main

November 17, 2022

[1]:
```
import matplotlib.pyplot as plt
from dask import dataframe as dd
from utils import utils
from IPython.display import display_html
import pandas as pd
import numpy as np
import os

PROVAS = ["NU_NOTA_CN","NU_NOTA_CH","NU_NOTA_MT","NU_NOTA_LC","NU_NOTA_REDACAO"]
OUTPUT = os.path.join(os.getcwd(), 'output')
YEARS = ['2019', '2020', '2021']

if not os.path.isdir(OUTPUT):
    os.makedirs(OUTPUT)

df_2019, df_2020, df_2021 = utils.setup()
```

[2]:
```
# DESCRIÇÃO DA MÉDIA GERAL

media_geral_2019 = df_2019["NU_MEDIA_GERAL"].describe(include=['object',␣
 ↪'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_2020 = df_2020["NU_MEDIA_GERAL"].describe(include=['object',␣
 ↪'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_2021 = df_2021["NU_MEDIA_GERAL"].describe(include=['object',␣
 ↪'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_2020, media_geral_2021],
                keys=['Descrição da Média Geral 2019',
                      'Descrição da Média Geral 2020',
                      'Descrição da Média Geral 2021'],
                axis=1)

df
```

[2]:

	Descrição da Média Geral 2019	Descrição da Média Geral 2020	\
count	3701909.00	2561304.00	
mean	522.62	526.60	
std	83.65	91.52	

```
min                                0.00                                0.00
25%                              464.04                              459.38
50%                              515.02                              516.88
75%                              576.74                              587.16
max                              850.82                              858.58

       Descrição da Média Geral 2021
count                     2238106.00
mean                          535.54
std                            88.96
min                             0.00
25%                           471.40
50%                           527.32
75%                           594.48
max                           862.68
```

[3]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE E

media_geral_classe_e_2019 = df_2019.query("Q006 == 'E'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_e_2020 = df_2020.query("Q006 == 'E'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_e_2021 = df_2021.query("Q006 == 'E'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_classe_e_2019,
                media_geral_2020, media_geral_classe_e_2020,
                media_geral_2021, media_geral_classe_e_2021],
                 keys=['Descrição da Média Geral 2019',
                       'Descrição da Média Geral 2019 - Classe E',
                       'Descrição da Média Geral 2020',
                       'Descrição da Média Geral 2020 - Classe E',
                       'Descrição da Média Geral 2021',
                       'Descrição da Média Geral 2021 - Classe E'],
                 axis=1)

df
```

[3]:
```
       Descrição da Média Geral 2019  \
count                     3701909.00
mean                          522.62
std                            83.65
min                             0.00
25%                           464.04
50%                           515.02
```

```
75%                           576.74
max                           850.82

       Descrição da Média Geral 2019 - Classe E  \
count                                1955389.00
mean                                     491.62
std                                       70.27
min                                        0.00
25%                                      445.68
50%                                      487.52
75%                                      534.70
max                                      817.06

       Descrição da Média Geral 2020  \
count                        2561304.00
mean                             526.60
std                               91.52
min                                0.00
25%                              459.38
50%                              516.88
75%                              587.16
max                              858.58

       Descrição da Média Geral 2020 - Classe E  \
count                                1351260.00
mean                                     492.53
std                                       77.15
min                                        0.00
25%                                      439.22
50%                                      484.70
75%                                      539.76
max                                      837.72

       Descrição da Média Geral 2021  Descrição da Média Geral 2021 - Classe E
count                        2238106.00                                1006470.00
mean                             535.54                                   499.00
std                               88.96                                    75.71
min                                0.00                                     0.00
25%                              471.40                                   448.30
50%                              527.32                                   492.14
75%                              594.48                                   544.86
max                              862.68                                   839.82
```

[4]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE D

media_geral_classe_d_2019 = df_2019.query("Q006 == 'D'")["NU_MEDIA_GERAL"].
 ↪describe(
```

```
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_d_2020 = df_2020.query("Q006 == 'D'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_d_2021 = df_2021.query("Q006 == 'D'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_classe_d_2019,
                media_geral_2020, media_geral_classe_d_2020,
                media_geral_2021, media_geral_classe_d_2021],
                 keys=['Descrição da Média Geral 2019',
                       'Descrição da Média Geral 2019 - Classe D',
                       'Descrição da Média Geral 2020',
                       'Descrição da Média Geral 2020 - Classe D',
                       'Descrição da Média Geral 2021',
                       'Descrição da Média Geral 2021 - Classe D'],
                 axis=1)

df
```

[4]:

	Descrição da Média Geral 2019	Descrição da Média Geral 2019 - Classe D	Descrição da Média Geral 2020
count	3701909.00	1109327.00	2561304.00
mean	522.62	535.56	526.60
std	83.65	76.85	91.52
min	0.00	0.00	0.00
25%	464.04	482.58	459.38
50%	515.02	531.98	516.88
75%	576.74	586.30	
max	850.82	826.30	

```
75%                            587.16
max                            858.58

       Descrição da Média Geral 2020 - Classe D  \
count                                802390.00
mean                                    542.73
std                                      84.90
min                                       0.00
25%                                     481.54
50%                                     537.94
75%                                     599.84
max                                     847.82

       Descrição da Média Geral 2021  Descrição da Média Geral 2021 - Classe D
count                        2238106.00                                737361.00
mean                             535.54                                   543.51
std                               88.96                                    81.55
min                                0.00                                    64.00
25%                              471.40                                   485.98
50%                              527.32                                   538.50
75%                              594.48                                   597.30
max                              862.68                                   845.04
```

[5]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE C

media_geral_classe_c_2019 = df_2019.query("Q006 == 'C'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_c_2020 = df_2020.query("Q006 == 'C'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_c_2021 = df_2021.query("Q006 == 'C'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_classe_c_2019,
                media_geral_2020, media_geral_classe_c_2020,
                media_geral_2021, media_geral_classe_c_2021],
                 keys=['Descrição da Média Geral 2019',
                       'Descrição da Média Geral 2019 - Classe C',
                       'Descrição da Média Geral 2020',
                       'Descrição da Média Geral 2020 - Classe C',
                       'Descrição da Média Geral 2021',
                       'Descrição da Média Geral 2021 - Classe C'],
                 axis=1)

df
```

[5]:
```
       Descrição da Média Geral 2019  \
count                     3701909.00
mean                          522.62
std                            83.65
min                             0.00
25%                           464.04
50%                           515.02
75%                           576.74
max                           850.82

       Descrição da Média Geral 2019 - Classe C  \
count                                 460888.00
mean                                     583.91
std                                       80.75
min                                        0.00
25%                                      528.38
50%                                      586.22
75%                                      641.78
max                                      845.00

       Descrição da Média Geral 2020  \
count                     2561304.00
mean                          526.60
std                            91.52
min                             0.00
25%                           459.38
50%                           516.88
75%                           587.16
max                           858.58

       Descrição da Média Geral 2020 - Classe C  \
count                                 291823.00
mean                                     597.33
std                                       87.30
min                                        0.00
25%                                      536.24
50%                                      599.94
75%                                      660.58
max                                      858.58

       Descrição da Média Geral 2021  Descrição da Média Geral 2021 - Classe C
count                     2238106.00                                 358669.00
mean                          535.54                                    588.45
std                            88.96                                     85.47
min                             0.00                                      0.00
25%                           471.40                                    527.66
50%                           527.32                                    588.78
```

```
75%                                 594.48                                 649.60
max                                 862.68                                 862.68
```

[6]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE B

media_geral_classe_b_2019 = df_2019.query("Q006 == 'B'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_b_2020 = df_2020.query("Q006 == 'B'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_b_2021 = df_2021.query("Q006 == 'B'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_classe_b_2019,
                media_geral_2020, media_geral_classe_b_2020,
                media_geral_2021, media_geral_classe_b_2021],
                 keys=['Descrição da Média Geral 2019',
                       'Descrição da Média Geral 2019 - Classe B',
                       'Descrição da Média Geral 2020',
                       'Descrição da Média Geral 2020 - Classe B',
                       'Descrição da Média Geral 2021',
                       'Descrição da Média Geral 2021 - Classe B'],
                 axis=1)

df
```

[6]:
```
       Descrição da Média Geral 2019  \
count                     3701909.00
mean                          522.62
std                            83.65
min                             0.00
25%                           464.04
50%                           515.02
75%                           576.74
max                           850.82

       Descrição da Média Geral 2019 - Classe B  \
count                                 129646.00
mean                                     620.45
std                                       77.41
min                                        0.00
25%                                      571.26
50%                                      626.48
75%                                      676.10
max                                      850.82
```

```
       Descrição da Média Geral 2020  \
count                     2561304.00
mean                          526.60
std                            91.52
min                             0.00
25%                           459.38
50%                           516.88
75%                           587.16
max                           858.58

       Descrição da Média Geral 2020 - Classe B  \
count                                  84249.00
mean                                     630.08
std                                       84.47
min                                      125.08
25%                                      574.94
50%                                      636.28
75%                                      691.28
max                                      852.86

       Descrição da Média Geral 2021  Descrição da Média Geral 2021 - Classe B
count                     2238106.00                                 95692.00
mean                          535.54                                   619.80
std                            88.96                                    84.68
min                             0.00                                     0.00
25%                           471.40                                   562.92
50%                           527.32                                   624.51
75%                           594.48                                   681.34
max                           862.68                                   851.04
```

[7]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE A

media_geral_classe_a_2019 = df_2019.query("Q006 == 'A'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_a_2020 = df_2020.query("Q006 == 'A'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
media_geral_classe_a_2021 = df_2021.query("Q006 == 'A'")["NU_MEDIA_GERAL"].
 ↪describe(
    include=['object', 'float64', 'int64']).apply(lambda s: np.round(s, 2))
df = pd.concat([media_geral_2019, media_geral_classe_a_2019,
                media_geral_2020, media_geral_classe_a_2020,
                media_geral_2021, media_geral_classe_a_2021],
                 keys=['Descrição da Média Geral 2019',
                       'Descrição da Média Geral 2019 - Classe A',
                       'Descrição da Média Geral 2020',
```

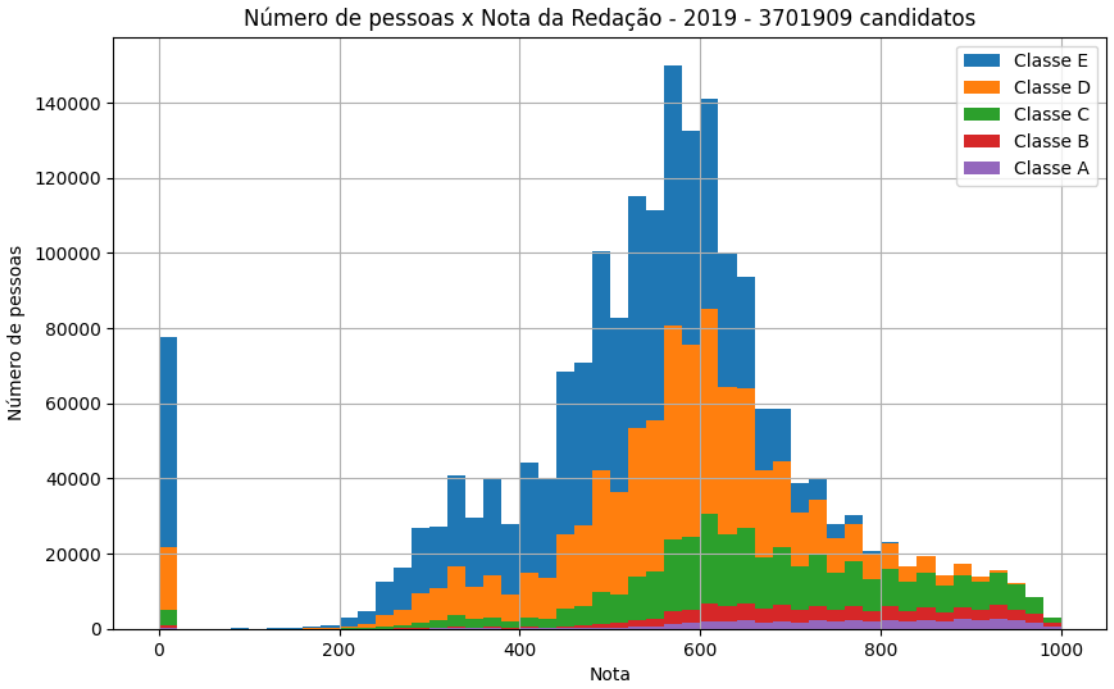
```
                        'Descrição da Média Geral 2020 - Classe A',
                        'Descrição da Média Geral 2021',
                        'Descrição da Média Geral 2021 - Classe A'],
                  axis=1)

df
```

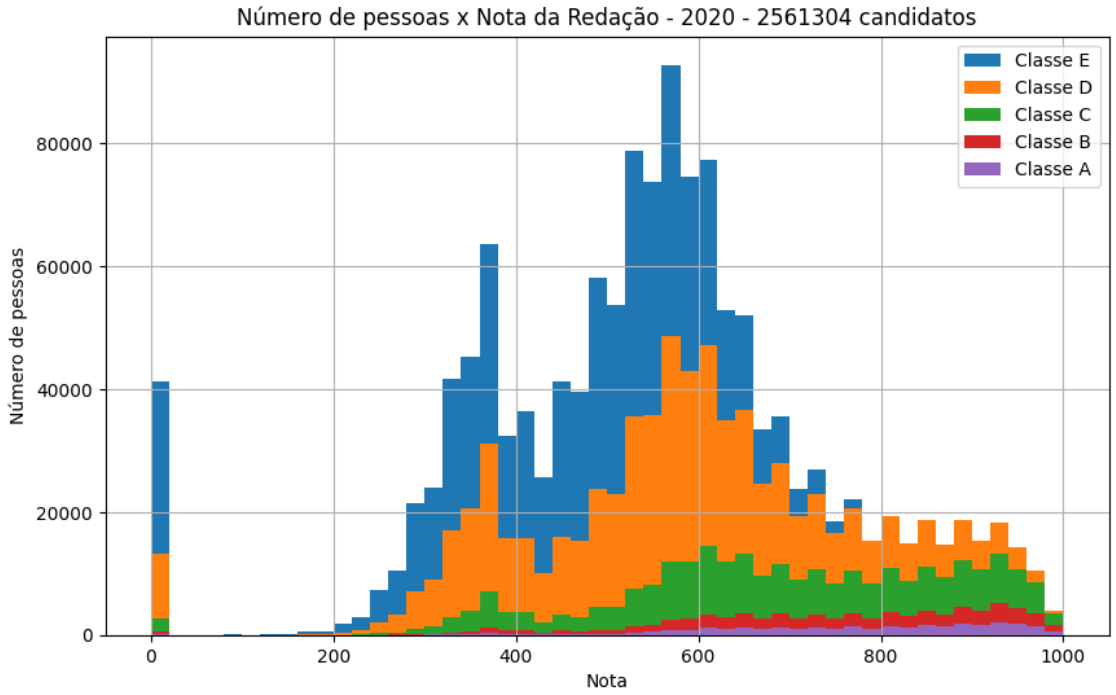
[7]:

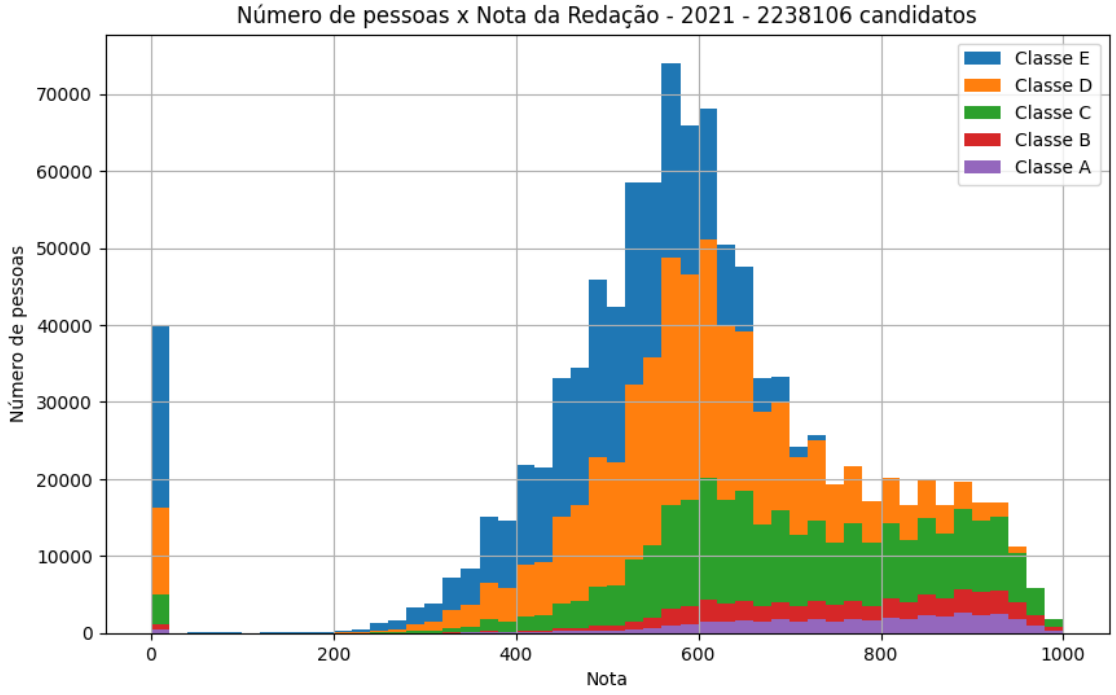
	Descrição da Média Geral 2019	Descrição da Média Geral 2019 - Classe A	Descrição da Média Geral 2020	Descrição da Média Geral 2020 - Classe A
count	3701909.00	46659.00	2561304.00	31582.00
mean	522.62	637.32	526.60	644.74
std	83.65	75.49	91.52	83.97
min	0.00	0.00	0.00	0.00
25%	464.04	592.66	459.38	591.84
50%	515.02	644.88	516.88	652.10
75%	576.74	690.94	587.16	705.76
max	850.82	837.48	858.58	852.76

	Descrição da Média Geral 2021	Descrição da Média Geral 2021 - Classe A
count	2238106.00	39914.00
mean	535.54	632.34
std	88.96	84.49
min	0.00	73.16
25%	471.40	576.34
50%	527.32	638.49
75%	594.48	693.86
max	862.68	859.96

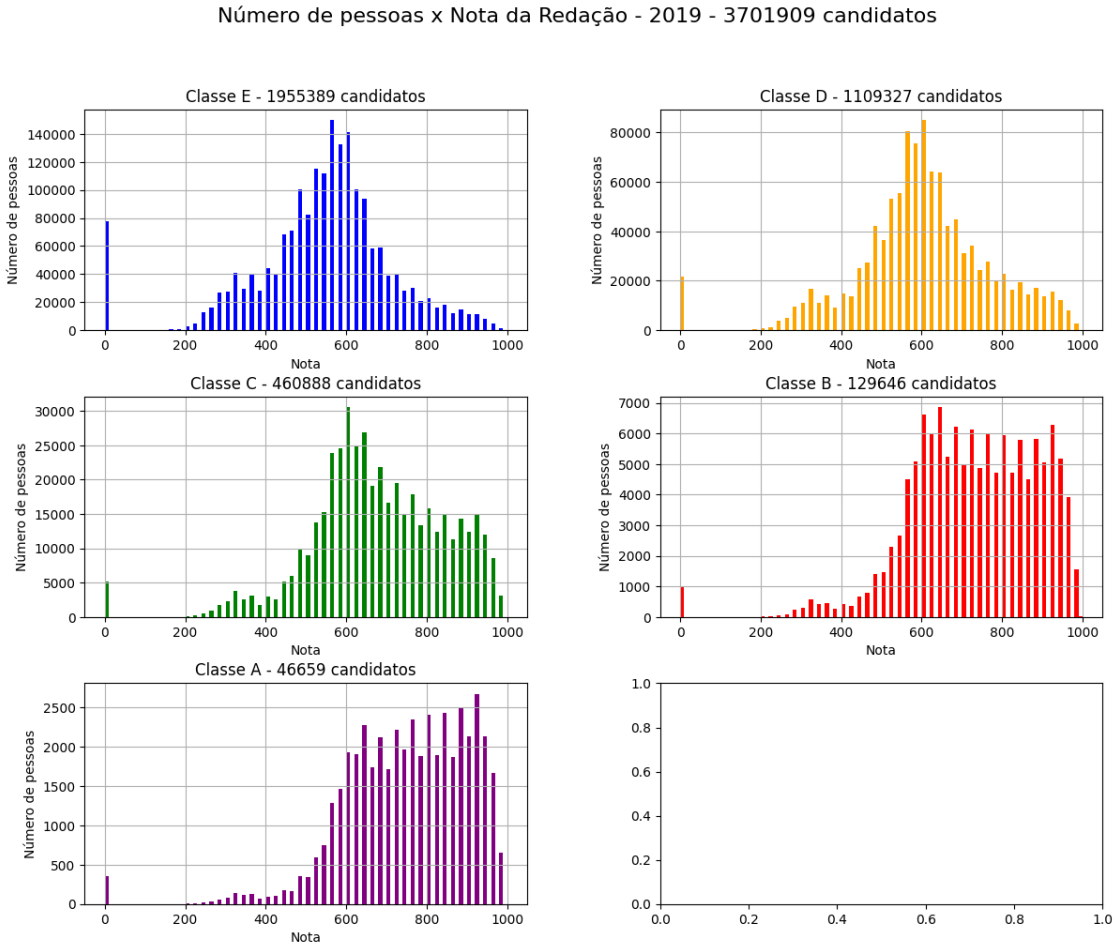
[8]:
```
utils.plot_hist_classes_all(df_2019, OUTPUT)
utils.plot_hist_classes_all(df_2020, OUTPUT)
utils.plot_hist_classes_all(df_2021, OUTPUT)
```

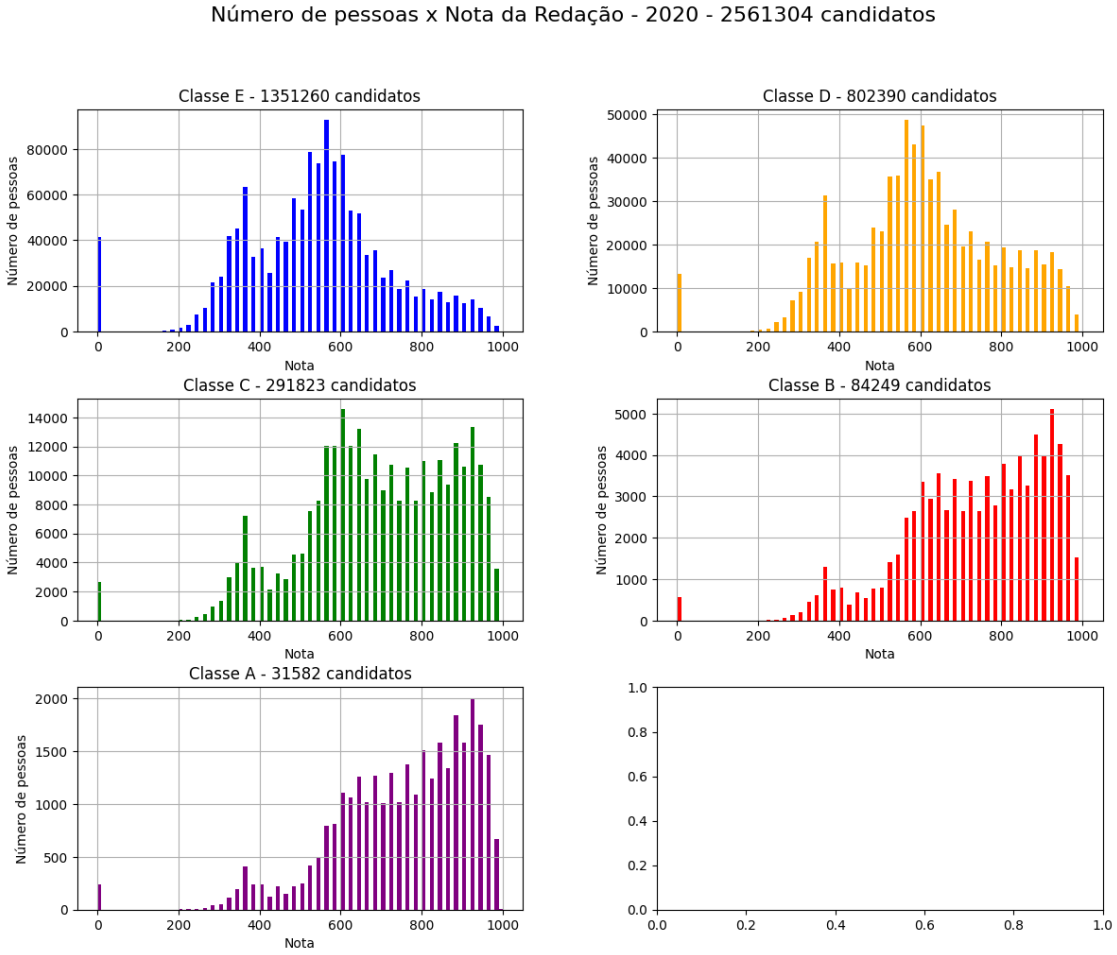


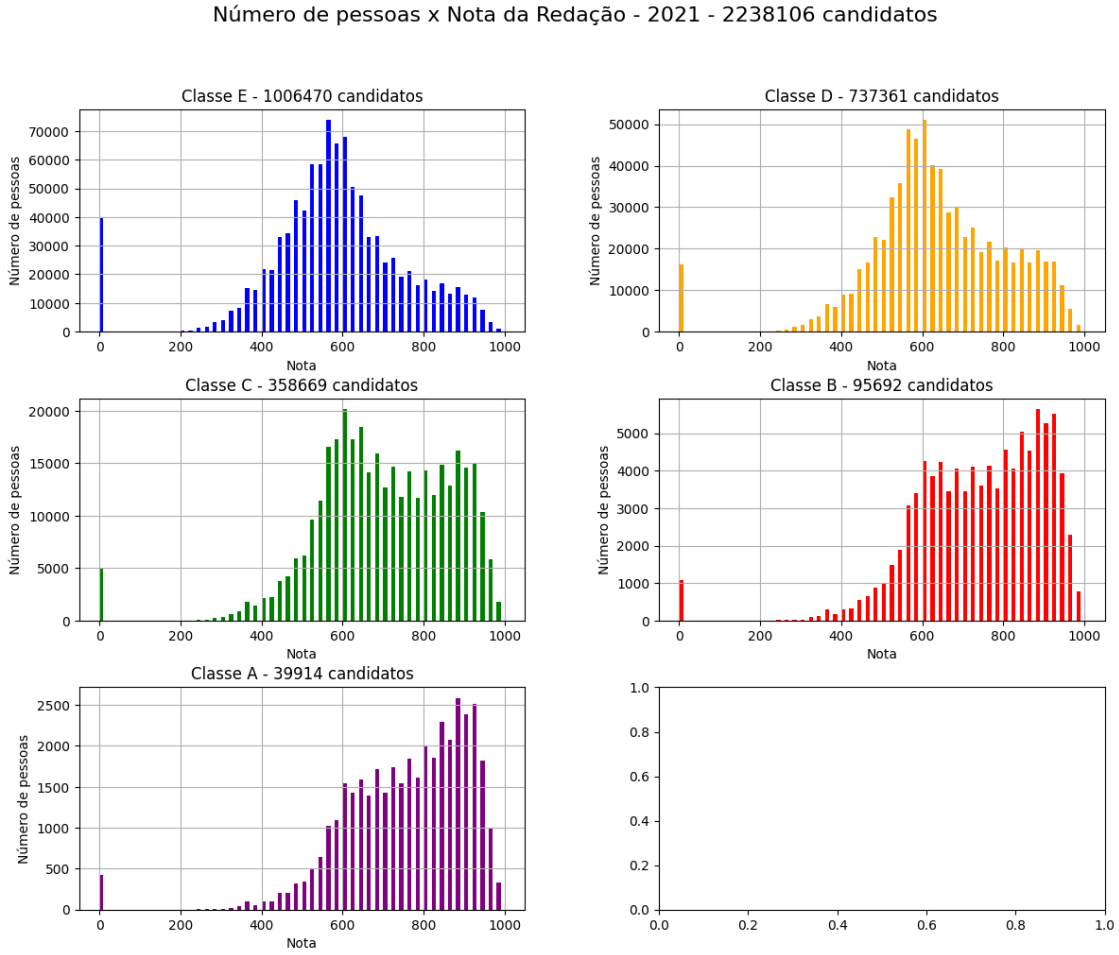

Número de pessoas x Nota da Redação - 2020 - 2561304 candidatos
Classe E
Classe D
Classe C
Classe B
Classe A
Número de pessoas
0
20000
40000
60000
80000
0
200
400
600
800
1000
Nota


Número de pessoas x Nota da Redação - 2021 - 2238106 candidatos
Classe E
Classe D
Classe C
Classe B
Classe A
Número de pessoas
0
10000
20000
30000
40000
50000
60000
70000
0
200
400
600
800
1000
Nota

[9]:
```
utils.plot_hist_classes(df_2019, OUTPUT)
utils.plot_hist_classes(df_2020, OUTPUT)
utils.plot_hist_classes(df_2021, OUTPUT)
```

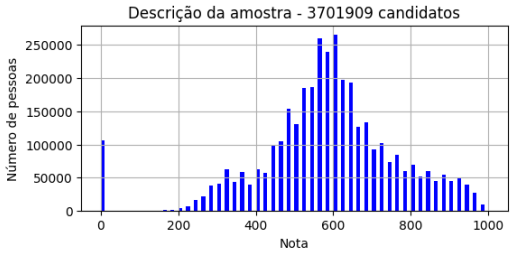


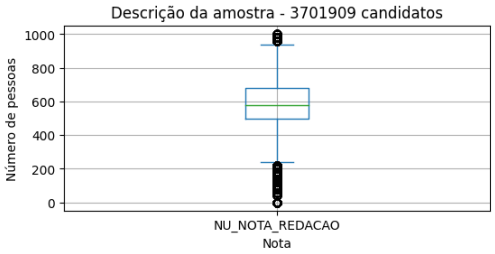

Número de pessoas x Nota da Redação - 2020 - 2561304 candidatos
Classe E - 1351260 candidatos
Classe D - 802390 candidatos
Classe C - 291823 candidatos
Classe B - 84249 candidatos
Classe A - 31582 candidatos
Número de pessoas
Nota

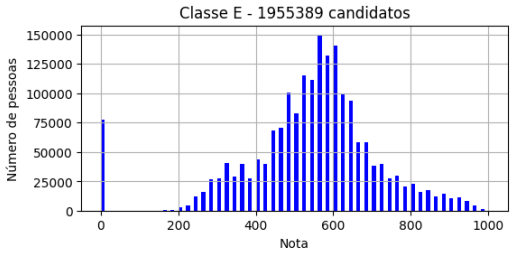


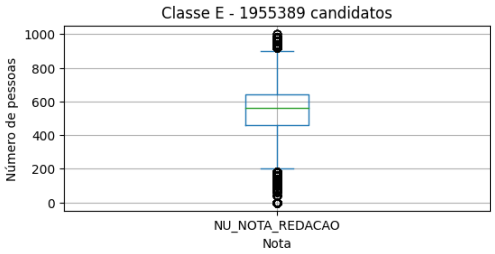
[10]:
```
utils.plot_hist_boxplot_classes(df_2019, OUTPUT)
utils.plot_hist_boxplot_classes(df_2020, OUTPUT)
utils.plot_hist_boxplot_classes(df_2021, OUTPUT)
```

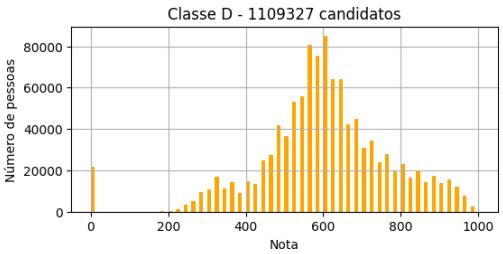
Número de pessoas x Nota da Redação - 2019 - 3701909 candidatos

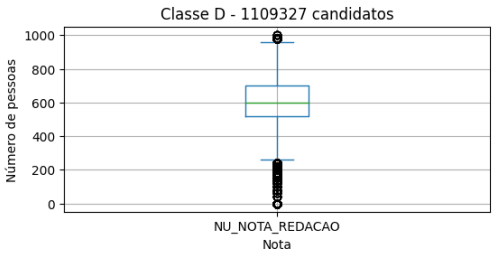


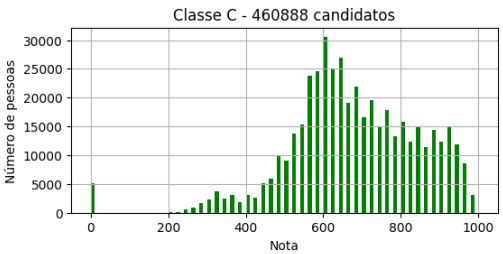


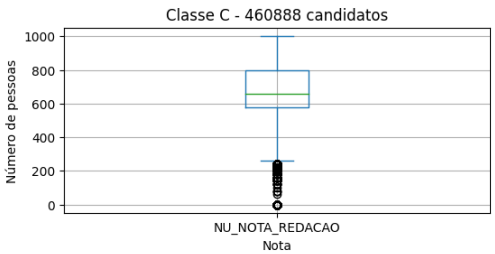


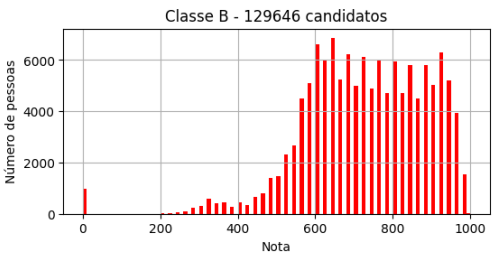


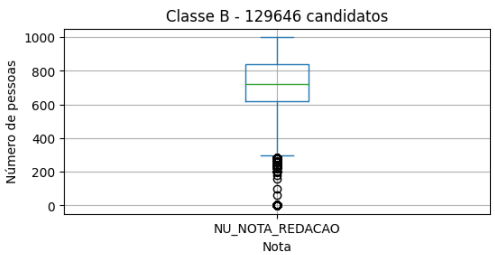


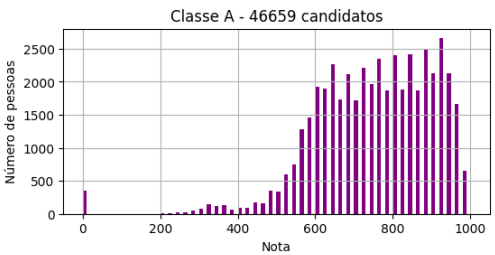


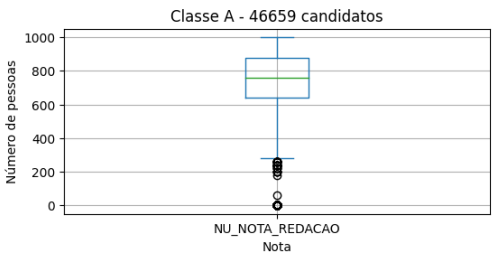




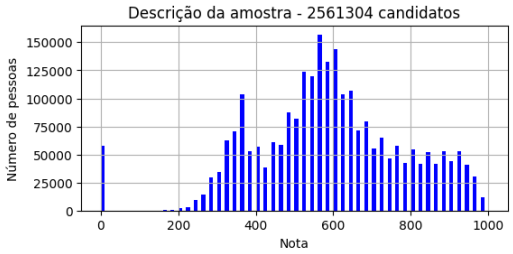


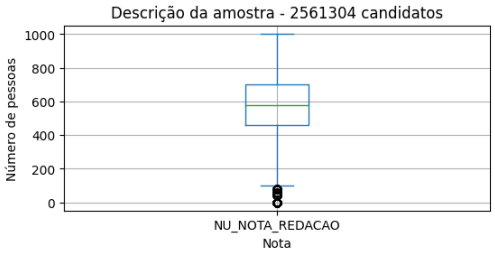


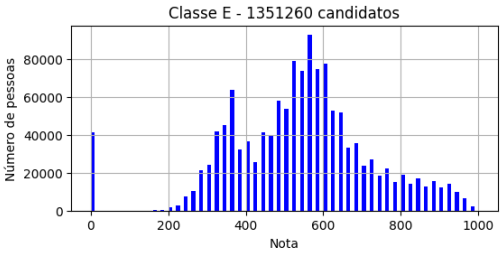


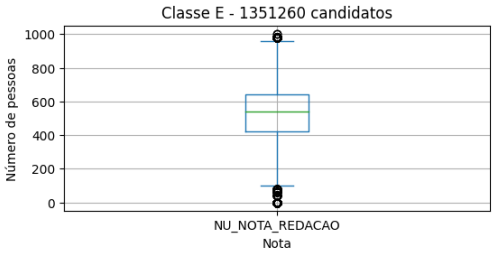


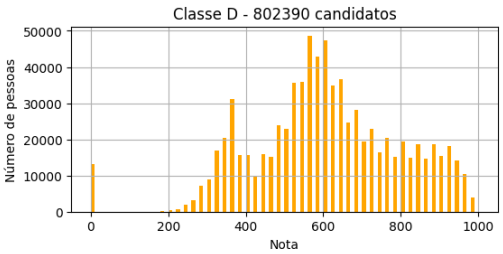
Número de pessoas x Nota da Redação - 2020 - 2561304 candidatos

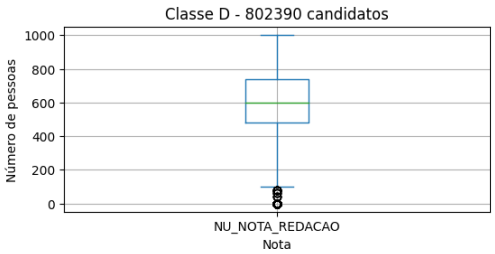


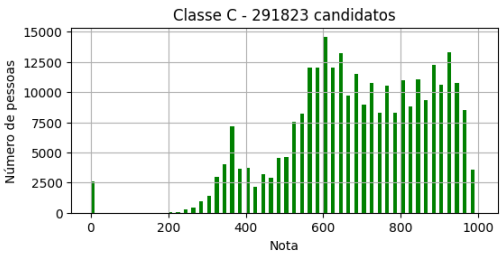


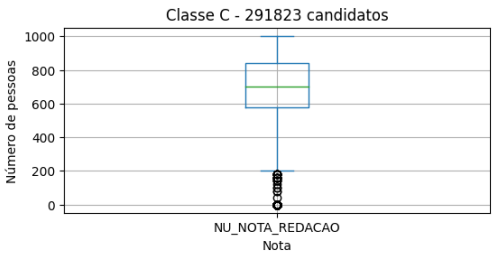


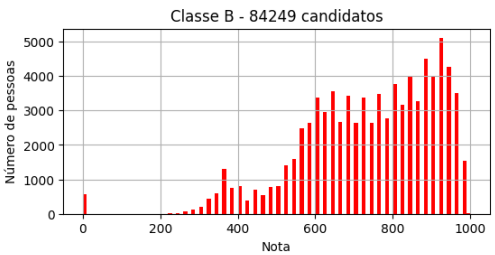


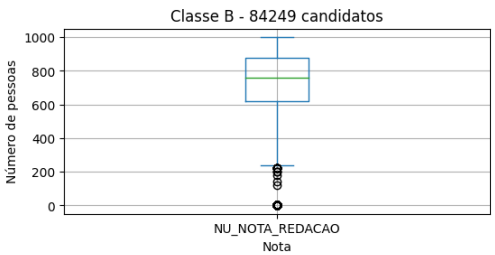


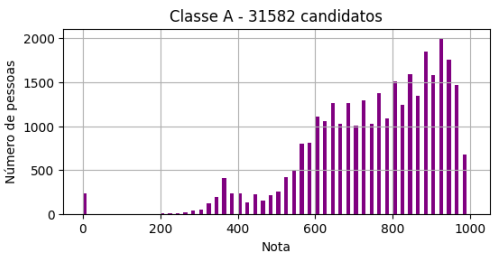


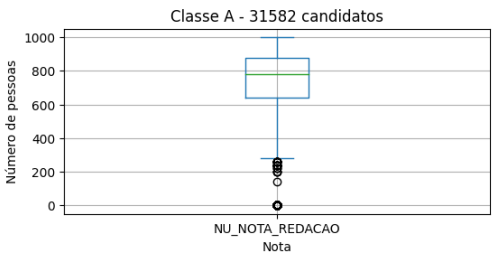




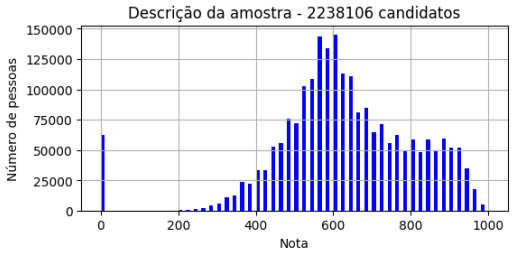


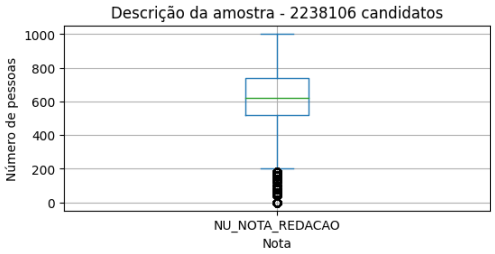


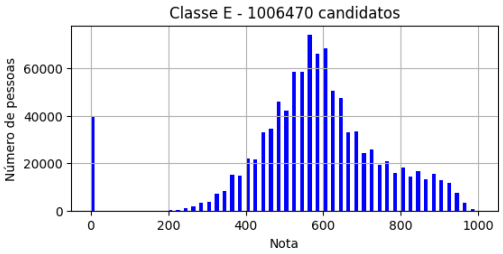


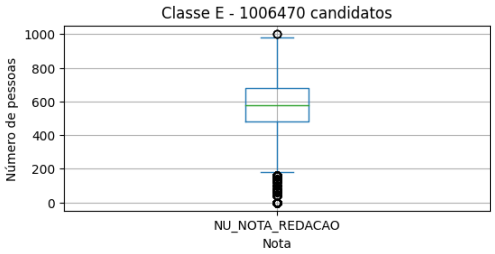


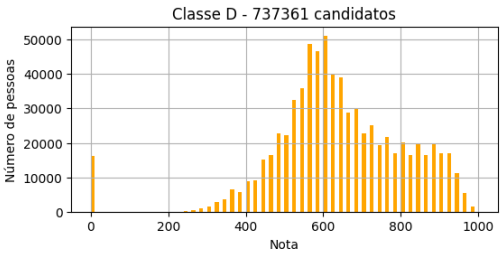
Número de pessoas x Nota da Redação - 2021 - 2238106 candidatos

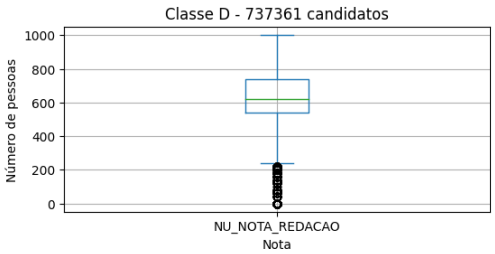


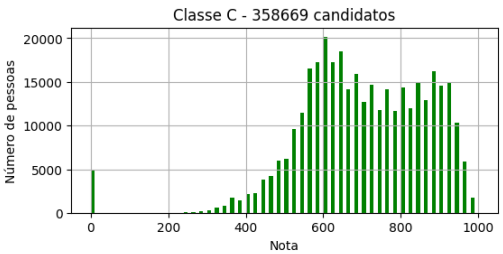


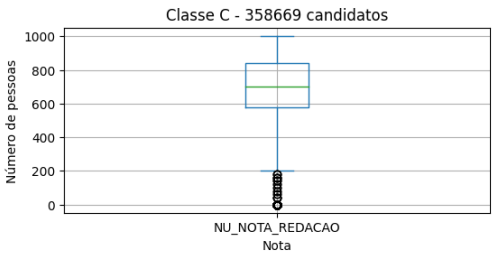


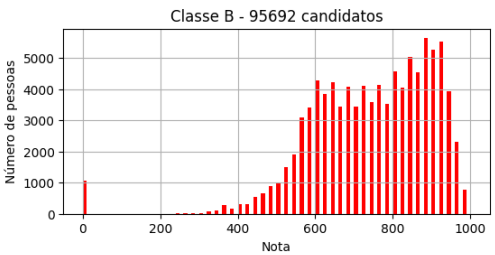


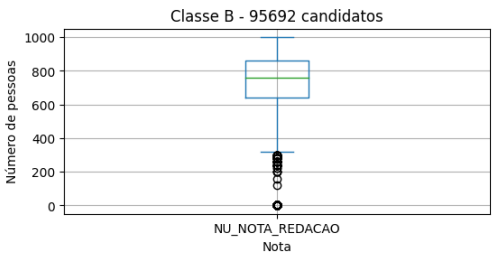


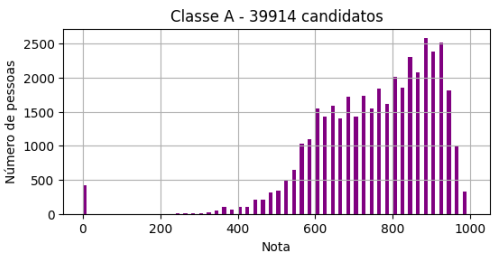


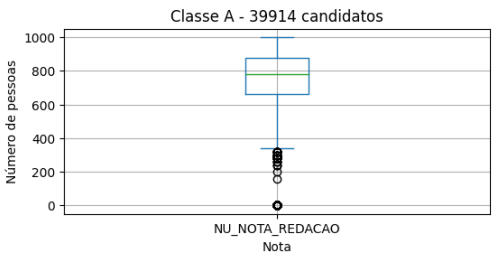




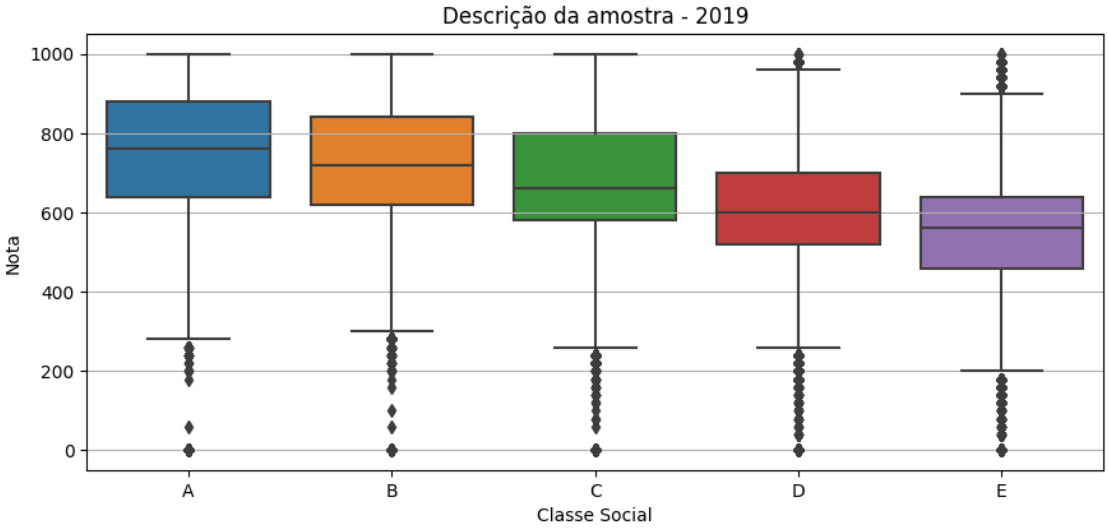


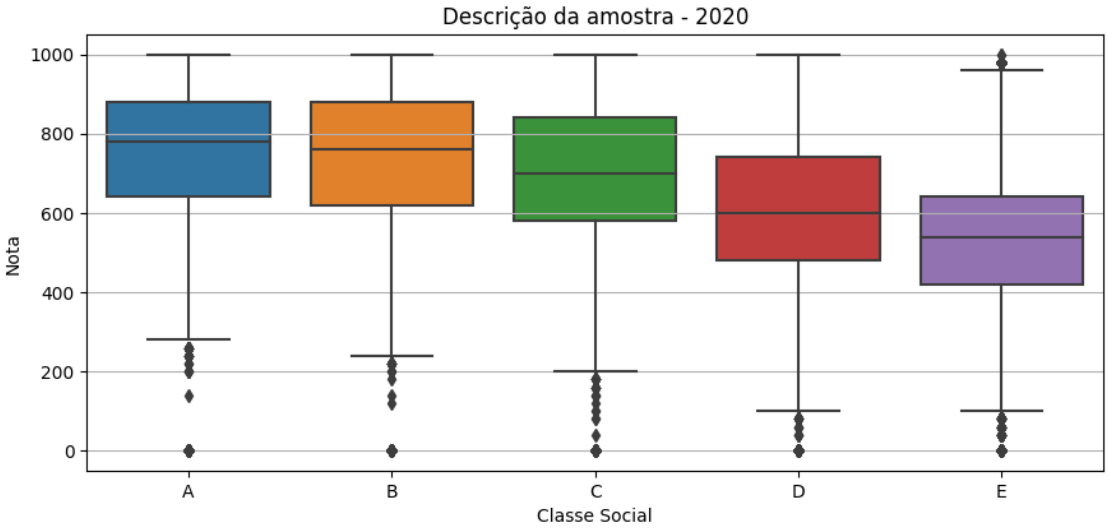


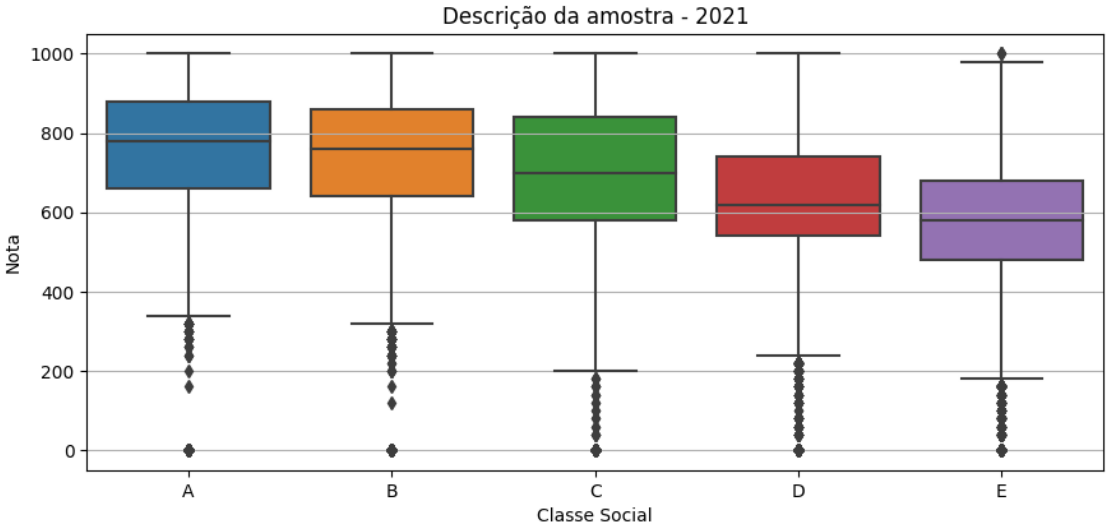




[11]:
```
utils.plot_boxplot_descricao_amostras([df_2019, df_2020, df_2021], OUTPUT)
```


Descrição da amostra - 2019
Nota
1000
800
600
400
200
0
A
B
C
D
E
Classe Social


Descrição da amostra - 2020
Nota
1000
800
600
400
200
0
A
B
C
D
E
Classe Social


Descrição da amostra - 2021
Nota
1000
800
600
400
200
0
A
B
C
D
E
Classe Social