Global Landslide Catalog Analysis

Shailja Kartik

December 06, 2021

# Introduction

The dataset we are using for my Final Project is the **“Global Landslide Catalog Export”** dataset obtained from NASA’s Open Data Portal.

A summary of the dataset is available at: <https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v>

The downloadable version of the dataset in the CSV format is from: <https://catalog.data.gov/dataset/global-landslide-catalog-export>

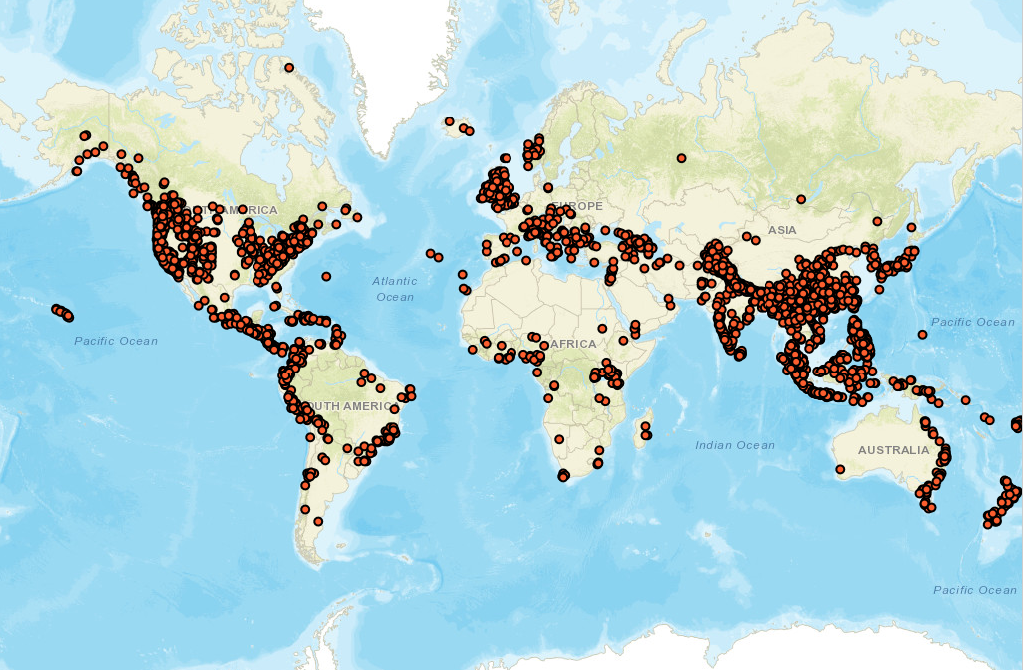
To learn more about NASA’s work on the landslide, please visit the homepage: <https://gpm.nasa.gov/landslides/index.html>

# Motivation behind using the Global Landslide Catalog Export (GLC)

Landslides cause billions of dollars in infrastructural damage and thousands of deaths worldwide. Data on past landslide events guides future disaster prevention, but we do not have a global picture of exactly when and where landslides occur.

NASA scientists have been building an open global inventory of landslides to address this problem. Knowing where and when landslides occur can help communities worldwide prepare for these disasters.

Below figure is just an overview of the collected landslides so far:



Through this project, we would welcome an opportunity to help make an informed decision that could save lives and property

# About the dataset

The Global Landslide Catalog (GLC) has been compiled since 2007 at NASA Goddard Space Flight Center.  
The GLC is a collection of **observational studies**.  
The GLC was developed to identify rainfall-triggered landslide events worldwide, regardless of size, impact, or location.  
The GLC considers all types of mass movements triggered by rainfall, which have been reported in the media, disaster databases, scientific reports, or other sources.

The dataset contains **31** Attributes and has **11033** observations.  
Each observation is a **Landslide**.

The list of all the Variables which are available for us to explore along with their types are listed below:

knitr::kable(attributes, "simple", col.names = c("Attribute Names", "Attribute Type"), align = c("l", "c"), caption = "Variables of the Dataset and their Types")

Variables of the Dataset and their Types

|  |  |
| --- | --- |
| Attribute Names | Attribute Type |
| source\_name | Text |
| source\_link | Website URL |
| event\_id | Number |
| event\_date | Date & Time |
| event\_time | Date & Time |
| event\_title | Text |
| event\_description | Text |
| location\_description | Text |
| location\_accuracy | Text |
| landslide\_category | Text |
| landslide\_trigger | Text |
| landslide\_size | Text |
| landslide\_setting | Text |
| fatality\_count | Number |
| injury\_count | Number |
| storm\_name | Text |
| photo\_link | Website URL |
| notes | Text |
| event\_import\_source | Text |
| event\_import\_id | Number |
| country\_name | Text |
| country\_code | Text |
| admin\_division\_name | Text |
| admin\_division\_population | Number |
| gazeteer\_closest\_point | Text |
| gazeteer\_distance | Number |
| submitted\_date | Date & Time |
| created\_date | Date & Time |
| last\_edited\_date | Date & Time |
| longitude | Number |
| latitude | Number |

However,we will be focusing only on the below Variables for my Project Analysis:

knitr::kable(attributes\_interest, "simple", col.names = c("Variable Names", "Variable Type"), align = c("l", "c"), caption = "Variables of the Dataset we will explore")

Variables of the Dataset we will explore

|  |  |
| --- | --- |
| Variable Names | Variable Type |
| source\_name | Text |
| event\_id | Number |
| event\_date | Date & Time |
| event\_time | Date & Time |
| event\_title | Text |
| event\_description | Text |
| location\_description | Text |
| location\_accuracy | Text |
| landslide\_category | Text |
| landslide\_trigger | Text |
| landslide\_size | Text |
| fatality\_count | Number |
| injury\_count | Number |
| storm\_name | Text |
| country\_name | Text |
| country\_code | Text |
| admin\_division\_name | Text |
| admin\_division\_population | Number |
| longitude | Number |
| latitude | Number |

# Goal

Through the below data analysis, we want to answer these questions:

1. How many people were killed in the largest landslide ever recorded?
2. Are the sizes of various landslides equally distributed in the dataset?
3. What are the countries with more than 50 injured recorded in any landslide?
4. Is there any correlation between the numerical variable?
5. Perform hypothesis testing to see if the mean of the fatality\_count of any two countries with the same number of landslides will be the same or not
6. Use Logistic Regression to predict the size of the landslide

# Analysis

## Loading the Package and Data

The first step is to load the necessary library, which contains all the datasets

If the package is not installed, please uncomment the below line of R code and execute it on your machine. The below statement installs the pre-requisite package

#install.packages("tidyverse")  
#install.packages("ggplot2")  
#install.packages("maps")

Now that the package is installed, we need to load the required library

# Loading required libraries #  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)

In the next step, we will create a local object that will hold our dataset.

The name of our local object will be “globalLandslide\_data”.

globalLandslide\_data <- as\_tibble(read.csv("D:\\SJSU\\1stSem\\Study\\ISE-201\\Submissions\\ProjectProposal-2\\originaldataset\\Global\_Landslide\_Catalog\_Export.csv"))

## Examining the Data

This section will use some basic steps to examine our data. Here we will also see if any changes are required in our dataset to make our work easy and smooth.

This stage can also be referred to as **“Cleaning the Data.”**

### 1. Evaluating the structure of the data

str(globalLandslide\_data)

## tibble [11,033 x 31] (S3: tbl\_df/tbl/data.frame)  
## $ source\_name : chr [1:11033] "AGU" "Oregonian" "CBS News" "Reuters" ...  
## $ source\_link : chr [1:11033] "https://blogs.agu.org/landslideblog/2008/10/14/the-lifan-landslide-from-natural-disaster-to-cover-up/" "http://www.oregonlive.com/news/index.ssf/2009/01/landslide\_plows\_through\_lake\_o.html" "https://www.cbsnews.com/news/dozens-missing-after-peru-landslides/" "https://in.reuters.com/article/idINIndia-41450420090731" ...  
## $ event\_id : int [1:11033] 684 956 973 1067 2603 4203 4290 225 236 873 ...  
## $ event\_date : chr [1:11033] "08/01/2008 12:00:00 AM" "01/02/2009 02:00:00 AM" "01/19/2007 12:00:00 AM" "07/31/2009 12:00:00 AM" ...  
## $ event\_time : logi [1:11033] NA NA NA NA NA NA ...  
## $ event\_title : chr [1:11033] "Sigou Village, Loufan County, Shanxi Province" "Lake Oswego, Oregon" "San Ramon district, 195 miles northeast of the capital, Lima, " "Dailekh district" ...  
## $ event\_description : chr [1:11033] "occurred early in morning, 11 villagers buried in 7 houses" "Hours of heavy rain are to blame for an overnight mudslide in Lake Oswego. " "(CBS/AP) At least 10 people died and as many as 80 were still missing Wednesday in central Peru after torrentia"| \_\_truncated\_\_ "One person was killed in Dailekh district, police said." ...  
## $ location\_description : chr [1:11033] "Sigou Village, Loufan County, Shanxi Province" "Lake Oswego, Oregon" "San Ramon district, 195 miles northeast of the capital, Lima, " "Dailekh district" ...  
## $ location\_accuracy : chr [1:11033] "unknown" "5km" "10km" "unknown" ...  
## $ landslide\_category : chr [1:11033] "landslide" "mudslide" "landslide" "landslide" ...  
## $ landslide\_trigger : chr [1:11033] "rain" "downpour" "downpour" "monsoon" ...  
## $ landslide\_size : chr [1:11033] "large" "small" "large" "medium" ...  
## $ landslide\_setting : chr [1:11033] "mine" "unknown" "unknown" "unknown" ...  
## $ fatality\_count : int [1:11033] 11 0 10 1 0 0 0 3 NA 2 ...  
## $ injury\_count : int [1:11033] NA NA NA NA NA NA NA NA NA NA ...  
## $ storm\_name : chr [1:11033] "" "" "" "" ...  
## $ photo\_link : chr [1:11033] "" "" "" "" ...  
## $ notes : chr [1:11033] "" "" "" "" ...  
## $ event\_import\_source : chr [1:11033] "glc" "glc" "glc" "glc" ...  
## $ event\_import\_id : num [1:11033] 684 956 973 1067 2603 ...  
## $ country\_name : chr [1:11033] "China" "United States" "Peru" "Nepal" ...  
## $ country\_code : chr [1:11033] "CN" "US" "PE" "NP" ...  
## $ admin\_division\_name : chr [1:11033] "Shaanxi" "Oregon" "JunÃ­n" "Mid Western" ...  
## $ admin\_division\_population: int [1:11033] 0 36619 14708 20908 798634 2404 2126 3191 2689 0 ...  
## $ gazeteer\_closest\_point : chr [1:11033] "Jingyang" "Lake Oswego" "San RamÃ³n" "Dailekh" ...  
## $ gazeteer\_distance : num [1:11033] 41.021 0.603 0.855 0.754 2.022 ...  
## $ submitted\_date : chr [1:11033] "04/01/2014 12:00:00 AM" "04/01/2014 12:00:00 AM" "04/01/2014 12:00:00 AM" "04/01/2014 12:00:00 AM" ...  
## $ created\_date : chr [1:11033] "11/20/2017 03:17:00 PM" "11/20/2017 03:17:00 PM" "11/20/2017 03:17:00 PM" "11/20/2017 03:17:00 PM" ...  
## $ last\_edited\_date : chr [1:11033] "02/15/2018 03:51:00 PM" "02/15/2018 03:51:00 PM" "02/15/2018 03:51:00 PM" "02/15/2018 03:51:00 PM" ...  
## $ longitude : num [1:11033] 107.5 -122.7 -75.4 81.7 123.9 ...  
## $ latitude : num [1:11033] 32.6 45.4 -11.1 28.8 10.3 ...

The output of the above query tells us the structure of the dataset.

We have a data frame with 11,033 observations on 31 variables. And the dataset is a mix of categorical, nominal, numerical, and continuous variables.

### 2. Peeking at the data

Looking at the first few observations in the dataframe

head(globalLandslide\_data)

## # A tibble: 6 x 31  
## source\_name source\_link event\_id event\_date event\_time event\_title   
## <chr> <chr> <int> <chr> <lgl> <chr>   
## 1 AGU https://blog~ 684 08/01/200~ NA "Sigou Vill~  
## 2 Oregonian http://www.o~ 956 01/02/200~ NA "Lake Osweg~  
## 3 CBS News https://www.~ 973 01/19/200~ NA "San Ramon ~  
## 4 Reuters https://in.r~ 1067 07/31/200~ NA "Dailekh di~  
## 5 The Freeman http://www.p~ 2603 10/16/201~ NA "sitio Baki~  
## 6 BusinessWorld Online http://www.b~ 4203 02/16/201~ NA "Paguite, A~  
## # ... with 25 more variables: event\_description <chr>,  
## # location\_description <chr>, location\_accuracy <chr>,  
## # landslide\_category <chr>, landslide\_trigger <chr>, landslide\_size <chr>,  
## # landslide\_setting <chr>, fatality\_count <int>, injury\_count <int>,  
## # storm\_name <chr>, photo\_link <chr>, notes <chr>, event\_import\_source <chr>,  
## # event\_import\_id <dbl>, country\_name <chr>, country\_code <chr>,  
## # admin\_division\_name <chr>, admin\_division\_population <int>, ...

Looking at the last few observations in the dataset

tail(globalLandslide\_data)

## # A tibble: 6 x 31  
## source\_name source\_link event\_id event\_date event\_time event\_title   
## <chr> <chr> <int> <chr> <lgl> <chr>   
## 1 St. Maries G~ http://www.gazet~ 10518 03/23/2017~ NA Mudslide abov~  
## 2 The Jakarta ~ http://www.theja~ 11109 04/01/2017~ NA Major landsli~  
## 3 Greater Kash~ http://www.great~ 10845 03/25/2017~ NA Barnari Sigdi~  
## 4 NBC Daily http://www.nbcda~ 10973 12/15/2016~ NA Landslide at ~  
## 5 AGU Landslid~ http://blogs.agu~ 10901 04/29/2017~ NA Mayor landsli~  
## 6 The Times of~ https://timesofi~ 10949 03/13/2017~ NA Kondapur Comm~  
## # ... with 25 more variables: event\_description <chr>,  
## # location\_description <chr>, location\_accuracy <chr>,  
## # landslide\_category <chr>, landslide\_trigger <chr>, landslide\_size <chr>,  
## # landslide\_setting <chr>, fatality\_count <int>, injury\_count <int>,  
## # storm\_name <chr>, photo\_link <chr>, notes <chr>, event\_import\_source <chr>,  
## # event\_import\_id <dbl>, country\_name <chr>, country\_code <chr>,  
## # admin\_division\_name <chr>, admin\_division\_population <int>, ...

### 3. Checking the summary of the Tibble

summary(globalLandslide\_data)

## source\_name source\_link event\_id event\_date   
## Length:11033 Length:11033 Min. : 1 Length:11033   
## Class :character Class :character 1st Qu.: 2785 Class :character   
## Mode :character Mode :character Median : 5563 Mode :character   
## Mean : 5599   
## 3rd Qu.: 8435   
## Max. :11221   
##   
## event\_time event\_title event\_description location\_description  
## Mode:logical Length:11033 Length:11033 Length:11033   
## NA's:11033 Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## location\_accuracy landslide\_category landslide\_trigger landslide\_size   
## Length:11033 Length:11033 Length:11033 Length:11033   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## landslide\_setting fatality\_count injury\_count storm\_name   
## Length:11033 Min. : 0.000 Min. : 0.000 Length:11033   
## Class :character 1st Qu.: 0.000 1st Qu.: 0.000 Class :character   
## Mode :character Median : 0.000 Median : 0.000 Mode :character   
## Mean : 3.219 Mean : 0.752   
## 3rd Qu.: 1.000 3rd Qu.: 0.000   
## Max. :5000.000 Max. :374.000   
## NA's :1385 NA's :5674   
## photo\_link notes event\_import\_source event\_import\_id   
## Length:11033 Length:11033 Length:11033 Min. :-111.2   
## Class :character Class :character Class :character 1st Qu.:2386.5   
## Mode :character Mode :character Mode :character Median :4773.0   
## Mean :4798.6   
## 3rd Qu.:7189.5   
## Max. :9669.0   
## NA's :1562   
## country\_name country\_code admin\_division\_name  
## Length:11033 Length:11033 Length:11033   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## admin\_division\_population gazeteer\_closest\_point gazeteer\_distance  
## Min. : 0 Length:11033 Min. : 0.000   
## 1st Qu.: 1963 Class :character 1st Qu.: 2.364   
## Median : 7365 Mode :character Median : 6.255   
## Mean : 157760 Mean : 11.874   
## 3rd Qu.: 34021 3rd Qu.: 15.816   
## Max. :12691836 Max. :215.449   
## NA's :1562 NA's :1562   
## submitted\_date created\_date last\_edited\_date longitude   
## Length:11033 Length:11033 Length:11033 Min. :-179.98   
## Class :character Class :character Class :character 1st Qu.:-107.87   
## Mode :character Mode :character Mode :character Median : 19.69   
## Mean : 2.52   
## 3rd Qu.: 93.95   
## Max. : 179.99   
##   
## latitude   
## Min. :-46.77   
## 1st Qu.: 13.92   
## Median : 30.53   
## Mean : 25.88   
## 3rd Qu.: 40.87   
## Max. : 72.63   
##

## Quality Check and Data Cleaning

In this section, we will be checking the quality of our data and cleaning our dataset, if required.

### 1. Checking the Variable Name

# Checking the names of the Columns   
names(globalLandslide\_data)

## [1] "source\_name" "source\_link"   
## [3] "event\_id" "event\_date"   
## [5] "event\_time" "event\_title"   
## [7] "event\_description" "location\_description"   
## [9] "location\_accuracy" "landslide\_category"   
## [11] "landslide\_trigger" "landslide\_size"   
## [13] "landslide\_setting" "fatality\_count"   
## [15] "injury\_count" "storm\_name"   
## [17] "photo\_link" "notes"   
## [19] "event\_import\_source" "event\_import\_id"   
## [21] "country\_name" "country\_code"   
## [23] "admin\_division\_name" "admin\_division\_population"  
## [25] "gazeteer\_closest\_point" "gazeteer\_distance"   
## [27] "submitted\_date" "created\_date"   
## [29] "last\_edited\_date" "longitude"   
## [31] "latitude"

Our variable names are meaningful, which is great. In addition, the clear variable names help us know the feature we are working on.

### 2. Removing the Columns which we are not including in our analysis

Let us remove the columns which we are not going to focus on for our further analysis:

globalLandslide\_df <- globalLandslide\_data[!(colnames(globalLandslide\_data) %in% c("source\_link", "landslide\_setting", "photo\_link", "notes",   
 "event\_import\_source", "event\_import\_id", "gazeteer\_closest\_point", "gazeteer\_distance", "submitted\_date", "created\_date",  
 "last\_edited\_date"))]

Looking at the structure of the dataframe we will be proceeding with:

str(globalLandslide\_df)

## tibble [11,033 x 20] (S3: tbl\_df/tbl/data.frame)  
## $ source\_name : chr [1:11033] "AGU" "Oregonian" "CBS News" "Reuters" ...  
## $ event\_id : int [1:11033] 684 956 973 1067 2603 4203 4290 225 236 873 ...  
## $ event\_date : chr [1:11033] "08/01/2008 12:00:00 AM" "01/02/2009 02:00:00 AM" "01/19/2007 12:00:00 AM" "07/31/2009 12:00:00 AM" ...  
## $ event\_time : logi [1:11033] NA NA NA NA NA NA ...  
## $ event\_title : chr [1:11033] "Sigou Village, Loufan County, Shanxi Province" "Lake Oswego, Oregon" "San Ramon district, 195 miles northeast of the capital, Lima, " "Dailekh district" ...  
## $ event\_description : chr [1:11033] "occurred early in morning, 11 villagers buried in 7 houses" "Hours of heavy rain are to blame for an overnight mudslide in Lake Oswego. " "(CBS/AP) At least 10 people died and as many as 80 were still missing Wednesday in central Peru after torrentia"| \_\_truncated\_\_ "One person was killed in Dailekh district, police said." ...  
## $ location\_description : chr [1:11033] "Sigou Village, Loufan County, Shanxi Province" "Lake Oswego, Oregon" "San Ramon district, 195 miles northeast of the capital, Lima, " "Dailekh district" ...  
## $ location\_accuracy : chr [1:11033] "unknown" "5km" "10km" "unknown" ...  
## $ landslide\_category : chr [1:11033] "landslide" "mudslide" "landslide" "landslide" ...  
## $ landslide\_trigger : chr [1:11033] "rain" "downpour" "downpour" "monsoon" ...  
## $ landslide\_size : chr [1:11033] "large" "small" "large" "medium" ...  
## $ fatality\_count : int [1:11033] 11 0 10 1 0 0 0 3 NA 2 ...  
## $ injury\_count : int [1:11033] NA NA NA NA NA NA NA NA NA NA ...  
## $ storm\_name : chr [1:11033] "" "" "" "" ...  
## $ country\_name : chr [1:11033] "China" "United States" "Peru" "Nepal" ...  
## $ country\_code : chr [1:11033] "CN" "US" "PE" "NP" ...  
## $ admin\_division\_name : chr [1:11033] "Shaanxi" "Oregon" "JunÃ­n" "Mid Western" ...  
## $ admin\_division\_population: int [1:11033] 0 36619 14708 20908 798634 2404 2126 3191 2689 0 ...  
## $ longitude : num [1:11033] 107.5 -122.7 -75.4 81.7 123.9 ...  
## $ latitude : num [1:11033] 32.6 45.4 -11.1 28.8 10.3 ...

We will be proceeding with globalLandslide\_df, which has **11,033 observations on 20 variables.**

### 3. Checking for missing data

# Calculate the total numbers of "Not Available" data   
sum(is.na(globalLandslide\_df))

## [1] 19656

There are many cells with missing data in our dataset. Let’s see which columns do not have data in them

names(which(colSums(is.na(globalLandslide\_df)) > 0))

## [1] "event\_time" "fatality\_count"   
## [3] "injury\_count" "country\_code"   
## [5] "admin\_division\_population"

### 4. Handling Missing Data

Let us address the above columns with missing data one-by-one

# Handling missing data in "injury\_count" column  
globalLandslide\_df$injury\_count[is.na(globalLandslide\_df$injury\_count)] = 0  
  
# Handling missing data in "admin\_division\_population" column  
globalLandslide\_df$admin\_division\_population[is.na(globalLandslide\_df$admin\_division\_population)] = 0  
  
# Handling missing data in "fatality\_count" column  
globalLandslide\_df[!is.na(globalLandslide\_df$fatality\_count),]

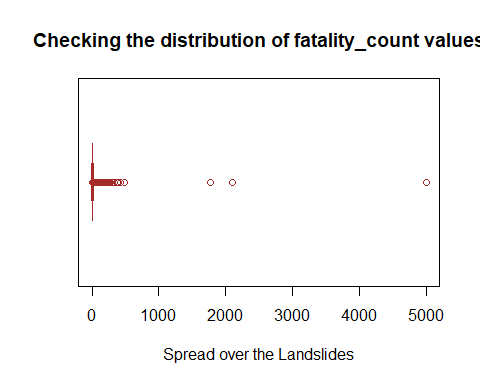
## # A tibble: 9,648 x 20  
## source\_name event\_id event\_date event\_time event\_title event\_descripti~  
## <chr> <int> <chr> <lgl> <chr> <chr>   
## 1 AGU 684 08/01/200~ NA "Sigou Vil~ "occurred early~  
## 2 Oregonian 956 01/02/200~ NA "Lake Oswe~ "Hours of heavy~  
## 3 CBS News 973 01/19/200~ NA "San Ramon~ "(CBS/AP) At le~  
## 4 Reuters 1067 07/31/200~ NA "Dailekh d~ "One person was~  
## 5 The Freeman 2603 10/16/201~ NA "sitio Bak~ "Another landsl~  
## 6 BusinessWorld Online 4203 02/16/201~ NA "Paguite, ~ "Thursdayâ€™s l~  
## 7 The Spokesman-Review 4290 03/30/201~ NA "Pend Orei~ "In Pend Oreill~  
## 8 CrÃ³nica Diaria 225 09/02/200~ NA "3 killed ~ "3 killed, incl~  
## 9 UPI 873 11/01/200~ NA "Lincang C~ "The report sai~  
## 10 BBC News 874 11/01/200~ NA "Kunming, ~ "Yunnan has so ~  
## # ... with 9,638 more rows, and 14 more variables: location\_description <chr>,  
## # location\_accuracy <chr>, landslide\_category <chr>, landslide\_trigger <chr>,  
## # landslide\_size <chr>, fatality\_count <int>, injury\_count <dbl>,  
## # storm\_name <chr>, country\_name <chr>, country\_code <chr>,  
## # admin\_division\_name <chr>, admin\_division\_population <dbl>,  
## # longitude <dbl>, latitude <dbl>

# Handling missing data for "country\_name"  
globalLandslide\_df <- globalLandslide\_df[grep('^[A-Za-z]', globalLandslide\_df$country\_name),]

### 5. Checking for outliers

To check if there are any outliers in our data-set, we will be using boxplots to easily view them.

## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some  
## notches went outside hinges ('box'): maybe set notch=FALSE



By looking at the above plot, we see that although the death count because of the Landslides over all the years was below 1000, there was one where the fatality count was 5000.  
It must have been an enormous disaster.

Hence, this solves our 1st Question: **5000 people were killed in the largest landslide disaster ever recorded**

## Data Transformation

### 1. Separating the Event\_Date to Date and Time

We saw in the above section that one of the columns with missing value is “event\_time”.

Let’s check how many cells of the “event\_time” are empty.

sum(is.na(globalLandslide\_df$event\_time))

## [1] 9471

There is no data in the event\_time variable as 11033 cells are empty.

This transformation step will get the time value from the “event\_date” variable.

# Checking the values of the event\_date column  
head(globalLandslide\_df$event\_date, 5)

## [1] "08/01/2008 12:00:00 AM" "01/02/2009 02:00:00 AM" "01/19/2007 12:00:00 AM"  
## [4] "07/31/2009 12:00:00 AM" "10/16/2010 12:00:00 PM"

By just checking the 5 records of the “event\_date”, we can say that it also contains the time. So, we will be using the Solution-2 in this Data Transformation Step

Let’s separate the time from the “event\_date” and store it into the “event\_time” column

# Splitting the event\_date by the first space  
ev\_dates <- as.data.frame(str\_split\_fixed(globalLandslide\_df$event\_date, " ", 2))  
colnames(ev\_dates) <- c("DATE", "TIME")  
head(ev\_dates, 10)

## DATE TIME  
## 1 08/01/2008 12:00:00 AM  
## 2 01/02/2009 02:00:00 AM  
## 3 01/19/2007 12:00:00 AM  
## 4 07/31/2009 12:00:00 AM  
## 5 10/16/2010 12:00:00 PM  
## 6 02/16/2012 12:00:00 AM  
## 7 03/30/2012 12:00:00 AM  
## 8 09/02/2007 12:00:00 AM  
## 9 09/05/2007 12:00:00 AM  
## 10 11/01/2008 12:00:00 AM

Now storing the “TIME” column of y to our “event\_time” and “DATE” column of y to our event\_date

globalLandslide\_df$event\_date <- ev\_dates$DATE  
globalLandslide\_df$event\_time <- ev\_dates$TIME

Now, let’s check the missing data in the event\_time

sum(is.na(globalLandslide\_df$event\_time))

## [1] 0

In the above step, we performed Data-Transformation for the “event\_date” and “event\_time” variable.

# Removing the landslide size which is 'unknown'  
globalLandslide\_df <- globalLandslide\_df[!grepl('unknown',globalLandslide\_df$landslide\_size),]

## Visual and Descriptive Analysis

### 1. Analyzing single Categorical Variables

Now, let us start by analyzing our 2 Categorical Variables - “landslide\_size” and “landslide\_category.”

The below query returns a table of frequency occurrences of data in each “landslide\_size” category

table(globalLandslide\_df$landslide\_size)

##   
## large medium small very\_large   
## 632 6039 1908 86

We can see that our data is not equally distributed among the different sizes of landslides by the output. And there are nine landslides whose size is not determined in our dataset.

### Let’s see this in a Visual format.

# Create a bar graph of wool observations  
ggplot(data = globalLandslide\_df) +   
 geom\_bar(mapping = aes(x = landslide\_size))



The above bar graph clarifies that the highest number of landslides that occurred so far is **medium-sized.** The size of the maximum occurring landslides in our dataset is \*\* 2 categories of Wool are equally distributed in our dataset. There are nine landslides with no measure in our dataset.

**This answers Question 2. No, the landslide sizes are not equally distributed in our dataset.**

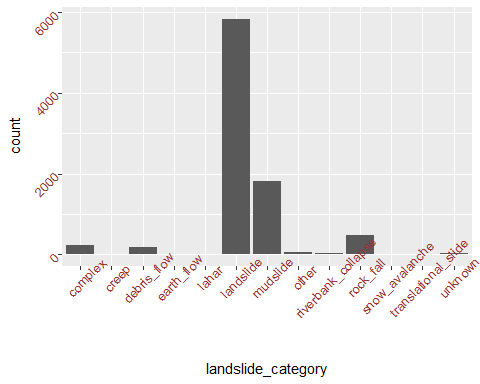
In the above plot, we can see that the number of very\_large landslides is minimal. So, we can easily merge the large and very\_large landslides as “large.”

globalLandslide\_df$landslide\_size[globalLandslide\_df$landslide\_size == "very\_large"] <- "large"

### 2. Analyzing the spread of landslides based on the category

Now, let us plot the landslide based on the category

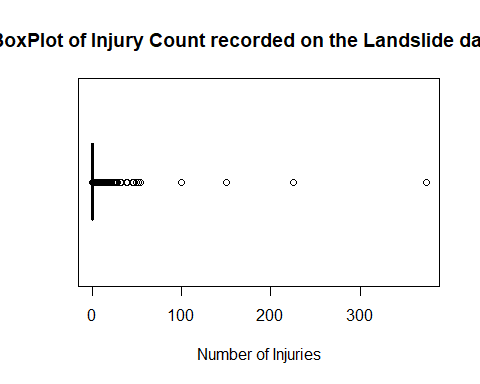
# Create a bar graph of wool observations  
  
p <- ggplot(data = globalLandslide\_df) +   
 geom\_bar(mapping = aes(x = landslide\_category))  
  
  
p + theme(axis.text.x = element\_text(color="#993333",   
 size=10, angle=45),  
 axis.text.y = element\_text(color="#993333",   
 size=10, angle=45))



From the above plot, we can see that most landslides are categorized as landslides. We also have significant mudslides and rock-fall types of landslides in our dataset.

### 3. Analyzing the number of injuries

# Create a boxplot of breaks  
boxplot(  
 x = globalLandslide\_df$injury\_count,  
 xlab = "Number of Injuries",  
 horizontal = TRUE,  
 main = "BoxPlot of Injury Count recorded on the Landslide dataset"  
)



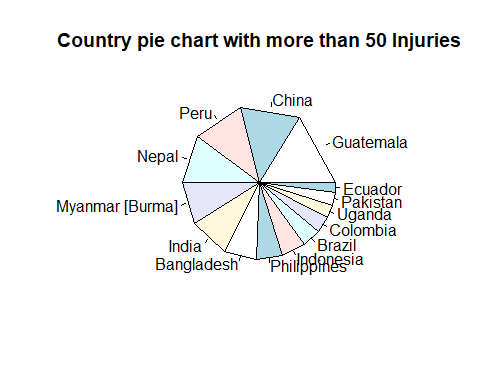
The above Box Plot shows the spread of the injury count across our data frame

### 3. Plotting only those coutries with total injury more than 50 on a Pie Plot

landslide\_distribution <- globalLandslide\_df %>%   
 group\_by(country\_name) %>%   
 summarise(sum\_injuries = sum(injury\_count)) %>%   
 arrange(-sum\_injuries)  
  
landslide\_distribution

## # A tibble: 141 x 2  
## country\_name sum\_injuries  
## <chr> <dbl>  
## 1 Guatemala 408  
## 2 China 318  
## 3 Peru 277  
## 4 Nepal 257  
## 5 Myanmar [Burma] 224  
## 6 India 217  
## 7 Bangladesh 170  
## 8 Philippines 138  
## 9 Indonesia 129  
## 10 Brazil 103  
## # ... with 131 more rows

graph\_data <- landslide\_distribution[apply(landslide\_distribution[,-1], 1, function(x) !all(x<=50)),]  
x <- c(graph\_data$sum\_injuries)  
yy <- c(graph\_data$country\_name)  
  
pie(x, yy, main = "Country pie chart with more than 50 Injuries", edges = 10)



The above plot shows the countries which have more than 50 injury\_count in any of the landslides.

## Hypothesis testing

In this section, we will move on to finding further answers to the hypothesis question listed above.

Now, let’s find out about the variables with character datatypes

data\_char<-globalLandslide\_df %>% dplyr::select(where(is.character))  
  
for(i in colnames(data\_char)){  
 print(unique(data\_char[i]))  
}

## # A tibble: 3,319 x 1  
## source\_name   
## <chr>   
## 1 AGU   
## 2 Oregonian   
## 3 CBS News   
## 4 Reuters   
## 5 The Freeman   
## 6 BusinessWorld Online  
## 7 The Spokesman-Review  
## 8 CrÃ³nica Diaria   
## 9 MagicValley.com   
## 10 UPI   
## # ... with 3,309 more rows  
## # A tibble: 2,670 x 1  
## event\_date  
## <chr>   
## 1 08/01/2008  
## 2 01/02/2009  
## 3 01/19/2007  
## 4 07/31/2009  
## 5 10/16/2010  
## 6 02/16/2012  
## 7 03/30/2012  
## 8 09/02/2007  
## 9 09/05/2007  
## 10 11/01/2008  
## # ... with 2,660 more rows  
## # A tibble: 269 x 1  
## event\_time   
## <chr>   
## 1 12:00:00 AM  
## 2 02:00:00 AM  
## 3 12:00:00 PM  
## 4 10:24:00 PM  
## 5 08:30:00 PM  
## 6 01:41:00 AM  
## 7 01:00:00 AM  
## 8 06:00:00 AM  
## 9 08:50:00 AM  
## 10 01:00:00 PM  
## # ... with 259 more rows  
## # A tibble: 8,406 x 1  
## event\_title   
## <chr>   
## 1 "Sigou Village, Loufan County, Shanxi Province"   
## 2 "Lake Oswego, Oregon"   
## 3 "San Ramon district, 195 miles northeast of the capital, Lima, "  
## 4 "Dailekh district"   
## 5 "sitio Bakilid in barangay Lahug"   
## 6 "Paguite, Abuyog, Leyte"   
## 7 "Pend Oreille County, State Route 20 near Usk, OR"   
## 8 "3 killed in Acapulco"   
## 9 "Warm Springs Road, Idaho"   
## 10 "Lincang City, Yunnan, Yunnan-Tibet No. 214 highway."   
## # ... with 8,396 more rows  
## # A tibble: 7,879 x 1  
## event\_description   
## <chr>   
## 1 "occurred early in morning, 11 villagers buried in 7 houses"   
## 2 "Hours of heavy rain are to blame for an overnight mudslide in Lake Oswego. "  
## 3 "(CBS/AP) At least 10 people died and as many as 80 were still missing Wedne~  
## 4 "One person was killed in Dailekh district, police said."   
## 5 "Another landslide in sitio Bakilid in barangay Lahug also left two families~  
## 6 "Thursdayâ€™s landslides were noted in Barangays Burubudan, Tadoc and Paguit~  
## 7 "In Pend Oreille County, a mudslide on State Route 20 near Usk forced Washin~  
## 8 "3 killed, including 2 children when rocks fell on their homes"   
## 9 "5 feet deep mud, Hotshot crew was trapped while cleaning debris from fire"   
## 10 "The report said heavy rainfall since Oct. 24 had hit 13 cities and counties~  
## # ... with 7,869 more rows  
## # A tibble: 8,322 x 1  
## location\_description   
## <chr>   
## 1 "Sigou Village, Loufan County, Shanxi Province"   
## 2 "Lake Oswego, Oregon"   
## 3 "San Ramon district, 195 miles northeast of the capital, Lima, "  
## 4 "Dailekh district"   
## 5 "sitio Bakilid in barangay Lahug"   
## 6 "Paguite, Abuyog, Leyte"   
## 7 "Pend Oreille County, State Route 20 near Usk, OR"   
## 8 "calle Granjas, AmpliaciÃ³n Miguel de la Madrid, Acapulco"   
## 9 "Warm Springs Road, Idaho"   
## 10 "Lincang City, Yunnan, Yunnan-Tibet No. 214 highway."   
## # ... with 8,312 more rows  
## # A tibble: 9 x 1  
## location\_accuracy  
## <chr>   
## 1 unknown   
## 2 5km   
## 3 10km   
## 4 25km   
## 5 1km   
## 6 50km   
## 7 exact   
## 8 100km   
## 9 250km   
## # A tibble: 13 x 1  
## landslide\_category   
## <chr>   
## 1 landslide   
## 2 mudslide   
## 3 complex   
## 4 rock\_fall   
## 5 debris\_flow   
## 6 riverbank\_collapse   
## 7 unknown   
## 8 lahar   
## 9 other   
## 10 snow\_avalanche   
## 11 creep   
## 12 earth\_flow   
## 13 translational\_slide  
## # A tibble: 16 x 1  
## landslide\_trigger   
## <chr>   
## 1 rain   
## 2 downpour   
## 3 monsoon   
## 4 tropical\_cyclone   
## 5 unknown   
## 6 continuous\_rain   
## 7 mining   
## 8 no\_apparent\_trigger   
## 9 snowfall\_snowmelt   
## 10 flooding   
## 11 dam\_embankment\_collapse  
## 12 earthquake   
## 13 construction   
## 14 other   
## 15 volcano   
## 16 freeze\_thaw   
## # A tibble: 3 x 1  
## landslide\_size  
## <chr>   
## 1 large   
## 2 small   
## 3 medium   
## # A tibble: 201 x 1  
## storm\_name   
## <chr>   
## 1 ""   
## 2 "Supertyphoon Juan (Megi)"   
## 3 "Tropical Storm Henrietta"   
## 4 "Hurricane Dora"   
## 5 "Agaton"   
## 6 "Typhoon Nina"   
## 7 "Tropical Depression 16"   
## 8 "Typhoon No. 2 and March 11th earthquake"  
## 9 "Typhoon Nepartak"   
## 10 "Tropical Storm Alma"   
## # ... with 191 more rows  
## # A tibble: 141 x 1  
## country\_name   
## <chr>   
## 1 China   
## 2 United States  
## 3 Peru   
## 4 Nepal   
## 5 Philippines   
## 6 Mexico   
## 7 Algeria   
## 8 Malaysia   
## 9 Indonesia   
## 10 Sierra Leone   
## # ... with 131 more rows  
## # A tibble: 140 x 1  
## country\_code  
## <chr>   
## 1 CN   
## 2 US   
## 3 PE   
## 4 NP   
## 5 PH   
## 6 MX   
## 7 DZ   
## 8 MY   
## 9 ID   
## 10 SL   
## # ... with 130 more rows  
## # A tibble: 888 x 1  
## admin\_division\_name  
## <chr>   
## 1 Shaanxi   
## 2 Oregon   
## 3 JunÃ­n   
## 4 Mid Western   
## 5 Central Visayas   
## 6 Eastern Visayas   
## 7 Washington   
## 8 Sinaloa   
## 9 Idaho   
## 10 Yunnan   
## # ... with 878 more rows

### Studying correlation

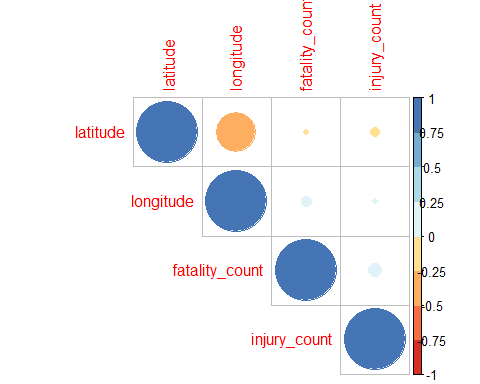
In order to study the correlation in the dataset, let’s take the numerical variables of importance

numerical\_data <- globalLandslide\_df[, c('fatality\_count', 'injury\_count', 'longitude', 'latitude')] # Numerical variables  
  
# Removing na values  
numerical\_data <- na.omit(numerical\_data)

library(corrplot)

## corrplot 0.90 loaded

library(RColorBrewer)  
corr <-cor(numerical\_data)  
corrplot(corr, type="upper", order="hclust",  
 col=brewer.pal(n=8, name="RdYlBu"))



The above plot solves answers our Question-3. There is a very minimum correlation between the fatality\_count and injury\_count. However, the only correlation we can find is between latitude and longitude.

### Hypothesis 1: Is the mean value of fatality\_count in one country equal to the mean value of fatality\_count of another country with the same number of landslide

First, let’s take the frequency table of the various countries

country\_tbl <- table(data\_char$country\_name)  
country\_tbl <- sort(country\_tbl, decreasing = T)  
country\_tbl

##   
## United States India   
## 2224 1261   
## Philippines Nepal   
## 669 479   
## China Indonesia   
## 425 350   
## United Kingdom Brazil   
## 225 214   
## Canada Malaysia   
## 173 166   
## Pakistan Vietnam   
## 141 116   
## New Zealand Australia   
## 106 105   
## Colombia Mexico   
## 101 86   
## Guatemala Japan   
## 82 82   
## Thailand Costa Rica   
## 77 76   
## Sri Lanka Taiwan   
## 75 66   
## Trinidad and Tobago Bangladesh   
## 65 58   
## Peru Italy   
## 58 55   
## Kenya Uganda   
## 53 45   
## Panama Myanmar [Burma]   
## 44 43   
## Georgia Honduras   
## 39 39   
## Jamaica Bulgaria   
## 37 35   
## Fiji Ecuador   
## 34 32   
## Kyrgyzstan Nicaragua   
## 32 31   
## Ireland El Salvador   
## 23 22   
## France Haiti   
## 22 22   
## Norway Azerbaijan   
## 22 21   
## Papua New Guinea South Africa   
## 21 21   
## Turkey Venezuela   
## 21 21   
## Bhutan Tajikistan   
## 20 20   
## Switzerland Dominican Republic   
## 19 17   
## Dominica Afghanistan   
## 16 15   
## Nigeria Brunei   
## 15 14   
## Chile Spain   
## 14 14   
## Russia Bosnia and Herzegovina   
## 13 12   
## Bolivia South Korea   
## 11 11   
## Austria Lebanon   
## 10 9   
## Saint Lucia Argentina   
## 8 7   
## Ghana Ivory Coast   
## 7 7   
## Madagascar Puerto Rico   
## 7 7   
## Sierra Leone Yemen   
## 7 7   
## Iran Portugal   
## 6 6   
## Rwanda American Samoa   
## 6 5   
## Greece Saint Vincent and the Grenadines   
## 5 5   
## Saudi Arabia Serbia   
## 5 5   
## Solomon Islands Tanzania   
## 5 5   
## Armenia Cameroon   
## 4 4   
## Guinea Iceland   
## 4 4   
## Isle of Man Laos   
## 4 4   
## Macedonia North Korea   
## 4 4   
## Angola Cuba   
## 3 3   
## Democratic Republic of the Congo Germany   
## 3 3   
## Romania Ukraine   
## 3 3   
## Bermuda Croatia   
## 2 2   
## Czechia East Timor   
## 2 2   
## Ethiopia Grenada   
## 2 2   
## Hong Kong Israel   
## 2 2   
## Liberia Luxembourg   
## 2 2   
## Namibia Poland   
## 2 2   
## Slovakia U.S. Virgin Islands   
## 2 2   
## Vanuatu Albania   
## 2 1   
## Algeria Barbados   
## 1 1   
## Belize Burkina Faso   
## 1 1   
## Burundi Cambodia   
## 1 1   
## Czech Republic Egypt   
## 1 1   
## Gabon Guam   
## 1 1   
## Jersey Jordan   
## 1 1   
## Kazakhstan Malawi   
## 1 1   
## Mauritius Mongolia   
## 1 1   
## Montenegro Morocco   
## 1 1   
## Oman Paraguay   
## 1 1   
## Republic of the Congo Saint Kitts and Nevis   
## 1 1   
## Singapore Slovenia   
## 1 1   
## Sudan Swaziland   
## 1 1   
## United Arab Emirates Uzbekistan   
## 1 1   
## Zambia   
## 1

Above is the frequency count of how many times a country had a landslide recorded.

#### CASE-1

Looking at the above distribution, we see that the ‘Guatemala’ and ‘Japan’ had same number of landslides Now, let’s see if there is any relation between their fatality counts

data\_globalLandslide\_Japan <- filter(globalLandslide\_df,country\_name== "Japan")  
data\_globalLandslide\_Guatemala <- filter(globalLandslide\_df, country\_name =="Guatemala")

Now, let’s perform the T-Test

(mean(data\_globalLandslide\_Japan$fatality\_count,na.rm=TRUE))

## [1] 3.097222

(mean(data\_globalLandslide\_Guatemala$fatality\_count,na.rm=TRUE))

## [1] 9.525641

(var(data\_globalLandslide\_Japan$fatality\_count,na.rm=TRUE))

## [1] 99.52563

(var(data\_globalLandslide\_Guatemala$fatality\_count,na.rm=TRUE))

## [1] 1789.707

(t.test(data\_globalLandslide\_Guatemala$fatality\_count,data\_globalLandslide\_Japan$fatality\_count, alternative = "two.sided", var.equal = FALSE))

##   
## Welch Two Sample t-test  
##   
## data: data\_globalLandslide\_Guatemala$fatality\_count and data\_globalLandslide\_Japan$fatality\_count  
## t = 1.3033, df = 86.218, p-value = 0.1959  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.376251 16.233089  
## sample estimates:  
## mean of x mean of y   
## 9.525641 3.097222

By looking at the mean of X and Y, we can say that although the number of landslides was the same, the mean of fatality count of these two countries is very different, and the alternative hypothesis is true

#### CASE-2

Now, let’s take another two countries with same number of landslides - Afghanistan and Nigeria with 15 landslides

data\_globalLandslide\_Afghanistan <- filter(globalLandslide\_df,country\_name== "Afghanistan")  
data\_globalLandslide\_Nigeria <- filter(globalLandslide\_df, country\_name =="Nigeria")

Now, let’s perform the T-Test on the injury\_count

(mean(data\_globalLandslide\_Afghanistan$fatality\_count,na.rm=TRUE))

## [1] 191.1667

(mean(data\_globalLandslide\_Nigeria$fatality\_count,na.rm=TRUE))

## [1] 2.272727

(var(data\_globalLandslide\_Afghanistan$fatality\_count,na.rm=TRUE))

## [1] 362483.2

(var(data\_globalLandslide\_Nigeria$fatality\_count,na.rm=TRUE))

## [1] 6.418182

(t.test(data\_globalLandslide\_Afghanistan$fatality\_count,data\_globalLandslide\_Nigeria$fatality\_count, alternative = "two.sided", var.equal = FALSE))

##   
## Welch Two Sample t-test  
##   
## data: data\_globalLandslide\_Afghanistan$fatality\_count and data\_globalLandslide\_Nigeria$fatality\_count  
## t = 1.0868, df = 11, p-value = 0.3004  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -193.6423 571.4302  
## sample estimates:  
## mean of x mean of y   
## 191.166667 2.272727

Here, as well there is a big difference on the fatality\_count of the two countries And the alternative hypothesis is true

### Fitting a Logistic Regression model to predict landslide\_size based on Country name and fatality\_count

sizeTbl <- table(data\_char$landslide\_size)  
sizeTbl <- sort(sizeTbl, decreasing = T)  
sizeTbl

##   
## medium small large   
## 6039 1908 718

**As we have 3 categories in our landslide\_size, we will be using “multinorm”**

require(nnet)

## Loading required package: nnet

globalLandslide\_df$landslide\_size <- as.factor(globalLandslide\_df$landslide\_size)  
  
globalLandslide\_df$landslide\_size <- relevel(globalLandslide\_df$landslide\_size, ref = "small")  
  
  
(test <- multinom(landslide\_size ~ fatality\_count+country\_name, data = globalLandslide\_df))

## # weights: 408 (270 variable)  
## initial value 8048.433627   
## iter 10 value 5038.296265  
## iter 20 value 4870.109640  
## iter 30 value 4803.653120  
## iter 40 value 4768.915852  
## iter 50 value 4731.029628  
## iter 60 value 4727.591801  
## iter 70 value 4724.689782  
## iter 80 value 4723.435549  
## iter 90 value 4723.303577  
## iter 100 value 4723.088146  
## final value 4723.088146   
## stopped after 100 iterations

## Call:  
## multinom(formula = landslide\_size ~ fatality\_count + country\_name,   
## data = globalLandslide\_df)  
##   
## Coefficients:  
## (Intercept) fatality\_count country\_nameAlbania country\_nameAlgeria  
## large -0.578098 0.2300878 -2.515561 -3.067708  
## medium 1.543283 0.1469364 7.845234 8.521871  
## country\_nameAmerican Samoa country\_nameAngola country\_nameArgentina  
## large -3.25849 -3.196976 12.59746  
## medium 12.07988 11.332459 12.21304  
## country\_nameArmenia country\_nameAustralia country\_nameAustria  
## large 0.4118777 -2.2002789 -3.294978  
## medium -0.9465639 -0.9870048 12.970072  
## country\_nameAzerbaijan country\_nameBangladesh country\_nameBelize  
## large 0.5367374 -1.756978 -15.46752  
## medium 0.5124465 -1.428416 -25.56755  
## country\_nameBermuda country\_nameBhutan country\_nameBolivia  
## large -15.449782 -13.055964 15.89792  
## medium -1.543018 0.794617 14.38909  
## country\_nameBosnia and Herzegovina country\_nameBrazil country\_nameBrunei  
## large -12.9279139 0.7411658 -16.508627  
## medium 0.7996407 1.3354296 -0.479314  
## country\_nameBulgaria country\_nameBurkina Faso country\_nameBurundi  
## large -1.788034 14.378477 -3.438488  
## medium -0.937253 -6.785308 8.971612  
## country\_nameCambodia country\_nameCameroon country\_nameCanada  
## large -2.792770 -3.014343 -1.549402  
## medium 8.187121 10.134804 -1.306708  
## country\_nameChile country\_nameChina country\_nameColombia  
## large -12.9914060 0.6695142 1.4672715  
## medium 0.7178465 0.7860384 0.6235825  
## country\_nameCosta Rica country\_nameCroatia country\_nameCuba  
## large -0.9082834 -3.014343 -3.196976  
## medium -0.4974509 10.134804 11.332459  
## country\_nameCzech Republic country\_nameCzechia  
## large -2.515561 -15.449782  
## medium 7.845234 -1.543018  
## country\_nameDemocratic Republic of the Congo country\_nameDominica  
## large 7.383624 -0.4178538  
## medium 7.590342 -0.8044751  
## country\_nameDominican Republic country\_nameEast Timor  
## large -2.4087788 -3.765105  
## medium -0.2519003 10.815946  
## country\_nameEcuador country\_nameEgypt country\_nameEl Salvador  
## large -0.6601672 -5.894841 -0.8680328  
## medium -0.4858372 12.308950 -0.5965775  
## country\_nameEthiopia country\_nameFiji country\_nameFrance  
## large 8.082781 -0.7133941 -1.485874  
## medium 8.497245 -0.8018322 -1.075178  
## country\_nameGabon country\_nameGeorgia country\_nameGermany  
## large -2.792770 -0.200735808 -3.070858  
## medium 8.187121 0.009777963 10.191661  
## country\_nameGhana country\_nameGreece country\_nameGrenada  
## large -3.351408 12.13112 -18.29449  
## medium 12.637225 11.12895 -25.82497  
## country\_nameGuam country\_nameGuatemala country\_nameGuinea  
## large -15.46752 -0.7129694 12.32142  
## medium -25.56755 0.6217765 11.06572  
## country\_nameHaiti country\_nameHonduras country\_nameHong Kong  
## large -0.6918849 -1.737976 -3.12791  
## medium -0.5143962 -0.869441 10.24737  
## country\_nameIceland country\_nameIndia country\_nameIndonesia  
## large 1.270953 -0.6106655 1.220030  
## medium -1.544228 0.1160878 1.126124  
## country\_nameIran country\_nameIreland country\_nameIsle of Man  
## large -14.795593 -1.0594665 -15.449782  
## medium -0.850085 -0.8716125 -1.543018  
## country\_nameIsrael country\_nameItaly country\_nameIvory Coast  
## large -3.12791 0.5530436 -1.128185  
## medium 10.24737 0.6898521 -0.907184  
## country\_nameJamaica country\_nameJapan country\_nameJersey  
## large -16.2030949 -0.2702735 -15.46752  
## medium 0.4955841 -0.2608066 -25.56755  
## country\_nameJordan country\_nameKazakhstan country\_nameKenya  
## large 17.182869 -2.515561 0.92307215  
## medium -9.921294 7.845234 0.04609626  
## country\_nameKyrgyzstan country\_nameLaos country\_nameLebanon  
## large -2.43992 9.735334 -3.070224  
## medium 15.24450 8.445189 13.313990  
## country\_nameLiberia country\_nameLuxembourg country\_nameMacedonia  
## large -3.014343 -3.014343 -3.25849  
## medium 10.134804 10.134804 12.07988  
## country\_nameMadagascar country\_nameMalawi country\_nameMalaysia  
## large 11.62043 -2.515561 -0.6686410  
## medium 11.53541 7.845234 -0.3115964  
## country\_nameMauritius country\_nameMexico country\_nameMongolia  
## large -15.46752 0.6674792 18.30748  
## medium -25.56755 0.7724658 -11.54534  
## country\_nameMontenegro country\_nameMorocco country\_nameMyanmar [Burma]  
## large -2.515561 -2.976083 12.09529  
## medium 7.845234 8.410691 11.71660  
## country\_nameNamibia country\_nameNepal country\_nameNew Zealand  
## large -15.449782 0.2216214 -0.7733699  
## medium -1.543018 0.6563372 -0.2068546  
## country\_nameNicaragua country\_nameNigeria country\_nameNorth Korea  
## large -0.5550472 0.08686268 -3.014343  
## medium -0.1612596 0.37172164 10.134804  
## country\_nameNorway country\_namePakistan country\_namePanama  
## large 0.3240648 0.2468113 -1.4258222  
## medium 1.2387180 0.4932025 -0.8757357  
## country\_namePapua New Guinea country\_namePeru country\_namePhilippines  
## large 0.8863399 1.1068262 -0.2340531  
## medium 0.6762315 0.8673936 0.4631328  
## country\_namePoland country\_namePortugal country\_namePuerto Rico  
## large -3.014343 10.00997 -16.8816586  
## medium 10.134804 10.34692 -0.8502976  
## country\_nameRomania country\_nameRussia country\_nameRwanda  
## large -3.196976 0.4460222 10.95641  
## medium 11.332459 -0.3597444 10.06126  
## country\_nameSaint Kitts and Nevis country\_nameSaint Lucia  
## large -15.46752 -3.323585  
## medium -25.56755 12.982608  
## country\_nameSaint Vincent and the Grenadines country\_nameSaudi Arabia  
## large 12.38428 12.46179  
## medium 11.73991 11.74287  
## country\_nameSerbia country\_nameSierra Leone country\_nameSlovakia  
## large -14.3787353 -1.163464 -3.014343  
## medium -0.4446624 -1.431597 10.134804  
## country\_nameSolomon Islands country\_nameSouth Africa  
## large -3.25849 -21.130129  
## medium 12.07988 -1.729319  
## country\_nameSouth Korea country\_nameSpain country\_nameSri Lanka  
## large 13.77458 -18.0733129 0.4103205  
## medium 13.32137 -0.8737196 1.2120124  
## country\_nameSudan country\_nameSwaziland country\_nameSwitzerland  
## large -3.067708 -2.608403 -0.8785222  
## medium 8.521871 7.960248 -1.7192888  
## country\_nameTaiwan country\_nameTajikistan country\_nameTanzania  
## large 1.296273 0.08543273 11.27755  
## medium 1.442904 -0.49442824 10.92888  
## country\_nameThailand country\_nameTrinidad and Tobago country\_nameTurkey  
## large -1.361164 -2.287774 -0.76323462  
## medium -1.067582 -1.137821 0.03199308  
## country\_nameU.S. Virgin Islands country\_nameUganda country\_nameUkraine  
## large -2.515561 12.47660 11.57391  
## medium 7.845234 12.51994 10.20178  
## country\_nameUnited Kingdom country\_nameUnited States  
## large -1.744838 -2.352257  
## medium -1.298424 -1.739923  
## country\_nameUzbekistan country\_nameVanuatu country\_nameVenezuela  
## large -2.976083 10.455074 0.2809421  
## medium 8.410691 8.458132 1.2415234  
## country\_nameVietnam country\_nameYemen country\_nameZambia  
## large 0.8859336 12.46473 -3.344718  
## medium 1.4174911 11.85391 8.857517  
##   
## Residual Deviance: 9446.176   
## AIC: 9986.176

The Residual Deviance here depicts how much the curve cannot fit and is very high.

In this case, our Logistic Regression failed. Hence, using the fatality\_count and country\_name, we were not able to predict the landslide\_size as medium and large, keeping small as base

## Summary of findings and questions for further analysis

1. We found the solution to Question-1(What was the landslide’s maximum death count so far?)  We found the answer by plotting the fatality\_count and looking at the maximum outlier. It could be an erroneous record, but the dataset can also have the correct number, and the death count could be **5000**
2. We found out the solution to Question-2(Are the sizes of various landslides equally distributed in the dataset)  
   No, the dataset is highly skewed towards landslides’ “medium” size. The maximum values in the dataset were recorded for the “medium” landslide. Hence the various landslide sizes are not equally distributed in the dataset.
3. We found out the solution to Question-3(What are the countries with more than 50 injured recorded in any landslide?) and used a pie-chart above to show the countries with more than 50 people wounded during any global landslide.
4. We found out the solution to Question-4(Is there any correlation between the numerical variable)  
   We only took the fatality\_count, injury\_count, latitude, and longitude to study the correlation between the numerical variables. However, we found no significant correlation between the location(depicted by latitude and longitude) and fatality\_count or injury\_count. We also found a very minimum correlation between the count of injured and demised people.
5. For our Question-5(Perform hypothesis testing to see if the mean of the fatality\_count of any two countries with the same number of landslides will be the same or not), we performed a T-test on two sets of countries with the same number of landslides. In both cases, we found that the mean value of fatality\_count is no match. Even if those two countries have the same number of landslides recorded, one of them lost more lives than the other.
6. For our Question-5(Use Logistic Regression to predict the size of the landslide), we used logistic regression to predict the size of the landslide based on the country name and fatality\_count. But our model failed. Hence, we could not predict the size of the landslide and could not build an efficient model using this dataset.

## Conclusion

With this study on the GLC dataset, we could identify some of the relevant queries we could form using this global data. And we were also able to find solutions to some of them. The result was that the landslide size or the landslide count does not predict the number of deaths. After seeing the records, we also found that so many medium-size landslides cause more fatality after seeing the records. Unfortunately, we could not create a helpful prediction model of the landslide size based on the various countries.

## Limitations

During this project, I realized I kept going back to the materials to clarify my understanding of the extensive dataset analysis. However, I also found that I lack my knowledge of R programming. There were some objectives for which I had to code many lines manually, but I’m sure there are many libraries of R which could make the work simple in one sentence. With this project, I also got practical exposure to understanding massive datasets. This GLC dataset can be used to create predictive models to identify potential landslide regions and their impact.