INTERNSHIP REPORT

A Report Submitted to

Jawaharlal Nehru Technological University Kakinada, Kakinada in partial fulfillment for the award of the degree of

## BACHELOR OF TECHNOLOGY

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

Submitted by

**B. PAVAN KALYAN**

**(20KN1A4407)**

**S. SRI VARSHA**

**(20KN1A4454)**

**P. SIVA NAGA RAJU**

**(20KN1A4434)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# NRI INSTITUTE OF TECHNOLOGY

**Autonomous**

##### (Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada) Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with ‘A’ Grade ISO 9001: 2015 Certified Institution

**Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist., PIN: 521212, A.P, India.**

**2020-2021**

|  |  |
| --- | --- |
| NRI | **NRI INSTITUTE OF TECHNOLOGY**  (An Autonomous Institution, Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada)  Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with ‘A’ Grade ISO 9001: 2015 Certified Institution  Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist., PIN: 521212, A.P, India. |

**CERTIFICATE**

This is to certify that the “**Internship report”** submitted by **B. PAVAN KALYAN** **(Regd. No.:20KN1A4407),** S. SUMA SRI VARSHA **(Regd. No.:20KN1A4454), P. SIVA NAGA RAJU (Regd. No.:20KN1A4434)** is work done by her and submitted during YEARS academic year, in partial fulfillment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING,** at **BLACK BUCK, LOCATION.**



**INTERNSHIP COORDIANTOR Head of the Department**

(R. KATHYAYANI) ( Dr. D. SUNEETHA)

**EXTERNAL EXAMINER**

## CERTIFICATE OF INTERNSHIP

## ACKNOWLEDGEMENT

I take this opportunity to thank all who have rendered their full support to my work. The pleasure, the achievement, the glory, the satisfaction, the reward, the appreciation and the construction of my project cannot be expressed with a few words for their valuable suggestions.

I am expressing my heartfelt thanks to **Head of the Department, Dr. D. SUNEETHA**

garu for her continuous guidance for completion of my Project work.

I am extending our sincere thanks to **Dean of the Department, Dr. K. V. SAMBASIVA RAO** for his continuous guidance and support to complete my project successfully.

I am thankful to the **Principal, Dr. C. NAGA BHASKAR** Garu for his encouragement to complete the Project work.

I am extending my sincere and honest thanks to the **Chairman, Dr. R. VENKATA RAO Garu & Secretary, Sri K. Sridhar** Garu for their continuous support in completing the Project work.

Finally, I thank the Administrative Officer, Staff Members, Faculty of Department of CSE, NRI Institute of Technology and my friends, directly or indirectly helped us in the completion of this project.

**B. PAVAN KALYAN**

**20KN1A4407**

**S. SUMA SRI VARSHA**

**20KN1A4454**

**P. SIVA NAGA RAJU**

**20KN1A4434**

# ABSTRACT

Supermarkets are expanding in most populated cities, and market competition is fierce. Future sales forecasting is an essential component of any business. Accurate forecasting of future sales assists businesses in developing and improving business strategies, as well as gaining proper market knowledge. Companies can use standard sales projections to analyze historical scenarios and then apply client purchases. Inferences are used prior to budgeting to detect shortfalls and weaknesses, as well as to develop a good strategy for the following year. A thorough understanding of previous opportunities enables one to plan for future market demands and increase one's chances of success. The dataset is one of the historical sales of a supermarket company that was recorded in three different branches over a three-month period. Predictive data analytics methods are easy to apply with this dataset.

In this project we used various Machine Learning Regression techniques to forecast the sales of the supermarkets referred above. The Machine Learning Techniques we used in this project are Linear Regression, K-nearest Neighbors, Support Vector Regression, Decision Tree Regression, Long Short-term Memory Regression etc. Even though we used various regression techniques, the main goal of this project is to find the ML model that predicts super market sales with the highest accuracy.

We followed 4 steps:

1. Data Collection
2. Data Pre-processing
3. Exploratory Data Analysis
4. Fitting the Model

## Organization Information:

Blackbuck Engineers is started in 2013 with the aim of creating a great ecosystem of academia, research, industry, and individuals. Blackbucks is a premier partner to Govts International Institute of Digital Technologies, and IITs. Blackbuck delivers the TAPTAP AI Driven employability platform to transform the journey of students towards their dream goals while helping 𝗛𝗥𝘀 𝗵𝗶𝗿𝗲 𝗿𝗶𝗴𝗵𝘁 𝘀𝘁𝘂𝗱𝗲𝗻𝘁𝘀.

Blackbuck has the largest chain of excellence in emerging tech across India.

Blackbucks runs post-graduation programs in AI, ML and Data Science

www.theblackbucks.com

## Programs and opportunities:

#### This ground-up approach helps us deliver not only the solution to our clients but also add value to at the core Blackbuck Engineers which operates in Five specific domains namely TapTap - AI Driven, Post-Graduation Programs, Center of Excellence, Virtual Programming Labs and Happie Days - A social Networking site for the students. TapTap offer services in Campus Recruitment drives for the Employers as well as College authorities. Recruiters can Conduct Customized Online Assessments secured with Best-in-class Proctoring and Schedule the end-to-end hiring process. Under each division we further provide specific industry solutions on focused domains with cutting edge technologies. Blackbuck Engineers emphasize on building relationships with our clients by delivering projects on time and within the budget

# Learning Objectives/Internship Objectives

* Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.
* An objective for this position should emphasize the skills you already possess in the area and your interest in learning more
* Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.
* Some internship is used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.
* Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

# INDEX

|  |  |  |
| --- | --- | --- |
| **S.NO** | **DESCRIPTION** | **PAGE NO** |
| 1 | Introduction | 6 |
| 2 | Motivation | 9 |
| 3 | System Analysis | 11 |
| 4 | Software Requirements Specifications | 13 |
| 5 | Literature Review | 15 |
| 6 | Keywords and Definition | 17 |
| 7 | Methodologies | 20 |
| 8 | Coding | 30 |
| 9 | Results | 32 |
| 10 | Challenges Faced | 36 |
| 11 | Conclusion | 38 |
| 12 | Future Scope | 41 |
| 13 | References & Appendix | 43 |

# WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES

|  |  |  |  |
| --- | --- | --- | --- |
| 1st WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 13.06.2022 | Monday | Introduce the Topic & the Problem Statement |
| 14.06.2022 | Tuesday | Introduce the Topic & the Problem Statement |
| 15.06.2022 | Wednesday | Introduce the Topic & the Problem Statement |
| 16.06.2022 | Thursday | Introduce the Topic & the Problem Statement |
| 17.06.2022 | Friday | Introduce the Topic & the Problem Statement |

|  |  |  |  |
| --- | --- | --- | --- |
| 2nd WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 20.06.2022 | Monday | Abstract Building |
| 21.06.2022 | Tuesday | Abstract Building |
| 22.06.2022 | Wednesday | Abstract Building |
| 23.06.2022 | Thursday | Abstract Building |
| 24.06.2022 | Friday | Abstract Submission |

|  |  |  |  |
| --- | --- | --- | --- |
| 3rd WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 27.06.2022 | Monday | Abstract Submission |
| 28.06.2022 | Tuesday | Abstract Submission |
| 29.06.2022 | Wednesday | Explain your Approach to Solving Problem |
| 30.06.2022 | Thursday | Explain your Approach to Solving Problem |
| 01.07.2022 | Friday | Explain your Approach to Solving Problem |

|  |  |  |  |
| --- | --- | --- | --- |
| 4th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 04.07.2022 | Monday | Explain your Approach to Solving Problem |
| 05.07.2022 | Tuesday | Explain Structure of Project |
| 06.07.2022 | Wednesday | Explain Structure of Project |
| 07.07.2022 | Thursday | Explain Structure of Project |
| 08.07.2022 | Friday | Explain Structure of Project |

|  |  |  |  |
| --- | --- | --- | --- |
| 5th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 11.07.2022 | Monday | Data Preprocessing |
| 12.07.2022 | Tuesday | Data Preprocessing |
| 13.07.2022 | Wednesday | Data Preprocessing |
|  | 14.07.2022 | Thursday | Data Preprocessing |
|  | 15.07.2022 | Friday | Data Preprocessing |

|  |  |  |  |
| --- | --- | --- | --- |
| 6th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 18.07.2022 | Monday | Perform Analysis |
| 19.07.2022 | Tuesday | Perform Analysis |
| 20.07.2022 | Wednesday | Perform Analysis |
|  | 21.07.2022 | Thursday | Perform Analysis |
|  | 22.07.2022 | Friday | Perform Analysis |

|  |  |  |  |
| --- | --- | --- | --- |
| 7th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 25.07.2022 | Monday | PPT Preparation |
| 26.07.2022 | Tuesday | PPT Preparation |
| 27.07.2022 | Wednesday | PPT Preparation |
|  | 28.07.2022 | Thursday | PPT Preparation |
|  | 29.07.2022 | Friday | PPT Preparation |

|  |  |  |  |
| --- | --- | --- | --- |
| 8th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 01.08.2022 | Monday | PPT Submission |
| 02.08.2022 | Tuesday | PPT Submission |
| 03.08.2022 | Wednesday | Mid Review |
|  | 04.08.2022 | Thursday | Mid Review |
|  | 05.08.2022 | Friday | Mid Review |

|  |  |  |  |
| --- | --- | --- | --- |
| 9th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 08.08.2022 | Monday | Mid Review |
| 10.08.2022 | Tuesday | Mid Review |
| 11.08.2022 | Wednesday | Building & Applying Algorithm |
|  | 12.08.2022 | Thursday | Building & Applying Algorithm |

|  |  |  |  |
| --- | --- | --- | --- |
| 10th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 16.08.2022 | Tuesday | Building & Applying Algorithm |
| 17.08.2022 | Wednesday | Building & Applying Algorithm |
| 19.08.2022 | Friday | Building & Applying Algorithm |

|  |  |  |  |
| --- | --- | --- | --- |
| 11th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 22.08.2022 | Monday | Concluding Project |
| 23.08.2022 | Tuesday | Concluding Project |
| 24.08.2022 | Wednesday | Concluding Project |
|  | 25.08.2022 | Thursday | Concluding Project |
|  | 26.08.2022 | Friday | Concluding Project |

|  |  |  |  |
| --- | --- | --- | --- |
| 12th WEEK | DATE | DAY | NAME OF THE TOPIC/MODULE COMPLETED |
| 29.08.2022 | Monday | Final Review |
| 30.08.2022 | Tuesday | Final Review |
| 01.09.2022 | Wednesday | Final Review |
|  | 02.09.2022 | Thursday | Final Review |
|  | 05.09.2022 | Friday | Final Review |

**CHAPTER – 1**

**INTRODUCTION**

# 1.INTRODUCTION

* **CONTEXT:**

The growth of supermarkets in most populated cities are increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

* **ATTRIBUTE INFORMATON:**

Invoice id: computer generated sales slip invoice identification number

Branch: Branch of supercentre (3 branches are available identified by A, B, C).

Customer type: Type of customers, recorded by the members for customers using member card and normal for without member card.

* City: Location of supercentres.
* Gender: Gender type of customer.

Production Line: General item categorization in groups.

Unit price: Price of each product in $.

Quantity: Number of products purchased by customer.

Tax: 5% tax fee for customer buying.

Total: Total price including tax.

Date: Date of purchase (Record available from January 2019 to March 2019)

Time: Purchase time (10am to 9pm)

Payment: Payment used by customer for purchase

COGS: Cost of goods sold

Gross margin percentage: Gross margin percentage.

Rating: customer satisfaction rating on their overall shopping experience (1 -10)

**1.1 CLASSIFICATION OF TECHNIQUES:**

1. Linear regression

2. Lassocv Regression

3. Random Forest Regression

4. Decision Tree Regression

5. Long Short-Term Memory Regression

6. Ridge Regression

7. Support Vector Regression

8. K Nearest Neighbor Regression

**CHAPTER – 2**

**MOTIVATION**

# 2. MOTIVATION

The success of a supermarket can be gauged by its sales. The growth of retail store & supermarket sales is directly proportional to the customer experience as **sales are linked to consumer satisfaction**. The higher the consumer satisfaction, the better your sales performance will be.

**Inventory** is an important asset sitting on the shelves of the supermarket. It plays an instrumental role in the fate of the store.

**Strong desire**, the most important element in sales, will provide the incentive to make the changes even when it is difficult or uncomfortable." Commitment is having a plan to achieve that goal. "Commitment is a measure of whether you will do whatever it takes to achieve success in sales.

By using this project whoever wants to improve their supermarket’s business they use these insights. By using this information supermarket owners can improve their business.





**CHAPTER – 3**

**SYSTEM ANALYSIS**

# 3.SYSTEM ANALYSIS

**2.1 EXISTING SYSTEM:**

The existing system is very paper based in small as well as medium supermarkets. Even though the

paper work and manpower requirement is less, the existing system is not very economical for these

markets. Relevant and irrelevant information are entered and stored in the same place, which is very

clumsy and untidy process.

In case of big supermarkets, the existing system is computerized to some extent, but it is not fully

automated to cover all the aspects of the supermarket. The data entry, storing, and retrieval procedure is

very inefficient. Further, there are chances of data misplacement and wrong data entry. The system is

still very insecure and inflexible to adapt to user requirements.

**2.2 PROPOSED SYSTEM:**

The proposed supermarket management system aims at full automation of big, medium, and mini

supermarkets by making the system reliable, fast, user-friendly, and informative. It reduces paperwork,

manpower requirement, and increases the productivity of the supermarket. Using this application, one

can add, modify, update, save, delete, and print details. There’s also a search feature to find products

available in the supermarket.

**CHAPTER – 4**

**SOFTWARE REQUIREMENTS**

# 4.SOFTWARE REQUIREMENTS

## System configurations

The software requirement specification can produce at the culmination of the analysis task. The function and performance allocated to software as part of system engineering are refined by established a complete information description, a detailed functional description, a representation of system behavior, and indication of performance and design constrain, appropriate validate criteria, and other information pertinent to requirements.

## Software Requirements:

* Operating system : Windows 10 Ultimate.
* Coding Language : Python
* Platform : Visual Studio Code.
* Data Base : Kaggle Database.

## Hardware Requirement:

* System : Core i5 2.4 GHz.
* Hard Disk : 1TB.
* Ram : 8GB.

**CHAPTER – 5**

**LITERATURE REVIEW**

# 5.LITERATURE REVIEW

**Python** is a [high-level](https://en.wikipedia.org/wiki/High-level_programming_language), [interpreted](https://en.wikipedia.org/wiki/Interpreter_(computing)), [general-purpose programming language](https://en.wikipedia.org/wiki/General-purpose_programming_language). Its design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with the use of [significant](https://en.wikipedia.org/wiki/Off-side_rule) [indentation](https://en.wikipedia.org/wiki/Off-side_rule).

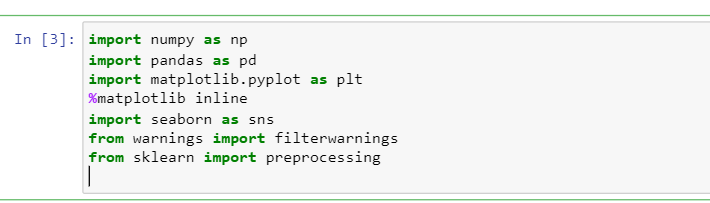
Python is [dynamically-typed](https://en.wikipedia.org/wiki/Type_system#DYNAMIC) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple

**Programming** including [structured](https://en.wikipedia.org/wiki/Structured_programming) (particularly [procedural](https://en.wikipedia.org/wiki/Procedural_programming)), [object-](https://en.wikipedia.org/wiki/Object-oriented_programming) [oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). It is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

**JUPYTER NOTBOOK**: The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

If so, then you can use a handy tool that comes with Python called pip to install Jupyter Notebook like this:

$ pip install jupyter



**CHAPTER – 6**

**KEYWORDS & DEFINITIONS**

# 6.KEYWORDS and DEFINITION

|  |  |
| --- | --- |
| PIP | pip is the standard package manager for Python. It allows you to install and manage additional packages that are not part of the Python standard  library. |
| from | The from keyword is used to import only a  specified section from a module. |
| Import | import imports the whole library. from import  imports a specific member or members of the library. |
| Import\* | It just means that you import all (methods,  variables, ...) in a way so you don't need to prefix them when using them |
| NumPy | NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. |
| Updater | The update() method inserts the specified items to the dictionary. The specified items can be a dictionary, or an iterable object with key value pairs. |
| Pandas | pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. |
| Matplotlib | Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, |
| Machine Learning | Machine Learning is a subfield of Artificial Intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behaviour. |
| Regression | Regression is a technique for investigating relationship between dependent variable or outcome and independent features or variables. |
| Dataset | A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. |
| Algorithm | A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer. |

**CHAPTER – 7**

**METHODOLOGIES**

# 7.METHODOLOGIES

We follow a structured methodology for our projects which starts from designing the solution to the implementation phase. Well planned Project reduces the time to deliver the project and any additional ad-hoc costs to our clients, hence we dedicate majority of our time understanding our client’s business and gather requirements. This ground up approach helps us deliver not only the solution to our clients but also add value to your investments.

**Data Preprocessing:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

It involves the following steps:

1. Data collection
2. Importing Libraries
3. Importing Dataset
4. Finding missing data
5. Encoding Categorical values
6. Exploratory Data Analysis
7. Splitting Dataset into train-test data
8. Fitting the model
9. Model Evaluation

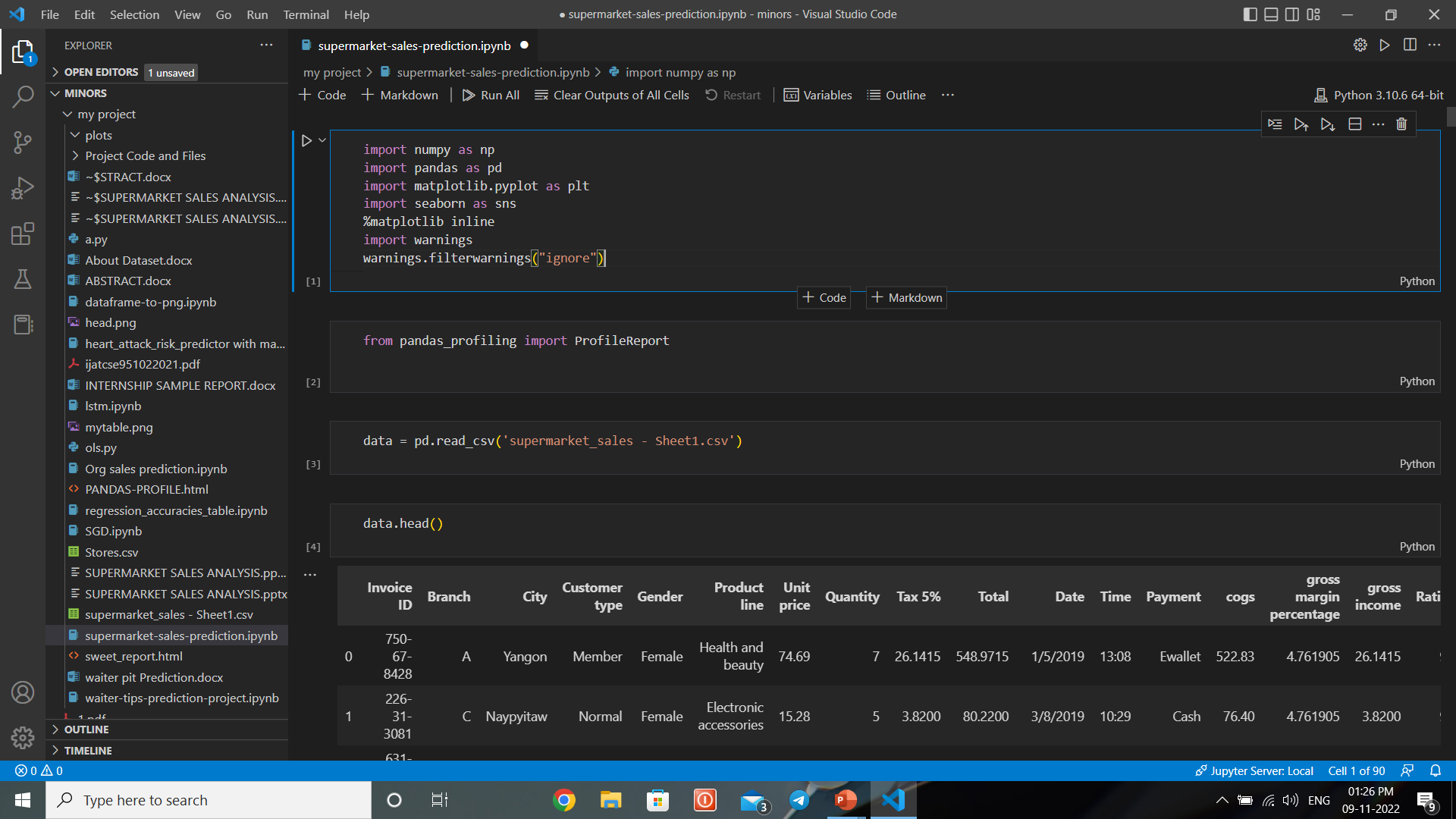
**Data Collection:**

Data collection is an essential part of exploratory data analysis. It refers to the process of finding and loading data into our system. Good, reliable data can be found on various public sites or bought from private organizations. Some reliable sites for data collection are Kaggle, GitHub, Machine Learning Repository, etc.

The data depicted below represents the supermarket sales dataset that is available on Kaggle. It contains information on supermarket and its sales.

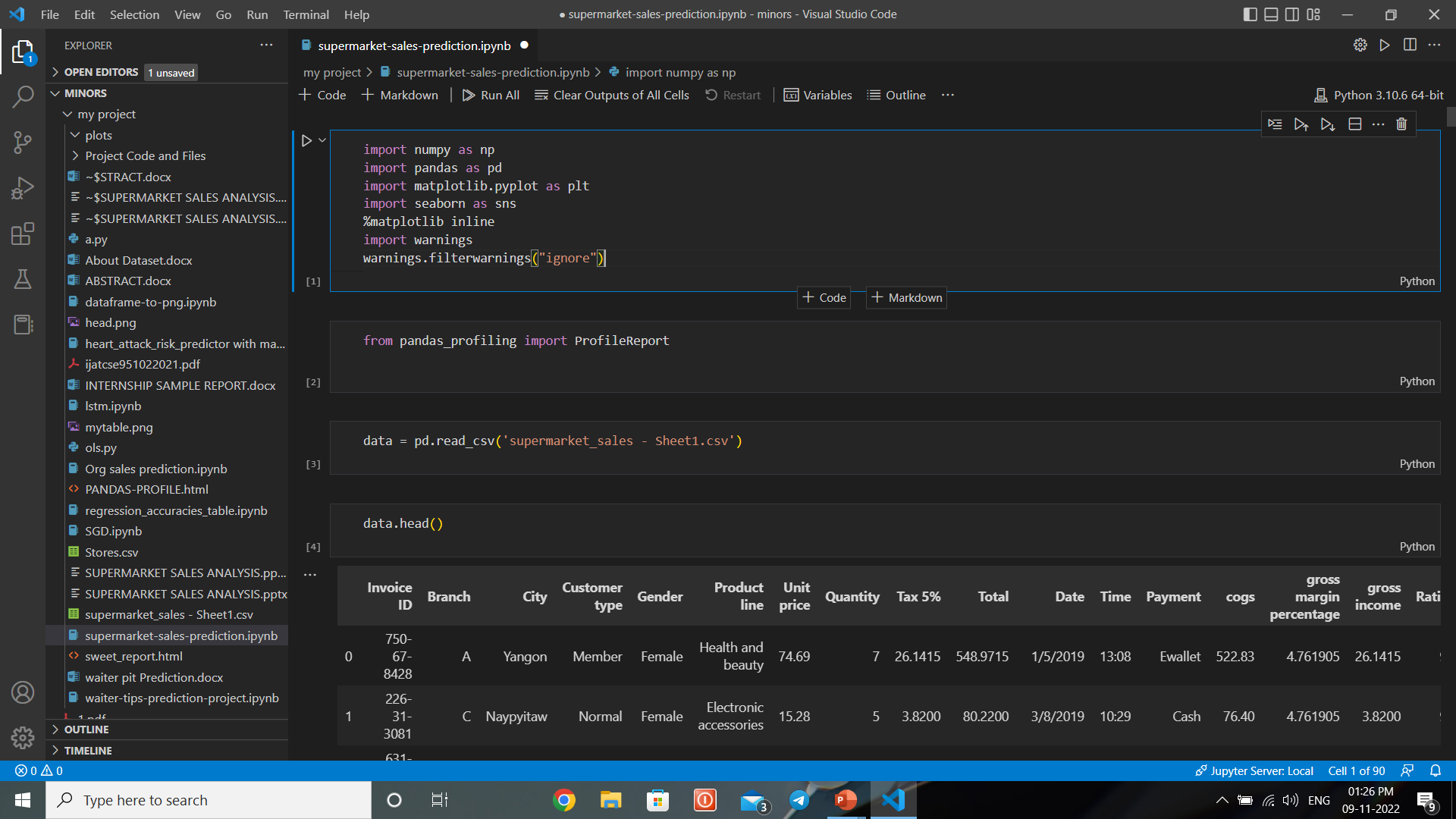
**Importing Libraries:**

To import the required libraries, we use the following code:

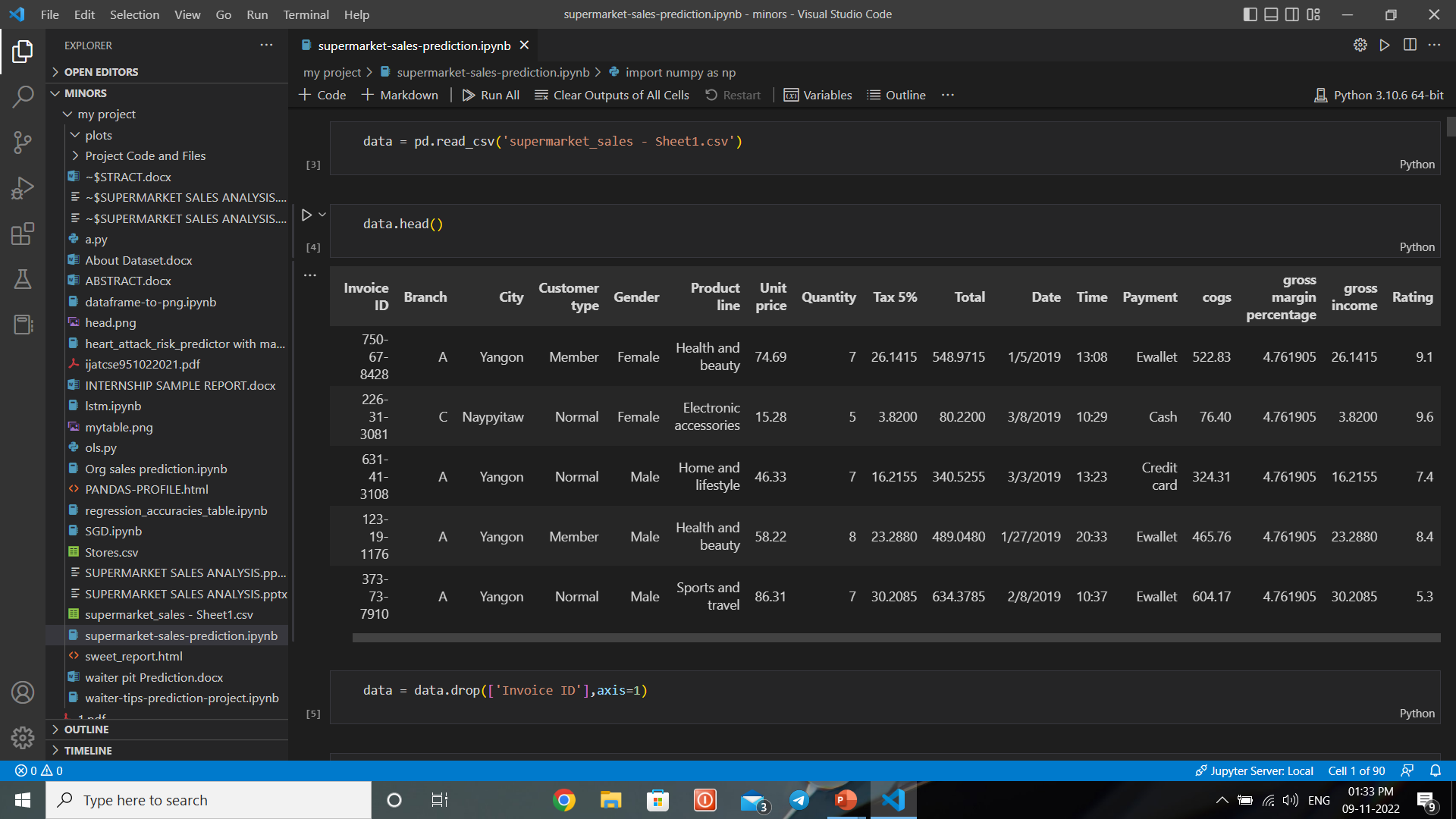


**Importing Dataset:**

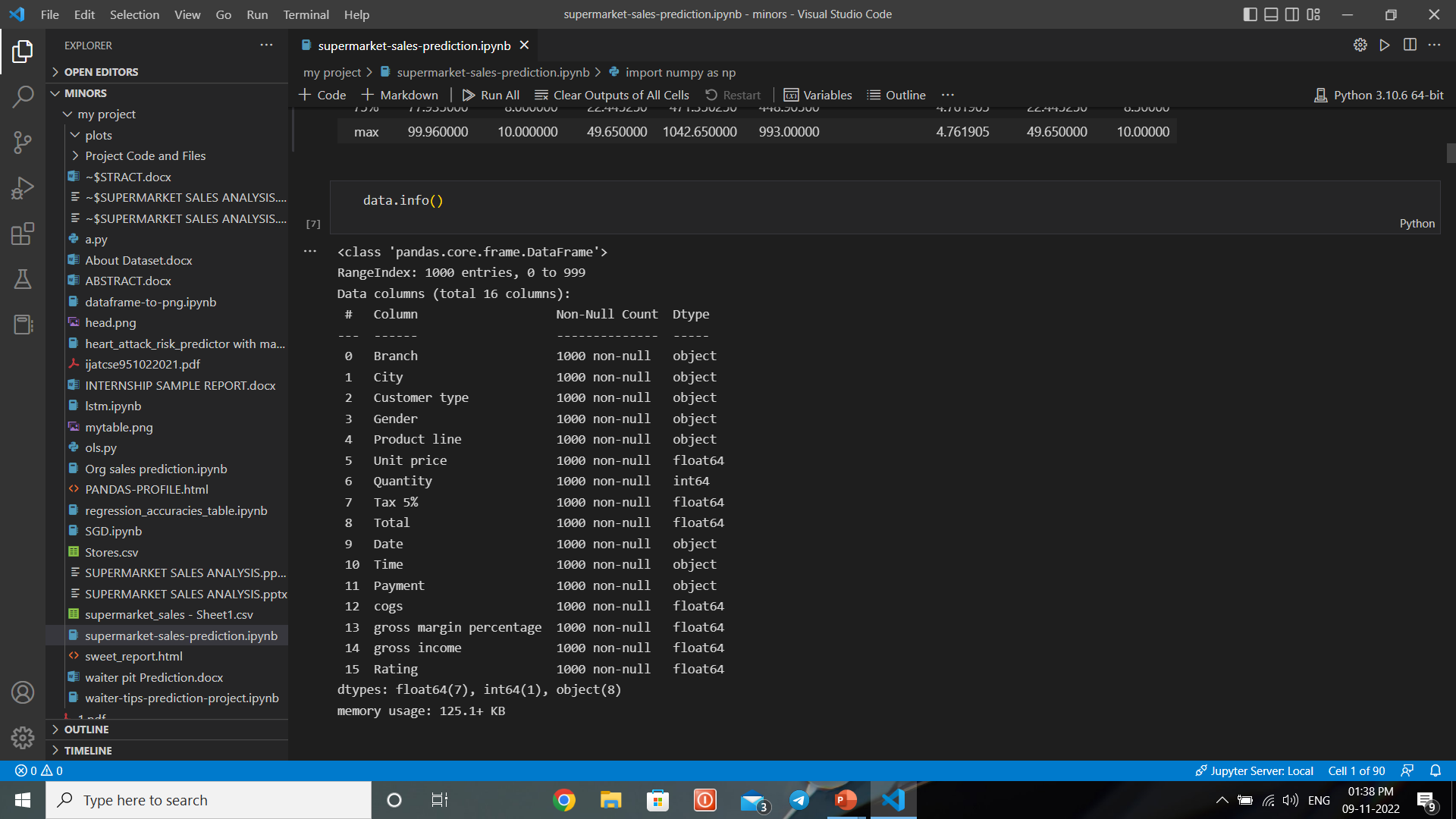
To import the dataset, use the following code:



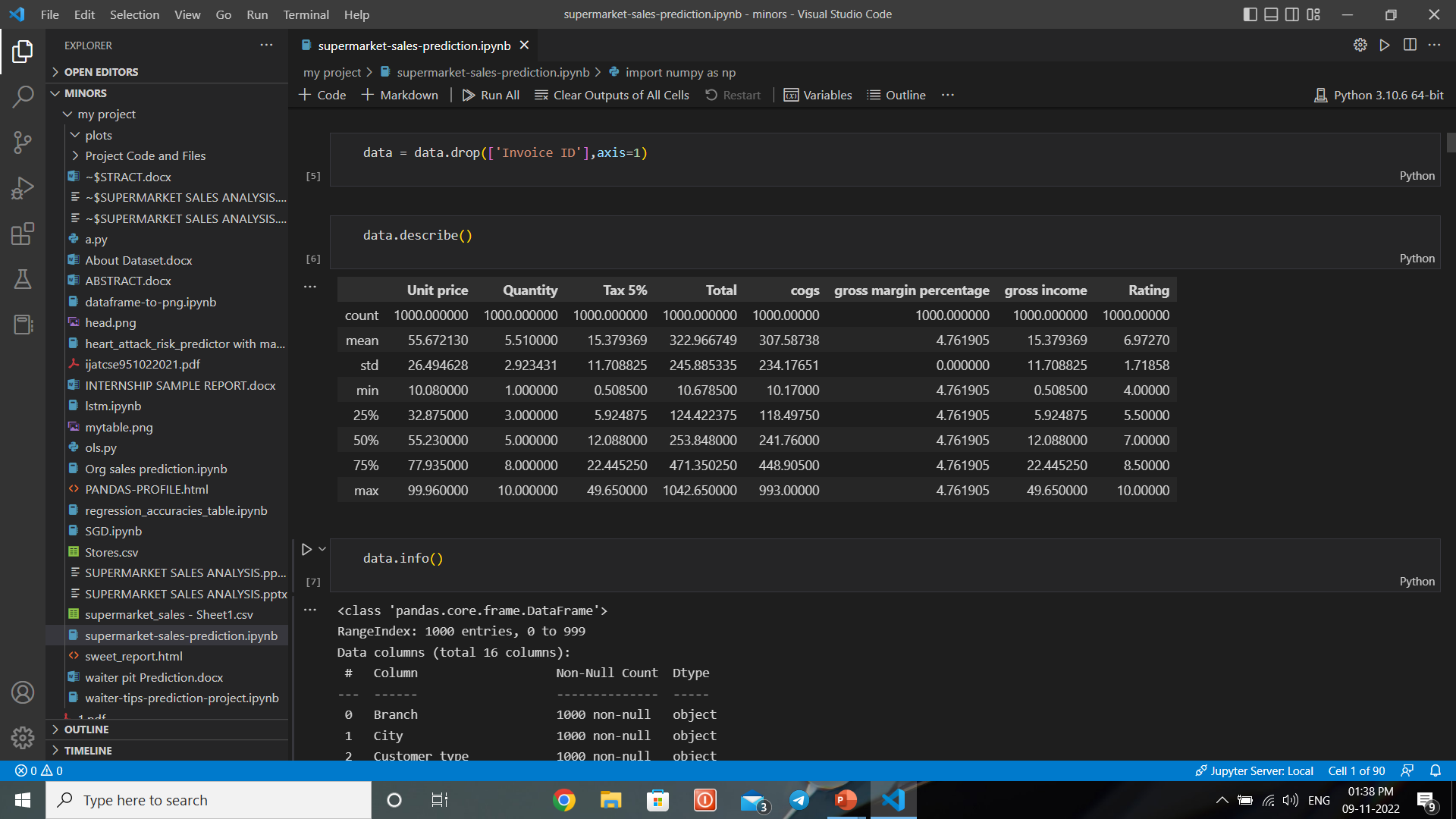
To get a brief look into our Dataset we used the following code:



Information about the dataset can be obtained by using the below code:



Python provides us description about Dataset as follows:



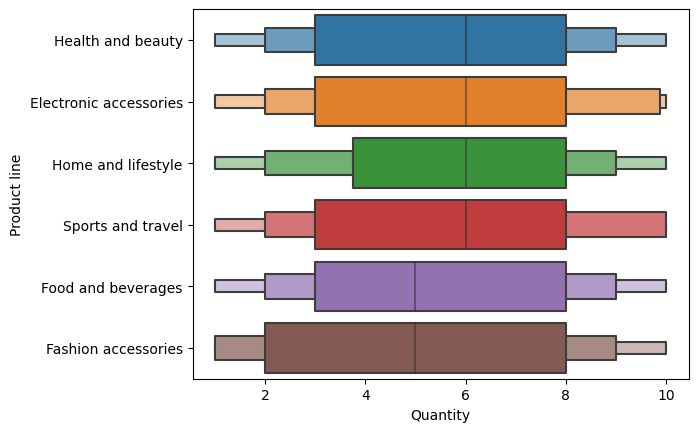
Since the supermarket sales dataset doesn’t have any null values but categorical data, we convert them into numerical values by using “Pandas” library.

**Exploratory Data Analysis:**

The visualization techniques that are used in this project are

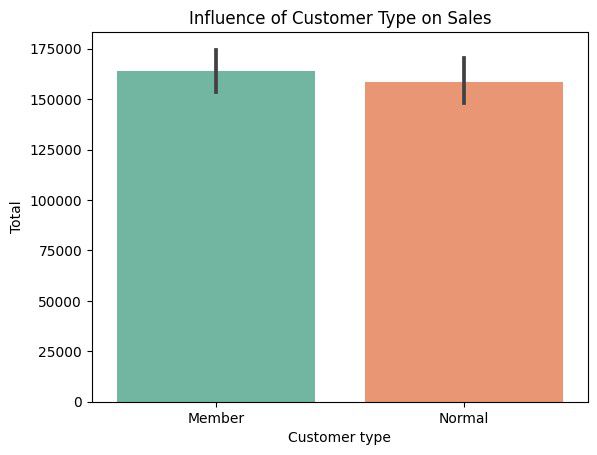
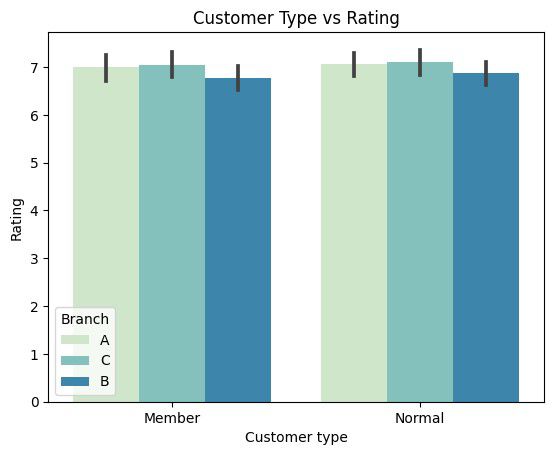
1. Box Plot
2. Bar Plot
3. Heat Map
4. Line Plot
5. **Box Plot:**

**Boxplot** is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. In addition to the box on a box plot, there can be lines (which are called *whiskers*) extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the **box-and-whisker plot** and the **box-and-whisker diagram**. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution (though Tukey's boxplot assumes symmetry for the whiskers and normality for their length.



1. **Bar Plot:**

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

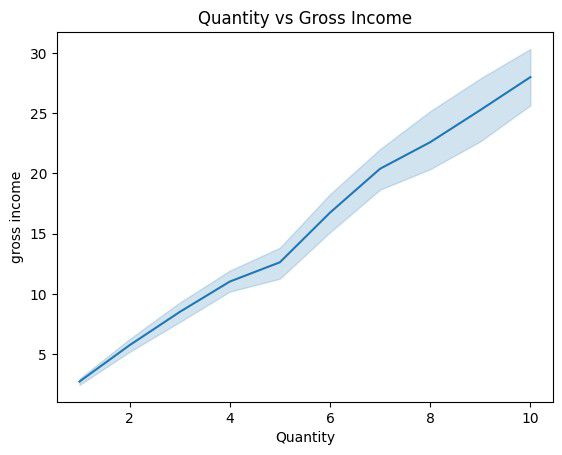
1. **Heat Map:**

A **heat map** (or **heatmap**) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by [hue](https://en.wikipedia.org/wiki/Hue) or [intensity](https://en.wikipedia.org/wiki/Brightness), giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. There are two fundamentally different categories of heat maps: the cluster heat map and the spatial heat map. In a cluster heat map, magnitudes are laid out into a matrix of fixed cell size whose rows and columns are discrete phenomena and categories, and the sorting of rows and columns is intentional and somewhat arbitrary, with the goal of suggesting clusters or portraying them as discovered via statistical analysis.



1. **Line Plot:**

A **line chart** or **line graph** or **curve chart** is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically. In these cases, they are known as run charts.



**Splitting Dataset into Train-Test data:**

The train-test split is used to estimate the performance of machine learning algorithms that are applicable

for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such

that we can compare our own machine learning model results to machine results. By default, the Test set

is split into 30 % of actual data and the training set is split into 70% of the actual data. Scikit-learn

alias “**sklearn”** is the most useful and robust library for machine learning in Python. The **scikit-learn**

**library** provides us with the “model\_selection” module in which we have the splitter function

“train\_test\_split().”

**Model Fitting:**

In this project we used most of the Regression Algorithms like

1. Linear Regression

2. Ridge Regression

3. LassoCV Regression

4. Decision Tree Regression

5. Random Forest Regression

6. OLS Regression

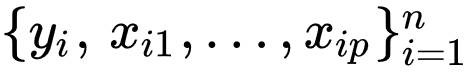
7. LSTM Regression

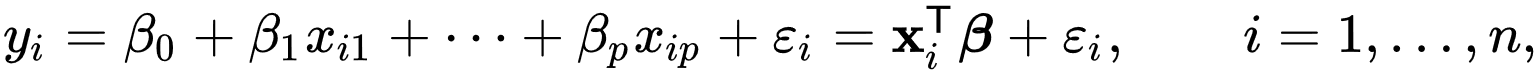
**1.Linear Regression:**

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called **multiple** linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Given a data set {\displaystyle \{y\_{i},\,x\_{i1},\ldots ,x\_{ip}\}\_{i=1}^{n}} of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable *y* and the *p*-vector of regressors **x** is linear. This relationship is modeled through a *disturbance term* or *error variable* *ε* — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus, the model takes the form

{\displaystyle y\_{i}=\beta \_{0}+\beta \_{1}x\_{i1}+\cdots +\beta \_{p}x\_{ip}+\varepsilon \_{i}=\mathbf {x} \_{i}^{\mathsf {T}}{\boldsymbol {\beta }}+\varepsilon \_{i},\qquad i=1,\ldots ,n,} 

**2. Ridge Regression:**

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. Also known as Tikhonov regularization, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems. it is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias (see bias–variance trade-off).

**3. LassoCV Regression:**

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso Regression uses L1 regularization technique. It is used when we have more features because it automatically performs feature selection. The word “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.

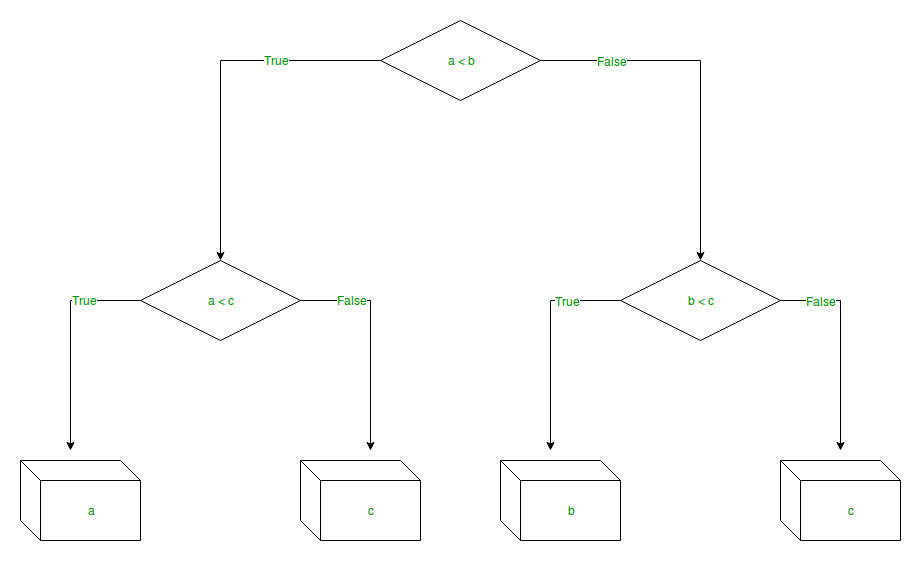
1. **Decision Tree Regression:**

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.

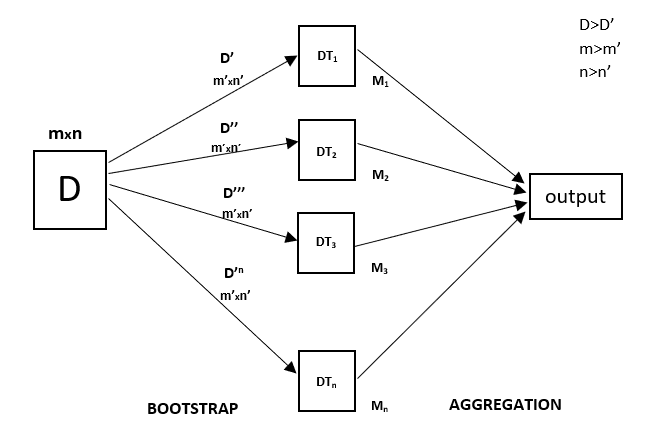
The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. 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 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.

**5. Random Forest:**

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn’t depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**. Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models.

**6.Ordinary Least Squares Regression:**

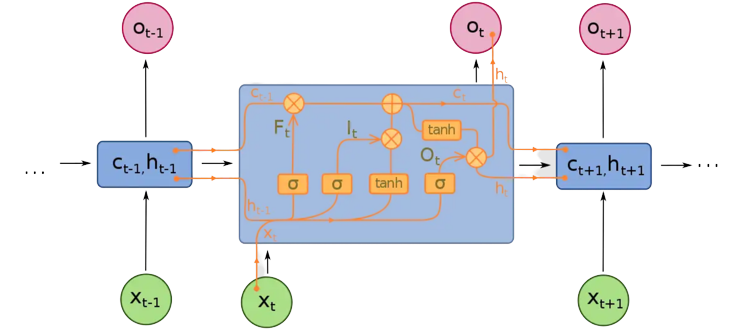
In statistics, ordinary least squares (OLS) is a type of linear least squares method for choosing the unknown parameters in a linear regression model (with fixed level-one effects of a linear function of a set of explanatory variables) by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the input dataset and the output of the (linear) function of the independent variable.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface—the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a simple linear regression, in which there is a single regressor on the right side of the regression equation.

The OLS estimator is consistent for the level-one fixed effects when the regressors are exogenous and forms perfect collinearity (rank condition), consistent for the variance estimate of the residuals when regressors have finite fourth moments and—by the Gauss–Markov theorem—optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed with zero mean, OLS is the maximum likelihood estimator that outperforms any non-linear unbiased estimator.

**7. LSTM Regression:{\displaystyle \mathbf {y} =X{\boldsymbol {\beta }}+{\boldsymbol {\varepsilon }},\,}{\displaystyle {\boldsymbol {\beta }}={\begin{pmatrix}\beta \_{0}\\\beta \_{1}\\\beta \_{2}\\\vdots \\\beta \_{p}\end{pmatrix}},\quad {\boldsymbol {\varepsilon }}={\begin{pmatrix}\varepsilon \_{1}\\\varepsilon \_{2}\\\vdots \\\varepsilon \_{n}\end{pmatrix}}.}**

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network that can learn order dependence. The output of the previous step is used as input in the current step in RNN. Hochreiter & Schmidhuber created the LSTM. It addressed the issue of RNN long-term dependency, in which the RNN is unable to predict words stored in long-term memory but can make more accurate predictions based on current data. RNN does not provide an efficient performance as the gap length rises. The LSTM may keep information for a long time by default. It is used for time-series data processing, prediction, and classification.



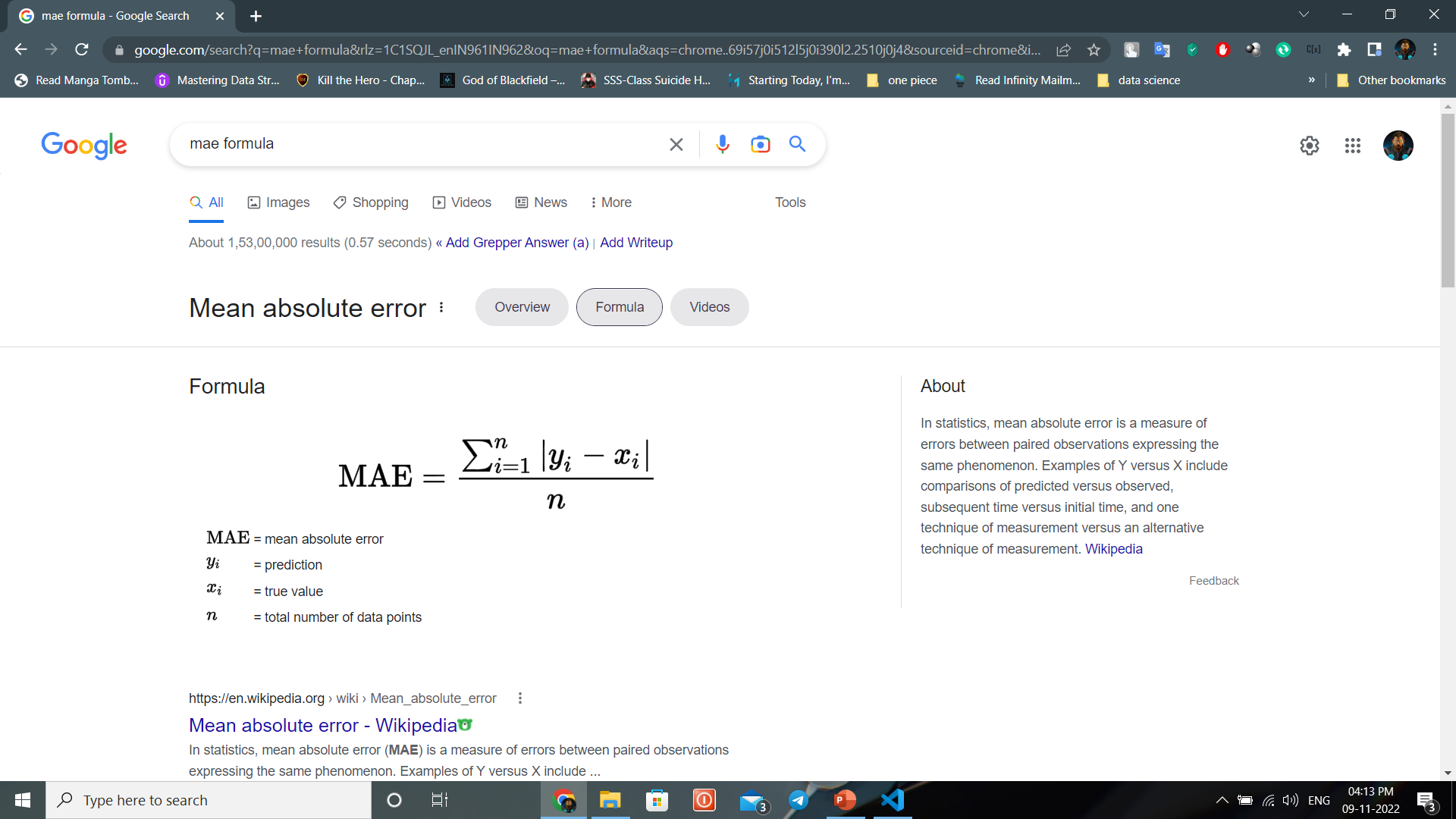
LSTM has feedback connections, unlike conventional feed-forward neural networks. It can handle not only single data points (like photos) but also complete data streams (such as speech or video). LSTM can be used for tasks like unsegmented, linked handwriting recognition, or speech recognition.

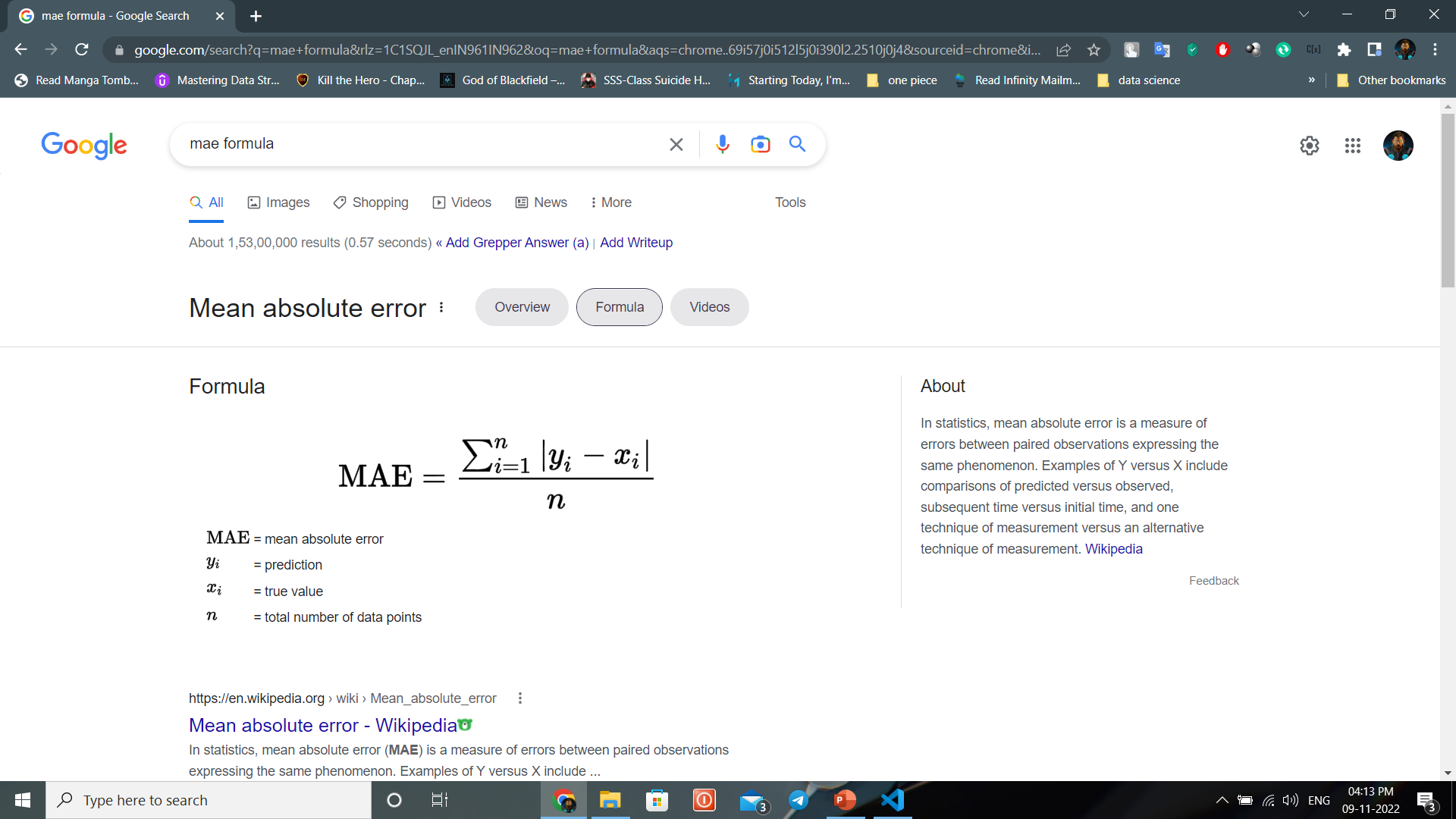
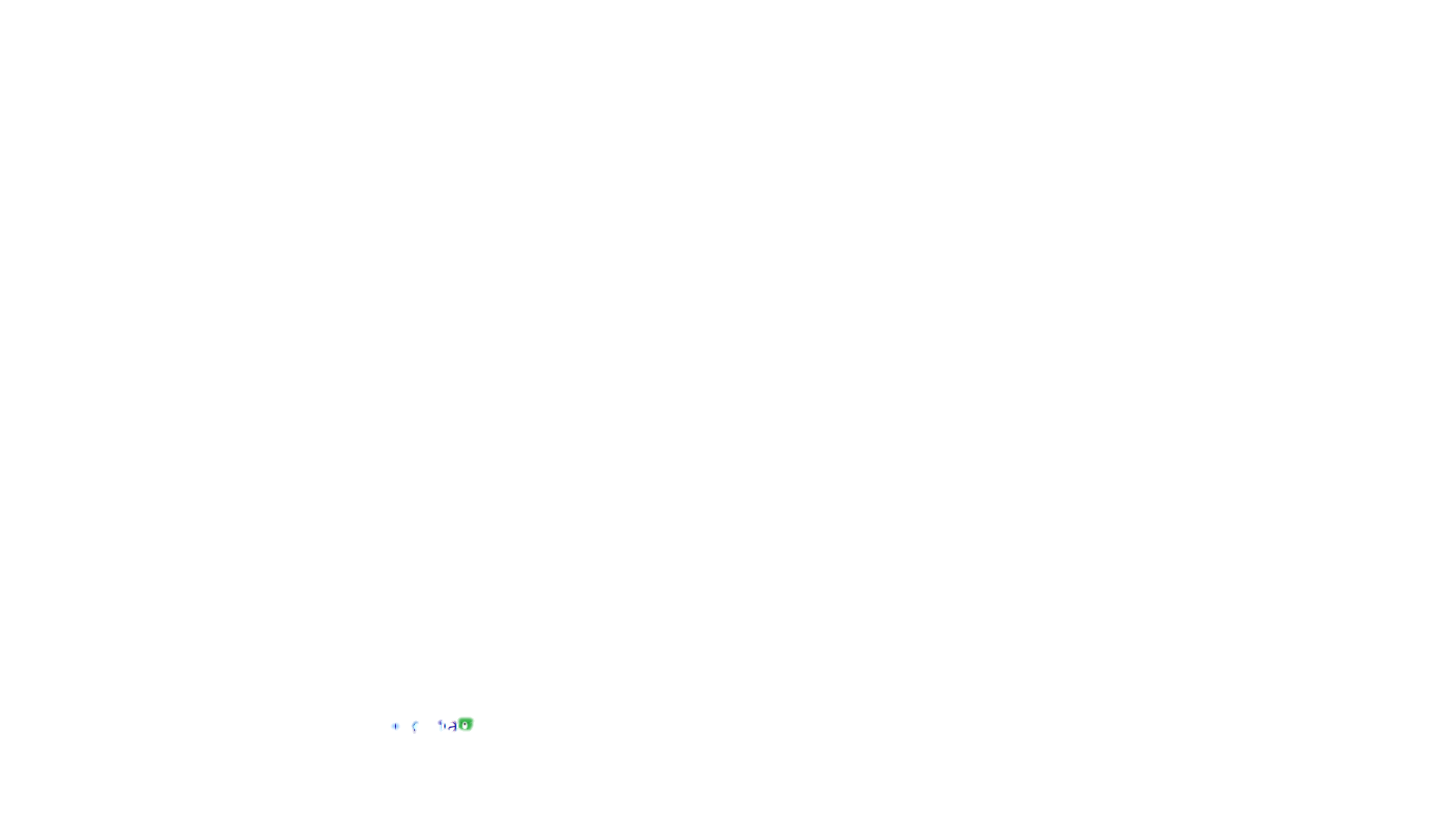
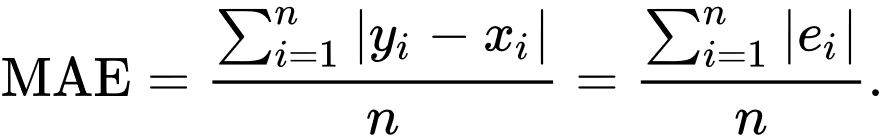
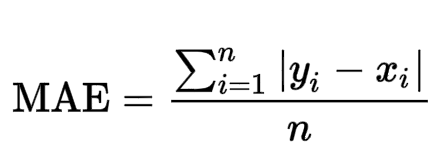
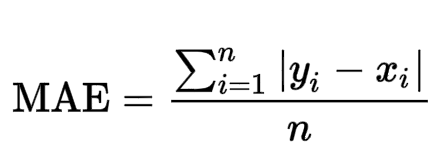
**Model Evaluation:**

The evaluation metrics that are used in this project are:

1. Mean Absolute Error
2. Mean Squared Error
3. Root Mean Squared Error

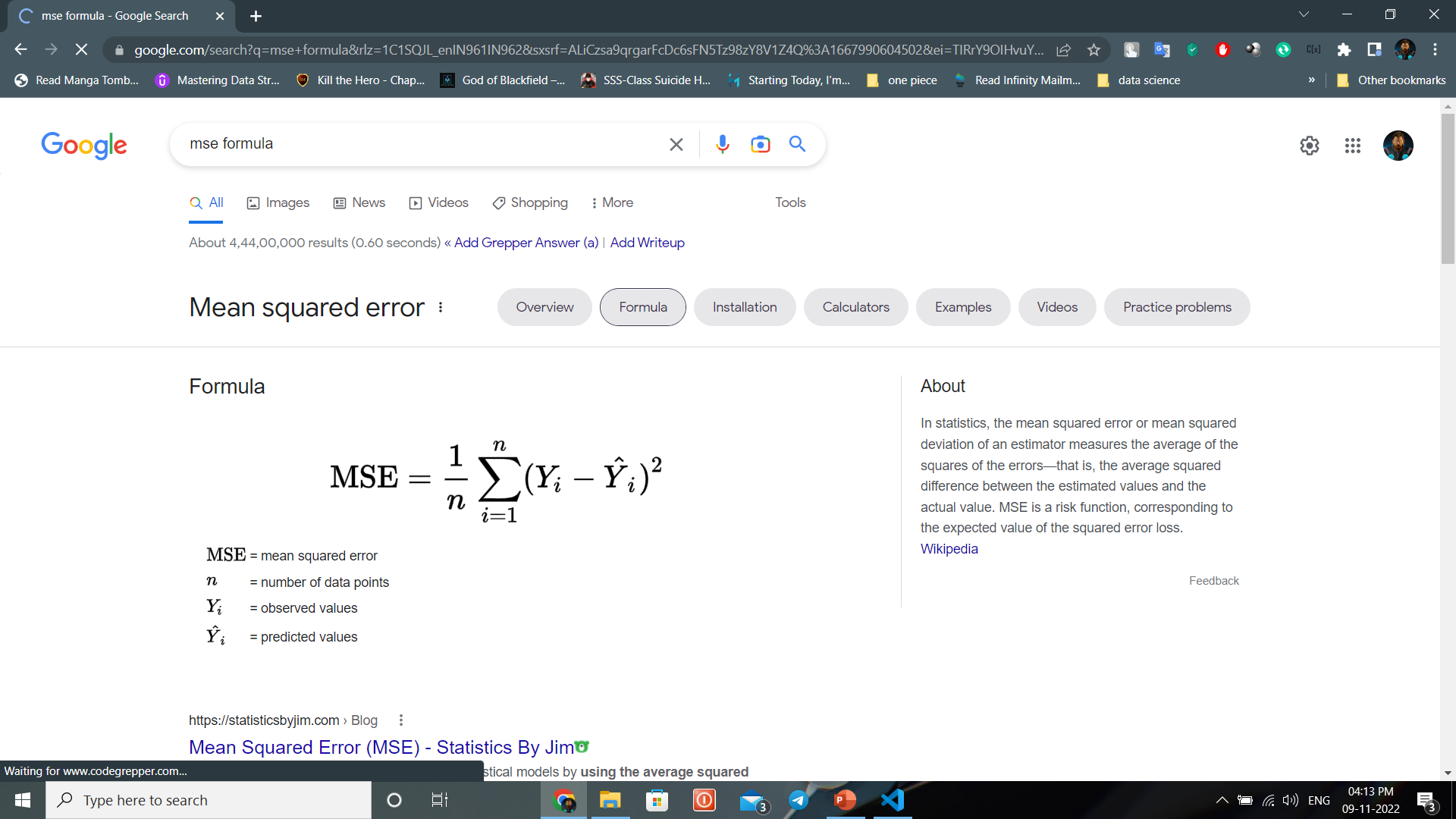
**1.Mean Absolute Error:**

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size

{\displaystyle \mathrm {MAE} ={\frac {\sum \_{i=1}^{n}\left|y\_{i}-x\_{i}\right|}{n}}={\frac {\sum \_{i=1}^{n}\left|e\_{i}\right|}{n}}.}

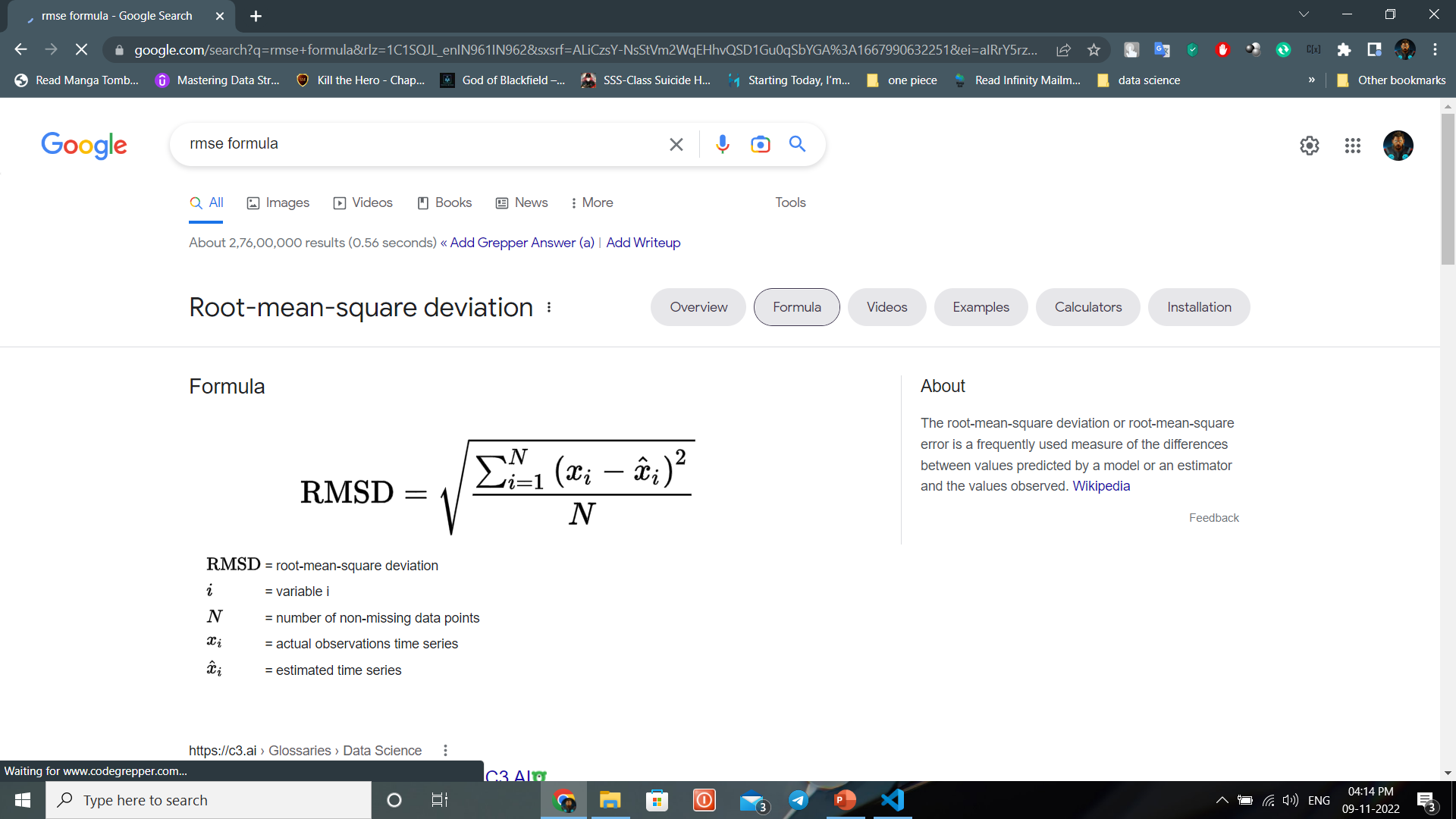
**2.Mean Squared Error:**

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. In machine learning, specifically empirical risk minimization, MSE may refer to the *empirical* risk (the average loss on an observed data set), as an estimate of the true MSE (the true risk: the average loss on the actual population distribution).

****

**3.Root Mean Squared Error:**

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called *residuals* when the calculations are performed over the data sample that was used for estimation and are called *errors* (or prediction errors) when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.



**CHAPTER – 8**

**CODING**

# 8.CODING

**Code to import Regression Techniques:**

from sklearn.ensemble import RandomForestRegressor

from sklearn.linear\_model import LinearRegression

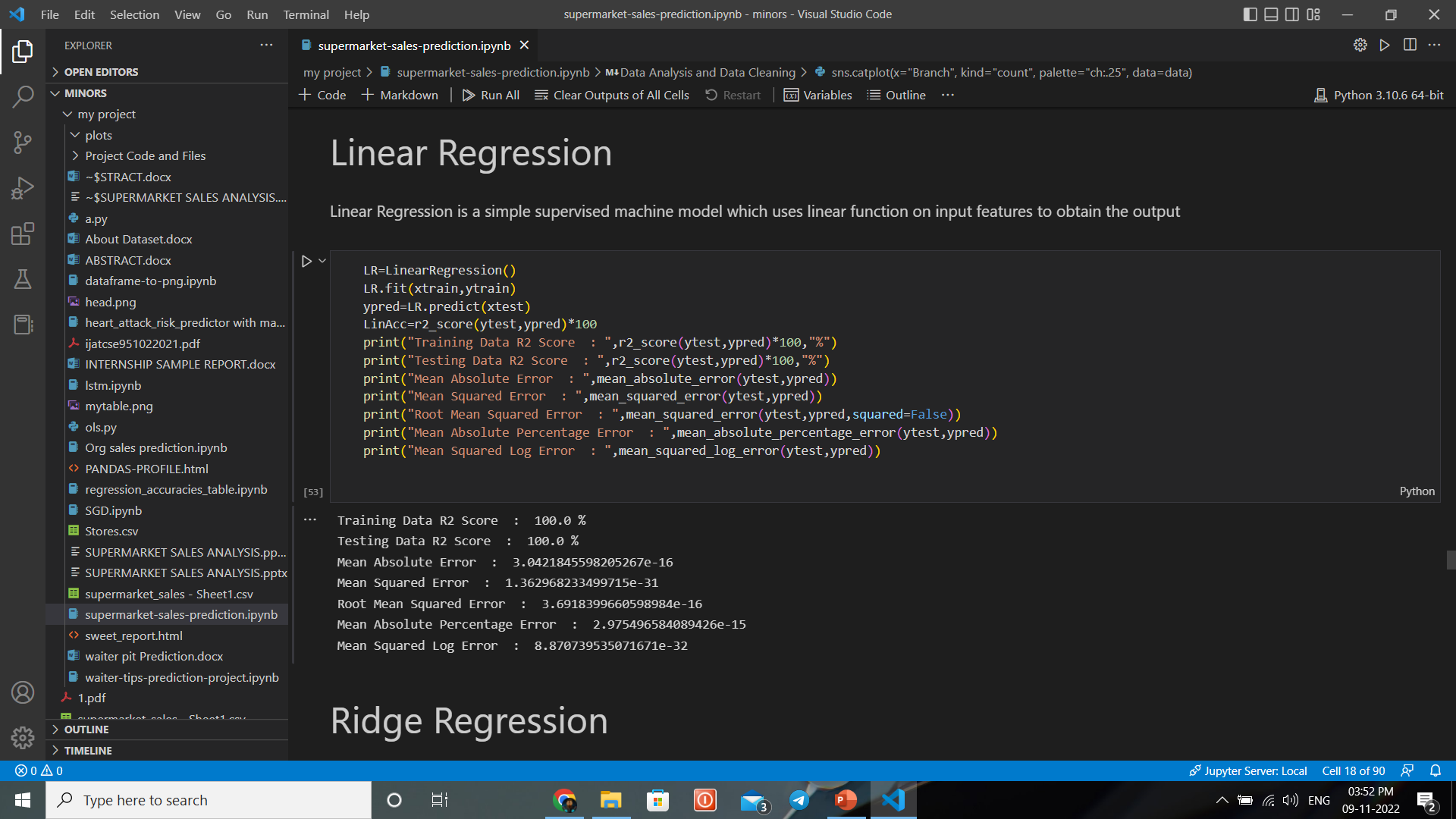
from sklearn.neighbors import KNeighborsRegressor

from sklearn.tree import DecisionTreeRegressor

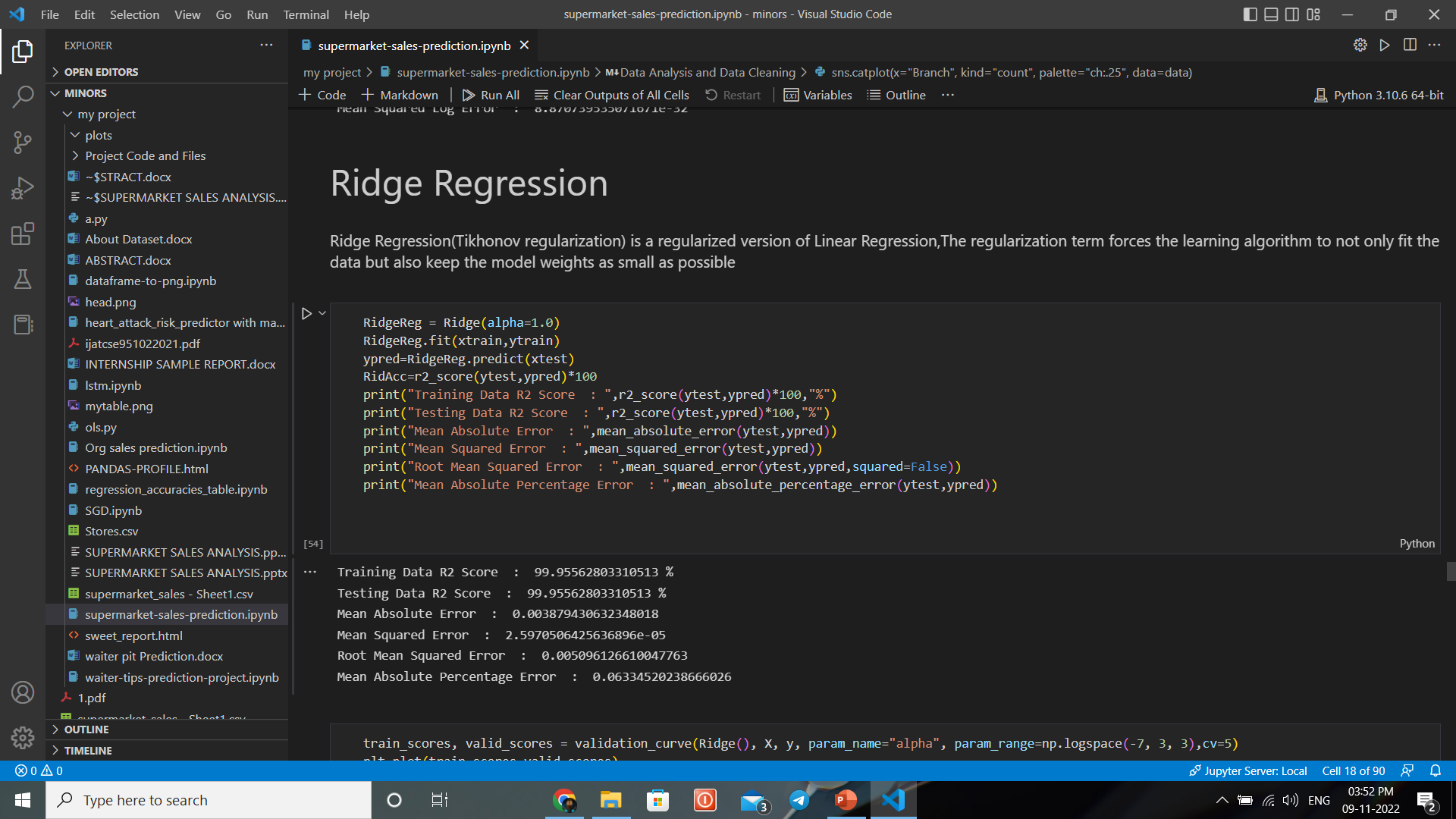
from sklearn.svm import SVR

from sklearn.metrics import r2\_score, mean\_absolute\_error, mean\_squared\_error

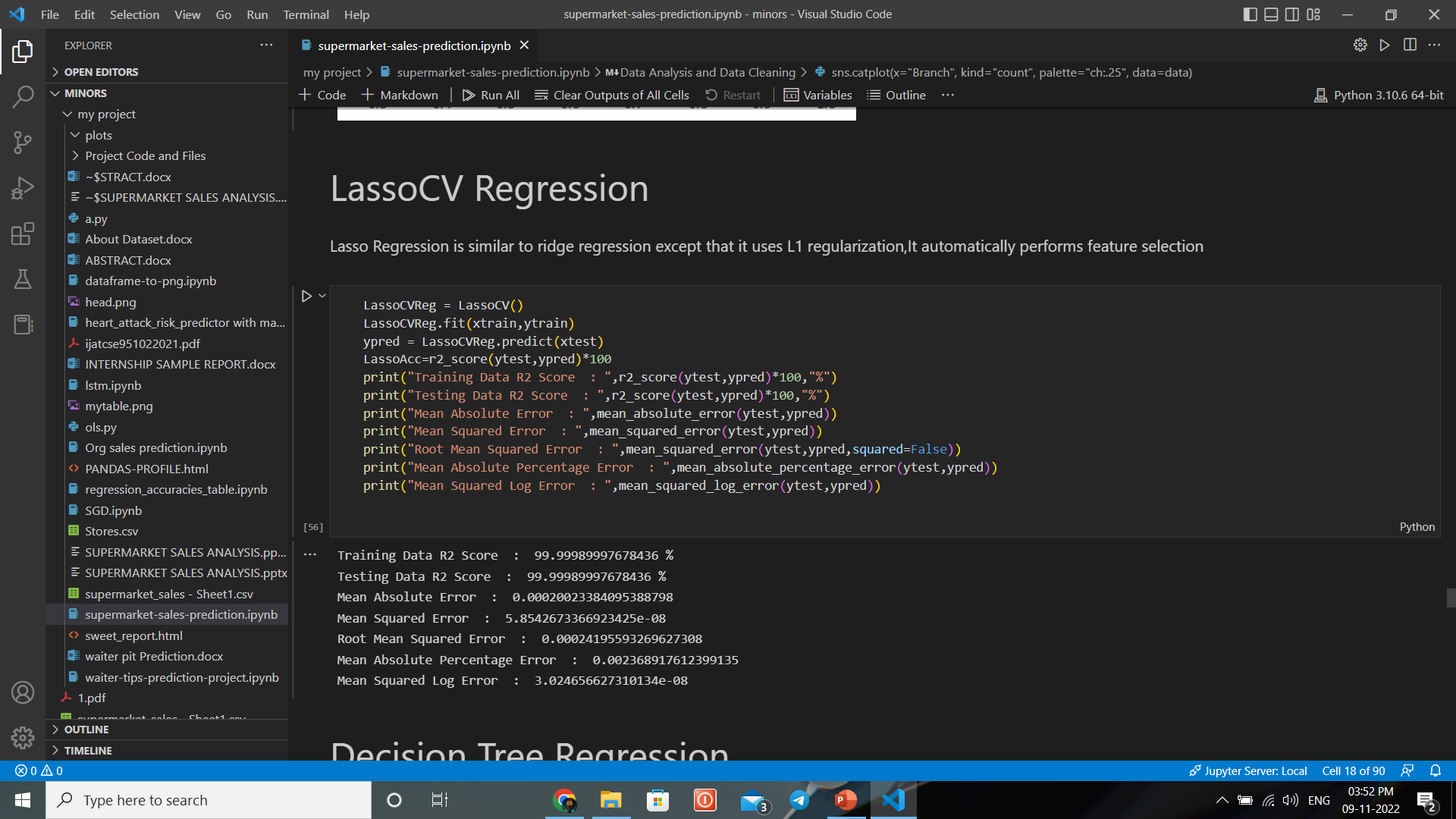
**Implementing Linear Regression:**



**Implementing Ridge Regression:**



**Implementing LassoCV Regression:**



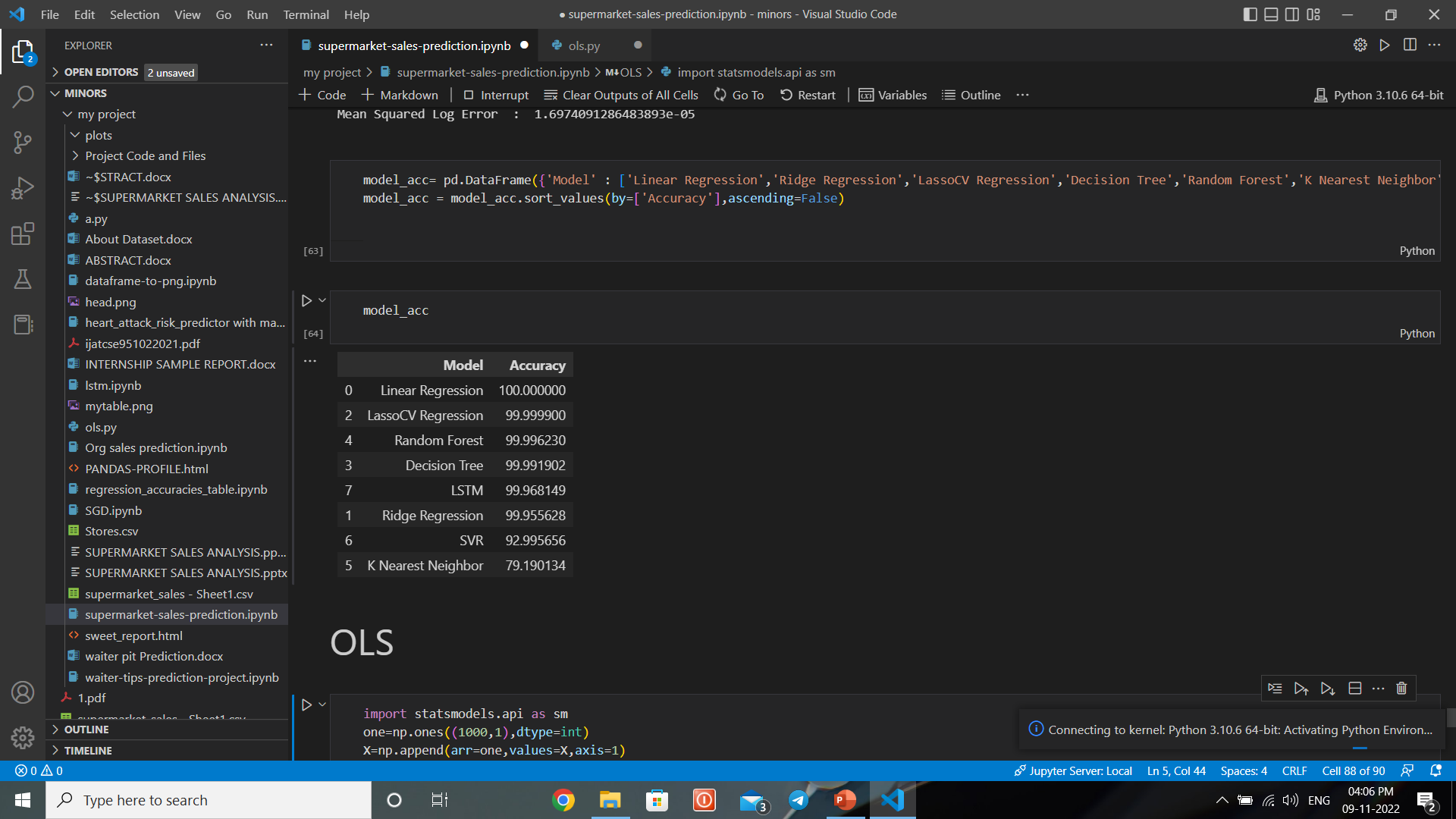
In addition to the above Regression Algorithms, we included some other techniques like Decision Tree Regression, Random Forest, LSTM Regression, OLS Regression, K-Nearest Neighbors.

**CHAPTER – 9**

**RESULTS**

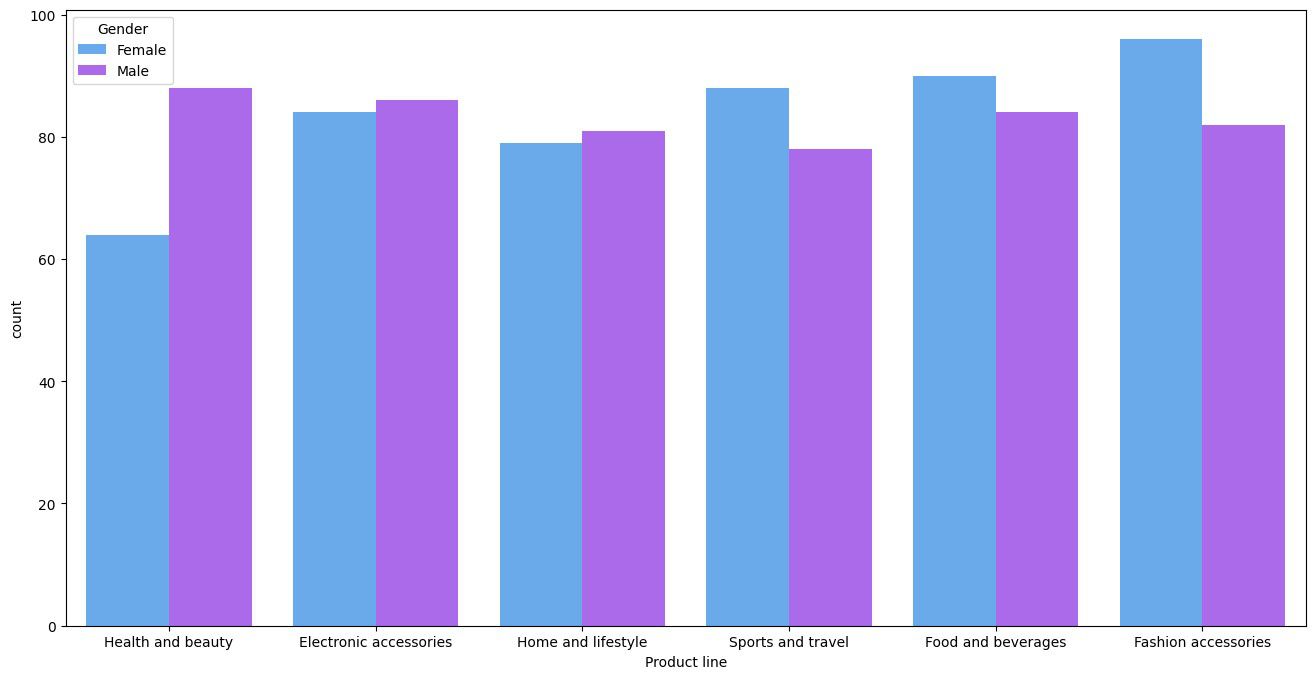
# 9.RESULTS

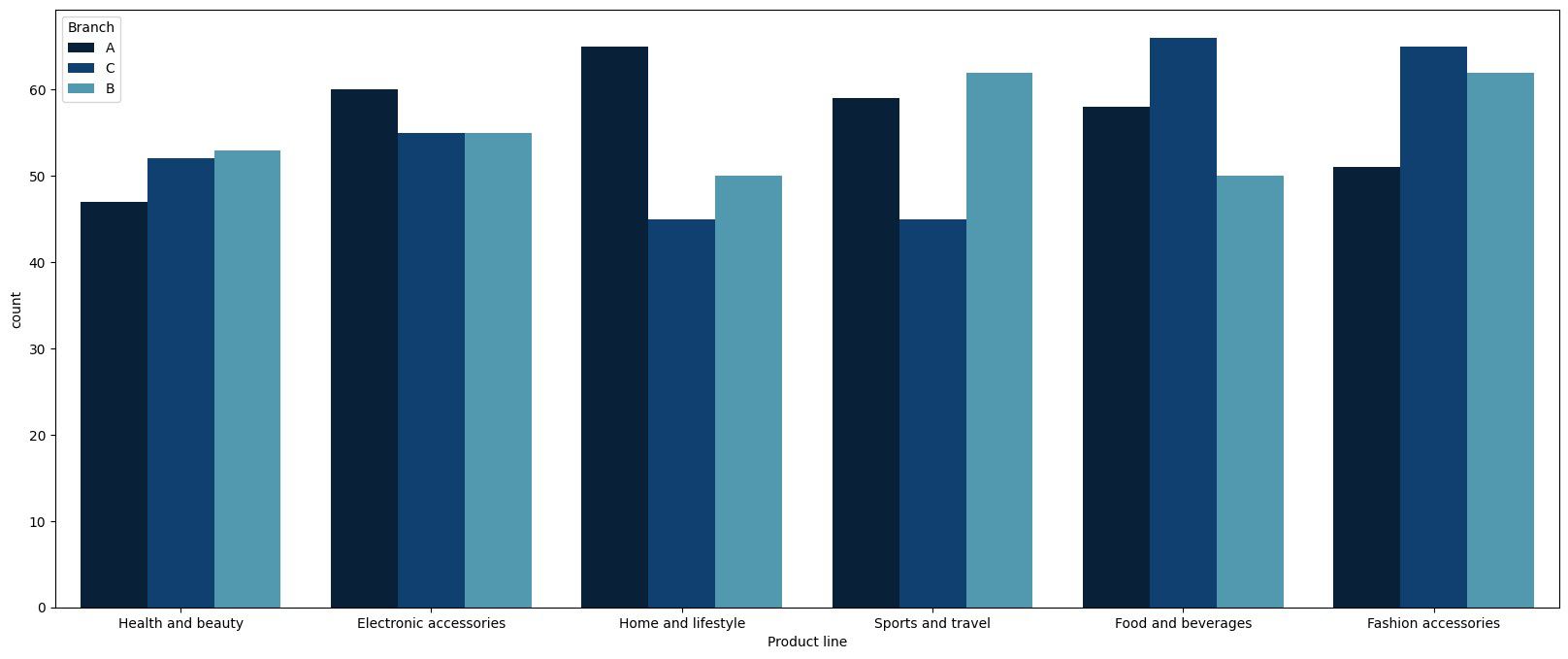
Out if all the Machine Learning Algorithms we have executed, Linear Regression Model has the highest Accuracy & K-Nearest Neighbor has the Lowest Accuracy.

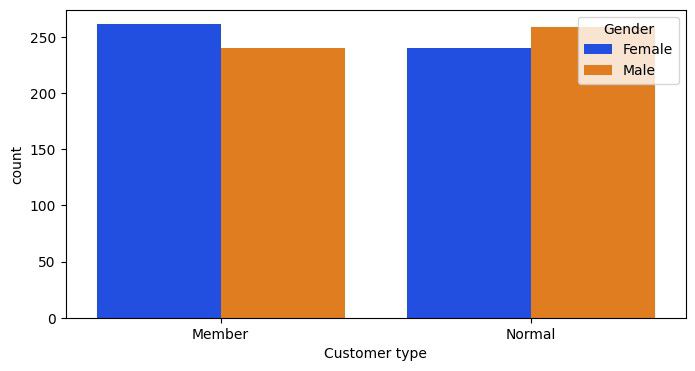
The Accuracies of all the Machine Learning Algorithms are as follows:

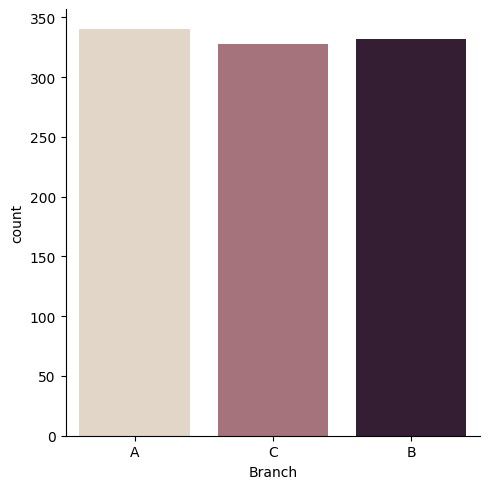
So, by taking all the accuracies into note, we selected Linear Regression algorithm as the ideal Machine Learning Regression Algorithm for prediction of future sales of Supermarkets.

**Insights**:





****

****

**CHAPTER – 10**

**CHALLENGES FACED**

# 10.CHALLENGES FACED

* When we first collected our dataset from Kaggle database we found that there are no null values in it. But it had Categorical data which has to be converted into numerical values. So, in order to covert categorical data into numerical values we used “pd.get\_dummies()” function.
* This technique creates a separate column for each categorical value. Although it might be quite useful, it creates insanely large number of variables or columns.
* If a certain column contains large number of categories, the “pd.get\_dummies()” function will create the columns that are equal to the number of categories thereby increasing the size of the dataset.
* The computation time of the program will directly depend on the size of the data that is used in it. If the dataset size is huge then the time complexity of the program will increase thereby making the program less efficient.
* The substitution of the “pd.get\_dummies()” is to use Label Encoder to encode the categorical data. This technique will not create any additional columns and thereby not effecting the size of the dataset.

**CHAPTER – 11**

**CONCLUSION**

# 11.CONCLUSION

The project as a whole describes the scope and viability of the agriculture industry and mainly of the financial, technical and its market potential. The project guarantee sufficient fund to repay the loan and also give a good return on capital investment. When analyzing the social- economic impact, this project is able to generate an employment of 5 and above. It will cater the demand of Agriculture and thus helps the other business entities to increase the production and service which provide service and support to this industry. Thus, more cyclic employment and livelihood generation. So, in all ways, we can conclude the project is technically and socially viable and commercially sound too.

When we take a close look at the Debt Service Coverage Ratio (DSCR), the avg: DSCR is 4.07 : 1, which is at a higher proposition and proposes a stable venture The Profit and Loss shows a steady growth in profit throughout the year and the firm has a higher Current Ratio (average) of 10.48, this shows the current assets and current liabilities are managed & balanced well.

This project helps supermarket owners who wants develop their business. We have plotted many

Visualizations from the supermarket sales dataset. From those visualizations we have to select useful insights and make use of them in order to increase sales in the future. We have applied most of the Regression models and from those models, Linear Regression Model gives predictions with the highest accuracy.

**CHAPTER – 12**

**FUTURE SCOPE**

**12.FUTURE SCOPE:**

* It reduces the time and manpower required for management and maintenance of different tasks.
* It reduces the paper work in existing system; hence it is economical and efficient.
* With this system, customers get quality of service; customers can even give feedback which can be

stored in the database.

* As the entire system is fully computerized, records of daily and monthly purchases

and sales can be recorded and analyzed.

* This system is very secure, user-friendly, and reliable.

**CHAPTER – 13**

**REFERENCES & APPENDIX**

**13.REFERENCES**

* Kaggle.com
* GeeksforGeeks
* Tutorials Point
* Wikipedia
* Javatpoint
* Analytics Vidya
* Towards Data science
* John Paul Muller and Luca Massaron, “Deep Learning for Dummies”

**APPENDIX**

Resources for Model Building, Data Analysis and Data Visualization:

1. Python
2. Jupyter Notebook
3. Visual Studio Code 2022
4. Anaconda