databricks Programming_Assignment_2

(https://databricks.com)

1

```
from pyspark.sql.functions import monotonically_increasing_id, col, struct, expr, array
from itertools import product
```

Programming Assignment 2

Wyatt Blair

12/2/2024

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.

Load the data

4

```
%sql
use catalog `hive_metastore`; select * from `default`.`class_9___12____data_for_programming___environmental___vshort_3_csv`
limit 100;
```

Table

	_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8
1	Alberta	ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL
2	Average Temperature (F)	36.8	10.6	15.8	25.3	39.1	49.5	56.7	60.9
3	Average High Temperature (F)	48.3	21.2	27	36.2	51.2	62.1	68.8	73.6
4	Average Low Temperature (F)	25.8	0.9	5	14.5	27.4	36.9	44.7	48.5
5	Average Precipitation (in)	18.2	0.9	0.7	0.9	1.1	2	3.2	3
6	null	null	null	null	null	null	null	null	null
7	British Columbia	ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL
8	Average Temperature (F)	43.7	27.2	30.5	36.7	43.8	50.9	56.8	61.2
9	Average High Temperature (F)	52.2	32.9	37.6	45.1	53.5	61.3	67.1	72.2
10	Average Low Temperature (F)	35.2	21.5	23.4	28.2	34.1	40.6	46.5	50.1
11	Average Precipitation (in)	49	7.1	4.3	4	3.3	2.8	2.8	2.2
12	null	null	null	null	null	null	null	null	null

13	Manitoba	ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL
14	Average Temperature (F)	34.6	-0.3	5.9	18.5	36.2	49.7	59.6	64.7
15	Average High Temperature (F)	44.6	9.2	15.9	28.5	47.1	61.6	70.7	75.8

77 rows

This result is stored as `_sqldf` and can be used in other Python cells.

5

```
raw_data = _sqldf.withColumn('row_index', monotonically_increasing_id())
```

Clean up data

7

```
provinces = raw_data.filter(col('row_index') % 6 == 0).drop('row_index').select('_c0').withColumnRenamed('_c0', 'Province')
n_provinces = provinces.count()

display(provinces)
```

Table

	Province
1	Alberta
2	British Columbia
3	Manitoba
4	New Brunswick
5	Newfoundland
6	Northwest Territories
7	Nova Scotia
8	Nunavut
9	Ontario
10	Prince Edward Island
11	Quebec
12	Saskatchewan
13	Yukon

13 rows

8

```
headers = raw_data.collect()[0][1:-2]
headers = [header.strip().replace('# ', 'N_') for header in headers]
print(headers)
```

```
['ANNUAL', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'YEARS', 'N_CITIES']
```

9

```
measurements = raw_data.filter(
  (col('row_index') > 0)
  &
  (col('row_index') < 5)
  ).select('_c0').withColumnRenamed('_c0', 'Measurement')
display(measurements)
```

Table

	Measurement
1	Average Temperature (F)
2	Average High Temperature (F)
3	Average Low Temperature (F)
4	Average Precipitation (in)

4 rows

10

```
data = raw_data.filter(
  (col('row_index') % 6 > 0)
  &
  (col('row_index') % 6 < 5)
  )
data = data.drop('_c0', '_c16', 'row_index')

for i, header in enumerate(headers):
  data = data.withColumnRenamed(f'_c{i+1}', header.strip())

display(data)
```

Table

	ANNUAL	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
18	45.5	25.8	25.7	31.8	40.5	50.4	59.1	66.8	67.2	60
19	30.9	11	9.9	16.8	27	34.5	41.9	50	51.2	44.8
20	49.2	4.5	3.9	3.8	3.5	3.5	3.9	3.9	4	4.5
21	18.2	-14.9	-12.5	-5	13.4	31.8	47.4	53.5	49.8	39.1

22	25.5	-7.8	-4.6	4.3	22.7	40.2	56.1	62.2	57.8	45.3
23	10.7	-22.2	-20.8	-14.5	4	23.6	38.6	45	42.1	32.9
24	10	0.6	0.5	0.4	0.4	0.6	1	1.4	1.6	1.2
25	43.3	22.6	22.9	29.6	38.8	48.5	57.3	64	64.1	57.3
26	51.5	30.5	31	37.1	46.6	57.7	66.9	73.1	73.1	66.1
27	35.1	14.7	14.7	22	31.3	39.3	47.8	54.8	55.1	48.7
28	53.5	5.2	4.2	4.7	4.3	4.1	3.8	3.6	3.7	4.3
29	9.5	-20.5	-21.2	-15.4	-0.7	18	34.6	43.9	41	30.9
30	15.5	-14.1	-14.6	-8.4	6.7	24	39.9	50.5	46.5	34.9
31	3.7	-27.1	-27.8	-22.5	-8	12.1	29.7	37.3	35.6	27.1
32	9.3	0.4	0.3	0.4	0.5	0.7	0.7	1.3	1.5	1.3
33	41.4	14	16.8	26.4	40.3	52.3	61.7	66.8	64.9	56.7
34	50.7	22.6	26.1	35.8	50.1	63.2	72.4	77.4	75.1	66.2

52 rows

11

```
multi_index = provinces.crossJoin(measurements)
display(multi_index)
```

Table

	Province	Measurement
1	Alberta	Average Temperature (F)
2	Alberta	Average High Temperature (F)
3	Alberta	Average Low Temperature (F)
4	Alberta	Average Precipitation (in)
5	British Columbia	Average Temperature (F)
6	British Columbia	Average High Temperature (F)
7	British Columbia	Average Low Temperature (F)
8	British Columbia	Average Precipitation (in)
9	Manitoba	Average Temperature (F)
10	Manitoba	Average High Temperature (F)
11	Manitoba	Average Low Temperature (F)
12	Manitoba	Average Precipitation (in)
13	New Brunswick	Average Temperature (F)
14	New Brunswick	Average High Temperature (F)
15	New Brunswick	Average Low Temperature (F)

52 rows

Create DataFrame with multi-index for organization's sake

13

```
multi_index = multi_index.withColumn("row_index", monotonically_increasing_id())
full_data = data.withColumn("row_index", monotonically_increasing_id())

full_data = multi_index.join(full_data, on="row_index").drop("row_index")

for header in headers:
  full_data = full_data.withColumn(header, full_data[header].cast('double'))

display(full_data)
```

Table

	Province	Measurement	1.2 ANNUAL	1.2 JAN	1.2 FEB	1.2 MAR	1.2 APR	1.2 MAY
1	Alberta	Average Temperature (F)	36.8	10.6	15.8	25.3	39.1	49.5
2	Alberta	Average High Temperature (F)	48.3	21.2	27	36.2	51.2	62.1
3	Alberta	Average Low Temperature (F)	25.8	0.9	5	14.5	27.4	36.9
4	Alberta	Average Precipitation (in)	18.2	0.9	0.7	0.9	1.1	2
5	British Columbia	Average Temperature (F)	43.7	27.2	30.5	36.7	43.8	50.9
6	British Columbia	Average High Temperature (F)	52.2	32.9	37.6	45.1	53.5	61.3
7	British Columbia	Average Low Temperature (F)	35.2	21.5	23.4	28.2	34.1	40.6
8	British Columbia	Average Precipitation (in)	49	7.1	4.3	4	3.3	2.8
9	Manitoba	Average Temperature (F)	34.6	-0.3	5.9	18.5	36.2	49.7
10	Manitoba	Average High Temperature (F)	44.6	9.2	15.9	28.5	47.1	61.6
11	Manitoba	Average Low Temperature (F)	24.5	-9.7	-4	8.6	25.3	37.9
12	Manitoba	Average Precipitation (in)	20.4	0.9	0.7	1	1.1	2.2
13	New Brunswick	Average Temperature (F)	40.5	14	16.5	26.2	37.8	49.8
14	New Brunswick	Average High Temperature (F)	50.1	23.6	26.6	35.4	46.9	60.6
15	New Brunswick	Average Low Temperature (F)	31.2	4.7	6.6	17	29.1	39.1

52 rows

Weight values by `N_CITIES` column

15

```
non_numerical_cols = ["Province", "Measurement", "YEARS", "N_CITIES"]
numerical_cols = [c for c in full_data.columns if c not in non_numerical_cols]

meas_df = full_data
for numerical_col in numerical_cols:
  meas_df = meas_df.withColumn(f'WEIGHTED_{numerical_col}', col(numerical_col) * col('N_CITIES'))
  meas_df = meas_df.drop(numerical_col)

meas_df = meas_df.drop('Province', 'YEARS')
display(meas_df)
```

Table

	Measurement	1.2 N_CITIES	1.2 WEIGHTED_ANNUAL	1.2 WEIGHTED_JAN	1.2 WEIGHTED_FEB	1.2 WEIGHTE
1	Average Temperature (F)	245	9016	2597	3871	
2	Average Temperature (F)	236	11398.8	5003.2	6372	
3	Average Temperature (F)	236	6088.8	212.4	1180	
4	Average Temperature (F)	277	5041.4	249.3	193.89999999999998	
5	Average Temperature (F)	471	20582.7	12811.199999999999	14365.5	
6	Average Temperature (F)	469	24481.800000000003	15430.099999999999	17634.4	
7	Average Temperature (F)	469	16508.800000000003	10083.5	10974.599999999999	
8	Average Temperature (F)	517	25333	3670.7	2223.1	
9	Average Temperature (F)	144	4982.400000000001	-43.199999999999996	849.6	
10	Average Temperature (F)	140	6244	1288	2226	
11	Average Temperature (F)	140	3430	-1358	-560	
12	Average Temperature (F)	181	3692.3999999999996	162.9	126.69999999999999	
13	Average Temperature (F)	83	3361.5	1162	1369.5	
14	Average High Temperature (F)	81	4058.1	1911.6000000000001	2154.6	
15	Average High Temperature (F)	81	2527.2	380.7	534.6	

52 rows

Group by `Measurement` column and take the weighted average

17

```
result_df = meas_df.groupBy("Measurement").sum()
for field in result_df.schema.fields:

  if field.name not in ['Measurement', 'sum(N_CITIES)']:
    result_df = result_df.withColumn(f'avg({field.name.removeprefix("sum(WEIGHTED_").removesuffix(")")})', col(field.name)
    ('sum(N_CITIES)'))
    result_df = result_df.drop(field.name)

result_df = result_df.drop('sum(N_CITIES)')
result_df = result_df.filter((col('Measurement') == 'Average Temperature (F)') | (col('Measurement') == 'Average Precipitat
(in)'))
display(result_df)
```

Table

	ABC Measurement	1.2 avg(ANNUAL)	1.2 avg(JAN)	1.2 avg(FEB)	1.2 avg(MAR)	1.2 avg(APR)	1.
1	Average Temperature (F)	38.84745011086475	14.209839246119733	16.85873059866962	22.82649667405765	30.637305986696234	3
2	Average Precipitation (in)	34.69959394610557	4.680324843115541	7.785529715762274	16.088261351052047	27.43831672203765	

2 rows

18

```
Average Temperature (F)
        --> avg(ANNUAL): 38.85
        --> avg(JAN): 14.21
        --> avg(FEB): 16.86
        --> avg(MAR): 22.83
        --> avg(APR): 30.64
        --> avg(MAY): 37.38
        --> avg(JUN): 42.63
        --> avg(JUL): 45.80
        --> avg(AUG): 45.00
        --> avg(SEP): 39.08
        --> avg(OCT): 31.65
        --> avg(NOV): 22.22
        --> avg(DEC): 15.68
Average Precipitation (in)
        --> avg(ANNUAL): 34.70
        --> avg(JAN): 4.68
        --> avg(FEB): 7.79
        --> avg(MAR): 16.09
        --> avg(APR): 27.44
        --> avg(MAY): 36.75
```