3/03/2022
Stat 223

Stat 223 Final Project Proposal
Group members: Cal Warshaw, Tomas Georgsson, Alex Friedrichsen

About the data:
Dataset: https://www.kaggle.com/dcohen21/8anu-climbing-logbook

Our data comes in the form of an SQLite database with four tables, with data divided as can be seen below in the following schema.

TableName	Total Rows	Total Columns	Primary_Key/Foreign_Key
Ascent	4111877	28	PMK: id, FKS: user_id, grade_id, method_id
Grade	83	14	id/grade_id
Method	5	4	id/method_id
User	62593	22	id/user_id

We foresee dealing with a couple of initial challenges in working with and attempting to draw conclusions from our data. Firstly, the data likely suffers from a self-selection bias, as the climbers who have chosen to upload their data are very serious and not a representative sample of all climbers. In addition, the data will perhaps suffer from geographic bias due to its origin on a European site.

Data Cleaning:
Since tables Ascent and User have 50,000+ observations we have decided to only look at the top 10,000 in order to reduce overall processing time from our limited machines. From a quick look through the data we can see a lot of 0 values for weight, we are only going to be looking at sports climbing so we will only be looking at flash, onsight, and redpoint climbing methods.

Possible Questions:
Ideal body type for sports climbers
Climb locations
Number of times taken to complete a climb
Climb types and who climbs more (m/f)
What where the popular climbs and how they changed over time

Modeling, Analysis, and Visualization Plan:
Make a heatmap to look at correlation between factors in the data.
PCA on the number of climbing tries to see the biggest determinant in contributing to the amount of tries
K-means on heights of climbers based on certain climbs. Are there climbs that are better suited for certain heights?