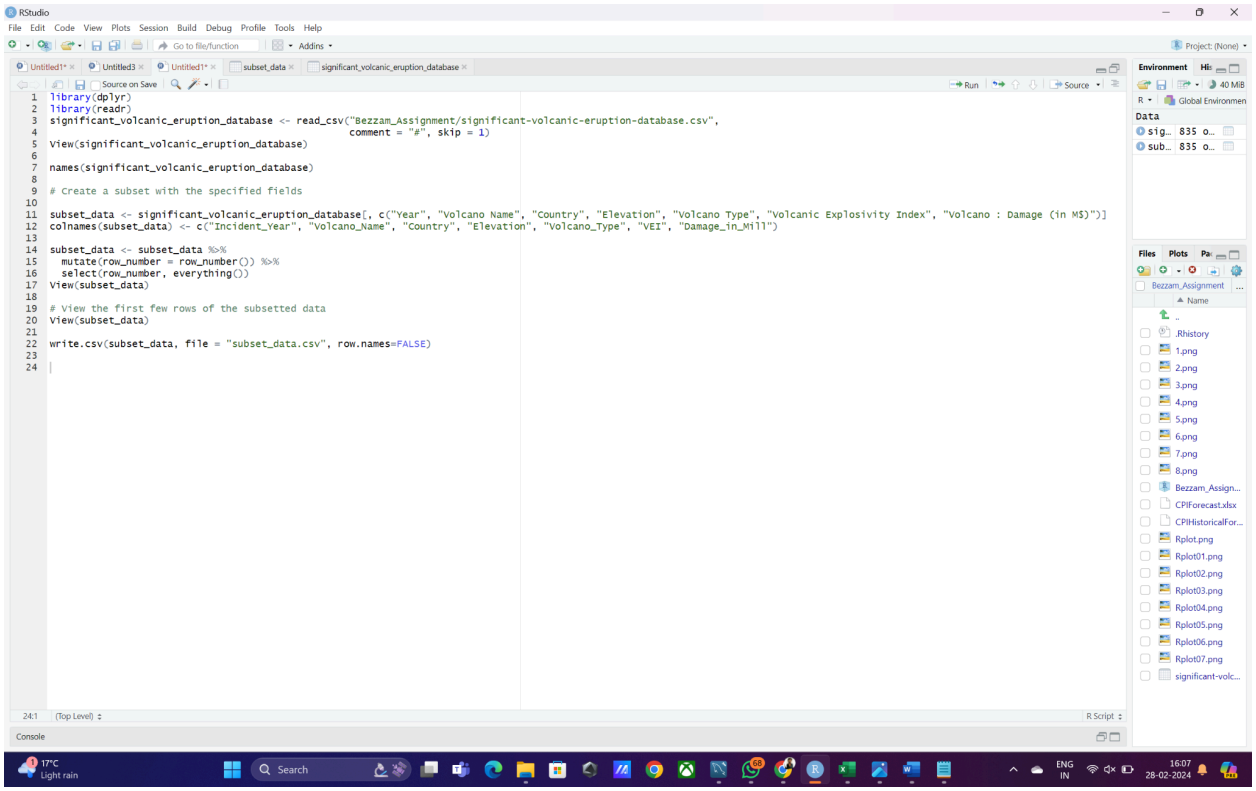
AIT ASSIGNMENT 5
CHRISSIE RAJ
G01465544

Purpose:
Demonstrate exploration of data via creation of statistical tables using RDBMS/SQL; matching appropriate summary statistics to the dataset's NOIR data types.

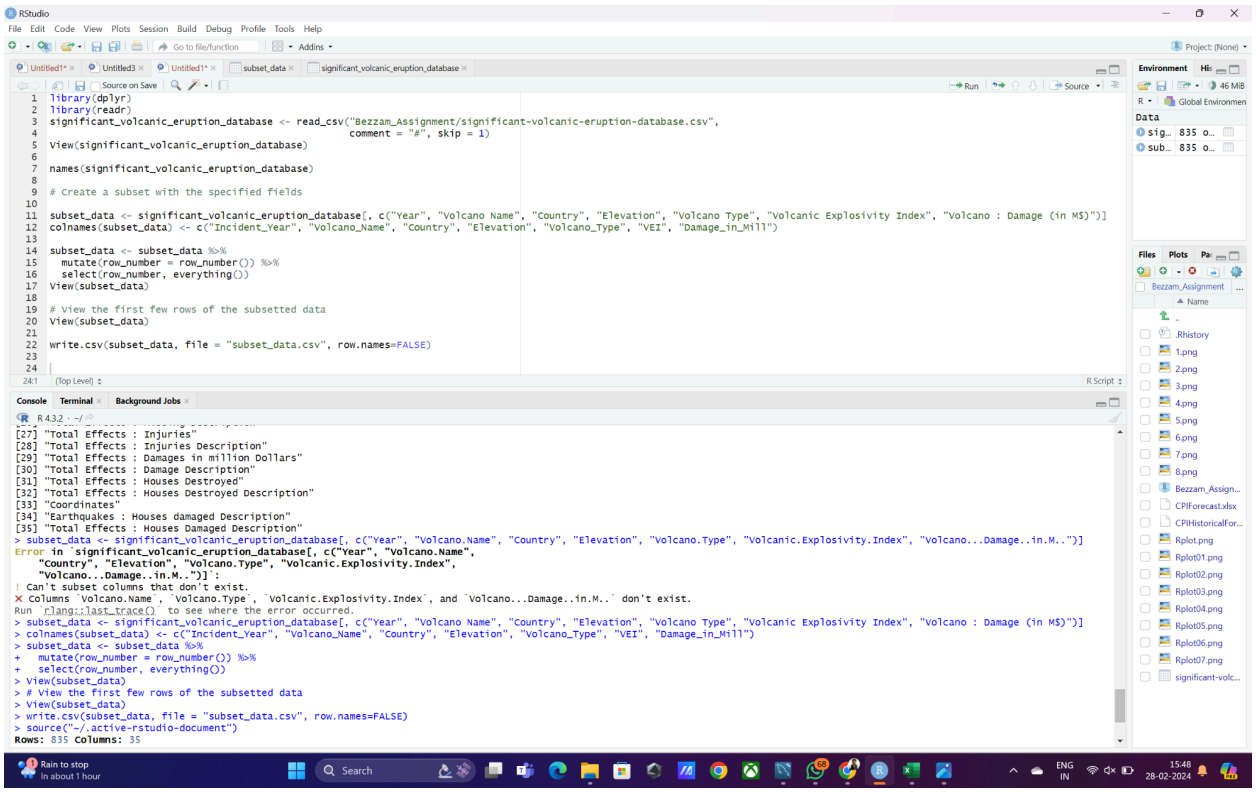
Points: 50

Deliverables:
- **Review IDMA Chapter 6 and author slide presentation**
  - **Pages 211-252 and 268-272**
- **Use the significant-volcanic-eruption-database.csv dataset (in the Code & Data folder)**
  - **Source: see Code and Data folder**
  - **Read the details about the dataset**
  - **Remove the # comments from the dataset**
  - **Create a subset of the dataset containing the following fields**
    Year, Volcano Name, Country, Elevation, Volcano Type, VEI, Damage in M$,
  - **Add a row number column to the dataset**
- **Create a SQL database schema and table for the subset dataset using any RDBMS (Oracle, MySQL, etc.)**
- **Load the dataset into the table; use an SQL query to display a few records**
- **Query the database and interpret the results, displaying:**
  - **The Country with the greatest number of eruptions**
  - **The Volcano with the greatest number of eruptions**
  - **The Volcano Type with the greatest VEI**
  - **The top 10 Volcano Names, Countries, & Volcano Types with the greatest amount of Damages in M$**

R:

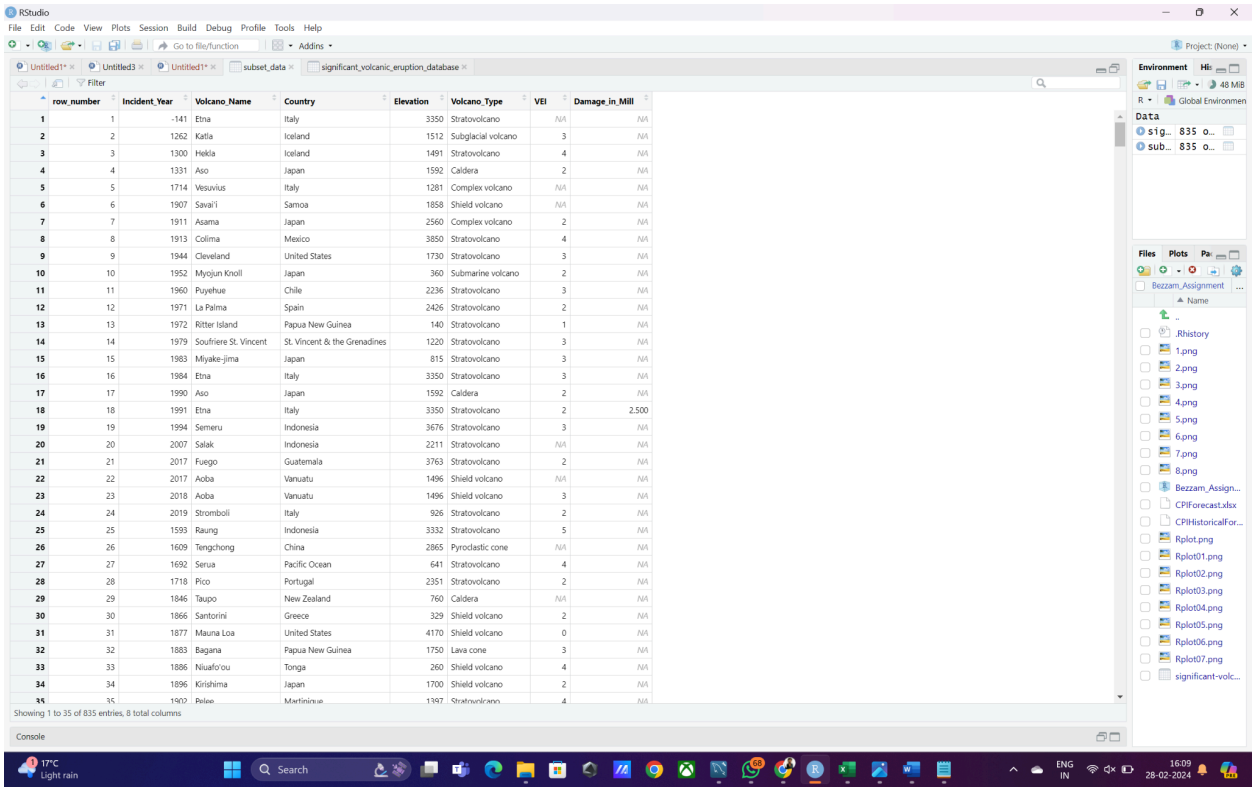


```
library(dplyr)
library(readr)
significant_volcanic_eruption_database <- read_csv("Bezzam_Assignment/significant-volcanic-eruption-database.csv",
                                                   comment = "#", skip = 1)
View(significant_volcanic_eruption_database)

names(significant_volcanic_eruption_database)

# Create a subset with the specified fields

subset_data <- significant_volcanic_eruption_database[, c("Year", "Volcano Name", "Country", "Elevation", "Volcano Type", "Volcanic Explosivity Index", "Volcano : Damage (in M$)")]
colnames(subset_data) <- c("Incident_Year", "Volcano_Name", "Country", "Elevation", "Volcano_Type", "VEI", "Damage_in_Mill")

subset_data <- subset_data %>%
  mutate(row_number = row_number()) %>%
  select(row_number, everything())
View(subset_data)

# View the first few rows of the subsetted data
View(subset_data)

write.csv(subset_data, file = "subset_data.csv", row.names=FALSE)
```



```
library(dplyr)
library(readr)
significant_volcanic_eruption_database <- read_csv("Bezzam_Assignment/significant-volcanic-eruption-database.csv",
                                                   comment = "#", skip = 1)
View(significant_volcanic_eruption_database)

names(significant_volcanic_eruption_database)

# Create a subset with the specified fields

subset_data <- significant_volcanic_eruption_database[, c("Year", "Volcano Name", "Country", "Elevation", "Volcano Type", "Volcanic Explosivity Index", "Volcano : Damage (in M$)")]
colnames(subset_data) <- c("Incident_Year", "Volcano_Name", "Country", "Elevation", "Volcano_Type", "VEI", "Damage_in_Mill")

subset_data <- subset_data %>%
  mutate(row_number = row_number()) %>%
  select(row_number, everything())
View(subset_data)

# View the first few rows of the subsetted data
View(subset_data)

write.csv(subset_data, file = "subset_data.csv", row.names=FALSE)
```

```
[27] "Total Effects : Injuries"
[28] "Total Effects : Injuries Description"
[29] "Total Effects : Damages in million Dollars"
[30] "Total Effects : Damage Description"
[31] "Total Effects : Houses Destroyed"
[32] "Total Effects : Houses Destroyed Description"
[33] "Coordinates"
[34] "Earthquakes : Houses damaged Description"
[35] "Total Effects : Houses Damaged Description"
> subset_data <- significant_volcanic_eruption_database[, c("Year", "Volcano.Name", "Country", "Elevation", "Volcano.Type", "Volcanic.Explosivity.Index", "Volcano...Damage..in.M..")]
Error in `significant_volcanic_eruption_database[, c("Year", "Volcano.Name",
    "Country", "Elevation", "Volcano.Type", "Volcanic.Explosivity.Index",
    "Volcano...Damage..in.M..")]`:
! Can't subset columns that don't exist.
✖ Columns `Volcano.Name`, `Volcano.Type`, `Volcanic.Explosivity.Index`, and `Volcano...Damage..in.M..` don't exist.
Run `rlang::last_trace()` to see where the error occurred.
> subset_data <- significant_volcanic_eruption_database[, c("Year", "Volcano Name", "Country", "Elevation", "Volcano Type", "Volcanic Explosivity Index", "Volcano : Damage (in M$)")]
> colnames(subset_data) <- c("Incident_Year", "Volcano_Name", "Country", "Elevation", "Volcano_Type", "VEI", "Damage_in_Mill")
> subset_data <- subset_data %>%
+   mutate(row_number = row_number()) %>%
+   select(row_number, everything())
> View(subset_data)
> # View the first few rows of the subsetted data
> View(subset_data)
> write.csv(subset_data, file = "subset_data.csv", row.names=FALSE)
> source("~/.active-rstudio-document")
Rows: 835 Columns: 35
```



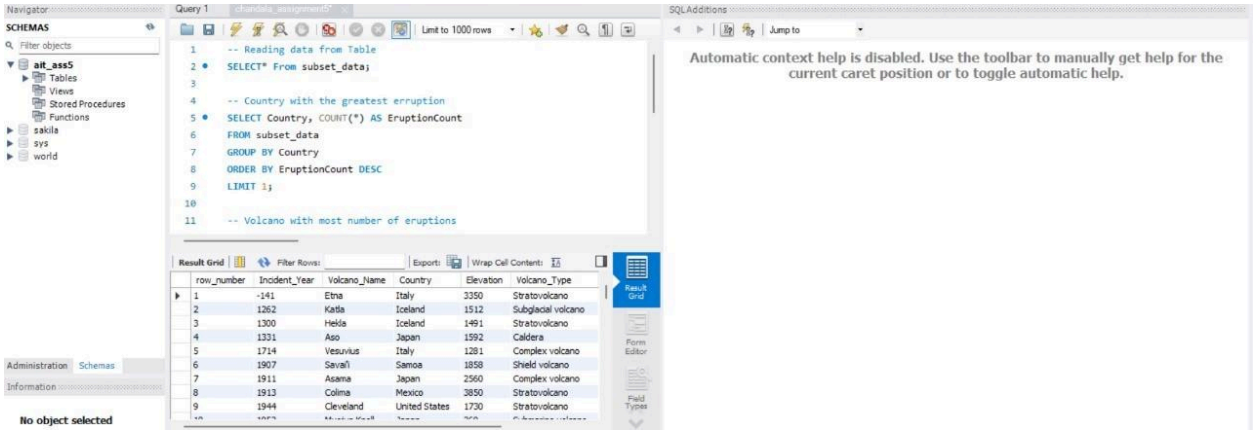
This code snippet demonstrates data manipulation and preparation in R using the `dplyr` and `readr` packages. It begins by loading the necessary libraries with `library(dplyr)` and `library(readr)`. The CSV file "significant-volcanic-eruption-database.csv" is then read using `read_csv`, skipping any lines starting with "#" as comments and also skipping the first row which typically contains headers. The dataset is then viewed using `View(significant_volcanic_eruption_database)` to inspect its contents. Next, a subset of the data is created, selecting specific columns such as "Year", "Volcano Name", "Country", "Elevation", "Volcano Type", "Volcanic Explosivity Index", and "Volcano : Damage (in M$)". The columns are then renamed for easier handling. The subset is further modified by adding a new column `row_number` which assigns a unique row number to each entry. This is accomplished using `mutate` from `dplyr` and reordering the columns to place `row_number` at the beginning with `select`. The resulting subset

data frame is viewed again with `View(subset_data)` for inspection. Finally, the modified subset is written to a new CSV file named "subset_data.csv" in the current working directory using `write.csv(subset_data, file = "subset_data.csv", row.names=FALSE)`.

This entire process showcases a common workflow in data analysis, illustrating the use of `dplyr` and `readr` functions to efficiently manipulate, subset, and prepare data for further analysis or visualization in R.

SQL: Create a SQL database schema and table for the subset dataset using any RDBMS (Oracle, MySQL, etc.)

- **Load the dataset into the table; use an SQL query to display a few records**
- **Query the database and interpret the results, displaying:**
  - **The Country with the greatest number of eruptions**
  - **The Volcano with the greatest number of eruptions**
  - **The Volcano Type with the greatest VEI**
  - **The top 10 Volcano Names, Countries, & Volcano Types with the greatest amount of Damages in M$**



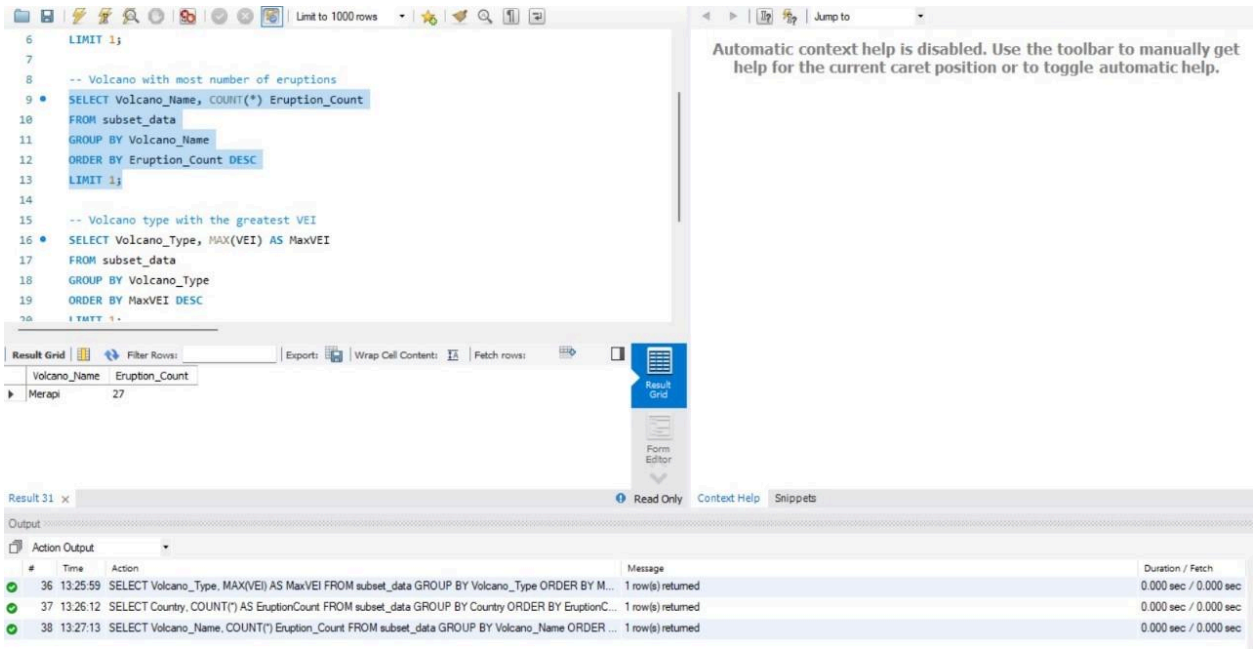
- **The Country with the greatest number of eruptions**

```
3       -- Country with the greatest erruption
4 •     SELECT Country, COUNT(*) AS EruptionCount
5       FROM subset_data
6       GROUP BY Country
7       ORDER BY EruptionCount DESC
8       LIMIT 1;
9
0       -- Volcano with the greatest VEI
1 •     SELECT Volcano_Type, MAX(VEI) AS MaxVEI
```

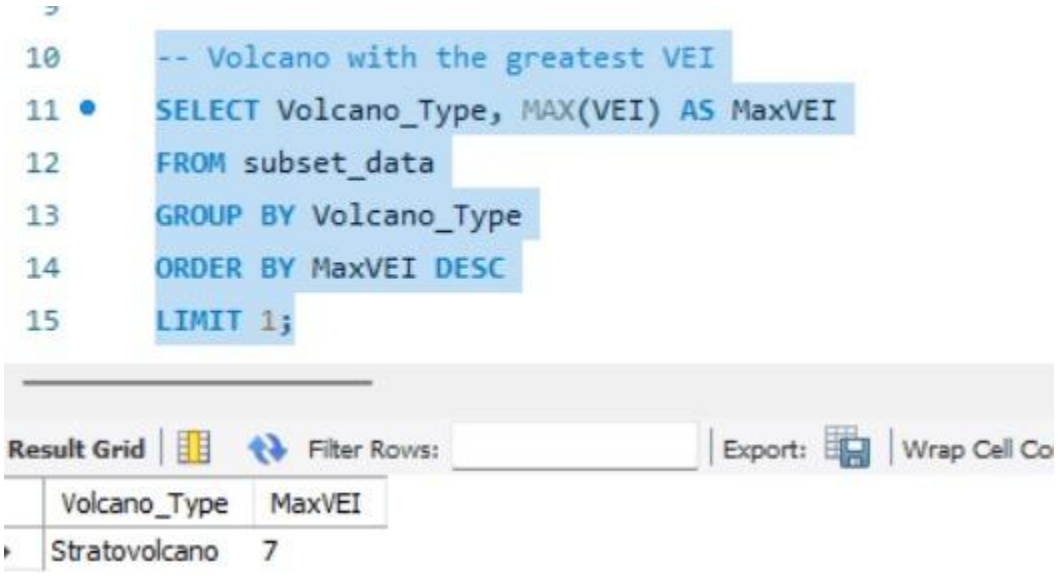
sult Grid | Filter Rows: | Export: | W

Country	EruptionCount
Indonesia	194

- **The Volcano with the greatest number of eruptions**



- **The Volcano Type with the greatest VEI**



- **The top 10 Volcano Names, Countries, & Volcano Types with the greatest amount of Damages in M$**

```
8     LIMIT 1;
9
10    -- Volcano with the greatest VEI
11 •  SELECT Volcano_Type, MAX(VEI) AS MaxVEI
12    FROM subset_data
13    GROUP BY Volcano_Type
14    ORDER BY MaxVEI DESC
15    LIMIT 1;
16
17    -- The top 10 Volcano Names, Countries, and Volcano Types
18 •  SELECT  Volcano_Name, Country, Volcano_Type, Damage_in_Mil
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Volcano_Name	Country	Volcano_Type	Damage_in_Mill
St. Helens	United States	Stratovolcano	2000
Merapi	Indonesia	Stratovolcano	600
Kilauea	United States	Shield volcano	370
Fuego	Guatemala	Stratovolcano	120
Sinabung	Indonesia	Stratovolcano	100
Rabaul	Papua New Guinea	Pyroclastic shield	86
Taal	Philippines	Stratovolcano	67
Kilauea	United States	Shield volcano	15
Galunggung	Indonesia	Stratovolcano	15
Ulawun	Papua New Guinea	Stratovolcano	14