Gălățeanu Andrei-Dan 311CC

30.05.2024

Cerinta 1: Cititi informatiile din fisierul train.csv si examinati structura acestora. Pentru acest lucru, trebuie să determinati programatic (utilizând cod Python) următoarele: numărul de coloane, tipurile datelor din fiecare coloană, numărul de valori lipsă pentru fiecare coloană, numărul de linii, dacă există linii duplicate.

Output:

```
891
12
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

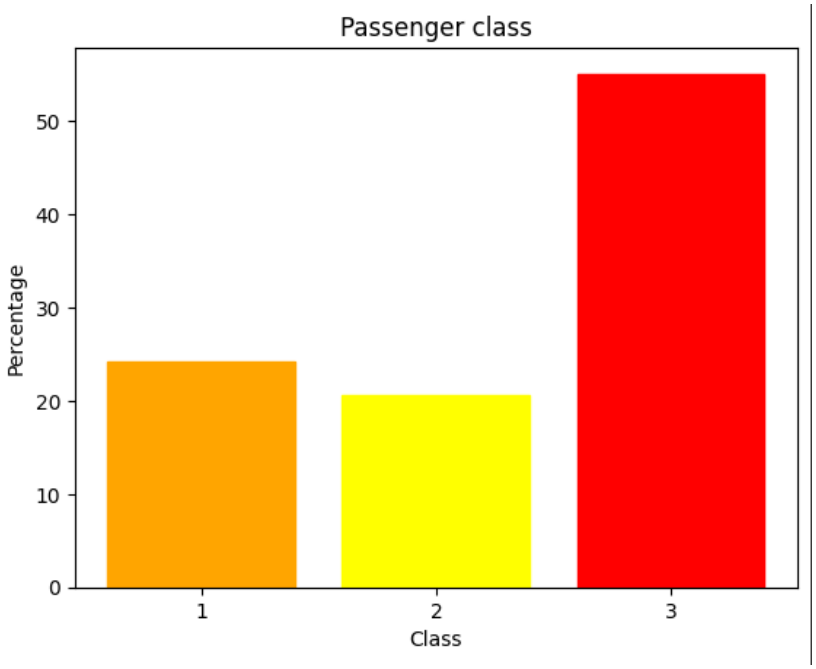
->12 coloane

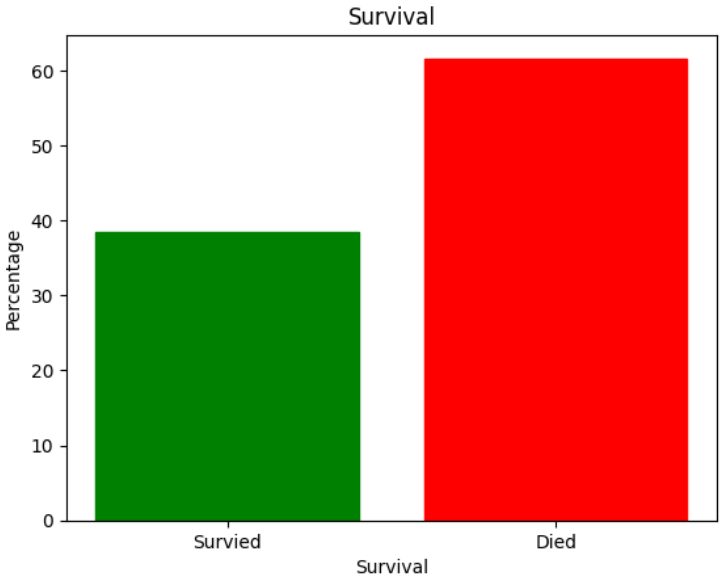
->891 linii

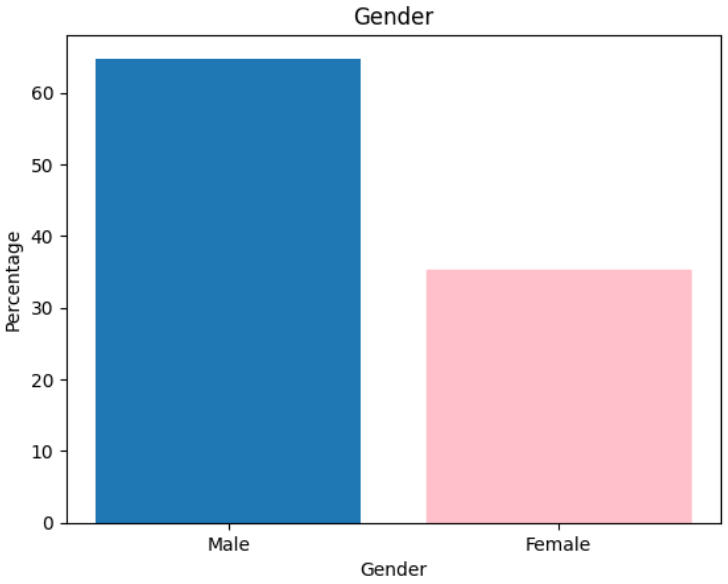
Cerinta 2: Determinati care este procentul persoanelor care au supravietuit si procentul persoanelor care nu au supravietuit. Determinati care este procentul pasagerilor pentru fiecare tip de clasă (coloana Pclass). Determinati care este procentul bărbatilor si care este procentul femeilor. Realizati un grafic potrivit pentru prezentarea acestor rezultate.

Output:

```
Percentage of people who didn't survive: 61.62%
Percentage of 1st class: 24.24%
Percentage of 2nd class: 20.65%
Percentage of 3rd class: 55.11%
```

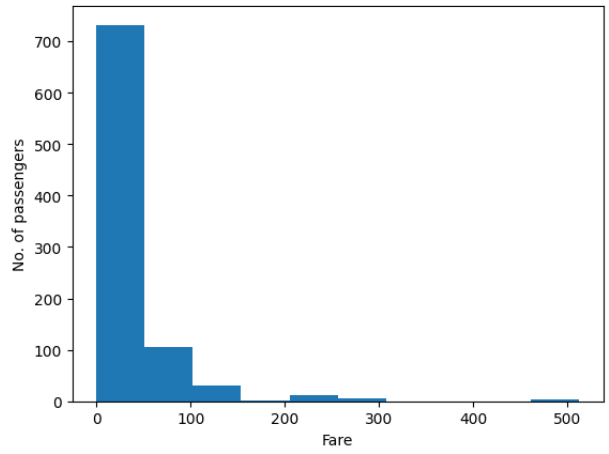


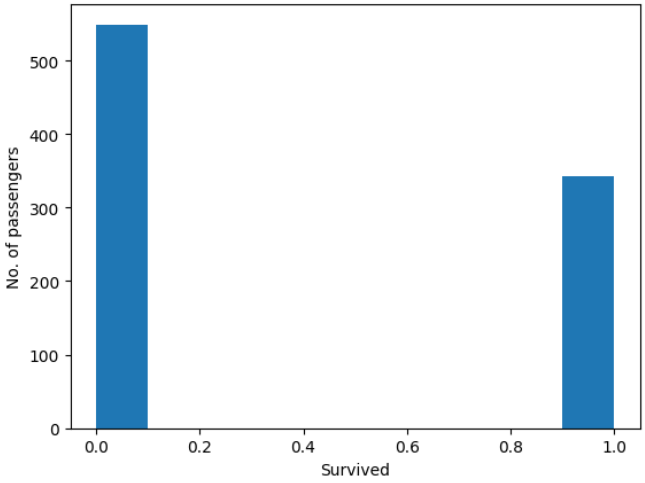


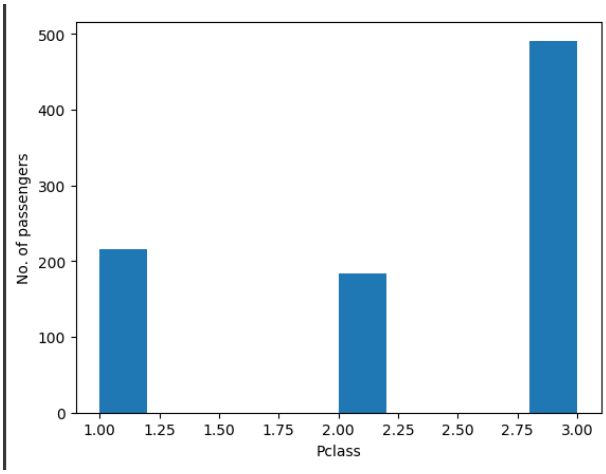


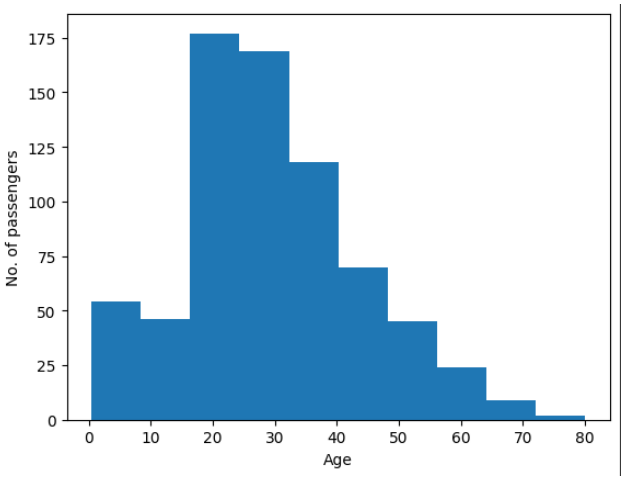
Cerinta 3: Această cerintă implică generarea de histograme pentru fiecare coloană cu valori numerice din setul de date Titanic. O histogramă este o reprezentare grafică a distributiei frecventelor unei variabile continue. Pe axa orizontală sunt incluse intervalele de valori ale variabilei, iar pe axa verticală se reprezintă numărul de exemple din setul de date care sunt incluse în fiecare interval. Histograma oferă o imagine vizuală a modului în care valorile sunt distribuite si permite identificarea tendintelor si a modelului de distributie al datelor. În cadrul acestei cerinte, pentru fiecare coloană numerică din setul de date Titanic, se va realiza o histogramă pentru a vizualiza distributia datelor si a evidentia caracteristicile importante ale acestora.

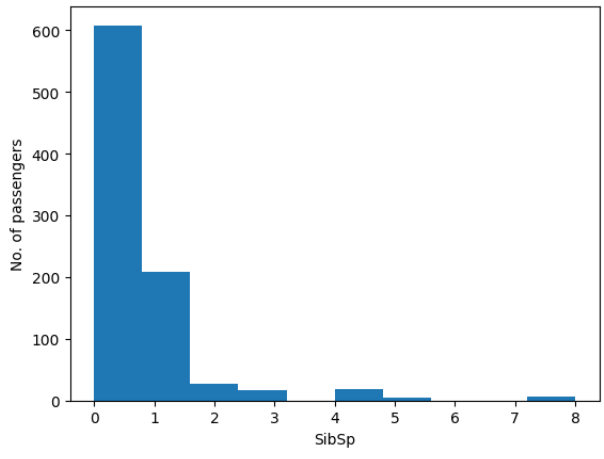
Output:

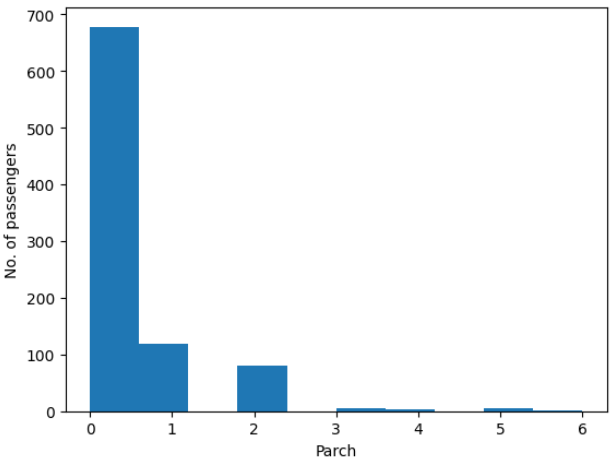












Cerinta 4: Identificati coloanele pentru care există valori lipsă. Apoi, pentru fiecare coloană identificată determinati numărul si proportia valorilor lipsă. Determinati care este procentul acestora pentru fiecare dintre cele două clase (coloana Survived).

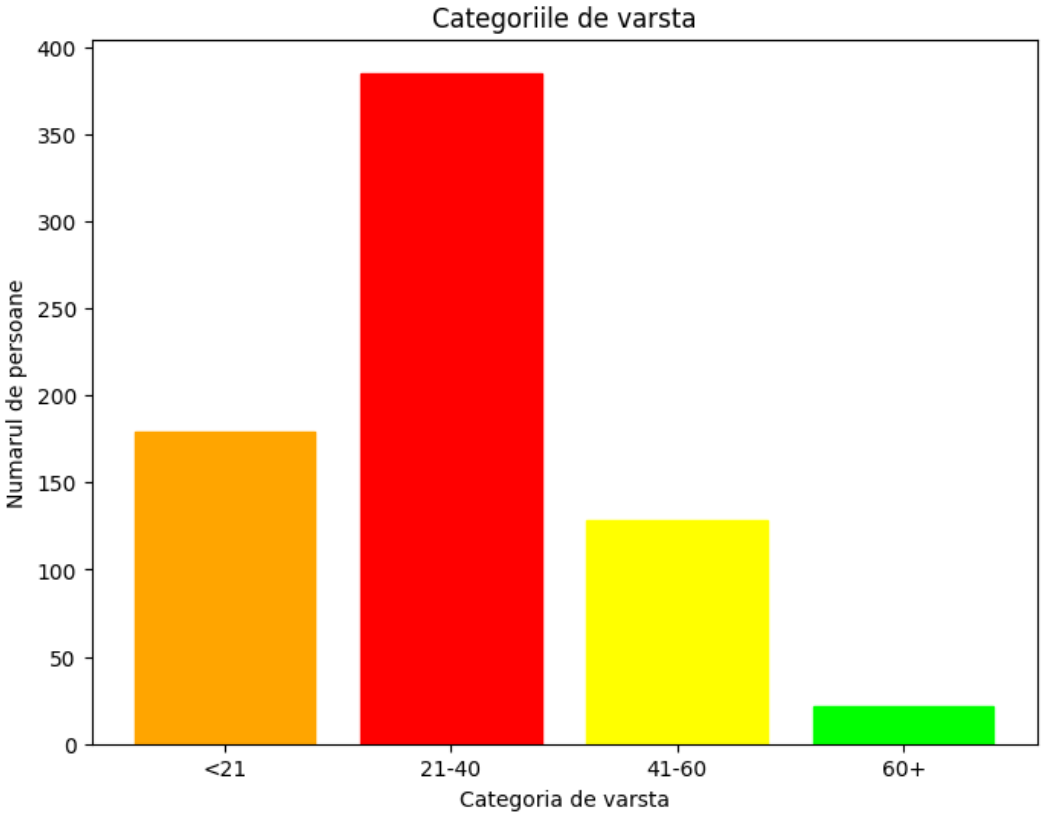
Output:

```
['Age', 'Cabin', 'Embarked']
Coloana Age are 177 valori lipsa, o proportie de 0.20.
Pt cei care au supravietuit sunt 15.20% si pt restul un procent de 22.77%

Coloana Cabin are 687 valori lipsa, o proportie de 0.77.
Pt cei care au supravietuit sunt 60.23% si pt restul un procent de 87.61%

Coloana Embarked are 2 valori lipsa, o proportie de 0.00.
Pt cei care au supravietuit sunt 0.58% si pt restul un procent de 0.00%
```

Cerinta 5: Considerăm patru categorii de vârstă: [0, 20], [21, 40], [41, 60], [61, max]. Determinati câti pasageri avem pentru fiecare din această categorie. Introduceti o coloană suplimentară si determinati pentru fiecare exemplu din setul de date indexul categoriei din care face parte. Realizati un grafic potrivit pentru a evidentia aceste rezultate.

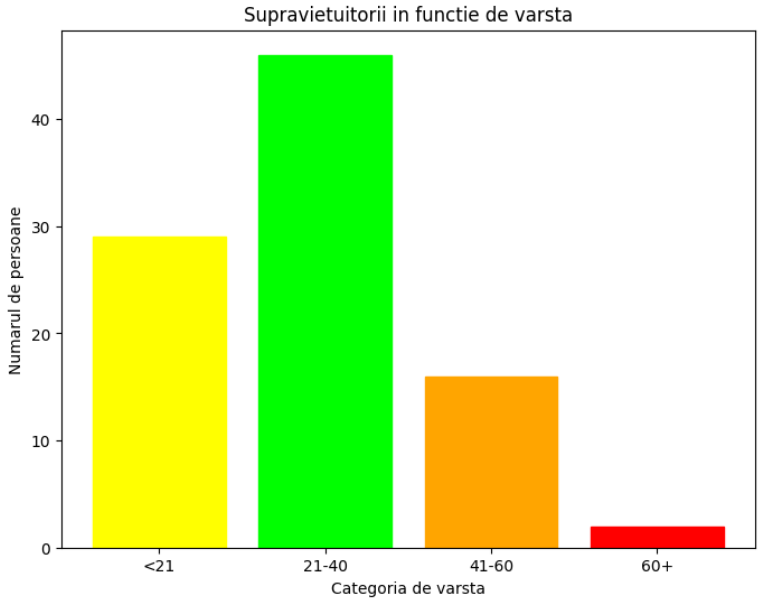
Output:

```
Sunt 179 de persoane cu varsta pana in 20 de ani
Sunt 385 de persoane cu varsta intre 21 si 40 de ani
Sunt 128 de persoane cu varsta intre 41 si 60 de ani
Sunt 22 de persoane cu varsta mai mare de 60 de ani
```



Cerinta 6: Determinati câti bărbati au supravietuit pentru fiecare dintre cele 4 categorii de vârstă propuse anterior. Realizati un grafic în care să evidentiati cum influentează vârsta procentul de supravietuire al bărbatilor, pe baza informatiilor pe care le avem în setul de date.

Output:

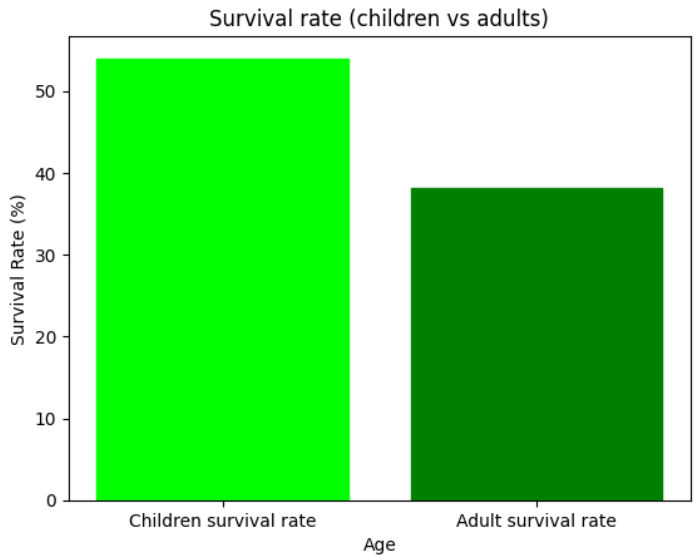


	AgeGroup	Survived	Age
AgeGroup	1.000000	-0.068878	0.914957
Survived	-0.068878	1.000000	-0.077221
Age	0.914957	-0.077221	1.000000

Cerinta 7: Determinati procentul copiilor aflati la bord (considerăm copii persoane cu vârsta < 18 ani). Realizati un grafic în care să evidentiati rata de supravietuire pentru copii si pentru adulti.

Output:

```
Au fost 113 copii la bord, un procent de 12.68
Au fost 601 adulti la bord
```



Cerinta 8: Completati valorile lipsă cu cele obtinute pentru media pasagerilor care fac parte din aceeasi clasă. Spre exemplu, dacă există o înregistrare pentru un pasager care supravietuieste, dar pentru care nu cunoastem vârsta, completăm vârsta cu media pasagerilor care au supravietuit. În cazul în care avem o coloană cu valori categoriale, determinăm cea mai frecventă valoare pentru respectiva clasă.

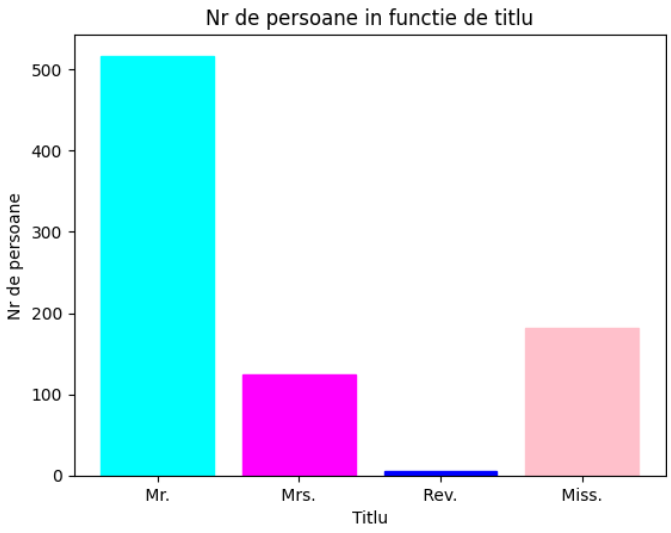
Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          891 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
 12  AgeGroup     891 non-null    category
dtypes: category(1), float64(2), int64(5), object(5)
memory usage: 84.7+ KB
```

Cerinta 9: Verificati dacă titlurile de noblete regăsite în coloana Name (Mr., Mrs., Don, etc.) corespund cu sexul persoanei respective. Reprezentati grafic câte persoane corespund fiecărui titlu.

Output:

```
Sunt 517 persoane cu titlul  Mr.
Toate persoanele au sexul corespunzator
Sunt 125 persoane cu titlul  Mrs.
Toate persoanele au sexul corespunzator
Sunt 0 persoane cu titlul  Don
Toate persoanele au sexul corespunzator
Sunt 0 persoane cu titlul  Sir
Toate persoanele au sexul corespunzator
Sunt 6 persoane cu titlul  Rev.
Toate persoanele au sexul corespunzator
Sunt 182 persoane cu titlul  Miss.
Toate persoanele au sexul corespunzator
Sunt 0 persoane cu titlul  Master
Toate persoanele au sexul corespunzator
```

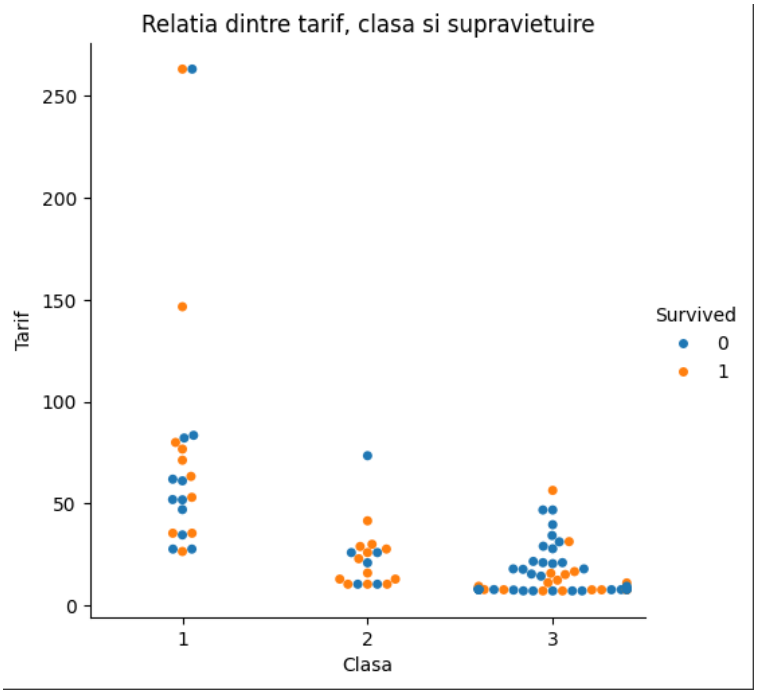


Cerinta 10: A influentat starea de a fi singur pe Titanic (nu are deloc rude pe vas) sansele de supravietuire? Histograma ar putea ajuta la investigarea acestui aspect. Investigati relatia dintre tarif, clasă si starea de supravietuire pentru primele 100 de înregistrări folosind catplot() din seaborn. (sugestie: folositi kind='swarm' pentru a vedea detalii pe grafic).

Output:

Observatii: se observa ca rata de supravietuire a fost de aprox 50% pt cei care nu erau singuri pe vapor. Cei care erau singuri au avut o rata de supravietuire mai mica de 30%.

	Survived	Fare	Pclass
Survived	1.000000	0.257307	-0.338481
Fare	0.257307	1.000000	-0.549500
Pclass	-0.338481	-0.549500	1.000000



Observatii: se observa ca tariful a fost in general mai mare pentru cei de la clasa 1. (1 > 2 > 3). De asemenea, putem observa ca majoritatea celor de la clasa 3 nu au supravietuit.

Cea mai mare rata de supravietuire pare a fi in clasa a 2a.