Mục tiêu

Sử dụng thư viện Seaborn để vẽ và tinh chỉnh được một số biểu đồ trên bộ dữ liệu Online Retail

Bộ dữ liệu Online Retail mô tả tình hình kinh doanh của một doanh nghiệp bán hàng online mà bạn đã được làm quen từ các buổi trước.

Trong bài thực hành này, chúng ta sẽ sử dụng thư viện Seaborn vừa học được để vẽ một số biểu đồ trên bộ dữ liệu này

Khai báo thư viện cần dùng

```
import pandas as pd
import seaborn as sns
```

Đọc dữ liệu

```
df = pd.read_csv("OnlineRetail.csv", encoding = "ISO-8859-1")
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

```
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

```
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Loại bỏ dữ liệu bị khuyết

```
df = df.dropna()
```

Tính giá của mỗi mã sản phẩm ở các đơn hàng

```
df["Price"] = df["Quantity"] * df["UnitPrice"]
```

```
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Vẽ biểu đồ phân bố

Biểu đồ phân bố cho giá sản phẩm

```
sns.violinplot(y = "UnitPrice", data=df)
```

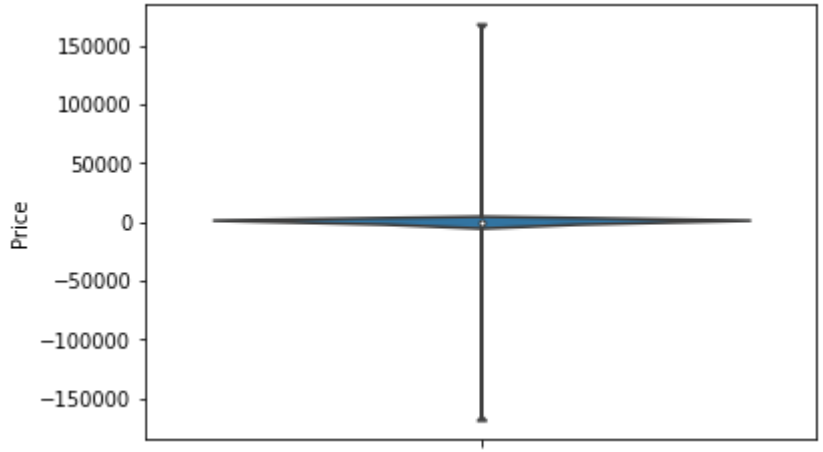
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae76d177d0>
```

Có thể nhận thấy dữ liệu phân bố không được đều, nhiều sản phẩm giá thấp và ít sản phẩm giá cao.

Biểu đồ phân bố cho tổng giá mỗi sản phẩm ở các đơn

```
sns.violinplot(y = "Price", data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae687c12d0>
```



Tính số lượng sản phẩm ở mỗi đơn hàng

```
df2 = df.groupby(['InvoiceNo'])['Quantity'].sum().reset_index()
```

```
df2.head()
```

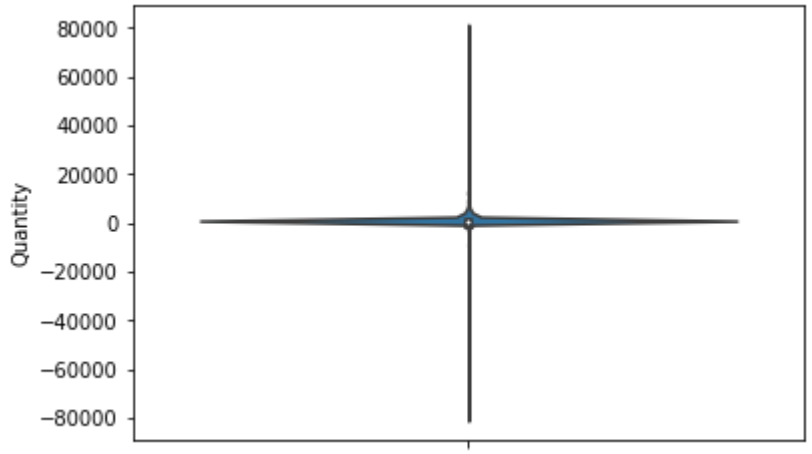
	InvoiceNo	Quantity
0	536365	40
1	536366	12
2	536367	83

Biểu đồ phân bố số lượng sản phẩm trên mỗi đơn

4	536369	3

```
sns.violinplot(y="Quantity", data=df2)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae6846d250>
```



Vẽ biểu đồ tần số

```
df3 = df.groupby(['Country'])['Quantity'].sum().reset_index()
```

```
df3.head()
```

	Country	Quantity
0	Australia	83653
1	Austria	4827
2	Bahrain	260
4	Brazil	356

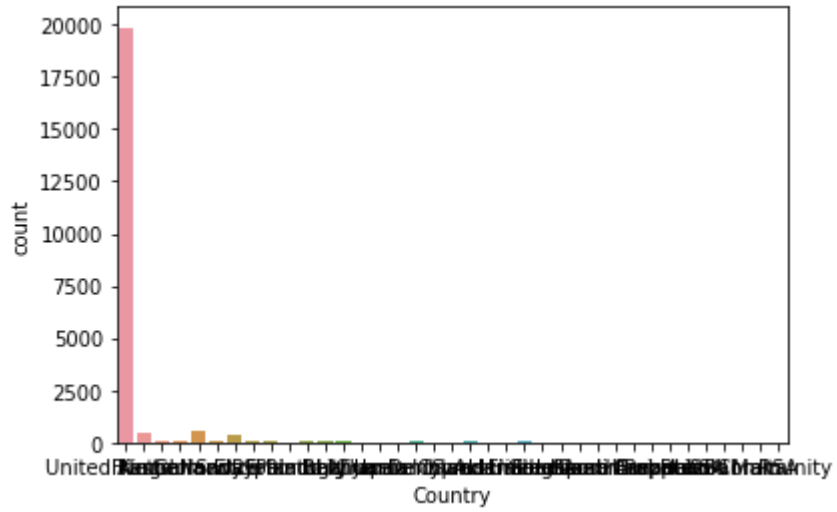
Loại bỏ dữ liệu về hóa đơn trùng lặp

```
df1 = df.drop_duplicates(subset = 'InvoiceNo')
```

Vẽ biểu đồ tần số cho số hóa đơn theo quốc gia

```
sns.countplot(x = "Country", data = df1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae76caf410>
```



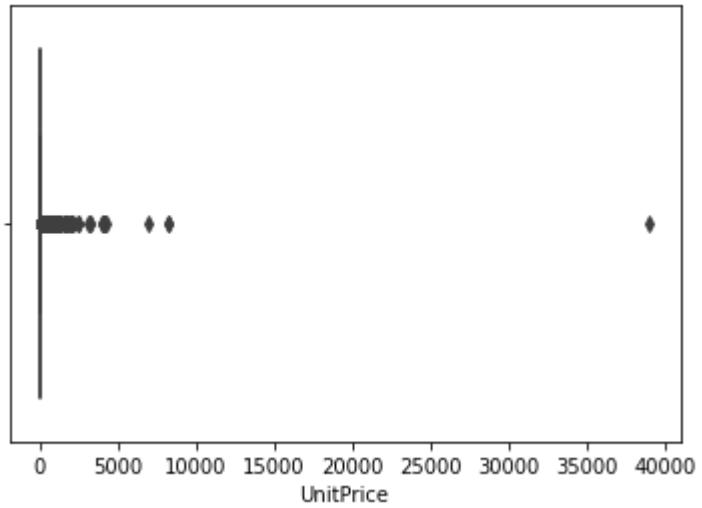
Có thể nhận thấy rằng số sản phẩm và số hóa đơn được mua rất nhiều ở Anh, áp đảo so với phần còn lại

Vẽ biểu đồ box plot

Vẽ biểu đồ box plot cho thuộc tính giá sản phẩm

```
sns.boxplot(x=df["UnitPrice"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae6d464c90>
```



Tương tự như trên đã nhận xét, giá sản phẩm phân bố rất không đồng đều, tập trung ở giá thấp

Biểu đồ box plot cho số lượng sản phẩm mỗi đơn

```
sns.boxplot(x=df2["Quantity"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fae683f3390>
```

Tổng kết

Qua bài thực hành này, chúng ta đã ôn lại cách vẽ các biểu đồ đã được học với bộ dữ liệu Food Price in Turkey