

Ciência dos Dados

Aula 02

Análise Exploratória dos Dados

Objetivos de Aprendizagem

Os alunos devem ser capazes de:

- Analisar variáveis qualitativas de forma uni e bivariadas utilizando tabelas de frequências e gráficos.

Aula de hoje

- Mostrar novo exemplo para definir distribuição conjunta e distribuição condicional (consequência de uma tabela cruzada).

Empresa de TV

PROBLEMA:

- 👉 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 cada plano para enviar malas diretas aos potenciais clientes de cada tipo de plano.

Empresa de TV

Essa 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, A ou B.
- *EC: apresenta estado civil do cliente no momento da adesão ao plano, Casado, Solteiro e Outros.
- *SATISFACAO: grau de satisfação do cliente pelo plano, Muito satisfeito, Satisfeito, Indiferente, Insatisfeito e Muito insatisfeito.
- *RENDA: renda pessoal do cliente, em milhares de reais.

Explorando cada variável qualitativa

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()`

Explorando cada variável qualitativa

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:

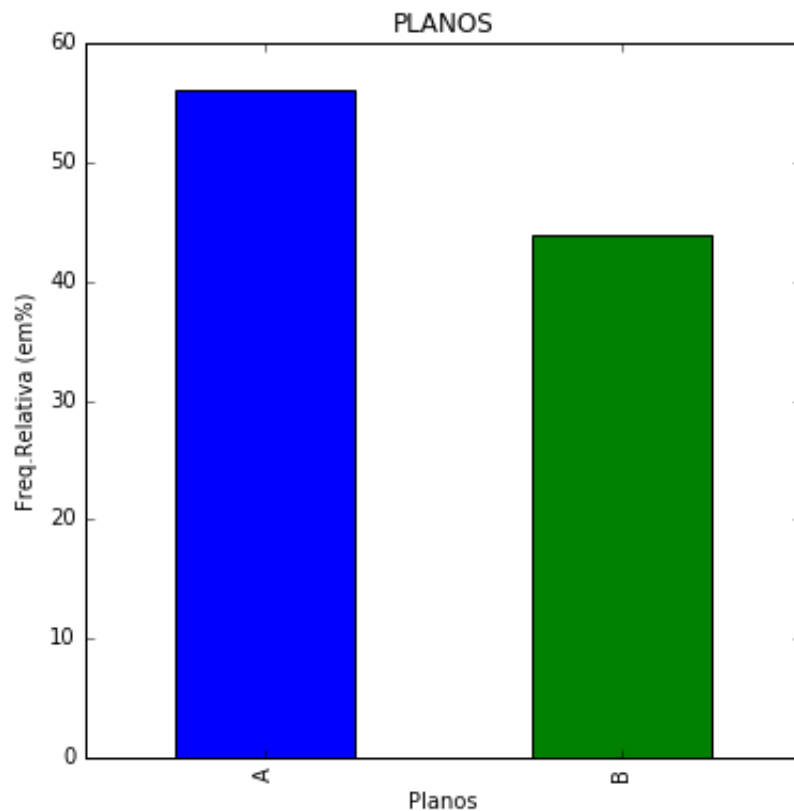
Muito Insatisfeito	9,8
Insatisfeito	19,5
Indiferente	23,2
Satisfeito	32,9
Muito Satisfeito	14,6

Comando Python:

`variável.value_counts(True)*100` ₇

Explorando cada variável qualitativa

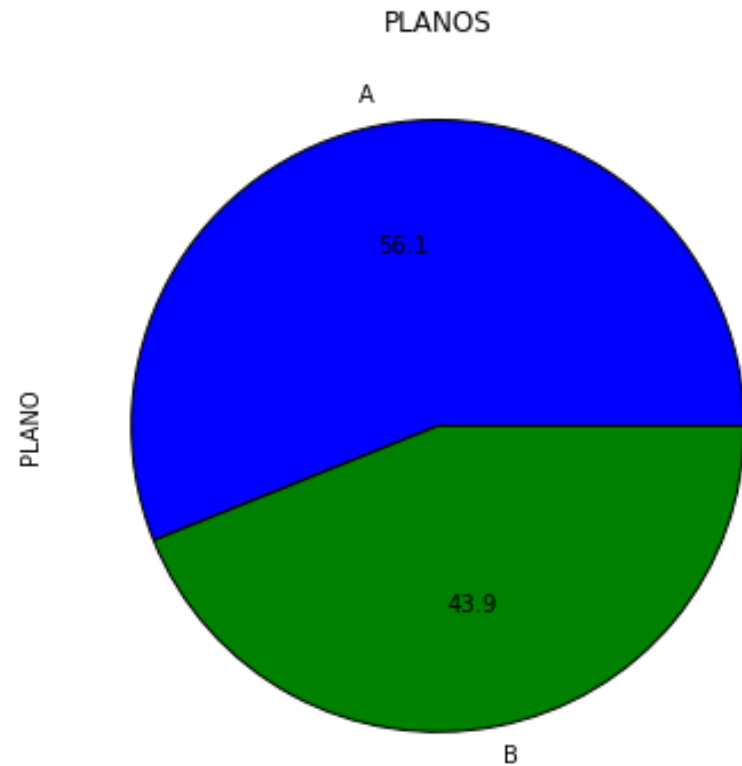
Gráfico de Barras



Comando Python:

```
variável.value_counts(True).plot(kind='bar')
```

Gráfico de Pizza

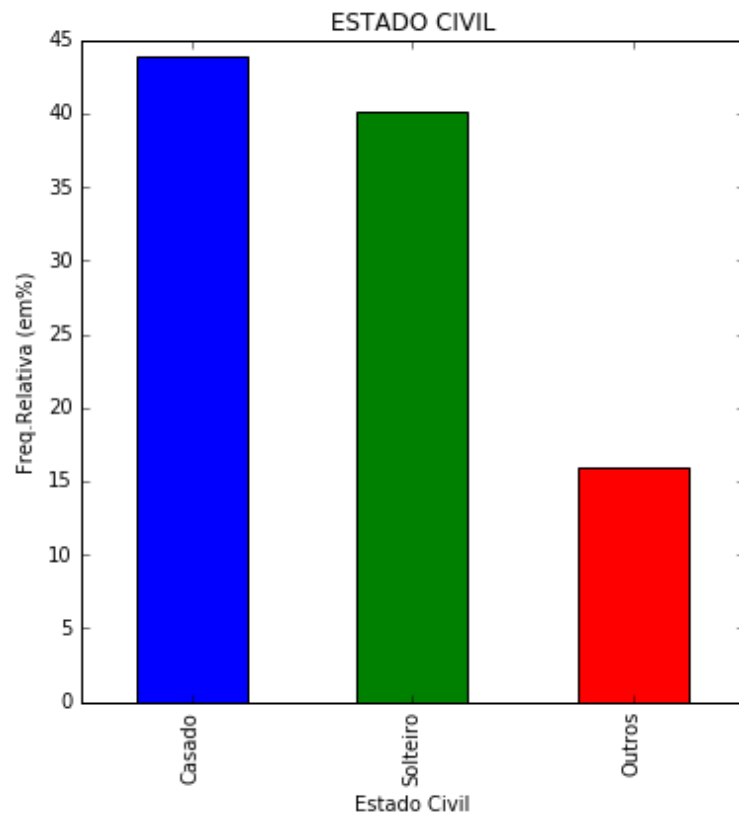


Comando Python:

```
variável.value_counts (True).plot(kind='pie')
```


Explorando cada variável qualitativa

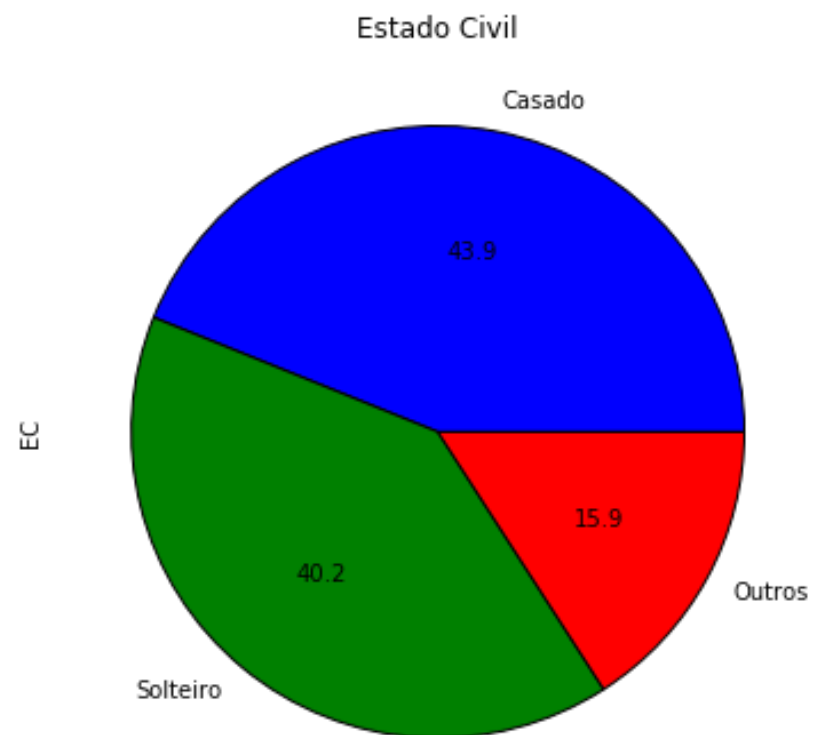
Gráfico de Barras



Comando Python:

```
variável.value_counts(True).plot(kind='bar')
```

Gráfico de Pizza

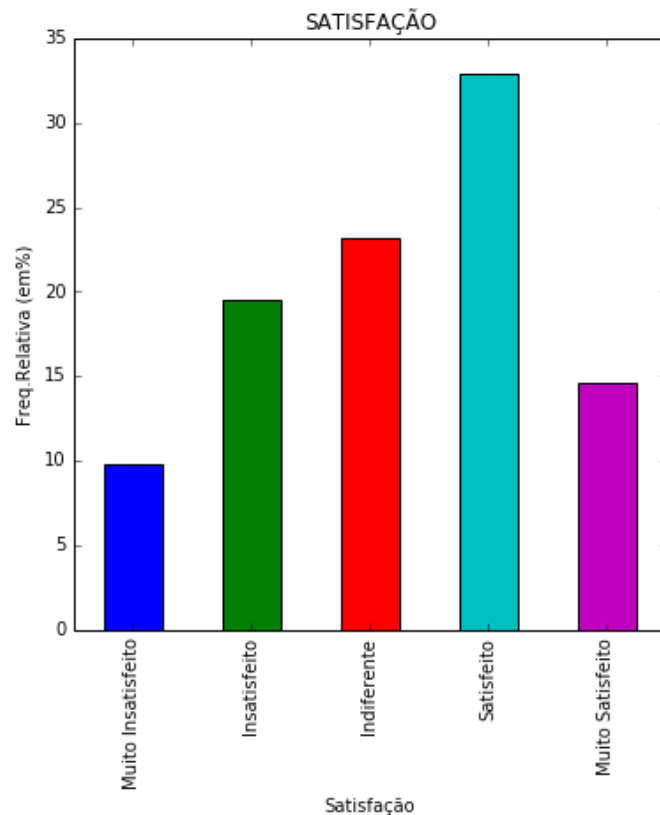


Comando Python:

```
variável.value_counts (True).plot(kind='pie')
```

Explorando cada variável qualitativa

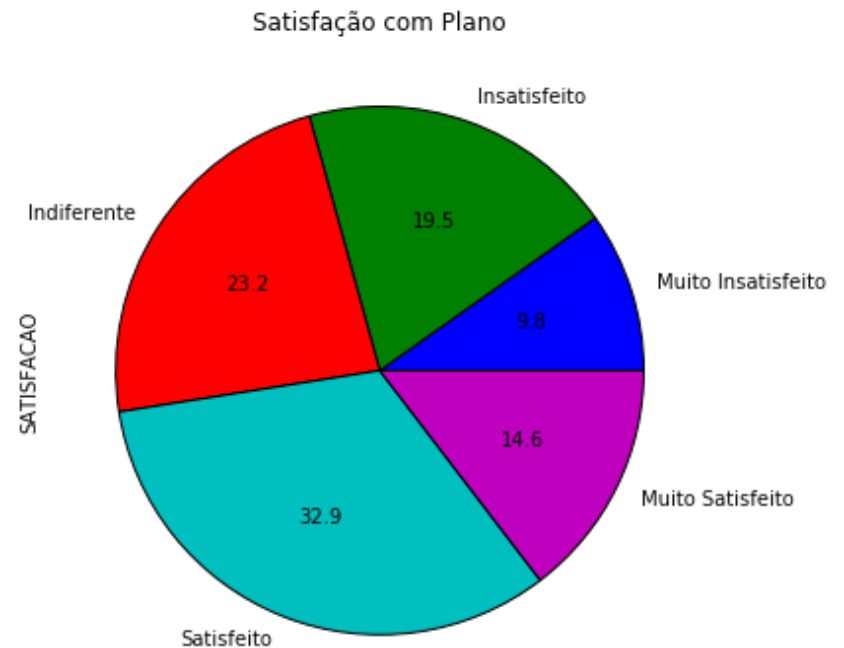
Gráfico de Barras



Comando Python:

```
variável.value_counts(True).plot(kind='bar')
```

Gráfico de Pizza



Comando Python:

```
variável.value_counts (True).plot(kind='pie')
```

Explorando tabela cruzada

Considerando os 82 clientes que fazem parte da amostra, os resultados da tabela dividem os clientes quanto as variáveis Estado Civil e Plano.

Estado Civil	Plano		Total
	A	B	
Casado	26	10	36
Solteiro	13	20	33
Outros	7	6	13
Total	46	36	82

Comando Python:

```
import pandas as pd  
pd.crosstab(variável linha, variável coluna)
```

Explorando tabela cruzada

Distribuição marginal: avaliação do comportamento dos clientes em uma variável.

Distribuição conjunta: avaliação do comportamento conjunto dos clientes nas duas variáveis.

Estado Civil	Plano		Total
	A	B	
Casado	26	10	36
Solteiro	13	20	33
Outros	7	6	13
Total	46	36	82

Distribuição conjunta

Distribuição marginal

Explorando tabela cruzada

Frequências relativas associadas ao problema:

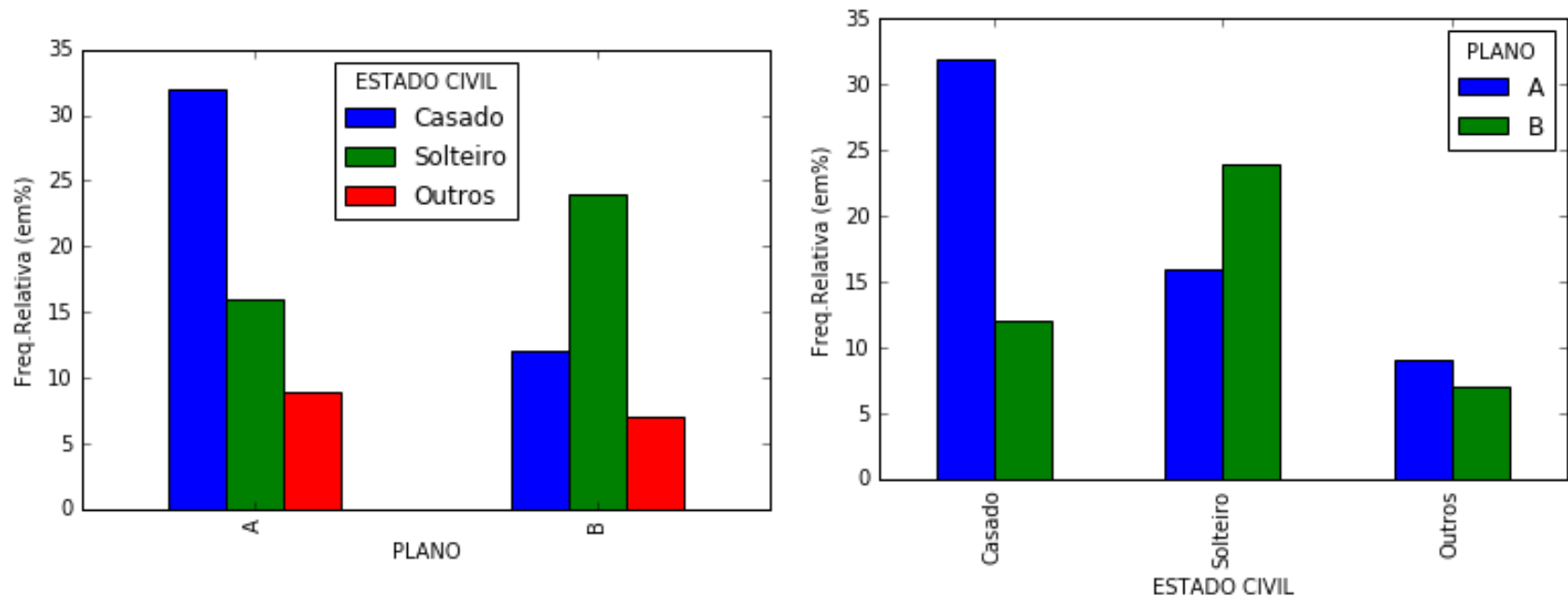
Estado Civil	Plano		Total
	A	B	
Casado	32%	12%	44%
Solteiro	16%	24%	40%
Outros	9%	7%	16%
Total	57%	43%	100%

Em vermelho: frequências relativas conjuntas

Em azul: frequências relativas marginais

Explorando tabela cruzada

Gráfico de Barras (% no total geral)



Comando Python:

```
pd.crosstab(variável linha, variável coluna).plot(kind='bar')
```

Explorando tabela cruzada

Frequências relativas associadas ao problema:

Estado Civil	Plano		Total
	A	B	
Casado	72%	28%	100%
Solteiro	39%	61%	100%
Outros	54%	46%	100%
Total	57%	43%	100%

Em verde: frequências relativas por linha

Em azul: frequências relativas marginais

Explorando tabela cruzada

Frequências relativas associadas ao problema:

Estado Civil	Plano		Total
	A	B	
Casado	57%	28%	44%
Solteiro	28%	56%	40%
Outros	15%	17%	16%
Total	100%	100%	100%

Em laranja: frequências relativas por coluna

Em azul: frequências relativas marginais

Explorando tabela cruzada

Gráfico de Barras

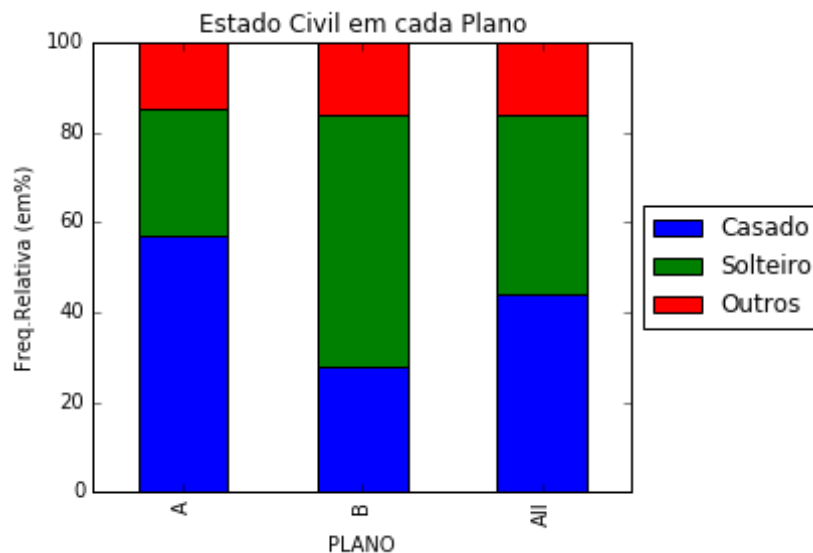
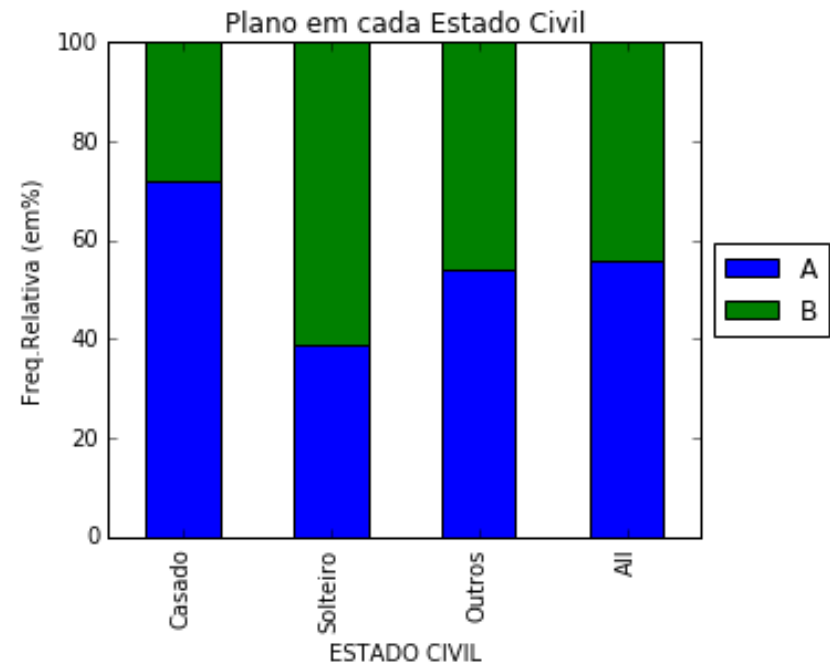


Gráfico de Barras



Comando Python:

```
pd.crosstab(variável linha, variável coluna).plot(kind='bar', stacked=True)
```

Explorando tabela cruzada

CUIDADO com gráficos abaixo!!

Gráfico de Barras

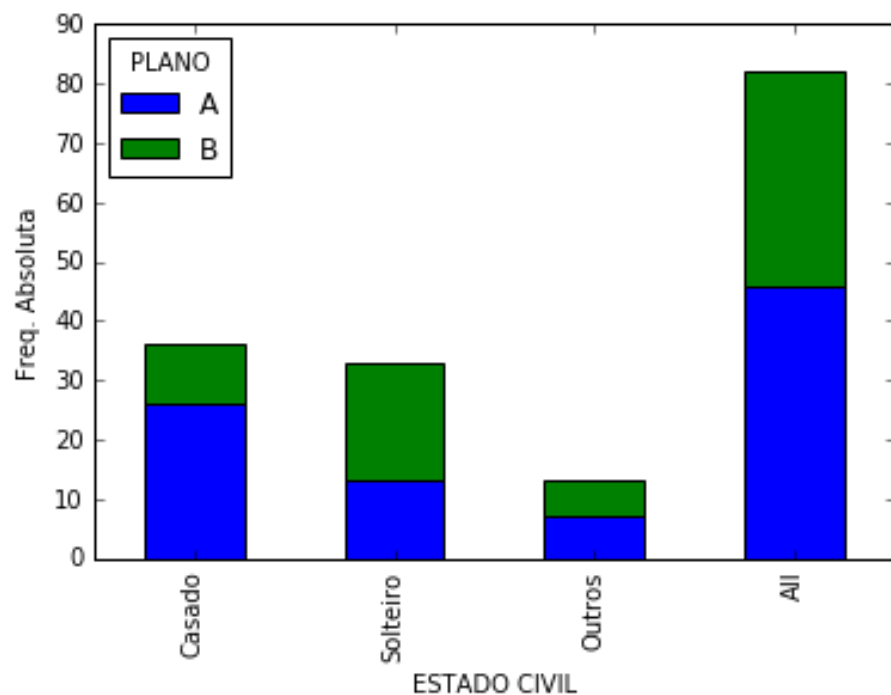
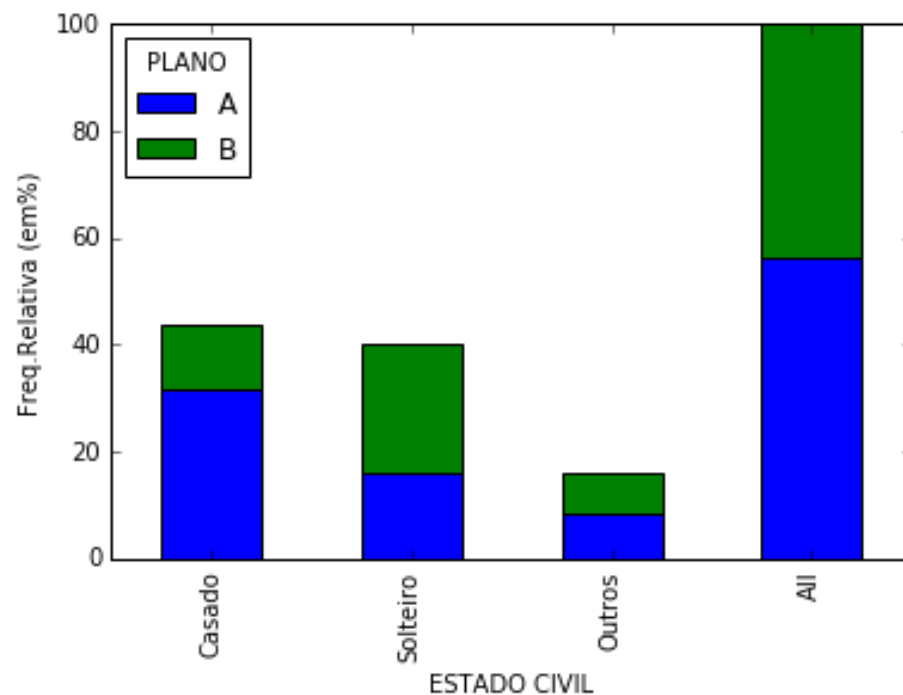


Gráfico de Barras



Explorando tabela cruzada

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)`

Preparo para próxima aula

Os alunos devem se preparar com:

1. Leitura prévia necessária: Magalhães e Lima (7ª. Edição): pág. 9 a 16 – destacando para variáveis quantitativas.
2. Python.