Homework 5

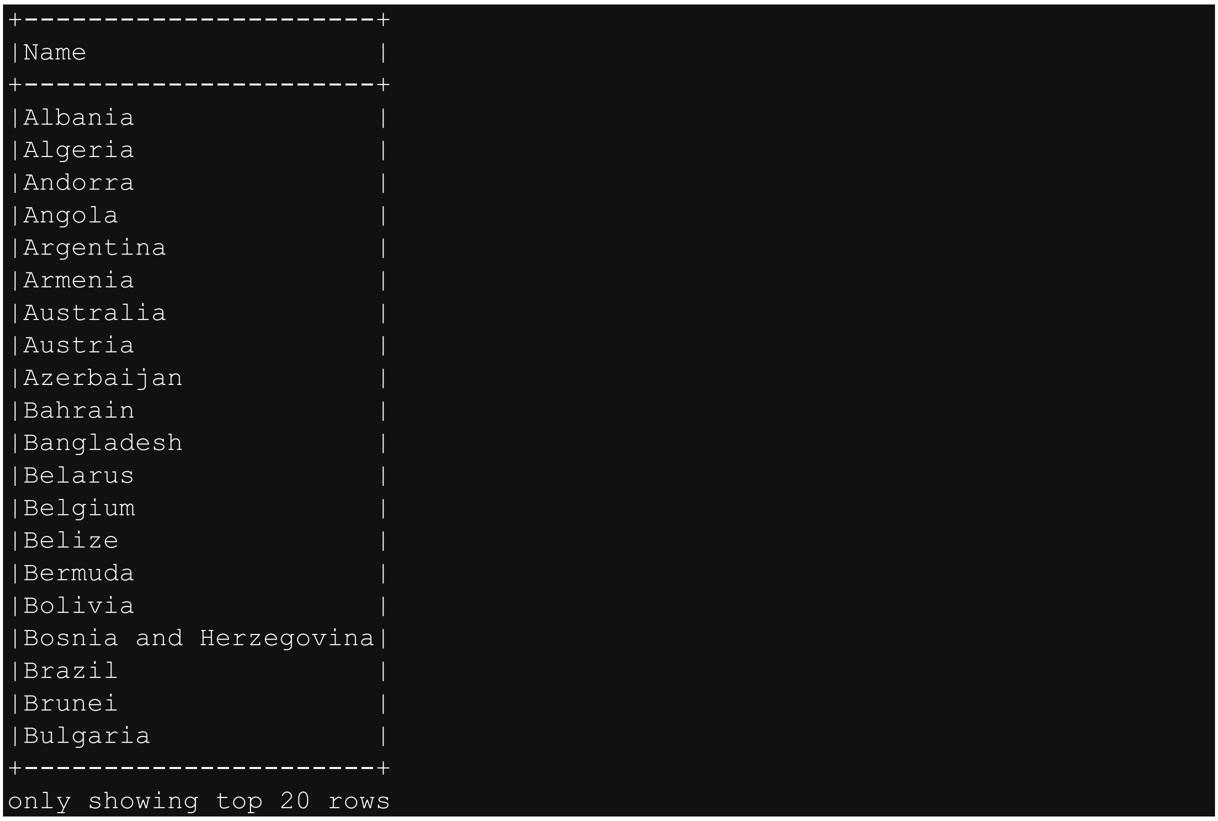
Question 2

```
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)
```

A

I

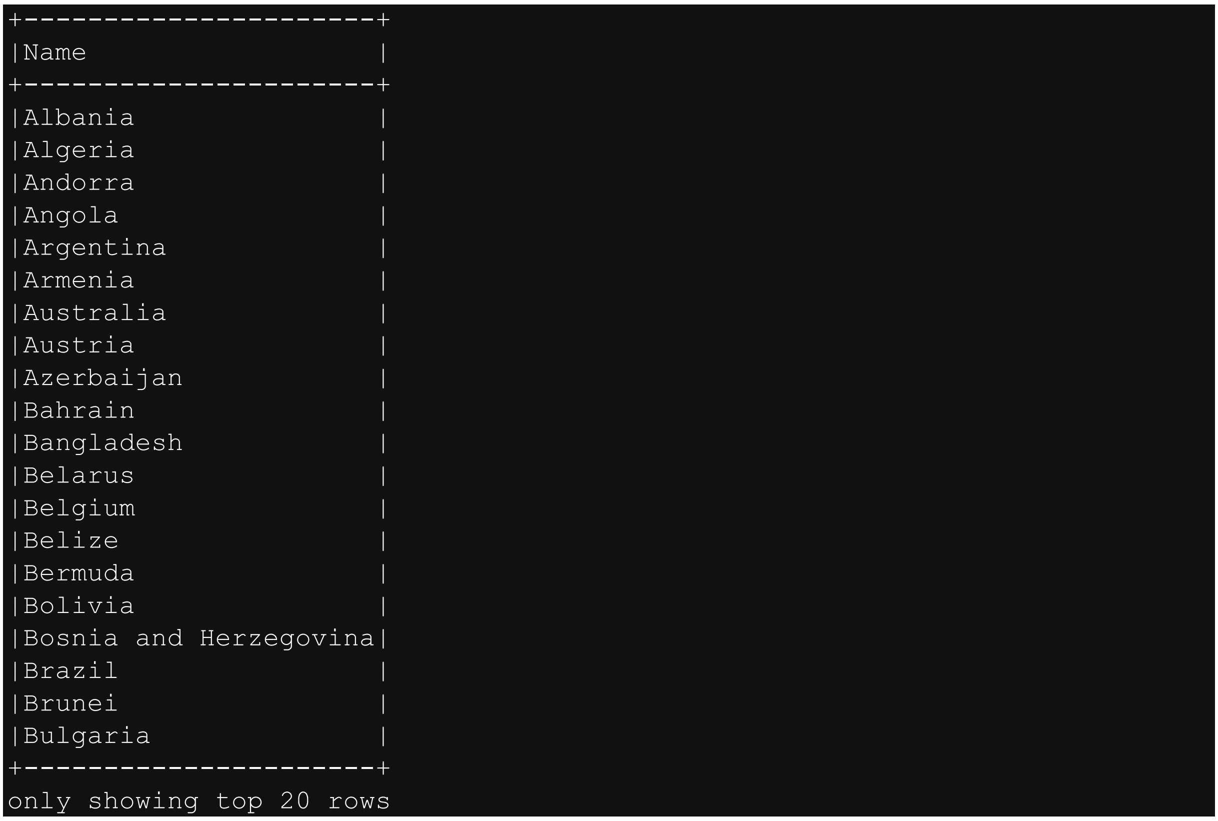
```
joined_df = aqi_df.join(country_df, aqi_df.country == country_df.Name)
result_df = joined_df.select("Name").distinct().sort("Name")
result_df.show(truncate=False)
```



```
+----------------------+
|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
```

II

```
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)
```



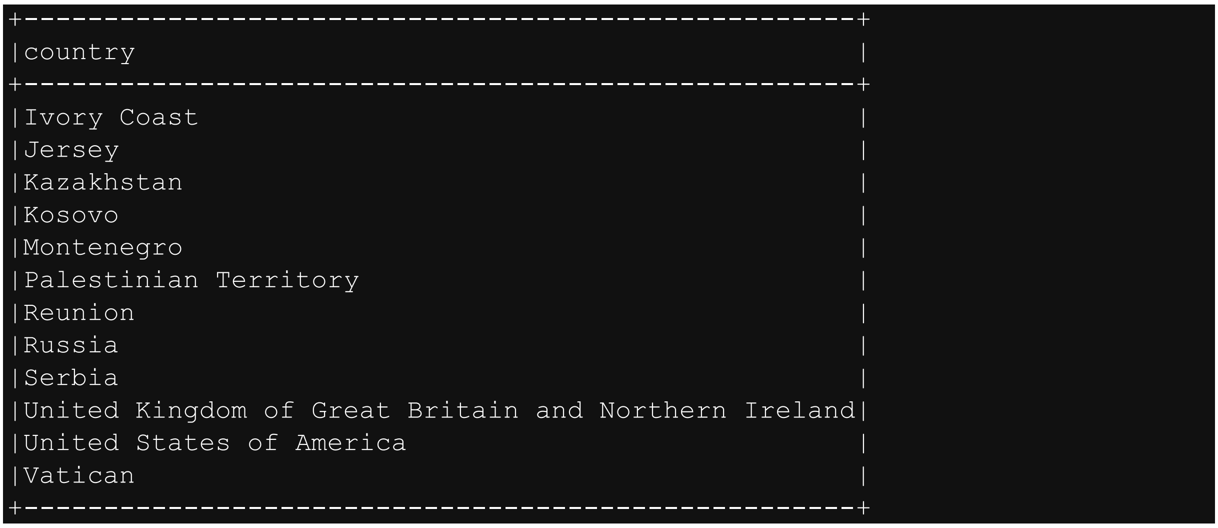
```
+----------------------+
|Name                  |
+----------------------+
|Albania               |
|Algeria               |
|Andorra               |
|Angola                |
|Argentina             |
|Armenia               |
|Australia             |
|Austria               |
|Azerbaijan          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

B

I

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)

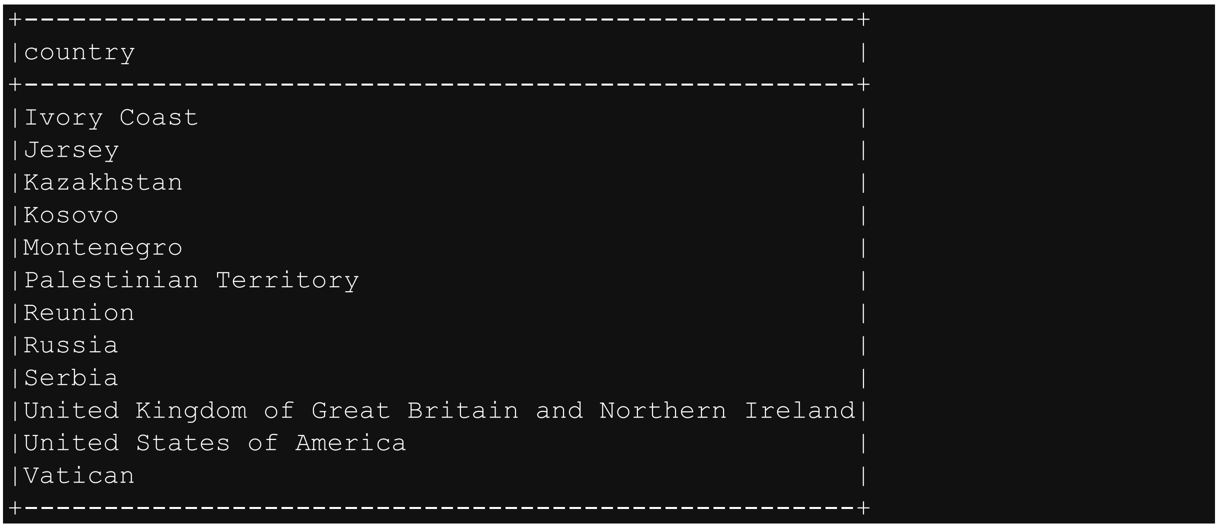


```
+----------------------------------------------------+
|country                                             |
+----------------------------------------------------+
|Ivory Coast                                         |
|Jersey                                              |
|Kazakhstan          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

II

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)



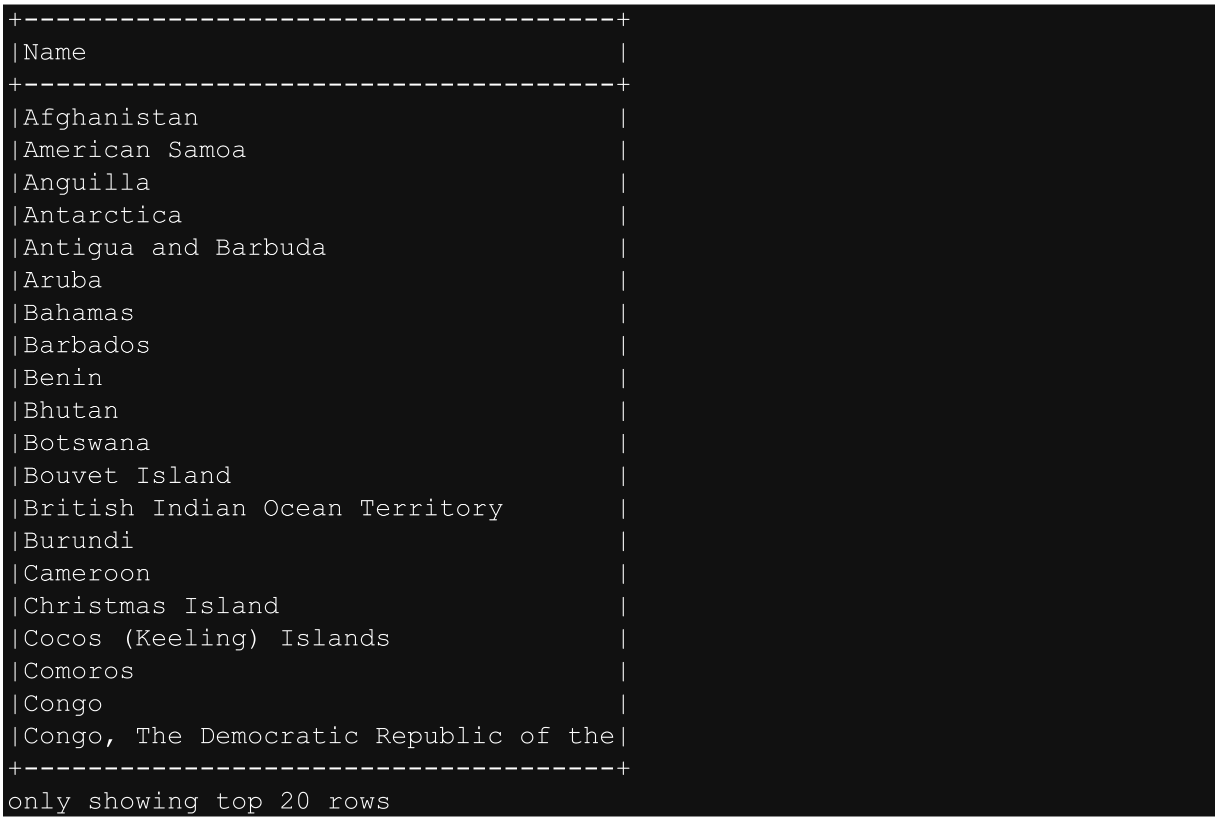
```
+----------------------------------------------------+
|country                                             |
+----------------------------------------------------+
|Ivory Coast                                         |
|Jersey                                              |
|Kazakhstan          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

C

I

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

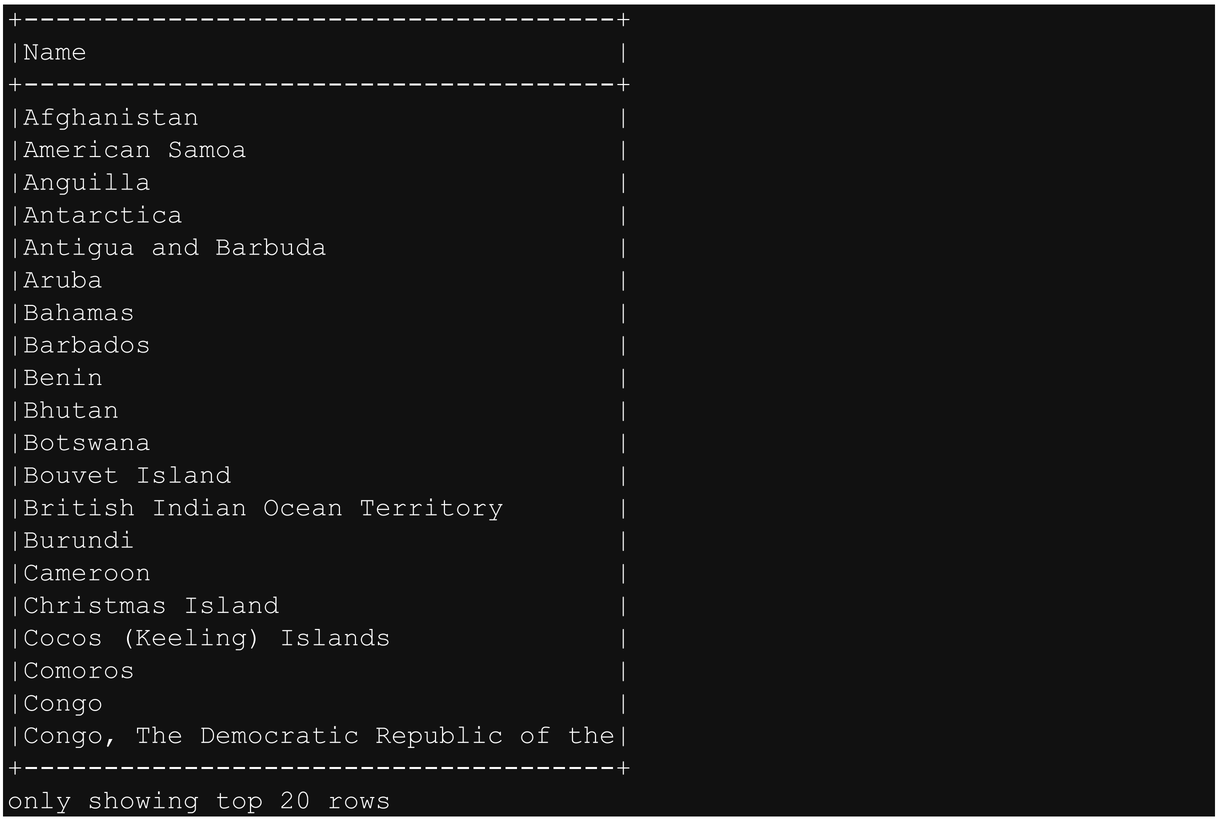
result_df = joined_df.select("Name").distinct().sort("Name")
result_df.show(truncate=False)



```
+-------------------------------------+
|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
```

II

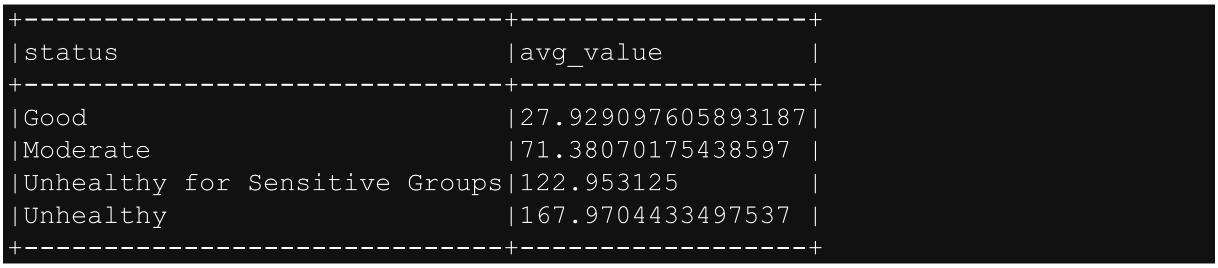
```
country_set = country_df.select("Name").distinct()
aqi_set = aqi_df.select("country").distinct()

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
```

D

```
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 |
+------------------------------+------------------+
```