

Name: Vedant Bothikar Roll no 31115

TE-1 Batch: K-1

DSDBAL Assignment No: 3

Title: Descriptive Statistics: Measures of central tendency & variability

Problem statement: Perform the following operations on any open source dataset (e.g., data.csv)

1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris versicolor' of iris.csv dataset.

Provide the codes with outputs and explain everything that you do in this step.

Learning objectives: To understand the concept of statistical

operation

S/w & H/w requirements:

Ubuntu 20.04 64 bit

programming language: python 3.8

jupyter notebook

8gb ram intel i7 processor

Theory: We have used 2 datasets for this arrangement

dataset 1: nba.csv

the dataset consists of following columns

dataset2: iris.csv

following columns: id,

SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, Species

Measures of central tendency:

Mean: represents the arithmetic average of the data

Median: The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

*Mode: The mode is the most frequent score in our data set. On a histogram, it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option.*

*Measures of variability:*

*Range: The range of a dataset is the difference between the largest and smallest values in that dataset.*

*standard deviation: The standard deviation is a statistic that measures the*

*dispersion of a dataset relative to its mean and is calculated as the square root of the variance.*

*Conclusion: We successfully implemented assignment using statistical measures and understood the concept of descriptive statistics*