Dima Alexandru – grupa 312CC

Proiect PCLP3 – Titanic

- partea 1 -

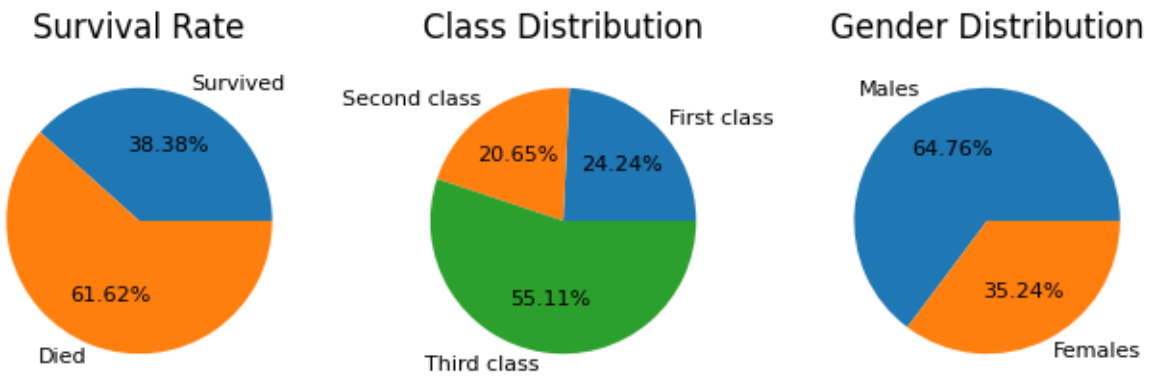
Acest proiect presupune prelucrarea faimosului set de date Titanic. Partea I din cerinta vizeaza partea din setul de date destinata antrenarii. Am folosit module cunoscute precum pandas, matplotlib si seaborn.

Cerinta 1

Pentru rezolvarea acestei cerinte se citesc datele din fisierul .csv intr-o structura DataFrame corespunzatoare bibliotecii pandas. In continuare se vor prelucra atribute ale acestei structuri precum shape, dtype, iar numarul de coloane cu valori lipsa si numarul de linii duplicate se vor determina prin metode precum dropna(), count() si duplicated().

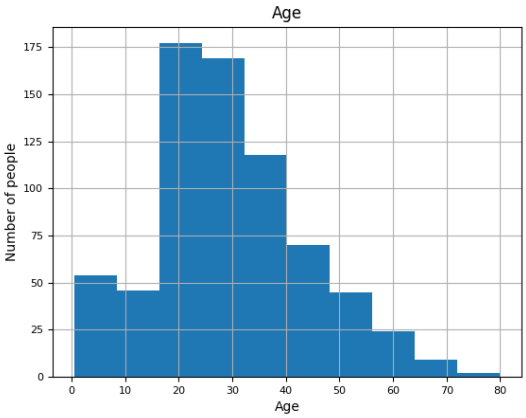
Cerinta 2

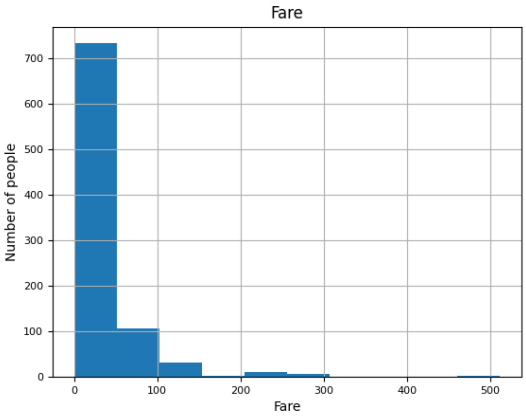
Pentru rezolvarea acestei cerinte se vor determina procentele persoanelor care au / nu au supravietuit, persoanelor din fiecare clasa (1, 2, 3) si persoanelor de sex masculin / feminin. Se obtine urmatoarea figura, cu 3 grafice de tip pie, realizate folosind biblioteca matplotlib.pyplot:

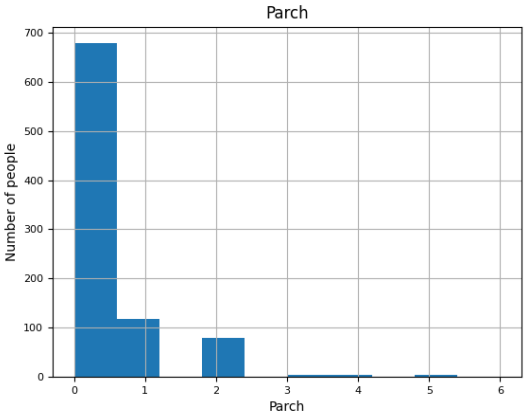


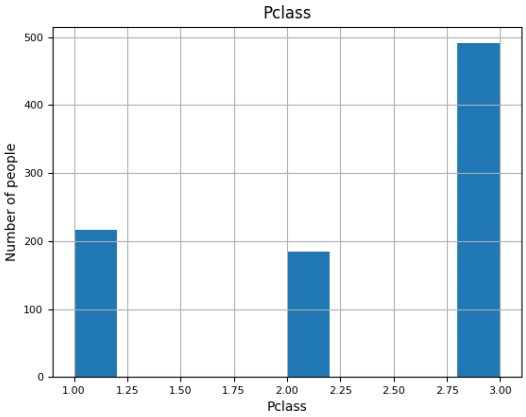
Cerinta 3

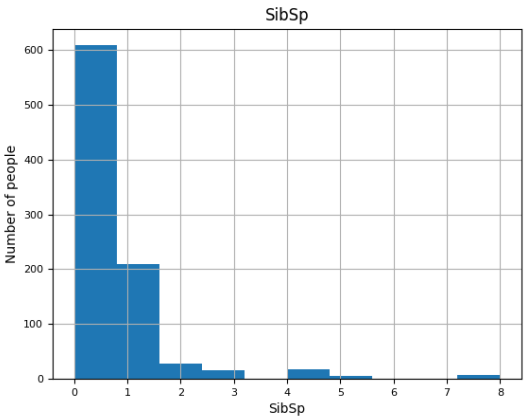
Pentru rezolvarea acestei cerinte vom folosi atributul dtype al fiecarei coloane din DataFrame pentru a determina coloanele cu valori numerice (int64 sau float64). Se obtine cate o histograma pentru fiecare coloana astfel determinata prin metoda hist():

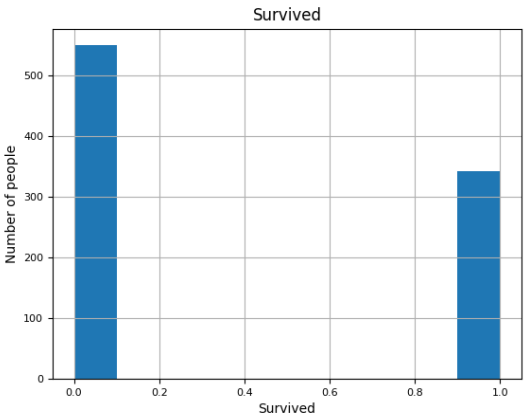

Age
Number of people
Age


Fare
Number of people
Fare


Parch
Number of people
Parch


Pclass
Number of people
Pclass


SibSp
Number of people
SibSp

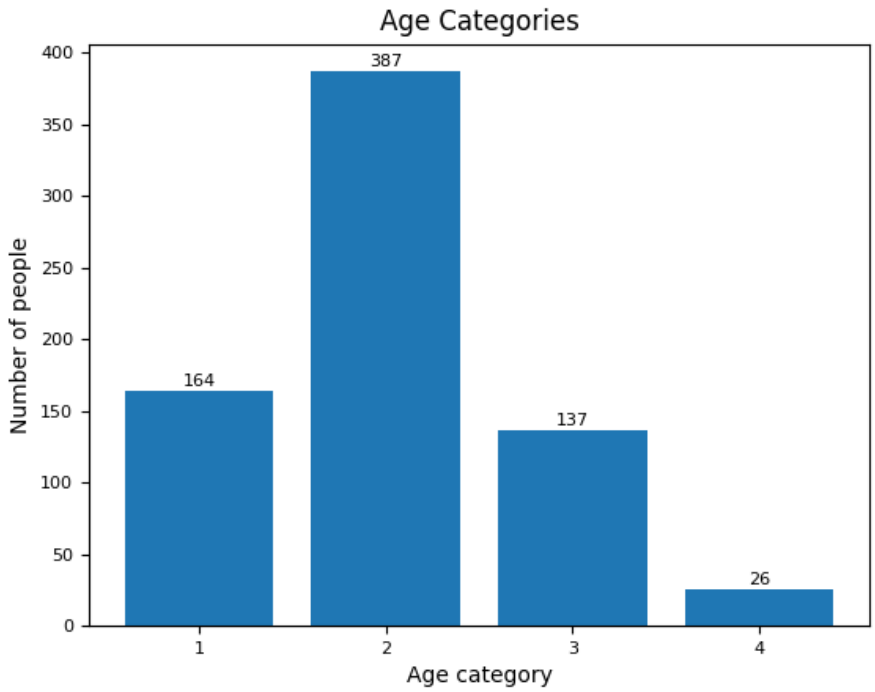

Survived
Number of people
Survived

Cerinta 4

Folosind combinatia de metode dropna() si count() se determina coloanele cu valori lipsa. Se calculeaza proportia valorilor lipsa, iar apoi se calculeaza din nou tinand cont si de starea de supravietuire a persoanelor.

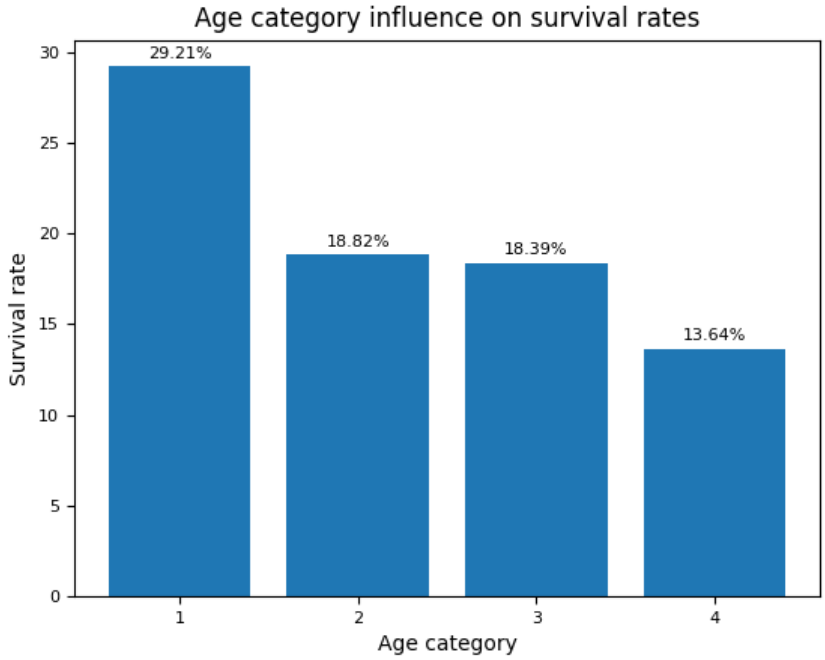
Cerinta 5

Dupa delimitarea categoriilor de varsta si setarea indicilor corespunzatori fiecarei categorii se va folosi metoda cut() pentru a crea o noua coloana in DataFrame, ce va contine indexul corespunzator categoriilor de varsta in care se incadreaza fiecare persoana. Cu ajutorul metodelor value_counts() si sort_index() se determina numarul de persoane din fiecare categorie de varsta, si se obtine urmatorul grafic:



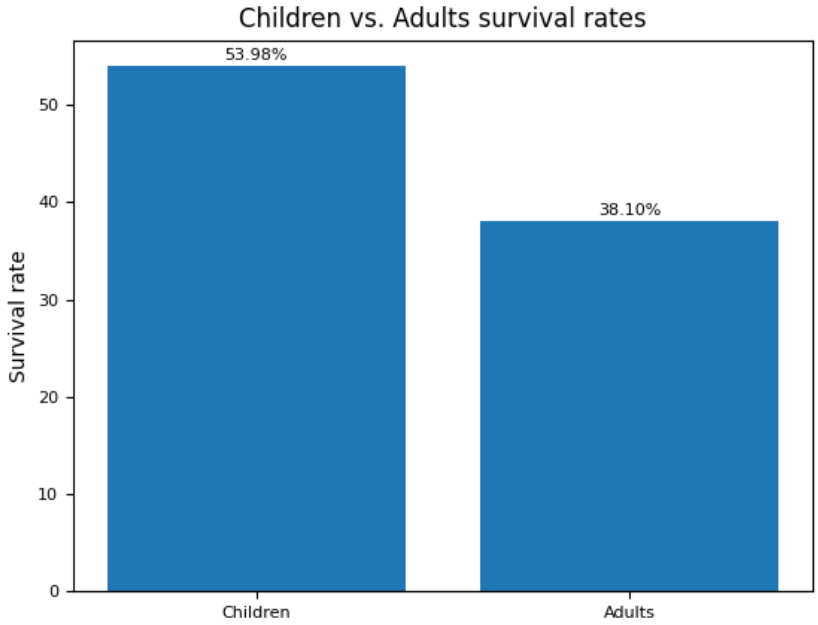
Cerinta 6

Folosind aceleasi metode ca la cerinta 5 si filtrand DataFrame-ul in functie de coloana ‘Survived’ se determina cati barbati din fiecare categorie de varsta au supravietuit, obtinand urmatorul grafic:



Cerinta 7

Pentru a determina procentul copiilor aflati la bord se va filtra DataFrame-ul in functie de varsta (titanic[‘Age’] < 18) si se vor numara liniile. Se determina ratele de supravietuire atat pentru copii cat si pentru adulti, rezultand urmatorul grafic:

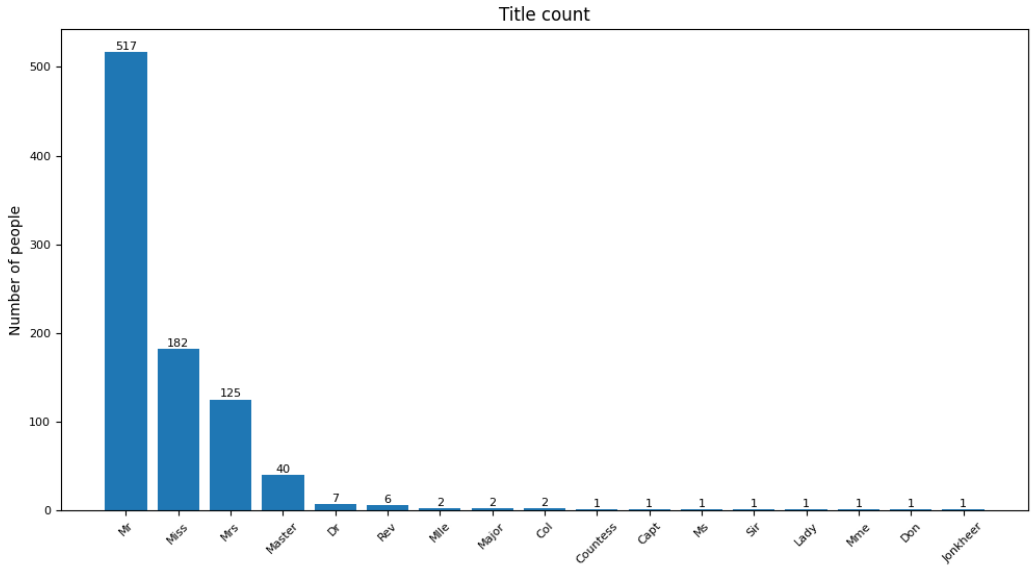


Cerinta 8

Se determina intai coloanele cu valori lipsa, in mod asemanator cerintei 4, iar apoi se determina tipul de date din acea coloana. Daca tipul este int64, se foloseste metoda fillna() pentru a umple coloana cu partea intreaga a mediei pe acea coloana (metoda mean()). Daca tipul este float64, procedeul este asemanator cu exceptia ca se umple coloana direct cu media obtinuta. Daca tipul de date este categoric, se folosesc metodele de la cerinta 5 pentru a determina cea mai frecventa valoare, iar coloana este umpluta cu ajutorul metodei fillna() cu valoarea determinata. Se vor salva noile date in fisierul de la calea ‘data/new_train.csv’.

Cerinta 9

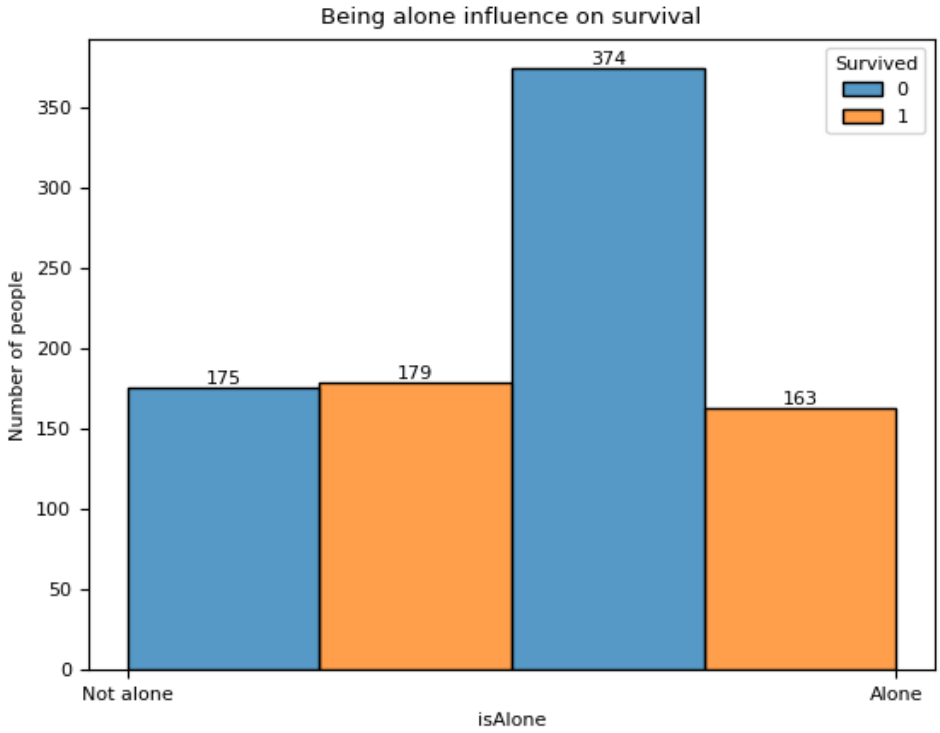
Pentru rezolvarea acestei cerinte am implementat o functie care se foloseste de expresii regulate pentru a extrage titlul din numele unei persoane. Aceasta functie este apelata prin metoda apply() pentru a crea o noua coloana in DataFrame ce contine titlurile aferente fiecarei persoane. Se determina cate persoane de fiecare sex corespund fiecarui titlu si se obtine urmatorul grafic:



Cerinta 10

Pentru rezolvarea acestei cerinte se creeaza o noua coloana in DataFrame, anume ‘isAlone’. Aceasta coloana contine valori de 0 sau 1 ce arata daca persoana este singura la bord sau nu (se iau in calcul fratii si sotii – SibSp, cat si parintii si copiii – Parch).

Folosind histplot() din biblioteca seaborn se obtine urmatorul grafic:



De asemenea, se obtine folosind catplot() din seaborn urmatorul grafic, ce urmareste relatia dintre tarif, clasa si starea de supravietuire pentru primele 100 de inregistrari: