A
PROJECT REPORT
ON

STUDY OF SUPERMARKET SALES PREDICTION

Submitted to

Savitribai Phule Pune University

In Partial Fulfillment of the Requirement for
Award of the Degree of

MASTER OF BUSINESS ADMINISTRATION

Submitted By

KIRAN SUNDARLAL UIKEY
MASTER OF BUSINESS ADMINISTRATOR
(BUSINESS ANALITICS)

Under the Guidance of

Prof. YOGITA KADBANE

SINHGAD INSTITUTE OF MANAGEMENT

(Academic **year:2021-2023)**

DECLARATION

I **Kiran Sundarlal Uikey**, student of MBA from Sinhgad Institute of Management, Pune hereby declares that the project on "**Study of supermarket sales prediction"** is submitted to Savitribai Phule Pune University is the record of work done by me under the guidance of Mrs.Yogita Kadban**e** in the partial fulfilment for the requirement of the award of degree of MASTER OF BUSINESS ADMINISTRATION.

This is the original work and has not been submitted to any other Institution for any other degree/diploma/certificate in this University or any other University.

Kiran S.Uikey

Place: Pune

Date :

Collage certificate

MENTORED MINDS

Internship Completed at Mentored Minds

This is to certify that **KIRAN UIKEY** Has Completed 2 Months of Internshipat Mentored Minds Pvt Ltd For the Role of **Data Analytics Intern**

During this internship the tasks given were performed with best efforts and the work wassincerely done to deliver the work packets.

We wish you all the best and you will have a wonderful Career ahead!

Harshal Mody
Contact - +919923342240
Mentored Minds
Pune, Maharashtra, India - 411038

Mentored Minds Private Limited
Director

CONTENTS

ACKNOWLEDGEMENT

I take this opportunity and privilege to express my deep sense of gratitude to Professor **M.N. Navale**, Honorable Founder Presidents **, Dr. (Mrs.) Sunanda M. Navale,** Founder Secretary, The Sinhgad Technical Education Society**,** Pune and **Dr.Daniel Penkar**, Director SIOM. They have been a source of inspiration to me and I am indebted to them for initiation me in the field of research.

I take this opportunity of submitting this report to express my regards towards those who offered their individual guidance in the hour of need.

I am deeply indebted to Faculty Member, SIOM **Prof. Yogita Kadbane** , my research guide at Sinhgad Institute of Management, Pune, without whose help completion of the project was highly impossible.

He gave knowledgeable insights about the topic, which helped me a lot. I also would like to thank **MR. Harshal Mody** CEO and Founder Mentored Minds for giving me this opportunity to work on this topic; it surely has given me insights into areas I was not much familiar earlier. I have tried to share those insights with you in this report.

I wish to express a special thanks to all teaching and non-teaching staff members of sinhgad Institute of Management, Pune for their continuous support. I would like to acknowledge all my family member, relatives and their help and encouragement.

Kiran S. Uikey

Place: Sinhgad institute of management

Date:

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

During my Two months of summer internship training in Pune at mentored minds. I have done my project work on "**Study of supermarket sales prediction**"
Mentored Minds is Leading platform for students and work professional can access and work on situations in the form of internship.

The internship started from 09-Nov-2022 and was completed on 09 Januray 2022. An executive summary for a report on supermarket sales prediction would provide a brief overview of the key findings and conclusions of the report. It would summarize the methods used to make the prediction, such as the types of data that were analyzed and the specific algorithms or statistical techniques that were employed.

It would also provide an overview of the results of the prediction, such as the forecasted sales figures and any key insights or recommendations for the supermarket management. In summary, an executive summary would give a quick and brief information about the report content without going into the details of the report.

The supermarket industry is highly competitive, with retailers constantly seeking ways to improve sales and increase profits. Predicting sales is a critical aspect of this effort, as it allows supermarkets to optimize inventory, staffing, and pricing decisions. In this executive summary, we present an overview of a study on supermarket sales prediction.

Overall, the study demonstrates the potential of machine learning techniques in predicting supermarket sales. By incorporating a range of variables and identifying the most significant factors affecting sales, the model can help supermarkets optimize inventory, staffing, and pricing decisions, leading to increased profitability and competitiveness.

CHAPTER 1

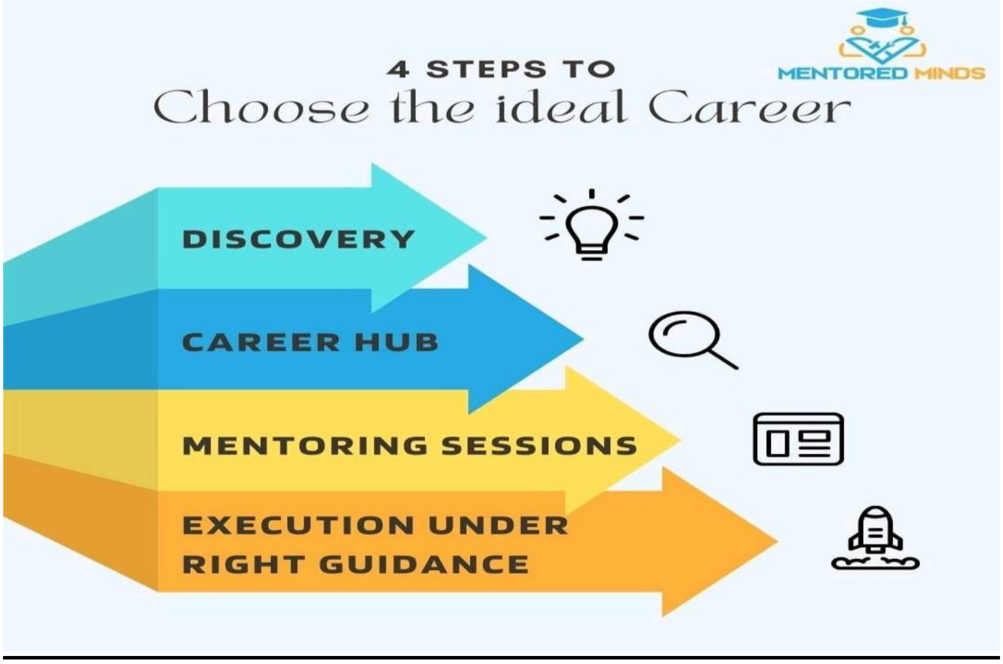
INTRODUCTION

INTRODUCTION

Sales analysis allows you to better understand your customers, the products they are buying and the reasoning behind this behaviour. In doing so, it becomes much easier to highlight your most profitable customers, and keeping these customers engaged with your business can be the key to increasing overall profitability.

A supermarket is self-service shop offering a wide variety of food, beverages and household products, organized into sections. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market.

The growth of supermarkets in most populated cities are increasing and market competitions are also high. Hence for our project we chose the supermarket sales dataset. In this project we have used different techniques to analyses the sales data set of supermarket such as Data Visualization, Hypothesis Testing, and Regression techniques such as Multiple Linear Regression, Ridge Regression, Support Vector Machine, Random Forest classifier for prediction of Total sales.



Guided discussions are designed to facilitate learning and improve knowledge and skills. The facilitator asks learners questions to stimulate and guide refection and critical thinking. These discussions usually complement other methods, such as a presentation, research or a case-based exercise.

Guided discussions also facilitate communication and knowledge sharing among learners. For example, after conducting individual research on food security information systems, learners may be asked to describe to the facilitator and the other learners how those systems work in their own countries.

Learners work together to perform different types of activity, such as evaluation, analysis or development of an assignment or a project. This method requires learners to collaborate, listen to each other, argue and negotiate; they develop interpersonal skills other than domain-specific and problem-solving skills. For example, learners may be divided into small groups and tasked with evaluating the impact of a food security program by applying the principles learned during the course. Each group must provide an evaluation report as an outcome of the assignment.

Mentored minds is an e-learning company that provides datasets of various companies and provide them technological solutions and insights to achieve specific objective.

What was once a simple way of doing business is transforming into a highly sophisticated form of management and marketing. Retail marketing consistently features more efficient, more meaningful and more profitable marketing practices. Retail is the accumulation of various marketing practices directed towards providing the best merchandise available. It consists of the sale of goods or merchandise, from a fixed location such as a big department store or a small store, in small or individual lots for direct consumption by the purchaser. Retailing may include subordinated services, such as delivery. A retailer buys goods or products in large quantities from manufacturers or importers, either directly or through a wholesaler, and then sells smaller quantities to the end-user i.e., the consumer or the end buyer. In the supply chain, retailers come at the end, just before the consumer

Manufacturer wholesaler Retailer consumers Retailing includes all activities involved in selling goods or services directly to final consumers for personal non-business use. A retailer or retail store is any business enterprise whose sales volume comes primarily from retailing. Retail is Indian’s largest industry, accounting for over 10 percentages of the country’s GDP and around eight percent of the employment. Retail industry in India is at the cross roads. It has emerged as one of the most dynamic and fast paced industries with several players entering the market.

OBJECTIVES

The following objectives are selected for the present study:

- To predict the total sale of Supermarket.
- To visualize how explanatory variables such as Branch, Customer type, Total sale, Rating, Payment type affect to study variable sales.
- To check the Independency of two attributes.
- To fit appropriate model for prediction of total sales.
- To know the innovative retail marketing practices adopted by Supermarket in India.
- To focus on organizational structure, departmental hierarchy, management techniques, marketing strategies of Supermarket.
- To know the role and importance and success of Supermarket in retail marketing in India and Maharashtra.
- To understand and identify the Customer Relationship Management Practices followed by the Supermarket
- To focus on the future plans and various marketing strategies of Supermarket

SCOPE

- Build a predictive model and find out the sales of each of the products at a particular store
- The Big Mart can use this model to understand the properties of the products which plays a key role in increasing the sales
- The project will be help in the increase the sales.
- Data collection is helpful to the future planning for the warehouse.
- Data Quality: The accuracy and completeness of historical sales data are critical for building accurate predictive models. If the data is incomplete or inaccurate, the predictions may be unreliable

LITERATURE REVIEW

This chapter contains a brief overview of previous work related to the problem this thesis was investigating. The main objective of this section was to understand the current depth in this field, the amount of academic research existing, how that research has been executed, and where there exist possible gaps in the literature. Furthermore, a secondary objective was to dig deeper into the existing research to conclude which algorithms, features, and evaluation metrics that appear frequently throughout the literature. In India, market reform and opening to FDI, along with prospects for 7% yearly growth

In retail sales in a market of 1.2 billion people, have generated billions of dollars of planned investment in supermarkets by local and multi-national firms, including Walmart and carrefour. Yet supermarket shares in India are currently very low (around 2%), due to the country's massive and complex small retail sector. Supermarket there face the 20/20/20 challenge: they must grow their food sales by 20% market share. Such unprecedented growth would still leave more traditional channels holding 80% of food market.

DOMESTIC AND REGIONAL MARKETS AS A FOCUS OF GROWTH:

Non-traditional agricultural exports have received large amounts of analytical attention over the past decades. Donor support to market oriented agriculture for smallholder farmers has also focused heavily on export markets, while "domestic food markets remain undercapitalized, risky, rudimentary, and relatively thin" (World Bank, 2007).

K. Bhaskar et al**.** The Indian Retail Industry is gradually inching its way towards becoming the next boom industry. Today, organized retail operations, chain stores and international investment are starting to move in, leading at least part of the retail sector to dramatically increase its scale of operations and integrate itself more closely into the international economy, potentially reducing farm-to-market losses of agricultural products, encouraging infrastructure improvement, and driving the training of the middle segments of the labor force. The India Retail Industry is the largest among all the industries accounting for over 10 percent of the country's GDP and around 8 percent of the employment. The Retail Industry in India has come forth as one of the most dynamic and fast paced industries with several players entering the market. But all of them have not yet tasted success because of the heavy initial investments that are required to break even with other companies and compete with them

Studies show that machine learning models are useful in the retail industry to gain knowledge of a business like a supermarket. Hence, analysis of the products bought by the customers is widely done, especially in the food industry. The most widely used methods for analysis described in the literature survey are SVM, MLP and RBFN. Less work has been observed to be done using the latest algorithms like VAR(Vector Autoregression) and SARIMA(Seasonal

Auto Regressive Integrated Moving Average) which cover the seasonality features. The seasonality feature plays a vital role to predict the sales and trends of the data. Example:

Seasonality feature refers to the sales of a product season wise for instance in Diwali season a lot of decoration, lighting etc would be bought by the people and the sales trends will change likewise. Similarly, in summer season mangoes are frequently bought by people and hence sales for mangoes would go up in summers but the Diwali decoration won't be of much use. In a framework has been described which is used to predict the obesity factor in customers from the grocery data. The paper is divided into two parts where the first part explains deriving dietary intake patterns from the grocery data. The second part includes making predictions related to the former part using suitable data mining tools. There is a limitation of this work that it is performed in a small sample size of self-selected households.

Supermarkets are an important aspect of modern life and have a significant impact on the economy. Accurate sales forecasting is essential for supermarkets to ensure that they have enough stock on hand to meet customer demand while minimizing waste and reducing costs. In this literature review, we will explore various methods and techniques used in supermarket sales forecasting.

Supermarkets are a crucial part of the retail industry, and sales prediction plays a critical role in optimizing their operations. Sales prediction enables supermarkets to manage their inventory, plan promotions, and forecast future revenue accurately. In this literature review, we explore various techniques and methods used for supermarket sales prediction.

Supermarket sales have been a popular research topic for many years, and a vast body of literature exists on this topic. This literature review will provide an overview of some of the key findings from this research.

One of the main factors affecting supermarket sales is pricing. Several studies have found that consumers are sensitive to prices and will switch to cheaper alternatives if prices increase. In addition, promotions and discounts can be effective in increasing sales, particularly for products with high profit margins. Furthermore, pricing strategies such as dynamic pricing, where prices change based on demand, have been found to be effective in increasing sales.

Another important factor affecting supermarket sales is the layout of the store. Research has shown that consumers tend to shop in a predictable pattern, starting from the right-hand side of the store and moving in a counterclockwise direction. Therefore, retailers can strategically place high-profit items in these areas to increase sales. Moreover, retailers can use signage and displays to draw attention to certain products, and can create themed areas to increase interest and promote sales.

The availability and quality of products is also important in driving supermarket sales. Customers expect a wide variety of products to choose from, including fresh produce, meat, and dairy. In addition, customers expect high-quality products, and are willing to pay a premium for them. Therefore, retailers must carefully manage their inventory to ensure that they have a wide variety of products available, and that the products are fresh and of high quality.

Customer service is another important factor that can affect supermarket sales. Customers expect friendly and helpful staff, and retailers that provide good customer service are more likely to build customer loyalty and increase sales. Furthermore, retailers can use technology such as self-checkout and mobile apps to provide a convenient shopping experience for customers, which can also increase sales.

In summary, several factors affect supermarket sales, including pricing, store layout, product availability and quality, and customer service. Retailers that effectively manage these factors are more likely to increase sales and build customer loyalty.

- Time series analysis

  Time series analysis is one of the most commonly used methods for sales prediction in supermarkets. It involves analyzing historical sales data to identify trends and patterns. Time series analysis can be used to predict future sales based on past sales trends.

- Machine learning

  Machine learning techniques have also been used for sales prediction in supermarkets. Machine learning algorithms can learn from historical sales data and make predictions based on the learned patterns. Techniques such as Random Forest, Support Vector Machines, and Artificial Neural Networks have been used for sales prediction in supermarkets.

- Deep learning

  Deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have also been used for sales prediction in supermarkets. These techniques can learn from large datasets and make accurate predictions based on the learned patterns. Deep learning models have been found to be effective in predicting sales for supermarkets.

- Hybrid learning

  Hybrid methods involve combining multiple techniques to improve the accuracy of predictions. For example, a hybrid approach that combines time series analysis and machine learning has been used for sales prediction in supermarkets. Hybrid methods can leverage the strengths of different techniques and improve the accuracy of predictions

CHAPTER 2

ORGANISATIONAL PROFILE

COMPANY PROFILE

MENTORED MINDS PVT.LTD.

Mentored Minds is a platform where candidates can seek internships through virtual or real mode in a variety of domain to seek industry level training, Perform practical real-world business problems, seek career guidance using Career Explorer test & the Career Hub, Build their skills and experience on performing analysis on different platforms and business tools. Candidates earn Certificate for their Internships and gain a lot of practical business exposure.

Mentored Minds provide platform required to produce and deliver e-learning. Digital tools and technologies are used in a variety of ways to support learning, teaching and assessment. A mix of digital learning tools, devices, platforms and applications is making learning more flexible and convenient.

The tutor or the instructor can task learners with conducting research on a specific subject. The instructor can guide the learner in collecting and organizing information (guided research). For example, learners may be asked to conduct research on the food security information systems (FSIS) in their own countries. The instructor provides suggestions to learners on how to find the required information and how to illustrate the FSIS using a diagrams

MENTORED MINDS

CEO: Harshal Mody

Office address: Kothrud, Pune

OFFICIAL Website:

www.mentoredminds.com

Email ID:

operations@mentoredminds.com

Linked In:

https://www.linkedin.com/company/mentored-minds/

BLOGS:

https://www.mentoredminds.com/blog

HEADQUATER:
Headquarted
in Pune, Maharashtra India – 411051

CONTACT: **9370295631**

VISION AND MISSION:

Vision:

To become the leading provider of accurate and reliable supermarket sale predictions, using cutting-edge technology and data-driven approaches to help our clients make informed business decisions.

Mission:

Our mission is to leverage the latest developments in machine learning and predictive analytics to provide our clients with highly accurate forecasts of supermarket sales. We aim to offer a comprehensive solution that incorporates both historical sales data and real-time market trends to deliver actionable insights that help our clients optimize their supply chain, improve inventory management, and maximize profitability. We strive to maintain the highest standards of professionalism, transparency, and ethics in all our interactions with clients, stakeholders, and partners.

Products

Health & Beauty

For all your beauty requirements:

- Oral care – toothpastes, toothbrushes, mouth washes
- Hair care – shampoos, conditioners and lotions
- Toiletries – soaps, hand washes and sanitizers
- Baby care – diapers, toiletries, accessories
- Skin care – lotions, face washes, scrubs, special needs
- Deodorants and perfumes
- Feminine care – napkins and tampons
- Mens grooming – shaving creams, gels, blades, face washes,

Over the counter items – antiseptics, ear buds, contraceptives, supplements, sprays and ointments

Electronic accessories

Electronic accessories are products that are designed to complement and enhance the functionality of electronic devices. They can include a wide range of items, such as chargers, cases, screen protectors, cables, adapters, headphones, speakers, and more. Some common electronic accessories include:

- Phone cases: These are designed to protect your phone from scratches, dust, and other damage. They come in various materials, such as plastic, silicone, leather, and metal.
- Power banks: These are portable batteries that allow you to charge your electronic devices on the go. They come in different sizes and capacities
- Screen protectors: These are thin films that you can place over the screen of your device to protect it from scratches and cracks.
- Headphones: These can be wired or wireless and are used for listening to music, watching videos, or making calls.
- Cables and adapters: These are used to connect your electronic devices to other devices or to charge them. They can include USB cables, HDMI cables, lightning cables, and more.
- Speakers: These are used to enhance the sound quality of your electronic devices, such as phones, tablets, and laptops.
- Smartwatches: These are wearable devices that can be used to track fitness, receive notifications, and control other devices.

Home and lifestyle

Overall, electronic accessories can enhance the functionality and usability of your electronic devices, making them more convenient and enjoyable to use.

Home and lifestyle products refer to a range of items that are used in and around the home to improve comfort, convenience, and functionality. These products can include furniture, appliances, decor, and other household items. Some common home and lifestyle products are:

- Furniture: This includes items such as sofas, chairs, beds, and tables, which are used to make your home comfortable and functional.

- Appliances: This includes items such as refrigerators, washing machines, and ovens, which are used to make household chores easier.

- Home decor: This includes items such as rugs, curtains, and artwork, which are used to decorate your home and create a pleasant living space.

- Lighting: This includes items such as lamps, light bulbs, and fixtures, which are used to provide light and create a specific atmosphere in your home.

- Kitchenware: This includes items such as pots, pans, and utensils, which are used to prepare and serve meals.

Sports and travel

- Luggage: This includes items such as suitcases, backpacks, and duffel bags, which are used to store and transport your belongings while traveling.

- Sports equipment: This includes items such as balls, bats, and rackets, which are used for various sports and outdoor activities.

- Travel accessories: This includes items such as travel pillows, eye masks, and earplugs, which can make long journeys more comfortable.

- Camping gear: This includes items such as tents, sleeping bags, and camping stoves, which are used for camping and other outdoor activities.

- Athletic clothing: This includes clothing designed specifically for sports and outdoor activities, such as running shorts, hiking pants, and swimwear.

DATA DESCRIPTION

We have used secondary data. The data of supermarket sales were collected from Kaggle website. The dataset is one of the historical sales of Supermarket Company which has recorded in 3 different branches for 3 months data. The data were collected from 1January 2019 to 30 March 2019.

Supermarket_sales.csv

This dataset contains 1000 records with 17 different variables out of these response variable is Total sale. In the dataset individual branch distribution being 340 Branch A (Mumbai) records, 328 Branch B (Delhi) records, and 332 Branch C (Pune) records over 3 months from January-March 2019. The customer type and gender are equally distributed with 501 Member and 499 Normal customer records, 499 Male and 501 Female records. The product line has a wide distribution with 170 Electronic accessories, 178 Fashion accessories, 174 Food and beverages, 152 Health and beauty, 160 Home and lifestyle, and 166 Sports and travel records. The payment method is distributed into 3 categories with 345 e-wallet, 344 cash, and 311 credit card records.

Variable	Information	Variable Type
Invoice id	Computer generated sales slip invoice identification number	Numerical
Branch	Branch of supermarket	Categorical
City	Location of supermarkets	Categorical
Customer type	Type of customers, recorded by Members for customers using member card and Normal for without member card.	Categorical
Gender	Gender type of customer	Categorical
Product line	General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel	Categorical
Unit price	Price of each product in $	Numerical
Quantity	Number of products purchased by customer	Numerical
Tax	5% tax fee for customer buying	Numerical
Total	Total price including tax	Numerical
Date	Date of purchase	Numerical
Time	Purchase time (10am to 9pm)	Numerical
Payment	Payment used by customer for purchase	Categorical
COGS	Cost of goods sold	Numerical
Gross margin percentage	Gross margin percentage	Numerical
Gross income	Gross income	Numerical
Rating	Customer stratification rating on their overall shopping experience	Numerical

CHAPTER 3

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

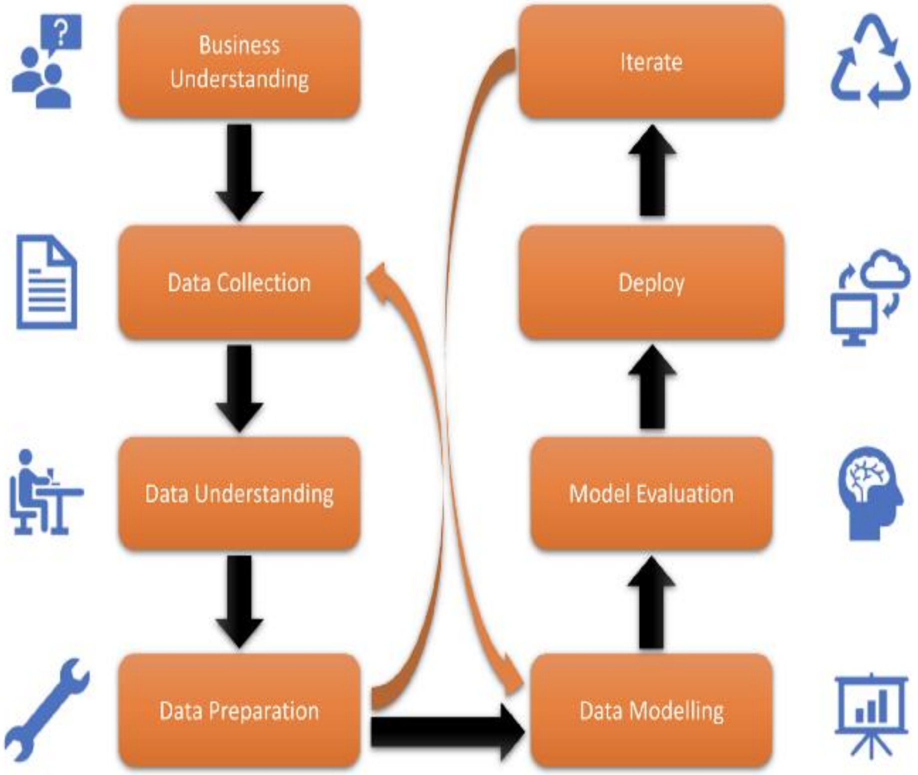


Fig.1.1

Goal of this project is to predict total sales of the supermarket which helps the seller to decide different discount offers and to make supermarket shopping paradise for the buyers and a marketing solutions for the sellers as well.
We have used secondary data. The data of supermarket sales were collected from Kaggle website. The dataset is one of the historical sales of Supermarket Company which has recorded in 3 different branches for 3 months data. The data were collected from 1January 2019 to 30 March 2019.

This dataset contains 1000 records with 17 different variables out of these response variable is Total sale. In the dataset individual branch distribution being 340 Branch A (Mumbai) records, 328 Branch B (Delhi) records, and 332 Branch C (Pune) records over 3 months from January-March 2019. The customer type and gender are equally distributed with 501 Member and 499 Normal customer records, 499 Male and 501 Female records. The product line has a wide distribution with 170 Electronic accessories, 178 Fashion accessories, 174 Food and beverages, 152 Health and beauty, 160 Home and lifestyle, and 166 Sports and travel records. The payment method is distributed into 3 categories with 345 e-wallet, 344 cash, and 311 credit card records.

DATA PREPROCESSING

Data Preprocessing is a technique that used to improve the quality of the data before analysis, so that data will lead to high quality results. Data preprocessing include data cleaning, data integration, data transformation, and data reduction.

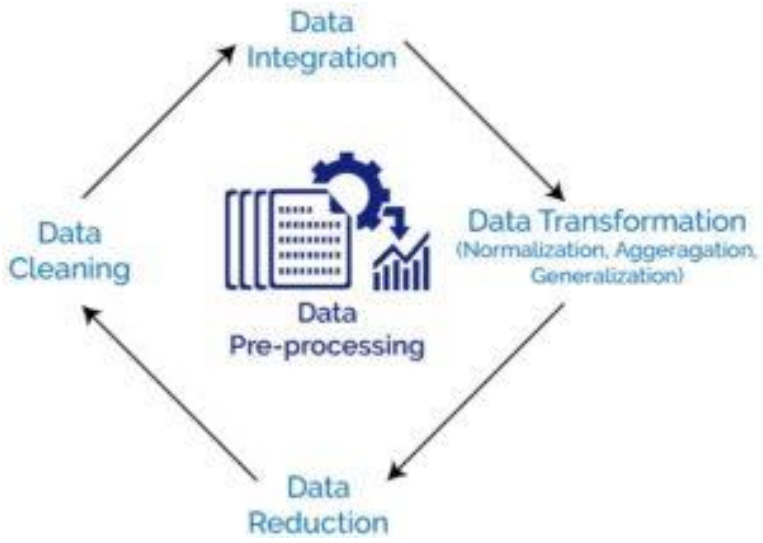


Fig.1.2

Data Cleaning: Data cleaning can be applied to remove noise and correct inconsistencies in the data.

Data integration: Data integration merges data from multiple sources in to a coherent data store, such as a data warehouse.

Data transformations: Data transformations such as normalization, may be applied for example, normalization may improve the accuracy and efficiency of algorithms involving distance measurements.

Data reduction: Data Reduction can reduce the data size by aggregating, eliminating redundant features, for instance.
These techniques are not mutually exclusive. They may work together.

Need of data preprocessing incomplete, noisy and inconsistent data are common place properties of large real world database and data warehouse.

Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data what where inconsistent with other recorded data may have been deleted. Furthermore recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing value for some

mining results. Therefore to improve the quality of data and, consequently, of the mining results, data preprocessing needed.

This dataset contains 1000 records with 17 variables. There are no missing values in the data. The Invoice ID column of data does not provide any information for analysis so we have dropped that column. Gross margin percentage is constant for all records so we have canceled that column. To check the outliers we use Boxplot technique.

Checking the Outliers:

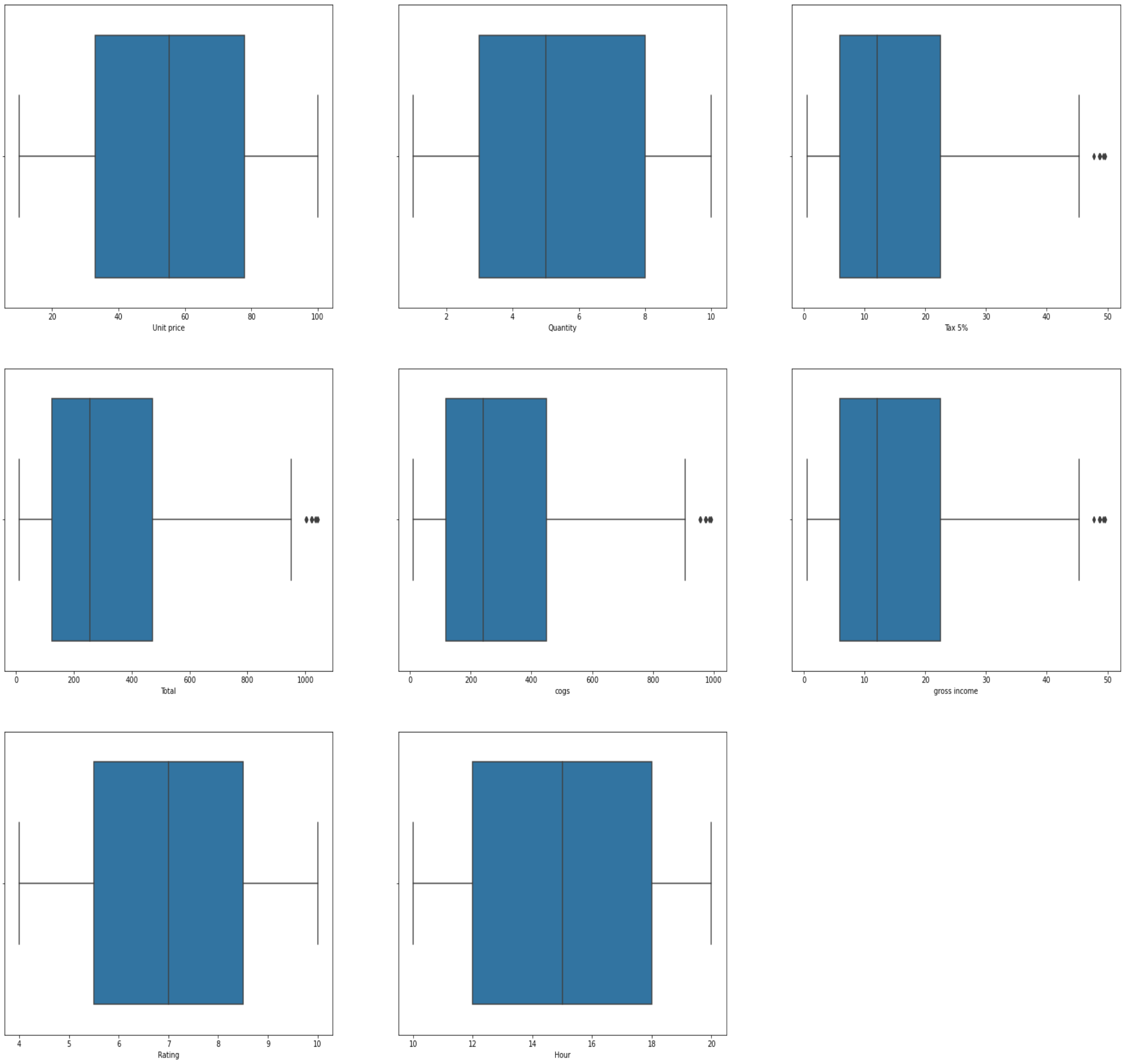


Fig.1.3

From the above boxplots, we can see that Tax 5%, cogs, Total and gross income have outliers. So now we have been remove outlier. Regarding to an attribute, if a value of it is out size 1.5 times IQR from mean, it is treated as outliers.

Normalisation of Data by removing Outliers:

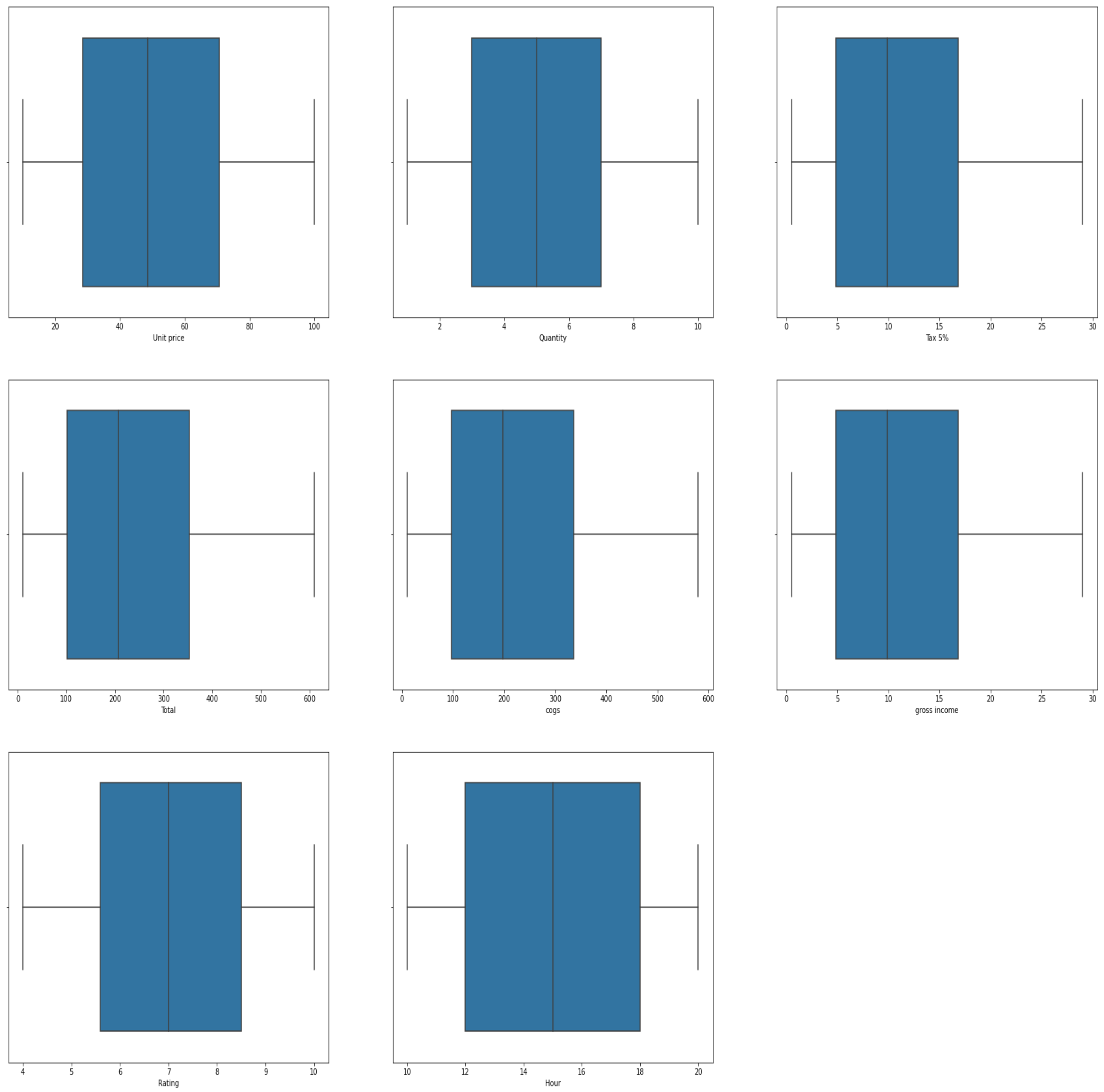


Fig.1.4

From above Boxplots we can clearly see that all outliers are remove and the data become Normal.

Pearson’s Correlation Heatmap:

By using pearson’s correlation, we can see the "linear" correlation between two attributes. By this way, we can remove the attribute that cause multicollinearity which is bad for modelling.

Python code:

```
corr = df.corr(method='pearson')
plt.figure(figsize=(15,12))
sns.heatmap(corr,annot=True,cmap='YlGnBu')
plt.show()
```

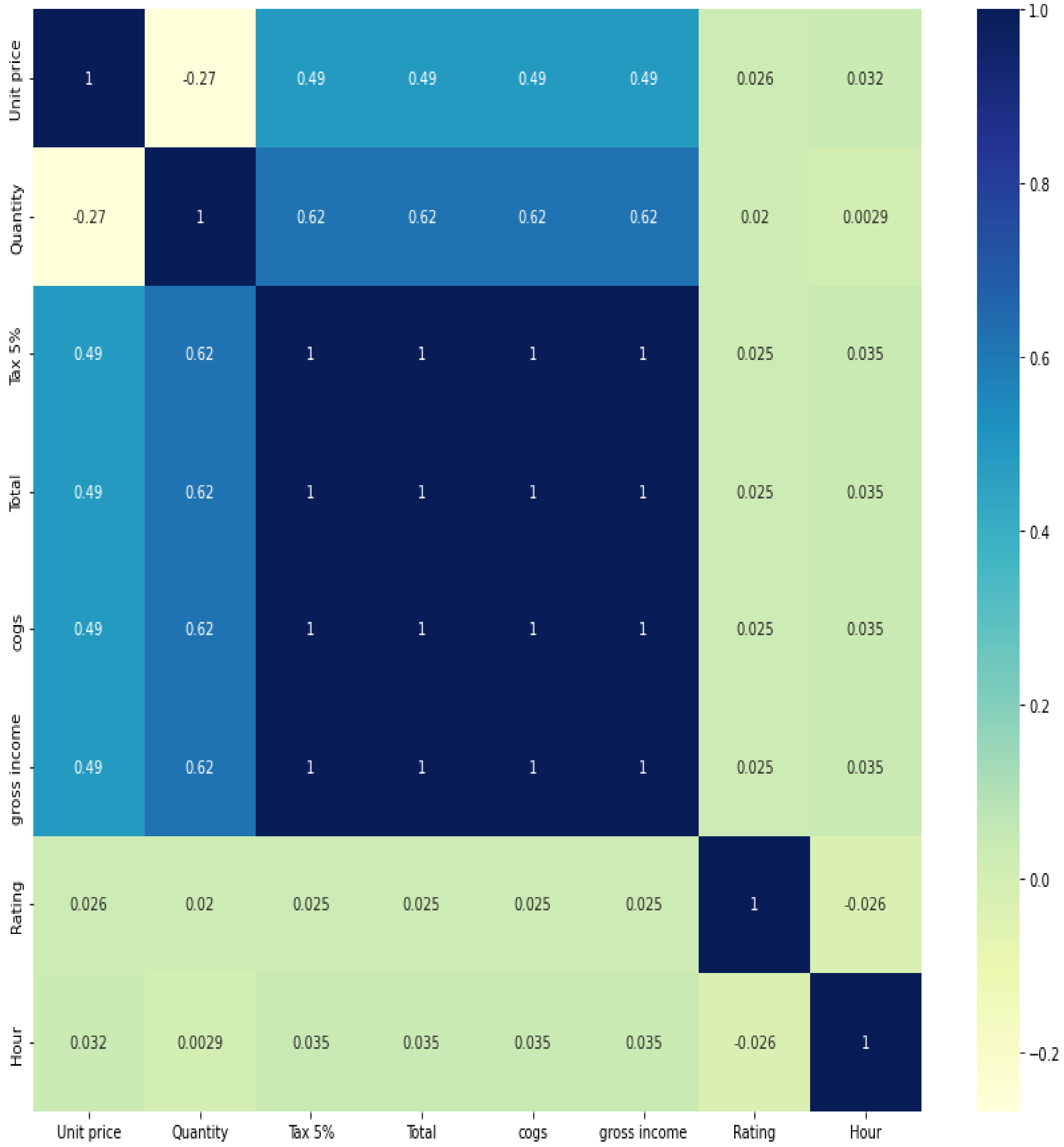


Fig.1.5

Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

You’re living in an era of large amounts of data, powerful computers, and artificial intelligence. This is just the beginning. Data science and machine learning are driving image recognition, development of autonomous vehicles, decisions in the financial and energy sectors, advances in medicine, the rise of social networks, and more. Linear regression is an important part of this.

Linear regression is one of the fundamental statistical and machine learning techniques. Whether you want to do statistics, machine learning, or scientific computing, there’s a good chance that you’ll need it. It’s best to build a solid foundation first and then proceed toward more complex methods.

By the end of this article, you’ll have learned:

- What linear regression is

Regression searches for relationships among variables. For example, you can observe several employees of some company and try to understand how their salaries depend on their features, such as experience, education level, role, city of employment, and so on.

This is a regression problem where data related to each employee represents one observation. The presumption is that the experience, education, role, and city are the independent features, while the salary depends on them.

- What linear regression is used for

Typically, you need regression to answer whether and how some phenomenon influences the other or how several variables are related. For example, you can use it to determine if and to what extent experience or gender impacts salaries.

Regression is also useful when you want to forecast a response using a new set of predictors. For example, you could try to predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and number of residents in that household.

Python code:

```
plt.figure(figsize=(10,8))
sns.pairplot(data=df,vars=['Total','cogs','gross income','Tax 5%'])
plt.show()
```

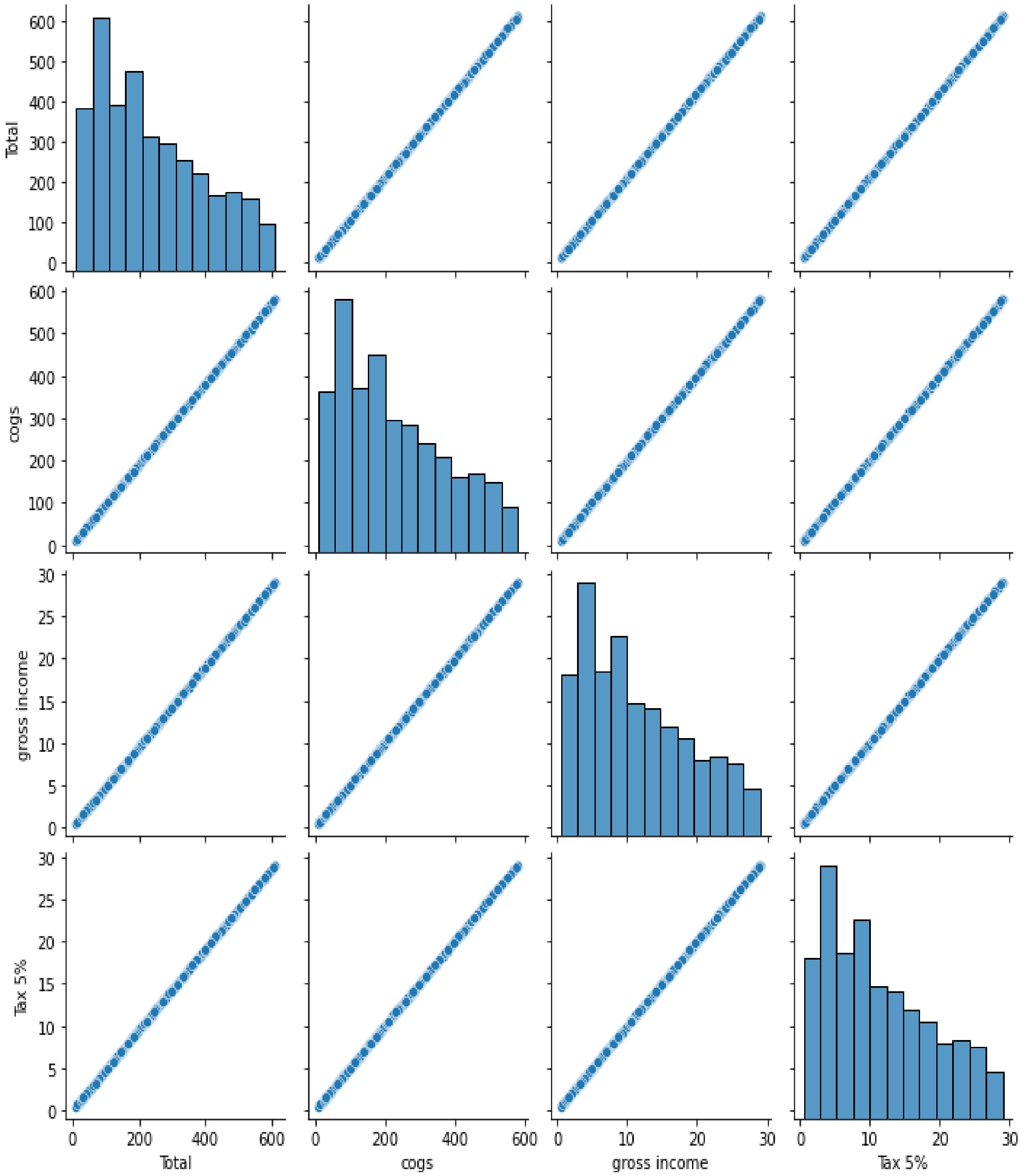


Fig.1.6

From the heatmap above, we can see that the Total, cogs, Tax 5%, gross income are perfectly correlated. This is very considerable, so let drop 3 of 4 attributes above (I choose cogs, gross income and Tax 5%).

Relationship of cities and branches

However, Pearson's correlation only works for continous value. For the category attribute, we can predict and check their correlation by scatter plot. For instance, let check the relationship between city and branch.

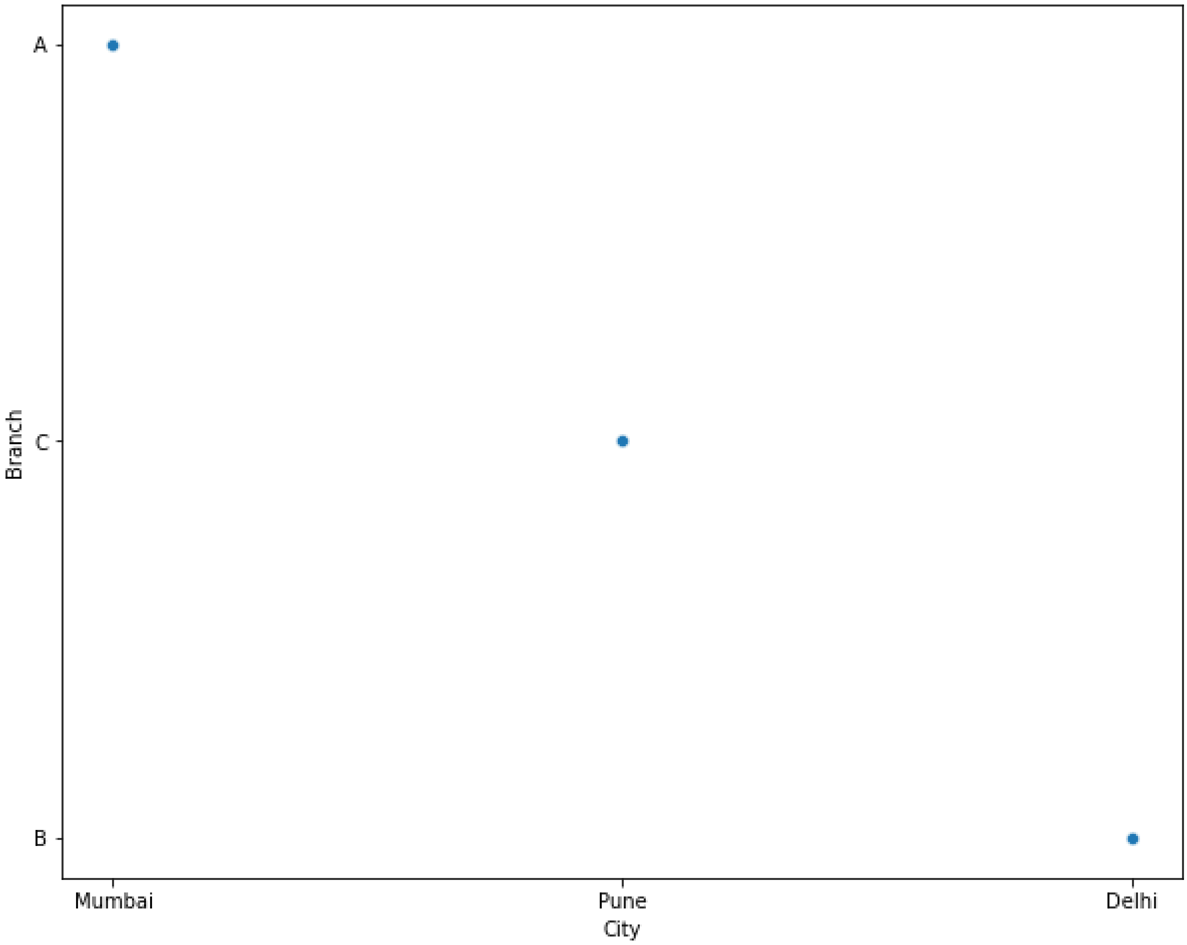


Fig.1.7

As a result, there is a relationship between city and branch. Each branch is located at a specific city. So, we can also remove 1 of those two (I remove City in this case).

Encoding the categorical variables:

In our dataset we have 5 columns categorical types which are branch, customer type, gender, product line and payment type. To use these categorical variables in our models we have to convert them into numeric type. So for that purpose we used OneHotEncoding from sklearn in python and used transformed data for further analysis.

Scalar Transformation:

In our data different features has different scales. We cannot use these different scales in one mathematical equation for our model. Hence to make each feature unitless and reliable to use in models we have use StandardScaler Transformation.

Summary statistics for all Numeric variables:

	Unit price	Quantity	Total	Rating	Hour	Day	Month
count	846.000000	846.000000	846.000000	846.000000	846.000000	846.000000	846.000000
mean	50.358452	4.913712	240.598142	7.022695	14.951537	15.234043	2.000000
std	24.962783	2.742327	158.206815	158.206815	3.202627	8.677909	0.835598
min	10.080000	1.000000	10.678500	4.000000	10.000000	10.000000	1.000000
25%	28.462500	3.000000	101.682000	5.600000	12.000000	8.000000	1.000000
50%	48.565000	5.000000	207.144000	7.000000	15.000000	15.000000	2.000000
75%	70.635000	7.000000	353.149125	8.500000	18.000000	23.000000	3.000000
max	99.890000	10.000000	609.168000	10.000000	20.000000	31.000000	3.000000

Form the five point summary of numerical columns we can notice that highest individual total sale noted was 609.16800 and highest individual quantity sold was 10. Overall Average Customer rating is 10.

Create Training And Testing Split :

In our data total number of observations are 1000. After removing outliers we get 846 observations. We have divide the data into 80% training data and 20% testing data
Training data summary :
After split the data 676 observations are training dataset.
Testing data summary : After split the data 170 observations are test dataset .

Python Library :

- Numpy
- Pandas
- Seaborn
- sklearn

NumPy is a Python library used for working with arrays. It stands for "Numerical Python

Pandas is a Python library used for data manipulation and analysis. It provides data structures for efficiently storing and manipulating large datasets, as well as tools for data cleaning, merging, and reshaping.

Seaborn is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating beautiful and informative statistical graphics.

Scikit-learn, also known as sklearn, is a Python library for machine learning. It provides a range of tools for machine learning tasks such as classification, regression, clustering, and dimensionality reduction.

CHAPTER 4

DATA ANALYSIS AND DATA INTERPRETATION

DATA ANALYSIS AND DATA INTERPRETATION

Table: Total sale per gender branch A

Gender	**Male**	**female**
Frequency	179	161
Percentage	48%	52%

Table no.1.1

Pie Chart:

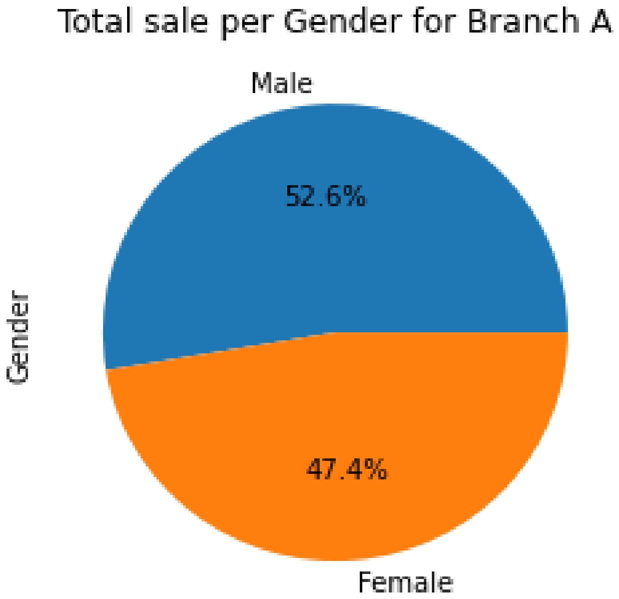


Fig.4.1

Interpretation:

The above pie chart shows that 48% of the total sales are from male customers and 52% are from female customers for branch A.

Table: Total sale per Gender for Branch B

Gender	**Male**	**Female**
Frequency	170	162
Percentage	51.2%	48.8%

Table no.1.2

Pie Chart:

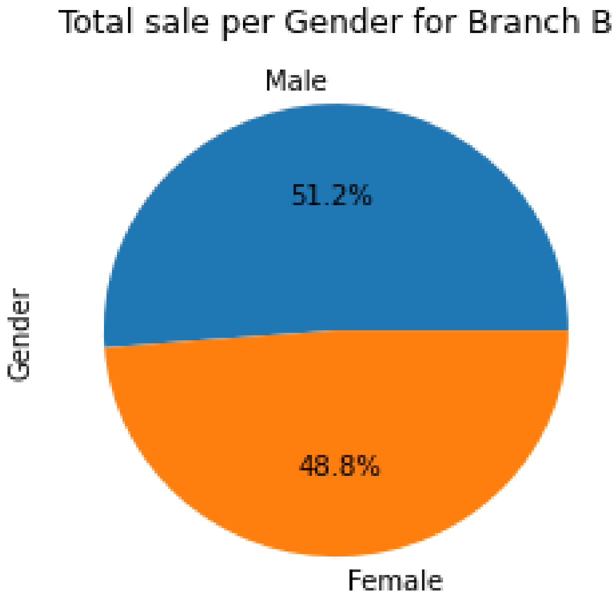


Fig.4.2

Interpretation:

The above pie chart shows that 49% of the total sales are from male customers and 51% are from female customers for branch B.

Table: Total sale per gender branch C

Gender	Male	female
Frequency	**178**	**150**
Percentage	**45.7%**	**54.3%**

Table no.1.3

Pie Chart:

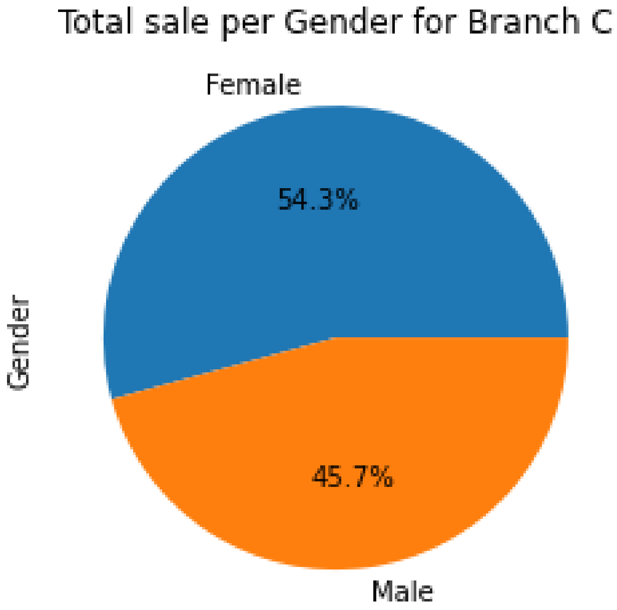


Fig.4.3

Interpretation:

The above pie chart shows that 48% of the total sales are from male customers and 52% are from female customers for branch C.

Table: payment method used by branch A customers

Payment Method of Branch A	**Percentage**
Cash	32.4%
Credit card	30.6%
E-Wallet	37.1%

Table no.1.4

Pie Chart:

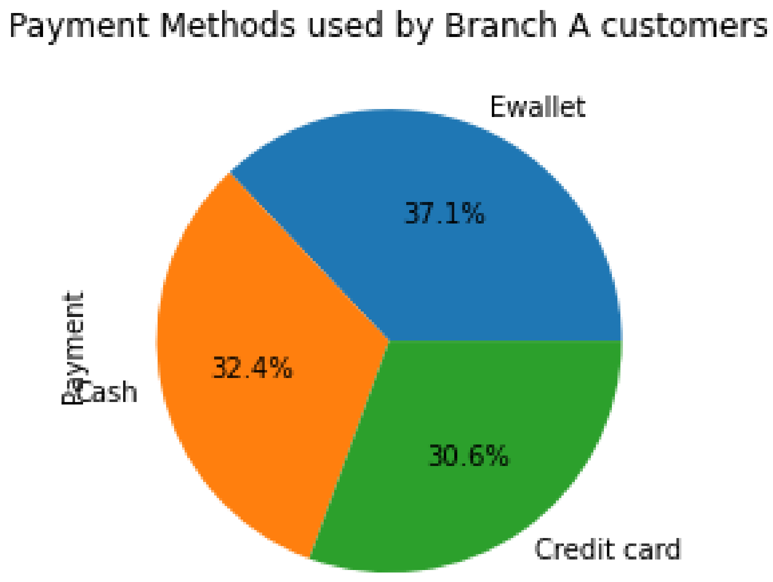


Fig.4.4

Interpretation:

In branch A, 32% customers used cash method for payment while 30% and 38% customers used credit card and E-wallet respectively. So E-wallet payment method is most used payment method than others.

Table: payment method used by branch B customers

Payment Method of Branch B	**Percentage**
Cash	33.1%
Credit card	32.8%
E-Wallet	34.8%

Table no.1.5

Pie chart:

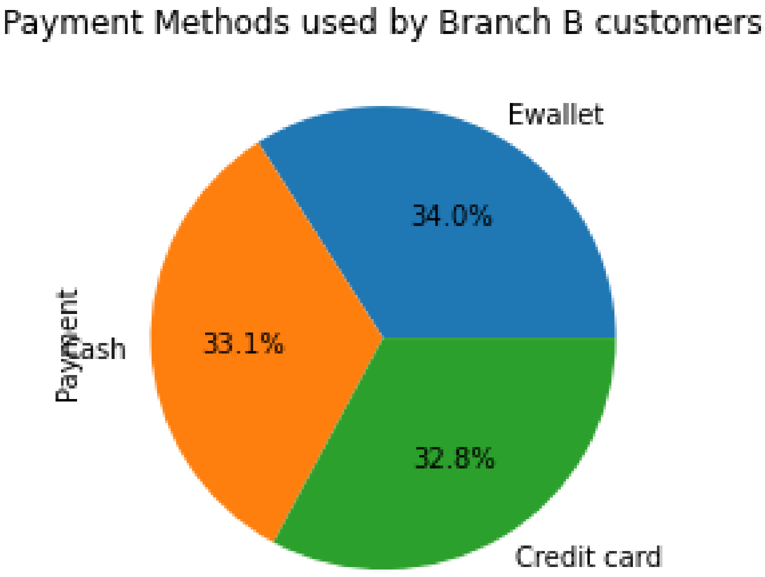


Fig.4.5

Interpretation:

In branch B, 34% customers used cash method for payment while 31% and 35% customers used credit card and Ewallet respectively. So Ewallet payment method is most used payment method than others.

Table: payment method used by branch C customers

Payment Method of Branch C	Percentage
Cash	37.8%
Credit card	29.9%
E-Wallet	32.3%

Table no.1.6

Pie chart:

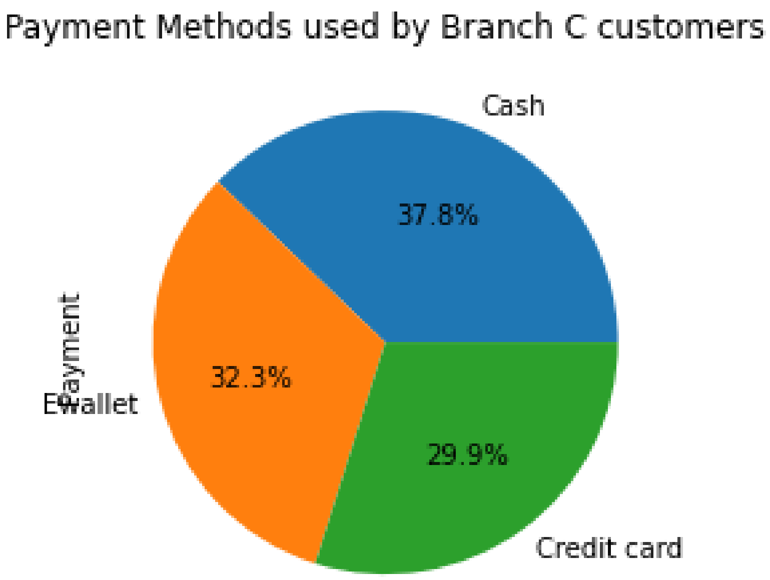


Fig.4.6

Interpretation:

In branch C, 36% customers used cash method for payment while 31% and 33% customers used credit card and Ewallet respectively. So cash payment method is most used payment method than others.

Table: Total Branch payment

Branch	**A,B,C**
Cash	34.4%
E-wallet	34.5%
Credit card	31.1%

Table no.1.7

Pie Chart:

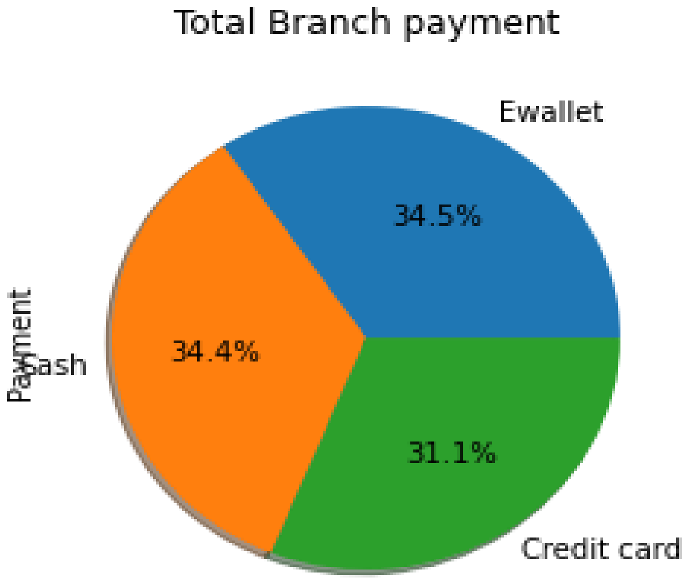


Fig.4.7

Interpretation:

In total branch of the customers use payment method are E-wallet payment 34.5% ,cash payment 34.4% and credit card payment is 31.1% so cash payment and E-wallet payment are almost similar.

Table : Average Total sale per product line

Branch	**A**	**B**	**C**
Electronic accessories	215	228	249
Fashion accessories	228	223	235
Food and beverages	242	249	239
Health and beauty	237	267	231
Home and lifestyle	257	239	239
Sports and travel	266	238	243

Table no.1.8

Bar Plot:

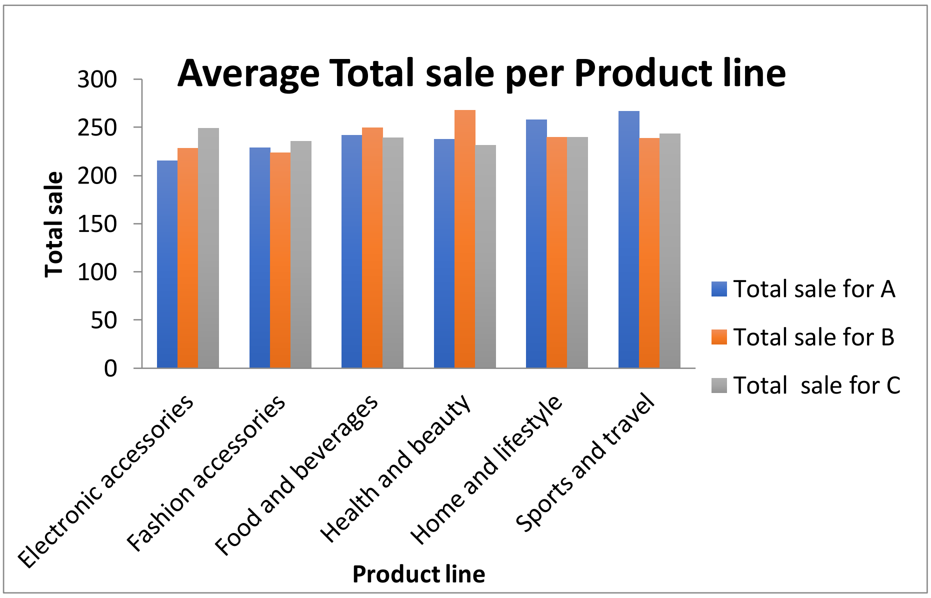


Fig.4.8

Interpretation:

The average total sales of all the product lines is almost same for all three branches. The above plot shows that the sales of electronic accessories is more in the branch C. In branch B the sales of health and beauty is more while sales of sports and travel is more in the branch A.

Table: Average Rating per product line

Branch	A	B	C
Electronic accessories	6.91	7.27	6.79
Fashion accessories	6.85	6.64	7.59
Food and beverages	7.40	6.94	6.99
Health and beauty	6.89	7.34	6.95
Home and lifestyle	7.05	6.54	7.06
Sports and travel	7.28	6.63	7.11

Table no.1.9

Bar plot:

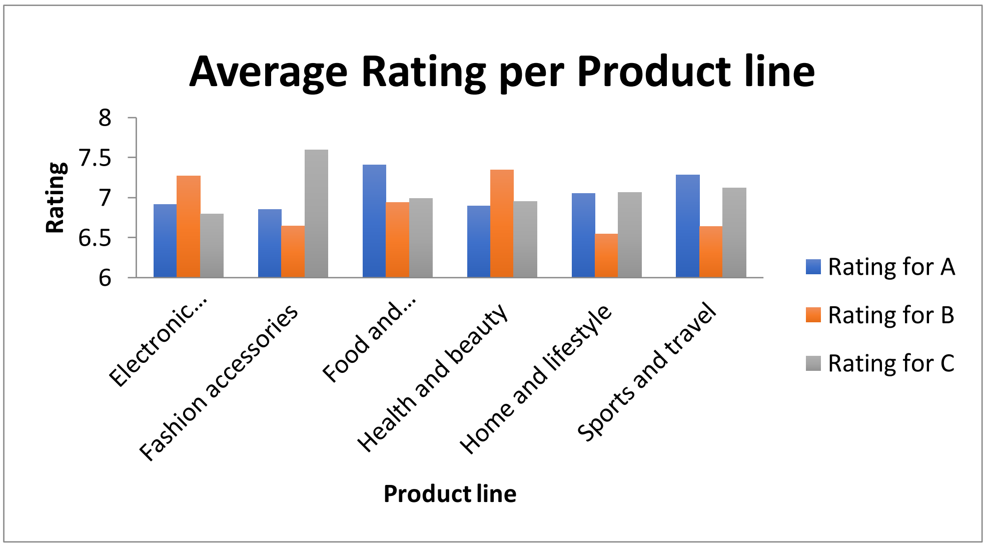


Fig.4.9

Interpretation:

From above joint bar plot, it is clear that average rating is high for Fashion accessories in Branch C. In branch B, average rating is high for Health and Beauty and in branch A average rating is high for Sports and Travel .

MODEL BUILDING

In this project we use machine learning techniques such as Ridge regression, Linear regression, SVM, Random Forest, XGBoost and compare all the algorithms to build a model which is best fit for our data.

Multiple Linear Regression

In Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be various cases in which the response variable is affected by more than one predictor variable; for such cases, the Multiple Linear Regression algorithm is used.

Moreover, Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable.

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

Assumptions for Multiple Linear Regression:

- A **linear relationship** should exist between the Target and predictor variables.
- The regression residuals must be **normally distributed**.
- MLR assumes little or **no multicollinearity** (correlation between the independent variable) in data.

Output of the model:

Multiple linear regression model r2 score : 0.8223131965471551

Conclusion:

Accuracy score for Multiple Linear Regression model is 82.23 %

Ridge Regression

Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.

Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.

Ridge regression is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.

Output
Ridge model r2 Score: 0.8223521073512319

Conclusion:
Accuracy score for Ridge Regression model is 82.23%

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Output
svm Model r2 Score: 0.7974942629301465

Conclusion:
Accuracy score for Support Vector Machine model is 79.75 %

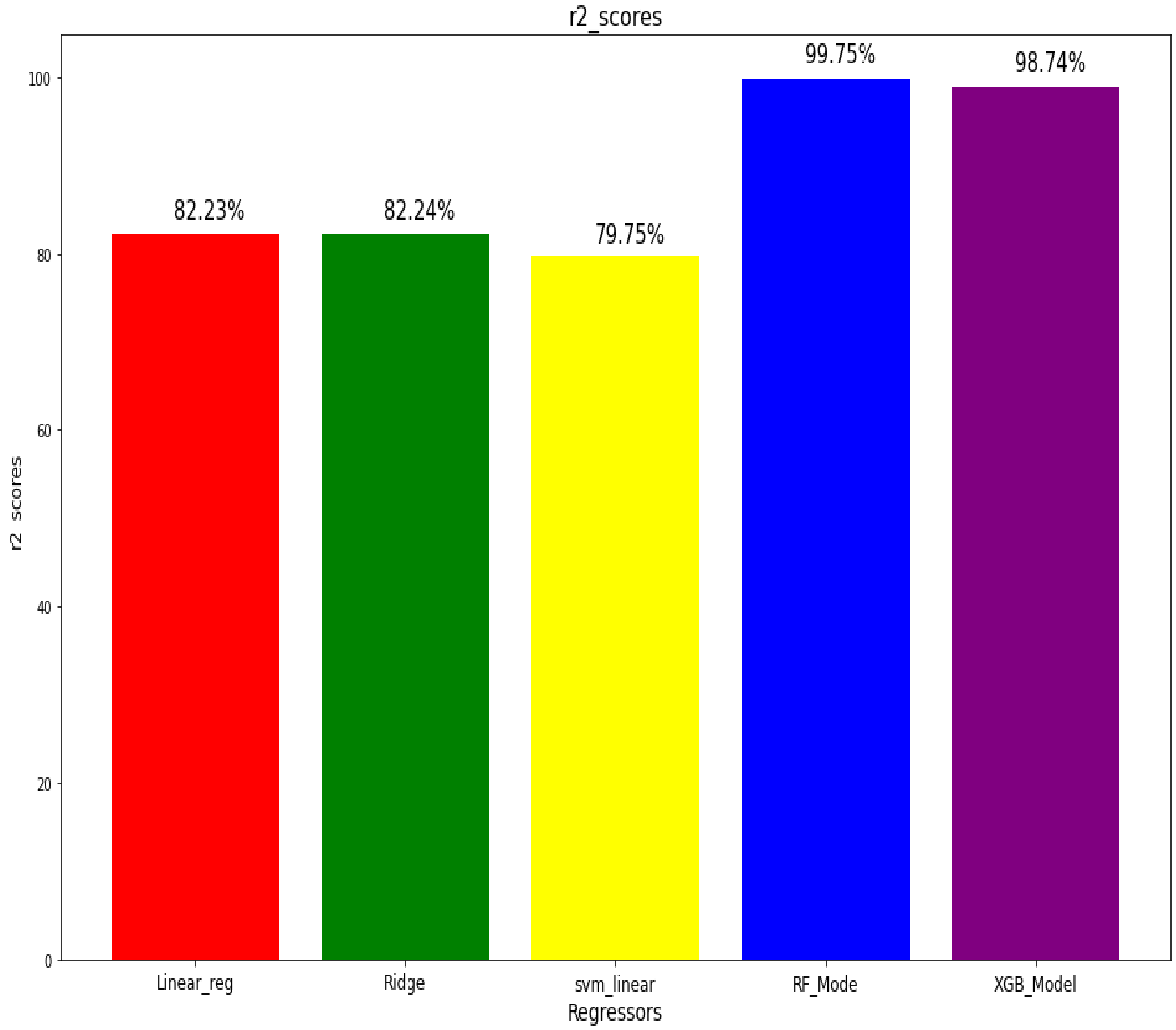


Fig.5.0

Interpretation:

From above graph it is clearly seen that accuracy score for Random forest model is 99.75% which is the highest accuracy among all the models fitted here. So, we choose Random forest model for prediction of total sale .

CHAPTER 5

CONCLUSIONS, FINDINGS, SUGGESTIONS AND LIMITATIONS:

FINDINGS:

- Seasonality: Sales tend to increase during holidays and other special occasions, as well as during certain times of the year (e.g., summer barbecues, back-to-school shopping, etc.).

- Promotions: Sales and discounts can attract customers and increase sales, but they can also cannibalize future sales if customers stock up during the promotion and delay their next purchase.

- Consumer behavior: Changes in consumer preferences, habits, and demographics can affect sales, as well as external factors such as the economy and public health crises.

- These feature are supermarket type grocery store, product price and supermarket opening years. The grocery store type seem to sell more and generally have higher sales than other type of store. The product price also affects sales as higher prices of products that sell more generally contributes to higher sales and finnaly the supermarket opening years , where newer store sell higher than older store.

SUGGESTIONS:

Customer segmentation: segment your customers based on demographics, purchasing behavior, and preference. This can help you understand the needs and wants of your customers and tailor your sales and marketing efforts accordingly

Inventory Management: Analise inventory data to determine which products sell quickly and which are slow-moving. This can help you optimize your inventory levels and reduce waste.

External Factors: Consider external factors that may affect sales, such as weather, holidays, and economic conditions. For example, sales of ice cream may increase during hot summer months, while sales of holiday decorations may increase during the holiday season.

LIMITATIONS OF THE STUDY

- This Supermarket sales data received from company may not be totally Original or Accurate. Hence, the Statistical analysis may not be justified in real life.

- Also, in this data, if the actual product was mentioned in place of the product category, a better “Market Basket Analysis” could have been performed.

- Data Quality: The accuracy and completeness of historical sales data are critical for building accurate predictive models. If the data is incomplete or inaccurate, the predictions may be unreliable.
- External Factors: External factors, such as changes in consumer preferences or economic conditions, can be difficult to predict and may affect sales in ways that are not captured in historical data.
- Seasonality: Seasonal fluctuations in sales can be challenging to predict accurately. For example, sales of ice cream may be high during the summer months, but predicting the exact timing and extent of these fluctuations can be difficult.
- Competition: Competition from other retailers can also impact sales and is challenging to predict. Changes in the competitive landscape, such as new store openings or price wars, can quickly disrupt sales patterns.

CHAPTER 6

CONCLUSION

The accuracy score for Random forest model is 99.75% which is the highest accuracy among all the models fitted here. Hence, we propose a model using the Random Forest algorithm and also we compare it with the other machine learning techniques such as Ridge regression, Linear regression, SVM. Based on the analysis of the supermarket sales data, it is possible to predict future sales with a reasonable degree of accuracy using various machine learning models.

The exploratory data analysis revealed that sales tend to be higher on weekends and holidays, and there is a strong positive correlation between the number of customers and the total sales. The analysis also identified certain product categories that contribute significantly to the total sales.

Several machine learning models were trained and evaluated using the sales data, including linear regression, decision tree, and random forest models. The results indicated that the random forest model outperformed the other models in terms of accuracy and was able to capture the nonlinear relationships between the sales data and the various predictors.

In conclusion, the supermarket can use the insights gained from the analysis and the predictive models to make informed decisions about inventory management, staffing levels, and promotional strategies. By leveraging machine learning techniques, the supermarket can optimize its operations and improve its profitability.

Female customers purchase more than male customers.

Customers prefer E-wallet payment type more to buy the products than other payment type.

The average total sales of all the product lines is almost same for all three branches. The above plot shows that the sales of electronic accessories is more in the branch C. In branch B the sales of health and beauty is more while sales of sports and travel is more in the branch A.

Average rating is high for Sports and Travel in Branch A. In branch B, average rating is high for Health and Beauty and in branch C average rating is high for fashion accessories.

BIBLlOGRAPHY

- Han J., Kamber M., and Pei J (2012) Data Mining: Concepts and Techniques. (Elsevier)

- Alex Smola and S.V.N. Vishwanathan (2008) Introduction to Machine Learning. (Third Edition) (Cambridge University Press)

- Ian H. Witten and Eibe Frank (2005) Data Mining: Practical Machine Learning Tools and Techniques. (Second Edition) (Elsevier)

- Mishra, S., & Mishra, P. (2021). Supermarket Sales Forecasting using Time Series Analysis and Machine Learning Techniques. International Journal of Advanced Computer Science and Applications, 12(2), 125-133.

- Das, R., Kumar, A., & Roy, P. K. (2017). Prediction of supermarket sales using machine learning algorithms. In 2017 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 604-608). IEEE.

- "Research Methodology: Methods and Techniques"- C.R. Kothari