Group A Assignment 3

Title:
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.

Objective:

1. If your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped bythe age groups.
2. Create a list that contains a numeric value for each response to the categorical variable.
3. 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.

Theory:
What is Statistics?

✓ Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and visualizing empirical data.

✓ Descriptive statistics and inferential statistics are the two major areas of statistics.

✓ Descriptive statistics are for describing the properties of sample and population data (what has happened).

✓ Inferential statistics use those properties to test hypotheses, reach conclusions, and make predictions (what can you expect).

Use of Statistics in Data Science

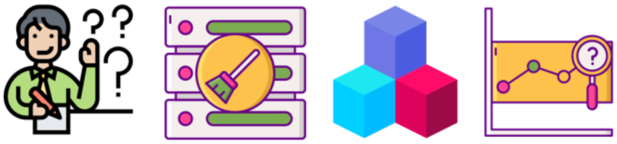
✓ Asking questions about the data

✓ Cleaning and preprocessing the data

✓ Selecting the right features

✓ Model evaluation

✓ Model prediction



Mean

✓ The arithmetic mean of a given data is the sum of all observations divided by the number of observations.

✓ For example, a cricketer's scores in five ODI matches are as follows: 12, 34, 45, 50, 24. To find his average score in a match, we calculate the arithmetic mean of data using the mean formula:

$$\text{Mean} = \frac{Sum\ of\ terms}{Number\ of\ terms}$$

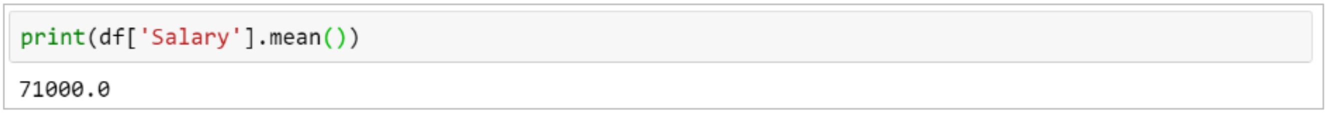
✓ Mean = Sum of all observations/Number of observations

Mean = (12 + 34 + 45 + 50 + 24)/5

Mean = 165/5 = 33

Mean is denoted by x̄ (pronounced as x bar).

To find the mean or the average salary of the employees, you can use the mean() functions in Python.



```
print(df['Salary'].mean())
```

```
71000.0
```

Mode

✓ The Mode refers to the most frequently occurring value in your data.

✓ You find the frequency of occurrence of each number and the number with the highest frequency is your mode. If there are no recurring numbers, then there is no mode in the data.

✓ Using the mode, you can find the most commonly occurring point in your data. This is helpful when you have to find the central tendency of categorical values, like the flavor of the most popular chip sold by a brand. You cannot find the average based on the orders; instead, you choose the chip flavor with the highest orders.

✓ Usually, you can count the most frequently occurring values and get your mean. But this only works when the values are discrete. Now, again take the example of class marks.

✓ Example: Take the following marks of students :

Marks = 35, 40, 45, 49, 34, 47, 39, 25, 19, 35, 28, 48

Over here, the value 35 occurs the most frequently and hence is the mode.

✓ But what if the values are categorical? In that case, you must use the formula below:

$$\text{Mode} = l + \left(\frac{f1 - f0}{2f1 - f0 - f2}\right) \times h$$

Where,

l = lower limit of modal class

h = lower limit of preceding modal class

f1 = frequency of modal class

f0 = frequency of class preceding modal class

f2 = frequency of class succeeding modal class

The modal class is simply the class with the highest frequency. Consider the range of frequencies given for the marks obtained by students in a class:

Marks	10-20	20-30	30-40	40-50
Number of Students	1	3	5	4

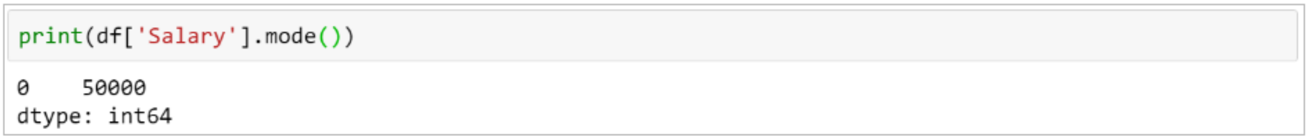
In this case, you can see that class 30-40 has the highest frequency, hence it is the modal class. The remaining values are as follows: l = 30, h = 20, f1 = 5, f0 = 3, f2 = 4

In that case, the mode becomes :

$$\text{Mode} = 30 + \left(\frac{5-3}{2*5-3-4}\right) \times 20$$

$$= 43.33$$

Hence, the mark which occurs most frequently is 43.33. The mode of salary from the salary data frame can be calculated as:



```
print(df['Salary'].mode())
```

```
0    50000
dtype: int64
```

Median

- ✓ Median refers to the middle value of a data. To find the median, you first sort the data in either ascending or descending order or then find the numerical value present in the middle of your data.
- ✓ It can be used to figure out the point around which the data is centered. It divides the data into two halves and has the same number of data points above and below.
- ✓ The median is especially useful when the data is skewed data. That is, it has high data distribution towards one side. In this case, the average wouldn't give you a fair mid-value but would lean more towards the higher values. In this case, you can use the middle data point as the central point instead.

- ✓ Consider n terms X_1, X_2, X_3,………… X_n. The basic formula for the median is by dividing the total number of observations by 2. This works fine when you have an odd number of terms because you will have one middle term and the same number of terms above and below. For an even number of terms, consider the two middle terms and find their average.

$$\text{Median} = \frac{n+1}{2} \text{ th term , n = odd}$$

$$\{ \frac{n}{2} \text{ th term} + \frac{n}{2} + 1 \text{ th term} \} / 2 \text{ , n = even}$$

Example: Consider following are students marks

Marks = 35, 40, 45, 49, 34, 47, 39, 25, 19, 35, 28, 48

To find the middle term, you first have to sort the data or arrange the data in ascending or descending order. This ensures that consecutive terms are next to each other.

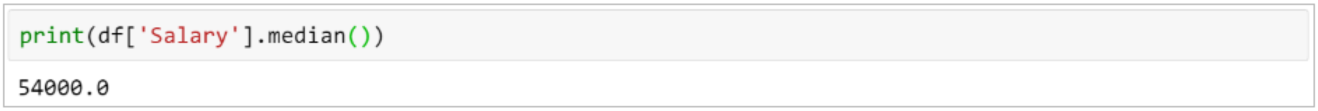
Sorted Marks = 19, 25, 28, 30, 34, 35, 39, 40, 45, 47, 48, 49

You can see that we have 12 data points, so use the median formula for even numbers.

$$\text{Median} = \{(\frac{12}{2} \text{ th term}) + (\frac{12}{2} + 1 \text{ th term}) \} / 2$$
$$= \{ 6^{th} + 7^{th}) /2 = (35 + 39) /2$$
$$= 37$$

So, the middle term in the range of marks is 37. This means that the other marks lie in a frequency range of around 37.

The median() function in Python can help you find the median value of a column. From the salary data frame, you can find the median salary as:



```
print(df['Salary'].median())
```

```
54000.0
```

Standard Deviation and Variance

- ✓ Deviation just means how far from the normal
- ✓ The Standard Deviation is a measure of how spread out numbers are.
- ✓ Its symbol is σ (the greek letter sigma)

✓ The formula is easy: it is the square root of the Variance.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

✓ The Variance is defined as: The average of the squared differences from the Mean.

✓ To calculate the variance follow these steps:

1. Work out the Mean (the simple average of the numbers)
2. Then for each number: subtract the Mean and square the result (the squared difference).
3. Then work out the average of those squared differences

✓ Variance is used to measure the variability in the data from the mean.

To calculate the Variance, take each difference, square it, and then average the result:

Variance =

$$\begin{aligned}
\sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\
&= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} \\
&= \frac{108520}{5} \\
&= 21704
\end{aligned}$$

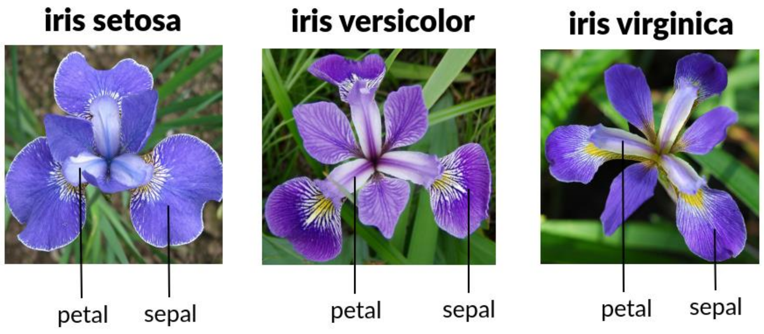
So the Variance is 21,704

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation =

$$\begin{aligned}
\sigma &= \sqrt{21704} \\
&= 147.32\ldots \\
&= \mathbf{147} \ \textit{(to the nearest mm)}
\end{aligned}$$

Finding statistical information of the iris flower dataset



Iris flower has three species - Setosa, Versicolor and Virginica

Load the dataset

url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv'

col_name = ['sepal-length','sepal-width','petal-length','petal-width','class']

df = pd.read_csv(url, names = col_name)

Display some information of the data

dataset.shape

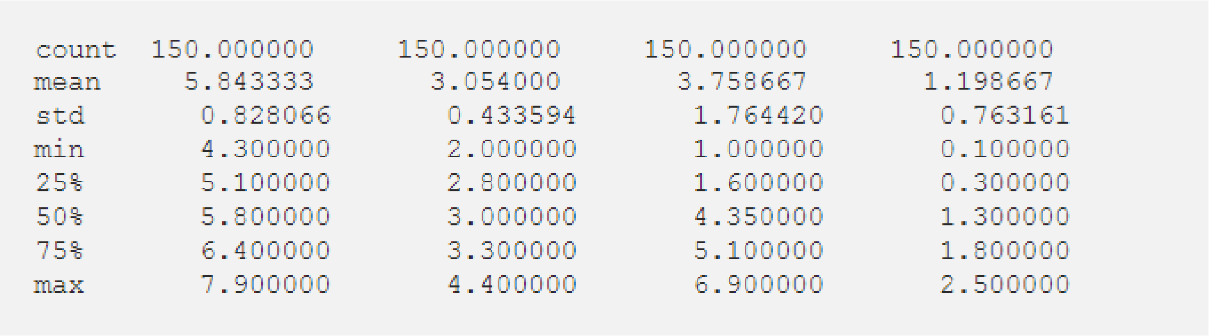
dataset.head()

dataset.info()

Display statistical information of the data

dataset.describe()

Output:



```
count  150.000000     150.000000     150.000000     150.000000
mean     5.843333       3.054000       3.758667       1.198667
std       0.828066       0.433594       1.764420       0.763161
min       4.300000       2.000000       1.000000       0.100000
25%       5.100000       2.800000       1.600000       0.300000
50%       5.800000       3.000000       4.350000       1.300000
75%       6.400000       3.300000       5.100000       1.800000
max       7.900000       4.400000       6.900000       2.500000
```

groupby()

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Example:

```
df = pd.DataFrame({'Animal': ['Falcon', 'Falcon',
...                           'Parrot', 'Parrot'],
...                'Max Speed': [380., 370., 24., 26.]})
>>> df
   Animal  Max Speed
0  Falcon      380.0
1  Falcon      370.0
2  Parrot       24.0
3  Parrot       26.0
>>> df.groupby(['Animal']).mean()
        Max Speed
Animal
Falcon      375.0
Parrot       25.0
```

Conclusion: Thus, we performed summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset with numeric variables grouped by the qualitative (categorical) variable.