Himalayan expeditions Analysis

Contexte

A tense context on the slopes of the Himalayas :

- ➔ Constant increase in the number of climbers each year.
- ➔ Increased need for data management and analysis to ensure the safety and success of expeditions.
- ➔ Creation of the Union of Himalayan Agencies (UHA) bringing together the main trekking agencies

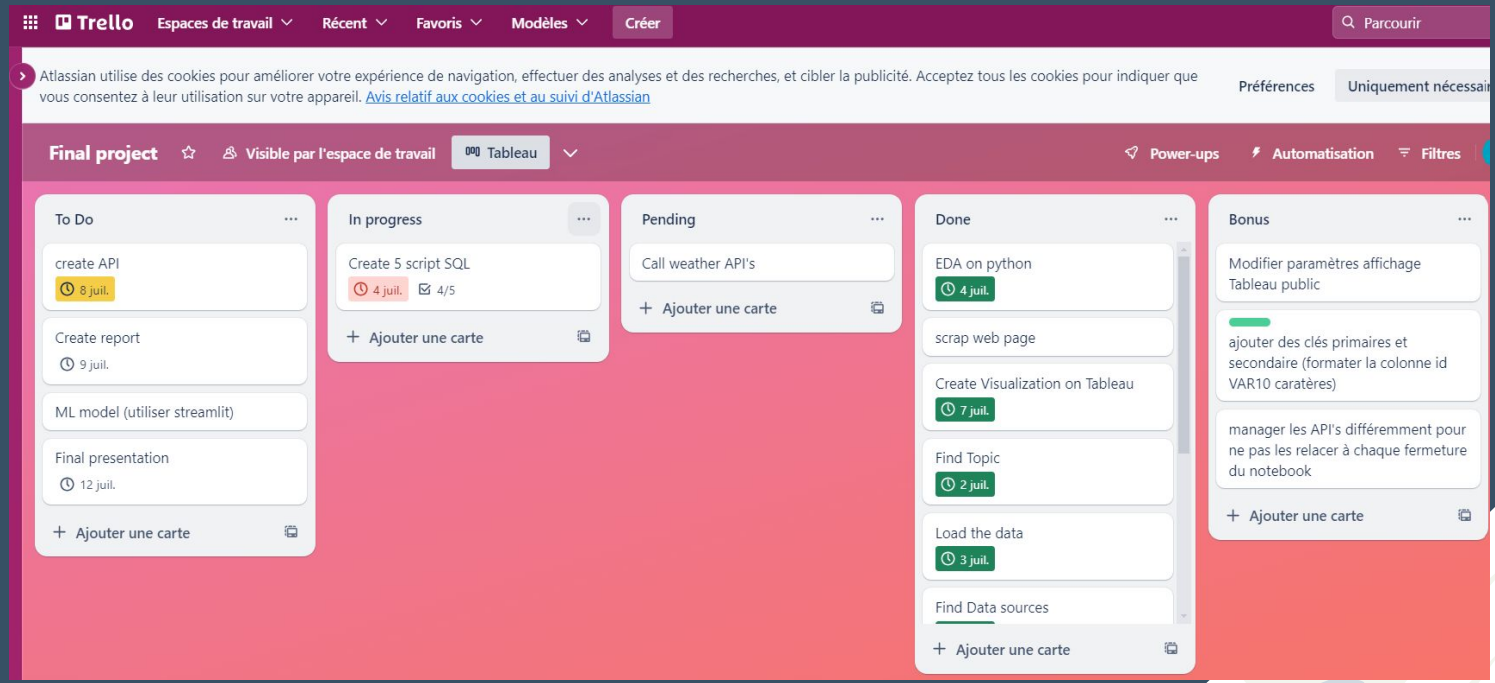
Objective

⇒ UHA has commissioned me for a freelance mission with a clear objective:

Analyze the success and risk factors of Himalayan expeditions to improve the planning and safety of future expeditions.

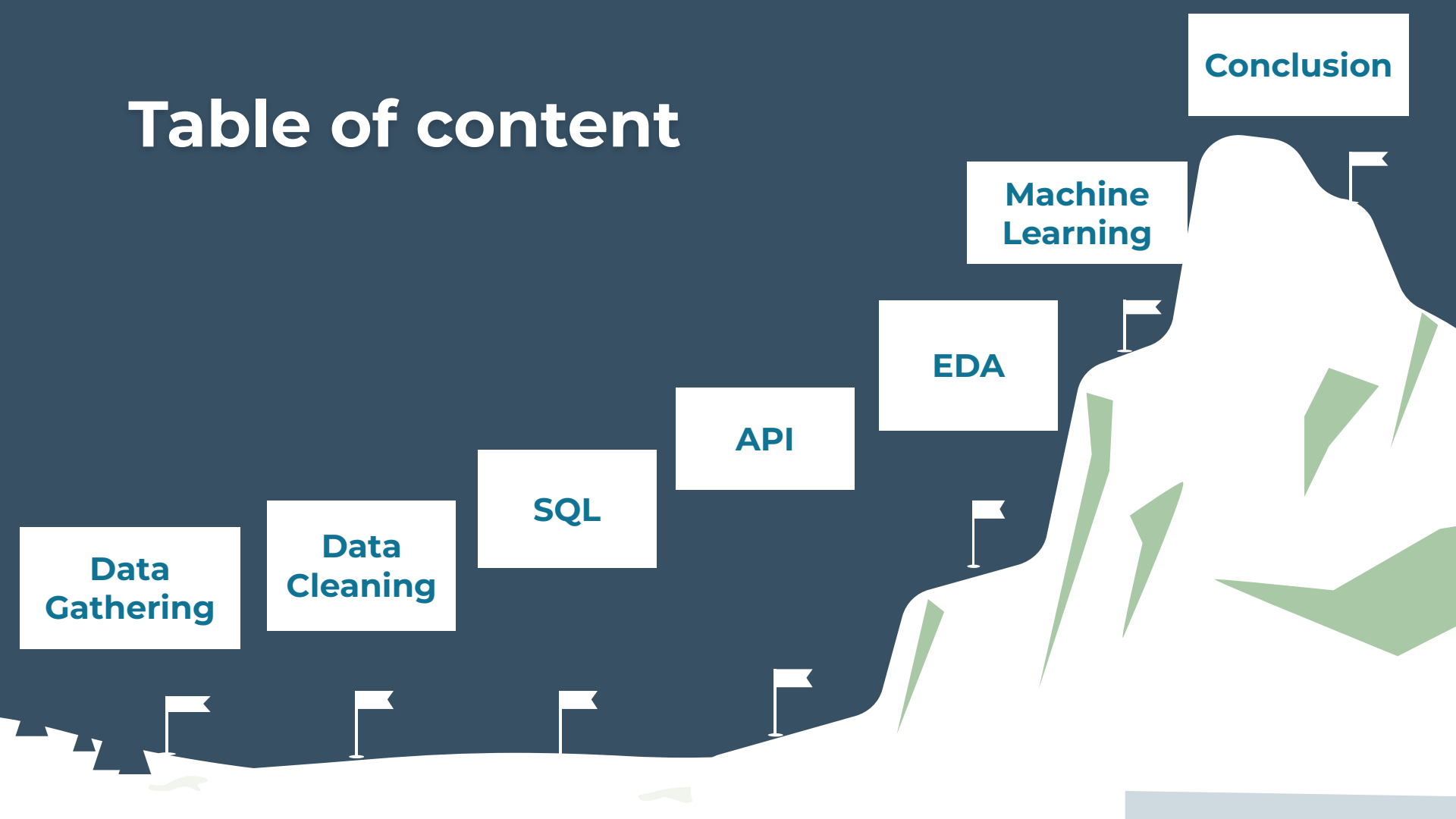
This study aims to provide crucial data for more sustainable and secure expedition management, thus preserving the beauty and integrity of these mythical peaks for future generations.

Trello : Kanban method



Steps the project stages were managed using the Kanban method and a Trello board

Table of content

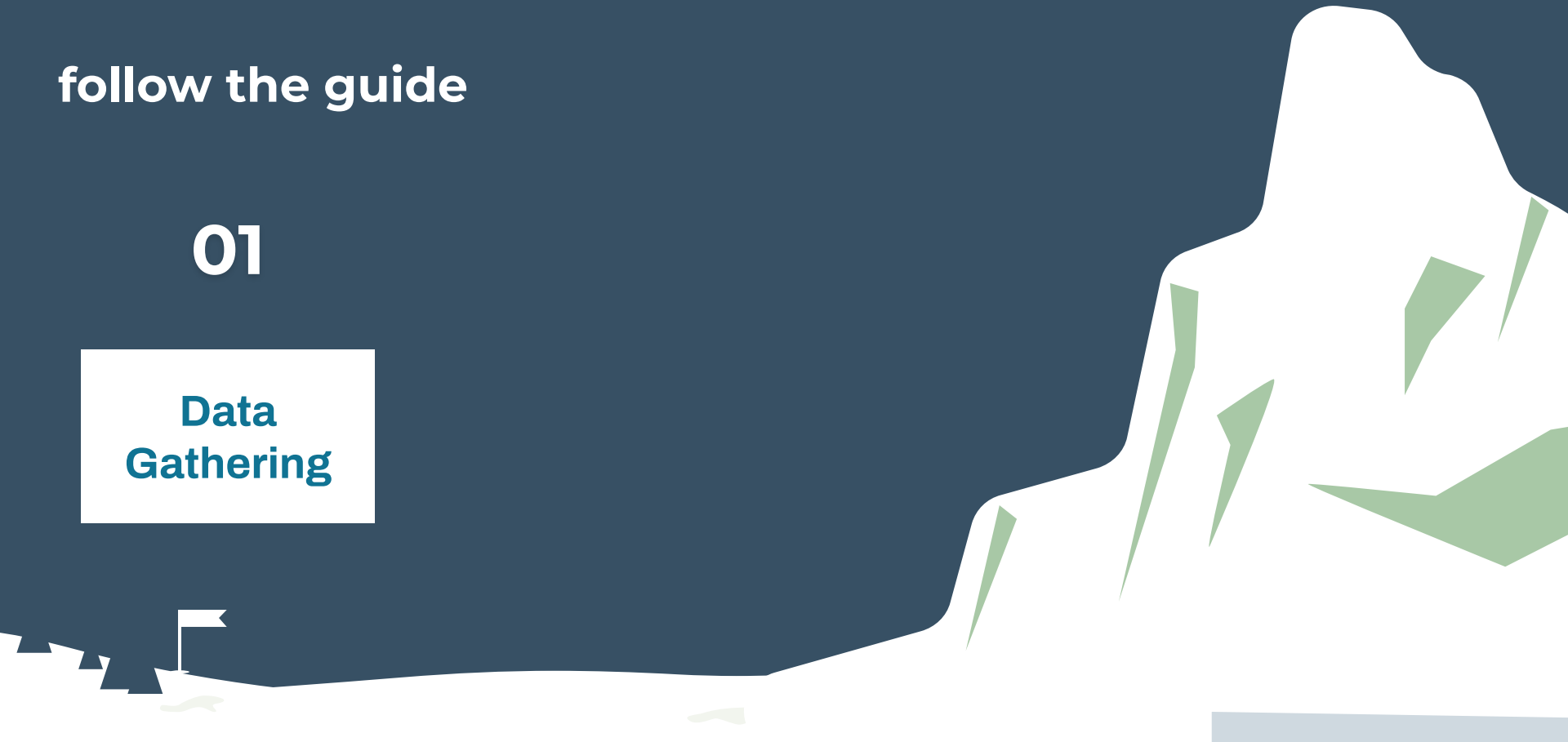


Let’s start the expedition...

follow the guide

01

Data Gathering



3 types of data sources

Flat File

The mean dataset was found on Kaggle based on the expedition archives of Elizabeth Hawley

API

Nominatim API was used to collect GPS coordinates of the peaks

Web scraping

The Topchinatravel website was used to scrape temperature and weather tables by season for Mount Everest.

02

Data Cleaning

DataFrame : expeditions_df

- Expedition_df.shape: (10364, 16)
- expedition_df.duplicated().sum() : 0
- expedition_df.head() :

expedition_id	peak_id	peak_name	year	season	basecamp_date	highpoint_date	termination_date	termination_reason	highpoint_metres	members	r
ANN260101	ANN2	Annapurna II	1960	Spring	1960-03-15	1960-05-17	NaT	Success (main peak)	7937.0	10	
ANN269301	ANN2	Annapurna II	1969	Autumn	1969-09-25	1969-10-22	1969-10-26	Success (main peak)	7937.0	10	
ANN273101	ANN2	Annapurna II	1973	Spring	1973-03-16	1973-05-06	NaT	Success (main peak)	7937.0	6	

DataFrame : expeditions_df
Managing null values

peak_name	1
basecamp_date	1095
highpoint_date	650
termination_date	2380
highpoint_metres	414
trekking_agency	1710

- Replace NaN by median for numerical column
- Replace NaN by "Unknown" for categorical column
- manage NaN basecamp_date (avg date per year + season)
- manage NaN hightpoint_date (avf diff)
- manage NaN termination_date (avg diff)
- Drop remaining NaN

DataFrame : peaks_df

- Peaks_df.shape: (468, 8)
- peaks_df.duplicated().sum() : 0
- peaks_df.isna().sum() :

```
peak_alternative_name          223
first_ascent_year              132
first_ascent_country           132
first_ascent_expedition_id     135
```

➢ Replace null value by "Unknown" for categorical column

DataFrame : members_df

- Members_df.shape : (76519, 21)
- members_df.duplicated().sum() : 0
- members_df.head() :

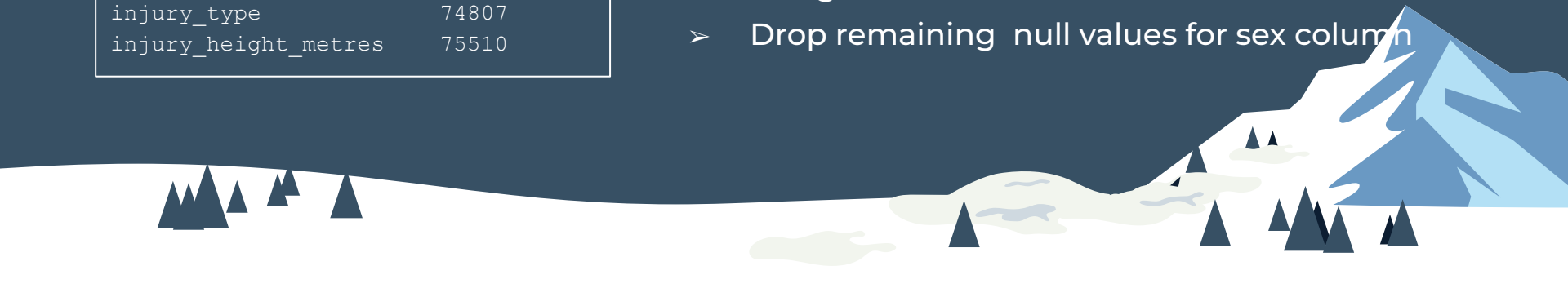
expedition_id	member_id	peak_id	peak_name	year	season	sex	age	citizenship	expedition_role	...	highpoint_metres	success	solo	oxygen_used
AMAD78301	AMAD78301-01	AMAD	Ama Dablam	1978	Autumn	M	40.0	France	Leader	...	NaN	False	False	False
AMAD78301	AMAD78301-02	AMAD	Ama Dablam	1978	Autumn	M	41.0	France	Deputy Leader	...	6000.0	False	False	False
AMAD78301	AMAD78301-03	AMAD	Ama Dablam	1978	Autumn	M	27.0	France	Climber	...	NaN	False	False	False
AMAD78301	AMAD78301-04	AMAD	Ama Dablam	1978	Autumn	M	40.0	France	Exp Doctor	...	6000.0	False	False	False

DataFrame : members_df

Managing null value

peak_name	15
sex	2
age	3497
citizenship	10
expedition_role	21
highpoint_metres	21833
death_cause	75413
death_height_metres	75451
injury_type	74807
injury_height_metres	75510

- replace missing values by 0 for 'death_height_metres' et 'injury_height_metres'
- replace missing values by median for 'age' et 'highpoint_metres'
- Replace null value by "Unknown" for other categorical column
- Drop remaining null values for sex column



API

- Using the Nominatim API to retrieve longitude and latitude based on the peak names
- Merging the newly created dataframe (containing the geographic coordinates) with the existing peak dataframe
- Approximately 50% of the coordinates were captured using the API.

Web scraping

2 tables created

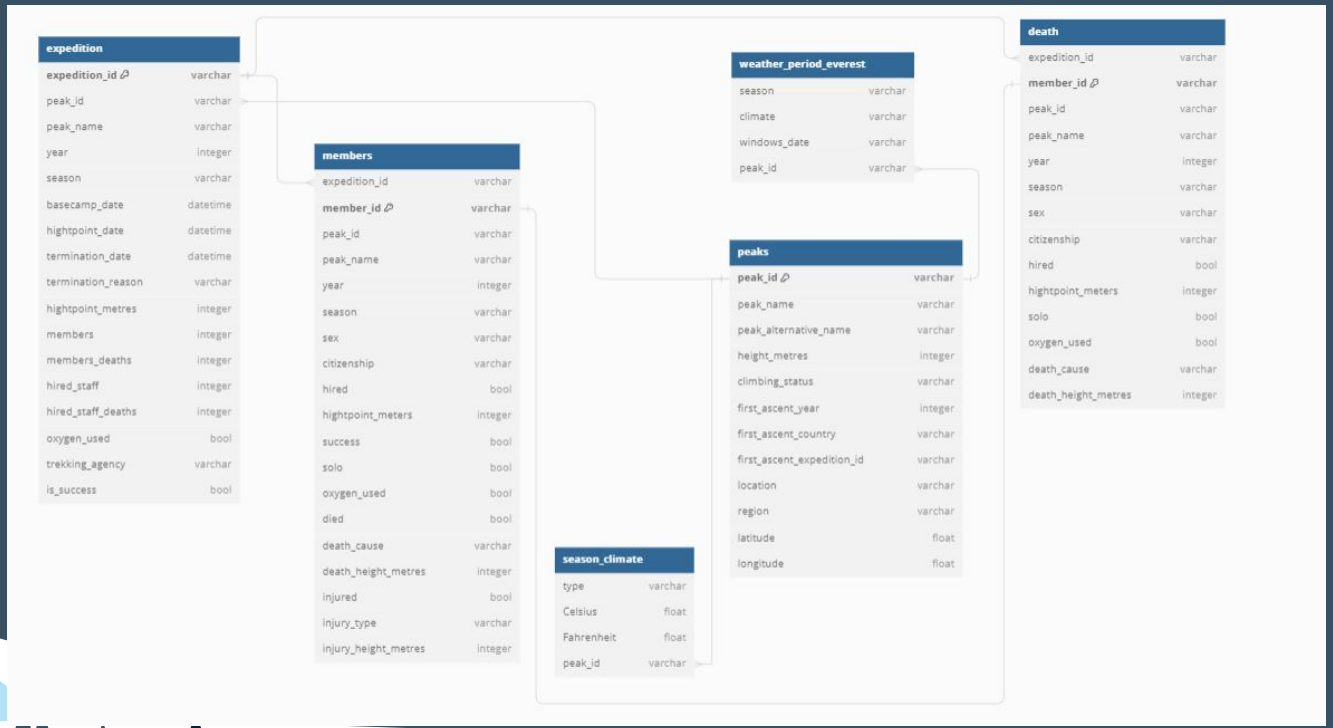
	season	climate	windows_date	peak_id
0	Summer	Very Wet	June 7 to Sep 30	EVER
1	Autumn Window	Dry, Warm, Calm	Oct 1 to Oct 20	EVER
2	Autumn	Very Windy, Cold, Very Dry, Dark	Oct 20 to Nov 30	EVER
3	Winter	Very Windy, Very Cold, Dry, Dark	Dec 1 to Feb 28	EVER
4	Spring	Windy, Cold, Dry	Mar 1 to May 20	EVER
5	Spring Window	Dry, Warm, Calm	May 20 to June 6	EVER

	Type	Celsius	Fahrenheit	peak_id
1	July	-18.0	-0.4	EVER
2	Aug	-18.0	-0.4	EVER
3	Sept	-21.0	-5.8	EVER
4	Oct	-27.0	-16.6	EVER
5	Nov	-30.0	-22.0	EVER
6	Dec	-34.0	-29.2	EVER
7	Jan	-36.0	-32.8	EVER
8	Feb	-35.0	-31.0	EVER
9	Mar	-32.0	-25.6	EVER
10	Apr	-31.0	-23.8	EVER
11	May	-25.0	-13.0	EVER
12	Jun	-20.0	-4.0	EVER

03

SQL

ERD


expedition
expedition_id varchar
peak_id varchar
peak_name varchar
year integer
season varchar
basecamp_date datetime
hightpoint_date datetime
termination_date datetime
termination_reason varchar
hightpoint_metres integer
members integer
members_deaths integer
hired_staff integer
hired_staff_deaths integer
oxygen_used bool
trekking_agency varchar
is_success bool
members
expedition_id varchar
member_id varchar
peak_id varchar
peak_name varchar
year integer
season varchar
sex varchar
citizenship varchar
hired bool
hightpoint_meters integer
success bool
solo bool
oxygen_used bool
died bool
death_cause varchar
death_height_metres integer
injured bool
injury_type varchar
injury_height_metres integer
weather_period_everest
season varchar
climate varchar
windows_date varchar
peak_id varchar
peaks
peak_id varchar
peak_name varchar
peak_alternative_name varchar
height_metres integer
climbing_status varchar
first_ascent_year integer
first_ascent_country varchar
first_ascent_expedition_id varchar
location varchar
region varchar
latitude float
longitude float
season_climate
type varchar
Celsius float
Fahrenheit float
peak_id varchar
death
expedition_id varchar
member_id varchar
peak_id varchar
peak_name varchar
year integer
season varchar
sex varchar
citizenship varchar
hired bool
hightpoint_meters integer
solo bool
oxygen_used bool
death_cause varchar
death_height_metres integer

Some queries :

Selected most popular trekking agency to expose it in the API (route : /statistics):

```sql
SELECT
harmonized_agency_name,
count(distinct expedition_id) as nb_expedition,
sum(is_success)/count(distinct expedition_id)  as success_rate
FROM himalaya.expedition
WHERE harmonized_agency_name != 'Unknown'
GROUP BY harmonized_agency_name
ORDER BY nb_expedition desc
```

harmonized_agency_name	nb_expedition	success_rate
Asian Trekking	832	0.5433
Thamserku Trekking	754	0.5159
Himalayan Guides	321	0.7227
Cosmo Treks	277	0.5307
Seven Summit Treks	274	0.6496
Monterosa Treks	230	0.4565
Cho Oyu Trekking	196	0.5408
Arun Treks	151	0.6954
Prestige Adventure	142	0.5563

Some queries :

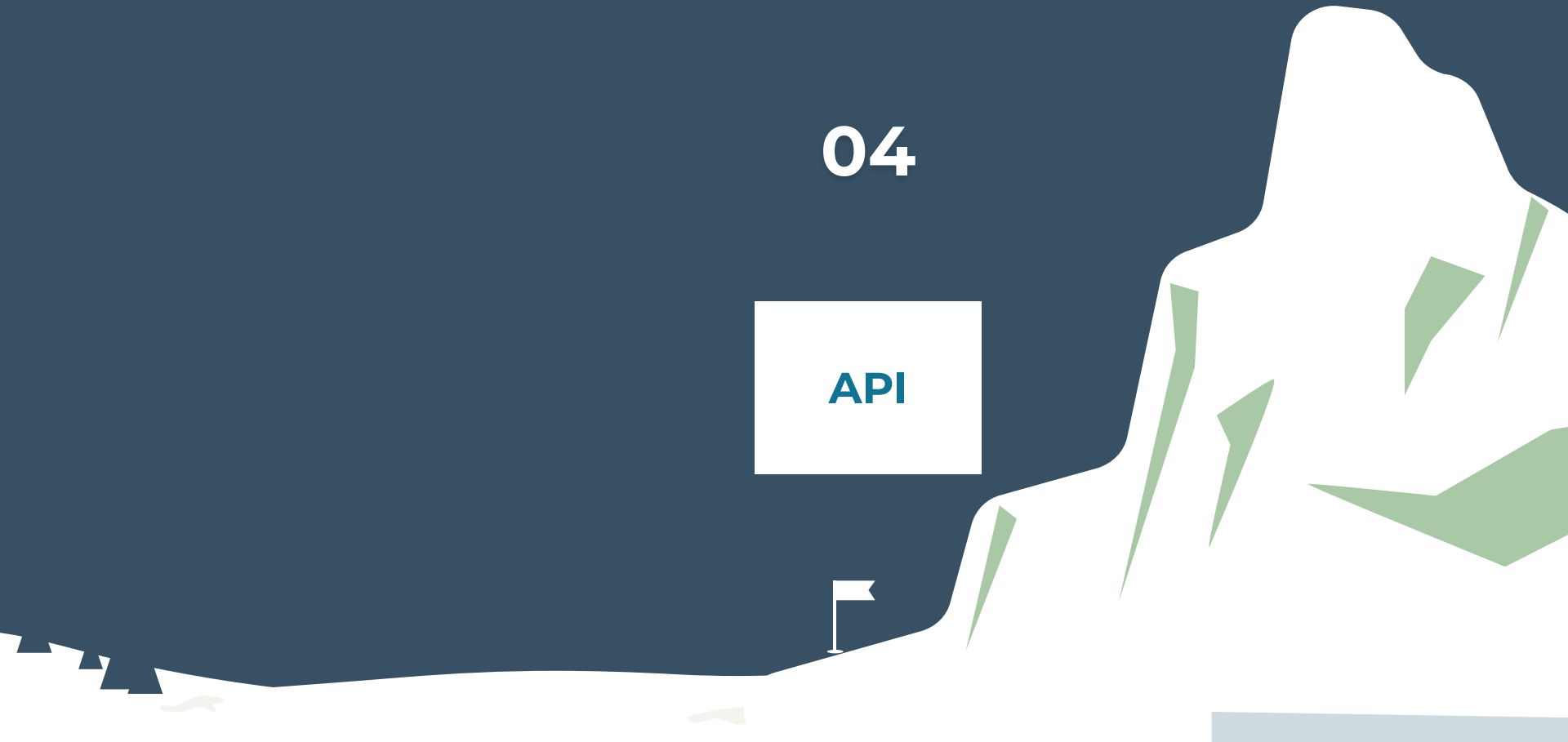
Create table with all of death detail from members table

```
Create table death
SELECT
expedition_id,
member_id,
m.peak_id,
m.peak_name,
year,
season,
sex,
age,
citizenship,
hired,
highpoint_metres,
solo,
oxygen_used,
death_cause,
death_height_metres,
p.longitude,
p.latitude
FROM himalaya.members m
LEFT JOIN himalaya.peaks p ON m.peak_id = p.peak_id
WHERE died = 1
```

expedition_id	member_id	peak_id	peak_name	year	season	sex	age	citizenship	hired	highpoint_metres	solo	oxygen_used	death_cause
AMAD79302	AMAD79302-04	AMAD	Ama Dablam	1979	Autumn	M	23	New Zealand	0	6100	0	0	Avalanche
AMAD83301	AMAD83301-01	AMAD	Ama Dablam	1983	Autumn	M	31	Switzerland	0	7400	0	0	Fall
AMAD83301	AMAD83301-13	AMAD	Ama Dablam	1983	Autumn	F	28	Switzerland	0	7400	0	0	Fall
AMAD85102	AMAD85102-03	AMAD	Ama Dablam	1985	Spring	M	32	Japan	0	6814	0	0	Fall
AMAD88102	AMAD88102-04	AMAD	Ama Dablam	1988	Spring	M	33	Canada	0	6300	0	0	Fall
AMAD92102	AMAD92102-01	AMAD	Ama Dablam	1992	Spring	M	36	Spain	0	6814	0	0	Fall
ANN170101	ANN170101-05	ANN1	Annapurna I	1970	Spring	M	32	UK	0	7315	0	0	Falling rock / ice

04

API



Flask API Development for Data Exposure

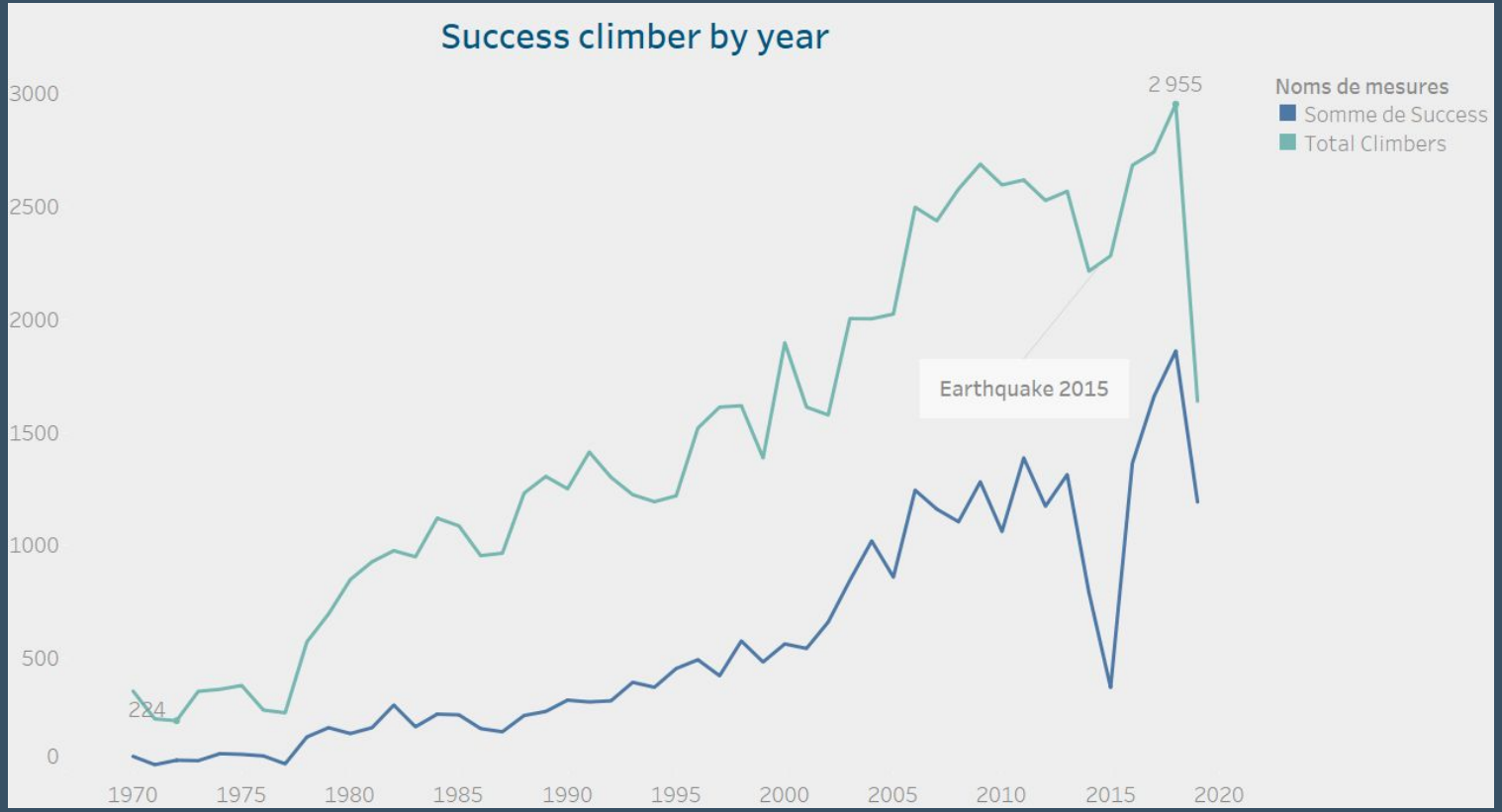
endpoints	description	parameters
/peaks	Return all the information about the peak in Himalaya	Page, page_size, include_detail, height_min
/peaks/<id>	Return all the information about the peak filtering by ID	peak_id
/expeditions	Return all the information about expeditions	Page, page_size, year

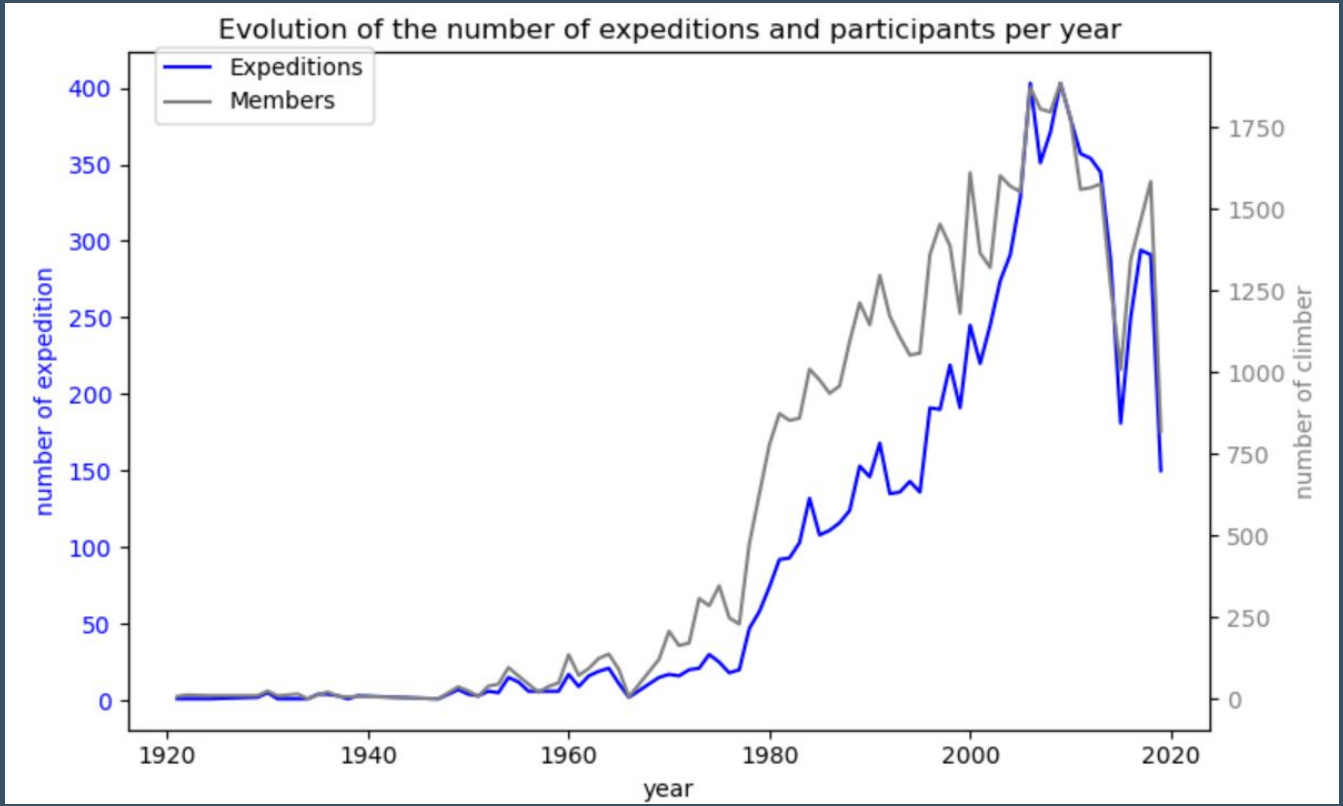
endpoints	description	parameters
/expeditions/<id>	Return information about specific expeditions	expedition_id
/statistics	Return some statistic about himalayan expedition	Start_year, end_year
/docs	Return swagger documentation	

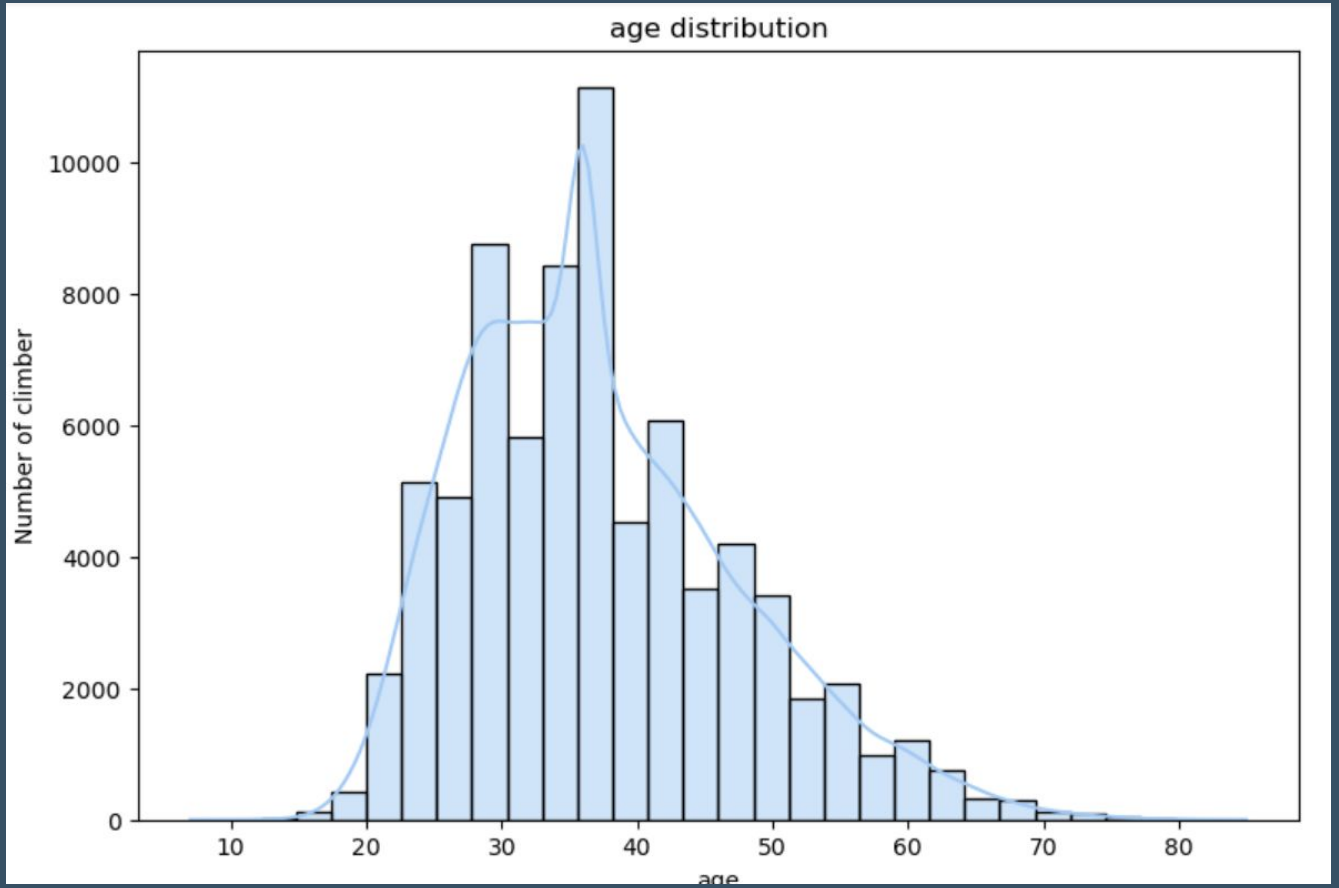
Demo : http://127.0.0.1:8080/

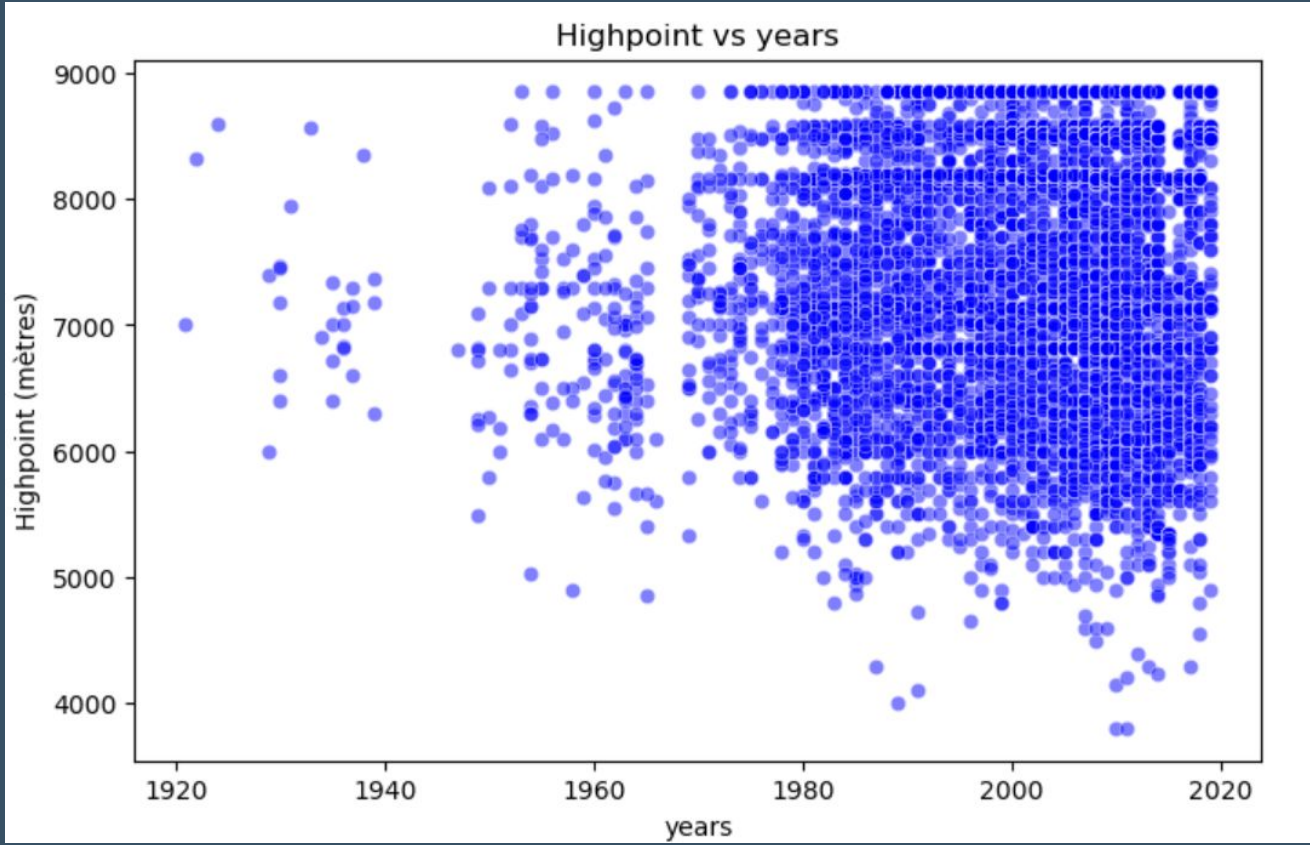
05

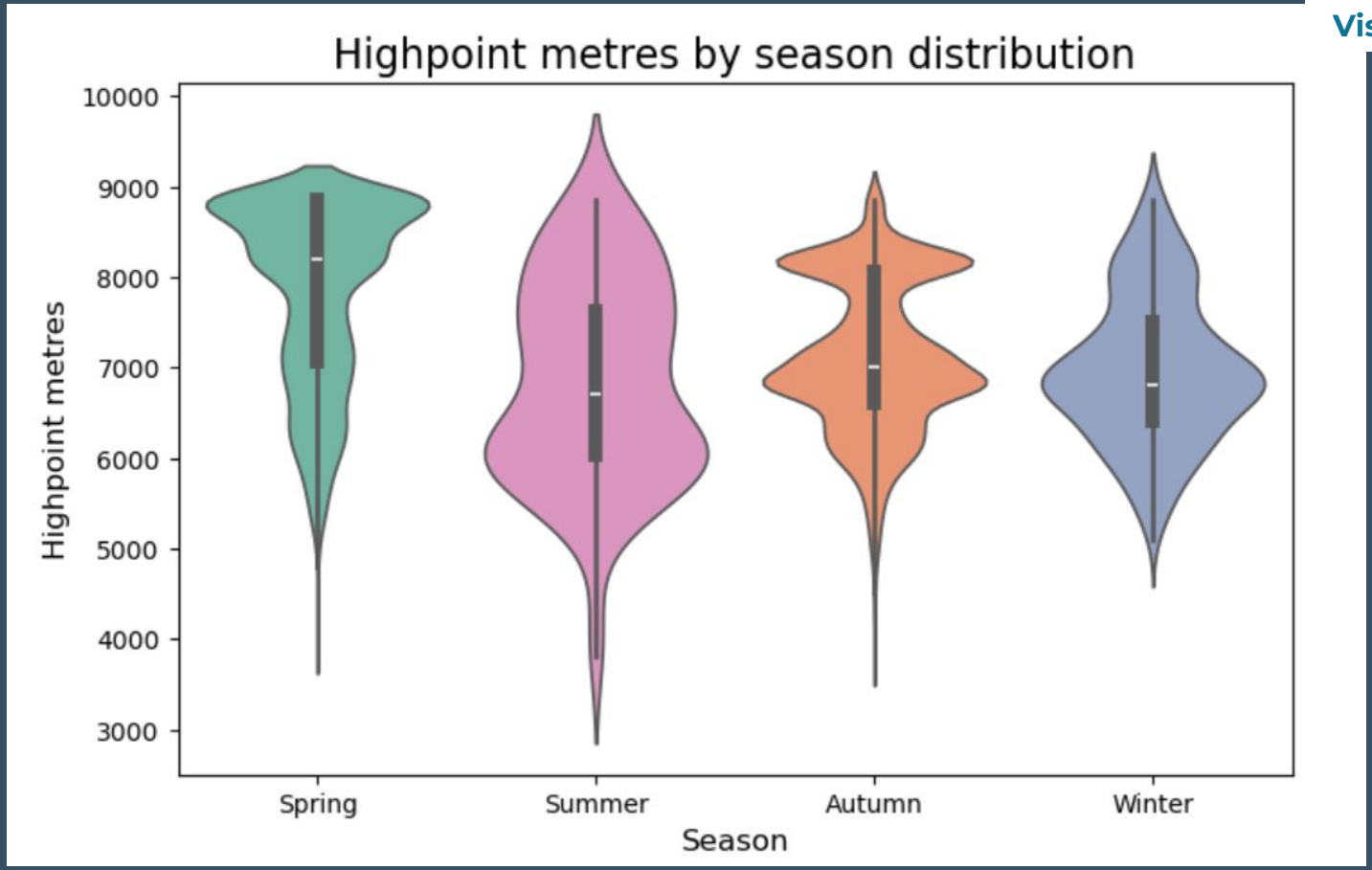
EDA & Visualization

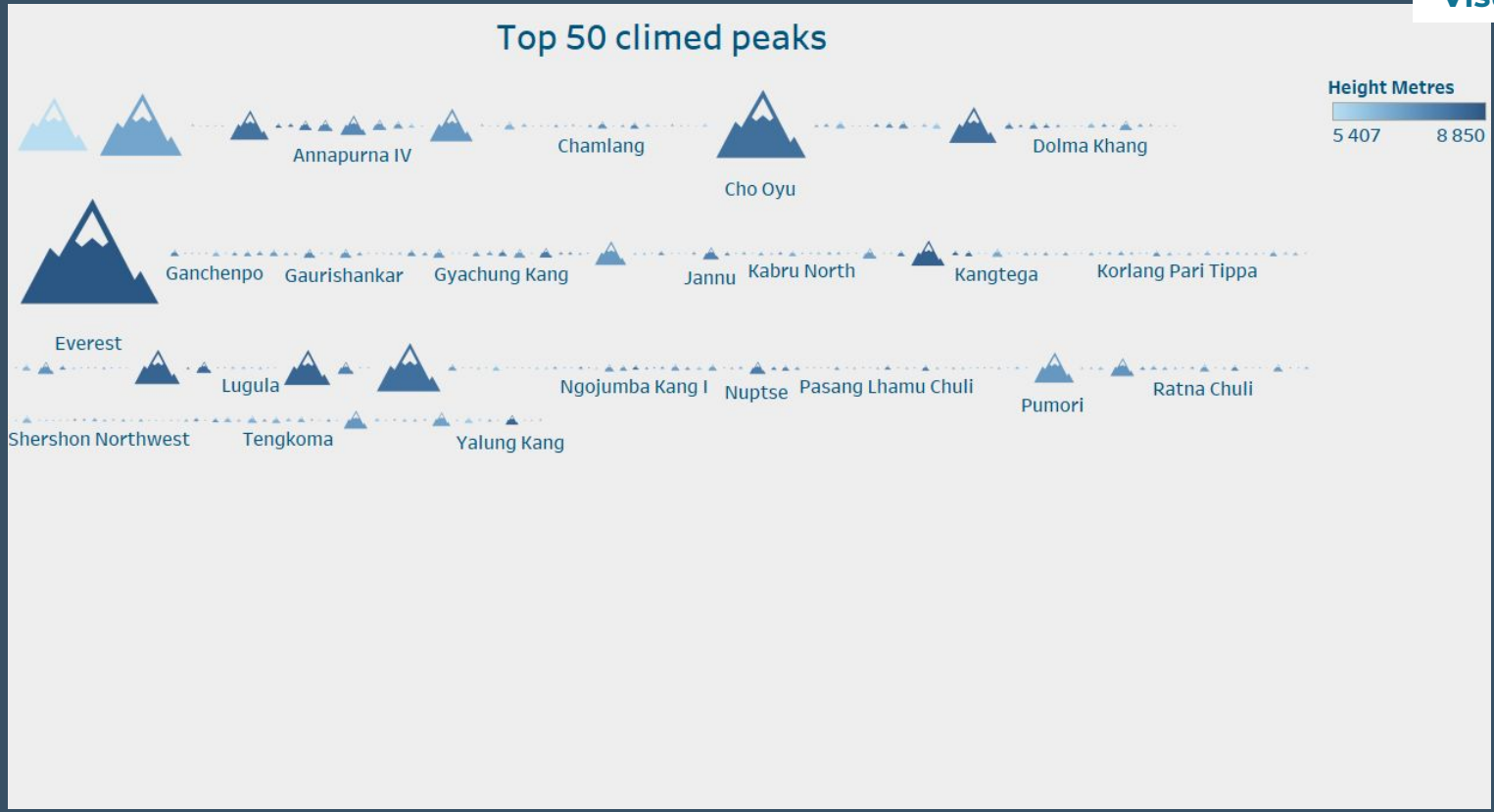

Success climber by year
3000
2500
2000
1500
1000
500
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2 955
224
Earthquake 2015
Noms de mesures
Somme de Success
Total Climbers


Evolution of the number of expeditions and participants per year
Expeditions
Members
number of expedition
number of climber
year
0
50
100
150
200
250
300
350
400
0
250
500
750
1000
1250
1500
1750
1920
1940
1960
1980
2000
2020


age distribution
Number of climber
0
2000
4000
6000
8000
10000
10
20
30
40
50
60
70
80
age


Highpoint vs years
Highpoint (mètres)
9000
8000
7000
6000
5000
4000
1920
1940
1960
1980
2000
2020
years

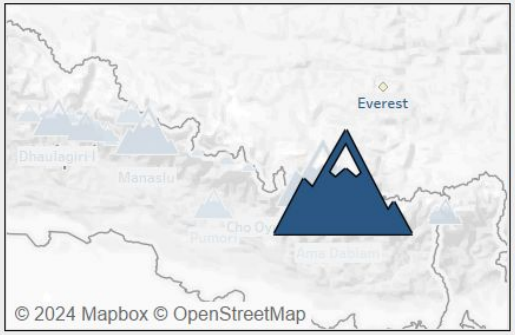

Highpoint metres by season distribution
Highpoint metres
10000
9000
8000
7000
6000
5000
4000
3000
Spring
Summer
Autumn
Winter
Season


Top 50 climed peaks
Height Metres
5 407
8 850
Annapurna IV
Chamlang
Cho Oyu
Dolma Khang
Ganchenpo
Gaurishankar
Gyachung Kang
Jannu
Kabru North
Kangtega
Korlang Pari Tippa
Everest
Lugula
Ngojumba Kang I
Nuptse
Pasang Lhamu Chuli
Pumori
Ratna Chuli
Shershon Northwest
Tengkoma
Yalung Kang

The Himalayan peak detail:

This dashboard, based on the Himalaya dataset, will allow us to dive into the risk and success factors of **Himalayan expeditions**.

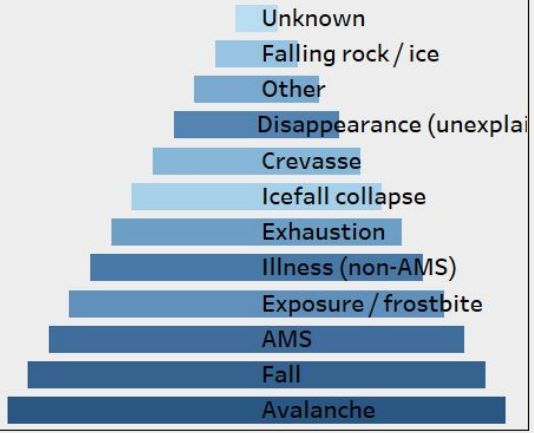
Select a mountain in the filter below or click directly on the desired peak on the adjacent map



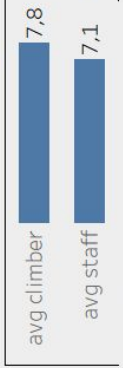
Please select Peak Name
Tout

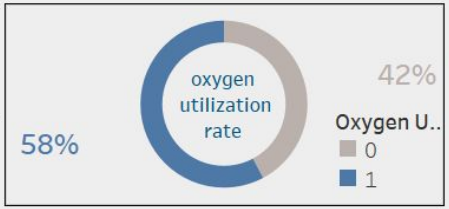
Everest	Climber	success_rate	avg highpoint	Dead climber	Death_rate	staff_rate
8 850m	20 622	46%	8 134	287	1,4%	83%

Deaths causes rank (asc)

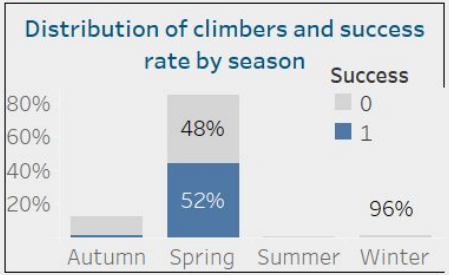


Team size (success)





Distribution of climbers and success rate by season

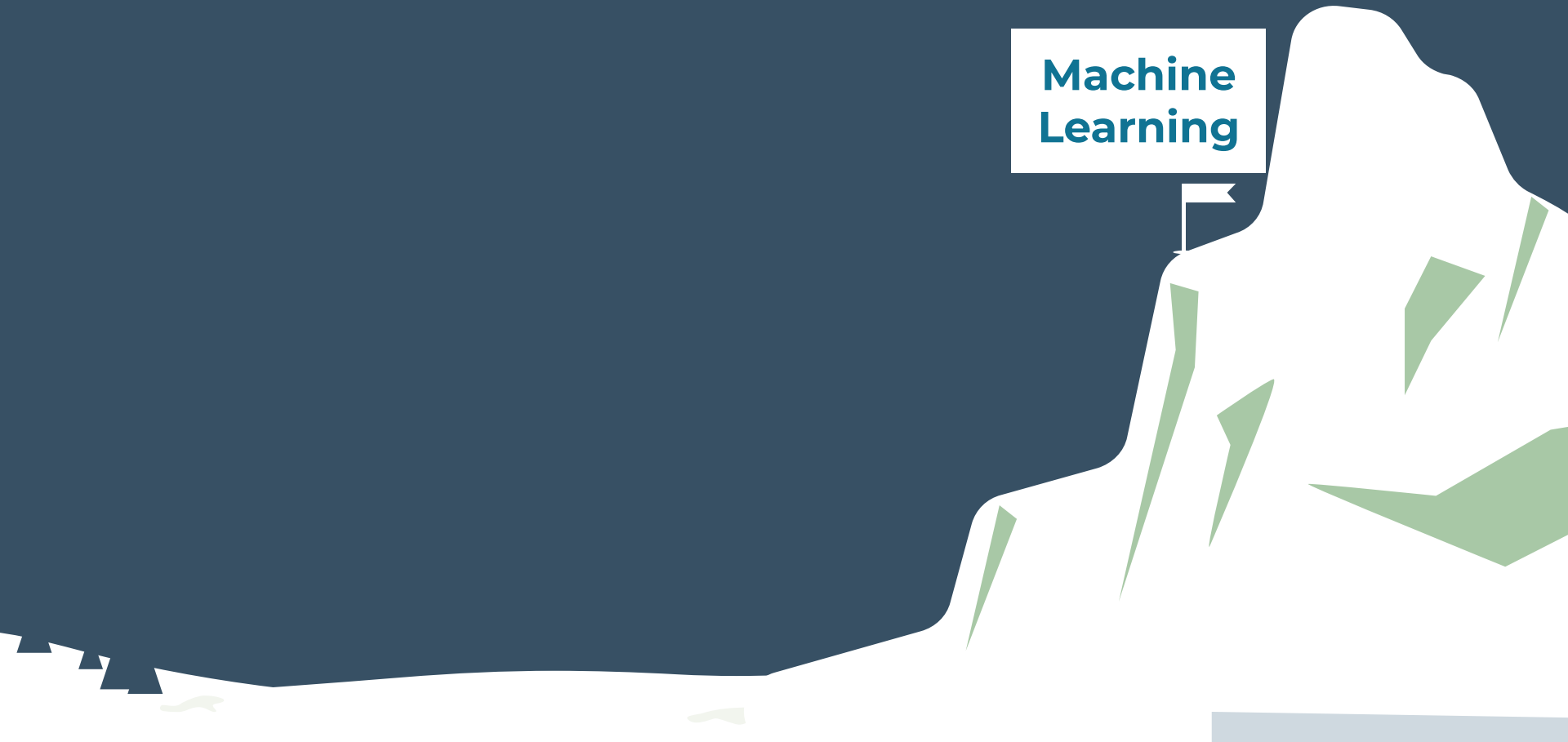


Gender distribution and avg age

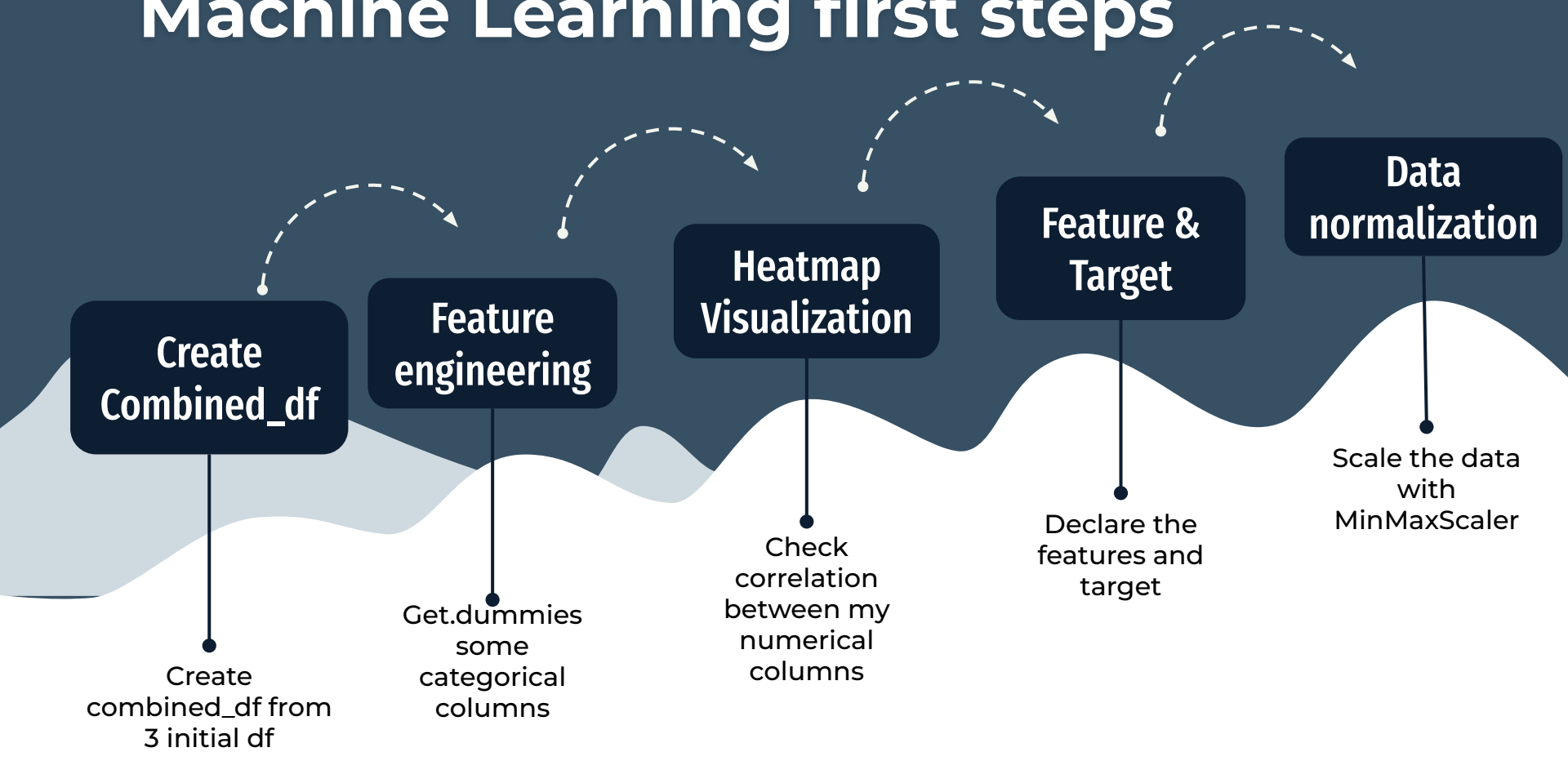
8%	92%
37	37

Turn on your oxygen for the last steps

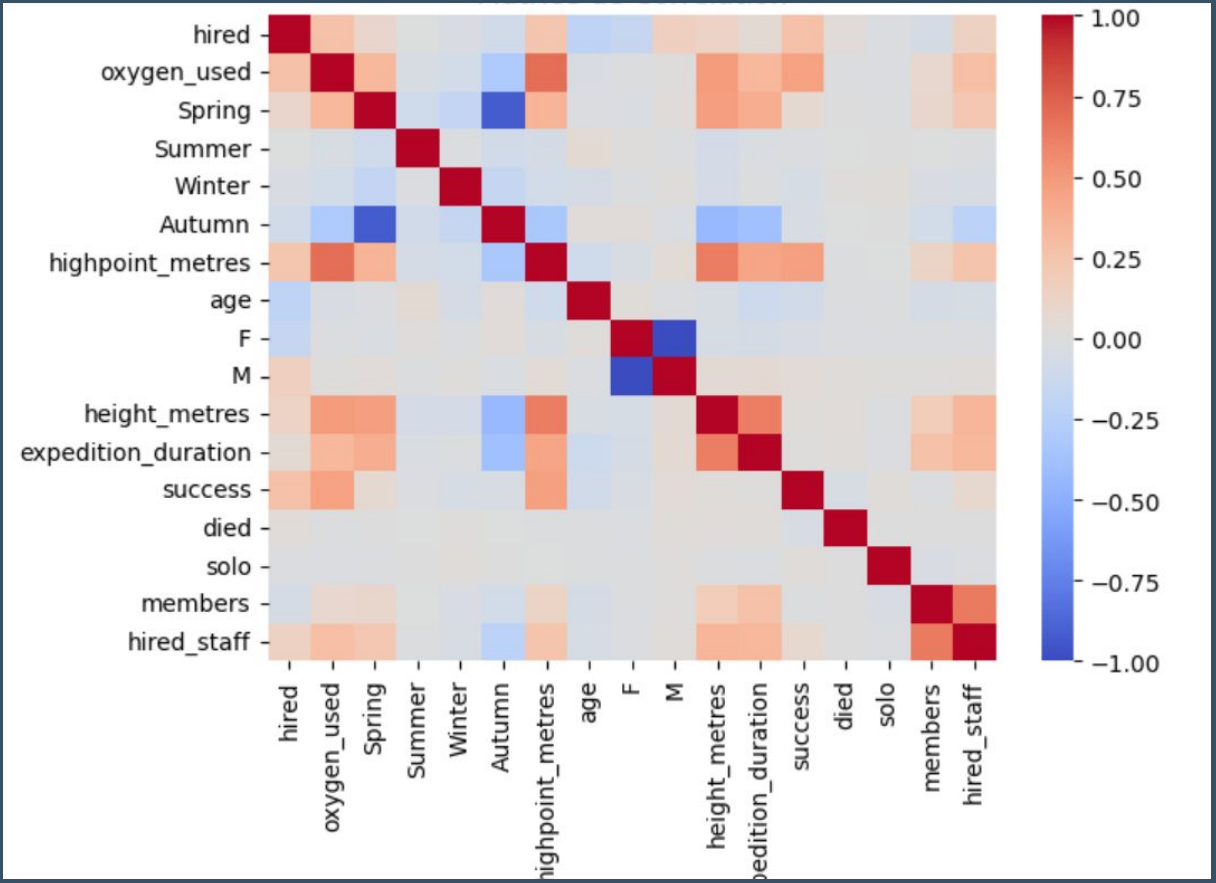
06



Machine Learning first steps



Heatmap correlation



Feature & Target declaration

```
features = combined_df[['hired', 'oxygen_used', 'Spring', 'Summer',
'Winter', 'Autumn', 'age', 'F', 'M', 'height_metres', 'solo',
'members', 'hired_staff', 'expedition_duration']]


target = combined_df['success']
```

I tested various combinations of features and this one yielded the best results

Modeling :

- I developed a function to test multiple machine learning models with different hyperparameters
- The function evaluates each model using normalized data and records performance scores (accuracy, precision, recall, F1-score) for each tested model
- The models tested include Random Forest, AdaBoost, Gradient Boosting, and Bagging Classifier

Evaluation

- For each model, hyperparameters were adjusted to maximize performance scores
- The evaluation results were compiled into a DataFrame for easy comparison and in-depth analysis

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest (100, 20, 2, 1, auto)	0.809592	0.776824	0.706900	0.740214
1	Random Forest (150, 25, 5, 2, sqrt)	0.811876	0.777823	0.713595	0.744326
2	Random Forest (200, 15, 10, 4, log2)	0.792321	0.759302	0.671750	0.712848
3	Bagging Classifier	0.780759	0.741665	0.657802	0.697221
4	Ada Boost	0.789252	0.733616	0.707830	0.720492
5	Gradient Boosting	0.800171	0.741337	0.736098	0.738708

Streamlit : the Himalayan Expedition Success Prediction Application

To help the UHA anticipate risks, we developed a Streamlit application based on our model. This app predicts the success or failure of expeditions using specific parameters

Key Features:

- Data-Driven Predictions
- User-Friendly Interface
- Real-Time Analysis

Benefits:

- Helps in planning safer expeditions
- Optimizes the use of resources
- Assists in minimizing the ecological footprint

Demo: http://localhost:8501/

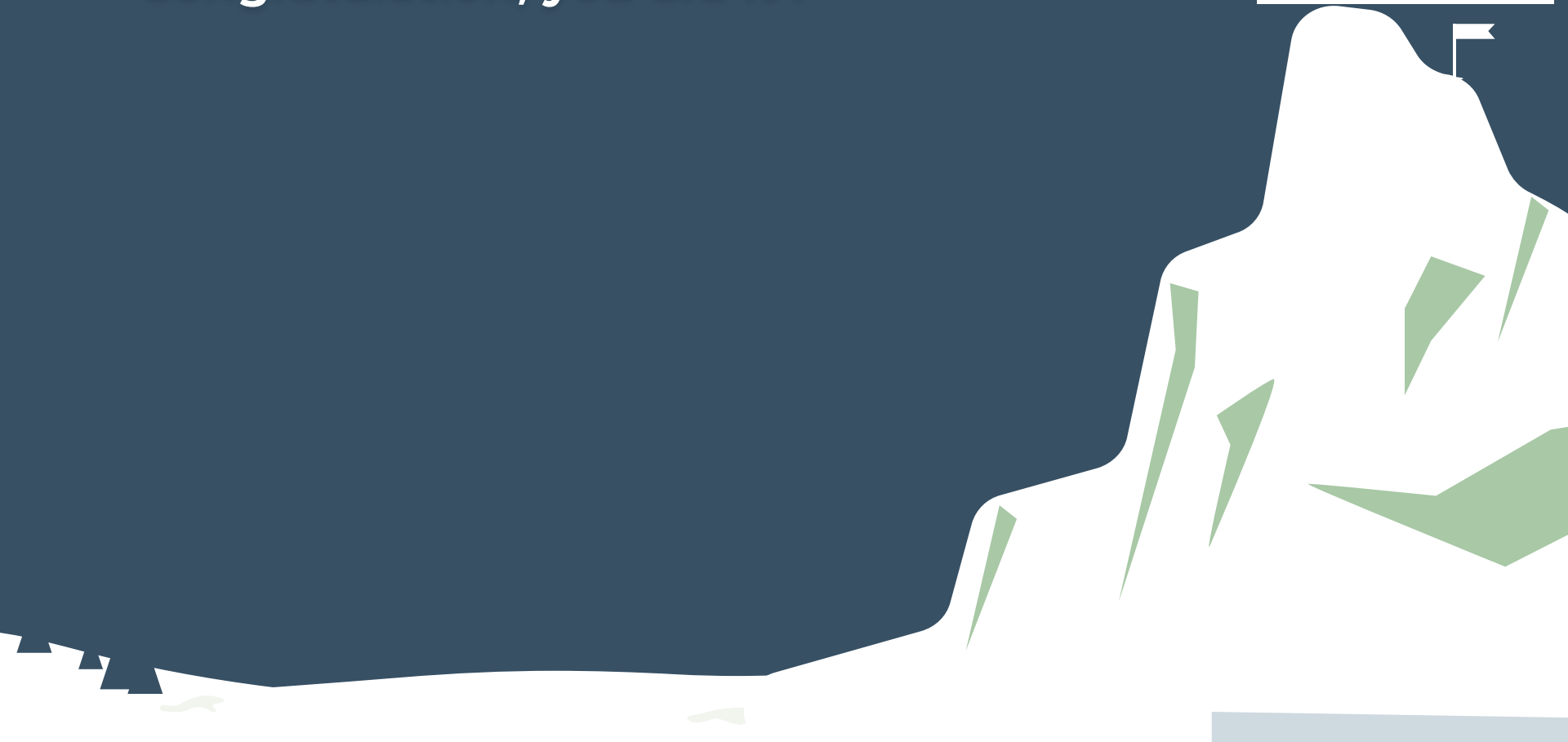
Challenge :

- Inability to retrieve historical weather data in the Himalayas via an API
- Difficulty in creating a high-performing model in the Streamlit
- Difficulty managing my time with the amount of work

07

Conclusion

Congratulation, you did it !



Conclusion

1- Risk and Evolution:

- Significant increase in the number of climbers and the use of oxygen leading to a multiplication of risk factors on high peaks.

2- Contribution of the Project:

- quipping the Union of Himalayan Agencies with advanced predictive tools to enhance sustainable and safe expedition management.

3- Vision for the Future:

- Maintaining access to the highest peaks responsibly for an authentic Himalayan mountaineering experience.
- Ensuring an exceptional experience for future generations.
- Reducing ecological footprint while maximizing climber safety.

- 2016 -

UNION OF HIMALAYAN AGENCIES

Thank You

Any questions ?