Seeing the Big Picture: Machine Learning Tools for Exploratory Pattern Discovery in Building Preservation Decisions

Rebecca Napolitano and Joe Kallas

December 2025

**Abstract:** Heritage practitioners now collect unprecedented volumes of data, such as environmental sensors, structural surveys, climate records, yet often struggle to translate this information into clear conservation priorities. This practice point demonstrates how machine learning techniques can identify the environmental and structural variables most strongly associated with deterioration, enabling practitioners to focus limited resources where they will have the greatest impact. We present factor analysis for exploratory pattern discovery and feature importance analysis for predictive modeling, with step-by-step interactive tutorials requiring no programming experience. These tools guide users from data upload through result interpretation, supporting evidence-based decision-making at scales ranging from individual components to entire districts.

# Introduction

Picture this: you manage 200 historic buildings in a downtown district. You have climate data from 15 weather stations spanning two decades, structural surveys documenting crack patterns and material decay, soil moisture readings, temperature fluctuations, and documentation of every intervention. Thousands of data points. Dozens of variables. One question: where do you start?

This scenario appears everywhere in preservation now. Digital documentation tools, environmental sensors, and GIS systems have become standard practice. In historic downtowns, this complexity is further compounded by the coexistence of buildings with different construction types, sizes, and construction periods. We drown in data but starve for insights! The old approach of eyeballing a handful of variables and making gut-check comparisons breaks down when you juggle dozens of interrelated parameters. Here is where dimensionality reduction techniques can help; these techniques take messy, complex datasets and distill them to what matters. Instead of examining each variable alone, they reveal underlying relationships and quantify which factors most influence outcomes: structural deterioration, material decay, climate vulnerability. This data-driven approach helps mitigate confirmation bias by testing assumptions against the entire dataset, revealing patterns that intuition alone might miss or misinterpret.

Data-driven analysis is not new to preservation; we have done quantitative work for decades.[1](#endnote-1) What has changed is scale and accessibility. Modern machine learning algorithms process datasets impossible to handle twenty years ago;[2](#endnote-2) cloud-based platforms like Google Colab[3](#endnote-3) eliminate the need for specialized software or expensive infrastructure; most important, we can now make these tools accessible to practitioners who have never written code.[4](#endnote-4)

This practice point introduces two complementary techniques: *factor analysis* for exploratory pattern discovery and *feature importance analysis* for predictive modeling. These serve distinct but related purposes. Factor analysis helps you understand the structure of your data, discovering, for example, that humidity, rainfall, and groundwater levels form a coherent “moisture-related" pattern. Feature importance analysis addresses a different question: given a specific outcome (like condition rating), which variables best predict it? These techniques can be applied at any scale. At the component level, they identify which environmental conditions most affect specific materials or structural elements; at the building scale, they prioritize interventions based on factors threatening overall integrity; at the district or city scale, they allocate resources across multiple sites by identifying shared vulnerabilities and site-specific risks.

In this practice point, we start with project planning to help you determine whether these techniques suit your project and what data you need. Then we explain factor analysis and feature importance mechanics, with interactive notebooks guiding you through implementation. We bridge the gap between analytical outputs and actionable conservation plans in the interpretation section. We close by addressing common challenges and pointing you toward resources for continued learning. Our goal is not to replace your preservation expertise with algorithms. Domain knowledge remains essential for framing questions, interpreting results, and translating findings into interventions. These computational tools help you extract more insight from data you already collect, so you can decide where to focus limited resources.

# Project Planning: Scoping Your Data-Driven Project

Before you open a computational notebook or run algorithms, plan. Just as you would not start a structural stabilization project without assessing the building’s condition and understanding failure mechanisms, do not jump into data-driven analysis without defining your objectives and evaluating your data.

Prerequisites for effective application of machine learning in heritage projects. This table summarizes what a practitioner must provide (Data Inputs) and what expertise they must bring (Practitioner Skills) to ensure robust results.

| **Category** | **Requirement** | **Description** |
| --- | --- | --- |
| **Data Inputs** | Structured Dataset | Data organized in rows (buildings or components) and columns (variables). |
|  | Sample Size | Recommended 30+ observations for reliable trends; 100+ for robust feature importance. |
|  | Consistency | Variables measured using consistent units and protocols across the dataset (e.g., all cracks measured in mm). |
| **Practitioner Skills** | Domain Expertise | Ability to identify relevant variables (e.g., distinguishing rising damp from condensation) and interpret the physical meaning of statistical results. |
|  | Data Literacy | Familiarity with basic spreadsheet operations, including CSV formatting and data cleaning. |
|  | Technical Barriers | **None**. No programming experience is required to use the provided notebooks. |

To clarify the intended audience and required expertise, these methods are best suited for interdisciplinary teams. **Preservation planners** can use feature importance to prioritize maintenance resources across large municipal inventories, while **structural engineers and architects** gain quantitative tools to identify dominant drivers of deterioration. **Conservators** can further refine these analyses by correlating environmental exposure data with specific decay mechanisms. Key interdisciplinary skills required include data curation (collecting consistent field data) and collaborative interpretation (combining statistical outputs with physical evidence).

## Defining Your Questions

Dimensionality reduction techniques are powerful but not appropriate for every challenge. They excel when you face complexity: many variables, unclear relationships between factors, uncertainty about which parameters matter most.[5](#endnote-5) Practical experience and statistical literature suggest minimum requirements. For factor analysis, aim for at least 5-10 observations per variable (e.g., 100 observations for 10-20 variables) to achieve stable factor structures.[6](#endnote-6) For Random Forest models, 30 observations is the absolute minimum, but 100+ observations yield more reliable importance rankings.[7](#endnote-7) With smaller samples, results become unstable and may not replicate.[8](#endnote-8)

Beyond quantity, data quality matters enormously. Were all variables measured using the same protocols, instruments, and units? Inconsistent measurement methods introduce noise that obscures real patterns. Heritage data present specific challenges: temporal autocorrelation (measurements over time are not independent), spatial dependencies (buildings in the same district share exposures), and survivorship bias (severely deteriorated buildings may have been demolished, removing extreme cases from your dataset). These issues can lead to spurious correlations. Where possible, account for temporal structure by including time lags or using time-series methods, and for spatial structure by including location variables or using spatial statistics. Lastly, watch for outliers, extreme or erroneous data points that distort results. Examine them carefully before removing them; in heritage data, “extremes" may represent significant events like storms or structural failures worth understanding rather than discarding. Figure [1](#fig:condition_dist) shows the distribution of condition ratings in the example dataset for this practice point, which helps identify potential imbalances that could bias predictions. Figure [2](#fig:condition) illustrates the qualitative meaning of each condition category, from Good (1) to Very Poor (5).



Distribution of building condition ratings in the dataset. Understanding the balance of your data (e.g., how many “Good" vs. “Very Poor" buildings) is essential before starting analysis. An imbalanced dataset can bias model predictions.

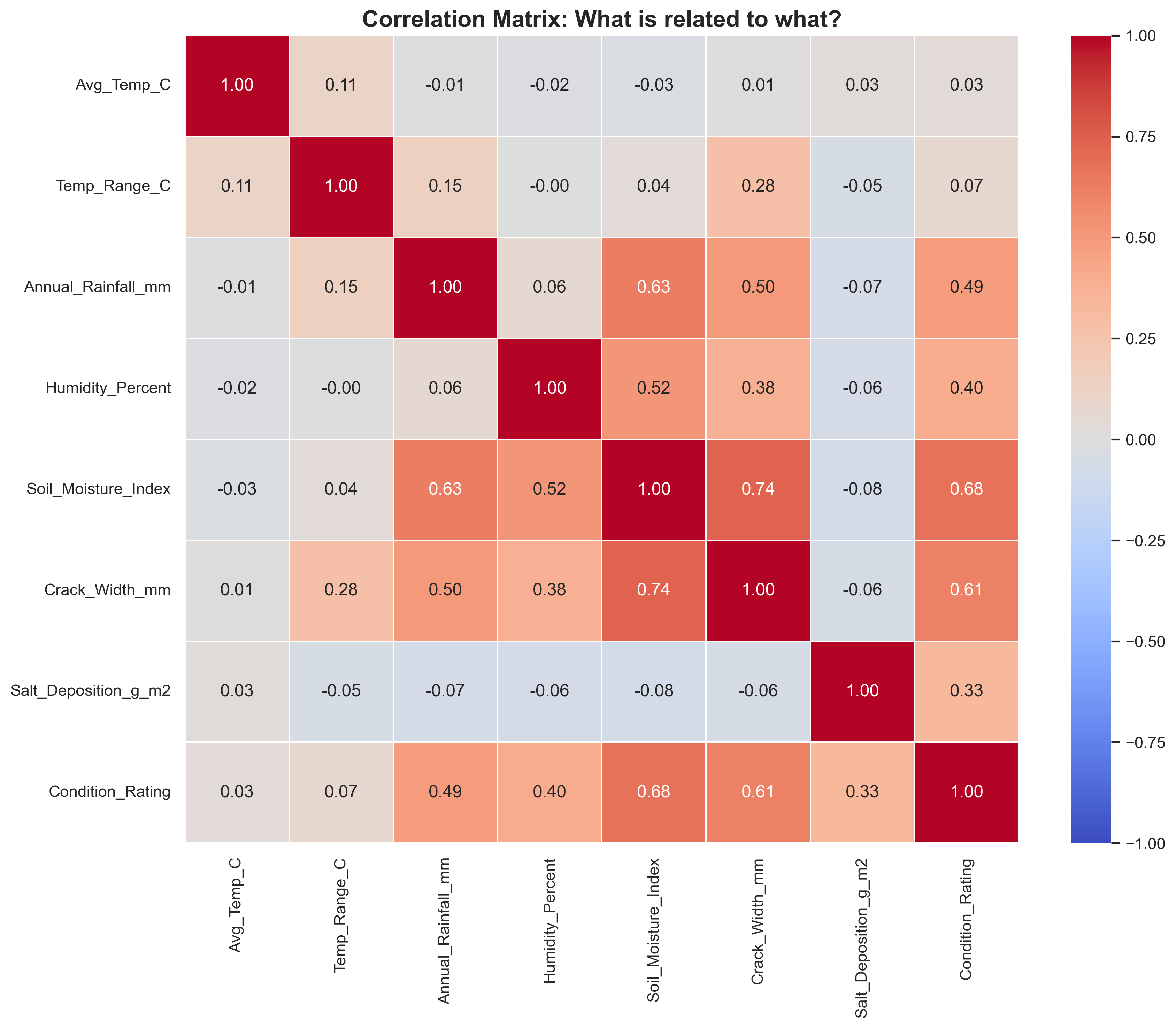


Visual representation of the five-level building condition rating scale used in this study, ranging from intact structures (1 – Good) to severe structural damage and partial collapse (5 – Very Poor).

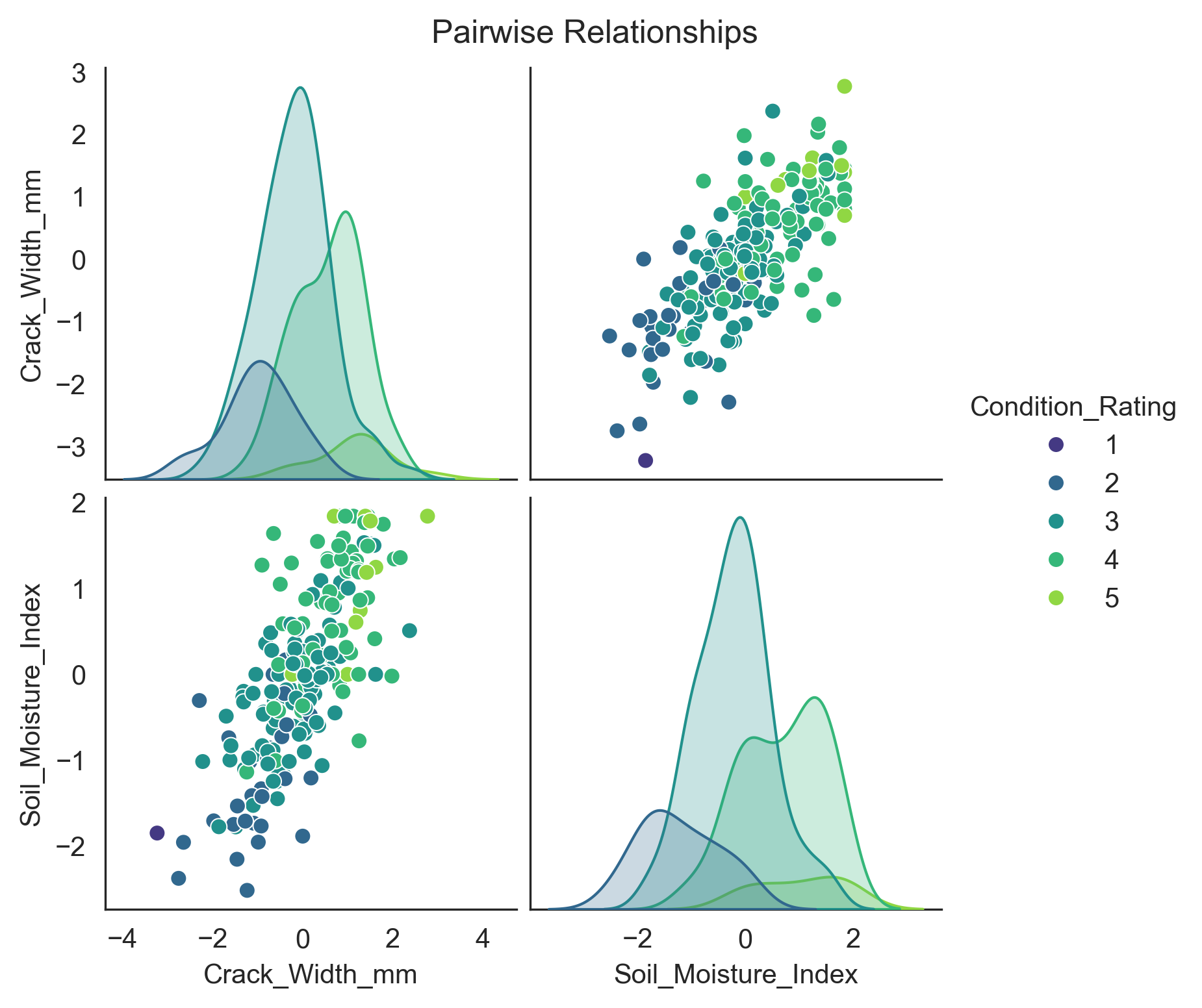
## Exploratory Data Visualization

Before applying dimensionality reduction techniques, examine your data’s structure through three complementary visualizations. The following exploratory visualizations (Figures [3](#fig:correlation) and [4](#fig:pairplot)) should be examined before proceeding to formal analysis, as they reveal data structure and potential issues requiring attention.

Start by examining correlation patterns (Figure [3](#fig:correlation)) to identify which variables move together; these visualizations are generated in the 04\_visualization.ipynb notebook and the 01\_simple\_statistics.ipynb notebook includes correlation analysis tools. For a more detailed view of relationships, pairwise scatter plots (Figure [4](#fig:pairplot)) reveal non-linear patterns and outliers that simple correlation coefficients miss. Before proceeding to analysis, check for multicollinearity using Variance Inflation Factors (VIF), which identify redundant variables that should be removed or combined; the 01\_simple\_statistics.ipynb notebook provides VIF calculation and interpretation guidance.



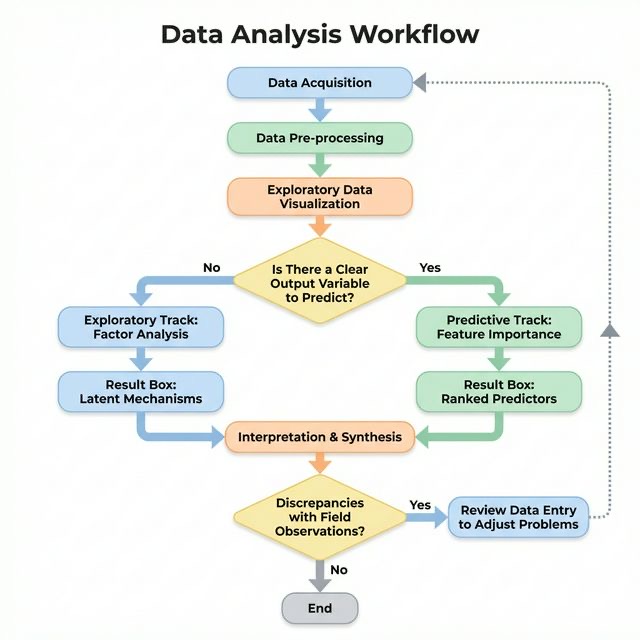
Correlation matrix heatmap showing relationships between variables. Darker red indicates strong positive correlation, while darker blue indicates strong negative correlation. This high-level overview helps identify potential patterns, such as moisture variables correlating with condition.



Pairwise scatter plots showing relationships among key continuous variables. The diagonal shows distributions of individual variables, while off-diagonal plots reveal bivariate relationships. This visualization helps identify non-linear patterns and outliers that correlation coefficients alone might miss. Note that all variables have been standardized (scaled to mean = 0, standard deviation = 1) for comparability, which is why some variables show negative values.

# The “How-To": Data Featurization

In this section we focus on two complementary analytical techniques: *factor analysis* and *feature importance analysis*. While the mathematical details happen behind the scenes in the computational notebooks, understanding what these techniques do and when to use them is essential for interpreting results correctly and applying them appropriately to preservation decisions. The workflow follows three stages. First, data preparation organizes your dataset, handles missing values, and standardizes variables for fair comparison. Second, analysis applies algorithms to identify patterns (factor analysis) or rank importance (feature importance). Third, interpretation translates numerical outputs into preservation insights (a process we detail in Section [4](#interpretation)). This section focuses on the analytical techniques themselves. Figure [5](#fig:flowchart) provides a visual roadmap of this entire workflow, illustrating how data preparation leads into the parallel tracks of exploratory and predictive analysis before converging on interpretation.



Proposed framework for data-driven prioritization in heritage preservation.

## Factor Analysis

### What It Does

Factor analysis explores patterns and relationships among parameters affecting structures, but note that these are *correlational* patterns, not *causal* mechanisms.[9](#endnote-9) When you measure many variables such as temperature, humidity, wind exposure, material properties, structural dimensions, these variables are rarely independent. Temperature and humidity vary together with seasonal patterns; wind exposure and salt deposition co-occur at coastal sites; soil moisture and groundwater levels fluctuate in coordination.

Factor analysis identifies clusters of variables that vary together, revealing hidden structures.[10](#endnote-10) We identify and extract factors that *correlate with* structural behavior. These factors represent overarching patterns that account for observed correlations among parameters. Understanding *why* these correlations exist requires domain expertise; the statistical method only reveals *that* they exist.

Think of factors as categories or groups of conditions that collectively *correlate with* structural behavior. These conditions are not considered alone but as part of broader, interrelated themes. Instead of separately tracking ten moisture-related variables (perhaps rainfall, humidity, condensation, groundwater level, soil saturation) factor analysis might reveal that these cluster into two meaningful factors: “atmospheric moisture" and “ground moisture." This consolidation simplifies subsequent analysis while preserving essential information. Each factor corresponds to a specific set of related conditions; parameters within the same factor share common characteristics in how they *co-vary with* structural impacts. The technique assigns a *loading* to each variable for each factor, a weight indicating how strongly that variable contributes to that factor. High loadings (typically above 0.6-0.7) indicate strong relationships; low loadings suggest the variable is not part of that factor’s pattern.

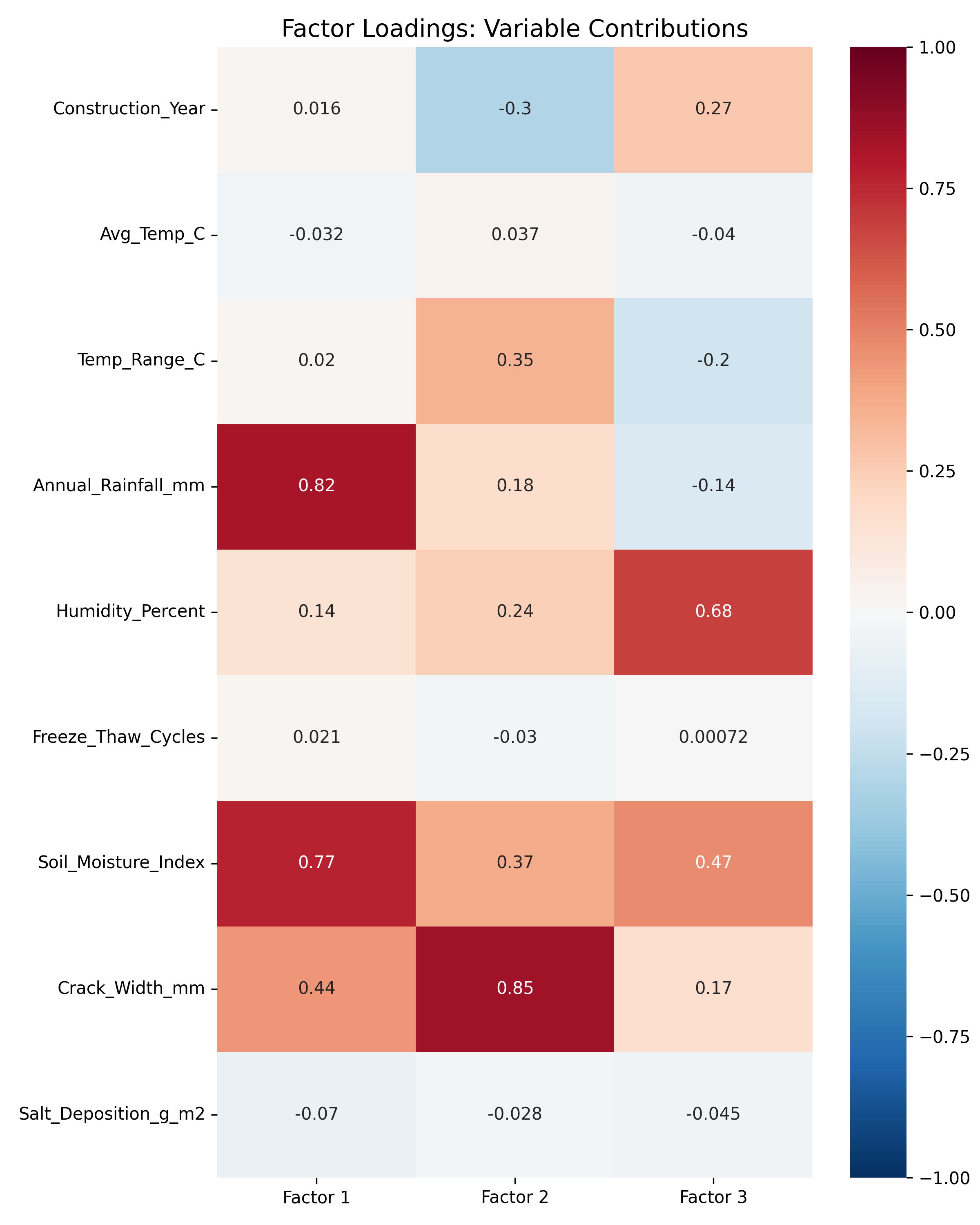
Factor analysis reduces dimensionality by condensing numerous parameters into fewer meaningful factors. If you begin with 30 variables, factor analysis might identify 5-8 underlying factors that capture 70-80% of the variation. This simplification focuses understanding on critical conditions *associated with* structural changes. Rather than monitoring and responding to 30 different variables, you focus conservation efforts on the handful of factors that show the strongest associations with deterioration patterns. The results are interpretable factors that represent underlying structural behavior drivers. These factors are derived mathematically, but *you* as the preservation professional give them meaning based on domain expertise. The algorithm identifies which variables cluster together; you name the factors. A factor with high loadings on temperature range, freeze-thaw cycles, and thermal expansion might be named “thermal stress;" a factor loading heavily on wind speed, rainfall intensity, and exposed surface area might be termed “weather exposure."

### When to Use Factor Analysis

Factor analysis is particularly valuable in several preservation scenarios. First, it can be useful when you have collected many variables but are uncertain which ones are most relevant or how they relate to each other. Secondly, it can be useful when ongoing monitoring collects numerous parameters but you need a smaller, more communicable set of metrics for practical decision-making. It is additionally useful when identifying deterioration mechanisms in the case where multiple environmental or structural stresses may be acting on a building, but you want to understand which combinations of stresses are most significant. Lastly, it can be useful when studying multiple buildings or sites and wanting to determine whether they experience similar or different patterns of stress. Factor analysis is less appropriate when you have few variables (fewer than 8-10), when your variables are largely independent, or when you already have strong theoretical reasons to know which variables matter most. In that last case, feature importance analysis (discussed next) may suit you better.

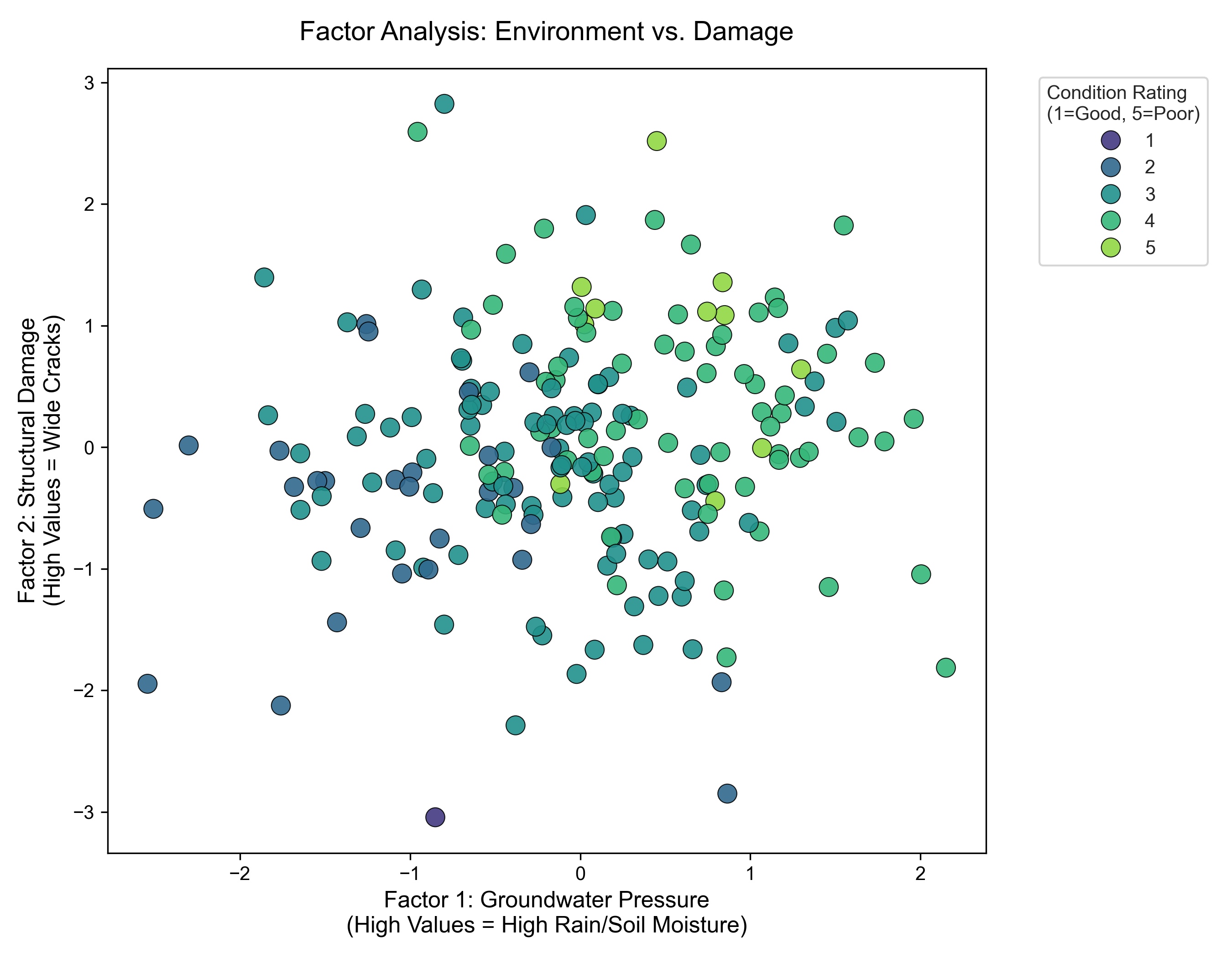
### The Process

Factor analysis proceeds through several steps.[11](#endnote-11) First, variables are standardized to ensure equal influence regardless of units. Second, the algorithm computes correlations between all variable pairs. Third, it extracts factors that explain correlation patterns. The first factor explains the most variance, the second explains the next-most, and so forth. Fourth, diagnostic tools determine how many factors to retain. A scree plot (generated automatically in the notebook) shows where additional factors stop adding meaningful information. Fifth, you interpret and name the factors based on which variables load strongly (typically 0.6 or higher). If soil moisture, rainfall, and crack width load on one factor, label it “Groundwater-Driven Deterioration" (Figure [6](#fig:loadings)). Each observation receives a factor score showing how strongly it exhibits each factor.



Factor loadings heatmap showing three extracted factors. Factor 1 captures Groundwater Environment (Rain, Soil Moisture), Factor 2 isolates Structural Damage (Crack Width), and Factor 3 represents Atmospheric Humidity. This separation allows us to test if environment (F1) actually predicts damage (F2).

Once factors are identified and named, you can visualize how individual buildings score on each factor. Figure [7](#fig:factor_scores) plots the Groundwater Factor (Factor 1) against the Structural Damage Factor (Factor 2), colored by condition rating. Visualizing these two specific factors against each other reveals the correlation between cause (environment) and effect (damage), with high-risk buildings (red) clustering where both factors are significant.



Factor scores scatter plot: Groundwater Environment (X-axis) vs. Structural Damage (Y-axis). The clustering of high-risk buildings (red) confirms that as groundwater pressure and rainfall intensity increase (Factor 1), structural damage indicators like crack width (Factor 2) also increase.

### Interactive Tutorial: Factor Analysis

The companion notebook (02\_factor\_analysis.ipynb) guides you through the complete workflow: data upload, statistical suitability testing, scree plot visualization, automatic factor number suggestion, Varimax rotation, and interpretation of factor loadings. The notebook uses a synthetic dataset of 200 heritage buildings with 9 environmental and structural variables, includes troubleshooting guidance, and saves results as CSV files for further analysis.

## Feature Importance

### What It Does

While factor analysis identifies groups of related variables, feature importance analysis ranks individual variables by their predictive power for specific outcomes[12](#endnote-12) (deterioration severity, maintenance costs, intervention timing). The technique trains machine learning models (in this case Random Forest[13](#endnote-13) or Gradient Boosting)[14](#endnote-14) to learn relationships between input variables and target outcomes. These ensemble methods construct multiple decision trees and aggregate results to determine which parameters matter most.[15](#endnote-15) Unlike simpler statistical approaches, these models automatically capture variable interactions.[16](#endnote-16) High humidity alone may not predict deterioration, but combined with temperature fluctuations it might strongly predict decay. After training, the models measure how much predictive power would be lost if each variable were removed. Variables whose removal causes large performance drops are deemed more important. Importance scores range from 0 to 1, with higher values indicating greater predictive power.

The ranking generated by these machine learning models guides the prioritization of conditions and factors within each category. Parameters identified as highly significant could be given priority attention in subsequent preservation and intervention strategies,,.[17](#endnote-17),.[18](#endnote-18).[19](#endnote-19) If the analysis reveals that groundwater level is the strongest predictor of foundation deterioration, monitoring and managing groundwater can become the top priority.

Feature importance analysis is an iterative process, where the models are trained and evaluated using historical data and domain expertise. The process uses part of your data to build the model (the training set) and reserves part to test how well predictions work on new data (the test set). This ensures that the rankings are refined and reliable for decision-making rather than simply describing patterns specific to your particular dataset. Feature importance contributes to dimensionality reduction not by grouping variables together (as factor analysis does), but by allowing you to filter out the least influential ones. This simplifies both interpretation and data collection since once you know which features drive outcomes, you can prioritize those in future monitoring campaigns or predictive models for this specific case of analysis.

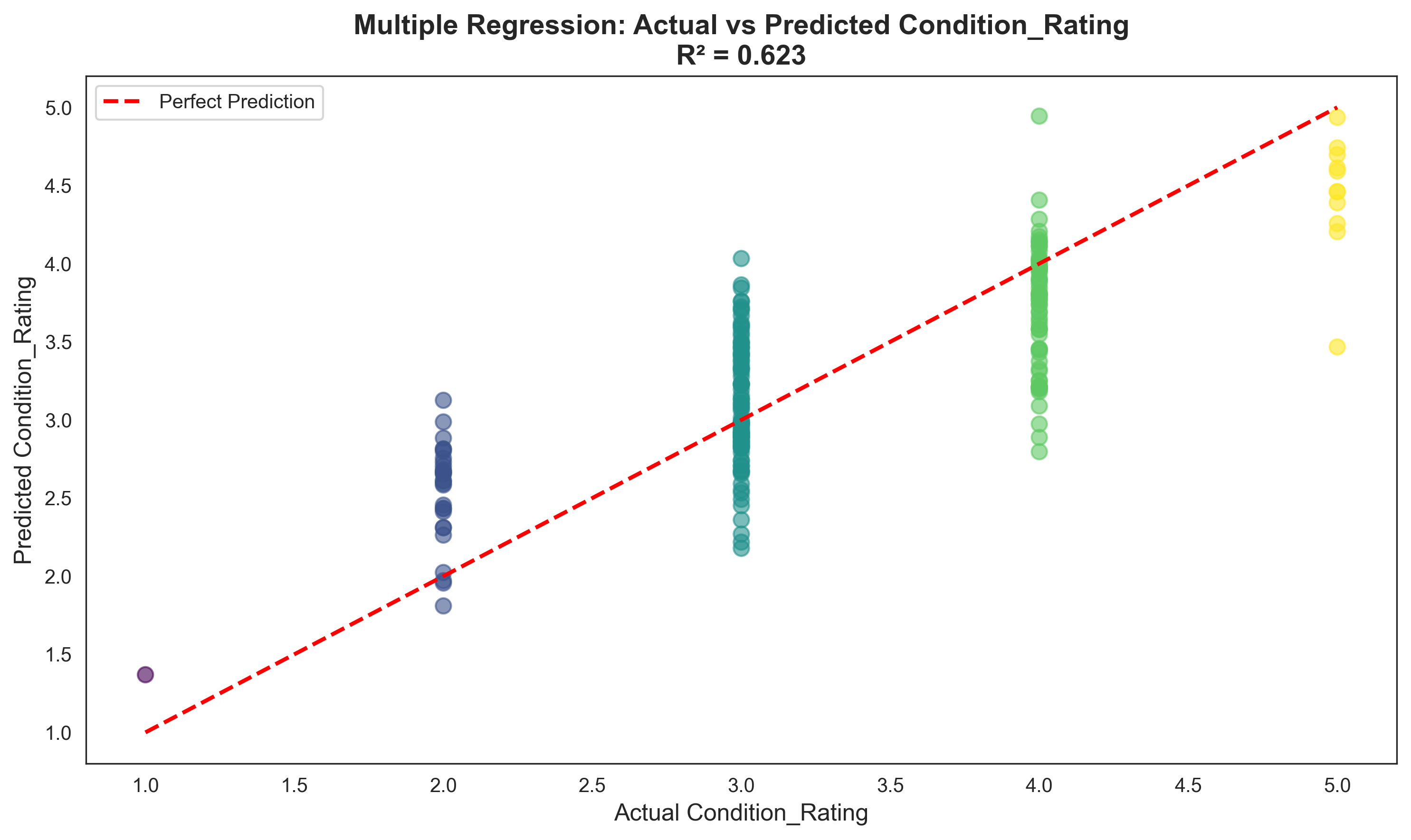
### When to Use Feature Importance

Feature importance analysis is particularly valuable in several scenarios. First, use it when you have a clear outcome to predict. The technique requires that you have both input variables (environmental conditions, material properties, structural characteristics) and an outcome variable you want to predict or explain (deterioration rating, intervention urgency, structural performance). If you only have input variables without outcomes, factor analysis is more appropriate. Second, this technique shines when prioritization is the goal. When you need to rank variables to guide resource allocation, monitoring priorities, or intervention strategies, questions like “Which three factors should we monitor most closely?" and “Where should we spend our limited maintenance budget?" are ones that feature importance can help answer. Third, use this technique when complex relationships are suspected. If multiple variables interact in nonlinear ways, traditional regression assumes linear relationships and requires you to specify interactions. Random Forests and Gradient Boosting automatically detect nonlinear patterns and interactions. Lastly, this method matters when predictive accuracy counts. If you want not just to understand relationships but also to predict future outcomes (“Given current conditions, which buildings will require intervention in the next five years?").

Feature importance is less appropriate when you lack outcome data, when you have very few observations (fewer than 30), or when relationships are simple enough that traditional statistical methods suffice.[20](#endnote-20) It’s also less useful when variables are highly collinear, when several variables measure nearly the same thing, because the model may distribute their importance unevenly. When variables are highly correlated (multicollinear), feature importance rankings become unstable. You can check for multicollinearity by calculating Variance Inflation Factor (VIF) for each variable. Values above 10 indicate severe collinearity.[21](#endnote-21) In such cases, consider removing redundant variables (e.g., keep annual rainfall but remove monthly rainfall averages), using factor scores from factor analysis as inputs, or accepting that importance rankings may vary but focusing only on the top 3-5 features that consistently rank high across different model runs. The notebooks implement both Random Forests and Gradient Boosting, which work similarly but have different strengths explored in the notebook. We recommend running both models and comparing their feature importance rankings. If both models identify the same top variables as most important, you can be confident in those findings. If the models disagree substantially, that suggests either data quality issues or complex relationships that merit closer investigation.

### The Process

While the provided tutorial notebook largely automates the mechanics of feature importance analysis, interpretation of the results requires a clear understanding of the underlying logic. The process begins with the definition of a target variable which must be cast in a numerical or ordinal format; this can be a measurable outcome such as damage severity, deterioration rate, or intervention cost. Simultaneously, feature variables representing environmental factors, material characteristics, or geometric properties are selected and standardized to ensure consistent numeric formatting. A distinction in heritage data analysis is the strategy for splitting data into training and testing sets. Unlike standard random splits, which can introduce data leakage, heritage datasets often necessitate temporal or spatial partitioning. This may involve training on historical periods (e.g., 2010–2017) to test on subsequent years (2018–2020), or training on specific geographic districts to test generalization to others. Once partitioned, ensemble algorithms like Random Forest and Gradient Boosting are employed. These methods construct multiple decision trees to capture the complex, non-linear interactions inherent in the built environment. Validation is then performed to assess prediction accuracy against domain-specific benchmarks. In the context of heritage datasets, an value between and typically indicates useful relationships, whereas a value above is excellent but uncommon. Figure [8](#fig:regression) shows actual versus predicted values for a multivariate model, while Figure [9](#fig:simple_regression) demonstrates the limited explanatory power of single-variable approaches.



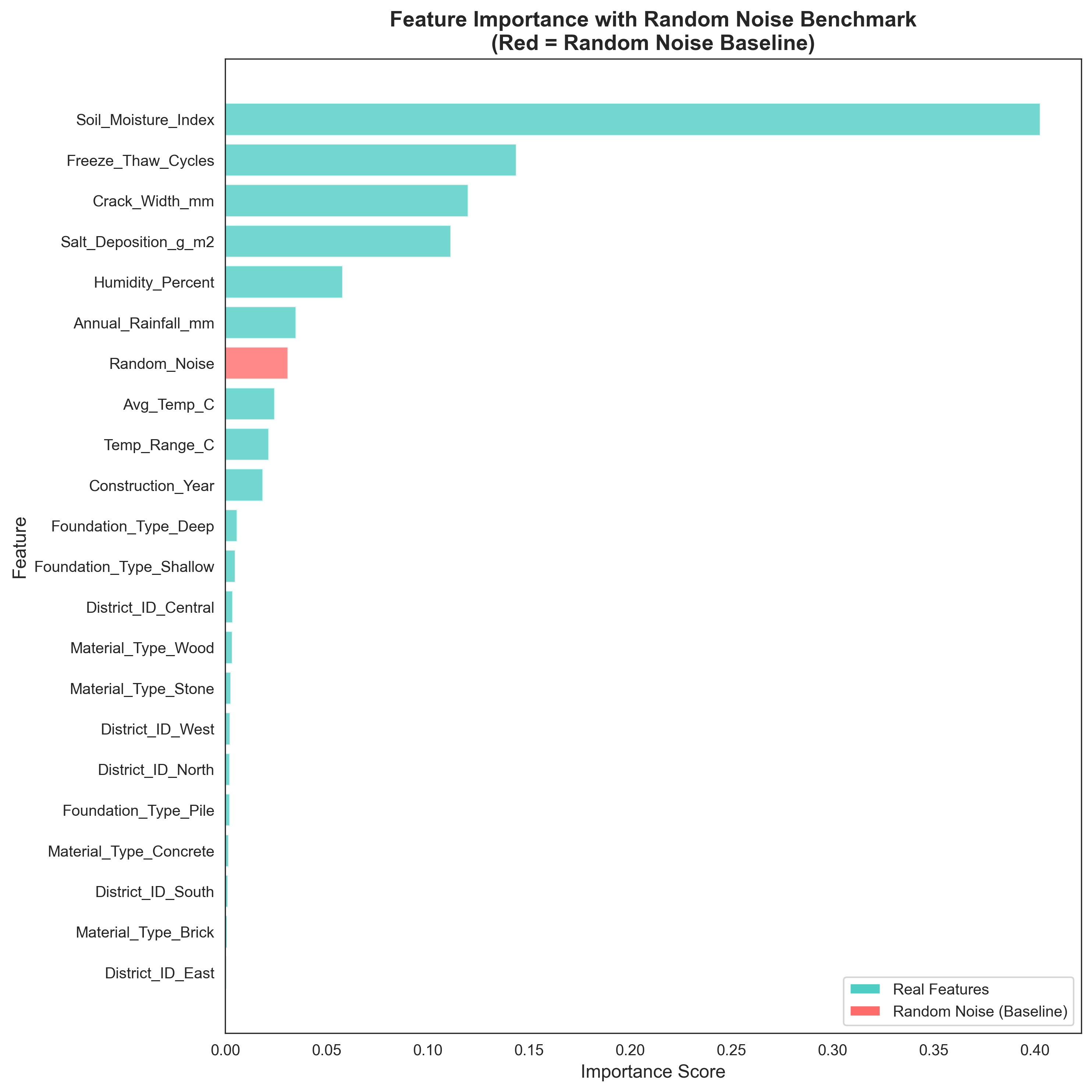
Actual vs. Predicted values from a regression model. Points clustering closely around the red dashed line indicate accurate predictions. Outliers far from the line represent buildings where the model’s prediction differed significantly from reality.



Simple linear regression of Condition Rating vs. Crack Width. This baseline model using a single predictor demonstrates why multivariate approaches are necessary. The scattered points show that crack width alone explains only part of the variation in building condition.

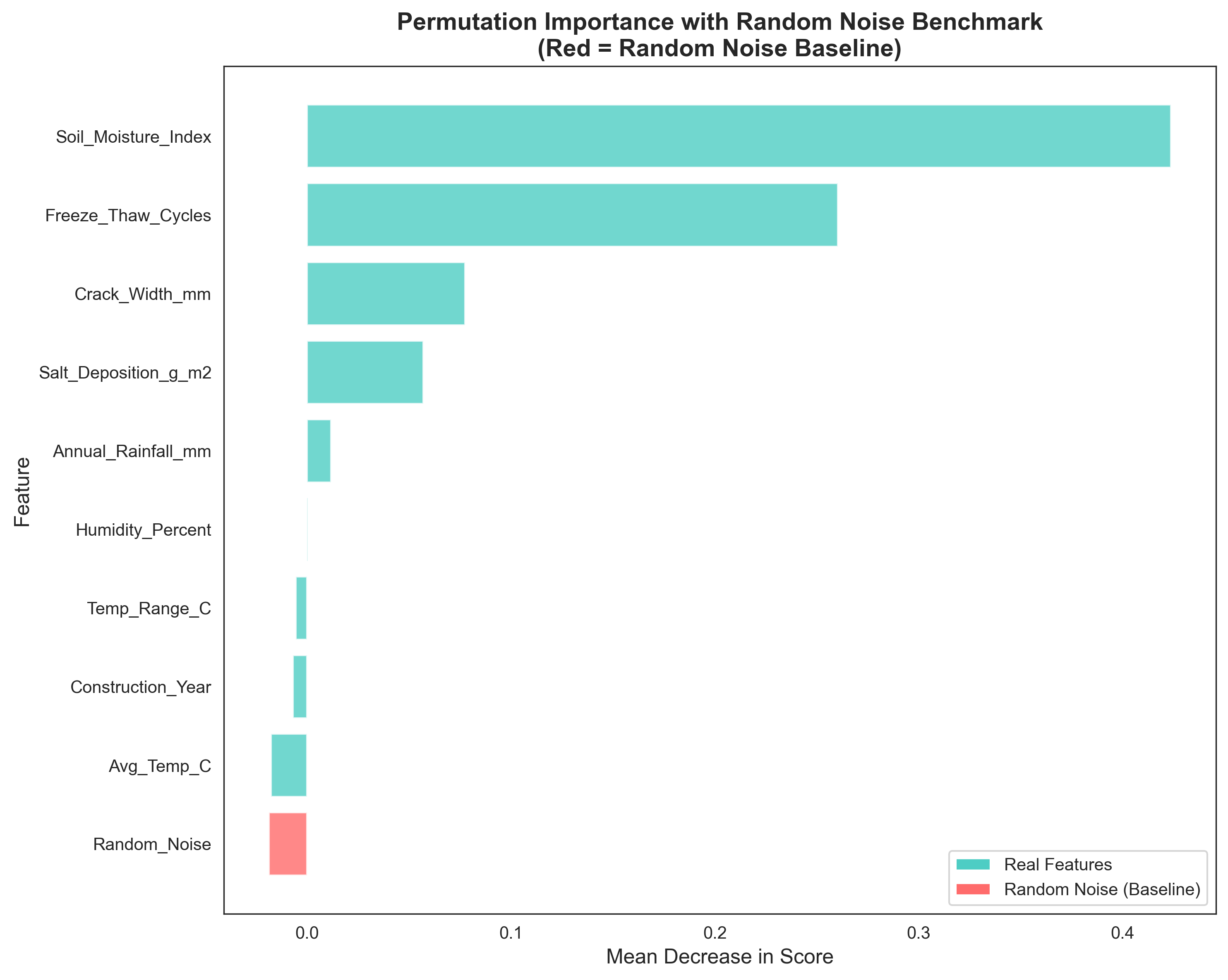
Similarly, a Mean Absolute Error (MAE) of rating points on a standard – scale is considered useful for prioritization, even if it lacks the precision required for individual diagnosis. To ensure the stability of these rankings, -fold or -fold cross-validation is recommended.[22](#endnote-22) Finally, feature importance is computed by measuring the performance loss occurred when a specific variable is removed (permutation importance).[23](#endnote-23) These results, ideally visualized as bar charts, should be interpreted in conjunction with preservation expertise.

To ensure that identified features represent genuine predictive relationships rather than spurious correlations, we include a random noise variable as a baseline benchmark. A random variable (drawn from a standard normal distribution) is added to the feature set before model training. Any real feature scoring below this noise variable indicates weak or unreliable predictive power and should be excluded from interpretation. This approach provides an objective threshold for distinguishing meaningful predictors from statistical artifacts.



Feature importance ranking with random noise benchmark (shown in red). Real features scoring above the noise baseline demonstrate genuine predictive power. This chart directly answers “What drives deterioration?” by ranking variables based on their predictive power. Variables like Soil Moisture and Freeze-Thaw Cycles appear as top drivers, well above the random noise threshold.

To validate that importance rankings reflect true predictive power rather than model artifacts, permutation importance (Figure [11](#fig:permutation)) measures performance loss when each variable is shuffled.



Permutation importance scores with random noise benchmark (shown in red). This validation technique confirms that the ranked variables truly contribute to predictions rather than being artifacts of the model’s internal structure. Features scoring above the noise baseline demonstrate reliable predictive power.

Finally, comparing multiple model types justifies algorithm selection and reveals whether simpler approaches might suffice. Furthermore, utilizing the uncertainty estimates provided by algorithms like Random Forests allows for reporting predictions with necessary confidence intervals (e.g., “”), resulting in a more transparent analytical output.

### When Interpreting Results

Never treat computational outputs as infallible; always validate results through site inspections and material testing. If model outputs conflict with field observations, re-examine your data. Feature importance provides rankings of influence, not proof of causation. High importance indicates correlation, not necessarily direct causation. Combine computational results with field expertise; view rankings as guides for investigation, not definitive conclusions. High-importance variables often cluster together, indicating broader mechanisms like moisture-related stress. The value lies in prioritization: deciding where to focus monitoring or intervention.

### Interactive Tutorial: Feature Importance

The companion notebook (03\_feature\_importance.ipynb) walks through model training, validation, and interpretation: train-test splitting with temporal/spatial considerations, hyperparameter explanations, dual model training (Random Forest and Gradient Boosting), performance evaluation, and feature importance visualization. The notebook includes model comparison tables, troubleshooting guidance for common issues, and saves all results as CSV files.

### Combining Factor Analysis and Feature Importance

Factor analysis and feature importance are complementary. Factor analysis reveals hidden structures like groups of related parameters representing broader processes like “moisture exposure" or “thermal stress." Feature importance quantifies how strongly each variable predicts specific outcomes. Used together, they can transform complex datasets into actionable insights. In terms of protocol, you can begin with factor analysis to identify underlying factors and inform hypotheses about relevant physical processes. Then you can apply feature importance to rank specific variables driving observed outcomes. When variables grouped by factor analysis also rank highly in feature importance, confidence in interpretation can increase whereas discrepancies can prompt further investigation.

### Demonstrated Applications in Post-Disaster Heritage Contexts

Although the workflow in this practice point is presented generically, closely related approaches; combining dimensionality reduction with feature importance analysis; have already been applied by the authors in several real post-disaster heritage contexts. In earlier work following the 2020 Beirut port explosion, dimensionality reduction strategies were used to manage large sets of correlated architectural, structural, and exposure-related variables, allowing dominant patterns in damage behavior to be identified before applying supervised learning models.[24](#endnote-24) While that work relied on feature clustering and selection rather than classical exploratory factor analysis, the objective was the same: to reduce complex, interrelated datasets into a smaller number of interpretable drivers that could be meaningfully linked to physical damage mechanisms. Feature importance analysis then quantified which attributes within those reduced sets most strongly influenced observed damage levels, revealing that social occupancy, exposure conditions and geometric characteristics often outweighed traditionally emphasized descriptors.

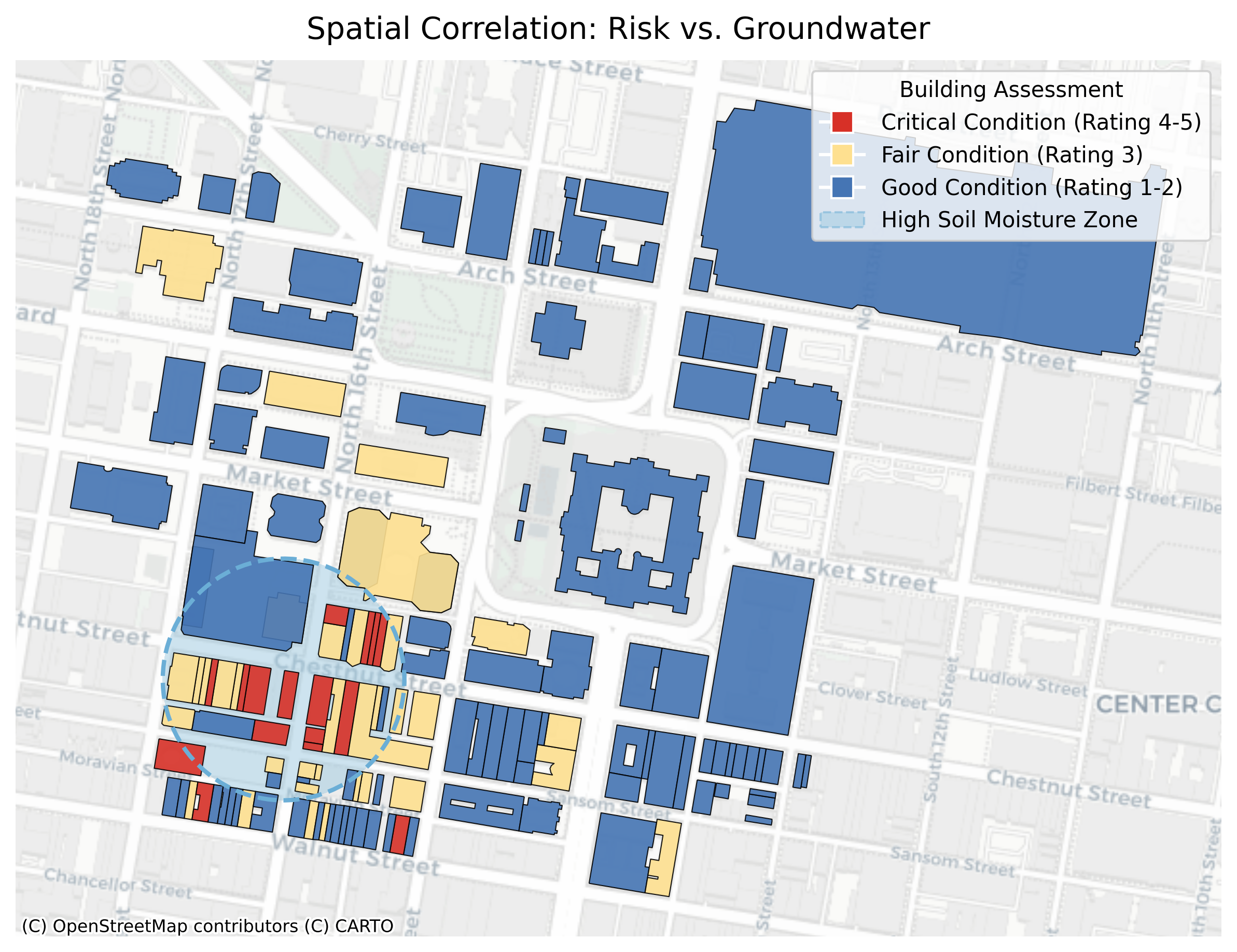
A related methodology was later applied in southern Ukraine after the 2023 Kakhovka Dam breach, where flooding affected historic masonry buildings across the Kherson region.[25](#endnote-25) In that study, dimensionality reduction was implemented through hierarchical clustering of correlated features to address multicollinearity and stabilize feature importance rankings, rather than through formal factor extraction, though the interpretation of correlated variables was guided by factor-based reasoning. The resulting reduced feature set enabled clearer interpretation of dominant damage drivers—such as duration of flood impact, proximity to the breach, foundation type, maintenance condition, social occupancy, and building orientation relative to flow direction—while reducing the risk that correlated inputs would distort feature importance rankings. Across these applications, dimensionality reduction and feature importance served complementary roles: the former structured complex datasets into interpretable groupings, while the latter translated those groupings into ranked, decision-relevant priorities that could inform rapid post-disaster assessment and intervention planning.

# Implementation: Interpreting and Using the Results

Running algorithms represents only the initial phase; the more significant challenge involves translating quantitative outputs into decisive action. In practice, these results are intended to inform collaborative decision-making among preservation architects, structural engineers, MEP specialists, and building envelope experts, rather than to prescribe solutions within a single disciplinary lens. While computational models excel at highlighting patterns, their findings require validation through professional judgment and comparison with field observations. To operationalize these results, practitioners can rank sites by risk using factor scores or predicted outcomes, establishing specific thresholds that trigger intervention. However, these data-driven priorities must be confirmed via site visits, as models may overlook recent repairs or specific local conditions. Furthermore, final decisions must weigh statistical risk against practical constraints like cost, technical feasibility, and cultural significance. Computational analysis should be viewed as a tool that supports, rather than supplants, core preservation principles. For illustrative scenarios demonstrating convergent evidence, model failure, and high uncertainty, refer to the ’Interpretation Guide’ within the feature importance notebook.[26](#endnote-26)

## Decision Framework: Reconciling Models with Field Observations

When translating model outputs into intervention priorities, visualizations help identify patterns that inform decision-making. Figure [12](#fig:spatial) maps predicted risk across building locations; geographic clustering of high-risk buildings suggests shared environmental exposures that warrant district-level rather than building-by-building interventions.



Spatial correlation between building risk and environmental factors. The clustering of high-risk buildings (red) overlaps with the high groundwater zone (blue shading), indicating a shared environmental driver. This visualization supports district-level interventions (e.g., drainage improvements) rather than isolated building repairs.

Discrepancies between model outputs and field observations are not uncommon in heritage practice, particularly when data are heterogeneous or temporally misaligned. Rather than indicating model failure, these conflicts often reveal data gaps or recent site changes. When model outputs conflict with field observations or domain knowledge, work through the following steps. First, quantify the disagreement. How large is the discrepancy? If a model ranks a building as low-risk but you observe active deterioration, document specifically what the model missed (recent water infiltration, localized spalling, etc.). Second, check data quality. Were the variables measured correctly for that building? Common issues include sensors placed in unrepresentative locations, measurements taken before recent damage occurred, data entry errors, or missing variables that would explain the discrepancy. Third, consider temporal mismatches. Models trained on historical data may not capture recent changes. A building may have been in good condition when measured but deteriorated since. Conversely, recent repairs may have improved condition beyond what historical patterns predict. Fourth, evaluate model limitations. Is the disagreement within the model’s expected error range? Fifth, weight evidence by strength. As a rough guide, assign approximate weights: direct field observation by a qualified professional (40%), model prediction with R² 0.7 and large sample (30%), historical deterioration records (20%), and model prediction with R² 0.5 or small sample (10%). These are guidelines, not rules; adjust based on data quality and expertise available. Finally, make a decision and document your reasoning. If field evidence outweighs model output, trust the field evidence but investigate why the model failed since this reveals data gaps or model limitations. If model output is strong and field evidence is ambiguous, use the model to prioritize further investigation; always document your reasoning for future reference and accountability.

# Challenges and Tips

Applying dimensionality reduction and feature importance analyses in real preservation projects requires care and judgment. Heritage data are rarely ideal since they’re often sparse, heterogeneous, and collected under varying conditions over long time spans. The goal is not to produce perfect models, but to extract useful patterns that support informed decision-making. Recognizing the limitations of both data and algorithms helps ensure that computational analysis enhances, rather than distorts, professional expertise.

Heritage datasets are often incomplete, with missing values scattered throughout. Imputation methods can help, but you should interpret results cautiously when more than 10-15% of your data required imputation. Complex models may fit noise rather than signal, a problem known as overfitting. The best defense is validation: always test your model on data it has never seen. If performance drops sharply on test data compared to training data, your model has likely memorized idiosyncrasies rather than learned generalizable patterns.

A highly accurate model has little value if you cannot explain its reasoning to practitioners or stakeholders. This matters especially in preservation, where decisions must often be justified to review boards, funding agencies, or the public. Favor transparent algorithms over black-box models when possible. Computational results should complement (not replace!) condition surveys and material analyses. The most robust decisions combine algorithmic insights with field observations and expert judgment.

Document your data sources, methods, and parameter settings so others can reproduce and verify your results. Transparency builds trust in data-driven preservation decisions and protects you if results are later questioned. What matters at the component level (moisture content in a specific wall) may differ from what matters at the district scale (groundwater rise affecting multiple buildings). Always interpret findings within the appropriate spatial and temporal context.

Models require periodic updating as new data accumulate. A reasonable schedule is annual retraining if you collect continuous monitoring data, or retraining when new data exceed 20-30% of your original training set. Monitor prediction errors over time. If MAE increases by 50% or more, your model may have become outdated due to changing conditions, a phenomenon called concept drift. For datasets spanning decades, recent 10-15 year rolling windows often perform better than cumulative data because they prioritize current conditions over historical patterns that may no longer apply.

# Interactive Tutorials

Five Google Colab notebooks provide hands-on training for practitioners with no programming experience. All code and data are available in the project repository: <https://github.com/rkn2/practicePointFeatureImp>. Start with Notebook 00 and work through sequentially. The interactive notebooks do not generate or prescribe conservation plans; instead, they apply analytical techniques to user-provided data and return interpretable outputs that practitioners evaluate using established conservation guidance, standards, and professional judgment. Selection and design of conservation actions remain the responsibility of the practitioner.

**Data Preparation** ([00\_data\_preparation.ipynb](https://colab.research.google.com/github/rkn2/practicePointFeatureImp/blob/main/00_data_preparation.ipynb)): File upload, data dictionary, handling missing values, one-hot encoding, standardization; includes optional multicollinearity check (VIF); outputs processed\_data.csv. Start here to clean and prepare your data.

**Simple Statistical Methods** ([01\_simple\_statistics.ipynb](https://colab.research.google.com/github/rkn2/practicePointFeatureImp/blob/main/01_simple_statistics.ipynb)): Correlation analysis, simple and multiple regression, group comparisons, VIF calculation with solutions for multicollinearity. Use this when you have fewer than 50 observations, 2–5 variables, or need transparent methods for stakeholder communication.

**Factor Analysis** ([02\_factor\_analysis.ipynb](https://colab.research.google.com/github/rkn2/practicePointFeatureImp/blob/main/02_factor_analysis.ipynb)): Bartlett’s and KMO tests, scree plot, Varimax rotation, factor loadings; includes discussion of Factor Analysis vs. PCA and rotation choice guidance (Varimax vs. Promax); outputs factor\_scores.csv and factor\_loadings.csv.

**Feature Importance** ([03\_feature\_importance.ipynb](https://colab.research.google.com/github/rkn2/practicePointFeatureImp/blob/main/03_feature_importance.ipynb)): Train-test split with temporal/spatial considerations, Random Forest and Gradient Boosting training, performance evaluation, model comparison; includes guidance on alternative models (Ridge, Lasso, Elastic Net, SVM, Neural Networks, KNN) with decision guide and learning resources; outputs feature\_importance.csv and model\_comparison.csv.

**Visualization** ([04\_visualization.ipynb](https://colab.research.google.com/github/rkn2/practicePointFeatureImp/blob/main/04_visualization.ipynb)): Correlation heatmaps, distribution plots, boxplots, pairplots, summary statistics; saves all plots as 300 DPI PNG files for publication.

All notebooks include beginner-friendly explanations, error handling, troubleshooting sections, and “Next Steps" guidance. The synthetic dataset includes 200 heritage buildings with 15 variables (construction year, district, material type, foundation type, temperature, rainfall, humidity, freeze-thaw cycles, soil moisture, crack width, salt deposition, condition rating, intervention urgency).

# Notes

1. Stephen D. Mikesell, “Historic Preservation That Counts: Quantitative Methods for Evaluating Historic Resources,” *The Public Historian* 8, no. 4 (1986): 61–74, <https://doi.org/10.2307/3377500>; Richard Veillon, *State of Conservation of World Heritage Properties – a Statistical Analysis (1979–2013)* (UNESCO World Heritage Centre, 2014), <https://whc.unesco.org/en/documents/134872>.
2. Veronika Samborska, “Scaling up: How Increasing Inputs Has Made Artificial Intelligence More Capable,” *Our World in Data*, 2025, <https://ourworldindata.org/scaling-up-ai>.
3. Google, *Google Colaboratory*, 2025, <https://colab.research.google.com>.
4. Trinh Nguyen, *Top 10 No Code Machine Learning Platforms to Use in 2024*, 2024, <https://www.neurond.com/blog/no-code-machine-learning-platforms>.
5. Aasim Ayaz Wani, “Comprehensive Review of Dimensionality Reduction Algorithms: Challenges, Limitations, and Innovative Solutions,” *PeerJ Computer Science* 11 (2025), <https://doi.org/10.7717/peerj-cs.3025>; Imola K. Fodor, *A Survey of Dimension Reduction Techniques* (Lawrence Livermore National Laboratory, 2002), <https://computing.llnl.gov/sites/default/files/148494.pdf>.
6. Karen Grace‐Martin, *How Big of a Sample Size Do You Need for Factor Analysis?*, 2021, <https://www.theanalysisfactor.com/sample-size-needed-for-factor-analysis>.
7. Huazhen Wang et al., “An Experimental Study of the Intrinsic Stability of Random Forest Variable Importance Measures,” *BMC Bioinformatics* 17 (2016): 60, <https://doi.org/10.1186/s12859-016-0900-5>.
8. M. Luz Calle and Víctor Urrea, “Letter to the Editor: Stability of Random Forest Importance Measures,” *Briefings in Bioinformatics* 12, no. 1 (2011): 86–89, https://doi.org/<https://doi.org/10.1093/bib/bbq011>.
9. Iván Palomares Carrascosa, *Understanding Correlation and Causation*, 2024, <https://www.statology.org/understanding-correlation-and-causation>; Jim Frost, *Correlation Vs Causation: Understanding the Differences*, 2023, <https://statisticsbyjim.com/basics/correlation-vs-causation>.
10. David Goretzko et al., “Exploratory Factor Analysis: Current Use, Methodological Developments and Recommendations for Good Practice,” *Current Psychology* 40, no. 6 (2021): 3510--3521, <https://doi.org/10.1007/s12144-019-00300-2>.
11. Brett Williams et al., “Exploratory Factor Analysis: A Five-Step Guide for Novices,” *Journal of Emergency Primary Health Care (JEPHC)* 8, no. 3 (2010): 1447–4999, <https://journals.sagepub.com/doi/pdf/10.33151/ajp.8.3.93>.
12. Mirka Saarela and Susanne Jauhiainen, “Comparison of Feature Importance Measures as Explanations for Classification Models,” *SN Applied Sciences* 3, no. 272 (2021), <https://doi.org/10.1007/s42452-021-04148-9>.
13. Leo Breiman, “Random Forests,” *Machine Learning* 45, no. 1 (2001): 5–32, <https://doi.org/10.1023/A:1010933404324>.
14. Jerome H Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics* 29, no. 5 (2001): 1189–232, <https://doi.org/10.1214/aos/1013203451>.
15. Zhengze Zhou and Giles Hooker, “Unbiased Measurement of Feature Importance in Tree-Based Methods,” *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, no. 2 (2021): 1–21, <https://doi.org/10.1145/3429445>.
16. Sumanta Basu et al., “Iterative Random Forests to Discover Predictive and Stable High-Order Interactions,” *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 8 (2018): 1943–48, <https://doi.org/10.1073/pnas.1711236115>.
17. Joe Kallas and Rebecca Napolitano, “Understanding Critical Masonry Building Attributes Shaping Vulnerability to Blast Loads: Data-Driven Insights from the 2020 Beirut Explosion,” *International Journal of Disaster Risk Reduction* 110 (2024): 104640, https://doi.org/<https://doi.org/10.1016/j.ijdrr.2024.104640>.
18. Joe Kallas and Rebecca Napolitano, “Understanding Key Building Attributes Influencing Flood Vulnerability in Masonry Structures After the 2023 Kakhovka Dam Breach in Ukraine,” *J. Comput. Cult. Herit.* (New York, NY, USA) 18, no. 3 (2025), <https://doi.org/10.1145/3744743>.
19. Joe Kallas and Rebecca Napolitano, “Beyond Field Surveys: Understanding the Role of 3D Spatial Attributes for Data-Driven Blast Vulnerability Assessment of Masonry Buildings,” *International Journal of Disaster Risk Reduction* 128 (2025): 105672, https://doi.org/<https://doi.org/10.1016/j.ijdrr.2025.105672>.
20. Jayita Gulati, *Logistic Vs SVM Vs Random Forest: Which One Wins for Small Datasets?*, 2025, <https://machinelearningmastery.com/logistic-vs-svm-vs-random-forest-which-one-wins-for-small-datasets/>; Fiona K. Ewald et al., *A Guide to Feature Importance Methods for Scientific Inference*, arXiv preprint arXiv:2404.12862, 2024, <https://arxiv.org/abs/2404.12862>.
21. Penn State STAT 462, *Detecting Multicollinearity Using Variance Inflation Factors*, 2025, <https://online.stat.psu.edu/stat462/node/180>.
22. Meenu Bhagat and Brijesh Bakariya, “A Comprehensive Review of Cross-Validation Techniques in Machine Learning,” *International Journal of Smart and Archived Technology* 16 (2025): 2229–7677, <https://doi.org/10.1234/ijsat.2025.01.1305>.
23. Saarela and Jauhiainen, “Comparison of Feature Importance Measures as Explanations for Classification Models.”
24. Kallas and Napolitano, “Understanding Critical Masonry Building Attributes Shaping Vulnerability to Blast Loads”; Kallas and Napolitano, “Beyond Field Surveys.”
25. Kallas and Napolitano, “Understanding Key Building Attributes Influencing Flood Vulnerability in Masonry Structures After the 2023 Kakhovka Dam Breach in Ukraine.”
26. Rebecca Napolitano and Joe Kallas, “Dimensionality Reduction for Heritage Preservation: Interactive Companion Notebooks,” in *GitHub Repository*, <https://github.com/rkn2/practicePoint>; GitHub, 2025.