Responsible Expeditions

Analysis of Himalayan expeditions for new travel agency

BACKGROUND

ANALYSIS

RECOMMENDATIONS

CONTEXT AND OBJECTIVE

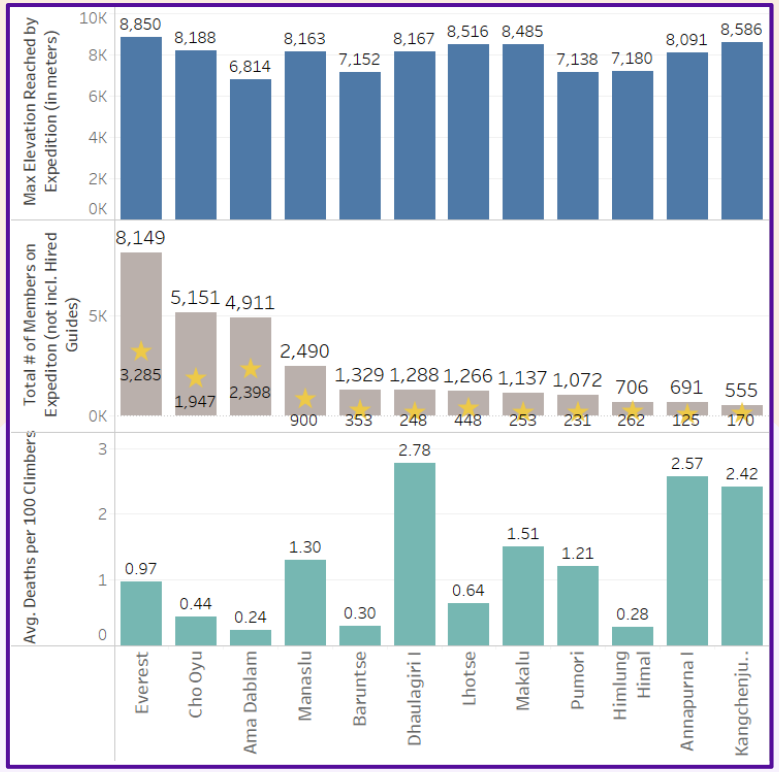
- Responsible Expeditions is a fictitious, new tourism agency focused on offering expedition tours through the Himalayas. The agency's top priorities are (1) offering expertise for climbing expeditions, (2) taking care of the peaks by understanding the amount of expedition traffic each season, and (3) employing local guides to support the local economy.
- As a data analyst for Responsible Expeditions, I've created a storyboard to give the team historical information about expeditions to help them with their staffing strategy and season planning strategy.
- Link to Project Brief

DATA SET INFO	• 2 datasets were sourced from Kaggle – expeditions.csv (each record represents an expedition; dataset includes all recorded expeditions from 1905 - 2020) and summiters.csv (each record represents a climber who reached the summit of the peak) Source: The Himalayan Database – The Expedition Archives of Elizabeth Hawley
TOOLS USED	Python Excel Tableau
SKILLS USED	• Sourcing open data for exploring variable relationships and analyzing time series data • Creating geographical visualizations and conducting geospatial analysis with Python • Conducting regression (supervised machine learning) and clustering (unsupervised machine learning) analysis in Python • Defining use-case and creating dashboard in Tableau based on Python analysis findings

BACKGROUND

ANALYSIS

RECOMMENDATIONS



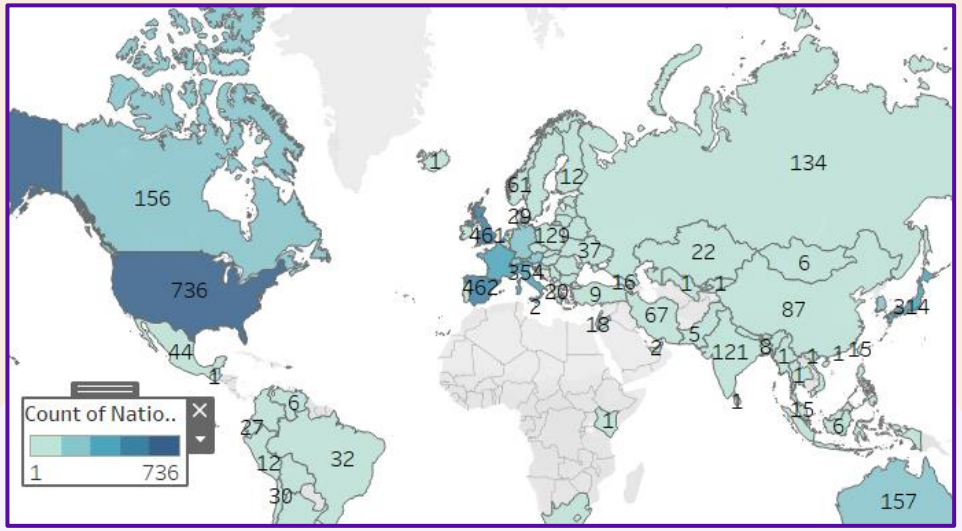
For our analysis, we're interested in expeditions that take place on **the tallest, most popular, or most dangerous peaks.** By filtering our records to focus on the 12 peaks that fall into these three categories, we're analyzing 84% of our 1990 - 2019 data (or **5,766 expeditions**).

Month of Date	Avg. Summits	Std. dev. of Summits
May	425.8	321.5
October	197.1	73.2
September	113.1	114.4
November	102.8	77.1
April	19.9	14.9
December	11.6	7.6
June	7.1	22.3
August	3.1	7.6
March	2.9	6.4
January	2.0	4.2
February	1.0	1.9
July	1.0	3.1

Looking at the average number of expeditions by month for the last 30 years, we can see that

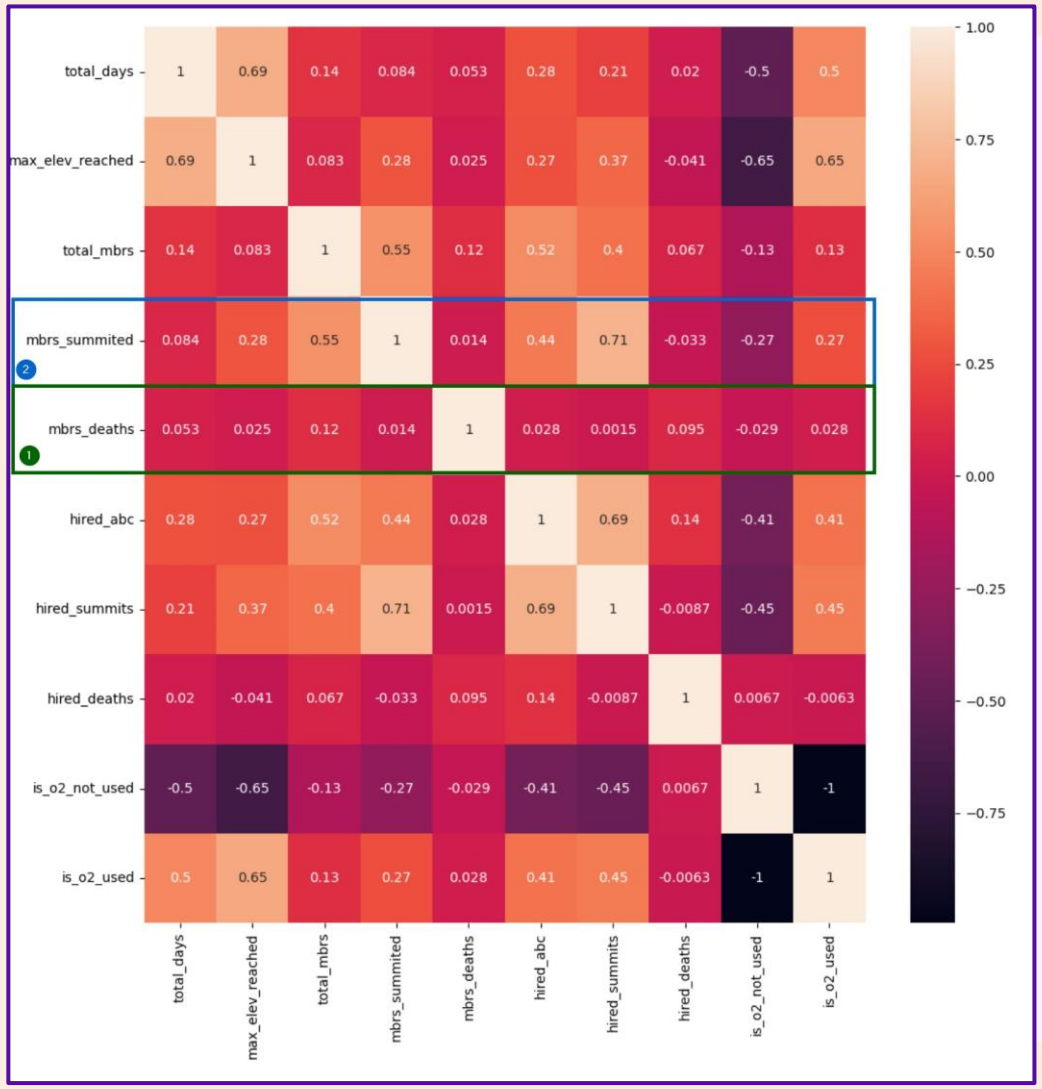
May, October, and September are the **busiest months**, while

July, February, and January are the **least busy** of months.



Our geospatial analysis shows us the following **nationalities** as having the **most expeditions** over the past 30 years:

1. United States
2. Spain
3. United Kingdom
4. Italy
5. France



First, we will look to see **how correlated our variables are** using the Coefficient Correlation Heatmap to the left.

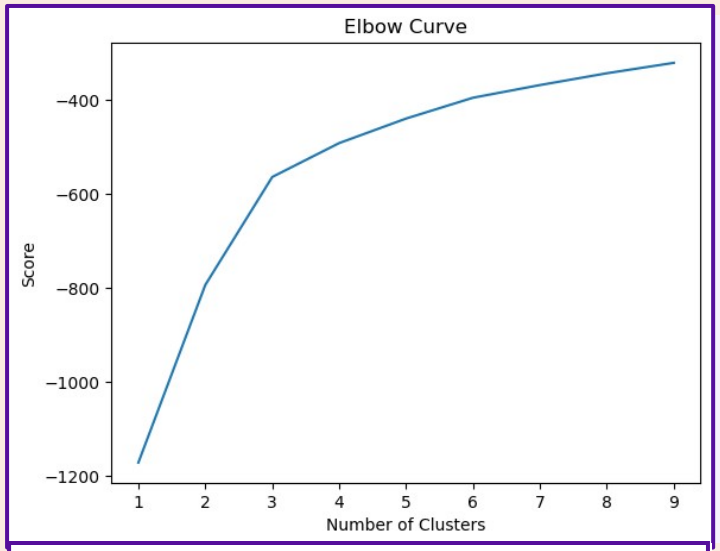
The correlation coefficient values tell us if there is:

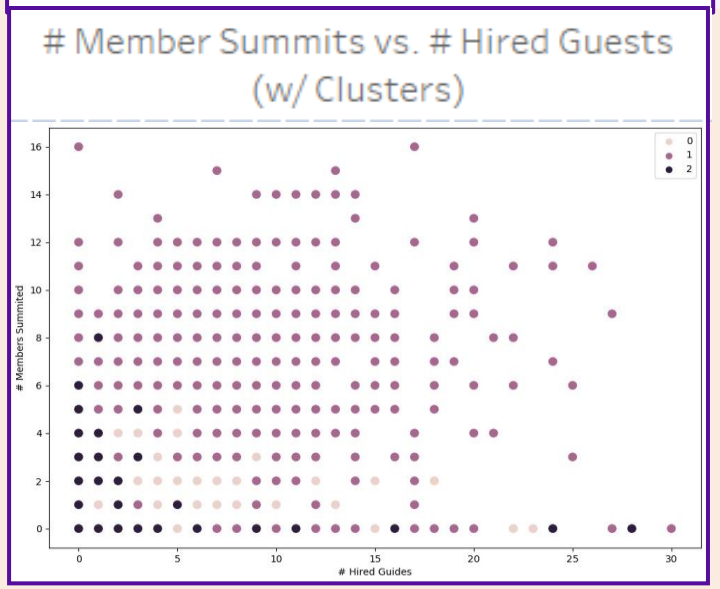
- A **strong relationship** between the two variables (0.5 - 1.0)
- A **moderately strong relationship** between the two variables (0.3 - 0.5)
- **No relationship** between the two variables (0.0 – 0.2)

Key Insights

1. There is no relationship between our member deaths variable and any other variable in our dataset. (All values for this line are <0.1)
2. We should look more into how the number of hired guides impact the number of members that summit. (Our coefficient for this relationship is 0.44 and is moderately strong).

ANALYSIS





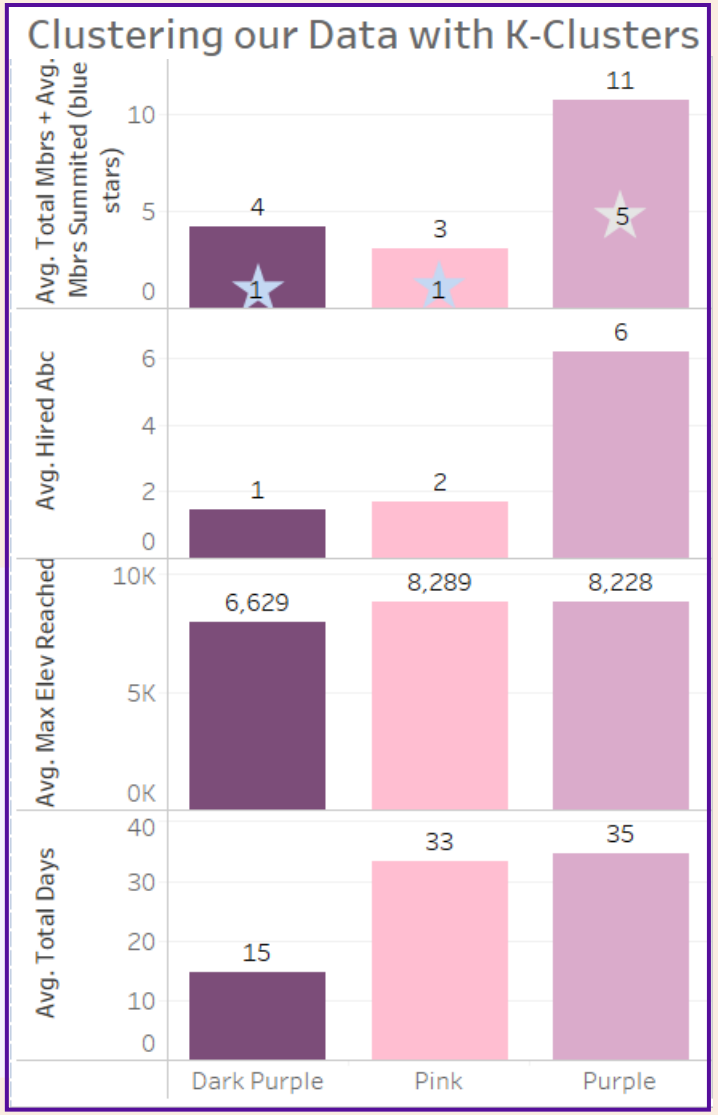
The curviness of our **Elbow Curve** and the tail continuing to climb instead of flattening out tells us our data may not be fit for the k-clustering method.

If we end up clustering our groups using 3 clusters (based on the 'bend' we see in the Elbow Curve), we can see what our groups look like (to the right).

Using the variable relationship **"# Members Summitted" vs. "# Hired Guides"**, we can see that our clusters are not very dense suggesting there's not a strong relationship between our points within each cluster.

The 3:2 (pink) and 6:11 (purple) groups reached the highest elevation on average. Our 4:1 (dark purple group) may be representing groups that stopped the expedition midway through (hence the lower max elev. reached and total days).

Our data suggests there **isn't an ideal ratio of hired guides to members based on this dataset**. We should try other clustering methods to better fit the shape of our data or investigate additional variables.



RECOMMENDATIONS

We should make sure we have **adequate staffing on the peaks of interest** outlined during the busier seasons. During the off-season, we can support our staff by providing training in expedition and environmental best practices and offer language lessons for the traveling nationalities we see most.

One limitation of our data is that it ends in 2019.
We should **obtain more recent data** to better understand the popularity of Himalayan expedition.
We can also look at the number of permits issued by the Nepalese government as an additional data point.

There are no variables in our dataset that are strongly correlated to the number of deaths that occur on an expedition. The clustering of our data set shows us that there isn't a sure way to generalize how many guides should be involved to influence the number of members who summit.

Not knowing a sure way to successfully summit and prevent accidents may be the nature of this extreme mountain range! However, we can **continue to look at other variables** (such as accident reasons, fatality reasons, and weather conditions) and **other analysis methods** (such as other methods in machine learning) to see what else we can learn!

For an in-depth look at this project, please visit the following links:

Tableau Storyboard

GitHub Repository