Luchian Albert-Calin 313CD

Proiect PCLP3 - Partea 1

Task 1:

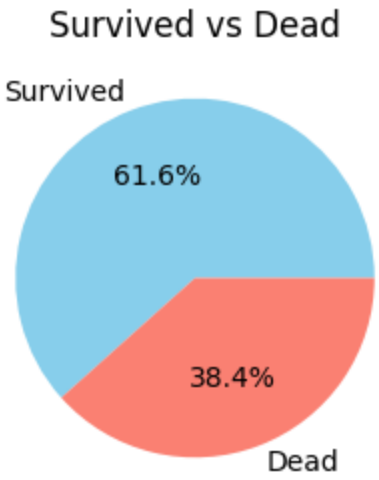
Taskul 1 este rezolvat in cadrul functiei do_task_1(df) unde param df este datafremul citit in main. Calculez nr de linii si nr de coloane folosind functia len pe df.axes[0], respectiv df.axes[1]. Retin in variabila data_types tipul fiecarei coloane. In total_missing_values calculez nr total de valori lipsa folosind functiile isnull() si sum() si in variabila duplicated_rows calculez nr total de randuri duplicate. Apoi afisez pe rand toate aceste variabile.

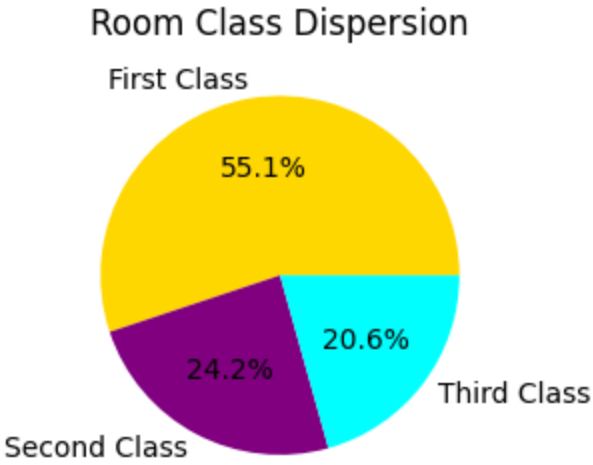
Task 2:

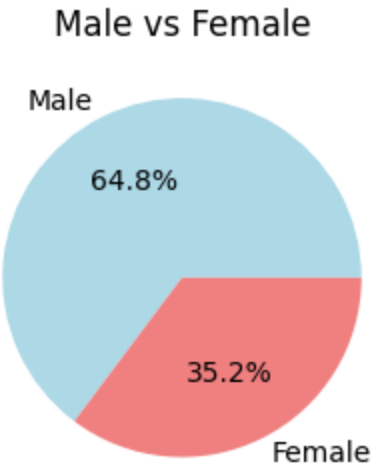
Taskul 2 este rezolvat in functia do_task_2(df). Calculez pe rand procentul pentru oamenii care au supravietuit vs oamenii care nu au supravietuit folosindu ma de param normalaize al fct value_counts si imultind rezultatul final cu 100, apoi rotunjesc rezultatul final la 2 zecimale. Apoi aplic acelasin proces pt a calcula nr de femei vs nr de barbati si pt procentul oamenilor in fiecare tip de camera . Creez un plot cu 3 subploturi fiecare afisand un piechart ce evidentiaza procentele calculate anterior.

Plot-ul rezultat :

Luchian Albert-Calin 313CD



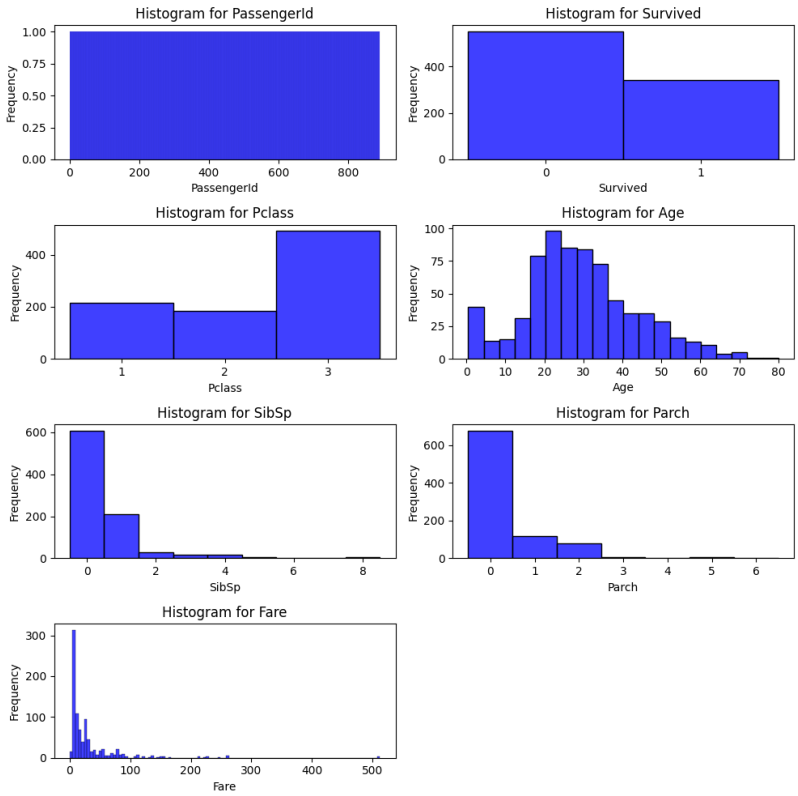




Task 3:

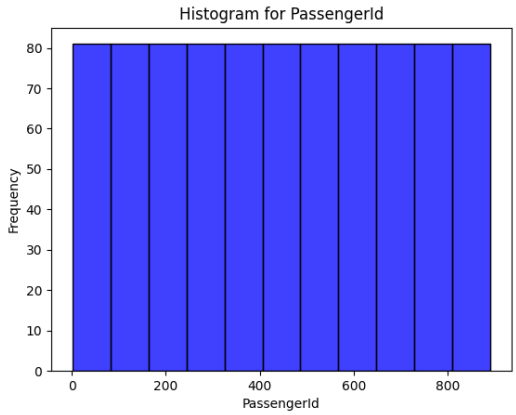
Luchian Albert-Calin 313CD

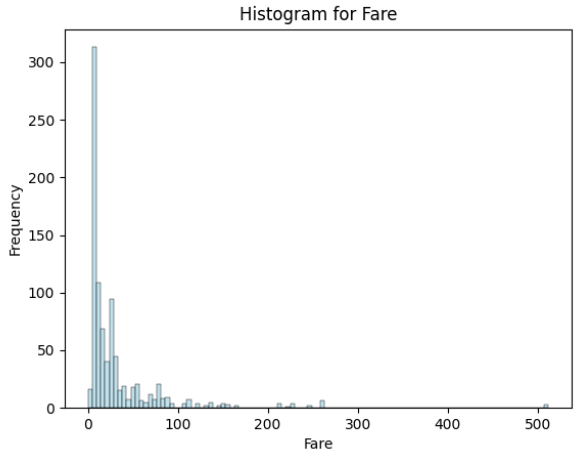
Taskul 3 este rezolvat in functia do_task_3(df). Retin in variabila numerical_cols toate coloanele care au tipul de date float64 sau int64 din baza de date. Apoi calculez nr de n_rows pentru plotul meu mare care este structurat pe formatul n_rows si 2 coloane, n_rows fiind determinat de formula n // 2 + (n % 2) – adaug Inca un rand in cazul in care nr de numerical_cols este impar. Apoi folosesc functia axs.ravel() pt a da flatten la axe si fac un for pe axe si indexul. Apoi pentru fiecare dintre coloanele numerice creez un plot de tipul sns.hist, iar in cele care contin valori intregi fac param discrete sa fie true si valorile afisate pe axa ox sa fie de tipul integer, in final sterg axele nefolosite(daca exista) si afisez graficele rezultate :



Separat pt pasagerId si Fare, deoarece nu se intelegeau bine din plot-ul mare

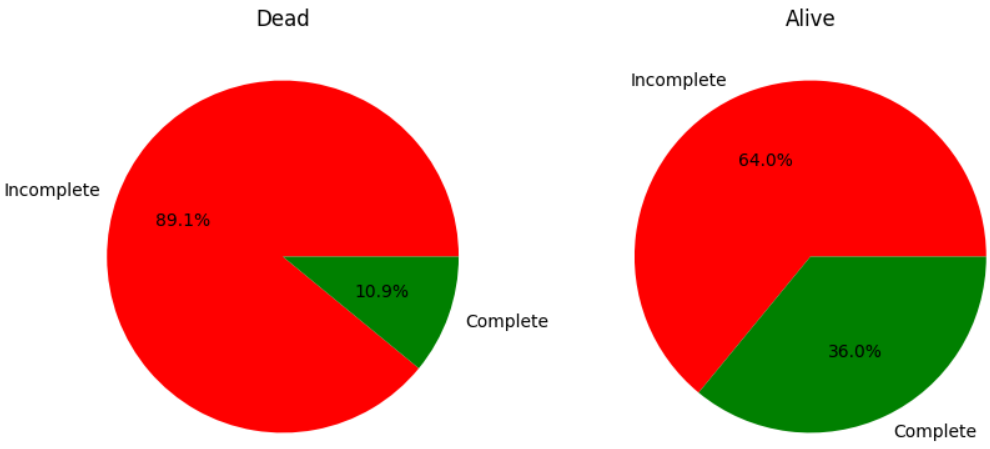
Luchian Albert-Calin 313CD





Task 4:

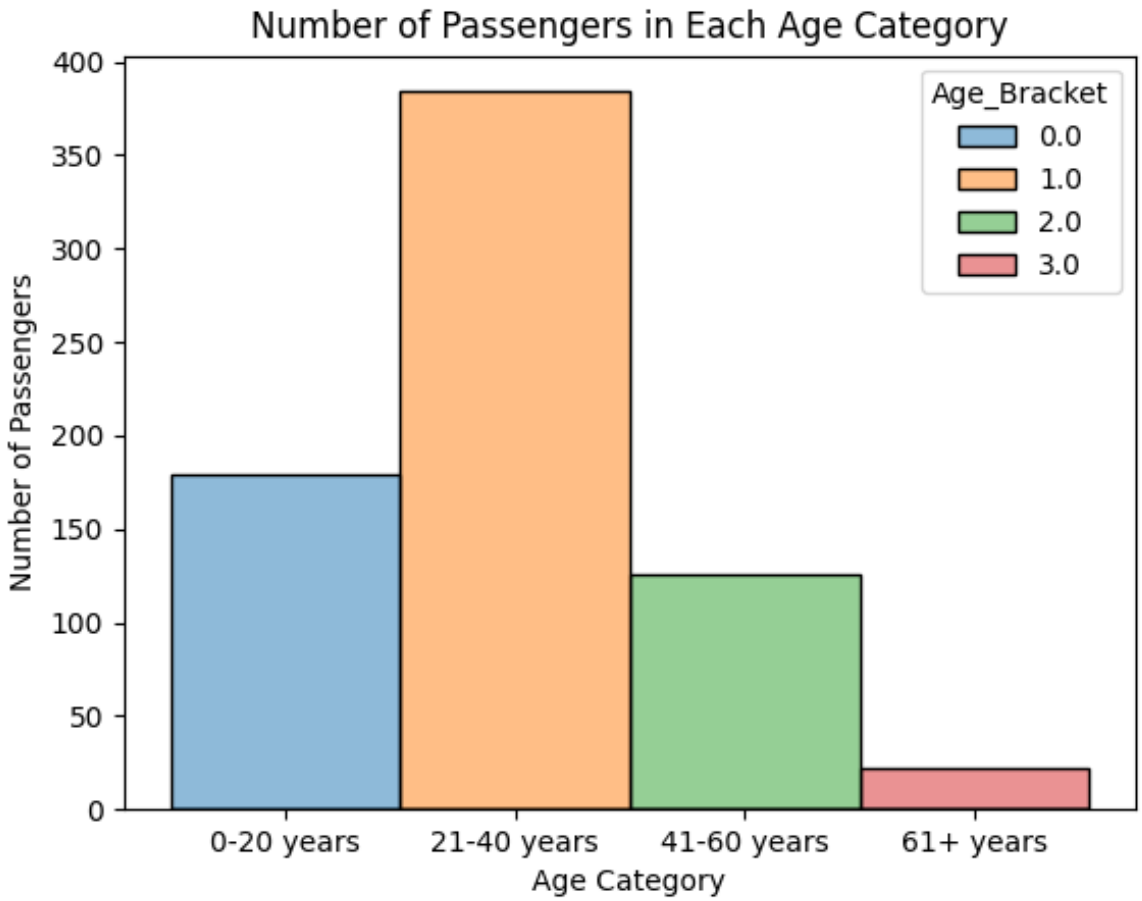
Pentru taskul 4 am retinut in var incomplete_data toate col-urile cu incomplete data si nr de incomplete data din ele. Si pt fiecare dintre aceste col-uri am calculate procentul de missing data din intregul col. Apoi am retinut iar in incomplete data,row-urile intregi cu incomplete data. Am caclulat apoi nr de oameni in viata cu incomplete data si nr de oameni morti cu incomplete data, precum si procentul lor din nr total de oameni morti/vii. Creez apoi 2 piecharturi pentru a evidentia asta.



Task 5:

Pentrul taskul 5 am folosit trei functii get_index, add_age_brackets si do_task_5, functia add_age_brackets itereaza prin valorile coloanei ‘Age’ a dataframeului, pt fiecare Age ia indexul corespunzator al valorilor din indexul de Age_Brackets sau intoarce None

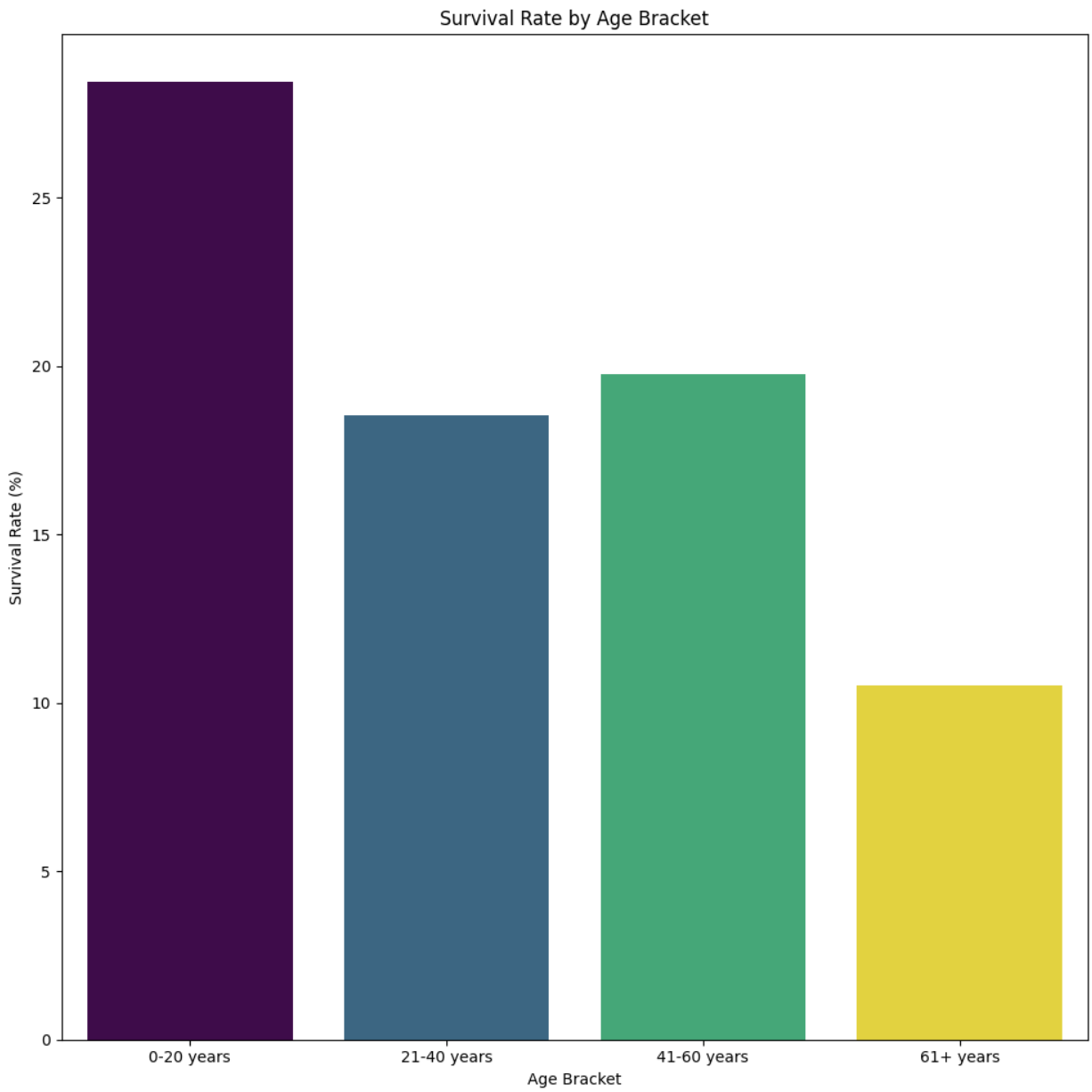
daca valoarea intiala este None. Dupa ce aceasta coloana este construita o inserez in dataFrameul original si construiesc un histplot pt a evidentia nr de pasageri din fiecare age_bracket. In final salvez df obtinut in fisierul data1.csv



Task 6:

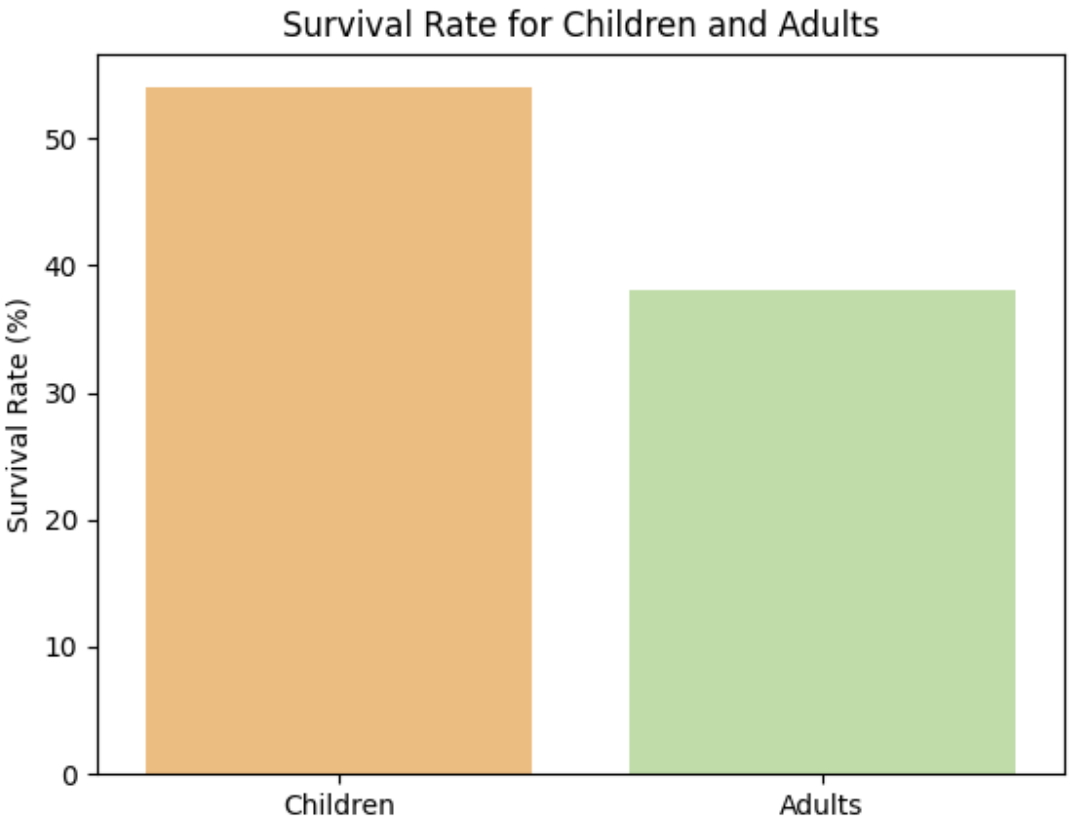
Calculez nr total de barbati care au supravietuit si nr total de barbati din fiecare age_bracket. Calculez procentul de barbati supravietuiti din fiecare age_bracket si creez un barplot pentru a evidentia survival_rate-ul pentru fiecare age_bracket :

Luchian Albert-Calin 313CD



Task 7:

Retin in variabila children_df toate randuri-le corespunzatoare persoanelor cu varsta mai mica de 18 ani din df, calculez procentul de copii din df si il afisez. Apoi calculez pe rand survival _rate-ul pe copii vs pentru adulti si creez un barplot pentru a compara aceste 2 rezultate :



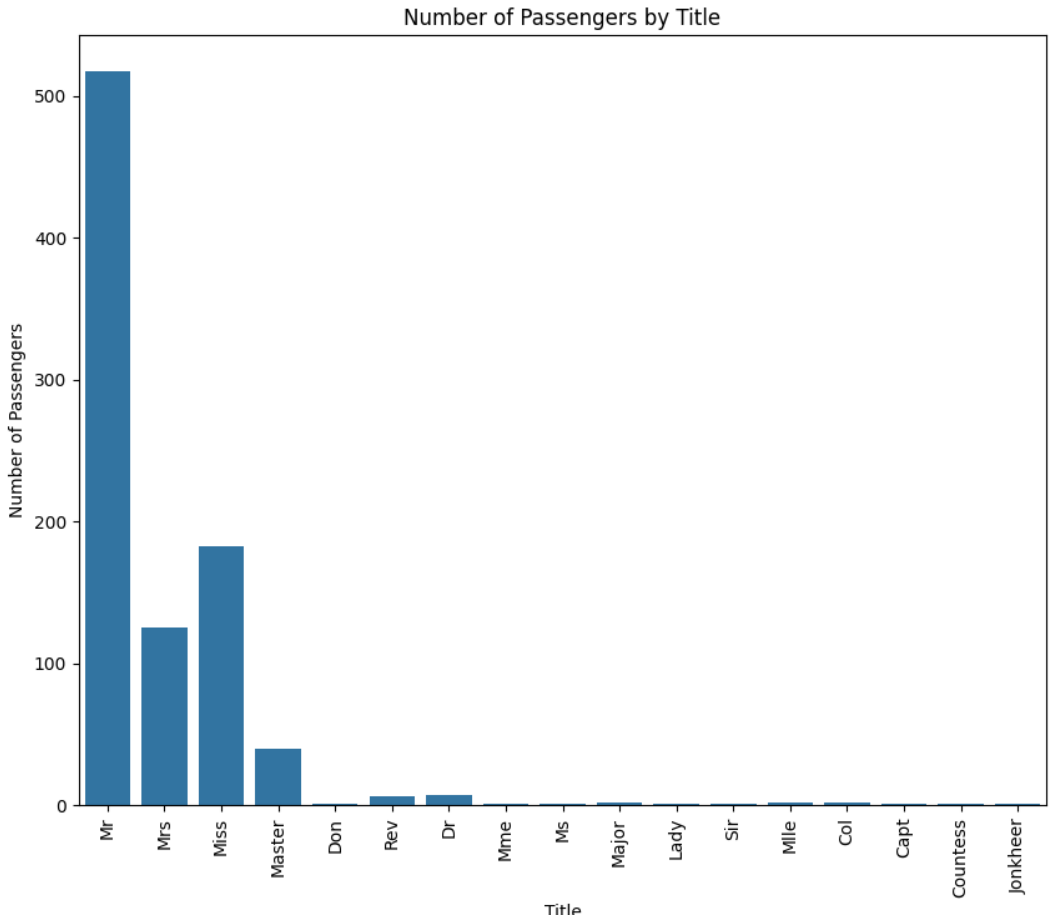
Task 8:

Pentrul task-ul 8 am folosit functia complete_df. Am luat o lista cu numele coloanelor incomplete din df. Am verificat daca tipul coloanei este numeric. Daca da am inlocuit valorile lipsa cu media val coloanei folosind comanda mean(), iar in cazul nu care sunt numerice am inlocuit cu cea mai frecventa valoare cu ajutorul comenzii .mode() si val de la indexul 0 din rezultat fiind cea mai frecventa val de pe aceea coloana. Am scris df-ul obtinut in fisierul data2.csv

Task 9:

Pentru taskul 9 am creeat un dictionar cu toate titlurile din df si gender-ul asociat lor, apoi am creeat o coloanal numita Titles care contine doar titlu-ul persoanei si o coloana expected gender care retine ce gender ar trebuii sa aiba persoana cu titlul respectiv. In final am comparat coloanal ‘Sex’ cu ‘Expected_Gender’ si am vazut cate valori difera(am ajuns la rezulatul 0). Apoi am creeat un countplot pentru a evidentia nr de aparitii al fiecarui titlu:

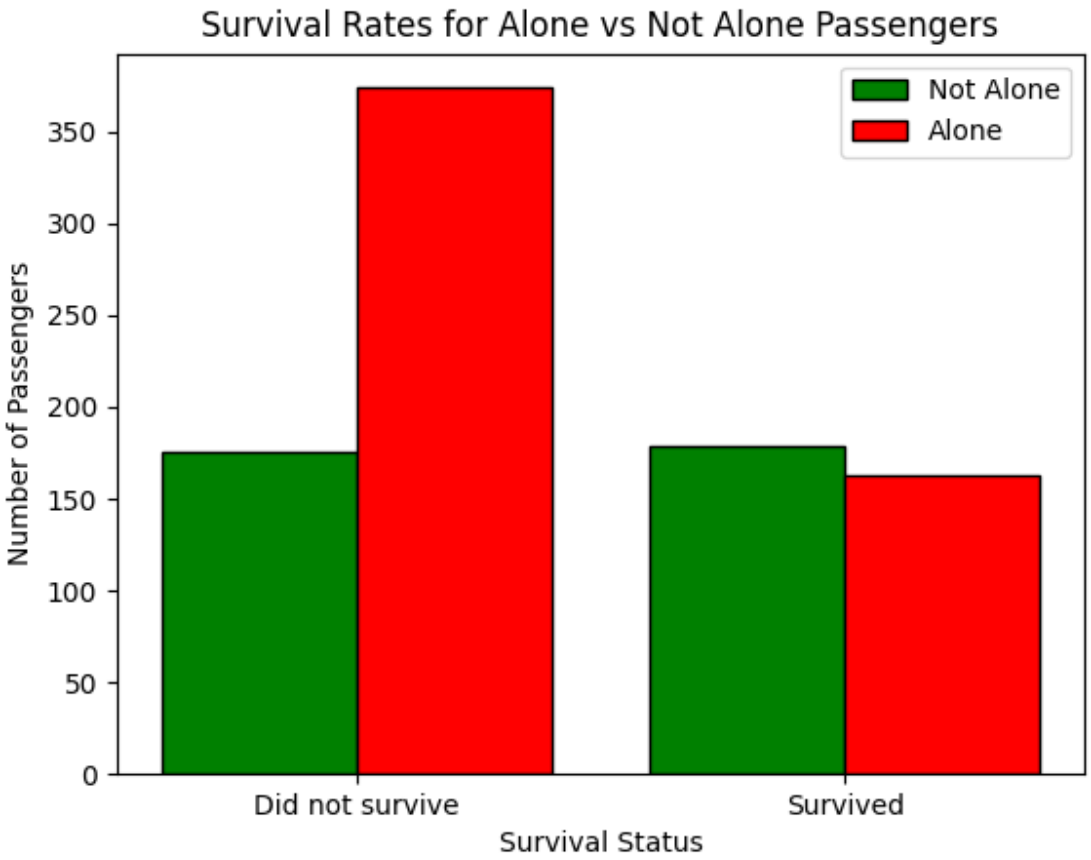
Luchian Albert-Calin 313CD



Task 10:

Pt task-ul 10 am folosit doua functii investigate_alone_survival si do_task_10. In investigate_alone_survival am folosit var alone pt persoanele cu 0 SibSp si 0 Parch (adica fara parinti, copii, frati, parteneri) si not_alone pt persoanele care au macar unul dintre cele doua categorii anterioare diferite de 0. Apoi am creeat o histograma pt a evidentia nr de persoane Alone care au murit vs nr de pers Not Alone care au murit si una pt cele care au trait.

Luchian Albert-Calin 313CD



Din acest graphic se poate observa clar ca persoanele singure au avut sansa mult mai mare de a muri decat cele care mai aveau si alte rude pe vas.

In functia do_task_10 am creeat un swarm plot pentru a evidentia relatia dintre Fare, tipul camerei in care au trait si statul de alive/Not Alive:

Luchian Albert-Calin 313CD