Registro da Análise e Limpeza dos dados

Fonte de Dados: UCI - Adults
http://archive.ics.uci.edu/ml/datasets/Adult

Objetivo

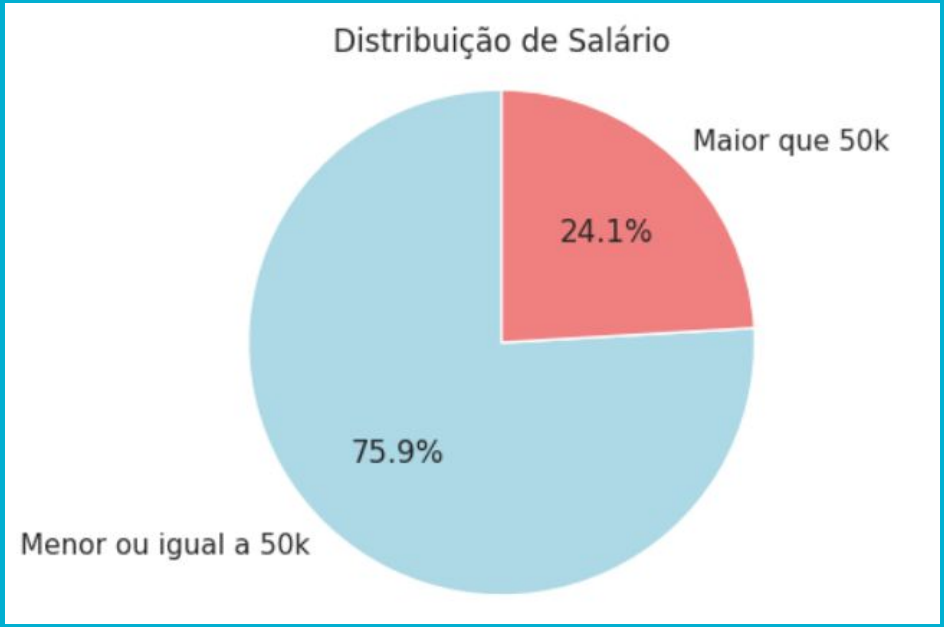
- Avaliar as características do perfil para apontar se a renda anual será até 50k ou acima.
- Precisamos construir uma Máquina Preditiva que, para ser usada na classificação de uma pessoa para participar ou não de um programa de intercâmbio.

Apresentação dos dados

```
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
```

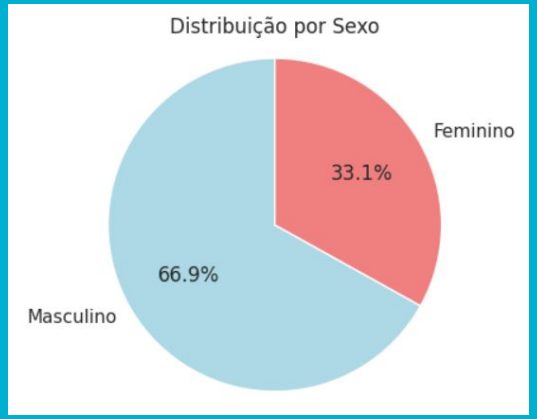
- Dados retirados de um censo de 1994 realizado nos EUA.
- 32.561 registro
- 14 colunas
- Dicionário de dados
  - age - Idade
  - workclass - Classificação do trabalho
  - fnlwgt - valor sequencial
  - education - nível educacional
  - education-num - valor relacionado ao nivel educacional
  - marital-status - Estado Civil
  - occupation - Ocupação
  - relationship - RElação da pessoa com o(a) dono(a) da residência.
  - race - Raça
  - sex - Sexo
  - capital-gain - Ganho de Capital
  - capital-loss - Perda de Capital
  - hours-per-week - Quantidade de horas trabalhadas por semana
  - native-country - Nacionalidae
  - salary - Renda anual de até 50k ou acima.

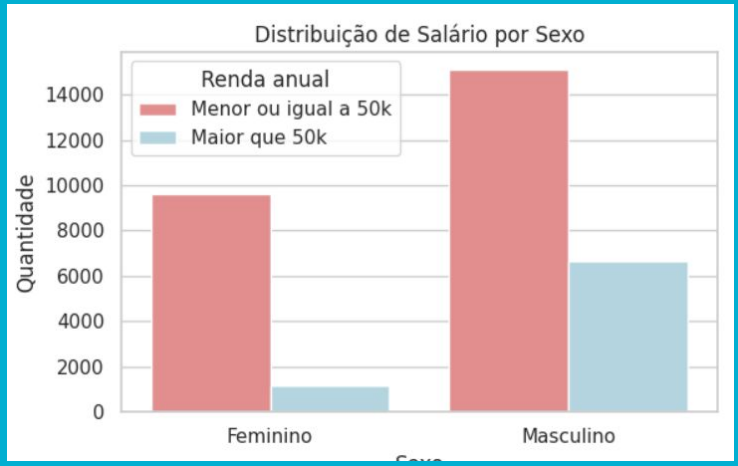
Distribuição da Renda Anual

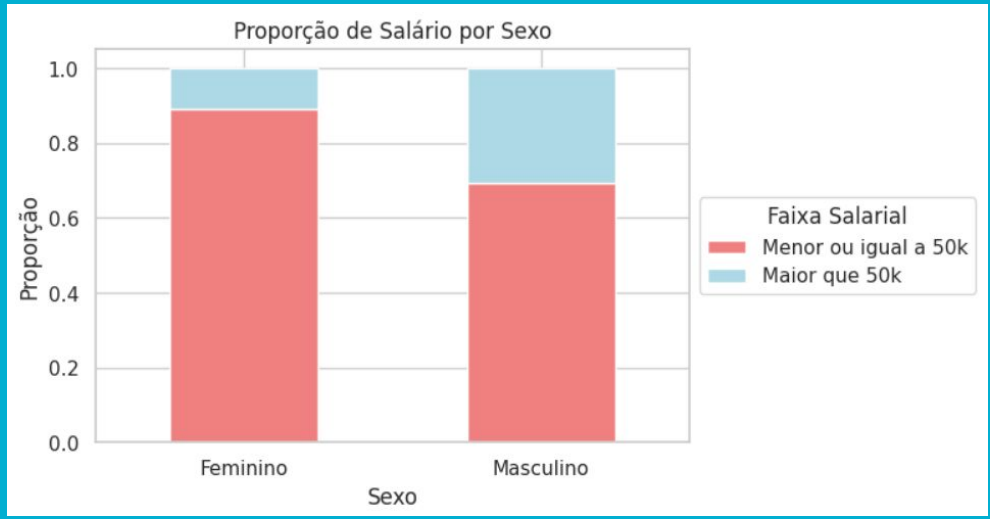


- ¾ das pessoas avaliadas possuem uma renda anual igual ou abaixo de 50k e apenas ¼ possui renda anual acima de 50k.

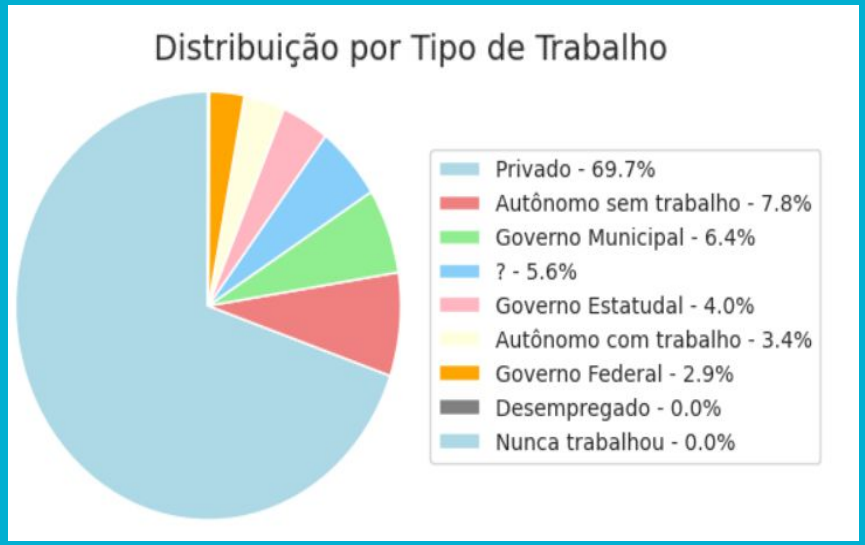
Sexo x Renda Anual

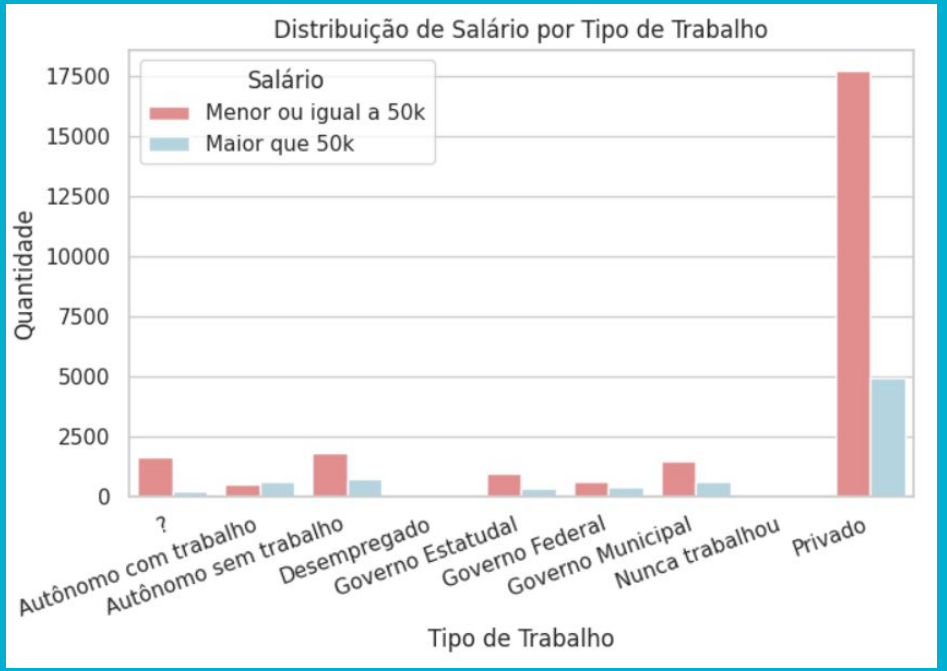






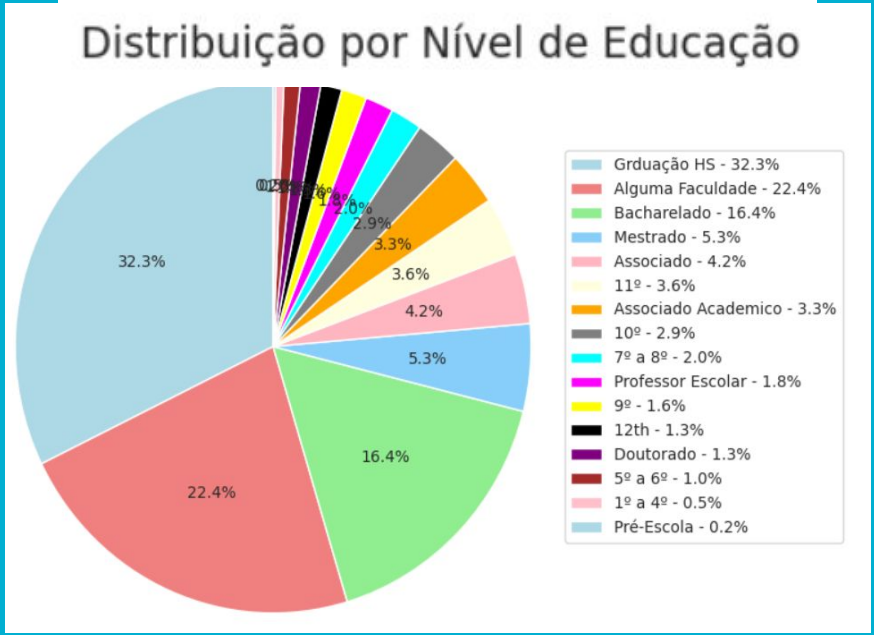
Tipo de Trabalho x Renda Anual

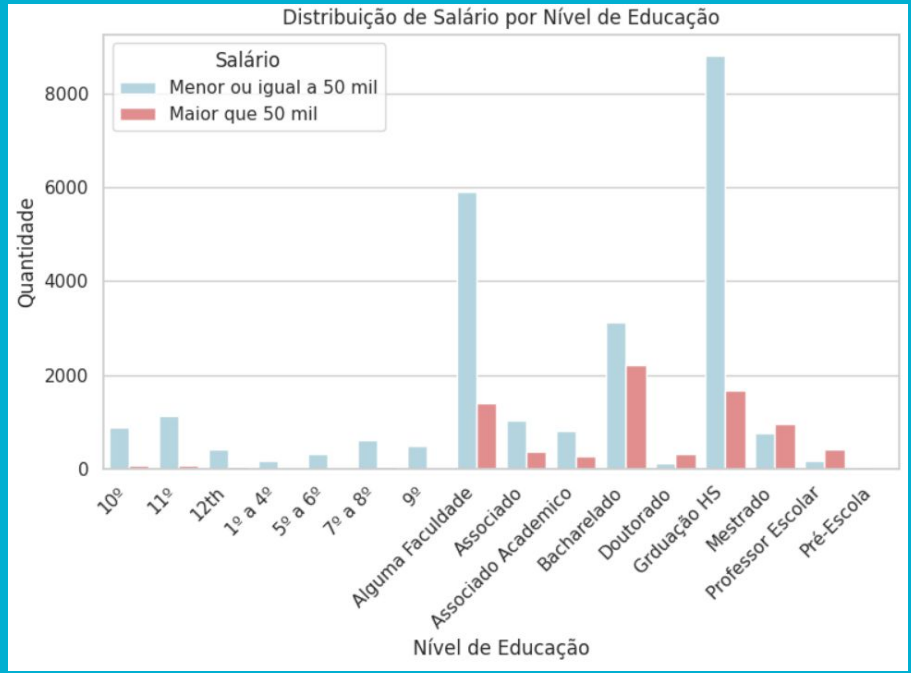




- Tratar ?

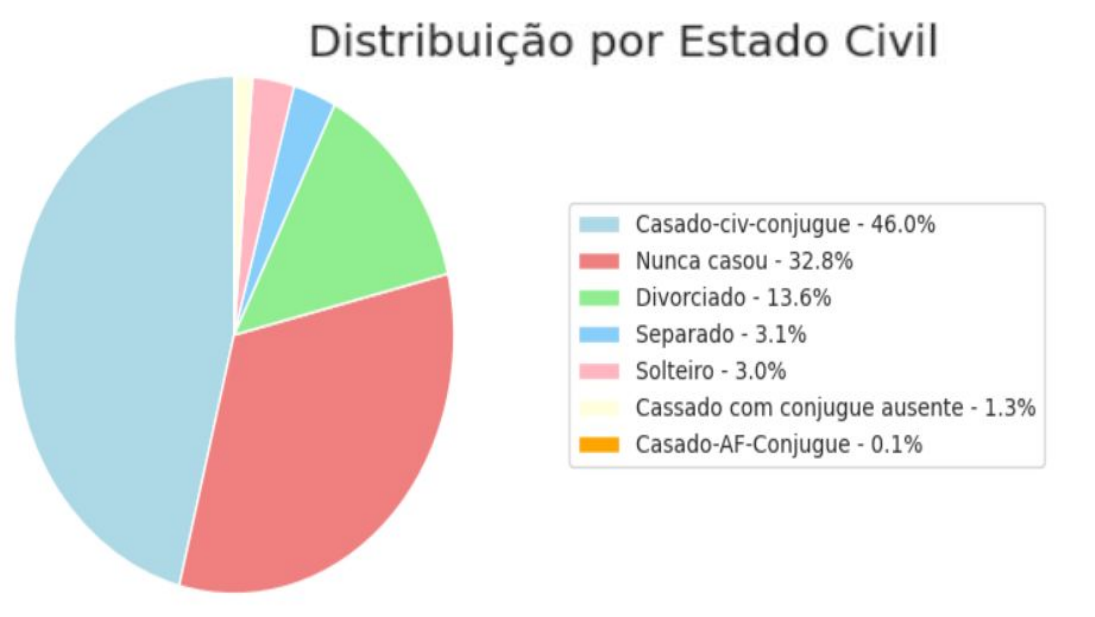
Nivel Escolar x Renda Anual

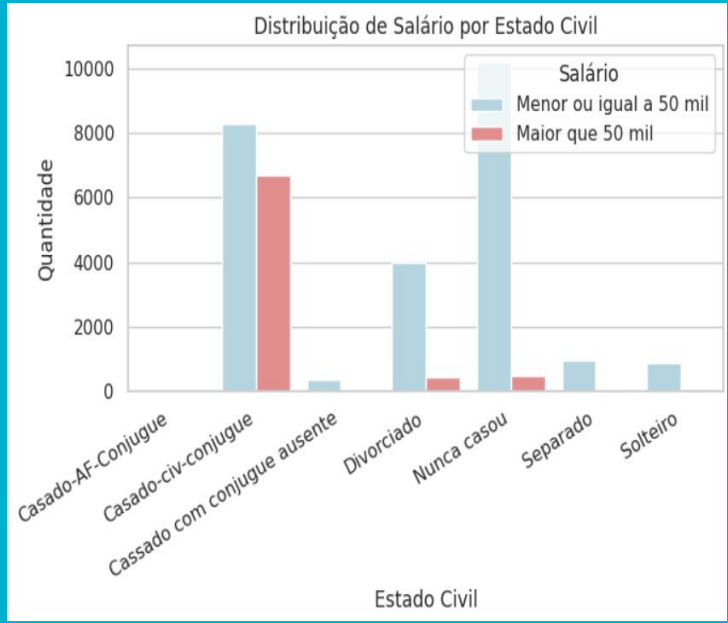




- Enquadrar series nos niveis adequados

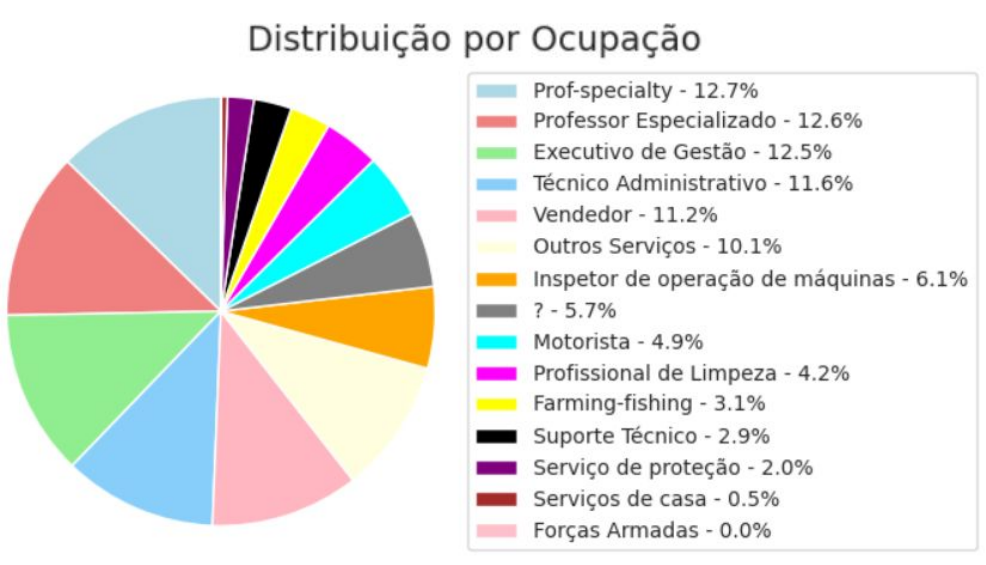
Estado Civil x Renda Anual

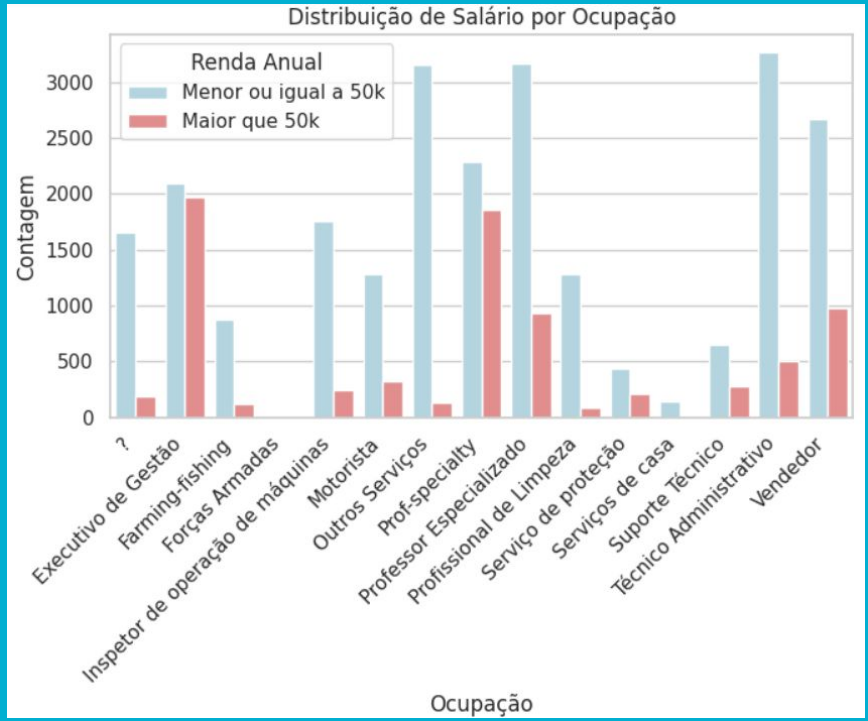




- Unir status de casado
- Unir nunca casou com solteiro

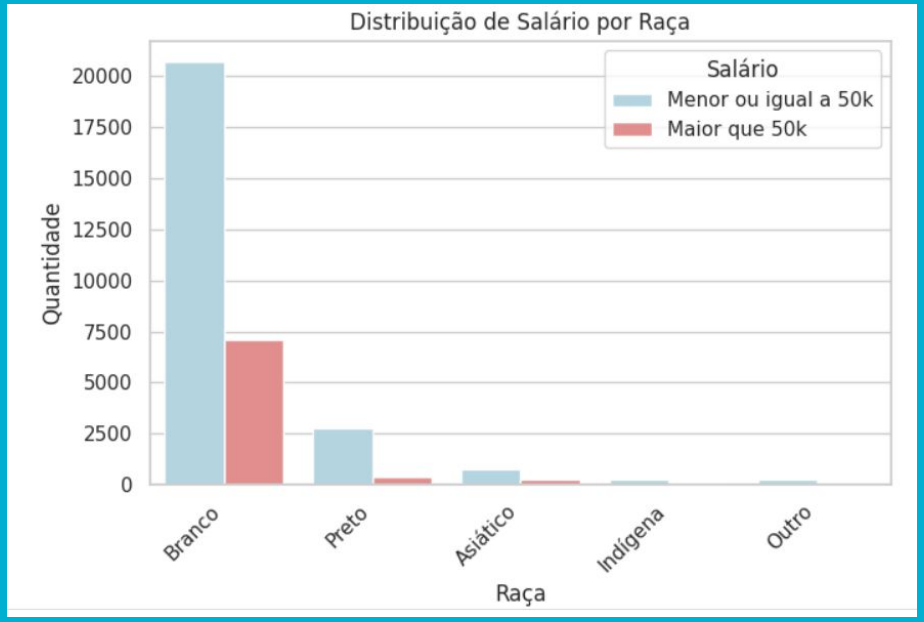
Ocupação x Renda Anual





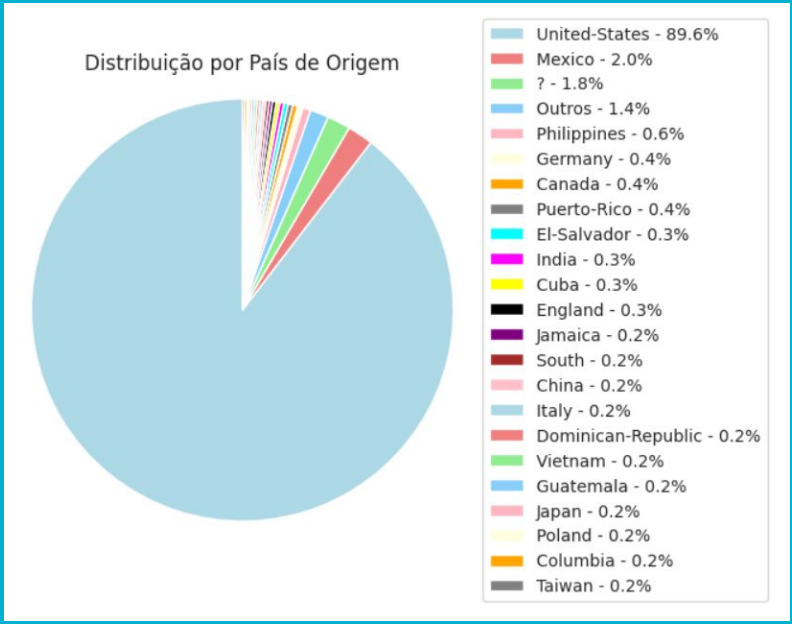
- Nesse caso o ? pode ser tratado como outros serviços.
-

Raça x Renda Anual



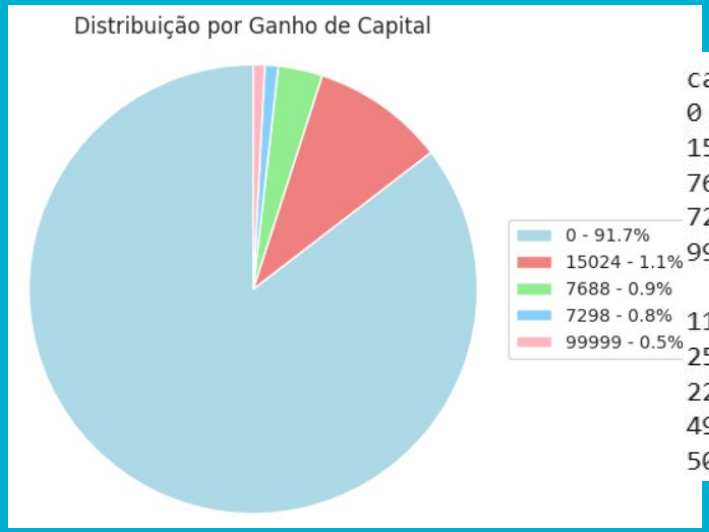
-
-

Nacionalidade

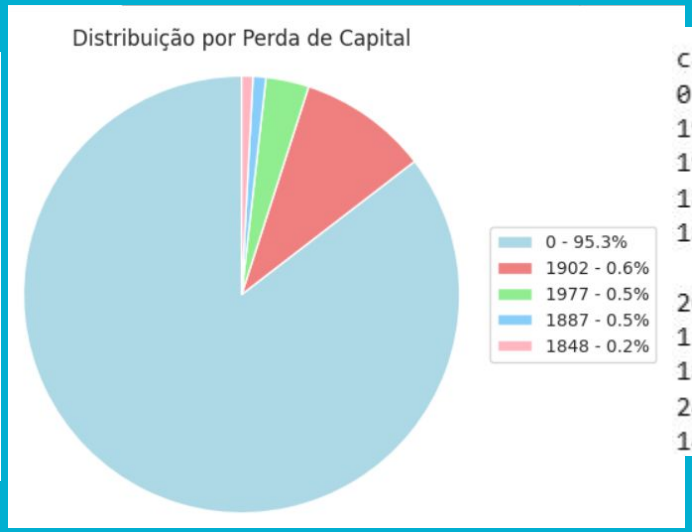


- 90% das pessoas são naturais do USA.
-

Ganho e Perda de Capital



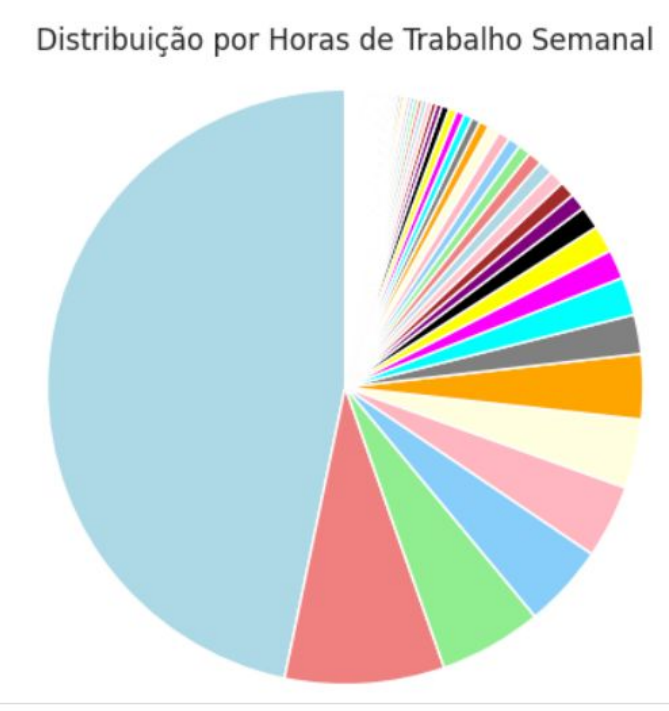
capital-gain	
0	91.67%
15024	1.07%
7688	0.87%
7298	0.76%
99999	0.49%
...	
1111	0.00%
2538	0.00%
22040	0.00%
4931	0.00%
5060	0.00%



capital-loss	
0	95.33%
1902	0.62%
1977	0.52%
1887	0.49%
1848	0.16%
...	
2080	0.00%
1539	0.00%
1844	0.00%
2489	0.00%
1411	0.00%

- 92% não teve ganho nenhum de capital
- 95% não teve perda de capital nenhuma

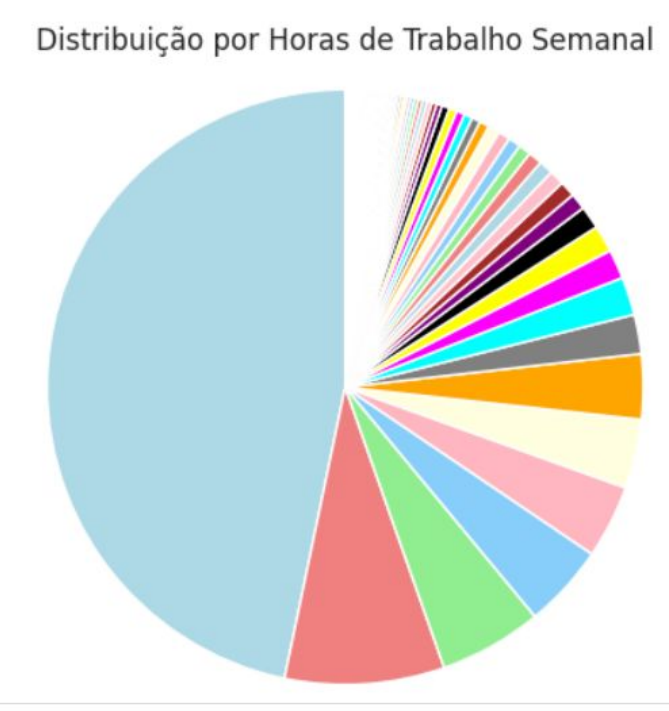
Horas por semana



hours-per-week	
40	46.73%
50	8.66%
45	5.60%
60	4.53%
35	3.98%
	...
82	0.00%
92	0.00%
87	0.00%
74	0.00%
94	0.00%

- 46% trabalham 40 horas semanais e os outros 54% estão distribuídos em 96 quantidade de horas trabalhadas diferentes

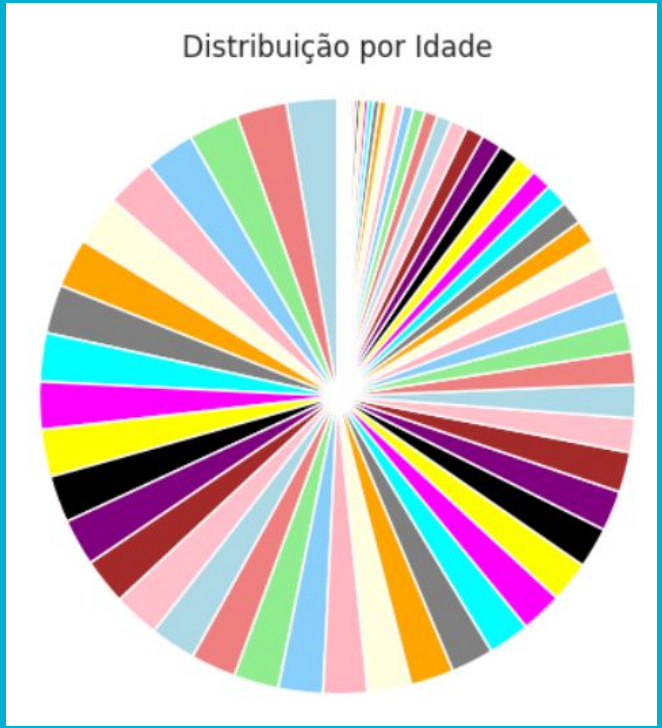
Horas por semana



hours-per-week	
40	46.73%
50	8.66%
45	5.60%
60	4.53%
35	3.98%
	...
82	0.00%
92	0.00%
87	0.00%
74	0.00%
94	0.00%

- 46% trabalham 40 horas semanais e os outros 54% estão distribuídos em 96 quantidade de horas trabalhadas diferentes

Idade




```
age
36     2.76%
s na saída da célula de código
34     2.72%
23     2.69%
35     2.69%
       ...
83     0.02%
88     0.01%
85     0.01%
86     0.00%
87     0.00%
Name: count, Length: 73, dtype: object
```

- Os dados de idade estão distribuídos em 73 valores distintos.

Conclusão

Variáveis que podem ser excluídas por não contribuírem para a definição de renda maior ou menor a 50k.

- Age
- fnlwgt
- education num
- relationship
- capital gain
- capital loss
- hour per week
- native country

Conclusão

Conteúdo a ser tratado:

- workclass - tratar ?
- em workclass unir ‘whithout-pay’ com ‘never-worked’
- education tratar as séries como ensino incompleto
- marital-status - unir todos os cassados.
- occupation - tratar ?