Questao_1

February 28, 2021

1 Questão 1

In [17]:
```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
```

In [18]:
```
path = "teste_smarkio_lbs.xls"
df1 = pd.read_excel(path, 'Análise_ML')
df2 = pd.read_excel(path, 'NLP')
```

1.1 1a aba de dados:

In [19]:
```
# Sanity check dos dados, para verificar, por exemplo, que nao existem
# probabilidades negativas ou maiores que 1
df1.describe()
```

Out[19]:

	Pred_class	probabilidade	True_class
count	643.000000	643.000000	181.000000
mean	52.712286	0.622436	38.574586
std	37.602068	0.266811	39.581017
min	2.000000	0.043858	0.000000
25%	12.000000	0.408017	0.000000
50%	59.000000	0.616809	24.000000
75%	81.000000	0.870083	74.000000
max	118.000000	1.000000	117.000000

In [20]:
```
df1.head()
```

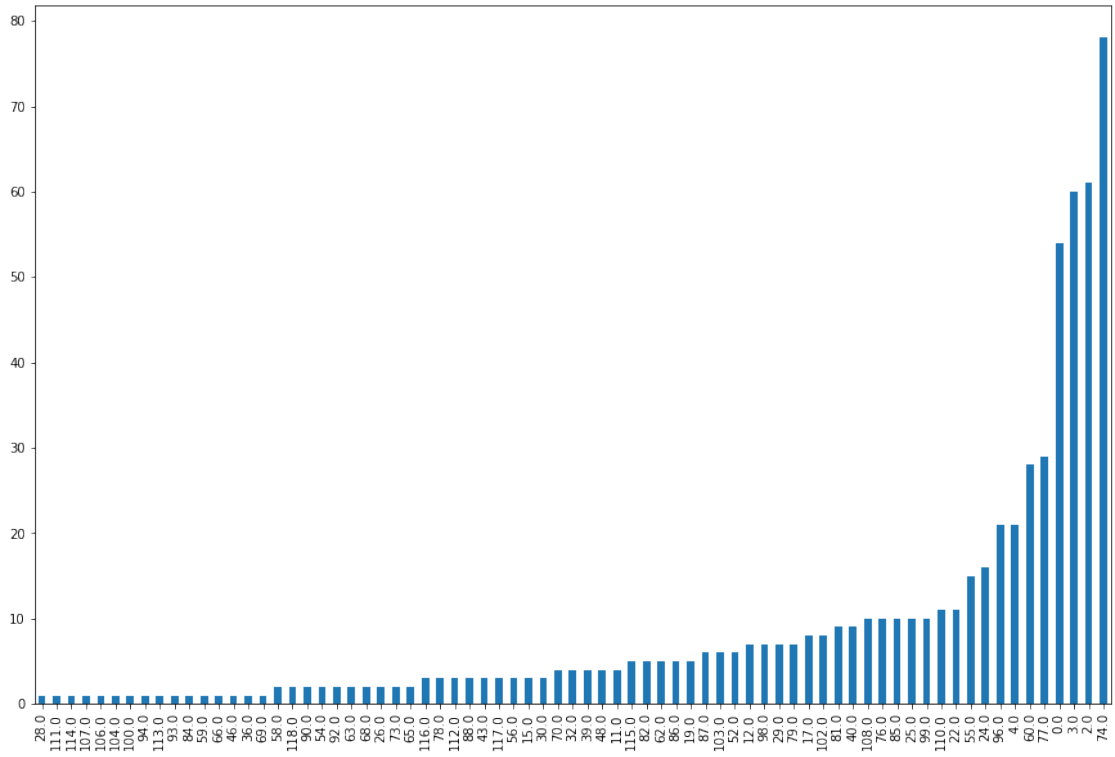
Out[20]:

	Pred_class	probabilidade	status	True_class
0	2	0.079892	approved	0.0
1	2	0.379377	approved	74.0
2	2	0.379377	approved	74.0
3	2	0.420930	approved	74.0
4	2	0.607437	approved	NaN

In [21]:
```
# Distribuição das classes verdadeiras: é possível ver que algumas são
# dominantes enquanto as demais aparecem poucas vezes
```

```
df1['True_class'] = np.where(df1['True_class'].isnull(),\
                             df1['Pred_class'],df1['True_class'])
df1['True_class'].value_counts().sort_values().plot(kind='bar',\
                                                    figsize=(15,10))
```

Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0xb22f30>



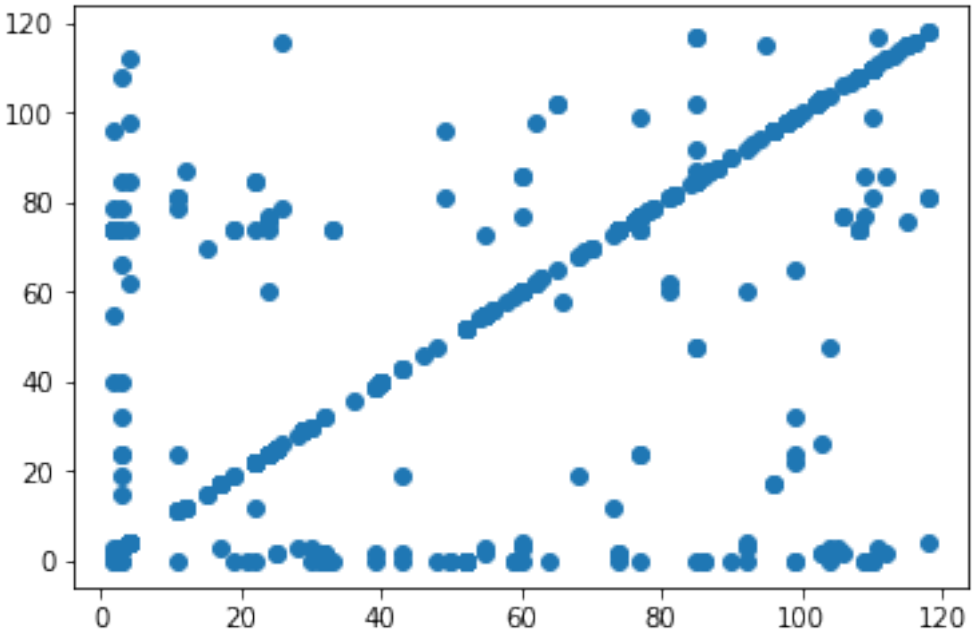
In [22]:
```
# Matriz de correlação das variaveis numericas, indicando correlação
# moderada entre as classes predita e verdadeira, que pode ser
# visualizada abaixo.
df1.corr()
```

Out[22]:

	Pred_class	probabilidade	True_class
Pred_class	1.000000	-0.123457	0.672541
probabilidade	-0.123457	1.000000	0.051251
True_class	0.672541	0.051251	1.000000

In [23]:
```
import matplotlib.pyplot as plt
plt.scatter(df1['Pred_class'],df1['True_class'] )
```

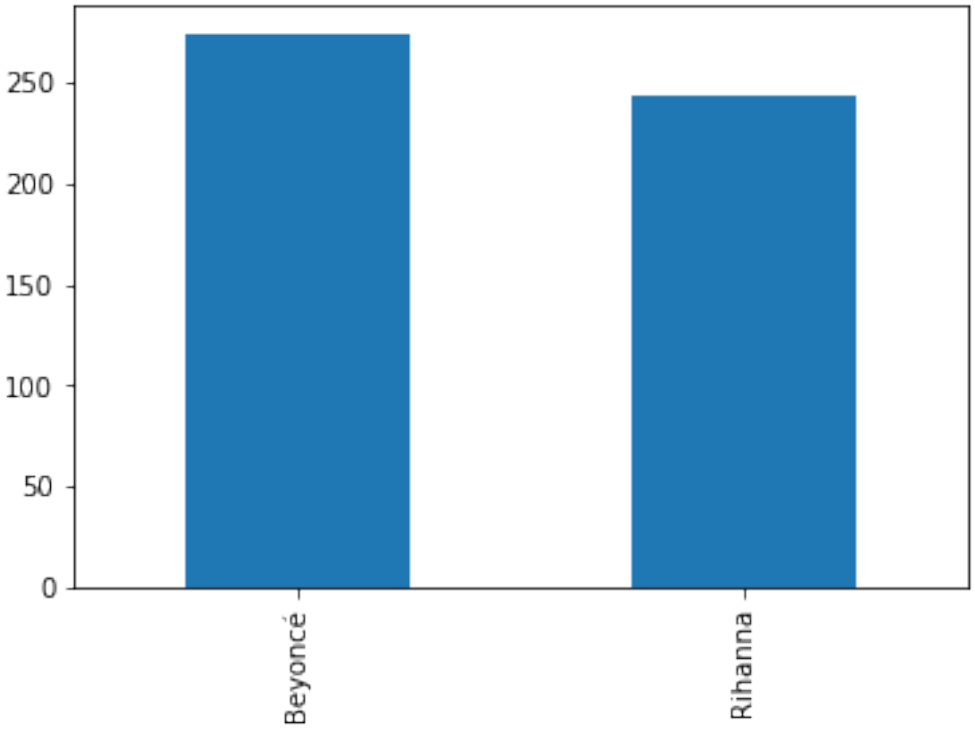
Out[23]: <matplotlib.collections.PathCollection at 0xb968b0>



1.2 2a aba de dados:

```
In [24]: # Em relação à segDistribuição da quantidade de musica de cada artista
         # a amostra está bem equilibrada
         df2['artista'].value_counts().plot(kind='bar')
```

```
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0xb96d90>
```


250
200
150
100
50
0
Beyoncé
Rihanna