Turning a request into an analysis

Data fellowship program

PATH

Turning a request into analysis– an overview

An unfortunate fact of life as a data analyst is that we are often asked produce analyses by people who have a limited understanding of the data available. Our job is to communicate clearly with whoever is asking for the analysis to produce the most useful output possible.

This involves

1. Identifying the Question: Accurately diagnosing the key questions hidden inside sometimes vague and general requests

2. Defining the Scope: Outlining what we can and can’t produce based on what we have available

3. Building a Timeline: Aligning the desired outputs with a reasonable timeframe for each step that will be required to produce them

4. Undertaking the Analysis: Using a variety of techniques to slice the data to get informative results, iterating at each step

5. Displaying the Results: Trying a few different potential visualizations and seeking feedback on the most effective versions

Within each of these steps are several substeps that we can use to guide our approach. We will use the example of previous work done by PATH MACEPA to illustrate how to turn a request into an analysis that is ready for presentation

Step 1: Identifying the Question

Identifying the key question(s) in an analysis request can be tricky

Oftentimes the request might be general, vague, or touching on several topics at once. Try to narrow down general questions to specific ones.

- Vague: "How are we doing in combating malaria?"
- Specific: "What is the trend in malaria test positivity rates over the past 5 years across high-burden districts?"
- Vague: "Are interventions working?"
- Specific: "What is the correlation between bed net coverage and malaria incidence by region over the last 3 years?"

Identify who exactly is asking for the analysis and why the are asking. Then try to determine what they most likely want to see at the end of the analysis.

- For example, if your country-lead comes to you with a request to analyze the distribution of malaria medications within your country, you may want to determine if that request is actually coming through them from an external funder (i.e. PMI or Global Fund) to satisfy a specific reporting element. This will have implications on how high-level or finely detailed your outputs should be.
- What is the overall objective of the funder? In this case, it might be to identify ways to streamline their distribution networks and ensure that medication stockouts don’t contribute to worsening malaria outcomes in key districts.

Attempt to link the overall questions to specific metrics.

- From the previous example, you might want to link “the distribution of malaria medications” to the percent of facilities stocked out of a given medication at the end of each month over time.

Step 2: Defining the Scope

- Based on the question(s) identified, what is feasible to produce? This will depend on the initial timeline of the ask, the data required, and the complexity of the steps.

- What data do you need vs what you already have?

- Are you familiar with the dataset or is this your first time working with it?

- Are there potential analysis steps you are unfamiliar with?

- Will you have to build a model?

- What partners or other staff will you be required to coordinate with?

    - Do you need to seek external expertise?

- Make sure to discuss these questions with your manager before bringing them to the stakeholder who requested the analysis.

- You will likely have to do some very initial exploratory data analysis to figure out what you are working with. For example, a dataset may be described to you as “complete” but in reality, you may discover that it needs additional work.

Step 3: Building a Timeline

Once the question(s) and the scope have been identified, it is worthwhile to reflect on its priority level when compared with ongoing work.

- Is this an immediate priority that will require you to push back existing deadlines? Are you able to complete it in the given timeframe with the other work you are already doing? If not, you may need to consider negotiating the timeframe.

- After prioritizing, try to build out a list of the steps required to analyze your question. This will vary depending on your analysis question, but here are some standard steps to consider.

    - Data acquisition, cleaning, and compilation
    - Initial analysis
    - Visualizing initial results
    - Gathering and implementing feedback
    - Preparing final visualizations and presentation materials

- For each of these steps, identify milestones or interim deliverables and then estimate the time required to complete those, building backwards from the final deliverable date if one has been given to you.

    - Keep in mind that you will have to incorporate some degree of stakeholder engagement throughout this process. For example, you may need to take time to get data from stakeholders you haven't worked with before. And when gathering and implementing feedback, remember to build in extra time because feedback responses are often delayed when working with busy funders or NMP officials.

- This timeline can be used as a draft work plan to share with whoever you are reporting to. Make sure to check if your timing and interim deliverables align with expectations for the project.

Step 4: Undertaking the Analysis

- Every analysis is unique, but there are some steps that you will likely build into each one.

Data acquisition, cleaning, and compilation

- This can be very time consuming but undertaking an analysis with incomplete or incorrect data will ultimately be more costly and frustrating than doing it properly from the beginning.
- Refer to Foundational Data Skills and Foundational R Skills for key data cleaning considerations.
- The goal should be to have one consolidated analysis dataset with all your variables included.

Initial analysis

- The first analysis passes should involve characterizing the completeness of the data, as well as descriptive analysis of basic trends. Slice and dice the data, stratifying from multiple levels to look for potential associations.

Visualizing initial results

- Build time into your workplan to **visualize iteratively**. This means that you will create preliminary and basic visualizations to start, showing these to a mentor or colleague to get feedback on what visualizations are effective. Don't be afraid to try things that are "out of the box", be creative!
- *This is a good step in the process to make sure your code is clean!* That means it is reproducible, runs cleanly, and is separated into manageably small chunks. **A good rule of thumb** is to have separate scripts for cleaning and for analysis, and then one code script per graph or set of graphs you are making to answer a question. That way, you will easily be able to find and edit your code later.

Gathering and implementing feedback

- Once you have a set of initial visualizations narrowed down to those you think are effective, show them to stakeholders such as your country lead or the project beneficiaries. Make it clear that these are preliminary and seek feedback on what could use more clarity.

Preparing final visualizations and presentation materials

- Implement the feedback that you received from key stakeholders. Refer to previous modules on data visualization and building slide decks for best practices.

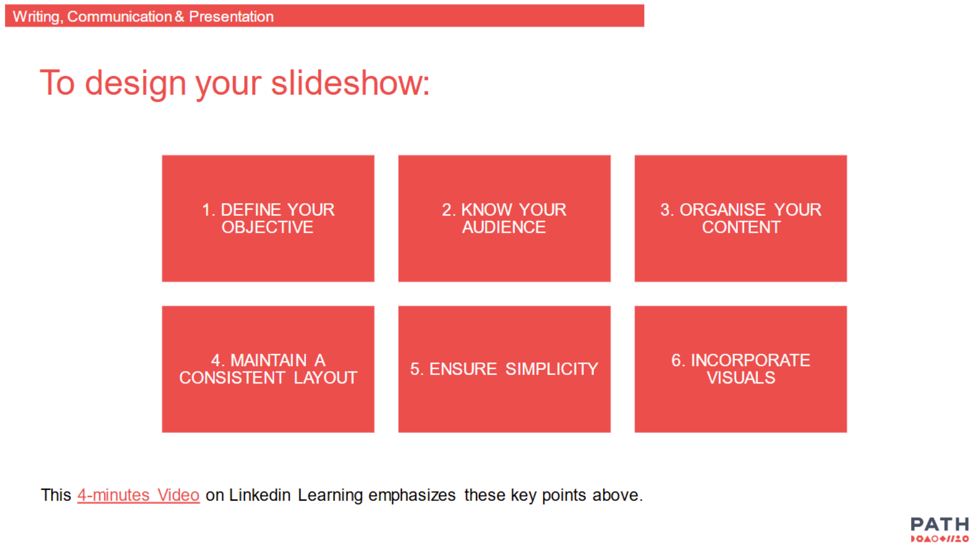
Step 5: Displaying the Final Results

Ensure the final report or presentation:

- Clearly answers the analysis questions you identified
- Is written in a manner accessible to non-technical audiences if necessary
- Informs decision making

The mark of a great presentation is using your data to tell a coherent story

- Start with providing enough background to ensure that people seeing this for the first time will be able to interpret your results
- Don't overexplain your data cleaning process, save those details for discussion unless it is necessary for understanding the results
- Organize your presentation around the analysis questions that you identified, make sure to answer each one on its own slide.
- Highlight key messages in your plots with visual cues to draw the eye
- Refer to the **Foundational Soft Skills** module for more tips on making great presentation slide decks



Writing, Communication & Presentation

To design your slideshow:

1. DEFINE YOUR OBJECTIVE

2. KNOW YOUR AUDIENCE

3. ORGANISE YOUR CONTENT

4. MAINTAIN A CONSISTENT LAYOUT

5. ENSURE SIMPLICITY

6. INCORPORATE VISUALS

This 4-minutes Video on Linkedin Learning emphasizes these key points above.

Example: Zambia Resurgence Analysis

1. Identifying the Question

Our initial question was “why is malaria increasing in Zambia?”

We broke this down into several subparts:

- By how much were malaria cases increasing compared with the previous transmission year?
- Were these increases uniform or heterogeneous in different parts of the country?
- What environmental factors, vector control patterns, or treatment capacity variables correlate with those case increases?

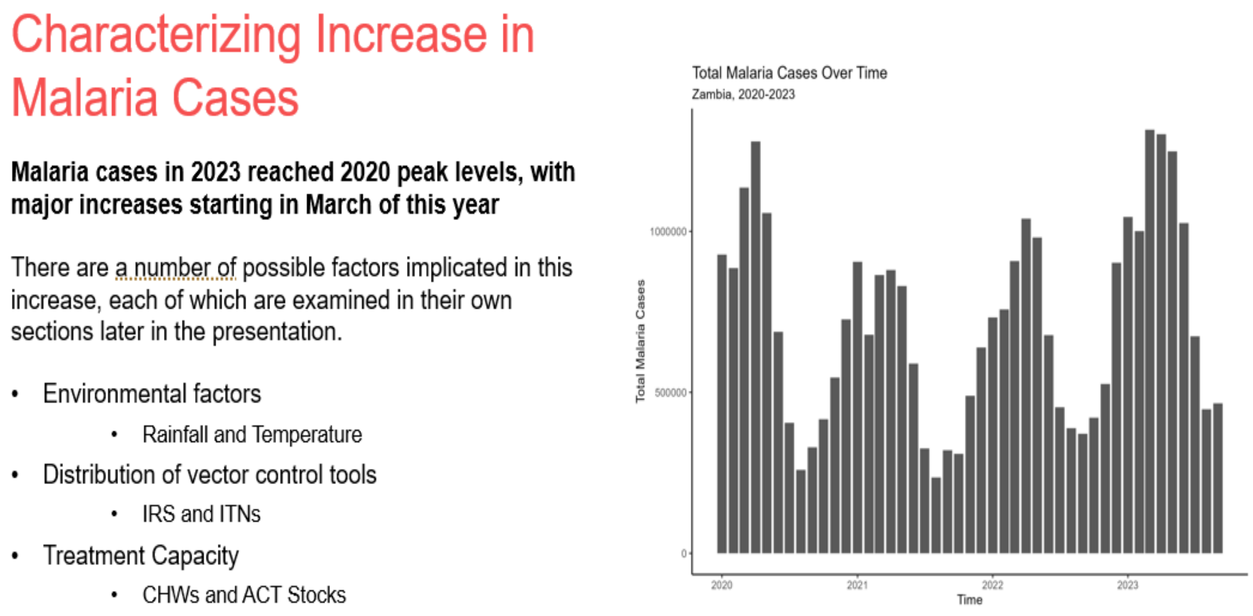
2. Defining the Scope

The scope was very broad and there were many data types that we needed to work with.

Data were sourced on IRS and ITNs from partners, environmental data from remote sensing sources, and case data from DHIS2.

We decided to start with a descriptive analysis that would identify potential causal factors that we could investigate further with a statistical model.

Final Presentation Screenshot



Characterizing Increase in Malaria Cases

Malaria cases in 2023 reached 2020 peak levels, with major increases starting in March of this year

There are a number of possible factors implicated in this increase, each of which are examined in their own sections later in the presentation.

- Environmental factors
  - Rainfall and Temperature
- Distribution of vector control tools
  - IRS and ITNs
- Treatment Capacity
  - CHWs and ACT Stocks

Example: Zambia Resurgence Analysis

3. Building a Timeline

This was an unusual case where the timeline evolved as we went. We began in March 2023 with an expectation of delivering descriptive results within a few months but ended up presenting to the NMCP in December.

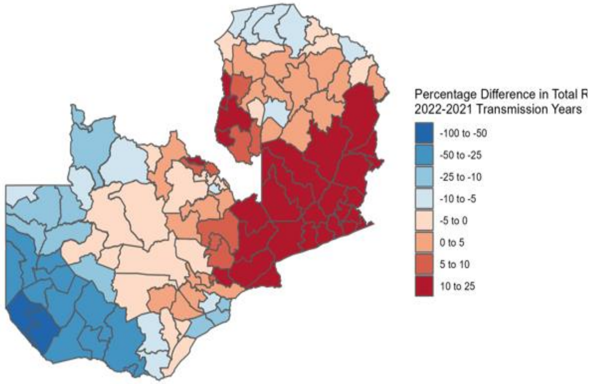
4. Undertaking the Analysis

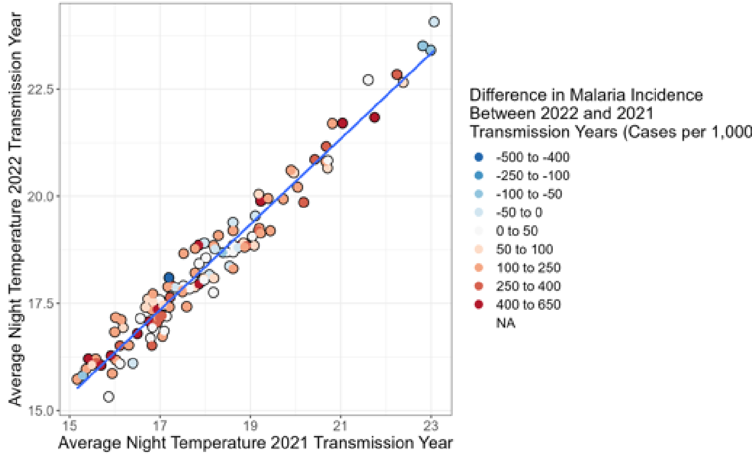
We created dozens of interim visualizations that were sent back and forth to the NMCP and partners to gather feedback.

Spatial visualizations, correlation plots, barcharts, and time series plots were all included at various points to try to illustrate key messages.

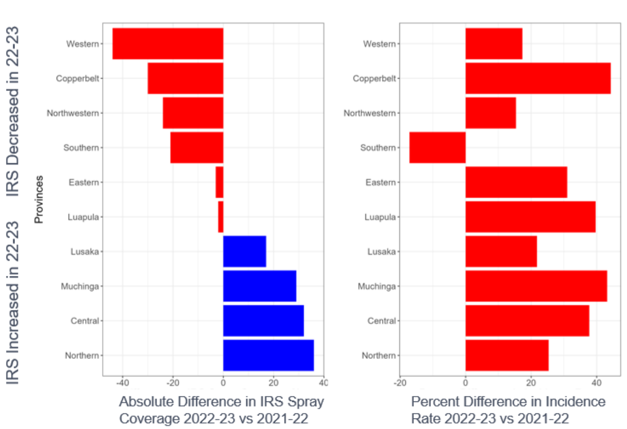
This analysis was very exploratory in nature, so we spent a lot of time following up on leads based on the initial associations we saw in the data and trying to make sure those patterns were clear and robust.

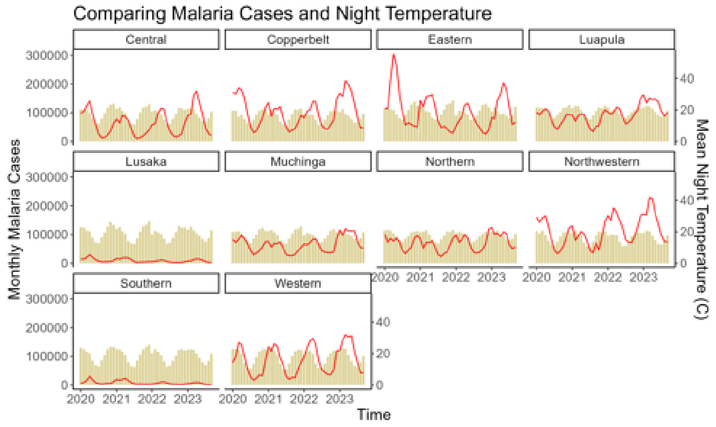
Interim Presentation Screenshots





*Provinces ordered (top to bottom) by largest difference between IRS coverage in 2021-22 and IRS coverage in 2022-23





Example: Zambia Resurgence Analysis

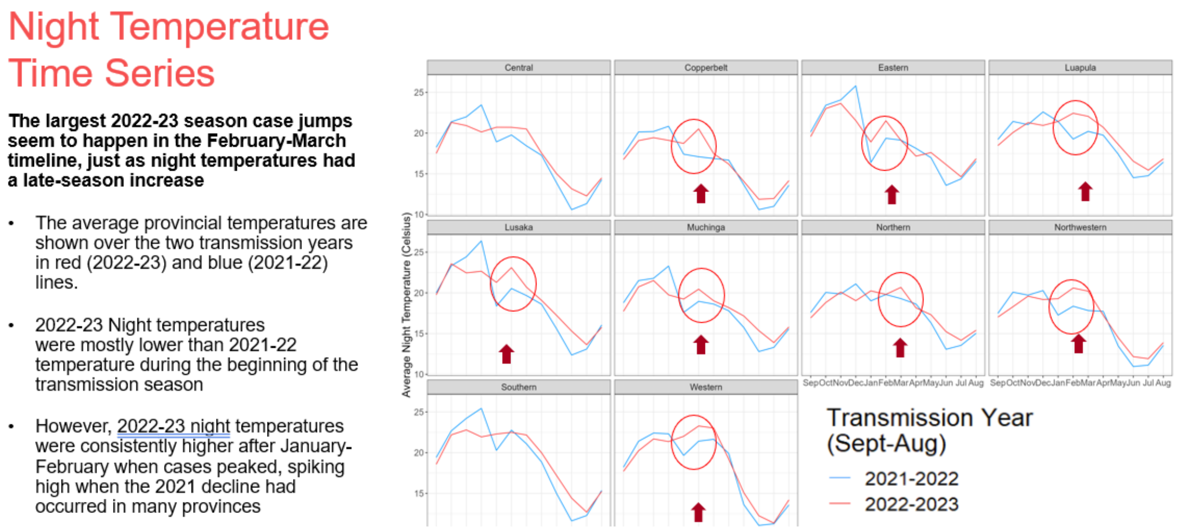
5. Displaying the Final Results

A primary focus for our presentation was communicating the message that there are many potential factors, and that our analysis wasn't final, but that we had strong suspicions around two potential explanatory factors for the case increases in Zambia.

We used visual cues on the slides to highlight the coincidence in timing of the case increase with the timing of a narrowing in the diurnal temperature range (lower day and higher night temperatures = better mosquito survival usually), and that both the 2023 and previous 2020 case increases happened 3 years after the last mass ITN distribution in Zambia, when the nets had outlived their usefulness.

This last point was critical because it presented a potential actionable point for policy makers, indicating that ITN distributions have to happen more frequently than every three years.

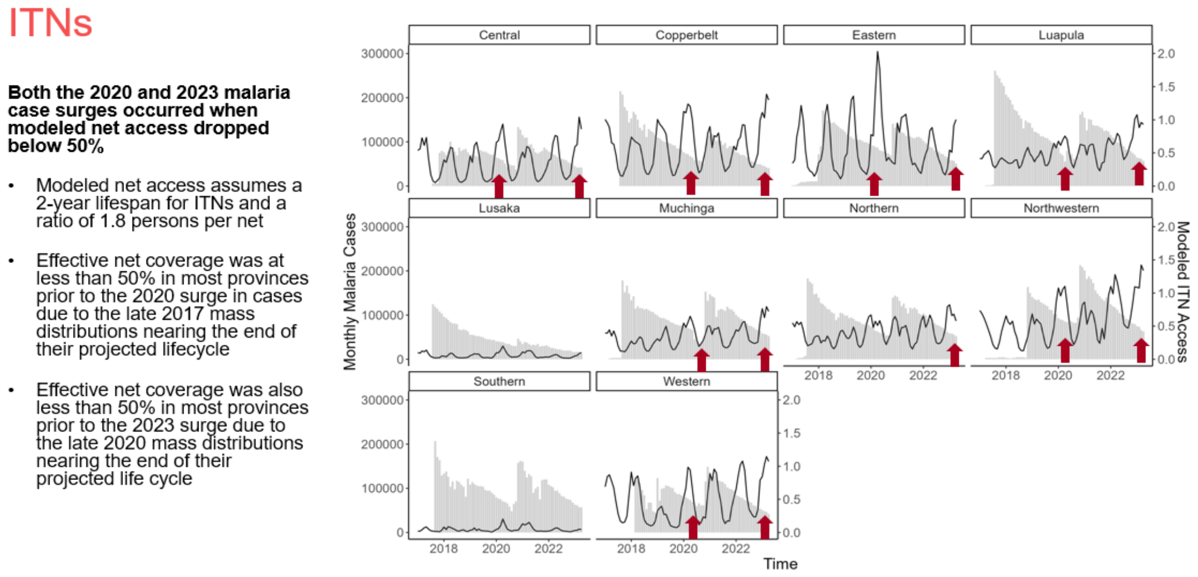
Final Presentation Screenshots


Night Temperature Time Series

The largest 2022-23 season case jumps seem to happen in the February-March timeline, just as night temperatures had a late-season increase

- The average provincial temperatures are shown over the two transmission years in red (2022-23) and blue (2021-22) lines.
- 2022-23 Night temperatures were mostly lower than 2021-22 temperature during the beginning of the transmission season
- However, 2022-23 night temperatures were consistently higher after January-February when cases peaked, spiking high when the 2021 decline had occurred in many provinces

Central
Copperbelt
Eastern
Luapula
Lusaka
Muchinga
Northern
Northwestern
Southern
Western
Average Night Temperature (Celsius)
25
20
15
10
Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug
Transmission Year (Sept-Aug)
2021-2022
2022-2023


ITNs

Both the 2020 and 2023 malaria case surges occurred when modeled net access dropped below 50%

- Modeled net access assumes a 2-year lifespan for ITNs and a ratio of 1.8 persons per net
- Effective net coverage was at less than 50% in most provinces prior to the 2020 surge in cases due to the late 2017 mass distributions nearing the end of their projected lifecycle
- Effective net coverage was also less than 50% in most provinces prior to the 2023 surge due to the late 2020 mass distributions nearing the end of their projected life cycle

Key takeaways

- Defining distinct analysis questions that are tied to data elements you can access is a key first step.

- Every step of the process is iterative! If you have a relationship with the person requesting the analysis, make sure to get feedback early and often. If not, do so with your mentor or another colleague.

- Think like a detective! Be creative and skeptical. Ask *why* you are seeing what you are seeing. Investigate trends and think about how factors like data quality might be impacting what you see. Assume there are data errors and hunt them out, especially when working with aggregated DHIS2 data.

- Make sure that your final product clearly answers the analysis questions you defined

- Try to tie your results to actionable insights for policymakers

PATH