# A Data Analysis of Tsunami Occurrences and Impacts.

By Chaitanya Sharma

# Summary

In this data analysis project, I will be investigating tsunamis around the world. My main focus will be on identifying the highest tsunami waves, the costliest tsunamis, the total number of deaths per country, and the number of tsunami occurrences globally. In the data collection phase, we will need to gather data related to tsunamis, including information on wave height, economic costs, number of deaths, and location of occurrences. This data will then be processed, cleaned and organized to allow for a comprehensive analysis. The findings of this analysis will be used to gain insights into the impact of tsunamis, the areas that are most susceptible, and the measures that can be taken to mitigate future disasters.

# 1.- Ask phase

The task of this analysis, is to answer the following questions:

Where are the areas where tsunamis are more frequent?

What are the tallest tsunami waves recorded since the 1900s?

Which tsunamis caused the biggest destructions and what were the economic costs?

What is the death toll of the tsunamis per country?

# 2.- Prepare phase

## Dataset used.

The dataset used in this analysis is publicly available from NOAA (National Oceanic and Atmospheric Administration) for anyone to use. The dataset is accessed throughGoogle's BigQuery.

## Information about the dataset.

The Global Historical Tsunami Database provides information on over 2,400 tsunamis from 200 BC to the present around the globe. The dataset includes two related tables.

The tables include information on the tsunami source and observations of the tsunamis such as the time, date, location, tsunami magnitude and intensity, maximum water height, number of fatalities, total damage (in U.S dollars), and other information.

## Data organization.

The dataset has two tables, historical_source_event which includes information on the tsunami source. The second table, historical_runups, contains information on the observations of the tsunamis, or "runups".

Both tables are related with an ID key which is used to join them.

# 3.- Process phase

## Data Querying with SQL

First, I began by creating a temporary table.

There are thousands of entries about the same tsunami occurrences, but they all duplicate core information. That's why I created a partition for every occurrence which happened the same day, month, year, and country, so that no tsunami record is repeated.

```
WITH tsunami AS (

SELECT ROW_NUMBER () OVER(PARTITION BY CONCAT (event.country, event.year, event.month, event.day)

ORDER BY CONCAT (event.country, event.year, event.month, event.day)) AS unique_id,

run.id AS id1,

run.tsevent_id AS id2,

run.day AS day,

run.year AS year,

run.month AS month,

event.maximum_water_height AS height,

run.longitude AS longitude,

run.latitude AS latitude,

event.event_validity,

event.country AS country,

CASE

WHEN event.deaths IS NULL THEN 0 ELSE event.deaths

END AS deaths_event,

CASE

WHEN run.deaths IS NULL THEN 0 ELSE run.deaths

END AS deaths_run,

CASE

WHEN event.total_damage_in_millions_dollars IS NULL THEN 0 ELSE event.damage_millions_dollars

END AS total_damage_million_dollars
```

```
FROM
bigquery-public-data.noaa_tsunami.historical_runups run
JOIN bigquery-public-data.noaa_tsunami.historical_source_event event
ON run.tsevent_id = event.id)

SELECT
unique_id,
id1 AS id,
year,
month,
day,
height,
longitude,
latitude,
country,
deaths_event + deaths_run AS total_deaths,
total_damage_million_dollars
FROM tsunami
WHERE unique_id = 1  #only the first entry per tsunami will be retrieved
AND event_validity = 4
AND year >= 1900
ORDER BY id1, year ASC, month ASC, day ASC
```

This SQL query will give me the information necessary for our analysis.

## Data cleaning.

After performing the last query on BigQuery, I downloaded the data as a CSV file for further cleaning. I performed the following operations using MS Excel:

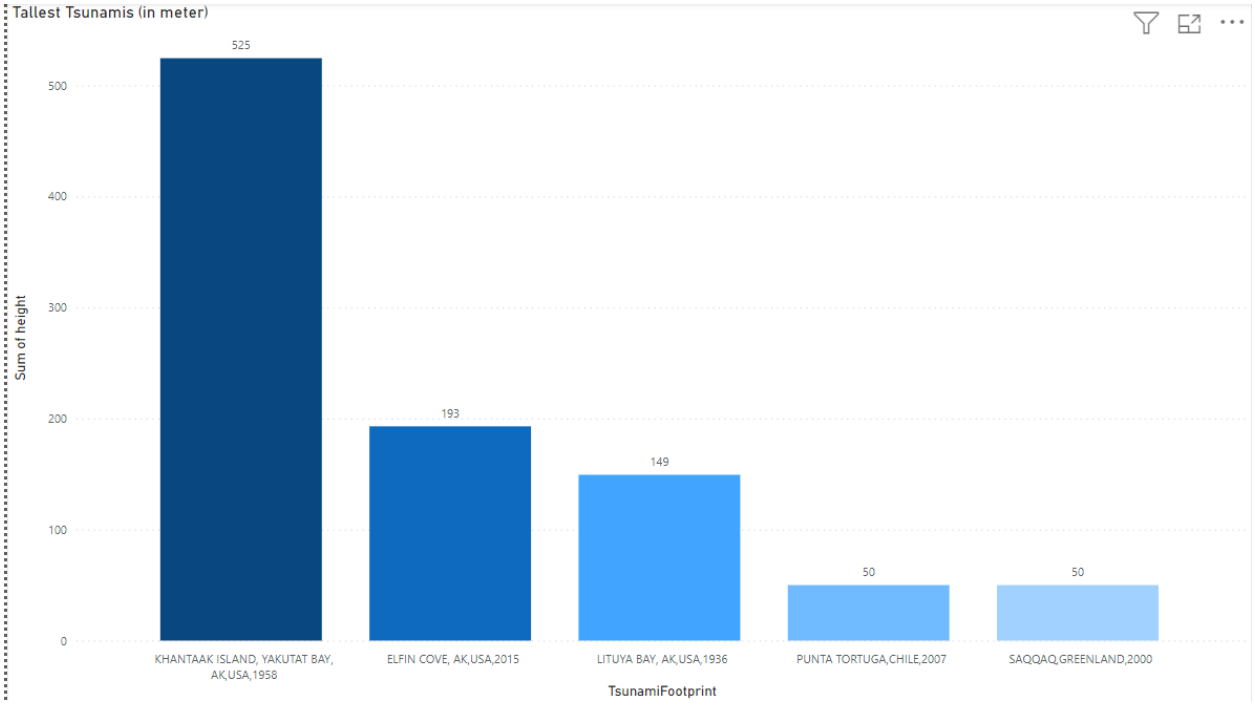
- Delete duplicate entries.
- Delete empty entries
- Give format to the data types, such as converting the Id column from integer to string data type.

# 4.- Analysis phase

After cleaning the dataset, I saved a copy of the dataset and imported it to Power BI to perform analysis and visualizations.

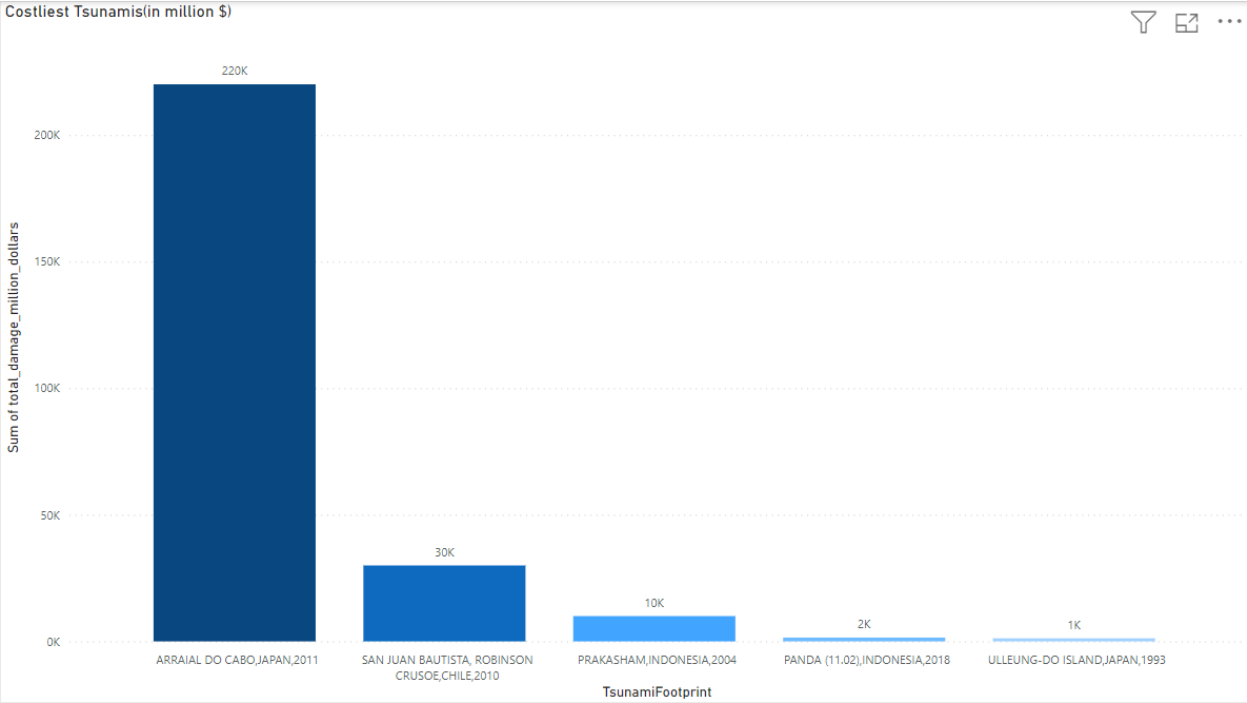
## Top tallest tsunami waves.

Our first analysis reveals the top 5 tallest tsunami waves.



## Top costliest tsunamis.

We found the top 5 costliest tsunamis since the year 1900.

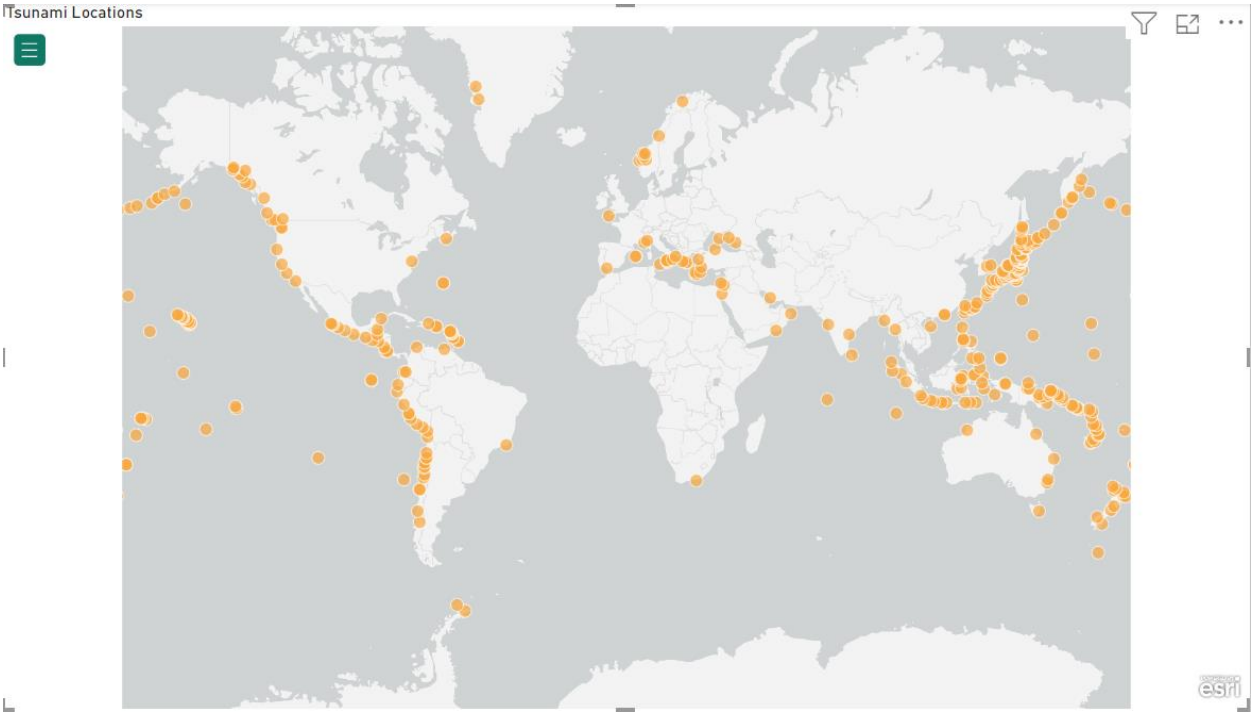


## Total deaths per country.

For the total deaths per country analysis, I created a map in which the color of the country represents the difference in deaths. I used a custom diverging color scheme from blue to red, the bluer the color, the fewer deaths. The redder the color, the more deaths.

## Tsunami occurrence.

For my last analysis, I decided use an ArcGIS map, which pin point the exact location where the tsunami occurred.




# 5.- Conclusion (Act phase)

We finish our analysis and answered all the original questions stated in the first phase of this studycase about the history of tsunamis around the world. If you want you can check the dashboard here.