ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

KHOA CÔNG NGHỆ THÔNG TIN

HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

-BÁO CÁO ĐỒ ÁN THỰC HÀNH-

|Giảng viên hướng dẫn|

Cô Tiết Gia Hồng

Cô Hồ Thị Hoàng Vy

Cô Nguyễn Ngọc Minh Châu

ĐỒ ÁN MÔN HỌC - HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

HỌC KỲ I – NĂM HỌC 2024-2025

BẢNG THÔNG TIN CHI TIẾT NHÓM

Mã nhóm:	CQ.BI.2425.15
Số lượng:	3 sinh viên
MSSV	**Họ tên**
21120474	Võ Đức Huy
21120484	Trần Nguyễn Minh Khôi
20120444	Nguyễn Chí Công

Công việc thực hiện	**Người thực hiện**	**Mức độ hoàn thành**
Nạp dữ liệu từ Source vào Stage, NDS, DDS, Viết MDX câu 3, 4, 9, 10, Vẽ biểu đồ và nhận xét, Data Mining	Võ Đức Huy	100%
Nạp dữ liệu từ Source vào Stage, NDS, DDS, Viết MDX câu 1, 6, 11, Vẽ biểu đồ và nhận xét, Data Mining	Nguyễn Chí Công	100%
Nạp dữ liệu từ Source vào Stage, NDS, DDS, Viết MDX câu 2, 5, 12, Vẽ biểu đồ và nhận xét câu 2, 5, 12, Data Mining	Trần Nguyễn Minh Khôi	100%

Nội dung

I. Phương pháp Incremental Extract 3
II. Quy trình từ Source vào Stage 3
III. Quy trình từ Stage vào NDS 4
IV. Quy trình từ NDS vào DDS 6
V. OLAP 8
VI. Phân tích dữ liệu 9
1. **Report the min and max of AQI value for each State during each quarter of years** 9
2. **Report the mean and the standard deviation of AQI value for each State during each quarter of years** 10
3. **Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County** 11
4. **For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County** 12
5. **For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters** 13
6. **Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California** 14
7. **Use a regional map to visually represent (by color) the mean AQI value in regions during a year** 15
8. **Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year** 15
9. **Report the mean AQI value by State, Category, DayLightSaving over years** 16
10. **Count the number of days by State, Category in each month** 18
11. **Report the number of days by Category and Defining Parameter** 18

VII. Data Mining 19
VIII. Kết luận 20
1. **Tổng quan AQI năm 2023** 20
2. **Những yêu cầu đã hoàn thành của đồ án** 20

IX. Github 21

I. Phương pháp Incremental Extract

Nhóm em thực hiện phương pháp Incremental Extract để nạp dữ liệu với các bước như sau:

- Bước 1: Tạo Metadata cho các bảng cần nạp dữ liệu vào, đặt giá trị mặc định cho CET và LSET
- Bước 2: trước khi bắt đầu ETL, cập nhật CET của bảng cần nạp dữ liệu bằng với ngày hiện tại
- Bước 3: Lấy LSET và CET của bảng cần nạp dữ liệu ra
- Bước 4: Thực hiện rút trích dữ liệu từ nguồn, lấy những dòng có CET > Created >= LSET hoặc CET > LastUpdated >= LSET
- Bước 5: Cập nhật lại LSET của bảng bằng với ngày hiện tại sau khi ETL thành công

II. Quy trình từ Source vào Stage

Trước khi bắt đầu nạp dữ liệu từ các file Excel vào Stage, nhóm em đã sử dụng python để điều chỉnh các dòng có Category không ứng với giá trị AQI (ví dụ AQI là 14 mà Category là Moderate), đổi kiểu dữ liệu của cột cột county_fips trong file “(2B)uscounties.xlsx” thành string.

Tiếp theo, nhóm em thực hiện Incremental extract để nạp dữ liệu từ Source vào Stage, gồm các bước như sau:

- **Bước 1**: Update CET của bảng Stage cần nạp dữ liệu vào bằng với ngày hiện tại, các thông tin như CET và LSET của từng bảng sẽ được lưu trong bảng Data_Flow của Metadata. Ví dụ cho một câu truy vấn:

```
UPDATE DATA_FLOW
SET CET = GETDATE()
WHERE  (TABLE_NAME = 'STATE_AQI')
```

- **Bước 2:** Thực hiện TRUNCATE bảng Stage cần nạp dữ liệu vào với câu truy vấn:

```
TRUNCATE TABLE AirQualityData_Stage
```

- **Bước 3**: Lấy ra LSET và CET từ bảng Data_Flow ra và gán vào các biến đã tạo sẵn trong Package

```
SELECT LSET, CET
FROM DATA_FLOW
WHERE TABLE_NAME='STATE_AQI'
```

- **Bước 4**: Thực hiện quá trình ETL: Lấy dữ liệu từ 3 file excel theo câu truy vấn sau, câu truy vấn này sẽ lấy những dòng có CET > Created >= LSET hoặc CET > Last Updated >= LSET

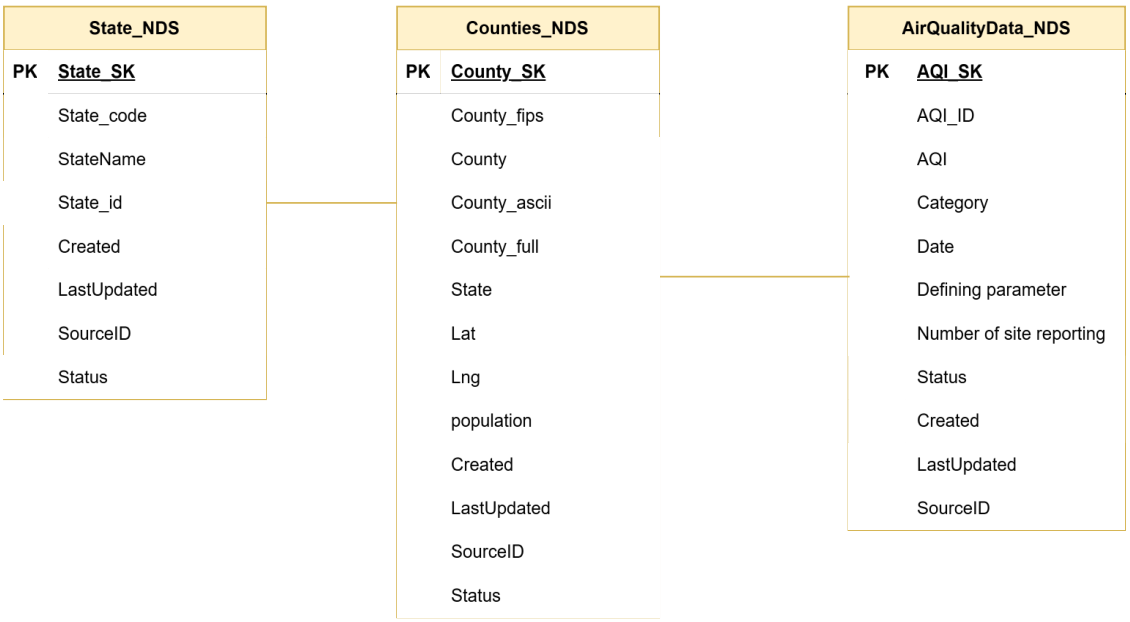
```
SELECT *
FROM [10_state_aqi_2021$]
WHERE ([Created] >= ? AND [Created] < ?) OR ([Last
Updated] >= ? AND [Last Updated] < ?)
```

- **Bước 5:** Thêm các cột cần thiết cho dữ liệu
- **Bước 6:** Nạp dữ liệu vào bảng Stage trong Database
- **Bước 7:** UPDATE lại LSET của bảng trong Metadata

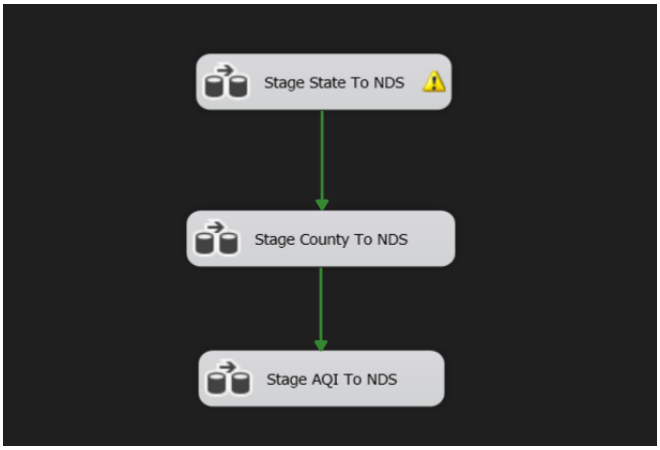
III. Quy trình từ Stage vào NDS

Link drawIO: [Project_BI.drawio - draw.io]

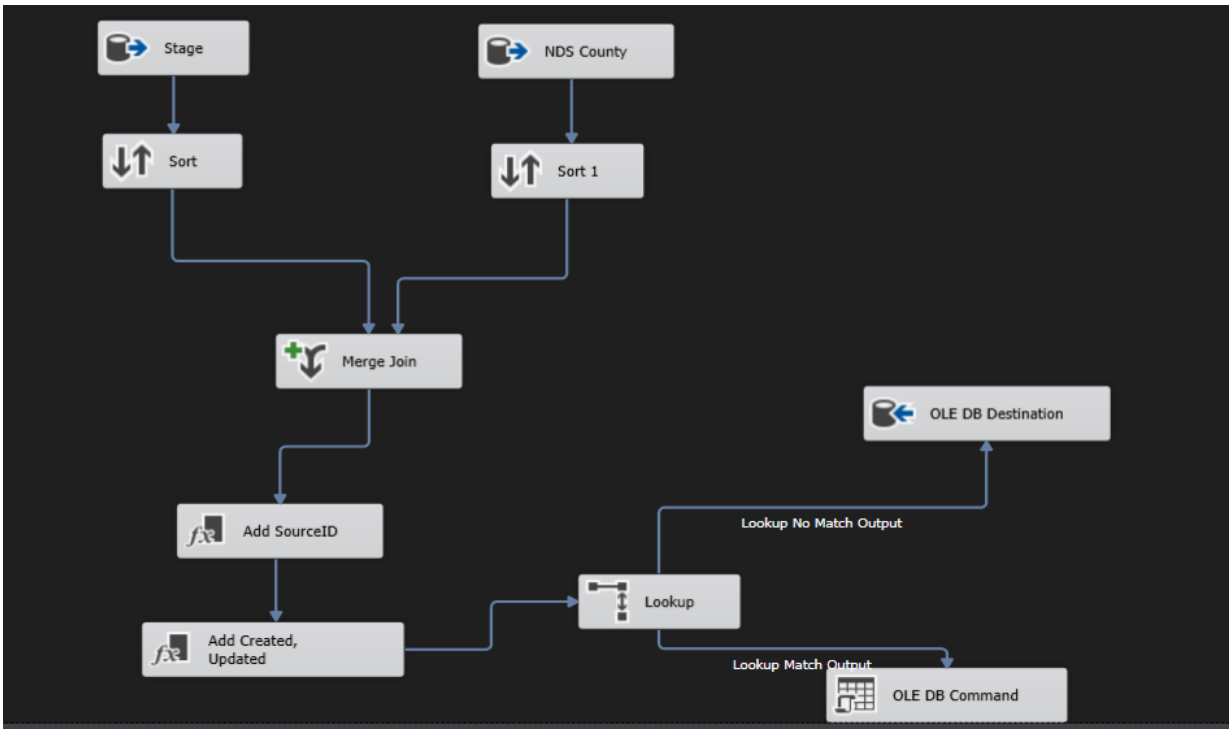
Lược đồ NDS:



Ở giai đoạn từ Stage vào NDS, nhóm em sẽ dùng 3 Data Flow Task lần lượt nạp dữ liệu từ Stage vào các bảng State_NDS, County_NDS và AirQualityData_NDS như sau:



Dưới đây là chi tiết một Data Flow nạp AirQualityData từ Stage vào NDS:



Ở Data Flow này, nhóm em thực hiện ETL như sau:

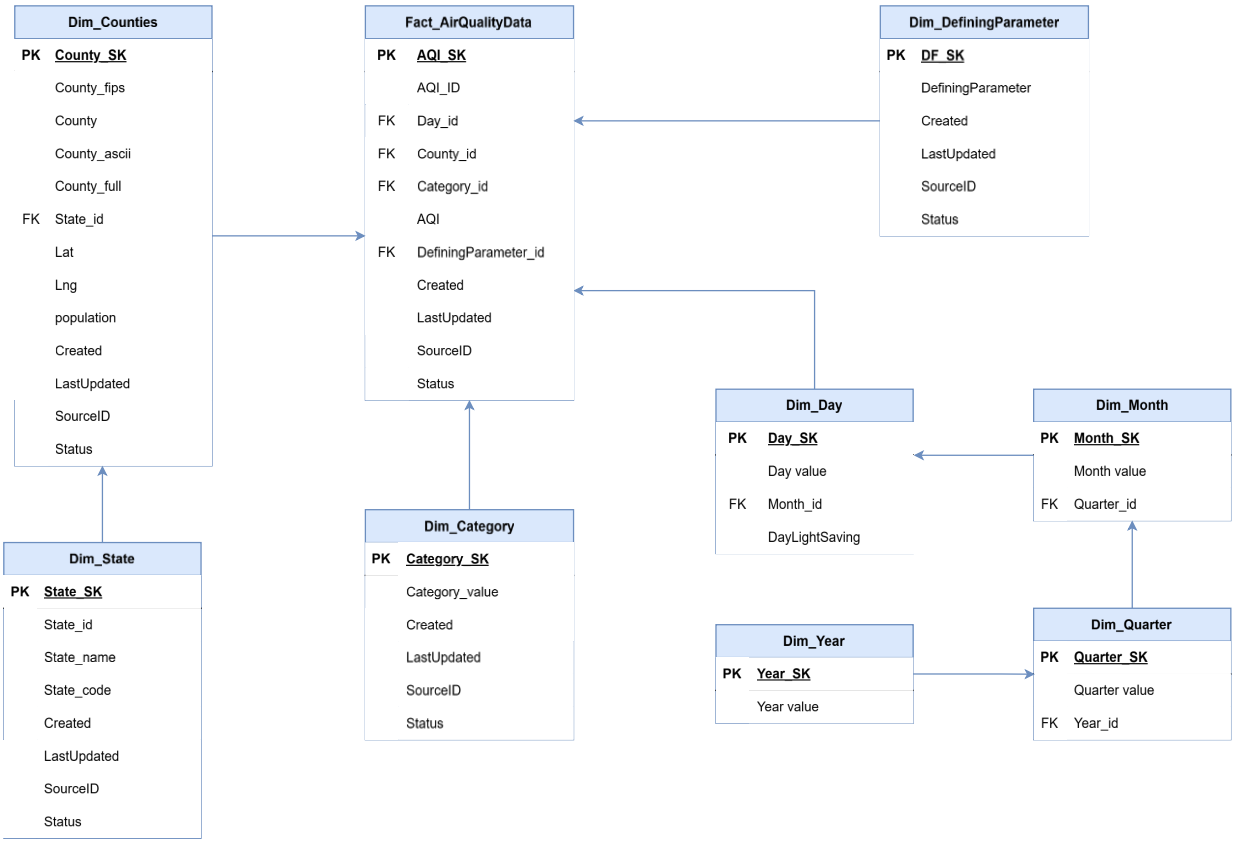
- Bước 1: Đầu tiên nhóm em sẽ lấy dữ liệu từ bảng AirQualityData_Stage từ CSDL với câu truy vấn như sau:

```
SELECT Row_ID, Date, AQI, Category, DefiningParameter,
DefiningSite, NumberOfSitesReporting, County_fips,
Status
FROM AirQualityData_Stage;
```

  , và lấy ra dữ liệu của bảng County trong NDS sau đó Sort cả hai bảng này với cột county_fips
- Bước 2: Merge Join lại để lấy ra được County_SK, sau đó thêm SourceID, Created, LastUpdated cho bảng này.
- Bước 3: Sử dụng Lookup để kiểm tra xem liệu dòng này đã tồn tại trong NDS chưa, nếu chưa thì thêm dữ liệu vào bảng, nếu đã tồn tại thì Update lại bảng

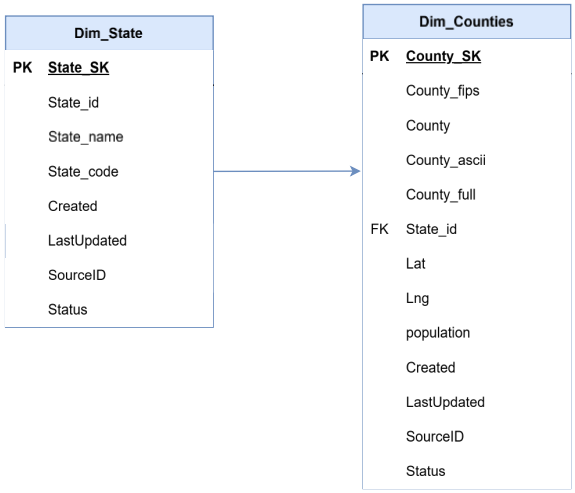
IV. Quy trình từ NDS vào DDS

Lược đồ DDS:

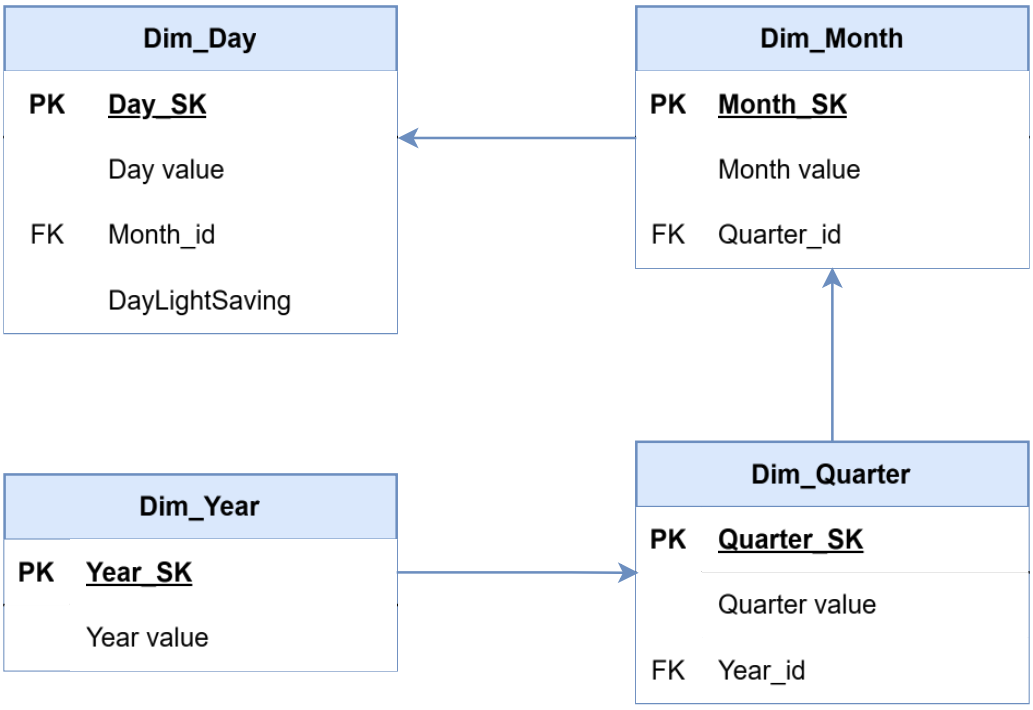


Ở quá trình ETL từ NDS sang DDS, nhóm em sẽ phân cấp chiều như sau:

- State -> County



- Year -> Quarter -> Month -> Day

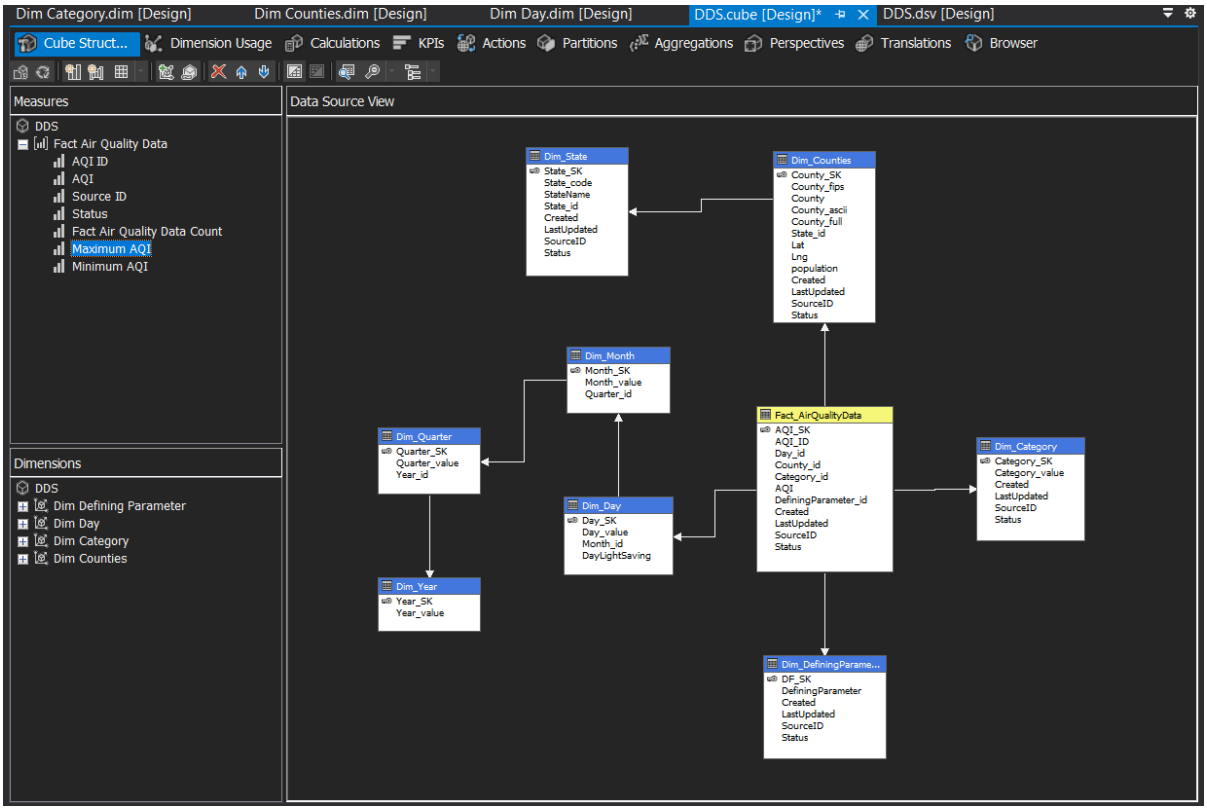


Nhóm em sử dụng Incremental Extract để nạp dữ liệu vào các bảng với thứ tự như sau:
Dim_State, Dim_County, Dim_Category, Dim_DefiningParameter, Dim_Year, Dim_Quarter, Dim_Month, Dim_Day, Fact_AirQualityData
Về phần Metdata, nhóm em sử dụng một bảng Data Flow để lưu trữ mô tả về các lần ETL, bao gồm Description, Source, Target, Transformation, Status, CET và LSET của từng bảng như sau:

DATA_FLOW	
PK	**FLOW_ID**
	DESCRIPTION
	SOURCE
	TARGET
	TRANSFORMATION
	STATUS
	ROW_INSERT
	ROW_UPDATE
	LSET
	CET

V. OLAP

Trong SSAS Project, nhóm em tạo một Data Source tên là “DDS” nối với cơ sở dữ liệu DDS đã tạo ở trên, sau đó tạo Data Source View và Cube dựa trên Data Source vừa tạo, các Measure và Dimension của Cube:



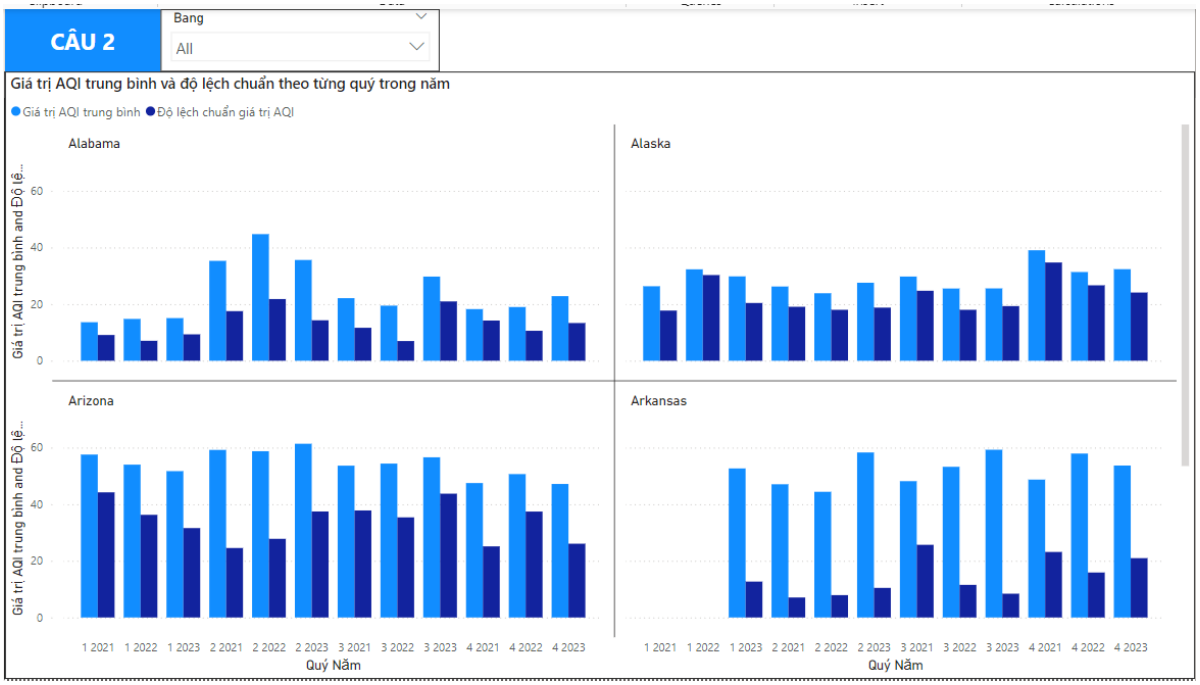
VI. Phân tích dữ liệu

1. Report the min and max of AQI value for each State during each quarter of years

Phân tích:

Tổng thể ở tất cả các bang: Giá trị AQI cao nhất thường nằm trong khoảng 100 đến 200 ở hầu hết các bang, tuy nhiên cũng có một số bang có chỉ số AQI cao nhất đạt mức từ 300 đến 500, cụ thể là 2 bang Arizona và California. Chỉ số AQI thấp nhất khá ổn định xuyên suốt trong năm, còn chỉ số AQI cao nhất thì biến đổi thường xuyên hơn, đa số sẽ giảm mạnh từ quý 2 năm 2021 đến quý 4 năm 2021 sau đó lại tăng lên. Nhìn chung chất lượng không khí có xu hướng dao động mạnh theo chu kỳ, với những đợt ô nhiễm nặng xen kẽ những đợt cải thiện. Tùy vào bang mà từng thời điểm trong năm chỉ số AQI sẽ thay đổi khác nhau.

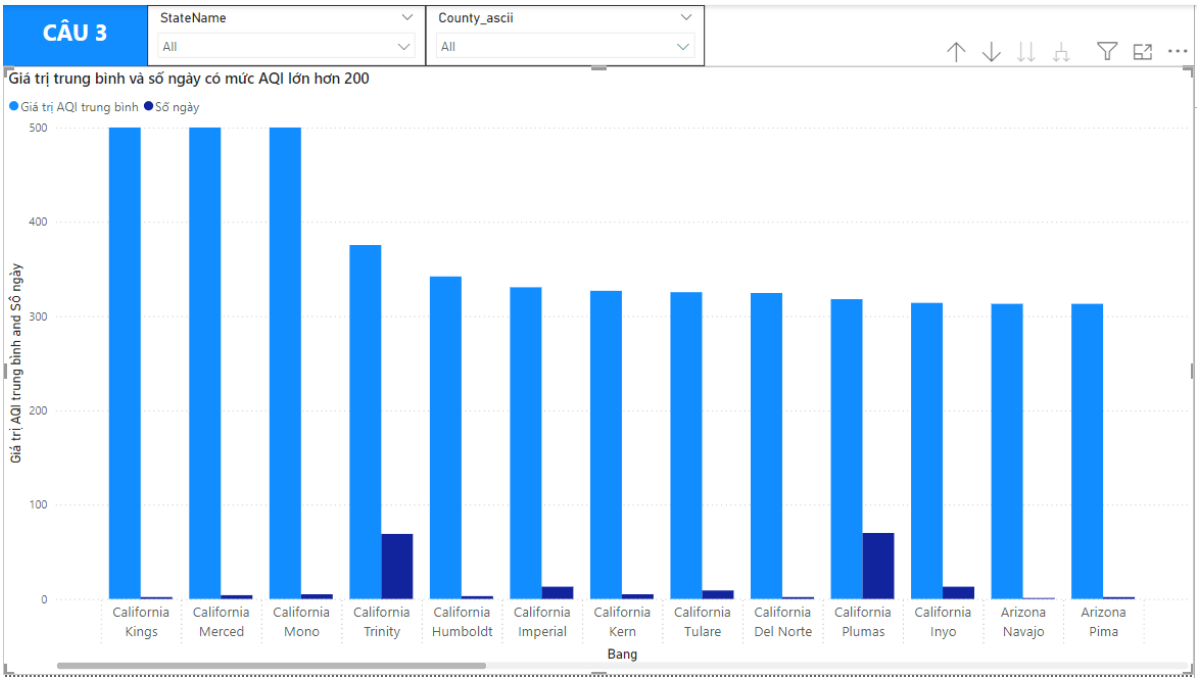
2. Report the mean and the standard deviation of AQI value for each State during each quarter of years



Phân tích:

- Năm 2021: Quý 3 có AQI trung bình cao nhất và độ lệch chuẩn cao nhất, quý 2 có AQI trung bình và độ lệch chuẩn thấp nhất, ở tất cả các quý đều ghi nhận AQI cao nhất là 500 cho thấy chất lượng không khí rất xấu
- Năm 2022: AQI trung bình tăng dần từ quý 1 đến quý 3 sau đó giảm nhẹ vào quý 4, quý 1 có AQI max là thấp nhất (297) trong khi các quý khác đều là 500, độ lệch chuẩn cũng tăng dần từ quý 1 đến quý 4
- Năm 2023: AQI trung bình có giảm nhẹ từ cuối năm 2022 đến đầu quý 1 năm 2023 nhưng lại tăng dần đến quý 3, sau đó lại giảm khi qua quý 4, độ lệch chuẩn biến động thường xuyên khi giảm từ Q1 – Q2 sau đó lại tăng từ Q2 – Q3 và tiếp tục giảm khi qua Q4, chỉ số AQI cao nhất cũng luôn là 500 ở tất cả các quý và AQI thấp nhất luôn là 0
- Xu hướng chung: chất lượng không khí có xu hướng ổn định hơn qua các năm, độ lệch chuẩn cũng có xu hướng giảm dần từ 2021 đến 2023 mặc dù giảm không quá nhiều, giá trị AQI cao nhất vẫn giữ nguyên ở mức 500 nhưng xuất hiện ít hơn

3. Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County

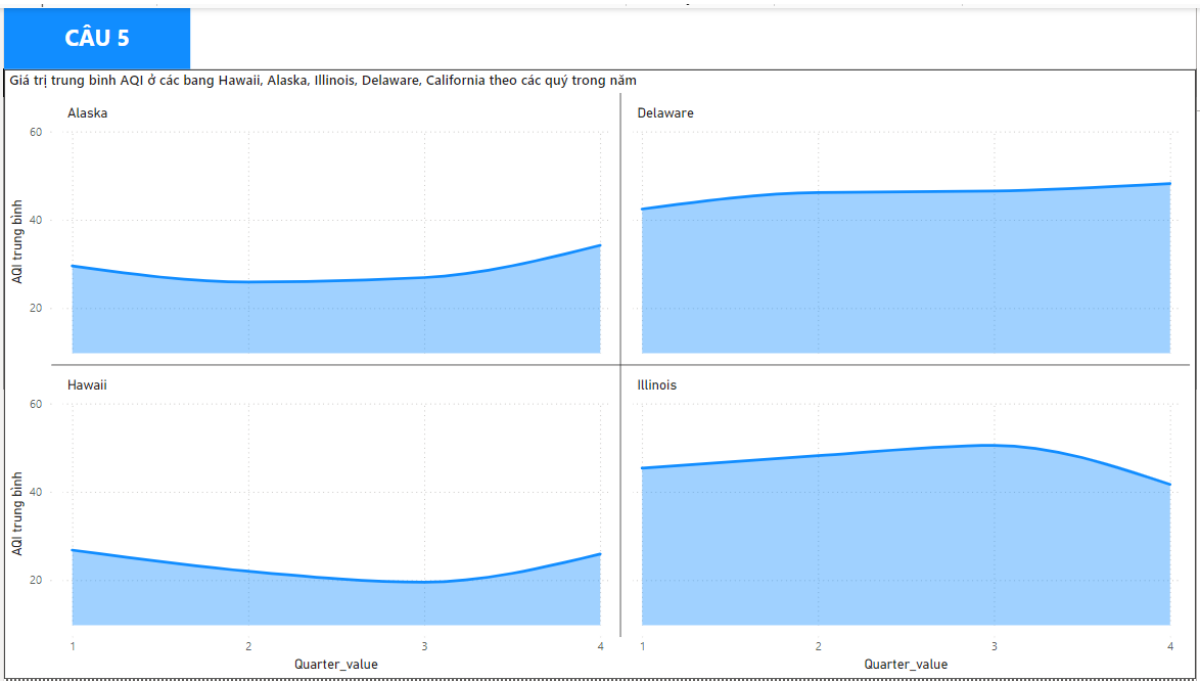


Phân tích:Có thể thấy bang California là bang xuất hiện nhiều nhất trong biểu đồ này, cho thấy mức độ AQI trung bình ở bang này đang nằm ở mức khá cao và có thể ảnh hưởng đến sức khỏe con người. Hạt có giá trị AQI trung bình cao nhất là Kings, Merced và Mono với giá trị là 500, tuy nhiên số ngày đo được mức AQI này rất thấp, nên có thể thấy mắc dù ở những hạt này có lúc giá trị AQI rất cao nhưng không diễn ra thường xuyên. Còn hạt Trinity, Plumas và San Bernardino là những nơi có AQI trung bình cao và diễn ra thường xuyên, ngoài bang California thì bang Arizona cũng có mức AQI trung bình nằm ở mức cao, cụ thể là 2 hạt La Paz có AQI trung bình là 215 với 73 ngày và Maricopa có AQI trung bình là 215 với 145.

4. For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County

Phân tích: Biểu đồ cho thấy hầu hết những hạt thuộc các bang Hawaii, Alaska, Illinois, Delaware có giá trị AQI chủ yếu nằm trong mức Good và Moderate, trong đó mức Good chiếm nhiều nhất, tuy nhiên ở 2 bang Illinois và Alaska, số ngày có AQI nằm trong mức Unhealthy và Unhealthy for Sensitive Groups vẫn còn khá cao, cụ thể là các hạt Cook, Denali, DuPage, FairBank North Star, bang Hawaii là bang có chỉ số AQI tốt nhất trong 4 bang.

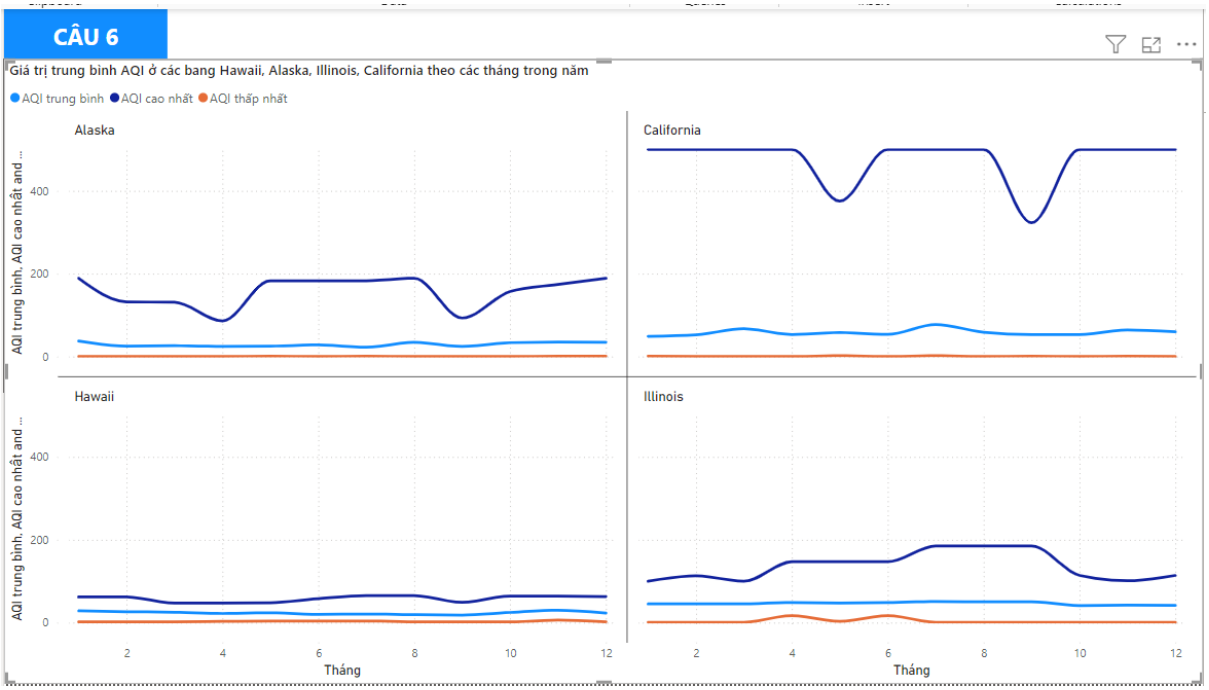
5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters



Phân tích:

- Bang Delaware và Illinois là hai bang có AQI trung bình cao nhất trong 4 bang, chỉ số AQI ở Illinois có xu hướng tăng dần từ quý 1 đến quý 3 sau đó giảm mạnh ở quý 2, còn ở Delaware, AQI trung bình có xu hướng tăng dần từ quý 1 đến quý 4, hai bang còn lại là Alaska và Hawaii có chỉ số AQI trung bình khá thấp, luôn ở dưới 40, chỉ số AQI ở hai bang này có xu hướng giảm từ quý 1 đến quý 3 và tăng lên ở quý 4

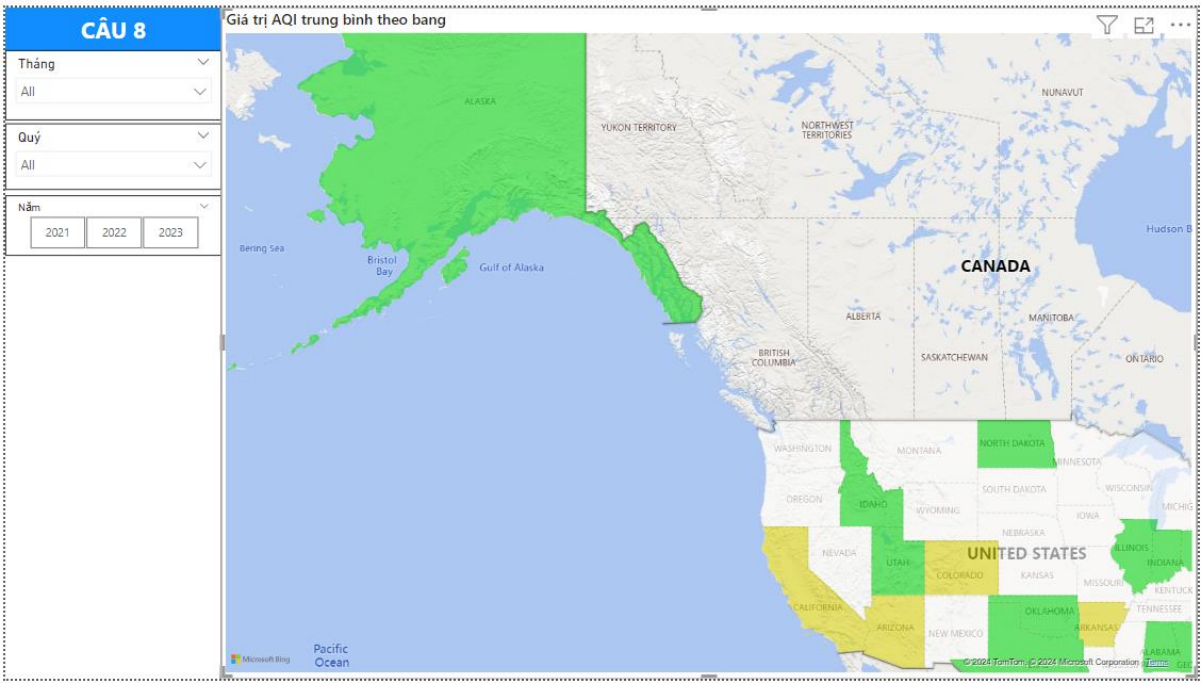
6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California



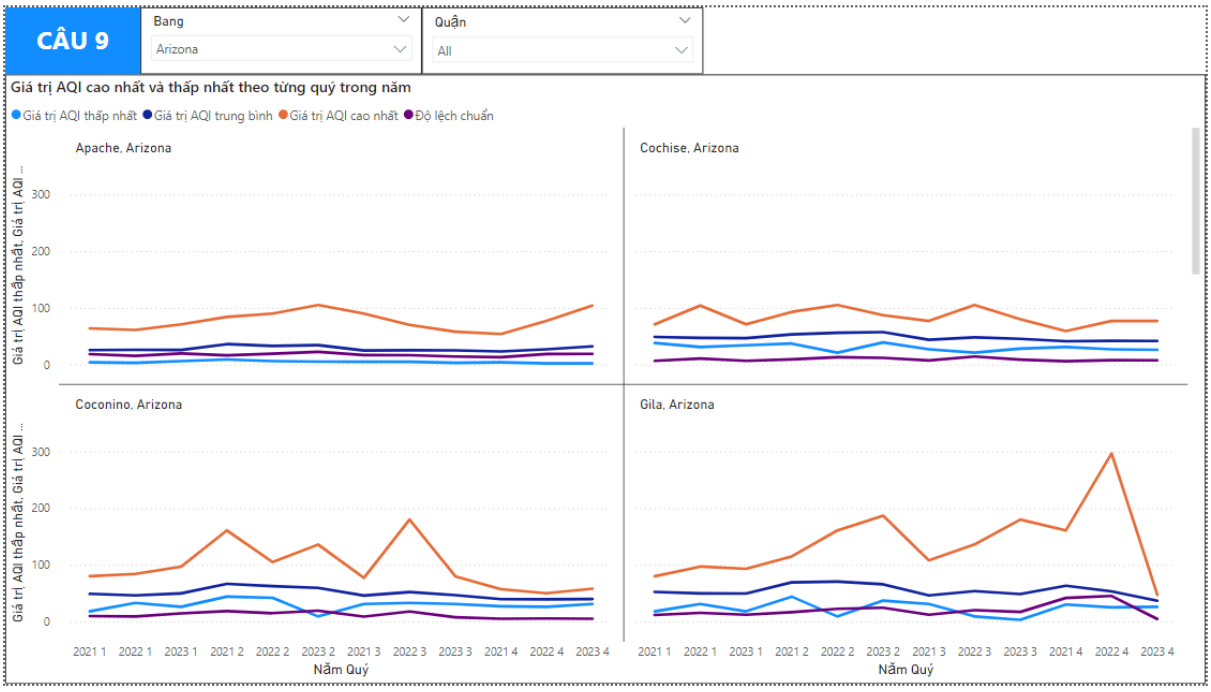
Phân tích:

- Ở bang Alaska, giá trị Aqi trung bình có thay đổi trong năm nhưng không quá nhiều và có xu hướng tăng dần, giá trị AQI thấp nhất không có thay đổi, giá trị AQI cao nhất có sự giảm dần từ tháng 1 đến tháng 4, sau đó tăng lên vào tháng 5 và ổn định đến tháng 8, sau đó lại giảm vào tháng 9 và tăng dần từ tháng 10
- Ở bang California có trung bình AQI khá cao và tăng nhẹ ở các tháng 3 và tháng 7, chỉ số AQI cao nhất thường ở mức 500, có 2 đợt giảm mạnh vào tháng 5 và tháng 9
- Bang Hawaii có các chỉ số AQI đều nằm ở mức thấp và ít khi biến đổi nhiều, cho thấy không khí ở đây có chất lượng rất tốt và ổn định
- Bang Illinois có AQI trung bình khá ổn định, nằm ở mức 40 -50, nhưng chỉ số AQI cao nhất có sự biến động rõ rệt, đặc biệt là sự tăng mạnh vào giữa năm. Điều này có thể liên quan đến các hoạt động công nghiệp, thời tiết hoặc các yếu tố mùa vụ.

7. Use a regional map to visually represent (by color) the mean AQI value in regions during a year



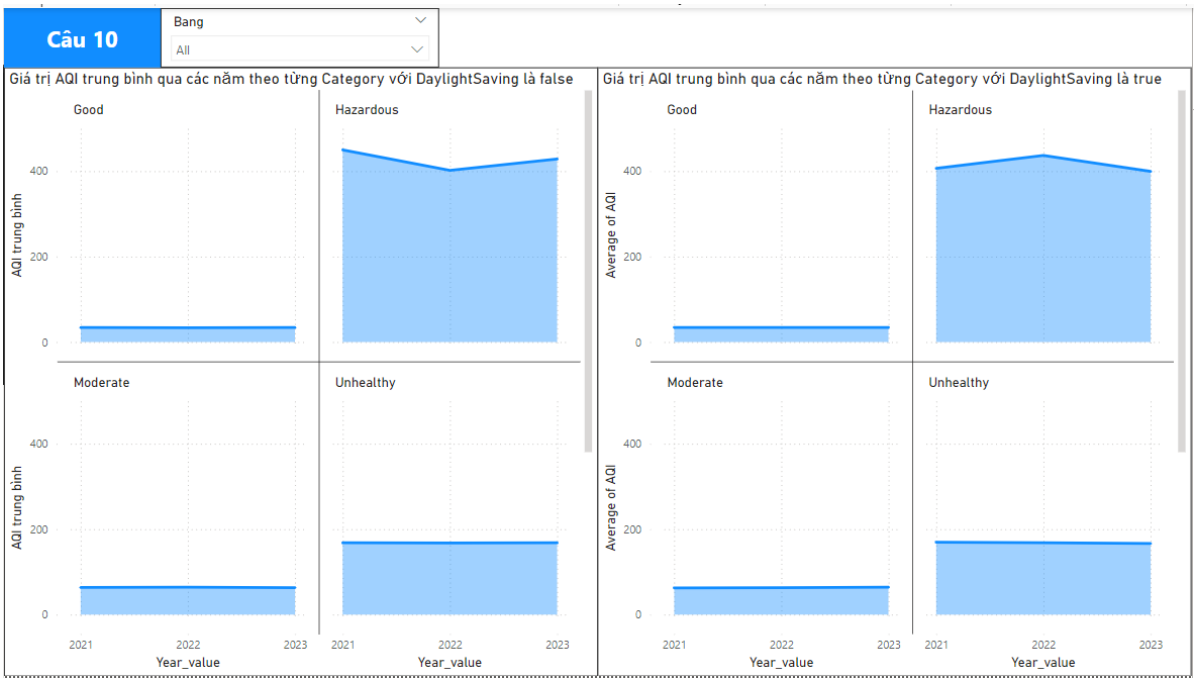
8. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year



Phân tích:

- Phần lớn các giá trị Aqi trung bình nằm trong khoảng 30-70, một số quận có giá trị AQI trung bình cao bất thường là Maricopa và La Paz của bang Arizona, Trinity và Imperial của bang California
- Phần lớn độ lệch chuẩn nằm trong khoảng 10-25, một số quận có độ lệch chuẩn cao bất thường là Maricopa và Gila của bang Arizona, La Paz và Imperial của bang California
- Hầu hết các bang đều có AQI nhỏ nhất nằm trong khoảng 10-30
- Hầu hết các giá trị AQI cao nhất nằm trong khoảng 70-150, tuy nhiên có một số quận ở bang Arizona và California có mức AQI cao nhất nằm trong khoảng 300 – 500, cho thấy mức độ ô nhiễm không khí ở hai bang này đang khá cao

9. Report the mean AQI value by State, Category, DayLightSaving over years

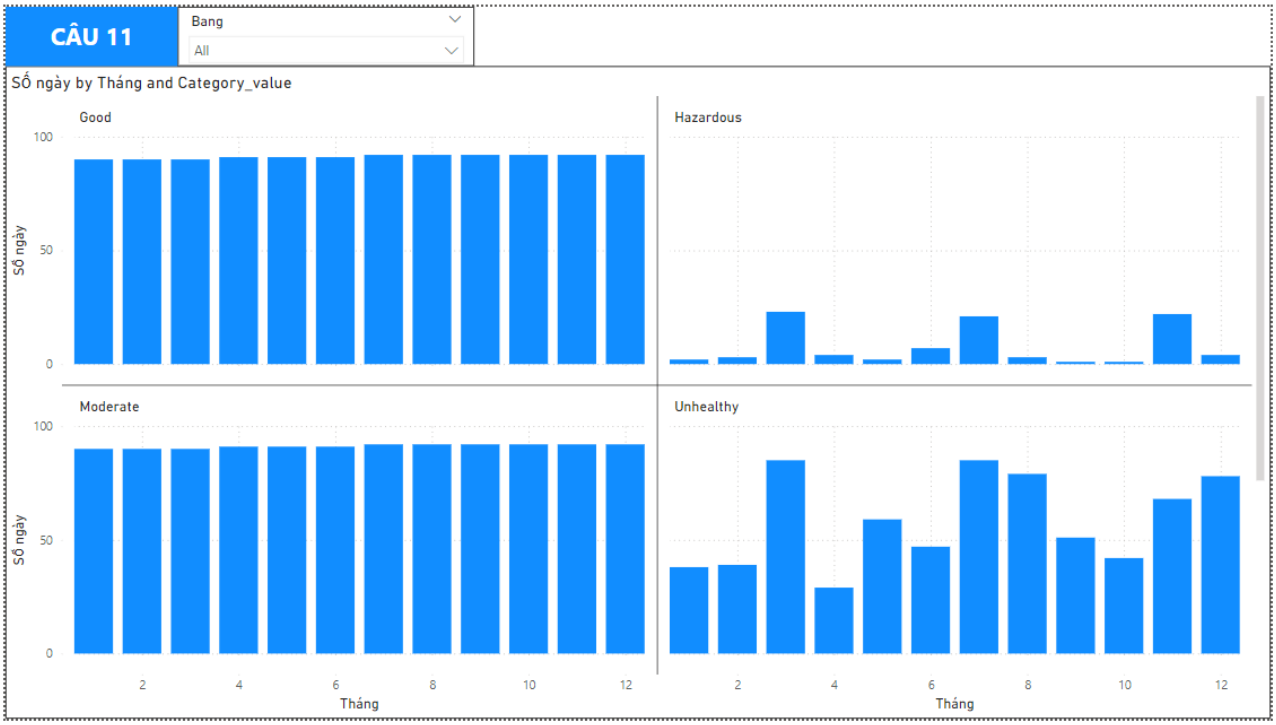


Phân tích:

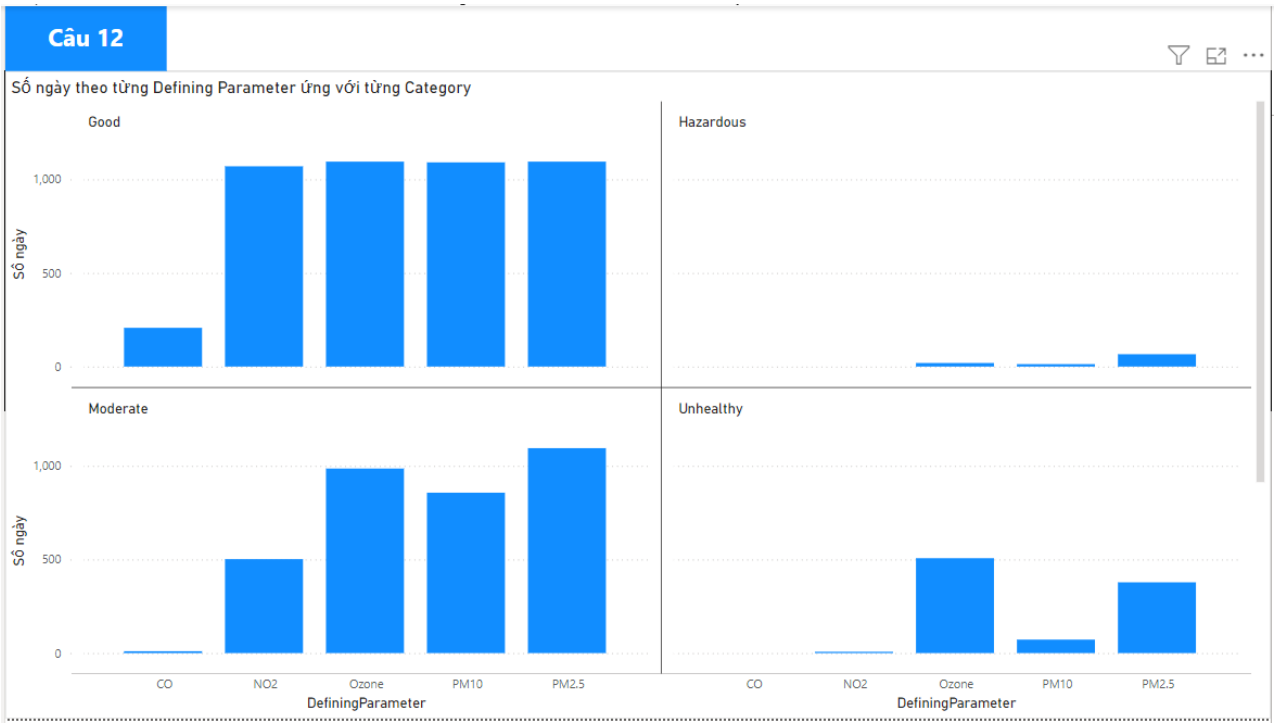
- Đối với mức "Good" (Tốt): Phần lớn các bang có chỉ số AQI trung bình tương đương giữa khoảng thời gian Daylight Saving Time và các thời điểm khác. Có một số khác biệt đáng chú ý như:
  + Alabama: AQI cao hơn trong khoảng thời gian Daylight Saving Time (25.74 so với 15.87).
  + Hawaii: AQI thấp hơn trong khoảng thời gian Daylight Saving Time (20.02 so với 24.28).

    + Illinois: AQI cao hơn trong khoảng thời gian Daylight Saving Time (37.5 so
    + với 33.47).
    + Utah: AQI cao hơn trong khoảng thời gian Daylight Saving Time (22.68 so với 17.72).
- Đối với mức "Moderate" (Trung bình): Hầu hết các bang chỉ có sự khác biệt nhỏ giữa hai khoảng thời gian. Một số khác biệt đáng chú ý:
    + Arizona: AQI cao hơn trong khoảng thời gian Daylight Saving Time (67.38 so với 63.22).
    + California: AQI thấp hơn trong khoảng thời gian Daylight Saving Time (65.74 so với 67.28).
    + Illinois: AQI cao hơn trong khoảng thời gian Daylight Saving Time (62.32 so
    + với 59.29).
- Đối với mức "Unhealthy for Sensitive Groups" : Một số bang có sự khác biệt như:
    + Alaska: AQI thấp hơn trong khoảng thời gian Daylight Saving Time (112 so với 122.89).
    + Arizona: AQI thấp hơn trong khoảng thời gian Daylight Saving Time (116.33
    + so với 127.17).
    + California: AQI thấp hơn trong khoảng thời gian Daylight Saving Time (119.53 so với 121.13).
    + Ohio: AQI cao hơn trong khoảng thời gian Daylight Saving Time (111.32 so với 107.5).
- Đối với các mức nghiêm trọng hơn (Unhealthy, Very Unhealthy, Hazardous): Ở mức này thường có ít dữ liệu và giá trị AQI cũng không quá khác biệt giữa hai khoảng thời gian
- Nhận xét tổng quát: Có thể thấy không có một xu hướng nhất quán về việc khoảng thời gian Daylight Saving Time, có những bang chỉ số AQI sẽ cao hơn trong khoảng Daylight Saving Time nhưng cũng có những nơi thấp hơn, chưa đủ bằng chứng để kết luận DayLight Saving Time có ảnh hưởng đến chất lượng không khí.

10. Count the number of days by State, Category in each month



11. Report the number of days by Category and Defining Parameter



Biểu đồ cho thấy các chất gây ảnh hưởng đến chỉ số AQI nhiều nhất là Ozone và PM2.5 với số ngày cao hơn hẳn các chất khác trong các Category Moderate, Hazadous, Unhealthy, Very Unhealthy và Unhealthy for Sensitive Groups, PM10 cũng gây ảnh hưởng nhưng không nhiều bằng hai chất trên,

do đó chính phủ nên có các biện pháp để kiểm soát nồng độ các chất này nhằm cải thiện chất lượng không khí

VII. Data Mining

Trong phần Data Mining, nhóm em sử dụng thuật toán Microsoft Time Series có sẵn trong SSAS project để thực hiện dự đoán giá trị AQI vào ngày, tháng hoặc năm tiếp theo. Thuật toán Time Series thường được sử dụng để dự đoán cho các dữ liệu thay đổi theo thời gian. Thuật toán Time Series là sự kết hợp của 2 phương pháp ARTXP và ARIMA, RTXP được tối ưu hóa cho các dự đoán tương lai gần. Khi cần dự đoán xa hơn thì các giá trị của thuật toán ARIMA đưa ra đúng hơn so với ARTXP. Các dữ liệu cần thiết cho một mô hình Time Series là:

- Một cột nhãn thời gian (ngày, tháng, năm)
- Một cột chứa giá trị dự đoán
- Các giá trị bổ sung

Để dự đoán chỉ số AQI, nhóm em đã chuẩn bị các bảng cho Data Mining như sau:

DataMining_Day	
PK	AirQualityData_Key
	Date
	AverageAQI
	MaxAQI
	MinAQI

DataMining_Month	
PK	AirQualityData_Key
	Month
	AverageAQI
	MaxAQI
	MinAQI

DataMining_Year	
PK	AirQualityData_Key
	Year
	AverageAQI
	MaxAQI
	MinAQI

Về cấu trúc của mô hình Data Mining, cột Date sẽ là Key, AverageAQI, MaxAQI và MinAQI sẽ là Input và Predict là AverageAQI

Structure	Data Mining Month
	Microsoft_Time_Series
Air Quality Data Key	Ignore
Average AQI	Predict
Max AQI	Input
Min AQI	Input
Month	Key

Kết quả dự đoán chỉ số AQI cho 3 tháng tiếp theo:

Expression

$TIME	Average AQI
01/01/2024 00:00:00	47
01/02/2024 00:00:00	48
01/03/2024 00:00:00	48

VIII. Kết luận

1. Tổng quan AQI năm 2023

Nhìn tổng thể năm 2023, chỉ số AQI trung bình có giảm nhưng không đáng kể, vẫn còn một số ngày có chỉ số AQI đạt mức “Hazardous”, trong đó bang có chất lượng không khí tốt nhất là Hawaii với 24,17, bang California và Arizona là 2 bang có chất lượng không khí tệ nhất. Điều này có thể giải thích bởi các điều kiện thời tiết cực đoan ở đây, các hoạt động công nghiệp và đô thị lớn, cháy rừng (đặc biệt là ở California)

2. Những yêu cầu đã hoàn thành của đồ án

Những yêu cầu mà nhóm đã hoàn thành trong đồ án:

- Thiết kế các lược đồ NDS, DDS
- Nạp dữ liệu từ Source vào Stage, từ Stage vào NDS và NDS vào DDS
- Tạo OLAP cube cho việc phân tích
- Viết truy vấn cho các câu MDX từ câu 1 đến câu 12
- Vẽ các biểu đồ để phân tích và nhận xét theo yêu cầu
- Xây dựng mô hình dự đoán chỉ số AQI trong ngày, tháng, năm tiếp theo

Những phần cần cải thiện

- Dự đoán chỉ số AQI cho quý tiếp theo
- Thiết kế DDS tốt hơn

IX. Github

Link github: https://github.com/dhuy00/project_BI.git

Link drawio: Project_BI.drawio - draw.io