Auto MPG - (Test 2)

bookmark_border

subject Machine Learning / AI casino 30 points

DESCRIPTION

Consider the automobile data set located at **/data/training/autompg.csv**
This data set consists of specification and performance details of cars

Here's a preview of the data under consideration:

mpg	cylinders	displacement	horsepower	weight	acceleration	year	name
18	8	307	130	3504	12	70	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	buick skylark 320
18	8	318	150	3436	11	70	plymouth satellite
16	8	304	150	3433	12	70	amc rebel sst
17	8	302	140	3449	10.5	70	ford torino
15	8	429	198	4341	10	70	ford galaxie 500
14	8	454	220	4354	9	70	chevrolet impala

Questions:

We need to analyse this data using Python programs as follows:
Load the data from **/data/training/autompg.csv**. Then perform the operations as described below:

1. Calculate the following statistics for **mpg** column from the data set:
Hint: Use *pandas* dataframe functions

- Mean
- Median
- Mode
- Standard Deviation

Print these values in a file named **output.csv** on separate rows

2. Calculate the correlation of **weight** with **mpg** using two methods – **Pearson** and **Kendall**

- First use df.corr and set method as **kendall** and ***print*** the value of resulting coefficient
- Then calculate the value using pearson method using function pearsonr() and ***print*** the value of its coefficient

- ***Print*** these two values on two rows of **output.csv** below the existing rows written in Step 1

3. Read a sample of the above data set from a file. First, the name of the file containing the sample must be read from a file named **testcaseauto.txt**
Example: **autosample1** or **autosample2.** Then read the sample data from **autosample1.csv** or **autosample2.csv** respectively

- Compute the mean and standard deviation of **mpg** column of this sample.
- Calculate and ***print*** the difference between the means & standard deviations of the sample and population data as: **< value for population > – < value for sample >**
- ***Print*** these two values in a file named **outputn.csv** on two separate rows

Input Format:

- First, you have to read data from a file named **autompg.csv** present at the location **/data/training/autompg.csv**
- The second file to be read is a file named **testcaseauto.txt** which is present at the location **/data/training/testcaseauto.txt**
- **testcaseauto.txt** has the following lines:
    - The first line contains the number of test cases **T**
    - From the second line, every line contains the name of the file containing sample data to be used in the calculation of required statistics such as **autosample1**
    - Then read the sample data from **/data/training/autosample1.csv**

Output Format:

- You have to create a file named **output.csv** at the location **/code/output/output.csv**
- This file should contain the following values on 6 rows
    - **Mean -** value rounded to **2 decimal places** such as **45.67**
    - **Median -** value converted to int such as **35**
    - **Mode -** value in the form of a list such as **[12.1]**
    - **Standard Deviation -** value rounded to **2 decimal places** such as **5.43**
    - **Correlation coefficient using Kendall -** value rounded to **2 decimal places**such as **0.23**

    - **Correlation Coefficient using Pearson -** value rounded to **2 decimal places**such as **0.45**
- In Step 3, 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 difference in **Means** and **Standard Deviations** as described, on two separate rows
- Both values need to be rounded to **2 decimal places** and then printed

Sample Test Cases:

testcaseauto.txt contains the following data:

2
autosample1
autosample2

Sample Output

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

	A
1	12.34
2	35
3	[25.4]
4	5.67
5	-0.23
6	0.45
7	
8	

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

	A
1	5.43
2	2.05
3	
4	
5	

DATASETS

- Training dataset