Volcanic Eruption during Holocene

Irene Viola

6/23/2020

Abstract

Volcanoes eruptions are a well-known natural phenomenon and a very rare one. That's why predict volcanic eruption is a big issue for geologist and scientist, even data scientist. In this work the volcanic eruption during the Holocene has been analysed and a prediction model has been proposed. The aim is to predict the volcano type based on the last 11700 years of volcanic activity all across the planet. Two different models have been used, knn and random forest, and the results have been compared and the best model has been improved as much as possible. The results are encouraging also if more studies and larger dataset would help to perfect the prediction and the model used.

1. Introduction

Volcanoes eruptions are a well-known natural phenomenon that affect the planet since its formation. These eruptive processes can be spectacular, massive and very dangerous, plus them are very rare. The prediction of a volcanic eruption is a very difficult task in science. The aim of this work is to find out a model that can predict volcanic eruption based on volcanoes type and improve the accuracy also if, because of the different variables to consider the rarity or the phenomenon and the large time scale, to reach 0.8-0.9 won't be possible and credible. To develop the model, Volcanic Eruption during Holocene database model by Smithsonian Institution on keggle.com has been used.

1.2 Holocene Period

Holocene epoch is a geological era, the younger of the Quaternary Period and the latest interval of geologic time (including our time, the present), covering approximately the last 11,700 years of Earth's history. This period is well known to be the less geological active of the Quaternary period, in fact most of the movement, earthquakes and tectonics stabilized during Pleistocene. During the Holocene the planet is stable and the continents are in the same position as we know them today. The only geological phenomena that keep occurring in this period is volcanism.

1.2 Volcanism

Volcanism is defined as the various processes and phenomena that occurs on Earth's surface and involve surficial discharge of molten rock, pyroclastic fragments, or hot water and steam, including volcanoes eruptions, geysers, and fumaroles. Volcanisms seems to occur all around the world randomly, instead geologist known that the volcanic activity is associated with several distinct geologic settings.

1.3 Geological settings

Geological setting is a description of geologic characteristics distinctive of a geographic area or phenomena. Regarding volcanisms geological setting is strictly connected to tectonic setting, such as the movement of opening and or colliding of geological plates that cover our planet and that compose the zest of the Earth: the lithosphere. It includes the crust, rigid and solid where we walk on, and the upper mantle, incandescent strata that is directly below the crust and is made of melted rock. Most geological settings are associated with the boundaries of the enormous rigid plates that make up the lithosphere (tectonic setting). The majority of active terrestrial volcanoes and related phenomena occur where two tectonic plates converge and one overrides the other (Subduction process), forcing it down into the mantle to be reabsorbed. Long curved chains of islands known as island arcs form at such subduction zones. A second major site of active volcanism is along the axis of the oceanic ridge system, where the plates move apart on both sides of the ridge and magma wells up from the mantle, creating new ocean floor along the trailing edges of both plates. Virtually all of this volcanic activity occurs underwater. In a few places the oceanic ridges are sufficiently elevated above the deep seafloor that they emerge from the ocean, and subaerial volcanism occurs. Iceland is the best-known example. A relatively small number of volcanoes occur within plates far from their margins because of plate movement over a "hot spot" from which magmas can penetrate to the surface. An even smaller number of volcanoes occur within a particular area of a plate, Rift Valley, due to an opening of the plate and consequent thinning of the plate that permits the magma to reach the surface.

1.4 Type of Volcanoes

There are many different type of volcanoes determined by the tectonic setting and magma composition. The most known and common are Stratovolcanoes. These are the classical volcanoes everyone can imagine, with a steep sides and symmetrical cone shape. Usually this type of volcanoes is associated to subduction zones. Then we can find Shield volcanoes that have a less impressive cone due to the low viscosity of the lava coming out, in fact the lava spreads all around covering a large area and forming very gentle slopes. These volcanoes are usually generated within the plate due to the hot spots influence. A very spectacular type of volcano is Fissure Vent. Here magma rises along the opening of two plates (the ridge) and find out its way out along fractures generating fountains of lava called fissure eruption. Usually this eruption can be seen only where the ridge emerges, like in Iceland. Isolated lava fountains along a fissure produce crater rows of small spatter. Another type of volcano is the spatter cone or cone that is very similar to stratovolcanoes but, since the magma rising contain too much gases, but not enough for an explosive event, it erupts as blobs magma and fall close to the vent forming a steep sided cone. This type of volcano is common in subduction zones and rift valley. When magma is stored in a magma chamber and the gas pressure overcome the equilibrium, the over pressure generates a very large explosive eruption that make all magma sprout out at once. The magma chamber, almost empty, collapses forming a depression or even bowl on the surface. The morphology originated is called Caldera volcano and can happens usually on rift valley and subduction zones. Together with these major classes of volcanoes, there are several intermediate typologies considered in this database. Pyroclastic shield is an uncommon type of shield volcano. Unlike most shield volcanoes, pyroclastic shields are formed mostly of pyroclastic and highly explosive eruptions rather than relatively fluid basaltic lava issuing from vents or fissures on the surface of the volcano. Pyroclastic cones are relatively small, steep (about 30°) volcanic landforms built of loose pyroclastic fragments, most of which are cinder-sized. A volcanic field is an area of the Earth's crust that is prone to localized volcanic activity. They usually contain 10 to 100 volcanoes such as cinder cones and are usually in clusters. Lava flows may also occur. They don't have to be confused with volcanic complex, or simply "complex", such as a mixed land form consisting of related volcanic centres and their associated lava flows and pyroclastic rock. They may form due to changes in eruptive habit or in the location of the principal vent area on a particular volcano. lava dome is a circular mound-shaped protrusion resulting from the slow extrusion of viscous lava from a volcano. A maar is a broad, low-relief volcanic crater caused by a phreatomagmatic eruption (an explosion which occurs when groundwater comes into contact with hot lava or magma). A maar characteristically fills with water to form a relatively shallow crater lake which may also be called a maar. Compound Volcano is a volcano with more than one cone. The cone in the Volcano is made up of alternating layers of lava and

ashes. Cones are among the simplest volcanic landforms. They are built by ejecta from a volcanic vent, piling up around the vent in the shape of a cone with a central crater. Tuff cone is a small monogenetic volcanic cone produced by phreatic (hydrovolcanic) explosions directly associated with magma migration to the surface. Tuff ring is a related type of small monogenetic volcano that is also produced like tuff cone, but are thinner and circular. Subglacial volcano is a volcanic form produced by subglacial eruptions or eruptions beneath the surface of a glacier or ice sheet which is then melted into a lake by the rising lava. Today they are most common in Iceland and Antarctica. Lava cone is a type of volcano composed primarily of viscous lava flows. The volcanic cone can contain a convex profile due to the flank flows of viscous lava. Explosion crater is a type of crater formed when material is ejected from the surface of the ground by an explosive event at or immediately above or below the surface. Last volcanic morphology type to consider are submarine volcanoes. These are underwater vents or fissures in the Earth's surface from which magma can erupt. Many submarine volcanoes are located near areas of tectonic plate formation, known as mid-ocean ridges. The volcanoes at mid-ocean ridges alone are estimated to account for 75% of the magma output on Earth. Although most submarine volcanoes are located in the depths of seas and oceans, some also exist in shallow water, and these can discharge material into the atmosphere during an eruption.

1.5 Dominant Rock Type

The igneous rocks or magmatic rocks are the rocks that form by volcanism effect. The most abundant igneous rock on Earth is basalt. This is an extrusive (forms when magma comes out and became lava and then solidify) igneous rock formed from the rapid cooling of silica-poor, magnesium-rich and iron-rich lava exposed at or very near to the surface. Rhyolite is again an extrusive igneous rock that, in this case, form from silica rich lava. If it cools down very quickly, the obsidian forms, if it cools down slowly several crystals form inside. Andesite is a medium term between basal and rhyolite, again extrusive igneous rock, formed by cool down of silica, magnesium and iron lava. Dacite, instead is the rock in between rhyolite and andesite, and trachyte is in between andesite and dacite. All the other type considered are peculiar mix in between these.

2. The Dataset

The dataset used in this work is the open Volcanic Eruptions in the Holocene Period in Kaggle.com (https://www.kaggle.com/smithsonian/volcanic-eruptions) uploaded by the Smithsonian Institution's for his Global Volcanism Program (GVP) and available on my GitHub. This database documents Earth's volcanoes and their eruptive history over the past 10,000 years. The GVP reports on current eruptions from around the world and maintains a database repository on active volcanoes and their eruptions. The GVP is housed in the Department of Mineral Sciences, part of the National Museum of Natural History, on the National Mall in Washington, D.C. The database has 1058 registered eruptions in rows and 12 columns with different variables: progressive number of each eruption, name of the volcano, country, region, type of volcano, latitude and longitude, altitude, activity, tectonic setting and dominant rock type

The database has been downloaded and uploaded in R.

2.1 Libraries Used

The libraries used in this project are: tidyverse

caret

data.table

readr

tidyr

ggmap

dplyr

maptools

plotly

maps

countrycode

ggplot2

recipes

tidymodels

themis

workflows

tune

vip

janitor

ranger

All the libraries are available as automatic installation calls.

3. Methods

3.1 Data correction and wrangling

First step after downloading the dataset is to inspect the database and check if the data need to be elaborated.

```
#Database structure
str(volcanic_eruption)
```

```
## tibble [1,508 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Number             : num [1:1508] 210010 210020 210030 210040 211001 ...
##  $ Name               : chr [1:1508] "West Eifel Volcanic Field" "Chaine des Puys" "Olot Volcanic Fi
##  $ Country            : chr [1:1508] "Germany" "France" "Spain" "Spain" ...
##  $ Region             : chr [1:1508] "Mediterranean and Western Asia" "Mediterranean and Western Asi
##  $ Type               : chr [1:1508] "Maar(s)" "Lava dome(s)" "Pyroclastic cone(s)" "Pyroclastic con
##  $ Activity Evidence  : chr [1:1508] "Eruption Dated" "Eruption Dated" "Evidence Credible" "Eruption
##  $ Last Known Eruption: chr [1:1508] "8300 BCE" "4040 BCE" "Unknown" "3600 BCE" ...
##  $ Latitude           : num [1:1508] 50.2 45.8 42.2 38.9 43.2 ...
##  $ Longitude          : num [1:1508] 6.85 2.97 2.53 -4.02 10.87 ...
##  $ Elevation (Meters) : num [1:1508] 600 1464 893 1117 500 ...
##  $ Dominant Rock Type : chr [1:1508] "Foidite" "Basalt / Picro-Basalt" "Trachybasalt / Tephrite Basa
##  $ Tectonic Setting   : chr [1:1508] "Rift Zone / Continental Crust (>25 km)" "Rift Zone / Continenta
##  - attr(*, "spec")=
##   .. cols(
##   ..   Number = col_double(),
##   ..   Name = col_character(),
##   ..   Country = col_character(),
```

```
##   ..   Region = col_character(),
##   ..   Type = col_character(),
##   ..   `Activity Evidence` = col_character(),
##   ..   `Last Known Eruption` = col_character(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   `Elevation (Meters)` = col_double(),
##   ..   `Dominant Rock Type` = col_character(),
##   ..   `Tectonic Setting` = col_character()
##   .. )
```

```
#Checking if `Name` is unique
unique(volcanic_eruption[duplicated(volcanic_eruption[, "Name"]), "Name"])
```

```
## # A tibble: 9 x 1
##   Name
##   <chr>
## 1 Borawli
## 2 Unnamed
## 3 Sumbing
## 4 Plosky
## 5 Santo Tomas
## 6 San Diego
## 7 Santa Isabel
## 8 Azul, Cerro
## 9 Flores
```

First thing to note is that the volcano eruptions considered came from different volcanoes, only nine of them repeat meaning that they gave more than one eruption in 11,700 years. Looking at the Country column, the countries names are not unique, it has been corrected using `countrycode` package and standardize.

Before moving on to check uniformity of the different value, it need to be considered if there are NAs.

There are 61 NAs. Now, before deciding if ignore or correct them and how, let's check in which variable are present.

The variables affected by NAs are "Activity Evidence", "Dominant Rock Type" and "Tectonic Setting". Looking at the voices of these variables, is clear that we can't ignore the NAs because, as said in the introduction, these are important variables to detect and identify volcanoes and their eruptions. Because of this we are going to replace NAs with more significant value for each variable.

Moving on inspecting the variables values another problem is given by the Type of volcanoes. In fact, the different types are not written univocally: the same type of volcano is written with and without s, es and other symbols. This is going to be a problem during the elaboration of data since every difference is considered a different value and in term of volcanoes type it doesn't make sense.

```
#the eruption type are written differently
Type_sum <- volcanic_eruption %>%
  select(Country) %>%
  group_by(volcanic_eruption$Type) %>%
  summarize(count = n())
#To have the best and clear correction let's correct each type
#"manually" to be sure that the names are univocal.
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Caldera(s)", "Caldera", Type))
```

```
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Complex(es)", "Complex", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Fissure vent(s)", "Fissure vent ", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Fissure vent\t", "Fissure vent  ", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Lava cone(s)", "Lava cone", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Lava dome(s)", "Lava dome", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Maar(s)", "Maar", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Pyroclastic cone(s)", "Pyroclastic cone", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Shield", "Pyroclastic shield", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Shield(s)", "Pyroclastic shield", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Stratovolcano?", "Stratovolcano", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Stratovolcano(es)", "Stratovolcano", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Submarine(es)", "Submarine", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Tuff cone(s)", "Tuff cone", Type))
volcanic_eruption <- volcanic_eruption %>%
  mutate(Type = ifelse(Type == "Volcanic field(s)", "Volcanic field", Type))
```

After this correction let's check again how many type of volcanoes were active in the last 11,700 years, during Holocene period.

Now instead of 33, the types of volcano are 20 (19 identified types plus "Unknown"). Until now, talking about times and eras it as always been used the geological notation "years". In fact, in geology, time count start from today (year1) and count increase moving to the past until Earth's formation estimated around $4.54 \times 10\hat{}9$ years $\pm$ 1%. In the Smithsonian database, another notation is considered for the /Last Known Eruption/. Looking at the column, we can see that the years are expressed in Common Era notation. Common Era (CE) is one of the notation systems for the world's most widely used calendar era. BCE (Before the Common Era or Before the Current Era) is the era before CE and those are equivalent to the Dionysian BC and AD system respectively. Notations refer to the Gregorian calendar (and its predecessor, the Julian calendar). The year-numbering system used by the Gregorian calendar is used throughout the world today, and is an international standard for civil calendars. However, it's definitely not correct when talking in geological time since it refers to human age (just a blink respect to geological ages). Because of this, a correction is needed to convert BCE and CE in "year" notation.

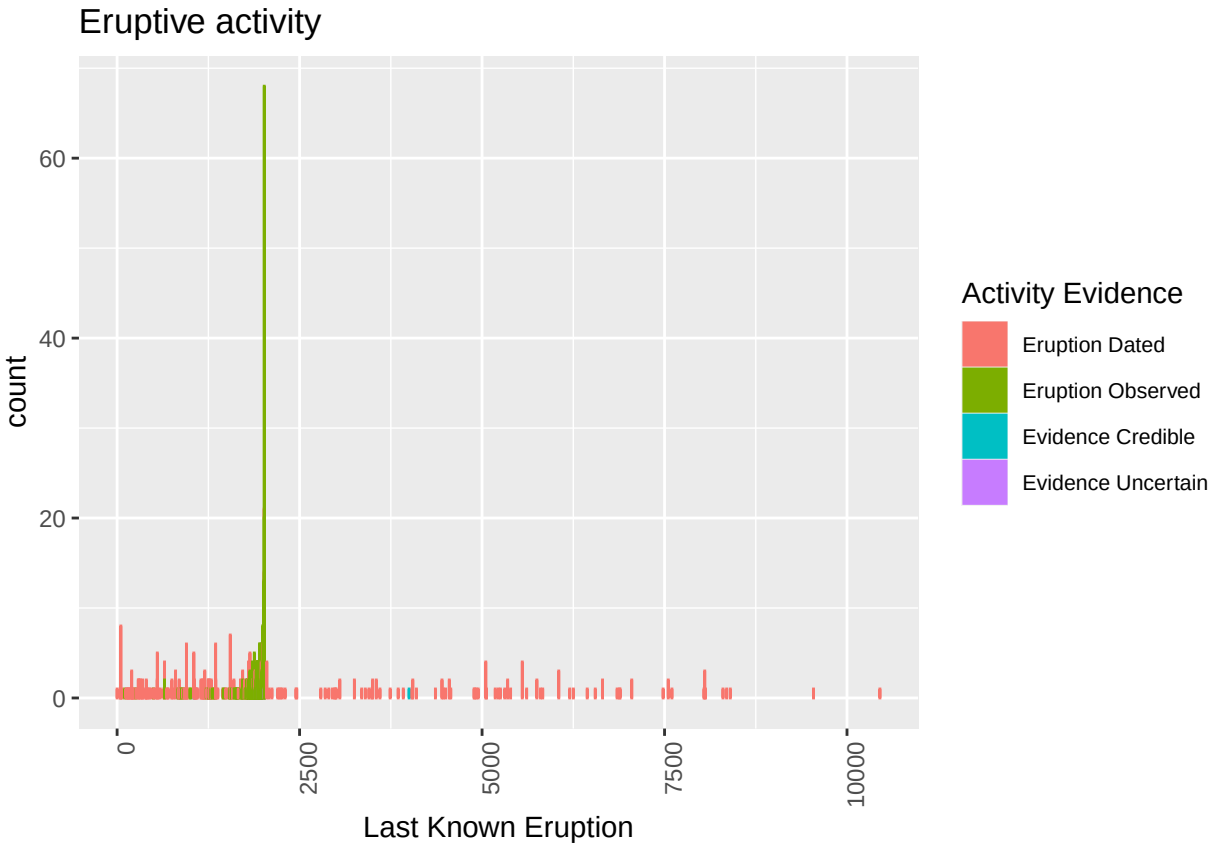
```
#First let's put NAs instead of "Unknown", this way, in the next step
#is possible to convert as.numeric and do calculation to have the
#correct notation
last_eruption <- ifelse(volcanic_eruption$`Last Known Eruption` ==
            "Unknown", NA , volcanic_eruption$`Last Known Eruption`)
#Define Years as numeric
year <- as.numeric(gsub("(^\\d+).+", "\\1", last_eruption))
#Define the factor of addiction for BCE and CE
adding_year <- ifelse(grepl("BCE", last_eruption), +2020,0) &
```

```
  ifelse(grepl("CE", last_eruption), -2016,0)
#Let's perform addition of 2020 to BCEs to have the kY used in
#geological scale and add 0 to the CEs.
year_geo <- adding_year + year
#Let's relace teh column values
volcanic_eruption$`Last Known Eruption`<-year_geo
```

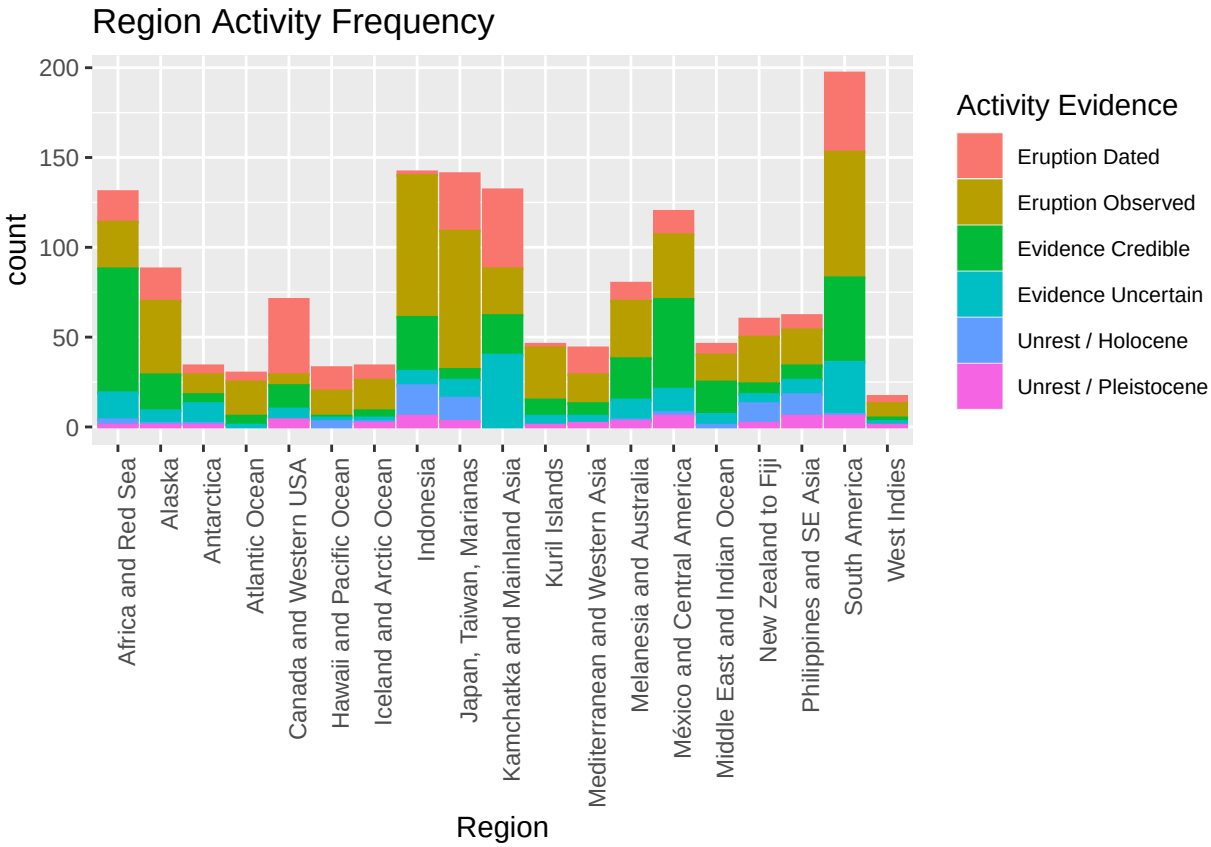
Now all "Last Known Eruption" are in correct geological scale notation.

3.2 Data investigation and Visualization

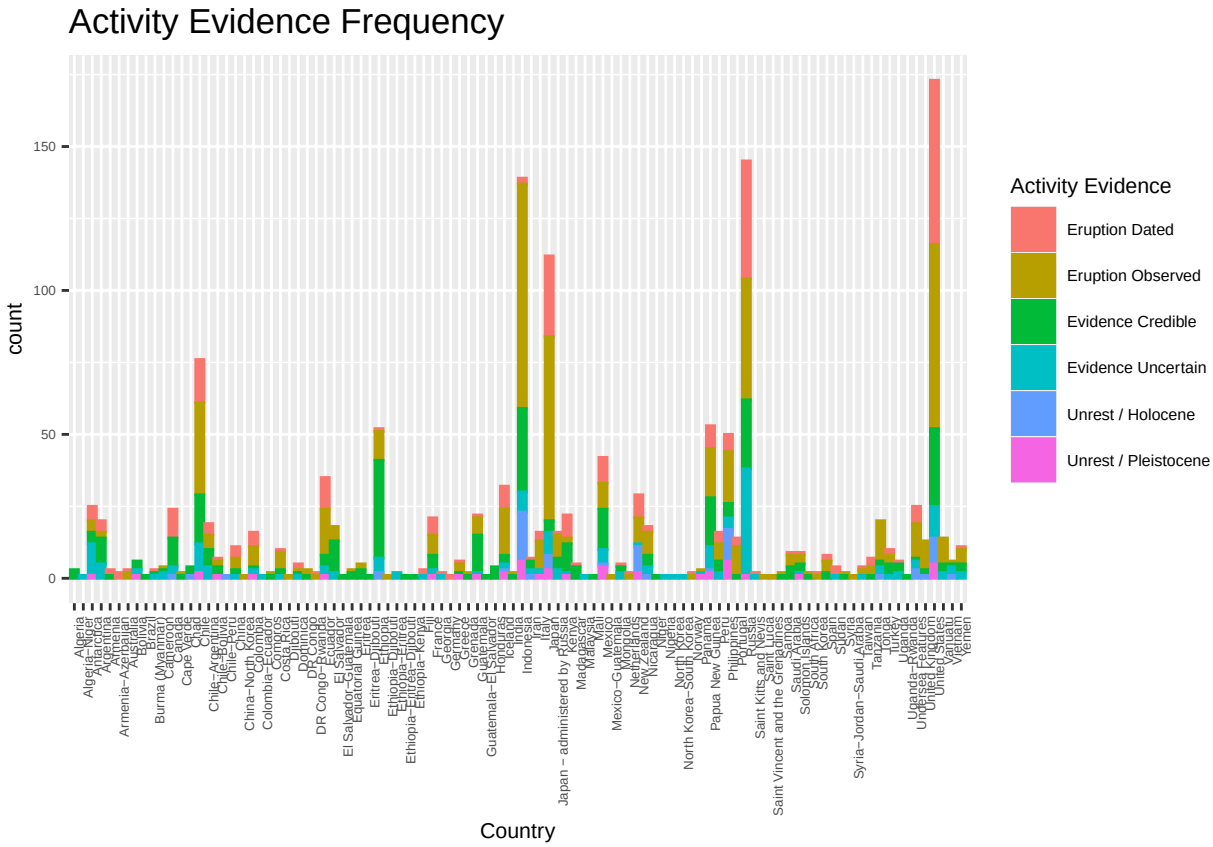
Now that the database is correct, let's use visualization to understand which method is better to use and which variables are more significant. Let's start visualize the data about the Holocene eruption, such as the year of the last known eruption of the volcanoes.



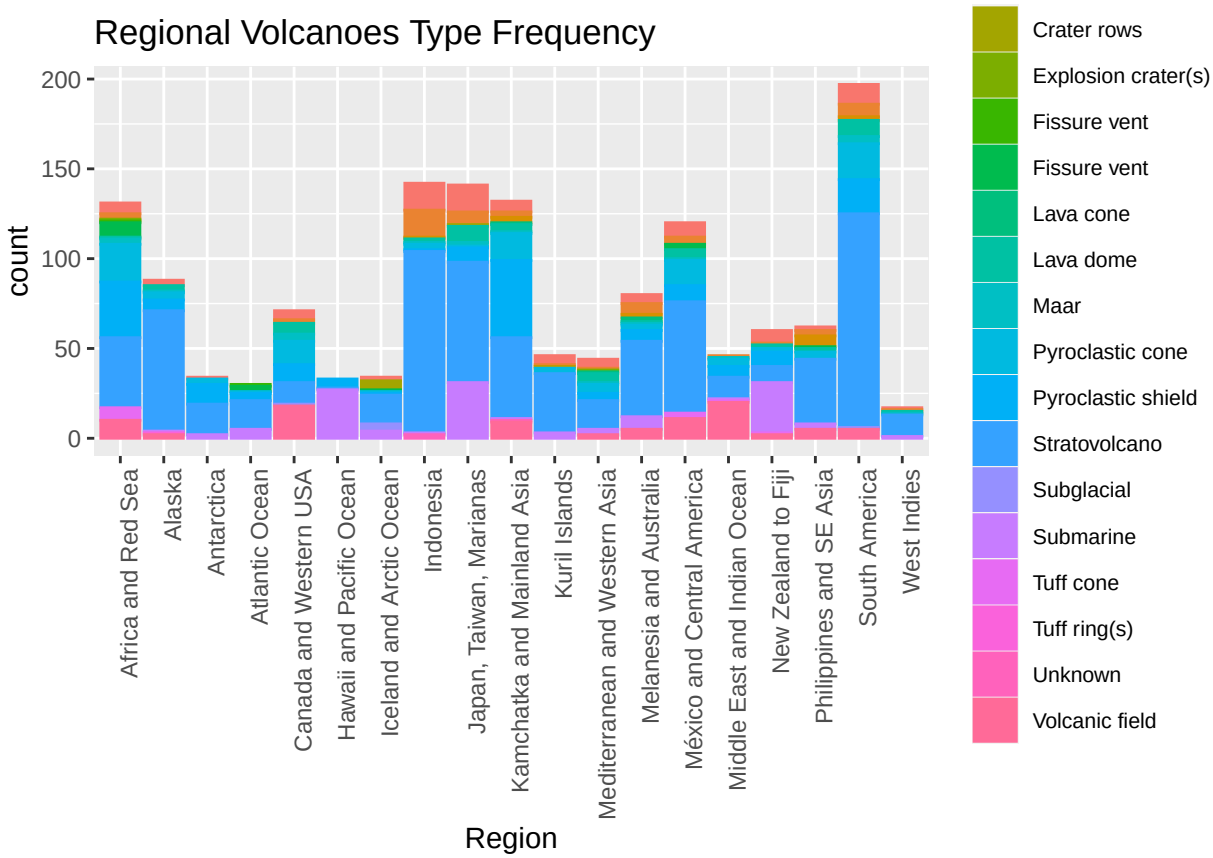
Is possible to see that there is a very intense period of observed eruption in the last 2000 years since it is the period in which man described and documented the phenomena. Let's see which Regions and Countries are the more active once.



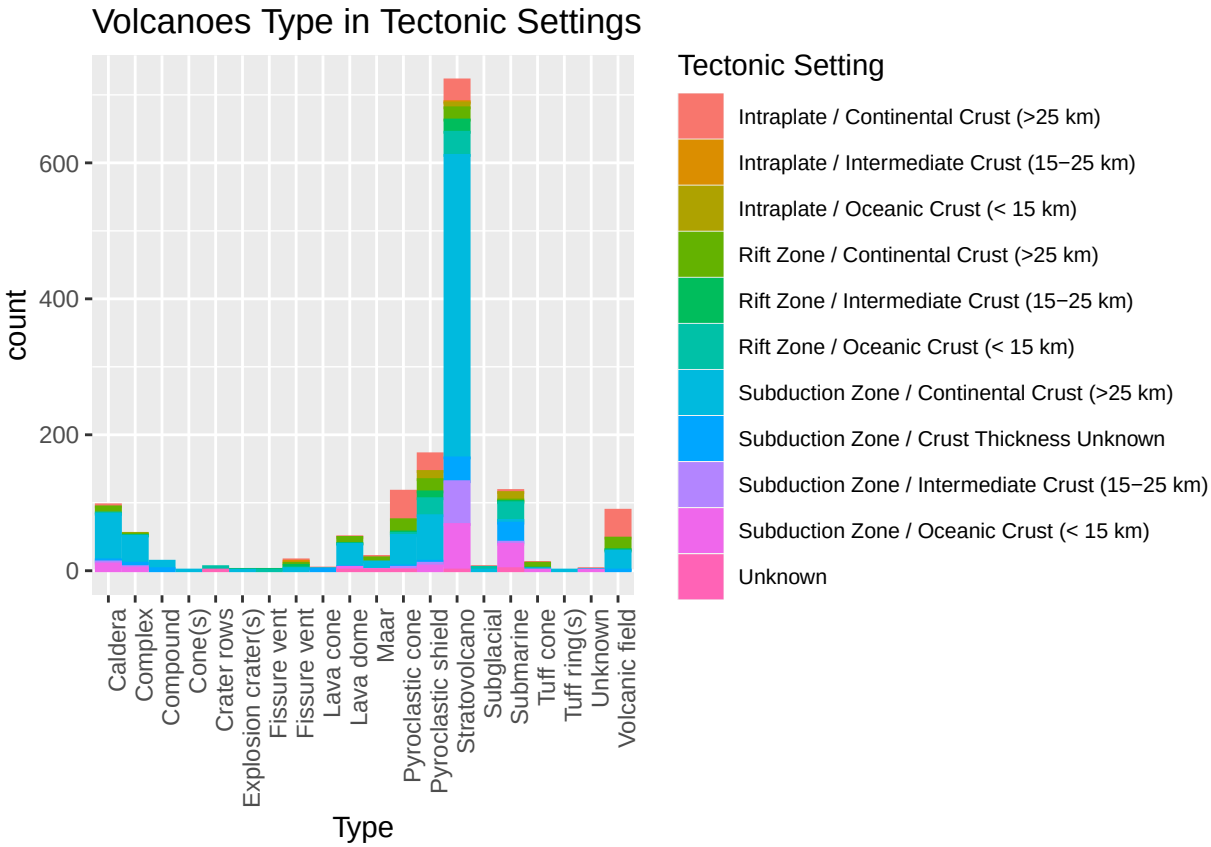
The most active region is South America with a very high number of eruption dated, observed and credible and low level of Uncertain.



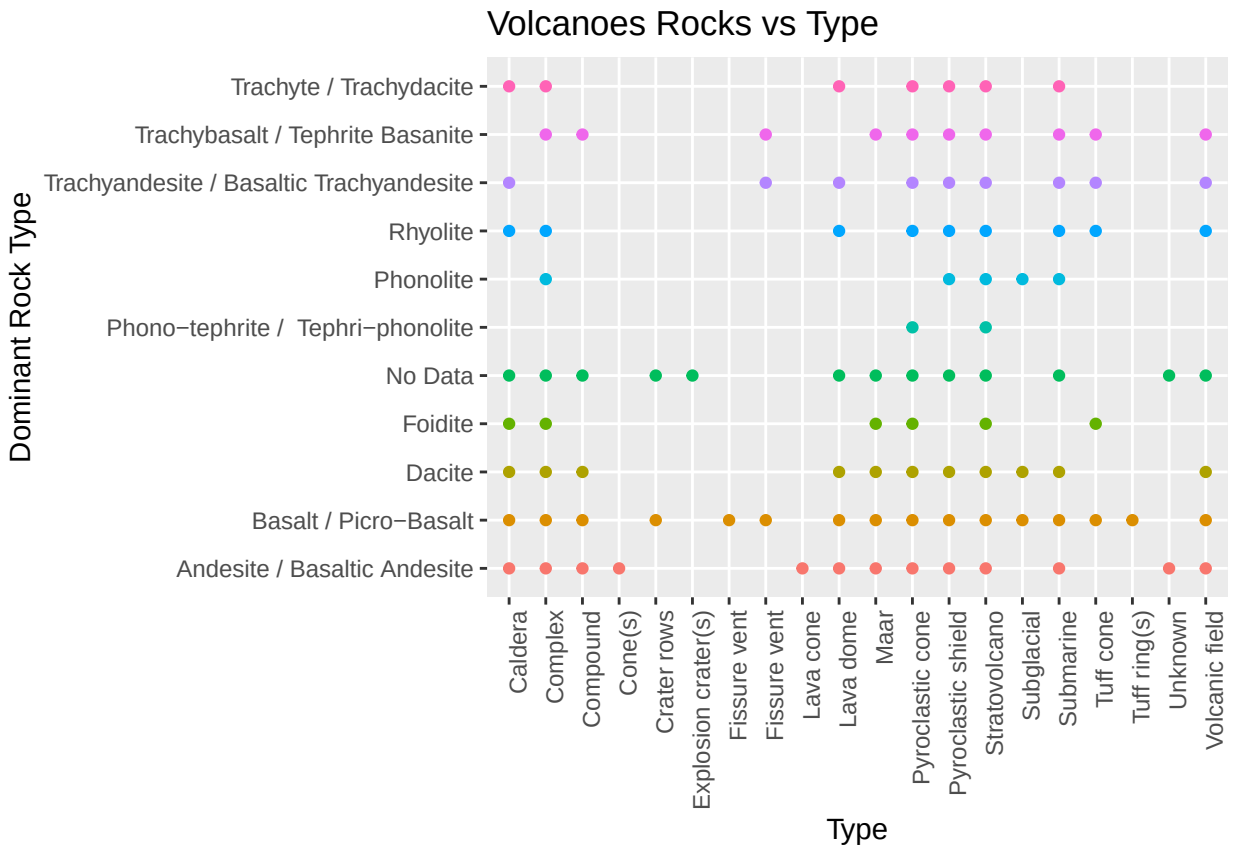
Moving to the countries, the most active one is United Stated followed by Russia and Indonesia. Now let's consider the type of volcanoes present in the different region.



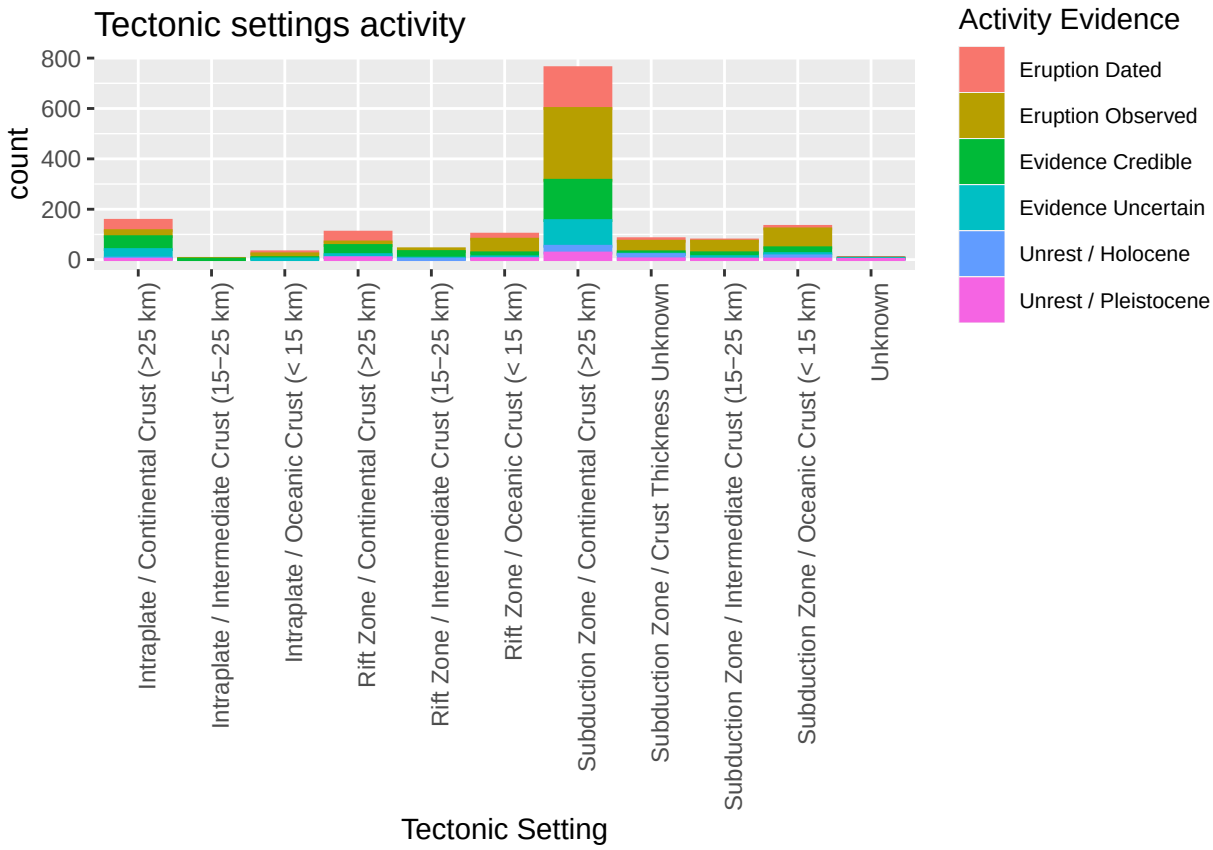
Almost everywhere the stratovolcanoes are the most present, followed by shield (pyroclastic or not). As said in the introduction, often the tectonic setting can give a specific type of volcano.



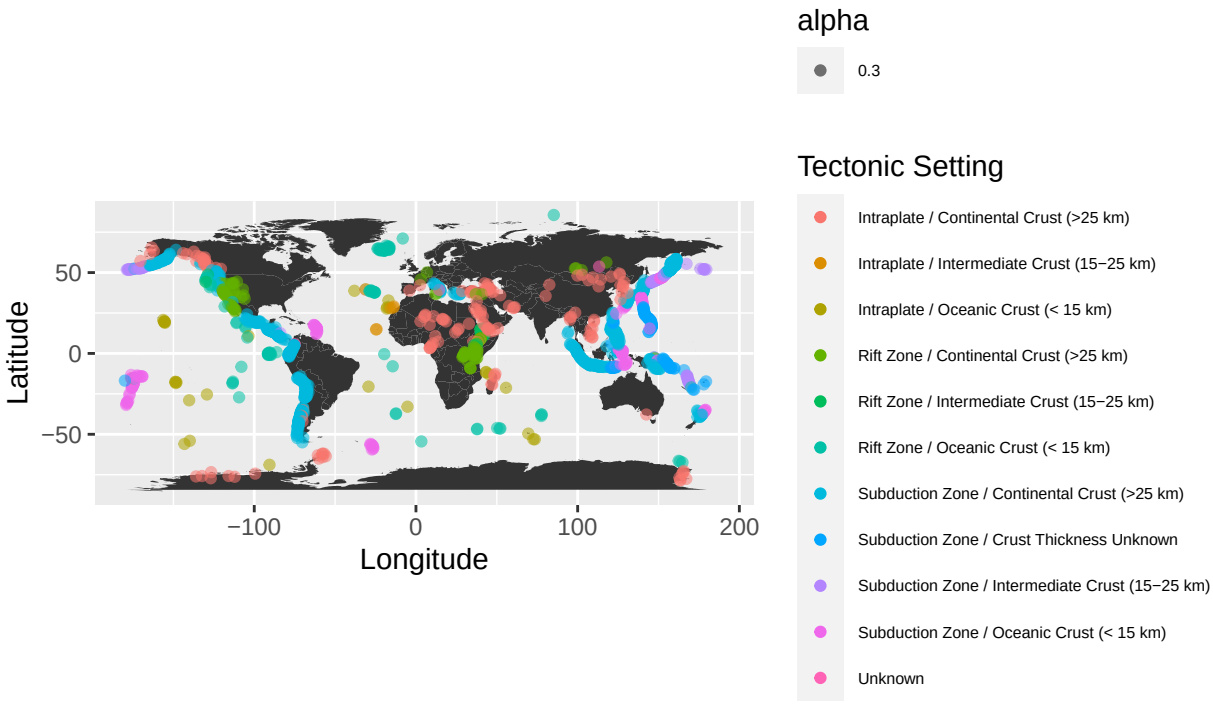
Stratovolcanoes are the most abundant and are easy to find in subduction thick zones, like where continental crust collide and the thickness of the plaque is >25km. Different rock type differs and in some case characterize volcanoes type, plotting Dominant Rock Type vs Volcano Type is possible to see these differences.

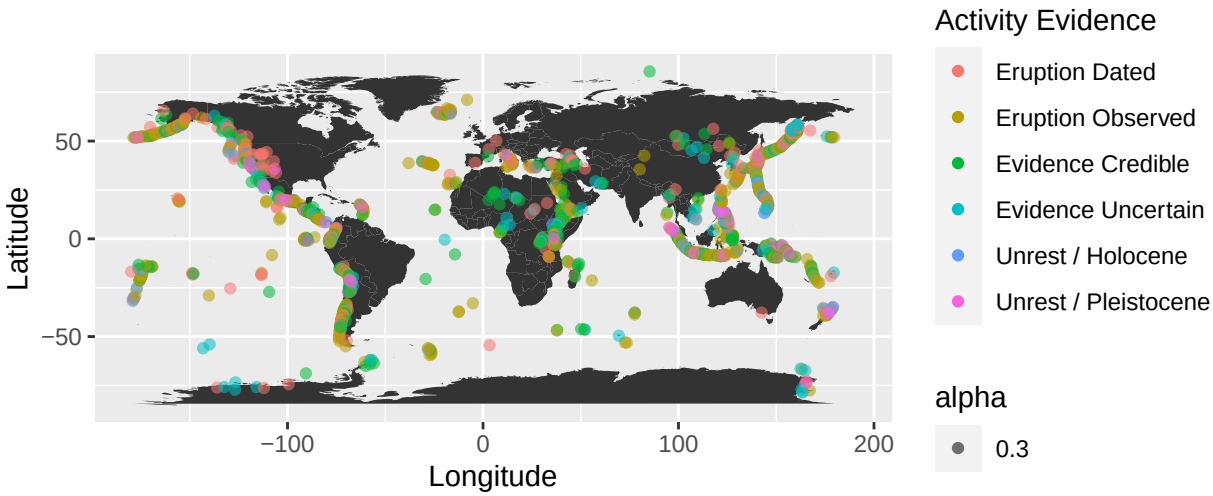


Stratovolcanoes are mostly formed by all the different type of rock and this is explainable due the cone strata formation. Instead there are volcanoes type formed mostly by one dominant rock. Crater rows and tuff rings are formed by basalt, cones instead are characterized by andesite. Tectonic setting is strictly connected to the activity evidence in fact some tectonic zones are more frequently active than others



Here is possible to see that teh tectonic setting Subduction Zone/Continental Crust (<25km) is the most active tectonic setting. Using latitude and longitude data is possible to see the distribution world-wide of the different volcanoes according to tectonic setting and according to activity evidence.





Comparing the two plots we can identify the areas where the presence and the activity of volcanoes is higher, the subduction zones.

3.4 Building the model

Until now all the values and parameters have been considered, but looking at the database some variables have a lot of voices that can be group together on prediction and model purpose. In "volcano type", there are 20 different voices, some of them with a very low number of eruption.

So let's create a new database where we consider the four most abundant type of volcanoes and let's group the remaining as "Other". In this database are added all the variables to use as predictor of the volcanic activity related to each type of volcanoes.

Instead of the names of the volcanoes, the ID number has been considered, together with latitude and longitude for geographical positioning. To identify the volcanoes type (grouped) have been taken into account also the last known eruption, the elevation, the tectonic setting and the dominant rock type as predictors. Looking at /tectonic setting/, we can group it in four categories: subduction zone, rift zone and intraplate plus the Unknown group for unknown data.

Last grouping to be considered is the dominant rock type grouping. In this case we group rocks based on the most abundant rock type in volcanology and simplify the classes.

Now that the new database to work with has been created, some small adjustments need to be done, like take out spaces in the column names to make it easier for the algorithms to use them, because spaces can give problems.

Now the last process to do is to create a partition of the database in training set and test set (90%-10%).

```
index <- createDataPartition(volcanoes_work4$volcano_type, p = 0.9, list = FALSE)
volcano_work_training <- volcanoes_work4[index,]
volcano_work_testing  <- volcanoes_work4[-index,]
```

3.5 Run the model

First algorithm considered is the K- nearest neighbors or Knn since it is easier to use in multiple dimensions. Knn is a method used for classification and regression that permit the pattern-recognition.

```r
#First Method: Knn method
fit <- train(volcano_type ~ ., method = "knn",
             tuneGrid = data.frame(k = seq(1, 15, 2)), data = volcano_work_training)
fit
```

```
## k-Nearest Neighbors
##
## 1360 samples
##    7 predictor
##    5 classes: 'Other', 'Pyroclastic cone', 'Pyroclastic shield', 'Stratovolcano', 'Submarine'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1360, 1360, 1360, 1360, 1360, 1360, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.5486986  0.3325427
##    3  0.5475952  0.3278635
##    5  0.5586391  0.3359464
##    7  0.5682487  0.3427793
##    9  0.5758191  0.3461570
##   11  0.5781513  0.3443742
##   13  0.5785985  0.3405557
##   15  0.5771519  0.3324285
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 13.
```

The accuracy of the model is near 0.6. Because of the large number of predictors and the large number of parameter that are going to define the probability and the prediction, random Forests is the best algorithm to use in this classification prediction.

```r
#Second Method: Random Forest
rf_fit <- train(as.factor(volcano_type)~ .,
                data = volcano_work_training,
                method = "ranger")
rf_fit
```

```
## Random Forest
##
## 1360 samples
##    7 predictor
##    5 classes: 'Other', 'Pyroclastic cone', 'Pyroclastic shield', 'Stratovolcano', 'Submarine'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1360, 1360, 1360, 1360, 1360, 1360, ...
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##     2   gini        0.4735393  0.0000000
##     2   extratrees  0.4735393  0.0000000
```

```
##   204   gini        0.6467598  0.4569701
##   204   extratrees  0.6283819  0.4222565
##   407   gini        0.6339388  0.4394691
##   407   extratrees  0.6250151  0.4236477
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 204, splitrule = gini
##  and min.node.size = 1.
```

```
#PRediction for voclavoes eruption with random forest
volcanoes_rf_pred <- predict(rf_fit, volcano_work_testing)
```

The accuracy improved in the model and give a prediction overall accuracy of 0.69. However, looking at the accuracy of each class, sensitivity and specificity is possible to see that also if random forest is a very good model to predict volcanoes type, it seems too perfect and need more correction in the database to perfection it. To check if the impression had about the model is correct, let's check centre and scale, perform Principal Component Analysis and identify and remove variables with near zero variance.

```
volcano_nzv_pca <- preProcess(select(volcanoes_work4, - volcano_type),
                              method = c("center", "scale", "nzv", "pca"))
#The result tells that some components have been ignored.
volcano_nzv_pca
```

```
## Created from 1508 samples and 7 variables
##
## Pre-processing:
##   - centered (4)
##   - ignored (3)
##   - principal component signal extraction (4)
##   - scaled (4)
##
## PCA needed 4 components to capture 95 percent of the variance
```

```
# identify which variables were ignored, centered, scaled, etc
volcano_nzv_pca$method
```

```
## $center
## [1] "Number"    "Latitude"  "Longitude" "Elevation"
##
## $scale
## [1] "Number"    "Latitude"  "Longitude" "Elevation"
##
## $pca
## [1] "Number"    "Latitude"  "Longitude" "Elevation"
##
## $ignore
## [1] "Last_eruption" "Tectonic"      "Rock_Type"
```

Some components have been ignored: "Last_eruption", "Tectonic" and "Rock_Type". These are important predictors in the analysis and is mandatory to understand why have been ignored.

These three variables are factor. To be considered in the analysis is needed to convert them as number. Since the conversion need to be accurate in its significance, dummies has been used.

```
volcano_recipe <- recipe(volcano_type ~ ., data = volcanoes_work4)
volcano_dummies<- volcano_recipe %>% step_dummy(Tectonic, Rock_Type,
                                      Last_eruption, one_hot = TRUE,)
volcano_dummies <-prep(volcano_dummies, training= volcanoes_work4)
volcano_work_d<- bake(volcano_dummies, new_data = volcanoes_work4)
```

Now the new dataset is correct and is possible to run again the model after the database has been partitioned again in test and training dataset.

```
set.seed(1)
test_index <- createDataPartition(volcano_work_d$volcano_type, p = 0.9, list = FALSE)
volcano_d_training <- volcano_work_d[test_index,]
volcano_d_testing  <- volcano_work_d[-test_index,]
```

Here Resampling method using traincontrol to control of printing and resampling for train has been used.

```
fit_control <- trainControl(
  method = "oob", #oob beacuse we 'll use random forest
  number = 10)
```

Now that the train control has been defined let's run again the random forest and redo the prediction for this model.

```
set.seed(1)
volcano_rf_fit <- train(as.factor(volcano_type) ~ .,
                        data = volcano_d_training,
                        method = "ranger",
                        trControl = fit_control)
volcano_rf_fit
```

```
## Random Forest
##
## 1360 samples
##  410 predictor
##    5 classes: 'Other', 'Pyroclastic cone', 'Pyroclastic shield', 'Stratovolcano', 'Submarine'
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##     2   gini        0.4779412  0.0000000
##     2   extratrees  0.4779412  0.0000000
##   206   gini        0.6698529  0.4934193
##   206   extratrees  0.6507353  0.4579882
##   410   gini        0.6602941  0.4789155
##   410   extratrees  0.6558824  0.4728285
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 206, splitrule = gini
##  and min.node.size = 1.
```

```
#Re-predict outcome on a test set
volcanoes_rf_pred2 <- predict(volcano_rf_fit, volcano_d_testing)
```

With the step dummies correction accuracy improved. A last attempt that has been done is to perform another random forest in a different way using as set_mode (Classification) and a workflow.

```
volcano_rf_fit2 <-
  rand_forest(trees = 1000) %>%
  set_mode("classification") %>%
  set_engine("ranger")
volcano_rf_fit2
```

```
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   trees = 1000
##
## Computational engine: ranger
```

```
#Since "classification" as been used as Mode in RF, a workflow is
#needed to keep this working
#First do a new recipe
volcano_recipe2 <- recipe(volcano_type ~ ., data = volcano_work_d)
#this time use the workflow function to constrain recipe and model
volcano_wf <-
  workflow() %>%
  add_recipe(volcano_recipe2) %>%
  add_model(volcano_rf_fit2)
#Bootstrap will define the resamples in the analysis
volcano_boot <-
  volcano_work_d %>%
  bootstraps()
#fit resample will run the method using the workflow defined and
#resaple using boothstrap instead of test and straining
volcano_res <-
  fit_resamples(
    volcano_wf,
    resamples = volcano_boot,
    control = control_resamples(save_pred = TRUE))
```

4. Results

In the table has been reported the results of the methods used and their prediction.

Models	accuracy
Knn	0.58
first random forest	0.65
Second random forest (dummies correction)	0.67
Third random forest (dummies+workflow)	0.63

Models	accuracy
Predictions	
First random forest prediction	0.68
Second random forest prediction	0.70

In general, is possible to see an improvement in accuracy in the models and in models prediction, excluded the last one. Looking at the confusion matrices of the random forest models is possible to see that both are correct for the gini effect, the K is very similar: first model k=0.45, second model k=0.48. Looking at the confusion matrices of the predictions, an increase in overall accuracy reflects an adjustment in the single class ones.

```
confusionMatrix(volcanoes_rf_pred, as.factor(volcano_work_testing$volcano_type))
```

```
## Confusion Matrix and Statistics
##
##                     Reference
## Prediction           Other Pyroclastic cone Pyroclastic shield Stratovolcano
##   Other                 15                4                  4             7
##   Pyroclastic cone       1                4                  2             1
##   Pyroclastic shield     1                0                  8             1
##   Stratovolcano         20                3                  3            63
##   Submarine              0                0                  0             0
##                     Reference
## Prediction           Submarine
##   Other                      0
##   Pyroclastic cone           0
##   Pyroclastic shield         0
##   Stratovolcano              0
##   Submarine                 11
##
## Overall Statistics
##
##                Accuracy : 0.6824
##                  95% CI : (0.6009, 0.7564)
##     No Information Rate : 0.4865
##     P-Value [Acc > NIR] : 1.129e-06
##
##                   Kappa : 0.5034
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Other Class: Pyroclastic cone
## Sensitivity                0.4054                 0.36364
## Specificity                0.8649                 0.97080
## Pos Pred Value             0.5000                 0.50000
## Neg Pred Value             0.8136                 0.95000
## Prevalence                 0.2500                 0.07432
## Detection Rate             0.1014                 0.02703
## Detection Prevalence       0.2027                 0.05405
```

```
## Balanced Accuracy          0.6351                0.66722
##                       Class: Pyroclastic shield Class: Stratovolcano
## Sensitivity                             0.47059               0.8750
## Specificity                             0.98473               0.6579
## Pos Pred Value                          0.80000               0.7079
## Neg Pred Value                          0.93478               0.8475
## Prevalence                              0.11486               0.4865
## Detection Rate                          0.05405               0.4257
## Detection Prevalence                    0.06757               0.6014
## Balanced Accuracy                       0.72766               0.7664
##                       Class: Submarine
## Sensitivity                    1.00000
## Specificity                    1.00000
## Pos Pred Value                 1.00000
## Neg Pred Value                 1.00000
## Prevalence                     0.07432
## Detection Rate                 0.07432
## Detection Prevalence           0.07432
## Balanced Accuracy              1.00000
```

In the first model confusion matrix is possible to see that the statistics by class are strange in some volcano type. Beside the balanced accuracy that is very high in submarine class (near 1), also sensitivity and specificity are near 1. The opposite trend is visible in Pyroclastic cone class, here balanced accuracy is the lowest in between the classes together with sensitivity and specificity is very high. All the other classes are in between. These results can make sense in class term since the pyroclastic cone class is the less abundant but submarine is not the most abundant. Since results are "too perfect" to be volcanic eruption predictions, the correction is needed.

```
confusionMatrix(volcanoes_rf_pred2,as.factor(volcano_d_testing$volcano_type))
```

```
## Confusion Matrix and Statistics
##
##                     Reference
## Prediction           Other Pyroclastic cone Pyroclastic shield Stratovolcano
##   Other                 19                6                  3             4
##   Pyroclastic cone       3                0                  0             0
##   Pyroclastic shield     0                0                  9             2
##   Stratovolcano         13                5                  5            66
##   Submarine              2                0                  0             0
##                     Reference
## Prediction           Submarine
##   Other                      1
##   Pyroclastic cone           0
##   Pyroclastic shield         0
##   Stratovolcano              0
##   Submarine                 10
##
## Overall Statistics
##
##                Accuracy : 0.7027
##                  95% CI : (0.6221, 0.775)
##     No Information Rate : 0.4865
##     P-Value [Acc > NIR] : 8.017e-08
```

```
##
##                   Kappa : 0.5323
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Other Class: Pyroclastic cone
## Sensitivity                0.5135                 0.00000
## Specificity                0.8739                 0.97810
## Pos Pred Value             0.5758                 0.00000
## Neg Pred Value             0.8435                 0.92414
## Prevalence                 0.2500                 0.07432
## Detection Rate             0.1284                 0.00000
## Detection Prevalence       0.2230                 0.02027
## Balanced Accuracy          0.6937                 0.48905
##                      Class: Pyroclastic shield Class: Stratovolcano
## Sensitivity                            0.52941               0.9167
## Specificity                            0.98473               0.6974
## Pos Pred Value                         0.81818               0.7416
## Neg Pred Value                         0.94161               0.8983
## Prevalence                             0.11486               0.4865
## Detection Rate                         0.06081               0.4459
## Detection Prevalence                   0.07432               0.6014
## Balanced Accuracy                      0.75707               0.8070
##                      Class: Submarine
## Sensitivity                   0.90909
## Specificity                   0.98540
## Pos Pred Value                0.83333
## Neg Pred Value                0.99265
## Prevalence                    0.07432
## Detection Rate                0.06757
## Detection Prevalence          0.08108
## Balanced Accuracy             0.94725
```

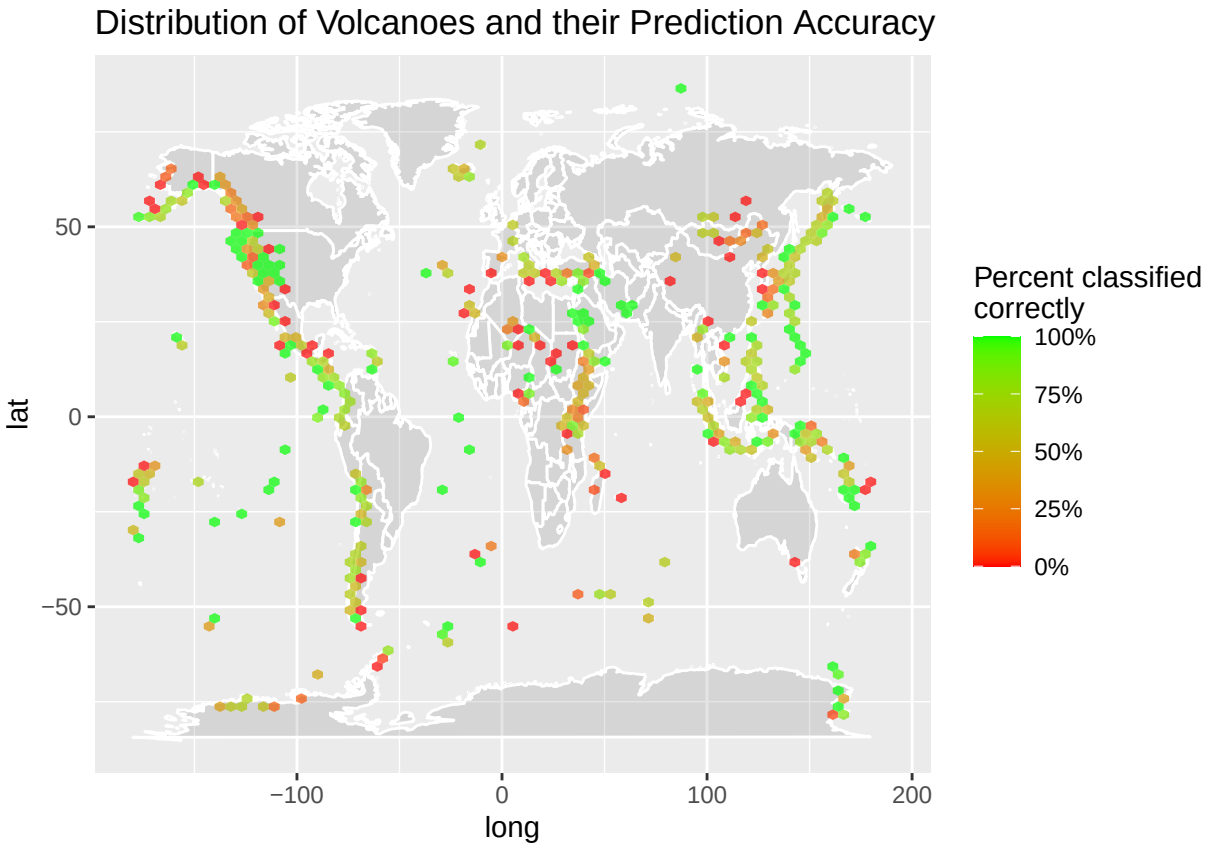
The second confusion matrix for the second prediction, have higher overall accuracy. Looking at the single class statistics, is possible to see that the higher accuracy is given by Submarine volcanoes again, also if it is lower than before, is near 1 again with a very high but more reasonable sensitivity and specificity. It makes sense that these value are very high since the submarine class cover up all the volcanoes eruption below sea-level independently to the type of volcano erupting. Pyroclastic cones are less abundant than the other classes and are difficult to predict since they are a peculiar type of cones so it makes sense that the sensitivity is 0 and the specificity is high, giving also a balanced accuracy of 0.50. Also Pyroclastic Shields are a peculiar type of shield volcanoes, but they are more abundant on the planet so the balanced accuracy near 0.7 together with the sensitivity and specificity make sense. Stratovolcanoes class represent the more abundant volcano type of the planet and most recognizable one, in fact the sensitivity is even higher that the submarine ones and the specificity is lower. The balanced accuracy for this class is very good. Last one to consider is the "other" class that include all the volcanoes remaining, it is not surprising, since it covers less abundant and more differenced typology that its statistics are these. However, the general increase in overall accuracy and decrease in single class statistics make the second random forest model the best one in predicting volcanoes eruption.

```
#Here the new results
volcano_res %>%
  collect_metrics()
```

```
## # A tibble: 2 x 5
##   .metric  .estimator  mean     n std_err
##   <chr>    <chr>      <dbl> <int>   <dbl>
## 1 accuracy multiclass 0.629    25 0.00362
## 2 roc_auc  hand_till  0.820    25 0.00190
```

Last results to discuss is the random forest model with classification output and the use of the workflow. The accuracy is lower than the other classes, however it is reasonable, over 0.6. It must be considered that the roc_auc is over 0.8 and the standard errors are quite low both for accuracy and roc_auc. In data science usually we consider the best model the one with higher accuracy in general, however in geology, the 0.6 is a very good prediction and seems very reasonable. Using this last random forest algorithm, is possible to construct a prediction and plot it in map to make it easyes to understand it.

```
volcano_pred <-
  volcano_res %>%
  collect_predictions() %>%
  mutate(correct = volcano_type == .pred_class) %>%
  left_join(
    volcano_work_d %>%
      mutate(.row = row_number())
  )
```



5. Conclusions

The best model to predict volcanoes eruption is the random forest, since all the prediction proposed is based on classification, in particular the one corrected using dummies. In this way all the predictors had

been considered and used to compute the models. A larger database and a stronger computer are needed to improve the model, since the model has been based only on 1058 observation that in data science are very few and the laptop used took 20-30 minutes to elaborate the script. The real big limitation of this kind of data is the limitation. This limitation is given by the phenomenon itself: volcanic eruptions are rare, geologically speaking, and cover a very large range of typology that cross from the geysers to Etna spectacular eruptions. Just think that the eruptions observed actually during Holocene are few and cover only the last 2000 year roughly. These events are also difficult to observe and date, for instance submarine events and Early Holocene events.

References

-Understanding Earth, by John Grotzinger and Tom Jordan (2010).

-Volcanoes, by Peter Francis (2003).

-Scienze della terra. Vol. 1: Elementi di geologia generale, by Pompeo L. Casati, (2012).

-Introduction to Data Science, Rafael A. Irizarry.