



Universidade Federal de Pernambuco – UFPE
Centro de Ciências Exatas e da Natureza – CCEN
Departamento de Estatística da UFPE

Análise Exploratória de Dados

Análise de massa de dados da tabela 1.1 do livro Noções de Probabilidade e Estatística dos autores Magalhães e Lima.

Disciplina: Análise Exploratória de Dados

Docente: Audrey Hellen Mariz de Aquino Cysneiros

Discentes: BRENDA BARROS ALVES DA SILVA

CLARISSE MILENA VICENTE MAGNATA

EVONIO DE BARROS CAMPELO JUNIOR

MARCELO JOSE SOARES SILVA FILHO

VITOR NEGROMONTE CABRAL DE OLIVEIRA

1. Apresentação da base de dados

Figura 1: visualização da tabela com as variáveis da base de dados a ser trabalhada, em python.

A variável (dataframe) criada vai trazer o banco com a forma em quantitativo estabelecido de linhas visualizadas (10).

```
dataframe.head(10)
```

	Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Tolernacia	Exer	Cine	OpCine	TV	OpTv
0	1	A	F	17	1.60	60.5	2	Nao	P	0	1	B	16	R
1	2	A	F	18	1.69	55.0	1	Nao	M	0	1	B	7	R
2	3	A	M	18	1.85	72.8	2	Nao	P	5	2	M	15	R
3	4	A	M	25	1.85	80.9	2	Nao	P	5	2	B	20	R
4	5	A	F	19	1.58	55.0	1	Nao	M	2	2	B	5	R
5	6	A	M	19	1.76	60.0	3	Nao	M	2	1	B	2	R
6	7	A	F	20	1.60	58.0	1	Nao	P	3	1	B	7	R
7	8	A	F	18	1.64	47.0	1	Sim	I	2	2	M	10	R
8	9	A	F	18	1.62	57.8	3	Nao	M	3	3	M	12	R
9	10	A	F	17	1.64	58.0	1	Nao	M	2	2	M	10	R

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Descrição das variáveis:

- Id: identificação do aluno
- Turma: turma a que o aluno foi alocado (A ou B)
- Sexo: F se feminino, M se masculino.
- Idade: idade em anos
- Alt: altura em metros
- Peso: peso em quilogramas
- Filhos: número de filhos na família
- Fuma: hábito de fumar, sim ou não
- Toler: tolerância ao cigarro:

(I) indiferente, (P) incomoda pouco e (M) incomoda muito

- Exerc: horas de atividade física, por semanal
- Cine: número de vezes que vai ao cinema por semana.
- OpCine: opinião a respeito das salas de cinema na cidade:

(B) regular a boa e (M) muito boa.

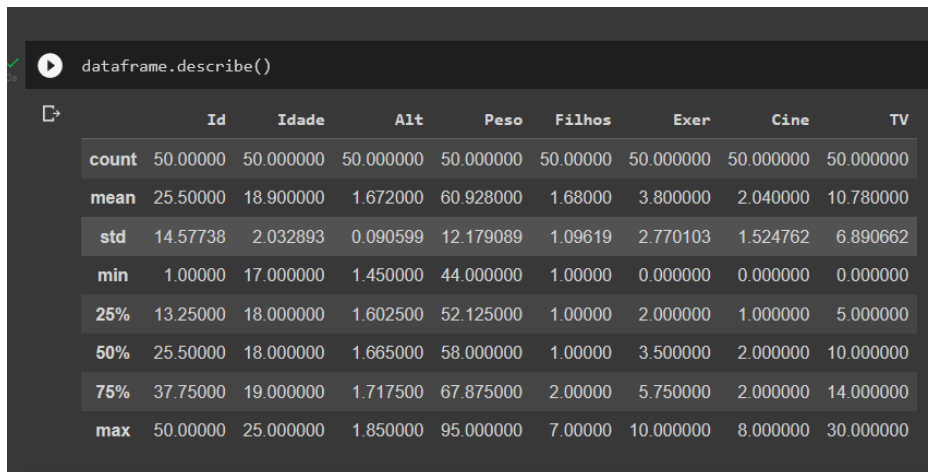
- TV: horas gastas assistindo TV, por semanal
- OpTV: opinião a respeito da qualidade da programação na TV:

(R) ruim, (M) média, (B) boa e (N) não sabe

2. Análise da Frequência Absoluta e Frequência Relativa em Percentual das variáveis quantitativas (Idade, Altura, Peso, Filhos, Exercício, Cinema, TV).

- Visão Geral:

Figura 2: descrição geral de toda as variáveis quantitativas, em python.



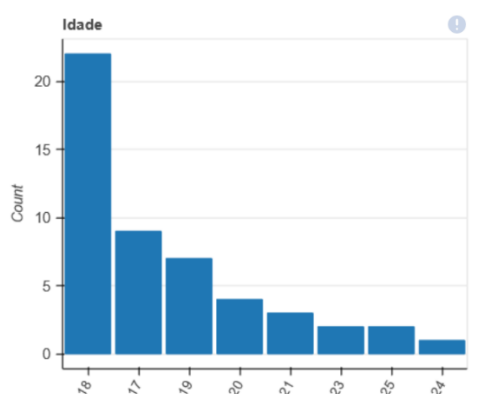
	Id	Idade	Alt	Peso	Filhos	Exer	Cine	TV
count	50.00000	50.000000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000
mean	25.50000	18.900000	1.672000	60.928000	1.68000	3.800000	2.040000	10.780000
std	14.57738	2.032893	0.090599	12.179089	1.09619	2.770103	1.524762	6.890662
min	1.00000	17.000000	1.450000	44.000000	1.00000	0.000000	0.000000	0.000000
25%	13.25000	18.000000	1.602500	52.125000	1.00000	2.000000	1.000000	5.000000
50%	25.50000	18.000000	1.665000	58.000000	1.00000	3.500000	2.000000	10.000000
75%	37.75000	19.000000	1.717500	67.875000	2.00000	5.750000	2.000000	14.000000
max	50.00000	25.000000	1.850000	95.000000	7.00000	10.000000	8.000000	30.000000

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

➤ Variável Idade

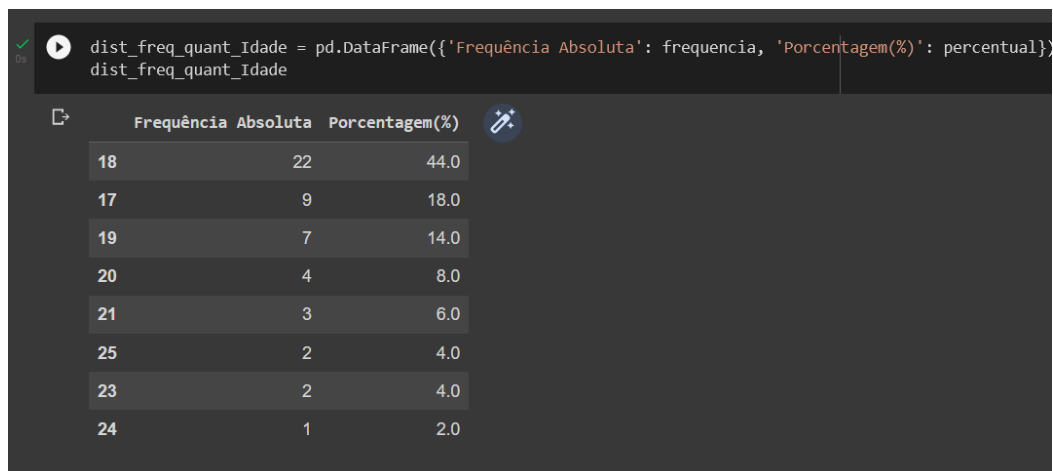
Valor Max = 25 anos, Valor Min = 17 anos, Média = 18,9 anos

Gráfico 1: gráfico da variável idade(x) x total(y), em excel.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 3: tabela de frequência da variável idade, em python.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Tabela 1: tabela de frequência da variável idade, em excell.

Tabela de frequência da variável idade			
Classes	Frequência Ab.	Frequência R.	Porcentagem
[17-20)	38	0,76	76,00%
[20-23)	7	0,14	14,00%
[23-25]	5	0,10	10,00%
Total	50	1	100,00%

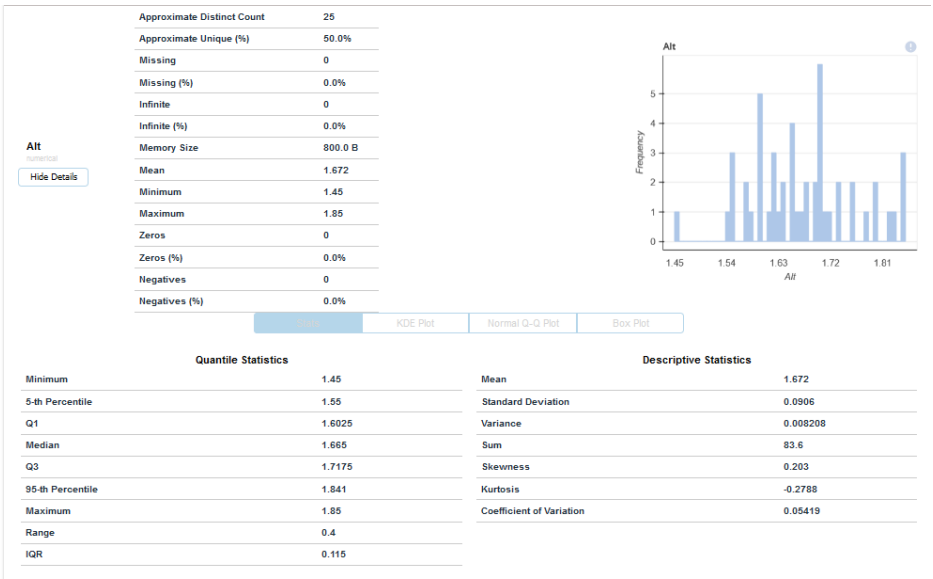
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Foi necessário fazer uma segunda tabela pois na tabela gerada através do código em Python não havia a divisão de classes.

A Analisando a tabela de frequência da variável idade, percebe-se que as maiores e menores observações são: 25 e 17 anos, respectivamente. Ademais, a maioria dos valores estão contidos na classe [17 - 20), totalizando cerca de 76% das observações. Em seguida, a classe [20 - 23) com 14% e, conseqüentemente, a menor classe é a [23 - 25] com 10%. A média e a mediana da variável, são, respectivamente, 18.9 e 18.

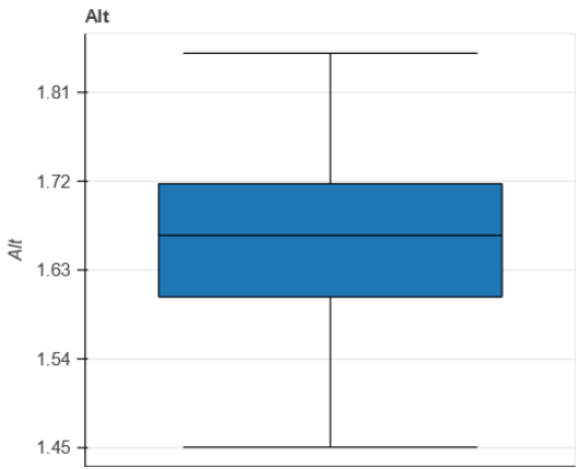
➤ Variável Altura

Figura 4: descrição estatística geral da variável altura, com quartis, max e min e seu devido gráfico.



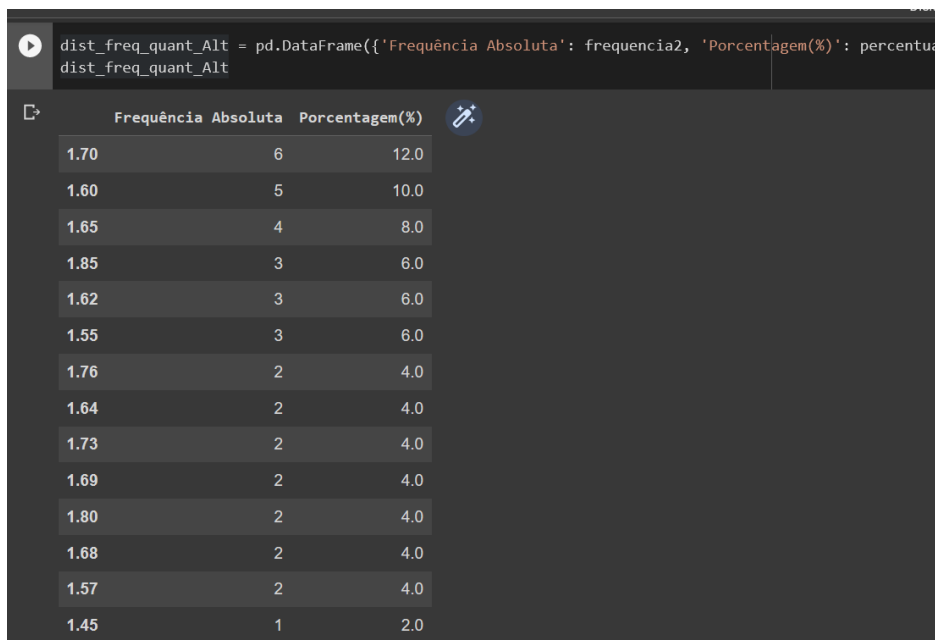
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 2: gráfico de boxplot da variável altura, em python.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 5: tabela de frequência da variável altura, em python.



The screenshot shows a Jupyter Notebook interface. At the top, a code cell contains the following Python code:

```
dist_freq_quant_Alt = pd.DataFrame({'Frequência Absoluta': frequencia2, 'Porcentagem(%)': percentua
dist_freq_quant_Alt
```

Below the code cell, the output of the DataFrame is displayed as a table:

	Frequência Absoluta	Porcentagem(%)
1.70	6	12.0
1.60	5	10.0
1.65	4	8.0
1.85	3	6.0
1.62	3	6.0
1.55	3	6.0
1.76	2	4.0
1.64	2	4.0
1.73	2	4.0
1.69	2	4.0
1.80	2	4.0
1.68	2	4.0
1.57	2	4.0
1.45	1	2.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Tabela 2: tabela de frequência da variável altura, em excel.

Tabela de frequência da variável Altura			
Classes(metros)	Frequência Ab.	Frequência R.	Porcentagem
[1,45 - 1,51)	1	0,02	2%
[1,51 - 1,57)	4	0,08	8%
[1,57 - 1,63)	12	0,24	24%
[1,63 - 1,69)	11	0,22	22%
[1,69 - 1,75)	12	0,24	24%
[1,75 - 1,81)	5	0,1	10%
[1,81 - 1,87)	5	0,1	10%
Total	50	1	100%

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

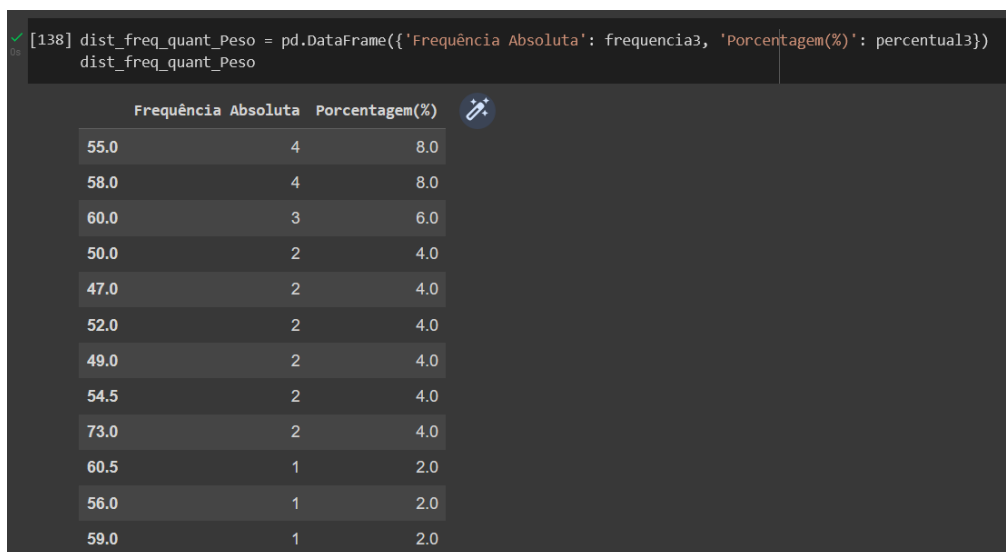
Observando a variável altura, nota-se que os valores máximo e mínimo são 1.45 e 1.85. A menor concentração é observada na classe [1.45 - 1.51) com apenas 2% das observações e a

maior na classe [1.57 - 1.75) com 70%. A média e a mediana são, 1.67 e 1.66, respectivamente. Dito isso, é perceptível uma assimetria positiva no histograma da variável, no entanto, esta assimetria é pequena devido a curta diferença entre a média e a mediana. Também é relevante observar a moda, encontrada como 1,70 e acima da média e mediana.

Analizando o box-plot, é possível observar a linha da mediana como 1.66 e ainda evidenciar uma assimetria negativa, devido à sua proximidade do Q3 da caixa. O Primeiro Quartil é encontrado como 1.60 e o terceiro Quartil é igual a 1.72, desta forma podemos definir a distância interquartil como 0.12 e concluir o limite inferior (1.42) e superior (1.9). Ambos os limites não foram atingidos e também não foram encontrados outliers.

➤ Variável Peso

Figura 6: tabela de frequência da variável peso, em python.



```
[138] dist_freq_quant_Peso = pd.DataFrame({'Frequência Absoluta': frequencia3, 'Porcentagem(%)': percentual3})
dist_freq_quant_Peso
```

	Frequência Absoluta	Porcentagem(%)
55.0	4	8.0
58.0	4	8.0
60.0	3	6.0
50.0	2	4.0
47.0	2	4.0
52.0	2	4.0
49.0	2	4.0
54.5	2	4.0
73.0	2	4.0
60.5	1	2.0
56.0	1	2.0
59.0	1	2.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002.

Tabela 3: tabela de frequência da variável peso, em excel.

Tabela de Frequência da Variável Peso			
Classes (kg)	Frequência Ab.	Frequência R.	Porcentagem
[44 - 52)	11	0,22	22%
[52 - 60)	19	0,38	38%
[60 - 68)	7	0,14	14%
[68 - 76)	7	0,14	14%
[76 - 84)	1	0,02	2%
[84 - 92)	4	0,08	8%
[92 - 100)	1	0,02	2%
Total	50	1	100%

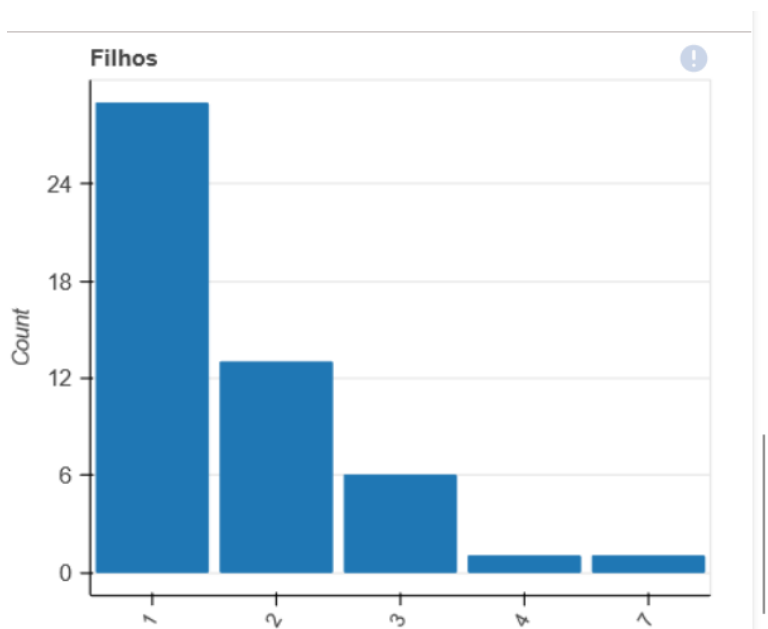
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando a tabela de frequência da variável peso, os valores máximo e mínimo são 100 e 44, respectivamente. Ademais, a maior concentração se encontra na classe [52 - 60) com 38%, seguida pela classe [44 - 52) que tem 22% das observações. A classes com menor concentração de indivíduos são as [76 - 84) e [92 - 100) com apenas 2%, cada. A média e a mediana são 60.9 e 67.8.

➤ Variável Filhos

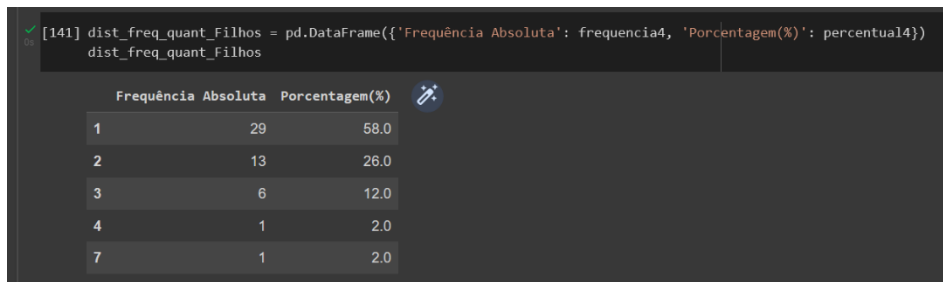
Valor Max = 7, Valor Min = 1, Média = 1,68

Gráfico 3: gráfico da variável filhos(x) x total (y), em python.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 7: tabela de frequência da variável filhos, em python.

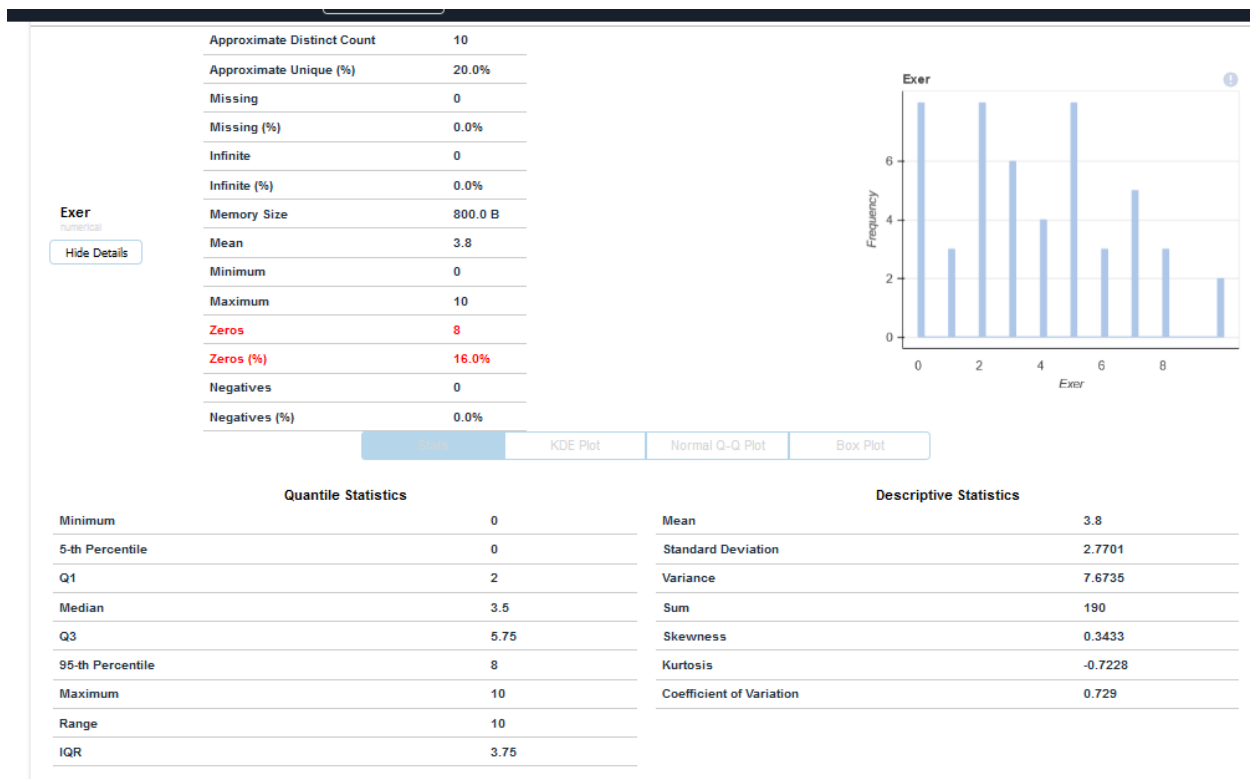


Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Na variável filhos, a maior concentração se dá no número de indivíduos com 1 filho na família, que totaliza 58% das observações e, em seguida, o número de indivíduos com 2 ou mais filhos, totalizando 42%. Além disso, a média e a mediana da variável, são, 1.68 e 1, respectivamente.

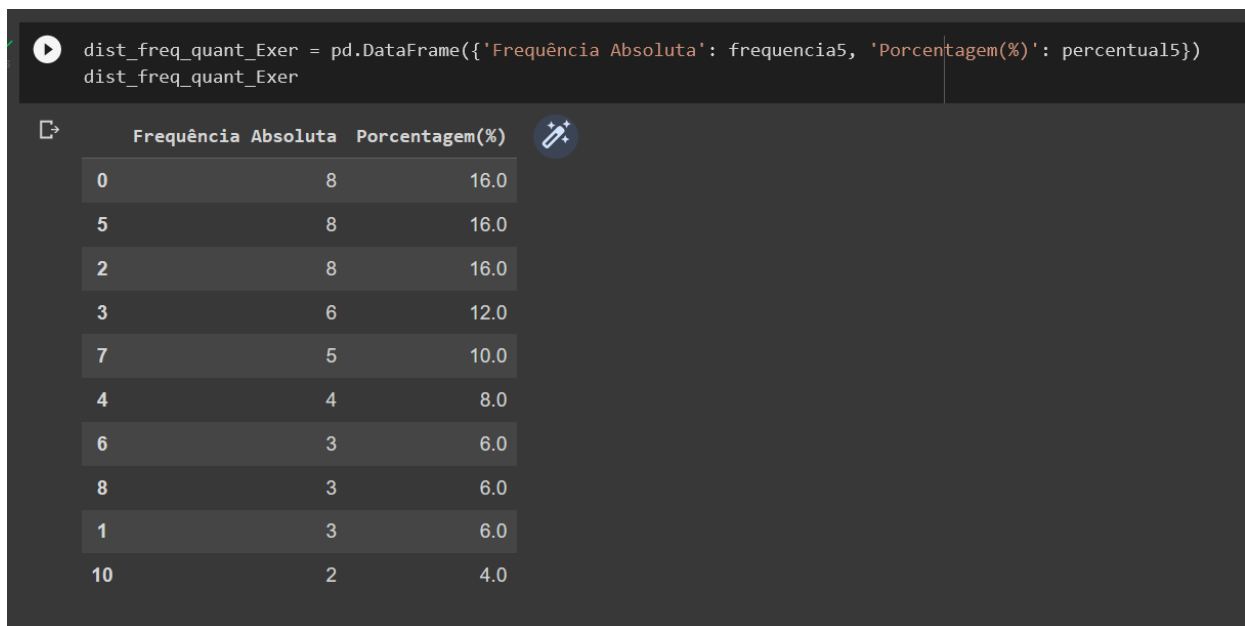
➤ Variável Exercício

Figura 8: descrição estatística geral da variável exercício, com quartis, max e min e seu devido gráfico.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 9: tabela de frequência da variável exercício, em python.



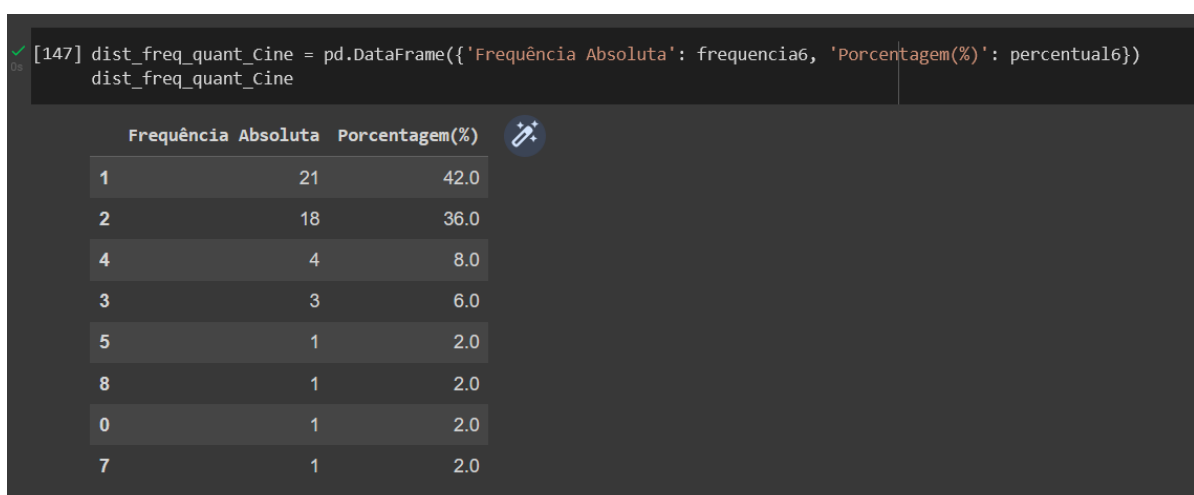
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Na variável exercícios, nota-se grande assimetria nos valores. Percebe-se, então, que a maior concentração de indivíduos se dá na classe [0-4), com 50% das observações e em seguida vem a classe [4-8) com 40%, logo, a classe [8-12] apresenta 10%. Os valores máximos e mínimos são, 0 e 12 horas por semana.

➤ Variável Cinema

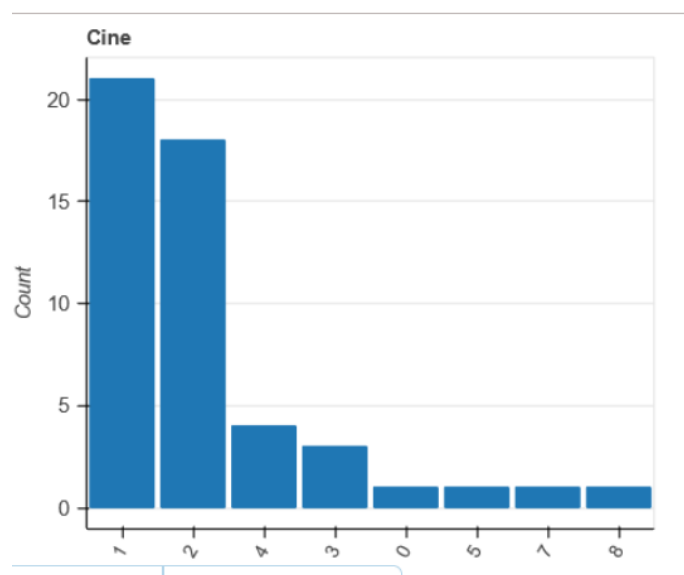
Valor Max = 8 , Valor Min=0 , Média = 2

Figura 10: tabela de frequência da variável Cinema, em python.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 4: gráfico da variável Cinema(x) x total (y), em python.

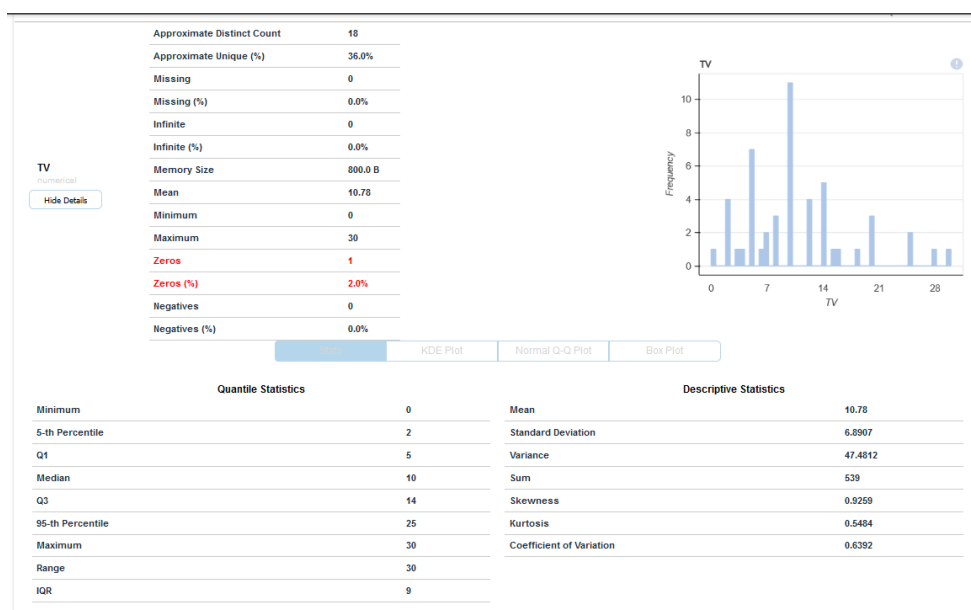


Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Ao analisarmos a tabela de frequência da variável Cine, notamos a grande concentração no número de pessoas que gastam entre 1 e 2 por semana no cinema, totalizando 78% dos entrevistados e 2% não vão ao cinema, portanto, 20% dos entrevistados gastam entre 4 e 8 horas semanais no cinema.

➤ Variável Tv

Figura 11: descrição estatística geral da variável Tv, com quartis, max e min e seu devido gráfico.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 12: tabela de frequência da variável Tv, em python.

```
[150] dist_freq_quant_TV = pd.DataFrame({'Frequência Absoluta': frequencia7, 'Porcentagem(%)': percentual7})
dist_freq_quant_TV
```

	Frequência Absoluta	Porcentagem(%)
10	11	22.0
5	7	14.0
14	5	10.0
2	4	8.0
12	4	8.0
20	3	6.0
8	3	6.0
25	2	4.0
7	2	4.0
16	1	2.0
28	1	2.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Tabela 4: tabela de frequência da variável TV, em excel.

Tabela de frequência de TV			
Classes	Frequência Abs	Frequência Rel	Porcentagem (%)
(0-5)	7	0,14	14%
(5-10)	13	0,26	26%
(10-15)	21	0,42	42%
(15-20)	2	0,04	4%
(20-25)	3	0,06	6%
(25-30)	4	0,08	8%
Total	50	1	100%

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando a tabela de frequência da variável TV, é notória a maior concentração de indivíduos que gastam entre 10 e 15 horas semanais assistindo televisão, totalizando 42%, enquanto a menor concentração está nos indivíduos que gastam entre 15 e 20 horas semanais, com apenas 4% dos entrevistados.

3. Análise da Frequência Absoluta e Frequência Relativa em Percentual das variáveis qualitativas (Turma, Sexo, Fuma, Tolerância, OpCine, Optv).

➤ Variável Turma

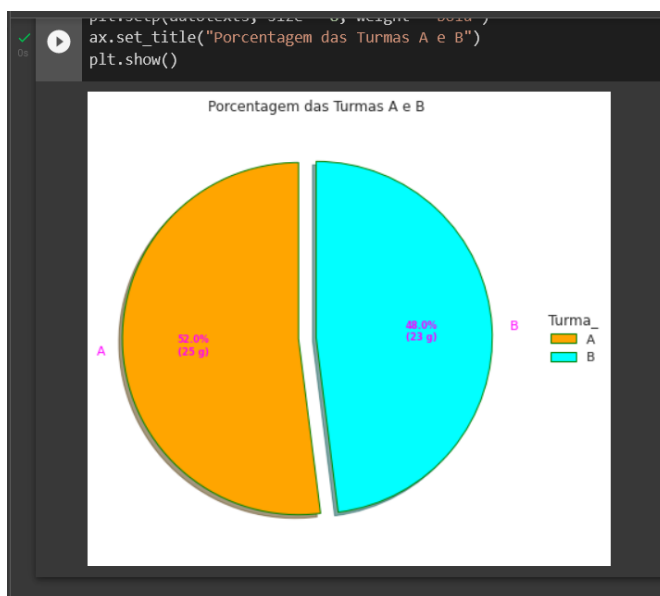
Figura 13: tabela de frequência da variável turma, em python.

```
[153] dist_freq_quali_Turma = pd.DataFrame({'Frequência Absoluta': frequencia8, 'Porcentagem(%)': percentual8})
dist_freq_quali_Turma
```

	Frequência Absoluta	Porcentagem(%)
A	26	52.0
B	24	48.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 5: gráfico de setores da variável Turma.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando a variável número de alunos em cada turma fica evidente uma maior concentração de alunos na turma A com 52% dos alunos, logo, a turma B engloba os demais 48%.

➤ Variável Sexo

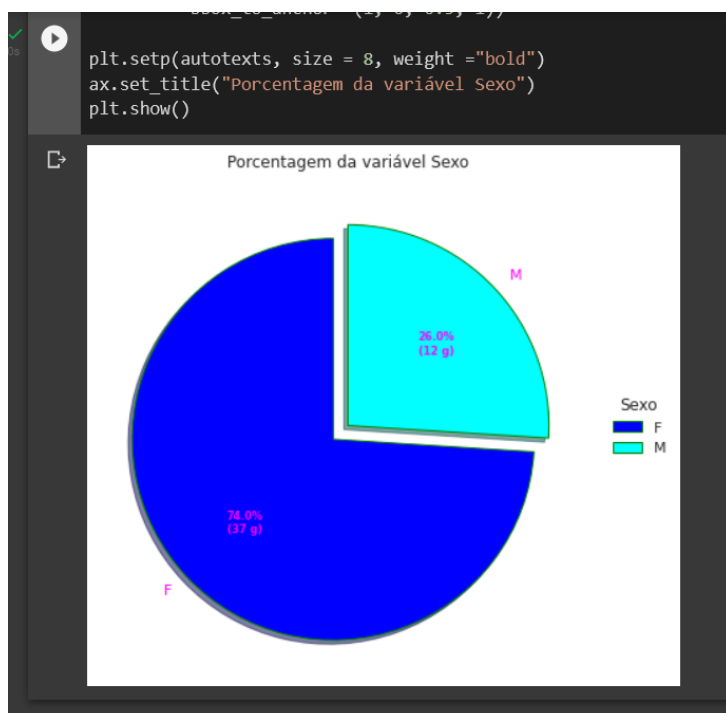
Figura 14: tabela de frequência da variável Sexo, em python.

```
[156] dist_freq_quali_Sexo = pd.DataFrame({'Frequência Absoluta': frequencia9, 'Porcentagem(%)': percentual9})
dist_freq_quali_Sexo
```

	Frequência Absoluta	Porcentagem(%)
F	37	74.0
M	13	26.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 6: gráfico de setores da variável Sexo.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002.

Ao analisarmos a variável sexo, é notória a concentração maior de indivíduos do sexo feminino, com 74% das pessoas, contra 26% de indivíduos do sexo masculino.

➤ Variável Fuma

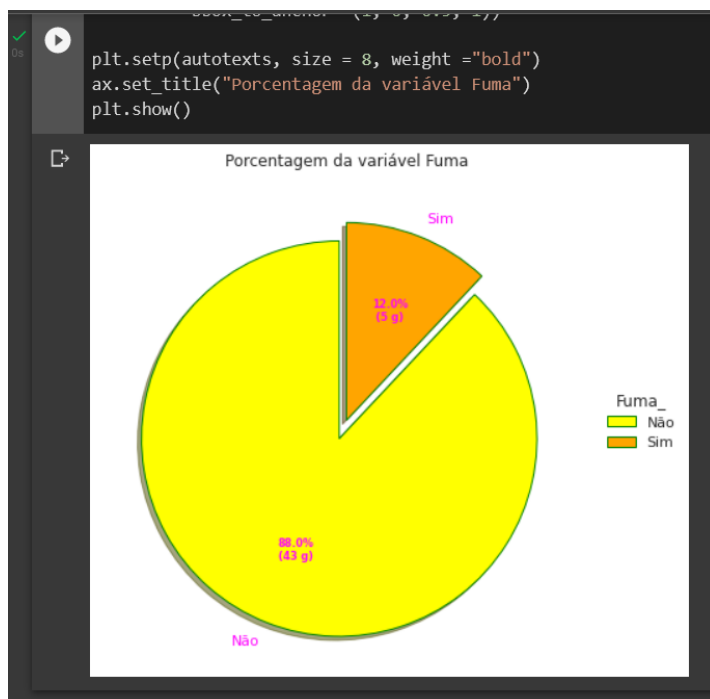
Figura 15: tabela de frequência da variável fuma, em python.

```
[159] dist_freq_quali_Fuma = pd.DataFrame({'Frequência Absoluta': frequencia10, 'Porcentagem(%)': percentual10})
dist_freq_quali_Fuma
```

	Frequência Absoluta	Porcentagem(%)
Nao	44	88.0
Sim	6	12.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 7: gráfico de setores da variável Fuma.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando a variável fumo, nota-se uma concentração extrema nos indivíduos que não possuem o hábito de fumar com 88% das observações contra apenas 12% de indivíduos que fumam. Em valores absolutos, 44 pessoas não fumam e apenas 6 indivíduos fumam.

➤ Variável Tolerância

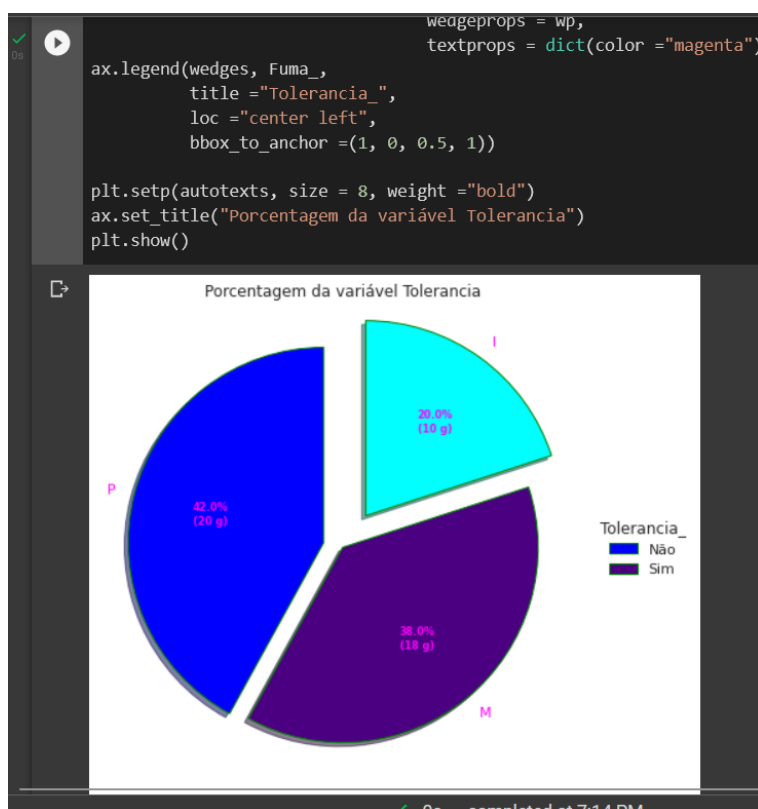
Figura 16: tabela de frequência da variável tolerância, em python.

```
[162] dist_freq_quali_Tolernacia = pd.DataFrame({'Frequência Absoluta': frequencia11, 'Porcentagem(%)': percentual11})
dist_freq_quali_Tolernacia
```

	Frequência Absoluta	Porcentagem(%)
P	21	42.0
M	19	38.0
I	10	20.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 8: gráfico de setores da variável tolerância.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Quanto a tolerância ao fumo, a maior concentração se dá no grupo de indivíduos que tem pouca tolerância, totalizando 42% e na sequência temos o grupo que tem muita tolerância ao cigarro, com 38%.

➤ Variável OpCine

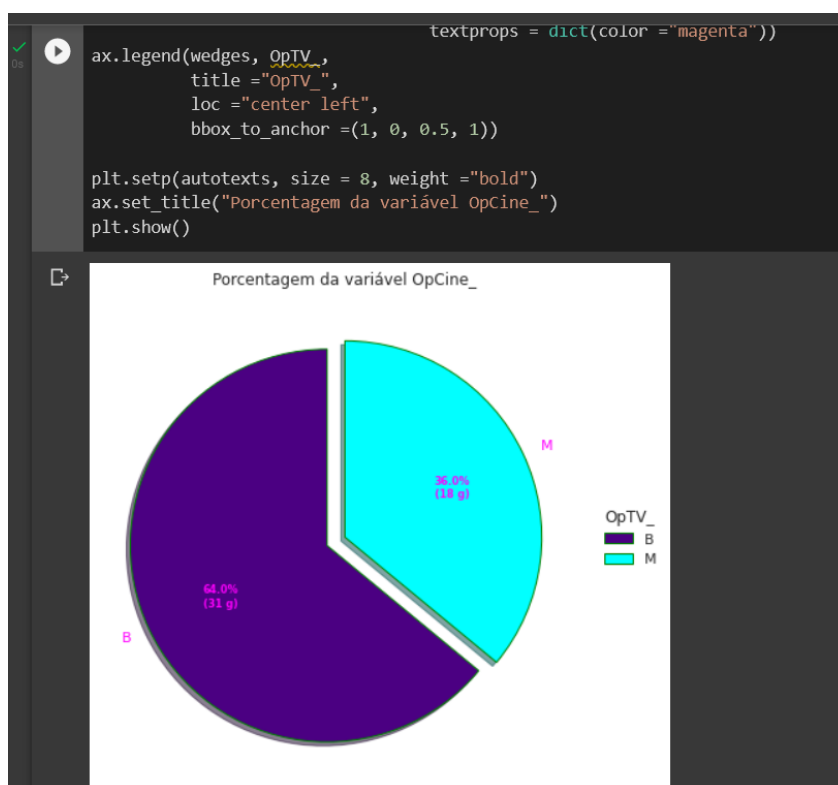
Figura 17: tabela de frequência da variável OpCine.

```
dist_freq_quali_OpCine = pd.DataFrame({'Frequência Absoluta': frequencia12, 'Porcentagem(%)': percentual12})
dist_freq_quali_OpCine
```

	Frequência Absoluta	Porcentagem(%)
B	32	64.0
M	18	36.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 9: gráfico de setores da variável OpCine.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Ao analisar a variável OpCine, é possível concluir que a maioria dos entrevistados considera a qualidade das salas de cinema da região como regular a boa (64%) e só 18 pessoas classificaram as salas como muito boas, ou seja, apenas 36% dos indivíduos.

➤ Variável OpTV

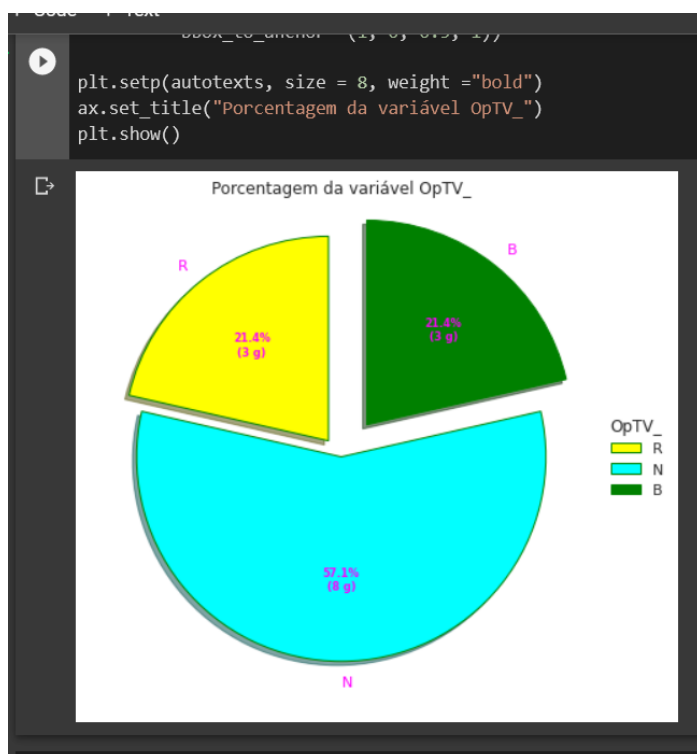
Figura 18: tabela de frequência da variável OpTV, em python.

```
[168] dist_freq_quali_OpTV = pd.DataFrame({'Frequência Absoluta': frequencia13, 'Porcentagem(%)': percentual13})
      dist_freq_quali_OpTV
```

	Frequência Absoluta	Porcentagem(%)
R	39	78.0
N	8	16.0
B	3	6.0

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 10: gráfico de setores da variável OpTV.

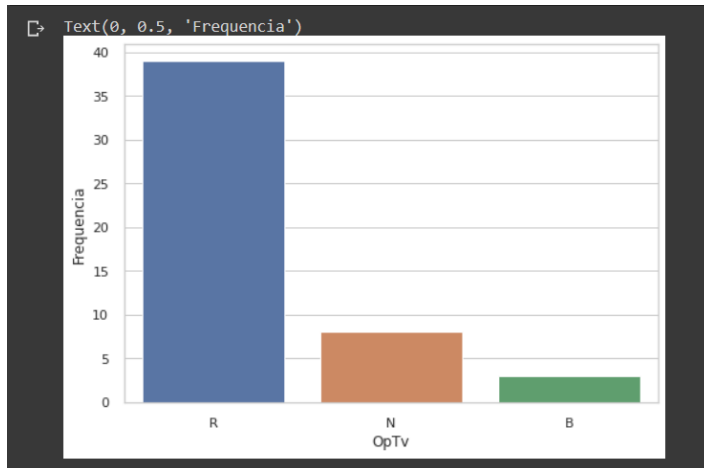


Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

A maioria dos entrevistados considera a qualidade da programação da TV com qualidade ruim (78%). Em seguida, 16% não soube classificar e 6% classificaram como boa.

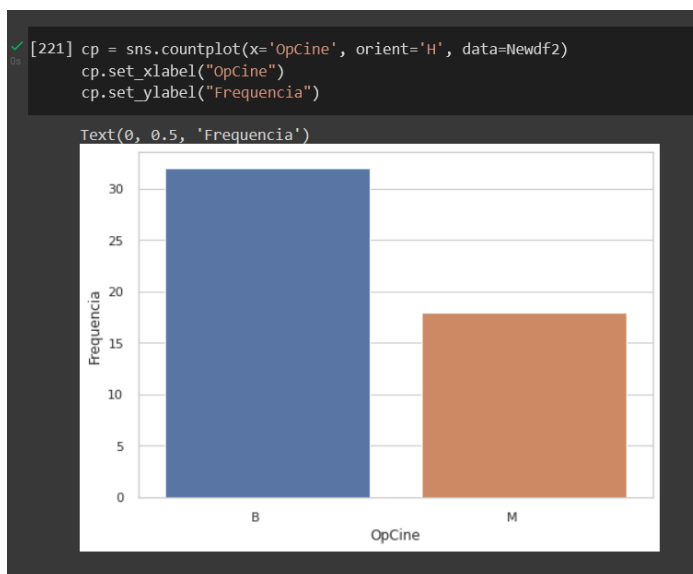
4. Análise de duas variáveis qualitativa e grau de dependência entre elas. Variáveis escolhidas: OpCine e OpTv

Gráfico 11: gráfico de barras OpTv x Frequência.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 12: gráfico de barras OpCine x Frequência.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Tabela 5: tabela de distribuição das opiniões dos alunos sobre TV e Cinema.

Qualificação de Opinião de TV e Cinema			
Opinião TV	Opinião Cinema		Total
	Regular a Boa	Muito Boa	
Ruim	27 (84%)	12 (67%)	39 (78%)
Média	0 (0%)	0 (0%)	0 (0%)
Boa	2 (6%)	1 (6%)	3 (6%)
Não Sabe	3 (10%)	5 (27%)	8 (16%)
Total	32 (64%)	18 (36%)	50 (100%)

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

A tabela acima mostra a distribuição das opiniões dos alunos sobre a TV e Cinema. Primeiramente, os dados mostram que, independente da variável relacionada, a maioria (78%) tem uma opinião ruim sobre a TV. Caso não haja dependência entre estas variáveis, as mesmas proporções seriam esperadas ao analisar cada opinião sobre cinema.

Observando a tabela, as proporções da opinião Regular a boa de Cinema (84%, 10%) distanciam-se das marginais (78% e 16%, respectivamente), enquanto as restantes (0% e 6%) são exatamente iguais às marginais. Essa observação parece indicar que existe um grau de dependência entre as variáveis, porém será fraco. Então as variáveis Opinião de Cinema e Opinião de TV parecem ser associadas.

Para quantificar estas associações, iremos considerar o coeficiente de correlação, usando o coeficiente de contingência. Para tal, iremos primeiramente assumir a independência das variáveis e simular quais deveriam ser os seus respectivos valores neste contexto, onde iremos assumir que os valores observados seguiriam a proporção esperada, dada pela porcentagem de Opinião Cinema (neste caso, 64% e 36%). Estes dados estão dados através desta tabela:

Tabela 6: tabela de valores esperados assumindo independência entre as variáveis.

Valores Esperados Assumindo Independência Entre as Variáveis			
Opinião TV	Opinião Cinema		Total
	Regular a Boa	Muito Boa	
Ruim	25 (64%)	14 (36%)	39 (78%)
Média	0	0	0
Boa	2 (64%)	1 (36%)	3 (6%)
Não Sabe	5 (64%)	3 (36%)	8 (16%)
Total	32 (64%)	18 (36%)	50 (100%)

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

E então poderemos verificar o desvio obtido através da diferença dos valores observados pelos valores esperados

Tabela 7: tabela de desvio entre observados e esperados.

Desvio Entre Observados e Esperados		
Opinião TV	Opinião Cinema	
	Regular a Boa	Muito Boa
Ruim	2 (0,16)	-2 (0,29)
Média	0	0
Boa	0	0
Não Sabe	-2 (0,8)	2 (1,33)

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Ao comparar ambas as tabelas, podemos observar que nas medidas que as proporções individuais se afastaram dos valores marginais, houve discrepância entre os valores observados e esperados. Como dito inicialmente, por indicar uma dependência fraca, os desvios da suposição de não-associação foram pequenos e semelhantes (2 e -2) e onde as proporções individuais foram iguais às marginais, o desvio foi igual a zero.

Embora os desvios possuam valores iguais, podemos observar que a medida da opinião Muito Boa x Não sabe (1,33) foi maior que todos os outros. Para determinar a medida de afastamento global, poderemos calcular o qui-quadrado (χ^2), obtida através da soma de todas as medidas.

$$\chi^2 = 0,16 + 0,29 + 0,8 + 1,33 = 2,58$$

Valores muito grandes de qui-quadrado indicam forte associação, o que não está presente neste caso.

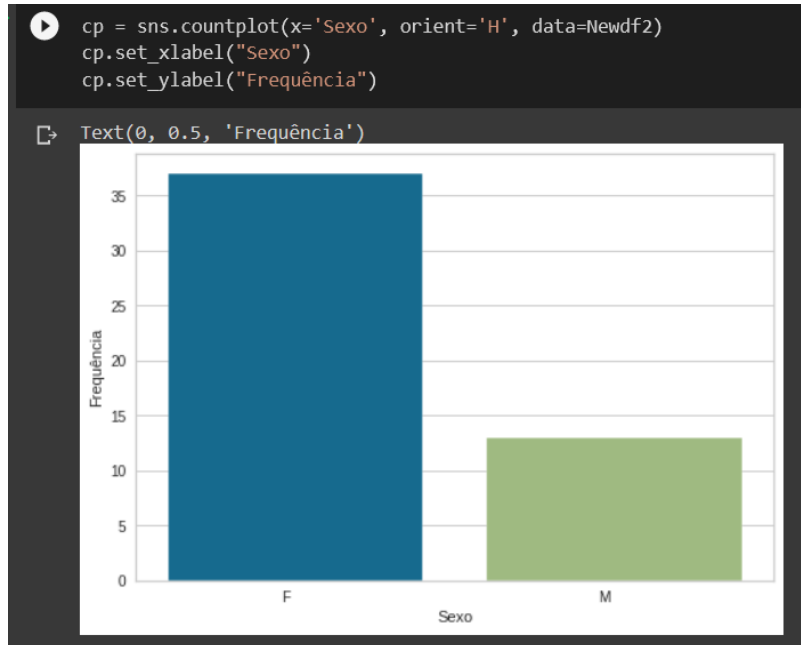
Para calcular uma medida entre 0 e 1 que nos indique um grau de dependência ou independência completa, podemos usar o coeficiente de contingência:

$$C = = = 0,22$$

Com nossas variáveis quantificadas a partir deste coeficiente e obtendo um valor próximo à zero, podemos concluir que as variáveis possuem uma fraca associação entre si.

5. Análise de uma variável quantitativa com uma qualitativa e grau de relação entre elas. Variáveis escolhidas: Sexo e TV.

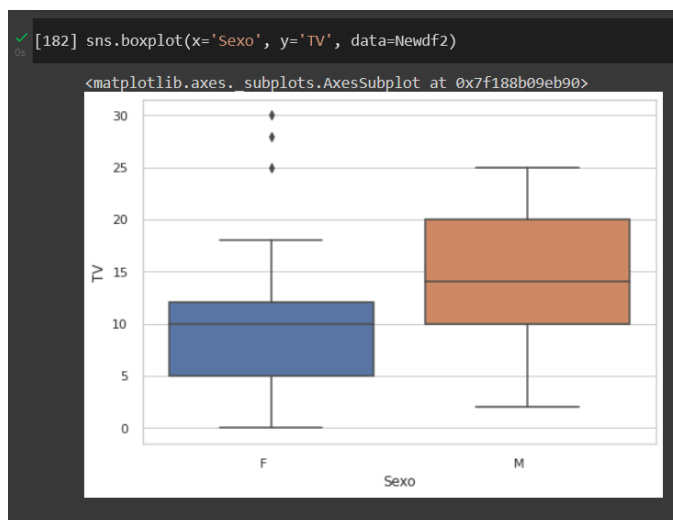
Gráfico 13: gráfico em barra das variáveis Sexo e frequência.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando o gráfico em barras, vê-se que as mulheres assistem TV por mais horas na semana que os homens. A amplitude entre as mulheres é de 30h, já entre os homens são de 25h.

Gráfico 14: gráfico de boxplot das variáveis Sexo e TV.



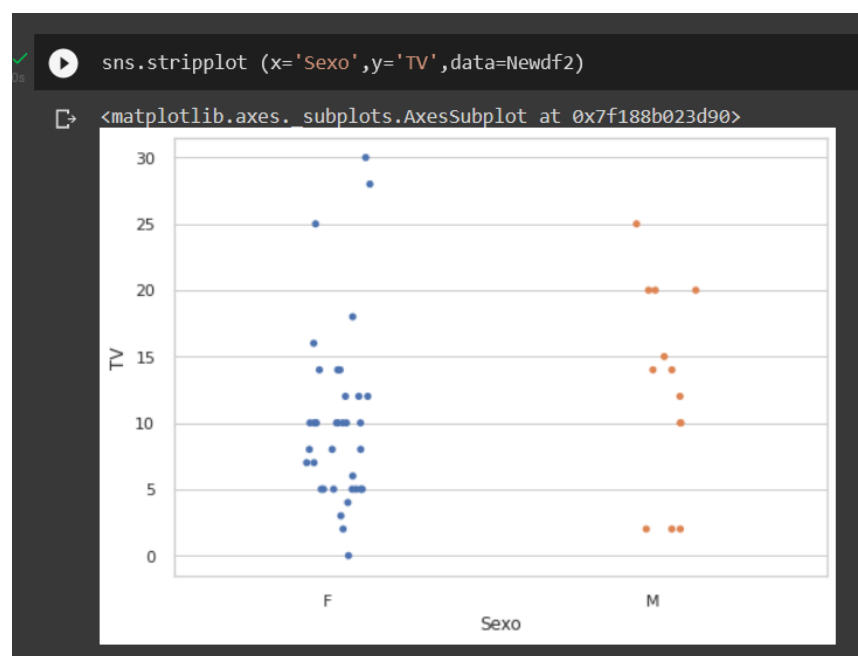
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando o box-plot, de forma separada a variável qualitativa, 75% das mulheres assistem mais que 5 horas por semana, apresentando uma mediana de 10 horas. Pela diferença dos intervalos interquartis, o conjunto dos dados femininos parece ter assimetria à esquerda (negativa, dispersão para valores menores). Trazendo a atenção para 3 valores que estão acima do limite superior, os outliers, dados que destoam do conjunto.

Agora analisando o sexo masculino, 75% deles assistem mais que 10 horas semanais, apresentando aproximadamente uma mediana de 14h. Calculando também a diferença entre os intervalos interquartis, o conjunto de dados referente aos homens, aparenta ter uma leve assimetria à direita (positiva, dispersão para valores maiores).

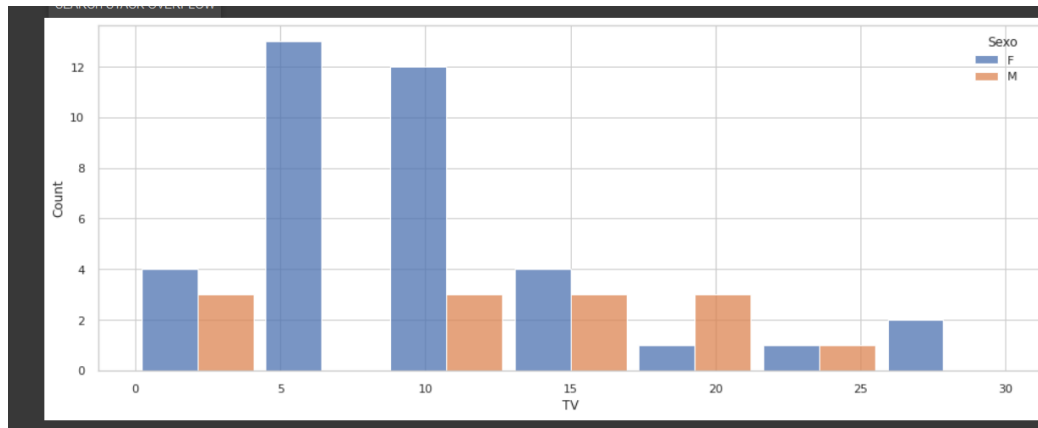
Comparando os dois gráficos, o sexo masculino tem maior mediana que o feminino, sendo assim 50% dos homens assistem mais TV que 75% das mulheres. Além disso, calculando a amplitude, conclui-se que o conjunto de dados do sexo feminino teve maior variância que o masculino, podendo ser explicada pelos dados discrepantes (outliers).

Gráfico 15: dotplot das variáveis Sexo e TV.



O gráfico de dispersão mostra que as horas assistidas semanalmente entre as mulheres tem maior concentração no intervalo [5,15] e dispersão para valores maiores. Já entre os homens, o intervalo de maior concentração está entre [10,15], aparentando ter maior dispersão para valores maiores.

Gráfico 16: histograma das variáveis Sexo e TV.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

.axes._subplots.AxesSubplot at 0x7f188b27ef10>

Analisando o histograma, referente ao sexo feminino, aparenta ser um gráfico unimodal (apresenta um único pico), evidenciando frequência maior em 5 e 10 horas de TV assistidas por semana. Em contra partida, o sexo masculino parecer ser mais um gráfico uniforme que o feminino, apresentando maior frequência em 0, 10, 15 e 20 horas semanais assistidas.

Em relação a simetria, o conjunto de dados referente às mulheres parece ter assimetria à direita, diferentemente no box-plot, e curva leptocúrtica (mais pontiaguda que a normal). Em relação aos homens uma leve assimetria à direita, como visto anteriormente no box-plot, e uma curva platicúrtica (mais achata que a a normal).

Em comparação, no geral as mulheres assistem mais horas de TV que os homens, por semana.

Grau de relação entre a variável Sexo e TV

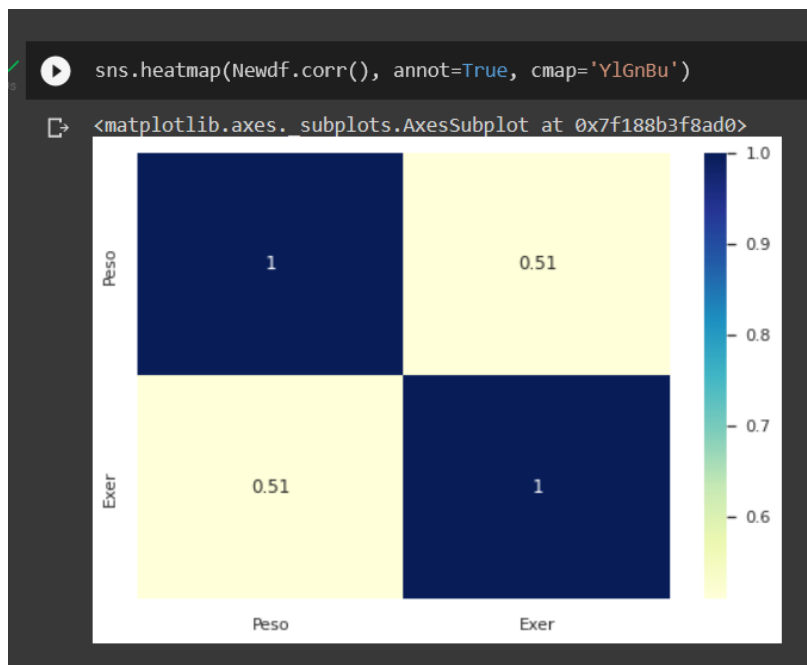
O grau de relação entre as variáveis sexo e TV não pode ser calculado a partir do coeficiente de correlação de Pearson e de Spearman, pelo fato de que são variáveis qualitativa e quantitativa, respectivamente. Por isso, se é utilizado o box-plot ou um gráfico de densidade para ver se há relação.

Como já temos o box-plot das variáveis, iremos analisar a partir dele.

Quando comparados os gráficos do sexo feminino e masculino, observa-se que seus dados estão posicionados em faixas mais afastadas um do outro, o que pode indicar uma relação. Como não há fatores que podem influenciar as variáveis, podemos concluir que há uma correlação de média para alta entre elas.

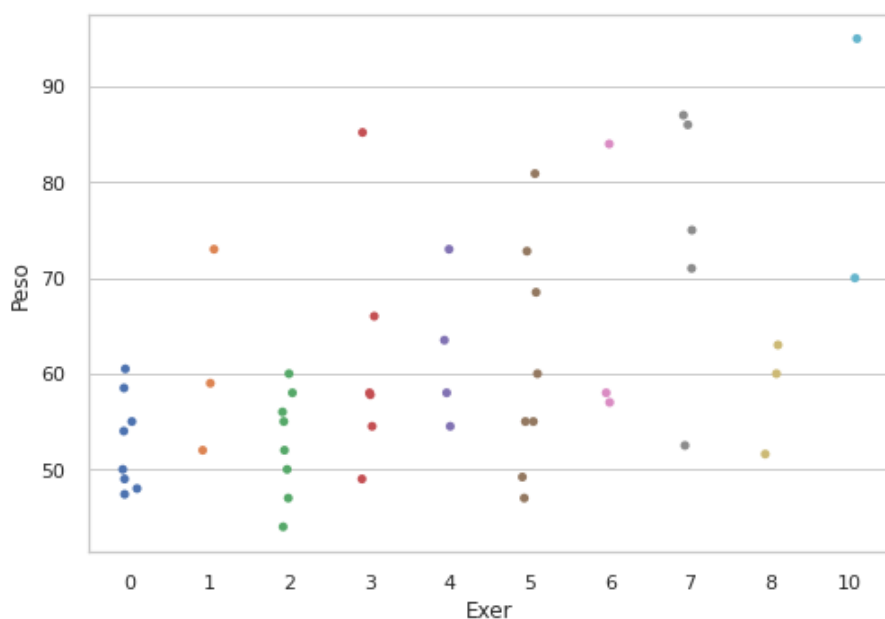
6. Análise de duas variáveis quantitativas e grau de relação linear entre elas (quantificar essa relação). Variável escolhida: Peso e Exercício.

Gráfico 16: gráfico de correlação linear entre as variáveis peso e exercício.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. São Paulo. EdUSP, 2002

Gráfico 17: dotplot das variáveis peso e exercício.



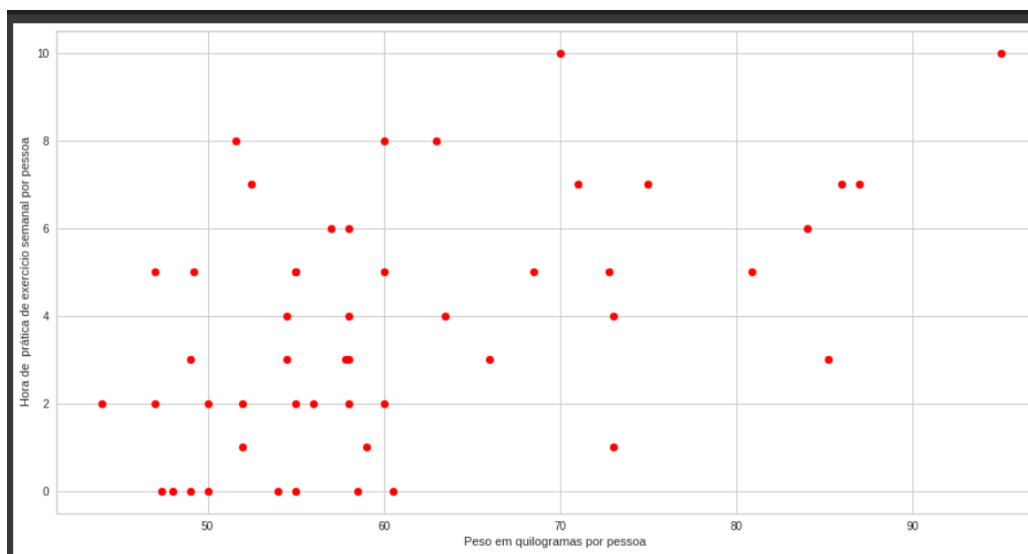
Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

O coeficiente de correlação de Pearson (r) foi de 0.51, representando uma correlação moderada entre o peso e exercícios.

Lembrando que coeficiente de correlação de Pearson mede o grau da correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados entre -1,0 e 1,0 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados.

Há Possível associação positiva o entre peso e exercícios.

Gráfico 19: gráfico de dispersão das variáveis Peso e Exercício.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Figura 19: ajuste da reta das variáveis peso x exercício, em python.

```
▶ X = dataframe['Peso'].values.reshape(-1,1)
  y = dataframe['Exer'].values.reshape(-1,1)

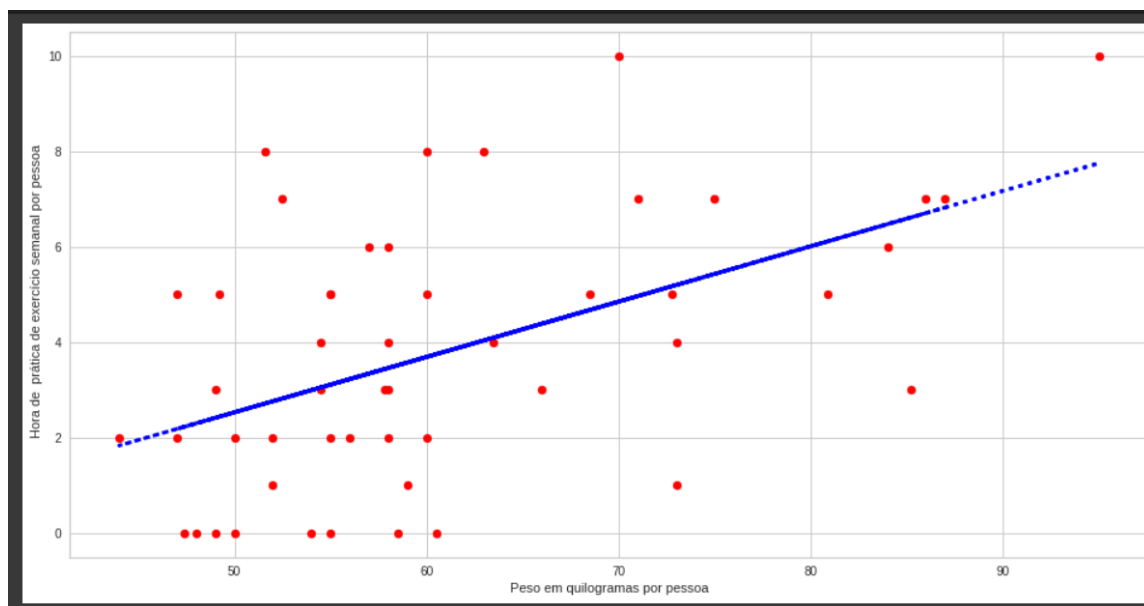
reg = LinearRegression()
reg.fit(X, y)

print("O modelo é: Peso x Exer = {:.5} + {:.5}X".format(reg.intercept_[0], reg.coef_[0][0]))

O modelo é: Peso x Exer = -3.2624 + 0.11591X
```

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 20: gráfico de dispersão das variáveis Peso e Exercício com a reta ajustada.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

$$Y = -3.2524 + 0.11591X$$

A reta acima implica que, a partir de 47 Kg, para cada 1 unidade de peso, haverá um aumento de y horas de exercícios.

7. Análise com 3 variáveis (a) duas quantitativas com uma qualitativa e (b) a relação linear existente entre as variáveis quantitativas).

Tabela 8: tabela de medida de resumo das variáveis altura x sexo.

Medidas de resumo altura x sexo				
Sexo	n	V(X)	Min	Max
Feminino	37	0,0043	1,45	1,82
Masculino	13	0,0026	1,7	1,85
Todos	50	0,0082	1,45	1,85

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Média do Sexo Feminino = 1,64

Média do Sexo Masculino = 1,78

Média das Variâncias = $(37 \cdot 0,0043) + (13 \cdot 0,0026) / 50 = 0,0038$

Grau de Associação = $1 - 0,0038 / 0,0082 = 0,5366$

O Grau de associação entre as variáveis altura e sexo é de 53%

Tabela 9: tabela de medida de resumo das variáveis peso x sexo.

Medidas de resumo peso x sexo				
Sexo	n	V(X)	Min	Max
Feminino	37	32,1	44	66
Masculino	13	83,0	55	95
Todos	50	148,32	44	95

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Média Peso Feminino = 40,7

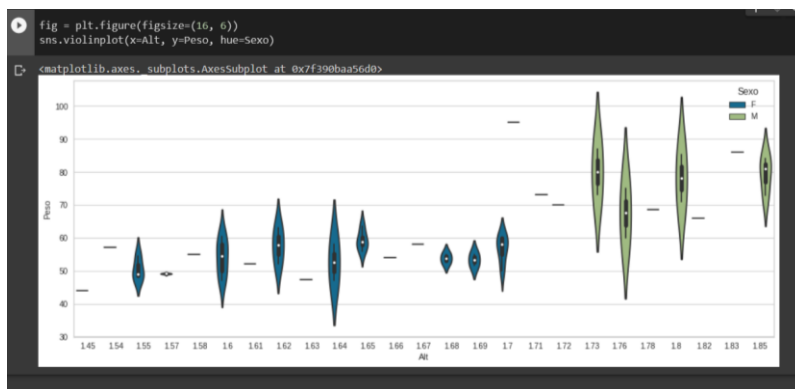
Média Peso Masculino = 20,22

Média das Variâncias = $(37 \cdot 31,1) + (13 \cdot 83) / 50 = 23,29$

Grau de Associação = $1 - 23,29 / 148,32 = 0,84$

O grau de associação entre as variáveis Peso e sexo é de 84%

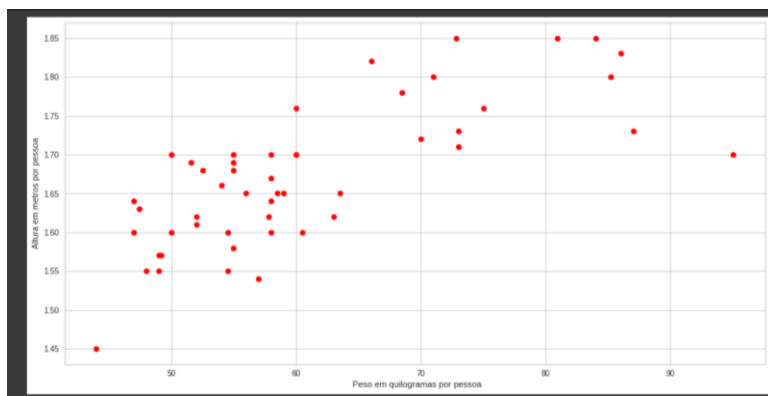
Gráfico 21: gráfico de violino das variáveis Peso e altura e sexo.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Verificando o gráfico de violino que cruza as 3 variáveis podemos notar que de fato, a variável peso e sexo e altura tem uma alta associação, pois quanto menor o peso e a altura temos classificações femininas e quanto maior o peso e a altura temos classificações masculinas.

Gráfico 22: gráfico de dispersão das variáveis Peso e altura.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando o gráfico de dispersão, notamos que parece haver uma associação linear entre as duas variáveis. Ou seja, a medida que aumenta a Altura, aumenta o peso. Podemos observar também um outline no último ponto presente no gráfico.

Figura 20: ajuste da reta das variáveis peso x altura, em python.

```
X = dataframe['Peso'].values.reshape(-1,1)
y = dataframe['Alt'].values.reshape(-1,1)

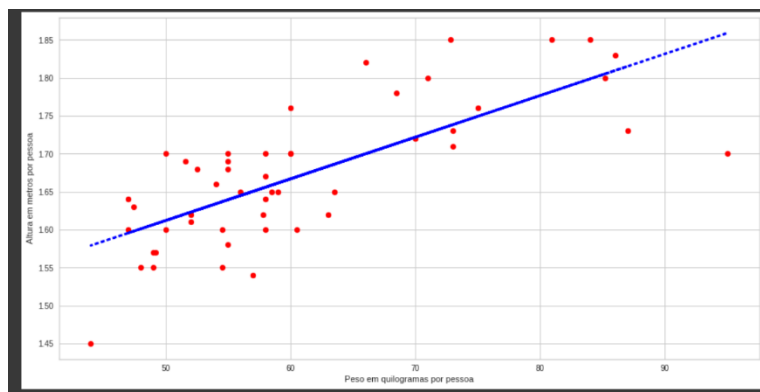
reg = LinearRegression()
reg.fit(X, y)

print("O modelo é: Peso x Alt = {:.5} + {:.5}X".format(reg.intercept_[0], reg.coef_[0][0]))
```

O modelo é: Peso x Alt = 1.3376 + 0.0054876X

Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Gráfico 23: ajuste da reta das variáveis Peso e altura.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Com o ajuste da reta podemos confirma a verificar essa associação que a partir de 44 a medida que vai aumentando a altura aumenta também o peso. Podendo ser previsto no ajuste $y = 1.3376 + 0.0054876x$.

Gráfico 23: gráfico de correlação linear das variáveis Peso e altura.



Fonte: Magalhães, Marcos Nascimento. Noções de probabilidade e estatística. 4.ed. SãoPaulo. EdUSP,2002

Analisando o coeficiente de correlação, podemos dizer que existe um alto grau correlação linear de pearson entre as duas variáveis, pois foi encontrado o valor de 0,74, proximo a 1.