Relatório de Análise V

Tratamento de dados faltantes

In [49]:
```
import pandas as pd
```

In [50]:
```
dados = pd.read_csv('../dados/aluguel_residencial.csv', sep=';')
dados.head(10)
```

Out[50]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN
2	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
3	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	NaN
4	Apartamento	Vista Alegre	3	1	0	70	1200.0	NaN	NaN
5	Apartamento	Cachambi	2	0	0	50	1300.0	301.0	17.0
6	Casa de Condomínio	Barra da Tijuca	5	4	5	750	22000.0	NaN	NaN
7	Casa de Condomínio	Ramos	2	2	0	65	1000.0	NaN	NaN
8	Apartamento	Centro	1	0	0	36	1200.0	NaN	NaN
9	Apartamento	Grajaú	2	1	0	70	1500.0	642.0	74.0

Método que verifica onde há valor nulo

In [51]:
```
dados.isnull()
```

Out[51]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	True	True
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	True
4	False	False	False	False	False	False	False	True	True
...	...	...	...	...	...	...	...	...	...
22575	False	False	False	False	False	False	False	False	False
22576	False	False	False	False	False	False	False	False	False
22577	False	False	False	False	False	False	False	False	False
22578	False	False	False	False	False	False	False	False	False
22579	False	False	False	False	False	False	False	False	True

22580 rows × 9 columns

Método que verifica onde não há valor nulo

In [52]:
```
dados.notnull()
```

Out[52]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	True	True	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True	False	False
2	True	True	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True	True	False
4	True	True	True	True	True	True	True	False	False
...	...	...	...	...	...	...	...	...	...
22575	True	True	True	True	True	True	True	True	True
22576	True	True	True	True	True	True	True	True	True
22577	True	True	True	True	True	True	True	True	True
22578	True	True	True	True	True	True	True	True	True
22579	True	True	True	True	True	True	True	True	False

22580 rows × 9 columns

Verificar Informações do Dataframe

In [53]:
```
# mostra a quantidade de registros que o dataframe possui
# mostra a quantidade de resgistros não nulos de cada variável
# podemos então identificar quais variáveis possuem registros nulos
dados.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22580 entries, 0 to 22579
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Tipo        22580 non-null  object 
 1   Bairro      22580 non-null  object 
 2   Quartos     22580 non-null  int64  
 3   Vagas       22580 non-null  int64  
 4   Suites      22580 non-null  int64  
 5   Area        22580 non-null  int64  
 6   Valor       22571 non-null  float64
 7   Condominio  20765 non-null  float64
 8   IPTU        15795 non-null  float64
dtypes: float64(3), int64(4), object(2)
memory usage: 1.6+ MB
```

- no caso acima, podemos ver que além das variáveis **Condomínio e IPTU**, a variável **Valor**, também possue alguns registros nulos

Verificando apenas os valores nulos de **Valor**

In [54]:
```
# passando a seleção direto para o dataframe
dados[dados['Valor'].isnull()]
```

Out[54]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
58	Apartamento	Barra da Tijuca	2	1	1	70	NaN	970.0	68.0
1492	Apartamento	Leme	2	0	0	75	NaN	878.0	NaN
1683	Casa	Campo Grande	3	4	3	363	NaN	NaN	NaN
2012	Apartamento	Botafogo	2	0	0	95	NaN	1010.0	170.0
2034	Apartamento	Copacabana	2	0	0	72	NaN	850.0	NaN
4941	Casa	Campo Grande	3	2	1	100	NaN	NaN	NaN
8568	Apartamento	Leme	2	0	1	75	NaN	878.0	NaN
8947	Apartamento	Glória	3	0	1	135	NaN	910.0	228.0
9149	Apartamento	Gávea	3	1	1	105	NaN	880.0	221.0

- a variável valor é a mais importante desse banco de dados
- portanto, não interessa manter esses registros nulos e serão removidos

Removendo registros nulos

- para melhor visualização, serão criadas variáveis para verificar o tamanho do df antes e depois da remoção

In [55]:
```
A = dados.shape[0]
# recebe o tamanho do df antes da remoção
dados.dropna(subset = ['Valor'], inplace=True)
# remove valores nulos de uma variável ou lista de variáveis
# é necessário o inplace para execução no df
B = dados.shape[0]
# recebe o tamanho do df depois da remoção
A - B
# mostra a diferença para saber o tanto de registros removidos
```

Out[55]: 9

In [56]:
```
# não retorna mais nenhum valor nulo
dados[dados.Valor.isnull()]
```

Out[56]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU

Tratamento Condicional

Tratamento Missings (continuação)

Verificando quantos registros nulos temos em uma variável

In [57]:
```
dados[dados['Condominio'].isnull()].shape[0]
```

Out[57]: 1813

Filtrando mais de uma variável

In [58]:
```
# cada filtro deve estar entre parênteses
selecao = (dados['Tipo'] == 'Apartamento') & (dados['Condominio'].isnull())
```

Descartando a seleção que não interessa

In [59]:
```
A = dados.shape[0]
dados = dados[~selecao]
# esse comando com o '~' inverte a Series booleana
# então pegamos um valor que não interessa e descartamos
B = dados.shape[0]
A - B
```

Out[59]: 745

In [60]:
```
# foram eliminados 9 registros que estavam com valor nulo
# mais 745 apartamentos com condomínio nulo
dados.shape[0]
```

Out[60]: 21826

In [61]:
```
# ainda permanecem registros nulos, mas não para apartamentos
dados[dados['Condominio'].isnull()].shape[0]
```

Out[61]: 1068

Atribuindo valores para registros nulos

In [62]:
```
# substitui todos os registros nulos do dataframe por 0
# é necessário o inplace para gravar no dataframe
# fillna() é o método que preenche valores nulos
dados.fillna(0)
```

Out[62]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	0.0	0.0
2	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
3	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	0.0
5	Apartamento	Cachambi	2	0	0	50	1300.0	301.0	17.0
...	...	...	...	...	...	...	...	...	...
22575	Apartamento	Méier	2	0	0	70	900.0	490.0	48.0
22576	Quitinete	Centro	0	0	0	27	800.0	350.0	25.0
22577	Apartamento	Jacarepaguá	3	1	2	78	1800.0	800.0	40.0
22578	Apartamento	São Francisco Xavier	2	1	0	48	1400.0	509.0	37.0

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
22579	Apartamento	Leblon	2	0	0	70	3000.0	760.0	0.0

21826 rows × 9 columns

Escolhendo variáveis e valores para registros nulos

Usando um Dicionário

- com o dicionário é possível escolher quais variáveis serão preenchidas e quais valores serão atribuídos

In [63]:
```
dados = dados.fillna({'Condominio': 0, 'IPTU': 0})
```

Verificando se há valores nulos

In [64]:
```
dados[dados['Condominio'].isnull()].shape[0]
```

Out[64]: 0

In [65]:
```
dados[dados['Condominio'].isnull()].shape[0]
```

Out[65]: 0

In [66]:
```
dados.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 21826 entries, 0 to 22579
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Tipo        21826 non-null  object 
 1   Bairro      21826 non-null  object 
 2   Quartos     21826 non-null  int64  
 3   Vagas       21826 non-null  int64  
 4   Suites      21826 non-null  int64  
 5   Area        21826 non-null  int64  
 6   Valor       21826 non-null  float64
 7   Condominio  21826 non-null  float64
 8   IPTU        21826 non-null  float64
dtypes: float64(3), int64(4), object(2)
memory usage: 1.7+ MB
```

Exportando e Sobrescrevendo o Dataframe

In [68]:
```
# para fins didáticos e manutenção dos notebooks, não sobrescreverei o arqui
dados.to_csv('../dados/aluguel_residencial_notnull.csv', sep=';', index = Fal
```