Proposal

Outliers: Matthew Hawkins, Aston Yong, Catherine Ning

October 26, 2021

Section 1

An introduction to the subject matter you're investigating

For this project, we are looking into the Himalayan database, which is a compilation of all expeditions that have climbed in the Nepal Himalaya. The data cover all expeditions from 1905 through Spring 2019 to more than 465 significant 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 main research question we are hoping to solve involves looking into various different factors which can predict the safety and success of the expedition.

The motivation for your research question (citing any relevant literature)

Many active climbers view the Himalayas as the proverbial "summit" of their climbing experience. These mountains provide a demanding and often dangerous task so, the most important goal for climbers is safety and success. We want to create a model that find the most significant factors for safe and successful expeditions so that future climbers can predict probabilities of safe, successful climbs. Climbing the Himalayas is still a dangerous and risky activity, and increased congestion is changing the dynamic of climbing (http://graphics.reuters.com/NEPAL-EVEREST/0100B4S22JR/index.html). Therefore, examining these dataset predictors and drawing insights from them will hopefully improve mountain climbing preparedness.

The general research question you wish to explore

What expedition factors are most predictive of safe and successful expeditions and what is the probability of success given these predictors?

Your hypotheses regarding the research question of interest.

We hypothesize expeditions with conditions like lower peaks, historically climbed peaks, more members, warm seasons, younger average age, and more recent year to have greater probability of safe success.

Section 2

In this section, you will describe the data set you wish to explore. This includes description of the observations in the data set

The data cover all expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. It is divided into 3 tables, one for peaks, expeditions, and members respectively. The peaks.csv file contains information about each Himalayan mountain with data on peak name, whether it has been climbed, height, and other variables. The expeditions.csv file contains data on individual expeditions, with variables like member count, season, and number of deaths. The members.csv file is the most narrow-scoped table with demographic and personal information for each member on an expedition. Multiple members can participate in one expedition, and multiple expeditions can climb one peak.

Description of how the data was originally collected (not how you found the data but how the original curator of the data collected it).

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 database is updated bi-annuallly, and member information comes from submitted permit applications.

Section 3

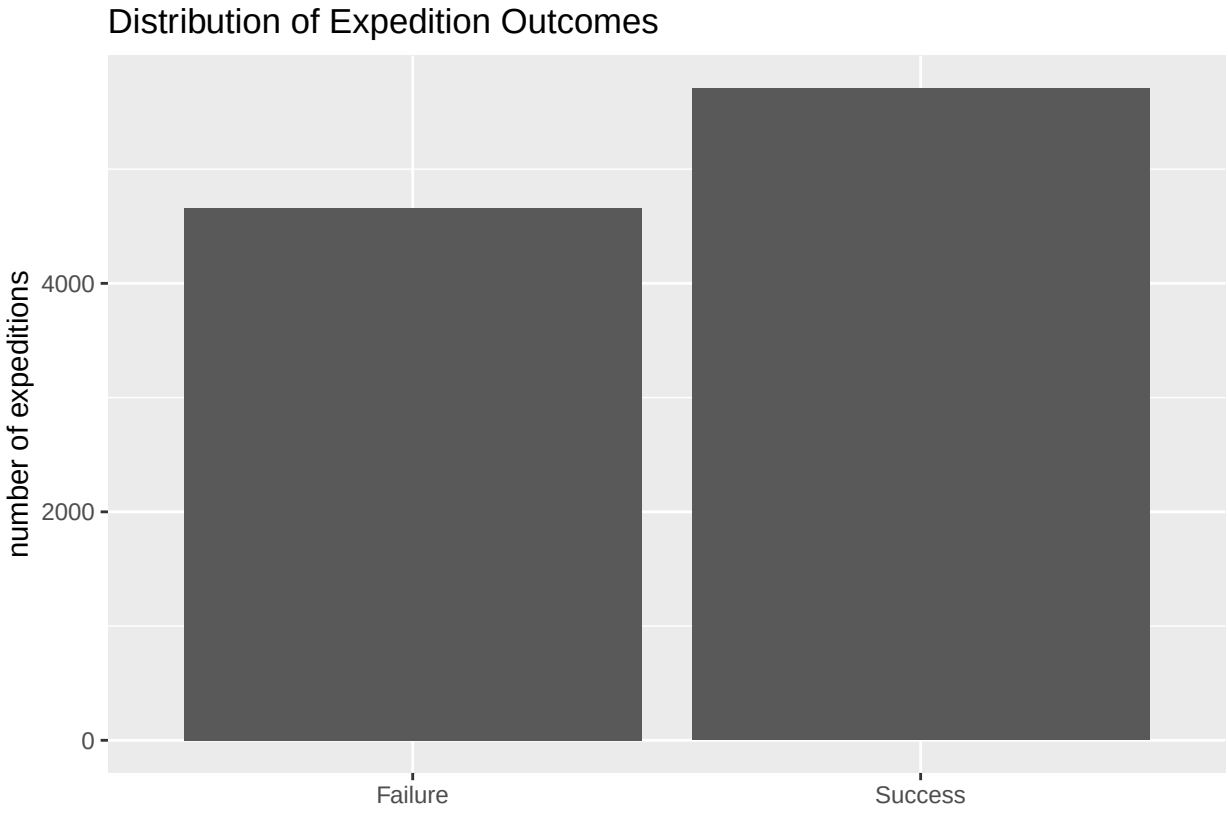
Description of the response variable.

The response variable will be whether or not an expedition was safe and successful. We will use the probability of a success and a threshold to predict expedition outcome. We can use the termination reason variable in the expeditions data set to create a new binary response variable, with the termination reason being success as one outcome and any other termination reason as the other outcome. One non-successful termination reason is Accident (death or serious injury), so safety outcomes are partially captured here.

Visualization

```
## # A tibble: 6 x 17
##   expedition_id peak_id peak_name     year season basecamp_date highpoint_date
##   <chr>         <chr>   <chr>        <dbl> <chr>  <date>        <date>
## 1 ANN260101     ANN2    Annapurna II  1960 Spring 1960-03-15    1960-05-17
## 2 ANN269301     ANN2    Annapurna II  1969 Autumn 1969-09-25    1969-10-22
## 3 ANN273101     ANN2    Annapurna II  1973 Spring 1973-03-16    1973-05-06
## 4 ANN278301     ANN2    Annapurna II  1978 Autumn 1978-09-08    1978-10-02
## 5 ANN279301     ANN2    Annapurna II  1979 Autumn NA            1979-10-18
## 6 ANN280101     ANN2    Annapurna II  1980 Spring 1980-03-25    1980-04-24
## # ... with 10 more variables: termination_date <date>,
## #   termination_reason <chr>, highpoint_metres <dbl>, members <dbl>,
## #   member_deaths <dbl>, hired_staff <dbl>, hired_staff_deaths <dbl>,
## #   oxygen_used <lgl>, trekking_agency <chr>, success_expedition <fct>
```

Note: success (claimed) was not included in our definition of success because the tidy tuesday README does not classify this as legitimate success




```
## # A tibble: 2 x 2
##   success_expedition     n
##   <fct>              <int>
## 1 0                   4657
## 2 1                   5707
```

```
## [1] 0.5506561
```

Within the dataset, 55.066% of expeditions were successful.

The original counts for each termination reason are displayed below. Success(subpeak) and Success(main peak) count as successes and all reasons count as a failure.

```
## # A tibble: 15 x 2
##    termination_reason                                                    n
##    <chr>                                                             <int>
##  1 Accident (death or serious injury)                                  299
##  2 Attempt rumoured                                                     12
##  3 Bad conditions (deep snow, avalanching, falling ice, or rock)      1097
##  4 Bad weather (storms, high winds)                                   1307
##  5 Did not attempt climb                                               233
##  6 Did not reach base camp                                              64
##  7 Illness, AMS, exhaustion, or frostbite                              458
##  8 Lack (or loss) of supplies or equipment                             220
##  9 Lack of time                                                         93
## 10 Other                                                               320
## 11 Route technically too difficult, lack of experience, strength, or moti~   438
## 12 Success (claimed)                                                    20
```

```
## 13 Success (main peak)                                          5581
## 14 Success (subpeak)                                             126
## 15 Unknown                                                        96
```

List of variables that will be considered as predictors

Possible predictor variables are peak_height, number_members, number_hired, oxygen_use, year, season, agency, solo or not, member_deaths, hired_deaths. From the members.csv file, data on avg_member_age, sex ratio, and injuries can be calculated from grouping by expedition.

Regression model technique (multiple linear regression and logistic regression)

We plan to use a logistic regression because we are predicting a categorical response variable and need probability of success to make predictions. Probabilities are at least 0 and at most 1, so a logistic regression model is most appropriate.