General Information Regarding Data.

Week 1

Aniket Vaishnav

2017BTEIT00062

Introduction

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. Data sets can also consist of a collection of documents or files.

We perform various operations upon this group of data
Such as:

1: sum
2: variance
3: min
4: max
5: mode
6: median
7: mean
8: count
9: change attribute

Data Set chosen here is nifty50-stock-market-data

Terminologies

Sum

The Arithmetic addition of an attribute in a dataset is called sum of an attribute.

sum = $\sum_{i=0}^{n} x$

Variance

It means the spread between the data sets itself. Denoted by symbol δ^2.

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Min / Max

The Minimum / Maximum in attribute

Mode

The most frequently appearing data value in data set is known as Mode.

Median

Within a sorted data the mid point of separation which can divide data set in two halves is known as median.

Code

Written in python can be found at GitHub

```
import csv
import glob
import pandas

def selectattribute(csv_file):
    print('Select an attribute you choose')
    for i in range(csv_file.keys().__len__()):
        print(i,' :  ',csv_file.keys()[i])
    attr = int(input())
    attr = csv_file.keys()[attr]
    print('attr is ', attr)
    return attr

if __name__ == '__main__':
    print('Choose one of the CSV from below :')
    dataset_dir = 'dataset'
    items = []
    items_size = 1
    for item in glob.glob(dataset_dir+'/*.csv'):
```

```
    print(str(items_size).ljust(4)+': '+item)
    items_size += 1
    items.append(item)

csv_file = pandas.read_csv(items[int(input())-1])
attr = selectattribute(csv_file)
print(' main selected attr : ',attr)
while True:
    print('''
        1: sum
        2: variance
        3: min
        4: max
        5: mode
        6: median
        7: mean
        8: count
        9: change attribute
        0: Exit
    choose an option : ''', end='')
    ch = int(input())
    if ch==0:
        break
    elif ch==1:
        print(csv_file[attr].sum())
    elif ch==2:
        print(csv_file[attr].var())
    elif ch == 3:
        print(csv_file[attr].min())
    elif ch==4:
        print(csv_file[attr].max())
    elif ch==5:
        print(csv_file[attr].mode())
    elif ch==6:
        print(csv_file[attr].median())
    elif ch==7:
        print(csv_file[attr].mean())
    elif ch==8:
        print(csv_file[attr].count())
    elif ch==9:
        attr = selectattribute(csv_file)
print(''' *********** THANK YOU *********** ''')
```

Dependencies

Python 3+
Pandas

Run via `python3 main.py`

References

https://en.wikipedia.org/wiki/Data_set
https://www.kaggle.com/rohanrao/nifty50-stock-market-data