Ryan Novak
Final Project Write-up

K-means Clustering on Avalanche Data

For my final project, I decided to analyze a dataset of Colorado's historic avalanche data. Going to college in Colorado, mountains are very prevalent. I have always enjoyed escaping to the mountains with my friends for snowboarding, hiking, and exploring. In the winter, while we frequently snowboard at resorts, we also enjoy going out for backcountry days. This is dangerous because the backcountry does not have ski patrol to help with accidents and the snow is not mitigated for avalanche danger. Just this year while riding in the backcountry, my friends and I were riding on very unstable snow. This caused mini avalanches as we tried to get down the mountain. These very small avalanches were enough to scare us into worrying about what would happen when a larger avalanche is triggered. When snowboarding in the backcountry, it is important to be aware of avalanche danger to avoid these situations. Avalanches can be extremely dangerous and are known for taking people's lives every year in Colorado. When deciding to expose yourself to the possibility of avalanches, it is important to know the chances of triggering an avalanche and what terrain is safest to ride. This is where my interest for this topic originated. By analyzing Colorado's avalanche data, I set out to cluster similar avalanche incidents together. Using historic avalanche data in this manner can tell me what avalanches over the years have similar characteristics. This can be taken to show trends in characteristics of avalanches over the years. What elevation do most avalanches occur? Are avalanches in the same elevation zone generally the same in width? What part does the slope angle play in avalanches? These questions can help people make smarter decisions when visiting the mountains to snowboard in avalanche prone areas. If large avalanches are clustered together showing similar width, starting elevation, and slope angle, it can tell someone to be cautious when deciding to ride these same conditions.

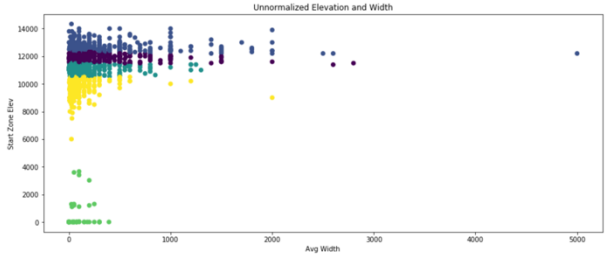
To complete this project, it is important to have a reliable data set. I sourced my data from the Colorado Avalanche Information Center's (CAIC) website. The CAIC is run by the Colorado Department of Natural Resources and aims to educate people on the dangers of avalanches and provide avalanche information. When visiting the CAIC website, they provide a database with access to recorded avalanches in Colorado. I was able to use this tool to compile a csv file of recorded avalanches in Colorado from the year 2010 to present. This report gave me 18,176 records of avalanches. The data includes observation IDs, backcountry zones, names of who reported the incident, the number of people effected, average slope angle, average width, start zone elevation, units for measurements, comments, avalanche types, dates of occurrence, and other information. For my project, I took the most interest in the average slope angle, average width, and start zone elevation since these are important factors in avalanches. I cleaned my data by discarding the rows of information I was not interested in and removed data that had missing values in the remaining rows of interest. This left me with two different data sets to examine. The first contained only the start zone elevation and average width of each avalanche. This data contained records of 2,572 avalanches after discarding unknown values. The second contained information on avalanches start zone elevation, average width, and average slope angle. This data set, after discarding unknown values, contained records of 1,560 avalanches. The start zone elevation is recorded in feet and refers to the elevation the avalanche initially started. The average width is also recorded in feet and gives an estimation of

how large the avalanche was. The average angle is recorded in degrees and refers to the angle of the face of the mountain in the zone where the avalanche occurred. Below I include a visualization of some of the data included in the uncleaned data set.

Obs ID	Date	BC Zone	Landmark	Elev	Asp	Type	Rsize	Dsize	Incident	Area Description	Avg Slope Angle	Start Zone Elev	Start Zone Elev units	Weak Layer	Weak Layer Type	Avg Width	Max Width	Width units
66783	2021/04/23	Vail & Summit County	buffalo Mountain	>TL	N	SS	R1	D1.5	NaN	Maximum Elvis	NaN	12400.0	ft	NaN	NaN	50.0	NaN	ft
66780	2021/04/23	Front Range	Herman Gulch	>TL	S	SS	R1	D1.5	NaN	Woods Mountain	NaN	12800.0	ft	NaN	NaN	100.0	NaN	ft

In the real world, one of the main companies that deals with this information is the CAIC. Using the same data I sourced from their website, the CAIC keeps detailed records of avalanches that occur in Colorado and across the US. Every year, the data collected is used to create statistical graphs of avalanche deaths in the US and by state, avalanche deaths by activity, and avalanche deaths every month. This information is used as educational material to show people the danger avalanches can pose to backcountry recreation and residents. The United States Geological Survey Agency (USGS) also works in avalanche research. The USGS has a project called the USGS Snow and Avalanche Project that aims to answer questions about avalanche activity. With avalanches causing multiple deaths across the US every year, the USGS studies the history of avalanches to try and forecast avalanche frequency. The project studies weather and avalanche history to determine what characteristics drive large avalanches. The USGS website for the Snow and Avalanche Project provides interesting information about their research. This includes the partners of the project, being the CAIC, US Forest Service National Avalanche Center, and the US Forest Service Rocky Mountain Research station. All of these government agencies are out to answer questions about why avalanches occur, what the biggest factors in avalanches are, and what weather patterns contribute to larger or more frequent avalanches. All this information is needed to increase the ability to safely recreate in the mountains during the winter and is used as educational material for people wanting to learn more about what to do in the case of an avalanche or how to look for warning signs of avalanche danger. The American Institute for Avalanche Research and Education (AIARE) is an organization that's purpose is to use information like this to educate the public on avalanche safety practices in hopes of lowering the chances of avalanches and raising the chance of survival when avalanches cannot be avoided.

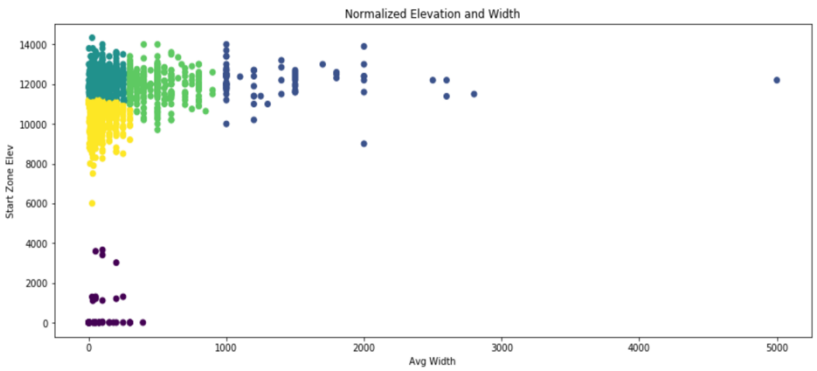
To begin answering the questions I set out with, I completed some exploratory analysis. I began by plotting the original data in terms of start zone elevation and average width of the avalanches. This gave me an idea of what the data I was working with looked like. I did further cleaning on my data set to see how the original data would be clustered off the start zone elevation and average width of the avalanches. I used k-means clustering to cluster the avalanche incidents into k=5 groups. The result of displaying the clusters and plotting start zone elevation vs average width can be seen in the figure on the right. The clusters appeared to be heavily reliant of the start zone elevation of the avalanche. Each cluster contains avalanches within a range of start zone elevations with very little regard to clustering in respect



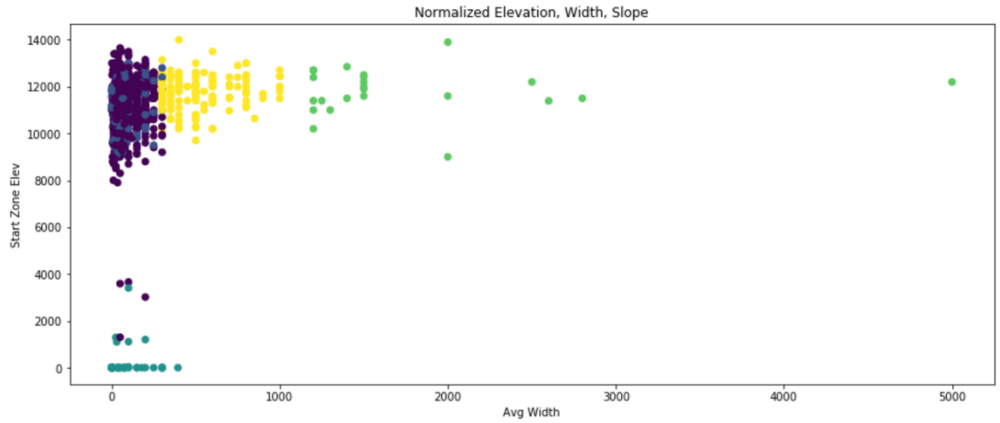
to the average width. A further analysis of the data I was working with showed a need for normalization of the data. The standard deviation of the start zone elevation was much larger than the standard deviation of the average width. This caused the clustering process to cluster with most regard to start zone elevation. To get results that rely on all the provided data, I needed to normalize my data going forward to receive more accurate results from all avalanche characteristics being considered.

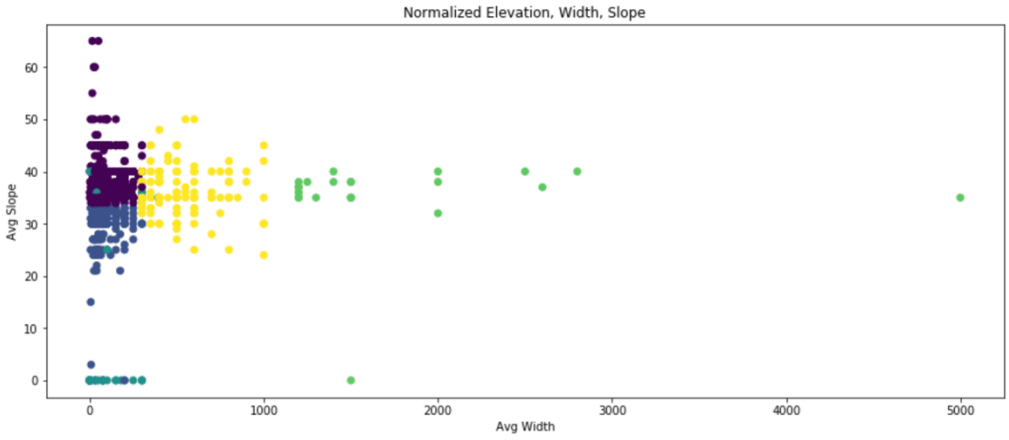
The method I used to analyze this data was k-means clustering. By clustering the avalanches, I can tell what similarity avalanches have to each other. The characteristics I cluster the avalanches by will be grouped into clusters of most similar avalanches. By examining the final clusters, I can see what characteristics make avalanches most similar. If all avalanches occurring at a certain slope angle are very similar in width, I can predict what slope an avalanche of a particular width would most likely occur at. This means the final clusters give a method of prediction for where an avalanche sharing a similar characteristic would be grouped. In my clustering function, I use a random initialization of beginning centroids. Five indexes within the avalanche data are chosen as the initial centroids. I then loop through the avalanche data and assign each avalanche to the closest cluster based off of the Euclidian distance from the centroids. After all of the data is assigned to a cluster, the centroids are updated to reflect the centroid of the current clusters. Repeating this process fifty times sorts all the avalanche data into the final clusters returned by the cluster function. This is important because I want to see what avalanches have been the most similar since 2010 and the final clusters will give me a way to visualize this. After completing the clustering function on data for start zone elevation and average width, I looked at the standard deviation of the data and realized the clustering was being done based off of start zone elevation more than average width. This led me to normalize my data before performing the clustering function. This will give a more equal consideration to all characteristics of the data rather than just one. To normalize my data, I used the method presented in class. This method normalizes by finding the mean and standard deviation of the data. It then takes the original data and subtracts the mean before dividing by the standard deviation. Now that the data is normalized, I again ran the clustering function on the data for start zone elevation and average width. This gave me a more expected clustering where avalanches with a similar start zone elevation and similar average widths were clustered together. Based off this realization, the data set with start zone elevation, average width, and slope angle was normalized before performing the clustering method.

My results gave me depictions of the final clustering of the avalanche data. Since the unnormalized data gave more consideration to specific characteristics, I only used the clusters from normalized data to examine my results. The first clusters I calculated were on the start zone elevation and average width of the avalanches depicted in the figure on the left. These clusters are predictable since I am only clustering off the two characteristics used to



plot the data. The clusters include the closest avalanches within a radius of the centroid. I would expect these clusters based on the spacing of data points representing avalanches on the plot. I was curious to see how these clusters would be affected by adding the slope angle into the clustering data. I created a new data set of slope angle, start zone elevation, and average width from the original data and cleaned out avalanches with unknown data. This data was normalized and sent to the clustering function. The final clusters are depicted in the figure to the left. Looking at the clusters, the blue cluster can be seen intertwined with the purple cluster. This is because the new clusters are including the characteristic of slope angle. This tells me that the previous clusters change because the similarity of slope angle varies across start zone elevations. In the real world, this can be because although the avalanches start in similar elevation zones, they are on different mountains in different areas. The terrain of mountains varies so mountain slopes are different between two mountains even when looking at the same elevation. To get a better visualization of the clusters, I plotted the same data but on axis for average slope vs average width. This new plot to the right shows the same clusters established in the last plot. Considering all of the plots of data together, it looks like two clusters of avalanches based on start zone elevation and average width plotted between 8,000ft to 14,000ft of start elevation are now intertwined. This can be because the max elevation of mountains varies and the slope along the faces of mountains also vary. Although avalanches happen at similar elevation, the slope is different at these elevations in different areas. Since I am using normalized data, the slope angle an avalanche occurs at has more correlation between avalanche similarity than the correlation gained by start zone elevation. This is important because in the real world, the slope angle being skied on in the backcountry should be given careful consideration for the avalanche potential it contains. This can help people





decide to avoid slopes of 25 to 45 degrees on high avalanche potential days and be confident their chances of triggering an avalanche are lower on slopes with angles outside of this range.

After cleaning my data by discarding unknown values and discarding other avalanche characteristics, the resulting data sets were smaller than I had hoped. In continuing work on this project, I want to find a more complete data set of avalanche history in Colorado. Discarding so many avalanches because of unknow characteristics made the data available to be cluster and examined very small compared to the initial available records. By finding a more complete data set, I can draw more accurate conclusions on my results from using all available records instead of only a fraction of the original records. If there is no more detailed information on Colorado avalanche history, I would want to transition from the historic avalanche data from Colorado to historic avalanche data of the US. By switching to US data, I would have more initial data leaving more data to be analyzed after cleaning. Another option would be to see if I can find more complete records of avalanches from a source other than the CAIC. I also want to include the Dsize from the initial data. The Dsize is a score for an avalanche representing the destruction potential behind the avalanche. This would give interesting insight into how destructive an avalanche is and if this depends on slope angle and elevation or if it would be correlated more to the average width. After analyzing my results, I can see that slope angle has more impact on avalanche similarity than compared to start zone elevation.