BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

**TRỰC QUAN HÓA DỮ LIỆU**


# BÁO CÁO BÀI TẬP MÔN HỌC

## BÀI TẬP: DV-TW-Lab02

**Giảng viên hướng dẫn** :Lê Nhựt Nam

**Lớp** :CQ2022/4

**Nhóm sinh viên thực hiện** : Đinh Viết Lợi- 22120188

Nguyễn Trần Lợi- 22120190

Nguyễn Nhật Long-22120194

Trần Minh Tâm- 22120323

*Hồ Chí Minh, ngày 20 tháng 11 năm 2024*

# MỤC LỤC

## Mục lục

# PHẦN 1: BÁO CÁO NHÓM

## I. Thông tin thành viên

| MSSV | Họ và tên | Email | Vai trò chính |
|---|---|---|---|
| 22120188 | Đinh Viết Lợi | 22120188@student.hcmus.edu.vn | Trưởng nhóm |
| 22120190 | Nguyễn Trần Lợi | 22120190@student.hcmus.edu.vn | |
| 22120194 | Nguyễn Nhật Long | 22120194@student.hcmus.edu.vn | |
| 22120323 | Trần Minh Tâm | 22120323@student.hcmus.edu.vn | |

## II. Phân chia công việc

Đồ án được xây dựng và hoàn thành trong khoảng thời gian từ : 13/11/2024 đến 24/11/2024.

| Nội dung công việc | Người phụ trách | Thời gian bắt đầu | Thời gian kết thúc | Kết quả mong muốn |
|---|---|---|---|---|
| Tìm nguồn dữ liệu | Tất cả thành viên | 13/11 | 13/11 | Nguồn dữ liệu phải đảm bảo chất lượng lẫn số lượng, độ uy tín cao. Đồng thời tập dữ liệu phải phù hợp với mục đích trực quan hóa theo chủ đề time series. |
| Quản lý, điều hành nhóm | Viết Lợi | 13/11 | 24/11 | Nhóm trưởng phải theo dõi tiến độ công việc, liên tục cập nhật tình hình của các thành viên và phân chia công việc phù hợp. |

| | | | | |
|---|---|---|---|---|
| Thảo luận, tìm hiểu các nội dung có thể khai thác | Tất cả thành viên | 15/11 | 15/11 | Các câu hỏi phải nhằm trả lời cho các nội dung chính đáng, quan trọng.<br>Nội dung các câu hỏi phải được dự đoán độ chính xác để được kiểm tra.<br>Nội dung phải phù hợp với nội dung chính của bài thực hành "Time Series" |
| Tìm hiểu và tiền xử lý dữ liệu | Viết Lợi | 16/11 | 17/11 | Ý nghĩa của dữ liệu phải được khai thác rõ ràng.<br>Tiền xử lý dữ liệu phải đảm bảo không làm mất mát dữ liệu, kiểu dữ liệu phải phù hợp.<br>Nội dung nằm tại tập tin preprocessing.iypnb & datetimeStatistic.ipynb |
| Khai thác các nội dung tổng quan đến chỉ số AQI | Viết Lợi | 17/11 | 22/11 | Nội dung phần bài làm được lưu tại tập tin Shilin_AQI.ipynb |
| Khai thác các nội dung đến các tương quan giữa các chất khí ô nhiễm | Trần Lợi | 17/11 | 22/11 | Nội dung phần bài làm được lưu tại tập tin<br>air_concentration_and_<br>status.ipynb |
| Khai thác các nội dung liên quan đến xu hướng biến đổi theo thời gian | Minh Tâm | 17/11 | 22/11 | Nội dung phần bài làm được lưu tại tập tin<br>analyse_air_quality_<br>with_aqi_index.ipynb |

| | | | | |
|---|---|---|---|---|
| Khai thác các nội dung về mối quan hệ giữa các chất ô nhiễm và chất lượng không khí | Nhật Long | 17/11 | 22/11 | |
| Xây dựng mô hình dự đoán | Nhật Long | 17/11 | 22/11 | |
| Kiểm tra chất lượng kết quả của toàn bài tập lớn. | Trần Lợi & Minh Tâm & Nhật Long | 22/11 | 23/11 | |
| Viết báo cáo bài tập nhóm | Viết Lợi | 22/11 | 23/11 | |

# PHẦN 2: TỔNG QUAN ĐỒ ÁN

## I. Yêu cầu đồ án

### a. Giới thiệu đồ án

- Đồ án được thực hiện dựa trên yêu cầu của đồ án DV–TW– Lab02 thuộc lớp "Trực quan hóa dữ liệu - CQ2022/4" trường đại học Khoa Học Tự Nhiên thuộc Đại học Quốc gia TP.HCM học kỳ I năm học 2024-2025.
- Tiếp nối nội dung của đồ án Lab01, đồ án Lab02 xoay quanh các nội dung liên quan đến kĩ năng và công cụ giúp trực quan hóa dữ liệu. Bên cạnh đó là làm quen với kiểu tập dữ liệu time series, phân tích các khái niệm, nội dung cần khai thác đối với kiểu dữ liệu này.
- Kết quả của đồ án là các mã nguồn được chú thích đầy đủ tên biểu đồ, ý nghĩa, kết luận rút ra từ dữ liệu, mô tả tổng quan về các khoảng thời gian được khai thác. Nội dung của các biểu đồ xoay quanh về chất lượng không khí tại quận Sỹ Lâm (Shilin) thuộc thành phố Đài Bắc, Đài Loan.

### b. Yêu cầu đồ án

- Bài làm không có có báo cáo sẽ không được chấm điểm.
- Các thành viên không đóng góp cho dự án sẽ không nhận được điểm.
- Các nguồn tham khảo (nếu có) cần được ghi đầy đủ trong báo cáo ở phần Tài liệu tham khảo. Lưu ý rằng cần phân biệt rõ giữa việc tham khảo và đạo văn.
- Cá nhân hoặc nhóm nào vi phạm gian lận và không trung thực sẽ nhận 0 điểm trong khóa học.
- Đối với buổi thực hành này, có một số hạn chế cụ thể mà chúng tôi yêu cầu bạn tuân thủ nghiêm ngặt:
  - Không sử dụng các giải pháp phần mềm nâng cao như Tableau cho mục đích trực quan hóa dữ liệu.
  - Nếu muốn sử dụng các thư viện bổ sung, vui lòng tham khảo ý kiến giảng viên trước để nhận được sự chấp thuận.
  - Trong khi việc áp dụng các thuật toán máy học đơn giản có thể cung cấp những hiểu biết sâu sắc hơn về dữ liệu của bạn, đây là yêu cầu tùy chọn và không phải là yêu cầu bắt buộc.

## II. Mức độ hoàn thành

- **Tổng quan**: nhóm hoàn thành đầy đủ các nội dung cơ bản được yêu cầu và thực hiện tìm hiểu vào các phần nội dung được khuyến khích. Bài làm được đầu tư cẩn thận về chất lượng và số lượng, báo cáo tổng hợp được tất cả quá trình làm việc và kết quả đạt được của nhóm.

- Bộ câu hỏi được chuẩn bị có tính phổ biến và ứng dụng cao, phù hợp với nhu cầu phân tích.
- Nhóm sử dụng đa dạng các loại biểu đồ để trực quan dữ liệu nhằm giải đáp các vấn đề được nêu ra. Tất cả thông tin cần thiết cho việc nắm bắt biểu đồ đều được trình bày rõ ràng, đầy đủ.

  ➔ Mức độ hoàn thành: 100%
- **Hạn chế:** nhóm chưa thật sự khai thác được hết tiềm năng của loại biểu đồ time series dẫn tơi việc các biểu đồ không có sự đột phá mạnh mẽ, chỉ xoay quanh mức cơ bản và vận dụng.

# PHẦN 3: BÁO CÁO KẾT QUẢ

## I. Thu thập dữ liệu

- Nguồn dữ liệu: Kaggle- Taiwan Air Quality Index Data 2016~2024.
- Ô nhiễm không khí và bảo vệ khí hậu là một trong những vấn đề được nhiều quốc gia và tổ chức quan tâm, trở thành một trong những chủ đề trọng tâm trong tình hình kinh tế phát triển nhanh chóng tại nhiều quốc gia trong thế kỷ 21. Chất lượng không khí trong những năm gần đây đang có xu hướng trở nên trầm trọng hơn bao giờ hết khi liên tục ghi nhận nhiều khoảng thời gian mức độ chất lượng không khí giảm mạnh và xuất hiện nhiều ca bệnh liên quan tới hô hấp. Nhiều quốc gia và tổ chức đã thực hiện nhiều quan sát, phân tích chất lượng không khí nhằm tìm ra giải pháp hạn chế sự tác động của không khí kém chất lượng.
- Tập dữ liệu trên được thu thập từ trang web Kaggle, một nền tảng chuyên cung cấp các nguồn tập tài liệu chất lượng với nhiều chủ đề. Tập dữ liệu " Taiwan Air Quality Index Data 2016-2024" thu thập 24 quan sát một ngày, liên tục từ ngày 25/11/2016 đến 31/8-2024 trên địa bàn của tất cả tỉnh thuộc Đài Loan.
- Nhằm giảm thiểu sai sót khi phân tích số lượng lớn đồng thời tập trung phân tích chất lượng hơn, nhóm chỉ tập trung phân tích một quận thuộc một tỉnh của Đài Loan. Đối tượng được nhóm quan sát phải đảm bảo mang lại nhiều giá trị từ kết quả phân tích. Từ việc phân tích các thông tin về văn hóa, địa lý, chính trị và kinh tế bên ngoài, nhóm đã quyết định chọn quận Sỹ Lâm (Shilin) thuộc tỉnh Đài Bắc (Taipei). Taipei là một quận hành chính quan trọng, có diện tích lớn và có nhiều địa điểm du lịch thuộc về thành phố Đài Bắc-Thủ đô của Đài Loan.
- Nguồn dữ liệu trên được sự cho phép tải và nghiên cứu bởi tác giả của nguồn dữ liệu cho mục đích nghiên cứu, tác giả khuyến khích sử dụng nguồn dữ liệu và gợi ý một số cách khai thác tiềm năng của tập dữ liệu.

## II. Khám phá dữ liệu

- Bộ dữ liệu được lưu dưới dạng một tập tin csv chứa một bảng thông tin về các quan sát được thực hiện đo đạc về đặc điểm của không khí sau mỗi giờ tại các địa điểm khắp Đài Loan.
- Mỗi dòng dữ liệu là một quan sát được thực hiện tại một thời gian cụ thể tại một thời điểm cụ thể tại Đài Loan. Mỗi quan sát này sẽ lưu lại các đặc điểm về các chỉ số đặc điểm của một số khí nhất định kèm theo đánh giá tổng quan của người thực hiện về chất lượng không khí tại thời điểm đó. Việc trùng lắp dữ liệu đối với thời gian và địa điểm quan sát sẽ được xử lý trước khi thực hiện phân tích. Các dòng dữ liệu phải cùng thực hiện các quan sát có nghiệp vụ tương tự nhau, các thông số, kiểu dữ liệu phải phù hợp với tất cả các dòng xung quanh.
- Nếu tồn tại các dòng dữ liệu nhiễu hoặc dữ liệu không phù hợp với cột dữ liệu tương ứng sẽ gây khó khăn cho việc phân tích và loại bỏ các giá trị này để đảm

bảo dữ liệu phù hợp trước khi phân tích.

- Bộ dữ liệu bao gồm một tập thuộc tính đa dạng về cả giá trị lẫn kiểu dữ liệu, cung cấp cho người sử dụng một nguồn thông tin đáng giá và cái nhìn tổng quan về vấn đề khi thực hiện phân tích. Các thuộc tính của dữ liệu và kiểu dữ liệu của chúng:
  - date- Chuỗi (object): thời điểm thực hiện quan sát chi tiết đến năm tháng ngày giờ.
  - sitename- Chuỗi (object): quận của thành phố.
  - county- Chuỗi (object): thành phố thuộc Đài Loan.
  - aqi- Số nguyên (int): chỉ số chất lượng không khí air quality index.
  - pollutant- Chuỗi (object): tình trạng ô nhiễm chính.
  - status- Chuỗi (object): đánh giá chất lượng không khí theo cấp độ.
  - so2- Số nguyên (float): nồng độ khí SO2, đơn vị ppb (parts per billion).
  - co- Số nguyên (float): nồng độ khí CO, đơn vị ppm (parts per million).
  - o3- Số nguyên (float): nồng độ khí O3, đơn vị ppb (parts per billion).
  - O3_8hr: nồng đồ khí O3 trung bình đo được sau 8 giờ liên tiếp.
  - pm10: nồng độ các hạt bụi mịn có đường kính nhỏ hơn 10 µm, đơn vị µg/m3.
  - pm2.5: nồng độ các hạt bụi mịn có đường kính nhỏ hơn 2.5 µm, đơn vị µg/m3
  - no2- Số nguyên (float): nồng độ khí NO2, đơn vị ppb (parts per billion).
  - nox- Số nguyên (float): nồng độ khí NOX, đơn vị ppb (parts per billion).
  - no- Số nguyên (float): nồng độ khí NO, đơn vị ppb (parts per billion).
  - windspeed- Số nguyên (float): tốc độ gió, đơn vị m/s.
  - winddirec- Số nguyên (int): hướng gió
  - pm2.5_avg- Số nguyên (float): trung bình động của bụi pm2.5, đơn vị đơn vị µg/m3.
  - pm10_avg- Số nguyên (float): trung bình động của bụi pm10, đơn vị đơn vị µg/m3.
  - so2_avg- Số nguyên (float): trung bình động của khí SO2, đơn vị đơn vị µg/m3.
  - longitude- Số nguyên (float): kinh độ của địa điểm thực hiện quan sát.
  - latitude- Số nguyên (float): vĩ độ của địa điểm thực hiện quan sát.
  - liteid- Số nguyên(int): mã địa điểm thực hiện quan sát.
- Nhìn chung hầu như tất cả các thuộc tính đều có kiểu dữ liệu phù hợp với nhu cầu hiểu về dữ liệu khi phần lớn đều là kiểu dữ liệu số nguyên. Đối với một số thuộc tính có giá trị chuỗi cũng hỗ trợ người phân tích dễ phân loại, đếm các quan sát.
- Tuy nhiên dữ liệu cũng tồn tại rất nhiều hạn chế, quá nhiều thuộc tính, nhiều thuộc tính khó nắm bắt, đơn vị của các thuộc tính cũng cần được lưu tâm. Tập dữ liệu nên cần trải qua nhiều bước tiền xử lý để thực sự có thể sử dụng. Một số

công việc chính nhóm thực hiện với bộ dữ liệu bao gồm:

- Trích xuất dữ liệu của Shilin, Đài Bắc: do tập dữ liệu quá lớn, nhóm quyết định chỉ chọn ra một quận duy nhất từ một thành phố để có thể tập trung khai thác hiệu quả hơn.
- Kiểm tra và xóa các dòng bị trùng lắp dữ liệu: tránh dữ liệu nhiễu.
- Loại bỏ một số thuộc tính không cần thiết: bộ thuộc tính này bao gồm [‘latitude’, ‘longtitude’, ‘siteid’, ‘unit’, ‘county’ do đã xác định được quận thực hiện quan sát duy nhất. Các thuộc tính còn lại tuy không được sử dụng toàn bộ nhưng không có vấn đề nghiêm trọng về việc ‘missing values’.
- Xử lý missing values: nhóm xóa đi các hàng có tồn tại missing values do tất các quan sát này đều được thực hiện trong ngày nên không có quá nhiều biến động. Nhóm thực hiện kiểm tra lại sau khi xóa, biểu đồ cho thấy không có bất kì ngày nào được xóa khỏi chuỗi thời gian liên tiếp. Đối với dữ liệu bị trống ở thuộc tính ‘pollutant’ nhóm thay thế các giá trị này bởi “Normal”.

# III. Trực quan và ý nghĩa

- Bộ dữ liệu sau khi tiền xử lý tuy có rất nhiều quan sát và thuộc tính giúp mang lại rất nhiều trực quan có thể thực hiện tuy nhiên nhóm chỉ lựa chọn một số biểu đồ hỗ trợ trực quan các vấn đề đáng quan tâm. Các vấn đề nhóm cho rằng quan trọng:
  - Chỉ số AQI tại Shilin, Đài Loan những năm gần đây có đặc điểm như thế nào: biểu đồ cột kiểm tra chỉ số AQI trung bình tại Shilin theo quý hằng năm.
  - Mật độ bụi PM2.5 tại Shilin, Đài Loan theo thời gian có chuyển biến như thế nào: biểu đồ đường, kiểm tra số ngày mật độ bụi này lớn hơn 40 ug/m3.
  - Đặc điểm không khí nào quyết định chỉ số AQI: biểu đồ nhiệt thể hiện ma trận tương quan giữa AQI và một số khí phổ biến. Kết quả cho biết AQI được đánh giá dựa trên chất khí nào.
  - Tình hình đặc điểm không khí nói chung tại Shilin theo thời gian: biểu đồ đường thể hiện sự tăng giảm của các loại khí theo thời gian.
  - Chất lượng không khí tại Shilin theo thời gian: biểu đồ cho biết các chuyên gia đánh giá tình hình không khí tại Shilin, liệu phần lớn thời gian nguồn khí ở đây có được coi là “tốt” không.
  - Các chất khí ô nhiễm ảnh hưởng thế nào đến chỉ số AQI: biểu đồ box plot cho cái nhìn cụ thể hơn từng chất khí sẽ quyết định chỉ số AQI chuyển biến thế nào.
  - Thời gian nào trong năm Shilin sẽ chịu ảnh hưởng nặng nề nhất của ô nhiễm không khí: Biểu đồ nhiệt phù hợp cho việc so sánh giữa các khoảng thời gian với nhau trong suốt 365 ngày mỗi năm. Kết quả của biểu đồ này còn giúp cho người dân hạn chế ra đường cũng như có biện pháp bảo vệ bản thân phù hợp
  - Thời gian nào trong ngày không khí tại Shilin sẽ có mức ô nhiễm cao nhất: biểu đồ Violin sẽ cho chính quyền tại thành phố có những biện pháp phù hợp nhằm giảm

thiểu mức độ ô nhiễm tại các khoảng thời gian nhất định

- Nồng đồ các chất ô nhiễm vào các thời gian trong ngày: biểu đồ nhiệt hỗ trợ kiểm tra mức độ ô nhiễm vào các thời gian trong ngày, qua đó người dân có thể lựa chọn chế độ sinh hoạt phù hợp.
- AQI thay đổi như thế nào dựa vào các chất ô nhiễm vào các thời gian trong ngày: biểu đồ kết hợp cho cái nhìn cụ thể hơn về các xu hướng thay đổi của AQI trong ngày với các thông tin của một vài khí ô nhiễm.
- Nồng độ các chất ô nhiễm phân bổ như thế nào trong ngày?
- AQI được quyết định bởi các nhân tố nào: biểu đồ radar phù hợp cho việc so sánh các nhân tố ảnh hưởng tới một đối tượng chính.

- Chi tiết về các biểu đồ, loại biểu đồ, ý nghĩa, nguyên nhân sử dụng và kết luận của từng biểu đồ được mô tả tại các file ipynb kèm giải thích của mã nguồn.

# PHẦN 4: TÀI LIỆU THAM KHẢO

## I. Tài liệu nhóm

[1] Đường dẫn github: https://github.com/Dzivilord/DV-TW-Lab02.git

[2] Tài liệu thảo luận: Question+ Problem - Google Docs

## II. Tài liệu tham khảo

[1] Python Data Visualization Cookbook.

[2] Aurélien Géron, 'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow'.

[3] Kyle Gallatin & Chris Albon, 'Machine Learning with Python Cookbook'.

[4] ProgrammingKnowleage, 'Matplotlib Tutorial for Begineers (Python)|Learn Data…' [Trực tuyến]. Địa chỉ: https://www.youtube.com/playlist?list=PLS1QulWo1RIZ3tcrdZodjuXTDTIIXH8EW

[5] GeeksforGeeks, 'Time Series Analysis & Visualization in Python' [Trực tuyến]. Địa chỉ: https://www.geeksforgeeks.org/time-series-data-visualization-in-python/