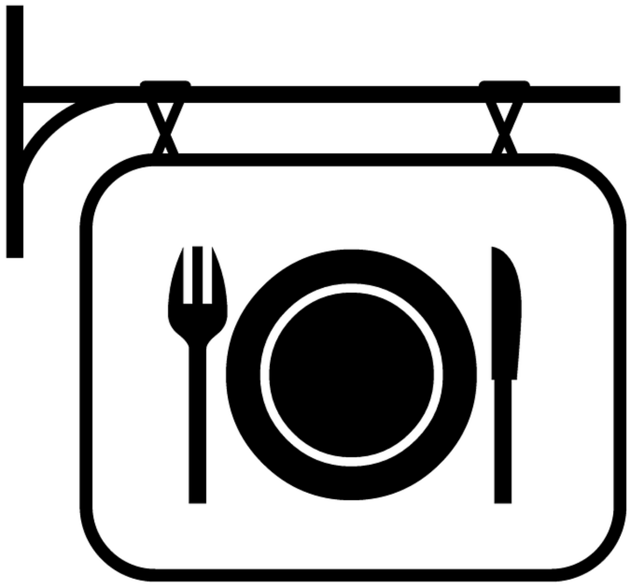
Food Establishment Recommender



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

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.

2. 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

I. An Introduction to Recommendation Systems in Software Engineering by Martin P. Robillard and Robert J. Walker
II. Amazon.com recommendation, Item to Item collaborative filtering by Greg Linden, Brent Smith & Jeremy York
III. A Literature Survey on Recommendation System Based on Sentimental Analysis by Achin Jain, Vanita Jain and Nidhi Kapoor
IV. Incorporating popularity in a personalized news recommender system by Nirmal Jonnalagedda, Susan Gauch, Kevin Labille and Sultan Alfarhood
V. Algorithms and Methods in Recommender Systems by Daniar Asanov
VI. 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.

2. **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.

3. **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.

3. 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

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

4. 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

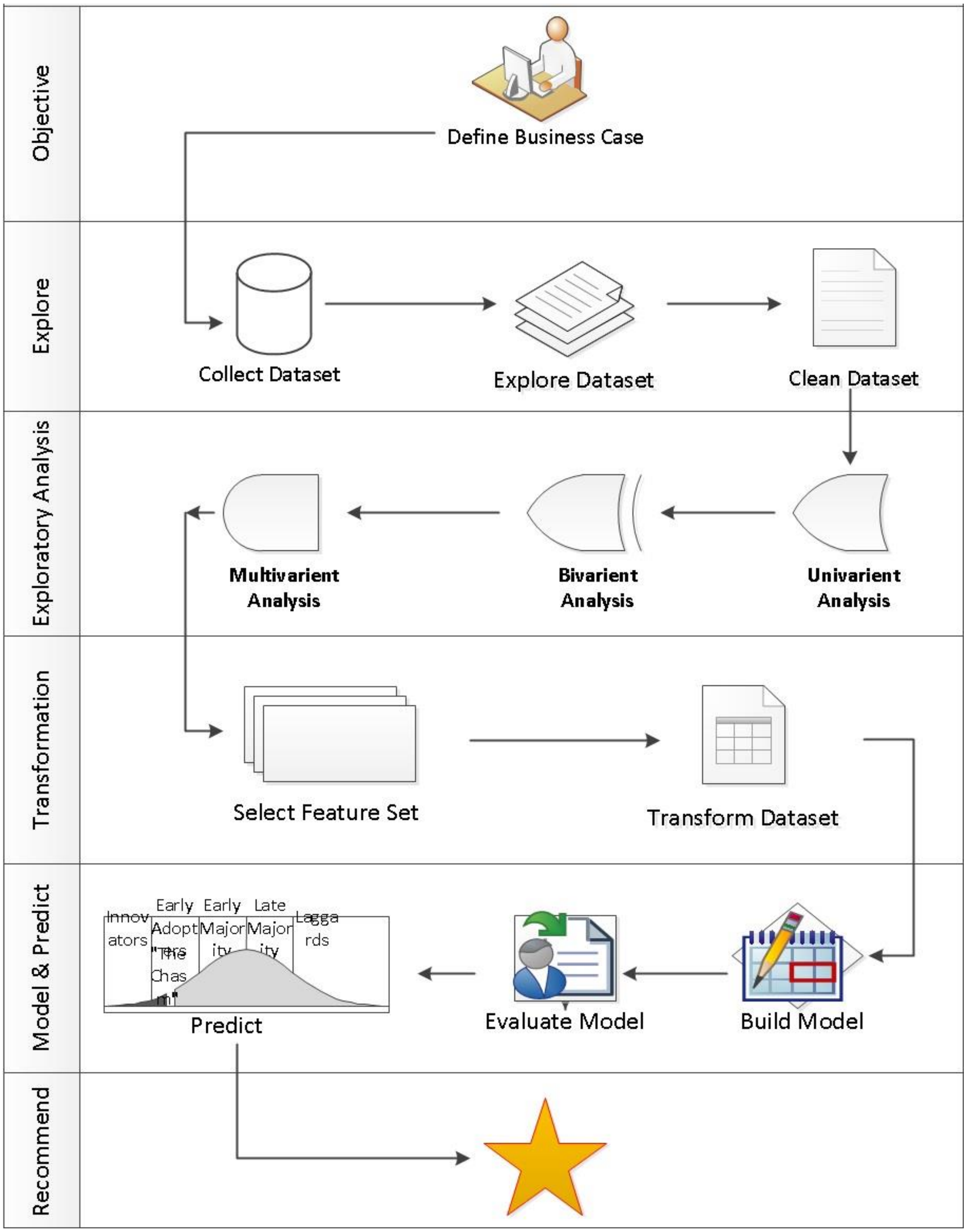
Step 5: Develop Predictive Model & Outcome

- 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.



5. Data Exploration

5.1 Initial dataset description

- Dinesafe dataset

```
 [1] "ROW_ID"                  "ESTABLISHMENT_ID"                  "INSPECTION_ID"
 [4] "ESTABLISHMENT_NAME"      "REVIEW"                            "VALUE"
 [7] "CUISINE_TYPE"            "ESTABLISHMENT_TYPE"                "ESTABLISHMENT_ADDRESS"
[10] "ESTABLISHMENT_STATUS"    "MINIMUM_INSPECTIONS_PERYEAR"       "INFRACTION_DETAILS"
[13] "INSPECTION_DATE"         "SEVERITY"                          "ACTION"
[16] "COURT_OUTCOME"           "AMOUNT_FINED"
```

- Address dataset

```
[1] "ESTABLISHMENT_ID"   "ESTABLISHMENT_NAME" "LONG_ADDRESS"      "SHORT_ADDRESS"      "DISTRICT"
[6] "CITY"               "POSTAL_CODE"
```

5.2 Dataset Summary

- Dinesafe dataset Summary

```
     ROW_ID       ESTABLISHMENT_ID     INSPECTION_ID            ESTABLISHMENT_NAME      REVIEW           VALUE
 Min.   :    1   Min.   : 1222579   Min.   :103179834   TIM HORTONS: 1135   Min.   :1.000   Min.   :1.000
 1st Qu.:22014   1st Qu.:10198651   1st Qu.:103542961   SUBWAY     :  919   1st Qu.:3.000   1st Qu.:1.000
 Median :42690   Median :10393868   Median :103666608   PIZZA PIZZA:  428   Median :3.000   Median :1.000
 Mean   :42345   Mean   :10107910   Mean   :103658272   MCDONALD'S :  383   Mean   :3.236   Mean   :1.526
 3rd Qu.:62202   3rd Qu.:10488300   3rd Qu.:103785830   SECOND CUP :  234   3rd Qu.:3.500   3rd Qu.:2.000
 Max.   :86941   Max.   :10584261   Max.   :103890691   FRESHII    :  222   Max.   :5.000   Max.   :4.000
                                                        (Other)    :12878   NA's   :11      NA's   :452
           CUISINE_TYPE                      ESTABLISHMENT_TYPE         ESTABLISHMENT_ADDRESS
 Cafe            :3194   Restaurant                 :10870   300 BOROUGH DR      :  147
 North American:2898     Food Take Out              : 2991   2300 YONGE ST       :  119
 Deli            :2353   Food Court Vendor          : 1672   1 DUNDAS ST W       :  103
 European        :2137   Bakery                     :  307   1800 SHEPPARD AVE E:   99
 Far Eastern     :1962   Bake Shop                  :  174   3401 DUFFERIN ST    :   98
 Mediterranean : 876     Ice Cream / Yogurt Vendors:   74   40 KING ST W        :   95
 (Other)         :2779   (Other)                    :  111   (Other)             :15538
         ESTABLISHMENT_STATUS MINIMUM_INSPECTIONS_PERYEAR
 Closed          :  101   Min.   :1.000
 Conditional Pass: 2973   1st Qu.:2.000
 Pass            :13125   Median :2.000
                          Mean   :2.255
                          3rd Qu.:3.000
                          Max.   :3.000
```

```
                                                                                                                    INFRACTION_DETAILS
Operator fail to properly wash surfaces in rooms                                                                              :1619
Operator fail to properly maintain rooms                                                                                      :1299
Operator fail to properly wash equipment                                                                                      :1114
Operator fail to properly maintain equipment(NON-FOOD)                                                                        : 523
Fail to ensure the presence of the holder of a valid food handler's certificate  -  Muncipal Code Chapter 545 Sec. 5G(17)(a): 389
(Other)                                                                                                                       :5607
NA's                                                                                                                          :5648
  INSPECTION_DATE                           SEVERITY                                  ACTION
25-10-2016:    72   C - Crucial         : 430   Notice to Comply                   :8031
04-10-2016:    64   M - Minor           :5560   Corrected During Inspection        :2190
17-05-2016:    62   NA - Not Applicable: 904    Ticket                             : 245
18-05-2016:    58   S - Significant     :3657   Summons                            :  48
24-10-2016:    58   NA's                :5648   Summons and Health Hazard Order:  19
19-01-2015:    57                               (Other)                            :  18
(Other)    :15828                               NA's                               :5648
             COURT_OUTCOME       AMOUNT_FINED
Pending              :  135   Min.   :   0.0
Conviction - Fined:  122      1st Qu.:  60.0
Charges Withdrawn :    25     Median : 120.0
Cancelled            :    6   Mean   : 208.1
Charges Quashed      :    2   3rd Qu.: 305.0
(Other)              :    2   Max.   :1875.0
NA's                 :15907   NA's   :16063
```

- Address dataset Summary

```
ESTABLISHMENT_ID                  ESTABLISHMENT_NAME                                                    LONG_ADDRESS
Min.   : 1222579   TIM HORTONS        :  272   2 STRACHAN AVE, TORONTO, ON M6K 3C3, CANADA            :  112
1st Qu.:10197086   SUBWAY             :  232   100 PRINCES' BLVD, TORONTO, ON M6K 3C3, CANADA         :   76
Median :10411080   PIZZA PIZZA        :  102   1 BLUE JAYS WAY, TORONTO, ON M5V 1J3, CANADA           :   59
Mean   :10113602   SHOPPERS DRUG MART:   76    300 BOROUGH DR, SCARBOROUGH, ON M1P 4P5, CANADA :   54
3rd Qu.:10515149   STARBUCKS          :   71   3401 DUFFERIN ST, NORTH YORK, ON M6A 2T9, CANADA:   43
Max.   :10584261   MCDONALD'S         :   70   (Other)                                               :15056
                   (Other)            :14730   NA's                                                  :  153
          SHORT_ADDRESS           DISTRICT                 CITY          POSTAL_CODE
2 STRACHAN AVE  :   112   METRO TORONTO:7387   RICHMOND HILL:    1   M6K 3C3:   247
100 PRINCES BLVD:    76   NORTH YORK   :2696   TORONTO      :15549   M9W    :    97
1 BLUE JAYS WAY :    59   SCARBOROUGH  :2492   VAUGHAN      :    3   M5J    :    89
300 BOROUGH DR  :    54   ETOBICOKE    :1770                         M2N    :    80
3401 DUFFERIN ST:    43   YORK         : 748                         M1B    :    64
40 BAY ST       :    43   EAST YORK    : 441                         M5V    :    63
(Other)         :15166   (Other)      :   19                         (Other):14913
```

5.3 Dataset Structure

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

- Dinesafe dataset Structure

```
'data.frame':   16199 obs. of  17 variables:
 $ ROW_ID                      : int  68185 50462 50463 50464 50465 30104 30105 30106 30107 44731 ...
 $ ESTABLISHMENT_ID            : int  10510325 10435255 10435255 10435255 10435255 10300086 10300086 10300086 10300086 10405624 ...
 $ INSPECTION_ID               : int  103505421 103490016 103550463 103750018 103824680 103490223 103551664 103750021 103824682 103522023 ...
 $ ESTABLISHMENT_NAME          : Factor w/ 864 levels "0109 Dessert + Chocolate",..: 1 2 2 2 2 3 3 3 3 4 ...
 $ REVIEW                      : num  3.5 3 3 3 3 3.5 3.5 3.5 3.5 3 ...
 $ VALUE                       : num  2 1 1 1 1 1 1 1 1 1 ...
 $ CUISINE_TYPE                : Factor w/ 17 levels "African","Bakeries",..: 15 4 4 4 4 8 8 8 8 5 ...
 $ ESTABLISHMENT_TYPE          : Factor w/ 12 levels "Bake Shop","Bakery",..: 12 9 9 9 9 9 9 9 9 12 ...
 $ ESTABLISHMENT_ADDRESS       : Factor w/ 1986 levels "1 ADELAIDE ST E",..: 668 4 4 4 4 4 4 4 4 1191 ...
 $ ESTABLISHMENT_STATUS        : Factor w/ 3 levels "Closed","Conditional Pass",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ MINIMUM_INSPECTIONS_PERYEAR: int  1 2 2 2 2 2 2 2 2 2 ...
 $ INFRACTION_DETAILS          : Factor w/ 226 levels "Altering number of washbasins in facility without inspector's approval O. Reg  562/90 Sec. 69",..: NA NA
NA NA NA NA NA NA NA NA ...
 $ INSPECTION_DATE             : Factor w/ 523 levels "01-02-2016","01-03-2016",..: 55 207 214 227 60 207 214 227 60 386 ...
 $ SEVERITY                    : Factor w/ 4 levels "C - Crucial",..: NA NA NA NA NA NA NA NA NA NA ...
 $ ACTION                      : Factor w/ 8 levels "Corrected During Inspection",..: NA NA NA NA NA NA NA NA NA NA ...
 $ COURT_OUTCOME               : Factor w/ 7 levels "Cancelled","Charges Quashed",..: NA NA NA NA NA NA NA NA NA NA ...
 $ AMOUNT_FINED                : int  NA NA NA NA NA NA NA NA NA NA ...
```

- Address dataset Structure

```
'data.frame':   15553 obs. of  7 variables:
 $ ESTABLISHMENT_ID  : int  9337616 10384957 10390332 10492908 10233710 10480531 10527234 10550136 10580268 10412094 ...
 $ ESTABLISHMENT_NAME: Factor w/ 12154 levels "*-SUNNYLEA COOP NURSERY SCHOOL",..: 9652 11211 4855 1802 9669 7984 6717 895 3154 9662 ...
 $ LONG_ADDRESS      : Factor w/ 10741 levels "1 ADELAIDE ST E, TORONTO, ON M5C 2V9, CANADA",..: 1 1 1 2 3 3 4 5 6 7 ...
 $ SHORT_ADDRESS     : Factor w/ 10885 levels "1 ADELAIDE ST E",..: 1 1 1 2 3 3 4 5 6 7 ...
 $ DISTRICT          : Factor w/ 8 levels "EAST YORK","Etobicoke",..: 4 4 4 4 5 5 4 4 4 4 ...
 $ CITY              : Factor w/ 3 levels "RICHMOND HILL",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ POSTAL_CODE       : Factor w/ 5139 levels "L3T","L4J","L4J 8J8",..: 2571 2571 2571 2876 1006 1006 3070 2173 2200 2891 ...
```

5.4 Dataset Sample

A sample of the two datasets using a head function

- Dinesafe dataset sample

ESTABLISHMENT_ID	ROW_ID	INSPECTION_ID	ESTABLISHMENT_NAME.x	REVIEW	VALUE	CUISINE_TYPE	ESTABLISHMENT_TYPE	ESTABLISHMENT_ADDRESS
<int>	<int>	<int>	<fctr>	<dbl>	<dbl>	<fctr>	<fctr>	<fctr>
10584093	86928	103889233	PIZZAIOLO	3.5	1	European	Restaurant	123 SPADINA AVE
10584149	86932	103889610	Thai Express	3.0	1	South East Asian	Food Take Out	320 FRONT ST W
10584240	86939	103890492	Starbucks Coffee	3.7	2	Cafe	Food Take Out	621 KING ST W
10584261	86940	103890691	GLAD DAY	4.5	2	Cafe	Restaurant	499 CHURCH ST
10584261	86941	103890691	GLAD DAY	4.5	2	Cafe	Restaurant	499 CHURCH ST

ESTABLISHMENT_STATUS	MINIMUM_INSPECTIONS_PERYEAR	INFRACTION_DETAILS	INSPECTION_DATE
<fctr>	<int>	<fctr>	<fctr>
Pass	2	NA	10-01-2017
Pass	2	NA	11-01-2017
Pass	2	NA	12-01-2017
Pass	2	FAIL TO PROVIDE THERMOMETER IN STORAGE COMPARTMENT O. REG 562/90 SEC. 21	12-01-2017
Pass	2	Operator fail to properly maintain rooms	12-01-2017

- Address dataset sample

ESTABLISHMENT_NAME	LONG_ADDRESS	SHORT_ADDRESS	DISTRICT	CITY	POSTAL_CODE
<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>
LIPSTICK & DYNAMITE	992 QUEEN ST W, TORONTO, ON M6J 1H1, CANADA	992 QUEEN ST W	METRO TORONTO	TORON...	M6J 1H1
FRANKIES BAR & CAFE	994 QUEEN ST W, TORONTO, ON M6J 1H1, CANADA	994 QUEEN ST W	METRO TORONTO	TORON...	M6J 1H1
PROGRESS PORTUGUESE BAKERY AND PASTRY	996 DOVERCOURT RD, TORONTO, ON M6H 2X5, CANADA	996 DOVERCOURT RD	METRO TORONTO	TORON...	M6H 2X5
MACELLERIA SAN GABRIELE BUTCHER & GRILL	998 ST CLAIR AVE W, TORONTO, ON M6E 1A2, CANADA	998 ST CLAIR AVE W	YORK	TORON...	M6E 1A2
FRIDA RESTAURANT & BAR	999 EGLINTON AVE W, YORK, ON M6C 2C7, CANADA	999 EGLINTON AVE W	YORK	TORON...	M6C 2C7

5.5 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.

```
'data.frame':   16199 obs. of  23 variables:
 $ ESTABLISHMENT_ID           : int  1222579 1222807 1222807 1222807 1222807 1222807 1222807 1222807 1223056 1223056 ...
 $ ROW_ID                     : int  1 4 5 6 7 8 9 10 11 12 ...
 $ INSPECTION_ID              : int  103868579 103472815 103537032 103537032 103616870 103702528 103732221 103874297 103541411 103647049 ...
 $ ESTABLISHMENT_NAME.x       : Factor w/ 864 levels "0109 Dessert + Chocolate",..: 683 645 645 645 645 645 645 645 658 658 ...
 $ REVIEW                     : num  5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3 3 ...
 $ VALUE                      : num  1 1 1 1 1 1 1 1 2 2 ...
 $ CUISINE_TYPE               : Factor w/ 17 levels "African","Bakeries",..: 16 9 9 9 9 9 9 9 8 8 ...
 $ ESTABLISHMENT_TYPE         : Factor w/ 12 levels "Bake Shop","Bakery",..: 9 12 12 12 12 12 12 12 12 12 ...
 $ ESTABLISHMENT_ADDRESS      : Factor w/ 1986 levels "1 ADELAIDE ST E",..: 1903 417 417 417 417 417 417 417 1623 1623 ...
 $ ESTABLISHMENT_STATUS       : Factor w/ 3 levels "Closed","Conditional Pass",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ MINIMUM_INSPECTIONS_PERYEAR: int  2 3 3 3 3 3 3 3 2 2 ...
 $ INFRACTION_DETAILS         : Factor w/ 226 levels "Altering number of washbasins in facility without inspector's approval O. Reg  562/90 Sec. 69",..: 159
159 120 150 NA NA NA NA NA NA ...
 $ INSPECTION_DATE            : Factor w/ 523 levels "01-02-2016","01-03-2016",..: 360 385 370 370 447 69 299 499 491 98 ...
 $ SEVERITY                   : Factor w/ 4 levels "C - Crucial",..: 2 2 4 2 NA NA NA NA NA NA ...
 $ ACTION                     : Factor w/ 8 levels "Corrected During Inspection",..: 3 3 1 3 NA NA NA NA NA NA ...
 $ COURT_OUTCOME              : Factor w/ 7 levels "Cancelled","Charges Quashed",..: NA NA NA NA NA NA NA NA NA NA ...
 $ AMOUNT_FINED               : int  NA NA NA NA NA NA NA NA NA NA ...
 $ ESTABLISHMENT_NAME.y       : Factor w/ 12154 levels "*-SUNNYLEA COOP NURSERY SCHOOL",..: 8793 7838 7838 7838 7838 7838 7838 7838 7993 7993 ...
 $ LONG_ADDRESS               : Factor w/ 10741 levels "1 ADELAIDE ST E, TORONTO, ON M5C 2V9, CANADA",..: 10207 2397 2397 2397 2397 2397 2397 2397 8874 8874
...
 $ SHORT_ADDRESS              : Factor w/ 10885 levels "1 ADELAIDE ST E",..: 10354 2403 2403 2403 2403 2403 2403 2403 8884 8884 ...
 $ DISTRICT                   : Factor w/ 8 levels "EAST YORK","Etobicoke",..: 6 5 5 5 5 5 5 5 3 3 ...
 $ CITY                       : Factor w/ 3 levels "RICHMOND HILL",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ POSTAL_CODE                : Factor w/ 5139 levels "L3T","L4J","L4J 8J8",..: 172 4019 4019 4019 4019 4019 4019 4019 4526 4526 ...
```

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

```
      African          Bakeries               Bar              Cafe         Caribbean 
           72                53               461              3194               234 
         Deli           Dessert          European       Far Eastern Juicery & Smoothies 
         2353               187              2137              1962               335 
Latin American     Mediterranean    Middle Eastern    North American          Pastries 
          213               876               115              2898               218 
  South Asian   South East Asian              <NA> 
          318               573                 0 
```

```
     Closed Conditional Pass             Pass              <NA> 
        101             2973            13125                 0 
```

```
EAST YORK     Etobicoke      ETOBICOKE  METRO TORONTO     NORTH YORK    SCARBOROUGH TORONTO ISLAND 
      364             0           1387           8290           3097           2517             10 
     YORK          <NA> 
      534             0 
```

```
  C - Crucial          M - Minor NA - Not Applicable     S - Significant 
          430               5560               904               3657 
```

6. Data Munging

6.1 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))

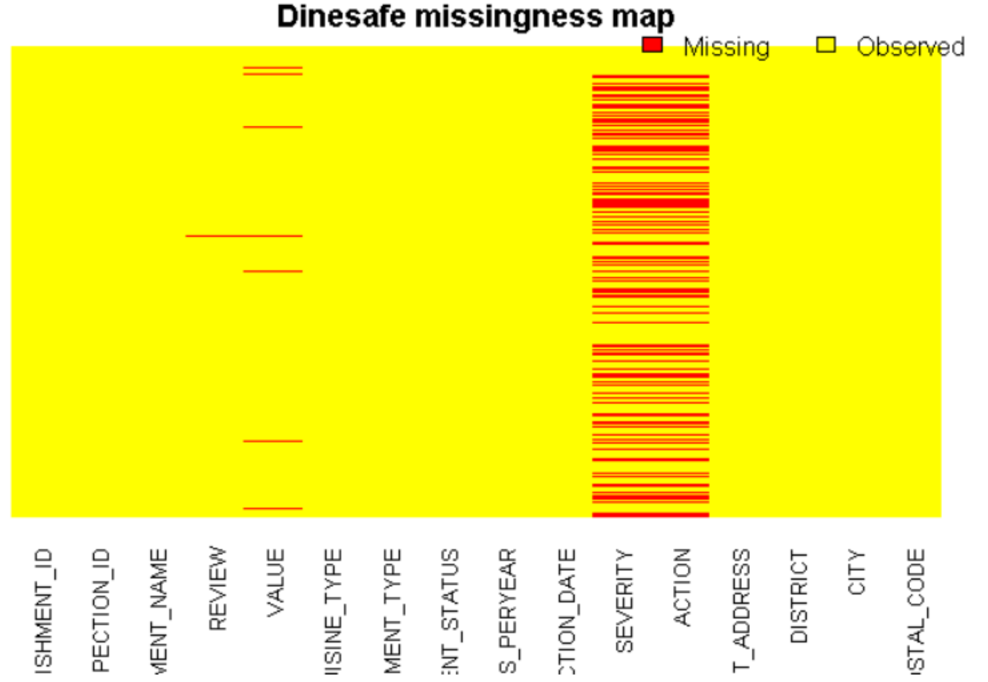
- *Rename “establishment_name.x” to “establishment_name”*

colnames(Dinesafe)[colnames(Dinesafe) == 'ESTABLISHMENT_NAME.x'] <- 'ESTABLISHMENT_NAME'

6.2 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.*

```
            ESTABLISHMENT_ID               INSPECTION_ID          ESTABLISHMENT_NAME
REVIEW
                           0                           0                           0
11
                       VALUE                CUISINE_TYPE          ESTABLISHMENT_TYPE
ESTABLISHMENT_STATUS
                         452                           0                           0
0
MINIMUM_INSPECTIONS_PERYEAR             INSPECTION_DATE                    SEVERITY
ACTION
                           0                           0                        5648
5648
               SHORT_ADDRESS                    DISTRICT                        CITY
POSTAL_CODE
                           0                           0                           0
0
```



6.3 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"))

6.4 Describe Dataset

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

```
Dinesafe$REVIEW
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   16188       11       22    0.921    3.236   0.6238      2.5      2.5      3.0      3.0      3.5      4.0      4.0

lowest : 1.0 1.5 2.0 2.5 2.8, highest: 4.2 4.3 4.5 4.6 5.0

Dinesafe$VALUE
       n  missing distinct     Info     Mean      Gmd
   15747      452        5    0.768    1.526   0.5575

Value        1.0   2.0   2.5   3.0   4.0
Frequency   7932  7398     4   364    49
Proportion 0.504 0.470 0.000 0.023 0.003
```

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

```
o   Complete_Dinesafe  <- Dinesafe[complete.cases(Dinesafe),]
o   nrow(Complete_Dinesafe)
```

6.5 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.

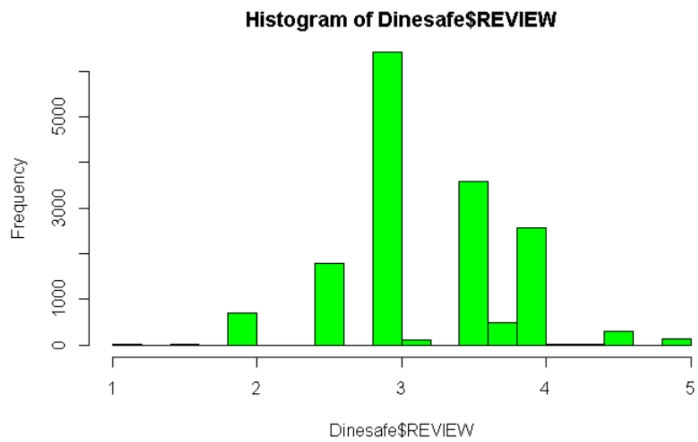
```
o   Dinesafe_NA  <- Dinesafe[!complete.cases(Dinesafe),]
o   nrow(Dinesafe_NA)
```

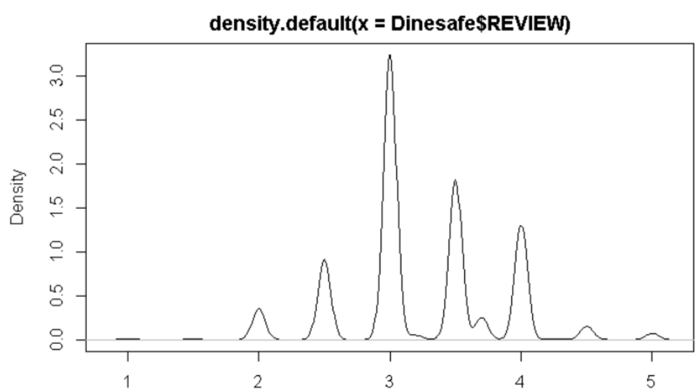
7. Data Exploratory Analysis & Visualization

7.1 Univariate Data Analysis

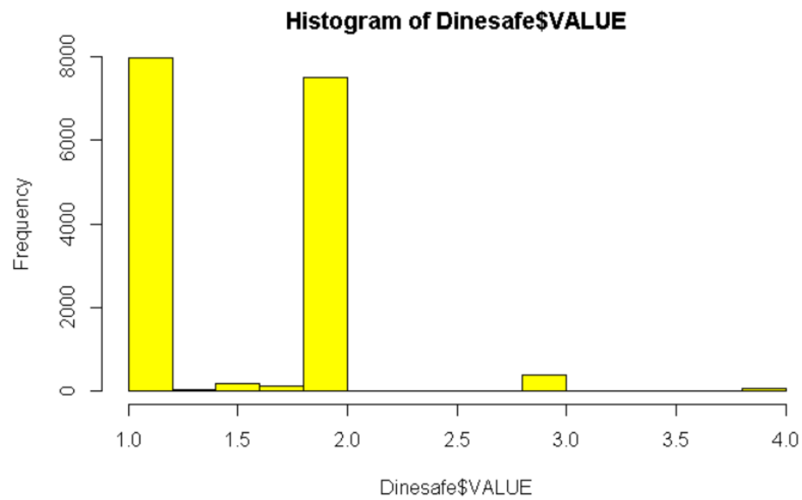
In this section a single variable from the dataset was analyzed to understand the data using histogram and density graphical representation

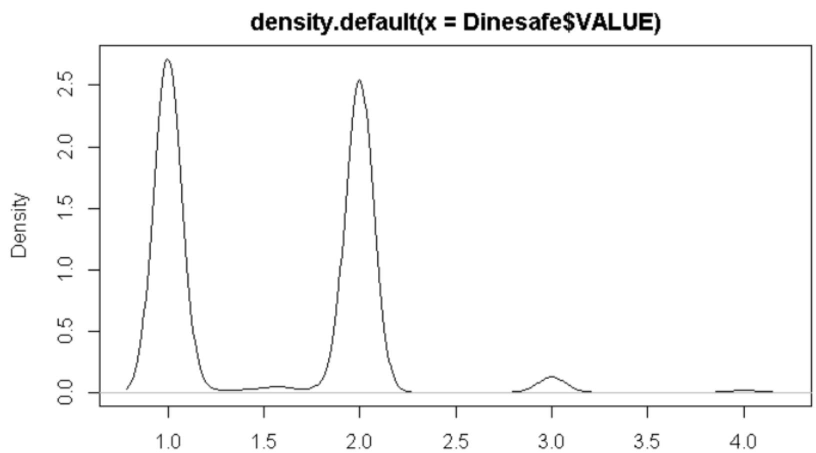
7.1.1 Review Variable : The data is normally distrusted



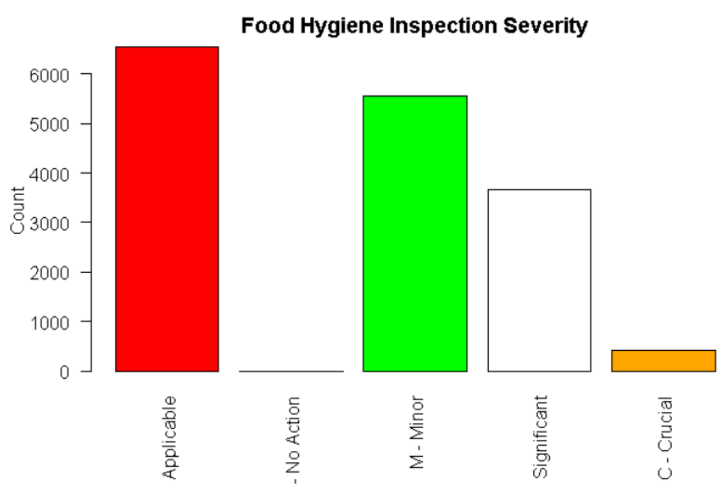


7.1.2 Value Variable : The data is skewed to the right

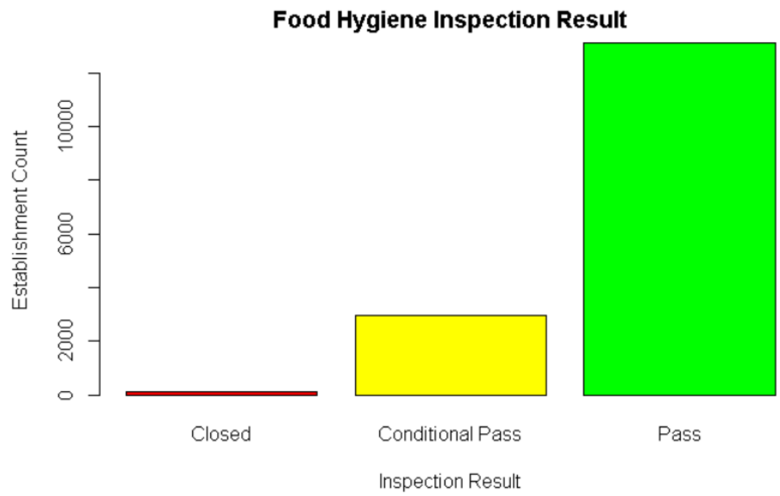




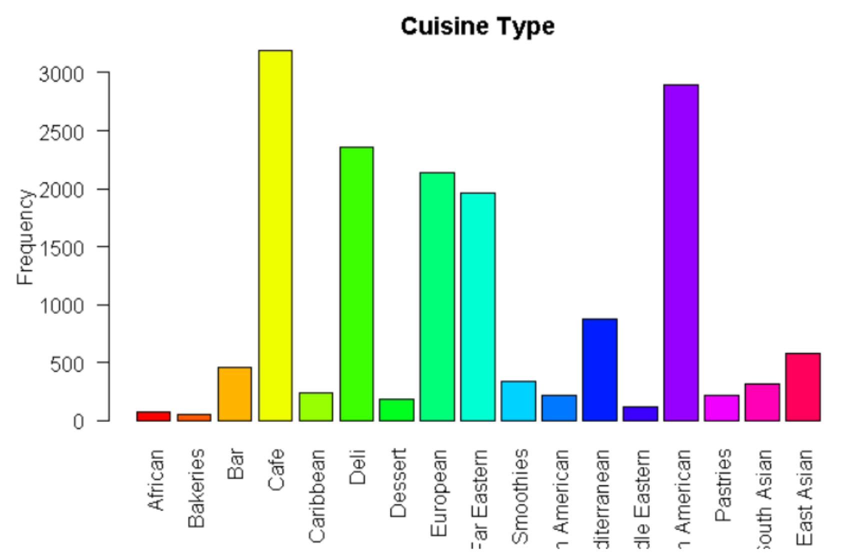
7.1.3 Food inspection severity graph



7.1.4 Food Hygiene Inspection Result

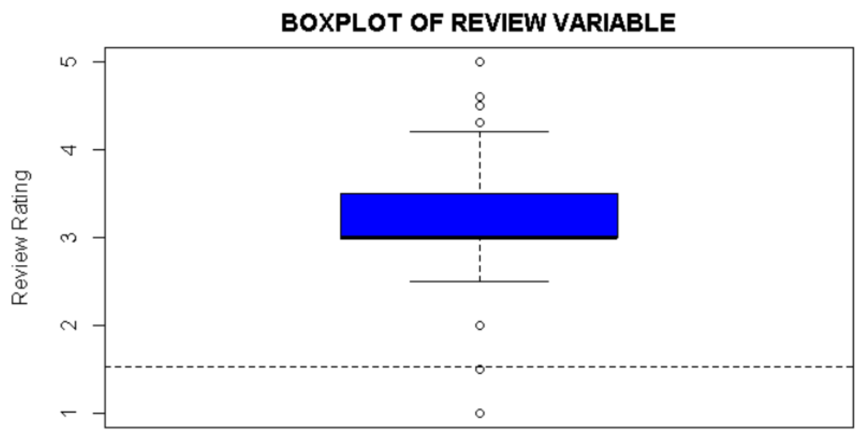


7.1.5 Establishment Cuisine Type



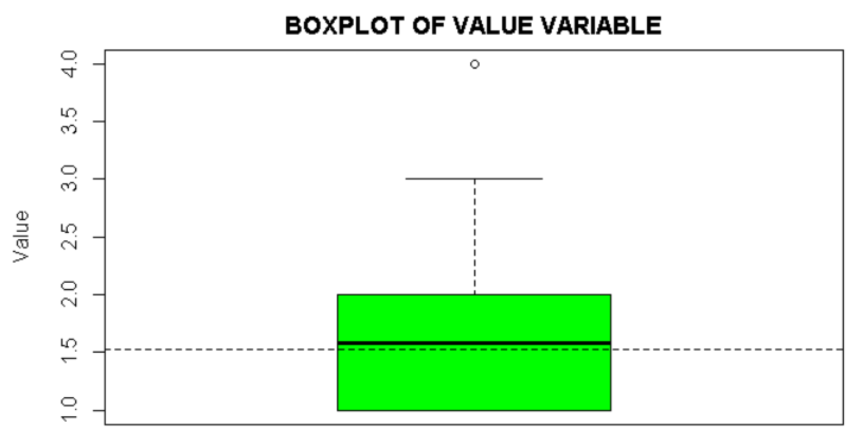
7.1.6 Boxplot of Review Variable

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



7.1.7 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.



7.2 Bivariate Data Analysis

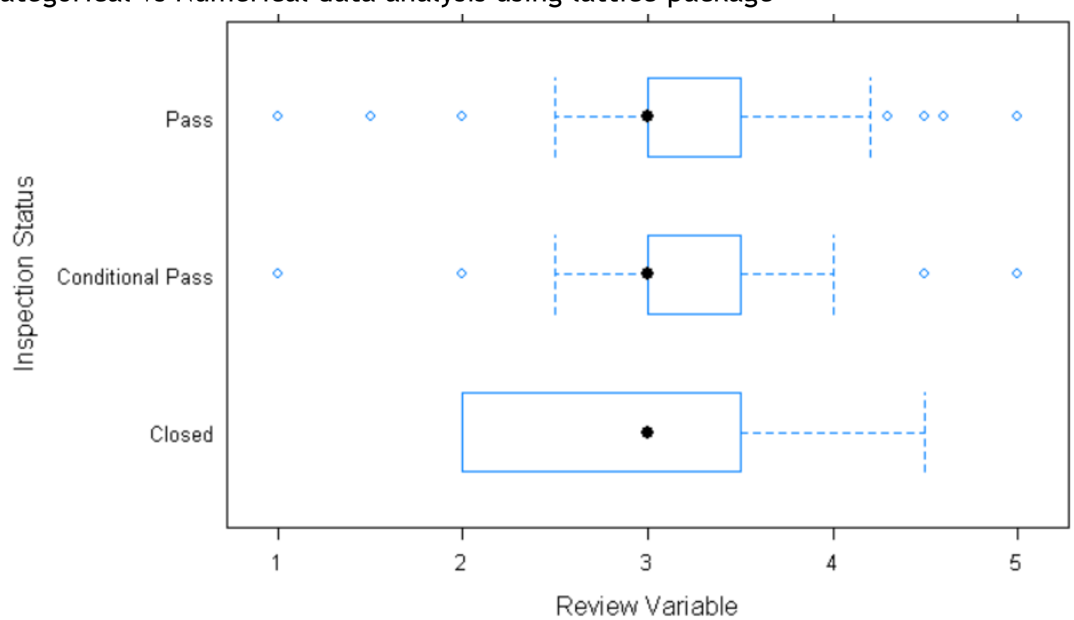
7.2.1 Mean and Standard Deviation

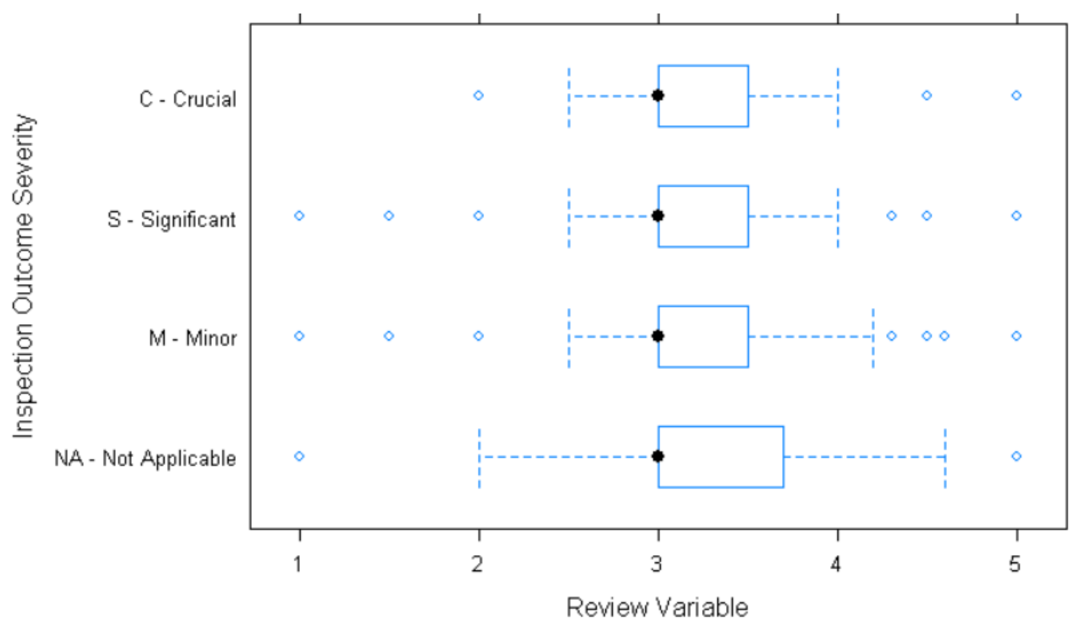
The mean and standard deviation value of food premises that failed inspection had a mean review value below those that passed inspection. Also failed food premises had a higher standard deviation value as compared to those who passed.

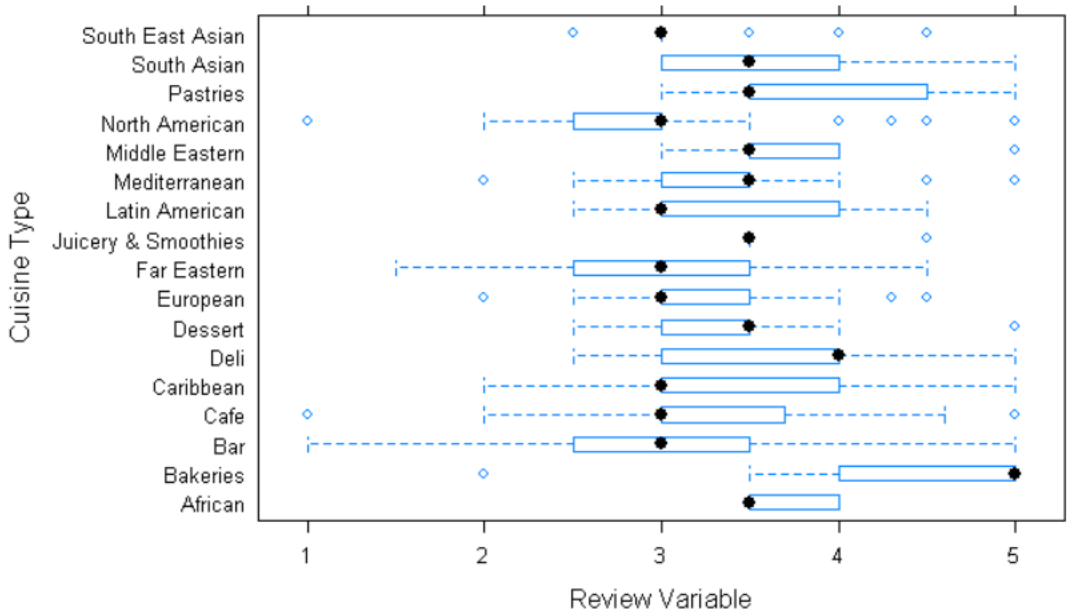
On the other hand the relationship between mean/standard deviation value and inspection outcome is not observer due to consistent result across all three values.

Mean Review data against establishment inspection status	Closed 2.871287	Conditional Pass 3.176495	Pass 3.252679
Standard Deviation of Review data against establishment inspection status	Closed 0.7471729	Conditional Pass 0.5964501	Pass 0.5700371
Mean value data against establishment inspection status	Closed 1.536582	Conditional Pass 1.580754	Pass 1.516429
Standard Deviation of Value data against establishment inspection status	Closed 0.4915077	Conditional Pass 0.5203282	Pass 0.5617093

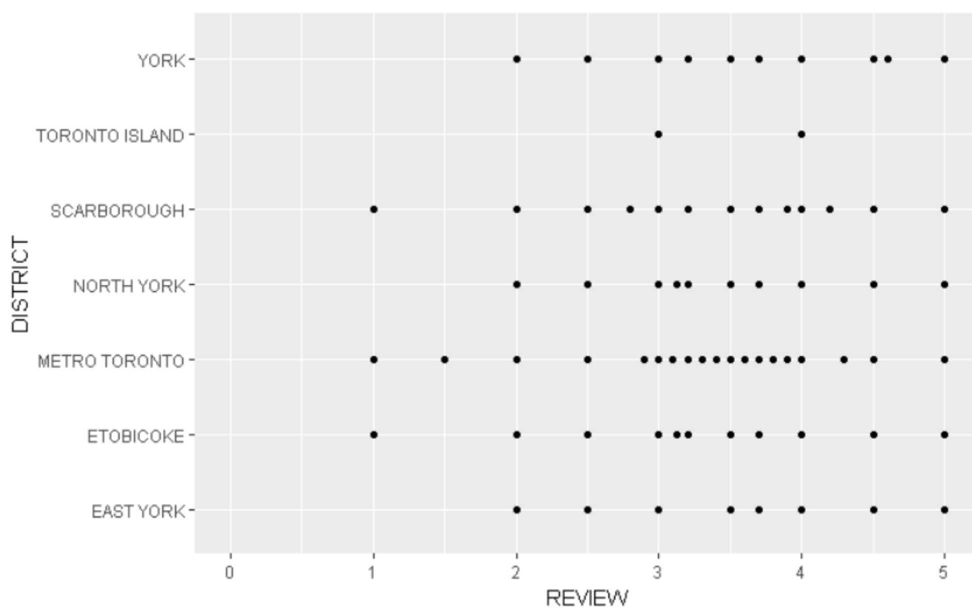
7.2.2 Categorical vs Numerical data analysis using lattice package

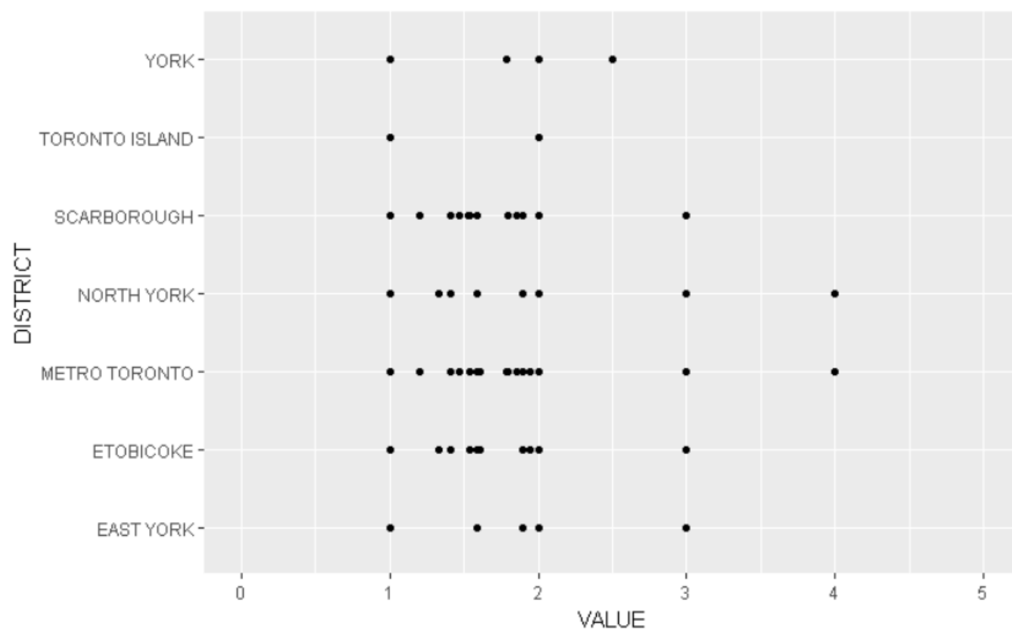






7.2.3 Categorical vs Numerical data analysis using ggplot2 package



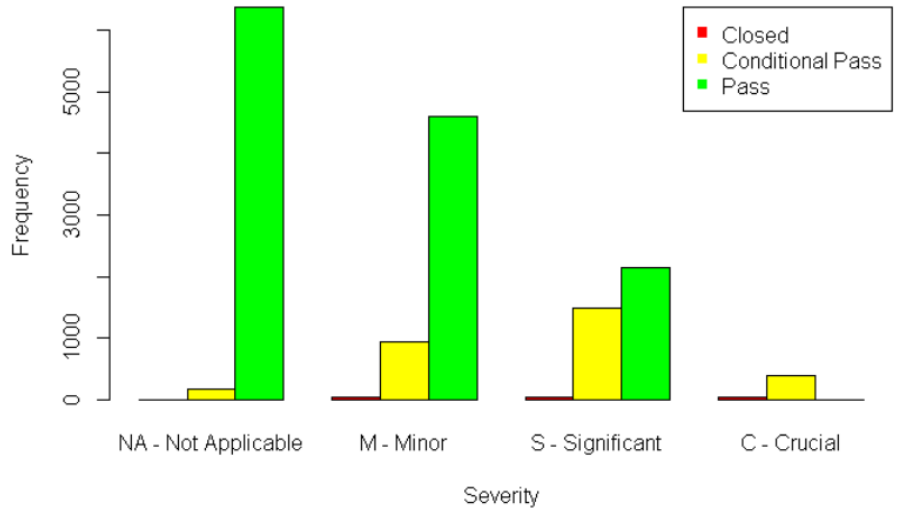


7.2.4 Crosstab analysis of “Severity” and “Inspection Status” analysis with Crosstab & barplot

```
------------------------------|---------------------|----------------------|---
                         Pass |                6383 |                 4601 |
2139 |                      2 |              13125 |
                              |             217.419 |                2.050 |
229.166 |             344.413 |                     |
                              |               0.486 |                0.351 |
0.163 |               0.000 |               0.810 |
                              |               0.974 |                0.828 |
0.585 |               0.005 |                     |
                              |               0.394 |                0.284 |
0.132 |               0.000 |                     |
------------------------------|---------------------|----------------------|---

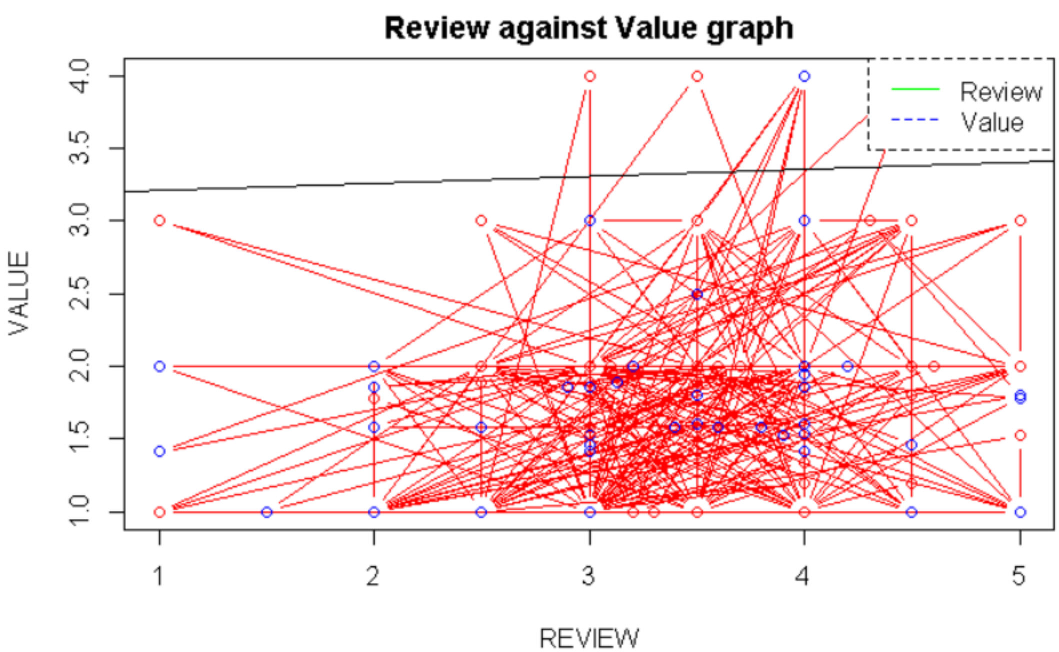
                 Column Total |                6552 |                 5560 |
3657 |                    430 |              16199 |
                              |               0.404 |                0.343 |
0.226 |               0.027 |                     |
------------------------------|---------------------|----------------------|---

                  y
x                  NA - Not Applicable M - Minor S - Significant C - Crucial
  Closed                             1         28            41          31
  Conditional Pass                 168        931          1477         397
  Pass                            6383       4601          2139           2
```



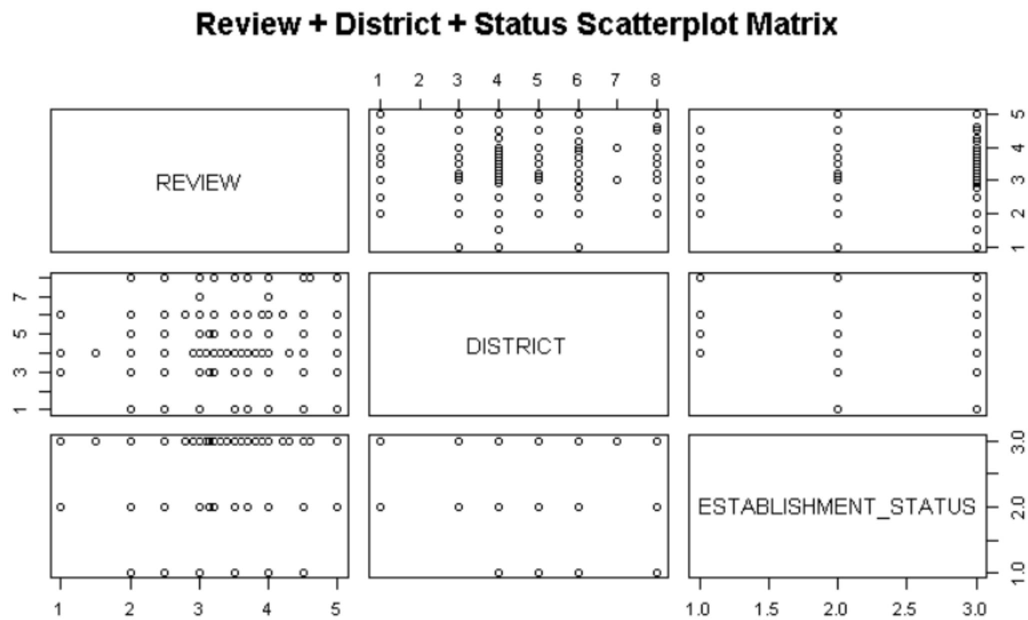
7.2.5 Relationship between "Review" and "Value" using scatter plot

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.

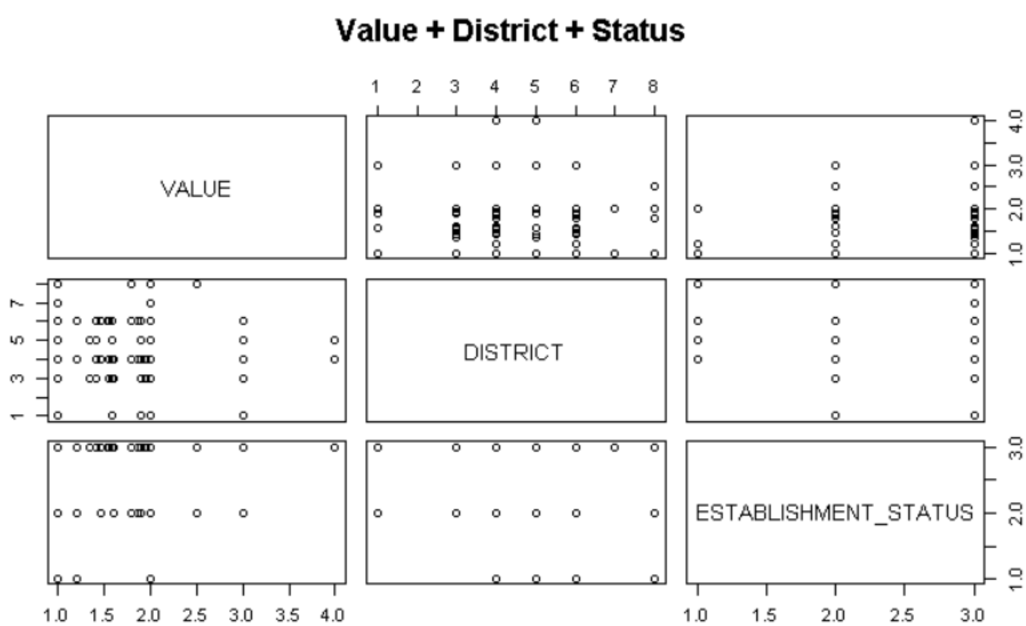


7.3 Multivariate Data Analysis

7.3.1 Simple Scatter Matrix on the relationship between Review, District and Status variables



7.3.2 Simple Scatter Matrix on the relationship between Value, District and Status variables



7.3.3 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)

	Dinesafe$ESTABLISHMENT_STATUS <fctr>	Dinesafe$CUISINE_TYPE <fctr>	Dinesafe$DISTRICT <fctr>	Dinesafe$REVIEW <dbl>
1	Pass	African	EAST YORK	3.500000
2	Pass	Bakeries	EAST YORK	5.000000
3	Conditional Pass	Cafe	EAST YORK	3.000000
4	Pass	Cafe	EAST YORK	3.217978
5	Conditional Pass	Deli	EAST YORK	4.000000
6	Pass	Deli	EAST YORK	3.771429
7	Pass	European	EAST YORK	3.010638
8	Pass	Far Eastern	EAST YORK	2.500000
9	Pass	Juicery & Smoothies	EAST YORK	3.500000
10	Pass	Latin American	EAST YORK	3.000000

1-10 of 10 rows

7.3.4 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)

	Dinesafe$ESTABLISHMENT_STATUS <fctr>	Dinesafe$CUISINE_TYPE <fctr>	Dinesafe$REVIEW <int>
1	Conditional Pass	African	7
2	Pass	African	65
3	Conditional Pass	Bakeries	5
4	Pass	Bakeries	48
5	Conditional Pass	Bar	60
6	Pass	Bar	401
7	Conditional Pass	Cafe	381
8	Pass	Cafe	2813
9	Closed	Caribbean	5
10	Conditional Pass	Caribbean	48

1-10 of 10 rows

8. 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

I. 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.

II. Created a unique rows based on the selected features. This reduces the number of rows from 16,199 to 2723

III. 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.

ESTABLISHMENT_ID <int>	ESTABLISHMENT_NAME <fctr>	REVIEW <dbl>	VALUE <dbl>	CUISINE_TYPE <fctr>	CUISINE_IDX <chr>
1222579	SAI-LILA KHAMAN DHOKLA HOUSE	5.0	1	South Asian	15
1222807	PHO BO TO	3.5	1	Far Eastern	9
1223056	PIZZA PIZZA	3.0	2	European	8
9000004	PAPINO'S PIZZA	4.0	1	European	8
9000026	2-4-1 PIZZA	2.5	2	European	8
9000029	2-4-1 PIZZA	2.5	2	European	8

IV. Changed CUISINE_TYPE from factor to numerical value

```
Dinesafe2$CUISINE_IDX <- as.numeric(Dinesafe2$CUISINE_IDX)
```

V. Created binary values for the "Cuisine Type" in order to create a binary attributes

```
Dinesafe2$African <- ifelse(Dinesafe2$CUISINE_TYPE == "African",1,0)
```

ESTABLISHMENT_ID	ESTABLISHMENT_NAME	REVIEW	VALUE	CUISINE_TYPE	CUISINE_IDX	African	Bakeries	Bar	Cafe	Caribbean	Deli	Dessert	European	FarEastern	Mediterranean	MidEastern	NAmerican	Juice
1222579	SAI-LILA KHAMAN DHOKLA HOUSE	5.0	1.000000	South Asian	15	0	0	0	0	0	0	0	0	0	0	0	0	0
1222807	PHO BO TO	3.5	1.000000	Far Eastern	9	0	0	0	0	0	0	0	0	1	0	0	0	0
1223056	PIZZA PIZZA	3.0	2.000000	European	8	0	0	0	0	0	0	0	1	0	0	0	0	0
9000004	PAPINO'S PIZZA	4.0	1.000000	European	8	0	0	0	0	0	0	0	1	0	0	0	0	0
9000026	2-4-1 PIZZA	2.5	2.000000	European	8	0	0	0	0	0	0	0	1	0	0	0	0	0
9000029	2-4-1 PIZZA	2.5	2.000000	European	8	0	0	0	0	0	0	0	1	0	0	0	0	0

```
'data.frame':   2723 obs. of  22 variables:
 $ ESTABLISHMENT_ID: int  1222579 1222807 1223056 9000004 9000026 9000029 9000031 9000046 9000109 9000116 ...
 $ REVIEW          : num  5 3.5 3 4 2.5 2.5 2.5 2.5 3 2 ...
 $ VALUE           : num  1 1 2 1 2 2 2 2 2 2 ...
 $ CUISINE_TYPE    : Factor w/ 17 levels "African","Bakeries",..: 16 9 8 8 8 8 8 8 3 4 ...
 $ CUISINE_IDX     : num  15 9 8 8 8 8 8 8 3 4 ...
 $ African         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Bakeries        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Bar             : num  0 0 0 0 0 0 0 0 1 0 ...
 $ Cafe            : num  0 0 0 0 0 0 0 0 0 1 ...
 $ Caribbean       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Deli            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Dessert         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ European        : num  0 0 1 1 1 1 1 1 0 0 ...
 $ FarEastern      : num  0 1 0 0 0 0 0 0 0 0 ...
 $ Mediterranean   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ MidEastern      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ NAmerican       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Juicery         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Pastries        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SouthAsian      : num  1 0 0 0 0 0 0 0 0 0 ...
 $ SEastAsian      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LAmerican       : num  0 0 0 0 0 0 0 0 0 0 ...
```

VI. Normalize the “Review”, “Value” & “Cuisine_Idx” labels

```
'data.frame':   2723 obs. of  20 variables:
 $ African      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Bakeries     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Bar          : num  0 0 0 0 0 0 0 0 1 0 ...
 $ Cafe         : num  0 0 0 0 0 0 0 0 0 1 ...
 $ Caribbean    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Deli         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Dessert      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ European     : num  0 0 1 1 1 1 1 1 0 0 ...
 $ FarEastern   : num  0 1 0 0 0 0 0 0 0 0 ...
 $ Mediterranean: num  0 0 0 0 0 0 0 0 0 0 ...
 $ MidEastern   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ NAmerican    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Juicery      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Pastries     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SouthAsian   : num  1 0 0 0 0 0 0 0 0 0 ...
 $ SEastAsian   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LAmerican    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ REVIEW       : num  1 0.625 0.5 0.75 0.375 0.375 0.375 0.375 0.5 0.25 ...
 $ VALUE        : num  0 0 0.333 0 0.333 ...
 $ CUISINE_IDX  : num  0.882 0.529 0.471 0.471 0.471 ...
```

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

9. 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

9.1 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.

9.2 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

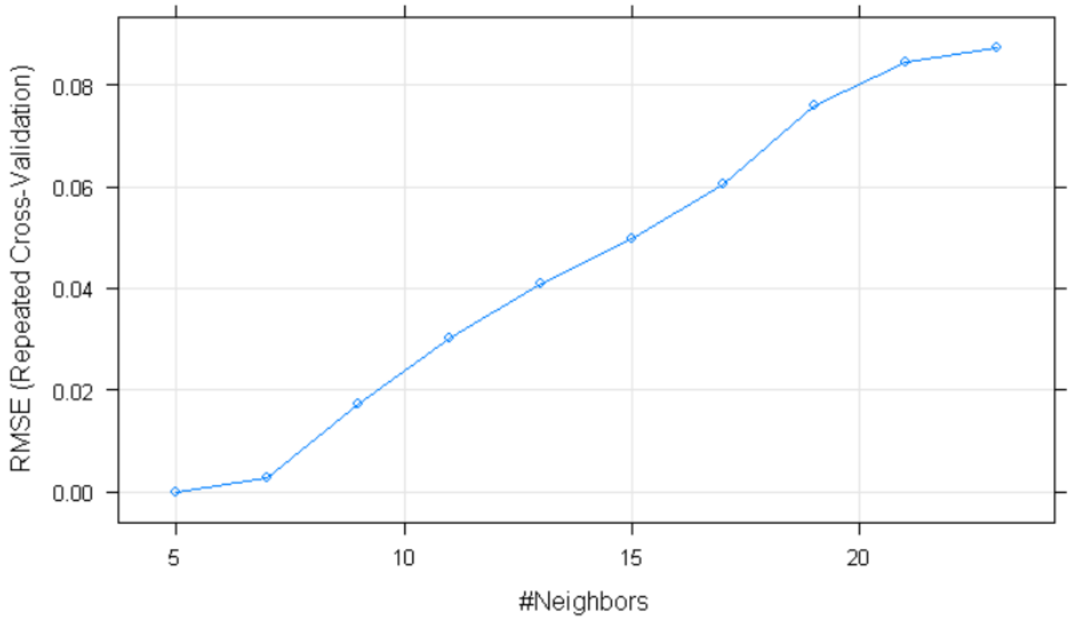
```
<truncated>k-Nearest Neighbors

2000 samples
  19 predictor

Pre-processing: centered (19), scaled (19)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1800, 1800, 1799, 1800, 1800, 1800, ...
Resampling results across tuning parameters:

  k   RMSE          Rsquared
   5  5.834634e-16  1.0000000
   7  2.668497e-03  0.9996632
   9  1.729579e-02  0.9857054
  11  3.022194e-02  0.9692707
  13  4.094209e-02  0.9503906
  15  4.964589e-02  0.9374691
  17  6.042854e-02  0.9077873
  19  7.599317e-02  0.8641550
  21  8.439008e-02  0.8395811
  23  8.720328e-02  0.8307749

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was k = 5.
```



9.3 Prediction

Confusion Matrix and Statistics

Prediction \ **Reference**	African	Bakeries	Bar	Cafe	Caribbean	Deli	Dessert	European	Far Eastern	Juicery	Latin American	Mediterranean	Middle Eastern	North American	Pastries	South Asian	South East Asian
African	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bakeries	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bar	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cafe	0	0	0	203	0	0	0	0	0	0	0	0	0	0	0	0	0
Caribbean	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
Deli	0	0	0	0	0	125	0	0	0	0	0	0	0	0	0	0	0
Dessert	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0
European	0	0	0	0	0	0	0	96	0	0	0	0	0	0	0	0	0
Far Eastern	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0
Juicery	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0
Latin American	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0
Mediterranean	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0
Middle Eastern	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
North American	0	0	0	0	0	0	0	0	0	0	0	0	0	91	0	0	0
Pastries	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0
South Asian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0
South East Asian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

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%

```
Overall Statistics

               Accuracy : 1
                 95% CI : (0.9949, 1)
    No Information Rate : 0.2808
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA
```

10. 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

$$\text{Euclidean Distance} = \sqrt{(x_1 - y_1)^2 + \ldots + (x_N - y_N)^2}$$

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

	12661	12672	12689	12694	12698	12701
12661	0.0000000	0.4166667	1.487697	1.4923399	1.487697	1.4552881
12672	0.4166667	0.0000000	1.449872	1.5372669	1.449872	1.4718948
12689	1.4876966	1.4498725	0.000000	1.4644975	0.000000	1.4595192
12694	1.4923399	1.5372669	1.464498	0.0000000	1.464498	1.4442951
12698	1.4876966	1.4498725	0.000000	1.4644975	0.000000	1.4595192
12701	1.4552881	1.4718948	1.459519	1.4442951	1.459519	0.0000000
12705	1.4952186	1.4788737	1.420945	1.4596496	1.420945	1.4790037
12710	1.5475996	1.4904541	1.430653	0.3764786	1.430653	1.4718948
12712	1.4552881	1.4718948	1.459519	1.4442951	1.459519	0.0000000

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”

```
```{r}
cuisine <- "African"
listing <- most.probable.recommend(cuisine, recommender, distances)
rownames(recommender)[listing[1:3]]
```
```

```
[1] "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.

11. 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

12. 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