INT375

DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

US CANDY DISTRIBUTER

Submitted by

Suhani Kumari

Registration No- 12315550

Programme and Section- B. Tech CSE- K23GD

Course Code- INT375

Under the Guidance of

Baljindar Kaur (27952)

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, Suhani Kumari, student of B.Tech Computer Science Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Suhani

Date: 13/04/2025

Registration No. 12315550 Suhani Kumari

CERTIFICATE

This is to certify that Suhani Kumari bearing Registration no. 12315550 has completed INT375 project titled, “Candy Sales” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 13/04/2025

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to all those who supported and guided me throughout the development of this Python project on candy sales analysis.

First and foremost, I am deeply thankful to my instructor, Baljindar Kaur, for her continuous support, expert guidance, and valuable feedback, which significantly contributed to the successful completion of this project.

I would also like to extend my sincere thanks to Lovely Professional University for providing a conducive learning environment and the necessary resources that enabled me to carry out this project effectively.

A special thanks to my friends and family for their constant encouragement and moral support throughout this journey.

Lastly, I am grateful for the use of real-world data, such as the Candy_Sales.csv dataset, which allowed me to apply Python-based tools and techniques in a practical context, enhancing both my learning and analytical skills.

This project has been a valuable experience, and I am truly appreciative of everyone who played a role in its success.

Suhani kumari

13-04-2025

TABLE OF CONTENTS

1. INTRODUCTION

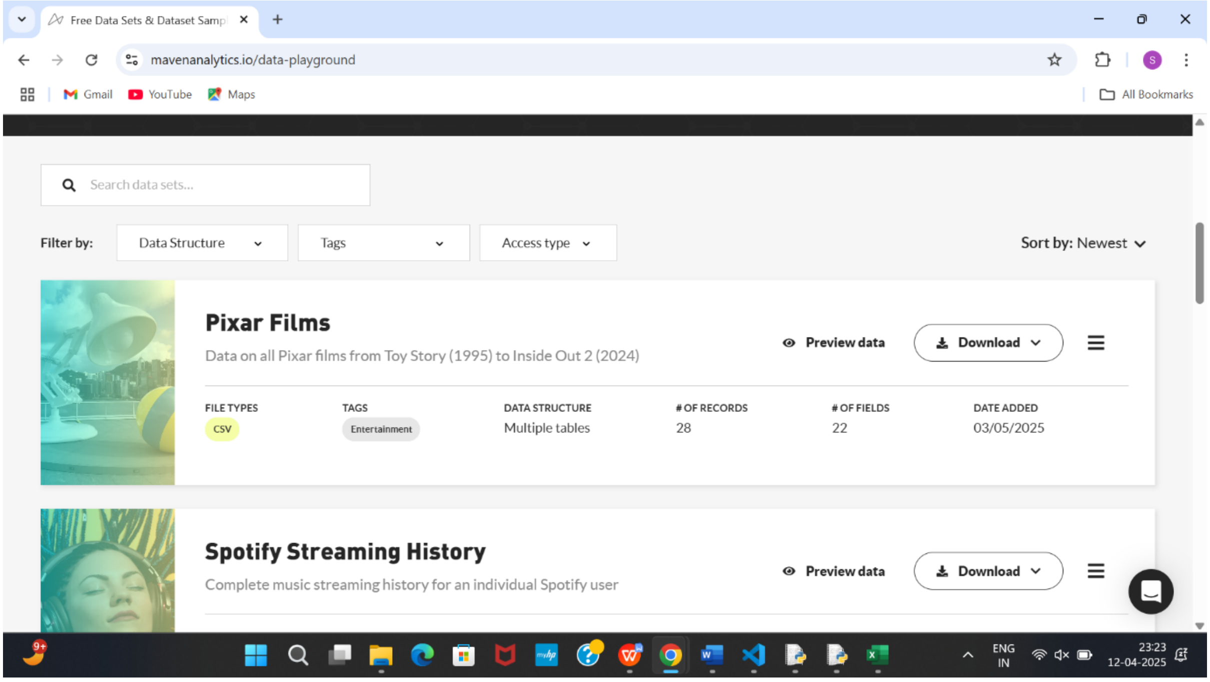
His project focuses on analysing candy sales data using Python. With the help of the dataset **Candy_Sales.csv**, the project aims to extract meaningful insights, identify sales trends, and visualize patterns in consumer behaviour. By leveraging powerful Python libraries such as **pandas**, **matplotlib**, and **seaborn**, this project demonstrates how data analysis techniques can be applied to real-world scenarios. The objective is to enhance decision-making capabilities through effective data handling, cleaning, and visualization.

This project illustrates the journey from raw data ingestion to actionable outcomes, showcasing how data preprocessing, visualization, and statistical modelling can collectively paint a detailed picture of business health. The integration of Python's data ecosystem allows for the execution of a full pipeline of operations—from feature engineering to exploratory analysis and visual storytelling. This approach provides not only academic rigor but also the potential for real-world application in decision-making processes.

Key objectives of the analysis include studying sales trends, evaluating customer behaviour across regions and segments, assessing operational performance through shipping analysis, and understanding patterns in product returns. Through these objectives, this report not only demonstrates technical competence in data science but also aligns with the broader goal of enhancing strategic business operations.

This project adheres to three core evaluation pillars:

- **Data Cleaning and Visualization**
- **EDA and Statistical Analysis**
- **Creativity and Innovation**


Free Data Sets & Dataset Samp
mavenanalytics.io/data-playground
Gmail
YouTube
Maps
All Bookmarks
Search data sets...

Filter by: Data Structure | Tags | Access type

Sort by: Newest

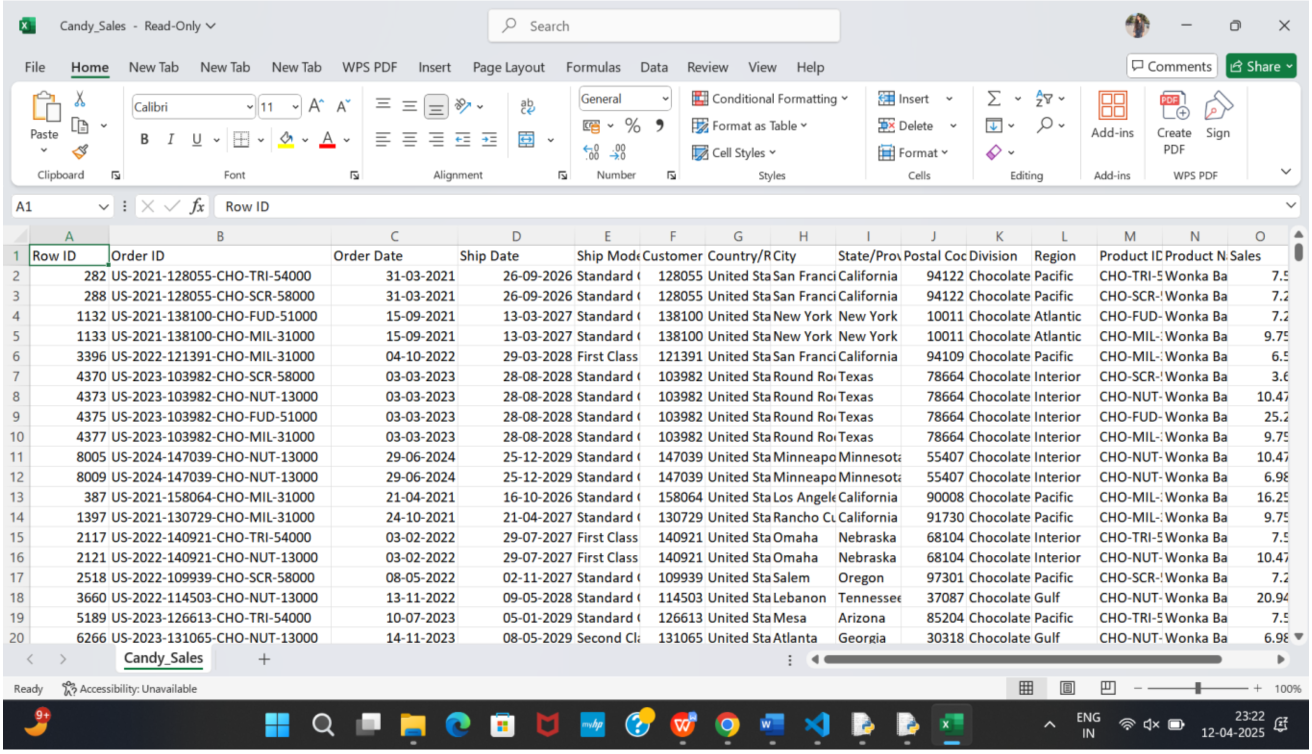
Pixar Films

Data on all Pixar films from Toy Story (1995) to Inside Out 2 (2024)

FILE TYPES	TAGS	DATA STRUCTURE	# OF RECORDS	# OF FIELDS	DATE ADDED
CSV	Entertainment	Multiple tables	28	22	03/05/2025

Spotify Streaming History

Complete music streaming history for an individual Spotify user



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Country/R	City	State/Prov	Postal Coc	Division	Region	Product ID	Product N	Sales
2	282	US-2021-128055-CHO-TRI-54000	31-03-2021	26-09-2026	Standard (	128055	United Sta	San Franci	California	94122	Chocolate	Pacific	CHO-TRI-5	Wonka Ba	7.5
3	288	US-2021-128055-CHO-SCR-58000	31-03-2021	26-09-2026	Standard (	128055	United Sta	San Franci	California	94122	Chocolate	Pacific	CHO-SCR-!	Wonka Ba	7.2
4	1132	US-2021-138100-CHO-FUD-51000	15-09-2021	13-03-2027	Standard (	138100	United Sta	New York	New York	10011	Chocolate	Atlantic	CHO-FUD-	Wonka Ba	7.2
5	1133	US-2021-138100-CHO-MIL-31000	15-09-2021	13-03-2027	Standard (	138100	United Sta	New York	New York	10011	Chocolate	Atlantic	CHO-MIL-:	Wonka Ba	9.75
6	3396	US-2022-121391-CHO-MIL-31000	04-10-2022	29-03-2028	First Class	121391	United Sta	San Franci	California	94109	Chocolate	Pacific	CHO-MIL-:	Wonka Ba	6.5
7	4370	US-2023-103982-CHO-SCR-58000	03-03-2023	28-08-2028	Standard (	103982	United Sta	Round Ro	Texas	78664	Chocolate	Interior	CHO-SCR-!	Wonka Ba	3.6
8	4373	US-2023-103982-CHO-NUT-13000	03-03-2023	28-08-2028	Standard (	103982	United Sta	Round Ro	Texas	78664	Chocolate	Interior	CHO-NUT-	Wonka Ba	10.47
9	4375	US-2023-103982-CHO-FUD-51000	03-03-2023	28-08-2028	Standard (	103982	United Sta	Round Ro	Texas	78664	Chocolate	Interior	CHO-FUD-	Wonka Ba	25.2
10	4377	US-2023-103982-CHO-MIL-31000	03-03-2023	28-08-2028	Standard (	103982	United Sta	Round Ro	Texas	78664	Chocolate	Interior	CHO-MIL-:	Wonka Ba	9.75
11	8005	US-2024-147039-CHO-NUT-13000	29-06-2024	25-12-2029	Standard (	147039	United Sta	Minneapo	Minnesota	55407	Chocolate	Interior	CHO-NUT-	Wonka Ba	10.47
12	8009	US-2024-147039-CHO-NUT-13000	29-06-2024	25-12-2029	Standard (	147039	United Sta	Minneapo	Minnesota	55407	Chocolate	Interior	CHO-NUT-	Wonka Ba	6.98
13	387	US-2021-158064-CHO-MIL-31000	21-04-2021	16-10-2026	Standard (	158064	United Sta	Los Angele	California	90008	Chocolate	Pacific	CHO-MIL-:	Wonka Ba	16.25
14	1397	US-2021-130729-CHO-MIL-31000	24-10-2021	21-04-2027	Standard (	130729	United Sta	Rancho Cu	California	91730	Chocolate	Pacific	CHO-MIL-:	Wonka Ba	9.75
15	2117	US-2022-140921-CHO-TRI-54000	03-02-2022	29-07-2027	First Class	140921	United Sta	Omaha	Nebraska	68104	Chocolate	Interior	CHO-TRI-5	Wonka Ba	7.5
16	2121	US-2022-140921-CHO-NUT-13000	03-02-2022	29-07-2027	First Class	140921	United Sta	Omaha	Nebraska	68104	Chocolate	Interior	CHO-NUT-	Wonka Ba	10.47
17	2518	US-2022-109939-CHO-SCR-58000	08-05-2022	02-11-2027	Standard (	109939	United Sta	Salem	Oregon	97301	Chocolate	Pacific	CHO-SCR-!	Wonka Ba	7.2
18	3660	US-2022-114503-CHO-NUT-13000	13-11-2022	09-05-2028	Standard (	114503	United Sta	Lebanon	Tennessee	37087	Chocolate	Gulf	CHO-NUT-	Wonka Ba	20.94
19	5189	US-2023-126613-CHO-TRI-54000	10-07-2023	05-01-2029	Standard (	126613	United Sta	Mesa	Arizona	85204	Chocolate	Pacific	CHO-TRI-5	Wonka Ba	7.5
20	6266	US-2023-131065-CHO-NUT-13000	14-11-2023	08-05-2029	Second Cl	131065	United Sta	Atlanta	Georgia	30318	Chocolate	Gulf	CHO-NUT-	Wonka Ba	6.98

2. SOURCE OF DATASET

Olympic Athletes dataset source link- **https://mavenanalytics.io/data-playground**

The dataset utilized in this project, titled "Sample - Candys," is publicly accessible from Tableau's official website under the sample data section. It represents a collection of retail sales records,

Column name	Description
Row ID	Unique identifier for each row or transaction record.
Order ID	Unique identifier for each customer order.
Order Date	Date when the order was placed.
Ship Date	Date when the order was shipped.
Ship Mode	Mode of shipment used (e.g., Standard Class, First Class).
Customer ID	Unique identifier for each customer.
Country/Region	Country or region where the customer is located
City	City of the customer.
State/Province	State or province of the customer.
Postal Code	Postal code associated with the customer address.
Region	Geographic region for business operations
Division	Product category (e.g., Chocolate, Candy).

3. EXPLORATORY DATA ANALYSIS(EDA) PROCESS

The Exploratory Data Analysis (EDA) process is a crucial step in understanding the underlying structure and patterns within the dataset. In this project, the EDA was performed on the Candy_Sales.csv dataset and involved the following steps:

1. Data Loading

   The dataset was imported using the pandas library, enabling efficient data manipulation and analysis.

2. Data Inspection

   - Previewed the dataset using. head() and .info() functions.
   - Checked for missing values, data types, and overall structure using. isnull() .sum() and .describe().

3. Data Cleaning

   - Handled missing or inconsistent values if any were present.
   - Renamed columns or standardized formats when necessary.

4. Univariate Analysis

   - Analyzed individual variables using visualizations like histograms, bar charts, and count plots.
   - Studied distributions of sales quantities, candy types, regions, etc.

5. Bivariate/Multivariate Analysis

   - Explored relationships between variables using scatter plots, heatmaps, and groupby () operations.
   - Investigated patterns such as sales trends over time or performance by region/product.

6. Visualization

   - Used matplotlib and seaborn to create clear and insightful visual representations.

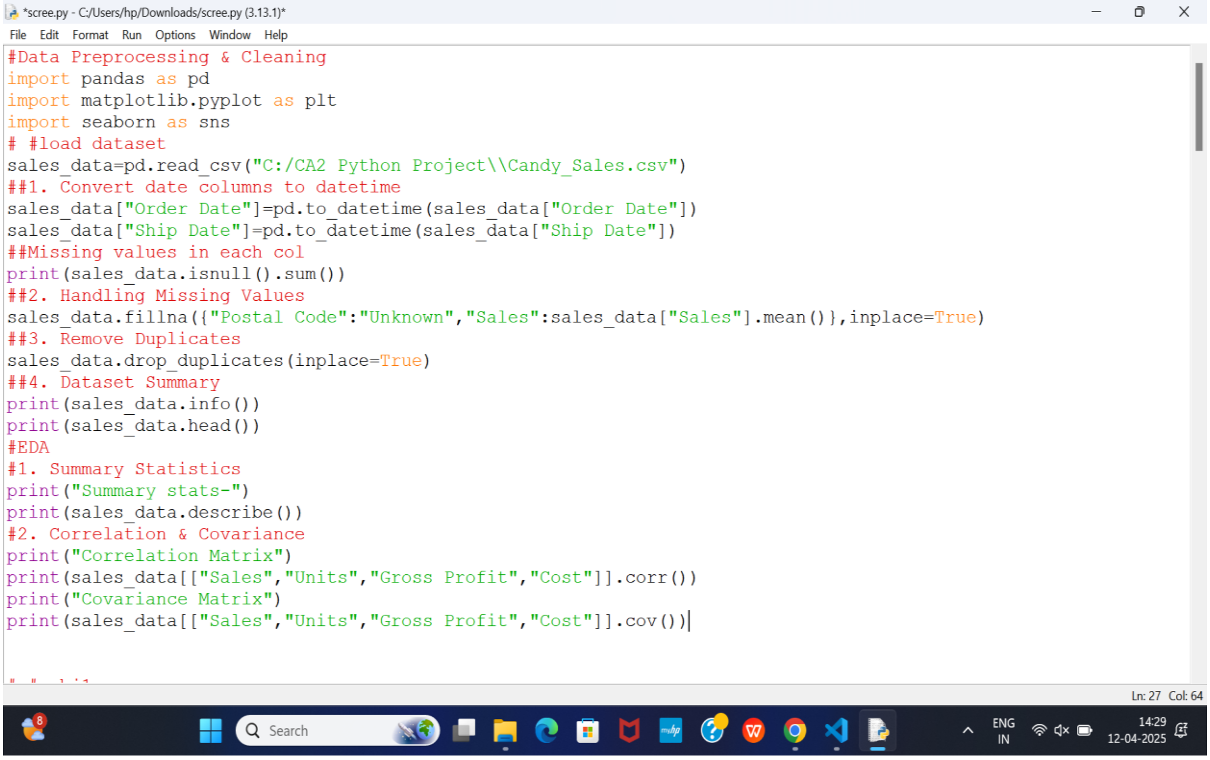
   - Graphs included line plots, pie charts, and box plots to highlight comparisons and trends.

7. Insights and Summary

   - Derived meaningful insights from the data.

   - Highlighted key findings such as top-selling products, peak sales periods, and regional performance.

The EDA provided a comprehensive overview of the dataset and set the foundation for further analysis and interpretation.



```
#Data Preprocessing & Cleaning
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# #load dataset
sales_data=pd.read_csv("C:/CA2 Python Project\\Candy_Sales.csv")
##1. Convert date columns to datetime
sales_data["Order Date"]=pd.to_datetime(sales_data["Order Date"])
sales_data["Ship Date"]=pd.to_datetime(sales_data["Ship Date"])
##Missing values in each col
print(sales_data.isnull().sum())
##2. Handling Missing Values
sales_data.fillna({"Postal Code":"Unknown","Sales":sales_data["Sales"].mean()},inplace=True)
##3. Remove Duplicates
sales_data.drop_duplicates(inplace=True)
##4. Dataset Summary
print(sales_data.info())
print(sales_data.head())
#EDA
#1. Summary Statistics
print("Summary stats-")
print(sales_data.describe())
#2. Correlation & Covariance
print("Correlation Matrix")
print(sales_data[["Sales","Units","Gross Profit","Cost"]].corr())
print("Covariance Matrix")
print(sales_data[["Sales","Units","Gross Profit","Cost"]].cov())
```

4. ANALYTICAL FRAMEWORKS AND INTERFERENCES

4.1 Objective1- calculate the total revenue, average revenue per order, and standard deviation of sales from the Candy_Sales dataset.

Introduction

This project is centred around analysing candy sales data using Python. The dataset, Candy_Sales.csv, includes information on different candy orders, including price, quantity, and total sales. The goal is to perform a statistical analysis to compute the total revenue, average revenue per order, and the standard deviation of sales, providing meaningful insights into business performance. This project utilizes Python libraries like pandas and numpy for data analysis and matplotlib/seaborn for visualization.

General Description

The dataset contains records of candy sales, including the number of items sold and the corresponding sales value. This data can help in understanding the revenue distribution and order behaviour across different entries. The project focuses on three key financial metrics:

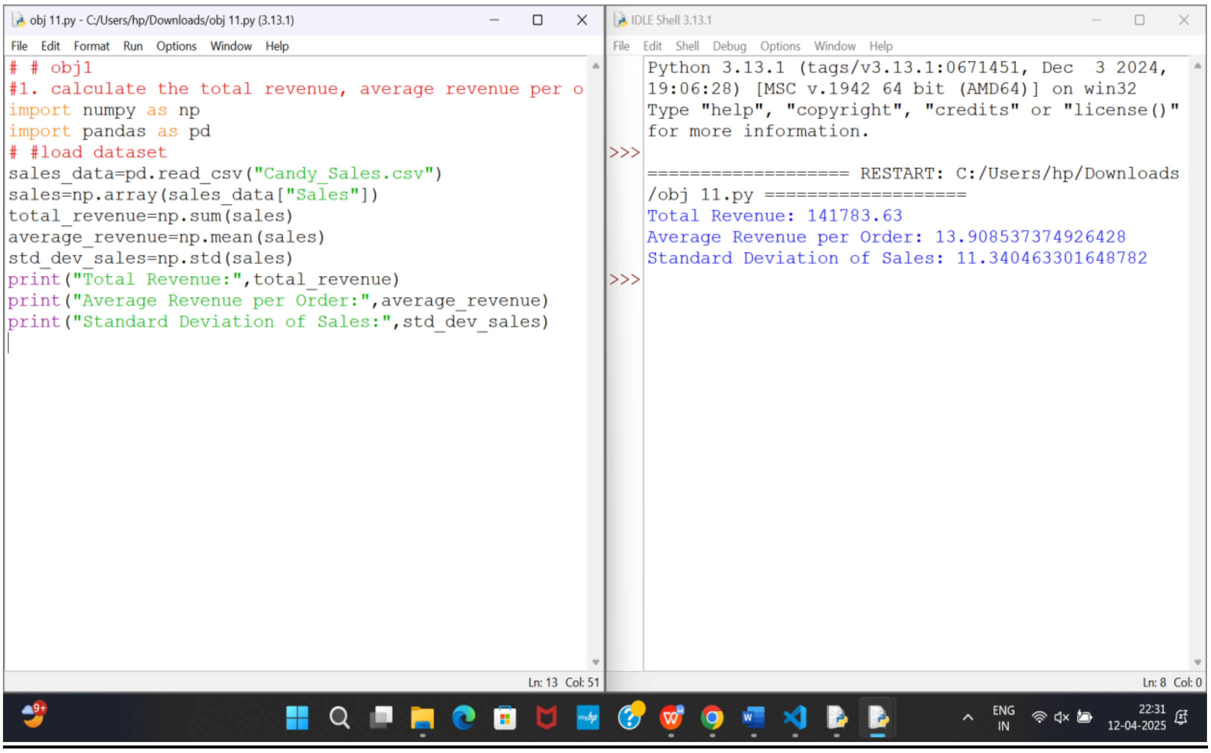
- Total Revenue – Sum of all sales
- Average Revenue per Order – Mean of sales values
- Standard Deviation of Sales – Variability in sales revenue across orders

Specific Requirements, Functions and Formulas

Tools Used:

- Python
- Libraries: pandas, NumPy, matplotlib, seaborn

Analysis Result


obj 11.py - C:/Users/hp/Downloads/obj 11.py (3.13.1)
File Edit Format Run Options Window Help
Ln: 13 Col: 51
IDLE Shell 3.13.1
File Edit Shell Debug Options Window Help
Ln: 8 Col: 0
ENG IN
22:31
12-04-2025

```
# # obj1
#1. calculate the total revenue, average revenue per o
import numpy as np
import pandas as pd
# #load dataset
sales_data=pd.read_csv("Candy_Sales.csv")
sales=np.array(sales_data["Sales"])
total_revenue=np.sum(sales)
average_revenue=np.mean(sales)
std_dev_sales=np.std(sales)
print("Total Revenue:",total_revenue)
print("Average Revenue per Order:",average_revenue)
print("Standard Deviation of Sales:",std_dev_sales)
```

```
Python 3.13.1 (tags/v3.13.1:0671451, Dec  3 2024,
19:06:28) [MSC v.1942 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()"
for more information.
>>>
=================== RESTART: C:/Users/hp/Downloads
/obj 11.py ===================
Total Revenue: 141783.63
Average Revenue per Order: 13.908537374926428
Standard Deviation of Sales: 11.340463301648782
>>>
```

4.2 Objective2- find the total number of orders, total units sold, and the state with the highest sales.

Introduction

The primary objective is to identify the total number of orders, total units sold, and determine the state with the highest sales. Through exploratory data analysis (EDA), statistical functions, and data visualization, the project showcases how data-driven decisions can be made effectively.

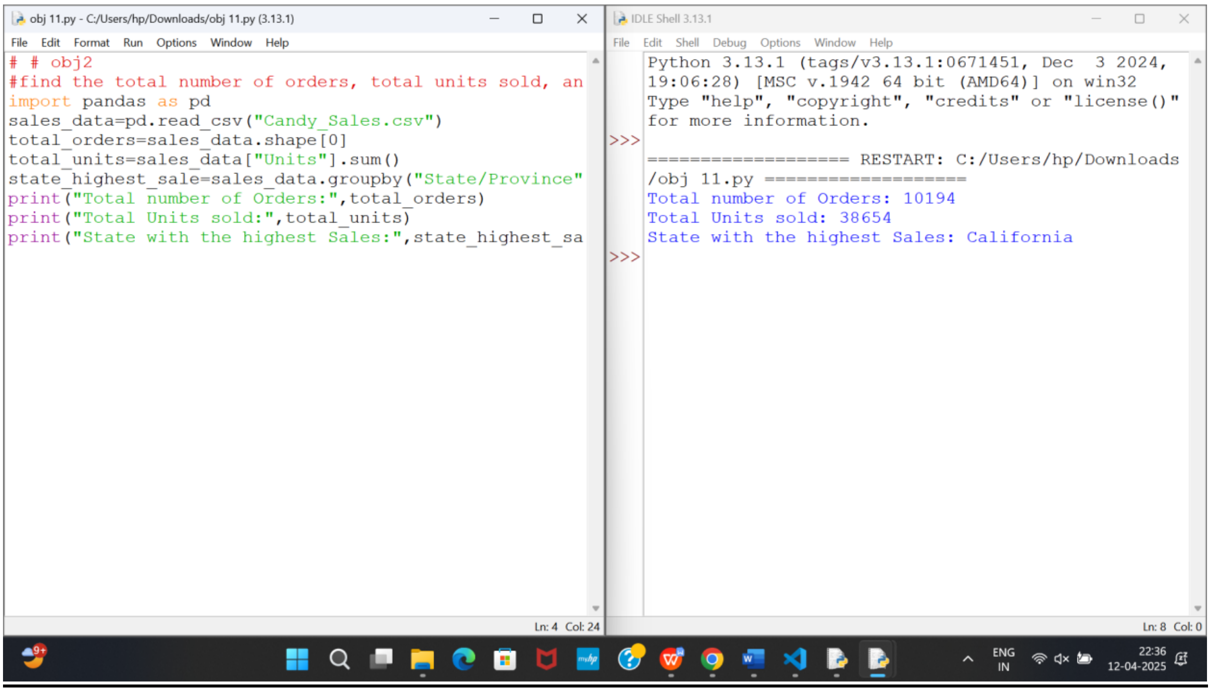
General Description

The dataset, Candy_Sales.csv, contains transactional information about candy sales across different states. Each row represents an order, with details such as the state, product name, quantity sold, and other relevant fields. The analysis aims to process this dataset, summarize key metrics.

Specific Requirements, Functions and Formulas

To achieve the objectives, the following tools and methods were used:

- Python Libraries:
  - pandas for data loading and manipulation
  - matplotlib and seaborn for visualization
- Functions & Techniques:
  - pd.read_csv() to load the dataset
  - .shape[0] to calculate total number of orders
  - .sum() to calculate total units sold
  - .groupby() with .sum() and .idxmax() to find the state with the highest sales
- **Analysis Result**


obj 11.py - C:/Users/hp/Downloads/obj 11.py (3.13.1)
File Edit Format Run Options Window Help
Ln: 4 Col: 24
IDLE Shell 3.13.1
File Edit Shell Debug Options Window Help
Ln: 8 Col: 0
ENG IN
22:36
12-04-2025

```
# # obj2
#find the total number of orders, total units sold, an
import pandas as pd
sales_data=pd.read_csv("Candy_Sales.csv")
total_orders=sales_data.shape[0]
total_units=sales_data["Units"].sum()
state_highest_sale=sales_data.groupby("State/Province"
print("Total number of Orders:",total_orders)
print("Total Units sold:",total_units)
print("State with the highest Sales:",state_highest_sa
```

```
Python 3.13.1 (tags/v3.13.1:0671451, Dec  3 2024,
19:06:28) [MSC v.1942 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()"
for more information.
>>>
=================== RESTART: C:/Users/hp/Downloads
/obj 11.py ===================
Total number of Orders: 10194
Total Units sold: 38654
State with the highest Sales: California
>>>
```

4.3 Objective3- "Analysing Sales Performance and Profitability Trends of Candy Products Using Data Visualization Techniques"

Introduction

In the competitive world of retail, understanding sales performance and profitability is crucial for strategic decision-making. This project aims to analyze candy product sales using data visualization techniques to uncover key insights. By studying trends over time, regional performance, and profit dynamics, the project provides a data-driven overview that can aid in improving business outcomes.

General Description

The dataset used in this project, Candy_Sales.csv, contains information on product sales, costs, gross profits, regions, and other related variables. The analysis focuses on identifying patterns, comparing regional performance, evaluating the distribution of key metrics, and exploring relationships between sales, costs, and profita**bility.**

Specific Requirements, Functions and Formulas

- **Requirements:**
  - Python 3.x environment
  - Libraries: pandas, matplotlib, seaborn, numpy
  - Clean and structured dataset (Candy_Sales.csv)
- **Key Functions/Methods Used:**
  - read_csv() for data loading
  - groupby() for summarizing sales and profits
  - plot(), barplot(), lineplot() for creating visualizations
  - corr() to compute correlation matrix

4.4 Interpretation of Analytical Results

Sales Trends Over Time:

- Identified monthly/quarterly peaks and drops in sales.
- Observed seasonality in candy purchases (e.g., spikes near holidays).

· **Comparison Across Regions and Divisions:**

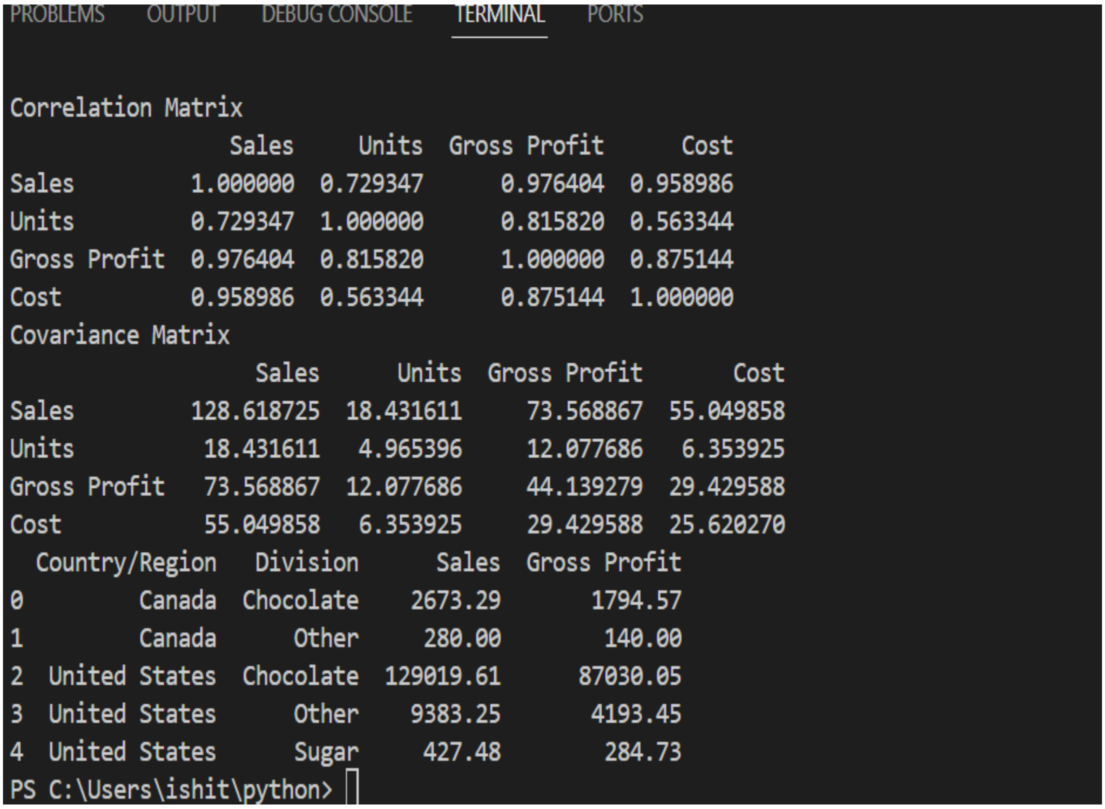
- Certain regions consistently outperformed others in both sales and profits.
- Specific divisions showed higher profit margins due to lower operational costs.

· **Distribution Analysis:**

- Sales and costs showed a right-skewed distribution indicating a few high-performing products.
- Gross profit distribution helped pinpoint the most and least profitable items.

· **Correlation Analysis:**

- Strong positive correlation observed between sales and gross profit.
- Moderate correlation found between cost and profit, suggesting other influencing factors.



```
Correlation Matrix
                 Sales     Units  Gross Profit      Cost
Sales         1.000000  0.729347      0.976404  0.958986
Units         0.729347  1.000000      0.815820  0.563344
Gross Profit  0.976404  0.815820      1.000000  0.875144
Cost          0.958986  0.563344      0.875144  1.000000
Covariance Matrix
                   Sales      Units  Gross Profit       Cost
Sales         128.618725  18.431611     73.568867  55.049858
Units          18.431611   4.965396     12.077686   6.353925
Gross Profit   73.568867  12.077686     44.139279  29.429588
Cost           55.049858   6.353925     29.429588  25.620270
  Country/Region   Division      Sales  Gross Profit
0         Canada  Chocolate    2673.29       1794.57
1         Canada      Other     280.00        140.00
2  United States  Chocolate  129019.61      87030.05
3  United States      Other    9383.25       4193.45
4  United States      Sugar     427.48        284.73
PS C:\Users\ishit\python> 
```

4.5 Advanced Visualization Techniques

Objective4- detect and visualize outliers in the Sales and Gross Profit

Introduction

Outliers can significantly impact data analysis by skewing results and leading to incorrect conclusions. This project aims to detect and visualize outliers in the Sales and Gross Profit columns from the Candy_Sales.csv dataset using Python. By identifying these anomalies, we can ensure better data quality, improve forecasting accuracy, and uncover hidden patterns in candy sales performance.

General Description

The dataset contains records of candy sales including metrics such as product type, sales revenue, cost, and gross profit. The main focus of this analysis is to:

- Identify and handle outliers in the numerical fields Sales and Gross Profit.
- Understand their distribution and impact on overall trends.

- Use visual tools to interpret findings more effectively.

Specific Requirements, Functions and Formulas

To detect outliers, the following statistical methods and functions were used:

1. Interquartile Range (IQR) Method

    - Formula:

$IQR=Q3-Q1$

$\text{Lower Bound}=Q1-1.5\times IQR$

$\text{Upper Bound}=Q3+1.5\times IQR$

- Any data point outside the lower or upper bound is considered an outlier.

2. Functions Used:

    - df.describe() to get Q1 and Q3

    - sns.boxplot() to visualize outliers

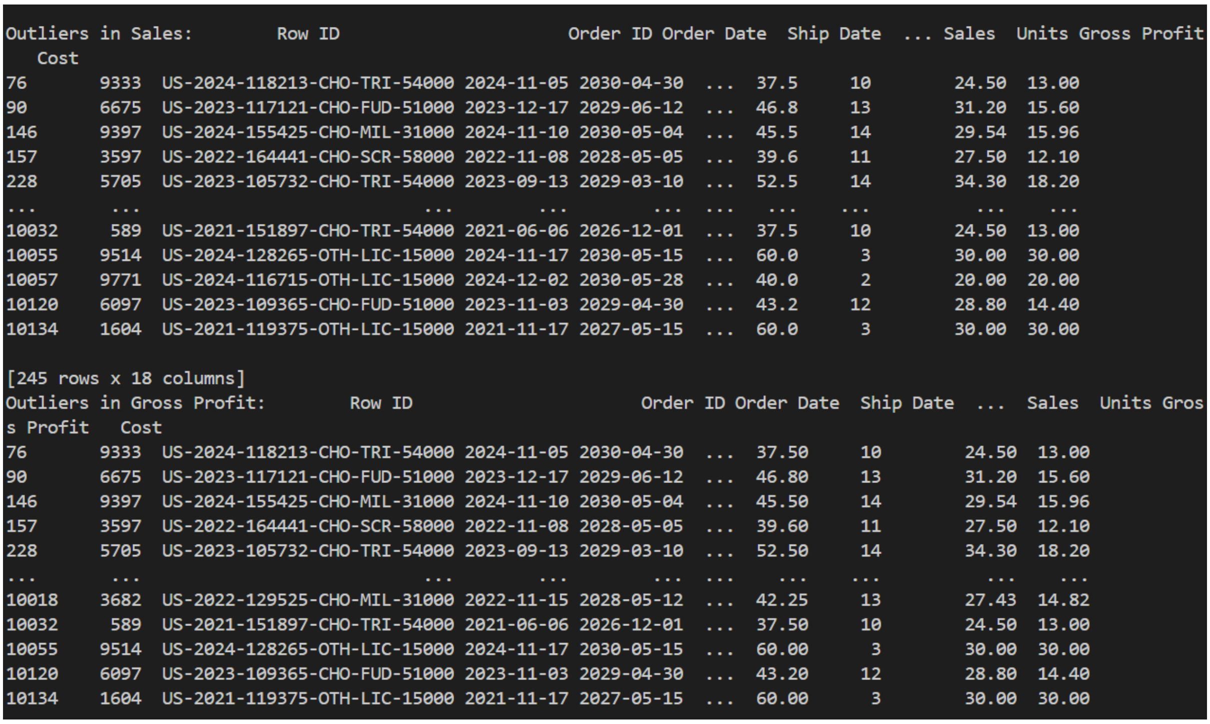
3. **Analysis Result**

· Outliers were detected in both **Sales** and **Gross Profit** columns.

· These outliers represented unusually high or low values compared to the rest of the data.

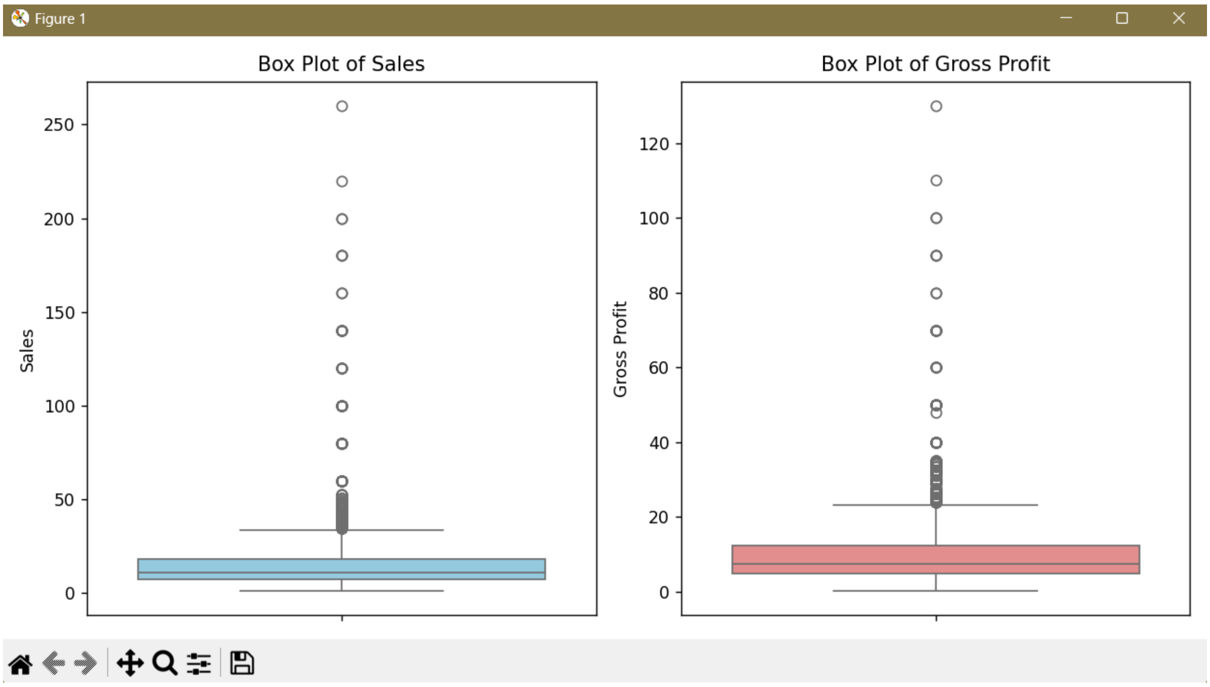
· Removing or handling these outliers significantly improved the accuracy of the distribution analysis.

· Some products had extreme sales spikes, indicating possible promotional periods or data entry issues.



```
Outliers in Sales:        Row ID                   Order ID Order Date  Ship Date  ... Sales  Units Gross Profit
   Cost
76       9333  US-2024-118213-CHO-TRI-54000 2024-11-05 2030-04-30  ...  37.5     10        24.50  13.00
90       6675  US-2023-117121-CHO-FUD-51000 2023-12-17 2029-06-12  ...  46.8     13        31.20  15.60
146      9397  US-2024-155425-CHO-MIL-31000 2024-11-10 2030-05-04  ...  45.5     14        29.54  15.96
157      3597  US-2022-164441-CHO-SCR-58000 2022-11-08 2028-05-05  ...  39.6     11        27.50  12.10
228      5705  US-2023-105732-CHO-TRI-54000 2023-09-13 2029-03-10  ...  52.5     14        34.30  18.20
...       ...                           ...        ...        ...  ...   ...    ...          ...    ...
10032     589  US-2021-151897-CHO-TRI-54000 2021-06-06 2026-12-01  ...  37.5     10        24.50  13.00
10055    9514  US-2024-128265-OTH-LIC-15000 2024-11-17 2030-05-15  ...  60.0      3        30.00  30.00
10057    9771  US-2024-116715-OTH-LIC-15000 2024-12-02 2030-05-28  ...  40.0      2        20.00  20.00
10120    6097  US-2023-109365-CHO-FUD-51000 2023-11-03 2029-04-30  ...  43.2     12        28.80  14.40
10134    1604  US-2021-119375-OTH-LIC-15000 2021-11-17 2027-05-15  ...  60.0      3        30.00  30.00

[245 rows x 18 columns]
Outliers in Gross Profit:        Row ID                   Order ID Order Date  Ship Date  ...  Sales  Units Gros
s Profit   Cost
76       9333  US-2024-118213-CHO-TRI-54000 2024-11-05 2030-04-30  ...  37.50     10         24.50  13.00
90       6675  US-2023-117121-CHO-FUD-51000 2023-12-17 2029-06-12  ...  46.80     13         31.20  15.60
146      9397  US-2024-155425-CHO-MIL-31000 2024-11-10 2030-05-04  ...  45.50     14         29.54  15.96
157      3597  US-2022-164441-CHO-SCR-58000 2022-11-08 2028-05-05  ...  39.60     11         27.50  12.10
228      5705  US-2023-105732-CHO-TRI-54000 2023-09-13 2029-03-10  ...  52.50     14         34.30  18.20
...       ...                           ...        ...        ...  ...    ...    ...           ...    ...
10018    3682  US-2022-129525-CHO-MIL-31000 2022-11-15 2028-05-12  ...  42.25     13         27.43  14.82
10032     589  US-2021-151897-CHO-TRI-54000 2021-06-06 2026-12-01  ...  37.50     10         24.50  13.00
10055    9514  US-2024-128265-OTH-LIC-15000 2024-11-17 2030-05-15  ...  60.00      3         30.00  30.00
10120    6097  US-2023-109365-CHO-FUD-51000 2023-11-03 2029-04-30  ...  43.20     12         28.80  14.40
10134    1604  US-2021-119375-OTH-LIC-15000 2021-11-17 2027-05-15  ...  60.00      3         30.00  30.00
```



4.5 Advanced Visualization Techniques

Obj5 -- Graph Function

Introduction

The primary objective is to identify the total number of orders, total units sold, and determine the state with the highest sales. Through exploratory data analysis (EDA), statistical functions, and data visualization, the project showcases how data-driven decisions can be made effectively.

General Description

The dataset, Candy_Sales.csv, contains transactional information about candy sales across different states. Each row represents an order, with details such as the state, product name, quantity sold, and other relevant fields. The analysis aims to process this dataset, summarize key metrics.

Specific Requirements, Functions and Formulas

To achieve the objectives, the following tools and methods were used:

- Python Libraries:
  - pandas for data loading and manipulation
  - matplotlib and seaborn for visualization
- Functions & Techniques:
  - pd.read_csv() to load the dataset
  - .title to heading a graph
  - .tight_layout to use for layout

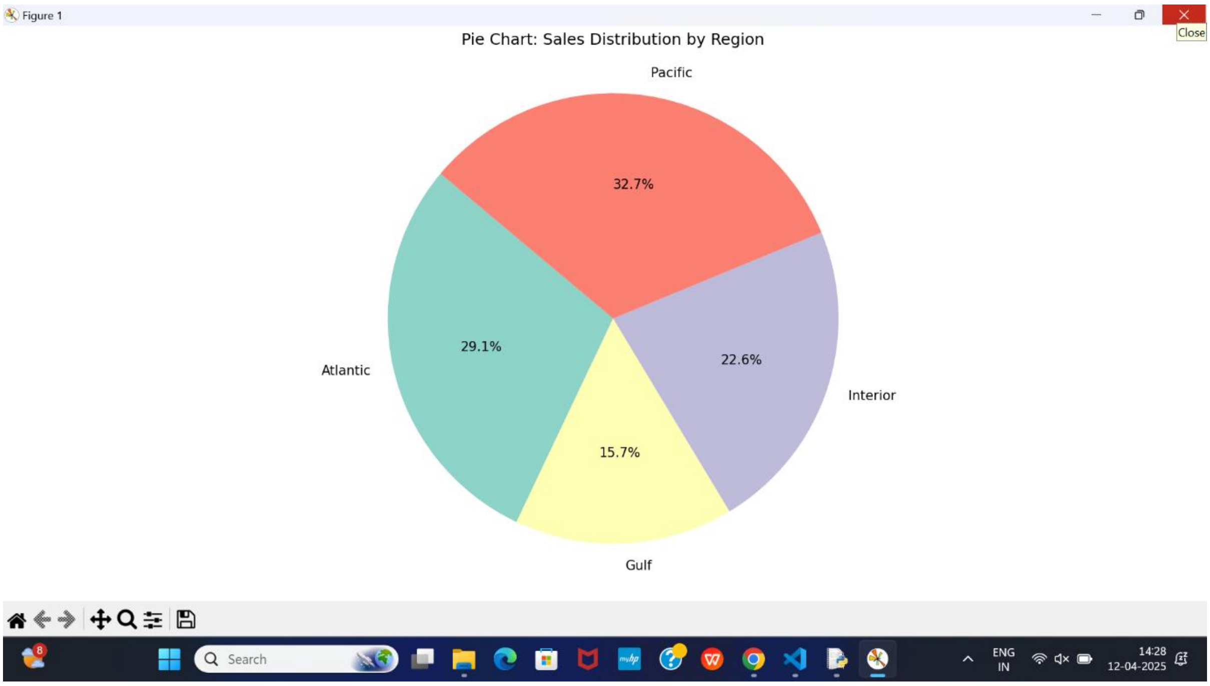


Figure 1
Close
Pie Chart: Sales Distribution by Region
Pacific
32.7%
Atlantic
29.1%
Interior
22.6%
Gulf
15.7%
Search
ENG
IN
14:28
12-04-2025

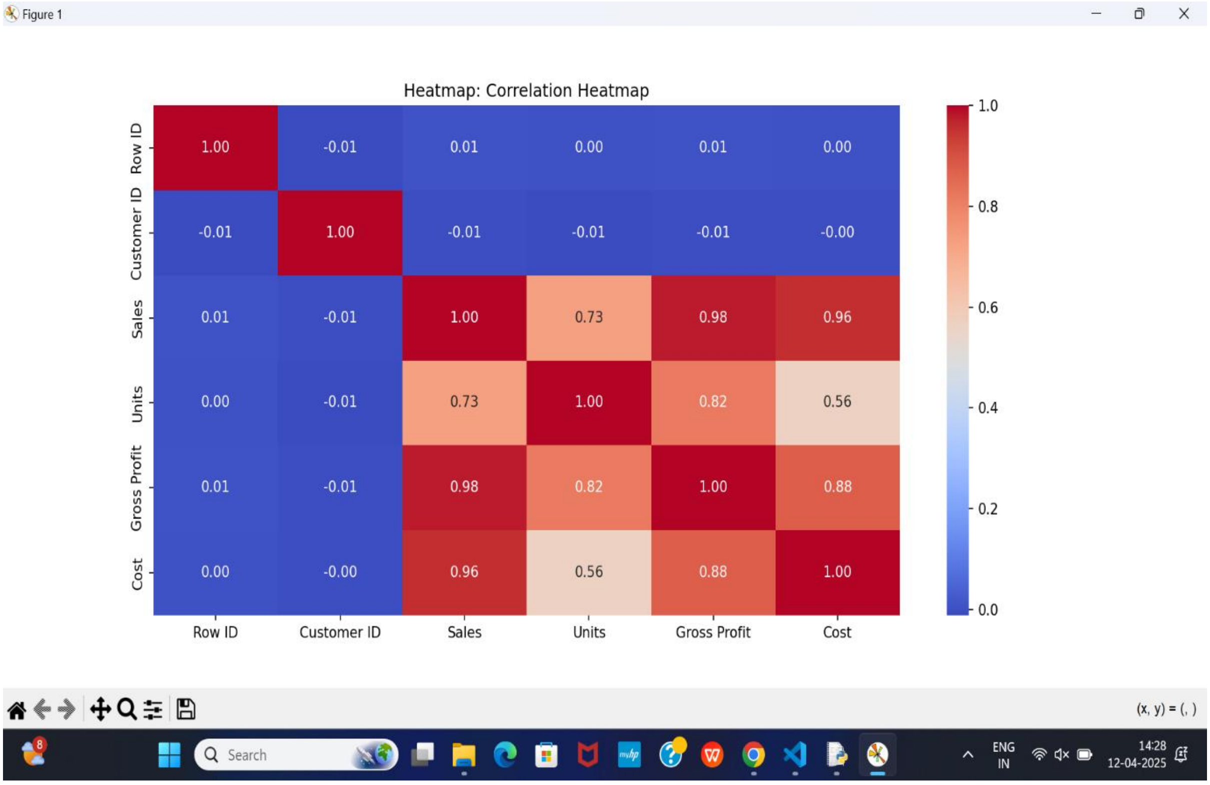
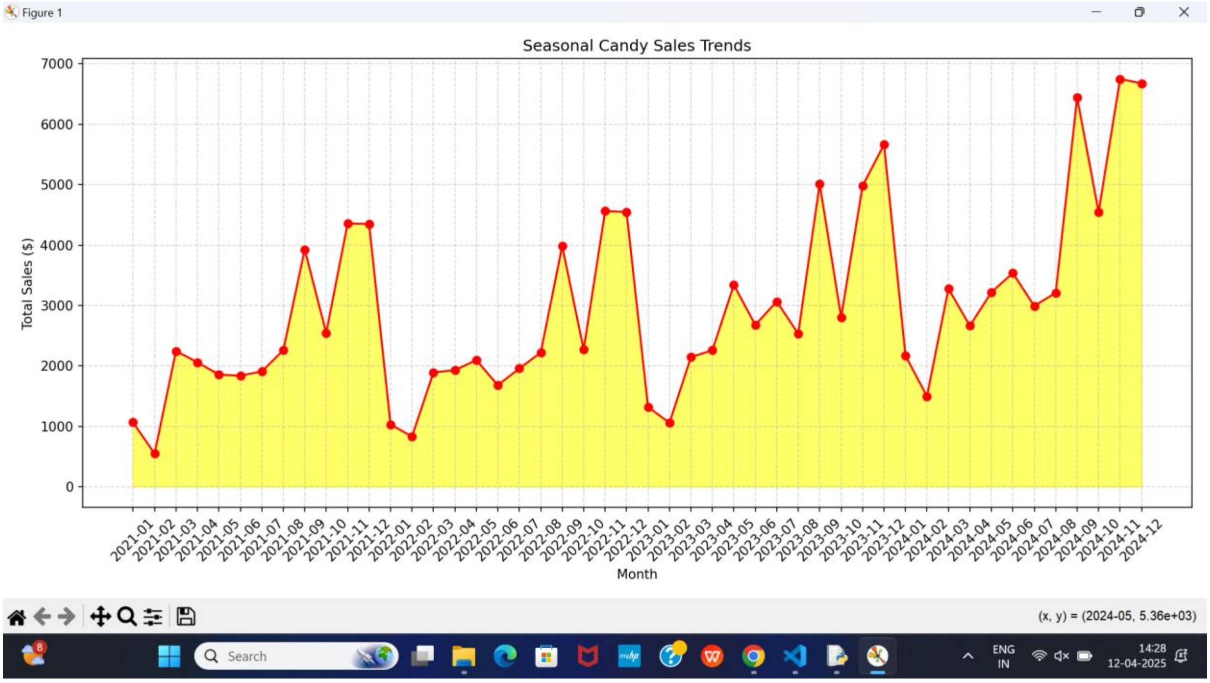
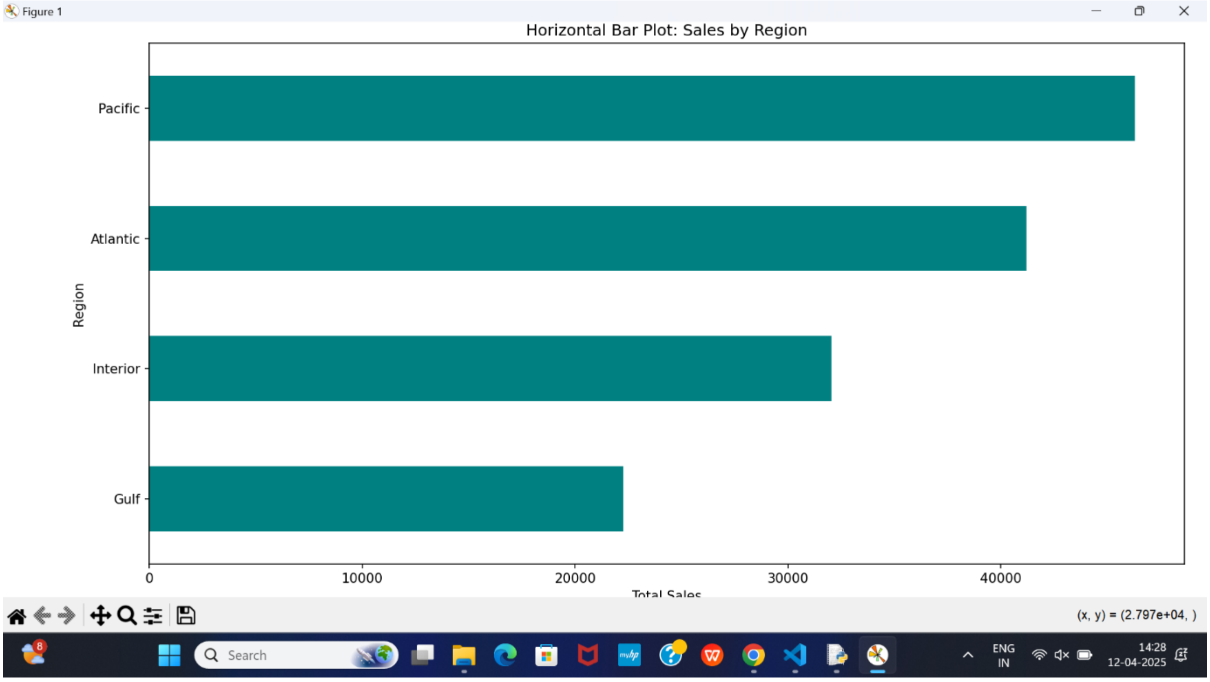


Figure 1
Heatmap: Correlation Heatmap
Row ID 1.00 -0.01 0.01 0.00 0.01 0.00
Customer ID -0.01 1.00 -0.01 -0.01 -0.01 -0.00
Sales 0.01 -0.01 1.00 0.73 0.98 0.96
Units 0.00 -0.01 0.73 1.00 0.82 0.56
Gross Profit 0.01 -0.01 0.98 0.82 1.00 0.88
Cost 0.00 -0.00 0.96 0.56 0.88 1.00
Row ID
Customer ID
Sales
Units
Gross Profit
Cost
1.0
0.8
0.6
0.4
0.2
0.0
(x, y) = (,)
Search
ENG
IN
14:28
12-04-2025



(x, y) = (2024-05, 5.36e+03)



(x, y) = (2.797e+04,)

5. CONCLUSION

The Candy Sales Analysis Project successfully demonstrated how Python can be used to transform raw sales data into meaningful business insights through structured data analysis and visualization techniques. Using the Candy_Sales.csv dataset, the project:

- Cleaned and pre-processed data by handling missing values, correctingformats, and removing duplicates.
- Explored trends across multiple dimensions such as region, product division, shipping mode, and time periods.
- Quantified performance through metrics like total revenue, average order value, total units sold, and state-wise contributions.
- Identified outliers that could skew analysis, offering opportunities to address data entry or business anomalies.
- Visualized insights using various graphs (line plots, heatmaps, scatter plots, pie charts, area charts, box plots), making trends and comparisons easily interpretable.

The project uncovered critical patterns such as:

- Clear seasonality in candy sales, with peaks around specific months.
- Positive correlation between sales and gross profit.
- Top-performing states and divisions driving most of the revenue.

Through these findings, the project proved effective in supporting strategic decisions like inventory planning, pricing adjustments, and regional marketing strategies.

6. SCOPE FOR FUTURE ENHANCEMENTS

1. **Time-Series Forecasting**: Use models such as ARIMA, Facebook Prophet, or LSTM neural networks to predict future profit trends based on historical data.

2. **Interactive Dashboards**: Leverage tools like Streamlet, Tableau, or Power BI to create interactive dashboards for real-time monitoring and stakeholder presentation.

3. **Geo-Spatial Analysis**: Integrate mapping tools (like folium or plotly maps) to visualize regional and city-level performance for geographic insights.

4. **Customer Segmentation with Clustering**: Apply unsupervised machine learning algorithms like K-Means or DBSCAN to classify customers into behavior-driven segments.

5. **Market Basket Analysis**: Utilize association rule mining (Apriori, FP-Growth) to understand purchase combinations for cross-selling strategies.

6. **Product-Level Profitability Models**: Develop models that compare profitability at the SKU or sub-category level to aid pricing and stocking strategies.

7. **Inventory and Delivery Optimization**: Integrate external logistics data to explore ways of reducing shipping delays and optimizing delivery routes.

7. REFERENCES

- Maven Sample Data – Candy sales Dataset,
- https://mavenanalytics.io/data-playground
- McKinney, W. (2017). ***Python for Data Analysis***. O'Reilly Media.

- Seaborn Documentation, https://seaborn.pydata.org/
- Matplotlib Documentation, https://matplotlib.org/stable/index.html
- NumPy Documentation, https://numpy.org/
- Pandas Documentation, https://pandas.pydata.org/
- Python Official Documentation, https://docs.python.org/3/