TASK:

Describe a scenario where the mean and median values of a dataset may differ significantly. What insights can be gained by examining this difference? Can you propose any solutions to handle this discrepancy? Additionally, how can Python be used to implement these solutions and what are some real-world examples of this phenomenon?

The mean and median values of a data set can differ significantly due to skewed data distribution which refers to a type of distribution in which the values or frequencies of data points are not evenly distributed around the mean. In a skewed distribution, the outliers in the tail pull the mean away from the center towards the longer tail. Skewed distributions are common in many real-world datasets. For example, the income distribution often has a right skew because there are relatively few individuals or households with extremely high incomes, which pushes the average upwards, while most people earn moderate incomes.

Examining the difference between the mean and the median can provide insight into the distribution of the data set and the presence of outliers. If the mean is much higher or lower than the median, this indicates that the data set is skewed. Understanding this difference is crucial to making accurate data-driven interpretations and decisions.

Identifying and Removing Outliers: Outliers can be identified using statistical methods such as z-scores or by visualizing the data using box plots. Once identified, outliers can be removed or adjusted accordingly.

Remove a certain percentage of extreme values from the dataset before analysis to reduce the influence of outliers on the distribution.

Transforming the Data: Transformations such as taking the logarithm or square root of the data can sometimes mitigate the effect of outliers and make the distribution more symmetric.

Using Robust Statistics: Robust statistical measures such as the trimmed mean or Winsorized mean are less sensitive to outliers and provide more accurate estimates of central tendency.

Datasets : https://www.kaggle.com/datasets/krishnaraj30/finance-loan-approval-prediction-data

References:

Goyal, C. (2021). Why you shouldn't just delete outliers. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/

Codebasic.(2020). Outlier detection and removal: z score, standard deviation | Feature engineering tutorial python. https://www.youtube.com/watch?v=KFuEAGR3HS4

Coursera. https://www.coursera.org/learn/foundations-of-data-science/lecture/2vmAG/welcome-to-module-2

Mirko, S.(2019). Python Statistics Fundamentals: How to Describe Your Data. https://realpython.com/python-statistics/

Clinfo. Stuck in the middle – mean vs. median. https://www.clinfo.eu/mean-median/

Jim, F. Statistics By Ji. Mean, Median, and Mode: Measures of Central Tendency. https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

Lumen. Introduction to Statistics. Skewness and the Mean, Median, and Mode. https://courses.lumenlearning.com/introstats1/chapter/skewness-and-the-mean-median-and-mode/