THE UNIVERSITY OF SYDNEY

Assignment Coversheet – GROUP ASSIGNMENT

Please fill in your details below. Use one form for each group assignment.

Personal Details of Students

Group Number	Group 26				
Family Name	Given Name (s)	Student Number (SID)	Unikey	Contribution + Percentage	Signature
Venkatakrishnan Idayakani	**Koushik**	**520634594**	kven4699	Complex visualisation and bonus visualisation. Contributed in making the graph data	Koushik
Datta	**Suhrid**	**520188710**	sdat6925	Middle visualisation and report formatting. Contributed in making the graph data	Suhrid
Pereira	**Reenal**	**520112117**	rper5801	Simple Visualisation and report writing. Contributed in making the graph data	Reenal
Gain	**Sanjukta**	**520372582**	sgai0121	Simple Visualisation and Report writing. Contributed in making the graph data	Sanjukta

Assignment Details:

Assignment Title		**Exploring the earthquake dataset**			
Assignment number		**02**			
Unit of Study Tutor		**Shijun Cai**			
Tutorial ID		**013**			
Due Date	**25/05/2023**	Submission Date	**25/05/2023**	Word Count	**7500 words**

Declaration:

1. I understand that all forms of plagiarism and unauthorised collusion are regarded as academic dishonesty by the university, resulting in penalties including failure of the unit of study and possible disciplinary action.
2. I have completed the **Academic Honesty Education Module** on Canvas.
3. I understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the ***University of Sydney By-Law 1999*** (as amended).
4. This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.
5. The assessment has not been submitted previously for assessment in this or any other unit, or another institution.
6. I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment may:
   a. Reproduce this assignment and provide a copy to another member of the school; and/or
   b. Use similarity detection software (which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking).
7. I have retained a duplicate copy of the assignment.

Please type in your group number here to acknowledge this declaration:	Group 26

COMP 5048

Suhrid Datta SID 520188710
Reenal Raina Pereira SID 520112117
Koushik Venkatakrishnan SID 520634594
Sanjukta Gain SID 520372582

Group Assignment Report

May 25, 2023

THE UNIVERSITY OF
SYDNEY

Contents

1 Introduction

1.1 Data set

The Earthquake Dataset, as hosted on Kaggle and curated by Warcoder, serves as a valuable resource for scientists, researchers, and analysts in the field of seismology and natural disaster management. By encompassing data from 2001-2022, the dataset provides a rich historical perspective on earthquake activity across the globe. With 782 individual earthquake events recorded, it offers ample data for meaningful analysis and the development of predictive models. Each record in the dataset is packed with essential information, including:

- DateTime: The date of the earthquake event along with the exact time, allowing for temporal analysis and the required temporal granularity for identification of patterns over time.
- Latitude: The geographical latitude of the earthquake's epicenter, enabling the study of geographical distribution and potential links to tectonic plate boundaries or geological features.
- Longitude: The geographical longitude of the earthquake's epicenter, further enhancing the spatial analysis capabilities. Depth: The depth of the earthquake below the Earth's surface, which can provide insights into the underlying geological processes and potential impact on infrastructure or populations.
- Magnitude: The earthquake's magnitude, typically measured on the moment magnitude scale or the Richter scale, which helps quantify the energy released by the event and assess its potential destructiveness.

The Earthquake Dataset's granular detail empowers researchers and analysts to investigate various aspects of earthquake events, including analyzing the frequency and distribution of earthquakes in specific regions or worldwide, which can help identify areas with higher seismic risk and inform disaster preparedness efforts. Studying correlations between various factors, such as depth and magnitude, to better understand the underlying geophysical processes and their implications for earthquake impacts. Identifying patterns or trends in earthquake activity over time, which can be used to develop early warning systems or inform infrastructure planning and resilience efforts. Developing predictive models for future earthquake events, utilizing machine learning algorithms and statistical techniques to better anticipate and prepare for potential seismic hazards. By providing a comprehensive and detailed view of global earthquake activity, the Earthquake Dataset is a valuable tool for advancing our understanding of these natural phenomena and improving our ability to manage their risks and consequences

1.2 Summary of Contrubution

- Task 1 - group contribution, Suhrid processed the data, reenal and sanjukta came up with plans on analysis, visualisation and Koushik implemented them in gephi and tableau.
- Task 1.1 - Koushik
- Task 2 - Koushik
- Task 3 - Suhrid
- Task 4 - Koushik
- Task 4.1 - Reenal
- Task 4.2 - Sanjukta

2 Design

2.1 Tasks

Overarching task Identify the patterns in impact statistic of most frequently occurring earthquakes and explore the trend of earthquakes over time.
Sub tasks to comprehend and provide an output for the overall task:

1. Explore earthquakes based on their similarity of impact statistics?
   Motivation: The objective of this task is to determine the most frequently occurring earthquake type based on their impact statistics.

   Which earthquakes are the most similar?

2. Identify the pattern or trend in the statistic of most frequently occurring earthquakes?
   Motivation: From this task we can get insights on the most significant impact metric and explore the trend of all statistics.

3. Identify correlation between impact statistics?
   Motivation: To define a statistic that can be used to investigate the dynamic trends present within the data set.

4. Explore the temporal dynamics of earthquakes based on a significantly important metric between 2001-2022.
   Motivation: The objective is to understand the development and impact of earthquakes over the years across multiple countries

   Top 5 most affected countries?

   Frequency of earthquakes over the years?

Task classification

Task	Data Type	Complexity
1	Graph Data	Middle
1.1	Graph data	Middle
2	High-Dimensional	Middle
3	High-Dimensional	Middle
4	Dynamic	Complex
4.1	High-Dimensional	Simple
4.2	Dynamic	Simple

Table 1: Classifying tasks based on their data type and complexity

2.2 Data processing

To explore the patterns of earthquakes we needed to create 3 different subset from the earthquake dataset.

2.2.1 Subset 1 - Graph data

The data processing for identifying the most commonly occurring earthquakes involves sub setting the data to select only relevant non-geographical features. The features, namely latitude, longitude, country, and continent, were eliminated from the dataset. The data was then cleaned by inspecting for missing values and handling them accordingly. This process aimed to ensure the data was suitably prepared for analysis and to identify any problems that could compromise the accuracy of the results.

The subset of the data was carefully prepared by including only the earthquake titles and numerical quantities that describe the impact caused by the earthquakes. This subset will be used for visualising Task 1 and Task 1.1.

2.2.2 Subset 2 - High-Dimensional data

From Task 1 visualisation, the nodes belonging to the largest cluster were extracted and saved into a separate Excel sheet. Using string regularization techniques, the names of these nodes were matched with

the original earthquake dataset. From this matching process, the corresponding statistics of these nodes or earthquakes were extracted.

Additionally, certain missing data points, such as country or continent were manually imputed based on common knowledge. These imputations were made to ensure that the data set was as complete as possible for further analysis and interpretation.

The subset will be used to visualise task 2.

2.2.3 Subset 3 - High-Dimensional data

In Python, a subset of earthquake data was constructed using the pandas library by selecting specific columns from the larger earthquake dataset and omitting rows that contain null values. The columns selected were "magnitude" ,"CDI", "mmi", "depth", "sig", and "nst". This subset will be used for visualising task 3.

2.2.4 Subset 4 - Dynamic data

From this subset we would extract data which explored the temporal dynamics of the earthquake. Thus, it focused on three essential features: 'magnitude', 'title', and 'country'.

In this section, a total of 22 subsets were generated. Each subset was created by merging the data from previous years with the data from the current year. Over time, multiple subsets were incrementally formed, incorporating the information from preceding years as new data was added. This approach was implemented to ensure that each year's dataset was initialized with the data from previous years, following the smart-initialization approach. This method facilitated a comprehensive examination of the relative changes in the impact of earthquakes over the years, specifically focusing on a particular country.

An important step in creating this subset involved identifying null values (NA) and their corresponding rows. These missing values were manually filled to ensure data completeness. For instance, a significant number of country names were initially null values, even though they were mentioned in the title. Therefore, the country names were extracted from the title and inserted into the country column. This data cleaning approach guarantees that subsequent analyses and visualizations are conducted using comprehensive and accurate data.

The cleaned data was then fed into an online Voronoi diagram generator [3]. The voronoi generator provided x and y values of the voronoi splits,along with a value column. For this particular subset, the Voronoi diagrams were constructed using the magnitude as the value, countries as the grouping criterion, and the title of the earthquake as the splitting factor.

This subset will be used to visualise tasks 4, 4.1 and 4.2.

2.3 Analysis

2.3.1 Task 1

- Clustering based on modularity community detection algorithm, represented by their distinct colors.

To analyze the similarity of impact metrics, we employed the modularity community detection algorithm to observe the clustering of nodes. The modularity community detection method is considered the most suitable approach for identifying clusters of nodes characterized by dense connections. Although this method may have limitations in detecting small clusters within a network, it is the optimal choice for our task since our objective is to identify the largest cluster of nodes.

- Node size based on the number of connected edges,degree centrality. To identify most connected earthquakes.

  The degree centrality of each node, which represents the number of connections it has, was visually emphasized by depicting it with a corresponding size. This clear representation allows for the easy identification of the earthquakes that are most similar based on their centrality in the network.

2.3.2 Task 1.1

- edge bundling to create a similar effect as weighted edges, thicker edges means more connections.

  To distinctly identify the most similar earthquakes, all their labels were enabled with edge bundling in the visualization. Edge bundling, which is particularly well-suited for this task, was chosen due to each node having a minimum of 60 or more edges. This approach not only enhanced the readability of the visualization but also conveyed the level of similarity through the thickness of the edges.

2.3.3 Task 4

- Clustering based on country as group, split by title of earthquakes.
- Smart-Initialisation method to explore the temporal dynamics of earthquakes in countries over each year.

2.4 Visualisation

2.4.1 Task 1: Overall Graph

The graph will be visualized using the Fruchterman-Reingold layout, which is particularly well-suited for visualizing large undirected graphs. This layout algorithm aims to position the nodes in a way that minimizes overlapping edges and maximizes the readability of the graph.

In addition to the layout, the color and size of the nodes will integrate the degree centrality analysis and modularity community detection algorithm, respectively. Degree centrality determines the number of connections each node has, and this information is represented by the color of the nodes. Nodes with higher degrees will have distinct colors, allowing for a visual identification of their centrality. On the other hand, the size of the nodes reflects the clusters identified through the modularity community detection algorithm. Larger nodes represent clusters with denser connections, making it easier to identify distinct groups within the graph.

By combining the Fruchterman-Reingold layout, degree centrality analysis, and modularity community detection algorithm, the visualization provides a comprehensive and informative representation of the large undirected graph, facilitating the analysis and interpretation of the data.

2.4.2 Task 1.1: Detailed Graph

The detailed graph will also be visualised using Fruchterman-Reingold method. edge bundling will be applied to minimise edge crossing and increase readability.

2.4.3 Task 2: Parallel Coordinates

Parallel coordinates is a visualization technique that is particularly useful for exploring and analyzing high-dimensional datasets. In this technique, each dimension is represented by a vertical axis, and individual data points are represented as lines that traverse these axes. By visually examining the intersections and connections between these lines, patterns and relationships within the data can be identified. Parallel coordinates enable the simultaneous exploration of multiple variables, facilitating the detection of correlations, trends, and outliers that may not be apparent in traditional scatter plots or other 2D visualizations. This technique is valuable for gaining insights into complex datasets and supporting data-driven decision-making processes in various domains, including finance, genetics, and environmental sciences.

2.4.4 Task 3: Pair Plot

The earthquake dataset is loaded using the pandas library, and specific columns of interest are selected for further analysis. The selected columns include magnitude, cdi, mmi, depth, sig, and nst. A pairplot is a type of scatterplot matrix that allows us to visualize the relationships between multiple variables simultaneously. In this case, the pairplot displays the pairwise relationships between the selected columns of interest. Each scatterplot in the matrix represents the relationship between two variables, and the diagonal of the matrix displays the distribution of each individual variable.

This visualization is particularly helpful in understanding the relationships and patterns between the different earthquake parameters. It allows us to identify potential correlations or trends between variables, which can provide insights into the nature and behavior of earthquakes. For example, we can examine whether there is a correlation between the magnitude of an earthquake and its reported intensity (cdi and mmi), or if there is any relationship between the depth of an earthquake and its significance (sig) or number of stations reporting it (nst).

2.4.5 Task 4: Dynamic data Voronoi

This task will be visualised using a Voronoi map. Voronoi diagrams partition a plane into regions based on distance to points in a specific subset of the plane. For the Voronoi map generation, the 'magnitude' was set as the value, the 'title' as the split, and 'country' as the group. This setup implies that each Voronoi region would represent a unique earthquake even (split by 'title'), grouped by the country it affected, and the size of the region would be determined by the magnitude of the earthquake.

The voronoi maps will provide a spatial representation of the earthquake data, providing details on the magnitude and location. By visualising a voronoi map for each year the evolution of seismic events over the 2001-2022 period can be examined clearly. Overall, this visualization method offers a unique and insightful perspective into the global pattern of seismic activity over the past two decades.

The cluster size in the visualization will represent the total number of earthquakes experienced by each country. To facilitate monitoring the evolution over time, the clusters are color-coded. Additionally, in order to enhance the clarity of the visualization and focus on the most destructive earthquakes, only countries that have experienced earthquakes greater than magnitude 8.0 are displayed on the chart.

Furthermore, to provide additional information on each earthquake, such as magnitude and location, a hover-over feature is implemented. Users can simply hover their mouse cursor over a specific earthquake node to access this essential information. This interactive feature enhances the clarity and usability of the visualization, allowing users to explore specific earthquake details on demand.

2.4.6 Task 4.1

This task can be effectively visualized using a bar chart, where the x-axis represents the countries and the y-axis represents the number of earthquakes. Each country will have a corresponding bar, allowing for easy comparison of earthquake frequencies. To simplify the extraction of information, the values representing the number of earthquakes can be displayed on top of each bar.

To enhance the visualization and provide context, each bar can be color-coded in conjunction with an overall Voronoi map that contains earthquakes from all the years. The Voronoi map overlays the bar chart, providing a visual representation of the geographical distribution of earthquakes. This integration helps in better understanding the spatial relationship between countries and their respective earthquake frequencies.

2.4.7 Task 4.2

We can use a frequency-time graph to visualize the occurrence of earthquakes. The x-axis of the graph can be divided into quarterly intervals within a year to identify any significant spikes in earthquake activity during specific time periods. By combining this visualization with a voronoi map, we can gain a clear understanding of the progression and changes in earthquake activity over time.

3 Implementation

3.1 Implementation of Graph data

R studio was used to identify patterns and similarities among earthquakes based on their statistical properties. The implementation process consisted of calculating the distance matrix using the dist() function, specifically the euclidean distance metric, and enforce a hard threshold for creating edges between highly similar earthquakes.

An undirected graph was constructed using the 'igraph' package by adding edges based on the distance matrix and threshold, the result was exported as '.gml' file to local computer.

The graph analysis techniques were performed on Gephi to identify clusters of highly similar earthquakes using algorithms such as community detection or clustering. This approach can provide insights into the underlying mechanisms of earthquake occurrence.

3.1.1 Implementation of Overview graph

Gephi was used to visualize and analyze earthquake data. The GML file was opened; which contained all the similar earthquakes. After importing the data, Gephi created a graph representation of the earthquake events, where nodes represent individual earthquakes and edges represent connections between similar earthquakes based on their impact metrics.

In order to group the most similar earthquakes,we use modularity as a community detection algorithm. Modularity is a suitable choice for this task because it measures the strength of the division of a network into communities, helping to identify clusters of nodes that are more densely connected to each other than to the rest of the network. Modularity was applied in Gephi, by clicking "run" on Modularity[5] in the "Statistics" panel. Furthermore, the resolution was adjusted to increase the communities in the graph.

To visually represent the earthquake communities, the nodes were color-coded based on their modularity class in Gephi. The process involved opening the "Appearance" panel and navigating to the "Nodes" tab. Then, the options for "ranking" and "palette" were used to change the node color. The "Modularity Class" was selected from the dropdown menu and applied to the nodes. This allowed for a clear and visual representation of the earthquake communities based on their modularity[6][4] class.

To represent the connectedness of each node, the node size was adjusted based on degree centrality in Gephi. Degree centrality measures the number of connections a node has within the network, indicating its level of connectivity. Nodes with higher degree centrality are considered more connected and can be interpreted as earthquakes with more similar events.

The node size was adjusted based on degree centrality by navigating to the "Nodes" tab in the "Appearance" panel. Next, the "Ranking" option was selected. From the dropdown menu, "Degree" was chosen as the parameter. The size of the nodes was then adjusted, with the minimum set to 5 and the maximum set to 50. This adjustment was made to emphasize the differences in degree centrality among the nodes, effectively representing their connectedness within the network.

3.1.2 Implementation of Red Cluster Subgraph

The figure 1) was constructed using both yEd and TULIP. Firstly, the overall graph for all similar earthquakes was opened in yEd. This required exporting the 1 in Gephi as GML. After opening the graph in yEd, the nodes which had 60 or more edges were isolated by utlisling "Select Elements" and setting a degree threshold of 60 to 500 in degree section, this option picked all nodes in the red cluster of 1; the community of interest in earthquake data.

Subgraph of red cluster nodes were created by following these steps:

1. Selected all the elements belonging to the red cluster.
2. Created a group from the selected elements.
3. Copied the group.
4. Pasted the group into a new graph file.

By following these steps, a sub-graph comprising only the nodes from the red cluster was created. This sub-graph allows for a focused analysis on the earthquake nodes within the red cluster that are most connected to each other. Then, the sub-graph was saved as a GML file.

The subgraph was opened in tulip. Where Fruchterman-Reingold[1] layout was applied to the graph. This force-directed technique minimises edge crossings and equally distributes nodes over the network to provide clear, attractive visualisations. This shows the red cluster's structure and interactions.

Edge bundling was also applied in tulip to improve visualisation. This method bundles edges with similar paths to help identify related nodes. Bundled edge thickness indicates node similarity. To increase readability and aesthetic appeal, the shape of edges were modified to "Cubic B-spline" curve from a "Polyline".

3.2 Implementation of Parallel Coordinates

The implementation began with importing the necessary libraries, namely 'pandas' and 'matplotlib.pyplot'. The high-dimensional CSV file was opened using the 'pd.readcsv()' function.

For the parallel coordinates plot, several key attributes were selected that provided insights into the intensity and impact of earthquakes. These attributes included magnitude, cdi (Community Internet Intensity Map), mmi (Modified Mercalli Intensity), tsunami, and sig (Significance). Each of these attributes contributed to our understanding of earthquake severity and the potential consequences.

The 'country' column serves as the basis for color-coding the lines in the parallel coordinates plot. Each line in the plot represents an earthquake event, while the color coding indicates different countries.

The exact process was repeated for the parallel coordinate plot, this time focusing on the averages of highly impacted countries. This allowed for the identification of statistical patterns across different countries.

3.3 Implementation of Pair plot

The implementation of the pair plot involves several steps to generate the visual representation of the relationships between variables in a dataset. First, the earthquake dataset is loaded into a pandas DataFrame using the 'readcsv()' function. The path to the CSV file is specified, and the data is imported, allowing for further manipulation and analysis.Next, specific columns of interest are selected from the DataFrame. These columns, which include magnitude, cdi, mmi, depth, sig, and nst, are chosen based on the information required for analysis.The seaborn library is then used to create a pair plot. The 'pairplot()' function is applied to the DataFrame, and the columns of interest are passed as an argument. This function generates a matrix of scatterplots, where each plot represents the relationship between two variables. The pair plot

provides a visual overview of the pairwise interactions between the selected variables, allowing for easy identification of trends, correlations, and outliers.Finally, the plot is displayed using the ‘show()‘ function from the matplotlib library. This function renders the generated pair plot and presents it on the screen.The pair plot implementation simplifies the process of analyzing and understanding relationships between multiple variables in a dataset. It provides an intuitive visual representation that aids in the exploration of data patterns and assists in making data-driven decisions. By leveraging the capabilities of pandas, matplotlib, and seaborn, the implementation efficiently handles data manipulation, plotting, and visualization tasks, making it easier for researchers and data analysts to extract valuable insights from their datasets.

3.4 Implementation of Voronoi Map

The initial step involved importing the CSV file for each respective year into Tableau. This was accomplished by navigating to the data source tab within Tableau and selecting the CSV file from the file explorer. It's important to note that the CSV files contain a number of key columns, including average x and average y, which were added to the columns and rows of the Tableau worksheet respectively.

Once the data was imported, the next step was to adjust the visual representation of the data. This was accomplished by changing the marks from their default setting of "automatic" to "polygon". The polygon mark type allows for the creation of a wide variety of custom shapes, in this case specifically, Voronoi polygons. This change in mark type is crucial to the overall visualization as it enables the creation of complex, irregular shapes based on the data.

The paths column from the CSV file was subsequently added to the path option in the marks card. This is a vital step as it is what allows the Voronoi maps polygons to appear on the visualization. Without this step, the polygons would not properly form based on the data, and the map would not be correctly visualized.

To enhance the visual differentiation between countries, the group column from the csv file was added to the color option within the marks card. By assigning different colors to each group, users are able to visually distinguish between different regions or countries.

In order to better represent the data spatially, the X column from the csv file was added to the columns section in Tableau. The visualization was then made dual-axis, and the axes were synchronized. This step ensures that both X values are plotted on the same scale, making the visualization more accurate and easier to interpret.

Lastly, the second x column (dual axis) was then changed to the mark type 'text'. This modification enables the value contained in the group and split to be displayed as text on the visualization. This is particularly useful when wanting to label the polygons with their respective group names or other relevant information.

3.5 Implementation of VA System

A tableau storyboard was constructed using the visualisation from all the tasks to effectively convey a visual story.The storyboard included:

1. Overall Graph of Similar Earthquakes
2. Detailed Graph of Similar Earthquakes
3. Parallel Coordinates Comparison of Impact Metrics
4. Pair plot of statistics
5. Evolution of Earthquakes Represented by the Voronoi Map

The integration of the overall graph was accomplished by copying the edge list from Gephi's data laboratory and constructing a new Excel file called "Tableau.csv". This file contained the source, target, ID, and weight of the graph's edges. Additionally, the Gephi file was exported as a GEXF file and subsequently opened in Microsoft Excel as an XML file. The XML file included the ID, labels, x and y positions, and colors of each node. This file was labeled "Nodes.csv".

From the "Nodes.csv" file, the ID, label, x, and y positions were copied and pasted into a new sheet of "Tableau.csv". This was done to facilitate the matching of nodes and edges based on their target names. Using the VLOOKUP function in Excel, the X and Y positions were matched to each edge by referencing the ID column. The X and Y positions are crucial to obtain the exact same layout result from fruchterman-Reingold method.

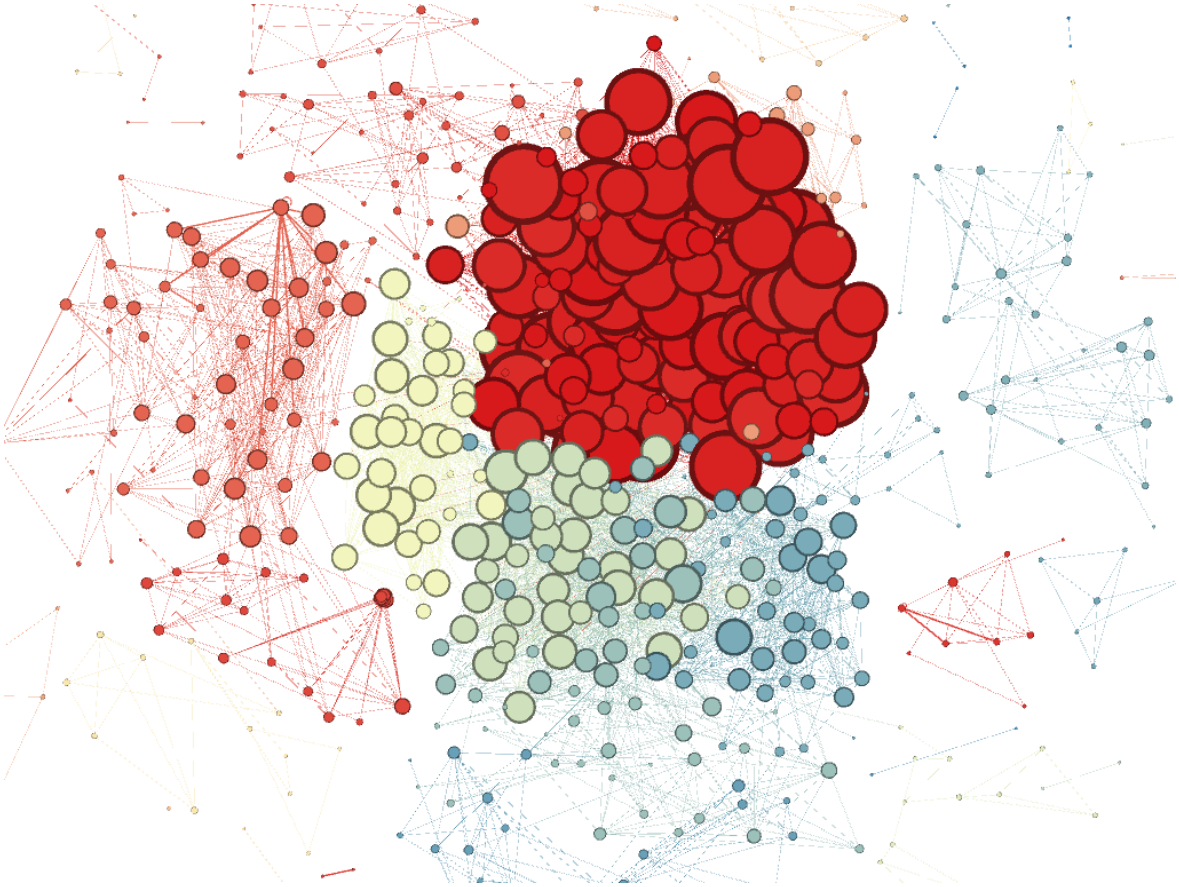
The next step involved duplicating the data in "Tableau.csv" to create lines connecting the dots represented by the X/Y coordinates. This was achieved by copying all the columns and pasting them beneath the final row. The original set of data was labeled as Base $= 1$, while the duplicated set was labeled as Base $= 2$, utilizing a new column for this purpose. Additionally, another column called "direction" was created by concatenating the target and source columns with an arrow in the middle. It is important to note that the direction specified here was solely for the purpose of Tableau visualization and does not transform the graph into a directed graph.

For the detailed graph of similar earthquakes, parallel coordinates plots, and pair plot, they were imported as images into a Tableau dashboard. The detailed graph was not implemented in Tableau similar to the overview graph because Tableau does not offer features like edge bundling and layout optimization. Additionally, the Voronoi map and its supporting tasks were already present in Tableau dashboards. All of these elements were integrated into one storyboard in Tableau, allowing for a cohesive presentation of the data.

4 Evaluation

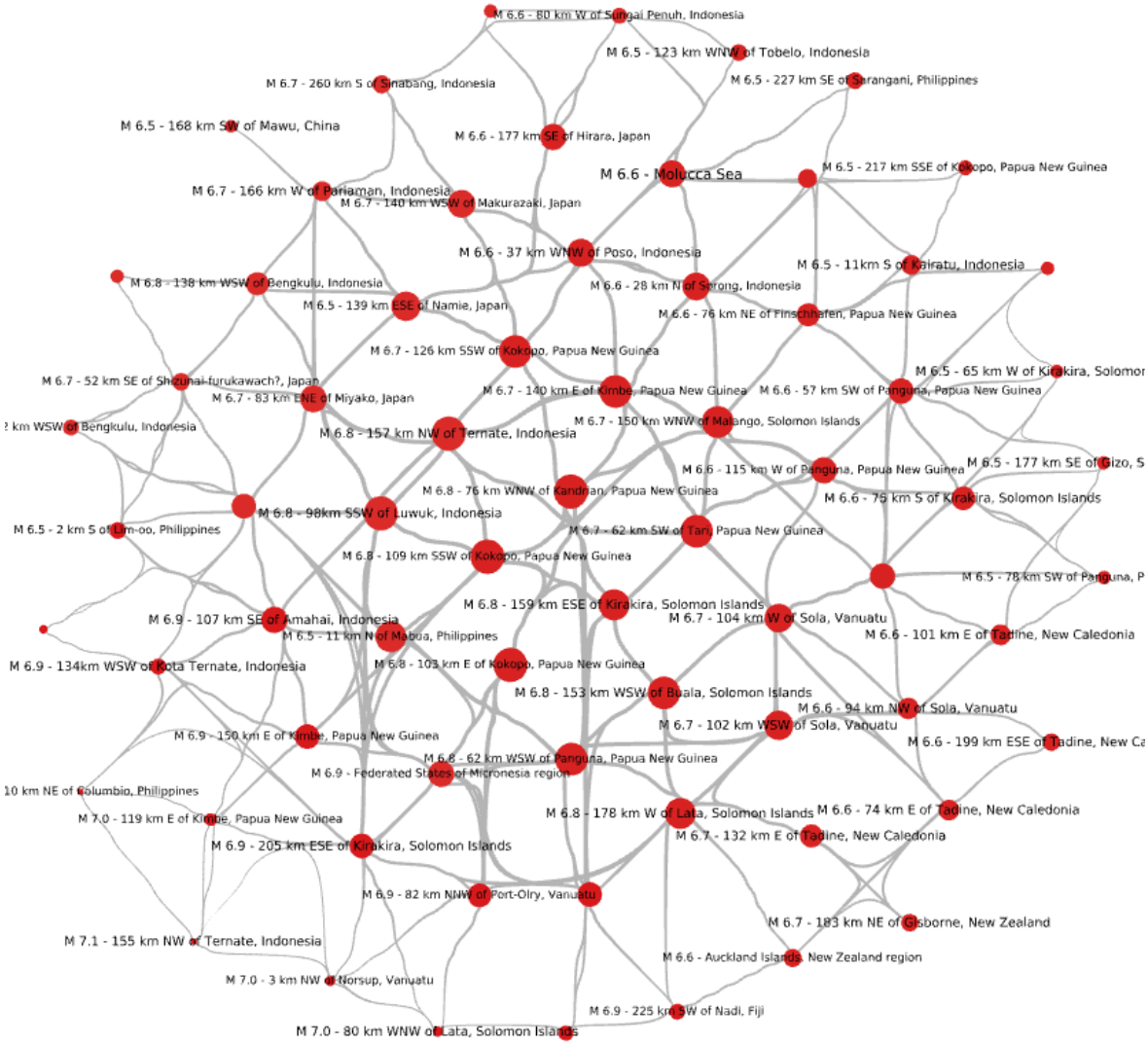
4.1 Results

4.1.1 Visualisation



(a) Overview graph of earthquakes with similar impact statistics

Figure 1: Overview of similar earthquakes



(b) Detailed zoom-in of the Red cluster from Figure 1a

Figure 1: Detailed graph of cluster

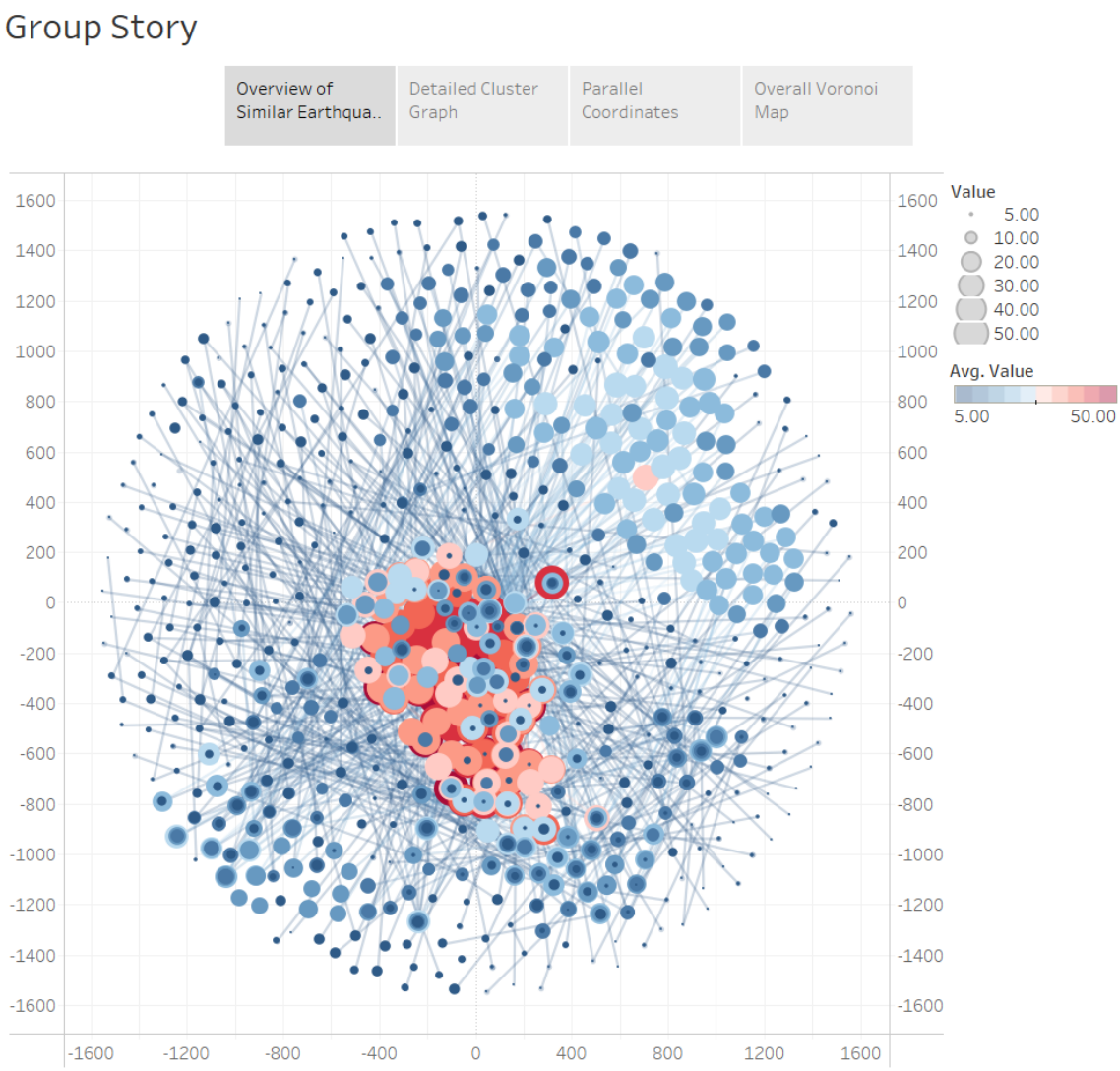


Figure 2: The depication of overview graph in tableau

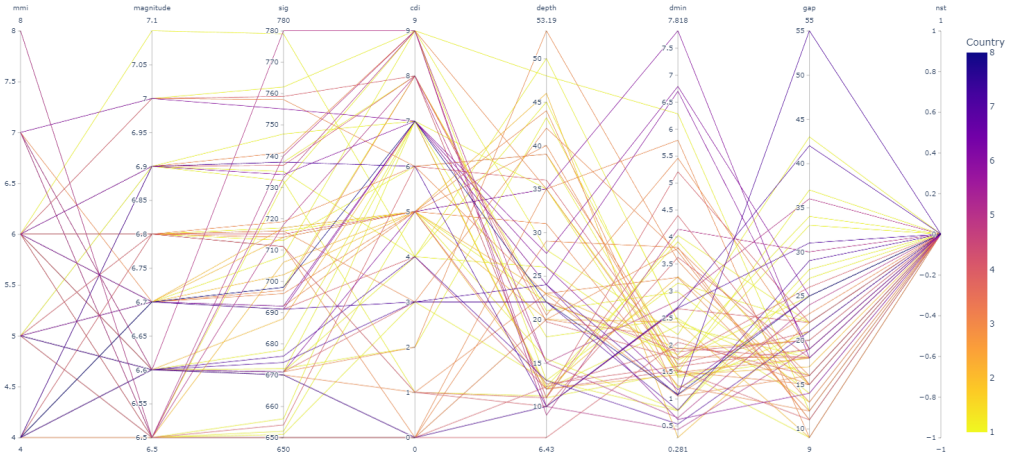


Figure 3: Parallel coordinates plot depicting the impact statistics of earthquakes

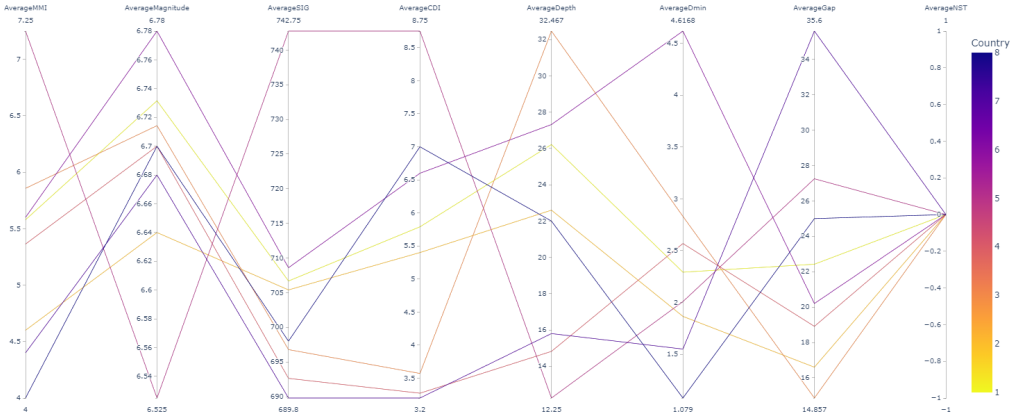


Figure 4: Parallel coordinates plot illustrating the impact statistics of highly impacted countries by earthquakes

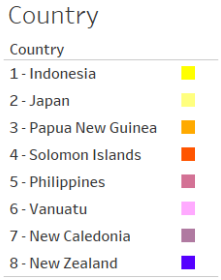


Figure 5: Parallel coordinates plot's legend

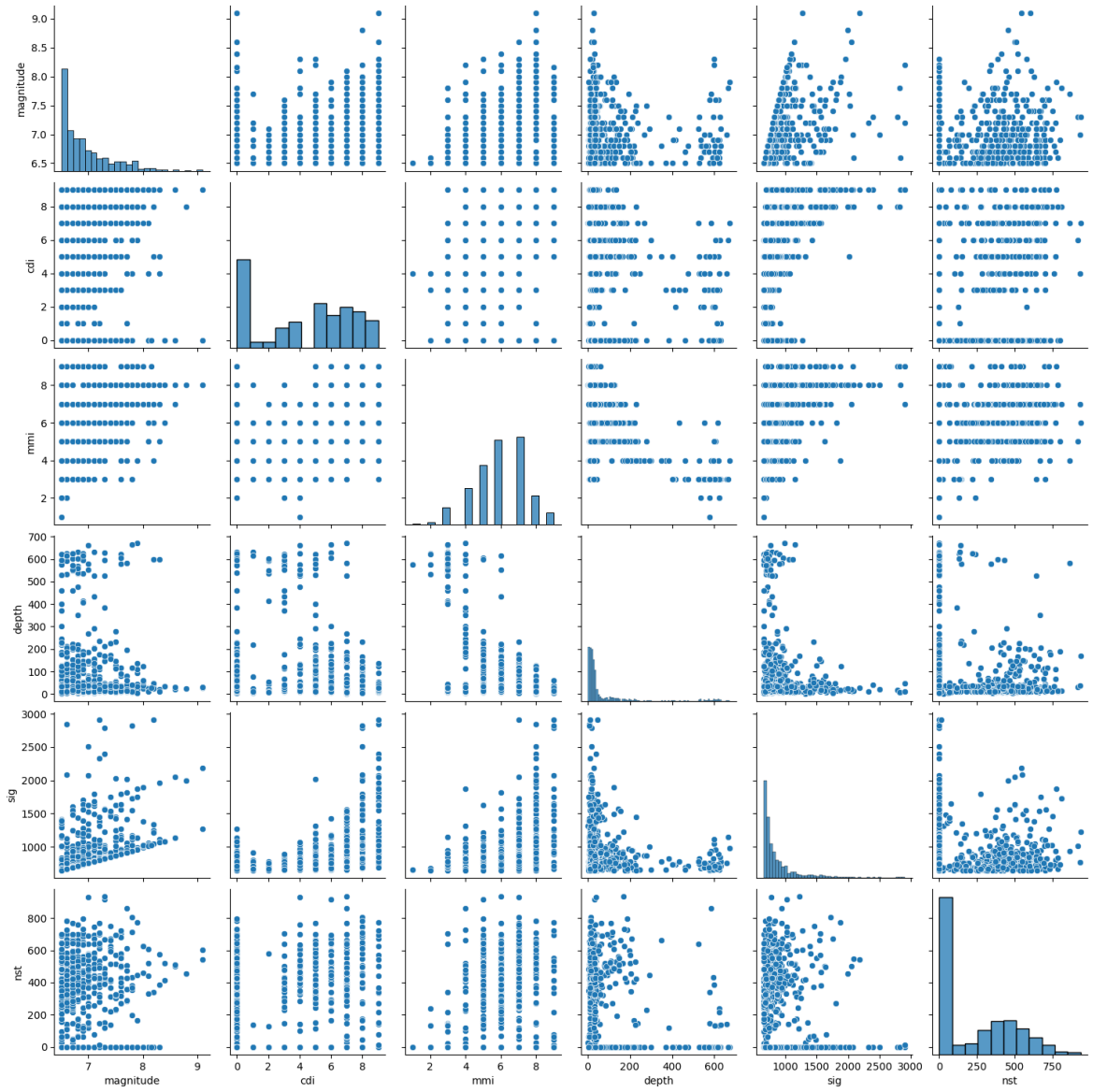


Figure 6: pair plot representing different stats of earthquake

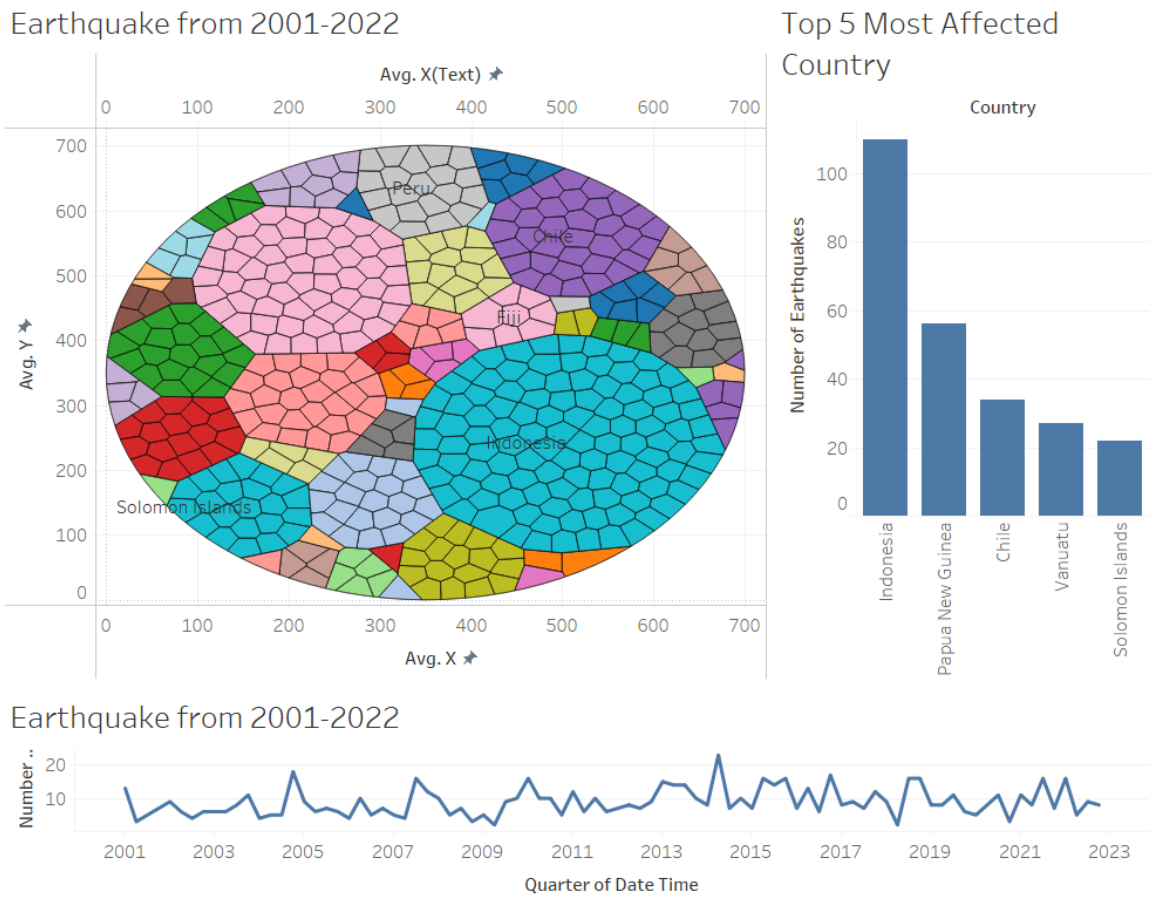


Figure 7: Dashboard of Dynamic Data

4.1.2 Storytelling/visual analysis

The Story Figure 1a provides an overview of earthquakes between the years 2001-2022, showcasing their similarities based on impact metrics. The visualization uses colors to classify the earthquakes, and the largest and most populous cluster is represented by the color red. To delve into the specifics of this cluster, we can refer to Figure 1b, which presents a detailed review of the red cluster observed in Figure 1a.

In Figure 1b, the red cluster stands out due to its prominent representation with the largest nodes and the highest number of nodes within its cluster. This suggests that the earthquakes in this cluster share significant similarities in terms of their impact metrics.

We identified that the most similar earthquakes are predominantly occurring in the same region, specifically Southeast Asia and Oceania. This region includes countries such as Indonesia, Japan, and Papua New Guinea. This finding is expected, as this particular region is known as the Ring of Fire, which is characterized by a high frequency of earthquakes.

This is followed up with figure 6 and figure 3 which display the overview of relationships between the

impact metrics and then detailed patterns of each highly similar earthquakes, since we are interested what makes the metrics so similar. Upon examining the parallel coordinates plot in Figure 3, it becomes evident that there is a correlation between magnitude and SIG (Significance). Each specific magnitude value is associated with a range of SIG values, indicating that the two variables are not independent of each other.

Furthermore, from 4 we can extract the metric pattern of each country specifically. It is clear from 4 that the Philippines (country specified by 5) stands out from the general trend observed in other countries. Unlike the majority of countries, the Philippines experiences relatively low-magnitude earthquakes on average. However, what sets it apart is the exceptionally high impact of these earthquakes, as indicated by measures such as CDI, MMI, and SIG.

While other countries may have higher-magnitude earthquakes, the Philippines demonstrates a unique pattern where even lower-magnitude earthquakes can result in significant impacts. This suggests that the vulnerability of the Philippines to seismic events is influenced by factors beyond just the magnitude of the earthquakes. The high impact indicated by CDI, MMI, and SIG likely reflects the population density, infrastructure quality, and overall preparedness of the country to handle seismic events.

Upon careful observation of the animation of the Voronoi maps, several significant findings emerge. It becomes apparent that the countries with the largest number of earthquakes are Indonesia, Papua New Guinea, Chile, Vanuatu, and Solomon Islands. Importantly, all of these countries are located within the renowned "Ring of Fire" region, an area characterized by intense seismic and volcanic activity along the boundaries of tectonic plates.

Furthermore, the analysis reveals that between the years 2011 and 2015, Papua New Guinea surpasses Indonesia in terms of the number of earthquakes recorded. However, for all other years, Indonesia consistently emerges as the largest contributor to seismic events among the countries considered. The unexpected dominance of Papua New Guinea in terms of earthquake occurrences for that particular time period highlights the dynamic nature of seismic activity within the Ring of Fire. It suggests that geological factors and processes affecting earthquake generation may undergo temporary shifts, causing localized variations in seismicity patterns.

Analysis The implementation of the pair plot allows us to generate scatter plots which allow us to establish the relationship between different feature columns that have been defined in the dataset. This allows us to further develop a correlation about how the different stats are being influenced and which amongst them have the highest influence on one another.

Evaluation of the graph visualizations can be assessed in terms of edge crossings and stress. These factors impact the readability and interpretability of the graphs, providing insights into the effectiveness of the chosen layout algorithms and visualization.

For the overview graph, edge crossings and stress are not evaluated due to the large scale of the graph. With a high number of nodes and edges present, it is expected that edge crossings and stress would be significantly high. Assessing these factors in the overview graph may not provide meaningful insights, as the main purpose of this graph is to provide a broad understanding of the earthquake data and its communities.

On the other hand, the detailed graph demonstrates a low number of edge crossings at 952, which can be attributed to the effective use of the Fruchterman-Reingold layout. This force-directed layout algorithm minimizes edge crossings by evenly distributing nodes across the graph, thereby reducing the likelihood of edge intersections. As a result, the detailed graph offers a clearer representation of the underlying structure and relationships within the red cluster, making it easier to analyze and interpret.

Upon evaluating the calculated graph density at a value of 0.3898, it becomes evident that the current layout effectively reduces edge crossings to approximately one-third of what could have been. This implies that the layout successfully mitigates the complexity and clutter associated with edge intersections, resulting in a significantly improved visual representation of the graph. The density value serves as a quantitative measure of this improvement, indicating that the chosen layout reduces edge crossings to a considerable extent, enhancing the clarity and interpretability of the graph.

Furthermore, it is worth noting that the calculated completeness value of 0.85 indicates a high level of accuracy in labeling elements within the same cluster. This implies that the chosen layout successfully identifies and groups together nodes that exhibit similar characteristics or connections. The high completeness value suggests that the cluster labels assigned to the graph's elements align well with the inherent structure or relationships present in the data.

Additionally, the stress in the detailed graph is relatively low at 319.2392 due to the good spacing between nodes and edges. This further contributes to the ease of readability and interpretability of the visualization. By utilizing the Fruchterman Reingold layout and edge bundling techniques, the graph effectively displays the relationships between the most connected nodes in the red cluster while minimizing visual clutter. Various layouts were attempted for this graph, but they resulted in less interpretable visuals with increased edge crossings. The exception being the stress minimization layout which slightly reduced stress.

The Voronoi map animation satisfies the criteria for a good animation, including preserving a mental map. This is achieved through smart initialization approach, which builds the Voronoi map for each time period, ensuring that the shape remains consistent and countries with the same color indicate similarity. While it may not fully preserve a mental map, it captures local dynamics and facilitates understanding of the geographic relationships in the data.

The animation excels in two aspects: smoothness and memorability. By preprocessing each time period using the prior time slice, it effectively bridges the transitions between different time slices. This approach ensures that the animation flows seamlessly from one frame to the next, resulting in a smooth visual experience. Additionally, the use of prior time slices eliminates any memory loss by providing a continuous reference point for viewers to track changes and retain context throughout the animation. As a result, the animation maintains smooth transitions and enables viewers to perceive the evolution of data without experiencing any significant memory gaps.

4.1.3 Pros/Cons

Pros

- All countries which have similar statistics can be easily identified from 1b.

- The evolution of earthquake occurrences can be followed from the voronoi map.
- The patterns of each countries' impact metric in the parallel coordinates are interpretable.
- Edge bundling in 1b aids in making visualisation readable by reducing the number of edge crossings.
- the colour of each country in the animation of voronoi map doesn't change, making it easier to see the evolution of earthquake occurrence for each specific country.
- The number of edge crossing and stress is low due to the effective usage of fruchterman-reingold layout.

Cons

- The storyboard is not interactive.
- The position of each country is not static during the animation of the voronoi map.
- In Figure 3, the lines of the parallel coordinates plot become cluttered, which hinders the identification of trends within each country.
- The overview graph could have been visualised using a matrix representation, as we know this visualisation improves readability by 30% for larger graphs.
- the voronoi animation doesnt have a fixed layout to explore the global dynamics.

4.2 Discussion: Summary and Limitations

In summary, our analysis reveals that the most similar earthquakes cluster within the Ring of Fire region. Moreover, this region consistently demonstrates a higher number of earthquakes compared to other regions worldwide. However, we also observed a surprising finding where Papua New Guinea experienced the highest number of earthquakes during a specific time period, indicating the variability of seismic activity within the Ring of Fire region over time.

A limitation of our current analytic system is its lack of interactivity, which prevents us from specifically tracking and examining the impact metrics of individual earthquakes. While we can observe and analyze earthquake patterns, trends, and general characteristics, we are unable to zoom in on a particular earthquake to explore its specific impact metrics in detail.

References

[1] David F Gleich. Spectral graph drawing. In *Handbook of Graph Drawing and Visualization*, pages 285–315. CRC Press, 2015.

[2] Tristan Guillevin. Beeswarm data generator, 2016.

[3] Tristan Guillevin. Voronoi treemap data generator, 2016.

[4] Y. Kaneko and et al. Anatomy of the 2011 mw 9.0 tohoku earthquake from inversion of seismic waveforms, gps, and leveling data. *Bulletin of the Seismological Society of America*, 101(3):1240–1255, 2011.

[5] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[6] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

5 Appendix

5.1 Weekly Meeting Record

5.1.1 March 30, 2023

All team members decided to meet at 12 pm at Koushik's residence.

- Choosing the appropriate dataset - Everyone
- Coming up with task questions for analysis - Suhrid
- Understanding Dataset parameters and providing short notes on the same - Reenal
- Classification of tasks - Sanjukta
- Data preprocessing - Koushik
- The next meeting is to be held on April 6.

5.1.2 April 6, 2023

- Deciding on the kinds and number of visualisations for the project - Everyone
- Developing Simple Visualizations – 2 each – Reenal and Sanjukta
- Necessary planning for Middle and Complex levels of visualizations - Suhrid and Koushik

5.1.3 April 20, 2023

- Simple visualizations – Bar chart, pie chart (Reenal), Line chart, Scatter plot (Sanjukta)

- Deciding on Simple Visualizations (bar chart, line chart) and necessary modifications – Suhrid and Koushik
- Report write-up on Simple Visualizations - Reenal and Sanjukta

5.1.4 April 27, 2023

- Developing Medium Complexity Visualizations – Everyone

5.1.5 May 4, 2023

All team members attended the meeting at 9.15 pm at ABS.

- Medium complexity viz - Network graph (Suhrid), geographic heatmap (Reenal and Sanjukta), Parallel coordinates (Koushik), Treemap (Suhrid)
- Deciding on medium Visualization (parallel coordinates) and improvising – Suhrid and Koushik.
- Report write-up on medium Complexity Visualizations - Koushik and Suhrid

5.1.6 May 11, 2023

All team members attended the meeting at 9.15 pm at ABS.

- Developing Complex level Visualizations – Suhrid and Koushik.
- Report write-up on voronoi map Visualization - Reenal
- Report write-up on Pair plot Visualization - Sanjukta
- Report write-up on Fruchterman-Reingold layout Visualization and adding more content to the report wherever necessary- Suhrid

5.1.7 May 18, 2023

- Deciding on Complex level Visualizations (Voronoi map, Pair plot viz, Fruchterman-Reingold layout) and improvising – Suhrid and Koushik
- Report write-up on Complex level Visualizations - Koushik and Suhrid
- More content to the was needed in report. Gathering some more info from the articles- Reenal and Sanjukta

5.1.8 May 22, 2023

- Storytelling or visual analysis – Suhrid and Koushik
- Bonus visualizations and additional content to the report - Koushik
- Final updates to the report – Everyone

5.2 Bonus Visualisation

5.2.1 Task

What is the relationship between the impact metrics of the most devastating earthquakes and the countries located in the Ring of Fire region?

5.2.2 Data Processing

To obtain a subset of earthquakes that occurred in the Ring of Fire region, we can apply a geographical filter to include only earthquakes within the designated area. Once we have this subset, we can establish arbitrary thresholds on impact metrics such as SIG, magnitude, MMI, CDI, and other relevant parameters to identify the most devastating earthquakes.

Additionally, we exclude any rows with null values in the selected impact metrics.

5.2.3 Analysis

To identify the countries in the analysis, we can use a color-coded scheme to represent each country in the Ring of Fire region. This means assigning a specific color to each country within the region to visually distinguish them on a map or chart.

In addition, to differentiate the size of each earthquake event, we can use varying circle sizes. The size of each circle will be proportional to the magnitude or impact of the earthquake. Larger circles can represent more devastating or significant earthquakes, while smaller circles can represent less severe events.

5.2.4 Visualisation

Beeswarm plots are well-suited for visualizing impact metrics because they represent individual data points without overlap, incorporate density estimation, and facilitate easy comparison. This allows for a clear and comprehensive understanding of the distribution and characteristics of earthquake events in the Ring of Fire region.

5.2.5 Implementation

The implementation process for the beeswarm visualization involved the following steps:

1. A CSV file was obtained from an online Beeswarm generator [2]. This file provided the necessary X and Y coordinates required to implement the beeswarm plot in Tableau.
2. The obtained X and Y coordinates were added to the rows and columns in Tableau to define the positions of the data points for each earthquake event. This allowed for the precise placement of the data points on the beeswarm plot.
3. The size of each earthquake event was determined by its magnitude. The magnitude values were assigned to the size parameter, ensuring that the visual representation of each event accurately reflected its impact.

4. A consistent color scheme was assigned to all the countries in the Ring of Fire region. This color scheme ensured that the colors remained the same for all comparisons, allowing for easier visual comparison and analysis of the impact metrics across different earthquakes.

5.2.6 Results

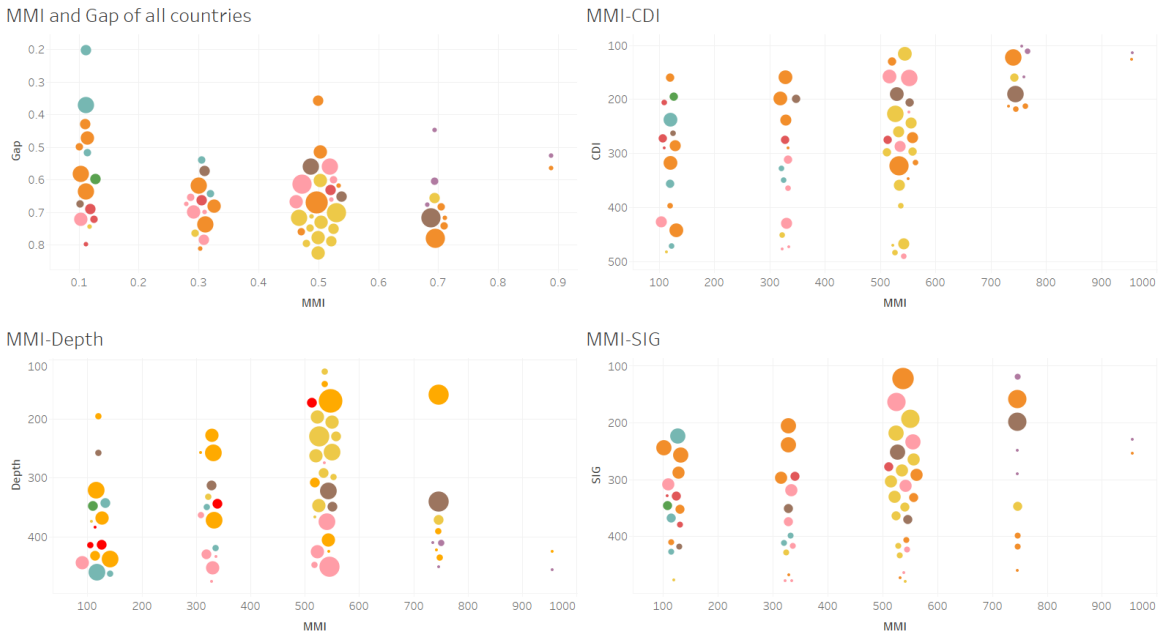


Figure 8: Beeswarm of MMI relationships with other statistics with magnitude as its size and country as its colour

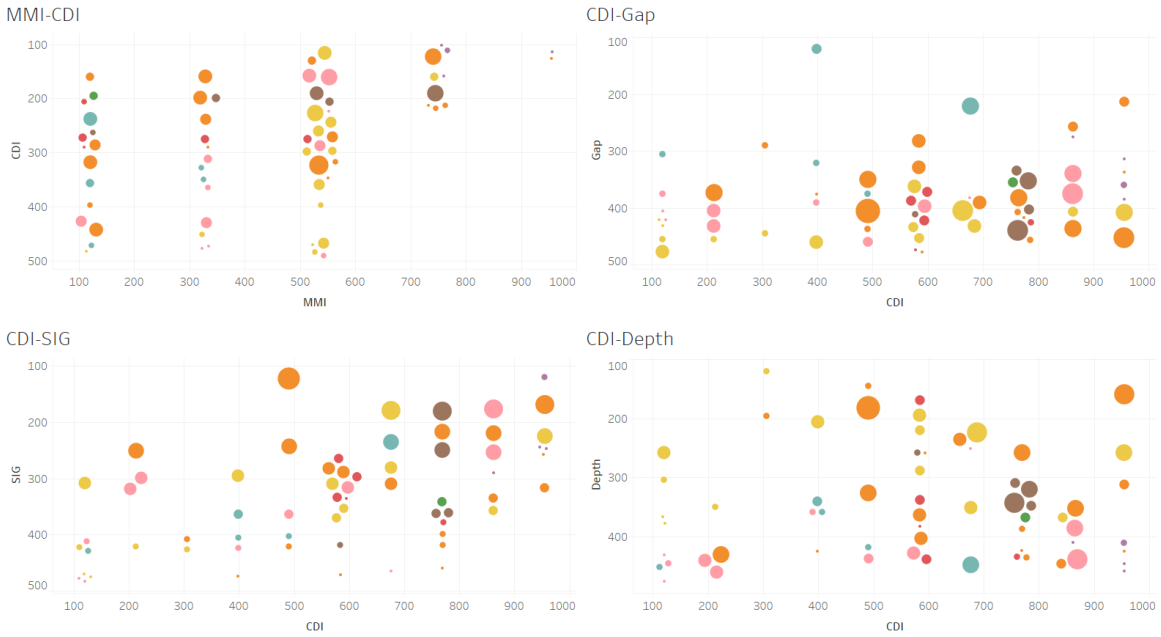


Figure 9: Beeswarm of CDI relationships with other statistics with magnitude as its size and countries as it colour

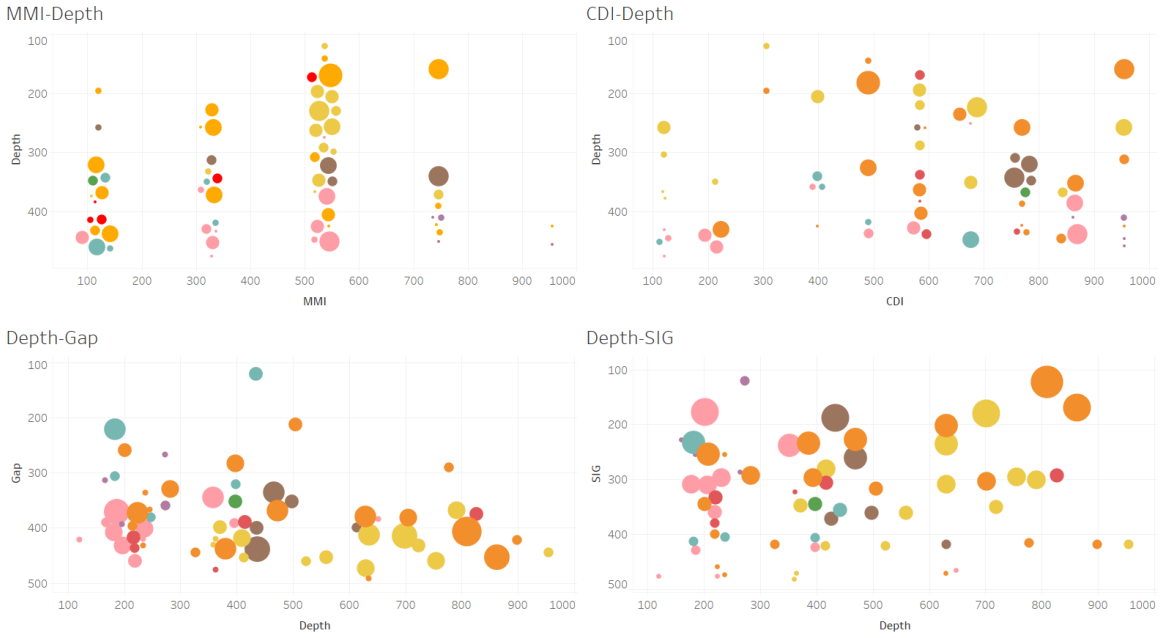


Figure 10: Beeswarm of Depth relationships with other statistics with magnitude as its size and country as its colour

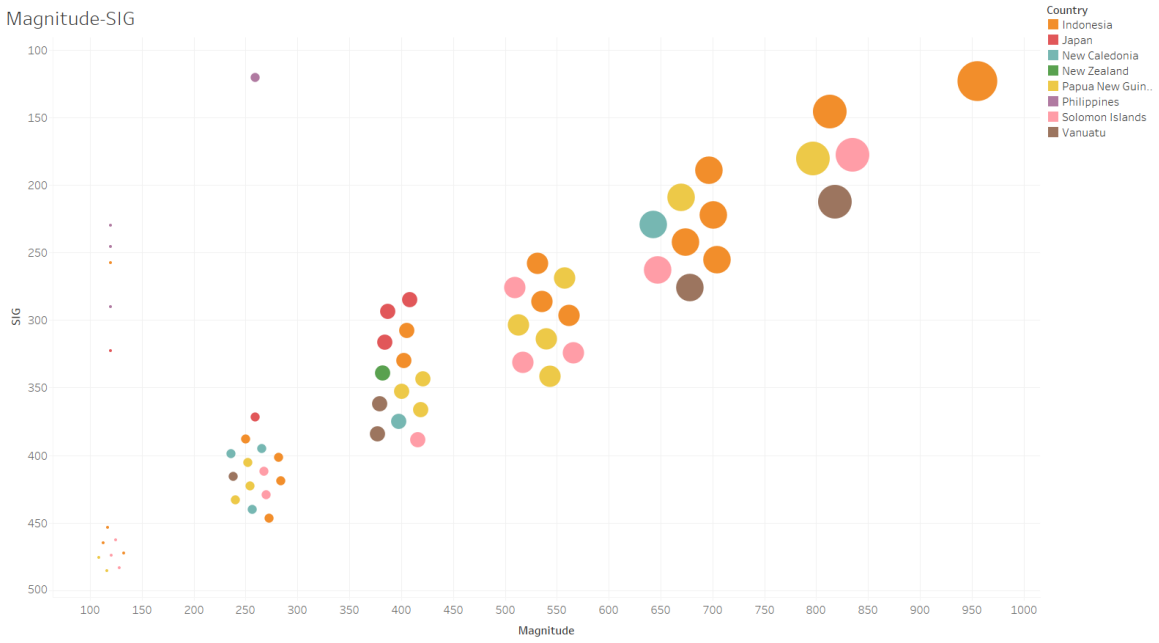


Figure 11: From other figures we see that magnitude and SIG are correlated which is true

5.2.7 Evaluation

The results of the beeswarm visualization indicate that there is no inherent correlation between most impact metrics, except for SIG and magnitude, which show a positive correlation as we can observe from figure 11.

The aim of this visualization is to identify the relationship between impact metrics of highly impacted countries. This visualization successfully accomplishes this task by effectively representing the impact metrics and allowing for a comprehensive analysis of their relationships.

When comparing a discrete metric with a continuous metric, the visualization appears clear, and the patterns can be easily discerned. However, when comparing two continuous metrics, the visualization tends to become messy and less interpretable.