CONTENTS

LIST OF FIGURES

CHAPTER 1

PROBLEM STATEMENT

The rapid growth of data collection has led to a new era of information. Data is being used to create more efficient systems and this is where Analysis and Prediction Systems come into practice. Prediction Systems are a type of information decision making systems as they provide clear understanding of characteristic of huge data sets. These systems use historical data and mathematical models to capture important trends.

Predictive modeling is then used on current data to predict what will happen next, or to suggest actions to take for optimal outcome. Almost every major tech company has applied them in some form or the other: Amazon uses it to suggest products to customers, YouTube uses it to decide which video to play next on auto-play, and Facebook uses it to recommend pages to like and people to follow. Moreover, companies like Netflix and Spotify depend highly on the effectiveness of their recommendation and prediction engines for their business and success.

CHAPTER 2

DATASET

The data is collected from a renowned website for data analytics named, Analytics Vidhya. This dataset contains the sales in four types of stores, Supermarket type 1, 2 and 3, and Grocery stores. The sales of these products depend on various factors and I have done some analyses to relate the sales to various factors. We can go for a bigger data-set as well but, since to process such large amounts of data; we will need a higher processing system. Therefore, we are working with a subset of the data, the extraction can equally be applied to larger chunks of data as well.

Following images will demonstrate the screenshot of the data-set to better facilitate our understanding:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Item_Iden	Item_Weig	Item_Fat_	Item_Visibility	Item_Type	Item_MRP	Outlet_Ide	Outlet_Est	Outlet_Siz	Outlet_Lo	Outlet_Type	Item_Outlet_
2	FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
3	DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
4	FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
5	FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
7	FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
8	FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
9	FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.764
10	FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.599
11	FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535
12	FDY07	11.8	Low Fat	0	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.027
13	FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.153
14	FDX32	15.1	Regular	0.1000135	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.265
15	FDS46	17.6	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.208
16	FDF32	16.35	Low Fat	0.0680243	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.426
17	FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.319
18	NCB42	11.8	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.889
19	FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
20	DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.668
21	FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket Type1	2748.422
22	FDN22	18.85	Regular	0.138190277	Snack Foods	250.8724	OUT013	1987	High	Tier 3	Supermarket Type1	3775.086
23	FDW12		Regular	0.035399923	Baking Goods	144.5444	OUT027	1985	Medium	Tier 3	Supermarket Type3	4064.043
24	NCB30	14.6	Low Fat	0.025698134	Household	196.5084	OUT035	2004	Small	Tier 2	Supermarket Type1	1587.267
25	FDC37		Low Fat	0.057556998	Baking Goods	107.6938	OUT019	1985	Small	Tier 1	Grocery Store	214.3876
26	FDR28	13.85	Regular	0.025896485	Frozen Foods	165.021	OUT046	1997	Small	Tier 1	Supermarket Type1	4078.025
27	NCD06	13	Low Fat	0.099887103	Household	45.906	OUT017	2007		Tier 2	Supermarket Type1	838.908
28	FDV10	7.645	Regular	0.066693437	Snack Foods	42.3112	OUT035	2004	Small	Tier 2	Supermarket Type1	1065.28
29	DRJ59	11.65	low fat	0.019356132	Hard Drinks	39.1164	OUT013	1987	High	Tier 3	Supermarket Type1	308.9312

Dataset Snap

CHAPTER 3

PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Steps Involved in Data Preprocessing:

1.Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

a) Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are: 1. Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple. 2. Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

b) Noisy Data: Noisy data is a meaningless data that can't be interpreted by ma- chines' can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

i. Binning Method: This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

ii. Regression: Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

iii. Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2.Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

a) Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

b) Attribute Selection: In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

c) Discretization: This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

d) Concept Hierarchy Generation: Here attributes are converted from level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3.Data Reduction: Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:

a) Data Cube Aggregation: Aggregation operation is applied to data for the construction of the data cube.

b) Attribute Subset Selection: The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.

c) Numerosity Reduction: This enable to store the model of data instead of whole data, for example: Regression Models.

d) Dimensionality Reduction: This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

The Big Mart sales data consists of 8523 rows and has 12 variables. The variables are described in the following table.

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

Dataset columns In the Big mart sales data following columns shoes missing values: “Item_Weight”, “Outlet_Size”. Following steps were taken to get rid of the missing values. For “Item_Weight” we replace the missing values with mean of the column and for “Outlet_Size” we replace the values with the mode of column.

Though this are the majority of ways in which data-preprocessing is carried out, the data-set which we have is already processed there are no missing values or other discrepancies present within it. Therefore, for our project we have opted out for this process.

3.1 Transformation

The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:

1. Smoothing: It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

   The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

2. Aggregation: Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.

   This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.

   The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies. For example, Sales, data may be aggregated to compute monthly and annual total amounts.

3. Discretization: It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes. Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values. Forexample, (1-10, 11-20) (age:- young, middle age, senior).

4. Attribute Construction: Where new attributes are created and applied to assist the mining process from the given set of attributes. This simplifies the original data and makes the mining more efficient.

5. Generalization: It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old). For example, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

6. Normalization: Data normalization involves converting all data variable into a given range.

3.2 Feature Engineering

Attribute “Item_Fat_Content” is a categorical attribute which had two categories:

Low Fat and Regular. However, the data had Low Fat, low fat, LF, reg and Regular which were then renamed Low Fat and Regular respectively.

Attribute “Outlet_Establishment_Year” did not had much intuitive meaning and hence it was replaced with how old the store is. This might help us determine better sales.

3.3 Recursive Feature Engineering

This algorithm is one of the popular methods for feature selection.
This method, ranks the importance of attributes and could help us determine which attributes should be eliminated.
This method creates subsets of data where each subset contains attributes number from 1 to n and desired algorithm is implemented.

Chapter 4

Mining Algorithm Used

4.1 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the liner regression between the explanatory (independent) variables and response (dependent) variable.

$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}+\epsilon$

y_i=dependent variable

x_i=explanatory variables

β_0=y-intercept (constant term)

β_p =slope coefficients for each explanatory variableϵ=the model's error.

The dataset contains factors like, the location of the outlet, type of outlet, visibility of the item in a store, weight of the product, the MRP of the product.

Model for predicting the future sales

Sales= a0 + a1*Item_MRP* + *a2*Item_Visibility + a3*Item_Establishment_Error + error

The linear regression for fitting the data is shown the same file.

4.2 Irrelevant Columns

Since the table is quite huge with large number of records. Mainly the main emphasis for each Filtering Algorithm differs based on the intention of Filtering.

For Example, columns like “Item type”, “Item sales” and “Out late location” are given most important. These fields are given the most importance data prediction and analysis.

Whereas in fields like “Outlet type” and “Outlet Establishment year” are given less importance as these columns don’t represent useful information.

Therefore there is not a straightforward answer for this section. Depending on the type of manipulation you do, the priority of the columns can change. It’s really upto the user depending on what kind of relevance or output he expects from the Data-set.

CHAPTER 5

TESTING DATA WITH UNKNOWN TUPLES

This is often a good step to find out hidden information from your Data-set. Since a data can reveal a lot of information is usually missed by people. In this way we might land up on getting the information, which we usually would not have thought of, but it is not necessary that we will always end up with some magical information.

Most of the times it happens that, we get crappy or unwanted information which is of no use whatsoever. This way of testing is also a good way of testing the accuracy of whatever model you have built so far.

If the system, responds in similar fashion, there is no need to worry you have written and algorithm that is consistent through-out. But it is not that easy to have a deterministic and consistent systems in real-life. Problems are so complex that often you will land-up with a piece of data that makes no sense. So, we need to be careful and take proper steps whenever we follow this in order to get good results.

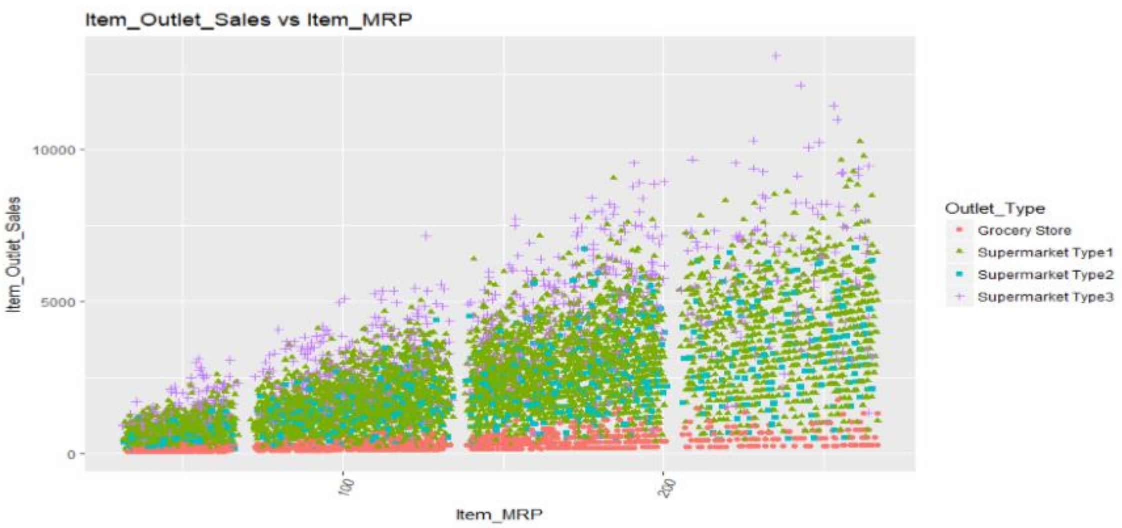
CHAPTER 6

DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

1. Item outlet sales vs Item MRP
   It can be seen in graph that in sales Supermarket Type 1 dominates compared to other type of outlets. However, it is interesting to see the gaps in the prices around 60,130 and 200. There could be many reasons for why there are gaps in the prices, it could be because prices for different categories differ and which led to the gaps.

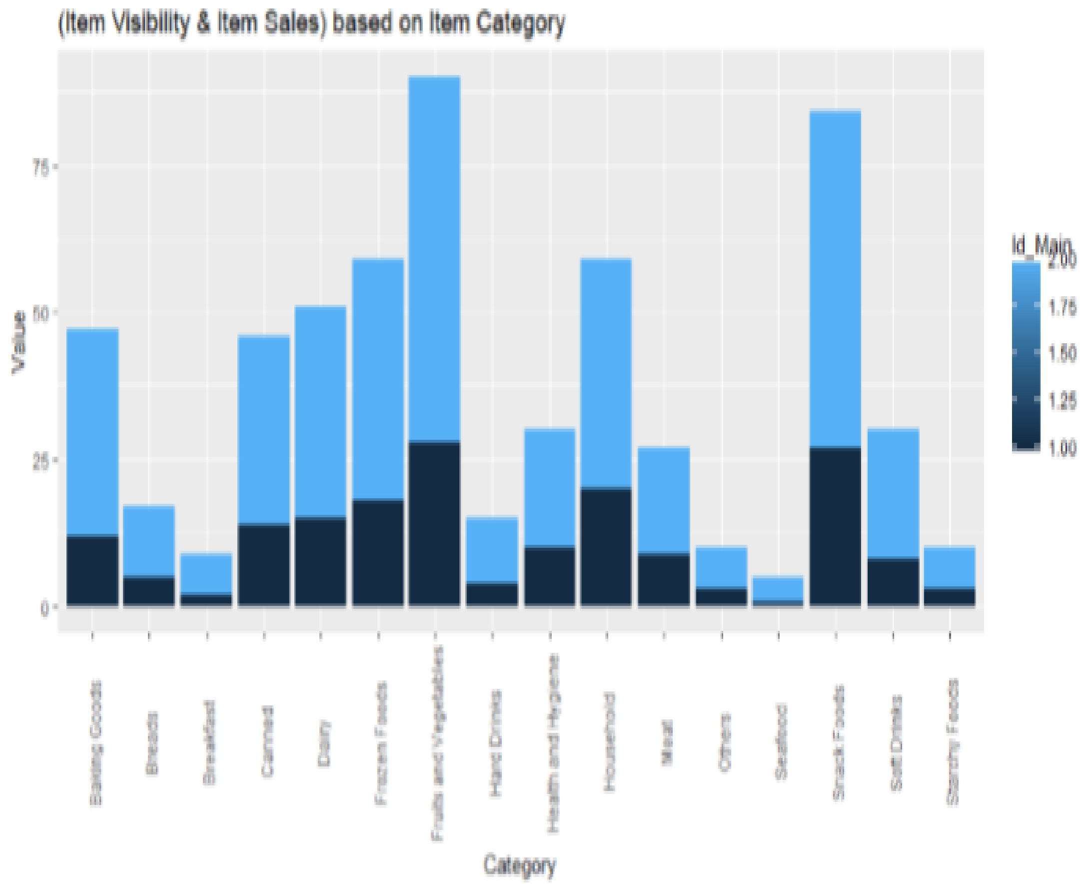


Item outlet sales vs Item MRP

2.Item Visibility vs Item Type:

Visibility and for an item category is directly proportional to aggregate sales for that particular category.

Those categories, who had the highest visibility had the highest sales.



Item Visibility vs Item Type

3.Item visibility vs Item Out late sales:

Item visibility vs Item Out late sales

CHAPTER 7

CONCLUSION

The project was done to facilitate our understanding of handling large chunks of data and applying appropriate visuals, so that the process of decision making is smooth. The set of algorithms which we have used are already present and available in market. Our intention was to test our understanding and application of that knowledge, on a smaller data-set. Going ahead we would like to apply same algorithms to a larger data-set and see, what results we get. One thing is for sure, to apply such processing to very large data-sets, python alone would not be sufficient. Therefore, in future we need to make sure that we make use of Big Data's Hadoop Ecosystem to facilitate our needs as they are much faster and suitable when dealing with very large chunks of data that needs to be processed in a parallel and efficient manner.