ROLLING SALES BROOKLYN

by

ASHISH KUMAR VEERANKI

U00820878

CS 7720 Data Mining Course Project

TABLE OF CONTENTS

[ABSTRACT 3](#_Toc449391156)

[INTRODUCTION 3](#_Toc449391157)

[BACKGROUND FOR THE STUDY 4](#_Toc449391158)

[STATEMENT OF THE PROBLEM 5](#_Toc449391159)

[PURPOSE OF THE STUDY 5](#_Toc449391160)

[DATA ANALSYSIS 5](#_Toc449391166)

* [FINDINGS 5](#_Toc449391167)
* [DATA CLEANING & OUTLIER DETECTION 7](#_Toc449391167)
* [DATA CLASSIFICATION 11](#_Toc449391167)
* [DATA CLUSTERING 19](#_Toc449391167)

[CONSLUSIONS 32](#_Toc449391167)

[SUMMARY 32](#_Toc449391168)

[REFERENCES 32](#_Toc449391169)

# ABSTRACT

This paper shows the experiences picked up from applying data mining techniques with the end goal of foreseeing the Brooklyn building deals values in view of assortment of variables. A dataset of more than 20K transactions in land properties was utilized. The dataset included around 19 attributes which caught data about the exchanges occurring for the past few years. The outcomes from applying the data mining strategies to foresee the building sales values are promising.

# INTRODUCTION

The variables that decide housing prices are important to urban organizers, designers, land experts and money related officials and the vast majority of the American Homeowners. As indicated by a 1998 Federal Reserve review, 66.2% of U.S. family units are property holders and housing investment adds up to 33% of family total assets. The quantity of new home deals and home resales are an essential part of the U.S. economy and information concerning these transactions is firmly followed with the end goal of gauging financial action. This paper looks at the elements that decide the housing prices in an sample of more than 20K home deals in Brooklyn, New York amid the time of Aug 2012 to Aug 2013.

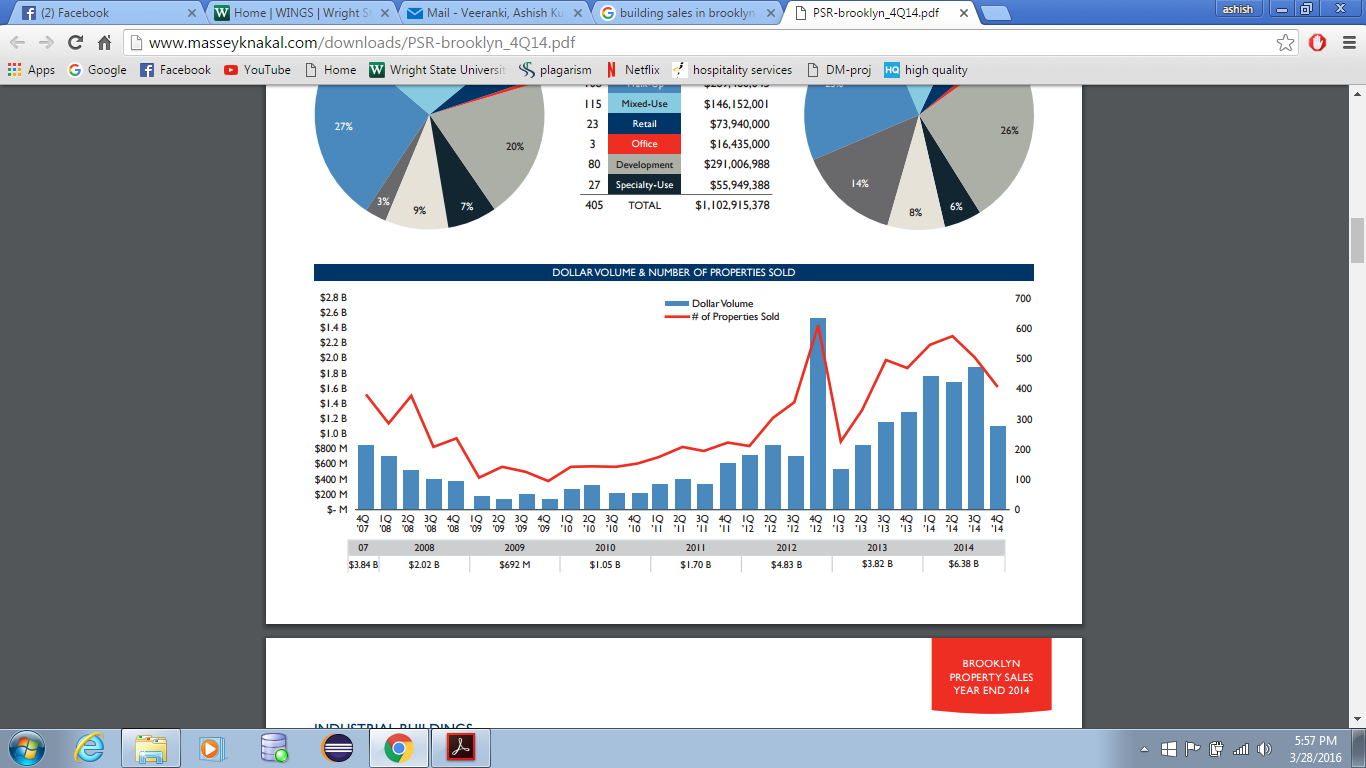
Offers of homes occur in the commercial center managed by the typical guidelines of supply and request. Since this is not an impeccable commercial center, there is an awesome scope for judgment in touching base at the offering value, along these lines this sort of employment is much similar to of a craftsmanship than science. There are 3 well known methodologies for assessing the property estimation: a) practically identical deals strategy, b)cost technique, c) income method. Based on the type of property, a degree of priority is assigned to each method used.

Essentially, none of the three strategies is foolproof and at last the appraiser needs to make a careful decision. However, more often than not, the clients who plan to purchase the property will take a gander at the encompassing and the duty class which will characterize the deal cost upto a gigantic degree. This study utilizes knowledge discovery techniques for example, data cleaning, classification and clustering with the assistance of different tools like WEKA and R studio.

# BACKGROUND FOR THE STUDY

Brooklyn is one of the 5 boroughs of New York and has the highest population when compared to other borough’s. In the past few years i.e., since 2006 there was no proper sales of properties until the third and fourth quarter of 2012 which had a dramatic change in the property sales. To analyse on what basis there was a dramatic change in the sales of buildings, that particular period from Aug 2012 to Aug 2013 is considered. In this project we are perofrming some data mining functionalities on a data set which has the details of Brooklyn building sales bought from Aug 2012 to Aug 2013. All the details such as building class category, address, year built, tax class at time of sale, sale price, sale date etc. are collected to perform several data mining activities. In this project several data mining functionalities such as data cleaning, classification and clustering are performed on the data set and the outputs are visualized to get a deep understanding of the data.

We are using R studio and WEKA tool to perform the mining activities on the data set considered. Finally the dataset has been migrated into the database (i.e., SQL Developer) and is being displayed.



# STATEMENT OF THE PROBLEM

This initial phase of a data mining project focuses on understanding the project objectives and requirements. Once you have specified the project from a business perspective, you can formulate it as a data mining problem and develop a preliminary implementation plan.

For example, business problem might be: "How can I sell more of the buildings to customers?" You might translate this into a data mining problem such as: "Which customers are most likely to purchase the buildings?" A model that predicts who is most likely to purchase the building must be built on data that describes the customers who have purchased that particular category of buildings in their appropriate locations in the past. Before building the model, you must assemble the data that is likely to contain relationships between customers who have purchased the buildings and customers who have not purchased the buildings. The various attributes which are required are the building class category, tax class at the time of built, present tax class, year of built, neighborhood, sale price, year built.

# PURPOSE OF THE STUDY

The purpose of the study is to analyze the sales of the buildings in the year Aug 2012 – Aug 2013 and predict a model which helps the customers, to urban organizers, designers, land experts and money related officials to decide on what factors the sales are impacted. Also the dataset considered has noisy and improper values which are affecting the model, so data cleaning is performed after which the data is classified and clustered based on the attributes of the dataset. Pictorial representation of the analysis is also being performed.

# DATA ANALSYSIS

1. **Findings**

**General Statistics of Data**: Data sets are comprised of Data items. Data items(data objects) are normally depicted by attributed. Data items(objects) can likewise be alluded to as tests, illustrations, examples, information focuses, or objects. Data Objects are known as data tuples if they are stored in the database. That is, the lines of a database relate to the data objects, and the columns are known as attributes. The type of an attribute is determined by the arrangement of conceivable qualities the attribute can have—nominal, binary, ordinal, or numeric. In our project, for the dataset selected we have 2 types of attributes- nominal and numeric.

**Nominal Attributes:** Nominal signifies "identifying with names." The estimations of an nominal attribute are images or names of things. Every quality represents some sort of class, code, or state, thus nominal attributes are additionally referred to as categorical.The qualities don't have any meaningful order.

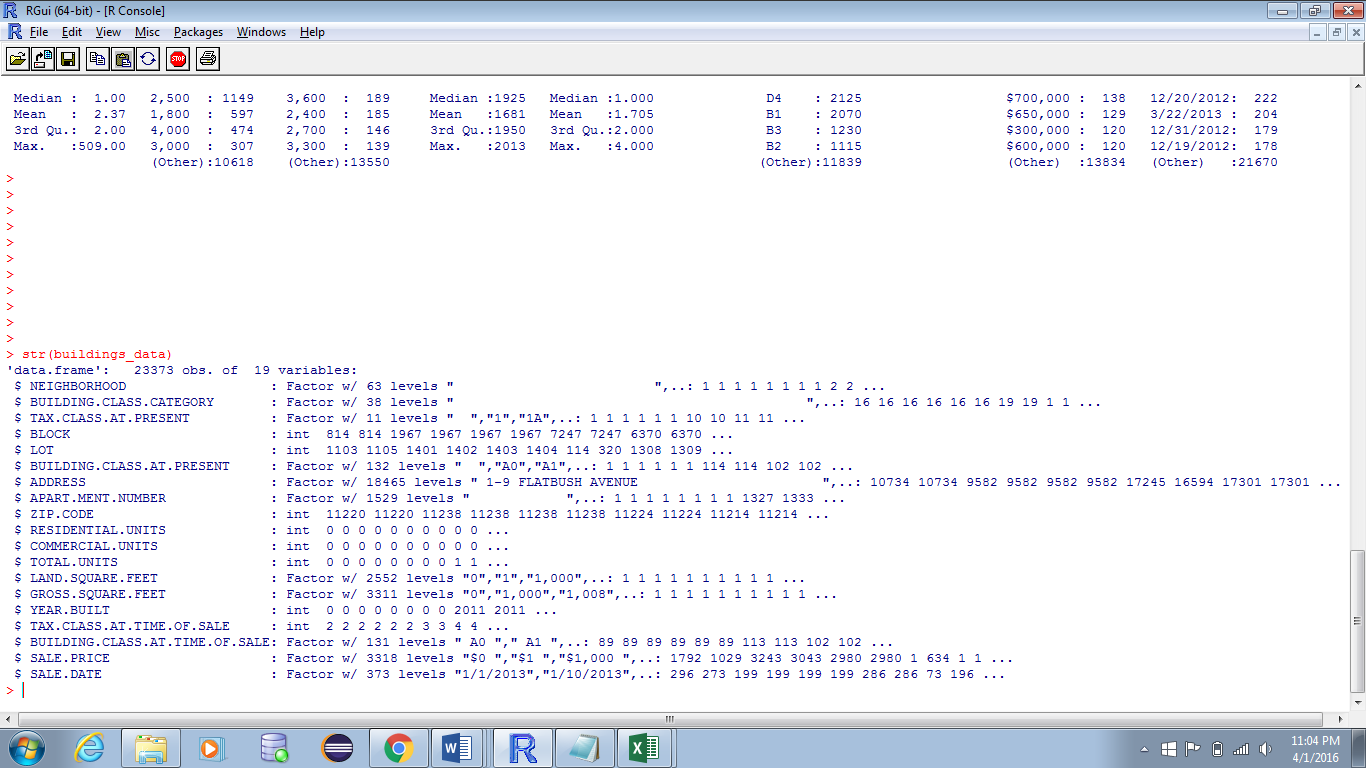
The nominal attributes which are present in the dataset considered are

Neighbourhood, Building Class Category, Tax Class at Present, Building Class at present, Address, Building Class at Time of Sale, Sale Price, Sale Date.

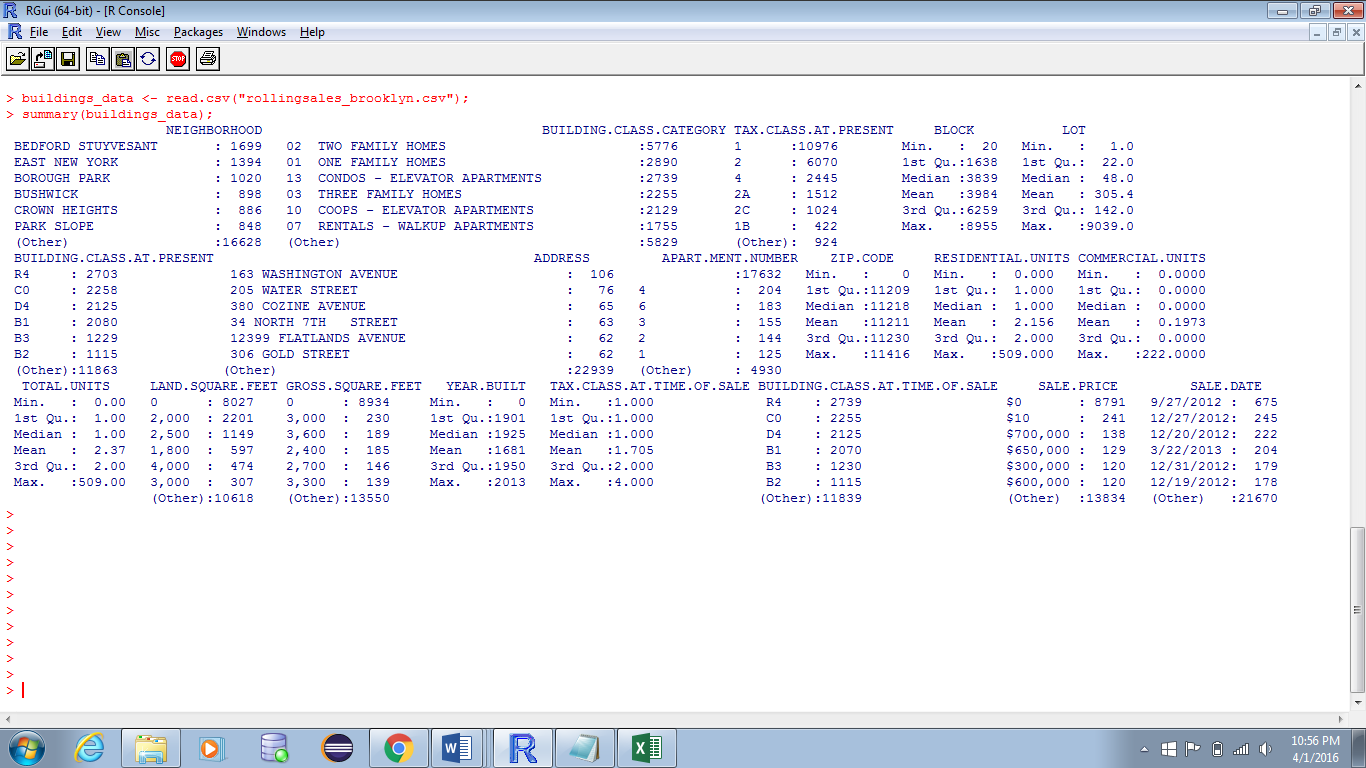
**Numerical Attributes:** A numeric attribute is quantitative; that is, it is a quantifiable amount, represented in number or genuine qualities. Numeric qualities can be interim scaled or proportion scaled.

The numeric attributes which are present in the dataset considered are

Block, Lot, Zipcode, Year Built, Residential Units, Commercial Units, Tax Class at Time of Sale, Total Units.



The summary of the entire data set before performing any activity.



From the above statistics we can find that we have some outliers and noisy data in Building Class category, tax class at present, building class at present, zip-code and year built attributes. These missing values are specified as “NA”. Also we are interested about the sales of buildings which were built after 1950 but the data set contains details of buildings built before 1950 which can be seen in the summary for year built attribute. Here, the data is correct but it is which is not required for us. These can be labelled as outliers.

Keeping in mind the end goal to make the information immaculate, we have to perform Data Cleaning, which is a conspicuous usefulness of information mining.

1. **Data Cleaning and Outlier Detection:**

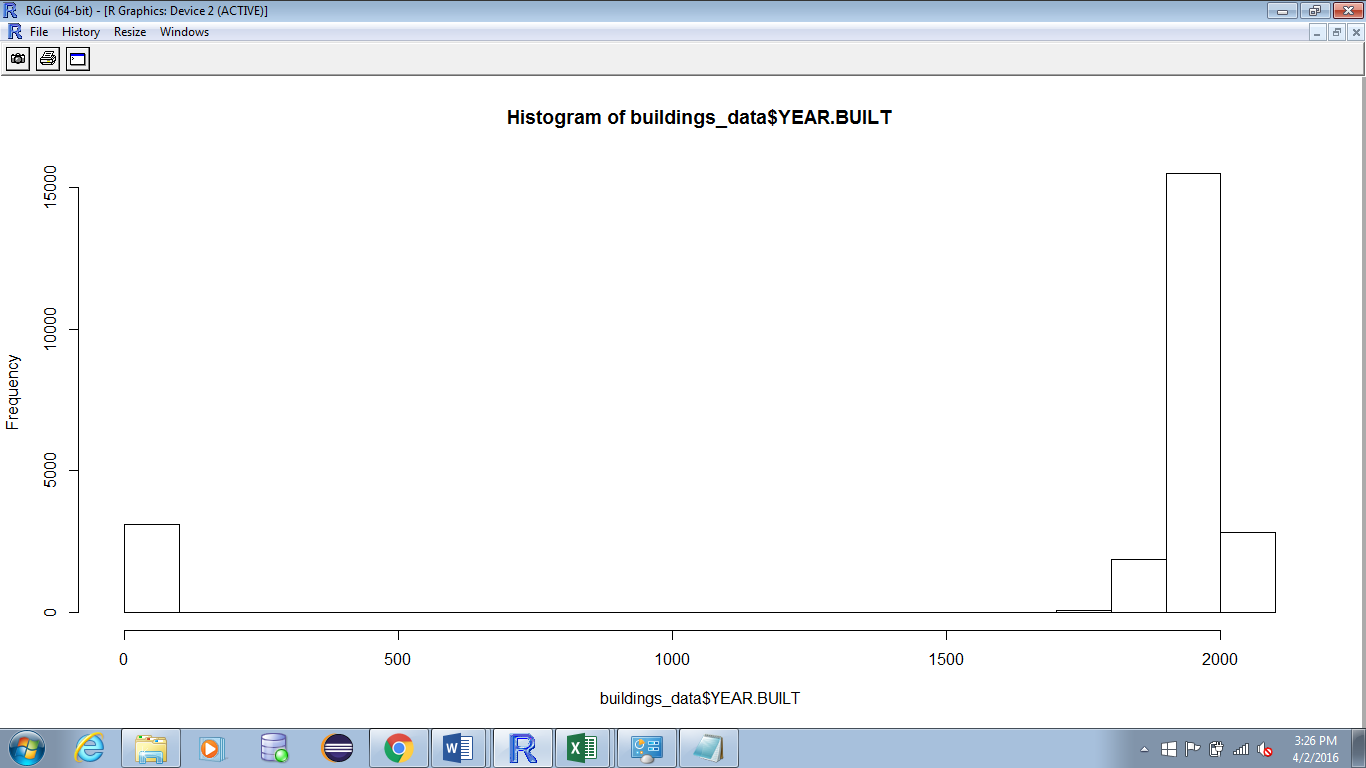
* **Outlier:**

Outliers are the information objects which go astray from the typical information objects. These data items are intriguing however they damage the techniques which produce ordinary information. Outliers ought to be researched deliberately. Regularly they contain important data about the procedure under scrutiny or the information assembling and recording process. Before considering the conceivable disposal of these focuses from the information, one ought to attempt to comprehend why they showed up and whether it is likely comparative qualities will keep on appearing. Obviously, exceptions are often terrible data values.

In our data set, we have many data objects for Year\_Built attribute which are less than 1950. Here, using R studio we are removing the outliers.

Below are the screen shots for the same.

**Year data before outlier removal**



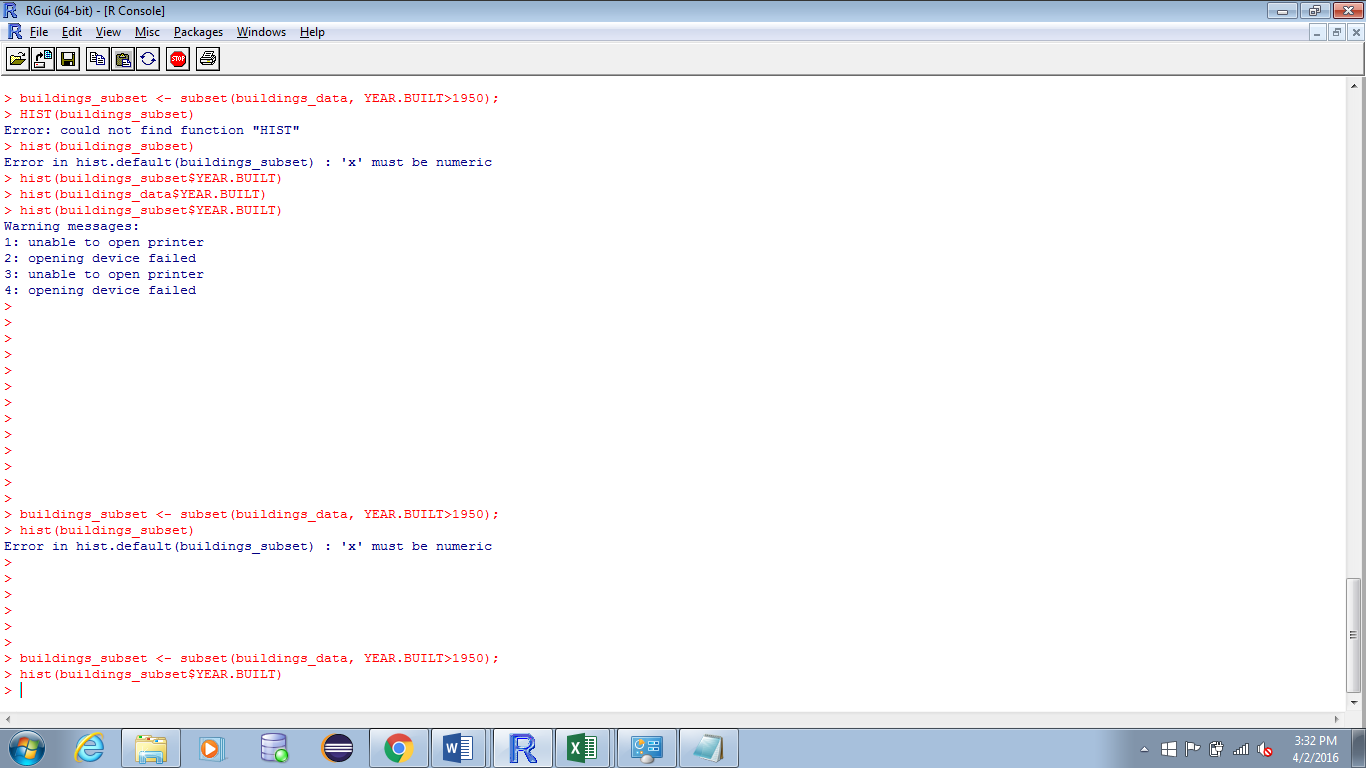
Reference pdf:

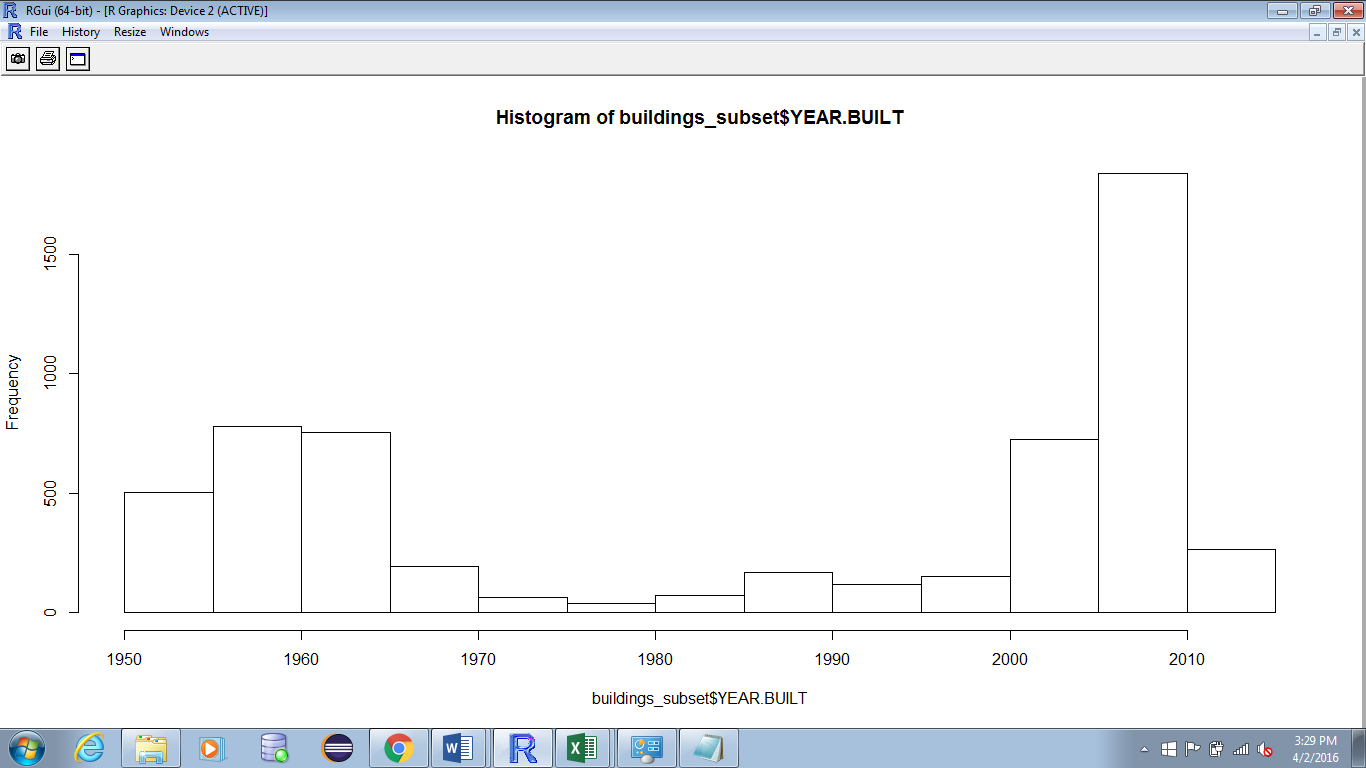


The above graph shows the frequency of buildings built verses the year of built before performing any data cleaning activity.

**Year data after outlier removal**

Using R studio the data points which are considered as outliers can be eliminated. The following R commands will help in removal of outliers and to generate the graph for that particular attribute.





Reference pdf:



The above graph shows the frequency of buildings built verses the year of built (after 1950) after performing any data cleaning activity.

* **Data Cleaning:**

Data is cleaned to check for the missing values, irrelevant values and other noisy data.

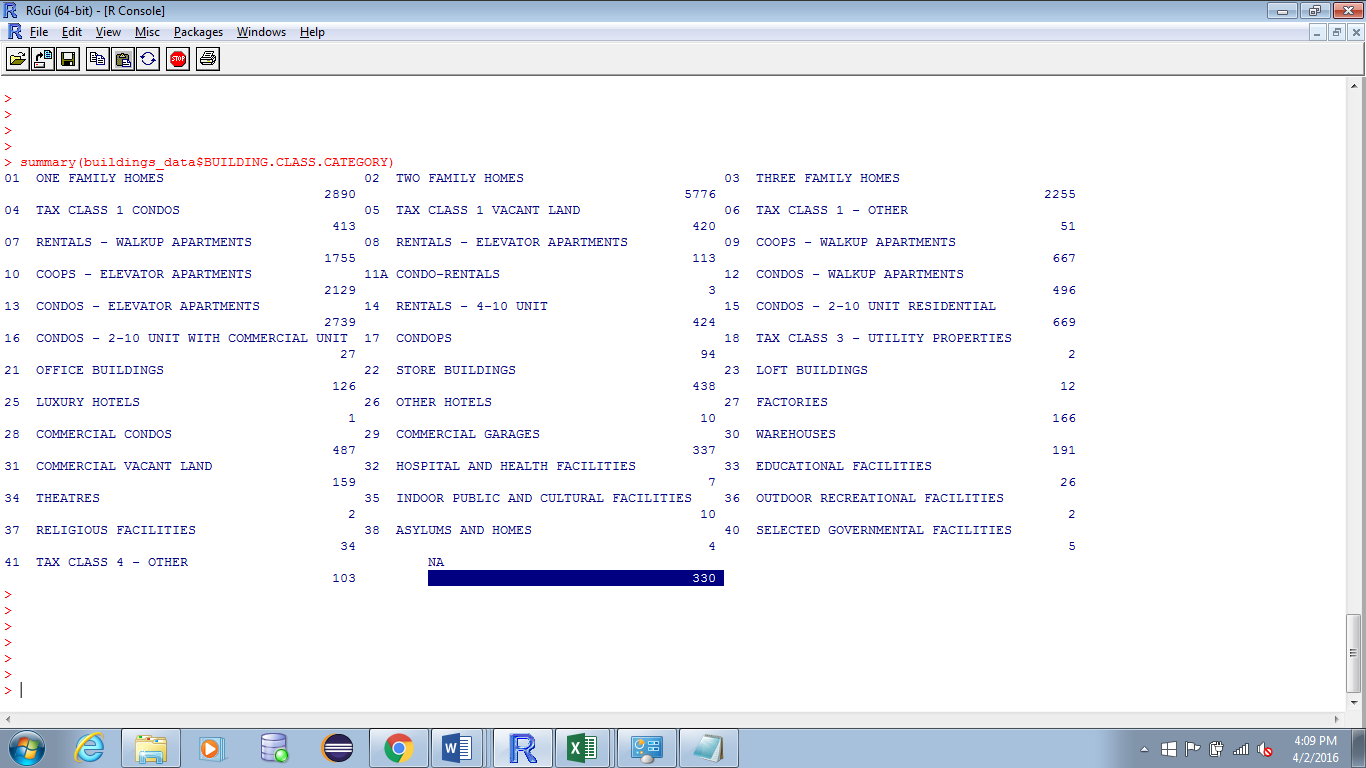
Data cleaning schedules work to "clean" the information by filling in missing qualities, smoothing boisterous information, recognizing or expelling outliers, and determining irregularities. Besides, filthy data can bring about disarray for the mining system, bringing about problematic yield. Information cleaning is normally executed as an iterative two-stage process comprising of discrepancy identification and data transformaton. Therefore, a valuable preprocessing step is to run your information through a few data cleaning routines.

Tool: R studio.

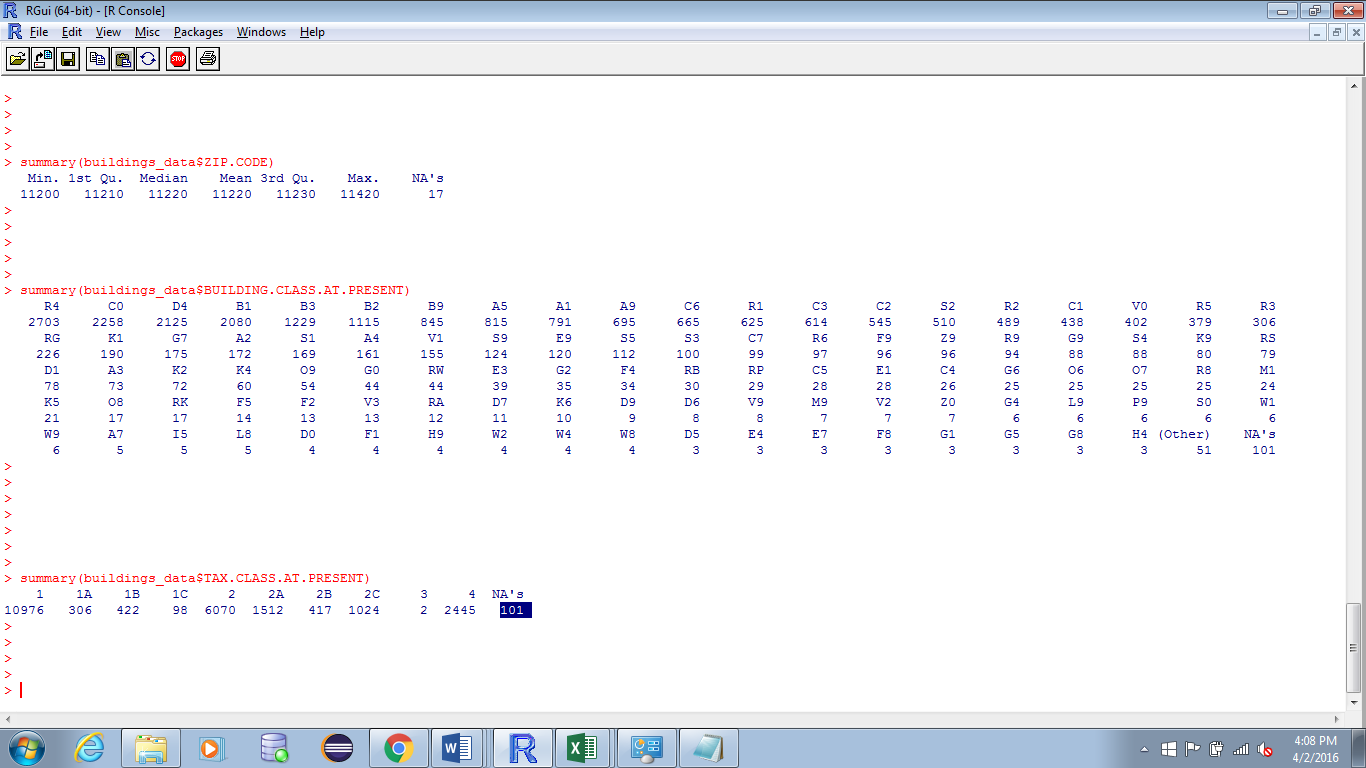
In our data set there are many data objects with missing values i.e., object which have a value of NA. Using R studio, we are removing such kind of noisy data. Below are the screen shots for the same.

Summary of few data set attributes before cleaning (Rows highlighted are those with noisy data).

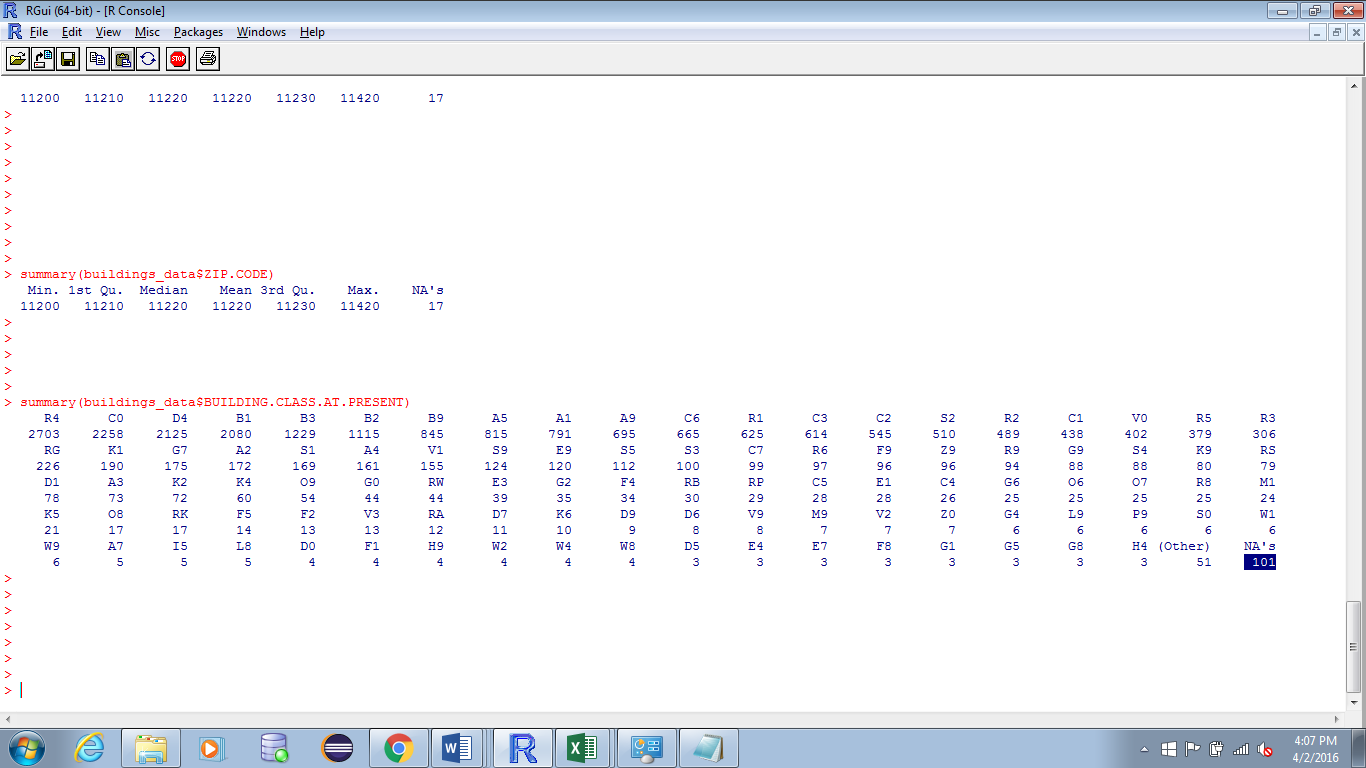
BUILDING CLASS CATEGORY (before cleaning)



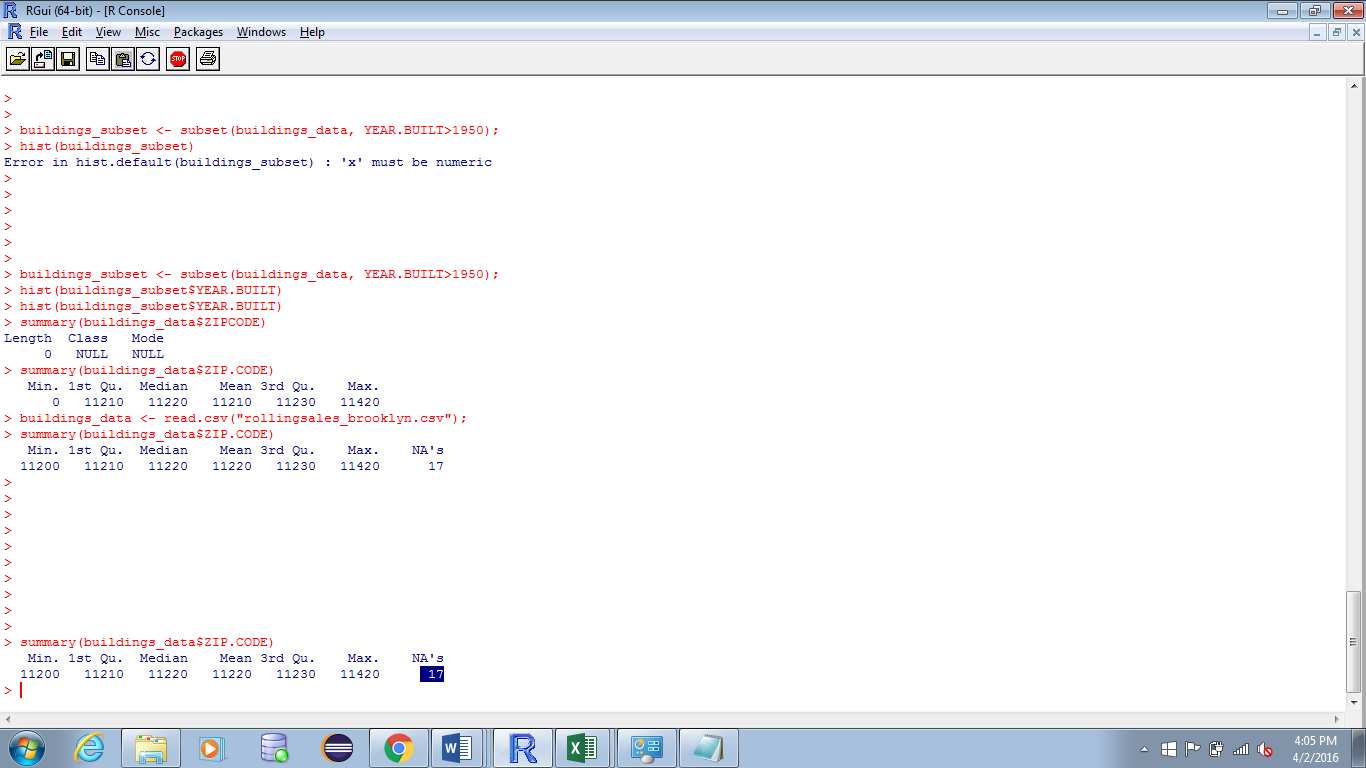
TAX CLASS AT PRESENT (before cleaning)



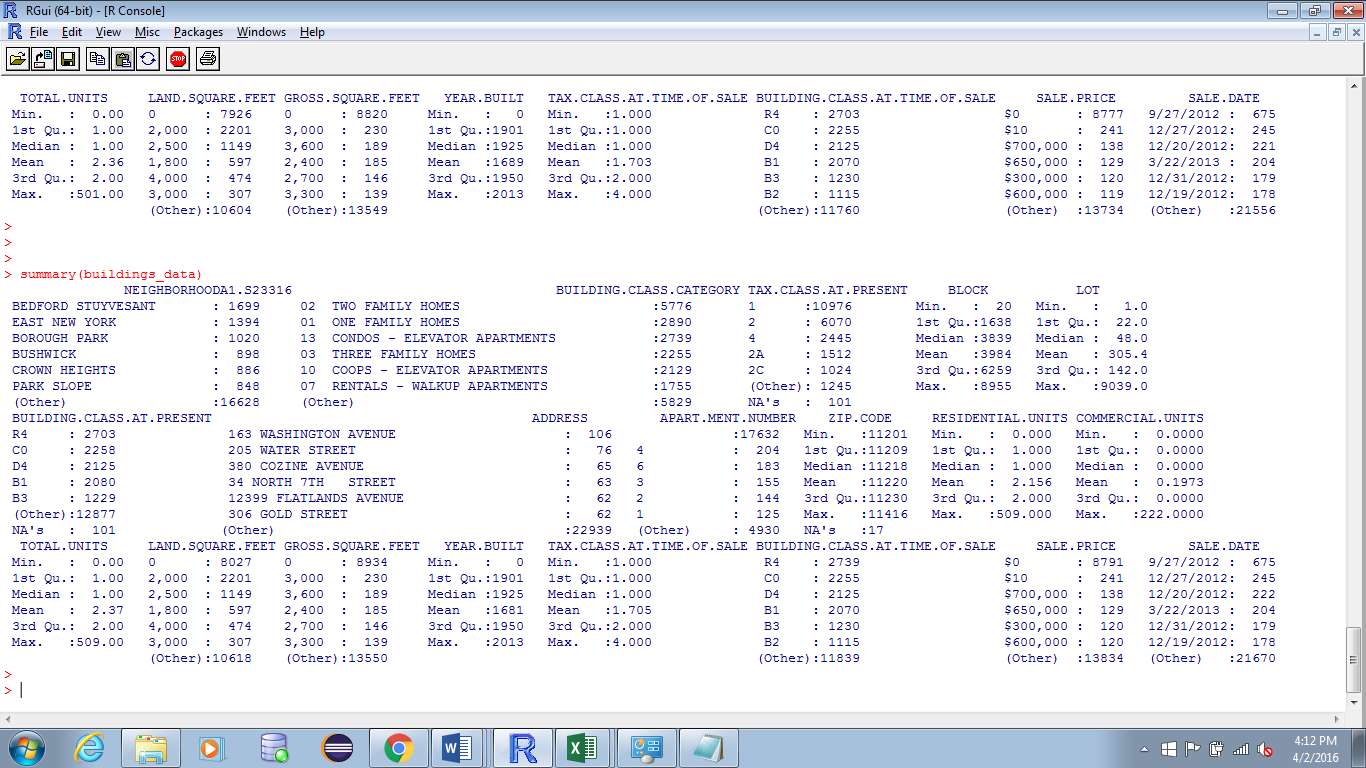
BUILDING CLASS AT PRESENT (before cleaning)



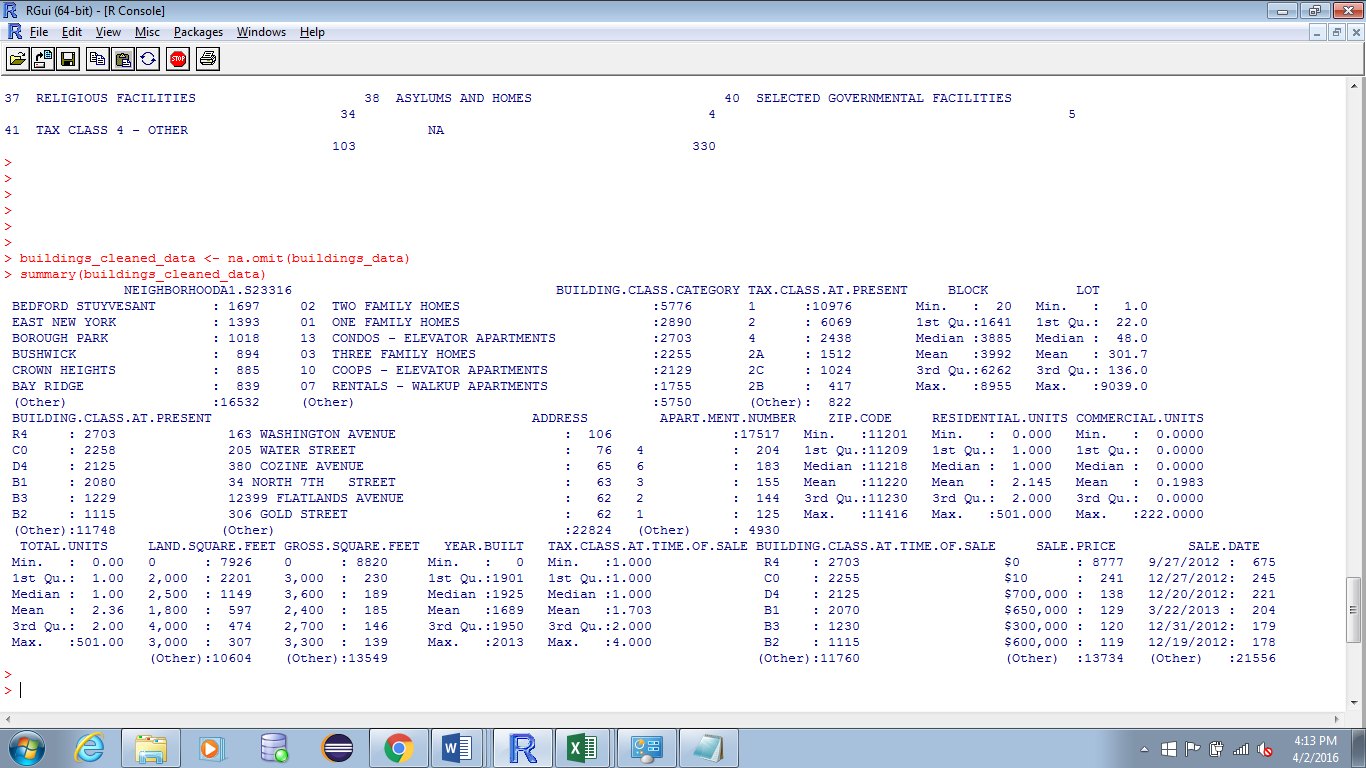
ZIPCODE (before cleaning)



Summary of the entire data set before applying any data cleaning operation.



Data Set after Cleaning. (You can observe that there are no missing values/ noisy data i.e., no NA’s are observed)



1. **Data Classification**

Classification of a data item to a specific class contingent upon the past samples of different articles is called data grouping. The data objects for which class must be appointed or anticipated is called target set and the past information objects set that are utilized to group new data articles is called training set.

* **K Nearest Neighbours (k n n):**

K nearest neighbour is a algorithm utilized for data arrangement and expectation. This calculation takes a training set and gauges the class of target set utilizing distance functions, for example, Euclidean separation, Manhattan separation and so forth. In k-NN classification, the result is a class enrolment. An item is characterized by a larger part vote of its neighbour, with the article being doled out to the class most regular among its k closest neighbour. The estimation of K is appointed by assessing the information. Suppose k = 1, then the item is just allocated to the class of that solitary closest neighbour. For any information set it is ideal to pick the k quality from 3 to 10 for good results. For our dataset the appropriate value of K is 8.

Tool: R studio

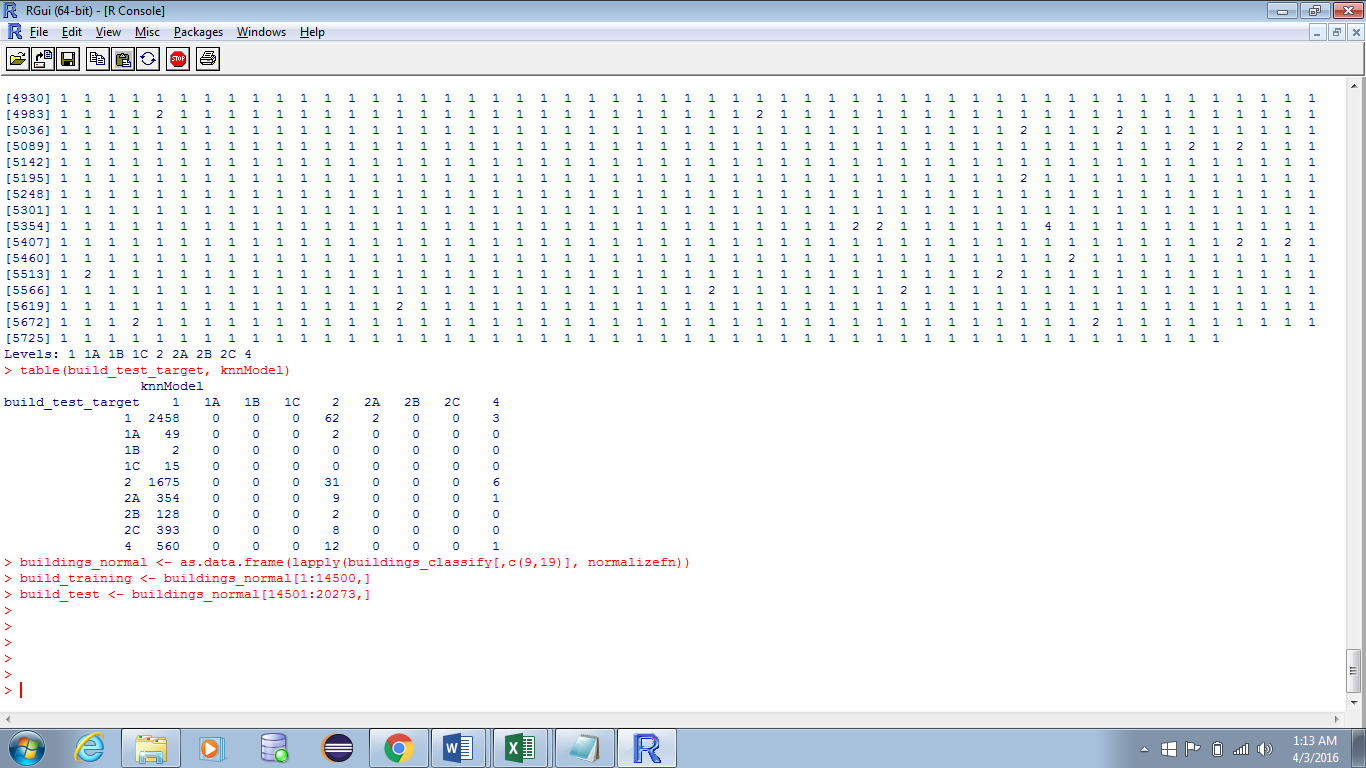
First of all we have to define a normalizing function that calculates the distance among the attributes.

**Loading the data set and generating a normalizing function.**



Then the next step is to divide the data set into 2 parts i.e., training set and test set. Normally the training set will contain 60% of the data set and the remaining records are included in test set.

**Setting the training data and test data.**



Then the final step is to load the class and calling the KNN function with the appropriate arguments along with the value of K.

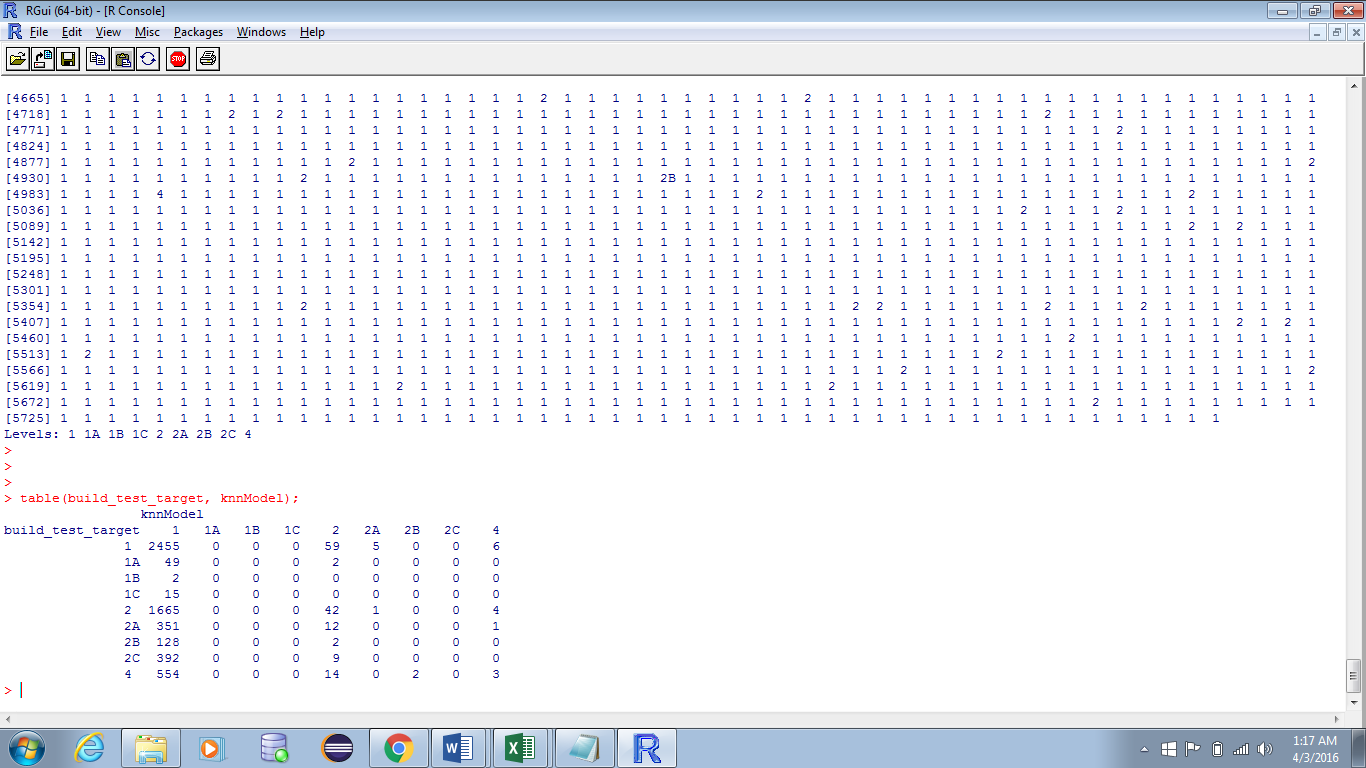
**Loading the class and calling the knn function with required arguments. The value of k is set to 8.**



**Output of the classifier.**



**Predictions of knn Algorithm on our data set.**



**R code for Knn algorithm:**

> set.seed(11090);

> gp <- runif(20273);

> buildings\_classify <- buildings\_data[order(gp),];

> head(buildings\_classify);

> normalizefn <- function(y){

+ return((y - min(y))/(max(y)-min(y)) ) }

> buildings\_normal <- as.data.frame(lapply(buildings \_classify[,c(6,7)], normalizefn));

> str(buildings \_normal);

> build\_training <- buildings \_normal[1:10000,];

> build\_test <- buildings \_normal[10001:16654,];

> build\_training\_target <- buildings \_data[1:10000, 3];

> build\_test\_target <- buildings \_data[10001:16654, 3];

> require(class);

> knnModel <- knn(train = build\_training, test = build\_test, cl = build\_training\_target, k = 8);

> knnModel;

> table(build\_test\_target, knnModel);

* **Naïve Bayes Algorithm:**

The Naive Bayesian classifier depends on Bayes' hypothesis with autonomy assumptions between indicators. A Naive Bayesian model can be easily constructed, with no muddled iterative parameter estimation which makes it especially helpful for huge datasets. Despite its straightforwardness, the Naive Bayesian classifier regularly does shockingly well and is generally utilized because it frequently outperforms more sophisticated classification methods.

Naïve Bayes algorithm takes the training data and finds its posteriori probability using the below formula,



where,

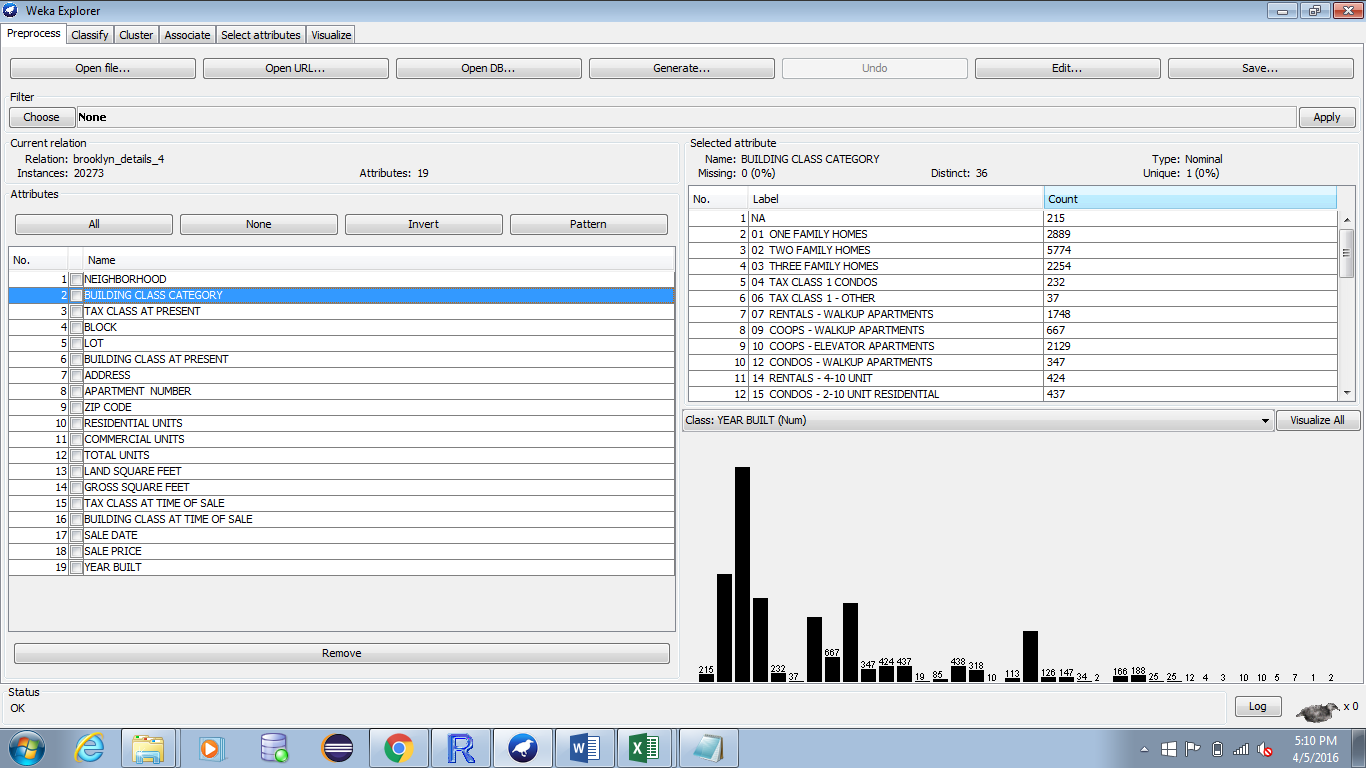
P(X|H) 🡪 Likelihood

P(H) 🡪 Class prior probability

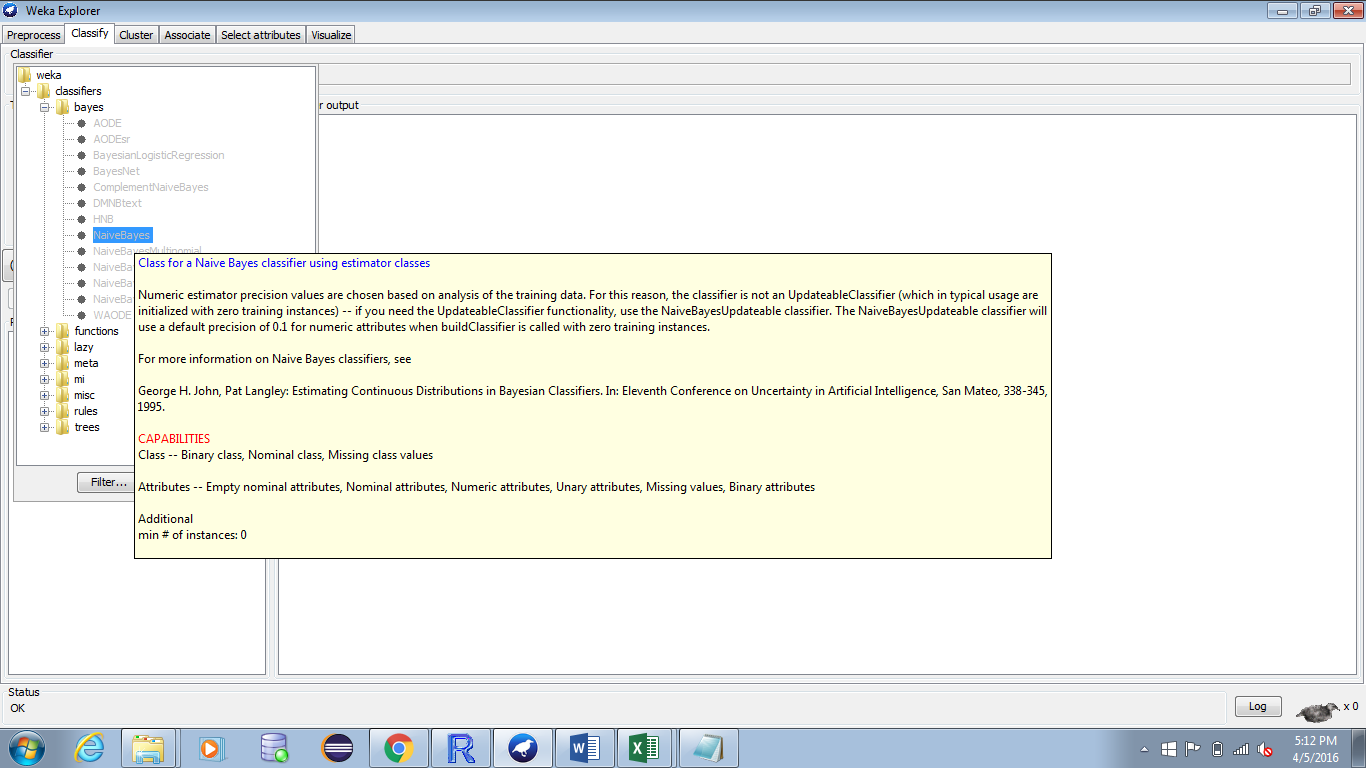
P(X) 🡪 Predictor prior probability

Tool: Weka

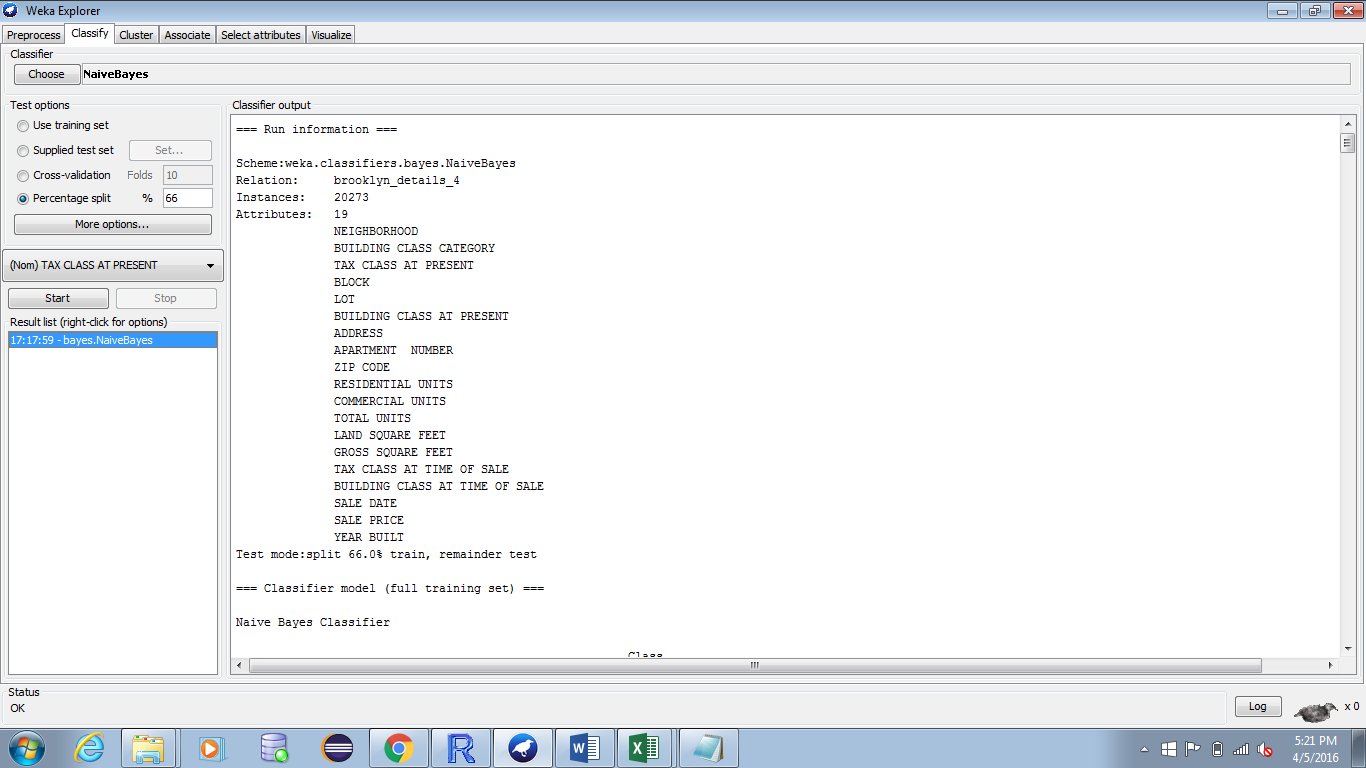
**Load the data set into Weka tool**

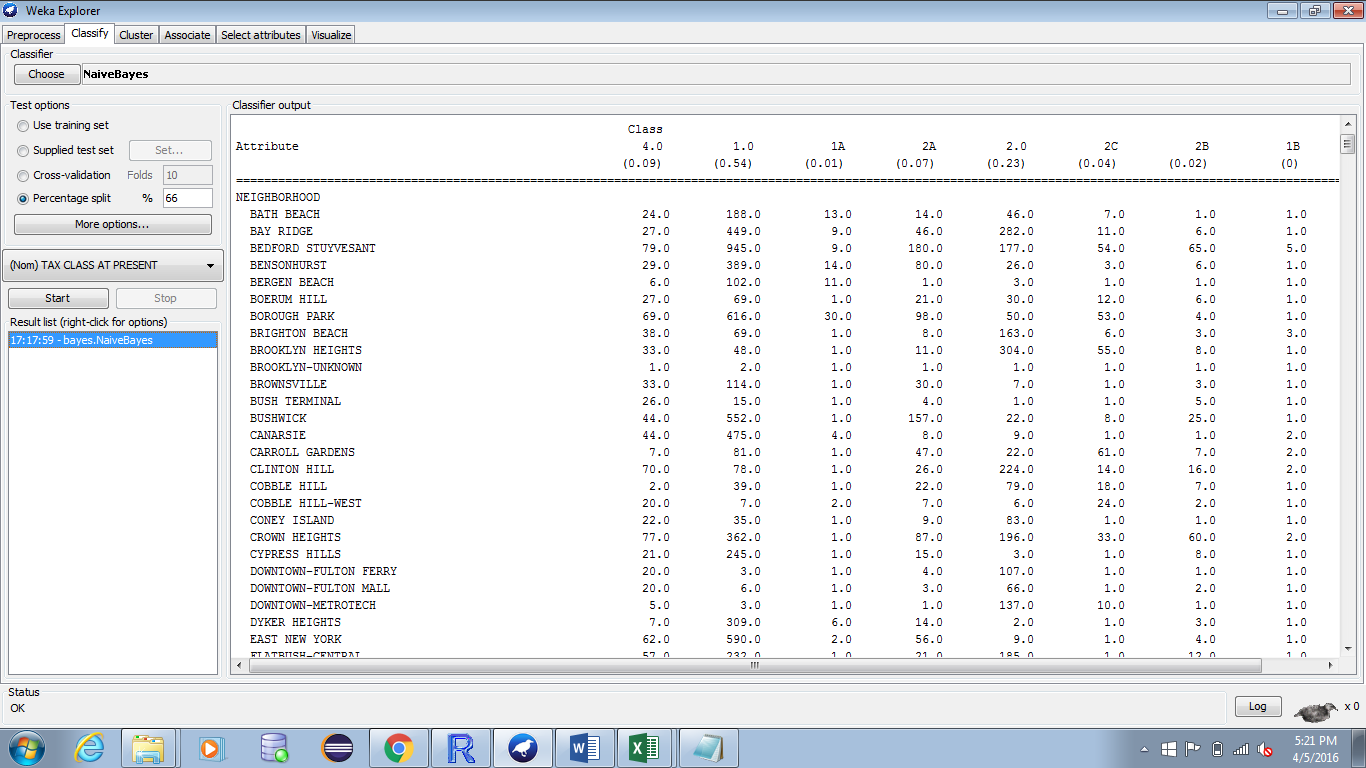


Choose the **naïve bayes classifier** from the classify tab. Set the test options to Percentage split i.e., it takes 66% data as training data and remaining data as test data.

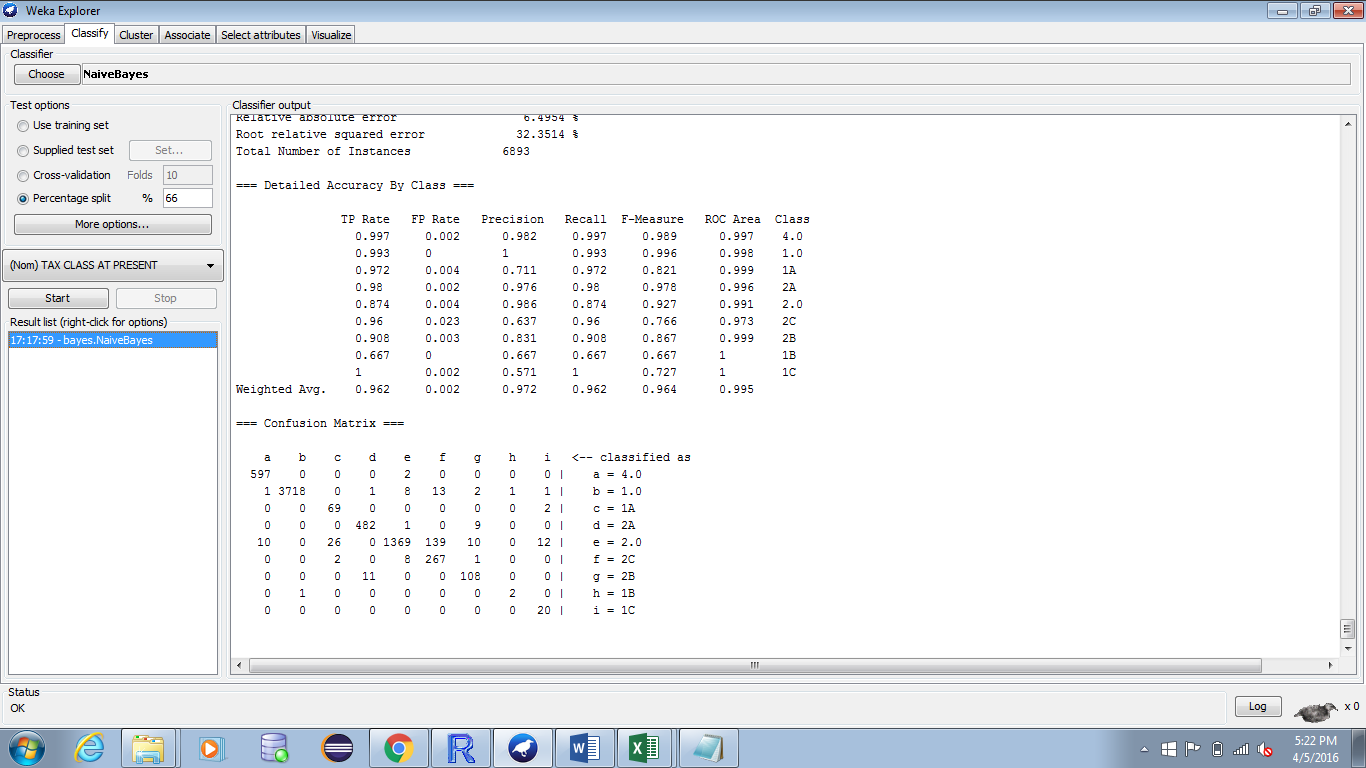


Output of the naïve bayes classification based on the **TAX CLASS AT PRESENT** attribute.





**Confusion Matrix (Classification based on Tax Class at present)**

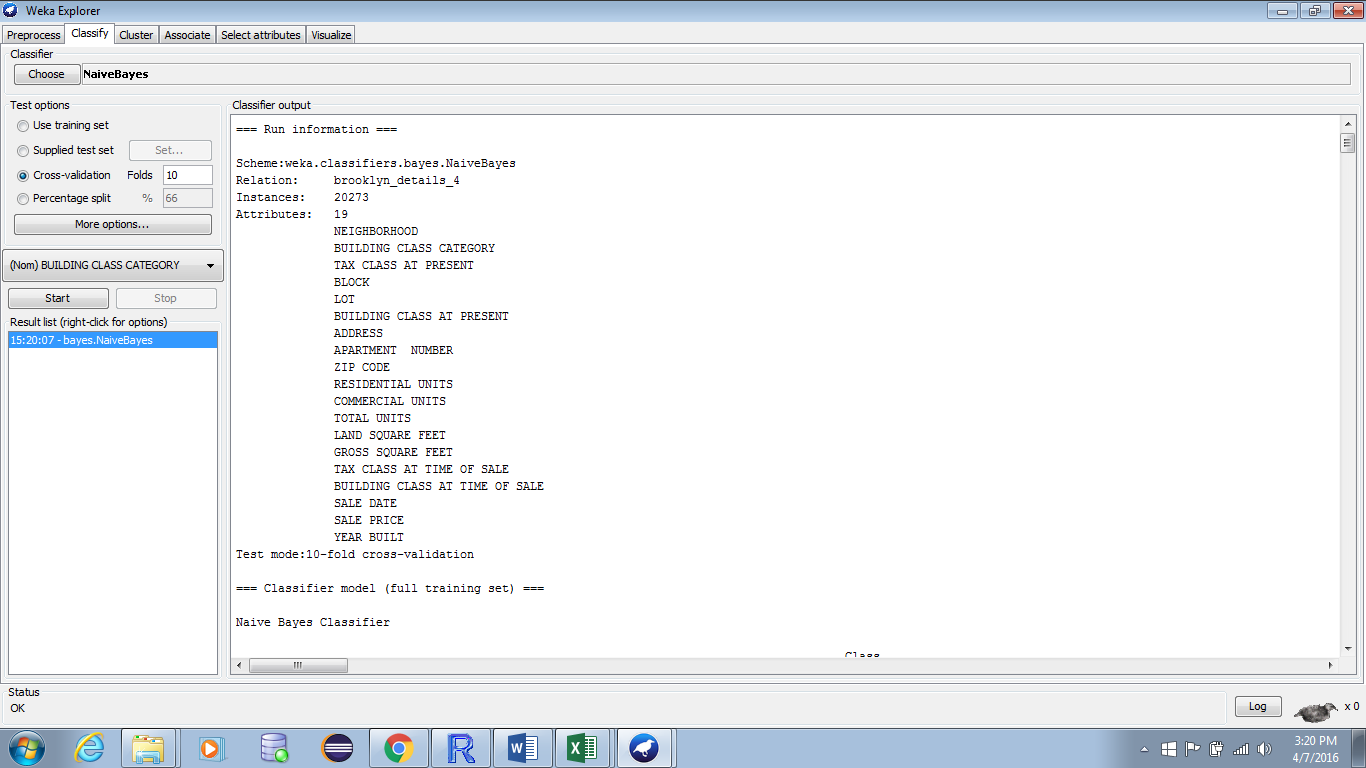


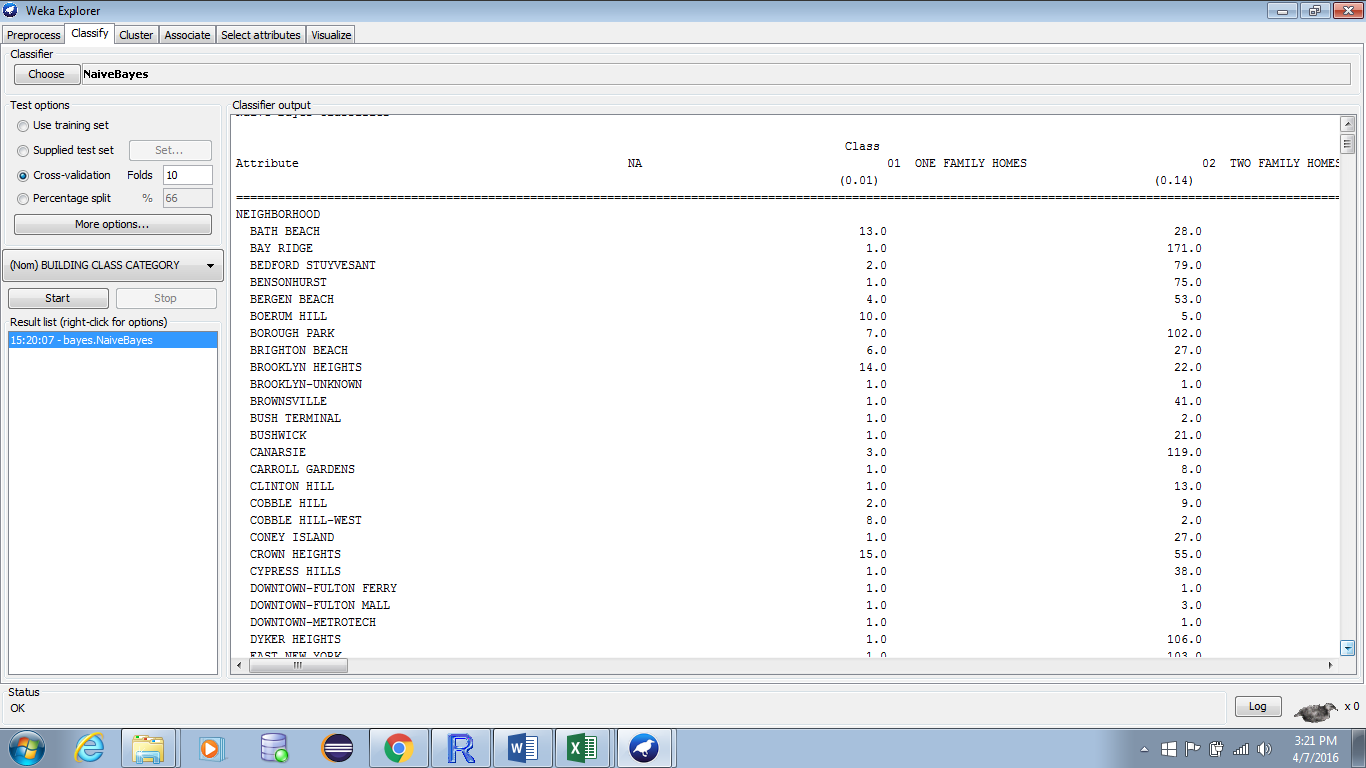
Reference Naïve Bayes output file:

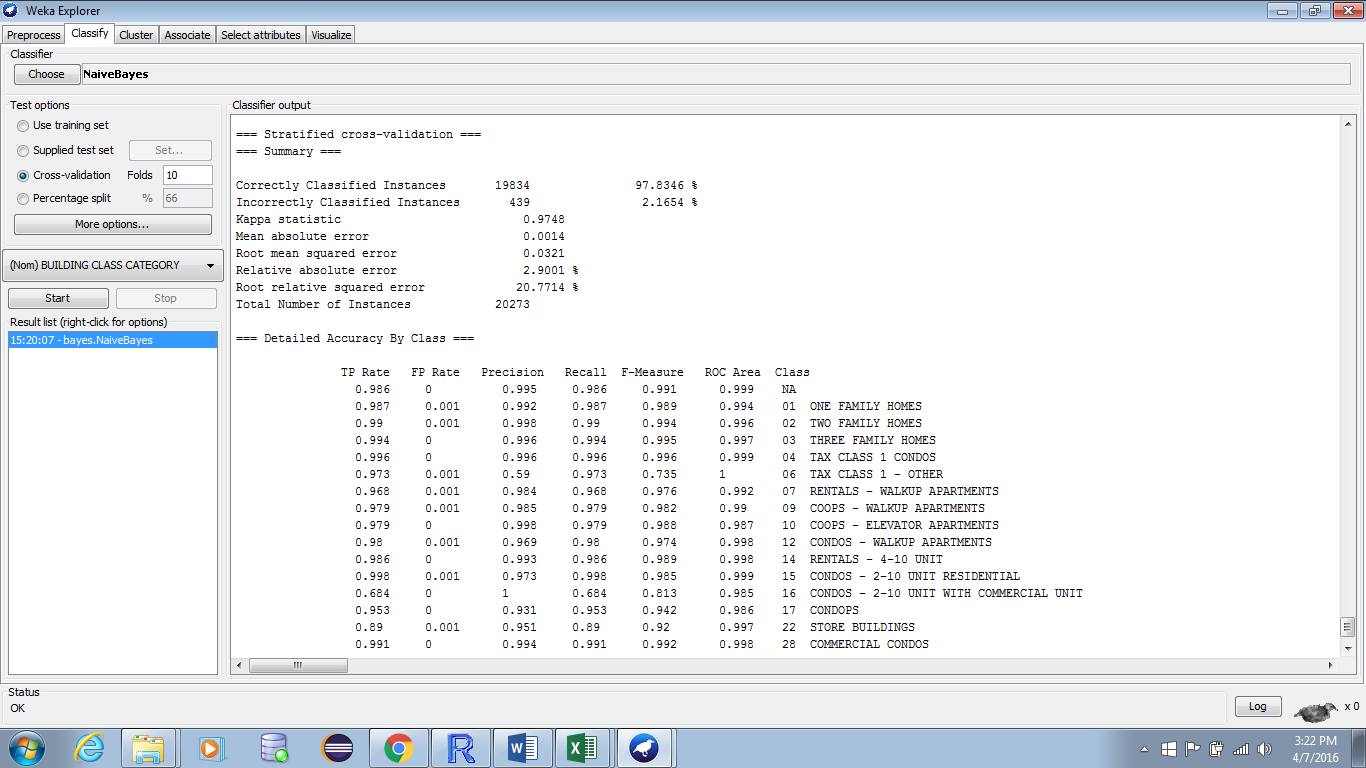


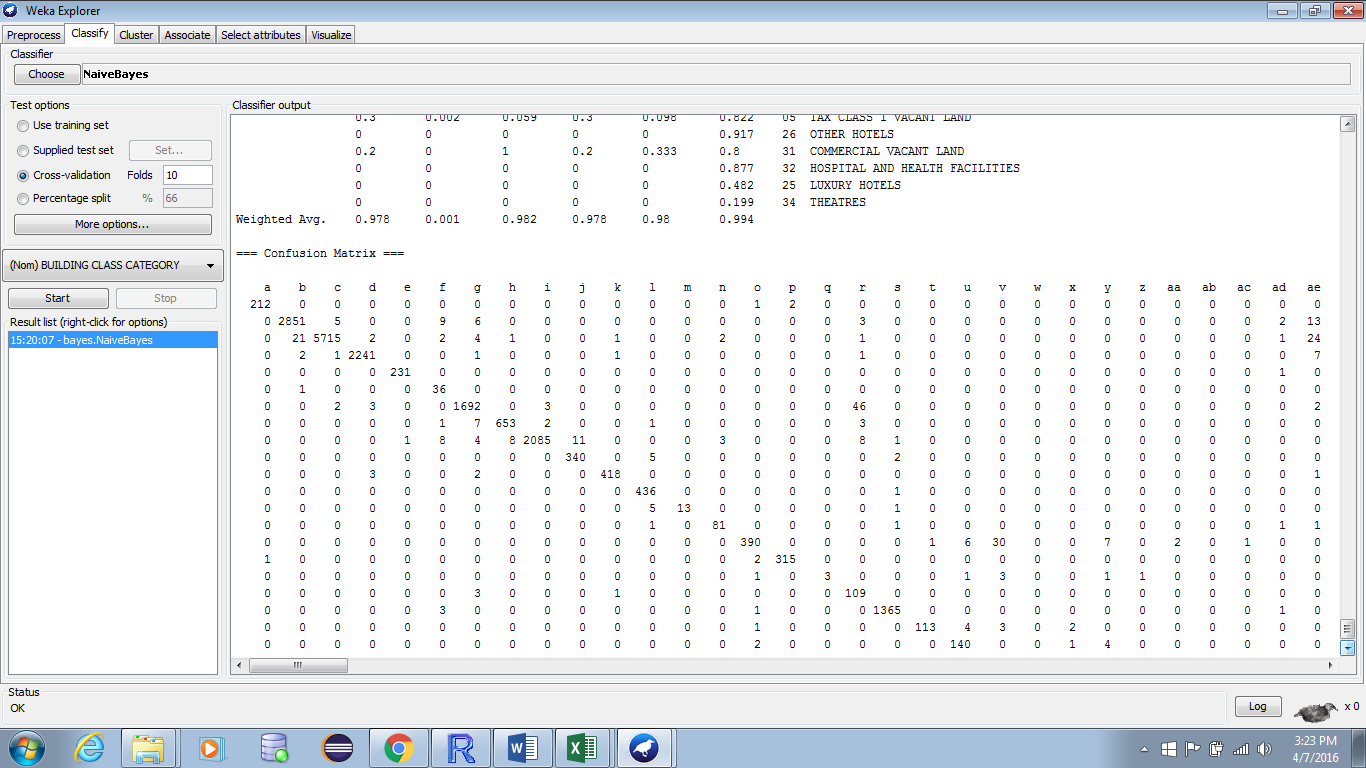
Tool: Weka

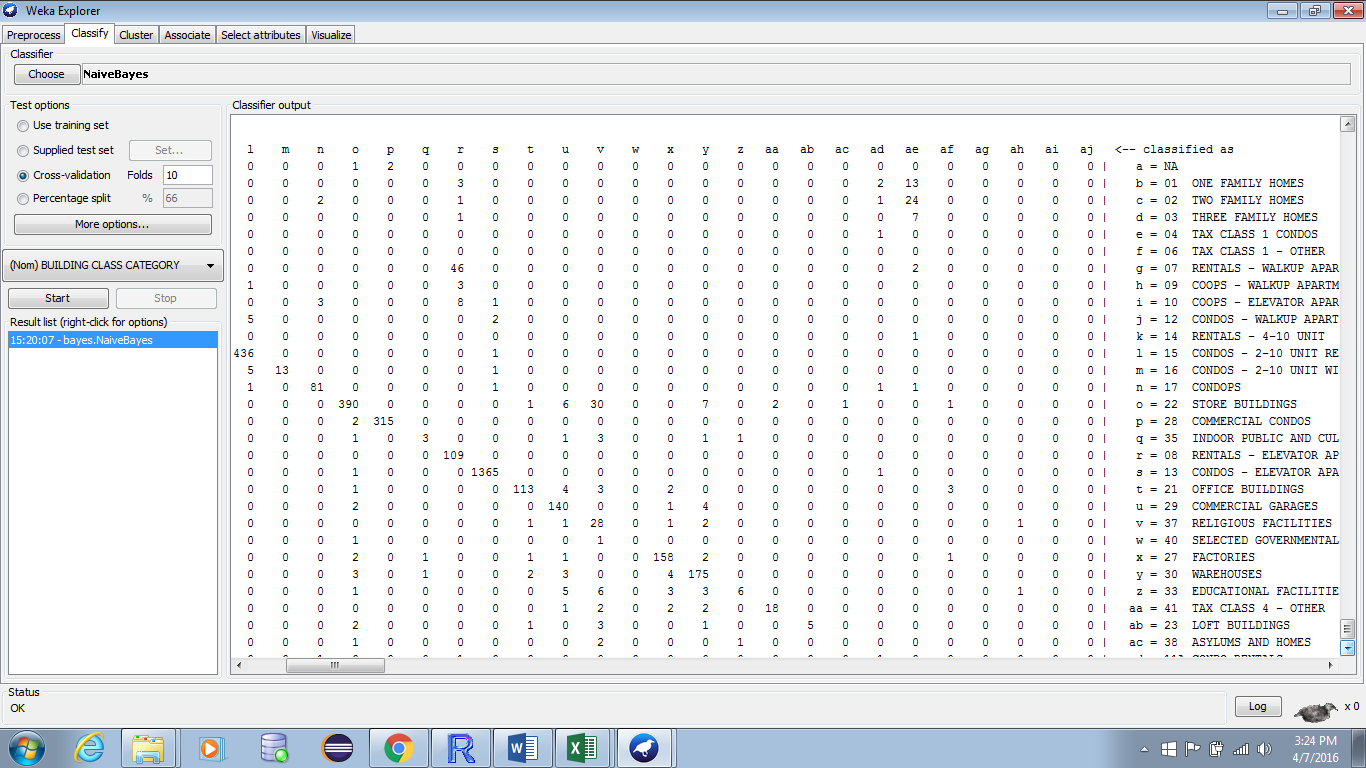
Load the data and choose the classifier in the same way which as mentioned for Naïve Bayes. **Here the classification is done based on building class category.**











Reference Output File:

****

1. **Clustering**

Clustering is the procedure of making a gathering of unique items into classes of comparative articles. Clustering is unsupervised learning in which similar data objects are grouped together. A cluster of data objects can be treated as one group. While doing cluster examination, we first parcel the arrangement of information into gatherings based on information similitude and after that assign the labels to the gatherings. The primary point of preference of clustering over classification is that, it is versatile to changes and singles out valuable components that recognize diverse gatherings. The similarity between the objects of different clusters should be as low as possible.

The following points will give an idea on why clustering is required in data mining −

* Adaptability − We require exceedingly versatile grouping calculations to manage vast databases.
* Capacity to manage various types of characteristics − Algorithms ought to be competent to be connected on any sort of information, for example, interim based (numerical) information, straight out, and binary data.
* Discovery of groups with trait shape − The clustering calculation ought to be equipped for identifying groups of subjective shape. They should not be limited to just separation measures that tend to discover round cluster of little sizes.
* High dimensionality − The clustering algorithm should not just have the capacity to handle low-dimensional information additionally the high dimensional space.
* Capacity to manage noisy information − Databases contain uproarious, lost or incorrect information. A few calculations are delicate to such information and may prompt low quality clusters.
* Interpretability − The clustering results ought to be interpretable, conceivable, and usable.
* **K means cluster:**

K-Means clustering expects to segment n objects into k groups in which every item has a place with the cluster with the closest mean. This technique creates precisely k diverse groups of most noteworthy conceivable qualification. The best number of clusters k prompting the best detachment (separation) is not known as from the earlier and must be figured from the data. The target of K-Means clustering is to minimize absolute intra-group difference, or, the squared mistake capacity.

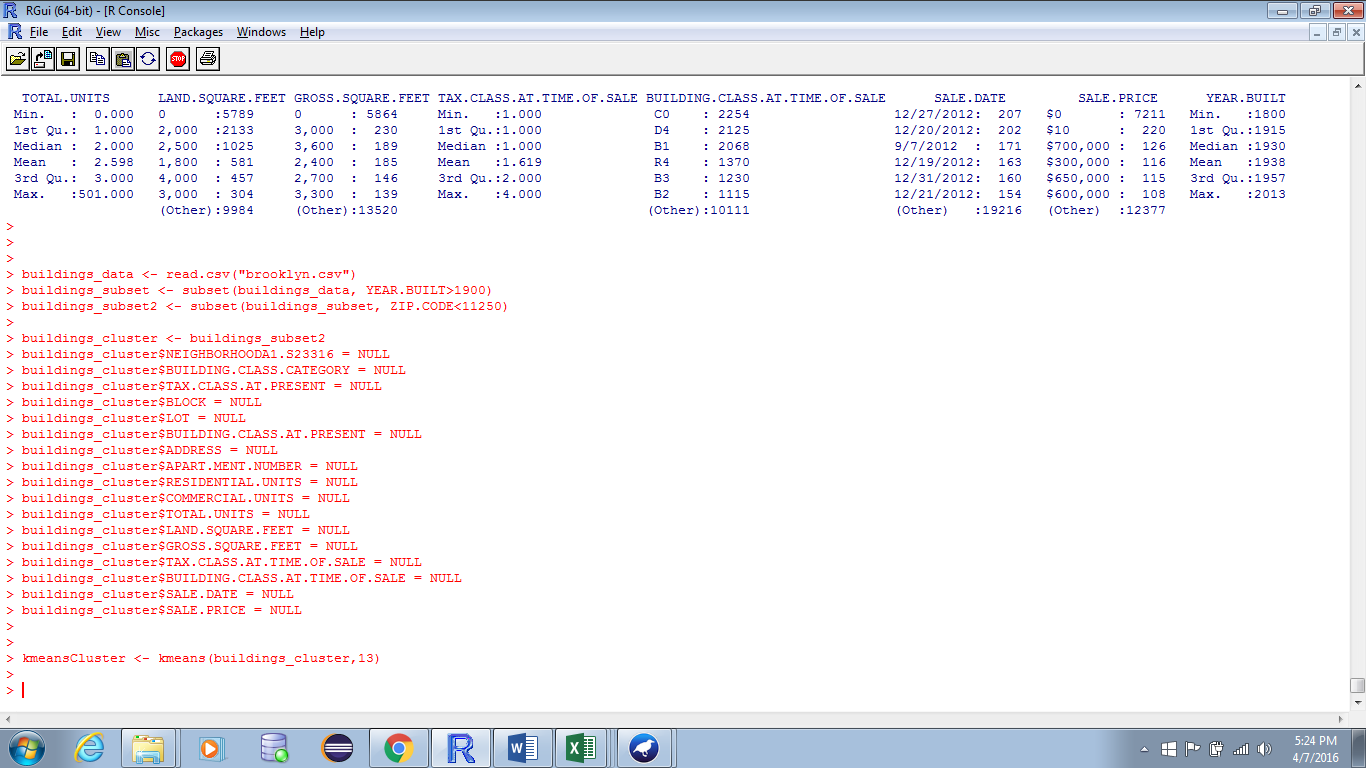
K-Means is generally a proficient technique. However, we have to determine the quantity of clusters, ahead of time and the final results are delicate to initialize and regularly ends at a nearby ideal. Unfortunately there is no worldwide hypothetical technique to locate the ideal number of groups. A handy methodology is to contrast the results of numerous runs and distinctive k and pick the best one in view of a predefined standard. In general, a large value of k will decrease the error but expands the danger of overfitting

The condition for this algorithm is that the value of k should be fixed apriori. In this algorithm k centroids are defined one for each cluster. This algorithm is a NP- hard problem.

Tool: R studio.

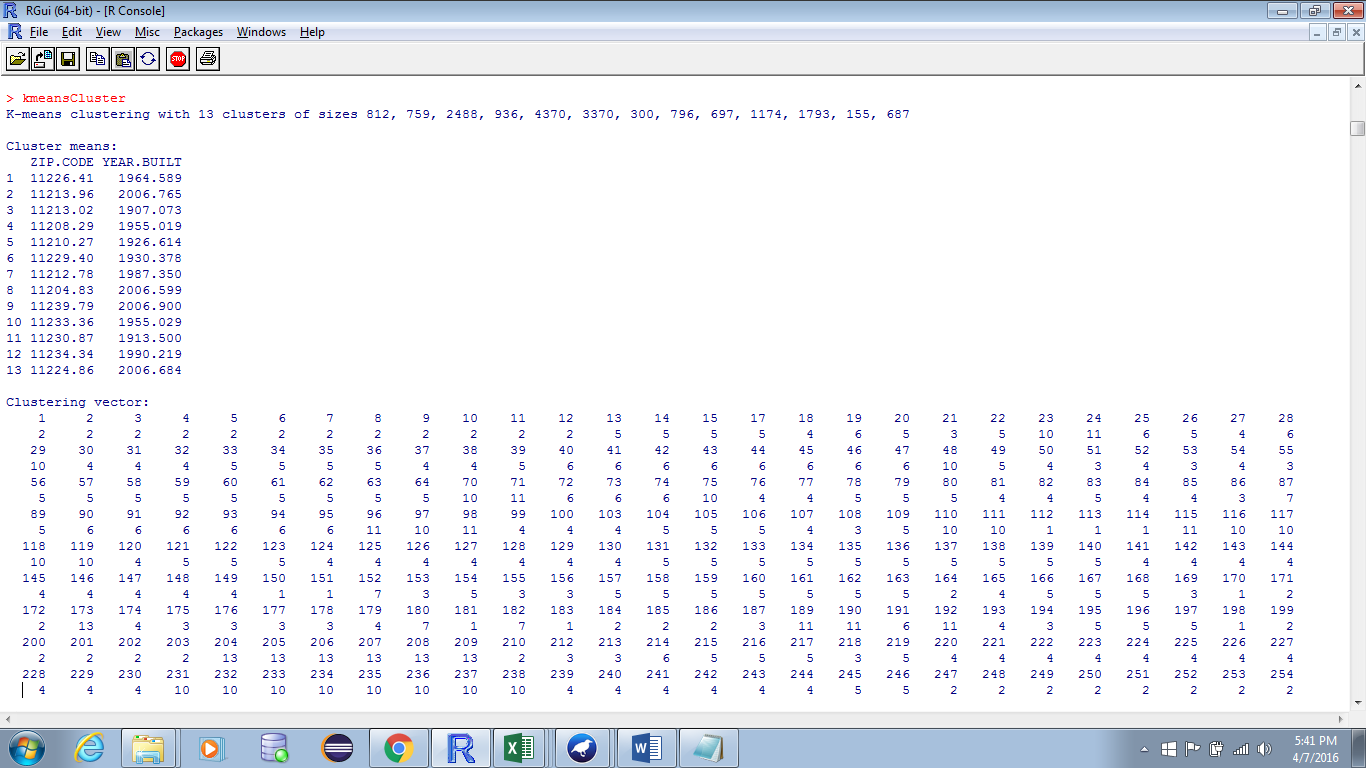
**Calling kmeans function on data set.**

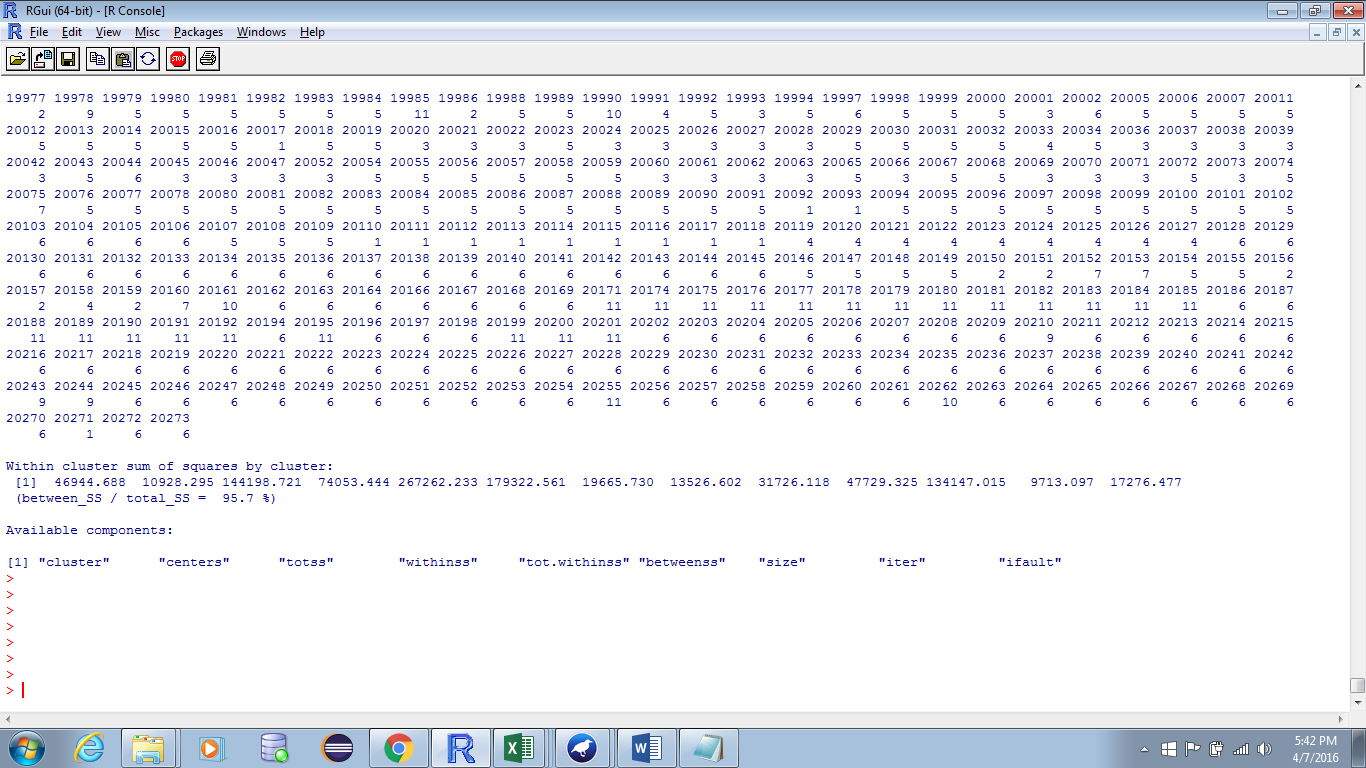
After loading the dataset in R, all the attributes which are not required for clustering are defined as null and the rest of the attributes are fed as input to the kmeans function along with the value of K.



**Cluster summary**

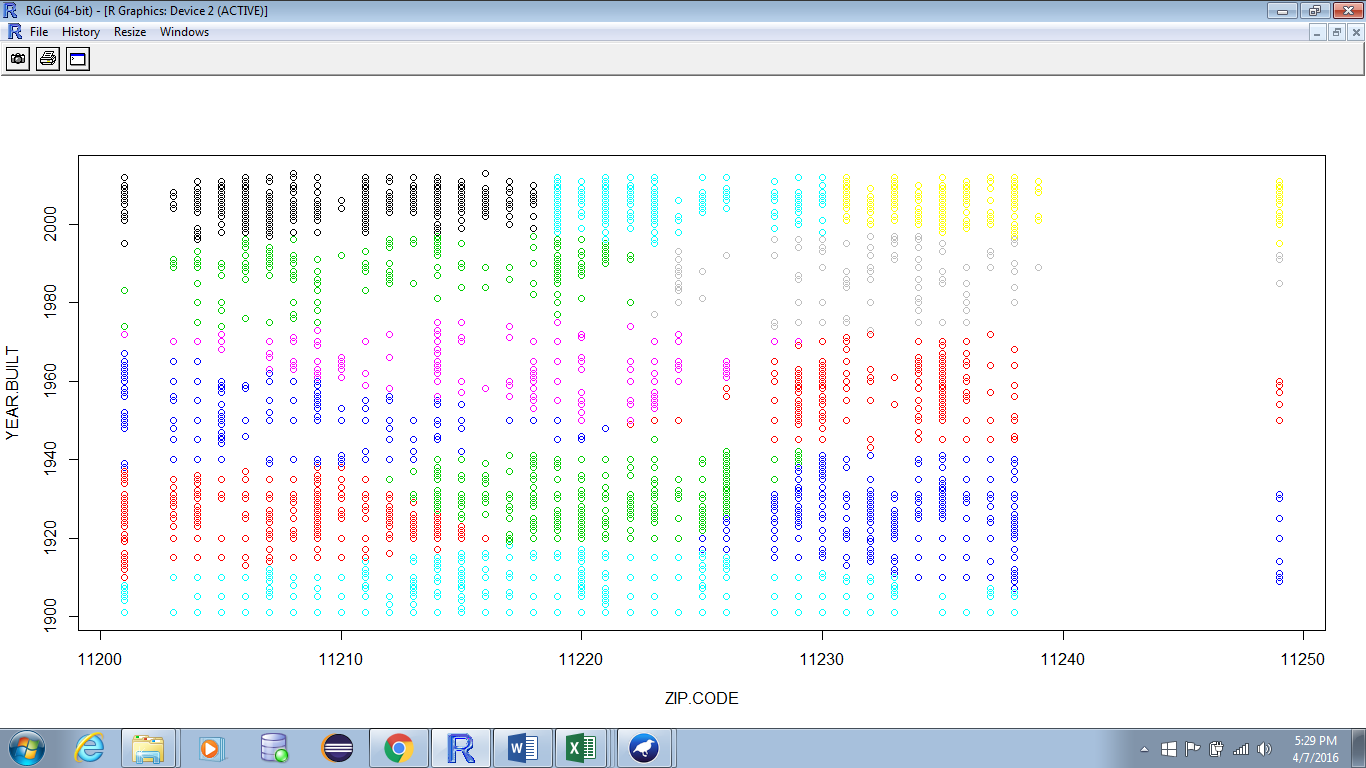
Kmeans cluster summary for the dataset based on the year built and the zipcode.





**Cluster Plot**

The following graph is the cluster plot for the kmeans function generated above with X- axis representing the Zipcode and the Y-axis representing the Year Built of the building.



Reference cluster pdf:



**R code for K Means algorithm:**

> buildings\_data <- read.csv("brooklyn.csv")

> buildings\_subset <- subset(buildings\_data, YEAR.BUILT>1900)

> buildings\_subset2 <- subset(buildings\_subset, ZIP.CODE<11250)

>

> building\_cluster <- buildings\_subset2

> building\_cluster$NEIGHBORHOODA1.S23316 = NULL

> building\_cluster$BUILDING.CLASS.CATEGORY = NULL

> building\_cluster$TAX.CLASS.AT.PRESENT = NULL

> building\_cluster$BLOCK = NULL

> building\_cluster$LOT = NULL

> building\_cluster$BUILDING.CLASS.AT.PRESENT = NULL

> building\_cluster$ADDRESS = NULL

> building\_cluster$APART.MENT.NUMBER = NULL

> building\_cluster$RESIDENTIAL.UNITS = NULL

> building\_cluster$COMMERCIAL.UNITS = NULL

> building\_cluster$TOTAL.UNITS = NULL

> building\_cluster$LAND.SQUARE.FEET = NULL

> building\_cluster$GROSS.SQUARE.FEET = NULL

> building\_cluster$TAX.CLASS.AT.TIME.OF.SALE = NULL

> building\_cluster$SALE.DATE = NULL

> building\_cluster$SALE.PRICE = NULL

>building\_cluster$BUILDING.CLASS.AT.TIME.OF.SALE = NULL

> summary(building\_cluster)

> kmeansCluster <- kmeans(building\_cluster,13)

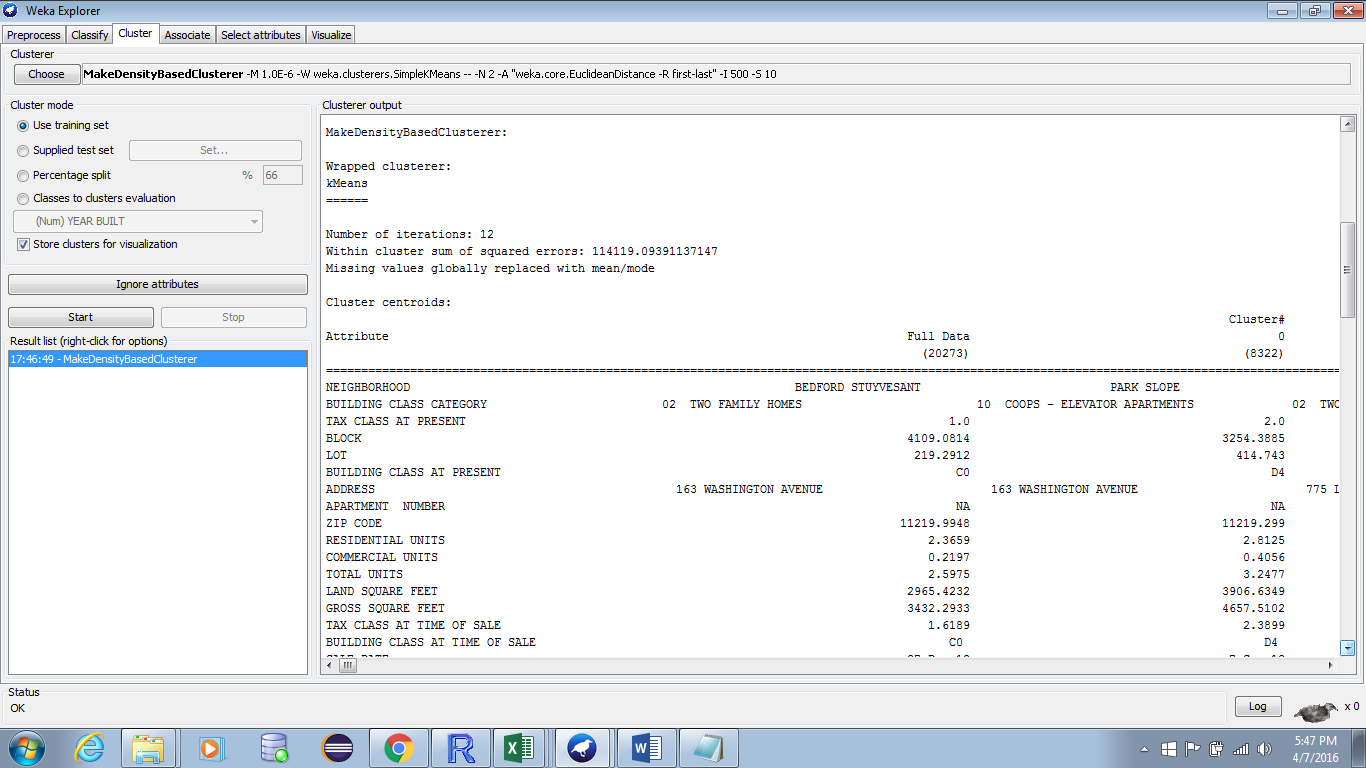
> kmeansCluster

> plot(building\_cluster, col= kmeansCluster$cluster);

* **Density based cluster:**

These are the clusters which are of arbitrary shape and can be modelled as dense regions in data space which are separated by sparse regions. In density-based clustering the clusters are of non-spherical shape.

Tool: Weka



Reference Dense Cluster Output:

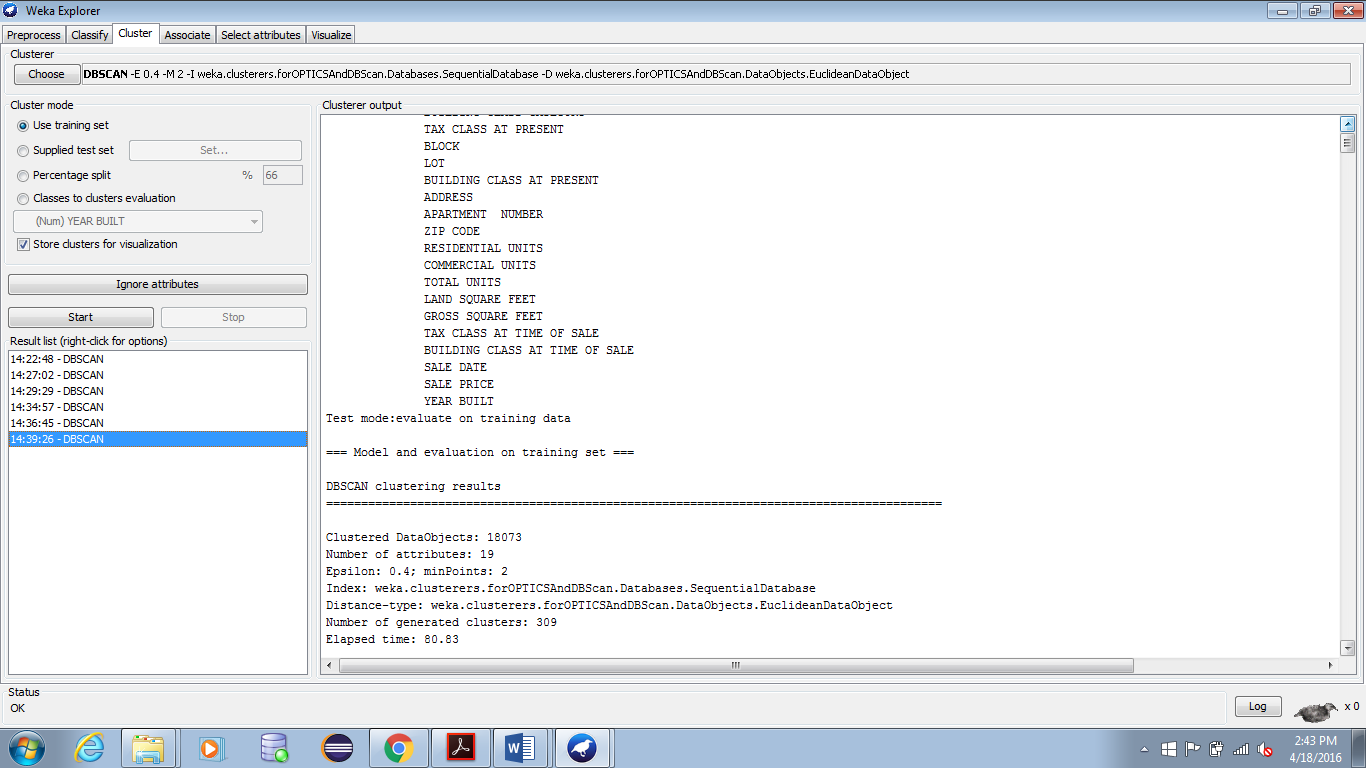


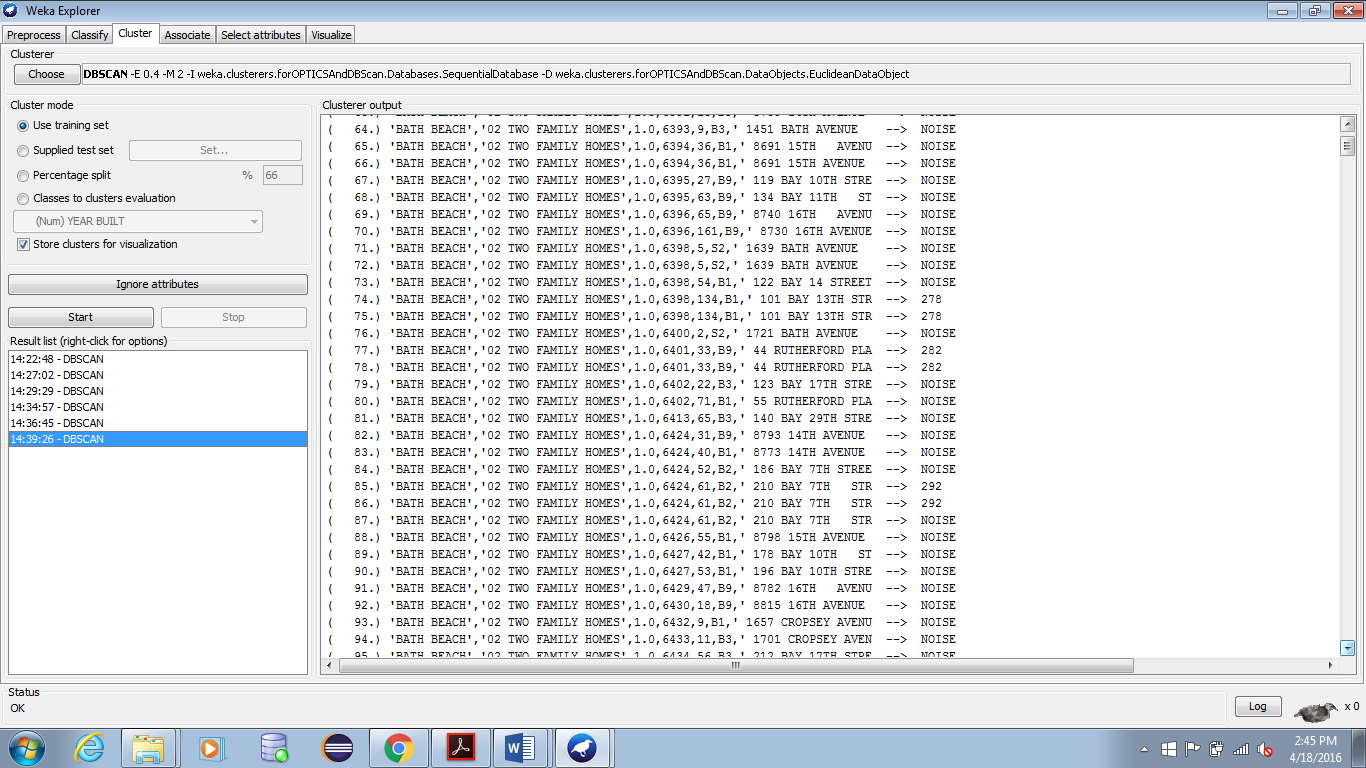
* **DBSCAN:**

DBSCAN stands for Density-based spatial clustering of applications with noise. In DBSCAN core objects are found which have dense neighbourhoods. In this method core objects and their neighbourhoods are connected to form dense and strong regions as clusters.

The DBSCAN calculation can recognize clusters in huge spatial data sets by having a look at the neighborhood thickness of database components, utilizing stand out information parameter. Besides, the user gets a proposal on which parameter esteem that would be reasonable. Therefore, minimal knowledge of the domain is required. The DBSCAN can likewise figure out what data should be named noisy or exceptions. Regardless of this, its working procedure is fast and scales exceptionally well with the extent of the database – directly. By utilizing the density distribution of hubs in the database, DBSCAN can sort these nodes into isolated clusters that characterize the diverse classes. DBSCAN can discover clusters of self-assertive shape. However, clusters that lie near each other have a tendency to belong to same class. The key drawback of DBSCAN is that they expect some sort of density drop to distinguish cluster outskirts. Additionally, they can't recognize intrinsic cluster structures which are predominant in the lion's share of genuine information

Tool: Weka



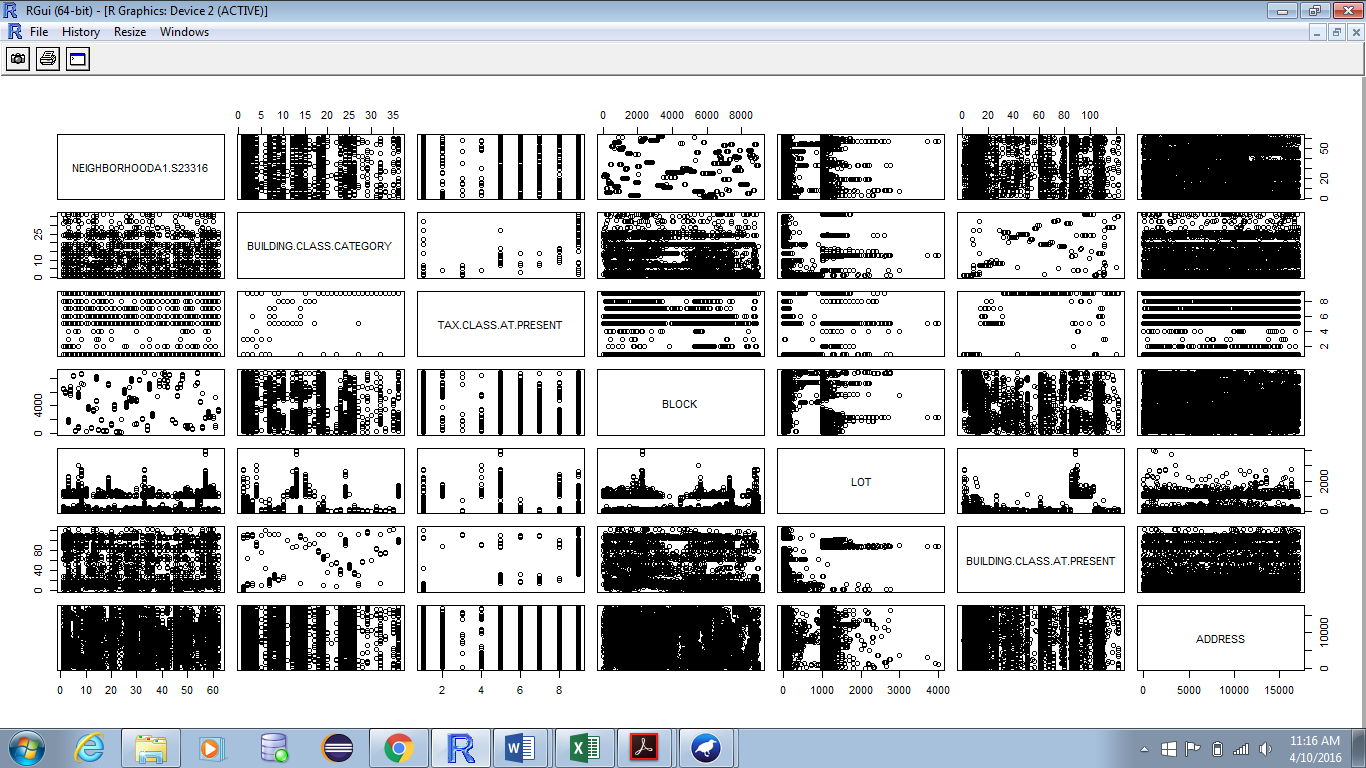


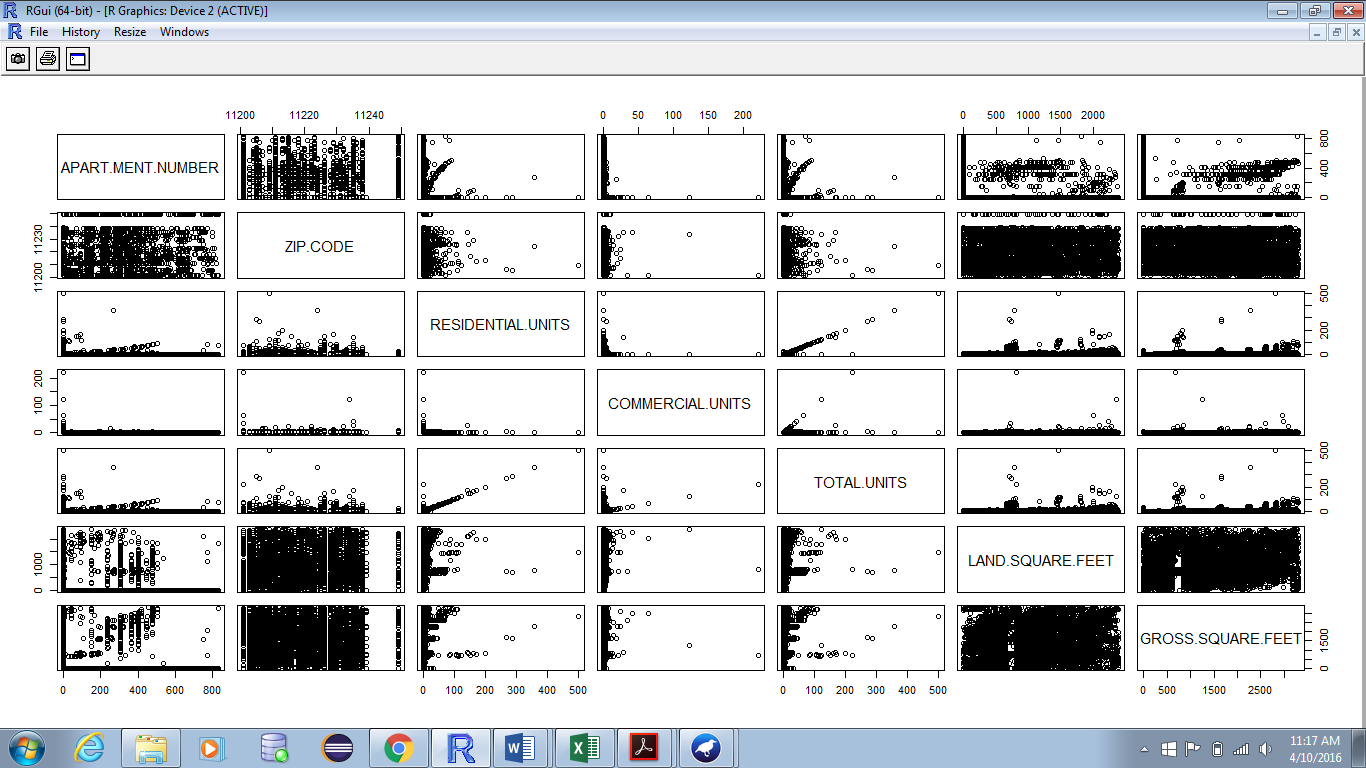
Reference DBSCAN output:

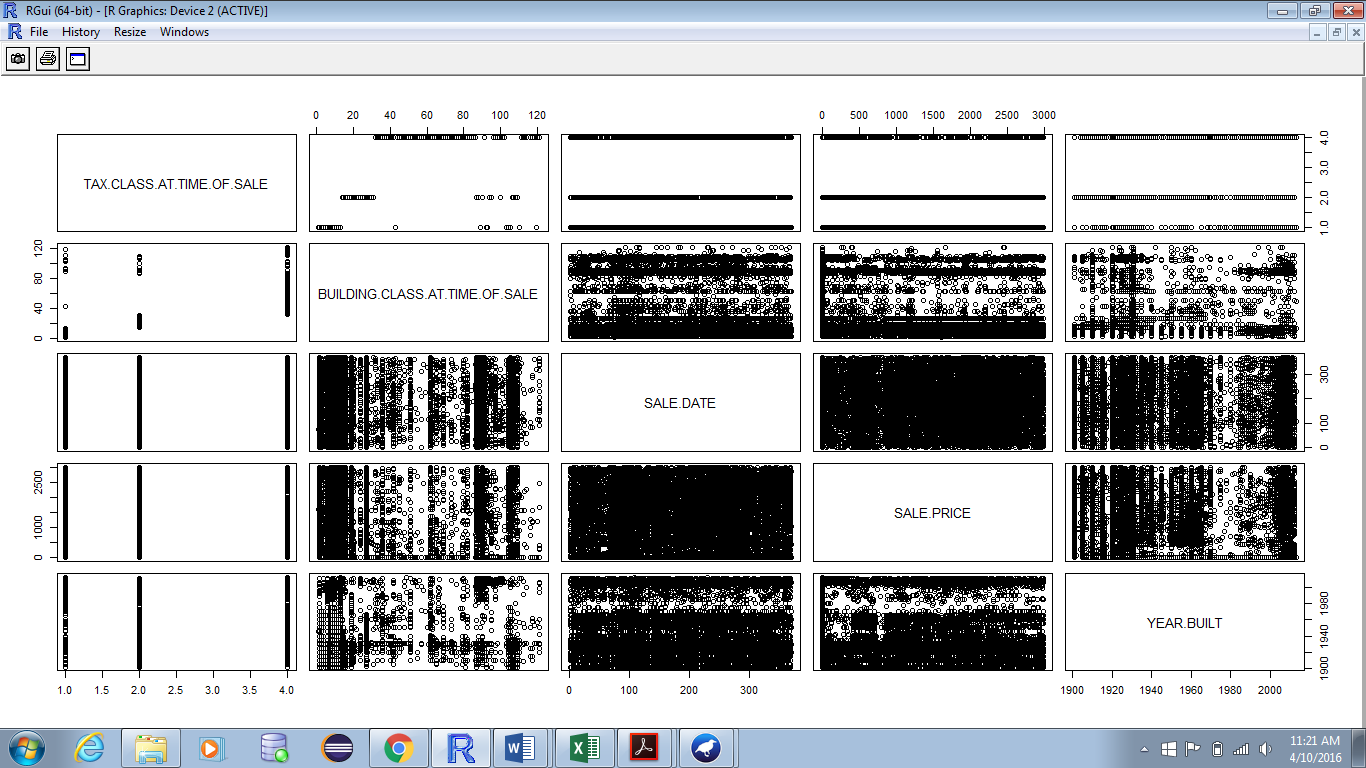


1. **Analysis**

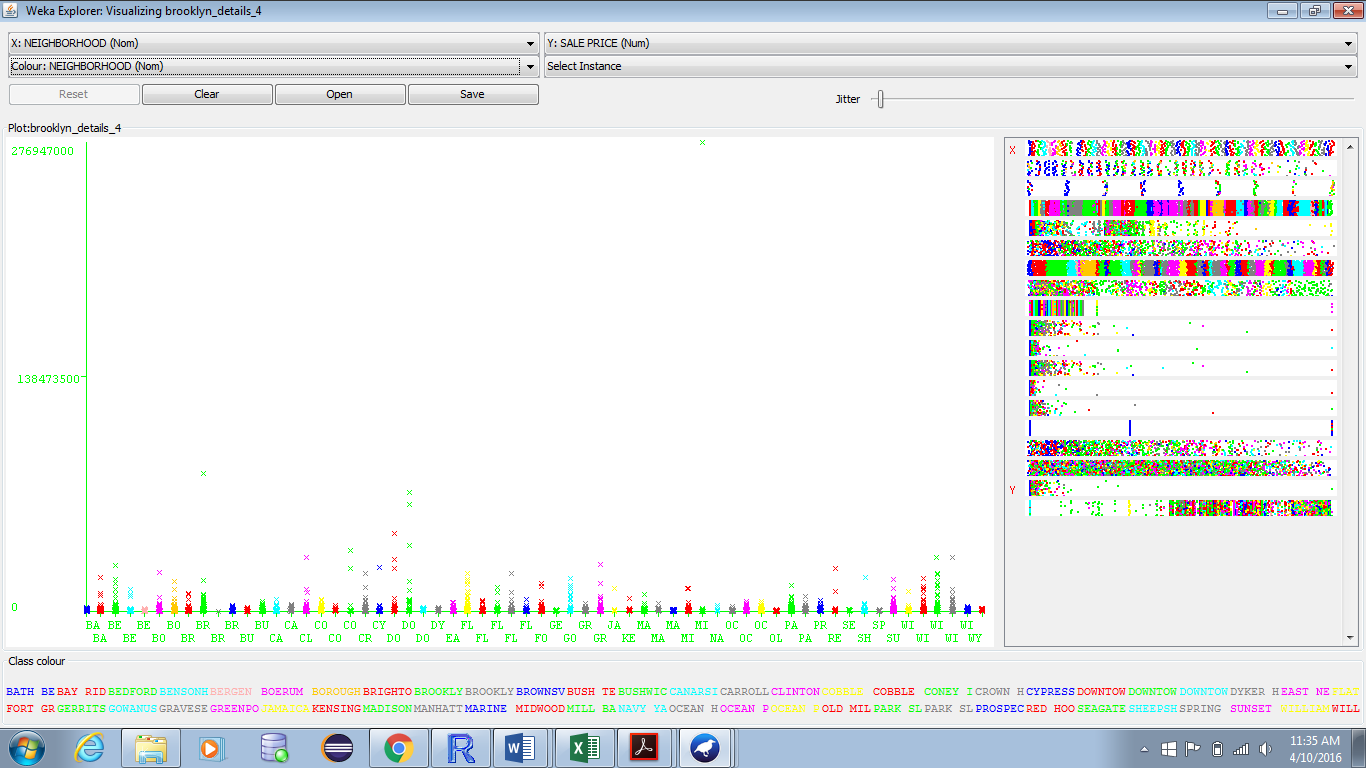
**Cluster Plot of entire data set**



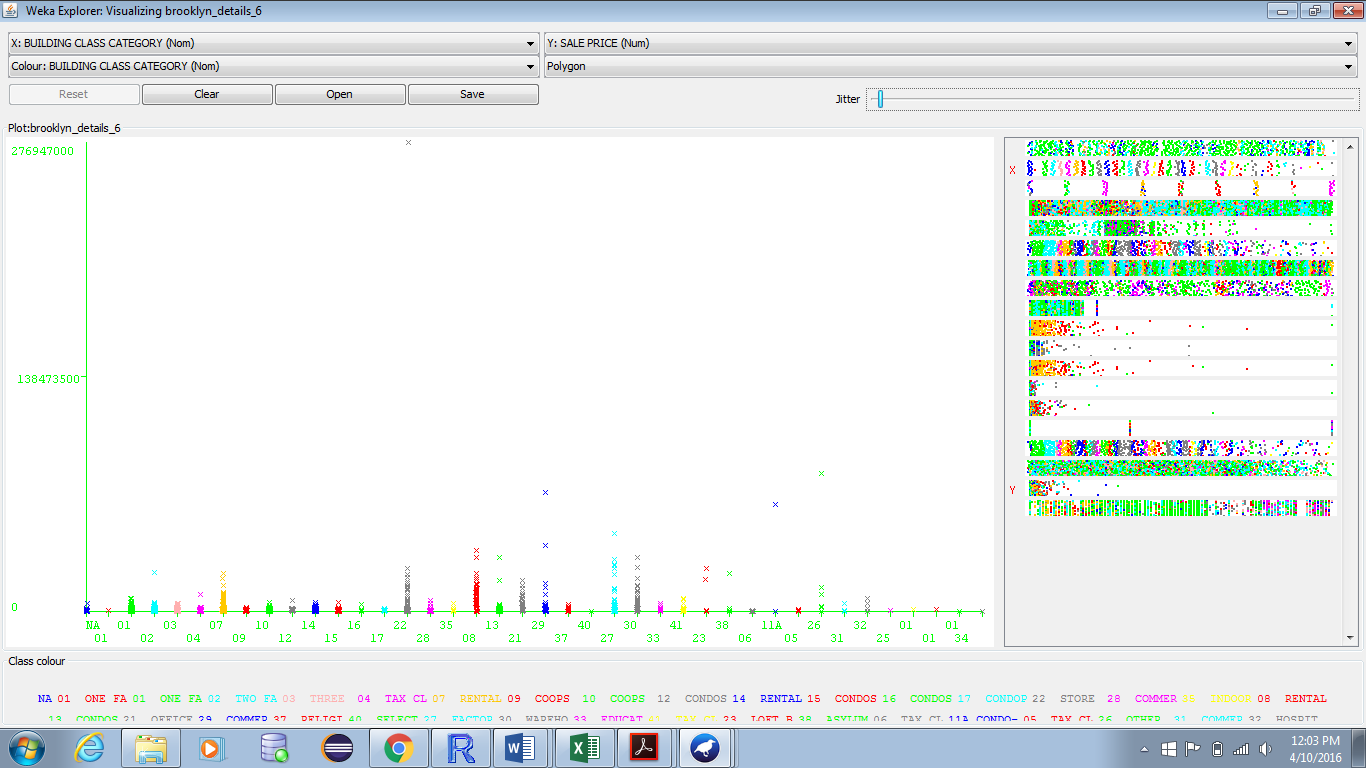


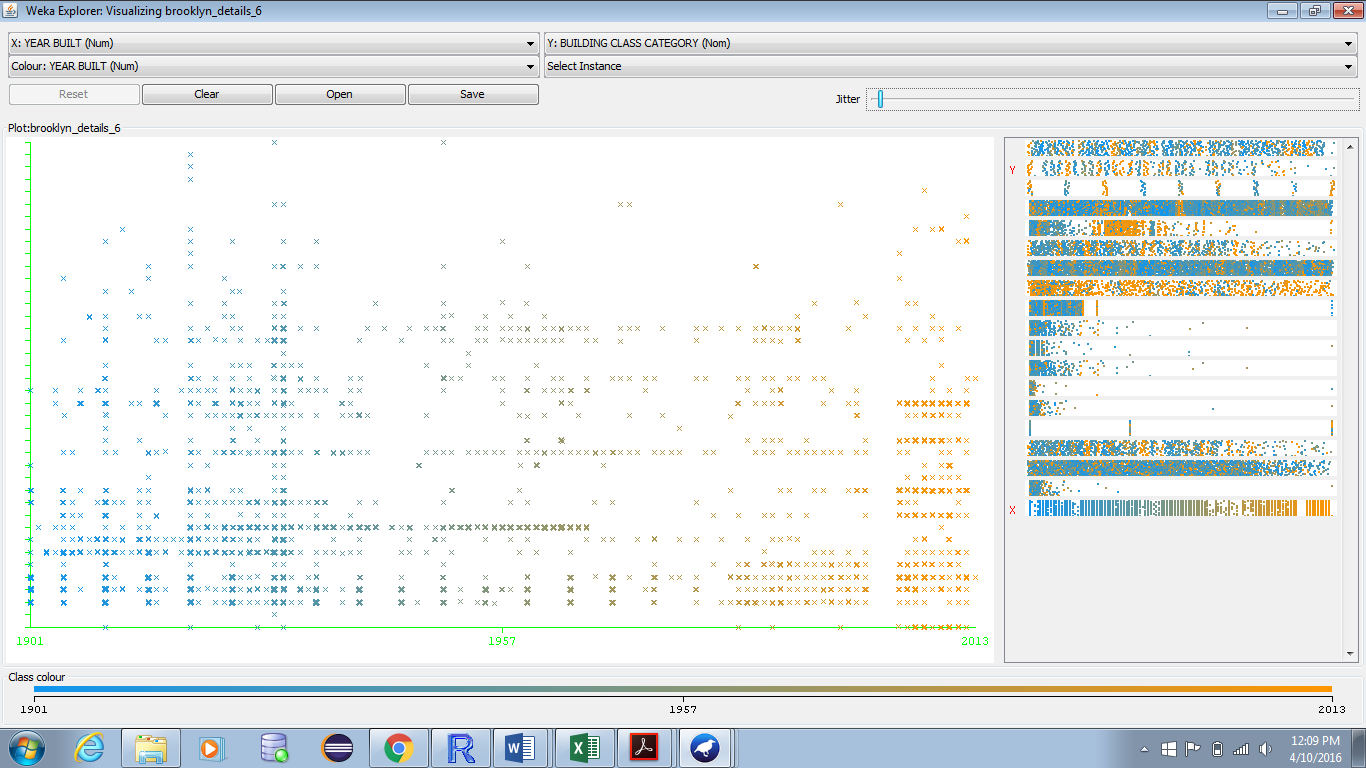


* Analysis based on neighbourhood. This segregates the number of buildings sold based on the locality and this helps the customers/clients to know in which locality does the buildings have higher sales.

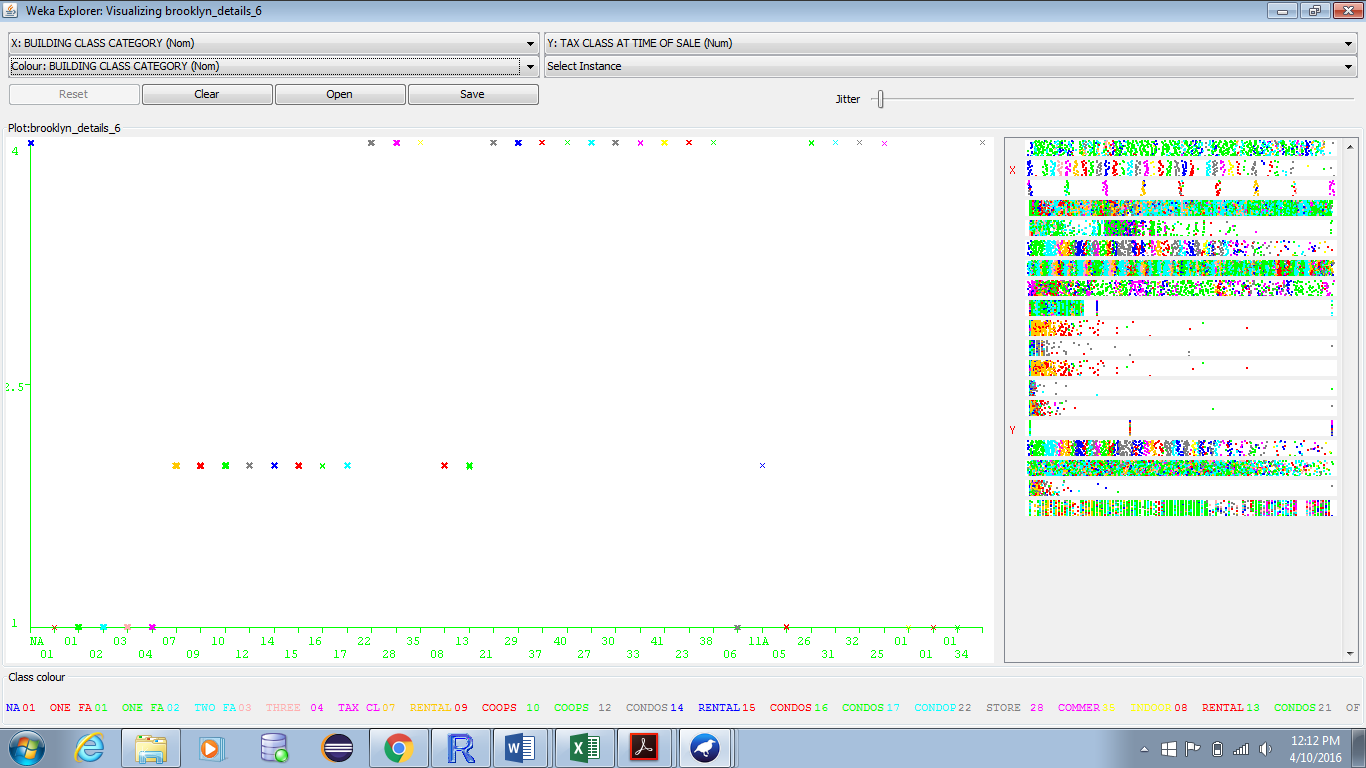


* Analysis based on building class category. This would help the companies to construct those kind of buildings which are most frequently bought.





* Analysis based on Tax class. This will determine the building class category based on the tax class like Class 1: Most residential property of up to three units, Class 2: All other property that is not in Class 1 and is primarily residential, Class 3: Most utility property, Class 4: All commercial and industrial properties.

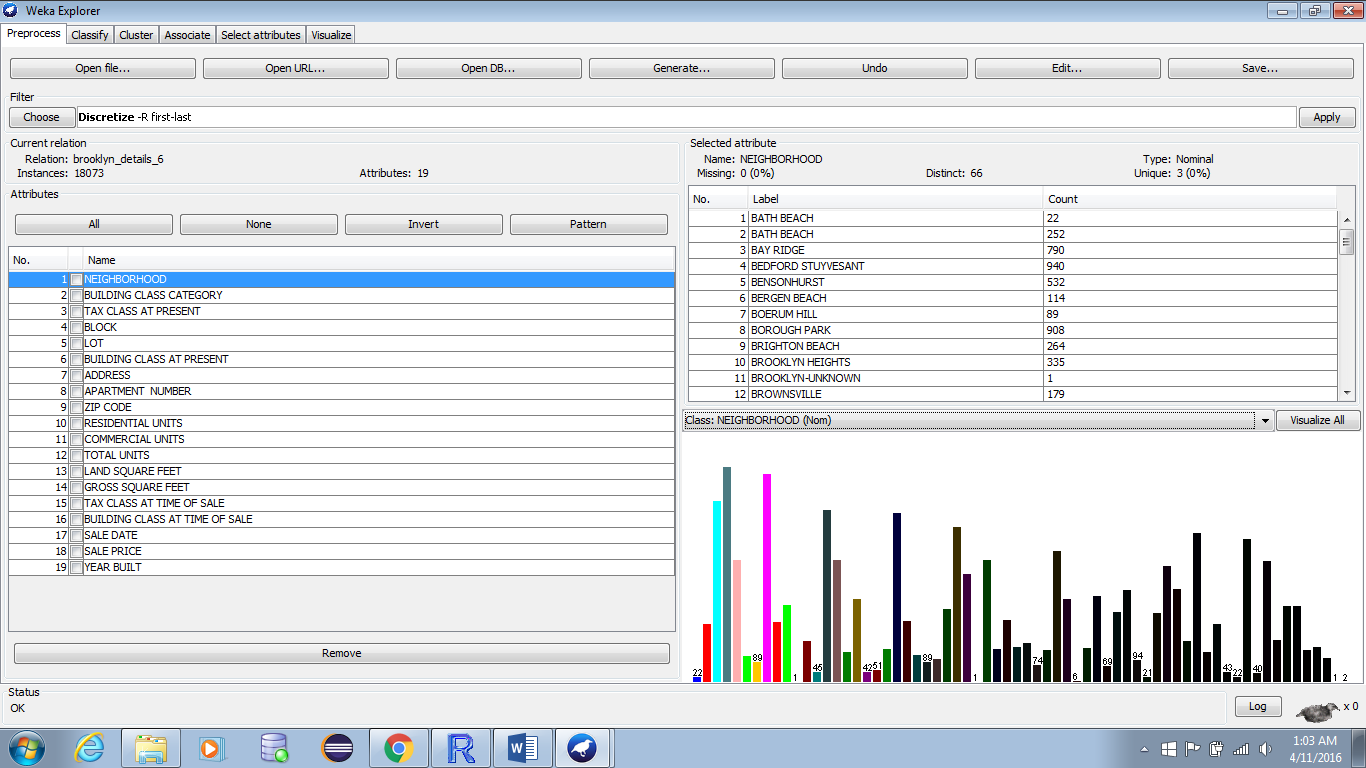


* Analysis based on neighbourhood and building class at time of sale. This gives the relation between the building class and the neighbourhood at time of sale.

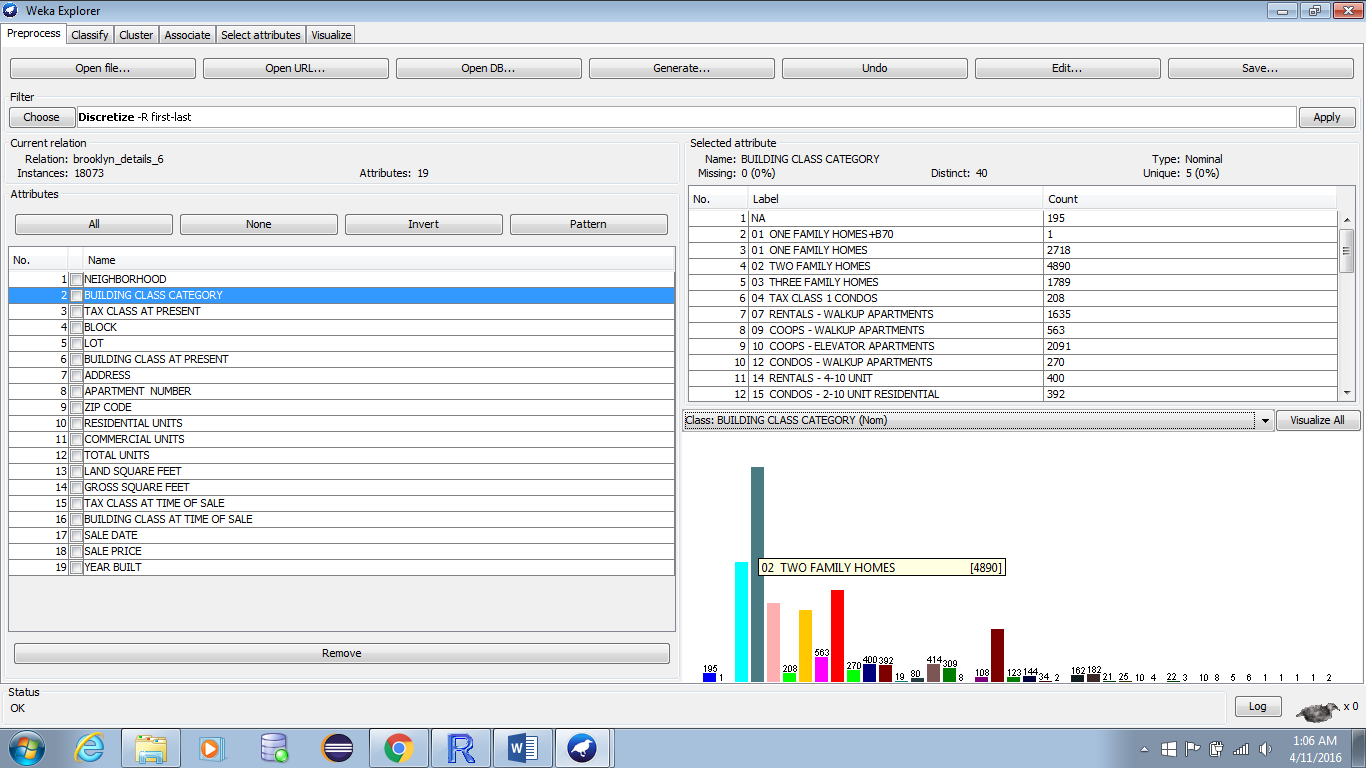


**Other Graphs:**

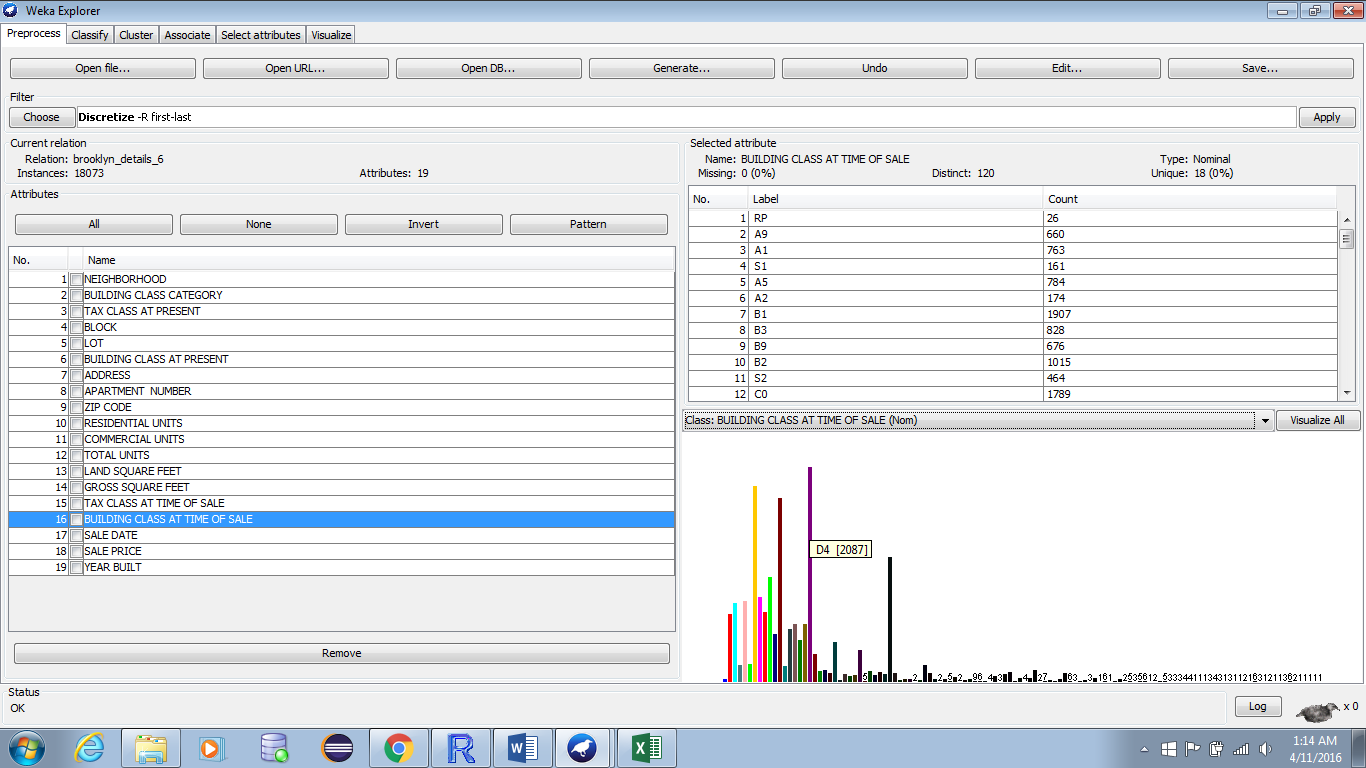
* Bar Graph for neighbourhood.



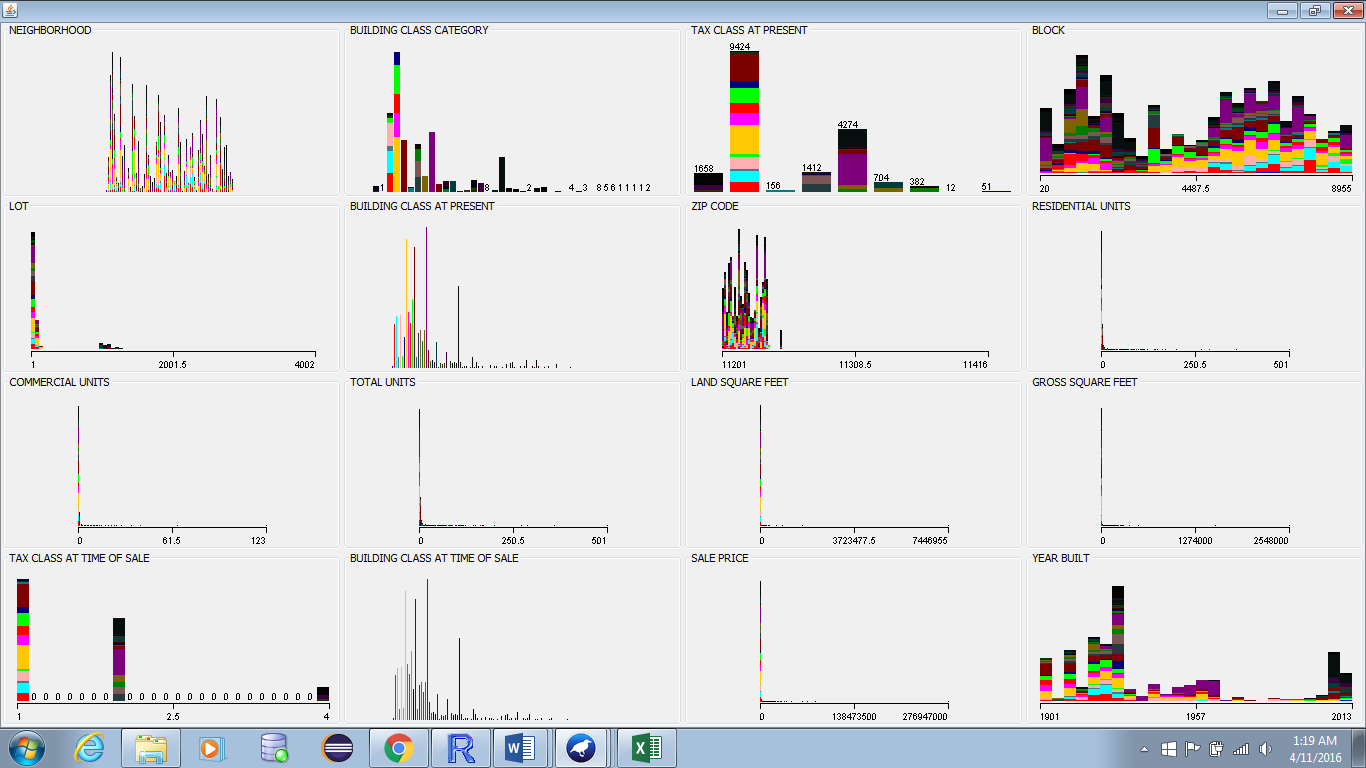
* Bar Graph for building class category.



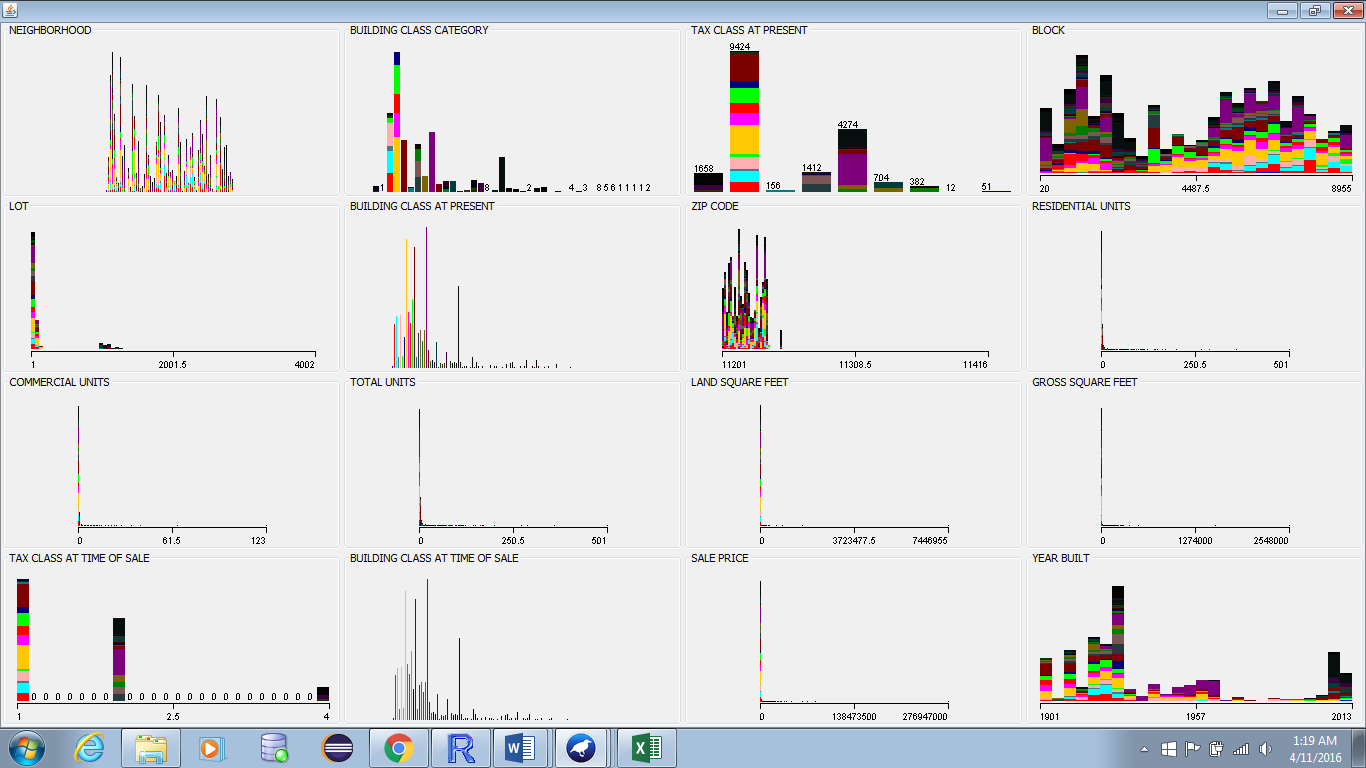
* Bar graph for building class at time of sale.



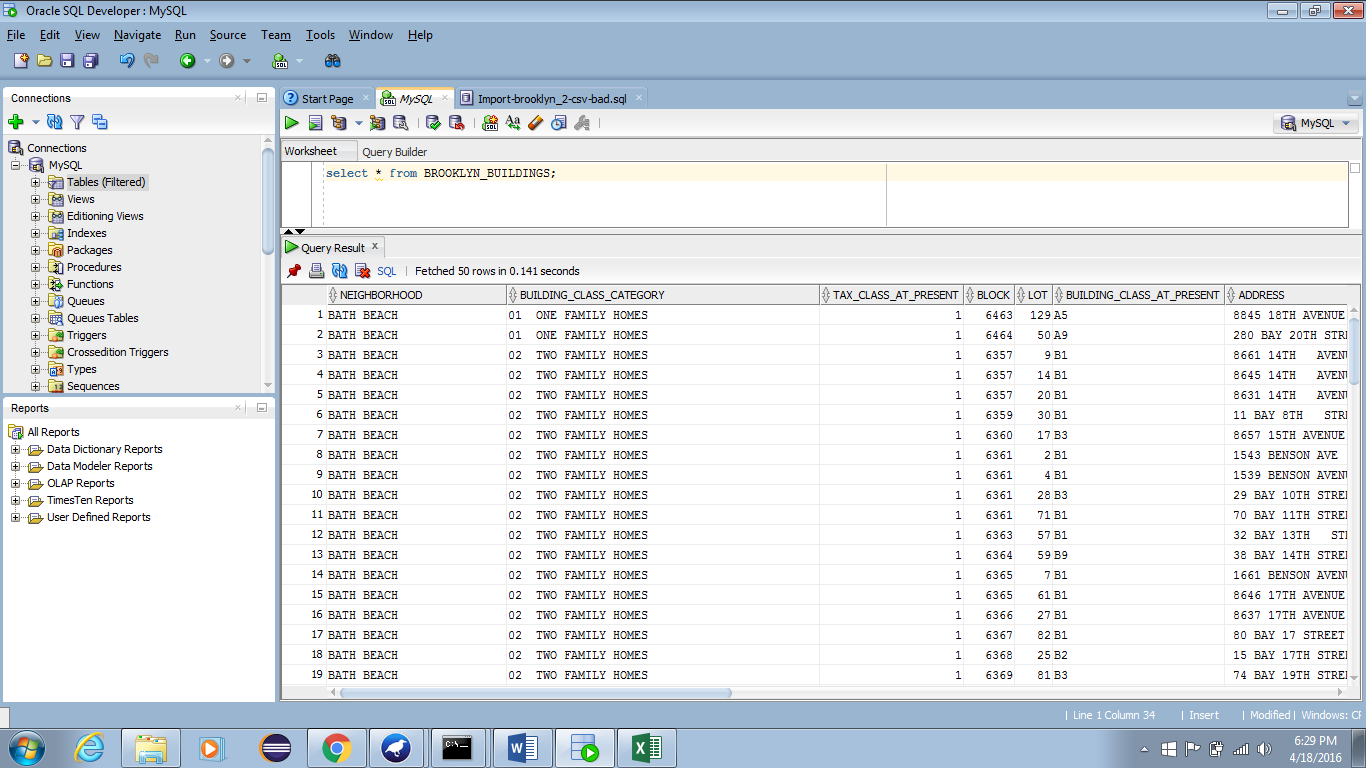
* Bar graph for Tax class at time of sale.



* Collective bar graph of all the attributes in data set.



* Data set migrated to database.



# CONCLUSIONS

This project work, we obviously classified and clustered the data from the general dataset considered to examine the sales of various categories of buildings that are considered in the data set. This aides in investigate and afterward in comprehension the building sales of Brooklyn in the data set. From the discoveries given in the above segment, unmistakably the classification and the clustering for the general data set considered are done impeccably. From the clustering, we are additionally ready to demonstrate the property sales based on the category selected.

# SUMMARY

* Data mining is the procedure of finding fascinating patterns from enormous measures of information. As a Knowledge discovery process, it commonly includes information cleaning, information combination, information selection, information change, pattern discovery, design assessment, and information presentation.
* Data sets are comprised of Data articles. A data object represents an entity. Data items are depicted by properties(attributes). Properties(attributes) can be nominal, binary, ordinal, or numeric.
* Data cleaning schedules endeavour to fill in missing qualities, smooth out commotion while distinguishing anomalies, and correct irregularities in the information. Information cleaning is normally executed as an iterative two-stage process comprising of discrepancy identification and data transformaton.
* **Classification** is a methodology which is used as a piece of data mining to recognize which set of characteristics these new entered observation has a place contingent upon the preparation set of data that fuses the cases which order this new perception has a spot.
* **Naïve Bayes algorithm** generally depends on the unexpected (or restrictive) probabilities. It will make usage of Bayes' hypothesis which is used to register probability or to discover likelihood by the strategy for counting each one of the frequencies that are associated with qualities moreover mix of authentic information. It is valuable in finding the likelihood (or probability) of two occasions.
* **Clustering** is the endeavour of accumulation of a game plan of things in a way that articles in the same get-together (called a group) are more comparable to each other than to those in various gatherings (groups).
* **K-means clustering algorithm** is the most definitely comprehended algorithm to settle the clustering inconveniences. The strategy that this algorithm takes after is general and uncomplicated way to deal with experience the grouping. It uses a clear and straightforward technique, first order the data and examines the result from classifier and uses that result for the cluster.
* A **Density Based Clustering** technique groups objects situated in light of the idea of thickness(density). It develops groups either as indicated by the thickness(density) of neighborhood items (e.g., in DBSCAN) or as per a thickness(density) capacity (e.g., in DENCLUE). OPTICS is a density based strategy that produces an augmented ordering of the information's bunching structure.

# REFERENCES

1. GitHub - oreillymedia/doing\_data\_science: This is the example code repository for Doing Data Science by Cathy O'Neil and Rachel Schutt (O'Reilly Media). (n.d.). Retrieved from [**https://github.com/oreillymedia/doing\_data\_science**](https://github.com/oreillymedia/doing_data_science)
2. Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA---Experiences with a Java Open-Source Project. The Journal of Machine Learning Research, 11, 2533-2541.
3. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques concepts and techniques. San Francisco: Morgan Kaufmann In.
4. Cushman & Wakefield (2014), Property Sales Report, Brooklyn, NY.
5. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.
6. McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).