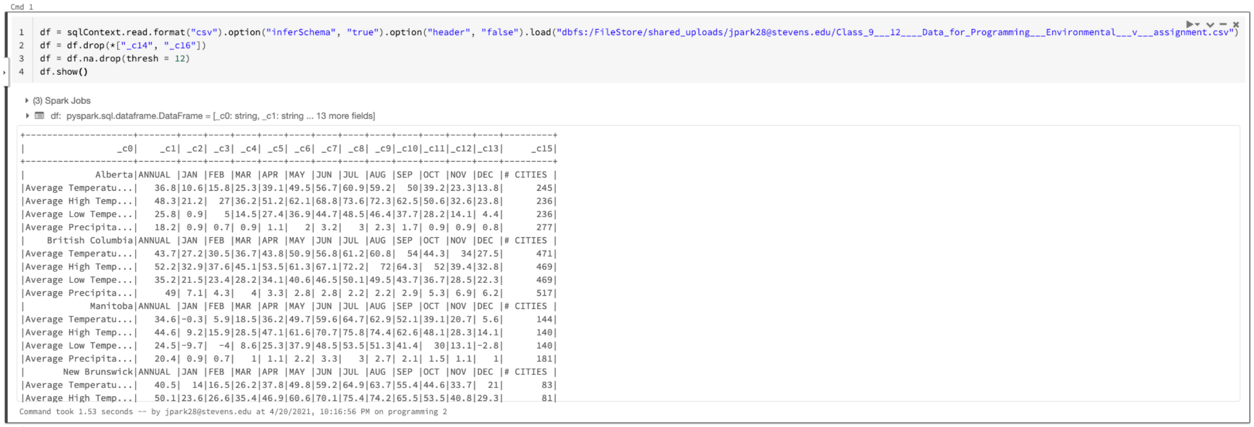
Programming Assignment 2
Jungho Park

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.

1. Import Data

Before importing, I converted “Class 9 - 12 - Data for Programming - Environmental - v – assignment.xlsx” file to csv file.



Cmd 1

```
df = sqlContext.read.format("csv").option("inferSchema", "true").option("header", "false").load("dbfs:/FileStore/shared_uploads/jpark28@stevens.edu/Class_9___12____Data_for_Programming___Environmental___v___assignment.csv")
df = df.drop(*["_c14", "_c16"])
df = df.na.drop(thresh = 12)
df.show()
```

(3) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 13 more fields]

```
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+---------+
|                 _c0|    _c1| _c2| _c3| _c4| _c5| _c6| _c7| _c8| _c9|_c10|_c11|_c12|_c13|      _c15|
+--------------------+-------+----+----+----+----+----+----+----+----+----+----+----+----+---------+
|             Alberta|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |# 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|       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|       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|       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|       277|
|    British Columbia|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |# 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|       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|       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|       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|       517|
|            Manitoba|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |# 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|       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|       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|       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|       181|
|       New Brunswick|ANNUAL |JAN |FEB |MAR |APR |MAY |JUN |JUL |AUG |SEP |OCT |NOV |DEC |# 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|        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|        81|
```

Command took 1.53 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:16:56 PM on programming 2

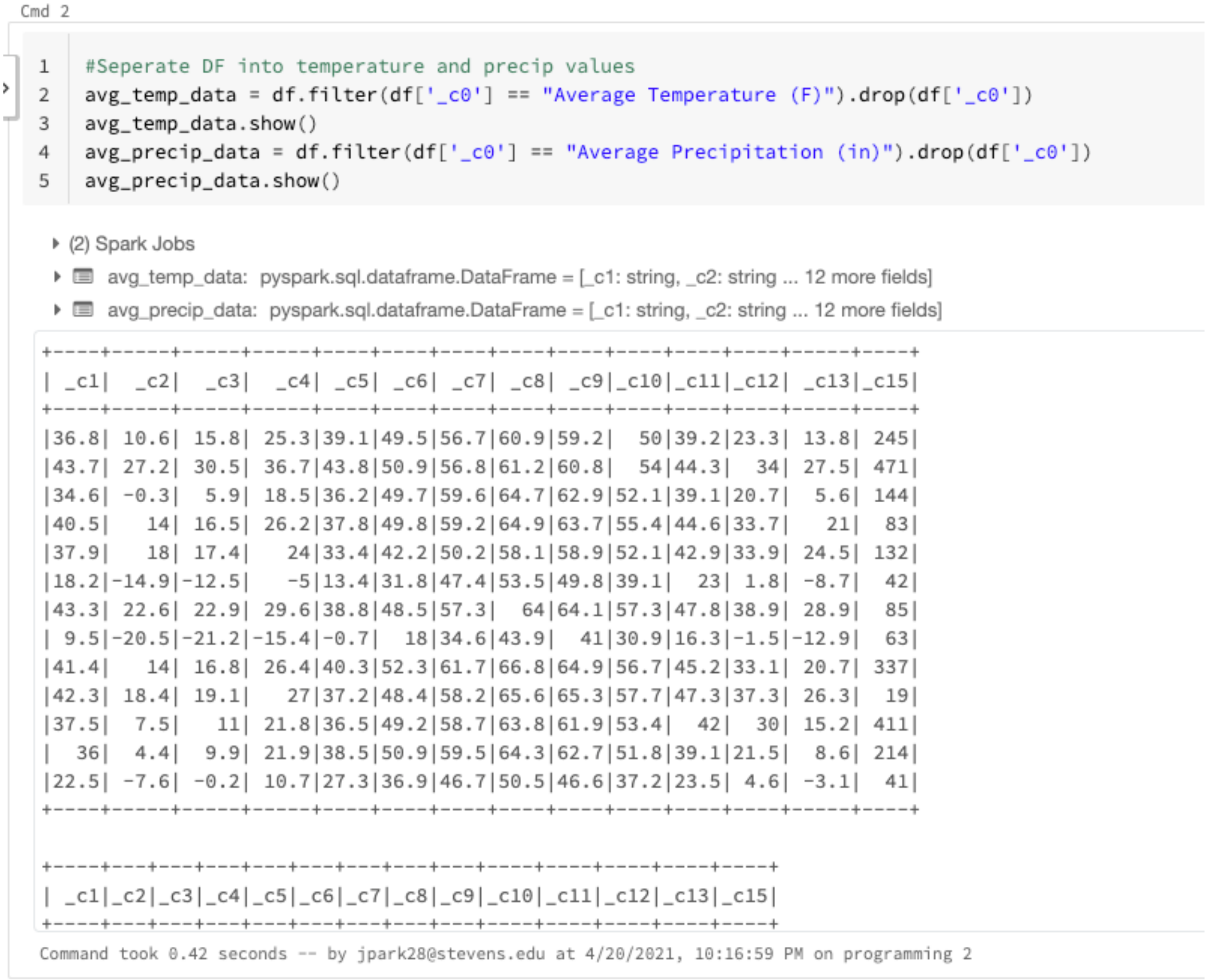
- Dropped YEAR and NULL columns (c14, c16)

Code:

df = sqlContext.read.format("csv").option("inferSchema", "true").option("header", "false").load("dbfs:/FileStore/shared_uploads/jpark28@stevens.edu/Class_9___12____Data_for_Programming___Environmental___v___assignment.csv")
df = df.drop(*["_c14", "_c16"])
df = df.na.drop(thresh = 12)
df.show()

Programming Assignment 2
Jungho Park

2. Cleaning Data
    a. Data Filtration
        i. Filtered average temperature data in to “avg_temp_data”
        ii. Filtered average precipitation data into “avg_precip_data”
        iii. Dropped index (_c0) column



Cmd 2

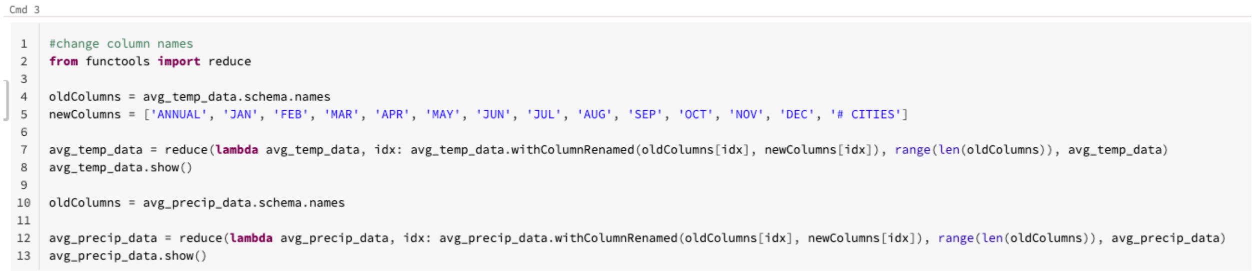
```
#Seperate DF into temperature and precip values
avg_temp_data = df.filter(df['_c0'] == "Average Temperature (F)").drop(df['_c0'])
avg_temp_data.show()
avg_precip_data = df.filter(df['_c0'] == "Average Precipitation (in)").drop(df['_c0'])
avg_precip_data.show()
```

```
avg_temp_data:  pyspark.sql.dataframe.DataFrame = [_c1: string, _c2: string ... 12 more fields]
avg_precip_data:  pyspark.sql.dataframe.DataFrame = [_c1: string, _c2: string ... 12 more fields]
```

```
+----+-----+-----+-----+----+----+----+----+----+----+----+----+-----+----+
| _c1|  _c2|  _c3|  _c4| _c5| _c6| _c7| _c8| _c9|_c10|_c11|_c12| _c13|_c15|
+----+-----+-----+-----+----+----+----+----+----+----+----+----+-----+----+
|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| 245|
|43.7| 27.2| 30.5| 36.7|43.8|50.9|56.8|61.2|60.8|  54|44.3|  34| 27.5| 471|
|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| 144|
|40.5|   14| 16.5| 26.2|37.8|49.8|59.2|64.9|63.7|55.4|44.6|33.7|   21|  83|
|37.9|   18| 17.4|   24|33.4|42.2|50.2|58.1|58.9|52.1|42.9|33.9| 24.5| 132|
|18.2|-14.9|-12.5|   -5|13.4|31.8|47.4|53.5|49.8|39.1|  23| 1.8| -8.7|  42|
|43.3| 22.6| 22.9| 29.6|38.8|48.5|57.3|  64|64.1|57.3|47.8|38.9| 28.9|  85|
| 9.5|-20.5|-21.2|-15.4|-0.7|  18|34.6|43.9|  41|30.9|16.3|-1.5|-12.9|  63|
|41.4|   14| 16.8| 26.4|40.3|52.3|61.7|66.8|64.9|56.7|45.2|33.1| 20.7| 337|
|42.3| 18.4| 19.1|   27|37.2|48.4|58.2|65.6|65.3|57.7|47.3|37.3| 26.3|  19|
|37.5|  7.5|   11| 21.8|36.5|49.2|58.7|63.8|61.9|53.4|  42|  30| 15.2| 411|
|  36|  4.4|  9.9| 21.9|38.5|50.9|59.5|64.3|62.7|51.8|39.1|21.5|  8.6| 214|
|22.5| -7.6| -0.2| 10.7|27.3|36.9|46.7|50.5|46.6|37.2|23.5| 4.6| -3.1|  41|
+----+-----+-----+-----+----+----+----+----+----+----+----+----+-----+----+


+----+---+---+---+---+---+---+---+---+----+----+----+----+----+
| _c1|_c2|_c3|_c4|_c5|_c6|_c7|_c8|_c9|_c10|_c11|_c12|_c13|_c15|
+----+---+---+---+---+---+---+---+---+----+----+----+----+----+
```

Command took 0.42 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:16:59 PM on programming 2

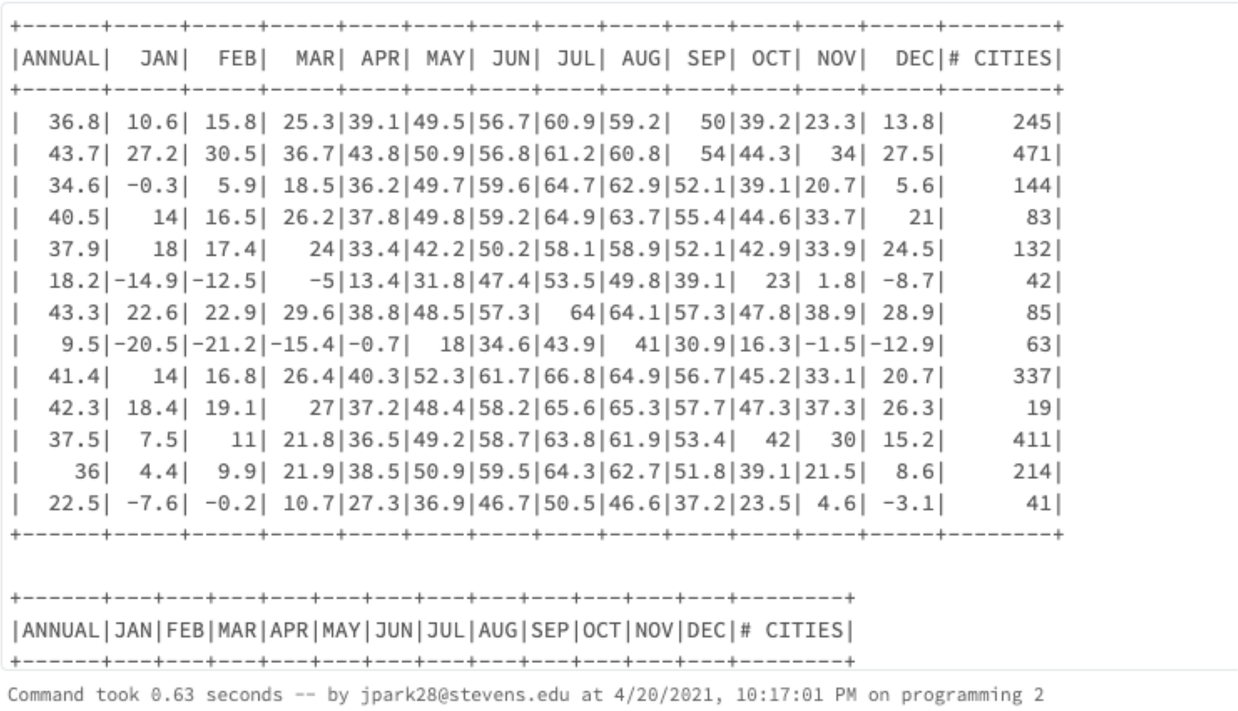
    b. Change column name



Cmd 3

```
#change column names
from functools import reduce

oldColumns = avg_temp_data.schema.names
newColumns = ['ANNUAL', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG', 'SEP', 'OCT', 'NOV', 'DEC', '# CITIES']

avg_temp_data = reduce(lambda avg_temp_data, idx: avg_temp_data.withColumnRenamed(oldColumns[idx], newColumns[idx]), range(len(oldColumns)), avg_temp_data)
avg_temp_data.show()

oldColumns = avg_precip_data.schema.names

avg_precip_data = reduce(lambda avg_precip_data, idx: avg_precip_data.withColumnRenamed(oldColumns[idx], newColumns[idx]), range(len(oldColumns)), avg_precip_data)
avg_precip_data.show()
```

Programming Assignment 2
Jungho Park



```
+------+-----+-----+-----+----+----+----+----+----+----+----+----+-----+--------+
|ANNUAL|  JAN|  FEB|  MAR| APR| MAY| JUN| JUL| AUG| SEP| OCT| NOV|  DEC|# CITIES|
+------+-----+-----+-----+----+----+----+----+----+----+----+----+-----+--------+
|  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|     245|
|  43.7| 27.2| 30.5| 36.7|43.8|50.9|56.8|61.2|60.8|  54|44.3|  34| 27.5|     471|
|  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|     144|
|  40.5|   14| 16.5| 26.2|37.8|49.8|59.2|64.9|63.7|55.4|44.6|33.7|   21|      83|
|  37.9|   18| 17.4|   24|33.4|42.2|50.2|58.1|58.9|52.1|42.9|33.9| 24.5|     132|
|  18.2|-14.9|-12.5|   -5|13.4|31.8|47.4|53.5|49.8|39.1|  23| 1.8| -8.7|      42|
|  43.3| 22.6| 22.9| 29.6|38.8|48.5|57.3|  64|64.1|57.3|47.8|38.9| 28.9|      85|
|   9.5|-20.5|-21.2|-15.4|-0.7|  18|34.6|43.9|  41|30.9|16.3|-1.5|-12.9|      63|
|  41.4|   14| 16.8| 26.4|40.3|52.3|61.7|66.8|64.9|56.7|45.2|33.1| 20.7|     337|
|  42.3| 18.4| 19.1|   27|37.2|48.4|58.2|65.6|65.3|57.7|47.3|37.3| 26.3|      19|
|  37.5|  7.5|   11| 21.8|36.5|49.2|58.7|63.8|61.9|53.4|  42|  30| 15.2|     411|
|    36|  4.4|  9.9| 21.9|38.5|50.9|59.5|64.3|62.7|51.8|39.1|21.5|  8.6|     214|
|  22.5| -7.6| -0.2| 10.7|27.3|36.9|46.7|50.5|46.6|37.2|23.5| 4.6| -3.1|      41|
+------+-----+-----+-----+----+----+----+----+----+----+----+----+-----+--------+

+------+---+---+---+---+---+---+---+---+---+---+---+---+--------+
|ANNUAL|JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC|# CITIES|
+------+---+---+---+---+---+---+---+---+---+---+---+---+--------+
```

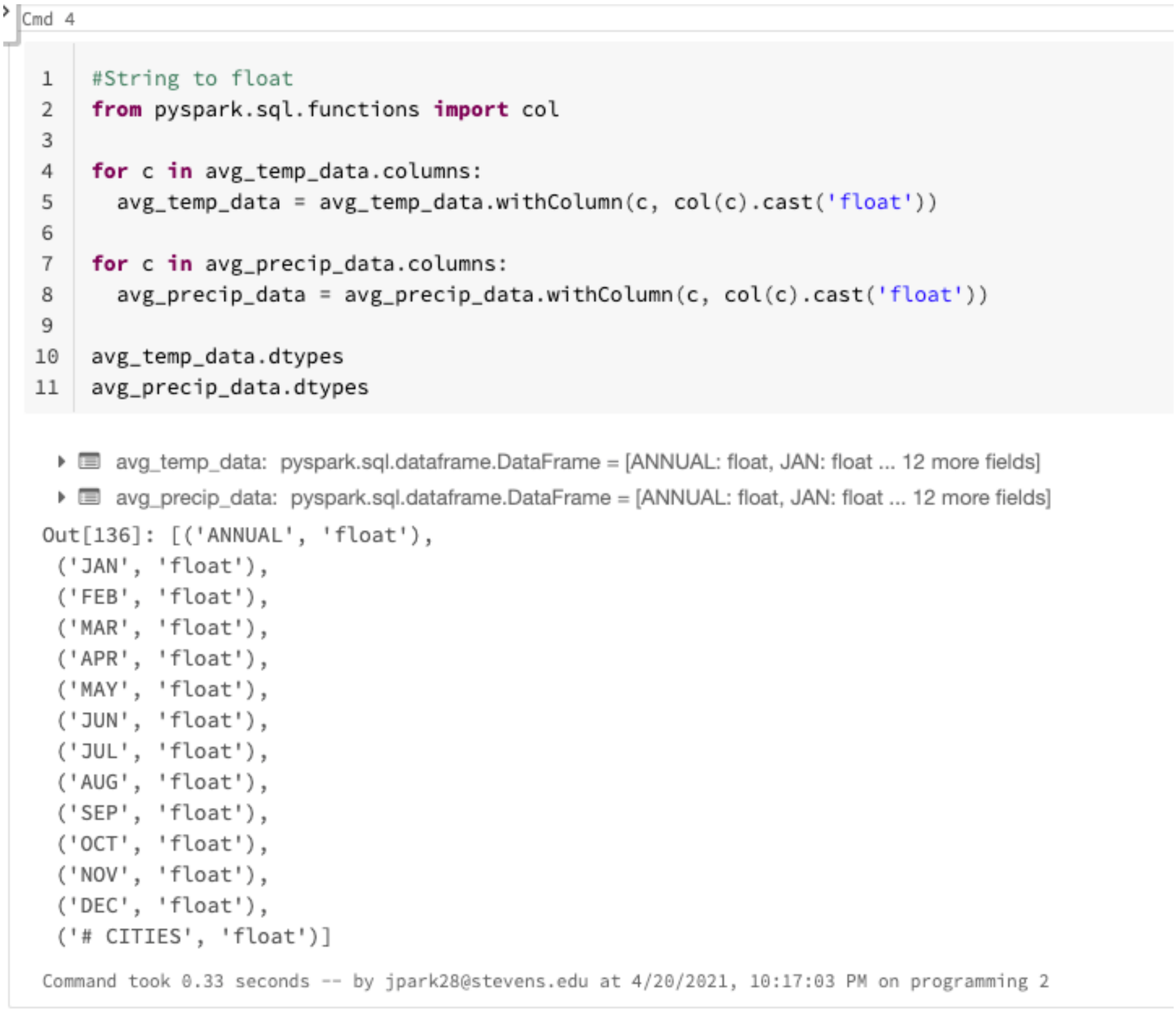
Command took 0.63 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:01 PM on programming 2

Code:

```
#change column names
from functools import reduce

oldColumns = avg_temp_data.schema.names
newColumns = ['ANNUAL', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG', 'SEP', 'OCT',
'NOV', 'DEC', '# CITIES']

avg_temp_data = reduce(lambda avg_temp_data, idx:
avg_temp_data.withColumnRenamed(oldColumns[idx], newColumns[idx]),
range(len(oldColumns)), avg_temp_data)
avg_temp_data.show()

oldColumns = avg_precip_data.schema.names

avg_precip_data = reduce(lambda avg_precip_data, idx:
avg_precip_data.withColumnRenamed(oldColumns[idx], newColumns[idx]),
range(len(oldColumns)), avg_precip_data)
avg_precip_data.show()
```

c. Change String to float

i. Needed to change all string dtypes to float in order to do aggregations & calculations.



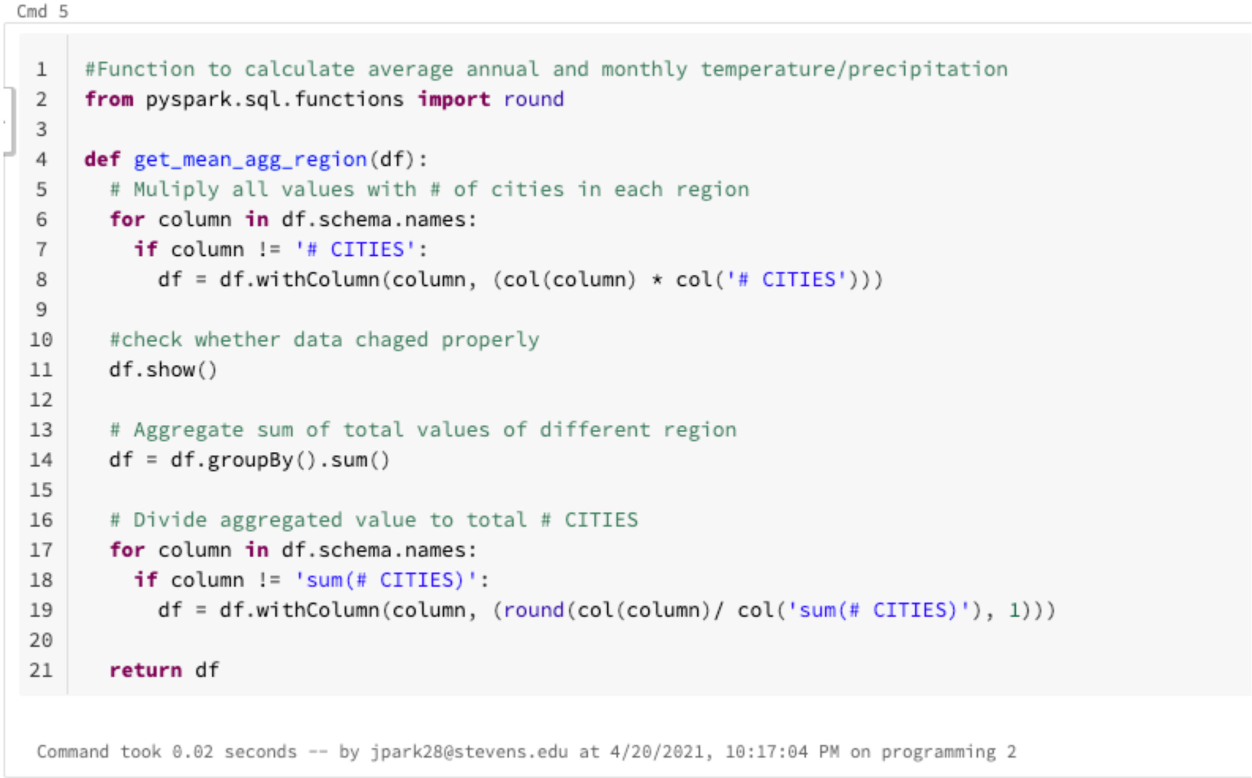
```
#String to float
from pyspark.sql.functions import col

for c in avg_temp_data.columns:
  avg_temp_data = avg_temp_data.withColumn(c, col(c).cast('float'))

for c in avg_precip_data.columns:
  avg_precip_data = avg_precip_data.withColumn(c, col(c).cast('float'))

avg_temp_data.dtypes
avg_precip_data.dtypes
```

```
avg_temp_data:  pyspark.sql.dataframe.DataFrame = [ANNUAL: float, JAN: float ... 12 more fields]
avg_precip_data:  pyspark.sql.dataframe.DataFrame = [ANNUAL: float, JAN: float ... 12 more fields]
Out[136]: [('ANNUAL', 'float'),
 ('JAN', 'float'),
 ('FEB', 'float'),
 ('MAR', 'float'),
 ('APR', 'float'),
 ('MAY', 'float'),
 ('JUN', 'float'),
 ('JUL', 'float'),
 ('AUG', 'float'),
 ('SEP', 'float'),
 ('OCT', 'float'),
 ('NOV', 'float'),
 ('DEC', 'float'),
 ('# CITIES', 'float')]
```

Command took 0.33 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:03 PM on programming 2

Programming Assignment 2
Jungho Park

3. Define function to find mean
    a. Formula: Σ(value * # of cities) / Σ(# of cities)
    b. Return calculated values
    c. Checked process by df.show()
    d. Round to 1st decimal point

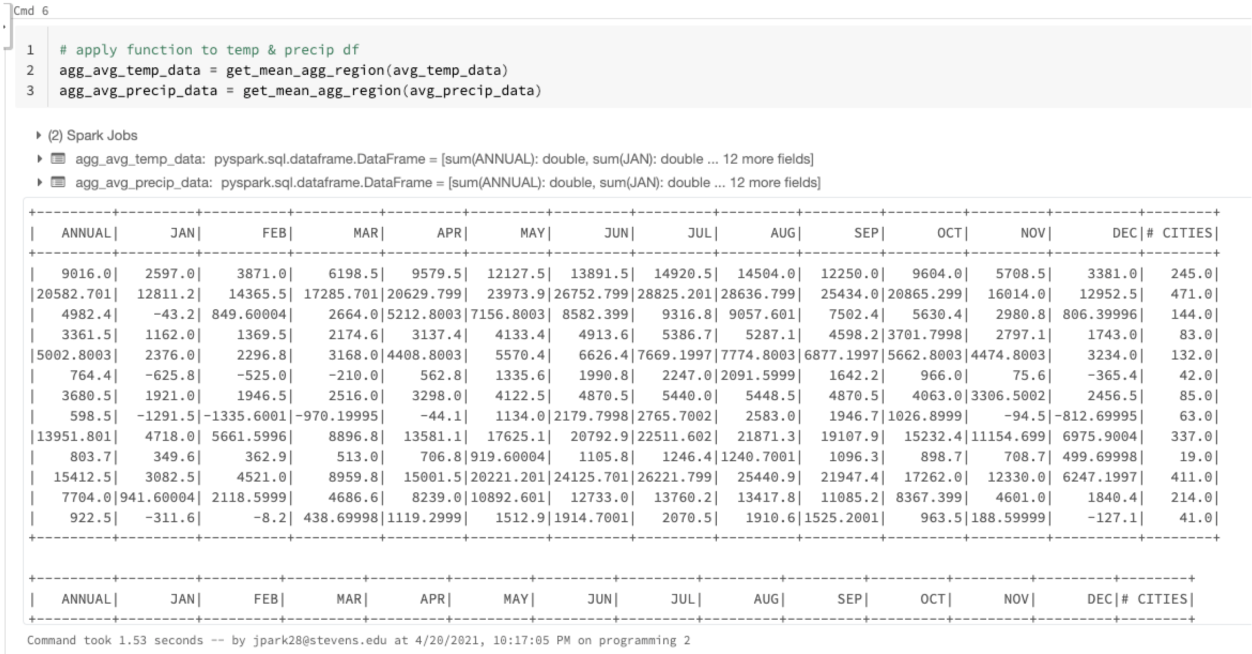


Cmd 5

```
#Function to calculate average annual and monthly temperature/precipitation
from pyspark.sql.functions import round

def get_mean_agg_region(df):
  # Muliply all values with # of cities in each region
  for column in df.schema.names:
    if column != '# CITIES':
      df = df.withColumn(column, (col(column) * col('# CITIES')))

  #check whether data chaged properly
  df.show()

  # Aggregate sum of total values of different region
  df = df.groupBy().sum()

  # Divide aggregated value to total # CITIES
  for column in df.schema.names:
    if column != 'sum(# CITIES)':
      df = df.withColumn(column, (round(col(column)/ col('sum(# CITIES)'), 1)))

  return df
```

Command took 0.02 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:04 PM on programming 2

4. Apply function to temperature and precipitation dataframe



Cmd 6

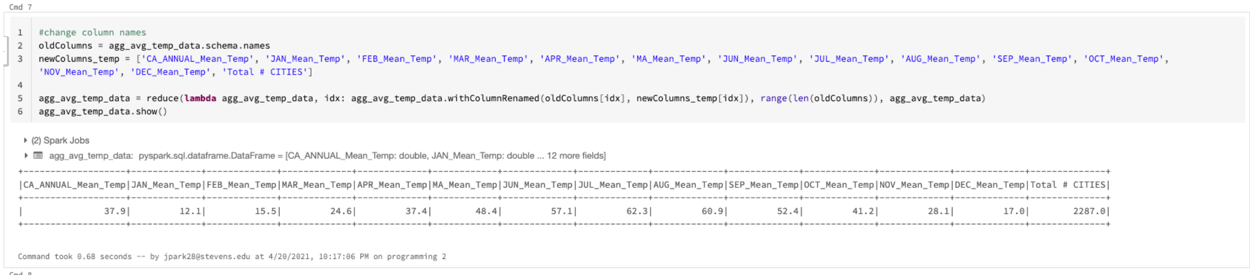
```
# apply function to temp & precip df
agg_avg_temp_data = get_mean_agg_region(avg_temp_data)
agg_avg_precip_data = get_mean_agg_region(avg_precip_data)
```

▸ (2) Spark Jobs
▸ agg_avg_temp_data: pyspark.sql.dataframe.DataFrame = [sum(ANNUAL): double, sum(JAN): double ... 12 more fields]
▸ agg_avg_precip_data: pyspark.sql.dataframe.DataFrame = [sum(ANNUAL): double, sum(JAN): double ... 12 more fields]

```
+---------+---------+----------+----------+---------+---------+---------+---------+---------+---------+---------+---------+----------+--------+
|   ANNUAL|      JAN|       FEB|       MAR|      APR|      MAY|      JUN|      JUL|      AUG|      SEP|      OCT|      NOV|       DEC|# CITIES|
+---------+---------+----------+----------+---------+---------+---------+---------+---------+---------+---------+---------+----------+--------+
|   9016.0|   2597.0|    3871.0|    6198.5|   9579.5|  12127.5|  13891.5|  14920.5|  14504.0|  12250.0|   9604.0|   5708.5|    3381.0|   245.0|
|20582.701|  12811.2|   14365.5| 17285.701|20629.799|  23973.9|26752.799|28825.201|28636.799|  25434.0|20865.299|  16014.0|   12952.5|   471.0|
|   4982.4|    -43.2| 849.60004|    2664.0|5212.8003|7156.8003| 8582.399|   9316.8| 9057.601|   7502.4|   5630.4|   2980.8| 806.39996|   144.0|
|   3361.5|   1162.0|    1369.5|    2174.6|   3137.4|   4133.4|   4913.6|   5386.7|   5287.1|   4598.2|3701.7998|   2797.1|    1743.0|    83.0|
|5002.8003|   2376.0|    2296.8|    3168.0|4408.8003|   5570.4|   6626.4|7669.1997|7774.8003|6877.1997|5662.8003|4474.8003|    3234.0|   132.0|
|    764.4|   -625.8|    -525.0|    -210.0|    562.8|   1335.6|   1990.8|   2247.0|2091.5999|   1642.2|    966.0|     75.6|    -365.4|    42.0|
|   3680.5|   1921.0|    1946.5|    2516.0|   3298.0|   4122.5|   4870.5|   5440.0|   5448.5|   4870.5|   4063.0|3306.5002|    2456.5|    85.0|
|    598.5|  -1291.5|-1335.6001|-970.19995|    -44.1|   1134.0|2179.7998|2765.7002|   2583.0|   1946.7|1026.8999|    -94.5|-812.69995|    63.0|
|13951.801|   4718.0| 5661.5996|    8896.8|  13581.1|  17625.1|  20792.9|22511.602|  21871.3|  19107.9|  15232.4|11154.699| 6975.9004|   337.0|
|    803.7|    349.6|     362.9|     513.0|    706.8|919.60004|   1105.8|   1246.4|1240.7001|   1096.3|    898.7|    708.7| 499.69998|    19.0|
|  15412.5|   3082.5|    4521.0|    8959.8|  15001.5|20221.201|24125.701|26221.799|  25440.9|  21947.4|  17262.0|  12330.0| 6247.1997|   411.0|
|   7704.0|941.60004| 2118.5999|    4686.6|   8239.0|10892.601|  12733.0|  13760.2|  13417.8|  11085.2| 8367.399|   4601.0|    1840.4|   214.0|
|    922.5|   -311.6|      -8.2| 438.69998|1119.2999|   1512.9|1914.7001|   2070.5|   1910.6|1525.2001|    963.5|188.59999|    -127.1|    41.0|
+---------+---------+----------+----------+---------+---------+---------+---------+---------+---------+---------+---------+----------+--------+


+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--------+
|   ANNUAL|      JAN|      FEB|      MAR|      APR|      MAY|      JUN|      JUL|      AUG|      SEP|      OCT|      NOV|      DEC|# CITIES|
+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--------+
```

Command took 1.53 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:05 PM on programming 2

5. Display dataframe with changed column names
    a. Average Temperature



Cmd 7

```
#change column names
oldColumns = agg_avg_temp_data.schema.names
newColumns_temp = ['CA_ANNUAL_Mean_Temp', 'JAN_Mean_Temp', 'FEB_Mean_Temp', 'MAR_Mean_Temp', 'APR_Mean_Temp', 'MA_Mean_Temp', 'JUN_Mean_Temp', 'JUL_Mean_Temp', 'AUG_Mean_Temp', 'SEP_Mean_Temp', 'OCT_Mean_Temp',
'NOV_Mean_Temp', 'DEC_Mean_Temp', 'Total # CITIES']

agg_avg_temp_data = reduce(lambda agg_avg_temp_data, idx: agg_avg_temp_data.withColumnRenamed(oldColumns[idx], newColumns_temp[idx]), range(len(oldColumns)), agg_avg_temp_data)
agg_avg_temp_data.show()
```

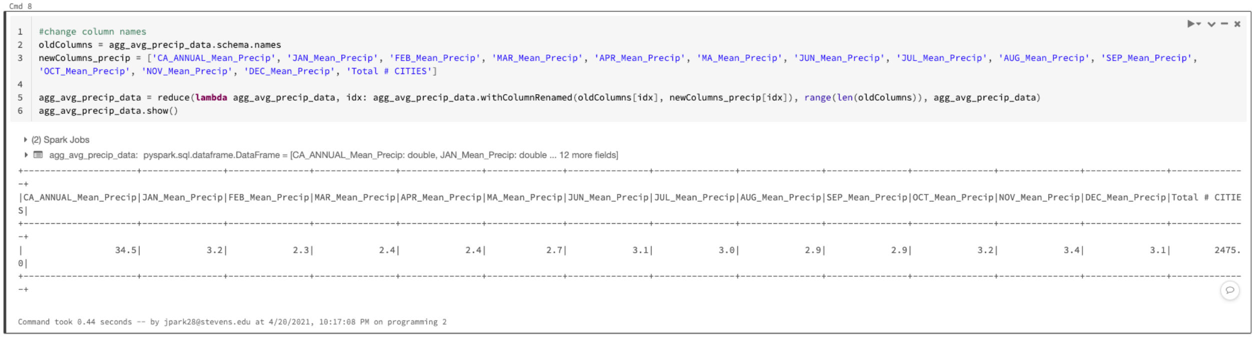
```
▸ (2) Spark Jobs
▸ agg_avg_temp_data:  pyspark.sql.dataframe.DataFrame = [CA_ANNUAL_Mean_Temp: double, JAN_Mean_Temp: double ... 12 more fields]
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+
|CA_ANNUAL_Mean_Temp|JAN_Mean_Temp|FEB_Mean_Temp|MAR_Mean_Temp|APR_Mean_Temp|MA_Mean_Temp|JUN_Mean_Temp|JUL_Mean_Temp|AUG_Mean_Temp|SEP_Mean_Temp|OCT_Mean_Temp|NOV_Mean_Temp|DEC_Mean_Temp|Total # CITIES|
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+
|               37.9|         12.1|         15.5|         24.6|         37.4|        48.4|         57.1|         62.3|         60.9|         52.4|         41.2|         28.1|         17.0|        2287.0|
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+

Command took 0.68 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:06 PM on programming 2
```

Code:

#change column names
oldColumns = agg_avg_precip_data.schema.names
newColumns_precip = ['CA_ANNUAL_Mean_Precip', 'JAN_Mean_Precip', 'FEB_Mean_Precip', 'MAR_Mean_Precip', 'APR_Mean_Precip', 'MA_Mean_Precip', 'JUN_Mean_Precip', 'JUL_Mean_Precip', 'AUG_Mean_Precip', 'SEP_Mean_Precip', 'OCT_Mean_Precip', 'NOV_Mean_Precip', 'DEC_Mean_Precip', 'Total # CITIES']

agg_avg_precip_data = reduce(lambda agg_avg_precip_data, idx: agg_avg_precip_data.withColumnRenamed(oldColumns[idx], newColumns_precip[idx]), range(len(oldColumns)), agg_avg_precip_data)
agg_avg_precip_data.show()

    b. Average Precipitation



Cmd 8

```
#change column names
oldColumns = agg_avg_precip_data.schema.names
newColumns_precip = ['CA_ANNUAL_Mean_Precip', 'JAN_Mean_Precip', 'FEB_Mean_Precip', 'MAR_Mean_Precip', 'APR_Mean_Precip', 'MA_Mean_Precip', 'JUN_Mean_Precip', 'JUL_Mean_Precip', 'AUG_Mean_Precip', 'SEP_Mean_Precip',
'OCT_Mean_Precip', 'NOV_Mean_Precip', 'DEC_Mean_Precip', 'Total # CITIES']

agg_avg_precip_data = reduce(lambda agg_avg_precip_data, idx: agg_avg_precip_data.withColumnRenamed(oldColumns[idx], newColumns_precip[idx]), range(len(oldColumns)), agg_avg_precip_data)
agg_avg_precip_data.show()
```

```
▸ (2) Spark Jobs
▸ agg_avg_precip_data:  pyspark.sql.dataframe.DataFrame = [CA_ANNUAL_Mean_Precip: double, JAN_Mean_Precip: double ... 12 more fields]
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+
|CA_ANNUAL_Mean_Precip|JAN_Mean_Precip|FEB_Mean_Precip|MAR_Mean_Precip|APR_Mean_Precip|MA_Mean_Precip|JUN_Mean_Precip|JUL_Mean_Precip|AUG_Mean_Precip|SEP_Mean_Precip|OCT_Mean_Precip|NOV_Mean_Precip|DEC_Mean_Precip|Total # CITIE
S|
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+
|                 34.5|            3.2|            2.3|            2.4|            2.4|           2.7|            3.1|            3.0|            2.9|            2.9|            3.2|            3.4|            3.1|        2475.
0|
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+

Command took 0.44 seconds -- by jpark28@stevens.edu at 4/20/2021, 10:17:08 PM on programming 2
```

Code:

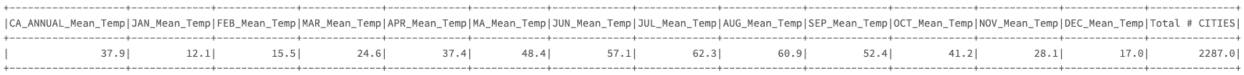
#change column names
oldColumns = agg_avg_precip_data.schema.names
newColumns_precip = ['CA_ANNUAL_Mean_Precip', 'JAN_Mean_Precip', 'FEB_Mean_Precip', 'MAR_Mean_Precip', 'APR_Mean_Precip', 'MA_Mean_Precip', 'JUN_Mean_Precip', 'JUL_Mean_Precip', 'AUG_Mean_Precip', 'SEP_Mean_Precip', 'OCT_Mean_Precip', 'NOV_Mean_Precip', 'DEC_Mean_Precip', 'Total # CITIES']

Programming Assignment 2
Jungho Park

```
agg_avg_precip_data = reduce(lambda agg_avg_precip_data, idx:
agg_avg_precip_data.withColumnRenamed(oldColumns[idx], newColumns_precip[idx]),
range(len(oldColumns)), agg_avg_precip_data)
agg_avg_precip_data.show()
```

6. Answers

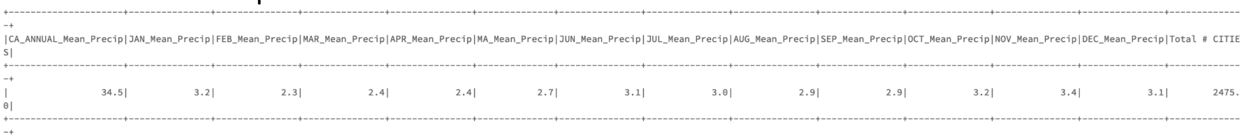
   a. Temperature



```
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+
|CA_ANNUAL_Mean_Temp|JAN_Mean_Temp|FEB_Mean_Temp|MAR_Mean_Temp|APR_Mean_Temp|MA_Mean_Temp|JUN_Mean_Temp|JUL_Mean_Temp|AUG_Mean_Temp|SEP_Mean_Temp|OCT_Mean_Temp|NOV_Mean_Temp|DEC_Mean_Temp|Total # CITIES|
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+
|               37.9|         12.1|         15.5|         24.6|         37.4|        48.4|         57.1|         62.3|         60.9|         52.4|         41.2|         28.1|         17.0|        2287.0|
+-------------------+-------------+-------------+-------------+-------------+------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+--------------+
```

ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
37.9	12.1	15.5	24.6	37.4	48.4	57.1	62.3	60.9	52.4	41.2	28.1	17.0

   b. Precipitation



```
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+
|CA_ANNUAL_Mean_Precip|JAN_Mean_Precip|FEB_Mean_Precip|MAR_Mean_Precip|APR_Mean_Precip|MA_Mean_Precip|JUN_Mean_Precip|JUL_Mean_Precip|AUG_Mean_Precip|SEP_Mean_Precip|OCT_Mean_Precip|NOV_Mean_Precip|DEC_Mean_Precip|Total # CITIE
S|
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+
|                 34.5|            3.2|            2.3|            2.4|            2.4|           2.7|            3.1|            3.0|            2.9|            2.9|            3.2|            3.4|            3.1|        2475.
0|
+---------------------+---------------+---------------+---------------+---------------+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
-+
```

ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
34.5	3.2	2.3	2.4	2.4	2.7	3.1	3.0	2.9	2.9	3.2	3.4	3.1