Double-click (or enter) to edit

1 . Download the dataset from Kaggle

```
from google.colab import files
uploaded = files.upload()
```

Choose Files Titanic-Dataset.csv

- **Titanic-Dataset.csv**(text/csv) - 61194 bytes, last modified: 12/24/2021 - 100% done

```
Saving Titanic-Dataset.csv to Titanic-Dataset.csv
```

2 a. Import The library

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

b. Import the dataset

```
dataset=pd.read_csv("Titanic-Dataset.csv")
```

```
dataset
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S

```
dataset.head(3)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

```
dataset.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine	female	NaN	1	2	W./C. 6607	23.45	NaN	S

```
dataset.shape
```

```
(891, 12)
```

```
dataset.info()
```

```
<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
```

```
dataset.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

c. Checking for Null values

```
dataset.isnull().any()
```

```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

```
dataset.isnull().sum()
```

```
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
```

there are null values in age,cabin and embarked column

```
dataset.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
				Futrelle,								

d . Data Visualization

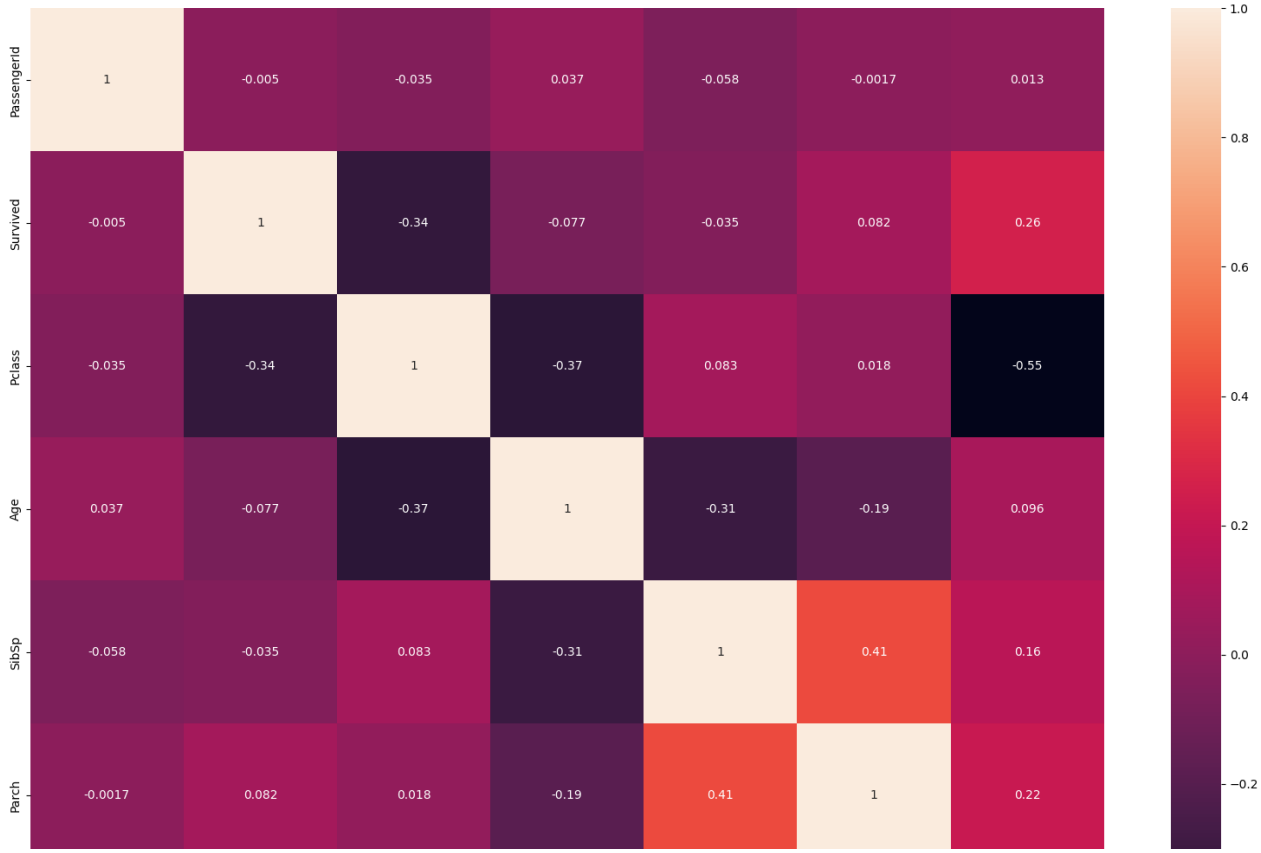
```
corr=dataset.corr()
corr
```

```
<ipython-input-15-f22ca9e9dc13>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecat
  corr=dataset.corr()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

```
plt.subplots(figsize=(20,15))
sns.heatmap(corr,annot=True)
```

```
<Axes: >
```




```
dataset.Survived.value_counts()
```

```
0    549
1    342
Name: Survived, dtype: int64
```

PassengerId Survived Pclass Age SibSp Parch Fare

```
dataset.Age.value_counts()
```

```
24.00    30
22.00    27
18.00    26
19.00    25
28.00    25
         ..
36.50     1
55.50     1
0.92      1
23.50     1
74.00     1
Name: Age, Length: 88, dtype: int64
```

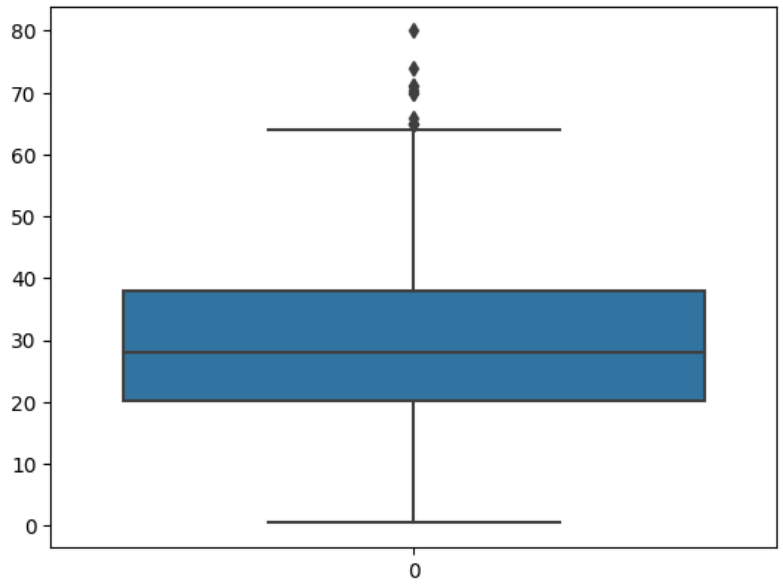
```
dataset.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
				Futrelle,								

e. Outlier Detection

```
sns.boxplot(dataset.Age)
```

```
<Axes: >
```



f. Splitting dependent and independent Variable

```
x=dataset.iloc[:,3:13]
y=dataset.iloc[:,13:14]
```

```
x.head()
```

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

```
y.head()
```

0
1
2
3
4

```
dataset.shape
```

```
(891, 12)
```

```
x.shape
```

```
(891, 9)
```

```
y.shape
```

```
(891, 0)
```

g. Perform Encoding

Label encoding on sex column

```
from sklearn.preprocessing import LabelEncoder
```

```
le=LabelEncoder()
```

```
x["Sex"]=le.fit_transform(x["Sex"])
```

```
x["Sex"]
```

```
0      1
1      0
2      0
3      0
4      1
      ..
886    1
887    0
888    0
889    1
890    1
Name: Sex, Length: 891, dtype: int64
```

```
x["Sex"].value_counts()
```

```
1    577
0    314
Name: Sex, dtype: int64
```

```
x["Sex"].nunique()
```

```
2
```

```
x.head()
```

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	NaN	S
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C85	C
2	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	C123	S

```
x.Cabin.value_counts()
```

```
B96 B98        4
G6             4
C23 C25 C27    4
C22 C26        3
F33            3
              ..
E34            1
C7             1
C54            1
E36            1
C148           1
Name: Cabin, Length: 147, dtype: int64
```

One hot encoding on cabin column

```
x.shape
```

```
(891, 9)
```

```
cabin=pd.get_dummies(x["Cabin"],drop_first=True)
```

```
cabin
```

	A14	A16	A19	A20	A23	A24	A26	A31	A32	A34	...	E8	F E69	F G63	F G73	F2	F33	F38	F4	G6	T
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
887	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
888	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
889	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
890	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

891 rows × 146 columns

```
x=pd.concat([x,cabin],axis=1)
```

```
x.head()
```

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	A14	...	E8	F E69	F G63	F G73	F2	F33	F38
0	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	NaN	S	0	...	0	0	0	0	0	0	0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C85	C	0	...	0	0	0	0	0	0	0
2	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	0	...	0	0	0	0	0	0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	C123	S	0	...	0	0	0	0	0	0	0
4	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	NaN	S	0	...	0	0	0	0	0	0	0

5 rows × 155 columns

```
x.drop(["Cabin"],axis=1,inplace=True)
```

```
x.head(10)
```

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	A14	A16	...	E8	F E69	F G63	F G73	F2	F33	F38
0	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	S	0	0	...	0	0	0	0	0	0	0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C	0	0	...	0	0	0	0	0	0	0
2	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	S	0	0	...	0	0	0	0	0	0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	S	0	0	...	0	0	0	0	0	0	0
4	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	S	0	0	...	0	0	0	0	0	0	0
5	Moran, Mr. James	1	NaN	0	0	330877	8.4583	Q	0	0	...	0	0	0	0	0	0	0
6	McCarthy, Mr. Timothy J	1	54.0	0	0	17463	51.8625	S	0	0	...	0	0	0	0	0	0	0
7	Palsson, Master. Gosta Leonard	1	2.0	3	1	349909	21.0750	S	0	0	...	0	0	0	0	0	0	0
	Johnson, Mrs																	

```
x.shape
```

```
(891, 154)
```

h. Splitting data into train and test

9	Nicholas	0	14.0	1	0	237736	30.0708	C	0	0	...	0	0	0	0	0	0	0

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
```

```
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

```
((623, 154), (268, 154), (623, 0), (268, 0))
```

```
a=[1,2,3,4,5,6]
b=[1,0,1,5,6,3]

for i in range(5):
    a_train,a_test,b_train,b_test=train_test_split(a,b,test_size=0.3,random_state=100)
    print("with random state",a_train)
```

```
 with random state [5, 4, 6, 1]
 with random state [5, 4, 6, 1]
 with random state [5, 4, 6, 1]
 with random state [5, 4, 6, 1]
 with random state [5, 4, 6, 1]
```

```
a=[1,2,3,4,5,6]    # 4 values for training and 2 for testing
b=[1,0,1,5,6,3]

for i in range(5):
    a_train,a_test,b_train,b_test=train_test_split(a,b,test_size=0.3)
```

```
print("without random state",a_train)
```

```
without random state [3, 6, 4, 1]
without random state [2, 5, 1, 3]
without random state [4, 3, 1, 6]
without random state [4, 3, 6, 1]
without random state [3, 2, 1, 6]
```

i. Feature Scaling

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
```

```
x_train
```

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	A14	A16	...	E8	F E69	F G63	F G73	F2	F33	F3
857	Daly, Mr. Peter Denis	1	51.0	0	0	113055	26.5500	S	0	0	...	0	0	0	0	0	0	
52	Harper, Mrs. Henry Sleeper (Myna Haxtun)	0	49.0	1	0	PC 17572	76.7292	C	0	0	...	0	0	0	0	0	0	
386	Goodwin, Master. Sidney Leonard	1	1.0	5	2	CA 2144	46.9000	S	0	0	...	0	0	0	0	0	0	
124	White, Mr. Percival Wayland	1	54.0	0	1	35281	77.2875	S	0	0	...	0	0	0	0	0	0	
578	Caram, Mrs. Joseph (Maria Elias)	0	NaN	1	0	2689	14.4583	C	0	0	...	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
835	Compton, Miss. Sara Rebecca	0	39.0	1	1	PC 17756	83.1583	C	0	0	...	0	0	0	0	0	0	
192	Andersen-Jensen, Miss. Carla Christine Nielsine	0	19.0	1	0	350046	7.8542	S	0	0	...	0	0	0	0	0	0	
629	O'Connell, Mr. Patrick D	1	NaN	0	0	334912	7.7333	Q	0	0	...	0	0	0	0	0	0	
559	de Messemaeker, Mrs. Guillaume Joseph (Emma)	0	36.0	1	0	345572	17.4000	S	0	0	...	0	0	0	0	0	0	
684	Brown, Mr. Thomas William Solomon	1	60.0	1	1	29750	39.0000	S	0	0	...	0	0	0	0	0	0	

623 rows × 154 columns

```
y_train
```

857
52
386
124
578
...
835
192
629
559
684

623 rows × 0 columns

✓ 0s completed at 12:31 PM