Homework 5

Question 3

```
import pyspark.sql.functions as fc

# Load Data to dataframes
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)

# Convert Spark Dataframe to RDD
country_rdd = country_df.rdd
city_rdd = city_df.rdd
country_lang_rdd = country_lang_df.rdd

aqi_rdd = aqi_df.rdd
```

A

I

```
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()
```

```
['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']
```

||

```
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()
```

```
['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']
```

B

I

```
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()
```

```
['Ivory Coast',
 'Jersey',
 'Kazakhstan',
 'Kosovo',
 'Montenegro',
 'Palestinian Territory',
 'Reunion',
 'Russia',
 'Serbia',
 'United Kingdom of Great Britain and Northern Ireland',
 'United States of America',
 'Vatican']
```

II

```
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()
```

```
['Ivory Coast',
 'Jersey',
 'Kazakhstan',
 'Kosovo',
 'Montenegro',
 'Palestinian Territory',
 'Reunion',
 'Russia',
 'Serbia',
 'United Kingdom of Great Britain and Northern Ireland',
 'United States of America',
 'Vatican']
```

C

I

```
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()
```

```
['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']
```

||

```
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()
```

```
['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']
```

D

```
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()
```

```
[('Good', 27.929097605893187),
 ('Moderate', 71.38070175438597),
 ('Unhealthy for Sensitive Groups', 122.953125),
 ('Unhealthy', 167.9704433497537)]
```