

Name: Necula Mihail  
Group: 313CAa

## Readme Project PCLP3

The link of the project from the github is the following:  
[https://github.com/Mhail027/Proiect\\_PCLP3](https://github.com/Mhail027/Proiect_PCLP3)

### A. TASK 1

Reads the dataset and store it in a dataframe using the function `read_csv` from the module "panda". After:

- determinate the number of lines and of columns from dataset
- verify if the dataset has duplicate rows
- find the type of the values from every column
- find the number of missing values from every column

```
***** TASK 1 *****
Number of lines: 891
Number of columns: 12

Doesn't exist duplicates.

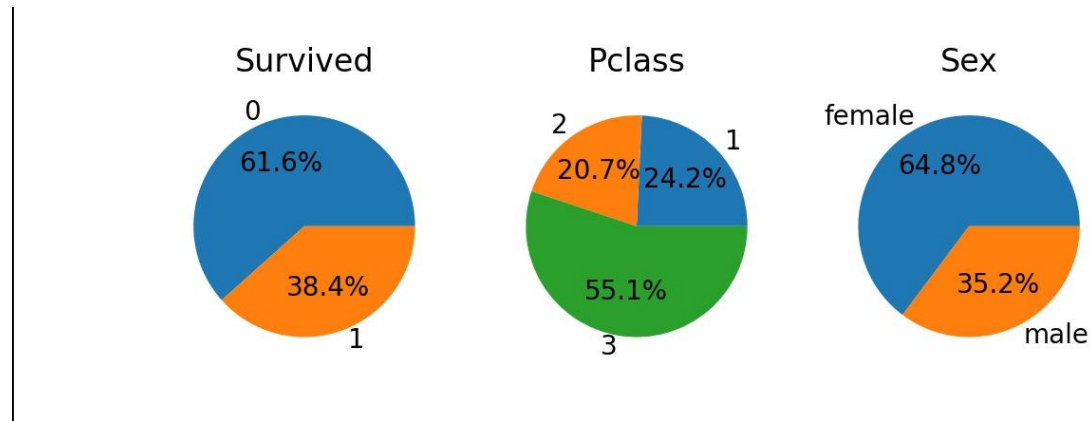
Column      Type
PassengerId  int64
Survived     int64
Pclass       int64
Name         object
Sex          object
Age         float64
SibSp        int64
Parch        int64
Ticket       object
Fare         float64
Cabin        object
Embarked     object
dtype: object

Column      Missing values
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
```

## B. TASK 2

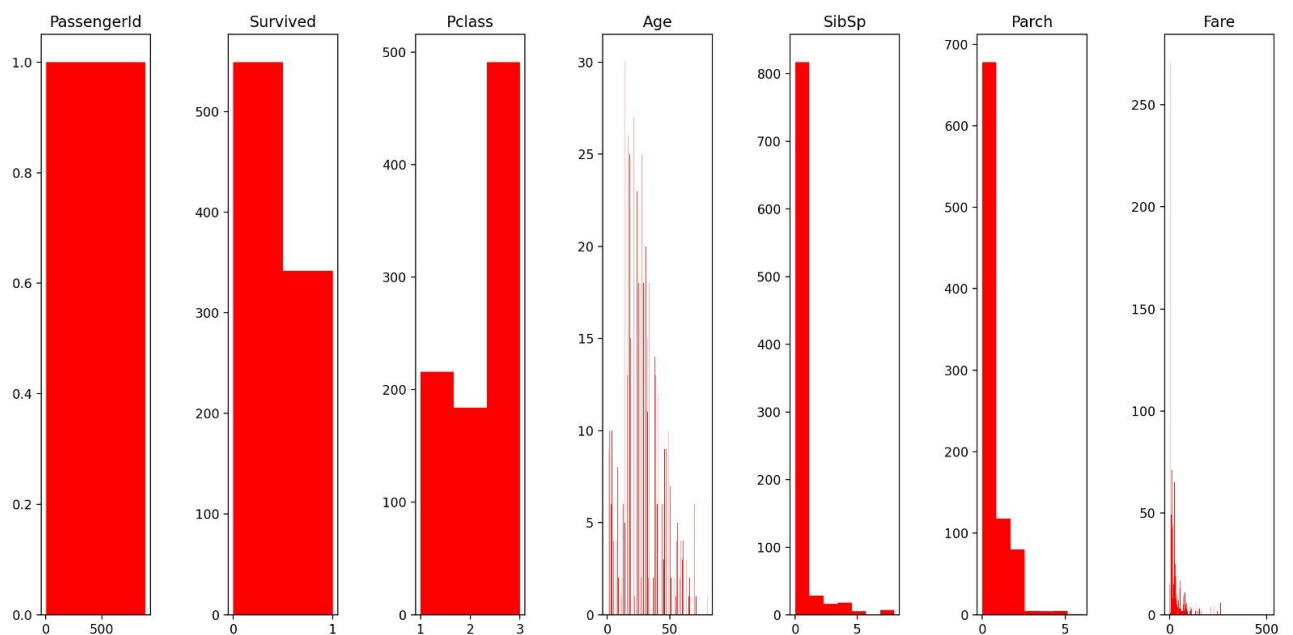
Process the columns "Survived", "Pclass" and "Sex". For every processed column is made a graphic of type pie. **The operation of processing a column includes the next steps:**

- find all the options / values from the column
- find of how many times every option appears in the column



## C. TASK 3

Process the columns which have just numerical values. For every processed column is made a histogram.



## D. TASK 4

Find the number of missing values from every column. For every column which has holes, is printed on the screen how many they are and the percentage of holes from all values which should be.

After, find the numbers of characteristics / values which are missing for the persons

which survived, respectively for the people that died. On the screen, we print the percentage of missing values from all values which should be for every class from the category "Survived".

```
***** TASK 4 *****
Columns with missing values
Age: 177 values - 19.86%
Cabin: 687 values - 77.1%
Embarked: 2 values - 0.22%

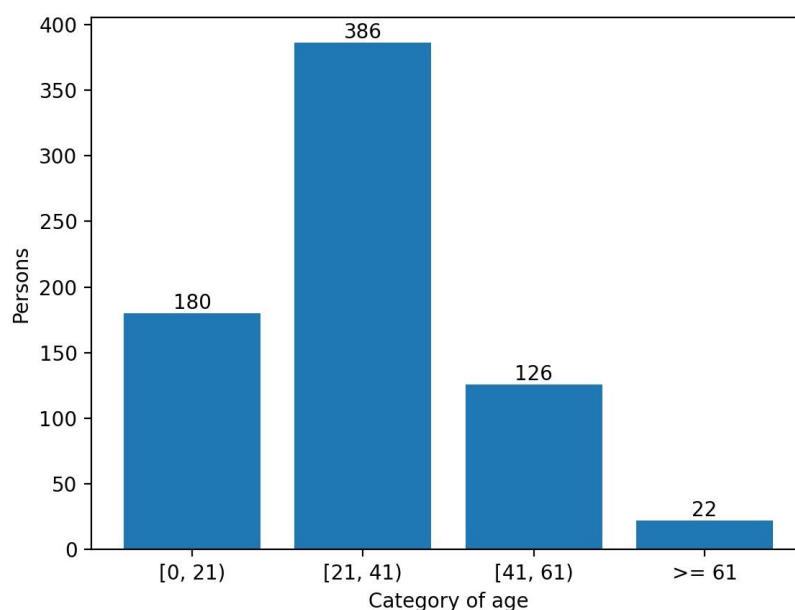
Percentage of missing values for deads
0.09 %
Percentage of missing values for survivors
0.06 %
```

#### E. TASK 5

Create a list which contains the category of age for every person. The categories of age are:

- [0, 21) years -> category 0
- [21, 41) years -> category 1
- [41, 61) years -> category 2
- over 61 years -> category 3

This list is added in the dataframe as a new column. After we do this, we count the number of persons from every category of age and make a graphic which contains these informations.

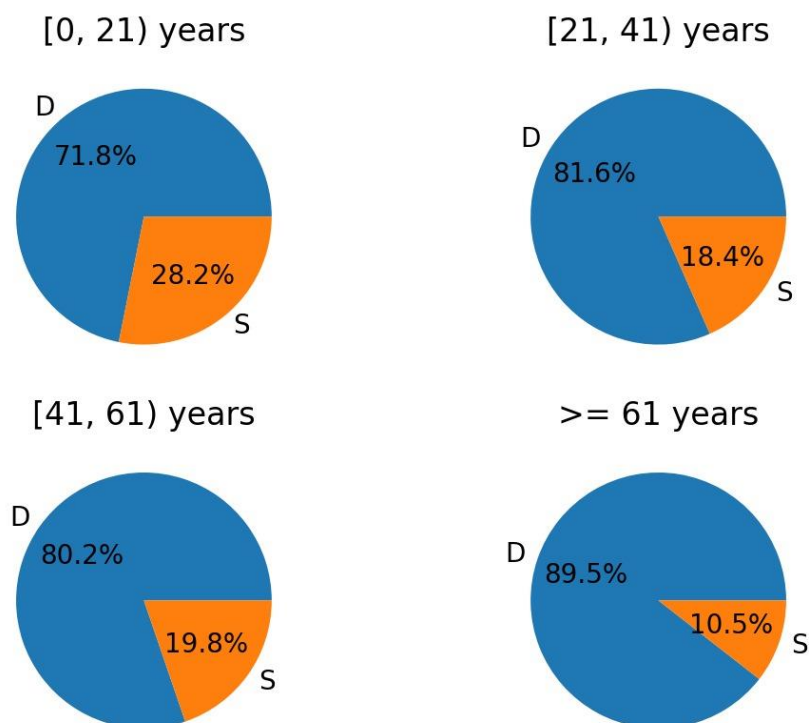


#### F. TASK 6

Add the column "Category of age" in dataframe and count how many male survived and died in every category of age. Print the number of male survivors on screen, for every category, and make a graphic with this informations.

```
**** TASK 6 ****
Male survivors
[0, 21) years: 29
[21, 41) years: 46
[41, 61) years: 16
>= 61 years: 2
```

Men



#### G. TASK 7

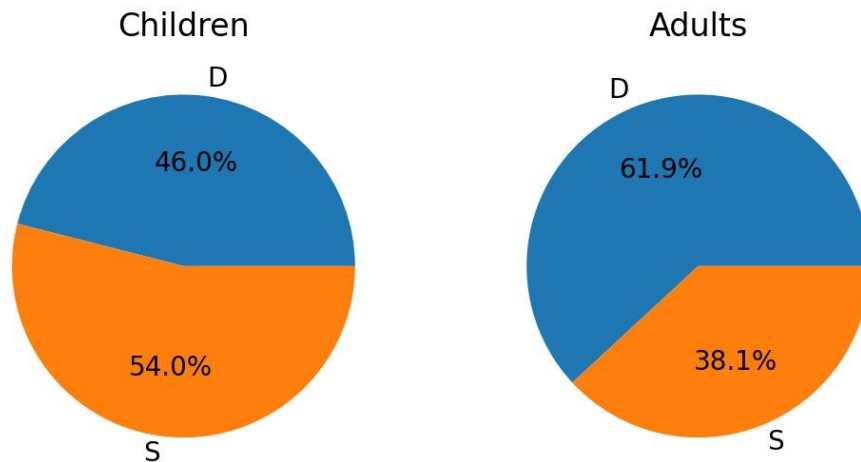
Find the number of children (< 18 years) and adults which survived and died. Calculate the percentage of children from the ship and print the result on screen.

After, we do a graphic of type pie which contains the informations about the adults and their existence after Titanic. We do, the same thing for children.

## \*\*\*\*\* TASK 7 \*\*\*\*\*

### Percentage of children from ship: 0.15%

Rate of survival



#### H. TASK 8

We fill up the holes from the dataframe. We have 3 columns with missing values: "Age", "Cabin", "Embarked".

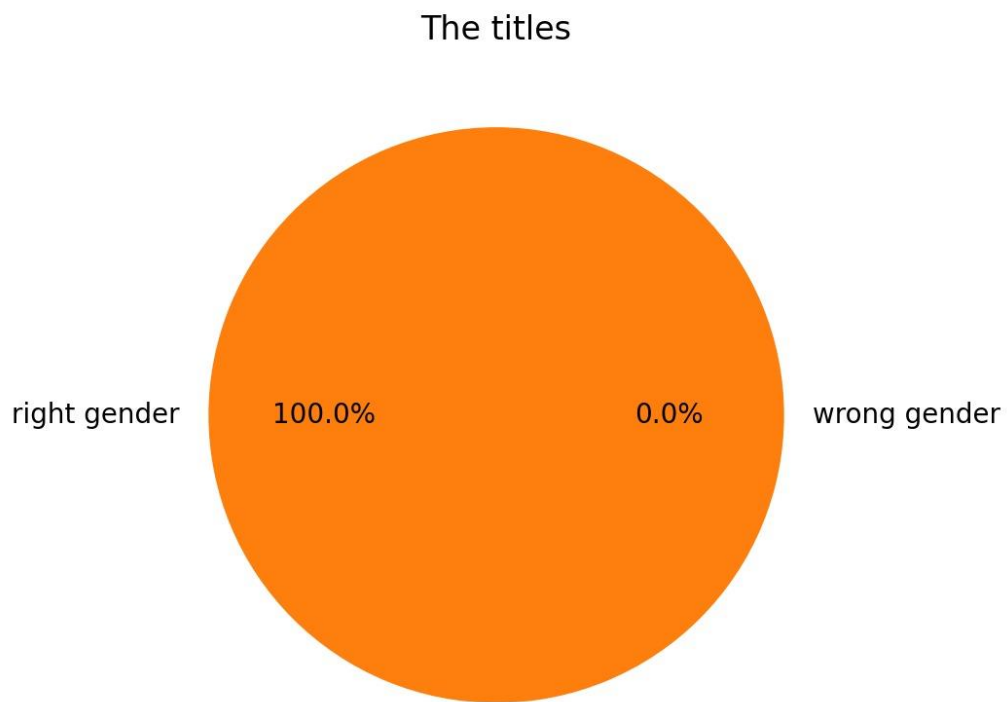
Age - We calculate the medium age for a survivor and for a person who died. If a person survived, but we don't know his age, we put the medium age of a survivor. The same thing is done and for a person who died, but have hasn't the age known.

Cabin, Embarked - Because these columns have string values, we must work a little differently. For every column, firstly we determine the most frequent option of survivors and fill the survivors's holes from the column. Secondly, we do the same thing for the person which, unfortunately, died.

The completed dataframe is saved in the file with the next name : "completed\_dataframe.csv"

#### I. TASK 9

Split the column "Name" to do a column with the titles of the people. We do 3 lists: one with the titles for men, another with the titles for women, and the last with the neutral titles. We go through the column of "Title" and "Sex" and count the number of wrong and right pairs / titles. We plot the results.



#### J. TASK 10

Take first 100 persons from the dataset and do a graphic with the columns: "Survived", "Pclass" and "Fare". The purpose of this graphic is to analyze how the class and the fare influenced the life of the persons.

The fare didn't influenced much, but the class yes. We see an increasing rate of death from second class to first class and from first class to third class.

