In [ ]:
```
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()
```

In [ ]:
```
# DESCRIÇÃO DA MÉDIA GERAL

media_geral_2019 = df_2019["NU_MEDIA_GERAL"].describe(include=['object', 'float64', 'int64']).apply(lambda s: np
media_geral_2020 = df_2020["NU_MEDIA_GERAL"].describe(include=['object', 'float64', 'int64']).apply(lambda s: np
media_geral_2021 = df_2021["NU_MEDIA_GERAL"].describe(include=['object', 'float64', 'int64']).apply(lambda s: np
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
```

Out[ ]:

	Descrição da Média Geral 2019	Descrição da Média Geral 2020	Descrição da Média Geral 2021
count	3701909.00	2561304.00	2238106.00
mean	522.62	526.60	535.54
std	83.65	91.52	88.96
min	0.00	0.00	0.00
25%	464.04	459.38	471.40
50%	515.02	516.88	527.32
75%	576.74	587.16	594.48
max	850.82	858.58	862.68

In [ ]:
```
# 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
```

Out[ ]:

	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
count	3701909.00	1955389.00	2561304.00	1351260.00	2238106.00	1006470.00
mean	522.62	491.62	526.60	492.53	535.54	499.00
std	83.65	70.27	91.52	77.15	88.96	75.71
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	464.04	445.68	459.38	439.22	471.40	448.30
50%	515.02	487.52	516.88	484.70	527.32	492.14
75%	576.74	534.70	587.16	539.76	594.48	544.86
max	850.82	817.06	858.58	837.72	862.68	839.82

In [ ]:
```
# 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
```

Out[ ]:

	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
count	3701909.00	1109327.00	2561304.00	802390.00	2238106.00	737361.00
mean	522.62	535.56	526.60	542.73	535.54	543.51
std	83.65	76.85	91.52	84.90	88.96	81.55
min	0.00	0.00	0.00	0.00	0.00	64.00
25%	464.04	482.58	459.38	481.54	471.40	485.98
50%	515.02	531.98	516.88	537.94	527.32	538.50
75%	576.74	586.30	587.16	599.84	594.48	597.30
max	850.82	826.30	858.58	847.82	862.68	845.04

In [ ]:
```
# 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
```

Out[ ]:

	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
count	3701909.00	460888.00	2561304.00	291823.00	2238106.00	358669.00
mean	522.62	583.91	526.60	597.33	535.54	588.45
std	83.65	80.75	91.52	87.30	88.96	85.47
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	464.04	528.38	459.38	536.24	471.40	527.66
50%	515.02	586.22	516.88	599.94	527.32	588.78
75%	576.74	641.78	587.16	660.58	594.48	649.60
max	850.82	845.00	858.58	858.58	862.68	862.68

In [ ]:
```
# 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
```

Out[ ]:

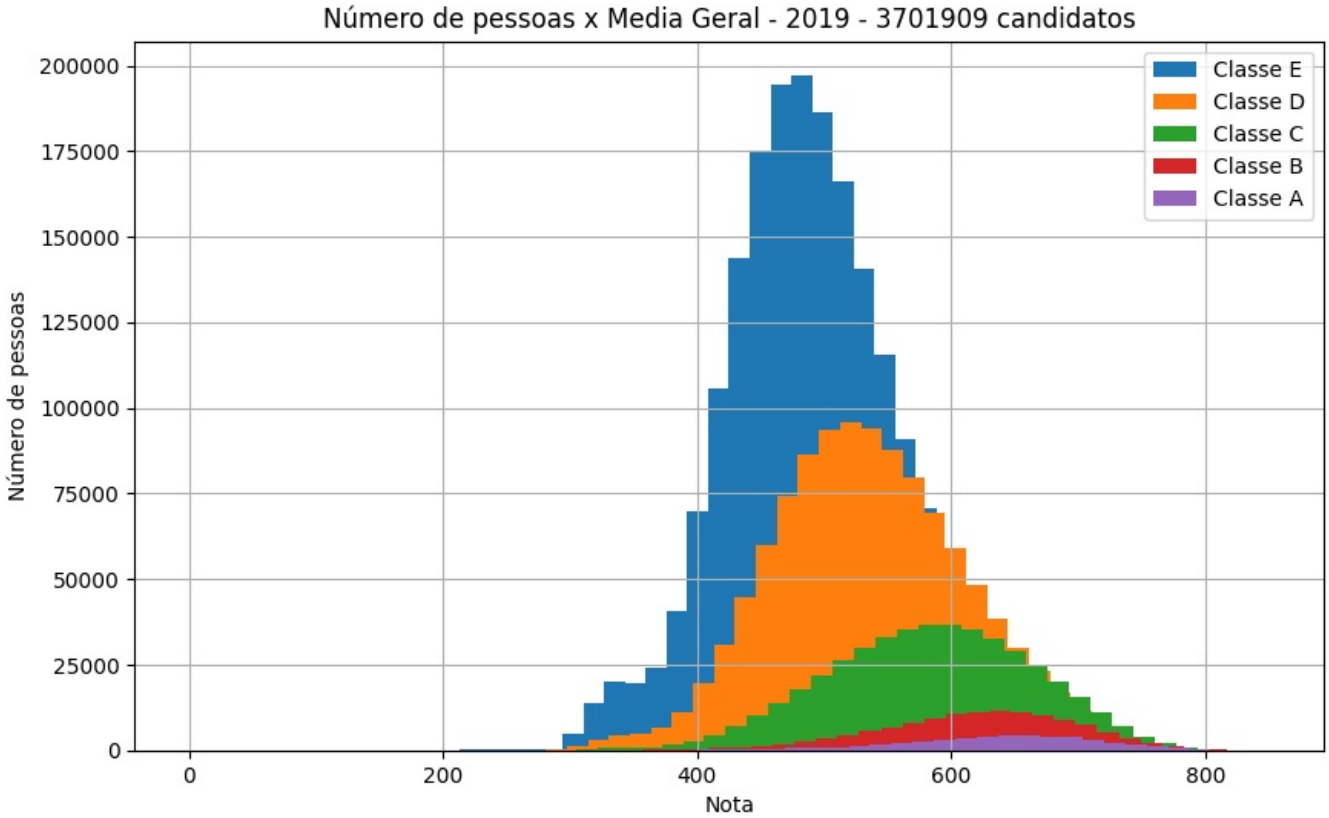
	Descrição da Média Geral 2019	Descrição da Média Geral 2019 - Classe B	Descrição da Média Geral 2020	Descrição da Média Geral 2020 - Classe B	Descrição da Média Geral 2021	Descrição da Média Geral 2021 - Classe B
count	3701909.00	129646.00	2561304.00	84249.00	2238106.00	95692.00
mean	522.62	620.45	526.60	630.08	535.54	619.80
std	83.65	77.41	91.52	84.47	88.96	84.68
min	0.00	0.00	0.00	125.08	0.00	0.00
25%	464.04	571.26	459.38	574.94	471.40	562.92
50%	515.02	626.48	516.88	636.28	527.32	624.51
75%	576.74	676.10	587.16	691.28	594.48	681.34
max	850.82	850.82	858.58	852.86	862.68	851.04

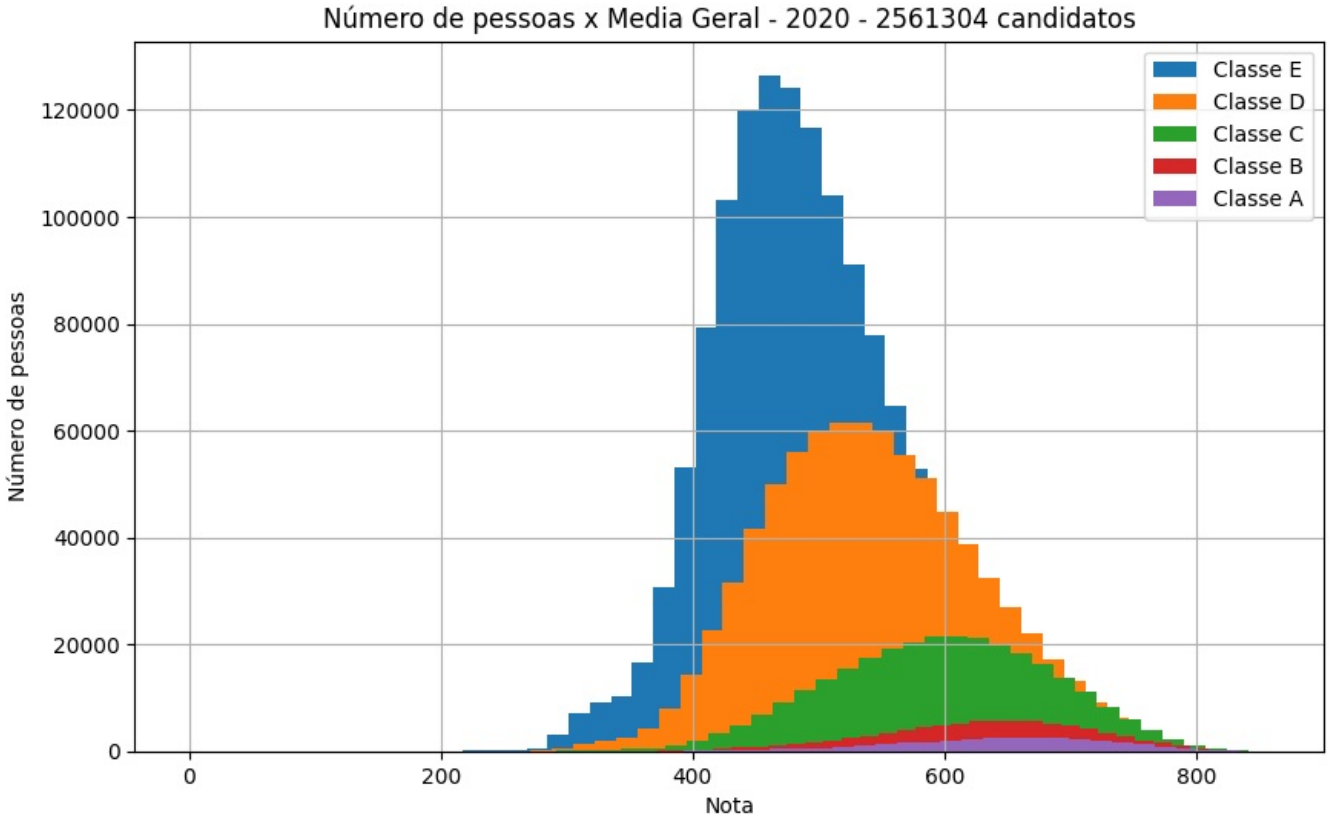
In [ ]:
```
# DESCRIÇÃO DA MÉDIA GERAL CLASSE A

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
```

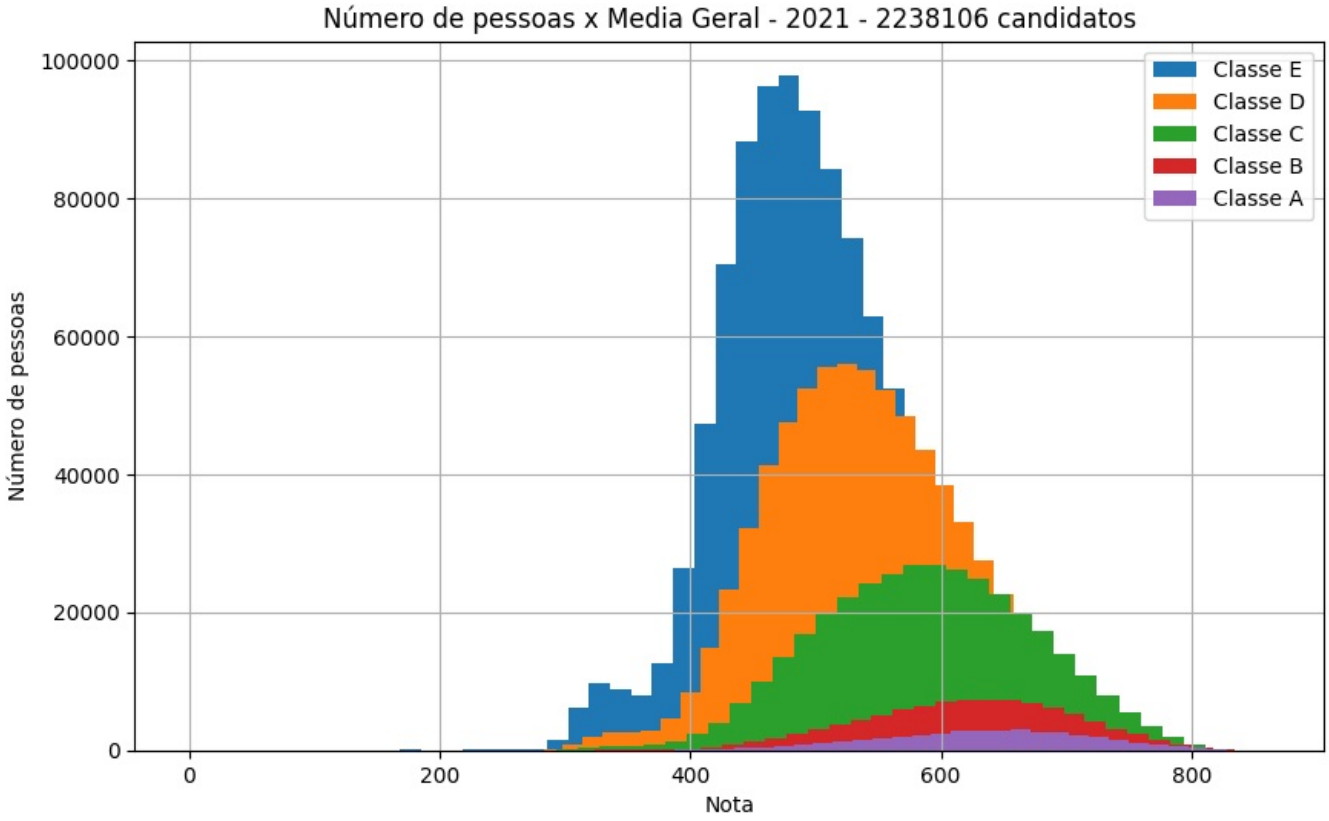
Out[ ]:

	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
count	3701909.00	46659.00	2561304.00	31582.00	2238106.00	39914.00
mean	522.62	637.32	526.60	644.74	535.54	632.34
std	83.65	75.49	91.52	83.97	88.96	84.49
min	0.00	0.00	0.00	0.00	0.00	73.16
25%	464.04	592.66	459.38	591.84	471.40	576.34
50%	515.02	644.88	516.88	652.10	527.32	638.49
75%	576.74	690.94	587.16	705.76	594.48	693.86
max	850.82	837.48	858.58	852.76	862.68	859.96

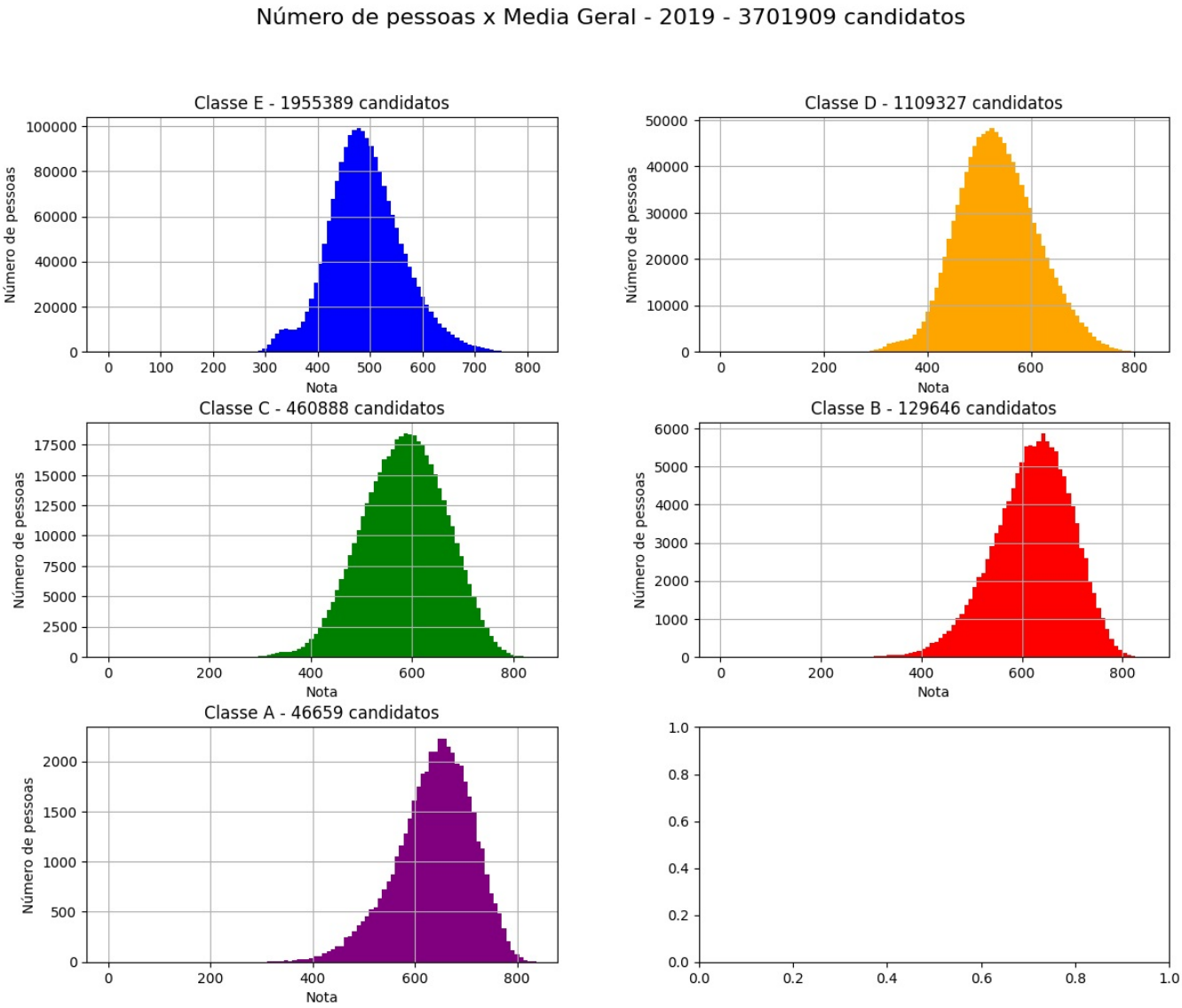
In [ ]:
```
utils.plot_hist_classes_all(df_2019, OUTPUT)
utils.plot_hist_classes_all(df_2020, OUTPUT)
utils.plot_hist_classes_all(df_2021, OUTPUT)
```


Número de pessoas x Media Geral - 2019 - 3701909 candidatos
Classe E
Classe D
Classe C
Classe B
Classe A
Número de pessoas
Nota
0
25000
50000
75000
100000
125000
150000
175000
200000
0
200
400
600
800

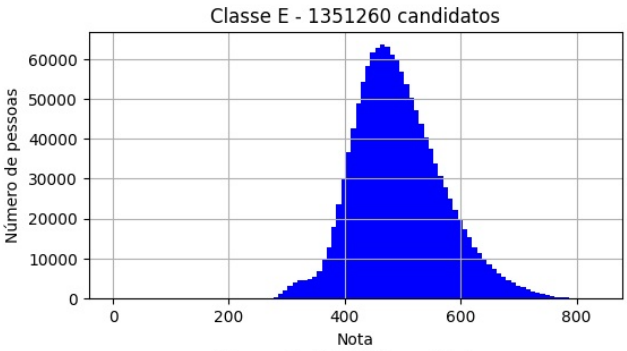

Número de pessoas x Media Geral - 2020 - 2561304 candidatos
Classe E
Classe D
Classe C
Classe B
Classe A
Número de pessoas
Nota
0
20000
40000
60000
80000
100000
120000
0
200
400
600
800

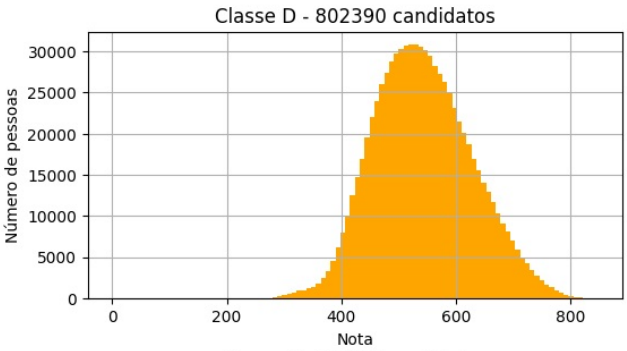


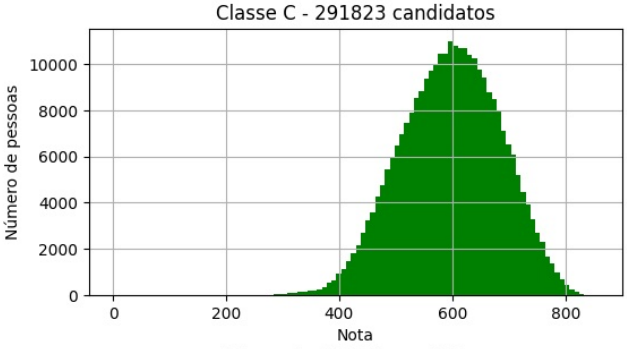
In [ ]:
```
utils.plot_hist_classes(df_2019, OUTPUT)
utils.plot_hist_classes(df_2020, OUTPUT)
utils.plot_hist_classes(df_2021, OUTPUT)
```

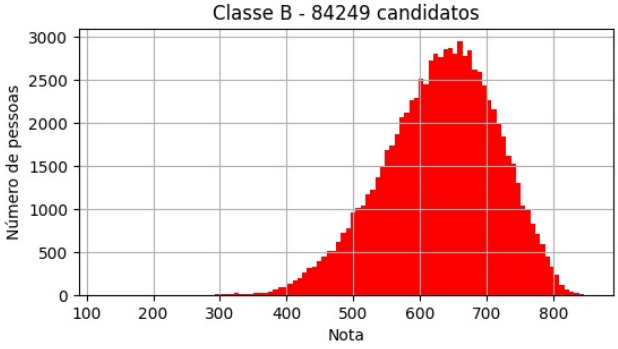


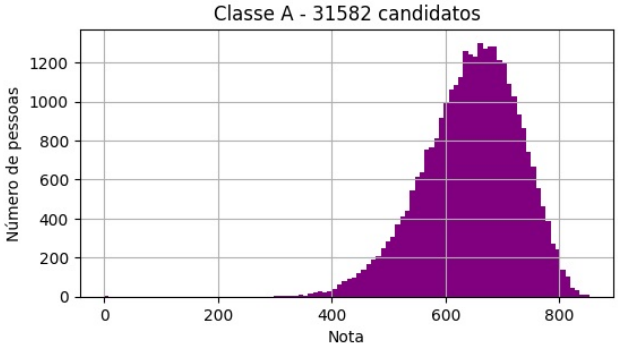
Número de pessoas x Media Geral - 2020 - 2561304 candidatos

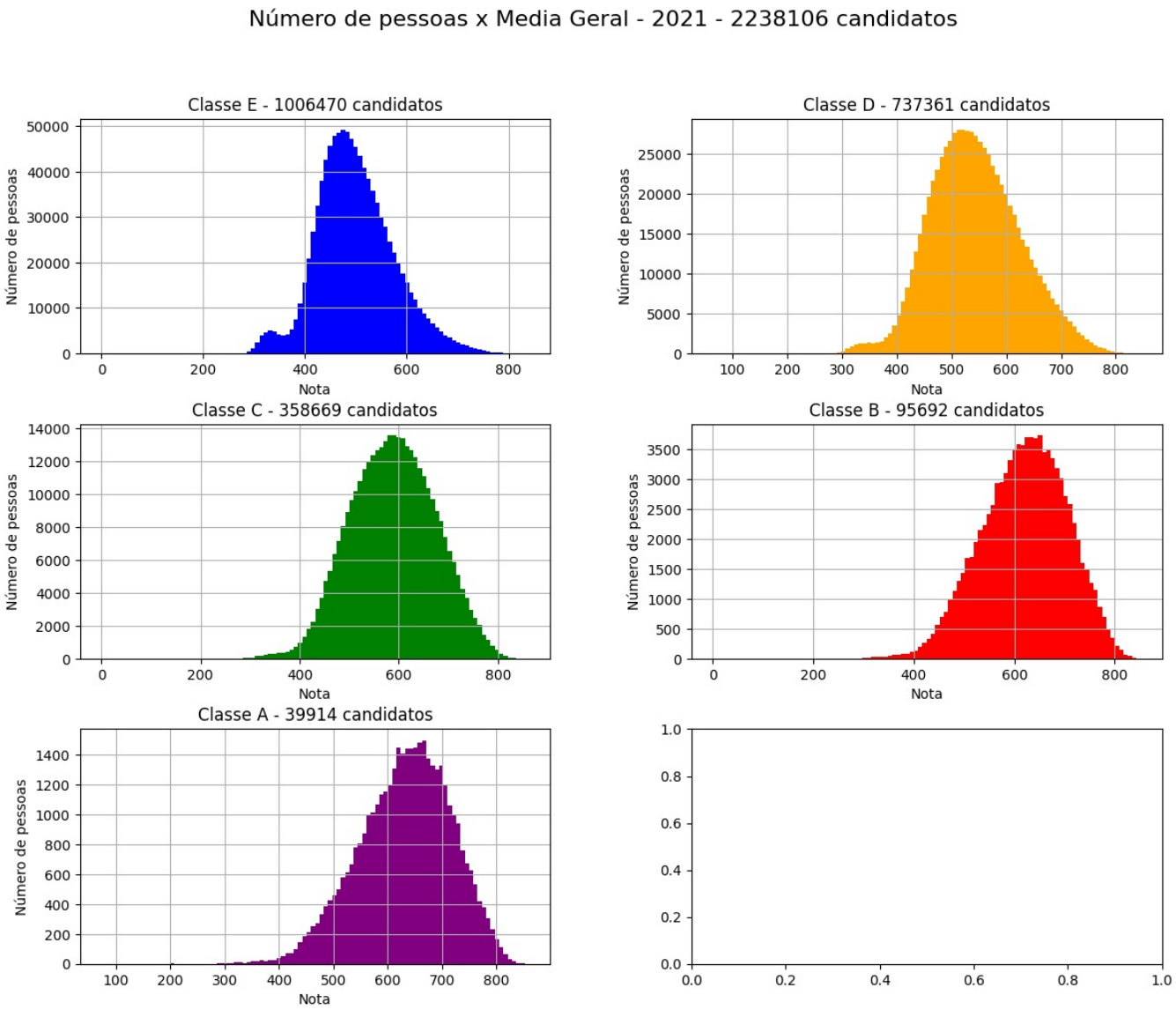








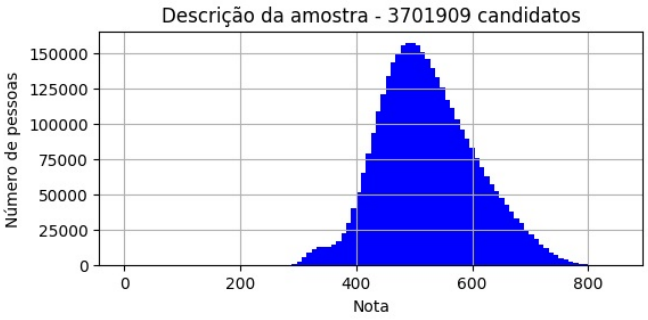


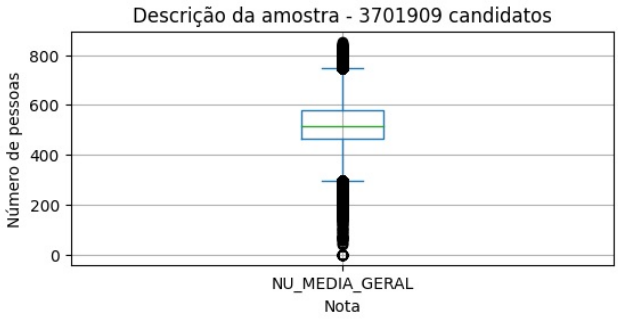


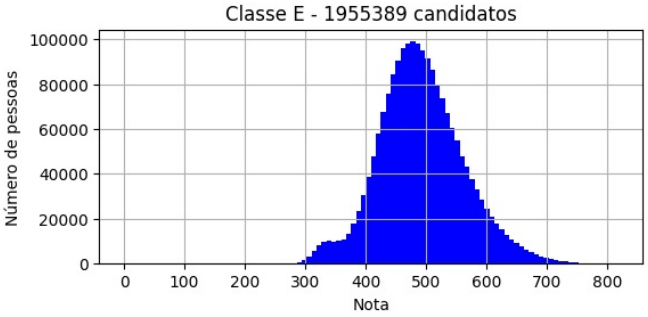
In [ ]:
```
utils.plot_hist_boxplot_classes(df_2019, OUTPUT)
utils.plot_hist_boxplot_classes(df_2020, OUTPUT)
```

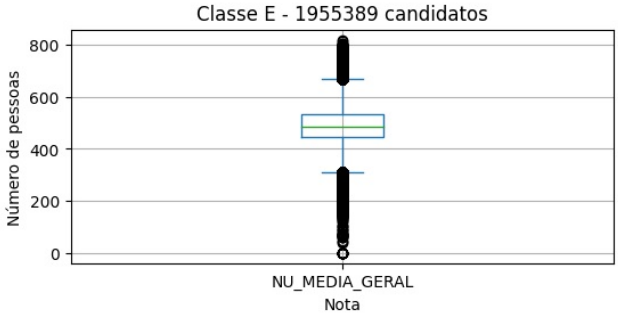
```
utils.plot_hist_boxplot_classes(df_2021, OUTPUT)
```

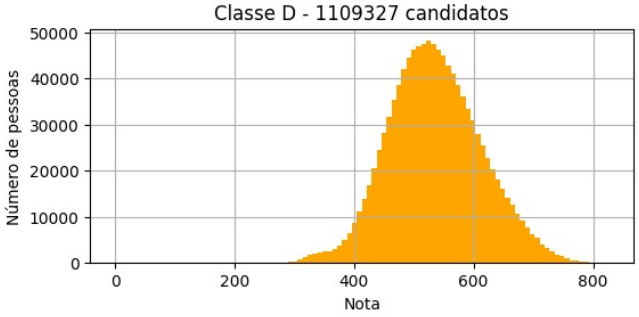
Número de pessoas x Media Geral - 2019 - 3701909 candidatos

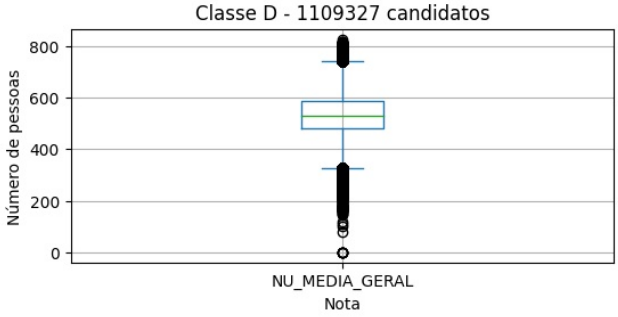


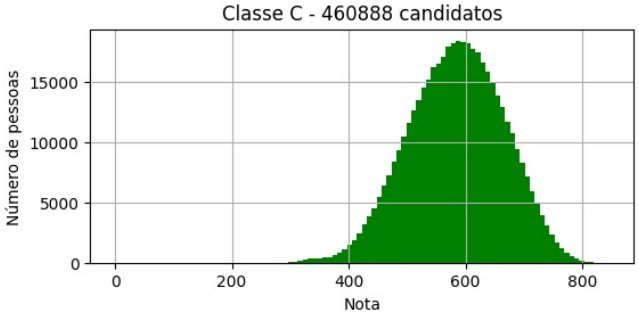


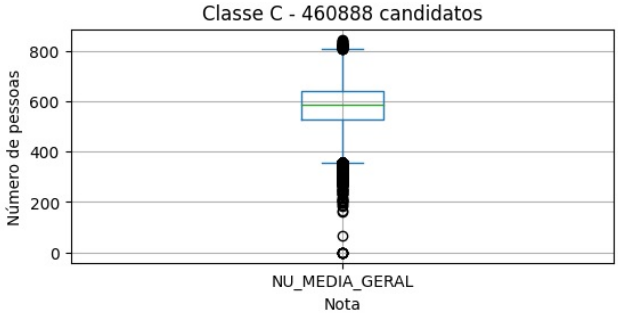


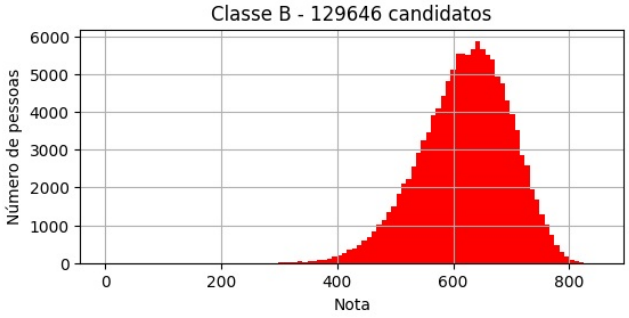


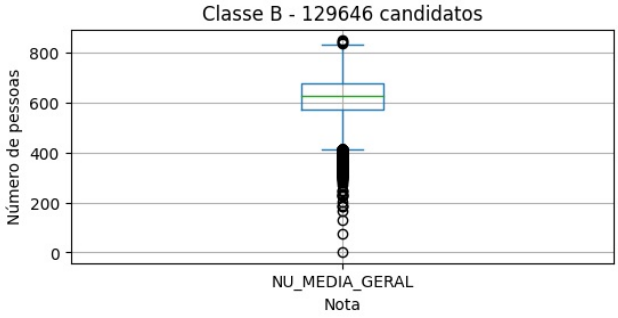


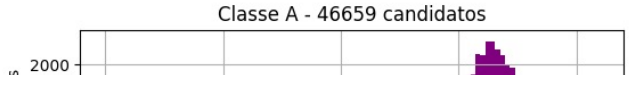


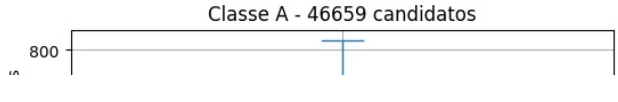


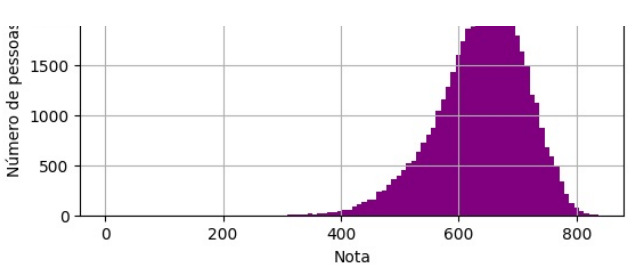


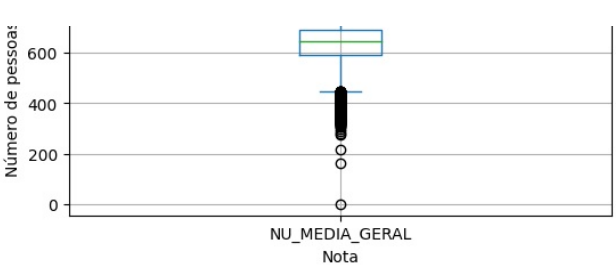




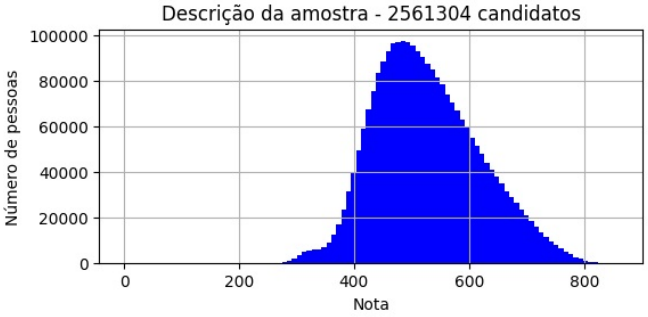


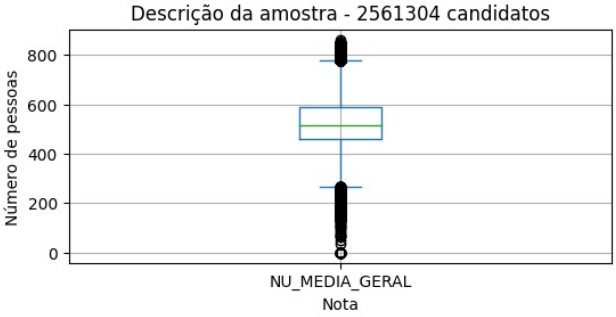


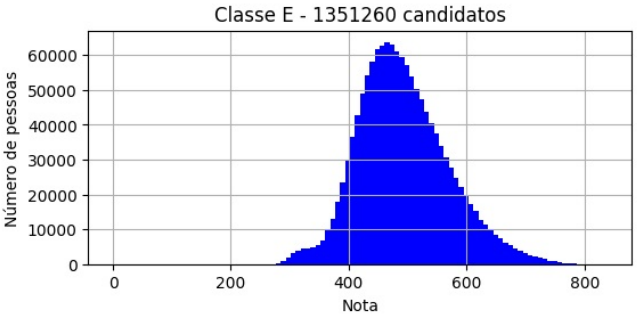


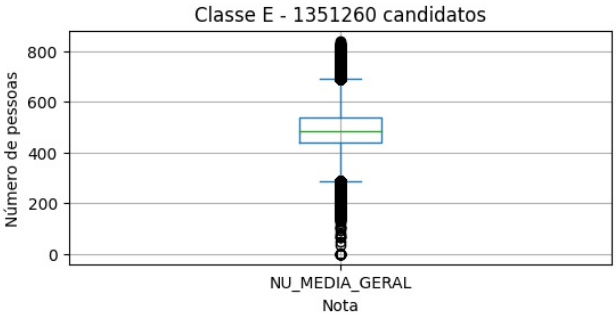


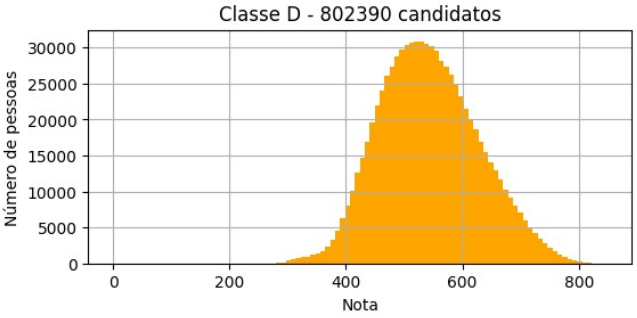
Número de pessoas x Media Geral - 2020 - 2561304 candidatos

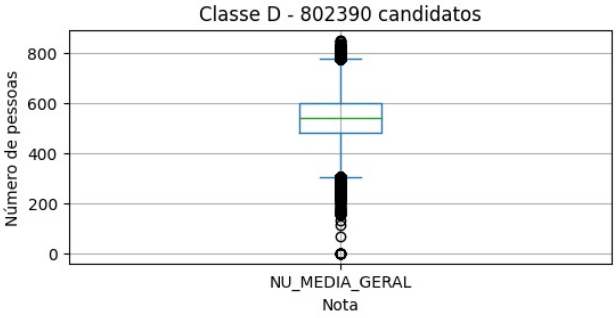


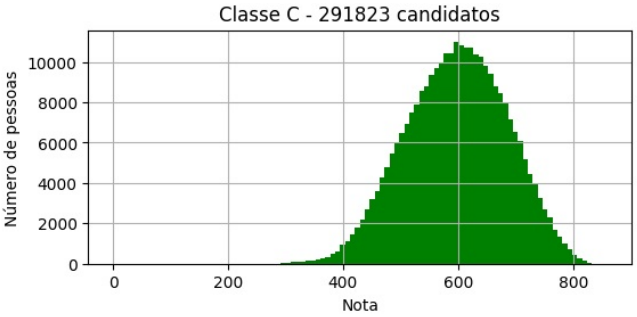


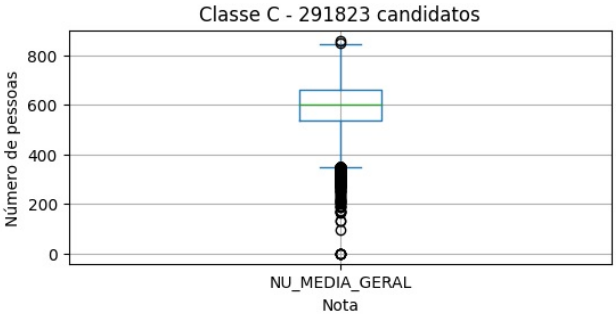


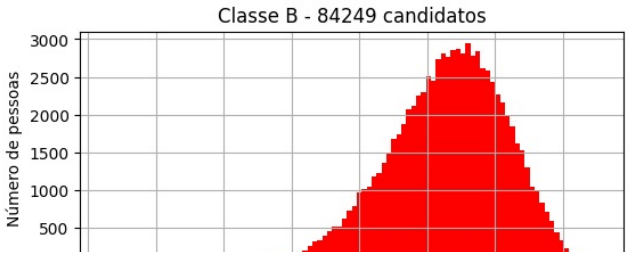


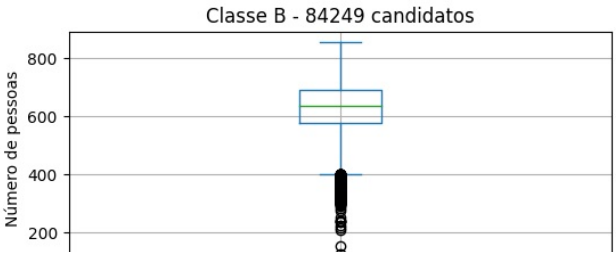


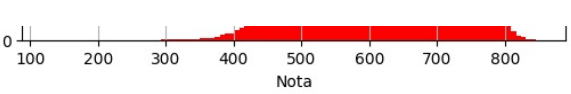


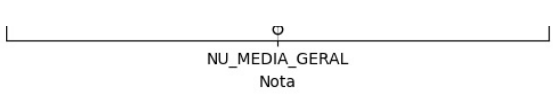


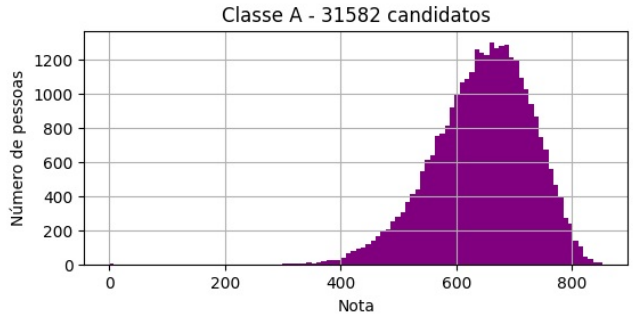


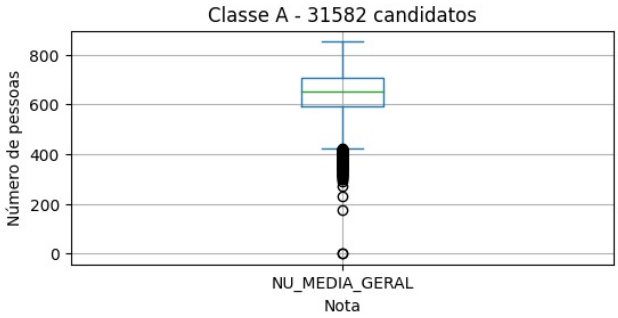




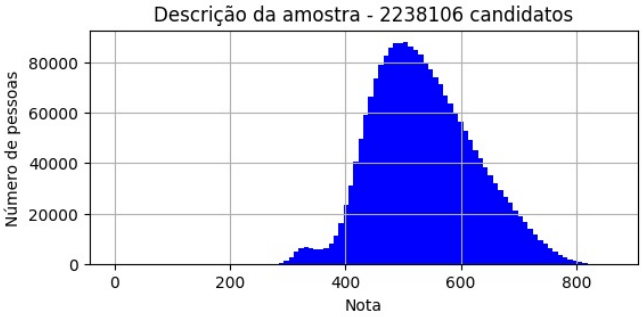


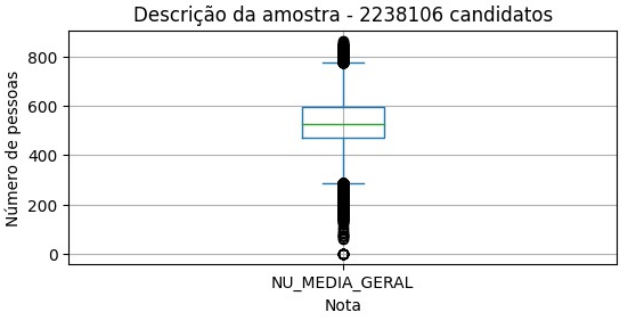


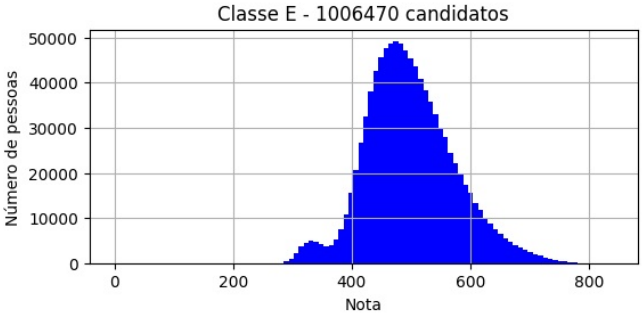


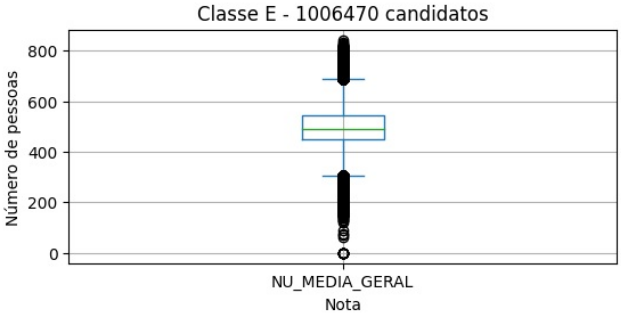


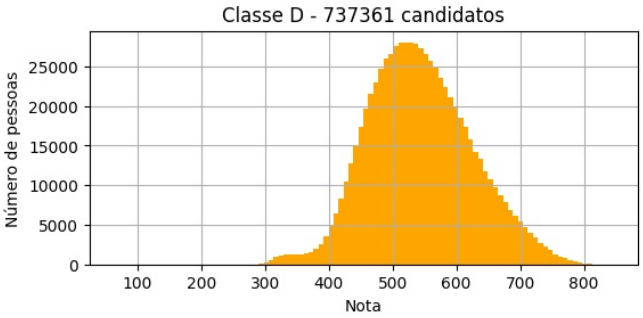
Número de pessoas x Media Geral - 2021 - 2238106 candidatos

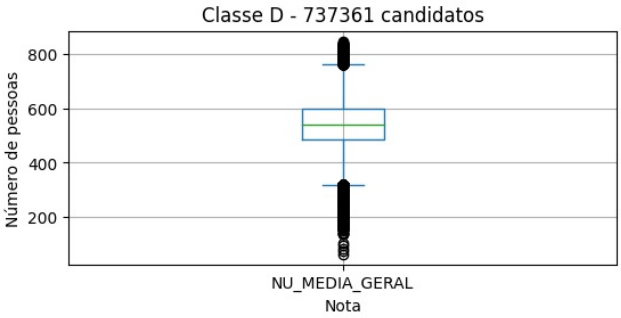


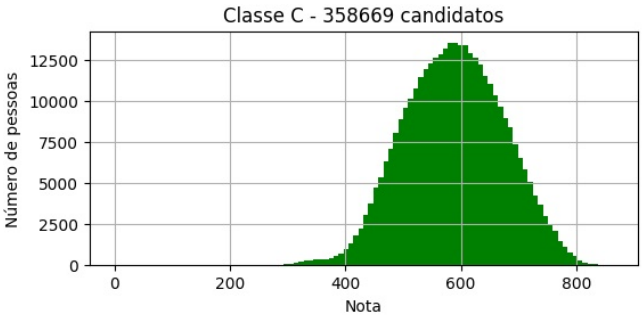


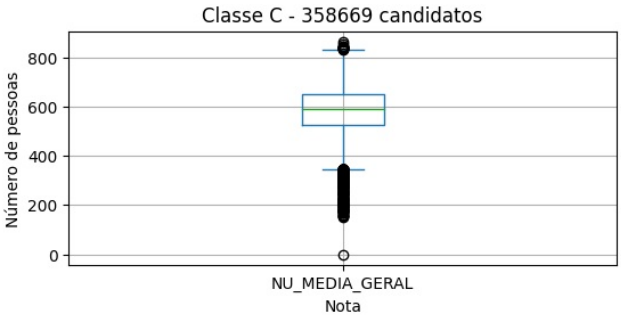


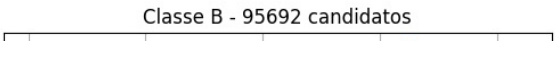


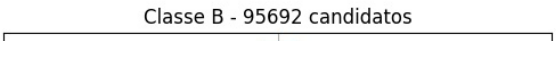


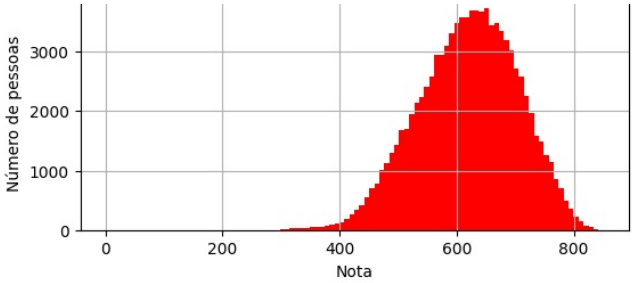


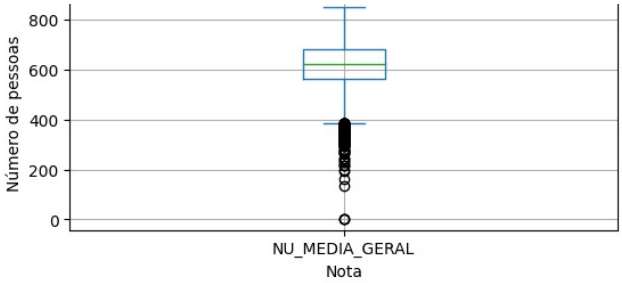


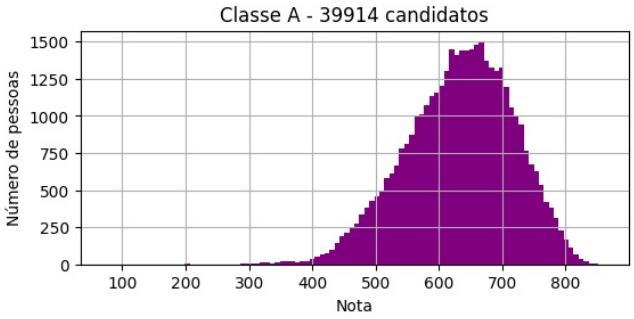


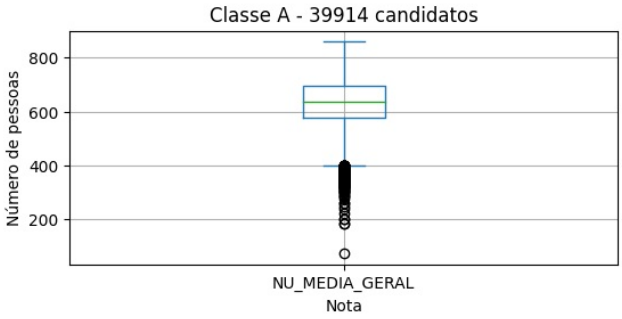












In [ ]:
```
utils.plot_boxplot_descricao_amostras([df_2019, df_2020, df_2021], OUTPUT)
```