ASSIGNMENT 24-25

JANUARY 11, 2024

Air Quality Data Analysis Using Apache Spark

Introduction

The objective of this analysis is to examine air quality data from sensors worldwide, focusing on identifying trends in AQI (Air Quality Index). Using Apache Spark for processing, the analysis highlights trends in air quality improvement, clusters regions by AQI, and calculates streaks of good air quality. Data is sourced from Sensor Community's JSON dataset, containing 24-hour averaged readings.

Data Preparation

1. **Data Ingestion:** The JSON dataset, which includes fields such as country, latitude, longitude, timestamp, and sensor values, was loaded into Spark with a predefined schema.
2. **Data Transformation:** Latitude and longitude fields were cast as floats and explode was applied to break down sensor data into individual records.

AQI Calculation

AQI was calculated using the UK AQI standard ranges. A custom function defined AQI values based on sensor readings, applied as a User-Defined Function (UDF) in Spark to create AQI scores per sensor reading. The values were grouped by date to obtain daily AQI for each country.

Task 1: Top 10 Countries in AQI Improvement

To determine the top 10 countries with the most AQI improvement:

- **Methodology:** Window functions were used to calculate daily AQI improvements by comparing consecutive day readings.
- **Result:** A table of the top 10 countries with the greatest AQI improvement and their current average AQI values was produced.

TOP 10 COUNTRIES IN TERMS OF AVERAGE AIR QUALITY

Index	Country	Current_Average_AQI
1	CN	8.4
2	NP	8.0
3	UZ	7.230769230769231
4	D	7.166666666666667
5	AZ	7.0
6	AM	6.5
7	PK	6.5
8	RS	6.443575042158516
9	CY	6.391304347826087
10	ID	6.285714285714286

only showing top 10 rows

Task 2: Clustering and Regional AQI Analysis

Using the KMeans algorithm to cluster data based on latitude and longitude, the analysis formed regional clusters:

- **Clustering:** The KMeans algorithm grouped data into 100 clusters based on geographical proximity.
- **Result:** A table shows the top 50 regions with the highest AQI improvement, highlighting areas with the best air quality trends.

TOP 50 REGIONS IN TERMS OF AIR QUALITY

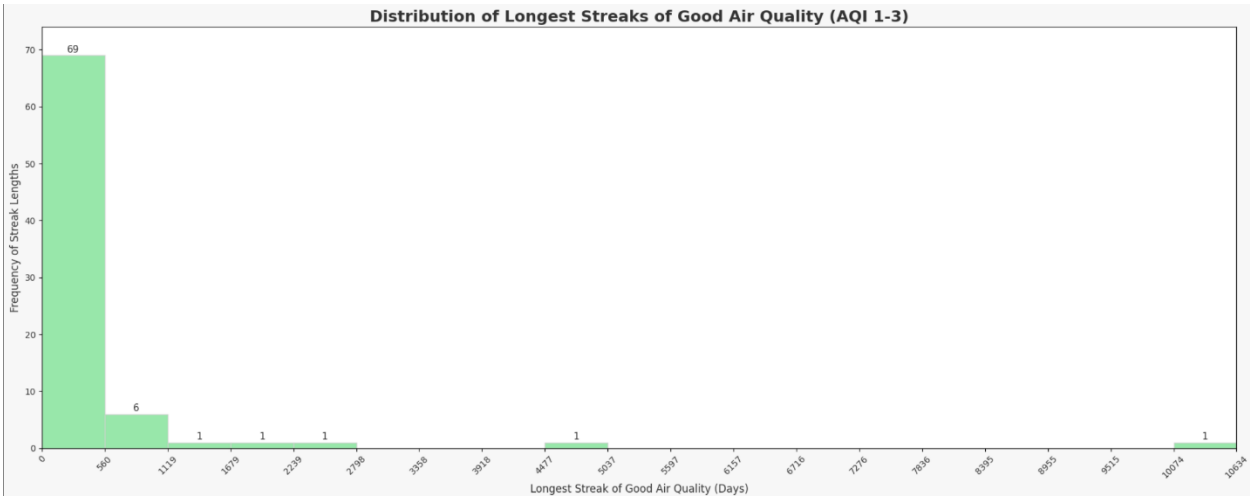
Index	Region	Current_Average_AQI
1	3	7.171428571428572
2	16	6.75
3	43	6.681792237442922
4	78	6.521739130434782
5	37	6.5
6	26	6.462172442941673
7	85	6.416666666666667
8	67	6.385185185185185
9	34	6.352517985611511
10	2	6.336040251249839
11	58	6.285714285714286
12	74	6.264942528735633
13	64	6.260273972602739
14	56	6.211675579322638
15	45	6.190627687016336
16	22	6.147058823529412
17	29	6.068595854548080 5
18	52	6.066666666666666
19	30	5.942857142857143
20	24	5.91764705882353
21	17	5.890625
22	71	5.873563218390805
23	68	5.855977640885238
24	87	5.852673796791444
25	70	5.84
26	32	5.833333333333333
27	35	5.810699083169613
28	82	5.7976190476190474
29	46	5.795023696682464
30	66	5.674983421750664
31	91	5.666666666666667
32	19	5.626991150442478
33	61	5.610687022900763 5
34	41	5.595505617977528
35	93	5.594098883572568
36	48	5.583333333333333
37	36	5.521862119530755
38	28	5.485294117647059
39	21	5.4676514188903
40	13	5.388232074438971
41	54	5.375
42	76	5.282947284345048
43	98	5.22
44	75	5.203703703703704
45	9	5.173745173745174
46	99	5.158333333333333
47	50	5.125
48	60	5.079175210754158
49	11	5.068181818181818
50	8	5.02906976744186

only showing top 50 rows

Task 3: Longest Streaks of Good Air Quality

To calculate the longest streaks of AQI values in the "good" range (1-3):

- **Methodology:** Spark's Window functions tracked consecutive good AQI days.
- **Visualization:** A histogram was generated with 20 bins, representing the distribution of streak lengths.

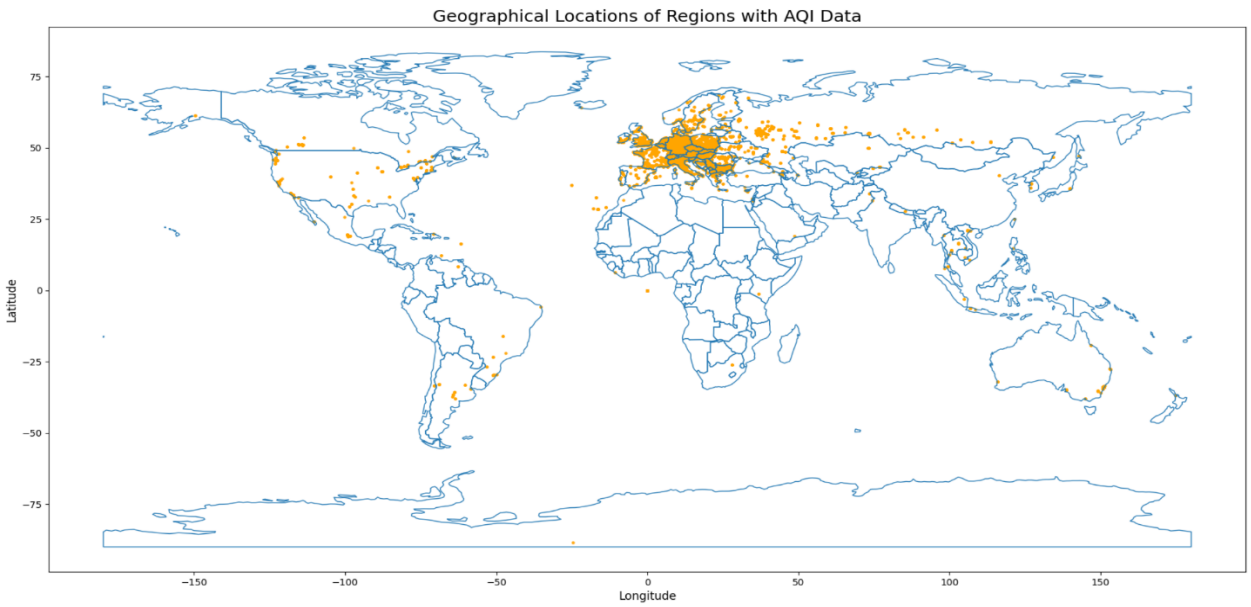


Geographical Distribution of AQI Data Points

Using GeoPandas, AQI data points were mapped to visualize their geographical distribution. The map overlays sensor readings onto country boundaries, showing regional AQI trends and highlighting areas with consistently high or low AQI.

The following files have been included to support the geographical mapping of AQI data:

- ne_110m_admin_0_countries.dbf
- ne_110m_admin_0_countries.prj
- ne_110m_admin_0_countries.shp
- ne_110m_admin_0_countries.shx
- ne_110m_admin_0_countries.cpg



Conclusion

This analysis provides insights into air quality improvements globally, identifies areas with prolonged good air quality, and offers a geographical perspective on AQI trends.