Visualização de Dados com Exemplos em Python

Rogério de Oliveira

Índice

1 Introdução

Existem ao menos dois tipos de visualização de dados. Uma primeira é voltada para que você explore dados e faça descobertas como padrões, relações e processos em ação sobre os dados. Um outro tipo de visualização é aquele que fornece ilustrações informativas, claras e visualmente atraentes de seus resultados, e que podem ser apresentados a outras pessoas ou ser incluídos em uma publicação.

Ambos os tipos de visualização podem ser feitos com `Python`, assim como em outras linguagens e ferramentas. Nossa opção pelo uso de `Python` é sua ampla popularidade, a adoção em um grande número de cursos de graduação e seu uso nas mais diversas áreas como Ciências da Computação, Engenharias, Biologia e Ciências da Saúde.

A **Visualização de Dados** envolve uma série de questões de como fazer descobertas sobre os dados ou apresentar informações. Essas questões vão muito além da linguagem ou ferramentas empregadas para a produção gráfica e constituem a trilha de Análise da Visualização de Dados. Neste texto introdutório, você irá aprender como empregar corretamente gráficos para responder questões relevantes sobre os dados, ao mesmo tempo que aprenderá a construir representações gráficas úteis com `Python`. Essas duas trilhas de apresentação serão desenvolvidas com o estudo de casos e aplicações em diferentes áreas como Finanças, Comércio, Saúde, Ciência e Tecnologia. Você poderá seguir apenas pela

trilha das questões de Visualização de Dados caso não queira se aprofundar no uso da linguagem `Python`, ou deixá-la para um segundo momento.

Objetivos do Capítulo

Neste capítulo você:

1. Entenderá o por quê da necessidade de visualização dos dados
2. Será capaz de reconhecer as principais questões sobre dados
3. Empregará gráficos simples na solução de Análise de Dados

1.1 Importância da Visualização de Dados

O **Quarteto de Anscombe** (F.J. Anscombe, 1973) talvez seja o exemplo mais conhecido que ilustra o valor da visualização dos dados, mesmo diante de várias informações que podemos obter, como variáveis estatísticas dos dados. Ele é formado por quatro conjuntos de dados que aparentam ser idênticos quando descritos por técnicas de estatística descritiva como a média e a variância, mas que são muito distintos quando exibidos graficamente.

Os quatro conjuntos de dados exibidos abaixo, pares (x, y), apresentam com até 3 casas decimais as mesmas médias e variâncias de x e y e correlação de 0.816, levando todos a uma mesma regressão linear:

$$y = 3 + 0.5x$$

[0]:
```
import seaborn as sns
sns.set(style="ticks")

# Load the example dataset for Anscombe's quartet
df = sns.load_dataset("anscombe")

# Show the results of a linear regression within each dataset
sns.lmplot(x="x", y="y", col="dataset", hue="dataset", data=df,
           col_wrap=2, ci=None, palette="muted", height=4,
           scatter_kws={"s": 50, "alpha": 1})
```

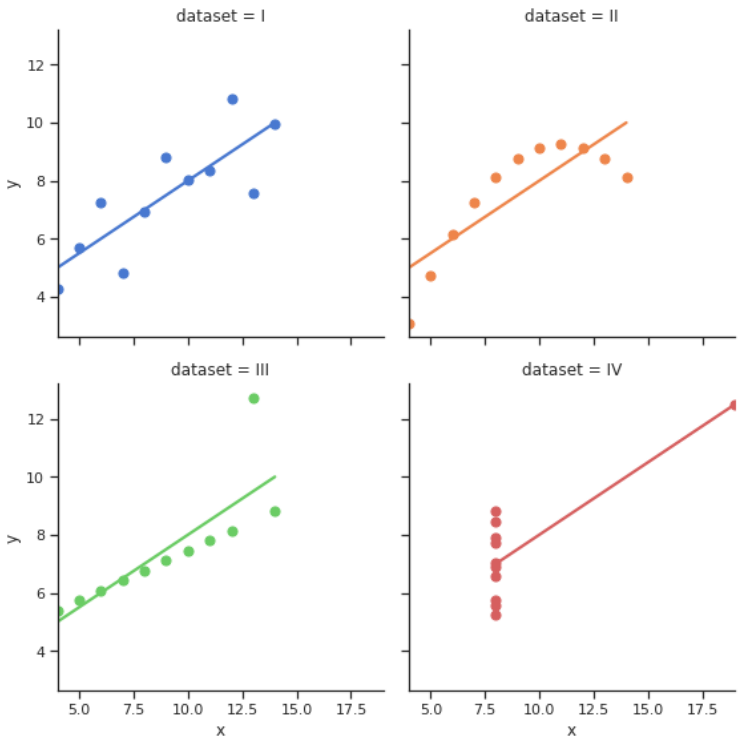

Fig 1: Quarteto de Anscombe (F.J. Anscombe, 1973).

Esse exemplo simples mostra a importância da Visualização de Dados na descoberta de relações entre os dados que dificilmente poderíamos identificar de outro modo.

Um exemplo mais elaborado pode ser dado pelo gráfico abaixo.

[0]:
```
# Libraries
import seaborn as sns
import pandas as pd
from matplotlib import pyplot as plt

# Data set
url = 'https://python-graph-gallery.com/wp-content/uploads/mtcars.csv'
df = pd.read_csv(url)
df = df.set_index('model')
del df.index.name
df

sns.clustermap(df, metric="euclidean", standard_scale=1, method="ward",␣
↪cmap="Blues")
```

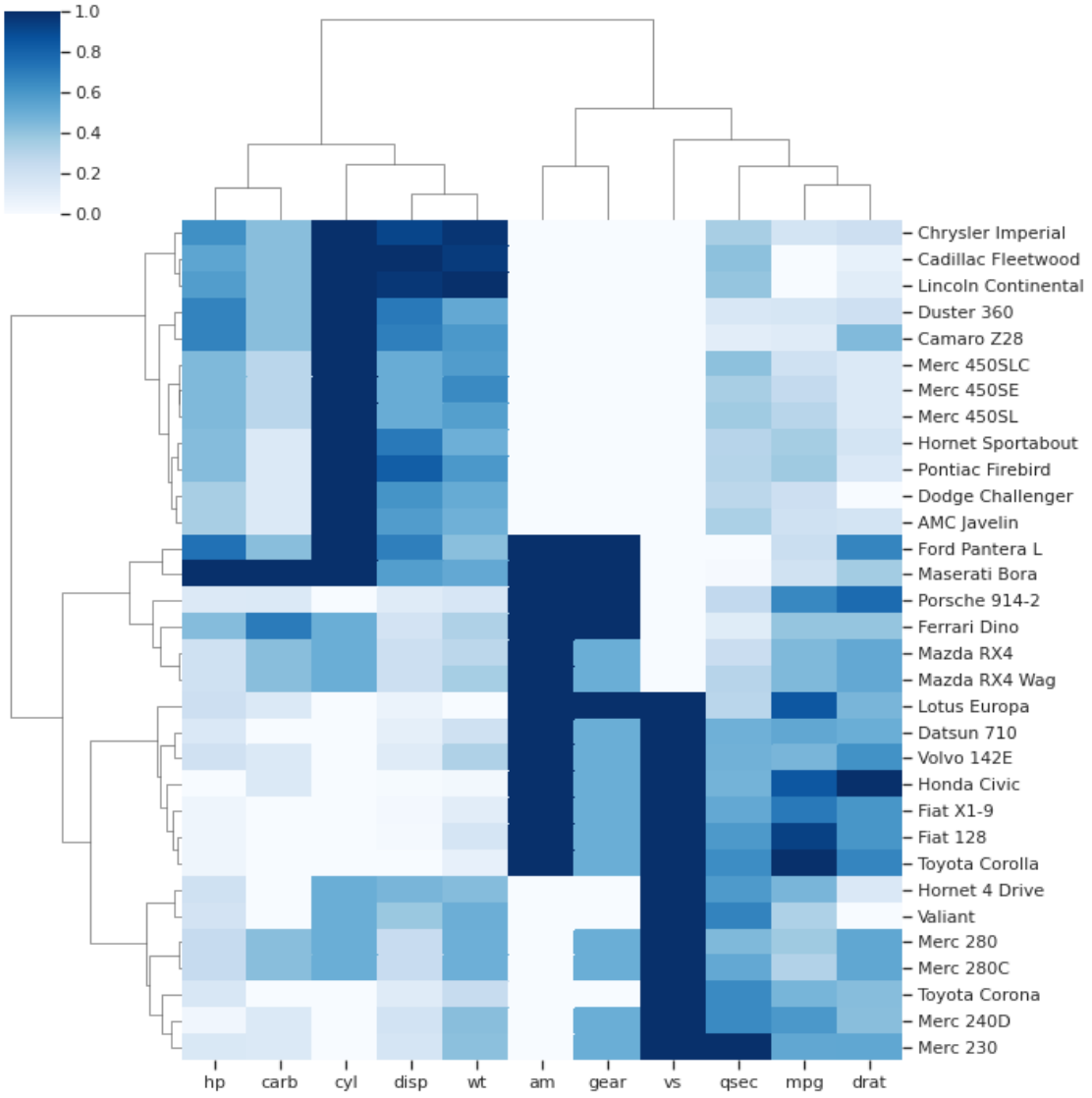

Fig 2: *Mapa de Calor* (ou *Heat Map*). A intensidade ou diferentes cores indicam a escala de valores.

Esse gráfico, conhecido como *Mapa de Calor* (ou *Heat Map*), não só serve para a descoberta de relações sobre os dados, como ilustra como o conjunto encontra-se dividido de acordo com características dos veículos (`hp`, `mpg` etc.) em dois grupos distintos. Ele apresenta a informação de um modo que mesmo um leigo poderá compreender a relação existente e, dificilmente, poderíamos apresentar essa informação de modo claro que não uma forma visual.

[0]:
```
# Libraries
from wordcloud import WordCloud
import matplotlib.pyplot as plt
plt.figure(figsize=(10,5))

# Read a Web content
text = read.url(meusite.mackenzie.br/rogerio/thisPythonBook.htm)

# Create the wordcloud object
wordcloud = WordCloud(width=880, height=880, margin=0).generate(text)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
```

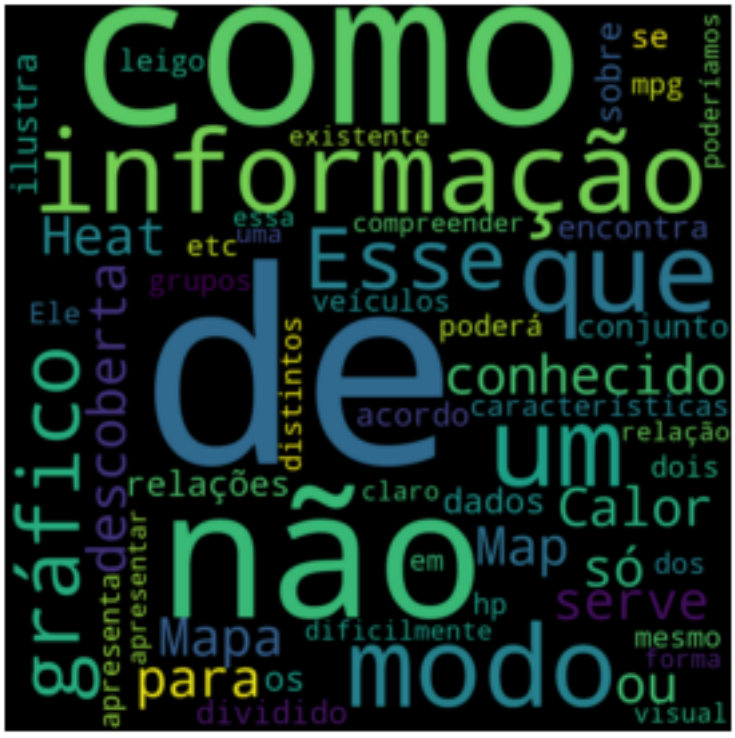

Fig 3: *Word Cloud*. Visualização clara de termos mais frequentes em textos, páginas Web, e-mails são usados para classificação automática de documentos.

O gráfico acima é conhecido como `Word Cloud` e é bastante popular na Web. Seu propósito? Identificar de modo fácil as maiores ocorrências de termos. Isso tem uma aplicação muito grande em negócios se você imaginar por exemplo esse gráfico sobre o texto das reclamações de clientes, ou sobre um catálogo de produtos, permitindo identificar facilmente as ocorrências mais comuns.

1.2 Fazendo Perguntas Relevantes sobre os Dados

Agora que entendemos a importância da Visualização dos Dados precisamos entender quais são as perguntas que podemos fazer sobre os dados e que gráficos devemos empregar para responder essas perguntas. Boa parte desse livro se dedica a entrar em detalhes sobre essas perguntas e que gráficos empregar em cada caso.

As perguntas que podemos fazer sobre os dados podem ser agrupadas em algumas categorias relevantes, cada uma delas tendo gráficos mais adequados para você obter e apresentar respostas à sua questão:

Evolução (ou *Tendência dos Dados*): gráficos de linha, área, séries múltiplas

Distribuição: histogramas, gráficos de distribuição de densidade, *boxplot*

Ranking: gráficos de barras, *word cloud*, *spider*

Correlação: gráficos de dispersão, *heat map*, *density 2D*

Partes de um todo: *Tree map*, diagramas Venn, *pie chart*

Gráficos mais especializados ainda podem envolver análises geográficas (*maps*), fluxo de dados em redes sociais e complexas estruturas de *Grafos*.

Para os nossos propósitos será suficiente tratarmos apenas as categorias acima e dedicaremos, cada um dos capítulos seguintes, à solução de casos envolvendo cada uma das categorias de questões.

Mas antes de nos aprofundarmos, vamos ver alguns exemplos dessas questões e como obtemos respostas gráficas à elas.

1.3 Evolução

Muitos dados apresentam uma evolução ao longo tempo. Nesses casos, frequentemente queremos saber sobre a tendência dessa evolução, sobre a probabilidade de se alcançar um valor ou ainda comparar a forma de evolução de variáveis que possam ter alguma relação.

Quando essas variáveis são numéricas, um gráfico de linhas é frequentemente usado para visualizar a tendência nos dados em intervalos de tempo e responder essas perguntas.

> **Estudo de Caso 1: Investigando Variações de Ações da Vale e Petrobrás.** Você possui dados recentes do índice da Bolsa, do dólar e do valor de ações de duas grandes empresas (Petrobrás e Vale). Você está interessado em saber se as ações das empresas têm uma evolução causada por mecanismos próprios de cada empresa ou do setor, como o aumento da demanda de aço, ou seguem uma tendência geral do mercado.

Esses dados apresentam o seguinte *layout*:

[0]:
```
mystocksn = pd.read_csv('mystocksn.csv')
mystocksn.head(10)
```

[0]:
```
          data      IBOV  VALE3  PETR4    DOLAR
0   2020-02-03  114629.0  11.90  14.20  4.28260
1   2020-02-04  115557.0  12.23  14.37  4.24680
2   2020-02-05  116028.0  12.44  14.43  4.25370
3   2020-02-06  115190.0  12.40  14.63  4.23360
...
    2020-02-12  116674.0  12.12  14.92  4.32753
    2020-02-13  115662.0  11.93  14.63  4.35450
```

Empregando um gráfico de linhas como abaixo você pode identificar que tanto o índice da bolsa, como as ações das duas empresas, seguem uma evolução bastante semelhantes indicando. Isso é uma forte indicação de que os valores têm uma causa comum para sua evolução (no caso o cenário externo) e, portanto, estão menos relacionados a causas específicas de cada empresa ou setor (como uma greve dos petroleiros ou um acidente em uma barragem de exploração de minério).

[0]:
```
mystocksn_norm = mystocksn

for item in ['IBOV','VALE3','PETR4','DOLAR']:
  mystocksn_norm[item] = mystocksn[item]/mystocksn[item].max()
  sns.lineplot(x='data',y=item,data=mystocksn_norm,label=item)

# Add legend
plt.legend(loc=3, ncol=1)
```

```
# Add titles
plt.title("Ações Normalizadas Bovespa", loc='left', fontsize=18,␣
 ↪fontweight=0, color='gray')
plt.xlabel("Data")
plt.ylabel("Pontos")

plt.xticks(rotation=45)
plt.show()
```

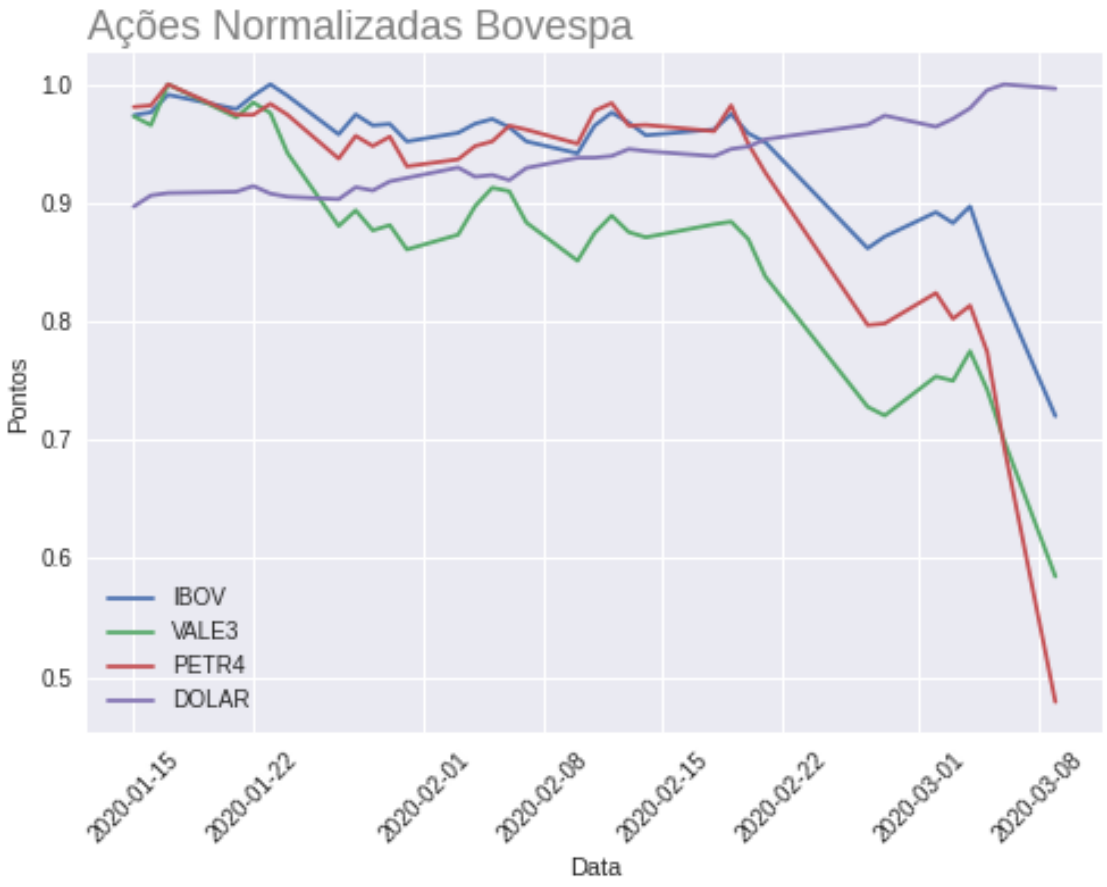

Fig 4: Gráficos de linhas múltiplas. Múltiplos valores de indíces e ações **normalizados** permitem comparar as séries e identificar possíveis causas comuns na evolução dos dados.

1.3.1 Normalizando os Dados

Mas para chegar nesse resultado você terá de aprender, além da escolha apropriada do gráfico para exibir esses resultados, como preparar os dados para exibir informações corretas. O 'mesmo' gráfico acima seria pouco informativo se não tivessemos feito antes uma **normalização** dos dados. Sendo o IBOV um valor em geral no intervalo de milhares $60.000 - 150.000$ e as ações e dólar valores em intervalos bem menores $3 - 60$, se os valores não fossem normalizados, não seria possível distinguir os valores menores, que seriam exibidos em uma *aparentemente* linha única. Retomaremos esse e outros pontos com mais detalhe quando estudarmos a **Evolução ou Tendência dos Dados** no capítulo 2.

1.4 Distribuição

Os dados a seguir foram obtidos da **WHO World Health Organization** (*Life expectancy and Healthy life expecancy*). São indicadores de saúde de diversos países como o índice de massa corpórea médio da população (BMI), expectativa de vida e índices de mortalidade. Uma pergunta relevante sobre a saúde global poderia ser sobre a distribuição desses índices entre os países, como por exemplo o caso a seguir.

> **Caso 2: WHO - BMI e Expectativa de Vida Média Globais.** Como se distribuem os valores de BMI e de Expectativa de Vida Média dos indivíduos entre os diferentes Países?

Os dados disponíveis da Organização Mundial de Saúde apresentam o seguinte *laytout*:

[0]:
```
WHO.head()
```

[0]:
```
       Country  Year  ... Income composition of resources  Schooling
0  Brazil  2015  ...                              0.479       10.1
1  Brazil  2014  ...                              0.476       10.0
2  Brazil  2013  ...                              0.470        9.9
...
   Brazil  2011  ...                              0.454        9.5

[5352 rows x 22 columns]
```

1.4.1 Histograma

Em muitos casos é muito mais importante você conhecer a distribuição de valores de uma variável que unicamente sua estatística descritiva (média, variância etc.). O histograma é um gráfico de distribuição e envolve os valores de uma única variável e sua frequência (quantidade de ocorrências) para esses valores.

A série completa de dados de BMI e série com dados somente do ano de 2010 apresentam ambas uma ***distribuição bimodal***. Isso nos diz que existem dois grupos distintos de países de acordo com o índice médio de massa corpórea, e essa informação seria muito difícil de se obter sem a visualização da distribuição dos dados. É uma informação que, além de relevante, permite abrir outras hipóteses de análise como: *Quais outras características em comum trazem esses dois grupos de países?*.

[0]:
```
sns.distplot(WHO.BMI, color="g")
sns.distplot(WHO2010.BMI, color="b")

# Add titles
plt.title("BMI de Diferentes Países (from WHO)", loc='left', fontsize=18,␣
 ↪fontweight=0, color='gray
plt.show()
```

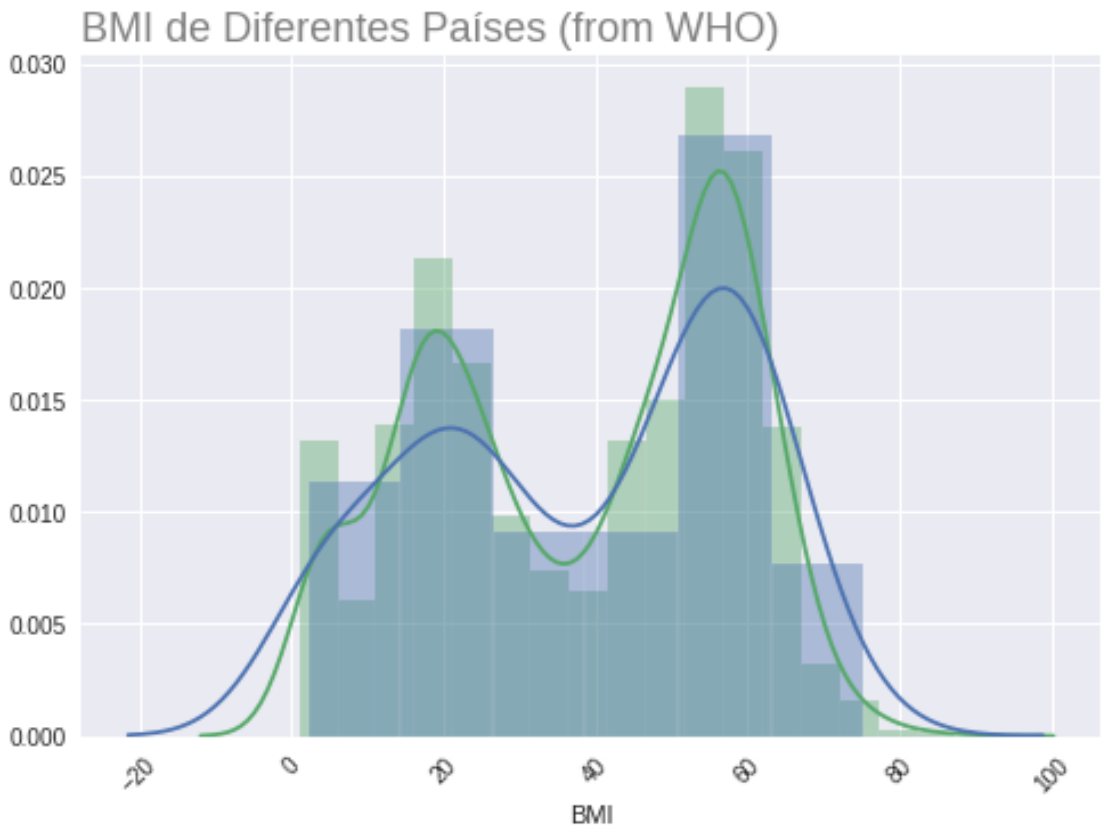

Fig 5: **Distribuição Bimodal** dos índices BMI entre os diferentes Páises indica haver dois grupos bastante distintos de países.

Já a Expectativa de Vida Média não apresenta a mesma distribuição. No gráfico seguinte você poderá observar uma **Distribuição em Calda à Esquerda**. ´

[0]:
```
sns.distplot(WHO.Life_expectancy, kde=True, color="g")
sns.distplot(WHO2010.Life_expectancy, kde=True, color="b")

# Add titles
plt.title("Expectativa de Vida de Diferentes Países (from WHO)",␣
 ↪loc='left', fontsize=18, fontweight=0, color='gray')

plt.xticks(rotation=45)
plt.show()
```

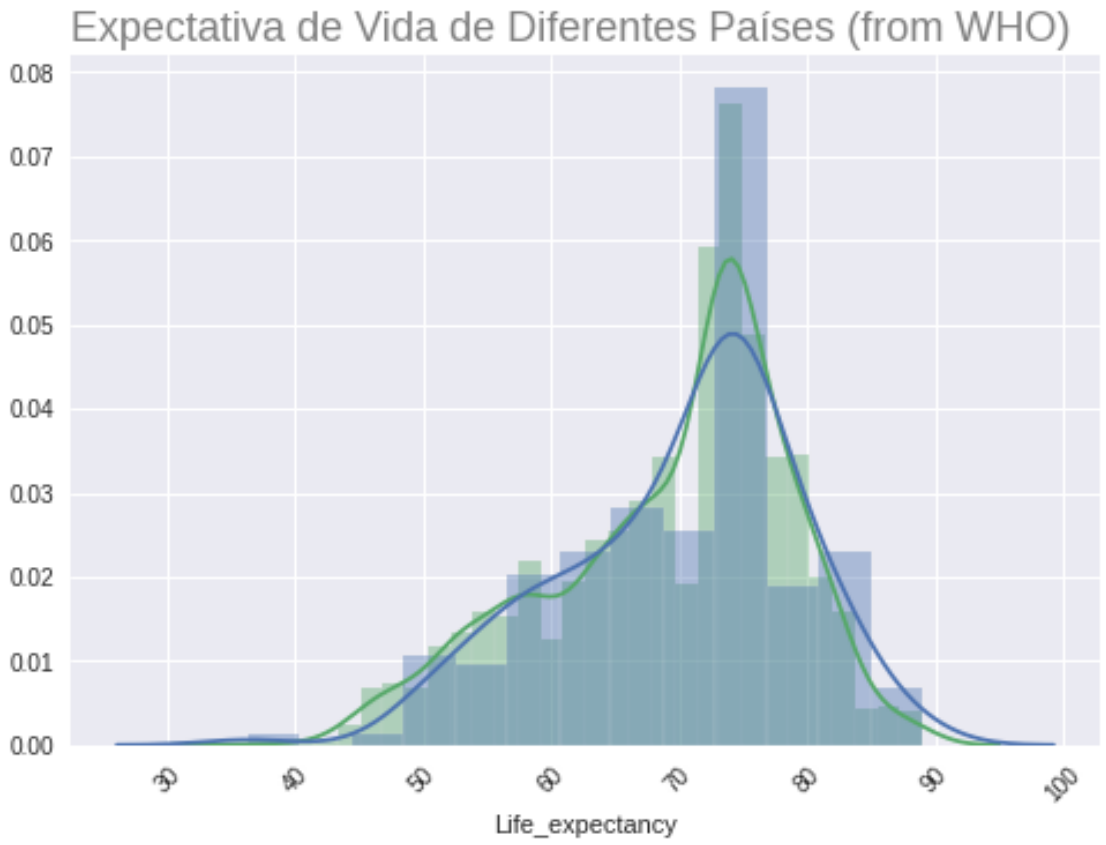

Fig 6: **Distribuição em Calda** da Expectativa de Vida Média. Típica de valores associados a fenômenos de desigualdade.

O que você pode concluir desses gráficos? Embora em termos de índice de massa corpórea os países estejam divididos claramente em dois grupos distintos, em termos da expectativa de vida as expectativas menores (< 75 anos) apresentam um espectro muito maior de valores. Isso é típico de situações de desigualdade e, muito provalvemente, reflete a desigualdade social entre os países.

Mas isso não pode ser inferido somente pelo gráfico acima e precisaria ser verificado com outros dados, como o produto interno ou a renda média de cada país. Isso é o que faremos no capítulo 3 de **Análise de Distribuição** dos dados. Por hora, é importante você entender que esse é um princípio bastante comum na análise de dados: *a cada análise, novas hipóteses podem ser levantadas e requerem novos dados*. Isso é um aspecto básico do método científico e você encontrará muitos exemplos disso neste livro. Não sem motivo, denominamos atualmente essa e outras análises de dados pelo termo comum de **Ciências de Dados**.

1.5 Correlação

Relações entre os dados são muitas vezes a parte mais importante dentre as descobertas que buscamos nos dados. Elas são denominadas de modo geral como correlações. Mas é importante notar que a correlação estatística tem um significado bastante específico e, em geral, é associada a **correlação linear**. Graficamente, entretanto, estamos livres para buscar quaisquer relações entre os dados, sejam elas lineares ou não.

> **Caso 3: Pokémon, winner?** Alguma característica torna um personagem de **Pokémon** mais apto a vencer as batalhas?

Embora possa parecer um conjunto de dados de brinquedo, dados de jogos como o Pokémon compartilham uma série de características com problemas bastante concretos. Cada um dos atributos de um personagem, ou *features*, poderiam ser igualmente entendidos como propriedades de produtos, clientes ou equipamentos. E podemos estar interessados no *melhor* Pokémon e sua melhor característica, do mesmo modo que podemos buscar os melhores produtos, clientes ou suas características.

Os **Gráficos de Dispersão** são normalmente a primeira escolha para a busca dessas relações. Veja agora o conjunto de dados de personagens Pokémons em que buscaremos essas relações:

[0]:
```
pokemon = pd.read_csv('Pokemon.csv',index_col=0)
```

```
[0]:                    Name Type 1  Type 2  ...  Speed  Generation  Legendary
    #                                        ...
    1               Bulbasaur  Grass  Poison  ...     45           1      False
    2                 Ivysaur  Grass  Poison  ...     60           1      False
    3                Venusaur  Grass  Poison  ...     80           1      False
    ...
                  Charmander   Fire     NaN  ...     65           1      False
```

Gráficos múltiplos são bastante empregados para descobrir relações entre os dados. Como não sabemos as relações que existem a ideia é buscarmos relações candidatas e, então, aprofundar a análise sobre elas. Gráficos como `pairplot` estão presentes em várias linguagens e ferramentas para visualização de dados. Eles permitem exibir simultaneamente vários Gráficos de Dispersão de diferentes pares de variáveis e buscarmos relações de interesse.

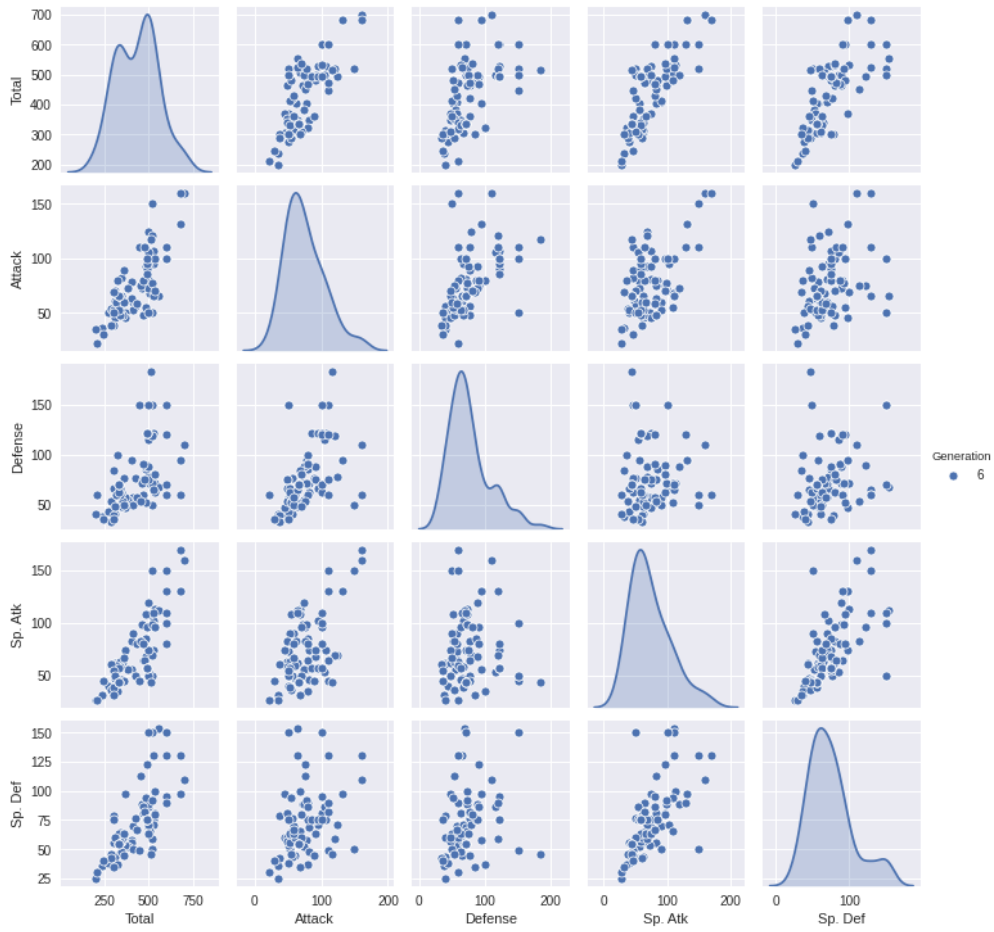

Fig 7: **Pairplots** exibem relações candidatas entre diversos pares de variáveis.

```
[0]: sns.pairplot(data=pokemon[pokemon.Generation == 6][['Generation',
'Sp. Atk', 'Sp. Def' ]],hue='Generation',aspect=1, height=2)
```

Se você já jogou Pokémon sabe que é um jogo bastante simples e que, ao final, o valor Total é em geral reponsável pela vitória do personagem. Assim, dentro da pergunta inicial buscarmos características que tornam um Pokémon mais apto podemos:

Quais pares de atributos apresentam correlação (relação linear) maior?

Não, não é evidente, mas isso é bastante próximo do que encontramos em casos reais (e este é um caso real!). Correlações claramente lineares, acima de 0.9, são bastante raras de serem encontradas. Mas se você considerar problemas complexos, em que várias variáveis contribuem para explicar o valor umas das outras, encontrar relações lineares de 0.6 ou 0.7 é bastante relevante.

Se você respondeu 'Sp. Atk' (*Speed Attack*) à pergunta anterior você acertou. No gráfico a seguir 'Total' e 'Sp. Atk.' apresentam uma linha de regressão com valor de correlação linear de 0.79!

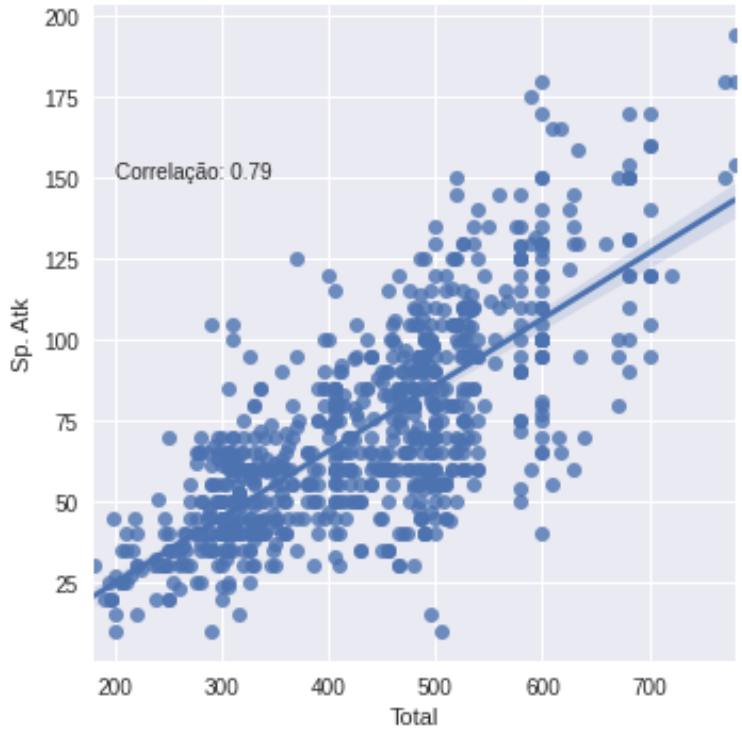

Fig 8: Dentre outras relações, relações lineares entre os dados são particularmente importantes.

[0]:
```
sns.lmplot(x='Total',y='Sp. Atk',data=pokemon)
plt.text(200,150,'Correlação: '+ str( round(pokemon[pokemon.Generation ==␣
↪6].Total.corr(pokemon[pokemon.Generation == 6]['Sp. Atk']),2)  ))
```

Embora tenhamos tratado de relações lineares, Gráficos de Dispersão permitem exibir quaisquer outras relações entre os dados. Um relação simples na forma de $y = x^2$ pode apresentar $corr(x, y) = 0$ e, mesmo assim, será claramente exibida em um Gráfico de Dispersão. Por isso eles desempenham um papel fundamental na descoberta de relações entre os dados.

1.6 Sumário

Aqui você pode entender a importância da **Visualização dos Dados** para busca de padrões, relações e processos em ação sobre os dados. Muitas dessas informações dificilmente poderiam ser encontradas de outra forma. Ela também tem um papel importante na comunicação de resultados.

O domínio de linguagens e ferramentas, como o `Python`, é importante para que você obtenha resultados na Visualização de Dados. Mas isso é um aspecto secundário quando comparado ao aspecto analítico da Visualização de Dados, ou o *Pensamento Analítico* sobre o dados. Em primeiro lugar é necessário saber quê perguntas são de interesse sobre os dados (evolução, distribuição, *ranking*, correlações ou partes de um todo) e como você pode visualizar as respostas a essas questões. Feito isso, você poderá então buscar como produzir a visualização desejada dos dados. Seu *Pensamento Analítico* deve ainda seguir princípios científicos onde, a cada análise de uma hipótese, seguiram novas hipóteses a serem analisadas, em um processo de refinamento contínuo.

Referências

Data Visualization: A Practical Introduction (2019) by Kieran Healy

Making Data Visual: A Practical Guide to Using Visualization for Insight (2017) by Danyel Fisher, Miriah Meyer

The Visual Display of Quantitative Information Hardcover (2001) by Edward R. Tufte

Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures (2019)
by Claus O. Wilke

Storytelling with Data: A Data Visualization Guide for Business Professionals Paperback (2015) by Cole Nussbaumer Knaflic

Python Data Visualization: An Easy Introduction to Data Visualization in Python with Matplotlip, Pandas, and Seaborn (2019) by Samuel Burns

Cronograma

Visualização de Dados com Exemplos em Python

[Jun 2020] Início dos trabalhos

Apresentação

1. Introdução

2. Evolução ou Tendência dos Dados
> gráficos de linha, área, séries múltiplas

3. Análise de Distribuição
> histogramas, gráficos de distribuição de densidade, *boxplot*

[Ago 2020] Início do capítulo 4

4. *Ranking*
> gráficos de barras, *word cloud*, *spider*

5. Correlações dos Dados
> gráficos de dispersão, *heat map*, *density 2D*

[Nov 2020] Até capítulo 6

6. Partes de um todo
> *Tree map*, diagramas Venn, *pie chart*

7. Conclusão

Referências

Apêndice 1. *Python Essencial*

Apêndice 2. *Recursos*

[Fev 2021] Conclusão e envio do Original

Glossário

Termos: Visualização de Dados; Ciência de Dados; Análise de Dados; Suporte à Decisão; *Big Data*; Aprendizado de Máquina; *Machine Learning*; Estatística; *Python; Matplotlib; Seaborn; Pandas*.