Page No.

Date

Name :- Rohini Janardan Devkar

Roll no :- 23272

PRN no :- 72030818G

TE-2

DSBDA Lab.

Practical No. 3(1)
Measures of Central Tendencies

Aim:-

Perform the 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 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.

Theory:-

*) Measures of Central Tendency :-

A measure of central tendency is a summary statistic that represents the centre point or typical value of dataset. In statistics the three most common measures of central tendancy are the mean, median and mode.

① Mean :- The mean is the arithmetic average and it is probably the measure of central tendency.

$$\text{Mean} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

② Median :- The median is the middle value. It is the value that splits the dataset in half. To find the median, order your data from smallest to largest and then find the data point that has an equal amount of values above it and below it.

$$\text{Median} = \left\{\frac{(n+1)}{2}\right\}^{th} \text{term}$$ ... when calculation is odd

and

$$\text{Median} = \frac{\left\{\frac{n}{2}\right\}^{th} \text{term} + \left\{\frac{n}{2}+1\right\}^{th} \text{term}}{2}$$

...when the calculation is even.

③ Mode :- - Mode is the value category that occurs most often within the dataset.

- If the data have multiple values that are tied for occuring the most frequently, you have a multimodal distribution.

$$\text{Mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] h$$

l = lesser limit of modal class

f_m = frequency possessed by the modal class

f_1 = frequency possessed by the class before the modal class.

f_2 = frequency possessed by the class after the modal class.

h = width of the class.

④ Standard Deviation :- Standard deviation is a number that describes how spread out the observations are standard deviation is a measure of uncertainty.

$$S.D. = \sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

N = Total number of frequency.

Data Science And Big Data Analytics Practical 3(1)

==

Name:- Rohini Janardan Devkar

PRN no:- 72030818G

Roll no:- 23272

Class :- TE2(COMP)

==

Problem Statement:-

Perform the following operations on any open source dataset (eg. 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.

==

In [19]:

```
import pandas as pd
df = pd.read_csv ('wages.csv')
```

In [20]:

```
df.describe()
```

Out[20]:

	earn	height	ed	age
count	1379.000000	1379.000000	1379.000000	1379.000000
mean	32446.292622	66.592640	13.354605	45.328499
std	31257.070006	3.818108	2.438741	15.789715
min	-98.580489	57.340000	3.000000	22.000000
25%	10538.790721	63.720000	12.000000	33.000000
50%	26877.870178	66.050000	13.000000	42.000000
75%	44506.215336	69.315000	15.000000	55.000000
max	317949.127955	77.210000	18.000000	95.000000

In [21]:

```
df.shape
```

```
Out[21]: (1379, 6)
```

```
In [22]: df.size
```

```
Out[22]: 8274
```

Measures of central tendency

```
In [24]: # Min
         df.min()
```

```
Out[24]: earn      -98.580489
         height         57.34
         sex           female
         race           black
         ed                 3
         age               22
         dtype: object
```

```
In [25]: # Max
         df.max()
```

```
Out[25]: earn      317949.127955
         height            77.21
         sex                male
         race              white
         ed                   18
         age                  95
         dtype: object
```

```
In [26]: # Mode
         df['age'].mode()
```

```
Out[26]: 0    36
         dtype: int64
```

```
In [27]: # Mean
         df['age'].mean()
```

```
Out[27]: 45.328498912255256
```

```
In [29]: # Median
         df['age'].median()
```

```
Out[29]: 42.0
```

```
In [30]: # Std Deviation
         round(df['age'].std(),2)
```

```
Out[30]: 15.79
```

Summary Statistics

In [32]:

```
# Summary statistics for all numerical columns
round(df.describe(),2)
```

Out[32]:

```
            earn   height       ed      age
count    1379.00  1379.00  1379.00  1379.00
 mean   32446.29    66.59    13.35    45.33
  std   31257.07     3.82     2.44    15.79
  min     -98.58    57.34     3.00    22.00
  25%   10538.79    63.72    12.00    33.00
  50%   26877.87    66.05    13.00    42.00
  75%   44506.22    69.32    15.00    55.00
  max  317949.13    77.21    18.00    95.00
```

In [33]:

```
# Summary statistics by groups
df['age'].groupby(df['ed']).describe()
```

Out[33]:

```
    count       mean        std   min    25%   50%    75%   max
ed
 3    1.0  68.000000        NaN  68.0  68.00  68.0  68.00  68.0
 4    2.0  67.000000   1.414214  66.0  66.50  67.0  67.50  68.0
 5    5.0  69.800000  13.367872  55.0  62.00  66.0  77.00  89.0
 6    7.0  67.571429  11.443443  44.0  66.00  71.0  73.50  79.0
 7    3.0  53.000000   8.000000  45.0  49.00  53.0  57.00  61.0
 8   28.0  57.071429  17.090020  24.0  46.00  60.0  68.50  87.0
 9   23.0  53.782609  17.929043  25.0  46.00  55.0  65.50  88.0
10   37.0  52.459459  21.595517  22.0  35.00  52.0  72.00  91.0
11   39.0  46.333333  15.863452  22.0  34.00  42.0  56.50  76.0
12  520.0  44.644231  15.938900  22.0  32.00  41.0  55.00  95.0
13  119.0  44.403361  16.024515  22.0  32.00  41.0  52.00  87.0
14  192.0  44.197917  15.370550  22.0  32.00  41.0  54.25  85.0
15   66.0  43.515152  16.020412  24.0  30.00  42.0  51.75  87.0
16  187.0  42.229947  13.778804  23.0  33.00  38.0  47.00  95.0
17   70.0  43.571429  11.917070  27.0  34.25  43.0  47.00  83.0
18   80.0  49.225000  12.323668  29.0  40.75  48.0  56.25  86.0
```

* Conclusion:-

Thus, I have studied and perform the basic statistical measures of mean, median and standard deviation.