Assignment 4: Data Exploration & R

Chrissie Raj

G01465544
chrissie.raj123@gmail.com

AIT 580-010

Assignment 4: Data Exploration & R

Purpose:
Demonstrate exploration of data via creation of statistical tables and visualizations using R; match appropriate summary statistics and graphs to the dataset's NOIR data types.

Points: 100

Deliverables:

Review IDMA Chapter 4 and author slide presentation

Review IDMA Chapter 8 (R) and author slide presentation

Load the significant-volcanic-eruption-database.csv dataset into an R data frame; display a few records

o Source: see Code and Data folder

o Read the details about the dataset

o Remove the # comments from the dataset

o Add a row number column to the dataset

Use R to answer and interpret the following (duplicate & check the Python results:

o Display appropriate and labeled summary statistics and visualizations for: Country

Volcano Type

Elevation

Volcanic Explosivity Index (VEI)

Volcano : Deaths

Relationship between Volcano Type and VEI

Relationship between VEI and Deaths

Files:

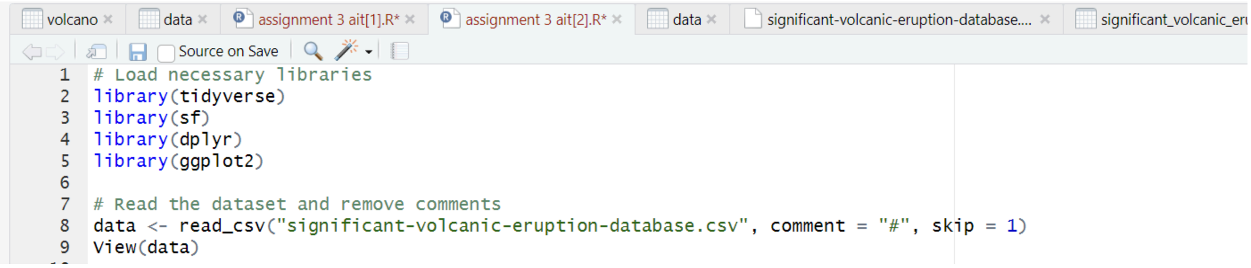
9781284180923_SLID_CH04.pptx 9781284180923_SLID_CH08B.pptx

Tools:

R IDE (your choice, but RStudio is easiest)

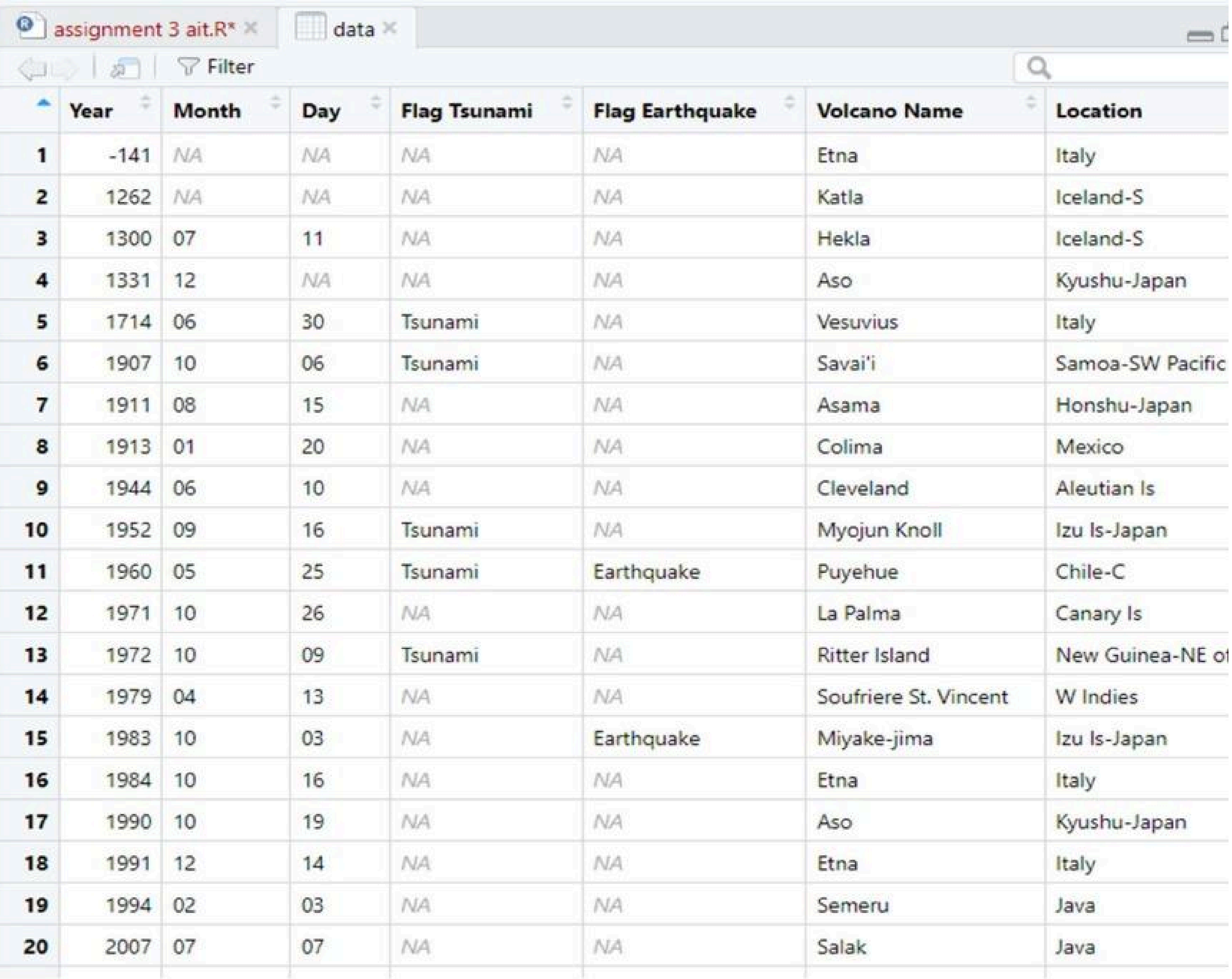
ANSWER:

#1

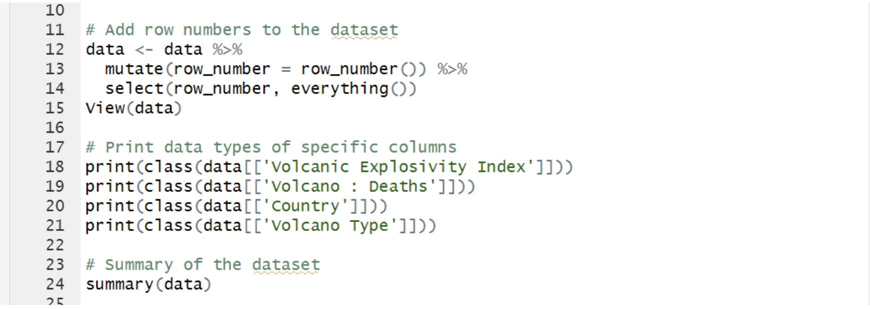


```
# Load necessary libraries
library(tidyverse)
library(sf)
library(dplyr)
library(ggplot2)

# Read the dataset and remove comments
data <- read_csv("significant-volcanic-eruption-database.csv", comment = "#", skip = 1)
View(data)
```

This code block loads necessary R packages (`tidyverse`, `sf`, `dplyr`, `ggplot2`), reads a CSV file ("significant-volcanic-eruption-database.csv") excluding lines starting with `#`, and displays the resulting dataset (`data`) for inspection using `View(data)`.

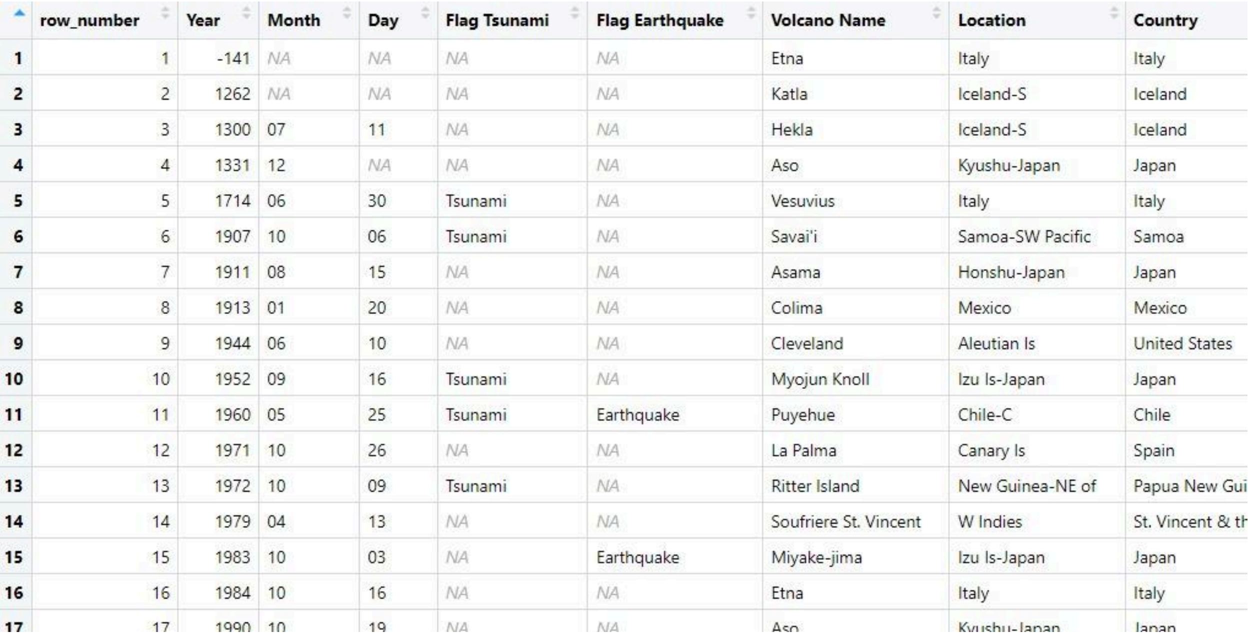


	Year	Month	Day	Flag Tsunami	Flag Earthquake	Volcano Name	Location
1	-141	NA	NA	NA	NA	Etna	Italy
2	1262	NA	NA	NA	NA	Katla	Iceland-S
3	1300	07	11	NA	NA	Hekla	Iceland-S
4	1331	12	NA	NA	NA	Aso	Kyushu-Japan
5	1714	06	30	Tsunami	NA	Vesuvius	Italy
6	1907	10	06	Tsunami	NA	Savai'i	Samoa-SW Pacific
7	1911	08	15	NA	NA	Asama	Honshu-Japan
8	1913	01	20	NA	NA	Colima	Mexico
9	1944	06	10	NA	NA	Cleveland	Aleutian Is
10	1952	09	16	Tsunami	NA	Myojun Knoll	Izu Is-Japan
11	1960	05	25	Tsunami	Earthquake	Puyehue	Chile-C
12	1971	10	26	NA	NA	La Palma	Canary Is
13	1972	10	09	Tsunami	NA	Ritter Island	New Guinea-NE o
14	1979	04	13	NA	NA	Soufriere St. Vincent	W Indies
15	1983	10	03	NA	Earthquake	Miyake-jima	Izu Is-Japan
16	1984	10	16	NA	NA	Etna	Italy
17	1990	10	19	NA	NA	Aso	Kyushu-Japan
18	1991	12	14	NA	NA	Etna	Italy
19	1994	02	03	NA	NA	Semeru	Java
20	2007	07	07	NA	NA	Salak	Java

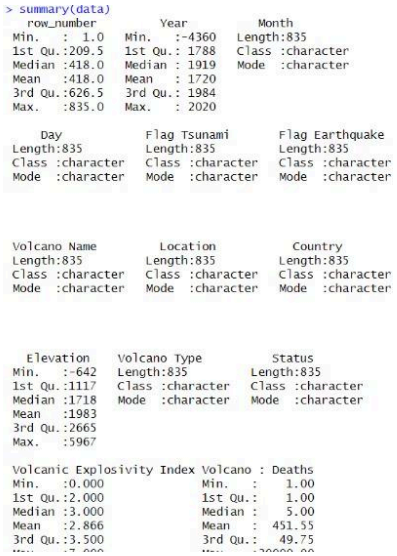


```

# Add row numbers to the dataset
data <- data %>%
  mutate(row_number = row_number()) %>%
  select(row_number, everything())
View(data)

# Print data types of specific columns
print(class(data[['Volcanic Explosivity Index']]))
print(class(data[['Volcano : Deaths']]))
print(class(data[['Country']]))
print(class(data[['Volcano Type']]))

# Summary of the dataset
summary(data)
```

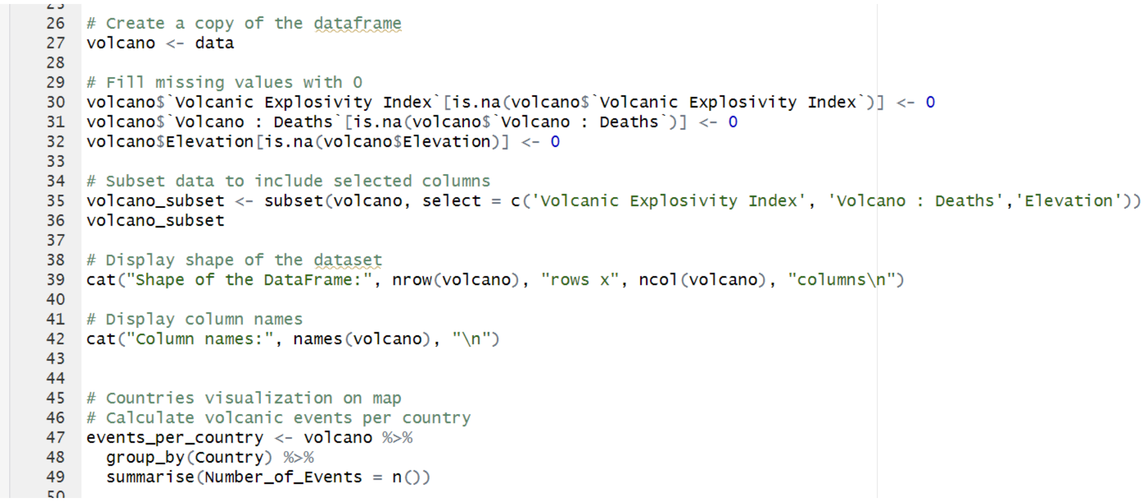
This code adds a new `row_number` column to `data`, giving each row a unique number. It then displays `data` in a viewer for inspection, prints the data types of selected columns, and summarizes the entire data frame with `summary()`. This sequence allows for a quick overview of `data`'s structure, variable types, and basic statistics.



	row_number	Year	Month	Day	Flag Tsunami	Flag Earthquake	Volcano Name	Location	Country
1	1	-141	NA	NA	NA	NA	Etna	Italy	Italy
2	2	1262	NA	NA	NA	NA	Katla	Iceland-S	Iceland
3	3	1300	07	11	NA	NA	Hekla	Iceland-S	Iceland
4	4	1331	12	NA	NA	NA	Aso	Kyushu-Japan	Japan
5	5	1714	06	30	Tsunami	NA	Vesuvius	Italy	Italy
6	6	1907	10	06	Tsunami	NA	Savai'i	Samoa-SW Pacific	Samoa
7	7	1911	08	15	NA	NA	Asama	Honshu-Japan	Japan
8	8	1913	01	20	NA	NA	Colima	Mexico	Mexico
9	9	1944	06	10	NA	NA	Cleveland	Aleutian Is	United States
10	10	1952	09	16	Tsunami	NA	Myojun Knoll	Izu Is-Japan	Japan
11	11	1960	05	25	Tsunami	Earthquake	Puyehue	Chile-C	Chile
12	12	1971	10	26	NA	NA	La Palma	Canary Is	Spain
13	13	1972	10	09	Tsunami	NA	Ritter Island	New Guinea-NE of	Papua New Gui
14	14	1979	04	13	NA	NA	Soufriere St. Vincent	W Indies	St. Vincent & th
15	15	1983	10	03	NA	Earthquake	Miyake-jima	Izu Is-Japan	Japan
16	16	1984	10	16	NA	NA	Etna	Italy	Italy
17	17	1990	10	19	NA	NA	Aso	Kyushu-Japan	Japan

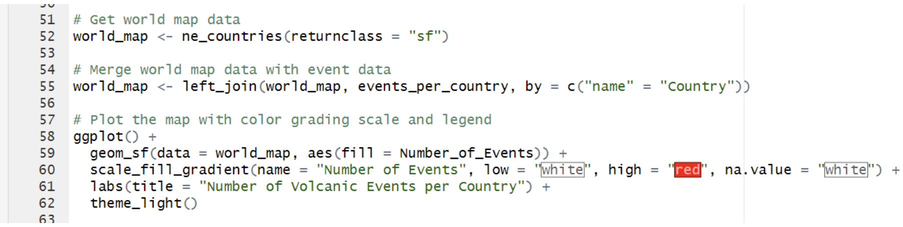


```
> summary(data)
  row_number         Year           Month
 Min.   :  1.0   Min.   :-4360   Length:835
 1st Qu.:209.5   1st Qu.: 1788   Class :character
 Median :418.0   Median : 1919   Mode  :character
 Mean   :418.0   Mean   : 1720
 3rd Qu.:626.5   3rd Qu.: 1984
 Max.   :835.0   Max.   : 2020

     Day            Flag Tsunami       Flag Earthquake
 Length:835         Length:835         Length:835
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



 Volcano Name         Location           Country
 Length:835         Length:835         Length:835
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



   Elevation     Volcano Type          Status
 Min.   :-642   Length:835         Length:835
 1st Qu.:1117   Class :character   Class :character
 Median :1718   Mode  :character   Mode  :character
 Mean   :1983
 3rd Qu.:2665
 Max.   :5967

 Volcanic Explosivity Index Volcano : Deaths
 Min.   :0.000              Min.   :    1.00
 1st Qu.:2.000              1st Qu.:    1.00
 Median :3.000              Median :    5.00
 Mean   :2.866              Mean   :  451.55
 3rd Qu.:3.500              3rd Qu.:   49.75
 Max.   :7.000              Max.   :30000.00
```



```
# Create a copy of the dataframe
volcano <- data

# Fill missing values with 0
volcano$`Volcanic Explosivity Index`[is.na(volcano$`Volcanic Explosivity Index`)] <- 0
volcano$`Volcano : Deaths`[is.na(volcano$`Volcano : Deaths`)] <- 0
volcano$Elevation[is.na(volcano$Elevation)] <- 0

# Subset data to include selected columns
volcano_subset <- subset(volcano, select = c('Volcanic Explosivity Index', 'Volcano : Deaths','Elevation'))
volcano_subset

# Display shape of the dataset
cat("Shape of the DataFrame:", nrow(volcano), "rows x", ncol(volcano), "columns\n")

# Display column names
cat("Column names:", names(volcano), "\n")


# Countries visualization on map
# Calculate volcanic events per country
events_per_country <- volcano %>%
  group_by(Country) %>%
  summarise(Number_of_Events = n())
```

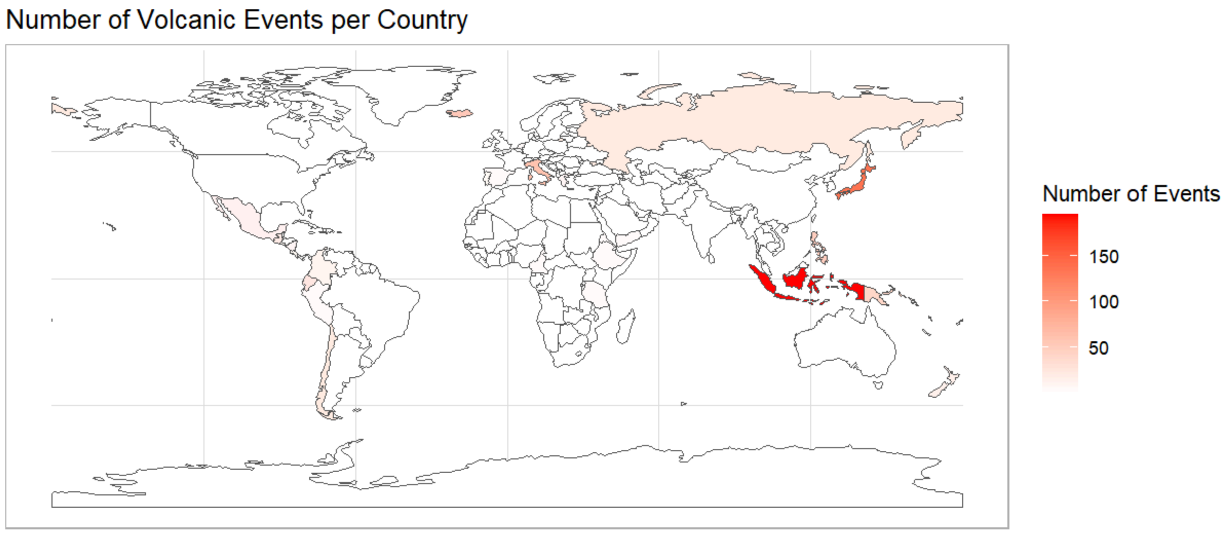
This code creates a copy of `data` called `volcano`, fills missing values in three columns with 0, displays these columns in `volcano_subset`, prints the data frame's shape (number of rows and columns), and then prints the column names of `volcano`. This helps to quickly check for missing values, view specific columns, and get an overview of the data frame's structure and variables.

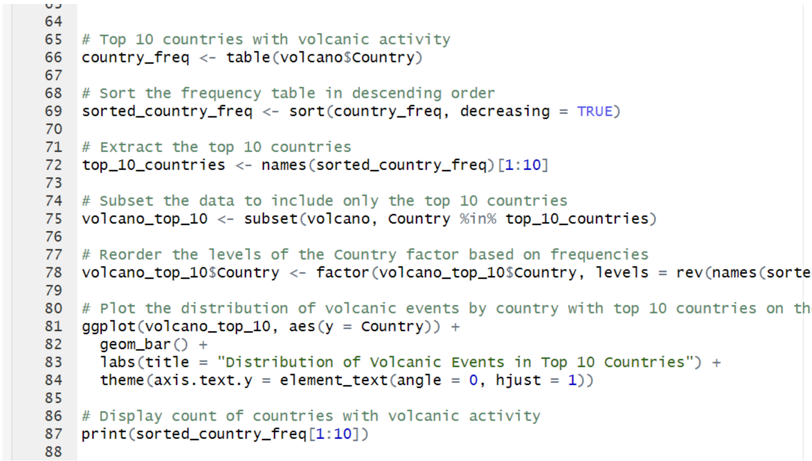


```
# Get world map data
world_map <- ne_countries(returnclass = "sf")

# Merge world map data with event data
world_map <- left_join(world_map, events_per_country, by = c("name" = "Country"))

# Plot the map with color grading scale and legend
ggplot() +
  geom_sf(data = world_map, aes(fill = Number_of_Events)) +
  scale_fill_gradient(name = "Number of Events", low = "white", high = "red", na.value = "white") +
  labs(title = "Number of Volcanic Events per Country") +
  theme_light()
```

This code calculates the number of volcanic events per country, then creates a world map visualization. It uses `volcano` data to count events by country, merges this data with world map data, and plots the map with colored countries representing event frequencies. This concise sequence provides a quick visual overview of volcanic event distribution across countries.

Plot:

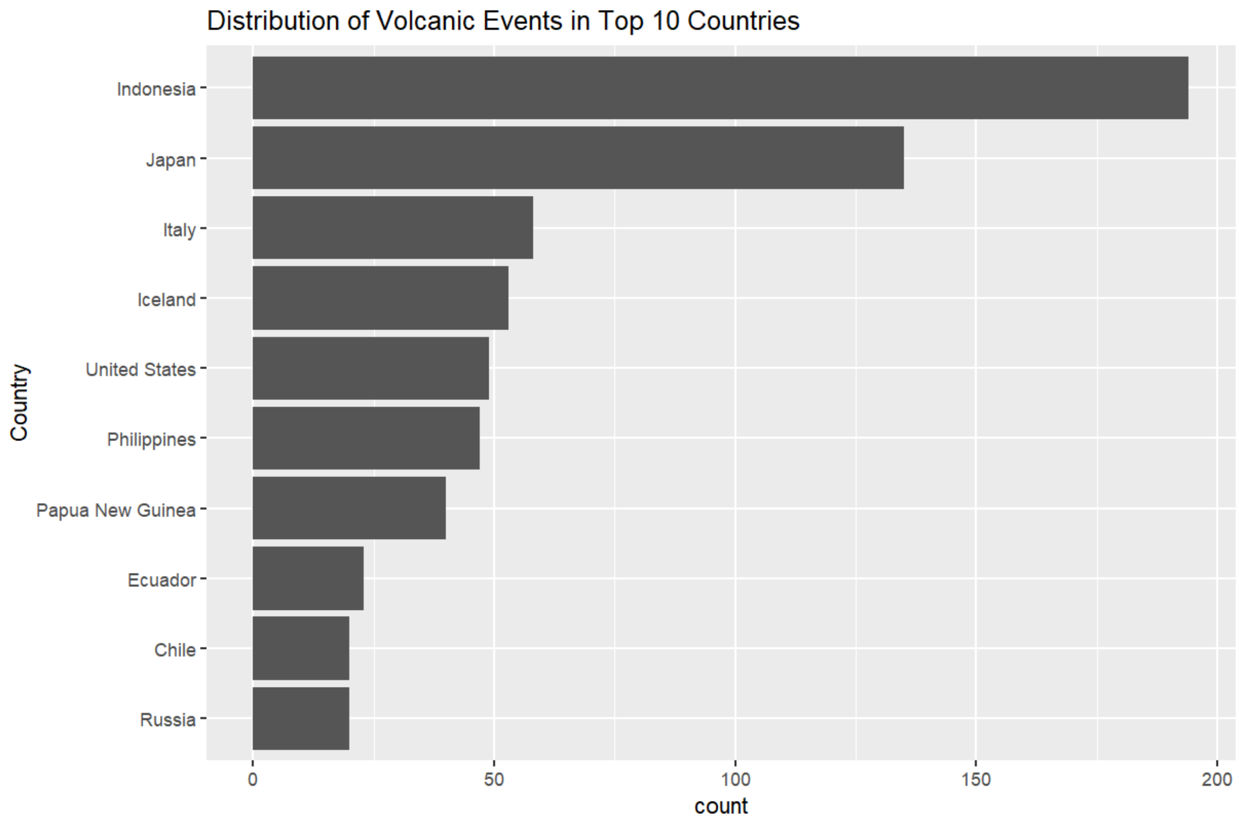

Number of Volcanic Events per Country
Number of Events
150
100
50

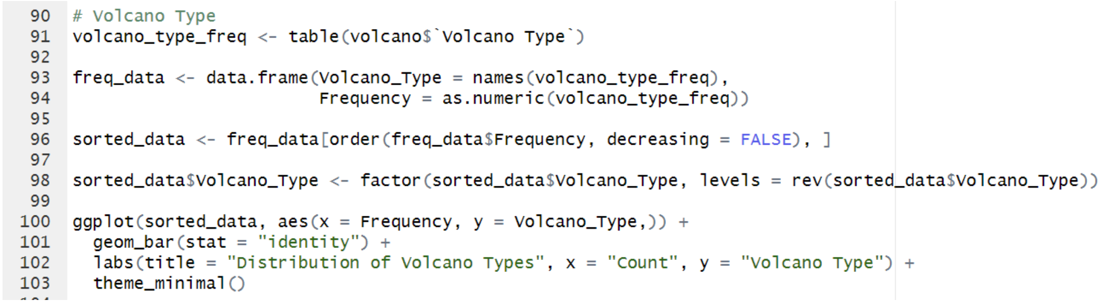


```
# Top 10 countries with volcanic activity
country_freq <- table(volcano$Country)

# Sort the frequency table in descending order
sorted_country_freq <- sort(country_freq, decreasing = TRUE)

# Extract the top 10 countries
top_10_countries <- names(sorted_country_freq)[1:10]

# Subset the data to include only the top 10 countries
volcano_top_10 <- subset(volcano, Country %in% top_10_countries)

# Reorder the levels of the Country factor based on frequencies
volcano_top_10$Country <- factor(volcano_top_10$Country, levels = rev(names(sorted_country_freq)))

# Plot the distribution of volcanic events by country with top 10 countries on the y-axis
ggplot(volcano_top_10, aes(y = Country)) +
  geom_bar() +
  labs(title = "Distribution of Volcanic Events in Top 10 Countries") +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))

# Display count of countries with volcanic activity
print(sorted_country_freq[1:10])

```

This code calculates the frequency of volcanic activity for each country, selects the top 10 countries with the most activity, and plots a bar graph showing the distribution of events. The plot is titled "Distribution of Volcanic Events in Top 10 Countriers)" with the y-axis representing the countries. Additionally, it prints the count of volcanic events for these top 10 countries. This succinct script provides a quick insight into the distribution and frequency of volcanic events across the top 10 countries.

PLOT:

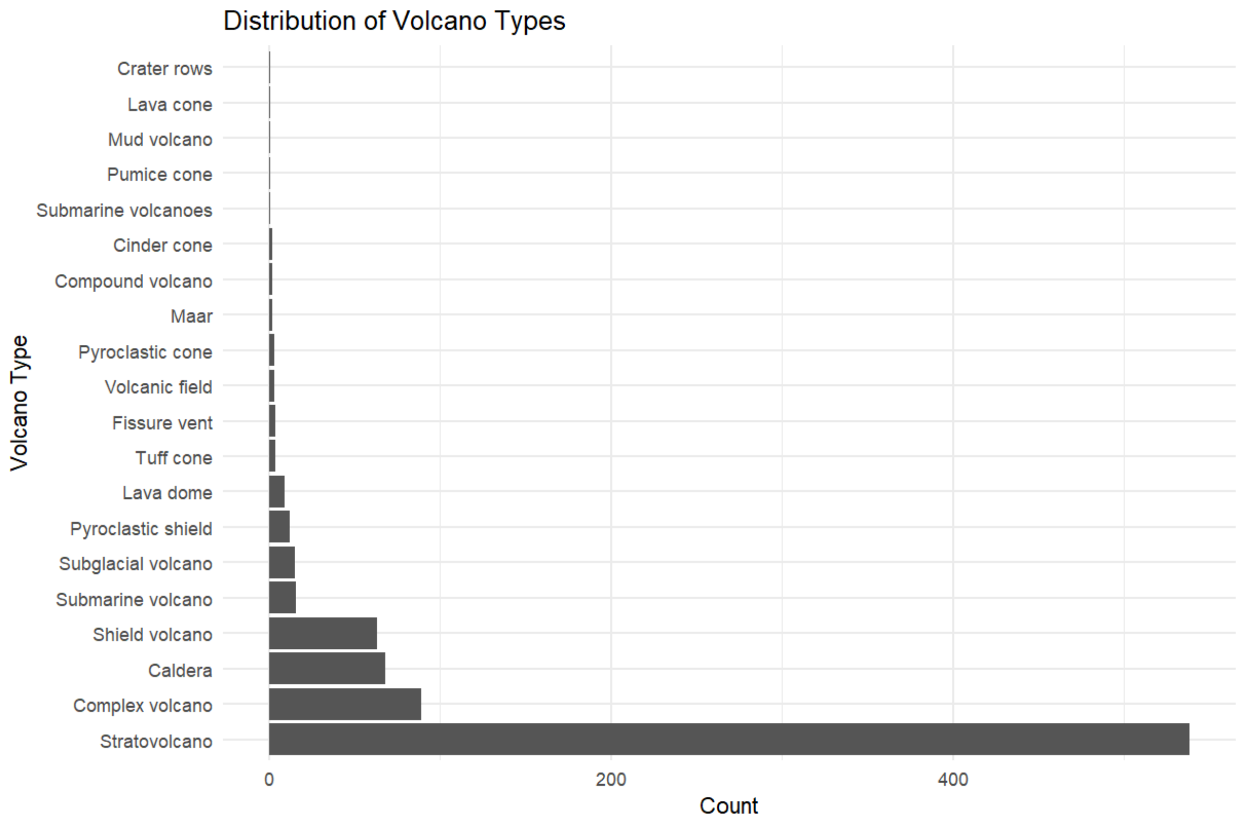


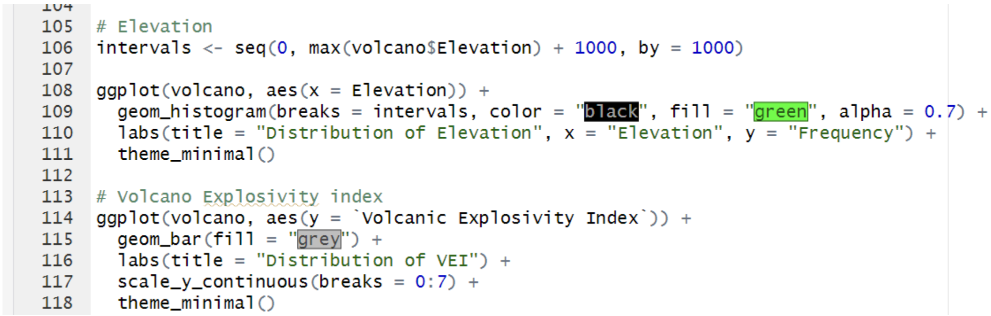


```
# Volcano Type
volcano_type_freq <- table(volcano$`Volcano Type`)

freq_data <- data.frame(Volcano_Type = names(volcano_type_freq),
                        Frequency = as.numeric(volcano_type_freq))

sorted_data <- freq_data[order(freq_data$Frequency, decreasing = FALSE), ]

sorted_data$Volcano_Type <- factor(sorted_data$Volcano_Type, levels = rev(sorted_data$Volcano_Type))

ggplot(sorted_data, aes(x = Frequency, y = Volcano_Type,)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Volcano Types", x = "Count", y = "Volcano Type") +
  theme_minimal()
```

This code creates a frequency table for volcano types in the `Volcano Type` column of the `volcano` data frame. It then converts this table into a sorted data frame and reverses the factor levels for plotting. This succinctly prepares the data for visualizing the distribution of volcano types, offering a quick overview of their occurrence frequencies.

PLOT:

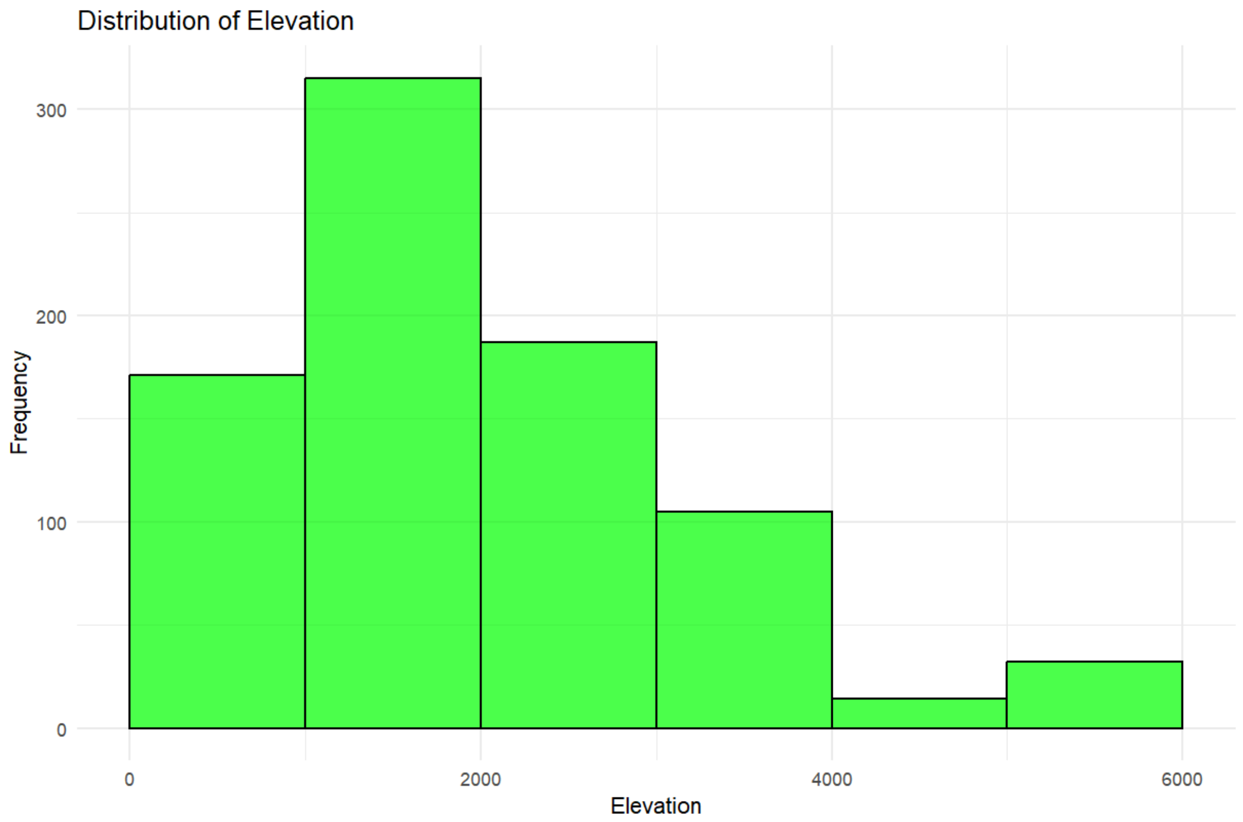




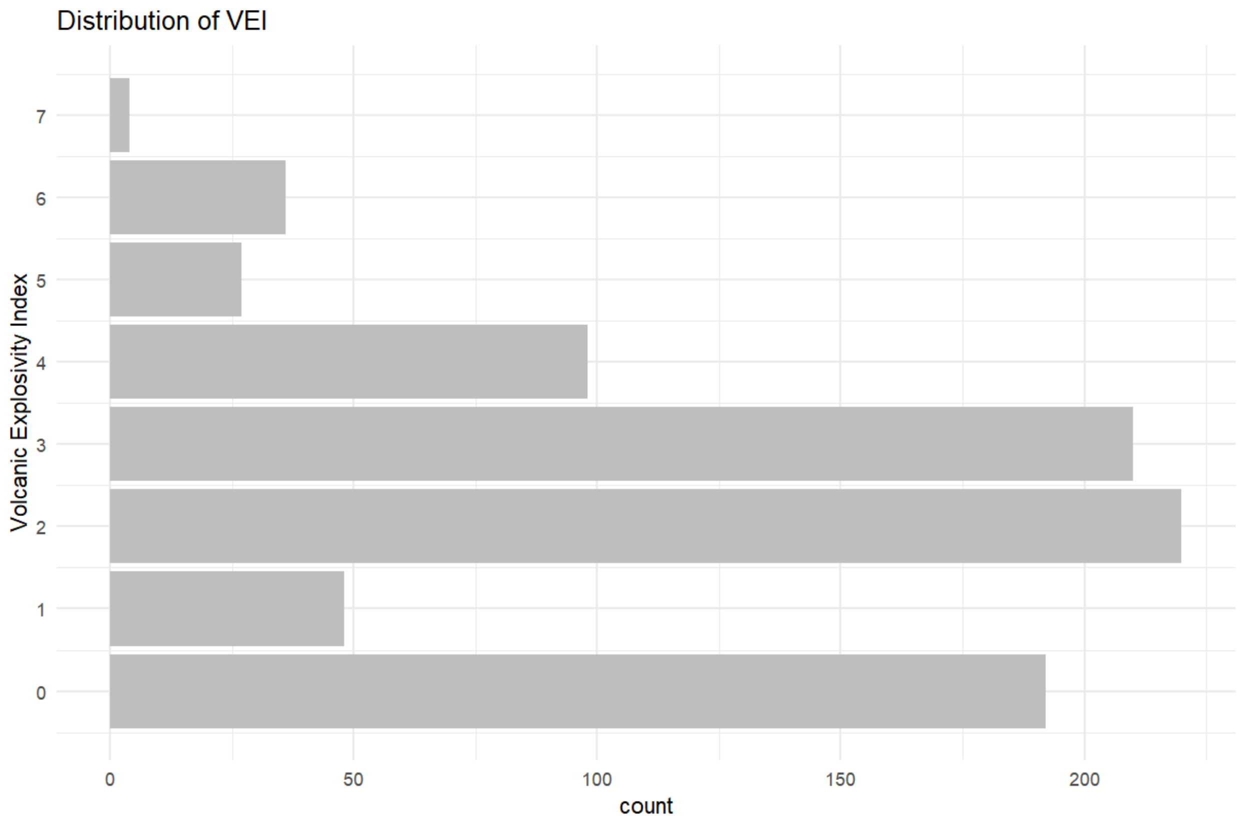
```
# Elevation
intervals <- seq(0, max(volcano$Elevation) + 1000, by = 1000)

ggplot(volcano, aes(x = Elevation)) +
  geom_histogram(breaks = intervals, color = "black", fill = "green", alpha = 0.7) +
  labs(title = "Distribution of Elevation", x = "Elevation", y = "Frequency") +
  theme_minimal()

# Volcano Explosivity index
ggplot(volcano, aes(y = `Volcanic Explosivity Index`)) +
  geom_bar(fill = "grey") +
  labs(title = "Distribution of VEI") +
  scale_y_continuous(breaks = 0:7) +
  theme_minimal()
```

This code block creates histograms and bar plots for the `volcano` dataset. The first part defines `intervals` for the elevation histogram, then plots a histogram of `Elevation` with specified breaks and colors. The second part plots a bar plot of `Volcanic Explosivity Index` (`VEI`) with a grey fill, providing a distribution visualization for VEI. Both plots are styled using the `theme_minimal()` theme.

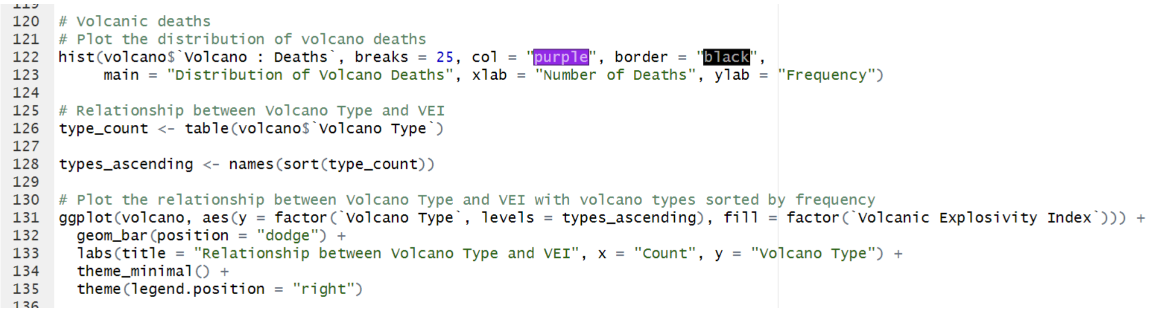
PLOT:


Distribution of Elevation
Frequency
300
200
100
0
0
2000
4000
6000
Elevation

PLOT:



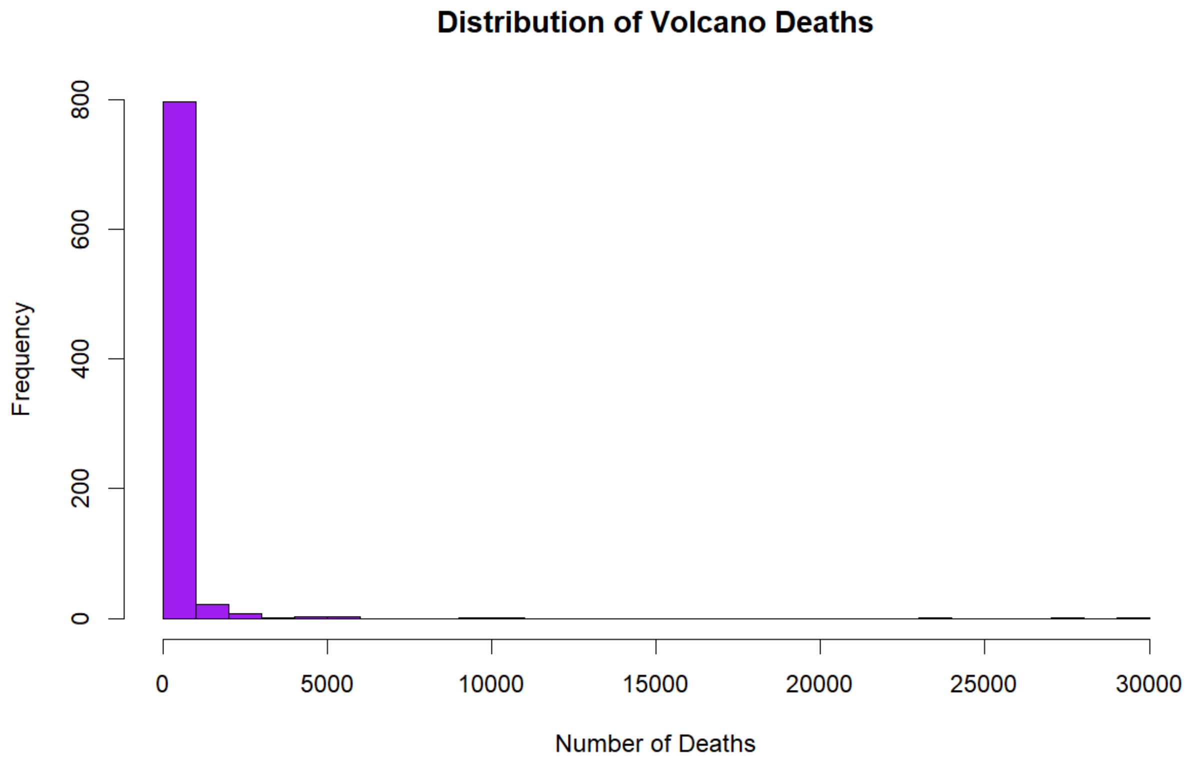
The distribution of Volcanic Explosivity Index (`VEI`) from the `volcano` data frame as a bar plot titled "Distribution of VEI", scaling the y-axis from 0 to 7. It then displays a histogram of volcano deaths from `volcano` with 25 breaks, titled "Distribution of Volcano Deaths". Finally, it prepares the `Volcano Type` data for analysis by sorting types in ascending order of frequency. These visualizations provide quick insights into the distribution of VEI, volcano deaths, and prepare the volcano type data for further examination.



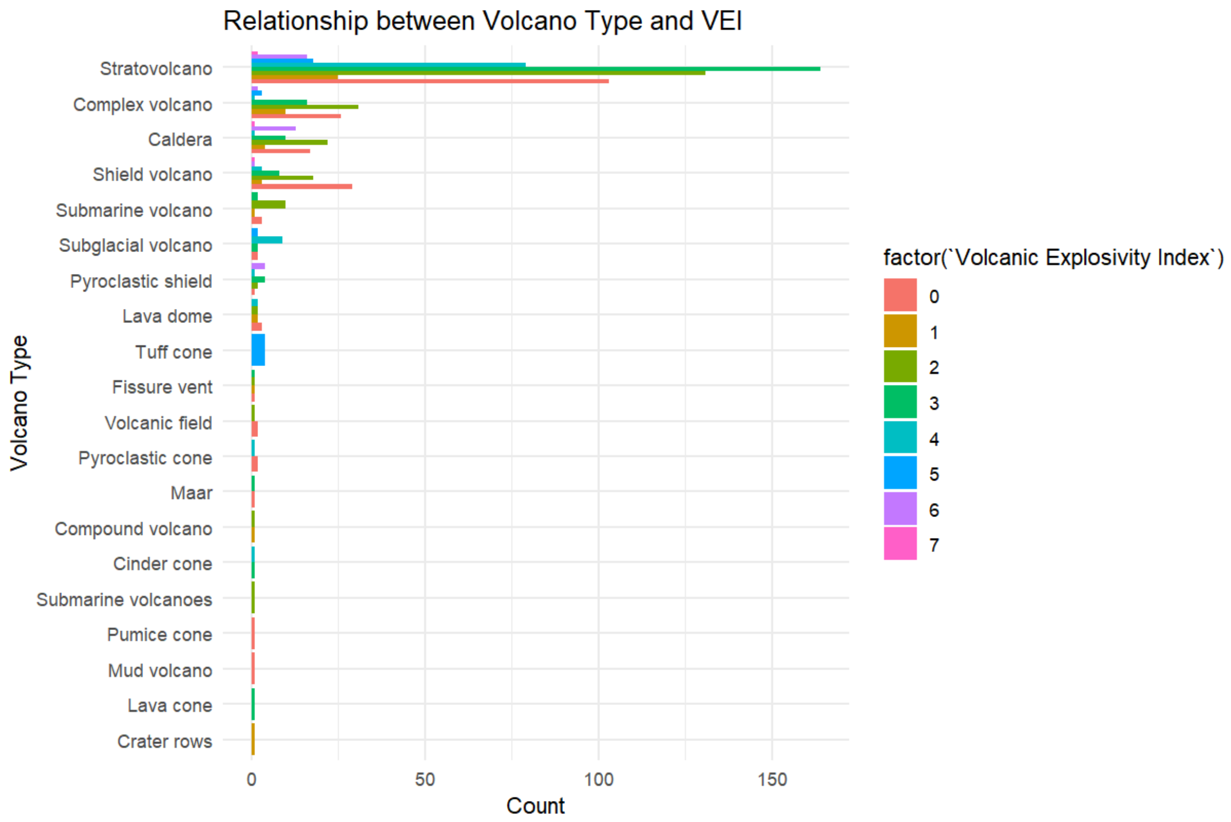
```r
# Volcanic deaths
# Plot the distribution of volcano deaths
hist(volcano$`Volcano : Deaths`, breaks = 25, col = "purple", border = "black",
     main = "Distribution of Volcano Deaths", xlab = "Number of Deaths", ylab = "Frequency")

# Relationship between Volcano Type and VEI
type_count <- table(volcano$`Volcano Type`)

types_ascending <- names(sort(type_count))

# Plot the relationship between Volcano Type and VEI with volcano types sorted by frequency
ggplot(volcano, aes(y = factor(`Volcano Type`, levels = types_ascending), fill = factor(`Volcanic Explosivity Index`))) +
  geom_bar(position = "dodge") +
  labs(title = "Relationship between Volcano Type and VEI", x = "Count", y = "Volcano Type") +
  theme_minimal() +
  theme(legend.position = "right")
```

This code block analyzes the `volcano` dataset, starting with a histogram displaying the distribution of volcano-related deaths (`Volcano : Deaths`). It then computes the frequency of each `Volcano Type` and sorts them in ascending order to prepare for a bar plot showing the relationship between `Volcano Type` and `Volcanic Explosivity Index` (`VEI`). The plot is designed with bars positioned using "dodge" and a minimal theme, highlighting the distribution. Finally, a scatter plot illustrates the relationship between `VEI` and `Volcano : Deaths`, colored red with 50% transparency, emphasizing the correlation between volcanic explosivity and fatalities. This succinctly visualizes key associations within the volcanic data.

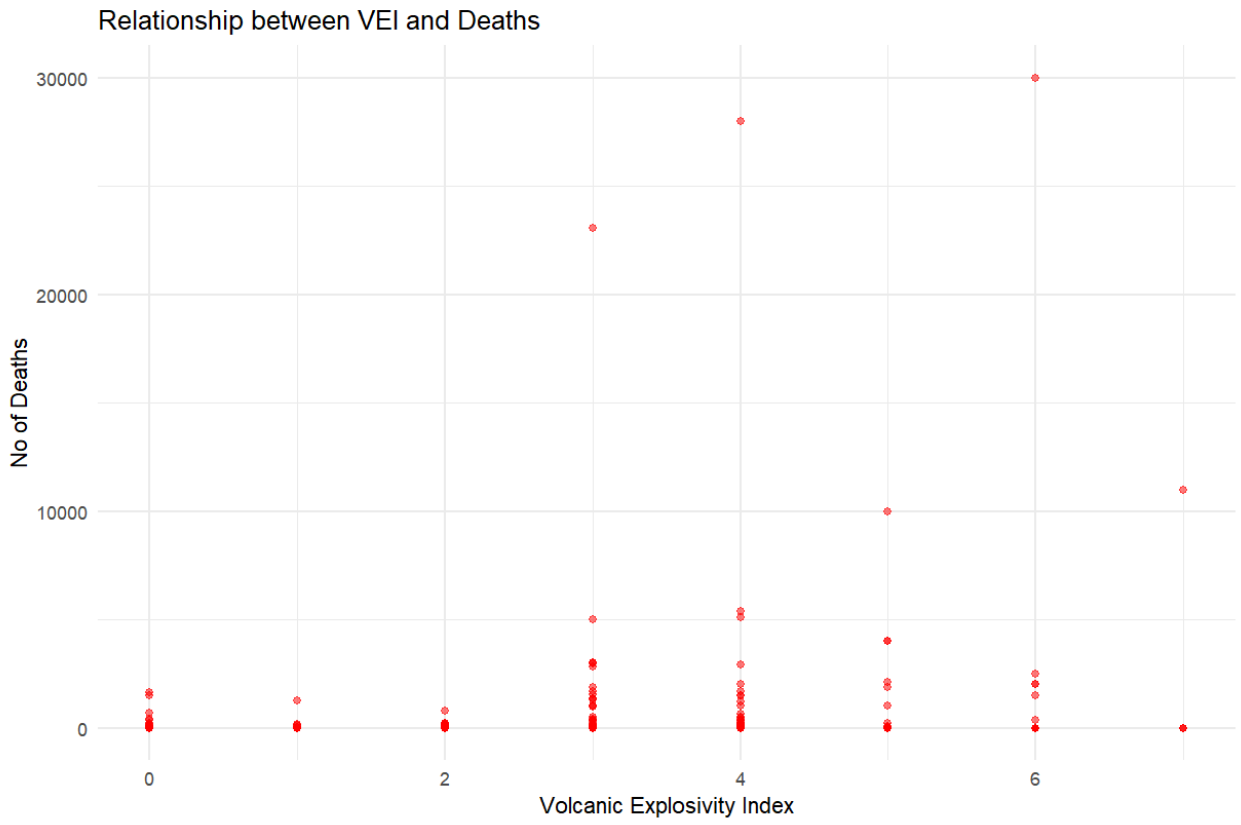
PLOT:


Distribution of Volcano Deaths
Frequency
0
200
400
600
800
0
5000
10000
15000
20000
25000
30000
Number of Deaths

PLOT:


Relationship between Volcano Type and VEI
Volcano Type
Stratovolcano
Complex volcano
Caldera
Shield volcano
Submarine volcano
Subglacial volcano
Pyroclastic shield
Lava dome
Tuff cone
Fissure vent
Volcanic field
Pyroclastic cone
Maar
Compound volcano
Cinder cone
Submarine volcanoes
Pumice cone
Mud volcano
Lava cone
Crater rows
0
50
100
150
Count
factor(`Volcanic Explosivity Index`)
0
1
2
3
4
5
6
7

PLOT:


Relationship between VEI and Deaths
No of Deaths
30000
20000
10000
0
0
2
4
6
Volcanic Explosivity Index