BIG DATA PROGRAMMING ASSIGNMENT 2

Using the environmental data for each of the provinces in Canada and weighting each piece of data by the number of cities in the province, calculate the mean temperature and mean precipitation for all of Canada for annual and each month.

PROGRAM:

```
from pyspark import SparkConf

from pyspark import SparkContext

import pandas as pd

import xlrd

sc = SparkContext.getOrCreate();


months=['ANNUAL','JAN','FEB','MAR','APR','MAY','JUN','JUL','AUG','SEP','OCT','NOV','DEC']

data=pd.read_excel("data.xlsx",header=None,skiprows=1,
names=['Alberta','ANNUAL','JAN','FEB','MAR','APR','MAY','JUN','JUL','AUG','SEP','OCT','NOV','DEC','YEARS','# CITIES'])


for n in months:

   data1= data[[n , '# CITIES']]

   data1 = data1.apply (pd.to_numeric, errors='coerce')

   final_data=data1.dropna()


   data_par = sc.parallelize(final_data[n])

   data_par2 = sc.parallelize(final_data['# CITIES'])

   col_data=data_par.collect()

   ci_data=df_par2.collect()

   wgt = []

   for j in range(0, len(col_data)):
    wgt.append(col_data[j] * ci_data[j])

def returnIth(lst, n1, i):
    return lst[n1::i]

wgt_avgtp = []
at_cities = []
totalwgtSum = 0
tot_cities=0

for j in range(0, len(wgt)):
    wgt_avgtp=returnNth(wgt, 0, 4)
    for k in range(0, len(wgt_avgtp)):
        totalwgtSum = totalwgtSum + wgt_avgtp[k]

for k in range(0, len(ci_data)):
    at_cities=returnNth(ci_data, 0, 4)
    for j in range(0, len(at_cities)):
        tot_cities = tot_cities + at_cities[j]

avg_temp = totalwgtSum/tot_cities
wgt_pretp = []
pt_cities = []
tot_sum_pretp = 0
total_cities_pretp = 0

for k in range(0, len(wgt)):
    wgt_pretp=returnNth(wgt, 3, 4)
```

```
        for j in range(0, len(wgt_pretp)):
            tot_sum_pretp = tot_sum_pretp + wgt_pretp[j]
    for k in range(0, len(ci_data)):
        pretp_cities=returnNth(ci_data, 3, 4)
        for j in range(0, len(pretp_cities)):
            total_cities_pretp = total_cities_pretp + pretp_cities[j]

    prep_temp = tot_sum_pretp/total_cities_pretp
    print ("Average Temperature for "+i+"\t"+ str(round(Average_temp, 2)), "\tF")
    print ("Average Precipitation "+i+"\t", str(round(prep_temp, 2)), "\tIN")
```

```
OUTPUT:
Average Temperature for ANNUAL          37.95   F
Average Precipitation ANNUAL    34.47   IN
Average Temperature for JAN     12.11   F
Average Precipitation JAN       3.21    IN
Average Temperature for FEB     15.52   F
Average Precipitation FEB       2.29    IN
Average Temperature for MAR     24.63   F
Average Precipitation MAR       2.42    IN
Average Temperature for APR     37.36   F
Average Precipitation APR       2.35    IN
Average Temperature for MAY     48.42   F
Average Precipitation MAY       2.74    IN
Average Temperature for JUN     57.05   F
Average Precipitation JUN       3.15    IN
Average Temperature for JUL     62.26   F
```

Average Precipitation JUL	3.05	IN
Average Temperature for AUG	60.89	F
Average Precipitation AUG	2.88	IN
Average Temperature for SEP	52.42	F
Average Precipitation SEP	2.93	IN
Average Temperature for OCT	41.21	F
Average Precipitation OCT	3.16	IN
Average Temperature for NOV	28.09	F
Average Precipitation NOV	3.44	IN
Average Temperature for DEC	16.98	F
Average Precipitation DEC	3.15	IN