Introdução à Ciência dos Dados
Ciência da Computação – UFV/Florestal

Aula Prática 06
Visualização de Dados

Objetivo: entender comandos para explorar dados visualmente por meio de gráficos
Pré-requisitos: linguagem de programação Python, Linux, estatística.
Meta: ao final da prática, o aluno será capaz de analisar um conjunto de dados visualmente para ajudar na tomada de decisões.

Roteiro

- Importar as bibliotecas

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

- Ler os dados em formato Excel

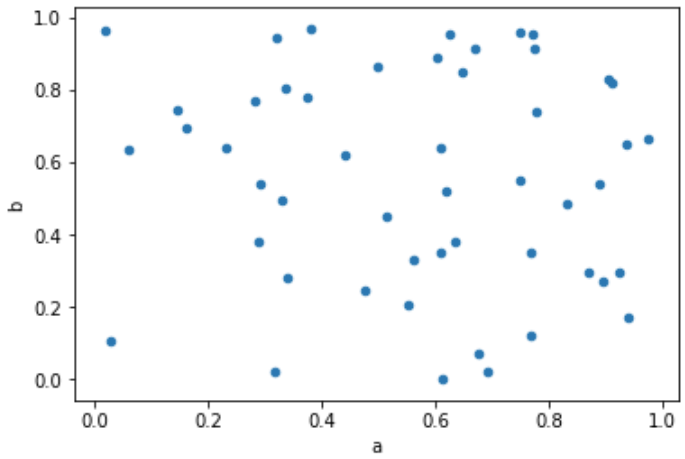
```
# Ler dados em formato excel
#df = pd.read_excel('http://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx')
df = pd.read_excel('Online%20Retail.xlsx')
```

- Filtrar quantidades e preços unitários negativos

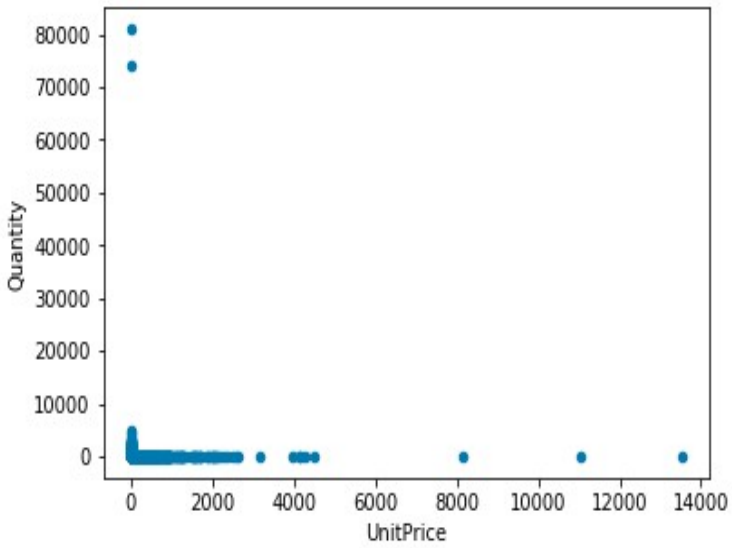
```
# Filtro para remover quantidade e preço unitário negativos
df=df[(df['Quantity'] > 0) & (df['UnitPrice'] > 0)]
```

- Gráfico de Pontos (Scatter Plot)

```
df_random = pd.DataFrame(np.random.rand(50, 4), columns=['a', 'b', 'c', 'd'])
df_random.plot.scatter(x='a', y='b');
plt.show()
```

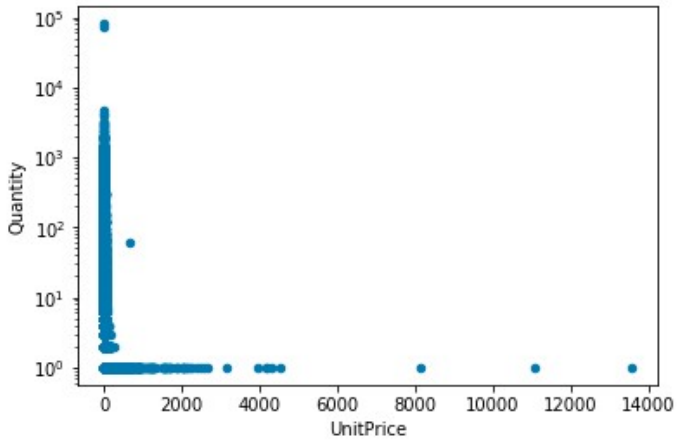


```
df.plot.scatter(x='UnitPrice', y='Quantity')
plt.show()
```



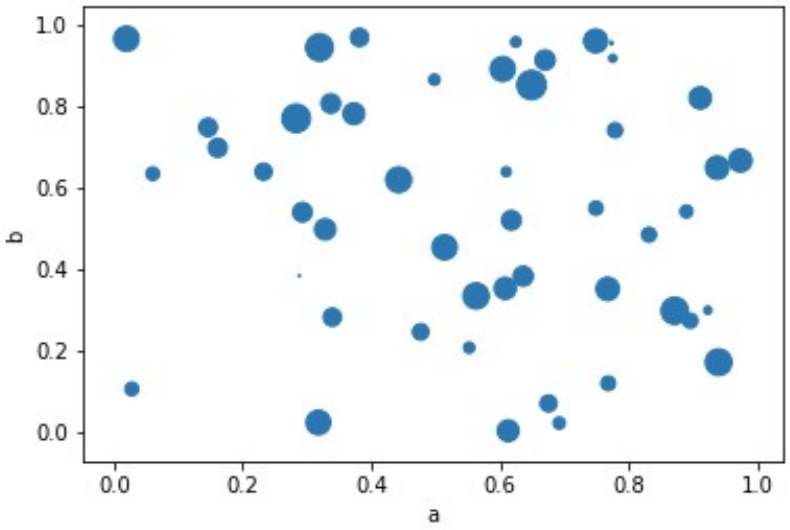
- Usar escala Log

```
# Colocar o eixo y em escala log (Atenção para valores menores que 1)
df.plot.scatter(x='UnitPrice', y='Quantity')
plt.xscale('linear')
plt.yscale('log')
plt.show()
```



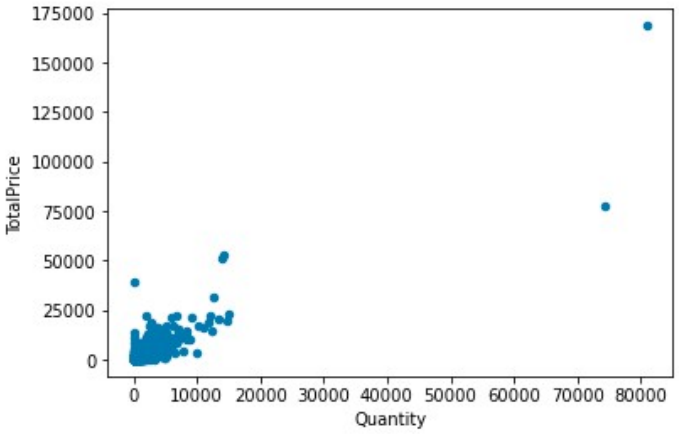
- Usar um terceiro atributo para definir o tamanho do marcador

```
df_random.plot.scatter(x='a', y='b', s=df_random['c']*200);
plt.show()
```



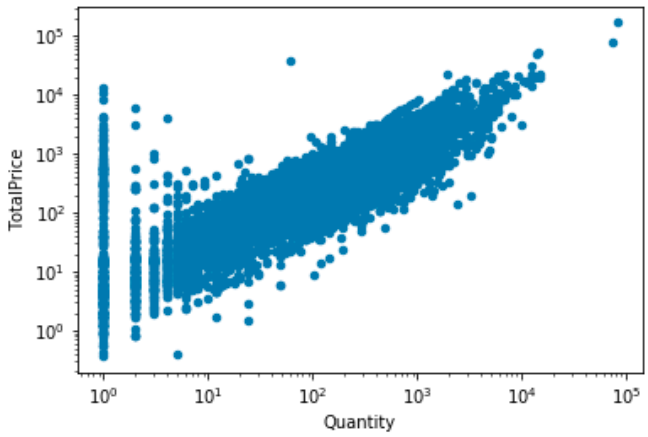
- Faz agrupamento antes de exibir o gráficos

```
# Calcula o TotalPrice para cada registro
df['TotalPrice']=df['Quantity']*df['UnitPrice']
# Agrupa por invoiceNo
df_group_invoice=df.groupby('InvoiceNo')
# Soma os atributos
df_group_invoice_sum=df_group_invoice.sum()
# Mostra a relação Quantidade x Total
df_group_invoice_sum.plot.scatter(x='Quantity',y='TotalPrice')
plt.show()
```



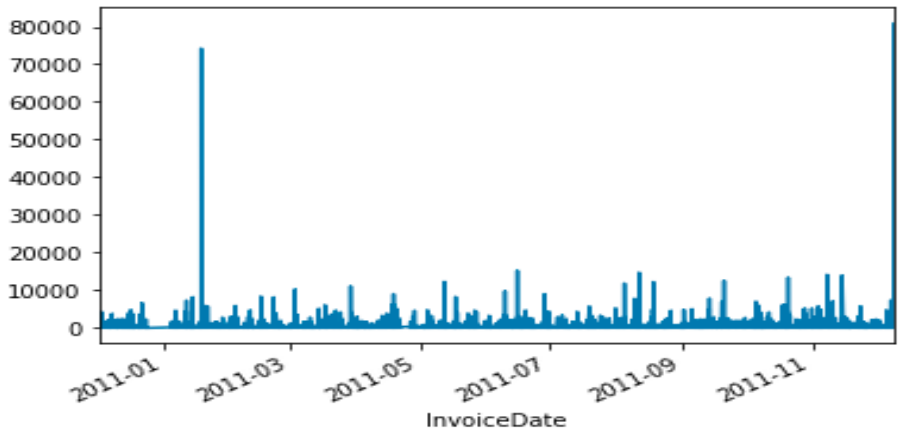
- Mesmo gráfico anterior, mas na escala log nos dois eixos

```
# Mostra a relação Quantidade x Total com ambos eixos na escala log
df_group_invoice_sum.plot.scatter(x='Quantity',y='TotalPrice')
plt.xscale('log')
plt.yscale('log')
plt.show()
```



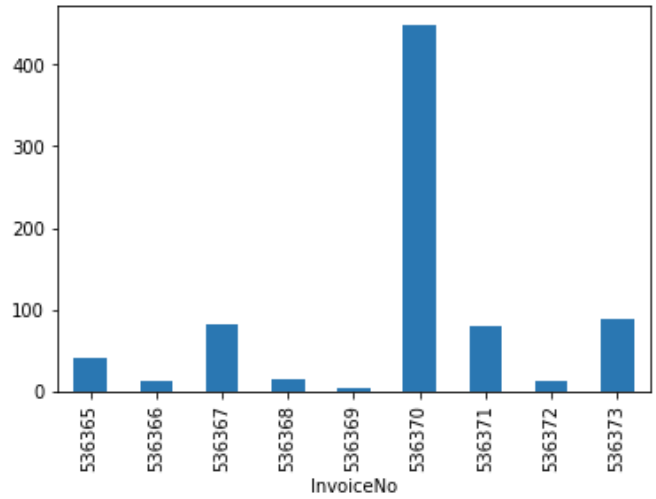
- Gráfico de linha para avaliar tendência. Agrupa as compras por data e mostra a quantidade de itens por data.

```
# Agrupa por Data
df_group_date=df.groupby('InvoiceDate')
# Soma os atributos
df_group_date_sum=df_group_date.sum()
# Mostra a relação Quantidade x Data
df_group_date_sum['Quantity'].plot()
plt.show()
```



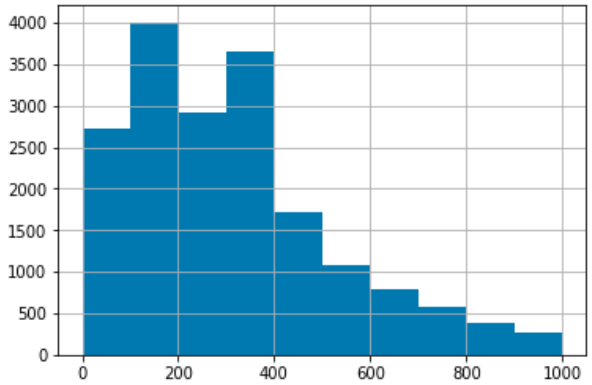
- Gráfico de barras para mostrar valores para atributos categóricos

```
# Filtra os 10 primeiros invoces agrupados, e plota as barras da quantidade
df_group_invoice_sum.iloc[0:9]['Quantity'].plot(kind='bar')
plt.show()
```

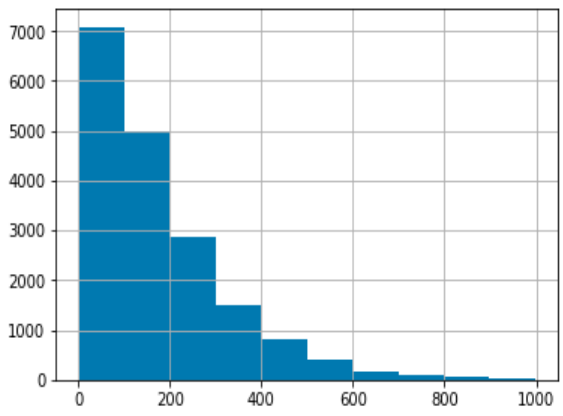


- Histograma

```
# Filtra TotalPrice < 1000 e faz o histograma
df_group_invoice_sum.loc[(df_group_invoice_sum['TotalPrice'] < 1000)]['TotalPrice'].hist()
plt.show()
```

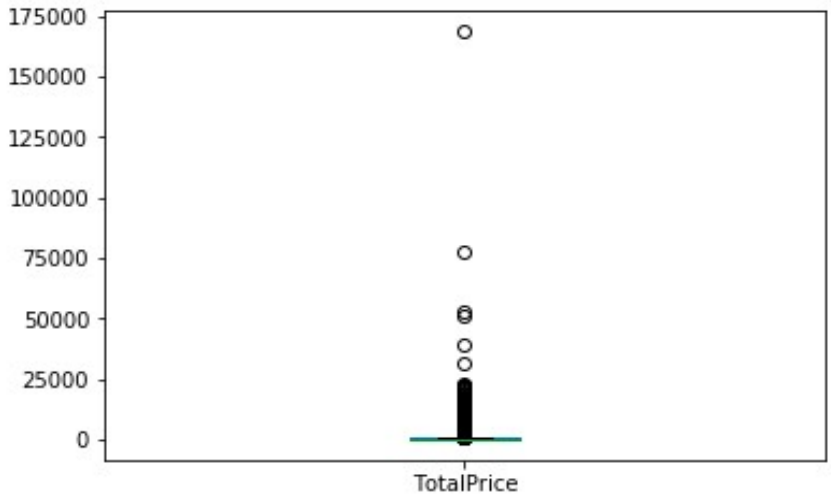



```
# Filtra TotalPrice < 1000 e Quantity < 1000 e faz o histograma
df_group_invoice_sum.loc[(df_group_invoice_sum['TotalPrice'] < 1000) \
                         & (df_group_invoice_sum['Quantity'] < 1000)]['Quantity'].hist()
plt.show()
```

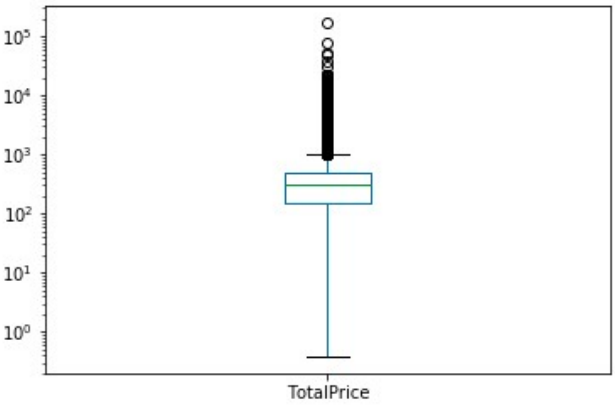


- Boxplot

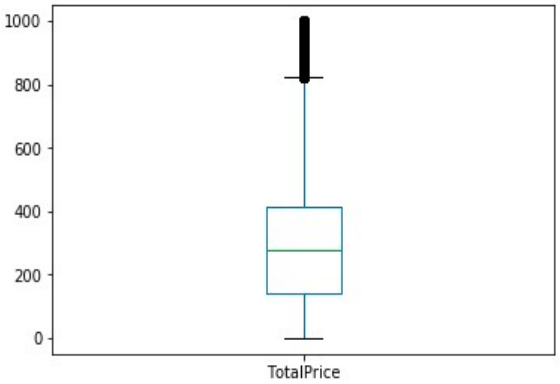
```
# Boxplot do TotalPrice
df_group_invoice_sum['TotalPrice'].plot.box()
plt.show()
```




```
# Boxplot do TotalPrice
df_group_invoice_sum['TotalPrice'].plot.box()
plt.yscale('log')
plt.show()
```




```
# Boxplot do TotalPrice
df_group_invoice_sum.loc[(df_group_invoice_sum['TotalPrice'] < 1000)]['TotalPrice'].plot.box()
plt.show()
```



Atividades:

1. Faça um gráfico de barras mostrando a quantidade vendida total de cada um dos 10 primeiros produtos do Dataframe.
2. Faça um Boxplot dos preços unitários dos produtos. Considere somente os 100 produtos mais vendidos.
3. Faça um gráfico de linhas mostrando o faturamento (total de vendas) por dia.
4. Faça o histograma da média dos preços unitários dos produtos.
5. Faça dois gráficos de barras, sendo um da quantidade de compras e outro do total de faturamento por país.
6. Considerando os 100 produtos mais vendidos, é possível visualizar alguma diferença entre a distribuição dos preços unitários desses produtos ao se comparar os países “Australia” e “United Kingdom”? Faça um gráfico que ajude a responder.

Referências:

https://pandas.pydata.org/pandas-docs/stable/visualization.html