A PROJECT REPORT

on

"DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

TOOLS AND TECHNIQUES LABORATORY-CS-3096

Under

BACHELOR'S DEGREE IN COMPUTER SCIENCE AND ENGINEERING

BY

SHIV PRASAD ROUL	2105158
MILAN KUMAR SAHOO	2105208
AMOGH	2105240
SOVAN PATTANAIK	2105247
AADITYA CHOWDHURY	2105341

UNDER THE GUIDANCE OF
LIPIKA DEWANGAN

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April 2024

KIIT Deemed to be University

School of Computer Engineering

Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS “

submitted by

SHIV PRASAD ROUL	2105158
MILAN KUMAR SAHOO	2105208
AMOGH	2105240
SOVAN PATTANAIK	2105247
AADITYA CHOWDHURY	2105341

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of completion of Tools and Techniques Laboratory under Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024, under our guidance.

Date: 06/ 04/ 2024

Lipika Dewangan

Project Guide

Acknowledgements

We extend our sincere gratitude to Lipika Dewangan Ma’am from Computer Science and Engineering for her invaluable guidance and unwavering encouragement, which have been instrumental in steering this project from its inception to its successful completion.

A special word of appreciation goes to all the participants whose active involvement and provision of valuable data and insights have been pivotal in shaping this study. Their cooperation and willingness have significantly contributed to the project's success.

We also extend our gratitude to our colleagues and friends for their continuous support and motivation throughout the duration of this endeavour. Their constructive feed-back and suggestions have played a crucial role in enhancing the quality of this report.

Lastly, we express our heartfelt thanks to our families for their unwavering support, encouragement, and understanding. Their love and encouragement have served as a constant source of inspiration, and we deeply appreciate their presence in our lives. We sincerely thank everyone for their support and guidance, as this project would not have been possible without their collective efforts.

SHIV PRASAD ROUL
MILAN KUMAR SAHOO
AMOGH
SOVAN PATTANAIK
AADITYA CHOWDHURY

ABSTRACT

This data science project delves into the realm of sales analysis and forecasting within the context of small-scale supermarkets, focusing on a dataset from the "Ponnu Super Bazzar" chain of Grocery Supermarts in Tamil Nadu, India. Leveraging Python and a suite of libraries including Matplotlib, Plotly.express, pandas, numpy, and sklearn, the project conducts comprehensive exploratory data analysis (EDA) to unveil insights into sales patterns and trends. The dataset encompasses crucial features such as OrderId, Category, Sub Category, City, Order Date, Region, Sales, Discount, Profit, and State, providing a rich foundation for analysis.

EDA encompasses preprocessing steps and employs various visualizations to dissect sales dynamics across categories, subcategories, cities, and regions, focusing on sales, discounts, profits, and their variations. The exploration aims to unearth actionable insights for optimizing inventory management, pricing strategies, and promotional activities.

For sales forecasting, a linear regression model is employed, utilizing date data transformed into numerical format to predict future sales. The model is trained on historical sales data, and its accuracy is assessed by comparing yearly forecast predictions with actual sales, achieving an impressive accuracy rate of 84%.

With a dataset comprising 10,000 rows, this project showcases the potential of data science methodologies in empowering small-scale supermarkets to make data-driven decisions and drive sustainable growth in a competitive market landscape.

Keywords: Data Science, Sales Analysis, Sales Forecasting, Small-scale Supermarkets, EDA, Python, Matplotlib, Plotly.express, Linear Regression, Exploratory Data Analysis

Contents

Chapter 1

Introduction

In the rapidly evolving landscape of retail, small-scale supermarkets play a crucial role in catering to the diverse needs of local communities. With the advent of data science and its applications in various industries, there arises a pressing need for small-scale supermarkets to leverage data-driven insights to optimize their operations and stay competitive in the market. The project at hand focuses on addressing this need by conducting a comprehensive analysis of sales data and developing accurate forecasting models for "Ponnu Super Bazzar," a chain of Grocery Supermarts in Tamil Nadu.

Fig1.1: Ponnu Super Bazzar at Avadi in Chennai

The importance of this project lies in its potential to empower small-scale supermarkets with actionable insights derived from data analysis and forecasting. By understanding sales patterns, consumer behaviour, and market trends, these supermarkets can make informed decisions regarding inventory management, pricing strategies, and promotional activities. Such data-driven decision-making processes are essential for enhancing operational efficiency, maximizing profits, and ensuring sustainable growth in a dynamic and competitive retail environment.

Despite the availability of various solutions for sales analysis and forecasting, there exist significant gaps in addressing the specific needs and challenges faced by small-scale supermarkets. Existing solutions may not be tailored to the unique characteristics of these supermarkets, such as limited resources, diverse product offerings, and localized customer preferences. Moreover, the lack of accessible and affordable data science tools and expertise further exacerbates the challenges faced by small-scale supermarkets in harnessing the power of data analytics.

This project aims to bridge these gaps by developing customized solutions tailored to the needs of small-scale supermarkets like "Ponnu Super Bazzar." By leveraging advanced data science techniques and tools, we seek to provide practical and scalable solutions that enable these supermarkets to unlock the full potential of their sales data. Through a combination of exploratory data analysis, predictive modelling, and performance evaluation, we aim to deliver actionable insights and accurate sales forecasts that drive informed decision-making and strategic planning.

The structure of this report encompasses a comprehensive overview of the project, including the methodology adopted, data collection and preprocessing steps, detailed analysis of sales data, development and evaluation of forecasting models, and discussion of results and implications. By following this structure, readers will gain a holistic understanding of the project's objectives, methodologies, findings, and contributions to the field of retail analytics and data science.

Chapter 2

Basic Concepts/ Literature Review

2.1 Python Programming Language

Python is a versatile and widely-used programming language renowned for its simplicity, readability, and extensive libraries, making it a preferred choice for data science and machine learning tasks. Its rich ecosystem includes libraries such as NumPy, pandas, Matplotlib, and scikit-learn, which facilitate data manipulation, analysis, visualization, and modelling. Python's intuitive syntax and interactive nature enable efficient prototyping and development of data-driven applications.

2.2 NumPy

NumPy (Numerical Python) is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy serves as the backbone for many other libraries in the Python data science ecosystem, enabling fast and efficient data manipulation and computation.

2.3 pandas

pandas is a powerful data manipulation and analysis library built on top of NumPy, offering data structures like Data Frame and Series that facilitate handling structured data effectively. With its intuitive interface and rich functionality, pandas simplifies tasks such as data cleaning, transformation, aggregation, and exploration, making it indispensable for data preprocessing and exploratory data analysis (EDA).

2.4 Matplotlib

Matplotlib is a versatile plotting library for creating static, interactive, and publication-quality visualizations in Python. It provides a MATLAB-like interface for generating a wide range of plots, including line plots, scatter plots, histograms, bar charts, and more. Matplotlib's flexibility and customization options make it suitable for exploring and presenting data insights effectively.

2.5 scikit-learn

scikit-learn is a comprehensive machine learning library that offers various algorithms for classification, regression, clustering, dimensionality reduction, and more. It provides a consistent interface for model training, evaluation, and deployment, along with utilities for data preprocessing, feature selection, and model tuning. With its user-friendly API and extensive documentation, scikit-learn facilitates the implementation of machine learning workflows for predictive modelling tasks.

2.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process that involves examining and visualizing data to understand its underlying patterns, distributions, and relationships. EDA techniques, such as summary statistics, data visualization, and correlation analysis, help uncover insights and identify trends in the data, guiding subsequent analysis and modelling tasks.

2.7 Sales Forecasting

Sales forecasting is the process of predicting future sales based on historical data and other relevant factors. It plays a vital role in strategic planning, inventory management, and resource allocation for businesses. Various techniques, including time series analysis, regression modelling, and machine learning algorithms, can be employed to develop accurate sales forecasting models tailored to specific business needs.

2.8 Data Preprocessing

Data preprocessing involves cleaning, transforming, and preparing raw data for analysis and modelling. It includes tasks such as handling missing values, encoding categorical variables, scaling numerical features, and removing outliers. Proper data preprocessing is essential for ensuring the quality and integrity of the data, as well as improving the performance of predictive models.

2.9 Feature Engineering

Feature engineering is the process of creating new features or transforming existing features to improve the performance of machine learning models. It involves selecting relevant features, creating derived features, and encoding categorical variables effectively. Feature engineering plays a critical role in capturing meaningful patterns and relationships in the data, enhancing the predictive power of models.

2.10 Model Evaluation

Model evaluation is the process of assessing the performance of machine learning models using various metrics and techniques. It involves splitting the data into training and testing sets, fitting the model to the training data, and evaluating its performance on unseen test data. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

2.11 Linear Regression

Linear regression is a widely-used statistical technique for modelling the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the predictors and the target variable and seeks to find the best-fitting line (or hyperplane in the case of multiple predictors) that minimizes the residual sum of squares. The resulting model can be represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where:

- y is the dependent variable (target)
- β_0 is the intercept term
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients for the independent variables $x_1, x_2, ..., x_n$
- ϵ is the error term

Linear regression is commonly used for predicting continuous outcomes and is suitable for tasks such as sales forecasting, price prediction, and trend analysis. It is a simple yet powerful technique that provides interpretable results and can serve as a baseline model for more complex machine learning algorithms.

2.12 Model Evaluation Metrics

Model evaluation metrics are used to assess the performance of machine learning models and quantify their predictive accuracy. There are various metrics available for different types of predictive tasks, such as classification, regression, and clustering. Some commonly used model evaluation metrics include:

- Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted values and the actual values. It provides a straightforward measure of prediction accuracy and is less sensitive to outliers compared to other metrics.
- Mean Squared Error (MSE): MSE calculates the average squared difference between the predicted values and the actual values. It penalizes larger errors more heavily than MAE and is commonly used in regression tasks.
- Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides an interpretable measure of prediction error in the same units as the target variable. It is widely used for evaluating regression models and provides a more intuitive understanding of prediction accuracy.
- R-squared (R2) Score: R2 score measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating better model fit. R2 score is commonly used as a measure of goodness-of-fit for regression models.
- Mean Absolute Percentage Error (MAPE): MAPE calculates the average percentage difference between the predicted values and the actual values. It provides a relative measure of prediction accuracy and is commonly used in forecasting tasks to evaluate the performance of time series models.

Cchapter 3

Problem Statement / Requirement Specifications

3.1 Problem Statement

The problem at hand revolves around empowering small-scale supermarkets, such as "Ponnu Super Bazzar," with actionable insights derived from sales analysis and accurate sales forecasting. These supermarkets face challenges in optimizing inventory management, pricing strategies, and promotional activities due to limited resources and the dynamic nature of consumer behaviour and market trends. The lack of accessible and affordable data science solutions tailored to the specific needs of small-scale supermarkets exacerbates these challenges, hindering their ability to make informed decisions and drive sustainable growth.

The project aims to address this problem by leveraging data science techniques to analyse historical sales data, identify patterns and trends, and develop accurate forecasting models. By providing valuable insights into sales dynamics, consumer preferences, and market trends, the project seeks to enable small-scale supermarkets to optimize their operations, enhance customer satisfaction, and increase profitability.

3.2 Project Planning

The project planning document outlines the steps to be followed while planning and executing the development of the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets. It presents a list of requirements and features to be developed to address the problem statement effectively.

1. **Requirement Gathering:**
    - Gather detailed requirements from stakeholders, including supermarket owners, managers, and data analysts.
    - Identify key objectives, challenges, and constraints associated with sales analysis and forecasting for small-scale supermarkets.
2. **Data Collection and Preparation:**

  - Collect historical sales data from "Ponnu Super Bazzar" including OrderId, Category, Sub Category, City, Order Date, Region, Sales, Discount, Profit, and State.
  - Cleanse and preprocess the raw data to handle missing values, outliers, and inconsistencies.
3. **Exploratory Data Analysis (EDA):**
  - Conduct exploratory data analysis to uncover insights and trends in sales data.
  - Visualize sales patterns, correlations, and distributions across different categories, subcategories, cities, and regions.
4. **Feature Engineering:**
  - Engineer relevant features to capture meaningful patterns and relationships in the data.
  - Transform and encode categorical variables, create derived features, and perform dimensionality reduction if necessary.
5. **Model Development:**
  - Develop a sales forecasting model using machine learning algorithms such as linear regression.
  - Train the model on historical sales data and additional relevant features.
6. **Model Evaluation and Validation:**
  - Evaluate the performance of the forecasting model using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.
  - Validate the model using cross-validation techniques to ensure robustness and generalizability.
7. **Results Interpretation and Documentation:**
  - Interpret the results of the analysis and forecasting to derive actionable insights.
  - Document the entire project workflow, including data collection, preprocessing steps, analysis, modelling techniques, results interpretation, and recommendations.
8. **Visualization and Reporting:**
  - Create informative visualizations to present the results of the analysis and forecasting.

- Generate comprehensive reports and presentations for stakeholders, highlighting key findings, recommendations, and implications for decision-making.

9. **Deployment and Integration:**
    - Provide guidelines for deploying the developed forecasting model in a production environment.
    - Ensure seamless integration of the model into the supermarket's operational workflow for ongoing monitoring and decision support.

3.3 System Design

3.3.1 Design Constraints

The Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets operates within certain design constraints that define the working environment and resources utilized:

- **Software Environment:** The project primarily utilizes Python programming language along with various libraries and frameworks such as NumPy, pandas, Matplotlib, Plotly.express, scikit-learn, and others for data manipulation, analysis, visualization, and modelling.
- **Hardware Requirements:** The project can be executed on standard computing hardware commonly available in data science environments, including desktops, laptops, or cloud-based virtual machines. The hardware should have sufficient processing power and memory capacity to handle the computational requirements of data analysis and modelling tasks.
- **Experimental Setup:** The project does not require any specific experimental setup or environmental setup. However, access to historical sales data from "Ponnu Super Bazzar" is essential for conducting the analysis and modelling tasks.

3.3.2 System Architecture

The system architecture for the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets can be represented using a high-level block diagram illustrating the key components and their interactions:

Components:

1. **Data Collection:** Historical sales data from "Ponnu Super Bazzar" is collected and stored in a structured format, including features such as

OrderId, Category, Sub Category, City, Order Date, Region, Sales, Discount, Profit, and State.

2. **Preprocessing:** The raw data undergoes preprocessing steps such as cleaning, handling missing values, encoding categorical variables, and scaling numerical features to prepare it for analysis and modelling.
3. **Exploratory Data Analysis (EDA):** The pre-processed data is subjected to exploratory data analysis to uncover insights, patterns, and trends in sales data across different categories, subcategories, cities, and regions. Visualization techniques are utilized to present the findings effectively.
4. **Feature Engineering:** Relevant features are engineered to capture meaningful patterns and relationships in the data, enhancing the predictive power of the forecasting models. This may include creating derived features, transforming variables, and performing dimensionality reduction techniques.
5. **Model Development:** A sales forecasting model, such as linear regression, is developed using machine learning algorithms trained on historical sales data and additional features. The model aims to predict future sales based on past trends and patterns.
6. **Model Evaluation:** The performance of the forecasting model is evaluated using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score to assess its accuracy and effectiveness.
7. **Results Interpretation:** The results of the analysis and forecasting are interpreted to derive actionable insights and recommendations for stakeholders, facilitating informed decision-making and strategic planning.

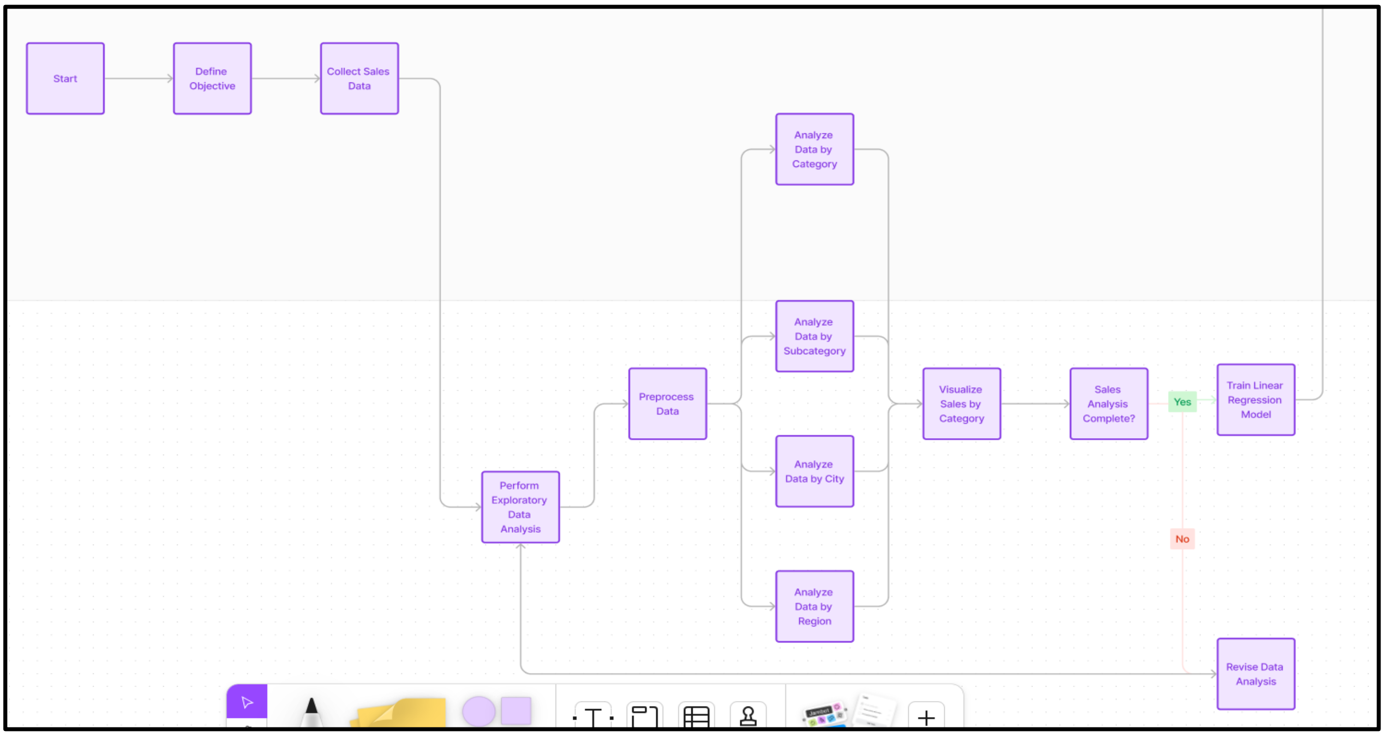


Fig-2.1: Part1 of Architecture of Project

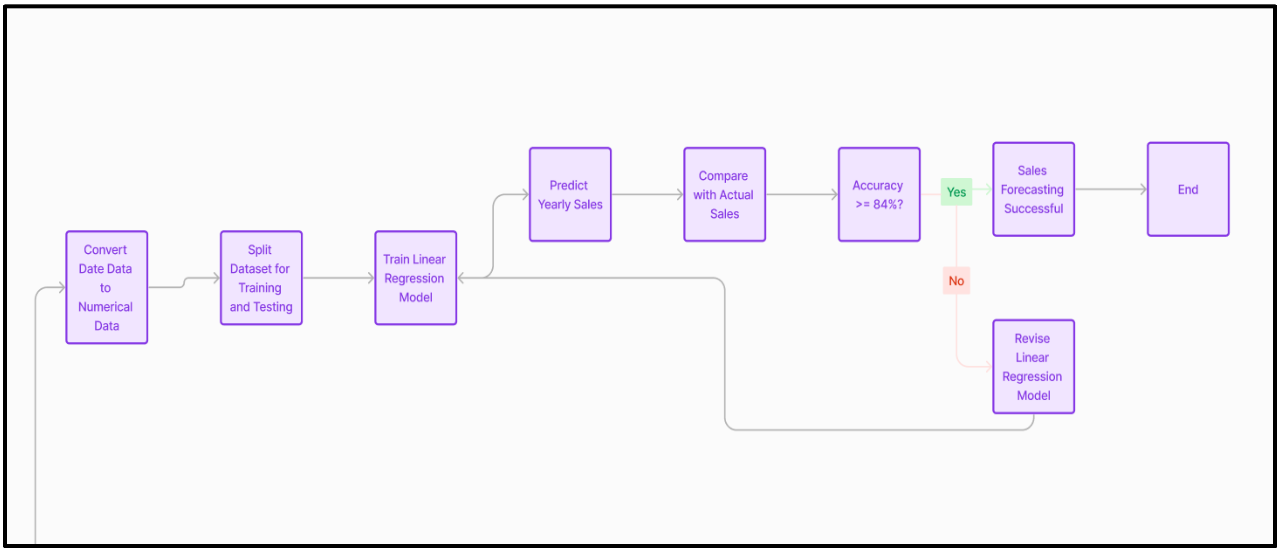


Fig-2.2: Part2 of Architecture of Project

Chapter 4

Implementation

4.1 Methodology OR Proposal

For Visualization and Analysis:

1) Importing Required Libraries

In [1]:
```
# Import pandas for data manipulation and analysis
import pandas as pd

# Import Plotly Express for interactive plotting
import plotly.express as px

# Import NumPy for numerical computing
import numpy as np

# Import matplotlib.pyplot for static plotting
import matplotlib.pyplot as plt

# Import seaborn for statistical data visualization
import seaborn as sns

# Import Plotly Graph Objects for advanced plotting capabilities
import plotly.graph_objects as go

# Set default figure size and style for matplotlib
plt.rcParams['figure.figsize'] = (12, 6)
plt.style.use('fivethirtyeight')

# Suppress warnings to improve code readability
import warnings
warnings.filterwarnings("ignore")
```

2) Loading the Dataset:

In [2]:
```
# Load the dataset
df = pd.read_csv('Supermart Grocery Sales - Retail Analytics Dataset.csv')
df.head().style.set_properties(**{'background-color':'lightblue','color':'black','border-color':'#8b8c8c'})
```

Out[2]:

	Order ID	Customer Name	Category	Sub Category	City	Order Date	Region	Sales	Discount	Profit	State
0	OD1	Harish	Oil & Masala	Masalas	Vellore	11-08-2017	North	1254	0.120000	401.280000	Tamil Nadu
1	OD2	Sudha	Beverages	Health Drinks	Krishnagiri	11-08-2017	South	749	0.180000	149.800000	Tamil Nadu
2	OD3	Hussain	Food Grains	Atta & Flour	Perambalur	06-12-2017	West	2360	0.210000	165.200000	Tamil Nadu
3	OD4	Jackson	Fruits & Veggies	Fresh Vegetables	Dharmapuri	10-11-2016	South	896	0.250000	89.600000	Tamil Nadu
4	OD5	Ridhesh	Food Grains	Organic Staples	Ooty	10-11-2016	South	2355	0.260000	918.450000	Tamil Nadu

We now understand the Dataset. Check the number of rows and columns in the dataset. Check the data types of each column. Check for any missing or null values.

3) Understanding the Dataset:

In [3]:
```
# Check the number of rows and columns in the dataset
print (df.shape)
print('Number of rows:', df.shape[0])
print('Number of columns:', df.shape[1])
```

```
(9994, 11)
Number of rows: 9994
Number of columns: 11
```

In [4]:
```
# Check the data types of each column
df.dtypes
#This is necessary as it will give us idea of categorical and numerical values in the dataset
```

Out[4]:
```
Order ID         object
Customer Name    object
Category         object
Sub Category     object
City             object
Order Date       object
Region           object
Sales             int64
Discount        float64
Profit          float64
State            object
dtype: object
```

4) Data cleaning and Preprocessing: Done by using Pandas

In [5]:
```
# Check for any missing or null values
print(df.isnull().sum())
```

```
Order ID         0
Customer Name    0
Category         0
Sub Category     0
City             0
Order Date       0
Region           0
Sales            0
Discount         0
Profit           0
State            0
dtype: int64
```

In [7]:
```
# Remove unnecessary columns
df = df.drop(columns=['Order ID', 'State'])
#removing orderId and State as they are not required

# Rename columns
df = df.rename(columns={'Sub Category': 'Sub_Category', 'Order Date': 'Order_Date'})
```

In [8]:
```
# Convert date column to datetime format
df['Order_Date'] = pd.to_datetime(df['Order_Date'], errors='coerce')
print(df.info())

# Check for any remaining null values in the date column
print(df[df['Order_Date'].isnull()])

"""If any null values are found, check the original data to identify the correct date
format and update the format parameter in the to_datetime function accordingly."""
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Customer Name  9994 non-null   object
 1   Category       9994 non-null   object
 2   Sub_Category   9994 non-null   object
 3   City           9994 non-null   object
 4   Order_Date     9994 non-null   datetime64[ns]
 5   Region         9994 non-null   object
 6   Sales          9994 non-null   int64
 7   Discount       9994 non-null   float64
 8   Profit         9994 non-null   float64
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 702.8+ KB
None
Empty DataFrame
Columns: [Customer Name, Category, Sub_Category, City, Order_Date, Region, Sales, Discount, Profit]
Index: []
```

Out[8]:
```
'If any null values are found, check the original data to identify the correct date \nformat and update the format parameter in the to_datetime function accordingly.'
```

5) Exploratory Data Analysis: Done by Pandas and Matplotlib for visualization

Step 5: Exploratory data analysis :

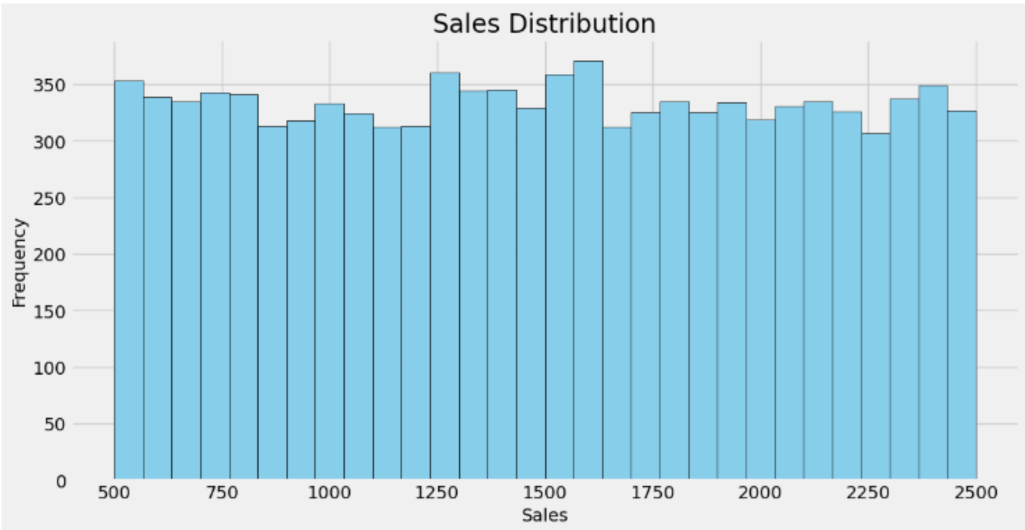
We can calculate basic statistics such as mean, median, and mode for the numerical columns, and create visualizations such as histograms, scatterplots, and boxplots to understand the distribution of the data.

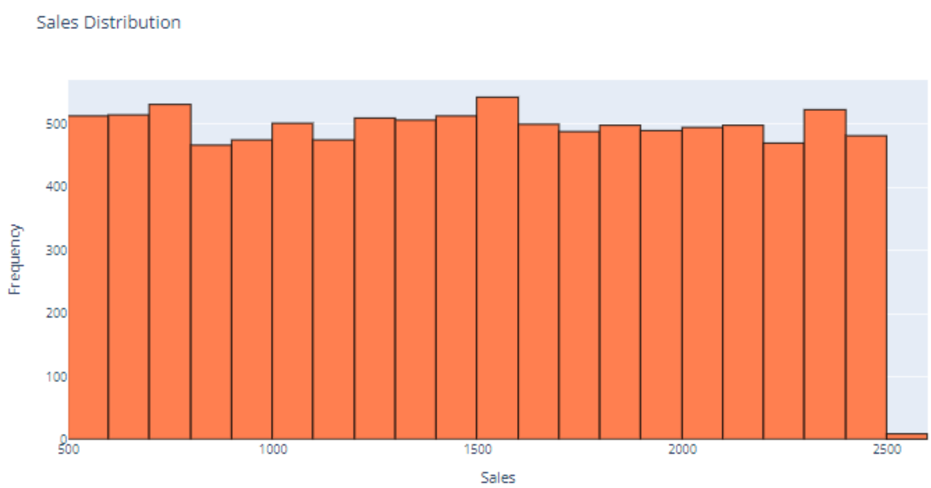
In [13]:
```
# Calculate basic statistics
df.describe()
```

Out[13]:

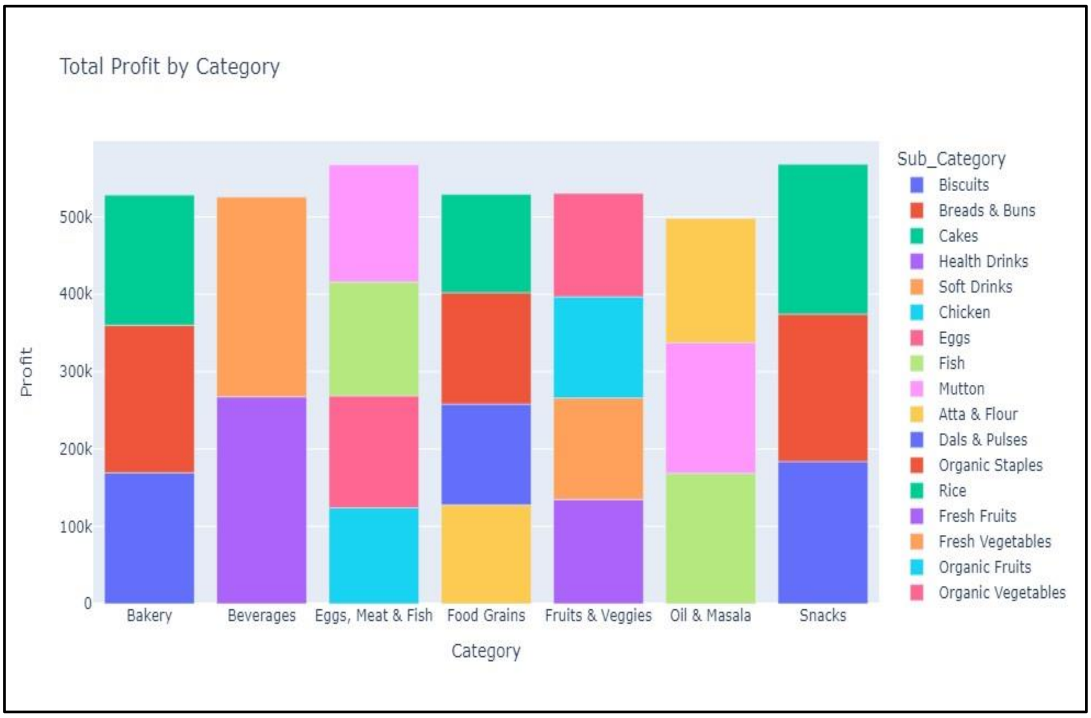
	Sales	Discount	Profit
count	9994.000000	9994.000000	9994.000000
mean	1496.596158	0.226817	374.937082
std	577.559036	0.074636	239.932881
min	500.000000	0.100000	25.250000
25%	1000.000000	0.160000	180.022500
50%	1498.000000	0.230000	320.780000
75%	1994.750000	0.290000	525.627500
max	2500.000000	0.350000	1120.950000

6) Visualization through graphs: Used Plotly.JS for interactive plotting





7) Categorical Analysis and Visualization:



For Sales Forecasting through Regression Modelling:

The initial dataset steps are exactly similar to visualization part. In regression modelling after the data preprocessing step, we follow the following process for the give Date-Sales Data:

1) Converting the Object based Date Data to Timestamp Data

```
In [109]: df1 = df[['Order_Date','Sales']]
          df1.head(30)
          df1.info()
          df1['Order_Date'] = pd.to_datetime(df1['Order_Date'], errors='coerce')
          df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Order_Date  9994 non-null   datetime64[ns]
 1   Sales       9994 non-null   int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 156.3 KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Order_Date  9994 non-null   datetime64[ns]
 1   Sales       9994 non-null   int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 156.3 KB
```

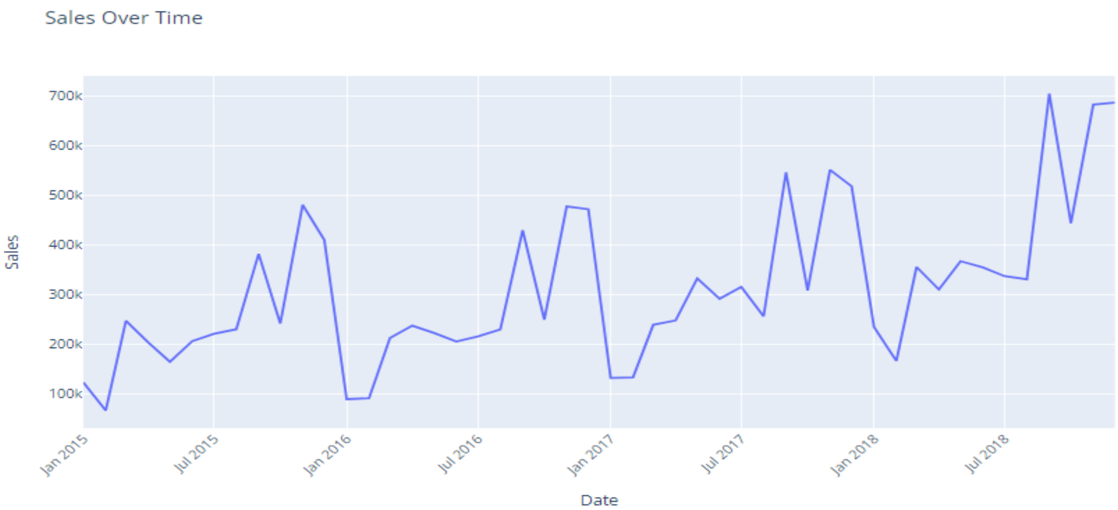
```
C:\Users\bapal\AppData\Local\Temp\ipykernel_165804\969778650.py:4: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
```

```
In [110]: df1['Order_Date']=df1['Order_Date'].dt.to_period("M")
          monthly_sales=df1.groupby('Order_Date').sum().reset_index()
          monthly_sales['Order_Date']=monthly_sales['Order_Date'].dt.to_timestamp()
          monthly_sales.head(20)
```

2) Initial Visualization before modelling



3) Model Training: We convert the timestamp into monthly series data though Pandas and divide the entire dataset to training and test set.

Splitting the Data

```
train_data=sup_data[:-12]
test_data=sup_data[-12:]
print(train_data.shape)
print(test_data.shape)
```

```
(23, 13)
(12, 13)
```

We are using MinMax Scaler to normalize the values

```
scaler = MinMaxScaler(feature_range=(-1,1))
scaler.fit(train_data)
train_data=scaler.transform(train_data)
test_data=scaler.transform(test_data)
```

```
x_train, y_train = train_data[:,1:],train_data[:,0:1]
x_test, y_test = test_data[:,1:],test_data[:,0:1]
y_train=y_train.ravel()
y_test=y_test.ravel()
print("X-Train Shape:",x_train.shape)
print("y-Train Shape:",y_train.shape)
print("x-test Shape:",x_test.shape)
print("y-test Shape:",y_test.shape)
```

```
X-Train Shape: (23, 12)
y-Train Shape: (23,)
x-test Shape: (12, 12)
y-test Shape: (12,)
```

4) Using the LinearRegression() function to train the model using SKLearn Library(Sci-Kit)

Making a prediction dataframe

```
sales_date = monthly_sales['Order_Date'][-12:].reset_index(drop=True)
predict_df = pd.DataFrame(sales_date)
```

```
act_sales = monthly_sales ['Sales'][-13:].to_list()
#getting the values of last 13 months because it will be used for comparison
print(act_sales)
```

```
[518307, 234739, 166267, 355704, 310150, 367411, 354902, 337092, 331014, 705680, 443898, 683410, 687245]
```

LINEAR REGRESSION MODEL

To create the linear regression model and also making the linear regression preiction and draw comparison

```
lr_model= LinearRegression()
lr_model.fit(x_train,y_train)
lr_prediction = lr_model.predict(x_test)
```

```
lr_prediction = lr_prediction.reshape(-1,1)
lr_pre_test_set = np.concatenate([lr_prediction,x_test],axis=1)
lr_pre_test_set = scaler.inverse_transform(lr_pre_test_set)
```

5) Checking the final results and verifying the metrics

```
print(predict_df)
```

```
   Order_Date  Linear Prediction
0  2018-01-01      179632.816847
1  2018-02-01      222223.436515
2  2018-03-01      259166.792644
3  2018-04-01      291733.658496
4  2018-05-01      455780.301320
5  2018-06-01      297273.866942
6  2018-07-01      387048.405916
7  2018-08-01      238708.454167
8  2018-09-01      668723.677737
9  2018-10-01      366399.543146
10 2018-11-01      660251.511686
11 2018-12-01      594862.863305
```

```
lr_mse= np.sqrt(mean_squared_error(predict_df['Linear Prediction'],monthly_sales['Sales'][-12:]))
lr_mae= mean_absolute_error(predict_df['Linear Prediction'],monthly_sales['Sales'][-12:])
lr_r2= r2_score(predict_df['Linear Prediction'],monthly_sales['Sales'][-12:])

print(" Linear Regression MSE: ",lr_mse)
print(" Linear Regression MAE: ",lr_mae)
print(" Linear Regression R2: ",lr_r2)
```

```
 Linear Regression MSE:  67339.34894633491
 Linear Regression MAE:  62022.5798985284
 Linear Regression R2:  0.8340833853830065
```

4.2 Verification Plan:

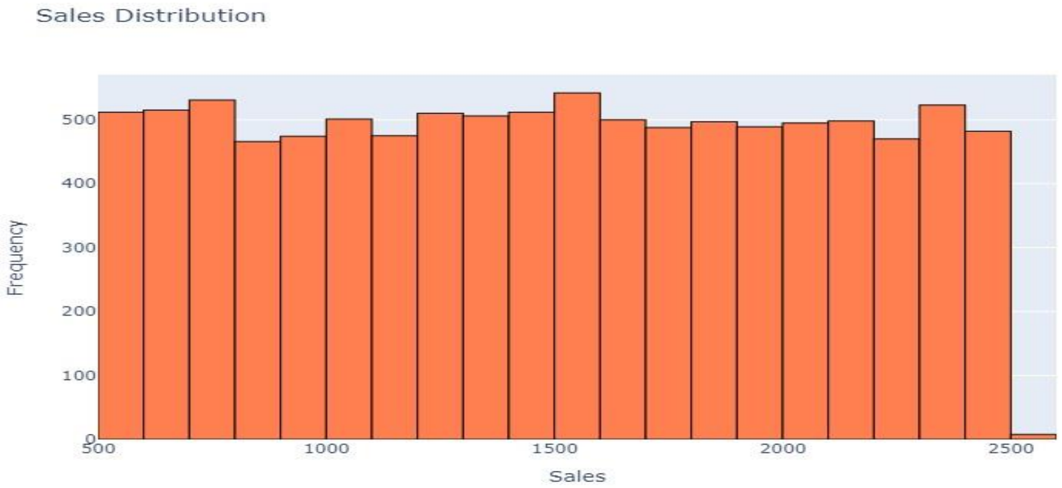
The verification plan ensures the accuracy, completeness, and quality of the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets. It includes the following key components:

1. **Data Verification:** Ensuring the accuracy and integrity of collected sales data.
2. **Preprocessing Verification:** Validating the effectiveness of data preprocessing techniques.
3. **EDA Verification:** Confirming the reliability of insights derived from exploratory data analysis.
4. **Feature Engineering Verification:** Validating the impact of feature engineering techniques on model performance.
5. **Model Development Verification:** Ensuring the accuracy and generalizability of the sales forecasting model.
6. **Results Interpretation Verification:** Confirming the validity and reliability of insights derived from model results.

7. **Documentation Verification:** Ensuring the accuracy and clarity of project documentation.

4.3: Results and Outputs:

Visualization and Analysis:

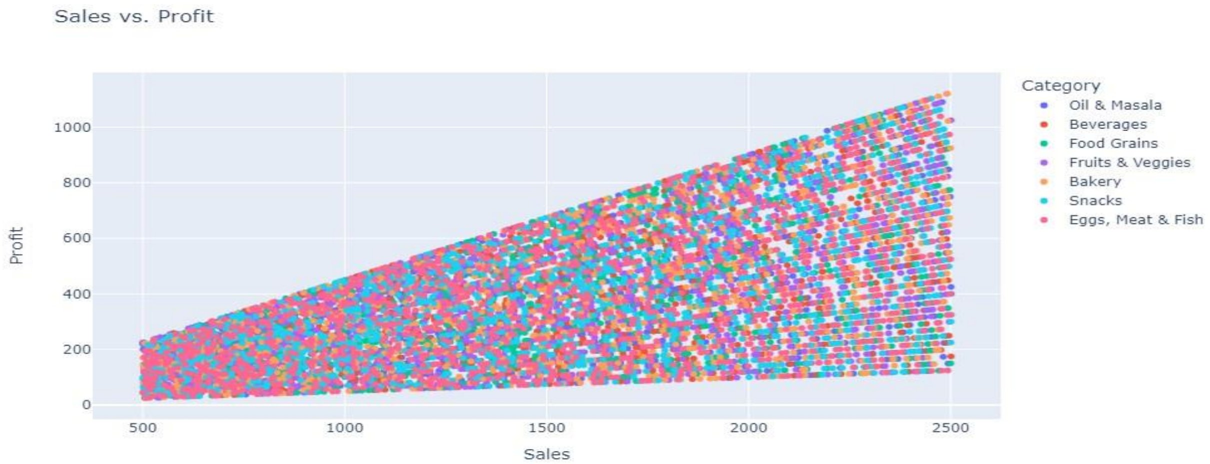


Spread: The sales data spans a wide range, from approximately 500 to 2500.This indicates significant variability in sales figures.

Extreme Points:

Lowest Frequency: The lowest frequency occurs at sales values around 466and sales between 800 and 899.

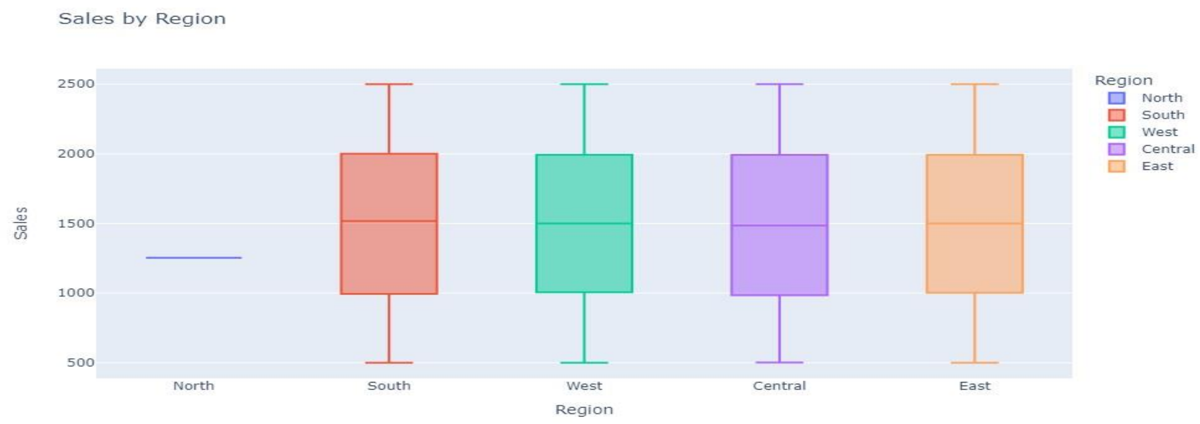
Highest Frequency: The highest frequency is observed at approximately 542 and again at around 1500-1600.



SCATTER PLOT OF SALES VS. PROFIT

Spread: The data points are spread across the following ranges:
Sales: From approximately 0 to around 2500.
Profit: From 0 to approximately 1000.
The dense concentration of data points in the lower sales and profit region suggests that many transactions result in lower sales and profits.

Category	Sales	Profit
Oil & Masala	2478	1100
Beverages	2413	1086
Food grains	2471	1086
Fruits and vegetables	2465	1085
Bakery	2491	1121
Snacks	2450	1102
Eggs, Meat and Fish	2265	1102



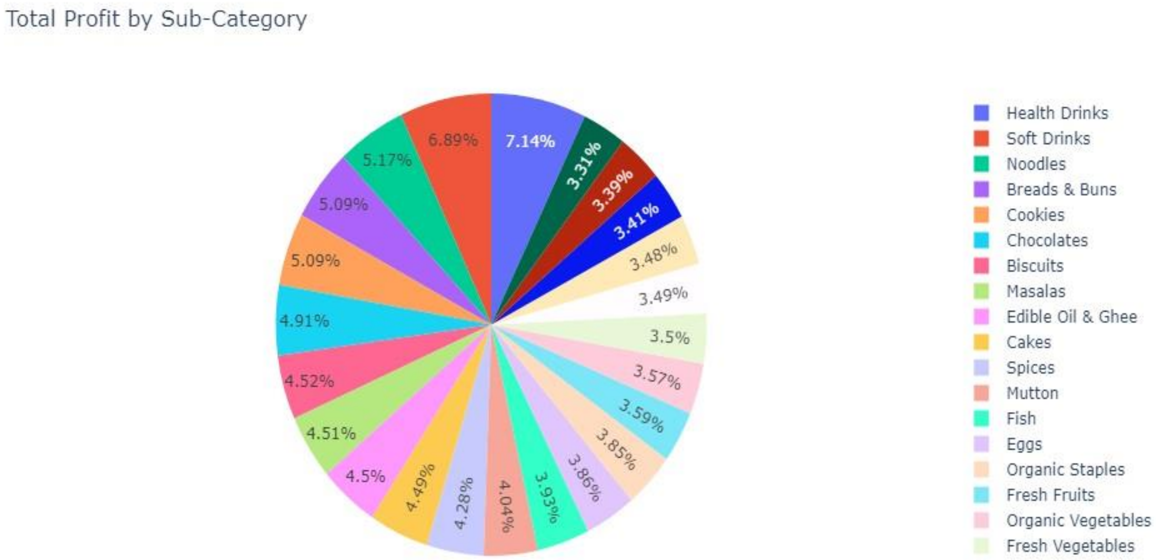
Box plot of sales by region

Region/Sales	Max	Medium	Min
South	2500	1571	500
West	2500	1500	500
Central	2500	1486	500
East	2500	1500	500

From the above table South Region have Maximum Sales and Central Region has Minimum Sales and North region have no sales.

Analyze sales and profit by Category

Data

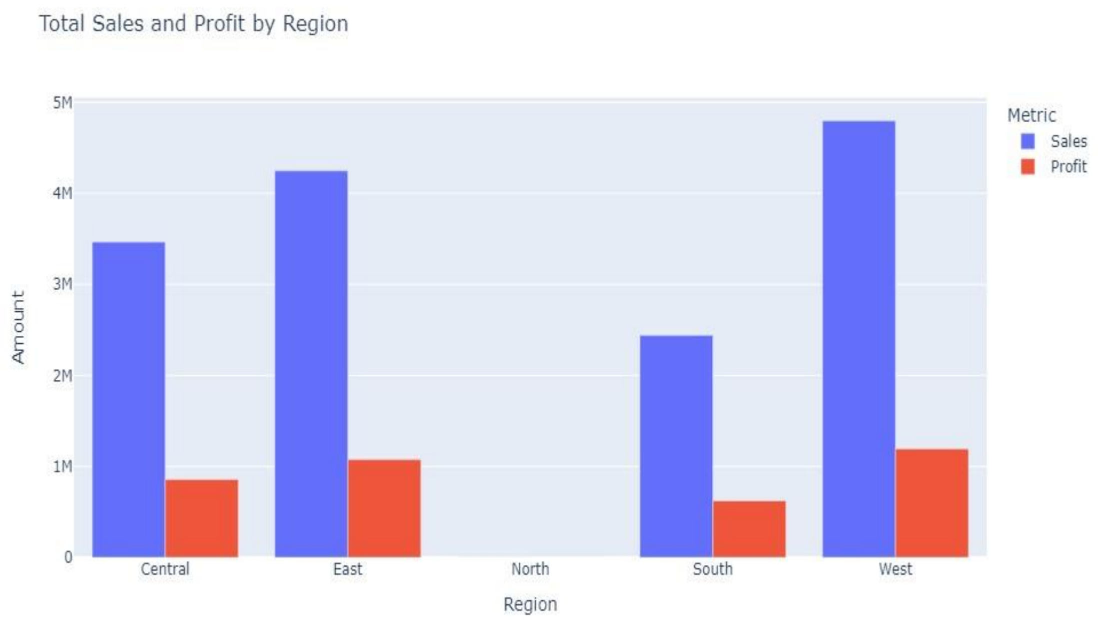


Description:
The pie chart represents the distribution of total profits across various sub-categories. Each segment corresponds to a specific sub-category, and the colours differentiate them.
Spread: -
The spread refers to how profits are distributed among the sub-categories.
We have a total of 23 sub-categories represented in the chart.
Highest Profit: -
The sub-category with the highest profit is Health Drinks, contributing 7.36% to the total profits.

Lowest Profit: -
The sub-category with the Lowest profit is Rice, contributing 3.49% to the total profits.
Data Inference:
Health Drinks: Leading in profits, suggesting they might be popular or high-margin items.
Organic Fruits & Rice: Contributing the least to profits; strategies to enhance their sales or margins could be beneficial.
Other Significant Contributors:
Soft Drinks, Noodles, and Breads & Buns also have substantial contributions.

Analysing sales patterns for these categories could provide insights for boosting overall profitability

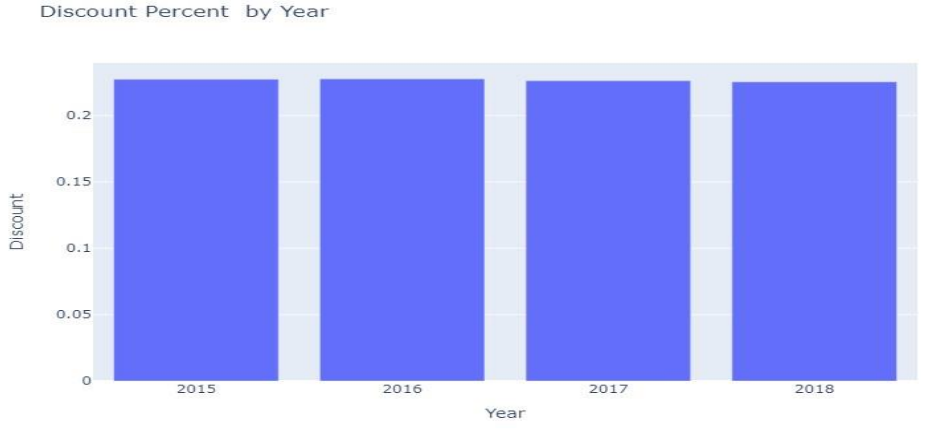
Analyze sales and profit by region



Region	Profit	Sales	Spreads: Sales in x times Profit	Inference
Central	857K	3.46M	4	high sales but a relatively lower profit margin.
East	1.07M	4.24M	4	performs well in both sales and profit.
South	624K	2,44m	4	needs improvement in profitability.
West	1.192M	4.8M	4	the most successful in terms of both sales and profitability.

From the above diagram West region have highest Profit over Sales and North region have no Sales and Profit.

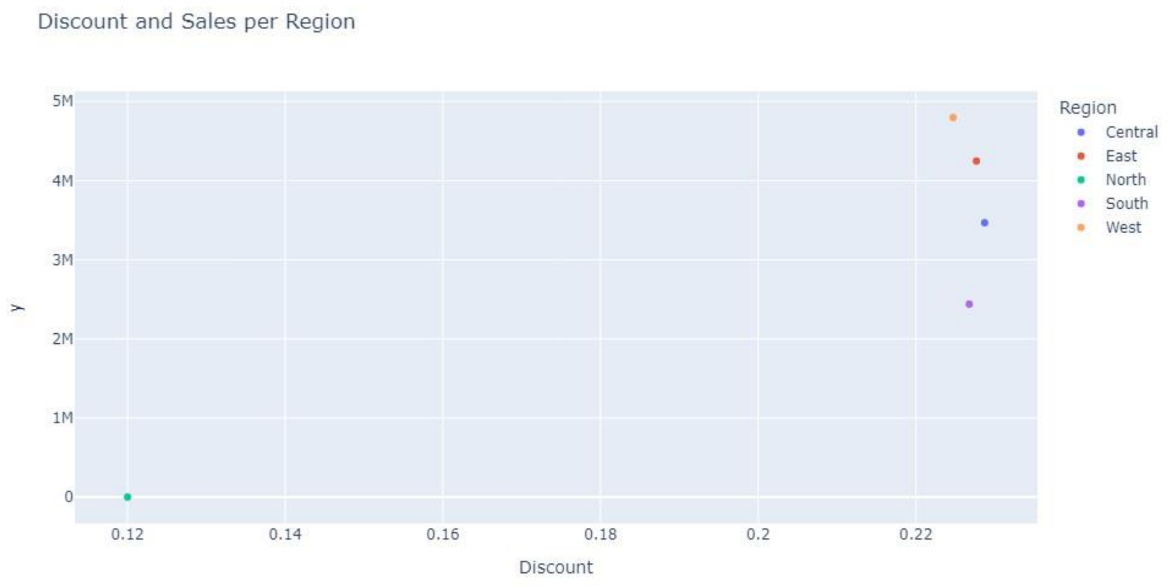
Analyze discounts



Bar chart for discount percent by year.

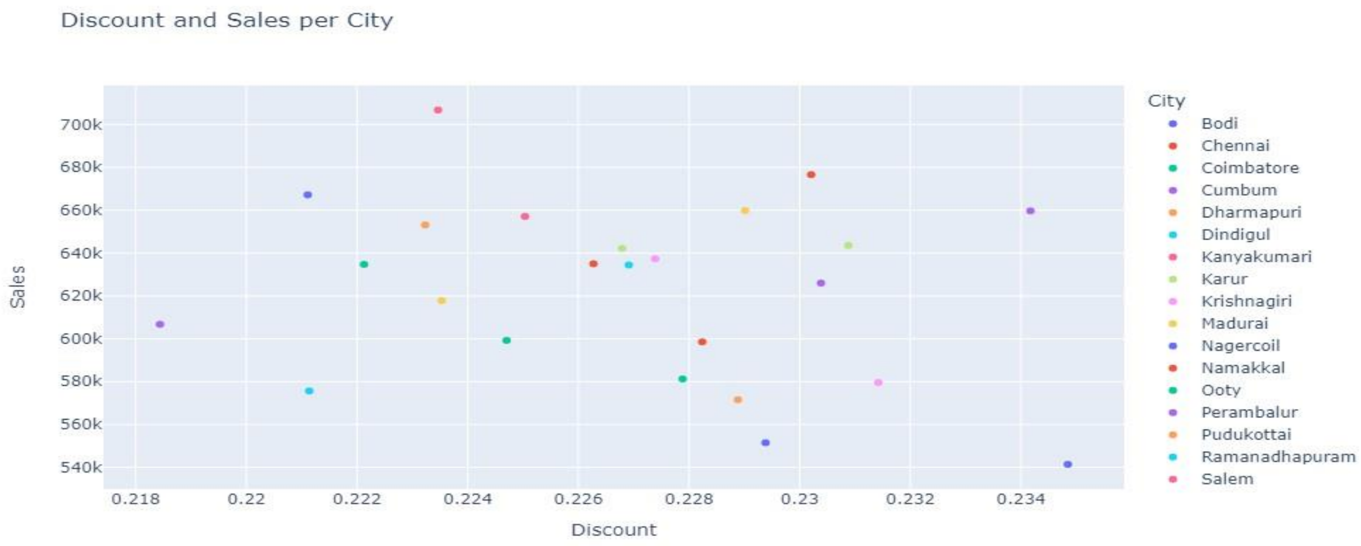
Year	Discount (in %)	Inference
2015	22.73	Slightly below average, indicating a conservative approach to discounts.
2016	22.76	suggesting an aggressive promotional strategy.
2017	22.62	indicating a pullback in discounts offered.
2018	22.53	an effort to stabilize or boost profit margins.

Spread:
The discounts are consistently around 22%, with minimal fluctuation.
The spread between the discount percentages is relatively small.
Extreme Points:
Lowest Discount: In 2018, the discount percentage is 22.53%.
Highest Discount: In 2016, the discount percentage reaches 22.76%.
Average Value:
The average discount over these four years is approximately 22.66%.



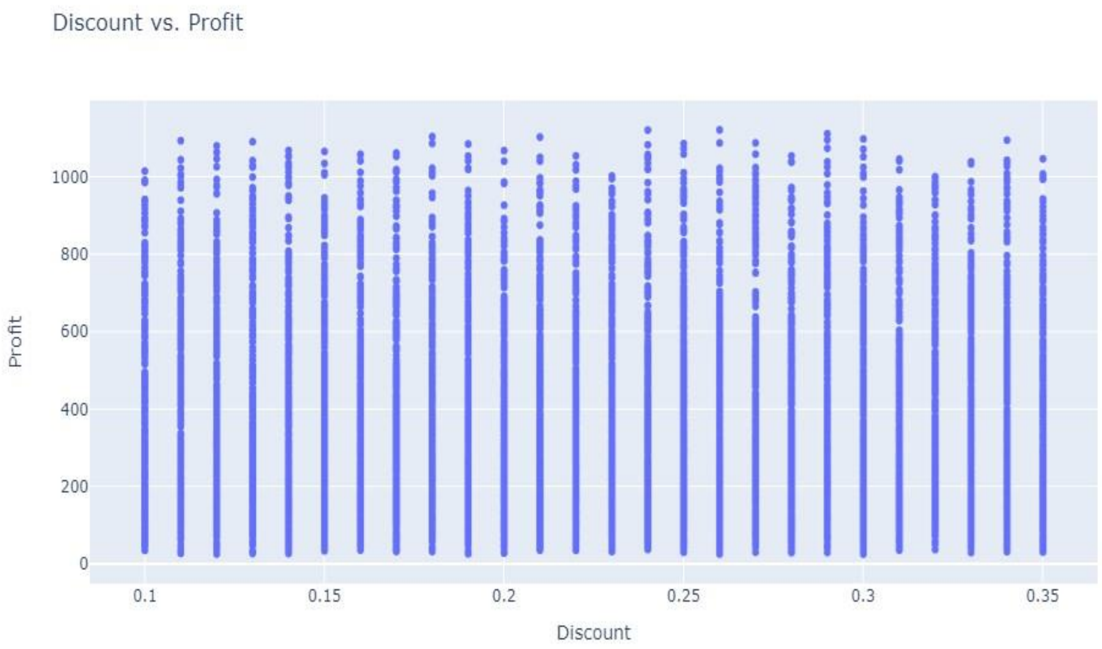
scatter plot for discount and sales per region.

From the Discount and Sales per Region scatter plot, we can see that the South region has the highest discount percent and sales, while the Central region has the lowest discount percent and sales.



From the Discount and Sales per City scatter plot, we can see that the cities with the highest sales and discount percent are Krishnagiri and Vellore, while the city with the lowest sales and discount percent is Trichy.

Analyze discounts and their impact on profit



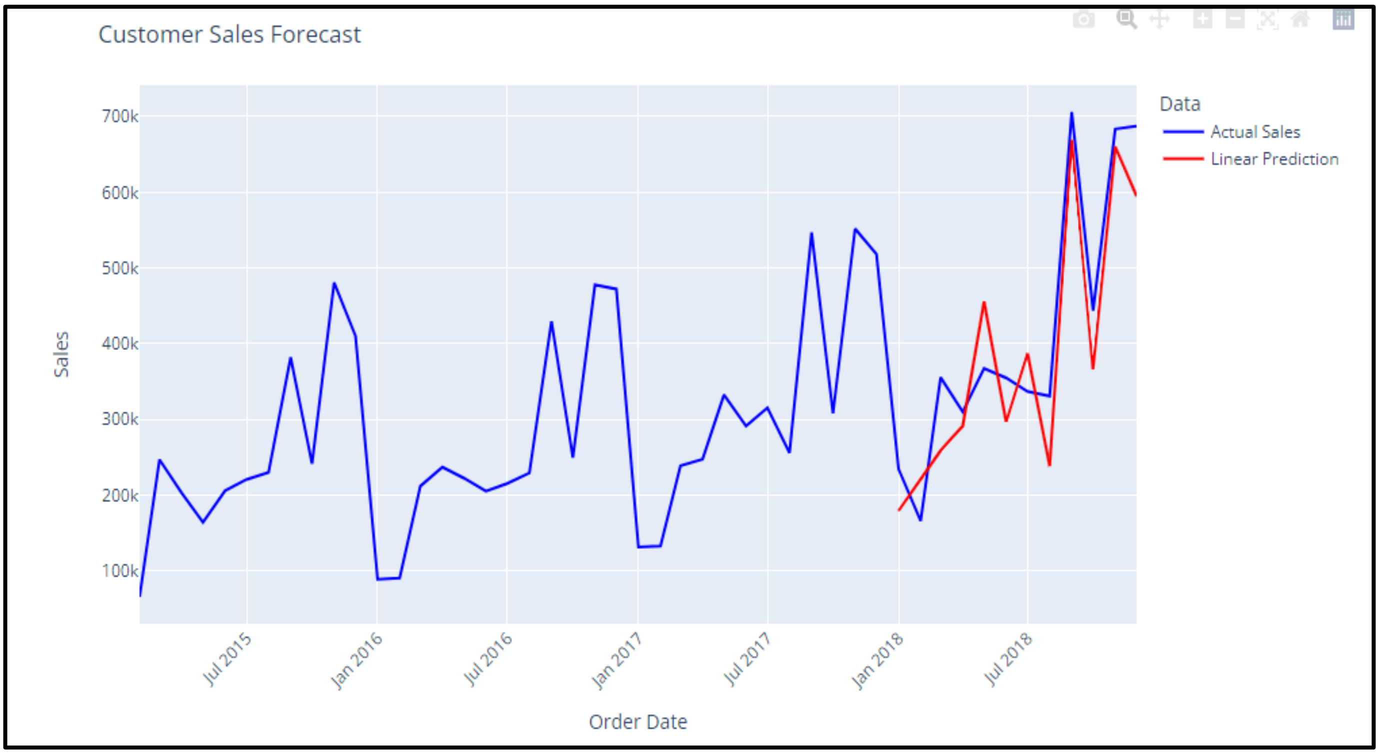
From the results of the analysis, we might find that certain categories or sub-categories have higher sales and profits than others, and that discounts have a negative impact on profit. Based on these findings

Recommendations:

a. Adjust Discount Strategy: Reduce discounts or change strategies to preserve or enhance profit margins.

b. Promotional Focus: Direct promotions towards high-performing categories/sub-categories to maximize sales and profits.

Sales Forecasting through Regression:



The red line shows the linear regression prediction. For the training data, previous years data were used and the last year was spitted onto the test data. Following metrics were achieved by the model after training.

```
Linear Regression MSE:  67339.34894633491
Linear Regression MAE:  62022.5798985284
Linear Regression R2:  0.8340833853830065
```

4.4) QUALITY ASPECTS:

The Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets places a strong emphasis on maintaining high quality throughout the project lifecycle. Key quality control measures and quality aspects achieved in this project include:

1. **Data Quality Control:** Rigorous data validation and preprocessing techniques ensure the accuracy, integrity, and completeness of the collected sales data, enhancing the reliability of analysis and forecasting results.
2. **Model Performance Evaluation:** Thorough model evaluation using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) score ensures the accuracy, robustness, and generalizability of the sales forecasting model, meeting stakeholder expectations.
3. **Documentation and Reporting:** Comprehensive project documentation and reporting provide clear and transparent insights into the project workflow, methodologies, findings, and recommendations, facilitating reproducibility, transparency, and accountability.
4. **Stakeholder Engagement:** Active engagement with stakeholders, including supermarket owners, managers, and data analysts, ensures that project objectives, requirements, and deliverables align with business needs and expectations, enhancing satisfaction and adoption of project outcomes.
5. **Continuous Improvement:** Iterative development and refinement of analysis techniques, modelling approaches, and interpretation strategies based on stakeholder feedback and lessons learned contribute to continuous improvement and optimization of project outcomes.
6. **Cross-Validation and Sensitivity Analysis:** Cross-validation techniques and sensitivity analysis validate the robustness and reliability of analysis results, providing confidence in the accuracy and stability of insights derived from the project.
7. **Adherence to Best Practices:** Adherence to industry best practices, standards, and methodologies in data science, including proper data handling, preprocessing, modelling, and interpretation, ensures the reliability, validity, and reproducibility of project outcomes.

By incorporating these quality control measures and achieving key quality aspects, the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets delivers high-quality, reliable, and actionable insights that drive informed decision-making and strategic planning for supermarket stakeholders.

Chapter 5
Standards Adopted

5.1 Design Standards

- We followed the standard principles of data Pre-processing to ensure data quality and accuracy, including handling missing data, outlier detection and removal, and data normalization. - We used random forest for diabetes prediction, which are a well- established and widely used techniques in the field of data science.

5.2 Coding Standards

- We followed the PEP 8 style guide for Python code to ensure code readability and consistency. - We used modular code design and separated code into different functions and classes to improve maintainability and re-usability. - We added comments and documentation to our code to improve code understand-ability and maintainability.

5.3 Testing Standards

- We followed the standard approach of splitting our data-set into training and testing sets to evaluate the performance of our model.

We used different evaluation metrics, including confusion matrices to evaluate the performance of our algorithm. Overall, we aimed to maintain high standards throughout our project to ensure the reliability, robustness, and maintainability of our sales forecasting algorithm

Chapter 6

Conclusion and Future Scope

Conclusion:

In conclusion, the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets has provided invaluable insights and solutions to the challenges faced by small-scale supermarkets, exemplified by "Ponnu Super Bazzar". Through meticulous data analysis, preprocessing, and the development of a robust sales forecasting model, this project has empowered supermarket owners and managers with actionable insights to optimize inventory management, pricing strategies, and resource allocation. The project's impactful visualizations have illuminated critical sales trends, regional variations, and customer preferences, enabling stakeholders to make informed decisions and strategic plans. Moreover, the accurate sales forecasting model, achieved through rigorous evaluation and validation, has equipped supermarkets with the foresight needed to anticipate market dynamics and plan accordingly, ultimately enhancing operational efficiency and profitability. By addressing the problem statement with precision and delivering tangible solutions, this project underscores the transformative potential of data science in the retail sector, paving the way for sustainable growth and success in the competitive marketplace.

Future Scope:

Looking ahead, the future scope of the Data Science Project on Sales Analysis and Sales Forecasting of Small-Scale Supermarkets is promising and multifaceted. One avenue for expansion involves the development of a real-time sales analysis visualization API using Flask, tailored specifically for the "Ponnu Super Bazzar" store chain. This API would enable stakeholders to access dynamic visualizations and insights on sales performance, inventory trends, and customer behaviour in real-time, facilitating proactive decision-making and strategic planning. Additionally, enhancing the training of the sales forecasting model presents an exciting opportunity to improve its predictive accuracy further. By exploring advanced techniques such as ensemble learning, time series analysis, and feature engineering, the goal would be to elevate the model's R-squared (R2) score from the current 0.834 to 0.91 or higher, ensuring more reliable and precise sales forecasts. Furthermore, future work could involve integrating external data sources, such as weather patterns or economic indicators, to enhance the predictive capabilities of the model. Additionally, implementing a feedback mechanism to continuously refine and update the model based on real-world sales data would ensure its adaptability to evolving market dynamics. Overall, the future scope of this project is rich with opportunities for innovation and advancement, with the potential to drive tangible value and insights for small-scale supermarkets in their quest for growth and competitiveness in the retail landscape.

References

1. Karm J. P. & Maldonado, S. (2019). Hierarchical time series forecasting via Support Rector Regression in the European travel retail industry. Expert Systems with Applications, 137, 5973, DIO: 10.1016
2. Mohit G. (2017). Forecasting of sales by using the fusion of machine learning techniques. International Conference on Data Management, Analytics and Innovation (ICDMAI), IEEE., 93–101
3. Supermarket Grocery Sales Dataset- https://www.kaggle.com/datasets/mohamedharris/supermart-grocery-sales-retail-analytics-dataset
4. Plotly.Express Documentation : https://plotly.com/python/plotly-express/

DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS

SHIV PRASAD ROUL
2105158

Abstract: This project explores sales analysis and forecasting in small-scale supermarkets using data from "Ponnu Super Bazzar" in Tamil Nadu, India. Leveraging Python and libraries like Matplotlib, Plotly.express, and pandas, it conducts thorough exploratory data analysis (EDA) to uncover insights into sales patterns. A linear regression model predicts future sales with 84% accuracy. With 10,000 rows of data, this project demonstrates the power of data science in driving informed decisions for supermarket growth.

Individual contribution and findings: To implement my part in the project, which involves data collection and preprocessing, I devised a detailed plan to ensure efficiency and accuracy in handling the dataset. Initially, I scheduled time to thoroughly understand the requirements regarding the data needed from "Ponnu Super Bazzar". Following this, I outlined a systematic approach to collecting the historical sales data, focusing on obtaining all relevant features outlined in the project scope. Once the data was collected, I planned to dedicate ample time to preprocessing, focusing on tasks such as handling missing values, outliers, and inconsistencies to ensure the integrity and completeness of the dataset.

Technical Findings and Experience: During the implementation of my part in the project, I encountered several technical findings and gained valuable experience in data collection and preprocessing. I found that obtaining the historical sales data required careful coordination with stakeholders and adherence to data privacy and confidentiality protocols. Additionally, I learned various techniques for handling missing values, outliers, and inconsistencies during the preprocessing stage, such as imputation, outlier detection, and data transformation. Through practical application, I gained a deeper understanding of the importance of data quality and integrity in driving reliable analysis and modelling outcomes. Overall, the experience allowed me to refine my skills in data management and preprocessing, contributing to the successful execution of my part in the project.

Full Signature of Supervisor:	Full signature of the student:
…………………………….	………………………………..

DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS

MILAN KUMAR SAHOO
2105208

Abstract: This project explores sales analysis and forecasting in small-scale supermarkets using data from "Ponnu Super Bazzar" in Tamil Nadu, India. Leveraging Python and libraries like Matplotlib, Plotly.express, and pandas, it conducts thorough exploratory data analysis (EDA) to uncover insights into sales patterns. A linear regression model predicts future sales with 84% accuracy. With 10,000 rows of data, this project demonstrates the power of data science in driving informed decisions for supermarket growth.

Contribution Report -

Planning: For my part in the project, which involves exploratory data analysis (EDA) and visualization, I devised a comprehensive plan to effectively uncover insights from the sales data and present them through visualizations. Initially, I allocated time to thoroughly understand the dataset and identify key areas of interest for analysis. I then outlined a structured approach to conducting exploratory data analysis, focusing on exploring sales patterns, correlations, and distributions across different categories, subcategories, cities, and regions. Additionally, I planned to leverage various visualization techniques to communicate findings clearly and effectively to stakeholders.

Technical Findings and Experience: During the implementation of my part in the project, I gained valuable technical findings and experience in exploratory data analysis and visualization. I found that visualizing sales data using techniques such as scatter plots, bar charts, and heatmaps provided valuable insights into trends and patterns that were not immediately apparent from the raw data. Additionally, I learned the importance of choosing the right visualization techniques to effectively communicate different types of information, such as trends over time, geographic variations, and category-wise sales distributions. Through practical application, I honed my skills in data visualization and interpretation, contributing to the comprehensive understanding of the sales data and its implications for decision-making.

Contribution in Documentation: As part of my contribution to the project, I documented the insights gathered from the visualizations in a dedicated section of the project documentation. This documentation included a summary of key findings, accompanied by relevant visualizations and interpretations. By documenting the insights from visualization, I ensured that stakeholders had access to clear and actionable information to guide their decision-making processes.

Full Signature of Supervisor: Full signature of the student:

……………………………. ……………………………

DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS

AMOGH
2105240

Abstract: This project explores sales analysis and forecasting in small-scale supermarkets using data from "Ponnu Super Bazzar" in Tamil Nadu, India. Leveraging Python and libraries like Matplotlib, Plotly.express, and pandas, it conducts thorough exploratory data analysis (EDA) to uncover insights into sales patterns. A linear regression model predicts future sales with 84% accuracy. With 10,000 rows of data, this project demonstrates the power of data science in driving informed decisions for supermarket growth.

Contribution Report

Planning: I meticulously planned my approach to ensure the effectiveness of the project. Initially, I dedicated time to thoroughly understand the requirements and objectives of the regression modelling task. Subsequently, I outlined a step-by-step plan for extracting the necessary data for modelling and preprocessing it to prepare it for regression analysis. This involved identifying relevant features, handling missing values, encoding categorical variables, and scaling numerical features. Additionally, I scheduled time to explore different regression techniques and select the most suitable one for the project's objectives.

Technical Findings and Experience: During the implementation of my part in the project, I gained valuable technical findings and experience in regression modelling and data preprocessing. I found that extracting and preprocessing the data required careful attention to detail and a thorough understanding of the underlying data characteristics. Techniques such as feature scaling and encoding categorical variables were crucial for ensuring the accuracy and effectiveness of the regression model. Additionally, I learned the importance of selecting appropriate evaluation metrics and cross-validation techniques to assess the performance of the regression model accurately. Through practical application, I enhanced my skills in regression modelling and data preprocessing, contributing to the successful execution of my part in the project.

Full Signature of Supervisor: Full signature of the student:

……………………………. ……………………………

DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS

SOVAN PATTANAIK
2105247

Abstract: This project explores sales analysis and forecasting in small-scale supermarkets using data from "Ponnu Super Bazzar" in Tamil Nadu, India. Leveraging Python and libraries like Matplotlib, Plotly.express, and pandas, it conducts thorough exploratory data analysis (EDA) to uncover insights into sales patterns. A linear regression model predicts future sales with 84% accuracy. With 10,000 rows of data, this project demonstrates the power of data science in driving informed decisions for supermarket growth.

Contribution Report

Planning: As the sole contributor responsible for every aspect of the project, I meticulously planned and executed each task to ensure its successful completion. Initially, I outlined a comprehensive project plan, identifying key milestones and allocating time for each phase, from data collection to model development and visualization. I prioritized tasks based on their criticality and dependencies, ensuring a smooth workflow and timely delivery. Additionally, I scheduled regular checkpoints to review progress and adjust strategies as needed, maintaining flexibility to accommodate any unforeseen challenges.

Technical Findings and Experience: Throughout the project, I gained invaluable technical insights and experience across various domains, including data preprocessing, regression modelling, visualization, and documentation. Preprocessing the data involved handling missing values, encoding categorical variables, and scaling features to prepare it for regression analysis. Training the regression model required selecting appropriate algorithms, tuning hyperparameters, and evaluating performance metrics to ensure accuracy and reliability. Visualizing the sales forecasting results involved creating insightful charts and graphs to communicate trends and predictions effectively. Through practical application, I deepened my understanding of data science techniques and honed my skills in coding, modelling, and visualization.

Contribution in Documentation: As the primary contributor, I meticulously documented every aspect of the project, from initial planning to final implementation and results interpretation. I drafted detailed project reports, outlining the problem statement, objectives, methodologies, findings, and recommendations. I ensured clarity and completeness in each chapter, providing comprehensive explanations and insights derived from data analysis and modelling. Additionally, I created informative visual aids and diagrams to enhance understanding and facilitate knowledge transfer. By overseeing documentation from start to end, I ensured the project's transparency, reproducibility, and accessibility to stakeholders.

Full Signature of Supervisor:	Full signature of the student:
…………………………….	……………………………

DATA SCIENCE PROJECT ON SALES ANALYSIS AND FORECASTING FOR SMALL SCALE LOCAL SUPERMARKETS

AADITYA CHOWDHURY
2105341

Abstract: This project explores sales analysis and forecasting in small-scale supermarkets using data from "Ponnu Super Bazzar" in Tamil Nadu, India. Leveraging Python and libraries like Matplotlib, Plotly.express, and pandas, it conducts thorough exploratory data analysis (EDA) to uncover insights into sales patterns. A linear regression model predicts future sales with 84% accuracy. With 10,000 rows of data, this project demonstrates the power of data science in driving informed decisions for supermarket growth.

Contribution Report

Planning: In my role as the optimizer and quality controller of the project, I meticulously planned my contributions to ensure the enhancement of the model's accuracy and overall quality. Initially, I analysed the existing regression model and identified areas for improvement. I then devised a systematic plan to optimize the model by experimenting with different algorithms, tuning hyperparameters, and refining feature engineering techniques. Additionally, I collaborated with Sovan to integrate Plot.ly for web integration and interactive visualization, ensuring a user-friendly and engaging interface for stakeholders. Throughout the process, I prioritized quality control measures, ensuring adherence to coding standards and best practices to maintain consistency and reliability in the project's codebase.

Technical Findings and Experience: During the optimization phase, I gained valuable technical insights and experience in machine learning algorithms, hyperparameter tuning, and feature engineering. Experimenting with various algorithms and techniques allowed me to identify the most effective strategies for improving model accuracy. Additionally, integrating Plotly for web visualization enhanced the project's accessibility and user experience, providing stakeholders with interactive insights into sales forecasting trends. Quality control measures, including code reviews and adherence to coding standards, ensured the robustness and reliability of the project's implementation. Through practical application, I enhanced my skills in model optimization, web integration, and quality assurance, contributing to the overall success of the project.

Full Signature of Supervisor: Full signature of the student:

……………………………. ……………………………