Process of Data Mining and Exploratory Data Analysis

Vinita Tomar, Piyush Devliyal

Department of Computer Science, Maharaja Surajmal Institute, C-4 Market, Fire Station Rd, Janakpuri,New Delhi, Delhi 110058 India

Abstract

In data mining we extract the vital information from the given data by following various steps . These steps are to be followed in a systematic manner and they help in making our workflow easier.
Data Mining becomes a crucial part when we have a huge database , as handling it becomes a tedious task, so after mining from them improved decision making could be done that may help us increase our gains. This paper discusses about the steps of data mining and application of them on a real dataset

Keywords: (Knowledge Discovery in Databases)KDD, Data Visualization, Data Mining, Data Analysis

Introduction

There's a rich amount of availability of the data over the worldwide web but having just the raw data doesn't help us in any way . The data needs to be processed properly so that we can get some relevant and required information that can help us in analyzing the data and making better decisions for future possibilities. As there's emerging technology of Artificial Intelligence , it helps in improving the data mining process a lot , we can do better statistical analysis and have models of higher accuracy that can help in getting much better predictions. Aim of data mining is to retrieve knowledge from large databases. Various data mining techniques are used depending on the type of data set available.
Some people treat data mining as a term similar to KDD whereas some consider it as just a part of the process of KDD. The steps involved in KDD followed in sequential manner are as follows:

1. Data Cleaning (remove inconsistent and irrelevant data)
2. Data Integration (data from various sources is merged together)
3. Data Selection (data relevant for analysis is selected)
4. Data Transformation (data is transformed and summarized into forms so that the data is ready to be mined)
5. Data Mining (Various techniques and methods are performed so as extract maximum patterns)
6. Pattern Evaluation (the interesting patterns are identified representing the knowledge based on interestingness measures)
7. Knowledge Presentation (Visualization and knowledge representation techniques are used to show to retrieved knowledge)

Even though Data Mining is just one step in the Knowledge Discovery process , it's still a very important one as after this step only the patterns are discovered. The sources for data are various like the internet, data warehouses, or dynamic data.

Mining Frequent Patterns

Frequent Patterns means that the patterns that occur frequently in our database.
There are various kinds of frequent patterns , including frequent sub sequences, frequent sub structures, frequent itemsets. A frequent itemset refers to a set of items that may appear together in a dataset like for example mobile phone, back cover , screen guard these items are frequently bought together whenever a customer purchases a new mobile phone. A frequent subsequence refers to a frequently occurring subsequence such as if a person purchases a camera , followed by a tripod stand and then more accessories . A frequent substructure refers to different structural forms that are in combination with itemsets or subsequences.
When data is mined there's a possibility of generating a number of patterns , not all the patterns that are generated would be useful to the user. Patterns that are classified as useful are the ones that could be easily understood by the end users, and can be used to generate results on test data with some level of accuracy. The patterns that are found to be interesting or useful represent the knowledge.

Steps of KDD

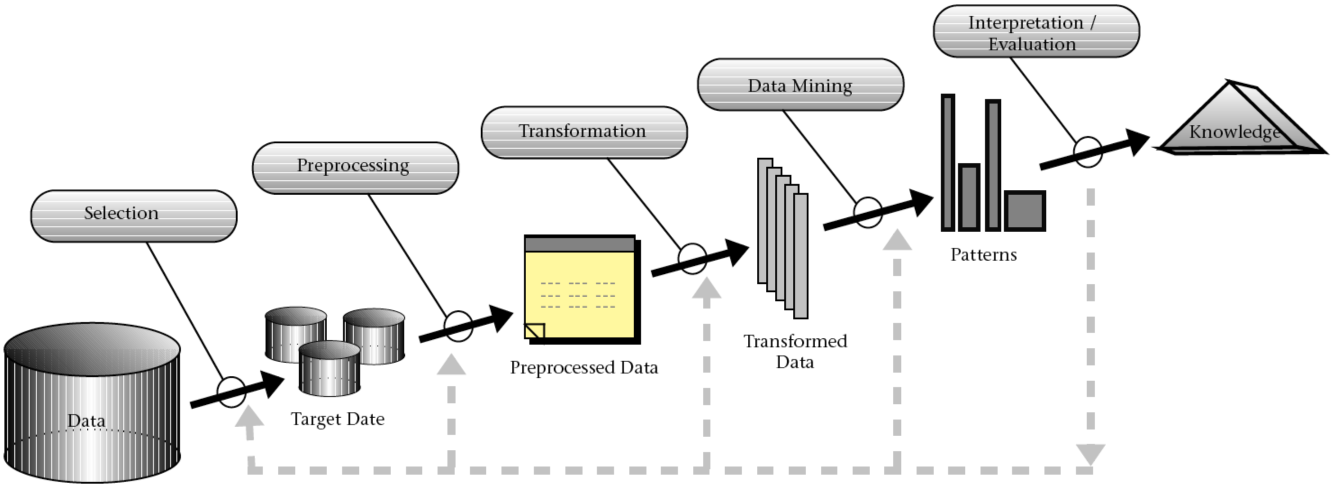


Figure 1 Steps of KDD

Data Cleaning

This is the first process of data preprocessing . By data cleaning we mean that by removing inconsistencies and redundancies from the data , we can either fill in the missing values or we can smoothen out the dataset by replacing it with a suitable central tendency value . If we start doing data mining on dirty or unprocessed data , then the results or patterns obtained from it won't be any useful. Moreover the dirty data can cause some undesirable outputs.

Data Integration

In Data Integration we’re using the data from multiple sources that are stored in data warehouses and it might have some inconsistencies in it , like if one of the table contains a column of Last Name whereas in other table that column is named as Surname , so if we integrate such data without any inspection then it will cause redundancies and inconsistency. In this case having a large database won't help in getting better results rather they would cause creating some confusing patterns that would reduce the quality of knowledge obtained.
During the process of data integration only , the process of data reduction is also completed . If we’re having a huge database then it's sure it will slow down our entire process of mining , so by data reduction we’re working on a smaller set of data that will provide us similar results as compared to large datasets. Data Reduction could be in done in two ways : Dimensionality Reduction (some data encoding schemes are applied to obtain a summarized representation of our huge data) , Numerosity Reduction (data are replaced with smaller representations using parametric and non parametric models)

Data Selection

In the Data Selection process we’re trying to reduce the size of the dataset by eliminating the unnecessary attributes or dimensions . The primary goal of data selection is to find a small subset of our dataset which can help in producing the almost same quality of results or patterns as with the original dataset.
When we do mining on a small dataset the results are obtained much faster and as the redundant attributes don't take a part during the process of pattern discovery , so it becomes much easier to understand the patterns.
Greedy Methods like Forward Selection , Backward Elimination and Decision Tree Induction are used for selection.

Data Transformation

In the Data transformation step of KDD , the data are transformed into appropriate forms so that they could be mined , like for example we can have dataset of weights of person of all around the world , there could be use of different weight measures like for in India we will be using a Kg unit whereas for UK we will be having a unit of Lbs , even though our dataset is not having missing values but the results obtained from it would be too much inaccurate due to difference in units used so it's important to have the same unit used throughout the whole dataset.
Few of the strategies used in data mining are :

1. Smoothing: Here the aim is to reduce or eliminate the noise from the data. Techniques like clustering , binning and regression are used.
2. Attribute Construction: A new attribute is constructed from existing attributes that will help us in a better mining process, like for example If we have an attribute for date of birth , so it could be converted into Age by doing some basic calculation over the values.

3. Normalization : The data which we’re having might not be uniform or smooth , so to make the data values fall in a particular range -1.0 to 1.0 or 0.0 to 1.0 normalization is used.
4. Discretization: When the data is in non continuous or discrete form then the numeric values of attributes could be replaced by interval labels. For example if we have data of marks of students then we categorize them into ranges like (40-50, 50-60,60-70, and so on) if required we further break the interval into smaller intervals.
5. Concept hierarchy generation for nominal data: In this technique the attributes are generalized into higher level concepts. For example in the address of a person, the street attribute can be generalized into district attribute and if we want to go higher in the conceptual hierarchy then as city.

Data Mining

Data Mining means extracting or retrieving vital information from a large dataset. It is a process of discovering patterns from a large number of databases . The main goal of data mining process is to retrieve the information and transform it into an understandable manner so that it can be used for future purposes. Data mining has helped in improving the decision making by insightful analysis. Data Mining Techniques can be divided into two main purposes; either they can describe the dataset or they can predict outcomes through the use of machine learning algorithms.
Some of the Data mining Techniques:

1. Association Rules: It’s a rule based method in which relationships between variables in a dataset are found. Used for : Market Basket Analysis, Products Relation Analysis.
2. Neural Networks: It’s a set of connected input and output units and each connection has some weight associated with it. It’s helpful when the dataset's trends are too complex and can’t be evaluated by humans or any other techniques.
3. Prediction: It uses a combination of other techniques like classification, clustering etc. It analyses the past instances in the right sequence to predict the future event.
4. Outlier Detection: In this technique the data items that don’t match with the expected pattern are observed. Majority of the datasets have outliers, they help in fraud detection and interruption identification.

Pattern Evaluation

All the patterns that are generated by data mining techniques are not interesting . Interestingness of a pattern solely depends on the user. Techniques are required to discover the interestingness of discovered patterns based on a few measures.When we’re using user specified constraints , then we may generate more interesting patterns.

Knowledge Presentation

This is the last step in KDD , where the knowledge discovered is presented to the user using various visualization techniques and tools so that it can be easily understood. We can represent the obtained knowledge in forms of graphs , trees , tables , tree maps, plots etc.

Data Description

Dataset of Restaurant Analysis has been taken from Kaggle. In this Restaurant Analysis Dataset there were a total of 12133 samples and 104 columns . After analyzing the dataset relevant columns were selected for the mining process.
After the selection of subset from the dataset our dataset table looked like this

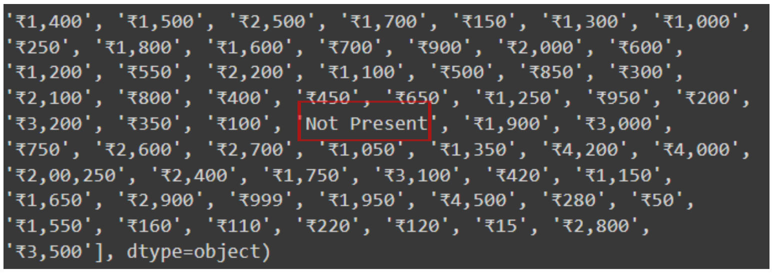
#	Column	Non-Null Count	Datatype
0	Restaurant Name	12133	object
1	Locality	12133	object
2	Type	12133	object
3	Rating	12133	object
4	Number of votes	12133	object
5	Cuisines	12133	object
6	Cost for two	12133	object
7	Payment Modes	12133	object
8	Free Parking	12133	int64
9	Home Delivery	12133	int64
10	Valet Parking Available	12133	int64

As you may observe, it has many columns as datatype of object , so we need to further work upon them to convert them to their designated data type.

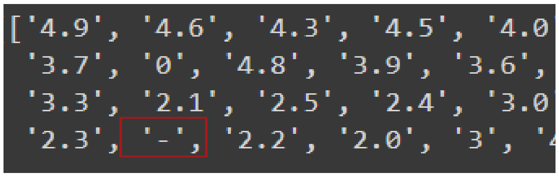
Process for Exploratory Analysis

Data Cleaning:

Even though you may observe that we have the same number of entries for each column but having just a value doesn't help us in any way, it needs to be a relevant one so that it can contribute to the process of knowledge retrieval. This is a vital step as accuracy of analysis depends on quality of data. For example in the column of Cost for two there were a few rows having the entry “Not Present” , similarly few entries in Rating column has an entry of ‘-’ , these won't anyhow help us in the process of knowledge discovery so we can either drop those entries or we can substitute them with the mean value of the column.



```
'₹1,400', '₹1,500', '₹2,500', '₹1,700', '₹150', '₹1,300', '₹1,000',
'₹250', '₹1,800', '₹1,600', '₹700', '₹900', '₹2,000', '₹600',
'₹1,200', '₹550', '₹2,200', '₹1,100', '₹500', '₹850', '₹300',
'₹2,100', '₹800', '₹400', '₹450', '₹650', '₹1,250', '₹950', '₹200',
'₹3,200', '₹350', '₹100', 'Not Present', '₹1,900', '₹3,000',
'₹750', '₹2,600', '₹2,700', '₹1,050', '₹1,350', '₹4,200', '₹4,000',
'₹2,00,250', '₹2,400', '₹1,750', '₹3,100', '₹420', '₹1,150',
'₹1,650', '₹2,900', '₹999', '₹1,950', '₹4,500', '₹280', '₹50',
'₹1,550', '₹160', '₹110', '₹220', '₹120', '₹15', '₹2,800',
'₹3,500'], dtype=object)
```



```
['4.9', '4.6', '4.3', '4.5', '4.0
 '3.7', '0', '4.8', '3.9', '3.6',
 '3.3', '2.1', '2.5', '2.4', '3.0
 '2.3', '-', '2.2', '2.0', '3', '
```

Data Integration

As we have used the dataset from only source so we're saved from performing this step, if there would have been multiple sources of data then this process becomes essential

Data Selection

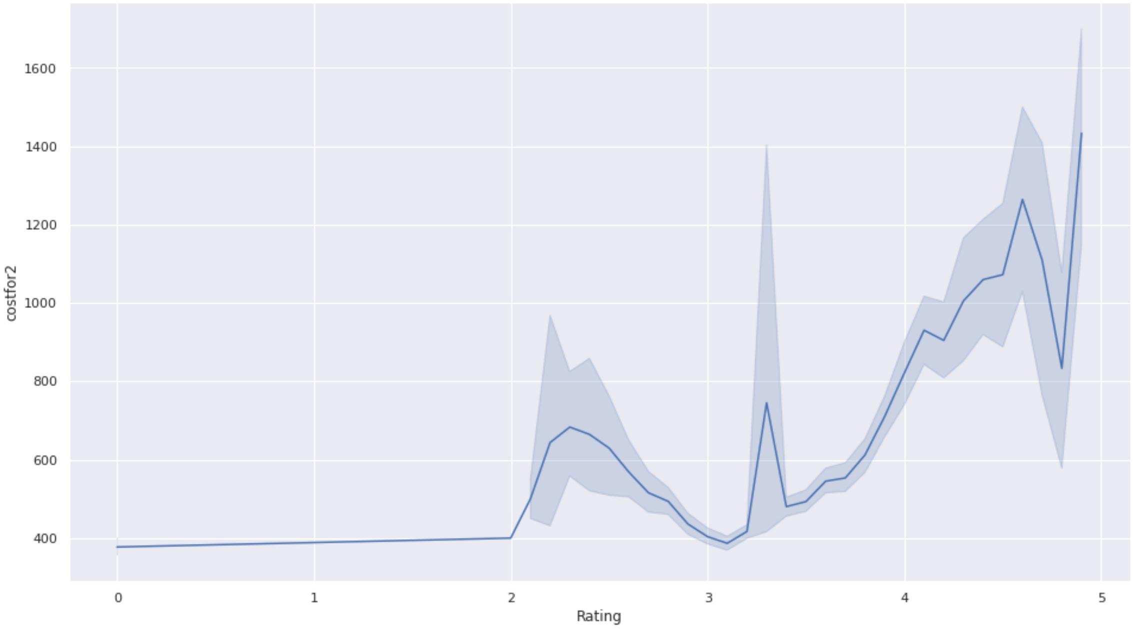
In the process of Data Selection we try to create a subset of our data such that it is capable of giving the same results as with the larger dataset.
Like there are a few locations where we just have 1-5 Restaurants located ,keeping this kind of entries in the dataset would just cause redundancy in pattern discovery, now we have a choice: we can either group them together and categorize them as 'miscellaneous' , or we can truncate those entries. In our case we categorized such entries' locations as 'misc'.

Data Transformation

Now we need to convert the datatype of our columns to relevant one's so that we can make the data operable .The datatype of 'cost for two' and 'rating' should be a float or an integer, otherwise it's just considered as a string and on strings we can't apply any mathematical operations, similarly the 'number of votes' should be of an integer datatype.

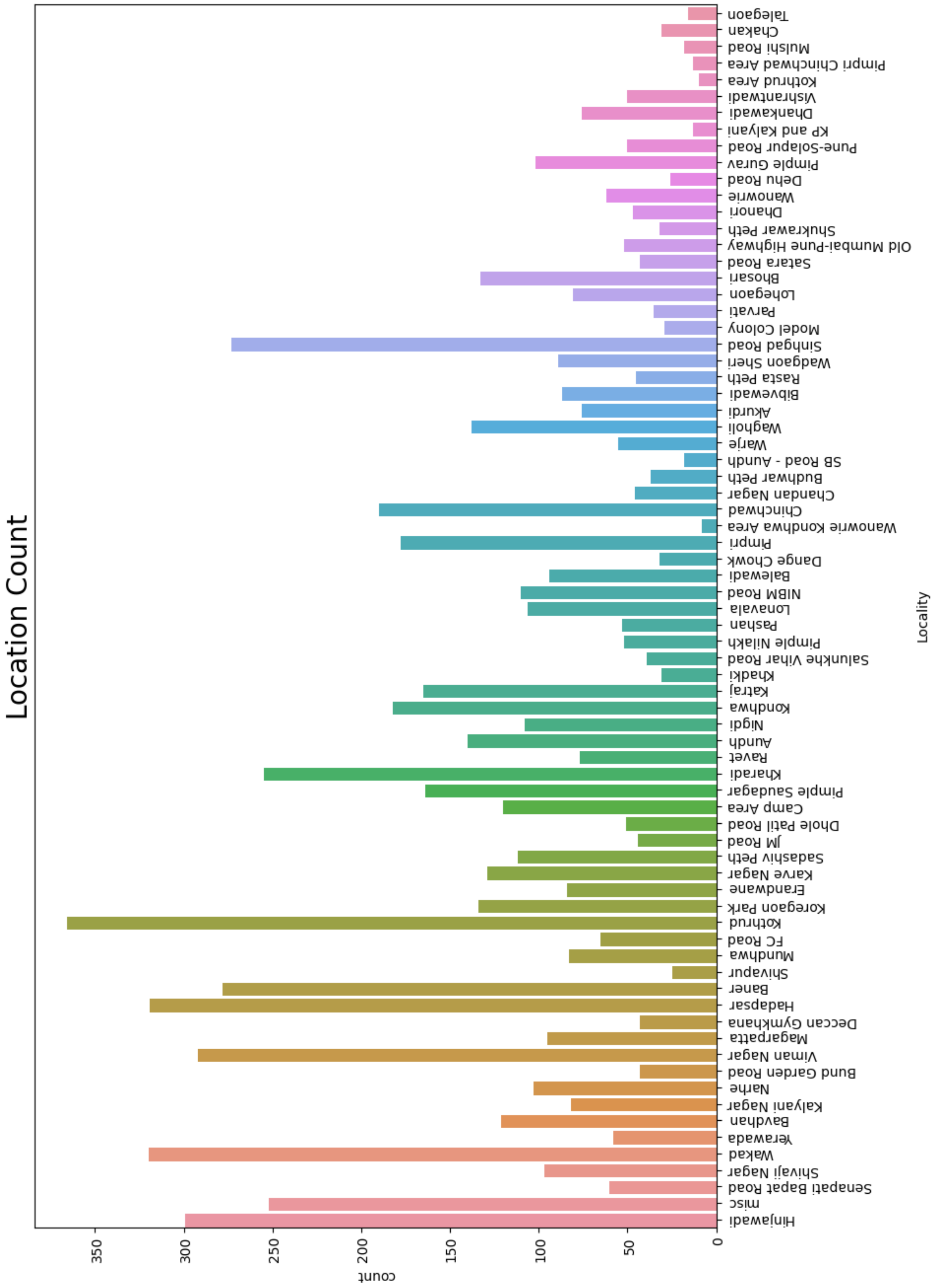
Analysis Results

1)Cost for two v/s Rating



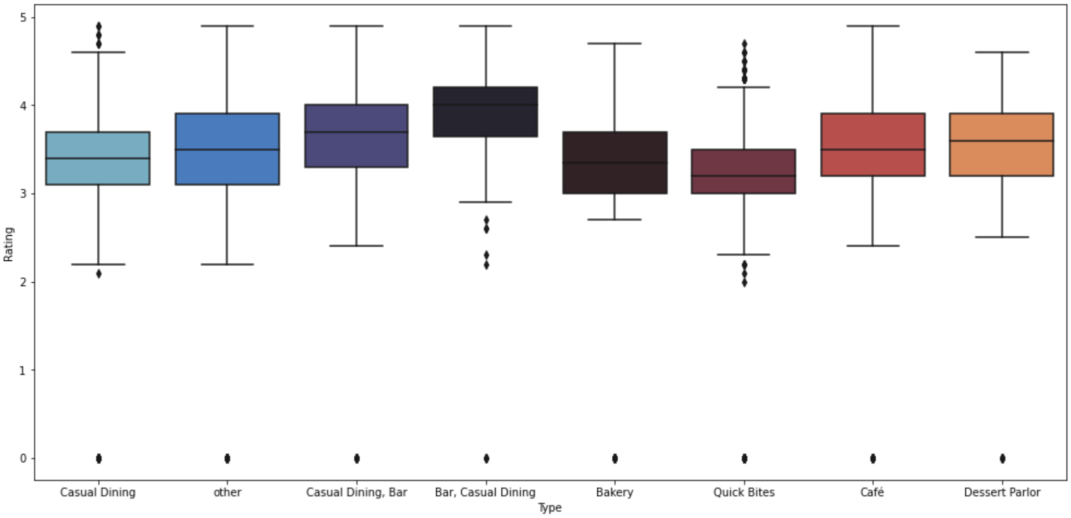
We have observed that if cost for two is 400 or less then people don't even tend to rate or the max they give is a rating of three , generally restaurants that have 'cost for two' 800 and above gets a rating between 4 and 5 , and for the highest 'cost for two' i.e. 1400 in this case people give a rating of nearest to 5.0. , So if opening a restaurant then we should try to keep cost for two between 800 and 1400 for getting a good rating

2)Location Count



If we want to open a restaurant in the given city then we can observe from the graph that Kothrud has the maximum number of restaurants open so definitely it won't be our choice , we need to choose which locality does have the minimum no of restaurants open , like we can choose Wanowrie , Kalyani , Talegaon

3)Box plot of Type of Restaurant and Rating

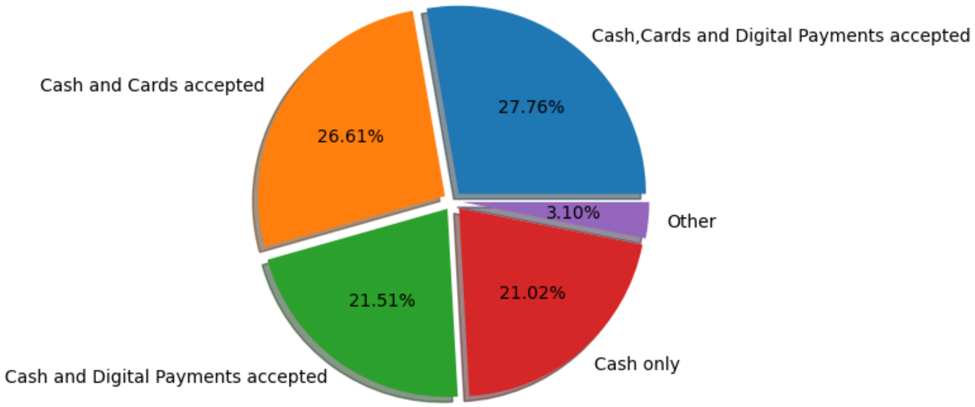


(Casual Dining, Bar) and (Bar, Casual Dining) are not same type

From the above box plot we can observe that on an average the type: “Bar and Casual Dining” gets the highest rating among all the types of restaurants. The least popular type is : “Quick Bites”

---

4)Payment Modes accepted by restaurants

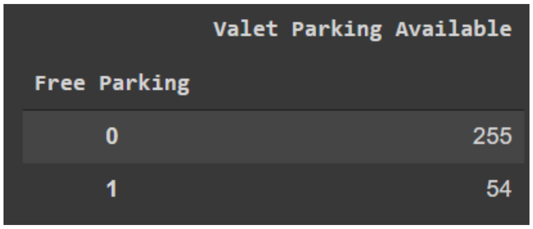


Majority of the restaurants accept Cash, Cards and digital payments , however there are a few restaurants that accept some other payment methods too like meal coupons .

---

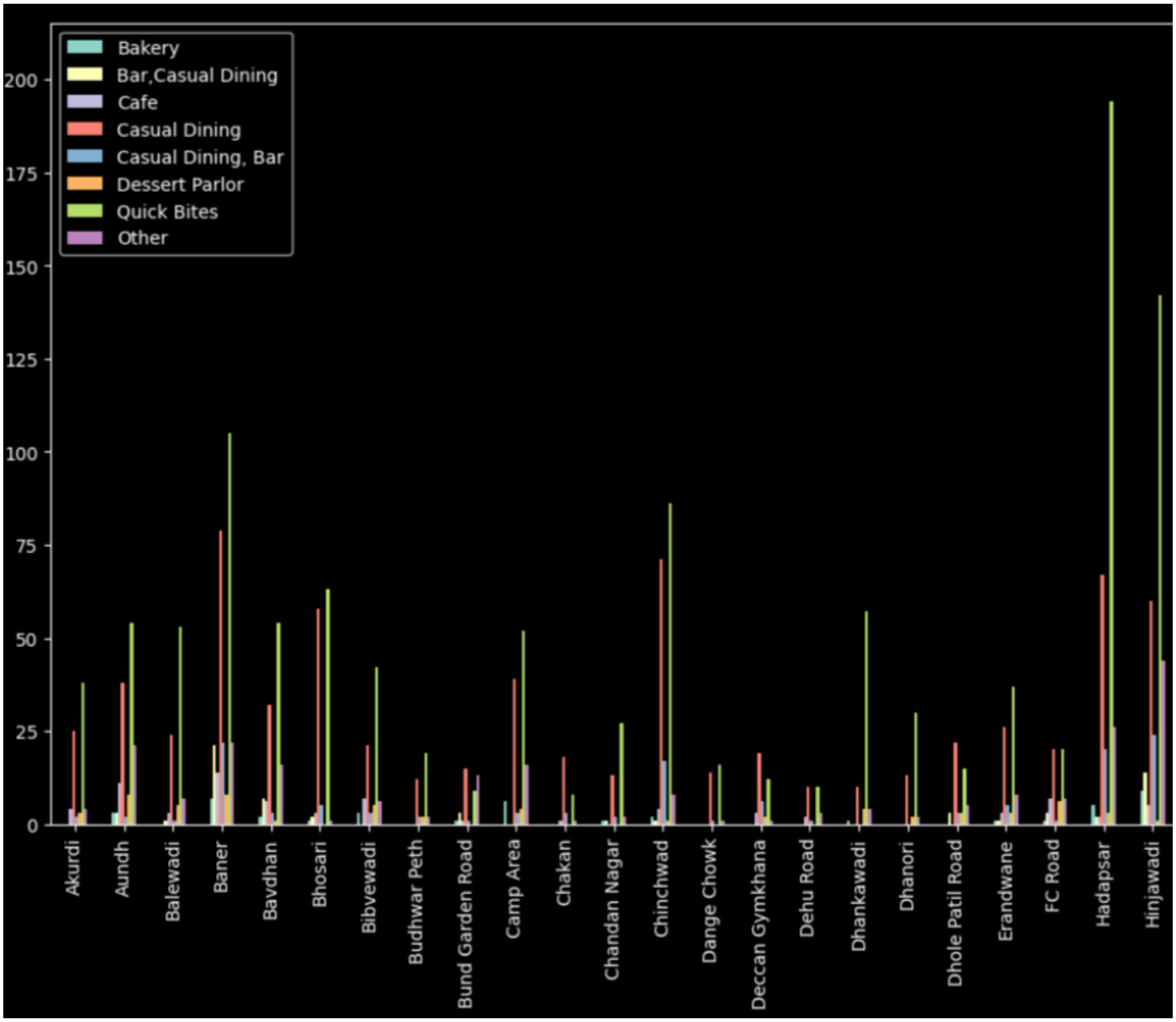
5)Valet Parking v/s Free parking

Only a few of the restaurants offer valet parking and out of which approximately only 55 offer it for free,
It means that restaurants don’t prefer to offer valet parking and if they even do so they go for paid parking.

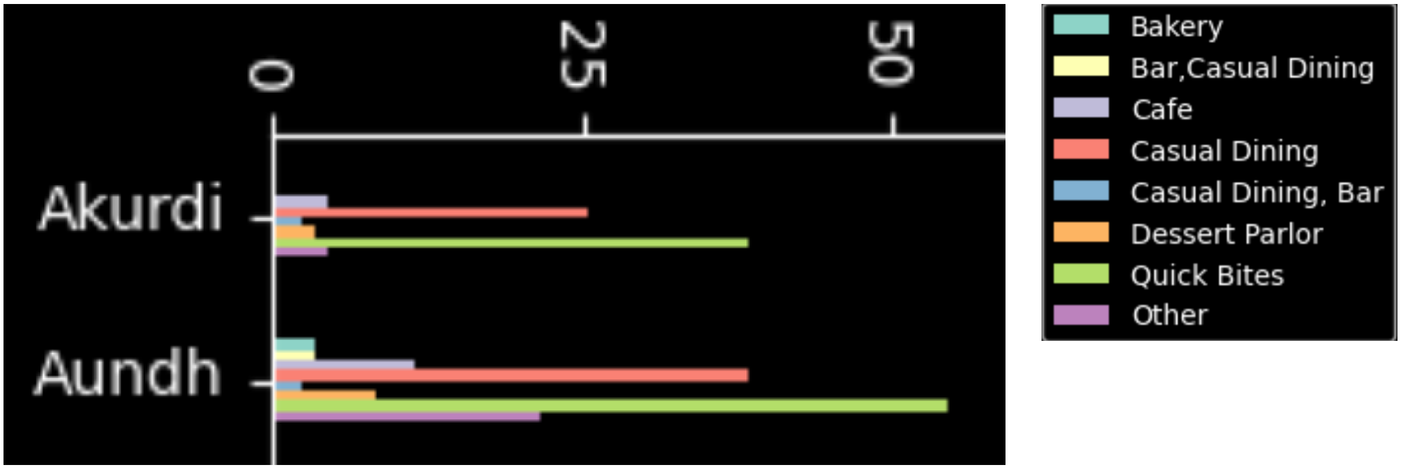


	Valet Parking Available
Free Parking	
0	255
1	54

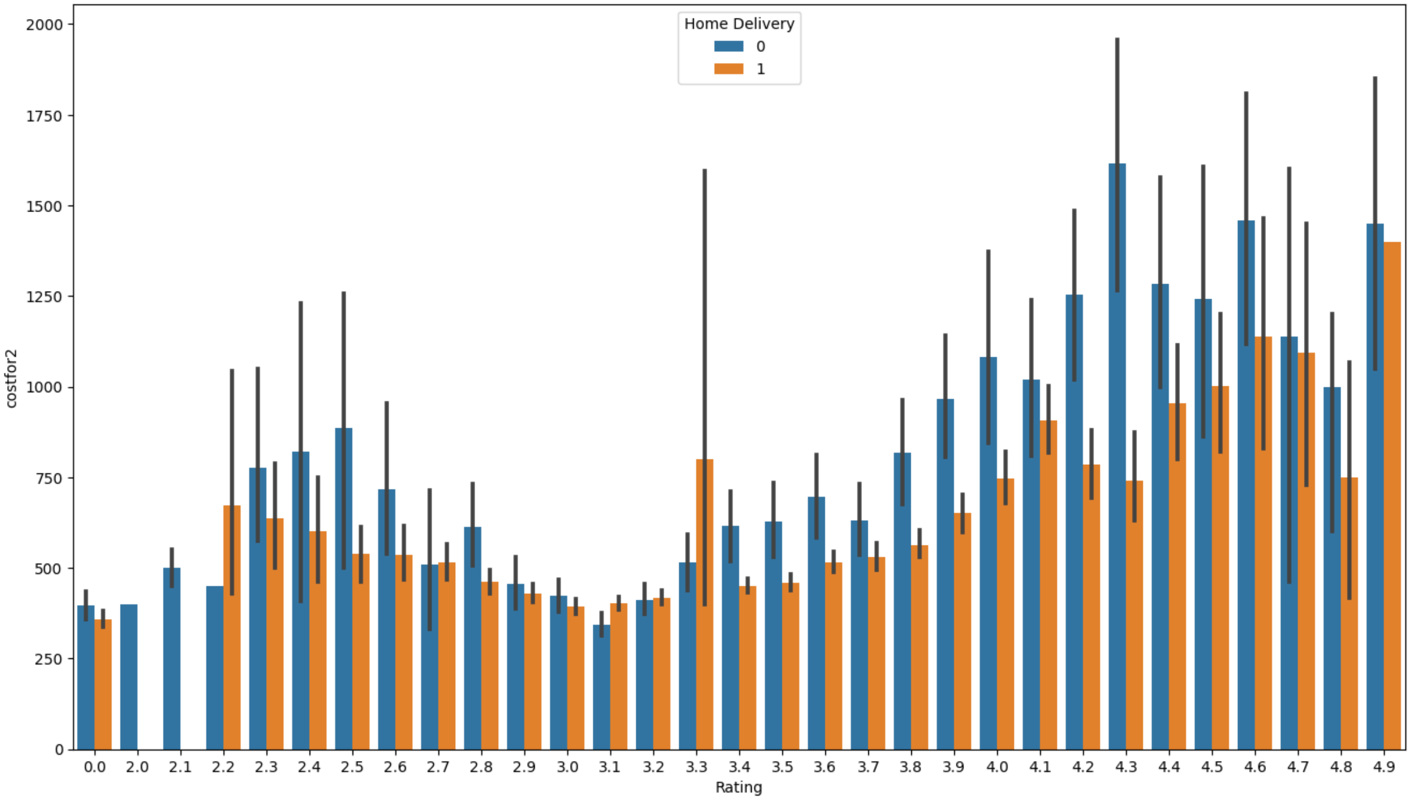
6)Location Wise count of restaurants of each type



From the above graph we can identify that if opening restaurant in the given locality then which type of restaurant we can open , we need to find the ‘type’ that doesn’t have higher number of restaurants open in the required locality, Like in case of Aundh we can choose to open , here the number of bakeries and casual dining bars are lesser , so we can choose either of these as our type of restaurant.



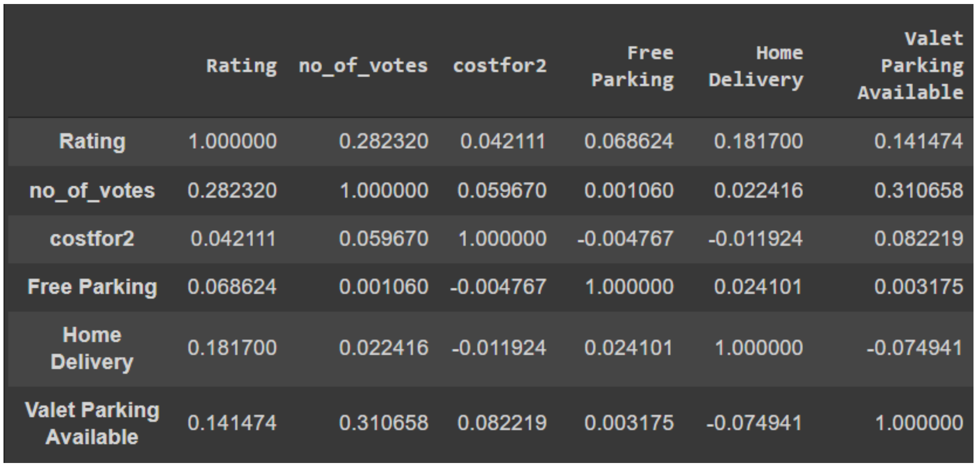
7)Cost for two v/s Rating in comparison with Home Delivery



From the above bar graph we can observe that the cost for two people is higher if they dine out in comparison with home delivered food and giving the same ratings . Most of the time people giving similar rating are paying less if the home delivery facility is available .

---

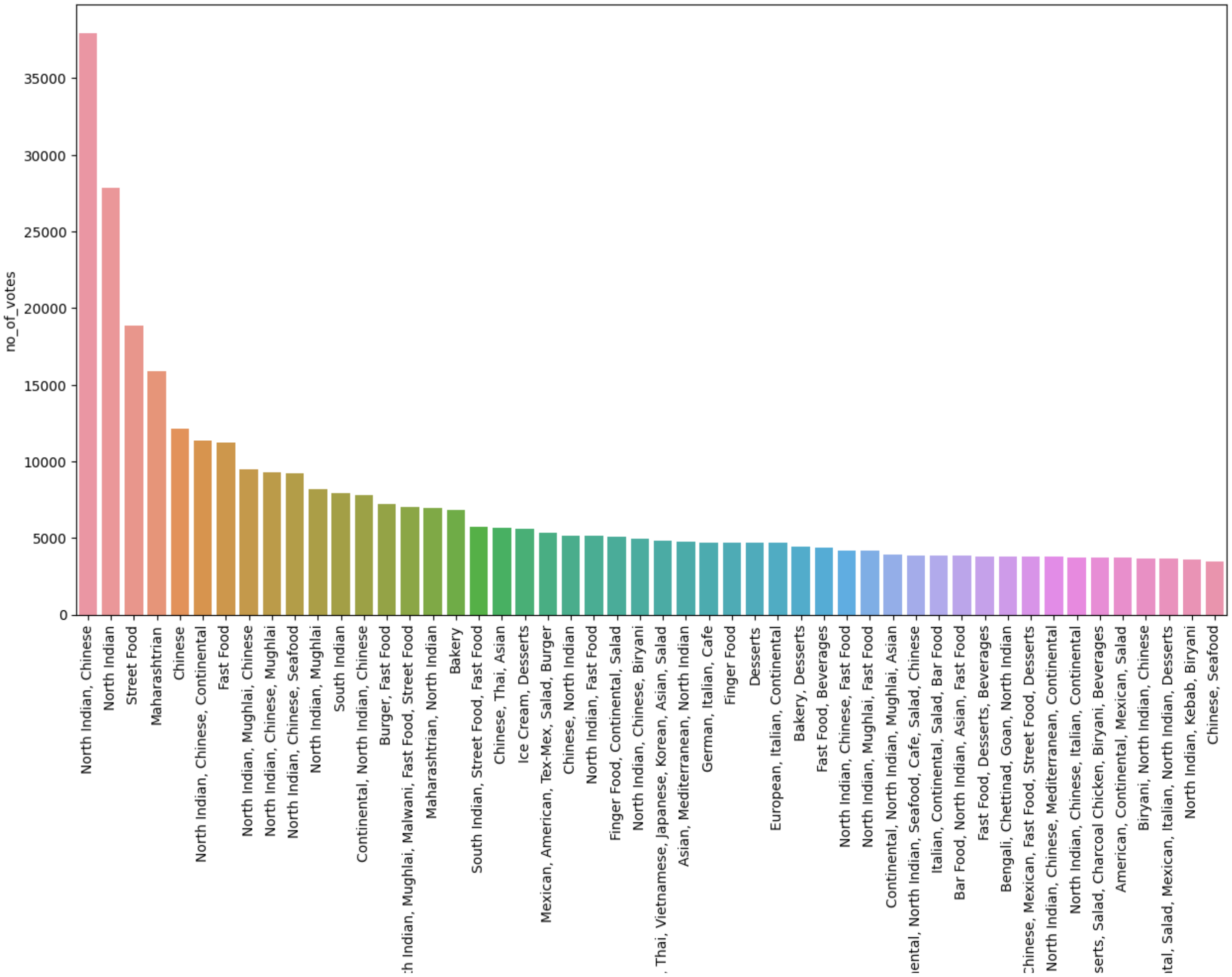
8)Correlation Table



	Rating	no_of_votes	costfor2	Free Parking	Home Delivery	Valet Parking Available
Rating	1.000000	0.282320	0.042111	0.068624	0.181700	0.141474
no_of_votes	0.282320	1.000000	0.059670	0.001060	0.022416	0.310658
costfor2	0.042111	0.059670	1.000000	-0.004767	-0.011924	0.082219
Free Parking	0.068624	0.001060	-0.004767	1.000000	0.024101	0.003175
Home Delivery	0.181700	0.022416	-0.011924	0.024101	1.000000	-0.074941
Valet Parking Available	0.141474	0.310658	0.082219	0.003175	-0.074941	1.000000

We can’t observe any strong correlation (value > 0.5) in our dataset.

8)No. of Votes v/s Cuisines



We can clearly see that the most popular cuisines are north indian and chinese , followed by street food and Maharahstrian food , so if opening a restaurant then choosing cuisine from these types of cuisines is a good choice.

Conclusion

Without a data mining process it's not possible to get so much insight into the datasets. Data mining helps in finding patterns and filtering them to get interesting patterns so that these patterns can be used to make better decisions for future jobs. It's flourished into all the sectors whether its education , health, business or sports , data mining is used in every domain as data is generated by them and as the data is generated then it's required to be mined to get beneficial results from it.

References

1. Data mining : concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed.
2. Data set source: https://www.kaggle.com/datasets/vedant8390rex/zomatoforproject232323
3. Library used: seaborn: statistical data visualization — seaborn documentation (pydata.org)
4. Image reference of figure 1 https://infovis-wiki.net/w/images/4/4d/Fayyad96kdd-process.png
5. https://www.ibm.com/cloud/learn/data-mining