Médecins Sans Frontières (MSF)

Employer Project: Technical Report

By Raksha Nama

Table of Contents

1. Context

Médecins Sans Frontières (MSF) operates in more than 70 countries, providing urgent medical care to populations affected by conflict, epidemics, disasters, or exclusion from healthcare. To maximise operational impact, MSF Brussels redesigned its "End of Assignment" (EoA) survey to better capture the experiences of frontline staff. The redesigned survey aims to improve learning across projects, strengthen the quality of care delivered, and safeguard personnel operating in high-risk, high-pressure environments.

Problem Statement: This project analyses the new EoA survey data to identify which projects achieve the greatest impact and which factors predict performance on MSF's core KPIs: **Relevance** and **Do No Harm (DNH)**. It visualises these findings to support data-driven decision-making in MSF.

2. Project Development Process

2.1 Project Scope

The analysis sought to:

1. Determine which MSF projects deliver the highest impact in their local context.
2. Identify key drivers predicting strong performance on Relevance and Do No Harm, the two KPIs that define "impact" within MSF's framework.
3. Use these findings to inform dashboard design for operational decision-making.

The Effectiveness KPI was excluded due to flaws in the current question structure, which does not measure impact in a way suitable for quantitative analysis.

2.2 Data Overview

The dataset consisted of synthetic survey data, constructed by MSF based on expected responses from the new EoA survey. Although simulated, the data was treated as valid to develop analytical processes and dashboard prototypes that MSF can apply once real data collection begins. Future iterations will require updating predictions and KPI scoring based on actual trends.

2.3 Data Cleaning and Reshaping

Python was selected for cleaning, transformation, and modelling due to its scalability, data manipulation libraries, and ease of replication across future datasets.

Key cleaning steps:

- Total dataset: 5,000 rows and 58 columns; final null row removed (resulting in 4,999 rows).
- Data was in long format. Using melting and pivoting, the dataset was restructured into a wide format (i.e. one respondent per row) to enable regression, descriptive analysis, and visualisation.

- 24 columns were removed due to duplication or identical responses across all entries. Duplication may be due to the synthetic data and will need re-evaluation once applied to real data.
- Columns were renamed for clarity, using snake_case. KPI columns use a q#_kpi pattern (e.g. 1_relevance, 3_effectiveness).
- All ID and question values were converted to integers.
- Final dataset shape after processing: 4,999 rows × 66 columns.

A major issue was misalignment between KPI calculation and survey structure: ordinal Likert-scale data (1–5) from the survey is treated as continuous (mean average was used in KPI calculation). This informed our descriptive analysis techniques (Section 2.5).

2.4 Exploratory Data Analysis (EDA)

EDA examined:

- KPI score distribution by country, project and employee title.
- Median scores per KPI.
- Frequency of high and low ratings (1-2 vs. 4-5).

This revealed heavy clustering of responses at "3", a symptom of Likert midpoint bias, which led to challenges in predictive modelling.

2.5 Descriptive Analysis

Descriptive statistics examined:

- Performance against impact KPIs.
- Projects and countries performing particularly well (scores 4-5) vs. poorly (scores 1-2).
- Drivers of high and low performance.

This informed business findings 1-2.

2.6 Predictive Analysis

Three modelling approaches were used to determine which KPIs best predict impact.

1) <u>Linear Regression (OLS)</u>

OLS was tested on Relevance but returned a low explanatory power (~8%) because it assumes numerical continuity, inappropriate for Likert-scale ordinal data.

2) <u>Ordinal Logistic Regression (OLR)</u>

OLR was applied to both Relevance and DNH:

- Relevance was most strongly predicted by DNH: increasing DNH scores by 1 point has a 56% chance of increasing Relevance scores.

- DNH was most strongly predicted by Relevance: increasing Relevance scores by 1 point has a 60% chance of increasing DNH scores.
- The Staff Support & Care KPI showed statistical significance for both: increasing Staff Support scores by 1 point has a 56% chance of increasing Relevance scores and a 31% chance of increasing DNH scores.

An additional regression combining Relevance and DNH into a single "impact" score (2–10) provided no statistically significant results so was discarded.

3) <u>Decision Tree Models</u>

Decision trees were used to identify thresholds which predict high or low impact scores. Due to clustering at "3", the data was rebalanced using SMOTE.

- Relevance model accuracy: 57%, with overprediction of "high" scores — logistic regression remained superior.
- DNH model accuracy: 68%, with overprediction of "low" scores. Important predictors of "low" scores were:

Key Predictor	Threshold
Coherent with MSF Values	1
Staff Support	1
Agile Decisions AND Population Centredness	$\leqq$ 4 AND $\leqq$ 3

These thresholds informed the dashboard's risk flag rules. See Appendix 2 for the full decision tree used.

2.7 Dashboard Development

Power BI was selected because it aligns with MSF's systems and supports DAX-based KPI logic. Dashboard design decisions:

Design Element	Rationale
Median instead of mean	Avoids distortion of ordinal data
Simple bar charts only	Straightforward interpretation for field teams
Neutral colour palette	Avoids confusion with MSF's emergency red
KPI cards	Fast insight for decision-making
Dynamic filters (year, country, project)	Supports operational comparison
Auto risk flags	Uses decision-tree thresholds to flag projects

Checklist-style responses were separated using delimiter splitting (|) to analyse the frequency of qualitative themes.

The dashboard is designed for replication, real-time updates, and future scaling once true data becomes available. The steps to create the dashboard and use of DAX code can be viewed in Appendix 1.

3. Technical Overview of the Code

3.1 Rationale Behind Tool Choice

- **Python** (pandas, stats models, scikit-learn, imblearn) enabled scalable manipulation of thousands of rows and repeatable analysis.
- **Power BI** allowed KPI visualisation through DAX, integrating with MSF infrastructure and producing accessible decision tools.
- **Combined approach** ensured rigour in modelling while remaining usable for non-technical end-users.

3.2 Data Cleaning & Transformation Code Logic

Key decisions:

1. **Wide Transformation**

Python code: df_wide = df.pivot_table(index="staff_id", columns="question", values="score")

2. **Column Standardisation**

Regular expressions and .str.replace() created consistent naming to reduce downstream function failures and manual debugging.

3. **Constant Removal**

Python code: df_wide = df_wide.loc[:, df_wide.nunique() > 1]

This avoided meaningless variables influencing regressions or tree splits - a necessary step given synthetic data.

4. **Datatype Casting**

Casting KPI scores to integers ensured compatibility with ordinal models.

3.3 Modelling

1. **Ordinal Logistic Model**

Python code: model = OrderedModel(df["relevance"], predictors, distr='logit')

- Logit link used to interpret coefficients as odds ratios.
- KPI coefficients were converted post-model to provide clear insight for stakeholders.

2. **Decision Trees**

Python code: tree = DecisionTreeClassifier(max_depth=best_depth)

- Gridsearch iteration optimised depth.
- Feature extraction identified thresholds for dashboard logic.

3. **SMOTE Balancing**

Python code: smote = SMOTE()

- Addressed midpoint clustering at score “3” to prevent statistical bias.

3.4 Visualisation Logic in Power BI

Design principles behind DAX and charts:

- Visibility first: Field teams must quickly interpret insights, even in emergencies.
- Accessibility: font >11px, no red except risk flags, tooltips for qualitative drill-downs.
- Interactivity: Filtering by year, country, and project type surfaces operational comparisons.
- Reproducibility: All measures (medians, flags, %, breakdowns) written as reusable DAX functions.

Example DAX:

Median_Relevance = MEDIAN('scores'[relevance])

RiskFlag = IF([Median_Relevance] < 3, 1, 0)

3.5 Technical Limitations

- Synthetic data produced misleading duplicates and repetitive qualitative answers (only 3 unique responses).
- KPI clustering at “3” impeded model precision.
- Effective KPI unusable given the current structure.
- Future real-data modelling may significantly change predictor significance and threshold tuning.

4. Patterns, Trends, and Insights

4.1 Key Findings

Finding	Evidence
Survey response rates are critically low	Only 37% response rate; all countries below MSF 50% target
Certain countries consistently underperform	Relevance: DRC & Haiti; Do No Harm: Yemen & South Sudan
Do No Harm and Relevance mutually reinforce each other	Each is the strongest predictor of the other
Staff Support is a core driver of impact	Increase by 1 point has: +56% likelihood of higher Relevance +31% likelihood of higher DNH
Psychological support, safety & living conditions are major pain points	Top themes in qualitative responses

4.2 Implications for MSF

- Improving staff support mechanisms directly strengthens impact delivery.
- Certain contexts require additional resources, accountability, or operational redesign.
- Risk detection of poor performing projects through dashboard monitoring.
- Qualitative data is essential to understand the drivers of scores.

5. Recommendations

5.1 Technical Recommendations

- Align KPI scoring with survey structure (use medians or % scoring 4-5, not means).
- Redesign the Effectiveness KPI for measurability.
- Build a Project Scorecard combining Relevance, DNH, Staff Support, and other key KPIs identified in the predictive modelling.
- Assign a Lead Data Analyst to apply modelling and dashboard to real EoA data, promote use of dashboard in decision-making and keep the dashboard alive.
- Conduct sentiment analysis to capture richer staff insights once real qualitative data becomes available.

5.2 Business Recommendations

- Prioritise frontline support: psychological help, safety protocols, and living conditions.
- Use a risk tracking system to target underperforming countries such as DRC, Haiti, Yemen, and South Sudan.
- Integrate evidence from the EoA into operational review cycles and management accountability through dashboard use.
- Boost participation through manager follow-ups, obligation reminders, or incentives.

6. Appendix

6.1 Appendix 1: Dashboard Design and DAX Code

Steps to create the dashboard using DAX code:

1. Created a new table for total internationally hired staff (see image below) - this was so that we could calculate survey response %

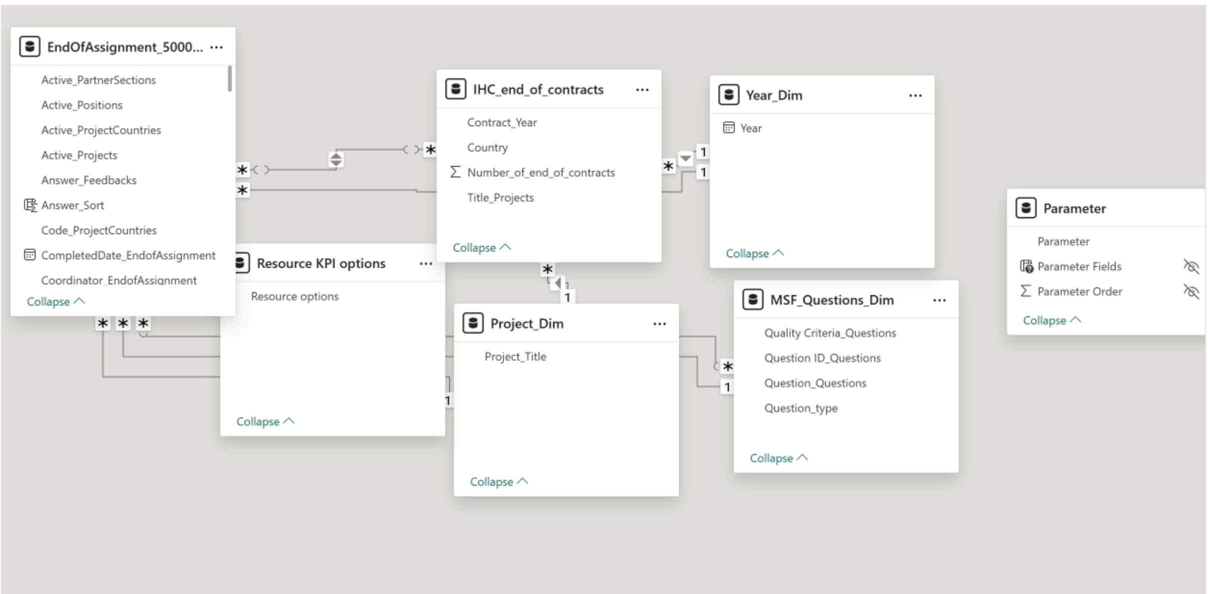
```
IHC_end_of_contracts =
DATATABLE(
    "Country", STRING,
    "Contract_Year", INTEGER,
    "Title_ProjectCountries", STRING,
    "Number_of_end_of_contracts", INTEGER,
    {
        -- Afghanistan
        {"Afghanistan", 2024, "Project 1 in Afghanistan", 25},
        {"Afghanistan", 2024, "Project 2 in Afghanistan", 27},
        {"Afghanistan", 2025, "Project 1 in Afghanistan", 27},
        {"Afghanistan", 2025, "Project 2 in Afghanistan", 29},

        -- Democratic Republic of the Congo
        {"Democratic Republic of the Congo", 2024, "Project 1 in Democratic Republic of the Congo", 22},
        {"Democratic Republic of the Congo", 2024, "Project 2 in Democratic Republic of the Congo", 23},
        {"Democratic Republic of the Congo", 2025, "Project 1 in Democratic Republic of the Congo", 24},
        {"Democratic Republic of the Congo", 2025, "Project 2 in Democratic Republic of the Congo", 25},

        -- Haiti
        {"Haiti", 2024, "Project 1 in Haiti", 22},
        {"Haiti", 2024, "Project 2 in Haiti", 29},
        {"Haiti", 2025, "Project 1 in Haiti", 27},
        {"Haiti", 2025, "Project 2 in Haiti", 22},

```

2. Created an entity relationship diagram (see below) in Power BI so we could link parameters, titles, and visualisations easily



3. Cleaned the data for checklist questions - the responses had many permutations, all separated by '|' (see sample below). These needed to be separated and counted individually.

Fraud and Corruption| HR Planning| Inventory Control
Communication with Management| Health and Medical S
Budget Management| Financial Accountability| Supply Ch

6.2 Appendix 2: Decision Tree

Final decision tree analysis that informed the 'thresholds' which predict whether the Do No Harm score was low (3 or below). This was pruned to 3 levels based on test-train model accuracy, to avoid overfitting of train data and to create simpler rules for MSF to understand.