naimuri
A QINETIQ company

Senior Data Scientist Tech Challenge

The Scenario

Welcome! We're excited to see how you approach problems.

Imagine you are a Senior Data Scientist at Naimuri. We've been approached by a government partner who needs help understanding and forecasting COVID-19 trends to inform policy. They have provided us with a historical dataset (22/01/2020 to 29/08/2020). Your task is to perform an initial investigation and propose a path forward.

Your deliverable is a Jupyter Notebook that not only shows your analysis but also answers key strategic questions. Your audience is a Naimuri Project Lead and a non-technical stakeholder from the partner.

Data

The data is available in CSV format attached to the email. Here is a sample:

	ObservationID	ObservationDate	State	Country	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

Your Task (To be completed in a Jupyter Notebook)

Your notebook should be well-commented, clearly structured, and walk the reader through your thought process. Please structure your work into the following sections.

Part 1: Data Triage & Exploratory Data Analysis (EDA)

- Load, clean, and pre-process the data.
- Perform EDA to uncover initial insights. What are the key trends? What interesting patterns do you see?
- **Crucially, identify and document any major data quality issues, limitations, or potential biases in this dataset. What key questions *can't* be answered with this data alone?**

Part 2: Baseline Forecasting & Experiment Design

This section should be structured as a formal experiment.

1. **Hypothesis & Methodology:**
    - State a clear hypothesis for your forecasting task (e.g., "A basic statistical model like ARIMA will be outperformed by an ML model like Prophet or an LSTM for predicting 'Confirmed' cases in the UK").
    - Select at least **two** different modeling approaches for comparison. Justify why you chose them.
2. **Execution & Evaluation:**
    - Train your models on the provided data.
    - Forecast the results for a specific country or region for the 30 days *after* 29/08/2020.
    - Define and apply a clear evaluation metric (e.g., RMSE, MAE) to compare your models.
3. **Validation:**
    - Do your own external research to find the *actual* data for your chosen country/period.
    - Compare your forecasts to the real-world data. How did your models perform? Which was more accurate and why?

Part 3: Strategic Recommendations

This is the most important part. Please answer the following questions in clear, concise markdown cells within your notebook.

1. **Stakeholder Summary (for the non-technical partner):**
    - Write a **one-paragraph summary** of your findings. Explain what your model does, its primary limitation, and your single biggest recommendation, all in plain English.
2. **Production & Lifecycle (for the Project Lead):**
    - Imagine this is not a one-off analysis but the start of a **live, automated forecasting system**. Briefly outline the architecture you would propose.
    - Consider: How would you ingest new data? What tools would you use to build the pipeline? How would you deploy and *monitor* the model for data or concept drift?
3. **Next Steps & Data Strategy:**
    - What **other data** would you ask the partner for to improve this model?
    - What **new features** would you prioritise engineering next?

What We're Looking For

- **A "Conscientious and Scientific Approach":** How you handle data quality, structure your experiment, and justify your decisions.
- **Clarity:** A clean, readable notebook with good visualisations that tell a story.
- **Strategic Thinking:** Your answers in Part 3 are as important as your code. We want to see how you think about the *full lifecycle* and *business value* of your work.
- **Pragmatism:** A working, well-justified solution is better than an overly complex one that doesn't run.

Use of AI

It is fully expected and welcomed that you may use AI assistants (like ChatGPT, Copilot, etc.) to help you, just as you would any other tool like an IDE. As with all work, we expect you to be open about its use.

Please add a final section to your notebook titled "AI Validation" that briefly describes:

1. How you used AI.
2. **How you validated its output**.