

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**
 - o **Pages 211-252 and 268-272**
- **Use the significant-volcanic-eruption-database.csv dataset (in the Code & Data folder)**
 - o **Source: see Code and Data folder**
 - o **Read the details about the dataset**
 - o **Remove the # comments from the dataset**
 - o **Create a subset of the dataset containing the following fields**
Year, Volcano Name, Country, Elevation, Volcano Type, VEI, Damage in M\$,
 - o **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:**
 - o **The Country with the greatest number of eruptions**
 - o **The Volcano with the greatest number of eruptions**
 - o **The Volcano Type with the greatest VEI**
 - o **The top 10 Volcano Names, Countries, & Volcano Types with the greatest amount of Damages in M\$**

R:

The image shows the RStudio interface with a script editor containing the following code:

```
1 library(dplyr)
2 library(readr)
3 significant_volcanic_eruption_database <- read_csv("Bezzam_Assignment/significant-volcanic-eruption-database.csv",
4                                                    comment = "#", skip = 1)
5 View(significant_volcanic_eruption_database)
6
7 names(significant_volcanic_eruption_database)
8
9 # Create a subset with the specified fields
10
11 subset_data <- significant_volcanic_eruption_database[, c("Year", "Volcano Name", "Country", "Elevation", "Volcano Type", "Volcanic Explosivity Index", "Volcano : Damage (in M$)")]
12 colnames(subset_data) <- c("Incident_Year", "Volcano_Name", "Country", "Elevation", "Volcano_Type", "VEI", "Damage_in_M11")
13
14 subset_data <- subset_data %>%
15   mutate(row_number = row_number()) %>%
16   select(row_number, everything())
17 View(subset_data)
18
19 # View the first few rows of the subsetted data
20 View(subset_data)
21
22 write.csv(subset_data, file = "subset_data.csv", row.names=FALSE)
23
24
```

The right sidebar shows the Environment pane with 'subset_data' and 'significant_volcanic_eruption_database' listed. The Files pane shows a list of files including 'Bezzam_Assignment', 'Rhistory', and various plots.

The image shows the RStudio interface with the same script as above. The console displays the following error message:

```
R 4.3.2 ~~~~~
[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.
X 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_M11")
> 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
```

The right sidebar shows the Environment pane with 'subset_data' and 'significant_volcanic_eruption_database' listed. The Files pane shows a list of files including 'Bezzam_Assignment', 'Rhistory', and various plots.

row_number	Incident Year	Volcano Name	Country	Elevation	Volcano Type	VEI	Damage in Mill
1	1	-141 Etna	Italy	3350	Stratovolcano	N/A	N/A
2	2	1262 Katla	Iceland	1512	Subglacial volcano	3	N/A
3	3	1300 Hekla	Iceland	1491	Stratovolcano	4	N/A
4	4	1331 Aso	Japan	1592	Caldera	2	N/A
5	5	1714 Vesuvius	Italy	1281	Complex volcano	N/A	N/A
6	6	1907 Sava'i	Samoa	1858	Shield volcano	N/A	N/A
7	7	1911 Asama	Japan	2560	Complex volcano	2	N/A
8	8	1913 Colima	Mexico	3850	Stratovolcano	4	N/A
9	9	1944 Cleveland	United States	1730	Stratovolcano	3	N/A
10	10	1952 Myojun Knoll	Japan	360	Submarine volcano	2	N/A
11	11	1960 Puyehue	Chile	2236	Stratovolcano	3	N/A
12	12	1971 La Palma	Spain	2426	Stratovolcano	2	N/A
13	13	1972 Ritter Island	Papua New Guinea	140	Stratovolcano	1	N/A
14	14	1979 Soufriere St. Vincent	St. Vincent & the Grenadines	1220	Stratovolcano	3	N/A
15	15	1983 Miyake-jima	Japan	815	Stratovolcano	3	N/A
16	16	1984 Etna	Italy	3350	Stratovolcano	3	N/A
17	17	1990 Aso	Japan	1592	Caldera	2	N/A
18	18	1991 Etna	Italy	3350	Stratovolcano	2	2,500
19	19	1994 Semeru	Indonesia	3676	Stratovolcano	3	N/A
20	20	2007 Salak	Indonesia	2211	Stratovolcano	N/A	N/A
21	21	2017 Fuego	Guatemala	3763	Stratovolcano	2	N/A
22	22	2017 Aoba	Vanuatu	1496	Shield volcano	N/A	N/A
23	23	2018 Aoba	Vanuatu	1496	Shield volcano	3	N/A
24	24	2019 Stromboli	Italy	926	Stratovolcano	2	N/A
25	25	1593 Raung	Indonesia	3332	Stratovolcano	5	N/A
26	26	1609 Tengchong	China	2865	Pyroclastic cone	N/A	N/A
27	27	1692 Serua	Pacific Ocean	641	Stratovolcano	4	N/A
28	28	1718 Pico	Portugal	2351	Stratovolcano	2	N/A
29	29	1846 Taupo	New Zealand	760	Caldera	N/A	N/A
30	30	1866 Santorini	Greece	329	Shield volcano	2	N/A
31	31	1877 Mauna Loa	United States	4170	Shield volcano	0	N/A
32	32	1883 Bagana	Papua New Guinea	1750	Lava cone	3	N/A
33	33	1886 Niuafo'ou	Tonga	260	Shield volcano	4	N/A
34	34	1896 Kirishima	Japan	1700	Shield volcano	2	N/A
35	35	1902 Pelée	Martinique	1197	Stratovolcano	4	N/A

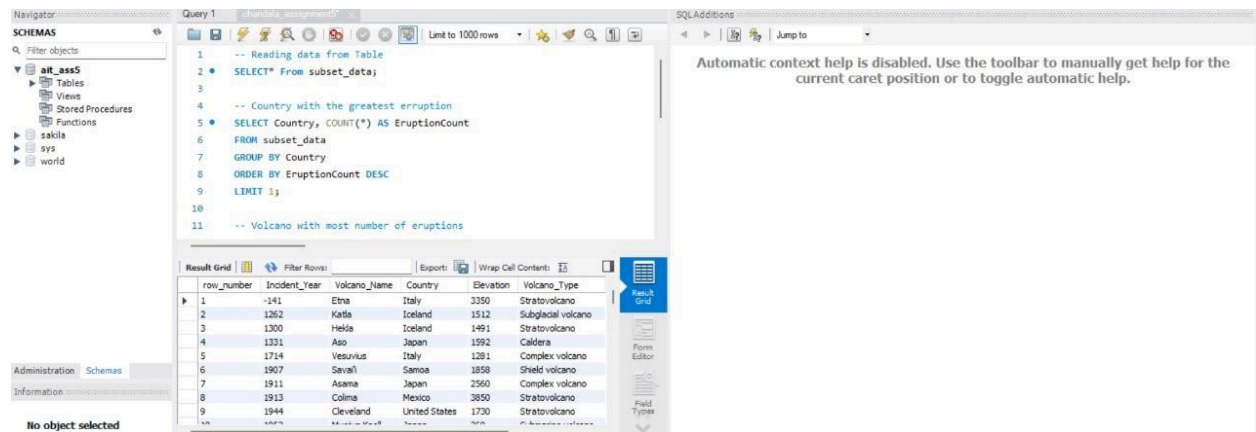
This code snippet demonstrates data manipulation and preparation in R using the `dplyr` and `readr` packages. It begins by loading the necessary libraries with `library(dplyr)` and `library(readr)`. The CSV file "significant-volcanic-eruption-database.csv" is then read using `read_csv`, skipping any 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:**
 - o **The Country with the greatest number of eruptions**
 - o **The Volcano with the greatest number of eruptions**
 - o **The Volcano Type with the greatest VEI**
 - o **The top 10 Volcano Names, Countries, & Volcano Types with the greatest amount of Damages in M\$**



o The Country with the greatest number of eruptions

```

3  -- Country with the greatest eruption
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

```

Result Grid	Filter Rows:	Export:
Country	EruptionCount	
Indonesia	194	

o The Volcano with the greatest number of eruptions

The screenshot shows a SQL IDE with a query editor and a results pane. The query editor contains the following SQL code:

```
6 LIMIT 1;
7
8 -- Volcano with most number of eruptions
9 • SELECT Volcano_Name, COUNT(*) Eruption_Count
10 FROM subset_data
11 GROUP BY Volcano_Name
12 ORDER BY Eruption_Count DESC
13 LIMIT 1;
14
15 -- Volcano type with the greatest VEI
16 • SELECT Volcano_Type, MAX(VEI) AS MaxVEI
17 FROM subset_data
18 GROUP BY Volcano_Type
19 ORDER BY MaxVEI DESC
20 LIMIT 1;
```

The results pane shows a table with two columns: Volcano_Name and Eruption_Count. The first row is Merapi with 27 eruptions.

Volcano_Name	Eruption_Count
Merapi	27

The bottom pane shows the Action Output with the following log:

#	Time	Action	Message	Duration / Fetch
36	13:25:59	SELECT Volcano_Type, MAX(VEI) AS MaxVEI FROM subset_data GROUP BY Volcano_Type ORDER BY M...	1 row(s) returned	0.000 sec / 0.000 sec
37	13:26:12	SELECT Country, COUNT(*) AS EruptionCount FROM subset_data GROUP BY Country ORDER BY EruptionC...	1 row(s) returned	0.000 sec / 0.000 sec
38	13:27:13	SELECT Volcano_Name, COUNT(*) Eruption_Count FROM subset_data GROUP BY Volcano_Name ORDER ...	1 row(s) returned	0.000 sec / 0.000 sec

o The Volcano Type with the greatest VEI

The screenshot shows a SQL IDE with a query editor and a results pane. The query editor contains the following SQL code:

```
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;
```

The results pane shows a table with two columns: Volcano_Type and MaxVEI. The first row is Stratovolcano with a MaxVEI of 7.

Volcano_Type	MaxVEI
Stratovolcano	7

- o 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_Mil
▶	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