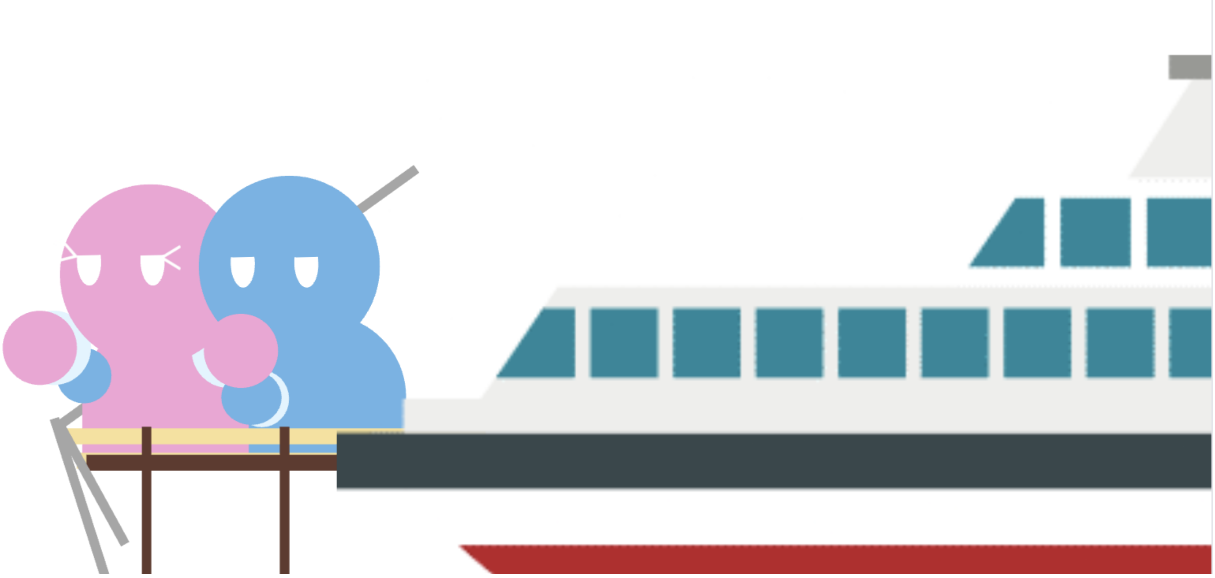
英語版または日本語版のいずれかにご記入ください。

Assignment Lesson 1-2

- Data Science Process -

Prior Knowledge and Data Preparation processes are needed before we further go to model the data. In this assignment, be creative and try to explain the two processes of the data science process with the provided template. Here is the case:



***The Titanic** is one of the most infamous shipwrecks in history. During her maiden voyage, the Titanic sank after colliding with an iceberg on April 15, 1912. Out of 2224 passengers and crew, 1502 died. In the aftermath of this sensational tragedy, there were better safety regulations for ships enacted by the international community. Survival depends in part on luck, but we want to know whether some groups of people seem to have been more likely to survive than others. [Here is the dataset](). (In Rapidminer, it can also be found under this repository: Samples > data > Titanic). To complete this assignment feel free to choose any statistical tool you preferred. (Rapidminer, Python, Excel, Google Sheets, etc.)*

Prior Knowledge

● Objective 10/10

Write down the objective of this case.

Find if specific groups of people have been more likely to survive than others

● Subject Area 10/10

Write some subject areas that are ideally needed to understand the dataset:

It is important to know:

- Which population was represented on board the Titanic?
- Who was the main population on board the Titanic?
- The plan of the boat (accessibility of the lifeboats, the cabins…)
- If there were enough lifeboats?
- Is there any difference between a passenger and a crew member?
- How did people get rescued?
- If people did not get on a lifeboat, did they die before or we simply don't have the

● Data:5/5

How many examples and attributes it has?

There are 1309 examples (809 who did not survive and 500 who did) and 12 attributes

Is the dataset enough to answer the objective? Why? **10/10**

The dataset represents

- 59% of the whole population of the Titanic
- 53% of the people who died
- 70% of the survivors

More than half of the population on board the Titanic is represented. However, the percentage of survivors is quite high in our dataset compared to the whole population. The dataset does not seem to be a good representative sample of the whole population. It may be better to increase the dataset.

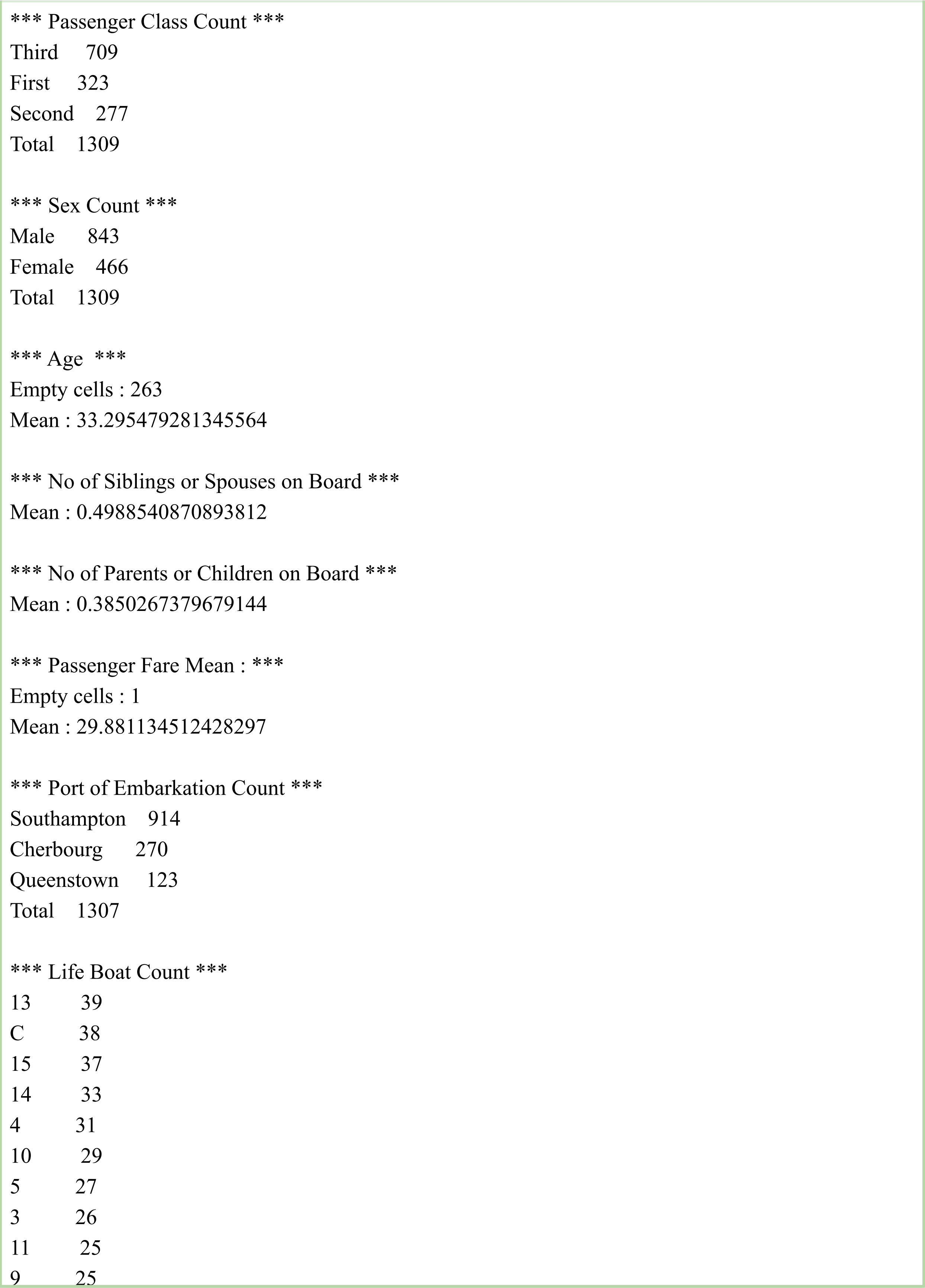
Identify the roles of the attributes (ID, label, regular, prediction, if any).**4/5**

```
Name: Regular (there are homonyms)
Passenger Class: Regular
Sex: Regular
Age: Regular
No of Siblings or Spouses on Board: Regular
No of Parents or Children on Board: Regular
Ticket Number: Regular
Passenger Fare: Regular
Cabin: Regular
Port of Embarkation: Regular
Life Boat: Regular
Survived: Label
```

Data Preparation

- **Data Exploration 10/10**

Write down the summary (count for categorical attributes, mean for numerical attributes) of all columns. (other than Name, Ticket Number, and Cabin). Just a screenshot of that information from your preferred statistical tools is ok.



```
*** Passenger Class Count ***
Third     709
First     323
Second    277
Total     1309

*** Sex Count ***
Male      843
Female    466
Total     1309

*** Age  ***
Empty cells : 263
Mean : 33.295479281345564

*** No of Siblings or Spouses on Board ***
Mean : 0.4988540870893812

*** No of Parents or Children on Board ***
Mean : 0.3850267379679144

*** Passenger Fare Mean : ***
Empty cells : 1
Mean : 29.881134512428297

*** Port of Embarkation Count ***
Southampton    914
Cherbourg      270
Queenstown     123
Total     1307

*** Life Boat Count ***
13        39
C         38
15        37
14        33
4         31
10        29
5         27
3         26
11        25
9         25
```

Missing Value 0/3 6/7

By seeing which part of the data that are missing, can you explain how you plan to handle them in order to reach the objective?

For the numerical attributes, we can replace the missing values with the mean if there are no outliers.
If there are outliers and if we don't remove them, we can replace the missing values with the median.
For the Life Boat, it might be better to not use this attribute as there are more empty cells than filled ones.
For the Port of Embarkation, it might be better to fill the empty cells with a different value such as "Unknown".

Data Types: 10/10

Identify the types of each attribute.

```
Name                                   string (Nominal)
Passenger Class                        string (Nominal)
Sex                                    string (Binominal)
Age                                   float64 (Numerical)
No of Siblings or Spouses on Board      int64 (Numerical)
No of Parents or Children on Board      int64 (Numerical)
Ticket Number                          object (Nominal)
Passenger Fare                        float64 (Numerical)
Cabin                                  object (Nominal)
Port of Embarkation                    object (Nominal)
Life Boat                              object (Nominal)
Survived                               object (Binominal)
```

Data Exploration: Grouping 14/15

Separate the "Age" attribute into 5 groups of the same interval. Check how many survived for each of those groups.

If we don’t count the empty cells:

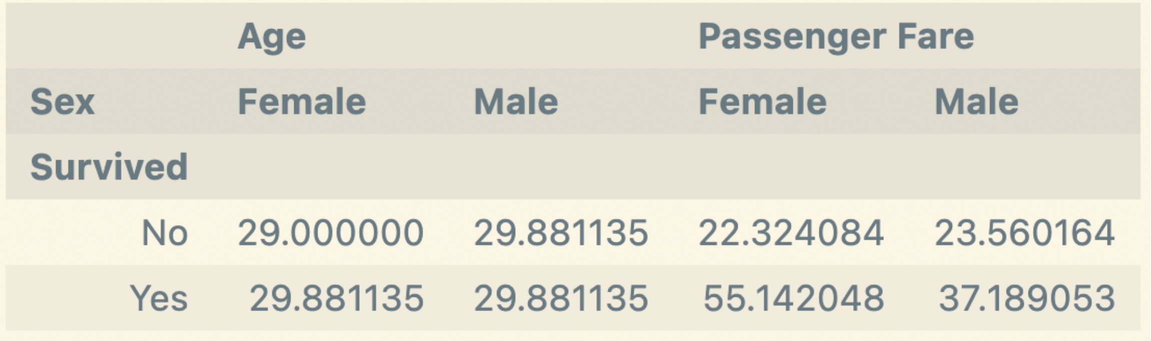
```
Age_Label      Survived
(16.1, 32.1]   No          327
               Yes         197
(32.1, 48.1]   No          162
               Yes         107
(0.09, 16.1]   Yes          74
               No           60
(48.1, 64.0]   No           59
               Yes          47
(64.0, 80.0]   No           11
               Yes           2
```

If we replace the empty cells with the mean value:

```
Age_Label      Survived
(16.1, 32.1]   No          517
               Yes         270
(32.1, 48.1]   No          162
               Yes         107
(0.09, 16.1]   Yes          74
               No           60
(48.1, 64.0]   No           59
               Yes          47
(64.0, 80.0]   No           11
               Yes           2
```

● Data Exploration: Pivoting 12/15

Make a pivot table to check the average passenger fare and the median age for each sex with survival as its column attributes.



	Age		Passenger Fare	
Sex	Female	Male	Female	Male
Survived				
No	29.000000	29.881135	22.324084	23.560164
Yes	29.881135	29.881135	55.142048	37.189053