Implementarea temei a durat, in total, 4 zile. Am ales sa fac partea I, deoarece am considerat-o mai realizabila intr-un timp asa scurt.

**Cerinta 1**

Am citit datele, apoi am afisat numarul de coloane, tipul datelor din fiecare coloana, numarul de valori lipsa pentru fiecare coloana, numarul de linii si am verificat daca exista linii duplicate:

Numarul de coloane: 12

Tipurile datelor din 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

Numarul de valori 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

Numarul de linii: 891

Exista linii duplicate? Nu

**Cerinta 2**

Am calculat procentul persoanelor care au supravietuit si procentul persoanelor care nu au supravietuit, procentul pasagerilor pentru fiecare tip de clasa, procentul barbatilor si procentul femeilor, iar apoi am facut un grafic in care sunt prezentate rezultatele:

Procentul persoanelor care au supravietuit: 38.38%

Procentul persoanelor care nu au supravietuit: 61.62%

Procentul pasagerilor pentru fiecare tip de clasa:

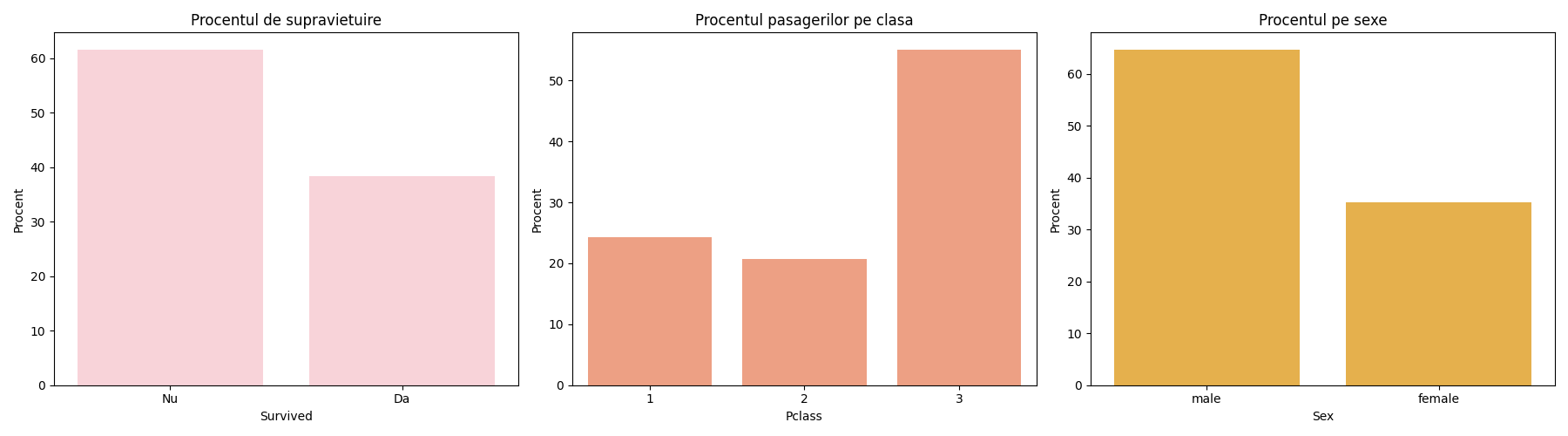
Clasa 1: 24.24%

Clasa 2: 20.65%

Clasa 3: 55.11%

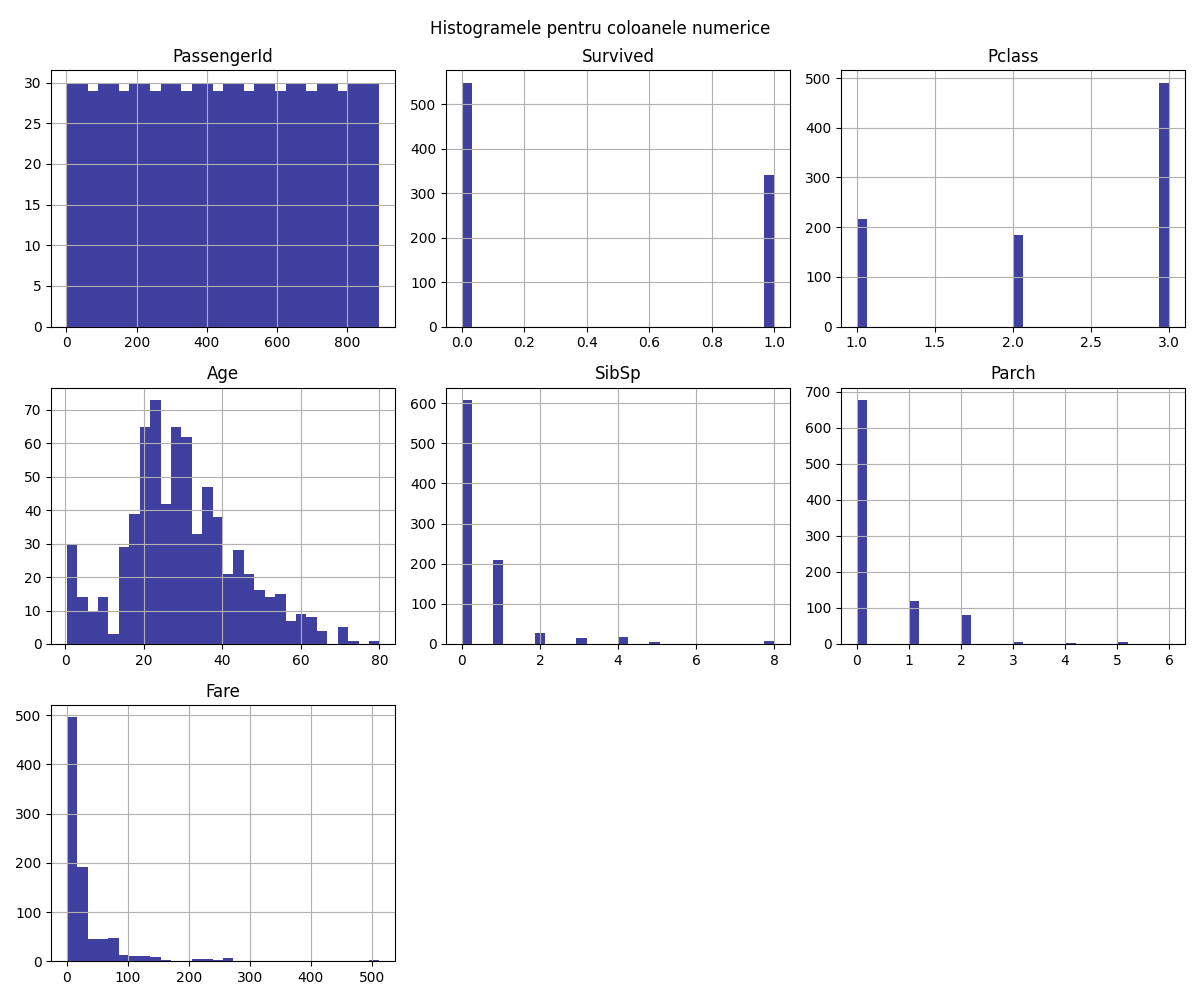
Procentul barbatilor: 64.76%

Procentul femeilor: 35.24%



**Cerinta 3**

Am generat 7 histograme, una pentru fiecare coloana cu valori numerice (int64 / float64). Pe axa OX sunt incluse intervalele de valori ale variabilei, iar pe axa OY e reprezentat numarul de exemple din setul de date care sunt incluse in fiecare interval:



**Cerinta 4**

Am identificat coloanele pentru care exista valori lipsa, iar pentru fiecare coloana identificata, am determinat numarul si proportia valorilor lipsa. Dupa, am determinat procentul acestora pentru fiecare dintre cele 2 clase ale coloanei Survived (0, 1):

Coloanele pentru care exista valori lipsa:

Age

Cabin

Embarked

Numarul si procentul valorilor lipsa pentru fiecare coloana:

Coloana Age: 177 valori lipsa, in proportie de 19.87%

Coloana Cabin: 687 valori lipsa, in proportie de 77.10%

Coloana Embarked: 2 valori lipsa, in proportie de 0.22%

Procentul valorilor lipsa pentru fiecare dintre cele doua clase (coloana Survived):

Clasa 0: 22.77% pentru coloana Age

Clasa 1: 15.20% pentru coloana Age

Clasa 0: 87.61% pentru coloana Cabin

Clasa 1: 60.23% pentru coloana Cabin

Clasa 0: 0.00% pentru coloana Embarked

Clasa 1: 0.58% pentru coloana Embarked

**Cerinta 5**

Am stabilit categoriile de varsta si am determinat cati pasageri avem pentru fiecare in parte. Am introdus o coloana suplimentara, Index, apoi am determinat pentru fiecare exemplu din setul de date, indexul categoriei din care face parte (1 / 2 / 3 / 4). Am realizat un grafic pentru a evidentia rezultatele. Informatiile noi, cu tot cu noua coloana adica, le-am salvat in train1.csv, atasat in subdirectorul Date:

Numarul de pasageri pentru fiecare categorie de varsta:

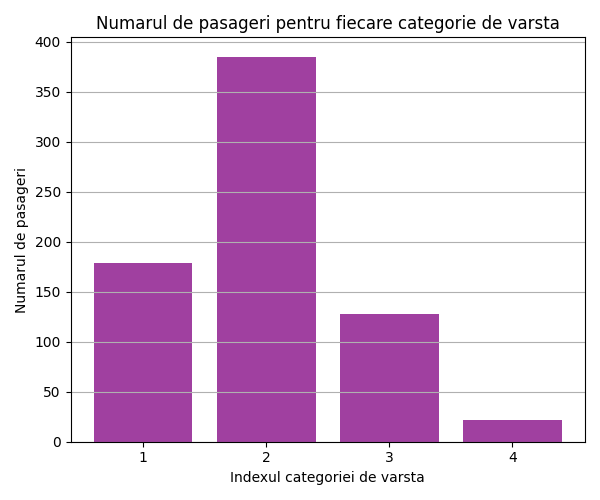
0-20 179

21-40 385

41-60 128

61+ 22

Name: count, dtype: int64



**Cerinta 6**

Am calculate cati barbati au supravietuit pentru fiecare dintre cele 4 categorii de varsta si am realizat un grafic in care am evidentiat influenta varstei aupra procentului de supravietuire a barbatilor:

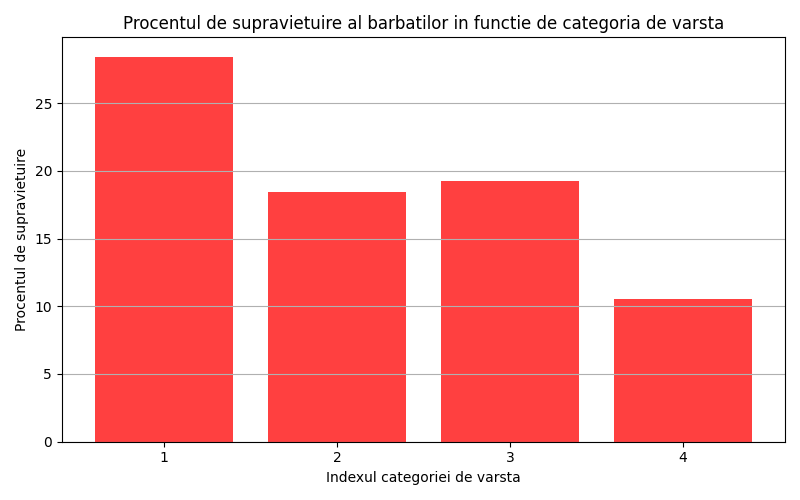
Numarul de barbati care au supravietuit pentru fiecare dintre cele 4 categorii de varsta:

Categoria 1.0: 29 barbati

Categoria 2.0: 46 barbati

Categoria 3.0: 16 barbati

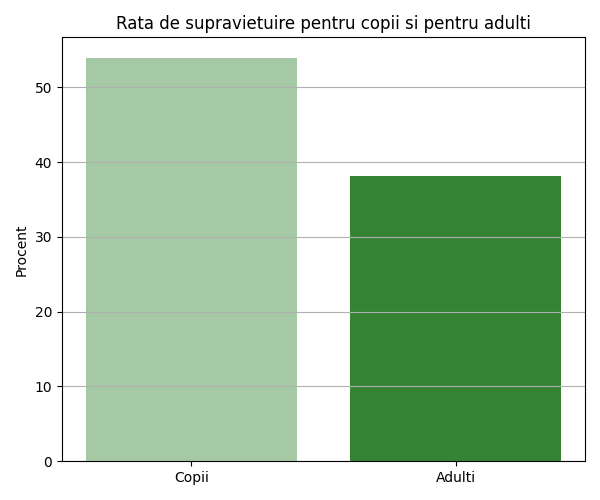
Categoria 4.0: 2 barbati



**Cerinta 7**

Am calculate procentul copiilor aflati la bord si am realizat un grafic care evidentiaza rata de supravietuire pentru copii si pentru adulti:

Procentul copiilor aflati la bord: 12.68%

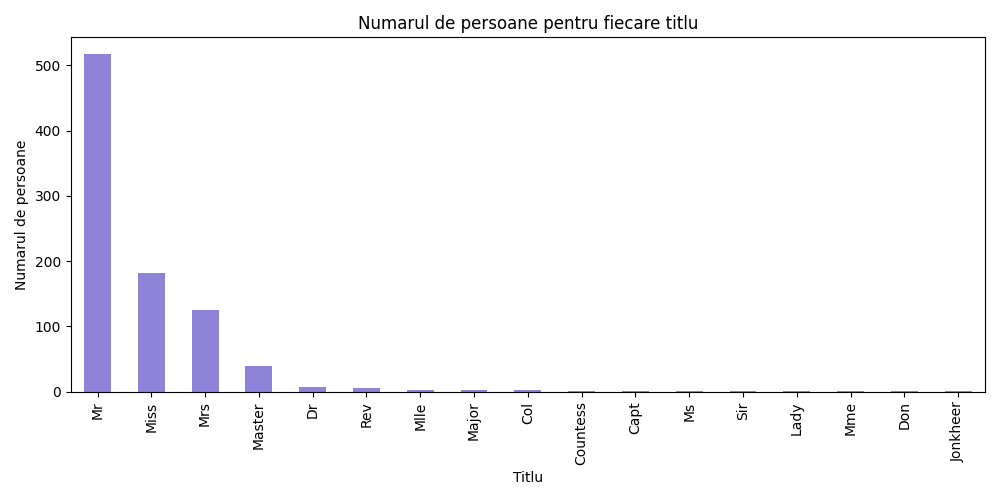


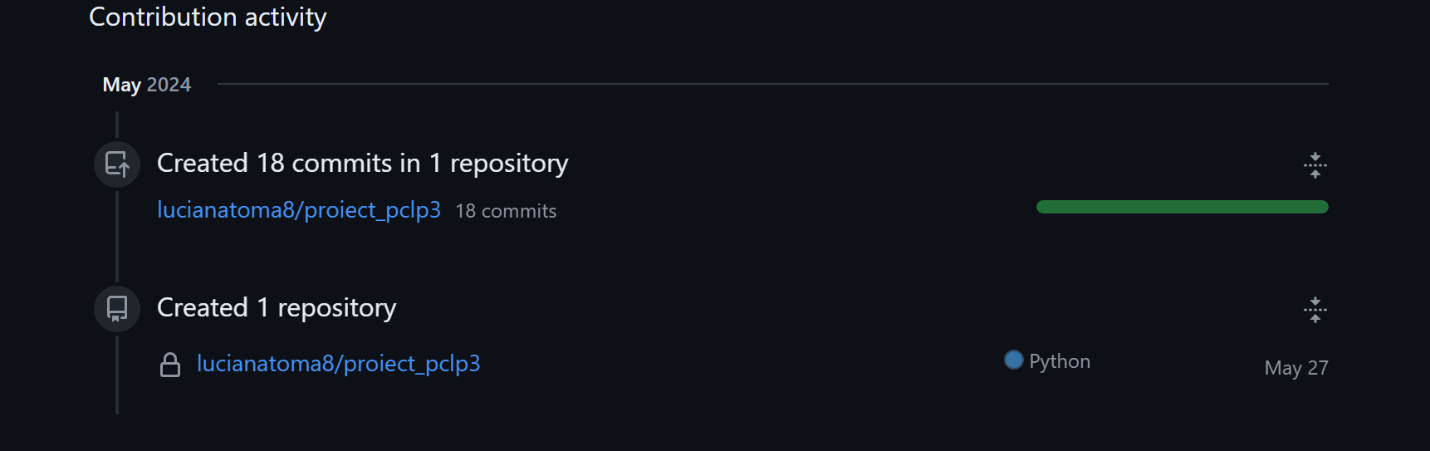
**Cerinta 8**

Am completat valorile lipsa cu cele obtinute pentru media pasagerilor care fac parte din aceeasi clasa. Valori lipsa existau doar pe coloanele Age, Cabin si Embarked. Pentru un pasager care supravietuieste, dar caruia nu i se cunoaste varsta, am completat varsta cu media pasagerilor care au supravietuit si invers. In cazul coloanelor cu valori categoriale(Cabin / Embarked), am determinat cea mai frecventa valoare pentru respectiva clasa. In subdirectorul Date, train1.csv nu are nimic completat la Index pentru persoanele fara varsta, iar in train2.csv, am sters de tot coloana Index.

**Cerinta 9**

Am verificat daca titlurile de noblete regasite in coloana Name corespund cu sexul persoanei respective si am reprezentat grafic cate persoane corespund fiecarui titlu. Pentru asta, am extras titlurile, le-am mapat cu sexul corespunzator pe o noua coloana, Titlul\_Regasit, si am facut verificarea. Am creat o noua coloana, Titlul\_Potrivit, care indica daca titlul corespunde sexului persoanei sau nu si am salvat rezultatul in verificare\_titluri.csv:



 Am folosit utilitarul git de-a lungul timpului in c are am implementat tema.

