UNIVERSIDADE DE SÃO PAULO - USP

SME0320 - Estatística I

Trabalho 1 - Análise Descritiva

Pedro Avellar Machado - 9779304

SÃO CARLOS
31 de Maio de 2020

1. INTRODUÇÃO

Este trabalho foi desenvolvido para a disciplina de Estatística I e tem como objetivo colocar em prática o conteúdo abordado de Análise Descritiva. Para isso foi disponibilizado um conjunto de dados a ser analisado, usando medidas, gráficos e tabelas, e interpretações destes.

O relatório está organizado da seguinte forma: apresentação do conjunto de dados e das variáveis escolhidas para análise; a análise em si e as interpretações; e as considerações finais.

A implementação utilizada para obter os resultados apresentados foi feita em linguagem *Python* através da ferramenta *Google Colab*. O *notebook* com os códigos, execuções, comentários e notas está disponível em https://github.com/pedroavellar/Estatistica-DataAnalysis/blob/master/T1Estat.ipynb

2. CONJUNTO DE DADOS

O conjunto de dados, fornecido para este trabalho, apresenta informações de funcionários de uma empresa X. Contém 50 observações, com as seguintes variáveis:

- Identificação do Funcionário
- **Idade - Quantitativa Discreta**
- **Salário (em salários mínimos) - Quantitativa Contínua**
- Número de filhos - Quantitativa Discreta
- Altura - Quantitativa Contínua
- Gênero - Qualitativa Nominal
- **Escolaridade - Qualitativa Ordinal**
- Horas trabalhadas na última semana - Quantitativa Discreta
- Região de Procedência - Qualitativa Nominal
- Peso - Quantitativa Contínua

A última observação do conjunto de dados deveria ser completada com valores, de forma que cada grupo tenha um conjunto diferente. Os valores foram escolhidos arbitrariamente e são:

Identificação do Funcionário	Idade	Salário (em salários mínimos)	Número de filhos	Altura	Gênero	Escolaridade	Horas trabalhadas na última semana	Região de Procedência	Peso
50	28	3,3	2	1,80	Masculino	médio	40	São Paulo	81,2

Três variáveis, de diferentes tipos (qualitativa, quantitativa discreta e quantitativa contínua), deveriam ser escolhidas para a análise. As variáveis escolhidas foram: **Idade, Salário e Escolaridade**

3. ANÁLISE E INTERPRETAÇÃO

- **Idade**

Medidas Descritivas:

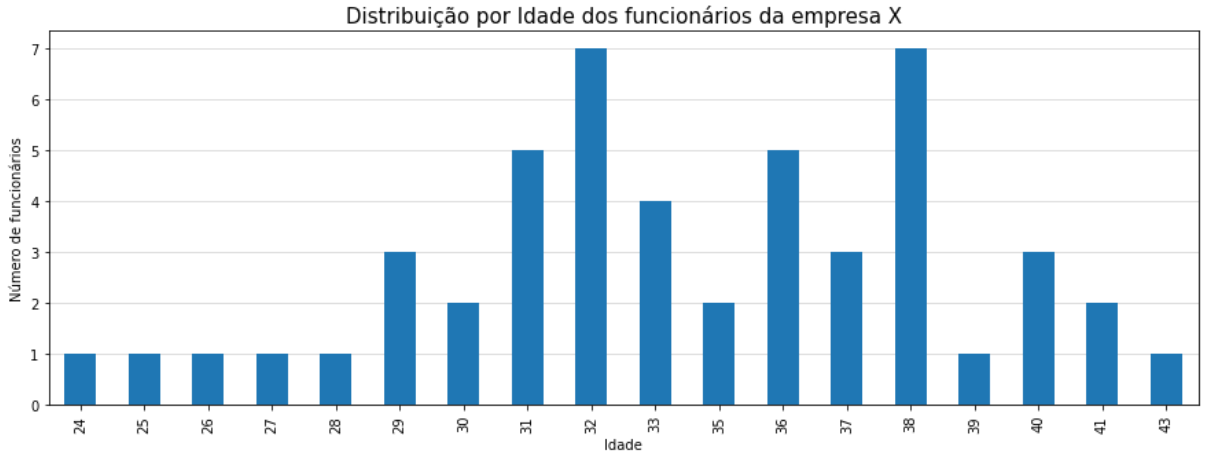
Medidas Gerais		
Quantidade	Mínimo	Máximo
50	24	43

Medidas de Posição		
Média	Mediana	Moda
33,98	33	32 e 38
Q1	Q2	Q3
31	33	38

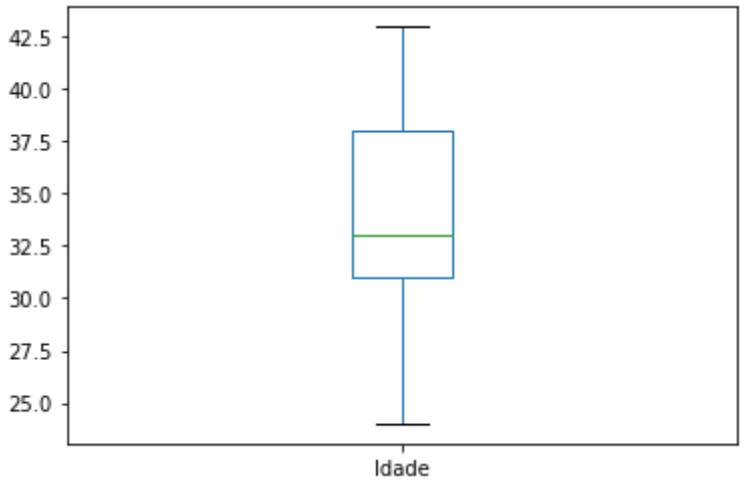
OBS: Q1, Q2, Q3: Primeiro, Segundo e Terceiro quartis, respectivamente

Medidas de Dispersão		
Amplitude		Intervalo interquartil (Q3-Q1)
19		7
Variância	Desvio padrão	Coeficiente de Variação
20,142	4,488	13,2%

Tabela de frequência		
Idade	Frequência absoluta	Frequência relativa
24	1	0.02
25	1	0.02
26	1	0.02
27	1	0.02
28	1	0.02
29	3	0.06
30	2	0.04
31	5	0.10
32	7	0.14
33	4	0.08
35	2	0.04
36	5	0.10
37	3	0.06
38	7	0.14
39	1	0.02
40	3	0.06
41	2	0.04
43	1	0.02
Total	50	1



Boxplot da Idade



O intervalo de idades desses funcionários está entre 24 e 43, existindo ocorrências em todas as idades deste intervalo. As maiores ocorrências são de 32 e 38 anos, 7 vezes cada. Percebemos que esta variável está bem distribuída, porém mais concentrada próximo ao meio do intervalo, como pode ser bem observado no *boxplot* acima.

- **Salário**

Medidas Descritivas:

Medidas Gerais		
Quantidade	Mínimo	Máximo
50	1,6	8,5

Medidas de Posição		
Média	Mediana	Moda
5,374	5,5	5,5
Q1	Q2	Q3
3,875	5,5	6,75

OBS: Q1, Q2, Q3: Primeiro, Segundo e Terceiro quartis, respectivamente

Medidas de Dispersão		
Amplitude		Intervalo interquartil (Q3-Q1)
6,9		2,875
Variância	Desvio padrão	Coeficiente de Variação
3,427	1,851	34,45%

OBS: Os intervalos e classes a seguir foram definidos de duas formas, as duas com número de intervalos, k = 8. Na primeira, como recomendado, com limite inicial da primeira classe no valor mínimo do conjunto e limite superior da última classe como valor máximo do conjunto. Como a seguir:

Porém, por conveniência, assim como também recomendado na disciplina, arredondamos LI_1 e *h* (amplitude da classe). Já que parece ser mais interessante a visualização de classes de 1|-- 2 salários mínimo, ao invés de 1,6 |-- 2,462. Gerando o novo histograma a seguir:

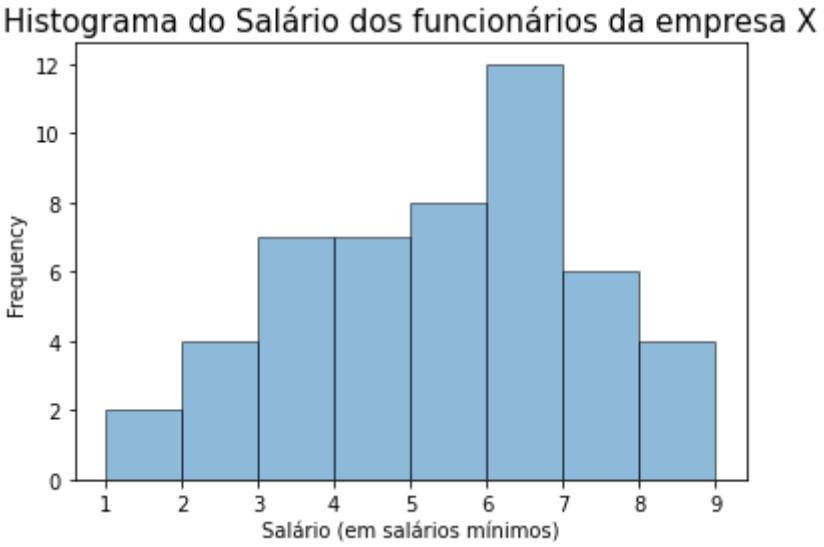
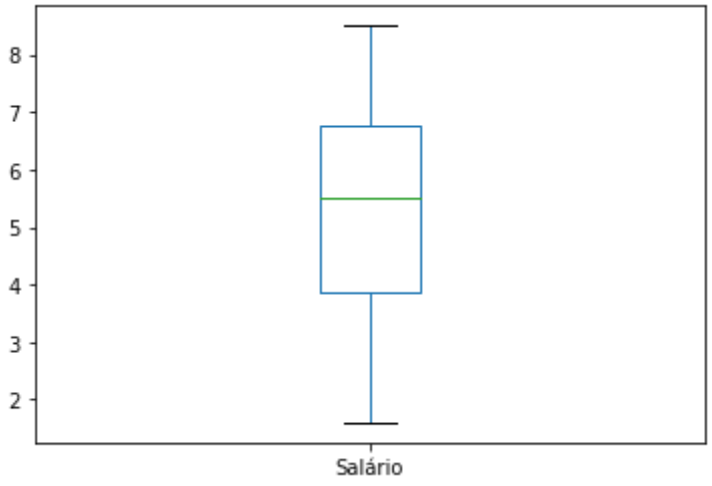


Tabela de frequência						
Ordem	Classe	Ponto médio	Frequência	Frequência relativa	Frequência acumulada	Frequência relativa acumulada
1	1 |-- 2	1,5	2	0.04	2	0.04
2	2 |-- 3	2,5	4	0.08	6	0.12
3	3 |-- 4	3,5	7	0.14	13	0.26
4	4 |-- 5	4,5	7	0.14	20	0.40
5	5 |-- 6	5,5	8	0.16	28	0.56
6	6 |--7	6,5	12	0.24	40	0.80
7	7 |-- 8	7,5	6	0.12	46	0.92
8	8 |-- 9	8,5	4	0.08	50	1

Boxplot do Salário



O menor salário desses funcionários é 1,6 salário mínimo, enquanto que o maior é 8,5. Da mesma forma que a variável anterior, a maior concentração da distribuição está no meio desta, principalmente entre 3 e 7. Vimos que o intervalo de salários com mais ocorrências é entre 6 e 7, correspondendo a 12 funcionários.

- **Escolaridade**

Tabela de frequência		
Grau de instrução	Frequência absoluta	Frequência relativa
‘fundamental’	3	0,06
‘médio’	28	0,56
‘superior’	19	0,38
Total	50	1

A maior parte dos funcionários deste conjunto tem nível de escolaridade de nível médio. Poucos funcionários têm apenas nível fundamental.

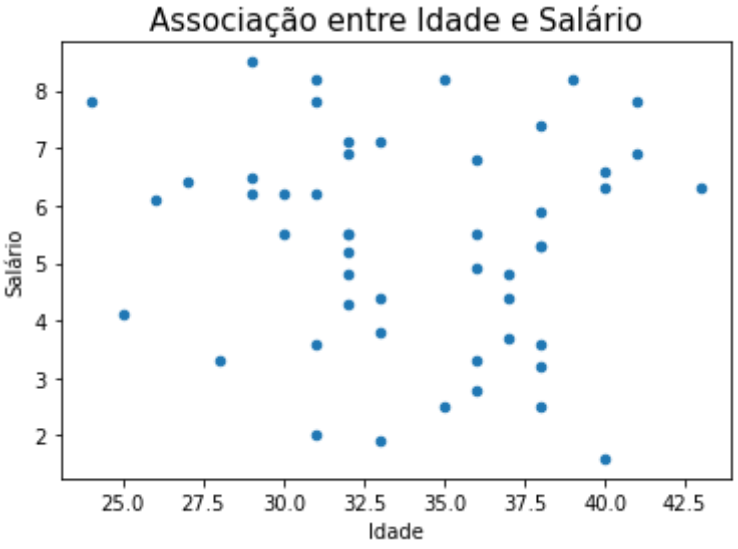
- **Análise Bivariada (Idade e Salário)**

Matriz de covariância:

	Idade	Salário
Idade	20.142449	-0.786245
Salário	-0.786245	3.427269

Matriz de correlação:

	Idade	Salário
Idade	1.00000	-0.09463
Salário	-0.09463	1.00000



Pode se observar que estas variáveis estão pouco relacionadas, com coeficiente de correlação (r) próximo de zero (-0,09). Por isso, o gráfico *scatter* é um caos sem apresentar nenhuma linearidade.

4. CONCLUSÃO

Após fazer a análise de dados de um conjunto, são obtidas informações importantes que podem não estar explícitas a primeira vista. As medidas calculadas nos mostram um comportamento geral da população ou amostra, permitindo mais relevância para compreensão ou tomada de decisões relacionadas a este grupo.

Entender os dados e suas características pode não ser muito trivial, para isso métodos de visualização, usando tabelas e gráficos adequados para determinada representação de diferentes tipos de dados, são muito úteis.