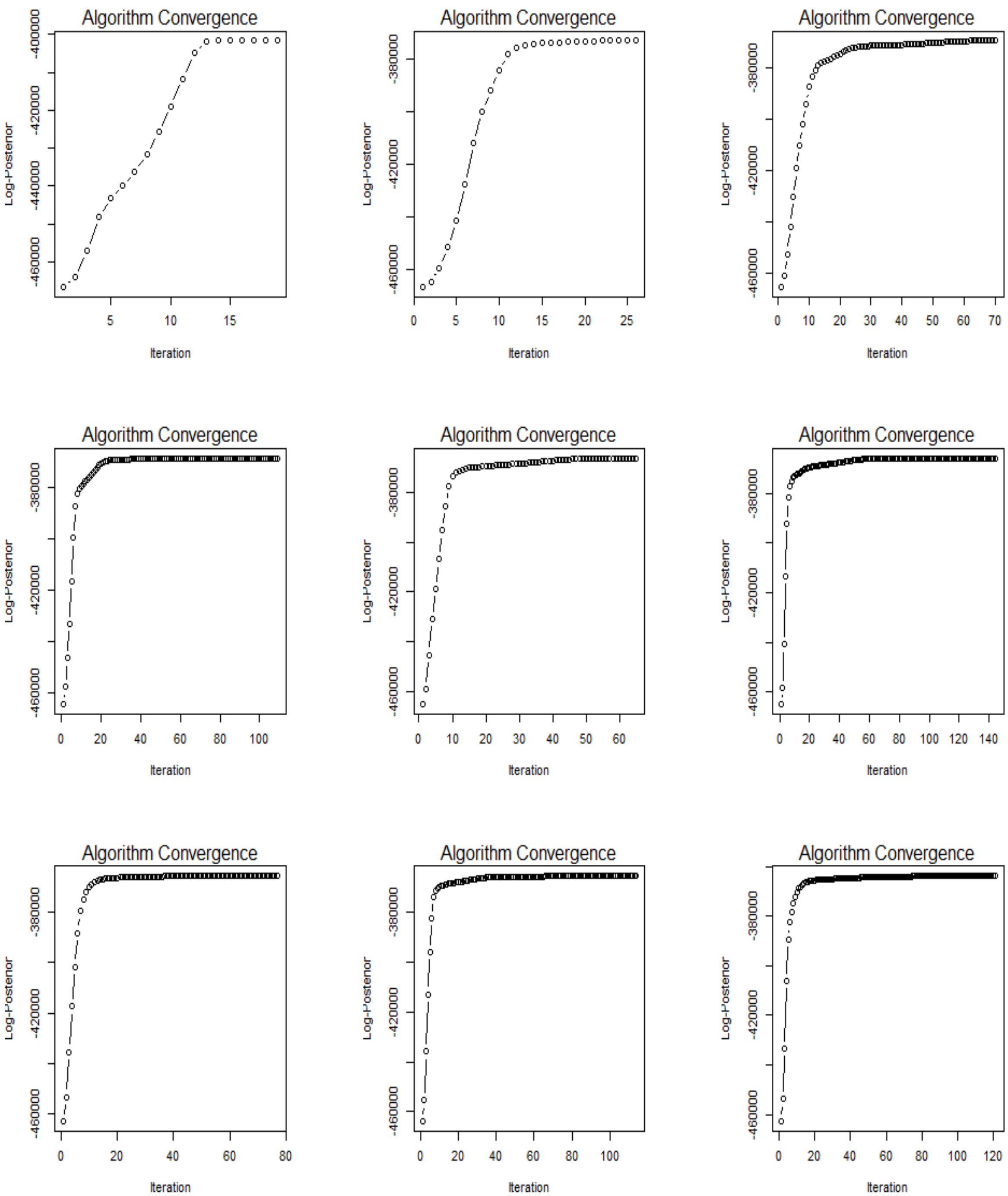



Figure 16: Graphs showing convergence for fit models for various G values