TSUNAMI

EXECUTIVE SUMMARY:

The objective is to analyze the historic data about Tsunami and if climatic change caused by global warming has impacted the tsunami intensity or tsunami validity before and after 1900.

DATASET SOURCE:

The NCEI/WDS Global Historical Tsunami Database contains tsunami source information. The tsunami source data is related to tsunami runup data which contains information on locations where tsunami effects were observed.

NCEI archives and assimilates tsunami, earthquake, and volcano data to support research, planning, response, and mitigation. Long-term data can be used to establish the history of natural hazard occurrences and help mitigate future events. The natural hazards datasets are available through the HazEL (Hazardous Event Lookup) interface, developed by NCEI.

It has the following columns: Year, Month, Tsunami Event Validity, Tsunami Cause Code, Deposits, Country, Location Name, Number of Runups, Earthquake Magnitude, Latitude, Longitude, Maximum Water Height, Tsunami Intensity, Total Damage Description and After 1990.

Citation: National Geophysical Data Center / World Data Service: NCEI/WDS Global Historical Tsunami Database. NOAA National Centers for Environmental Information.
doi:10.7289/V5PN93H7

COLUMN EXPLANATION:

1. Year – The year Tsunami occurred.
2. Month – The month Tsunami occurred.
3. 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:
    a. -1 - erroneous entry (removed)
    b. 0 - event that only caused a seiche or disturbance in an inland river/lake
    c. 1 - very doubtful tsunami
    d. 2 - questionable tsunami
    e. 3 - probable tsunami
    f. 4 - definite tsunami
4. Tsunami Cause Code: Valid values: 0 to 11, The source of the tsunami:
    a. 0 - Unknown
    b. 1 - Earthquake
    c. 2 - Questionable Earthquake
    d. 3 - Earthquake and Landslide

e. 4 - Volcano and Earthquake
f. 5 - Volcano, Earthquake, and Landslide
g. 6 - Volcano
h. 7 - Volcano and Landslide
i. 8 - Landslide
j. 9 - Meteorological
k. 10 - Explosion
l. 11 - Astronomical Tide

5. Deposits: Criteria commonly used to identify tsunami deposits includes sharp, erosive contact with underlying material, one or more layers of material that fine upward (grain size gets smaller toward the top of the layer), layers that thin landward and The Deposit is numbered based on the above information by geologists.
6. Country: Country name where the Tsunami occurred.
7. Location Name: The area along the coastline.
8. Number of Runups: Total Number of locations along the coastline where the tsunami wave reached a maximum height above the normal sea level.
9. Earthquake Magnitude: Size or Strength of the Earthquake (Using Richter Scale)
10. Latitude and Longitude: The location where the Tsunami originated.
11. Maximum Water Height: Maximum Elevation above the normal sea level that a tsunami wave reached at a specific location.
12. Tsunami Intensity: Potential impact of a tsunami event obtained from maximum water height," "runup height," and "inundation area". Defined by Soloviev and Go (1974) as, I = log (SQRT (2) * h) where h is the average runups height.
13. Total Damage Description:
    a. 0 - None
    b. 1 - Limited (<$1 million)
    c. 2 - Moderate (~$1 to $5 million)
    d. 3 - Severe (~>$5 to $24 million)
    e. 4 - Extreme (~$25 million or more)
14. After 1900: (Created this to compare the Tsunami intensity before and after 1900.
    a. 0 – Any year before 1900
    b. 1 – Any year after 1900

DATA CLEANING:

The dataset has 1433 rows x 47 columns. It had a lot of NaN values which needed to be removed. We used Python to remove the NaN.

Step I:

```python
df =  pd.read_table('tsunamis-2023-04-29_13-58-14_-0500.tsv') #loading raw data
```

Python

```python
df
```

Python

	Search Parameters	Year	Mo	Dy	Hr	Mn	Sec	Tsunami Event Validity	Tsunami Cause Code	Earthquake Magnitude	...	Total Missing	Total Missing Description	Total Injuries	Total Injuries Description	Total Damage ($Mil)	Total Damage Description	Total Houses Destroyed	Total Houses Destroyed Description	Tota House Damage
0	["0 <= Year >= 2023","Probable Tsunami <= Vali...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal
1	NaN	46.0	NaN	NaN	NaN	NaN	NaN	4.0	6.0	6.2	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal
2	NaN	79.0	8.0	24.0	7.0	NaN	NaN	3.0	4.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal
3	NaN	103.0	NaN	NaN	NaN	NaN	NaN	3.0	1.0	7.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal
4	NaN	142.0	NaN	NaN	NaN	NaN	NaN	3.0	1.0	7.0	...	NaN	NaN	NaN	NaN	NaN	3.0	NaN	3.0	Nal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
1428	NaN	2022.0	11.0	22.0	2.0	3.0	7.0	4.0	1.0	7.0	...	NaN	NaN	3.0	1.0	NaN	2.0	NaN	1.0	Nal
1429	NaN	2022.0	12.0	4.0	NaN	NaN	NaN	4.0	6.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal
1430	NaN	2023.0	1.0	9.0	17.0	47.0	35.0	4.0	1.0	7.6	...	NaN	NaN	NaN	NaN	NaN	2.0	37.0	1.0	327.
1431	NaN	2023.0	2.0	6.0	1.0	17.0	35.0	4.0	1.0	7.8	...	NaN	NaN	119200.0	4.0	39300.0	4.0	166581.0	4.0	Nal
1432	NaN	2023.0	3.0	16.0	0.0	56.0	2.0	4.0	1.0	7.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal

1433 rows × 46 columns

Step II:

```python
df = df[['Year','Tsunami Event Validity','Tsunami Cause Code','Deposits','Country','Location Name','Number of Runups','Mo','Earthquake Magnitude','Latitude','Longitude','Maximum Wat
df = df.dropna()     #Droping NaN values
```

Python

```python
df
```

Python

	Year	Tsunami Event Validity	Tsunami Cause Code	Deposits	Country	Location Name	Number of Runups	Mo	Earthquake Magnitude	Latitude	Longitude	Maximum Water Height (m)	Tsunami Intensity	Total Damage Description
14	684.0	4.0	1.0	2.0	JAPAN	NANKAIDO	2.0	11.0	8.4	32.500	134.000	3.00	1.00	3.0
32	1026.0	3.0	1.0	1.0	JAPAN	OFF MASUDA, SHIMANE PREFECTURE	2.0	6.0	7.5	34.800	131.800	10.00	3.00	3.0
39	1096.0	4.0	1.0	3.0	JAPAN	ENSHUNADA	5.0	12.0	8.4	34.000	137.500	7.00	2.00	2.0
62	1433.0	4.0	1.0	0.0	JAPAN	OFF IZU-OSHIMA	1.0	11.0	7.1	34.900	139.500	2.00	1.00	3.0
64	1481.0	4.0	1.0	0.0	GREECE	ISLAND OF RHODES	1.0	5.0	7.1	36.000	28.000	3.00	3.00	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1117	1997.0	4.0	1.0	0.0	SOLOMON ISLANDS	SANTA CRUZ IS. VANUATU	41.0	4.0	7.7	-12.584	166.676	3.00	-2.00	1.0
1146	2002.0	4.0	1.0	0.0	VANUATU	VANUATU ISLANDS	2.0	1.0	7.2	-17.600	167.856	3.00	-4.14	1.0
1151	2002.0	4.0	1.0	0.0	PAPUA NEW GUINEA	BISMARCK SEA	88.0	9.0	7.6	-3.260	142.940	5.50	1.08	3.0
1155	2003.0	4.0	1.0	1.0	MEXICO	S. MEXICO	6.0	1.0	7.5	18.770	-104.104	0.61	0.50	3.0
1156	2003.0	4.0	1.0	1.0	ALGERIA	N. ALGERIA	39.0	5.0	6.8	36.964	3.634	3.00	1.50	4.0

280 rows × 14 columns

Step III: Adding a column ‘After 1900’ and values are 0 – if before 1900, 1 – if after 1900.

```python
def classifi(inp):                                  #classification tsunami sources
    if (inp > 1900.0):
        return 1

    else:
        return 0


l = list(df['Year'])
l = [classifi(i) for i in l ]
df['After 1990'] = l            # create new column which if year >1900 returns 1
print(l)
```

Python

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 
```

Cleaned Dataset: Now the dataset has 280 rows x 15 columns.

Obs	Year	Month	Tsunami Event Validity	Tsunami Cause Code	Deposits	Country	Location Name	Number of Runups	Earthquake Magnitude	Latitude	Longitude	Maximum Water Height (m)	Tsunami Intensity	Total Damage Description	After 1900
1	684	11	4	1	2	JAPAN	NANKAIDO	2	8.4	32.5	134	3	1	3	0
2	1026	6	3	1	1	JAPAN	OFF MASUDA, SHIMANE PREFECTURE	2	7.5	34.8	131.8	10	3	3	0
3	1096	12	4	1	3	JAPAN	ENSHUNADA	5	8.4	34	137.5	7	2	2	0
4	1433	11	4	1	0	JAPAN	OFF IZU-OSHIMA	1	7.1	34.9	139.5	2	1	3	0
5	1481	5	4	1	0	GREECE	ISLAND OF RHODES	1	7.1	36	28	3	3	3	0
6	1495	9	4	1	0	JAPAN	KAMAKURA, SAGAMI BAY, TOKAIDO	2	7.1	35.1	139.5	5	2	1	0
7	1498	9	4	1	5	JAPAN	ENSHUNADA SEA	6	8.3	34	138.1	10	4	3	0
8	1509	9	3	1	0	TURKEY	MARMARA SEA	1	7.7	40.8	28.1	6	3	4	0
9	1510	9	4	1	0	JAPAN	OSAKA BAY	1	6.7	34.6	135.4	2	1	1	0
10	1520	4	4	1	0	JAPAN	KII, KUMANANONADA	1	7	33.6	136.3	2	1	1	0
11	1586	7	4	1	0	PERU	CENTRAL PERU	6	8.5	-12.3	-77.7	26	3.5	3	0
12	1596	9	4	3	0	JAPAN	BEPPU BAY, KYUSHU	2	6.9	33.3	131.7	5	2	3	0
13	1604	11	4	1	0	PERU	S. PERU	6	8.5	-17.88	-70.94	5	3.5	3	0
14	1605	2	4	1	4	JAPAN	NANKAIDO	10	7.9	33	134.9	10	3	3	0
15	1611	12	4	1	4	JAPAN	SANRIKU	11	8.1	39	144.5	25	4	3	0
16	1615	6	4	1	0	UKRAINE	BLACK SEA	1	5.7	44.9	35.5	1	2	1	0
17	1640	7	4	7	3	JAPAN	SE. HOKKAIDO ISLAND	3	6.5	42.07	140.68	8	1	1	0
18	1650	9	4	7	5	GREECE	THERA ISLAND (SANTORINI)	7	6.3	36.404	25.396	30	6	1	0
19	1662	10	4	1	0	JAPAN	HIUGANADA	3	7.6	31.7	132	1	2.5	2	0
20	1670	12	4	1	0	JAPAN	BOSO, JAPAN	1	6.4	35.5	141	2	1	1	0
21	1674	2	4	3	0	INDONES	BANDA SEA	33	6.8	-3.75	127.75	100	1.5	3	0
22	1677	4	4	1	0	JAPAN	SANRIKU	4	8.1	40	144	6	2	1	0
23	1677	11	4	1	2	JAPAN	OFF SE. BOSO PENINSULA	27	7.4	35	141.5	8	3	3	0
24	1693	1	4	1	3	ITALY	ISLAND OF SICILY	11	7.4	37.133	15.017	12	4	4	0
25	1703	12	4	1	2	JAPAN	OFF SW BOSO PENINSULA	48	8.2	34.7	139.8	11.7	3	4	0
26	1707	10	4	1	5	JAPAN	NANKAIDO	31	8.4	33.2	134.8	25.7	4	4	0
27	1711	9	4	1	0	INDONES	BANDA SEA	2	7	-4	129	1.2	1.5	1	0
28	1711	12	4	1	0	JAPAN	SEIONAIKAI, JAPAN	1	6.7	34.3	134	2	1	3	0

STATISTICAL ANALYSIS:

1. First, we wanted to see the relationship between all the variables (except After 1900, Latitude, Longitude, Month and Year) and Tsunami Intensity. For this we decided to run a multiple linear regression model and see if we have significant model, find which variable is causing more impact and find how much variation in Tsunami Intensity is explained by all variables.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	184.94564	26.42081	10.14	<.0001
Error	272	708.73129	2.60563		
Corrected Total	279	893.67694			

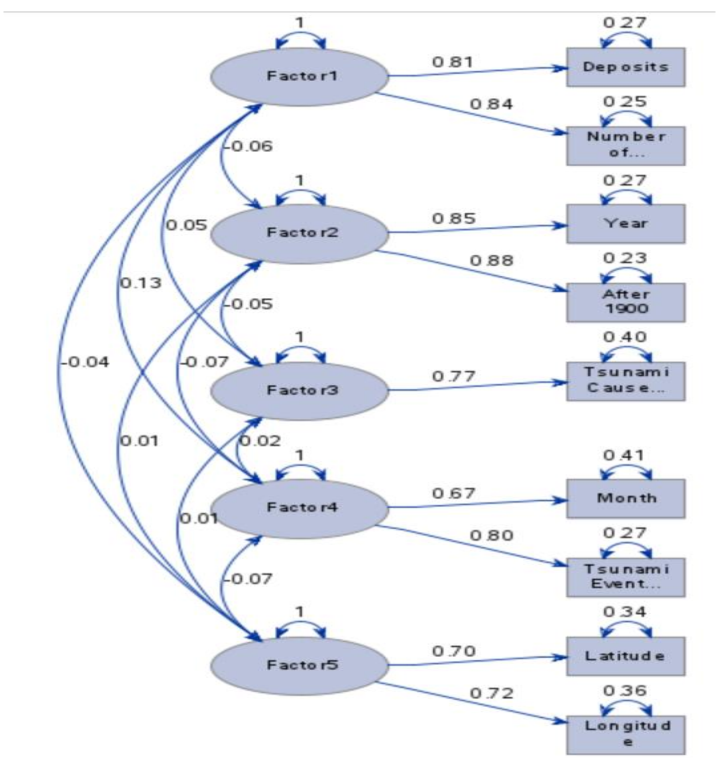
Root MSE	1.61420	R-Square	0.2069
Dependent Mean	1.52229	Adj R-Sq	0.1865
Coeff Var	106.03767		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-1.32685	1.40893	-0.94	0.3472	0	.	0
Tsunami Event Validity	1	-0.29146	0.27828	-1.05	0.2959	-0.05937	0.90729	1.10218
Tsunami Cause Code	1	0.13041	0.11864	1.10	0.2727	0.06505	0.83260	1.20106
Deposits	1	0.11936	0.05518	2.16	0.0314	0.18333	0.40591	2.46360
Number of Runups	1	-0.00065208	0.00150	-0.44	0.6634	-0.03674	0.41001	2.43897
Earthquake Magnitude	1	0.36391	0.16432	2.21	0.0276	0.14295	0.69982	1.42894
Maximum Water Height (m)	1	0.01037	0.00305	3.40	0.0008	0.19349	0.90064	1.11032
Total Damage Description	1	0.38276	0.09660	3.96	<.0001	0.22860	0.87601	1.14154

2. The p value < 0.05, The regression model is significant. About R^2 = 18.65 % of variation in Tsunami Intensity is explained by the variables.
3. The Deposits, Earthquake Magnitude, Maximum Water Height, and Total Damage Description are having significant impact on the Tsunami intensity.
4. After this we decided to reduce the data using Factor Analysis.

Rotated Factor Pattern (Standardized Regression Coefficients)					
	Factor1	Factor2	Factor3	Factor4	Factor5
Number of Runups	0.84312	0.22566	0.07598	0.00243	0.02335
Deposits	0.81212	0.06340	0.24191	-0.02067	0.05805
Total Damage Description	0.57139	-0.14426	-0.08910	-0.01601	-0.03549
Earthquake Magnitude	0.55140	-0.10992	-0.36356	0.36790	-0.24463
After 1900	0.06933	0.88084	0.00695	0.07960	-0.01405
Year	0.06646	0.84532	0.00727	-0.03632	-0.13589
Tsunami Intensity	0.43215	-0.48219	0.23430	0.06349	-0.07191
Tsunami Cause Code	-0.04721	-0.03139	0.76837	-0.08528	0.01241
Maximum Water Height (m)	0.15956	-0.04001	0.58782	0.15213	-0.04483
Tsunami Event Validity	0.11045	0.17320	-0.06857	0.79821	0.29540
Month	-0.28198	-0.13115	0.14235	0.66821	-0.22289
Longitude	0.01680	-0.15016	-0.33960	0.01239	0.71923
Latitude	-0.02231	-0.01503	0.39932	0.04937	0.70281

5. Factor 1 – Effects of Tsunami (Deposits, Number of Runups), Factor 2 – Timeline (Year, After 1900), Factor 3 – Tsunami Cause Code, Factor 4 – (Month, Tsunami Event Validity) and Factor 5 – Location (Latitude and Longitude).



6. We saved the output of the factor analysis dataset as Tsunami_Factors with the factors as columns with the existing columns. (280 rows x 20 columns)
7. For the Tsunami_Factors dataset, we ran multiple linear regression between the new factors, Total Damage Description, Earthquake Magnitude and Maximum Water Height(m).

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	506.71180	63.33897	44.36	<.0001
Error	271	386.96514	1.42792		
Corrected Total	279	893.67694			

Root MSE	1.19495	R-Square	0.5670
Dependent Mean	1.52229	Adj R-Sq	0.5542
Coeff Var	78.49737		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	6.82057	1.55247	4.39	<.0001	0	.	0
Factor1	1	1.11219	0.12513	8.89	<.0001	0.62143	0.32688	3.05922
Factor2	1	-0.93472	0.07530	-12.41	<.0001	-0.52227	0.90273	1.10776
Factor3	1	0.25277	0.11305	2.24	0.0262	0.14123	0.40048	2.49697
Factor5	1	-0.24973	0.07967	-3.13	0.0019	-0.13953	0.80633	1.24019
Factor4	1	0.28891	0.08793	3.29	0.0012	0.16143	0.66190	1.51079
Maximum Water Height (m)	1	-0.00082727	0.00292	-0.28	0.7772	-0.01543	0.53843	1.85726
Earthquake Magnitude	1	-0.66954	0.20020	-3.34	0.0009	-0.26301	0.25836	3.87056
Total Damage Description	1	-0.12246	0.08635	-1.42	0.1573	-0.07314	0.60074	1.66460

8. The p value < 0.05, The regression model is significant. About R^2 = 55.42 % of variation in Tsunami Intensity is explained by the variables.
9. All the factors are having significant impact on Tsunami Intensity except Maximum Water Height and Earthquake Magnitude. Factor 1 (Effects of Tsunami - Deposits, Number of Runups) is having the highest impact on Tsunami Intensity.
10. Now we use this dataset to run a Independent Sample t test to compare the average Tsunami Intensity before and after the year 1900.

After 1900	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		105	2.2219	1.3982	0.1365	-3.0000	6.0000
1		175	1.1025	1.8695	0.1413	-4.1400	5.0000
Diff (1-2)	Pooled		1.1194	1.7085	0.2109		
Diff (1-2)	Satterthwaite		1.1194		0.1964		

After 1900	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		2.2219	1.9513	2.4925	1.3982	1.2313	1.6179
1		1.1025	0.8236	1.3814	1.8695	1.6920	2.0889
Diff (1-2)	Pooled	1.1194	-Infty	1.4674	1.7085	1.5775	1.8633
Diff (1-2)	Satterthwaite	1.1194	-Infty	1.4436			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	278	5.31	1.0000
Satterthwaite	Unequal	264.72	5.70	1.0000

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	174	104	1.79	0.0014

Independent Two Sample T Test Case III (Left Tail)	
Sample 1 Mean, $\overline{x}_1$	2.219
Sample Standard Deviation, s_1	1.398
Sample 1 Size, n_1	105
Sample 2 Mean, $\overline{x}_2$	1.1025
Sample 2 Standard Deviation, s_2	1.8695
Sample 2 Size, n_2	175
Degree of Freedom, MIN(n_1-1,n_2-1)	104
Significance level, α	**0.05**
Finding t Score t	**5.684**
Using p Value Approach	
Finding p value	**1.000**
Using Critical Value Approach	
Finding $t_{critical}$	**-1.660**

11. The p value > 0.05 implying that the mean of Tsunami intensity before 1900 was greater than the mean of Tsunami intensity after 1900.

CONCLUSION

We were able to conclude that climate change has no Impact on Tsunami across the globe based on the Data available in NCEI.

Done By:

Sharath Muruganandam

Sathya Keshav Arigela

Varun Mohankumar Jayasree