Mount Rainier Report and Analysis

1. Introduction

This report contains analyses conducted on data collected about Mount Rainier. Specifically, there were two datasets utilized in these analyses: 1. A dataset containing the date, routes, climbs attempts, climb successes, and a summary variable of success percentage. There were 4077 rows in this original dataset. 2. A dataset containing the date and a number of weather variables- battery voltage, temperature, humidity, wind speed, wind direction, and solar radiation. This original dataset included 464 rows. These data covered the years 2014-2015.

The questions that were asked were:

1. What is the success rate of each route on Mount Rainier? (story 175)
2. What is the relationship between weather and success rate? (story 176)

This report will aim to answer these questions and provide further details about the data used, methods, overall results, and conclusions. The first question is exploratory in nature and will be addressed through use of visualization representation. The hypothesis for the second question is that there will be a relationship between success rate and weather variables (null hypothesis being that there is no relationship between success rate and the weather).

2. Body

2a. Data

For question 1, I utilized jupyter notebook for all data manipulation and visualization. Initially, I looked at the columns to see if there were any that immediately did not seem relevant. I opted to leave all of the columns in initially and move on to cleaning the data. This began with searching for null values. The data set did not have any of these so no removal or imputation of data due to missing values was needed. In addition, variables did not need to be scaled or centered in any way given that I would not be dealing with issues like multicollinearity in answering this question. Finally, there was a restricted range on the success percentage variable (0-1) so outliers were not relevant to find. However, looking at the summary data it was clear there were some errors in the data as the max percentage was over 1 (meaning over 100%). I removed these rows from the data (n = 10). Following this I transformed the success percentage variable column to an actual percentage by multiplying the column by 100 as this would make it easier to read tables/graphs.

Since this question asked about success by route, I then grouped data by route and created a new data frame with this information. Following this, I then removed columns ‘attempted’ and ‘succeeded’ to clean up the data set as these would not be useful. Finally, I created a data table where I presented each route and its success percentage in ascending order.

For question 2, I needed to combine data about Mount Rainier weather with the success percentage from the climbing statistics data set. To do this I needed to find a common column they shared which was the ‘Date’ column. In the weather dataset, for each day there was one single number for each weather variable which was an average of that weather variable for the day. Whereas in the climbing statistics data set, there were many days where there was data on multiple climb attempts for that day. This meant that in addition to some of the data cleaning and transformation discussed above, I would also have to aggregate the data so that there was only one success percentage point per day if I was going to combine the datasets appropriately. I did this by calculating the sum of both the “attempted” and “succeeded” columns. From there I then recalculated the “success percentage” column, and again removed rows where the percentage was greater than 100. Following this, I combined the weather and climbing statistics data using a join statement and used the date as the column to link these data sets. This resulted in removal of a fair amount of data from each data set with the combined data set containing 203 rows. With the combined data set created, I then created a variable for month so that I would be able to look at aspects of the data by month as I hypothesized that there would be trends with successful climbs and aspects of the weather that varied meaningfully by month. Having this variable would make it easy to graph data and get summary stats by month.

2b. Analysis

To visualize aspects of the data, I utilized graphs and tables which are presented below in the results. I also utilized descriptive statistics to summarize certain aspects of the data. I used python, R and power BI for these tasks.

I utilized a linear regression model with Success Percent as the dependent variable and weather variables (battery voltage, temperature, humidity, wind speed and direction and solar radiation) as independent variables. I chose linear regression as all of the variables are numeric and continuous. An important assumption of linear regression is that there is not multicollinearity between independent variables. Therefore, prior to running the regression model, I examined correlations between the independent variables. There were no variables that had correlation coefficients over .7. To make even further sure that multicollinearity would not be an issue and to improve interpretation of the results of the regression, I centered the independent variables. A limitation to using linear regression in this example was that some of the variables may not have been linearly distributed. Therefore, there may have been models to chose from for certain independent variables; however, for the sake of time for this brief project, this was not further explored. I utilized R for these analyses.

2c. Results

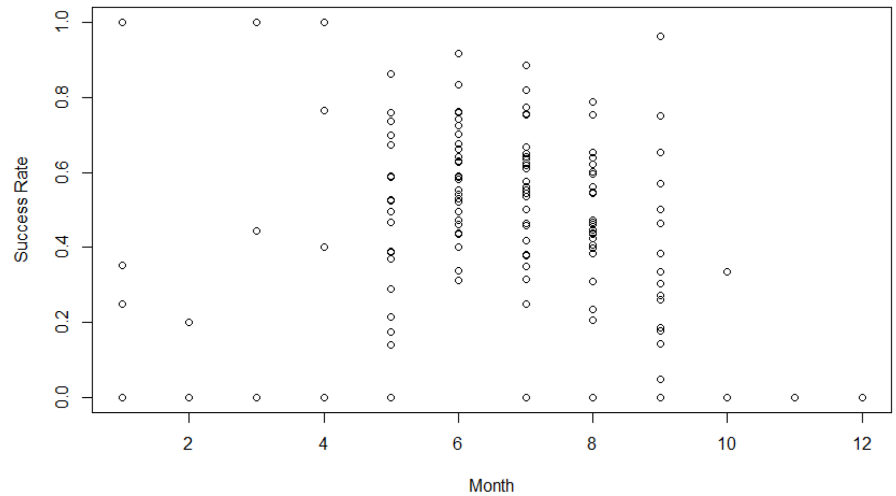
To answer the first question (story 175), I created a table where I grouped the data by route climbed. This clearly shows that success differs greatly depending on the route chosen to climb. In particular, there are several routes which were not successfully climbed between when the data was collected in 2014-2015. Meanwhile, the route ‘Tahoma Cleaver’ had a 100% success rate.

With respect to success rate by time of year, it is clear upon visual inspection that certain times of year are more and less advantageous to try to climb Mount Rainier. In particular, late spring to early fall appear to be associated with the highest rates of success. June (m = 60.27%, SD = 14.72) and July (m = 54.71%, SD = 18.22) had the highest rates of success while December (m = 0, SD = 0) and November (m = 0, SD = 0) had the lowest rates of success. It is also important to take into account that number of attempts were highest in the summer months and lower in winter months which naturally impacts the success rate.

Table 1. Success Rate by Route

Route	SuccessPercent
Sunset Amphitheater	0.00
Liberty Wall	0.00
Edmonds HW	0.00
Wilson Headwall	0.00
Kautz Headwall	0.00
Sunset RIngraham Directge	0.00
Nisqually Glacier	0.00
glacier only - no summit attempt	4.55
Fuhrers Finger	11.11
Ingraham Direct	13.21
Gibralter Chute	20.00
Unknown	23.24
Mowich Face	25.00
Kautz Cleaver	26.19
Gibralter Ledges	26.97
Fuhrer's Finger	37.18
Tahoma Glacier	42.42
Little Tahoma	47.27
Curtis RIngraham Directge	50.00
Success Cleaver	50.00
Disappointment Cleaver	51.95
Liberty RIngraham Directge	52.17
Kautz Glacier	52.69
Ptarmigan RIngraham Directge	53.03
Emmons-Winthrop	53.39
Tahoma Cleaver	100.00

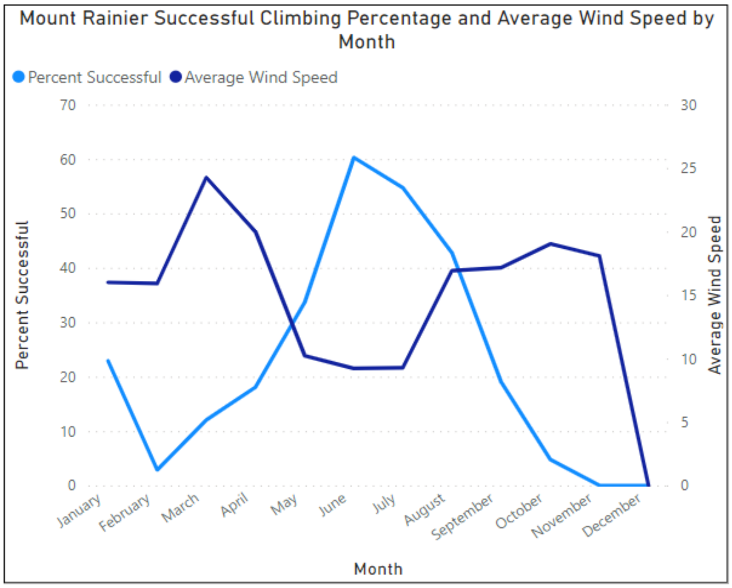
Graph 1. Success Rate by Month



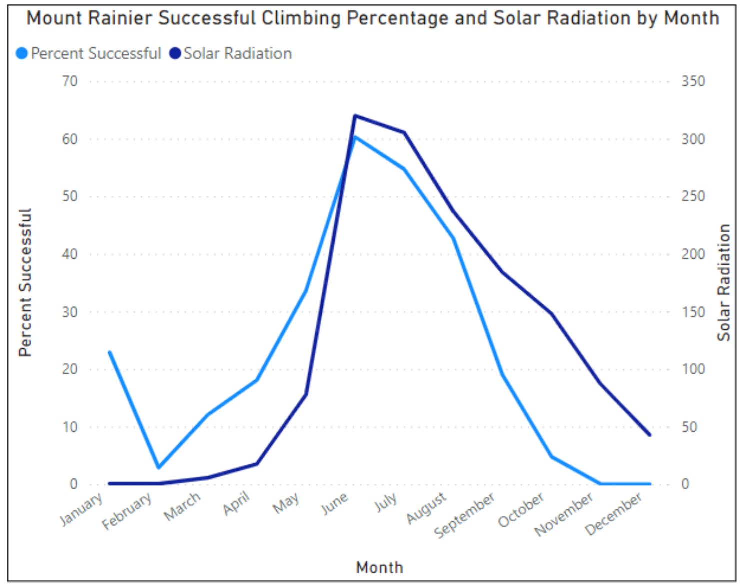
A multiple regression was run to predict Success Percentage from battery voltage, temperature, humidity, wind speed, wind direction and solar radiation. These variables significantly predicted Success Percentage, $F(6, 196) = 15.77$, $p < .001$, $R^2 = .32$. Only wind speed ($\beta = -.04$, $p < .05$) and solar radiation ($\beta = .11$, $p < .001$) added statistically significantly to the prediction. What this means is that this model explained 32% of the variance in Success Percentage. However, there were only two weather variables – wind speed and solar radiation- were significant predictors of Success Percentage.

To better visualize the relationship between Success Percentage and Wind Speed and Solar Radiation, I created time series graphs using power BI and plotted both variables by month together. The allowed for clearly seeing relationships between these variables. Specifically, it appears that average wind speed is lower in summer months and higher in the winter, and, therefore, success percentage is higher when wind speed is lower. Alternatively, looking at the graph for success percentage and solar radiation, it is clear that these variables overlap significantly; that is, as solar radiation increases in the summer months, climbing success percentage increases.

Graph 3. Success Percentage and Wind Speed by Month



Graph 4. Success Percentage and Solar Radiation by Month



3. Conclusions

In summary, this report summarizes analyses and results answering the questions – 1. What is the success rate of climbs by route on Mount Rainier? and 2. Is there a relationship between climbing success rate and weather on Mount Rainier? The data show that if you want to have the highest likelihood of successfully climbing Mount Rainier, you would want to climb the Tahoma-Cleaver route in June when the solar radiation is high and wind speed is low. You would be least successful in climbing a route like Sunset Amphitheater during December or November when the wind speeds are higher on average and solar radiation is low.

There were some limitations to these analyses, some of which were discussed above. There were some months where there were very few climb attempts which resulted in little data for those months and therefore, success rates may have been artificially low. However, this is likely a representation of reality as climbing Mount Rainier in winter is dangerous due to frequent severe storms. Secondly, time of year is likely functioning as a confounding variable in the data. Weather metrics change by time of year as can be seen clearly from the graphical representations of wind direction and solar radiation. Climbing success rates are also clearly associated with time of year. To examine if weather variables play a role in climbing success rates, it would likely be beneficial to restrict the data to the months where it is climbed regularly (i.e., May- September). This may give a better estimate of if weather variables were relevant to successful climbs and resolve issues with the confounding aspect of time of year. In addition, with more time, additional types of regression models may be used to fit the data and/or corrections to the data distributions. For instance, there are ways to correct for seasonality trends which can aid in improving modeling and forecasting with machine learning methods.