artists

December 29, 2023

[3]:
```
import os
import pandas as pd
from config_file import data_path

path = os.path.join(data_path, "artists.jsonl")
df = pd.read_json(path, orient="records", lines=True)
```

[4]:
```
df.head()
```

[4]:
```
                      id             name  \
0  72578usTM6Cj5qWsi471Nc      Raghu Dixit
1  7b6Ui7JVaBDEfZB9k6nHL0  The Local Train
2  5wJ1H6ud777odtZl5gG507    Vishal Mishra
3  0n4a5imdLBN24fIrBWoqrv          Because
4  4gdMJYnopf2nEUcanAwstx        Anuv Jain

                                              genres
0     [filmi, indian folk, indian rock, kannada pop]
1  [desi pop, hindi indie, indian indie, indian r…
2                        [desi pop, modern bollywood]
3  [opm, pinoy hip hop, pinoy r&b, pinoy trap, ta…
4  [hindi indie, indian indie, indian singer-song…
```

[4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27650 entries, 0 to 27649
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      27650 non-null  object
 1   name    27650 non-null  object
 2   genres  22101 non-null  object
dtypes: object(3)
memory usage: 648.2+ KB
```

[5]:
```
df.describe()
```

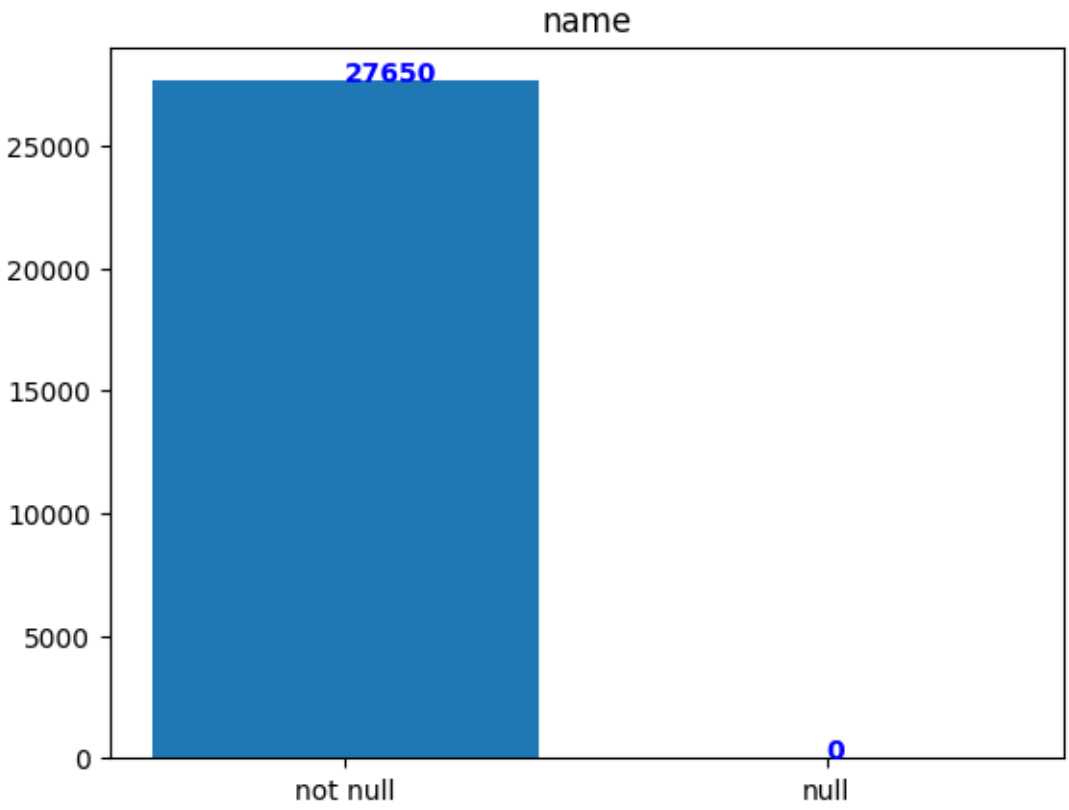
[5]:

	id	name	genres
count	27650	27650	22101
unique	22163	27542	11645
top	-1	TNT	[classic thai pop]
freq	5488	4	59

[9]:
```
from matplotlib import pyplot as plt
def plot_hist(df, col):
    col_info = [df[col].notnull().sum(), df[col].isnull().sum()]
    print(f"not null: {col_info[0]}, null: {col_info[1]}")
    plt.bar(["not null", "null"], col_info)
    for i, v in enumerate(col_info):
        plt.text(i, v, str(v), color="blue", fontweight="bold")
    plt.title(col)
    plt.show()
```

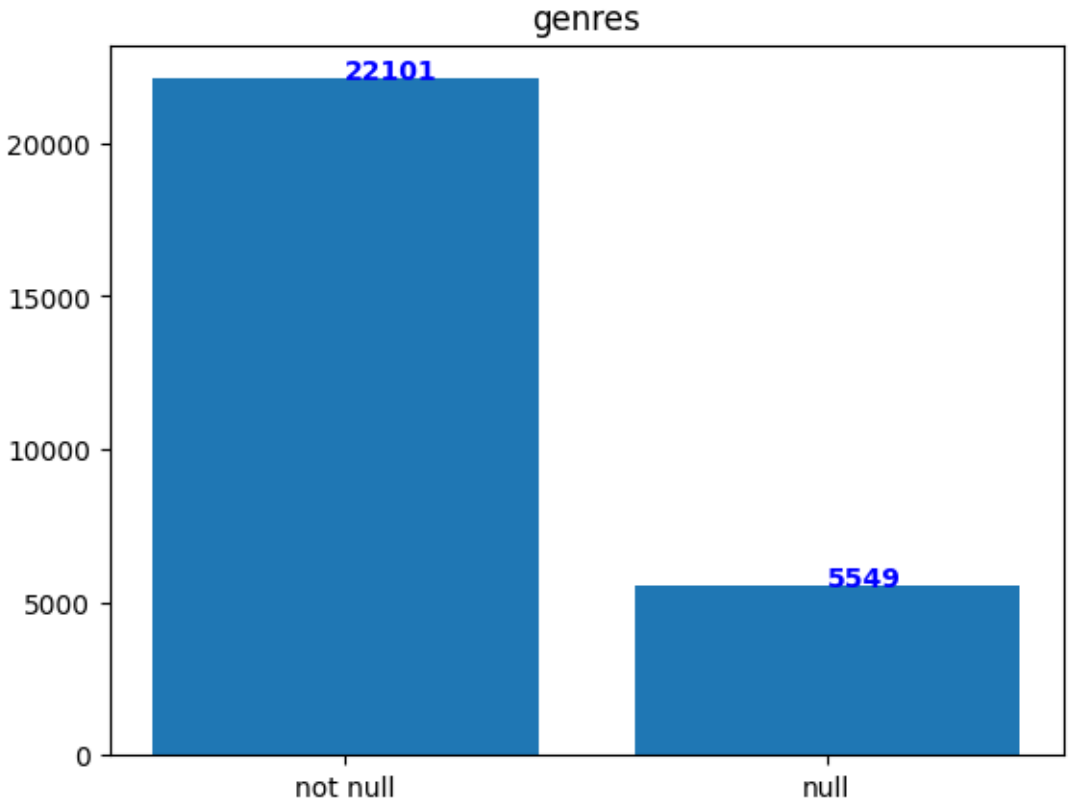
[10]:
```
plot_hist(df, "name")
```

```
not null: 27650, null: 0
```



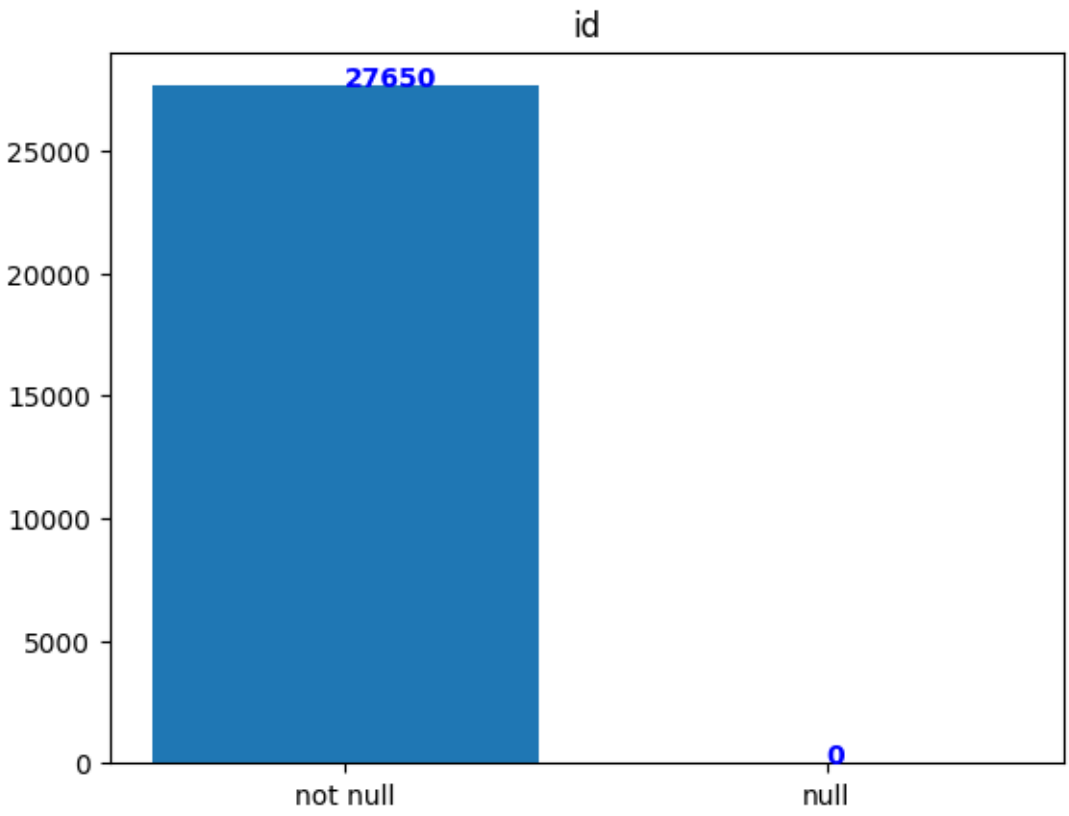
[11]:
```
plot_hist(df, "genres")
```

```
not null: 22101, null: 5549
```



[13]:
```
plot_hist(df, "id")
```

```
not null: 27650, null: 0
```



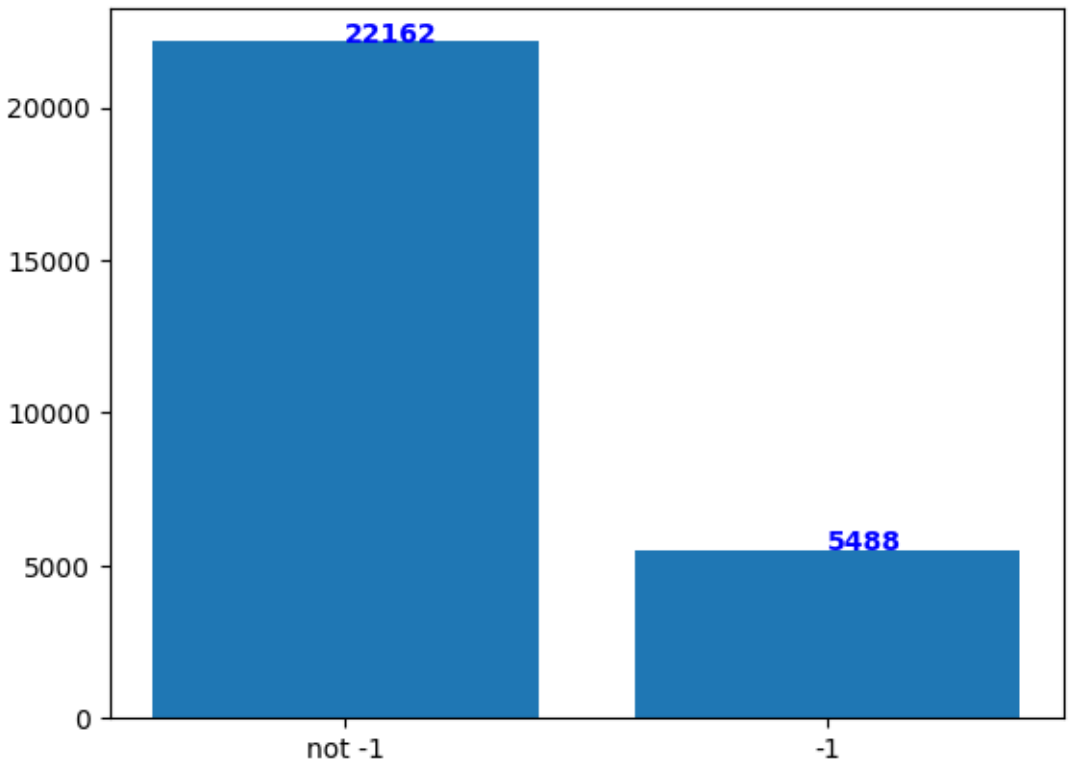
[15]:
```
# df where id is -1
df[df["id"] == -1]
```

[15]:
```
        id                              name  \
6       -1                       Los Arroyenos
12      -1                Los Hermanos Cardozo
29      -1                           Tren Loco
30      -1                                Orly
33      -1  Sergio Galleguillo Y Los Amigos
…     ..                                 …
27623   -1                           Morttagua
27628   -1                           Fon.Leman
27640   -1                              Sponge
27641   -1
27649   -1


                                                 genres
6                                          [villancicos]
12                        [chamame, folklore argentino]
29     [argentine heavy metal, argentine metal, latin…
30                        [cuarteto, musica blumenauense]
```

```
33                                    [folklore argentino]
…                                                      …
27623                        [focus trance, melodic techno]
27628     [progressive trance, progressive trance house]
27640  [alternative rock, grunge, pop rock, post-grunge]
27641                                                 None
27649                                        [thai indie]

[5488 rows x 3 columns]
```

[28]:
```
col_info = [df[df["id"] != -1].shape[0], df[df["id"] == -1].shape[0]]
print(f"not -1: {col_info[0]}, -1: {col_info[1]}")
plt.bar(["not -1", "-1"], col_info)
for i, v in enumerate(col_info):
    plt.text(i, v, str(v), color="blue", fontweight="bold")
plt.show()
```

```
not -1: 22162, -1: 5488
```



0.1 Podsumowanie

- gatunek wykonywany przez artystę często wynosi null

- id artysty wynosi często -1