MULTIMEDIA UNIVERSITY

TDS2101
Data Science Fundamentals

Earthquake Clustering Analysis

Tutorial Section: TT4L

ID	Name	Email Address
1191103225	Muhammad Waiee bin Zainol	1191103225@student.mmu.edu.my

1. Overview

1.1. Introduction

Earthquakes are natural disasters that can cause significant loss of life and property damage. Predicting earthquakes can help mitigate their impact on society.The aim of this project is to identify patterns in earthquake occurrences over time and identify areas with a high risk of earthquake occurrence.The results can help in disaster preparedness and planning, enabling authorities to take preventive measures and minimize the impact of earthquakes.

1.2. Objectives

1. To identify patterns in earthquake occurrences over time.
2. To identify areas and locations with a high risk of earthquake occurrences.
3. To provide insights and recommendations for earthquake preparedness and risk management measures in the identified regions.

2. Exploratory Data Analysis

2.1. Descriptive Analysis

For this clustering analysis, we will be using “Significant_Earthquakes.csv” as our main dataset.

Earthquake Occurrences by Year

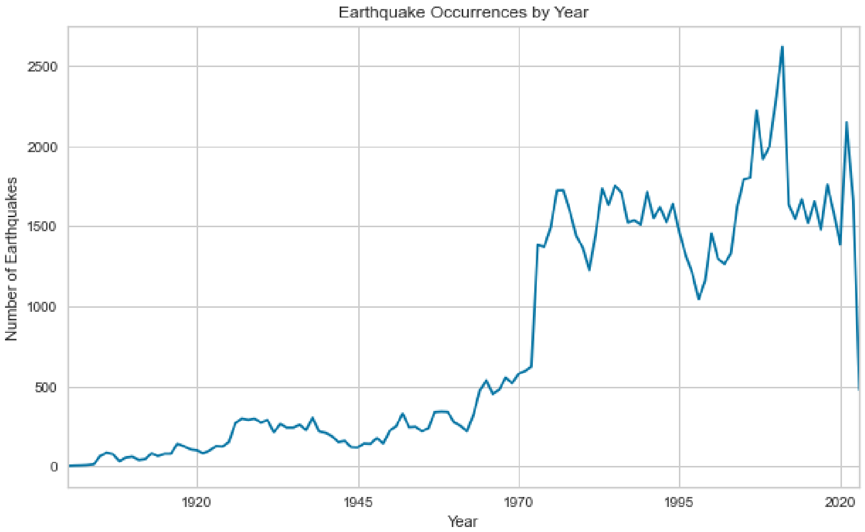


Fig 2.1.1 Earthquake Occurrences by Year

Graph above represents the Earthquake Occurrences by Year. Based on the graph, we can observe that the frequency of earthquakes has been increasing since the early 1900s, which could be due to better detection technology and population growth in seismically active areas. We also can observe that earthquake occurrences are at peak from the year 2000 and above.

Earthquake Occurrences by Month

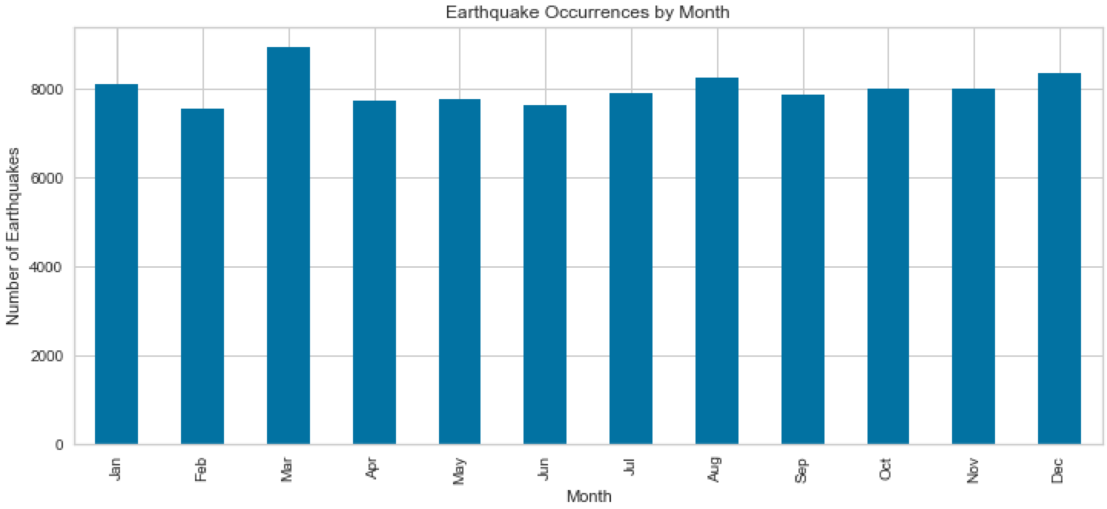


Fig 2.1.2 Earthquake Occurrences by Month

Graph above represents the Earthquake Occurrences by Month. Based on the graph, we can observe that there is no obvious pattern in frequency of the earthquakes based on months. The highest amount of earthquake occurrences is in March since 1900. Meanwhile, the lowest amount of earthquake occurrences is in February. Overall, the amount of earthquakes is distributed and occurred with quite a balanced amount.

Density Distribution of Earthquake Magnitude

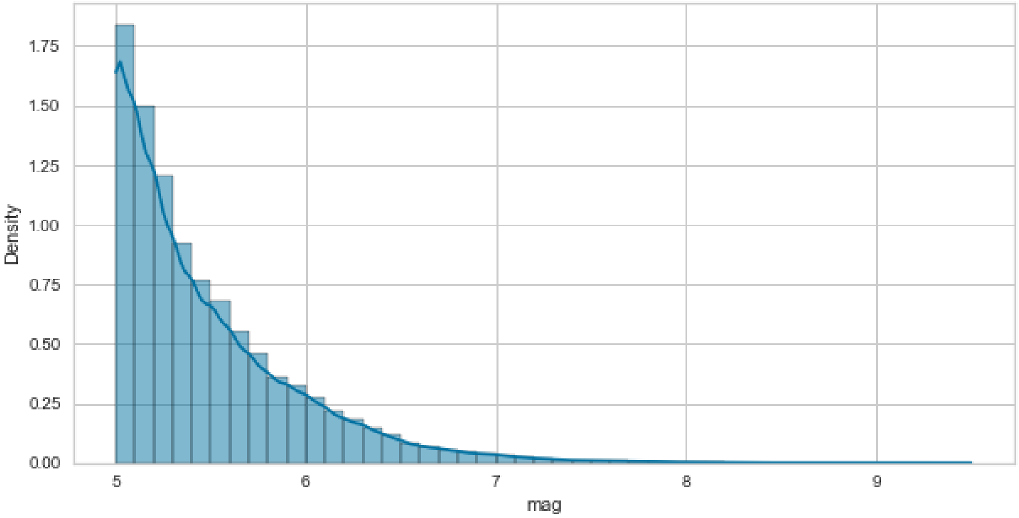


Fig 2.1.3 Density Distribution of Earthquake Magnitude

This plot shows the density distribution of earthquake magnitudes since 1900. The distribution is highly skewed to the left, indicating that the majority of earthquakes had a magnitude closer to 5, while only a few are very strong. The plot also reveals that the distribution follows a logarithmic scale, which means that an increase in magnitude by one unit corresponds to a ten-fold increase in the strength of the earthquake. The distribution curve helps us understand the frequency and severity of earthquakes around the world.

Density Distribution of Earthquake Depth

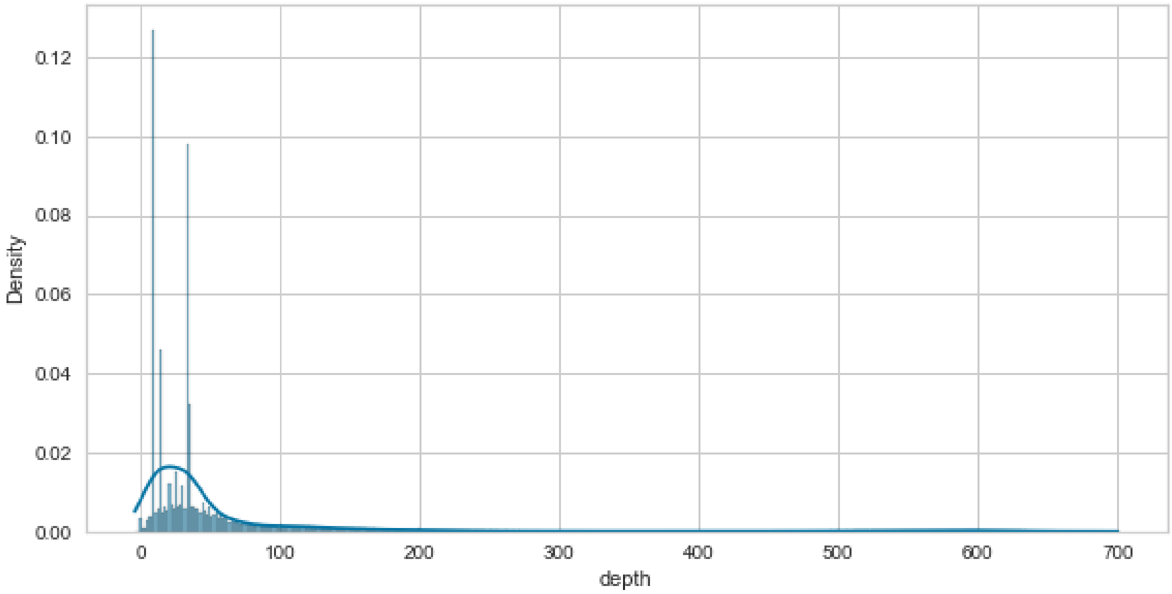


Fig 2.1.4 Density Distribution of Earthquake Depth

This plot shows the density distribution of earthquake depths since 1900. The distribution is also highly skewed to the left, indicating that the majority of earthquakes occur at shallow depths, while only a few occur at deeper depths. The plot reveals that most earthquakes occur within the first 100 kilometers of the Earth's crust. However, some regions, such as the Ring of Fire, experience more frequent and stronger earthquakes at greater depths. The distribution curve helps us understand the spatial and temporal patterns of earthquakes around the world.

2.2. Data Transformation

When performing clustering analysis, it is often beneficial to apply data transformation techniques such as scaling or normalization to ensure that all variables have the same scale or magnitude. Clustering algorithms are sensitive to the scale of variables, and when variables have different units or scales, it can lead to biased results.

For this analysis, I used the commonly used method for data normalization which is **StandardScaler**. **StandardScaler** transforms the data by subtracting the mean and scaling to unit variance. It calculates the z-scores of the data, resulting in transformed data with a mean of zero and a standard deviation of one.

It also helps in addressing issues related to different scales and units of measurement in your dataset. By standardizing the variables, we can ensure that variables with larger variances do not dominate the clustering algorithm. This allows for a fair assessment of their relative contributions to the clustering process. Moreover, it makes the clustering algorithm less sensitive to the magnitude of the variables, leading to more accurate and reliable results.

2.3. Outliers and Missing Values

Outliers

For outlier detection, **boxplot** is used which is a common and straightforward method. It allows us to visually identify potential outliers by analyzing the distribution of our variables as they can have a notable impact on clustering results, also they may introduce noise or skew the distances between data points.

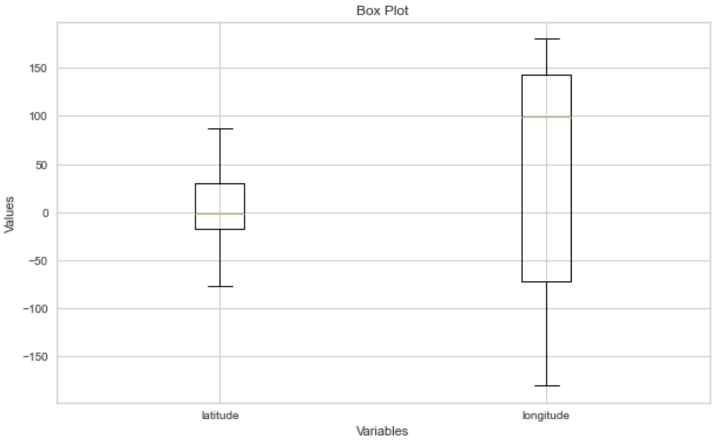


Fig 2.3.1 Boxplot of Latitude and Longitude

Boxplot above represents two main variables that are related to our main objectives, which are latitude and longitude. Based on the boxplot above, we can conclude that there are no obvious extreme values outside the whiskers (the lines that extend from the box). Therefore, we can consider that there are no potential outliers for these variables.

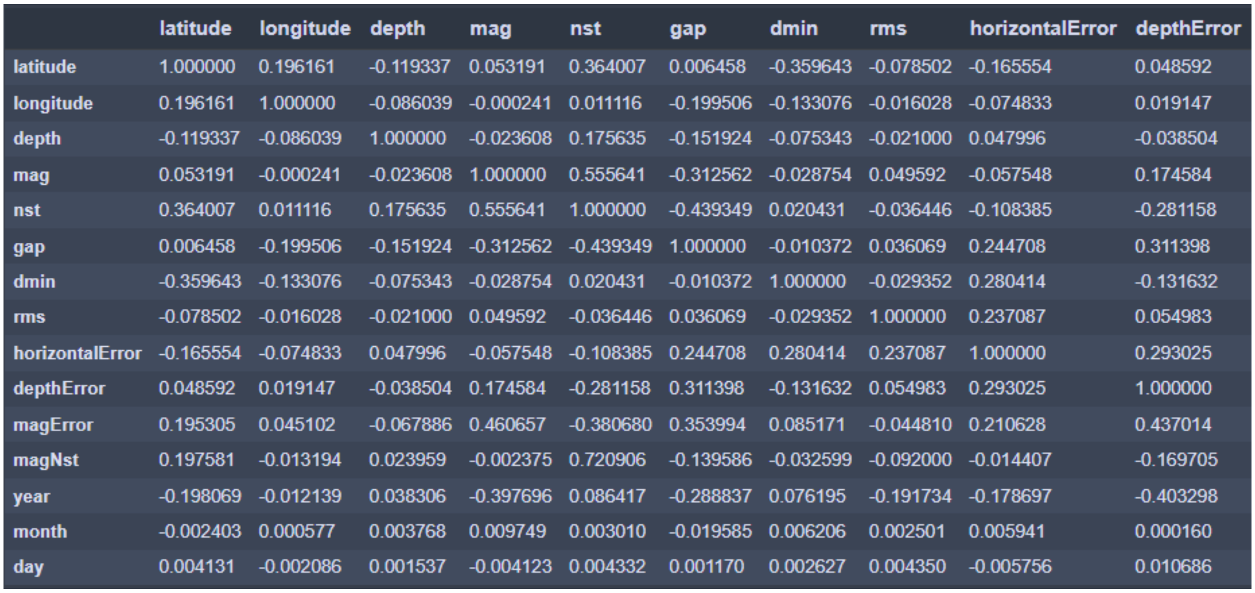
Null/Missing Values

In this analysis, the approach for handling null/missing values is by **interpolating them** using the interpolate() function. This method helps in filling the gaps in the data by estimating values based on the existing data points. Also, the dropna() function is used to **drop remaining missing values** as it has a small amount and might not give a significant impact to our analysis.

2.4. Relationship between variables

To explore the relationships between four main variables (latitude, longitude, depth, mag) in this dataset, I used **correlation analysis** and **heatmaps** as they are effective tools for understanding the strength and direction of relationships between variables.

Correlation Analysis

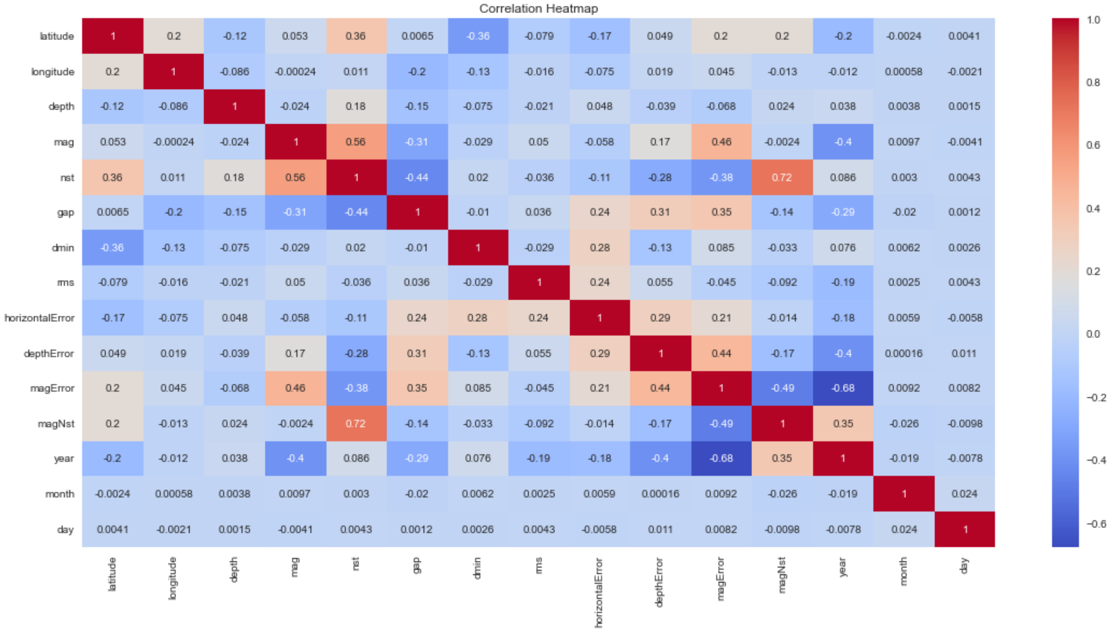


	latitude	longitude	depth	mag	nst	gap	dmin	rms	horizontalError	depthError
latitude	1.000000	0.196161	-0.119337	0.053191	0.364007	0.006458	-0.359643	-0.078502	-0.165554	0.048592
longitude	0.196161	1.000000	-0.086039	-0.000241	0.011116	-0.199506	-0.133076	-0.016028	-0.074833	0.019147
depth	-0.119337	-0.086039	1.000000	-0.023608	0.175635	-0.151924	-0.075343	-0.021000	0.047996	-0.038504
mag	0.053191	-0.000241	-0.023608	1.000000	0.555641	-0.312562	-0.028754	0.049592	-0.057548	0.174584
nst	0.364007	0.011116	0.175635	0.555641	1.000000	-0.439349	0.020431	-0.036446	-0.108385	-0.281158
gap	0.006458	-0.199506	-0.151924	-0.312562	-0.439349	1.000000	-0.010372	0.036069	0.244708	0.311398
dmin	-0.359643	-0.133076	-0.075343	-0.028754	0.020431	-0.010372	1.000000	-0.029352	0.280414	-0.131632
rms	-0.078502	-0.016028	-0.021000	0.049592	-0.036446	0.036069	-0.029352	1.000000	0.237087	0.054983
horizontalError	-0.165554	-0.074833	0.047996	-0.057548	-0.108385	0.244708	0.280414	0.237087	1.000000	0.293025
depthError	0.048592	0.019147	-0.038504	0.174584	-0.281158	0.311398	-0.131632	0.054983	0.293025	1.000000
magError	0.195305	0.045102	-0.067886	0.460657	-0.380680	0.353994	0.085171	-0.044810	0.210628	0.437014
magNst	0.197581	-0.013194	0.023959	-0.002375	0.720906	-0.139586	-0.032599	-0.092000	-0.014407	-0.169705
year	-0.198069	-0.012139	0.038306	-0.397696	0.086417	-0.288837	0.076195	-0.191734	-0.178697	-0.403298
month	-0.002403	0.000577	0.003768	0.009749	0.003010	-0.019585	0.006206	0.002501	0.005941	0.000160
day	0.004131	-0.002086	0.001537	-0.004123	0.004332	0.001170	0.002627	0.004350	-0.005756	0.010686

Fig 2.4.1 Correlation Analysis Table

In this correlation analysis, I used the **corr()** function to obtain a correlation matrix. It provides a numerical measure of the linear relationship between pairs of variables. Based on correlation analysis in the table above, the highest positive correlation is **0.720906**, between magNst and nst. Other than that, the highest negative correlation is **-0.676992**, between year and magError.

Heatmaps



	latitude	longitude	depth	mag	nst	gap	dmin	rms	horizontalError	depthError	magError	magNst	year	month	day
latitude	1	0.2	-0.12	0.053	0.36	0.0065	-0.36	-0.079	-0.17	0.049	0.2	0.2	-0.2	-0.0024	0.0041
longitude	0.2	1	-0.086	-0.00024	0.011	-0.2	-0.13	-0.016	-0.075	0.019	0.045	-0.013	-0.012	0.00058	-0.0021
depth	-0.12	-0.086	1	-0.024	0.18	-0.15	-0.075	-0.021	0.048	-0.039	-0.068	0.024	0.038	0.0038	0.0015
mag	0.053	-0.00024	-0.024	1	0.56	-0.31	-0.029	0.05	-0.058	0.17	0.46	-0.0024	-0.4	0.0097	-0.0041
nst	0.36	0.011	0.18	0.56	1	-0.44	0.02	-0.036	-0.11	-0.28	-0.38	0.72	0.086	0.003	0.0043
gap	0.0065	-0.2	-0.15	-0.31	-0.44	1	-0.01	0.036	0.24	0.31	0.35	-0.14	-0.29	-0.02	0.0012
dmin	-0.36	-0.13	-0.075	-0.029	0.02	-0.01	1	-0.029	0.28	-0.13	0.085	-0.033	0.076	0.0062	0.0026
rms	-0.079	-0.016	-0.021	0.05	-0.036	0.036	-0.029	1	0.24	0.055	-0.045	-0.092	-0.19	0.0025	0.0043
horizontalError	-0.17	-0.075	0.048	-0.058	-0.11	0.24	0.28	0.24	1	0.29	0.21	-0.014	-0.18	0.0059	-0.0058
depthError	0.049	0.019	-0.039	0.17	-0.28	0.31	-0.13	0.055	0.29	1	0.44	-0.17	-0.4	0.00016	0.011
magError	0.2	0.045	-0.068	0.46	-0.38	0.35	0.085	-0.045	0.21	0.44	1	-0.49	-0.68	0.0092	0.0082
magNst	0.2	-0.013	0.024	-0.0024	0.72	-0.14	-0.033	-0.092	-0.014	-0.17	-0.49	1	0.35	-0.026	-0.0098
year	-0.2	-0.012	0.038	-0.4	0.086	-0.29	0.076	-0.19	-0.18	-0.4	-0.68	0.35	1	-0.019	-0.0078
month	-0.0024	0.00058	0.0038	0.0097	0.003	-0.02	0.0062	0.0025	0.0059	0.00016	0.0092	-0.026	-0.019	1	0.024
day	0.0041	-0.0021	0.0015	-0.0041	0.0043	0.0012	0.0026	0.0043	-0.0058	0.011	0.0082	-0.0098	-0.0078	0.024	1

Fig 2.4.2 Correlation Heatmap

Based on the heatmap above, it shows us almost the same result as previous correlation analysis. The highest positive correlation is **0.72**, between magNst and nst with brighter color. Other than that, the highest negative correlation is **-0.68** between year and magError with the darkest color.

3. Feature Selection

3.1. Univariate Feature Selection

Univariate feature selection is a commonly used method in feature selection, and it can be a suitable choice for this dataset because it is relatively **straightforward to implement and interpret**. It assesses the relationship between each feature and the target variable independently, making it easy to understand which features have the strongest correlation with the target. It also utilizes statistical tests to **measure the significance of the relationship between each feature and the target variable**. This helps in identifying features that have a strong impact on the target variable and are likely to be informative for clustering analysis. By selecting features based on their relationship with the target variable, this feature selection aims to **retain the most relevant and informative features for the clustering analysis**. This can help in capturing the essential characteristics of the earthquake data and identifying the significant factors contributing to earthquake occurrence.

3.2. Optimal Feature Set

From this selection technique, we can find optimal features by ranking the features based on their relevance or statistical scores. Based on the result, the top features that are relevant are **latitude**, **longitude**, **magnitude**, and **depth**.

4. Model Construction and Comparison

4.1. Particular model used

K-Means Clustering is a commonly used model for clustering analysis because it is simple, easy to implement, and computationally efficient. It works well when the data has distinct clusters and the objective is to partition the data points into groups based on their similarity.

4.2. Output of the model

Scatter Plot

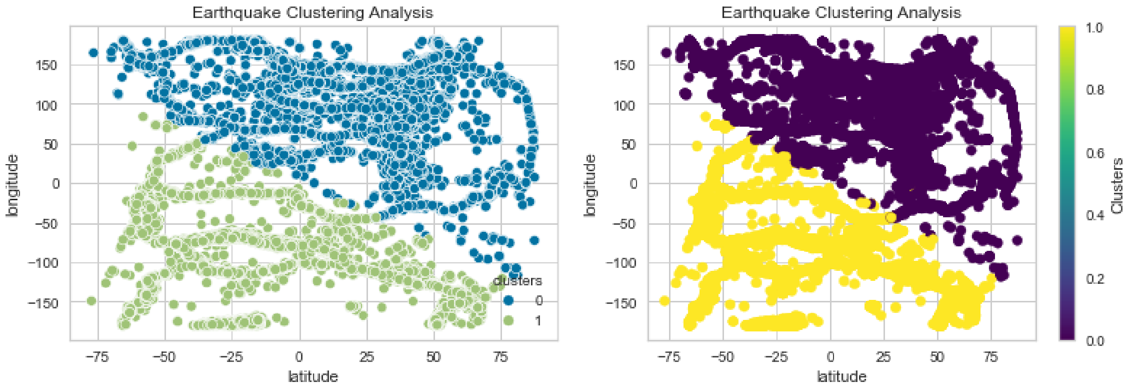


Fig 4.2.1 Earthquake Clustering Analysis Scatter Plot

Both plots above represent Earthquake Clustering analysis which consist of latitude and longitude. Based on the figures above, we can observe that the majority of points are **well-clustered and distinct**, indicating that the clustering algorithm has successfully identified patterns in the data. However, there are a **few points that appear to be overlapped with each other**, potentially indicating some level of ambiguity or similarity between those data points. It also indicates there is a possibility of outliers or noisy data points.

Silhouette Plot

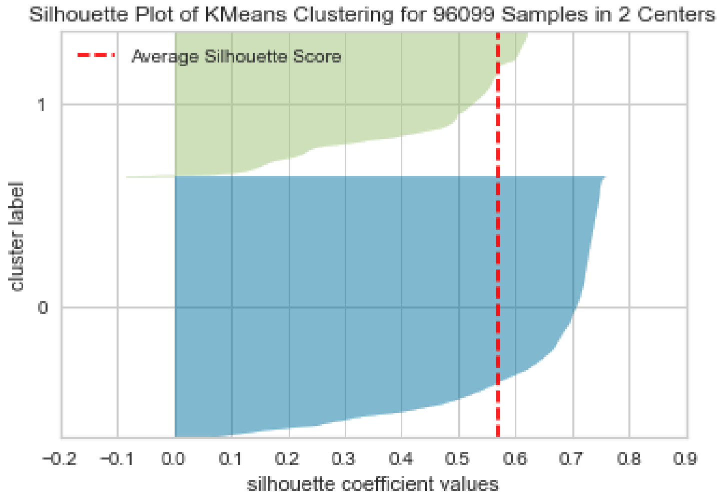


Fig 4.2.2 Silhouette Plot

Based on the silhouette plot above, the average silhouette score ranges from **0.5 to 0.6**. This indicates a moderate level of separation and cohesion within the clusters. A silhouette score closer to 1 suggests well-defined and distinct clusters, while scores closer to 0 indicate overlapping or poorly separated clusters.

In the silhouette plot, it is observed that one of the clusters has a larger figure size compared to the other cluster. This suggests that the larger cluster contains more data points than the smaller cluster. The imbalance in cluster sizes could potentially impact the overall clustering performance and interpretation of the results. In this case, the silhouette plot suggests that the clustering algorithm has succeeded in creating distinct clusters.

Basemap

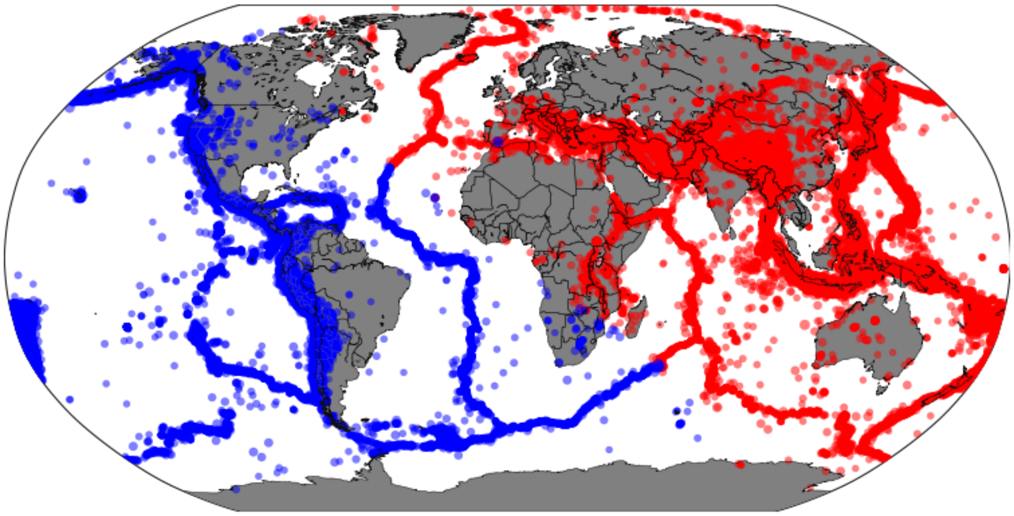


Fig 4.2.3 Basemap

This clustered map displays every earthquake that has occured around the globe since 1900. Each point on the map represents the location (latitude and longitude) of the earthquake which is plotted based on their epicenter coordinates. From this figure, we can identify areas and locations with a high risk of earthquake occurrences. The map clearly shows that earthquakes are clustered into two major groups. Both clusters label have different characteristics which we can measure their differences by magnitude and depth. The red color indicates high risk location of earthquake occurrences while blue color indicates low risk location of earthquake occurrences.

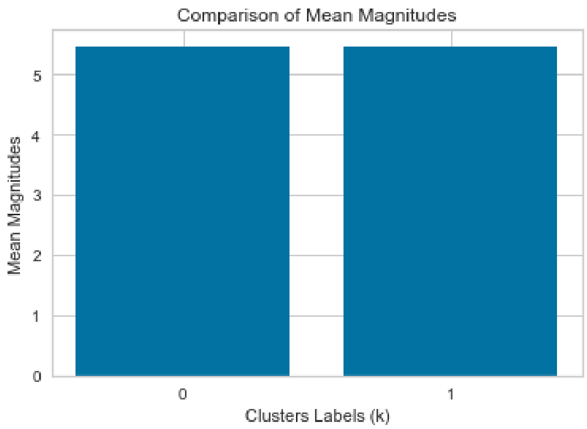


Fig 4.2.4 Comparison of Mean Magnitude

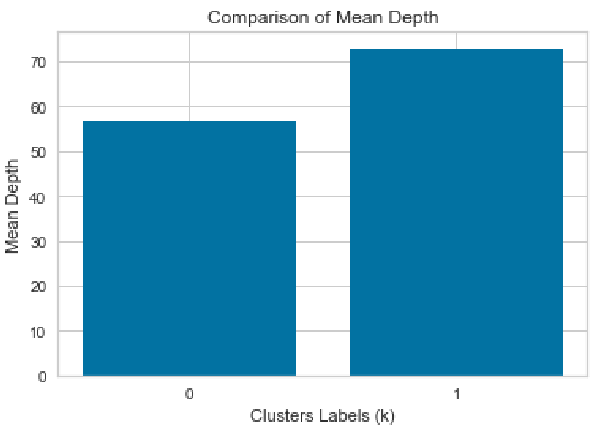
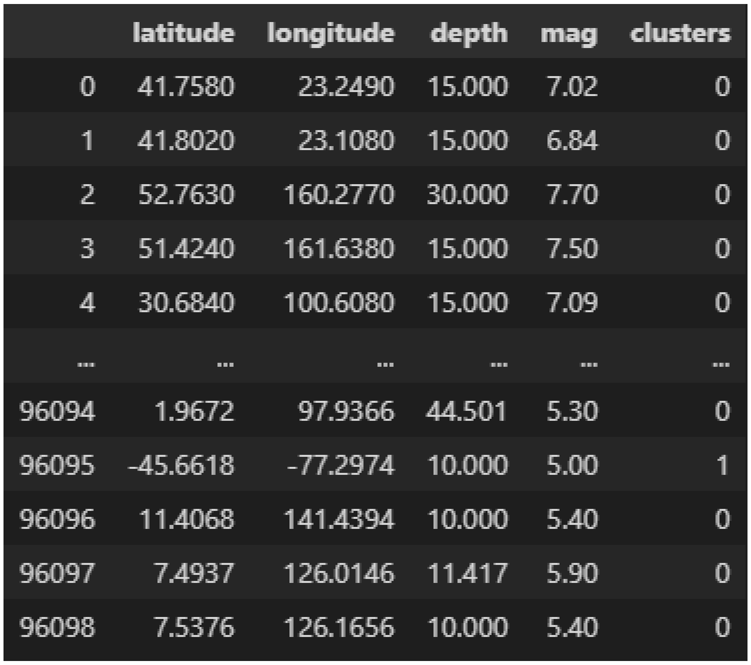


Fig 4.2.5 Comparison of Mean Depth

Figure 4.2.4 represents Comparison of Mean Magnitudes of k=2.. Based on the graph, we can observe that cluster label 1 has slightly higher mean magnitude (**5.461928395135152**) compared to label 0 (**5.453090283336001**). Moreover, Figure 4.2.5 represents Comparison of Mean Depth of k=2. Based on the graph, we can observe that cluster label 1 has higher mean depth (**72.87603063277942**) compared to label 0 (**56.677855162478004**).

Based on the result, we can observe that cluster label 1 has higher mean magnitude and mean depth. Thus, we can conclude that **cluster label 1 has higher risk of earthquake occurrences compared to label 0**.

Final Dataframe



	latitude	longitude	depth	mag	clusters
0	41.7580	23.2490	15.000	7.02	0
1	41.8020	23.1080	15.000	6.84	0
2	52.7630	160.2770	30.000	7.70	0
3	51.4240	161.6380	15.000	7.50	0
4	30.6840	100.6080	15.000	7.09	0
...	...	...	...	...	...
96094	1.9672	97.9366	44.501	5.30	0
96095	-45.6618	-77.2974	10.000	5.00	1
96096	11.4068	141.4394	10.000	5.40	0
96097	7.4937	126.0146	11.417	5.90	0
96098	7.5376	126.1656	10.000	5.40	0

Fig 4.2.6 Final Dataframe

Through this clustering analysis, the identified regions have been categorized into clusters that indicate the level of earthquake risk. This classification **provides valuable insights into the spatial distribution of earthquake occurrences and helps prioritize risk management efforts**.

Given the identified high-risk areas, it is **recommended to review and update building codes and regulations**. Implementing stricter standards for construction and ensuring compliance through inspections can significantly improve the structural integrity of buildings, reducing the potential impact of earthquakes.

By **establishing and expanding early warning systems** enables residents to receive advance notice of impending earthquakes. This, coupled with enhanced emergency response capabilities, such as drills and training for first responders, ensures a rapid and effective response during seismic events.Critical infrastructure, including hospitals, schools, and transportation networks, should undergo regular seismic vulnerability assessments. Identifying potential vulnerabilities and implementing mitigation measures enhances the resilience of key infrastructure components.

Furthermore, **effective communication and coordination among relevant stakeholders are vital** for successful earthquake preparedness and risk management. Collaboration with government agencies, emergency management organizations, community groups, and businesses ensures a cohesive approach to address earthquake risks in the identified regions.

4.3. Validate and Compare Models

To validate the models, we can use the Elbow Method to find optimal K(number of clusters) values and Silhouette method to compare the clustering scores.

<u>Elbow Method</u>

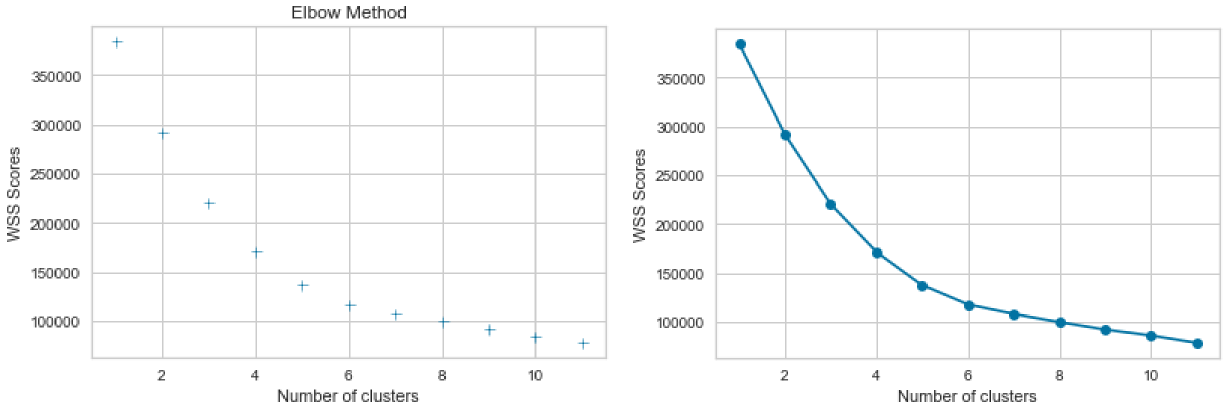


Fig 4.3.1 Elbow Method

The plot above shows the Elbow method which consists of WSS(Within-Cluster Sum of Squares) Scores and number of clusters. The goal is to find which cluster has elbow shape. Based on the plot above, we can observe that the “elbow” is in range 3-4. After that, the number of clusters decreases continuously, which indicates that there is no potential “elbow” in that range. Overall, we can conclude that the **optimal k(number of clusters) values is in the 3-4**.

Silhouette Method

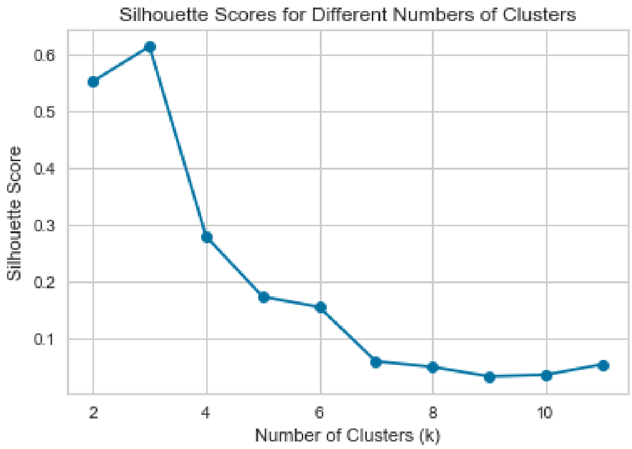
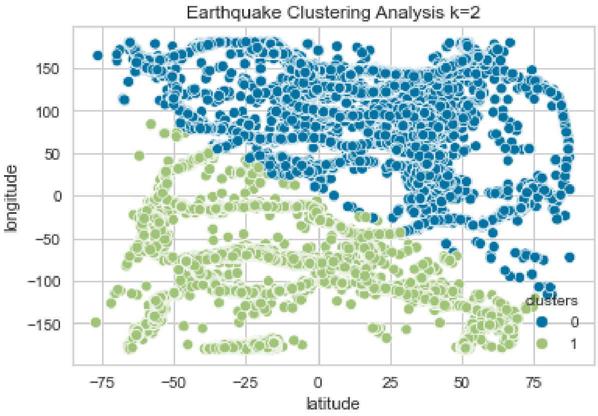


Fig 4.3.2 Silhouette Method

The plot above represents Silhouette Scores for different numbers of clusters. From the plot, we can observe that k=3 has the highest silhouette score, which is **0.612689** followed by k=2 with silhouette score of **0.550902.** This result shows that the **ideal k(number of clusters) values are between 2 and 3**. Now, we can compare models using both clusters.

Models Comparison



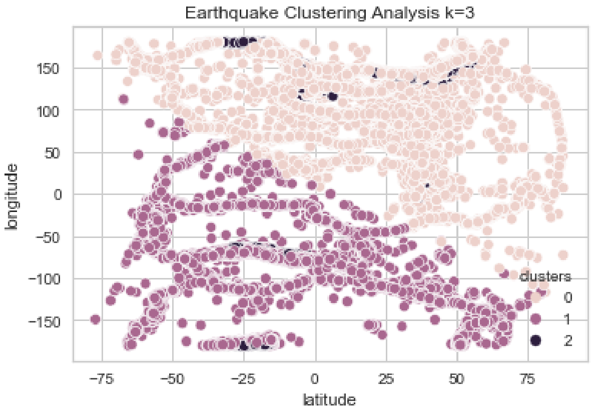


Fig 4.3.3 Clustering k=2

Fig 4.3.4 Clustering k=3

The graph above represents clustering analysis where k=2 and clustering analysis where k=3.
Based on the plot, we can observe that the Fig 4.3.3 has less overlapped points compared to the Fig 4.3.4 which indicates better performance in clustering. Moreover, we can see an obvious pattern in the first figure compared to the second figure which shows the data is clustered well. Thus, we can conclude that

the Fig 4.3.3 has better clustering results since it has a small amount of overlapped data points and has an obvious pattern. Since our objective is just to identify high and low risk locations for earthquake occurrences, therefore **we choose k = 2 in this case**. Further studies are needed in future in order to obtain better and more accurate results.

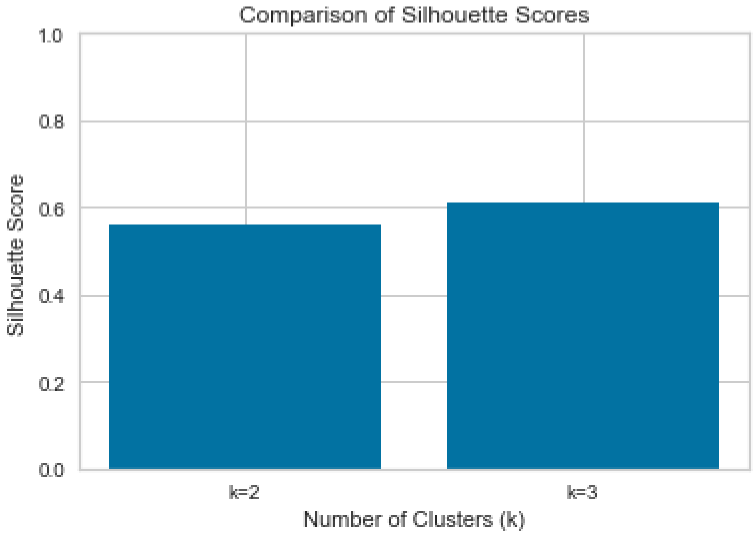


Fig 4.3.5 Comparison of Silhouette Scores

Histogram above shows the comparison of silhouette scores between k values which are k=2 and k=3. Based on the graph, we can observe that k=3 has a higher silhouette score (**0.6127978332976023**) compared to k=2 (**0.5600029858262092**) which shows that it is the best value for k. However, due to overlapping data points and less obvious patterns, thus we choose k=2 as the k value in this case because of not too much difference in terms of silhouette score.

4.4. Impact of Features towards Modeling

Some features may have a stronger influence on the clustering outcomes, while others may contribute less or introduce noise. Identifying the impactful features allows for more focused and efficient analysis.
The impact of features **can affect the performance of the clustering model**. Certain features may lead to better separation and more distinct clusters, while others may introduce noise or confusion.

Methods to assess the impact of features in clustering analysis may include **correlation analysis, feature importance measures, and dimensionality reduction techniques**.

4.5. Hyper-parameter Tuning

K-Means Clustering has a single hyper-parameter, which is the number of clusters (K). We can perform hyper-parameter tuning by **trying different values of K** and **evaluating the clustering performance** using metrics like the silhouette score or WCSS (as mentioned on models comparison section). Typically, it would iterate over a range of K values and choose the one that maximizes the clustering quality.

5. Deployment

5.1. Hosted on Streamlit

This clustering analysis is hosted on Streamlit. It can be assessed by this link:
https://earthquake-analysis-k-means-clustering-waiee.streamlit.app

5.2. Performance and strategy

To address the issue of overlapping points in this clustering analysis, we may consider adjusting the clustering algorithm parameters, and exploring alternative clustering algorithms that may better handle overlapping data points. Additionally, outlier detection and removal techniques can be employed to mitigate the impact of outliers on the clustering results.

It would also be beneficial to further investigate the characteristics of the clusters, such as their centroid locations, spread, and proximity to each other. Furthermore, evaluating other clustering algorithms and adjusting the number of clusters may help improve the separation and balance of the clusters.