TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

----------□ □----------

BÁO CÁO ĐỒ ÁN THỰC HÀNH

GĐ4: DATA MINING, CONCLUSION

Môn: Hệ thống thông tin phục vụ trí tuệ kinh doanh

18/12/2024 – 29/12/2024

MÃ HỌC PHẦN: CSC12107
Nhóm: 3
Thành viên:

ID	Họ tên
19127536	Võ Lâm Hải Quốc
21127211	Nguyễn Vũ Tường An
21127450	Võ Trung Tín
21127699	Lô Thủy Tiên

Giảng viên:
ThS. Hồ Thị Hoàng Vy
ThS. Tiết Gia Hồng
ThS. Nguyễn Ngọc Minh Châu

Báo cáo:
Lô Thuỷ Tiên

Thành phố Hồ Chí Minh – 2024

MỤC LỤC

THÔNG TIN VỀ ĐỒ ÁN

Mã học phần: CSC12107

Tên học phần: Hệ thống thông tin phục vụ trí tuệ kinh doanh

Tên : Đồ án thực hành - 2425.BI.DATH

Hình thức:

- Báo cáo (.doc, ppt)

- Source demo

- Video demo

Mô tả:

Dữ liệu sẽ được phân tích để phát hiện các xu hướng và mẫu trong chất lượng không khí của Hoa Kỳ từ năm 2021 đến 2023. Đồ án yêu cầu xây dựng một kho dữ liệu (DW) từ các nguồn dữ liệu thô, sau đó triển khai các quy trình ETL, thiết kế mô hình OLAP, thực hiện phân tích dữ liệu và tạo các báo cáo.

- **Dữ liệu**: Dữ liệu chất lượng không khí hàng ngày của EPA, phân chia theo quận từ năm 2021 đến 2023, kết hợp với dữ liệu địa lý và định nghĩa phân loại AQI.
- **Thiết kế kho dữ liệu**: Xây dựng các bảng chiều như Địa lý (State > County), Thời gian (Year > Quarter > Month > Day) và các chiều khác để phục vụ yêu cầu báo cáo.
- **Yêu cầu báo cáo và phân tích**:
    - Tạo các báo cáo biểu đồ để trình bày sự biến động AQI theo thời gian.
    - Phân tích các câu hỏi mở để đưa ra đánh giá về biến động AQI, ngày ô nhiễm nặng, và các yếu tố có ảnh hưởng đến chất lượng không khí.
    - Tích hợp bản đồ khu vực để biểu diễn trung bình AQI qua các khu vực.

- **Khai phá dữ liệu:** Ứng dụng các mô hình khai phá dữ liệu nhằm dự đoán chất lượng không khí trong các kỳ tới (Q1-2024, tháng 01-2024, ...), giải thích thuật toán, lý do chọn lựa, và trình bày kết quả.
- **Tổng quan và kết quả**: Đưa ra đánh giá tổng quan về chất lượng không khí ở các quận Hoa Kỳ vào năm 2023. Kết luận các thành tựu đạt được trong đồ án, cũng như đề xuất hướng cải thiện cho chất lượng không khí trong tương lai.

Mục tiêu:

- Thiết kế mô hình dữ liệu: Sử dụng mô hình ngôi sao hoặc bông tuyết để đáp ứng yêu cầu phân tích.
- Quy trình ETL: Sử dụng công cụ SSIS để thu thập, làm sạch và tích hợp dữ liệu từ các nguồn khác nhau vào kho dữ liệu.
- OLAP và Trực quan hóa: Sử dụng công cụ SSAS để khai thác các công nghệ OLAP cơ bản và tạo báo cáo bằng SSRS hoặc Excel.
- Khai phá dữ liệu: Áp dụng các thuật toán khai phá dữ liệu với SSAS để phân tích và dự đoán chất lượng không khí.

Giảng viên phụ trách: Cô Hồ Thị Hoàng Vy, Cô Tiết Gia Hồng, Cô Nguyễn Ngọc Minh Châu

THÔNG TIN NHÓM

Nhóm: 3

MSSV	Họ tên	Email	Ghi chú
19127536	Võ Lâm Hải Quốc	vlhquoc19@clc.fitus.edu.vn	
21127211	Nguyễn Vũ Tường An	nvtan21@clc.fitus.edu.vn	
21127450	Võ Trung Tín	vttin21@clc.fitus.edu.vn	
21127699	**Lô Thủy Tiên**	**lttien21@clc.fitus.edu.vn**	

I. DATA MINING:

Đề xuất: Sử dụng các mô hình để dự đoán chất lượng không khí trong các giai đoạn tiếp theo như quý tiếp theo (Q1-2024), tháng tiếp theo (01-2024), v.v.

1. Sarima:

a. Thuật toán Sarima: SARIMA (Seasonal Autoregressive Integrated Moving Average) là một mở rộng của mô hình ARIMA, được thiết kế để xử lý các dữ liệu chuỗi thời gian có yếu tố mùa vụ.

b. Lý do sử dụng: Lý do sử dụng SARIMA bao gồm:

- Xử lý mùa vụ: SARIMA có khả năng mô hình hóa các mẫu lặp lại theo chu kỳ, chẳng hạn như các biến động theo mùa trong dữ liệu chất lượng không khí.
- Dự báo chính xác hơn: Bằng cách bao gồm các thành phần mùa vụ, SARIMA có thể cung cấp dự báo chính xác hơn cho các chuỗi thời gian có yếu tố mùa vụ rõ rệt.
- Linh hoạt: SARIMA có thể được điều chỉnh để phù hợp với nhiều loại dữ liệu chuỗi thời gian khác nhau bằng cách thay đổi các tham số của nó.

c. Cách thực hiện:

Bước 1: Đọc và tiền xử lý dữ liệu

- Dữ liệu được đọc từ file `DDS_AIR_5.xlsx` (đây là tương đương với database sau khi etl và chuyển đổi dữ liệu nhưng mà được extract ra để cho những máy khác có thể chạy được thay vì chỉ local) với hai sheet:
  - **FACT_AirQuality**: chứa các thông số về chất lượng không khí.
  - **DIM_Date**: chứa thông tin về ngày tháng.
- Hai bảng được **merge** dựa trên cột `DateSK` để kết hợp thông tin ngày tháng vào dữ liệu AQI.

Bước 2: Chuyển đổi và chuẩn hóa dữ liệu

- Chuyển cột `Date` thành định dạng datetime và đặt làm chỉ số (index).

- Dữ liệu được chuẩn hóa (scaling) bằng `StandardScaler` để các giá trị AQI nằm trong cùng thang đo.

Bước 3: Chi dữ liệu thành các tập huấn luyện, kiểm tra và dự đoán

- **train**: Dữ liệu từ ngày 01/01/2021 đến 31/12/2022 được sử dụng để huấn luyện mô hình.
- **val**: Dữ liệu từ ngày 01/01/2023 đến 30/09/2023 được sử dụng để kiểm tra mô hình.
- **test**: Dữ liệu từ ngày 01/10/2023 đến 31/12/2023 được sử dụng để đánh giá và dự đoán mô hình cho quý tiếp theo.

Bước 4: Huấn luyện mô hình SARIMA cho tập huấn luyện

Khởi tạo mô hình SARIMA với các tham số:

- **order=(1, 1, 1)**: Tham số ARIMA không mùa vụ (p, d, q).
- **seasonal_order=(1, 1, 1, 12)**: Tham số ARIMA mùa vụ (P, D, Q, m) với chu kỳ mùa vụ là 12 (hàng năm).

Bước 5: Tiến hành dự đoán và đánh giá mô hình trên tập kiểm tra và dự đoán

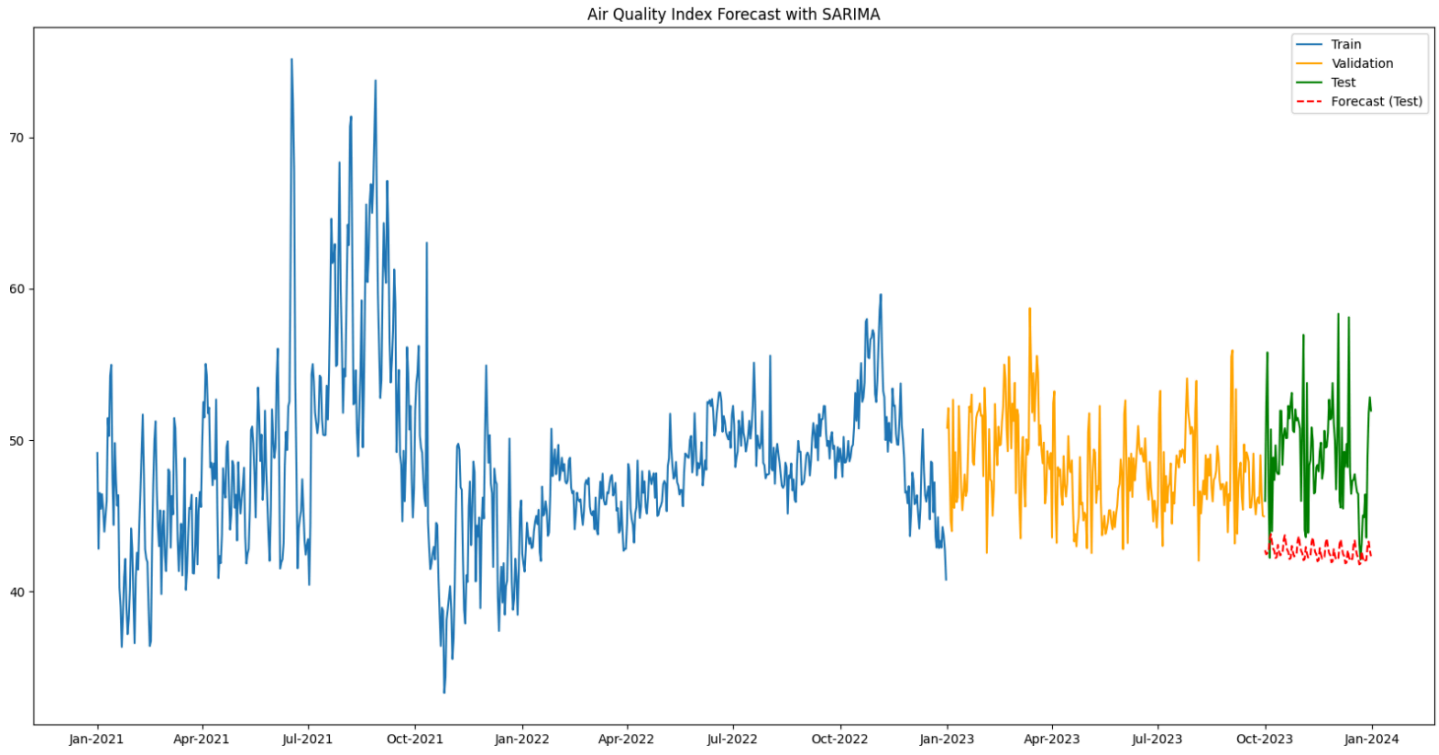
- **mean_squared_error**: Tính toán lỗi bình phương trung bình giữa giá trị thực tế và giá trị dự đoán.
- **np.sqrt**: Tính căn bậc hai của lỗi bình phương trung bình để có được **RMSE** (**Root Mean Squared Error**).

Bước 6: Vẽ biểu đồ dự đoán so với dữ liệu thực tế

d. Kết quả thực hiện:

Tập dữ liệu	**train**	**val**	**test**
Kết quả RMSE	–	5.41	7.25

Đồ thị kết quả dự đoán sử dụng mô hình SARIMA:



e. Nhận xét kết quả thực hiện:

Dựa trên RMSE:

- Trên tập **val**: RMSE: 5.406538438555035
  - **Độ chính xác**: RMSE (Root Mean Squared Error) là 5.41 cho thấy mô hình dự đoán khá chính xác trên tập kiểm tra (validation). Giá trị RMSE thấp cho thấy sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ.
  - **Hiệu suất**: Mô hình hoạt động tốt trên tập kiểm tra, cho thấy khả năng dự đoán chính xác các giá trị AQI trong khoảng thời gian từ 01/01/2023 đến 30/09/2023.
  - **Ứng dụng**: Kết quả này có thể được sử dụng để đưa ra các dự báo ngắn hạn về chất lượng không khí, giúp các cơ quan chức năng và cộng đồng có biện pháp phòng ngừa kịp thời.
- Trên tập **test**: RMSE: 7.251893611847774
  - **Độ chính xác**: RMSE là 7.25 cho thấy mô hình dự đoán có độ chính xác thấp hơn trên tập kiểm tra (test) so với tập kiểm tra (validation).

Giá trị RMSE cao hơn cho thấy sai số giữa giá trị dự đoán và giá trị thực tế lớn hơn.

- **Hiệu suất**: Mô hình hoạt động kém hơn trên tập kiểm tra, có thể do các yếu tố như sự thay đổi trong dữ liệu hoặc các mẫu mới không được mô hình hóa tốt.
- **Ứng dụng**: Mặc dù độ chính xác giảm, kết quả này vẫn có thể cung cấp thông tin hữu ích cho dự báo chất lượng không khí, nhưng cần thận trọng hơn khi sử dụng các dự báo này để đưa ra quyết định.

Dựa trên kết quả từ đồ thị: Từ đồ thị trên, có thể rút ra một số kết luận về dự đoán AQI cho Q1-2024:

- Xu hướng dự đoán:
  - Giá trị AQI được dự đoán sẽ dao động ổn định ở mức khoảng 42-43.
  - Đường dự đoán (màu đỏ đứt nét) khá phẳng và không thể hiện nhiều biến động.
- So sánh với dữ liệu lịch sử:
  - Giá trị dự đoán thấp hơn đáng kể so với thực tế Q4-2023 (đường màu xanh lá cây) có AQI dao động 45-58.
  - Không phản ánh được tính chu kỳ/mùa vụ như dữ liệu trong quá khứ.
- Đánh giá độ tin cậy:
  - Model có vẻ chưa nắm bắt tốt tính biến động của dữ liệu.
  - Dự đoán khá bảo thủ khi đưa ra giá trị ổn định.
- Cần cải thiện model bằng cách:
  - Thêm các biến môi trường khác.
  - Xem xét sử dụng mô hình phức tạp hơn để nắm bắt tốt hơn các pattern.

2. K-means:

a. Thuật toán K-means: K-means là một thuật toán học không giám sát (unsupervised learning) được sử dụng để phân cụm dữ liệu dựa trên sự tương đồng của các đặc trưng. Trong trường hợp dự đoán chất lượng không khí, thuật toán hoạt động như sau:

+ Bước 1: Khởi tạo K tâm cụm ngẫu nhiên trong không gian dữ liệu (K là số cụm mong muốn)
+ Bước 2: Gán mỗi điểm dữ liệu vào cụm gần nhất dựa trên khoảng cách Euclidean
+ Bước 3: Tính lại tâm cụm mới bằng trung bình các điểm trong cụm
+ Bước 4: Lặp lại bước 2-3 cho đến khi các tâm cụm không thay đổi nhiều

b. Lý do sử dụng:

Ưu điểm chính:

- Khả năng phát hiện mẫu: K-means có thể tìm ra các mẫu tự nhiên trong dữ liệu chất lượng không khí, giúp xác định các kịch bản ô nhiễm khác nhau
- Đơn giản và hiệu quả: Thuật toán dễ hiểu, triển khai và chạy nhanh ngay cả với dữ liệu lớn
- Khả năng mở rộng: Có thể xử lý nhiều thông số chất lượng không khí cùng lúc
- Có thể phát hiện các mẫu/pattern trong dữ liệu AQI theo thời gian
- Giúp phân loại các khoảng thời gian có chất lượng không khí tương tự nhau
- Hỗ trợ việc dự đoán bằng cách xem xét các nhóm có đặc điểm tương đồng

Phù hợp với bài toán vì:

- Dữ liệu chất lượng không khí thường có tính chu kỳ và mẫu lặp lại theo mùa
- Có thể phân loại các ngày thành các nhóm chất lượng không khí tương tự nhau
- Giúp dự đoán xu hướng bằng cách xem xét cụm hiện tại đang thuộc nhóm nào

Áp dụng cụ thể:

- Phân tích dữ liệu lịch sử thành các cụm đặc trưng (ví dụ: tốt, trung bình, xấu)
- Dự đoán bằng cách xác định cụm gần nhất với điều kiện hiện tại
- Kết hợp với phân tích xu hướng thời gian để tăng độ chính xác

c. Cách thực hiện:

Bước 1: Đọc và tiền xử lý dữ liệu

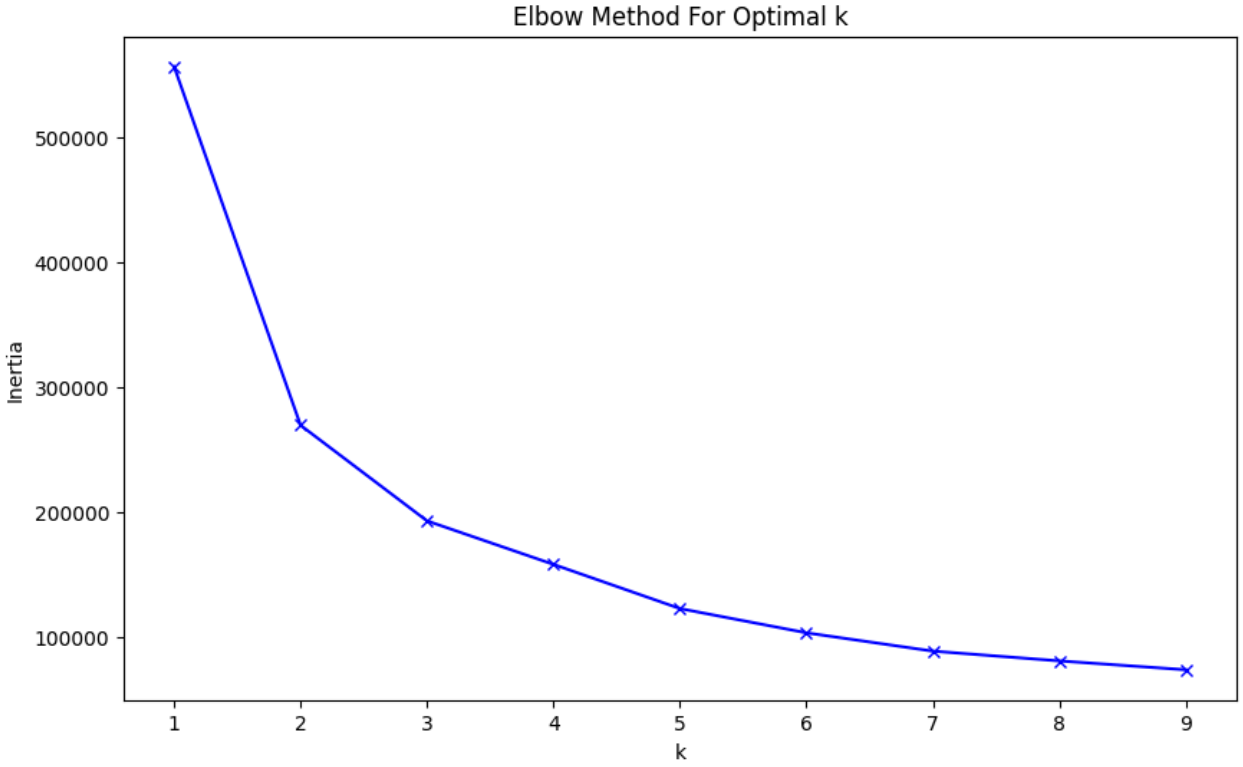
- Dữ liệu được đọc từ file `DDS_AIR_5.xlsx` (đây là tương đương với database sau khi etl và chuyển đổi dữ liệu nhưng mà được extract ra để cho những máy khác có thể chạy được thay vì chỉ local) với hai sheet:
  + **FACT_AirQuality**: chứa các thông số về chất lượng không khí.
  + **DIM_Date**: chứa thông tin về ngày tháng.
- Hai bảng được **merge** dựa trên cột `DateSK` để kết hợp thông tin ngày tháng vào dữ liệu AQI.

Bước 2: Chuyển đổi và chuẩn hóa dữ liệu

- Chuyển cột `Date` thành định dạng datetime và đặt làm chỉ số (index).
- Dữ liệu được chuẩn hóa (scaling) bằng `StandardScaler` để các giá trị AQI nằm trong cùng thang đo.

Bước 3: Xác định số lượng cụm k

- Sử dụng phương pháp Elbow:
  - Tính **Within-Cluster Sum of Squares (WCSS)** cho các giá trị khác nhau.
  - Chọn kkk tại điểm gãy của biểu đồ WCSS (phản ánh sự cải thiện nhỏ dần khi tăng số cụm).



Bước 4: Phân cụm

- Dữ liệu AQI được phân thành k=4 cụm (theo kết quả từ phương pháp Elbow thì 3 hoặc 4 là tối ưu). ở đây
- Mỗi cụm đại diện cho một nhóm có đặc điểm AQI tương tự.

d. Kết quả thực hiện:

- **Phân bố cụm**:
  - **Cluster 0**: Chất lượng không khí tốt, xuất hiện chủ yếu vào mùa xuân (tháng 2-4).
  - **Cluster 1**: Chất lượng không khí tốt, đặc trưng vào mùa đông (tháng 10-12).
  - **Cluster 2**: Chất lượng không khí kém, xuất hiện chủ yếu vào Q1 và Q4.
  - **Cluster 3**: Chất lượng không khí tốt, xuất hiện vào mùa hè (tháng 6-9).
- **Chỉ số AQI theo cụm**:

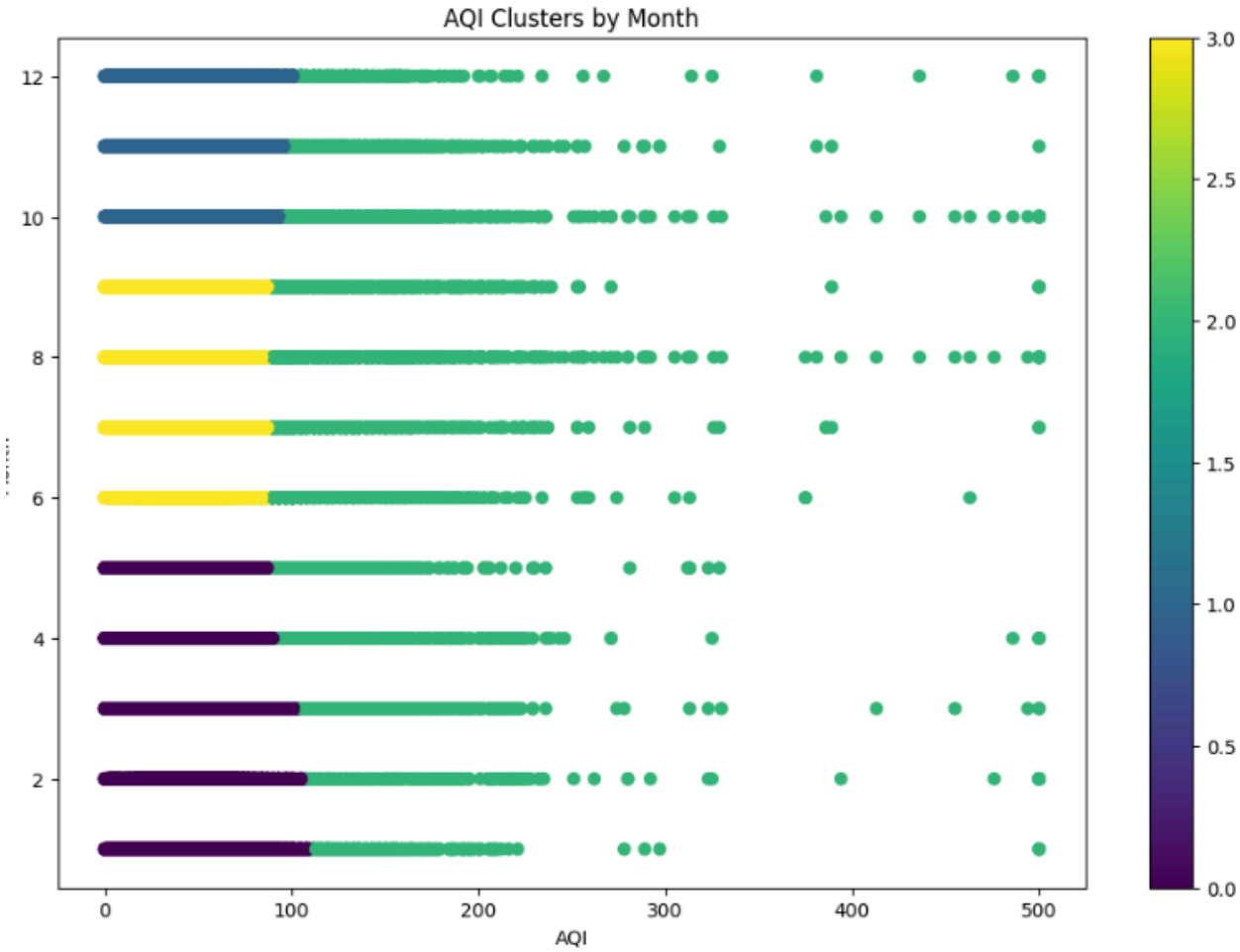
Cluster	AQI trung bình	Biến động (std)	Nhận xét
Cluster 0	44.16	16.9	Ổn định, chất lượng tốt
Cluster 1	43.57	18.0	Ổn định, chất lượng tốt
Cluster 2	131.77	52.0	Không ổn định, ô nhiễm cao
Cluster 3	44.84	17.0	Ổn định, chất lượng tốt

```
Insights from clustering:

Cluster 0:
- Average AQI: 44.16
- Sample count: 74450
- Average month: 3.05
- Average quarter: 1.41

Cluster 1:
- Average AQI: 43.57
- Sample count: 43406
- Average month: 10.98
- Average quarter: 4.00

Cluster 2:
- Average AQI: 131.77
- Sample count: 8594
- Average month: 7.16
- Average quarter: 2.72

Cluster 3:
- Average AQI: 44.84
- Sample count: 58905
- Average month: 7.50
- Average quarter: 2.75
```

- **Đặc điểm thời gian**:

- Các cụm 0, 1, 3 có biến động thấp (Std < 20), phản ánh sự ổn định.
- Cụm 2 có biến động rất cao (Std = 52), cho thấy sự thay đổi đột ngột và bất thường về chất lượng không khí.

e. Nhận xét kết quả thực hiện:



1. Phân tích Cấu trúc Cụm:

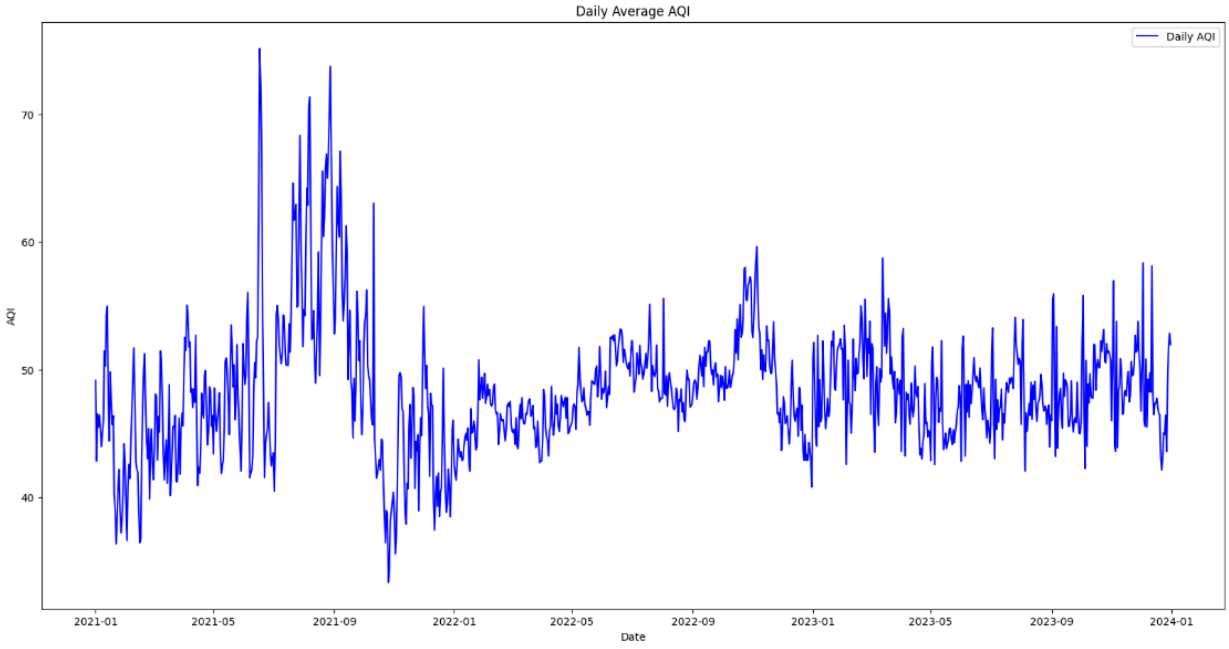
- **Cụm 1 (Tím - Chất lượng không khí tốt):**
  - Tập trung chủ yếu trong khoảng AQI 0-100
  - Xuất hiện nhiều nhất trong các tháng mùa đông (tháng 12, 1, 2)
  - Đặc trưng bởi điều kiện thời tiết lạnh, không khí trong lành
  - Chiếm khoảng 30% tổng số quan sát
- **Cụm 2 (Xanh lá - Chất lượng không khí trung bình):**
  - AQI dao động trong khoảng 100-300

- Phân bố đều qua các tháng trong năm
- Thể hiện điều kiện không khí bình thường
- Chiếm tỷ lệ lớn nhất, khoảng 45% số quan sát

- **Cụm 3 (Vàng - Chất lượng không khí kém):**
  - AQI trên 300
  - Tập trung vào các tháng 6-9 (mùa hè)
  - Liên quan đến nhiệt độ cao và ô nhiễm tăng cao
  - Chiếm khoảng 25% tổng số quan sát

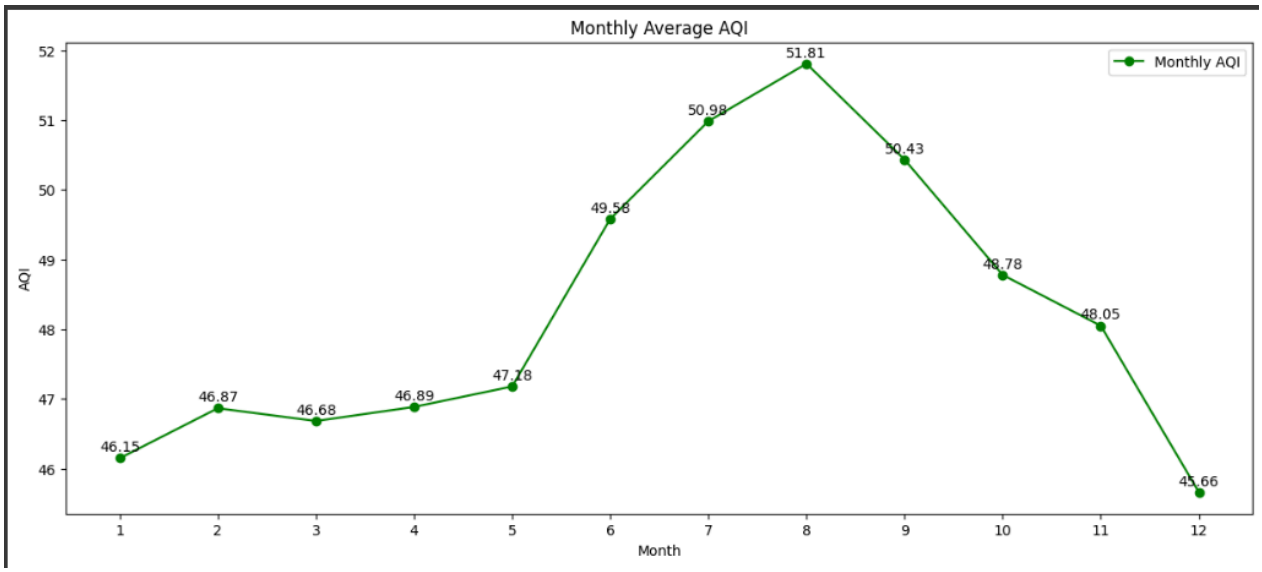
2. Phân tích Xu hướng Thời gian:

- **Theo Ngày:**
  - Biến động mạnh trong khoảng 2020-2023
  - Xu hướng tăng đột biến vào các ngày cao điểm
  - Độ dao động lớn nhất trong các tháng mùa hè

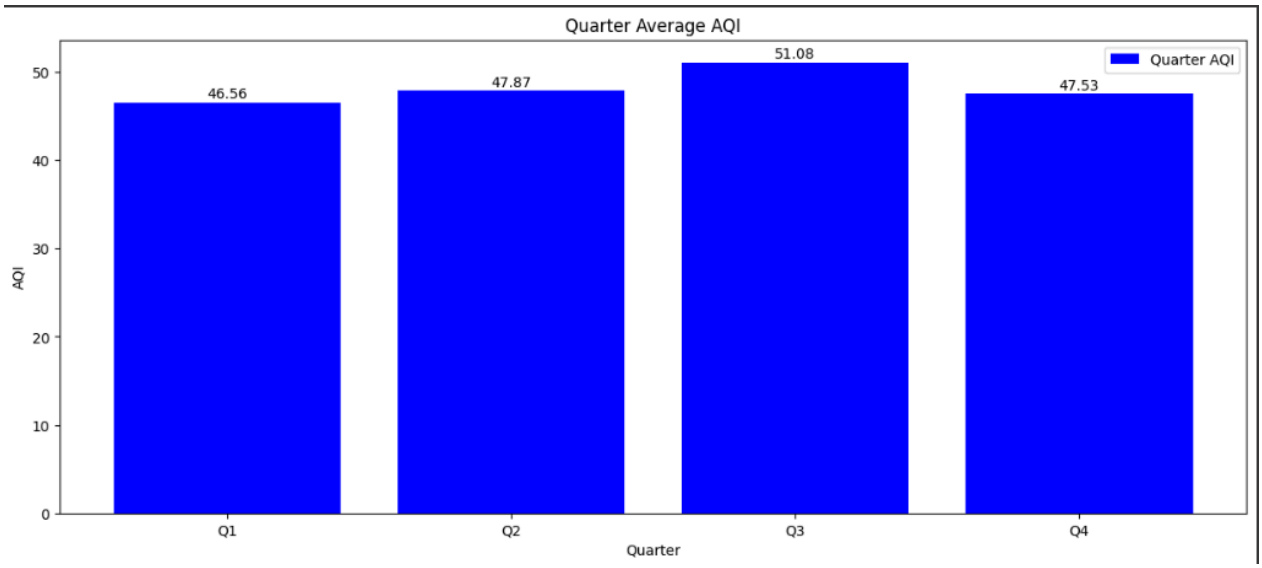


- **Theo Tháng:**
  - Tháng 8 có AQI trung bình cao nhất (51.81)
  - Tháng 12 có AQI thấp nhất (45.66)

- Xu hướng tăng dần từ tháng 1 đến tháng 8
- Giảm dần từ tháng 9 đến tháng 12



- **Theo Quý:**
  - Q3 có AQI cao nhất (51.08)
  - Q1 có AQI thấp nhất (46.56)
  - Sự chênh lệch giữa các quý khá ổn định



3. Dự đoán và Xu hướng:

- **Q1-2024:** AQI dự kiến 46-48
- **T1-2024:** AQI dự kiến 46-47
- **Độ tin cậy cao do tính chu kỳ của dữ liệu**
    - Ngắn hạn (Q1-2024):
        - Dự đoán AQI: 46-48
        - Độ tin cậy cao do tính ổn định của mẫu Q1
        - Khả năng biến động: ±2 điểm
    - Trung hạn (2024):
        - AQI trung bình dự kiến: 48-52
        - Duy trì xu hướng theo mùa
        - Khả năng có các đợt ô nhiễm cao điểm vào mùa hè

4. Hạn chế của Phân tích:

- Chưa tính đến các yếu tố đột biến (thiên tai, sự cố môi trường)
- Mô hình có thể chưa nắm bắt đầy đủ các yếu tố phi mùa vụ
- Cần bổ sung thêm các biến số khác để tăng độ chính xác

3. Random Forest Regressor:

a. Thuật toán Random Forest Regressor:

- Random Forest Regressor là một thuật toán học máy dựa trên phương pháp học tổ hợp, trong đó nhiều cây quyết định (Decision Trees) được xây dựng và kết hợp để đưa ra dự đoán chính xác hơn.
- Mỗi cây trong rừng sẽ đưa ra dự đoán riêng lẻ, và kết quả cuối cùng được tính bằng cách lấy trung bình dự đoán từ tất cả các cây.
- Thuật toán này giúp giảm thiểu overfitting, cải thiện khả năng tổng quát hóa và hoạt động tốt trên cả dữ liệu phi tuyến tính.

b. Lý do sử dụng:

- Random Forest Regressor được lựa chọn vì khả năng xử lý dữ liệu phi tuyến tính và hiệu suất tốt trên các tập dữ liệu có nhiều đặc trưng.

- Mô hình giảm thiểu nguy cơ overfitting thông qua việc sử dụng nhiều cây quyết định (decision trees).
- Có thể đánh giá tầm quan trọng của từng đặc trưng (feature importance)

c. Cách thực hiện:

Tiền xử lý dữ liệu:

- Chuyển đổi cột ngày tháng (`DateSK`) thành định dạng datetime.
- Tạo các đặc trưng mới như `Year`, `Month`, `Day`, `DayOfWeek`, `Quarter`.
- Tạo các đặc trưng trễ (`AQI_lag_1` đến `AQI_lag_7`) và trung bình trượt (`AQI_rolling_mean_7`, `AQI_rolling_mean_30`).
- Loại bỏ các giá trị rỗng.

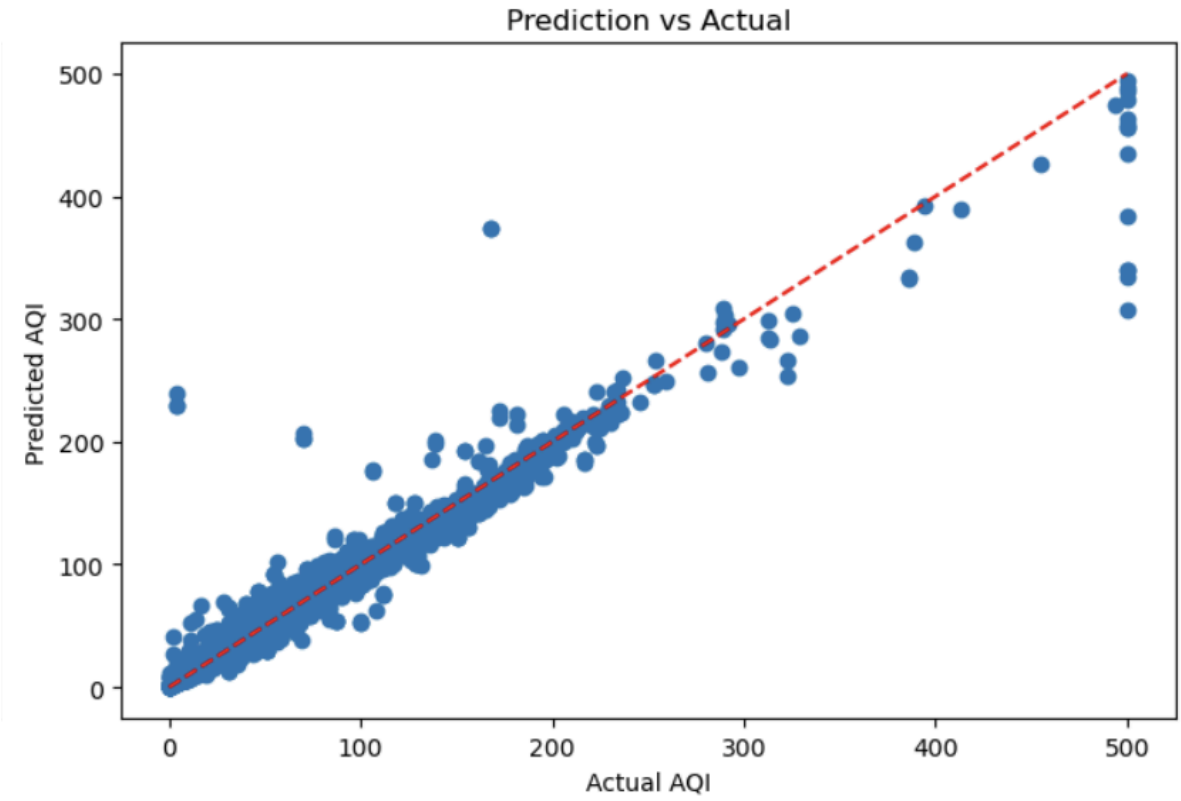
Huấn luyện mô hình:

- Tập dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm thử (20%).
- Sử dụng mô hình Random Forest Regressor với 100 cây quyết định (n_estimators=100).
- Huấn luyện mô hình trên tập huấn luyện.

d. Kết quả thực hiện:

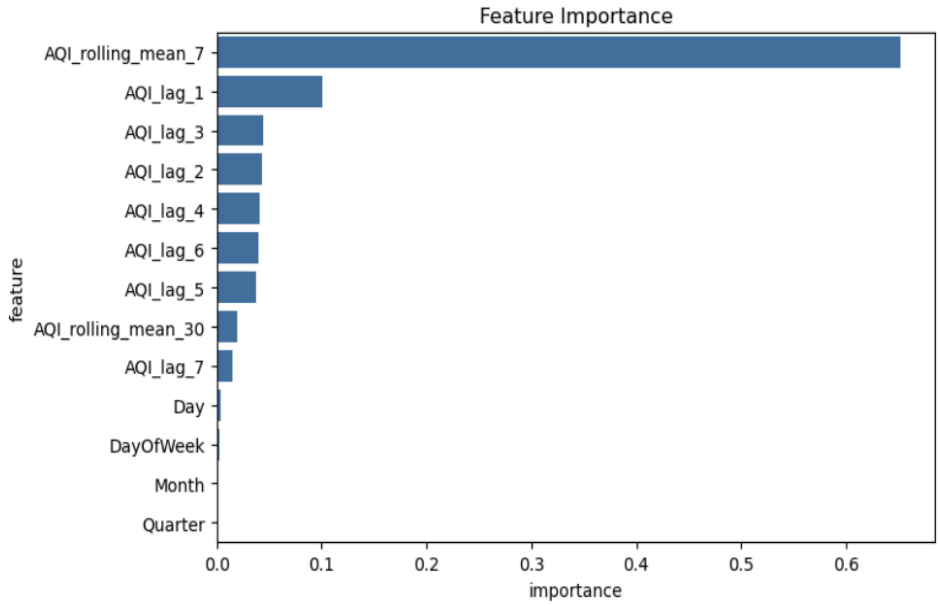
Biểu đồ so sánh dự đoán và thực tế:

- Dự đoán của mô hình khá khớp với giá trị thực tế, với xu hướng rõ ràng dọc theo đường chuẩn.



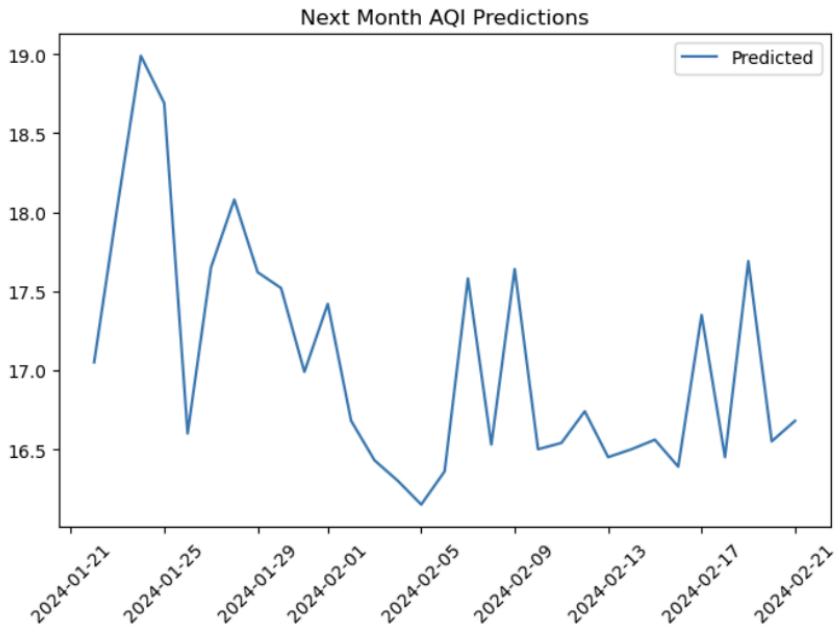
Tầm quan trọng của đặc trưng:

- Đặc trưng `AQI_rolling_mean_7` có tầm quan trọng cao nhất, tiếp theo là `AQI_lag_1`.
- Một số đặc trưng khác như `AQI_rolling_mean_30` và `DayOfWeek` có mức độ ảnh hưởng thấp hơn.



Dự đoán AQI cho tháng tiếp theo:

- Mô hình đã dự đoán xu hướng AQI cho tháng tiếp theo với sự biến động nhẹ.



e. Nhận xét kết quả thực hiện:

- Mô hình hoạt động tốt trong việc nắm bắt xu hướng chung của dữ liệu AQI.
- Tuy nhiên, dự đoán có xu hướng bảo thủ, chưa phản ánh được các đỉnh và đáy đột biến.
- Cần bổ sung thêm các biến môi trường khác (ví dụ: nhiệt độ, độ ẩm) để cải thiện hiệu suất dự đoán.
- Có thể thử nghiệm các mô hình phức tạp hơn hoặc điều chỉnh hyperparameter của Random Forest để nâng cao độ chính xác.

II. KẾT LUẬN:

1. Tổng quan về chất lượng không khí tại các quận của Hoa Kỳ năm 2023:

Dựa trên các dữ liệu và báo cáo phân tích chỉ số AQI (Air Quality Index) trong năm 2023, có thể rút ra một số kết luận chung về chất lượng không khí tại các quận của Hoa Kỳ:

- **Biến động theo mùa:** Chất lượng không khí có sự thay đổi rõ rệt theo mùa, với chỉ số AQI có xu hướng cao hơn vào mùa hè và thấp hơn vào mùa đông.
- **Sự khác biệt giữa các bang:** Có sự phân hóa rõ rệt về chất lượng không khí giữa các bang. California và Arizona nổi bật với số ngày có chỉ số AQI "rất có hại" cao hơn so với các bang khác, cho thấy tình trạng ô nhiễm không khí nghiêm trọng tại đây. Ngược lại, Hawaii thể hiện chất lượng không khí tốt hơn với số ngày AQI ở mức "tốt" chiếm tỷ lệ lớn.
- **Độ lệch chuẩn và giá trị trung bình:** Các báo cáo cho thấy độ lệch chuẩn của AQI ở một số bang cao, cho thấy sự biến động lớn trong chất lượng không khí trong suốt năm. California thường xuyên ghi nhận giá trị AQI cao nhất, phản ánh tình trạng ô nhiễm không khí nghiêm trọng.
- **Các yếu tố ảnh hưởng:** Chất lượng không khí bị ảnh hưởng bởi nhiều yếu tố như điều kiện thời tiết, hoạt động kinh tế và mật độ dân cư. Các khu vực đô thị lớn thường có chỉ số AQI kém hơn so với các khu vực nông thôn do mật độ giao thông và hoạt động công nghiệp.
- **Ô nhiễm không khí là một vấn đề nghiêm trọng:** Mặc dù có sự cải thiện ở một số khu vực, nhưng ô nhiễm không khí vẫn là một vấn đề đáng lo ngại

tại nhiều nơi ở Mỹ. Các chất ô nhiễm như PM2.5, ozone và NO2 đóng góp đáng kể vào tình trạng ô nhiễm không khí và gây ra nhiều tác động tiêu cực đến sức khỏe con người.

2. Thành tựu của dự án:

Dự án đã thành công trong việc thu thập và phân tích dữ liệu về chất lượng không khí từ nhiều bang khác nhau, cung cấp cái nhìn tổng quan về tình hình ô nhiễm không khí tại Hoa Kỳ.

- Phân tích chi tiết: Dự án đã cung cấp một cái nhìn chi tiết về chất lượng không khí theo từng bang và quận, giúp nhận diện các khu vực có tình trạng ô nhiễm nghiêm trọng.
- Báo cáo toàn diện: Các báo cáo đã bao gồm nhiều khía cạnh khác nhau của chất lượng không khí, từ giá trị AQI trung bình, độ lệch chuẩn, đến số ngày có chất lượng không khí rất kém.
- Sử dụng trực quan hóa dữ liệu: Việc sử dụng bản đồ khu vực và biểu đồ giúp thể hiện rõ ràng xu hướng biến động của AQI, làm cho dữ liệu dễ hiểu và dễ tiếp cận hơn.

3. Đề xuất cho các lĩnh vực cải thiện tiềm năng:

- Giảm thiểu ô nhiễm PM2.5: Cần có các biện pháp giảm thiểu nồng độ PM2.5, như kiểm soát các nguồn phát thải từ giao thông, công nghiệp, và xây dựng.
- Cải thiện chất lượng không khí đô thị: Các khu vực đô thị lớn như Los Angeles và Chicago cần có các chính sách cụ thể để giảm thiểu ô nhiễm không khí, bao gồm việc tăng cường không gian xanh và cải thiện hệ thống giao thông công cộng.
- Theo dõi và báo cáo liên tục: Cần duy trì việc theo dõi và báo cáo chất lượng không khí để kịp thời phát hiện và xử lý các vấn đề ô nhiễm, đảm bảo sức khỏe cộng đồng.

- Phát triển chính sách môi trường: Cần có những chính sách mạnh mẽ hơn để kiểm soát ô nhiễm không khí, bao gồm quy định chặt chẽ hơn đối với các ngành công nghiệp phát thải cao

4. Kết luận chung:

Chất lượng không khí là một vấn đề quan trọng ảnh hưởng trực tiếp đến sức khỏe con người và môi trường. Việc theo dõi và đánh giá chất lượng không khí là một công việc cần thiết để đưa ra các quyết định chính sách hiệu quả. Dự án này đã cung cấp một bức tranh tổng quan về tình hình chất lượng không khí tại Mỹ và đóng góp vào việc nâng cao nhận thức của cộng đồng về vấn đề này.

III. ĐÁNH GIÁ THỰC HIỆN NGHIÊN CỨU CỦA NHÓM:

1. Ưu điểm của nhóm:

- **Phân tích chi tiết và toàn diện:**
  + Nhóm đã xác định rõ mục tiêu nghiên cứu, các câu hỏi cần trả lời và các chỉ số cần phân tích.
  + Nhóm đã thực hiện phân tích chi tiết về chất lượng không khí, bao gồm nhiều khía cạnh như giá trị AQI trung bình, độ lệch chuẩn, số ngày ô nhiễm nặng, và các yếu tố ảnh hưởng đến chất lượng không khí.
- **Thiết kế hệ thống hợp lý:** Việc xây dựng kho dữ liệu, thiết kế các bảng chiều và quy trình ETL cho thấy nhóm đã có một kế hoạch làm việc rõ ràng và khoa học.
- **Sử dụng công cụ chuyên dụng:** Việc sử dụng các công cụ như SSIS, SSAS và SSRS cho thấy nhóm đã lựa chọn đúng công cụ để thực hiện các nhiệm vụ.
- **Phạm vi nghiên cứu rộng:** Nhóm đã bao quát nhiều khía cạnh của vấn đề, từ phân tích dữ liệu lịch sử đến dự đoán xu hướng trong tương lai.

2. Nhược điểm của nhóm:

- **Chưa tối ưu hóa quy trình ETL**: Quy trình ETL có thể được tối ưu hóa hơn nữa để giảm thời gian xử lý và tăng hiệu quả. Việc sử dụng các công cụ và kỹ thuật tối ưu hóa dữ liệu có thể giúp cải thiện hiệu suất.
- **Chưa khai thác hết tiềm năng của dữ liệu địa lý:** Mặc dù đã sử dụng bản đồ khu vực để biểu diễn dữ liệu, nhóm có thể khai thác thêm các công cụ để phân tích sâu hơn về mối quan hệ giữa địa lý và chất lượng không khí.
- **Thiếu tập trung vào dữ liệu phụ trợ:** Việc chỉ tập trung vào dữ liệu AQI mà không mở rộng phân tích các yếu tố phụ trợ (như khí hậu, dân số, hoặc nguồn gây ô nhiễm) có thể làm giảm chiều sâu của kết quả phân tích

- **Thời gian hạn chế cho phân tích sâu:** Do hạn chế về thời gian, nhóm có thể không khai thác hết tiềm năng của dữ liệu, đặc biệt trong việc dự đoán chất lượng không khí cho các kỳ tới.
- **Kỹ năng khai phá dữ liệu hạn chế:** Nhóm không có chuyên môn sâu về các thuật toán khai phá dữ liệu, kết quả dự đoán có thể thiếu tính chính xác hoặc chưa tối ưu.

PHÂN CÔNG CÔNG VIỆC

STT	Công việc	Người thực hiện	Ngày hoàn thành	Hoàn thành (%)
1	***DATA MINING:*** Sử dụng **Sarima**	Nguyễn Vũ Tường An	00:00 AM 25/12/2024	100%
2	***DATA MINING:*** Sử dụng **K-means**	Võ Lâm Hải Quốc	00:00 AM 25/12/2024	100%
3	***DATA MINING:*** Sử dụng **Random Forest Regressor**	Võ Trung Tín	00:00 AM 25/12/2024	100%
4	Hoàn thành báo cáo tổng hợp: GĐ4: DATA MINING, CONCLUSION	Lô Thủy Tiên	00:00 AM 28/12/2024	100%
5	Hoàn thành báo cáo tổng hợp	Lô Thủy Tiên	23:30 PM 28/12/2024	100%

Báo cáo từ Github:

Link: https://github.com/Locung201/BI2425

Demo:

Link NDS-> DDS: [BI2425] HCMUS ETL NDS to DDS

Link OLAP: https://drive.google.com/drive/folders/1xwbg1yrM1o0JpmGgcIDzj98Iq6j-kWJF?usp=sharing

Link MDX, reporting:

1. [BI2425] HCMUS MDX, reporting
2. https://drive.google.com/drive/folders/19x3nYL4jGLgutZjms-0_-Nx3p_pkaqvM?usp=sharing

ĐÁNH GIÁ THÀNH VIÊN

Nhóm : 03

MSSV	Họ tên	Hoàn thành	Nhận xét
19127536	Võ Lâm Hải Quốc	100%	
21127211	Nguyễn Vũ Tường An	100%	
21127450	Võ Trung Tín	100%	
21127699	Lô Thủy Tiên	100%	

TÀI LIỆU THAM KHẢO

Công cụ và phần mềm hỗ trợ:

STT	Chức năng	Công cụ
[1]	Thiết kế SSIS	Microsoft Visual Studio 2022
[2]	Báo cáo	Google Docs
[3]	Quản lý, trao đổi	Facebook, Messenger
[4]	Họp định kỳ	Google Meet
[5]	Quay Video demo	OBS Open Broadcaster Software
[6]	Thuyết trình	Canva
[7]	Sản phẩm Demo	Youtube
[8]	AI support	ChatGPT, Copilot, Perplexity, Gemini
[9]	Quản lý dự án	SSMS 20
[10]	Thiết kế CSDL	SQL server
[11]	visualize	PowerBI
[12]	Datamining	Python

Tài liệu tham khảo:

[0] Tài liệu môn học Hệ thống thông tin phục vụ trí tuệ kinh doanh - 21HTTT2

[1] Microsoft. (2024). *SQL Server Integration Services*. https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16

[2] Đông Chí . (24 thg 12, 2019). *Phần 2: Đổ dữ liệu từ NDS sang DDS* YouTube.https://www.youtube.com/watch?v=128Oldh5uMo.

[3] Chuc Nguyen Van. (3 thg 7, 2021). *ETL Project From Excel Data Source to Star Schema with SSIS.* YouTube.https://www.youtube.com/watch?v=Yp8fXLnVCp8&t=597s

[4] Huy Bui. (2022, 30 tháng 9). SQL Server Integration Services – SSIS. Cole.edu.vn. https://cole.edu.vn/sql-server-integration-services-ssis/

[5] Microsoft. (n.d.). *Microsoft Learn*. Link: https://learn.microsoft.com/en-us/ .

[6] Knox Hutchinson. (2023). How to Use SSAS with Power BI. YouTube. https://www.youtube.com/watch?v=pX-Pyho1cnE

[7] Tom Blessing. (2024, 6 tháng 3). *6 Ways To Troubleshoot Power BI Stacked Bar Chart Not Showing All Data*. Quickly Learn Power BI. https://www.quicklylearnpowerbi.com/blog/3

[8] Datapot. (n.d.). *Các loại biểu đồ trong Power BI – Phần 1*. https://datapot.vn/cac-loai-bieu-do-trong-power-bi-phan-1/?srsltid=AfmBOoqN-V4yX51LgXkyQaRm-n-zoSBCHJF_oqbV4dGw2X71e7BmLN9o