**PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE**

**ACADEMIC YEAR: 2021-22**

## **DEPARTMENT of COMPUTER ENGINEERING DEPARTMENT**

**CLASS: T.E. SEMESTER: I**

**SUBJECT: DSBDAL**

|  |  |
| --- | --- |
| **ASSINGMENT NO.** | 3 |
| **TITLE** | Descriptive Statistics - Measures of Central Tendency and variability |
| **PROBLEM STATEMENT /DEFINITION** | 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- verginica’ of iris.csv dataset.   Provide the codes with outputs and explain everything that you do in this step. |
| **OBJECTIVE** | 1. To provide summary statistics for a sample dataset 2. To identify some basic statistical details of specific columns of a dataset |
| **OUTCOME** | After completion of this assignment the students will be able to:   1. calculate mean, median, minimum, maximum, standard deviation for a dataset 2. identify some basic statistical details of specific columns of a dataset |
| **S/W PACKAGES AND**  **HARDWARE APPARATUS USED** | Jupyter notebook, Pandas libraries, Windows/ Linux Operating System, I5 machines/ Laptops |
| **REFERENCES** | [www.kaggle.com](http://www.kaggle.com) |
| **STEPS** | PartI: This is one sample set of steps, students may try any other wayOpen Jupyter notebook fileImport NumpyImport PandasRead “nba.csv” sample dataset in pandas data frame with the help of read\_csv function, dataset available in the following link<https://drive.google.com/file/d/1kH6mN8y0Eo6u7zlYvcEqP9uIaIPFdYMQ/view?usp=sharing>Use shape, head, dtypes, isnull to get familiar with the datasetCalculate mode using mode() function on a sample column, eg. collegeCalculate mean using mean() function on a sample column, eg. SalaryUse value\_counts on sample column, eg. AgeUse groupby on a sample column, eg. Age and SalaryPartI: This is one sample set of steps, students may try any other wayOpen Jupyter notebook fileImport NumpyImport PandasRead “Iris.csv” sample dataset in pandas data frame with the help of read\_csv function, dataset available in the following link<https://drive.google.com/file/d/1U8ks4kGRSMKSFkU1lwpf1SncwXaX29IZ/view?usp=sharing>Use shape, head, dtypes, isnull to get familiar with the datasetUse value\_counts on Species column to get count of ‘Iris-setosa’, ‘Iris-versicolor’ and ‘Iris- verginica’ of iris.csv datasetWrite a function to check for each species and use describe().transpose() function to display some basic statistical details like percentile, mean, standard deviation etc. of all species, sepal length, sepal width, petal length and petal width |
| **INSTRUCTIONS FOR**  **WRITING JOURNAL** | 1. Date  2. Assignment no.  3. Problem definition  4. Learning objective  5. Learning Outcome  6. Concepts related Theory  7. Algorithm  8. Test cases  10. Conclusion/Analysis |

**Prerequisites:** Basic knowledge of DBMS

**Concepts related Theory:**

**Descriptive statistics:** It summarizes and organizes characteristics of a data set.

The [**central tendency**](https://www.scribbr.com/statistics/central-tendency/)concerns the averages of the values.

The [**variability**](https://www.scribbr.com/statistics/variability/)or dispersion concerns how spread out the values are.

**Measures of central tendency**

It estimates the center, or average, of a data set. The mean, median, mode are 3 ways of finding the average.

Mean: The mean, or *M*, is the most commonly used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called *N*.

**Example:**

Mean number of library visits

Data set: 15, 3, 12, 0, 24, 3

Sum of all values: 15 + 3 + 12 + 0 + 24 + 3 = 57

Total number of responses N = 6

Mean: Divide the sum of values by N to find M: 57/6 = **9.5**

**Median:** The median is the value that’s exactly in the middle of a data set.

To find the median, order each response value from the smallest to the biggest. Then, the median is the number in the middle. If there are two numbers in the middle, find their mean.

**Example:**

Median number of library visits

Ordered data set: 0, 3, 3, 12, 15, 24

Middle numbers: 3, 12

Median: Find the mean of the two middle numbers: (3 + 12)/2 = **7.5**

Mode: The [mode](https://www.scribbr.com/statistics/mode/) is the simply the most popular or most frequent response value. A data set can have no mode, one mode, or more than one mode.

To find the mode, order your data set from lowest to highest and find the response that occurs most frequently.

**Example:**

|  |  |
| --- | --- |
| **Mode number of library visits** | |
| **Ordered data set** | : 0, 3, 3, 12, 15, 24 |
| **Mode** | Find the most frequently occurring response: **3** |

**Measures of variability**

It gives you a sense of how spread out the response values are. The range, standard deviation and variance each reflect different aspects of spread.

**Range**

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.

**Example:**

Range of visits to the library in the past year

**Ordered data set:** 0, 3, 3, 12, 15, 24

**Range:** 24 – 0 = **24**

**Standard deviation**

It is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by *N* – 1.
6. Find the square root of the number you found.

**Example:**

Standard deviations of visits to the library in the past year

In the table below, you complete **Steps 1 through 4**.

Raw data Deviation from mean Squared deviation

15 15 – 9.5 = 5.5 30.25

3 3 – 9.5 = -6.5 42.25

12 12 – 9.5 = 2.5 6.25

0 0 – 9.5 = -9.5 90.25

24 24 – 9.5 = 14.5 210.25

3 3 – 9.5 = -6.5 42.25

M = 9.5 Sum = 0 Sum of squares = 421.5

**Step 5:** 421.5/5 = 84.3

**Step 6:** √84.3 = 9.18

From learning that **s =9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

**Conclusion:** Students learned descriptive Statistics - Measures of Central Tendency and variability by performing various operations on given dataset using Pandas library.

**Review Questions**:

1. What is mean?
2. What is mode?
3. What is median?
4. What do you mean by central tendency?
5. What do you mean by variability?
6. Explain groupby function
7. Explain transpose function
8. Explain value\_count function
9. What are measures of central tendency?
10. What are measures of variability?