

```
##KV.VARSHA
##21BCB7074
##varsha.21bcb7074@vitapstudent.ac.in

In [ ]: Data Preprocessing
        o Import the Libraries.
        o Importing the dataset.
        o Checking for Null Values.
        o Data Visualization.
        o Outlier Detection
        o Splitting Dependent and Independent variables
        o Perform Encoding
        o Feature Scaling.
        o Splitting Data into Train and Test

In [1]: ### Importing the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

In [3]: ### Importing the dataset
dataset=pd.read_csv("Titanic-Dataset.csv")

In [4]: dataset

Out[4]:
   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  Heikinen, 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
887         888         1       1  Graham, Miss. Margaret Edith    female  19.0    0    0    112053  30.0000   B42      S
888         889         0       3  Johnston, Miss. Catherine Helen "Carrie" female  NaN     1    2  W./C. 6607   23.4500   NaN      S
889         890         1       1  Behr, Mr. Karl Howell      male  26.0    0    0    111369  30.0000  C148      C
890         891         0       3  Dooley, Mr. Patrick      male  32.0    0    0    370376   7.7500   NaN      Q

891 rows x 12 columns

In [5]: dataset.tail()

Out[5]:
   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 Helen "Carrie" female  NaN     1    2  W./C. 6607   23.45   NaN      S
889         890         1       1  Behr, Mr. Karl Howell      male  26.0    0    0    111369  30.00  C148      C
890         891         0       3  Dooley, Mr. Patrick      male  32.0    0    0    370376   7.75   NaN      Q

In [6]: dataset.shape

Out[6]: (891, 12)

In [7]: 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                 284 non-null    object
11   Embarked              889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB

In [8]: dataset.describe()

Out[8]:
   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    0.383638    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    6.000000   512.329200

In [9]: corr=dataset.corr(numeric_only=True)

In [10]: corr

Out[10]:
   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

In [11]: plt.subplots(figsize=(20,15))
sns.heatmap(corr,annot=True)

Out[11]:
<Axes: >


```