*An Internship Report on*

# DATA SCIENCE

*Submitted in partial fulfillment of the requirements for the award of the degree of*

### BACHELOR OF TECHNOLOGY

*in*

### COMPUTER SCIENCE & ENGINEERING

*from*

### JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

By

**Bayyapu Srilekha**(21C11A05G0)

Under the guidance of

**Mrs. M. Anusha Ph.D**.

Assistant Professor

|  |
| --- |
| **DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING** |

|  |  |  |
| --- | --- | --- |
| anurag logo | **ANURAG ENGINEERING COLLEGE**  **(An Autonomous Institution)**  (Affiliated to JNTUH, Hyderabad & Approved by AICTE, New Delhi)  Ananthagiri (V & M), Kodad, Suryapet (Dt.), Telangana -508206.  **2024 – 25** |  |

**ANURAG ENGINEERING COLLEGE**

**(An Autonomous Institution)**

(Affiliated to JNTUH, Hyderabad & Approved by AICTE, New Delhi)

Ananthagiri (V & M), Kodad, Suryapeta (Dt), Telangana -508206.

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**



**CERTIFICATE**

*This is to certify that the project work entitled “****DATA SCIENCE****” is a Bonafide work done by “* ***Bayyapu Srilekha(21C11A05G0)****” in the partial fulfillment for the award of Bachelor of Technology in Computer Science & Engineering from JNTUH, Hyderabad during the year* ***2024-25****.*

*This work has not been submitted to any other university or institute or organization for the award of any degree or diploma.*

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Mrs.M.Anusha  **Supervisor** |  | (Dr. Y.V.R. Naga Pawan)H.O.D. |
|  |  |  |
|  |  |  |
| **External Examiner** |  | Dr. T. Suresh Kumar  **Principal** |

###### 

**Internship Certificate**

###### 

# ACKNOWLEDGEMENTS

This report will certainly not be completed without due acknowledgement paid to all those who helped me during this project wo

I would like to express my gratefulness to the project trainer **Mr.Uttam Grade, Embrizon Technologies, Hyderabad** for the valuable guidance in carrying out this project work.

I express sincere thanks to my supervisor **Mrs.M.Anusha,** for giving moral support, kind attention and valuable guidance throughout this project work.

It is my privilege to thank **Dr. Y. V. R. Naga Pawan***,* Head of the Department, for his encouragement during the progress of this project work.

I am thankful to both Teaching and non-teaching staff members of **Department of Computer Science & Engineering** for their kind cooperation and all sorts of support in bringing out this project work successfully.

I derive great pleasure in expressing my sincere gratitude to the principal **Dr. T. Suresh Kumar** for his timely suggestions, which helped me to complete this work successfully.

I am thankful to the **Management** of **ANURAG Engineering College** for providing required facilities during the project work.

I would like to thank my parents and my friends for being supportive all the time, and I am very much obliged to them.

**Bayyapu Srilekha**

(21C11A05G0)

**CONTENTS**

|  |  |
| --- | --- |
| Problem Statement | 1 |
| List Of Figures | 2 |
| List Of Tables | 3 |
| List of Acronyms | 4 |
| Abstract | 5 |
| CHAPTER 1  Introduction | 6-19 |
| CHAPTER 2  About The Organization | 20-21 |
| CHAPTER 3  Objectives | 22-23 |
| CHAPTER 4  System Architecture | 24-25 |
| CHAPTER 5  Technologies used  5.1 Traditional machine Learning  5.2 Libraries and Tools used  5.2.1 python Libraries | 26 |
| CHAPTER 6  Software Requirements  6.1 python  6.2 why python data science | 27-28 |
| CHAPTER 7  Hardware Requirements | 29 |
| CHAPTER 8  Frontend and Backend Requirements | 30-31 |
| CHAPTER 9  Modules | 32-34 |
| CHAPTER 10  Challenges and solutions  Result and Discussion | 35-44 |
| Internship Highlights  Key takeaways and Impact of the Internship  Conclusion and References | 45-48 |

**Abstract**

The Big Mart Sales Prediction project aims to forecast sales by leveraging historical sales data and employing various ML(machine learning) techniques, with a primary focus on linear regression. The project involves data collection and preprocessing, including cleaning, handling missing values, and feature engineering to create relevant predictors. (LR)Linear regression is utilized to model the relationship between features. The model is trained on the processed dataset, and its performance is evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-square(R2) values. The ultimate goal is to provide actionable insights for inventory management and sales optimization, thereby enhancing the overall profitability of Big Mart.

**Problem Statement**

Sales Prediction Project Idea: The Big Mart sales dataset is a treasure trove of learning opportunities. It consists of 2013 sales data for 1559 products across ten outlets in different cities.my goal in this ML projects to build a regression model that can predict the sales of each these 1559 products for the following year in each of the 10 different big mart outlets. the dataset also includes specific attributes for each product and store, providing valuable insights into the factors influencing sales.

## List Of Figures

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Description** | **Page No** |
| Fig 1 | Steps for sales prediction | 15 |
| Fig 2 | System architecture | 24 |
| Fig 3 | Correlation matrix | 38 |
| Fig 4 | Prediction accuracy | 39 |
| Fig 5 | Bar graph of item\_ outlet\_ ales vs outlet \_identifier | 40 |
| Fig 6 | Bar graph of item \_outlet \_sales vs outlet \_size | 41 |
| Fig 7 | Bar graph of item \_outlet \_sales vs outlet \_type | 42 |
| Fig 8 | Bar graph of item \_outlet \_sales vs outlet \_year | 43 |
| Fig 9 | Scatter plot of item \_outlet \_sales vs item \_mrp | 44 |
| Fig 10 | Scatter plot of item \_outlet sales vs item \_visibility | 45 |

**List Of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Description** | **Page No.** |
| 1 | Dataset | 18 |
| 2 | Hardware Requirements | 29 |

|  |  |
| --- | --- |
|  |  |
|  | **CHAPTER 1**  **INTRODUCTION** |
| **1.1** | **1.1 DATA SCIENCE** |
|  | Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. In Wikipedia, Data Science is defined as a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.1.2 why Data Science?Data is the oil for today’s world. With the right tools,technologies, algorithms, we can use data and convert it into a distinct business advantageData Science can help us to detect fraud using advancedmachine learning algorithmsAllows to build intelligence ability in machinesIt enables us to take better and faster decisionsIt helps us to recommend the right product to the right customer to enhance your business1.3 why it is important?Data science helps brands to understand their customers in a much enhanced and empowered manner. Customers are the soul and base of any brand and have a great role to play in their success and failure. With the use of data science, brands can connect with their customers in a personalized manner, thereby ensuring better brand power and engagement.One of the reasons why data science is gaining so much of attention is because it allows brands to communicate their story in such a engaging and powerful manner. When brands and companies utilize this data in a comprehensive manner, they can share their story with their target audience, thereby creating better brand connect. After all, nothing connects with consumers like an effective and powerful story, that can inculcate all human emotions.Big Data is a new field that is constantly growing and evolving. With so many tools being developed, almost on a regular basis, big data is helping brands and organizations to solve complex problems in IT, human resource , and resource management in an effective and strategic manner. This means effective use of resources, both material and non-material. |
|  | **1.4 Core Components of Data Science :**   1. **Data Collection**:   Gathering raw data from various sources like databases, APIs, or web scraping. This step is essential for ensuring the dataset's relevance and quality.   1. **Data Preprocessing:**   Cleaning and transforming raw data by handling missing values and normalizing formats. This prepares the dataset for accurate analysis and modeling.   1. **Feature Selection:**   Identifying and selecting the most relevant features that contribute to model performance. This reduces dimensionality and helps prevent overfitting.   1. **Model Building:**   Creating or Building models using selected algorithms and training them on the prepared dataset for optimal results.   1. **Model Evaluation:**   Assessing model performance using metrics like accuracy and RMSE on a validation dataset. This determines how well the model generalizes to new data.     1. **Data Visualization:**   Using graphical representations to present data and model results clearly. This aids in understanding patterns and facilitating informed decision-making.    As the demand for data-driven insights continues to rise, the role of data scientists has become one of the most sought-after professions in the job market. Organizations across sectors, from healthcare to finance, are investing in data science initiatives to harness the power of their data. This trend is expected to grow, as businesses increasingly recognize the importance of data science in driving innovation and achieving long-term success.  BigMart is a popular retail chain in India that operates multiple stores across the country. With the increasing competition in the retail industry, it is essential for BigMart to optimize its sales and revenue. Sales prediction is a critical component of retail analytics that helps businesses like BigMart to forecast future sales and make informed decisions about inventory management, pricing, and marketing strategies. Accurate sales prediction can help BigMart to reduce stockouts, overstocking, and waste, resulting in significant cost savings and improved customer satisfaction.  The Bigmart Sales Prediction project aims to develop a linear regression model that can accurately predict the sales of items at Bigmart outlets based on various item attributes and outlet characteristics. The project involves a comprehensive analysis of the dataset, feature engineering, model development, and evaluation. The goal is to identify the most important features that influence sales and develop a model that can provide accurate sales forecasts, enabling Bigmart to optimize its inventory management, reduce stock outs and overstocking, and enhance customer satisfaction.  The project will involve several steps, including data preprocessing, feature engineering, model development, model training, and prediciton. The dataset will be preprocessed to handle missing values, outliers, and categorical variables. Feature engineering techniques such as more related features will be applied to improve the model's performance. . The dataset will be split into training and testing sets, and a linear regression model will be developed using the training set. The model will be evaluated using metrics such as mean squared error (MSE) and coefficient of determination (R-squared) to assess its performance. Feature importance will be analyzed to identify the most important features that influence sales. |
|  | **1.5 Sales prediction:**  A sales forecast is an expression of expected sales revenue. A sales forecast estimates how much your company plans to sell within a certain time period (like quarter or year). The best sales forecasts do this with a high degree of accuracy, and they’re only as accurate as the data that fuels them.  The daily competition between different malls as well as big malls is becoming more and  more intense because of the rapid rise of international supermarkets and online shoppings.  Every mall or mart tries to provide personal and short-term donations or benefits to attract  more and more customers on a daily basis, such as the sales price of everything which is  usually predicted to be managed through different ways such as corporate asset management,  logistics, and transportation service, etc. Current machine learning algorithms that are very  complex and provide strategies for predicting or predicting long-term demand for a company's  sales, which now also help in overcoming budget and computer programs.  we basically discuss the subject of specifying a large mart sale or predicting an item for a customer’s future need in a few supermarkets in various locations and products that support the previous record. Various ML algorithms such as linear regression, random forest, etc. are used to predict sales volume.  Regular sales forecasting research can help in-depth analysis of  pre-existing conditions and conditions and then, assumptions are often used in terms of  customer acquisition, lack of funding, and strength before setting budgets and marketing plans  for the coming year. |
|  |  |
|  |  |
|  | Fig 1:Steps for Sales prediction |
|  |  |
|  | **1.6 Regression:**  Regression is a statistical method used to establish a relationship between two or more variables, where the goal is to predict the value of one variable based on the values of others. In other words, regression analysis helps us understand how changes in one or more independent variables (predictors) affect a dependent variable (outcome or response).  **1.6.1 Types of Regression:**   1. **Linear regression**: It is a fundamental machine learning algorithm used to predict a continuous output variable based on one or more input features.   There are two types of linear regressions they are:   1. **Simple Linear Regression:** It is a type of regression where only one independent variable is used to predict the dependent variable. 2. **Multiple Linear Regression:** Multiple linear regression is a type of regression where more than one independent variable is used to predict the dependent variable.   **2. Non-Linear Regression**: Non-linear regression is a type of regression where the relationship between the independent and dependent variables is not linear.  **3. Logistic Regression:** Logistic regression is a type of regression used for binary classification problems, where the dependent variable is a binary outcome  (0 or 1, yes or no, etc.).  **4. Ridge Regression:** Ridge regression is a type of regression that is used to handle multicollinearity between the independent variables.  **5. Lasso Regression:** It is a type of regression that is used to handle multicollinearity between the independent variables by adding a penalty term to the cost function  **1.6.2 Linear regression:**  **DEF:** It is a type of supervised learning algorithm that predicts a continuous output variable based on one or more input features. The goal of linear regression is to create a linear equation that best predicts the value of the target variable based on the input features. In other words, it tries to find the best-fitting linear line that minimizes the difference between the predicted and actual values. There are two types of linear regressions they are: |
|  |  |
|  | As we know Regression can be termed as a parametric technique which means we can predict a continuous or dependent variable on the basis of a provided datasets of independent variables.  The Equation of simple LR is:  **Y = βo + β1X + ∈------------ (1)**  where,  **Y :** It is basically the variable which we used as a predicted value.  **X :** It is a variable(s) which is used for making a prediction.  **βo :** It is said to be a prediction value when **X=0**.    **β1 :** when there is a change in **X** value by **1** unit then **Y** value is also changed. It can also be said as slope term ∈.   1. **Simple Linear Regression**   Simple linear regression is a type of linear regression that involves only one input feature. It is used to predict a continuous output variable based on a single input feature. The equation for simple linear regression is: y = β0 + β1x + ε, where y is the target variable, x is the input feature, β0 is the intercept or bias term, β1 is the slope coefficient, and ε is the error term. The goal is to find the best values for β0 and β1 that minimize the error term.   1. **Multiple Linear Regression**   Multiple linear regression is a type of linear regression that involves more than one input feature. It is used to predict a continuous output variable based on multiple input features. The equation for multiple linear regression is:  y = β0 + β1x1 + β2x2 + … + βnxn + ε, where y is the target variable, x1, x2, …, xn are the input features, β0 is the intercept or bias term, β1, β2, …, βn are the slope coefficients, and ε is the error term. The goal is to find the best values for β0, β1, β2, …, βn that minimize the error term.  **DataSet:** The dataset for this project consists of 8523 item-outlet combinations, each represented by 12 features, including**Table 1: Dataset** |
|  |  |
|  |  |
|  | **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.)  Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis Split into training and evaluation sets.  **Model Selection**:  We used decision tree regression machine learning algorithm , We got a accuracy of 95.7% on test set so we implemented this algorithm.  **Decision tree regression**  Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either: Conditions [Decision Nodes] Result [End Nodes] The branches/edges represent the truth/falsity of the statement and take makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:  **Decision Tree Regression:**  Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.  **Analyze and Prediction**:  In the actual dataset, we chose only 9 features: 1.ItemWeight ---- 2.ItemFatContent ---- Weight Whether the product of is low fat product or not 3.ItemVisibility ---- The % of the total display area of all products in a store allocated to 4.ItemType ---- 5.ItemMRP ---- The the category to particular which the product product belongs Maximum Retail Price (list price) of the product 6.OutletEstablishmentYear ---- The year in which the store was established 7.OutletSize ---- The size of the store in terms of ground area covered 8.OutletLocationType ---- The type of city in which the store is located 9.\*Outlet Type ---- Whether the outlet is just a grocery store or some sort of supermarket  **Accuracy on test set:**  We got an accuracy of 95.80% on test |
|  |  |
|  |  |

**CHAPTER 2**

##### **ABOUT THE ORGANIZATION**

**Name of the Organization:** Embrizon Technologies

**CEO:** Prabhakar Samadder

**Establishment year:** 2023

Embrizon is a company that provides industrial training, virtual industrial training, and paid internships in India. It is headquartered in Hyderabad.

Embrizon offers training in a variety of technologies, including artificial intelligence, machine learning, data science, and cloud computing. The company also offers placement assistance to its students.

Embrizon has been the subject of some controversy, with some people questioning the legitimacy of its programs. However, the company has defended its programs and its partnerships with Microsoft and IIT Guwahati.

Here is some additional information about Embrizon:

The company was founded in 2023.It is a limited liability partnership firm. Its designated partners are Prabhakar samadder and Jakeer shaik. Its last financial year end date was sep 13, 2023.Its current status is active. Embrizon is an EdTech company that provides online learning experiences, focusing on industrial training, live projects, and internships. Here are some key details:

**2.1 Core Offerings:**

**Live Industrial Training:** Offers hands-on training in various technologies like AI, ML, Data Science, Cloud Computing, etc.

**Live Projects:** Provides opportunities to work on real-world projects with industry experts.

**Internships:** Offers paid internships to gain practical experience and build a professional network.

**Placement Assistance:** Helps students secure jobs with top companies.

**Unique Features:**

**Industry-Aligned Curriculum:** The courses are designed to meet the latest industry standards.

**Expert Mentorship:** Students are mentored by experienced industry professionals.

**Flexible Learning:** Offers flexible learning options to accommodate different schedules

**Community-Driven Learning:** Fosters a strong learning community among students.

**2.2 Company Details:**

**Incorporation:** Incorporated as a Limited Liability Partnership (LLP) in 2023.

**Headquarters:** Hyderabad, India.

**Partnerships:** Collaborates with industry leaders like Microsoft and IIT Guwahati.

**Team:** A team of experienced educators, industry experts, and career counselors.

Overall, Embrizon appears to be a legitimate company that offers valuable training and internship opportunities.

**CHAPTER 3**

**OBJECTIVES**

* Predicting future sales.
* Identifying key factors influencing sales.
* Developing a robust and scalable model.
* Improving business decision-making.
* Evaluating model performance.
* Monitoring and updating the model.
* Providing insights and recommendations.

**1. Predicting Future Sales**

Accurate sales prediction is crucial for inventory management, financial planning, and marketing strategies. By leveraging historical sales data and various influencing factors, the model aims to forecast future sales trends, enabling the business to align its resources and strategies effectively.

**2. Identifying Key Factors Influencing Sales:**

Understanding the key drivers of sales helps businesses tailor their marketing efforts and product offerings. The project involves analyzing various features such as product type, store location, seasonality, and promotional activities to determine which factors significantly impact sales performance.

**3. Developing a Robust and Scalable Model:**

Creating a model that not only performs well on existing data but can also adapt to new data is essential for long-term success. The project focuses on selecting appropriate algorithms, optimizing hyperparameters, and ensuring that the model can handle increasing data volumes without compromising performance.

**4. Improving Business Decision-Making:**

The insights derived from the sales prediction model can guide strategic decisions across various departments, including marketing, supply chain, and finance. By providing data-driven recommendations, the model empowers stakeholders to make informed choices that enhance operational efficiency and profitability.

**5. Evaluating Model Performance:**

Regular assessment of the model's accuracy and reliability is vital to ensure it meets business objectives. This involves using various performance metrics, such as RMSE and R-squared, along with cross-validation techniques to validate the model's predictions against actual sales data.

**6. Monitoring and Updating the Model:**

Sales patterns can change due to market dynamics, consumer behavior, or external factors, necessitating ongoing model maintenance. The project includes a framework for continuous monitoring of model performance and periodic updates to incorporate new data and refine predictions, ensuring sustained accuracy.

**7. Providing Insights and Recommendations:**

Beyond just predicting sales, the project aims to deliver actionable insights that can drive business strategies. By analyzing the results and trends identified by the model, stakeholders can receive tailored recommendations on inventory management, pricing strategies, and promotional The insights derived from the sales prediction model can guide strategic decisions across various departments, including marketing, supply chain, and finance. By providing data-driven recommendations, the model empowers stakeholders to make informed choices that enhance operational efficiency and profitability.

**5. Evaluating Model Performance:**

Regular assessment of the model's accuracy and reliability is vital to ensure it meets business objectives. This involves using various performance metrics, such as RMSE and R-squared, along with cross-validation techniques to validate the model's predictions against actual sales data.

**6. Monitoring and Updating the Model:**

Sales patterns can change due to market dynamics, consumer behavior, or external factors, necessitating ongoing model maintenance. The project includes a framework for continuous monitoring of model performance and periodic updates to incorporate new data and refine predictions, ensuring sustained accuracy.

**7. Providing Insights and Recommendations:**

Beyond just predicting sales, the project aims to deliver actionable insights that can drive business strategies. By analyzing the results and trends identified by the model, stakeholder. **CHAPTER 4**

**SYSTEM ARCHITECTURE**

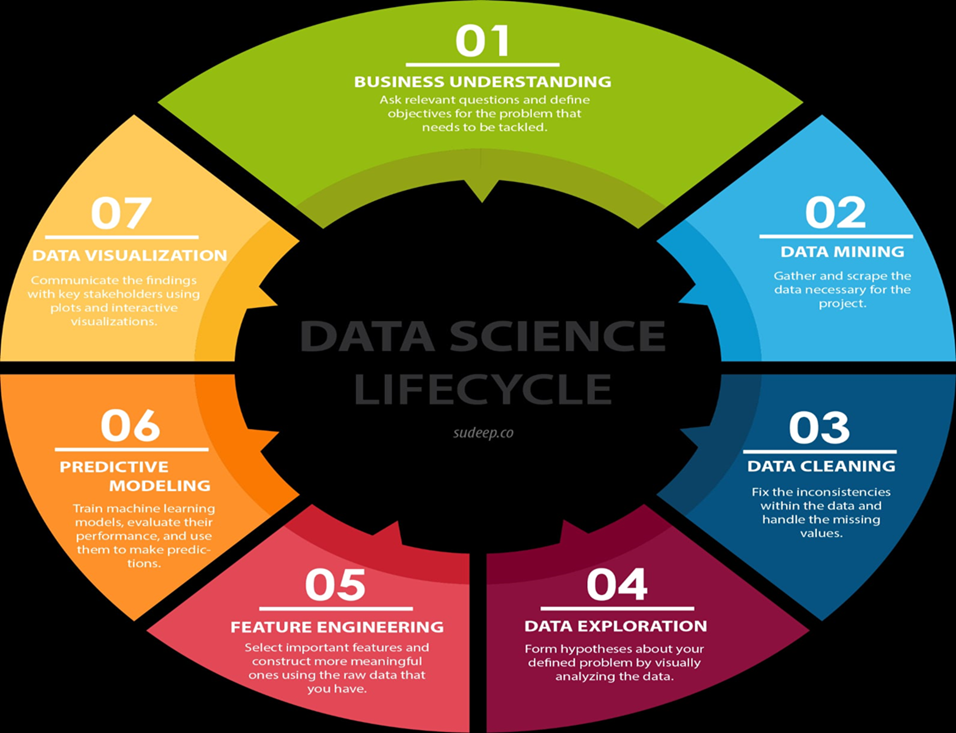


fig 2:System Architecture

* **Business Understanding:** The first step is to understand the business problem you are trying to solve. This involves asking relevant questions and defining clear objectives.
* **Data Mining:** Once you have a clear understanding of the problem, you need to gather and collect the data needed to solve it. This step involves data mining, where you gather and scrape the relevant data from various sources.
* **Data Cleaning:** After you have collected the data, you need to clean it. This involves fixing inconsistencies, handling missing values, and preparing the data for further analysis.

* **Data Exploration:** Once your data is clean, you can begin exploring it. This involves visually analyzing the data to gain insights and form hypotheses about your problem.
* **Feature Engineering:** In this step, you select important features from your data and construct new ones that are more meaningful. This process is crucial for building effective machine learning models.
* **Predictive Modeling:** Now, you can start building machine learning models to predict the outcome you are interested in. This involves training models, evaluating their performance, and using them to make predictions.
* **Data Visualization:** The final step is to communicate your findings to stakeholders using plots and interactive visualizations. This helps them understand the insights you have gained from your analysis and make informed decisions**.**

**CHAPTER 5**

**TECHNOLOGIES USED**

Here are some of the key technologies used in sales prediction:

**5.1 Traditional Machine Learning:**

**Random Forest:** An ensemble learning method that combines multiple decision trees to improve accuracy.

**5.2 Libraries and Tools Used**

**5.2.1 Python Libraries:**

**NumPy:** Provides efficient numerical operations on arrays and matrices, essential for scientific computing.

**Pandas:** A powerful library for data manipulation and analysis, offering data structures like Data Frames and Series.

**Matplotlib:** A versatile plotting library for creating customizable visualizations.

**Seaborn:** Built on top of Matplotlib, it offers high-level statistical data visualization.

**Jupyter Notebook:** A Powerful Tool for Data Science and Beyond Jupyter Notebook is a web-based interactive environment for creating and sharing documents that contain live code, equations, visualizations, and narrative text. It's become a popular tool for data scientists, researchers, and educators due to its flexibility and ease of use.

evaluating various models, including classifiers for sales prediction.

**Tensorflow:** TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML-powered applications.

**Matplotlib**: A plotting library for Python that is used to visualize training results to help on better understanding.

**Seaborn**: Built on top of Matplotlib, Seaborn is used for statistical data visualization, allowing for more aesthetically pleasing and informative visual representations of data, particularly useful for exploring datasets and analyzing results.

**CHAPTER 6**

##### **SOFTWARE REQUIREMENTS**

**6.1 Software Requirements :**

Operating system : Windows 10.

Programming Language : Python

Integrated Development Environments (IDEs) : Google colab

**6.2 Programming Language:**

**Python:**

Python is a high-level, general-purpose programming language known for its simplicity and readability. Python is a versatile and user-friendly programming language known for its readability and simplicity.

It's widely used in various fields, including:

Web Development: Building dynamic websites and web applications with frameworks like Django and Flask.

Data Science and Machine Learning: Analysing and interpreting large datasets, creating predictive models, and developing AI applications.

Python's Key Characteristics:

Interpreted Language: Python code is executed line by line, making it easier to debug and test.

High-Level Language: Python abstracts away many low-level details, allowing developers to focus on problem-solving.

Object-Oriented Programming: Python supports object-oriented principles, promoting code reusability and modularity.

Dynamic Typing: Variable types are determined at runtime, adding flexibility to coding.

Extensive Standard Library: Python comes with a rich collection of built-in modules and functions for various tasks.

Large and Active Community: A strong community provides support, resources, and third-party libraries.

Python's readability and efficiency make it a popular choice for both beginners and experienced programmers. Its wide range of applications and supportive community contribute to its continued growth and relevance in the programming world.

**6.3 Why Python for Data Science?**

**Readability:** Python's syntax is clean and easy to understand, making it accessible to both beginners and experienced programmers.

**Versatility:** It's used for a wide range of data science tasks, from data cleaning and analysis to machine learning and data visualization.

**Powerful Libraries:** Python boasts a robust collection of libraries specifically designed for data science

**Libraries Used:**

* Pandas
* NumPy
* Scikit-Learn
* Matplotlib
* seaborn

**CHAPTER 7**

**HARDWARE REQUIREMENTS**

The hardware requirements for a Sales prediction system can vary widely depending on the scale and complexity of the project. Here are some general guidelines:

**For Small-Scale:**

**Processor:** A modern CPU with multiple cores (e.g., Intel Core i5 or i7, AMD Ryzen 5 or 7)

**RAM:** 8GB or more

**Storage:** Solid-state drive (SSD) for faster data access

**For Medium-Scale:**

**Processor:** High-performance CPU (e.g., Intel Xeon, AMD Ryzen Threadripper)

**RAM:** 16GB or more

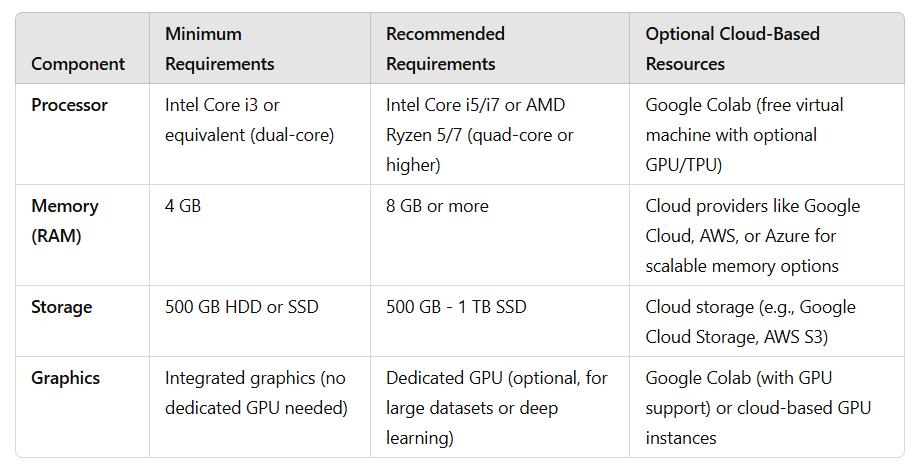
**Storage:** High-capacity SSD or HDD

**For Large-Scale:**

**Processor:** Multiple high-performance CPUs or a powerful CPU with many cores

**RAM:** 32GB or more

**Storage:** High-capacity SSDs or a combination of SSDs and HDDs



**Table 2:HARDWARE REQUIREMENTS**

**CHAPTER 8**

**FRONTEND AND BACKEND REQUIREMENTS**

**Frontend Requirements :**

**1.** User Interface (UI) Frameworks:

* React

1. Data Visualization Libraries:

* Plotly
* Matplotlib

1. HTML/CSS:

* HTML
* CSS

**BACKEND REQUIREMENTS:**

1. Programming Languages:

* Python

1. Web Frameworks:

* Flask (Python)
* Django (Python)

1. Database Management Systems:

* SQL Databases

1. Data Processing Libraries:

* Pandas (Python)
* NumPy (Python)

1. Machine Learning Libraries:

* Scikit-learn (Python)

**CHAPTER 9**

**MODULES**

##### 

##### **9.1 MODULES:**

##### Data Collection

##### Data Preprocessing

##### Feature Selection

##### Model Buliding

##### Model Evaluation

##### Prediction

##### Data Visualization

##### **9.1.1 Data Collection:**

##### This module involves gathering data from various sources that are relevant to the problem at hand. Data can be collected from structured sources like databases or unstructured sources like text files and web scraping.

##### **Key Activities:**

##### Identifying relevant data sources (e.g., sales records, customer demographics, product information).

##### Ensuring data quality and integrity during the collection process.

##### **9.1.2 Data Preprocessing:**

##### This module focuses on cleaning and transforming the raw data into a format suitable for analysis. It addresses issues such as missing values, outliers, and inconsistencies in the dataset.

##### **Key Activities:**

##### Handling missing data.

##### Normalizing or standardizing numerical features to bring them to a common scale.

##### Encoding categorical variables using techniques like one-hot encoding or label encoding.

##### Removing duplicates and irrelevant features.

##### **9.1.3 Feature Selection:**

##### This module involves selecting the most relevant features from the dataset that contribute significantly to the predictive model. The goal is to improve model performance and reduce overfitting**.**

##### **Key Activities:**

##### Using statistical tests to evaluate feature importance.

##### Analyzing correlations between features to eliminate redundant variables.

##### **9.1.4 Model Building:**

##### In this module, various machine learning algorithms are applied to the training data to create predictive models. The choice of algorithm depends on the nature of the problem (e.g., regression, classification).

##### **Key Activities:**

##### Splitting the dataset into training and validation sets to evaluate model performance.

##### Selecting appropriate algorithms (e.g., linear regression, decision trees, random forests) based on the problem type..

##### **9.1.5 Model Evaluation:**

##### This module assesses the performance of the trained models using various evaluation metrics. The goal is to ensure that the model generalizes well to unseen data.

##### **Key Activities:**

##### Using metrics such as Mean Absolute Error (MAE), R-squared(R2), or accuracy (for classification) to evaluate model performance.

##### **9.1.6 Prediction:**

##### This module involves using the trained model to make predictions on new, unseen data. The predictions should be actionable and relevant to the business context.

##### **Key Activities:.**

##### Feeding the processed data into the model to generate predictions.

##### Interpreting the predictions and providing insights that can inform business decisions.

##### **9.1.7 Data Visualization:**

##### This module focuses on presenting the results of the analysis and model predictions through visual means. Effective visualization helps stakeholders understand complex data and insights easily.

##### **Key Activities:**

##### Creating visualizations such as bar charts, line graphs, scatter plots, and heatmaps to represent data trends and model performance.

##### Using visualization tools (e.g., Matplotlib, Seaborn, Tableau) to enhance the presentation of findings and facilitate decision-making.

##### 

##### **CHAPTER 10**

##### **Challenges and solutions**

##### **Data Quality Issues:** The dataset may contain missing values, duplicates that can skew the results of the predictive models.

##### **Solution:**

##### **Imputation:** Fill missing values using mean, median, or mode.

##### **Duplication:** Remove duplicate entries to ensure data integrity

##### **Feature Selection:** Identifying the most relevant features among many can be challenging.

##### **Solution:** Perform correlation analysis and use feature importance from

##### models like Random Forest to select relevant features.

##### **Overfitting:** The model may perform well on training data but poorly on unseen data.

##### **Solution:** Apply regularization techniques(eg:lasso,rigid)and use cross-

##### ensure the model generalizes well.

##### 

##### **Underfitting**: This occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor performance.

##### **Solution:** The solution is to increase model complexity by using more

##### Sophisticat Algorithms or adding relevant features..

##### **Non-Linearity:** The model may perform well on training data but poorly on unseen data.

##### **Solution:** Utilize non-linear models (e.g., Decision Trees, Random Forests) or

##### feature transformations.

##### **Model Evaluation:** The model may perform well on training data but poorly on unseen data.

##### **Solution:** Use multiple evaluation metrics (e.g., R², RMSE, MAE) and visualizations

##### assess model performance.

##### **Scalability:** The model may perform well on training data but poorly on unseen data.

##### **Solution:** Choose efficient algorithms and consider dimensionality

##### techniques like PCA. These challenges and solutions are crucial for improving the accuracy

##### and reliability of the sales prediction model

##### .

##### 

##### **RESULT AND DISCUSSION**

##### 

##### **FIG 3: Correlation Matrix**

##### The image shows a correlation matrix. This matrix is used to visualize the correlation between different variables. In this case, the variables are related to a dataset about sales of items at different outlets.

##### The matrix shows the correlation coefficient between each pair of variables. The correlation coefficient is a number between -1 and 1 that measures the strength and direction of the linear relationship between two variables.

##### Positive correlation: A positive correlation means that as one variable increases, the other variable also tends to increase. This is represented by a value closer to 1.

##### Negative correlation: A negative correlation means that as one variable increases, the other variable tends to decrease. This is represented by a value closer to -1.

**Less correlated features :**

\* Item\_ identifier

\* Item \_mrp

\* Outlet\_ identifier

\* Outlet \_size

\* Outlet \_year

**More correlated features :**

\* Item\_ visibility

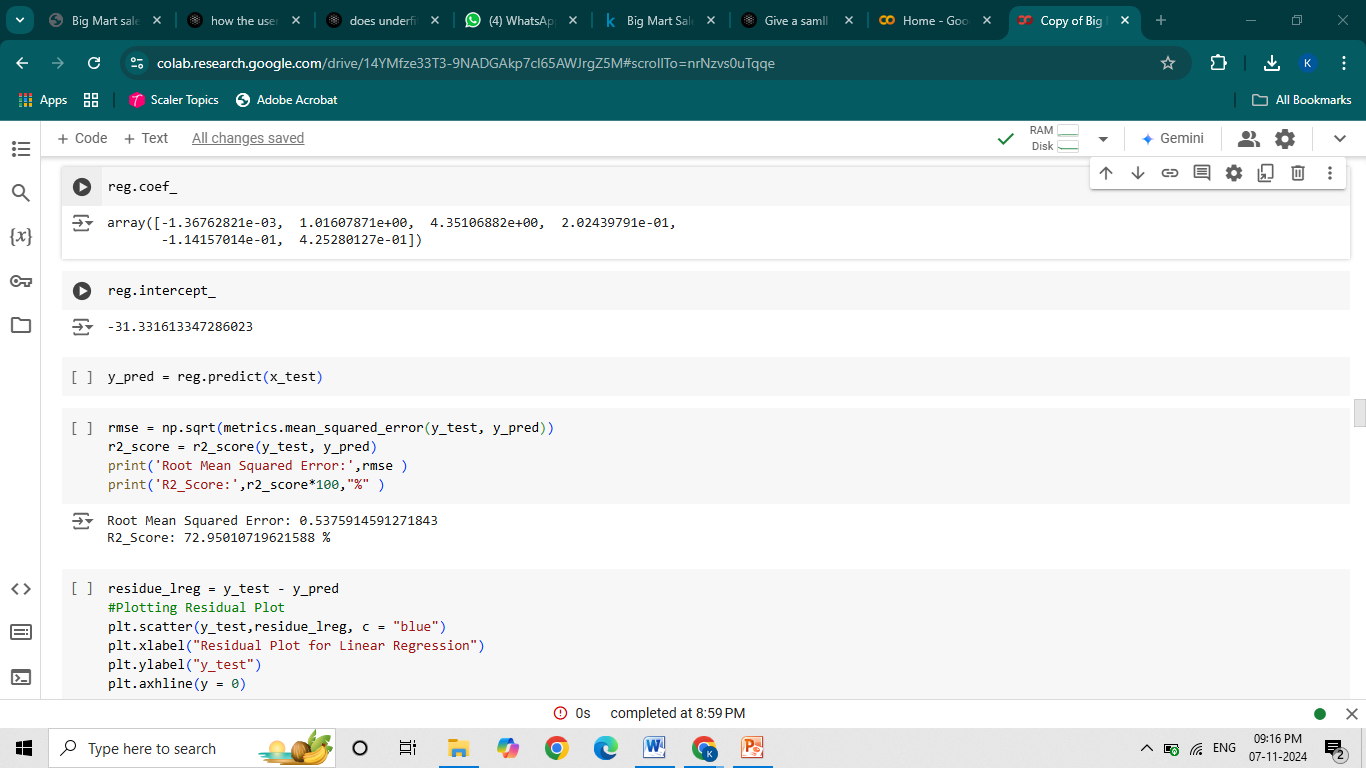
\* Item \_mrp

\* Outlet \_identifier

\* Outlet \_size

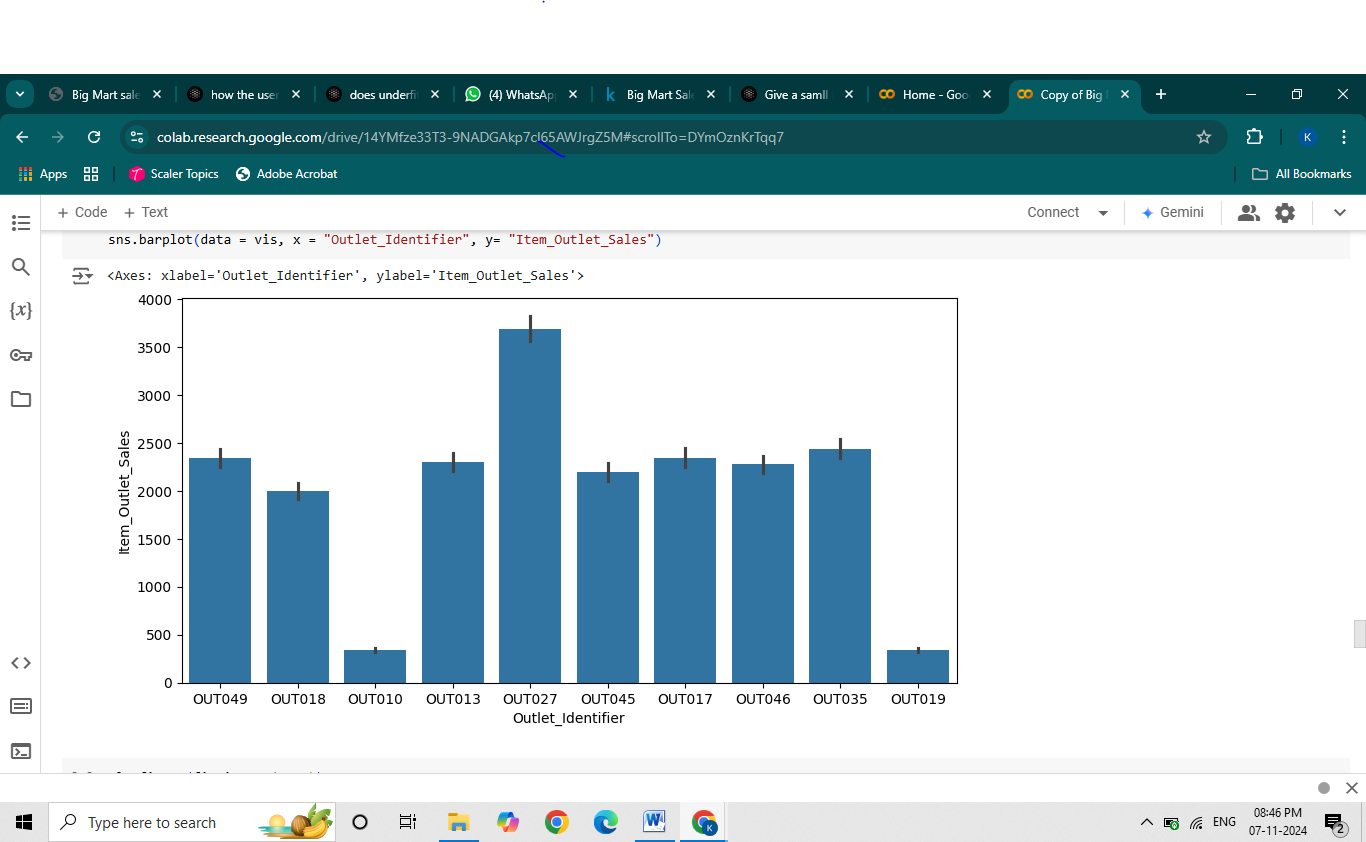
\* Outlet \_year

**Prediction Accuracy:**

****

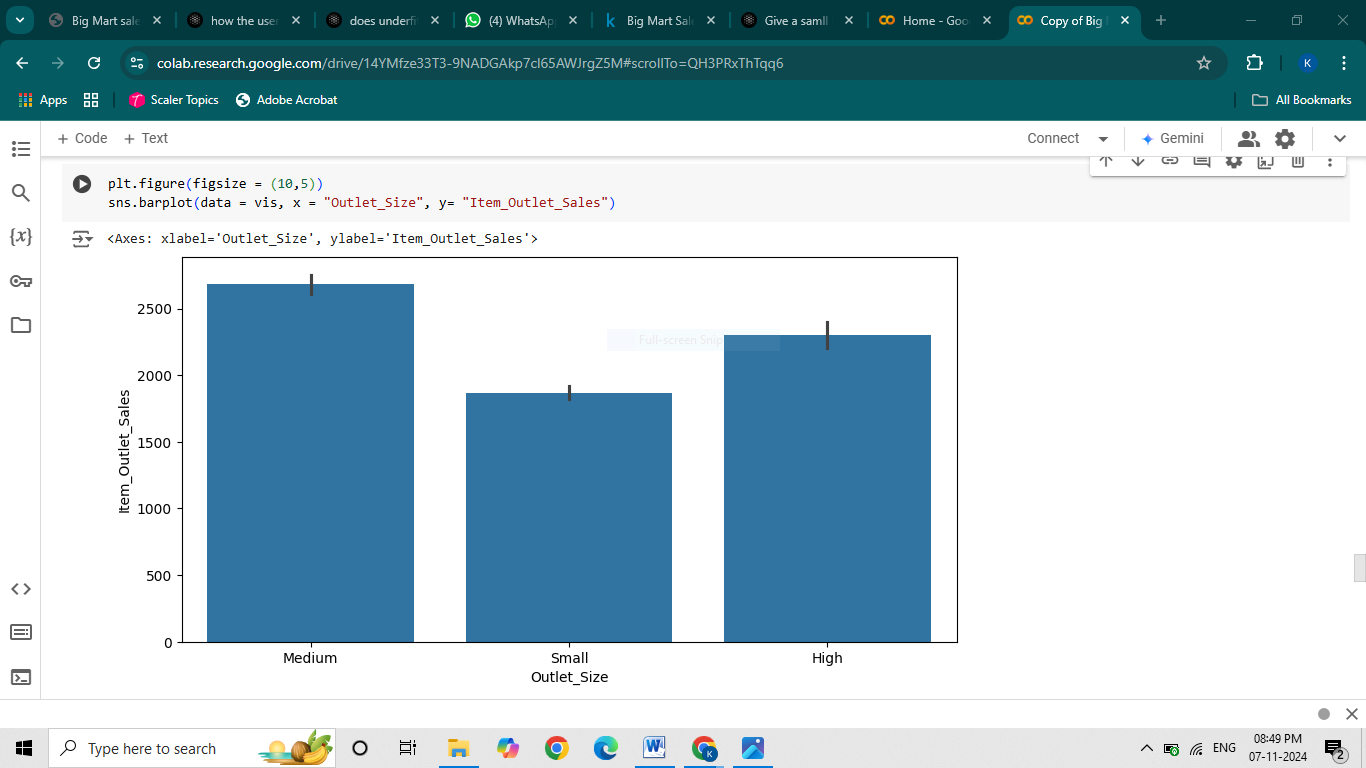
**FIG 4 : Prediction Accuracy**

The predicton accuracy I got by using Linear Regression algorithm is72.95 % .

****

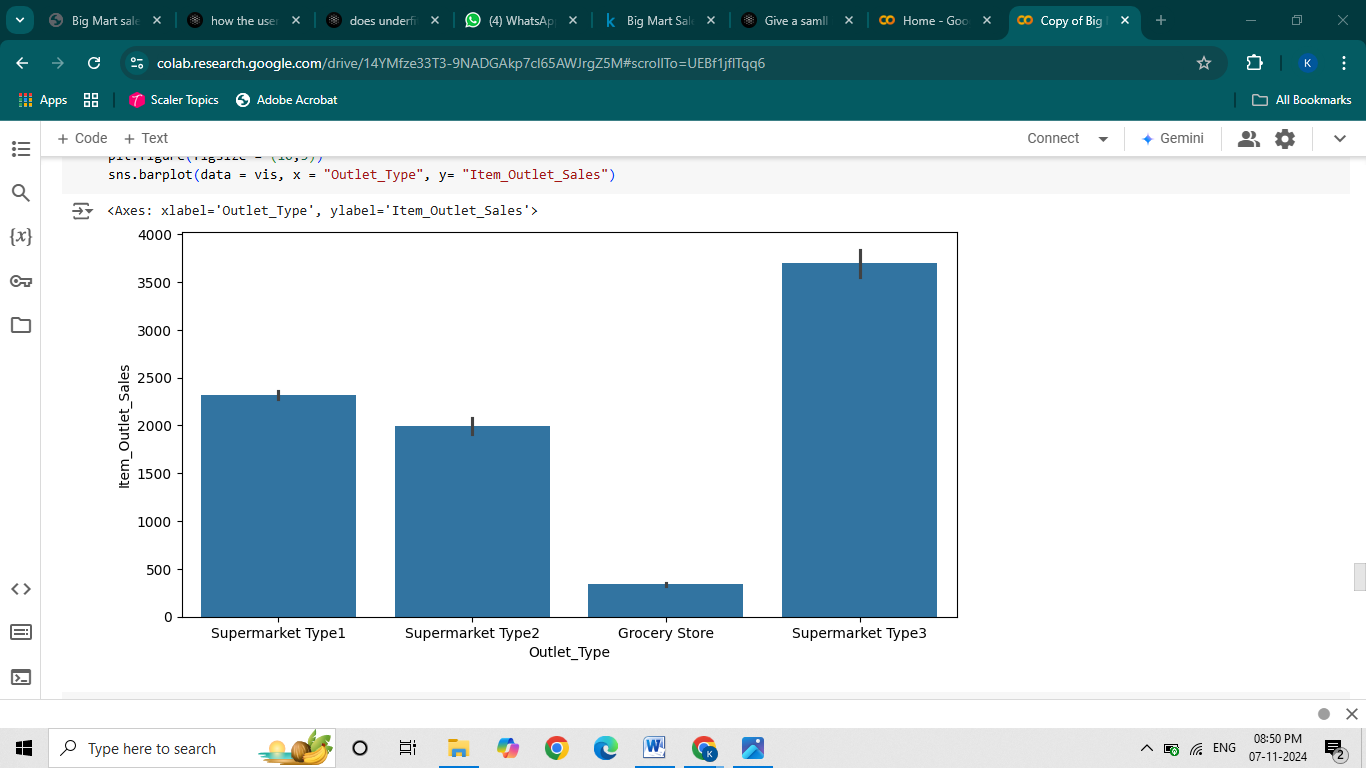
**FIG 5: Bar Graph of Item\_ outlet \_Sales vs Outlet\_ Identifier**

* The image is a bar chart that shows the average item outlet sales for different outlets.
* X-axis: The x-axis represents the Outlet\_ Year, which is a unique for each outlet.
* Y-axis: The y-axis represents the Item\_ Outlet \_Sales, which is the average sales generated by each every year.
* According to the above Bar Graph Outlet OUT027 have highest sales and OUT010 have lowest sales .

****

**FIG 6: Bar Graph of Item\_ outlet\_ Sales vs Outlet\_ Size**

* The image is a bar chart that shows the average sales of items in different outlet sizes.
* X-axis: The x-axis represents the Outlet\_ Size for each outlet.
* Y-axis: The y-axis represents the Item \_Outlet \_Sales, which is the average sales generated by outlet

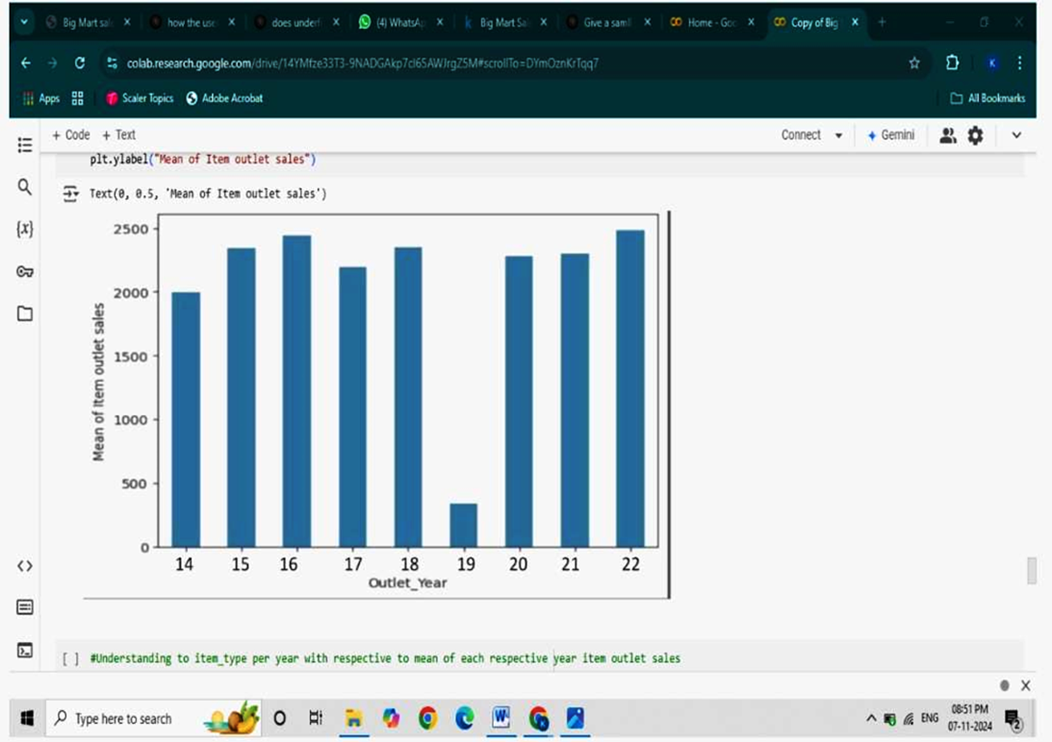
****

**FIG 7: Bar Graph of Item\_outlet\_Sales vs Outlet\_Type**

* The image is a bar chart that shows the average sales of items in different outlet types.
* X-axis: The x-axis represents the Outlet\_ Type for each outlet.
* Y-axis: The y-axis represents the Item\_ Outlet \_Sales, which is the average sales generated by Outlet.
* The chart shows that the average sales of items are highest in Supermarket Type 3 outlets, followed by Supermarket Type 1 outlets, followed by Supermarket Type 2 outlets, and then

Grocery Store outlets. This suggests that Supermarket Type 3 outlets are are earning more

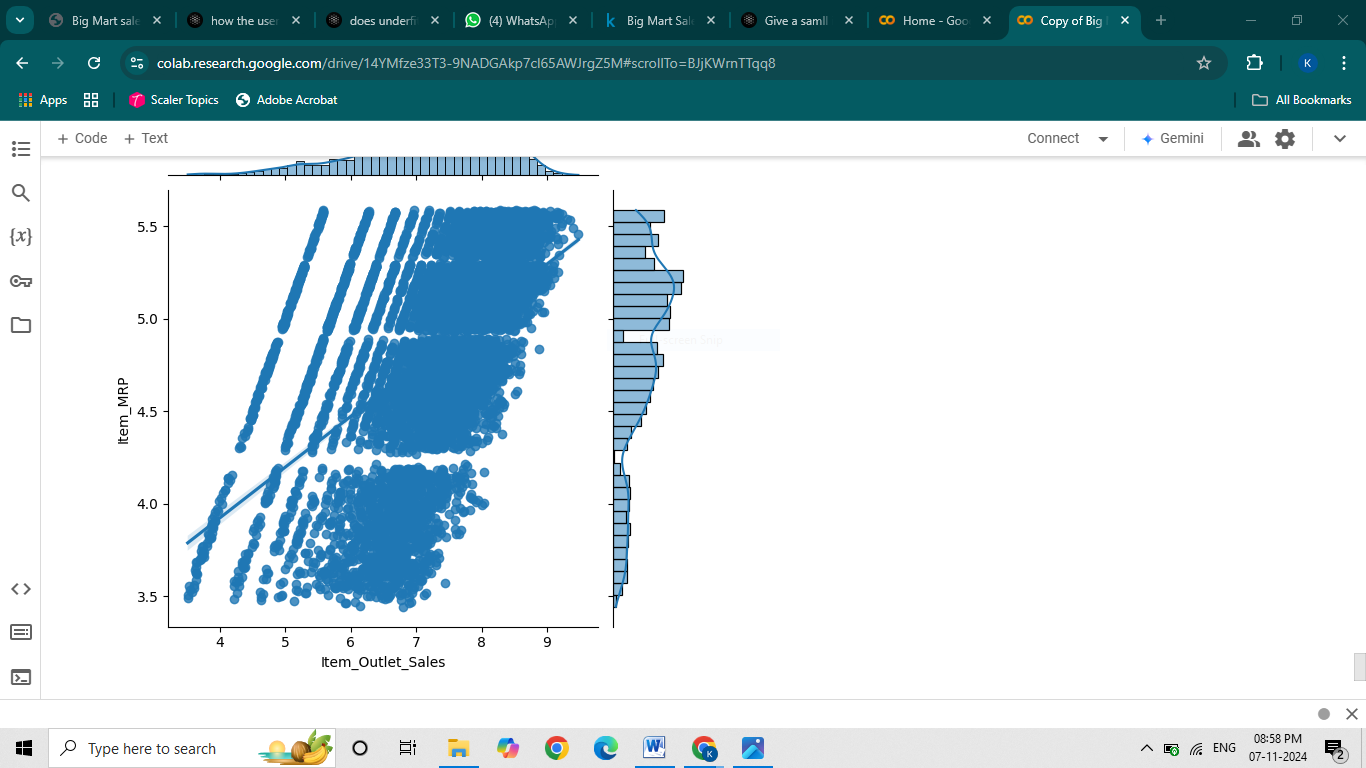
* The chart shows that the average sales of items are highest in medium-sized outlets, followed by high-sized outlets, and then small-sized outlets. This suggests that medium-sized outlets are the most profitable for the business.

****

Ygouhwxb14

**FIG 8 : Bar Graph of Item\_ outlet\_ Sales vs Outlet \_Year**

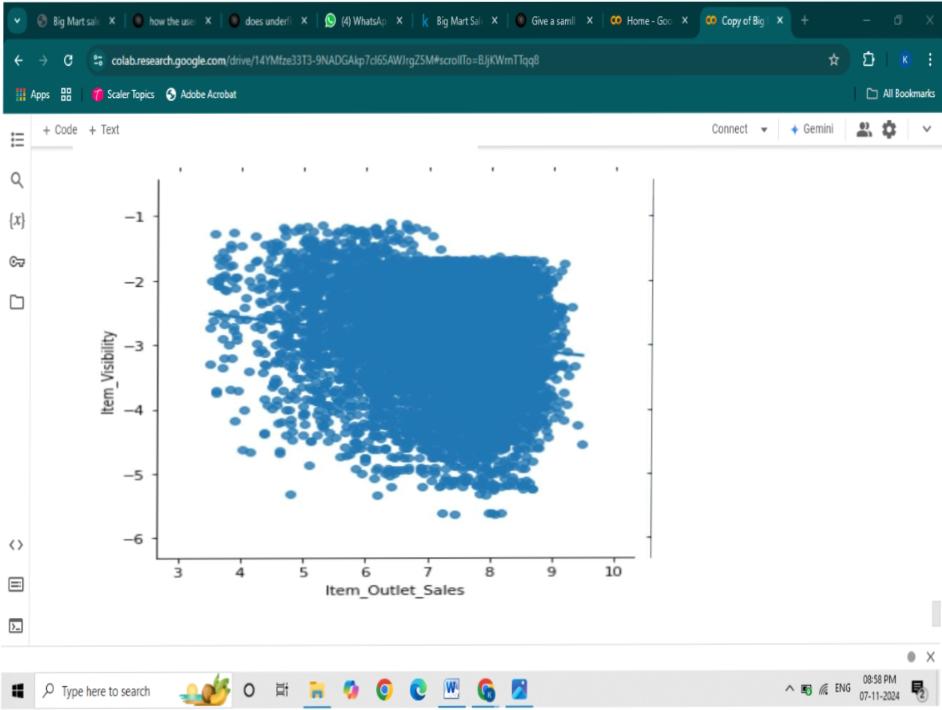
* The image shows a bar chart that represents the mean of item outlet sales for different outlet years.
* X-axis: The x-axis represents the Outlet\_ Year, which is a unique for each outlet.
* Y-axis: The y-axis represents the Item\_ Outlet\_ Sales, which is the average sales generated by each every year.
* The chart shows that the mean of item outlet sales is highest for outlet year 22 and lowest for outlet year 19. This indicates that outlets established in 2022 had the highest average sales, while those established in 2019 had the lowest average sales.

****

**FIG 9: Scatter plot of Item\_ outlet\_ Sales vs Item\_ MRP**

* The plot shows a scatter plot of the relationship between the 'Item\_ Outlet \_Sales' and the 'Item\_ MRP'.
* X-axis: "Item\_ Outlet \_Sales" represents the sales of items in different outlets.
* Y-axis: "Item\_ MRP" represents the MRP of each item within the store.
* There is a clear positive linear relationship between the two variables.

This means that as the sales of the items increase, the MRP (Maximum Retail Price) also increases.



**FIG 10 : Scatter plot of Item \_outlet\_ Sales vs Item\_ Visibility**

* The image shows a scatter plot of the relationship between "Item\_ Outlet\_ Sales" and "Item\_ Visibility."
* X-axis: "Item \_Outlet \_Sales" represents the sales of items in different outlets.
* Y-axis: "Item\_ Visibility" represents the visibility of each item within the store, likely a measure of shelf space or placement.
* This relationship suggests that certain items, despite having less prominent shelf space or lower visibility, might have a higher demand or appeal to customers.

**Internship Highlights**

During my data science internship, I had the opportunity to work on several impactful projects that enhanced my analytical and technical skills. I collaborated with a team to clean and preprocess large datasets, ensuring data quality for analysis. I applied machine learning algorithms to develop predictive models. Additionally, I gained hands-on experience with data visualization tools like Seaborn and Matplotlib, creating interactive dashboards that effectively communicated insights to stakeholders. This internship not only deepened my understanding of data science methodologies but also allowed me to contribute to real-world business solutions, reinforcing my passion for the field.

**Key Takeaways and Impact of the Internship**

1. **Practical Application of Theoretical Knowledge:**

Gained hands-on experience applying data science concepts and techniques learned

in academic settings.

1. **Enhanced Technical Skills:**

Improved proficiency in programming languages (e.g., Python, R) and data manipulation

libraries (e.g., Pandas, NumPy).

1. **Understanding of Data Pipeline:**

Learned the end-to-end data pipeline, from data collection and cleaning to model

deployment and evaluation.

1. **Collaboration and Teamwork:**

##### Developed skills in working collaboratively with cross-functional teams, enhancing

##### communication and project management abilities.

**Conclusion**

In this project, I focused on predicting sales for Big Mart using a linear regression algorithm, which yielded an accuracy of 72.95%. This result indicates that the model is capable of capturing a significant accuracy of sales data, although there is still potential for improvement. The BigMart sales prediction project has provided valuable insights into the factors influencing sales performance and inventory management. By employing the linear regression algorithm, I achieved an accuracy of 72.95%, which demonstrates a reasonable level of predictive capability for the dataset at hand.

**References**

* Ayesha Syed, Asha Jyothi Kalluri, Venkateswara Reddy Pocha, Venkata Arun Kumar Dasari, B.Ramasubbaiah (2020, FEBRUARY). “BIGMART SALES USING MACHINE LEARNING WITH DATA ANALYSIS”.
* “PREDICTION OF BIG MART SALES USING MACHINE LEARNING” in IRJMETS(Volume:03/Issue:09/September-2021)
* Predicitive Analysis for Big Mart Sales Using Machine Learning Algorithm” in IJRASET Volume 10 Issue VIII August 2022
* Nayana R, Chaithanya G, Meghana T, Narahari K S, Sushma M “Predictive Analysis for Big Mart Sales using Machine Learning Algorithms” in IJERTRTCSIT – 2022 (Volume 10 – Issue 12)
* <https://www.researchgate.net/publication/381826788_Big_Mart_Sales_Prediction_using_Machine_Learning>
* [**https://www.researchgate.net/publication/380146323\_Bigmart\_Sales\_Prediction\_Based\_on\_Voting\_Classifier\_Algorithm\_and\_Linear\_Regression?\_sg=HnLMfY\_PI3PSIhN-Nw8QvcqRPmUjnW0lFDtu6w5SHMlbolbL0cA65KXItJCWMpPov3bV0-ToMhyhjvg&\_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6Il9kaXJlY3QifX0https://www.researchgate.net/publication/380146323\_Bigmart\_Sales\_Prediction\_Based\_on\_Voting\_Classifier\_Algorithm\_and\_Linear\_Regression?\_sg=HnLMfY\_PI3PSIhN-Nw8QvcqRPmUjnW0lFDtu6w5SHMlbolbL0cA65KXItJCWMpPov3bV0-ToMhyhjvg&\_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6Il9kaXJlY3QifX0**](https://www.researchgate.net/publication/380146323_Bigmart_Sales_Prediction_Based_on_Voting_Classifier_Algorithm_and_Linear_Regression?_sg=HnLMfY_PI3PSIhN-Nw8QvcqRPmUjnW0lFDtu6w5SHMlbolbL0cA65KXItJCWMpPov3bV0-ToMhyhjvg&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6Il9kaXJlY3QifX0)