Programming Assignment 2

CWID : 10469491

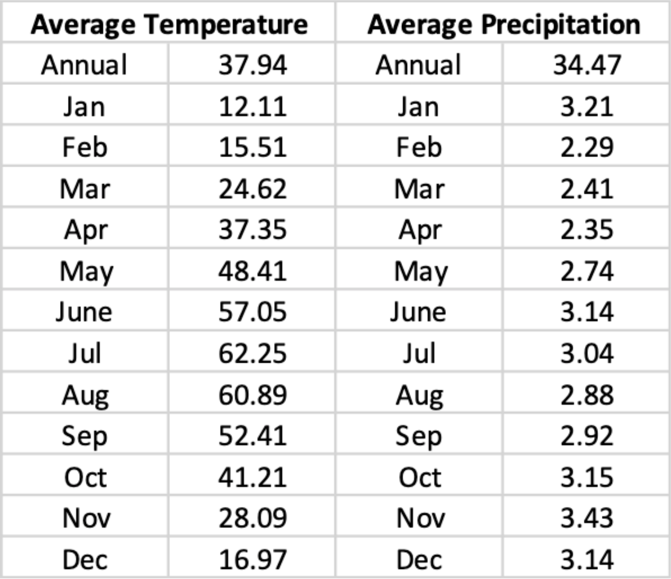
Source code

```
! pip install pyspark

import pyspark.sql.functions
import pandas as pd
from pyspark.sql.types import StructType,StructField, StringType, IntegerType
from pyspark.sql.functions import when

from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

filepath = "Class 9 - 12  - Data for Programming - Environmental - vshort.csv"
df = spark.read.format("csv")\
.option("inferSchema","true")\
.option("header","true")\
.option("sep", ",")\
.load(filepath)

df1 = df.drop("_c16")
dataframe = df1.drop(*["Years"]).na.drop()
dataframe.show()

oldname = dataframe.columns[-1]
newname = "Cities"
dataframe = dataframe.withColumnRenamed(oldname, newname)
dataframe.printSchema()

dataframe = dataframe.withColumn('Cities',
    when(dataframe.Cities.endswith('# CITIES\xa0'),
regexp_replace(dataframe.Cities,'# CITIES\xa0','0')) \
   .otherwise(dataframe.Cities))
```

```
dataframe.show()


df_temp = dataframe.filter((dataframe.Alberta != 'Average High Temperature (F)') & \
                           (dataframe.Alberta != 'Average Low Temperature (F)') & \
                           (dataframe.Alberta != 'Average Precipitation (in)') )

df_preci = dataframe.filter((dataframe.Alberta != 'Average High Temperature (F)') & \
                          (dataframe.Alberta != 'Average Low Temperature (F)') & \
                          (dataframe.Alberta != 'Average Temperature (F)'))

weights_sum_temp = df_temp.rdd.map(lambda x: (x[0],float(x[-1])))\
.reduceByKey(lambda x,y: x+y)\
.collect()[0][1]

weights_sum_temp


weights_sum_precp = df_preci.rdd.map(lambda x: (x[0], float (x[-1])))\
.reduceByKey(lambda x,y: x+y)\
.collect()[0][1]

weights_sum_precp


df_temprature = df_temp.rdd.map(lambda x: (x[0], x[1:16]))\
.flatMap(lambda x: x[1:]).map(lambda x:[(i, x[i], x[-1]) for i in range(len(x)-2)])

df_precipitation = df_preci.rdd.map(lambda x: (x[0], x[1:16]))\
.flatMap(lambda x: x[1:]).map(lambda x:[(i, x[i], x[-1]) for i in range(len(x)-2)])

temp_list = df_temprature.collect()
temp_fobject = list(filter(lambda x: x[0][1] != 'ANNUAL\xa0', temp_list))

precp_list = df_precipitation.collect()
precp_fobject = list(filter(lambda x: x[0][1] != 'ANNUAL\xa0', precp_list))

temp_yy = list(map(lambda x:[(i, float(x[i][1])*float(x[i][-1])) for i in range(len(x))],temp_fobject))

precp_yy = list(map(lambda x:[(i, float(x[i][1])*float(x[i][-1])) for i in range(len(x))],precp_fobject))

temp_rdd = spark.sparkContext.parallelize(temp_yy)
```

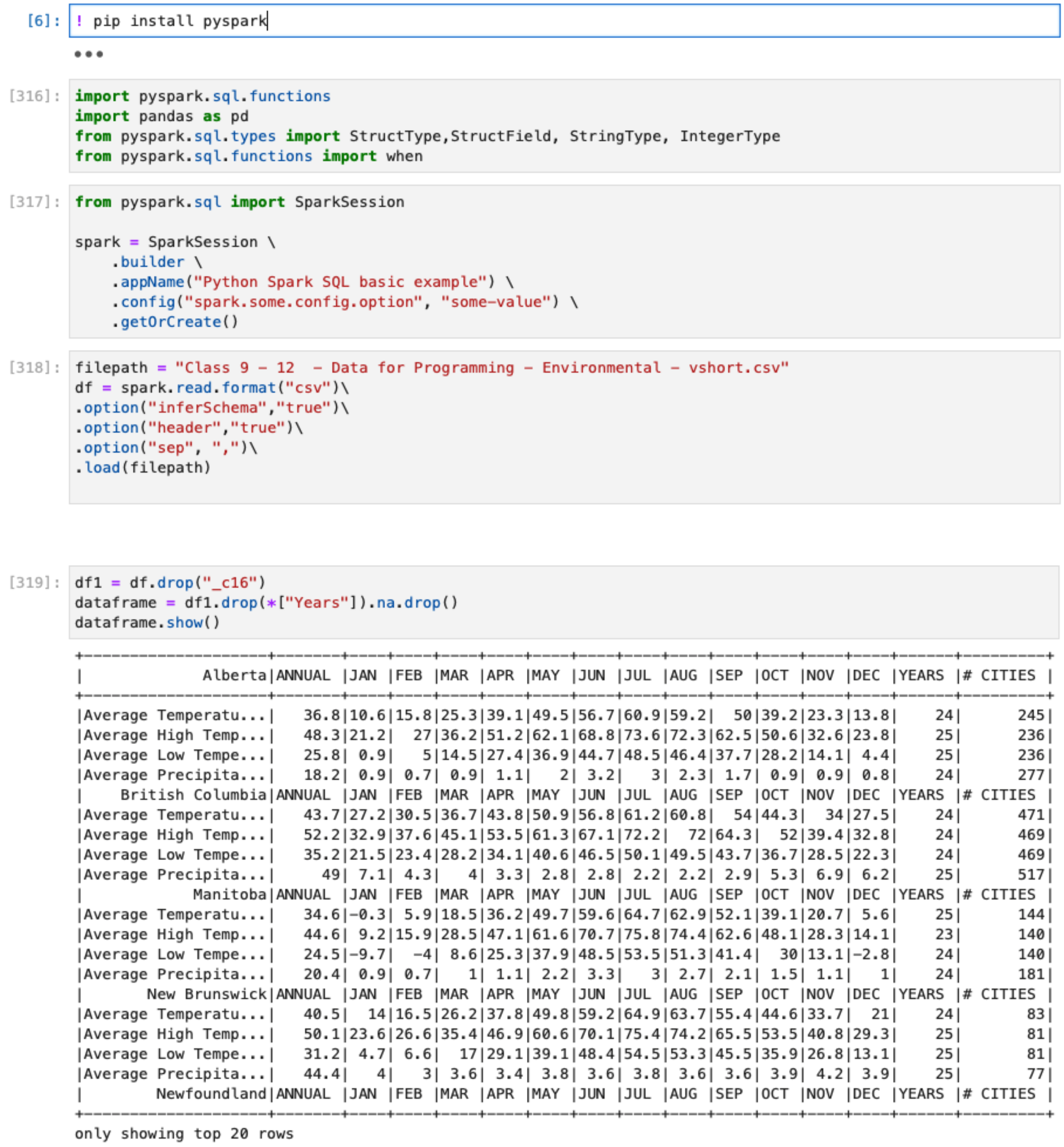
```
precp_rdd = spark.sparkContext.parallelize(precp_yy)



average_temps = temp_rdd.flatMap(lambda x:x)\
.reduceByKey(lambda x,y: x+y)\
.mapValues(lambda x: round(x/weights_sum_temp,4))

average_precp = precp_rdd.flatMap(lambda x:x)\
.reduceByKey(lambda x,y: x+y)\
.mapValues(lambda x: round(x/weights_sum_precp,4))

average_temps.toDF().show()
average_precp.toDF().show()
# 0=annual, 1 = jan, 2 = feb, 3 = mar, 4 = april, ...... 12 = dec
```

Results



Average Temperature		Average Precipitation	
Annual	37.94	Annual	34.47
Jan	12.11	Jan	3.21
Feb	15.51	Feb	2.29
Mar	24.62	Mar	2.41
Apr	37.35	Apr	2.35
May	48.41	May	2.74
June	57.05	June	3.14
Jul	62.25	Jul	3.04
Aug	60.89	Aug	2.88
Sep	52.41	Sep	2.92
Oct	41.21	Oct	3.15
Nov	28.09	Nov	3.43
Dec	16.97	Dec	3.14

Screenshots



[6]:

```
! pip install pyspark
```

•••

[316]:

```
import pyspark.sql.functions
import pandas as pd
from pyspark.sql.types import StructType,StructField, StringType, IntegerType
from pyspark.sql.functions import when
```

[317]:

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

[318]:

```
filepath = "Class 9 - 12  - Data for Programming - Environmental - vshort.csv"
df = spark.read.format("csv")\
.option("inferSchema","true")\
.option("header","true")\
.option("sep", ",")\
.load(filepath)
```

[319]:

```
df1 = df.drop("_c16")
dataframe = df1.drop(*["Years"]).na.drop()
dataframe.show()
```

```
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+---------+
|             Alberta|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |# CITIES |
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+---------+
|Average Temperatu...|   36.8|10.6|15.8|25.3|39.1|49.5|56.7|60.9|59.2|  50|39.2|23.3|13.8|    24|      245|
|Average High Temp...|   48.3|21.2|  27|36.2|51.2|62.1|68.8|73.6|72.3|62.5|50.6|32.6|23.8|    25|      236|
|Average Low Tempe...|   25.8| 0.9|   5|14.5|27.4|36.9|44.7|48.5|46.4|37.7|28.2|14.1| 4.4|    25|      236|
|Average Precipita...|   18.2| 0.9| 0.7| 0.9| 1.1|   2| 3.2|   3| 2.3| 1.7| 0.9| 0.9| 0.8|    24|      277|
|    British Columbia|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |# CITIES |
|Average Temperatu...|   43.7|27.2|30.5|36.7|43.8|50.9|56.8|61.2|60.8|  54|44.3|  34|27.5|    24|      471|
|Average High Temp...|   52.2|32.9|37.6|45.1|53.5|61.3|67.1|72.2|  72|64.3|  52|39.4|32.8|    24|      469|
|Average Low Tempe...|   35.2|21.5|23.4|28.2|34.1|40.6|46.5|50.1|49.5|43.7|36.7|28.5|22.3|    24|      469|
|Average Precipita...|     49| 7.1| 4.3|   4| 3.3| 2.8| 2.8| 2.2| 2.2| 2.9| 5.3| 6.9| 6.2|    25|      517|
|            Manitoba|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |# CITIES |
|Average Temperatu...|   34.6|-0.3| 5.9|18.5|36.2|49.7|59.6|64.7|62.9|52.1|39.1|20.7| 5.6|    25|      144|
|Average High Temp...|   44.6| 9.2|15.9|28.5|47.1|61.6|70.7|75.8|74.4|62.6|48.1|28.3|14.1|    23|      140|
|Average Low Tempe...|   24.5|-9.7|  -4| 8.6|25.3|37.9|48.5|53.5|51.3|41.4|  30|13.1|-2.8|    24|      140|
|Average Precipita...|   20.4| 0.9| 0.7|   1| 1.1| 2.2| 3.3|   3| 2.7| 2.1| 1.5| 1.1|   1|    24|      181|
|       New Brunswick|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |# CITIES |
|Average Temperatu...|   40.5|  14|16.5|26.2|37.8|49.8|59.2|64.9|63.7|55.4|44.6|33.7|  21|    24|       83|
|Average High Temp...|   50.1|23.6|26.6|35.4|46.9|60.6|70.1|75.4|74.2|65.5|53.5|40.8|29.3|    25|       81|
|Average Low Tempe...|   31.2| 4.7| 6.6|  17|29.1|39.1|48.4|54.5|53.3|45.5|35.9|26.8|13.1|    25|       81|
|Average Precipita...|   44.4|   4|   3| 3.6| 3.4| 3.8| 3.6| 3.8| 3.6| 3.6| 3.9| 4.2| 3.9|    25|       77|
|        Newfoundland|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |# CITIES |
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+---------+
only showing top 20 rows
```
]

[320]:

```
oldname = dataframe.columns[-1]
newname = "Cities"
dataframe = dataframe.withColumnRenamed(oldname, newname)
dataframe.printSchema()
```

```
root
 |-- Alberta: string (nullable = true)
 |-- ANNUAL : string (nullable = true)
 |-- JAN : string (nullable = true)
 |-- FEB : string (nullable = true)
 |-- MAR : string (nullable = true)
 |-- APR : string (nullable = true)
 |-- MAY : string (nullable = true)
 |-- JUN : string (nullable = true)
 |-- JUL : string (nullable = true)
 |-- AUG : string (nullable = true)
 |-- SEP : string (nullable = true)
 |-- OCT : string (nullable = true)
 |-- NOV : string (nullable = true)
 |-- DEC : string (nullable = true)
 |-- YEARS : string (nullable = true)
 |-- Cities: string (nullable = true)
```

[321]:

```
dataframe = dataframe.withColumn('Cities',
    when(dataframe.Cities.endswith('# CITIES\xa0'),regexp_replace(dataframe.Cities,'# CITIES\xa0','0')) \
   .otherwise(dataframe.Cities))

dataframe.show()
```

```
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+------+
|             Alberta|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |Cities|
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+------+
|Average Temperatu...|   36.8|10.6|15.8|25.3|39.1|49.5|56.7|60.9|59.2|  50|39.2|23.3|13.8|    24|   245|
|Average High Temp...|   48.3|21.2|  27|36.2|51.2|62.1|68.8|73.6|72.3|62.5|50.6|32.6|23.8|    25|   236|
|Average Low Tempe...|   25.8| 0.9|   5|14.5|27.4|36.9|44.7|48.5|46.4|37.7|28.2|14.1| 4.4|    25|   236|
|Average Precipita...|   18.2| 0.9| 0.7| 0.9| 1.1|   2| 3.2|   3| 2.3| 1.7| 0.9| 0.9| 0.8|    24|   277|
|    British Columbia|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |     0|
|Average Temperatu...|   43.7|27.2|30.5|36.7|43.8|50.9|56.8|61.2|60.8|  54|44.3|  34|27.5|    24|   471|
|Average High Temp...|   52.2|32.9|37.6|45.1|53.5|61.3|67.1|72.2|  72|64.3|  52|39.4|32.8|    24|   469|
|Average Low Tempe...|   35.2|21.5|23.4|28.2|34.1|40.6|46.5|50.1|49.5|43.7|36.7|28.5|22.3|    24|   469|
|Average Precipita...|     49| 7.1| 4.3|   4| 3.3| 2.8| 2.8| 2.2| 2.2| 2.9| 5.3| 6.9| 6.2|    25|   517|
|            Manitoba|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |     0|
|Average Temperatu...|   34.6|-0.3| 5.9|18.5|36.2|49.7|59.6|64.7|62.9|52.1|39.1|20.7| 5.6|    25|   144|
|Average High Temp...|   44.6| 9.2|15.9|28.5|47.1|61.6|70.7|75.8|74.4|62.6|48.1|28.3|14.1|    23|   140|
|Average Low Tempe...|   24.5|-9.7|  -4| 8.6|25.3|37.9|48.5|53.5|51.3|41.4|  30|13.1|-2.8|    24|   140|
|Average Precipita...|   20.4| 0.9| 0.7|   1| 1.1| 2.2| 3.3|   3| 2.7| 2.1| 1.5| 1.1|   1|    24|   181|
|       New Brunswick|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |     0|
|Average Temperatu...|   40.5|  14|16.5|26.2|37.8|49.8|59.2|64.9|63.7|55.4|44.6|33.7|  21|    24|    83|
|Average High Temp...|   50.1|23.6|26.6|35.4|46.9|60.6|70.1|75.4|74.2|65.5|53.5|40.8|29.3|    25|    81|
|Average Low Tempe...|   31.2| 4.7| 6.6|  17|29.1|39.1|48.4|54.5|53.3|45.5|35.9|26.8|13.1|    25|    81|
|Average Precipita...|   44.4|   4|   3| 3.6| 3.4| 3.8| 3.6| 3.8| 3.6| 3.6| 3.9| 4.2| 3.9|    25|    77|
|        Newfoundland|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |YEARS |     0|
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+------+------+
only showing top 20 rows
```

[322]:

```
df_temp = dataframe.filter((dataframe.Alberta != 'Average High Temperature (F)') & \
                            (dataframe.Alberta != 'Average Low Temperature (F)') & \
                            (dataframe.Alberta != 'Average Precipitation (in)') )
```

[323]:

```
df_preci = dataframe.filter((dataframe.Alberta != 'Average High Temperature (F)') & \
                            (dataframe.Alberta != 'Average Low Temperature (F)') & \
                            (dataframe.Alberta != 'Average Temperature (F)'))
```

[324]:
```
weights_sum_temp = df_temp.rdd.map(lambda x: (x[0],float(x[-1])))\
.reduceByKey(lambda x,y: x+y)\
.collect()[0][1]

weights_sum_temp
```

[324]:
```
2287.0
```

[325]:
```
weights_sum_precp = df_preci.rdd.map(lambda x: (x[0], float (x[-1])))\
.reduceByKey(lambda x,y: x+y)\
.collect()[0][1]

weights_sum_precp
```

[325]:
```
2475.0
```

[326]:
```
df_temprature = df_temp.rdd.map(lambda x: (x[0], x[1:16]))\
.flatMap(lambda x: x[1:]).map(lambda x:[(i, x[i], x[-1]) for i in range(len(x)-2)])

df_precipitation = df_preci.rdd.map(lambda x: (x[0], x[1:16]))\
.flatMap(lambda x: x[1:]).map(lambda x:[(i, x[i], x[-1]) for i in range(len(x)-2)])
```

[327]:
```
temp_list = df_temprature.collect()
temp_fobject = list(filter(lambda x: x[0][1] != 'ANNUAL\xa0', temp_list))

precp_list = df_precipitation.collect()
precp_fobject = list(filter(lambda x: x[0][1] != 'ANNUAL\xa0', precp_list))
```

[328]:
```
temp_yy = list(map(lambda x:[(i, float(x[i][1])*float(x[i][-1])) for i in range(len(x))],temp_fobject))

precp_yy = list(map(lambda x:[(i, float(x[i][1])*float(x[i][-1])) for i in range(len(x))],precp_fobject))
```

[329]:
```
temp_rdd = spark.sparkContext.parallelize(temp_yy)

precp_rdd = spark.sparkContext.parallelize(precp_yy)
```

[330]:

```
average_temps = temp_rdd.flatMap(lambda x:x)\
.reduceByKey(lambda x,y: x+y)\
.mapValues(lambda x: round(x/weights_sum_temp,4))

average_precp = precp_rdd.flatMap(lambda x:x)\
.reduceByKey(lambda x,y: x+y)\
.mapValues(lambda x: round(x/weights_sum_precp,4))
```

[331]:

```
average_temps.toDF().show()
average_precp.toDF().show()
# 0=annual, 1 = jan, 2 = feb, 3 = mar, 4 = april, ...... 12 = dec
```

```
+---+-------+
| _1|     _2|
+---+-------+
|  0|37.9463|
|  1|12.1062|
|  2|  15.52|
|  3|24.6268|
|  4|37.3558|
|  5|48.4152|
|  6|57.0529|
|  7|62.2569|
|  8|60.8941|
|  9|52.4194|
| 10|41.2087|
| 11|28.0917|
| 12|16.9792|
+---+-------+

+---+------+
| _1|    _2|
+---+------+
|  0|34.472|
|  1| 3.208|
|  2|2.2918|
|  3|2.4156|
|  4|2.3515|
|  5|2.7435|
|  6|3.1485|
|  7|3.0476|
|  8|2.8841|
|  9|2.9286|
| 10|3.1577|
| 11| 3.437|
| 12|3.1457|
+---+------+
```