Detecting Geo-Spatial Faultlines across Nepal

1st Hemang Seth
Contributor 1
IIIT-Bangalore
Hemang.Seth@iiitb.ac.in

2nd Tanish Pathania
Contributor 2
IIIT-Bangalore
Tanish.Pathania@iiitb.ac.in

3rd Vasu Aggarwal
Contributor 3
IIIT-Bangalore
Vasu.Aggarwal@iiitb.ac.in

I. Visual Analytics Workflow

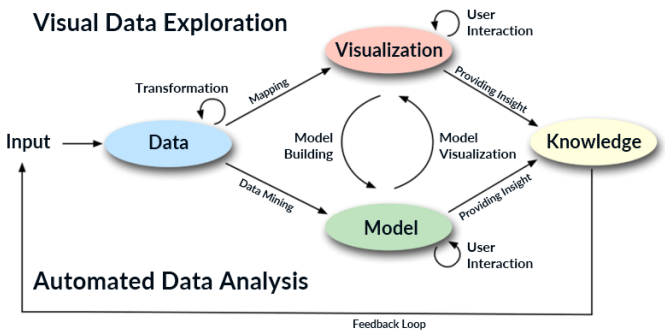


Fig. 1: Visual Analytics Workflow and Processes [1]

We start by summarizing **Task 1**, focusing on the visualization methods used and the inferences drawn. This marks the first iteration of our workflow, providing a foundation for further improvements. Then we expand our knowledge base by providing accurate analysis of the events that followed.

Key areas for refinement are identified based on insights gained, and these are incorporated into subsequent iterations of the workflow. The approach shown in **Figure 1**, inspired by Ivan Bozov [1], serves as a guide for structuring and optimizing our methods effectively.

II. Task-I-Seismic Hazard Assessment

A. *Prior Visualization and Inferences Taken From Previous Work*

From previous works, we thoroughly analyzed the impact of seismic **Fig 2** events in terms of fatalities and injuries. It was observed that economically disadvantaged populations were disproportionately affected, with a higher concentration of casualties in their regions. Regarding risk assessment, our findings revealed that measures to mitigate seismic risks and provide assistance to affected populations were minimal, underscoring a significant gap in equitable disaster relief efforts. To build upon this, we decided to conduct a district-by-district analysis.

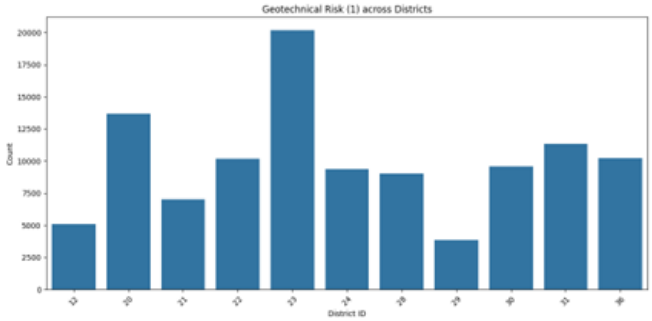


Fig. 2: Geotechnical risk accross the districts

B. *Pre-Processing the Dataset*

- **Data Cleaning and Integrity**:
  - Missing values were removed using the `.dropna()` function to maintain data accuracy.
  - Key numerical metrics such as **total fatalities and injuries** were identified for detailed analysis.
- **Categorization and Transformation**:
  - Continuous seismic data was categorized using the `numpy.histogram` function.
  - Data was grouped into bins and transformed into a **categorical format**, enabling **insightful visualization**.
- **Enhanced Visualization Features**:
  - Consistent color palettes were used to differentiate between fatalities and injuries.
  - Refined border styling improved visual clarity.

C. *Data Context Provided*

We worked with four primary datasets:

- **foundation_type.csv**: Contains information on the basic foundation of the house
- **condition_post_eq**: Contains conditions of the ground post earthquake so that it can be used later
- **land_surface_condition.csv**: Contains information on the land surface condition how it is so that it can be visualised for further purposes
- **ground_floor_type.csv**: Which type of construction prooves to be the most dangerous

These datasets were used to analyze earthquake impacts and geotechnical risks across various administrative levels.

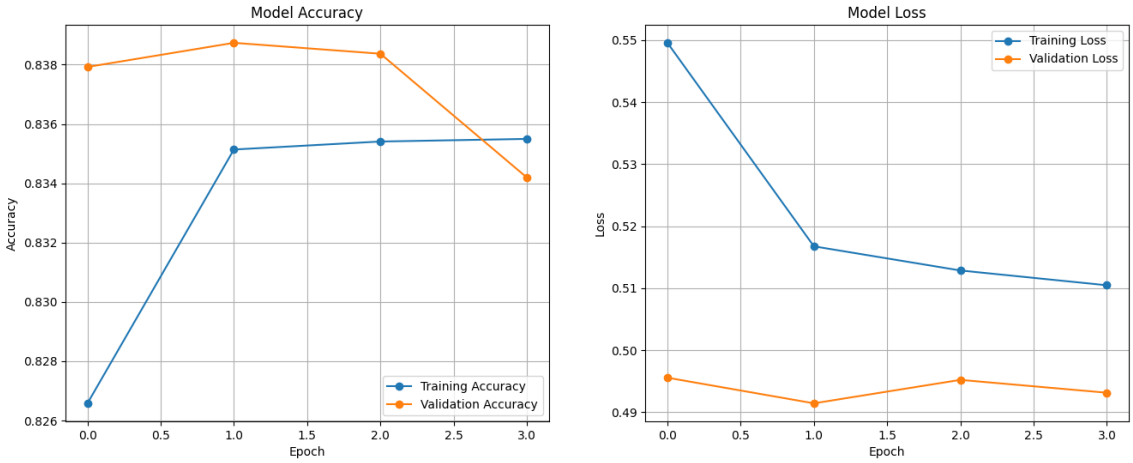


Fig. 3: Model accuracy when we ran it first time.

D. *Data Manipulation Techniques Used*

a) . Mapping and Initial Setup:

- Created mappings for `district_id` to `district_name` and `vdcmun_id` to `vdcmun_name`, ensuring duplicates were removed.
- Initialized and cleaned a dataframe with regional identifiers, sorted by `district_id` and `vdcmun_id`.

b) Earthquake Impact Aggregation:

- Extracted and aggregated data on deaths and injuries at district and VDC/municipality levels.
- Computed total fatalities, injuries, and cases, integrating these metrics into the main dataset.

c) Population and type of geographical conditions:

- Mapped population data from the demographics dataset to each district and VDC/municipality.
- The calculated what percentage of houses are in various kinds of different land types.

d) Geotechnical Risk Assessment:

- Aggregated geotechnical risks (e.g., land settlement, liquefaction, landslides, and floods) from building damage data.
- Added these risk metrics to the main dataframe for district and VDC/municipality levels.

e) Final Data Preparation:

- Consolidated the dataset to include regional identifiers, earthquake impact metrics, population counts, per capita data, and geotechnical risks for analysis.

This streamlined preprocessing pipeline provided a comprehensive structure for assessing seismic risks and vulnerabilities across administrative regions.

E. *Experiments with Machine Learning Models*

This section elaborates on the steps undertaken to prepare, train, and evaluate a machine learning model to classify seismic risk assessments. We describe the preprocessing techniques, the dataset features, and the evaluation results.

1) Data Format Given to Us:

- **Datasets:** The given datasets mainly included information on **Foundation type**, how it is made and what are the geological risk factors involved in the the ground levels. The dataset contained **Textual** and **Numerical** data in it making pre processing a bit difficult.
- **Data Cleaning:**
  - Removed duplicate rows to ensure unique entries in the dataset.
  - Addressed missing values by removing incomplete rows for critical fields.
- **Feature Selection:**
  - Selected features highly relevant to seismic risk using statistical correlation metrics.
  - Focused on structural and land-related attributes that significantly influence risk levels.

2) k-Nearest Neighbors (k-NN) Classifier:

- **Preprocessing for Model Training:**
  - Removed irrelevant columns such as identifiers (`building_id`) and purely descriptive fields.
  - Label-encoded categorical features for model compatibility.
  - Ensured a clean dataset by removing rows with missing target values.
- **k-Nearest Neighbors (k-NN):**
  - Configured a k-NN classifier with:
    - `metric: manhattan`, `n_neighbors: 9`, `weights: distance`.
  - Split the dataset into 80% training and 20% testing subsets for evaluation.
- **Evaluation:**
  - **Accuracy Score:** Measured the model's overall predictive accuracy.
  - **Classification Report:** Included metrics like precision,

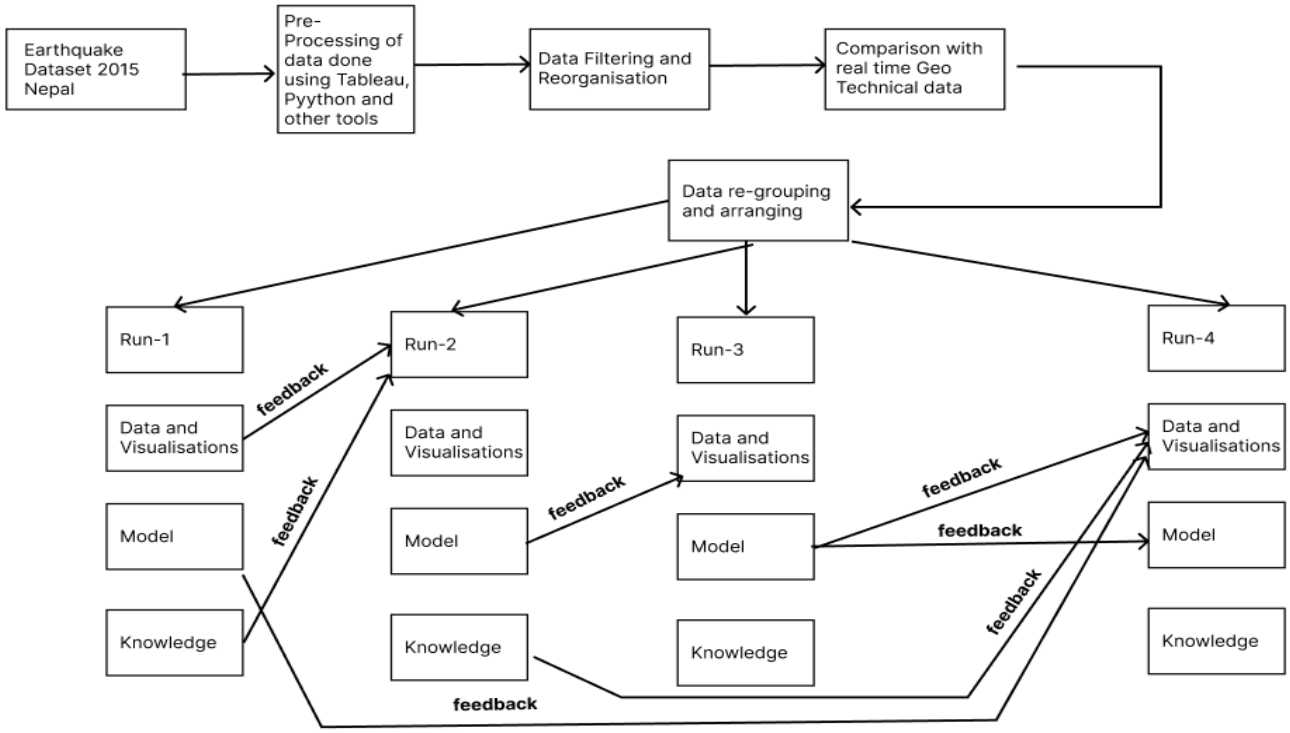


Fig. 4: Workflow diagram of the Nepal earthquake dataset analysis for Risk And assessment how step by step the entire analysis was done

recall, F1 score, and support for each risk category.

– **Plotting** The final plotting done in Run 2 and Run 4 were the main incorporation with the map

F. Describing the Visual Analysis

Refer the **Fig 4** for analysing the workflow of how I strutured and went step by step. Taking the feedbacks gained from every run

*1) **Run 1:** :*

(i) **Data:**

- Utilized datasets capturing district-level **risk factors** such as:
  - Land settlement, fault cracks, liquefaction
  - Landslides,rockfalls,floods,other risks
- Aggregated data to calculate the **total contributions** of each risk type across districts.

(ii) **Visualization:**

- **Doughnut Chart:** Represented risk proportions with segmented areas, using the `tab20` color palette for clarity. Refer **Fig 5**
- **Treemap:** Used rectangle sizes to reflect cumulative risk values, emphasizing the dominance of certain risks over others. Which has **Land flood** risk which has **Landslide** risk etc, Refer **Fig 6**

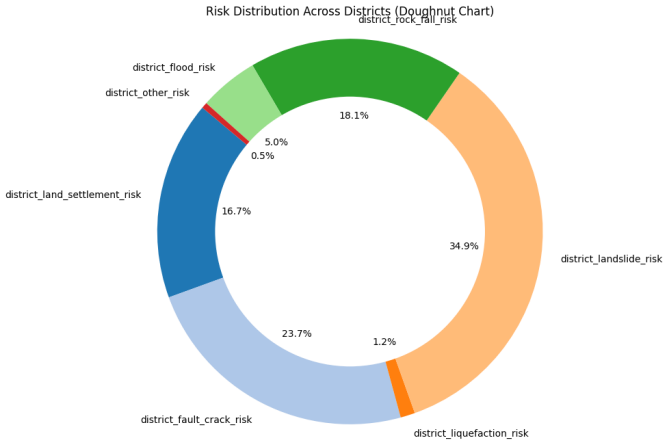


Fig. 5: Risk Distribution Across Districts (Doughnut Chart)

- **Radar Chart:** Used a radar chart to visualise the comparitive damage accross districts. This refers to the total damage done overall Refer **Fig 7**

(iii) **Knowledge:**

- Highlighted the relative **prevalence** of various risks across districts.
- Identified which **risks dominate** the overall seismic

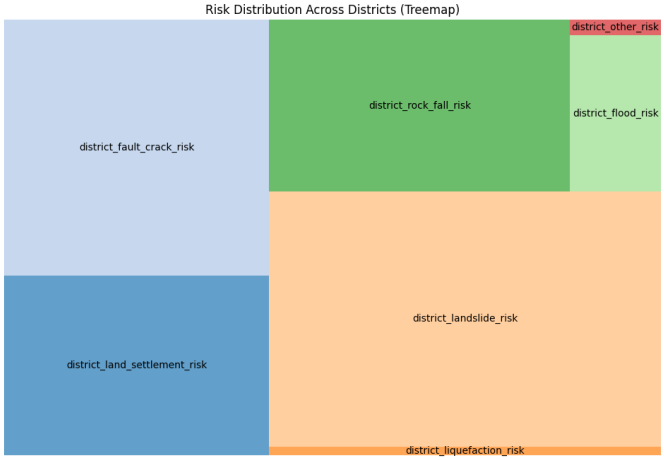


Fig. 6: Risk Distribution Across Districts (Treemap)

TABLE I: Table showing Color grades of the radar chart

Color Number	Color	Damage (property and sum total)
1	Blue	More than 200,000
2	Green	150,000-200,000
3	Forest Green	100,000-150,000
4	Yellow	Less than 100,000

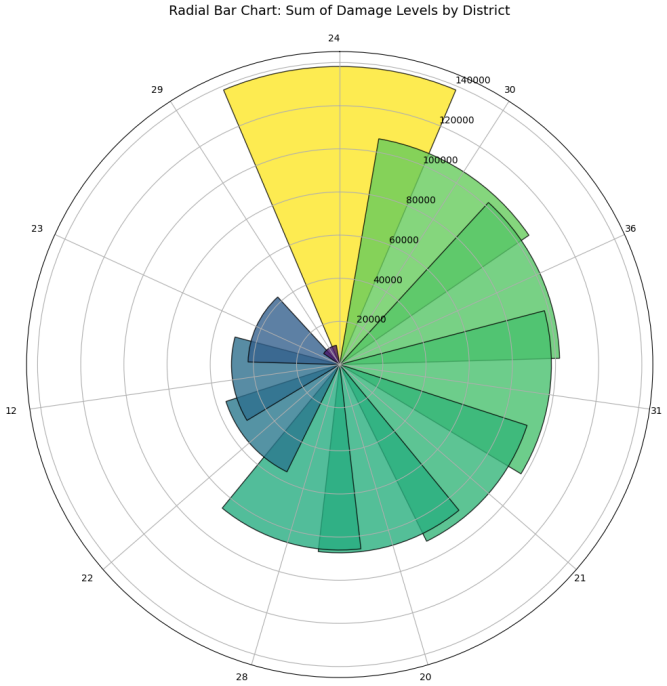


Fig. 7: Radar chart illustrating the distribution of risk factors across districts. This highlights the key areas of damage

risk profile.

- Analyzed the geographical distribution of these risks.

(iv) **Model:**

- No predictive model was employed for Run 1.
- Applied **categorical aggregation** and summation techniques.
- Prepared the data for effective visualization.

(v) **Feedback:** The Doughnut Chart and Treemap provided clear risk overviews but lacked district-specific granularity. The radar chart topped by providing the in depth analysis of the entire project. Future work should incorporate more granular data to identify district-level **hotspots**.

*2) **Run 2** :*

(i) **Data:**

- **Extracted district-level data** to examine the cumulative and relative contributions of various risks, emphasizing comparative analysis across districts.
- **Our focus** now main was to provide a comparitive district level analysis. To specifically pin point which district the Nepal govt. should target on

.

(ii) **Visualization:**

- **Stacked Bar Chart:** District-wise bars segmented by risk categories, enabling comparison of overall and relative risks across districts.As seen in **Fig 8**

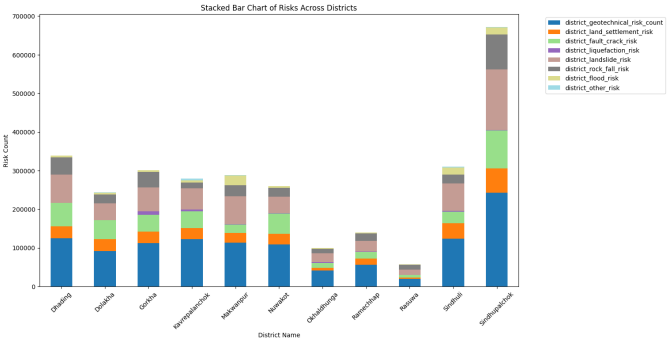


Fig. 8: Stacked Bar Chart of Risks Across Districts

- **Sunburst Chart:** Depicted hierarchical distributions of risks by district and type, providing an intuitive breakdown of seismic vulnerabilities. As seen in **Fig 9**

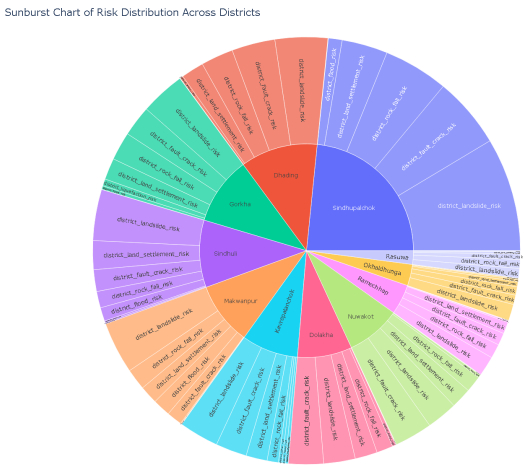


Fig. 9: Sunburst Chart of Risk Distribution Across Districts

- **Cluster Of District Damage:** District wise cluster plotted that shows which districts have more Geotechnical risks. As shown in **Fig. 10**

TABLE II: Cluster-Wise Damage Categorization on the Map

Cluster Number	Cluster Color	Damage Type
1	Blue	Less Severe
2	Yellow	Moderate Severe
3	Red	Very Severe

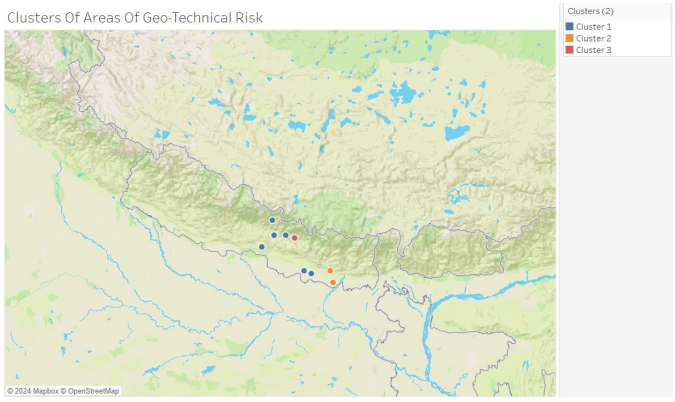


Fig. 10: Cluster Of Geo technical risk

- **Cluster Of District Damage plotted in a physical map:** This map included plotting in a physical map with the relief features in order to show that **hilly regions** have more damage (This sets up the story for further runs). As shown in **Fig. 11**

TABLE III: Cluster-Wise Damage Categorization on the Physical Map

Cluster Number	Cluster Color	Damage Type
1	Blue	Less Severe
2	Yellow	Moderate Severe
3	Red	Very Severe

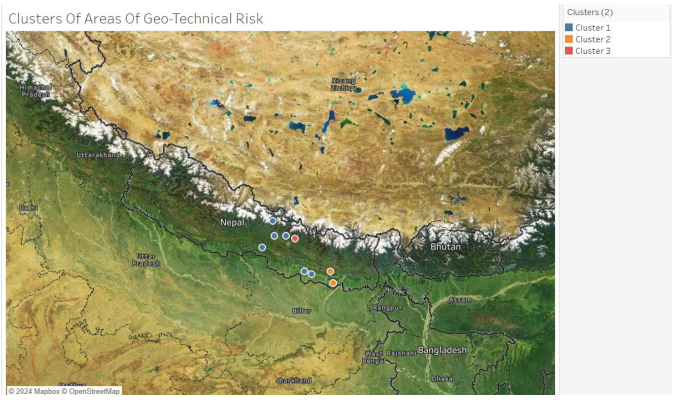


Fig. 11: Clusters of districts of geo technical risks

(iii) **Knowledge:**

- **Revealed disparities** in risk burdens among districts and identified districts with dominant vulnerabilities, such as land settlement and fault cracks.
- **Identified** the main kind of damage whether it was land liquefaction or fault cracks or what

(iv) **Model:** Implemented categorical breakdowns to structure data for comparison. Relative proportions were computed for each risk type within districts. The correlation parameters for each district was identified till we got the required results. The model gave us insights on **Which districts** have got risks which we plotted on Tableau

(v) **Feedback:** While visualizations effectively conveyed inter-district comparisons, they could benefit from incorporating temporal data to analyze how risk patterns evolve over time. Not only that they gave an insight on where which districts have more damage in clusters

*3) **Run 3** :*

(i) **Data:**

- **Clustered geotechnical risk** data based on factors such as liquefaction, landslides, and settlement risks.
- **Additionally, correlation metrics** were computed to identify interdependencies among risk factors.
- **This interpretion** of the data helped us analyse better

(ii) **Knowledge:**

- **Highlighted strong** correlations among specific **geotechnical risks** and grouped wards into clusters based on similar risk profiles. Provided a basis for targeted mitigation strategies.
- **Here we did** a district wise analysis of the entire districts using Radar plots and found places like **Makwanpur and Goorkha** have got maximum damage which matches the data of Geo technical risks of Nepal.

(iii) **Model:** Applied hierarchical clustering algorithms to group wards based on their geotechnical risk profiles. Correlation analysis identified key relationships among risk factors.

(iv) **Visualization:**

- **Sunburst Chart (Significant Correlations):** Represented positive and negative correlations among

risk factors, with segment sizes reflecting correlation strengths.As shown in **Fig. 12**

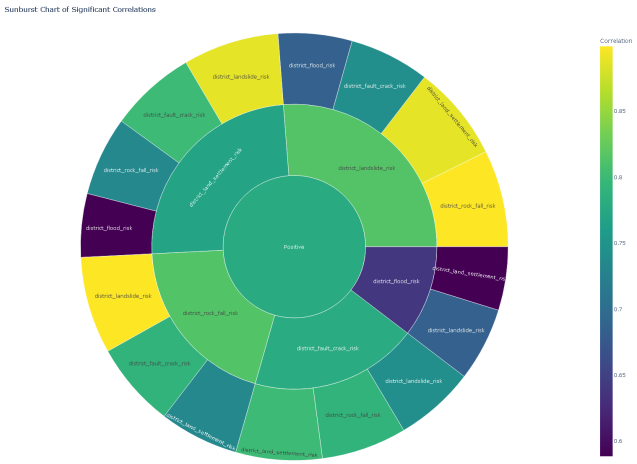


Fig. 12: Sunburst Chart of Significant Correlations

- **Pie Chart of damages done** Represented the pie chart of all the damages done in various areas of high geotechnical risks.As shown in **Fig. 13**

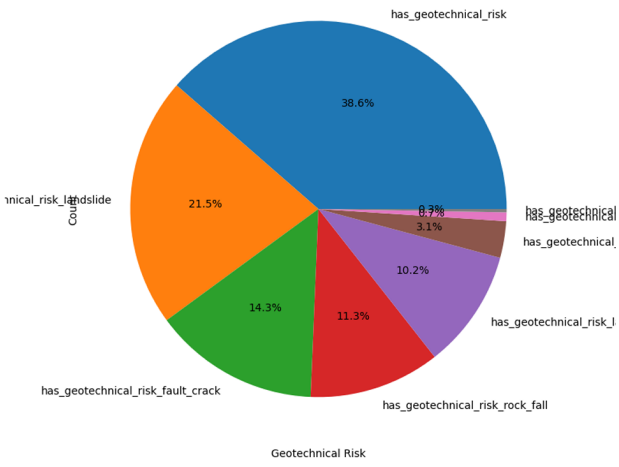


Fig. 13: Most Frequent Geotechnical Risks plotted Bar chart

- **Bar Chart of risks** Represented the pie chart of all the damages done in various areas of high geotechnical risks. **Landslides** have more this matches as it is a hilly region. As shown in **Fig. 14**

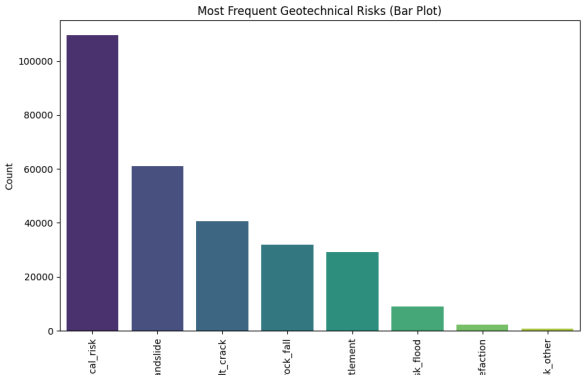


Fig. 14: Most Frequent Geotechnical Risks plotted Bar wise

- **Radar Charts:** Plotted Comparitive **Radar charts** of the top 6 damage districts. To visualise the damage caused in the main districts **MAKWANPUR** and **KAVREPALCHOK** came to be in the first places. Each axis of the radar chart corresponds to a specific risk type, and the enclosed polygon reflects the relative intensity of these risks for each district. As seen in **Fig. 15**

(v) **Feedback:** The correlation and clustering analyses provided actionable insights but could be enhanced by incorporating more advanced clustering techniques, such as k-means or DBSCAN, to validate results. Future iterations should include detailed interpretations of cluster characteristics. As we had obtained the **District wise data** Now our main focus shifted on plotting centers of High GEOTECHNICAL risks across the country.

*4) **Run 4:** :*

(i) **Data:**

- **Data was filtered** to include the six districts with the highest cumulative risk values.
- **Risk categories** analyzed included land settlement, fault cracks, liquefaction, landslides, rockfalls, floods, and other associated risks.
- **Based on the existing data** special clusters of risks were plotting in certain wards and municipalities of these districts.

(ii) **Knowledge:**

- **Focused on understanding district-level** vulnerabilities by comparing dominant risks. The analysis provided insights into the specific risk factors contributing most to the overall risk profile in high-risk districts.
- **On considering the top districts** from the previous visualisations we got an idea of where more damage is done. There we focussed our K means and got the needed results.

(iii) **Visualization:**

- **Plotting the Risk Distribution** On the final run we took the **Pie charts** from the previous runs and plotted the final Risk distribution to finally know which type

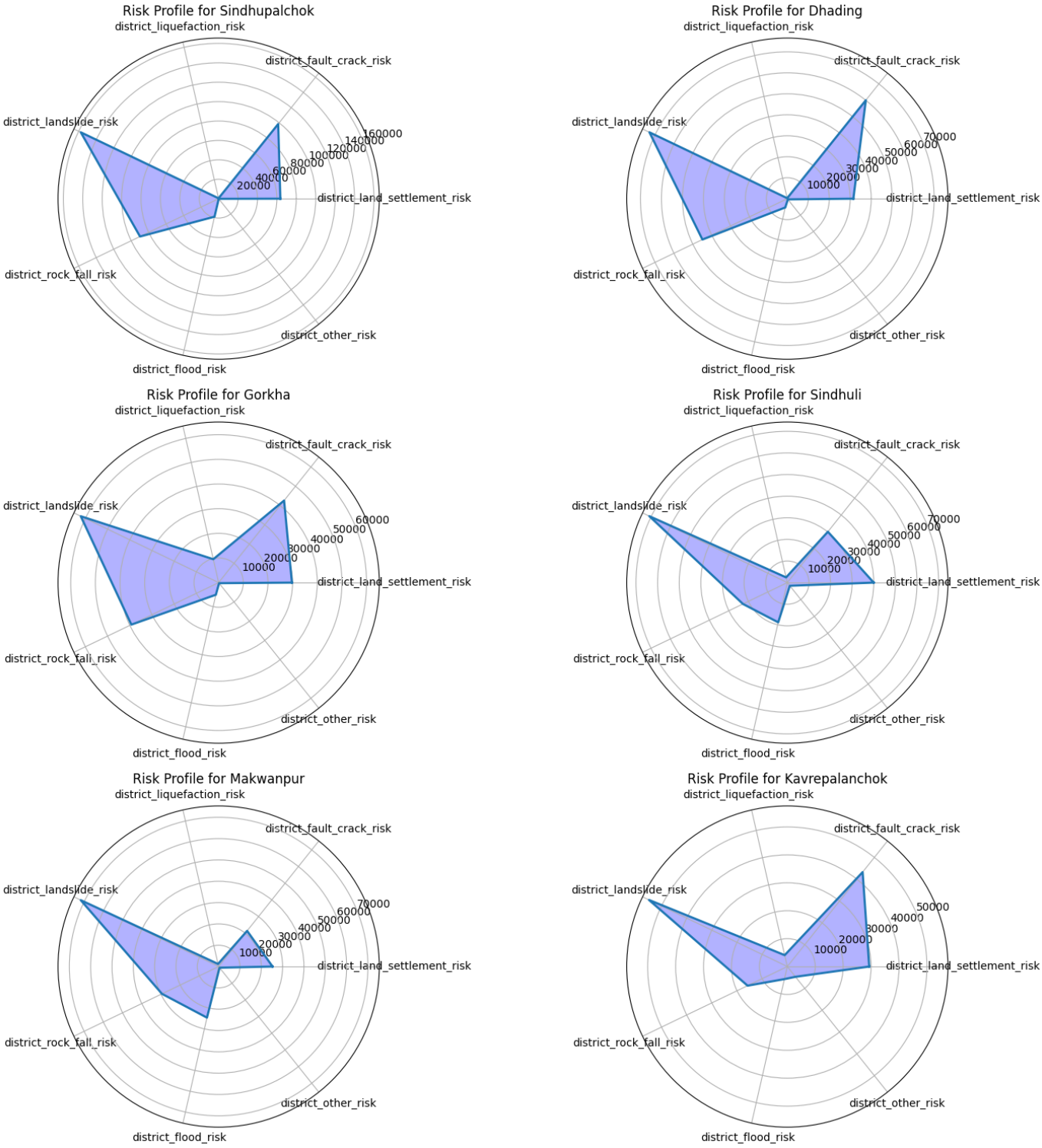


Fig. 15: Risk Profiles for the Top 6 Districts with the Highest Total Risks.

of risks are the most dangerous. **Refer Fig. 16**

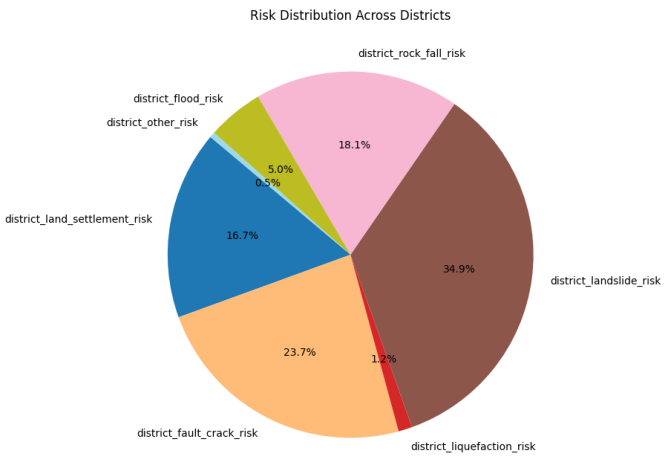


Fig. 16: Pie chart of risk visualisation

- **Plotting Of the risk clusters** We plotted the risk clusters on the map of Nepal to in those specified districts obtained from **Run 3** there we saw the regions of high moderate and severe technical risks **Refer Fig. 17**

TABLE IV: Cluster-Wise Damage Categorization ward wise

Cluster Number	Cluster Color	Damage Type
1	Blue	Less Severe
2	Yellow	Moderate Severe
3	Red	Very Severe

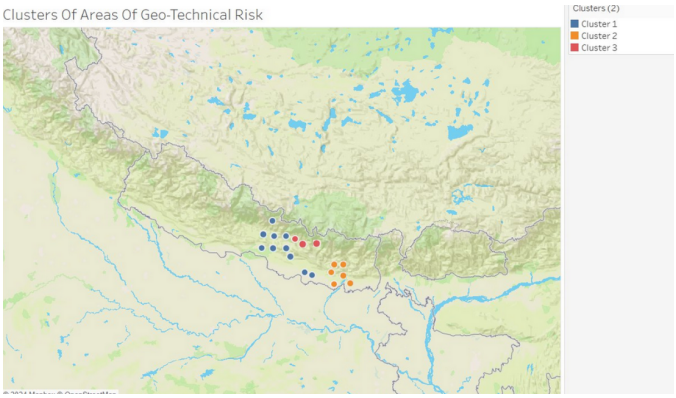


Fig. 17: View of Clusters Describing Geotechnical Risk of Various Wards

- **Plotting of risk clusters on the physical map** Plotting it next to the physical map gave us better analysis on why these clusters were concentrated in specific regions. This was due to **Varied Topography** and as it was in hilly regions soo the risk of **Landslides** was more . This matched with our previous runs. **Refer Fig. 18**

TABLE V: Cluster-Wise Damage Categorization ward wise plotted Physical wise

Cluster Number	Cluster Color	Damage Type
1	Blue	Less Severe
2	Yellow	Moderate Severe
3	Red	Very Severe

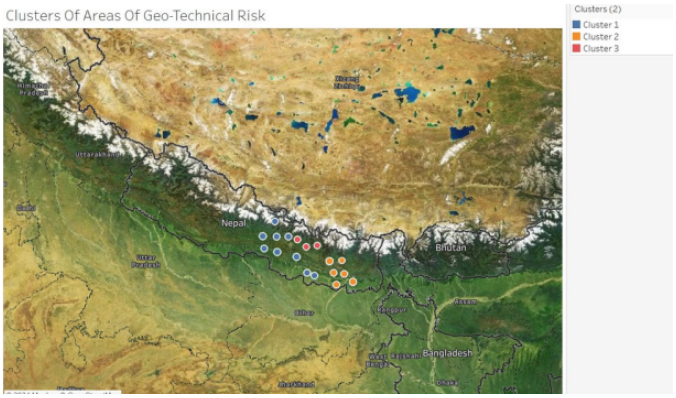


Fig. 18: Physical View of Clusters Describing Geotechnical Risk of Various Wards

- **Plotting the Risk Distribution In a node link diagram** On the final run we took the **Node Link** to establish the final correlation between the damage types (RUN -2). This gave us more analysis between the correlation **Refer Fig. 20**

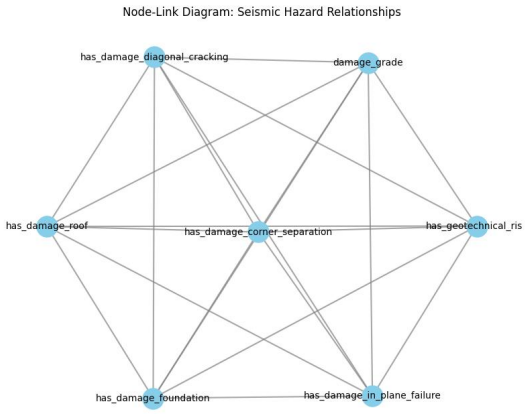


Fig. 19: Node Link showing us a strong correlation between the factors

(iv) **Model:**

- **Calculated cumulative risk** for each district and normalized data across various risk categories. Data was structured and then plotted district wise then ward wise for better analysis.

- **On plotting** the **Node link diagram** we saw the strong interelation between the risk factors finding the **MAIN RISK IS LANDSLIDES**. This matches as **Nepal is a hilly region**

(v) **Feedback:**

- **The radar charts** effectively highlighted district-specific vulnerabilities and dominant risks, offering a comprehensive visual summary for stakeholders. These were taken from the run 3 and then visualised in run 4 for further analysis. Hence finally correlating with all the tasks.
- **The Cluster** maps provided an in depth analysis of the entire plots
- **Lastly the node link** diagrams plotted the strong correlations

G. *Story of the Entire Task*

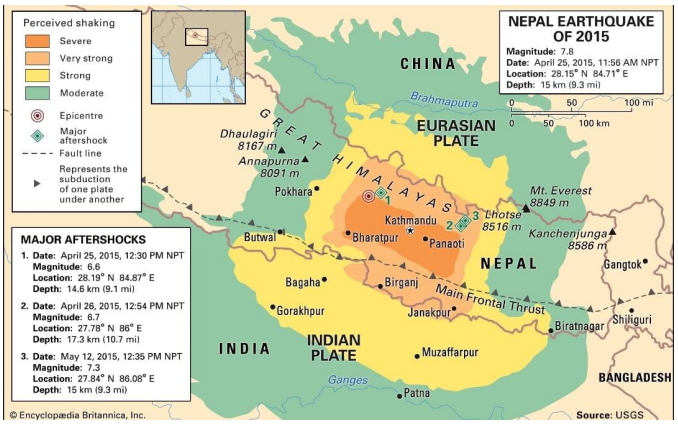


Fig. 20: Geo technical faultlines and risk [7]

1) Our Initial Approach: When I first started the task. I analysed the entire region's faultlines and the geo technical risks after the tremors of the 2015 Earthquake and how it had impacted the area. On refering the entire Fig 20 we found out the region of Goorkha and Kathmandu had been deeply affected. We took those regions and ran our ML Models on it . This helped us to identify the regions where more damages had been done. Not only that it also helped us in shaping our work as based

2) Taking Inferences from the Previous Feedbacks:

- **Run 1** the first and the basic run gave me an overall idea of the data and the visualisation. Helped me analyse the risky districts and the various types of damages.
- **Run 2** I took those and plotted them in the map using **Tableau** where i could easily plot them district wise.There I combined them with the various types of risks
- **Run 3** There on finding the 6 most risky districts I plotted them in the Radar Chart to find the various grades of damages. There I focussed my main research on the main districts and what type of damages had been done combining them in the ML Model and ten plotting the necessary visualisations.
- **Run 4** On combining all I finally plotted the needed graphs and everything. I plotted the combined graphs of ward wise risks along with the various types. THIS GAVE US THE COMBINED plots and figures

3) Final Conclusion For A-3 Task-1:

- I then came to the conclusion that **Nepal** is in more risk of landlsides than any other place especially the unstable belt of **Himalayas**. Not only that we found out the regions of **Makwanpur and Kavrepalchok** they are in higher risk than the regions where the Earthquake was actuallly happened.
- Lastly on finally plotting the needed clusters I found out that the risk factors could be easily bifurcated into several 3 main clusters where the most dangerious ones are the ones in the **Foothils of Himalays** this was prooved with the topography of NEPAL

III. Task-II-Death and Destruction Caused

A. *Prior Visualization and Inferences Taken From Previous Work*

From Previous Works, we thoroughly analyzed the impact of the damage in terms of deaths and injuries. It was observed that the poor were **disproportionately affected**, with a higher concentration of fatalities and injuries in their regions. Regarding the risk analysis, our findings revealed that measures to mitigate risks and assist those who suffered deaths and injuries were minimal, highlighting a gap in equitable relief efforts. We decided to go further providing a district by district analysis

B. *Important Dataset Analysis Columns Used*

For the analysis, we utilized two datasets:

1) **mapping.csv**: This file provided key relationships and mappings between different entities involved in the study.
2) **ward_vdcmun_district_name_mapping.csv**: This dataset contained the administrative hierarchies, including ward, VDC/municipality, and district names, which were crucial for regional analysis.

These datasets facilitated a detailed understanding of the geographical and administrative dimensions of the damage and its impact.

C. *Pre-Processing the Dataset*

- Collected three CSV files related to **death and destruction**.
- Mapped the data to individual *demographics* and *households* for collective analysis.
- Analyzed key columns, including `Household ID` and `District ID`, for data linking.
- Performed data processing using Python libraries:
  - `Pandas` for data manipulation and cleaning.

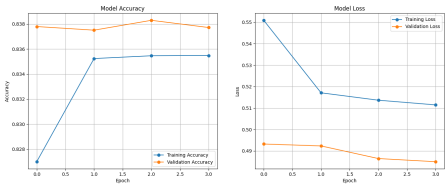


Fig. 21: Comparision Of Model Accuracy vs Model Loss Of Random Forest (Not considered due to their model loss

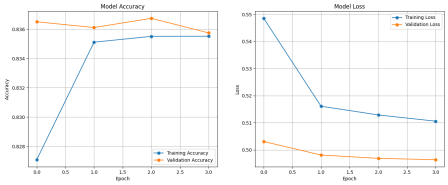


Fig. 22: Comparision Of Model Accuracy vs Model Loss Of Decision Trees (Not considered due to their low model accuracy

D. *Data Context Provided*

We worked with four primary datasets:

- **csv_building_damage_assessment.csv**: Contains information on building damage assessments.
- **ward_vdcmun_district_name_mapping.csv**: Maps districts and VDC/municipality names to their respective IDs.
- **csv_household_earthquake_impact.csv**: Records earthquake-related impacts such as deaths and injuries.
- **csv_individual_demographics.csv**: Provides demographic data at the individual level.

These datasets were used to analyze earthquake impacts and geotechnical risks across various administrative levels.

E. *Data Manipulation Techniques Used*

a) ***1. Mapping District and VDC/Municipality Names to IDs****:*

- `NumPy` for numerical operations and transformations.
- Created a new DataFrame named `clustering_df` from the three files for thorough analysis.
- Applied techniques such as merging, grouping, and pivoting to prepare the data for visualization.
- Utilized data visualization libraries:
  - `Matplotlib` for creating static plots and charts.
  - `Seaborn` for enhanced visual representation and statistical graphics.
- Conducted exploratory data analysis (EDA) to identify trends and patterns.
- Visualized *damage assessment* to communicate findings effectively.

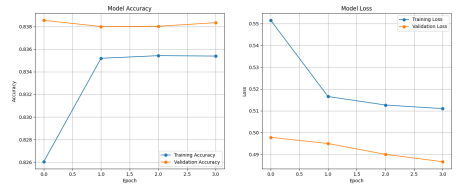


Fig. 23: Comparision Of Model Accuracy vs Model Loss Of KNN

- Created a dictionary mapping `district_id` to `district_name` and another mapping `vdcmun_id` to `vdcmun_name`.
- Identified and removed duplicates in the mapping file based on `vdcmun_id`.

b) ***2. Initial Dataframe Setup****:*

- Initialized a dataframe with `vdcmun_id`, `district_id`, `vdcmun_name`, and `district_name`.
- Removed duplicates and sorted data by `district_id` and `vdcmun_id`.

c) ***3. Aggregating Earthquake Impact Data****:*

- Extracted columns related to deaths and injuries from the earthquake impact dataset.
- Calculated aggregated metrics at both district and VDC/-municipality levels:
  - Total deaths and injuries.
  - Total cases where deaths or injuries occurred.
- Merged these metrics into the main dataframe.

d) ***4. Population Data Integration****:*

- Calculated population counts for each district and VDC/-municipality from the demographics dataset.
- Mapped population data into the main dataframe.

e) ***5. Calculating Per Capita Metrics****:*

- Computed deaths and injuries per 1,000 people for both districts and VDC/municipalities.

f) ***6. Assessing Geotechnical Risks****:*

- Aggregated geotechnical risks from the building damage dataset at both district and VDC/municipality levels.
- Risks included:
  - Land settlement, fault cracks, liquefaction, landslides, rockfalls, floods, and other risks.
- Added risk metrics to the main dataframe.

g) ***7. Final Data Preparation****:*

- Cleaned and organized the dataset to include:
  - Regional identifiers (districts and VDCs).
  - Aggregated earthquake impact metrics.
  - Population counts and per capita metrics.
  - Geotechnical risk assessments.

This preprocessing pipeline ensured a robust structure for analyzing earthquake impacts and vulnerabilities across administrative levels.

F. *Experiments with Machine Learning Models*

This section elaborates on the steps undertaken to prepare, train, and evaluate machine learning models to classify earthquake-induced building damages. We describe the preprocessing techniques, the models used, and the evaluation results.

1) Data Format Given to Us:

- **Datasets:**
  - The primary datasets used were:
    - `csv_building_damage_assessment.csv` – containing structural and geographical data about damaged buildings.
    - `ward_vdcmun_district_name_mapping.csv` – mapping administrative units to their corresponding names.
- **Data Cleaning:**
  - Removed duplicate rows and ensured unique mappings for administrative identifiers such as `vdcmun_id`.
  - Enriched the dataset by mapping `vdcmun_id` to `vdcmun_name` and `district_name`.
- **Feature Selection:**
  - Selected categorical features using **Cramér's** V to identify strong associations with the target variable (`damage_grade`).
  - Retained features with a Cramér's V value greater than 0.2, constructing a reduced feature set.

2) Models Used:

- **Prior Models Used** Models like Random Forest, Decision Trees were also considered but removed due to their less accuracy and more model loss as per **Fig.13 ,14**
- **Preprocessing for Model Training:**
  - Irrelevant columns such as `building_id`, `district_name`, and `ward_id` were dropped.
  - Rows with missing values in the target variable (`damage_grade`) were removed.
  - All categorical features were label-encoded for model compatibility.
- **k-Nearest Neighbors (k-NN):**
  - Configured a k-NN classifier with the following parameters:
    - `metric: manhattan`, `n_neighbors: 9`, `weights: distance`.
  - Trained on an 80-20 split of the dataset into training and testing subsets.
- **Evaluation:**
  - **Accuracy Score:** The overall predictive accuracy of the model.
  - **Classification Report:** Precision, recall, F1 score, and support metrics for each class.
  - **Confusion Matrix:** A graphical representation of the predictions and errors of the model.

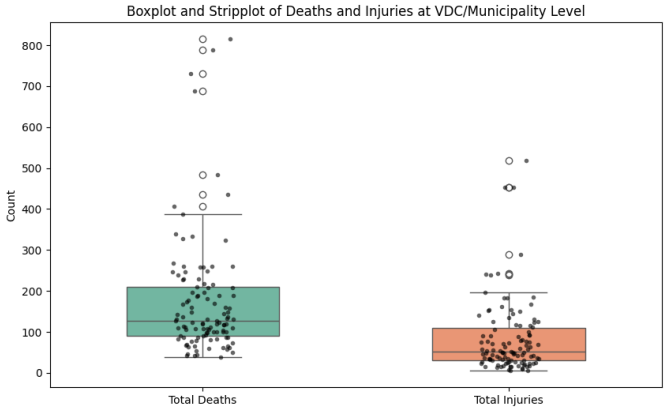


Fig. 24: Boxplot and Stripplot of Deaths and Injuries at VDC or Municipality Level

G. *Describing the Visual Analysis*

You can check our workflow in **Fig. 25** This step by step we approached on analysing the damages and inferences were gained. I went and took the feedbacks after each run and analysed and imporved my model

*1) **First Run**:*

(i) **Data**

The dataset provides insights into deaths and injuries across various **VDCs (Village Development Committees)** or Municipalities, along with correlations between risk factors and population-based metrics. This forms the basis for analyzing the spatial and statistical patterns of the data.

(ii) **Knowledge**

Through the analysis, we aim to identify patterns, trends, and relationships. For example, we can see in **fig : 24** the combination of **boxplots and stripplots** highlights the distribution and variability of deaths and injuries at the VDC or Municipality level. While the boxplot summarizes the central tendency and outliers, the stripplot offers a granular view of individual data points, uncovering clustering or patterns that may not be evident in aggregate views.

We also saw in cluster mapping a relatively more deaths in the hilly footfalls regions such as **District 23** and all as explained in **Fig: 28, 29** there we could get a good idea that the demography of Nepal makes it impervious to mountains thats why more deaths have occured over there

Furthermore, a heatmap (**Fig 26**) Illustrating the correlation between risk factors reveals critical interdependencies. For instance, strong positive or negative correlations provide actionable insights for prioritizing interventions.

(iii) **Model**

The model visualizes deaths and injuries per 1000 population by district using a line chart **Fig. 27** .This approach effectively compares the magnitudes across districts, emphasizing areas with disproportionately high

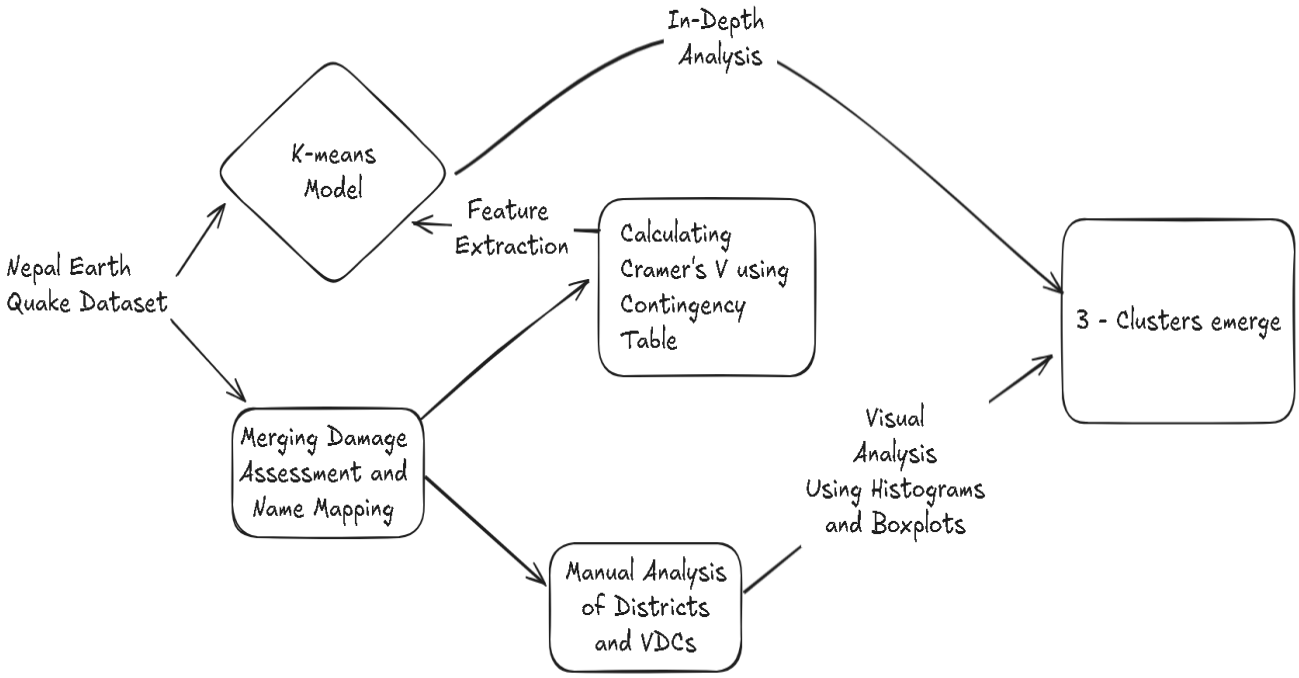


Fig. 25: Workflow diagram of the Nepal earthquake dataset analysis for Damage Analysis.

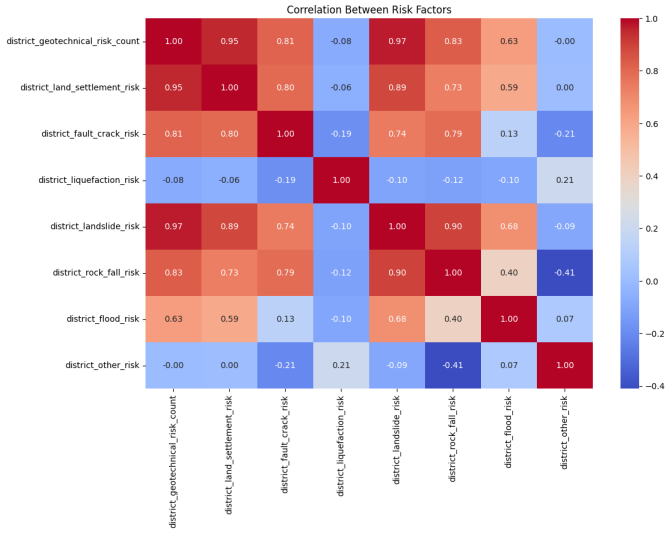


Fig. 26: Correlation Between Risk Factors

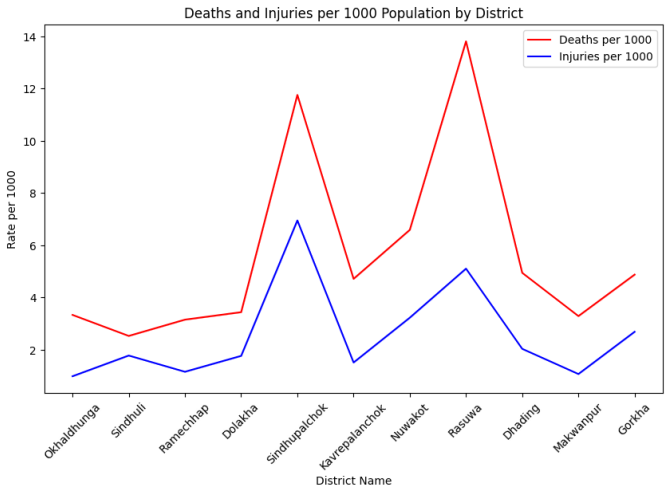


Fig. 27: Deaths and Injuries per 1000 Population by District

mortality rates. These insights guide resource allocation and highlight districts requiring urgent attention.

(iv) **Visualization**

Spatial visualizations enhance the story by plotting deaths and injuries as clusters on a map. **Fig. 28** and **Fig. 29** These clusters are overlaid with physical features of Nepal to contextualize the impact of geography on the observed trends. Together, these layers provide a comprehensive understanding of the spatial distribution of risk factors and outcomes.

The clustering visually categorizes districts based on the severity of deaths: - **Cluster 1** (Blue) highlights areas with relatively low mortality rates (fewer than 1000

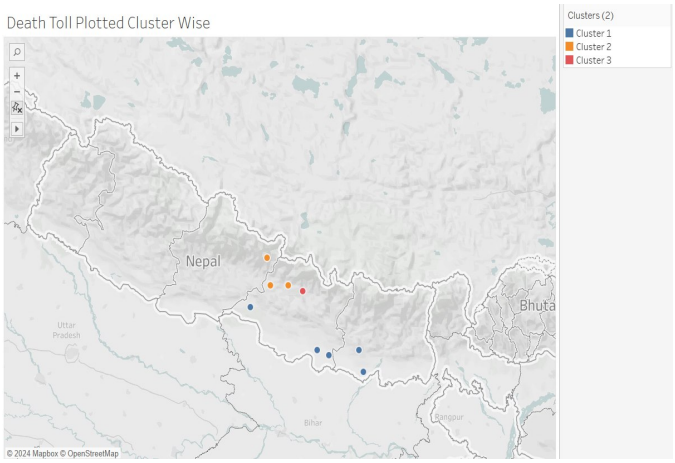


Fig. 28: Cluster-Wise Map Plot Showing Deaths in Clusters

deaths). - **Cluster 2** (Yellow) indicates regions with moderate severity, ranging from 1000 to 2000 deaths. - **Cluster 3** (Red) pinpoints critical areas with over 2000 deaths, suggesting the need for urgent interventions.
These clusters offer actionable insights into mortality trends and help prioritize resource allocation and response planning.

Fig. 29: Death Clusters per districts Overlaid with Physical Features of Nepal

(v) **Feedback**
The workflow demonstrates how integrating multiple visualization techniques can yield deeper insights. For future visualizations, ensure the following:

- Use combined approaches (e.g., heatmaps and line charts) to uncover correlations and trends.
- Overlay data with contextual spatial features for enhanced geographical understanding.
- Focus on storytelling by linking visualizations to actionable insights, prioritizing simplicity and clarity.

This feedback serves as a foundation for improving subsequent analyses, ensuring they remain impactful and aligned with decision-making needs.

*2) **Second Run:** :*

(i) **Inference Taken From First Run:** Initial exploration revealed districts with significant population-death discrepancies and highlighted clusters of high mortality rates. These findings guided the design of this second run, focusing on refining visualizations to enhance interpretability and uncover deeper insights.

(ii) **Data:** The dataset encompasses information on district populations, total deaths, and mortality rates. It includes key attributes such as population size, normalized deaths, deaths per 1000 population, and risk factor correlations. This data serves as the foundation for deriving meaningful insights and building the visual analytics workflow.

(iii) **Knowledge:** Insights from the first run informed a more nuanced analysis in this iteration. For example:

- Notable patterns in mortality rates per 1000 population across districts.
- Correlations between population sizes and mortality rates.
- Risk factor interdependencies identified through a correlation matrix.

These observations shaped the models and visualizations developed in this second run.

(iv) **Model:** Analytical techniques employed include:

- Normalization to standardize population and death metrics.
- Regression analysis to uncover population-death trends.
- Correlation analysis for exploring relationships among risk factors.

Together, these approaches provide a comprehensive understanding of the dataset.

(v) **Visualisation:** This iteration employs the following visual formats to construct a coherent narrative:

- **District Population vs Total Deaths (Circular View):** This polar representation **Fig 30** highlights normalized populations and deaths for each district, enabling pattern recognition.
- **Distribution of death and figures over districts:** This polar representation **Fig 31** highlights deaths and injuries accross the injuries in a combined curve for better analysis
- **Municipality-wise Deaths Per 1000 Population:** Mortality rates are illustrated using a pie chart for proportions and a scatter plot for detailed district-level comparisons. **Fig. 32**

TABLE VI: Cluster-Wise Death Categorization on the Physical Map

Cluster Number	Cluster Color	Deaths Range
1	Blue	Less than 1000 deaths
2	Yellow	1000 to 2000 deaths
3	Red	More than 2000 deaths

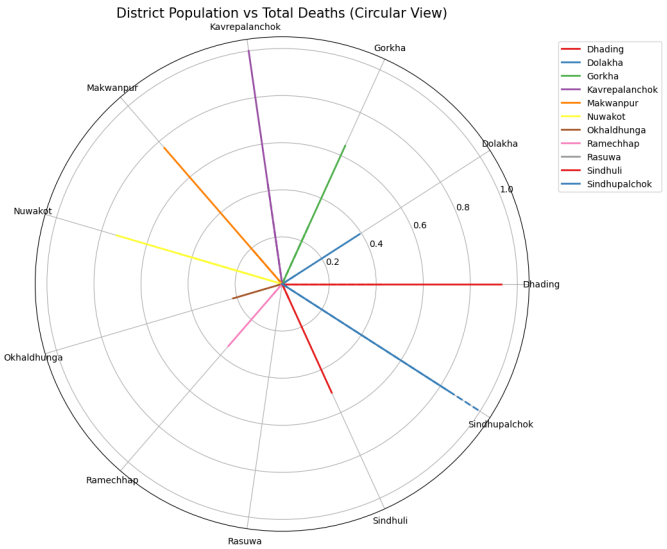


Fig. 30: District Population vs Total Deaths (Circular View)

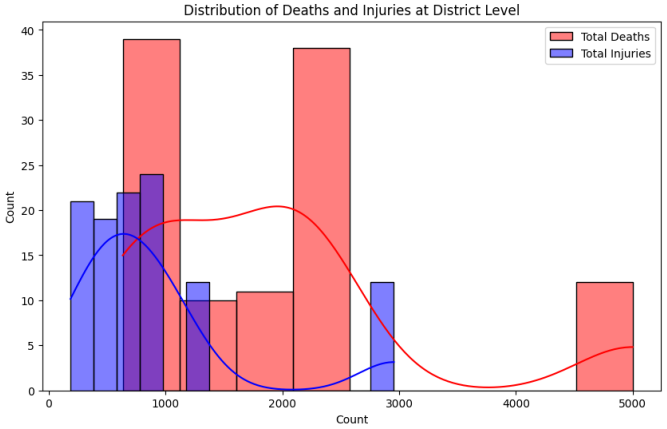


Fig. 31: Distribution of death and injuries in a combined graph

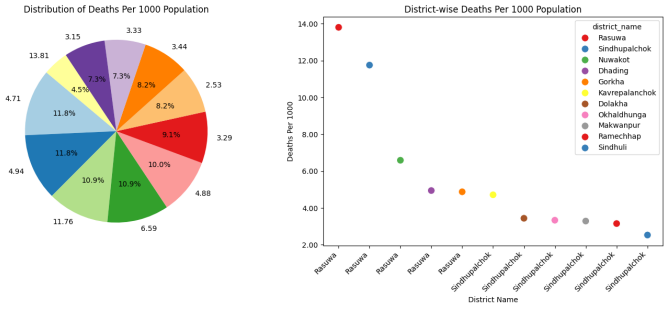


Fig. 32: Municipality-wise Deaths Per 1000 Population: Pie Chart and Point Chart

- **Integrated Scatter Plot with Regression Line:** This plot demonstrates population-death correlations and provides a clear trendline for interpreting overall relationships. **Fig. 33**

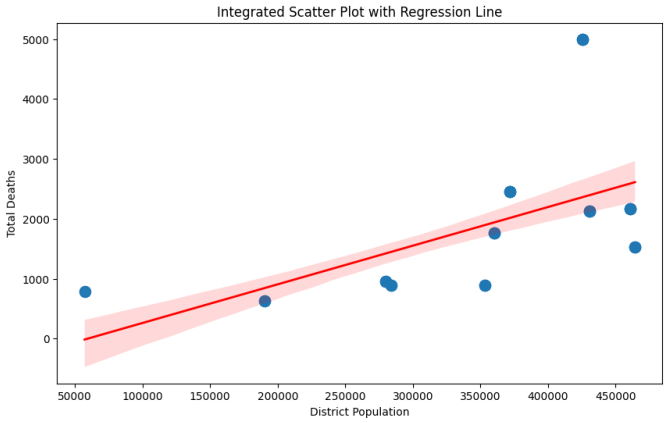


Fig. 33: Integrated Scatter Plot with Regression Line Total Death w.r.t Total Population

- **Node-Link Diagram of Risk Factor Correlations:** A heatmap reveals detailed correlations, while the node-link diagram illustrates risk factor relationships in a network format. **Fig. 34**

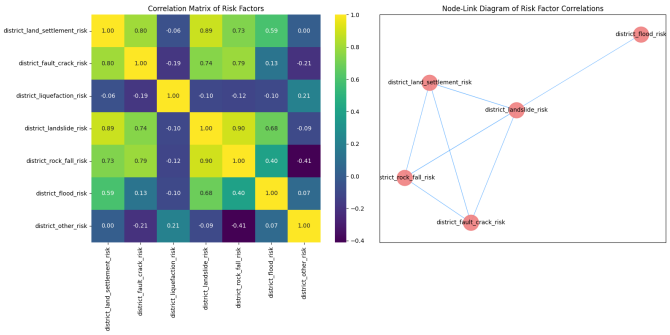


Fig. 34: Node-Link Diagram of Risk Factor Correlations

(vi) **Feedback:** This workflow demonstrates significant advancements over the first run by integrating refined visualizations and deeper analyses. Key improvements include:

- Clearer population-death patterns through circular and regression plots.
- Enhanced granularity in mortality rate visualizations.
- Actionable insights into risk factor correlations, providing a foundation for strategic decision-making.

For future iterations:

- Introduce temporal analyses for tracking changes over time.
- Incorporate interactivity in visualizations for dynamic exploration.
- Expand the dataset with additional socioeconomic and demographic factors.

These enhancements will strengthen the workflow's adaptability and application to public health scenarios.

*3) **Third Run** :*

(i) **Inference Taken From Second Run:** Building upon the insights from the second run, this iteration focuses on refining the analysis of population-death-injury relationships while adding clustering to identify risk patterns across VDCs. These new findings enhance our understanding of mortality and injury distribution, allowing for better-informed resource allocation and targeted interventions.

(ii) **Data:** The dataset includes total population, deaths, injuries, and detailed district-level risk factors. It also incorporates Rahat data to evaluate the effectiveness of relief efforts across districts. These data points are crucial for the comparative analysis between districts and risk profiles.

(iii) **Knowledge:** Key insights from this run include:

- Relationships between population, deaths, and injuries across districts.
- Clustering of VDCs based on total deaths and injuries.
- Identification of districts and VDCs with higher vulnerability to risks and casualties.

These insights have been integrated into the visualizations and clustering models presented in this iteration.

(iv) **Model:** Analytical techniques applied in this iteration include:

- K-means clustering for identifying patterns in casualties across VDCs.
- Comparative analysis of death and injury profiles by district.
- Integration of radar charts for visualizing combined risk and death profiles.

These methods provide deeper insights into the vulnerabilities of districts and VDCs.

(v) **Visualization:** The following visual formats were utilized to enhance understanding:

- **Superimposed Scatter Plot of Population, Total Deaths, and Total Injuries:** This plot combines data on district populations, total deaths, and injuries, with distinct markers for each type of casualty. **Fig. 35**
- **Risk and Death Profiles (Radar Chart) and Total Rahat by District (Bar Chart):** This visualization integrates a bar chart and radar chart to assess risk and relief distribution across districts. **Fig. 36**
- **Combined Risk and Death Profiles of Top Districts (Radar Chart):** A radar chart visualizes the risk and death profiles of the six highest-ranking districts in terms of total risk and death metrics. **Fig. 37**
- **Clustering of VDCs Based on Total Deaths and Injuries:** Using K-means clustering, this visualization groups VDCs based on their total deaths and injuries, highlighting patterns across regions. **Fig. 38**

(vi) **Feedback:** The third run introduces key improvements:

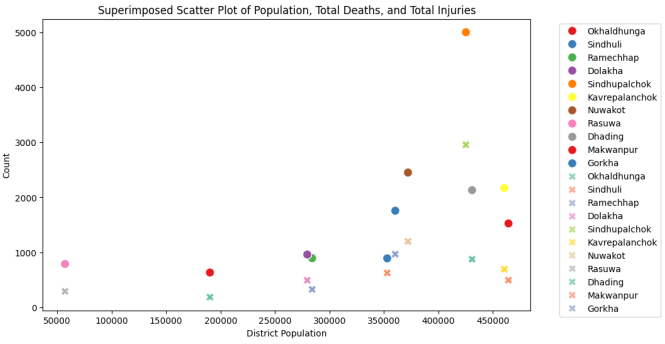


Fig. 35: Superimposed Scatter Plot of Population, Total Deaths, and Total Injuries

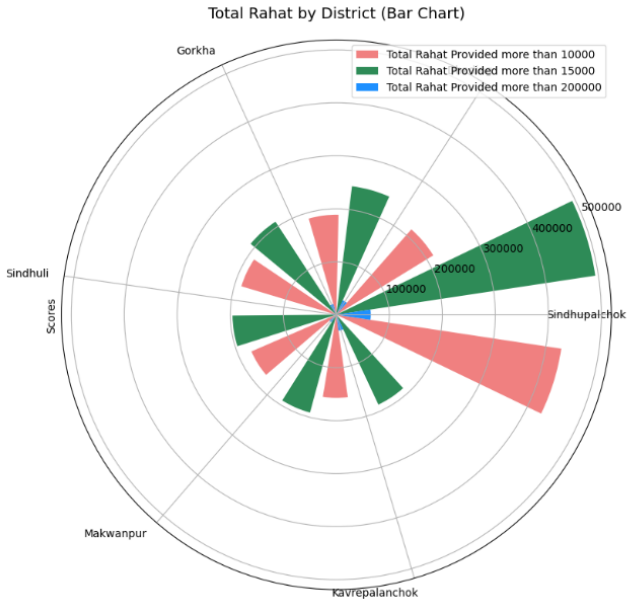


Fig. 36: Risk and Death Profiles (Radar Chart) and Total Rahat by District (Bar Chart)

- Enhanced insights into the distribution of deaths and injuries across districts.
- The addition of clustering analysis enables a clearer understanding of high-risk areas.
- Combining radar charts and bar charts provides a holistic view of risk and relief efforts.

For future iterations:

- Integrate temporal changes in death and injury rates.
- Introduce machine learning models to predict high-risk areas based on historical data.
- Expand clustering to include socioeconomic factors.

These improvements will further refine the analysis, leading to more targeted and efficient intervention strategies.

*4) **Fourth Run**:*

(i) **Inference Taken From Third Run:** Building upon insights from the third run, this iteration aims to refine the

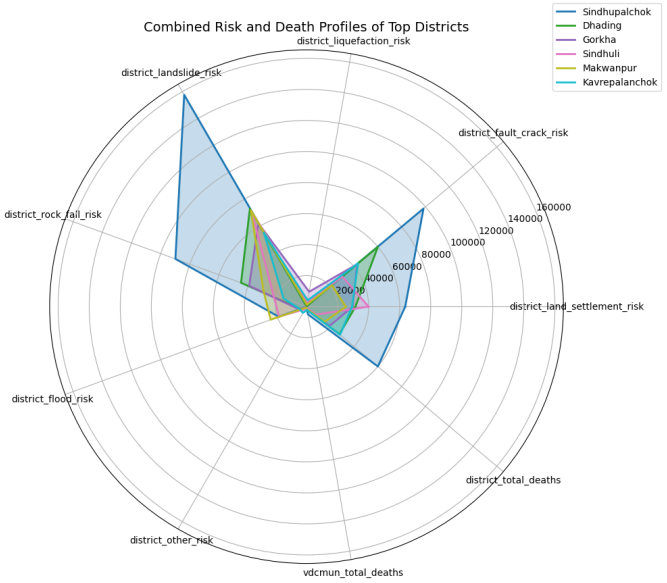


Fig. 37: Combined Risk and Death Profiles of Top Districts (Radar Chart)

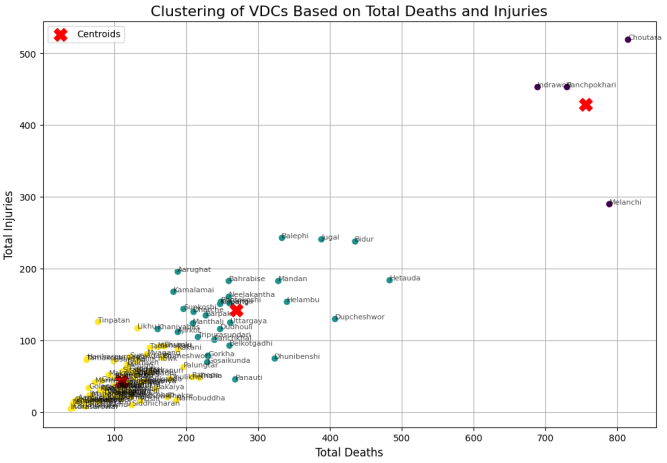


Fig. 38: Clustering of VDCs Based on Total Deaths and Injuries

clustering of districts based on the total number of deaths and injuries, adding a more detailed geographical context. The analysis now focuses on clustering districts based on casualties and integrating this with geographic data to identify regions with extreme vulnerability. The clustering methodology helps to distinguish high-risk districts and offers a clearer pathway for resource allocation and focused interventions.

(ii) **Data**: The dataset used in this iteration includes total population, deaths, injuries, district-level risk factors, and additional geographic data. This new data incorporates **VDC** (**Village Development Committee**)-specific information for more granular analysis. The dataset also integrates Rahat relief distribution to evaluate the effectiveness of post-disaster relief efforts.

(iii) **Knowledge**: Key insights from this run include:

- Enhanced clustering of districts based on total deaths and injuries.
- Identification of high-risk districts that require urgent interventions.
- Geographic factors influencing the distribution of deaths and injuries across districts.

These insights have been integrated into the updated clustering models and visualizations.

(iv) **Model**: Analytical techniques applied in this iteration include:

- **K-means clustering** for grouping districts based on total casualties.
- Comparative analysis using additional geographic data to highlight spatial patterns.
- Use of radar charts and bar charts to visualize risk and relief efforts in districts.

These methods refine the understanding of district-level vulnerabilities, aiding in more focused disaster response.

(v) **Visualization**: The following visual formats were utilized to enhance understanding:

- **Clustering of Districts Based on Total Deaths and Injuries (Clustered Scatter Plot):** This scatter plot visualizes the clustering of districts based on their casualties. Each district is color-coded according to its cluster. **Fig. 39**

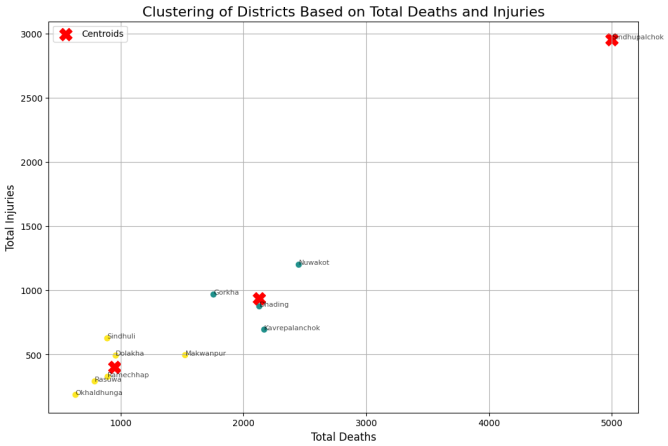


Fig. 39: Clustering of Districts Based on Total Deaths and Injuries

TABLE VII: Cluster-Wise Death Categorization

Cluster Number	Cluster Color	Deaths Range
1	Blue	Less than 400 deaths
2	Yellow	400 to 800 deaths
3	Red	More than 1200 deaths

- **Physical Map of Cluster Plotting Based on Most Affected Wards:** A map showing the physical location of districts and VDCs categorized by their cluster profiles. **Fig 40, Fig- 41**

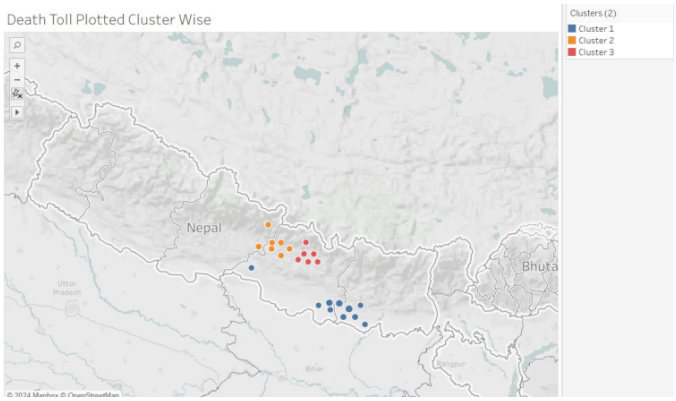


Fig. 40: Cluster Plotting Of Total Deaths and Injuries Cluster Wise As per the most affected Wards

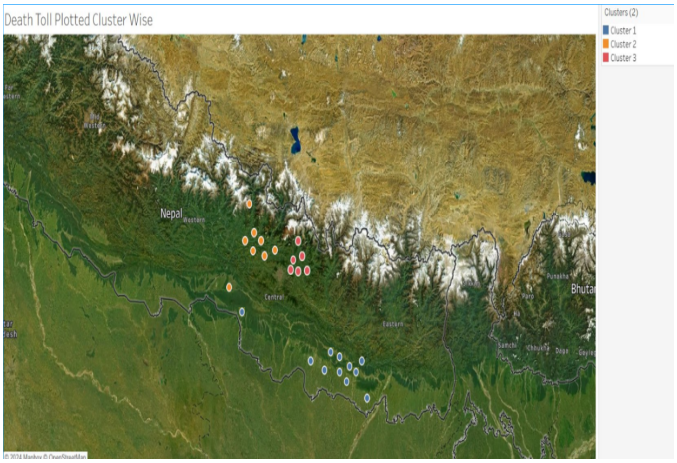


Fig. 41: Physical Map Of Cluster Plotting of Total Deaths and Injuries Cluster Wise As per the Most Affected Wards

We can see the relief features and the topography affects greatly in the *Damaging Lives*. More safer areas are downside where there are less deaths

(vi) **Feedback**: The fourth run introduces key improvements:

- Enhanced insights into the spatial distribution of deaths and injuries across districts.
- The clustering analysis now includes geographic factors for more detailed insights.
- Clearer visualization of high-risk areas using radar charts, scatter plots, and geographic maps.

For future iterations:

TABLE VIII: Cluster-Wise Death Categorization Physical Map

Cluster Number	Cluster Color	Deaths Range
1	Blue	Less than 400 deaths
2	Yellow	400 to 800 deaths
3	Red	More than 1200 deaths

- Incorporate temporal analysis of casualties to detect evolving patterns over time.
- Use machine learning models to predict future high-risk areas based on historical trends.
- Expand the clustering model to include socioeconomic data for a more holistic understanding of risk.

These improvements will further refine the analysis, leading to more efficient and targeted disaster response strategies.

H. *Story of the Entire Task*

*1) **Our Initial Approach**:* Our analysis began by examining the correlation matrix of various risk factors. This revealed critical insights into the relationship between the risk of landslides and other contributing factors. Understanding these correlations allowed us to identify key causes of death and destruction. To visualize these findings, we created informative box plots that showcased the distribution and severity of risk factors.

Subsequently, we performed a district-wise clustering of deaths to identify the most affected areas. Radar charts further refined our analysis by focusing on individual municipalities, enabling us to compute results per 1,000 people. These steps provided a granular view of the disaster's impact across different regions.

*2) **Taking Inferences from Previous Feedback**:* In our first iteration, we incorporated heatmaps, spatial features, and linked visualizations to map the disaster's impact. These visualizations highlighted significant patterns and relationships in the data.

During the second iteration, regression plots revealed death patterns that aligned with our clustering analysis. This iterative process enhanced our insights and allowed us to refine our models, resulting in more targeted and meaningful inferences.

*3) **Final Conclusion for Task 2 Assignment-3**:* For the final phase, we implemented a clustering model that combined data on deaths and injuries. This approach enabled us to plot clusters across Nepal, revealing that the hilly regions were the most at risk. These findings were consistent with the geotechnical risk patterns identified in our initial analysis.

The clustering model was further enhanced using K-Nearest Neighbors (KNN), which provided the final results. This comprehensive approach not only validated our earlier inferences but also highlighted high-risk regions with greater precision, guiding potential mitigation efforts effectively.

IV. Task-III-Damage Grade Prediction And Rahat Assessment

A. *Prior Visualization and Inferences Taken From Previous Work*

In Task 1, we conducted a detailed analysis of the earthquake's damage and its impact on the socio-economic status of the affected population. A key observation from the analysis was that the poorer segments of the population were disproportionately affected. This was evident from the fact that

the most significant damage was concentrated in the regions predominantly inhabited by the economically disadvantaged. Furthermore, an examination of the relief efforts, particularly in terms of *rahat* (relief assistance), revealed that the assistance provided to the poorer regions was minimal. This highlighted a noticeable gap in the equity of the relief efforts, emphasizing the need for more targeted and efficient relief distribution to those most in need.

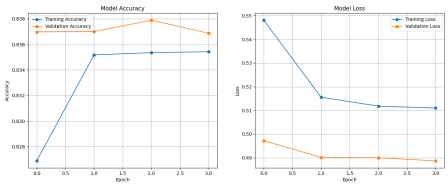


Fig. 42: Model Accuracy and Model Loss Of Decision Based Tree Classifier

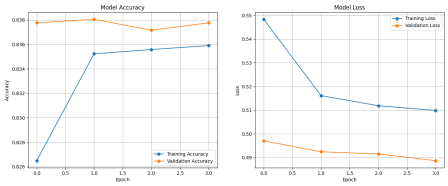


Fig. 43: Model Accuracy and Model Loss Of KNN Classifier

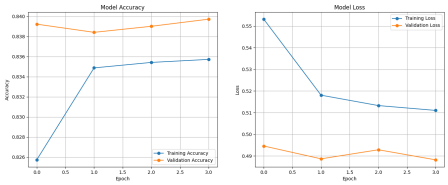


Fig. 44: Model Accuracy and Model Loss Of XG-Boost

B. Additional Dataset Used

In addition to the primary dataset for the 2015 earthquake [6], we integrated an auxiliary dataset containing Richter scale values recorded during the event [5]. This supplementary dataset was chosen due to its alignment with the primary dataset, particularly regarding the damage grade assessments. By merging these datasets, we enhanced the predictive capability of our analysis, allowing for a more accurate characterization of the earthquake's impact. The combined dataset facilitated a comprehensive understanding of the seismic event, enabling more precise predictions for future damage assessments and intervention strategies based on the differential regional impacts.

C. Pre-Processing The Dataset

The dataset was loaded using `pandas`, and missing values were removed with the `dropna()` function to maintain data integrity. The numerical column, such as *age*, was discretized into **12 equal bins** using `numpy.histogram` to transform continuous data into categories. These bins were normalized and mapped to polar coordinates for creating a Nightingale chart using `matplotlib.PolarAxes`. Aesthetic improvements, including consistent color schemes from `seaborn.color_palette()` and border enhancements with `matplotlib.patches`, were applied for clarity.

D. Data Context Provided

We extended the analysis by integrating a supplementary `Richter scale` dataset with the original data using `pandas.merge()`. This integration provided a more nuanced understanding of the spatial and socio-economic disparities in earthquake damage and relief distribution. The enhanced dataset, after preprocessing, enabled deeper insights into the correlation between earthquake intensity and socio-economic impact, visualized using `matplotlib` and `seaborn`.

E. Data Manipulation Techniques Used

*1) **Data Grouping Methods**:* The dataset was grouped based on several criteria to facilitate further analysis. Various statistical measures such as mean, standard deviation, minimum, maximum, and sum were computed for numerical columns. It was grouped on the basis of **damage grades** and **Rahat recieved**

F. Experiments With Machine Learning Models

1) Data Format Given To Us: The dataset comprises a combination of **numerical, categorical, and binary columns**. Numerical columns include *building_id*, which uniquely identifies each building, and *geo_level_1_id*, *geo_level_2_id*, and *geo_level_3_id*, representing hierarchical geographical locations. Structural attributes such as *count_floors_pre_eq* (number of floors before the earthquake), *age* (building's age in years), *area_percentage*, and *height_percentage* quantify physical characteristics. Categorical variables include *land_surface_condition*, *foundation_type*, *roof_type*, and *plan_configuration*, which describe construction features. Binary columns like *has_superstructure_adobe_mud* and *has_secondary_use_agriculture* indicate the presence or absence of specific structural or usage attributes.

2) Models Used: The machine learning models employed in this study are as follows:

- **Decision Tree Classifier**: A tree-based model that splits the data iteratively based on feature thresholds to classify the target variable.**Fig 42**
- **K-Nearest Neighbors (KNN)**: This model classifies a sample based on the majority class among its nearest neighbors in the feature space. **Fig 43**
- **Random Forest Classifier**: An ensemble model combining multiple decision trees, trained on random subsets of data and features, to improve classification accuracy and reduce overfitting.

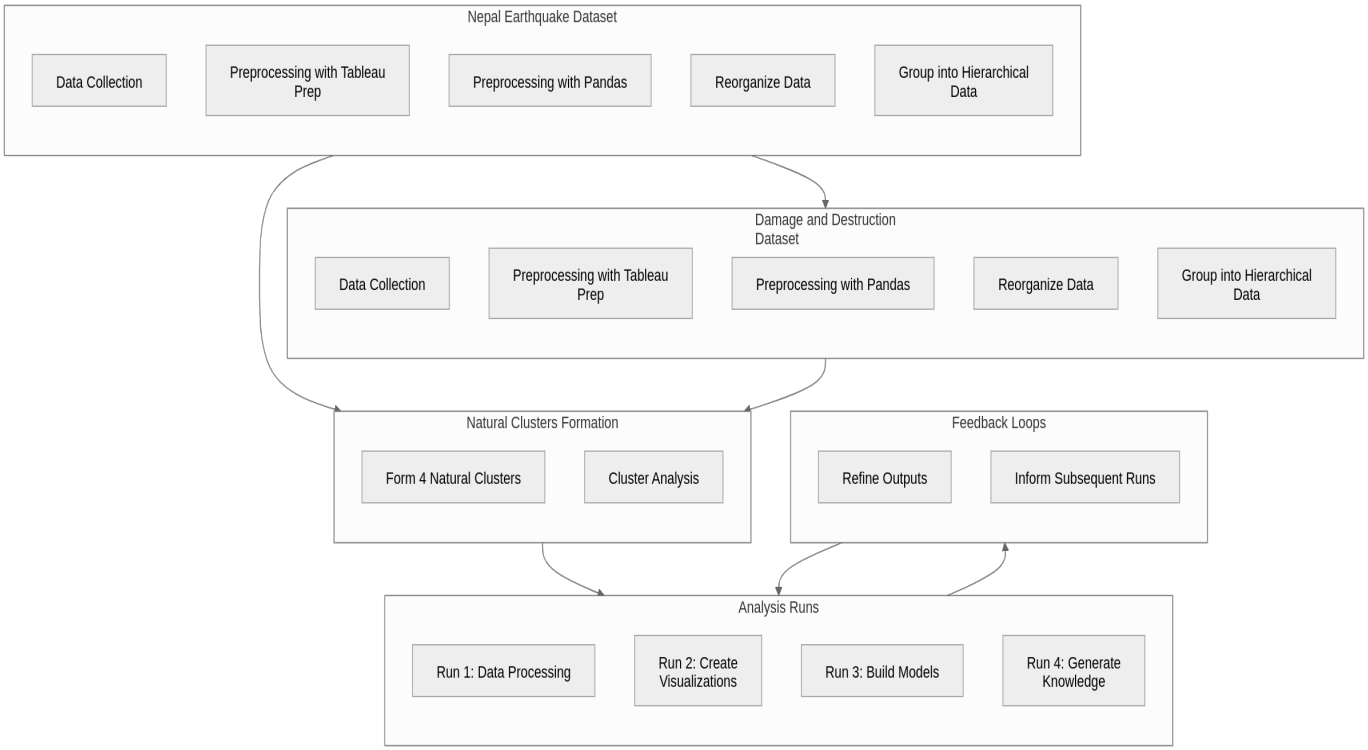
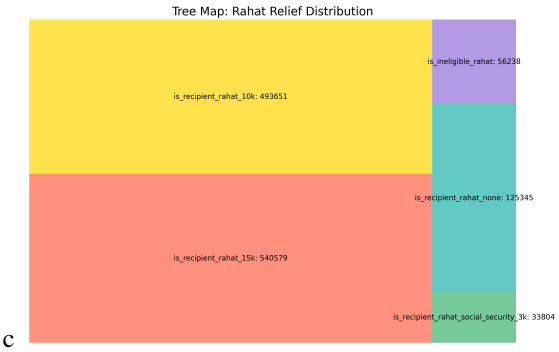


Fig. 45: Workflow diagram of the Nepal earthquake dataset analysis.



c

Fig. 46: Rahat Distribution (Run 1)

- **Gradient Boosting Classifier**: A boosting method that builds trees sequentially, minimizing the errors of previous models, to achieve strong predictive performance.
- **XGBoost Classifier**: An optimized implementation of gradient boosting with additional features such as regularization and parallel processing, used for its high accuracy and efficiency. **Fig 44**

Each model was chosen for its specific advantages in handling the given classification problem. **K-Nearest Neighbors (KNN)** was selected due to its simplicity and effectiveness in classifying samples based on the proximity of other data points, particularly when data points are well-separated in the feature space. **Decision Tree Classifier** was used for its interpretability and ability to model non-linear relationships in the data, as it builds clear decision rules based on feature values. **Random Forest Classifier**, being an ensemble of decision trees, was employed to improve model stability and accuracy by averaging predictions from multiple trees, reducing overfitting that may occur in individual decision trees. **Gradient Boosting Classifier** was used for its ability to sequentially build trees, each correcting errors from the previous tree, which results in improved model performance, especially on imbalanced datasets. Lastly, **XGBoost Classifier** was chosen for its efficiency and scalability, with features like parallel processing, regularization, and handling of missing values, making it ideal for high-performance predictions in large and complex datasets.

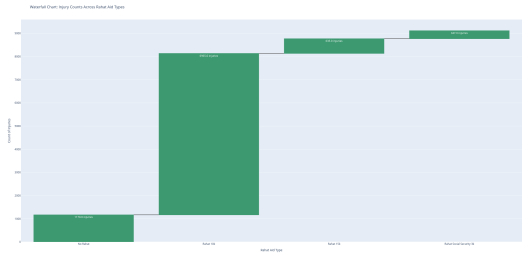


Fig. 47: WaterFall Graph to visualize the Rahat (Run 1)

G. *Describing The Visual Analytics Workflow*

You can look at **Fig 45** for getting an idea of the approach of our Visual Analytics for the task of plotting the damage grade assessment and rahat measures initiated

*1) **First Run**:*

(i) **Data:** In the initial data processing stage, the primary target variable `damage_grade` was encoded for classification. Missing values in the dataset were imputed using appropriate strategies, ensuring data integrity. Categorical variables were numerically encoded to suit machine learning algorithms, and feature selection was conducted to retain relevant variables.

(ii) **Visualization:** Refer to the A1 report for detailed visualizations. Key visualizations include:

- **Waterfall Graph Of Rahat Distribution Distribution (Run 1):** A waterfall chart **(Fig. 47)** shows the waterfall of Rahat recieved in one of the most affected districts.
- **Damage Grade Distribution: (Run 1)** A bar chart **(Fig. 48)** shows the frequency of each damage grade.

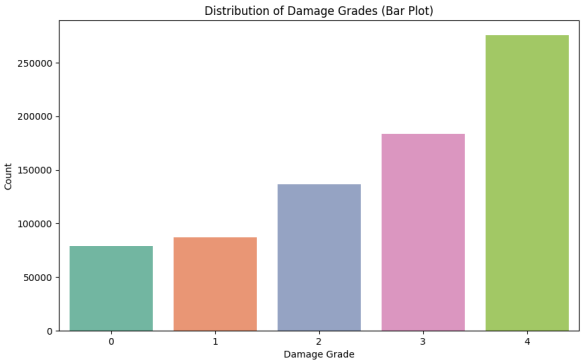


Fig. 48: Distribution of Damage Grades (Run 1)

- **Correlation Heatmap: (Run 1)** Displays relationships between damage-related variables and geotechnical risks**(Fig. 49)**.
- **Most Frequent Damage Types: (Run 1)** A bar chart **(Fig. 50)** illustrates common types of damage.

(iii) **Models:** No machine learning models were applied in this phase. The focus was on visual exploration to guide model selection for later stages.

(iv) **Knowledge:**

- **Damage Grade Distribution:** Insights into building conditions and damage trends are shown in Fig. 48.
- **Geotechnical Risks:** Highlighting environmental factors like liquefaction to inform mitigation strategies.

(v) **Feedback:** After doing a thorough analysis we got a load of distribution of the damage grades and the relavent factors by the correlation matrix

*2) **Second Run**:*

(i) **Inference taken from first run**

Key findings from the first run guided improvements in this iteration. It was observed that areas with lower

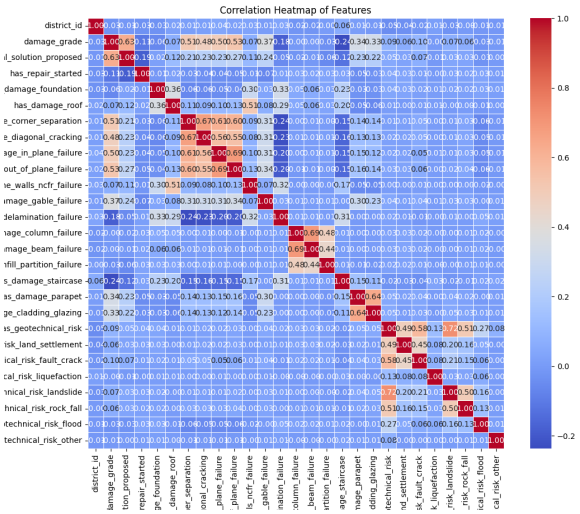


Fig. 49: Correlation Heatmap of Features (Run 1)

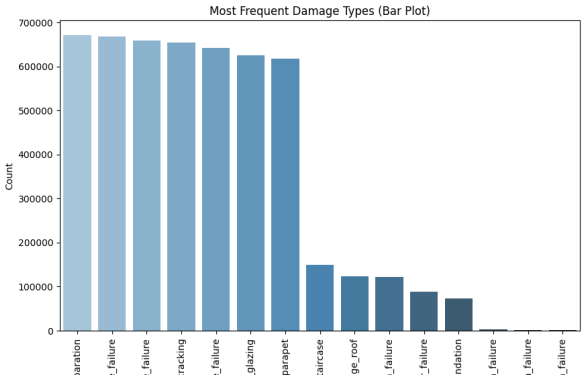


Fig. 50: Most Frequent Damage Types (Run 1)

relief amounts correlated strongly with regions reporting less damage severity. This insight helped refine cluster formation, ensuring more accurate representation of relief distribution. Additionally, feedback highlighted the need for **incorporating socioeconomic factors**, which were integrated into the second run for a more comprehensive analysis.

(ii) **Data**

In this second iteration, the dataset underwent additional cleaning and advanced feature engineering. Key enhancements included the inclusion of more features to better capture the variability in damage assessments. The data also incorporated new dimensions for improved analysis, such as the impact of socioeconomic factors and relief measures. These refinements aimed to enable deeper insights into the spatial and statistical relationships between damage severity, relief allocation, and resource distribution.

(iii) **Visualization**

To gain insights, multiple visualizations were created:

- **Damage Grade Distribution: Fig. 51** shows the frequency distribution of damage grades.

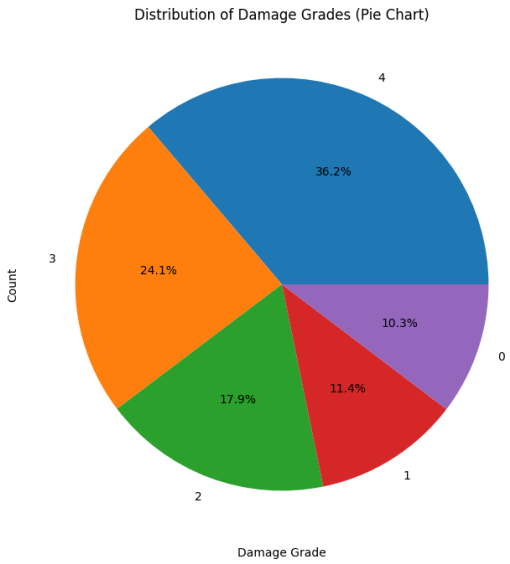


Fig. 51: Distribution of Damage Grades (Run 2)

- **Correlation Heatmap:** A heatmap was generated to identify correlations among features, showcasing the impact of additional features. **Fig. 52.**

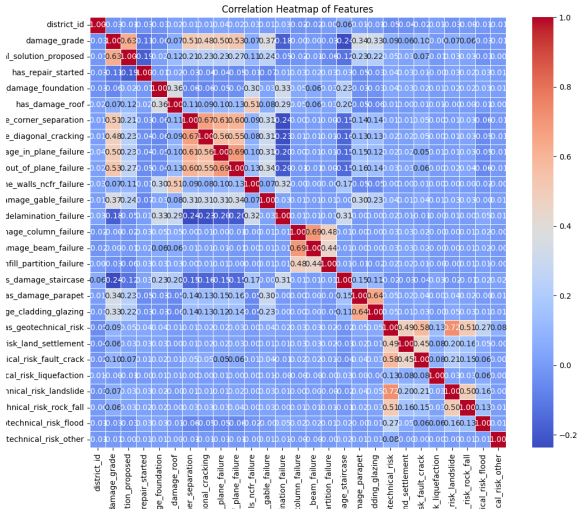


Fig. 52: Correlation Heatmap of Features (Run 2)

- **Most Frequent Damage Types: Fig. 53** highlights the most common damage types across the dataset.
- **Cluster Visualisation in Map: Fig. 55, Fig. 54** We took the map of Nepal and plotted the districts where Rahat has been recieved in the denominations of Rs 10,000 and Rs 15,000

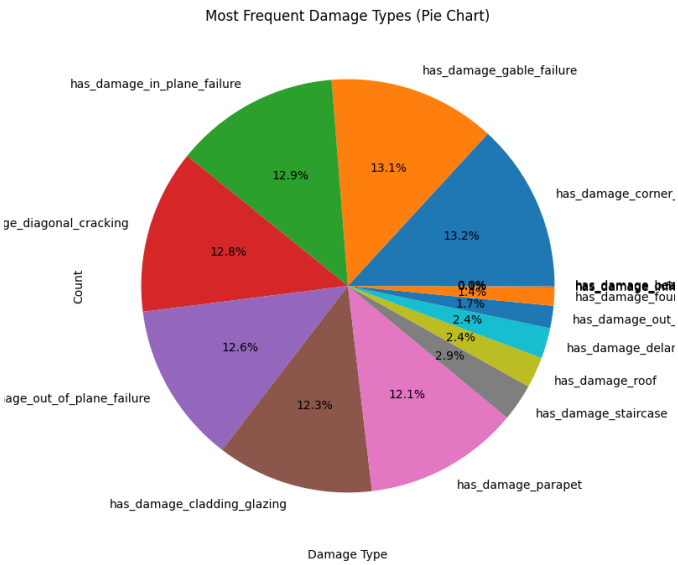


Fig. 53: Most Frequent Damage Types Based on the structure (economic aspect) (Run 2)

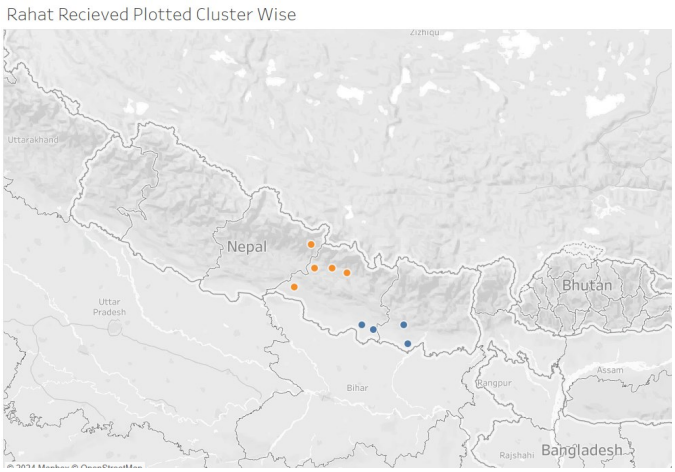


Fig. 54: Clustering of plots of Rahat recieved District wise in a gray scale map (Run 2)

(iv) Model

Cluster Number	Cluster Color	Relief Received Range
1	Blue	Rahat till 10,000
2	Orange	Rahat till 15,000

TABLE IX: Relief Clusters by District in a grayscale

The KMeans clustering model was used for grouping the dataset based on damage levels. The optimal number of

clusters was determined using the elbow method, ensuring a balance between model complexity and accuracy. The PCA visualization helped reduce dimensionality, providing a clearer separation of clusters.

Fig. 55: Cluster plotting of the Rahat provided District wise in a Physical Map of India (Run 2)

(v) **Knowledge**
Analysis of the second run provided the following insights:

Cluster Number	Cluster Color	Relief Received Range
1	Blue	Rahat till 10,000
2	Orange	Rahat till 15,000

TABLE X: Relief Clusters by District

- Areas with higher damage grades were associated with denser clusters in the PCA plot, reflecting severity.
- Correlations revealed strong links between socioeconomic factors and damage severity, guiding targeted relief efforts.
- Clustering highlighted geographic hotspots of destruction, aiding prioritization of aid.
- By plotting the district that have recieved the Rahat we could exactly pin point where the immediate need was needed for attention. Who have got more Rahat who have less.
- We found out that Blue have recieved less **Because the damage happened was less** this was later verifed with the analysis of Geo technical risks and the Death and destruction,

(vi) **Feedback**
Feedback from this iteration emphasized the importance of integrating temporal data to analyze trends over time. Future iterations could include:

- Adding time-series data for relief distribution to monitor consistency.
- Including demographic variables to ensure equitable resource allocation.
- Exploring machine learning models for predictive analytics, aiding proactive disaster management.

*3) **Third Run**::*

(i) **Inference from Second Run:** Based on the analysis and feedback from the second run, the third iteration of the model introduced several improvements in both data preprocessing and visualization. The feature engineering steps were refined to better capture relationships within the dataset, including better handling of missing values and outliers. The clustering approach was enhanced to provide more distinct and meaningful groupings of damage grades, which allowed for better predictive performance in subsequent modeling steps.

(ii) **Data:** Building on the progress and feedback from earlier runs, additional data transformations were introduced to prepare the dataset for more complex modeling:

- **Feature Scaling:** Min-Max scaling and standardization techniques were applied to ensure all numerical variables were on comparable scales.
- **Outlier Detection and Handling:** Outliers in numerical features were detected. Identified outliers were either capped to reduce their impact.
- **Data Transformation:** Skewed features were log-transformed to improve symmetry and better align with modeling assumptions.

(iii) **Visualization:** Advanced visualizations were generated

- **Damage Grade Distribution (Run 3):** The distribution of damage grades for Run 3 is visualized using a bar chart (Fig. 56).

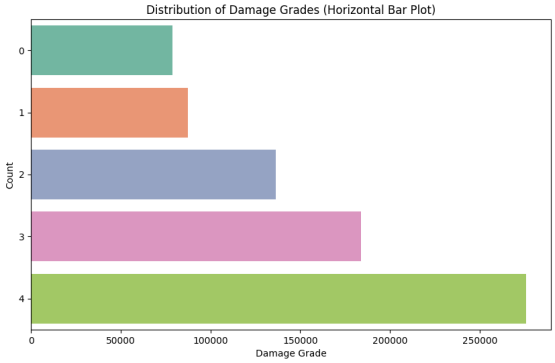


Fig. 56: Distribution of Damage Grades (Run 3).

- **Correlation Heatmap (Run 3):** The relationships between damage grade features and other are visualised in Run 3 **an advanced version from Run 2 and 1** are visualized using a correlation heatmap (Fig. 57).
- **Most Frequent Damage Types (Run 3):** The most frequently reported damage types for Run 3 are shown in Fig. 58.
- **PCA Visualization (Run 3):** The PCA visualization for Run 3 is shown in **Fig. 59**. This plot highlights the structure of the dataset after dimensionality reduction. This combines and gives us better results on combining it with K Means
- **Clustering (Run 3):** The clustering results using the

KMeans algorithm for Run 3 are visualized in **Fig. 60.**

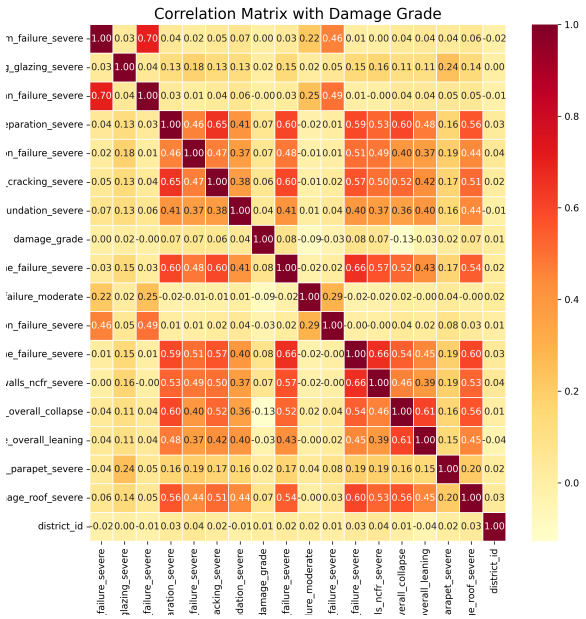


Fig. 57: Correlation Heatmap of Features With Damage Grade (Run 3).

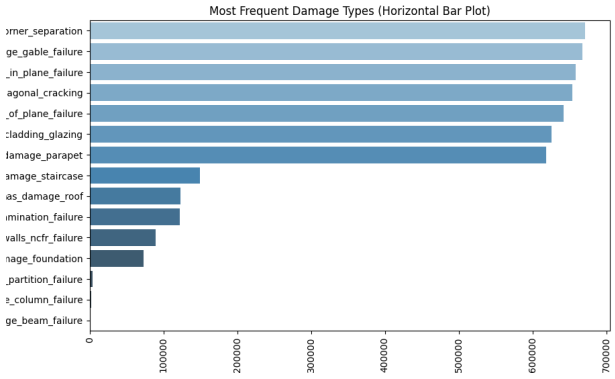


Fig. 58: Most Frequent Damage Types (Run 3).

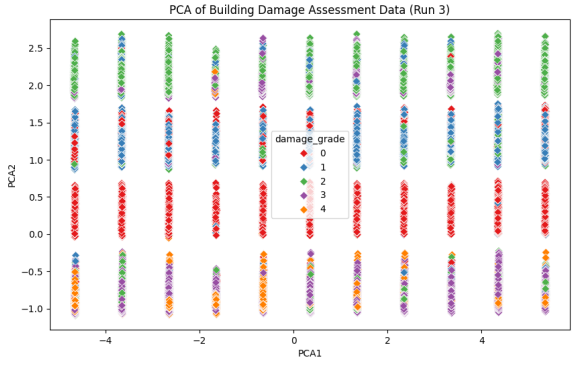


Fig. 59: PCA of Building Damage Assessment Data (Run 3).

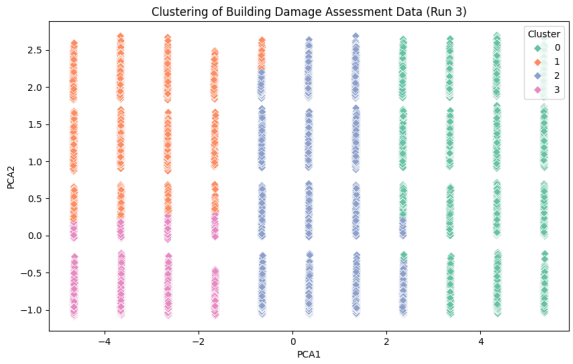


Fig. 60: Clustering of Building Damage Assessment Data (Run 3).

- **Damage Grade by Clusters (Run 3):** The distribution of damage grades across clusters for Run 3 is presented in **Fig. 61.**

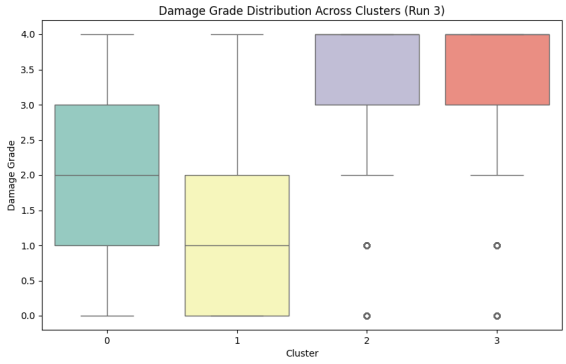


Fig. 61: Damage Grade Distribution Across Clusters (Run 3).

- **Geotechnical Risks (Run 3):** The distribution of geotechnical risks for Run 3 is shown in **Fig. 62.**

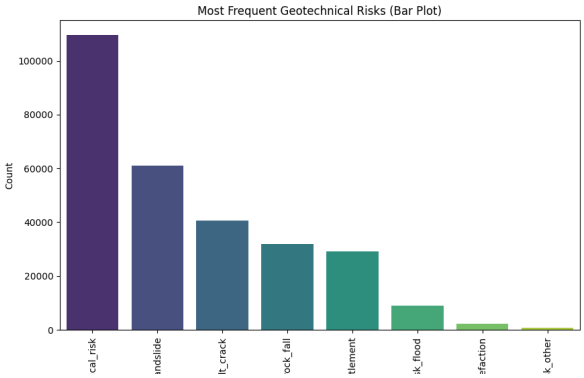


Fig. 62: Most Frequent Geotechnical Risks (Run 3).

(iv) **Model** We finally applied the **XG-Boost classifier** on several of these datasets increasing the number of columns of analysis. On every analysis we ran our moel got more insights and obtained better results

(v) **Knowledge:**

- **Damage Grade and Clusters:** Refined clustering highlighted regions and patterns with concentrated damage grades and geotechnical risks. These insights can guide interventions and resource allocation.
- **Predictive Modeling:** Feature importance analysis revealed the key predictors of damage grades, offering actionable insights for building risk assessments and mitigation strategies.

(vi) **Feedback:**

- Explore alternative clustering techniques for better group separation.
- Further refine feature importance analysis.
- Investigate the role of geotechnical risks in more complex modeling.
- Plot a Rahat distribution map ward wise for the final clustering

*4) **Fourth Run**:*

(i) **Inference from Third Run:** Building upon the insights from the third run, we refined our focus on the damage grade distribution and its correlation with **geotechnical** risks. We also deepened the clustering analysis by incorporating additional features that enhance our ability to predict damage outcomes. The key takeaway from Run 3 was the importance of better feature scaling and transformation techniques to address skewed data. After taking all the features into our account we finally got the **ward wise analysis** of cluster plotting

(ii) **Data:** For Run 4, we continued refining the dataset through additional **preprocessing steps**. Features were further standardized, and missing values were addressed with more advanced imputation techniques.**Outliers** were handled efficiently These steps helped ensure that the model could make more accurate predictions by reducing the impact of extreme values and improving data distribution.

(iii) **Visualization:** In this run, we enhanced the visualizations to further clarify patterns in the dataset and predictions:

- **Damage Grade Distribution Area Curve (Run 4):** The distribution of damage grades with the district for Run 4 is visualized using a area curve which gets **(Fig. 63)**.
- **Most Frequent Damage Types with including the socio economic factors (Run 4):** The most frequently reported damage types for Run 4 are shown in **Fig. 64**.
- **Geotechnical Risks (Run 4):** The distribution of geotechnical risks for Run 4 is shown in **Fig. 65**. We see that more Rahat has been recieved in those areas where more damage has taken place hence aligning with our data.

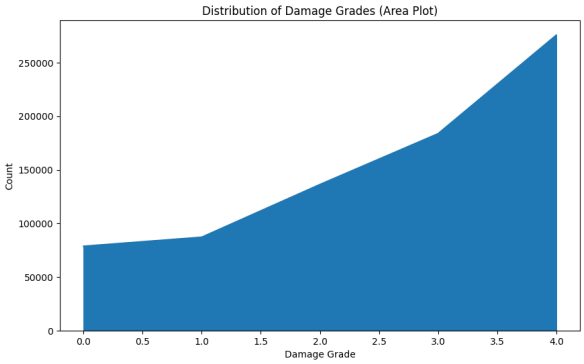


Fig. 63: Distribution of Damage Grades (Run 4).

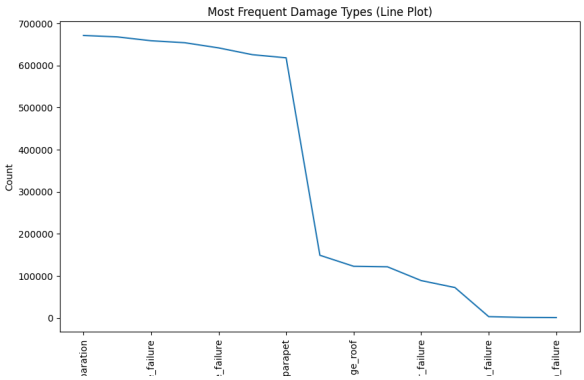


Fig. 64: Most Frequent Damage Types (Run 4).

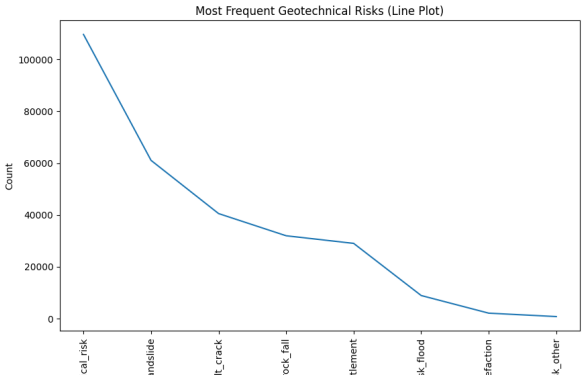


Fig. 65: Most Frequent Geotechnical Risks Correlated with the damage types (Run 4).

(iv) **Model:** In this run, we implemented advanced machine learning models to further predict damage grades and evaluate feature importance:

- **XG Boosting Classifier:** A XG Boosting Classifier was introduced, providing better predictive power by boosting weak learners and reducing bias. This model showed improvements in accuracy compared to previous models.

TABLE XI: Table showing Rahat estimate

Cluster Number	Cluster Color	Deaths Range
1	Blue	Less than Rs10,000
2	Yellow	More than Rs15,000

(v) **Knowledge:** The key insights from this run focused on the refinement of damage grade predictions, driven by better feature engineering and clustering. The XG Boosting Classifier's improved performance offers deeper insights into the relationship between damage types, geotechnical risks, and damage grades. Understanding these relationships is crucial for accurate risk assessment and resource allocation. When we went and finally plotted the clusters ward wise we found out that the **Death and Destruction** data was matching with the Rahat recieved.

(vi) **Feedback:** We can conclude that the **Nepal** govt. has given Rahat in the regions most needed. The Rahat distribution has matched with the entire geo technical risks. As the hilly regions had more damage (Prooved by the death and destruction) we could easily see that it was more impacted hence in need off more Rahat

H. Story Of The Entire Task

*1) **Our Initial Approach**:* Initially, our objective was to analyze the geographical impact of damage and destruction caused by a calamity, specifically using the WDCUM (Weighted Destruction and Cluster Utilization Method). The first runs gave us an impact of the earthquake how damage it has caused and what all fault lines have been happened. Then after combining various parameters and models we came to several conclusions after analysing various correlation matrices. Hence our ML Model quitely predicted the accurate values of where damage has happened and where more action is neeeded.

TABLE XII: Table showing Rahat estimate

Cluster Number	Cluster Color	Deaths Range
1	Blue	Less than Rs10,000
2	Yellow	More than Rs15,000

*2) **Taking Inferences from the Previous Feedbacks**:* Feedback indicated the need for actionable insights linking the damage levels to relief distribution. Consequently, in the second plot, we incorporated data on relief measures ("Rahat") received in the affected regions, derived using the WDCUM approach. This plot clusters the areas based on relief distribution, showing a correlation between damage intensity and the volume of relief provided. Orange clusters (heavily impacted areas) align with regions receiving substantial relief, while blue clusters (less affected regions) received comparatively less.

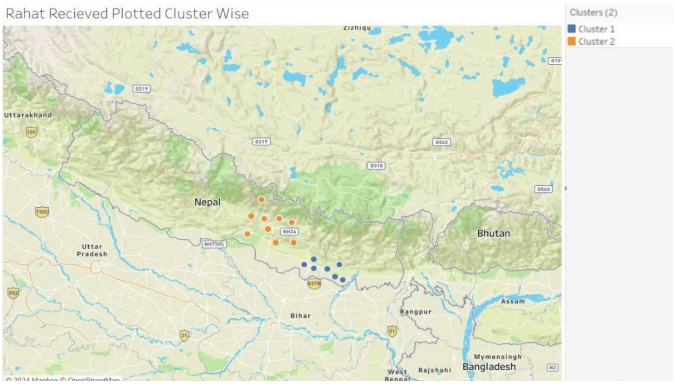


Fig. 66: Physical cluster map estimate of the rahat obtained ward wise

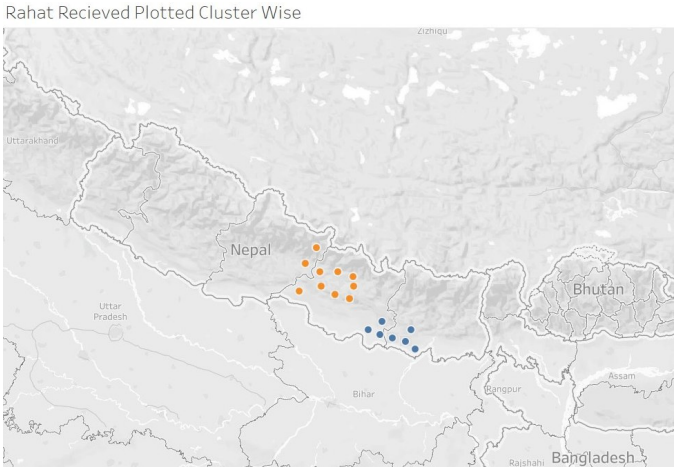


Fig. 67: Physical cluster map estimate of the rahat obtained ward wise

*3) **Final Conclusion For A-3 Task-3**:* By comparing both plots, a clear inference can be drawn: regions that faced higher destruction (orange areas in the first plot) were prioritized for relief efforts (orange areas in the second plot). This suggests an efficient allocation strategy based on the severity of the calamity, leveraging the WDCUM method for better resource management. Such insights are critical for assessing the effectiveness of disaster response mechanisms.

V. Scope For Future Expansion Of the Report

1) **Inclusion of Temporal Data:** Analyzing how relief distribution evolved over time to ensure consistency.
2) **Integration of Socioeconomic Factors:** Incorporating demographic and economic data to assess the equity of relief distribution.
3) **Detailed Resource Analysis:** Studying the types of relief resources provided and their adequacy for each cluster.

4) **Machine Learning Models:** Using predictive analytics to simulate future disaster-relief scenarios and optimize strategies.

VI. Overall Conclusions from all the tasks

1) **Task 1** Analyzing the geotechnical risks and which districts are in the most potential danger if in case **A future earthquake happens** (**Makwanpur and Kavrepalchok**). Those and the hilly foothills districts and wards.
2) **Task 2** Those districts where there has been more risk has been proved to be the source of maximum destruction and they need to be immediatly addressed . Those districts where maximum death happened there only maximum damage and those were the regions and districts with maximum risks (**Makwanpur and Kavrepalchok**)
3) **Task 3** Studying the Rahat recieved in those districts found out the **Economically less** prospering districts havent recieved much RAHAT. Also we found out that the Damage grade and the correlation with the geo technical risk. We found out that landslides are the most common ones

VII. Individual Contribution

1) **Hemang Seth** Completed the entire workflow and analysis of the **Task 1 geo technical risks** and implemented the tableau.
2) **Tanish Pathania** Completed the entire workflow and analysis of the **Task 2 death and destructions** and implemented the tableau.
3) **Vasu Aggarwal** Completed the entire workflow and analysis of the **Task 3 Rahat provided and the damage grade** and implemented the tableau.

References

[1] Visual Analytics Workflow and Processes *Framework for Visual Analytics*. By **Ivan Bozov** Available at : vizmind.edu. Accessed: 2024-11-25.
[2] Seismic Hazard Assessment and Risk Management *Analyses on 2015 Goorkha Earthquake*. By **Team Deadline Missed** Available at : Assignment 1. Accessed: 2024-09-10.
[3] Challenges in Visual Data Analysis By **Kiem** Available at :
[4] Earthquake Dataset. *Earthquake Dataset*. Available at: Earthquake dataset
[5] Supplement Ritcher Scale Dataset. *Ritcher Earthquake Dataset*. Available at: Ritcher Earthquake dataset
[6] Earthquake Dataset. *Earthquake Dataset*. Available at: Earthquake dataset
[7] Analysis Earthquake Pic *Fault Pic* Available at: Here