

**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

**DIWALI SALES ANALYSIS**

**Submitted by**

**NAME:** MUTHA HASITH

**REGISTRATION NO:** 12301431

**PROGRAMME AND SECTION:** K23GN03

**COURSE CODE** **:** CSE375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Mutha.Hasith** bearing Registration no. **12301431** has completed **CSE375** project titled, **“Mrs.Aashima”** 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**

Lovely Professional University

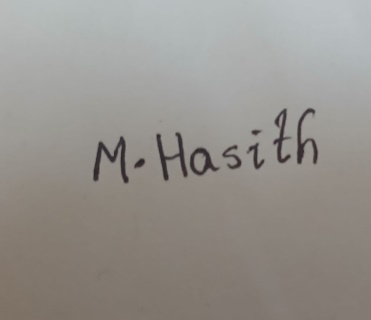
Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Mutha.Hasith student of CSE (Program name) 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.

Date: 11-04-2025 Signature :



Registration No: 12301431 Name of the student: Mutha.Hasith

**Table of Contents**

|  |  |
| --- | --- |
| **TITLE** | **PAGE NO** |
| **1.Introduction** | **5** |
| **2. What is EDA?** | **5** |
| **3. Why EDA Is Important For Diwali Sales** | **5** |
| **4. Source of Dataset** | **6** |
| **5. Step-by-Step EDA Process** | **6** |
| **6. Dataset Preprocessing** | **7** |
| **7. Univariate Analysis** | **7** |
| **8. Bivariate Analysis** | **7** |
| **9. Multivariate Analysis** | **8** |
| **10. Outlier Detection** | **8** |
| **11. Correlation Analysis** | **8** |
| **12. Analysis on Dataset** | **8 - 18** |
| **12.1. Objective 1: Sales by Age Group** | **8 – 9** |
| **12.2. Objective 2: Top States by Orders** | **9 – 10** |
| **12.3. Objective 3: Purchase Pattern by Age-Gender Groups** | **10 – 12** |
| **12.4. Objective 4: State-wise Sales Trends** | **12 – 13** |
| **12.5. Objective 5: Average Order Value by Age & Gender** | **13 – 14** |
| **12.6. Objective 6: Relationship Between Amount and Orders** | **15 – 17** |
| **12.7. Objective 7: Outlier Detection** | **17 - 18** |
| **13. Conclusion** | **19** |
| **14. Future Scope** | **20** |
| **15. References** | **20** |

**1. Introduction**

The Diwali festival is one of the largest and most celebrated festivals in India. During this period, consumer buying patterns undergo significant changes due to festive offers, gift-giving traditions, and increased purchasing power. This report presents an in-depth Exploratory Data Analysis (EDA) of Diwali Sales data aimed at uncovering hidden trends and extracting insights that can support data-driven decisions.

The analysis focuses on key customer attributes like age, gender, and location to understand their influence on purchasing behavior. By understanding these patterns, companies can create targeted marketing campaigns, stock popular items, and forecast demand more accurately.

EDA is not just a technical step; it is a journey through the data that brings us closer to actionable insights. It allows businesses to make decisions not based on assumptions, but grounded in actual behavior.

**2. What is EDA?**

Exploratory Data Analysis (EDA) is a crucial first step in the data analysis process. It involves examining datasets to summarize their main characteristics, often with visual methods. EDA is used to:

* Get a sense of the structure, patterns, and relationships in data
* Identify anomalies, missing values, and outliers
* Generate hypotheses and guide further data modeling
* Understand the distribution of variables

**Techniques used in EDA:**

* Descriptive Statistics: Mean, median, mode, range, standard deviation
* Data Visualization: Histograms, bar plots, scatter plots, box plots
* Data Cleaning: Handling null values, duplicates, formatting
* Feature Engineering: Creating new columns, segmenting categories
* Correlation & Relationships: Using statistical tools to assess interaction between variables

**3. Why EDA is Important for Diwali Sales**

Understanding Diwali sales through EDA is essential because it helps uncover patterns in how customers behave during the festive season. Here’s why it is important:

* Demand Forecasting: EDA helps predict what products are likely to be in demand.
* Targeted Marketing: By knowing which age group or gender spends more, marketing strategies can be tailored.
* Revenue Optimization: Recognizing high-value states or segments allows more focused sales efforts.
* Stock Planning: EDA helps determine where to allocate inventory efficiently.

This analytical approach ensures that festive campaigns are not based on assumptions but are backed by insights derived from real data.

**4. Source of Dataset**

The dataset was collected from a CSV file that records sales data from the Diwali festive season.

* File Name: Diwali Sales Data.csv
* Format: CSV (Comma-Separated Values)
* Encoding: Latin1
* Fields in the Dataset:
  + Gender: Male or Female
  + Age: Age of customer
  + State: State where the purchase was made
  + Amount: Total purchase value
  + Orders: Number of items ordered

**5. Step-by-Step EDA Process**

**EDA in this report follows these detailed steps:**

1. Import Libraries: Pandas, NumPy, Matplotlib, Seaborn
2. Load Dataset: Read CSV file using Pandas
3. Initial Data Inspection: Check data types, shape, head, and summary
4. Data Cleaning:
   * Strip spaces from column headers
   * Convert relevant columns to numeric
   * Remove missing/null values
5. Feature Engineering:
   * Create custom Age\_Group brackets
   * Combine columns for analysis like Age\_Gender
6. Univariate Analysis: Analyze each variable on its own
7. Bivariate Analysis: Study relationships between two variables
8. Multivariate Analysis: Explore interactions among three or more variables
9. Outlier Detection: Identify extreme values using IQR and box plots
10. Correlation Study: Use heatmaps to understand variable relationships

**6. Dataset Preprocessing**

**Preprocessing steps include:**

* Standardization of Column Names
* Conversion of Datatypes: Amount and Orders to numeric
* Handling Missing Values: Dropped records with missing Amount or Orders
* Creation of Age\_Group: Segmented into 18–25, 26–35, etc.

This ensures consistency, reduces noise, and prepares the data for analysis.

**7. Univariate Analysis**

We analyzed one variable at a time to understand distribution:

* Gender: Male vs Female proportions
* Age: Most active buyer age group
* Amount: Range of purchases
* Orders: Quantity trends

Graphs: Pie charts, bar plots, histograms

**8. Bivariate Analysis**

We studied interaction between two variables:

* Age vs Amount
* Gender vs Amount
* State vs Orders

This reveals how two features influence each other. For example, higher spending in certain states or age groups.

**9. Multivariate Analysis**

Here we examined Age + Gender + State:

* Who buys more?
* Which combination is most profitable?

Visualizations used: Heatmaps, stacked bar graphs, group plots

**10. Outlier Detection**

We used:

* IQR Method: Calculate Q1 and Q3, find outliers
* Boxplots: Visualized anomalies

Outliers were mostly large purchases—likely business clients

**11. Correlation Analysis**

We created:

* Correlation Matrix using .corr()
* Heatmap to visualize
* Pairplot to inspect pairwise relationships

Found strong correlation between Orders and Amount

**12. Analysis on Dataset**

**12.1 Objective 1: Sales by Age Group**

* **General Description:**

We analysed total sales across different age brackets to identify the most active buyer groups. This objective helps businesses understand which customer segments contribute the most to revenue during the festive season. By examining how different age groups behave in terms of spending, companies can tailor their promotions, offers, and product placements more effectively to maximize conversions.

* **Specific Requirements:**

To fulfil this objective, we grouped customers based on predefined age brackets such as 18–25, 26–35, 36–45, 46–50, and 51–55. For each group, we calculated the total Amount spent during the Diwali sales period.

* **Analysis Results:** The age group **26-35** emerged as the top buyer segment.
* **Python Code Used:**
* **Visualization:**

def plot\_sales\_by\_age(df):

age\_order = ['<18', '18-25', '26-35', '36-45', '46-55', '56-65', '65+']

plt.figure(figsize=(8, 5))

sns.barplot(x='Age\_Group', y='Amount', data=df, estimator=sum, order=age\_order, hue='Age\_Group', palette='viridis', legend=False, errorbar=None)

plt.title('Total Sales by Age Group')

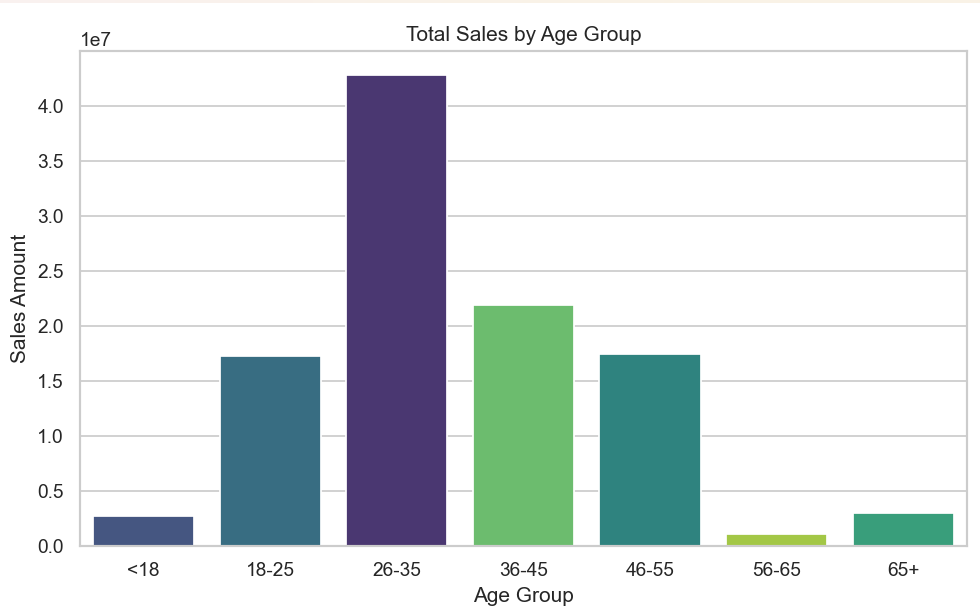
plt.xlabel('Age Group')

plt.ylabel('Sales Amount')

plt.tight\_layout()

plt.show()

plot\_sales\_by\_age(df)



**12.2 Objective 2: Top States by Orders**

* **General Description:**

We investigated state-wise order volume to identify high-performing regions in terms of sales activity. This analysis helps businesses understand geographical demand during the Diwali season and guides decisions related to inventory allocation**,** regional promotions, and delivery optimization.

* **Specific Requirements:**

The data was grouped by State, and total Orders were calculated. Then, the states were sorted in descending order to extract the top 10 states based on order count.

* **Analysis Results:** States like **Uttar Pradesh**, **Maharashtra**, and **Karnataka** recorded the highest number of orders.
* **Python Code Used:**

def plot\_top\_states\_by\_orders(df):

top\_states = df.groupby('State')['Orders'].sum().sort\_values(ascending=False).head(10).reset\_index()

plt.figure(figsize=(10, 5))

sns.barplot(x='Orders', y='State', data=top\_states, hue='State', palette='Blues\_r', legend=False)

plt.title('Top 10 States by Orders')

plt.xlabel('Orders')

plt.ylabel('State')

plt.tight\_layout()

plt.show()

plot\_top\_states\_by\_orders(df)

* **Visualization:**

A graph with blue bars

AI-generated content may be incorrect.

**12.3 Objective 3: Purchase Pattern by Age-Gender Groups**

* **General Description:**

To understand the impact of both age and gender on sales, we created a combined category called Age\_Gender. This helped identify the most profitable customer segments by analyzing who spends the most during the Diwali season.

* **Specific Requirements:**
* Create a new column combining Age Group and Gender (e.g., 18-25\_Male)
* Group by this new column and sum the total Amount spent
* Rank segments to identify top spender
* **Analysis Results:** Younger males and females between 18-35 were the leading contributors to revenue.
* **Python Code Used:**

def plot\_simple\_age\_gender\_pie(df):

df['Age\_Gender'] = df['Age'].astype(str) + ' | ' + df['Gender'].astype(str)

age\_gender\_sales = df.groupby('Age\_Gender')['Amount'].sum().sort\_values(ascending=False)

top6 = age\_gender\_sales.head(6)

colors = sns.color\_palette('Set3')

plt.figure(figsize=(7, 7))

plt.pie(top6.values, labels=top6.index, autopct='%1.1f%%', startangle=140, colors=colors)

plt.title('Top 6 Age-Gender Groups by Purchase Amount')

plt.tight\_layout()

plt.show()

plot\_simple\_age\_gender\_pie(df)

* **Visualization:**

A pie chart with numbers and a few different colored circles

AI-generated content may be incorrect.

**12.4 Objective 4: State-wise Sales Trends**

* **General Description:**

Analyzing state-wise revenue provides valuable insights into geographical performance during the Diwali season. This analysis helps businesses optimize regional marketing, tailor logistics, and allocate inventory efficiently

* .**Specific Requirements:**
* Group the dataset by State
* Calculate total Amount spent in each state
* Visualize the top contributing states using bar and line plots
* **Analysis Results:** Northern and western states exhibited higher sales amounts.
* **Python Code Used:**
* **Visualization:**

def plot\_statewise\_sales\_line(df):

state\_sales = df.groupby('State')['Amount'].sum().sort\_values(ascending=False)

plt.figure(figsize=(10, 6))

plt.plot(state\_sales.values, state\_sales.index, marker='o', color='darkorange')

plt.title('State-wise Sales (Amount on X-axis)')

plt.xlabel('Total Sales Amount')

plt.ylabel('State')

plt.grid(True)

plt.tight\_layout()

plt.show()

plot\_statewise\_sales\_line(df)

A graph with orange lines

AI-generated content may be incorrect.

**12.5 Objective 5: Average Order Value by Age & Gender**

* **General Description:**
* Evaluating the Average Order Value (AOV) helps in understanding customer spending behaviour. By analysing AOV across age and gender, we can uncover who is making high-value purchases and tailor strategies like premium offers or discounts accordingly.
* **Specific Requirements:**
* Calculate AOV = Total Amount / Total Orders
* Group by Age Group and Gender
* Visualize patterns across different demographics
* **Analysis Results:** Older age groups tend to place higher-value orders. Gender also shows mild variation in spending.
* **Python Code Used:**

def plot\_avg\_order\_value(df):

df['Average\_Order\_Value'] = df['Amount'] / df['Orders']

df.dropna(subset=['Age', 'Gender', 'Average\_Order\_Value'], inplace=True)

df['Age'] = df['Age'].astype(str).str.strip()

df['Gender'] = df['Gender'].astype(str).str.strip()

grouped = df.groupby(['Age', 'Gender'])['Average\_Order\_Value'].mean().reset\_index()

grouped['Age'] = grouped['Age'].astype(int)

plt.figure(figsize=(10, 6))

sns.scatterplot(data=grouped, x='Age', y='Average\_Order\_Value', hue='Gender', s=100, palette='Set1')

plt.title('Average Order Value by Age Group and Gender')

plt.xlabel('Age')

plt.ylabel('Average Order Value (₹)')

plt.grid(True)

plt.tight\_layout()

plt.show()

plot\_avg\_order\_value(df)

* **Visualization:**

A graph with red and blue dots

AI-generated content may be incorrect.

**12.6 Objective 6: Relationship Between Amount and Orders**

* **General Description:**

This analysis explores whether there's a direct relationship between the number of orders and total spending. In simpler terms: do more orders result in more money spent?

* **Specific Requirements:**
* Evaluate the correlation between Orders and Amount
* Use Correlation Heatmap and Pairplot to visually interpret the relationship
* **Analysis Results:** A strong positive correlation was found between Orders and Amount.
* **Python Code Used:**

def plot\_pairplot(df):

sns.set\_style('whitegrid')

pair = sns.pairplot(df[['Amount', 'Orders']], height=3)

pair.fig.suptitle('Pairplot of Amount & Orders', fontsize=14, weight='bold', y=1.0)

plt.show()

plot\_pairplot(df)

def plot\_basic\_heatmap(df):

corr = df[['Amount', 'Orders']].corr()

plt.figure(figsize=(8, 5))

plt.imshow(corr, cmap='coolwarm')

plt.colorbar()

plt.xticks([0, 1], ['Amount', 'Orders'])

plt.yticks([0, 1], ['Amount', 'Orders'])

plt.title('Correlation Heatmap')

plt.tight\_layout()

plt.show()

plot\_basic\_heatmap(df)

* **Visualization:** A diagram of a heatmap

  AI-generated content may be incorrect.

A graph of a number of numbers

AI-generated content may be incorrect.

**12.7 Objective 7: Outlier Detection**

* **General Description:**

This analysis focuses on identifying unusual sales values that fall significantly outside the typical purchase range. These outliers could be either extremely high or low purchase amounts, possibly indicating bulk orders, system anomalies, or high-value customers.

* **Specific Requirements:**
* Use the IQR (Interquartile Range) method to detect outliers in the Amount column
* Visualize outliers using a boxplot
* **Analysis Results:** Some extreme purchase amounts were identified as outliers, possibly due to bulk or unusual transactions.
* **Python Code Used:**

def plot\_outliers\_boxplot(df):

q1 = df['Amount'].quantile(0.25)

q3 = df['Amount'].quantile(0.75)

iqr = q3 - q1

lower\_bound = q1 - 1.5 \* iqr

upper\_bound = q3 + 1.5 \* iqr

outliers = df[(df['Amount'] < lower\_bound) | (df['Amount'] > upper\_bound)]

plt.figure(figsize=(6, 4))

sns.boxplot(x=outliers['Amount'], color='tomato')

plt.title('Boxplot of Outlier Amounts')

plt.xlabel('Amount')

plt.tight\_layout()

plt.show()

plot\_outliers\_boxplot(df)

* **Visualization:**

A graph of a number of numbers

AI-generated content may be incorrect.

**13. Conclusion**

The Diwali Sales Analysis project successfully showcased how Exploratory Data Analysis (EDA) can uncover meaningful patterns in festive consumer behavior. By methodically cleaning, transforming, and visualizing the dataset, we were able to derive critical business insights that can guide future decision-making.

**Key Findings:**

* The 26–35 age group emerged as the most active and valuable buyer segment, contributing the highest revenue. This suggests that marketing campaigns should be designed with this demographic in mind.
* States like Uttar Pradesh, Maharashtra, and Karnataka consistently recorded higher order volumes, making them strong markets for festive sales. These regions should be prioritized in inventory distribution and promotional efforts.
* Young adults between 18–35 years old were responsible for the majority of purchases, regardless of gender. This highlights the importance of crafting offers that appeal to younger, digitally active consumers.
* There was a strong positive correlation between the number of orders and the purchase amount, confirming that customers who buy more also tend to spend more. This insight supports bundling strategies, combo deals, and personalized upselling.
* Outlier detection revealed high-value transactions that could be attributed to bulk buyers or B2B customers. These customers represent opportunities for loyalty programs, exclusive deals, and corporate sales initiatives.
* Impact of EDA: By applying structured EDA techniques—like univariate, bivariate, multivariate analysis, and outlier detection—we turned raw data into insights. These findings are not just academic; they can help real businesses make strategic decisions during high-sales periods like Diwali.

**14. Future Scope**

* Incorporate **time-based trends** to understand purchase behaviour by day/week.
* Use **machine learning models** to predict future sales.
* Conduct **product-category level analysis** for deeper insights.
* Integrate customer feedback or reviews if available.

**15. References**

* Diwali Sales Data (CSV File) for git hub:<https://github.com/rishabhnmishra/Python_Diwali_Sales_Analysis/blob/main/Diwali%20Sales%20Data.csv>
* Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
* Statistical Techniques: Correlation, Z-score, IQR method