Assignment No.3

Name - Kudharan Sumit Dattatraya
Class - TE
Div - 4
Subject - DSBDAL

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 values 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' & 'Iris-versicolor' of iris.csv dataset.

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

Theory -

1) Explain need of statistics in data science.

→ When the data is big & unorganised statistics plays a powerful role in that situation. When a company uses statistics to find insights, it makes the tedicious task look minimalist & easy in front of the big & buffer information that was Provided earlier.

Some ways in which statistics helps in Data Science are

1) Prediction & classification.

2) Help to create probability distribution & estimation.

3) Pattern detection & grouping

4) Powerful Insights.

5) Segmentation & optimization

2) Explain measure of central tendancy with example

→ This measure is an important way to summarize the dataset with one representative value. This measure provides a rough picture of where data points are centered.

The commonly used measures of central tendancy are:

- Mean - "Average" value is termed as the mean of the dataset.
- Median - The middle value of the sorted dataset.
- Mode - The most frequently occuring value in the dataset.

Example - Consider the weight (in kg) of 5 children as 36, 40, 32, 42, 30. Let's compute mean, median & mode

→ Mean = $(36+40+32+42+30)/5 = 180/5 = 36$ kg.

Meadion = $30, 32, \boxed{36}\; 40, 42 = 36$ kg

Mode = 36 kg occurs most number of times so mode is 36kg

3) Explain measures of dispersion with example.

→ Measures of dispersion measures the scatter of the data that is how far values in the distribution are. Dispersion is the measure of the extent to which the points of the distribution differ from the average of distribution.

Types of measures dispersion -

① Absolute Measures of dispersion

② Relative Measures of dispersion.

① Absolute Measures of dispersion -

Types of absolute measures of dispersion -

① Range -

Range is the measure of the difference between the largest & smallest value of the data variability.

Example - 1, 2, 3, 4, 5, 6, 7

Range = Highest value - Lowest value = 7-1 = 6

② Mean (μ) -

Mean is calculated as the averoge of the numbers.

Example - 1, 2, 3, 4, 5, 6, 7, 8

Mean (μ) = (sum of all the terms / total No. of terms)

= (1+2+3+4+5+6+7+8) / 8 = 36/8 = 4.5

i) Variance (σ^2) -

Variance can be calculated by obtaining the sum of the squared distance of each term in the distribution from the mean, & then dividing this by the total number of terms in the distribution.

Formula - $(\sigma^2) = \Sigma(x-\mu)^2/N$

ii) standard Deviation -

It can be represented as the square root of variance

Formula - $\sqrt{\sigma}$

iii) Quartile -

Quartile divide the list of numbers of data into quartos.

iv) quartile Deviation -

quartile deviation is the measure of the difference between upper & lower quartile.

Formula - $Q_3 - Q_1$

③ Mean Deviation -

It is also known as averoge deviation

Formula -

Mean Deviation using Mean = $\Sigma|X-M|/N$

Mean Deviation using Median = $\Sigma|x-x_1|/N$

② Relative measures of dispersion -

Relative Measure of dispersion in statistics are the values without units. A relative measure of dispersion is used to compare the distribution of two or more datasets.

Types of relative measure dispersion:

1) Coefficient of Range -

It is calcuated as the ratio of the difference between the largest & smallest terms of the distribution to the sum of the largest & smallest terms of the distribution.

formula - $L-S/L+S$

2) Co-efficient of variation -

The co-efficient of variation is used to compare the 2 data with respect to homogeneity or consistency.

Formula - $C.V = (\sigma/x)100$

3) Co-efficient of standard deviation -

It is the ratio of standard deviation with the mean of the distribution of terms.

Formula - $\sigma = (\sqrt{(x-\bar{x}_i))}/(N-1))$

4) Co-efficient of Quartile Deviation -

It is the ratio of the difference between the upper quartile & the lower quartile to sum of the upper quartile & lower quartile.

formula - $(Q_3 - Q_3)/(Q_3 + Q_1)$

5) Co-efficient of mean deviation -

It can be computed using the mean or median of the data

Mean deviation using Mean - $\Sigma|x-m|/N$

Mean deviation using Mean - $\Sigma|x-x_1|/N$

4) What is mean, mode, median and standard deviation. Solve example.

→ ① Mean -

Mean is calculated as the average of numbers.

② Mode -

Mode is the value that occurs most frequently.

③ Median -

Median is the middle number in an ordered dataset.

④ Standard Deviation -

It is nothing but the square root of variance.

Example -

Sample Dataset - 154, 139, 154, 192, 180, 140, 154, 155, 192.

→ $$\text{Mean} = \frac{\text{sum of all terms}}{\text{No. of terms}}$$

$$= \frac{154+139+154+192+180+140+154+155+192}{9}$$

$$\boxed{\text{Mean} = \frac{1460}{9} = 162.2}$$

Mode =

In the dataset 154 occured 3 times which is maximum.

$\boxed{\therefore \text{ Mode} = 154}$

Median =

To find median first sort the dataset as ordered.

139	140	154	154	154	155	180	192	192

Now we have formula to find middle position

$$\text{Position} = \frac{n+1}{2} = \frac{9+1}{2} = \frac{10}{2} = 5$$

139	140	154	154	154	155	180	192	192
				↑ 5				

$\boxed{\text{Median} = 154}$

Standard Deviation= $\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$ Here, $\bar{x}$ = mean = 162.2

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
154	-8.2	67.24
139	-23.2	538.24
154	-8.2	67.24
192	29.8	888.04
180	17.8	316.84
140	-22.2	492.84
154	-8.2	67.24
155	-7.2	51.84
192	29.8	888.04
	Sum =	3377.56

$$= \sqrt{\frac{3377.56}{9-1}}$$

$$= \sqrt{\frac{3377.56}{8}}$$

$$= \sqrt{422.195}$$

$$= 20.55$$

$$\boxed{\text{Standard Deviation} = 20.55}$$

5) Explain dataset describe() method.

→ The describe() method is used for calcuelating Some statiscal data like percentile, mean and std of the numerical values of the series or Dataframe.

Syntax -

Dataframe.describe(percentiles = None, include=None, exclude=None)

Parameters -

percentile - It is an optional parameter which is a list like data types of numbers that should fall between 0 & 1. Its default value is [.25, .5, .75].

include- It is also an optional parameter that includes the list of the data types while describing the Dataframe. Its default value is None.

exclude - It is also an optional porameter that exclude the list of data type while describing Datafrome. Its default value is None.

Example -

```
import pandas as pd
a1 = pd.Series ([1,2,3])
a1.describe()
```

Output -

```
count     3.0
mean      2.0
std       1.0
min       1.0
25%       1.5
50%       2.0
75%       2.5
max       3.0
dtype     float64
```

6) Variance & what are the steps to calcuate Variance

→ The term variance refers to a ~~Stastical~~ statistical measurement of the spread between numbers in a data set.

Steps for calculating Variance -

Dataset	46	69	32	60	52	41

Step1 find the mean -

$$\text{mean}(\bar{x}) = \frac{46+69+32+60+52+41}{6} = 50$$

step 2 - find each data value's deviation from the mean.

Data	Deviation from the mean
46	$46 - 50 = -4$
69	$69 - 50 = 19$
32	$32 - 50 = -18$
60	$60 - 50 = 10$
52	$52 - 50 = 2$
41	$41 - 50 = -9$

Step 3 - Square each deviation from the mean

Squared deviation from the mean

$(-4)^2 = 16$

$(19)^2 = 381$

$(-18)^2 = 324$

$(10)^2 = 100$

$(2)^2 = 4$

$(-9)^2 = 81$

Step 4 - find sum of squares

$16 + 381 + 324 + 100 + 4 + 81 = 886$

step 5 - Divide the sum of squares by $n-1$

$$\text{Variance} = \frac{886}{6-1}$$

$$= \frac{886}{5}$$

$$\boxed{\text{Variance} = 177.2}$$