Analysis of Avalanches in the Pacific Northwest

Josh Bullers

March 26, 2016

The Problem

Avalanches are an intimidating force of nature and seem to be a randomly occuring event in nature. While they seem random, avalanches are heavily dependent on the structure of the snow and external forces. Many avalanche predictions are based on an expert going out into the field and cutting sample blocks from the snow to observe the structure of the snow in that region and determine if there is currently a risk of avalanches. This process is not only time consuming but limited to the specific region the expert is able to visit that day.

The main issue with this type of measurement is its scalability. Physical examination of the snow is fairly accurate in the ability to determine the risk of avalanche, but is costly in the sense that it requires one expert per sample site per day.

It is because of this reason that an analysis of weather data should be done to determine if avalanches can be better predicted based on the main dependent variable which is weather.

The Data

Weather data comes in many formats and can become an extensive project if attempting to use an assortment of resources. The exact sources are listed at the end of this paper, but the main resources utilized were the Northwest Avalanche Center (NWAC) and Snotel Weather Reports provided by National Resource Conservation Service (NRCS).

This data focuses on independent variables such as minimum and maximum temperatures by day and the daily snow to water depth, which is provided in Inches. These data points are gathered from Snotel Stations that provide a seperate .txt file for each variable. When using all of Washington State's Snotel Sites, there are roughly 70 different locations to download, with each site containing data in seperate .txt files for each variable.

The data on avalanches is provided by the NWAC and the observations are not in an easily downloadable format, but rather HTML on their site. To extract this data easily, a Python html crawler was created using the package BeautifulSoup. This package allows the user to easily extract specific elements from the HTML. BeautifulSoup in combination with Python Pandas, allows for the scraped avalanches to be output into a cleaned .csv, which can then be consumed by an R script for further analysis.

Assumptions, Limitations, and Possible Bias

The main focus of this analysis was on the weather, so datestamps were used as more of an index than to create timeseries data. With this focus, an assumption was made that avalanches typically occur between the months of November and early June and that a year round view of the weather data was not needed. Weather outside of the winter months where snow levels were at 0.

A limitation of this analysis is in the weather data used. Weather data can involve hundreds of different variables that can be collected from various sources. With so many variables and sources the process of collecting all of this data could take an extended period of time, that is not within the limits of this project. It is that reason, that this analysis focuses on only minimum and maximum temperature and snow to water equivalent measurements.

Possible bias occurs in the observation of avalanches. There is currently no data on every avalanche that occurs in the Pacific Northwest or even daily data on every avalanche within a specific location to create a

sample set. Given the proper amount of time, a better data set could be built by using weather monitoring equipment and cameras to observe a specific location 24 hours a day for a full winter season.

With that acknowledged, this analysis attempts to compensate for that bias by using multiple observations sites across the Pacific Northwest, and combining these observations into one main data set.

The Hypothesis

It is the hypothesis of this analysis, that there is a difference in weather between the sample of no avalanches(s1) and the sample of avalanche observations(s2). In terms of testing the difference between sample sets, the hypothesis is that s1 not equal to s2. Therefore the null hypothesis is s1 = s2 and this analysis will attempt to disprove this later in the paper.

In addition to the difference of means, the hypothesis is also that a combination of these variables will help predict avalanches.

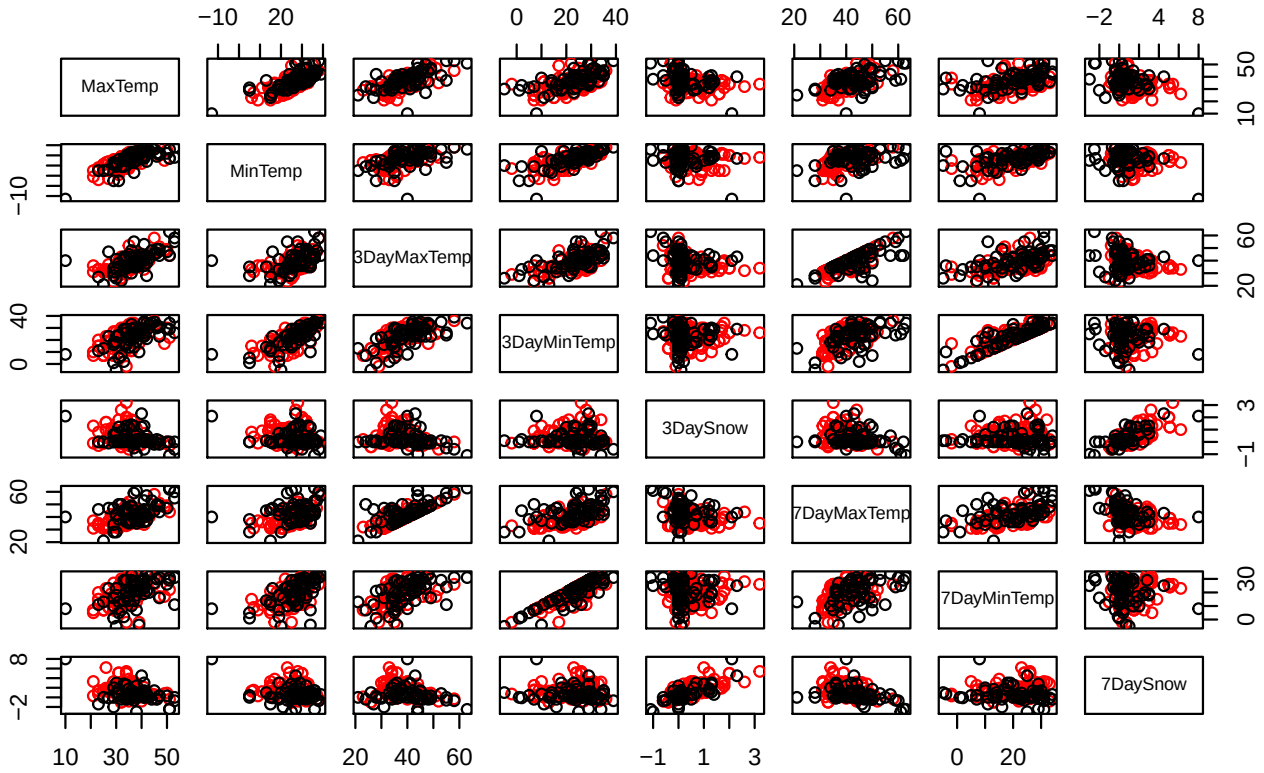
Describing the Data

The data used contains two sample sets. The first sample contains random observations from individuals in the mountains, when avalanches were observed. The observations have been paired with associated weather data. The variables associated with the observations are:

- 7 Day High Temperature
- 7 Day Low Temperature
- 7 Day Snow Gain/Loss
- 3 Day High Temperature
- 3 Day Low Temperature
- 3 Day Snow Gain/Loss

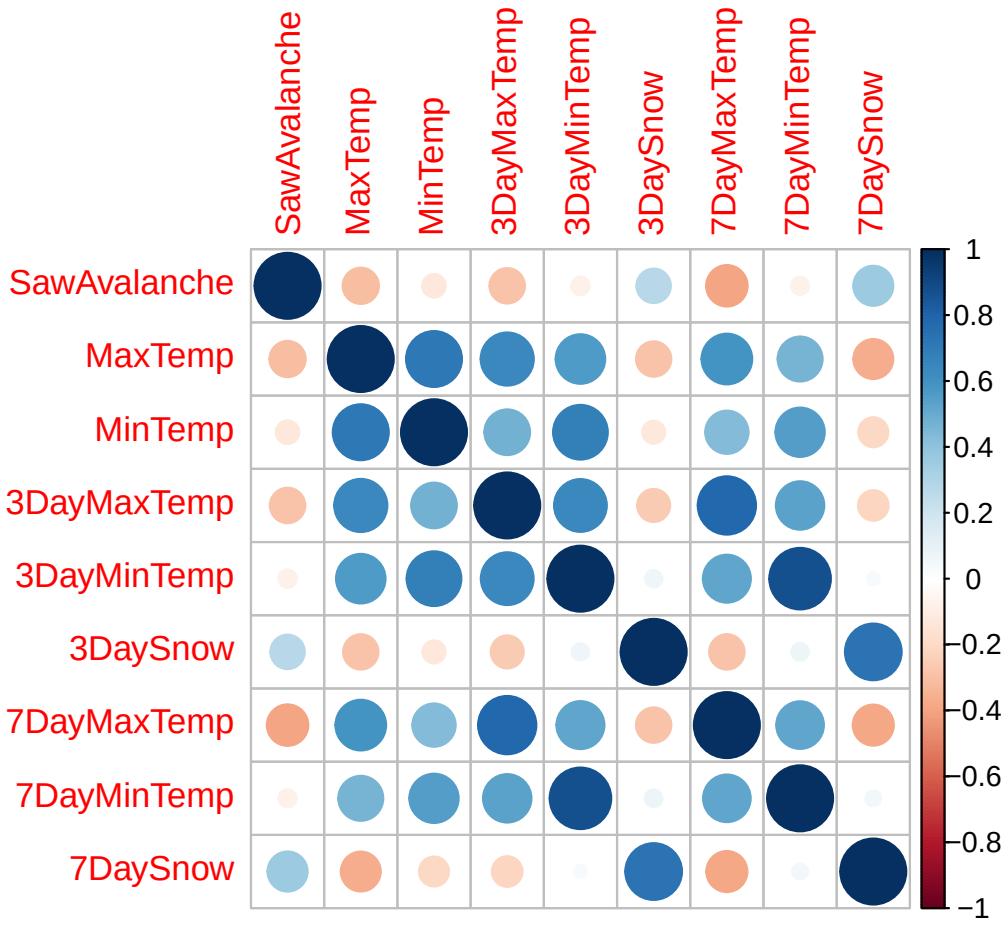
The choice was made to use high and low temperatures since this gives a better picture of the extremes that effect the snow. Using an average would smooth the observations, making it more difficult to get a full picture of what is going on.

To begin looking into the data, it is helpful to plot the points on a scatterplot. A basic scatter can only discover insights about two variable combinations at a time, so to get a full view of what is happening, a scatterplot matrix is useful.



From the matrix it appears that there may be some interesting findings between snow change data and temperatures when looking for differences in the two sample sets. There does not seem to be as much happening when plotting between only temperatures and the two sample sets.

A second view into the data that may review important information is a correlation matrix to attempt to discover any correlation between the data. In using correlation, this project assumes that a correlation of -.5/.5 to -1/1 indicates a strong correlation, -.3/.3 to -.5/.5 indicates a moderate correlation, -.1/.1 to -.3/.3 indicates a weak correlation and -.1 to .1 indicates no correlation.



The first thing to note about the correlation matrix is that not all of the variables are independent from each other. Some of the variables have a strong interdependence, such as 3DayMinTemp and 7DayMinTemp. This is to be expected as it is quite possible that the minimum temperature for the 7 day window is the same as the 3 day window.

Another interesting takeaway, is that the initial thought that there may be some interesting takeaways between avalanches and both temperature and snow change, may be true. It can be seen that there is some moderate correlation between avalanches and MaxTemp (-.3), 7DayMaxTemp (-.4), and 7DaySnow (.37).

With these correlations present, it would be helpful to find the mean of these variables for comparing the sample sets. The descriptive statistics for the sample sets are as follows:

Avalanche Sample Descriptive Statistics

```
##                 MaxTemp  7DayMaxTemp    7DaySnow
## nbr.val      98.0000000   98.0000000  98.0000000
## nbr.null      0.0000000    0.0000000   2.0000000
## nbr.na        0.0000000    0.0000000   0.0000000
## min          21.0000000   31.0000000  -1.0000000
## max          51.0000000   58.0000000   6.2000000
## range        30.0000000   27.0000000   7.2000000
## sum        3330.0000000 3850.0000000 146.3000000
## median       34.5000000   39.5000000   1.2000000
## mean         33.9795918   39.2857143   1.4928571
```

```
## SE.mean         0.5939912    0.5544230   0.1445282
## CI.mean.0.95    1.1789080    1.1003761   0.2868484
## var            34.5768988   30.1237113   2.0470619
## std.dev         5.8802125    5.4885072   1.4307557
## coef.var        0.1730513    0.1397075   0.9584009
```

No Avalanche Sample Descriptive Statistics

```
##                   MaxTemp  7DayMaxTemp   7DaySnow
## nbr.val        60.0000000   60.0000000 60.0000000
## nbr.null        0.0000000    0.0000000  7.0000000
## nbr.na          0.0000000    0.0000000  0.0000000
## min            10.0000000   21.0000000 -3.0000000
## max            53.0000000   63.0000000  8.0000000
## range          43.0000000   42.0000000 11.0000000
## sum          2246.0000000 2640.0000000 33.1000000
## median         37.0000000   44.0000000  0.3000000
## mean           37.4333333   44.0000000  0.5516667
## SE.mean         0.9913087    1.0722488  0.2116053
## CI.mean.0.95    1.9836042    2.1455649  0.4234212
## var            58.9615819   68.9830508  2.6866073
## std.dev         7.6786445    8.3056036  1.6390874
## coef.var        0.2051285    0.1887637  2.9711553
```

Analyzing the Descriptive Statistics

It can be seen that there are in fact differences between the means of the two sample groups. To find out if the difference between the sample means is significantly different, an independent test of difference between means should be conducted. An independent test is used in this case since there is no overlap between the samples in each sample set.

When conducting this test in R, a Welch Two Sample t-test can be used to produce a 95% confidence interval and determine if the null hypothesis can be rejected.

MaxTemp Mean Test

```
##
##  Welch Two Sample t-test
##
## data:  noAvalanche and avalanche
## t = 2.9886, df = 101.05, p-value = 0.003519
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.161263 5.746220
## sample estimates:
## mean of x mean of y
##  37.43333  33.97959
```

7DayMaxTemp Mean Test

```
##
##  Welch Two Sample t-test
##
## data:  noAvalanche and avalanche
## t = 3.9054, df = 90.817, p-value = 0.0001806
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.316455 7.112116
## sample estimates:
## mean of x mean of y
##  44.00000  39.28571
```

7DaySnow

```
##
##  Welch Two Sample t-test
##
## data:  noAvalanche and avalanche
## t = -3.6729, df = 112.05, p-value = 0.0003692
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4489186 -0.4334623
## sample estimates:
## mean of x mean of y
## 0.5516667 1.4928571
```

Difference of Sample Means Conclusions

As can be seen from the Welch Two Sample t-test, the null hypothesis can be rejected for all three variables tested. This analysis can therefore conclude that there is a significant difference between the means of the maximum temperature, 7 day maximum temperature, and 7 day snow change, when comparing the avalanche and no avalanche sample sets.

While all three variables have been accepted as having different means, it is worth noting the the p-value is significantly lower for the 7 day variables than that of the observation days' maximum temperature.

Predicting Avalanches

With three variables showing a statistically significant difference of means, it is worth attempting to predict avalanches with these variables. Since the outcome is a logical variable rather than continuous, a logical regression will be used, rather than a linear regression.

The first step of running a logical regression on this data is to shuffle the sample rows together so that a training set and test set can be used. Next the model that will be tested should be considered. Since there is interdependence between MaxTemp and 7DayMaxTemp, one should be thrown out. Looking back out the test of difference between means, it appears that 7DayMaxTemp had a lower p-value so the 7DayMaxTemp variable will be kept. The only two variables left to be used in a predictive model are the 7DayMaxTemp and 7DaySnow.

```
##
## Call:
## glm(formula = SawAvalanche ~ ., family = "binomial", data = modelSubset)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4504  -1.0838   0.6631   0.9278   1.7418
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.42078    1.27278   2.688  0.00720 **
## `7DayMaxTemp` -0.07912    0.02921  -2.708  0.00676 **
## `7DaySnow`     0.33693    0.14795   2.277  0.02276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.81  on 157  degrees of freedom
## Residual deviance: 186.45  on 155  degrees of freedom
## AIC: 192.45
##
## Number of Fisher Scoring iterations: 4
```

Interpreting the Model

As can be seen from the produced model, both variables are significant in predicting if an avalanche will occur. The result is interpretted as log odds when performing a binomial regression. This model indicates that for every 1 degree increase in the maximum temperature, the log odds of an avalanche drops by .079. As for snow change, for every 1 inch increase in snow over a seven day period, the log odds of an avalanche will increase by .337.

Conclusions of Avalanche Data Analysis

In conclusion, this analysis has been able to determine that the orginal hypothesis that the means of weather data is different between the avalanche observations sample and the no avalanche sample, can be accepted. It was determined that maximum temperature the day of observation, 7 day maximum temperature, and 7 day snow change, all are statistically significantly different. In addition, a predictive model in the form of a logistic regression was formed with 7 day maximum temperature and 7 day snow change.

In the future, this study can be improved by obtaining a better data set by having a sample site that provides 24 hour monitoring and reporting for a full winter season.

The full project can be found here: https://github.com/joshbullers/avalanche_project

Data Sources

Northwest Avalanche Center
https://www.nwac.us/observations/

National Resources Conservation Service
http://www.wcc.nrcs.usda.gov/snow/snow_map.html