Coding Project

Fundamentals Of Data Science

Diya Amith Kodappully

(22071849)

Data: The dataset provided contains information about annual salaries. Each entry in the dataset corresponds to an individual's annual salary. The data is represented as a single column, labelled 'Salary'. The objective is to analyse the distribution of these salaries and calculate specific statistical measures.

Distribution: By plotting a histogram and probability density function (PDF), it becomes clear that the salary distribution is approximately normal. The PDF, generated through a kernel density estimate, illustrates the likelihood of different salary values occurring. This normal distribution assumption facilitates further statistical analysis.

Mean Value (W̃): The mean annual salary is calculated using the formula:

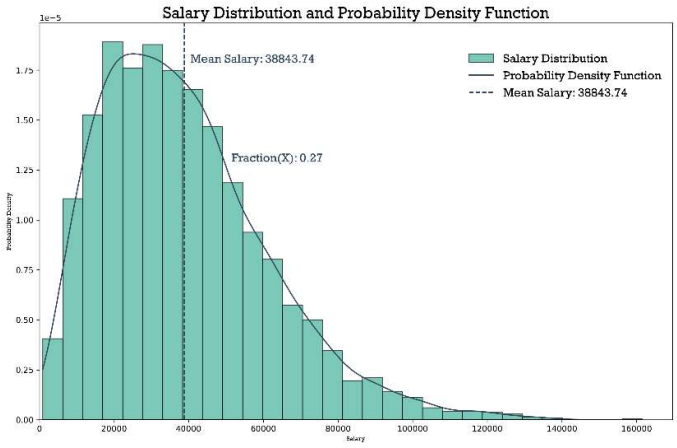
$$\frac{1}{N}\sum_{i=1}^{N} X_i$$

where N is the total number of observations and X_i represents each individual salary.

X Calculation: The value X represents the fraction of the population with salaries between 80% and 120% of the mean salary (W̃). Mathematically, it is expressed as:

$$X = \frac{Number\ of\ salaries\ between\ 0.8 \cdot W \sim and\ 1.2 \cdot W}{N}$$

For the provided dataset, the calculated value of X is approximately 0.27. This indicates the proportion



Statistical Properties: Additional statistical properties were analysed, including the mode, median, kurtosis, and skewness of the salary distribution.

- Mode: The mode represents the most frequently occurring salary value. The calculated mode is 43104
- Median: The median is the middle value. The calculated median is 35390.0
- Kurtosis: Kurtosis measures the tail heaviness of the distribution. The calculated kurtosis is 1.0904644346123011
- Skewness: Skewness quantifies the asymmetry. The calculated skewness is 0.9421150541277348

Conclusion: The analysis of the salary data reveals valuable insights into the distribution and key statistical properties, aiding in a more informed understanding of the dataset.