**Term Project on Predicting Hiked Times**

**Timothy Robbins**

**DSC530-Data Exploration and Analysis**

**11/11/2019**

**Statistical/Hypothetical Question**

Can hiking time be predicted based on the following variables: length of hike, maximum elevation, elevation gain, maximum speed, and/or difficulty level of the hike?

**Outcome of your EDA**

I found that most of the variables did, in fact, have a positive correlation with hiking (moving) time. The strongest correlations with moving time were the length of the hike (.85) and the elevation gain (.62). There were also fairly strong correlations with moving time and the following variables: maximum elevation (.33) and the difficulty level of the hike (.21).

The optimum regression model tested included the hike time as the outcome variable, with the distance hiked and elevation gain being the predictor variables. The model yielded an R-squared of 0.76, and a p-value of less than .00. I thus believe they could be used to predict hiking times for the general population.

**What do you feel was missed during the analysis?**

I was surprised to find a negative correlation between hiking time and maximum hiking speed. I think it would have been a good idea to take a deeper dive into the dataset to analyze maximum hiking speeds for more potential outliers or invalid data.

**Were there any variables you felt could have helped in the analysis?**

I could have calculated an average hiking speed with the available variables. However, the resulting variable would obviously have a very strong correlation to hiking time (perhaps even 1:1), so would probably not be a good candidate for testing.

**Were there any assumptions made you felt were incorrect?**

I did make several assumptions about the maximum values of the variables (i.e. assumed that maximum elevation was <= 20,000 feet). Based on my hiking experience, I was fairly comfortable with my decisions that values exceeding the maximums were errors, but there is a small possibility they may be off; there may be some incredibly good hikers out there.

**What challenges did you face, what did you not fully understand?**

The biggest challenge for me was the graphs. I had some difficulty understanding some of the code provided by the author, as it was more difficult than what I had learned in my introductory Python class. I found some DataCamp classes on matplotlib and pandas, that were pretty easy to pick up on. Also, I wasn’t too familiar with PowerPoint, so it took me a while to take screenshots of the graphs generated in Jupyter Notebook and get them into a presentable format in PowerPoint.