

# **Análise de Dados**

Sergio Pedro Rodrigues Oliveira

10 January 2025

# SUMÁRIO

<b>1</b>	<b>PRINCIPAIS TÓPICOS</b>	<b>1</b>
<b>2</b>	<b>DADOS</b>	<b>2</b>
2.1	O que são dados? . . . . .	2
2.2	Informação . . . . .	2
2.3	Tabela banco de dados termos . . . . .	2
2.4	Teoria . . . . .	3
2.5	Tipos de variáveis . . . . .	3
<b>3</b>	<b>Estatística Básica (Teoria medidas de posição e dispersão)</b>	<b>4</b>
3.1	Preparação dos dados para aplicação de estatística básica . . . . .	4
3.1.1	Teoria . . . . .	4
3.1.2	Preparação dos dados (sumariar dados coletados) . . . . .	7
3.1.2.1	Variável Quantitativa Discreta . . . . .	8
3.1.2.2	Variável Quantitativa Contínua . . . . .	9
3.1.2.3	Variáveis Qualitativas . . . . .	14
3.2	Medidas de posição . . . . .	14
3.3	Medidas de dispersão . . . . .	14
3.4	Análise Estatística . . . . .	15
<b>4</b>	<b>EXCEL</b>	<b>16</b>
4.1	Ferramentas do Excel . . . . .	16
4.2	Filtro Excel . . . . .	16
4.3	Tabela dinâmica . . . . .	16
4.4	Gráficos . . . . .	16
4.5	Bloco if-else - SE() . . . . .	17
4.6	Cruzar dados . . . . .	17
<b>5</b>	<b>PRIMEIRA ANÁLISE - PYTHON</b>	<b>18</b>
5.1	Imprimir na tela - print() . . . . .	18
5.2	Variável . . . . .	18
5.3	Bloco IF-ELSE . . . . .	18
5.4	Alguns comandos de powershell do Windows . . . . .	19
5.4.1	Comandos Básicos . . . . .	19
5.4.2	Alterar restrições de segurança . . . . .	19
5.4.3	Serviços e Processos . . . . .	19

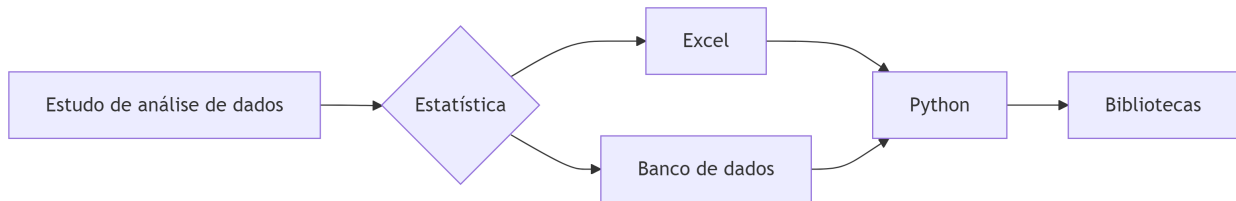
## LISTA DE FIGURAS

1	Estatística descritiva. . . . .	4
2	Tipos de variáveis. . . . .	6
3	Distribuição tabular quantitativo discreta. . . . .	8
4	Distribuição de frequências em classes. . . . .	9
5	Intervalo de classes, para distribuição de frequência quantitativa contínua. . . .	9
6	Premissas da distribuição de frequências quantitativa contínua. . . . .	10
7	Tabela de distribuição de frequência quantitativa contínua. . . . .	12

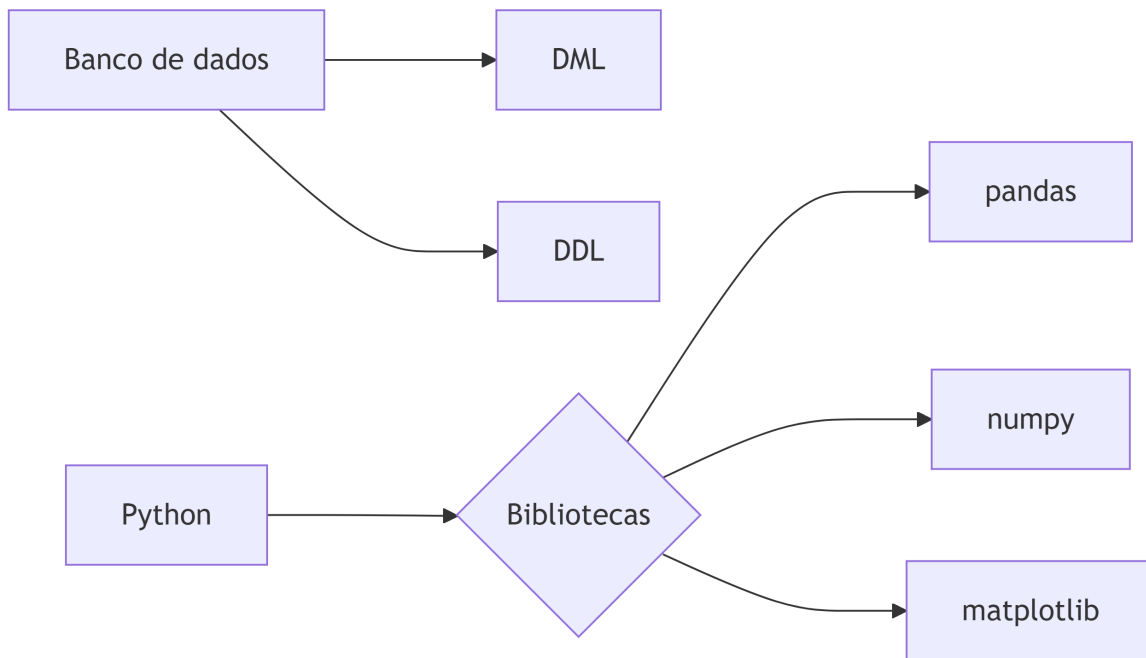
## LISTA DE TABELAS

1	Bnaco de dados nomeclaturas . . . . .	2
2	Comandos básicos de powershell . . . . .	19
3	Comandos para alterar restrições de segurança powershell . . . . .	19
4	Comandos de serviços e processos powershell . . . . .	19

# 1 PRINCIPAIS TÓPICOS



Plano de estudo de análise de dados



Destrinchando tópicos

## 2 DADOS

### 2.1 O que são dados?

Dados são valores brutos atribuídos a algo.

### 2.2 Informação

- Informação é a ordenação e organização dos dados de forma a transmitir significado e compreensão dentro de um contexto.
- Informação é o significado que a gente obtém a partir dos dados.
- Informação = fazer perguntas para os dados (responder pergunta).
- Nem sempre podemos confiar nos dados, é preciso entender o contexto dos dados.
  - De onde eles vem?
  - Quem são as pessoas que responderam?
  - O que são esses dados?

### 2.3 Tabela banco de dados termos

```
from IPython.display import Markdown
from tabulate import tabulate
table = [["Coluna(s)", "Campo(s)"],
         ["Linha(s)", "Registro(s)"]]
Markdown(tabulate(
    table,
    headers=["Nomeclatura", "Nomeclatura técnica"]
))
```

Table 1: Bnaco de dados nomeclaturas

Nomeclatura	Nomeclatura técnica
Coluna(s)	Campo(s)
Linha(s)	Registro(s)

## 2.4 Teoria

- Análise descritiva dos dados através da tabulação das variáveis e cálculo de medidas descritivas (média, desvio-padrão, etc).
- Análise descritiva dos dados (Informações preliminares):
  - Contagem dos resultados observados em cada variável do conjunto de dados.
  - Natureza descritiva dos dados, tipo de variáveis (categórica ou numérica).
  - Três objetivos principais:
    - \* Verificar erros e anomalias.
    - \* Compreender a distribuição de cada uma das variáveis isoladamente.
    - \* Compreender a natureza e a força das relações entre as variáveis.
- Após essas etapas, estabelecer um modelo estatístico formal e relatar suas conclusões.

## 2.5 Tipos de variáveis

- Variável numérica:
  - Continua  
Se seus valores pertencer ao conjunto dos números reais.  
Ex.: Temperatura corporal, saldo em caixa, peso da carga de um caminhão, etc.
  - Discreta  
Se seus valores pertencer ao conjunto dos números inteiros.  
Ex.: Número de pessoas com febre, número de empresas, número de caminhões, etc.
- Variável categórica:
  - Ordinal  
Se seus valores podem ser ordenados do menor para o maior.  
Ex.: Temperatura (baixa, média ou alta), saldo em caixa (negativo, nulo ou positivo), etc.
  - Nominal  
Quando não for possível estabelecer ordenamento.  
Ex.: Sexo do indivíduo, atividade fim da empresa, marca/modelo do caminhão, etc.

### 3 Estatística Básica (Teoria medidas de posição e dispersão)

#### 3.1 Preparação dos dados para aplicação de estatística básica

##### 3.1.1 Teoria

- Definição de Estatística:

A Estatística de uma maneira geral compreende aos métodos científicos para **COLETA**, **ORGANIZAÇÃO**, **RESUMO**, **APRESENTAÇÃO** e **ANÁLISE** de Dados de Observação (Estudos ou Experimentos), obtidos em qualquer área de conhecimento. A finalidade é a de obter conclusões válidas para tomada de decisões.

- Estatística Descritiva

Parte responsável basicamente pela **COLETA** e **SÍNTESE** (Descrição) dos Dados em questão.

Disponibiliza de técnicas para o alcance desses objetivos. Tais Dados podem ser provenientes de uma **AMOSTRA** ou **POPULAÇÃO**.

- Estatística Inferencial

É utilizada para tomada de decisões a respeito de uma população, em geral fazendo uso de dados de amostrais.

Essas decisões são tomadas sob condições de **INCERTEZA**, por isso faz-se necessário o uso da **TEORIA DA PROBABILIDADE**.

- O fluxograma da estatística descritiva pode ser espesso da seguinte forma:



Figure 1: Estatística descritiva.

- A representação tabular (Tabelas de Distribuição de Frequências) deve conter:



- Cabeçalho

Deve conter o suficiente para que as seguintes perguntas sejam respondidas “**o que?**” (Relativo ao fato), “**onde?**” (Relativo ao lugar) e “**quando?**” (Correspondente à época).

- Corpo

É o lugar da Tabela onde os dados serão registrados. Apresenta colunas e sub colunas.

- Rodapé

Local destinado à outras informações pertinentes, por exemplo a Fonte dos Dados.

- População e Amostras

- População

É o conjunto de todos os itens, objetos ou pessoas sob consideração, os quais possuem pelo menos uma característica (Variável) em comum. Os elementos pertencentes à uma População são denominados “Unidades Amostrais”.

- Amostras

É qualquer subconjunto (não vazio) da População. É extraída conforme regras pré-estabelecidas, com a finalidade de obter “estimativa” de alguma Característica da População.

- Tipos de variáveis

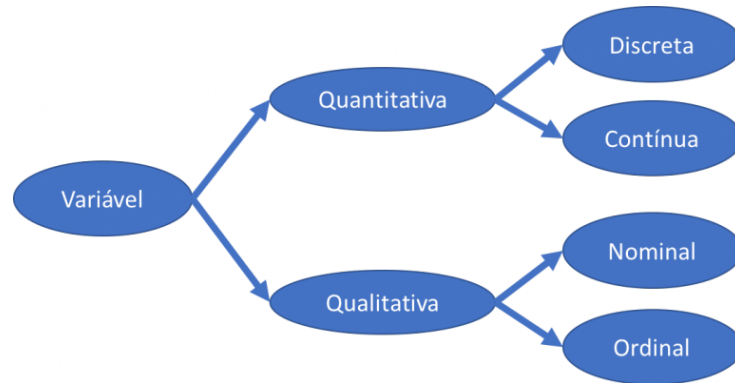


Figure 2: Tipos de variáveis.

- *Qualitativo nominal:*  
Não possuem uma ordem natural de ocorrência.
- *Qualitativo ordinal:*  
Possuem uma ordem natural de ocorrência.
- *Quantitativo discreta:*  
Só podem assumir valores inteiros, pertencentes a um conjunto finito ou enumerável.
- *Quantitativo contínua:*  
Podem assumir qualquer valor em um determinado intervalo da reta dos números reais.

### 3.1.2 Preparação dos dados (sumariar dados coletados)

- Frequência (conceito):  
É a quantidade de vezes que um valor é observado dentro de um conjunto de dado.
- Distribuição em frequências:
  - A distribuição tabular é denominada: “Tabela de Distribuição de Frequências”.
  - Podemos separar em 3 modelos de distribuição tabular:
    - \* Variável Quantitativa Discreta.
    - \* Variável Quantitativa Contínua.
    - \* Variáveis Qualitativas.

### 3.1.2.1 Variável Quantitativa Discreta

- Passos da preparação dos dados:
  - 1º Passo - **DADOS BRUTOS**:  
Obter os dados da maneira que foram coletados.
  - 2º Passo - **ROL**:  
Organizar os DADOS BRUTOS em uma determinada ordem (crescente ou decrescente).
  - 3º Passo - **CONSTRUÇÃO TABELA**:  
Na primeira coluna são colocados os valores da variável, e nas demais as respectivas frequências.
    - Frequência absoluta simples (Nº de vezes que cada valor da variável se repete).
- Principais campos da **distribuição tabular de variáveis quantitativas discreta**:
  - $n$  é o número total de elementos da amostra.
  - $x_i$  é o número de valores distintos que a variável assume.
  - $F_i$  é a Frequência Absoluta Simples.
  - $f_i$  é a Frequência Relativa Simples.
  - $f_i\%$  é a Frequência Relativa Simples Percentual.  $f_i\% = f_i \cdot 100\%$ .
  - $F_a$  é a Frequência Absoluta Acumulada.

$x_i$	$F_i$	$f_i$	$f_i\%$	$F_a \downarrow$	$F_a \uparrow$	$f_a \downarrow$	$f_a \uparrow$
0	6	0,2	20	6	30	0,2	1
1	11	0,37	37	17	24	0,57	0,8
2	8	0,27	27	25	13	0,84	0,43
3	2	0,07	7	27	5	0,91	0,16
4	2	0,06	6	29	3	0,97	0,09
6	1	0,03	3	30	1	1	0,03
Total	30	1	100	-	-	-	-

Figure 3: Distribuição tabular quantitativo discreta.

Obs.: As setas simbolizam ordem crescente ou decrescente.

### 3.1.2.2 Variável Quantitativa Contínua

- Teoria:
  - A construção da representação tabular é realizada de maneira análoga ao caso das variáveis discretas.
  - As frequências são agrupadas em classes, denominadas de “Classes de Frequência”.
  - Denominada “Distribuição de Frequências em Classes” ou “Distribuição em Frequências Agrupadas”.

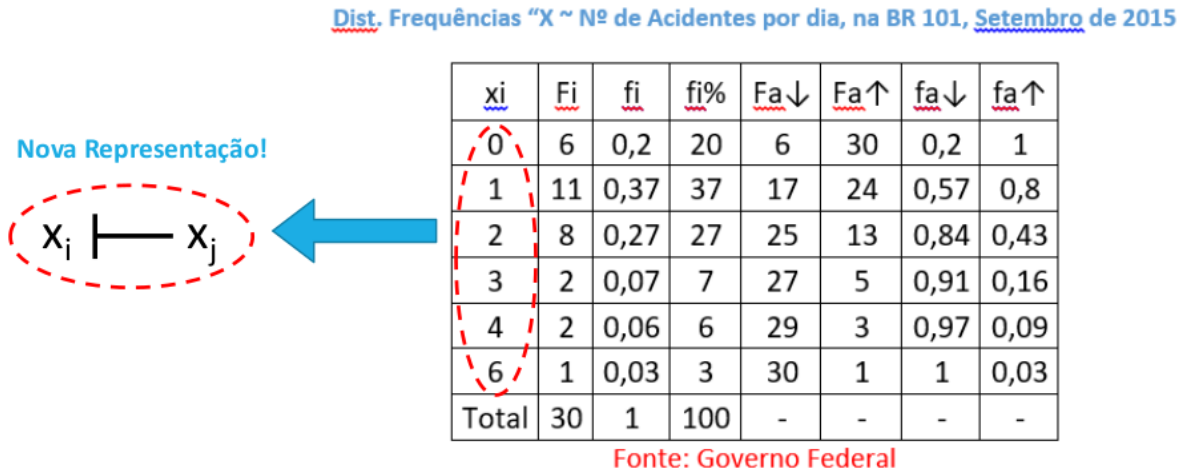


Figure 4: Distribuição de frequências em classes.

- Convencionar o tipo de intervalo para as classes de frequência:
  - Intervalo “exclusive – exclusive”:  $x_i \text{ — } x_j$
  - Intervalo “inclusive – exclusive”:  $x_i \text{ — } x_j$
  - Intervalo “inclusive – inclusive”:  $x_i \text{ — } x_j$
  - Intervalo “exclusive – inclusive”:  $x_i \text{ — } x_j$

OBS.:  $x_i$  - Limite Inferior (LI) de Classe;  
 $x_j$  - Limite Superior (LS) de Classe;

Figure 5: Intervalo de classes, para distribuição de frequência quantitativa contínua.

## Premissas


- 
- i) As classes têm que ser exaustivas, isto é, todos os elementos devem pertencer a alguma classe;
  - ii) As classes têm que ser mutualmente exclusivas, isto é, cada elemento tem que pertencer a uma única classe

Figure 6: Premissas da distribuição de frequências quantitativa contínua.

Passos para contruir a **Tabela Distribuição de Frequências Contínua**:

1. Como estabelecer o **número de classes** ( $k$ ):

- Normalmente varia de 5 a 20 classes.
- Critério fórmula de Sturges:

$$k \cong 1 + 3,3 \cdot \log(n)$$

- Critério da Raiz quadrada:

$$k \cong \sqrt{n}$$

Onde  $n$  é o número de elementos amostrais.

1. Como calcular a **Amplitude Total** ( $AT_x$ ):

- Diferença entre o maior e o menor valor observado.
- Intervalo de variação dos valores observados.
- Aproximar valor calculado para múltiplo do n<sup>o</sup> classes ( $k$ ).
- Garantir inclusão dos valores mínimo e máximo.
- Cálculo:

$$AT_x = Mx(X_i) - Mn(X_i)$$

Onde,

$AT_x$  é a Amplitude Total.

$Mx(X_i)$  é o *valor máximo das amostras*.

$Mn(X_i)$  é o *valor mínimo das amostras*.

- Exemplo:

Se  $k = 5$ ,

$$AT_x = 28$$

Logo, arredondando  $AT_x = 30$ , para aproximar o valor  $AT_x$  de um múltiplo de  $k$ .

1. Como calcular a **Amplitude das classes da frequência** ( $h$ ):

- As classes terão amplitudes iguais.
- Cálculo:

$$h = h_i = \frac{AT_x}{k}$$

Onde,  $k$  é o **número de classes** e  $AT_x$  é a **Amplitude Total**.

1. Como determinar o ponto médio das classes, representatividade da classe ( $p_i$ ):

$$p_i = \frac{(LS_i - LI_i)}{2}$$

Onde,

$LS_i$  é o limite superior da classe.

$LI_i$  é o limite inferior da classe.

2. Passos da preparação dos dados:

- 1º Passo - **DADOS BRUTOS**:

Obter os dados da maneira que foram coletados.

- 2º Passo - **ROL**:

Organizar os DADOS BRUTOS em uma determinada ordem (crescente ou decrescente).

- 3º Passo - **CONSTRUÇÃO TABELA**:

Na primeira coluna são colocados as classes, e nas demais as respectivas frequências.

- Exemplo:

Nº Classe	Classes (xi)	Fi	fi	fi%	Fa↓	Fa↑	fa↓	fa↑	fa↓%	pi
1	45  --- 52	3	0,08	8	3	40	0,08	1	100	48,5
2	52  --- 59	7	0,18	18	10	37	0,26	0,92	92	55,5
3	59  --- 66	11	0,28	28	21	30	0,53	0,75	75	62,5
4	66  --- 73	10	0,25	25	31	19	0,78	0,47	47	69,5
5	73  --- 80	4	0,10	10	35	9	0,88	0,22	22	76,5
6	80  --- 87	4	0,10	10	39	5	0,98	0,12	12	83,5
7	87  --- 94	1	0,02	2	40	1	1,00	0,02	2	90,5
Total		40	1,00	100	-	-	-	-		-

Fonte: Dados Fictícios

Figure 7: Tabela de distribuição de frequência quantitativa contínua.

$X_i$  são as classes.

$F_i$  é a Frequência Absoluta Simples.

$f_i$  é a Frequência Relativa Simples.

$f_i\%$  é a Frequência Relativa Simples Percentual.

$F_a$  é a Frequência Absoluta Acumulada.



$f_a$  é a Frequência Absoluta Acumulada Simples.

$f_a \%$  é a Frequência Absoluta Acumulada Simples Percentual.

$p_i$  é a Representatividade da classe (ponto médio das classes).

#### **3.1.2.3 Variáveis Qualitativas**

### **3.2 Medidas de posição**

### **3.3 Medidas de dispersão**

### 3.4 Análise Estatística

- Para fazer uma Análise Estatística eficiente de dados, precisamos:

- Limpar os dados:

Remover os *OUTLIER* (valores atípicos, inconsistentes).

- Aplicar Estatística Descritiva aos dados:

As medidas de posição (**Média**, **Mediana** e **moda**) e dispersão (**Amplitude Total**, **Desvio**, **Desvio Médio**, **Variância**, **Desvio-padrão** e **Coefficiente de Variação**) são maneiras de descrever os dados.

- Comparar as medidas dos dados:

Principalmente medidas de dispersão, me especial **Coefficiente de Variação**, são ótimas para comparar dados.

- Previsão de dados:

A principal técnica é de **Regressão**, porém para aplicar, necessita que os dados estejam limpos e com pouca dispersão (quanto menor, melhor).

## 4 EXCEL

### 4.1 Ferramentas do Excel

Algumas ferramentas do Excel que podem ajudar na análise da dados:

- Filtro
- Tabela (tabela dinâmica)
- Gráficos

### 4.2 Filtro Excel

- Inserir filtros na primeira linha (campo):
  - Célula na primeira linha
  - Aba “Dados” > “Classificar e Filtrar” > “Filtro”

### 4.3 Tabela dinâmica

- Inserir Tabela dinâmica:
  - Selecionar toda tabela;
    - \* Selecionar primeira célula (“A1”);
    - \* Comandos: CTRL + SHIFT + ↓ + →;
  - Aba “Inserir” > “Tabelas” > “Tabela dinâmica”;
  - Opção “Nova planilha”.
- Agrupar informações com tabela dinâmica:
  - Linha/Registro: informação que queremos;
  - Valores: Normalmente registros únicos (primary key, exemplo: “ID”).

### 4.4 Gráficos

- Criar um gráfico rápido com base na tabela dinâmica:
  - Clickar na tabela dinâmica criada;
  - Aba “Inserir” > “Gráficos” > “Gráficos recomendados”.

## 4.5 Bloco if-else - SE()

- Podemos usar o bloco **if-else** no Excel usando a função **SE()**.
- Na função **SE()**, usamos como argumentos:
  - Expressão a ser avaliada;
  - Ação caso verdadeira;
  - Ação caso falso.
- Para usar uma **função no Excel** na barra de fórmulas inserimos o sinal de = antes da expressão/função para o Excel saber que se trata de uma expressão.

Exemplo:

=SE(\$T2="TRUE"; "Pessoa Gestora"; \$X2)

## 4.6 Cruzar dados

- Cruzar dados usando uma tabela dinâmica no Excel:

Podemos cruzar os dados usando tabela dinâmica escolhendo cuidadosamente as informações que estarão contidas nas colunas e nas linhas, assim cruzando as informações.
- Podemos filtrar os dados apresentados nas linhas e/ou colunas para melhor visualizar a informação.

## 5 PRIMEIRA ANÁLISE - PYTHON

### 5.1 Imprimir na tela - print()

O comando `print()` imprime na tela uma mensagem.

Exemplo:

```
print("Hello world!")
```

Hello world!

### 5.2 Variável

### 5.3 Bloco IF-ELSE

## 5.4 Alguns comandos de powershell do Windows

### 5.4.1 Comandos Básicos

Table 2: Comandos básicos de powershell

Comandos	Explicação
cd	Navegar pelas pastas.
dir	Listar arquivos de uma pasta.
cls ou clear	Limpa a tela.
Get-Help <comando>	Obter ajuda com relação a um comando do powershell.
mkdir <nome_diretório>	Criar um novo diretório.
'Copy-Item -Path "caminho com extensão" -Destination "caminho destino"'	Copiar arquivo ou diretório.
Remove-Item "<caminho com extensão>"	Excluir arquivo ou diretório.

### 5.4.2 Alterar restrições de segurança

Table 3: Comandos para alterar restrições de segurança powershell

Comandos	Explicação
Get-ExecutionPolicy -List	Para exibir as políticas de execução para cada escopo.
Set-ExecutionPolicy Unrestricted	Permite executar todo e qualquer script.
Set-ExecutionPolicy All Signed	Todos os scripts devem ser assinados por alguém confiável.
Set-ExecutionPolicy Remote Signed	Todos os scripts que forem baixados da Internet devem ser assinados por alguém confiável.
Set-ExecutionPolicy Restricted	Não permite a execução de nenhum script.

### 5.4.3 Serviços e Processos

Table 4: Comandos de serviços e processos powershell

Comandos	Explicação
Get-Service	Lista de serviços em execução.
Start-Service <nome do serviço>	Iniciar serviço.
Stop-Service <nome do serviço>	Parar serviço.
Suspend-Service <nome do serviço>	Suspender serviço.
Restart-Service <nome do serviço>	Reiniciar serviço.