SVM and Cross-Validation using R Studio

Nainika Kaushik, Manjot Kaur Bhatia, Sonali Rastogi

Abstract: Each passing day data is getting multiplied. It is difficult to extract useful information from such big data. Data Mining is used to extract useful information. Data mining is used in majorly all fields like healthcare, marketing, social media platforms and so on. In this paper, data is loaded and preprocessed by dealing with some missing values. The dataset used is of Airbnb, the platform used for lodging and tourism industry. Analyzing the data by plotting correlation using spearman method. Further, applying PCA and Support Vector Machine classification technique on the dataset. There are various applications of SVM, it is used in face-detection, text and hypertext categorization, classification of images, bioinformatics and so on. SVM has high dimensional input space, sparse document vectors and regularization parameters therefore it is appropriate to use SVM. Cross-validation gives more accurate result. The dataset is divided into folds. The end product is the test set which is similar to full dataset. Confusion matrix is evaluated, grid approach is followed for building the matrix at various seeds and kernels (RBF, Polynomial). The aim of this research is to see which is the best kernel for the dataset.

Keywords: Big Data, Data mining, Machine learning Rattle, RStudio, Support Vector Machine.

I. INTRODUCTION

SVM is a supervised machine learning model used for classification. SVM is used in text and hypertext categorization. It is also used for classification of newsfeed as – sports, entertainment and politics. [1]SVM has various extensions like support vector clustering (SVC), Multiclass SVM [2], Transductive SVM, Structured SVM, etc. some common use of support vector machine is face detection text and hypertext categorisation of classification of images bioinformatics protein food and remote homology detection, handwriting recognition generalized predictive control GPC. Airbnb is one of the most popular online market places where people book accommodation. Airbnb provides a platform to the people, offering lodging, hospitality and homestays. Support Vector Machine technique classifies rooms, on the basis of the type and rooms can be classified as entire home, private room, shared room. Support vector machine and K-means clustering is applied on the dataset and then to analyze it using a grid approach. Results are calculated at various kernels with different cost, seed value, degree parameters and correlation are plotted using spearman method.

II. RELATEDWORK

Many authors have proposed work on predictive analysis using different techniques. Jarou (2009) applied support vector machine and skin colour models for adult image detection[3] .Support vector machine is used in areas of facial detection but it is a cumbersome model. Skincare models use skin ratio, which is calculated using RGB information. A Hybrid scheme that is the combination of support vector machine and skin colour models is used to detect adult images. M. Dave (2009) proposed a parameter which acted as a churn, in the telecom industry. The decision tree model was built using R Studio. And from the computations, “gender” parameter was excluded, as it did not have any major impact on the data[4]. Wind power has become the most crucial parameter in this global warming environment. Wind power is free of cost and renewable source. Mathew S. (2019) worked on the dataset of NREL (National Renewable Energy Laboratory) using R Studio, and analysed various models on parameters- wind speed, wind direction, air temperature, air pressure and air density[5]. It was able to forecast one hour ahead. Nowadays, WhatsApp is the most usable application on our mobile phones[11]. It’s the first and the last thing we check during our day[6]. WhatsApp messages can be used analyze sentiments of people and to control circulation of fake news. Joshi S. (2016), performed sentiment analysis on WhatsApp chat database using R Studio. Wang Y. (2020) analyzed twitter data with 5,208 tweets containing hashtags such as #distracteddriving, #textanddrive, #textinganddriving using RStudio and tableau[7]. It gave the visualization of the attitudes and opinions on using mobile phones while driving.

Revised Manuscript Received on September 05, 2020.

Nainika Kaushik*, Assistant Professor, Department of Information Technology, Jagan Institute of Management Studies - JIMS Rohini, Delhi, India.

Dr. Manjot Bhatia, Professor, Jagan Institute of Management Studies - JIMS Rohini, Delhi, India.

Sonali Rastogi, Computer Science Engineering Graduate, Jagannath University Bahadurgarh, Haryana, India.

III. PROPOSED METHODOLOGY

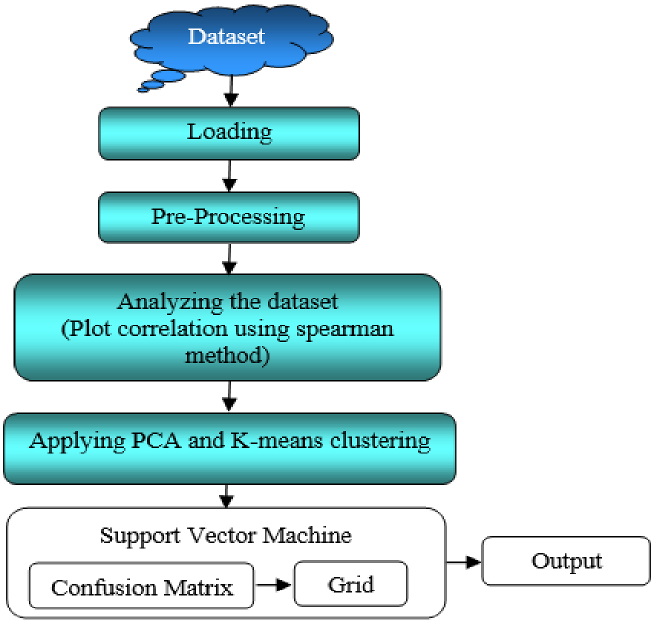


Figure 1: System Workflow

In the proposed work dataset is loaded into the R Studio. Then pre-processing takes place in which data is cleaned. All the missing values need to be treated.

The next phase is analyzing the dataset by plotting correlation using spearman method. Further applying PCA technique and K-means clustering algorithms. The last step is to implement SVM and compute confusion and grid approach. Figure 1 depicts the steps of the proposed work.

Data Collection and Description

The database of the Airbnb was fetched for 500 records for the analyzation process, it consists of the following parameters or attributes such as the house name, hostname, location, room type, number of reviews given, etc.

Data Pre-processing

Clean or pre-process the data to deal with missing values, mean was computed and entered into missing value where datatype was numerical. Otherwise for categorical data, deleted the obsolete row. So, a cleaned dataset of 500 records was created.

Support Vector Machine

SVM is used in both solving classification and regression problem statements. In linear separable system, a hyperplane is created to classify datasets into groups. Two parallel lines are created parallel to hyperplanes known as marginal planes. These marginal planes pass through nearest points of the groups. The distance between the two marginal planes is referred to as marginal distance. The points passing through marginal planes are called support vectors. In non-separable systems, SVM kernels are used. SVM kernels convert low dimensional systems to high dimensional systems. Always create a generalized model for better accuracy and the aim is to maximize marginal distance.

Software Tool Used R

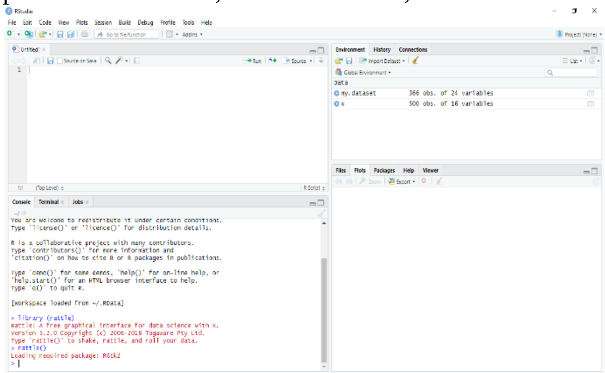
R is an IDE, that is used for statistical computing and graphics. It can work with many operating systems such as Windows NT, macOS, Ubuntu, Fedora, Red hat Linux. The languages that are used are Java, C++ and JavaScript. R- rattle is a Data mining and Data modelling statistical computing environment. One can download it from the comprehensive R archive network (CRAN) repository. R studio has only one data mining interface that is Rattle.

R: Rattle package

Include RGtk2 based graphical user interface. It allows to load CSV files. It provides functionality for data mining. To install Rattle by writing down two commands in R prompt.

>library(rattle)>rattle ()

Build number of models such as k means, tree, forest, support vector machine, Neural networks, etc.



```
Data
my.dataset     366 obs. of 24 variables
x              500 obs. of 16 variables
```

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> library (rattle)
Rattle: A free graphical interface for data science with R.
Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
Loading required package: RGtk2
> |
```

Fig. 2. R Studio Interface

IV. PROPOSED METHODOLOGY

The dataset was uploaded in the R. Studio. Rattle is used for GUI analyses. The .csv file was loaded, and partition was done as 70 % training and 30% testing.

Attributes in Dataset

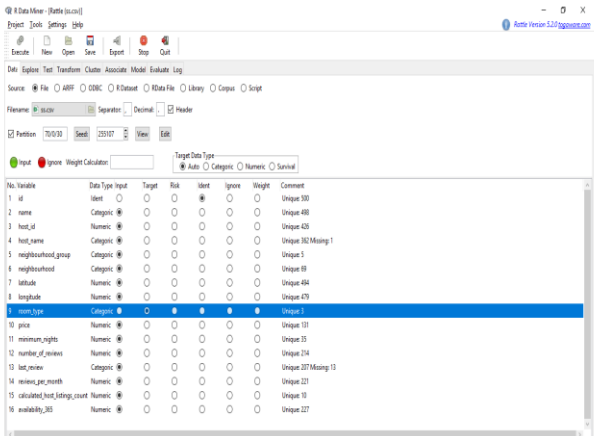
The attributes which are there in the data set are mentioned here.

Table 1: List of attributes in the used dataset.

S.No.	Attribute Name
1	Id
2	Name
3	host_id
4	host_name
5	neighborhood_group
6	Neighborhood
7	Latitude
8	Longitude
9	room_type
10	Price
11	minimum_nights
12	Number_of_reviews
13	last_review
14	reviews_per_month
15	calculated_host_listings_count
16	availability_365

Loading the dataset

Select "Data" tab on that load the csv file , then change the partition to 70 % training and 30 % testing .Click on "Execute" all the attributes will be shown with its data types , input ,target ,risk, ident, ignore, weight, comments parameters .Select target variable , selected room_type as the target variable as it is categoric. SVM can be applied only on categoric attribute.



R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute
New
Open
Save
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: ss.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/0/30 Seed: 255107 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	id	Ident	○	○	○	●	○	○	Unique: 500
2	name	Categoric	●	○	○	○	○	○	Unique: 498
3	host_id	Numeric	●	○	○	○	○	○	Unique: 426
4	host_name	Categoric	●	○	○	○	○	○	Unique: 362 Missing: 1
5	neighbourhood_group	Categoric	●	○	○	○	○	○	Unique: 5
6	neighbourhood	Categoric	●	○	○	○	○	○	Unique: 69
7	latitude	Numeric	●	○	○	○	○	○	Unique: 494
8	longitude	Numeric	●	○	○	○	○	○	Unique: 479
9	room_type	Categoric	○	●	○	○	○	○	Unique: 3
10	price	Numeric	●	○	○	○	○	○	Unique: 131
11	minimum_nights	Numeric	●	○	○	○	○	○	Unique: 35
12	number_of_reviews	Numeric	●	○	○	○	○	○	Unique: 214
13	last_review	Categoric	●	○	○	○	○	○	Unique: 207 Missing: 13
14	reviews_per_month	Numeric	●	○	○	○	○	○	Unique: 221
15	calculated_host_listings_count	Numeric	●	○	○	○	○	○	Unique: 10
16	availability_365	Numeric	●	○	○	○	○	○	Unique: 227

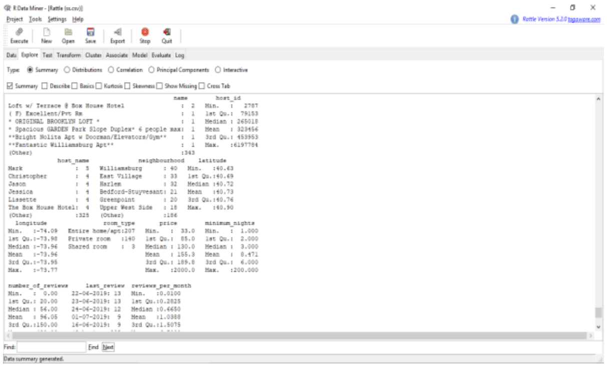
Fig. 3. Attributes in the Dataset with Target Variable

Summary of the dataset

In "Explore" tab, selecting "Summary", All the details of the dataset will be visible in detail such as datatype of attributes and statistical information such as Min, First Quartile, Median, Mean, Third Quartile and Max., for each attribute[8].

Retrieval Number: 100.1/ijeat.A16731010120
DOI:10.35940/ijeat.A1673.1010120
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

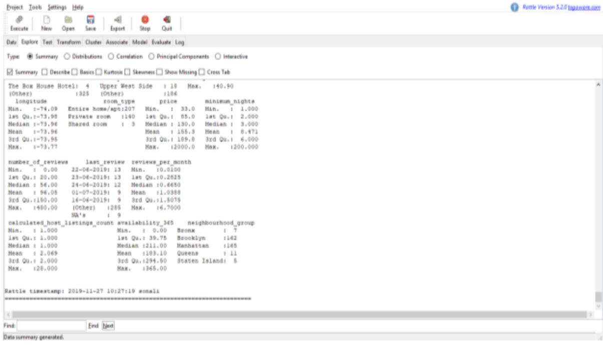
Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

☑ Summary □ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
                                                  name         host_id
Loft w/ Terrace @ Box House Hotel               :  2   Min.   :   2787
( F) Excellent/Pvt Rm                           :  1   1st Qu.:  79153
* ORIGINAL BROOKLYN LOFT *                      :  1   Median : 265018
* Spacious GARDEN Park Slope Duplex* 6 people max:  1   Mean   : 323456
**Bright Nolita Apt w Doorman/Elevators/Gym**   :  1   3rd Qu.: 453953
**Fantastic Williamsburg Apt**                  :  1   Max.   :6197784
(Other)                                         :343
          host_name          neighbourhood     latitude
Mark               :  5   Williamsburg     : 40   Min.   :40.63
Christopher        :  4   East Village     : 33   1st Qu.:40.69
Jason              :  4   Harlem           : 32   Median :40.72
Jessica            :  4   Bedford-Stuyvesant: 21   Mean   :40.73
Lissette           :  4   Greenpoint       : 20   3rd Qu.:40.76
The Box House Hotel:  4   Upper West Side  : 18   Max.   :40.90
(Other)            :325   (Other)          :186
  longitude             room_type       price          minimum_nights
Min.   :-74.09   Entire home/apt:207   Min.   :  33.0   Min.   :  1.000
1st Qu.:-73.98   Private room   :140   1st Qu.:  85.0   1st Qu.:  2.000
Median :-73.96   Shared room    :  3   Median : 130.0   Median :  3.000
Mean   :-73.96                         Mean   : 155.3   Mean   :  8.471
3rd Qu.:-73.95                         3rd Qu.: 189.8   3rd Qu.:  6.000
Max.   :-73.77                         Max.   :2000.0   Max.   :200.000

number_of_reviews     last_review   reviews_per_month
Min.   :  0.00    22-06-2019: 13   Min.   :0.0100
1st Qu.: 20.00    23-06-2019: 13   1st Qu.:0.2825
Median : 56.00    24-06-2019: 12   Median :0.6650
Mean   : 96.05    01-07-2019:  9   Mean   :1.0388
3rd Qu.:150.00    16-06-2019:  9   3rd Qu.:1.5075
```

Data summary generated.

Fig. 4(a). Summary of Dataset in detail.


Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

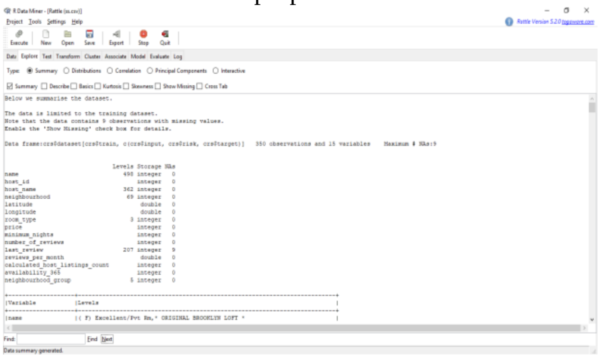
☑ Summary □ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
The Box House Hotel:  4   Upper West Side  : 18   Max.   :40.90
(Other)            :325   (Other)          :186
  longitude             room_type       price          minimum_nights
Min.   :-74.09   Entire home/apt:207   Min.   :  33.0   Min.   :  1.000
1st Qu.:-73.98   Private room   :140   1st Qu.:  85.0   1st Qu.:  2.000
Median :-73.96   Shared room    :  3   Median : 130.0   Median :  3.000
Mean   :-73.96                         Mean   : 155.3   Mean   :  8.471
3rd Qu.:-73.95                         3rd Qu.: 189.8   3rd Qu.:  6.000
Max.   :-73.77                         Max.   :2000.0   Max.   :200.000

number_of_reviews     last_review   reviews_per_month
Min.   :  0.00    22-06-2019: 13   Min.   :0.0100
1st Qu.: 20.00    23-06-2019: 13   1st Qu.:0.2825
Median : 56.00    24-06-2019: 12   Median :0.6650
Mean   : 96.05    01-07-2019:  9   Mean   :1.0388
3rd Qu.:150.00    16-06-2019:  9   3rd Qu.:1.5075
Max.   :480.00    (Other)   :285   Max.   :6.7000
                  NA's      :  9
calculated_host_listings_count availability_365   neighbourhood_group
Min.   : 1.000                 Min.   :  0.00   Bronx        :  7
1st Qu.: 1.000                 1st Qu.: 39.75   Brooklyn     :162
Median : 1.000                 Median :211.00   Manhattan    :165
Mean   : 2.069                 Mean   :183.10   Queens       : 11
3rd Qu.: 2.000                 3rd Qu.:294.50   Staten Island:  5
Max.   :28.000                 Max.   :365.00


Rattle timestamp: 2019-11-27 10:27:19 sonali
======================================================================
```

Data summary generated.

Fig. 4(b). Summary of Dataset in detail.

Fig 4(c). describes about the dataset limited to training dataset with detailed information about data frames, levels that are chosen for the said purpose.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

☑ Summary □ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

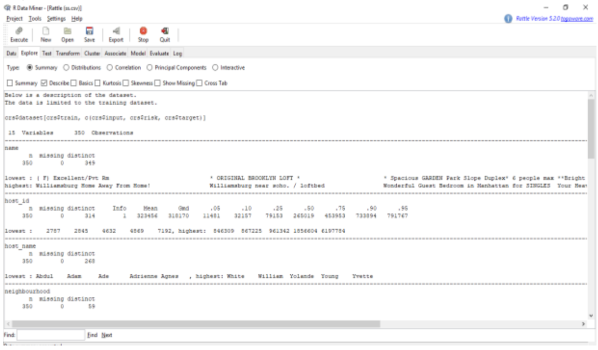
```
Below we summarise the dataset.

The data is limited to the training dataset.
Note that the data contains 9 observations with missing values.
Enable the 'Show Missing' check box for details.

Data frame:crs$dataset[crs$train, c(crs$input, crs$risk, crs$target)]   350 observations and 15 variables    Maximum # NAs:9


                               Levels Storage NAs
name                              498 integer   0
host_id                               integer   0
host_name                         362 integer   0
neighbourhood                      69 integer   0
latitude                               double   0
longitude                              double   0
room_type                           3 integer   0
price                                 integer   0
minimum_nights                        integer   0
number_of_reviews                     integer   0
last_review                       207 integer   9
reviews_per_month                      double   0
calculated_host_listings_count        integer   0
availability_365                      integer   0
neighbourhood_group                 5 integer   0

+-------------------+------------------------------------------------------------+
|Variable           |Levels                                                      |
+-------------------+------------------------------------------------------------+
|name               |( F) Excellent/Pvt Rm,* ORIGINAL BROOKLYN LOFT *            |
```

Data summary generated.

Fig. 4(c). Summary of Dataset in detail.

Describing the data

Under the explore tab and summary radio button by selecting the describe option, it shows all the attributes with the available observations and missing observations of the attributes along with their statistical interpretation such as Mean, Median etc[9].

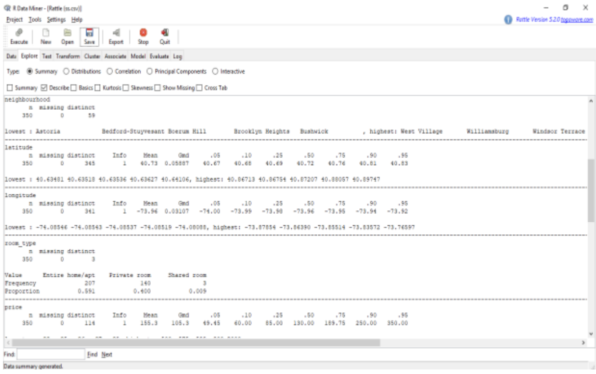

R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary ☑ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
Below is a description of the dataset.
The data is limited to the training dataset.

crs$dataset[crs$train, c(crs$input, crs$risk, crs$target)]

 15  Variables      350  Observations
--------------------------------------------------------------------------------
name
       n  missing distinct
     350        0      349

lowest : ( F) Excellent/Pvt Rm                   * ORIGINAL BROOKLYN LOFT *              * Spacious GARDEN Park Slope Duplex* 6 people max **Bright
highest: Williamsburg Home Away From Home!        Williamsburg near soho. / loftbed        Wonderful Guest Bedroom in Manhattan for SINGLES  Your Mea
--------------------------------------------------------------------------------
host_id
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      314        1   323456   318170    11481    32157    79153   265019   453953   733894   791767

lowest :    2787    2845    4632    4869    7192, highest:  846309  867225  961342 1856604 6197784
--------------------------------------------------------------------------------
host_name
       n  missing distinct
     350        0      268

lowest : Abdul    Adam     Ade      Adrienne Agnes   , highest: White    William  Yolande  Young    Yvette
--------------------------------------------------------------------------------
neighbourhood
       n  missing distinct
     350        0       59
```

Fig. 5(a). Description of Dataset in Detail


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

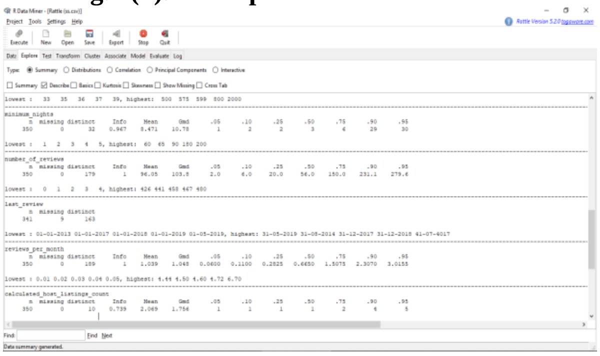
Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary ☑ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
neighbourhood
       n  missing distinct
     350        0       59

lowest : Astoria            Bedford-Stuyvesant Boerum Hill        Brooklyn Heights   Bushwick          , highest: West Village       Williamsburg       Windsor Terrace
--------------------------------------------------------------------------------
latitude
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      345        1    40.73  0.05887    40.67    40.68    40.69    40.72    40.76    40.81    40.83

lowest : 40.63481 40.63518 40.63536 40.63627 40.64106, highest: 40.86713 40.86754 40.87207 40.88057 40.89747
--------------------------------------------------------------------------------
longitude
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      341        1   -73.96  0.03107   -74.00   -73.99   -73.98   -73.96   -73.95   -73.94   -73.92

lowest : -74.08546 -74.08543 -74.08537 -74.08519 -74.00088, highest: -73.87854 -73.86390 -73.85514 -73.83572 -73.76597
--------------------------------------------------------------------------------
room_type
       n  missing distinct
     350        0        3

Value      Entire home/apt    Private room     Shared room
Frequency              207             140               3
Proportion           0.591           0.400           0.009
--------------------------------------------------------------------------------
price
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      114        1    155.3    105.3    49.45    60.00    85.00   130.00   189.75   250.00   350.00
```

Data summary generated.

Fig. 5(b). Description of Dataset in Detail


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

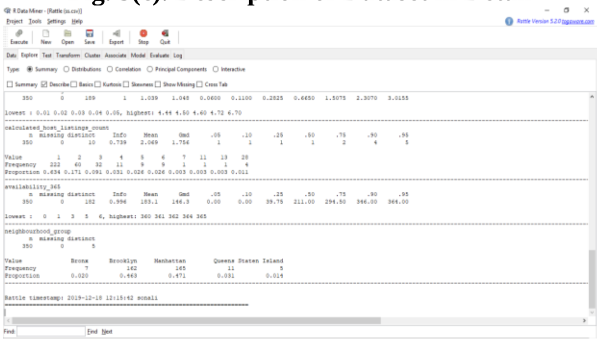
Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary ☑ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
lowest :   33   35   36   37   39, highest:  500  575  599  800 2000
--------------------------------------------------------------------------------
minimum_nights
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0       32    0.967    8.471    10.78        1        2        2        3        6       29       30

lowest :   1   2   3   4   5, highest:  60  65  90 180 200
--------------------------------------------------------------------------------
number_of_reviews
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      179        1    96.05    103.8      2.0      6.0     20.0     56.0    150.0    231.1    279.6

lowest :   0   1   2   3   4, highest: 426 441 458 467 480
--------------------------------------------------------------------------------
last_review
       n  missing distinct
     341        9      163

lowest : 01-01-2013 01-01-2017 01-01-2018 01-01-2019 01-05-2019, highest: 31-05-2019 31-08-2014 31-12-2017 31-12-2018 41-07-4017
--------------------------------------------------------------------------------
reviews_per_month
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      189        1    1.039    1.048   0.0600   0.1100   0.2825   0.6650   1.5075   2.3070   3.0155

lowest : 0.01 0.02 0.03 0.04 0.05, highest: 4.44 4.50 4.60 4.72 6.70
--------------------------------------------------------------------------------
calculated_host_listings_count
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0       10    0.739    2.069    1.756        1        1        1        1        2        4        5
```

Data summary generated.

Fig. 5(c). Description of Dataset in Detail


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

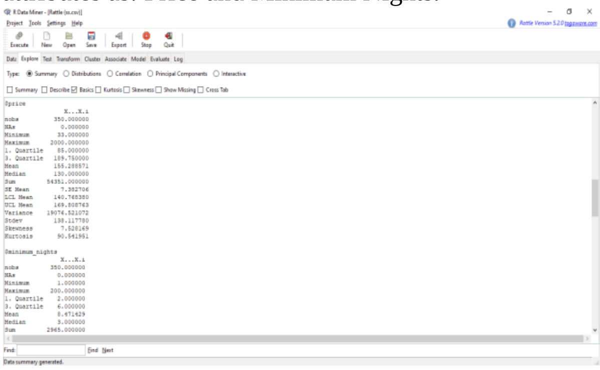
□ Summary ☑ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
     350        0      189        1    1.039    1.048   0.0600   0.1100   0.2825   0.6650   1.5075   2.3070   3.0155

lowest : 0.01 0.02 0.03 0.04 0.05, highest: 4.44 4.50 4.60 4.72 6.70
--------------------------------------------------------------------------------
calculated_host_listings_count
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0       10    0.739    2.069    1.756        1        1        1        1        2        4        5

Value          1     2     3     4     5     6     7    11    13    28
Frequency    222    60    32    11     9     9     1     1     1     4
Proportion 0.634 0.171 0.091 0.031 0.026 0.026 0.003 0.003 0.003 0.011
--------------------------------------------------------------------------------
availability_365
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     350        0      182    0.996    183.1    146.3     0.00     0.00    39.75   211.00   294.50   346.00   364.00

lowest :   0   1   3   5   6, highest: 360 361 362 364 365
--------------------------------------------------------------------------------
neighbourhood_group
       n  missing distinct
     350        0        5

Value              Bronx      Brooklyn     Manhattan        Queens Staten Island
Frequency              7           162           165            11             5
Proportion         0.020         0.463         0.471         0.031         0.014
--------------------------------------------------------------------------------

Rattle timestamp: 2019-12-18 12:15:42 sonali
======================================================================
```

Fig. 5(d). Description of Dataset in Detail

Basics about the Data

Click on "Explore" tab, Selecting Type (radio buttons) and choosing the appropriate checkbox as "BASICS". Basics tells about each numerical value of data presenting statistical information for each attribute as Skewness, Kurtosis, Variance and etc. Fig. 6(a). describes the statistical data about the attributes as: Price and Minimum Nights.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rattle Version 5.2.0 togaware.com
Find: Find Next

Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary □ Describe ☑ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
$price
                 X...X.i
nobs         350.000000
NAs            0.000000
Minimum       33.000000
Maximum     2000.000000
1. Quartile   85.000000
3. Quartile  189.750000
Mean         155.288571
Median       130.000000
Sum        54351.000000
SE Mean        7.382706
LCL Mean     140.768380
UCL Mean     169.808763
Variance   19076.521072
Stdev        138.117780
Skewness       7.528169
Kurtosis      90.541961

$minimum_nights
                X...X.i
nobs        350.000000
NAs           0.000000
Minimum       1.000000
Maximum     200.000000
1. Quartile   2.000000
3. Quartile   6.000000
Mean          8.471429
Median        3.000000
Sum        2965.000000
```

Data summary generated.

Fig. 6(a). Basics of Attribute

Fig. 6(b). describes the statistical data about the attributes as: Minimum Nights, Number of reviews and Reviews per Month.

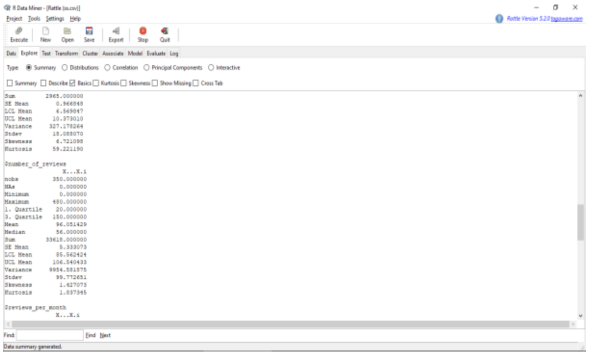

```
Sum          2965.000000
SE Mean         0.966848
LCL Mean        6.569847
UCL Mean       10.373010
Variance      327.178264
Stdev          18.088070
Skewness        6.721098
Kurtosis       59.221190

$number_of_reviews
                 X...X.i
nobs          350.000000
NAs             0.000000
Minimum         0.000000
Maximum       480.000000
1. Quartile    20.000000
3. Quartile   150.000000
Mean           96.051429
Median         56.000000
Sum         33618.000000
SE Mean         5.333073
LCL Mean       85.562424
UCL Mean      106.540433
Variance     9954.581875
Stdev          99.772651
Skewness        1.427073
Kurtosis        1.837345

$reviews_per_month
                X...X.i
```
Data summary generated.

Fig. 6(b). Basics of Attribute

Fig. 6(c). describes the statistical data about the attributes as: Reviews per Month, Calculated Host Listing Counts.

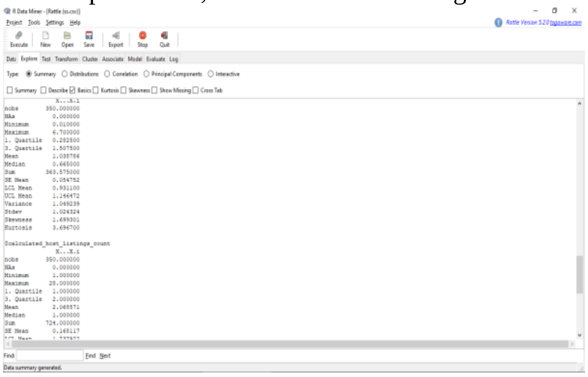

```
               X...X.i
nobs        350.000000
NAs           0.000000
Minimum       0.010000
Maximum       6.700000
1. Quartile   0.282500
3. Quartile   1.507500
Mean          1.038786
Median        0.665000
Sum         363.575000
SE Mean       0.054752
LCL Mean      0.931100
UCL Mean      1.146472
Variance      1.049239
Stdev         1.024324
Skewness      1.699301
Kurtosis      3.696700

$calculated_host_listings_count
               X...X.i
nobs        350.000000
NAs           0.000000
Minimum       1.000000
Maximum      28.000000
1. Quartile   1.000000
3. Quartile   2.000000
Mean          2.068571
Median        1.000000
Sum         724.000000
SE Mean       0.168117
LCL Mean      1.737922
```
Data summary generated.

Fig. 6(c). Basics of Attribute

Fig. 6(d). describes the statistical data about the attributes as: Calculated Host Listing Counts, Availability_365(i.e. Throughout the year).

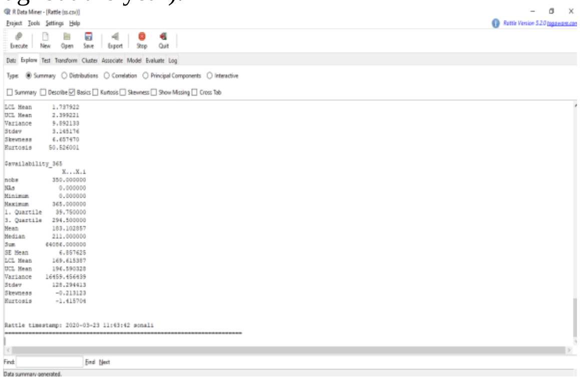

```
LCL Mean       1.737922
UCL Mean       2.399221
Variance       9.892133
Stdev          3.145176
Skewness       6.657470
Kurtosis      50.526001

$availability_365
                 X...X.i
nobs          350.000000
NAs             0.000000
Minimum         0.000000
Maximum       365.000000
1. Quartile    39.750000
3. Quartile   294.500000
Mean          183.102857
Median        211.000000
Sum         64086.000000
SE Mean         6.857625
LCL Mean      169.615387
UCL Mean      196.590328
Variance    16459.456439
Stdev         128.294413
Skewness       -0.213123
Kurtosis       -1.415704


Rattle timestamp: 2020-03-23 11:43:42 sonali
======================================================================
```
Data summary generated.

Fig. 6(d). Basics of Attribute

Knowing more about the Data

Under the explore tab and summary button by selecting the kurtosis option, it shows the peakness of all the variables around the mean. The kurtosis measure of variables can be positive and as well as negative.

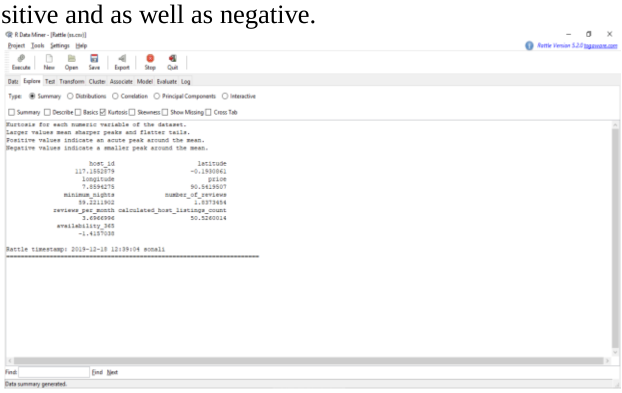

```
Kurtosis for each numeric variable of the dataset.
Larger values mean sharper peaks and flatter tails.
Positive values indicate an acute peak around the mean.
Negative values indicate a smaller peak around the mean.

           host_id                       latitude
       117.1562879                     -0.1930861
         longitude                          price
         7.8594275                     90.5419507
    minimum_nights                number_of_reviews
        59.2211902                      1.8373454
 reviews_per_month calculated_host_listings_count
         3.6966996                     50.5260014
  availability_365
        -1.4157038

Rattle timestamp: 2019-12-18 12:39:04 sonali
======================================================================
```
Data summary generated.

Fig. 7. Kurtosis of Attributes

Similarly, under the explore tab and summary button by selecting the skewness option, it shows the skewness of all the variables. The skewness tells about the tail of the distribution of the data. The skewness measure of variables can be positive, negative and as well as undefined.

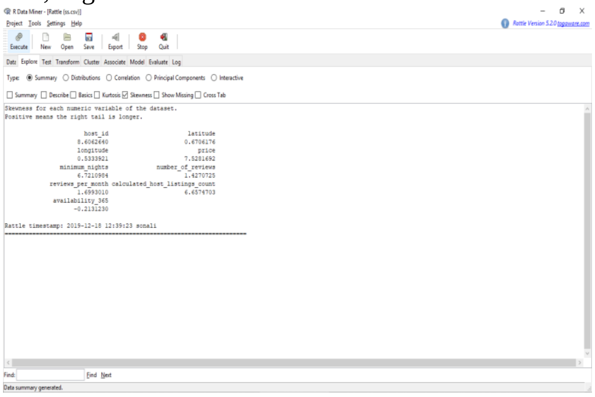

```
Skewness for each numeric variable of the dataset.
Positive means the right tail is longer.

           host_id                       latitude
         8.6062640                      0.6706176
         longitude                          price
         0.5333921                      7.5281692
    minimum_nights                number_of_reviews
         6.7210984                      1.4270725
 reviews_per_month calculated_host_listings_count
         1.6993010                      6.6574703
  availability_365
        -0.2131230

Rattle timestamp: 2019-12-18 12:39:23 sonali
======================================================================
```
Data summary generated.

Fig. 8. Kurtosis of Attributes

Under the explore tab and summary radio button by selecting the show missing checkbox, missing value summary shows up.

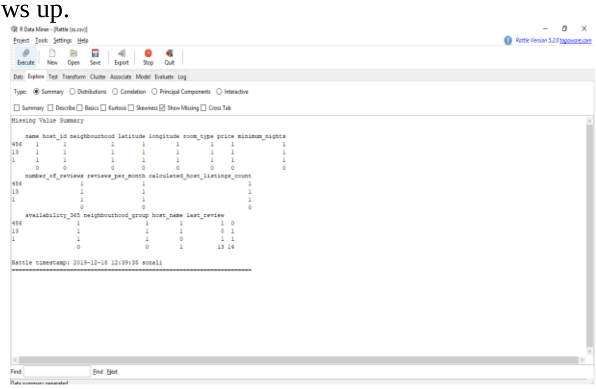

```
Missing Value Summary

    name host_id neighbourhood latitude longitude room_type price minimum_nights
486    1       1             1        1         1         1     1              1
13     1       1             1        1         1         1     1              1
1      1       1             1        1         1         1     1              1
       0       0             0        0         0         0     0              0
    number_of_reviews reviews_per_month calculated_host_listings_count
486                 1                 1                              1
13                  1                 1                              1
1                   1                 1                              1
                    0                 0                              0
    availability_365 neighbourhood_group host_name last_review
486                1                   1         1           1  0
13                 1                   1         1           0  1
1                  1                   1         0           1  1
                   0                   0         1          13 14

Rattle timestamp: 2019-12-18 12:39:35 sonali
======================================================================
```
Data summary generated.

Fig. 9. Missing Summary of Attributes

Under the explore tab and correlation button by selecting one of the methods among listed as spearman, Karl Pearson and many more. The correlation tells about the statistical dependence of the random variables.

Selecting the appropriate method, the dependence among the variables is shown the below Graph.

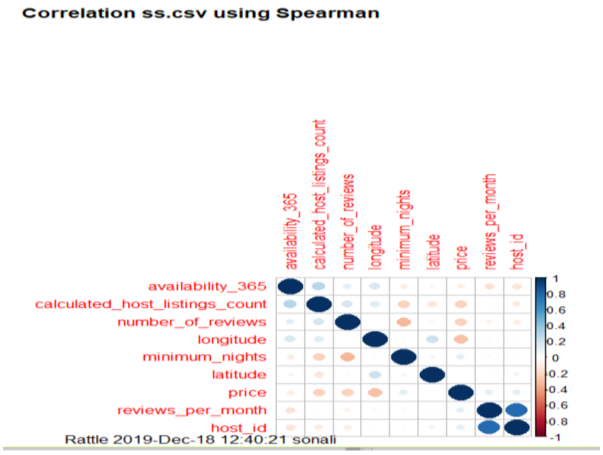


Fig. 10(a). Ordered Correlation Graph through Spearman's Rank Method

Retrieval Number: 100.1/ijeat.A16731010120
DOI:10.35940/ijeat.A1673.1010120
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

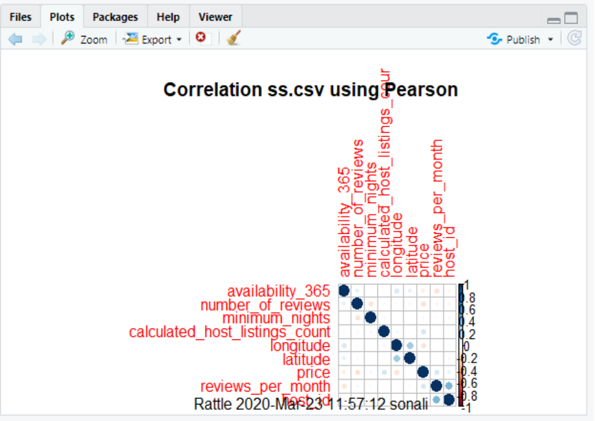


Fig. 10(b). Ordered Correlation Graph through Pearson's Method

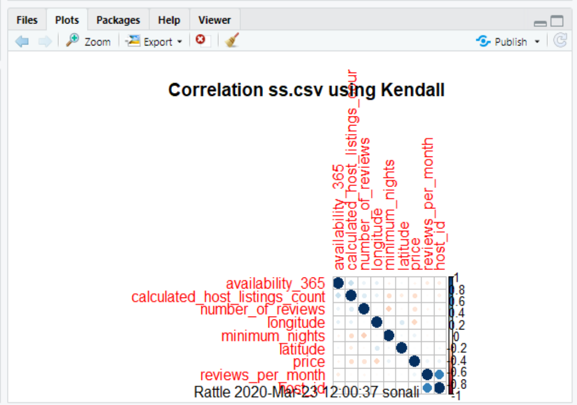


Fig. 10(c). Ordered Correlation Graph through Kendall's Method

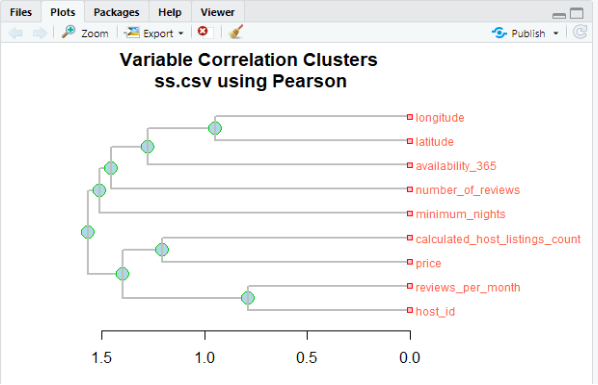


Fig. 10(d). Ordered Correlation Clusters through Pearson's Method

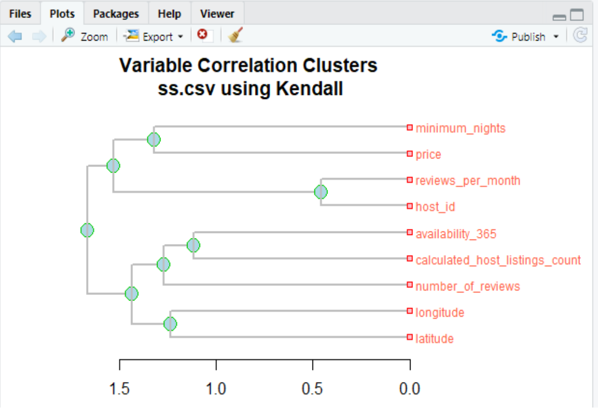


Fig. 10(e). Ordered Correlation Clusters through Kendall's Method

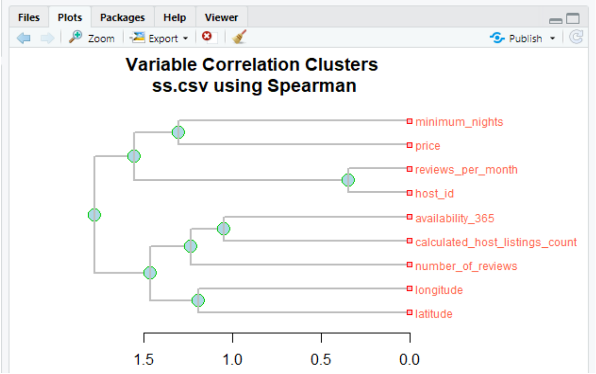
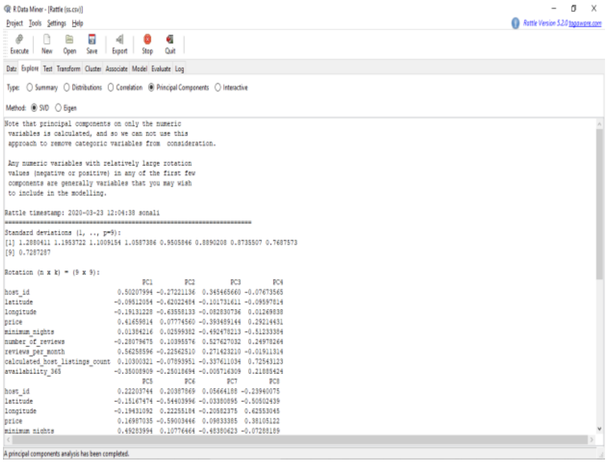


Fig. 10(f). Ordered Correlation Clusters through Spearman's Method

Principal Components About Data

Principal Components Analysis (PCA) is a method to convert the correlated variables or said 'attributes' into a smaller number of uncorrelated variables or attributes. PCA is only applicable to numerical components and so cannot be applied to categorical data.

Fig. 11(a). describes the PCA through SVD method for the dataset.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

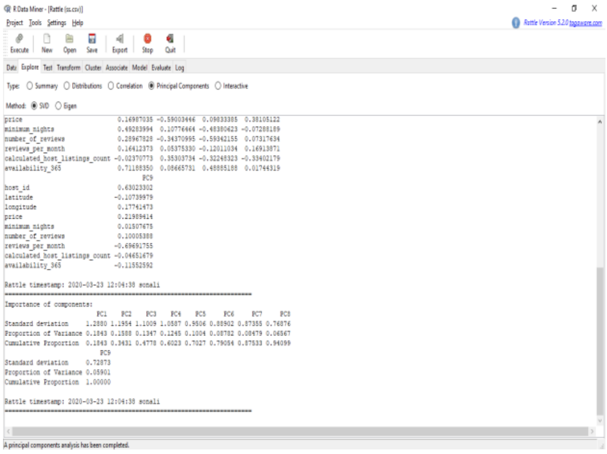
Type: ○ Summary ○ Distributions ○ Correlation ● Principal Components ○ Interactive

Method: ● SVD ○ Eigen

```
Note that principal components on only the numeric
 variables is calculated, and so we can not use this
 approach to remove categoric variables from  consideration.

 Any numeric variables with relatively large rotation
 values (negative or positive) in any of the first few
 components are generally variables that you may wish
 to include in the modelling.

Rattle timestamp: 2020-03-23 12:04:38 sonali
======================================================================
Standard deviations (1, .., p=9):
[1] 1.2880411 1.1953722 1.1009154 1.0587386 0.9505846 0.8890208 0.8735507 0.7687573
[9] 0.7287287

Rotation (n x k) = (9 x 9):
                                     PC1         PC2          PC3         PC4
host_id                       0.50207994 -0.27221136  0.345465660 -0.07673565
latitude                     -0.09512054 -0.62022484 -0.101731611 -0.09597814
longitude                    -0.19131228 -0.63558133 -0.082830736  0.01269838
price                         0.41659814  0.07774560 -0.393489144  0.29214431
minimum_nights                0.01384216  0.02599382 -0.492478213 -0.51233384
number_of_reviews            -0.28079675  0.10395576  0.527627032  0.24978264
reviews_per_month             0.56258596 -0.22562510  0.271423210 -0.01911314
calculated_host_listings_count  0.10300321 -0.07893951 -0.337611034  0.72543123
availability_365             -0.35008909 -0.25018694 -0.005716309  0.21885424
                                     PC5         PC6         PC7         PC8
host_id                       0.22203744  0.20387869  0.05664188 -0.23940075
latitude                     -0.15167474 -0.54403996 -0.03380895 -0.50502439
longitude                    -0.19431092  0.22255184 -0.20582375  0.62553045
price                         0.16987035 -0.59003446  0.09833385  0.38105122
minimum_nights                0.49283994  0.10776464 -0.48380623 -0.07288189
```

A principal components analysis has been completed.

Fig. 11(a). PCA by SVD Method


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Summary ○ Distributions ○ Correlation ● Principal Components ○ Interactive

Method: ● SVD ○ Eigen

```
price                         0.16987035 -0.59003446  0.09833385  0.38105122
minimum_nights                0.49283994  0.10776464 -0.48380623 -0.07288189
number_of_reviews             0.28967828 -0.34370995 -0.59342155  0.07317634
reviews_per_month             0.16412373  0.05375330 -0.12011034  0.16913871
calculated_host_listings_count -0.02370773  0.35303734 -0.32248323 -0.33402179
availability_365              0.71188350  0.08665731  0.48885188  0.01744319
                                     PC9
host_id                       0.63023302
latitude                     -0.10739979
longitude                     0.17741473
price                         0.21989414
minimum_nights                0.01507675
number_of_reviews             0.10005388
reviews_per_month            -0.69691755
calculated_host_listings_count -0.04651679
availability_365             -0.11552592

Rattle timestamp: 2020-03-23 12:04:38 sonali
======================================================================
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.2880 1.1954 1.1009 1.0587 0.9506 0.88902 0.87355 0.76876
Proportion of Variance 0.1843 0.1588 0.1347 0.1245 0.1004 0.08782 0.08479 0.06567
Cumulative Proportion  0.1843 0.3431 0.4778 0.6023 0.7027 0.79054 0.87533 0.94099
                           PC9
Standard deviation     0.72873
Proportion of Variance 0.05901
Cumulative Proportion  1.00000

Rattle timestamp: 2020-03-23 12:04:38 sonali
======================================================================
```

A principal components analysis has been completed.

Fig. 11(b). PCA by SVD Method

Fig. 11(c). describes the plotted graph among the principal components of dataset.

Retrieval Number: 100.1/ijeat.A16731010120
DOI:10.35940/ijeat.A1673.1010120
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

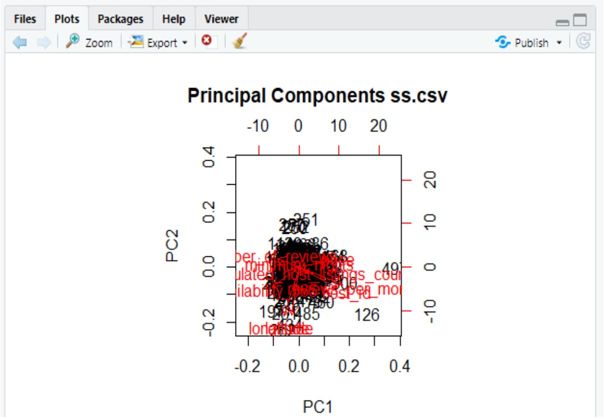
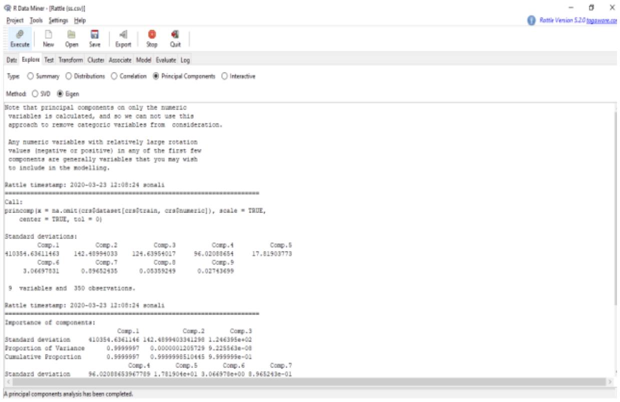


Fig. 11(c). Principal Components of Attributes

Fig. 11(d). describes the PCA through Eigen method for the dataset.



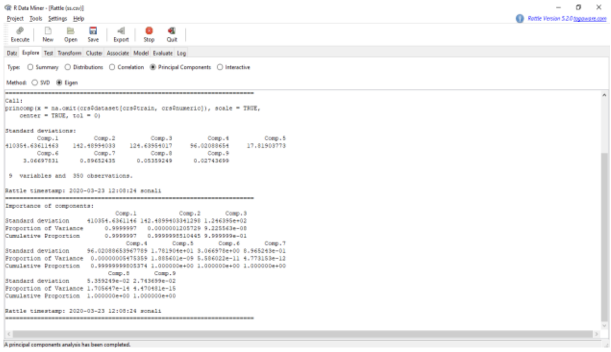
Type: ○ Summary ○ Distributions ○ Correlation ● Principal Components ○ Interactive

Method: ○ SVD ● Eigen

```
Note that principal components on only the numeric
 variables is calculated, and so we can not use this
 approach to remove categoric variables from  consideration.

 Any numeric variables with relatively large rotation
 values (negative or positive) in any of the first few
 components are generally variables that you may wish
 to include in the modelling.

Rattle timestamp: 2020-03-23 12:08:24 sonali
======================================================================
Call:
princomp(x = na.omit(crs$dataset[crs$train, crs$numeric]), scale = TRUE,
    center = TRUE, tol = 0)

Standard deviations:
        Comp.1          Comp.2          Comp.3          Comp.4          Comp.5
410354.63611463    142.48994033    124.63954017     96.02088654     17.81903773
        Comp.6          Comp.7          Comp.8          Comp.9
    3.06697831      0.89652435      0.05359249      0.02743699

 9  variables and  350 observations.

Rattle timestamp: 2020-03-23 12:08:24 sonali
======================================================================
Importance of components:
                           Comp.1            Comp.2       Comp.3
Standard deviation     410354.6361146 142.4899403341298 1.246395e+02
Proportion of Variance      0.9999997   0.0000001205729 9.225563e-08
Cumulative Proportion       0.9999997   0.9999998510445 9.999999e-01
                              Comp.4       Comp.5       Comp.6       Comp.7
Standard deviation     96.02088653967789 1.781904e+01 3.066978e+00 8.965243e-01
```

A principal components analysis has been completed.

Fig. 11(d). PCA by Eigen Method



Type: ○ Summary ○ Distributions ○ Correlation ● Principal Components ○ Interactive

Method: ○ SVD ● Eigen

```
======================================================================
Call:
princomp(x = na.omit(crs$dataset[crs$train, crs$numeric]), scale = TRUE,
    center = TRUE, tol = 0)

Standard deviations:
        Comp.1          Comp.2          Comp.3          Comp.4          Comp.5
410354.63611463    142.48994033    124.63954017     96.02088654     17.81903773
        Comp.6          Comp.7          Comp.8          Comp.9
    3.06697831      0.89652435      0.05359249      0.02743699

 9  variables and  350 observations.

Rattle timestamp: 2020-03-23 12:08:24 sonali
======================================================================
Importance of components:
                           Comp.1            Comp.2       Comp.3
Standard deviation     410354.6361146 142.4899403341298 1.246395e+02
Proportion of Variance      0.9999997   0.0000001205729 9.225563e-08
Cumulative Proportion       0.9999997   0.9999998510445 9.999999e-01
                              Comp.4       Comp.5       Comp.6       Comp.7
Standard deviation     96.02088653967789 1.781904e+01 3.066978e+00 8.965243e-01
Proportion of Variance  0.00000005475359 1.885601e-09 5.586022e-11 4.773153e-12
Cumulative Proportion   0.99999999805374 1.000000e+00 1.000000e+00 1.000000e+00
                             Comp.8       Comp.9
Standard deviation     5.359249e-02 2.743699e-02
Proportion of Variance 1.705647e-14 4.470481e-15
Cumulative Proportion  1.000000e+00 1.000000e+00

Rattle timestamp: 2020-03-23 12:08:24 sonali
======================================================================
```

A principal components analysis has been completed.

Fig. 11(e). PCA by Eigen Method

Fig. 11(f). describes the plotted graph among the principal components of dataset.

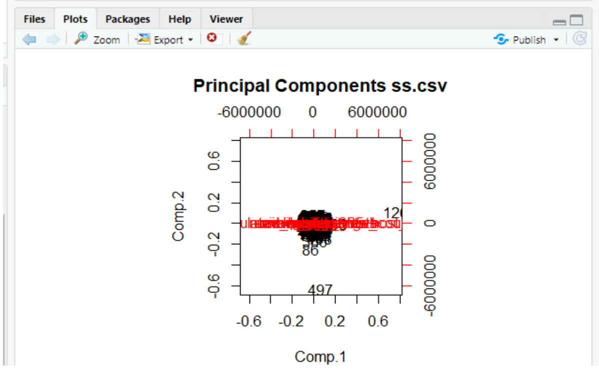


Fig. 11(f). Principal Components of Attributes

Fig. 11(g). describes the plotted graph for the principal component's importance of dataset with variance through Eigen Method.

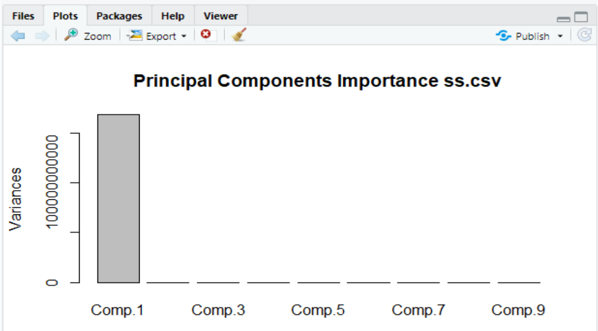


Fig. 11(g). Principal Components of Attributes

Fig. 11(h). describes the plotted graph for the principal component's importance of dataset with variance through SVD Method.

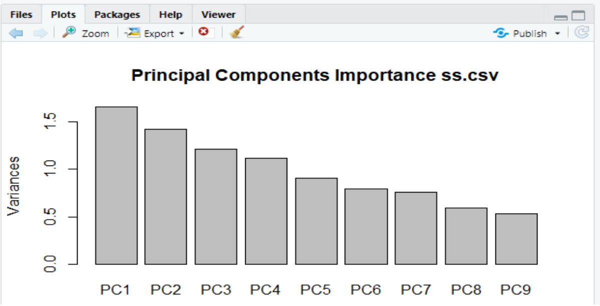
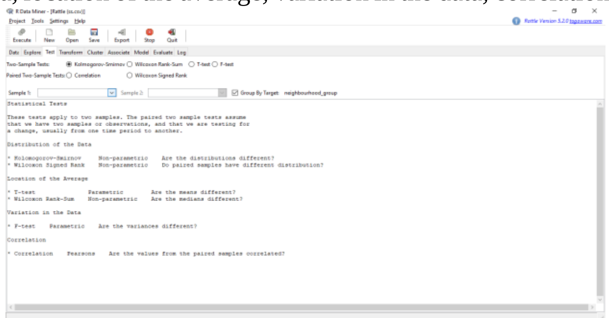


Fig. 11(h). Principal Components of Attributes Testing the Data

Test the data using the prebuild testing methods in rattle. Each testing technique has different basis for judging the technique. Such as they test the data on basis of distribution of data, location of the average, variation in the data, correlation.



Two-Sample Tests: ● Kolmogorov-Smirnov ○ Wilcoxon Rank-Sum ○ T-test ○ F-test

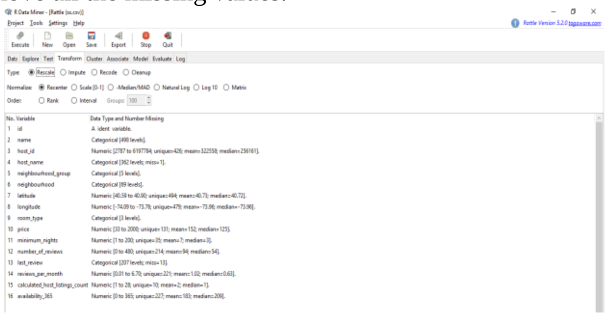
Paired Two-Sample Tests: ○ Correlation ○ Wilcoxon Signed Rank

Sample 1: Sample 2: ☑ Group By Target: neighbourhood_group

```
Statistical Tests

These tests apply to two samples. The paired two sample tests assume
that we have two samples or observations, and that we are testing for
a change, usually from one time period to another.

Distribution of the Data

* Kolmogorov-Smirnov     Non-parametric    Are the distributions different?
* Wilcoxon Signed Rank   Non-parametric    Do paired samples have different distribution?

Location of the Average

* T-test                Parametric        Are the means different?
* Wilcoxon Rank-Sum     Non-parametric    Are the medians different?

Variation in the Data

* F-test    Parametric    Are the variances different?

Correlation

* Correlation    Pearsons    Are the values from the paired samples correlated?
```

Fig. 12. Testing the Attributes Transform the data

Rescale the data, it shows the details of all the variables, its datatype and number missing. Clean the data at this stage, remove all the missing values.



Type: ● Rescale ○ Impute ○ Recode ○ Cleanup

Normalize: ● Recenter ○ Scale [0-1] ○ -Median/MAD ○ Natural Log ○ Log 10 ○ Matrix

Order: ○ Rank ○ Interval Groups: 100

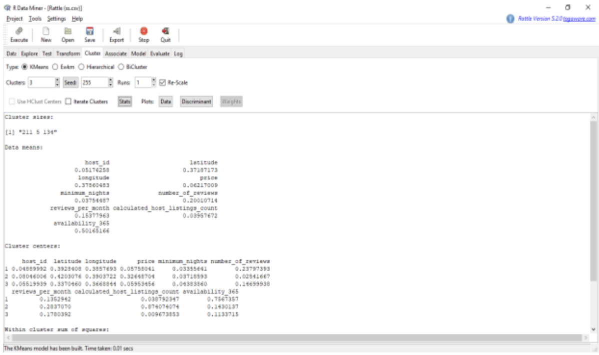
```
No. Variable                      Data Type and Number Missing
1   id                            A ident variable.
2   name                          Categorical [498 levels].
3   host_id                       Numeric [2787 to 6197784; unique=426; mean=322558; median=256161].
4   host_name                     Categorical [362 levels; miss=1].
5   neighbourhood_group           Categorical [5 levels].
6   neighbourhood                 Categorical [89 levels].
7   latitude                      Numeric [40.59 to 40.90; unique=494; mean=40.73; median=40.72].
8   longitude                     Numeric [-74.09 to -73.78; unique=479; mean=-73.96; median=-73.96].
9   room_type                     Categorical [3 levels].
10  price                         Numeric [33 to 2000; unique=131; mean=152; median=125].
11  minimum_nights                Numeric [1 to 200; unique=35; mean=7; median=3].
12  number_of_reviews             Numeric [0 to 480; unique=214; mean=94; median=54].
13  last_review                   Categorical [207 levels; miss=13].
14  reviews_per_month             Numeric [0.01 to 6.70; unique=221; mean=1.02; median=0.63].
15  calculated_host_listings_count Numeric [1 to 28; unique=10; mean=2; median=1].
16  availability_365              Numeric [0 to 365; unique=227; mean=183; median=206].
```

Fig. 13. Rescaling of Attributes

Cluster Formation

Under the Cluster tab by selecting the suitable model for clustering such as Kmeans in the below picture. Necessary details such as number of clusters to be formed, seeds, runs, re-scaling option, Use HClust Centers, Iterate Clusters, stats tab, and plotting tab: data, discriminant, weights have to be entered accordingly to the dataset entered.

Kmeans method is used for cluster formation and enter details such as required clusters as three, setting the seed level and number of runs and rescaling it.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Stats
Plots: Data Discriminant Weights

Type: ● KMeans ○ Ewkm ○ Hierarchical ○ BiCluster

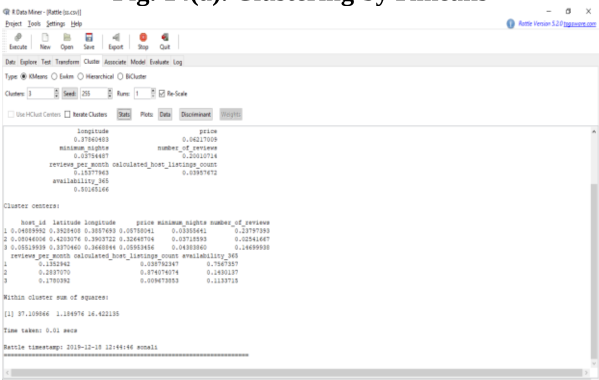
Clusters: 3 Seed: 255 Runs: 1 ☑ Re-Scale

□ Use HClust Centers □ Iterate Clusters

```
Cluster sizes:

[1] "211 5 134"

Data means:

                 host_id                      latitude
              0.05176258                    0.37187173
               longitude                         price
              0.37860483                    0.06217009
          minimum_nights             number_of_reviews
              0.03754487                    0.20010714
       reviews_per_month calculated_host_listings_count
              0.15377963                    0.03957672
        availability_365
              0.50165166

Cluster centers:

     host_id  latitude longitude      price minimum_nights number_of_reviews
1 0.04889992 0.3928408 0.3857693 0.05758041     0.03355641        0.23797393
2 0.08046006 0.4203076 0.3903722 0.32648704     0.03718593        0.02541667
3 0.05519939 0.3370460 0.3668844 0.05953456     0.04383860        0.14699938
  reviews_per_month calculated_host_listings_count availability_365
1         0.1352942                    0.038792347        0.7567357
2         0.2837070                    0.874074074        0.1430137
3         0.1780392                    0.009673853        0.1133715

Within cluster sum of squares:
```

The KMeans model has been built. Time taken: 0.01 secs

Fig. 14(a). Clustering by Kmeans


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Stats
Plots: Data Discriminant Weights

Type: ● KMeans ○ Ewkm ○ Hierarchical ○ BiCluster

Clusters: 3 Seed: 255 Runs: 1 ☑ Re-Scale

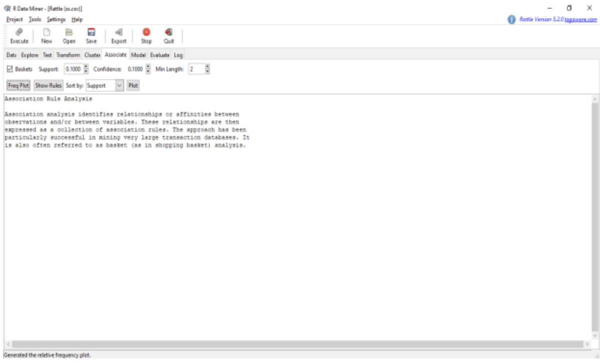
□ Use HClust Centers □ Iterate Clusters

```
               longitude                         price
              0.37860483                    0.06217009
          minimum_nights             number_of_reviews
              0.03754487                    0.20010714
       reviews_per_month calculated_host_listings_count
              0.15377963                    0.03957672
        availability_365
              0.50165166

Cluster centers:

     host_id  latitude longitude      price minimum_nights number_of_reviews
1 0.04889992 0.3928408 0.3857693 0.05758041     0.03355641        0.23797393
2 0.08046006 0.4203076 0.3903722 0.32648704     0.03718593        0.02541667
3 0.05519939 0.3370460 0.3668844 0.05953456     0.04383860        0.14699938
  reviews_per_month calculated_host_listings_count availability_365
1         0.1352942                    0.038792347        0.7567357
2         0.2837070                    0.874074074        0.1430137
3         0.1780392                    0.009673853        0.1133715

Within cluster sum of squares:

[1] 37.109866  1.184976 16.422135

Time taken: 0.01 secs

Rattle timestamp: 2019-12-18 12:44:46 sonali
======================================================================
```

Fig. 14(b). Clustering by Kmeans

Association of Data

Association Analysis is widely used in Data Mining. Rattle is supporting Association Mining rules through the tab of "Association" of unsupervised paradigm. It is based on market basket analysis and has two dependent factors as Support and Confidence. Either ident or Target Variable is used by rattle for the association analysis. Retail business widely use this approach for their purpose.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Freq Plot
Show Rules
Plot

☑ Baskets Support: 0.1000 Confidence: 0.1000 Min Length: 2

Sort by: Support

```
Association Rule Analysis

Association analysis identifies relationships or affinities between
observations and/or between variables. These relationships are then
expressed as a collection of association rules. The approach has been
particularly successful in mining very large transaction databases. It
is also often referred to as basket (as in shopping basket) analysis.
```

Generated the relative frequency plot.

Fig. 14(c). Association Rule Analysis

Association rule mining is used and below is the plotted relation among the variables.

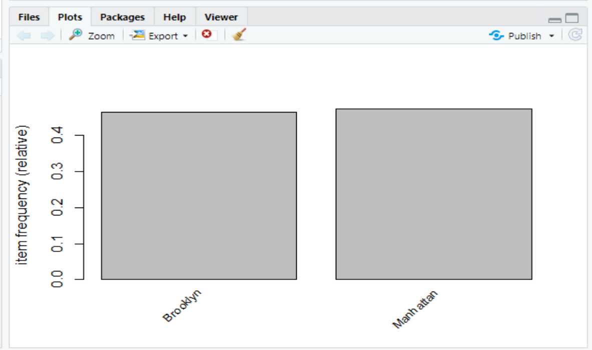
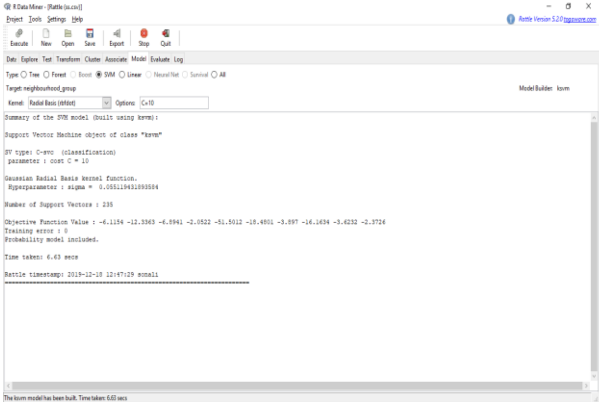


Fig. 14(d). Association Analysis between the variables

Model Selection

Here under the model bar select the suitable type of model and set the required parameters such as kernel, degree etc.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Tree ○ Forest ○ Boost ● SVM ○ Linear ○ Neural Net ○ Survival ○ All

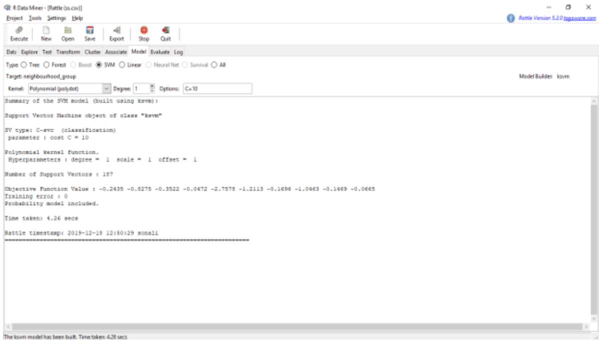
Target: neighbourhood_group Model Builder: ksvm

Kernel: Radial Basis (rbfdot) Options: C=10

```
Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 10

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0551194318935B4

Number of Support Vectors : 235

Objective Function Value : -6.1154 -12.3363 -6.8941 -2.0522 -51.5012 -18.4801 -3.897 -16.1634 -3.6232 -2.3726
Training error : 0
Probability model included.

Time taken: 6.63 secs

Rattle timestamp: 2019-12-18 12:47:29 sonali
======================================================================
```

The ksvm model has been built. Time taken: 6.63 secs

Fig. 15(a).SVM by Radial Basis (rbfdot)


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Tree ○ Forest ○ Boost ● SVM ○ Linear ○ Neural Net ○ Survival ○ All

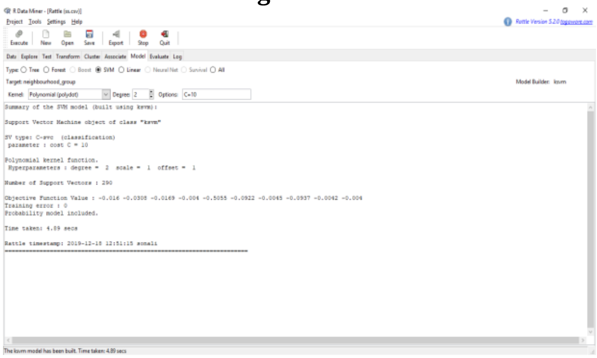
Target: neighbourhood_group Model Builder: ksvm

Kernel: Polynomial (polydot) Degree: 1 Options: C=10

```
Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 10

Polynomial kernel function.
 Hyperparameters : degree =  1  scale =  1  offset =  1

Number of Support Vectors : 187

Objective Function Value : -0.2435 -0.8275 -0.3522 -0.0472 -2.7578 -1.2113 -0.1696 -1.0463 -0.1469 -0.0665
Training error : 0
Probability model included.

Time taken: 4.26 secs

Rattle timestamp: 2019-12-18 12:50:29 sonali
======================================================================
```

The ksvm model has been built. Time taken: 4.26 secs

Fig. 15(b). SVM by Polynomial (POLYDOT) with degree =1


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Tree ○ Forest ○ Boost ● SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: neighbourhood_group Model Builder: ksvm

Kernel: Polynomial (polydot) Degree: 2 Options: C=10

```
Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 10

Polynomial kernel function.
 Hyperparameters : degree =  2  scale =  1  offset =  1

Number of Support Vectors : 290

Objective Function Value : -0.016 -0.0308 -0.0169 -0.004 -0.5055 -0.0922 -0.0045 -0.0937 -0.0042 -0.004
Training error : 0
Probability model included.

Time taken: 4.89 secs

Rattle timestamp: 2019-12-18 12:51:15 sonali
======================================================================
```

The ksvm model has been built. Time taken: 4.89 secs

Fig. 15(c). SVM by Polynomial (POLYDOT) with degree =2

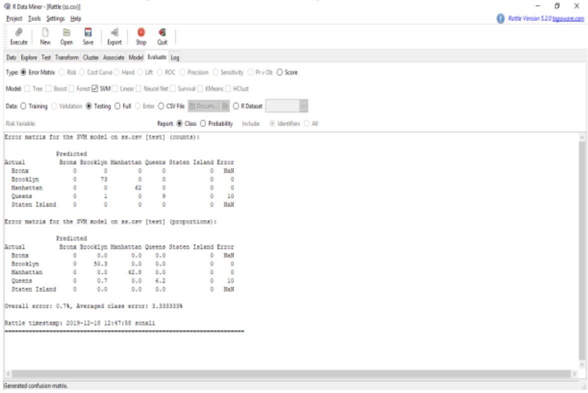
Evaluating the data

Here under the Evaluate bar choose the type and model and select the data mode i.e. testing, training, etc., risk variable and report type. Note that the error matrix generated will be different for each seed and depending upon the model chosen.

The evaluation tab basically is used for the evaluation of the performance of the models chosen for the selected dataset. Evaluation tab offers various types of evaluation models such as Error matrix, risk, cost curve, hand, left, ROC, sensitivity , Pr v Ob, Score and models that are available in the evaluation tab follows as- decision tree, Ada boost algorithm, random forest algorithm, SVM, Neural networks, survival, K means, HClust[10].

To select the type of evaluation for the selected data set one can choose among the types of radio buttons shown. The model for the evaluation can be chosen by selecting the appropriate checkbox. Also, for the data one has to select the radio button among the listed ones. Also select the appropriate report button among the listed ones.

Fig. 16(a) and Fig.16(b) display the error matrix for the dataset. The error matrix displays the predicted as well as the overall error along with the average class error.


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest ☑ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

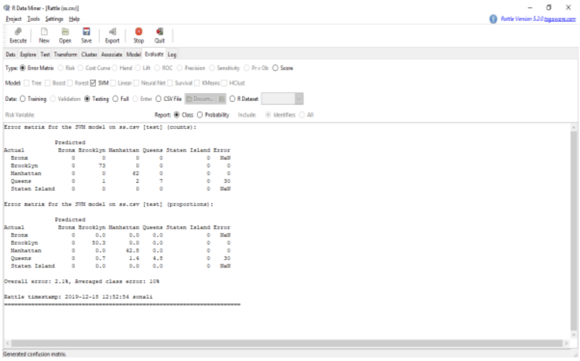
Data: ○ Training ○ Validation ● Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

```
Error matrix for the SVM model on ss.csv [test] (counts):

               Predicted
Actual          Bronx Brooklyn Manhattan Queens Staten Island Error
  Bronx             0        0         0      0             0   NaN
  Brooklyn          0       73         0      0             0     0
  Manhattan         0        0        62      0             0     0
  Queens            0        1         0      9             0    10
  Staten Island     0        0         0      0             0   NaN

Error matrix for the SVM model on ss.csv [test] (proportions):

               Predicted
Actual          Bronx Brooklyn Manhattan Queens Staten Island Error
  Bronx             0      0.0       0.0    0.0             0   NaN
  Brooklyn          0     50.3       0.0    0.0             0     0
  Manhattan         0      0.0      42.8    0.0             0     0
  Queens            0      0.7       0.0    6.2             0    10
  Staten Island     0      0.0       0.0    0.0             0   NaN

Overall error: 0.7%, Averaged class error: 3.333333%

Rattle timestamp: 2019-12-18 12:47:58 sonali
======================================================================
```

Generated confusion matrix.

Fig. 16(a). Error Matrix through SVM


R Data Miner - [Rattle (ss.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest ☑ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ○ Training ○ Validation ● Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

```
Error matrix for the SVM model on ss.csv [test] (counts):

               Predicted
Actual          Bronx Brooklyn Manhattan Queens Staten Island Error
  Bronx             0        0         0      0             0   NaN
  Brooklyn          0       73         0      0             0     0
  Manhattan         0        0        62      0             0     0
  Queens            0        1         2      7             0    30
  Staten Island     0        0         0      0             0   NaN

Error matrix for the SVM model on ss.csv [test] (proportions):

               Predicted
Actual          Bronx Brooklyn Manhattan Queens Staten Island Error
  Bronx             0      0.0       0.0    0.0             0   NaN
  Brooklyn          0     50.3       0.0    0.0             0     0
  Manhattan         0      0.0      42.8    0.0             0     0
  Queens            0      0.7       1.4    4.8             0    30
  Staten Island     0      0.0       0.0    0.0             0   NaN

Overall error: 2.1%, Averaged class error: 10%

Rattle timestamp: 2019-12-18 12:52:54 sonali
======================================================================
```

Generated confusion matrix.

Fig. 16(b). Error Matrix through SVM

Grid Approach

Grid matrix is a structured approach of displaying the data in a two-dimensional format. One can change the size of a matrix accordingly to the data set or dynamically. One can form grid Matrix in R studio by importing a package of grid layout function[12].

Seed is a vector in R that stores random number that are declared by normal generator for random number generator RNG. Useful in reproducing random objects and creating simulations.

In the Table 2, it can be clearly seen that 5 splits which corresponds to different seeds value. Here data set is analyzed by using various different kernels, degree and cost parameters. It can be seen that polynomial kernel gives the least overall average error (18.64%) with degree as 2, cost parameters as 1 and 10.

Table 2: Grid Matrix

Split	Seed	RBF C=1	RBF C=10	Poly D=1, C=1	Poly D=1, C=10	Poly D=2, C=1	Poly D=2, C=10
Split 1	42	20.70%	22.10%	23.50%	23.50%	21.40%	21.40%
Split 2	255107	24.10%	21.50%	18.70%	18.70%	22.10%	22.10%
Split 3	894393	13%	13.60%	15%	15%	11.50%	11.50%
Split 4	114762	19.30%	21.40%	22.10%	22.10%	21.40%	21.40%
Split 5	307054	20.20%	16.10%	17.50%	17.50%	16.80%	16.80%
Average		19.46%	18.94%	19.36%	19.36%	18.64%	18.64%

V. CONCLUSION

Paper analyses a dataset of 500 records in this support vector machine model and K-means clustering is applied. The dataset is first cleaned and then confusion matrix is evaluated. Then grid approach is applied at different seeds. Various kernels are evaluated such as polynomial, radial basis at different degree and cost parameters for SVM. Observed that the best kernel for the dataset is polynomial with degree 2 and cost parameters as 1,2.

REFERENCES

1. Mayor, Shweta & Pant, Bhasker. (2012). Document Classification Using Support Vector Machine. International Journal of Engineering Science and Technology.
2. Duan, Kai-Bo; Keerthi, S. Sathiya (2005). "Which Is the Best Multiclass SVM Method? An Empirical Study" (PDF). Multiple Classifier

Systems. LNCS. 3541.pp. 278-285. CiteSeerX 10.1.1.110.6789. doi: 10.1007/11494683_28. ISBN 978-3-540-26306-7.

3. B. Choi, B. Chung and J. Ryou, "Adult Image Detection Using Bayesian Decision Rule Weighted by SVM Probability," 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, 2009, pp. 659-662.
4. Vani Kapoor Nijhawan, Mamta Madan, Meenu Dave (2009). A Comparative Analysis Using RStudio for Churn Prediction .International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,Volume-8 Issue-7S2, May 2019
5. Pawar A., Jape V.S., Mathew S. (2019) Wind Power Forecasting Using Support Vector Machine Model in RStudio. In: Mallick P., Balas V., Bhoi A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore
6. Joshi S. (2019) Sentiment Analysis on WhatsApp Group Chat Using R. In: Shukla R., Agrawal J., Sharma S., Singh Tomer G. (eds) Data, Engineering and Applications. Springer, Singapore
7. Qian C., Li Y., Zuo W., Wang Y. (2020) Analysis of Driving Safety and Cellphone Use Based on Social Media. In: Stanton N. (eds) Advances in Human Factors of Transportation. AHFE 2019. Advances in Intelligent Systems and Computing, vol 964. Springer, Cham
8. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2017). URL https://www.R-project.org/
9. Zhao, Y.: R Reference Card for Data Mining. http://www.Rdatamining.com Online. Access 10 June 2017
10. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Cambridge (2016)
11. Patil, S.: WhatsApp group data analysis with R. Int. J. Comput. Appl. 154(4) (2016)
12. Torgo, L.: Data Mining with R Learning with Case Studies. CRC Press, Taylor & Francis Group an Informa Business (2011)

AUTHORS PROFILE

Nainika Kaushik She is working as an assistant professor in department of information technology, JIMS. She has completed her MTech from IGDTUW with the specialization in mobile and pervasive computing and B. tech from GGSIPU in computer science and engineering. Her areas of interest are data mining, mobile computing and software testing. She has published many papers in international journals.

Manjot Kaur Bhatia She is working as a professor in department of information technology, JIMS. She has completed her M.C.A., M.Phil. and Ph.D. from University of Delhi. Her research topic is in the area of Information Security. She has more than 20 years of teaching experience. She is actively involved in teaching and research in the areas of Security, Databases, Linux, Operating System and Information Security. Her other areas of interest include Cloud Computing, Steganography, Data Hiding, and Software testing. She has been guiding many Ph. D. scholars in their research work and motivated MCA students also towards the research work. She has published many research papers in various refereed International journals and has presented number of papers in international conferences. She has been the session chair in various international conferences.

Sonali Rastogi She has completed her B. Tech from Jagan Nath University, Bahadurgarh (JIMS) with the specialization in computer science and engineering. She has recently published a paper in international journal. Her areas of interest are influential marketing, data mining, and software testing.