CS450 - Final exam

Katya Griffiths-Julien, Jerry Lau

Problem I: MapReduce implementation using Python and mrjob

1. Total Incomes
   The totalIncome function mapper splits the lines into individual items and emits a key-value pair with a null key and the income as the value. Reducer then calculates the total sum of incomes.
   ```
   Total Income (Trial): 63168
   Total Income (Test): 210015551664
   ```

2. Mean
   The meanIncomes function calculates the average income using mapper which extracts individual incomes from each line, emitting key-value pairs with a null key and the income as the value. The reducer then computes the mean by summing up all incomes and dividing by the total number of incomes.
   ```
   Mean Income (Trial): 63.168
   Mean Income (Test): 21001.5551664
   ```

3. Generalized Mean
   For the generalizedMeanIncomes function, mapper extracts individual incomes, emitting key-value pairs with a null key and the income as the value. Reducer then calculates the generalized mean using the given order. This demonstrates the flexibility of MapReduce in handling generalized mean calculations for different orders. We set p = 3 since a p = 2 would give the test dataset NaN since you cannot do a square root of a negative number.
   ```
   Generalized Mean (p=3) (Trial): 1670.5617751539044
   Generalized Mean (p=3) (Test): 6464.506503065638
   ```

4. Maximum
   The maxIncome function mapper extracts individual incomes, emitting key-value pairs with a null key and the income as the value. Reducer then determines the maximum income from the received values and outputs the result. MapReduce efficiently handled finding maximum income in the large dataset.
   ```
   Maximum Income (Trial): 13473
   Maximum Income (Test): 164016448792
   ```

5. Minimum
   The minIncome mapper extracts individual incomes, emitting key-value pairs with a null key and the income as the value. Reducer determines the minimum income from the received values.

```
Minimum Income (Trial): 1
Minimum Income (Test): 1
```

6. Standard deviation
   The stdDevIncomes function mapper extracts individual incomes, emitting key-value pairs with a null key and the income as the value. Reducer calculates the mean and then computes the variance and standard deviation.

```
Standard Deviation (Trial): 1.9975475523751618
Standard Deviation (Test): 6.641274873150054
```

Problem II: Sparkifying world cities

1. Simple cleaning worldcitiespop
   This task required us to remove all entries where the population was null. Using the filter() function, we kept the values that are not null and discarded the rest.

2. Statistics
   This is the output of the cities and their population

```
+---------------+
|sum(population)|
+---------------+
|     2289584999|
+---------------+

+---------------+
|min(population)|
+---------------+
|              7|
+---------------+

+---------------+
|max(population)|
+---------------+
|       31480498|
+---------------+

+-----------------+
|  avg(population)|
+-----------------+
|47719.57063359733|
+-----------------+
```

3. Histograms
   In this task, we did a log and then floored the values then called a count to group them for the result

rounded_value	count
0	5
1	174
2	2187
3	20537
4	21550
5	3248
6	269
7	10

4. TopK
   These were the top 10 cities in the dataset

Country	City	AccentCity	Region	Population	Latitude	Longitude
jp	tokyo	Tokyo	40	31480498	35.685	139.751389
cn	shanghai	Shanghai	23	14608512	31.045556	121.399722
in	bombay	Bombay	16	12692717	18.975	72.825833
pk	karachi	Karachi	05	11627378	24.9056	67.0822
in	delhi	Delhi	07	10928270	28.666667	77.216667
in	new delhi	New Delhi	07	10928270	28.6	77.2
ph	manila	Manila	D9	10443877	14.6042	120.9822
ru	moscow	Moscow	48	10381288	55.752222	37.615556
kr	seoul	Seoul	11	10323448	37.5985	126.9783
br	sao paulo	São Paulo	27	10021437	-23.473293	-46.665803

5. Total population of top K cities
   These top 10 cities with New Delhi in duplicate accounts for ~5% of the dataset population

```
0.05827942402587343
```

6. Total population of each region
   Region 04 has the most population. The provided is the top 20 regions with the most population

region	total_population
04	112249869
07	99634521
02	90271668
05	88272385
16	71422264
08	69386610
09	68742631
11	61477197
01	60187116
06	58191980
15	57366367
10	52912641
13	50762945
19	49067537
23	48056875
30	47565528
25	44541463
12	44192013
40	43122879
27	42540035

7. Sum of population in countries where there is are cities with over 10 million population
   The dataset returned with a total of 41 countries with the requirement

8. Average population of the for cities above certain threshold
   We set the threshold to be at 1 million, the dataset returned with a total of 89 countries

9. Cities that are higher than avg country population
   We set the threshold to be at 1 million, the data set returns with a total of 277 cities.

10. Re-cleaning

    We removed the duplicate values, mainly Dehli and New Dehli, the new dataset now contains 47736 items compared with 47980. A reduction of 244 items. This is the new top 20 cities by population

Country	City	AccentCity	Region	Population	Latitude	Longitude	rounded_value
jp	tokyo	Tokyo	40	31480498	35.685	139.751389	7
cn	shanghai	Shanghai	23	14608512	31.045556	121.399722	7
in	bombay	Bombay	16	12692717	18.975	72.825833	7
pk	karachi	Karachi	05	11627378	24.9056	67.0822	7
in	delhi	Delhi	07	10928270	28.666667	77.216667	7
ph	manila	Manila	D9	10443877	14.6042	120.9822	7
ru	moscow	Moscow	48	10381288	55.752222	37.615556	7
kr	seoul	Seoul	11	10323448	37.5985	126.9783	7
br	sao paulo	São Paulo	27	10021437	-23.473293	-46.665803	7
tr	istanbul	Istanbul	34	9797536	41.018611	28.964722	6
ng	lagos	Lagos	05	8789133	6.453056	3.395833	6
mx	mexico	Mexico	09	8720916	19.434167	-99.138611	6
id	jakarta	Jakarta	04	8540306	-6.174444	106.829444	6
us	new york	New York	NY	8107916	40.7141667	-74.0063889	6
cd	kinshasa	Kinshasa	06	7787832	-4.3	15.3	6
eg	cairo	Cairo	11	7734602	30.05	31.25	6
pe	lima	Lima	15	7646786	-12.05	-77.05	6
cn	peking	Peking	22	7480601	39.928889	116.388333	6
gb	london	London	H9	7421228	51.514125	-0.093689	6
co	bogota	Bogotá	34	7102602	4.649178	-74.062827	6

only showing top 20 rows