An Exploration of Mountain Climbing Over The Last Century

MIDS W200 Fall '24 | Thurs 2:00 PM
Presented on: December 12, 2024
DateSet: Tidy Tuesday Himalayan Climbing Expeditions
Team Members: Alec Heyde, Josh Pearce, Ender Ricart, Vai Vasudeva
Presentation: **Datasci 200 Project 2**
Github Link: Click here

Objective & Questions:

We are looking at over 100 years of data related to mountain climbing expeditions in the Himalayas. Our primary objective for this project is to identify factors associated with successful expeditions. A secondary objective is to broadly understand how climbing changed over the last 100+ years. The data will be cleaned up, investigated and grouped by common variables, [expedition ID], and further analysis will be done, as noted below.

Our project focuses on providing an analysis to the following questions:

1) What are the factors that lead to successful expeditions and what correlation does each have to success?
    a) Success is being defined as follows:
        i) Reached a peak yes or no (expeditions dataset)
        ii) No one died (expeditions dataset)
2) What major changes have occurred in the sport of climbing in the last 100+ years?
    a) How has the demographic makeup changed over time?
        i) Country of origin
        ii) Sex
        iii) Age
    b) How has the use of oxygen impacted successful expeditions? Has this changed over time?
    c) How has the proportion of successful ascents changed over time?
    d) Are the heights of peaks that were ascended for the first time recently (last 50 years) higher than those first ascended historically (51-100+ years)?

Assumptions:

There were a number of assumptions made to enable our exploration and analysis. They included the following:

- The data inputted in the three data-sets are different representations of similar or same events related to Himalayan mountain climbs and therefore cross-referenceable.
- That the recorded data is largely accurate, with correct informational input such as sex, age, and deaths.
- The data column referred to as "success" in the data sets refers to the successful ascent of the peak, and is independent of team member deaths.

- Trends observed in frequency of climbs over time is due to actual drops in volume of climbs and not a failure of data entry.
- Any missing data is Missing Completely At Random (MCAR)

Data Sources (raw data):

Members data: This dataset contains detailed information about individual climbers participating in mountaineering expeditions. It includes demographic data (e.g., age, sex, citizenship), expedition details (e.g., role, season, peak name), outcomes (e.g., success, injury, death cause), and specific metrics like oxygen use and high points reached. It contains information from the years 1905 to spring 2019.

Peaks data: Contains peak-specific information related to the Himalayan expeditions from 1905 to Spring 2019.

Expeditions data: Contains expedition-specific information related to the Himalayan expeditions from 1905 to Spring 2019. It includes expedition-level data such as dates (basecamp and termination), peak climbed, outcomes (e.g., deaths, injuries, high points reached), team details (e.g., members, hired staff), and logistical information like trekking agency and oxygen use.

Our selection consists of three distinct data-sets but each has overlapping data fields such as expedition_id and peak_id or peak_name, enabling us to merge and perform analysis across the three data sets. More information about the datasets used can be found in Appendix E.

Data Cleaning & Exploration:

Before starting our analysis, we performed a series of explorations on the data and its variables to ensure quality and integrity and identify areas worth a more in-depth analysis to unravel relevant correlations. This included data cleaning and gut-checks which can be found in each of the dataset EDA, an essential step before merging the data. We performed all data cleaning prior to creating the merged dataset, so that only clean data was included.

Members Dataset:

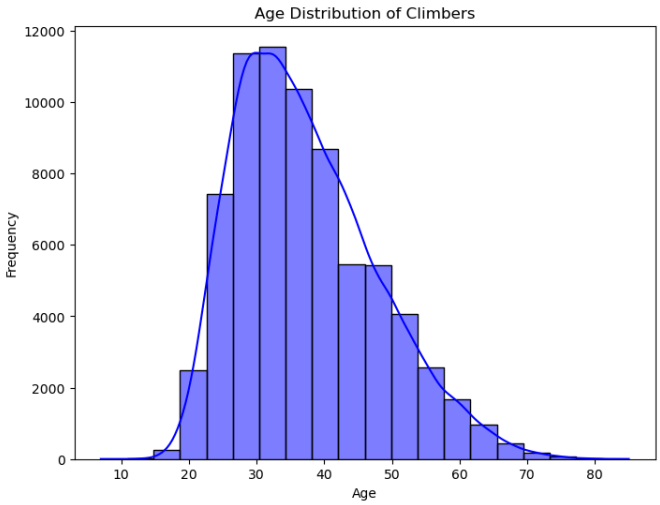
Within the Members dataset we were particularly interested in the following fields, identifying and accounting for any missing or unexpected values:

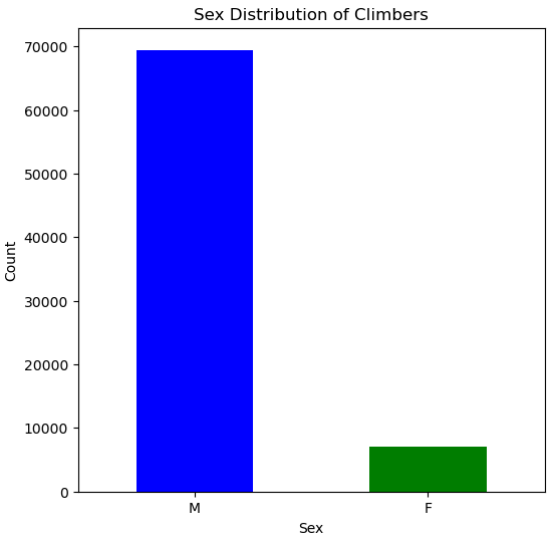
- ‘age’, ‘sex’, ‘citizenship’, ‘success’, ‘highpoint_metres’, ‘died’, ‘peak_name’.

We checked for column completeness using the info(), isnull(), and sum() functions. In particular we noted columns with a high percentage of null or “NaN” values and excluded them from initial analyses.

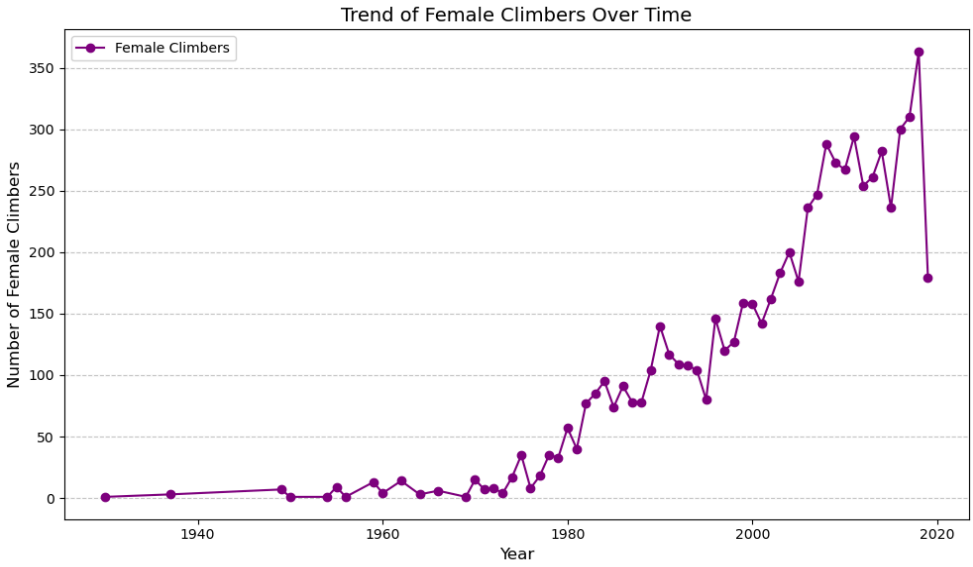
Members Preliminary Analysis

Many climbing expeditions involved multiple team members, however the “Members” data-set gives us an opportunity to focus on some of the individual climbers and their age, role, nationality, and more. Looking into the age of individuals in the Members data set we find a clear bell curve emerging that peaks between the late 20s and early 30s.

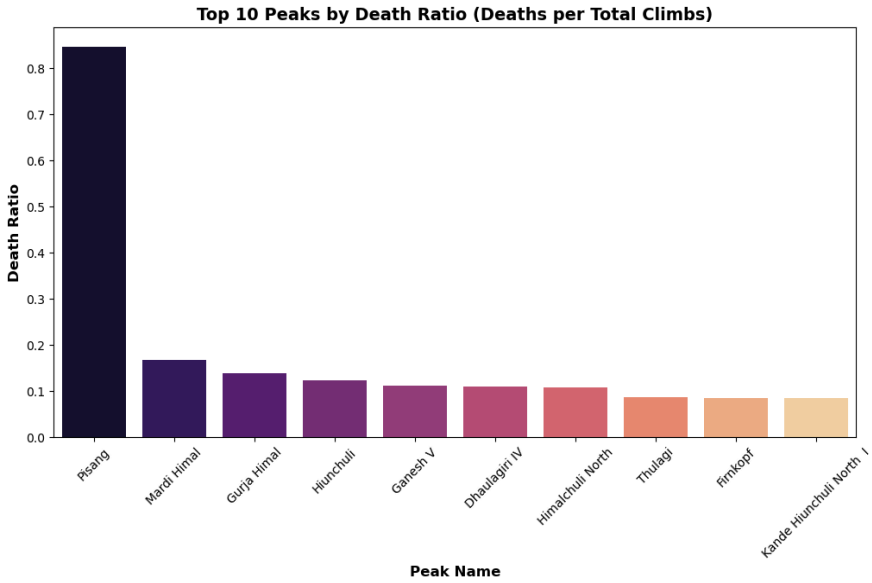




We also discover that Himalayan climbing expeditions remain a highly sexed industry, with nearly 70,000 climbers sexed Male and around 7,000 sexed Female. Visualized, it is a striking asymmetry. If we look at trends over time in the quantity of female climbers we see a sharp climb in numbers around the 1990s.

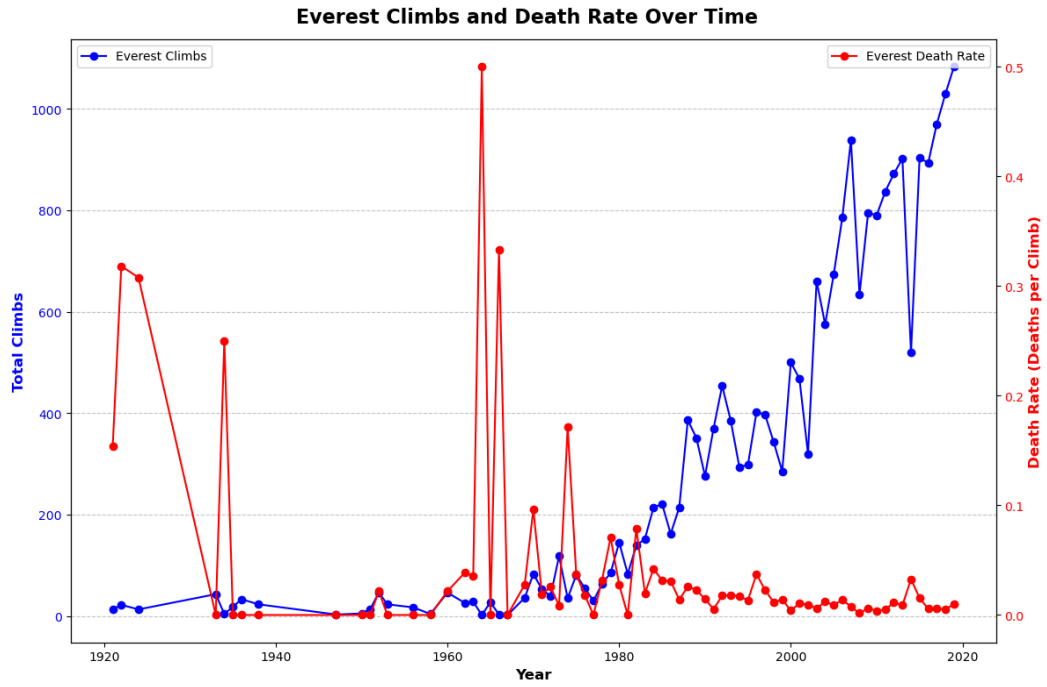


Though we see a rise in female representation in climbing missions, there are no solo climbs made by females within our members data-set, though preliminary exploration online notes that French citizen Chantal Mauduit made solo climbs of two Himalayan summits - Lhotse and Manaslu - in the 1990s. Chantal is not recorded in our dataset.



Before moving on, we wanted to show data on the deadliest mountain in the Himalayas. Based on the above graph, it is without a doubt Pisang that is the most deadly of the mountains in the Himalayas - nearly all climbers who have attempted it died.

If we slice instead by the most frequently climbed peak in the Himalayas we find that Mt. Everest is the winner with 21,813 out of the 76,519 member entries. Here we see that the frequency of Mt. Everest climbs increase, the death rate doesn't scale with it. In fact, it has a downward trend save for several spikes in deaths around the time that we see the rise in climb frequency of Everest start to tick up.



The vast majority of attempts are group expeditions (count: 76,397), while solo climbs are extremely rare (count: 121). Solo climbers appear to have a higher proportion of success compared to group climbers. The limited number of solo expeditions makes it challenging to draw definitive conclusions about whether solo climbing is genuinely more effective (see Appendix B).

It's possible that solo climbers are exceptionally skilled or selective in their attempts, which could contribute to their higher success rate, rather than solo climbing itself being inherently advantageous.

Peaks Dataset:

These are the variables that were checked and cleaned in the Peaks dataset:

- ‘peak_id’, ‘peak_name’, ‘peak_alternative_name’, ‘height_metres’, ‘climbing_status’, ‘first_ascent_year’,’first_ascent_country’, ‘first_ascent_expedition_id’.

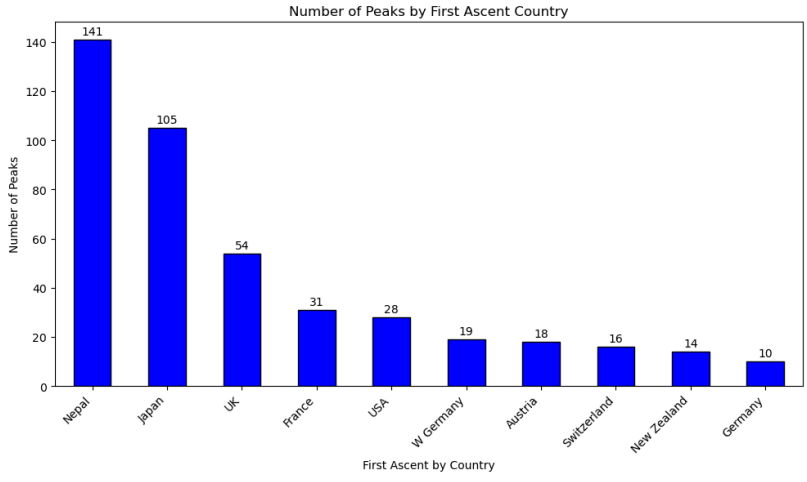
We ensured that the number of unique peak_ids and peak_names matched the total number of rows in the file, confirming there was one record per ID. We also verified that all first_ascent_year values fell within the range of 1905–2019. Notably, one record (peak_id = SPH2) had a first_ascent_year of ‘201’, likely an input error, which was flagged as a data quality issue.

As part of the cleaning, we identified the number of NaN values in each column and examined the relationship between missing values in first_ascent_year, first_ascent_country, and first_ascent_expedition_id. We expected the number of NaNs in these fields to be equal and to correspond to the number of records with a climbing_status of "Climbed," indicating complete data. However, we found eight records with unexpected missing information across various columns.

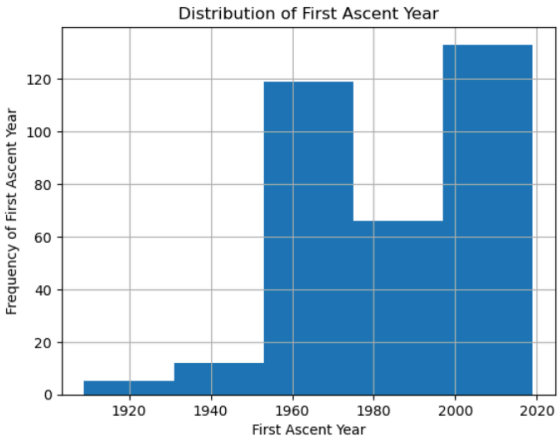
Despite this, we decided to retain these records in the dataset because they contained relevant information about peaks and climbing successes.

Peaks Preliminary Exploration

In an analysis of mountain ascent by countries, we find in the raw data that there are some entries that list multiple countries such as - “Japan, Nepal” and “New Zealand, USA, UK” for example.



If we break apart these multi-country climbing expeditions and individually count each country towards its total we find that Nepal has by far the highest count of first ascent climbs as compared with other countries. The majority of Nepal's ascent numbers are within multi-country climbing expeditions and not independent ones. Nepal has only 8 first-ascents independently, and 133 conjoined with another country. We assumed that this is reflective of Nepal's proximity to the Central Himalayas and the role of mountaineering in their economy.



When we look at the distribution over time of first ascents, we see a dramatic rise in expeditions around the 1960s, going from under 20 first ascents to just under 120 in frequency. There is a brief dip in first ascent counts between the late 1970s up to the late 1990s before frequency spikes up again around 2000 at 135 count (See Appendix A).

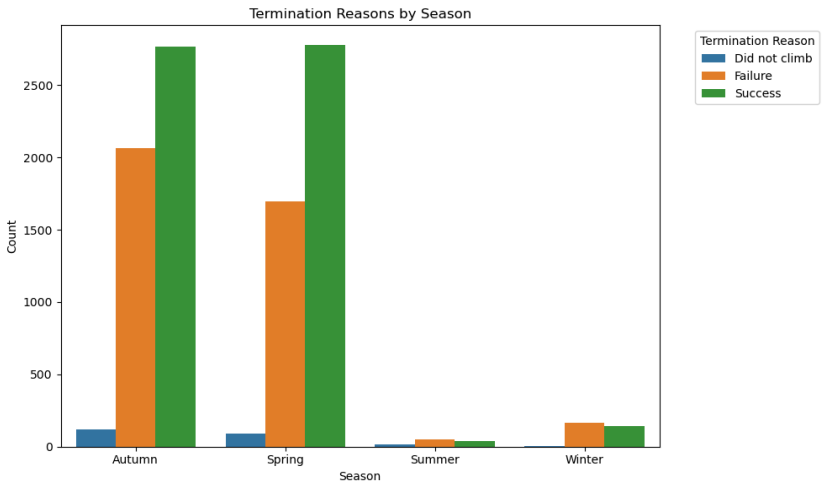
Expeditions Dataset:

These are the variables that were checked for this particular dataset:

- “season”, “basecamp_date”, “highpoint_date”, “termination_date”, “peak_name”, “members”, “member_deaths”, “hired_staff”, “hired_staff_deaths”, “termination_reason”

We ensured all `expedition_id` values were unique and consistent across rows with no duplicates, verified and cleaned `season` values by removing rows labeled as "Unknown," consolidated and clarified ambiguous categories in the `termination_reason` column, cross-referenced `peak_id` and `peak_name` with the Peaks dataset to confirm consistency, standardized all dates to datetime format, addressed missing values (NaNs) in essential columns, and corrected logical inconsistencies by ensuring deaths did not exceed the total number of members or staff, updating the respective counts when necessary.

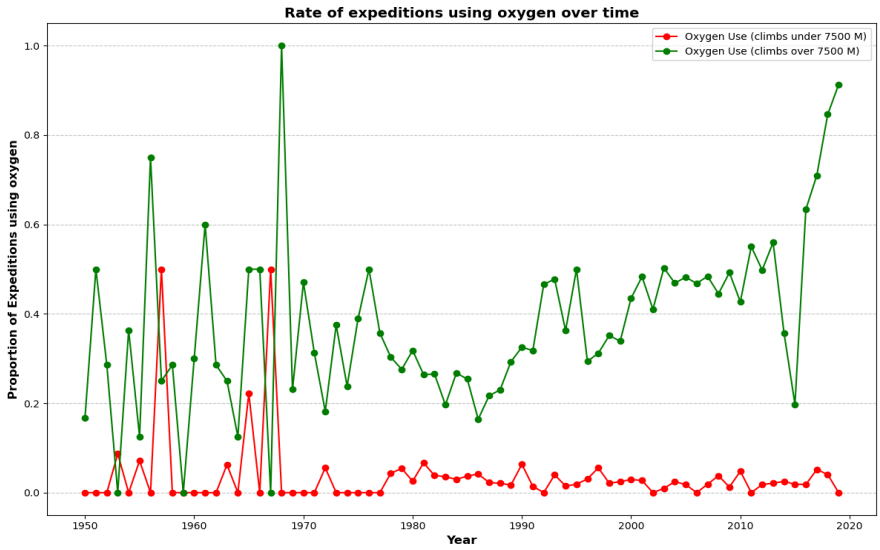
Expeditions Analysis



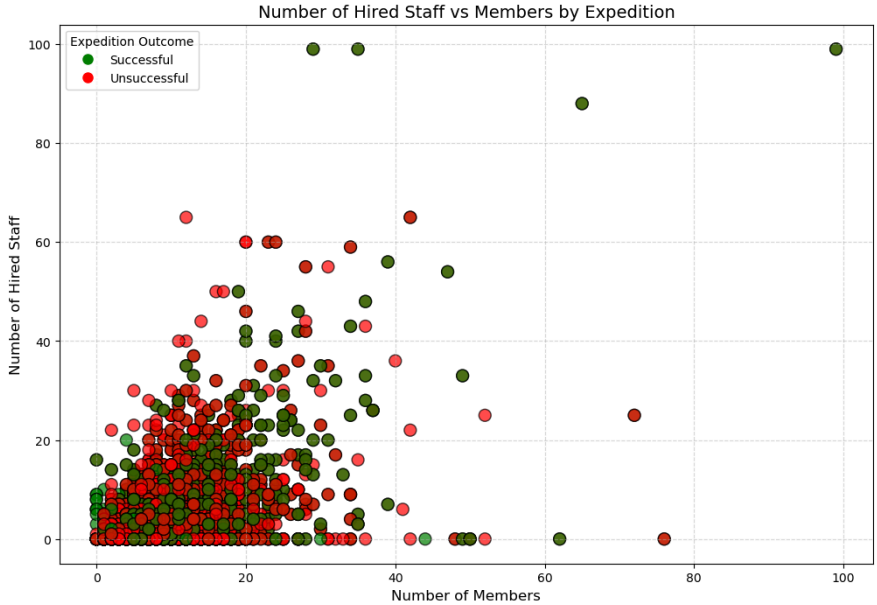
Initially, we looked at success rate by season, and found that spring and autumn are far and away the most likely seasons for expeditions, while summer (the Monsoon season) and winter (the season for high altitude storms and snow) have many fewer expeditions. Expeditions in Spring also have a higher likelihood of success than expeditions at other times during the year.

We also looked into the use of oxygen over time, separated by height of the mountain climbed. We found that after dropping through the 1980s, oxygen use has consistently risen since 1985 for climbs of over

7500 metres. Many of these climbs enter what is known as the “death zone” in alpinism at 8000 metres, above which a human body cannot survive without oxygen for an extended period. Previously, the frontier in mountain climbing was to attempt climbs to high peaks without the use of oxygen as a way to increase the challenge of the climb, but it seems this ethos has decreased in recent years. Climbs under 7,500 metres rarely see the use of oxygen.

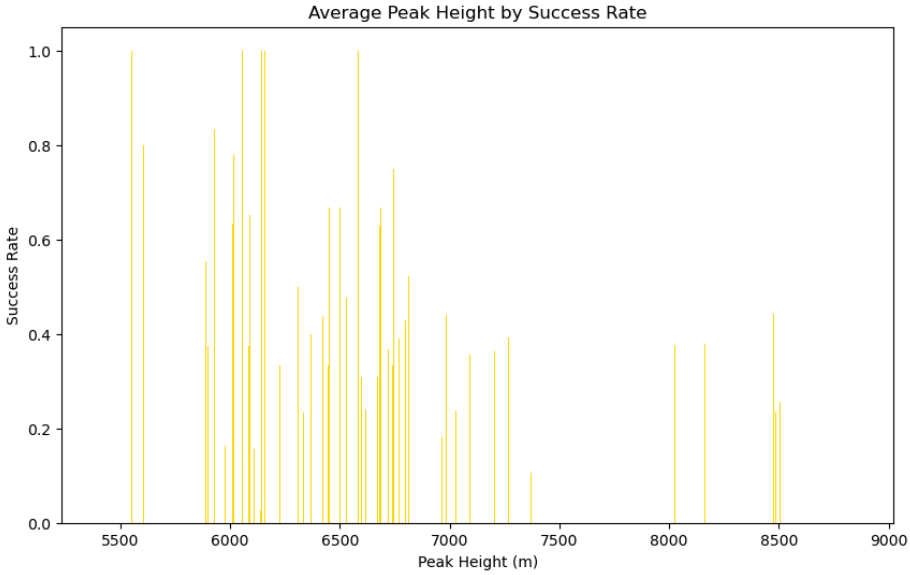


Merged Dataset:

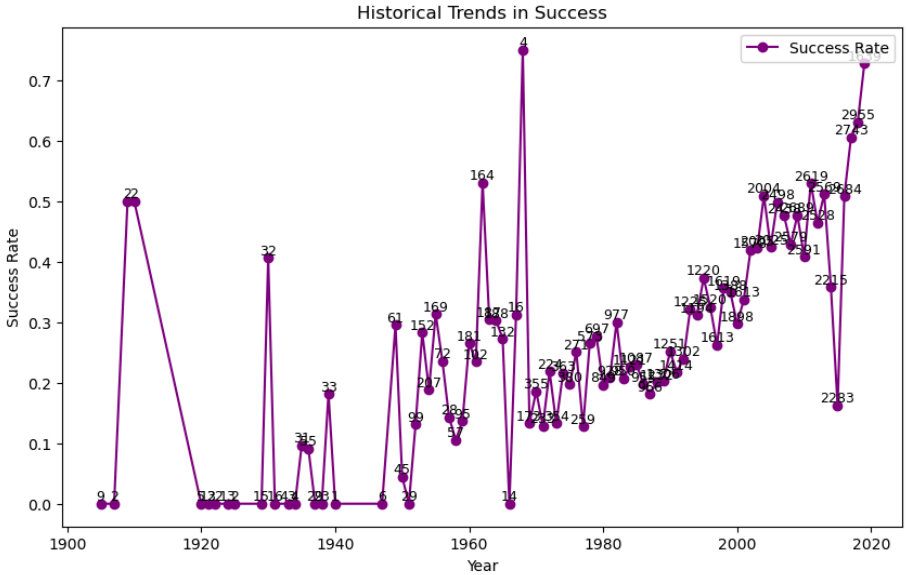


As can be seen above, successful expeditions tend to occur more frequently as the number of members and hired staff increases, especially for groups larger than 20, suggesting a positive correlation between expedition scale and success. In contrast, smaller expeditions dominate the unsuccessful outcomes, though a few larger-scale expeditions (with more than 40 members or staff) also failed. A few outliers with unusually high numbers of members or staff (e.g., 60–100) are mostly successful, indicating that

well-resourced expeditions are more likely to achieve their goals. We further explored the background of staff hired (Appendix C).



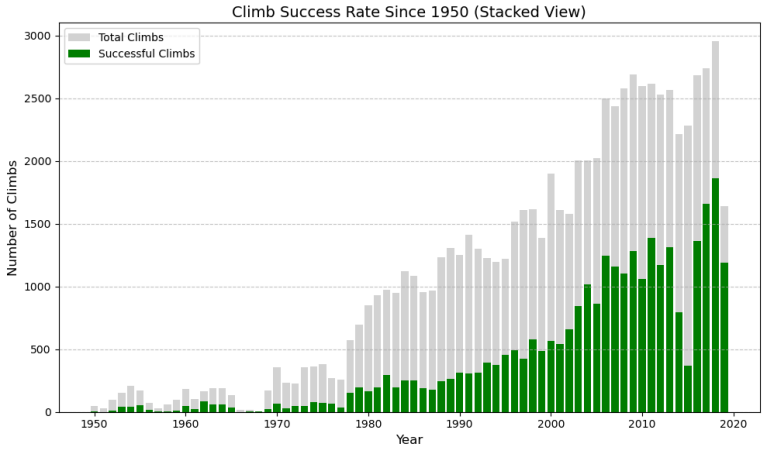
Success rates for peaks below 6000 meters tend to cluster near higher values, suggesting that lower-altitude peaks are generally easier to summit. Peaks above 8000m show notably lower success rates, highlighting the increased difficulty associated with extreme altitudes. Overall, there is a trend of decreasing success rates as peak height increases, with significant variability in mid-range altitudes.



Early expeditions (pre-1940s) had low and inconsistent success rates, likely due to limited experience, and fewer overall expeditions, leading to more variability. The mid-20th century shows notable spikes in success rates, though these are based on fewer expeditions. After the 1980s, success rates stabilized and gradually improved, reflecting advancements in gear, planning, and knowledge about peak climbing. The overall trend highlights the steady increase in expedition success rates over the decades, with modern expeditions achieving significantly higher and more consistent success.

Findings and Analysis:

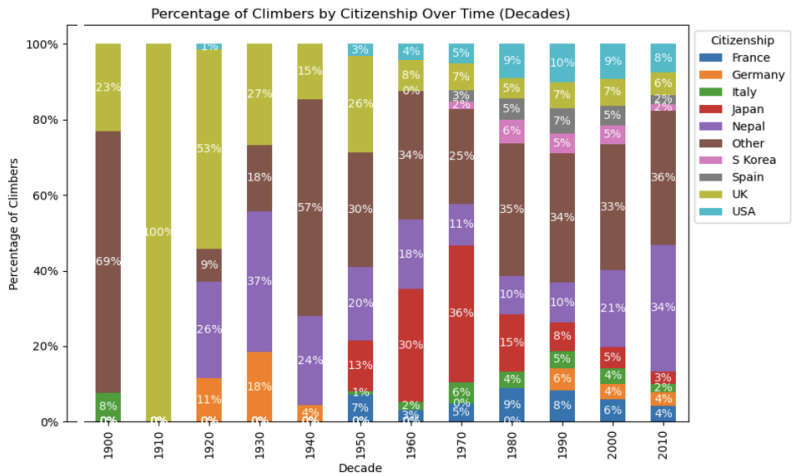
Returning to our analysis objective - what are possible factors/elements that contribute to (or detract from) the success of an expedition? We can start by looking into Members and Expedition data where “Success” is a listed category, defined in the data as the completed ascent of the peak. In Members data we see an increase in the volume of total climbs as well as an increase in the ratio of successful number of climbs when adjusted for climb frequency. Interestingly, while the ratio of successful climbs continues to increase, we do see the ratio of failed climbs decrease around the late 1970s onwards.



Moving to our secondary objective of understanding how the sport of climbing has evolved between 1900-2019. From a demographic perspective, we are concerned with citizenship, sex, and age of climbers. We start by looking at how the citizenship of the climbers has evolved over the period with the following two charts.

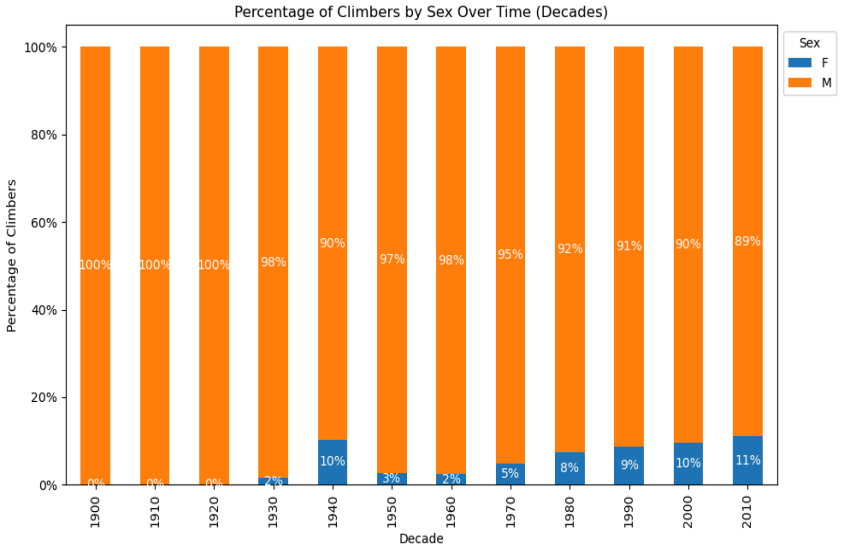
These figures illustrate how the citizenship makeup of climbers changed over time. To streamline the analysis, the Percentage of Climbers is calculated by dividing the number of climbers from each country (including hired and non-hired, successful and unsuccessful) by the total number of climbers in each decade. The “Other” category includes all citizenships outside of the top nine, including those with multiple citizenships.

Between 1900 and 1949, climbers from the UK, Nepal, Germany, and the Other category represented the largest share. In the 1950s, a new trend emerged with climbers from the USA, Japan, Italy, and France joining expeditions. This coincides with Nepal opening its borders to Western expeditions in 1950.



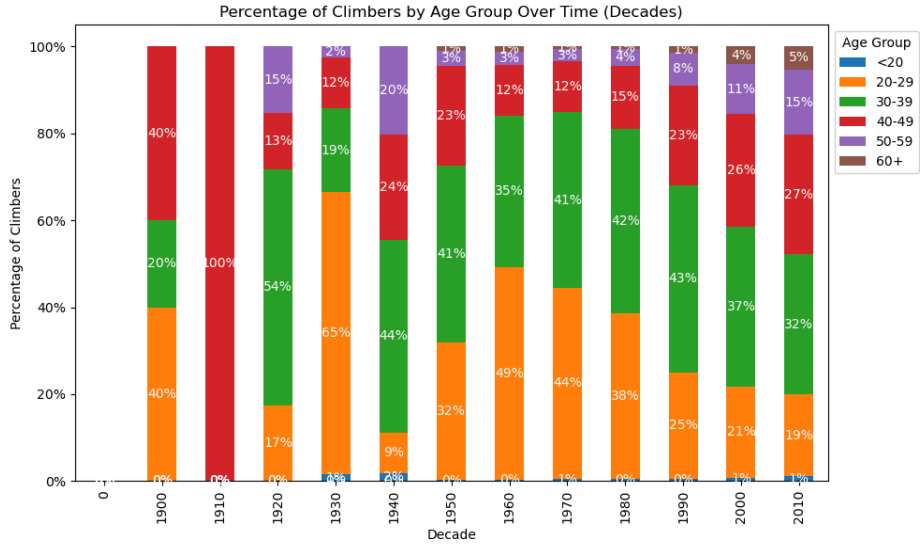
Between the 1950s and 1970s, climbers from Japan accounted for a significant portion of the total, reaching approximately 35% in the 1970s. In the period from the 1990s to the 2010s, the proportion of climbers from Nepal increased from 10% to 34%. This growth may be attributed to the increasing reliance on hired staff, who are often local Nepali individuals.

In the next charts, we dig into the proportion of climbers who are male vs. female over the observed period.



Between 1905 and 2019, female representation in climbing remained low, with at most 11% of climbing members identifying as female. However, the charts reveal a few notable insights: In the 1940s, females made up 10% of climbers—a significant increase compared to the preceding and following decades. This period saw a sharp decline in overall climbing activity, likely due to World War II, where males may have been more likely to be in the armies at the time. Between the 1960s and 1980s, the proportion of female climbers increased notably from 2% to 8%. This rise may be linked to Second Wave Feminism (1960s–1980s), which reshaped societal norms and expanded opportunities for females.

After World War II, as climbing gained popularity, the most common age groups for climbers were 20-29, 30-39, and 40-49. Between the 1960s and 2010s, several interesting trends emerge and can be explained by several factors (See Appendix D):



- The 40-49 age group grew from 12% to 27%.
- The 20-29 age group declined significantly from 49% to 19%.
- The 50-59 age group increased from 3% to 15%.

Predictive Analysis

To better understand the relationships in the data, we made two exploratory logistic regression models, each with different dependent variables:

- Successful ascent: termination reason is listed as Success (main peak, sub peak, or claimed)
- No deaths: no deaths of expedition crews were reported.

A detailed description of the predictor variables we used are in Appendix F.

We chose to use logistic regression as this was the best modelling approach to predict binary outcomes such as expedition success and the deaths of expedition members. Results of exploratory regression analysis for **expedition success** are below.

Overall, we find that the prediction model for this variable was not very accurate, though we do find some support for several predictors of success:

```
Accuracy: 0.6729468599033817
                Feature  Coefficient
2          oxygen_used     1.528778
0            no_deaths     0.683935
6   hired_nepali_staff     0.571980
3 trekking_agnecy_used     0.125786
7              injured     0.025584
1      expedition_size     0.017337
5        past_20_years    -0.029487
4               season    -0.143167
8            over_7500    -0.588136
```

- Using oxygen (understandably) makes it more likely that an expedition reaches its goal
- Having no deaths increases likelihood of success
- Hiring Nepali staff raises the likelihood of a successful summit
- Attempting a 7,500 metre or higher peak is associated with a lower chance of success

We performed the same regression, trying to predict expeditions **without any deaths**:

```
Accuracy: 0.9289855072463769
                Feature  Coefficient
4        past_20_years     0.615413
2 trekking_agnecy_used     0.497107
1          oxygen_used     0.072800
3               season    -0.004645
0      expedition_size    -0.035928
5   hired_nepali_staff    -0.320984
6              injured    -0.415337
7            over_7500    -0.884238
```

We find that recent expeditions, and expeditions using a trekking agency are both linked to a lower likelihood that an expedition will result in death for one of its members. Death is much more likely for climbs over 7,500 metres, and injuries are predictive of death. Hiring Nepali staff is linked to a lower chance of having no deaths, though this may be because Nepali staff are hired for more dangerous climbs.

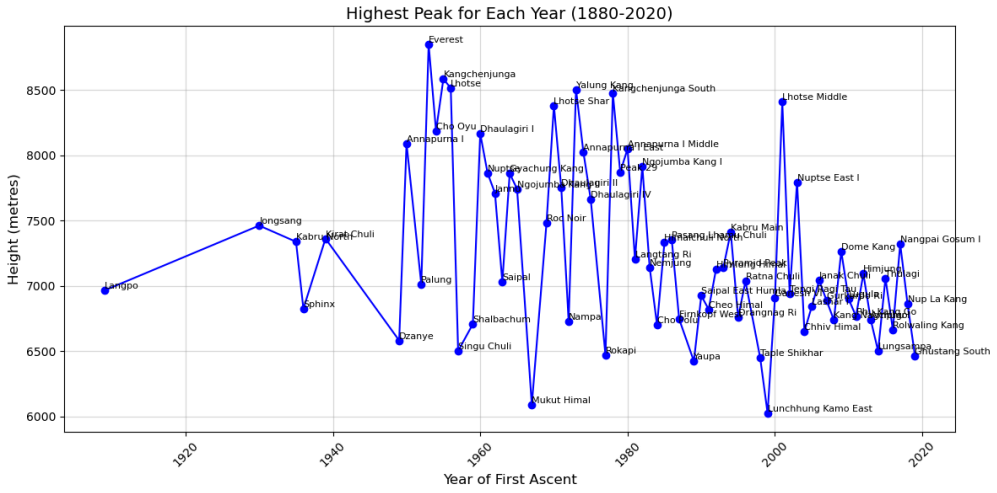
Appendix

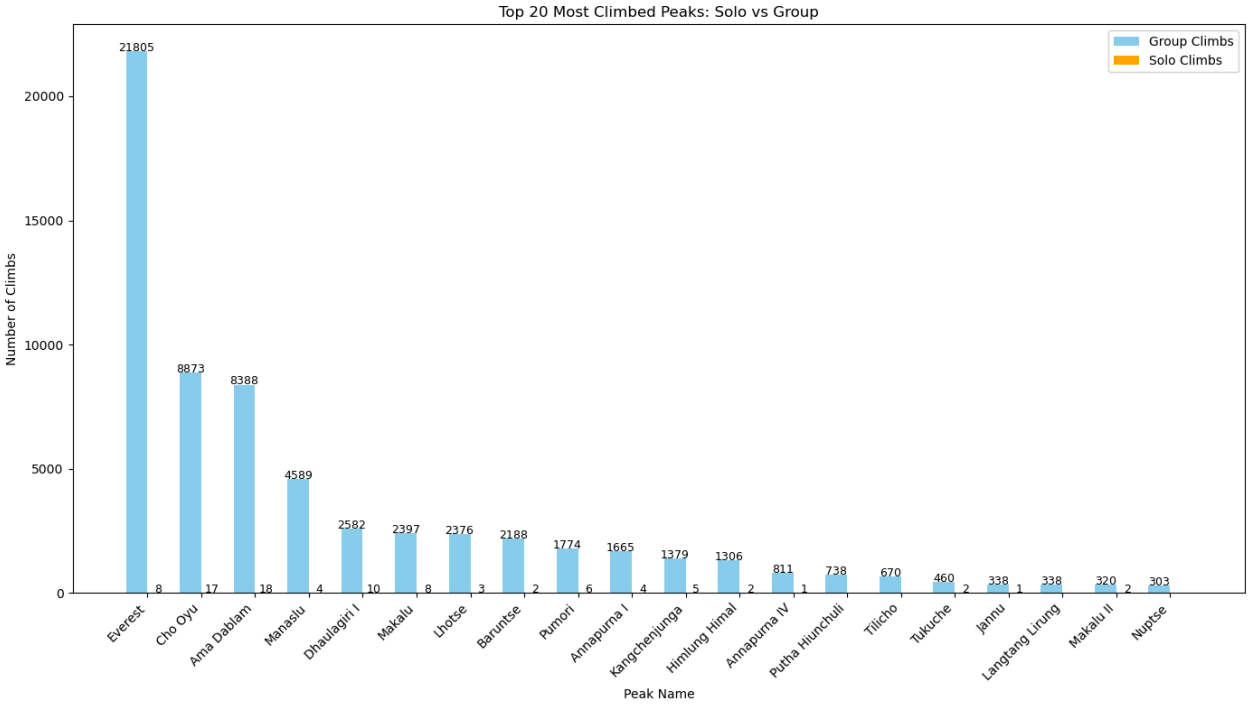
A. When consulting external literature regarding changes across these three inflection points that result in the sudden spike in frequency in the late 1950s through late 1970s, drop down around 1980 through the late 1990s, and the spike again in 2000s what could be an explanation, if any, for such trends? Possible answers include:

- **Surge in late 1950s**: Political changes played a crucial role in shaping Himalayan mountaineering expeditions. Before 1950, Nepal was closed to foreign climbers due to the isolationist policies of the ruling Rana dynasty. However, geopolitical tensions, including China's occupation of Tibet and fears of communist expansion, prompted Nepal to open its borders to Western expeditions, allowing mountaineering to flourish.
- This shift enabled landmark climbs like the 1950 French Annapurna I expedition, setting the stage for future Himalayan expeditions and international mountaineering competition.
- **Dip in late 1970s - 1980s**: Global Economic Recession (1970s-1980s): Economic downturns likely reduced funding for large-scale expeditions, which rely heavily on sponsors and national organizations.
- **Rebound in the 2000s:** The resurgence in the 2000s coincided with improved political stability in Nepal, better mountaineering technology, adventure tourism growth, and increased commercial expeditions on peaks like Everest and Annapurna.

Looking at the highest peak climbed by year from 1880 to 2020 to identify any possible pattern. Our entry hypothesis was that there would be a general trend upwards in terms of peak height over time. Interestingly this is not the case. Looking into the written history of changes in Himalayan mountain climbing we do find a possible answer for this trend is that once all of the main and highest summits had been ascended, climbers sought greater

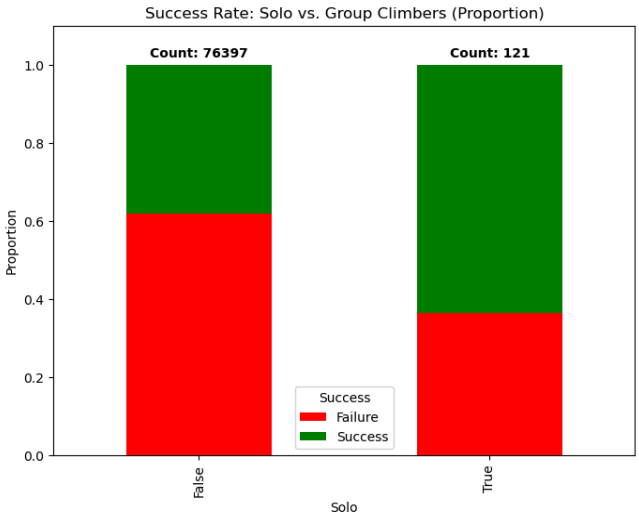
challenges by taking harder routes, using minimal gear, or avoiding supplemental oxygen[1].



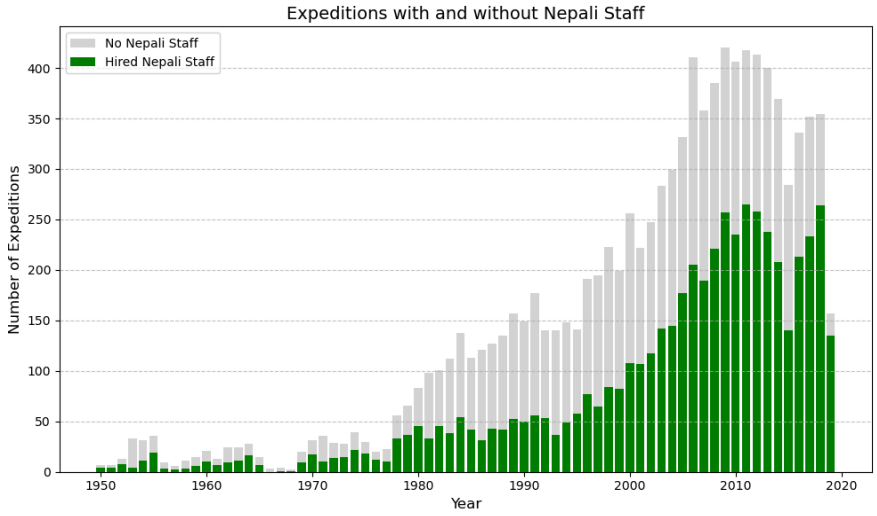


B. We looked into the success rate of solo vs. multi-person expeditions, and found that while there were many fewer solo climbs, the success rate was comparable, if not slightly better than group expeditions.

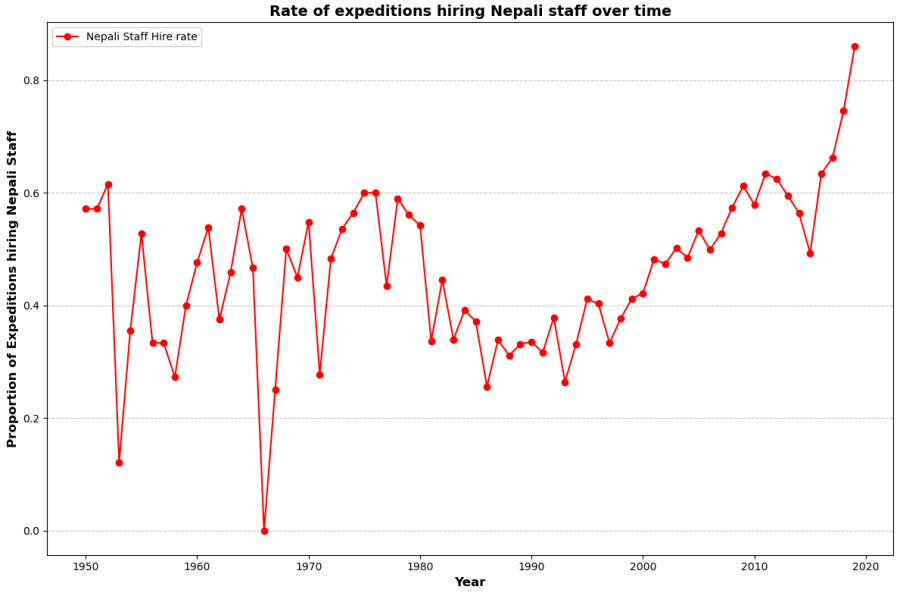
[1] Encyclopaedia Britannica. "Himalayas: Study and Exploration." *Britannica*. Last modified Dec 5, 2024 . Accessed December 10, 2024. https://www.britannica.com/place/Himalayas/Study-and-exploration.



C. We went on to investigate the usage of specifically Nepali hired staff over time, and found that the number of expeditions that use Nepali Sherpa hired staff has steadily increased over time, as the number of expeditions in the Himalayas has grown substantially.



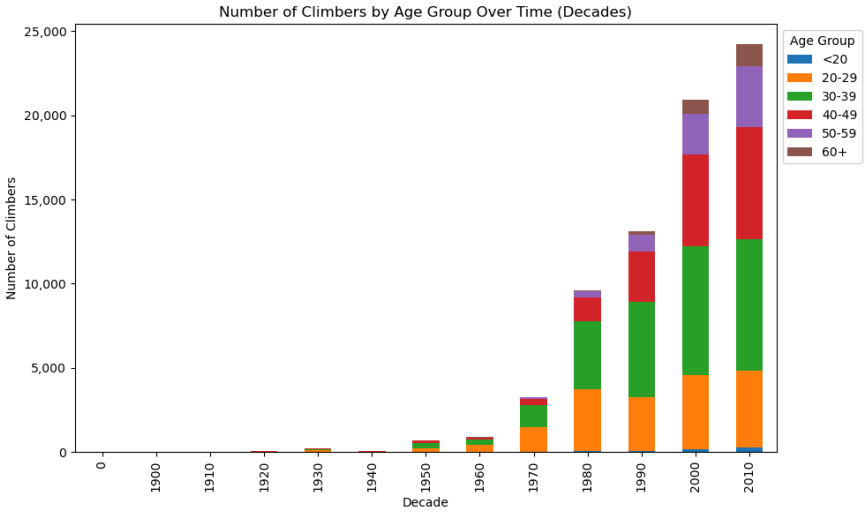
D. We also looked at the proportion of expeditions that hire specifically Nepali hired staff, and find that as time has gone on, more expeditions have decided to hire Nepali staff.

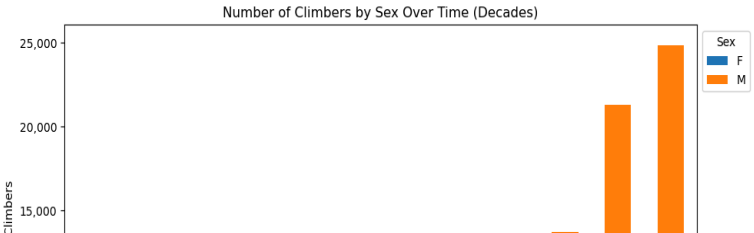


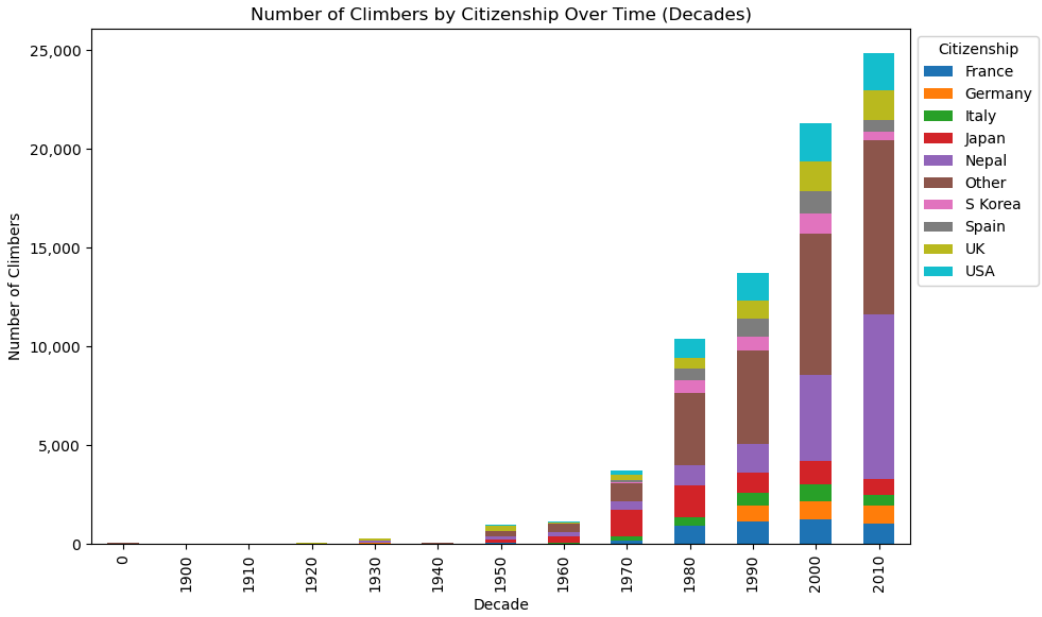
E. These shifts indicate that over time, climbing has increasingly become a sport for older individuals.

Several factors may explain this trend:

1. Advances in technology and healthcare have allowed people to maintain their physical fitness and health for longer periods.
2. Increased costs associated with climbing expeditions—due to advancements in climbing gear and logistics—may mean that older individuals are more likely to have the financial resources and time necessary to participate in these adventures







E. Datasets information:

Members data:

- This dataset contains 76518 rows and 12 columns
- Columns of this dataset:
    - ‘expedition_id', 'member_id', 'peak_id', 'peak_name', 'year', 'season', 'sex', 'age', 'citizenship', 'expedition_role', 'hired', 'highpoint_metres', 'success', 'solo', 'oxygen_used', 'died', 'death_cause', 'death_height_metres', 'injured', 'injury_type', 'injury_height_metres'
- Dataset description:

Peaks data:

- This dataset contains 469 rows and 8 columns.
- Columns of this dataset:
    - 'peak_id', 'peak_name', 'peak_alternative_name', 'height_metres', 'climbing_status', 'first_ascent_year', 'first_ascent_country', 'first_ascent_expedition_id'
- Dataset description: Contains peak-specific information related to the Himalayan expeditions from 1905 to Spring 2019.

Expeditions data:

- This dataset contains 10364 rows and 16 columns.
- Columns of this dataset:
    - 'expedition_id', 'peak_id', 'peak_name', 'year', 'season', 'basecamp_date', 'highpoint_date', 'termination_date', 'termination_reason', 'highpoint_metres', 'members', 'member_deaths', 'hired_staff', 'hired_staff_deaths', 'oxygen_used', 'trekking_agency'
- Dataset description: Contains expedition-specific information related to the Himalayan expeditions from 1905 to Spring 2019. It includes expedition-level data such as dates (basecamp and termination), peak targeted, outcomes (e.g., deaths, injuries, high points reached), team

details (e.g., members, hired staff), and logistical information like trekking agency and oxygen use.

F. Predictive analysis variables:

- Number of expedition members
- Oxygen Used (expedition level)
- Trekking agency (binary, used or not)
- Nepali Hired Staff (binary - did the expedition use hired staff of Nepali nationality)
- Season: Spring (binary - 1 if climb is in spring season, 0 otherwise, as March - April is the most stable season to climb, before the Monsoon in May, and the storms of fall and winter)
- Past 20 years (binary, 1 if expedition was made in the last 20 years, 0 otherwise, to capture advances in technique)
- Presence of injuries (0 if no injuries, 1 if any member was injured)
- Peak is over 7500 metres tall (1 if yes, 0 if no)