Tutorial da videoaula - Semana 3: Análise descritiva de dados (ADD)

Neste exercício, iremos fazer a análise descritiva de uma base de dados com o objetivo de aplicar o conteúdo visto nesta semana.

1. Crie um novo notebook e inclua uma descrição para ele.

2. Nesta atividade, além da biblioteca **pandas**, vamos importar as bibliotecas **seaborn** e **matplotlib** para gerar os gráficos.

```
import pandas as pd

import seaborn as sb

import matplotlib.pyplot as plt

sb.set(rc={'figure.figsize':(15,8)})
```

A base de dados deste exercício é a "Quantidade de alunos por tipo de ensino" da rede estadual de São Paulo de 2021. A base tem 25 atributos, entre os quais estão:

- **MUN**: município
- **ZONA**: zona (1 = urbana, 2 - rural)
- **ANOS INICIAIS**: Anos iniciais do Ensino Fundamental
- **ANOS FINAIS**: Anos finais do Ensino Fundamental
- **ENSINO MEDIO**: Anos iniciais do Ensino Fundamental

Os dados estão sendo importados diretamente do Portal de Dados Abertos da Educação do Estado de São Paulo. O dicionário de dados está disponível neste link

3. Importe a base de dados direto da URL a seguir e verifique as primeiras linhas. O arquivo contém 50 registros. Olhando os 20 primeiros, podemos observar que há valores ausentes nos atributos **idade**, **uf** e **renda**.

```
url =
'https://dados.educacao.sp.gov.br/sites/default/files/Quantidade%20de%20alunos%20por%20tipo%20de%20ensino%20da%20rede%20estadual%20-%202021.csv'

escolas = pd.read_csv(url)

escolas.head(20)
```

4. A função **info()**) mostra que há 5351 registros para a maioria dos atributos, portanto, é o número de escolas.

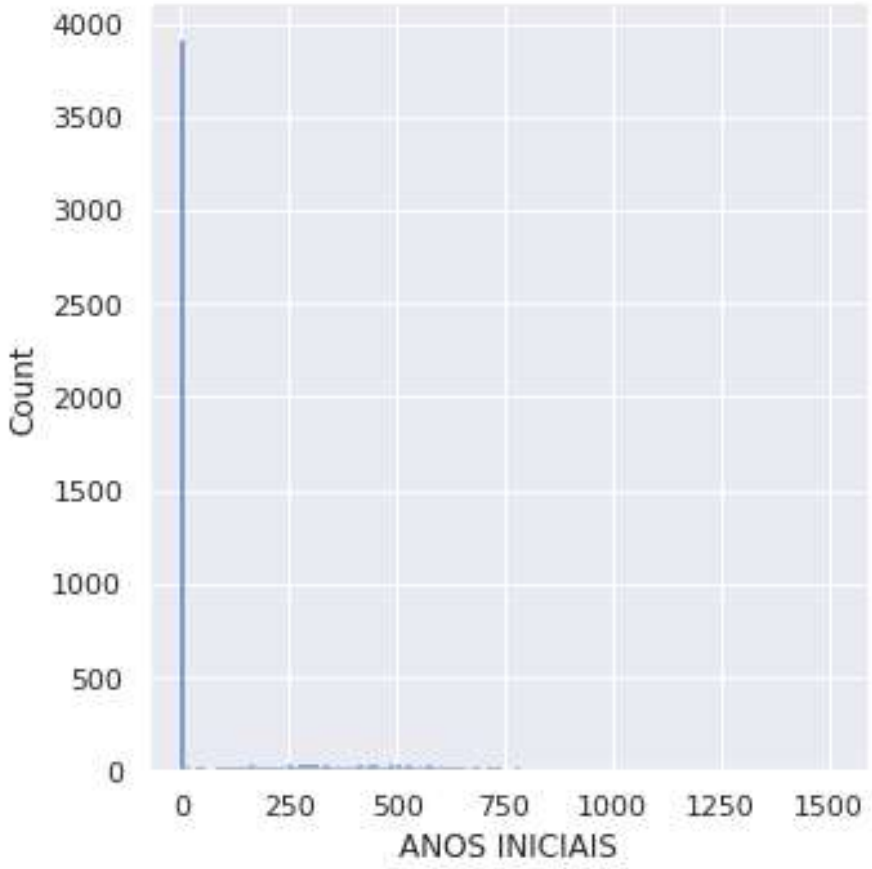
```
escolas.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5351 entries, 0 to 5350
Data columns (total 24 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   CDREDE             5351 non-null   int64  
 1   DE                 5351 non-null   object 
 2   CODMUN             5351 non-null   int64  
 3   MUN                5351 non-null   object 
 4   CATEG              5351 non-null   int64  
 5   COD_ESC            5351 non-null   int64  
 6   TIPOESC            5351 non-null   int64  
 7   CODVINC            259 non-null    float64
 8   NOMESC             5351 non-null   object 
 9   ENDESC             5351 non-null   object 
 10  NUMESC             5327 non-null   object 
 11  BAIESC             5348 non-null   object 
 12  EMAIL              5273 non-null   object 
 13  FONE1              5328 non-null   float64
 14  ZONA               5350 non-null   float64
 15  ED_INFANTIL        5351 non-null   int64  
 16  CLASSES ESPECIAIS  5351 non-null   int64  
 17  SALA DE RECURSO    5351 non-null   int64  
 18  ANOS INICIAIS      5351 non-null   int64  
 19  ANOS FINAIS        5351 non-null   int64  
 20  ENSINO MEDIO       5351 non-null   int64  
 21  EJA FUNDAMENTA_AI  5351 non-null   int64  
 22  EJA FUNDAMENTAL_AF  5351 non-null   int64  
 23  EJA ENSINO MEDIO   5351 non-null   int64  
dtypes: float64(3), int64(14), object(7)
```

Distribuição de frequências

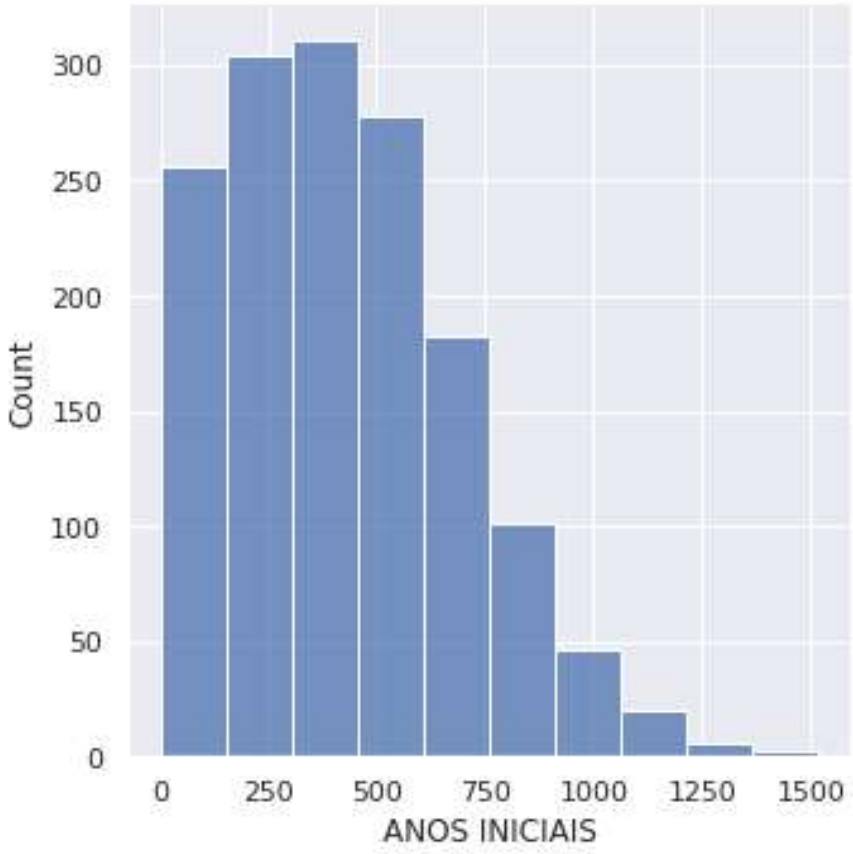
5. Vamos observar a distribuição de alguns atributos, gerando o histograma do atributo **ANOS INICIAIS** com a função **displot()**.

```
sb.displot(escolas['ANOS INICIAIS'])

plt.show()
```



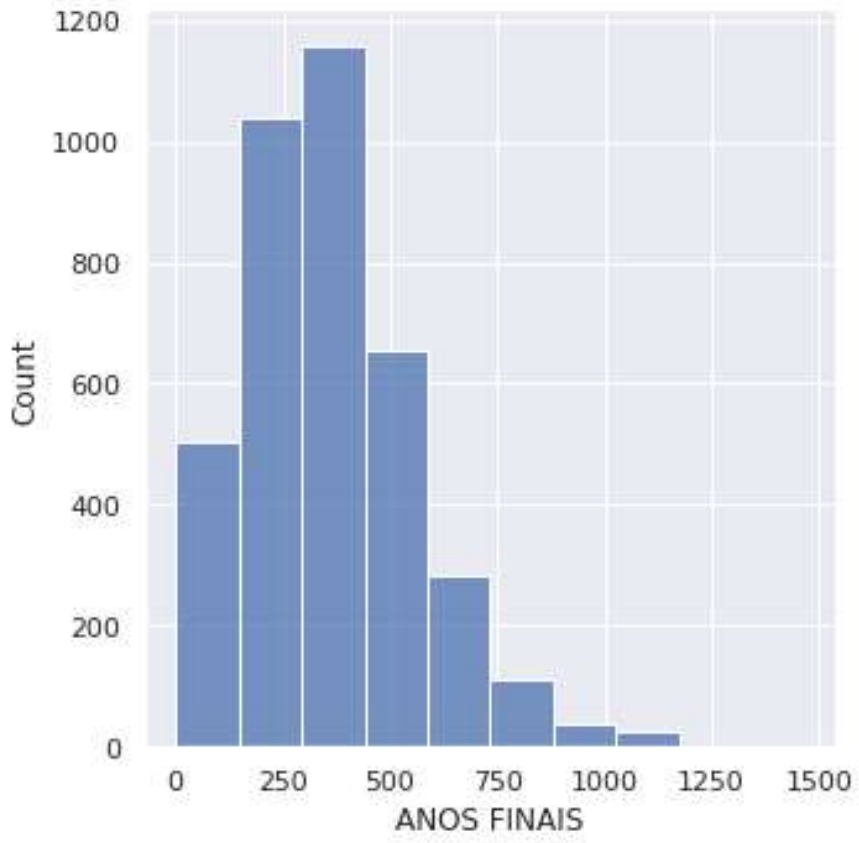
6. Os dados têm muitos valores 0 no campo **ANOS INICIAIS** porque há escolas que têm somente os anos finais do Ensino Fundamental, outras têm somente Ensino Médio etc. Vamos filtrar os dados e ver novamente selecionando o valor de 10 pastas para o histograma.

```
filtro_escolas = escolas[escolas['ANOS  INICIAIS'] > 0]

sb.displot(filtro_escolas['ANOS INICIAIS'],bins=10)

plt.show()
```



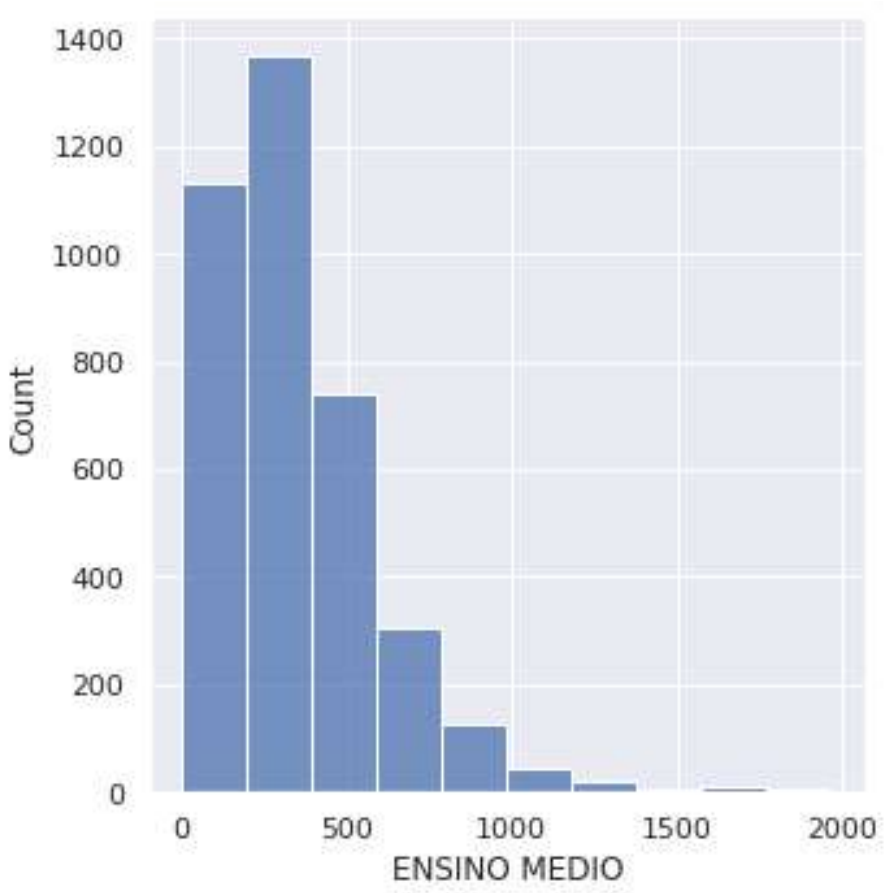
7. Vamos olhar o atributo **ANOS FINAIS**, já filtrando escolas sem alunos desses anos.

```
filtro_escolas = escolas[escolas['ANOS  FINAIS'] > 0]
sb.displot(filtro_escolas['ANOS FINAIS'],bins=10)
plt.show()
```



8. Para o atributo **ENSINO MEDIO** faremos igual.

```
filtro_escolas = escolas[escolas['ENSINO  MEDIO'] > 0]
sb.displot(filtro_escolas['ENSINO MEDIO'],bins=10)
plt.show()
```



Podemos observar que a maioria das escolas nas três faixas de ensino tem uma maior concentração de estudantes na faixa de 250 a 500 alunos.

9. Que tal agora olharmos a proporção de alunos no ensino regular dos anos iniciais e finais do Ensino Fundamental e do Ensino Médio?

Vamos somar o total de alunos matriculados em cada período e gerar um gráfico de setores usando a função **pie()** do mattplotlib.

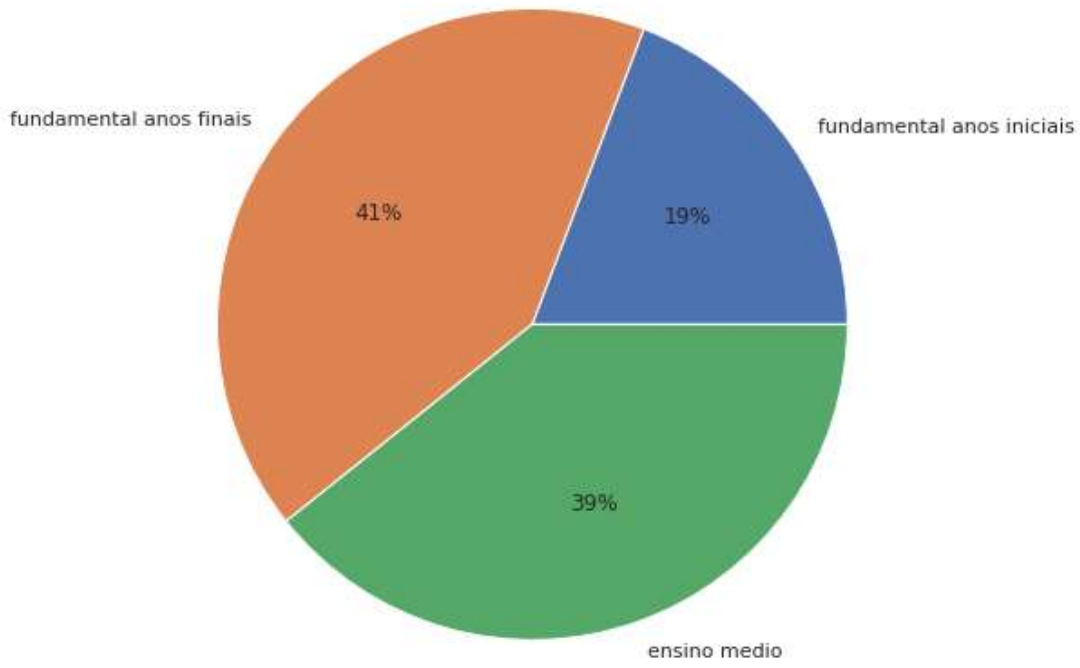
O total de alunos desses períodos é de mais de 3.3 milhões.

```
fundamental_iniciais = escolas['ANOS_INICIAIS'].sum()

fundamental_finais = escolas['ANOS_FINAIS'].sum()

medio = escolas['ENSINO_MEDIO'].sum()

(fundamental_iniciais + fundamental_finais + medio)
```

A soma dos valores de cada um dos três períodos serão os dados passados como uma lista **alunos** para a função **pie**.

Vamos passar também os rótulos para identificar os dados como uma lista **periodo**. O parâmetro **autopct** coloca os percentuais dos dados no gráfico no formato desejado.

```
alunos = [fundamental_iniciais,fundamental_finais,medio]

periodo = ['fundamental anos iniciais', 'fundamental anos finais', 'ensino medio']

plt.pie(alunos,labels=periodo,autopct = '%0.0f%%')

plt.show()
```



Assim, vemos que as maiores concentrações de alunos do ensino público estadual são do Ensino Fundamental nos anos finais, seguido de perto pelos alunos do Ensino Médio.

Para simplificar o exemplo, não estamos considerando aqui os alunos do EJA (Ensino de Jovens e Adultos) nem os alunos do Ensino Infantil.

Posição relativa

10. Vamos seguir olhando as medidas de posição relativa dos dados usando o diagrama de caixa (gráfico boxplot) do Ensino Médio.

```
filtro_escolas = escolas[escolas['ENSINO  MEDIO'] > 0]

sb.boxplot(y=filtro_escolas['ENSINO MEDIO'])

plt.show()
```

O boxplot nos mostra que há uma grande concentração de escolas com aproximadamente de 200 a 500 estudantes do Ensino Médio.

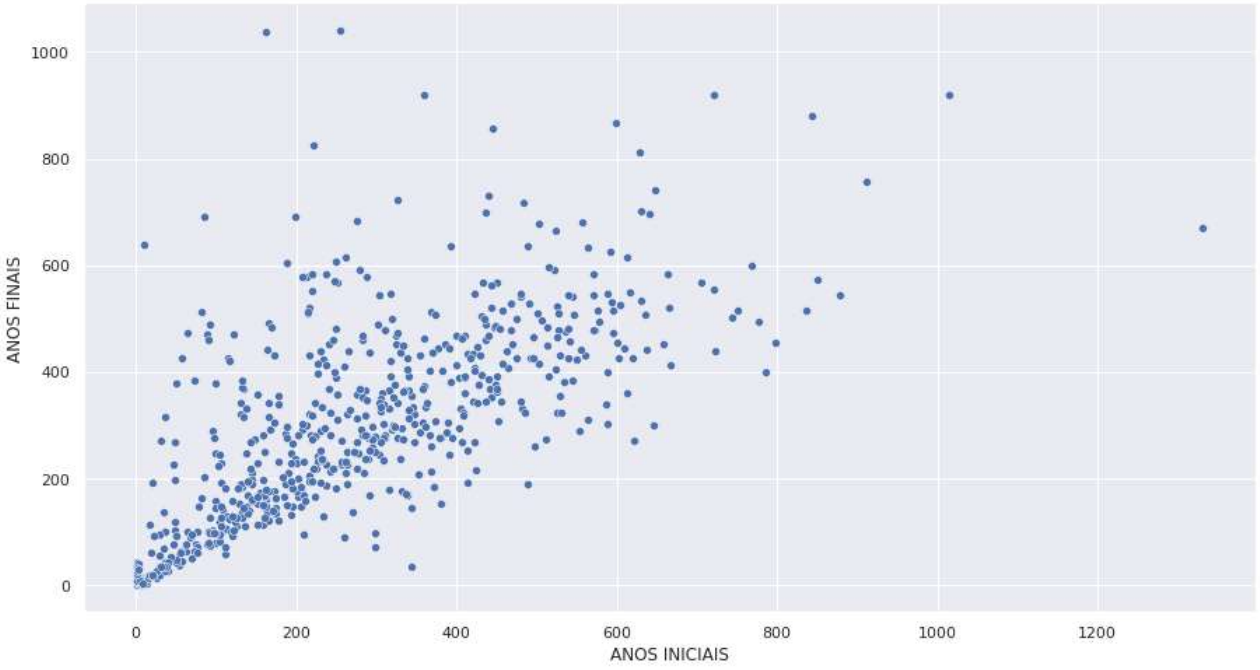
Tendência central e dispersão

11. A seguir, podemos ver as medidas de tendência central e dispersão desses dados com a função **describe()**, olhando a distribuição desses dados. Por exemplo, a média (mean) de alunos dos anos finais do Ensino Fundamental por escola é de ~257, com desvio-padrão de 234,97. Podemos ver também os valores dos quartis para esse atributo.

```
escolas['ANOS_FINAIS'].describe()
```

12. Vamos agora olhar se há alguma correlação entre atributos desses dados? Para isso, vamos usar o gráfico de dispersão para comparar os atributos com a quantidade de alunos das escolas com estudantes do Ensino Fundamental dos anos iniciais e finais. Primeiro, filtramos escolas cujo número de estudantes seja maior que zero em ambos os casos. Depois, usamos a função **scatterplot()** para plotar os dados.

```
filtro_escolas = escolas[(escolas['ANOS_INICIAIS'] > 0) & (escolas['ANOS_FINAIS'] > 0)]

sb.scatterplot(x=filtro_escolas['ANOS INICIAIS'],y=filtro_escolas['ANOS FINAIS'])

plt.show()
```



Embora haja alguma dispersão pelo gráfico, podemos ver uma grande concentração de escolas de ambos os períodos de Ensino Fundamental na faixa entre 1 e 600 estudantes, mostrando uma correlação positiva. Isso faz sentido, visto que filtramos escolas que abrigam os dois períodos de ensino e que, portanto, têm uma capacidade proporcional de alocar esses estudantes. Além do mais, a quantidade de estudantes tende a ser proporcional ao local na qual elas estão.

Com isso, concluímos o exercício de análise descritiva de dados desta semana.

Para praticar mais, sugiro que você escolha um outro conjunto de dados do seu interesse e use ADD para conhecer melhor as características dos dados.

Algumas sugestões de bases são:

- Portal Brasileiro de Dados Abertos
- Basedosdados.org
- Kaggle