# 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

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.

# 2 Methodology & 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

Mindat.org is a platform where both professionals and hobby mineral collectors mark locations on a map, indicating where they have found crystals and specifying the type of mineral discovered. To access this information, we had to extract the data from the map through web scraping.

We began by identifying the relevant HTML elements on the target website that contained the data we aimed to scrape. 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 accessible through BeautifulSoup.

We extended the script to gather all relevant links to localities and scrape data from them, storing the results in a structured table format. After a test run with three links, we expanded the script to collect data from 520 links.

Several challenges arose, particularly with the browser closing unexpectedly. We implemented error handling with try/except blocks and added rotating user agents, undetected_chromedriver, and randomized wait times to simulate human-like behavior. Despite this, the script still terminated after about one hour. Installing a VPN resolved the issue, allowing the script to run for five hours without interruptions. 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. . After adjusting the wait times, we were able to access full lists of localities. However, due to time constraints, we did not rerun the script to scrape all detailed pages.

Though Mindat offers an API, it did not provide the specific data we needed, so we proceeded with the dataset gathered through our customized scraping method.

## 2.2 Enrichment of Dataset – Altitude 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.

To enrich the dataset with elevation data, inaccurate or inconsistent coordinate entries were removed and valid entries were standardized to decimal format. The original data contained a mix of formats - such as decimal degrees and DMS - as well as incomplete or placeholder values, which were either corrected or discarded. The cleaned coordinate pairs were then processed through a number of external elevation APIs (Open-Elevation, OpenTopoData, Open-Meteo) using a fallback mechanism to ensure maximum coverage. The retrieved elevation values were cleaned and converted to a uniform integer format. The final dataset was exported as an Excel file with three sheets: raw data, cleaned coordinates, and enriched elevation data, ensuring both accuracy and added geographic context.

To add mineral categories to our dataset, we reviewed the mineral entries and identified appropriate categories for each. Since we are not professional mineralogists, we consulted generative AI (ChatGPT) to assist with this process.Based on these suggestions, we grouped minerals according to common rock-forming categories in Swiss geology, including quartz, feldspar, mica, and amphibole, which are frequently mentioned in scientific literature and fieldwork.

Remaining minerals were assigned to broader chemical–crystallographic groups (e.g., silicates, carbonates, sulfides) following conventional mineral classification systems.This hybrid classification approach combines practical field-based geological knowledge with academic frameworks, providing a clearer interpretation for visualizing the frequency and distribution of mineral types.

## 2.3 Transformation of Dataset

The transformation of the dataset for analysis consists of two steps; normalizing the dataset and cleansing of the mineral names. After the scraping process, each row of the dataset listed a scraped location along with all the minerals found there on a single line. For further processing, it was necessary to transform the dataset so that each row represents a single location-mineral pair.

Upon inspecting the dataset, we observed that some minerals appeared under multiple name variations or included minor distinctions that were irrelevant to our analysis. Therefore, we performed a cleansing of the mineral names, applying the following measures:

- Remove quotation marks
- Remove “var.” prefixes which highlighted that different types of the same mineral were found
- Remove group classification names (Group, Supergroup, Subgroup)
- Remove extra spaces
- Remove chemical elements in parentheses (e.g. “(Fe)”)

### 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.

### 2.5 Analysis Methods

To illustrate the spatial distribution of minerals in Switzerland, we developed an interactive map and heatmap using the folium and streamlit libraries. The heatmap applies a blue-to-green color scale to represent mineral occurrence intensity. A filter function allows users to select specific minerals and view their locations. To enhance accessibility for non-experts, representative mineral images are displayed using data retrieved from the Wikipedia API.

The analysis was based on categorizing all recorded minerals into two groups: rock-forming mineral groups and broader chemical-crystallographic categories. Frequencies were then calculated across the dataset and a bar chart was created to visualize the most common crystal types found in Switzerland.

Pearson's and Spearman's correlations were utilized to assess the relationships between crystal occurrences and elevation. Mineral types and categories were evaluated in relation to altitude and altitude categories. The analysis was complemented by linear regression. The use of visualizations was instrumental in elucidating the findings of the study.

## 3 Analysis & Results

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 Results

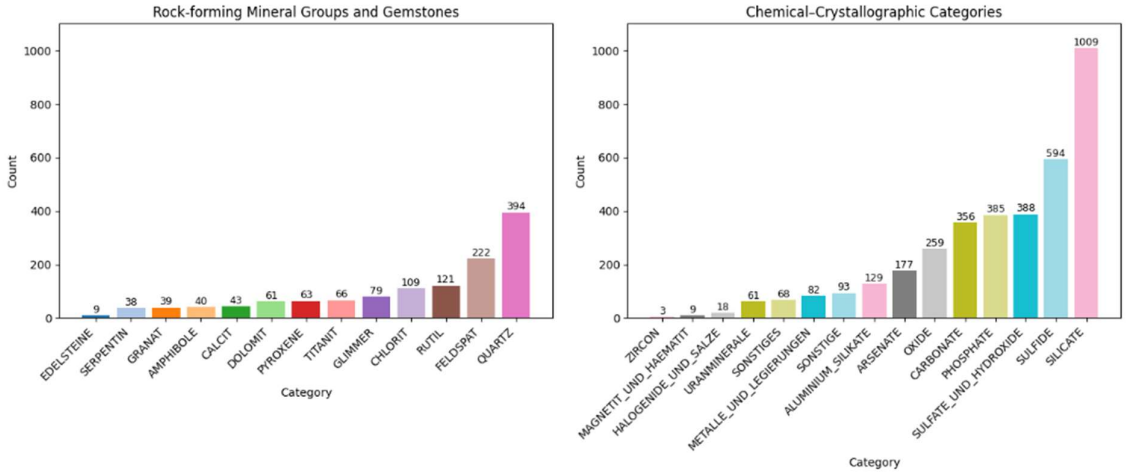
This chapter presents the results of the data analysis, including the distribution of crystal occurrences in Switzerland, the most frequently found crystal types, and the correlation between crystal occurrences and elevation.

### 3.2.1 Distribution of crystal occurrences in Switzerland

The web app demonstrates that minerals can be found throughout Switzerland wherever rocks are present. However, alpine regions—particularly the canton of Valais—show a notably higher concentration of mineral findings. All mineral types appear across these high-occurrence regions, and no clear spatial separation between different minerals can be observed. The visualizations highlight not only the geographic patterns but also the natural beauty of these minerals, showcased through detailed images. The app can be accessed online at swisscrystals.streamlit.app or run locally using the command streamlit run app.py in the respective directory of the app.py script.

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

The categories applied in our dataset consist of two main groups: first, the **rock-forming mineral groups,** *and second all remaining ones*, classified into broader **chemical–crystallographic categories.** To answer our research question – *What are the most common crystal types found in Switzerland?* – we created the following bar chart.

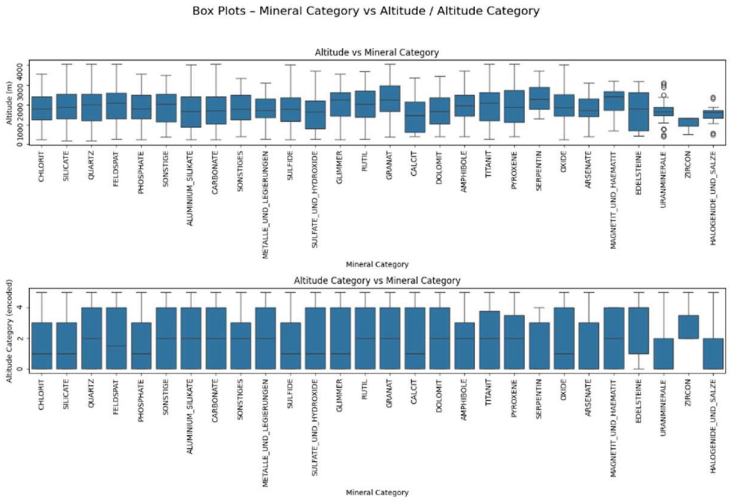



As we can see in the table, quartz and feldspar, are among the most commonly found crystals in Switzerland. This aligns with the well-known fact that these minerals are typical and widespread in the region. Since silicates are by far the most common mineral group – making up over 90% of the Earth's crust and being especially prevalent in mountainous regions like the Alps – our results are also consistent with this general knowledge.

At first glance, it may appear that only a small number of gemstones are listed. However, this is because many gemstones are categorized under broader mineral groups—for example, under "Quartz," which includes stones such as Rock Crystal, Smoky Quartz, Rose Quartz, and many others.

### 3.2.3 Correlation between crystal occurrences and elevation Level

Both Pearson and Spearman correlations were calculated to examine the relationship between crystal occurrence and elevation. Overall, the correlations were negligible, and most p-values exceeded 0.05, indicating no statistically significant relationships. One exception - Altitude Category vs. Mineral Category - was statistically significant ($p < 0.05$), but had a negligible effect size ($r \approx -0.035$, $R^2 = 0.001$), suggesting no practical relevance.



The box plots show broad, overlapping distributions across mineral categories, with no clear patterns related to elevation. Narrower ranges in some categories are likely due to small sample sizes rather than true trends.

However, Scatter plots were too dense to be clearly interpreted, while box plots and histograms showed overlapping or sampling driven patterns without strong trends. Regression analyses by mineral category yielded some statistically significant results, with 'GRANAT' showing the strongest effect ($R^2 = 0.168$), followed by weaker effects for 'DOLOMITE', 'SERPENTIN' and others. Most of the slopes were negative, suggesting that higher elevations have fewer of the more highly coded minerals.

In conclusion, no substantial or meaningful correlation was found between crystal occurrence and elevation, and elevation does not appear to be a determinant of mineral type.

# 4 Conclusion & Outlook

This study aimed to investigate the spatial distribution of crystals in Switzerland, the most frequent crystal types, and the correlation between crystal occurrence and elevation. Our analysis found that while there is no significant correlation between elevation and crystal occurrences, higher concentrations were observed in the Alpine regions, especially in the canton of Valais. The distribution of crystal types revealed that quartz and feldspar were the most commonly found minerals. However, due to time and technical constraints, we did not scrape all the pages from Mindat.org, meaning the dataset represents not a complete picture of crystal occurrences in Switzerland. Additionally, categorizing over 800 mineral types was challenging, and expert input would be more reliable than using AI for classification. This would have extended the scope of the project beyond our initial plan.

Future work could involve scraping the entire dataset from Mindat.org to get a more comprehensive understanding of crystal occurrences. Furthermore, refining the mineral classifications with expert knowledge would provide a more accurate and detailed dataset. Future studies could also explore additional geological factors or advanced techniques to identify further patterns and relationships in crystal distribution.