Exploring the Heights

Visualizing Trends in Himalayan Mountaineering Expeditions

Course: *INFO 511 - Foundations in Data Science (Spring 2025)*

Presented By: *Praveen Kumar Pappala, Sai Navya Reddy Busireddy, Gowtham Theeda*

Project Overview

This project explores the data of **Himalayan Mountaineering Expeditions**, documenting expeditions in Nepal from **2020-2024**. For our project we will be focusing on trends in climbing success, risk factors, peak accessibility, and expedition strategies. Through **data visualization**, we aim to uncover patterns in mountaineering achievements, the impact of environmental challenges, and the factors influencing expedition outcomes using Python.

Dataset – What, Where and Why?

The dataset we are using, **The Himalayan Mountaineering Expeditions Dataset**, from tidytuesday is a comprehensive archive that documents mountaineering expeditions in the **Nepal Himalaya**. It originates from **The Himalayan Database**, which was founded to continue the pioneering work of journalist **Elizabeth Hawley**, who meticulously recorded Himalayan climbing history.

This database is a vital resource for the climbing community, offering detailed records of expeditions. This study will not only allow us to apply our data science skills but also deepen our appreciation for the Himalayas as a symbol of human endurance and spiritual connection.

The dataset is publicly available. It is a well-documented, open-source archive of mountaineering expeditions. It consists of two main datasets:

- **Expeditions Data (exped_tidy.csv)** – Expedition details, peak summits, climbing routes, success rates, termination reasons, and expedition characteristics from **2020-2024**.
- **Peaks Data (peaks_tidy.csv)** – Peak characteristics, including **altitude, accessibility, restrictions, and first recorded ascents**.

Data Processing & Cleaning

🔗 Merging Datasets

- Combined `exped_tidy.csv` (expedition data) and `peaks_tidy.csv` (peak metadata) using the common key: `PEAKID`
- Resulting dataset includes **both expedition and peak-level attributes**: oxygen use, member deaths, summit dates, peak height, range, etc.

❌ Initial Missing Value Inspection

Expedition Dataset:

- 100% missing in secondary routes (`ROUTE3`, `ROUTE4`, `ASCENT3`, `ASCENT4`) → **dropped**
- 97%–99% missing in `ACHIEVMENT`, `ACCIDENTS`, `ASCENT1`, `ROUTE2` → **dropped**
- Up to 10% missing in summit date-related columns → **retained with caution**

Peaks Dataset:

- ~96% missing in `REFERMEMO`, ~94% in `TREKYEAR`, ~84% in `PSMTNOTE`, etc. → **non-essential → dropped**
- Remaining missing values (like `PYEAR`, `PEXPID`, `PSUMMITERS`) filled with `"Unknown"` or `0`

Exploratory Data Analysis (EDA)

📌 Objective:

To uncover **early trends and correlations** in expedition outcomes, team behavior, and mountain characteristics using summary statistics and targeted visualizations.

📈 Summary Statistics & Distribution Checks

- Reviewed **distribution of summit success (**`SUCCESS1`**)** across different:
  - Seasons (`SEASON_FACTOR`)
  - Oxygen use (`O2USED`)
  - Peaks (`PKNAME`, `HEIGHTM`)
  - Team size (`TOTMEMBERS`)
- Evaluated **death data**:
  - Member (`MDEATHS`) and hired personnel (`HDEATHS`)
  - Compared against `NOHIRED`, `TERMREASON_FACTOR`

🧭 Outcome of EDA

- Identified key predictors of success: `O2USED`, `SEASON_FACTOR`, `TOTMEMBERS`
- Confirmed `SMTDATE` could be used for time series
- Selected top peaks for **geospatial mapping**
- Built a clean foundation for **advanced visualizations**

🔬 Methodology - Descriptive Summarization

- Aggregated statistics per peak
- Time-based trend analysis
- Route-level success and seasonal patterns
- Interactive map via folium

Predictor Variables (Features)	Outcome Variables
SEASON, O2USED, TOTMEMBERS	SUCCESS1 (summit success)
HEIGHTM, CAMPS, ROUTE1	MDEATHS / HDEATHS
OPEN, PSTATUS_FACTOR	Summit trends

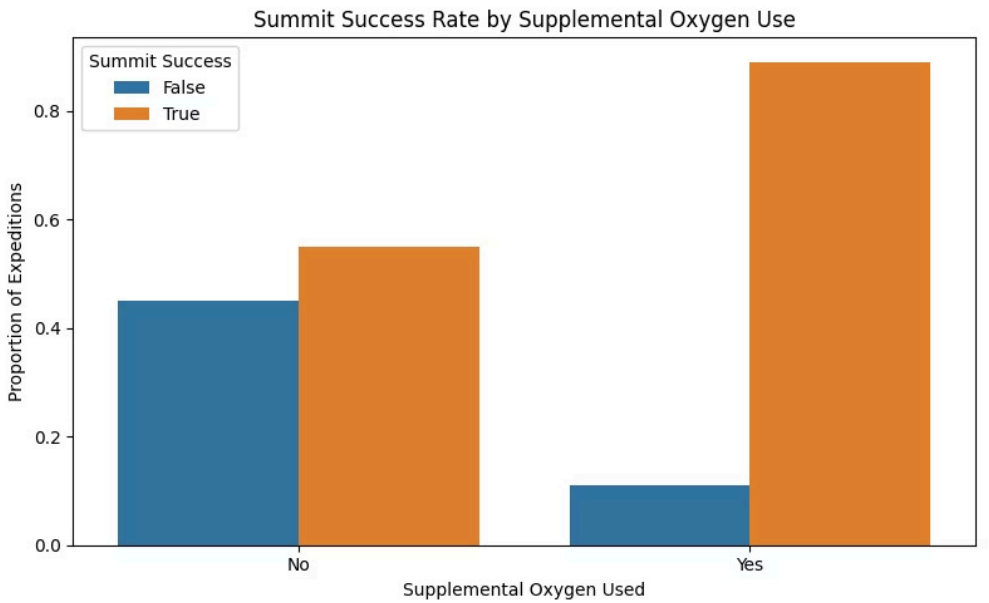
🧗 Climbing Success and Risk Analysis

Supplemental Oxygen vs Summit Success

✅ Research Question

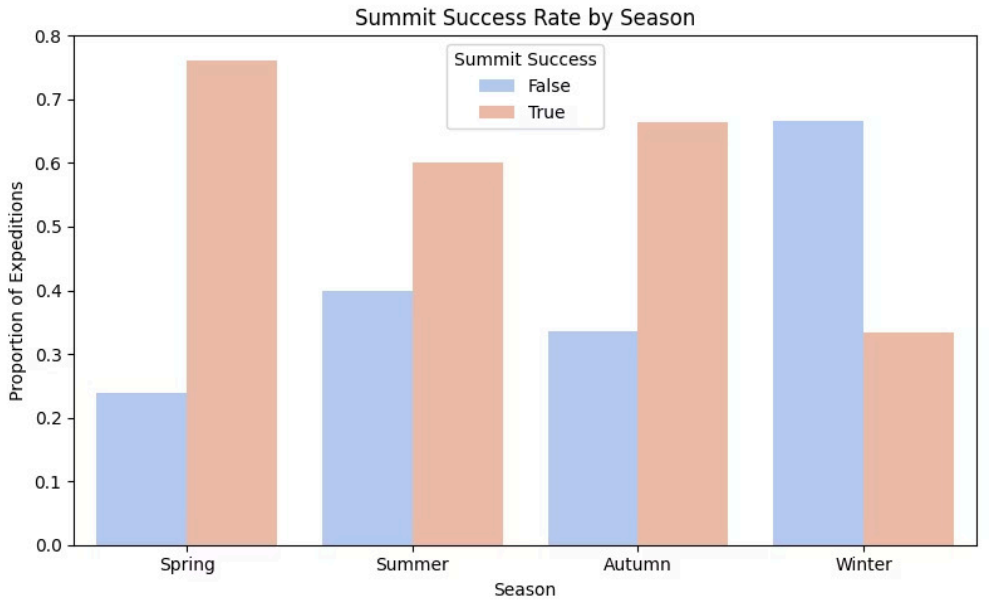
How does the use of supplemental oxygen affect summit success rates in Himalayan expeditions?

Result: Expeditions using supplemental oxygen had a ~90% success rate vs ~55% for those that didn't.
Conclusion: Oxygen use is a **critical enabler** for summit success.

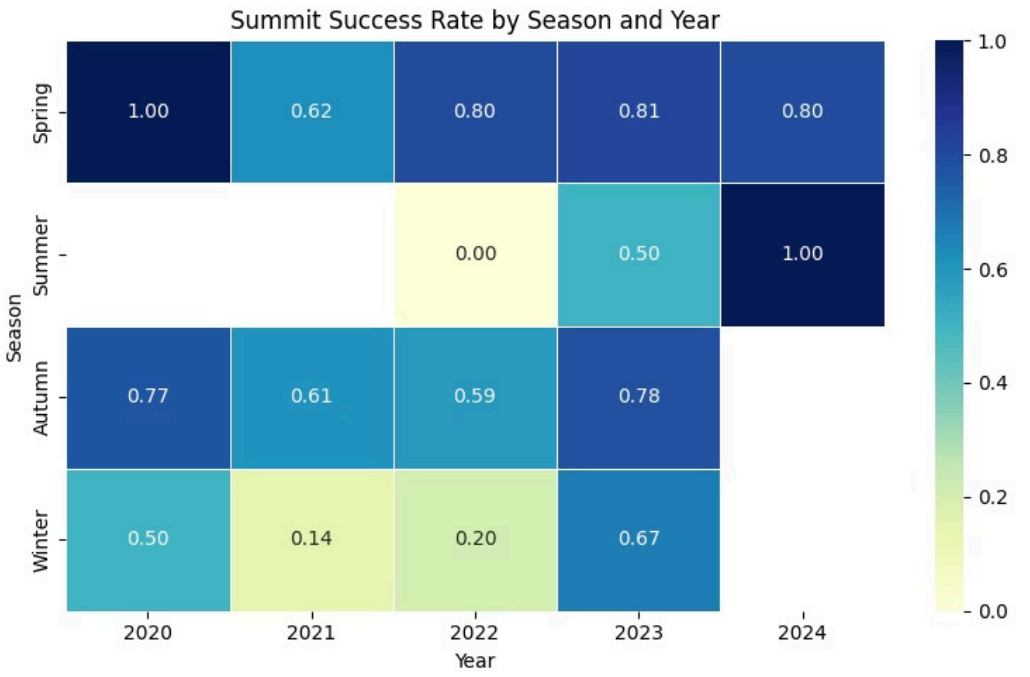


Grouped Bar Chart

Success Rate by Season & Year



Grouped Bar Chart

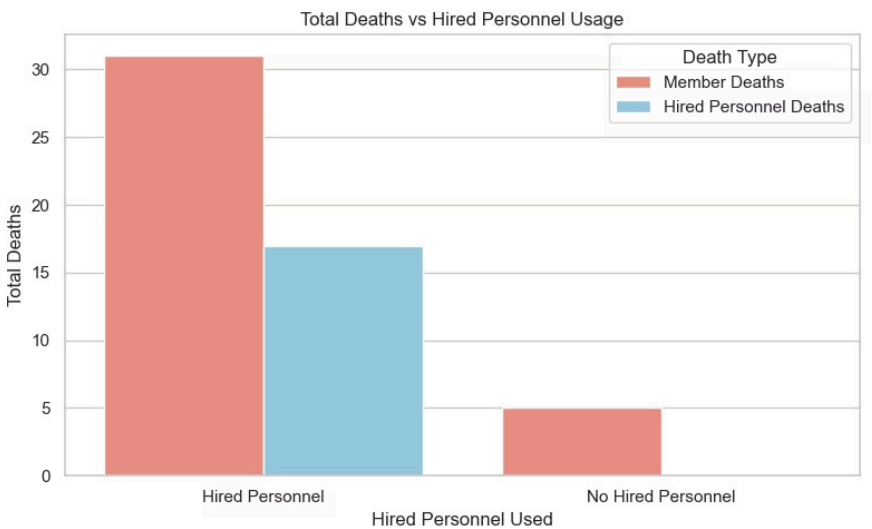


Heatmap

Result: Spring had the highest summit success (~76%), followed by Autumn. Winter had very low success (~33%).

Conclusion: Climbing season is strongly associated with success. Spring is the most favorable.

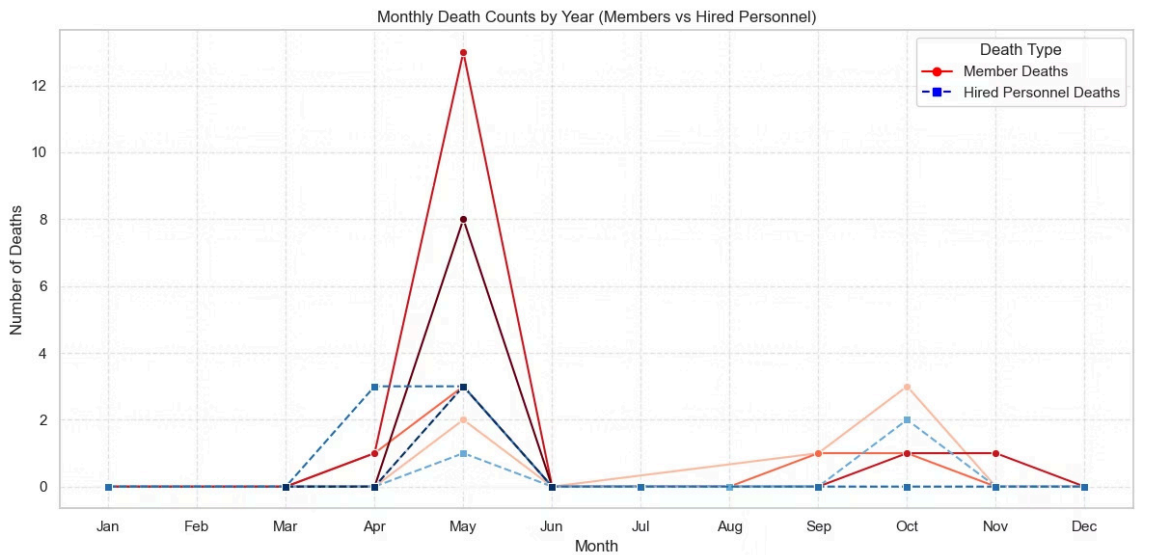
Death Analysis - Members & Hired Personnel



Grouped Bar Chart

Result: Expeditions **with hired personnel** showed higher average death rates among both members and staff.

Conclusion: Larger or more complex expeditions may face higher risk despite support.

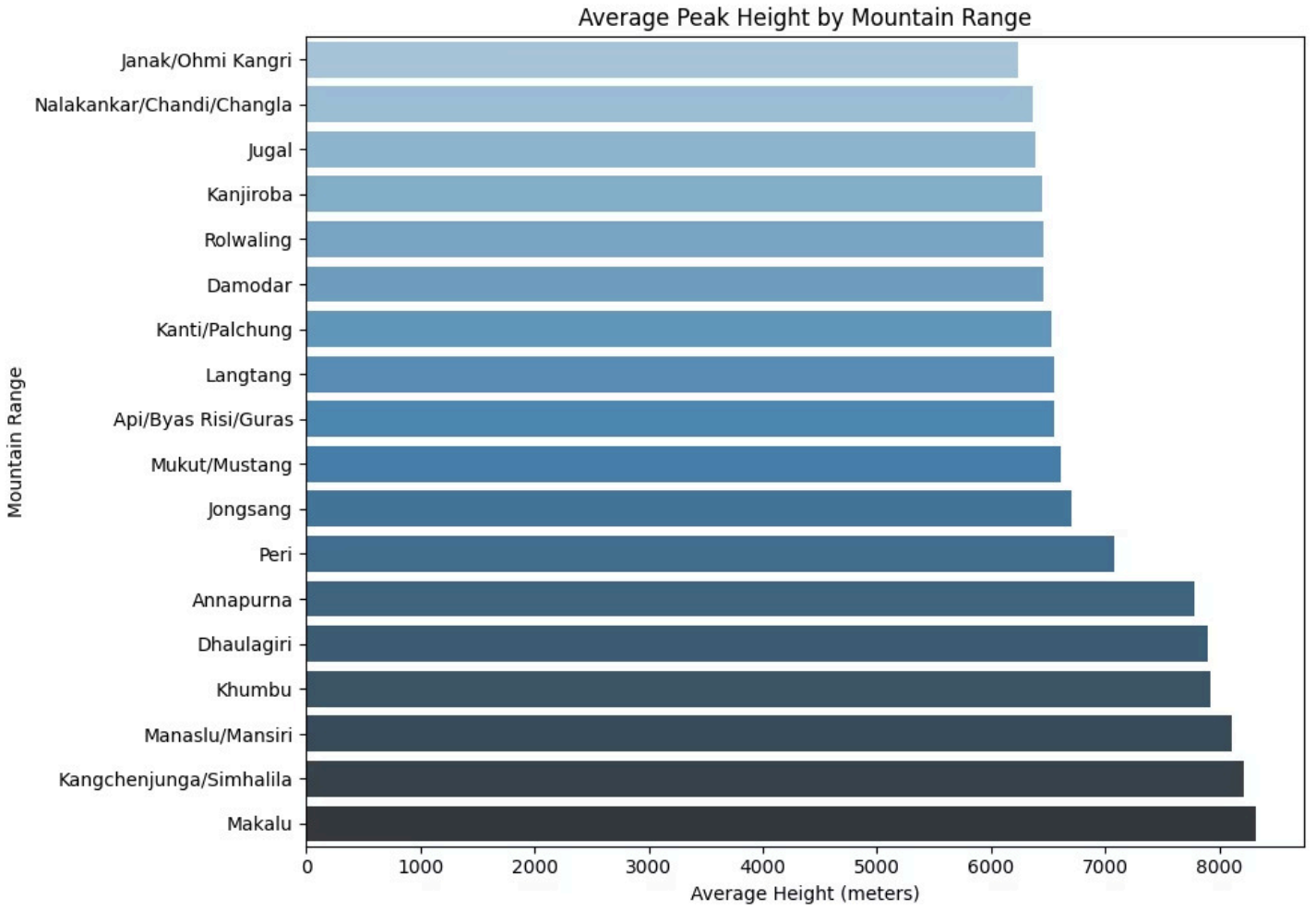


Time Series - Monthly Death Count Trend

Result: Deaths cluster in high-activity months, especially May and October. Hired personnel often see higher death variability.

Conclusion: Success and risk **co-occur during peak seasons**, emphasizing planning importance.

🏔️ Peak Characteristics and Accessibility



Horizontal bar chart

Average Peak Height by Mountain Range

✅ Research Question

Which mountain ranges in the Nepal Himalaya have the highest average peak heights?

Result: Makalu and Kangchenjunga ranges feature the highest peaks (>8200m).
Conclusion: Certain regions concentrate **tall, elite peaks** and receive more experienced teams.

Peak Height by Accessibility (Open vs Closed)

✅ Research Question

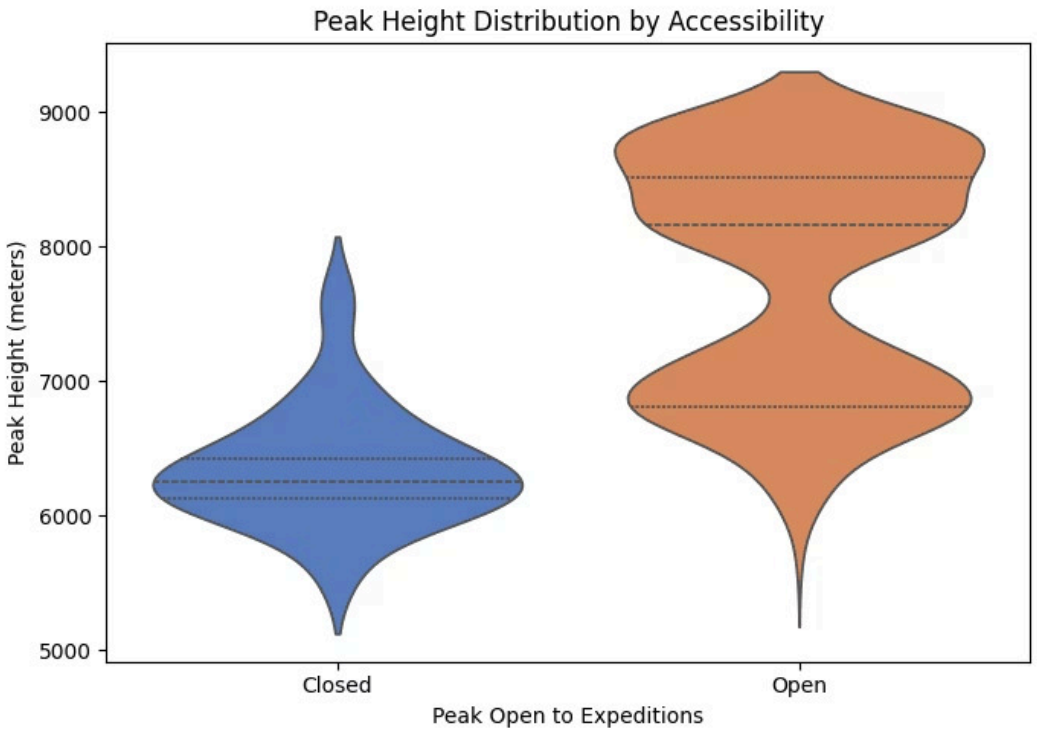
Is there a significant difference in peak height between open and closed Himalayan peaks?

🧪 Result:

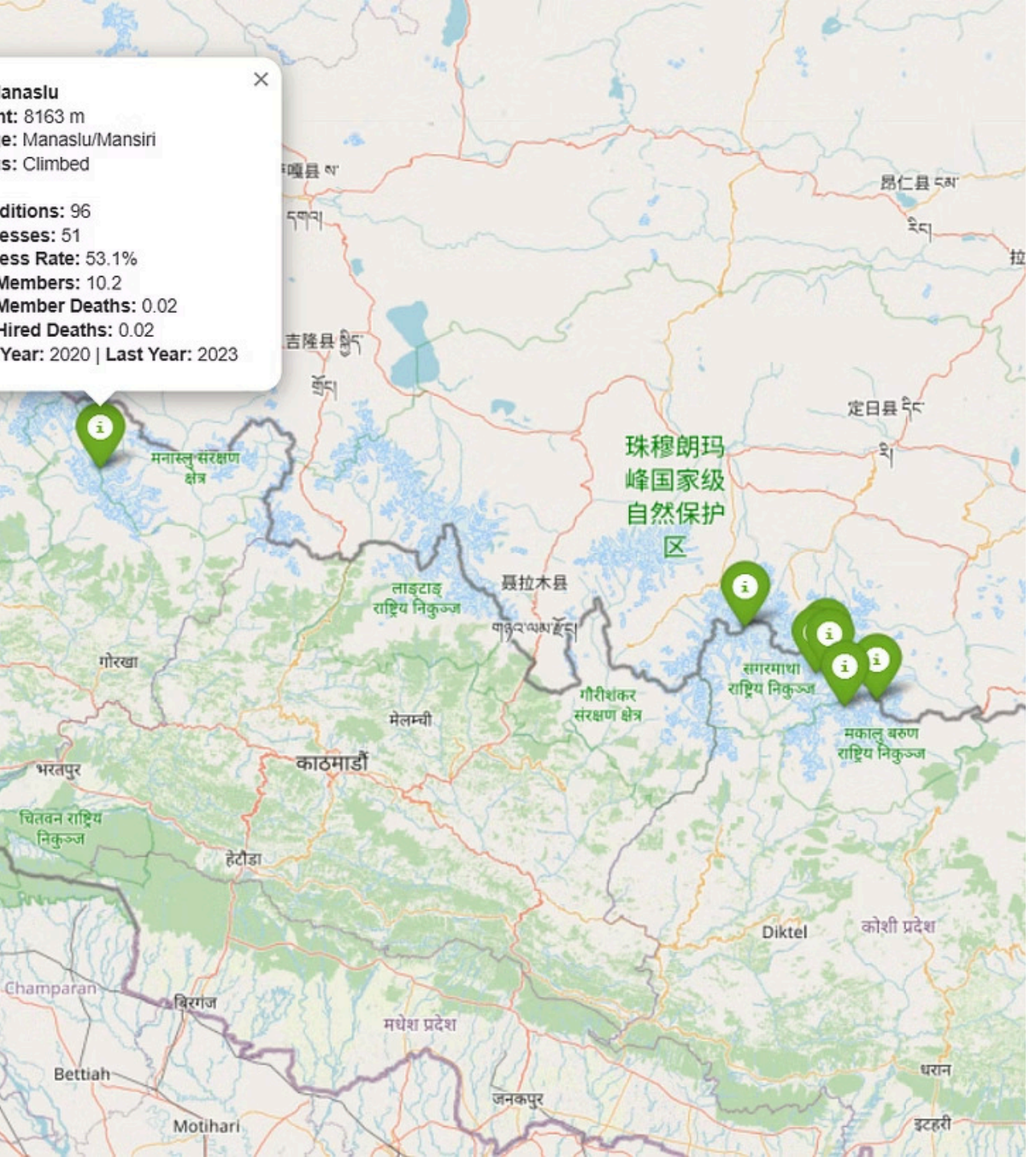
- **Open peaks** show a **broader range**, extending from 6000m up to Everest (~8800m).
- **Closed peaks** are **clustered** at lower elevations, with few exceeding 7000m.
- The **median height** of open peaks is significantly higher.

💡 Conclusion:

Open peaks are not necessarily easier — in fact, they are often **taller and more diverse**. This may be due to better documentation, popularity, or regulation. Closed peaks may be restricted due to **access issues, sacred status, or unestablished routes**, rather than height alone.



Violin plot



Geospatial Map of Top Peaks

Plot Type: Folium interactive map

Result: Everest, Annapurna, and Manaslu show high success and expedition activity. Each popup reveals expedition count, success rate, deaths, and team size.

Conclusion: Spatial analysis shows **risk/success hotspots** — and how peaks differ in accessibility and outcomes.

🗺️ Expedition Strategies & Route Success

Most Common Termination Reasons

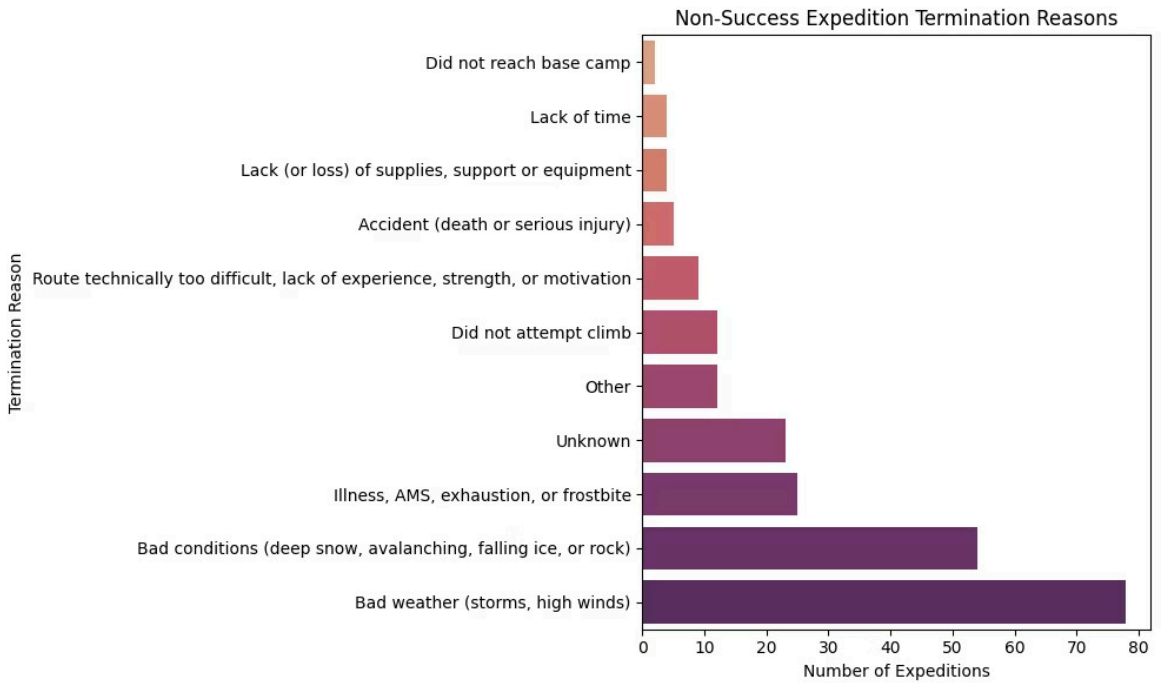
✅ Research Question

What are the most common reasons that Himalayan expeditions are terminated?

Plot Type - Horizontal Bar Chart

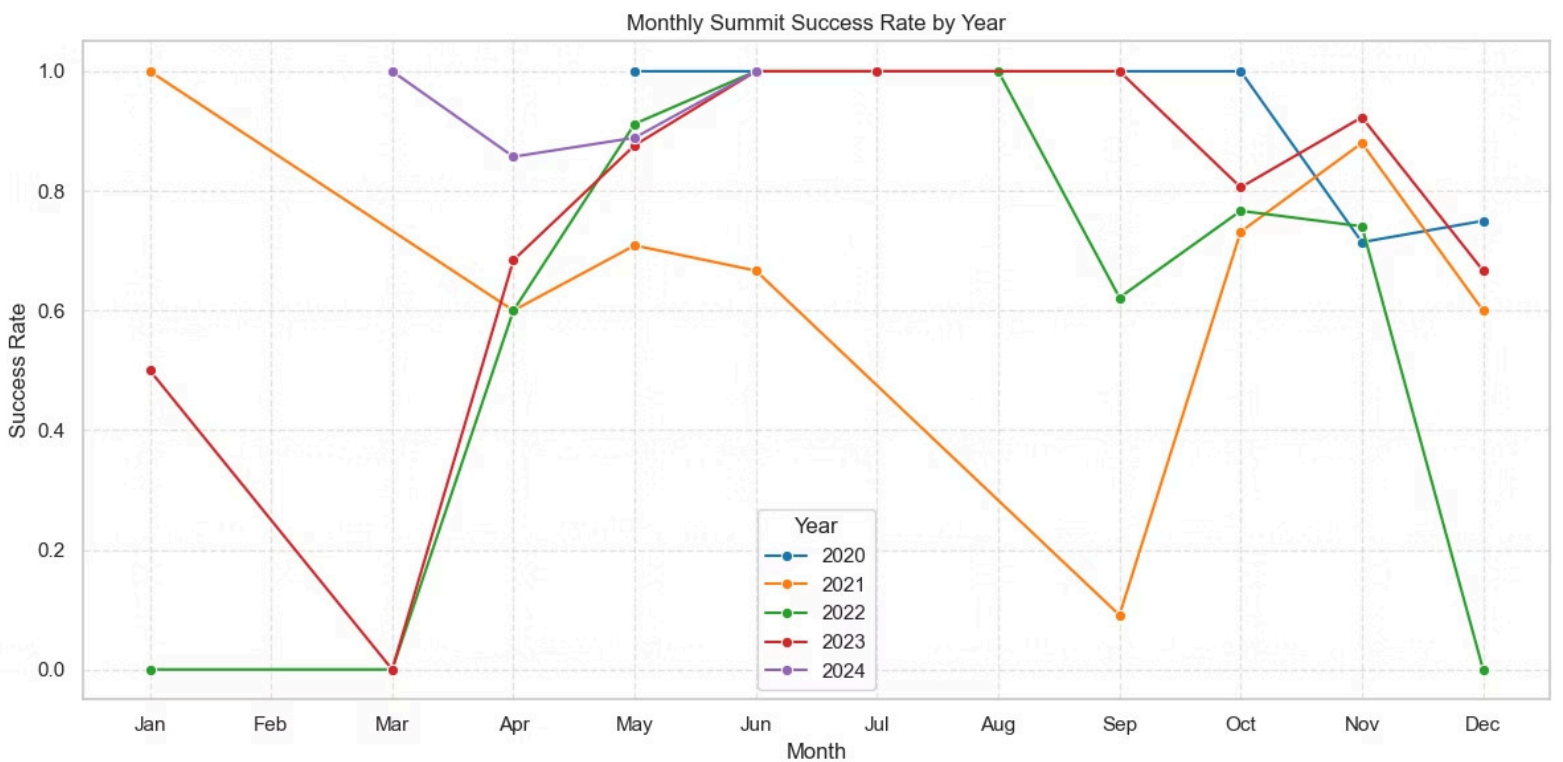
📊 Result:

- **Bad weather (storms, winds)** and **poor conditions** (snow, avalanche risk) are the most common.
- Other reasons include: **illness/frostbite**, **lack of supplies**, or **technical difficulty**.
- Very few expeditions terminate due to **accidents**, **time**, or **logistical issues**.



💡 **Conclusion:** Expedition failure is rarely due to dramatic disasters. Instead, **weather and physical conditions** remain the leading reasons for not reaching the summit. Many climbers turn back early due to **safety decisions**, not just inability.

Monthly Success Rate by Year



Time Series - Multi line plot

📊 Result:

May and October are consistently successful months; summer climbs show year-to-year volatility.

💡 Conclusion:

Month-level timing is just as important as the broader season — especially for safety and planning.

Route Success Rates by Mountain Range

✅ Research Question

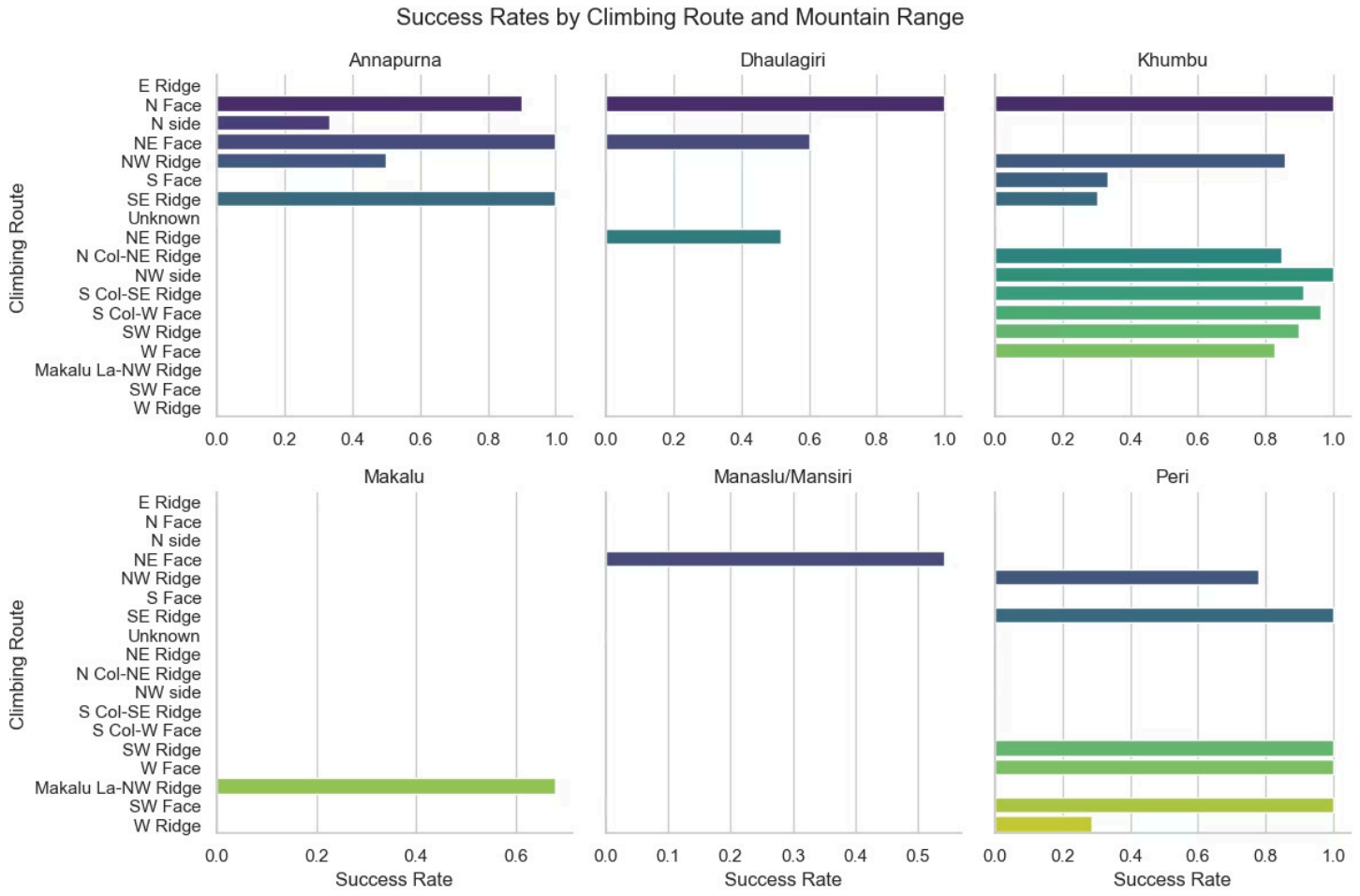
Which climbing routes have the highest success rates across different Himalayan mountain ranges?

📊Result:

Success rates vary widely by range and route. Some routes within the same range have **notably lower success**.

💡 Conclusion:

Route selection critically affects outcomes, even within the same peak or region.



Faceted Bar Chart

✅ Conclusion

> Successful Himalayan climbs are not defined by strength alone — but by **timing, strategy, equipment (like oxygen), and respect for risk**. Data reveals these are consistent drivers of outcome across peaks and seasons.

References:

TidyTuesday Project. (2025, January 2). *Exploring the heights: The Himalayan climbing database* [Data visualization project]. TidyTuesday / R for Data Science. Retrieved April 2025, from https://github.com/rfordatascience/tidytuesday

The Himalayan Database. (n.d.). *A comprehensive historical record of expeditions to the Nepal Himalaya*. Retrieved April 2025, from https://www.himalayandatabase.com

OpenStreetMap contributors. (n.d.). *Nominatim geocoding service* [Web application]. OpenStreetMap Foundation. Retrieved April 2025, from https://nominatim.openstreetmap.org

Wikipedia contributors. (2024). *List of highest mountains on Earth*. In Wikipedia, The Free Encyclopedia. Retrieved April 2025, from https://en.wikipedia.org/wiki/List_of_highest_mountains_on_Earth

Peakbagger.com. (n.d.). *Mountain data and GPS coordinates*. Retrieved April 2025, from https://www.peakbagger.com

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering*, 9(3), 90–95.

Waskom, M. (2021). *Seaborn: Statistical data visualization*. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021