resolucao

February 18, 2024

1 Proposta de resolução da TP1 (Análise de dados)

1.1 1.

A PORDATA, Base de Dados de Portugal Contemporâneo (https://www.pordata.pt/), constitui uma base de dados com "informação sobre múltiplas áreas da sociedade, para Portugal, municípios e países europeus". Em Pordata>Portugal>Ambiente, Energia e Território>Poluição Atmosférica e Clima, encontra o número de dias sem chuva em estações meteorológicas portuguesas, nas últimas décadas.

1.1.1 a)

Importe da folha de calculo os dados numéricos com os respetivos rótulos de linhas e de colunas. Selecione apenas as células que contém os dados relevantes.

1.1.2 Resolução:

[1]:
```
import pandas as pd
df = pd.read_excel('PORDATA_Dias_sem_chuva.xlsx', sheet_name='Quadro',skiprows␣
 ↪= 7,nrows= 61,usecols="A:J")

print(df)
```

```
    Unnamed: 0  Viana do Castelo  Bragança  Porto  Castelo Branco  Lisboa  \
0         1960                 0       195      0             243     212
1         1961                 0       212      0             263     253
2         1962                 0       269      0             290     267
3         1963                 0       212      0             242     225
4         1964                 0       256      0             283     262
..         …                 …       …    …               …      …
56        2016               212       228    218             253     245
57        2017               237       279    245             292     290
58        2018               181       231    185             245     255
59        2019               199       258    215             263     258
60        2020               180       234    201             240     247

    Beja  Faro  Funchal  Angra do Heroísmo
0    241     0      292                  0
1    271     0      317                  0
```

```
2    283     0      311                  0
3    232     0      291                  0
4    278     0      307                  0
..   …   …       …                  …
56   244   283      279                202
57   284   315      303                190
58   224   275      277                193
59   267   312      323                197
60     0   288      291                185

[61 rows x 10 columns]
```

1.1.3 b) Por visualização direta dos dados, verifique os rótulos das colunas. Corrija, se necessário, o(s) nome(s) dos rótulos das colunas.

1.1.4 Resolução:

- É conveniente renomear a 1ª coluna
- É igualmente aconselhavel eliminar os acentos os espaços e simplificar os nomes dos rótulos

[2]:
```
print(df.head(0), end='\n')
newcolrotulo={'Unnamed: 0': 'ano','Viana do Castelo':'Viana','Castelo Branco':
 ↪'C_Branco','Angra do Heroísmo':'Angra'}
df = df.rename(columns=newcolrotulo)


print(df)
```

```
Empty DataFrame
Columns: [Unnamed: 0, Viana do Castelo, Bragança, Porto, Castelo Branco, Lisboa,
Beja, Faro, Funchal, Angra do Heroísmo]
Index: []
     ano  Viana  Bragança  Porto  C_Branco  Lisboa  Beja  Faro  Funchal  Angra
0   1960      0       195      0       243     212   241     0      292      0
1   1961      0       212      0       263     253   271     0      317      0
2   1962      0       269      0       290     267   283     0      311      0
3   1963      0       212      0       242     225   232     0      291      0
4   1964      0       256      0       283     262   278     0      307      0
..   …    …       …    …       …      …   …    …       …    …
56  2016    212       228    218       253     245   244   283      279    202
57  2017    237       279    245       292     290   284   315      303    190
58  2018    181       231    185       245     255   224   275      277    193
59  2019    199       258    215       263     258   267   312      323    197
60  2020    180       234    201       240     247     0   288      291    185

[61 rows x 10 columns]
```

[3]:
```
print(df)
```

	ano	Viana	Bragança	Porto	C_Branco	Lisboa	Beja	Faro	Funchal	Angra
0	1960	0	195	0	243	212	241	0	292	0
1	1961	0	212	0	263	253	271	0	317	0
2	1962	0	269	0	290	267	283	0	311	0
3	1963	0	212	0	242	225	232	0	291	0
4	1964	0	256	0	283	262	278	0	307	0
..	…	…	…	…	…	…	…	…	…	…
56	2016	212	228	218	253	245	244	283	279	202
57	2017	237	279	245	292	290	284	315	303	190
58	2018	181	231	185	245	255	224	275	277	193
59	2019	199	258	215	263	258	267	312	323	197
60	2020	180	234	201	240	247	0	288	291	185

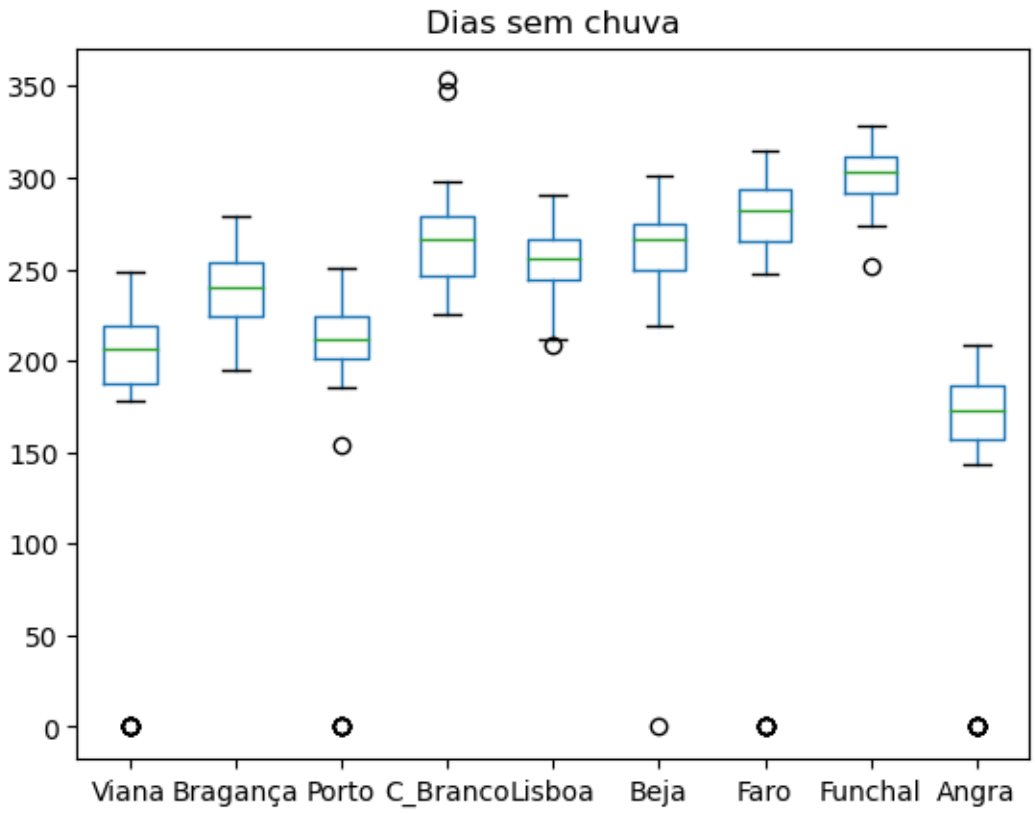
[61 rows x 10 columns]

1.1.5 c)

1.1.6 i) Construa um gráfico com os diagramas de extremos e quartis (box plot) que permita comparar os números de dias sem chuva nas estações meteorológicas da base de dados.

1.1.7 ii) Qual a estacão que registou mais e a que registou menos dias sem chuva nas últimas décadas? Analise o gráfico, referindo a concentração e a dispersão dos dados;

[4]:
```
estacoes=['Viana', 'Bragança', 'Porto', 'C_Branco','Lisboa', 'Beja', 'Faro',␣
 ↪'Funchal', 'Angra']
boxp1=df.boxplot(by =None, column = estacoes, grid = False)
import matplotlib.pyplot as plt
plt.title('Dias sem chuva')
plt.show()
```

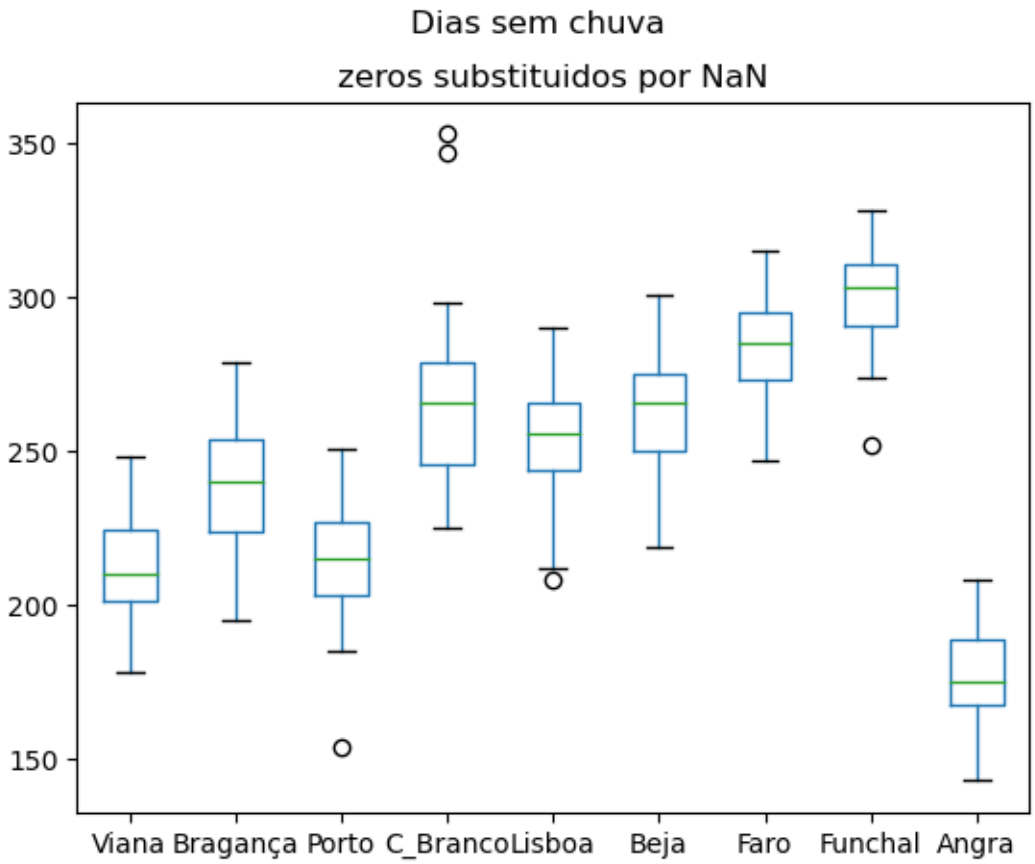


1.1.8 d) Comente a existência de zeros, nos dados e nos gráficos. Como podemos interpretar estes valores? Substitua todos os zeros por NaN e repita o exercício anterior. Quais as diferenças mais notórias?

1.1.9 Resolução:

Não é credivel que num ano não tenha havido pelo menos um dia de chuva. Os zeros significam que **não há dados relativos a esse anos/estações**.

[5]:
```
import numpy as np
df.replace(0, np.nan, inplace=True)
boxp2=df.boxplot(by = None, column = estacoes, grid = False)
plt.suptitle('Dias sem chuva')
plt.title('zeros substituidos por NaN')
plt.show()
```



[6]:
```
##### resumo estatistico
round((df.iloc[:,1:10]).describe(),2)
```

[6]:

	Viana	Bragança	Porto	C_Branco	Lisboa	Beja	Faro	Funchal	\
count	51.00	61.00	54.00	61.00	61.00	60.00	53.00	61.00	
mean	210.86	239.13	215.56	265.84	254.59	262.60	283.40	301.33	
std	16.87	18.18	19.06	24.25	17.59	17.51	16.36	15.72	
min	178.00	195.00	154.00	225.00	208.00	219.00	247.00	252.00	
25%	201.50	224.00	203.00	246.00	244.00	250.00	273.00	291.00	
50%	210.00	240.00	215.00	266.00	256.00	266.00	285.00	303.00	
75%	224.50	254.00	227.00	279.00	266.00	275.25	295.00	311.00	
max	248.00	279.00	251.00	353.00	290.00	301.00	315.00	328.00	

	Angra
count	51.00
mean	177.02
std	15.84
min	143.00
25%	167.50

```
50%    175.00
75%    189.00
max    208.00
```

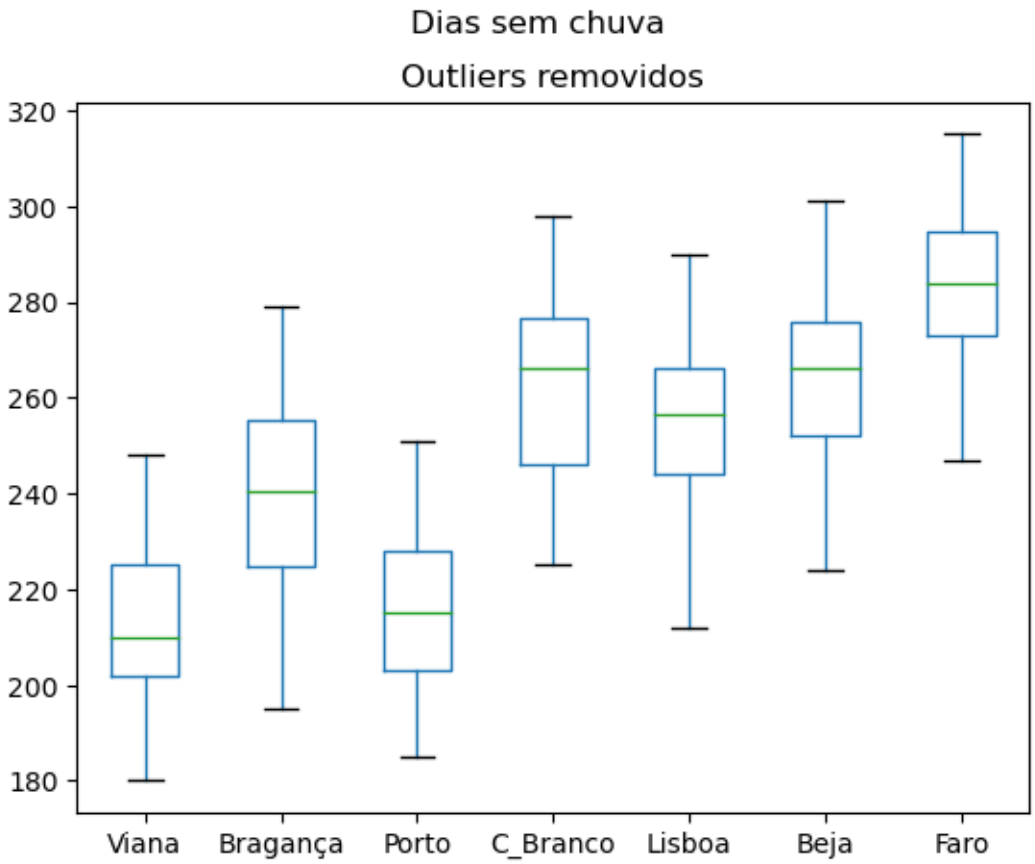
1.1.10 e) Refaça os gráficos restringindo às estações meteorológicas do continente e removendo os outliers. Comente os resultados;

[7]:
```
dfcont = df.iloc[:,0:8] # DataFrame com apenas as estações do continente
# remover outliers
```

Para remover os "outliers" de um DataFrame iremos remover as linhas que contenham "outliers".

[8]:
```
def remover_outliers(dataframe, threshold=1.5):
    Q1 = dataframe.quantile(0.25)
    Q3 = dataframe.quantile(0.75)
    AIQ = Q3 - Q1
    linhas_filtradas = ~((dataframe < (Q1 - threshold * AIQ)) | (dataframe >␣
 ↪(Q3 + threshold * AIQ))).any(axis=1)
    return dataframe[linhas_filtradas]

dfcont_sem_outliers = remover_outliers(dfcont, threshold=1.5)
```

[9]:
```
cols_list = (dfcont_sem_outliers.columns.tolist())[1:8]
boxp3=dfcont_sem_outliers.boxplot(by = None, column = cols_list, grid = False)
plt.suptitle('Dias sem chuva')
plt.title('Outliers removidos')
plt.show()
```



1.1.11 f) Descreva em termos de quartis a distribuição do número de dias sem chuva em Castelo Branco. Repita com os dados relativos ao Porto;

[10]:
```
dfcont_sem_outliers['Viana'].quantile([0,0.25, 0.5, 0.75,1])
```

[10]:
```
0.00    180.0
0.25    202.0
0.50    210.0
0.75    225.0
1.00    248.0
Name: Viana, dtype: float64
```

[11]:
```
dfcont_sem_outliers['Porto'].quantile([0,0.25, 0.5, 0.75,1])
```

[11]:
```
0.00    185.0
0.25    203.0
0.50    215.0
0.75    228.0
1.00    251.0
```

```
Name: Porto, dtype: float64
```

[12]:
```
dfcont_sem_outliers['Porto'].quantile([0,0.25, 0.5, 0.75,1])
```

[12]:
```
0.00    185.0
0.25    203.0
0.50    215.0
0.75    228.0
1.00    251.0
Name: Porto, dtype: float64
```

1.1.12 g) Construa uma tabela de frequências para o número de dias sem chuva no Porto. Repita o exercício com os dados em classes definidas empiricamente e compare com a utilização da regra de Sturges;

[13]:
```
Dias_sem_chuva_Porto = pd.crosstab(df['Porto'],'numero de dias')
Dias_sem_chuva_Porto
```

[13]:

col_0	numero de dias
Porto	
154.0	1
185.0	1
188.0	2
189.0	1
191.0	1
198.0	1
200.0	1
201.0	2
202.0	2
203.0	4
204.0	1
205.0	1
207.0	2
209.0	2
210.0	1
212.0	1
215.0	4
217.0	1
218.0	2
220.0	1
221.0	1
222.0	2
223.0	3
224.0	1
227.0	2
229.0	2
231.0	1

```
232.0              1
237.0              1
240.0              1
241.0              1
242.0              1
245.0              2
249.0              1
250.0              1
251.0              1
```

não aconselhável para dados numéricos... é conveniente personalizar as classes (intervalos). Por exemplo usar os intervalos com extremos nos pontos indicados na lista "Fclasses"

[14]:
```
Fclasses=[150,175,200,225,250,275,300]
# definir classes
Porto_myclasses = pd.cut(dfcont_sem_outliers['Porto'],bins=Fclasses)
cross_tab_result = pd.crosstab(index=Porto_myclasses, columns=['numero de␣
 ↪dias'])
cross_tab_result
```

[14]:
```
col_0       numero de dias
Porto
(175, 200]               6
(200, 225]              31
(225, 250]              13
(250, 275]               1
```

1.1.13 Classes definidas pela regra de Sturges:

Na regra de Sturges as classes, de um conjunto de dados numérico X, têm todas a mesma amplitude e o número de classes (nc) é dado por

$$\text{nc} = \text{int}(1 + \log_2(n)), \text{ onde, n é o tamanho do conjunto X}$$

e a amplitude de cada classe é dada por

$$xc = \frac{\max(X) - \min(X)}{nc}$$

[15]:
```
import numpy as np
import math

def sturge_rule_bins(data: np.array) -> int:
    """ Sturge's rule for optimal bin selection
    Parameters:
        data (np.array) - a one-dimensional array with data
    Returns:
        nbins (int) - number of bins
```

```
    """
    assert data.ndim == 1
    n = data.size
    width = 1.0 + np.log2(n)

    nbins = math.ceil((data.max() - data.min()) / width)
    nbins = max(1, nbins)

    return nbins

# transformar Pandas.Series data em np.array
```

[16]:
```
#Porto_nan_removed=dfcont_sem_outliers['Porto'].dropna() # remover NaN
Porto_nan_removed=(dfcont_sem_outliers['Porto'])[~dfcont_sem_outliers['Porto'].
 ↪isnull()]
Porto1=Porto_nan_removed.to_numpy(dtype=None, copy=False) # converter Pandas.
 ↪series para numpy.narray
print(Porto1)
```

```
[222. 212. 210. 237. 201. 242. 223. 231. 223. 188. 200. 207. 204. 223.
 215. 198. 221. 203. 203. 203. 218. 222. 227. 240. 202. 209. 205. 202.
 209. 232. 220. 191. 215. 203. 207. 249. 250. 229. 251. 217. 189. 215.
 245. 229. 224. 241. 218. 245. 185. 215. 201.]
```

[17]:
```
nc=sturge_rule_bins(Porto1)
Porto_sturges_classes = pd.cut(Porto_nan_removed,bins=10)
sturge_tab_result = pd.crosstab(index=Porto_sturges_classes, columns=['dias sem␣
 ↪chuva'])
sturge_tab_result
```

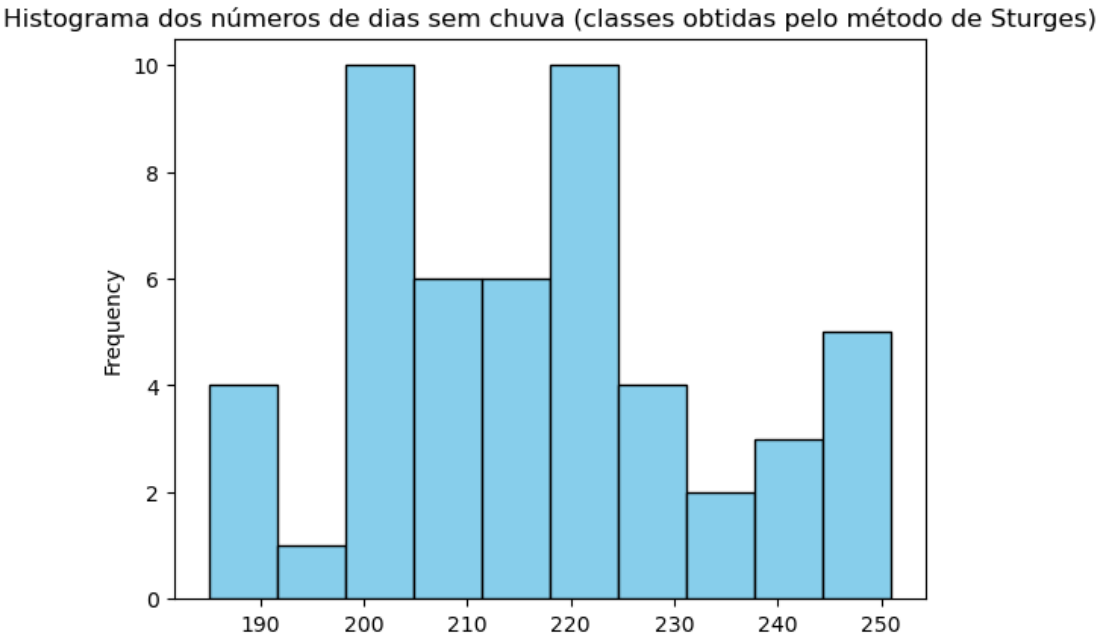
[17]:

col_0	dias sem chuva
Porto	
(184.934, 191.6]	4
(191.6, 198.2]	1
(198.2, 204.8]	10
(204.8, 211.4]	6
(211.4, 218.0]	8
(218.0, 224.6]	8
(224.6, 231.2]	4
(231.2, 237.8]	2
(237.8, 244.4]	3
(244.4, 251.0]	5

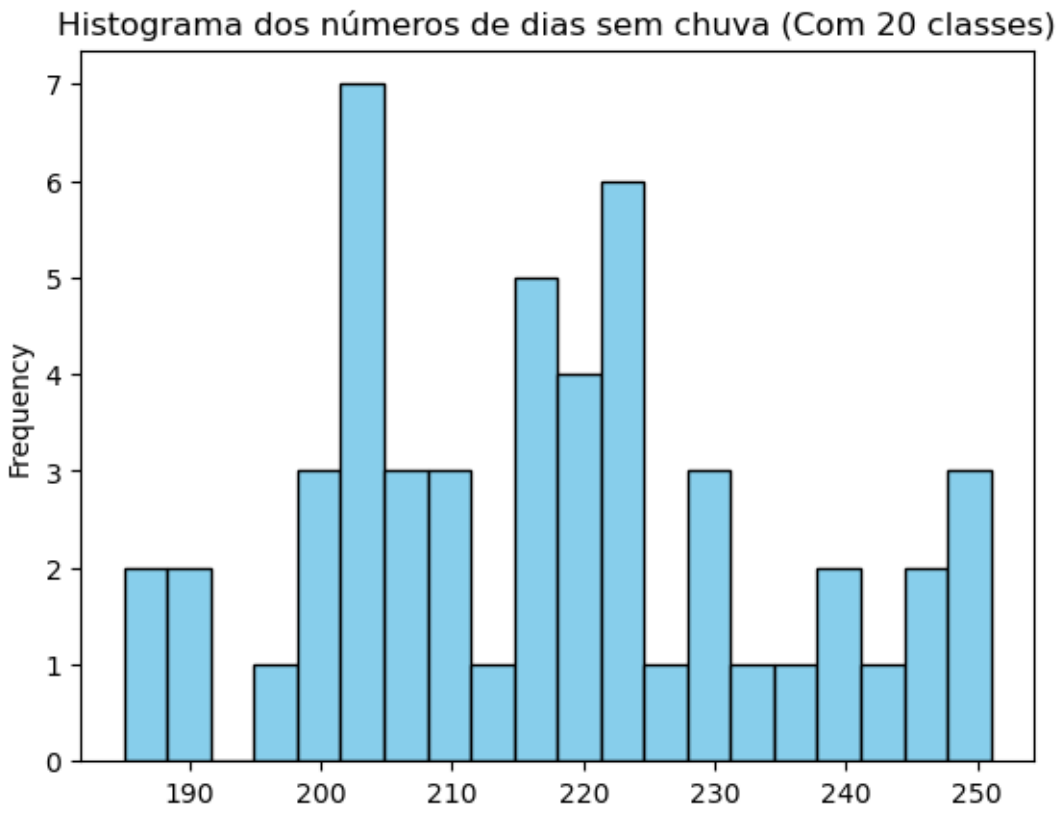
1.1.14 h)

Construa um gráfico que permita visualizar a forma da distribuição do número de dias sem chuva no Porto. O que pode observar no gráfico? Analise a tabela com as frequências observadas no gráfico;

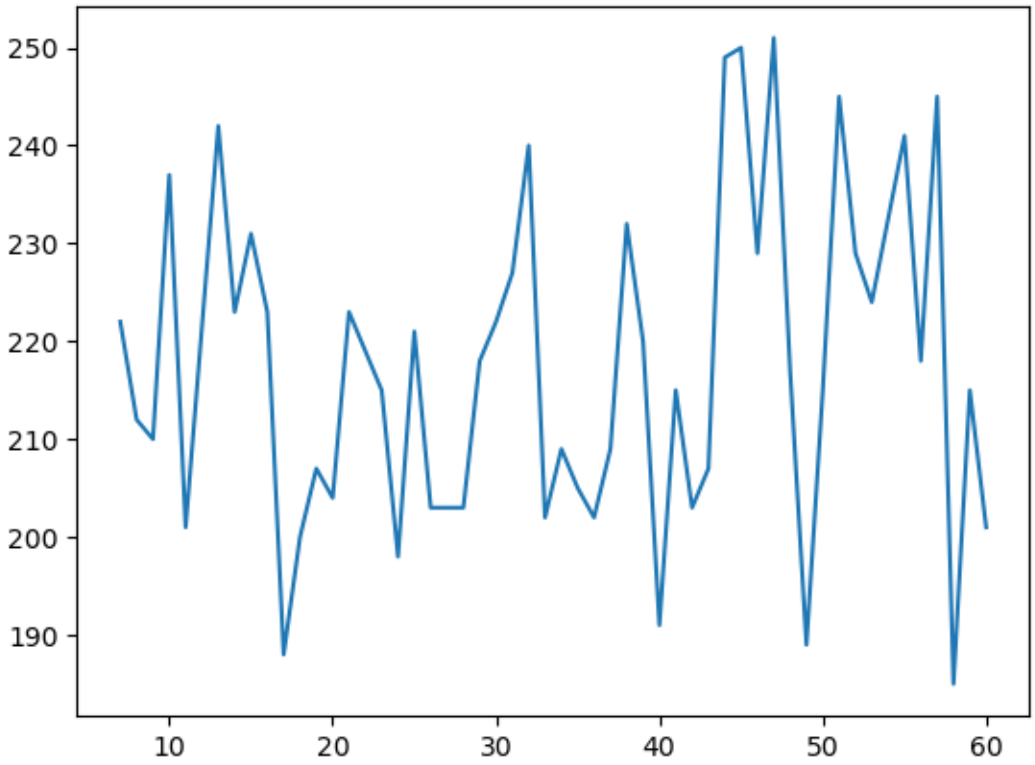
[18]:
```
Porto_nan_removed.plot(kind='hist', bins=nc, edgecolor='black', color='skyblue')
plt.title('Histograma dos números de dias sem chuva (classes obtidas pelo␣
 ↪método de Sturges)')
plt.show()
```



[19]:
```
Porto_nan_removed.plot(kind='hist', bins=20, edgecolor='black', color='skyblue')
plt.title('Histograma dos números de dias sem chuva (Com 20 classes)')
plt.show()
```



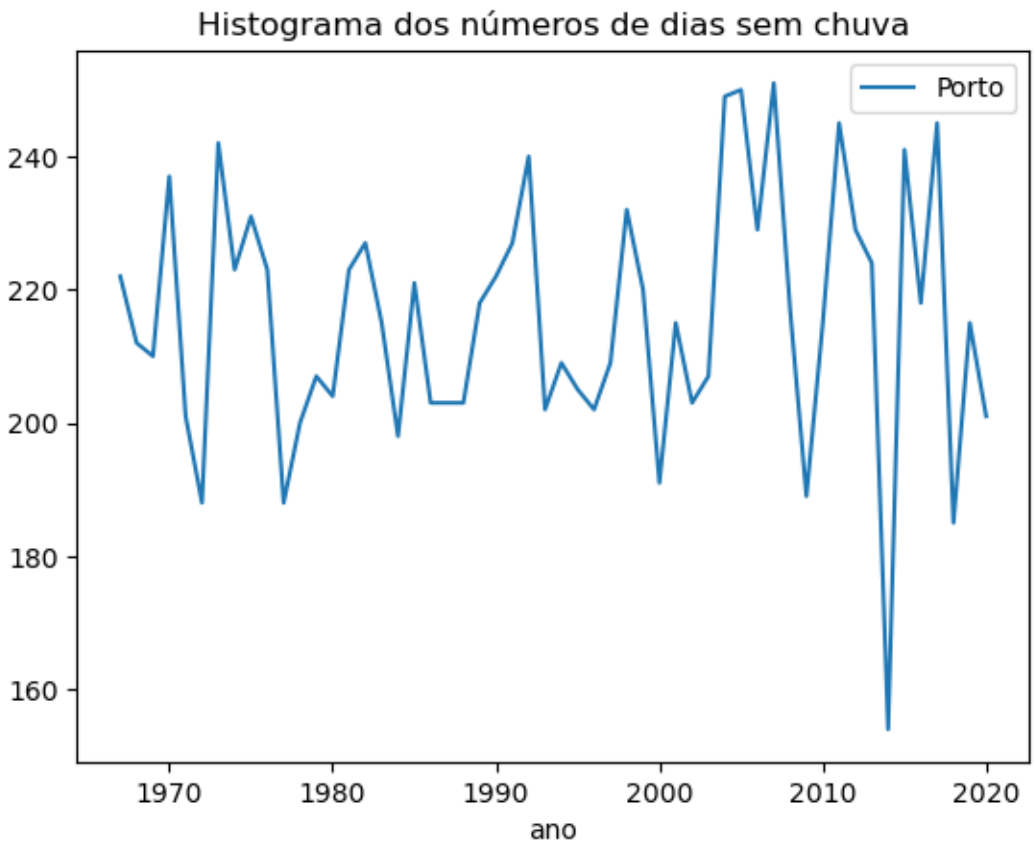
[20]:
```
Porto_nan_removed.plot(kind='line')

plt.show()
```



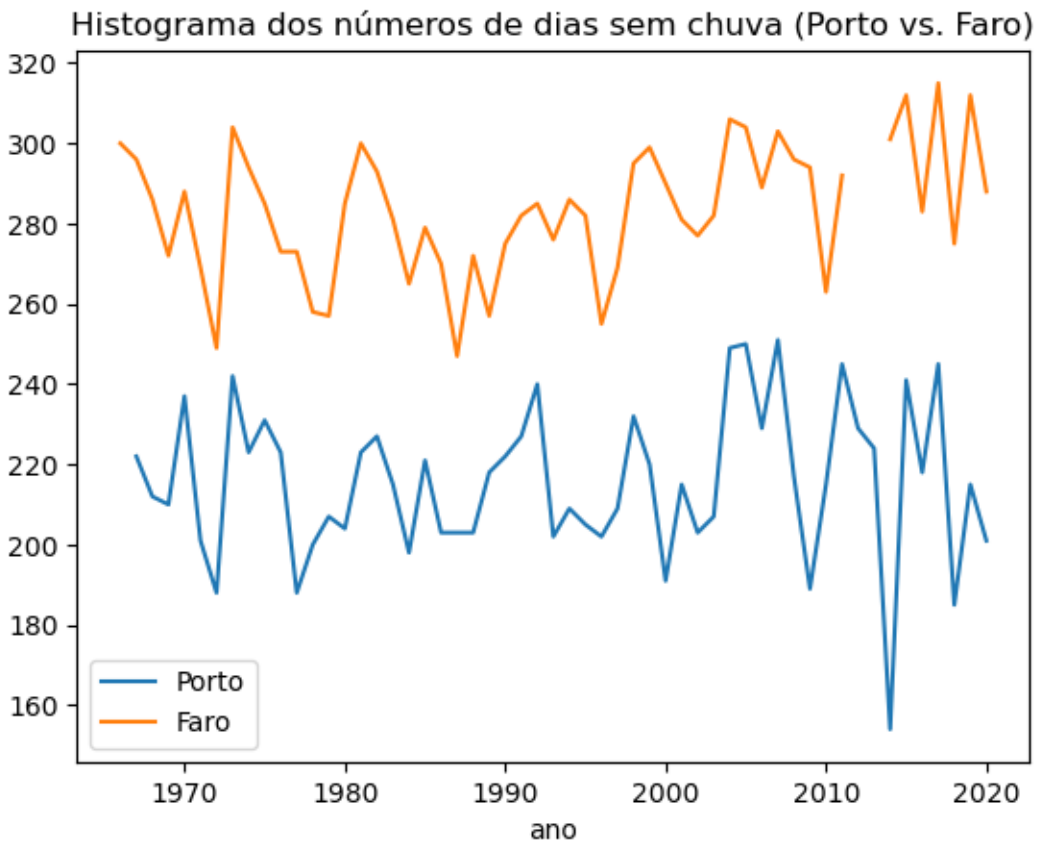
1.1.15 i)

1.1.16 Analise graficamente a evolução temporal do número de dias sem chuva no Porto. O que conclui? Repita a análise para o número de dias sem chuva em Faro.

[26]:
```
df.plot(x='ano',y='Porto',kind='line')
plt.title('Histograma dos números de dias sem chuva')
plt.show()
```



[29]:
```
df.plot(x='ano',y=['Porto','Faro'],kind='line')
plt.title('Histograma dos números de dias sem chuva (Porto vs. Faro)')
plt.show()
```


Histograma dos números de dias sem chuva (Porto vs. Faro)
320
300
280
260
240
220
200
180
160
Porto
Faro
1970
1980
1990
2000
2010
2020
ano

[ ]: