Occurrence of Crystals in Switzerland



CIP02 – Data Collection, Integration & Preprocessing

|  |  |
| --- | --- |
| Lucerne University of Applied Sciences & Arts  MSc in Applied Information and Data Science  Spring Semester 2025  **Authors:**  Andreas Goerre  Barbara Maier  Sheena Walker | **Lecturer Info**  Andreas Melillo  Ramón Christen |

**Table of Contents**

[1 Introduction 1](#_Toc194941927)

[2 Preprocessing 1](#_Toc194941928)

[3 Analysis 1](#_Toc194941929)

[3.1 Exploratory Data Analysis 1](#_Toc194941930)

[3.2 Further Analysis 1](#_Toc194941931)

[3.2.1 Distribution of crystal occurrences in Swizerland 1](#_Toc194941932)

[3.2.2 Highest frequency of occurrence of crystal types in Switzerland 1](#_Toc194941933)

[3.2.3 Statistical Correlation between crystal occurrences and elevation Level 1](#_Toc194941934)

[4 Conclusion 2](#_Toc194941935)

# Introduction

Switzerland's diverse geology and alpine terrain host a rich variety of crystals, from quartz to rare minerals. These formations are influenced by temperature, pressure and mineral composition, making altitude a key factor in their distribution. This project investigates the correlation between crystal occurrence and elevation in Switzerland. We aim to determine whether certain crystals are more abundant at certain altitudes and how geological formations influence their diversity.

Using a public data source (Mindat.org), we will apply Python-based data analysis to identify patterns. Our approach includes data collection, cleaning, (statistical) analysis, and visualization to uncover relationships between elevation and crystal formation.

The results will be valuable for mineralogists, geologists, and collectors, with potential applications in other mountainous regions.

# Preprocessing

An important part for making a valuable analysis of the occurrence of crystals and their types in Switzerland was the sourcing and preprocessing of the data.

To get an extensive and useful dataset of all mineral occurrences in Switzerland we used WebScraping with our main datasource mindat.org. Afterwards we processed, cleaned and enriched the data in several ways, such that we had one final dataset with all needed datapoints. Following the process is described in detail.

## 2.1 Web Scraping

We began by identifying the relevant HTML elements on the target website that contained the data we aimed to scrape from the tables – specifically the locinfodiv and newlocminlist sections. Initially, we wrote a basic script that was capable of scraping data from a single locality. This was only successful after switching from BeautifulSoup4 to Selenium, paired with the ChromeDriver, since the content was not rendered in the static HTML and therefore not accessible through BeautifulSoup.

Subsequently, we extended this code to allow it to first collect all relevant links to individual localities displayed on the map and then to extract the information from each of these links, storing the results in a structured table format. After a brief test run – where we initially limited the script to collecting only three links – we launched the extended code.

However, several issues arose. One of the main problems was that the browser kept closing unexpectedly without any error messages or explanation. In response, we implemented multiple try and except blocks to catch and handle potential errors during execution. While this mitigated some issues, it did not fully resolve the problem. Because the website is designed to prevent automated scraping, we made further adjustments to disguise the script’s behavior: we introduced rotating user agents, installed undetected\_chromedriver, and incorporated randomized wait times (between 5 and 15 seconds) between requests to simulate more human-like behavior. Despite these measures, the script still terminated unexpectedly after approximately one hour of execution. After installing a VPN provider, the issue was finally resolved, and the script was able to run for five consecutive hours, successfully collecting data from around 520 links.

Unfortunately, the scraping logic – designed to navigate through the raw HTML structure rather than zooming and scrolling the map manually – sometimes only captured summary pages instead of individual locality pages. This occurred in cases where Mindat displayed an overview of localities within a region, rather than listing them individually. To investigate this, we re-scraped one of the summary pages and observed that, initially, only the aggregate page was shown. However, after increasing the wait times before scraping individual localities, the full list of detailed localities became accessible. Due to time constraints, we did not rerun the script to scrape these pages individually, nor did we extend the code to reprocess all previously collected pages.

Finally, it is worth noting that Mindat does offer a public API, but it did not provide the specific data we required. Therefore, we decided to proceed with the dataset gathered through our customized scraping process.

## 2.2 Enrichment of Dataset – Elevation level and Categories

For doing the planned analysis we needed to add the elevation level based on the coordinates and also add a categorization for the mineral types.

In order to enrich the dataset with the elevation of each location, we needed to remove the inaccurate coordinate points and also clean up the coordinates so that the format was the same for all accurate entries. The original data contained a mix of coordinate formats, including decimal degrees and degrees-minutes-seconds (DMS), as well as some incomplete or placeholder values. These inconsistencies were addressed by standardising all valid entries into decimal format and discarding those that were either unrecognised, labelled 'unknown' or identified as rough approximations such as '46.00000,8.00000'.

Once the data was cleaned, each coordinate pair was enriched with elevation information using a number of external elevation APIs. A fallback mechanism was implemented to maximise data coverage: if the first service failed or returned no result, the system would automatically try the next. The APIs used in this sequence were Open-Elevation, OpenTopoData and Open-Meteo. The elevation values obtained from these sources were then cleaned and converted to a common integer format.

The final output was exported as an Excel file containing three sheets: one with the original raw data, one with the cleaned and validated coordinates, and one with the enriched data containing the elevation for each valid location. This process ensures that the resulting dataset is both reliable and enhanced with valuable topographic context.

For the enrichment of the categories for all mineral types....

## 2.3 Transformation of Dataset

....

## 2.4 Final Cleaning and Enrichment

To finalize the dataset we did some further minor cleaning for the column names.

After an initial analysis was done, it got clear very fast that the elevation levels were not sufficient as there were to many different numbers. Therefore, these values were also categorized based on the altitude zones. This made the analysis easier and clearer. The location types were not used for that matter as there were also to many different categories and it was not clear based on which information these were classified.

# Analysis

This chapter presents an analysis of crystal occurrences in Switzerland. It begins with an exploratory overview of the dataset, including its structure, key variables, and the number of unique minerals and locations. The subsequent analysis is organized around three main research questions: the spatial distribution of crystal sites, the most frequently occurring crystal types, and the correlation between crystal occurrences and elevation. Each of these topics is addressed using appropriate tools such as GIS mapping, frequency analysis, and statistical modelling to uncover patterns and provide geological insights.

## 3.1 Exploratory Data Analysis

The final dataset under consideration contains information on 319 unique crystal discovery locations and includes 864 distinct mineral types. The dataset is structured into eight columns, each containing specific geographical, geological, and classification data. Notable columns include Altitude (integer), Altitude Category, Mineral, and Mineral Category (all as object types). Location data is provided via Mindat Locality ID and Latitude & Longitude. The dataset also includes environmental context through Köppen climate type and Location Type. The dataset's overall structure is conducive to exploratory and correlation analysis.

## 3.2 Further Analysis

To structure our analysis, we follow 3 topics with each one research question.

### Distribution of crystal occurrences in Swizerland

1. **What is the spatial distribution of crystal occurrences across different regions in Switzerland?**To achieve this we will:

* Collect geolocation data on known crystal sites from mineralogical databases and geological surveys.
* Use GIS tools (e.g., GeoPandas, Folium) to map occurrences across different regions.
* Categorize sites by geographic features (e.g., Alpine vs. Jura regions).
* Identify spatial clustering patterns and regional mineral diversity.
* ***Pro Kategorie eine Farbe und dann auf eine CH-Karte einzeichnen (Andy)***

### Highest frequency of occurrence of crystal types in Switzerland

1. **What are the most common crystal types found in Switzerland?**For this we will:

* Extract and classify crystal occurrences by **mineral type**.
* Analyze frequency distributions of different crystal types across Switzerland.
* Compare findings with **geological literature** to validate classification.
* Identify **potential geological factors** influencing crystal variety.
* ***Auf Top 10 minerals und dann auf Kategorien (Barbara)***

### Correlation between crystal occurrences and elevation Level

In order to examine the relationship between crystal occurrences and elevation, both Pearson's and Spearman's correlation coefficients were utilized. Pearson's coefficient assesses linear relationships between variables, while Spearman's coefficient evaluates monotonic trends without making the assumption of linearity. The employment of both methods ensures a more robust understanding of potential associations, irrespective of the distribution or structure of the data.

Across all possible combinations of minerals, mineral categories, altitude and altitude categories, the correlation coefficients were found to be nearly negligible, with the majority of the p-values exceeding the commonly accepted significance threshold of 0.05. This finding suggests that there is no statistically significant relationship between elevation and the type of crystal or mineral found in the vast majority of cases.

The sole statistically significant outcome, which pertained to the relationship between Altitude Category and Mineral Category, revealed a modest negative correlation (r ≈ -0.035). Despite the p-value falling below 0.05, the extremely low effect size and minimal explained variance (R² = 0.001) indicate that this correlation is not practically meaningful and may be due to chance or external factors.

The scatter plots (RQ3-2) confirmed the absence of a trend; however, they were visually dense and limited in interpretability due to the number of unique mineral types. Conversely, the box plots (RQ3-3), which were centered on mineral categories, were more legible and exhibited overlapping distributions across altitude levels, thereby corroborating the modest statistical outcomes. The histogram (RQ3-4) facilitated the visualization of the frequency of mineral categories across altitude categories; however, any observed patterns are likely attributable to the concentration of sampling rather than to a genuine correlation. Due to the ambiguity and limited interpretive value of the plots, they were excluded from the final documentation.

To answer the research question – *Is there a statistically significant correlation between the occurrences of crystals and the elevation of the place found?* – a comprehensive statistical analysis and meticulous visual inspection have been conducted, leading to the conclusion that there is an absence of a substantial or significant correlation between the presence of crystals and elevation. While a modest statistical significance was identified in a single instance, it is of negligible practical relevance. Consequently, the dataset does not support the hypothesis that elevation is a determining factor for mineral type occurrence.

# Conclusion

# Outlook