Gălățeanu Andrei-Dan 311CC

30.05.2024

**Cerinta 1**: Cititi informatiile din fisierul train.csv si examinati structura acestora. Pentru acest lucru, trebuie să determinati programatic (utilizând cod Python) următoarele: numărul de coloane, tipurile datelor din fiecare coloană, numărul de valori lipsă pentru fiecare coloană, numărul de linii, dacă există linii duplicate.

Output:

A screenshot of a computer

Description automatically generated

->12 coloane

->891 linii

**Cerinta 2**: Determinati care este procentul persoanelor care au supravietuit si procentul persoanelor care nu au supravietuit. Determinati care este procentul pasagerilor pentru fiecare tip de clasă (coloana Pclass). Determinati care este procentul bărbatilor si care este procentul femeilor. Realizati un grafic potrivit pentru prezentarea acestor rezultate.

Output:

A screenshot of a computer screen

Description automatically generated

A graph of a number of different colored squares

Description automatically generated

A graph of a person and person

Description automatically generated

A green and red rectangular graph

Description automatically generated

**Cerinta 3**: Această cerintă implică generarea de histograme pentru fiecare coloană cu valori numerice din setul de date Titanic. O histogramă este o reprezentare grafică a distributiei frecventelor unei variabile continue. Pe axa orizontală sunt incluse intervalele de valori ale variabilei, iar pe axa verticală se reprezintă numărul de exemple din setul de date care sunt incluse în fiecare interval. Histograma oferă o imagine vizuală a modului în care valorile sunt distribuite si permite identificarea tendintelor si a modelului de distributie al datelor. În cadrul acestei cerinte, pentru fiecare coloană numerică din setul de date Titanic, se va realiza o histogramă pentru a vizualiza distributia datelor si a evidentia caracteristicile importante ale acestora.

Output:

A graph with numbers and a bar

Description automatically generatedA graph with blue bars

Description automatically generated

A blue graph with white text

Description automatically generatedA graph with blue bars

Description automatically generated

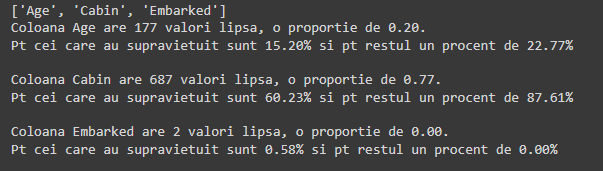
A graph with blue squares

Description automatically generated with medium confidenceA graph with numbers and a bar

Description automatically generated with medium confidence

**Cerinta 4**: Identificati coloanele pentru care există valori lipsă. Apoi, pentru fiecare coloană identificată determinati numărul si proportia valorilor lipsă. Determinati care este procentul acestora pentru fiecare dintre cele două clase (coloana Survived).

Output:

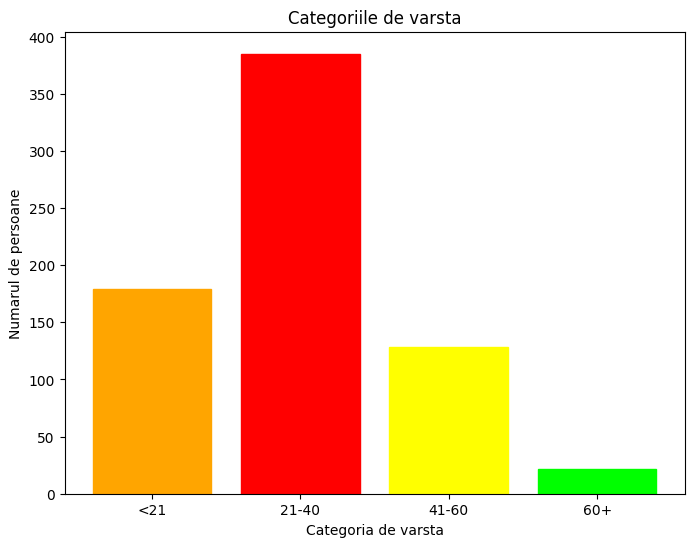


**Cerinta 5**: Considerăm patru categorii de vârstă: [0, 20], [21, 40], [41, 60], [61, max]. Determinati câti pasageri avem pentru fiecare din această categorie. Introduceti o coloană suplimentară si determinati pentru fiecare exemplu din setul de date indexul categoriei din care face parte. Realizati un grafic potrivit pentru a evidentia aceste rezultate.

Output:

A black background with white text

Description automatically generated



**Cerinta 6**: Determinati câti bărbati au supravietuit pentru fiecare dintre cele 4 categorii de vârstă propuse anterior. Realizati un grafic în care să evidentiati cum influentează vârsta procentul de supravietuire al bărbatilor, pe baza informatiilor pe care le avem în setul de date.

Output:

A graph with different colored squares

Description automatically generated

A screenshot of a graph

Description automatically generated

**Cerinta 7**: Determinati procentul copiilor aflati la bord (considerăm copii persoane cu vârsta < 18 ani). Realizati un grafic în care să evidentiati rata de supravietuire pentru copii si pentru adulti.

Output:

A graph of survival rate

Description automatically generated

**Cerinta 8**: Completati valorile lipsă cu cele obtinute pentru media pasagerilor care fac parte din aceeasi clasă. Spre exemplu, dacă există o înregistrare pentru un pasager care supravietuieste, dar pentru care nu cunoastem vârsta, completăm vârsta cu media pasagerilor care au supravietuit. În cazul în care avem o coloană cu valori categoriale, determinăm cea mai frecventă valoare pentru respectiva clasă.

Output:

A screenshot of a computer

Description automatically generated

**Cerinta 9**: Verificati dacă titlurile de noblete regăsite în coloana Name (Mr., Mrs., Don, etc.) corespund cu sexul persoanei respective. Reprezentati grafic câte persoane corespund fiecărui titlu.

Output:

A screenshot of a computer screen

Description automatically generated

A graph with different colored squares

Description automatically generated

**Cerinta 10**: A influentat starea de a fi singur pe Titanic (nu are deloc rude pe vas) sansele de supravietuire? Histograma ar putea ajuta la investigarea acestui aspect. Investigati relatia dintre tarif, clasă si starea de supravietuire pentru primele 100 de înregistrări folosind catplot() din seaborn. (sugestie: folositi kind=’swarm’ pentru a vedea detalii pe grafic).

Output:

A green and yellow graph

Description automatically generated

Observatii: se observa ca rata de supravietuire a fost de aprox 50% pt cei care nu erau singuri pe vapor. Cei care erau singuri au avut o rata de supravietuire mai mica de 30%.

A screenshot of a computer screen

Description automatically generated

A graph with numbers and dots

Description automatically generated

Observatii: se observa ca tariful a fost in general mai mare pentru cei de la clasa 1. (1 > 2 > 3). De asemenea, putem observa ca majoritatea celor de la clasa 3 nu au supravietuit.

Cea mai mare rata de supravietuire pare a fi in clasa a 2a.