# Project Overview:

This project focuses on analysing the business performance of a company using sales data provided in an Amazon S3 Bucket. Initially, the data was unstructured, containing extra columns, rows, multiple-column headers, extra spaces, spelling mistakes, inconsistencies, and blank values.

The columns in the raw data were transformed as follows:

| **Raw Data Columns** | **Transformed Columns** |
|---|---|
| Business Name | Business Name |
| Customer Name | Contact Person |
| Address | Local Address, City, PIN Code, State |
| Time | Meeting Time |
| Status | Meeting Status |
| Calling Date | Calling Date |
| Meeting Date | Meeting Date |
| Tele Caller Name | Tele Caller Name |
| BDM Name | BDM Name |
| Product Name | Product Proposal |
| Mapping | Map Status |
| Category | Business Category |
| BDM Feedback | BDM Feedback |
| Tele Caller Feedback | Tele Caller Feedback |
| Follow-up Date | Dropped the Column |

We Started by distributing the data to our teams and everyone was assigned different tasks

# Data Cleaning:

The data cleaning step included removing repeating column headers using Python the rows and columns that were blank between the data were removed.

- df=df.dropna(how="all") This formula removes all the columns having null values

**Address:**

Then, by separating the address columns to extract the PIN code, city, and state. Since the addresses were not consistent, some had PIN codes at the beginning and some at the end, and the same issue occurred with other elements of the address, then using the Excel Data tab's Text to Columns function. The address was separated using a comma as a delimiter. Then, the PIN code, city, and state were separated, and using the Concatenate function, the address was reassembled.

**Meeting Time:**

The time data was inconsistent; some rows had data such as "8 to 9 PM," "8.00pm," and "9:00 AM." The time was made consistent by replacing "." with ":" and structuring it properly in the format HH: MM PM. Blank values were filled using the mode.

**Meeting Status:**

Inconsistent data was segregated to confirm, call, and go

**Calling Date and Meeting Date:**

The date column consisted of values such as 9/8/2023, 9.8.2023 which was made consistent to 09-08-2023 also there was a lot of difference between the calling and meeting date like calling date was 05-02-2009 and the meeting date was 06-02-2023 which was made consistent by subtracting 7 days or adding 7 days in the not consistent value.

**Telecaller Name and BDM Name:**

The names were made consistent by replacing any spelling mistake, for example making Deepandra sir to Deepandra, and so on.

**Product Proposal:**

The data was made consistent as there were many variations in product proposals spelling mistakes, and extra spaces, for example, the entries were GMVT+Social Media, also Social Media+GMVT those were made consistent into GDP+Social Media and so on.

**Map Status:**

It consisted of entries such as Done, Not Done, JD Paid, Unverified, JD Paid map done these were changed to Done, Not Done, and Unverified by removing unwanted elements.

**BDM and Tele caller feedback:**

Were kept as it is since nothing much can be done to these columns.

**Final Data check:**

Checked for proper data types are assigned, no spelling mistakes, no blank values, no inconsistent data.

# Cleaning Sales data:

The sales Data file contains the company's Sales, Profit, and Expenses information, this file was dropped into the Amazon S3 bucket the file was then downloaded and the data cleaning process was started as per the observation this file had columns:

| **Initial Columns** | **Changed Into** |
|---|---|
| Business Name | Business Name |
| Keywords | Business Categories |
| Client Name | Contact Person |
| Address | Local Address, State, City, Pincode |
| Product | Product Proposal |
| Tele Caller | Telecaller Name |
| BDM Name | BDM Name |
| Login Date | Login Date |
| Sales Amount | Sales Amount |
| GST Amount | GST Amount |
| Advanced Amount | Advanced Amount |
| Payment Mode | Payment Mode |

**Business Category:**

The business category column was assigned by extracting the category from the keyword's column

**Local Address:**

From the local address PIN code, city, and state are extracted using text to column function of Excel blank state, city, and Pincode is filled by searching businesses online blank pin code are filled by using the mode value for a particular city

**Product:**

Blank values in the product column are filled using No Data

**Tele Caller and BDM Name:**

Spelling mistakes in Tele Caller and BDM names were corrected, and empty values are filled with telecallers from a specific city.

**Login Date:**

The format of the date is corrected

**Sales Amount and Advanced Amount:**

These were kept as it is, empty values were filled with 0

**Payment Mode:** The payment mode column is made standardized by including only Check, online, and cash, and no data values such as IMPS, RTGS, or Gpay were changed into Online Payment

**GST Number:**

The GST Number column is kept as it is, blanks are filled with no data

**Final Data check:**

Checked for proper data types are assigned, no spelling mistakes, no blank values, no inconsistent data.

## Assigning Business ID:

After cleaning the data, we assigned a unique business ID to each business to uniquely identify separate businesses, the business IDs were assigned by identifying unique businesses in the data by unique function of Excel then these unique businesses were given a specific Sr No and assigned to it using VLOOKUP function after which unique business ID were created by extracting 1 letter of each column something Like G5S9J1 these unique ID were assigned using Vlookup to the business.

## Additional Columns Added:

**Payment Mode New:** Standardised already existing products column, Adwords, Adward, adv changed to Google Ad, Facebook, FB, fb, Fb likes changed into Facebook Ad.

**Business Segment:** Further segregated the business category column into a wider category by standardizing Yoga, Yoga Classes, Zumba, Zumba Classes changed to a single category such as Health and Fitness likewise Dr Chetan, Dr Madhuri, Physiotherapy, Oncologist were categorized as Clinic.

**Profit:** 30% of Sales Amount

**Expenses:** 70% of Sales Amount

## Column Data Types assigned:

**Varchar:** Business Name, Business Category, Business Segment, Contact Person, Address, Map, Product Proposal, BDM Name, Tele caller Name, GST Number Feedback, Tele caller Feedback, Payment Mode and Status.

**Integer:** Sales Amount, Advanced Amount, GST Amount, Profit, Expenses

**Date and Time:** Meeting Date, Calling Date, Login Date, Time.

# Creating Table:

4 tables were created out of this data

**Table-1 Business Info**

Business_ID -------Primary Key

Business_Name

Contact_Person

Address

PinCode

City

State

GST_Number

**Table-2 Meeting Info**

Business_ID------ Foreign Key

Telecaller

BDM

Calling_Date

Meeting_Date

Meeting_Time

Meeting_Status

**Table-3 Product Info**

Business_ID---------Foreign Key

Business_Category

Map

Product_Proposal

**Table-4 Sales Data**

Business_ID--------Foreign Key

Login_Date

Sales_Amount

Advanced_Amount

GST_Amount

Payment_Mode

All this data was then loaded into SQL creating tables and joining them using the primary key and foreign key. From the table having a primary key extra duplicate columns were removed to follow the constraint of the primary key after loading the tables appropriate relationships between the tables were created and an ER Diagram was created then an initial analysis of the problem statement on SQL was solved and the final analysis was done making a power BI dashboard which is used to analyze different aspects of the business and how do they contribute in the profit.