Proiect PCLP3

Ciocodeica Antonio-Mihai, 314CD

Partea I.

Cerinta 1.

Pentru citirea si examinarea structurii fisierului train.csv, am folosit modulul Pandas. Metoda .read_csv() permite citirea unui dataframe aflat la adresa primita ca argument. Determinarea numarului de valori lipsa si a numarului de linii duplicate a fost realizata folosind metode precum .isnull() sau .duplicated().

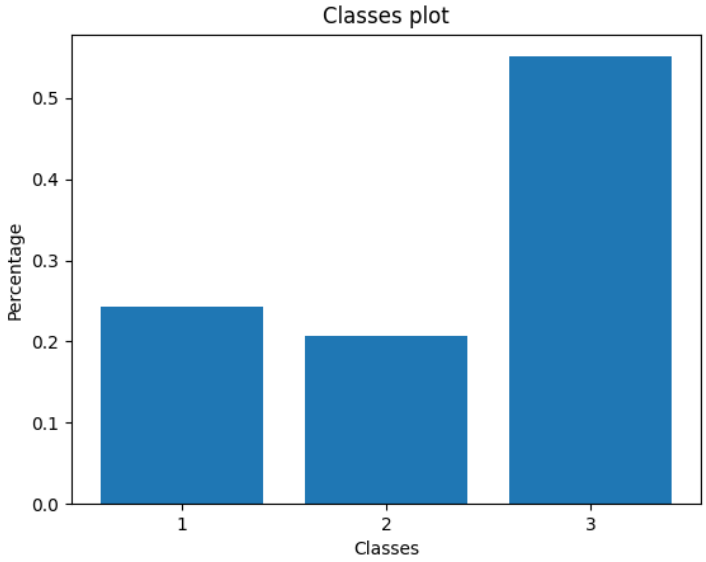
Cerinta 2.

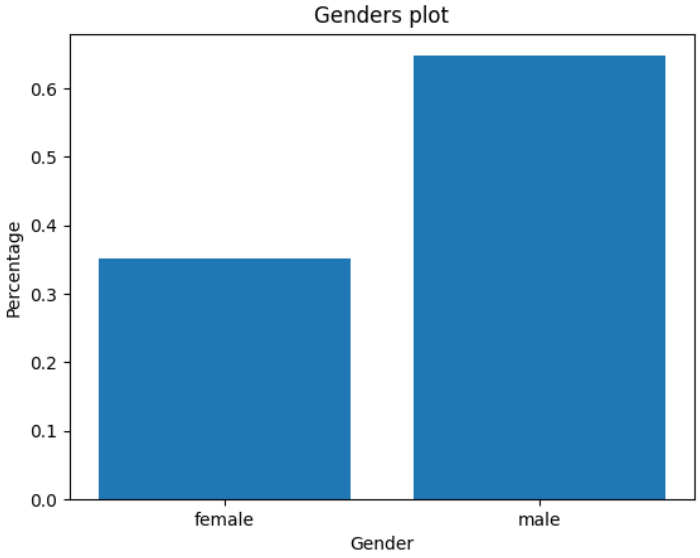
Determinarea procentului de persoane care au supravietuit se realizeaza prin impartirea numarului de supravietuitori la numarul total de pasageri. Pentru a afla numarul de supravietuitori, calculam suma elementelor de pe coloana ‘Survived’. Aceasta coloana contine valori de 1 pentru supravietuitori si valori de 0 in rest.

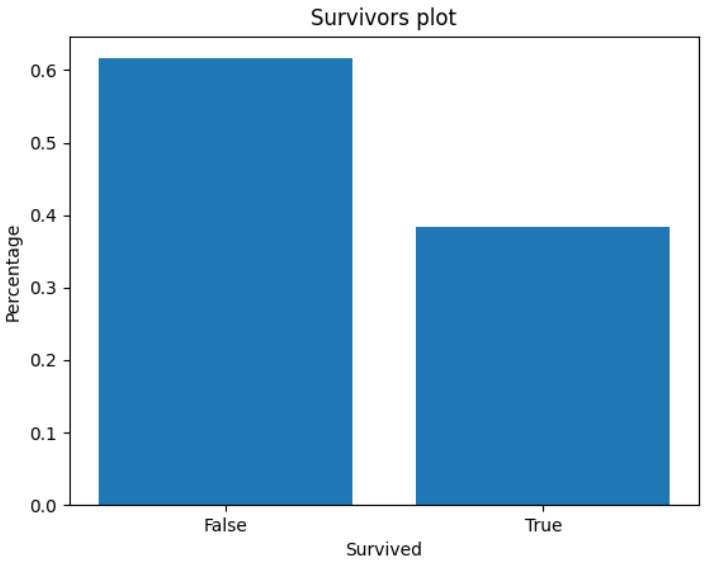
Pentru a determina procentul pasagerilor in functie de clasa, am creat o lista cu clasele unice din coloana ‘Pclass’, apoi am parcurs coloana din dataframe si am incrementat pentru fiecare clasa intalnita valoarea corespunzatoare din lista de aparitii totale. Procentul pasagerilor din fiecare clasa este egal cu numarul de pasageri din clasa respectiva impartit la numarul total de pasageri.

Determinarea procentului de barbati si de femei a fost realizat in acelasi mod cu algoritmul descris mai sus.

Graficele pentru reprezentarea rezultatelor au fost realizate folosind modulul “matplotlib.pyplot”. Pe axa orizontala sunt prezente categoriile in care pot fi incadrati pasagerii, iar axa verticala reprezinta procentul persoanelor care se incadreaza in fiecare categorie prezentata.

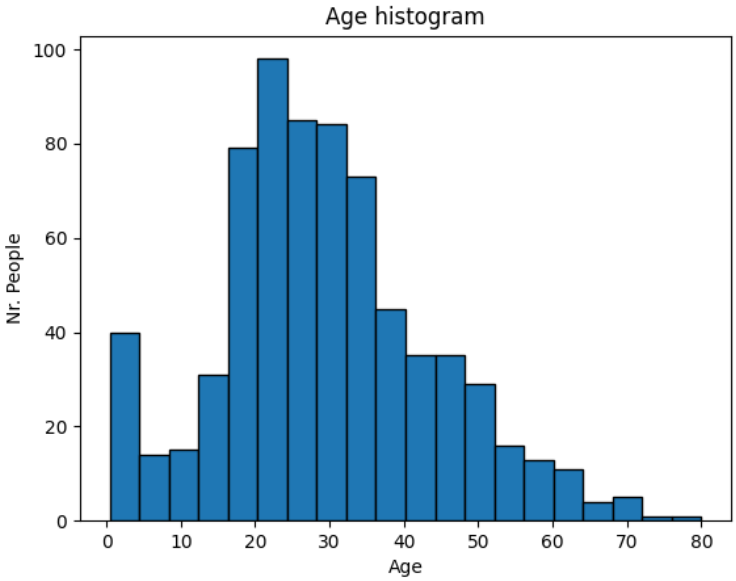


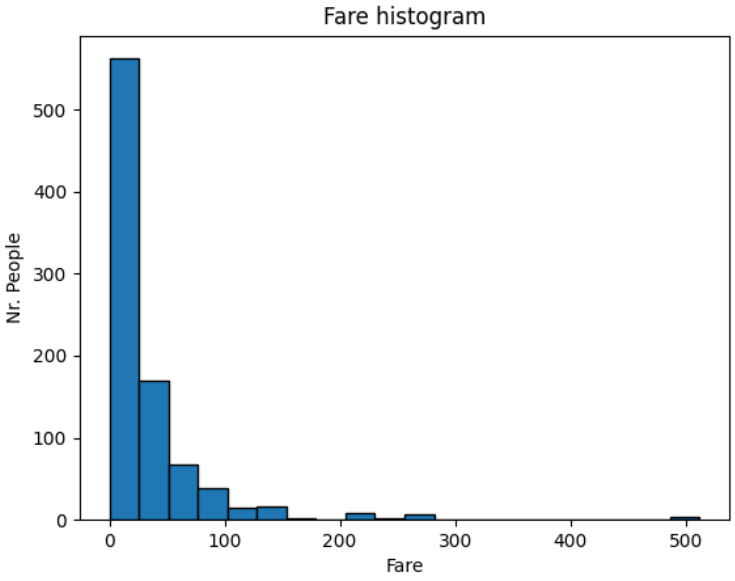


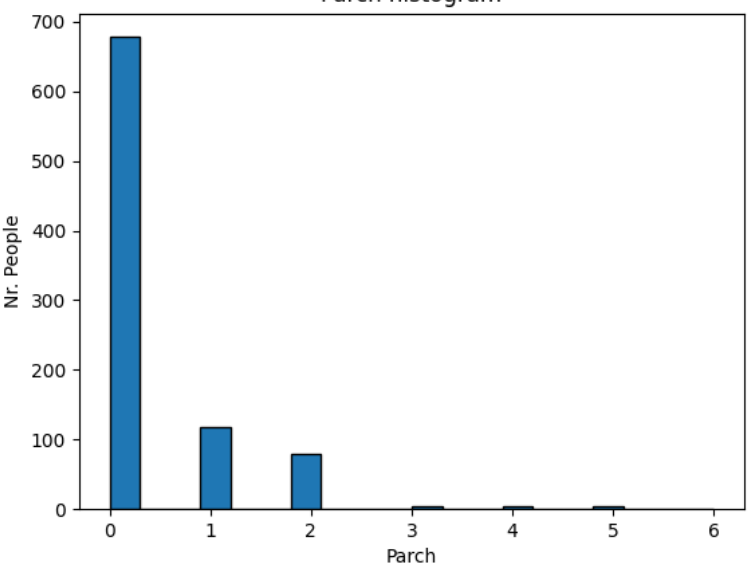


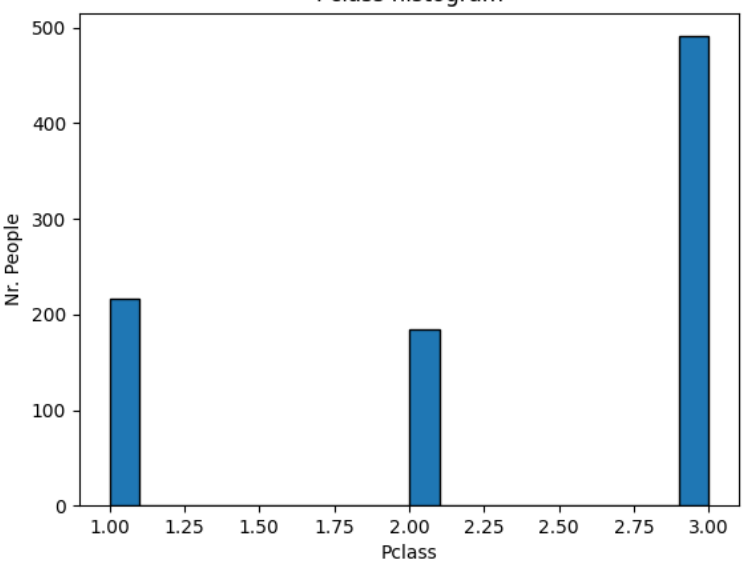
Cerinta 3.

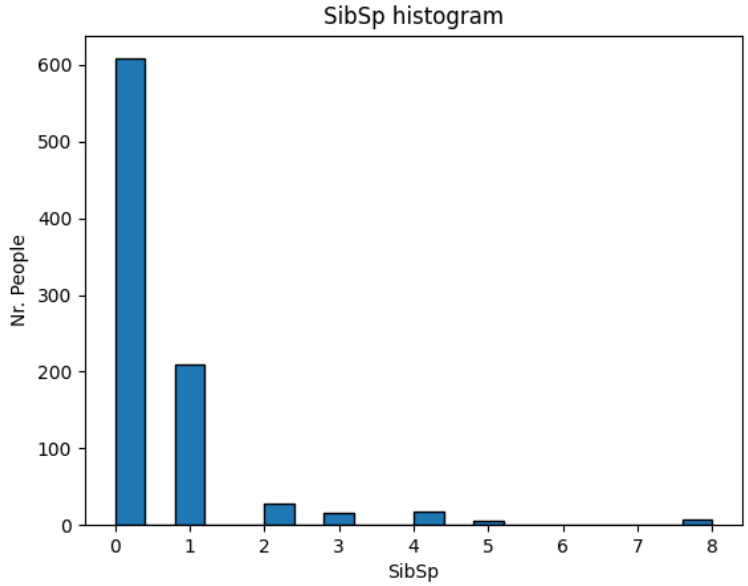
Cerinta implica generarea de histograme pentru coloanele din dataframe care contin valori numerice. Pentru a realiza acest lucru, am iterat prin fiecare coloana a dataframe-ului si am verificat daca tipul de date este ‘int’ sau ‘float’. Pentru coloanele care respecta aceasta conditie, am creat o histograma intr-un mod similar cu generarea histogramelor de la cerinta 2.

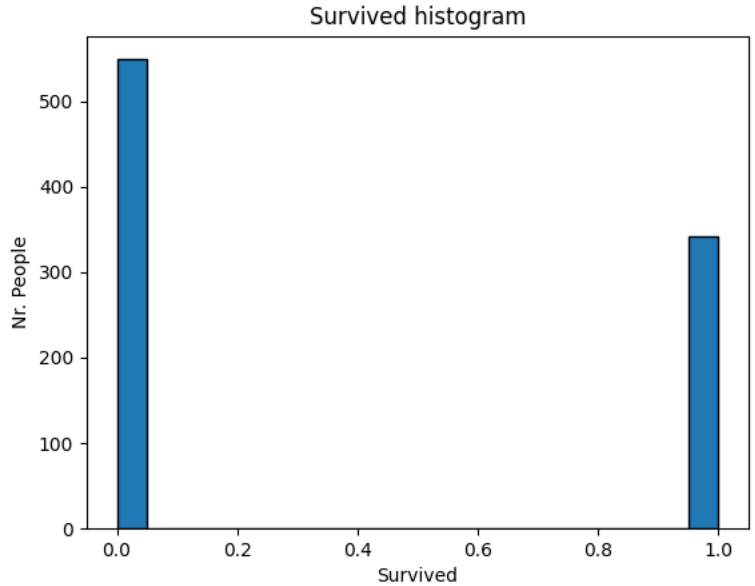

Age histogram
Nr. People
Age
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100


Fare histogram
Nr. People
Fare
0
100
200
300
400
500
0
100
200
300
400
500


Parch histogram
Nr. People
Parch
0
1
2
3
4
5
6
0
100
200
300
400
500
600
700


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


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


Survived histogram
Nr. People
Survived
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

Cerinta 4.

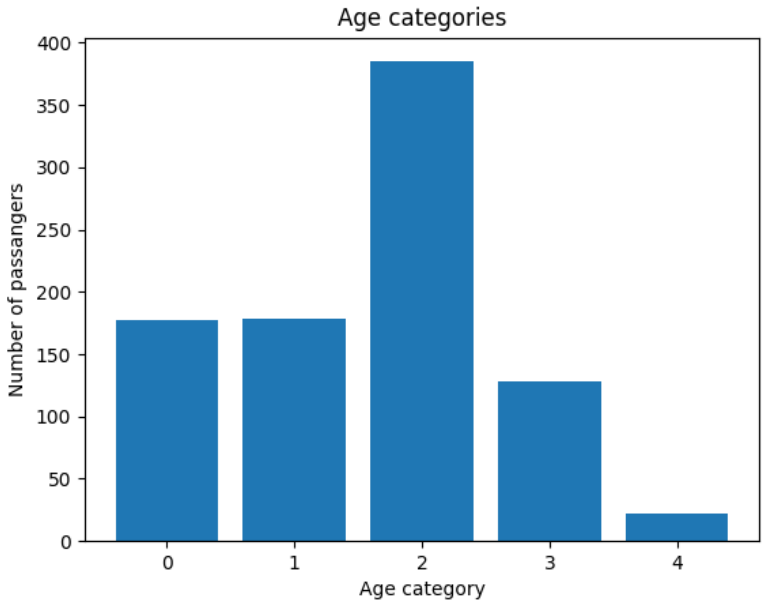
Pentru rezolvarea cerintei 4, trebuie sa identificam coloanele care contin valori lipsa si pentru fiecare coloana sa determinam numarul de valori lipsa si proportia acestora.

Pentru a realiza acest lucru, am iterat prin fiecare coloana a dataframe-ului si am verificat, folosint metoda .isnull() daca exista sau nu valori lipsa in coloana curenta. In cazul in care exista, am parcurs coloana si pentru fiecare valoare lipsa, am incrementat numarul de valori NaN gasite pentru clasa respectiva.

Cerinta 5.

Cerinta implica impartirea tuturor pasagerilor in categorii de varsta si adaugarea acestei categorii in dataframe. Pentru adaugarea unei coloane in dataframe, am folosit metoda .insert(), cu valoarea default 0. Categoriile de varsta prezentate in enunt au fost indexate incepand cu 1, iar pentru persoanele a caror varsta este necunoscuta au fost inclusi in categoria 0.

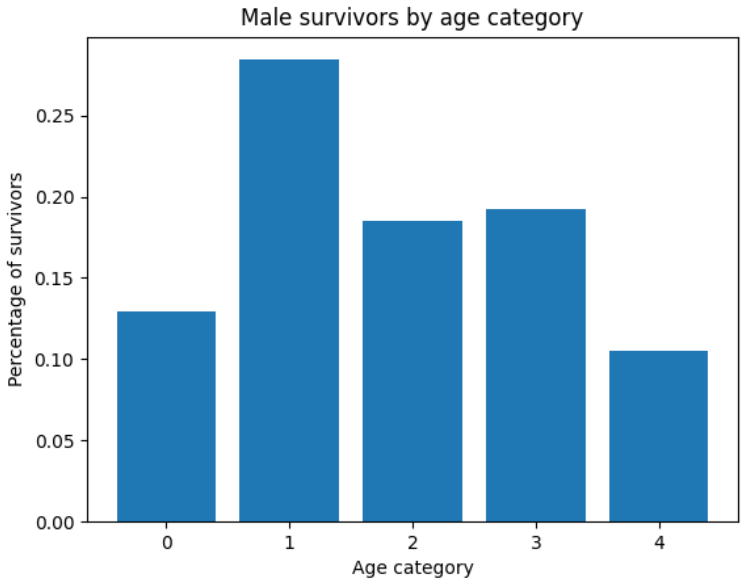
Pentru incadrarea persoanelor in fiecare categorie, am creat lista ‘category’ si am parcurs coloana ‘Age’ din dataframe, incrementand, pentru fiecare pasager, valoarea corespunzatoare din lista creata. Noul dataframe, care contine si coloana ‘AgeCategory’ a fost salvata cu numele ‘AgeCategory_train.csv’, folosind metoda .to_csv. Repartitia pasagerilor pe categorii de varsta a fost evidentiata si printr-un grafic.



Cerinta 6.

Cerinta consta in determinarea numarului de barbati care au supravietuit pentru fiecare dintre cele 4 categorii de varsta propuse anterior si evidentierea procentului de supravietuire printr-un grafic.

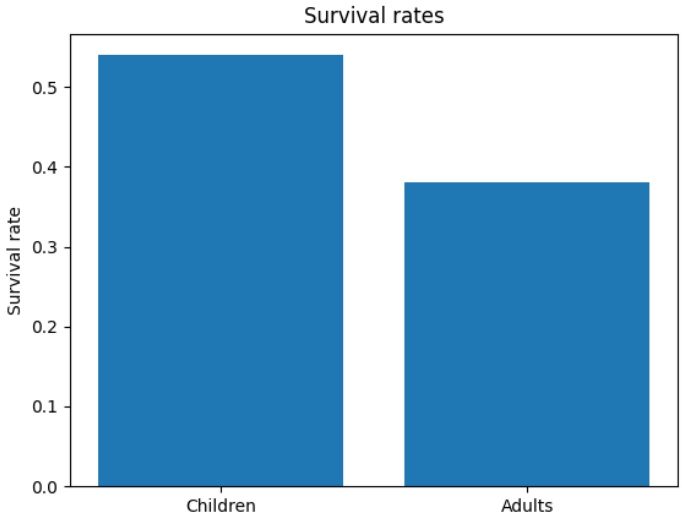
Pentru determinarea numarului de barbati din fiecare categorie de varsta, am parcurs dataframe-ul si pentru fiecare pasager barbat, am incrementat valoarea corespunzatoare din lista ‘total_males’. Mai mult, pentru determinarea procentului de supravietuire al fiecare clase, am incrementat si valoarea din ‘male_survivors’, in cazul in care pasagerul a supravietuit. In graficul obtinut se poate observa cum procentul de supravietuire este mult mai mare in cazul persoanelor mai tinere.



Cerinta 7.

Cerinta presupune determinarea procentului copiilor aflati la bord si realizarea unui grafic pentru a evidentia rata de supravietuire in cazul copiilor, dar si al adultilor.

Rezolvarea cerintei este foarte similara cu rezolvarea cerintei 6. Totusi, pentru o reprezentare cat mai exacta, am decis sa ignor persoanele a caror varsta este necunoscuta.



Cerinta 8.

Cerinta necesita completarea valorilor lipsa din dataframe cu media valorilor obtinute pentru pasagerii care fac parte din aceeasi clasa.

Pentru a rezolva cerinta, am selectat fiecare coloana din dataframe si am verificat daca exista valori lipsa. In cazul in care coloana contine valori lipsa, am iterat prin toata coloana si am calculat valoarea medie, atat pentru supravietuitori, cat si pentru cei decedati. In cazul in care tipul de date al coloanei nu este numeric, nu se poate calcula media, asa ca am determinat inregistrarea cu cel mai mare numar de aparitii. Dupa determinarea mediei, am iterat din nou prin toata coloana si am inlocuit valorile lipsa cu valoarea medie obtinuta.

Dataframe-ul rezultat a fost salvat in directorul ‘Date’ cu numele ‘mean_train.csv’.

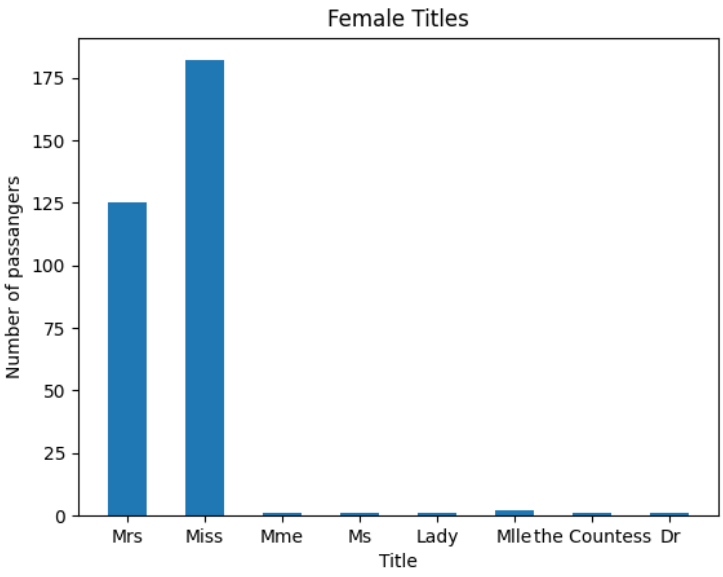
Cerinta 9.

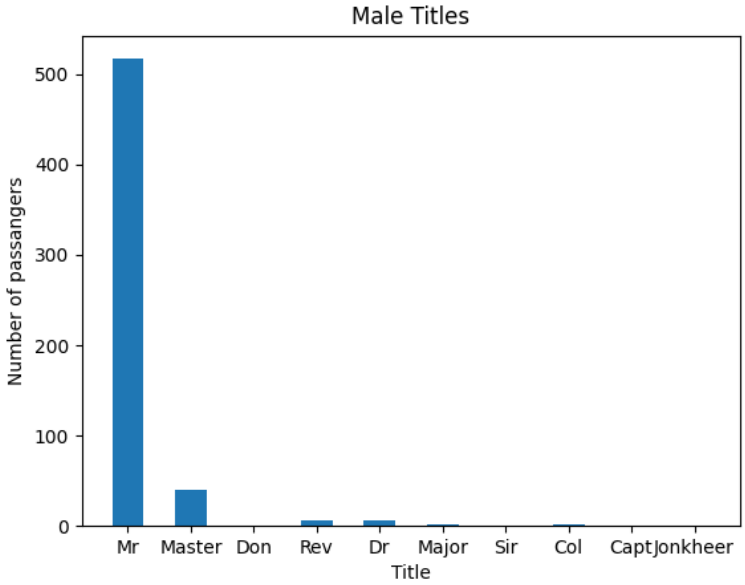
Cerinta presupune determinarea titlurilor de noblete in functie de sexul persoanelor si numarul de aparitii al acestora.

Pentru rezolvarea cerintei, am creat doua dictionare care vor contine drept chei titlurile gasite in dataframe, iar ca valori numarul de aparitii al fiecarui titlu. Am iterat prin coloana ‘Name’ a dataframe-ului si am extras titlul de noblete folosind regex (pentru fiecare inregistrare, titlul de noblete se afla dupa un spatiu si se termina cu caracterul “.”).

Avand titlul de noblete, obtinem si sexul persoanei si verificam daca exista in dictionar titlul curent. In cazul in care exista, incrementam valoarea cheii, in caz contrat, adaugam cheia in dictionar cu valoarea 1.

Reprezentarea grafica a rezultatului obtinut este similara cu reprezentarile prezentate la cerintele de mai sus.





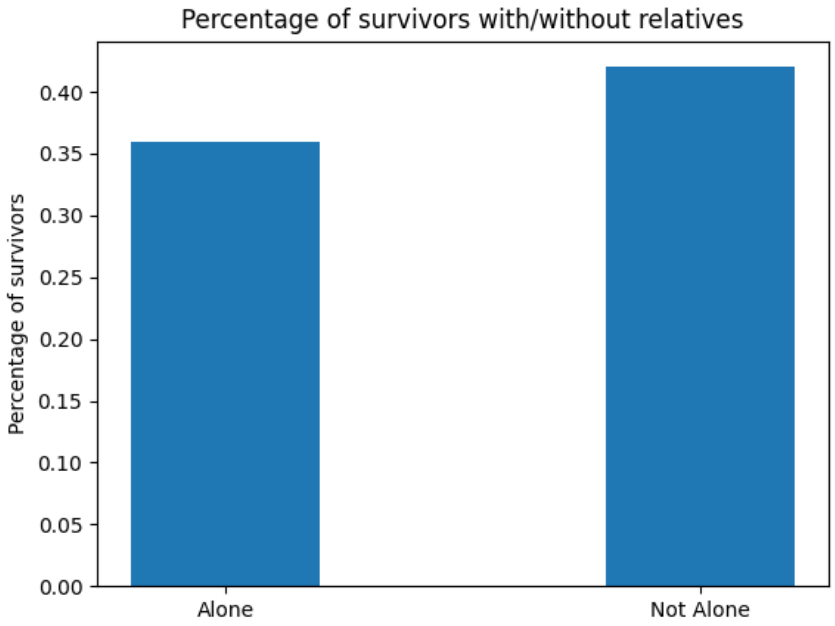
Cerinta 10.

Pentru a raspunde la intrebarea din cerinta, trebuie sa determinam procentul de supravietuire al persoanelor singure si al celor cu rude pe vas. Putem face acest lucru prin a crea un dictionar cu numele de familie, similar cu ce am facut pentru titlurile de la cerinta 9. Numele de familie al fiecarei persoane este reprezentat de primul cuvant din coloana ‘Name’.

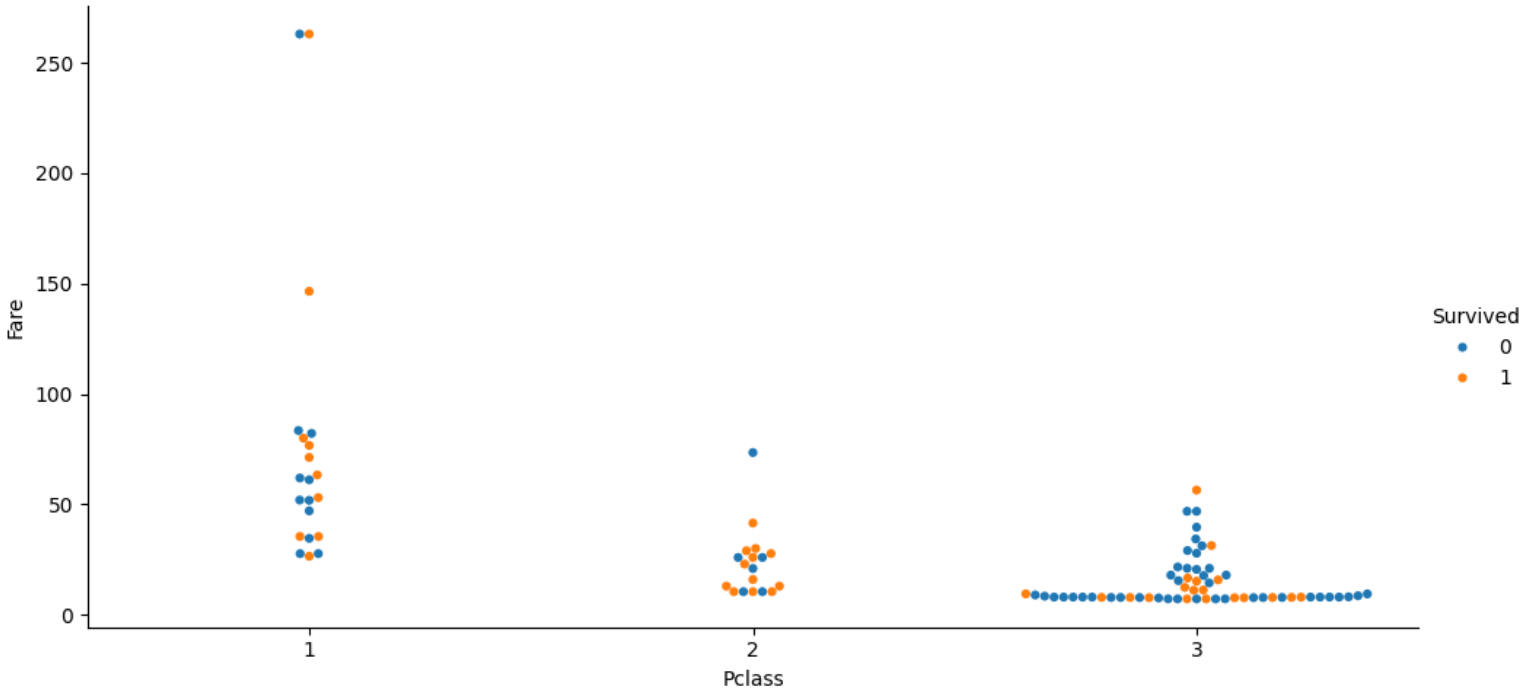
Dupa ce am creat dictionarul care contine toate numele de familie, impreuna cu numarul de aparitii al acestora, trebuie sa determinam numarul de persoane singure si numarul de persoane cu rude. Am parcurs tot dataframe-ul si pentru fiecare persoana, am verificat daca a fost singura si daca a supravietuit. Astfel, am creat un dictionar cu cheile “Alone” si “Not_Alone”. Fiecare cheie va avea ca valoare un alt dictionar cu cheile “True” si “False”, care indica daca au supravietuit sau nu.

Avand toate informatiile necesare, putem calcula rata de supravieturie pentru cele doua categorii de oameni. Pentru setul de date “train.csv”, procentul de supravietuire al persoanelor cu rude a fost de aproximativ 42%, pe cand cel al persoanelor singure de 36%.

In concluzie, starea de a fi singur pe Titanic a avut influente asupra sanselor de supravietuire.



A doua parte a cerintei presupune investigarea relatiei dintre tarif, clasa si starea de supravietuire pentru primele 100 de inregistrari. Pentru a realiza acest lucru, am folosit un grafic ce contine informatii despre cele 3 intregistrari ale fiecarei persoane.



Se poate observa cum numarul de puncte albastre este cel mai mare in cadrul clasei 3, fapt care indica o rata de supravietuire mai mica decat pentru clasele 1 si 2.

De asemenea, pentru clasa 1, tariful este mai mare decat pentru celelalte clase.