Introduction

Even though many climbers lost their lives on Everest expeditions, the statistic has been growing over time. Therefore, I have investigated safety issues, such as mortality rates and the effects of the host, in addition to revealing key historical trends in Everest climbing, such as variations in the number of climbers, the seasons, and demographics like age and gender. This study's dataset, which covers Everest expeditions from 1953 until 2020, was obtained via Kaggle [1]. With a better understanding of Everest expeditions, climbers, organizations, and researchers will be able to plan safer and more productive ascents.

After an extensive exploration of diverse modelling techniques—from linear and nonlinear regression to sophisticated methods like general additive models (GAM), tree-based approaches, and logistic regression—I've selected two primary models that best align with my dataset: the general additive model and logistic regression model. These choices are grounded in their superior compatibility with the dataset's characteristics and their exceptional predictive capacity for forecasting future expeditions. In addition, I have utilized a variety of statistical techniques, including means, regression analyses, correlation studies, variance assessments, and more. I hope to provide a thorough examination of expedition data over a range of time periods by utilizing these statistical tools and prediction models. This is not only a retrospective analysis, but it also provides important estimates for various variables related to Everest expeditions. In light of upcoming Everest expeditions, these projections offer insightful information that will facilitate improved strategic planning and decision-making.

I started researching Everest trips and discovered a fascinating—and sometimes depressing—story. Although the numbers showed a steady increase in the number of trips over time, they also highlighted the tragic deaths of climbers on these journeys. This led me to investigate the nuances of safety in more detail, paying close attention to death rates and the unique impact of host nations on these missions. I investigated the patterns that shape these expeditions, like the different climbers' participation rates and the seasons that have significant effects on these challenging climbs. Examining demographic factors such as the age and gender distribution of climbers provided additional perspectives on the dynamics of the expedition.

With the intention of guiding climbers, organizations, and researchers alike, it aimed to offer a comprehensive overview of Everest expeditions. This more profound understanding may provide useful direction for future ascents to Everest, making ascents safer and more effective.

Introduction To Dataset

I am going to use a dataset available on Kaggle, a web source for datasets used in analysis [1]. The dataset contains a comprehensive record of Everest expeditions from 1953 to 2020, encompassing a total of 10,184 entries. Each entry within this dataset is associated with various columns that encapsulate valuable information about the expeditions.

Name: This column encompasses the names or identities of the individuals who participated in the Everest expeditions.

Yr/Seas: Represents the year and season of the expedition, providing a temporal reference for the event. Example. 2013 spr.
Date: Indicates the specific date when the expedition took place, offering a detailed timeline for the activity.
Time: Specifies the time of the expedition, potentially providing information about the time of the day.
Citizenship: Indicates the nationality or citizenship of the individual undertaking the expedition.
Sex: Represents the gender of the individual involved in the expedition, distinguishing between male and female.
Age: Provides the age of the individual at the time of the expedition, offering demographic information.
Oxy: Likely represents the usage or availability of oxygen during the expedition, categorizing individuals based on oxygen usage (possibly 'Yes' or 'No').
Dth: Indicates mortality or death associated with the expedition, potentially categorizing incidents as 'Yes' or 'No' or possibly providing details related to fatalities.
Host: Represents the hosting organization or entity responsible for organizing or facilitating the expedition, potentially providing insights into the organizing body or expedition service provider.

This dataset serves as a comprehensive repository, offering rich insights into the multifaceted aspects of Everest expeditions, ranging from temporal trends and demographic profiles to safety measures.

Project Objectives

The primary aim of this project is to conduct a comprehensive analysis of Everest expedition data through the application of various statistical learning methodologies. By leveraging diverse statistical techniques, this analysis aims to extract invaluable insights from historical expedition records. The overarching goal is to not only gain a profound understanding of past expedition trends but also to utilize this knowledge for predictive purposes. In addition, this project makes projections about future expedition patterns based on historical data analysis. The goal is to provide accurate projections and insights into possible patterns or shifts in the dynamics of the Everest expedition by utilizing predictive models that are based on strong statistical methodologies. In the end, this project hopes to combine statistical modelling and data-driven insights to offer insightful predictions about how Everest expeditions will develop going forward.

This project aims the following category of the data analysis:

1. Climbing Trends Over Time

I aim to study the historical data to understand the fluctuations and trends in the number of climbers over different periods. This involves investigating whether there has been a steady increase, decrease, or any significant changes in climbing activity over time. Also, this involves identifying any seasonal patterns in climbing activity. Understanding how climbing trends vary throughout the year could provide insights into preferred climbing seasons and potential risks associated with specific times.

2. Demographic Analysis:

Age and Gender Distribution: We'll explore the demographic makeup of climbers, looking at age and gender distributions over time. This analysis helps in understanding the demographics of those participating in Everest expeditions. By correlating climbers' demographics (age, nationality) with their success rates in reaching the summit, we aim to identify patterns suggesting higher success rates among certain groups.

Nationality and Region Trends: Investigating climbers' nationalities or regions aims to uncover any shifts or consistency in the representation of different countries or continents across the years.

3. Safety Analysis:

Mortality Rate and Accident Patterns: Analysing the mortality rate helps in understanding the safety aspects of climbing Everest. This includes studying accident patterns, identifying potential factors contributing to accidents, and investigating the role of oxygen usage in accidents.

Host Organization Impact: This objective involves investigating whether the choice of host organization (expedition operator or guiding service) correlates with climbers' success rates. Understanding this relationship can shed light on the role of host organizations in climbers' achievements.

4. Predictive Models:

Forecasting Climber Numbers: Building predictive models to estimate the number of climbers for future years based on historical trends. This aids in understanding potential future demands for resources and infrastructure on Everest.

Climbing Success Prediction: Using regression analysis to predict the likelihood of climbers' success in reaching the summit. Factors like age, oxygen usage, and the climbing season will be considered to assess their impact on success rates.

Tools and Technology

I used a variety of data visualization methods, including pie charts, bar charts, and line charts, to effectively present and explain the expedition data using the Jupyter Notebook environment and Python programming. Furthermore, I employed diverse statistical techniques in my analysis, such as the application of sophisticated predictive models like General Additive Models (GAM) and logistic regression.

Jupyter Notebook

It's an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. I utilized Jupyter Notebook for its interactive interface, combining Python code with explanatory text to analyze the Everest expedition dataset collaboratively.

Programming in Python

Python is a high-level, multipurpose language that is renowned for being readable and simple to use. I made use of Python's vast library for analysis, visualization, and data manipulation. Its adaptability made it possible to apply a range of statistical methods and machine learning models.

Techniques for Presenting Data

Pie charts: These useful visual aids for displaying categorical data effectively show proportions as slices of a circle.
Bar charts are excellent for showing changes over time through bars of varying heights or for comparing categorical data.
Line charts are useful for showing relationships and trends in data over time; they are frequently employed in time-series analysis.

General Additive Models (GAMs)

Designed to handle complex relationships between variables, GAMs are a flexible statistical modelling technique that build upon generalized linear models (GLMs). They were helpful in identifying complex patterns in the expedition data and are useful for capturing nonlinear relationships. I have used this model to predict the death rate and future expeditions.

Logistic Regression

A statistical model called logistic regression is used to examine the relationship between one or more independent variables, such as age and oxygen consumption, and a categorical dependent variable, such as success or failure. It's frequently applied to binary classification problems and was utilized to determine how different factors affected the success of climbers. I have used this model to analyse the impact of host country and oxygen in the expeditions.

Methodology

The incremental model serves as the foundation for the development of this project. This paradigm combines the linear sequential model with the iterative prototype approach. New features will be added as each phase of development is finished. The phases of the linear sequential model are analysis, design, coding, and testing. The program iterated through these stages on a regular basis, providing incremental improvements for each new increment.

Project Flow

1. Data Reading and Processing

I've undertaken comprehensive data processing to enable in-depth analysis based on seasonal, yearly, and geographical perspectives, along with future predictions. To facilitate this analysis, I partitioned the 'Yr/Seas' column into 'Year' and 'Season', allowing for a granular exploration of data across years and seasons. Additionally, I introduced a 'Decade' column to observe trends spanning different decades.

Moreover, I've enhanced the dataset by introducing new columns such as 'Country', 'Continent', and 'Country Code', derived from the 'Citizenship' column. This enhancement involved refining the geographical data to standardize country names using the 'pycountry' Python library. This transformation aligns the dataset's geographical information with a validated dataset, enabling a more accurate depiction of the Everest expedition's geographical trends over time.

2. Implementing Different Model

I've chosen two main models that best fit my dataset after thoroughly investigating a variety of modeling approaches, including logistic regression and general additive models (GAM), as well as more complex techniques like linear and nonlinear regression and tree-based approaches. These decisions are based on how well they match the features of the dataset and how well they can predict the outcomes of upcoming expeditions. Furthermore, I have employed an assortment of statistical methods, such as means, regression analyses, correlation studies, variance evaluations, and additional ones. Using these statistical tools and prediction models, I aim to provide an extensive analysis of expedition data over various time periods.

Result

Expeditions trend over time

1. Overall expeditions trend over the decade by two different hosts
We can see the overall expedition numbers over the decade from the 1950s to the 2010s in the following bar chart. We can see that the popularity of the Everest climbing culture is increasing over time for both hosts. However, the data shows that more expeditions have been done from the host Nepal.

2. Climbing numbers in different Continents over time
As we know the fact, Tenzing Norgay from Nepal and Edmund Percival Hillary from New Zealand successfully climbed Mt. Everest in 1953. Among the different continents, it can be seen that the majority of submissions are done by Asians. It seems North America started to explore Everest in the 1960s and Europeans started in the 1970s, but the polarity of climbing Everest began in the 1980s for both North Americans and Europeans. Whereas, Africans and South Americans were the last in the race to explore Everest which is the 1990s. Overall, we can see the expeditions are getting increasing over time for all continents.

3. Climbing Numbers in Different Countries
Among the different country, it can be seen that Nepal are at the top of the list with more than half of the expeditions. And the USA at the second with 9.8% expeditions. China, India, and the UK are in the 3rd, 4th, and 5th positions respectively.

4. Gender and Age Distribution in Climbing
In climbing Everest gender distribution, very few numbers of climbers are Female. Also, the age between 25 to 40 is seeing top age group to for both Males and Females.

5. Future expedition prediction
I have tried different approaches like linear regression, Non-linear Regression, General additive models (GAM), Tree-based Methods, and more for future expedition prediction. However, I have chosen the GAM model by analyzing the scope of the model. I have found

the number of expeditions will increase over time. The GAM model predicts that the total number of expeditions might be more than 1200 in the year 2030. Note: The data for the years 2014, 2015, and 2020 are excluded from the prediction model because there were very few expeditions in 2014 and 2015 because of disasters and insufficient data for 2020.

Safety Analysis

1. Expedition safety related to Host selection and use of Oxygen
According to the death data distribution, it seems that the host country Nepal is a better choice than China. The death rate of expeditions hosted by China is higher than Nepal. I have selected the Logistic Regression Model to analyze the safety in terms of host selection and the impact of Oxygen in the Everest expeditions. It can be seen that the survival possibility will be high when people climb with oxygen.

1.1. Death Rate by Host
Expeditions from the Chinese side have a higher death rate than the Nepal side. Also, the death rate is on the pick in the 1990s but the rate is decreasing over time.

1.2.Death Rate prediction for the future years
I have implemented the General additive model (GAM) as a predictive model to analyze future death rates over time. Even though the number of expeditions is increasing over time, but death rate is decreasing which is good news for expidition lovers. The GAM model predicts that the death rate will be 0.00329 in the year 2040.

1.3. Survival probability with the use of Oxygen The majority of expeditions are done by using oxygen but there are some expeditions that are done without oxygen. I have analyzed the impact of oxygen on the death rate. I have found that the use of oxygen can improve the death rate. The death probability is higher when the explorer climbs without oxygen. Also, I have plotted the possibility of death by choosing a specific host. My result shows Nepal is pretty much safer than China.

Conclusion

Despite devastating losses, Everest expeditions from 1953 to 2020 showed growth, according to analysis. Examining mortality rates, the effects of hosts, and past trends in climber demographics provide insight into how the Everest landscape has changed over time. Stakeholders can now plan safer ascents and have a better understanding of the mountain's challenges for upcoming expeditions thanks to this Kaggle dataset.

Reference

1] “Mount Everest Ascent Data (1953-2020).” Accessed: Oct. 05, 2023. [Online]. Available: hlps://www.kaggle.com/datasets/ropandey12/mount-everest-ascent-data-19532020
[2] R. B. Huey, C. Carroll, R. Salisbury, and J.-L. Wang, “Mountaineers on Mount Everest: Effects of age, sex, experience, and crowding on rates of success and death,” PLOS ONE, vol. 15, no. 8, p. e0236919, Aug. 2020, doi: 10.1371/journal.pone.0236919.
[3] S. Tibshirani and H. Friedman, “Valerie and Patrick Has7e”.

[4] J. L. Westhoff, T. D. Koepsell, and C. T. LiIell, “Effects of experience and commercialisa7on on survival in Himalayan mountaineering: retrospec7ve cohort study,” The BMJ, vol. 344, p. e3782, Jun. 2012, doi: 10.1136/bmj.e3782.

[5] N. Kianfar and M. S. Mesgari, “GIS-based spa7o-temporal analysis and modeling of COVID-19 incidence rates in Europe,” Spat. Spa 10.1016/j.sste.2022.100498.