Wrangle Report

1.Missing data in Age :

We create new column called Title that have the title of every one by extracted it from the column “Name” using the function (str.extract)

We create a dataframe for only the title that contain ‘Mr’,new dataframe for ‘Miss’ ,new dataframe for ‘Mrs’ , new dataframe for ‘Master’ , new dataframe for ‘Dr’ and new dataframe for ‘Rev’

For cleaning the missing value for column ‘Age’ we decided to change every missing value in the age column to the mean of every age title so if the missing value of age in row that contain title ‘Mr’ we replaced it to the mean of the age for the dataframe that contain only mr , if the missing value of age in row that contain title ‘Miss’ we replaced it to the mean of the age for the dataframe that contain only miss , if the missing value of age in row that contain title ‘Masters’ we replaced it to the mean of the age for the dataframe that contain only masters and so on.

2.Missing data in 2 rows of Embarked Column :

so
we inplaced 2 missing values in embarked with the mode ’S’.
value counts ⇒ S = 644, C = 168, Q = 77
so the mode is ‘S’ as a mode of this column.

3.Missing data in 687 rows of Cabin Column:

Due to the high number of missing values, this variable is directly dropped from the analysis.
so
We dropped this column from the dataset.

4.

because of the data type of the column of Age is float and we want to be an integer number, But we cannot neglect or delete the number of months, especially if there are children less than a year old. we divided the age column into 2 columns, the first for the number of years and the second for the number of months, and we made the datatype to each of them as an integer number

---

5.

Each variable must form a column and Each observation must form a row .
SO
we divided the Embarked column to 3 columns , each of them represent a variable (S,C,Q),
and each row of them represent one observation .
people who embark in ‘S’ take 1 in ‘S column’ and zeros in the other columns , and so on.

6.

we can merge the 2 columns of sibsp and parch in one column called (Family Size).
because sibsp column represent the number of siblings / spouses of the person aboard the Titanic, and parch column represent the number of parents / children of the person aboard the Titanic
So they represent the same variable, then his/her family size = sibsp+parch

---

7.

we changed the datatype of ‘Sex’ column to integer by replacing ‘Male’ by 0, and ‘Female’ by 1 .

8.

We clean the ‘Name’ column by split it into 3 columns ’first_name’,’last_name’ and ‘Title’ and then we drop column ‘Name’
We cleaned the “frist_name “ column by removing the title from it so we used the method (str.split) and then created a new column that contains only the second value after the “.” And then we drop the “first_name” column

---

9.

We changed the datatype of column ‘Title’ by replacing each title in title column by a number ‘Mr’⇒1,’Miss’⇒2, 'Mrs'⇒3,'Master'⇒4 ,'Dr' ⇒ 5 'Rev’⇒6

['Col', 'Mlle', 'Major'] ⇒7 we replaced them by the same number because they have the same number of occurance (2)
['Countess', 'Lady', 'Capt', 'Sir', 'Jonkheer', 'Don' ,'Ms ', 'Mme']⇒ 8 we replaced them by the same number because they have the same number of occurance (1)
