PROIECT PCLP3

Avramoniu Calin-Stefan 312CC

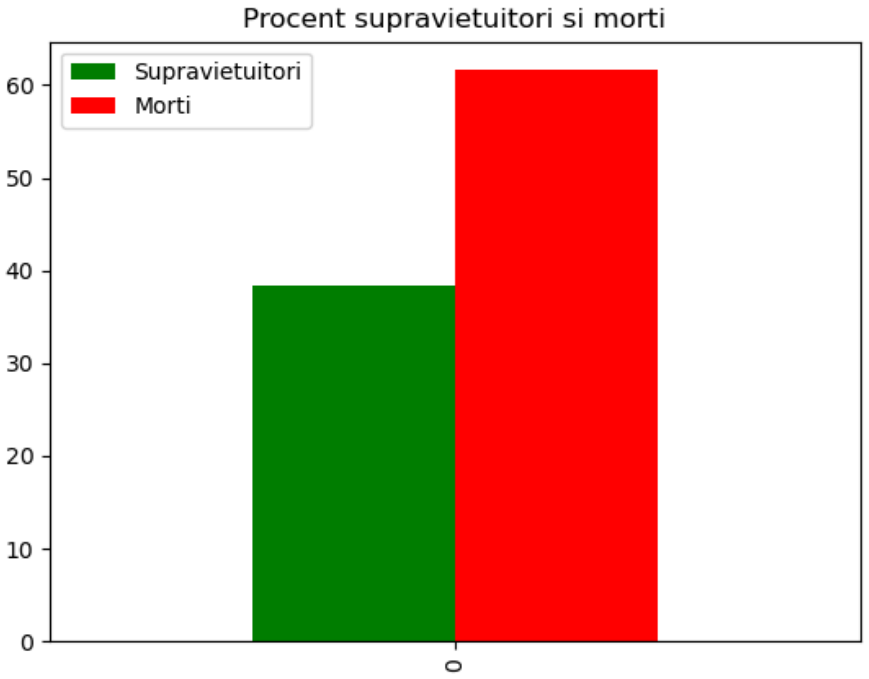
REZOLVARE PARTEA 1:

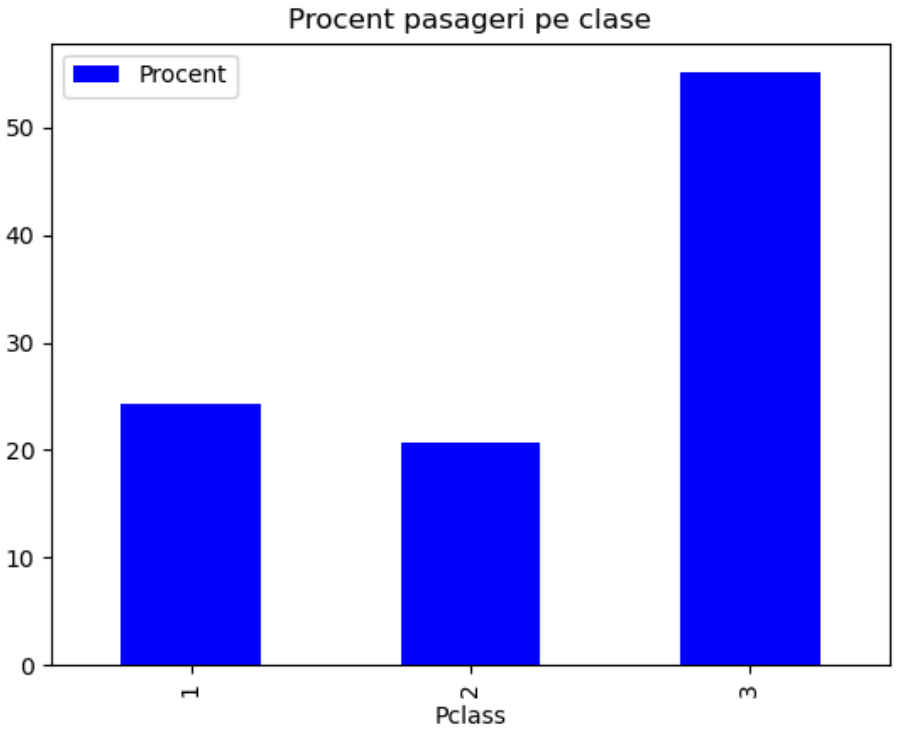
1. Citim fișierul csv și atribuim rezultatul unui dataframe, folosind pandas. Vom afișa variabilele *numar_coloane* , *numar_linii* si *numar_linii_duplicate*. De asemenea, vom afișa și numărul de valori lipsa pentru fiecare coloana.

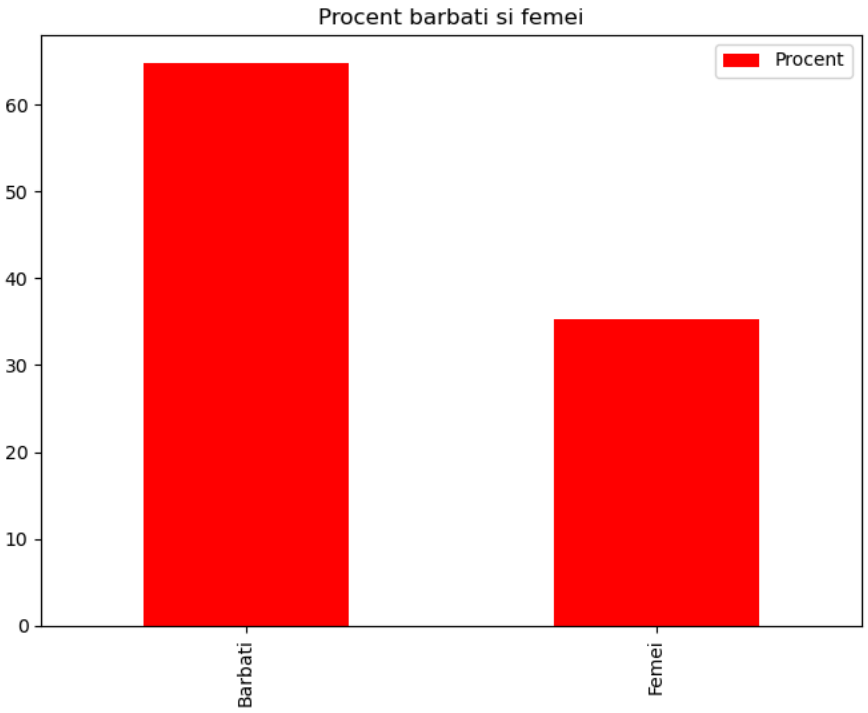
Output:

```
================================== CERINTA 1 ==================================


Nr col:  12
Nr linii:  891
Nr linii duplicate:  0


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

Valori lipsa pe fiecare coloana:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

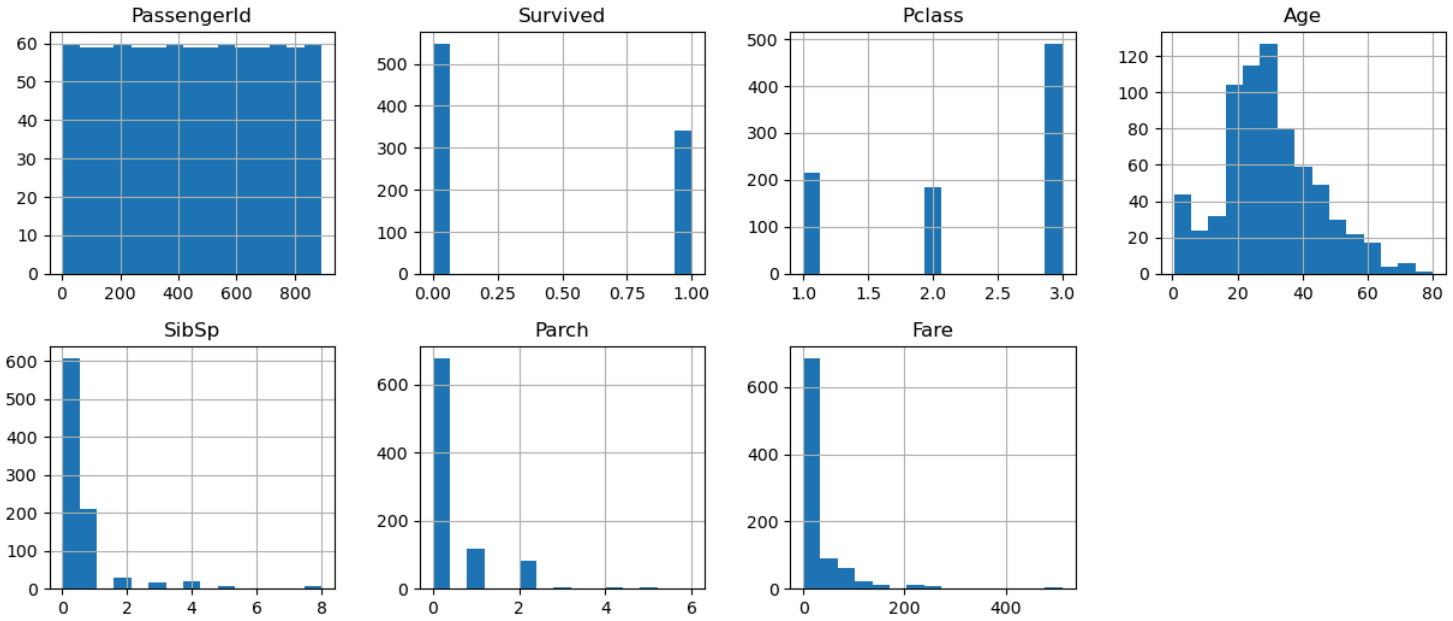
2. Vom calcula procentele persoanelor care au supraviețuit sau au murit, procentele pasagerilor pe clase, precum și procentele barbatilor si femeilor de pe vas. Apoi am creat cate un dataframe nou cu informații despre fiecare procent în parte și le-am afisat folosind matplotlib.







3. La cerința 3 am realizat histogramele pentru coloanele numerice din setul de date primit. Coloanele numerice sunt de tipurile: *int64* și *float64*. Histograma evidentiaza excelent cate persoane au murit și cate au supraviețuit, precum și distribuția pasagerilor atat pe clase cât și pe varste. Se observă astfel că sunt mai mulți pasageri la clasa a 3-a decât la celelalte 2 clase, iar intervalul predominant de varsta este 20-40 de ani.

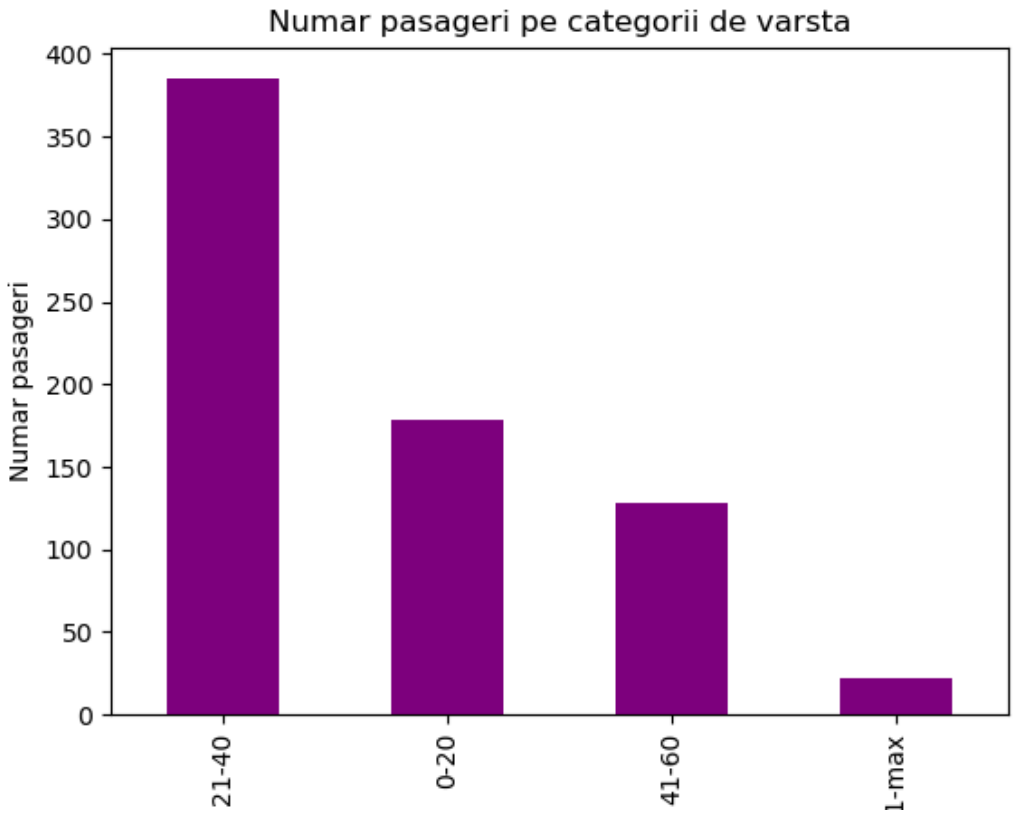


4. Pentru cerința 4 am identificat numărul de valori lipsa pentru fiecare coloana și procentul acestora din numărul total. Am afisat doar procentul valorilor lipsa, deoarece numărul acestora a fost afisat deja la cerința 1.

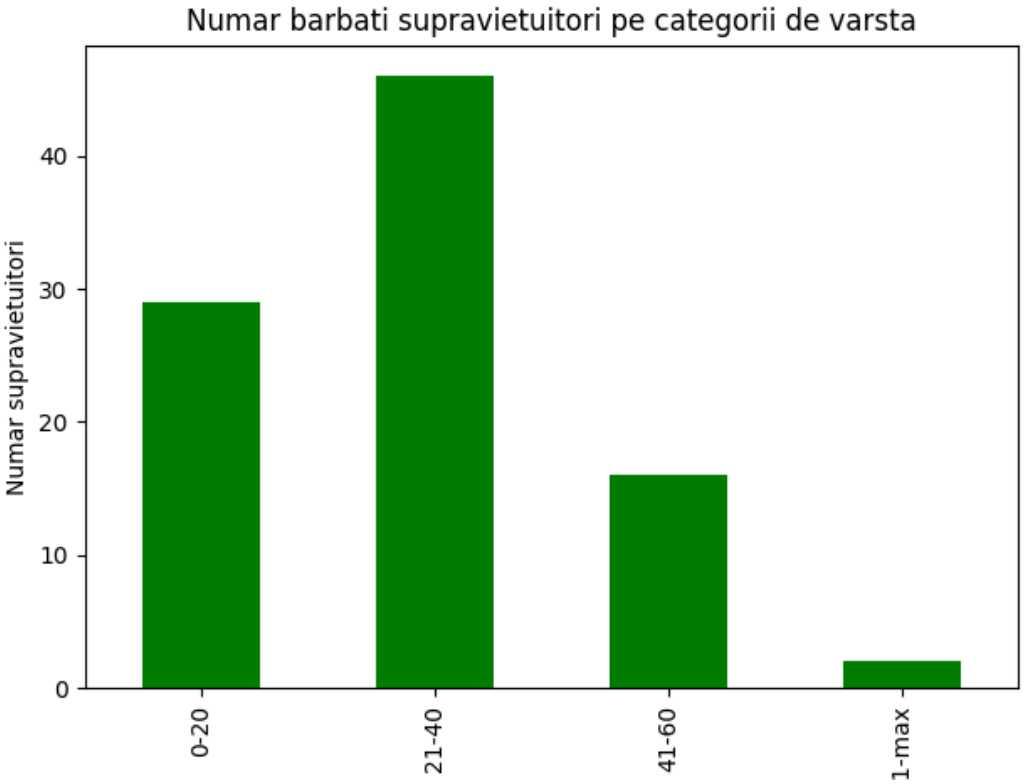
Output:

```
================================ CERINTA 4 ================================


Procent valori lipsa pe fiecare coloana:
PassengerId     0.00
Survived        0.00
Pclass          0.00
Name            0.00
Sex             0.00
Age            19.87
SibSp           0.00
Parch           0.00
Ticket          0.00
Fare            0.00
Cabin          77.10
Embarked        0.22
dtype: float64
```

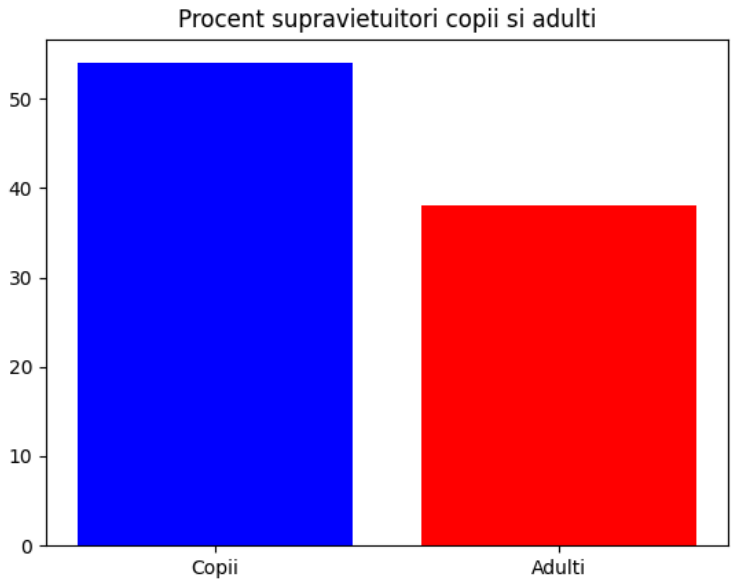
5. Am introdus coloana *'Categorie varsta'* în dataframe și am atribuit fiecărei persoane intervalul în care se situează varsta. Graficul de mai jos evidențiază distribuția pasagerilor pe categoriile de varsta.



6. La cerința 6 am realizat un grafic care evidentiaza cati barbati au supraviețuit pe fiecare categorie de varsta.



7. Am calculat numărul de minori și adulți care au supraviețuit, iar apoi am calculat procentajul acestora din numărul total de minori și adulți. Graficul următor arată distribuția supraviețuitorilor adulți și minori.



De asemenea, procentul copiilor la bord este de: **12.68%**

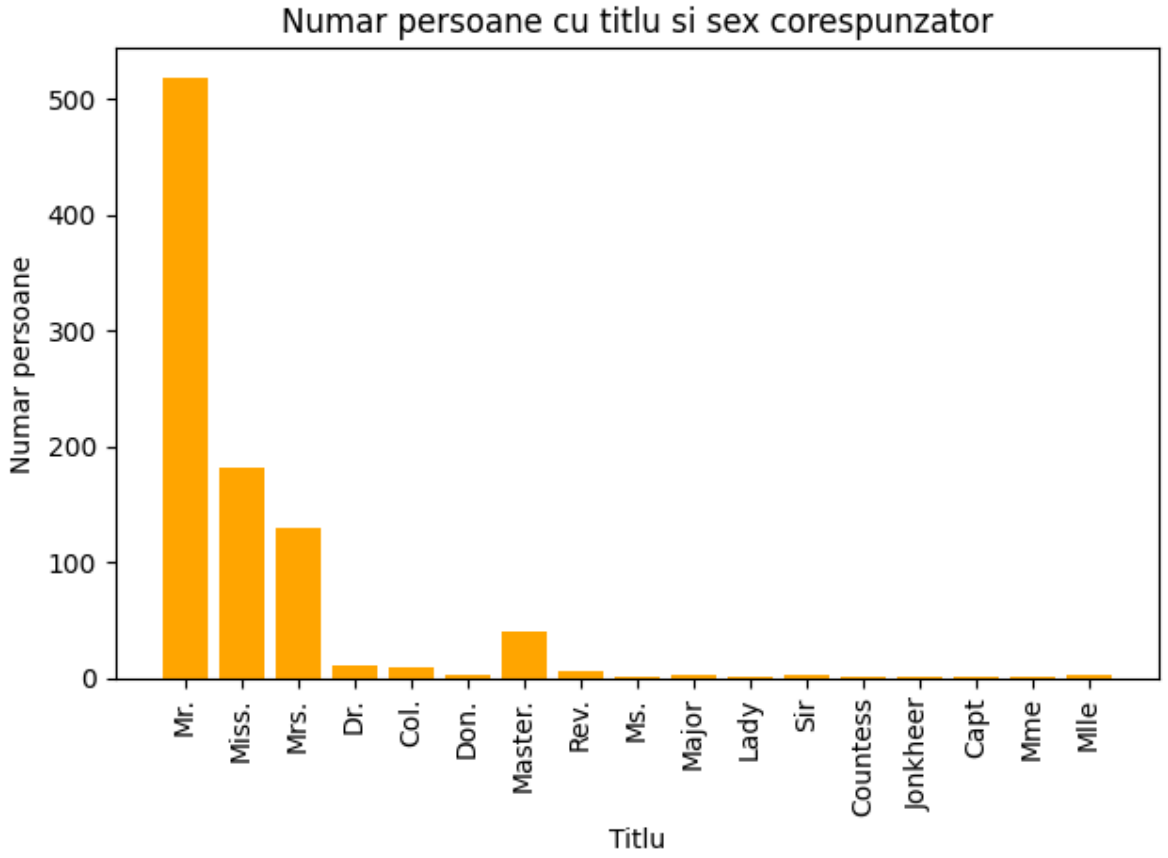
8. La cerința 8 am completat valorile lipsa din coloanele *'Fare'*, *'Age'*, *'Embarked'* și *'Categorie varsta'* conform enunțului. Astfel noul dataframe rezultat a fost salvat în fișierul *'rezultate_cerinta8.csv'*.

Output:

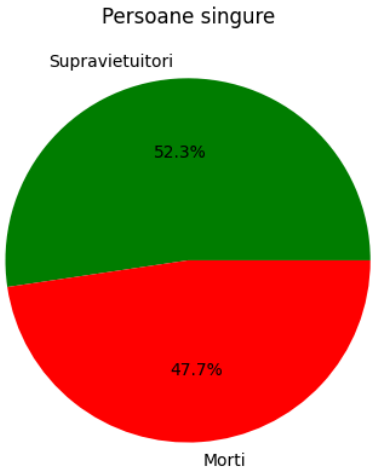
```
================================ CERINTA 8 ================================


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   PassengerId       891 non-null    int64
 1   Survived          891 non-null    int64
 2   Pclass            891 non-null    int64
 3   Name              891 non-null    object
 4   Sex               891 non-null    object
 5   Age               714 non-null    float64
 6   SibSp             891 non-null    int64
 7   Parch             891 non-null    int64
 8   Ticket            891 non-null    object
 9   Fare              891 non-null    float64
 10  Cabin             204 non-null    object
 11  Embarked          889 non-null    object
 12  Categorie varsta  714 non-null    category
dtypes: category(1), float64(2), int64(5), object(5)
memory usage: 84.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   PassengerId       891 non-null    int64
 1   Survived          891 non-null    int64
 2   Pclass            891 non-null    int64
 3   Name              891 non-null    object
 4   Sex               891 non-null    object
 5   Age               891 non-null    float64
 6   SibSp             891 non-null    int64
 7   Parch             891 non-null    int64
 8   Ticket            891 non-null    object
 9   Fare              891 non-null    float64
 10  Cabin             204 non-null    object
 11  Embarked          891 non-null    object
 12  Categorie varsta  891 non-null    category
dtypes: category(1), float64(2), int64(5), object(5)
memory usage: 84.7+ KB
None
```

Cele două df.info() din output evidențiază faptul că valorile lipsa au fost completate și salvate în dataframe.

9. Am creat un vector de titluri, iar apoi pentru coloana ‘Name’ am verificat care este countul fiecărui titlu și dacă acesta corespunde cu sexul persoanei (acolo unde este cazul, de exemplu dr. poate fi și femeie și bărbat). Graficul de mai jos evidențiază numărul de persoane care au titlul și sexul corespunzător.



10. Consider că, starea de a fi singur nu a influențat șansele de supraviețuire, așa cum evidentiaza si diagrama următoare.



De asemenea, am investigat și relația dintre tarif, clasa și starea de supraviețuire pentru primii 100 de pasageri. Putem observa ca cei de la clasa a 3-a au mai puțini membrii supraviețuitori, dar prețul biletelor este de asemenea mai mic comparativ cu celelalte două clase. Totodată, cei din clasa a 2-a au o rata mai buna de supraviețuire, deși prețul călătoriei este asemanator cu cel de la clasa a 3-a.

Rezultat folosind catplot ,kind=swarm :