Project Title

Alex Friedrichsen, Cal Warshaw

Abstract

In the past several years, climbing has been increasing popular as a sport around the globe. And as data collection becomes more available to the average person, these hundreds of thousands of climbers have logged climbs to transnational databases to share their climbs with peers around the world. Climbers should have the ability to analyze their climbs with other climbers as well as get a more in-depth analysis of climbers around the world making different types of climbs. This analysis uses the data to group the type of climber to the difficulty and type of climb being made. The weight, height, experience, age and sex of the climber is examined with the purpose of comparing the characteristics and experience of the climber to the type and difficulty of the climb. [conclusion]

**Introduction**

With more climbs being recorded than ever before, analyzing these databases becomes all the more important to able to classify how the individual is affected by the type of climb and difficulty of the climb. With this research, we will be able to sort the variety of people and the experience they have that are making these ascents. This information can be used by new climbers as well as experienced climbers to compare themselves with other climbers that have made a type of ascent before. This information can make climbing safer as well as an analysis of where climbing currently is at in the world. We will look the clustering of height, weight, age of the climber and the number of years climbed. We will analyze the classification of the type of climb, or shorthand using the sex, weight, height, age of the climber, the number of years climbed and the bmi of the climber. We will find which factor contributes the most to classify an induvial. And we will attempt to find the difference in the climber traits, types and difficulties of climbs based on sex.

**Data Description**

We obtained this data from a Kaggle containing a European database of over four million climbers. This data was collected from the world’s largest rock-climbing logbook, 8a. The statistics given included a number of variables that we will not be using such as the name of the climber and the competition that they were in. We created the variables bmi from the weight and the height and also created the age at ascent and the years climbed from the birthday of the climber, the year the climber started climbing and the year of the climb. The database was imported using SQL into R. The variables look at are:

Shorthand: The type of climb

usa\_routes: The usa standard difficulty of a climb

Height: height of the climber in centimeters

Weight: weight of the climber in kilos

Age\_at\_accent: The age the climber was at ascent in years

Years\_climbed\_at\_ascent: years the climber has been climbing at ascent

Bmi: the body mass index of the climber

Sex: the gender of the climber

**Descriptive Statistics**

Plot 1: Summary Statistics

Table

Description automatically generated with medium confidence

Plot 1 describes the summary statistics of the height, weight, age at ascent, years climbed at ascent and bmi variables in the data. It can be seen that the average height is 176.7 cm, the average weight is 68.56 kg, the average age of ascent is 29.09, the average years climbed is 9.336 and the average bmi is 21.92.

Plot 2: Covariance matrix

Text

Description automatically generated

Plot 2 describes the covariance matrix of the quantitative variables being analyzed. The weight and the height, and the age at ascent and years climbed are the most correlated variables in the dataset. This is to be expected due to biology with weight and height, and the older you are the more likely you have more years of experience climbing.

**Methods and Results**

Plot 3: k-means

Chart, bar chart, line chart

Description automatically generated

A k-means analysis was conducted using the height, weight, age of the climber and the number of years climbed. We chose 2 cluster to analyze these variables. These clusters are merged slightly so while the two clusters are separate there is little evidence to suggest a big difference between the two clusters. This means that while there is grouping, it is not complete.

Plot 4: QDA confusion matrix

A screenshot of a computer

Description automatically generated with medium confidence

A QDA analysis was conducted on the height, weight, bmi, age\_at\_ascent, years\_climbed\_at\_ascent to analyze the shorthand, or the type of climb ascended. The box M test reject the null hypothesis for homogeneity of variance, so the QDA analysis over the LDA analysis was conducted. It was found that the analysis predicted the type of climb by the climber to 53% accuracy.

Plot 5:

Chart, line chart

Description automatically generated

A principal component analysis was conducted on height, weight, bmi, age at ascent and years climbed at ascent. It was found that 2 principal components were to explain this data accurately.

Plot 6:

Graphical user interface, chart, histogram

Description automatically generated

A k nearest neighbors analysis was conducted to classify the sex of the climber from height, weight, bmi, age\_at\_ascent, and years\_climbed\_at\_ascent. We are still at the stages of accurately cleaning the data so that the knn analysis will be most accurate.

**Conclusion**

**Limitations and Recommendations**