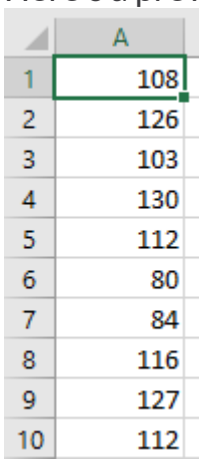
IQ Data - (Assignment 4 - Question 1)

- subject Machine Learning / AI

DESCRIPTION

The IQ data set containing 10000 data points is present at the location (**/data/training/iqdata.csv**)
The data set contains only the IQ values of people who participated in the survey across the world in a single column without header.

Here's a preview of the data under consideration:



	A
1	108
2	126
3	103
4	130
5	112
6	80
7	84
8	116
9	127
10	112

It contains IQ values with the below specifications:

- The average IQ is **around** 110
- There are a few super-intelligent people whose IQ is 192
- There are a few people with less IQ of 34
- The standard deviation is **around** 20
- The data points follow a **normal distribution**

Based on this data, create Python programs to perform the required analysis as described below:

1. Load the 10000 point data into a 1-D array. Then recalculate its **mean** & **standard deviation** to obtain their exact values. ***Print*** these two values.

- **Hint:** Use functions from **numpy** library
- **Note:** These two values are calculated for the entire data in all the cases

2. Calculate what percentage of people should have an IQ value between two values specified in the **/data/training/testcaseiq.txt**

- **Hint:** Since the data follows a normal distribution, use an appropriate function of **norm** from **scipy.stats** library
- Using this function, calculate the probability of an IQ score being smaller than the **upper** value specified in the **testcaseiq.txt**
- Similarly, calculate the probability of an IQ score being smaller than the **lower** value specified in the **testcaseiq.txt**
- Subtract the above two values to calculate the probability of IQ score falling between the **lower** and **upper** values
- Finally, print the result as percentage without the % sign
- Do this for all the testcases provided in **testcaseiq.txt**

3. A sample is drawn from this data is stored in different files such as **iqsample1.csv , iqsample2.csv** and so on. Read the name of the file (**<file_name>)** from **testcaseiq.txt** and then read the corresponding file from **/data/training/<file_name>.csv**

Consider a Null hypothesis that the mean of the sample is equal to the population mean of the above 10000-point data set. Test and decide whether the hypothesis can be accepted or rejected based on the p-value as:
- If p-value < 0.05, *print* as **"Reject"**
- Else *print* as **"Accept"**

Input Format:

- The first file to be read will be **iqdata.csv**, which contains the data as mentioned above. This file is in **.csv** format and is present at the location **(/data/training/iqdata.csv)**
- The second file to be read is **testcaseiq.txt** which is present at **(/data/training/testcaseiq.txt)**
- **testcaseiq.txt** has the following lines:
  - The first line contains the number of test cases T
  - From the second line, every set of three lines contain the lower value of the desired IQ range, the upper value of the desired IQ range and the name of the file containing samples to be used in the calculation of Null Hypothesis testing such as **iqsample1**
  - Then read the sample data from **(/data/training/iqsample1.csv)**

Output Format:

- For each test case **T**, create an output file, **output1.csv, output2.csv, ..., outputn.csv** where **n** represents the test case number

- **outputn.csv** should be present at the location (**/code/output/outputn.csv**) . This file should consist of the values for **Mean** and **Standard Deviation** on two separate rows, both values rounded to **2 decimal places.**

  Note: These two values are calculated for the entire data in all the cases

- The third line should contain the percentage value such as **34.567** of people with IQ in the specified range. The value should be rounded to **3 decimal places**
- The fourth line should contain the **result of the Null Hypothesis** in the format stated above
- **outputn.csv** should consist of the values on four **separate rows one below the other**

Sample Test Cases:

testcaseiq.txt contains the following data:

2
80
140
iqsample1
70
120
iqsample2

Sample Output:

Example: output1.csv will have data looking like this:

	A
1	110.12
2	19.23
3	34.567
4	Accept
5	
6	

DATASETS

- Training dataset