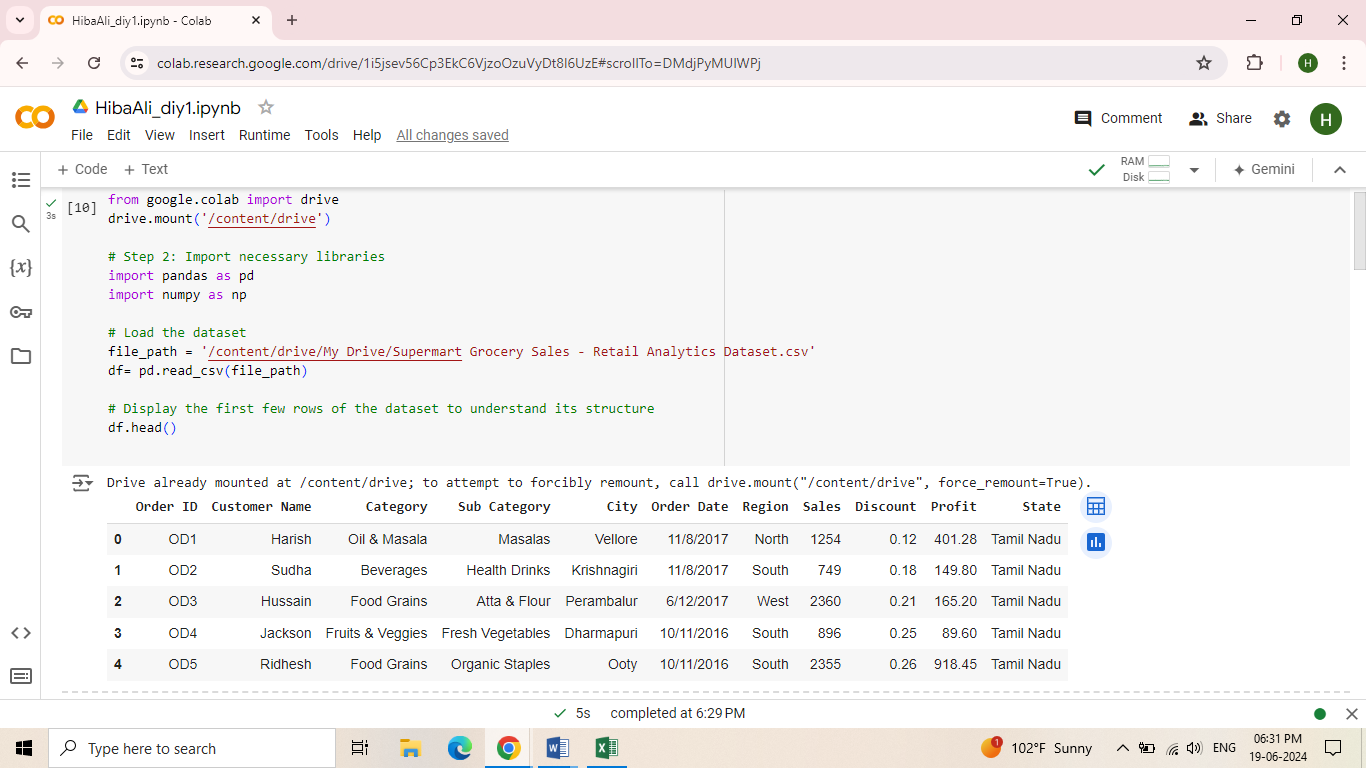
Supermarket Grocery sale – Retail Analytics Data set

1. 10 major KPI (Key performances indicator) from the data set are:
   * City wise, Region wise Total Sales Revenue
   * Region wise Gross Profit Margin
   * City wise Net Profit Margin
   * Region wise Sales Growth Rate
   * City wise purchase of sub-category
   * Year wise profit
   * City wise Average discount

* Region wise sales Average
* Category wise average profit
* Category, sub- category wise Average sales

1. Loading the Dataset and Performing Data Pre-processing, Outlier Detection and Exploratory Data Analysis of the dataset:

LOADING THE DATASET



The dataset contains the following columns:

Order ID: Unique identifier for each order.

Customer Name: Name of the customer.

Category: Product category.

Sub Category: Sub-category within the product category.

City: City where the order was placed.

Order Date: Date when the order was placed.

Region: Region where the order was placed.

Sales: Total sales amount for the order.

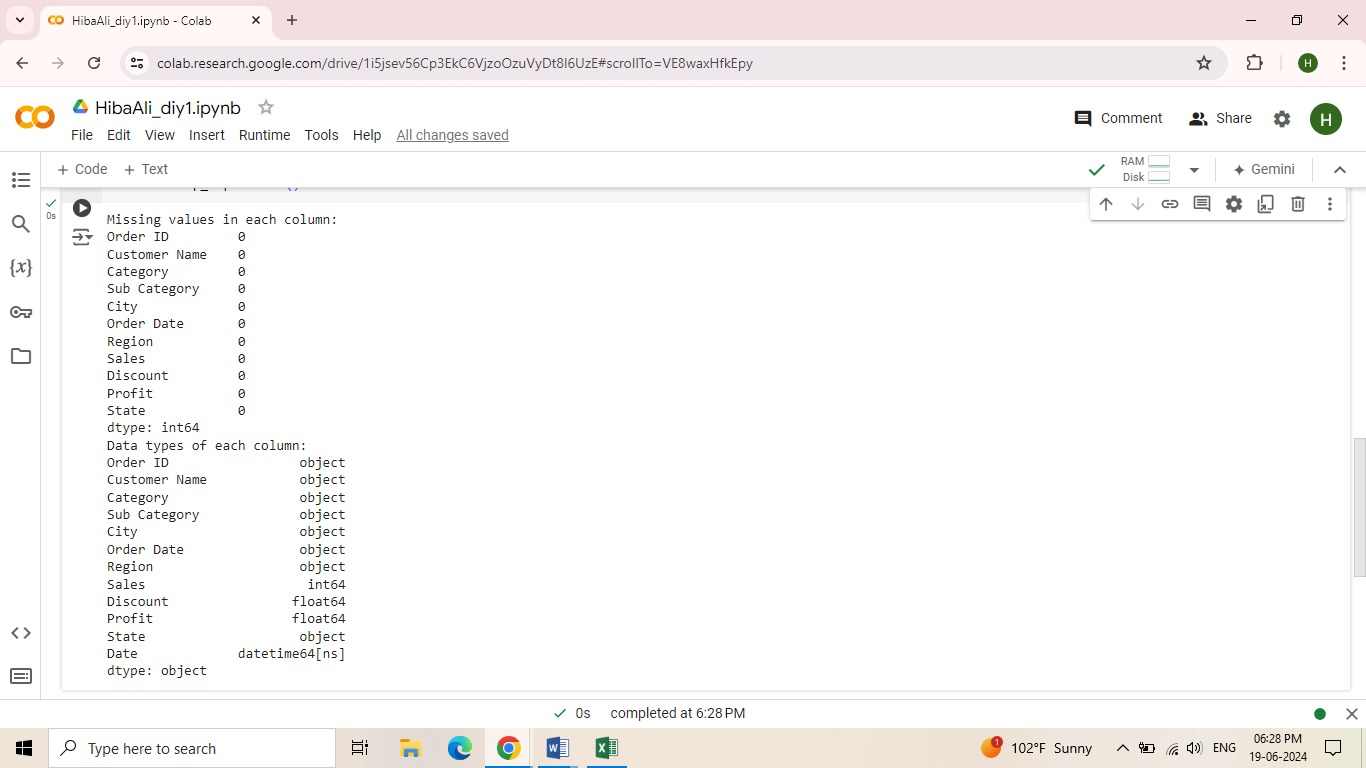
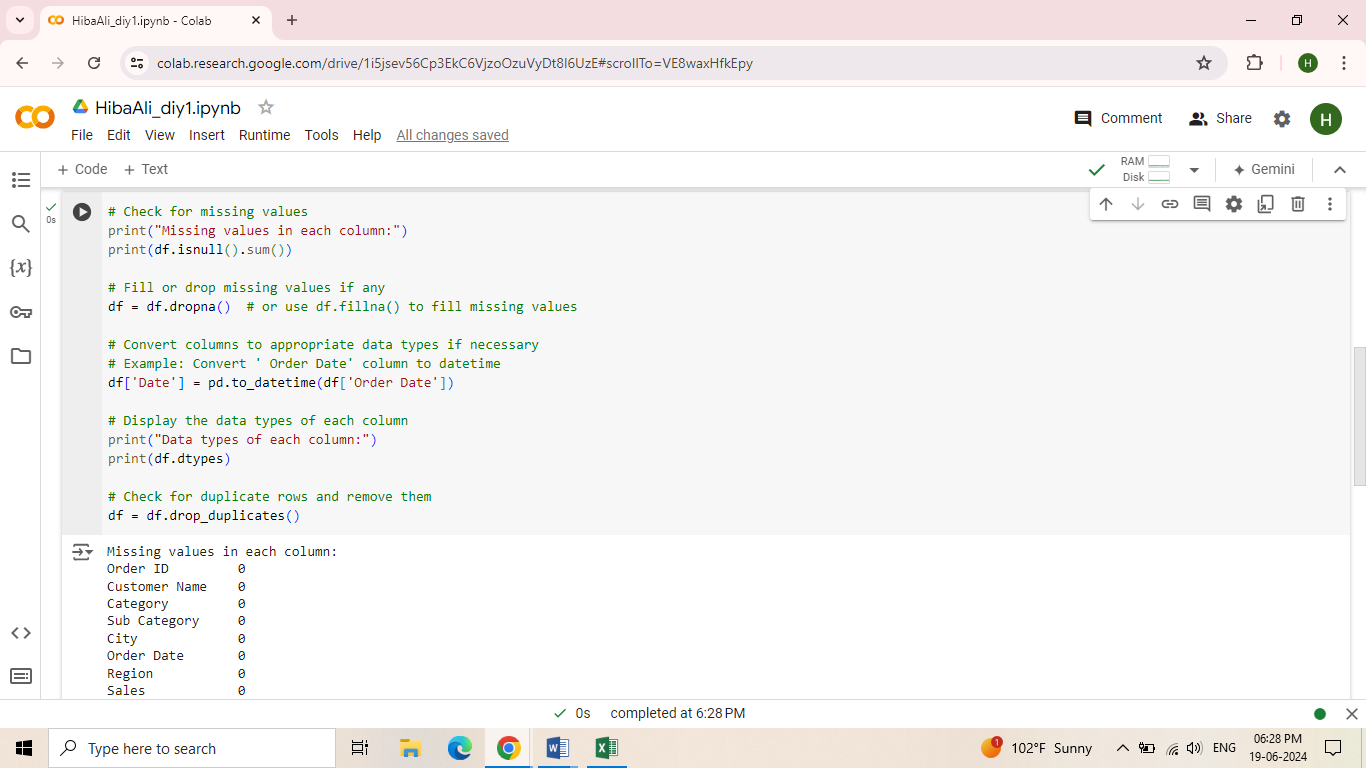
Discount: Discount applied to the order.

Profit: Profit earned from the order.

DATA PRE-PROCESSING:

Data pre-processing is a crucial step in data analysis that involves cleaning and transforming raw data into a format suitable for analysis. This ensures the data's accuracy, consistency, and reliability.

After loading the data set we will check for missing values if there are any missing values in the data set then we will remove it to simplify the process, then converted the data type of the column to appropriate data type and further then Removed the duplicate rows from the data set to ensure uniqueness and avoid skewed analysis results.



OUTLINER DETECTION:

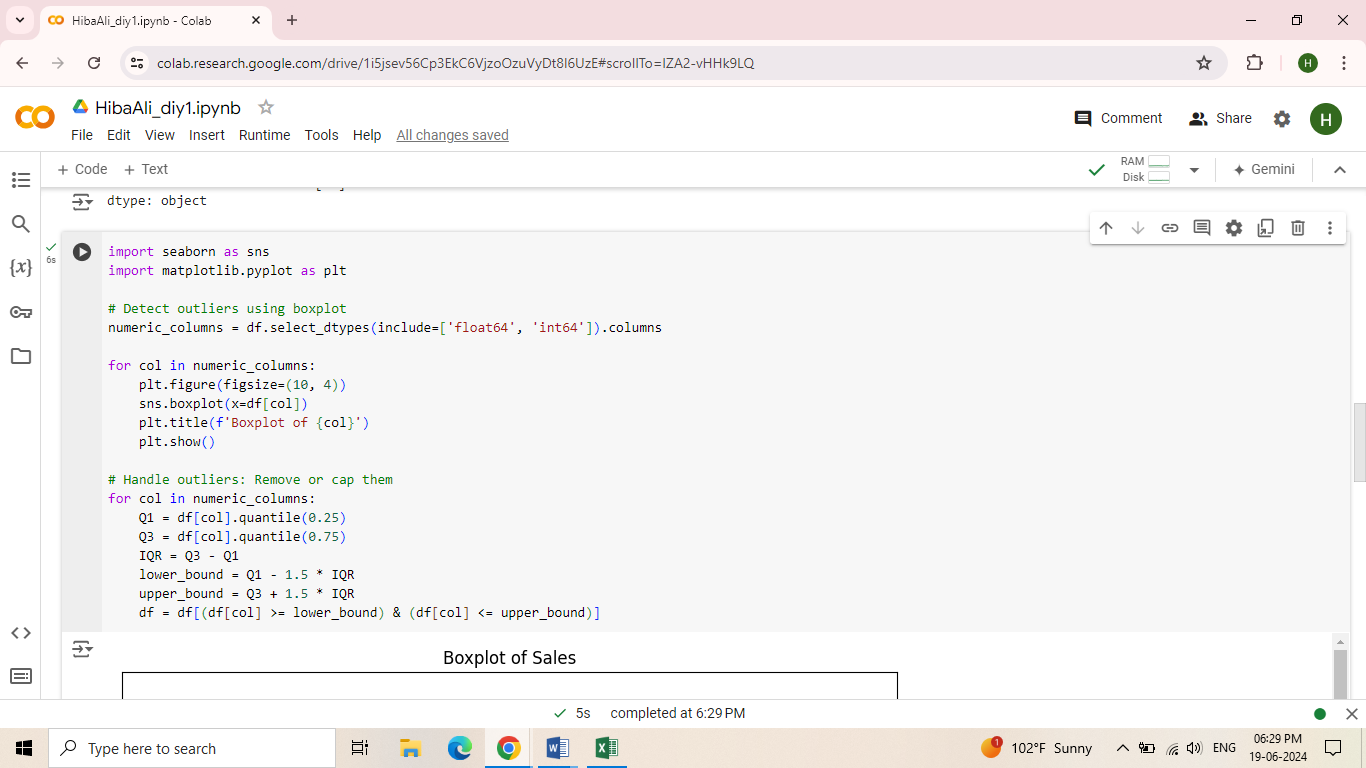
Outlier detection is the process of identifying and handling extreme values that deviate significantly from other observations. These outliers can distort analysis results and must be addressed.

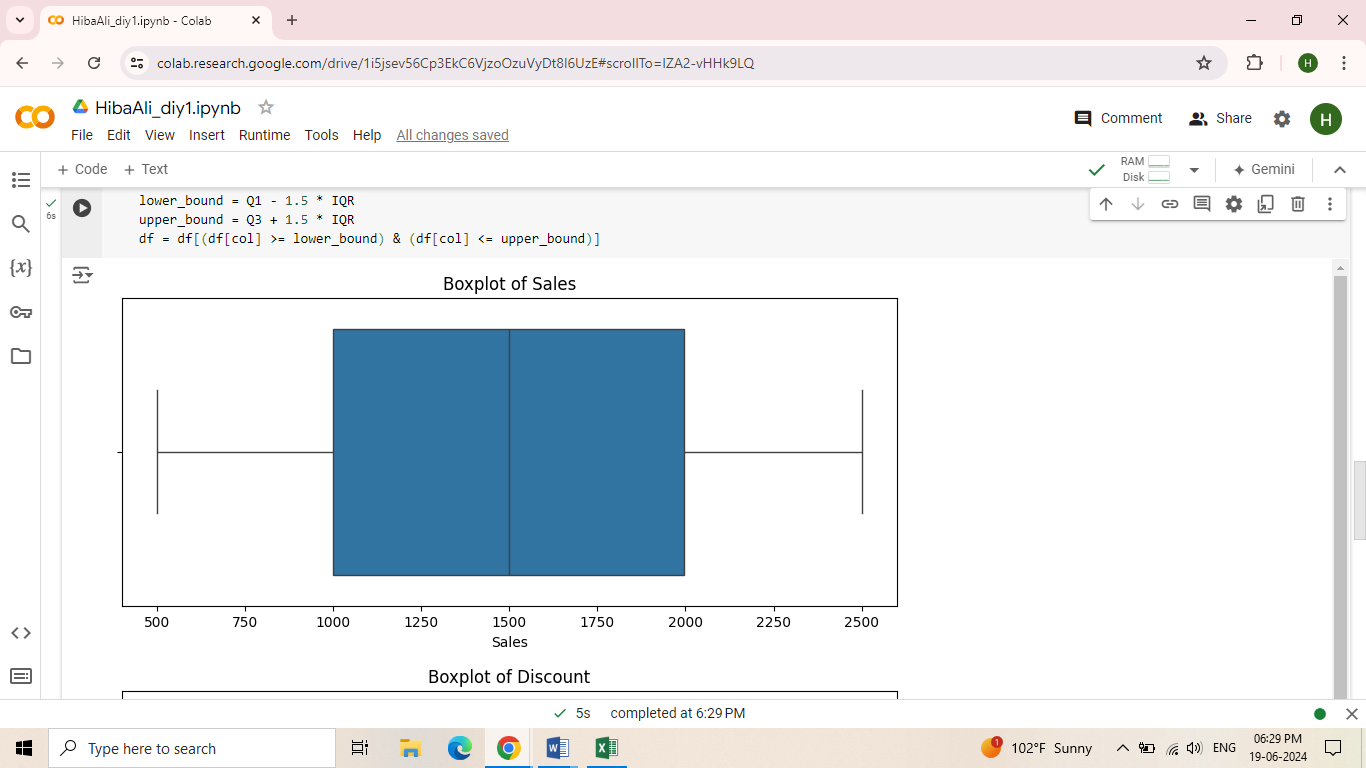
* Visual Detection:

Boxplots were used to visually identify outliers in numerical columns.

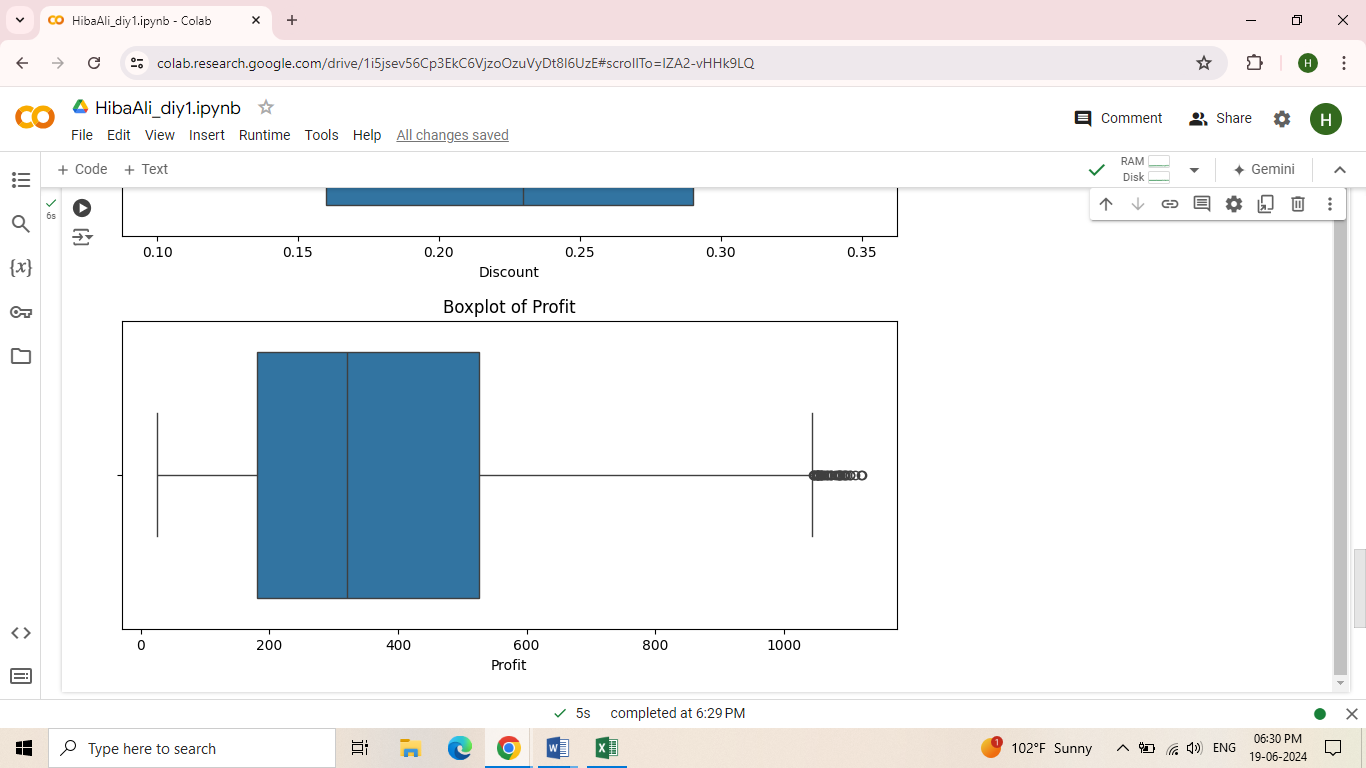
* Statistical Methods:

The Interquartile Range (IQR) was calculated to determine the bounds for outlier detection. Values outside these bounds were considered outliers.





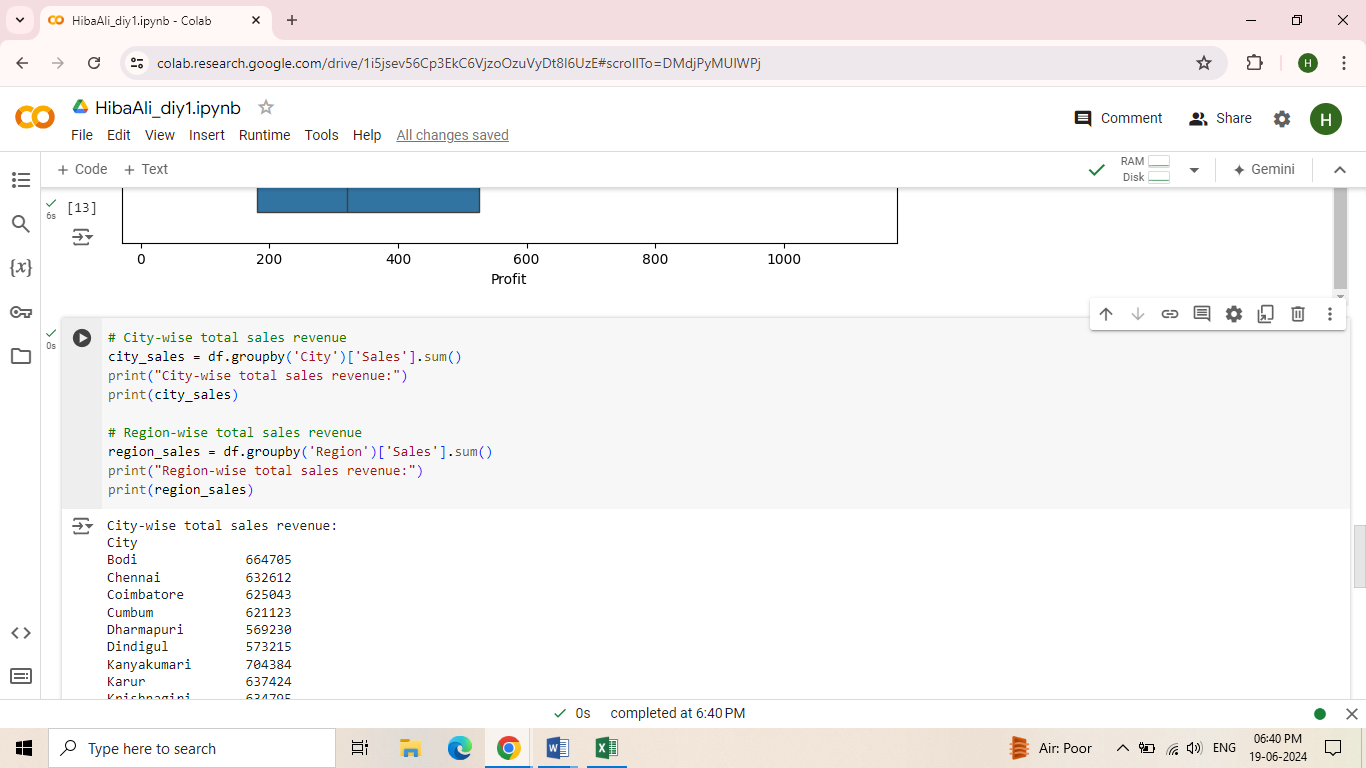


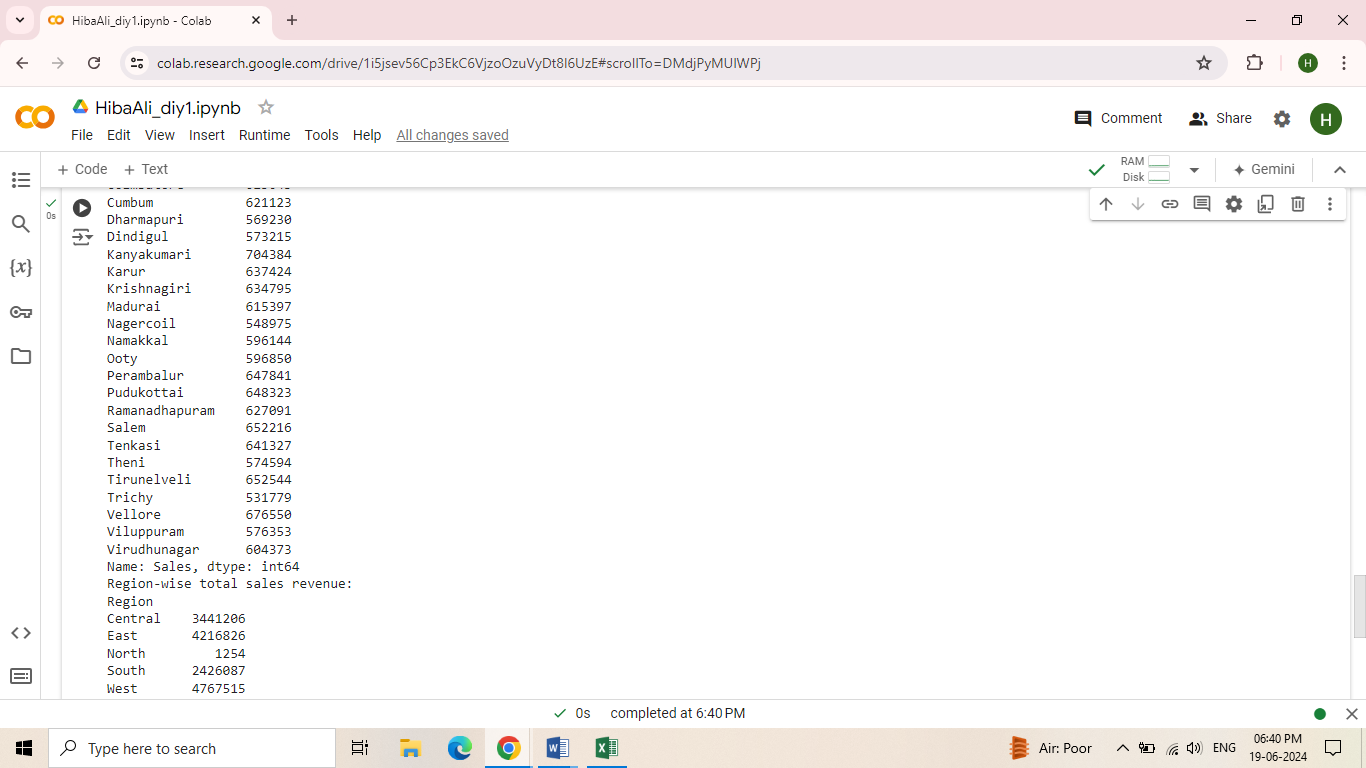


EXPLORATORY DATA ANALYSIS(EDA):

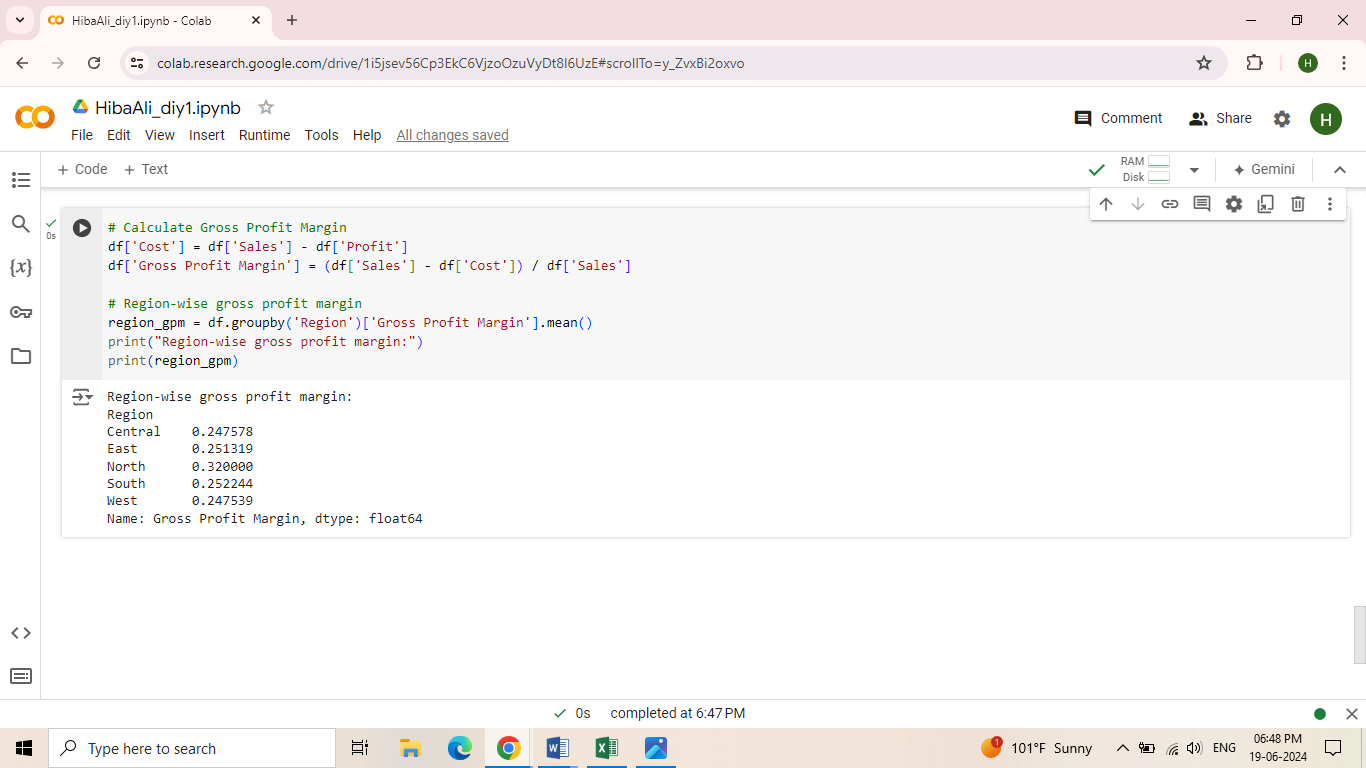
EDA involves analysing and summarizing the main characteristics of a dataset using statistical and visual methods. It helps uncover patterns, spot anomalies, and generate hypotheses for further analysis.

* City wise, Region wise Total sales revenue

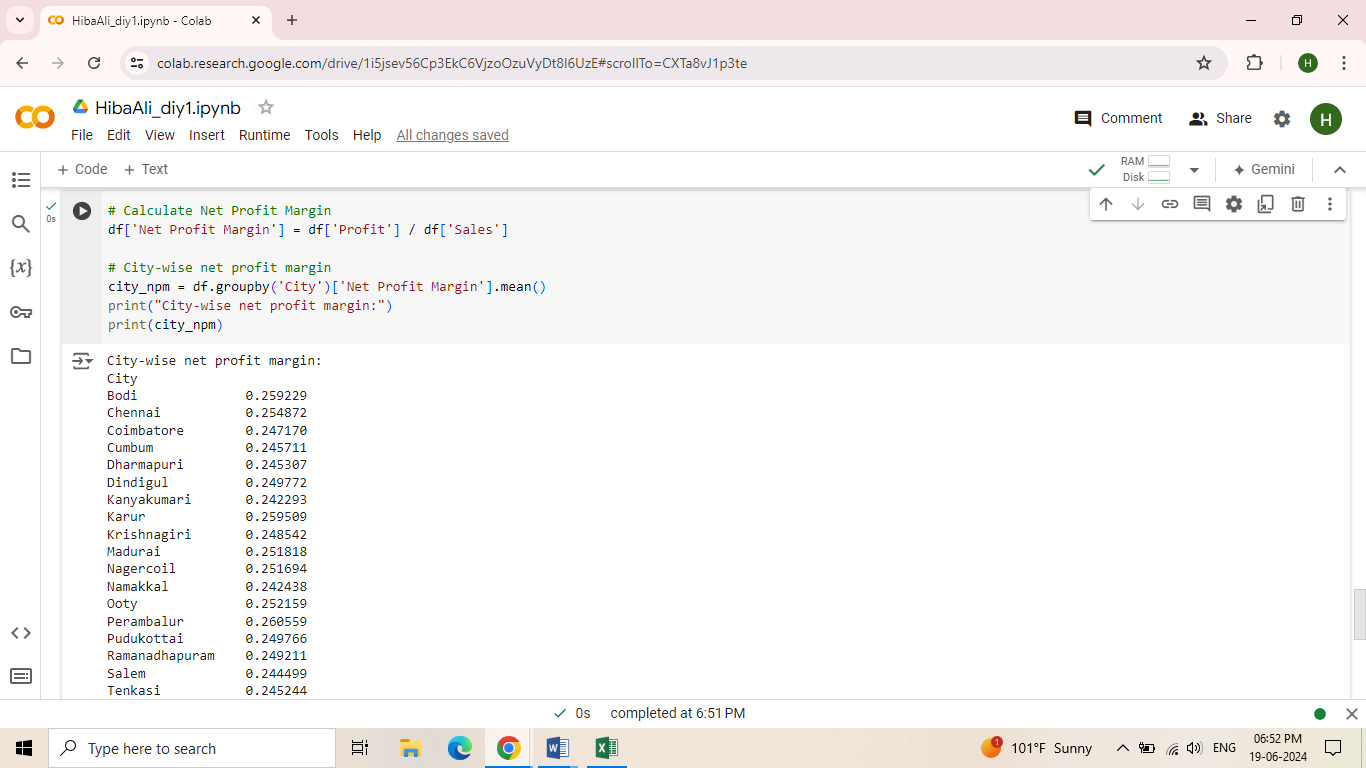


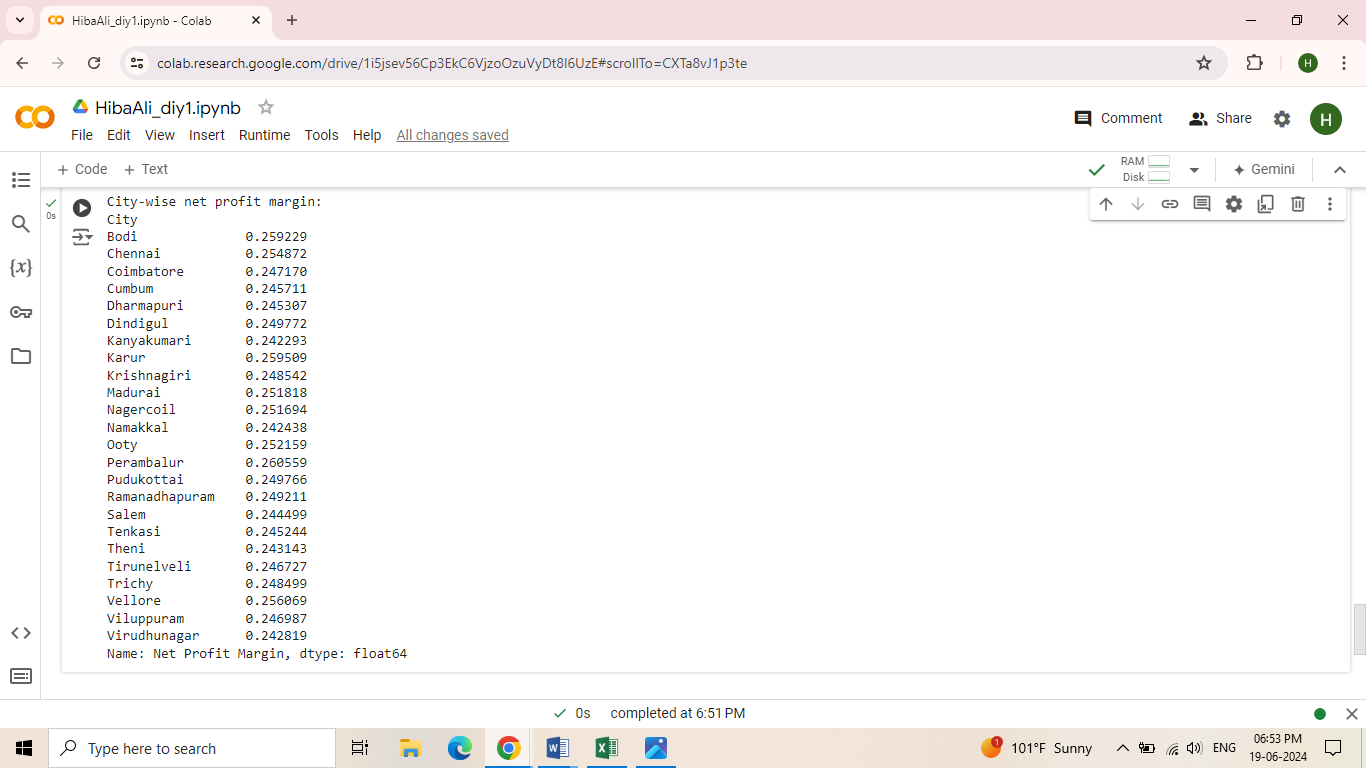


* Region wise gross profit margin

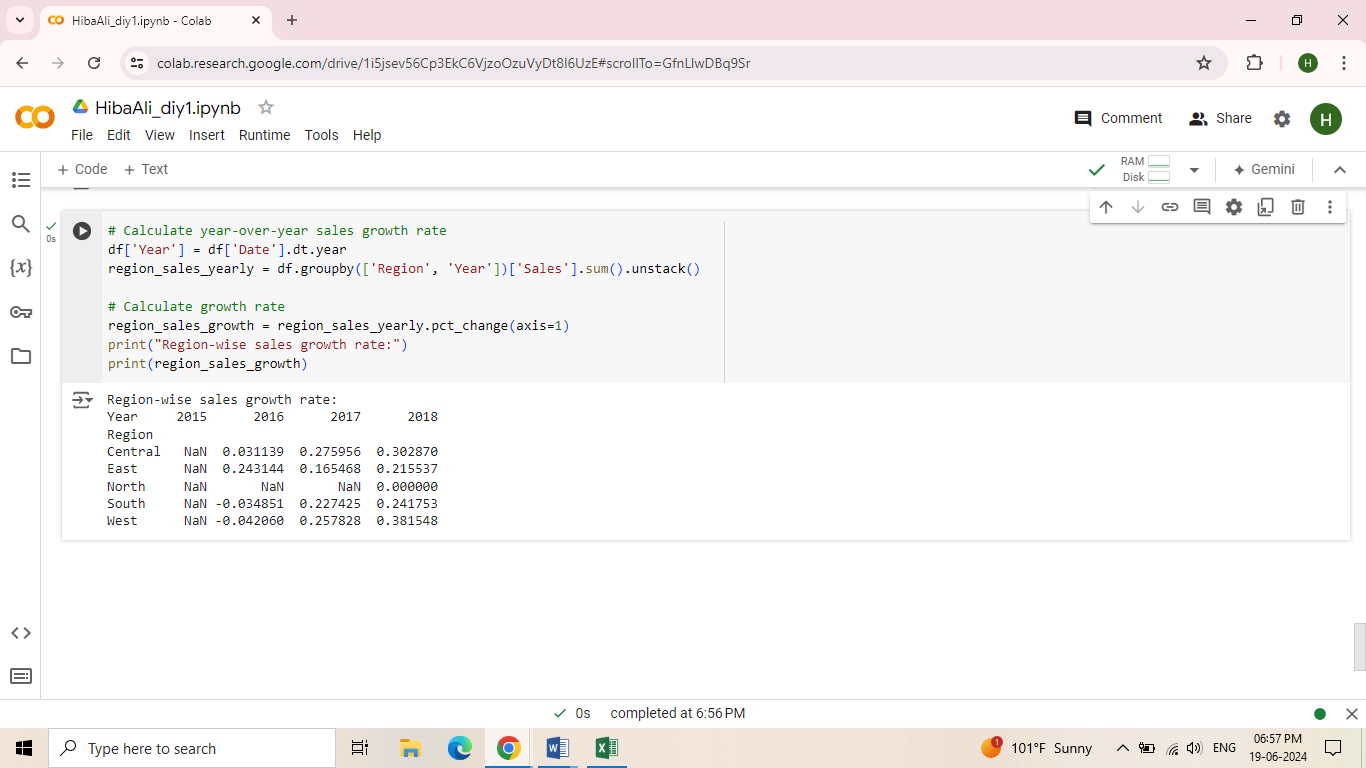


* City wise Net profit Margin

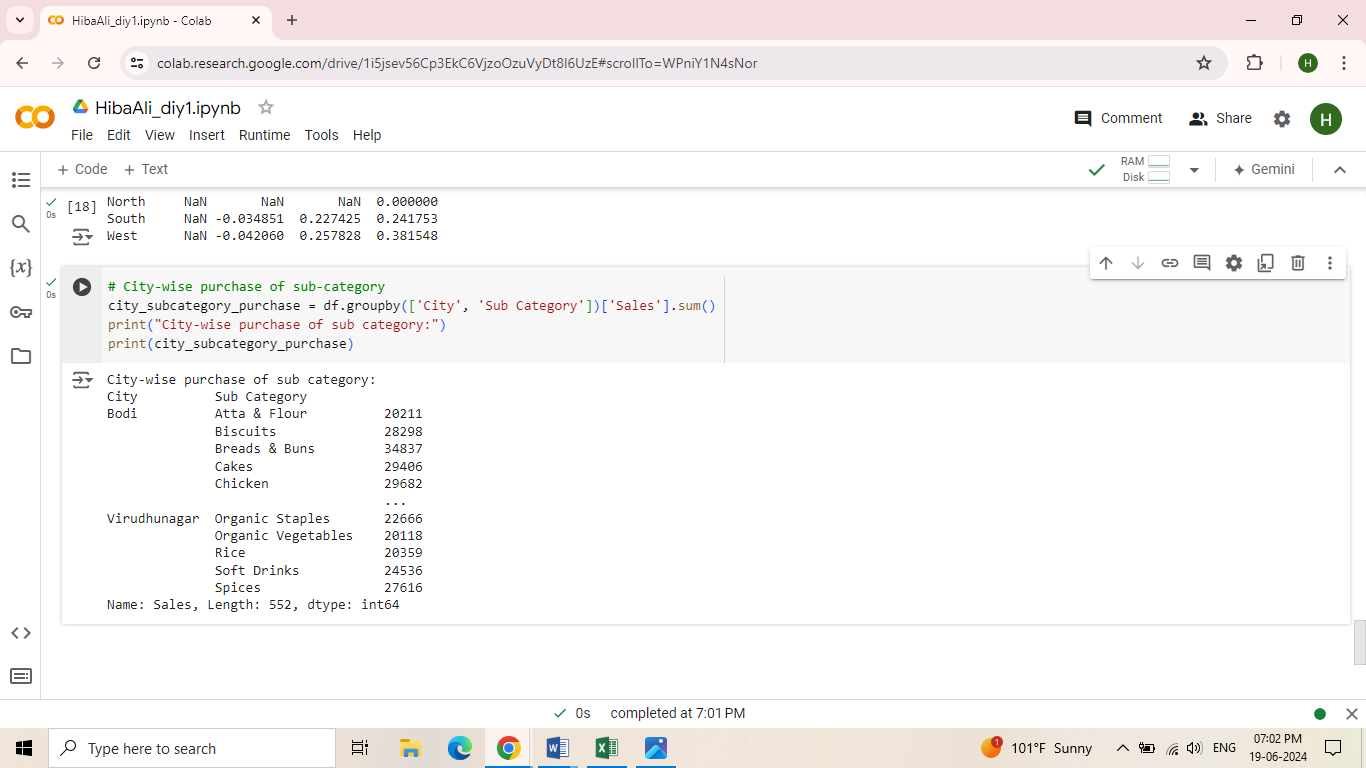




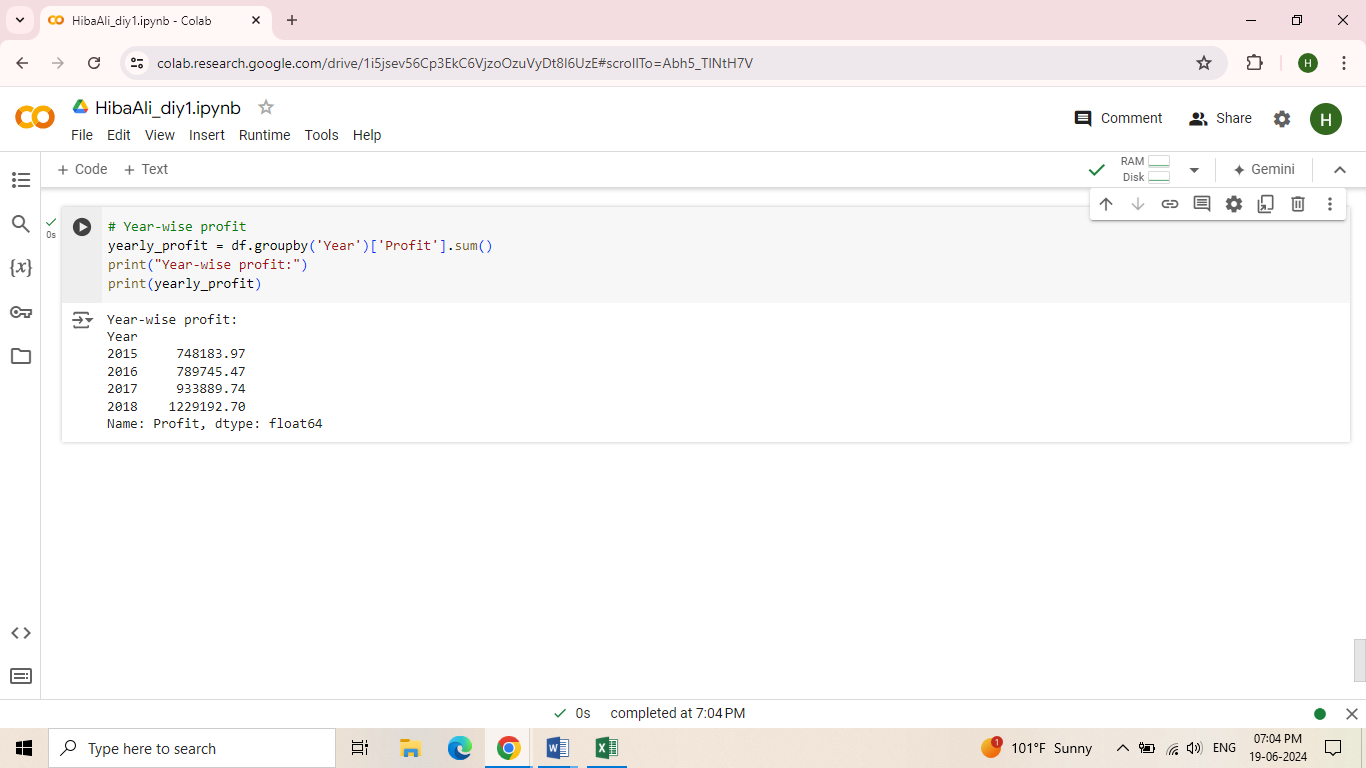
* Region wise Sales growth Rate



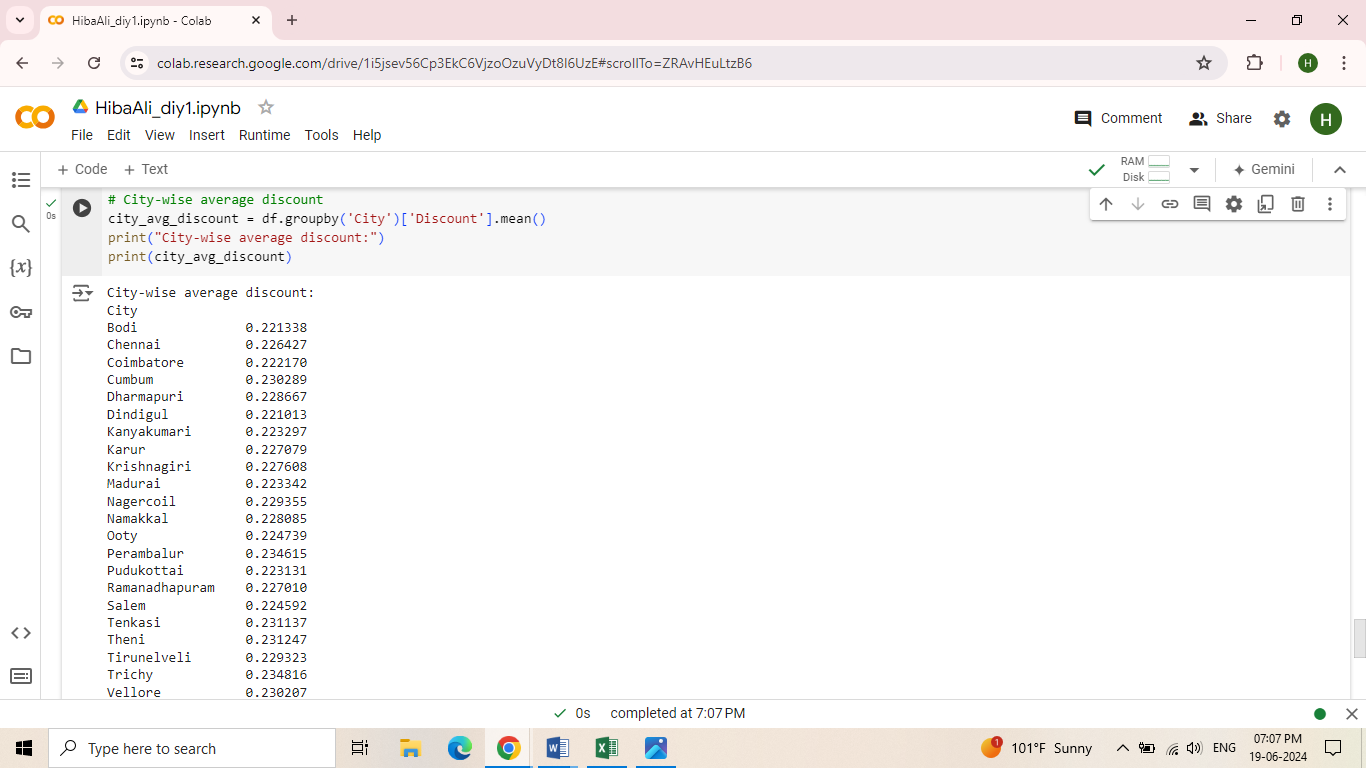
* City wise Purchase of sub category



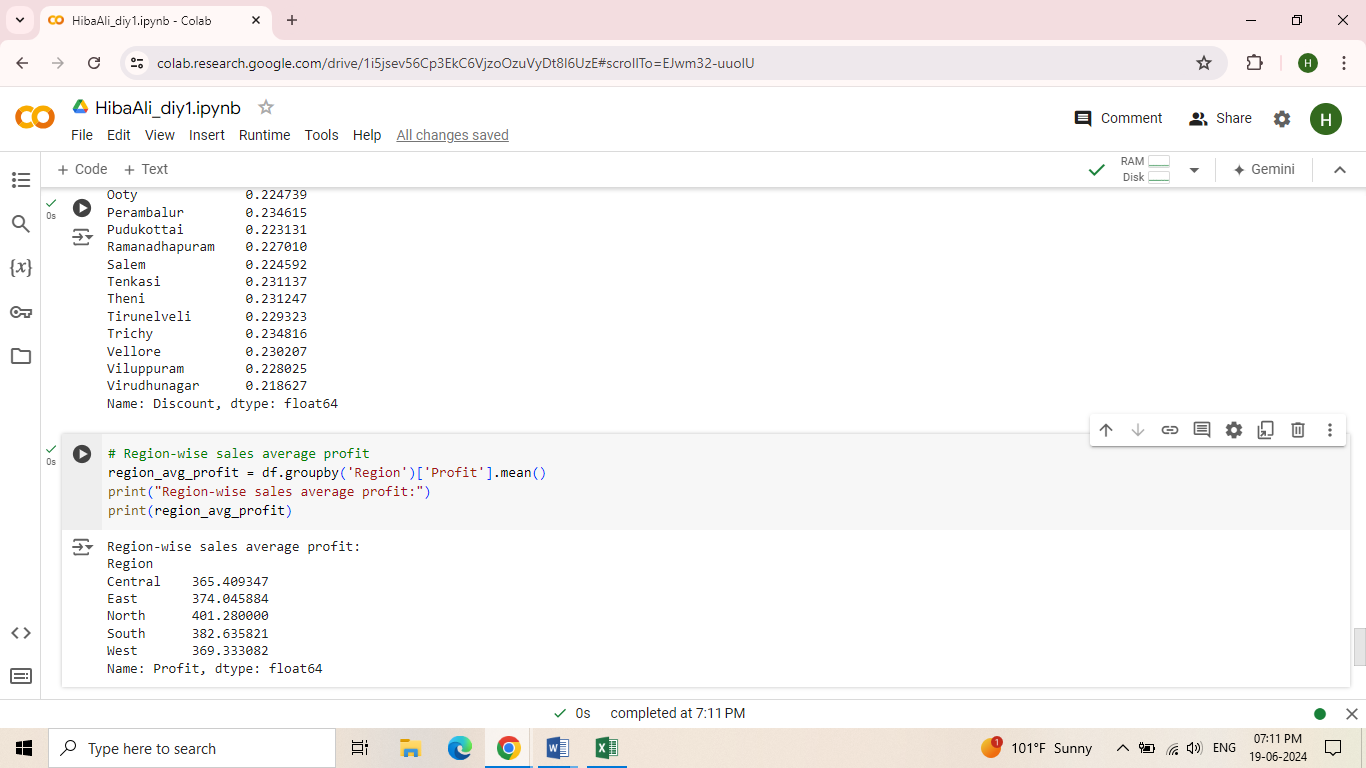
* Year wise Profit



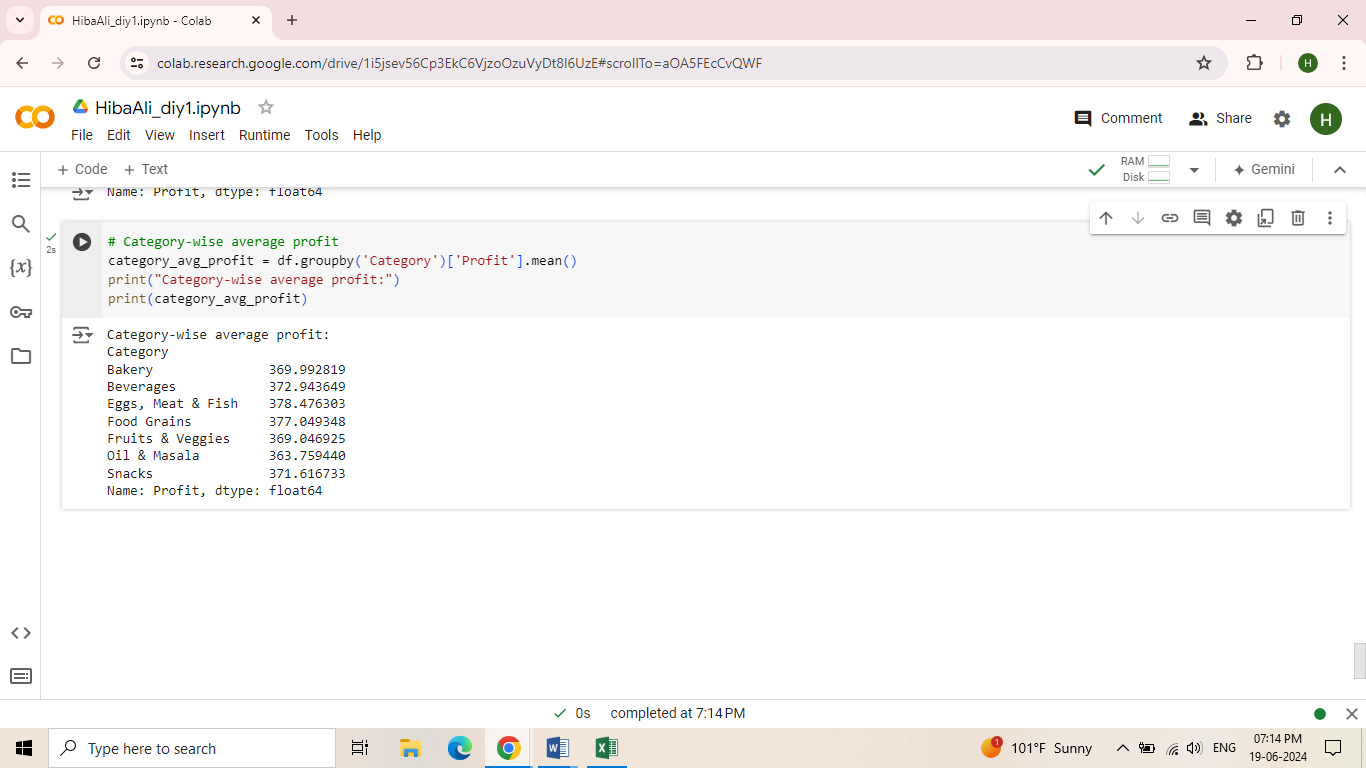
* City wise Average Discount



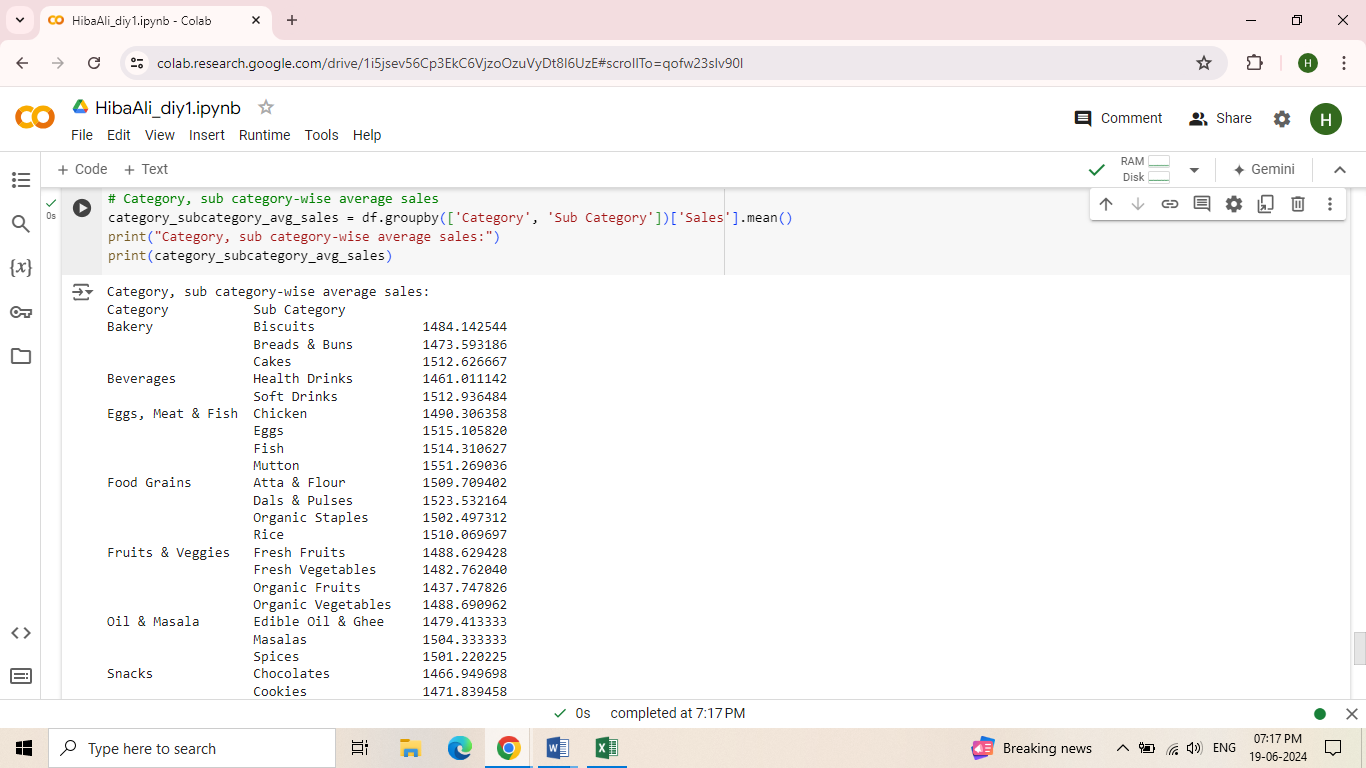
* Region Wise Sales Average profit

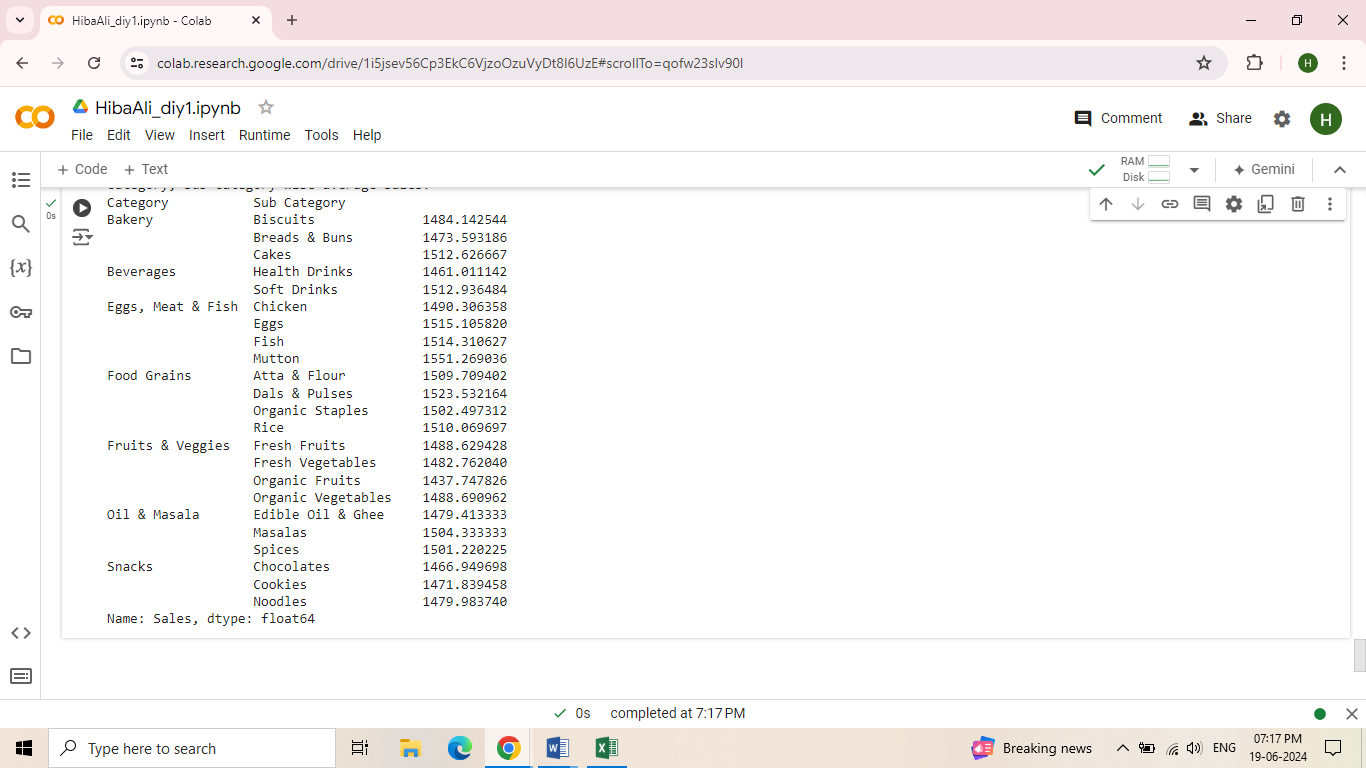


* Category wise Average profit



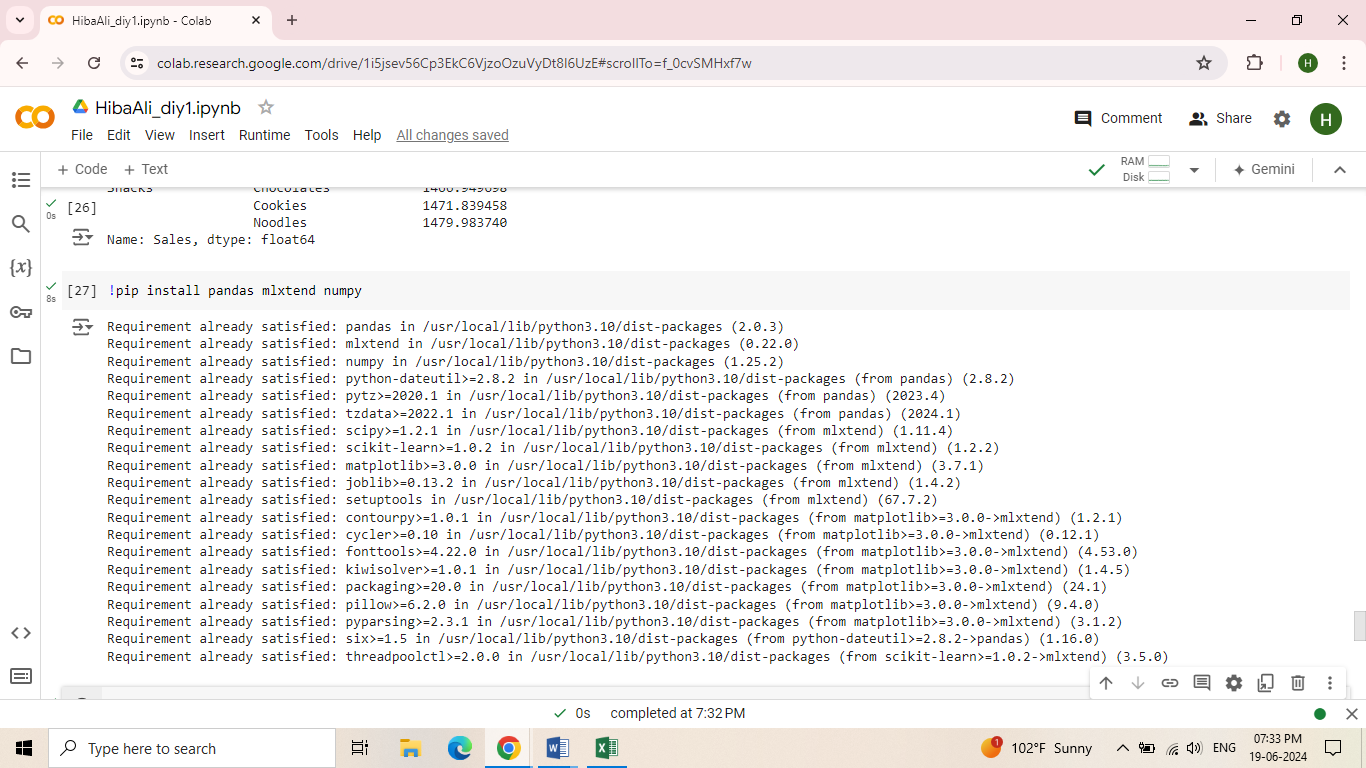
* Category, Sub category wise Average sale



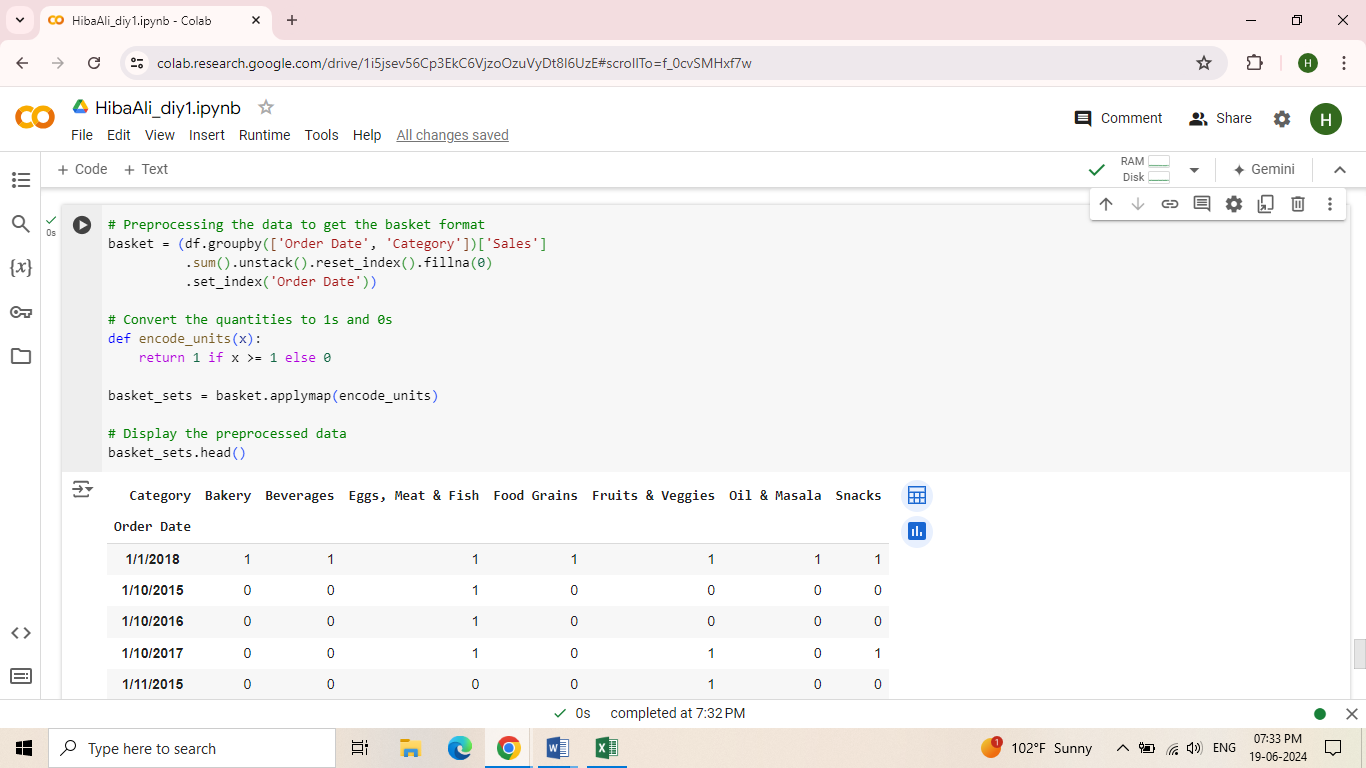


1. Using Association Rule Mining technique to identify the items frequently bought together and their demands:

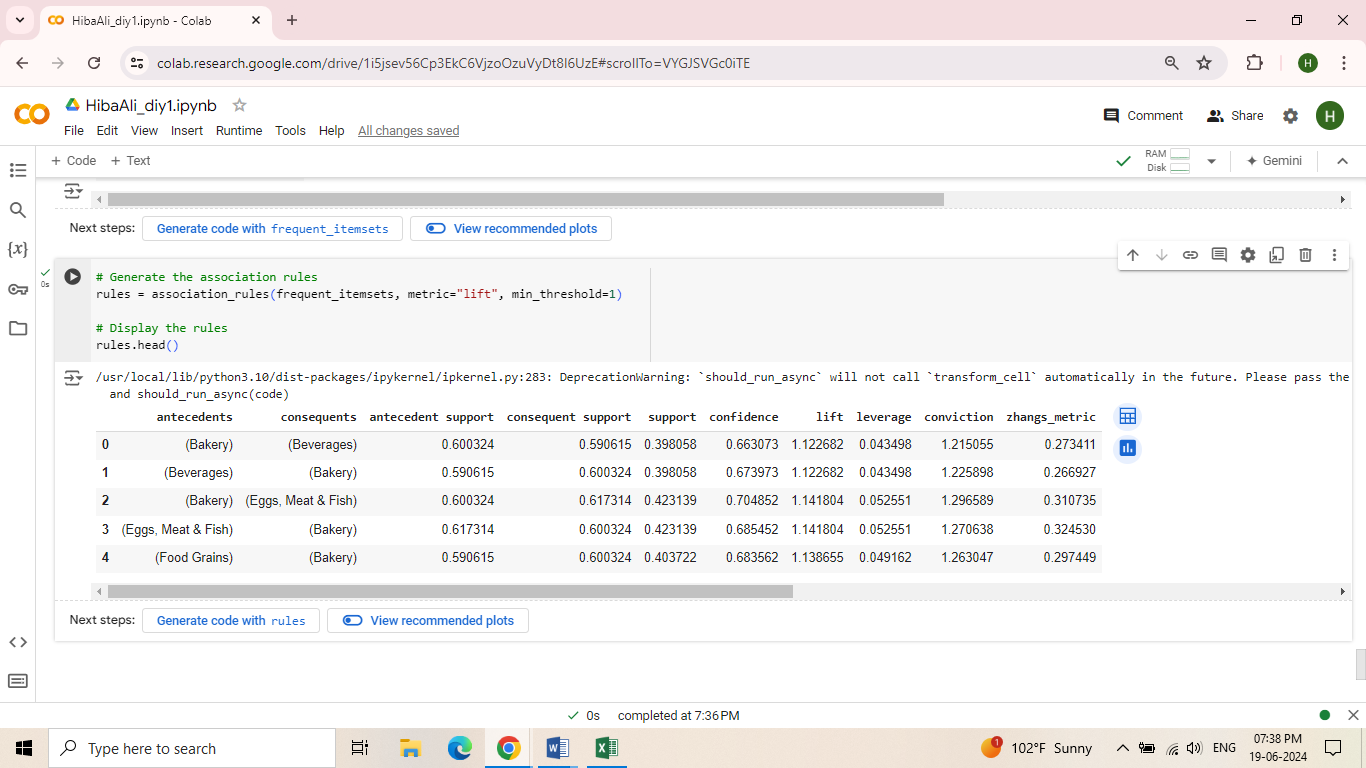
First, we have to install the necessary library “mlxtend”



We'll to pre-process the data to the format suitable for association rule mining. Typically, this involves creating a basket matrix where each row represents a transaction and each column represents an item, with values indicating whether the item was purchased in the transaction.





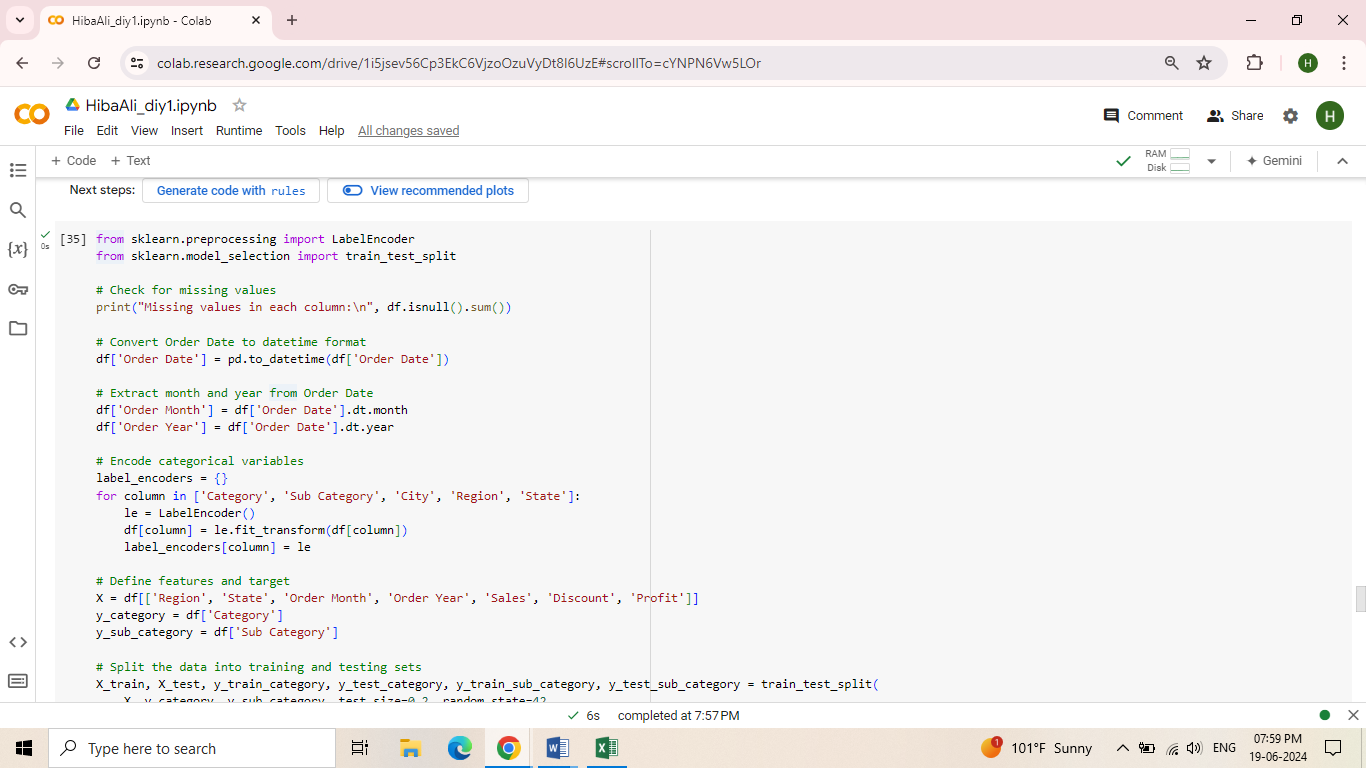


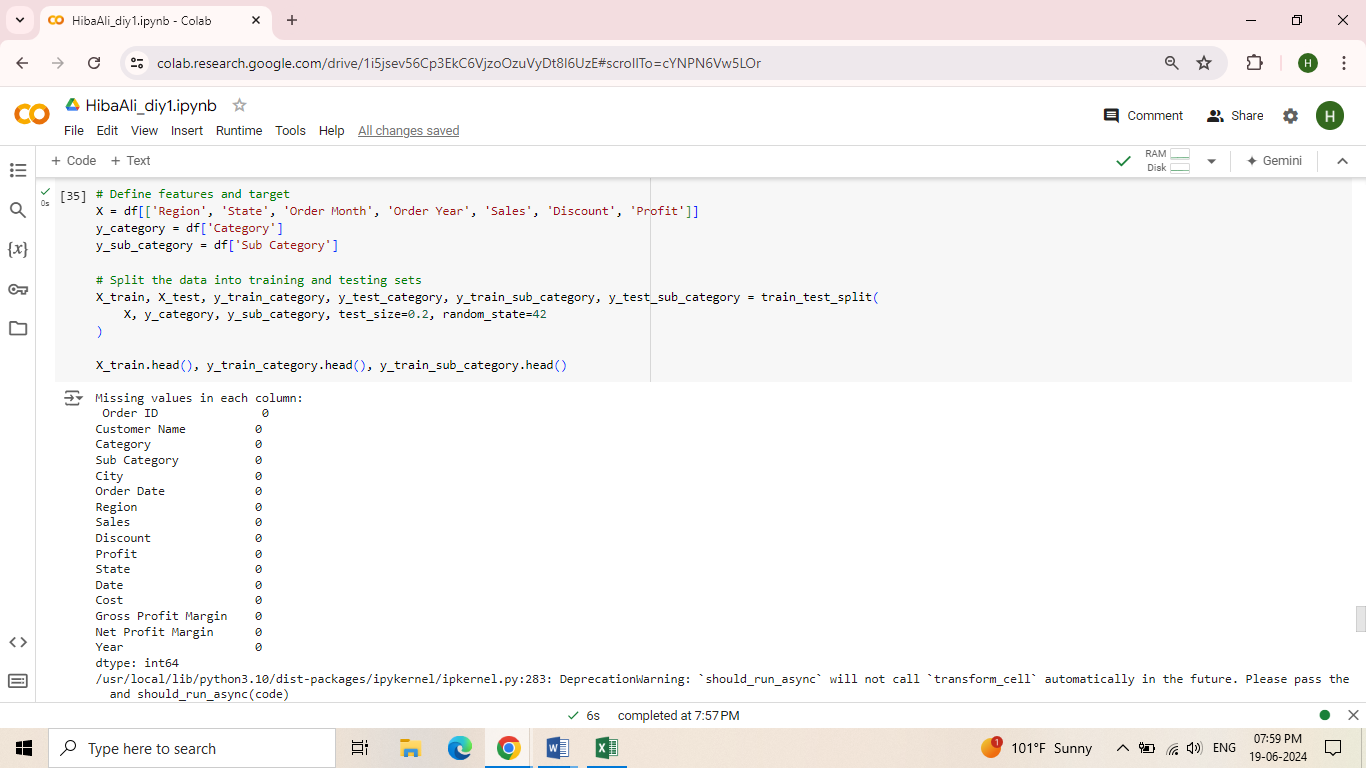
1. Using Classification techniques to develop a model and predict the item categories and sub-categories that would provide the highest sales and profit region wise/state wise:

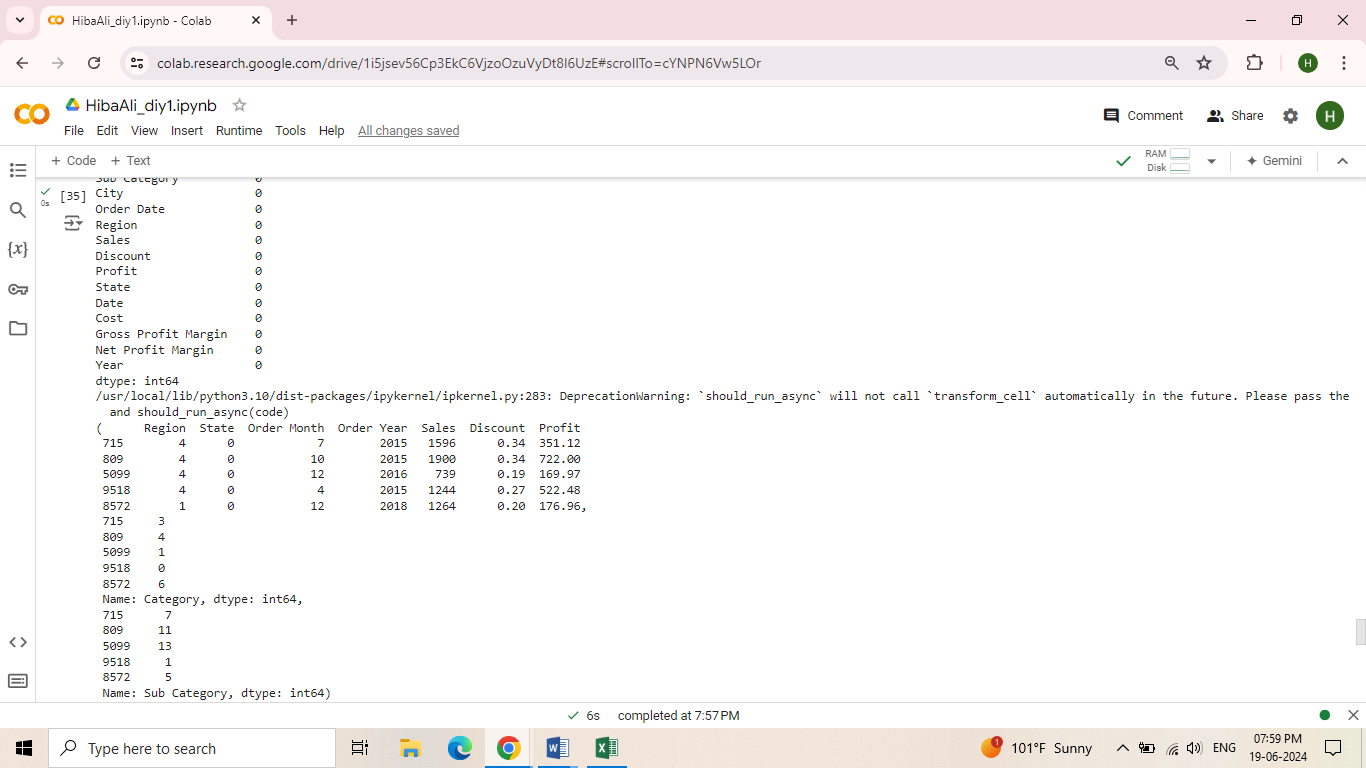
Our goal is to predict the item categories and sub-categories that provide the highest sales and profit region-wise/state-wise.

Pre-processing Steps

* Handle missing values: Check for and handle any missing values.
* Encode categorical variables: Convert categorical columns to numerical values using encoding techniques.
* Feature engineering: Create any additional necessary features (e.g., extracting month and year from Order Date).
* Split the data: Split the dataset into training and testing sets.



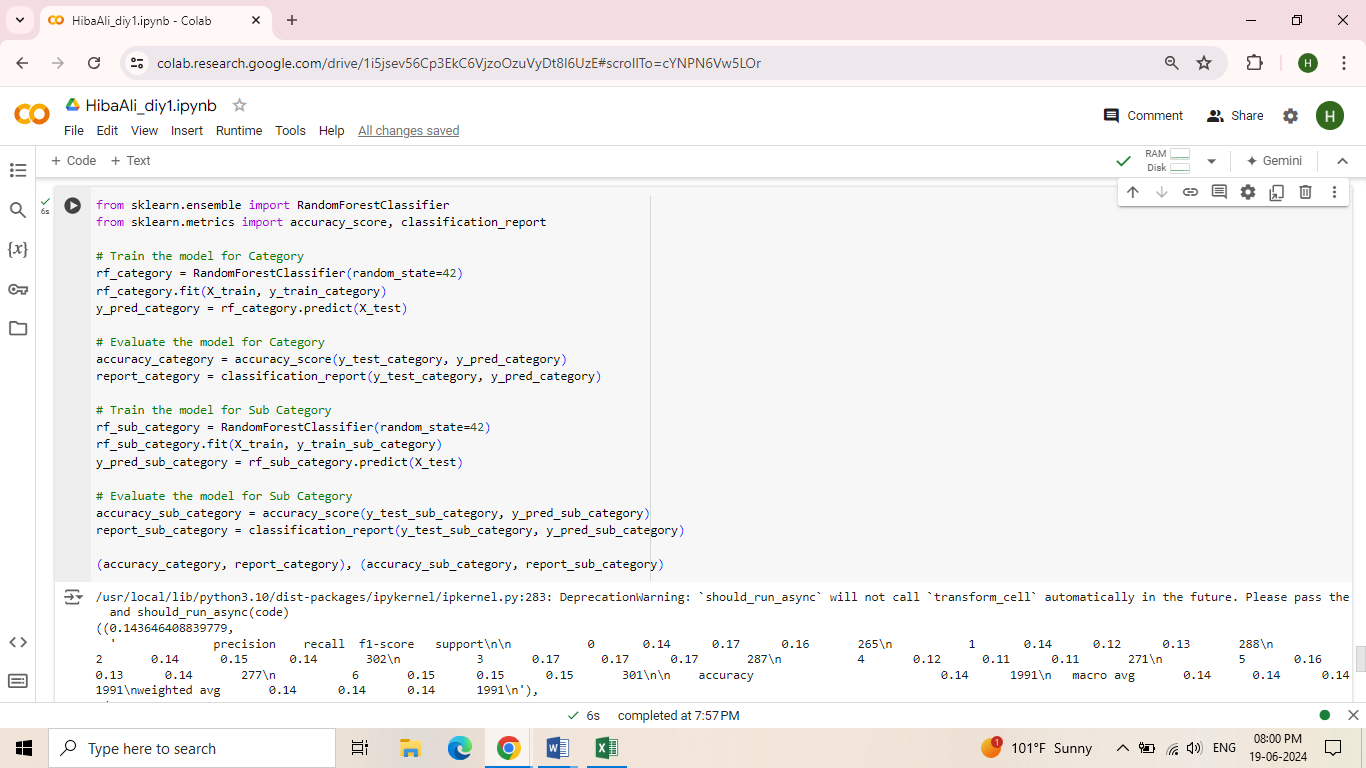


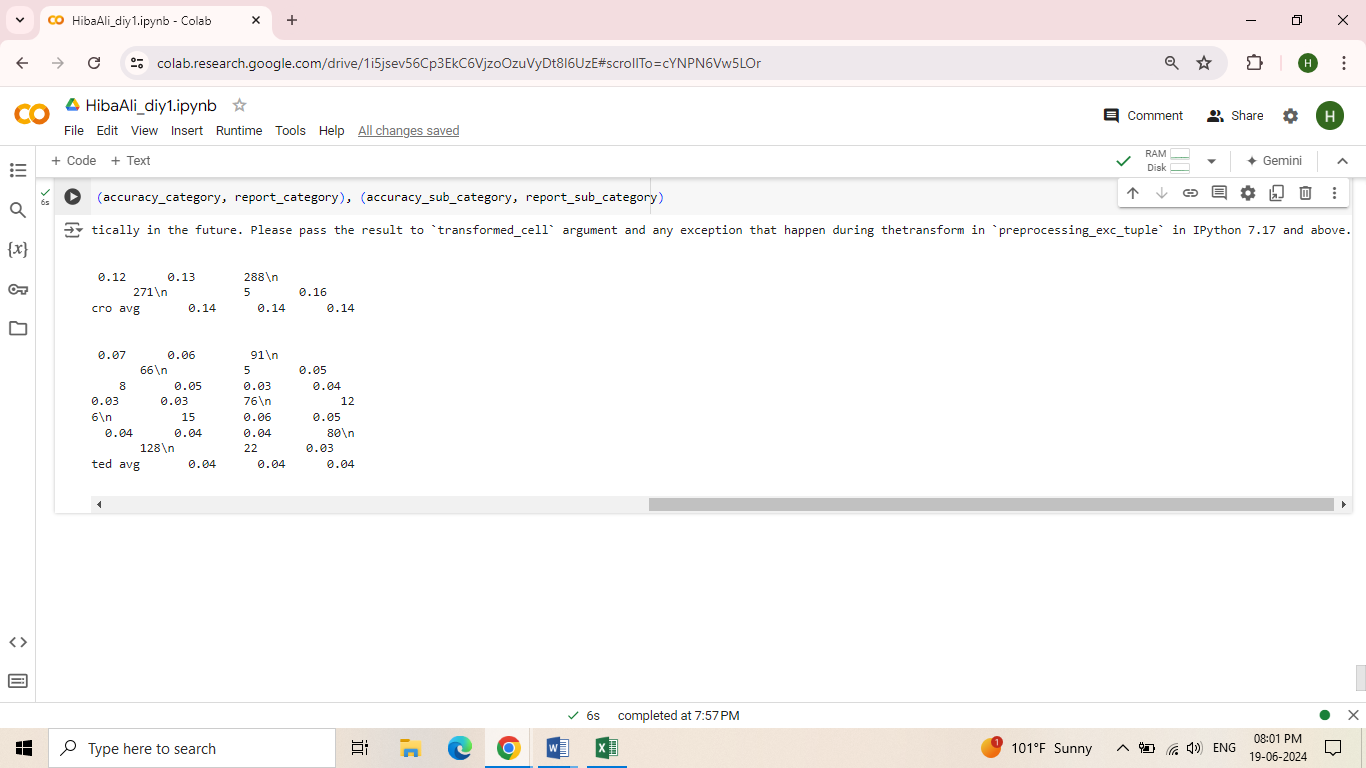


Next, we'll train classification models to predict the Category and Sub Category.

I will use the Random Forest classifier, which generally performs well for such tasks.

I will evaluate the models using accuracy scores.





1. Modifying the dataset to incorporate the Non-Volatile feature of data warehouse, where data cannot be updated or deleted but only can be modified:

To implement the non-volatile feature in our dataset where data cannot be updated or deleted but can only be modified, and all updates should be available for analysis, we can follow these steps:

* Create a Versioning System: Implement a versioning system to keep track of changes.
* Append Changes: Whenever there are updates, append these changes as new rows with version information rather than modifying the existing rows.

