Team: Outstanding Owls
Mary Kendig
Namrata Rao
Neeraj Shirname

Data Cleaning

To monitor potential volcanic eruptions, the Icelandic Meteorological Office (IMO) records the longitude/latitude, depth, magnitude, and relative distance to known volcanoes of earthquakes. This information is released publically covering a 48 hour period and is updated every five minutes. The dataset displayed on the Icelandic Meteorological Office site is near-real-time data. To access the data, the Outstanding Owls utilized a web crawler to monitor the data and aggregate all the data into a single dataset. The first two columns in the the dataset represent the date and time (GMT) of the earthquake. Columns 3 and 4 denote the location of the epicentre. The next two columns focuses on the depth and magnitude of the earthquake. Column 7 is the measure of earthquake quality and finally column 8 provides details on the location of the epicentre from a nearby location.

Licensing

Users are free to use data on this web-site; however, if the data are used for presentation or publication, reference must be made to the Icelandic Meteorological Office. The terms and conditions are listed here: http://en.vedur.is/about-imo/the-web/conditions.

The official data citation for this set is: Icelandic Meteorological Office (2016). Whole country – earthquakes during the last 48 hours [Dataset]. Date accessed September 14, 2016. Retrieved from http://en.vedur.is/earthquakes-and-volcanism/earthquakes#view=table

Metadata

Metadata is available on the IMO website under the 'instructions on using the earthquake pages' under the related topics section on the right hand side of the dataset. This page has various details describing how the earthquakes data is captured, information on the earthquake graphs and map available on the site along with information that each column within the dataset represents. The metadata provides some basic information on how earthquakes occur and defines terms related to earthquakes such as 'hypocentre' and 'epicenter' for better understanding of the topic. The metadata also provides detailed information on how the color-coding is done in the maps and graphs provided on the website.

Rationale Behind Remediating Data

The first column that required cleaning was the Date column. In the Date column, daily nomenclature and the calendar date were concatenated into the same column. For example, dates read as "Wednesday

11/2/2016” rather than “11/2/2016”. While daily nomenclature is similar to calendar dates, including the item limits the ability to filter dates or conduct analytics by month or days. To clean this data, Neeraj utilized an Excel function to separate the one column into two separate columns, thus splitting the nomenclature and calendar date. The second column that required cleaning was the earthquake location. The Icelandic Meteorological Office noted earthquake locations by recording the distance from known volcanoes, the direction, and the known volcanoes’ name within the same column. As each piece of data provides important insight, it is necessary to separate them like the Date column. Furthermore, distance is a quantitative statistic while volcano and direction are qualitative. Using R programming scripts, the data was separated into three separate columns. In this form, deeper analytics can be conducted on the distance, direction, and volcano name.

Issues with Data

There were two main issues with the data that we encountered.

- There were a few records for which the 'magnitude' column had missing values .
  It is worth noting that we do have records with magnitude values of '0' in the dataset.
  So, an empty cell for magnitude does not mean that it is a record with magnitude value '0' but it simply means that the data is missing. However, since the number of such records was small in relation to the entire dataset, we decided to discard these ambiguous records.
- Since the data is about earthquakes in Iceland, the 'Location' column has the Icelandic names of all the locations where these earthquakes were recorded. However, while importing the data into Excel, some of the characters in the names were replaced by junk characters. We had to manually find the original names of these locations and replace them in the final data set for the locations to be in accordance with the ones in the original data set.

Data Cleaning Process

Step 1:	Deleted the URL Column from the data set since we did not require it for our study.
Step 2:	Parsed the 'Date' column into two separate columns 'Day' and 'Date' for clarity.

Step 3:	Identified values in the 'Location' column which had junk characters and manually replaced them by the corresponding Icelandic names of the locations.
Step 4:	Checked for NULL values in the data set (Identified and discarded 12 records with 'magnitude' field empty).
Step 5:	Checked for duplicates in the data set (None found).

Word Count: 780