VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

PROGRESS CHECK-IN REPORT

HCMC WEATHER AND AIR QUALITY ANALYSIS

Class:	Introduction to Data Science, 22KHDL1
Supervisor:	
	TA. Lê Nhựt Nam
Group 10 - CapyData:	
22127234	Cao Hoàng Lộc
22127360	Võ Nguyễn Phương Quỳnh
22127440	Phan Võ Minh Tuệ
22127450	Phạm Anh Văn

Ho Chi Minh City - 2024

Table of Contents

I. CURRENT STATUS OF PROJECT

1.1. Data Collecting

Status: Completed

Description: Weather and air quality data for Ho Chi Minh City (latitude 10.823099, longitude 106.629664) were collected using the Open-Meteo API. The dataset spans October 1, 2022, to September 30, 2024, providing hourly data points.

Tools:

- Libraries: `requests`, `requests-cache`, `retry-requests`, and `pandas`.
- API Integration:
  - Weather data from the Open-Meteo Weather Archive API.
  - Air quality data from the Open-Meteo Air Quality API.

Result:

- 17,544 records each dataset
- Weather data saved as `../data/hcmc_weather_data.csv` (10 features).
- Air quality data saved as `../data/hcmc_air_quality_data.csv` (8 features).

1.2. Data Pre-processing

Status: Completed

Description: The merged dataset combining weather and air quality data has been fully processed to ensure data quality and readiness for analysis. The Data Combination Process:

- ***Handling missing values:*** The dataset contains 17544 non-null entries, so no missing values were found or handled.
- ***Mapping `weather_code` to `weather_status`:*** Replaced numerical `weather_code` with descriptive `weather_status` using WMO Weather Interpretation Codes for better interpretability.
- ***Validation of values:*** Verified that key variables fall within realistic ranges. → All values passed validation checks.
- ***Time continuity:*** No missing time intervals were detected; the dataset maintains hourly continuity.
- ***Outlier detection and handling:***
  - Identified: some example metrics:
    - `temperature_2m: 185 outliers (1.05%)`
    - `precipitation: 2183 outliers (12.44%)`
    - `dew_point_2m: 738 outliers (4.21%)`
  - Categorization:
    - Kept outliers: Represent real phenomena, such as extreme weather or high pollution levels. Examples include high precipitation, high pollutant levels (`pm10`, `pm2_5`, `carbon_monoxide`, etc.).
    - Handled outliers: Outliers deemed unrealistic (e.g., `temperature` or humidity values outside HCMC-specific ranges) were corrected or removed.
- ***Relationship validity:*** No issues detected between variable relationships.
- ***Time zone adjustment:*** Converted timestamps from UTC to UTC+7 (HCMC local time) for proper contextual analysis.

Output:

- 17,544 records (rows)
- 18 attributes (columns): 16 numerical and 2 objective columns.
- Cleaned data saved as `../data/clean_hcmc_waq.csv`.

1.3. Data EDA to Answering Question

Status: Completed

Correlations between features: Based on the correlation matrix, we assign that Weather and air quality in HCMC are linked together. Humidity inversely relates to vapour pressure deficit, while temperature and humidity impact ozone levels, and higher humidity may lower it.

Question 1: ***Is there a correlation between wind speed/direction and PM10 levels? Does wind from certain directions bring higher pollution levels?***

- Higher wind speeds (to the right on the x-axis) generally correlate with lower average PM10 concentrations. This trend suggests that as wind speed increases, it disperses particulate matter, lowering pollution concentrations in the area.
- Winds from the West and Northwest bring the highest PM10 pollution levels. Winds from the East and Southeast are associated with cleaner air, possibly due to fewer pollution sources or more effective pollutant dispersion in those directions.

Question 2: ***How do extreme weather events affect air quality parameters, and what are the lag effects on pollutant concentrations?***

- Extreme weather events have varying impacts on different pollutants, with SO_2 and O_3 being most significantly affected.
- The lag effects persist for considerable periods (1000-4000 hours), suggesting the need for extended monitoring and management strategies post-extreme weather events.
- This information is crucial for public health response planning and air quality management during and after extreme weather conditions.

Question 3: ***Are there distinct seasonal or monthly patterns in air quality metrics?***

- Carbon Monoxide (CO): This pollutant stands out with significantly higher concentrations than the others, fluctuating around 300–700 µg/m^3. It shows a noticeable seasonal trend, with peaks around the last months of the year and lower concentrations in early summer.
- Other Pollutants: The rest of the pollutants (PM10, PM2.5, NO_2, SO_2, O_3, and US AQI) have relatively low concentrations compared to CO, all staying below 100 µg/m^3 and stable through out the year.

Question 4: ***What is the relationship between precipitation and air quality? Does rainfall help reduce pollutant concentrations, and if so, to what extent?***

- Rainfall seems to contribute slightly to the reduction of particulate pollutants (PM10 and PM2.5), but the effect is relatively minor and not strong enough to significantly impact the overall air quality (as reflected by AQI).
- This finding implies that while rain can help reduce pollutant levels, it may not be sufficient to substantially improve air quality on its own.

Question 5: ***Are there specific times of day (morning, afternoon, evening) when pollution levels tend to be higher?***

- Ozone levels peak in the morning, likely due to favorable conditions for ozone formation, such as sunlight availability and less cloud cover.

- PM10 and PM2.5 are consistent with slightly higher concentrations in the afternoon.

Question 6: ***How do temperature patterns vary across different time periods (daily, monthly)? Are there significant anomalies in temperature trends?***

- While actual temperatures show moderate fluctuations, apparent temperatures can exceed actual temperatures by up to 10°C during peak 4-5-6 months, indicating significant heat stress conditions.
- The presence of strong daily cycles and monthly patterns, combined with notable anomalies in early 2024, suggests a changing temperature regime that could have important implications for urban planning and public health considerations.

1.4. Data Modeling

Status: On-going

Model 1: US-AQI Prediction

- Introduction: The goal is to predict **US-AQI** (Air Quality Index). The current aim is to evaluate the compatibility and performance of several popular time series prediction models with the dataset, with the expectation that model accuracy improves with increasing complexity and to identify the most effective model type and preprocessing approach for optimal training performance.

- Modelling approaches:

- **Approach 1: ARIMA**
  - **Objective:** Use ARIMA for US-AQI prediction due to the dataset's univariate nature.
  - **Outcome:** Achieved an MSE of **541.79**, but the results suggest it may not be suitable for accurate air quality forecasting.
  - **Evaluation:** While ARIMA fits the data structure, its performance indicates limited applicability for practical air quality prediction.

- **Approach 2: SARIMA**
  - **Objective:** Extend ARIMA to include seasonal components to better capture seasonality in the dataset.
  - **Outcome:** Result in an MSE of **610.64**, despite accounting for seasonality.
  - **Evaluation:** Slightly underperforming and potentially due to the use of suboptimal parameters for the model.

- **Approach 3: Random Forest Regressor**
  - **Objective:** Train a RandomForest model using all available features, including lagged variables and weather data.
  - **Outcome:** Random Forest performed exceptionally well with an MSE of **0.18**.
  - **Evaluation:** The results are impressive; however, it is likely due to model overfitting.

- **Approach 4: XGBoost Regressor**
  - **Objective:** Train an XGBoost model using all features, including lagged variables and rolling statistics.
  - **Outcome:** XGBoost achieved an MSE of **648.55**, but the errors remained significant despite using advanced modeling techniques.

  - **Evaluation:** The model's performance suggests that further tuning and feature engineering are necessary to improve prediction accuracy.

- **Approach 5: LSTM**
  - **Objective:** Train an LSTM model to capture sequential dependencies in the dataset.
  - **Outcome:** LSTM achieved an MSE of **745.24**, indicating no improvement in prediction accuracy despite its capability to model temporal dependencies.
  - **Evaluation:** Extensive tuning and computational resources did not yield better results, suggesting limitations in applying LSTM to this dataset without additional data or adjustments.

- **Overall Review:** The models tested did not achieve the desired prediction accuracy, as the error rates remained high and did not improve with increasing model complexity. This is likely due to suboptimal data preprocessing, such as inadequate handling of data types or scaling issues, which may have hindered the models' ability to learn effectively.

- **Future improvements:**
  - Try to find a better way to preprocess dataset for time series.
  - Collect more data to evaluate generalization and robustness of the models.
  - Explore hybrid approaches (e.g., combining SARIMA and LSTM) for improved performance.
  - Apply advanced hyperparameter tuning techniques like Bayesian optimization for LSTM and XGBoost.

Model 2: Weather Status Classification:

- Introduction: The goal is to develop a classification model with the accuracy of at least 98% to predict **'weather status'** using a dataset containing air quality and weather measurements for applications in environmental monitoring, health advisory systems, and urban planning.

- Modelling approaches:

- **Approach 1: Train DecisionTreeClassifier and RandomForestClassifier on full features.**
  - **Objective:** Train a DecisionTree and RandomForest model using all features in the dataset.
  - **Outcome:**
    - The classification report indicates almost perfect accuracy, precision, recall, and F1-score for all classes for both models but there are slight drops in precision and recall for minority classes in RandomForest.
  - **Evaluation:**
    - Decision Tree is likely to get potential overfitting, while RandomForest has better generalization for most classes.
- **Approach 2: Train DecisionTreeClassifier and RandomForestClassifier on top 10 best features.**
  - **Objective:** Use SelectKBest to reduce feature dimensionality
  - **Outcome:**
    - RandomForest has performed better than in approach 1 while DecisionTree still kept perfect classification.
  - **Evaluation:**
    - Further evaluation on future data is necessary to gain generalization ability.
- **Approach 3: Train following approach 2 but on balanced dataset.**

  - **Objective:** Apply SMOTE to address the imbalanced dataset and train on the top 10 features.
  - **Outcome:**
    - Achieved performance on all metrics almost similar to Approach 2. Except from that, RandomForest have seen an improvement.
  - **Evaluation:**
    - Both models are good enough to classify weather status, but DecisionTree may get overfitting problem. Further evaluation and testing on real data is needed.
- **Approach 4: RandomForest with Hyperparameter Tuning.**
  - **Objective:** Optimize RandomForest hyperparameters such as ‘n_estimators’,‘max_depth’, ‘min_samples_split’ using GridSearchCV.
  - **Outcome:**
    - No significant improvements in the performance of model.
  - **Evaluation:**
    - May use other hyperparameters tuning techniques to improve model.

- Future improvements:

- Collect more data to do evaluation to avoid overfitting and retrain if needed.
- Test more techniques and models to improve performance.

1.5 Review

Status: On-going

II. CURRENT ISSUES

2.1. Challenging Question

How can we identify high-risk periods for air pollution based on the combination of weather and pollution factors?

- **Description:** We previously considered exploring how to identify high-risk periods for air pollution based on a combination of weather and pollution factors. However, due to the complexity and challenges in analyzing multiple variables effectively, we decided to eliminate this question from our current objectives.
- **Impact:** By narrowing our focus, we can concentrate on more manageable issues that can still provide meaningful insights and actionable results. This decision allows us to allocate resources more effectively and prioritize objectives that align closely with our capabilities and project goals.

2.2. Analogous Questions

- **Description:** Both of these questions focus on the relationship between wind speed and direction with air pollution levels, specifically fine particulate matter (PM2.5 and PM10). Both questions hypothesize that wind factors can influence the distribution of air pollution, and certain wind directions may bring higher pollution levels, depending on the emission sources.
  - **Question 2:** Is there a correlation between wind speed/direction and PM2.5 or PM10 levels?
  - **Question 6:** How do changes in wind direction and speed correlate with variations in pollutant levels across different areas of the city?
- **Impact:** The redundancy of similar questions addressing the same core issue can lead to wasted resources and over-analysis, as efforts may be duplicated. While it can reinforce understanding and provide cross-validation, it may also dilute focus, delay progress, and prevent exploration of other important areas. Consolidating such questions into a single, clear inquiry would streamline the research process and optimize resource allocation.

2.3. Data Complexity in Visualization

- **Description:** The dataset has a high number of rows and columns, making it visually overwhelming when attempting to interpret trends or key points. Visualizations may appear cluttered, with too much information presented at once, or fail to convey clear insights due to the sheer data volume.
- **Impact:** Complex visualizations with too much data can cause confusion and misinterpretation, as it's difficult for viewers to identify important patterns or key information. The lack of clarity in visualization can reduce the effectiveness of data storytelling and ultimately affect decision-making.

III. SOME IDEAS TO SOLVE THESE ISSUES

3.1. Resolving Redundant Analysis Questions

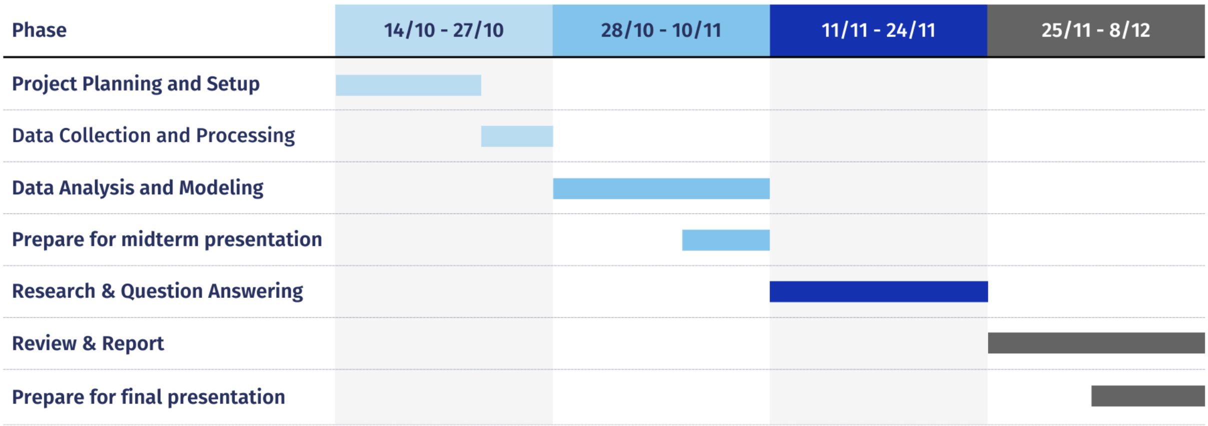
- Unified Research Framework: combine Q2 and Q6 into a single comprehensive inquiry: “Is there a correlation between wind speed/direction and PM2.5 or PM10 levels?”
- Structured Analysis Approach
  1. Analyze citywide wind patterns
  2. Map pollution levels against wind data
  3. Identify location-specific correlations
  4. Develop predictive insights

3.2. Managing Data Visualization Complexity

- Filtering Techniques
  - Implement time-based aggregation (daily/weekly averages)
  - Focus on statistically significant patterns
  - Use representative sampling for initial analysis
- Enhanced Visualization Methods: create hierarchical dashboards:
  1. Top-level summary views
  2. Intermediate detail layers
  3. Detailed data exploration options

IV. PLAN FOR REMAIN TIME

The project is on track to meet all objectives within the planned timeline:



4.1. Progress Overview

All tasks are progressing on schedule according to the Gantt chart. The following milestones have been completed:

- Collecting Data
- Data Pre-processing
- Exploratory Data Analysis (EDA) to Answer Key Questions

4.2. Upcoming Tasks:

Continue Data Modeling (25/11 - 1/12):

- Crawl more data to check over-fitting and improve models.
- Try more techniques and models to find the best solution to these two tasks.

Review and Report (2/12 - 8/12):

- Consolidate findings from EDA and data modeling.
- Prepare a detailed report, including insights, analysis results, and recommendations.
- Review the report for accuracy and completeness.

Prepare for Final Presentation (05/12 - 10/12):

- Create a visually appealing and concise presentation.
- Summarize key insights, methods, and results for easy understanding by stakeholders.
- Practice presenting to ensure clarity and confidence during the final delivery.