Alexander Gustafson
CU ID: algu6965
GitHub: alex-goose
Data Mining Project Milestone 4

Progress Report of Exploring Snow Depth Trends: Analysis, Prediction, and Insights Using Advanced Data Mining Techniques

Abstract

This study aims to explore the impact of elevation on snow depth variations in high-altitude German cities and contrasting regions across France. The objective is to understand how geographical elevation influences snow accumulation patterns, which is critical for addressing challenges related to climate change, urban planning, and environmental management. Using data mining techniques, specifically cluster analysis and regression analysis, we examined snow depth measurements from over 100 cities, focusing particularly on cities situated above 1000 meters in Germany. The results reveal a significant correlation between elevation and snow depth, with higher elevations consistently showing greater snow accumulation. This study not only confirms the expected impact of elevation on snow depth but also highlights regional variations that suggest other climatic factors at play. The findings provide valuable insights for environmental policy makers and contribute to the broader understanding of snow distribution patterns in European high-altitude regions.

Introduction

The primary questions asked in this project are how has average snow depths varied over the last 50 years, is there a significant difference in average snow depths at different elevations within a localized region, and how intensely are ski towns going to be affected by rising temperatures?

Understanding the variation in snow depths over the last fifty years, the differences in snow accumulation at various elevations, and the potential impact of rising temperatures on ski towns is crucial for multiple reasons. Firstly, analyzing historical snow depth changes can help us

detect climate patterns and predict future weather conditions. This information is vital for things such as agricultural planning, water resource management, and ski towns way of life. Secondly, recognizing how elevation influences snow depths within specific areas aids in environmental and urban planning, especially in regions prone to snowfall. It ensures that infrastructure is designed to cope with typical snow loads and helps in the management of natural habitats affected by snow levels. Lastly, assessing the impact of climate change on ski resorts is essential for the tourism industry. Ski towns rely heavily on consistent snowfall; understanding how rising temperatures might reduce snow reliability can spur necessary adaptations in these communities, potentially influencing economic strategies and sustainability practices.

Related Work

The previous work done on this data is from the paper "Observed snow depth trends in the European Alps: 1971 to 2019"[1] which describes where the data came from, what it contains, and their own analysis on the data. It came from the European Geosciences Union, which was able to receive the data from the six countries discussed in the data. This paper acknowledges that the data is more in depth in some countries than others, mainly due to some countries not having observation towers in many regions in the beginning. Most countries significantly increased the amount of observation areas several years to decades after when this study began. The study found that across all regions, the average mean snow depth was -8.4% per decade and that the maximum snow depth was down -5.6% per decade. It also notes that seasonal trends varied greatly across regions of similar elevations, which is an interesting pattern that I hope to replicate.

Most other work discussing the snow depth in the European Alps comes from another study that looked at trends over the last 600 years, which found an unprecedented decrease over that period of time. I hope to find similar patterns by looking at my more in depth data from a shorter time frame.

Data Set

[1]Matiu, Michael, et al. "Observed Snow Depth Trends in the European Alps: 1971 to 2019." The Cryosphere, Copernicus GmbH, 18 Mar. 2021, tc.copernicus.org/articles/15/1343/2021/.

The dataset is a combination of several datasets that record snow depth measurements as well as many other relevant measurements within Austria, France, Germany, Italy, Slovenia, and Switzerland. is composed of 20 sub-data sets, with 7 representing daily measurements, 12 representing monthly measurements, and one representing the meta data. Each sub-data set represents a region indicated by the country and the area. For example, "data_monthly_AT_HZB" represents the country of Austria and within it are over 100 different cities where data was collected.

The data was not collected from one entity, but was instead collected individually from each country. For example, the data used for Austria was taken from the Austrian Hydrological Service while France was taken from the national weather service Météo-France. This means that consistency will be an issue as several countries did not begin collecting the data at the same time and include different measurements. Also, the number of weather stations varied greatly, with Germany doubling the number of stations in the late 70s yet they steadily decreased in number beginning in the early 2000s. However, in total the weather stations grew and reached their peak numbers in the 80s through the early 2000s.

Due to the variability of the data and when the data collection began, clustering will be used to group together regions that had similar timescales for data collection. Also, pruning will be used to cut down on the amount of regions looked at by removing those with too small of a time scale to find patterns in. Certain regions may be entirely eliminated if they are deemed to be too sporadic, inconsistent, or too late starting.

The measurements that remain consistent include HNsum, which gives the measurement of total snow depth (in centimeters). HSsum represents the mean snow depth, and HSmac measures the maximum snow depth. This project will focus on the monthly data, which explains where the maximum and mean snow depths come from and how they relate to the total snow depth. There are several gap filled values such as HSmean_gapfill and Frac_gapfilled that indicate how many observations had to be filled in from other nearby stations due to missing values. While this is important information, this project will likely not use columns like this and will instead trust that the values given are close enough to the true values to perform a good analysis on. Lastly, Month and Year remain consistent between data sets, representing the year that data was collected and the month indicated by 1-12, with 1 indicating January and 12 indicating December.

Milestones for this project: The first milestone was to clean the data, which was done by the end of week 10. The data cleaning included pruning values and merging the datasets. Since the data came from many different sources this was a lot of work and the tools listed later were used to aid in this. The second milestone was data mining to find interesting patterns which was done by the end of the spring break, around March 31st. Once this was done, I began a data analysis on the patterns found which was done in early April, between weeks 11 and 12. The third milestone was using this newfound knowledge to create visualizations so that I can share the results with more people, as well as help myself gain a better understanding. Lastly, I attempted to use the patterns found to train some machine learning models to make predictions on the next few decades which proved troublesome, but I plan to keep working on it after this project. I hope to be able to expand on this project in more detail so that it can be part of my portfolio. Week 14 was used to do any last minute cleaning to make the project presentable and I practiced presenting the data.

Main Techniques Applied

Pandas was used to gain a general understanding of the dataset, and was crucial in determining how large the dataset is. A jupyter notebook was created on my local machine in VS Code, and the datasets have been imported in. Pandas was used to load in the csv files where general commands have been used such as .nunique(), .max(), and .head() to determine what the datasets looked like and what they contain. With millions of data points across all of the csv files, pandas was used more to combine the data into one general dataframe. The initial plan was to load the data into a SQL database, but the ease of use and simplicity of pandas was necessary. Also, it was helpful to have the commands and results displayed to the screen instead of in the terminal or within a python script, since a jupyter notebook was used. This allows a user to go through the notebook and see everything that was done to the data to filter it and how it was queried and utilized.

Data clustering was performed on several data sets, comparing elevation to the average snow depth measurement. The elbow method was used to determine how many clusters to use, and 4 was the ultimate choice. The results were both printed out in statistical evaluations as well as plotted for comparison which can be seen at the end of this paper.

I had spent most of my time merging datasets and found elevation and max monthly snow depth to be the most interesting data. Due to this, instead of looking at broad scale views of the data (all countries or all cities within one country), I pruned down the data to find regions with significantly higher elevations than most and looked at their long-term trends. I decided to split it up like this due to the intense variation of elevation from country to country. When k-means clustering was applied to elevation (in meters) within Germany, the two centroids were [577.13, 1614.17]. When k-means clustering was done in France, we ended up with [2004.25, 1272.69], which showed how much higher the regions being studied were within France. Even so, areas in France that were not at what would be considered extremely high elevations, like Chamonix which sits only at 3,396 feet, had very high snow depths compared to regions at the same elevations elsewhere.

An example of this new approach was, within the data_monthly_DE_DWD.csv, I found that the dataset began in 1937. I pruned the data to only include data from 1970 and later. After that, I merged this with the meta_all.csv to include elevation as a new column in the Germany dataset. Once that was done, I pruned once more to only include data where the elevation was above 1000 meters, since higher elevation regions were the focus given that snow depth trends were the motivation. I decided to focus on HSmax, which is the max recorded depth of that month (determined as the max of all daily samples) and used ‘dropna’ to drop any rows where this feature was ‘NaN’. Finally, I noticed almost all of the months had either a value of zero or a low value for most months, so I dropped all rows where the recorded month was not in November, December, January, or February since these months have the most and the most interesting snowfall patterns. I then chose the city ‘Zugspitze’ to focus on because its elevation of 2,964 meters showed heavy snowfall through the past few decades. I prepped the data for plotting with a few more sorting and cleaning methods and was able to use Plotly Express to plot the data. The graph showing the result from this can be found in the visualizations section at the end of this paper. A trend line was added to show change over time, and it is clear that there is a pattern where the max snow depth per month is slowly decreasing. This is not surprising, and it is very interesting to be able to use real data to see how climate change is affecting snow depth trends in the German Alps.

Key Results

The analysis conducted yielded two significant insights regarding variations in snow depth. First, the data confirmed a predictable pattern: an increase in elevation correlates with an increase in mean snow depth across different regions. Second, and more critically, it was found that average snow depths are decreasing over time, regardless of elevation. This trend persists across various geographical regions, indicating broader environmental changes that goes beyond local or regional climatic conditions.

Applications

The findings from this study offer valuable insights with practical and strategic implications. The observed decline in snow depths, irrespective of elevation, underscores the urgent need for environmental policies aimed at combating climate change. Policymakers can leverage this data to advocate for more stringent measures to reduce global warming, which is likely contributing to the reduced snowfall. This evidence is crucial for raising awareness about the changing climate in mountainous regions, often perceived as permanently snowy. Conservation efforts can also be tailored to preserve alpine ecosystems, which are particularly vulnerable to climate fluctuations.

For companies like Vail Resorts and Aspen, these insights are critical for strategic planning and long-term investments. Understanding that snow depths are declining even in traditionally snowy regions like the Alps can influence decisions on future resort purchases or expansions. The data serves as a basis for diversifying business models to include year-round tourist activities or investing in artificial snowmaking technologies, ensuring continued visitor engagement regardless of natural snowfall changes.

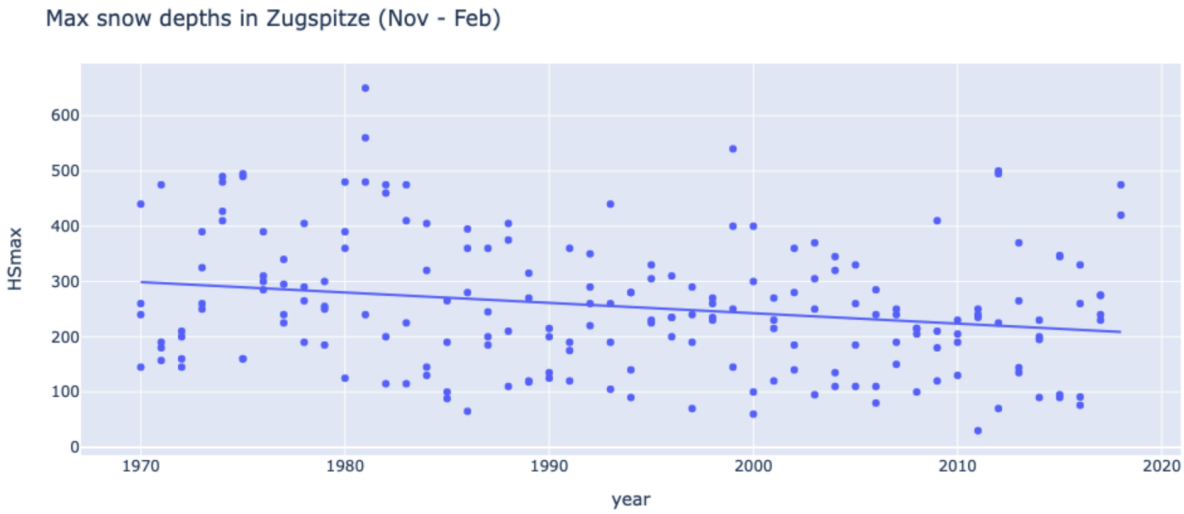
Moreover, the decreasing trend in snow depths can guide ski resort developers to consider locations where the decline is slower or where elevation might counterbalance the effects of warming temperatures. This strategic approach can optimize returns on investment by prioritizing areas with better long-term snowfall prospects. The concept of "climate resilience" can become a key factor in real estate and infrastructure planning within the ski industry, focusing on adaptive strategies to handle changing snowfall patterns.

A future research project could involve a detailed analysis of individual ski resort towns to determine which ones are experiencing the slowest rates of snow decline. This would provide more localized insights that could help predict which locations might offer the best conditions for snow sports in the coming decades. Expanding the dataset to include more variables, such as local climate mitigation efforts or changes in land use, could further refine the predictions and recommendations for ski resort management and development.

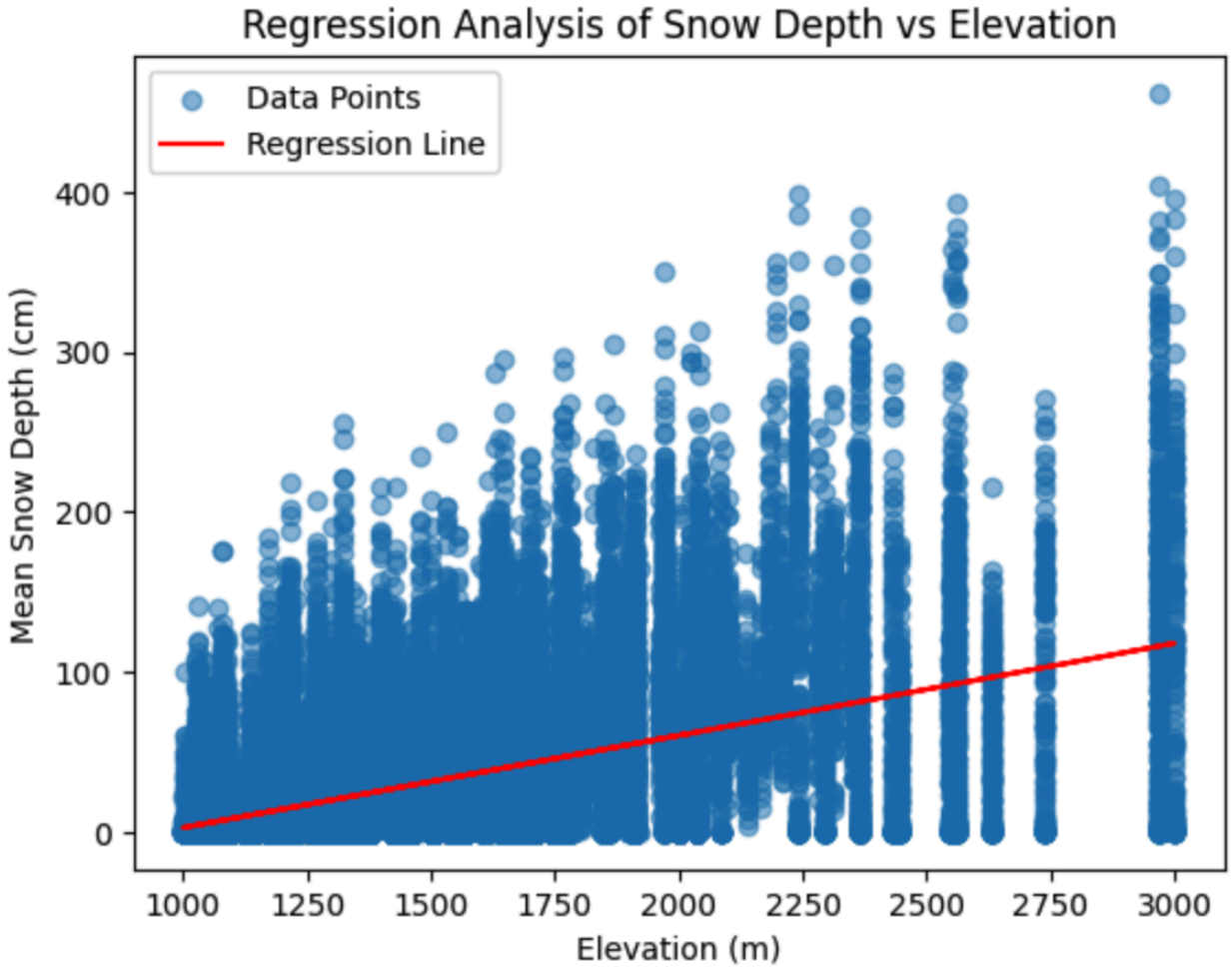
Overall, the knowledge gained from this study is crucial for understanding and adapting to climate impacts in high-elevation regions and for strategic business planning in industries reliant on natural snowfall, such as winter sports and tourism. By translating these insights into action, stakeholders can better prepare for future environmental conditions and make informed decisions that align with both economic goals and sustainability principles.

Visualizations

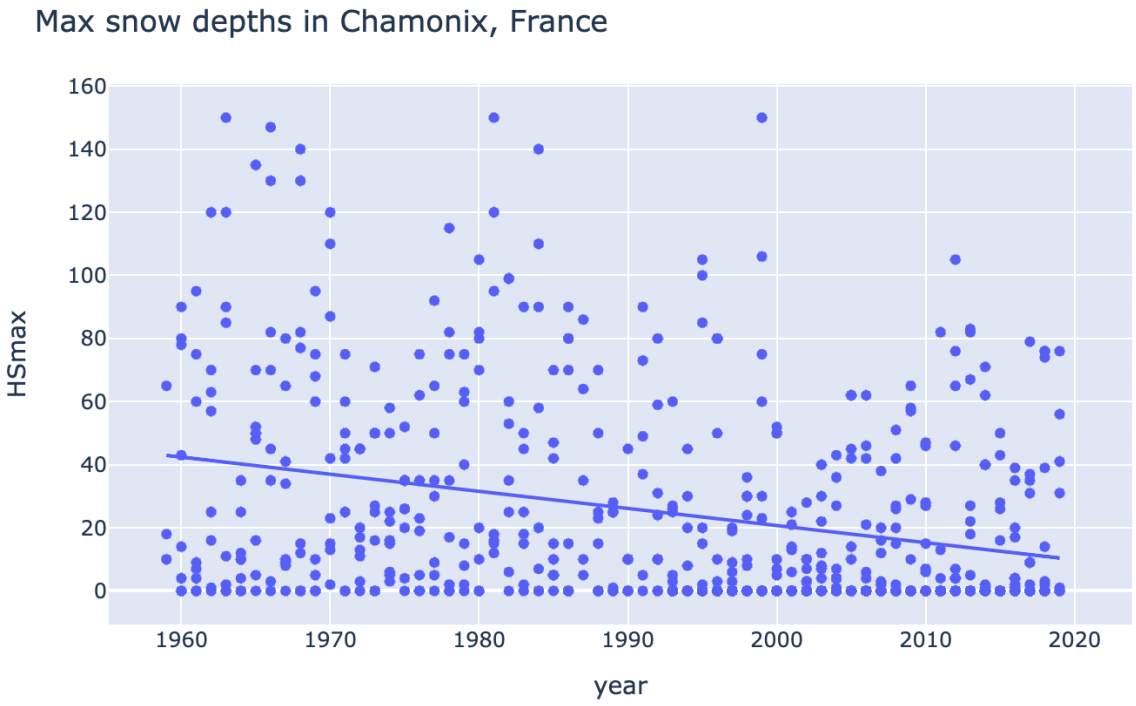
This visualization below shows a scatter plot of the max snow depths, within the months November - February for Zugspitze, Germany. This was an early analysis that practiced pruning the data into manageable chunks, while also trying to find what makes an interesting pattern. Due to the nature of snowfall mostly taking place in winter months, the included months were trimmed heavily to help identify decent patterns. This practice was dropped moving forward.



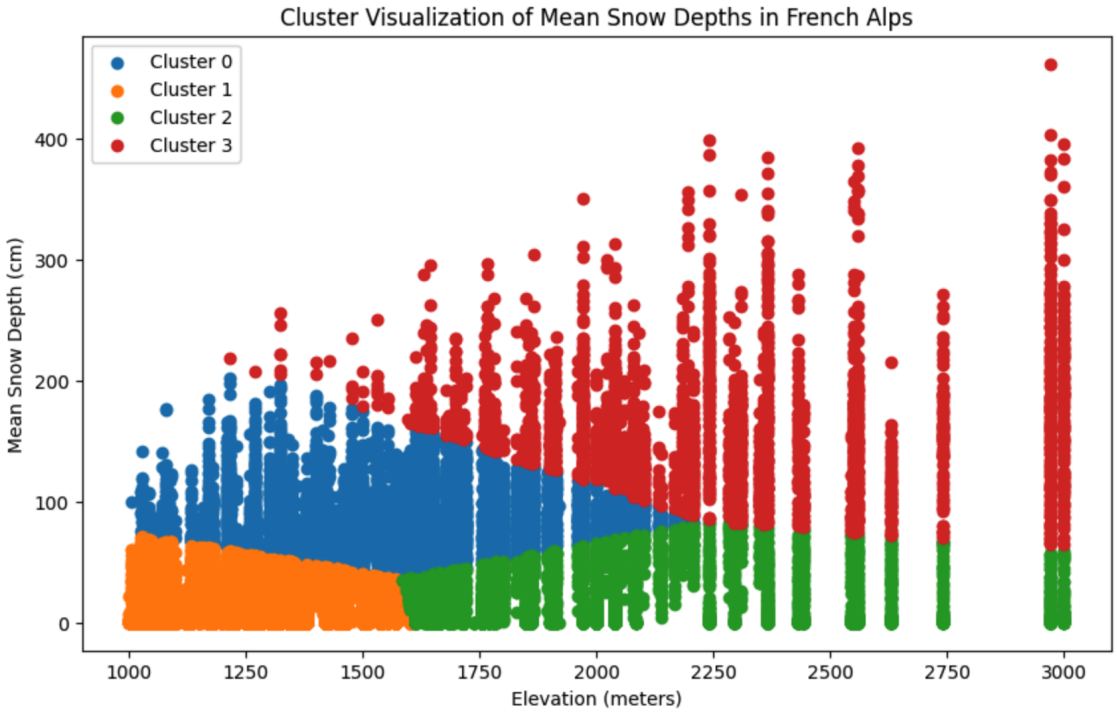
This plot below shows a regression analysis of snow depths versus elevation. This plot was used to better practice data mining tools as well as various python packages that allowed for this type of work. While the pattern is not particularly interesting, being the average snow depth per month based on elevation, it is still helpful to have a baseline understanding that an increase in elevation in the Alps is correlated with an increase in snowfall.



This next plot shows the max snow depths in the town of Chamonix, France. This graph puts together the best practices presented in earlier attempts. This graph is among several that look at ski towns in particular to examine how their snowfall amounts have fallen dramatically throughout the decades. There is a regression line which clearly shows the decrease, and even without it it is clear that snowfall has fallen. There are several high outliers between 1960-2000, and events of snowfall of similar magnitudes have not been seen in over 20 years.



Lastly, this plot shows a cluster analysis of elevation and mean snow depths. The elbow method was used to determine how many clusters to include, and then several python packages were utilized to plot the results with a different color for each cluster.



These visualizations are what I consider to be the best that were created throughout the project, as they show the progression of how they were created and used.