DATA SOURCING

SOURCE OF DATA

"Himalayan Expeditions" Accessed from https://www.kaggle.com/datasets/raskoshik/himalayan-expeditions?resource=download&select=expeditions.csv on May 18, 2023.

(See below for "Project Requirements Checklist.")

WHY HIMALAYAN EXPEDITIONS?

The Himalayas is a mountain range in Asia and home of the world's tallest mountain, Mount Everest. In addition to Mount Everest, there are hundreds of peaks in the Himalayas that climbers travel to each year from all over the world to summit. I would like to understand more about the peaks that make up the Himalayas as well as the successful and unsuccessful expeditions that take place each year. (Specifically what patterns are there for successful expeditions vs. failed expeditions or expeditions that result in fatalities?)

COLLECTION

The dataset was retrieved from https://www.himalayandatabase.com/ which is based on the archives of Elizabeth Hawley and is supplemented by information from books, alpine journals, and correspondence from Himalayan climbers. This database is considered trustworthy, as it is the most comprehensive database of expedition records in the Himalayas. Also, it is held in high regard by climbers planning expeditions, journalists and mountaineering historians, and researchers who study patterns of accidents, fatalities, and oxygen use.

CONTENTS

There are 4 data sets available on Kaggle sourced from the Himalayan Database.

The first dataset (and primary dataset I hope to use) is described here:

- "**Expeditions**" – The "expeditions" data shows a record for each expedition on a Himalayan mountain peak from 1905 – 2020 (10,494 records). Each record shows detailed information about the dates of the expedition, the peak, the countries of the climbers who make up the expeditions, the count of climbers and hired guides (sherpas), the count of climbers and guides that summitted, the count of climbers and guides that died, accidents, and the use of supplemental oxygen.

The remaining three datasets are described below and may be used as supplemental data (if needed):

- "**Peaks**" – The "peaks" data shows a record for each peak of interest in the Himlayan Database (462 records). Each record shows details about the peak such as alternate names, region information, height (in feet and meters), restrictions, if it has been climbed or unclimbed, and information regarding the first summit (such as when it was, who summitted it, and countries represented by the first summit expedition).
    - Note on merging: This data set would be interesting to merge with the "expeditions" data set on the **"peak_id"** field. I could then use the peak height information to compare to the expeditions' "max_elev_reached" field.

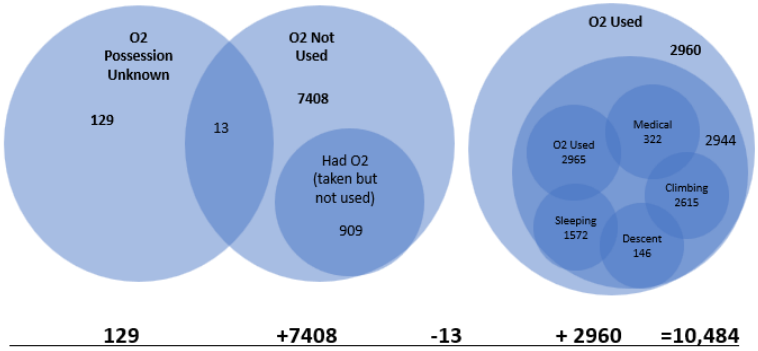
- "**Deaths**" – The "deaths" data shows a record for each climber that has died while on an expedition from 1905 – 2019 (1,104 records). Each record shows the peak name, the climber's name, information regarding when the climb took place, the climber's citizenship, the climber's gender, the climber's age, if oxygen was used, and the cause of death.
    - Note on merging: This data set would be interesting to use in addition to the "expeditions" data set. I would need to create a unique identifier for each record to match to the "expeditions" data. In addition, there could be multiple records in the "deaths" sheet that match a single record in the "expeditions" record (for example, an expedition is recorded as one line in the "expeditions" dataset, but if the expedition had multiple members who died, there will be multiple lines in the "deaths" dataset).
- "**Summiters**" – The "summiters" data shows a record for each climber that has reached the summit (or highest point) of the peak they are climbing from 1909 - 2020 (30,679 records). Each record shows the peak name, climber's name, information regarding when the climb took place, the climber's citizenship, the climber's gender, the climber's age, if the climber used O2, if the climber died on the descent, and the climber's host country.

DATA PROFILE ("EXPEDITIONS" DATA SET)

DATA CLEANING

Duplicates	10 full duplicates found (stored as "df_full_dups"). Created a subset to exclude these duplicates called "df_exp_clean". Confirmed that our number of records was reduced from 10,494 to 10,484.

Inaccurate and Inconsistent Data	How do the categorical O2 data interact with each other? ('expeditions.csv')
	• There are 8 variables that tell us if O2 was or was not used and to what capacity. They can be split into the following categories:
	  ○ POSITIVE - Indication that oxygen was used:
	  ▪ "is_o2_used"
	  ▪ "is_o2_medical"
	  ▪ "is_o2_sleeping"
	  ▪ "is_o2_climbing"
	  ▪ "is_o2_descent"
	  ○ NEGATIVE - Indication that oxygen was not used:
	  ▪ "is_o2_not_used"
	  ▪ "had_o2"
	  ○ UNKNWON - Indication that oxygen possession is unknown:
	  ▪ "is_o2_unkwn"
	• Looking into the positive categories, it looks as though "is_o2_used" is meant to be an umbrella category and is positive if any of the other positive categories are checked.
	• Looking into the negative categories, the name "had_o2" is a bit confusing as it means that the expedition had oxygen but did not use it. For this, **I've changed the column name "had_o2" to "had_o2_unused."**
	• Next, I looked to make sure that a positive category and a negative category were not both selected for a record. Confirmed that issue does not show up in our dataframe.

	• Last, I looked to see that each record had at least one of these O2 categorical variables selected (if not a positive or negative, the record should be listed as UNKNOWN.)
	  ○ There are 105 occurrences where there is not an oxygen category selected.
	  ○ For this, **I changed these 105 records to have a value of 1 for "is_o2_unkwn_2"** to indicate that we don't know what the oxygen situation was for that expedition.
	
	The breakdown for the oxygen categorical variables is as follows:
	 

Missing Data	There is no missing data, however there are categorical variables related to oxygen that should always have a value in one of the 8 oxygen-related columns (this is because variable "is_o2_unknown" is a 'catch-all' for if it is unknown if oxygen was used or not). I created a new list called "df_o2_check_2" to count all the instances where the sum of these 8 fields is 0. There are 105 occurrences that have a value of 0 for this field. I then added two new columns: • 'is_o2_unkwn_2' – this holds a 1 value for all expeditions that have a 1 value in the column 'is_o2_unkwn.' It also holds a 1 value for the 105 occurrences that do not have a value for any categorical o2 variables. • 'o2_check' – this holds the sum of all o2 categorical variables. These new variables are held in the main data frame "df_exp_clean_3."
Data Timeliness	There are 10,484 records that fall between 1905 and 2020. Below, I'm calculating how many records there are in the 21st century (2000-2020), and the most recent 10 years on record (2010-2020). Based on the questions I ask for the project and how many records are available, I could then make a subset for smaller range of years, if needed: • From 2000 – 2020, there are 6,882 records (66% of total records) • From 2010 – 2020, there are 3,681 records (35% of total records)

	It may be interesting to create a flag for 20th century vs. 21st century expeditions to compare patterns of using oxygen or if deaths are more or less common between these two time periods.
Mixed-Type Columns	There are no mixed-type columns.
Outliers	At the end of my script, I've identified outliers for the following variables: • **total_days** – 38 outliers (max outlier value = 280, min outlier value = 75) • **total_mbrs** – 360 outliers (max outlier value = 99, min outlier value = 18) • **mbrs_summited** – 523 outliers (max outlier value = 39, min outlier value = 8) • **mbrs_deaths** – 563 outliers (max outlier value = 10, min outlier value = 1) • **hired_abc** – 906 outliers (max outlier value = 99, min outlier value = 8) • **hired_summits** – 1333 outliers (max outlier value = 48, min outlier value = 3) • **hired_deaths** – 212 outliers (max outlier value = 11, min outlier value = 1) I have not excluded any outliers since each record is an instance of an expedition that has occurred. However, since my questions will be focusing on successful vs. unsuccessful (fatal) expeditions, I may want to split my data up between what an average expedition would look like vs. an expedition that's not typical (in regards to how many total members). With exception to 'total_days', the 'total_mbrs' variable could be influential on the other variables listed (for example, the argument could be made that the more total members on an expedition could mean the more member deaths to occur). I would like to analyze if this correlation exists and the significance of this correlation later in this project, and then split my data accordingly, if needed!

OVERVIEW OF VARIABLES

Column	Keep?	Quantitative? Qualitative?	Time Invariant / Variant	Data Type	Description
peak_id	X	Qualitative	Invariant	object	ID of a peak
peak_name	X	Qualitative	Invariant	object	Peak name
nationality	X	Qualitative	Invariant	object	Principle nationality of expedition
year	X	Qualitative	Variant	int64	Expedition year
season	X	Qualitative	Variant	object	Expedition season
host_cntr		Qualitative	Invariant	object	Country ascent begun
other_cntrs		Qualitative	Invariant	object	Members nationality
sponsor		Qualitative	Invariant	object	Expedition sponsor
leaders		Qualitative	Invariant	object	Leader/leaders of expedition
rte_1_name		Qualitative	Invariant	object	The first route
rte_2_name		Qualitative	Invariant	object	The second route
rte_3_name		Qualitative	Invariant	object	The third route
rte_4_name		Qualitative	Invariant	object	The fourth route

team_asc_1		Qualitative	Invariant	object	Time team ascended first route
team_asc_2		Qualitative	Invariant	object	Time team ascended second route
team_asc_3		Qualitative	Invariant	object	Time team ascended third route
team_asc_4		Qualitative	Invariant	object	Time team ascended fourth route
is_disputed		Qualitative	Invariant	int64	Success is disputed or unverified
is_claim		Qualitative	Invariant	int64	Expedition claims summited without any evidence
is_commercial_rte		Qualitative	Invariant	int64	If the expedition was commercial
is_standard_rte		Qualitative	Invariant	int64	If the route is standard
other_smts		Qualitative	Invariant	object	Other peaks summited
approach		Qualitative	Invariant	object	Main expedition intermediate points to reach the mountain
bc_arrived		Qualitative	Variant	object	Arrival date at the base camp
bc_left		Qualitative	Variant	object	Departure date at the base camp
total_days	X	Quantitative	Invariant	int64	Total # of days spent by the expedition
exp_result	X	Qualitative	Invariant	object	The result of the expedition
is_traverse		Qualitative	Invariant	int64	If the traverse has been made
is_ski_snowboard		Qualitative	Invariant	int64	If the descent was made using ski/snowboard
is_parapente		Qualitative	Invariant	int64	If the descent was made using parapente
term_note		Qualitative	Invariant	object	Summit termination details/results
summit_day		Qualitative	Invariant	object	Day of the summit
time		Qualitative	Variant	object	Summit time
max_elev_reached	X	Quantitative	Invariant	int64	Max elevation reached by expedition
summit_days		Qualitative	Variant	int64	Day of the summit according to total_days feature
total_mbrs	X	Quantitative	Invariant	int64	Total number of expedition members
mbrs_summited	X	Quantitative	Invariant	int64	Total number of expedition members summitted
mbrs_deaths	X	Quantitative	Invariant	int64	Total number of expedition members that died
high_camps		Quantitative	Invariant	int64	Total number of high camps
hired_abc	X	Quantitative	Invariant	int64	Total number of guides hired
hired_summits	X	Quantitative	Invariant	int64	Total number of guides to reach to top
hired_deaths	X	Quantitative	Invariant	int64	Total number of guides that died
rope_fixed		Quantitative	Invariant	int64	Total rope length that has been fixed on the mountain/route
is_no_hired_abc	X	Qualitative	Invariant	int64	If a guide was hired
is_o2_not_used	X	Qualitative	Invariant	int64	If O2 has not been used
is_o2_climbing	X	Qualitative	Invariant	int64	If O2 has been used during climbing
is_o2_descent	X	Qualitative	Invariant	int64	If O2 has been used only on descent
is_o2_sleeping	X	Qualitative	Invariant	int64	If O2 has been used while sleeping
is_o2_medical	X	Qualitative	Invariant	int64	If O2 used for medical reasons
is_o2_used	X	Qualitative	Invariant	int64	If O2 has been used
is_o2_unkwn	X	Qualitative	Invariant	int64	It's unknown whether O2 was used or not
had_o2	X	Qualitative	Invariant	int64	O2 was taken but unused
camp_sites		Qualitative	Invariant	object	Expedition camps locations

accidents		Qualitative	Invariant	object	Accident description
achievements		Qualitative	Invariant	object	Achievement description
agency		Qualitative	Invariant	object	Commercial agency name
members		Qualitative	Invariant	object	Expedition member names (name: L – Leader, S – Summitted, D – Died).

VARIABLE REVIEW AND DESCRIPTIVE STATISTICAL ANALYSIS

QUANTITATIVE VARIABLES

Looking at Quantitative Variables

```
In [13]: df_exp_reduced.describe()
```

Out[13]:

	year	total_days	max_elev_reached	total_mbrs	mbrs_summited	mbrs_deaths	hired_abc	hired_summits	hired_deaths
count	10494.000000	10494.000000	10494.000000	10494.000000	10494.000000	10494.000000	10494.000000	10494.000000	10494.000000
mean	2001.531065	20.280541	7106.911187	5.942920	1.867257	0.074805	2.787116	1.010101	0.030303
std	14.755142	17.655621	1763.608343	5.410798	2.894572	0.374835	5.066210	2.437417	0.282998
min	1905.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1994.000000	4.000000	6530.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2006.000000	18.000000	7200.000000	5.000000	1.000000	0.000000	1.000000	0.000000	0.000000
75%	2012.000000	32.000000	8188.000000	8.000000	3.000000	0.000000	3.000000	1.000000	0.000000
max	2020.000000	280.000000	8850.000000	99.000000	39.000000	10.000000	99.000000	48.000000	11.000000

Interesting takeaways from descriptive statistics:

- The mean year is 2009. With our range spanning from 1905 to 2020, this tells us that a large majority of the recorded expeditions are in the 21st century.
- The average # of days to climb a peak is 18, but this most likely varies depending on how height of the specific peak being summited.
- The max elevation reached is 8850 m, which is the height of the tallest peak – Mount Everest.
- The 99 total members on a single expedition looks to be associated with the "China-Japan-Nepal Friendship Expedition to Qomolangma/Sagarmatha" expedition which took place in 1988.
- The average number of deaths for both climbers and guides is 0 which could mean that deaths are not likely. (As with the elevation, it would be interesting to see if this mean changes based on the specific peak the expedition was held on).

QUALITATIVE VARIABLES

peak_name / peak_id	**nationality**	**season**	**exp_results**
Everest (EVER) has the most recorded expeditions (2155).	The USA has the most recorded expeditions (1167).	Most recorded expeditions occur in Autumn (5209).	Most recorded expeditions had a result recorded of "Success" (5692).
is_no_hired_abc	**is_o2_not_used**	**is_o2_climbing**	**is_o2_descent**
Most expeditions did NOT have a hired guide (7013).	7413 expeditions did not use oxygen.	2615 expeditions used oxygen while climbing.	146 expeditions used oxygen on the descent.

is_o2_sleeping	is_o2_medical	is_o2_used	is_o2_unkwn_2
1572 expeditions used oxygen while sleeping.	322 expeditions used oxygen for medical reasons.	2965 expeditions used oxygen for some undocumented reason.	There are 129 expeditions where it is unknown as to if oxygen was used.
had_o2_unused	**year**		
909 expeditions brought oxygen but did not use it.	The year with most recorded expeditions was 2009.		

LIMITATIONS

The collection of data regarding Himalayan expeditions has an amazing history:

- The data collection started with Elizabeth Hawley in 1960.
- In 1991, Richard Salisbury proposed the idea of transferring Hawley's records to a computer database.
- From 1993 – 2004, Neeta Karmacharya and Namita Shrestha took over the data entry.
- The initial database was published and contained data from 1905 – 2003.
- Elizabeth Hawley retired from her data collection in early 2016, and was succeeded by her long-time assistant Billi Bierling.
- There is now a small team of record collectors that continue the collection of Himalayan expeditions.
- There is also an online form where climbers can report their expeditions.

With the various ways the collection methods have changed since 1960, there are bound to be limitations to the available dataset. For example,

- It should be noted that this database is just a sample of all expeditions ever done. Though it is a vast database, it is more than likely that there have been expeditions that have not been reported (especially prior to Elizabeth Hawley's start in 1960).
- In addition, the Himalayas stretch across 1,500 miles so collecting the data of all expeditioners early on may not have been possible.
- Today, there are record collectors located around the Himalayas as well as an online form, however I could imagine expeditioners could opt out of sharing their information.
- All information regarding the expedition (such as max elevation reached, if the expedition reached the summit, if oxygen was used) is subject to the climber accurately reporting events as they happened.
- With the switch from written notes to a computer database, there is room for data entry errors (as there is the possibility of data entry errors today).

ETHICS

Regarding ethics, there is personal information that is recorded in the database (such as full names). For this purpose of this analysis, I've omitted columns that contain names. If I were to use the "deaths" or "summiters" datasets, these too contain first and last names as well as identifying information such as age and country. If I end up using these datasets, I will omit these columns, as well.

LIST OF QUESTIONS TO EXPLORE

Given the amount of data available, I want to learn more about both the **peaks** (such as tallest peaks, deadliest peaks, and most popular peaks), as well as information about the **expeditions** (such as O2 use, hired guides, and number of members in the expedition).

QUESTIONS ABOUT PEAKS

- What peak is recorded for most attempts?
- What are the tallest peaks?
- What are the deadliest peaks?
    - For the deadliest peaks, what are the top causes of death?
- What are the most popular peaks?

QUESTIONS ABOUT EXPEDITIONS

- What nationality is most represented by climbers?
- Is there a season most climbers choose to climb?
- How many expeditions happen each year?
    - By season?
- What percentage of expeditions bring O2?
    - What percentage of expeditions use O2?
- What percentage of expeditions have hired guides?
    - What percentage of expeditioners are hired guides? (Ratio)
- What is the average number of expedition members?

ADDITIONAL INFORMATION

PROJECT REQUIREMENTS CHECKLIST

EXPEDITIONS DATA SET

Requirement	Notes
Be open-source.	Found on Kaggle.
Authentic/authoritative source.	Sourced from The Himalayan Database (a non-profit organization) – see more in "Collection" section.
Include non-anonymized column names.	Yes.
Be recent. (Ideally no more than 3 years old, but, if necessary, no more than 10 years old.)	Data spans from 1905 – 2020.
Contain at least 2 – 3 continuous variables (apart from index variables, ID variables, dates, years, etc.)	List of continuous variables of interest: • Total days of expedition (total_days) • Max elevation reached on expedition (max_elev_reached) • Total members on the expedition (total_mbrs) • Total members to reach the top (mbrs_summited) • Total members on expedition to pass (mbrs_deaths)

	• Total number of hired guides ('hired_abc') • Total number of hired guides to reach the top ('hired_summits') • Total number of hired guides to pass ('hired_deaths')
Contain at least 2-3 categorical variables (apart from index variables, ID variables, dates, years, etc.)	List of categorical variables of interest: • Peak Name • Season of Expedition • Expedition Results • Were guides hired? ('is_no_hired_abc') • Was O2 used? • Was O2 not used? • Was O2 possession unknown? • Was O2 brought on expedition? • Was O2 used while climbing? • Was O2 used on descent? • Was O2 used while sleeping? • Was O2 used for medical purposes? • Agency
Contain at least 1,500 rows	# of records 1905 – 2020: # of records 2010 – 2020: # of records 2000 – 2020:
Include a geographical object of some kind.	Geographical object: • Nationality of expedition team • Other Countries represented on expedition team
Time series data set.	

ADDITIONAL RESOURCES

"Overcrowding on Mount Everest contributes to rise in deaths." Accessed from https://www.pbs.org/newshour/world/overcrowding-on-mount-everest-contributes-to-rise-in-deaths on May 19, 2023.

"Responsible Tourism in the Himalayas." Accessed from https://www.edreams.com/blog/responsible-tourism-in-the-himalayas/ on May 19, 2023.

"Should Everest be closed?" Accessed from https://www.theguardian.com/world/2006/oct/08/conservation.environment on May 19, 2023.

"Trash and Overcrowding at the Top of the World." Accessed from https://education.nationalgeographic.org/resource/trash-and-overcrowding-top-world/ on May 19, 2023.

"'Why Are You Doing This?' On Mountaineering in the 21st Century." Accessed from https://lithub.com/why-are-you-doing-this-on-mountaineering-in-the-21st-century/ on May 19, 2023.