A STUDY ON BIGMART SALES PREDICTION USING MACHINE LEARNING

INTERNSHIP REPORT SUBMITTED TO THE BHARATHIAR UNIVERSITY

FOR THE AWARD OF THE DEGREE OF

MASTER OF BUSINESS ADMINISTRATION

By

S. SETHURAMALAKSHMI

1P23MB030

Under the Guidance of

N.Vellingiri

M.C.A., B.Ed.,

Assistant Professor

SCHOOL OF MANAGEMENT STUDIES - PG

RVS COLLEGE OF ARTS AND SCIENCE (AUTONOMOUS)

Affiliated to Bharathiar University, Approved by AICTE
Re Accredited with 'A+' Grade by NAAC

Sulur, Coimbatore – 641 402.

NOVEMBER 2024

CERTIFICATE

This is to certify that the Internship report, entitled **"A Study on BigMart Sales Prediction Using Machine Learning"**, submitted to the Bharathiar University, in partial fulfilment of the requirements for the award of the **DEGREE OF MASTER OF BUSINESS ADMINISTRATION**, is a record of original work done by **Miss. SETHURAMALAKSHMI S** During the period **May 2024 to July 2024** Of his internship in School of Management Studies - PG, RVS College of Arts and Science (Autonomous), Coimbatore - 641402, under my supervision and guidance and the internship report has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or other similar title of any candidate of any University.

Date:

Director Signature of the Guide

Date of Viva-voce Examination held on _____.

Internal Examiner External Examiner

DECLARATION

I, **SETHURAMALAKSHMI S** hereby declare that the internship, entitled **"A Study on BigMart Sales Prediction Using Machine Learning",** submitted to the Bharathiar University, in partial fulfilment of the requirements for the award of the **DEGREE OF MASTER OF BUSINESS ADMINISTRATION** is a record of original and independent research work done by me during the period May 2024 to July 2024 under the supervision and guidance of **Mr.N.Vellingiri, M.C.A., B.Ed., Assistant Professor,** School of Management Studies - PG, RVS College of Arts and Science (Autonomous), Coimbatore – 641 402 and it has not formed the basis for the award of any other Degree / Diploma / Associateship / Fellowship or other similar title to any candidate of any University.

Date: Signature of the Candidate

Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

CERTIFICATE

OF INTERNSHIP

Ministry of MSME, Govt. of India

This is to Certify that

S.SETHURAMALAKSHMI

MBA (BUSINESS ANALYTICS)

RVS COLLEGE OF ARTS AND SCIENCE

has Successfully Completed the **45 Days** Internship on

Machine Learning

at Pantech e learning Pvt. Ltd.

Duration: From **15th May 2024** to **30th June 2024**

PEL-SI-2024-2602

CERTIFICATE NO

DIRECTOR, PANTECH E LEARNING
WWW.PANTECHELEARNING.COM

Acknowledgement

I would like to extend my sincere gratitude to **Pantech E-Learning** for the opportunity to intern at such an esteemed organization Their insightful feedback and constructive criticism have greatly helped in broadening my understanding of various concepts related to Machine Learning.

My deepest appreciation goes to **Mr.N.Vellingiri, M.C.A., B.Ed., Assistant Professor** for their exceptional guidance and mentorship throughout my internship. Their encouragement and constructive feedback motivated me to push my boundaries and explore new dimensions in data analytics.

I am immensely grateful to my academic institution, **RVS College of Arts and Science** for continuously motivating and supporting me during this internship journey. Their assistance in the preparation and approval of this internship made it possible for me to gain such a valuable experience.

Lastly, I would like to thank my family and friends for their unwavering support and encouragement throughout my internship period. Without their constant belief in my abilities, this would not have been possible.

Thank you all for making this a rewarding learning experience.

CONTENTS

CHAPTER-I

INTRODUCTION OF THE REPORT

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store to give accurate results. Big Mart has collected sales data from Kaggle, for various products across different stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business.

1.1 BACKGROUND OF STUDY

In the rapidly evolving retail landscape, the ability to accurately predict sales is vital for optimizing business operations and ensuring customer satisfaction. BigMart, a leading supermarket chain with a widespread presence across the country, recognizes the importance of leveraging data analytics to enhance its sales forecasting capabilities.

The study focuses on the development of a predictive model that utilizes historical sales data collected from numerous stores to forecast product sales at both the product and store levels. This approach is essential for addressing the complexities of inventory management, where timely availability of popular products is crucial for maximizing sales and minimizing costs associated with overstocking.

Sales forecasting at BigMart is influenced by a myriad of factors, including product characteristics, store attributes, promotional efforts, and seasonal trends. By analyzing this multifaceted dataset, the study aims to uncover the key drivers of sales performance.

Advanced machine learning algorithms will be applied to model the intricate relationships between these variables, enabling a more precise understanding of how different factors contribute to sales outcomes.

Moreover, the incorporation of geographic segmentation in the analysis will provide insights into regional variations in consumer behavior, allowing BigMart to tailor its marketing and inventory strategies to specific local demands.

Effective utilization of historical sales data will involve rigorous data preprocessing and exploratory data analysis (EDA) to identify trends and correlations within the dataset. Visualization tools will be employed to present findings in an accessible manner, facilitating interpretation and decision-making processes.

The primary objective of this study is to deliver an accurate and actionable sales prediction model that not only enhances inventory management but also informs strategic decisions regarding pricing, promotions, and product placements.

By successfully implementing this predictive model, BigMart aims to achieve several outcomes: improved inventory turnover rates, increased sales through better product availability, and enhanced responsiveness to market fluctuations.

Ultimately, this study seeks to position BigMart as a data-driven leader in the retail industry, enabling the company to harness the power of analytics to drive business success and adapt to the dynamic nature of consumer preferences and market conditions.

Through informed decision-making and a deeper understanding of sales dynamics, BigMart can strengthen its competitive advantage and foster long-term growth in an increasingly complex retail environment.

To facilitate this analysis, the study will employ various data pre-processing and visualization techniques to prepare and interpret the data effectively. The ultimate goal is to create an accurate and reliable sales prediction model that empowers BigMart to make informed decisions regarding inventory control, marketing strategies, and resource allocation. By achieving these objectives, the study aims to enhance BigMart's operational efficiency and overall competitiveness in the retail market.

The BigMart Sales Prediction study aims to bridge the gap between data and actionable business insights. By utilizing historical sales data and advanced analytics, the study endeavors to equip BigMart with the tools needed to enhance its operational efficiency, improve customer satisfaction, and secure a leading position in the competitive retail market. This comprehensive approach not only aims for immediate improvements in sales forecasting but also sets the stage for a data-centric culture within the organization, fostering continuous improvement and innovation.

1.2 PURPOSE OF STUDY

The primary purpose of this study is to develop a comprehensive predictive sales model for BigMart that utilizes historical sales data to enhance inventory management and optimize marketing strategies across its extensive network of supermarkets. In the competitive retail environment, accurately forecasting sales is critical for ensuring the availability of popular products while minimizing costs associated with overstocking. This study aims to identify and analyze key drivers of sales performance, including product characteristics, store attributes, promotional efforts, and seasonal trends.

By employing advanced data analytics and machine learning algorithms, the study will uncover complex relationships within the sales data, providing insights that can inform strategic decision-making. The insights derived from this analysis will help BigMart to tailor its inventory strategies, ensuring that products are stocked appropriately based on expected demand. Additionally, the predictive model will facilitate a more effective approach to pricing and promotional activities, allowing BigMart to maximize sales and enhance customer satisfaction.

Another key objective of the study is to explore the impact of geographic segmentation on sales performance. Understanding regional variations in consumer behavior will enable BigMart to develop targeted marketing strategies that resonate with local preferences and demands. By aligning inventory and promotional efforts with these insights, BigMart can improve its responsiveness to market fluctuations and drive sales growth.

Furthermore, the study aims to establish a framework for ongoing data analysis and sales forecasting within the organization. By fostering a data-driven culture, BigMart can continuously refine its strategies and adapt to changing consumer trends over time.

This study seeks to provide BigMart with a powerful sales prediction model that not only enhances inventory management but also drives strategic marketing efforts. By understanding the intricacies of sales dynamics and consumer behavior, BigMart can strengthen its competitive advantage and foster long-term success in an increasingly complex retail landscape. Through informed decision-making and data-driven insights, the company aims to enhance customer satisfaction and achieve greater profitability.

1.3 SCOPE OF WORK

The scope of this project will focus on developing a predictive sales model for BigMart using advanced analytics techniques to analyze sales performance across various product categories. The primary objectives will include the collection, pre-processing, and visualization of historical sales data to gain insights .

Gather historical sales data from BigMart's diverse stores, including product details, sales figures, promotions, and regional information. Pre-process the collected data to ensure accuracy and consistency, addressing any missing values and anomalies.

Utilize machine learning algorithms to develop a predictive sales model that forecasts future sales based on historical data. Train and validate the model to ensure accuracy and reliability in sales predictions.

1.4 METHODOLOGY

The data for the BigMart sales prediction study is categorized into two main types:

1. **PRIMARY DATA**
2. **SECONDARY DATA**.

Each category serves distinct purposes in understanding sales performance and developing the predictive model.

DATA SOURCES

1.PRIMARY DATA

- Primary data is collected directly by the BigMart team through various channels, including surveys and questionnaires distributed via social media and other digital platforms.

2.SECONDARY DATA

- Secondary data comprises historical sales records, product details, and demographic information sourced from BigMart’s internal databases and external market research reports.

CHAPTER-II

COMPANY DETAILS

2.1 OVERVIEW OF THE INDUSTRY

Education is the base for economical growth as well as social transformation of any country. Education and Training services is a broad category that encompasses job specific certification training, project training and classes emphasizing self-fulfilment and personal motivation. Many of the industries' programmes, classes and training services fall under the category of Career and Technical Education (CTE), also known as Vocational Education. Industrial training's aim is to improve the industrial knowledge among the students or professionals and also to develop their ability to comply with its regulatory requirements.

Global Education and training services companies are increasingly looking for new growth opportunities. Especially China and India rely on these services for their economy. Leading Education and Services firm include New Oriental Education and Technology group of China, NIIT Limited of India and Third Force of Ireland.

There are also firms which involve in Software Projects Development, perform Outsourcing activities and System integration services along with Education and Training services. Software Projects Development deals with Multimedia solutions and IT related projects development and carrying out outsourcing activities for large scale IT Enterprises. Firms also involve in providing Lab solutions to Engineering Colleges, say for example, Development of Evaluation boards, Elance boards and Webserver boards for electronics and communication department.

In the global marketplace, education and training services have been expanding rapidly due to technological advancements, the rise of the digital economy, and increased globalization. Countries like India and China are investing heavily in these services as they aim to upskill their populations to remain competitive in the global economy.

2.2 COMPANY PROFILE - INTRODUCTION

Pantech Solutions Pvt. Ltd. is one of the well-known and well-trusted solution providers in South India for Education and Training, IT and Electronics Applications. Today, Pantech stands as a source of reliable and innovative products that enhance the quality of customer's professional and personal lives.

Conceived in 2004, Pantech Solutions is rooted in Chennai and has its branches in Hyderabad, Bangalore, Pune, Cochin, Coimbatore and Madurai. Pantech is a leading solution provider in all technologies and has extensive experience in research and development. Its 260 employees in all the metros of South-India are active in the areas of Production, Software Development, Implementation, System integration, Marketing, Education and Training.

2.3 WHY PANTECH?

With a client list spanning nearly in all industries, and colleges, Pantech Solutions' product solutions have benefited customers of many different sizes, from non-profit organizations to companies.

- **Our Vision: "**To Gain Global Leadership in Providing Technological Solutions Through Sustained Innovation".
- **Core Values:** When we take on your project, we take the stewardship of the project with you in the director's seat. As stewards of your project, we consider ourselves successful not when we deliver your final product but when the product meets your business objectives. You'll see that our 6 core values are derived from our stewardship quality.
    - **Integrity** – Honesty in how we deal with our clients, each other and with the world.
    - **Candor** – Be open and upfront in all our conversations. Keep clients updated on the real situation. Deal with situations early; avoid last minute surprises.
    - **Service** – Seek to empower and enable our clients. Consider ourselves successful not when we deliver our client's final product but when the product is launched and meets success.

- **Kindness** – Go the extra mile. Speak the truth with grace. Deliver more than is expected or promised.
- **Competence** – Benchmark with the best in the business. Try new and better things. Never rest on laurels. Move out of comfort zones. Keep suggesting new things. Seek to know more.
- **Growth** – Success is a journey, not a destination. Seek to multiply/increase what we have – wealth, skills, influence, and our client's business.

2.3.1 PRODUCTS AND SERVICES

Pantech Solutions' business activities are divided into three broad areas:

1. **Solution**
2. **Service**
3. **Product**

Solutions

➢ Multimedia Solutions

Pantech Multimedia Solutions division specializes in website design and development, web-based information systems, flash and animations, e-commerce applications, Database creation, Web based applications, digital presentations and virtual tours.

➢ Technology Solutions

Pantech Technology Solutions is a consulting division that advices and introduces, cutting edge technology based solutions to clients. This division aims to open the Southern African Business and the IT Sector as a whole to a variety of niche markets.

➢ Technical Support

Pantech Technical Support Division not only Complements its other divisions by providing highly experienced technical engineers to support and maintain the various products and services but also outsource it's expertise to other IT companies and corporate. Whatever is the requirement, the Pantech team is ready to develop a solution using its structured project management approach to ensure that the project arrives on time and within budget.

Service

System Architecture - a flexible, scalable and cost-effective architecture is constructed by o Identifying, designing and interfacing the Hardware building blocks to realize the product in the block level.

- Defining Software building blocks and interfaces.
- Validating the implementation of the individual building blocks and their interfaces.
- Validation and fine-tuning of the entire architecture.
- Defining the Design requirements for each and every Hardware and Software building block and interface.
- Design for Manufacturability: Component Engineering to ensure the Manufacturability Selection of components, Availability and Replacement options for chosen components
- Design for Testability: Defining Test Methodologies and Diagnostics package development.

Product

Embedded Solutions for electronics and communication applications result in the following end products.

- **8051 EVALUATION BOARD** NXP's P89V51RD2, 8051 Kit is proposed to smooth the progress of developing and debugging of various designs encompassing of High speed 8-bit Microcontrollers.
- **ARM9 ELANCE BOARD** ATMEL's ARM9 AT91SAM9261, ARM Kit is High–end mobile technology, proposed to smooth the progress of developing and debugging of various designs encompassing of High speed 32-bit processors. It integrates on board TFT Display, Ethernet, Memories, USB device and host controller and audio codec to create a stand-alone versatile test platform.
- **ENC28J60 WEBSERVER BOARD** The PS-PIC-WEBSERVER development Board is developed to embed the PIC microcontroller into internet or intranet. It is well suited for the user to write TCP/UDP application with an 8-bit microcontroller. This enhanced board supports Microchip's 40-pin PIC micro controllers (16F/18F).

2.3.2 CLIENTELE

Over the past 7 years, Pantech Solutions have improved the quality of communication and satisfied customers earning their respect by providing excellent products and services.

In addition, the Company is flexible with services and financial structures for contracts aiming for mutually beneficial relationships with the customers. Their range of customers is like Large Corporate Offices, Universities, Educational Institutions, Factories, etc.

2.3.2.1 EDUCATION AND ACADEMIC

ISRO	Ahmedabad
Meenakshi Ramasamy Polytechnic College	Ariyalur
Arkay College of Engineering	Bodhan, Andhra
Anna University	Chennai
Bharath Polytechnic College	Chennai
CPCL Polytechnic College	Chennai
PSG Institute Of Management	Coimbatore

2.3.2.2 INDUSTRIES

Indian Space Research Organization(ISRO)	Bangalore
Defence Research Development Organization (DRDO)	Delhi
National Small Industries Corporation(NSIC)	Delhi
L&T	Chennai
ITI	Chennai
NIT	Trichy

2.4 ORGANIZATION MISSION

Over the new few years our goal is to harness our talents and skills by permeating our company further with process-centered management. In this way, once a customer's project enters our quality oriented process, it will exit as a quality product.

We will also strive to add to our knowledge and enhance our skills by creating a learning environment that includes providing internal technology seminars, attending conferences and seminars, building a knowledge library and encouraging learning in every way. Our in-house Intranet portal makes sure that knowledge is shared within the organization.

With our beliefs, the future can only look promising as we continue to build our team with the best Indian talent and mould them into our quality-oriented culture. We will find our niche in a competitive world by excelling at what we do, following our guiding principles and most importantly, listening to the needs of our customer.

Pantech e Learning

DIGITAL LEARNING SIMPLIFIED

CHAPTER – III

BROAD AREA OF WORK

In the business analysis process, **Machine Learning** acts as a powerful tool to derive insights, optimize operations, and facilitate data-driven decision-making. From predictive analytics and process automation to customer segmentation and financial forecasting, ML provides scalable solutions that help businesses address complex problems and capitalize on opportunities in a dynamic market environment.

Functional Areas of Work: **Machine Learning (Visual Code)**

This focuses on business analysis with an emphasis on Sales Prediction using Machine Learning Algorithms, there are several functional areas to explore. They are:

3.1 Data Collection and Preparation

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions and etc.

❖ Data Sourcing

Identifying relevant data sources, which could be internal databases, external APIs, Excel spreadsheets, or CSV files. The intern needs to ensure that the data sources are reliable and relevant to the business objectives.

❖ Data Cleaning

Before any analysis can be performed, data must be cleaned. This involves handling missing values, removing duplicates, correcting inconsistencies, and standardizing data formats.

❖ Data Transformation

The categorical variables (e.g., Outlet_Size, Outlet_Location_Type) likely underwent label encoding or one-hot encoding to convert them into numerical formats suitable for machine learning models.

❖ Splitting Data

The dataset is split into training and test sets (`X_train`, `Y_train`, `X_test`, `Y_test`), which allows for evaluating model performance on unseen data.

Data Preparation

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

3.2 Dataset

The dataset consists of 8523 individual data. There are 12 columns in the dataset, which are described below.

1. Item Identifier - Unique product ID

2.ItemWeight - Weight of product

3.ItemFatContent - Whether the product is low fat or not

4.ItemVisibility - The % of the total display area of all products in a -store allocated to the particular product

5.ItemType - The category to which the product belongs

6.ItemMRP - Maximum Retail Price (list price) of the product

7.OutletIdentifier - Unique store ID

8.OutletEstablishmentYear - The year in which the store was established

9.OutletSize - The size of the store in terms of ground area covered

10.OutletLocationType - The type of city in which the store is located

11.OutletType - The outlet is just a grocery store or some sort of supermarket

12.ItemOutletSales - sales of the product in t particular store. This is the outcome variable to be predicted.

3.3 Model Selection

A regression-based machine learning model (e.g., **XGBoost**) was trained to predict sales (Item_Outlet_Sales). The training process used features like product visibility, outlet size, and product type.

3.4 Model Training

1. **Feature Selection:** The first step in model training is to identify the important features that influence sales. These features could include:

    - Product characteristics like Item_Type and Item_MRP.
    - Store attributes like Outlet_Type, Outlet_Location_Type, and Outlet_Size.
    - Product visibility and marketing efforts, captured by Item_Visibility.

The selected features are then used as input to the model to predict sales.

2. **Splitting the Data:** The dataset is typically split into two sets:

    - **Training Set**: This is the portion of the data used to train the machine learning model. It allows the model to learn the relationships between the input features and the target (sales).
    - **Testing Set**: This part of the data is reserved for evaluating how well the model performs on unseen data.

3. **Model Selection:** Common models used for sales prediction, one of the model is **XGBoost.** XGBoost is an advanced implementation of the gradient boosting algorithm that aims to improve speed and performance. After training the model, it's important to evaluate how well it performs on unseen data (test set). Common evaluation metrics for regression problems like sales prediction.

4. **Training the Model:** During training, the model is fed the training data, which includes the selected features and the corresponding sales. The model adjusts its internal parameters to minimize the difference between the predicted sales and the actual sales. This process continues iteratively until the model achieves a satisfactory level of accuracy.

3.5 Model Evaluation

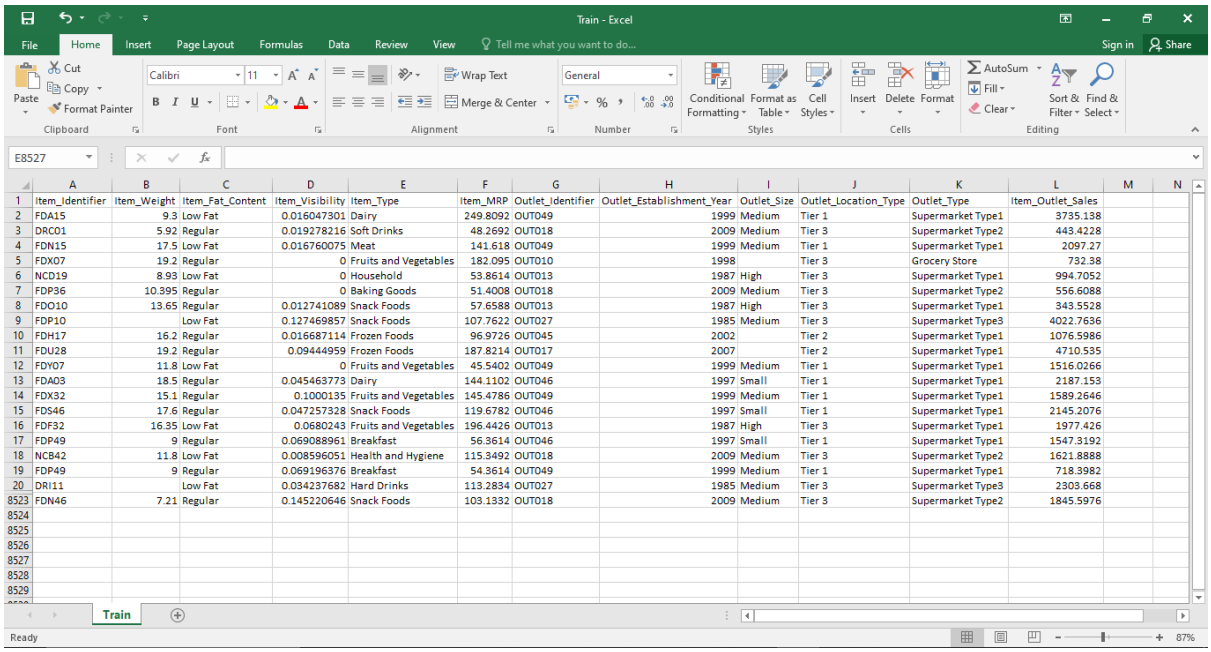
Evaluation Metrics: After training, the model is tested on the test dataset to check how well it generalizes to new, unseen data. Some common metrics used to evaluate the model's performance include:

- **R-squared (R^2)**: A metric that explains how well the model's predictions match the actual data. It shows the proportion of variance in the sales data that is predictable from the features.

CHAPTER – IV

ANALYSIS

4.1DATASET-BIGMART SALES

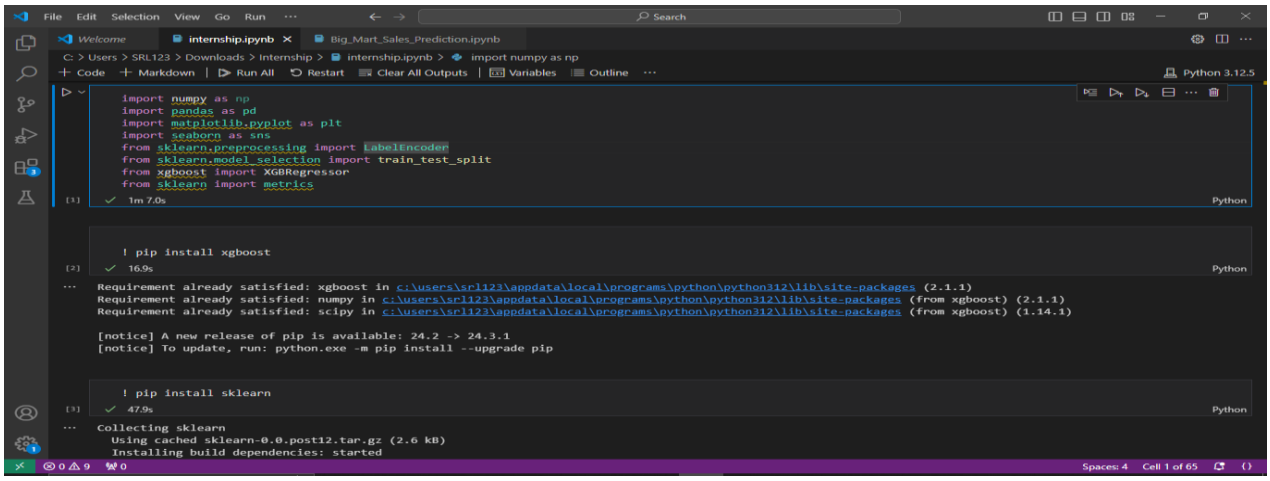


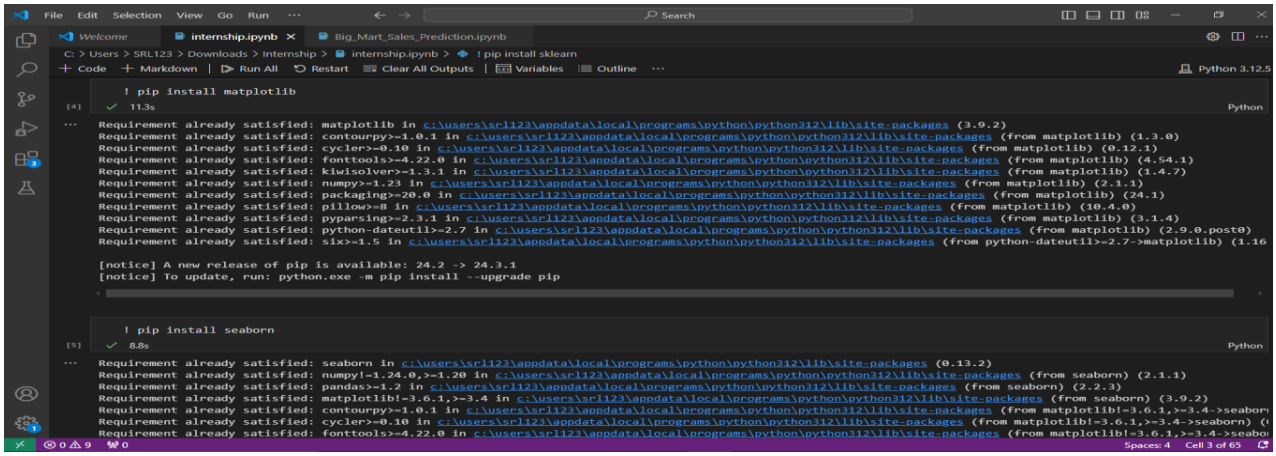
Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535
FDY07	11.8	Low Fat	0	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.0266
FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.2646
FDS46	17.6	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.426
FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.3192
NCB42	11.8	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.8888
FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.668
FDN46	7.21	Regular	0.145220646	Snack Foods	103.1332	OUT018	2009	Medium	Tier 3	Supermarket Type2	1845.5976

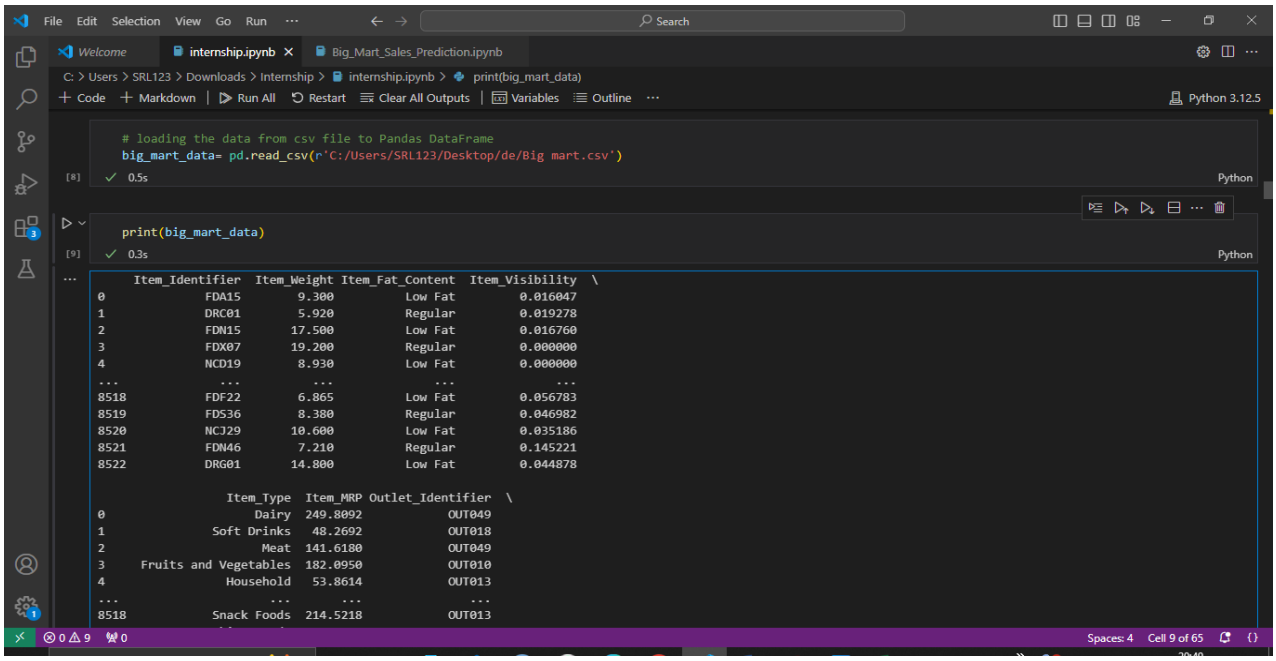
This dataset captures item-specific sales data across various outlets, including product attributes, outlet types, and sales figures. It's useful for analyzing factors influencing item sales and outlet performance.

The dataset consists of 8523 individual data. There are 12 columns in the dataset

4.2 ANALYSIS USING PYTHON (VISUAL CODE)







```python
# first 5 rows of the dataframe
big_mart_data.head()
```

[10] 0.1s

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outle
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Super
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Super
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Super
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	G
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Super

```python
# number of data points & number of features
big_mart_data.shape
```

[11] 0.0s

```
(8523, 12)
```

```python
# number of data points & number of features
big_mart_data.shape
```

[11] 0.0s

```
(8523, 12)
```

```python
# getting some information about thye dataset
big_mart_data.info()
```

[12] 0.3s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            8523 non-null   object
 1   Item_Weight                7060 non-null   float64
 2   Item_Fat_Content           8523 non-null   object
 3   Item_Visibility            8523 non-null   float64
 4   Item_Type                  8523 non-null   object
 5   Item_MRP                   8523 non-null   float64
 6   Outlet_Identifier          8523 non-null   object
 7   Outlet_Establishment_Year  8523 non-null   int64
 8   Outlet_Size                6113 non-null   object
 9   Outlet_Location_Type       8523 non-null   object
 10  Outlet_Type                8523 non-null   object
 11  Item_Outlet_Sales          8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
```

Categorical Features:

- Item_Identifier
- Item_Fat_Content
- Item_Type
- Outlet_Identifier
- Outlet_Size
- Outlet_Location_Type
- Outlet_Type

```python
# checking for missing values
big_mart_data.isnull().sum()
```

[13] 0.0s

```
Item_Identifier                 0
Item_Weight                  1463
Item_Fat_Content                0
Item_Visibility                 0
Item_Type                       0
Item_MRP                        0
Outlet_Identifier               0
Outlet_Establishment_Year       0
Outlet_Size                  2410
Outlet_Location_Type            0
Outlet_Type                     0
Item_Outlet_Sales               0
dtype: int64
```

```
Handling Missing Values

Mean --> average

Mode --> more repeated value
```

```python
# mean value of "Item_Weight" column
big_mart_data['Item_Weight'].mean()
```

[14] 0.2s

```
np.float64(12.857645184135976)
```

```python
# filling the missing values in "Item_weight column" with "Mean" value
big_mart_data['Item_Weight'].fillna(big_mart_data['Item_Weight'].mean(), inplace=True)
```

[15] 0.2s

```
C:\Users\SRL123\AppData\Local\Temp\ipykernel_11260\2509980927.py:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a cop

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to pe

  big_mart_data['Item_Weight'].fillna(big_mart_data['Item_Weight'].mean(), inplace=True)
```

```python
# mode of "Outlet_Size" column
big_mart_data['Outlet_Size'].mode()
```

[16] 0.0s

```
0    Medium
Name: Outlet_Size, dtype: object
```

```python
# filling the missing values in "Outlet_Size" column with Mode
mode_of_Outlet_size = big_mart_data.pivot_table(values='Outlet_Size', columns='Outlet_Type', aggfunc=(lambda x: x.mode()[0]))
```

[17] 0.2s

```python
print(mode_of_Outlet_size)
```

[18] 0.0s

```
Outlet_Type Grocery Store Supermarket Type1 Supermarket Type2  \
Outlet_Size         Small             Small            Medium

Outlet_Type Supermarket Type3
Outlet_Size            Medium
```

```python
miss_values = big_mart_data['Outlet_Size'].isnull()
```

[19] 0.1s

```python
print(miss_values)
```

```python
print(miss_values)
```

[20] 0.1s

```
0       False
1       False
2       False
3        True
4       False
        ...
8518    False
8519     True
8520    False
8521    False
8522    False
Name: Outlet_Size, Length: 8523, dtype: bool
```

```python
big_mart_data.loc[miss_values, 'Outlet_Size'] = big_mart_data.loc[miss_values,'Outlet_Type'].apply(lambda x: mode_of_Outlet_size[x])
```

[21] 0.4s

```python
# checking for missing values
big_mart_data.isnull().sum()
```

[22] 0.0s

```
Item_Identifier            0
Item_Weight                0
Item_Fat_Content           0
Item_Visibility            0
```

```python
# checking for missing values
big_mart_data.isnull().sum()
```

[22] ✓ 0.0s Python

```
Item_Identifier              0
Item_Weight                  0
Item_Fat_Content             0
Item_Visibility              0
Item_Type                    0
Item_MRP                     0
Outlet_Identifier            0
Outlet_Establishment_Year    0
Outlet_Size                  0
Outlet_Location_Type         0
Outlet_Type                  0
Item_Outlet_Sales            0
dtype: int64
```

Data Analysis

```python
big_mart_data.describe()
```

[23] ✓ 0.3s Python

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	8523.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914

Data Analysis

```python
big_mart_data.describe()
```

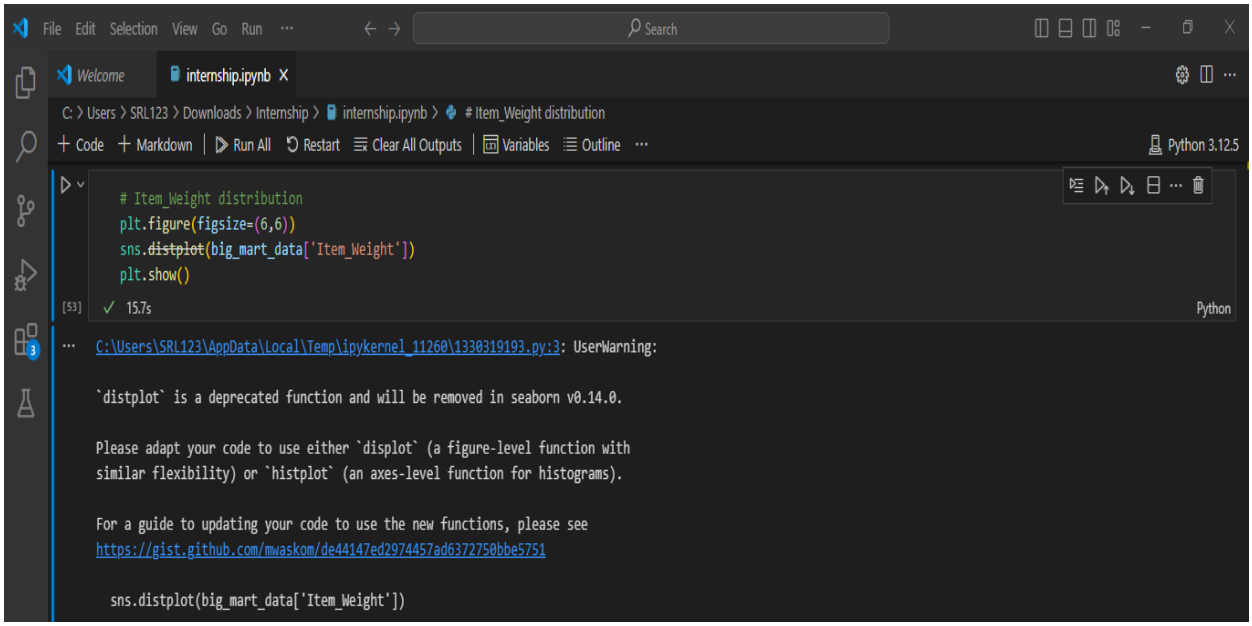
[23] ✓ 0.3s Python

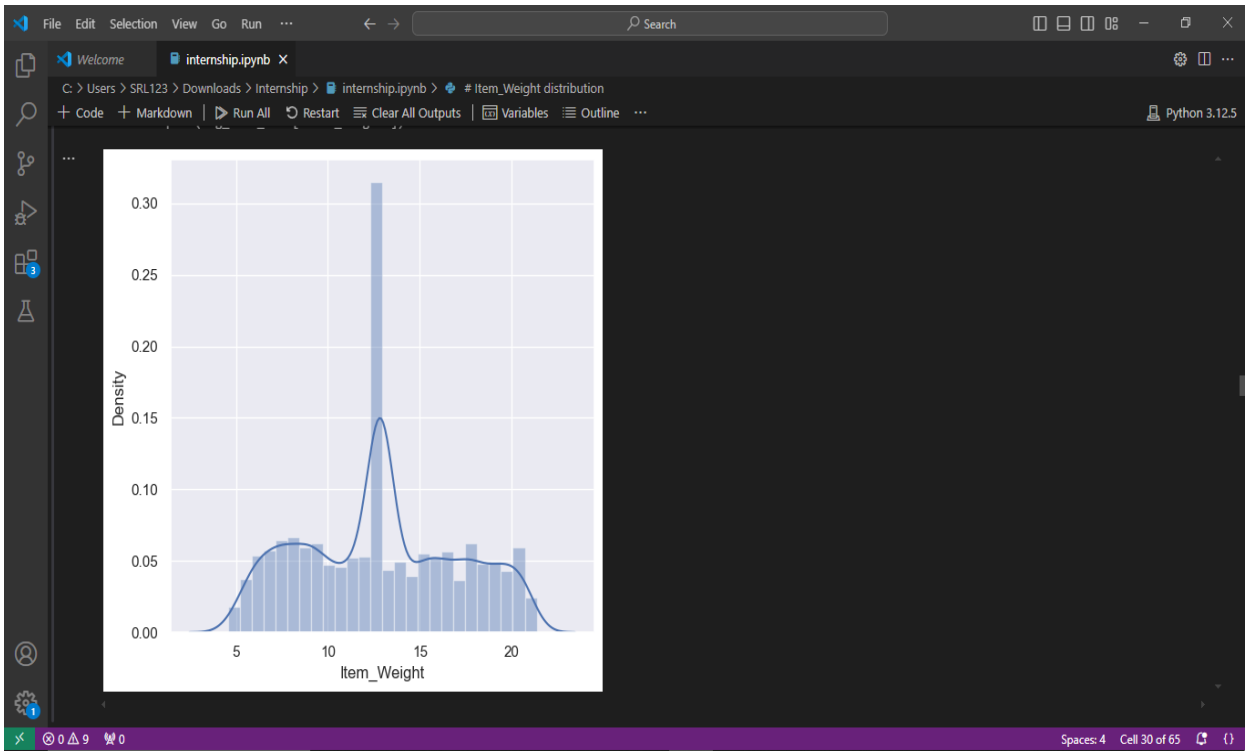
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	8523.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.226124	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	9.310000	0.026989	93.826500	1987.000000	834.247400
50%	12.857645	0.053931	143.012800	1999.000000	1794.331000
75%	16.000000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Numerical Features

```python
sns.set()
```

[24] ✓ 0.0s Python

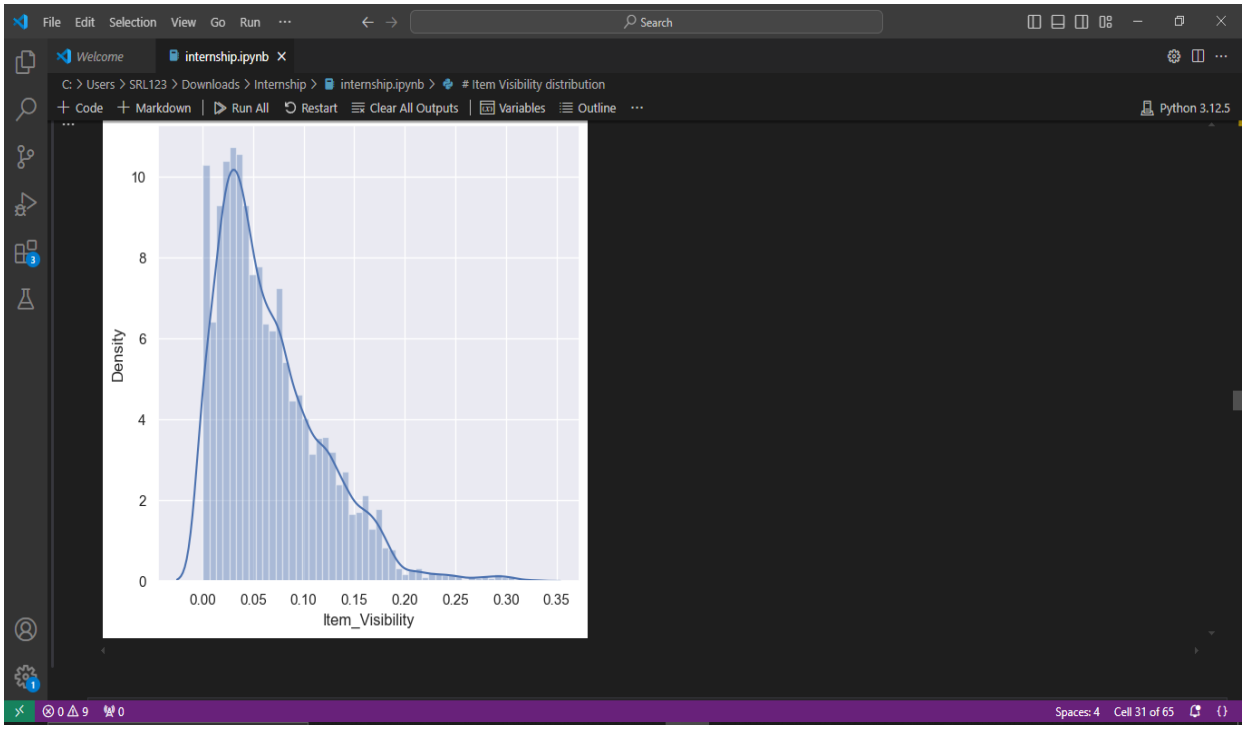




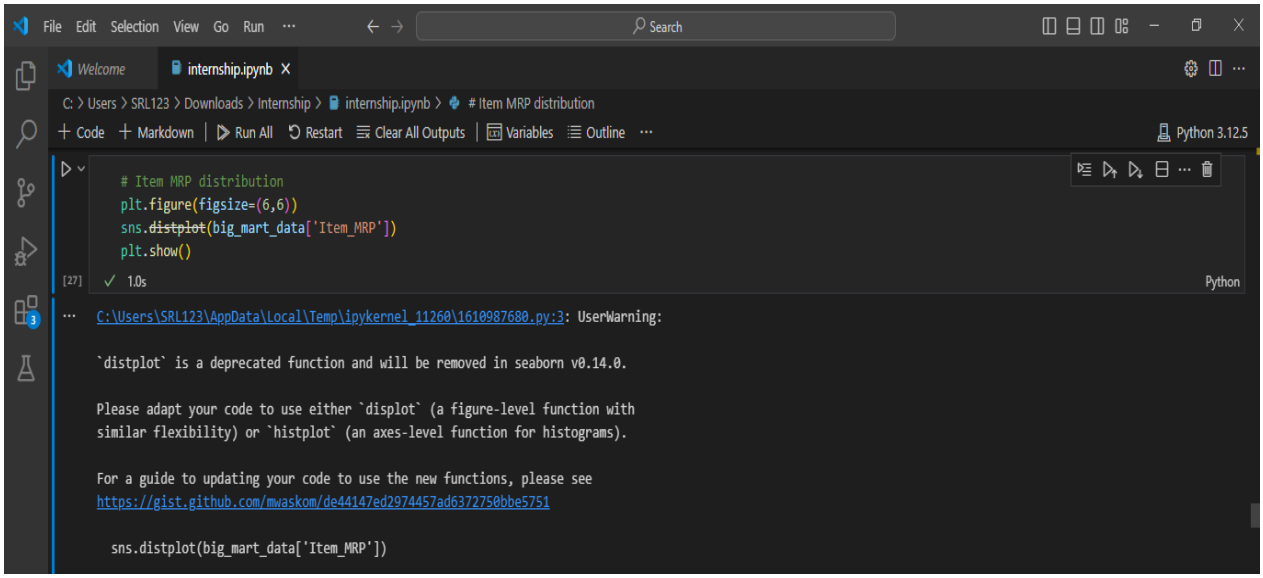
The density plot shows the distribution of **Item_Weight**, with a significant peak around a particular weight value. This suggests that many items have similar weights, while the rest of the weights are more evenly spread out across the range.

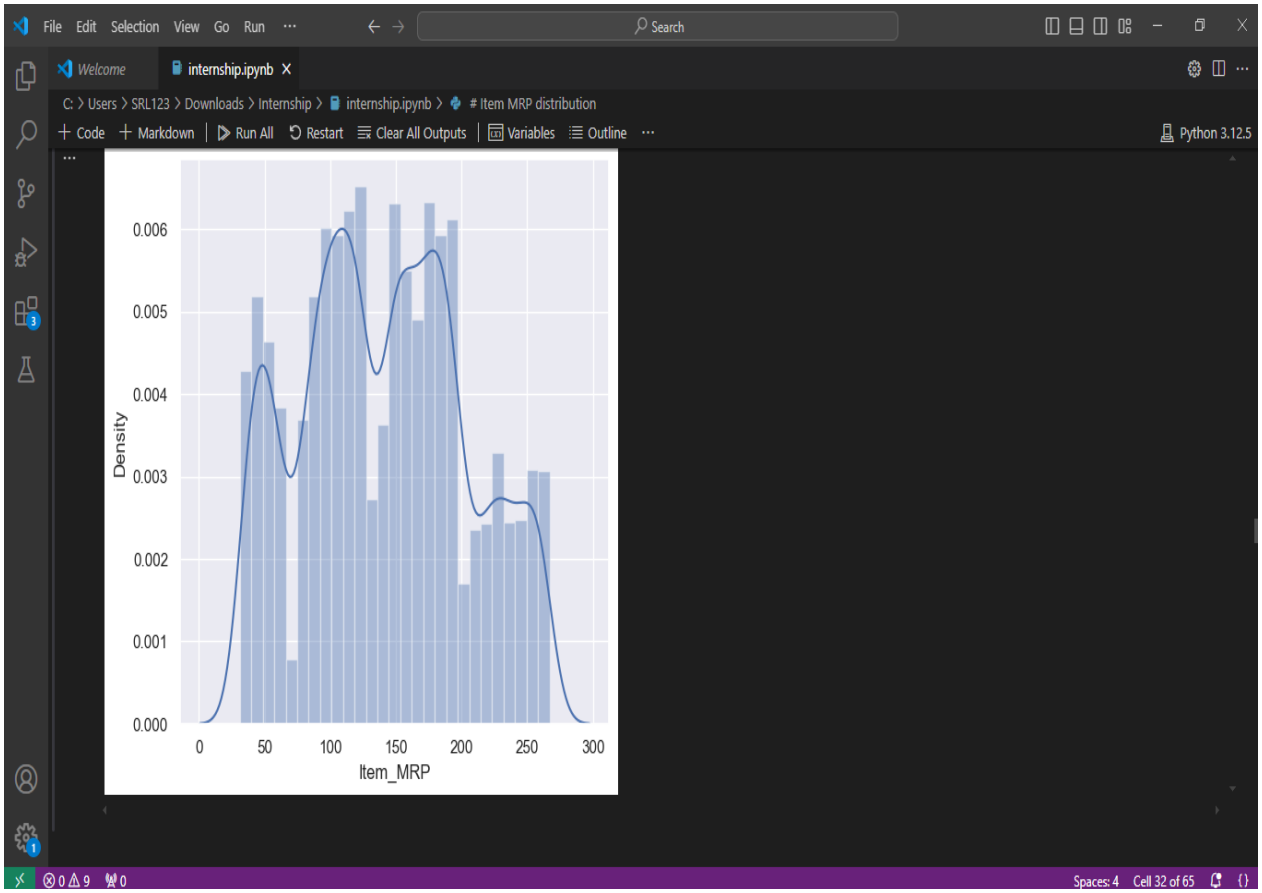
```
# Item Visibility distribution
plt.figure(figsize=(6,6))
sns.distplot(big_mart_data['Item_Visibility'])
plt.show()
```

```
C:\Users\SRL123\AppData\Local\Temp\ipykernel_11260\193435663.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(big_mart_data['Item_Visibility'])
```

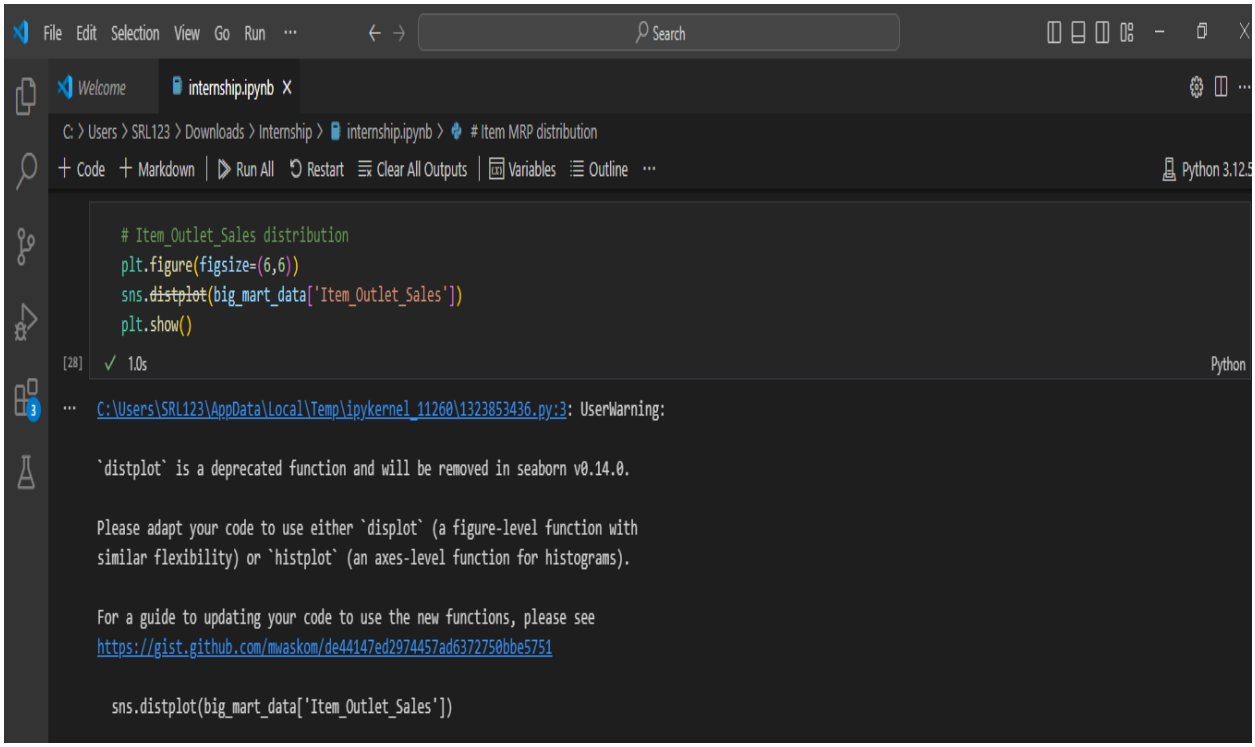


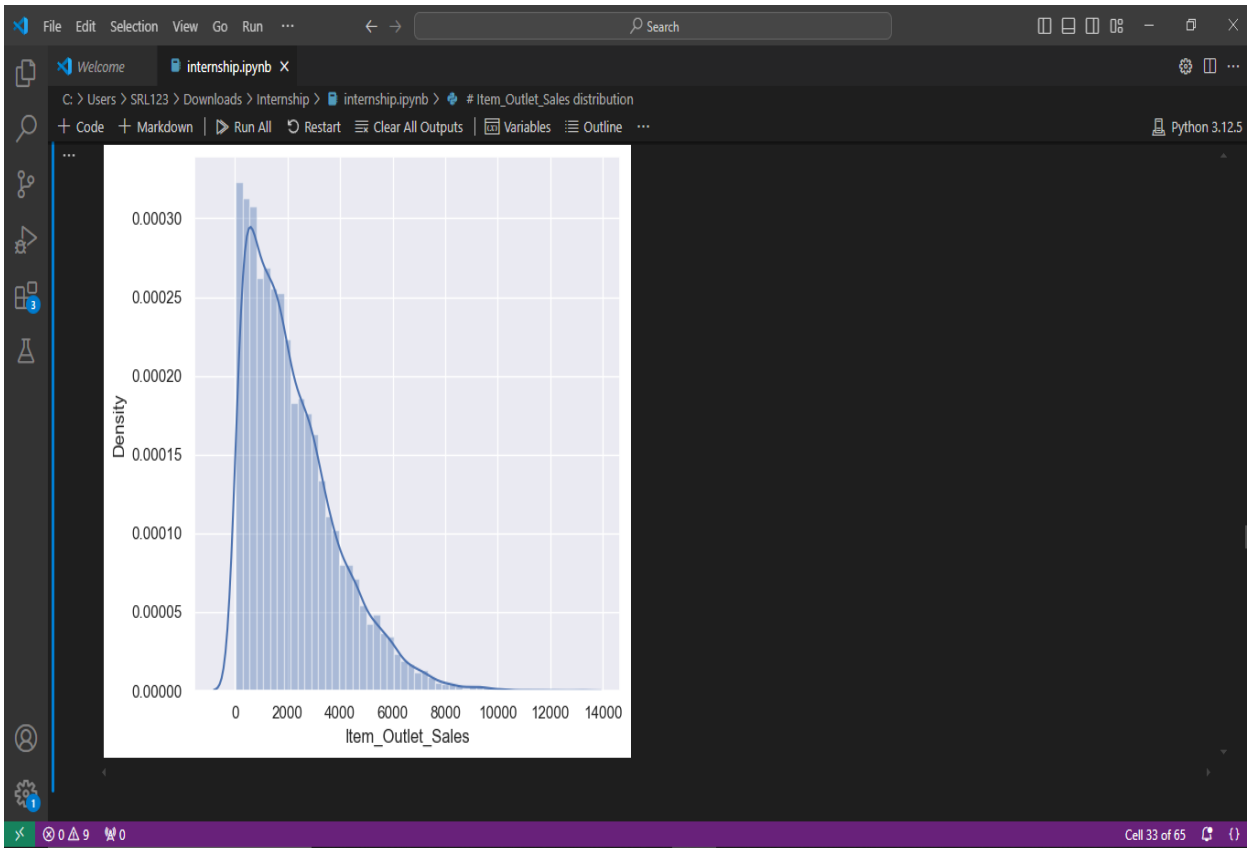
The plot shows that most items have low visibility values, with the density decreasing as visibility increases. This suggests that a majority of items are displayed with minimal visibility in the store.



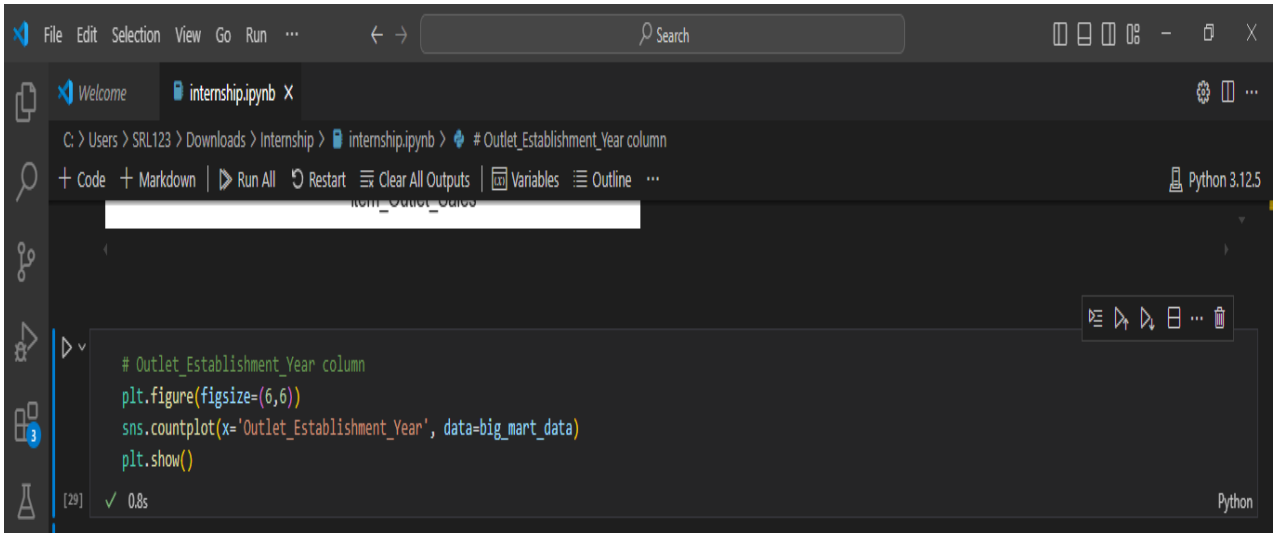


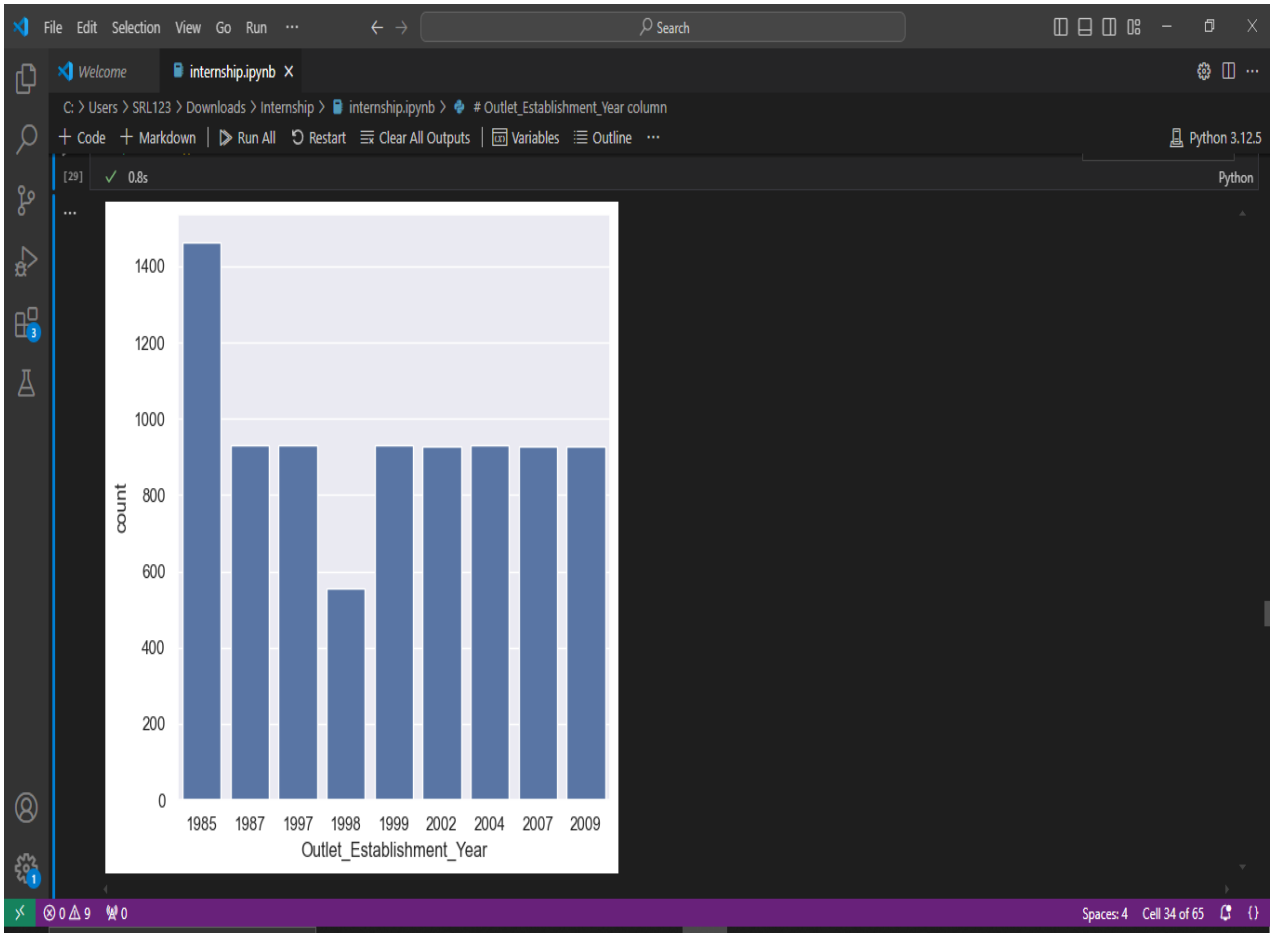
Item MRP Density Plot: The distribution shows common price clusters, with higher density around specific price points.



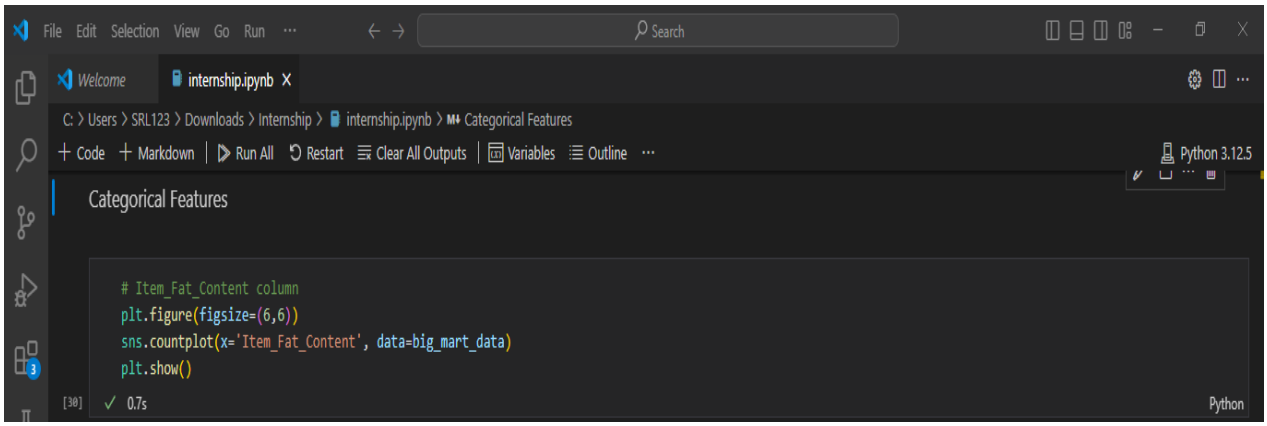


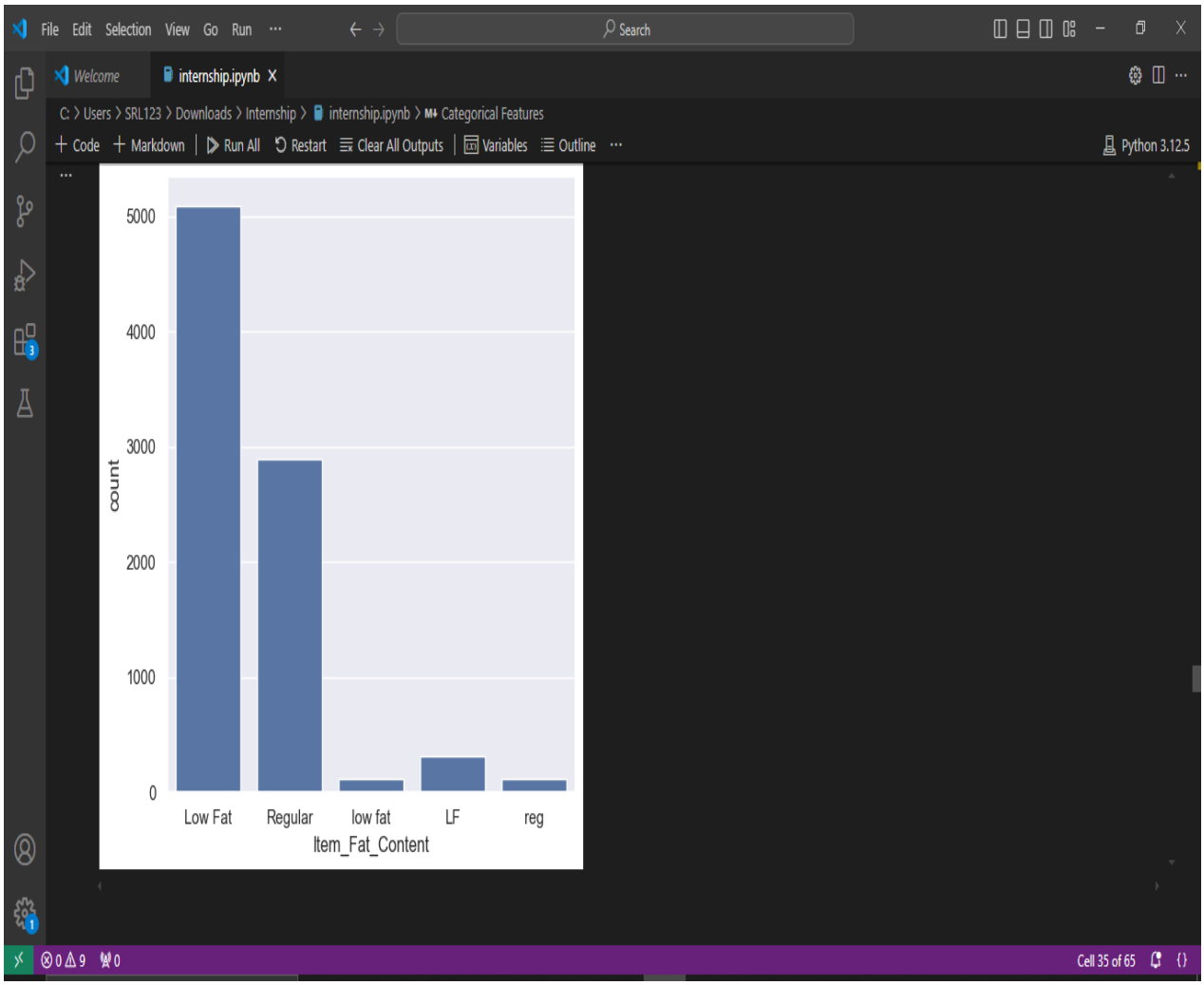
Item Outlet Sales Density Plot: Most items have low sales, with a few high-selling items creating a right-skewed distribution.



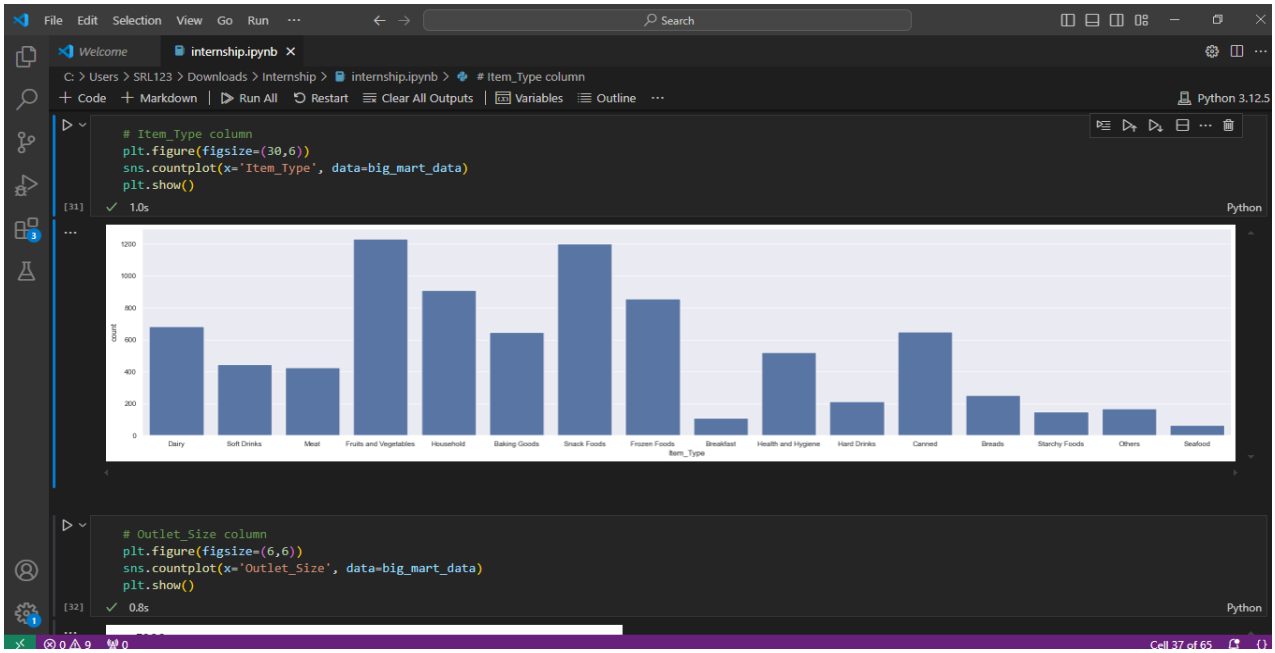


Outlet Establishment Year Count Plot: Most outlets were established in 1985, with fewer outlets added consistently in later years.

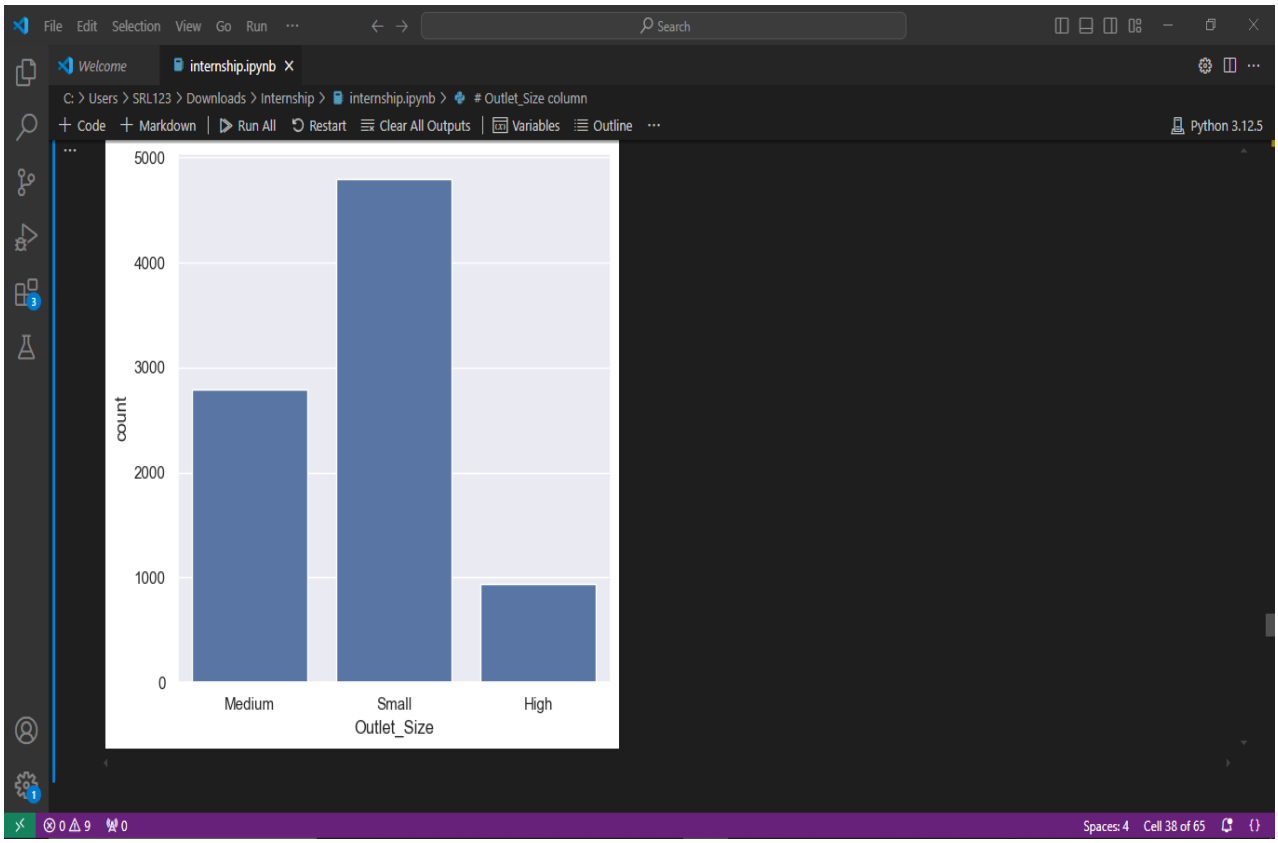




The bar chart shows that "Low Fat" items have the highest count, followed by "Regular" items, with other categories being minimal.



The bar chart displays various "Item Types" with "Fruits and Vegetables" having the highest count, while other item types have varying but lower frequencies.



The bar chart illustrates that "Medium" outlet sizes have the highest count, followed by "Small" and "High" sizes in decreasing order.

File Edit Selection View Go Run ··· Search

Welcome | internship.ipynb

C: > Users > SRL123 > Downloads > Internship > internship.ipynb > big_mart_data.replace({'Item_Fat_Content': {'low fat':'Low Fat','LF':'Low Fat', 'reg':'Regular'}}, inplace=True)

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ··· Python 3.12.5

Data Pre-Processing

```python
big_mart_data.head()
```

[33] ✓ 0.1s Python

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outle
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Super
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Super
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Super
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	Small	Tier 3	G
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Super

File Edit Selection View Go Run ··· Search

Welcome | internship.ipynb

C: > Users > SRL123 > Downloads > Internship > internship.ipynb > big_mart_data.replace({'Item_Fat_Content': {'low fat':'Low Fat','LF':'Low Fat', 'reg':'Regular'}}, inplace=True)

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ··· Python 3.12.5

```python
big_mart_data['Item_Fat_Content'].value_counts()
```

[34] ✓ 0.1s Python

```
Item_Fat_Content
Low Fat    5089
Regular    2889
LF          316
reg         117
low fat     112
Name: count, dtype: int64
```

```python
big_mart_data.replace({'Item_Fat_Content': {'low fat':'Low Fat','LF':'Low Fat', 'reg':'Regular'}}, inplace=True)
```

[35] ✓ 0.1s Python

```python
big_mart_data['Item_Fat_Content'].value_counts()
```

[36] ✓ 0.1s Python

```
Item_Fat_Content
Low Fat    5517
Regular    3006
Name: count, dtype: int64
```

File Edit Selection View Go Run ··· Search

Welcome | internship.ipynb

C: > Users > SRL123 > Downloads > Internship > internship.ipynb > big_mart_data['Item_Fat_Content'].value_counts()

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ··· Python 3.12.5

Label Encoding

```python
encoder = LabelEncoder()
```

[37] ✓ 0.0s Python

```python
big_mart_data['Item_Identifier'] = encoder.fit_transform(big_mart_data['Item_Identifier'])

big_mart_data['Item_Fat_Content'] = encoder.fit_transform(big_mart_data['Item_Fat_Content'])

big_mart_data['Item_Type'] = encoder.fit_transform(big_mart_data['Item_Type'])

big_mart_data['Outlet_Identifier'] = encoder.fit_transform(big_mart_data['Outlet_Identifier'])

big_mart_data['Outlet_Size'] = encoder.fit_transform(big_mart_data['Outlet_Size'])

big_mart_data['Outlet_Location_Type'] = encoder.fit_transform(big_mart_data['Outlet_Location_Type'])

big_mart_data['Outlet_Type'] = encoder.fit_transform(big_mart_data['Outlet_Type'])
```

[38] ✓ 0.1s Python

```python
big_mart_data.head()
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outle
0	156	9.30	0	0.016047	4	249.8092	9	1999	1	0	
1	8	5.92	1	0.019278	14	48.2692	3	2009	1	2	
2	662	17.50	0	0.016760	10	141.6180	9	1999	1	0	
3	1121	19.20	1	0.000000	6	182.0950	0	1998	2	2	
4	1297	8.93	0	0.000000	9	53.8614	1	1987	0	2	

Splitting features and Target

```python
X = big_mart_data.drop(columns='Item_Outlet_Sales', axis=1)
Y = big_mart_data['Item_Outlet_Sales']
```

```python
print(X)
```

```
      Item_Identifier  Item_Weight  Item_Fat_Content  Item_Visibility  \
0                 156        9.300                 0         0.016047
1                   8        5.920                 1         0.019278
2                 662       17.500                 0         0.016760
3                1121       19.200                 1         0.000000
4                1297        8.930                 0         0.000000
...               ...          ...               ...              ...
8518              370        6.865                 0         0.056783
8519              897        8.380                 1         0.046982
8520             1357       10.600                 0         0.035186
8521              681        7.210                 1         0.145221
8522               50       14.800                 0         0.044878

      Item_Type  Item_MRP  Outlet_Identifier  Outlet_Establishment_Year  \
0             4  249.8092                  9                       1999
1            14   48.2692                  3                       2009
2            10  141.6180                  9                       1999
3             6  182.0950                  0                       1998
4             9   53.8614                  1                       1987
...         ...       ...                ...                        ...
8518         13  214.5218                  1                       1987
8519          0  108.1570                  7                       2002
8520          8   85.1224                  6                       2004
8521         13  103.1332                  3                       2009
8522         14   75.4670                  8                       1997
...
8521          1                  2            2
8522          2                  0            1
```

Splitting the data into Training data & Testing Data

```python
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```python
print(X.shape, X_train.shape, X_test.shape)
```

```
(8523, 11) (6818, 11) (1705, 11)
```

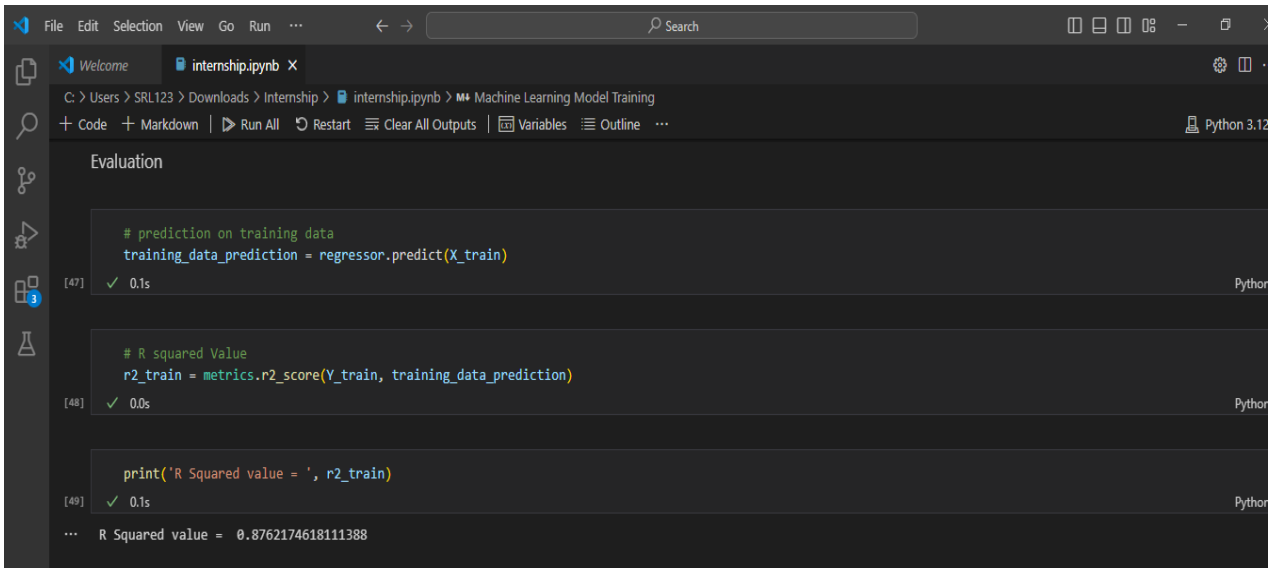
Machine Learning Model Training

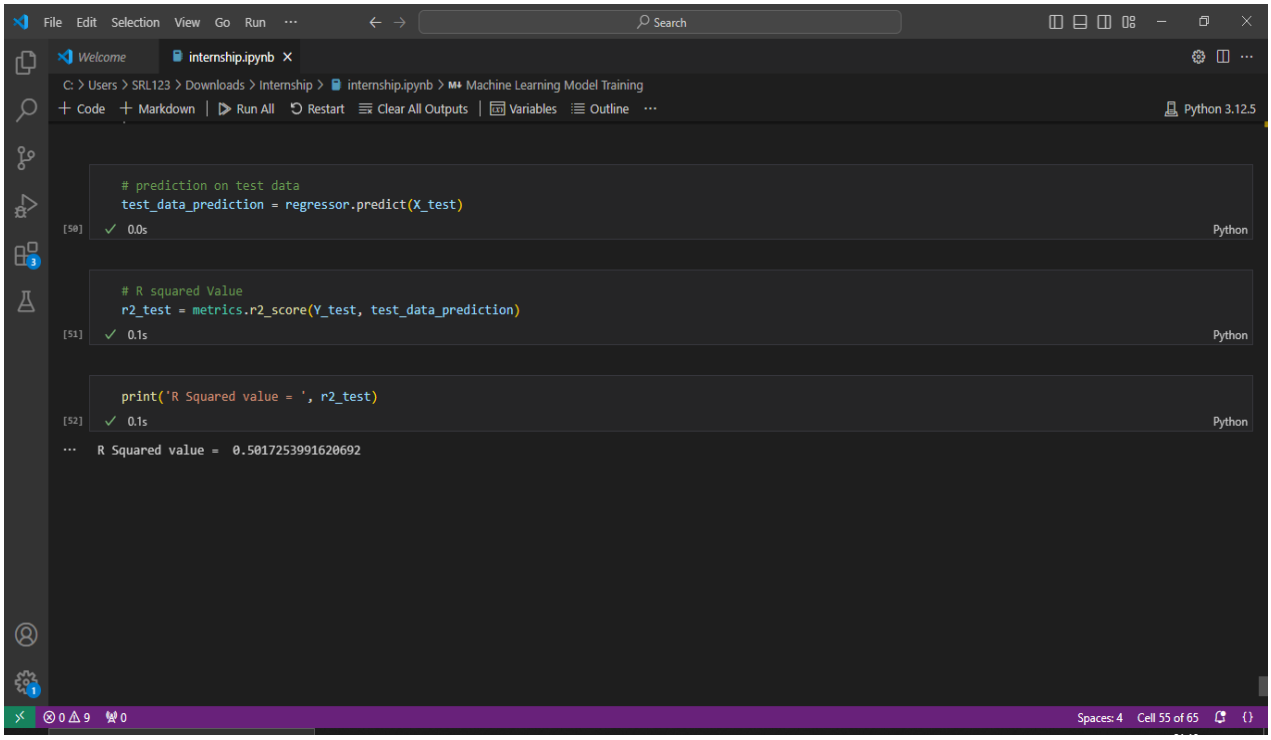
XGBoost Regressor

```python
regressor = XGBRegressor()
```

```python
regressor.fit(X_train, Y_train)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=None, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=None, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=None, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
```





The model's R-squared value of **0.5017** indicates that it explains about 50.17% of the variability in the target variable, showing moderate predictive accuracy.

CHAPTER-V

LEARNING OUTCOMES, CHALLENGES FACED AND RECOMMENDATIONS

5.1LEARNING OUTCOMES

1. **Data Pre-processing and Cleaning:** Gained hands-on experience in handling missing values, encoding categorical variables, and normalizing data for model training.

2. **Feature Engineering:** Learned the importance of feature selection and transformation to enhance model performance and accuracy.

3. **Model Evaluation and Tuning:** Acquired skills in evaluating models using metrics like R-squared, and learned how hyper parameter tuning impacts predictive accuracy.

4. **Practical Application of ML Algorithms:** Applied regression models (e.g., Linear Regression, Decision Trees) to predict sales, understanding their advantages and limitations.

5. **Data Interpretation for Business Insights:** Developed the ability to interpret model results to derive insights that can guide business strategies, such as which factors drive sales in different outlets.

5.2CHALLENGES FACED

1. **Data Quality and Missing Values:** Dealing with incomplete data and ensuring consistency across features required careful pre-processing.

2. **Low Model Accuracy:** Initial models had low predictive power, indicating the need for feature engineering and tuning to improve results.

3.**Overfitting and Under fitting:** Striking a balance between complex models and generalizability was challenging, particularly with limited data for validation.

4.**Feature Selection and Encoding:** Identifying the most impactful features among numerous categories and variables required iterative testing and analysis.

5. **Computational Constraints:** Running complex models and tuning parameters was computationally intensive, especially for large datasets.

5.3 RECOMMENDATIONS

1. **Add More Features:** Incorporate additional variables, such as customer demographics or regional economic data, to improve prediction accuracy.

2. **Use Advanced Models:** Explore more complex algorithms like Random Forests, Gradient Boosting, or Neural Networks, which may capture non-linear relationships better.

3. **Optimize Hyperparameters:** Systematically tune model parameters using techniques like Grid Search or Randomized Search for better performance.

4. **Cross-Validation:** Implement cross-validation techniques to improve model robustness and reduce the risk of overfitting.

5. **Data Enrichment:** Gather more data to better represent seasonal trends or regional differences in sales patterns, providing a richer context for predictions