



## ▼ Lab Análise Exploratória de Dados (EDA)

---

### Caso: Onde comprar imóveis em Melbourne (Austrália)?

Empregue os dados da URL:

[http://meusite.mackenzie.br/rogerio/TIC/Melbourne\\_housing\\_FULL.csv](http://meusite.mackenzie.br/rogerio/TIC/Melbourne_housing_FULL.csv)

A análise de dados tem uma grande aplicação no mercado de imóveis, seja para projetar oportunidades de negócios (como que tipo de imóvel e onde construir), seja para responder a questões mais simples (onde encontro melhores oportunidade de imóveis de 2 dormitórios). Ao final desse Lab você será capaz de responder algumas dessas perguntas explorando uma base com cerca de 35K registros sobre negócios de imóveis em Melbourne.

### Exercício 0. Semântica dos Dados

Essa é uma parte importante da exploração dos dados mas as informações precisam por algum tipo de documentação dos dados. Analise algumas informações dos dados aqui...

[Melbourne Housing Market](#)

## ▼ Exercício 1. Imports e Aquisição dos Dados

Faça aqui os imports para a construção da sua EDA. Em seguida faça a dos dados. Lembre-se de verificar a origem dos dados para empregar as funções e parâmetros corretos de leitura dos dados (extensão do arquivo, headers, separador etc.).

### ▼ imports

```
# seu código
import pandas          as pd
import numpy           as np
import matplotlib.pyplot as plt
import seaborn         as sns
import statsmodels.formula.api as sm
import warnings
```

```
warnings.filterwarnings("ignore") # Suppress all warnings?
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning
import pandas.util.testing as tm
```

## ▼ Aquisição dos dados

```
# seu código
houses = pd.read_csv('http://meusite.mackenzie.br/rogerio/TIC/Melbourne_housing_FULL.csv')
houses.head()
```

## ▼ Exercício 2. Explore a Estrutura dos Dados

Quantos registros e atributos tem? Qual o aspecto dos dados? Quais os tipos dos dados (categóricos, numéricos)?

```
# seu código
houses.shape
len(houses)
houses.columns
houses.head()
houses.shape
houses.info
houses.dtypes
```

```
(34857, 21)
```

## ▼ Exercício 3 Examine Estatísticas dos Dados Brutos

Verifique por exemplo:

1. Qual a média de preços
2. Qual o ano da construção mais antiga e a mais recente? O que você conclui?
3. Quantas regiões há e qual a região com mais casas à venda?

```
# seu código
pd.options.display.max_columns = 20
houses.describe(include='all')
```

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date
<b>count</b>	34857	34857	34857.000000	34857	2.724700e+04	34857	34857	34857
<b>unique</b>	351	34009	NaN	3	NaN	9	388	388
<b>top</b>	Reservoir	Charles St	NaN	h	NaN	S	Jellis	28/10/2021
<b>freq</b>	844	6	NaN	23980	NaN	19744	3359	1
<b>mean</b>	NaN	NaN	3.031012	NaN	1.050173e+06	NaN	NaN	NaN
<b>std</b>	NaN	NaN	0.969933	NaN	6.414671e+05	NaN	NaN	NaN
<b>min</b>	NaN	NaN	1.000000	NaN	8.500000e+04	NaN	NaN	NaN
<b>25%</b>	NaN	NaN	2.000000	NaN	6.350000e+05	NaN	NaN	NaN
<b>50%</b>	NaN	NaN	3.000000	NaN	8.700000e+05	NaN	NaN	NaN
<b>75%</b>	NaN	NaN	4.000000	NaN	1.000000e+06	NaN	NaN	NaN

Qual o ano da construção mais antiga e a mais recente? O que você conclui?

Aqui certamente há um erro nos dados. A EDA é uma etapa importante na verificação da qualidade dos dados.

## ▼ Exercício 4. Verifique Dados Faltantes

Existem dados faltantes? O que você pode afirmar sobre o número de linhas com valores nulos?

**DICA:** `_ .isnull(). _`

```
houses.isnull().sum()
houses.isnull().sum().sum()
```

```
Suburb      0
Address     0
Rooms       0
Type        0
Price      7610
Method      0
SellerG     0
Date        0
Distance    1
Postcode    1
Bedroom2    8217
Bathroom    8226
Car         8728
Landsize   11810
BuildingArea 21115
YearBuilt   19306
CouncilArea 3
Latitude    7976
Longitude   7976
Regionname  3
```

```
Propertycount      3
dtype: int64
```

## Exercício 4b. Obtenha a Média de Valores de Price (RESOLVIDO)

Em paralelo obtenha o percentual de valores nulos. O que você conclui?

```
houses.Price.mean()
houses.Price.isnull().sum() / houses.Price.count()

0.27929680331779644
```

É um percentual de valores nulos muito alto e que certamente compromete o valor da média. Por isso, mesmo em uma questão simples, é sempre bom verificar a qualidade dos dados.

## Exercício 5. Tratando Dados Nulos

Antes, discuta as estratégias de tratamento de nulos para `Price`. Em seguida, por simplicidade, aplique a estratégia de imputar os valores médios para `Price` e `Landsize`.

**DICA:** `_ replace(np.NaN, _)` ou `.fillna()`

```
houses.Price = houses.Price.replace(np.NaN, houses.Price.mean())
houses.Landsize = houses.Landsize.fillna(houses.Landsize.mean())

1050173.344955408
```

## Exercício 6

Por quantas Regiões estão distribuídas as casas de Melbourne? Qual o percentual da Região com mais casas vendidas?

**DICA:** use `_.unique()`, `_.value_counts()`, `.groupby()`

```
houses.Regionname.unique()
houses.Regionname.value_counts() / len(houses) * 100
```

Southern Metropolitan	33.955877
Northern Metropolitan	27.417735
Western Metropolitan	19.505408
Eastern Metropolitan	12.557019
South-Eastern Metropolitan	4.988955
Eastern Victoria	0.654101

```

Northern Victoria      0.582379
Western Victoria      0.329919
Name: Regionname, dtype: float64

```

```
houses.groupby('Regionname').Regionname.count().sort_values(ascending=False) / houses.Regi
```

```

Regionname
Southern Metropolitan      33.958800
Northern Metropolitan      27.420095
Western Metropolitan       19.507087
Eastern Metropolitan       12.558100
South-Eastern Metropolitan  4.989384
Eastern Victoria           0.654157
Northern Victoria          0.582430
Western Victoria           0.329948
Name: Regionname, dtype: float64

```

## ▼ Exercício 7

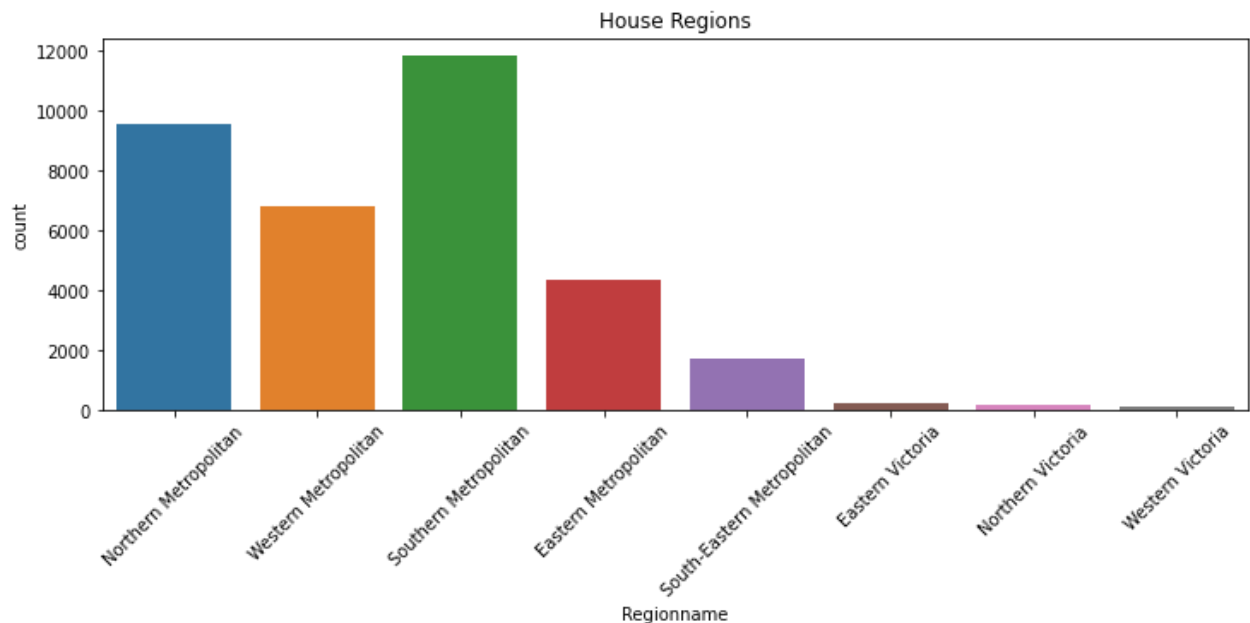
Examine a questão anterior graficamente.

**DICA:** `sns.countplot()` ou `sns.barplot`

```

plt.figure(figsize=(12,4))
sns.countplot(houses.Regionname)
plt.title('House Regions')
plt.xticks(rotation=45)
plt.show()

```



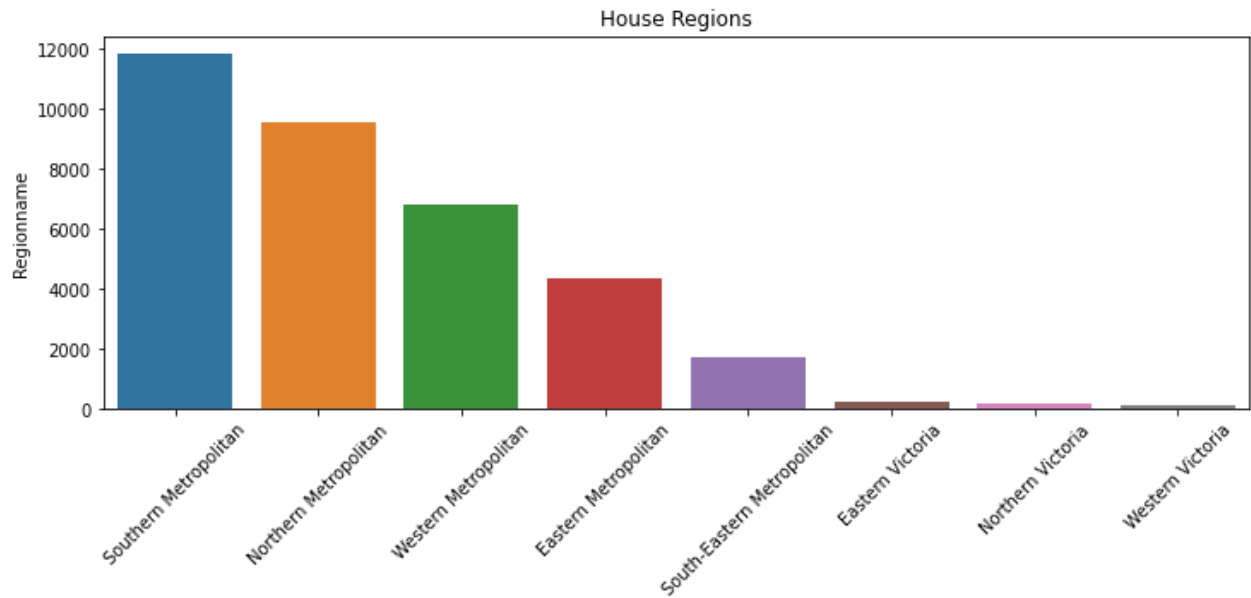
# Se você quiser ir mais fundo...

```

plt.figure(figsize=(12,4))
sns.barplot(houses.Regionname.value_counts().index, houses.Regionname.value_counts())

```

```
plt.title('House Regions')  
plt.xticks(rotation=45)  
plt.show()
```




```
# Veja que um gráfico de barras é melhor  
plt.figure(figsize=(8,8))  
plt.pie(houses.Regionname.value_counts(),  
        labels=houses.Regionname.value_counts().index,  
        autopct='%1.1f%%')  
plt.title('House Regions')  
plt.show()
```

## House Regions

## ▼ Exercício 8 (RESOLVIDO)


Quais os 5 preços maiores e menores preços dos imóveis?



```
houses.Price.nlargest(5)
houses.Price.nsmallest(5)
```

4378	85000.0
29669	112000.0
17529	121000.0
3063	131000.0
3290	145000.0

Name: Price, dtype: float64



## ▼ Exercício 9

Qual o maior e menor preço dos imóveis? De que região é cada um desses imóveis?

```
houses.nlargest(1, 'Price').Regionname
houses.nsmallest(1, 'Price').Regionname
```

4378	Western Metropolitan
------	----------------------

Name: Regionname, dtype: object

```
houses.Price.min()
houses.Price.max()
```

```
houses[ houses.Price == houses.Price.min() ].Regionname
houses[ houses.Price == houses.Price.max() ].Regionname
```

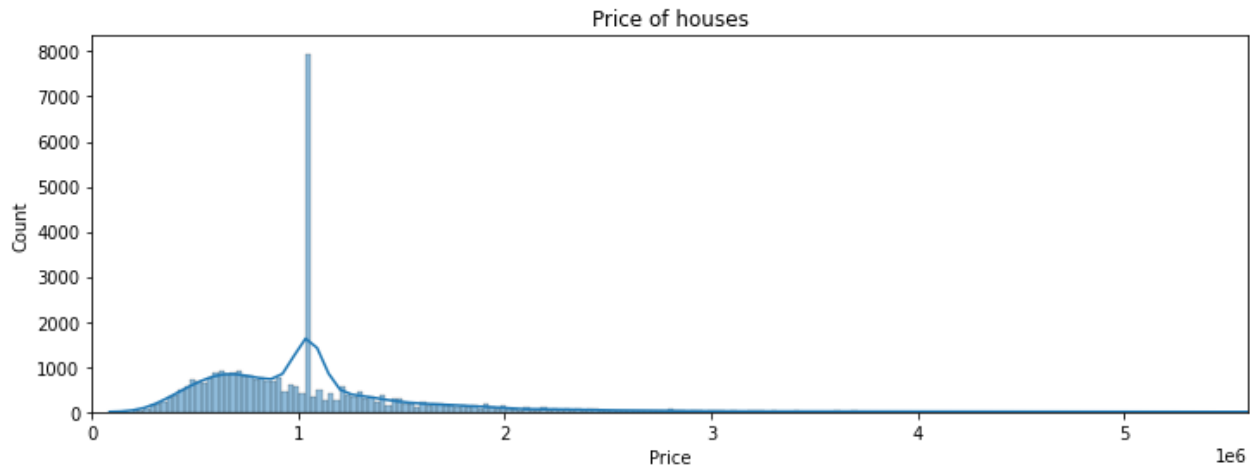
25635	Southern Metropolitan
-------	-----------------------

Name: Regionname, dtype: object

## ▼ Exercício 10

Como estão distribuídos os preços dos imóveis? (Faça um gráfico)

```
plt.figure(figsize=(12,4))
sns.histplot(houses.Price,kde=True)
plt.title('Price of houses')
plt.xlim([0, 0.5*houses.Price.max()])
plt.show()
```



## ▼ Exercício 11 Discussão

Você nota algo estranho nessa distribuição? Como você corrigiria isso?

Veja que existe um valor que se sobressai. É o valor médio que foi imputado para os valores nulos. Esse é o risco de se imputar os valores.

Uma estratégia seria a de imputar valores aleatórios com a mesma distribuição e média dos valores existentes. Algo não complexo, mas que requer alguma programação adicional. A solução está abaixo, mas não tem necessidade de você compreender o código, mas entenda ao menos o raciocínio empregado.

```
houses2 = pd.read_csv('http://meusite.mackenzie.br/rogerio/TIC/Melbourne_housing_FULL.csv')
houses2.Price.isnull().sum()
```

```
7610
```

```
for i in range(len(houses2)):
    if np.isnan(houses2.loc[i].Price):
        new_price = houses2.loc[np.random.randint(0, len(houses2))].Price
        while np.isnan(new_price):
            new_price = houses2.loc[np.random.randint(0, len(houses2))].Price
        houses2.at[i, 'Price'] = new_price
```

```
houses2.Price.isnull().sum()
```

```
0
```

```
plt.figure(figsize=(12,4))
sns.histplot(houses2.Price, kde=True)
plt.title('Price of houses')
plt.xlim([0, 0.5*houses2.Price.max()])
plt.show()
```





Mesmo assim, note que esse tipo de atribuição fará com que imóveis com mais metros quadrados e melhor região ainda possam receber um valor aleatório de um imóvel menor e de uma região menos valorizada! Neste caso uma interpolação de valores poderia fazer mais sentido. Mas não vamos seguir aqui com essa solução. Aqui o que importa é a discussão e a reflexão sobre os dados (*Data Thinking*).

## ▼ Exercício 12. (RESOLVIDO)

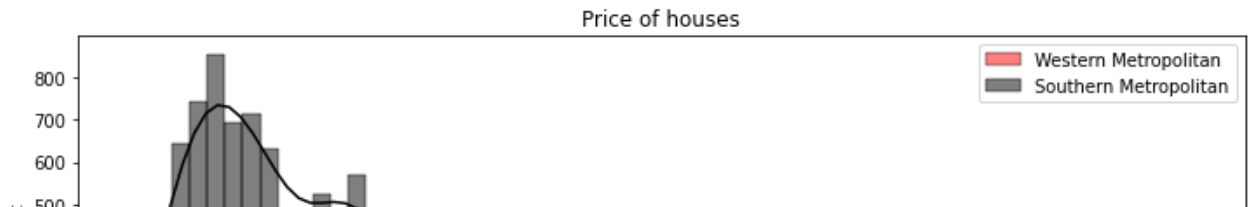
Compare as distribuições de preços das regiões de 'Western Metropolitan' e 'Southern Metropolitan'. O que você pode afirmar?

# Código resolvido, você não precisa codificar nada aqui. Apenas analise e responda a perg

```
plt.figure(figsize=(12,4))
sns.histplot(houses2[houses2.Regionname == 'Western Metropolitan'].Price,kde=True,color='r',
             label = 'Western Metropolitan')
sns.histplot(houses2[houses2.Regionname == 'Southern Metropolitan'].Price,kde=True,color='b',
             label = 'Southern Metropolitan')
plt.title('Price of houses')

plt.xlim([0, 0.5*houses2.Price.max()])
plt.legend()

plt.show()
```



Há mais venda de imóveis 'Southern Metropolitan' mas a distribuição de preços é bastante semelhante a da 'Western Metropolitan'.



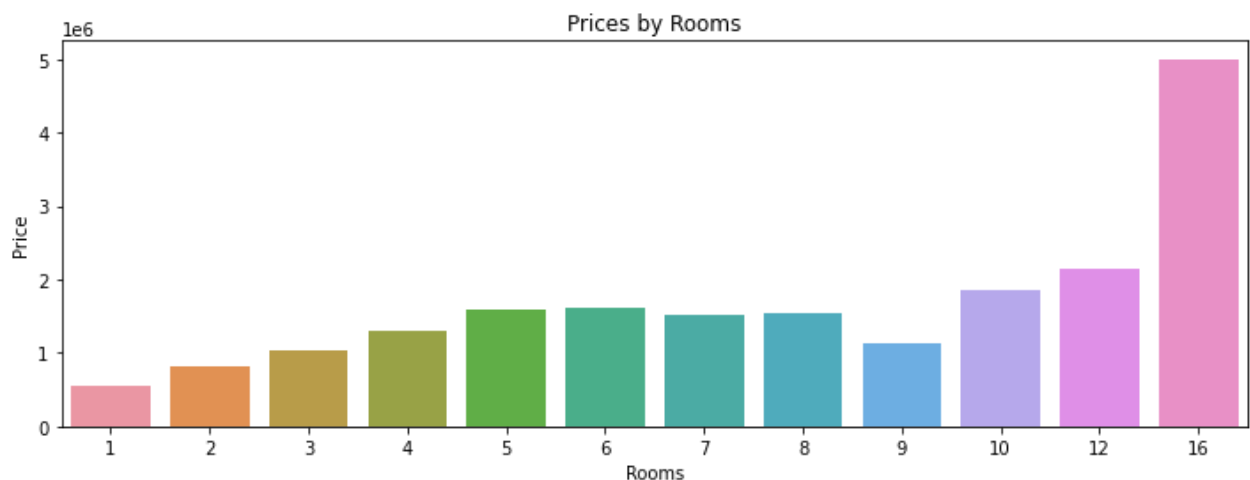
## ▼ Exercício 13. (RESOLVIDO)

Imóveis com mais dormitórios são mais caros? Exiba em um gráfico

**DICA:** `df.groupby()` e `sns.barplot()`

```
housesRooms = houses.groupby('Rooms').Price.mean()
```

```
plt.figure(figsize=(12,4))
sns.barplot(housesRooms.index, housesRooms)
plt.title('Prices by Rooms')
plt.show()
```



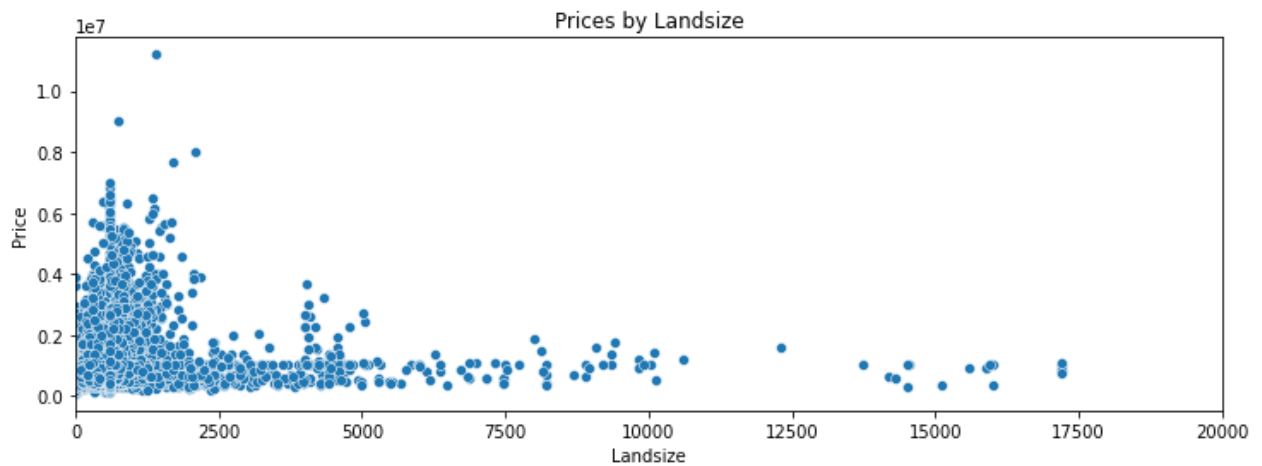
## ▼ Exercício 14.

Imóveis com maior área, são mais caros? Exiba em um gráfico

**DICA:** Aqui você não pode empregar o `barplot()` (por que?). Empregue um gráfico de *dispersão* do `sns` (qual é?)

```
plt.figure(figsize=(12,4))
sns.scatterplot(houses.Landsize, houses.Price)
plt.title('Prices by Landsize')
plt.xlim([0,20000])
```

```
plt.show()
```



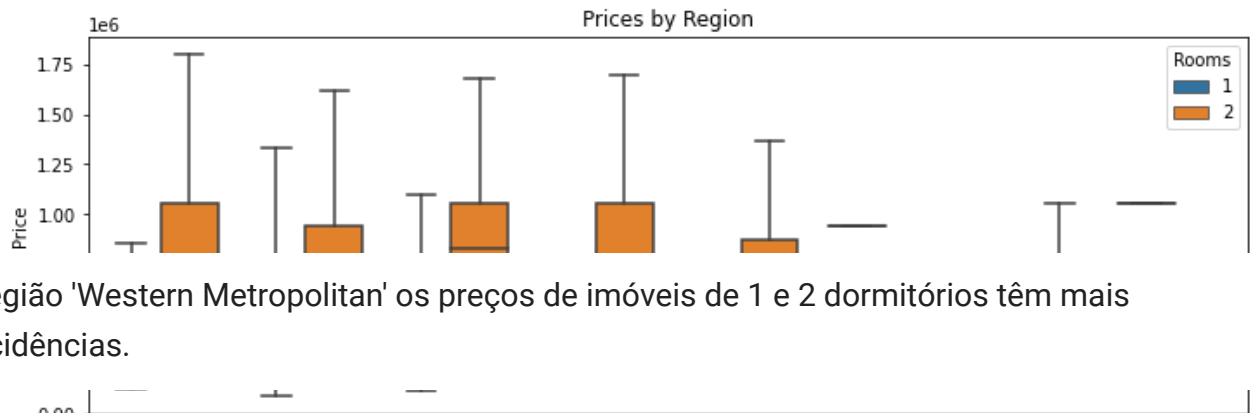
## ▼ Exercício 15. (RESOLVIDO)

Um Cliente quer comprar um apartamento de 1 dormitório. Ele deseja comprar em uma região em que o valor do imóvel seja mais próximo dos valores de imóveis com 2 dormitórios, pois pretende casar no próximo ano e isso seria portanto mais cômodo para uma troca. Qual região você recomendaria o cliente comprar?

**DICA:** Empregue um `boxplot()`

1. Crie um data frame somente com os registro de imóveis com 1 e 2 dormitórios
2. Faça um então um `boxplot()` do preço por região
3. Inclua o parâmetro `hue` para `Rooms` (ver Teoria Trilha 4)

```
houses2 = houses[houses.Rooms<=2]
plt.figure(figsize=(12,4))
sns.boxplot('Regionname', 'Price', data=houses2, hue='Rooms', showfliers=False)
plt.title('Prices by Region')
plt.xticks(rotation=45)
plt.show()
```



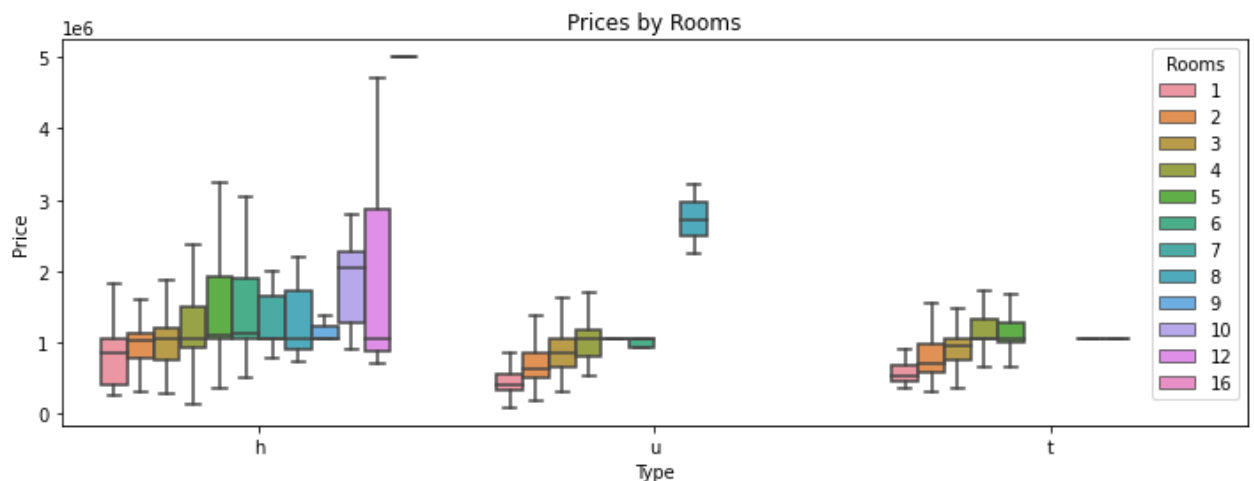
Na regi o 'Western Metropolitan' os pre os de im oveis de 1 e 2 dormit rios t m mais coincid ncias.

## Exerc cio 16.

Um cliente quer fazer um investimento e deseja comprar um im ovel com o maior n mero de c modos. Sendo im ovel para investimento ele procura algo com a maior variabilidade poss vel de pre o (pois pretende comprar pelo pre o menor e vender daqui um ano pelo pre o maior). Que tipo de im ovel voc  sugere ao cliente (Type = h(ouse), u(nit, ou apartamento), t(ower, sobrado)).

**DICA:** Empregue como modelo a solu  o do exerc cio anterior.

```
plt.figure(figsize=(12,4))
sns.boxplot('Type', 'Price', data=houses, hue='Rooms', showfliers=False)
plt.title('Prices by Rooms')
plt.show()
```



Casas com mais dormit rios apresentam variabilidade maior de pre o.

## Conclus o

Em geral um EDA apresenta algum tipo de conclusão. Sendo aqui apenas um exercício poderíamos concluir nossa análise:

Os imóveis da Região Sul são os mais caros havendo também a maior oferta de imóveis. O preços tem forte influência da região e as casas apresentam um espectro maior de preços (variação). A ausência de preço para um grande número de imóveis é um fator a ser revisado na análise.

