Alexander Gustafson (algu6965)
Data Mining Project Milestone 3

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

Problem Statement and Motivation

The motivation behind this project is to find interesting patterns and long term trends regarding European Alp snow depth data collected between 1971 to 2019 to help our understanding of climate change's effects on snow depths. Secondly, the motivation is to use this knowledge to make predictions for future years' snow depths and snowfall amounts. The data mining process of the dataset will allow us to find these hidden patterns, which cannot be found with other methods due to the size of the dataset. Previous knowledge includes programming abilities and data mining concepts that will be crucial for mining the data, being able to understand and use the results, and to create visualizations to further help put the found patterns into an easy to understand visual. I will apply this previous knowledge to this project by using my programming and mathematical skills to dig through the data and find patterns. The interesting trends I hope to find from this project include how significantly climate change has lessened the snow depths in the region, when this began to occur, if there were timeframes where snow depths dropped the most, if there are any year over year increases, and what will the snow depths look like in the future.

Literature Survey

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.

Proposed Work

There will be much to do in order to get the data ready for analysis due to how large the dataset is. For each of the six countries looked at, each one contains the data for every month over 40 years for over 100 cities. For example, the dataset for Germany contains all of the collective data over 40 years for 943 different towns and cities. To make the data more manageable, I will begin

[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/.

by finding the best methods of clustering. For one country, it would be ideal to group together regions with similar characteristics such as elevation. Instead of using each unique data point, the average value could be used for regions within certain predefined elevations such as 300-500, 500-750, etc. Similarly, since this project aims to look at the Alps themselves and not their smaller, lower elevated towns, a threshold will be used to eliminate regions that fall below a certain elevation. One city mentioned is Absdorf, Austria with an elevation of 597 feet, which seems too low for its snow depths to be significant. The dataset contains many cities will elevations well over 5,000 feet which is more useful.This dataset comes with a file "meta_all.csv" which contains the name of the city as well as it's coordinates, elevation, and snowfall and snow depths for the beginning and end of each year. This data can be merged with other datasets so that the elevation can be used to prune the data. In total, the meta_all file contains 2,979 unique cities, many of which likely do not fit in with the project.

The design of the ultimate dataset will be one dataset that contains the monthly observations for all regions above a certain elevation. The data will be inserted into a SQL table that can be used to query the data. The primary dataset will be made using pandas and will combine the monthly data from the 12 monthly datasets, and it will be merged with the meta_all dataset so that elevation will be included.

What sets this project apart from previous work done with the data is it will look at the data in a more broad sense to find patterns that cover the entire area. Previous work has looked primarily at each dataset separately and

Data Set

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.

Evaluation Methods

Evaluation methods will include box plots for visualizations of the data for each year or decade, helping to recognize outliers and variability. K Means clustering will be used to group together years with similar snow depth patterns into clusters to identify distinct trends with similar patterns and behavior. When it comes to making predictions, decision trees will be used to identify important predictors of snow depth variability and to better classify years into different categories, perhaps on environmental factors and elevation. Bayesian classification will be used to compute the information gain on environmental variables such as temperature, snow fall, and elevation. This approach will assist in understanding the significance of each variable in predicting snow depth trends accurately.

Tools

Pandas has already been used to gain a general understanding of the dataset, and has been helpful in determining how large the dataset is. A jupyter notebook has been created on my local machine in VS Code, and the datasets have been imported in. Pandas has been used to load in the csv files where general commands have been used such as .nunique(), .max(), and .head() to determine what the datasets looks like and what they contain. With millions of data points across all of the csv files, pandas will be used more to combine the data into one general dataframe. Once this has been created, pandas will still be used for some general querying, however, it will be loaded into a SQL table so that I can make more specific commands that combine many features.

The use of SQL will be important due to the variability of when data was and was not collected throughout the dataset. An example of SQL's utility in this will be to "SELECT HSmean FROM data WHERE Elevation > 750;". Commands like this could be beneficial in getting a broad understanding of the data without the worry of missing yearly values. I can perform other queries that use "WHERE year BETWEEN 1971 AND 1980;" to extract all data where the year meets the requirements. Once this data is pulled, that is when pandas will be used.

Lastly, Scikit-learn will be used to make future predictions, GitHub will be used for version control, and the data mining textbook used in class will be referenced for methods and algorithms. Plotly and Seaborn will be used for visualizations including box-plots and histograms towards the end of the project.

Milestones

The first milestone is to clean the data, which I hope to have done by the end of week 10. The data cleaning will include pruning values and merging the datasets into one. Since the data came from many different sources this will likely be a lot of work and the tools listed previously will be used to aid in this. The second milestone is to begin data mining to find interesting patterns which will be done by the end of the spring break, around March 31st. Once this is done, I can begin a data analysis on the patterns found which will be done in early April, between weeks 11 and 12. The third milestone will be to use 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 will use the patterns found to train some machine learning models to make predictions on the next few decades which will be done by the end of week 13. Week 14 will be used to do any last minute cleaning to make the project presentable and I will practice presenting the data.

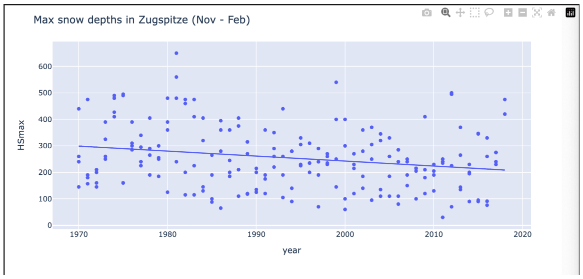
Progress Report Updates

The primary difference between the initial proposed work compared to the current state of the project is the absence of SQL tables. At this stage, Pandas has proven to be more useful. This is due to the inconsistencies within the data, which means it is easier to keep the majority of the data separated in their own independent dataframes. Different variations of dataframes have been made to be combined or compared with others, which would have been too complicated to perform with SQL.

Motivation Updates

At this stage, I have spent most of my time merging data sets and I have 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 have been pruning down the data to find regions with significantly higher elevations than most and looking at their long term trends. I have decided to split it up like this due to the intense variation of elevation from country to country. When k-means clustering is applied to elevation (in meters) within Germany, the two centroids are [577.13, 1614.17]. When k-means clustering is done in France, we end up with [2004.25, 1272.69], which shows how much higher the regions being studied are within France. Even so, areas in France that are not at what would be considered extremely high elevations, like Chamonix which sits only at 3,396 feet, the snow depths are very high when compared to regions at the same elevations elsewhere.

An example of this new approach is, within the data_monthly_DE_DWD.csv, I found that the data set 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 to 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 are the focus since snow depth trends are 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. Here is the result:



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.

Milestones Completed

Milestones completed thus far include cleaning the data which was done using pandas. I initially read all of the CSV files into dataframes in python within a VS code environment. Once these were read in I could begin to gain an understanding of their structure and what they contained. Most of the columns, roughly 16 out of the 23 are not ones that I plan to use in my project. Also, my project aims to use elevation as a key metric which was not initially present in the monthly data sets for each region. This information was kept within a meta csv that I was able to merge with the other dataframes. For one data frame of a country, it then includes the monthly data for each city. For example, it will list the city several hundred times as it gets through the data for that area, and then it will do this again for the next city/town. When considering how many cities/towns are in each csv, over 250 in some cases, this is why elevation is so important. It is a great metric to use for a cutoff point to prune out cities that are too low for their snow depths to be relevant to the project.

Milestones Todo

The main work left to do is to do more work similar to the trend finding of HSmax like what was done in Germany. This way I will be able to compare and see if the slow decrease in max snow depths is a trend for all European Alp countries. Also, I still need to complete a few more predictive models so that I can try to find interesting future patterns, which I currently predict will show a continuation of decreased snow depths.

Secondly, I would like to mostly use the rest of the time to focus on k-means clustering to find how temperature and elevation as a whole within a country impact the snow depth instead of focusing only on specific cities. I will also use the information gained from this for the predictive models I would like to have.

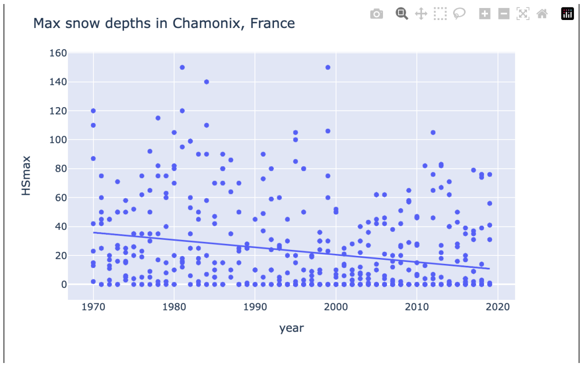
I still plan to train some machine learning models in python so that I can plot a trend line for the future to make predictions on how snow depth will likely continue to fall in the coming years.

Lastly, I would still like to use several other data mining principles that we have learned throughout the class, even if they are not extremely relevant to my project. I hope to be able to do so in order to gain practice using them and to be able to possibly find interesting results that I otherwise would not have found.

Results So Far

As seen in the work on Germany and France, I have so far been pruning the datasets in order to clean them and narrow down the data to only that which is useful for my goals. The initial millions of data points were initially quite overwhelming, and it has been pleasant to see the datasets become what I have wanted them to be. At this time, I have found trends from the pst 40-50 years for several German and French cities and they all show decreases in snowfall amounts.

I made the decision to filter the German data to only include the snowiest months, however, I figured it is just as interesting to include all months so that we can visualize the snow cover that sticks around after the winter as well as the early season snow. By including Spring and Summer data, it is possible to see trends in summer heat that melts snow and how it has gotten worse over the years. One popular ski destination, Chamonix has been plotted below.



This time series plot shows how much of a significant decrease Chamonix is seeing. Even without the trend line, it is possible to visualize how much of a drop in maximum snow depths has been. It is interesting to see how low the depths were from 2000-2010 and the increase that follows in 2010-2020.