NAME : Mansi Dwivedi
BRANCH : IT
UID : 2019140016
BATCH : B
COURSE : Data Analytics Lab
EXPERIMENT : 1

AIM : To perform EDA such as number of data samples, number of features, number of classes, number of data samples per class, removing missing values, conversion to numbers, using the seaborn library to plot different graphs.

PROBLEM STATEMENT : Choose any one of the given datasets, apply EDA on it and write a detailed inference.

CODE & OUTPUT:
https://colab.research.google.com/drive/1724MzZ0VEYuVHBOlltw8mQ4nIh0OecLA?usp=sharing

CONCLUSION :
After learning about these exploratory data analysis techniques and also implementing them, I was able to conclude that :

- In the process of EDA, we first start by removing the redundant variables, that is, the columns that we know will not contribute to the model building and learning.
- The next step can comprise selecting the variable, which means studying every column and the kind of data that they possess, and also looking for ways to remove the null values.
- Now we can study the distribution of the features like mean, min, and max. That will help us determine whether the data contains outliers or not, and then eventually look for ways to remove them. In my case, I studied the value distribution using the value_counts() method, which showed me the datapoints after which the frequency becomes comparatively negligible and hence can be termed outlying.
- Next I removed the rows that had a frequency less than a given amount because that would lead to a lot of unnecessary data, hence leading to model confusion while learning.

- Now I encoded the categorical features. For which I studied various types of encoding, like one hot encoding, label encoding, hashing encoding, binary encoding, target encoding, etc. For two of the features in my dataset, I went ahead with binary encoding since the data only had two categories.
- Finally after data cleaning and preprocessing, I analyzed the relationships between different variables and plotted various graphs and visualizations to understand the correlation between them. I used the seaborn library to plot graphs like heatmaps (it provides a coloured distribution that signifies correlation between all the numerical data in the dataset). The observation has been mentioned in the colab notebook itself. After that, I plotted scatter plots, which showed a precise relationship between any two variables.