Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

Name	Jay Salunke
UID	2022601007
Batch	A
Experiment no	1
Experiment name	Create basic charts using Tableau / Power BI / R / Python / D3.js to be performed on the dataset of Ecommerce field

- **AIM:** Complete all plots on practice dataset and reproduce on e-commerce dataset.
- Basic - Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot
- Calculate Product wise sales, region wise sales
- Write observations from each chart

1. Dataset

Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. However, The UCI Machine Learning Repository has made this dataset containing actual transactions from 2010 and 2011. The dataset is maintained on their site, where it can be found by the title "Online Retail".

Dataset Link:
https://www.kaggle.com/datasets/carrie1/ecommerce-data/data

Metadata:

Dataset Metadata:

1. **Number of Rows**: The dataset contains `541909` rows.
2. **Number of Columns**: The dataset consists of `8` columns.
3. **Column Names**:
   - `InvoiceNo`: The invoice number, a unique identifier for each transaction.
   - `StockCode`: The product (item) code.
   - `Description`: The product description.
   - `Quantity`: The quantity of each product per transaction.
   - `InvoiceDate`: The date and time when the transaction was generated.
   - `UnitPrice`: The price per product.
   - `CustomerID`: The unique identifier for each customer.
   - `Country`: The country where the customer resides.

Data Description:

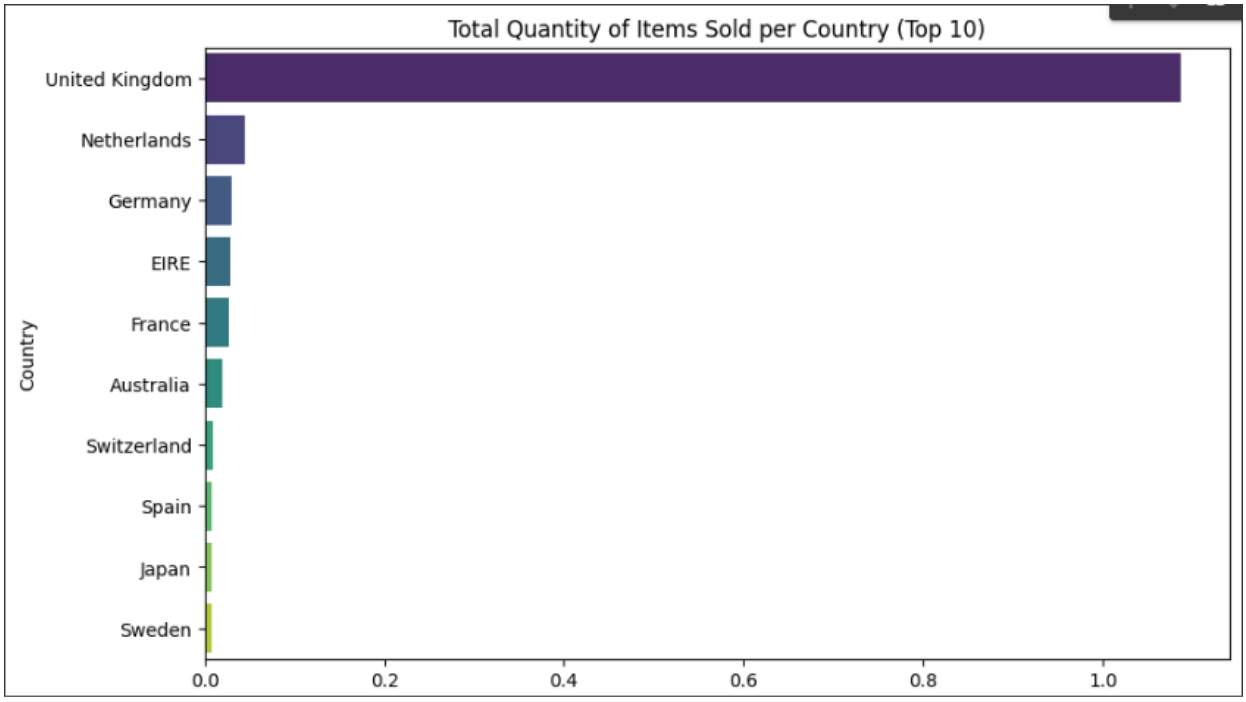
The dataset contains retail transaction data from an e-commerce store, with details on invoices, products, quantities, prices, and customer locations. It tracks purchases made by customers over

time, useful for analyzing sales patterns, customer behavior, and product performance.This data could be used to analyze the economic growth of different regions within Maharashtra over time by comparing the GDP values across years and districts.

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load your dataset
file_path = 'data.csv'  # Replace with the correct file path if needed
data = pd.read_csv(file_path, encoding='ISO-8859-1')

# Convert InvoiceDate to datetime with flexible date parsing
data['InvoiceDate'] = pd.to_datetime(data['InvoiceDate'], infer_datetime_format=True, errors='coerce')

# Drop rows where the date conversion failed
data = data.dropna(subset=['InvoiceDate'])

# 1. Bar Chart - Total Quantity of Items Sold per Country
country_sales = data.groupby('Country')['Quantity'].sum().sort_values(ascending=False).head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=country_sales.values, y=country_sales.index, palette='viridis')
plt.title('Total Quantity of Items Sold per Country (Top 10)')
plt.xlabel('Quantity Sold')
plt.ylabel('Country')
plt.show()
```

Analyzing a sales dataset to visualize the top 10 countries by total quantity of items sold. The code processes the data, handles date conversions, and creates a bar chart showing the quantity of items sold for each country, highlighting the top 10 countries with the highest sales.

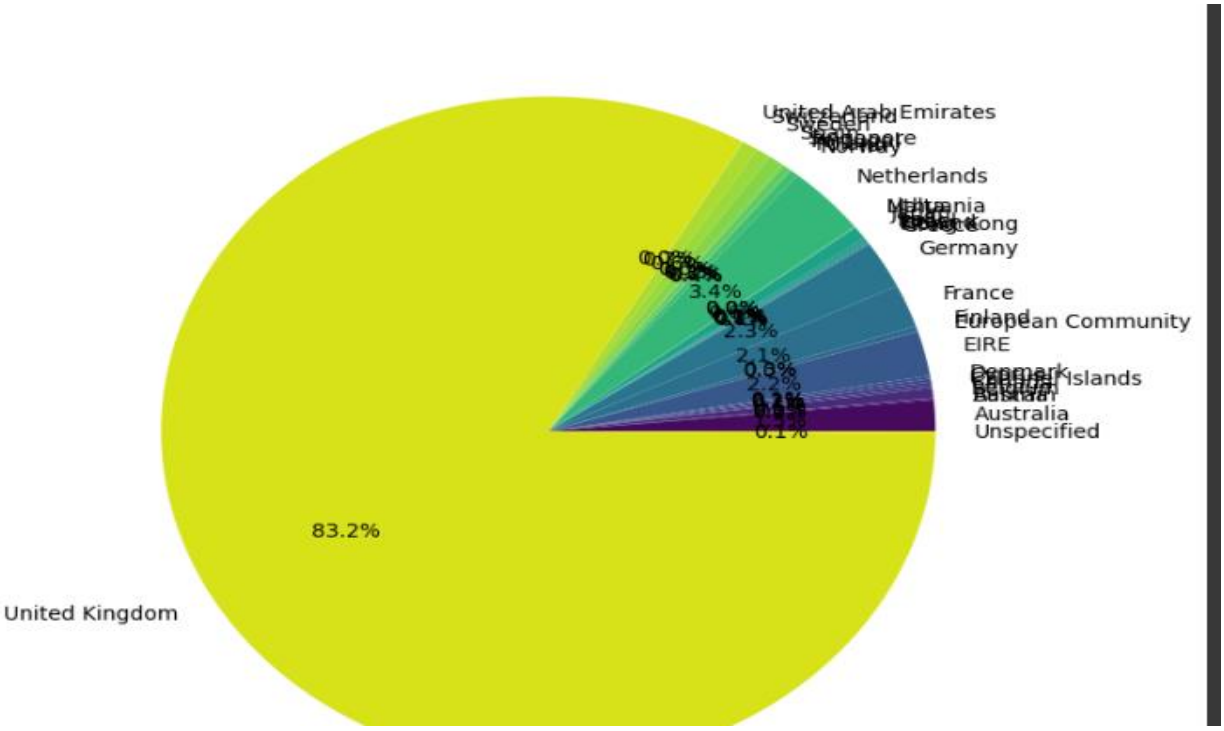


Code:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Aggregate data by country for the pie chart and filter out negative values
country_sales_pie = data.groupby('Country')['Quantity'].sum()
country_sales_pie = country_sales_pie[country_sales_pie > 0]  # Keep only positive values

# Plotting the pie chart
plt.figure(figsize=(8, 8))
plt.pie(country_sales_pie, labels=country_sales_pie.index, autopct='%1.1f%%', colors=sns.color_palette('viridis', len(country_sales_pie)))
plt.title('Distribution of Sales Across Countries')
plt.show()
```

Creating a pie chart to visualize the distribution of sales across different countries. The data is aggregated by country to sum up the total quantity sold. Negative values are filtered out to ensure only positive sales are considered. The pie chart displays the proportion of sales for each country, with segments colored using a `viridis` palette, and percentage labels are added to each segment for clarity.

PieChart



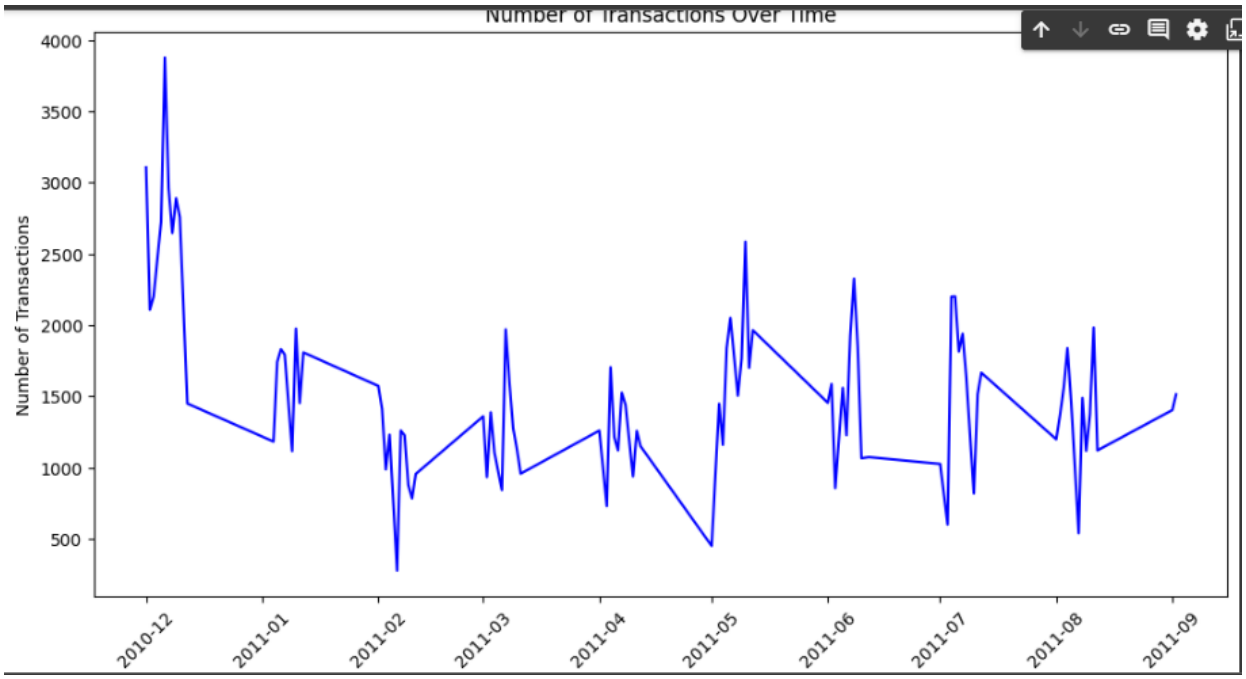
Code:

```
import matplotlib.pyplot as plt

# Extract date for daily transaction counts
data['Date'] = data['InvoiceDate'].dt.date
daily_transactions = data.groupby('Date')['InvoiceNo'].count()

# Plotting the timeline chart
plt.figure(figsize=(12, 6))
plt.plot(daily_transactions.index, daily_transactions.values, color='blue')
plt.title('Number of Transactions Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Transactions')
plt.xticks(rotation=45)
plt.show()
```

Creating a timeline chart to visualize the number of transactions over time. The data is aggregated by day to count the total transactions (`InvoiceNo`) per date. The resulting

time series is plotted with dates on the x-axis and the number of transactions on the y-axis. The chart shows how transaction volume varies over time, with date labels rotated for better readability.

TimeLine Chart

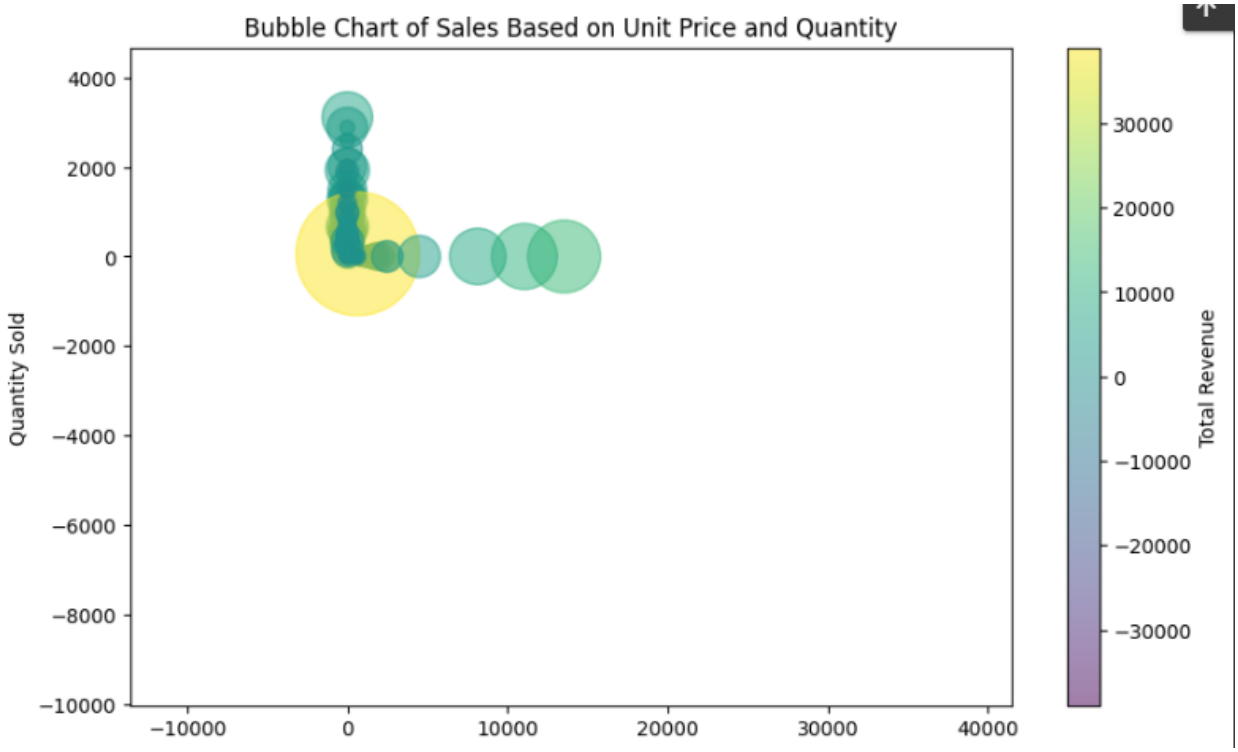


Code:

```
import matplotlib.pyplot as plt

# Calculate total revenue for bubble size
data['Revenue'] = data['UnitPrice'] * data['Quantity']

# Plotting the bubble chart
plt.figure(figsize=(10, 6))
plt.scatter(data['UnitPrice'], data['Quantity'], s=data['Revenue']/10, alpha=0.5, c=data['Revenue'], cmap='viridis')
plt.title('Bubble Chart of Sales Based on Unit Price and Quantity')
plt.xlabel('Unit Price')
plt.ylabel('Quantity Sold')
plt.colorbar(label='Total Revenue')
plt.show()
```

Creating a bubble chart to analyze sales data based on unit price and quantity. Each bubble represents a transaction, with the x-axis showing the unit price and the y-axis displaying the quantity sold. The size of each bubble corresponds to the total revenue from that transaction, scaled down for visibility. The color of the bubbles indicates the revenue amount, with a color gradient provided by the `viridis` colormap. The chart helps visualize the relationship between unit price, quantity sold, and revenue.

Bubble Chart

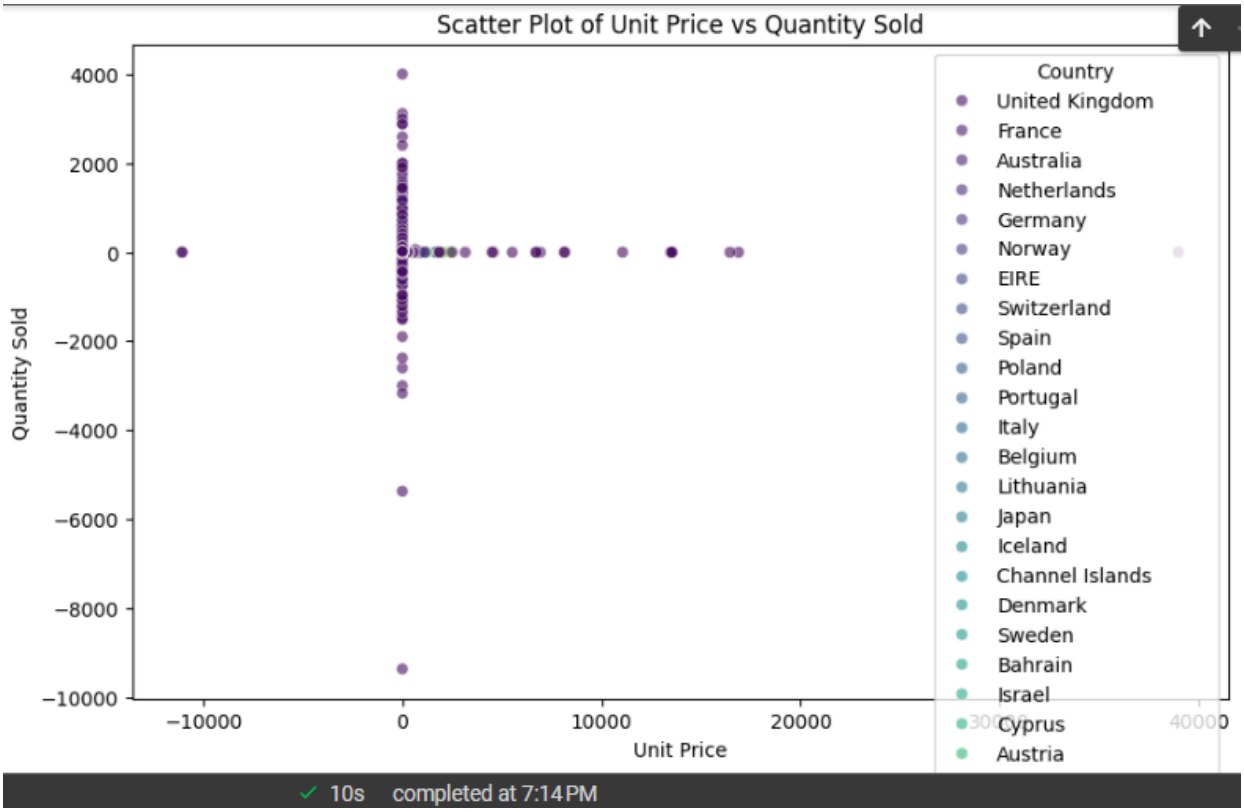


Code:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Plotting the scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='UnitPrice', y='Quantity', data=data, hue='Country', palette='viridis', alpha=0.6)
plt.title('Scatter Plot of Unit Price vs Quantity Sold')
plt.xlabel('Unit Price')
plt.ylabel('Quantity Sold')
plt.show()
```

creating a scatter plot to explore the relationship between unit price and quantity sold, with data points colored by country. Each point represents a transaction, with the x-axis showing the unit price and the y-axis representing the quantity sold. The color of each point indicates the country, using the `viridis` color palette to differentiate between them. The plot helps visualize how unit price and quantity sold vary across different countries.

Scatter Plot



Code:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Filter out extreme outliers for a clearer histogram
filtered_data = data[data['UnitPrice'] < data['UnitPrice'].quantile(0.95)]  # Filter out top 5% extreme prices

# Plotting the histogram with adjusted bins
plt.figure(figsize=(10, 6))
sns.histplot(filtered_data['UnitPrice'], bins=50, kde=True, color='purple')
plt.title('Distribution of Unit Prices (Filtered)')
plt.xlabel('Unit Price')
plt.ylabel('Frequency')
plt.show()
```

creating a histogram to visualize the distribution of unit prices, after filtering out extreme outliers. Specifically, you exclude the top 5% of the highest unit prices to focus on more typical values. The histogram is plotted with 50 bins and includes a kernel density estimate (KDE) to show the distribution's shape. The histogram uses a purple color to display the frequency of

different unit prices, providing a clearer view of the price distribution without the influence of extreme values.

Histogram plot

Questions asked :

1. **Bar Chart of Total Quantity of Items Sold per Country**:
   - **Question Answered**: Which countries have the highest total quantity of items sold?
2. **Pie Chart of Sales Distribution Across Countries**:
   - **Question Answered**: What is the proportion of total sales attributed to each country?
3. **Timeline Chart of Number of Transactions Over Time**:
   - **Question Answered**: How does the number of transactions vary over time?
4. **Bubble Chart of Sales Based on Unit Price and Quantity**:
   - **Question Answered**: What is the relationship between unit price, quantity sold, and total revenue for transactions?
5. **Scatter Plot of Unit Price vs. Quantity Sold**:
   - **Question Answered**: How do unit price and quantity sold relate to each other across different countries?
6. **Histogram of Unit Price Distribution**:
   - **Question Answered**: What is the distribution of unit prices, excluding extreme values?

Conclusion: The analyses reveal:

1. **Top Countries**: The top 10 countries contribute most to total sales.
2. **Sales Distribution**: Sales are spread unevenly across countries.
3. **Transaction Trends**: Daily transactions show fluctuations over time.
4. **Revenue Insights**: Revenue correlates with unit price and quantity sold.
5. **Pricing Patterns**: Most unit prices are within a typical range, excluding extreme values.

Submission
https://docs.google.com/forms/d/e/1FAIpQLSc9g-c226PEYBN02hiE0vgdLtDYVSnUgtN1RhjBcZfSxz10Bg/viewform?usp=sf_link