Name :Yehia Ahmed Hassan EL-Boudy (20191311656).
Ahmed Hany Mohamed Abdulawahab (20191310663).
Mamoun Mohamed Hassan Abdelbaey (20191312895).
ID Number : 20191311656.
Subject : Data science methodolgy.
To Professor :Dr/ Magda Matbouly

Project

- In this project we are working on a data set of the a ship called “titanic” in which it was a cruise trip in the ocean and in the middle of the way is sank.
- This data set contains various information about the passengers : considered as columns.
- The columns and their explanation:

1. PassengerId: Unique Id of a passenger
2. Survived: If the passenger survived(0-No, 1-Yes)
3. Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: Name of the passenger
5. Sex: Male/Female
6. Age: Passenger age in years
7. SibSp: No of siblings/spouses aboard
8. Parch: No of parents/children aboard
9. Ticket: Ticket Number
10. Fare: Passenger Fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

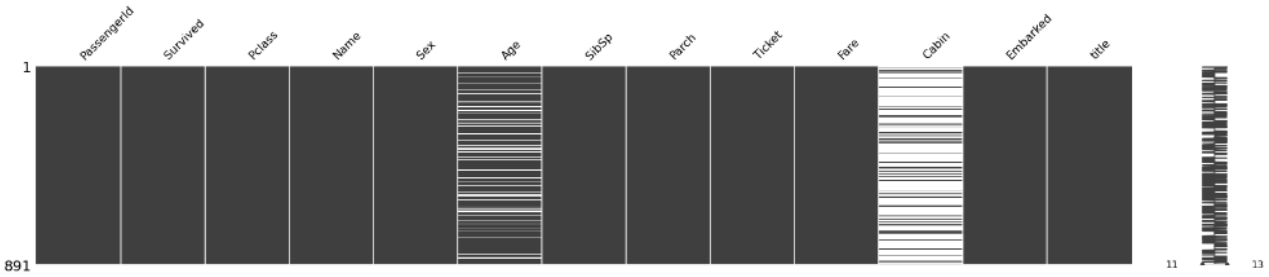
➢ Data wrangling:

- *Visualizing our nulls*:

```
In [172]: #this function helps ut to visualize the null value in each column
          missingno.matrix(train,figsize=(30,5))
```

❖ "missingno" is a function that helps us visualizing our nulls in the data.

```
In [132]: #this function helps ut to visualize the null value in each column
          missingno.matrix(train,figsize=(30,5))
Out[132]: <matplotlib.axes._subplots.AxesSubplot at 0x1610c3f5ee0>
```



❖ The columns that include the white color means it has nulls.

- *Creating a data frame:*
  - ❖ this data frame includes the data type of every feature, missing values, unique values, and count called "train_data_dict".

- *Setting the passenger id as the index.*

- *Filling the null values:*
  - ❖ First we will begin in this step with the fare column.

    When we checked the missing values in this column/feature, we didn't find any missing values. So we wanted to know if it includes any zero values ,because its not logical to be going on the trip for free.

```
In [136]: #checking if we have zero values in fare
          print((train['Fare']==0).sum())
15
```

    We found out 15 values equal to 0.
  - ❖ we replaced them with nulls with the replace function.

```
In [137]: #changing zeros to null
          train.Fare=train.Fare.replace(0,np.NAN)
```

❖ when we viewed the fare's column null values , we concluded that they are embarked from the same place, they were all male ,however, they we were from different classes. So we will be filling the nulls with respect to the Pclass.

❖ We created 3 data frames: each data frame includes each class, either its class one , two, or three.

```
In [145]: #Fare null values with respect to Pclass1
          train_p1=pd.DataFrame(train[train.Pclass==1])

In [146]: #Fare null values with respect to Pclass2
          train_p2=pd.DataFrame(train[train.Pclass==2])

In [147]: #Fare null values with respect to Pclass3
          train_p3=pd.DataFrame(train[train.Pclass==3])
```

❖ Filling each fare null value with the median that corresponds to its Pclass.

```
]: #Filling fare null values with respect to Pclass1
   train_p1.Fare.fillna(train_p1.Fare.median(),inplace=True)

]: #Filling fare null values with respect to Pclass3
   train_p2.Fare.fillna(train_p2.Fare.median(),inplace=True)

]: #Filling fare null values with respect to Pclass3
   train_p3.Fare.fillna(train_p3.Fare.median(),inplace=True)
```

❖ Then we concatenated the three data frames in just one data frame.

```
n [154]: train=pd.concat([train_p1,train_p2,train_p3])
         train.head()
```

ut[154]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Mrs.
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Mrs.
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	Mr.
12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S	Miss.
24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.5000	A6	S	Mr.

- *Filling the age null values:*
  - ❖ When we checked if age had any zero value , we didn't find any. So we decided to fill the age with the mean of the ages , since its normally distributed.

- *The cabin feature:*
  - ❖ We found out that 77% of the fields of the cabin column contains nulls.

```
In [160]: train.Cabin.isnull().mean()
Out[160]: 0.7710437710437711
```

- The Embarked feature:
  - ❖ There is only one null value.

<u>Conclusion:</u> we created a new data frame that includes all the features except the cabin feature (we dropped this column).

```
In [162]: #Droping the cabin feature because it has alot of null values
trainML=train[['Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket',
       'Fare', 'Embarked', 'title']]
```

In addition to that we also removed the row that had a null value in the embarked feature.

Finally we are having a clean data set called "TrainML".

```
trainML.head()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	title
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	Mr.
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	Mrs.
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	Miss.
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S	Mrs.
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S	Mr.

Remark: we divided our data set into two set , one to train our model and the other to test our model and see how accurate is this.

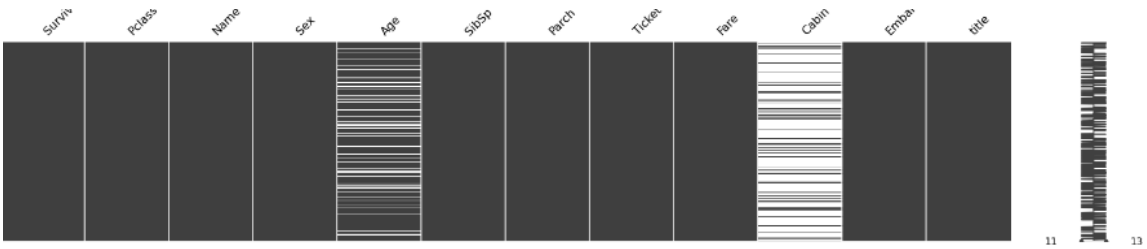
- **Test Data set:**
  -

```
In [168]: test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

  -
    - *Visualizing our nulls*:

```
In [*]: #this function helps ut to visualize the null value in each column
        missingno.matrix(test,figsize=(30,5))
```



- **Filling the nulls:**
  - In the age feature, we filled the nulls with the mean of the whole column values.
  - In the fare feature, we just found out one row having a null value ,so we dropped that row from the testing data set.
  - In the cabin feature, we had a lot of nulls . So we dropped the whole column.

```
In [184]: testML=test[['Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare','Embarked'
```

```
In [185]: testML.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 418 entries, 892 to 1309
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Pclass    418 non-null    int64
 1   Name      418 non-null    object
 2   Sex       418 non-null    object
 3   Age       418 non-null    float64
 4   SibSp     418 non-null    int64
 5   Parch     418 non-null    int64
 6   Ticket    418 non-null    object
 7   Fare      415 non-null    float64
 8   Embarked  418 non-null    object
dtypes: float64(2), int64(3), object(4)
memory usage: 32.7+ KB
```

At this point, we have cleaned the two data sets, testing and training, and we are ready to test our and view the prediction of our model.