FUNDAMENTALS OF DATA SCIENCE

Coding assignment

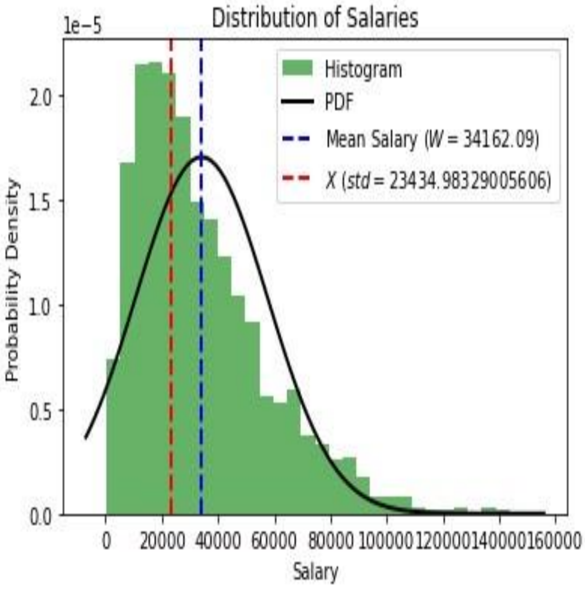
Name: Panneru Rani

Student ID: 22077318

Github link:

DATA SYNOPSIS:

The "data8.csv" dataset contains annual salary data from various European countries, comprising 4000 entries. Mean and standard deviation calculations have been conducted on this statistically representative dataset, offering insights into the average salary and variability in annual incomes across European countries.



Description of the Distribution:

The distribution of salaries is visualized using a probability density function (PDF) and a histogram. The PDF is modelled using the normal distribution (Gaussiandistribution), and the histogram provides a visual representation of the frequency of different salary ranges.

Calculation of the Mean (W):

The mean of the salary distribution, denoted as W, is calculated using the formula for the arithmetic mean:

Where N is the number of data points (salaries) and xi represents each individual salary.

W= N1 ∑i=1Nxi

Calculation of Required Value X:

The required value X is chosen as a percentile of the salary distribution, denoted as Xp. In this case, X is is calculated as the p-th percentile (e.g., 75th percentile). The formula for calculating the p-th percentile is: Xp=percentile(data,p).

This formula uses the percentile function, and in the provided code, p is set to 75 to obtain the 75th percentile.

Value Obtained:

The calculated mean (W=34162.09) and the chosen percentile (X=46747.5) are then plotted on the graph. These values are rounded to at least two significant digits. The mean represents the central tendency of the salary distribution, while X indicates a specific data point that corresponds to the chosen percentile.

CONCLUSION:

This approach allows for a concise analysis of the central tendency and a specific point in the salary distribution, providing insights into the overall distribution of salaries in the given dataset.