
CURSO DE ESTATÍSTICA - PARTE 1

Trabalho de Análise Descritiva de um Conjunto de Dados

Utilizando os conhecimentos adquiridos em nosso treinamento realize uma análise descritiva básica de um conjunto de dados retirados da Pesquisa Nacional por Amostra de Domicílios - 2015 do IBGE.

Vamos construir histogramas, calcular e avaliar medidas de tendência central, medidas separatrizes e de dispersão dos dados.

Siga o roteiro proposto e vá completando as células vazias. Procure pensar em mais informações interessantes que podem ser exploradas em nosso dataset.

DATASET DO PROJETO

Pesquisa Nacional por Amostra de Domicílios - 2015

A **Pesquisa Nacional por Amostra de Domicílios - PNAD** investiga anualmente, de forma permanente, características gerais da população, de educação, trabalho, rendimento e habitação e outras, com periodicidade variável, de acordo com as necessidades de informação para o país, como as características sobre migração, fecundidade, nupcialidade, saúde, segurança alimentar, entre outros temas. O levantamento dessas estatísticas constitui, ao longo dos 49 anos de realização da pesquisa, um importante instrumento para formulação, validação e avaliação de políticas orientadas para o desenvolvimento socioeconômico e a melhoria das condições de vida no Brasil.

Fonte dos Dados

<https://ww2.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2015/microdados.shtm>

Variáveis utilizadas

Renda

Rendimento mensal do trabalho principal para pessoas de 10 anos ou mais de idade.

Idade

Idade do morador na data de referência em anos.

Altura (elaboração própria)

Altura do morador em metros.

UF

Código	Descrição
11	Rondônia
12	Acre
13	Amazonas
14	Roraima
15	Pará
16	Amapá
17	Tocantins
21	Maranhão
22	Piauí
23	Ceará
24	Rio Grande do Norte
25	Paraíba
26	Pernambuco
27	Alagoas
28	Sergipe
29	Bahia
31	Minas Gerais
32	Espírito Santo
33	Rio de Janeiro
35	São Paulo
41	Paraná
42	Santa Catarina
43	Rio Grande do Sul
50	Mato Grosso do Sul
51	Mato Grosso
52	Goiás
53	Distrito Federal

Sexo

Código	Descrição
0	Masculino
1	Feminino

Anos de Estudo

Código	Descrição
1	Sem instrução e menos de 1 ano
2	1 ano

Código	Descrição
3	2 anos
4	3 anos
5	4 anos
6	5 anos
7	6 anos
8	7 anos
9	8 anos
10	9 anos
11	10 anos
12	11 anos
13	12 anos
14	13 anos
15	14 anos
16	15 anos ou mais
17	Não determinados
	Não aplicável

Cor

Código	Descrição
0	Indígena
2	Branca
4	Preta
6	Amarela
8	Parda
9	Sem declaração

Observação

Os seguintes tratamentos foram realizados nos dados originais:

1. Foram eliminados os registros onde a **Renda** era inválida (999 999 999 999);
2. Foram eliminados os registros onde a **Renda** era missing;
3. Foram considerados somente os registros das **Pessoas de Referência** de cada domicílio (responsável pelo domicílio).

Utilize a célula abaixo para importar as biblioteca que precisar para executar as tarefas

Sugestões: **pandas, numpy, seaborn**

```
In [1]: import pandas as pd
```

```
import numpy as np
import seaborn as sns
```

Importe o dataset e armazene o conteúdo em uma DataFrame

```
In [2]: dados = pd.read_csv('dados.csv')
```

Visualize o conteúdo do DataFrame

```
In [3]: dados.head()
```

```
Out[3]:
```

	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	0	23	8	12	800	1.603808
1	11	1	23	2	12	1150	1.739790
2	11	1	35	8	15	880	1.760444
3	11	0	46	2	6	3500	1.783158
4	11	1	47	8	9	150	1.690631

Para avaliarmos o comportamento da variável RENDA vamos construir uma tabela de frequências considerando as seguintes classes em salários mínimos (SM)

Descreva os pontos mais relevantes que você observa na tabela e no gráfico.

Classes de renda:

A ► Acima de 25 SM

B ► De 15 a 25 SM

C ► De 5 a 15 SM

D ► De 2 a 5 SM

E ► Até 2 SM

Para construir as classes de renda considere que o salário mínimo na época da pesquisa era de **R\$ 788,00**.

Siga os passos abaixo:

1º Definir os intervalos das classes em reais (R\$)

```
In [4]: sm = 788
classes = [dados.Renda.min(),
           2 * sm,
           5 * sm,
           15 * sm,
           25 * sm,
           dados.Renda.max()]

classes
```

```
Out[4]: [0, 1576, 3940, 11820, 19700, 200000]
```

2º Definir os labels das classes

```
In [5]: labels = ['E', 'D', 'C', 'B', 'A']
```

3º Construir a coluna de frequências

```
In [6]: frequencia = pd.value_counts(  
    pd.cut(x = dados.Renda,  
        bins = classes,  
        labels = labels,  
        include_lowest = True)  
    )  
frequencia
```

```
Out[6]: E    49755  
D    18602  
C     7241  
B      822  
A     420  
Name: Renda, dtype: int64
```

4º Construir a coluna de percentuais

```
In [7]: percentual = pd.value_counts(  
    pd.cut(x = dados.Renda,  
        bins = classes,  
        labels = labels,  
        include_lowest = True),  
    normalize = True)  
    * 100  
percentual
```

```
Out[7]: E    64.751432  
D    24.208745  
C     9.423477  
B     1.069755  
A     0.546590  
Name: Renda, dtype: float64
```

5º Juntar as colunas de frequência e percentuais e ordenar as linhas de acordo com os labels das classes

```
In [8]: dist_freq_renda = pd.DataFrame({  
    'Frequência': frequencia,  
    'Porcentagem (%)': percentual  
    })  
dist_freq_renda.sort_index(ascending = False)
```

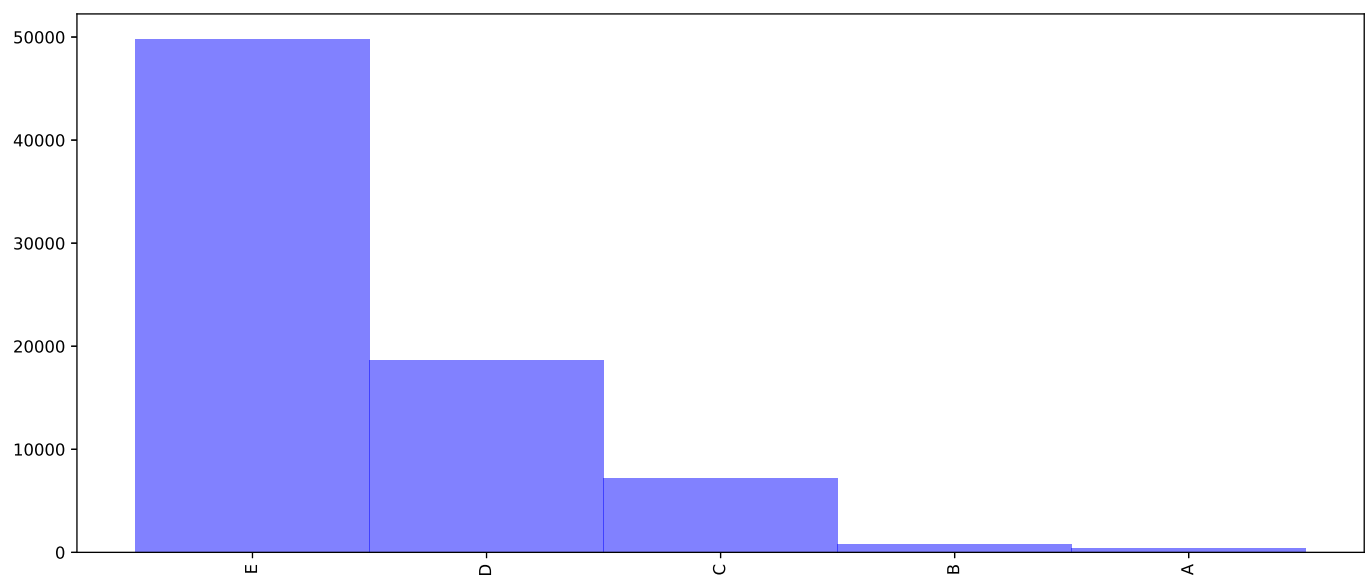
```
Out[8]:
```

	Frequência	Porcentagem (%)
A	420	0.546590
B	822	1.069755
C	7241	9.423477
D	18602	24.208745
E	49755	64.751432

Construa um gráfico de barras para visualizar as informações da tabela de frequências acima

```
In [9]: dist_freq_renda['Frequência'].plot.bar(width = 1, color = 'blue', alpha = 0.5, figsize=(14, 6))
```

Out[9]: <AxesSubplot:>



Conclusões

Conforme o gráfico mostra, a maioria da população se encontra na classe E (até 2 salários mínimos). Conforme a classe salarial aumenta a população em cada classe diminui consideravelmente. Até a classe B e A (de 15 a 25, e acima de 25 salários mínimos respectivamente).

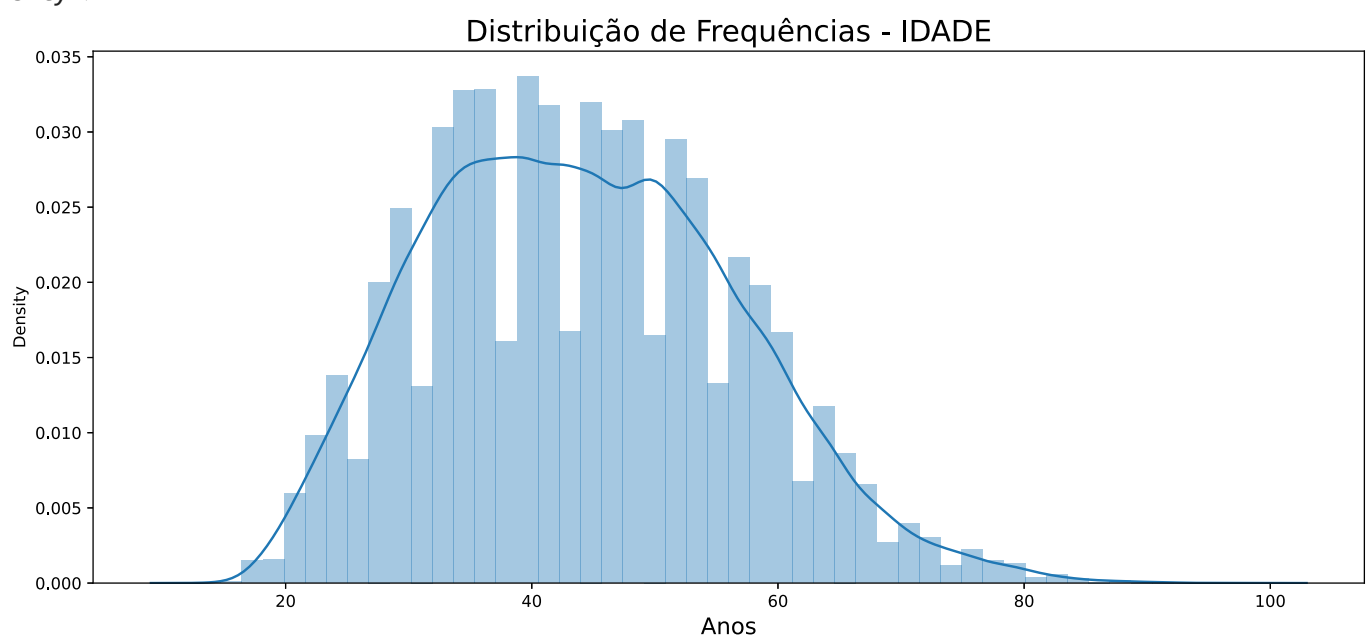
Crie um histograma para as variáveis QUANTITATIVAS de nosso dataset

Descreva os pontos mais relevantes que você observa nos gráficos (assimetrias e seus tipos, possíveis causas para determinados comportamentos etc.)

In [10]:

```
ax = sns.distplot(dados['Idade'])
ax.figure.set_size_inches(14, 6)
ax.set_title('Distribuição de Frequências - IDADE', fontsize=18)
ax.set_xlabel('Anos', fontsize=14)
ax
```

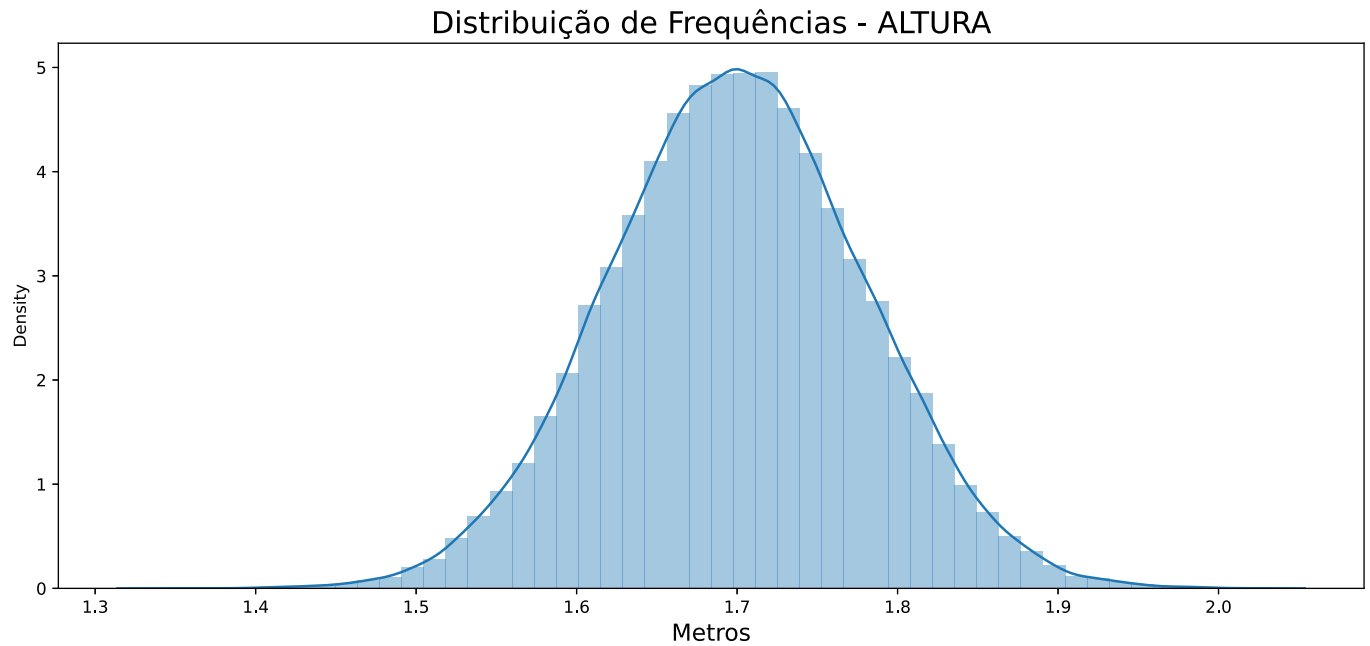
Out[10]: <AxesSubplot:title={'center':'Distribuição de Frequências - IDADE'}, xlabel='Anos', ylabel='Density'>



In [11]:

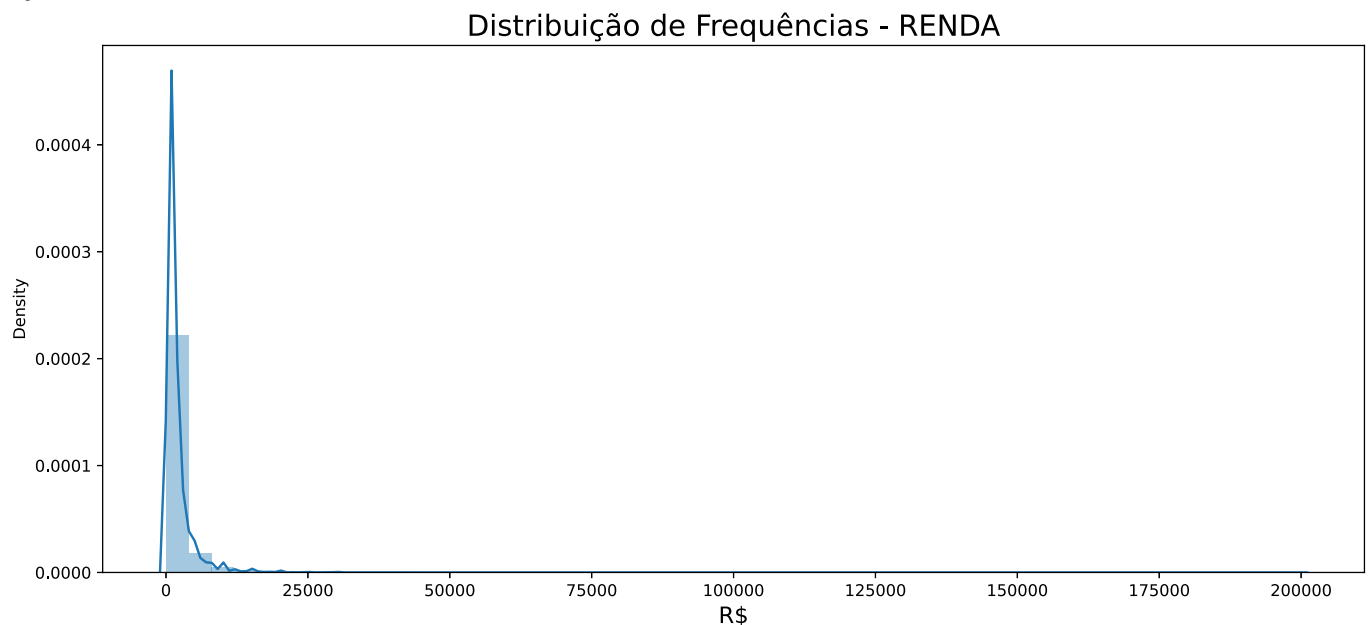
```
ax = sns.distplot(dados['Altura'])
ax.figure.set_size_inches(14, 6)
ax.set_title('Distribuição de Frequências - ALTURA', fontsize=18)
ax.set_xlabel('Metros', fontsize=14)
ax
```

```
Out[11]: <AxesSubplot:title={'center':'Distribuição de Frequências - ALTURA'}, xlabel='Metros', ylabel='Density'>
```



```
In [12]: ax = sns.distplot(dados['Renda'])
ax.figure.set_size_inches(14, 6)
ax.set_title('Distribuição de Frequências - RENDA', fontsize=18)
ax.set_xlabel('R$', fontsize=14)
ax
```

```
Out[12]: <AxesSubplot:title={'center':'Distribuição de Frequências - RENDA'}, xlabel='R$', ylabel='Density'>
```



Conclusões

Distribuição de Frequência - Idade:

Considerando que os dados mostram as respostas dadas pelos chefes das famílias o gráfico mostra da independência dos entrevistados. Poucas pessoas de 20 anos são chefes da própria família, esse número aumenta com a idade, chegando ao topo próximo aos 40 anos.

Distribuição de Frequência - Altura:

A distribuição de altura está bem uniforme na população observada. A maioria das pessoas tendo por volta de 1,7 metros.

Distribuição de Frenquência - Renda:

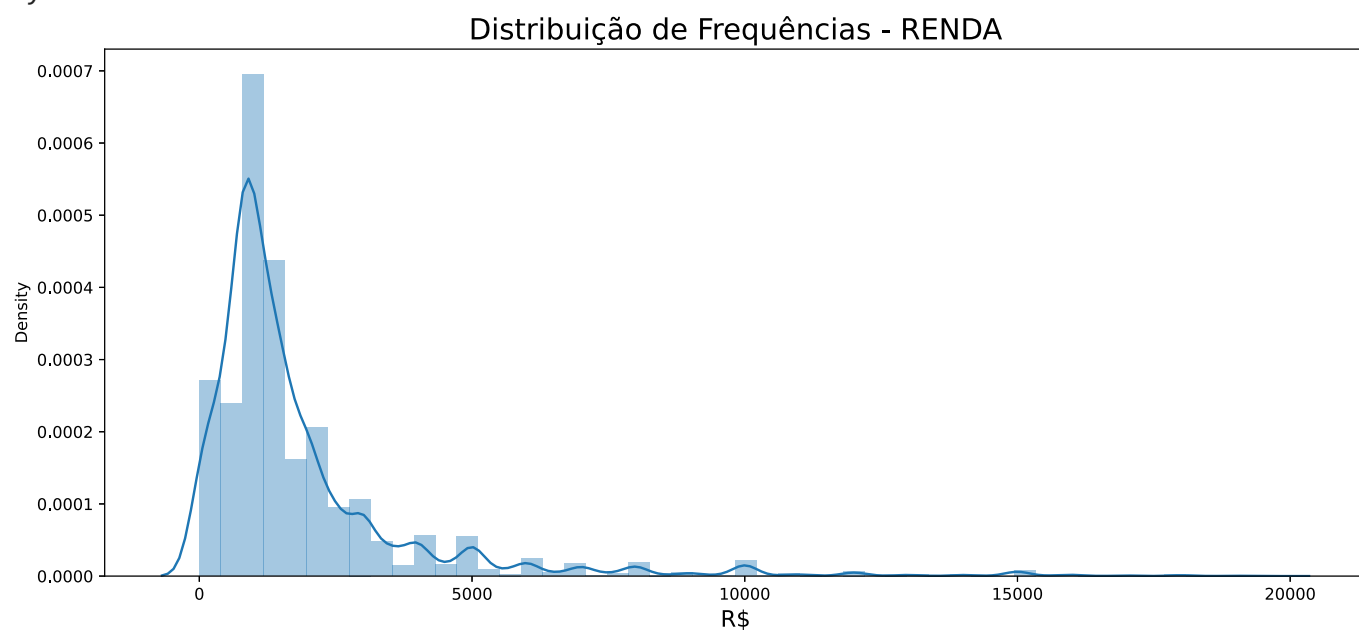
O gráfico apresenta uma forte assimetria à direita, dificultando até mesmo a análise do gráfico, é possível perceber somente que a maior parte da população se concentra a esquerda do gráfico, tendo uma renda baixa.

Para a variável RENDA, construa um histograma somente com as informações das pessoas com rendimento até R\$ 20.000,00

In [13]:

```
ax = sns.distplot(dados.query('Renda < 20000')['Renda'])
ax.figure.set_size_inches(14, 6)
ax.set_title('Distribuição de Frequências - RENDA', fontsize=18)
ax.set_xlabel('R$', fontsize=14)
ax
```

Out[13]: <AxesSubplot:title={'center': 'Distribuição de Frequências - RENDA'}, xlabel='R\$', ylabel='Density'>



Construa uma tabela de frequências e uma com os percentuais do cruzando das variáveis SEXO e COR

Avalie o resultado da tabela e escreva suas principais conclusões

Utilize os dicionários abaixo para renomear as linha e colunas das tabelas de frequências e dos gráficos em nosso projeto

In [14]:

```
sexo = {
    0: 'Masculino',
    1: 'Feminino'
}
cor = {
    0: 'Indígena',
    2: 'Branca',
    4: 'Preta',
    6: 'Amarela',
    8: 'Parda',
    9: 'Sem declaração'
}
anos_de_estudo = {
    1: 'Sem instrução e menos de 1 ano',
    2: '1 ano',
    3: '2 anos',
    4: '3 anos',
    5: '4 anos',
    6: '5 anos',
}
```



```

7: '6 anos',
8: '7 anos',
9: '8 anos',
10: '9 anos',
11: '10 anos',
12: '11 anos',
13: '12 anos',
14: '13 anos',
15: '14 anos',
16: '15 anos ou mais',
17: 'Não determinados'
}
uf = {
11: 'Rondônia',
12: 'Acre',
13: 'Amazonas',
14: 'Roraima',
15: 'Pará',
16: 'Amapá',
17: 'Tocantins',
21: 'Maranhão',
22: 'Piauí',
23: 'Ceará',
24: 'Rio Grande do Norte',
25: 'Paraíba',
26: 'Pernambuco',
27: 'Alagoas',
28: 'Sergipe',
29: 'Bahia',
31: 'Minas Gerais',
32: 'Espírito Santo',
33: 'Rio de Janeiro',
35: 'São Paulo',
41: 'Paraná',
42: 'Santa Catarina',
43: 'Rio Grande do Sul',
50: 'Mato Grosso do Sul',
51: 'Mato Grosso',
52: 'Goiás',
53: 'Distrito Federal'
}

```

```
In [15]: frequencia = pd.crosstab(dados.Sexo,
                                dados.Cor)
frequencia.rename(index = sexo, inplace = True)
frequencia.rename(columns = cor, inplace = True)
frequencia
```

```
Out[15]: Cor  Indígena  Branca  Preta  Amarela  Parda
```

Sexo					
Masculino	256	22194	5502	235	25063
Feminino	101	9621	2889	117	10862

```
In [16]: percentual = pd.crosstab(dados.Sexo,
                                dados.Cor,
                                normalize = True) * 100
percentual.rename(index = sexo, inplace = True)
percentual.rename(columns = cor, inplace = True)
percentual
```

Out[16]:	Cor	Indígena	Branca	Preta	Amarela	Parda
	Sexo					

	Cor	Indígena	Branca	Preta	Amarela	Parda
Sexo						
Masculino	0.333160	28.883394	7.160333	0.305830	32.617126	
Feminino	0.131442	12.520822	3.759761	0.152264	14.135867	

Conclusões

As frequências mostram que a pesquisa foi respondida majoritariamente por homens. As pessoas que responderam se declararam na maior parte Pardas ou Brancas, essas duas cores representando mais de 80% dos entrevistado.

Realize, para a variável RENDA, uma análise descritiva com as ferramentas que aprendemos em nosso treinamento

Obtenha a média aritmética

```
In [17]: dados.Renda.mean()
```

```
Out[17]: 2000.3831988547631
```

Obtenha a mediana

```
In [18]: dados.Renda.median()
```

```
Out[18]: 1200.0
```

Obtenha a moda

```
In [19]: dados.Renda.mode()
```

```
Out[19]: 0      788
dtype: int64
```

Obtenha o desvio médio absoluto

```
In [20]: dados.Renda.mad()
```

```
Out[20]: 1526.4951371644931
```

Obtenha a variância

```
In [21]: dados.Renda.var()
```

```
Out[21]: 11044906.00622118
```

Obtenha o desvio-padrão

```
In [22]: dados.Renda.std()
```

```
Out[22]: 3323.3877303470294
```

Obtenha a média, mediana e valor máximo da variável RENDA segundo SEXO e COR

Destaque os pontos mais importante que você observa nas tabulações

O parâmento *aggfunc* da função *crosstab()* pode receber uma lista de funções. Exemplo: *aggfunc = {'mean', 'median', 'max'}*

In [23]:

```
renda_por_sexo_e_cor = pd.crosstab(dados.Cor,
                                   dados.Sexo,
                                   values = dados.Renda,
                                   aggfunc = {'mean', 'median', 'max'})
renda_por_sexo_e_cor.rename(index = cor, inplace = True)
renda_por_sexo_e_cor.rename(columns = sexo, inplace = True)
renda_por_sexo_e_cor
```

Out[23]:

	max		mean		median	
Sexo	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino
Cor						
Indígena	10000.0	120000.0	1081.710938	2464.386139	797.5	788.0
Branca	200000.0	100000.0	2925.744435	2109.866750	1700.0	1200.0
Preta	50000.0	23000.0	1603.861687	1134.596400	1200.0	800.0
Amarela	50000.0	20000.0	4758.251064	3027.341880	2800.0	1500.0
Parda	100000.0	30000.0	1659.577425	1176.758516	1200.0	800.0

Conclusões

Em todos os casos a mediada é a medida mais baixa, isso mostra que a renda da maioria da população se concentra nos valores mais baixos, sendo as médias distorcidas pelos extremos.

Considerando as médias, Homens ganham mais do que as Mulheres, tendo uma média maior em quase todos os casos, com exceção da cor Indígena.

Branco e Amarelo também ganham em média mais do que as outras cores.

Mesmo os Amarelos ganhando mais em média, os Brancos chegam á valores mais extremos, com Homens Brancos tendo a maior renda da amostra, seguidos por Mulheres Indígenas, e Mulheres Brancas e Homens Pardos em terceiro lugar.

Obtenha as medidas de dispersão da variável RENDA segundo SEXO e COR

Destaque os pontos mais importante que você observa nas tabulações

O parâmento *aggfunc* da função *crosstab()* pode receber uma lista de funções. Exemplo: *aggfunc = {'mad', 'var', 'std'}*

In [24]:

```
dispersao_renda_por_sexo_e_cor = pd.crosstab(dados.Cor,
                                              dados.Sexo,
                                              values = dados.Renda,
                                              aggfunc = {'mad', 'var', 'std'})
dispersao_renda_por_sexo_e_cor.rename(index = cor, inplace = True)
dispersao_renda_por_sexo_e_cor.rename(columns = sexo, inplace = True)
dispersao_renda_por_sexo_e_cor
```

Out[24]:

	mad		std		var	
Sexo	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino
Cor						
Indígena	798.910889	3007.892952	1204.093490	11957.498292	1.449841e+06	1.429818e+08
Branca	2261.012346	1670.967106	4750.791872	3251.013154	2.257002e+07	1.056909e+07
Preta	975.602482	705.453357	1936.309271	1349.799809	3.749294e+06	1.821960e+06
Amarela	3709.597211	2549.146322	5740.824820	3731.173660	3.295707e+07	1.392166e+07
Parda	1125.827704	811.580946	2312.087184	1596.233048	5.345747e+06	2.547960e+06

Conclusões

A tabela mostra que em alguns casos existe uma dispersão muito grande de salários entre Homens e Mulheres da mesma cor.

A cor Indígena por exemplo, como visto na tabela anterior, as Mulheres ganham em média mais do que o dobro dos homens. Mas as medidas de dispersão mostram que a renda das Mulheres Indígenas é bem dispersa, o que pode indicar que há valores extremos distorcendo a média.

Mas na maioria dos casos a renda dos Homens é mais dispersa, indicando valores mais próximos dos extremos, enquanto as Mulheres costumam apresentar uma renda mais próxima da média.

Construa um box plot da variável RENDA segundo SEXO e COR

É possível verificar algum comportamento diferenciado no rendimento entre os grupos de pessoas analisados? Avalie o gráfico e destaque os pontos mais importantes.

1º - Utilize somente as informações de pessoas com renda abaixo de R\$ 10.000

2º - Para incluir uma terceira variável na construção de um boxplot utilize o parâmetro *hue* e indique a variável que quer incluir na subdivisão.

Mais informações: <https://seaborn.pydata.org/generated/seaborn.boxplot.html>

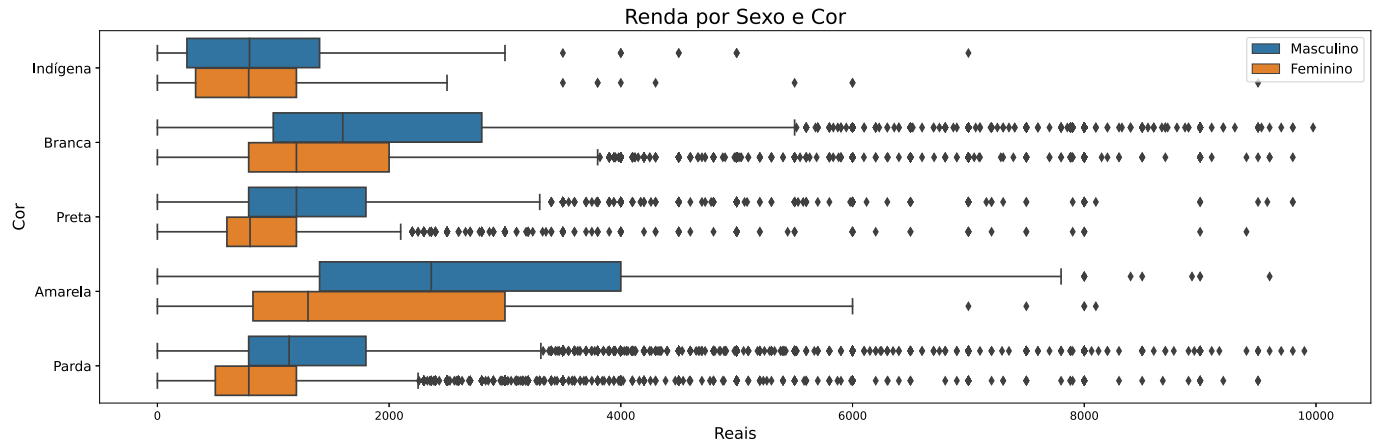
In [25]:

```
ax = sns.boxplot(x = 'Renda', y = 'Cor', hue = 'Sexo', data = dados.query('Renda < 10000'), ori
ax.figure.set_size_inches(20, 6)
ax.set_title("Renda por Sexo e Cor", fontsize = 18)
ax.set_xlabel("Reais", fontsize = 14)
ax.set_ylabel('Cor', fontsize=14)
ax.set_yticklabels(['Indígena', 'Branca', 'Preta', 'Amarela', 'Parda'], fontsize=12)

handles, _ = ax.get_legend_handles_labels()
ax.legend(handles, ['Masculino', 'Feminino'], fontsize=12)

ax
```

Out[25]: <AxesSubplot:title={'center':'Renda por Sexo e Cor'}, xlabel='Reais', ylabel='Cor'>



Conclusões

Considerando a faixa de renda utilizada, até R\$10.000,00, o gráfico confirma que Homens tendem a ter uma renda maior do que as Mulheres, sendo a Cor Indígena a única que apresenta um equilíbrio maior.

Na maior parte Homens tem uma concentração maior de Renda.

Considerando as Cores Preta, Amarela e Parda 25% dos Homens tem Renda maior do que 50% das Mulheres. E 50% dos Homens tem Renda maior do que praticamente 75% das Mulheres no caso das Cores Preta e Parda.

O gráfico mostra também as Cores Branca e Amarela tendo a maior concentração de Renda, ainda mais considerando os valores extremos.

DESAFIO

Qual percentual de pessoas de nosso *dataset* ganham um salário mínimo (R\$ 788,00) ou menos?

Utilize a função `percentileofscore()` do *scipy* para realizar estas análises.

Mais informações:

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.percentileofscore.html>

In [26]:

```
from scipy import stats

percentual = stats.percentileofscore(dados.Renda, 788, kind='weak')
print(f"{percentual:.2f}%")
```

28.87%

Qual o valor máximo ganho por 99% das pessoas de nosso *dataset*?

Utilize o método `quantile()` do *pandas* para realizar estas análises.

In [27]:

```
maximo_99 = dados.Renda.quantile(0.99)
print(f"R$ {maximo_99:.2f}")
```

R\$ 15000.00

Obtenha a média, mediana, valor máximo e desvio-padrão da variável RENDA segundo ANOS DE ESTUDO e SEXO

Destaque os pontos mais importante que você observa nas tabulações

O parâmento *aggfunc* da função *crosstab()* pode receber uma lista de funções. Exemplo:
aggfunc = ['mean', 'median', 'max', 'std']

```
In [28]: dispersao_renda_por_anos_de_estudo_e_sexo = pd.crosstab(dados['Anos de Estudo'],
                                                             dados['Sexo'],
                                                             values = dados.Renda,
                                                             aggfunc = {'mean', 'median', 'max', 'std'})
dispersao_renda_por_anos_de_estudo_e_sexo.rename(index = anos_de_estudo, inplace = True)
dispersao_renda_por_anos_de_estudo_e_sexo.rename(columns = sexo, inplace = True)
dispersao_renda_por_anos_de_estudo_e_sexo
```

Out[28]:

		max		mean		median		std	
Sexo		Masculino	Feminino	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino
Anos de Estudo									
Sem instrução e menos de 1 ano		30000.0	10000.0	799.49	516.20	700.0	390.0	1023.90	639.31
1 ano		30000.0	2000.0	895.63	492.77	788.0	400.0	1331.95	425.29
2 anos		40000.0	4000.0	931.18	529.91	788.0	450.0	1435.17	498.23
3 anos		80000.0	3500.0	1109.20	546.85	800.0	500.0	2143.80	424.12
4 anos		50000.0	10000.0	1302.33	704.28	1000.0	788.0	1419.82	629.55
5 anos		35000.0	8000.0	1338.65	781.39	1045.0	788.0	1484.65	635.78
6 anos		25000.0	6000.0	1448.88	833.73	1200.0	788.0	1476.63	574.55
7 anos		40000.0	9000.0	1465.50	830.75	1200.0	788.0	1419.71	602.04
8 anos		30000.0	18000.0	1639.40	933.62	1300.0	800.0	1515.58	896.78
9 anos		60000.0	20000.0	1508.04	868.02	1200.0	788.0	2137.66	973.22
10 anos		45000.0	6000.0	1731.27	925.92	1218.0	800.0	2078.61	620.61
11 anos		200000.0	100000.0	2117.06	1286.79	1500.0	1000.0	2676.54	1819.04
12 anos		30000.0	120000.0	2470.33	1682.31	1800.0	1200.0	2268.08	4851.83
13 anos		25000.0	20000.0	3195.10	1911.73	2400.0	1300.0	2797.12	2053.79
14 anos		50000.0	20000.0	3706.62	2226.46	2500.0	1600.0	3987.21	2064.08
15 anos ou mais		200000.0	100000.0	6134.28	3899.51	4000.0	2800.0	7447.61	4212.77
Não determinados		7000.0	3000.0	1295.76	798.17	1200.0	788.0	979.65	459.99

A tabela mostra claramente a diferença entre Renda de Homens e Mulheres com o mesmo tempo de estudo.

As médias mostram que Homens tem Renda maior do que Mulheres, essa diferença chega a ser quase o dobro em alguns casos.

As máximas mostram valores bem mais altos para Homens, o que pode indicar que alguns valores extremos podem estar distorcendo a média.

As medianas e os desvios padrões mostram metade da população de cada faixa de estudo ganhando bem a baixo das médias, e com um alto desvio padrão, sendo mais um indício de que as médias estão sendo distorcidas por valores extremos.

Construa um box plot da variável RENDA segundo ANOS DE ESTUDO e SEXO

É possível verificar algum comportamento diferenciado no rendimento entre os grupos de pessoas analisados? Avalie o gráfico e destaque os pontos mais importantes.

1º - Utilize somente as informações de pessoas com renda abaixo de R\$ 10.000

2º - Utilize a variável IDADE para identificar se a desigualdade se verifica para pessoas de mesma idade. Exemplo: `data=dados.query('Renda < 10000 and Idade == 40')` ou `data=dados.query('Renda < 10000 and Idade == 50')`

3º - Para incluir uma terceira variável na construção de um boxplot utilize o parâmetro `hue` e indique a variável que quer incluir na subdivisão.

Mais informações: <https://seaborn.pydata.org/generated/seaborn.boxplot.html>

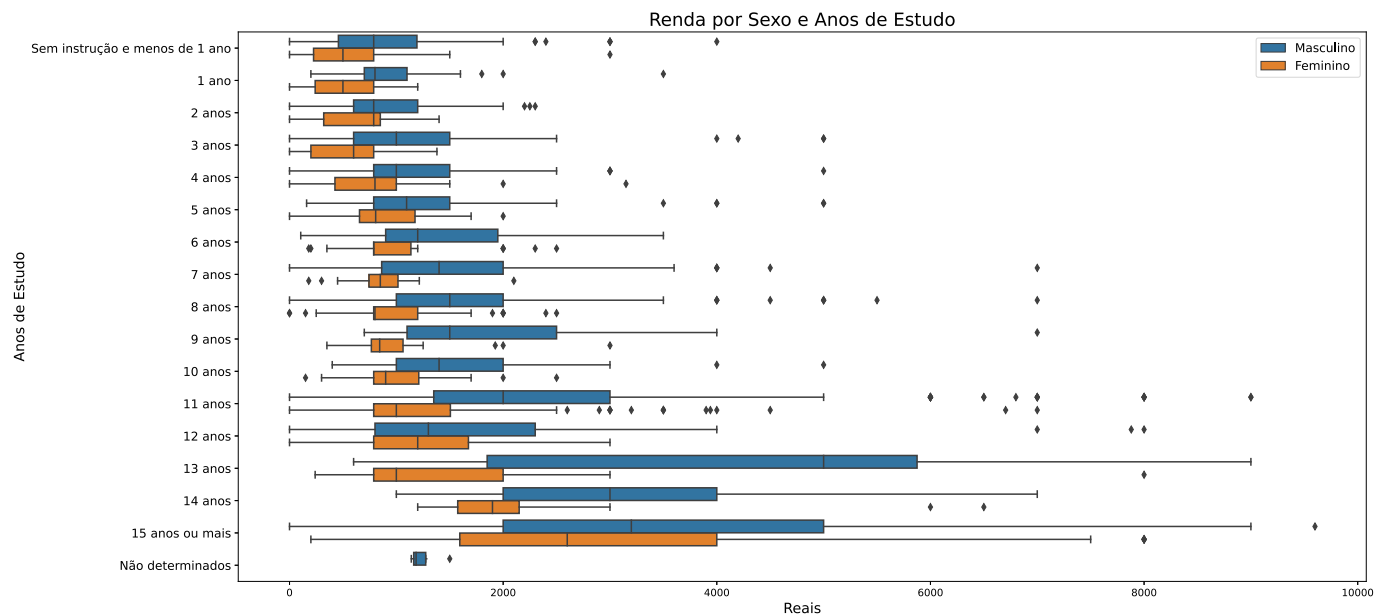
In [29]:

```
ax = sns.boxplot(x = 'Renda', y = 'Anos de Estudo', hue = 'Sexo', data = dados.query('Renda < 10000'))
ax.figure.set_size_inches(20, 10)
ax.set_title("Renda por Sexo e Anos de Estudo", fontsize = 18)
ax.set_xlabel("Reais", fontsize = 14)
ax.set_ylabel('Anos de Estudo', fontsize=14)
ax.set_yticklabels([key for key in anos_de_estudo.values()], fontsize=12)

handles, _ = ax.get_legend_handles_labels()
ax.legend(handles, ['Masculino', 'Feminino'], fontsize=12)

ax
```

Out[29]: <AxesSubplot:title={'center':'Renda por Sexo e Anos de Estudo'}, xlabel='Reais', ylabel='Anos de Estudo'>



Conclusões

Mesmo com os filtros de Renda e Idade utilizados o gráfico confirma as impressões tidas na tabela.

Homens tem Renda superior a das Mulheres com o mesmo tempo de estudo.

Conforme os Anos de Estudo aumentam a Renda também aumenta, e aumenta também a diferença de Renda entre os dois Sexos.

Obtenha a média, mediana, valor máximo e desvio-padrão da variável RENDA segundo as UNIDADES DA FEDERAÇÃO

Destaque os pontos mais importante que você observa nas tabulações

Utilize o método `groupby()` do `pandas` juntamente com o método `agg()` para contruir a tabulação. O método `agg()` pode receber um dicionário especificando qual coluna do

DataFrame deve ser utilizada e qual lista de funções estatísticas queremos obter, por exemplo: `dados.groupby(['UF']).agg({'Renda': ['mean', 'median', 'max', 'std']})`

```
In [34]: dispersao_renda_por_uf = dados.groupby('UF').agg({'Renda': ['mean', 'median', 'max', 'std']}).r
dispersao_renda_por_uf.rename(index = uf, inplace = True)
dispersao_renda_por_uf
```

Out[34]:

UF	Renda			
	mean	median	max	std
Rondônia	1789.76	1200	50000	2406.16
Acre	1506.09	900	30000	2276.23
Amazonas	1445.13	900	22000	1757.94
Roraima	1783.59	1000	20000	2079.66
Pará	1399.08	850	50000	2053.78
Amapá	1861.35	1200	15580	2020.69
Tocantins	1771.09	1000	60000	2934.59
Maranhão	1019.43	700	30000	1887.82
Piauí	1074.55	750	40000	2373.36
Ceará	1255.40	789	25000	1821.96
Rio Grande do Norte	1344.72	800	15500	1651.81
Paraíba	1293.37	788	30000	1950.27
Pernambuco	1527.08	900	50000	2389.62
Alagoas	1144.55	788	11000	1237.86
Sergipe	1109.11	788	16000	1479.00
Bahia	1429.65	800	200000	3507.92
Minas Gerais	2056.43	1200	100000	3584.72
Espírito Santo	2026.38	1274	100000	3513.85
Rio de Janeiro	2496.40	1400	200000	5214.58
São Paulo	2638.10	1600	80000	3503.78
Paraná	2493.87	1500	200000	4302.94
Santa Catarina	2470.85	1800	80000	3137.65
Rio Grande do Sul	2315.16	1500	35000	2913.34
Mato Grosso do Sul	2262.60	1500	42000	3031.42
Mato Grosso	2130.65	1500	35000	2542.63
Goiás	1994.58	1500	30000	2221.93
Distrito Federal	4241.95	2000	100000	5550.46

A tabela mostra a diferença de Renda entre os estados.

Em todos os estados a mediana mostra 50% da população abaixo da Média de Renda. A Renda máxima em cada estado é dez vezes maior do que a média. Indicando a média distorcida por valores extremos. O Desvio Padrão também sugere esta distorção.

Construa um box plot da variável RENDA segundo as UNIDADES DA FEDERAÇÃO

É possível verificar algum comportamento diferenciado no rendimento entre os grupos analisados? Avalie o gráfico e destaque os pontos mais importantes.

1º - Utilize somente as informações de pessoas com renda abaixo de R\$ 10.000

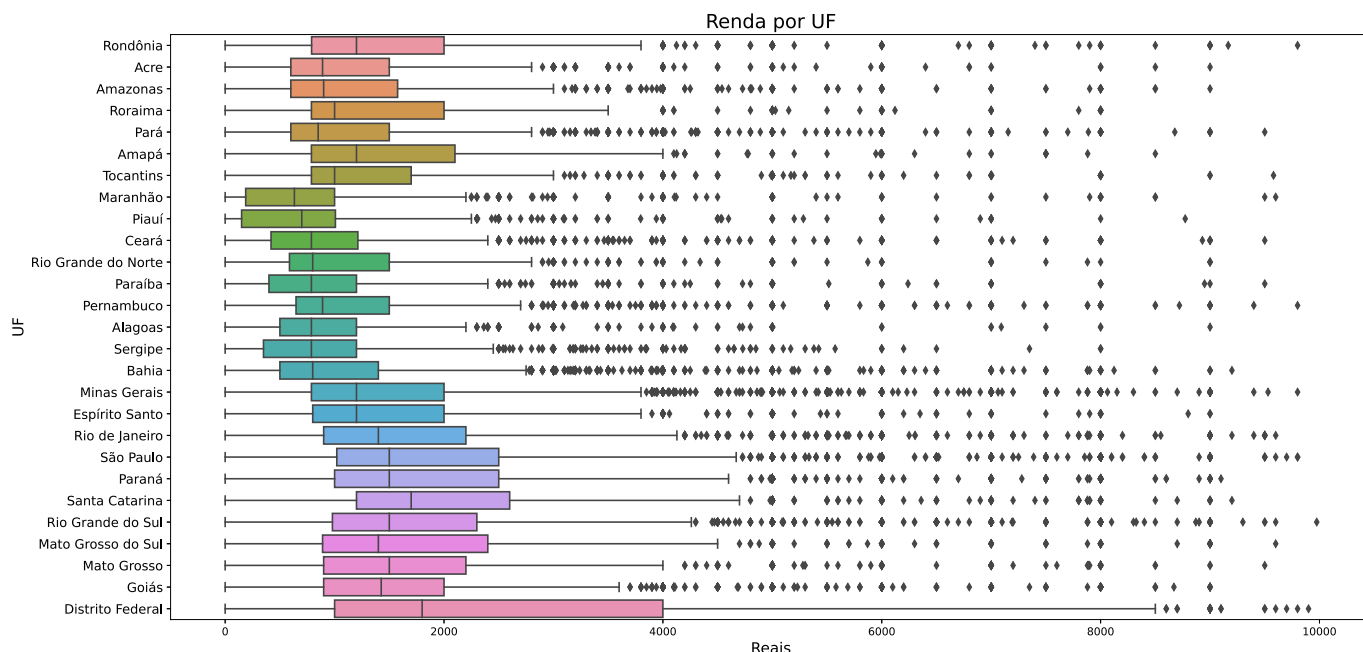
In []:

In [39]:

```
ax = sns.boxplot(x = 'Renda', y = 'UF', data = dados.query('Renda < 10000'), orient = 'h')
ax.figure.set_size_inches(20, 10)
ax.set_title("Renda por UF", fontsize = 18)
ax.set_xlabel("Reais", fontsize = 14)
ax.set_ylabel("UF", fontsize = 14)
ax.set_yticklabels([key for key in uf.values()], fontsize = 12)

ax
```

Out[39]: <AxesSubplot:title={'center': 'Renda por UF'}, xlabel='Reais', ylabel='UF'>



Conclusões

O gráfico mostra alguns estados tendo uma Renda bem superior aos outros.

A maior parte da população se concentra em uma faixa de renda que não varia muito. Mas todos apresentam valores muito altos distorcendo a Renda.

O Distrito Federal se destaca no gráfico, com a maior parte da população se concentrando nos valores mais altos, e com os maiores extremos.