

Proiect PCLP3

2023-2024

enunt:

https://curs.upb.ro/2023/pluginfile.php/322280/mod_resource/content/1/Enunt_Proiect_PCLP3_2024.pdf

Nume: Barbu Sebastian-Cornel

Serie: CA

Grupa: 314

link to github:

https://github.com/Sebastopol27/Proiect_Python_PCLP3

- TASK1

Pt a realiza tema am avut nevoie de biblioteca pandas

Cel mai simplu task din tema asta as putea spune, am folosit functii commune, ceea ce ar putea pune probleme totusi sunt functiile: `isnull()` `sum()` si `duplicated()`

-> `isnull()` selecteaza campurile din Data Frame care nu au valoarea, iar in combinatia cu functia `sum()`, numara cate spatii goale sunt

->`is duplicate()` face aceeaasi smecherie doar ca verifica daca exista linii duplicate si sunt numerate acestea. In dataframeul nostrum nu sunt totusi.

• TASK2

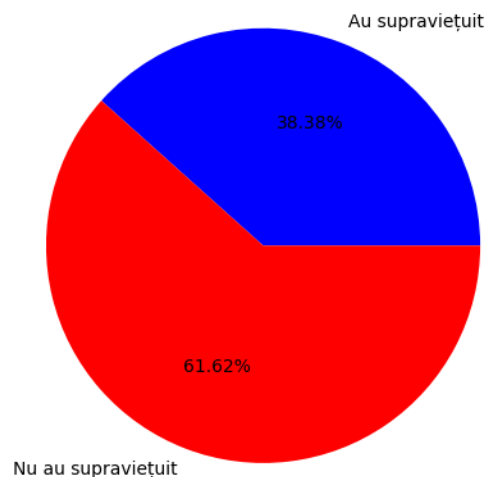
Taskul a constat in calcularea probabilitatilor cerute si in crearea unor grafice aferente.

->functia `value_counts()` am folosit o pentru a pastra informatiile despre toate tipurile de date pe care le am in fiecare coloana, apoi cu `.get()` mi am accesat datele asa cum stiam ca trebuie sa fie.

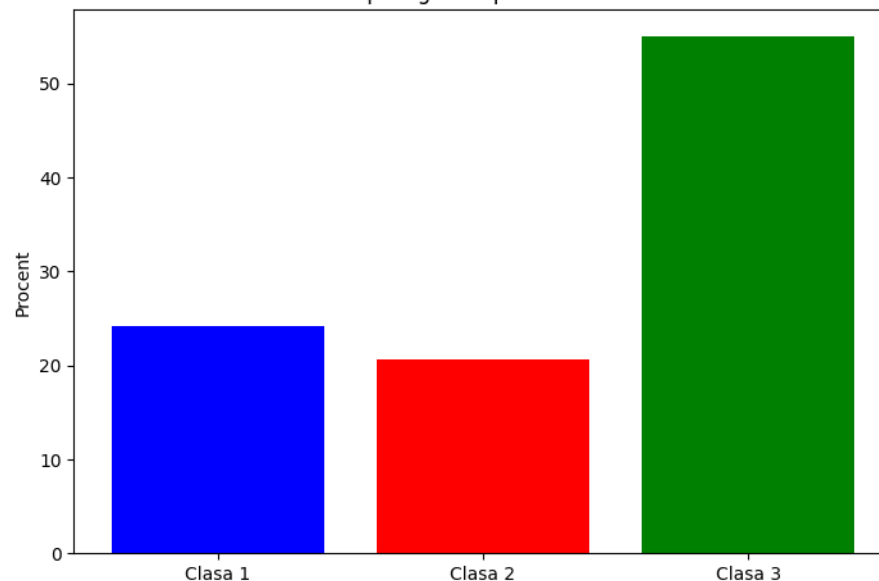
->procentele au fost calculate cu formula clasica si dupa ce am calculate toate procentele am afisat, sa vad daca au sens.

->mai apoi au fost create diagramele care arata astfel:

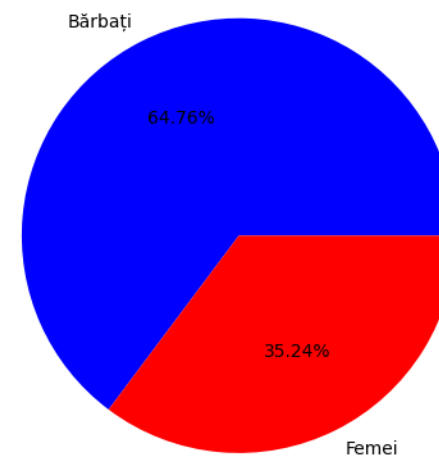
Procentul persoanelor care au supraviețuit



Procentul pasagerilor pentru fiecare clasă



Procentul bărbaților și femeilor

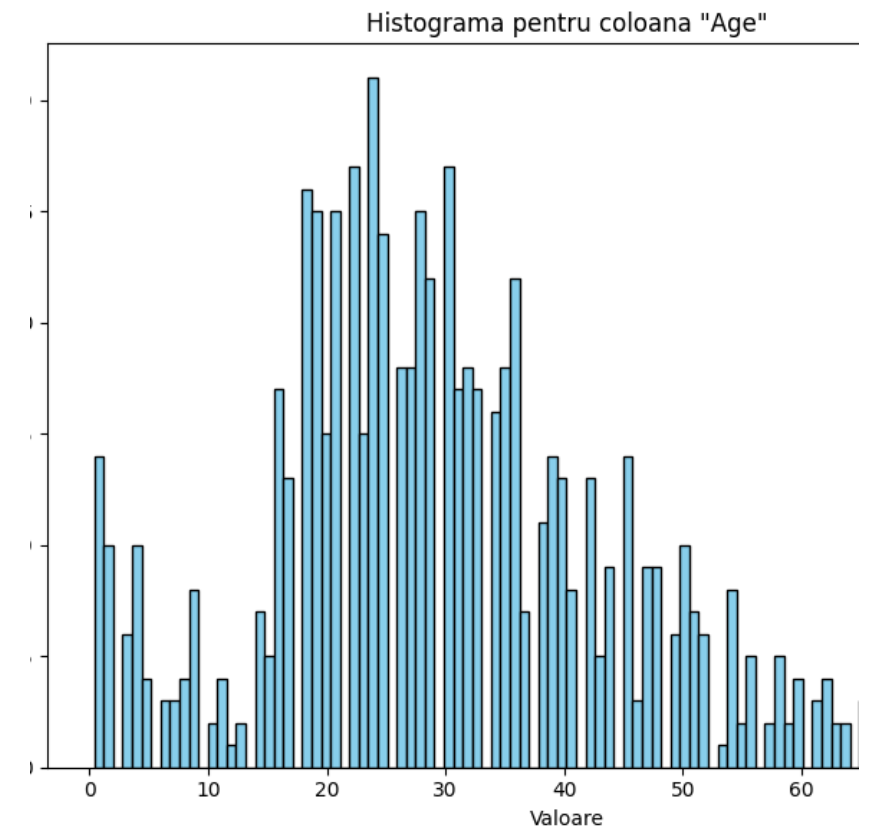


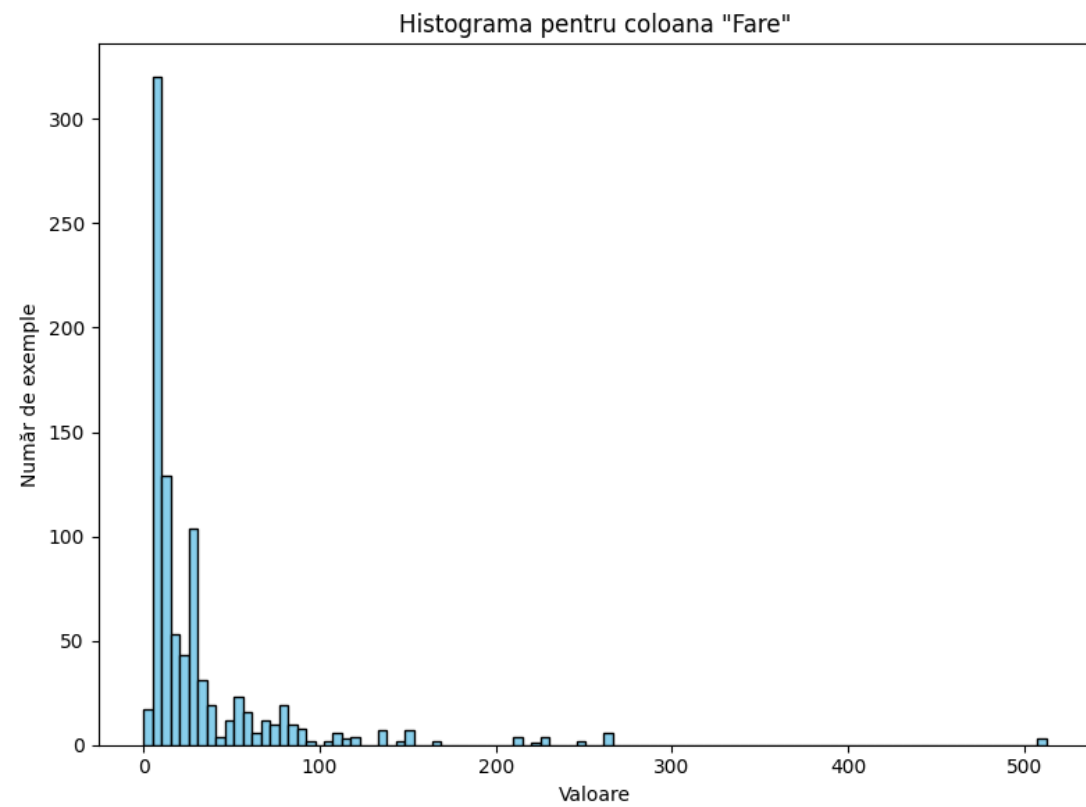
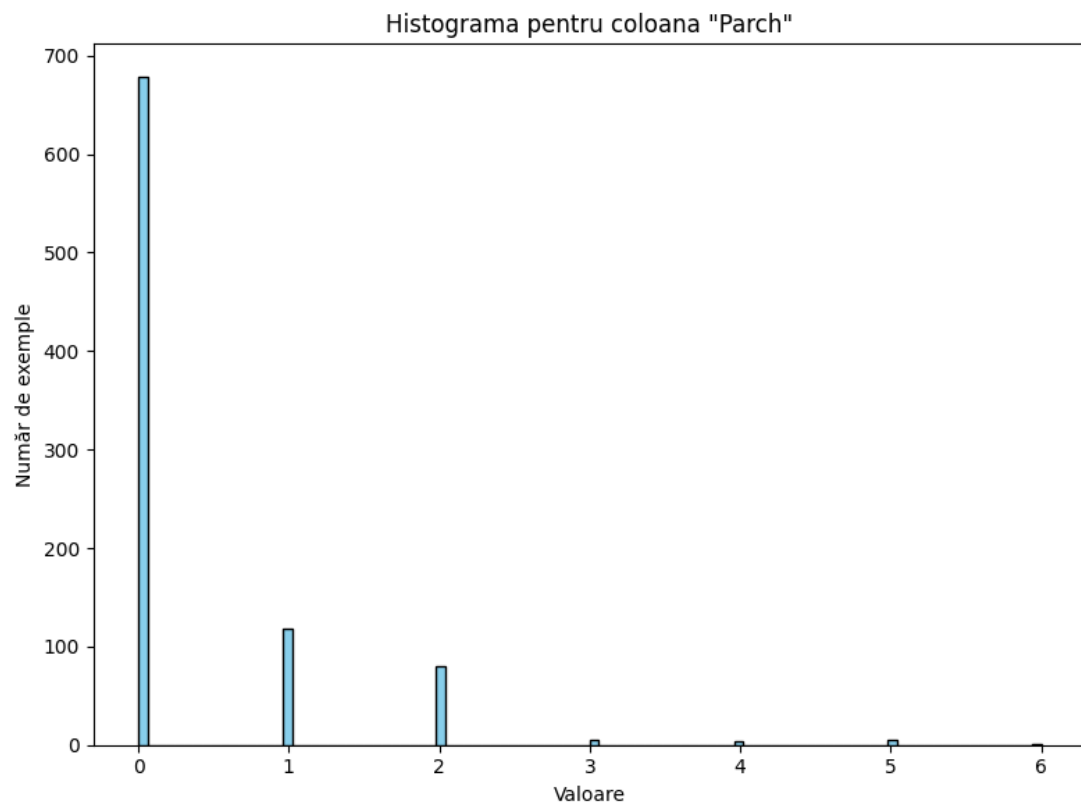
- TASK3

Un task aparent greu dar s a rezolvat destul de usor.

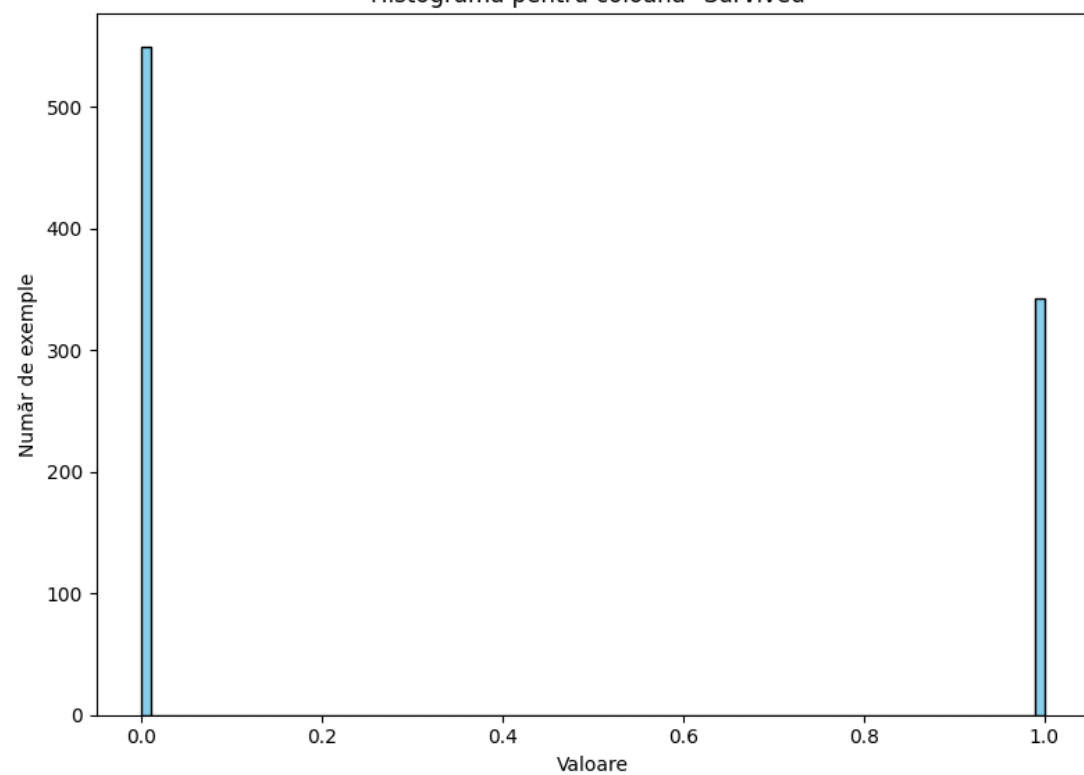
Am inceput prin a face un data frame mai mic care sa contina doar coloanele care au tipul de date numeric.

Apoi sub un for am afisat fiecare histograma, dupa numele ei. Acestea arata asa:

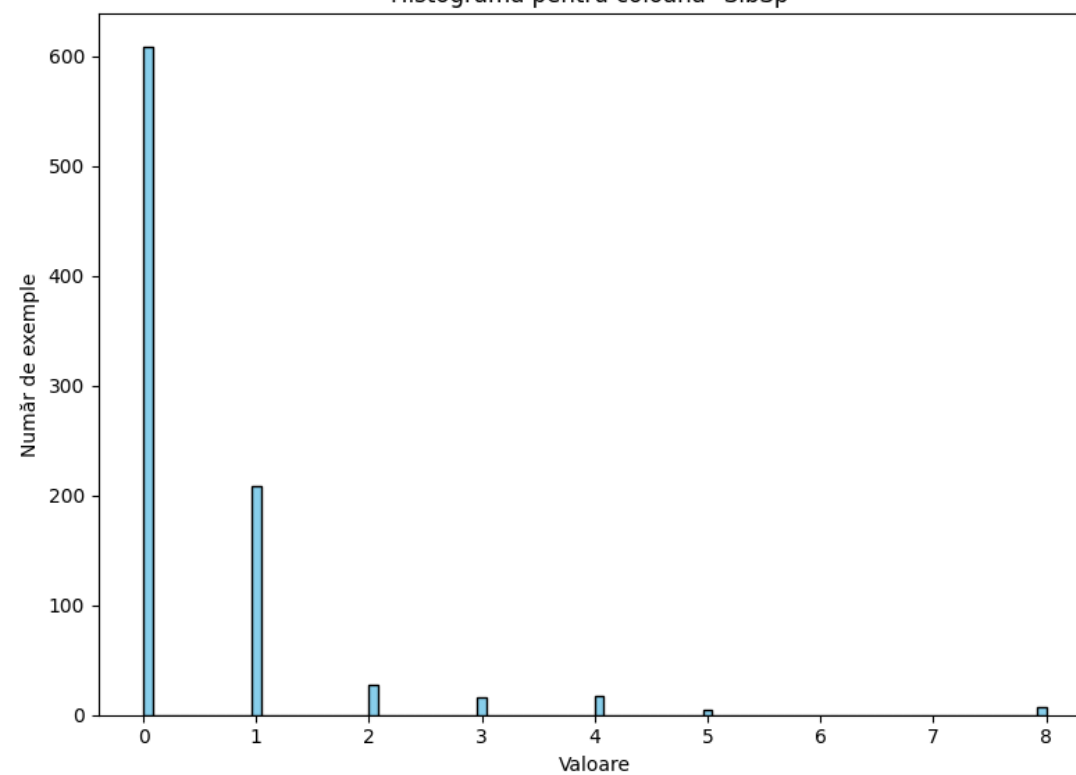




Histograma pentru coloana "Survived"



Histograma pentru coloana "SibSp"



- TASK4

Un task care porneste de la taskul 1, cautand mai intai coloanele din df care au valori lipsa.

Dupa acest lucru a trebuit sa calculam iarasi niste probabilitatile.
->ce este mai interesante este faptul ca am un vector de probabilitati pentru fiecare categorie:

```
Numărul de valori lipsă pentru fiecare coloană:
```

```
Age      177  
Cabin    687  
Embarked  2  
dtype: int64
```

```
Proportia valorilor lipsă:
```

```
[19.865319865319865, 77.10437710437711, 0.2244668911335578]
```

```
Numărul de valori lipsă pentru morti:
```

```
Age      125  
Cabin    481  
dtype: int64
```

```
Procentul lipsa pentru morti:
```

```
[14.029180695847362, 53.98428731762065]
```

```
Proportia valorilor lipsă pentru vii:
```

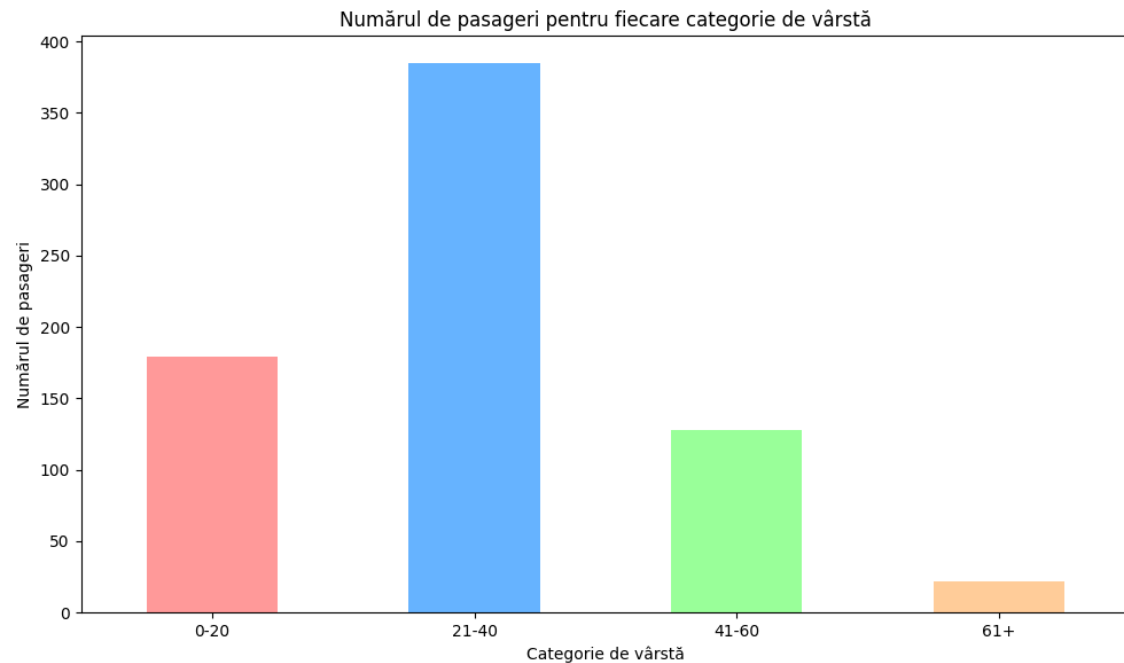
```
Age      52  
Cabin    206  
Embarked  2  
dtype: int64
```

```
Procentul lipsa pentru vii:
```

```
[5.836139169472503, 23.120089786756452, 0.2244668911335578]
```

- TASK5

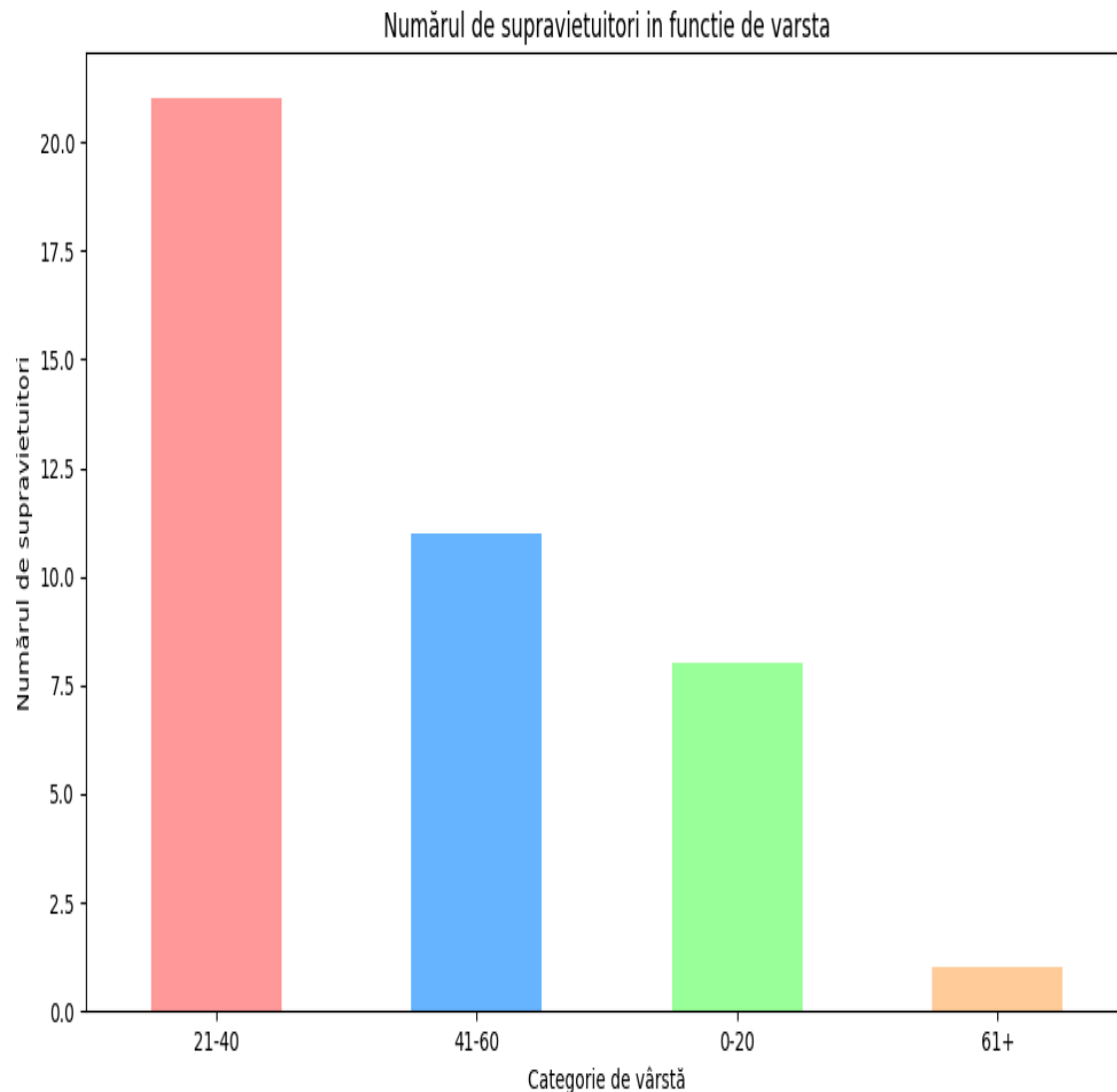
Am create o noua coloana care sa contina date cu privier la acest table. Au fost numerate categoriile de varsta sii dupa am create tabelul si modificat fisiercul csv astfel incat sa apara si aceasta.



- TASK6

A constat in limitarea tabelului doar la persoanele care au supravietuit iar mai apoi doar la barbati. Dupa aceea am numarat in functie de varsta si am afisat tabelul.

Acesta arata faptul ca barbatii cu varste intre 21 si 40 de ani au supravietuit cel mai mult, iar cei in varsta mai toti au murit.



- TASK7

O alta analiza interesanta a fost sa impartim oamenii intre copii(<18) si adulti(restul).

Am facut o coloana care are toti copii iar mai apoi am scos toti copii care au supravietuit si am facut acelasi lucru si pentru adulti si dupa am create aceste procente care arata cati copii si adulti au supravietuit

