##314CA Pal Roberto-Giulio

Partea 1

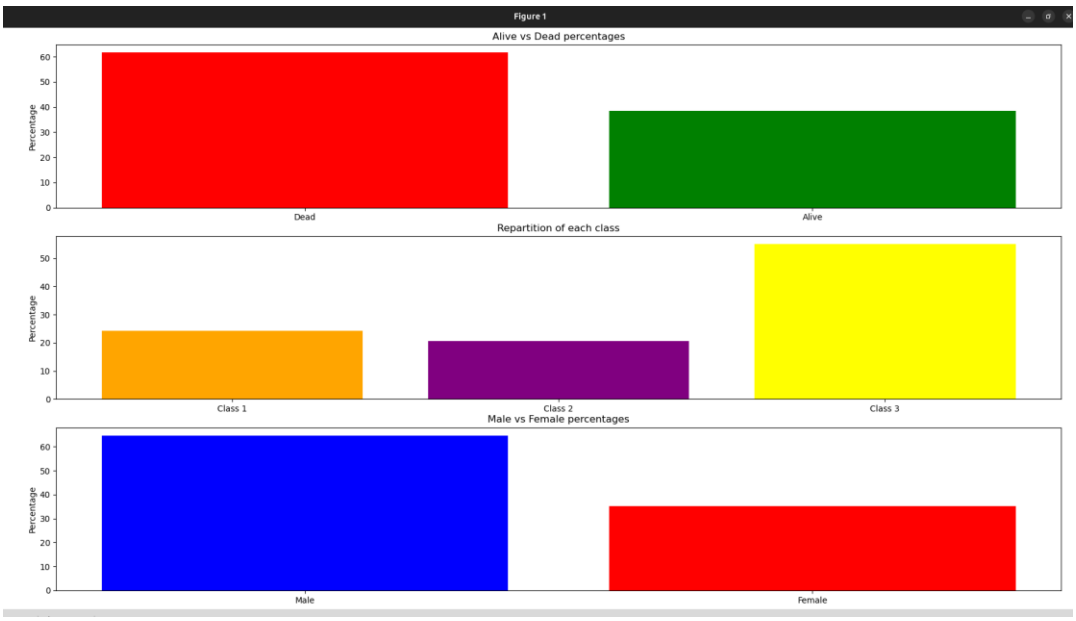
<Task 1>

In acet task vom afisa date despre fisierul train.csv folosindu-ne de functiile oferite de libraria pandas. Functia shape ne ofera atat numarul de linii cat si numarul de coloane.

Functia dtypes() ne ofera tipul de date al fiecarei coloane a fisierului. Functia duplicated() ne ofera liniile care apar de mai multe ori in data set-ul nostru.

<Task 2>

In acest task vom examina date referitoare la: procentul de persoane care au supravietuit vs celor care au murit, repartizarea oamenilor in cele 3 clase, si procentul de femei vs cel al barbatilor aflati la bord. Iata rezultatele obinute:



Din aceste grafice observam ca majoritatea oamenilor nu a supravietuit, majoritarea oamenilor era repartizata la clasa a 3-a, iar majoritatea pasagerilor era formata din barbati.

<Task 3>

In acest task vom realiza o histograma pentru fiecare coloana numerica, pentru a observa distributia datelor persoanelor aflate la bordul Titanicului. Iata rezultatele obtinute:

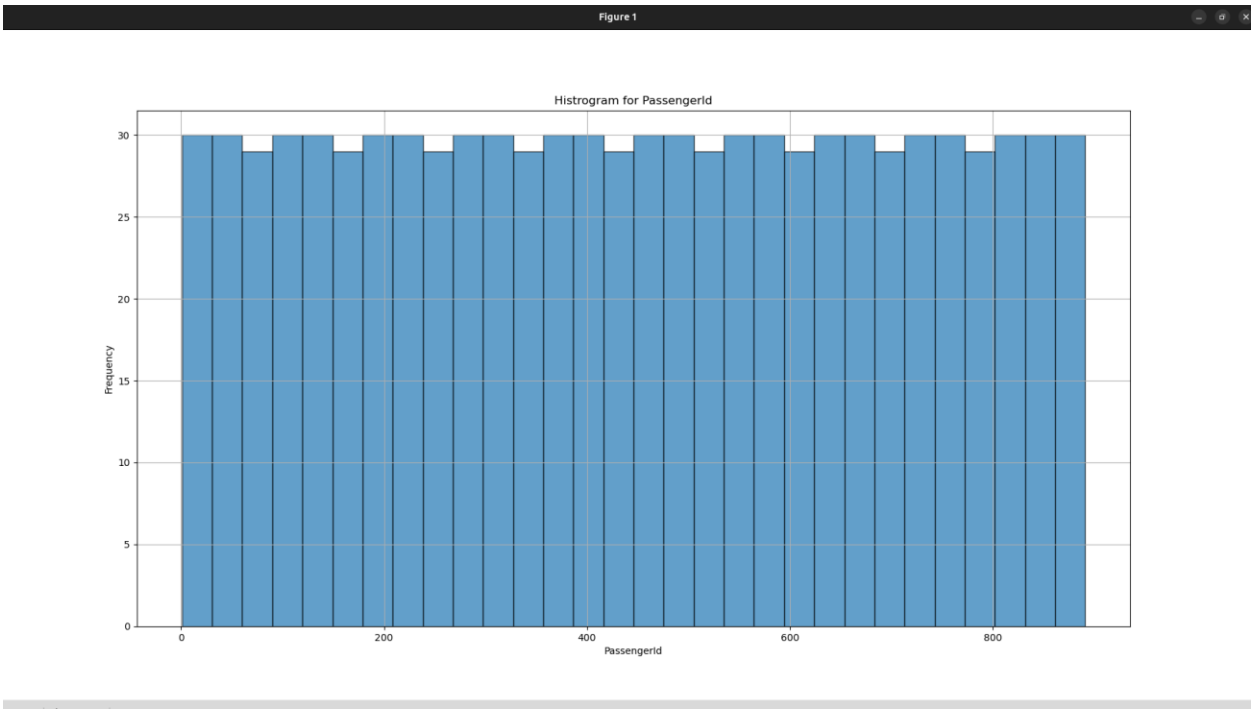


Figure 1
Histrogram for PassengerId
Frequency
PassengerId
0
5
10
15
20
25
30
0
200
400
600
800

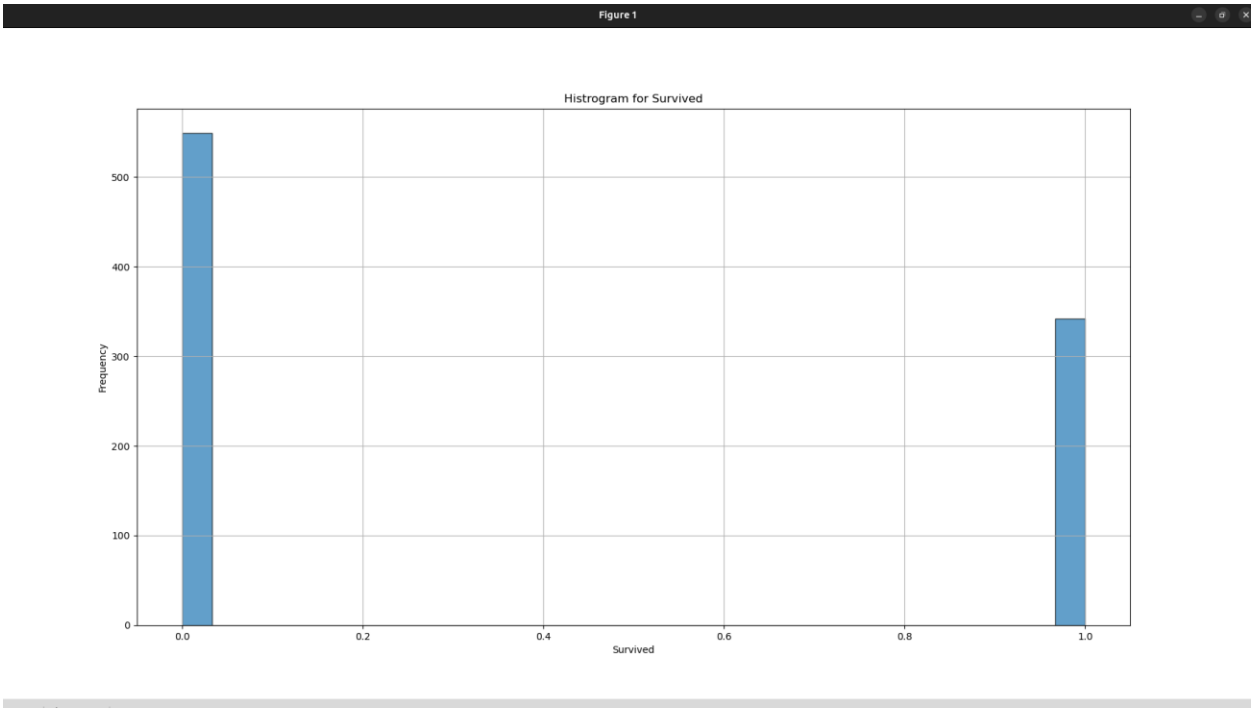


Figure 1
Histrogram for Survived
Frequency
Survived
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0

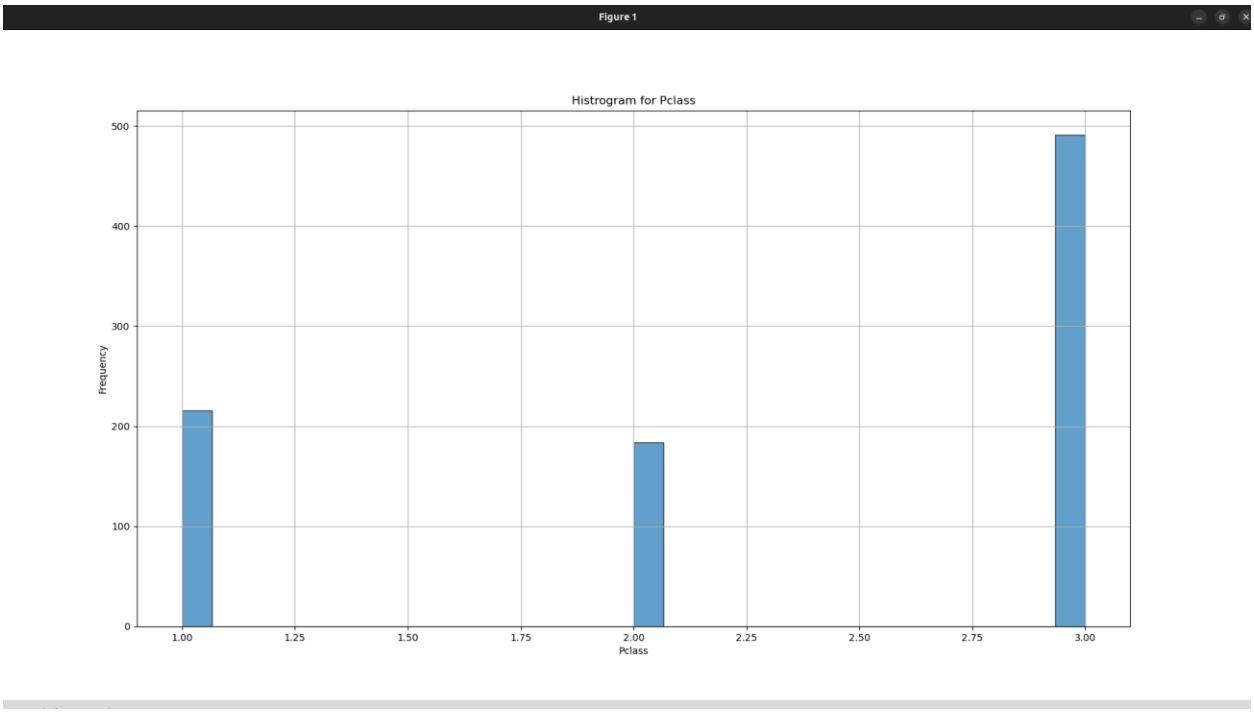


Figure 1
Histrogram for Pclass
Frequency
Pclass
0
100
200
300
400
500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

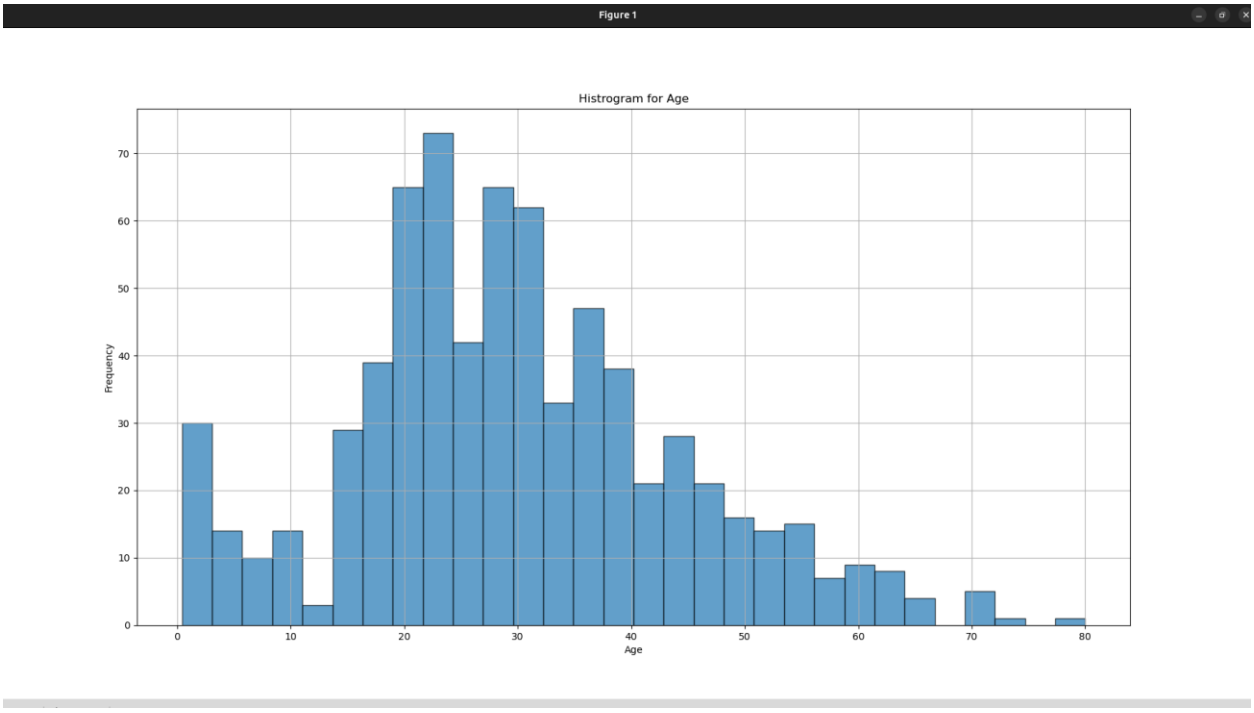


Figure 1
Histrogram for Age
Frequency
Age
0
10
20
30
40
50
60
70
80

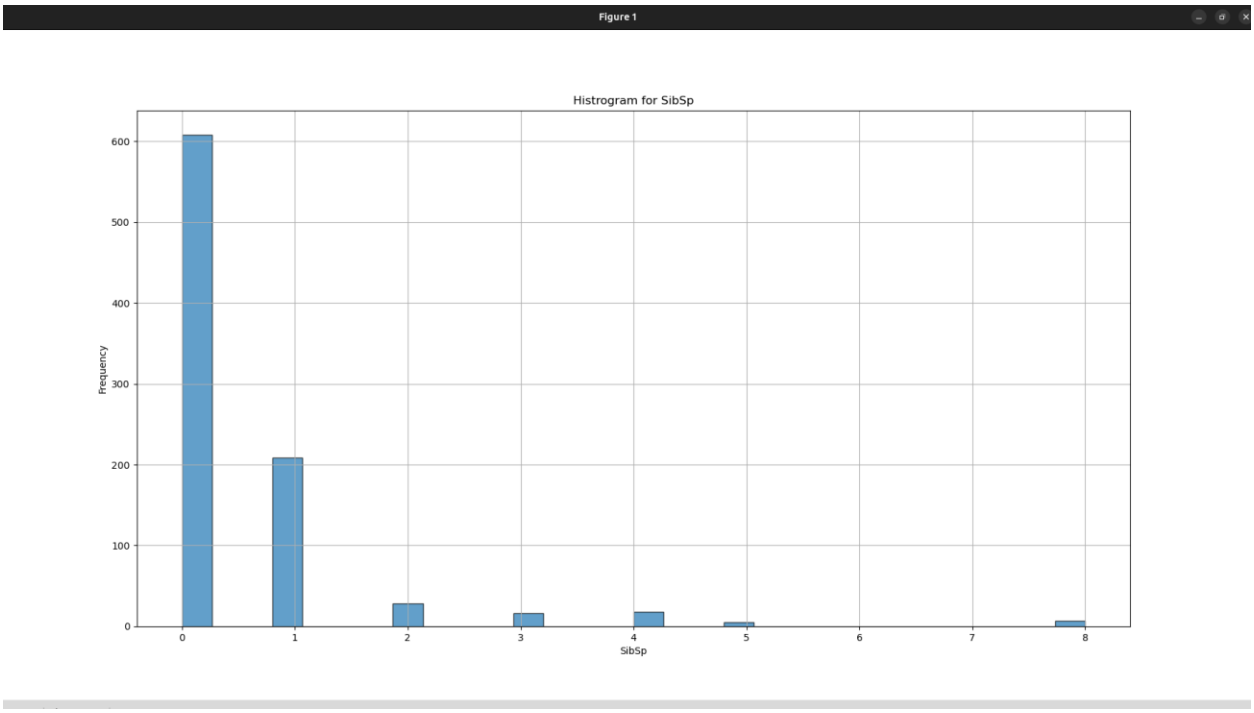


Figure 1
Histrogram for SibSp
Frequency
SibSp
0
1
2
3
4
5
6
7
8
100
200
300
400
500
600

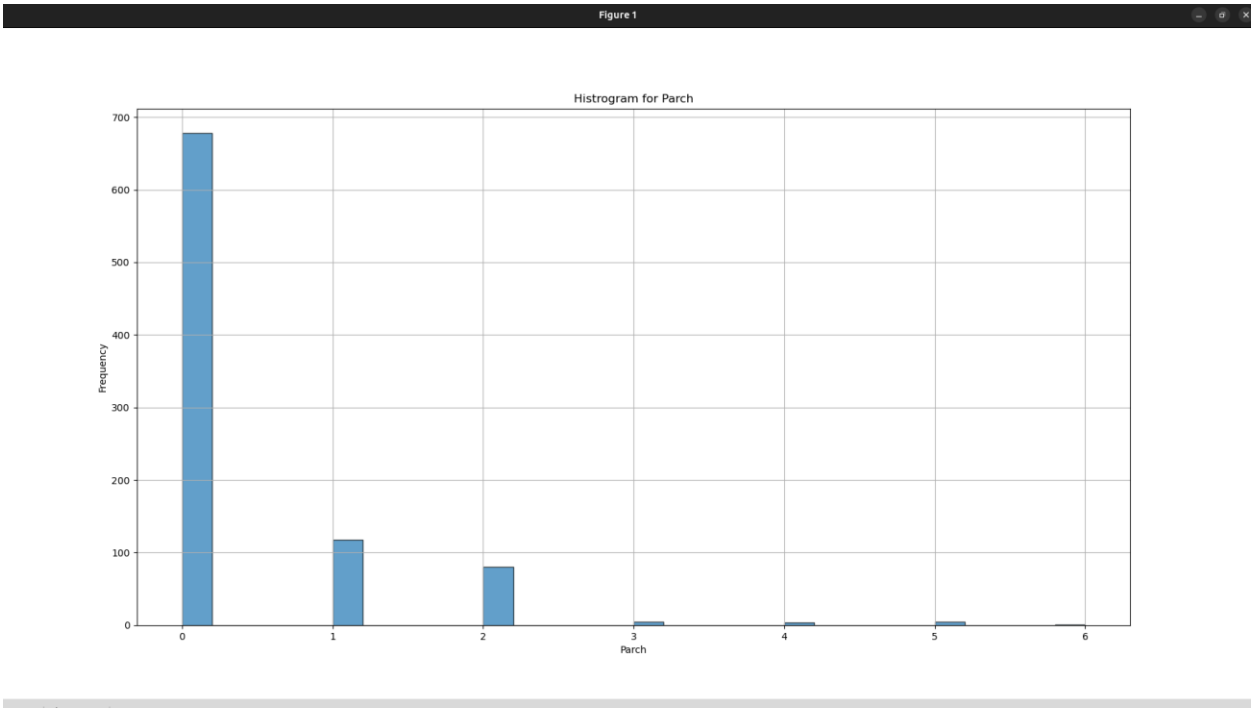
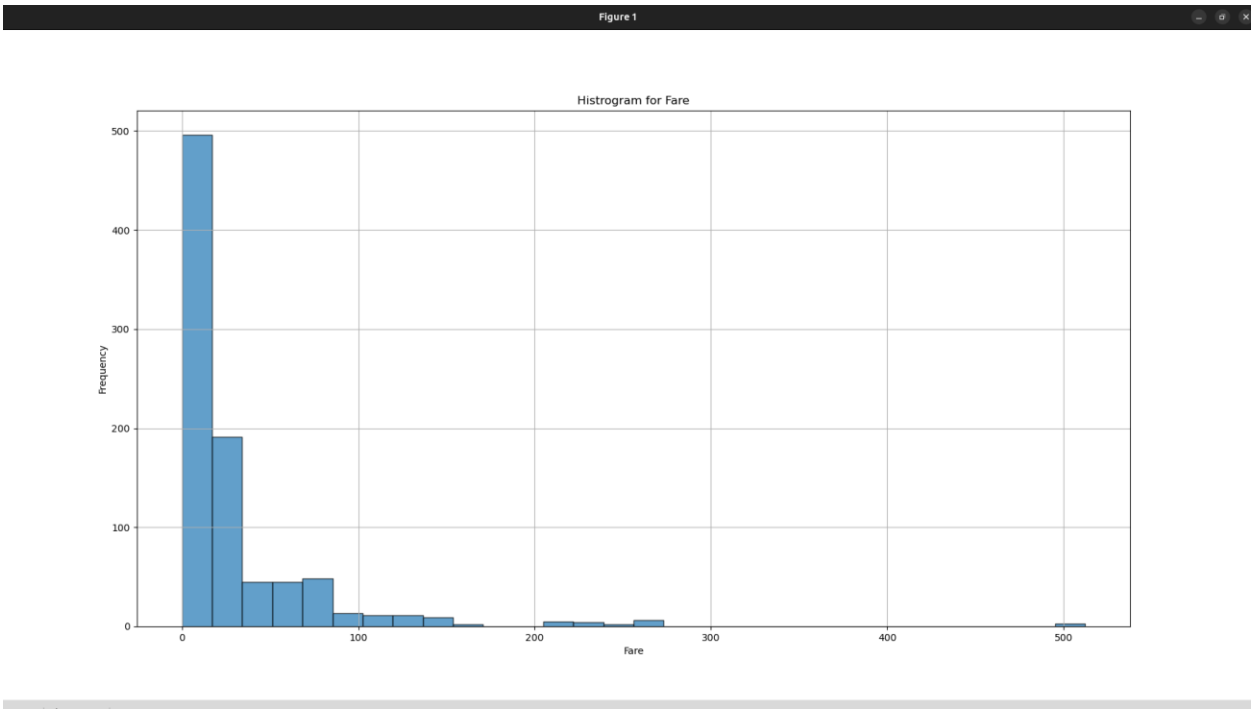


Figure 1
Histrogram for Parch
Frequency
Parch
0
1
2
3
4
5
6
100
200
300
400
500
600
700

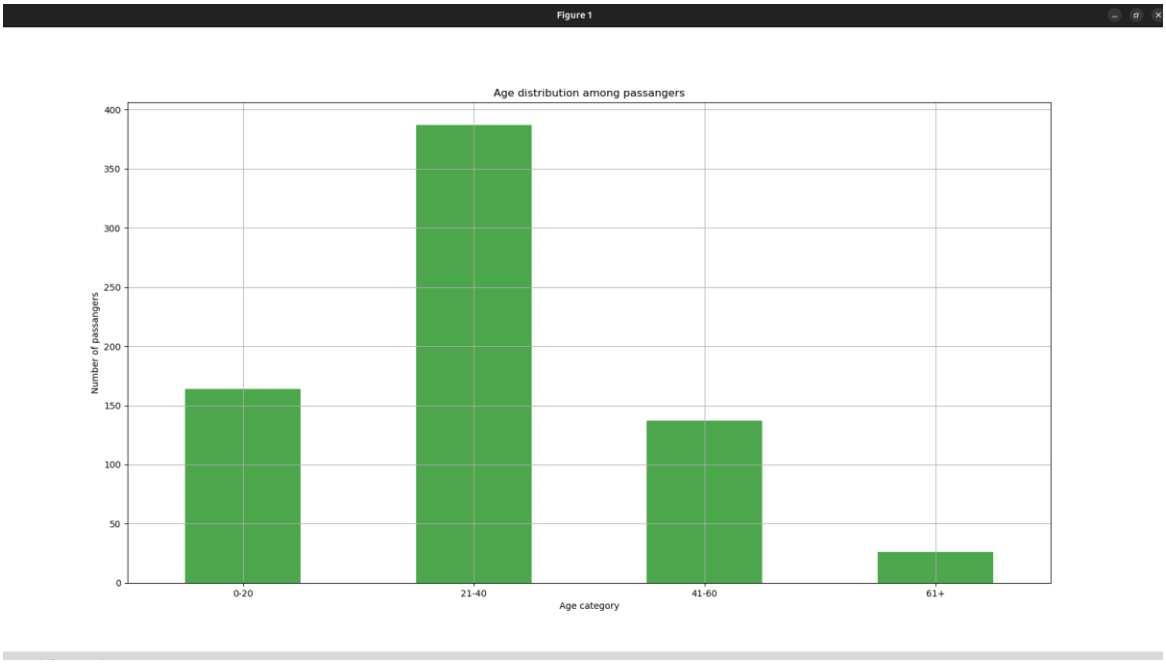


<Task 4>

In acest task vom identifica pentru fiecare coloana proportia valorilor lipsa(%) din cadrul acesteia, raportandu-ne de asemenea si la distributia acestora in cadrul persoanelor care au supravietuit sau nu. Din distributia datelor vom observa ca in mare parte lipsesc data referitoare la cabina in care se aflau pasagerii, atat in cazul celor care au supravietuit (~60%), cat si in cazul celor care au decedat(~87%). Un procent de date, dar de aceasta data ceva mai scazut, observam in cadrul varstei pasagerilor unde lipseste un procent de aproximativ 22% pentru persoanele care decedat si aproximativ 15% pentru pesoanele care au supravietuit.

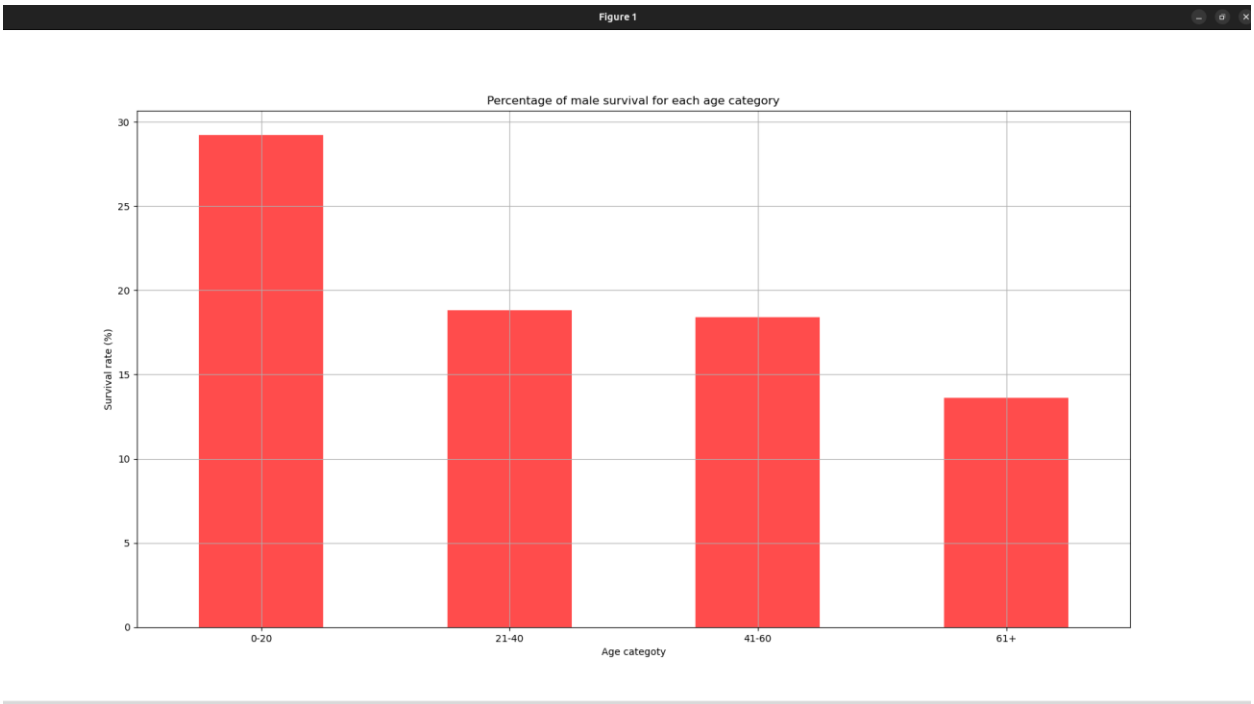
<Task 5>

In cadrul acestui task vom impartii pasagerii pe categorii de varste, impartindu-i in 4 intervale: [0, 20], [21 – 40], [41- 60], [61+]. Vom salva in fisierul task5.csv datele referitoare la repartizarea fiecarei persoane in una din categoriile anterior mentionate. Iata aici distributia pasagerilor:



<Task 6>

In acest task vom analiza cum varsta a fost un factor ce a influentat sansele de supravietuire ale barbatior, in functie de categoria de varsta din care fac parte. Iata aici rezultatele:

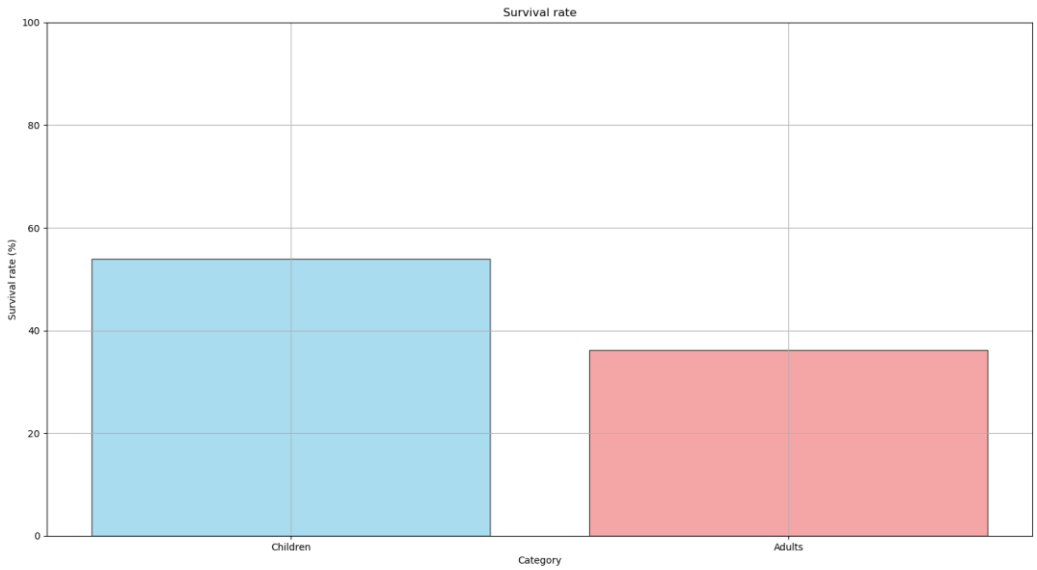


Observam ca aproape 30% din tinerii aflati pe vas au supravietuit, fiind lideri ai

acestui clasament. La polul opus se afla persoanele varstnice, care au supravietuit in procent de ~14%. Am obtinut procente similare la catergoriile de varsta [21-40], [41-60].

<Task 7>

In cadrul acestui task vom compara rata de supravietuire a copiilor cu cea a persoanelor adulte. Iata rezultatele:



Observam o rata mai mare de supravietuire in cazul copiilor, poate pentru ca adultii au prioritizat salvarea celor mai tineri.

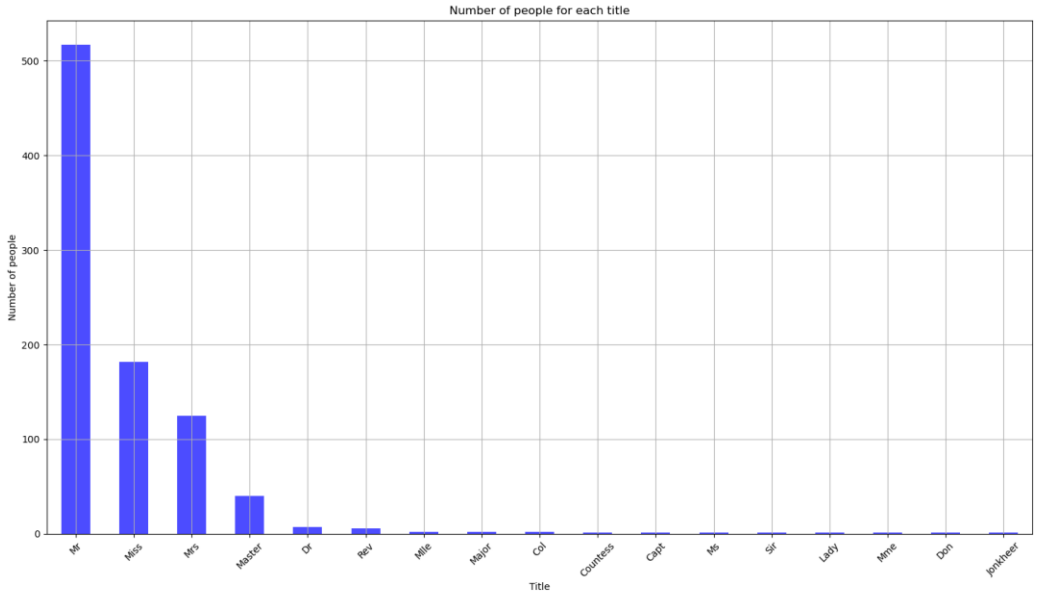
<Task 8>

In acest task vom incerca sa completam datele lipsa din cadrul fisierului, in functie de datele cunoscute pentru restul pasagerilor care au supravietuit sau nu. Vom salva rezultatul acestor predictii in cadrul fisierului task8.csv . Pentru fiecare categorie vom completa fiecare linie care lipseste, tinand cont de alti factori care se intalnesc in datele desrpre persoana respectiva, precum varsta, daca a decedat sau nu, categoria de varsta, clasa, etc. Functia missing_num calculeaza pentru coloanele care contin valori numerice valoare medie a acestora si o adauga la persoanele la care respectiva categorie lipseste. Functia missing_cats va completa in locurile lipsa cea mai des intalnita valoare din

coloana respectiva corespunzatoare persoanelor care au murit sau supravietuit(depinde de persoana curenta).

<Task 9>

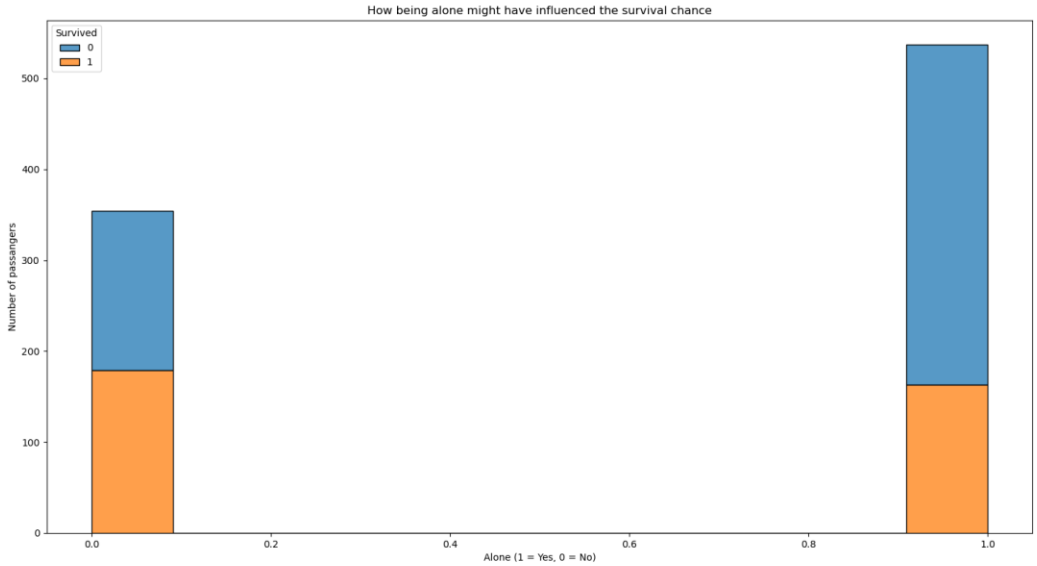
In acest task vom repartiza fiecarui titlu intalnit un numar de persoane care ii corespunde. Iata graficul:



Observam ca cei mai multi oameni prezenti pe vas aveau apelativul Mister(~520), urmati apoi de doamne(~180).

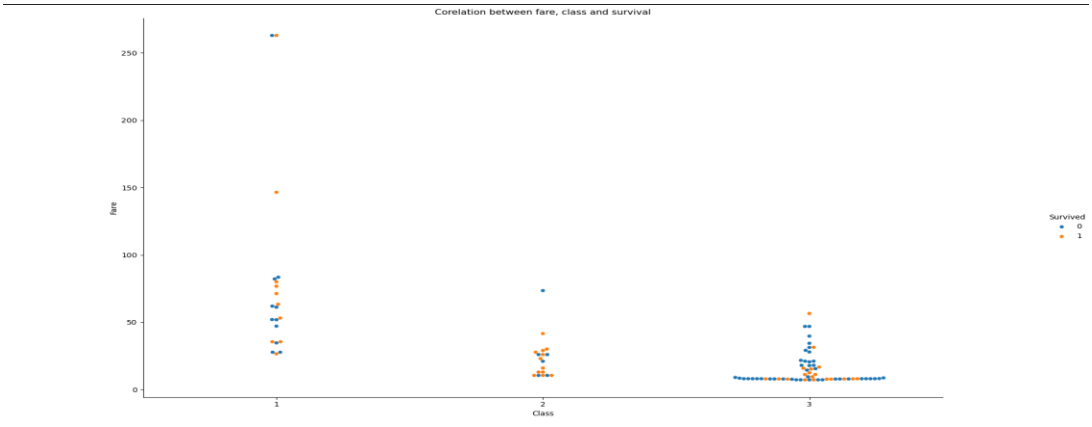
<Task 10>

In acest task vom investiga daca prezenta unor cunostinte pe vas a afectat sau nu sansele de supravietuire ale fiecarei persoane. Ia-ta graficul:



Observam ca printre persoanele insotite de alte cunostinte sansele de supravietuire erau aproape de 50%. In schimb, pentru persoanele singure, rata de supravietuire scade destul de drastic, indicand o tendinta prin care persoanele care mai aveau cunostinte la bord erau intr-o pozitie mai buna de a supravietui.

Urmeaza sa analizam corelatia dintre tarif, clasa si supravietuirea pasagerilor. Iata graficul:



Dupa cum observam si pe graifc, pare ca la clasele 1 si 2 distributia deceselor versus cea a supravietuitorilor este in balans. In schimb, in cazul pasagerilor de la clasa a 3-a, sansele de supravietuire s-au diminuat. Tariful pare sa aiba un amestec nesemnificativ in supravietuirea persoanelor.

Acest grafic poate fi extins si pentru situatii de actualitate din viata de zi cu zi, precum aflarea celui mai sigur loc dintr-un vehicul sau avion, sau cum clasa la care ne cumparam bilet intr-un avion poate influenta sansele de supravietuire in cazul unui eveniment nefavorabil.