Final Project: Analyzing Mountaineering Deaths on the World's Highest Peaks

(Exploring how mountain height relates to both death counts and causes of death)

Astrid Reker and Petra Radinkovic

MA346-1 Data Science

Professor Nathan Carter

Bentley University, Fall 2025

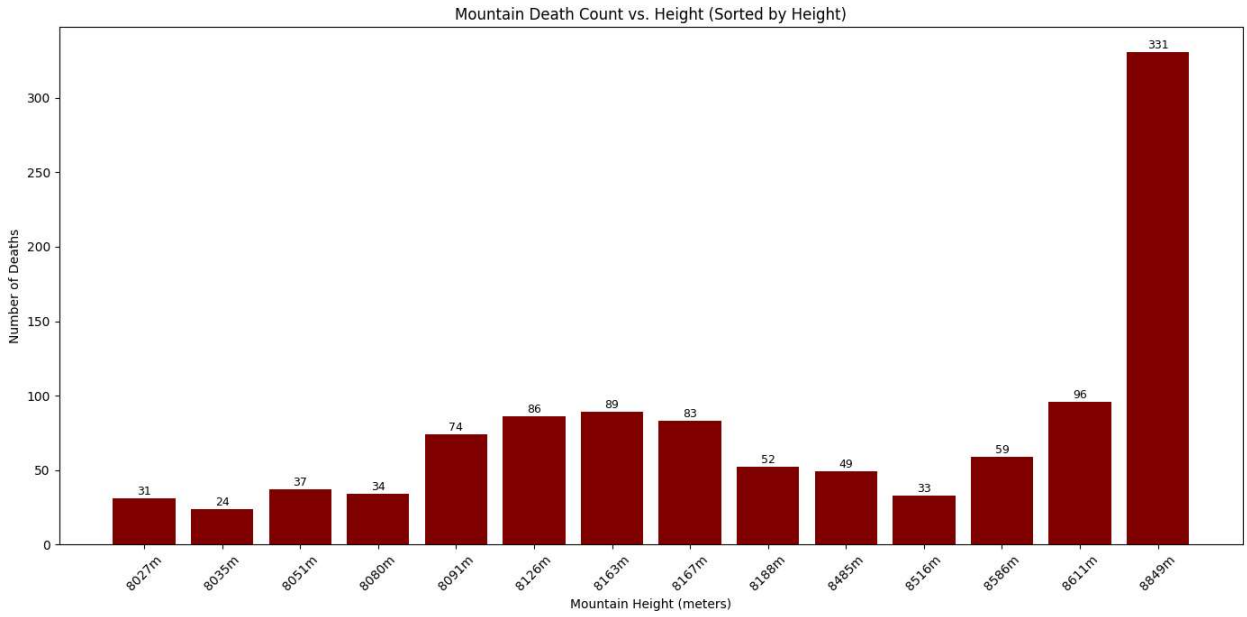
For this project, we investigated whether the heights of mountains correlate the number of climbers that have died on the mountain. We also briefly explored how the height of the mountain correlated with causes of deaths, but that was not our primary focus of this analysis. Using two data sets we determined at a 95% confidence level that the height of a mountain does in fact correlate with the number of climbers that die on the mountain. Specifically, for every additional one hundred meters on a mountain, approximately eighteen additional deaths occur.

We used two data sets: The first one we got from Kaggle, called “Deaths on eight thousanders” and starting from 1895, it details records of climbers that have died on mountains over 8000 meters tall. It also details the mountain where the accident occurred, their names, and nationalities. The second data set was from Gigasheet, called “List of Highest Mountains on Earth”. The second dataset contained information about each major peak, such as height, prominence, range, and ranking. Before merging, we cleaned and standardized the mountain names because the two sources used slightly different naming.

After fixing these consistencies, we then had to merge these two datasets into one bigger dataset that included all of the information from “Deaths on eight thousanders” and some of the more useful columns of the mountain height data set (such as height, range, parent mountain, and prominence). After fuzzy matching the mountain names on both data sets- and subsequently cleaning up the mess that left- we performed an inner merge on the two matching name columns. With this we were left with our new, much more useful data set which we lovingly named “df_mtn_death.” This enriched dataset allowed us to analyze our two main

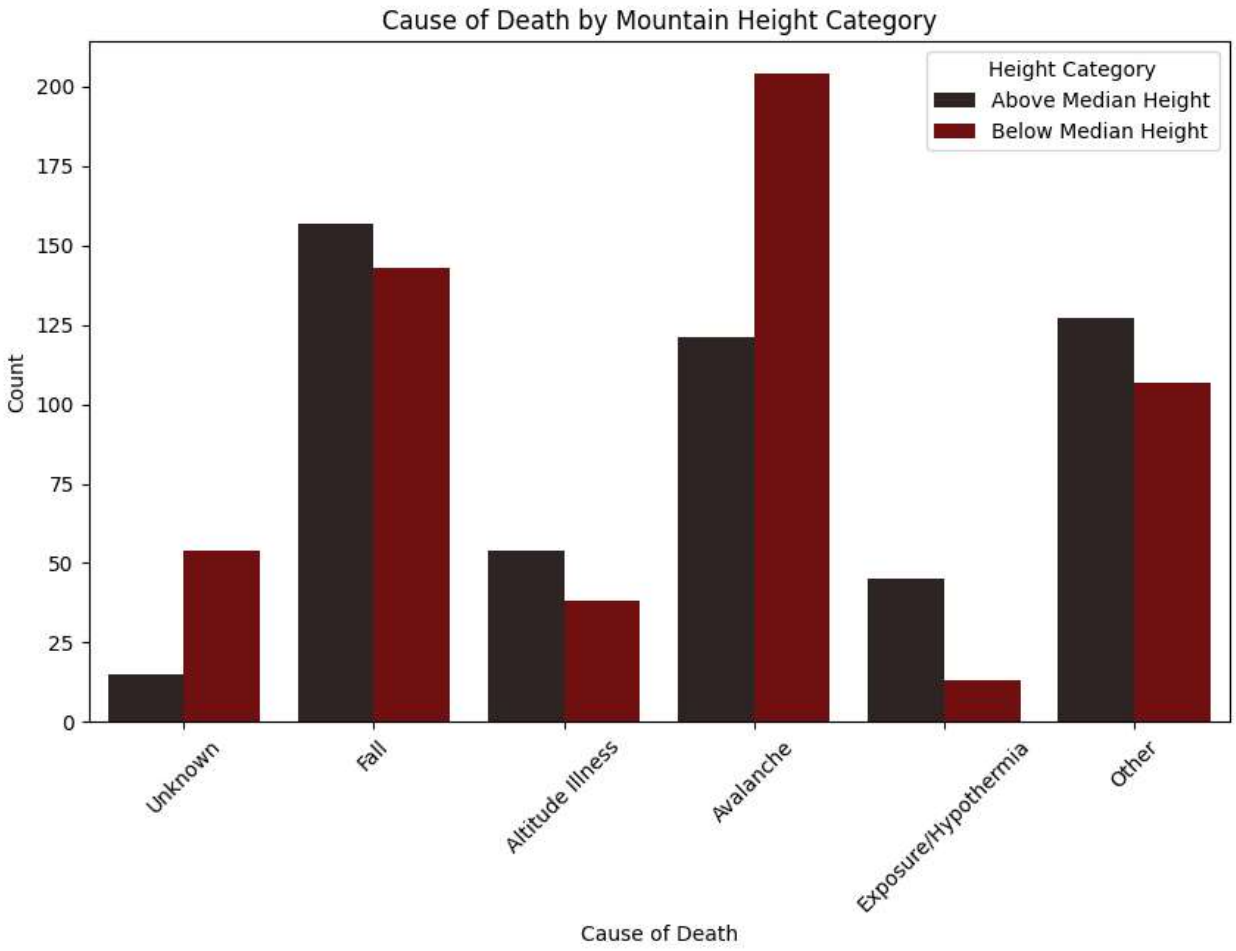
research questions: Does the height of mountains correlate with the amount of people that have died on them? Do those heights impact their causes of death?

With this dataset we then made some data visualizations. We found an ascending bar chart appropriate to show deaths by mountain height as it would give a clear representation of how many people are dying on these mountains as sorted by height:



In reference to our secondary question regarding causes of deaths in relations to heights of mountains, we sorted our mountains into two aptly named categories: “Above Median” and “Below Median.” We then simplified the incredibly detailed causes of deaths into broader categories such as “Avalanche,” “Fall,” “Altitude Illness,” “Exposure/Hypothermia,” “Other,” and “Unknown.” The reason we did these things was that we could put this information into a bar

chart with this information:



The final form of data analysis we performed on this data was a linear model that told us about the relationship between mountain height and number of climbers that have died on them:

OLS Regression Results

Dep. Variable:	Death Count	R-squared:	0.405
Model:	OLS	Adj. R-squared:	0.355
Method:	Least Squares	F-statistic:	8.163
Date:	Wed, 03 Dec 2025	Prob (F-statistic):	0.0144
Time:	02:40:56	Log-Likelihood:	-76.537
No. Observations:	14	AIC:	157.1
Df Residuals:	12	BIC:	158.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1437.2377	530.240	-2.711	0.019	-2592.532	-281.943
Height	0.1828	0.064	2.857	0.014	0.043	0.322

Omnibus:	5.433	Durbin-Watson:	0.972
Prob(Omnibus):	0.066	Jarque-Bera (JB):	2.394
Skew:	0.833	Prob(JB):	0.302
Kurtosis:	4.152	Cond. No.	2.66e+05

From these results we can see that these results are statistically significant with a P-Value of 0.014. The positive coefficient of 0.1828 tells us that for each additional meter on the mountain an additional 0.1828 death occurs on the mountain. This indicates that the height of a mountain is in fact positively correlated with climbers dying on the mountain. The OLS method works here because it assesses whether any of the predictors are significant as well as if each of the predictors are significant, showing whether the model is significant at all.

This analysis has several limitations that should be considered when interpreting the results. The data likely contains reporting bias, as highly monitored mountains such as Everest have more complete death records than remote or less frequently climbed peaks. Our analysis focused primarily on mountain height and did not account for other crucial factors, such as weather conditions and climbing routes. The data spans more than a century, during which climbing technology and safety standards have changed substantially, but our analysis did not

adjust for this. Lastly, because the data set is observational, we can identify associations but cannot establish causation. If we were to continue with this project, we would try to determine popularity of these mountains as well as how dangerous each climb is considered to be.