Cookbook for tidying avalanche data

LoadDate.R

This code loads the data and tidies it into two different data frames.

Downloading and Reading Data The data on the avalanches may be downloaded manually at: `https://utahavalanchecenter.org/avalanches/download`

The data on the weather is downloaded from the NOAA. For the data set for the Alta Guard Station located at Alta Ski Resort in Little Cottonwood Canyon go to `http://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00420072/detail` add to cart, select .csv file type, all possible date range, set units to metric and select all the variables and used R to remove the variables that were not useful due to lack of data. Weather data for Salt Lake city can be found at `http://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:US490006/detail`. I decided not to use this information since the Wasatch is a micro-climate that can have completely different weather than Salt Lake City. As a point of reference the link to the NOAA's datasets is `https://www.ncdc.noaa.gov/data-access`

The Alta Guard Station does not include data pertaining to wind events. Since wind is a fundamental cause of snow-pack destabilization I retrieved data for wind events from: `http://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USS0011J69S/detail`. The Louis Meadow station was chosen for the wind events, because it is nearest station located in the Wasatch at an altitude similar to Big and Little Cottonwood Canyons.

The data is then read into R, the NOAA encoded NA as -9999, so this is passed into the read.csv function.

Description of variables The following is a description of the variables in addition to the changes of the variables names into human readable forms.

- PRCP $->$ Precipitation - the amount of precipitation - (in tenths of mm)
- SNWD $->$ Snow_Depth - the depth of the snow pack - (in mm)
- SNOW $->$ Snowfall - amount of snowfall in 24 hour period - (in mm)
- TMAX $->$ Max_Temperature - the maximum temperature in 24 hour period - (in tenths of C deg)
- TMIN $->$ Min_Temperature - the minimum temperature in 24 hour period - (in tenths of C deg)
- TOBS $->$ Temperature_at_observation_time - the temperature at a random observation time - (in tenths of C deg)
- AWND $->$ Average_Wind_Speed - the average wind speed during a 24 hour period - (in meters per second)
- WSFI $->$ Max_Wind_Speed - the maximum wind speed during a 24 hour period - (in meters per second)
- Date - the date pertaining to the avalanche/weather/wind event - (in month-day-year)
- Region - the general location in Utah that avalanche occurred in
- Place - the locally or forest service name for the location of the avalanche

- Trigger - the cause of the avalanche
- Depth - the depth of the avalanche that occurred - (in inches)
- Width - the width of the avalanche from flank to flank that occurred (in feet)
- Vertical - the distance the avalanche traveled from crown to stauchwall (I believe, it could also be crown to debris pile) - (in feet)
- Aspect - the positioning of the slope in relation to the compass
- Elevation - the altitude the avalanche occurred at - (measured in feet)
- Caught - the number of people caught in the avalanche
- Carried - the number of people carried in the avalanche
- Injured - the number of people injured by the avalanche
- Killed - the number of people killed in the avalanche
- Latitude - the latitude of the avalanche - (in degrees)
- Longitude - the longitude of the avalanche - (in degrees)
- WeakLayer - the layer the avalanche failed on
- BuriedFully - the number of people who were fully buried by the avalanche
- BuriedPartly - the number of people were partly buried by the avalanche

Make dates variable "Date" The avalanche dates data came in mixed forms, but all of it was in month followed by day followed by year. So for each different format the dates are put into month-day-year. First the two digit 2000's are taken care of. Then the dates with two digits of the 2010s, then the two digit 1920s and up are taken care of. Finally the four digit dates are encoded.

The dates are then assigned the class of dates and checked.

Variables recorded as characters Some variables are thought to be characters despite being numeric so these variables are forced to be numeric. This necessarily introduces some NAs.

Fix the Depth units The Depth variable had observations that were in feet and inches. This is fixed and all foot measurements are converted into inches.

Split Latitude and Longitude The latitude and longitude were contained in one column, I split these out into their own columns.

Blanks set to NA All blank cells were converted to NAs

Removed variables The following is list of variables that were deemed unfit and removed from the data set. The reason they were removed is included

- STATION - Not useful, because we know what information came from which station.
- STATION_NAME - not useful for the same reason STATION is not useful
- MDPR - not enough observations to be included
- MDSF - not enough observations to be included
- DAPR - not enough observations to be included
- DASF - not enough observations to be included
- WT01 - not enough observations to be included
- WT06 - not enough observations to be included
- WT05 - not enough observations to be included
- WT11 - not enough observations to be included
- WT04 - not enough observations to be included
- WT03 - not enough observations to be included

Creation of data frames Two data frames were created one that only contains observations of the avalanches that occurred (aw_df) and the weather and wind information that correspond to each avalanche. Another data frame was created that contains all possible observations of dates (fl_df) during which avalanches could occur and the weather and wind that correspond is also made. Also the dates that are NA are remove.

The fl_df This data frame contains all the possible dates for which avalanches could occur. In other words, there are no observations for avalanches in August, so the month of August could safely be removed. All the possible range of dates for which avalanches have occurred are identified, and the complement of this set is removed from the data frame. This includes the years for which only an insignificant number of avalanches was recorded.

subsetDates.R

This script subsets the fl_df further so that only significant dates are included. Put another way all the outlier dates, such as avalanches that occurred in June are removed.

First this code loads the data frames. It then parses out the months into a vector, so that the occurrence of the months may be analyzed. This vector is then looped over mod i so that the fact the winter months are split by the new year does not affect the possibility that an outlier. The possible outliers are then checked for how much of the overall data they compose. It was found that May only has 0.855% of the avalanches that occurred, June only 0.132%, and October only 0.132%. Because of this, the months from May to October were removed from the data frame.

This same process is done for the years. So first the years were parsed to a vector, for which we could analyze for outliers. First the percentage of years was inspected, it was found that this compared to the

outliers determined by boxplot.stats were comparable. So these outliers are assigned to a vector to be removed from the data frame.

subsetNumAv_Dates.R

This script subsets the data into observations of dates. Because of this, a new variable is recorded that records the number of avalanches that occurred on the given date. This collapses the variables of the different avalanches that occurred on a given date, because of this the mean of the variables was taken.

First the script loads the data. It then counts the number of avalanches that occurred and assigns it to a temporary data frame. The fact that this was done correctly is then checked. All the dates previously in subsetDates.R that were decided to be outliers are removed, as are any NAs.

A new data frame that will contain each date as an observation is then created, which takes the mean of any variables for which multiple avalanches occurred. It then merges this data frame with the number of avalanches that occurred and assigns 0 to all variables for which there were no recorded avalanches.

subsetIK.R

This script subsets the data into observations of bodily harm.

First the data is loaded. The observation variables (Injured, and Killed) are passed to a new data frame which is then melted and merged to the aw_df data frame. All non harmed observations are then removed.

subsetCCB.R

This script subsets the data to observations of avalanches that took people for a ride.

The processing for this data frame is exactly the same as for subsetIK.R, except the variables parsed for the observations are Caught, Carried, BuriedPartly, and BuriedFully used.