

# Aula de hoje

1



Aula sobre:  
**Exploratória de  
Variáveis Qualitativas**



1º: uso do notebook  
**Aula03\_Atividade**

2º: uso do notebook  
**Aula03\_Exercício**



Insper

# **Ciência dos dados**

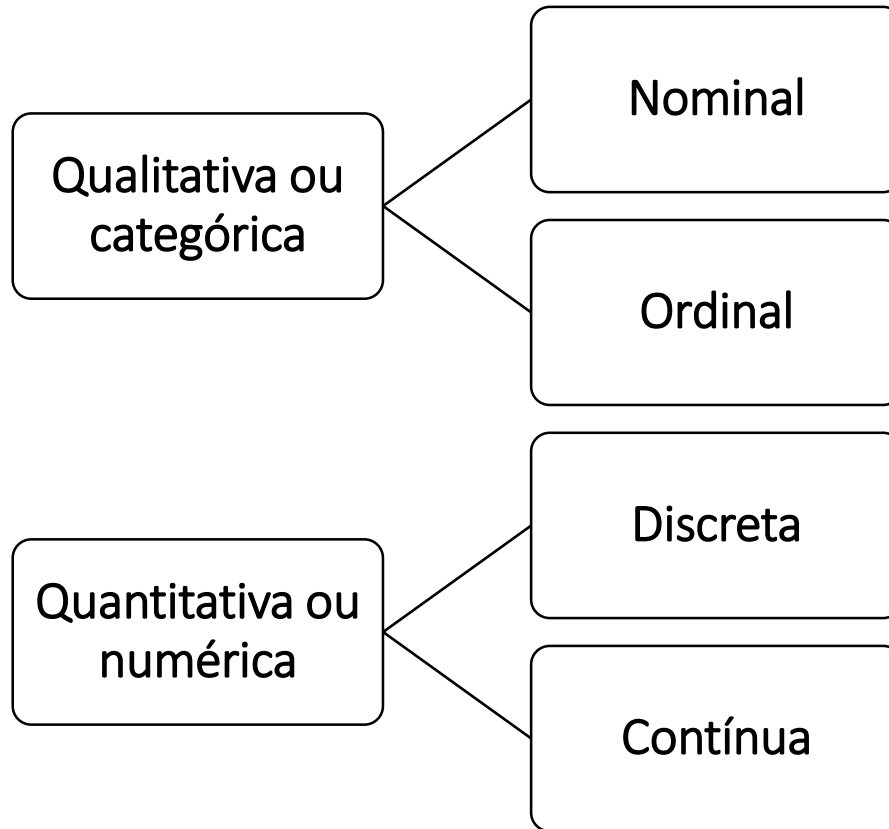
**Tipos de Variáveis e  
Medidas Resumo**

# Aula de hoje

Ao final desta aula, o aluno deve ser capaz de:

- Selecionar informações de bancos de dados, tratá-los e prepará-los para processamento
- Elaborar análises exploratórias de dados (univariadas e multivariadas), utilizando ferramentas estatísticas e computacionais adequadas.

# Tipos de variáveis



# Qualitativas ou categóricas

- **Nominais:**

não é possível estabelecer uma ordem natural entre os valores.

Ex.: Refrigerantes: “Mate Couro”, “Gengibirra”, “Itubaína”,  
“Laranjinha”;

Turma: A, B e C.

- **Ordinais**

é possível estabelecer uma ordem natural entre as categorias.

Ex.: Tamanho: Pequeno, Médio, Grande;

Classe social: Baixa, Média, Alta.

# Quantitativas

- **Discretas:**

Provenientes, usualmente, de contagens.

Ex.: número de membros de uma família; número de kb de um arquivo; número de conexões em uma rede

- **Contínuas:**

Provenientes de mensuração

Ex.: massa (em kg) de um produto;  
comprimento (em km) de uma estrada; etc

# Vamos praticar usando o *dataset* da aula anterior?

7

	Country	Population	GDPcapita	surface	region	landlocked
0	Albania	2901883	1915.424459	28750	europa_east	0
1	Algeria	36036159	2231.980246	2381740	africa_north	0
2	Angola	21219954	623.245275	1246700	africa_sub_saharan	0
3	Antigua and Barbuda	87233	10614.794315	440	america_north	0
4	Argentina	41222875	10749.319224	2780400	america_south	0

# Recursos de análise / visualização:

8

Objetivo	Estratégia
Qualitativa (Categórica)	<b>Tabela de frequências univariada ou cruzada</b> Gráficos de setor Gráficos de barra
Quantitativa	<b>Tabela de frequências</b> <b>Média, Mediana, Quartis, Amplitude, Desvio Padrão, entre outras medidas resumo</b> Histograma Boxplot



# Tipos de dados e conversões

## Tipos de dados disponíveis (extraído do link abaixo)

- `object`: utilizado para *strings* (isto é, sequências de caracteres)
- `int64`: usado para números inteiros
- `float64`: usado para números de *ponto-flutuante* (isto é, decimais e frações)
- `bool`: usado somente para valores True ou False
- `datetime64`: usado para valores relativos a datas
- `timedelta`: usado para representar a diferença entre datas
- `category`: usado para valores que usam um de um número limitado de opções disponíveis (não é mandatório, mas categorias podem ter ordenamento explícito)

Fonte: <https://www.vooo.pro/insights/guia-de-funcionalidades-avancadas-do-pandas/> extraído em 26/08/2020

# Conversões com Pandas

10

## Conversion

<code>Series.astype(self, dtype[, copy, errors])</code>	Cast a pandas object to a specified dtype dtype.
<code>Series.infer_objects(self)</code>	Attempt to infer better dtypes for object columns.
<code>Series.copy(self[, deep])</code>	Make a copy of this object's indices and data.
<code>Series.bool(self)</code>	Return the bool of a single element PandasObject.
<code>Series.to_numpy(self[, dtype, copy])</code>	A NumPy ndarray representing the values in this Series or Index.
<code>Series.to_period(self[, freq, copy])</code>	Convert Series from DatetimeIndex to PeriodIndex with desired frequency (inferred from index if not passed).
<code>Series.to_timestamp(self[, freq, how, copy])</code>	Cast to DatetimeIndex of Timestamps, at <i>beginning</i> of period.
<code>Series.to_list(self)</code>	Return a list of the values.
<code>Series.get_values(self)</code>	(DEPRECATED) Same as values (but handles sparseness conversions); is a view.
<code>Series.__array__(self[, dtype])</code>	Return the values as a NumPy array.

Fonte: <https://pandas.pydata.org/pandas-docs/version/0.25.3/reference/series.html#conversion> extraído em 26/08/2020

# Accessors com Pandas

## Accessors

Pandas provides dtype-specific methods under various accessors. These are separate namespaces within `Series` that only apply to specific data types.

Data Type	Accessor
Datetime, Timedelta, Period	<code>dt</code>
String	<code>str</code>
Categorical	<code>cat</code>
Sparse	<code>sparse</code>

Fonte: <https://pandas.pydata.org/pandas-docs/version/0.25.3/reference/series.html#accessors> extraído em 26/08/2020

# Categorical Accessors com Pandas

12

## Categorical accessor

Categorical-dtype specific methods and attributes are available under the `Series.cat` accessor.

<code>Series.cat.categories</code>	The categories of this categorical.
<code>Series.cat.ordered</code>	Whether the categories have an ordered relationship.
<code>Series.cat.codes</code>	Return Series of codes as well as the index.
<code>Series.cat.rename_categories(self, *args, ...)</code>	Rename categories.
<code>Series.cat.reorder_categories(self, *args, ...)</code>	Reorder categories as specified in <code>new_categories</code> .
<code>Series.cat.add_categories(self, *args, ...)</code>	Add new categories.
<code>Series.cat.remove_categories(self, *args, ...)</code>	Remove the specified categories.
<code>Series.cat.remove_unused_categories(self, ...)</code>	Remove categories which are not used.
<code>Series.cat.set_categories(self, *args, ...)</code>	Set the categories to the specified <code>new_categories</code> .
<code>Series.cat.as_ordered(self, *args, **kwargs)</code>	Set the Categorical to be ordered.
<code>Series.cat.as_unordered(self, *args, **kwargs)</code>	Set the Categorical to be unordered.

Fonte: <https://pandas.pydata.org/pandas-docs/version/0.25.3/reference/series.html#api-series-cat> extraído em 26/08/2020

# Empresa de TV

Uma empresa de TV via satélite criou recentemente dois tipos de planos de canais (A e B).

A empresa tem como objetivo estudar o perfil dos clientes que aderiram ao plano para enviar malas diretas aos potenciais clientes de cada tipo de plano.

A base de dados apresenta algumas variáveis para uma amostra de 82 clientes selecionados aleatoriamente dentre aqueles que aderiram aos planos. As variáveis têm os seguintes significados:

- CLIENTE: identificador do cliente.
- PLANO: apresenta o plano adquirido pelo cliente – (1=A ou 2=B).
- EC: apresenta estado civil do cliente no momento da adesão ao plano – (1=Casado, 2=Solteiro e 3=Outros).
- SATISFACAO: grau de satisfação do cliente pelo plano – (Muito insatisfeito, Insatisfeito, Indiferente, Satisfeito e Muito satisfeito).
- RENDA: renda pessoal do cliente, em milhares de reais.

O arquivo `EmpresaTV_Cod.xlsx` contém as variáveis descritas acima.

# Tabela de frequências para variável qualitativa 14

## Frequências absolutas por PLANO:

A	46
B	36

## Frequências absolutas por ESTADO CIVIL:

Casado	36
Solteiro	33
Outros	13

## Frequências absolutas por SATISFACAO:

Muito Insatisfeito	8
Insatisfeito	16
Indiferente	19
Satisfeito	27
Muito Satisfeito	12

**Comando Python:**  
`variável.value_counts()`

# Tabela de frequências para variável qualitativa 15

## Frequências **relativas** por PLANO:

A	56,1
B	43,9

## Frequências **relativas** por ESTADO CIVIL:

Casado	43,9
Solteiro	40,2
Outros	15,9

## Frequências **relativas** por SATISFACAO:

Muito Insatisfeito	9,8
Insatisfeito	19,5
Indiferente	23,2
Satisfeito	32,9
Muito Satisfeito	14,6

**Comando Python:**  
`variável.value_counts(True)*100`

# Tabela de frequências cruzada entre duas variáveis qualitativas

- PLANO A

ESTADO CIVIL	Casado	Solteiro	Outros	All
SATISFAÇÃO				
Muito Insatisfeito	4.0	0.0	0.0	4.0
Insatisfeito	4.0	0.0	7.0	11.0
Indiferente	7.0	7.0	2.0	15.0
Satisfeito	30.0	9.0	4.0	43.0
Muito Satisfeito	11.0	13.0	2.0	26.0
All	57.0	28.0	15.0	100.0

- PLANO B

ESTADO CIVIL	Casado	Solteiro	Outros	All
SATISFAÇÃO				
Muito Insatisfeito	6.0	8.0	3.0	17.0
Insatisfeito	6.0	14.0	11.0	31.0
Indiferente	6.0	25.0	3.0	33.0
Satisfeito	11.0	8.0	0.0	19.0
Muito Satisfeito	0.0	0.0	0.0	0.0
All	28.0	56.0	17.0	100.0

## Comando Python:

```
pd.crosstab(variável linha, variável coluna)
```



# Notebook Atividade – em aula

17

Manipulando base de dados reais:

- Download pelo Blackboard
- Fazer juntos e discutir em sala

# Notebook Exercício

Manipulando base de dados reais:

- Download pelo Blackboard
- Fazer individual e discutir na mesa