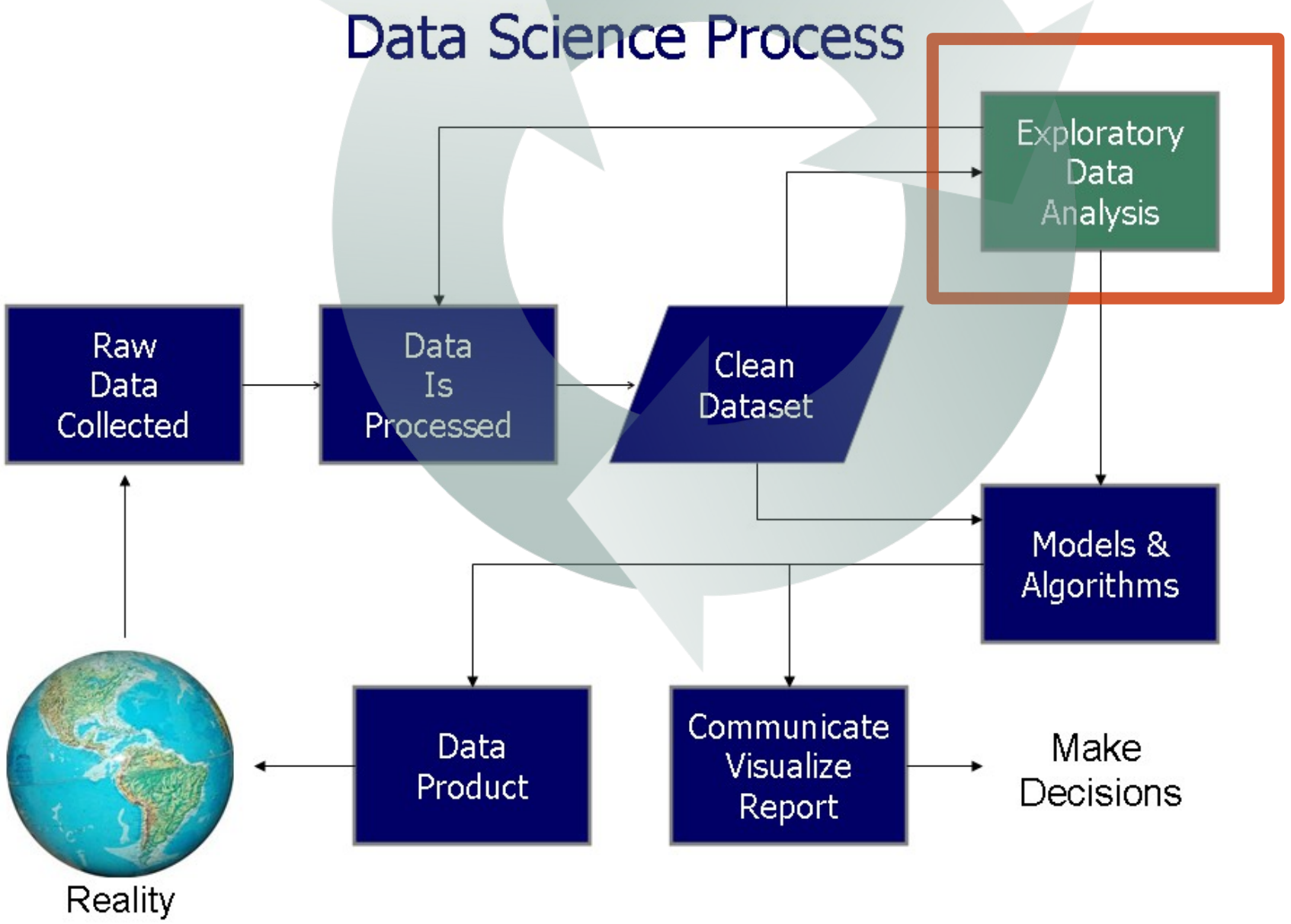
Pandas – Análise Exploratória

Luiz Celso Gomes-Jr

Análise Exploratória de Dados

- Um ramo da análise estatística proposto por John Tukey nos anos 70
- Uma abordagem para se analisar datasets e **sumarizar suas características principais**, usando **métodos visuais**
- Essencial quando o conhecimento a priori é limitado
- Modelos não visuais podem ser usados ou não, dependendo do objetivo

Expl. Analysis in context



Objetivos

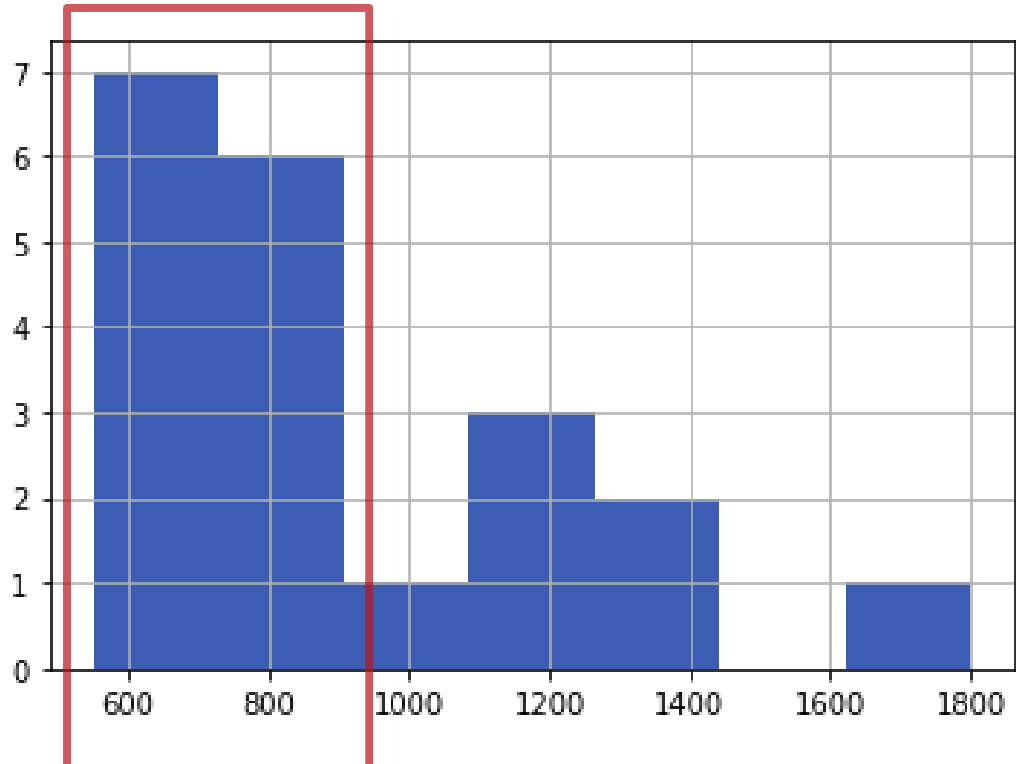
- **Sugestão de hipóteses** sobre as causas dos fenômenos observados
- **Avaliar premissas** sobre as quais as inferências estatísticas serão implementadas
- **Suporte a seleção** de ferramentas e técnicas apropriadas
- **Substanciar outras coletas** de dados para suplementar o dataset

Histograma

Histogramas mostram a quantidade de elementos com valores similares para uma variável. Por exemplo, no histograma abaixo há muitos apartamentos que custam entre R$ 600 e R$ 900.

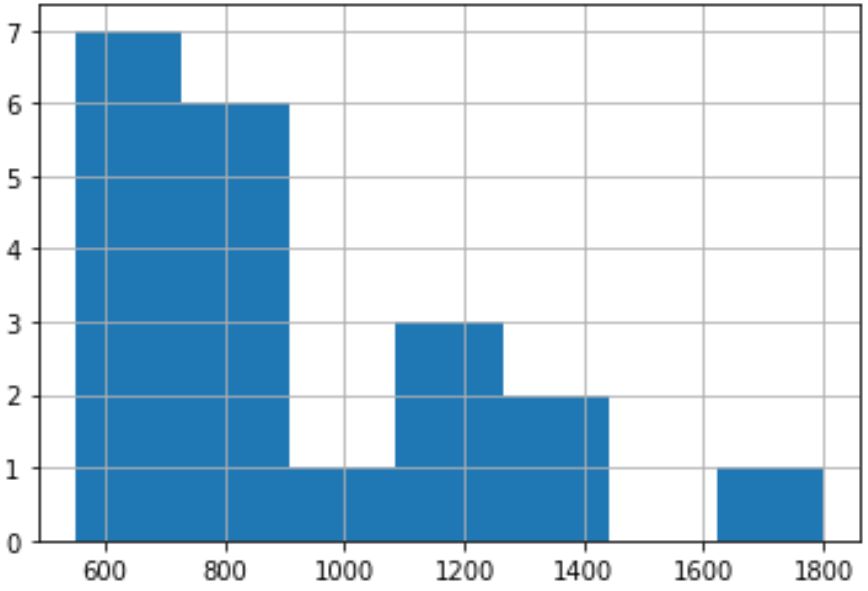
```
df['aluguel'].hist(bins=7)
```

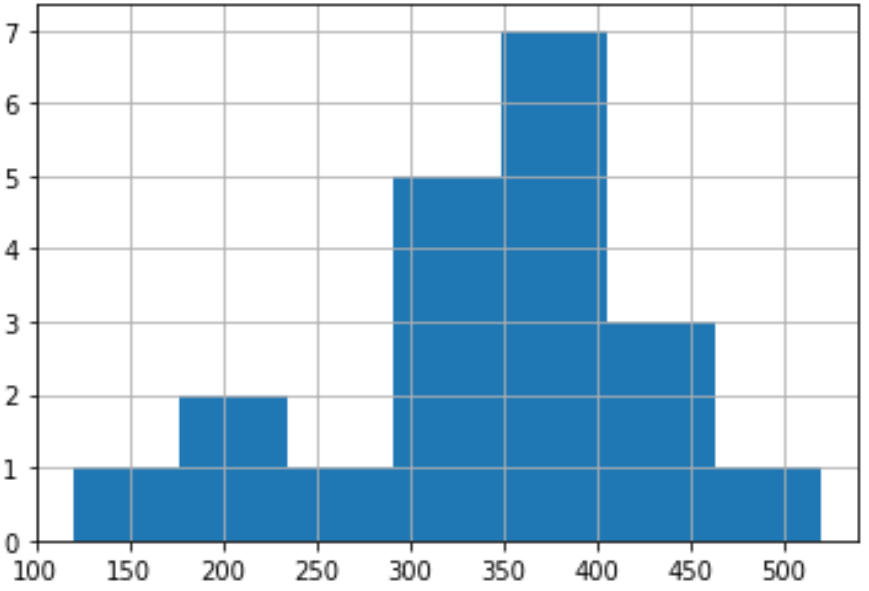
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f550fc122e8>
```



Histograma

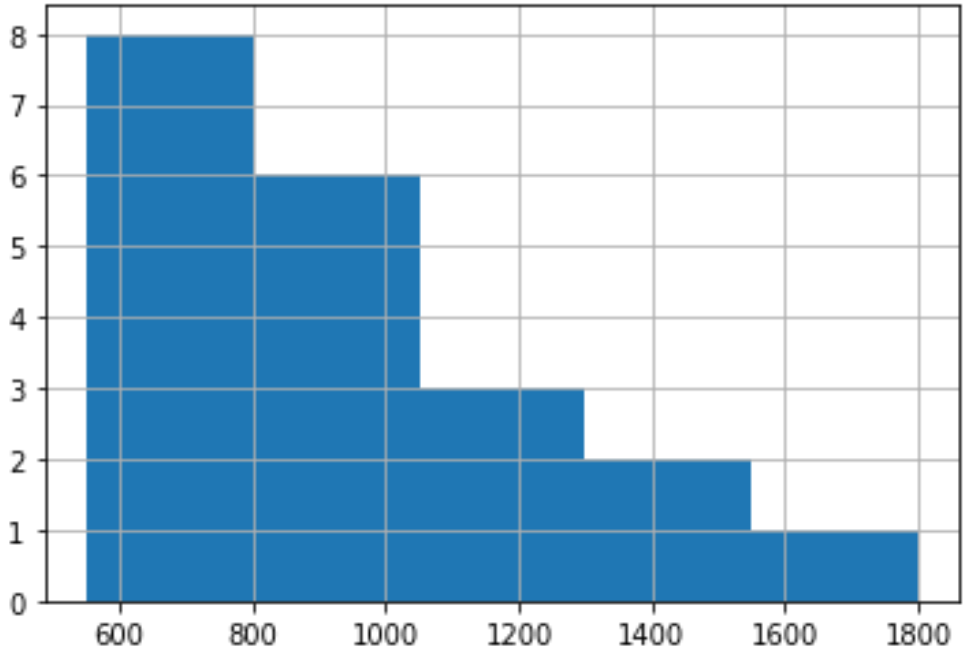
Aluguel X Condomínio



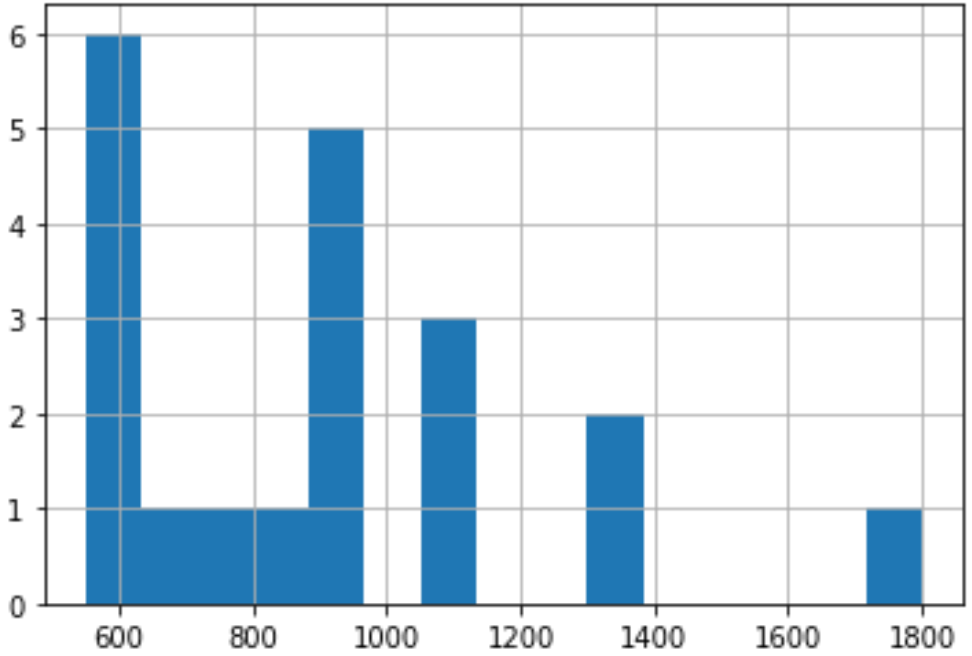


Histograma - bins

Para aumentar ou diminuir a granularidade de um histograma, basta ajustar o parâmetro `bins`. Este parâmetro controla o número de barras possíveis no intervalo de dados.



Aluguel – 5 bins

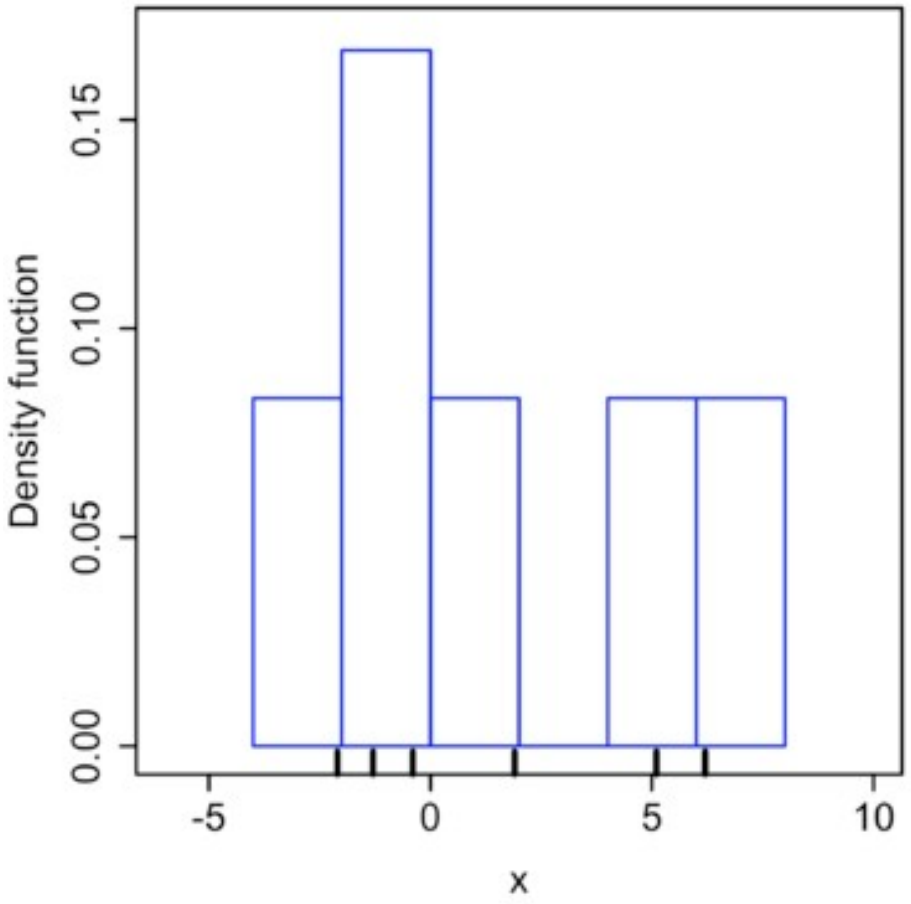


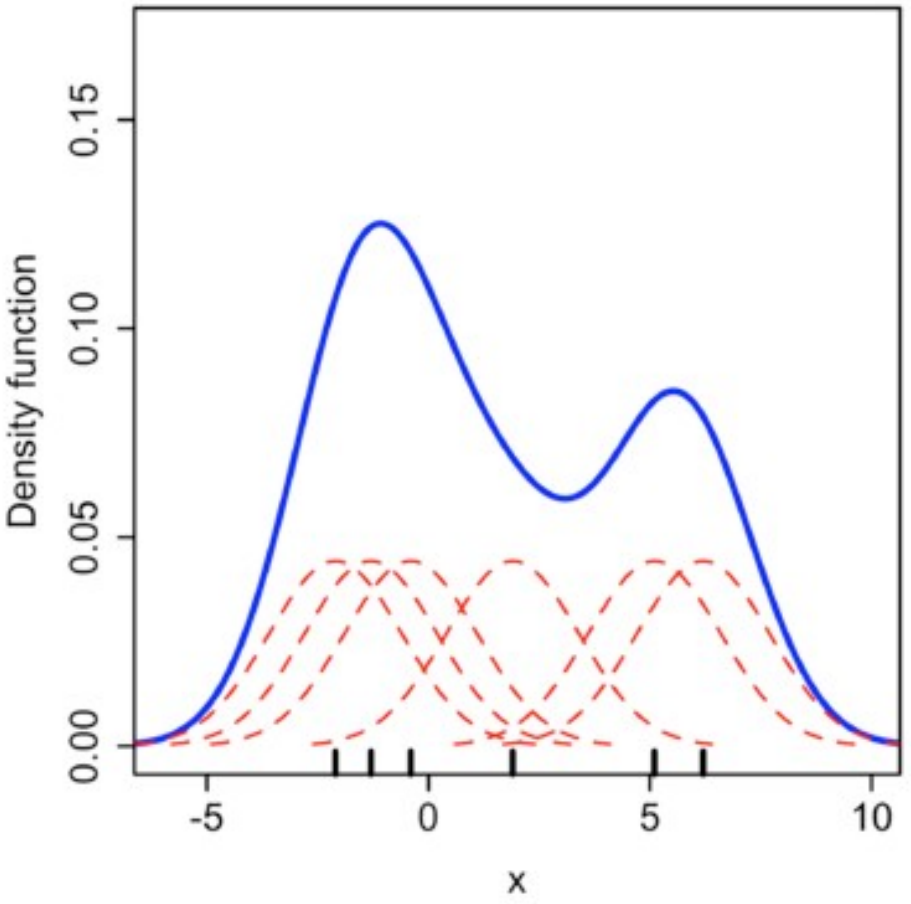
Aluguel – 15 bins

Kernel density estimation (KDE)

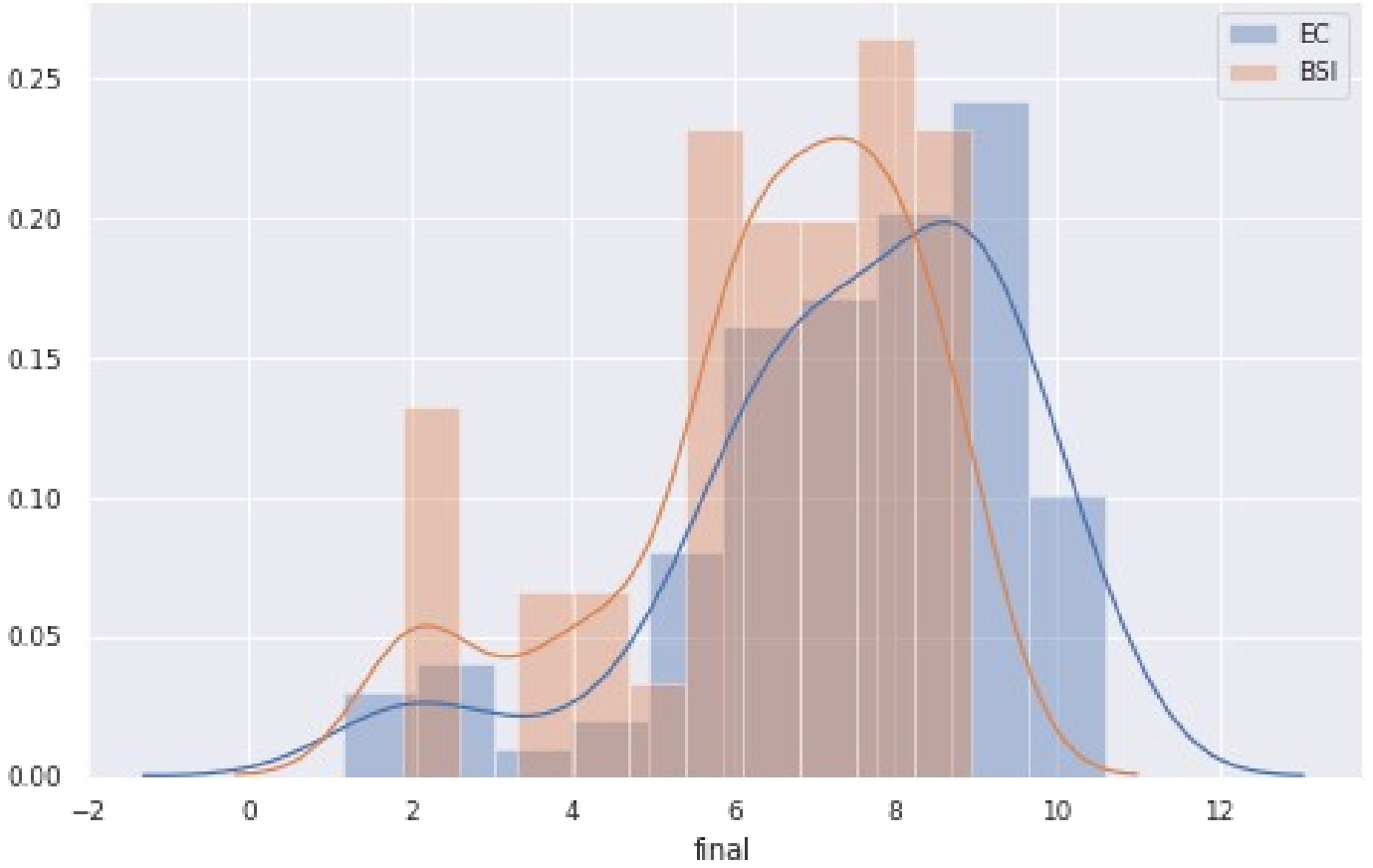
- Uma forma de estimar a função de densidade da probabilidade de uma variável
- Transforma observações discretas em uma distribuição contínua

Discreto → Contínuo





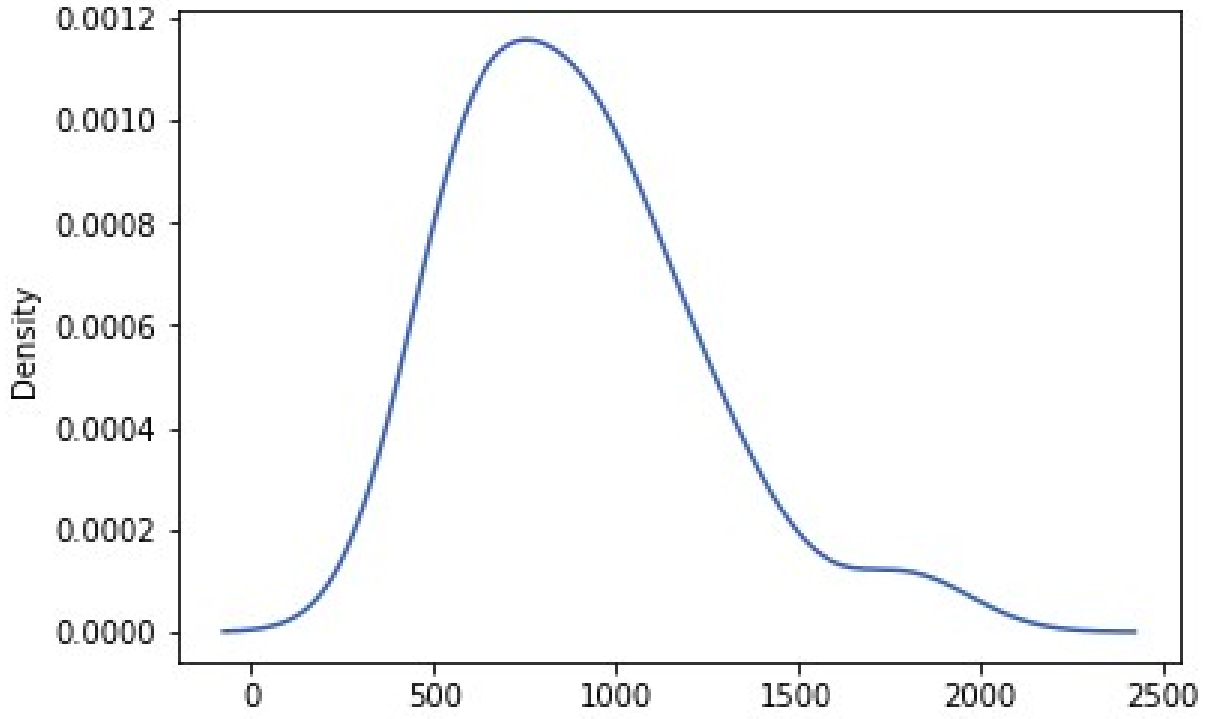
KDE para comparação de grupos



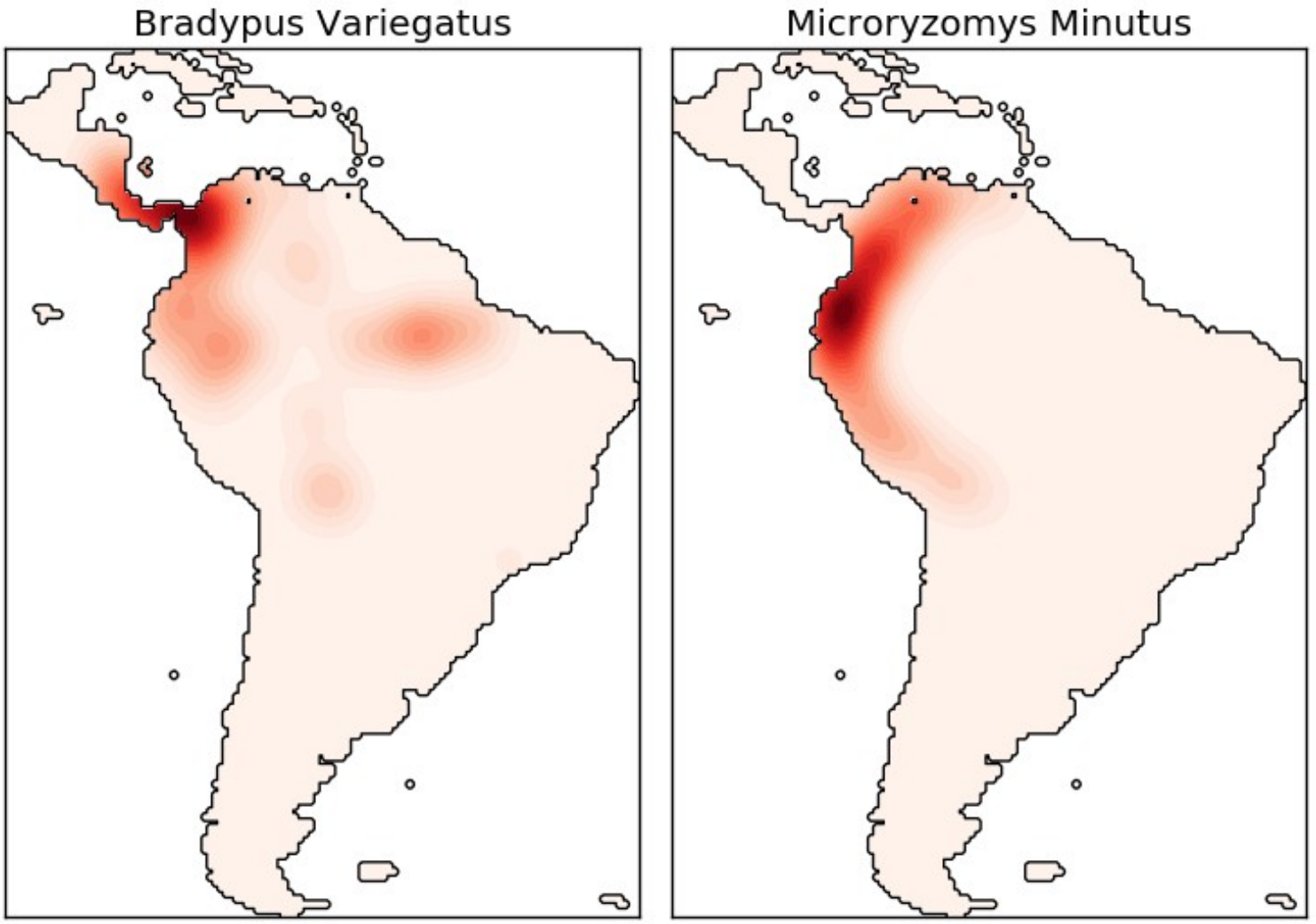
KDE

```
df['aluguel'].plot.kde()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fab5b295b38>
```



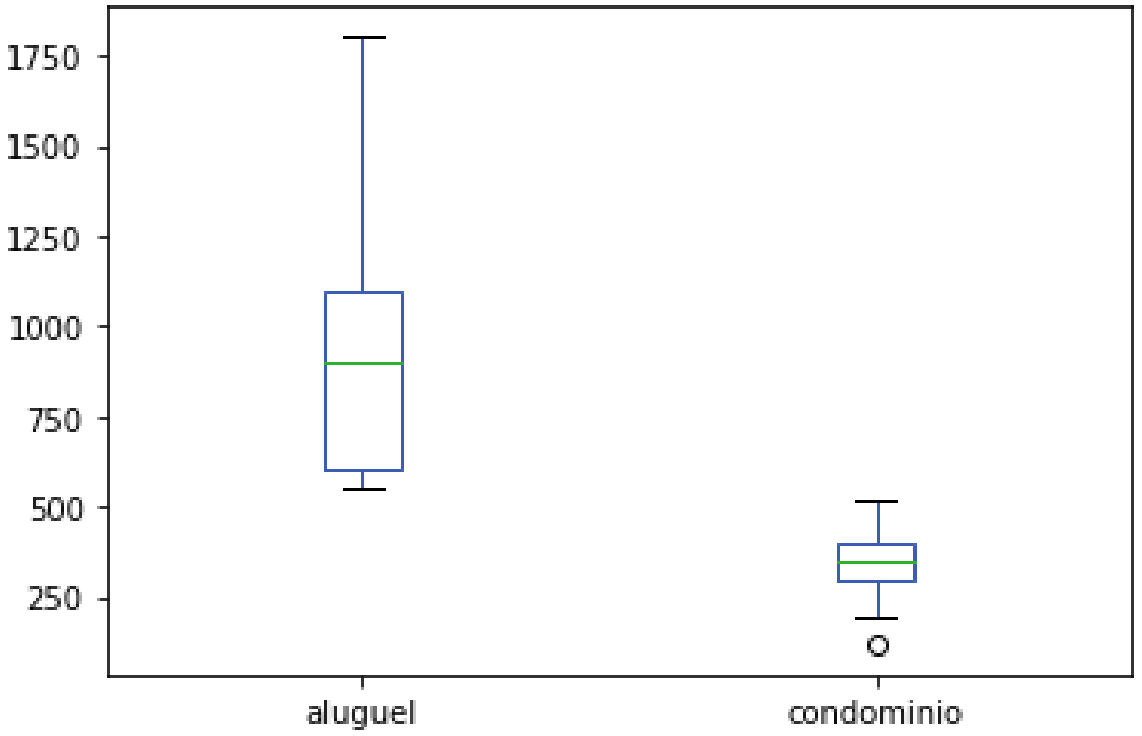
KDE em 2D



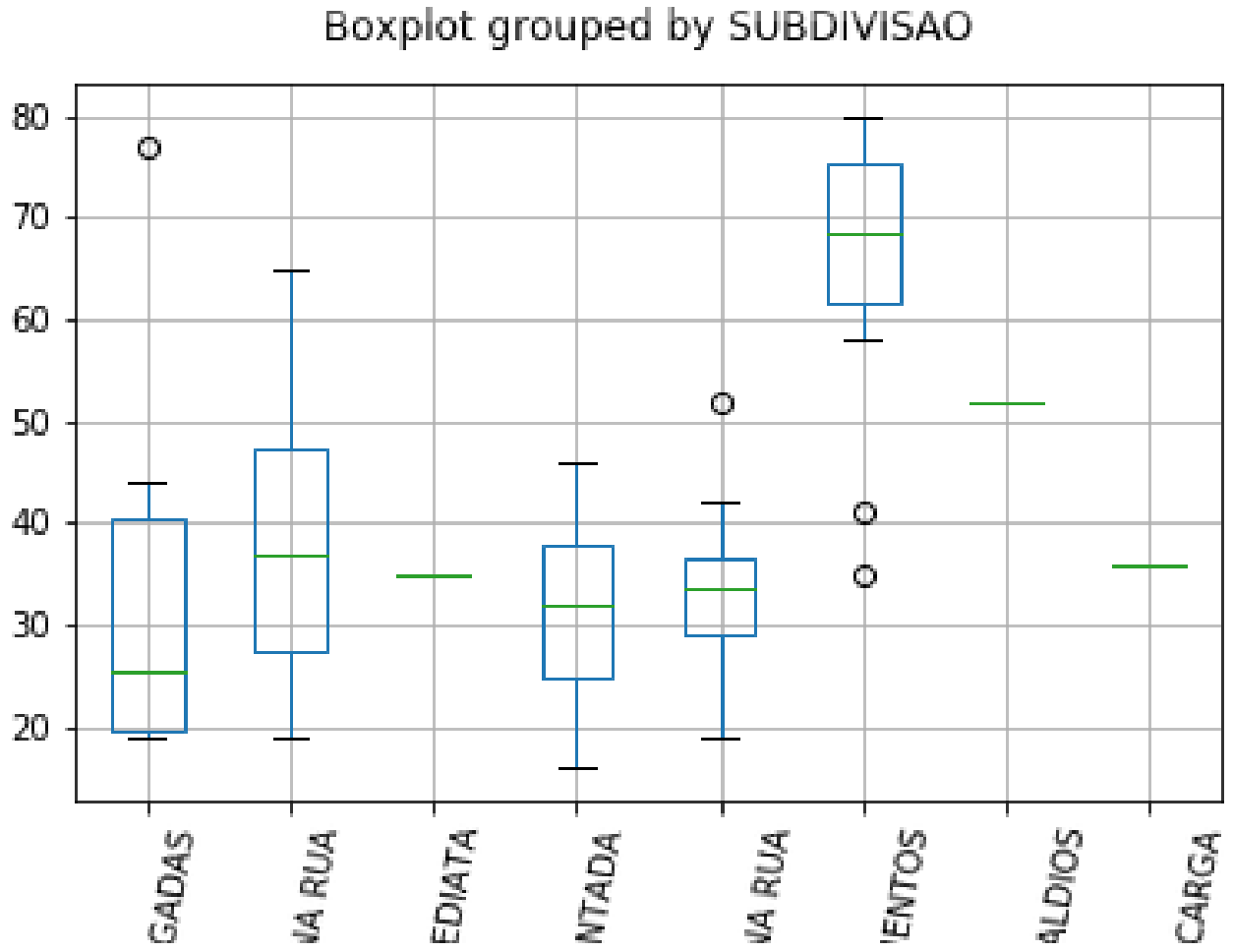
BoxPlots

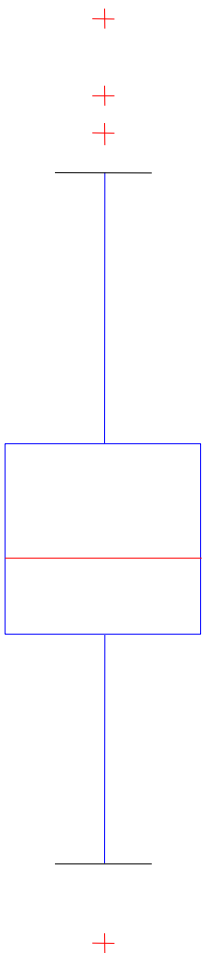
Boxplots resumem de forma concisa a distribuição das variáveis. Abaixo podemos ver as variáveis `aluguel` e `condomínio`. O gráfico mostra como os valores de aluguel variam mais comparados com os de condomínio.

```
ax = df[['aluguel', 'condominio']].plot.box()
```



Freq. Dist. - Box Plot





outlier

10th percentile

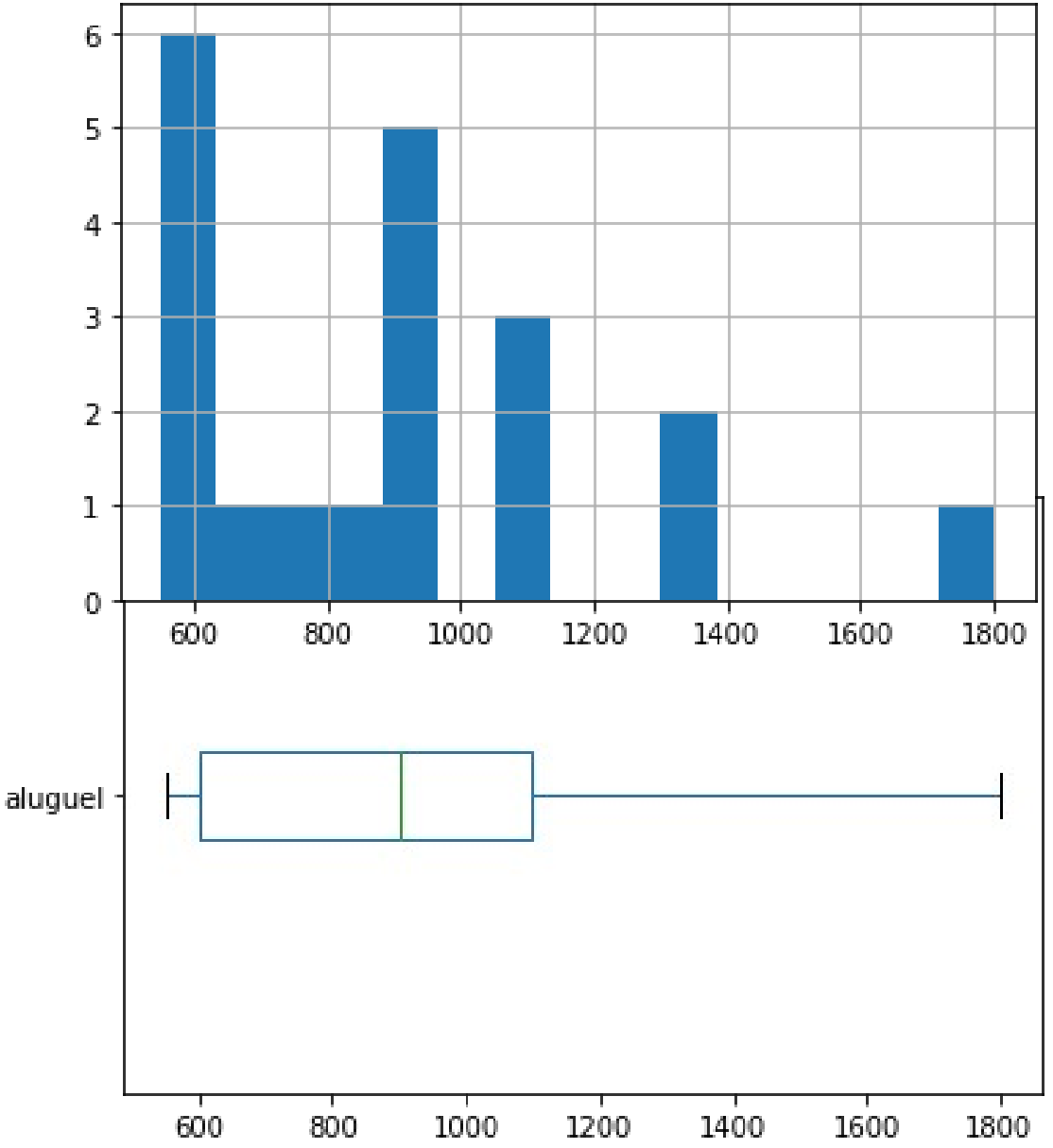
75th percentile

50th percentile

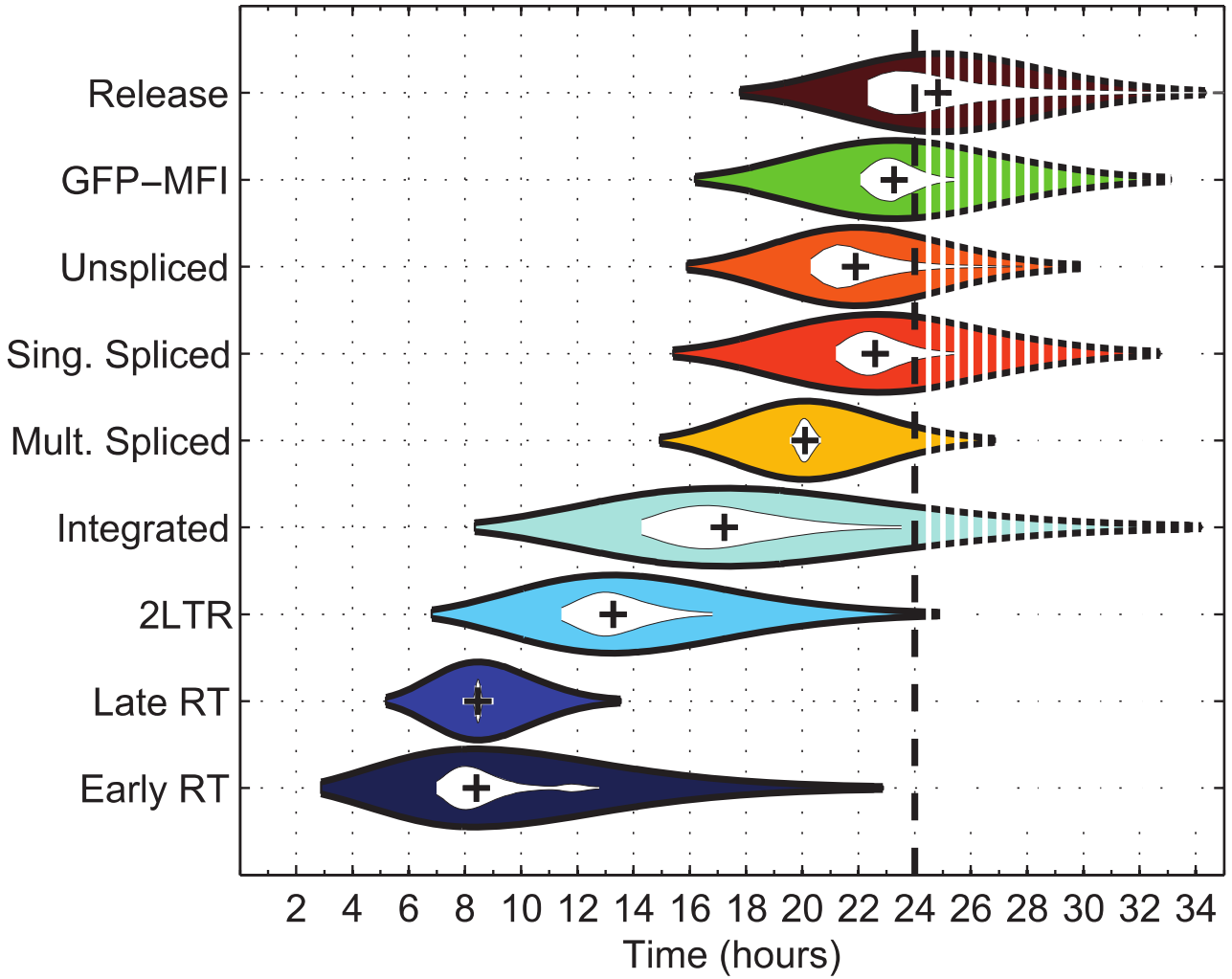
25th percentile

10th percentile

Boxplot X Histograma



Violin Plot

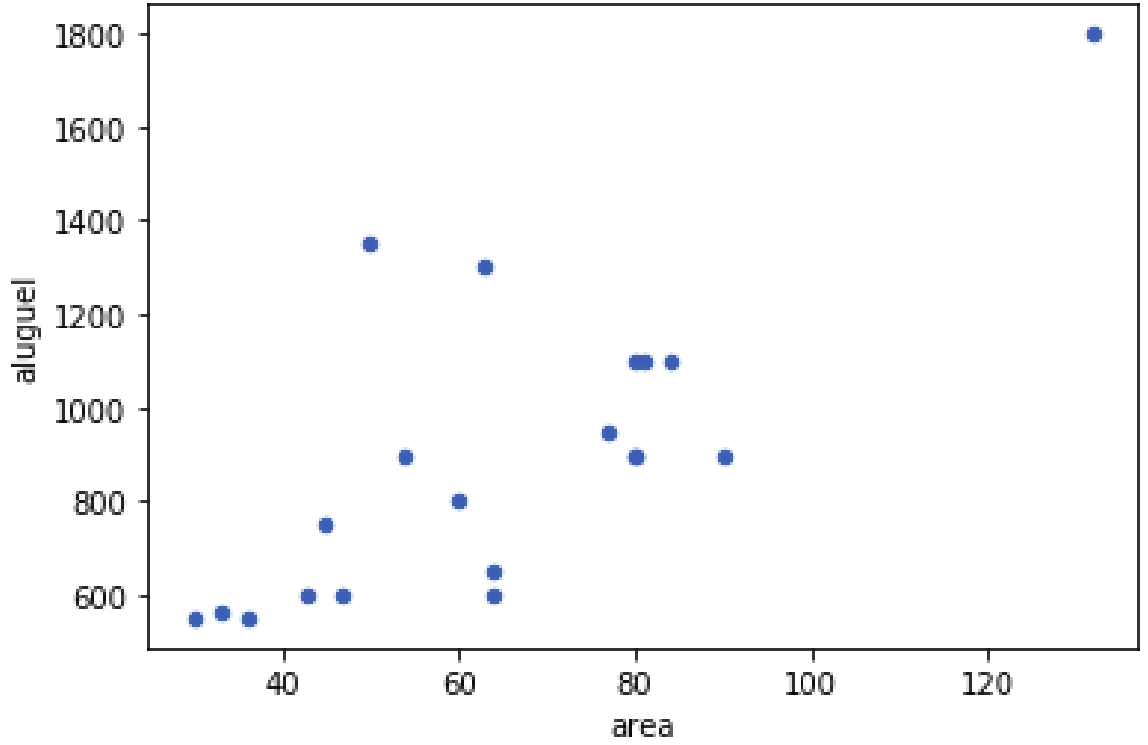


Cropped from figure 1 in Pejman Mohammadi, Sébastien Desfarges, István Bartha, Beda Joos, Nadine Zangger, Miguel Muñoz, Huldrych F. Günthard, Niko Beerenwinkel, Amalio Telenti, Angela Ciuffi (2013). "24 Hours in the Life of HIV-1 in a T Cell Line". PLOS Pathogens. DOI:10.1371/journal.ppat.1003161.

Scatter Plots

Scatter plots são úteis para se identificar padrões em duas variáveis. Por exemplo, abaixo podemos verificar alguns agrupamentos e também perceber a correlação entre as duas variáveis.

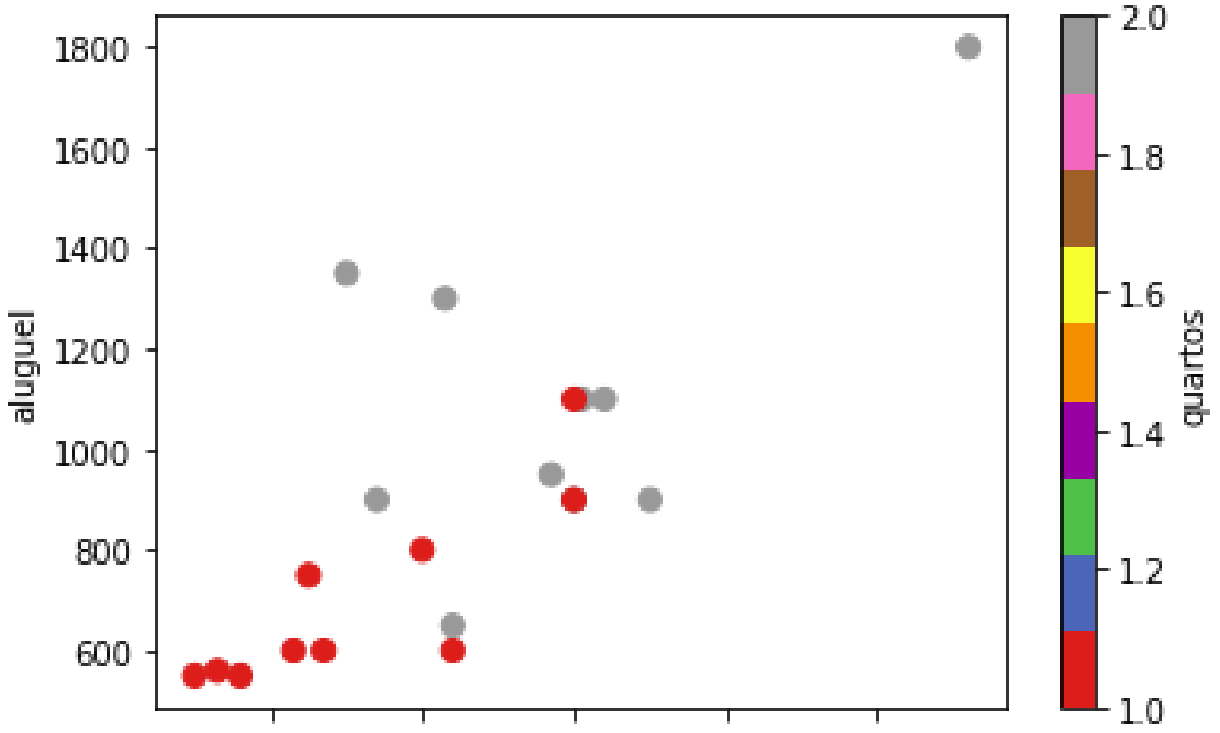
```
ax = df.plot.scatter(x='area', y='aluguel')
```



Scatter Plots

Podemos também usar cores para representar uma terceira variável como no exemplo abaixo (número de quartos).

```
ax = df.plot.scatter(x='area', y='aluguel',
                     s = 50, c='quartos', colormap='Set1')
```

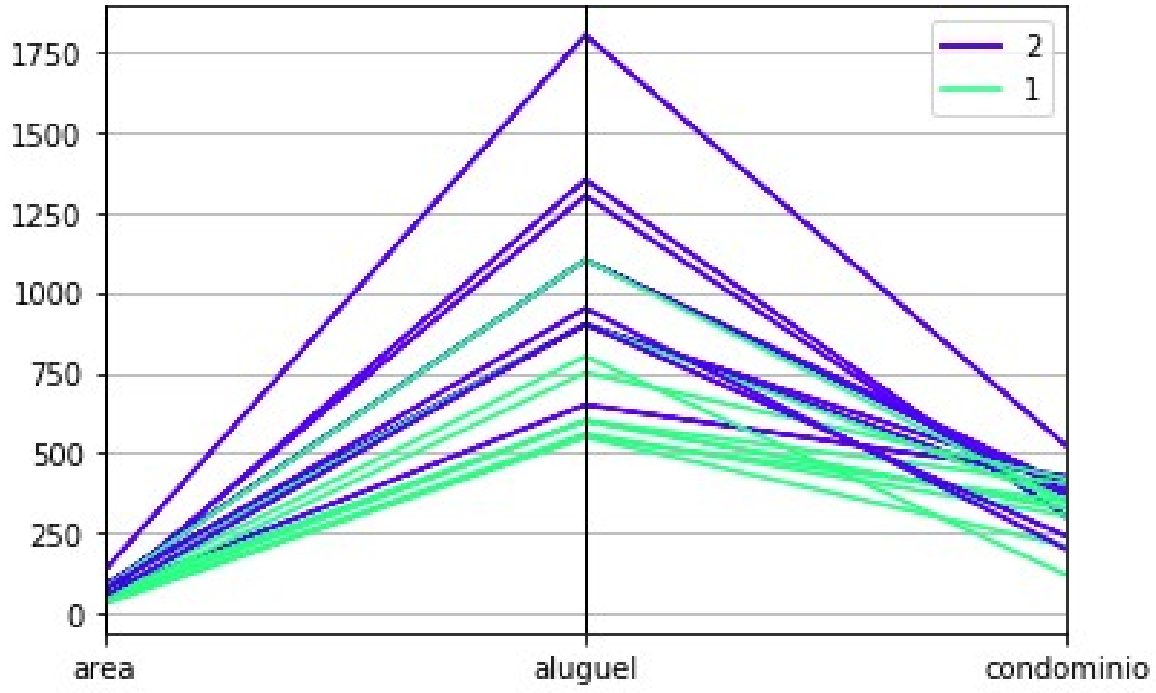


Parallel Coordinates

Uma visualização útil para comparar mais de duas variáveis é o Parallel Coordinates. Abaixo plotamos linhas em azul representando apartamentos com 2 quartos e linhas verdes para os de 1 quarto.

```
from pandas.plotting import parallel_coordinates

ax = parallel_coordinates(
    df[['area','aluguel','condominio', 'quartos']],
                        'quartos', colormap='winter')
```



Análise de Correlação

Entender como as variáveis estão correlacionadas é importante para a definição de hipóteses e para substanciar decisões na fase de construção do modelo. Abaixo usamos o método **corr()** para exibir as correlações entre todos os pares de variáveis.

```
df.corr()
```

	codigo	quartos	suite	area	vaga	aluguel	condominio
codigo	1.000000	-0.335195	-0.110545	-0.324856	-0.104901	-0.300360	0.070605
quartos	-0.335195	1.000000	0.229416	0.542466	-0.104828	0.619797	0.214173
suite	-0.110545	0.229416	1.000000	0.652274	0.312641	0.651048	0.470034
area	-0.324856	0.542466	0.652274	1.000000	0.533035	0.748196	0.466627
vaga	-0.104901	-0.104828	0.312641	0.533035	1.000000	0.251974	-0.087415
aluguel	-0.300360	0.619797	0.651048	0.748196	0.251974	1.000000	0.302494
condominio	0.070605	0.214173	0.470034	0.466627	-0.087415	0.302494	1.000000

Análise de Correlação

Abaixo podemos perceber que a correlação entre área e aluguel é alta. Já a correlação entre aluguel e condomínio, apesar de existir, não é tão expressiva.

```
df.corr()
```

	codigo	quartos	suite	area	vaga	aluguel	condominio
codigo	1.000000	-0.335195	-0.110545	-0.324856	-0.104901	-0.300360	0.070605
quartos	-0.335195	1.000000	0.229416	0.542466	-0.104828	0.619797	0.214173
suite	-0.110545	0.229416	1.000000	0.652274	0.312641	0.651048	0.470034
area	-0.324856	0.542466	0.652274	1.000000	0.533035	0.748196	0.466627
vaga	-0.104901	-0.104828	0.312641	0.533035	1.000000	0.251974	-0.087415
aluguel	-0.300360	0.619797	0.651048	0.748196	0.251974	1.000000	0.302494
condominio	0.070605	0.214173	0.470034	0.466627	-0.087415	0.302494	1.000000

Análise de Correlação

Uma forma útil de se visualizar correlações é através de mapas de calor, como no exemplo abaixo. Aqui, correlações mais altas possuem cor azul forte, enquanto as mais baixas tendem ao vermelho.

```
df_corr = df.corr()
df_corr = df_corr.style.background_gradient(cmap='RdBu')
df_corr
```

	codigo	quartos	suite	area	vaga	aluguel	condominio
codigo	1	-0.335195	-0.110545	-0.324856	-0.104901	-0.30036	0.0706052
quartos	-0.335195	1	0.229416	0.542466	-0.104828	0.619797	0.214173
suite	-0.110545	0.229416	1	0.652274	0.312641	0.651048	0.470034
area	-0.324856	0.542466	0.652274	1	0.533035	0.748196	0.466627
vaga	-0.104901	-0.104828	0.312641	0.533035	1	0.251974	-0.0874148
aluguel	-0.30036	0.619797	0.651048	0.748196	0.251974	1	0.302494
condominio	0.0706052	0.214173	0.470034	0.466627	-0.0874148	0.302494	1

Exercícios!

- Revise o conteúdo e faça os exercícios do notebook:
  03e-Pandas_Análise Exploratória.ipynb
- Faça os exercícios do notebook:
  03e1-Exercício-Pandas_Análise Exploratória