Understanding Factors Contributing to a Successful Himalayan Expedition*

Analysis using data from Himalayan expeditions from 1905 through Spring 2019 leveraging Bayesian Logistic Regression

Kaavya Kalani

April 18, 2024

This study analyses the relationship between different demographic, environmental and geographic factors and the successfulness of an attempt to summit a peak. Data from Himalayan expeditions in Nepal from 1905 through Spring 2019 is used and a Bayesian Logistic Regression model is leveraged to analyse the trends and factors influencing a successful summit. I find a strong relation between height of the peak, sex, age, season of expedition, if it was a solo ascent, and the success in summitting. Young age, being a man and embarking on the expedition in Spring or Autumn are some factors which increase one's chances of having a successful summit. The insights from this study aim to help future expedition planning, risk management, and safety protocols.

Table of contents

*Code and data supporting this analysis is available at: https://github.com/kaavyakalani26/himalayan-expeditions-analysis

1 Introduction

Mountaineering, with its blend of adventure and challenge, has captivated explorers for generations. It represents a pinnacle of human endeavor, testing physical prowess, mental fortitude, and strategic planning against some of the most formidable natural landscapes on Earth. Amidst the allure of conquering these majestic peaks, lies a critical question: what factors contribute to the likelihood of a successful summit attempt? This paper delves into precisely this inquiry.

There are numerous mountain ranges around the world where people embark on expeditions. One of the most famous mountain ranges are the Himalayas, which also consist of the highest mountain peak in the world, Mt. Everest. My paper uses the data from expeditions to the Himalayan mountain ranges in Nepal.

My estimand is the relationship between different demographic, environmental and geographic factors (such as height of the peak, sex, age, season of expedition and if it was a solo ascent) and the successfulness of climbing a summit. Using the analysis dataset, my goal is to identify trends and factors that influence a successful expedition and eventually conclude what factors help in a more successful attempt.

I use data from Cookson (2020) which are sourced from The Himalayan Expedition records (link), to understand these factors and trends. This is done by leveraging a Bayesian Logistic Regression model and then predicting the probability of a successful attempt over various demographic and environmental factors.

My analysis shows how mountaineering is a highly male dominated activity. It highlights that young age, being a man and embarking on the expedition in Spring or Autumn are some factors which increase one's chances of having a successful summit.

The findings highlighted by this research have practical implications for expedition planning, risk management, and safety protocols. Ultimately, the study aims to improve decision-making in high-altitude mountaineering, making it safer and more informed in one of the world's most challenging environments.

The paper is further organized into four sections. Section 2 discusses how the dataset to be used for the analysis was obtained and pre-processed. I will explain the variables of interest in the dataset used for the analysis. Section 3 describes the model being used for the analysis. Section 4 then highlights and discusses the trends and associations found during the analysis. Lastly, Section 5 talks about some interesting trends found in Section 4 in depth, link it to the real world and also highlight the weaknesses and future of my analysis.

2 Data

For this analysis, I have combined three datasets into one, which is used for analysis. The datasets were cleaned and analysed using the statistical programming software `R` (R Core Team 2023) along with the help `tidyverse` (Wickham et al. 2019), `knitr` (Xie 2014), `ggplot2` (Wickham 2016), `here` (Müller 2020), `dplyr` (Wickham et al. 2023), `rstanarm` (Goodrich et al. 2024), `arrow` (Richardson et al. 2024), `broom.mixed` (Bolker and Robinson 2022), `modelsummary` (Arel-Bundock 2022) and `kableExtra` (Zhu 2024).

2.1 Analysis Dataset

The raw datasets were obtained from Cookson (2020). I chose the ones cleaned for Himalayan expeditions. Alex got his datasets from The Himalayan Database (Salisbury (n.d.)).

The Himalayan Database is a compilation of records for all expeditions that have climbed in the Nepal Himalaya. The database is based on the expedition archives of Elizabeth Hawley, a longtime journalist based in Kathmandu, and it is supplemented by information gathered from books, alpine journals and correspondence with Himalayan climbers.

The original database currently covers all expeditions from 1905 through Spring-Summer 2023 to the most significant mountaineering peaks in Nepal. Also included are expeditions to both sides of border peaks such as Everest, Cho Oyu, Makalu and Kangchenjunga as well as to some smaller border peaks. Data on expeditions to trekking peaks are included for early attempts, first ascents and major accidents. The updates to this database are published bi-annually.

My dataset derived from Alex's contains the entries from 1905 through Spring 2019.

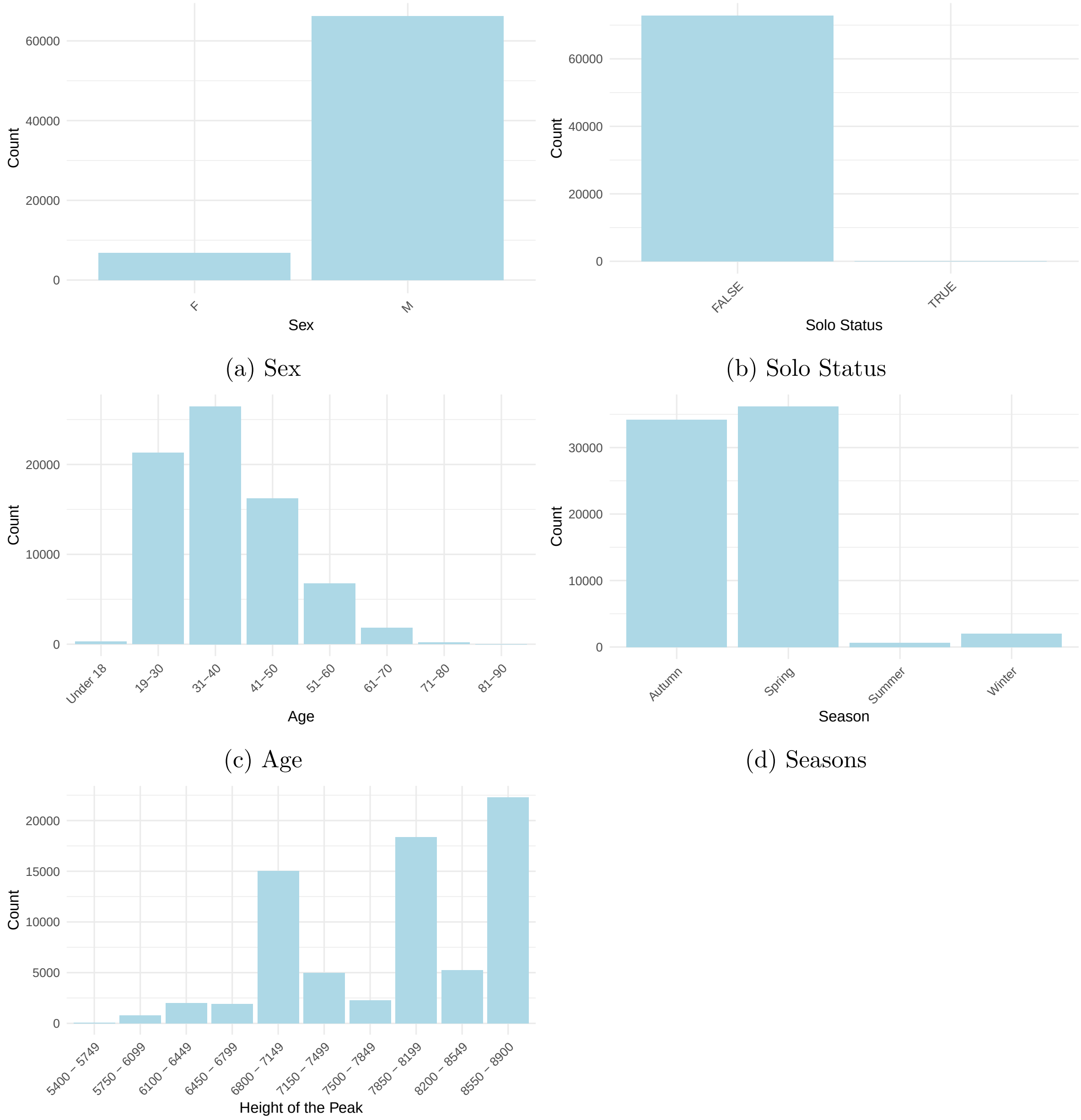
The three datasets I considered included information about all peaks, all expeditions on those peaks and all members on those expeditions. The data from these three datasets are combined to form the main analysis dataset. A person becomes an entry in my analysis dataset if, between 1905 and Spring 2019, they attempted to climb any one of the many Himalayan peaks in Nepal. It also included expeditions to both sides of border peaks as mentioned before.

Among the overall range of variables available, I chose the following to be included in the analysis dataset.

- Height is the height range in which the peak's height in metres falls. This is for the peak the person in the current entry is on an expedition for. The categories for this are 5400 - 5749, 5750 - 6099, 6100 - 6449, 6450 - 6799, 6800 - 7149, 7150 - 7499, 7500 - 7849, 7850 - 8199, 8200 - 8549 and 8550 - 8900.
- Seasons is the season the expedition is embarked in. This takes on either of the four values: Autumn, Spring, Winter, Summer.
- Sex is the sex reported by the expedition member and it is either male or female.
- Age is the age group in which the expedition member fell in at the time of the expedition. Depending on the best available data, this could be as of the summit date, the date of death, or the date of arrival at basecamp. The different categories for this are Under 18,19-30, 31-40, 41-50, 51-60, 61-70, 71-80 and 81-90.
- Successrepresents whether the person's expedition resulted in a successful summit.
- Solo represents whether the person attempted a solo ascent.
- Died represents whether the person died during the expedition.

Figure 1 shows the counts for different variables we are considering to model. We see that there are significantly higher men than women. This shows the existence of a gender imbalance skewed towards men in mountaineering in the Himalayas. We also see other trends like most of the expedition members choose to not do solo ascents, most of the members fall in the middle-aged range and the peak expedition seasons are the more pleasant autumn and spring compared to the extreme seasons like winter and summer. These difference in counts need to be kept in mind when analysing success proportions.

Figure 2 shows the proportion of expedition members who has a successful outcome or died during their expedition. For sex, we see more men had successful expeditions compared to women. This is interesting keeping in the mind the vast difference in the counts of these two sexes going on expeditions. It might spark an interesting discussion into the gender imbalance in mountaineering. Additionally, we see most of the people who go on expeditions solo have more deaths which might explain the difference in the counts of people attempting solo ascent versus not attempting a solo ascent which was observed earlier. Some other trend we observe are that success rates decline with age and winter season contributes to the most deaths. The success and death proportions based on the height are very varied and do not follow a specific trend. This goes to show that just the height of the peak alone doesn't define the probability of having a successful summit but there are other factors which go into it.



(a) Sex

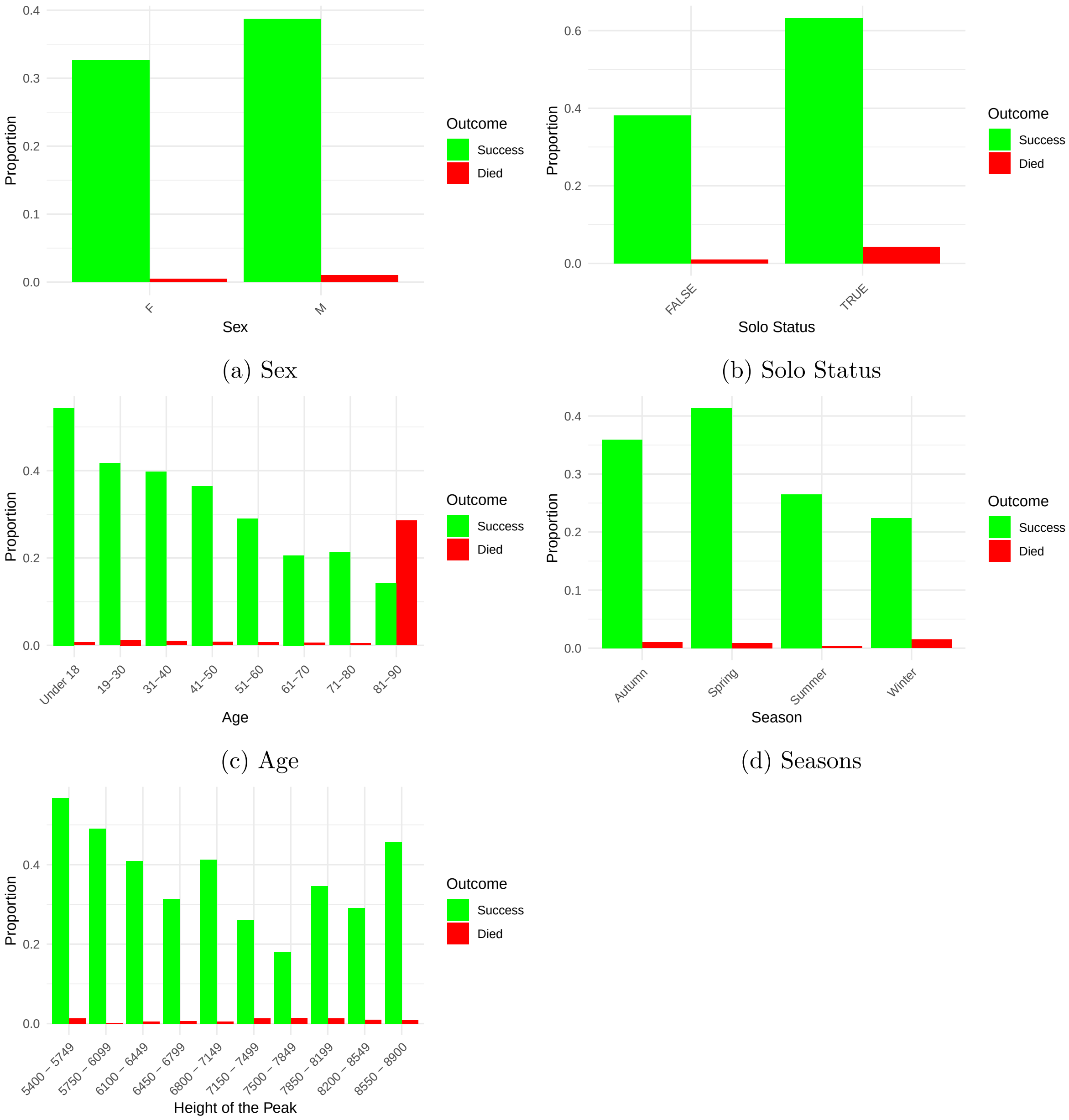
(b) Solo Status

(c) Age

(d) Seasons

(e) Height of the Peak

Figure 1: Counts for the variables of interest



(a) Sex

(b) Solo Status

(c) Age

(d) Seasons

(e) Height of the Peak

Figure 2: Proportions for the variables of interest and the outcome of their expedition

3 Model

I used a Bayesian Logistic Regression model to find the probability that someone will successfully summit the Himalayan peak they are on the expedition for. Logistic regression is a method used for binary classification to predict the probability of a categorical dependent variable.

My model will be based on five independent demographic variables: `height of the peak`, `sex`, `age`, `seasons` and `solo` and the dependent variable will be `success`.

The logistic regression model I will being using is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \times \text{height} + \beta_2 \times \text{sex} + \beta_3 \times \text{age} + \beta_4 \times \text{seasons} + \beta_5 \times \text{solo} \tag{1}$$

$$\begin{aligned}
\beta_0 &\sim \text{Normal}(0, 2.5)\\
\beta_1 &\sim \text{Normal}(0, 2.5)\\
\beta_2 &\sim \text{Normal}(0, 2.5)\\
\beta_3 &\sim \text{Normal}(0, 2.5)\\
\beta_4 &\sim \text{Normal}(0, 2.5)\\
\beta_5 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

where,

- $\hat{p}$ represents the probability that someone will successfully summit the peak they are on the expedition for.
- β_0 represents the intercept term of this logistical regression. It is the probability that someone will successfully summit the peak they are on the expedition for if the predictors' values are zero
- β_1 is the coefficient corresponding to height of the peak
- β_2 is the coefficient corresponding to sex of the person
- β_3 is the coefficients corresponding to age of the person
- β_4 is the coefficients corresponding to seasons of the person
- β_5 is the coefficients corresponding to if the person is attempting the summit alone

In my model, normal priors with a mean of 0 and a standard deviation of 2.5 are used for both the coefficients and the intercept. Setting the mean of the priors to 0 implies that there is no expectation of a particular direction or magnitude for the coefficients or intercept. I chose this as I have no expectation of the same. The standard deviation of 2.5 reflects the uncertainty or variability in the prior beliefs. I chose a moderately wide prior to allow for a reasonable amount of uncertainty.

The chosen priors allow the data to largely determine the posterior distribution as they are relatively non-informative. They don't heavily influence the results unless the data provide strong evidence to the contrary.

The use of moderately wide priors can also help regularize the model, preventing overfitting and providing more stable estimates, particularly when dealing with limited data.

Table 1: Summary of the model

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.76	0.25	0.33	1.18
height_range5750 - 6099	-0.58	0.26	-1.01	-0.15
height_range6100 - 6449	-1.07	0.25	-1.49	-0.64
height_range6450 - 6799	-1.61	0.25	-2.03	-1.18
height_range6800 - 7149	-1.14	0.25	-1.56	-0.72
height_range7150 - 7499	-1.89	0.25	-2.31	-1.46
height_range7500 - 7849	-2.40	0.25	-2.83	-1.97
height_range7850 - 8199	-1.48	0.25	-1.90	-1.06
height_range8200 - 8549	-1.79	0.25	-2.22	-1.37
height_range8550 - 8900	-1.09	0.25	-1.51	-0.66
sexM	0.27	0.03	0.22	0.31
soloTRUE	1.16	0.19	0.84	1.49
age_range31-40	-0.10	0.02	-0.13	-0.06
age_range41-50	-0.26	0.02	-0.30	-0.23
age_range51-60	-0.62	0.03	-0.67	-0.57
age_range61-70	-1.08	0.06	-1.18	-0.97
age_range71-80	-1.03	0.18	-1.34	-0.74
age_range81-90	-209.28	182.58	-597.95	-21.95
age_rangeUnder 18	0.39	0.13	0.19	0.60
seasonsSpring	0.12	0.02	0.09	0.16
seasonsSummer	-0.62	0.09	-0.78	-0.46
seasonsWinter	-0.82	0.06	-0.91	-0.73

Table 1 shows the coefficients for my Bayesian model along with the standard error and the 95% credible interval. The standard error (SE) is a measure of the precision with which a sample statistic estimates a population parameter. It quantifies the variability of sample statistics around the population parameter. A 95% credible interval means that there is a 95% probability that the true parameter lies within the interval, given the observed data and the model assumptions.

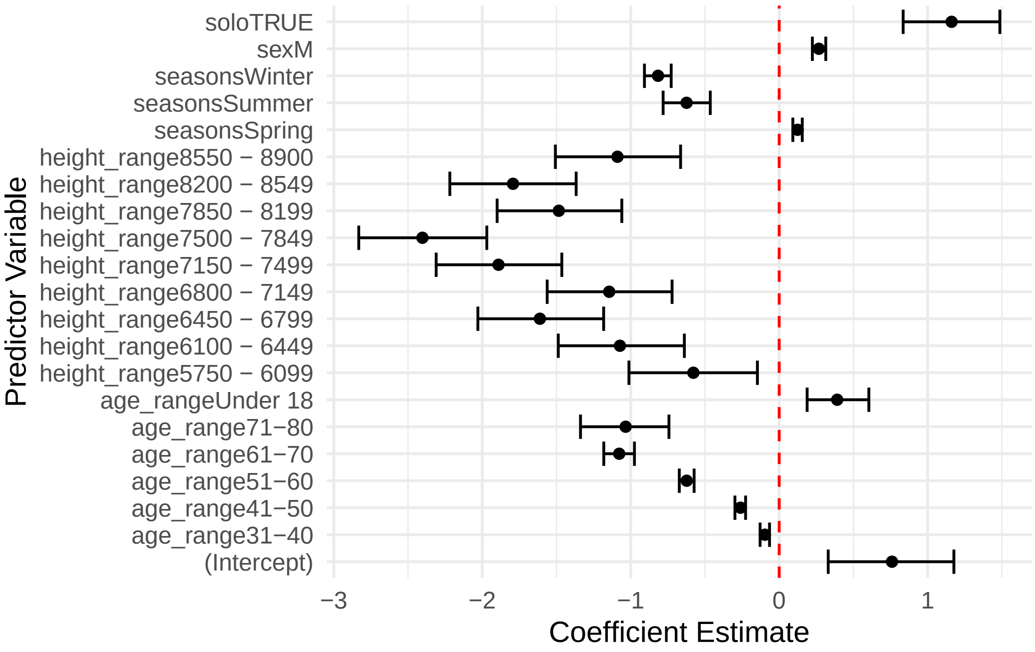


Figure 3: Coefficients of the model

4 Results

Figure 3 illustrates the coefficients and their associated 95% credible intervals for the predictor variables in the Bayesian model. Each point represents the estimated coefficient for a predictor variable, while the horizontal lines depict the credible interval around the estimate. Variables with coefficients to the right of zero indicate a positive association with the outcome variable, suggesting that an increase in the predictor variable corresponds to an increase in the outcome variable. Conversely, coefficients to the left of zero indicate a negative association, implying that an increase in the predictor variable is associated with a decrease in the outcome variable. These insights can help in understanding the direction and magnitude of the relationships between predictor variables and the outcome in the Bayesian model. I removed the coefficient for age_range81-90 which according to Table 1 was very small leading to all other coefficients being uninterpretable in the plot.

We see that the males have a slightly higher success probability compared to females. We also notice how expeditions in Summer and Winter have lesser success probability than Autumn whereas Spring has higher success probability than it. Additionally, we notice the success probabilities getting lower with increase in age.

Figure 4 shows the predicted success probability according the the age and sex of the expedition member. We notice that the success probabilities over all ages tend to be slightly higher in

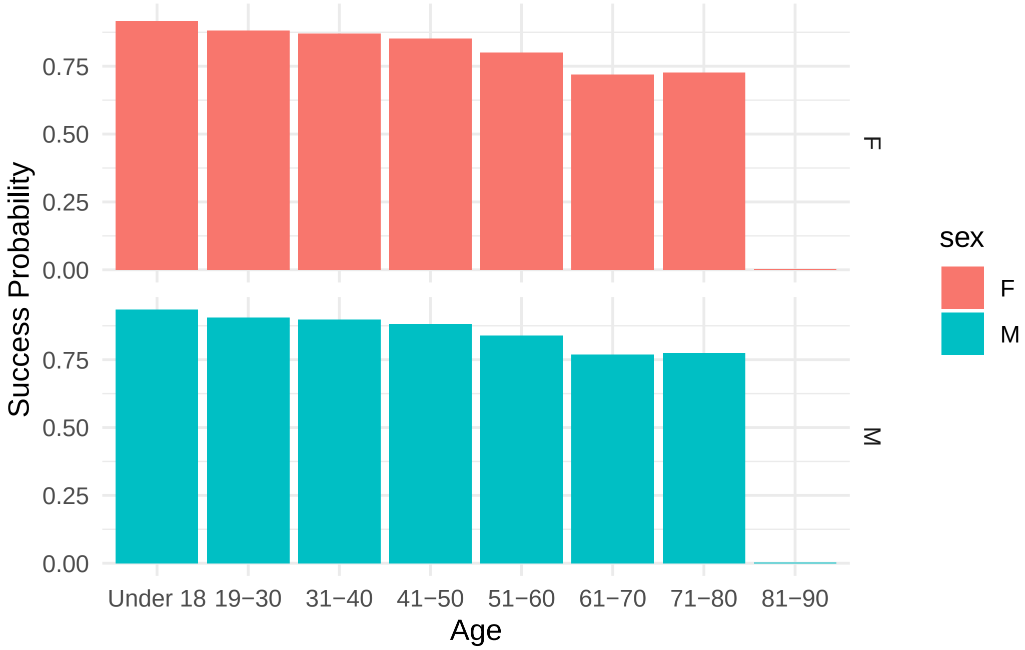


Figure 4: Predicted Success Probability by Age and Sex

males compared to females, just as we saw in Figure 3. But, when combined with age we see there differences are not constant. We see that the difference in the success probabilities for people under 50 are very close to each other but for people aged 51 or above the difference is much larger. The success probabilities are pretty high, crossing 0.8 for younger individuals but starts declining with increase in age. The steep decline in category 81-90 might stem from the significant less number of expedition members in that age group leading to a less varied result.

Figure 5 shows the predicted success probability according the the height of the peak being ascended and the season of expedition. We notice that the success probabilities are pretty high for shorter mountain ranges and become lesser with increase in height. This might have to do with the fact that less higher peaks could be comparatively easier to summit. This, however, cannot be established as the trend as we see a non-pattern being followed for heights 6800 metres and above. We see that winter season expeditions have lower success probabilities, probably owning to the extreme cold conditions at higher altitudes. We also notice interesting trend as an interaction of these two factors. One such trend would be the success probabilities during different seasons for peaks between 5400 metres and 5749 metres are all comparatively closer to each other than compared to peaks between 7500 metres and 7849 metres in height. All seasons' success probabilities for peaks between 5400 metres and 5749 metres fall between 0.85 to 0.9 whereas for for peaks between 7500 metres and 7848 metres, the success probabilities go from around 0.35 in Winter to around 0.6 in Autumn or Spring.

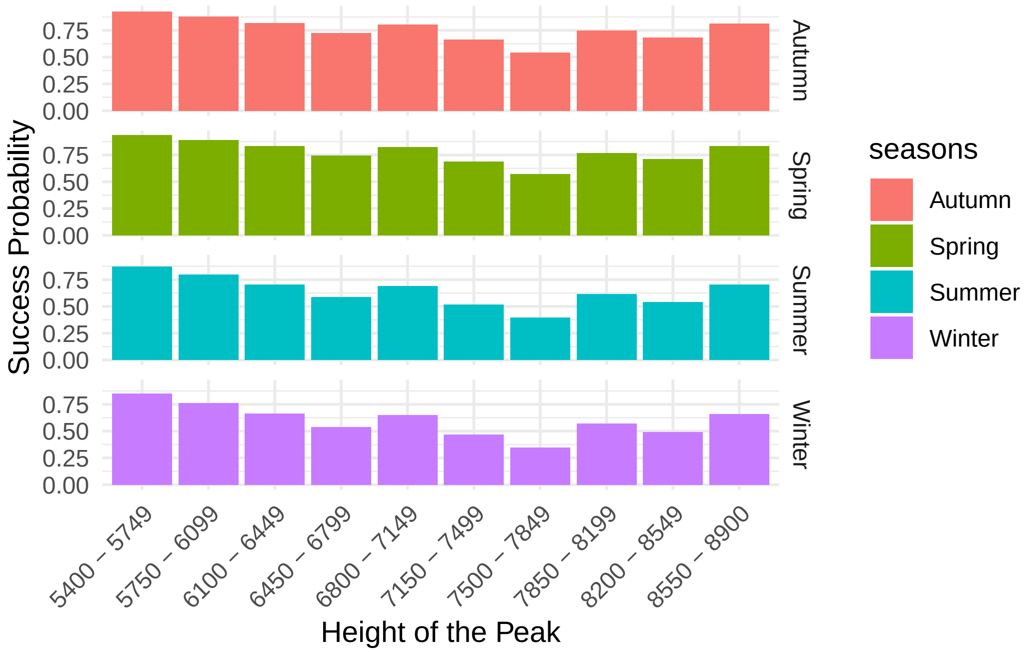


Figure 5: Predicted Success Probability by Height of the Peak and Seasons

5 Discussion

1) What is done in this paper?
2) What is something that we learn about the world?

5.1 Weaknesses and Limitations

5.2 Future Directions

6 Appendix

6.1 Cleaning

For the analysis data, the cleaning steps I took were:

1. Initial merging: The raw data from the expeditions and members datasets is merged based on a common identifier, expedition_id, consolidating information about expedition participants.

2. Column selection and renaming: Irrelevant columns are removed from the merged dataset, and the remaining columns (peak_id.x, season.x, sex, age, success, solo, died) are selected for further analysis. Additionally, column peak_id.x is renamed to peak_id.
3. Secondary merging: The cleaned expeditions dataset is merged with the peaks dataset based on a common identifier, peak_id, to incorporate information about the height of each peak climbed during expeditions.
4. Filtering out incomplete data: Rows with missing values for key variables such as sex or age are filtered out to ensure the integrity of the dataset.
5. New ranges: Height range categories are created based on predefined ranges in meters: 5400-5749, 5750-6099, 6100-6449, 6450-6799, 6800-7149, 7150-7499, 7500-7849, 7850-8199, 8200-8549, and 8550-8900. Age range categories are defined as follows: Under 18, 19-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91 or older.
6. Final dataset creation: season.x is renamed to seasons. and then the resulting dataset is further refined to include only relevant columns (height_range, seasons, sex, age_range, success, solo, died).

6.2 Analysis dataset

Here is a glimpse of the dataset used for analysis

Table 2: Analysis dataset

height_range	seasons	sex	age_range	success	solo	died
5750 - 6099	Autumn	M	19-30	TRUE	FALSE	FALSE
5750 - 6099	Autumn	M	19-30	TRUE	FALSE	FALSE
5750 - 6099	Autumn	M	19-30	TRUE	FALSE	FALSE
5750 - 6099	Autumn	M	19-30	TRUE	FALSE	FALSE
5750 - 6099	Autumn	M	19-30	TRUE	FALSE	FALSE
5750 - 6099	Autumn	M	51-60	TRUE	FALSE	FALSE

6.3 Model summary

6.4 Posterior distribution

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

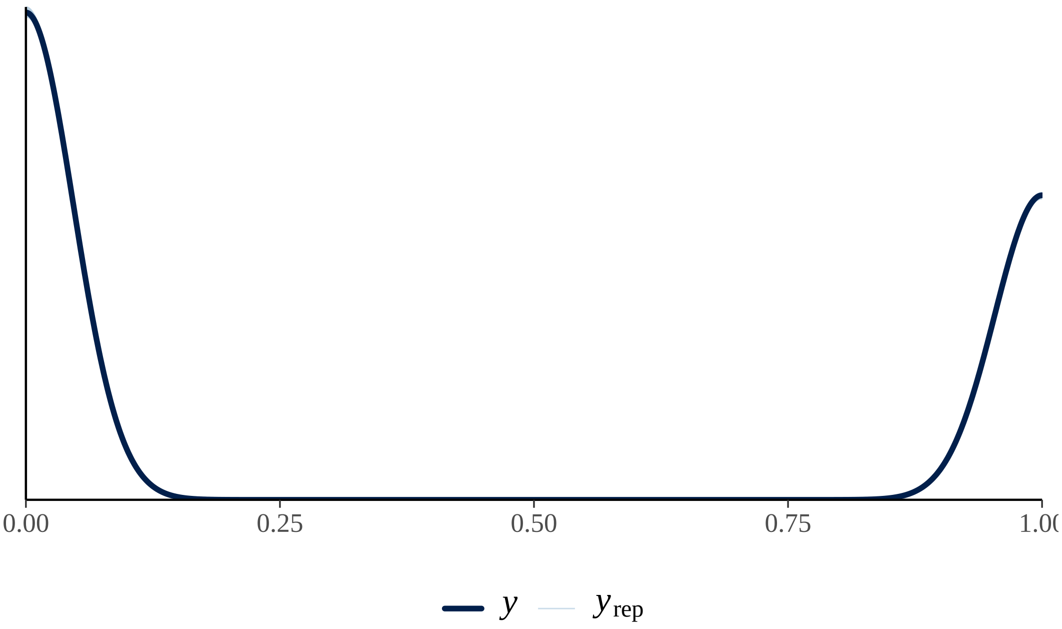


Figure 6: Posterior distribution for logistic regression model

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models.* https://CRAN.R-project.org/package=broom.mixed.

Cookson, Alex. 2020. “Data/Himalayan-Expeditions at Master · Tacookson/Data.” *GitHub.* https://github.com/tacookson/data/tree/master/himalayan-expeditions.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. “Rstanarm: Bayesian Applied Regression Modeling via Stan.” https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://here.r-lib.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'.* https://CRAN.R-project.org/package=arrow.

Salisbury, Richard. n.d. “The Himalayan Database, the Expedition Archives of Elizabeth Hawley.” *The Himalayan Database, The Expedition Archives of Elizabeth Hawley.* https://www.himalayandatabase.com/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.