**1. Introduction and problem statement**

The objective of this analysis is to gain insights from the Sales data of ABC Company over the last three years (2017, 2018 & 2019). The dataset includes information on Sales transactions, Targets set for Salespersons, Customer details, Product information, and dates. Through this analysis, we aim to identify the sales trends, evaluate salesperson performance, understand customer behaviour, and assess the overall business performance.

**2. About the Data**

The dataset comprises several tables, including sales data for 2017, 2018, and 2019, a targets table with monthly targets for salespersons, customer details, product group information, and a date table.

Some more information on the different tables:

**Sales Data (2017, 2018, 2019)**:

This table contains detailed information about sales transactions over the years 2017, 2018, and 2019. It includes data such as invoice numbers, customer IDs, salesperson IDs, product IDs, quantities sold, unit prices, and net weights for each transaction.

**Targets Table:**

The targets table stores monthly sales targets assigned to each salesperson. It includes columns for salesperson IDs, months, and corresponding target values. This table helps track the performance of salespersons against their set targets.

**Customer Details:**

This table provides comprehensive customer information, including the unique IDs, company names, cities, and states. It serves as a reference for analyzing customer demographics and geographic distribution.

**Product Information:**

The product information table contains details about various products, including their IDs, names, and groupings. It helps categorize products and analyze sales performance across different product categories.

**3. Initial Findings**

* The sales data spans three years, with significant variations observed in sales volumes and revenues across different periods.
* No missing values were identified in the dataset
* Outliers were present in the sales data, but the variance was because of the sales transactions based on unit price and quantities sold.
* The targets table had Months as the column names and Salesperson IDs as the row names. The Months column will pose a problem when connecting it with the dates table during the data modelling phase.
* Individual product names are not available, only Product group names are.

**4. Data Distribution**

**Python code:**

**Histogram:**

num\_cols = ['Qty Itens', 'Unit Price', 'Net Weight']

plt.figure(figsize=(12, 6))

for i, col in enumerate(s17\_num\_cols, 1):

    plt.subplot(2, 2, i)

    sns.histplot(data=s17, x=col)

    plt.title(f'Histogram of {col}')

    plt.xlabel(col)

    plt.ylabel('Frequency')

    plt.xlim(s17[col].quantile(0.05), s17[col].quantile(0.5))

plt.tight\_layout()

plt.show()

**Boxplot:**

for col in s17\_num\_cols:

    plt.figure(figsize=(8, 6))

    sns.boxplot(data=s17[col])

    q1 = s17[col].quantile(0.25)

    q2 = s17[col].median()

    q3 = s17[col].quantile(0.75)

    plt.ylim(s17[col].quantile(0.05), s17[col].quantile(0.95))

    plt.title(f'Box plot of {col}')

    plt.xlabel(col)

    plt.ylabel('Values')

    plt.show()

**Outlier counts:**

outlier\_counts = {}

for col in s17\_num\_cols:

    q1 = s17[col].quantile(0.25)

    q3 = s17[col].quantile(0.75)

    iqr = q3 - q1

    lb\_iqr = q1 - 1.5 \* iqr

    ub\_iqr = q3 + 1.5 \* iqr

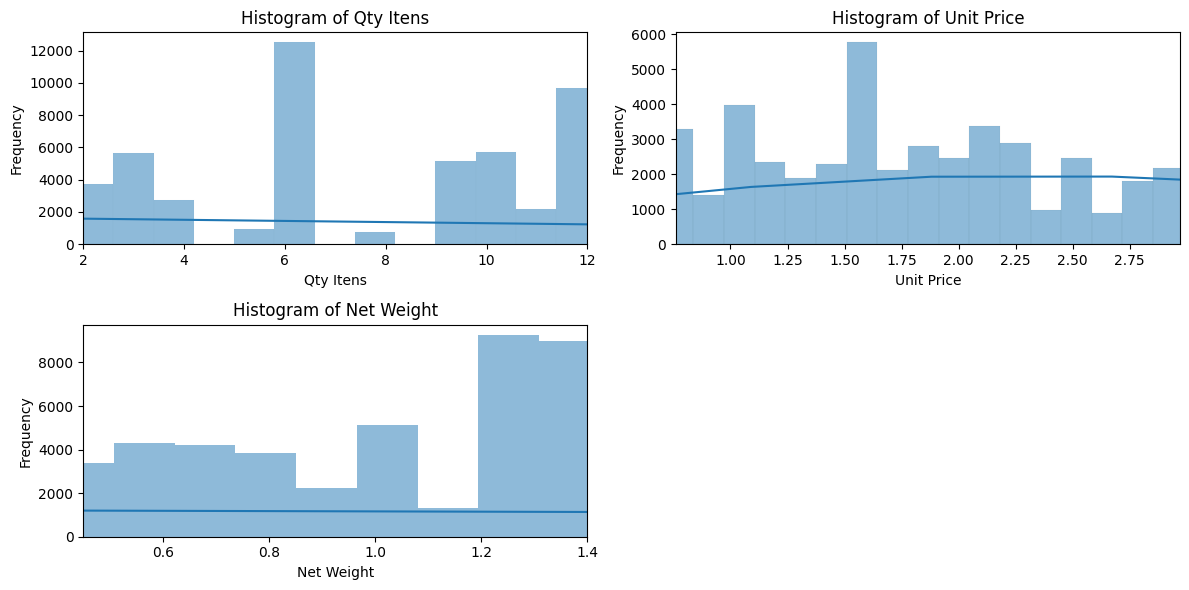
    outliers = s17[(s17[col] < lb\_iqr) | (s17[col] > ub\_iqr)]

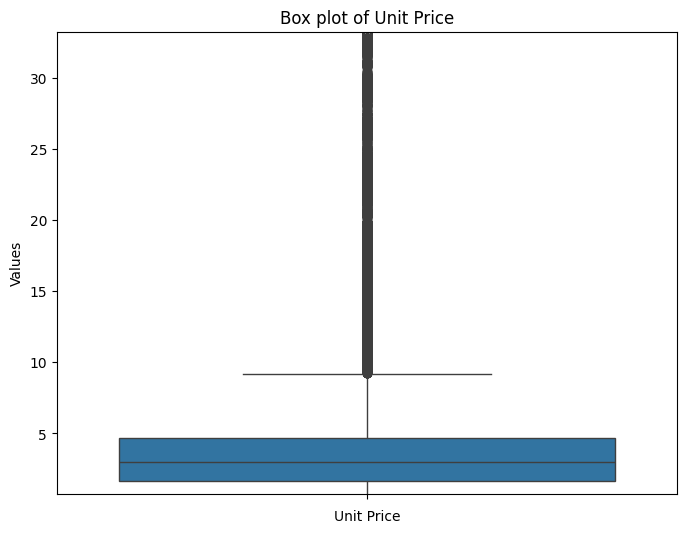
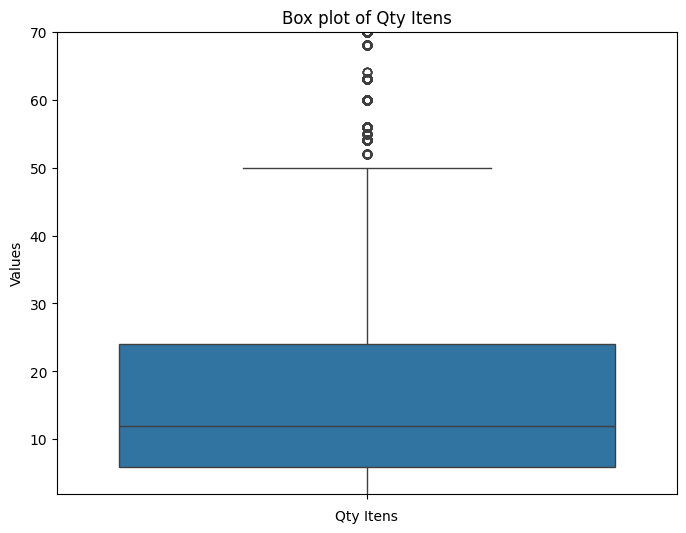
    outlier\_counts[col] = len(outliers)

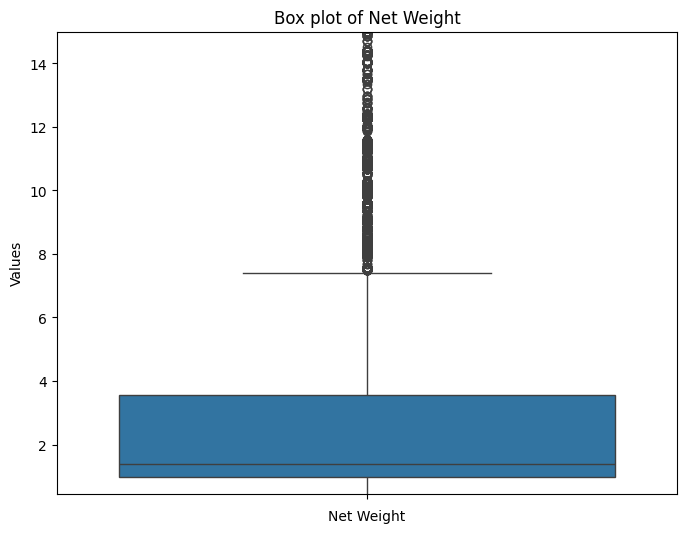
for col, count in outlier\_counts.items():

    print(f"Outlier count in '{col}': {count}")

**2017 Data:**







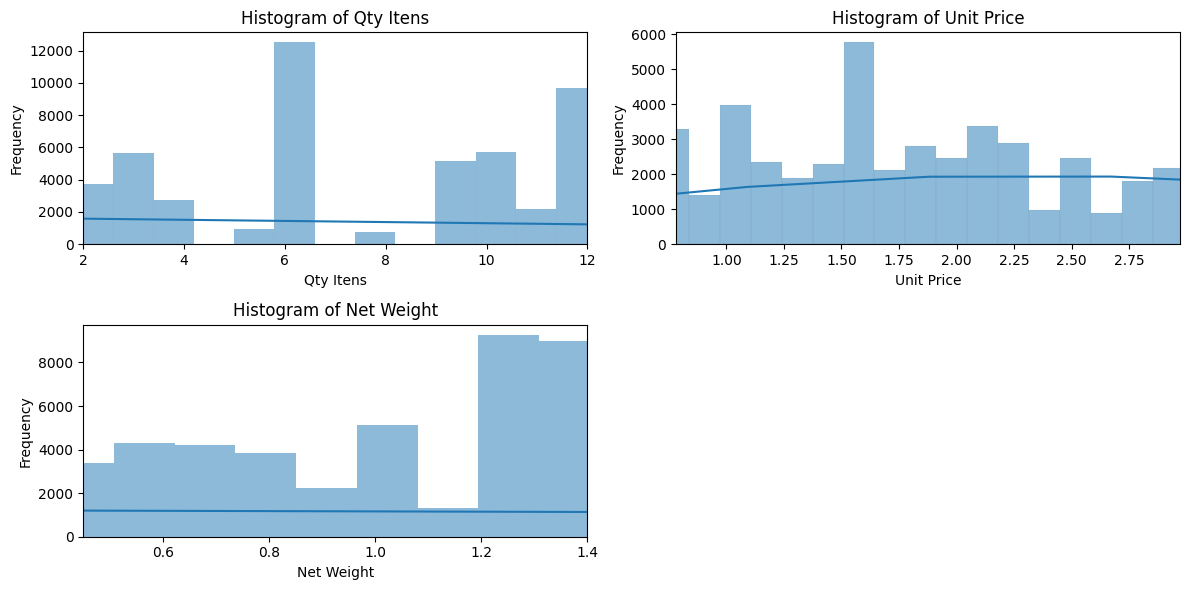
**Outlier Count:**

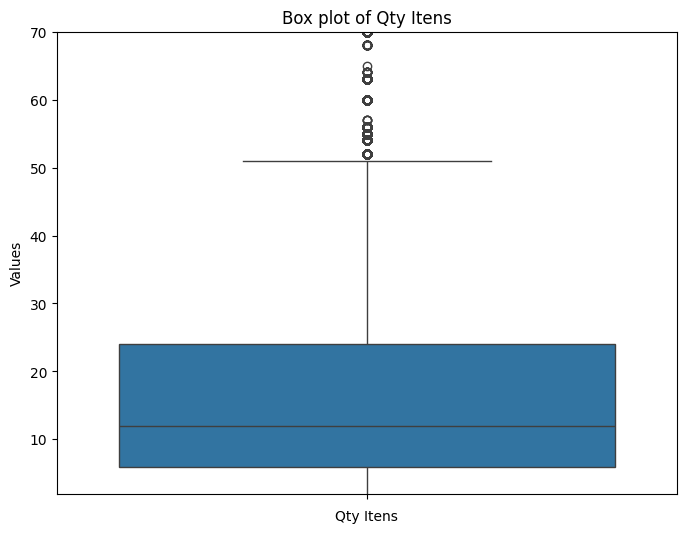
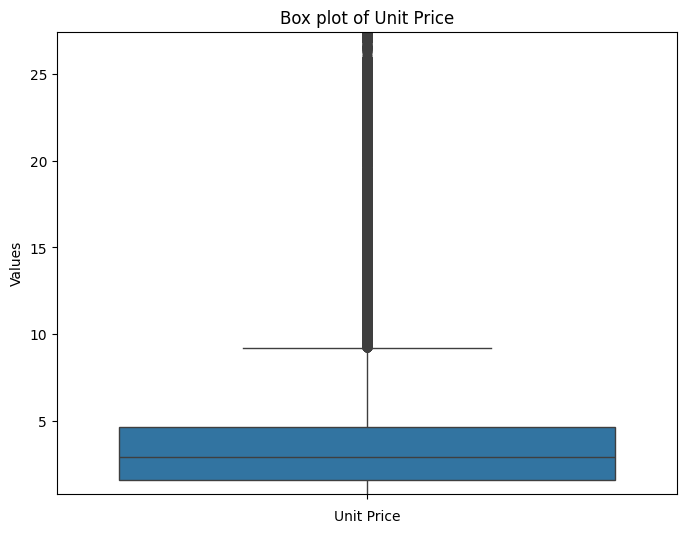
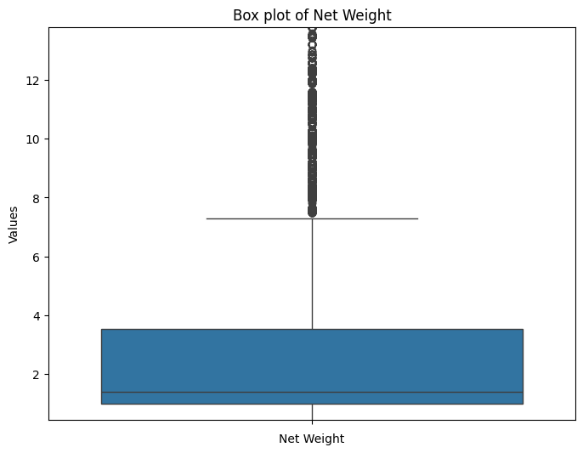
Outlier count in 'Qty Itens': 7179

Outlier count in 'Unit Price': 13411

Outlier count in 'Net Weight': 11155

**Sales 2018 Data:**



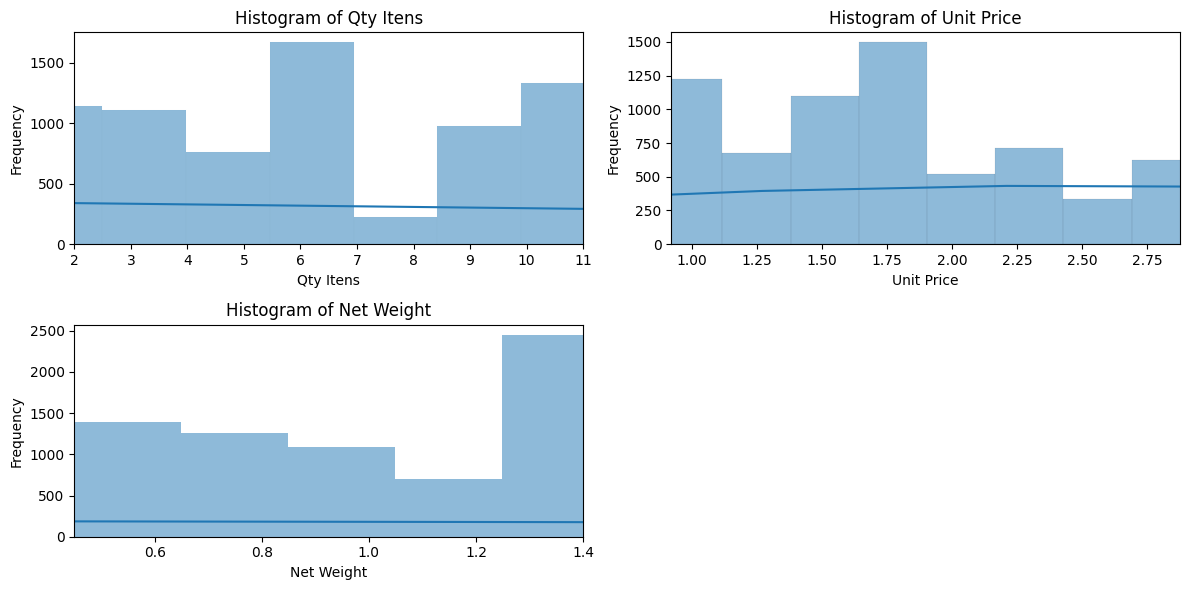
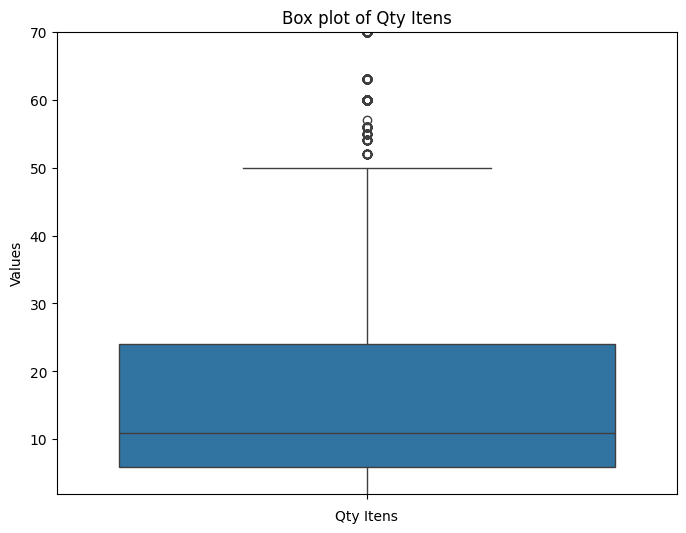
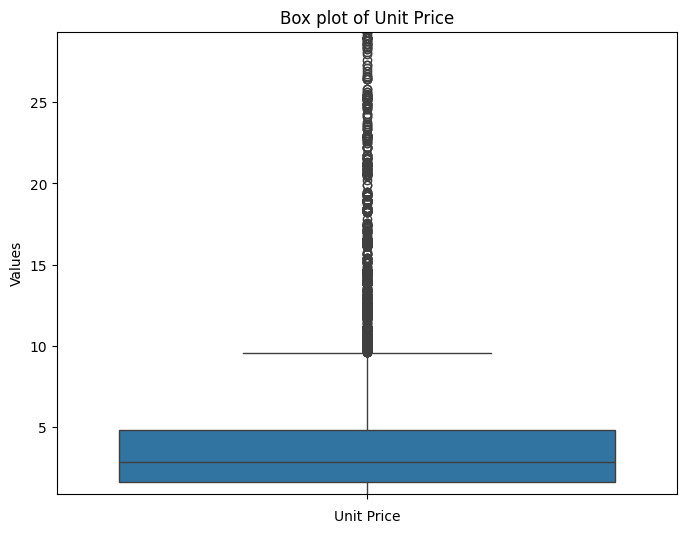
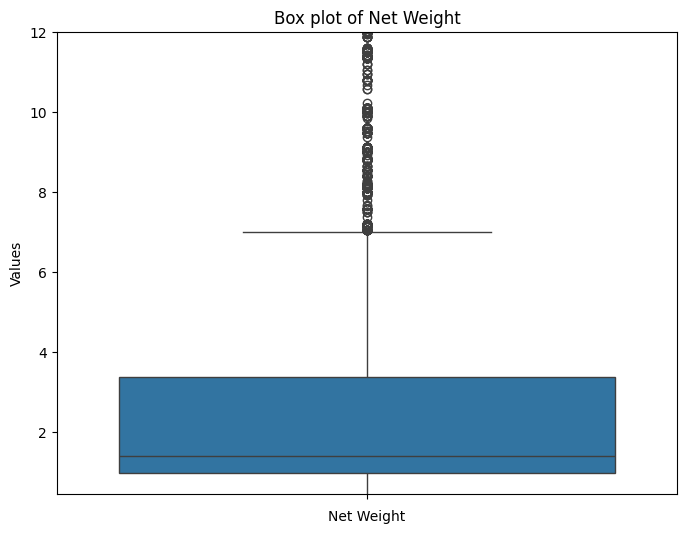
**Outlier count:**

Outlier count in 'Qty Itens': 12105

Outlier count in 'Unit Price': 18325

Outlier count in 'Net Weight': 17837

**Sales 2019 Data:**

**5. Analysis Charts with Textual Findings**

**Bivariate Analysis:**

plt.figure(figsize=(8, 6))

plt.scatter(s17['Qty Itens'], s17['Unit Price'])

plt.xlabel('Qty Itens')

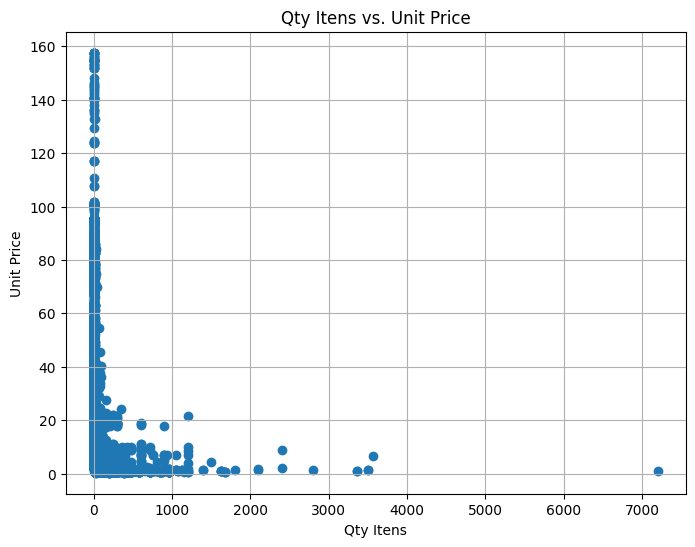
plt.ylabel('Unit Price')

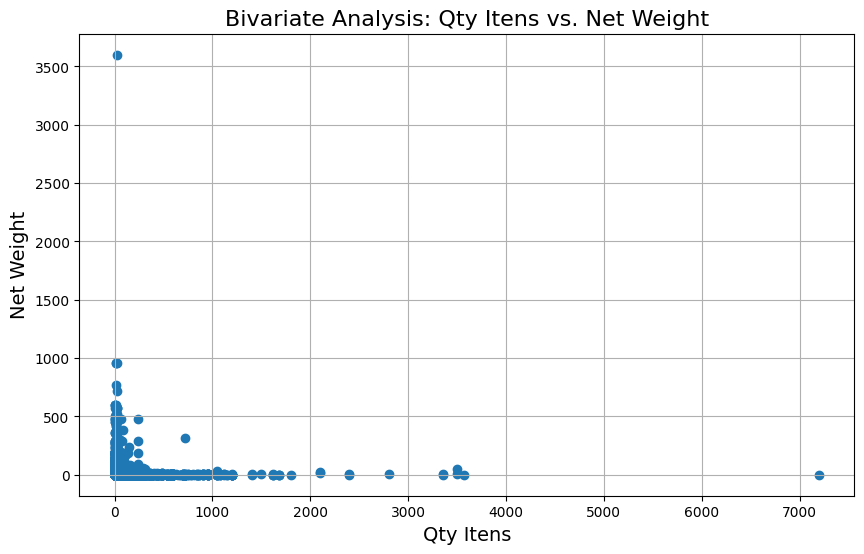
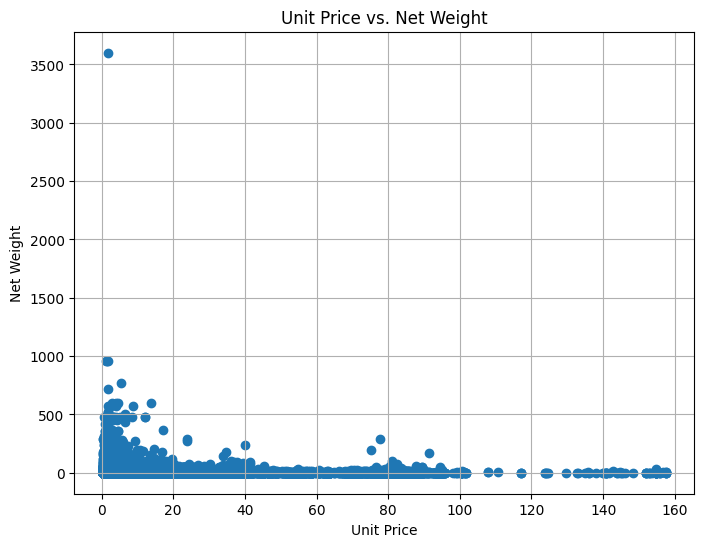
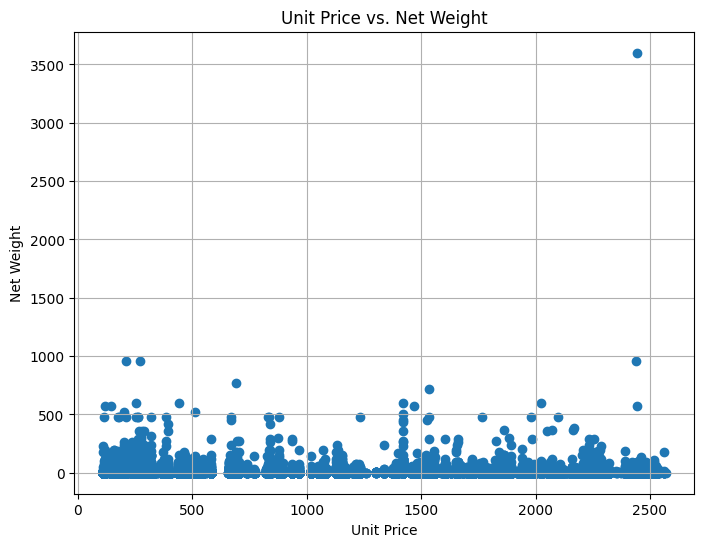
plt.title('Qty Itens vs. Unit Price')

plt.grid(True)

plt.show()

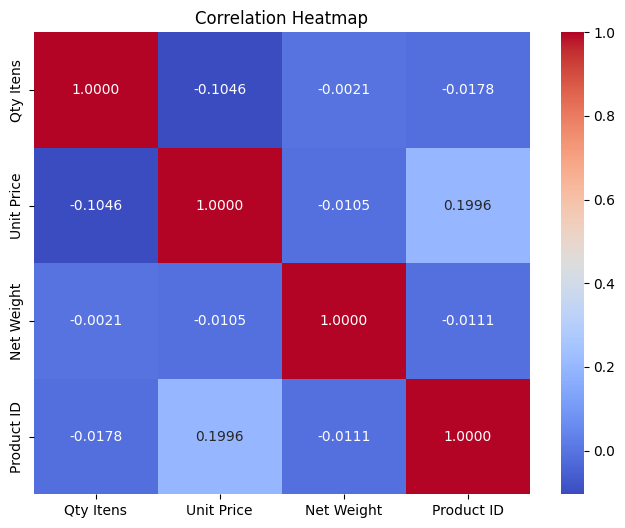
**2017 Data**:



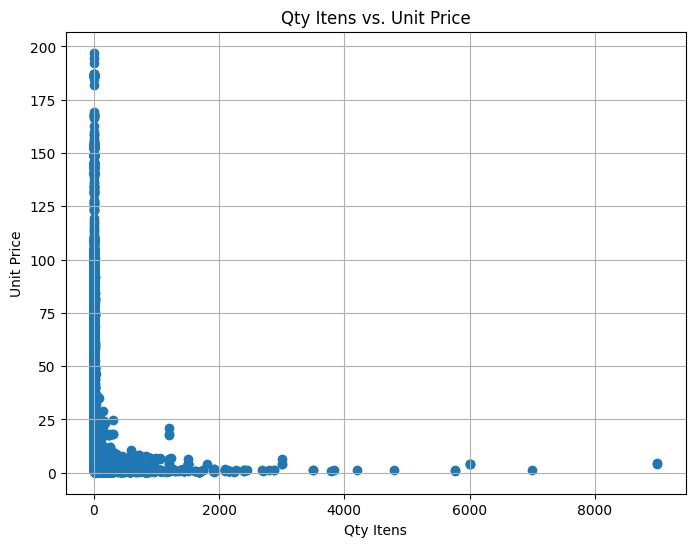
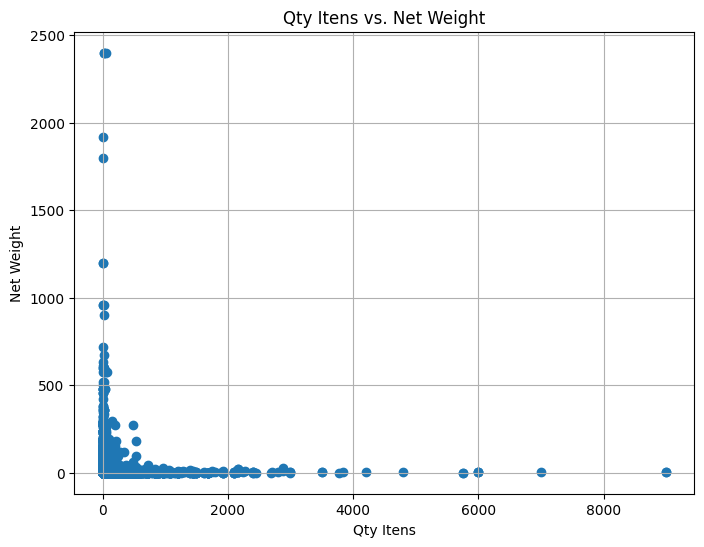
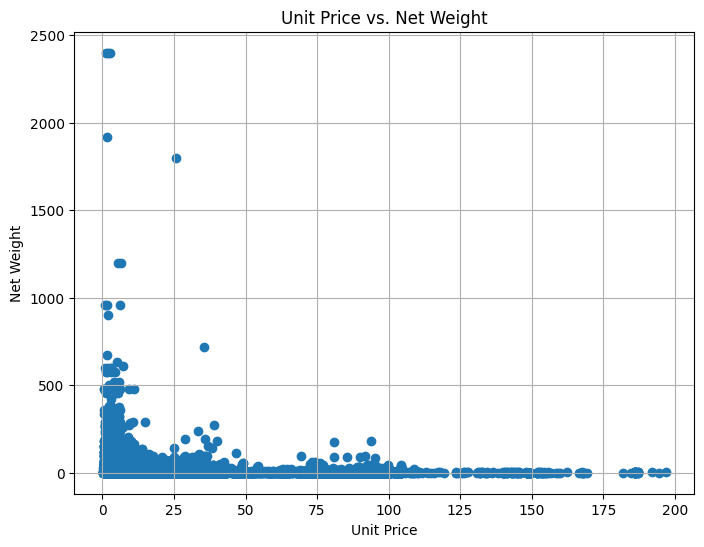
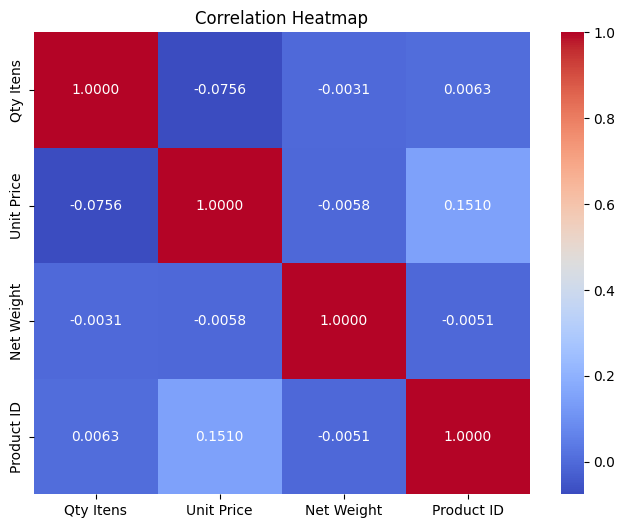
  

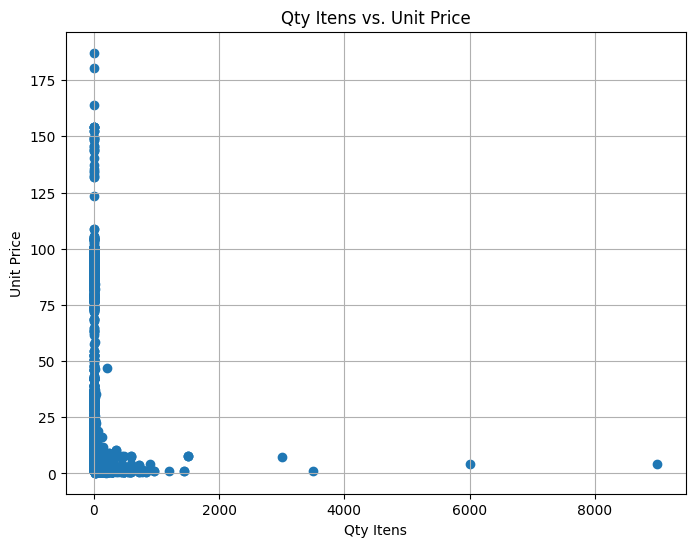
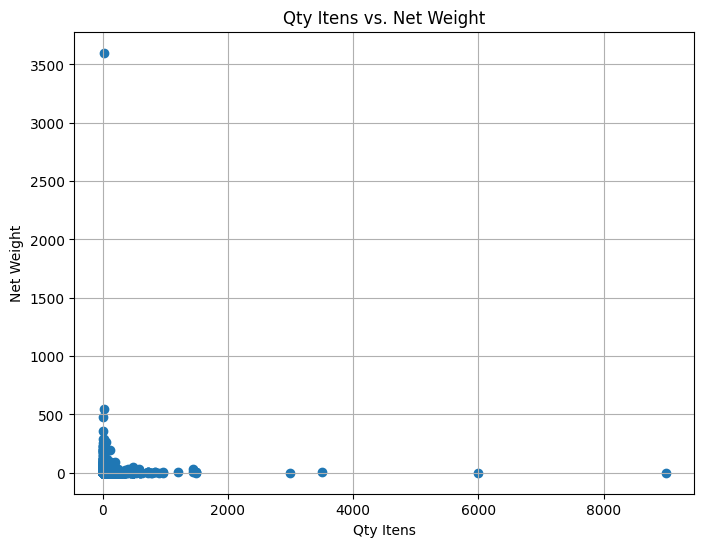
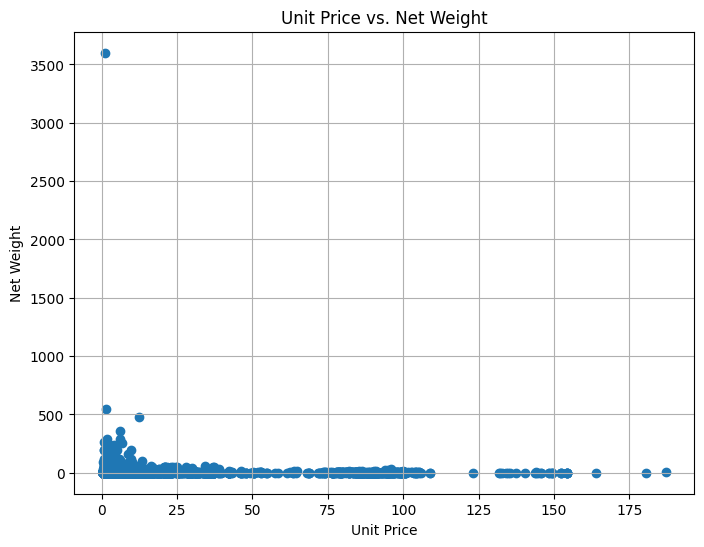
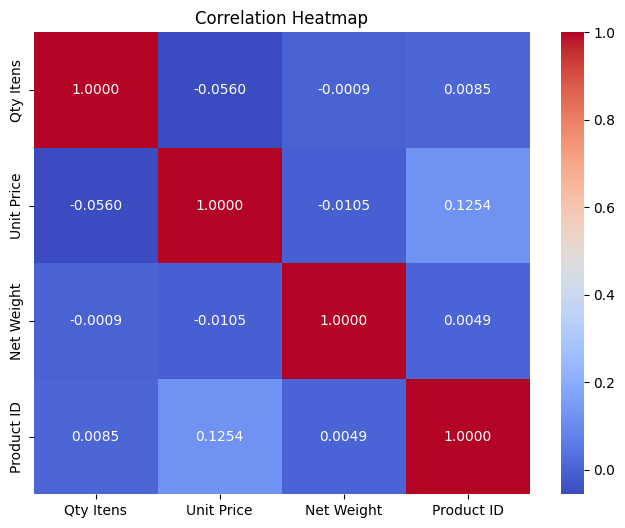
**Correlation Matrix:**

|  | **Qty Itens** | **Unit Price** | **Net Weight** | **Product ID** |
| --- | --- | --- | --- | --- |
| **Qty Itens** | 1.000000 | -0.104552 | -0.002132 | -0.017844 |
| **Unit Price** | -0.104552 | 1.000000 | -0.010493 | 0.199617 |
| **Net Weight** | -0.002132 | -0.010493 | 1.000000 | -0.011082 |
| **Product ID** | -0.017844 | 0.199617 | -0.011082 | 1.000000 |



**2018 Data:**

**2019Data:**   

6. Final Thoughts (Conclusion)

In conclusion, the exploratory data analysis provided valuable insights into the sales performance of ABC Company. By analyzing sales trends, evaluating employee performance, and understanding customer behaviour, actionable recommendations can be made to optimize sales strategies, improve targeting efforts, and enhance overall business outcomes.