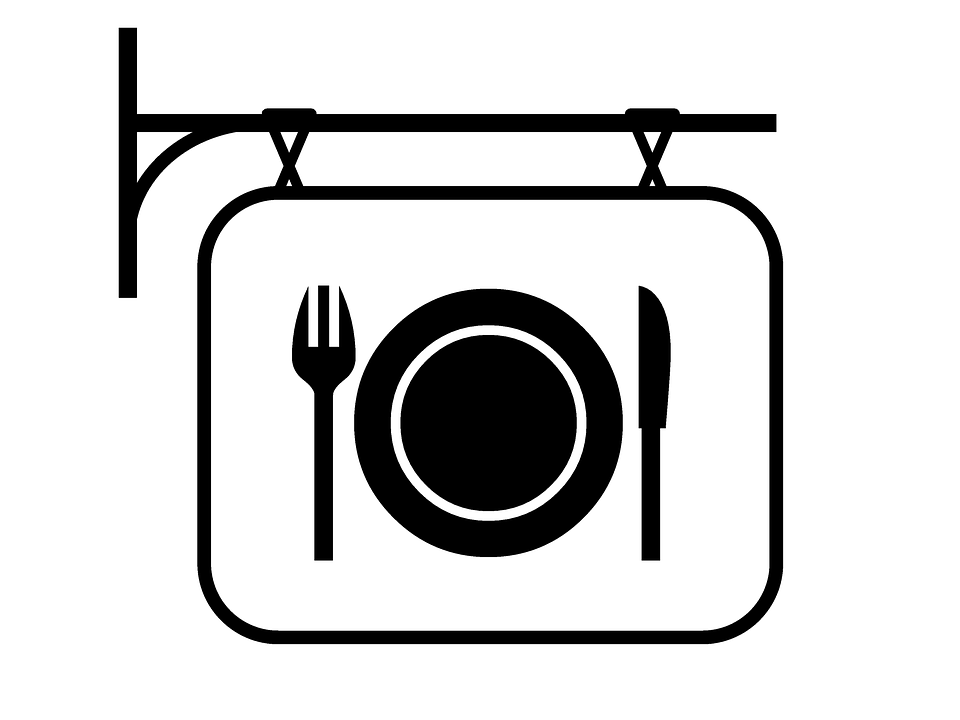
Food Establishment Recommender



CKME136 - Capstone report on food establishment dataset predictive and recommendation with supervised machine learning algorithm

**Presented By - Mohammed Amir**

**Supervised By – Dr. Tamer Abdou**

**Certificate in big data and predictive analytics**

**Ryerson University School of Continuing Education, 350 Victoria Street, Toronto Canada**

**April 17, 2017**

1. Introduction 4

2. Literature Review 5

3. Dataset Review 6

Dataset Adaptation 6

4. Methodology 9

Step 1: Define Objective 9

Step 2: Prepare & Explore Data 9

Step 3: Explorative Analyze Data 9

Step 4: Transform Data 9

Step 5: Develop Predictive Model & Outcome 9

Step 5: Create Recommendation 9

5. Data Exploration 11

5.1 Initial dataset description 11

5.2 Dataset Summary 11

 Dinesafe dataset Summary 11

 Address dataset Summary 12

5.3 Dataset Structure 12

 Dinesafe dataset Structure 12

 Address dataset Structure 12

5.4 Dataset Sample 13

A sample of the two datasets using a head function 13

 Dinesafe dataset sample 13

 Address dataset sample 13

5.5 Merged Dataset Summary 13

6. Data Munging 15

6.1 Remove duplicates 15

6.2 Missingness 15

6.3 Format Data Types 16

6.4 Describe Dataset 16

6.5 Impute Missing Values 17

7. Data Exploratory Analysis & Visualization 18

7.1 Univarient Data Analysis 18

7.1.1 Review Variable : The data is normally distrusted 18

7.2 Bivarient Data Analysis 21

7.2.1 Mean and Standard Deviation 21

7.3 Multivarient Data Analysis 26

8. Data Transformation 28

9. Predictive Analysis 30

9.1 Algorithm selection 30

9.2 Model Building 30

9.3 Prediction 31

10. Recommender System 32

11. Conclusion 33

12. Reference 34

# Introduction

In the digital world we live in, humans’ daily live is integrated to a digital technology in many different forms such as communication, entertainment, shopping, travel, social media etc. The common theme among technology base service providers is the reliance of a historical user data or/and product attributes in order to predict & recommend products and services to customers that are similar to the one that they are currently purchased. Recommender systems primary advantage is filtering a large set of data, item and/or product in order to provide much relevant and personal service customers in order to enhance their experience.

For the capstone project a city of Toronto Dinesafe food hygiene dataset in combination of yelp and travel advisor websites food premises customer review & rating data to create a predictive and recommender system.

# Literature Review

In the information age we live in, the creation of data grows exponentially and not all the data is in a structured format. For individuals to shift through this large data in order to retrieve a relevant information that is suitable for their consumption is time consuming and tedious.

Information scientist developed a technique using statistics, machine learning and sentiment analysis to identify a relationship between items in order to provide a richer experience for users by providing only relevant information.

The primary articles that was reviewed in preparation of the capstone are

1. An Introduction to Recommendation Systems in Software Engineering by Martin P. Robillard and Robert J. Walker
2. Amazon.com recommendation, Item to Item collaborative filtering by Greg Linden, Brent Smith & Jeremy York
3. A Literature Survey on Recommendation System Based on Sentimental Analysis by Achin Jain, Vanita Jain and Nidhi Kapoor
4. Incorporating popularity in a personalized news recommender system by Nirmal Jonnalagedda, Susan Gauch, Kevin Labille and Sultan Alfarhood
5. Algorithms and Methods in Recommender Systems by Daniar Asanov
6. Basic Approaches in Recommendation Systems by Alexander Felfernig, Michael Jeran, Gerald Ninaus, Florian Reinfrank, Stefan Reiterer, and Martin Stettinger

There are three main techniques recommender systems are implemented on

1. **Collaborative Filtering:** This is a domain independent technique that analysis users profile attribute against item attributes to generate a recommendation. Recommendation is provided based on a similarity of user profiles and item profile using historical preference data.

Collaborative filtering is considered as the most basic and easiest recommender system technique. The disadvantage of this technique is with a cold start, this refers to lack of user profile data when users are new with no existing profile in the recommender system.

1. **Content Based Filtering:** This is a domain dependent technique that analysis attributes of items in order to generate a recommendation. This technique is used when there is a cold start, where the user has no profile. The recommendation depends on attribute similarities between items with no user profile input; therefore it is capable of recommending items to users that are new or has no historic data.

The second advantage of content based filtering is that, the technique is good in handling data sparsity, data sparsity refers to a lack of user rating or reviews on items. The disadvantage of this technique is when there is no enough item attributes, it fails to recommend the item to a user.

1. **Hybrid Filtering:** This technique is a combination multiple techniques such as collaborative, content based & context based techniques to take the strength of both techniques and improve the performance of the recommendation.

# Dataset Review

As part of City of Toronto Open Data Initiative, the Toronto Public Health food safety inspection DineSafe data is available online for public use and this dataset will be used in this exercise.

<http://www.toronto.ca/health/dinesafe/index.htm>

In this project a subset of the **dinesafe** dataset has over 16,199 rows of historical inspection result, with 2,715 food premises for the year 2015 and 2016. The data attributes and description are provided below.

|  |  |
| --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** |
| ROW\_ID | Represents the Row Number |
| ESTABLISHMENT\_ID | Unique identifier for an establishment |
| INSPECTION\_ID | Unique identifier for each Inspection |
| ESTABLISHMENT\_NAME | Business name of the establishment |
| ESTABLISHMENTTYPE | Establishment type ie restaurant, mobile cart |
| ESTABLISHMENT\_ADDRESS | Municipal address of the establishment |
| ESTABLISHMENT\_STATUS | Pass, Conditional Pass, Closed |
| MINIMUM\_INSPECTIONS\_PERYEAR | Every eating and drinking establishment in the City of Toronto receives a minimum of 1, 2, or 3 inspections each year depending on the specific type of establishment, the food preparation processes, volume and type of food served and other related criteria |
| INFRACTION\_DETAILS | Description of the Infraction |
| INSPECTION\_DATE | Calendar date the inspection was conducted |
| SEVERITY | Level of the infraction, i.e. S – Significant, M – Minor, C – Crucial |
| ACTION | Enforcement activity based on the infractions noted during a food safety inspection |
| COURT\_OUTCOME | The registered court decision resulting from the issuance of a ticket or summons for outstanding infractions to the Health Protection and Promotion Act |
| AMOUNT\_FINED | Fine determined in a court outcome |

## Dataset Adaptation

Dinesafe dataset is suitable for a predictive analytics, however it doesn’t contain any customer oriented attributes such as user profile, rating, postal code and other attributes that are necessary for recommender system.

In order to adopt the data for an enhanced analytics & recommender system, customer rating, dollar value and cuisine type information was added to the dataset manually based on yelp and travel adviser customer rating. Also zip code and district information was extracted from google geocode for all the premesis based on their street address.

|  |  |
| --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** |
| ROW\_ID | Represents the Row Number |
| ESTABLISHMENT\_ID | Unique identifier for an establishment |
| INSPECTION\_ID | Unique identifier for each Inspection |
| ESTABLISHMENT\_NAME | Business name of the establishment |
| REVIEW | Customer satisfaction rating (1-5), 1 low, 5 high |
| VALUE | Value for money (1 - 5), 1 cheap , 5 expensive |
| CUISINE TYPE | Cuisine Type such as North American, European, African, Latin American, South Asian, Far Eastern etc… |
| ESTABLISHMENTTYPE | Establishment type ie restaurant, mobile cart |
| ESTABLISHMENT\_ADDRESS | Municipal address of the establishment |
| ESTABLISHMENT\_STATUS | Pass, Conditional Pass, Closed |
| MINIMUM\_INSPECTIONS\_PERYEAR | Every eating and drinking establishment in the City of Toronto receives a minimum of 1, 2, or 3 inspections each year depending on the specific type of establishment, the food preparation processes, volume and type of food served and other related criteria |
| INFRACTION\_DETAILS | Description of the Infraction |
| INSPECTION\_DATE | Calendar date the inspection was conducted |
| SEVERITY | Level of the infraction, i.e. S – Significant, M – Minor, C – Crucial |
| ACTION | Enforcement activity based on the infractions noted during a food safety inspection |
| COURT\_OUTCOME | The registered court decision resulting from the issuance of a ticket or summons for outstanding infractions to the Health Protection and Promotion Act |
| AMOUNT\_FINED | Fine determined in a court outcome |
| ADDRESS | Full premises address |
| DISTRICT | Toronto district (Metro Toronto, York, North York, East York,  Etobicoke, Scarborough) |
| CITY | Toronto |
| POSTAL CODE | Toronto postal codes |

# Methodology

In this project R language on RStudio was used in the implementation of project. The procedure that was followed in the analysis and development of a proof of concept is outlined below

## Step 1: Define Objective

* + The objective of this investigation is to produce an efficient mechanism to predict & recommend food premises such as restaurants, coffee shops, deli, bakery across Toronto based on Toronto Public Health historical DineSafe inspection dataset.

## Step 2: Prepare & Explore Data

* + Collect and explore dataset
  + Clean dataset by removing institutions, convenience stores, groceries, schools etc
  + Create data consistency by removing typo errors, missing
  + Identify missing attributes & retrieved from yelp, traveladvisor & google
  + Merge missing attributes with the dinesafe dataset

## Step 3: Explorative Analyze Data

* + Analyze data structure, missingness, dimension & description
  + Perform univariant data analysis
  + Perform bivariant data analysis
  + Perform multivariant data analysis

## Step 4: Transform Data

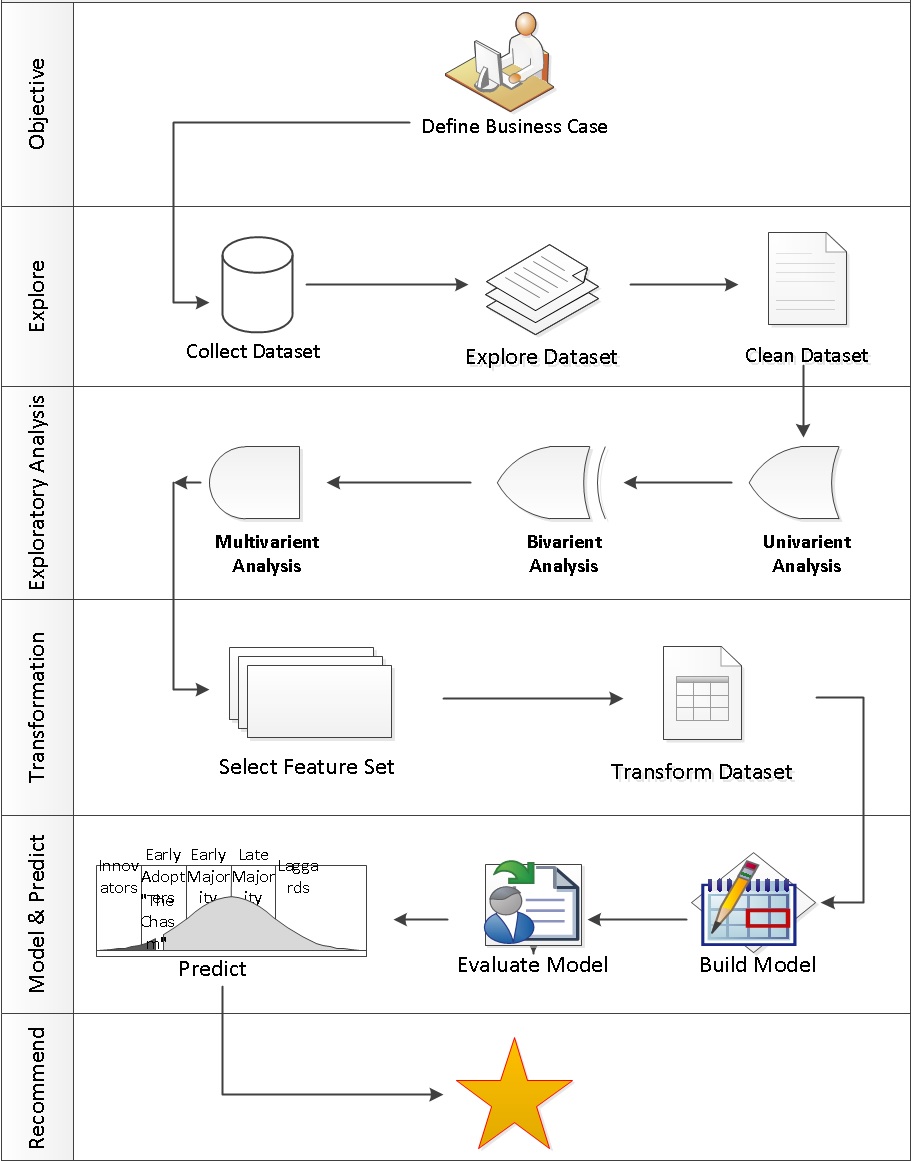
* + Define dataset as supervised or non-supervised algorithm
  + Analyze predictive & recommender algorithms to use
  + Remove duplicate premises data
  + Select labels from subset of the dataset
  + Transform nominal categorical data into a numerical nominal value
  + Normalize the data

## Step 5: Develop Predictive Model & Outcome

* + Select machine learning algorithm
  + Split data into training and testing
  + Cross validation dataset
  + Build a model
  + Evaluate & validate the model
  + Calculate model accuracy
  + Improve accuracy
  + Apply model on test dataset & observe outcome

## Step 5: Create Recommendation

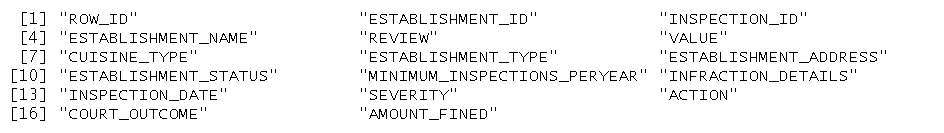
* + Build a recommender model
  + Apply data set on a recommender model and validate the prediction result.



# Data Exploration

## Initial dataset description

* Dinesafe dataset

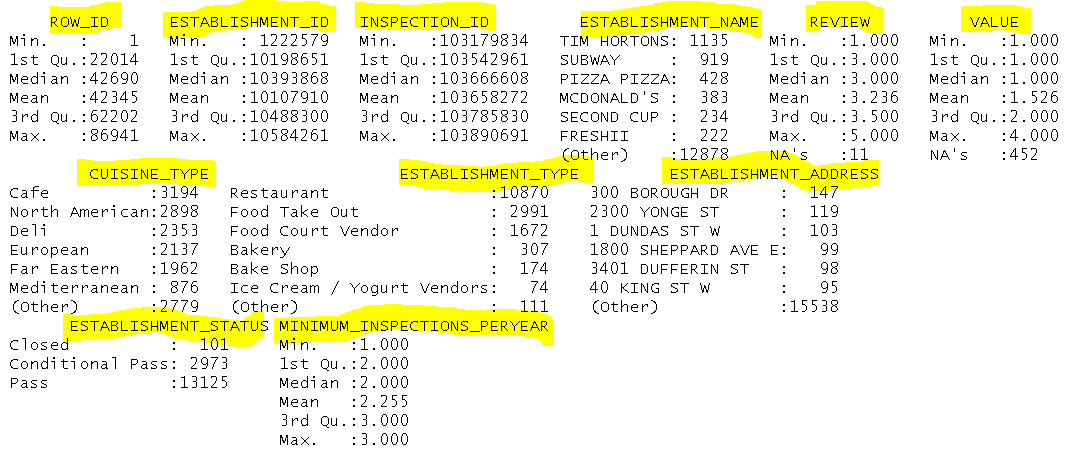


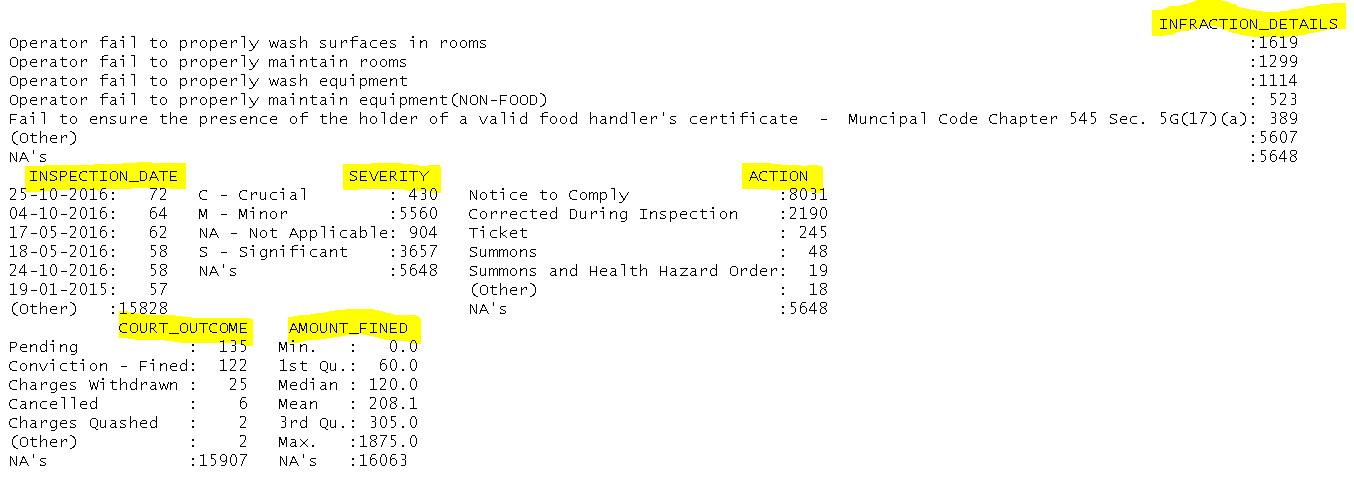
* Address dataset



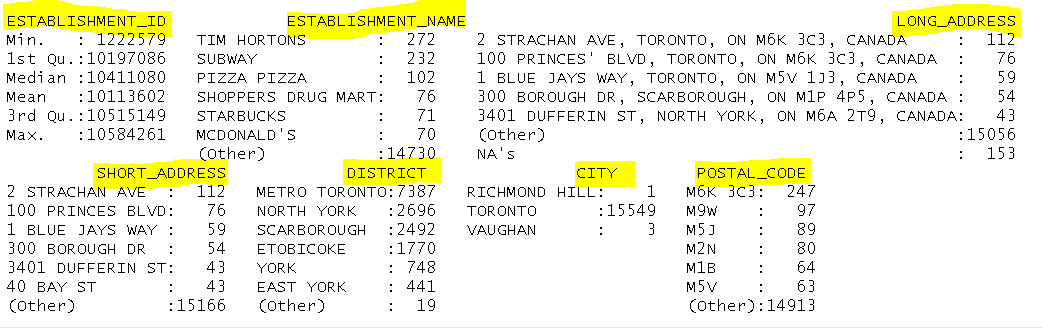
## Dataset Summary

### Dinesafe dataset Summary





### Address dataset Summary



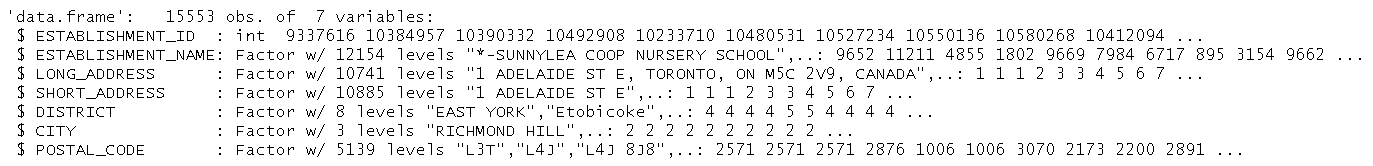
## Dataset Structure

Data structure of Dinesafe and Address datasets which has numeric and factor values

### Dinesafe dataset Structure

### 

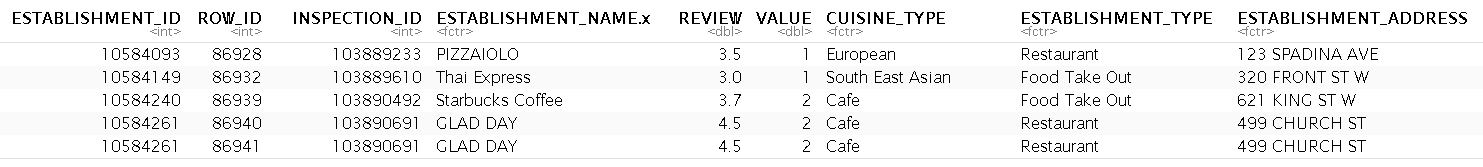
### Address dataset Structure

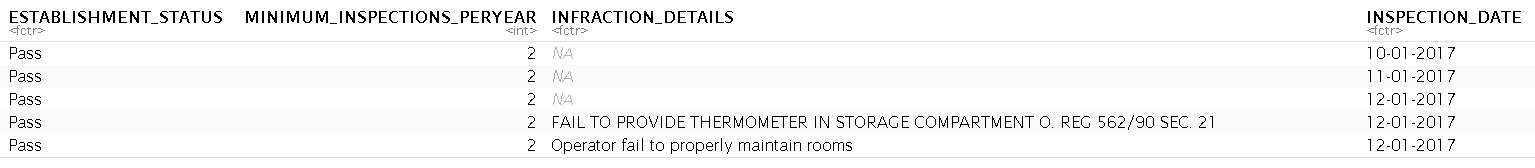


## Dataset Sample

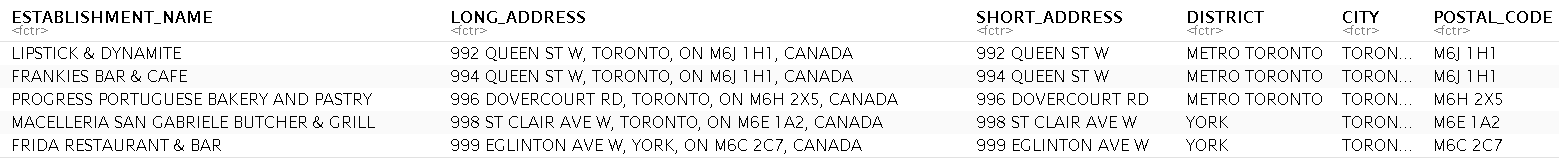
### A sample of the two datasets using a head function

### Dinesafe dataset sample



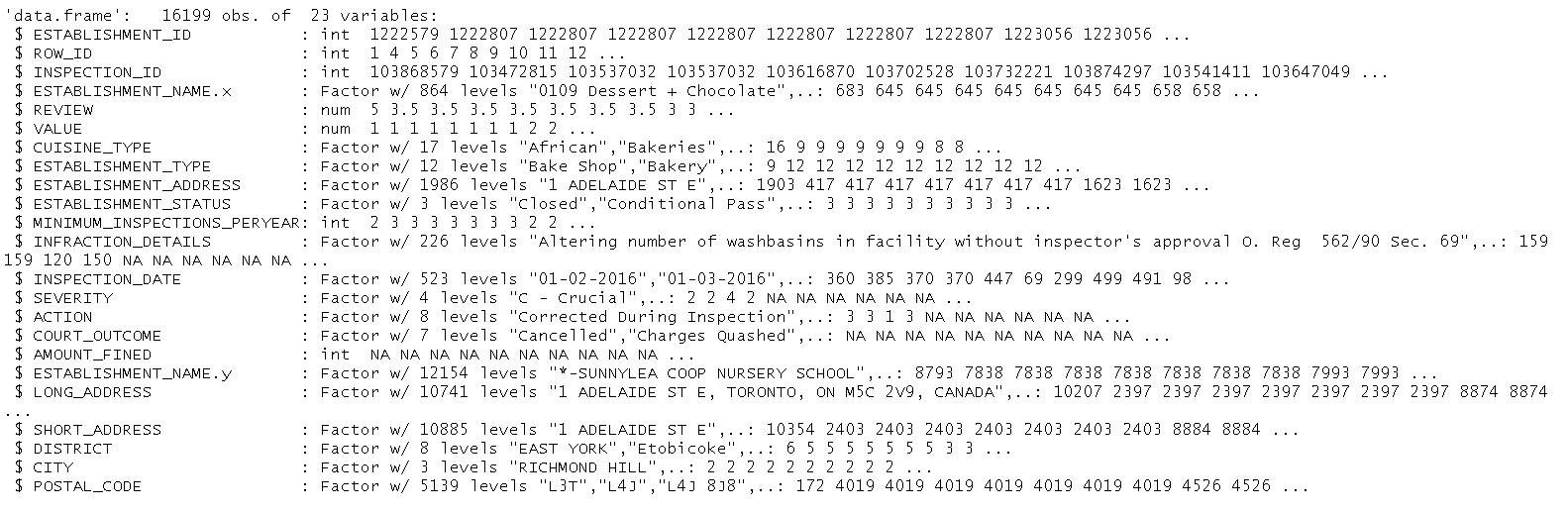


### Address dataset sample

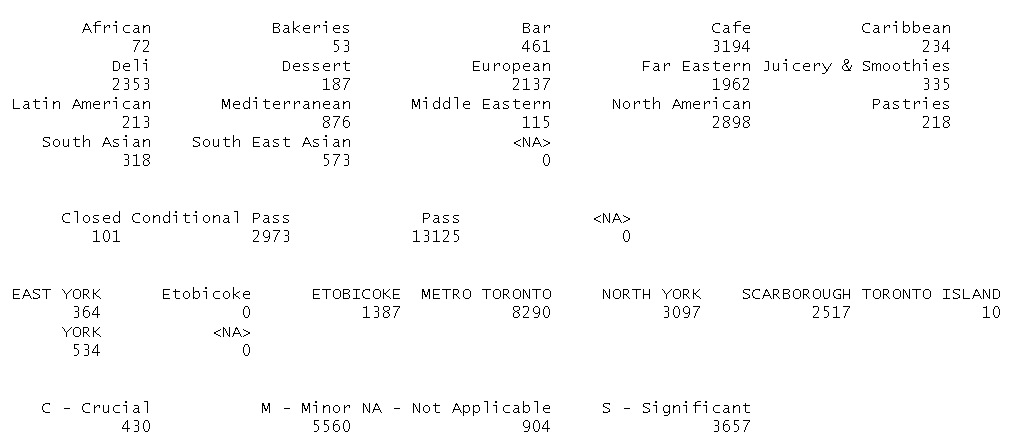


## Merged Dataset Summary

The dinesafe and address datasets were merged based on establishment id. The new dataset structure includes establishment information, inspection outcome and geographical location.



Explore cuisine type, inspection outcome & its severity and establishment location.



# Data Munging

## Remove duplicates

* + Remove duplicate columns from the two dataset merger such as “establishment name” & “establishment address”

*Dinesafe <- subset(Dinesafe, select = -c(ESTABLISHMENT\_NAME.y, ESTABLISHMENT\_ADDRESS) )*

* + Remove data columns that are not relevant to the analysis such as “court outcome”, “amount fined” and “infraction detail”

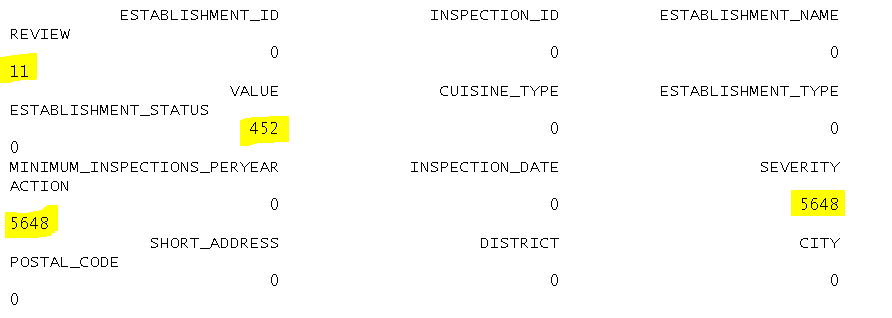
*Dinesafe <- subset(Dinesafe, select = -c(ROW\_ID, COURT\_OUTCOME,AMOUNT\_FINED,LONG\_ADDRESS, INFRACTION\_DETAILS) )*

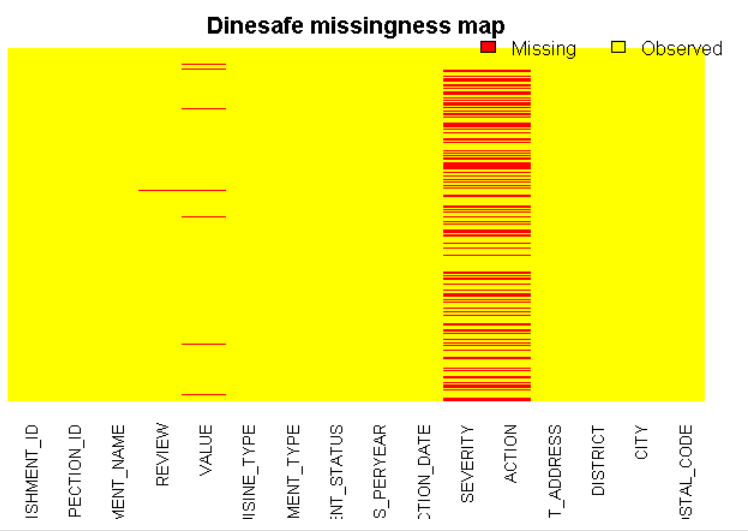
* + *Rename “establishment\_name.x” to “establishment\_name”*

*colnames(Dinesafe)[colnames(Dinesafe) == 'ESTABLISHMENT\_NAME.x'] <- 'ESTABLISHMENT\_NAME'*

## Missingness

* + *Identify & quantify missingness in the dataset, the “review”, “value”, “action” and “severity” columns has missing values that need to be imputed. This is represented in the missmap graph shown below in red using the Amelia package.*





## Format Data Types

* + Convert Action column from factor to character type to avoid error during data imputation

Dinesafe$ACTION = as.character(Dinesafe$ACTION)

* + Set Categorical Data Type Level for Establishment Status column

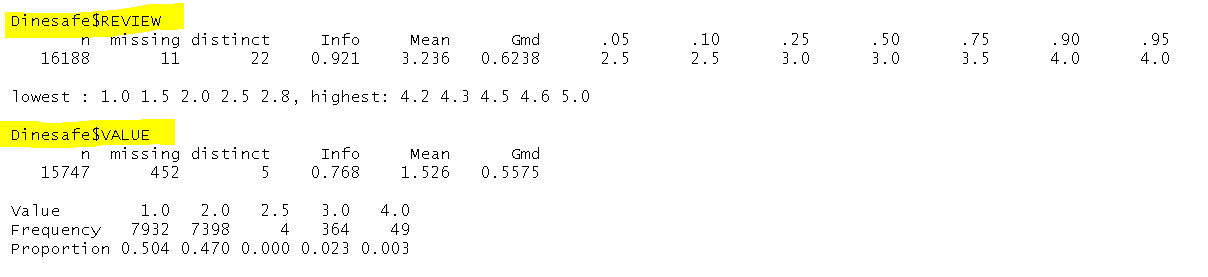
Dinesafe$ESTABLISHMENT\_STATUS = factor(Dinesafe$ESTABLISHMENT\_STATUS,levels=c("Closed","Conditional Pass", "Pass"))

* + Set Categorical Data Type Level for Severity column

Dinesafe$SEVERITY <- factor(Dinesafe$SEVERITY, levels = c("NA - Not Applicable", "N - No Action", "M - Minor", "S - Significant", "C - Crucial"))

## Describe Dataset

* + Describe quantitative values in “Review” and “Value” columns using HMISC library



Identify complete rows with no missing (NA) value using complete case function returning 10195 rows.

* + Complete\_Dinesafe <- Dinesafe[complete.cases(Dinesafe),]
  + nrow(Complete\_Dinesafe)

## Impute Missing Values

Impute missing values in “review”, “value”, “severity” & “action” columns

* + Impute “Review” column using the mean review value for the specific cuisine type, the below script demonstrates this for an “African” cuisine type

Dinesafe$REVIEW[is.na(Dinesafe$REVIEW) & Dinesafe$CUISINE\_TYPE=="African"] = mean(Dinesafe$REVIEW[Dinesafe$CUISINE\_TYPE=="African"], na.rm=TRUE)

* + Impute “Value” column using the mean value for the specific cuisine type, the below script demonstrates this for an “African” cuisine type

Dinesafe$VALUE[is.na(Dinesafe$VALUE) & Dinesafe$CUISINE\_TYPE=="African"] = mean(Dinesafe$VALUE[Dinesafe$CUISINE\_TYPE=="African"], na.rm=TRUE)

* + In Severity column, the only missing values were for “Pass” establishment status, therefore the missing value in severity column was imputed with “Not applicable”

Dinesafe$SEVERITY[is.na(Dinesafe$SEVERITY) & Dinesafe$ESTABLISHMENT\_STATUS == "Pass"] = "NA - Not Applicable"

* + In Action column, the only missing values were for “Pass” establishment status, Dinesafe$ACTION[is.na(Dinesafe$ACTION) & Dinesafe$ESTABLISHMENT\_STATUS == "Pass" & Dinesafe$SEVERITY == "NA - Not Applicable"] = "No Action Required"

Finally checking for incompleteness it returns zero value confirming there is no missing data.

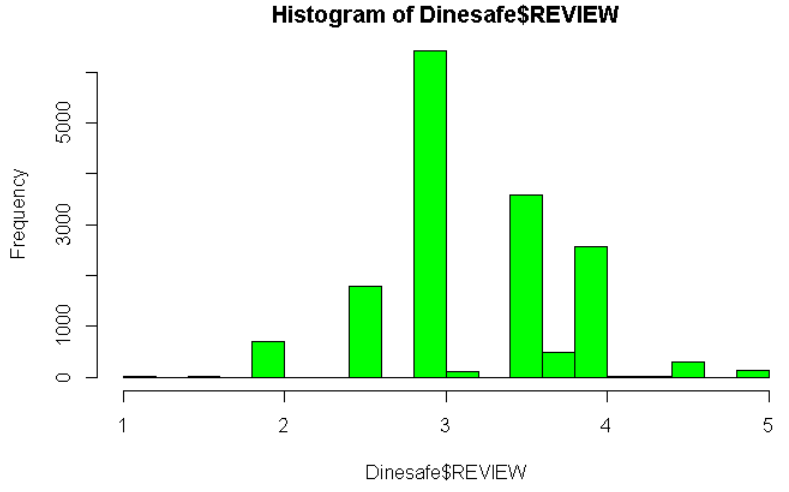
* + Dinesafe\_NA <- Dinesafe[!complete.cases(Dinesafe),]
  + nrow(Dinesafe\_NA)

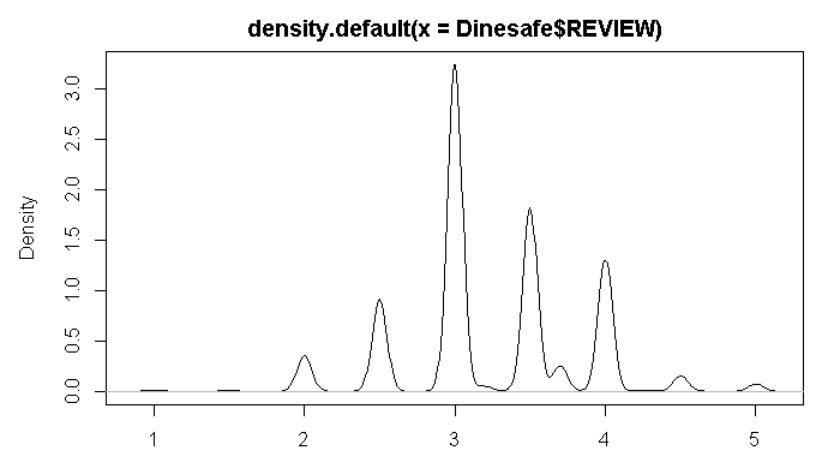
# Data Exploratory Analysis & Visualization

## Univariate Data Analysis

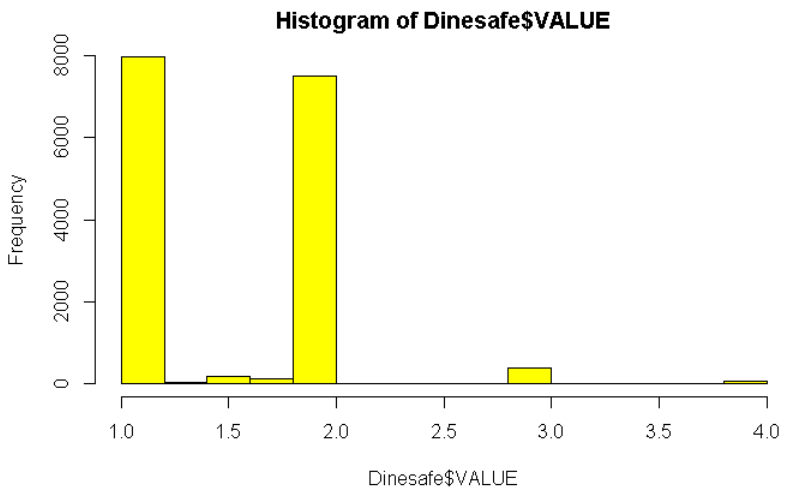
In this section a single variable from the dataset was analyzed to understand the data using histogram and density graphical representation

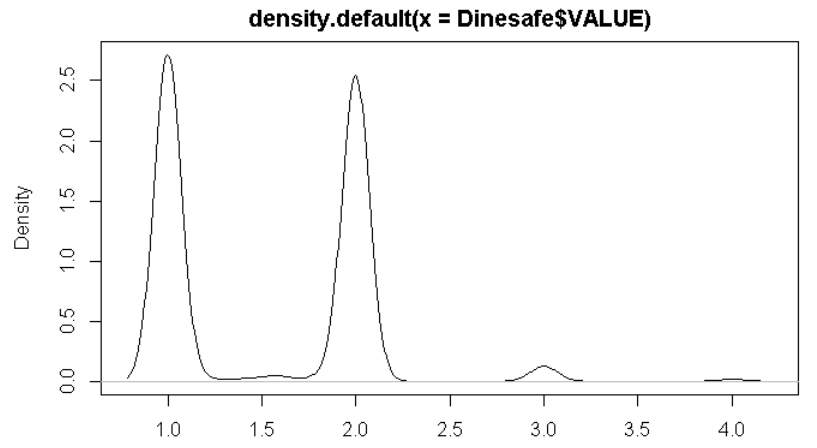
### Review Variable : The data is normally distrusted



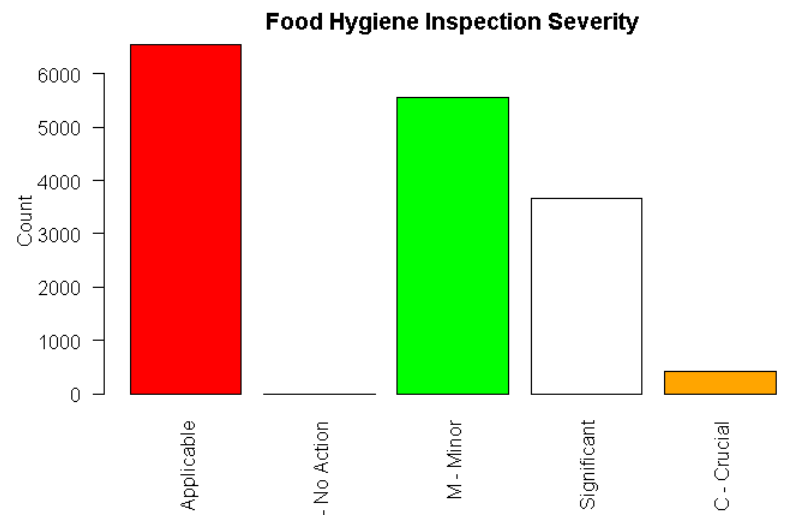


* + 1. Value Variable : The data is skewed to the right

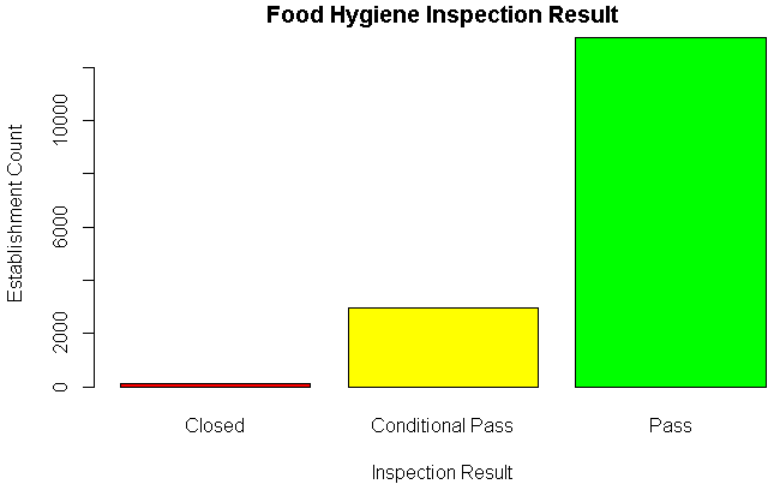




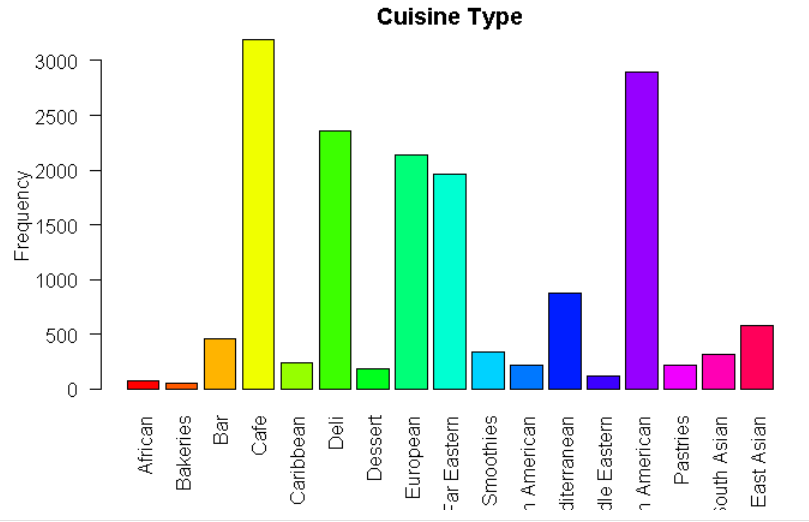
* + 1. Food inspection severity graph



* + 1. Food Hygiene Inspection Result

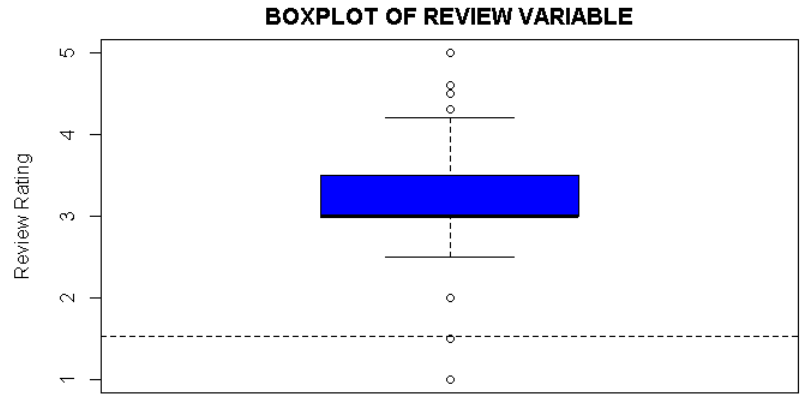


* + 1. Establishment Cuisine Type



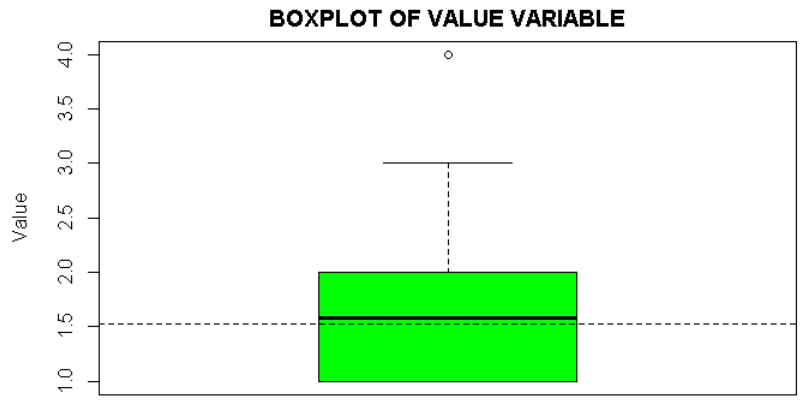
* + 1. Boxplot of Review Variable

The data graph shows that the mean and median values (Horizontal dot line) are far apart and most of the values are lying between 3 and 3.5 with outlier value below 2.5 and above 3.5



* + 1. Boxplot of Value Variable

The data graph shows that the mean and median values (Horizontal dot line) are close to each other at 1.5 and most of the values are lying between 1 and 2 with outlier value at 4.

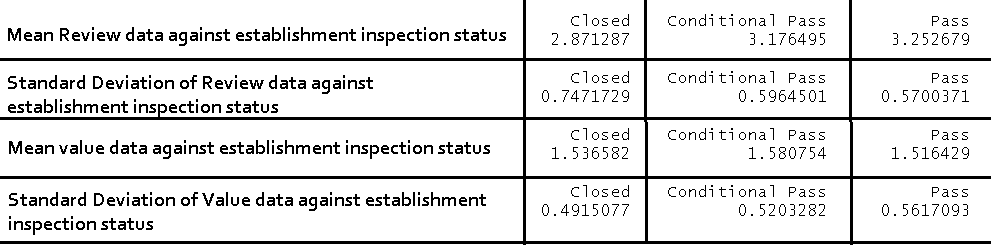


## Bivariate Data Analysis

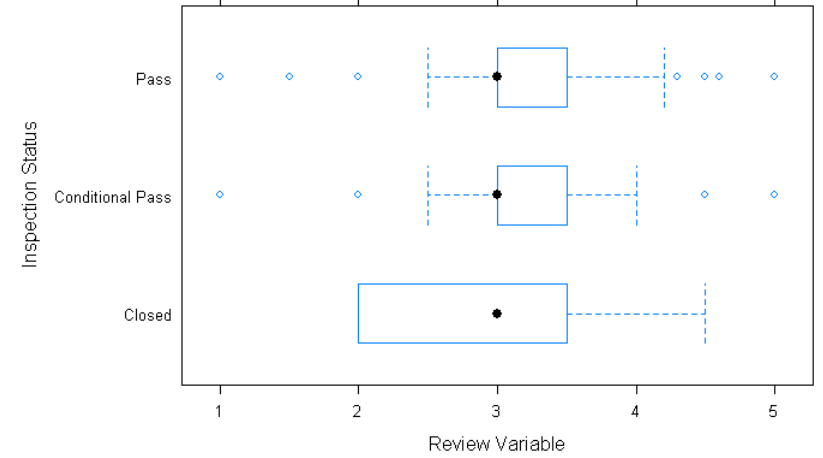
### Mean and Standard Deviation

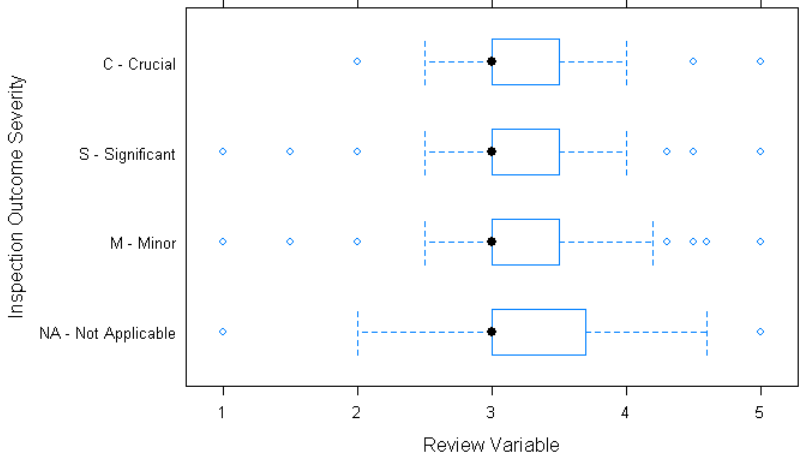
The mean and standard deviation value of food premises that failed inspection had a mean review value below those that passed inspection. Also failed food premises had a higher standard deviation value as compared to those who passed.

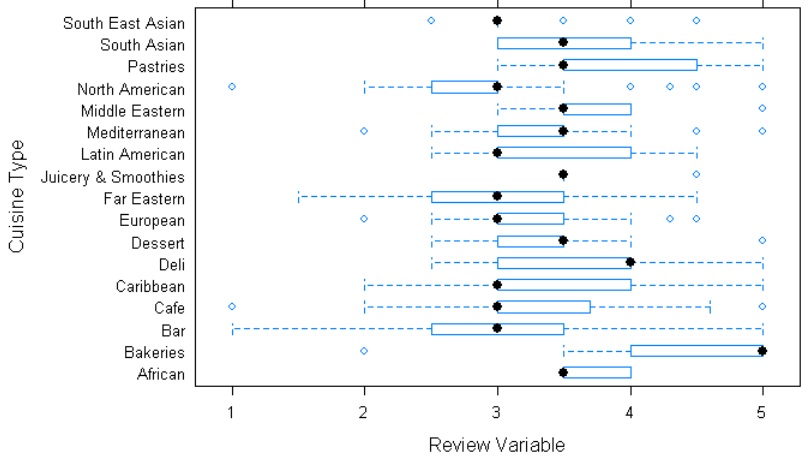
On the other hand the relationship between mean/standard deviation value and inspection outcome is not observer due to consistent result across all three values.



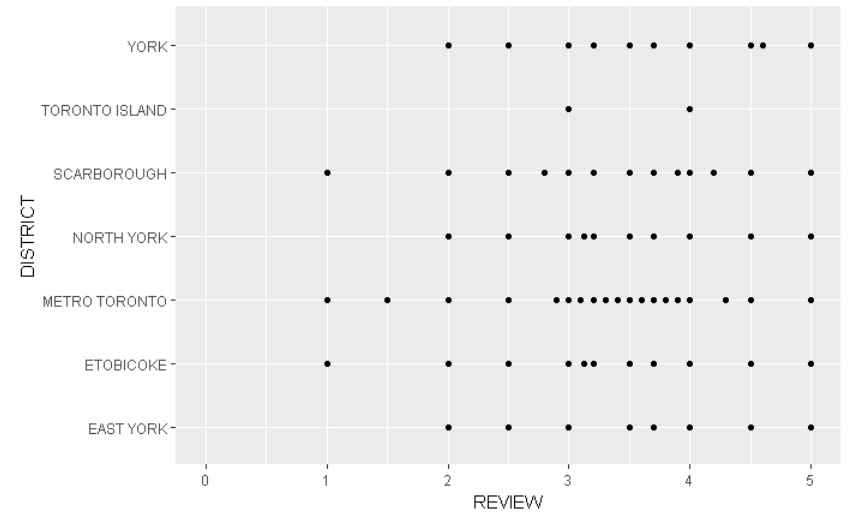
* + 1. Categorical vs Numerical data analysis using lattice package

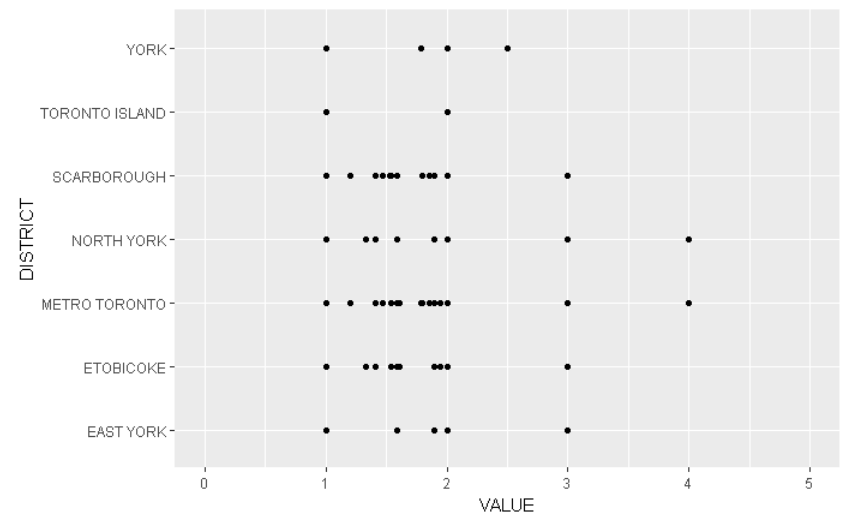




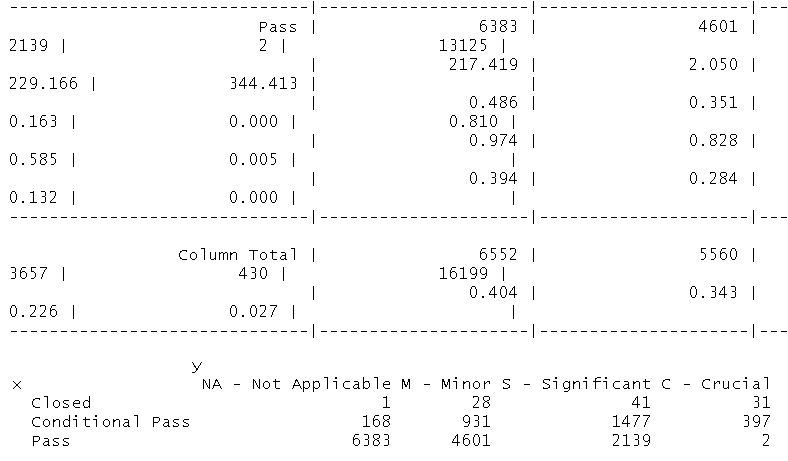


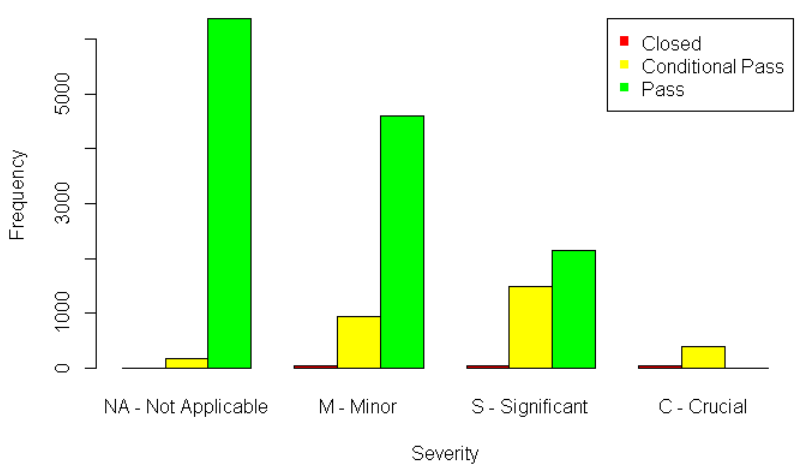
* + 1. Categorical vs Numerical data analysis using ggplot2 package





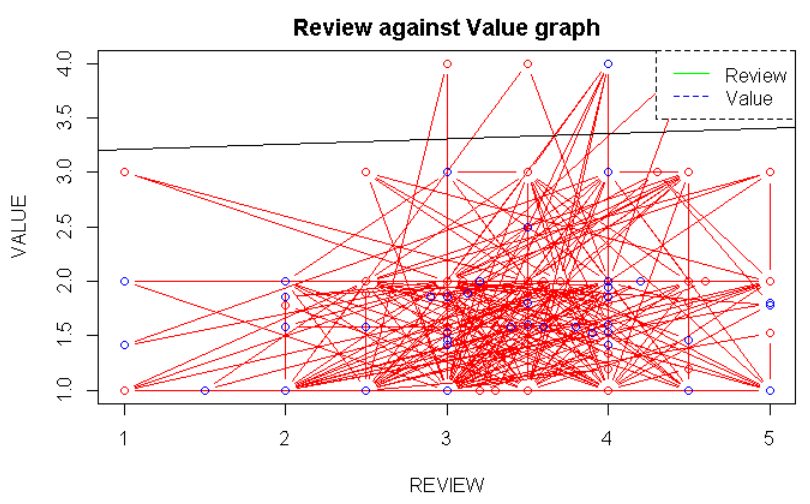
* + 1. Crosstab analysis of “Severity” and “Inspection Status” analysis with Crosstab & barplot





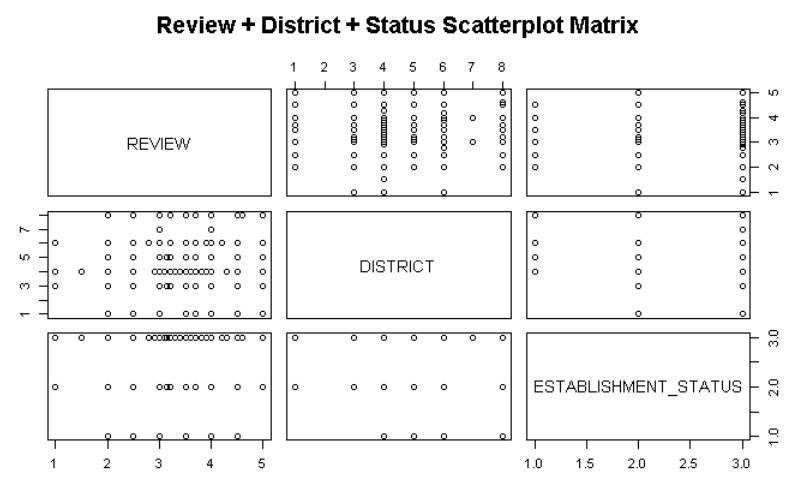
* + 1. Relationship between “Review” and “Value” using scatter plot

As shown on the graph below there is no linear relationship between a restaurant review and value for money variables. The values are scattered all over the box and doesn't follow the simple linear regression model line.

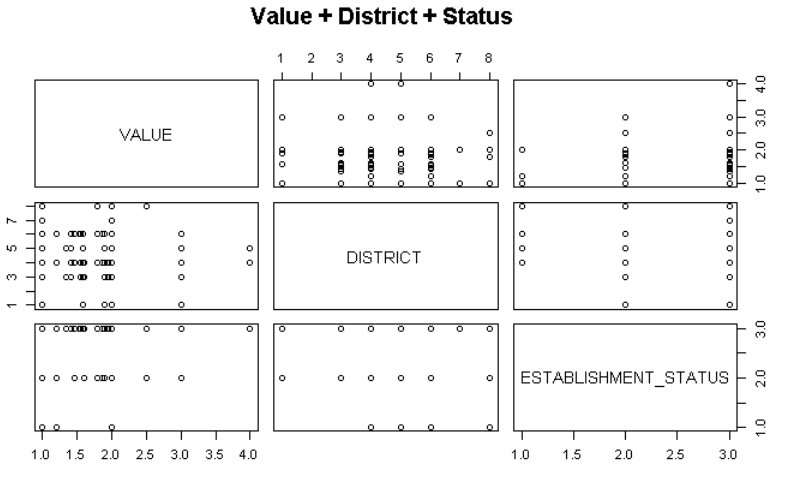


## Multivariate Data Analysis

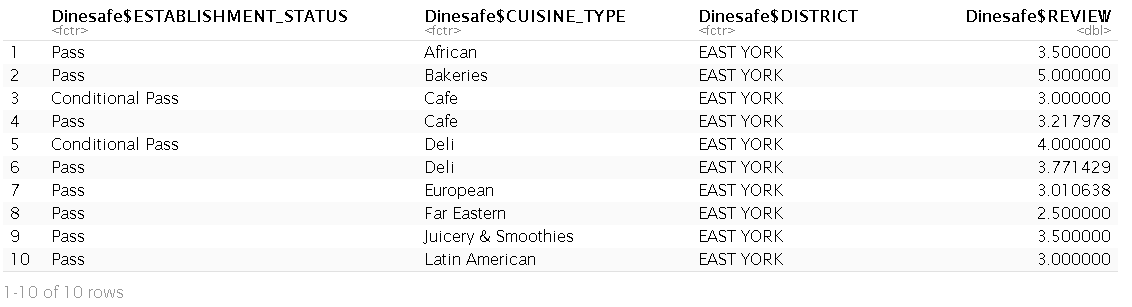
* + 1. Simple Scatter Matrix on the relationship between Review, District and Status variables



* + 1. Simple Scatter Matrix on the relationship between Value, District and Status variables

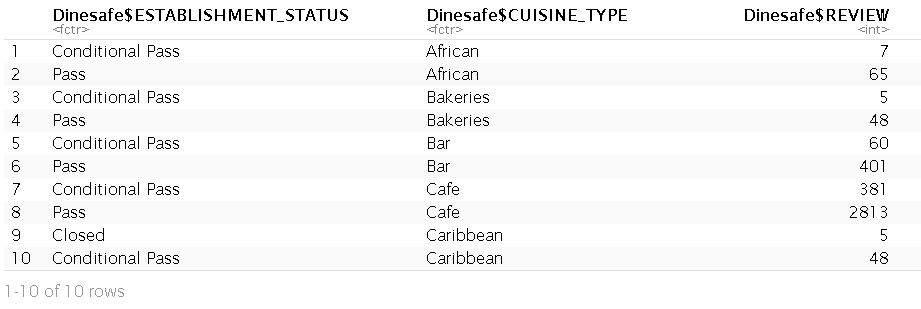


* + 1. Aggregation of “Review” against “Status”, “Cuisine Type” and “District as categorical and numerical values.

head(aggregate(Dinesafe$REVIEW ~ Dinesafe$ESTABLISHMENT\_STATUS + Dinesafe$CUISINE\_TYPE + Dinesafe$DISTRICT, FUN=mean),10) 

* + 1. Aggregation of “Review” against “Status” and “Cuisine Type” values as categorical and numerical values.

head(aggregate(Dinesafe$REVIEW ~ Dinesafe$ESTABLISHMENT\_STATUS + Dinesafe$CUISINE\_TYPE, FUN=length),10)



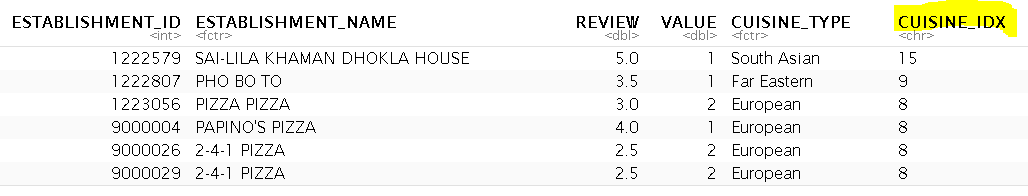
# Data Transformation

Following data exploration and analysis, the next step will be to perform data transformation in preparation for prediction and recommendation. The transformation processes are

* 1. Selected appropriate variables to create a feature

The labels we are important for prediction and recommendation are “Establishment ID”, “Establishment Name”, “Review”, “Value” and “Cuisine Type”. The rest of the labels are not relevant in creating an attribute of the establishment.

* 1. Created a unique rows based on the selected features. This reduces the number of rows from 16,199 to 2723
  2. Using “if else” function changed the “Cuisine Type” label from qualitative nominal value to quantitative nominal value Transform labels to the appropriate data type ranging from 1 to 17. These values are not ordinal and are treated as an index and it will be used an input to a predictive algorithm since only numeric values are accepted.

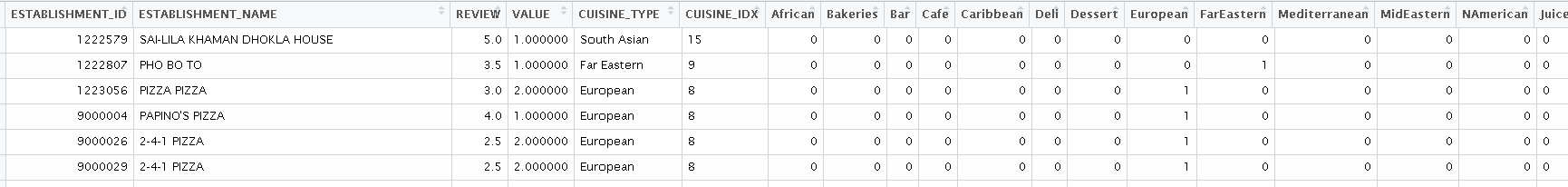


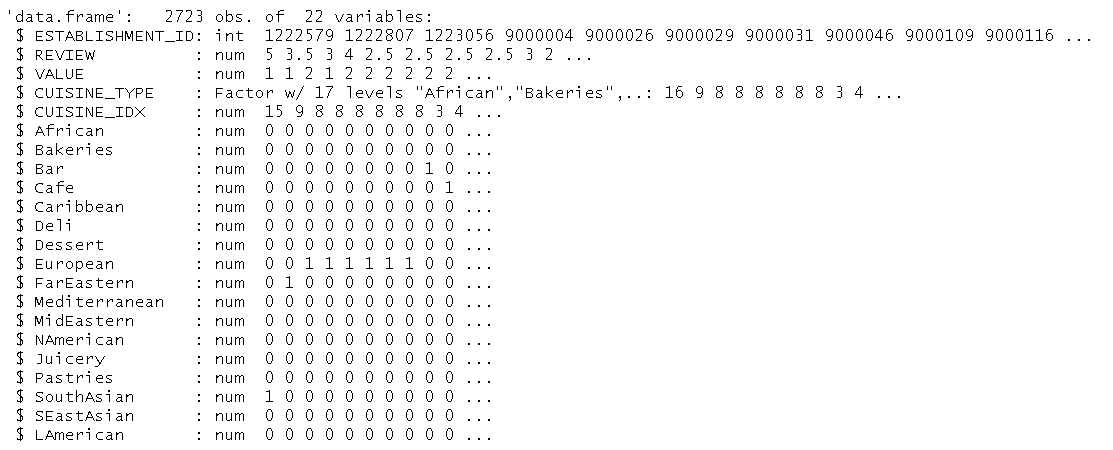
* 1. Changed CUISINE\_TYPE from factor to numerical value

Dinesafe2$CUISINE\_IDX <- as.numeric(Dinesafe2$CUISINE\_IDX)

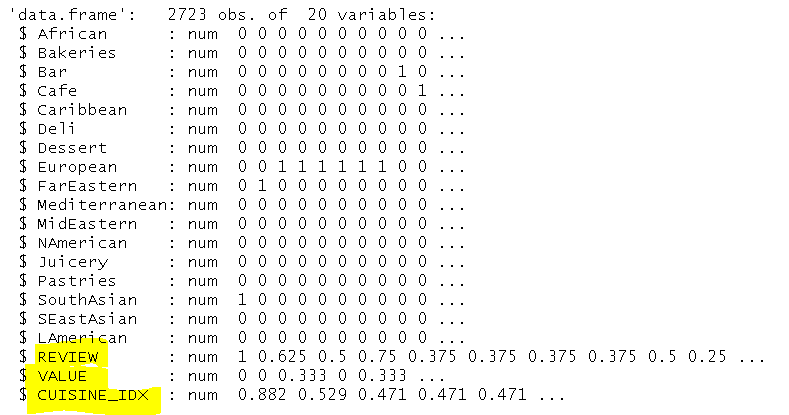
* 1. Created binary values for the “Cuisine Type” in order to create a binary attributes

Dinesafe2$African <- ifelse(Dinesafe2$CUISINE\_TYPE == "African",1,0)





* 1. Normalize the “Review”, “Value” & “Cuisine\_Idx” labels



* 1. Randomly split the dataset into two for training and testing to be used in predictive analysis

# Predictive Analysis

Predictive analysis is a process of making prediction of an outcome based on existing features using historical data. The data analysis, cleansing and transformation steps that were applied in the earlier steps are used in this predictive step

## Algorithm selection

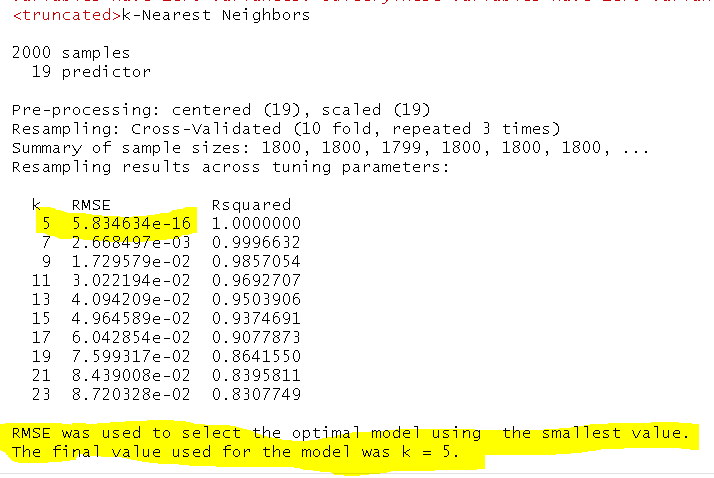
The primary objective in this task is to classify the food premises into a number of classes based on its attributes such as the cuisine type. This scenario is a good example of a supervised learning algorithm since the outcome value is provided during the training phase.

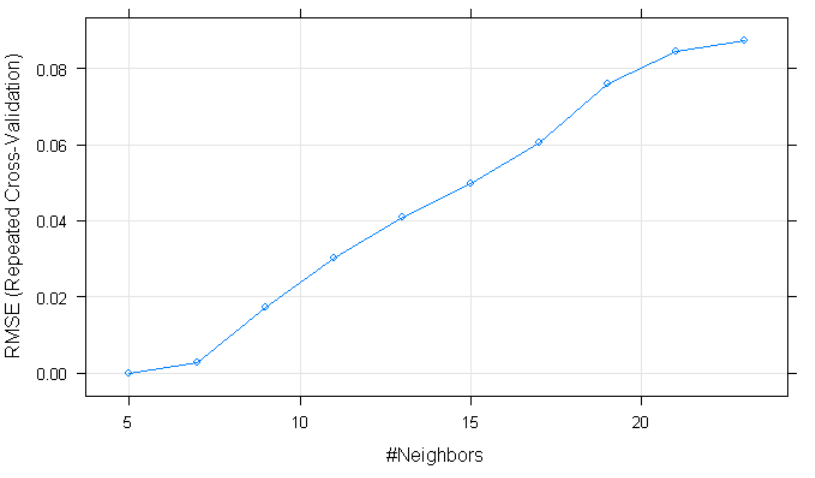
## Model Building

The first phase of building the KNN model is to perform a cross validation to determine the optimum K value for the given dataset in order to create a more accurate outcome. 10 fold cross validation with three repeats and the outcome was plotted.

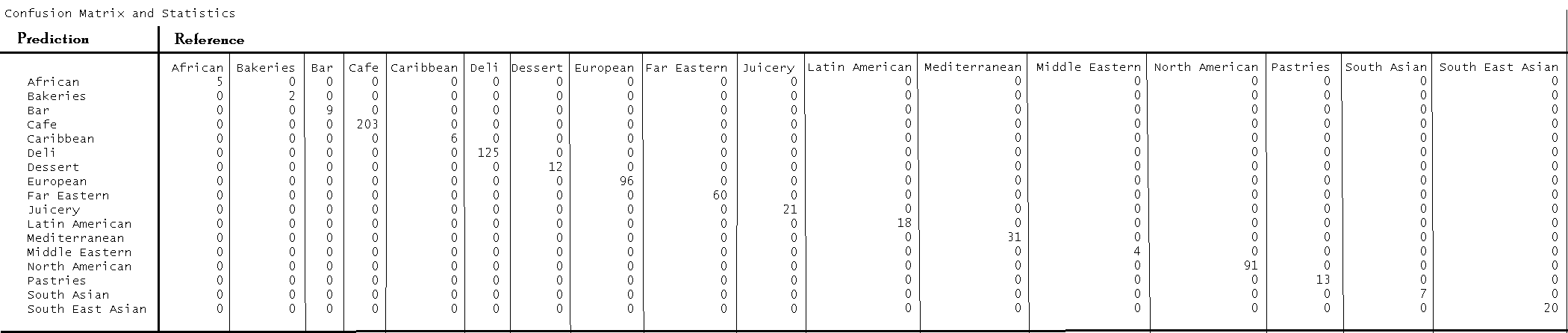
Caret and Class package were used to build the model

The smallest RMSE value indicates the most optimized K values to use and as shown below K = 5 was selected



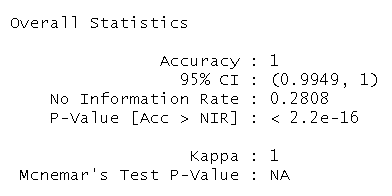


## Prediction



Classification accuracy in KNN is a ration of correct prediction to a total prediction made. To measure the accuracy of our result “confusion matrix” is applied in order to summarize the prediction result.

Prediction result is 100%



# Recommender System

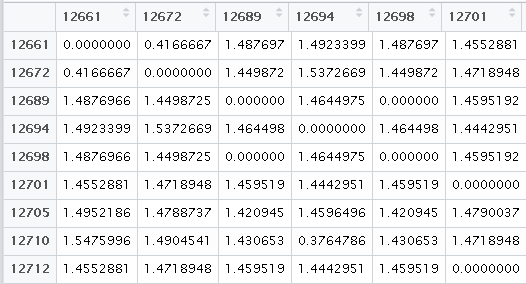
The most common technique used in a recommender system to identify similarity between two items feature vectors.

The most common technique used in a recommender system is identifying similarity between two items feature vectors based on how close it is distance. The smaller the distance implies a higher similarity.

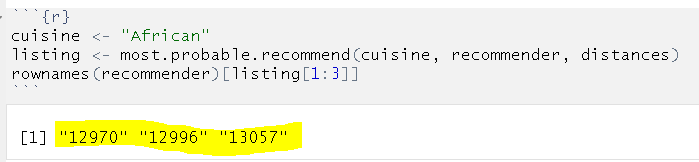
The distance between two items is calculated using the euclidean distance formula



distances <- as.matrix(dist(recommender , method="euclidean"))



Recommend three restaurants with African cuisine based on the recommender matrix and Euclidian distance between each items. The recommender output is restaurant id “12970”, “12996” & “13057”



This is a good example of content based recommender system where similarities are defined by item attributes in the absence of user profile. This recommender types is used to overcome cold start.

# Conclusion

In this exercise the following tasks were accomplished

* Data exploration & preparation
* Data analysis
* Predictive analytics
* Implementation of recommendation system

Next Step

* Improve the recommendation accuracy
* Implement alternative supervised algorithm for recommender system

# Reference

* Multivariate Analysis Using R

By BN Mandal (IASRI Library Ave, New Delhi)

* Plot Graphs in R

By Bret Larget

* A LITERATURE SURVEY ON RECOMMENDATION SYSTEM BASED ON SENTIMENTAL ANALYSIS

Achin Jain1, Vanita Jain2and Nidhi Kapoor3

BharatiVidyapeeth College of Engineering, New Delhi

* Advanced Computational Intelligence

An International Journal (ACII), Vol.3, No.1, January 2016

* An Introduction to Recommendation Systems in Software Engineering

Martin P. Robillard and Robert J. Walker

* Amazon.com Recommendations, Item-to-Item Collaborative Filtering

Greg Linden, Brent Smith, and Jeremy York • Amazon.com

* Incorporating popularity in a personalized news recommender system

Nirmal Jonnalagedda, Susan Gauch, Kevin Labille and Sultan Alfarhood

Computer Science and Computer Engineering, University of Arkansas, Fayetteville, Arkansas, United States

* YELP - A Preference-Based Restaurant Recommendation System

Sumedh Sawant & Gina Pai

Stanford University

* Exploiting Dining Preference for Restaurant Recommendation

Fuzheng Zhang†, Nicholas Jing Yuan†, Kai Zheng∗,Defu Lian‡, Xing Xie†, Yong Rui†

†Microsoft Research

* Recommender Systems

By Aggarwal C.C

ISBN 978-3-319-29657-9

* Recommendation with Knn by Ferran Marti

https://rpubs.com/ferranmt/80166