 # Rusanescu Andrei-Marian 313CC

https://github.com/andr31154/PCLP3-Project

1. Incarc datele din fisierul train.csv. Afisez numarul coloanelor cu metoda columns, tipurile de date cu metoda .dtypes. Pentru a calcula valorile lipsa pentru fiecare coloana folosesc metoda isnull() si apoi suma, pentru a calcula numarul de valori lipsa pentru fiecare coloana. Cu metoda .duplicated gasesc duplicatele, si apoi pentru a afisa numarul acestora, se foloseste .sum().

Sunt 12 coloane

Tipurile de date pentru fiecare coloana

PassengerId   int64

Survived   int64

Pclass   int64

Name   object

Sex   object

Age   float64

SibSp   int64

Parch   int64

Ticket   object

Fare   float64

Cabin   object

Embarked   object

dtype: object

Valorile lipsa pentru fiecare coloana

PassengerId   0

```
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In total sunt 891 linii

Nu sunt linii duplicate

2. In cadrul acestui task, pentru a afisa datele am folosit rotunjirea la 2 zecimale folosind functia round. Toate datele sunt reprezentate in acelasi grafic si salvate in fisierul cer2.png.

Au supravietuit: 38.38%

Nu au supravietuit: 61.62%

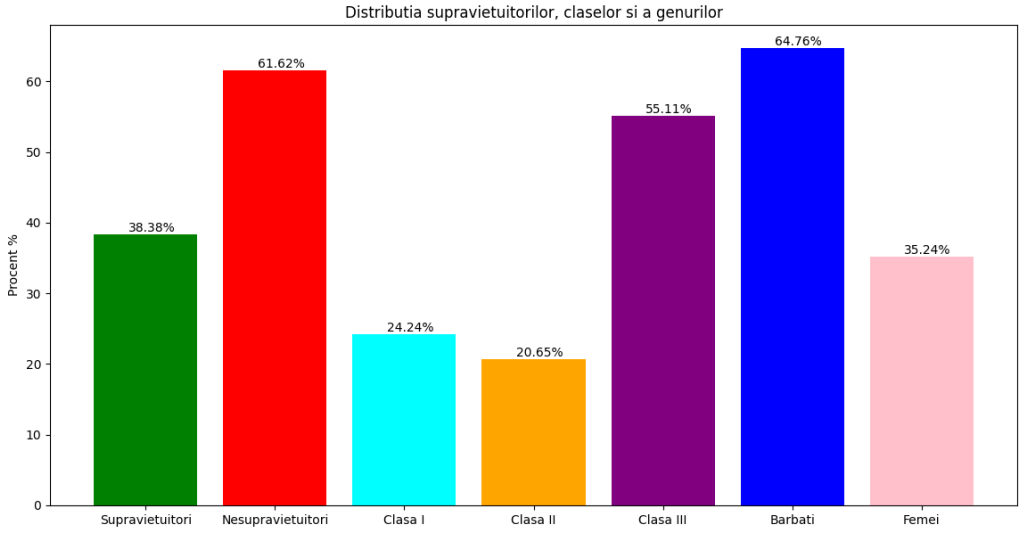
Pasageri de clasa I: 24.24%

Pasageri de clasa II: 20.65%

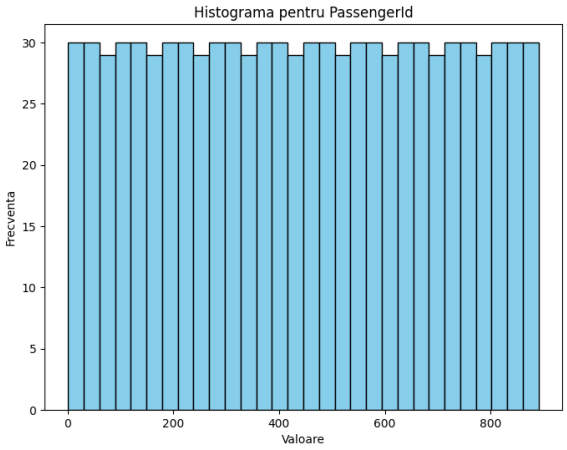
Pasageri de clasa III: 55.11%

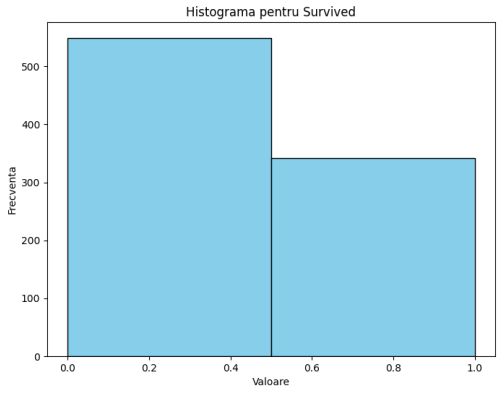
Barbati: 64.76%

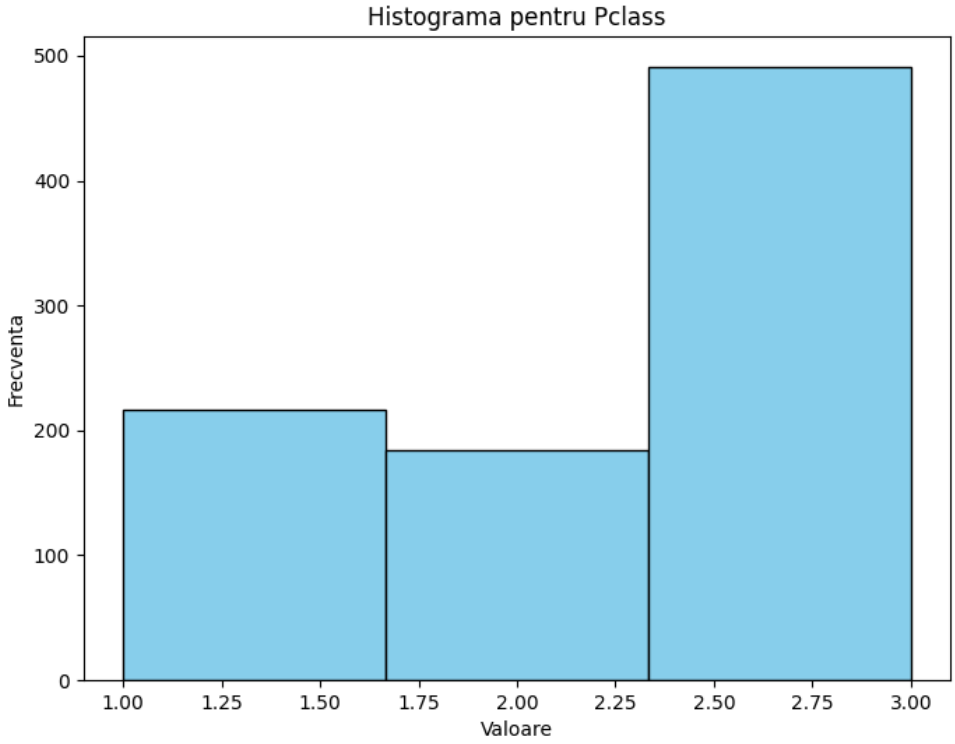
Femei: 35.24%

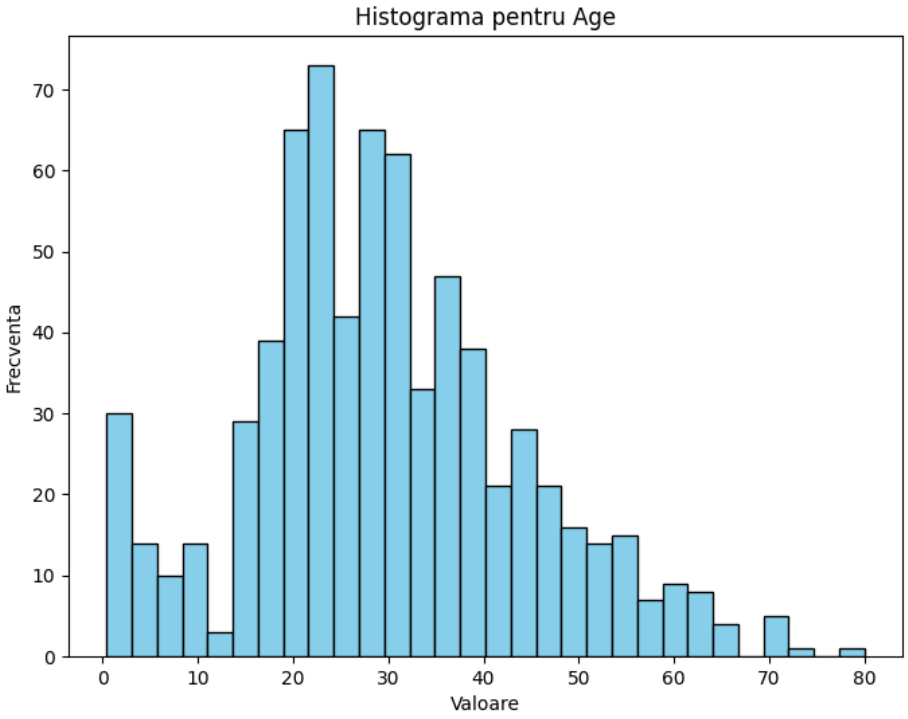


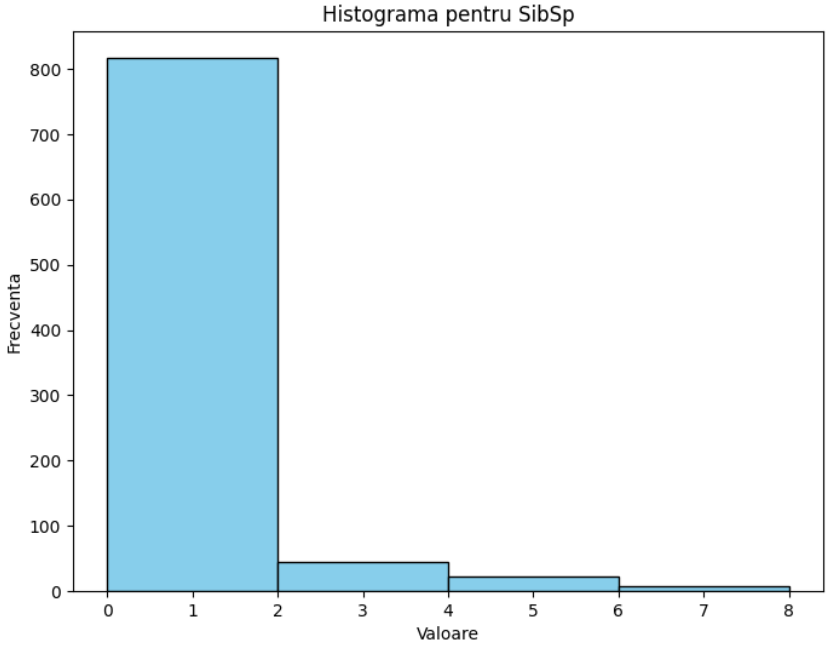
3. Pentru a selecta toate coloanele care contin valori numerice, se foloseste metoda select_dtypes cu parametrul number. Se itereaza prin fiecare coloanal si in functie de ce coloana e, se afiseaza o histograma cat mai potrivita. Aceasta se salveaza in fisierul corespunzator de forma cer3_$i.png, i = 0 : n.

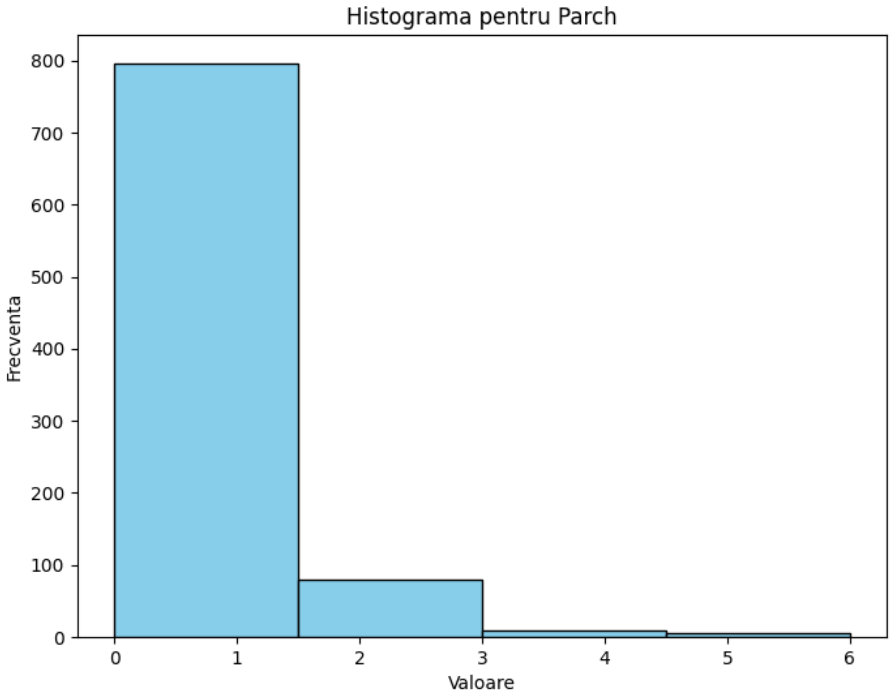


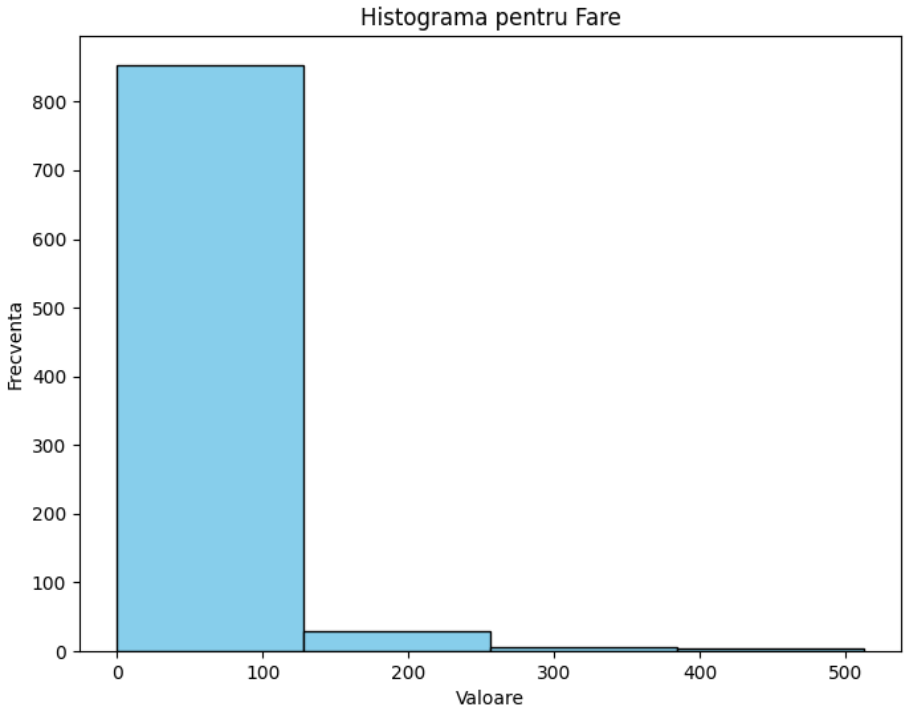



Histograma pentru Pclass
Frecventa
Valoare
0
100
200
300
400
500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Histograma pentru Age
Frecventa
Valoare
0
10
20
30
40
50
60
70
80


Histograma pentru SibSp
Frecventa
Valoare
0
100
200
300
400
500
600
700
800
0
1
2
3
4
5
6
7
8


Histograma pentru Parch
Frecventa
Valoare
0
100
200
300
400
500
600
700
800
0
1
2
3
4
5
6



4. Se extrag coloanele care contin valori lipsa cu aceeasi metoda descrisa mai sus. Se afiseaza numarul de valori lipsa si proportia acestora.

Coloanele cu valori lipsa si proportia lor:

Age: 177 valori lispsa, 19.87%

Cabin: 687 valori lispsa, 77.1%

Embarked: 2 valori lispsa, 0.22%

Apoi acestea sunt grupate dupa coloana Survived si se afiseaza relatia dintre aceste date si daca cate persoane au supravietuit sau nu.

Age

Survived = 0: 125 valori lipsa (22.77%)

Survived = 1: 52 valori lipsa (15.2%)

Cabin

Survived = 0: 481 valori lipsa (87.61%)

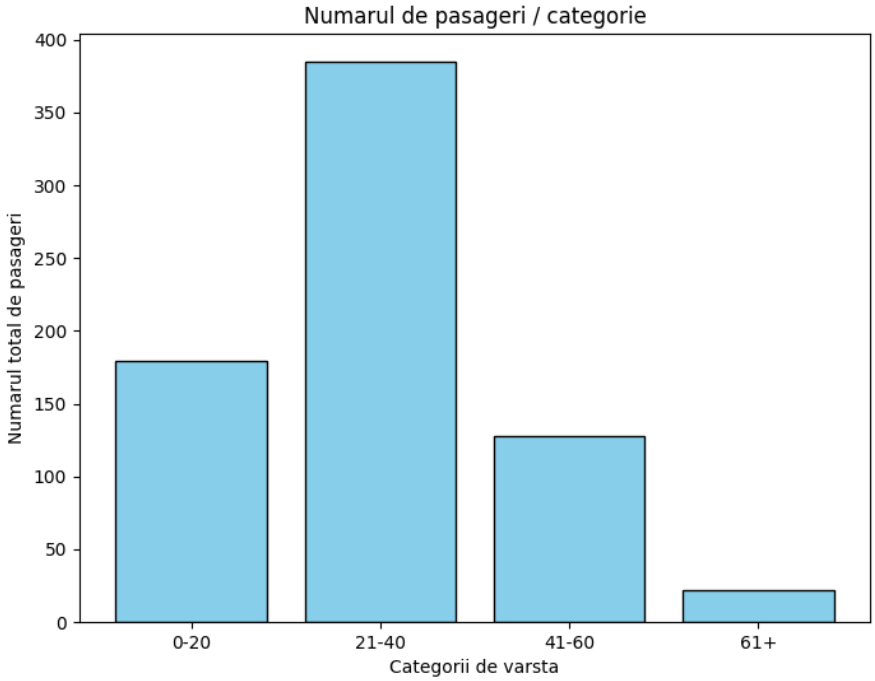
Survived = 1: 206 valori lipsa (60.23%)

Embarked

Survived = 0: 0 valori lipsa (0.0%)

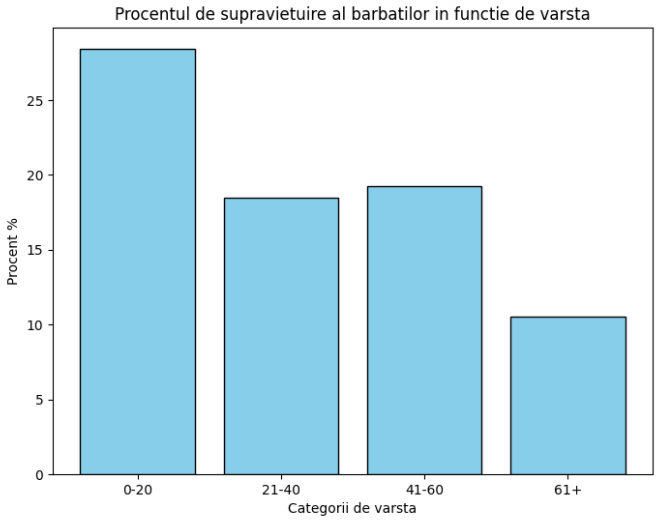
Survived = 1: 2 valori lipsa (0.58%)

5. Noua coloana de varsta este adaugata folosind un vector de intervale si unul de index pentru fiecare interval. Se face o freceventa a valorilor varstelor pentru a putea prezenta pe grafic numarul persoanelor de o anumita varsta.



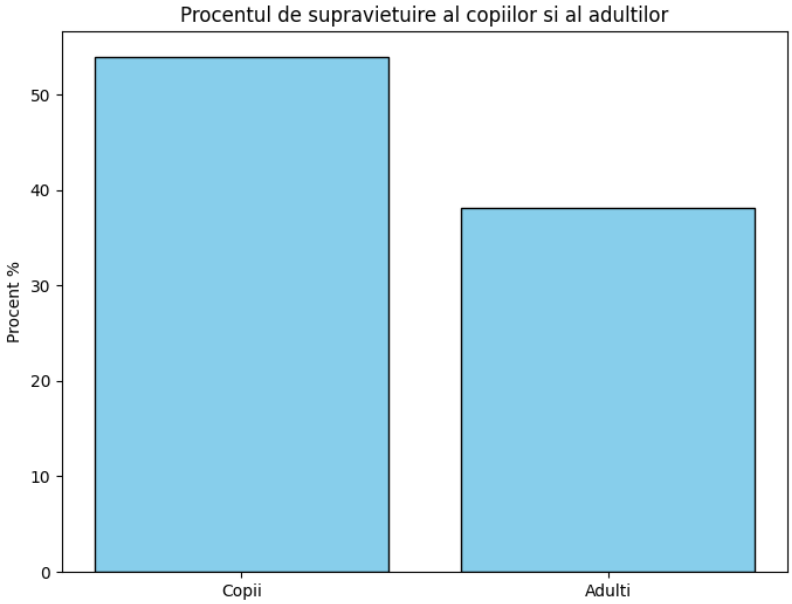
6. Procentul de supravietuire al barbatilor in functie de varsta:

28.43, 18.47, 19.28, 10.53



7. Procentul de supravietuire al copiilor: 53.98

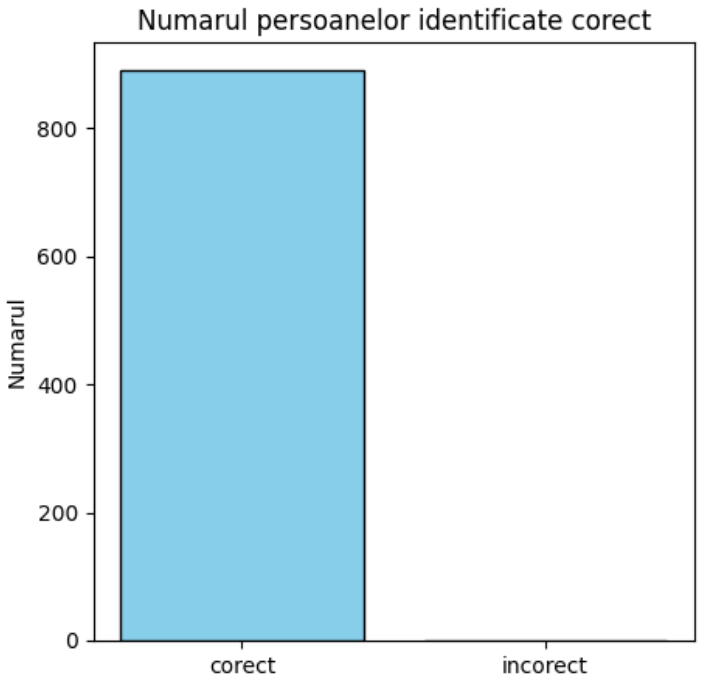
Procentul de supravietuire al adultilor: 38.1



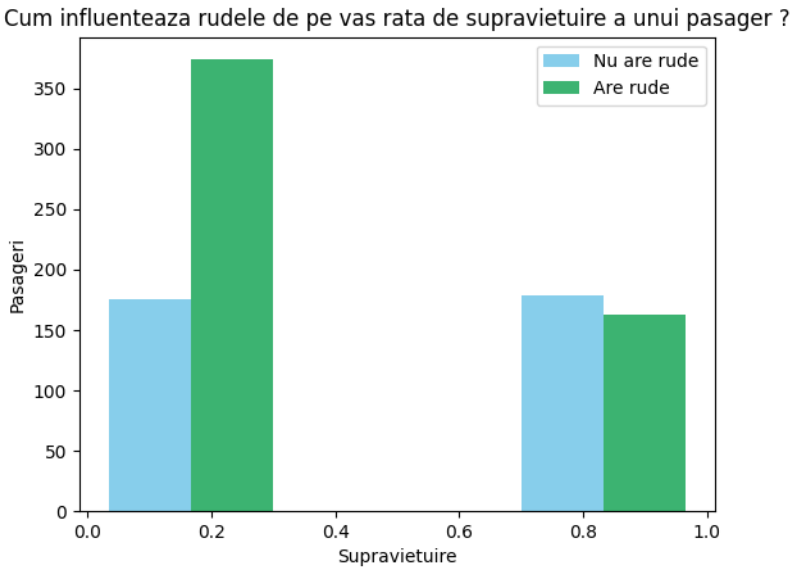
8. Se filtreaza coloanele care contin valori lipsa, si in functie de tipul de date, se umplu spatiile goale fie cu media valorilor, fie cu cea mai frecventa componenta din coloana.

9. Folosind comanda titles = data['Name'].str.split(', ').str[1].

str.split('.').str[0].unique(), am gasit toate titlurile posibile. Apoi am creat un vector de astfel de titluri si am iterat prin linii, verificand pentru fiecare persoana daca genul este corespunzator titlului. 1 singura persoana identificata incorect.

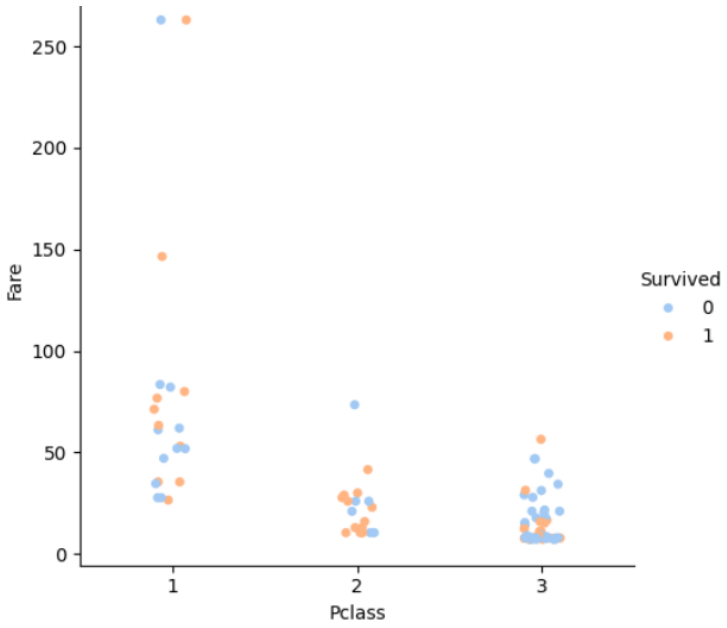


10.



Se observa o oarece corelatie intre faptul ca o persoana nu are rude si rata de supravietuire dar nu sunt destul de multe informatii in acest sens cat sa se poata decida ca este o regula.

Relatia dintre tarif, clasa si supravietuirea pasagerului



Se observa ca persoanele de clasa a 2-a au supravietuit in numar cel mai mare, in timp ce persoanele de la clasa a 3-a in numar cel mai putin.