Open in Colab
(https://colab.research.google.com/gist/KayvanShah1/2ced1e5560b0d088df28446c28cdb0fb/pyspark_assignment.ipynb)

Installation

In [179…

```
!pip install pyspark
!pip install ipython-autotime
%load_ext autotime
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.9/dist-packages (3.4.0)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.9/dist-packages (from pyspark) (0.10.9.7)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: ipython-autotime in /usr/local/lib/python3.9/dist-packages (0.3.1)
Requirement already satisfied: ipython in /usr/local/lib/python3.9/dist-packages (from ipython-autotime) (7.34.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autotim
e) (0.2.0)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.9/dist-pack
ages (from ipython->ipython-autotime) (3.0.38)
Requirement already satisfied: jedi>=0.16 in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autoti
me) (0.18.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autot
ime) (0.7.5)
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-
autotime) (67.6.1)
Requirement already satisfied: pygments in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autotim
e) (2.14.0)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autot
ime) (4.8.0)
Requirement already satisfied: decorator in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-autotim
e) (4.4.2)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.9/dist-packages (from ipython->ipython-au
totime) (5.7.1)
Requirement already satisfied: matplotlib-inline in /usr/local/lib/python3.9/dist-packages (from ipython->ipython
-autotime) (0.1.6)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /usr/local/lib/python3.9/dist-packages (from jedi>=0.16->ip
ython->ipython-autotime) (0.8.3)
Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.9/dist-packages (from pexpect>4.3->ipyth
on->ipython-autotime) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.9/dist-packages (from prompt-toolkit!=3.0.0,!=3.
0.1,<3.1.0,>=2.0.0->ipython->ipython-autotime) (0.2.6)
The autotime extension is already loaded. To reload it, use:
  %reload_ext autotime
time: 9.19 s (started: 2023-04-21 01:13:34 +00:00)
```

In [180…

```
%%bash
java --version
```

```
openjdk 11.0.18 2023-01-17
OpenJDK Runtime Environment (build 11.0.18+10-post-Ubuntu-0ubuntu120.04.1)
OpenJDK 64-Bit Server VM (build 11.0.18+10-post-Ubuntu-0ubuntu120.04.1, mixed mode, sharing)
time: 99.8 ms (started: 2023-04-21 01:13:43 +00:00)
```

Spark Session

In [181…

```
# Import SparkSession
from pyspark.sql import SparkSession
# Create a Spark Session
spark = SparkSession.builder.master("local[*]").getOrCreate()
# Check Spark Session Information
spark
```

Out[181]: **SparkSession - in-memory**

SparkContext

Spark UI (http://9eafe957c520:4040)

Version	v3.4.0
Master	local[*]
AppName	pyspark-shell

time: 10.9 ms (started: 2023-04-21 01:13:43 +00:00)

In [182…
```
import pyspark.sql.functions as fc
```

time: 295 µs (started: 2023-04-21 01:13:43 +00:00)

Assignment Solution

Load Data to dataframes

In [183…
```
country_df = spark.read.json("country.json")
city_df = spark.read.json("city.json")
country_lang_df = spark.read.json("countrylanguage.json")

aqi_df = spark.read.csv("aqi.csv", header=True, inferSchema=True)
```

time: 1.16 s (started: 2023-04-21 01:13:43 +00:00)

Sample View

In [184…
```
country_df.show(5)
```

```
+-------+----+-----+-------------+------+------+--------------------+--------------------+---------+-------------
-+--------------------+-----------+----------+--------------------+-----------+
|Capital|Code|Code2|    Continent|   GNP|GNPOld|      GovernmentForm|         HeadOfState|IndepYear|LifeExpectanc
y|           LocalName|       Name|Population|              Region|SurfaceArea|
+-------+----+-----+-------------+------+------+--------------------+--------------------+---------+-------------
-+--------------------+-----------+----------+--------------------+-----------+
|    129| ABW|   AW|North America| 828.0| 793.0|Nonmetropolitan T...|             Beatrix|        0|          78.
4|               Aruba|      Aruba|    103000|           Caribbean|      193.0|
|      1| AFG|   AF|         Asia|5976.0|   0.0|     Islamic Emirate|       Mohammad Omar|     1919|          45.
9|Afganistan/Afqane...|Afghanistan|  22720000|Southern and Cent...|   652090.0|
|     56| AGO|   AO|       Africa|6648.0|7984.0|            Republic|JosÃ© Eduardo dos...|     1975|          38.
3|              Angola|     Angola|  12878000|      Central Africa|  1246700.0|
|     62| AIA|   AI|North America|  63.2|   0.0|Dependent Territo...|        Elisabeth II|        0|          76.
1|            Anguilla|   Anguilla|      8000|           Caribbean|       96.0|
|     34| ALB|   AL|       Europe|3205.0|2500.0|            Republic|      Rexhep Mejdani|     1912|          71.
6|          ShqipÃ«ria|    Albania|   3401200|     Southern Europe|    28748.0|
+-------+----+-----+-------------+------+------+--------------------+--------------------+---------+-------------
-+--------------------+-----------+----------+--------------------+-----------+
only showing top 5 rows

time: 165 ms (started: 2023-04-21 01:13:44 +00:00)
```

In [185…
```
city_df.show(5)
```

```
+-----------+-------------+---+-------------+----------+
|CountryCode|     District| ID|         Name|Population|
+-----------+-------------+---+-------------+----------+
|        AFG|        Kabol|  1|        Kabul|   1780000|
|        AFG|     Qandahar|  2|     Qandahar|    237500|
|        AFG|        Herat|  3|        Herat|    186800|
|        AFG|        Balkh|  4|Mazar-e-Sharif|    127800|
|        NLD|Noord-Holland|  5|    Amsterdam|    731200|
+-----------+-------------+---+-------------+----------+
only showing top 5 rows

time: 200 ms (started: 2023-04-21 01:13:44 +00:00)
```

In [186…
```
country_lang_df.show(5)
```

```
+-----------+----------+----------+----------+
|CountryCode|IsOfficial|  Language|Percentage|
+-----------+----------+----------+----------+
|        ABW|         T|     Dutch|       5.3|
|        ABW|         F|   English|       9.5|
|        ABW|         F|Papiamento|      76.7|
|        ABW|         F|   Spanish|       7.4|
|        AFG|         F|   Balochi|       0.9|
+-----------+----------+----------+----------+
only showing top 5 rows

time: 116 ms (started: 2023-04-21 01:13:45 +00:00)
```

In [187…

```
aqi_df.show(5)
```

```
+----------+---------+-------------------+-----+
|      date|  country|             status|value|
+----------+---------+-------------------+-----+
|2022-07-21|  Albania|               Good|   14|
|2022-07-21|  Algeria|           Moderate|   65|
|2022-07-21|  Andorra|           Moderate|   55|
|2022-07-21|   Angola|Unhealthy for Sen...|  113|
|2022-07-21|Argentina|           Moderate|   63|
+----------+---------+-------------------+-----+
only showing top 5 rows

time: 79.2 ms (started: 2023-04-21 01:13:45 +00:00)
```

Question 2

a.i

In [188…

```
joined_df = aqi_df.join(country_df, aqi_df.country == country_df.Name)
joined_df.show(10)
```

```
+----------+----------+--------------------+-----+-------+----+-----+-------------+--------+--------+------------
--------+--------------------+---------+--------------+--------------------+----------+----------+--------------
-----+-----------+
|      date|   country|              status|value|Capital|Code|Code2|    Continent|     GNP|  GNPOld|      Govern
mentForm|         HeadOfState|IndepYear|LifeExpectancy|           LocalName|      Name|Population|              R
egion|SurfaceArea|
+----------+----------+--------------------+-----+-------+----+-----+-------------+--------+--------+------------
--------+--------------------+---------+--------------+--------------------+----------+----------+--------------
-----+-----------+
|2022-07-21|   Albania|                Good|   14|     34| ALB|   AL|       Europe|  3205.0|  2500.0|
Republic|      Rexhep Mejdani|     1912|          71.6|          ShqipÃ«ria|   Albania|   3401200|     Southern E
urope|    28748.0|
|2022-07-21|   Algeria|            Moderate|   65|     35| DZA|   DZ|       Africa| 49982.0| 46966.0|
Republic|Abdelaziz Bouteflika|     1962|          69.7|Al-JazaÂ’ir/AlgÃ©rie|   Algeria|  31471000|     Northern A
frica|  2381741.0|
|2022-07-21|   Andorra|            Moderate|   55|     55| AND|   AD|       Europe|  1630.0|     0.0|Parliamentar
y Cop...|                    |     1278|          83.5|             Andorra|   Andorra|     78000|     Southern E
urope|      468.0|
|2022-07-21|    Angola|Unhealthy for Sen...|  113|     56| AGO|   AO|       Africa|  6648.0|  7984.0|
Republic|JosÃ© Eduardo dos...|     1975|          38.3|              Angola|    Angola|  12878000|      Central A
frica|  1246700.0|
|2022-07-21| Argentina|            Moderate|   63|     69| ARG|   AR|South America|340238.0|323310.0|    Federal
Republic| Fernando de la RÃºa|     1816|          75.1|           Argentina| Argentina|  37032000|       South Am
erica|  2780400.0|
|2022-07-21|   Armenia|            Moderate|   76|    126| ARM|   AM|         Asia|  1813.0|  1627.0|
Republic|   Robert KotÂšarjan|     1991|          66.4|            Hajastan|   Armenia|   3520000|         Middle
East|    29800.0|
|2022-07-21| Australia|            Moderate|   56|    135| AUS|   AU|      Oceania|351182.0|392911.0|Constitution
al Mo...|        Elisabeth II|     1901|          79.8|           Australia| Australia|  18886000|Australia and N
ew...|  7741220.0|
|2022-07-21|   Austria|                Good|   45|   1523| AUT|   AT|       Europe|211860.0|206025.0|    Federal
Republic|      Thomas Klestil|     1918|          77.7|         Ã–sterreich|   Austria|   8091800|      Western E
urope|    83859.0|
|2022-07-21|Azerbaijan|                Good|   12|    144| AZE|   AZ|         Asia|  4127.0|  4100.0|    Federal
Republic|     HeydÃ¤r Ã„liyev|     1991|          62.9|         AzÃ¤rbaycan|Azerbaijan|   7734000|         Middle
East|    86600.0|
|2022-07-21|   Bahrain|           Unhealthy|  165|    149| BHR|   BH|         Asia|  6366.0|  6097.0|  Monarchy
(Emirate)|Hamad ibn Isa al-...|     1971|          73.0|          Al-Bahrayn|   Bahrain|    617000|         Middl
e East|      694.0|
+----------+----------+--------------------+-----+-------+----+-----+-------------+--------+--------+------------
--------+--------------------+---------+--------------+--------------------+----------+----------+--------------
-----+-----------+
only showing top 10 rows

time: 392 ms (started: 2023-04-21 01:13:45 +00:00)
```

In [189…

```
result_df = joined_df.select("Name").distinct().sort("Name")
result_df.show(truncate=False)
```

```
+----------------------+
|Name                  |
+----------------------+
|Albania               |
|Algeria               |
|Andorra               |
|Angola                |
|Argentina             |
|Armenia               |
|Australia             |
|Austria               |
|Azerbaijan            |
|Bahrain               |
|Bangladesh            |
|Belarus               |
|Belgium               |
|Belize                |
|Bermuda               |
|Bolivia               |
|Bosnia and Herzegovina|
|Brazil                |
|Brunei                |
|Bulgaria              |
+----------------------+
only showing top 20 rows

time: 519 ms (started: 2023-04-21 01:13:45 +00:00)
```

a.ii

In [190…

```
country_set = country_df.select("Name").distinct()
aqi_set = aqi_df.select("country").distinct()

result_set = country_set.intersect(aqi_set).sort("Name")

result_set.show(truncate=False)
```

```
+----------------------+
|Name                  |
+----------------------+
|Albania               |
|Algeria               |
|Andorra               |
|Angola                |
|Argentina             |
|Armenia               |
|Australia             |
|Austria               |
|Azerbaijan            |
|Bahrain               |
|Bangladesh            |
|Belarus               |
|Belgium               |
|Belize                |
|Bermuda               |
|Bolivia               |
|Bosnia and Herzegovina|
|Brazil                |
|Brunei                |
|Bulgaria              |
+----------------------+
only showing top 20 rows

time: 752 ms (started: 2023-04-21 01:13:46 +00:00)
```

b.i

In [191…

```
joined_df = aqi_df.join(country_df, aqi_df.country == country_df.Name, "left_anti")

result_df = joined_df.select("country").distinct().sort("country")
result_df.show(truncate=False)
```

```
+-------------------------------------------------+
|country                                          |
+-------------------------------------------------+
|Ivory Coast                                      |
|Jersey                                           |
|Kazakhstan                                       |
|Kosovo                                           |
|Montenegro                                       |
|Palestinian Territory                            |
|Reunion                                          |
|Russia                                           |
|Serbia                                           |
|United Kingdom of Great Britain and Northern Ireland|
|United States of America                         |
|Vatican                                          |
+-------------------------------------------------+

time: 443 ms (started: 2023-04-21 01:13:47 +00:00)
```

b.ii

In [192…

```
country_set = country_df.select("Name").distinct()
aqi_set = aqi_df.select("country").distinct()

result_set = aqi_set.subtract(country_set).sort("country")

result_set.show(truncate=False)
```

```
+-------------------------------------------------+
|country                                          |
+-------------------------------------------------+
|Ivory Coast                                      |
|Jersey                                           |
|Kazakhstan                                       |
|Kosovo                                           |
|Montenegro                                       |
|Palestinian Territory                            |
|Reunion                                          |
|Russia                                           |
|Serbia                                           |
|United Kingdom of Great Britain and Northern Ireland|
|United States of America                         |
|Vatican                                          |
+-------------------------------------------------+

time: 528 ms (started: 2023-04-21 01:13:47 +00:00)
```

c.i

In [193…

```
joined_df = country_df.join(aqi_df, aqi_df.country == country_df.Name, "left_anti")

result_df = joined_df.select("Name").distinct().sort("Name")
result_df.show(truncate=False)
```

```
+-----------------------------------+
|Name                               |
+-----------------------------------+
|Afghanistan                        |
|American Samoa                     |
|Anguilla                           |
|Antarctica                         |
|Antigua and Barbuda                |
|Aruba                              |
|Bahamas                            |
|Barbados                           |
|Benin                              |
|Bhutan                             |
|Botswana                           |
|Bouvet Island                      |
|British Indian Ocean Territory     |
|Burundi                            |
|Cameroon                           |
|Christmas Island                   |
|Cocos (Keeling) Islands            |
|Comoros                            |
|Congo                              |
|Congo, The Democratic Republic of the|
+-----------------------------------+
only showing top 20 rows

time: 638 ms (started: 2023-04-21 01:13:48 +00:00)
```

c.ii

In [194…

```
country_set = country_df.select("Name").distinct()
aqi_set = aqi_df.select("country").distinct()

result_set = country_set.subtract(aqi_set).sort("Name")

result_set.show(truncate=False)
```

```
+-----------------------------------+
|Name                               |
+-----------------------------------+
|Afghanistan                        |
|American Samoa                     |
|Anguilla                           |
|Antarctica                         |
|Antigua and Barbuda                |
|Aruba                              |
|Bahamas                            |
|Barbados                           |
|Benin                              |
|Bhutan                             |
|Botswana                           |
|Bouvet Island                      |
|British Indian Ocean Territory     |
|Burundi                            |
|Cameroon                           |
|Christmas Island                   |
|Cocos (Keeling) Islands            |
|Comoros                            |
|Congo                              |
|Congo, The Democratic Republic of the|
+-----------------------------------+
only showing top 20 rows

time: 642 ms (started: 2023-04-21 01:13:48 +00:00)
```

d

In [195…

```
aqi_df = aqi_df.withColumn('year', fc.year('date'))
aqi_df = aqi_df.withColumn('month', fc.month('date'))

aqi_aug_2022_df = aqi_df.filter(
    (aqi_df['year'] == 2022) & (aqi_df['month'] == 8)
)

aqi_aug_2022_df = aqi_aug_2022_df.groupBy('status')

aqi_aug_2022_df = aqi_aug_2022_df.agg(fc.avg('value').alias('avg_value'))

aqi_aug_2022_df = aqi_aug_2022_df.filter(fc.count('value') >= 100)
aqi_aug_2022_df = aqi_aug_2022_df.sort("avg_value")
aqi_aug_2022_df.show(truncate=False)
```

```
+-----------------------------+------------------+
|status                       |avg_value         |
+-----------------------------+------------------+
|Good                         |27.929097605893187|
|Moderate                     |71.38070175438597 |
|Unhealthy for Sensitive Groups|122.953125        |
|Unhealthy                    |167.9704433497537 |
+-----------------------------+------------------+

time: 586 ms (started: 2023-04-21 01:13:49 +00:00)
```

Question 3

Convert Spark Dataframe to RDD

In [196…

```
country_rdd = country_df.rdd
city_rdd = city_df.rdd
country_lang_rdd = country_lang_df.rdd

aqi_rdd = aqi_df.rdd
```

```
time: 188 ms (started: 2023-04-21 01:13:50 +00:00)
```

Sample View

In [197…

```
aqi_rdd.take(5)
```

Out[197]:
```
[Row(date=datetime.date(2022, 7, 21), country='Albania', status='Good', value=14, year=2022, month=7),
 Row(date=datetime.date(2022, 7, 21), country='Algeria', status='Moderate', value=65, year=2022, month=7),
 Row(date=datetime.date(2022, 7, 21), country='Andorra', status='Moderate', value=55, year=2022, month=7),
 Row(date=datetime.date(2022, 7, 21), country='Angola', status='Unhealthy for Sensitive Groups', value=113, year=
2022, month=7),
 Row(date=datetime.date(2022, 7, 21), country='Argentina', status='Moderate', value=63, year=2022, month=7)]
time: 174 ms (started: 2023-04-21 01:13:50 +00:00)
```

In [198…
```
country_rdd.take(5)
```

Out[198]:
```
[Row(Capital=129, Code='ABW', Code2='AW', Continent='North America', GNP=828.0, GNPOld=793.0, GovernmentForm='Non
metropolitan Territory of The Netherlands', HeadOfState='Beatrix', IndepYear=0, LifeExpectancy=78.4, LocalName='A
ruba', Name='Aruba', Population=103000, Region='Caribbean', SurfaceArea=193.0),
 Row(Capital=1, Code='AFG', Code2='AF', Continent='Asia', GNP=5976.0, GNPOld=0.0, GovernmentForm='Islamic Emirat
e', HeadOfState='Mohammad Omar', IndepYear=1919, LifeExpectancy=45.9, LocalName='Afganistan/Afqanestan', Name='Af
ghanistan', Population=22720000, Region='Southern and Central Asia', SurfaceArea=652090.0),
 Row(Capital=56, Code='AGO', Code2='AO', Continent='Africa', GNP=6648.0, GNPOld=7984.0, GovernmentForm='Republi
c', HeadOfState='JosÃ© Eduardo dos Santos', IndepYear=1975, LifeExpectancy=38.3, LocalName='Angola', Name='Angol
a', Population=12878000, Region='Central Africa', SurfaceArea=1246700.0),
 Row(Capital=62, Code='AIA', Code2='AI', Continent='North America', GNP=63.2, GNPOld=0.0, GovernmentForm='Depende
nt Territory of the UK', HeadOfState='Elisabeth II', IndepYear=0, LifeExpectancy=76.1, LocalName='Anguilla', Name
='Anguilla', Population=8000, Region='Caribbean', SurfaceArea=96.0),
 Row(Capital=34, Code='ALB', Code2='AL', Continent='Europe', GNP=3205.0, GNPOld=2500.0, GovernmentForm='Republi
c', HeadOfState='Rexhep Mejdani', IndepYear=1912, LifeExpectancy=71.6, LocalName='ShqipÃ«ria', Name='Albania', Po
pulation=3401200, Region='Southern Europe', SurfaceArea=28748.0)]
time: 92.3 ms (started: 2023-04-21 01:13:50 +00:00)
```

a.i

In [199…
```
country_key_rdd = country_rdd.map(lambda x: (x["Name"], x)).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: (x["country"], x)).distinct()

country_aqi_join_rdd = country_key_rdd.join(aqi_key_rdd).map(lambda x: x[0]).distinct()
country_aqi_join_rdd = country_aqi_join_rdd.sortBy(lambda x: x)

country_aqi_join_rdd.collect()
```

Out[199]: ['Albania',
'Algeria',
'Andorra',
'Angola',
'Argentina',
'Armenia',
'Australia',
'Austria',
'Azerbaijan',
'Bahrain',
'Bangladesh',
'Belarus',
'Belgium',
'Belize',
'Bermuda',
'Bolivia',
'Bosnia and Herzegovina',
'Brazil',
'Brunei',
'Bulgaria',
'Burkina Faso',
'Cambodia',
'Canada',
'Cape Verde',
'Cayman Islands',
'Central African Republic',
'Chad',
'Chile',
'China',
'Colombia',
'Costa Rica',
'Croatia',
'Cyprus',
'Czech Republic',
'Denmark',
'Dominican Republic',
'Ecuador',
'Egypt',
'El Salvador',
'Estonia',
'Ethiopia',
'Finland',
'France',
'French Guiana',
'Gabon',
'Gambia',
'Georgia',
'Germany',
'Ghana',
'Gibraltar',
'Greece',
'Grenada',
'Guadeloupe',
'Guam',
'Guatemala',
'Honduras',
'Hong Kong',
'Hungary',
'Iceland',
'India',
'Indonesia',
'Iran',
'Iraq',
'Ireland',
'Israel',
'Italy',
'Japan',
'Jordan',
'Kenya',
'Kuwait',
'Kyrgyzstan',
'Laos',
'Latvia',
'Lebanon',
'Liberia',
'Liechtenstein',
'Lithuania',

```
 'Luxembourg',
 'Macao',
 'Macedonia',
 'Madagascar',
 'Malaysia',
 'Malta',
 'Martinique',
 'Mexico',
 'Moldova',
 'Monaco',
 'Mongolia',
 'Myanmar',
 'Nepal',
 'Netherlands',
 'New Caledonia',
 'New Zealand',
 'Nigeria',
 'Norway',
 'Pakistan',
 'Peru',
 'Philippines',
 'Poland',
 'Portugal',
 'Puerto Rico',
 'Qatar',
 'Romania',
 'San Marino',
 'Saudi Arabia',
 'Senegal',
 'Singapore',
 'Slovakia',
 'Slovenia',
 'South Africa',
 'South Korea',
 'Spain',
 'Sri Lanka',
 'Sudan',
 'Sweden',
 'Switzerland',
 'Taiwan',
 'Tajikistan',
 'Thailand',
 'Togo',
 'Trinidad and Tobago',
 'Turkey',
 'Turkmenistan',
 'Uganda',
 'Ukraine',
 'United Arab Emirates',
 'Uzbekistan',
 'Venezuela',
 'Vietnam',
 'Zambia']
time: 2.31 s (started: 2023-04-21 01:13:50 +00:00)
```

a.ii

In [200…

```
country_key_rdd = country_rdd.map(lambda x: x["Name"]).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: x["country"]).distinct()

country_aqi_join_rdd = country_key_rdd.intersection(aqi_key_rdd).distinct()
country_aqi_join_rdd = country_aqi_join_rdd.sortBy(lambda x: x)

country_aqi_join_rdd.collect()
```

```
Out[200]: ['Albania',
           'Algeria',
           'Andorra',
           'Angola',
           'Argentina',
           'Armenia',
           'Australia',
           'Austria',
           'Azerbaijan',
           'Bahrain',
           'Bangladesh',
           'Belarus',
           'Belgium',
           'Belize',
           'Bermuda',
           'Bolivia',
           'Bosnia and Herzegovina',
           'Brazil',
           'Brunei',
           'Bulgaria',
           'Burkina Faso',
           'Cambodia',
           'Canada',
           'Cape Verde',
           'Cayman Islands',
           'Central African Republic',
           'Chad',
           'Chile',
           'China',
           'Colombia',
           'Costa Rica',
           'Croatia',
           'Cyprus',
           'Czech Republic',
           'Denmark',
           'Dominican Republic',
           'Ecuador',
           'Egypt',
           'El Salvador',
           'Estonia',
           'Ethiopia',
           'Finland',
           'France',
           'French Guiana',
           'Gabon',
           'Gambia',
           'Georgia',
           'Germany',
           'Ghana',
           'Gibraltar',
           'Greece',
           'Grenada',
           'Guadeloupe',
           'Guam',
           'Guatemala',
           'Honduras',
           'Hong Kong',
           'Hungary',
           'Iceland',
           'India',
           'Indonesia',
           'Iran',
           'Iraq',
           'Ireland',
           'Israel',
           'Italy',
           'Japan',
           'Jordan',
           'Kenya',
           'Kuwait',
           'Kyrgyzstan',
           'Laos',
           'Latvia',
           'Lebanon',
           'Liberia',
           'Liechtenstein',
           'Lithuania',
```

```
 'Luxembourg',
 'Macao',
 'Macedonia',
 'Madagascar',
 'Malaysia',
 'Malta',
 'Martinique',
 'Mexico',
 'Moldova',
 'Monaco',
 'Mongolia',
 'Myanmar',
 'Nepal',
 'Netherlands',
 'New Caledonia',
 'New Zealand',
 'Nigeria',
 'Norway',
 'Pakistan',
 'Peru',
 'Philippines',
 'Poland',
 'Portugal',
 'Puerto Rico',
 'Qatar',
 'Romania',
 'San Marino',
 'Saudi Arabia',
 'Senegal',
 'Singapore',
 'Slovakia',
 'Slovenia',
 'South Africa',
 'South Korea',
 'Spain',
 'Sri Lanka',
 'Sudan',
 'Sweden',
 'Switzerland',
 'Taiwan',
 'Tajikistan',
 'Thailand',
 'Togo',
 'Trinidad and Tobago',
 'Turkey',
 'Turkmenistan',
 'Uganda',
 'Ukraine',
 'United Arab Emirates',
 'Uzbekistan',
 'Venezuela',
 'Vietnam',
 'Zambia']
time: 1.29 s (started: 2023-04-21 01:13:52 +00:00)
```

b.i

In [224…

```
country_key_rdd = country_rdd.map(lambda x: (x["Name"], x)).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: (x["country"], x)).distinct()

aqi_country_join_rdd = aqi_key_rdd.leftOuterJoin(country_key_rdd).filter(lambda x: x[1][1] is None)
aqi_country_join_rdd = aqi_country_join_rdd.map(lambda x: x[0]).distinct()
aqi_country_join_rdd = aqi_country_join_rdd.sortBy(lambda x: x)

aqi_country_join_rdd.collect()
```

Out[224]:
```
['Ivory Coast',
 'Jersey',
 'Kazakhstan',
 'Kosovo',
 'Montenegro',
 'Palestinian Territory',
 'Reunion',
 'Russia',
 'Serbia',
 'United Kingdom of Great Britain and Northern Ireland',
 'United States of America',
 'Vatican']
time: 1.94 s (started: 2023-04-21 01:33:39 +00:00)
```

b.ii

In [204…
```
country_key_rdd = country_rdd.map(lambda x: x["Name"]).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: x["country"]).distinct()

country_aqi_set_rdd = aqi_key_rdd.subtract(country_key_rdd).distinct()
country_aqi_set_rdd = country_aqi_set_rdd.sortBy(lambda x: x)

country_aqi_set_rdd.collect()
```

Out[204]:
```
['Ivory Coast',
 'Jersey',
 'Kazakhstan',
 'Kosovo',
 'Montenegro',
 'Palestinian Territory',
 'Reunion',
 'Russia',
 'Serbia',
 'United Kingdom of Great Britain and Northern Ireland',
 'United States of America',
 'Vatican']
time: 819 ms (started: 2023-04-21 01:14:55 +00:00)
```

c.i

In [227…
```
country_key_rdd = country_rdd.map(lambda x: (x["Name"], x)).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: (x["country"], x)).distinct()

aqi_country_join_rdd = aqi_key_rdd.rightOuterJoin(country_key_rdd).filter(lambda x: x[1][0] is None)
aqi_country_join_rdd = aqi_country_join_rdd.map(lambda x: x[0]).distinct()
aqi_country_join_rdd = aqi_country_join_rdd.sortBy(lambda x: x)

aqi_country_join_rdd.collect()
```

Out[227]:
```
['Afghanistan',
 'American Samoa',
 'Anguilla',
 'Antarctica',
 'Antigua and Barbuda',
 'Aruba',
 'Bahamas',
 'Barbados',
 'Benin',
 'Bhutan',
 'Botswana',
 'Bouvet Island',
 'British Indian Ocean Territory',
 'Burundi',
 'Cameroon',
 'Christmas Island',
 'Cocos (Keeling) Islands',
 'Comoros',
 'Congo',
 'Congo, The Democratic Republic of the',
 'Cook Islands',
 'Cuba',
 'CÃ´te dÂ’Ivoire',
 'Djibouti',
 'Dominica',
 'East Timor',
 'Equatorial Guinea',
 'Eritrea',
 'Falkland Islands',
 'Faroe Islands',
 'Fiji Islands',
 'French Polynesia',
 'French Southern territories',
 'Greenland',
 'Guinea',
 'Guinea-Bissau',
 'Guyana',
 'Haiti',
 'Heard Island and McDonald Islands',
 'Holy See (Vatican City State)',
 'Jamaica',
 'Kazakstan',
 'Kiribati',
 'Lesotho',
 'Libyan Arab Jamahiriya',
 'Malawi',
 'Maldives',
 'Mali',
 'Marshall Islands',
 'Mauritania',
 'Mauritius',
 'Mayotte',
 'Micronesia, Federated States of',
 'Montserrat',
 'Morocco',
 'Mozambique',
 'Namibia',
 'Nauru',
 'Netherlands Antilles',
 'Nicaragua',
 'Niger',
 'Niue',
 'Norfolk Island',
 'North Korea',
 'Northern Mariana Islands',
 'Oman',
 'Palau',
 'Palestine',
 'Panama',
 'Papua New Guinea',
 'Paraguay',
 'Pitcairn',
 'Russian Federation',
 'Rwanda',
 'RÃ©union',
 'Saint Helena',
 'Saint Kitts and Nevis',
```

```
 'Saint Lucia',
 'Saint Pierre and Miquelon',
 'Saint Vincent and the Grenadines',
 'Samoa',
 'Sao Tome and Principe',
 'Seychelles',
 'Sierra Leone',
 'Solomon Islands',
 'Somalia',
 'South Georgia and the South Sandwich Islands',
 'Suriname',
 'Svalbard and Jan Mayen',
 'Swaziland',
 'Syria',
 'Tanzania',
 'Tokelau',
 'Tonga',
 'Tunisia',
 'Turks and Caicos Islands',
 'Tuvalu',
 'United Kingdom',
 'United States',
 'United States Minor Outlying Islands',
 'Uruguay',
 'Vanuatu',
 'Virgin Islands, British',
 'Virgin Islands, U.S.',
 'Wallis and Futuna',
 'Western Sahara',
 'Yemen',
 'Yugoslavia',
 'Zimbabwe']
time: 1.48 s (started: 2023-04-21 01:42:33 +00:00)
```

c.ii

In [206…

```
country_key_rdd = country_rdd.map(lambda x: x["Name"]).distinct()

aqi_key_rdd = aqi_rdd.map(lambda x: x["country"]).distinct()

country_aqi_set_rdd = country_key_rdd.subtract(aqi_key_rdd).distinct()
country_aqi_set_rdd = country_aqi_set_rdd.sortBy(lambda x: x)

country_aqi_set_rdd.collect()
```

Out[206]:
```
['Afghanistan',
 'American Samoa',
 'Anguilla',
 'Antarctica',
 'Antigua and Barbuda',
 'Aruba',
 'Bahamas',
 'Barbados',
 'Benin',
 'Bhutan',
 'Botswana',
 'Bouvet Island',
 'British Indian Ocean Territory',
 'Burundi',
 'Cameroon',
 'Christmas Island',
 'Cocos (Keeling) Islands',
 'Comoros',
 'Congo',
 'Congo, The Democratic Republic of the',
 'Cook Islands',
 'Cuba',
 'CÃ´te dÂ’Ivoire',
 'Djibouti',
 'Dominica',
 'East Timor',
 'Equatorial Guinea',
 'Eritrea',
 'Falkland Islands',
 'Faroe Islands',
 'Fiji Islands',
 'French Polynesia',
 'French Southern territories',
 'Greenland',
 'Guinea',
 'Guinea-Bissau',
 'Guyana',
 'Haiti',
 'Heard Island and McDonald Islands',
 'Holy See (Vatican City State)',
 'Jamaica',
 'Kazakstan',
 'Kiribati',
 'Lesotho',
 'Libyan Arab Jamahiriya',
 'Malawi',
 'Maldives',
 'Mali',
 'Marshall Islands',
 'Mauritania',
 'Mauritius',
 'Mayotte',
 'Micronesia, Federated States of',
 'Montserrat',
 'Morocco',
 'Mozambique',
 'Namibia',
 'Nauru',
 'Netherlands Antilles',
 'Nicaragua',
 'Niger',
 'Niue',
 'Norfolk Island',
 'North Korea',
 'Northern Mariana Islands',
 'Oman',
 'Palau',
 'Palestine',
 'Panama',
 'Papua New Guinea',
 'Paraguay',
 'Pitcairn',
 'Russian Federation',
 'Rwanda',
 'RÃ©union',
 'Saint Helena',
 'Saint Kitts and Nevis',
```

```
 'Saint Lucia',
 'Saint Pierre and Miquelon',
 'Saint Vincent and the Grenadines',
 'Samoa',
 'Sao Tome and Principe',
 'Seychelles',
 'Sierra Leone',
 'Solomon Islands',
 'Somalia',
 'South Georgia and the South Sandwich Islands',
 'Suriname',
 'Svalbard and Jan Mayen',
 'Swaziland',
 'Syria',
 'Tanzania',
 'Tokelau',
 'Tonga',
 'Tunisia',
 'Turks and Caicos Islands',
 'Tuvalu',
 'United Kingdom',
 'United States',
 'United States Minor Outlying Islands',
 'Uruguay',
 'Vanuatu',
 'Virgin Islands, British',
 'Virgin Islands, U.S.',
 'Wallis and Futuna',
 'Western Sahara',
 'Yemen',
 'Yugoslavia',
 'Zimbabwe']
time: 1.41 s (started: 2023-04-21 01:15:26 +00:00)
```

d

In [238…

```
aqi_aug_2022_rdd = aqi_rdd.filter(lambda x: x['date'].year==2022 and x['date'].month==8)

aqi_aug_2022_rdd = aqi_aug_2022_rdd.map(lambda x: (x["status"], x["value"]))
aqi_aug_2022_rdd = aqi_aug_2022_rdd.aggregateByKey(
    (0,0),
    lambda total, count: (total[0] + count, total[1] + 1),
    lambda total1, total2: (total1[0] + total2[0], total1[1] + total2[1])
)

aqi_aug_2022_rdd = aqi_aug_2022_rdd.filter(lambda x: x[1][1] >= 100).mapValues(lambda x: x[0] / x[1])

aqi_aug_2022_rdd.collect()
```

Out[238]:

```
[('Good', 27.929097605893187),
 ('Moderate', 71.38070175438597),
 ('Unhealthy for Sensitive Groups', 122.953125),
 ('Unhealthy', 167.9704433497537)]
time: 296 ms (started: 2023-04-21 02:02:33 +00:00)
```