Day Four: Data Fundamentals and Intro to RStudio Environment

SDS 192: Introduction to Data Science

Lindsay Poirier Statistical & Data Sciences, Smith College

Spring 2022

```
global_landslide <- read.csv("https://data.nasa.gov/api/views/dd9e-wu2v/rows.csv")
```

1. Run the code chunk above, and then start to review the dataset using strategies we discussed in class. You might check its dimensions (number of columns and number of row), column names, and print the first 6 rows. You can use the code chunk below to help you.

```
# Check number of columns
ncol(global_landslide)
```

```
## [1] 31
```

```
# Check the column names
names(global_landslide)
```

```
##  [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"
```

```
# Print the first six rows
head(global_landslide)
```

```
##           source_name
## 1                 AGU
## 2           Oregonian
```

```
## 3             CBS News
## 4              Reuters
## 5          The Freeman
## 6 BusinessWorld Online
##
## 1                                          https://blogs.agu.org/landslideblog/2008/10/14/the-lifan
## 2                                                http://www.oregonlive.com/news/index.ss
## 3                                                                    https://www.cbsnews.c
## 4                                                                                https://in
## 5                                                                 http://www.philstar.com/cebu
## 6 http://www.bworldonline.com/content.php?section=Nation&title=-death-toll-from-rains-rises-nationwi
##   event_id             event_date event_time
## 1      684 08/01/2008 12:00:00 AM         NA
## 2      956 01/02/2009 02:00:00 AM         NA
## 3      973 01/19/2007 12:00:00 AM         NA
## 4     1067 07/31/2009 12:00:00 AM         NA
## 5     2603 10/16/2010 12:00:00 PM         NA
## 6     4203 02/16/2012 12:00:00 AM         NA
##                                                 event_title
## 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
##
## 1
## 2
## 3
## 4
## 5 Another landslide in sitio Bakilid in barangay Lahug also left two families homeless. Lilibeth Mags
## 6
##                                         location_description
## 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
##   location_accuracy landslide_category landslide_trigger landslide_size
## 1           unknown          landslide              rain          large
## 2              5km           mudslide          downpour          small
## 3             10km          landslide          downpour          large
## 4           unknown          landslide           monsoon         medium
## 5              5km          landslide  tropical_cyclone         medium
## 6              5km          landslide          downpour         medium
##   landslide_setting fatality_count injury_count               storm_name
## 1              mine             11           NA
## 2           unknown              0           NA
## 3           unknown             10           NA
## 4           unknown              1           NA
## 5           unknown              0           NA Supertyphoon Juan (Megi)
## 6           unknown              0           NA
##   photo_link notes event_import_source event_import_id  country_name
```

```
## 1                                 glc             684         China
## 2                                 glc             956 United States
## 3                                 glc             973          Peru
## 4                                 glc            1067         Nepal
## 5                                 glc            2603   Philippines
## 6                                 glc            4203   Philippines
##   country_code admin_division_name admin_division_population
## 1           CN             Shaanxi                         0
## 2           US              Oregon                     36619
## 3           PE               Junín                     14708
## 4           NP         Mid Western                     20908
## 5           PH     Central Visayas                    798634
## 6           PH     Eastern Visayas                      2404
##   gazeteer_closest_point gazeteer_distance         submitted_date
## 1               Jingyang          41.02145 04/01/2014 12:00:00 AM
## 2            Lake Oswego           0.60342 04/01/2014 12:00:00 AM
## 3              San Ramón           0.85548 04/01/2014 12:00:00 AM
## 4                Dailekh           0.75395 04/01/2014 12:00:00 AM
## 5              Cebu City           2.02204 04/01/2014 12:00:00 AM
## 6           Balinsacayao           2.28967 04/01/2014 12:00:00 AM
##             created_date       last_edited_date longitude latitude
## 1 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM  107.4500  32.5625
## 2 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM -122.6630  45.4200
## 3 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM  -75.3587 -11.1295
## 4 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM   81.7080  28.8378
## 5 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM  123.8978  10.3336
## 6 11/20/2017 03:17:00 PM 02/15/2018 03:51:00 PM  124.9668  10.7004
```

```
# View the data frame
#View(global_landslide)
```

2. Find one nominal variable in the dataset, one ordinal, one discrete, and one continuous. Uncomment the lines below, and store the entire column of values in the variables.

```
#There are many including source_link, event_id, event_title, event_description, landslide_category, et
nominal <- global_landslide$source_name

#There are a few including location_accuracy and landslide_size
ordinal <- global_landslide$landslide_size

#There are a few including fatality_count, injury_count, and admin_division_population
discrete <- global_landslide$fatality_count

#There are a few including latitude and longitude
continuous <- global_landslide$latitude

#Feel free to ask if you selected other variables and are unsure!
```

3. Check the class of each of these vectors. Are they what you expected?

```
# Check the class of the nominal variable below.

class(nominal)
```

```
## [1] "character"
```

```
# Check the class of the ordinal variable below.

class(ordinal)
```

```
## [1] "character"
```

```
# Check the class of the continuous variable below.

class(continuous)
```

```
## [1] "numeric"
```

```
# Check the class of the discrete variable below.

class(discrete)
```

```
## [1] "integer"
```

4. Determine the number of distinct possible values in the nominal and ordinal variables.

```
# Calculate the number of distinct values in the nominal variable.

length(unique(nominal))
```

```
## [1] 3918
```

```
# Calculate the number of distinct values in the ordinal variable.

length(unique(ordinal))
```

```
## [1] 7
```