README
Maxim Anca Stefania, 314CC

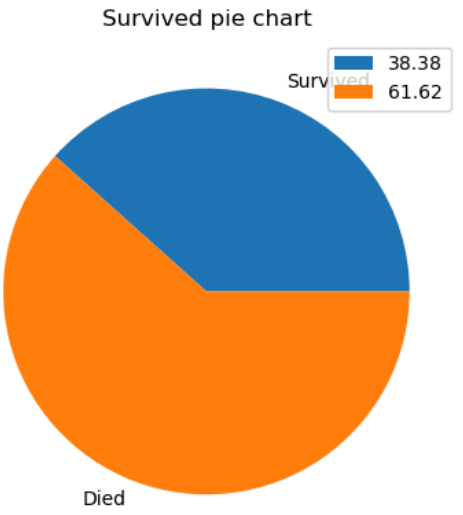
Git: https://github.com/theonlytruealex/PCLP3/tree/main_I
Partea I se afla pe branch-ul main_I.

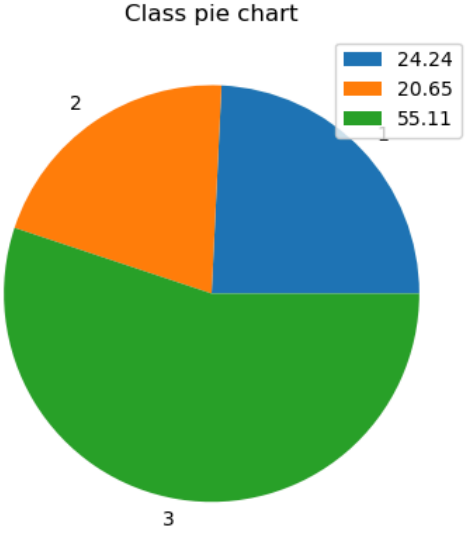
Cerința 1:

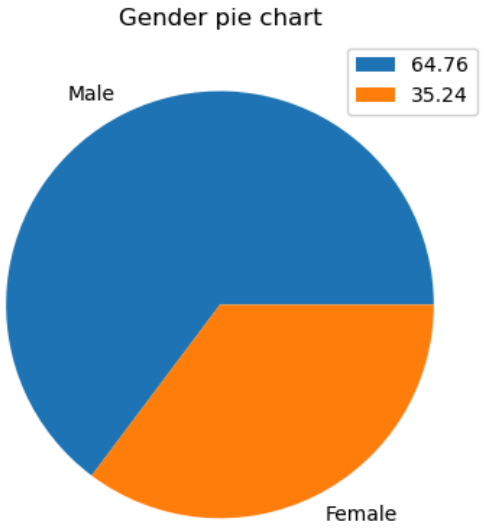
Am determinat, conform cerintei, numărul de coloane, tipul lor, numărul de spații din fiecare coloana, numărul de linii și numărul de linii duplicate.

Cerința 2:

Am determinat separat numărul de campuri care aveau valoarea 1 pe coloana Survived și calculat procentul aferent. Pentru fiecare tip de clasa (1, 2, 3), am aflat numărul și procentul pasagerilor corespunzatori clasei respectiva (aplicand aceeasi metoda ca pt Survived). Similar, am calculat numărul și procentul de bărbați și femei aflați la bord. Graficele rezultate sunt:

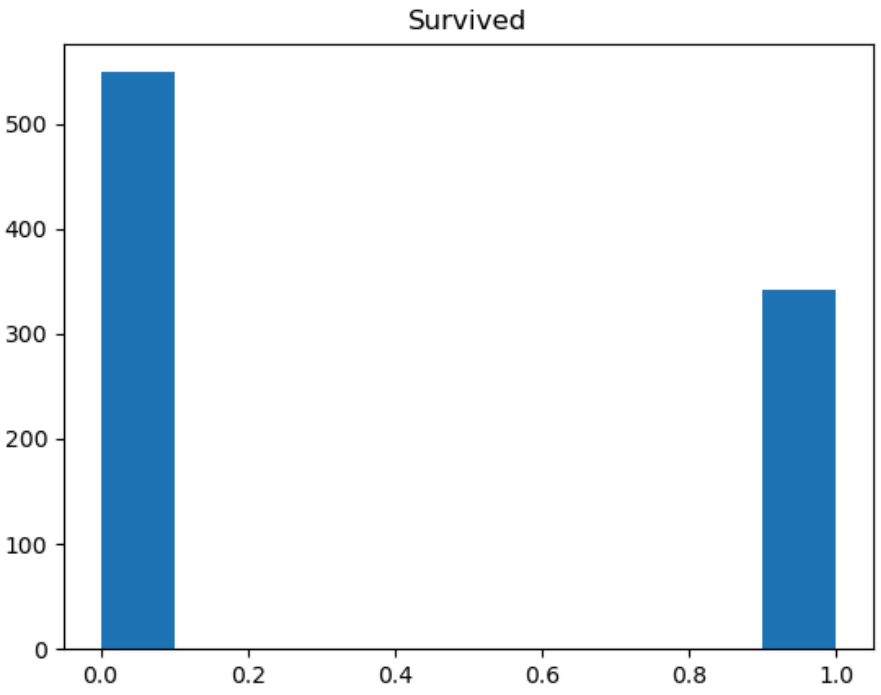


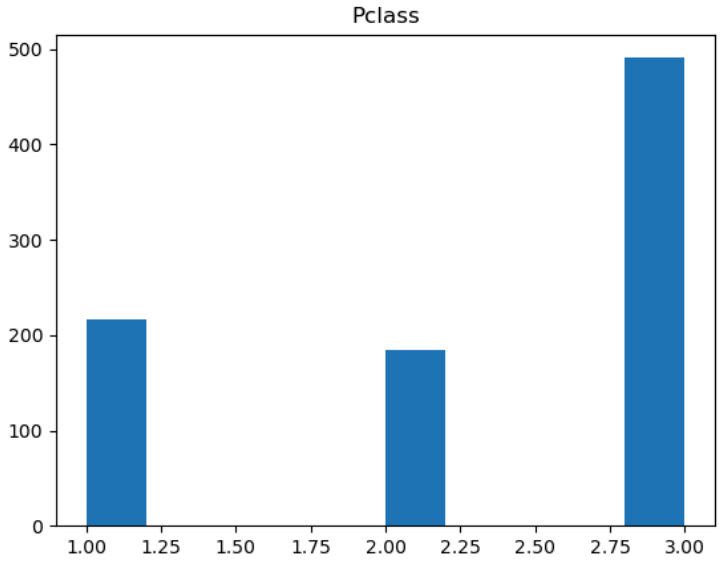


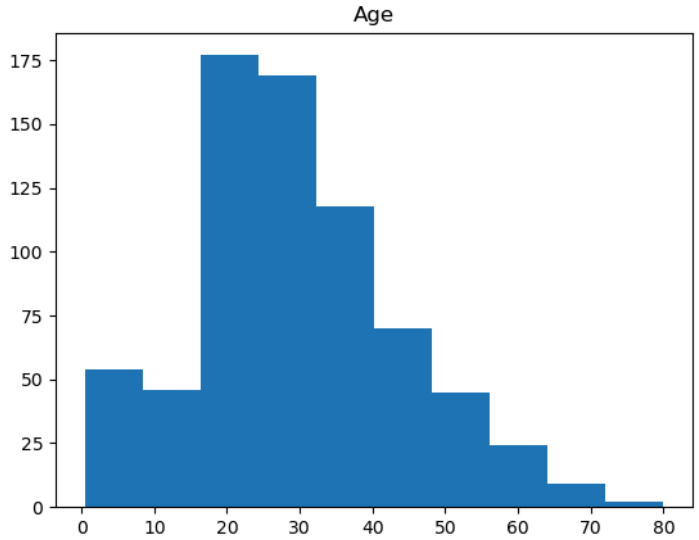


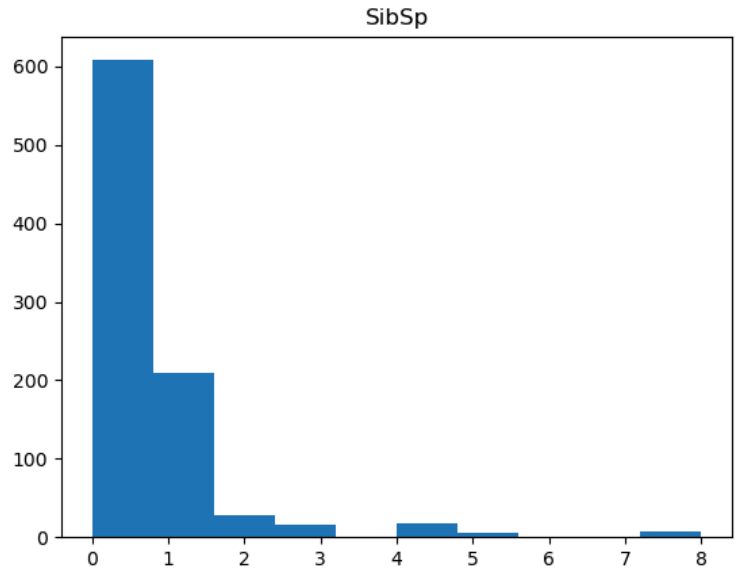
Cerința 3:

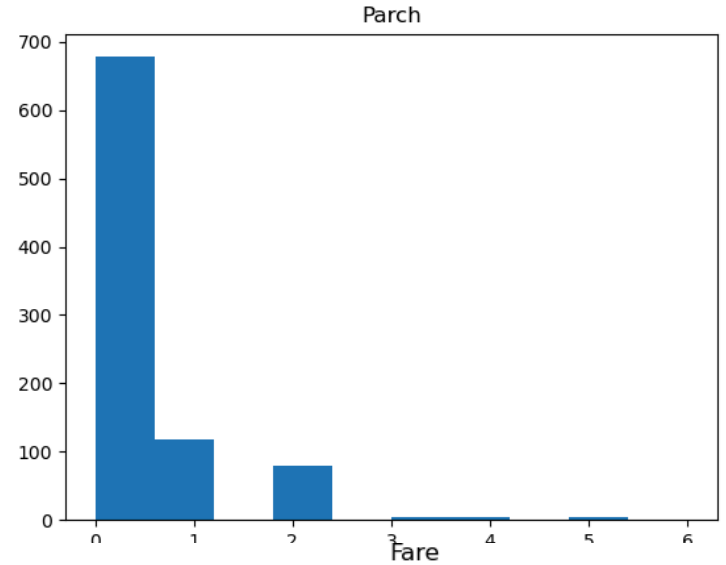
Am preluat doar coloanele cu date numerice din fișierul train.csv, asemanator , și am realizat cate o histograma pentru fiecare

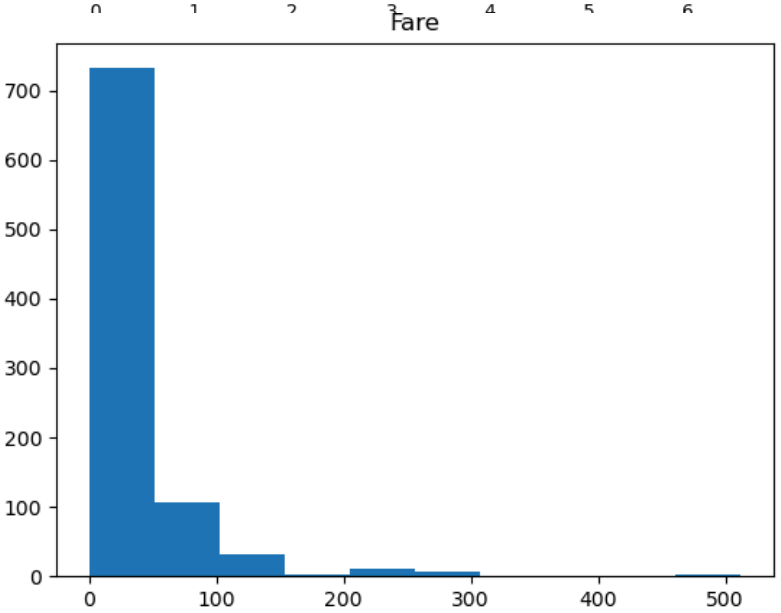


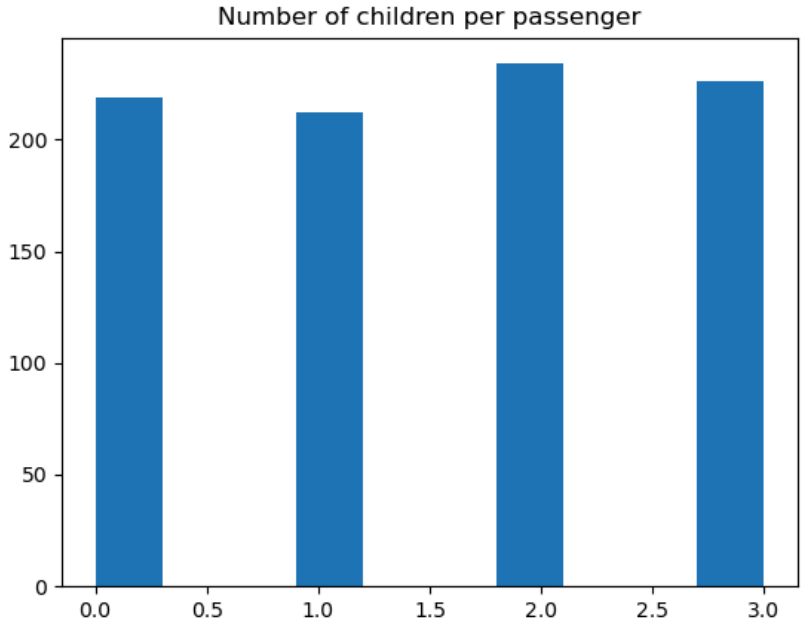

Pclass
500
400
300
200
100
0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Age
175
150
125
100
75
50
25
0
0
10
20
30
40
50
60
70
80


SibSp
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8


Parch
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6


Fare
700
600
500
400
300
200
100
0
0
100
200
300
400
500

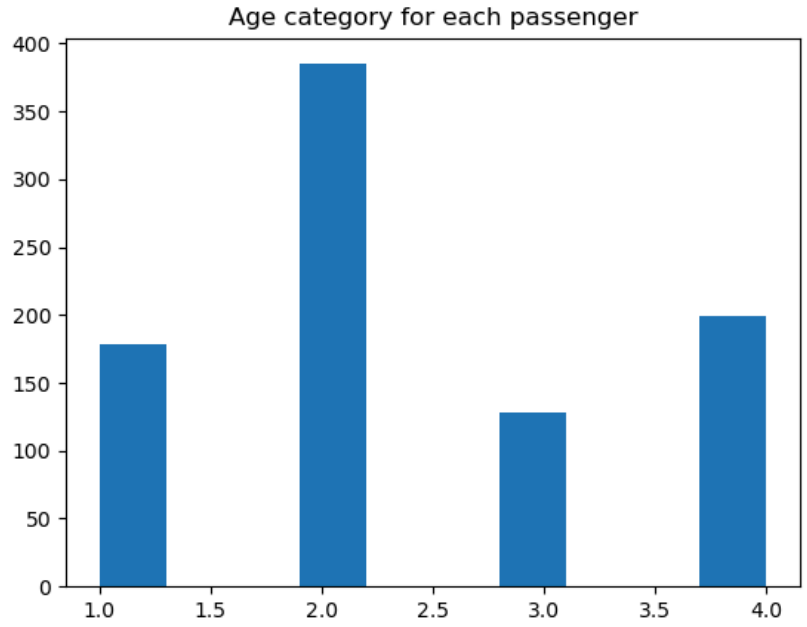

Number of children per passenger
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Cerința 4:

Am selectat coloanele cu date lipsa folosind functia gaps (creata de mine), unde am construit o lista cu coloanele care contin date lipsa, in urma folosirii funcției isnull() și am calculat numărul de date lipsa si procentajul acestora, pentru fiecare coloana obtinuta. Am apelat în mod repetat funcția gaps pentru a obține datele aferente pentru pasagerii supraviețuitori și nesupraviețuitori.

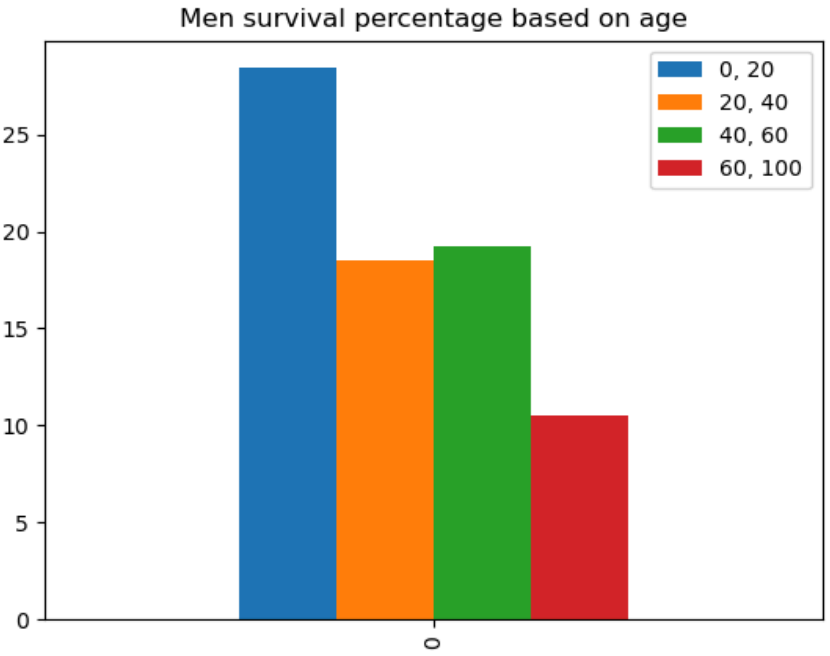
Cerința 5:

Am calculat numărul de persoane aparținând fiecărei categorii de varsta, după care am construit coloana suplimentară, unde, iterand prin fiecare rand al DataFrame ului, am determinat indexul corespunzător categoriei de varsta. La final, am afisat histograma corespunzătoare indecsilor obtinuti anterior.



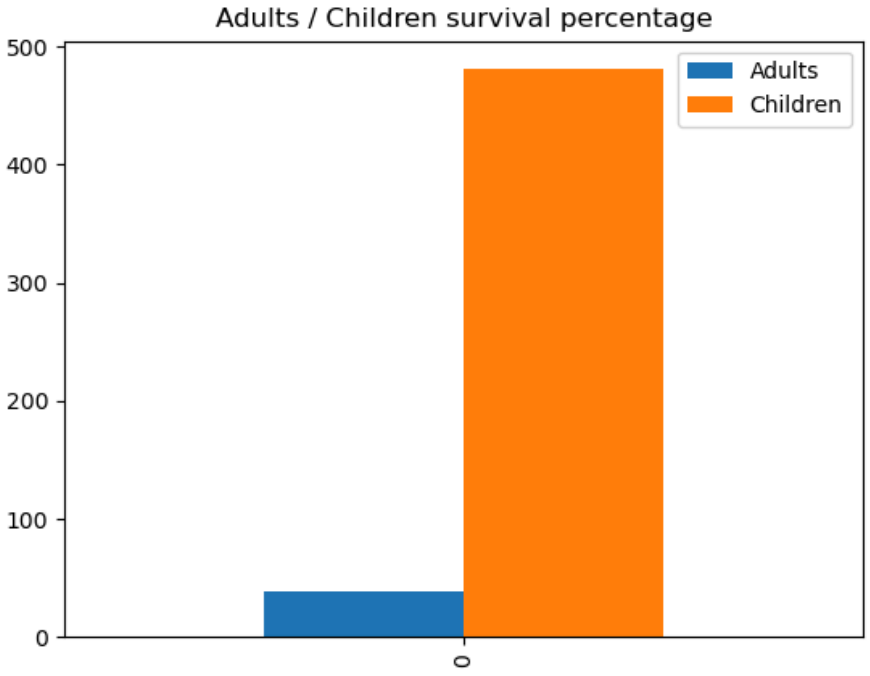
Cerința 6:

Folosind rezultatele obținute anterior, am calculat numărul bărbaților supraviețuitori din fiecare categorie. Pentru fiecare dintre cele 4 valori obținute, am creat un nou dataFrame în care am asociat numărul obtinut cu intervalul de varsta din care face parte și am reprezentat rezultatele sub forma unui grafic de tip ‘bar’.



Cerința 7:

Am calculat numărul si procentul copiilor aflați la bord si al copiilor supraviețuitori, dupa care, numărul si procentul adultilor si al adulților supraviețuitori. Similar cu Cerința 6, am construit un nou dataFrame pentru a reprezenta datele obtinute, mapand procentele cu categoria din care fac parte (Adults / Children) si le-am reprezentat sub forma de grafic de tip ‘bar’.



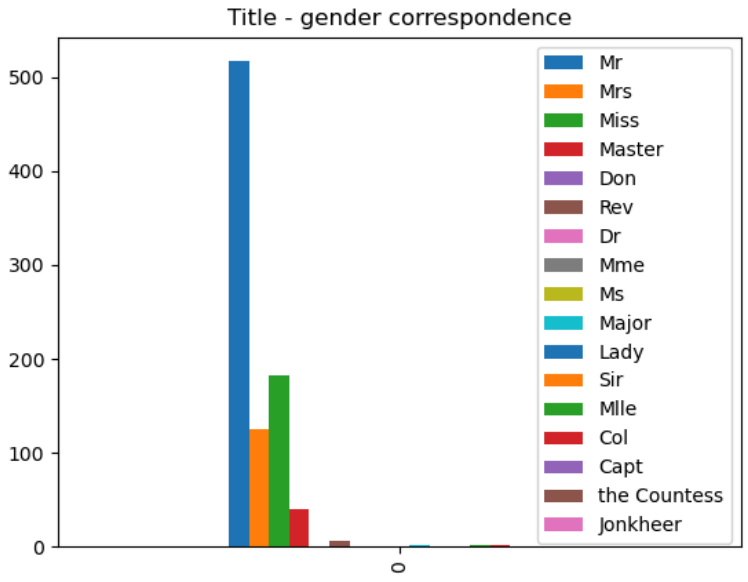
Cerința 8:

Pentru a obține golurile din fiecare coloana, iterez prin fiecare coloana, preiau golurile cu metodele .isnull() și .any() (isnull() imi afiseaza o serie de valori bool, aleg doar seria care are măcar o valoare True cu .any()). Pentru fiecare pasager determinat anterior, calculez media valorilor pasagerilor care aparțin aceleiași clase si înlocuiesc rezultatul in poziția valorii lipsa (o localizez folosind metoda .isna()). Asemanator coloanelor cu valori numerice, fac si pentru cele cu valori categoriale, folosind metoda .mode() (care imi returnează cel mai frecvent rezultat).

Salvez noul dataFrame intr-un fișier csv.

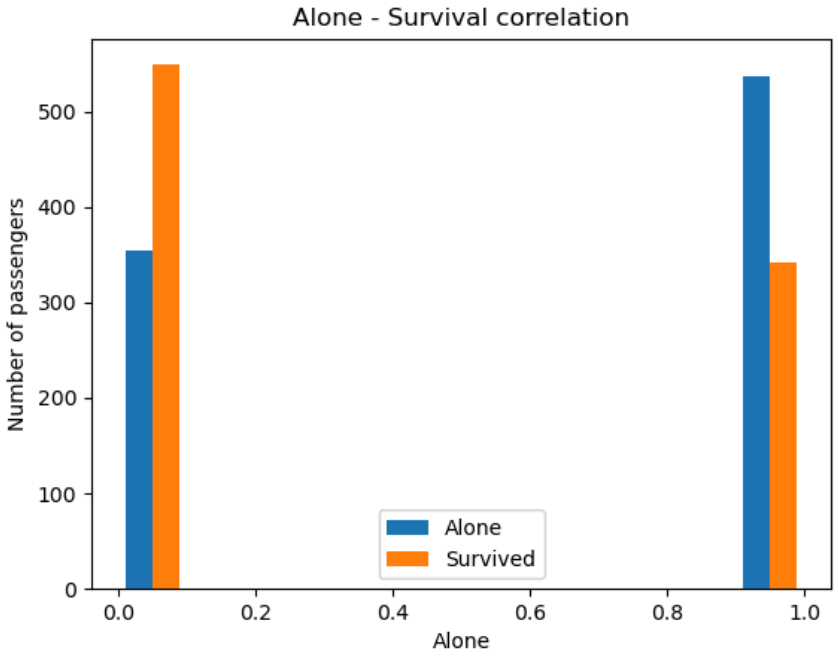
Cerința 9:

Pentru a extrage toate titlurile posibile, am folosit metoda split in mod repetat asupra coloanei Name din dataFrame. După ce am determinat titlurile, le-am mapat cu sexul corespunzător (manual). Pentru fiecare coloana din setul de date, am verificat corespondență si am păstrat numărul acestora într-un dataFrame. Graficul rezultat este:



Cerința 10:

Preiau doar pasagerii care nu au rude la bord (coloanele SibSp si Parch au valorile 0) si creez un nou dataFrame in care pun si valorile de pe coloana Survived. Am reprezentat corelatia celor doua valori prin urmatoarea histograma:



Pentru primele 100 de valori din fisierul train.csv, evidentiez corelatia valorilor de pe coloanele Fare, PClass si Survival printr-un grafic de tip catplot, astfel: