**VIETNAM NATIONAL UNIVERSITY HCMC**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

****

**Subject: Decision support and business intelligence applications**

**Topic: Analysis the Global Historical Tsunami Database**

**Lecturer:** Do Phuc

**Lecturer:** Nguyen Thi Kim Phung

**Group:** 2

**Members:**

* Pham Thuy Dung - 20521214
* Dau Dinh Quang Anh - [20521059](mailto:%3205%3210%359@%67%6d%2e%75it.e%64%75%2e%76%6e)

**COMMENT**

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

…………………………………………………………………………………………………………………………

…………………………………………………………………………………………………………………………

**Table of Contents**

[1 Topic Overview 5](#_Toc168647662)

[2 Dictionary of Data 6](#_Toc168647663)

[3 Data Warehouse Design 10](#_Toc168647664)

[4 Buid Data Warehouse (SSIS) 11](#_Toc168647665)

[5 Data Warehouse Analysis (SSAS) 14](#_Toc168647666)

[5.1 Create Data Warehouse 16](#_Toc168647667)

[5.2 Create Data Source View 17](#_Toc168647668)

[5.3 Create Data Cube 18](#_Toc168647669)

[5.4 Edit the Dimension 19](#_Toc168647670)

[5.4.1 DimDate 19](#_Toc168647671)

[5.4.2 DimTime 20](#_Toc168647672)

[5.4.3 DimLocation 21](#_Toc168647673)

[5.4.4 DimDescription 22](#_Toc168647674)

[5.4.5 DimDetailDamage 23](#_Toc168647675)

[5.4.6 DimTotalDamage 24](#_Toc168647676)

[5.4.7 Run the Process for Deployment 25](#_Toc168647677)

[6 Query MDX 26](#_Toc168647678)

[6.1 Query Tsunami year and description 26](#_Toc168647679)

[6.1.1 SQL Query 27](#_Toc168647680)

[6.1.2 Get SQL result 27](#_Toc168647681)

[6.2 Query Number of Deaths and their description 27](#_Toc168647682)

[6.2.1 SQL Query 27](#_Toc168647683)

[6.2.2 Get SQL Result 28](#_Toc168647684)

[6.3 Query Tsunami intensity over the year 28](#_Toc168647685)

[6.3.1 SQL query 28](#_Toc168647686)

[6.3.2 Get SQL Result 29](#_Toc168647687)

[6.4 Query total of deaths over the year with country 29](#_Toc168647688)

[6.4.1 SQL Query 29](#_Toc168647689)

[6.4.2 Get SQL Result 29](#_Toc168647690)

[6.5 Query number of houses damaged and their descriptions 30](#_Toc168647691)

[6.5.1 SQL Query 30](#_Toc168647692)

[6.5.2 Get SQL Result 30](#_Toc168647693)

[6.6 Query for the total number of tsunamis by time of day with descriptions 31](#_Toc168647694)

[6.6.1 SQL Query 31](#_Toc168647695)

[6.6.2 Get SQL Result 31](#_Toc168647696)

[6.7 Query Countries and location has tsunami over the months 32](#_Toc168647697)

[6.7.1 SQL Query 32](#_Toc168647698)

[6.7.2 Get SQL Result 32](#_Toc168647699)

[6.8 Top 10 Countries by Event Count in 2020 33](#_Toc168647700)

[6.8.1 SQL Query 33](#_Toc168647701)

[6.8.2 Get SQL Result 33](#_Toc168647702)

[6.9 Top 3 Years by Event Count 34](#_Toc168647703)

[6.9.1 SQL Query 34](#_Toc168647704)

[6.9.2 Get SQL Result 34](#_Toc168647705)

[6.10 Query for the total damage ($) by country and year 34](#_Toc168647706)

[6.10.1 SQL Query 34](#_Toc168647707)

[6.10.2 Get SQL Result 35](#_Toc168647708)

[7 Report Data (by using PowerBI) 35](#_Toc168647709)

[7.1 Sum of Deaths by Year 35](#_Toc168647710)

[7.2 Sum of Tsunami Intensity by Year 36](#_Toc168647711)

[7.3 Count of Total Houses Damaged by Year 36](#_Toc168647712)

[7.4 Sum of Maximum Water Height (m) by Year 37](#_Toc168647713)

[7.5 Sum of Vol by Year 37](#_Toc168647714)

[7.6 Sum of Earthquake Magnitude by Year 38](#_Toc168647715)

[7.7 Sum of Number of Runups and Sum of Earthquake Magnitude by Year 38](#_Toc168647716)

[7.8 Count of Total Damage ($Mil) by Country 39](#_Toc168647717)

[7.9 Sum of Tsunami Intensity in Month by Country 39](#_Toc168647718)

[7.10 Count of Tsunami Days in Months 40](#_Toc168647719)

[8 Pivot Table 40](#_Toc168647720)

[8.1 Dataset 40](#_Toc168647721)

[8.2 Statistics on the number of deaths and causes over the years 41](#_Toc168647722)

[8.2.1 Top 10 Number of Tsunami Event Validity per Country 41](#_Toc168647723)

[8.2.2 Number of Earthquake magnitude over the year 41](#_Toc168647724)

[8.2.3 Number of Tsunami Intensity over the month 42](#_Toc168647725)

[8.2.4 Top 10 of Country has Damage ($Mil) 43](#_Toc168647726)

[8.2.5 Sum of Total House damage description 43](#_Toc168647727)

[8.2.6 Count of Country per Tsunami Cause 43](#_Toc168647728)

[8.2.7 Top 10 of Country has Volcano 44](#_Toc168647729)

[8.2.8 Top 10 of Country has total runnup highest 45](#_Toc168647730)

[8.2.9 Top 10 of Year has number of deaths highest 45](#_Toc168647731)

[8.2.10 Top 10 of Country has Maximum Water Height 46](#_Toc168647732)

[9 Data Mining 46](#_Toc168647733)

[9.1 Load Dataset 46](#_Toc168647734)

[9.2 Visualization of Dataset 47](#_Toc168647735)

[9.2.1 Visualization Number of Tsunami per Country 48](#_Toc168647736)

[9.2.2 Visualization number of Tsunami per Cause 48](#_Toc168647737)

[9.2.3 Number of Tsunamis per Month (Sorted from High to Low) 49](#_Toc168647738)

[9.2.4 Number of Earthquakes by Magnitude Range 50](#_Toc168647739)

[9.2.5 Number of Tsunamis by Maximum Water Height (m) Range 50](#_Toc168647740)

[9.2.6 Number of Tsunamis by Number of Runups Range 51](#_Toc168647741)

[9.2.7 Number of Tsunamis by Event Validity 52](#_Toc168647742)

[9.2.8 Number of Tsunamis by Total Death Description 52](#_Toc168647743)

[9.2.9 Number of Tsunamis by Total Damage Description 53](#_Toc168647744)

[9.3 Clustering Dataset by using some algorithms 53](#_Toc168647745)

[9.3.1 Clustering by using K-means 53](#_Toc168647746)

[9.3.2 Clustering by using Agglomerative 55](#_Toc168647747)

[9.3.3 Clustering by using Birch 55](#_Toc168647748)

[9.3.4 Comparison between alogirthm 56](#_Toc168647749)

[10 Conclusion 56](#_Toc168647750)

# Topic Overview

This report conducts research on the causes, consequences, and damage of tsunamis based on the Global Historical Tsunami Database.

This dataset represents historical tsunamis and related information, providing data on over 2,400 tsunamis from 2100 BC to the present in the Atlantic, Indian, and Pacific Oceans, as well as the Mediterranean and Caribbean Seas. This dataset download from [NCEI Global Historical Hazard Database (noaa.gov)](https://www.ngdc.noaa.gov/hazel/view/hazards/tsunami/event-data)

# Dictionary of Data

|  |  |
| --- | --- |
| Attribute | Description |
| Year | The year in which the event occurred. |
| Mo (Month) | The month in which the event occurred. |
| Dy (Day) | The day on which the event occurred. |
| Hr (Hour) | The hour at which the event occurred. |
| Mn (Minute) | The minute at which the event occurred. |
| Sec (Second) | The second at which the event occurred. |
| Tsunami Event Validity | Valid values: -1 to 4  Validity of the actual tsunami occurrence is indicated by a numerical rating of the reports of that event:  -1: erroneous entry  0: event that only caused a seiche or disturbance in an inland river/lake  1: very doubtful tsunami  2: questionable tsunami  3: probable tsunami  4: definite tsunami |
| Tsunami Cause Code | Valid values: 0 to 11  The source of the tsunami:  0: Unknown  1: Earthquake  2: Questionable Earthquake  3: Earthquake and Landslide  4: Volcano and Earthquake  5: Volcano, Earthquake, and Landslide  6: Volcano  7: Volcano and Landslide  8: Landslide  9: Meteorological  10: Explosion  11: Astronomical Tide |
| Earthquake Magnitude | Valid values: 0.0 to 9.9  The value in this column contains the primary earthquake magnitude. Magnitude measures the energy released at the source of the earthquake. Magnitude is determined from measurements on seismographs. For pre-instrumental events, the magnitudes are derived from intensities. There are several different scales for measuring earthquake magnitudes. The primary magnitude is chosen from the available magnitude scales in this order:   * Mw Magnitude * Ms Magnitude * Mb Magnitude * Ml Magnitude * Mfa Magnitude * Unknown Magnitude |
| Vol | Volcano identification number if a volcanic eruption caused the tsunami. |
| More Info | Additional information or notes about the event. |
| Deposits | Information about deposits left by the tsunami. |
| Country | The Country where the tsunami source occurred (For example enter: Japan or Russia). |
| Location Name | The Country, State, Province or Island where the tsunami source occurred (For example enter: Japan or Honshu |
| Latitude (kinh độ) | Valid values: **-90 to 90**  Latitude: 0 to 90 (Northern Hemisphere) -90 to 0 (Southern Hemisphere) |
| Longitude (vĩ độ) | Valid values: **-180 to 180**  Longitude: 0 to 180 (Eastern Hemisphere) -180 to 0 (Western Hemisphere) |
| Maximum Water Height (m) | Maximum Water Height (m)   * The maximum water height above sea level in meters for this event. If the type of measurement of the runup was a: * Tide Gauge - half of the maximum height (minus the normal tide) of a tsunami wave recorded at the coast by a tide gauge. * Runup Height - the maximum elevation the wave reaches at the maximum inundation. |
| Number of Runups | The total number of runups link will display the runup locations associated with a particular tsunami event. |
| Tsunami Magnitude (Abe) | Tsunami magnitude scale by Abe. |
| Tsunami Magnitude (Iida) | Tsunami magnitude scale by Iida. |
| Tsunami Intensity | Intensity of the tsunami. |
| Deaths | Whenever possible, numbers of deaths are listed. |
| Death Description | Valid values: 0 to 4  When a description was found in the historical literature instead of an actual number of deaths, this value was coded and listed in the Deaths column. If the actual number of deaths was listed, a descriptor was also added for search purposes.  0: None  1: Few (~1 to 50 deaths)  2: Some (~51 to 100 deaths)  3: Many (~101 to 1000 deaths)  4: Very many (over 1000 deaths) |
| Missing | Number of people missing after the tsunami. |
| Missing Description | Description of missing people. |
| Injuries | Whenever possible, numbers of injuries from the tsunami are listed. |
| Injuries Description | Valid values: 0 to 4  When a description was found in the historical literature instead of an actual number of injuries, this value was coded and listed in the Injuries column. If the actual number of injuries was listed, a descriptor was also added for search purposes.  0: None  1: Few (~1 to 50 injuries)  2: Some(~51 to 100 injuries)  3: Many (~101 to 1000 injuries)  4: Very many (over 1000 injuries) |
| Damage ($Mil) | The value in the Damage column should be multiplied by 1,000,000 to obtain the actual dollar amount. |
| Damage Description | Valid values: 0 to 4  For those events not offering a monetary evaluation of damage, the following five-level scale was used to classify damage (1990 dollars) and was listed in the Damage column. If the actual dollar amount of damage was listed, a descriptor was also added for search purposes.  0: NONE  1: LIMITED (roughly corresponding to less than $1 million)  2: MODERATE (~$1 to $5 million)  3: SEVERE (~$5 to $25 million)  4: EXTREME (~$25 million or more) |
| Houses Destroyed | Whenever possible, numbers of houses destroyed are listed. |
| Houses Destroyed Description | Valid values: 0 to 4  For those events not offering an exact number of houses damaged, the following four-level scale was used to classify the damage and was listed in the Houses Destroyed column. If the actual number of houses destroyed was listed, a descriptor was also added for search purposes.  0: None  1: Few (~1 to 50 houses)  2: Some (~51 to 100 houses)  3: Many (101 to 1000 houses)  4: Very Many (~over 1000 houses) |
| Houses Damaged | Number of houses damaged. |
| Houses Damaged Description | Description of the damage to houses. |
| Total Deaths | Whenever possible, total number of deaths from the tsunami and secondary effects are listed. |
| Total Death Description | Valid values: 0 to 4  When a description was found in the historical literature instead of an actual number of deaths, this value was coded and listed in the Deaths column. If the actual number of deaths was listed, a descriptor was also added for search purposes.  0: None  1: Few (~1 to 50 deaths)  2: Some (~51 to 100 deaths)  3: Many (~101 to 1000 deaths)  4: Very many (over 1000 deaths) |
| Total Missing | Total number of missing including secondary effects. |
| Total Missing Description | Description of total missing. |
| Total Injuries | Total number of injuries including secondary effects. |
| Total Injuries Description | Description of total injuries. |
| Total Damage ($Mil) | Total Damage in Millions of Dollars from the Tsunami and the Source Event |
| Total Damage Description | Valid values: 0 to 4  For those events not offering a monetary evaluation of damage, the following five-level scale was used to classify damage (1990 dollars) and was listed in the Damage column. If the actual dollar amount of damage was listed, a descriptor was also added for search purposes.  0: NONE  1: LIMITED (roughly corresponding to less than $1 million)  2: MODERATE (~$1 to $5 million)  3: SEVERE (~$5 to $25 million)  4: EXTREME (~$25 million or more) |
| Total Houses Destroyed | Total number of houses destroyed. |
| Total Houses Destroyed Description | Description of total houses destroyed. |
| Total Houses Damaged | Total number of houses damaged. |
| Total Houses Damaged Description | Description of total houses damaged. |

# Data Warehouse Design

A computer screen shot of a computer

Description automatically generated

# Buid Data Warehouse (SSIS)

A screenshot of a computer

Description automatically generated

A diagram of a task

Description automatically generated

A screenshot of a computer

Description automatically generated

A computer screen shot of a diagram

Description automatically generated

The source data used for analysis is a csv file, so we use flat file sources to integrate the data, use sort to remove and order the lines.

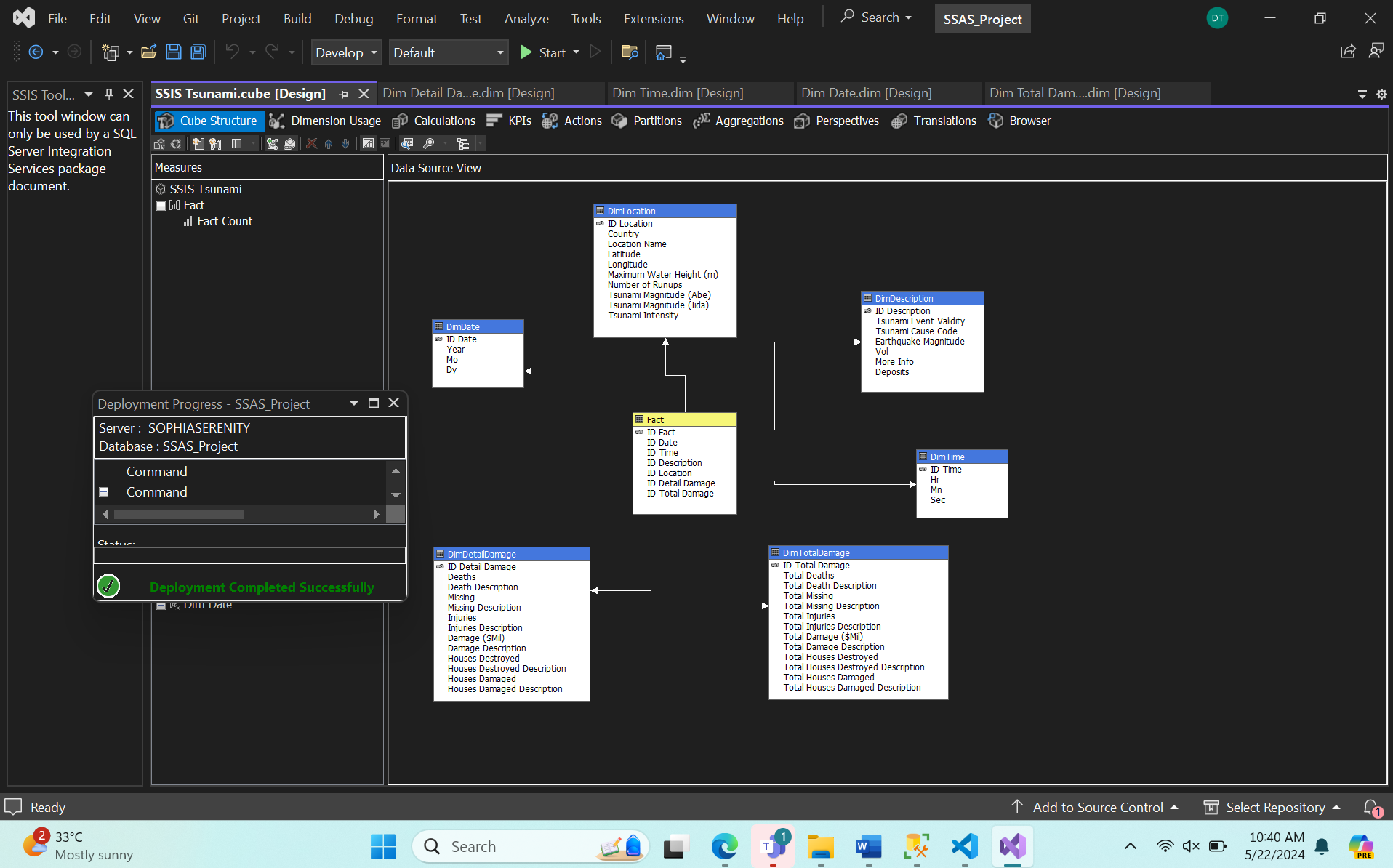
# Data Warehouse Analysis (SSAS)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

## Create Data Warehouse

A screenshot of a computer

Description automatically generated

## Create Data Source View

A screenshot of a computer screen

Description automatically generated

## Create Data Cube

A computer screen shot of a computer

Description automatically generated

## Edit the Dimension

### DimDate

A screenshot of a computer program

Description automatically generated

### DimTime

A screenshot of a computer program

Description automatically generated

### DimLocation

A screenshot of a computer

Description automatically generated

### DimDescription

A screenshot of a computer

Description automatically generated

### DimDetailDamage

A screenshot of a computer

Description automatically generated

### DimTotalDamage

A screenshot of a computer

Description automatically generated

### Run the Process for Deployment

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Query MDX

## Query Tsunami year and description

### SQL Query

A screenshot of a computer program

Description automatically generated

### Get SQL result

A white screen with a black border

Description automatically generated

## Query Number of Deaths and their description

### SQL Query

A screenshot of a computer

Description automatically generated

### Get SQL Result

A white screen with a black border

Description automatically generated

## Query Tsunami intensity over the year

### SQL query

A screenshot of a computer

Description automatically generated

### Get SQL Result

A screenshot of a computer

Description automatically generated

## Query total of deaths over the year with country

### SQL Query

A screenshot of a computer

Description automatically generated

### Get SQL Result

A screenshot of a computer

Description automatically generated

## Query number of houses damaged and their descriptions

### SQL Query

A computer screen shot of a computer code

Description automatically generated

### Get SQL Result

A white screen with a black border

Description automatically generated

## Query for the total number of tsunamis by time of day with descriptions

### SQL Query

A screenshot of a computer program

Description automatically generated

### Get SQL Result

A white screen with a black border

Description automatically generated

## Query Countries and location has tsunami over the months

### SQL Query

A computer screen shot of a computer code

Description automatically generated

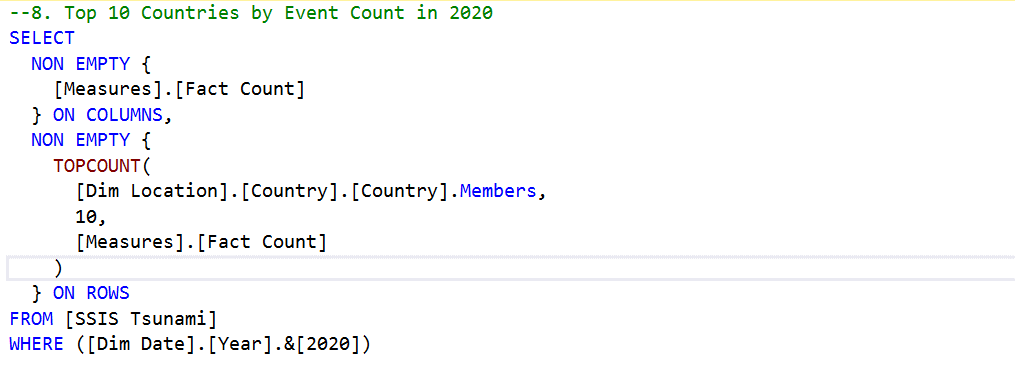
### Get SQL Result

A screenshot of a computer

Description automatically generated

## Top 10 Countries by Event Count in 2020

### SQL Query

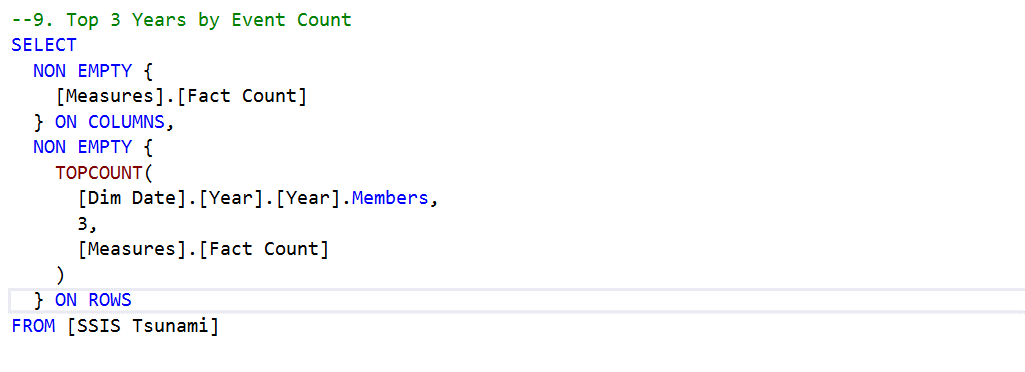


### Get SQL Result



## Top 3 Years by Event Count

### SQL Query

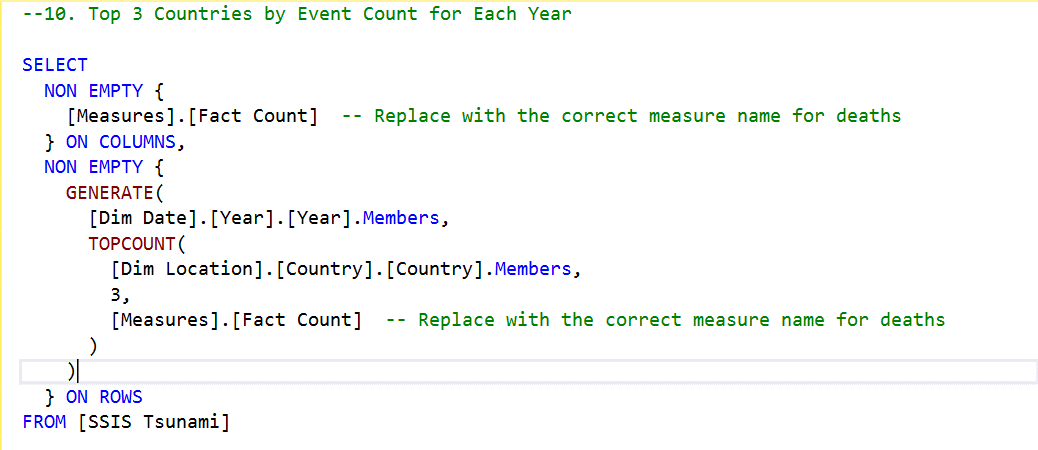


### Get SQL Result

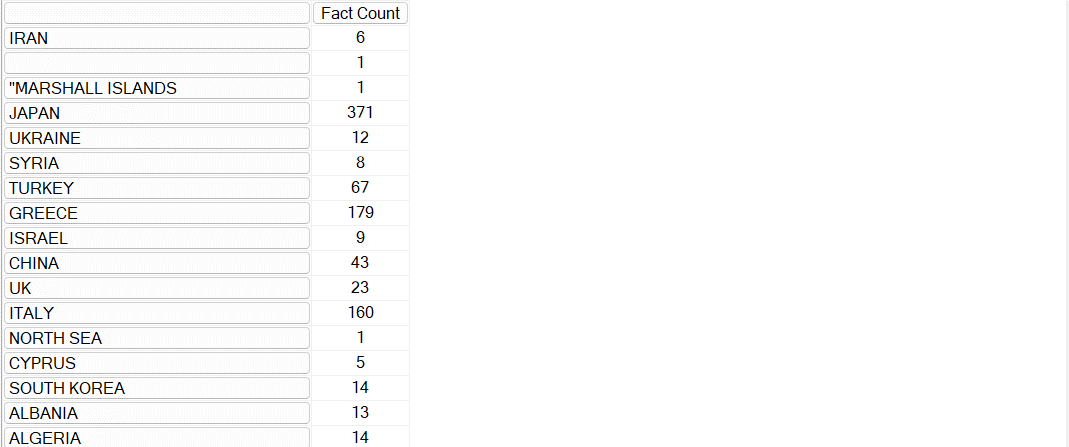


## Query for the total damage ($) by country and year

### SQL Query



### Get SQL Result



# Report Data (by using PowerBI)

## Sum of Deaths by Year

A graph showing a line of a graph

Description automatically generated with medium confidence

## Sum of Tsunami Intensity by Year

A graph showing a line of blue lines

Description automatically generated with medium confidence

## Count of Total Houses Damaged by Year

A graph showing a graph

Description automatically generated

## Sum of Maximum Water Height (m) by Year

A graph showing the growth of the year

Description automatically generated

## Sum of Vol by Year

A graph showing a line

Description automatically generated

## Sum of Earthquake Magnitude by Year

A graph showing a graph

Description automatically generated

## Sum of Number of Runups and Sum of Earthquake Magnitude by Year

A white graph with black text

Description automatically generated

## Count of Total Damage ($Mil) by Country

A screen shot of a graph

Description automatically generated

## Sum of Tsunami Intensity in Month by Country

A screenshot of a computer

Description automatically generated

## Count of Tsunami Days in Months

A grid of numbers on a white background

Description automatically generated

# Pivot Table

## Dataset

A screen shot of a computer

Description automatically generated

## Statistics on the number of deaths and causes over the years

### Top 10 Number of Tsunami Event Validity per Country

A screenshot of a graph

Description automatically generated

### Number of Earthquake magnitude over the year

A screenshot of a graph

Description automatically generated

### Number of Tsunami Intensity over the month

A graph on a computer screen

Description automatically generated

### Top 10 of Country has Damage ($Mil)

A screenshot of a graph

Description automatically generated

### Sum of Total House damage description

A screenshot of a computer

Description automatically generated

### Count of Country per Tsunami Cause

A screenshot of a graph

Description automatically generated

### Top 10 of Country has Volcano

A screenshot of a graph

Description automatically generated

### Top 10 of Country has total runnup highest

A screenshot of a graph

Description automatically generated

### Top 10 of Year has number of deaths highest

A graph on a graph

Description automatically generated

### Top 10 of Country has Maximum Water Height

A screenshot of a computer screen

Description automatically generated

# Data Mining

## Load Dataset

A screenshot of a computer

Description automatically generated

## Visualization of Dataset

### Visualization Number of Tsunami per Country

A graph with a number of countries/regions

Description automatically generated

### Visualization number of Tsunami per Cause

A graph with different colored bars

Description automatically generated

### Number of Tsunamis per Month (Sorted from High to Low)

A graph of a number of tsunamis

Description automatically generated

### Number of Earthquakes by Magnitude Range

A graph of a number of earthquake

Description automatically generated

### Number of Tsunamis by Maximum Water Height (m) Range

A graph of water levels

Description automatically generated with medium confidence

### Number of Tsunamis by Number of Runups Range

A graph of a bar graph

Description automatically generated

### Number of Tsunamis by Event Validity

A graph with different colored squares

Description automatically generated

### Number of Tsunamis by Total Death Description

A graph with different colored squares

Description automatically generated

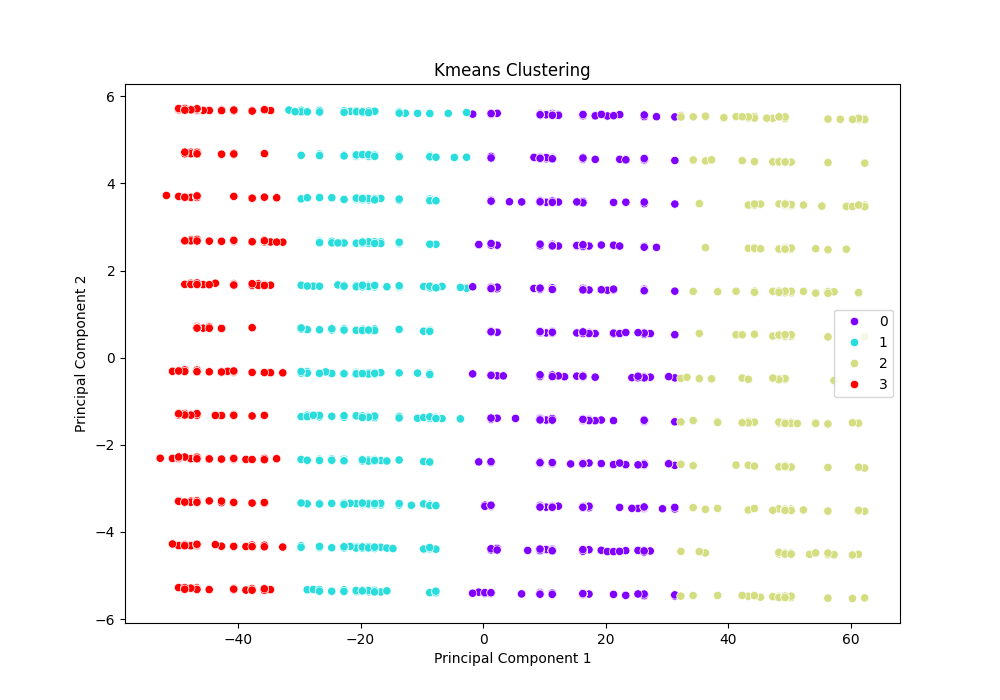
### Number of Tsunamis by Total Damage Description

A graph with different colored squares

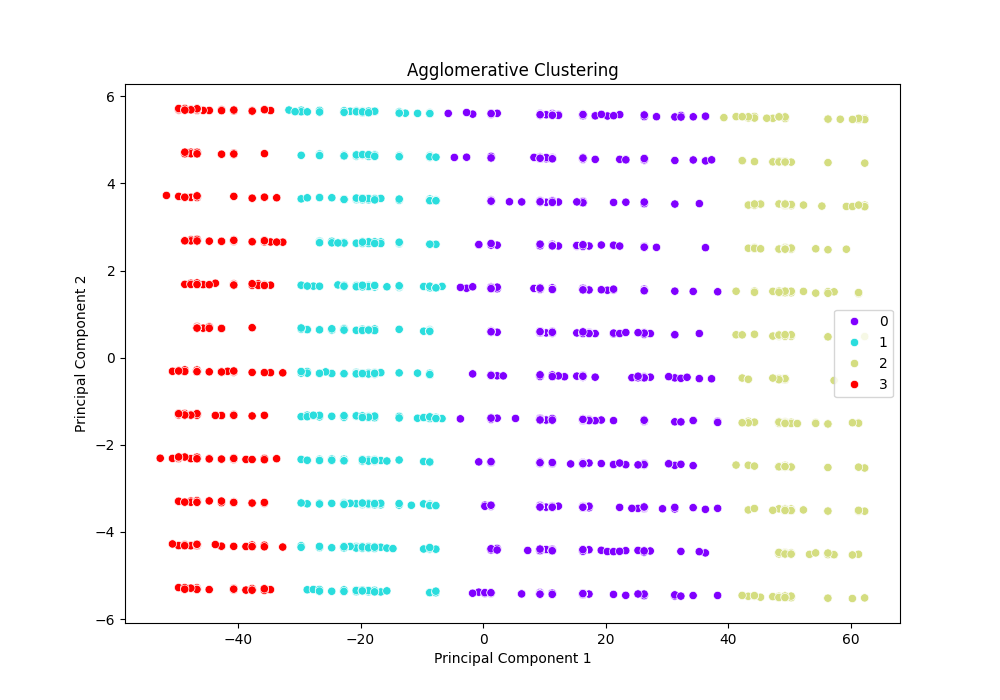
Description automatically generated with medium confidence

## Clustering Dataset by using some algorithms

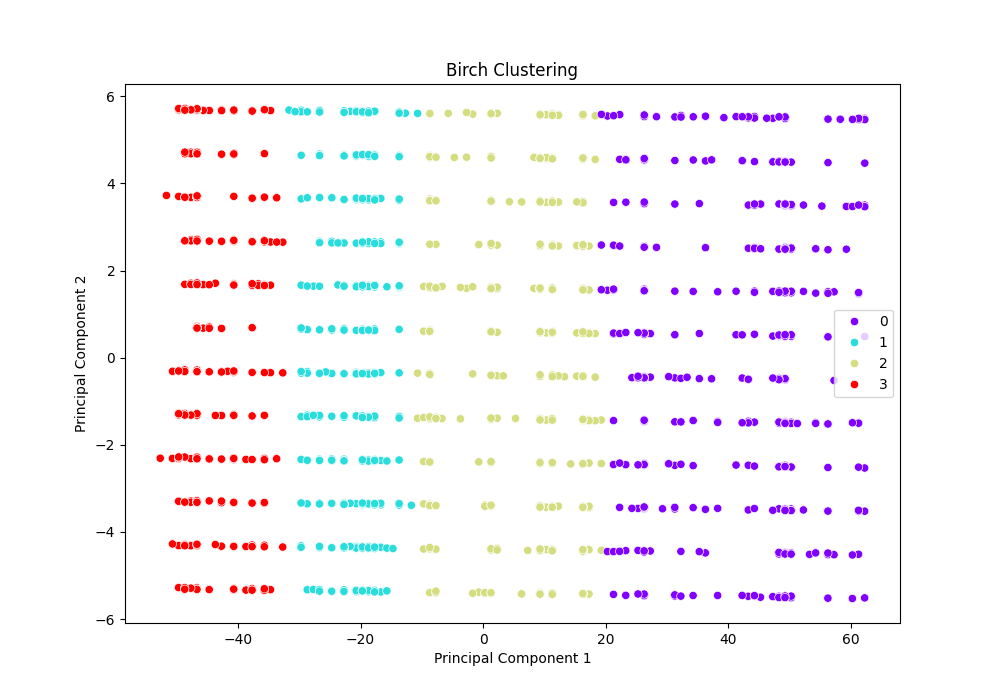
### Clustering by using K-means



### Clustering by using Agglomerative



### Clustering by using Birch



### Comparison between alogirthm

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Silhouette Score | Davies-Bouldin Index | Calinski-Harabasz Index |
| KMeans | 0.652535 | 0.420917 | 14819.58 |
| Agglomerative | 0.64618 | 0.417298 | 13931.45 |
| Birch | 0.582364 | 0.512778 | 10347.99 |

**Overall Conclusion:**

* Kmeans performs the best in terms of the Silhouette Score and Calinski-Harabasz Index, suggesting it forms the most well-defined and separated clusters.
* Agglomerative clustering performs the best in terms of the Davies-Bouldin Index, indicating the most optimal clustering with respect to intra-cluster similarity.
* Birch consistently shows lower performance across all three metrics compared to kmeans and agglomerative clustering.

# Conclusion

After analysis, we have some conclusions. Firstly, the main cause of tsunamis is earthquakes, the year with the highest number of deaths by tsunami is the 2000s. Secondly, Japan has the highest number of Tsunami. Moreover, Most of Tsunami has Earthquake magnitude range in 7-8. Then, Tsunamis occur most frequently in August each year. Lastly, in data mining, the data used for analysis gives quite good results, highlighting visualization of data on the causes, deaths, and damages of Tsunamis. In Comparison, Kmeans appears to be the most balanced and effective clustering method overall, with agglomerative clustering also being a strong contender, especially in terms of intra-cluster similarity. Birch is less effective in this comparison.