Exploring the Heights

Visualizing Trends in Himalayan Mountaineering Expeditions

📌 Research Purpose

The purpose of this project is to examine trends and patterns in Himalayan mountaineering expeditions from 2020–2024. Using cleaned and merged data from the Himalayan Database, we explored how environmental conditions, expedition strategies, and mountain characteristics relate to expedition outcomes. This project is rooted in the broader objective of understanding high-altitude expedition dynamics—an area crucial to public safety, tourism planning, and mountaineering logistics. The analysis was exploratory and descriptive, aligned with the data science principle of using available datasets to discover meaningful insights for policy makers, researchers, and climbers.

🗃️ Dataset – What, Where & 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**.

Since the dataset is freely available online, **no IRB approval or special clearance** is required for data access.

🗃️ Data Cleaning and Preprocessing

The datasets were merged on a common identifier, PEAKID, resulting in a rich combined dataset of 4,500+ expedition records. We performed the following data cleaning steps:

- **Dropped Columns with >90% Missingness**: Columns such as ROUTE3, ROUTE4, ASCENT3, ACHIEVEMENT, and ACCIDENTS were dropped due to high null counts and redundancy.
- **Categorical Normalization**: Fields like SEASON, PSTATUS, and TERMREASON were cleaned and converted to factor variables.

- **Summit Dates and Time Series Prep**: Date fields were standardized to support monthly and yearly analysis.
- **Missing Data Handling**:
    - Minor gaps in numeric fields were filled with 0 or retained with caution.
    - Categorical gaps (e.g., PYEAR, PEXPID) were replaced with "Unknown" for analysis continuity.

This preprocessing allowed us to conduct robust Exploratory Data Analysis (EDA) and create features aligned with our research questions.

🗄️ Modeling and Analysis

Our approach was **descriptive** rather than inferential or predictive. No machine learning models were used due to the nature of the questions and the course's focus. However, we did apply structured aggregation, filtering, and grouping operations using Python (pandas, seaborn, plotly, folium).

Key variables explored:

- **Predictors**: Season (SEASON), oxygen use (O2USED), total team size (TOTMEMBERS), range (RANGE), route type (ROUTE1), and peak accessibility (PSTATUS).
- **Outcomes**: Summit success (SUCCESS1), deaths (MDEATHS, HDEATHS), and termination reason (TERMREASON).

Geospatial mapping was conducted using the **folium** package to visualize expedition density and outcomes by peak location.

☑️ Research Questions

📊 Climbing Success and Risk Analysis – *by Praveen Kumar Pappala*

🕸️ Visualization 1: Supplemental Oxygen vs Summit Success

Question: *Does using supplemental oxygen make a noticeable difference in whether a climber reaches the summit?*

Code Summary: We began by grouping all expeditions by whether or not oxygen was used (`O2USED`). For each group, we calculated the proportion of climbers who succeeded (`SUCCESS1 = 1`) versus those who didn't (`SUCCESS1 = 0`). The clever combination of `groupby()` and `value_counts(normalize=True)` gave us relative percentages, not just counts — allowing for a fair comparison. Next, we **reshaped the data** using `.melt()` to make it suitable for plotting in Seaborn. Finally, we created a grouped bar chart where the x-axis is oxygen use (Yes/No), and the bars show the proportions of successful vs unsuccessful climbs:

Result: The visual result clearly differentiated expeditions with and without supplemental oxygen. While we don't claim statistical causality, the chart visually shows a higher proportion of summit success among those who used oxygen.

Visualization 2: Success Rate by Season and Year

Question: *Is there a seasonal pattern in when climbers are more likely to succeed?*

Code Summary: Using `SEASON_FACTOR` (which includes Spring, Summer, Autumn, Winter), we grouped expeditions and calculated the **proportion of successful climbs per season**, similar to the oxygen plot. We visualized this with a grouped bar chart, showing success and failure rates per season side-by-side.

To explore deeper temporal trends, we created a **pivot table** showing success rate across **both season and year** and then visualized this using a **heatmap**.

Result: The visuals revealed that Spring stood out across most years with relatively higher summit rates, while Winter consistently showed lower proportions. It set the stage for further questions about weather patterns and climber strategy.

Visualization 3: Member and Hired Personnel Deaths

Question: *Do expeditions with hired personnel experience more fatalities? And how do deaths vary across the calendar year?*

Code Summary: We computed **average deaths** for members (`MDEATHS`) and hired staff (`HDEATHS`) by grouping on `NOHIRED` (indicator of hired personnel). Then, we visualized this comparison in a simple grouped bar chart.

For the **monthly trends**, we created a line chart with two lines — one for member deaths, and one for hired staff — grouped by `MONTH`.

Result: The deaths tended to peak during the most active climbing months (May and October). The trend lines visually suggested that support personnel might face higher risks, potentially due to heavier logistical duties.

Peak Characteristics and Accessibility *– by Sai Navya Reddy Busireddy*

Visualization 4: Average Peak Height by Mountain Range

Question: *Which Mountain ranges in the Nepal Himalaya host the tallest peaks on average?*

Code Summary: We used `groupby()` to calculate the **mean elevation** (`HEIGHTM`) for each unique mountain range (`HIMAL_FACTOR`). After converting it into a DataFrame, we plotted a horizontal bar chart:

Result: Makalu and Kangchenjunga emerged visually as having the tallest average peaks. The color gradient helped guide the eye, emphasizing elevation disparities across ranges.

🧩 Visualization 5: Peak Height by Accessibility (Open vs Closed)

Question: *Do closed peaks differ in height compared to open ones?*

Code Summary: We used a **violin plot** to compare the distribution of `HEIGHTM` across `PSTATUS_FACTOR` (open/closed). Violin plots display the distribution shape, median, and spread, all at once — making it ideal for comparing categories visually.

Result: Open peaks showed a broader and taller distribution than closed ones. But as per feedback, we described this as a **visual trend** and **did not claim significance**, since no statistical test was performed.

🧩 Visualization 6: Geospatial Map of Top Peaks

Question: *Where are the most frequently climbed and successful peaks located?*

Code Summary: Using **folium**, we added **interactive markers** at the latitude and longitude of each peak. Each popup showed key data like: Expedition count, Summit success rate, Member and staff deaths etc

Result: High activity clusters emerged around Everest, Annapurna, and Manaslu. The map provided an intuitive spatial overview of climbing activity and outcomes.

🎯 Expedition Strategies and Route Patterns *– by Gowtham Theeda*

🧩 Visualization 7: Most Common Termination Reasons

Question: *Why do expeditions typically stop short of summiting?*

Code Summary: We simply used `value_counts()` to tally up how often each termination reason appeared. Then, we plotted it using a horizontal bar chart with descending frequency.

Result: Weather and bad conditions dominated as the leading termination causes. The chart helped visually rank the challenges that climbers often face.

🧩 Visualization 8: Monthly Summit Success Rate by Year

Question: *Are some months consistently more successful than others, year after year?*

Code Summary: We grouped by `YEAR` and `MONTH` to calculate average summit success. A multi-line plot was created where each line represented a different year.

Result: The months of May and October appeared as consistent peaks in summit success across years. The plot offered a temporal narrative of seasonal timing.

Visualization 9: Route Success Rates by Mountain Range

Question: *Do certain routes within a mountain range perform better than others?*

Code Summary: We grouped by both `HIMAL_FACTOR` and `ROUTE1` to compute success rates. This was then visualized using **faceted bar charts** — one small chart per mountain range.

Result: The comparison suggested wide variability in route success even within the same range. It underscored the importance of route planning.

Future Research Directions

- Perform formal **inferential statistics** to test observed visual differences.
- Study **climate data** alongside expedition dates to explore weather impacts.
- Expand the dataset beyond 2020–2024 for longitudinal trends.
- Use logistic regression or decision trees to predict success likelihood based on conditions.

Conclusion

This project uses descriptive data science to analyze mountaineering expeditions in the Nepal Himalaya. Through visual exploration of oxygen use, team composition, route success, and seasonal conditions, we identified patterns that can support risk assessment and expedition planning. However, no causal claims are made, and all findings are visual summaries, not inferential analyses. This reinforces the power of data storytelling while recognizing the limits of descriptive methods.

Our analysis shows that **successful climbs depend not just on strength**, but on **planning, equipment, timing, and risk awareness**. These findings support better decisions for future climbers, tour operators, and public safety advisors. With more advanced modeling, this work can evolve into tools for expedition planning and risk forecasting.

References

- TidyTuesday Project. (2025, January 2). *Exploring the Heights: The Himalayan Climbing Database* [Data visualization project]. 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]. Retrieved April 2025, from https://nominatim.openstreetmap.org
- Wikipedia contributors. (2024). *List of highest mountains on Earth*. In Wikipedia. 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

🤖 Generative AI Tool Use Acknowledgement

This report was authored by the student team. All visualizations, data cleaning, and model logic were implemented and reviewed manually by the student team.

- **Gamma AI** was used exclusively for enhancing the visual layout of the final presentation slides. No analysis, content generation, or interpretation was performed by Gamma AI.
- AI assistance was not used for inferential conclusions or result interpretation, in alignment with course guidelines.

👥 Researcher Biosketches

Praveen Kumar Pappala
Role: Research Section – Climbing Successes and Risk Analysis, final report authoring and project coordination

Sai Navya Reddy Busireddy
Role: Research Section – Peak Characteristics and Accessibility, Data Cleaning & preprocessing and presentation development

Gowtham Theeda
Role: Research Section – Expedition Strategies and Route Success, Exploratory data analysis, code integration

🔁 Peer Review Recommendations Response Page

Feedback:
- Improve ADA accessibility of charts.
- Avoid 'significant' without statistical testing.
- Clarify visual trends are descriptive.

Response:
- ✅ Accepted: Enhanced visual contrast and labeling for accessibility.
- ✅ Accepted: Reworded all instances of 'significant' to 'visual trend'.
- ✅ Accepted: Clarified all interpretations are descriptive, not inferential.