Analiza setului de date Titanic
– Partea 1 –

Obiective

- Implementarea scriptului pentru determinarea diferitelor rezultate, a graficelor / histogramelor se realizeaza cu scopul analizei conditiilor si factorilor de supravietuire sau deces a persoanelor aflate in setul de date.
- Pregatirea datelor care are loc in cadrul acestei parti poate fi folosita pentru crearea unui model care sa prezica, de exemplu, daca o persoana a supravietuit sau pentru gasirea outlie-urilor din cadrul setului de date.

Implementare

1.Analiza informatiilor din fisierul ce contine setul de date

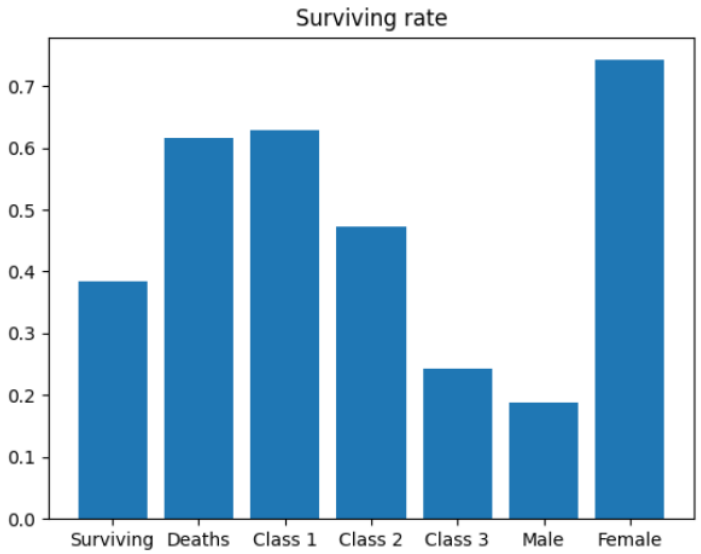
Prin selectarea, ca input, a analizei fisierului .csv, scriptul va oferi detalii precum numarul de coloane, tipul de date al fiecarei coloane, coloanele cu valori lipsa si existent liniilor duplicate.

```
Number of columns: 12
Column types :
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
```

```
Missing cells on columns:
PassengerId 0
Survived 0
Pclass 0
Name 0
Sex 0
Age 177
SibSp 0
Parch 0
Ticket 0
Fare 0
Cabin 687
Embarked 2
No dublicated rows
Number rows: 891
```

2.Rata de supravietuire / deces, in general, dupa sex si dupa clasa

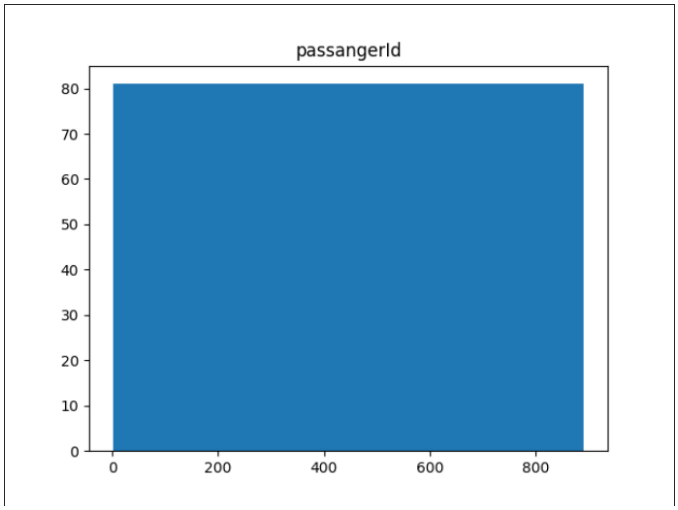
Selectarea comenzii de afisare a procentajelor de supravietuire va construi un grafic si va afisa pe ecran fiecare dintre acestea.

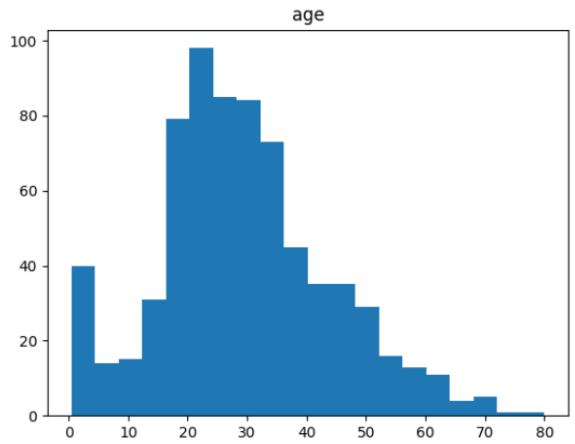


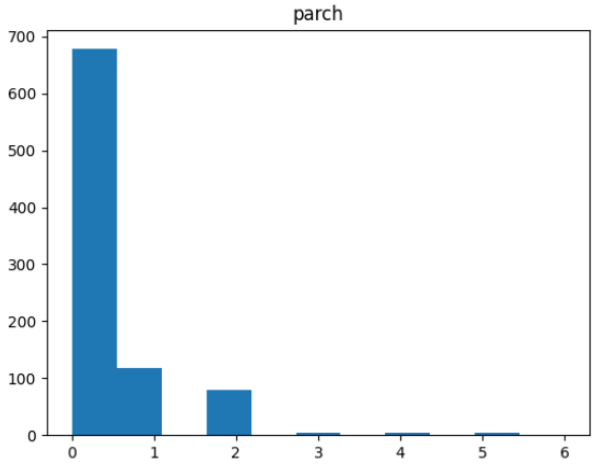

```
Survivors: 0.3838383838383838
Deaths: 0.6161616161616161
Class : 1 percentage : 0.6296296296296297
Class : 2 percentage : 0.47282608695652173
Class : 3 percentage : 0.24236252545824846
Male percentage: 0.18890814558058924
Female percentage: 0.7420382165605095
```

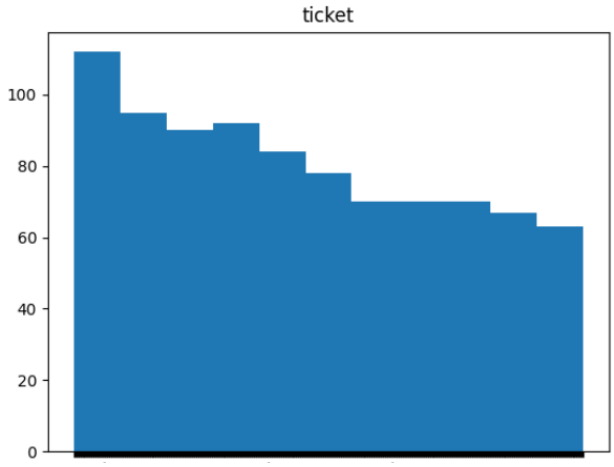
3.Afisarea histogramelor corespunzatoare coloanelor cu valori numerice

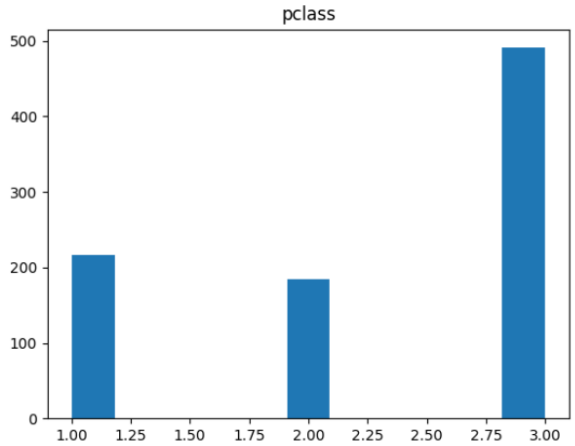
Pentru fiecare coloana care contine valori numerice (fie este numar intreg, fie zecimal, fie complex), se va construi care o histograma corespunzatoare acesteia.

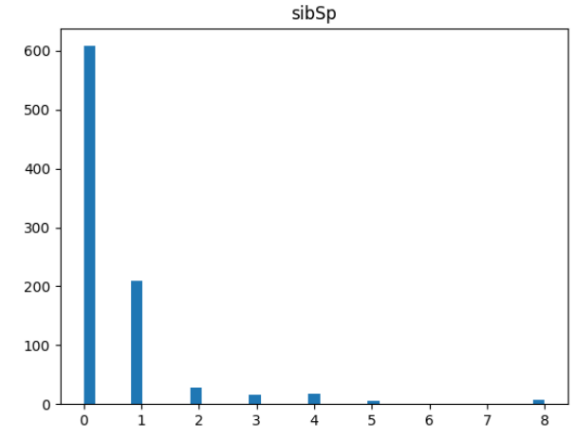


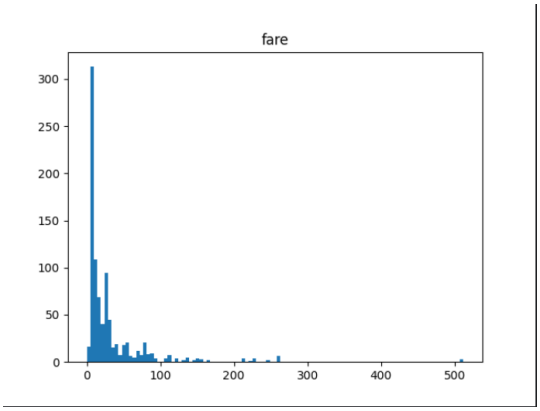


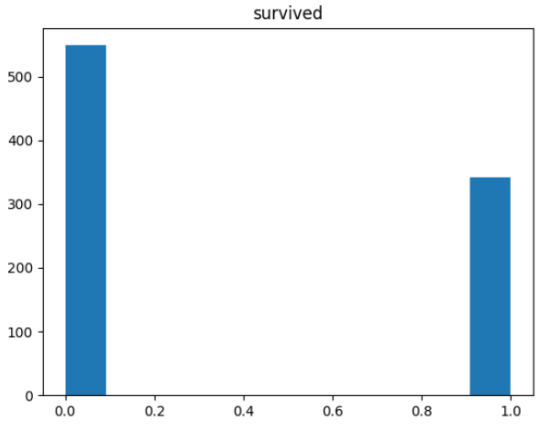












4.Afisarea numarului lipsa de valori de pe fiecare coloana

Prin selectarea comenzii de afisare a numarului lipsa de valori, pentru fiecare coloana, se vor numara acestea si se va calcula procentajul acestora, la general, dar si in particular pentru persoanele decedate si cele care au supravietuit.

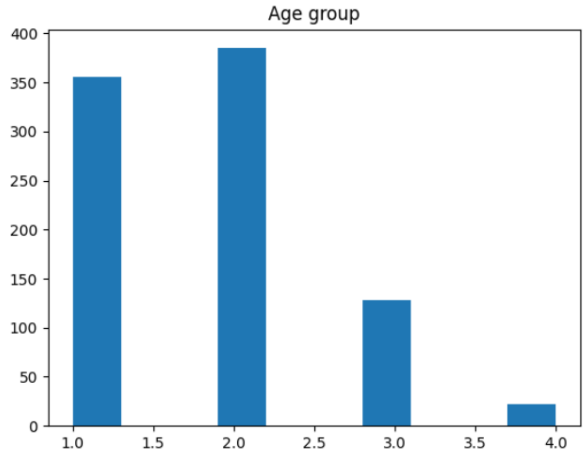
```
Missing values
name 0.0
passangerId 0.0
pclass 0.0
sex 0.0
age 0.19865319865319866
sibSp 0.0
parch 0.0
ticket 0.0
fare 0.0
cabin 0.7710437710437711
embarked 0.002244668911335578
survived 0.0
```

```
Missing values survived
name 0.0
passangerId 0.0
pclass 0.0
sex 0.0
age 0.15204678362573099
sibSp 0.0
parch 0.0
ticket 0.0
fare 0.0
cabin 0.6023391812865497
embarked 0.005847953216374269
survived 0.0
```

```
Missing values dead
name 0.0
passangerId 0.0
pclass 0.0
sex 0.0
age 0.22768670309653916
sibSp 0.0
parch 0.0
ticket 0.0
fare 0.0
cabin 0.8761384335154827
embarked 0.0
survived 0.0
```

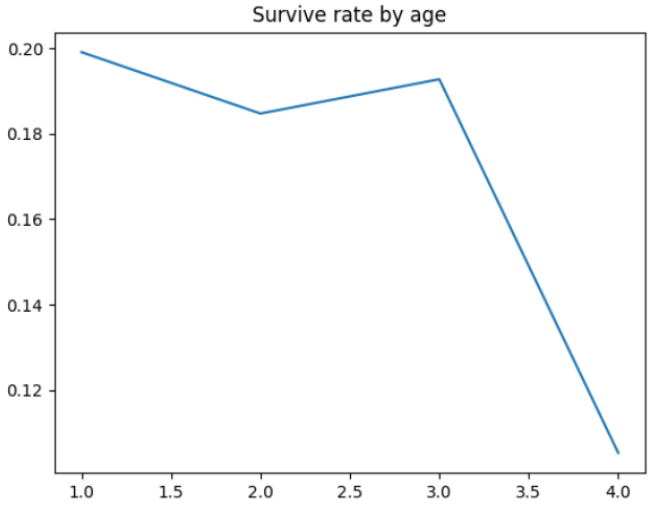
5.Adaugarea unei coloane care sa specific grupa de varsta caruia ii apartine fiecare pasager

Aceasta comanda va adauga o coloana care va indica daca un pasager apartine grupei de varste 1 (pana in 20 de ani), 2 (21 – 40 de ani), 3 (40 – 60 de ani), 4 (peste 60 de ani). De asemenea, se va afisa o histograma care va corespunde numarului de personane din fiecare grupa.



6.Afisarea unui graf a numarului de supravietuitori in functie de varsta

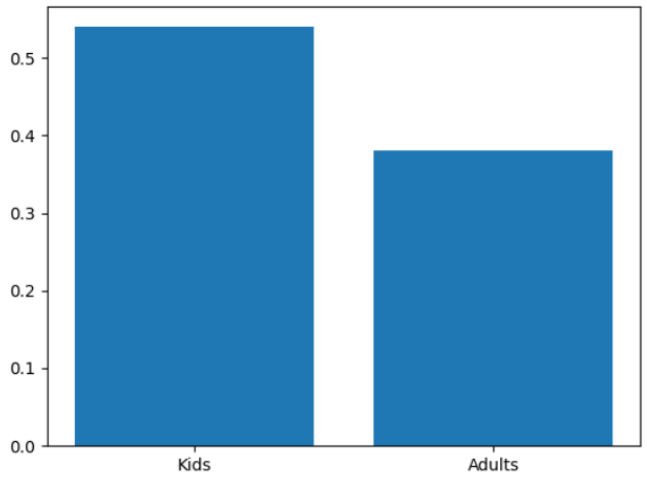
Scriptul va numara pentru fiecare grupa de varsta cati oameni au supravietuit si va construi un grafic pentru acestea.



7. Calcularea procentajului de copii de pe bord si compararea numarului de supravituitori copii vs adulti

Comanda va afisa pe ecran procentajul de copii (persoane sub 18 ani) aflati la bord si va reprezenta corespunzator grafic procentajul de copii supravietuitori, in comparatie cu numarul de adulti care nu au murit.

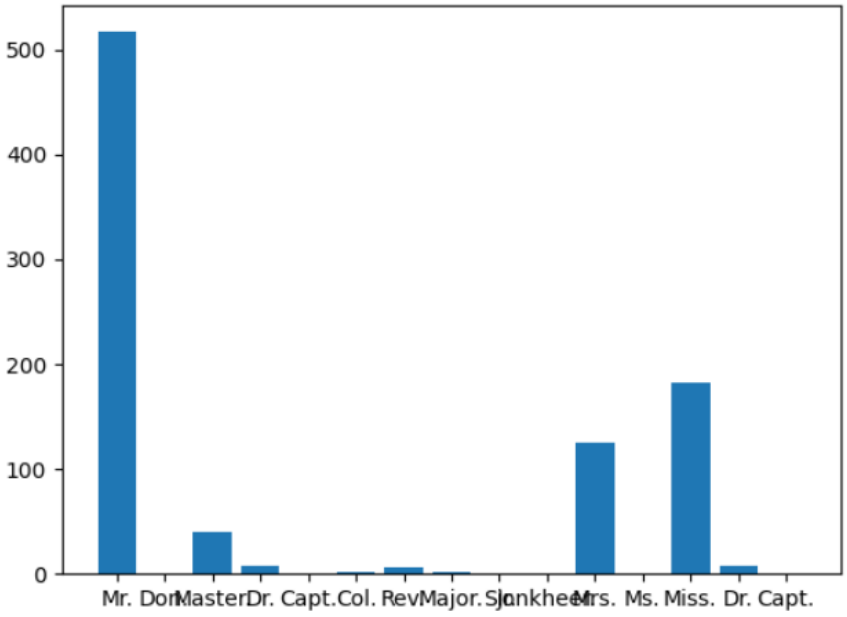
```
Kids percentage : 0.12682379349046016
```



8. Verificarea titlurilor de nume in functie de sex

La primirea acestei comenzi, scriptul va verifica daca setul de date are asociat fiecarei persoane un titlu corespunzator in functie de sex. Pentru setul Titanic, acest lucru se respecta, fiecare persoana avand un titlu corect. De asemenea, se va afisa un grafic care contine numarul de titluri folosite corect.

```
Name titles are correct : True
```



9.Inlocuirea valorilor lipsa

Scriptul va determina care sunt coloanele cu valorile lipsa si va incerca sa prezica ce valori s-ar potrivi cel mai bine acolo. Daca valorile sunt categoriale, de exemplu, clasa sau coloana de supravietuire, se va inlocui valoarea lipsa cu cea mai frecventa valoare a celorlalte linii. Daca valorile nu sunt categoriale, cum ar fi pretul calatoriei, varsta etc., se va inlocui valoarea lipsa cu media valorilor de pe celelalte linii. Cu aceste noi informatii se va crea fisierul no_missing_values.csv care nu va contine nicio valoare lipsa.

10.Analiza deceselor in functie de pretul calatoriei si clasa

Aceasta comanda va afisa grafic modul cum sunt distribuite decesele in functie de pretul calatoriei si clasa. Valoarea 0 va spune ca o persoana apartinand unei anumite clase, care a platit un anumit pret, a murit, iar valoare 1, ca acea persoana a supravietuit. Se observa ca persoanele apartinand unei clase mai

inalte, au avut o rata de supravietuire mai mare decat cei care au fost la clase inferioare.