APPLIED STATISCAL MODELLING- MAIN ASSIGNMENT

TABLE OF CONTENTS

1.OVERVIEW

The publicly available yelp dataset is used for analysis and dataset contains online reviews of business platforms. The dataset is in JSON format which has subset of Toronto, Canada. Mainly data includes two tables Business open_Toronto and review table. Business open_Toronto has information about the individual Business restaurant in the toronto city. Restaurant information like business_id, neighbourhood, review_Count, is_open, ratings, address of the restaurant are present. Review Json file contains information about the reviews that users write on yelp website and has variables like business id, review date, for which review has written, rating and review text.

2 Comparison of Ratings of neighbourhood using Gibbs sampling

2.1 Data Management and Data handling:

Initial first level data set is shown in Fig1

Filter

	business_id	name	neighborhood	address
1	l09JfMeQ6ynYs5MCJtrcmQ	Alize Catering	Yonge and Eglinton	2459 Yonge St
2	1K4qrnfyzKzGgJPBEcJaNQ	Chula Taberna Mexicana	Leslieville	1058 Gerrard Street E
3	nbhBRhZtdaZmMMeb2i02pg	Sunnyside Grill		2777 Steeles Avenue W
4	FXHfcFVEfI1vVngW2gVOpw	Bampot House of Tea & Board Games		201 Harbord Street
5	VXH7zXcZzXlmAVN8GSjGRQ	Thai Express		4700 Keele Street
6	dTWfATVrBfKj7Vdn0qWVWg	Flavor Cuisine	Scarborough	8 Glen Watford Drive
7	1nhf9BPXOBFBkbRkpsFaxA	Mirage Grill & Lounge	Yonge and Eglinton	117 Eglinton Avenue E
8	sJ0MYSAIVK28cMzh-s-NPA	Amaya Express	Downtown Core	Eaton Centre, 220 Yonge St
9	JmKgz6n7zn24F-WkgT-kiA	Maki My Way	St. Lawrence	7 King Street E
10	t8yi2l7pZF43Rlf9_lHdDA	Hero Certified Burgers - King & Yonge	Downtown Core	79A Yonge Street

Fig 1 View of Business json data frame

The dataset is arranged in ascending order of business id index. Business yelp dataset is in nested form. It has different classes under one column. For example, Category has more than 70 different categories which need to be refined to have same level across data. Nested data structure is converted into tidy representation. To have tidy dataset, Categories care converted into separate columns and other data frames are converted into separate columns example. Attributes, Business.

Question 1

2.2 I want you to compare the ratings of currently open Indian restaurants in the neighbourhoods of Scarborough and Etobicoke. Which neighbourhood is best for this kind of food? How much better?

The yelp business dataset comprises significant features such as business_id, hours, stars, neighbourhood, is_open, review_Count and other significant attributes. To contrast two neighbourhoods with respect to Indian restaurant category, a subset of data extracted from business dataset that's compatible with the hierarchical model function. Subgroup of data mainly includes column names such as "neighbourhood","is_open", "Indian"," stars".

df_ck <- subset (df_TO_tidy_cat, (neighborhood == "Scarborough" | neighborhood == "Etobicoke") &Indian==1&is_open==1, select = c("stars", "neighborhood","Indian","is_open"))

Approach: Bayesian inference is employed to model the ratings of the open Indian restaurant. Hierarchical modelling is applied to find overall mean, precision between groups and precision within groups.

In the data frame, there are 34 Indian restaurants present in the Scarborough and 14 in Etobicoke neighbourhood. Overall mean of these two neighbourhood is 3.67 and 3.35 respectively. Box plot of neighbourhood with respect to stars are shown in Fig 2

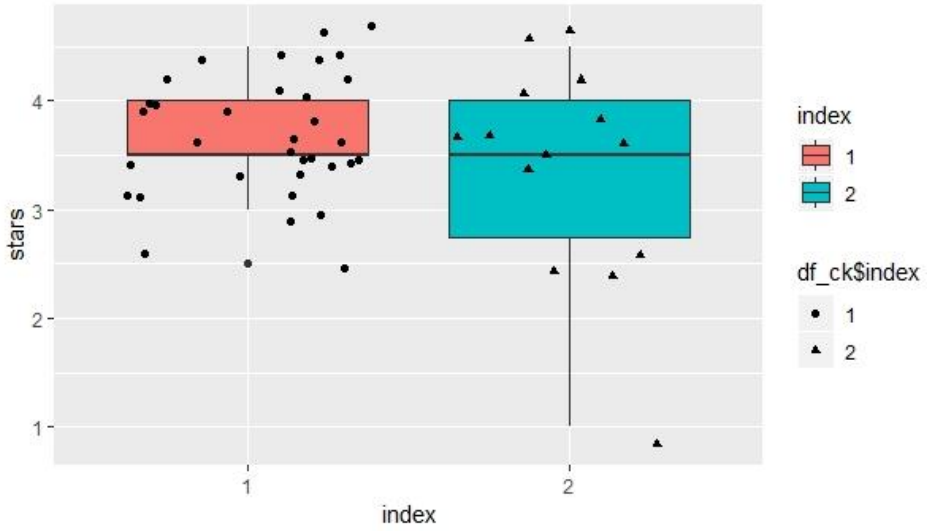


Fig 2 Box plot of neighbourhood with stars

MODELLING:

Parameters

- Mean of stars in neighbourhood "Scarborough" is defined by $\theta_1=\mu+\delta$
- Mean of stars in neighbourhood "Etobicoke" is defined by $\theta_2=\mu-\delta$
- Common precision is given by ~ G (a0, b0)
- Overall mean is given by ~N (mu_0, tau_0)
- Mean difference is given by ~ N(d0,1/r0)

Gibbs sampling is used to model difference in the means of two groups. The posterior probability conditioned on overall mean, common precision of the samples.

Selecting initial priors for the Gibbs sampling is essential. Range of stars in the data is 1-5. On the account of mean is normal distributed, prior mean value assigned as 2.5. As 2 standard deviations from mean contribute 95% interval of the data, prior standard deviation is estimated as 1.265625 (2.5 + 2*sigma). Initial mean and tau_0 are 2.5, 1/ 1.265625.

Priors for a0 and b0 follows gamma distribution. Assigned value for a0 and b0 are 2 and 2.53125 respectively. Posterior estimate of overall mean obtained after Gibbs sampling is 3.517625 can be referred from the graph Fig3.

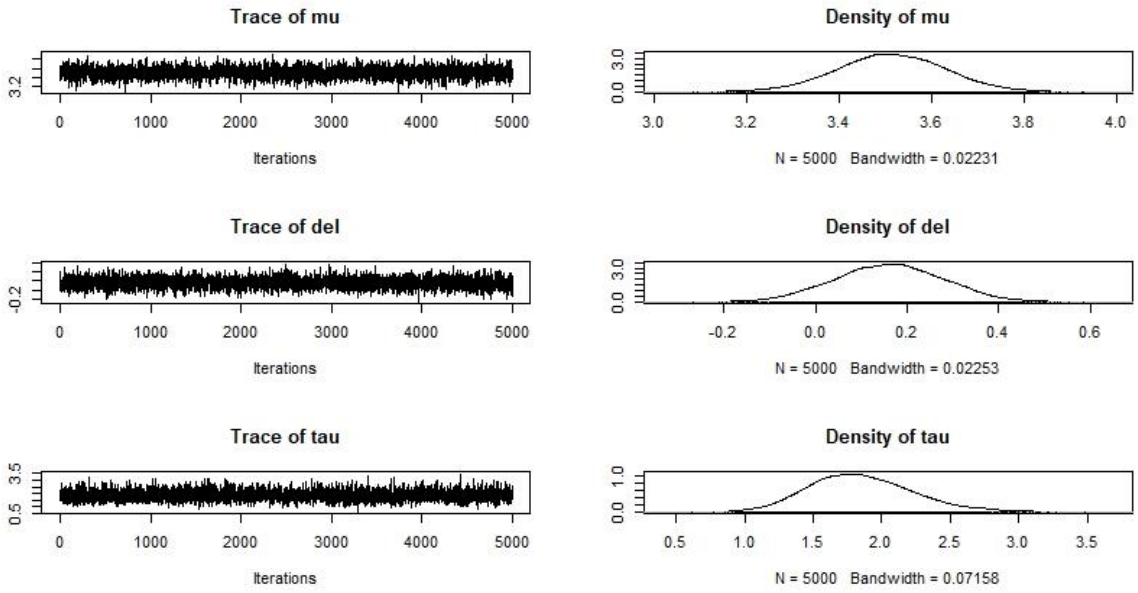


Fig 3 Gibbs sampling mean, precision of both neighbourhood

The sample mean and posterior mean of the both neighbourhoods can be observed from the table 1

Neighbourhood	**Sample Mean**	**Posterior Mean**
Scarborough	3.67	3.65
Etobicoke	3.35	3.35

Table 1 Sample and posterior mean

The sample difference between two posterior samples can be seen from the Fig 4 sample difference plot.

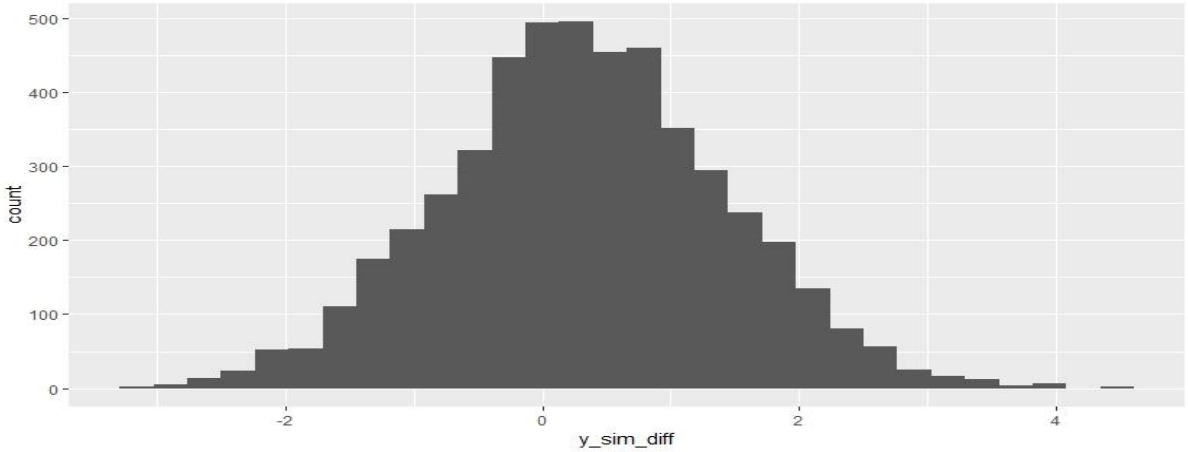


Fig 3 Sample difference

Inference:

From the fig 3 and fig 4, it can be observed that Scarborough's mean is 0.6128 greater than mean of Etobicoke mean.

y1_sim <- rnorm(5000, fit[, 1] + fit[, 2], sd = 1/sqrt(fit[, 3]))

y2_sim <- rnorm(5000, fit[, 1] - fit[, 2], sd = 1/sqrt(fit[, 3]))

ggplot(data.frame(y_sim_diff = y1_sim - y2_sim)) + stat_bin(aes(y_sim_diff))

mean (y1_sim > y2_sim)

0.6128

Comparing the mean of posterior samples of two neighbourhood, Indian restaurants in Scarborough's are not assured to be better than Etobicoke.

Summary:

- The ratings range from 1-5. Mean of two neighbourhoods is compared from the posterior samples.
- There is no inference stating that Scarborough is clearly superior than Etobicoke.

2.3 Compare the ratings of (open) restaurants across multiple different neighbourhoods in the city. Are any neighbourhoods clearly superior to others? If so, by how much?

To compare different neighbourhoods, subset of data frame including neighb ourhood, is_open, Indian, stars columns are extracted. Using Bayesian inference, M-groups neighbourhood are modelled.

Parameters

- μ, the overall mean across neighbourhoods;
- τb, the precision (inverse variability) between neighbourhoods;
- τw, the precision (inverse variability) within neighbourhoods;
- θm, the mean stars rated to neighbourhoods m.

Selecting initial priors for the Gibbs sampling is essential. Range of stars in the data i s 1-5. On the account of mean is normal distributed, prior mean value assigned as 2.5 . As 2 standard deviations from mean contribute 95% interval of the data, prior stand ard deviation is estimated as 1.265625 (2.5 + 2*sigma). Initial mean and tau_0 are 2. 5, 1/ 1.265625.

Using Gibbs sampling, mean, precision between groups and precision within group a re calculated as seen from table2 below

mu	Tau_w	Tau_b
3.512662	1.706760	8.328037

Table 2

The mean ratings of neighbourhoods range from 2.5 to 4.25.A fact can be inferred from Box-plot that South hill has second lowest mean of 2 but it has largest sample rating of 4.25. Box-plot shows clear statement that **Ossington Strip had the highest posterior mean among all neighbourhoods.**

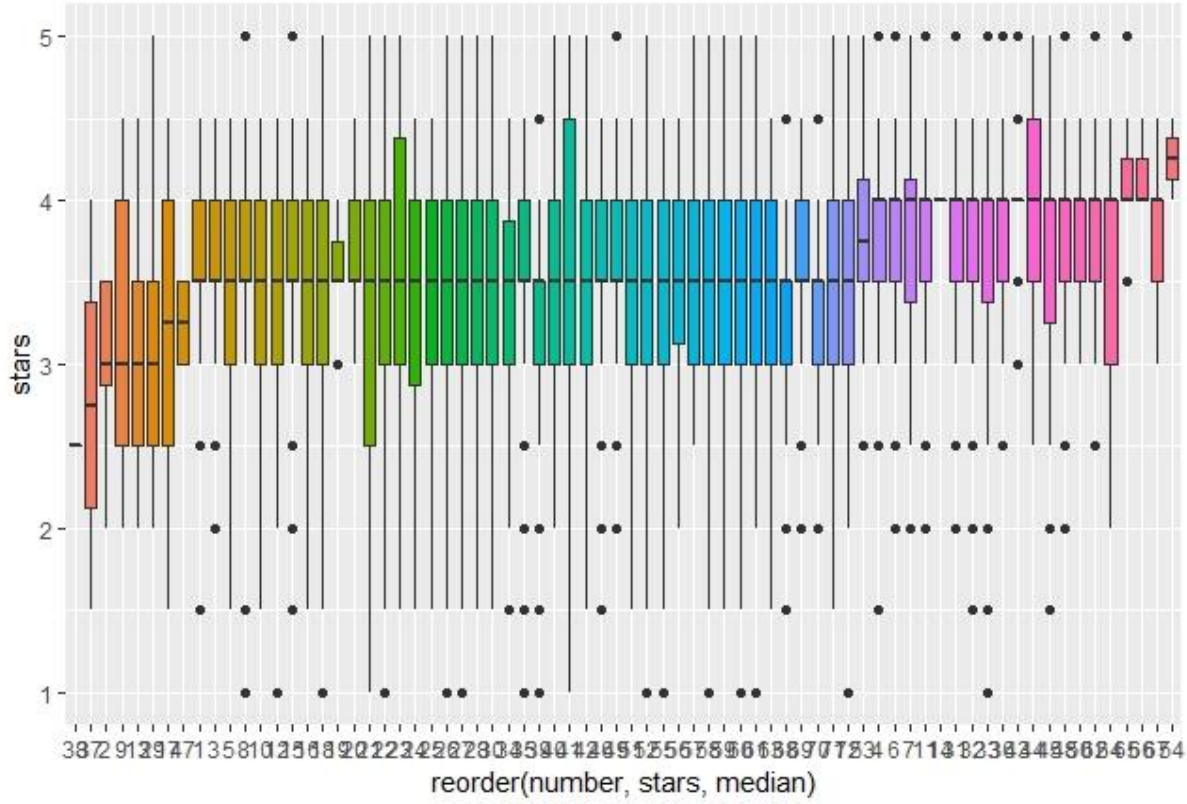


Fig4 Box plot of ratings of open neighbourhood

Posterior Estimates		
Mean	Precision within group	Precision between groups
3.5102646190335	1.90	15.69

Ossington strip has 35 restaurants with posterior theta of 3.89 than south hill which has 2 Ossington strip with rating of 3.6. Posterior mean of Ossington strip is influenced by sample size which is not same in case of south hill of small sample size.

South hill has lowest number of restaurants. The difference between sample mean and theta is greater than posterior mean that is influenced by overall mean. Because of sample size increase, the sample mean, and theta comes closer.

The basis why Ossington strip is clearly superior to all other neighbourhood can be proved by comparing mean of Ossington with overall mean of all neighbourhoods.

theta_hat <- apply (fit2$theta, 2,mean)

mean(fit2$theta[,41]>mean(theta_hat))

0.9998

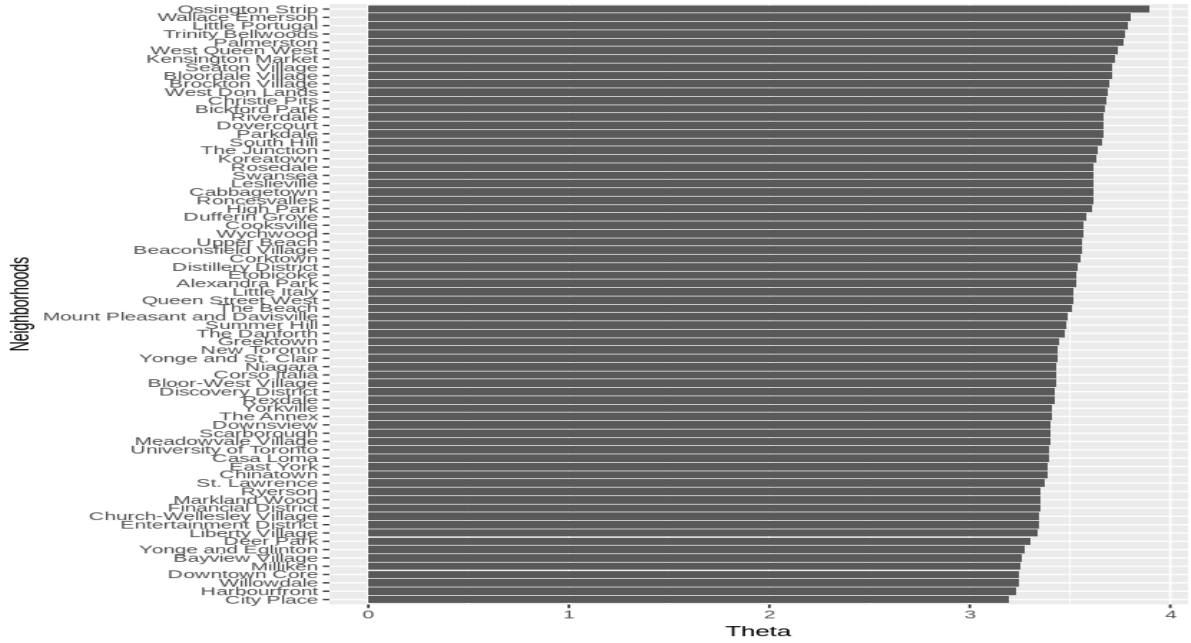


Fig 5 Comparison of mean(theta) for open neighbourhoods

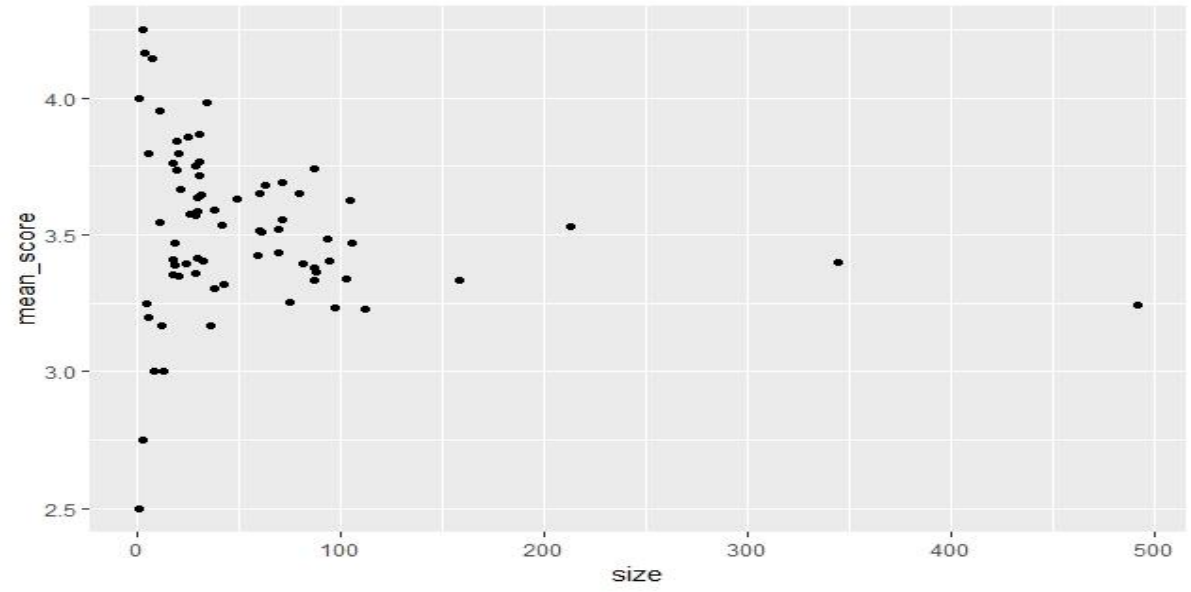


Fig 6 size and difference between theta and sample mean.

Another way of showing superiority is comparing Ossington strip with city place which has low theta value.

```
Ossington _sim <- rnorm(5000, fit[, 1] + fit[, 2], sd = 1/sqrt(fit[, 3]))
city _sim <- rnorm(5000, fit[, 1] - fit[, 2], sd = 1/sqrt(fit[, 3]))
ggplot(data.frame(y_sim_diff = y1_sim - y2_sim)) + stat_bin(aes(y_sim_diff))
sum (Ossington_Sim > city_sim)/5000
```

0.7562

Question 2

3.1 What are the factors most strongly associated with restaurants being closed?

Initial analysis is done to explore more about parameters or significant features. Initial hypothesis taken into consideration.

```
#  Based on Higher review count, is there any relationship between quality of restaurant?
# hypothesis 1: higher number of review count is contributing to quality of restaurant?
```

The restaurants have highest number of review counts contributing to class of the restaurant. The class status is either “good” or “bad”. The class “GOOD” is assigned by calculating the restaurant stars is greater than 3.5 to and restaurant which are less than 3.5 are assigned as “BAD”. To explore this, RestaruantResearvation is chosen. By assuming high quality restaurant have reservation. Based on this assumption, the class of each restaurant is estimated if the RestaruantResearvation is “TRUE”.
Hypothesis is proven by taking stars, review Count, class and plotting in the box plot. As seen from the graph, more the review count with good stars contribute in determine whether the restaurant is good or bad.

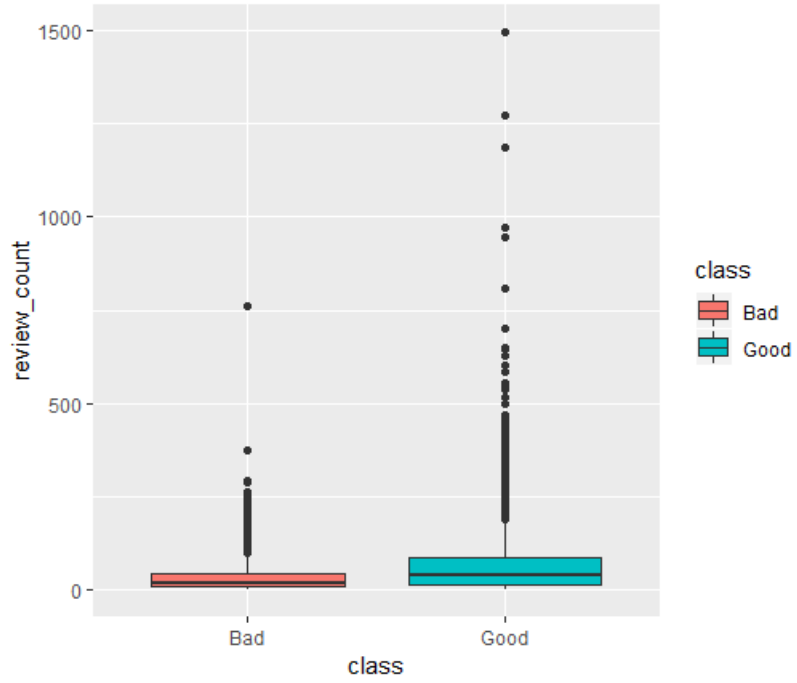


Fig7 Boxplot of class of restaruant

Feature selection:

Exploratory analysis gives interesting facts about significant features and contribute in selecting the features for modelling purposes. Nested structure of business dataset is converted into columns to maintain same level across the data. Total of 155 features are selected for the analysis. Let 's analyses each variable one by one.

One -value attributes VS Is_open:

- State attribute present in the dataset contains only value of Ontario as ON. As yelp dataset is analysed only for Ontario city. This doesn't have any impact on predicting whether the restaurant is open or not

```
> head(df_TO_tidy_cat$state)
[1] "ON" "ON" "ON" "ON" "ON" "ON"
```

- As similar as state, City is not considered as significant variable since city value is filtered to explore only Toronto.

```
> head(df_TO_tidy_cat$city)
[1] "ON" "ON" "ON" "ON" "ON" "ON"
```

- As similar as state, Postal code is not considered as significant variable since it has dif ferent levels of postal code which doesn't affect or impact on is_open.

```
head(df_bus_TO$postal_code)
[1] "M4P 2H6" "M4M 3A6" "M3J 3K5" "M5S 1H6" "M3J 1P3" "M1S 2C1"
```

Hours Vs Is_open

The hours parameter is split into opening and closing hours of Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. From the boxplot, different timing affect both stars and is_open. Mainly two open_hours and closed_hours of both weekend& weekdays are considerd.

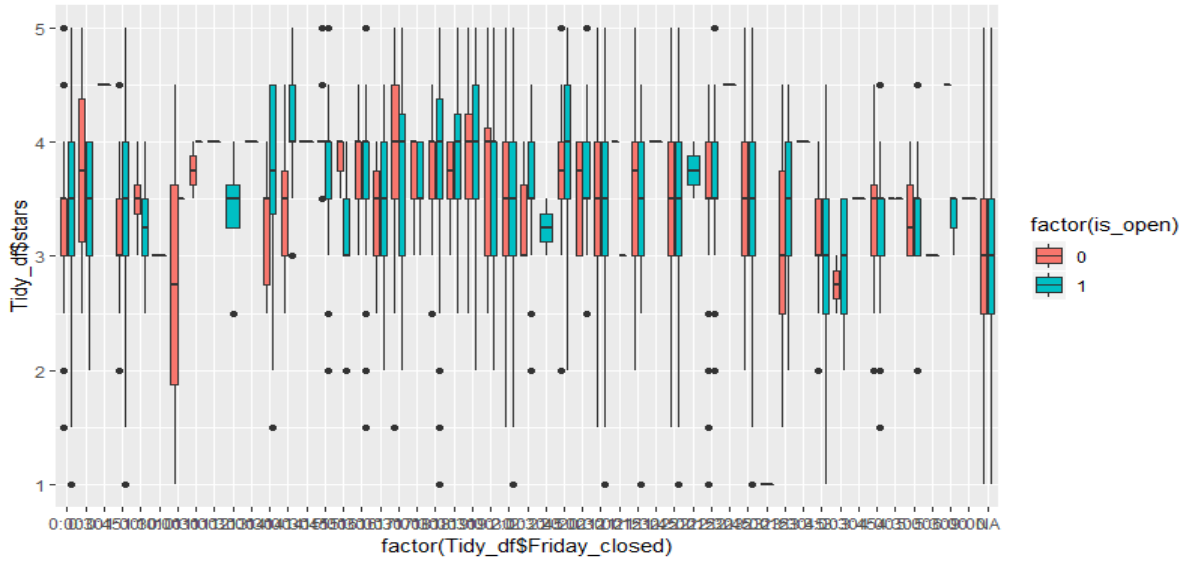


Fig 8 Hours VS Is_open

Categories vs Is_open

There are more than 70 different categories under categories column. Each category is plotted against is_open and stars. There are over 50 categories that doesn't have impact on is_open or ratings. There is mean variation is observed for restaurant being open or closed. It can be inferred from the Fig 9.

Since it is difficult to include each category in predicting whether restaurant is being closed or not. Over 50 categories are removed from the data frame

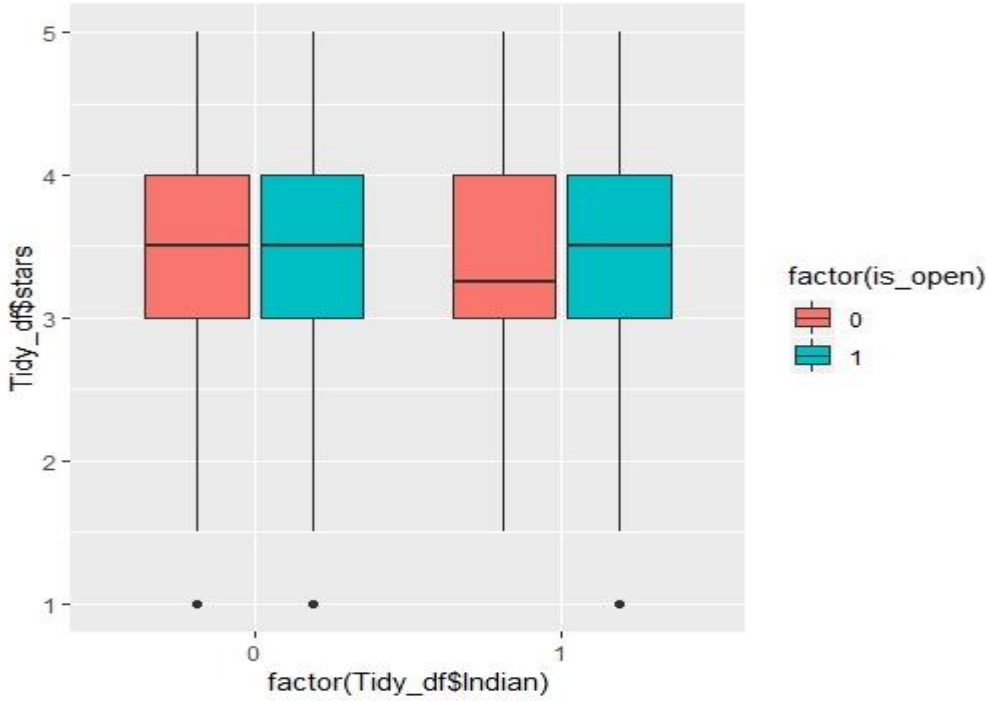


Fig 9 Is_open vs Indian vs Stars

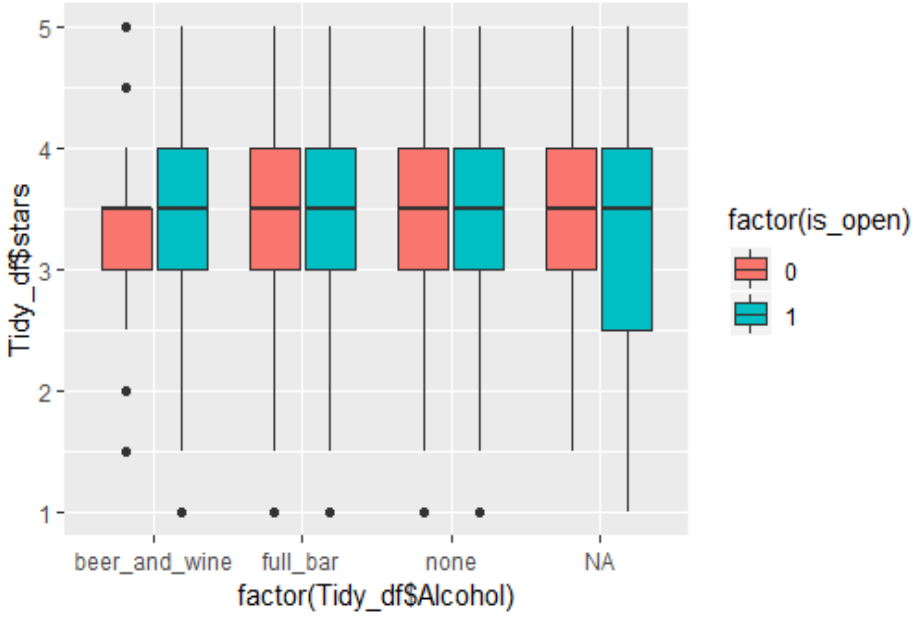


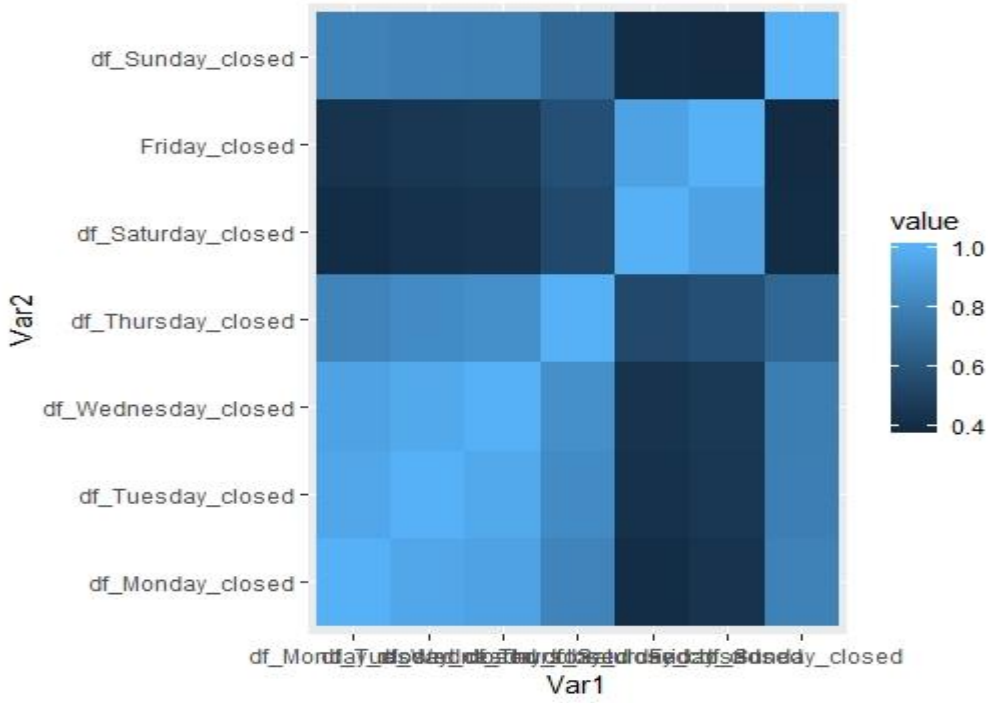
Fig 10 Is_open vs Alcohol

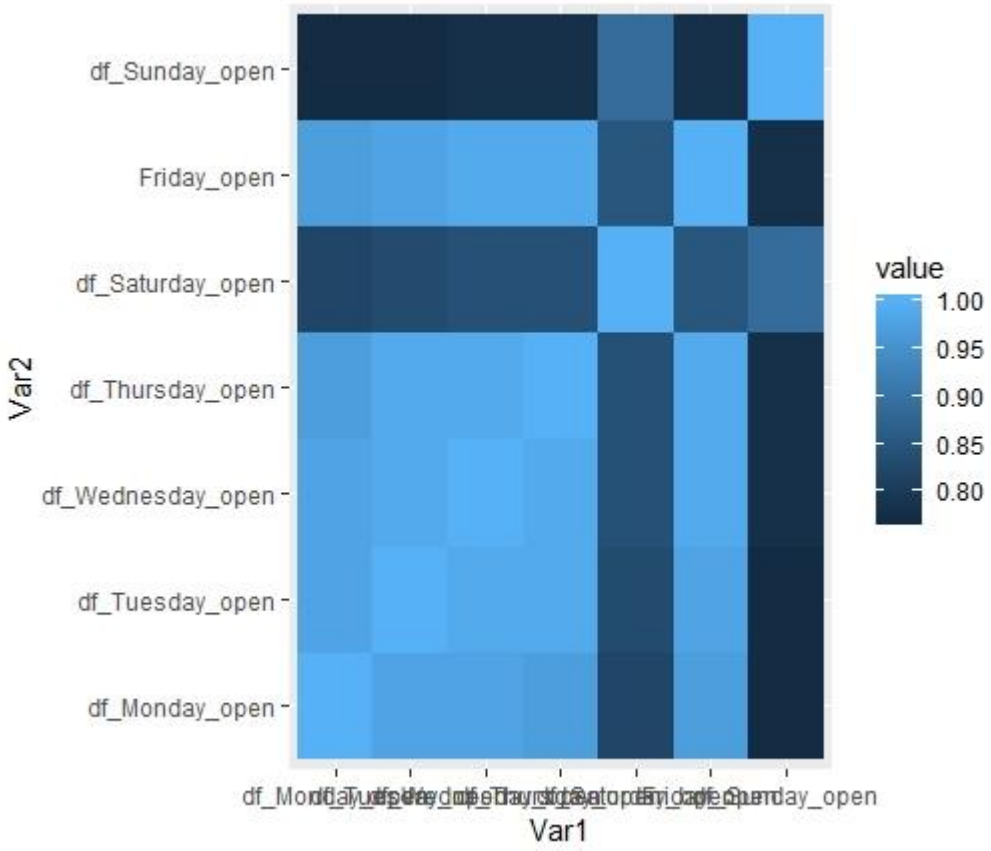
Eliminating the following variables that doesn't have significance impact on rest being closed because there are no mean variations are observed.

Nightlife,Canadian(new),Sanwidches,Breakfast &brunch, Italien,Cafes,Coffe &tea,Fast food, Japanese , Middle Eastern.Mediterranean, Korean,pubs, GoodForMeal_dessert","GoodForMeal_lunch","BusinessParking_Valet","BusinessParking_validated","GoodForMeal_dinner","GoodForMeal_latenight","BikeParking","GoodForMeal_brunch","Ambience_classy","Ambience_hipster","lot","OutdoorSeating","RestaurantsTableService","RestaurantsReservations","RestaurantsGoodForGroups","NoiseLevel","HasTV","GoodForKids

Co-relation removal.

As weekdays and weekends have mostly common opening and closing time, it is high probability that opening and closing hours are highly corelated. It can be inferred from the





It can be inferred from the heatmap of both opening and closing hours of days. Friday Closed, Saturday open, Thursday_closed, Friday_open are chosen as part of feature selection.

Chi-2 test.

Null-Hypothesis: Two variables are independent
Chi-2 test is performed on variables with respect to is_open to accept the null hypothesis. Performing chi-2 test on attributes showed the dependency of the variables over the restaurant is being closed or not.

For some of the attributes, P value is observed more than 0.05. As the variables are independent and not impacting is_open, insignificant parameters are eliminated.

```
tbl1<-table (Regul$is_open, Regul$`Sushi Bars`)
chi2 = chisq.test(tbl1, correct=F)
sqrt(chi2$statistic / sum(tbl))
```

0.7483

The following attributes are removed after performing chi-2 test.

- Burgers
- Sushi_bars
- Asian Fusion
- Thai
- Mexican
- Seafood.

Data imputation:

Missing data imputation:

As part of missing data evaluation, variables have more missing “NA” values than 3000 are not considered and doesn’t have impact on restaurant being closed.

Caters	3054
WheelchairAccessible	4492
DogsAllowed	6148
GoodForDancing	6427
Music_background_music	6469
Music_no_music	6469
Music_karaoke	6469
Music_live	6469
Music_video	6469
Music_jukebox	6469
Ambience_romantic	6469
HappyHour	6476
CoatCheck	6477
Smoking	6478
BestNights_tuesday	6638
BestNights_friday	6638
BestNights_wednesday	6638
BestNights_thursday	6638
BestNights_sunday	6638
BestNights_saturday	6638
GoodForMeal_dessert	6638
DriveThru	6661
ByAppointmentOnly	7106
DietaryRestrictions_gluten-free	7136
DietaryRestrictions_vegan	7136
DietaryRestrictions_kosher	7136
DietaryRestrictions_halal	7136
DietaryRestrictions_soy-free	7136
DietaryRestrictions_vegetar-ian	7136
vegetarian	7136
AgesAllowed	7142
Open24Hours	7146
RestaurantsCounterService	7147
HairSpecializesIn_curly	7147
HairSpecializesIn_perms	7147
HairSpecializesIn_kids	7147
HairSpecializesIn_extensions	7147
Music_dj	7147
BusinessParking_garage	7148

KNN imputation:
To impute data, KNN imputation is performed on opening and closing hours of Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. For each imputation different k are chosen because of the levels in the attributes,

```
impute1 <- kNN(Data, variable="Friday_open",k=8)
summary(impute1)
```

3.2How accurately can you predict when a restaurant in the dataset will be closed?

To predict when a restaurant in the dataset will be closed, Logistic regression with lasso is applied

The data frame is split into train and test. Train data contains 70% of the data to train the model. The 30 % of data as test is utilized to perform cross-validation. The glm library is applied on training features to predict the whether the restaurants are open or closed. The class label "1 "represents the restaurant is being open and "0" otherwise.

The following features are chosen as per feature selection performed above.

- "neighborhood"
- "latitude"
- "longitude"
- "stars"
- "review_count"
- "is_open"
- "Chinese"
- "Pizza"
- "American (Traditional)"
- "df_Thursday_closed"
- "Friday_closed"
- "df_Saturday_open"
- "Friday_open"
- "RestaurantsPriceRange2"
- "Alcohol"
- "RestaurantsTakeOut"
- "Ambience_casual"
- "BusinessAcceptsCreditCards"

```
logitMod <- glm (is_open~neighborhood+longitude+latitude+stars+review_count+Chinese
        +Pizza+`American (Traditional)`+df_Thursday_closed+Friday_open+Friday_closed
        +df_Saturday_open+RestaurantsPriceRange2+Alcohol+RestaurantsTakeOut+Ambience_casual+BusinessAcceptsCreditCards
        ,data=train_d)
```

Call: glm(formula = is_open ~ neighborhood + longitude + latitude +
stars + review_count + Chinese + Pizza + `American (Traditional)` +
df_Thursday_closed + Friday_open + Friday_closed + df_Saturday_open +
RestaurantsPriceRange2 + Alcohol + RestaurantsTakeOut + Ambience_casual +
BusinessAcceptsCreditCards, data = train_d)

Coefficients:

Degrees of Freedom: 5003 Total (i.e. Null); 4757 Residual
Null Deviance: 1061
Residual Deviance: 877.7 AIC: 5986

Predicting with testing data helps to explore more interesting results. For example Chinese restaurant

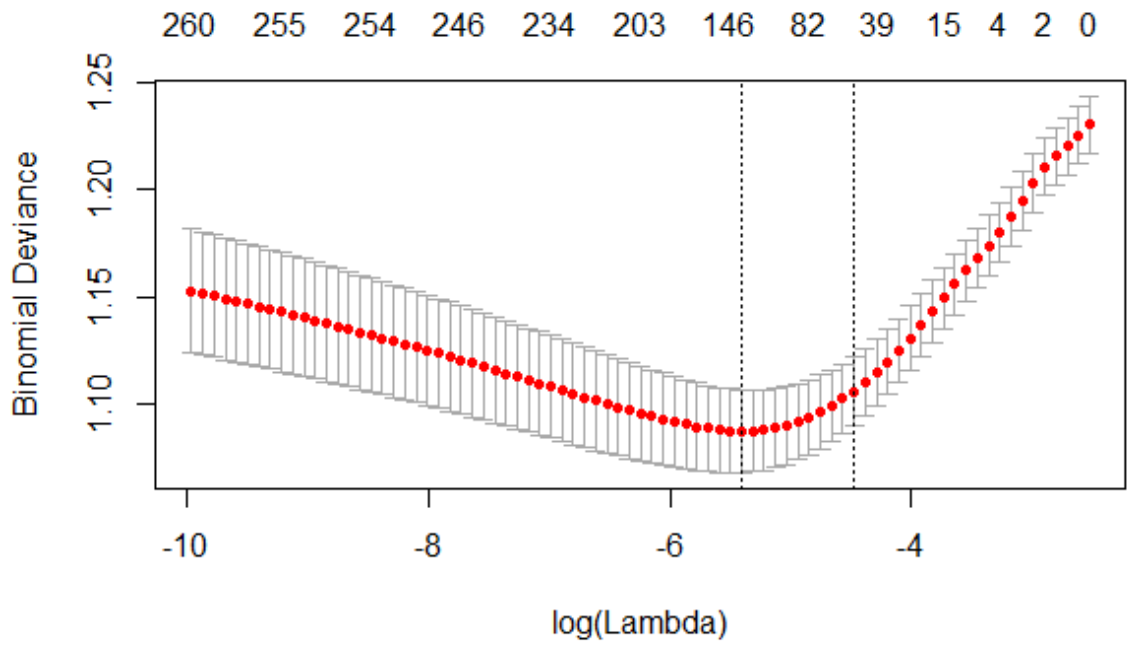
negative co-efficient -0. 1432927.The negative co-relation is observed between Chinese and the restaurant being closed. Most likely the Chinese restaurant will be closed rather than being open.

The co-efficient of the -2.423e-02. This implies that most of the restaurants with price r ange 3 have high probability of being closed.

```
> confusionMatrix(data = factor(as.numeric(pdata>0.5)), reference = factor(test_d$is_o
n))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  303  478
         1  351 1012

               Accuracy : 0.6133
                 95% CI : (0.5923, 0.634)
    No Information Rate : 0.695
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1351
 Mcnemar's Test P-Value : 1.208e-05

            Sensitivity : 0.4633
            Specificity : 0.6792
         Pos Pred Value : 0.3880
         Neg Pred Value : 0.7425
             Prevalence : 0.3050
         Detection Rate : 0.1413
   Detection Prevalence : 0.3643
      Balanced Accuracy : 0.5712

       'Positive' Class : 0
```

Model was able to predict the neighbourhood with accuracy of 0.6133. Tuning hyperparameters and performing regularization with lasso takes significant features into account and predicts with accuracy of 0.72. Both can be inferred from the Fig a and fig b. The graph shows number of non-zero coefficients change with λ value changing to 0..

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  178  116
         1  476 1374

               Accuracy : 0.7239
                 95% CI : (0.7044, 0.7427)
    No Information Rate : 0.695
    P-Value [Acc > NIR] : 0.001816

                  Kappa : 0.2298
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.27217
            Specificity : 0.92215
         Pos Pred Value : 0.60544
         Neg Pred Value : 0.74270
             Prevalence : 0.30504
         Detection Rate : 0.08302
   Detection Prevalence : 0.13713
      Balanced Accuracy : 0.59716

       'Positive' Class : 0
```



Question 3

4. Restaurants are organised by neighbourhood. There are a lot of neighbourhoods in the data. Using the longitude and latitude of each restaurant, can you find other ways to organise restaurants together, e.g., using a clustering model? Can you find any interesting associations with other elements of the data using this clustering?

To organise the restaurants together, another approach is clustering model. The subgroup of data frame including latitude and longitude are extracted. Model based clustering is applied to cluster the neighbourhood based on co-ordinates. The data is fit using gaussian mixture models over range of groups G via EM algorithm.

Mostly related with AIC, Bayesian information criterion is chosen as metric to select best model which fits the cluster with minimum uncertainty. Using "mclust" package, BIC is calculated to select number of groups that fits the data best. Highest BIC value with minimum uncertainty is preferred over other top 2 models.

By default, mclust estimates BIC score for 9 number of components or clusters which can be seen from the graph 8. BIC information gives the best 3 models with scores.

Top 3 models based on the BIC criterion:

VVV,9	VVV,8	VVV,7
58229.81	57646.91	57576.59

It can be observed from the table that model "VVV,9" component has maximum BIC score out of all other clusters. There might be a possibility that there are other models are lying on the edge of 9th cluster.

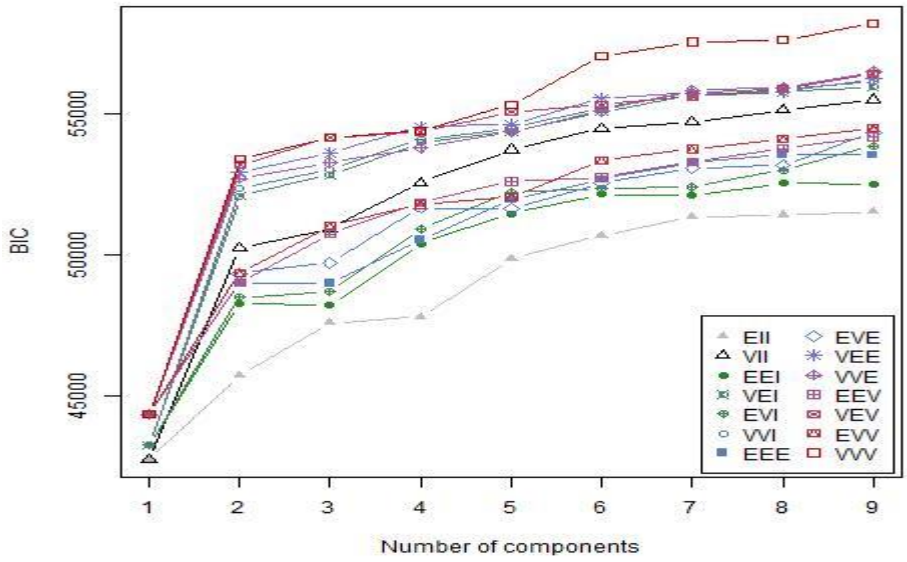


Fig 14. BIC plot of Clusters (Default)

To test the assumption, “mclust” is fitted for more clusters ranging value from 1 to 20. From the figure 9, it can be observed that, 20 is best fit model. There might be a possibility that other models can be lying on the edge.

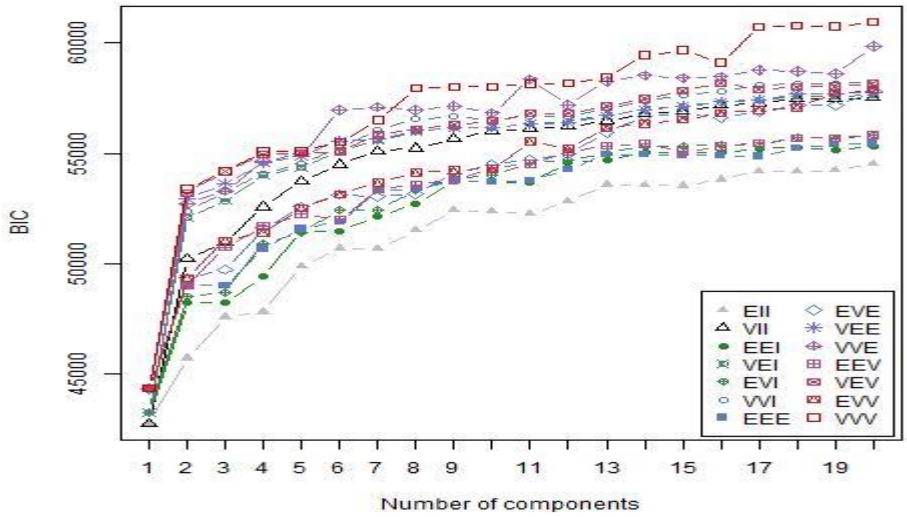


Fig 15. BIC plot for clusters (1-20)

Again, mclust is fitted for more models ranging from 1 to 40. Fit gives the best top 3 models with score. Best model 36 is chosen with minimum uncertainty. Uncertainty plot can be referenced from the fig

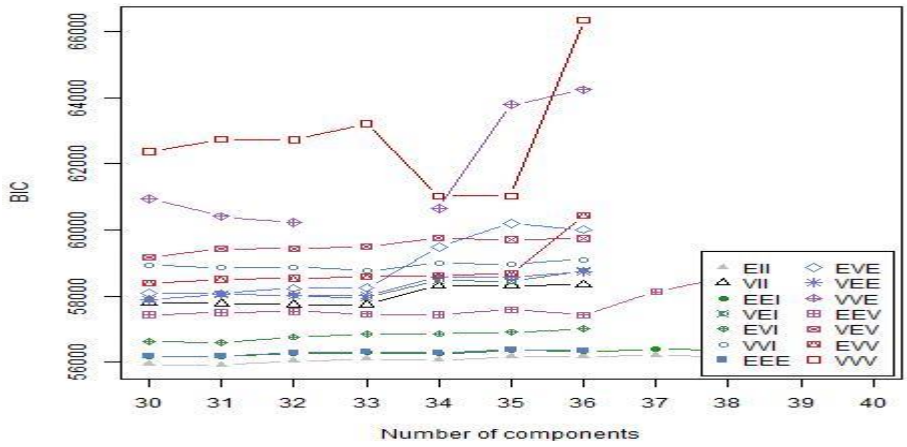


Fig 15. BIC plot for clusters (3-40)

Classification of Model VVV can be seen from the fig 15. Visualization of 9 components are shown in the graph Fig 17.

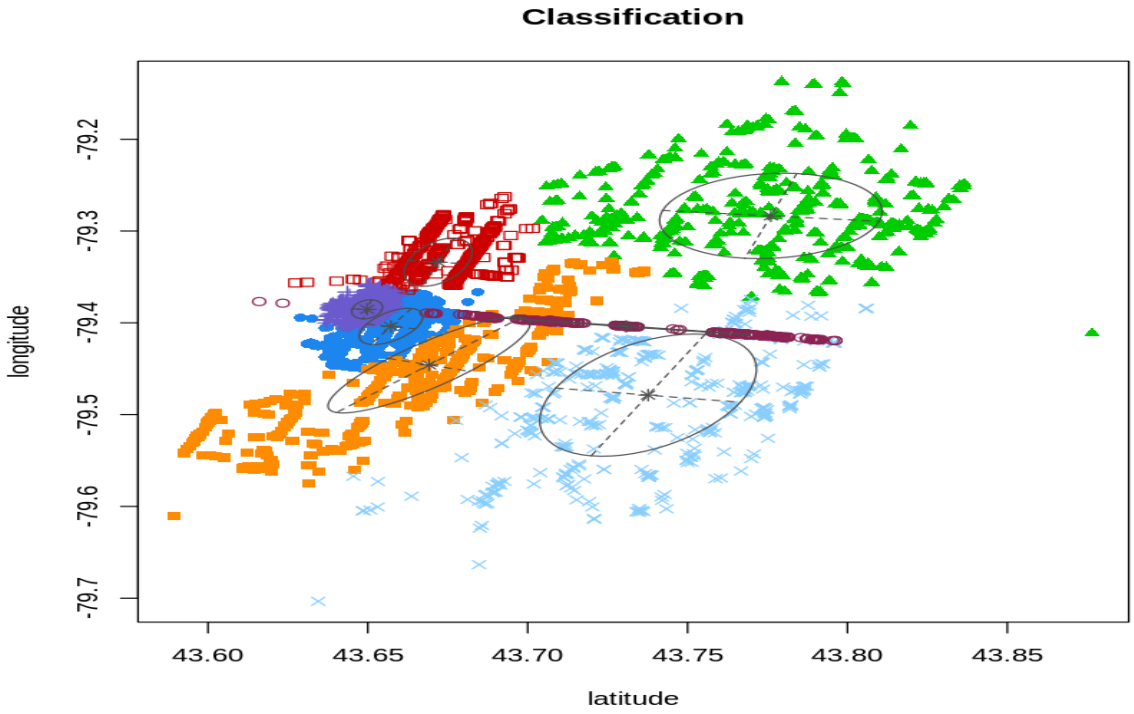


Fig 17 classification of clusters for default 9 clusters

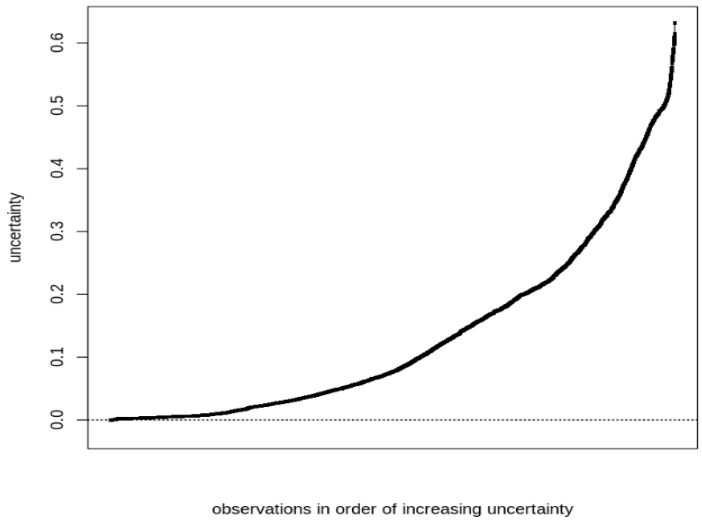


Fig 18 Uncertainty graph

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

Hypothesis1: using logistic regression with clusters fit, predicting whether the restaurant is open.

Interesting associations can be derived by testing the above hypothesis. Performing logistic regression in predicting the is_open using clusters fit.

The VVV model with 35 clusters fit is used as predictor variable to predict whether the restaurant is open or not.

model <- multinom(is_open~ cluster_fit, data =datacluster_neigh,MaxNWts = 845544).

From the below confusion matrix table, interesting fact can be seen that logistic regression is able to predict more open restaurant than closed restaurant since data is skewed. The precision of predicting open restaurant is more than predicting closed restaurants.

This pattern is observed is due to the data is skewed and cluster fit only relies on latitude and longitude.

Precision	Recall	F1
0.38547	0.46885	0.43208
0.6789	0.695016	0.820076

Hypothesis: using logistic regression with clusters fit, predicting the neighbourhood.

The above hypothesis can be tested by performing logistic regression with regularization in predicting the neighbourhood with clusters fit, latitude and longitude.

The VVV model with 35 clusters fit is used as predictor variable to predict whether the neighbourhood.

model <- multinom(neighbourhood ~ cluster_fit, data =datacluster_neigh,MaxNWts = 845544).

The accuracy of predicting the neighbourhood using cluster fit is 0.401371. Even though , cluster fit mainly based on latitude and longitude, prediction accuracy is low rather compared to is_open.

```
> sum(diag) / n
[1] 0.401371
```