Data Science Assignment 4(Group 9)
Critical Infrastructure Vulnerability due to Landslides

Pritesh Maheshwari, Nick Meyer, Anindita Ghosh,
Vasundhara Acharya, Sarvesh Patidar, Himanshu Dey

December 9, 2021

Introduction

The primary goal of our project is to identify the vulnerability of Critical infrastructure (mainly power plants) due to Landslides in the North-Eastern regions of the US. Our workflow can be broadly divided into data acquisition, data retrieval, data analysis, and data archival. Furthermore, each step has data storage, reflection phase, and documentation step in common. Our group started by identifying and collecting data from various sources like precipitation [NAS], landslide, and earthquake data was collected from NASA. The data about Critical infrastructure was collected from the Homeland security website. The collected data was then verified to eliminate errors in data capturing. The next big task was to merge the different data sets, for which we converted all our data in a single format. Finally, based on latitude and longitude all our data sets were merged. With the merged data we then performed exploratory data analysis and visualization to understand the relation between our data sets. We also applied different unsupervised machine learning algorithms to find the vulnerable locations. The identified features and findings are detailed in the report. We have extensively referred to the class notes[Mun] for this assignment. We have also referred to our previous assignments.

1 Choose an investigation and identify pre-existing source of data that can address a particular data science goal

1.1 Choose, and state, the goal and reasons why the datasets were chosen and how they were found and managed, Min 3-4 sentences.

Our goal for this project is to assess the risk of landslides, precipitation, and earthquake in the north-eastern part of the USA and provide sufficient information to assist in critical infrastructure planning, development, and strategy upgrading. To achieve the above objective, data related to topography, precipitation, and other geographically related factors such as earthquake was collected. We have collected our data from NASA's website and other U.S. Government websites, the source of our datasets are detailed below:

Data	Map	Format	Source	Link
Elevation and Slope Data	Google Earth Map	CSV	Google Earth Pro,GPS Visualizer"	[gps]
Precipitation Data	Global Precipitation Map	NetCDF	Goddard Earth Sciences Data and Information Services Center (GES DISC)	[NAS]
EarthQuake Data	EarthQuake Maps	CSV	The USGS Earthquake Hazards Program	[USG]
Power Plants Data	Geospatial map	Shape File	Homeland Infrastructure Foundation-Level Data (HIFLD)	[HIF]

All the datasets were stored as .csv and .h5 files. The datasets were added on Google Drive as a sharing medium among the team members. It is also uploaded on the Github repository for version control and data archival.

1.2 Document and discuss the data formats and any metadata standards/ conventions in use, and the method(s) of discovery and access and how they helped or hindered the process, Min 3-4 sentences.

Since all our datasets were gathered from NASA and the U.S. Government websites, the data formats and conventions have been used without any major changes in the project work. The date format used in our datasets was the ISO standard of mm/dd/yyyy. Other information about the formats and conventions used in the project is uploaded to the Metadata file that is uploaded to the Github repository. The metadata for all the datasets was easily discoverable on their respective websites. These metadata are maintained and updated regularly by NASA and other Government websites. Therefore, all the datasets gathered were clean and required very few or no changes in the format. These datasets were easily usable for the analysis part as their formats and standards were adaptable through the use of the Python libraries.

2 Data Analysis

2.1 Develop and state two particular questions/hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study (preliminary, full and post) to answer these questions and document the analysis design, Min 3-4 sentences

Climate change is causing an increased impact and frequency of natural disasters all over the globe. Among the natural hazards, landslides are among the most dangerous natural disasters, resulting in significant economic damage and human deaths worldwide. To deal with these threats of natural hazards a comprehensive disaster management strategy must be devised and implemented.

Questions/Hypothesis

- Driven by this goal our first question is to assess the vulnerability of critical infrastructures across the northeastern part of the US caused due to landslides. Compute a risk index for each of the critical infrastructures.
- What is the relationship between factors such as the precipitation, elevation, slope with the landslide probability.

The following workflow created in PowerPoint describes our team's analysis design.

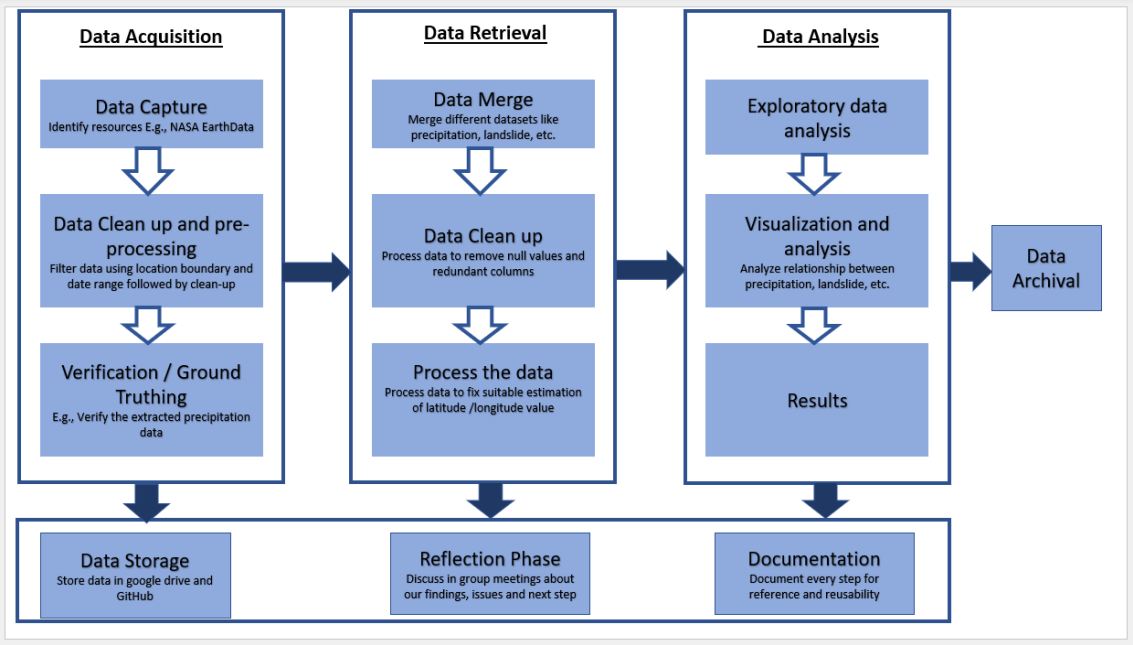

Figure 1: Workflow Diagram

Our workflow can be broadly divided into data acquisition, data retrieval, data analysis, and data archival. Furthermore, each step has data storage, reflection phase, and documentation step in common. Our group started by identifying and collecting data from various sources like precipitation [NAS], landslide, and earthquake data was collected from NASA. The data about Critical infrastructure was collected from the Homeland security website. The collected data was then verified to eliminate errors in data capturing.

The next big task was to merge the different data sets, for which we converted all our data in a single format. For example, precipitation data were converted from netcdf4 to CSV format, landslide data was converted from GeoTIFF to CSV format, etc. All the data clean-up and pre-processing were done using python and we shared our colab notebook among our team. Finally, based on latitude and longitude all our data sets were merged.
With the merged data we then performed exploratory data analysis and visualization to understand the relation between our data sets. We also applied different unsupervised machine learning algorithms to find the vulnerable locations. The identified features and findings were then included in the report. At all steps the data is stored in our shared google drive and GitHub. The reflection phase comprised group meetings where we discussed our progress, failure, and success and decided on our next step. All our work was documented for future reference and also to ensure data re-usability.
The project report will be archived in GitHub with all necessary documentation to ensure data preservation and data stewardship.

2.2 Provide a description of the choices of tools/methods used or a description of any code or scripts written, and describe how your results were stored and managed, Min 3-4 sentence. Submit your code to course GitHub repository for evaluation

We used python coding to downloaded, pre-process and merge the datasets. We have stored our final dataset in an HDF file. Our primary aim was to find a relation between the critical infrastructure vulnerability, landslide, precipitation, and earthquake datasets. For this, we first captured similar data for the Northwestern regions of the US and used it as our training set. We then used a Random Forest Regression Algorithm to find the critical infrastructure vulnerability of the northeast.

For our visualizations and exploratory data analysis we plotted our data using scatter plots and bar plots. All our python codes, data files, visualization plots, and python scripts are stored on GitHub to ensure data preservation and data re-usability.

2.3 Perform the analysis in a form that can be validated and describe the steps and results your group took to ensure this validation, Min 3-4 sentences

Great care was taken to ensure that each stage of analysis and data preparation was validated and checked. The two major types of workflows done in this project involved processing data, or creating models. Each category had a methodology for validation.

In the data processing workflow, several different operations were performed. These included row level changes, merging, and filtering. When row level changes were performed, manual checks were made to ensure that the row level operation did what was expected. Using descriptive statistics for that row also indicated if there were any failures. On merge operations, the length of the dataset can easily change. When merging it was important to look at that cardinality before and after the merge to make sure it met the expectations. Many failed merges were caught by this check. When filtering, a combination of the two checks were performed to catch failed column or row drops. By employing these steps at each stage of the data processing workflow, many errors were caught.

It was also important to make sure each stage of model development was validated. When creating our models, the workflow can be separated into pre-processing, model development, and model analysis stages. Since the pre-processing steps can have a great impact on the dataset as a whole, great care was taken to ensure that the processing did not reduce the number of rows, or lose any information. This was done in the same way as the data workflow. When it came to developing the models, each relevant model had metrics like accuracy and mean squared error applied to their training. Additionally, roc-auc curves were used to validate those metrics and make sure that there was no overfitting. Finally, when looking at model outputs, predictions, and Monte Carlo processing, descriptive statistics were used to make sure that the model was creating predictions within the expected range.

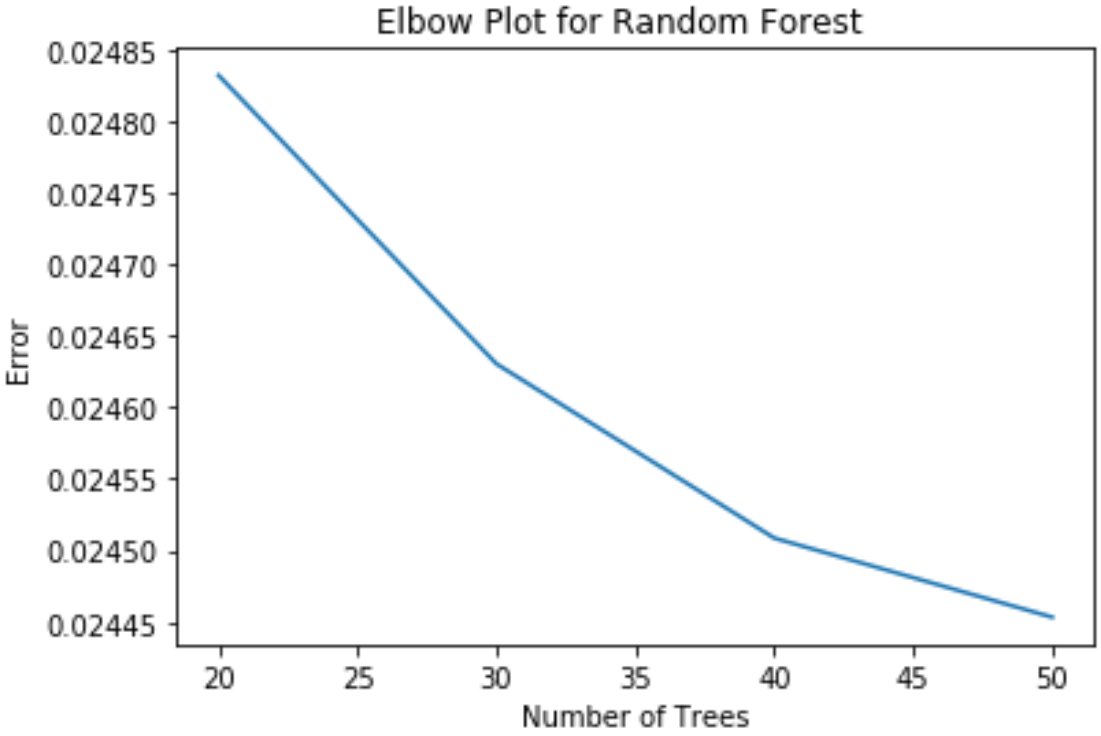

Figure 2: ROC-AUC Curve for Random Forest

There were also validation steps for the entire workflow. The intent of this project is to create reproducible workflows using our input data so that other data users could recreate it. For that reason, it was important to make sure the most important deliverable, the data workflow and process, was effective and usable. When changes were made to the process, the entire subsection of the workflow would be rechecked and rerun from its input files to make sure that no breaking changes were made. Additionally, there were several times where the entire workflow was run from start to finish to ensure that it worked. This was incredibly time consuming but helped ensure that the project was sure to be successful.

3 Presentation/Visualization

3.1 Prepare presentation / visualization of both the data (and any meta-data, information) and the results of the analysis and describe them, Min 2-3 sentences.

Response: Data Analysis is the process of analyzing, cleaning, manipulating, and modeling data with the objective of identifying usable information and informing conclusions.The final merged dataset was analysed using a number of tools.

- Identification of the missing data and the outliers: The dataset was inspected to find any of the missing values. The isnull() [isn19] function in python was utilized. There were no missing values in the dataset. The size of the dataset initially was **73476052**rows. To obtain statistically significant results, we eliminated the outliers. The boxplot was plotted for every feature to see if outliers exist. The points outside the whiskers of the box plot were considered to be outliers. The plot of the features is shown in the figure 3. We can see that there are many points beyond the whisker and they must be eliminated before we proceed further. The outliers were eliminated

using the Winsorize method [CKSC00]. The effect of the outliers was reduced by replacing the largest and smallest value with observations closest to them. We could have trimmed the outliers but that would make us to lose datapoints (dataset would be biased). The boxplot of the features after the elimination of the outliers is shown in the figure 4. We can see that after this step, there are no points beyond the whisker.

- We plotted the distribution of the data after the elimination of the outliers. We can see that the features elevation and slope almost share similar distribution here. This is because the slope was computed based on the elevation values. The figure 5 shows the plot of the distribution of the numerical features.

- Next, we wanted to see which are the features that are strongly correlated to the landslide probability. Correlation computation was very much required to find meaningful information,interdependence and connection between the various features and landslide probability. It would provide us good insights even though the features were extracted from different datasets. The correlation was visualized in the form of matrix and heatmaps. The python's function corr() was used to plot the pairwise correlation between the columns using Pearson method. The plot of the matrix is shown the figure 6. From the figure, we can see that there is a strong correlation between the precipitation and landslide probability. We also see that it is a positive correlation. As the rainfall increases (precipitation increases), we will be seeing higher landslide probability. We also see that there is a positive correlation between elevation and landslide probabilities. The higher the elevation, higher is the frequency of landslides. To achieve, better visualization we have also plotted the heatmap as shown in the figure 7. The darker the color, the higher is the correlation of the feature with the landslide probability.

- Since the landslide probabilities were continuous numeric data, we had to bin them for further plotting and analysis. We utilized the pandas cut function [cut21]. Three different bins were created as follows in the following intervals: [(-0.00129, 0.43] < (0.43, 0.86] < (0.86, 1.29] The values that fell within the first bin were labeled as the **"Low Risk"**. The The values that fell within the second bin were labeled as the **"Moderate Risk"**. The values that fell within the last bin were labeled as the **"High Risk"**.

- A violin plot of the highly correlated features is shown in the figure 8.

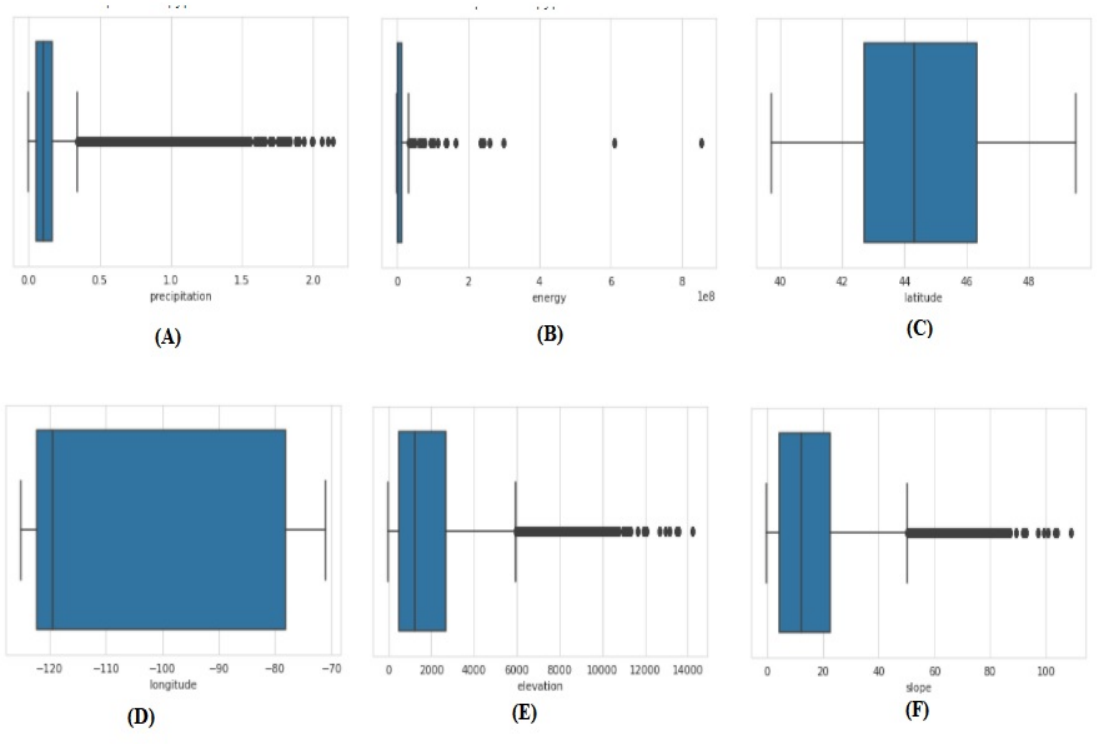

Figure 3: Outliers. (A). Box plot of precipitation. (B). Box plot of energy. (C). Box plot of latitude. (D). Box plot of longitude. (E). Box plot of elevation. (F). Box plot of slope

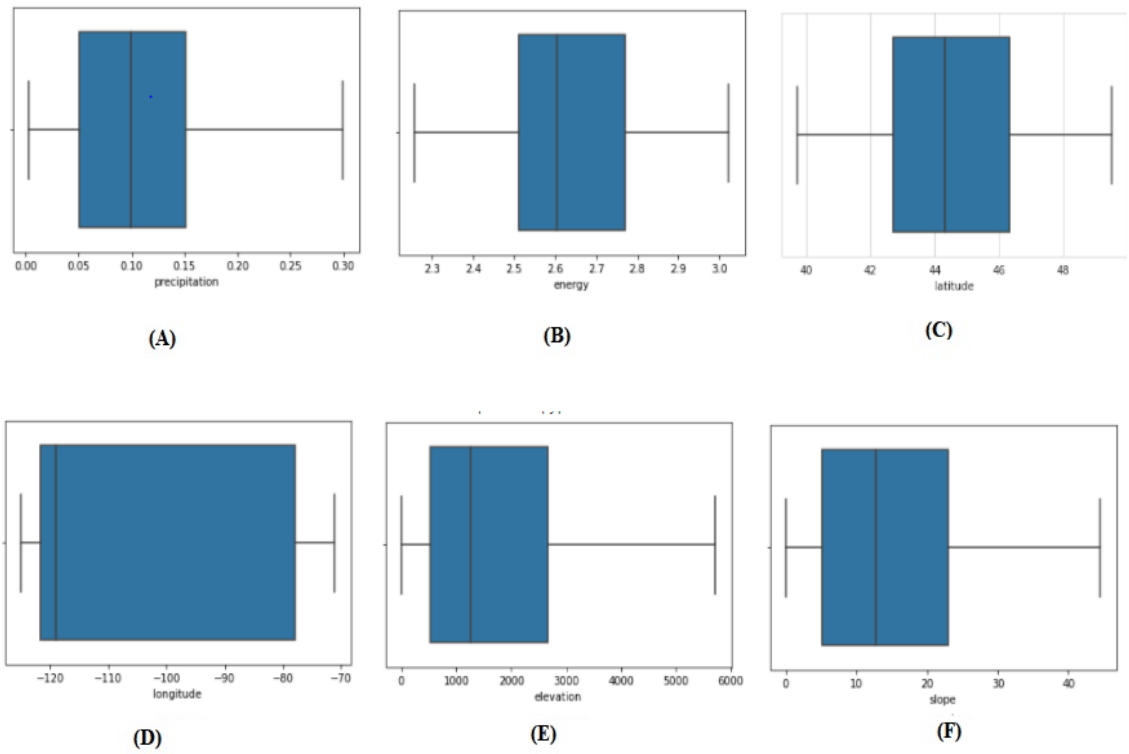

Figure 4: After elimination of Outliers. (A). Box plot of precipitation. (B). Box plot of energy. (C). Box plot of latitude. (D). Box plot of longitude. (E). Box plot of elevation. (F). Box plot of slope

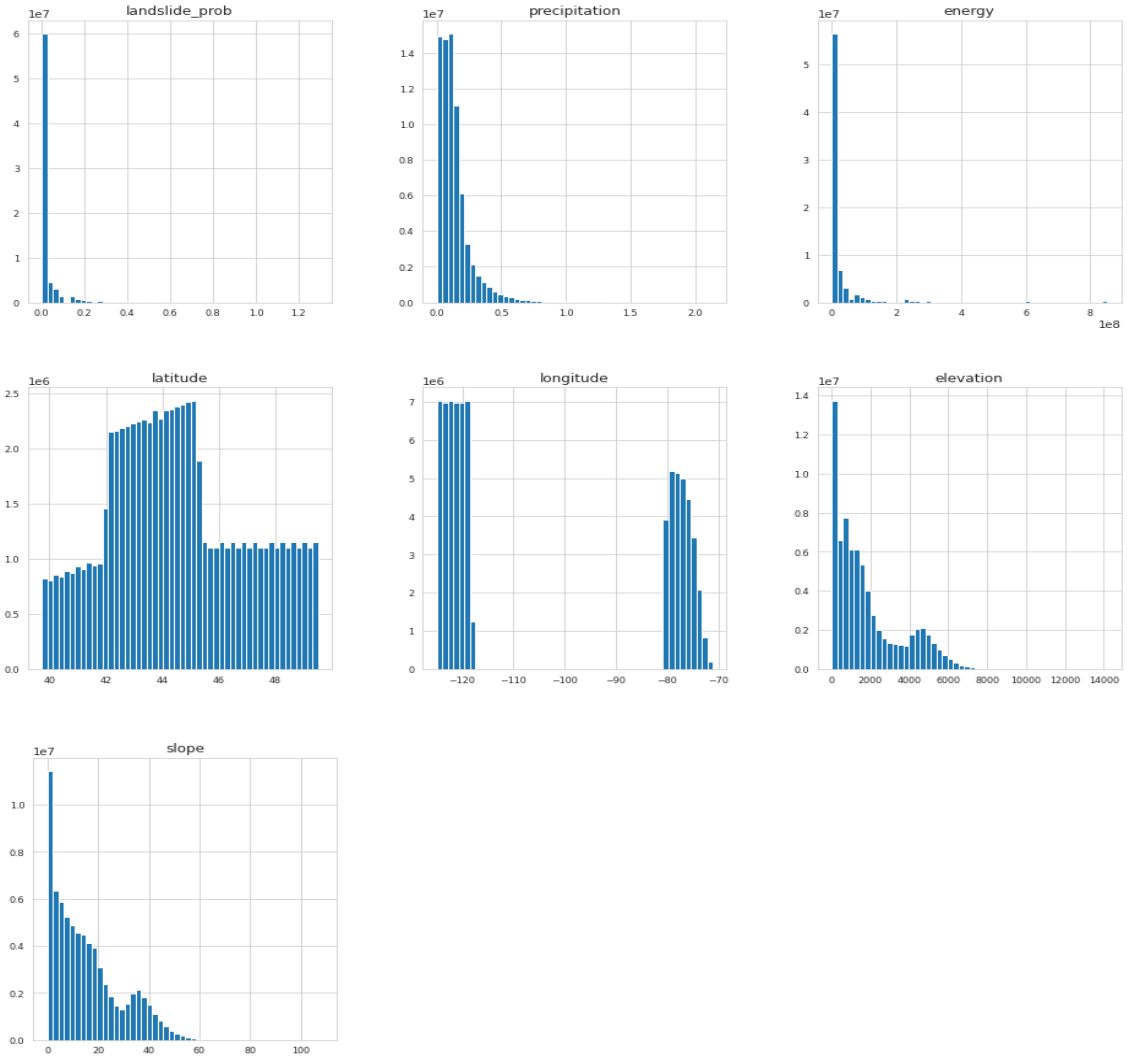

Figure 5: Data Distribution

	landslide_prob	precipitation	energy	latitude	longitude	elevation	run	slope
landslide_prob	1.000000	0.239832	0.063006	0.069174	-0.131648	0.102372	0.129903	0.087323
precipitation	0.239832	1.000000	0.081829	-0.121491	0.272923	-0.266793	-0.274953	-0.233792
energy	0.063006	0.081829	1.000000	0.285470	-0.345650	-0.190577	0.342838	-0.239789
latitude	0.069174	-0.121491	0.285470	1.000000	-0.590008	0.129006	0.617797	0.040858
longitude	-0.131648	0.272923	-0.345650	-0.590008	1.000000	-0.420460	-0.999235	-0.290438
elevation	0.102372	-0.266793	-0.190577	0.129006	-0.420460	1.000000	0.420512	0.984160
run	0.129903	-0.274953	0.342838	0.617797	-0.999235	0.420512	1.000000	0.289989
slope	0.087323	-0.233792	-0.239789	0.040858	-0.290438	0.984160	0.289989	1.000000

Figure 6: Correlation Matrix

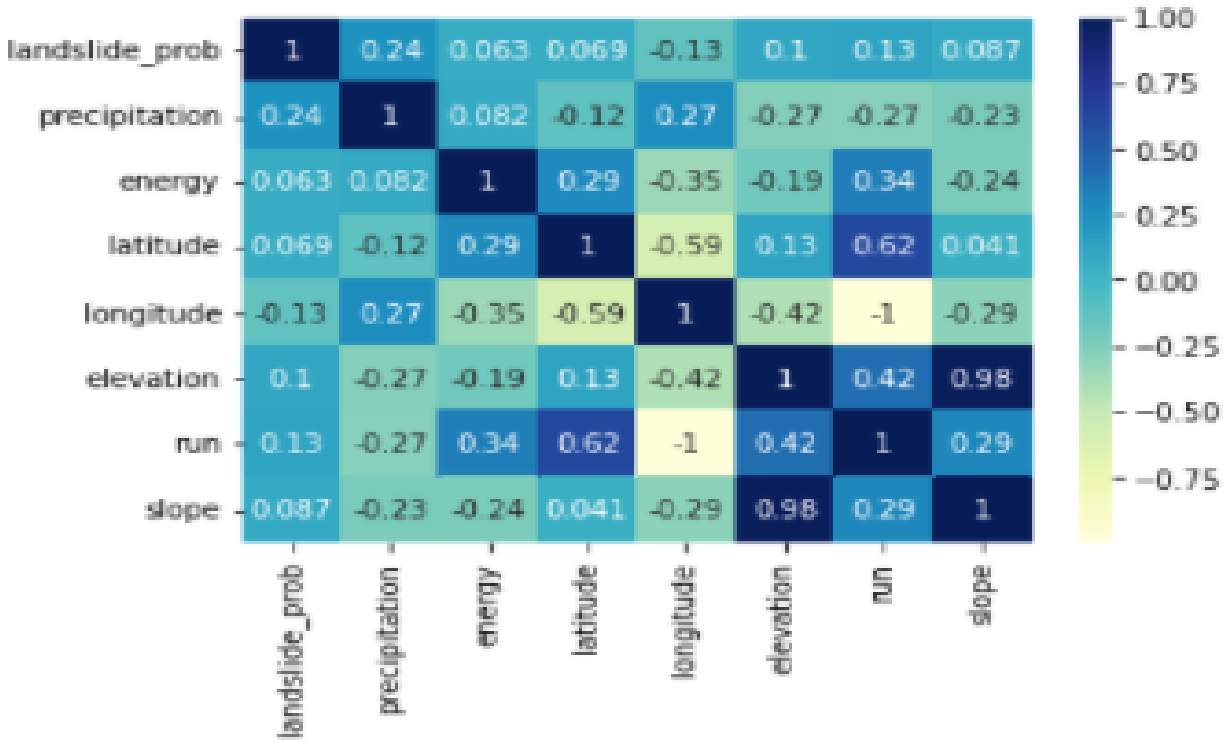

Figure 7: Correlation Heatmap

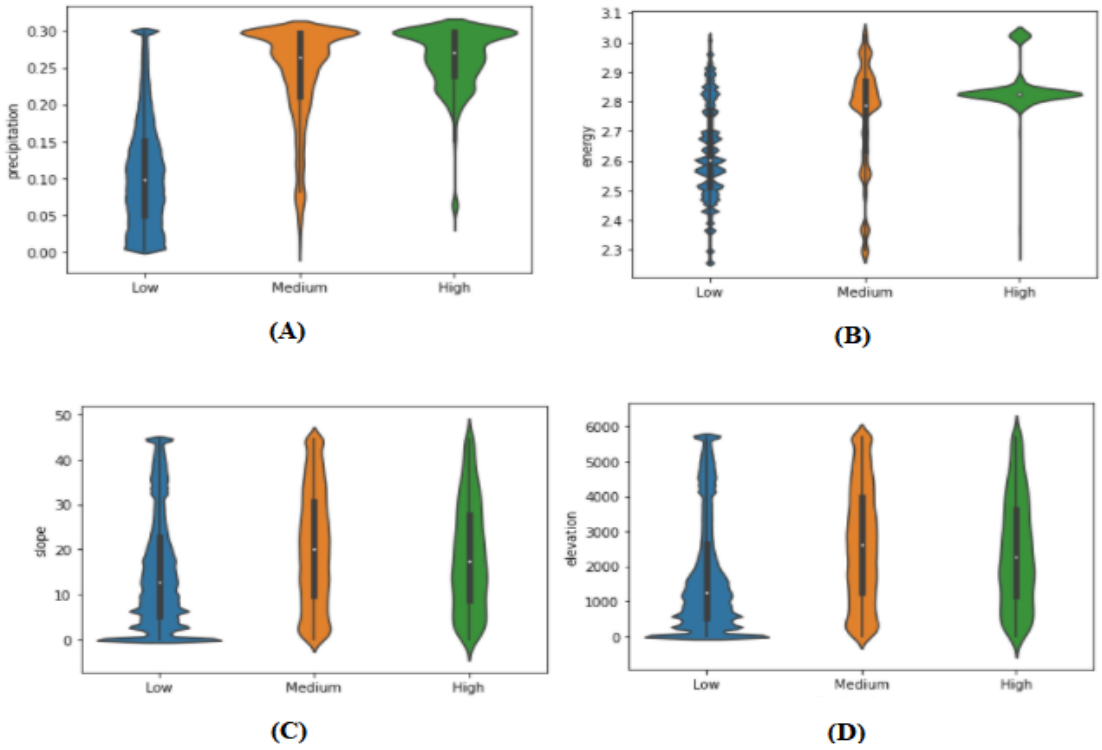

Figure 8: Violin Plot.(A).Precipitation versus Landslide probability. (B). Energy versus Landslide probability. (c). Slope versus Landslide probability. (D). Elevation versus Landslide Probability

3.2 Document the management of the presentation / visualization products and any associated metadata, etc. Min 2-3 sentences

All of our data, including plots, data, poster are uploaded in the course GitHub repository. All of the plots have been given appropriate labelling and titles. Also, legends are there to have correct interpretation. Since all the data is store on the GitHub repository so that long-term storage can be ensured.The storage on Git ensures that the data, visualizations and analysis are archived for future use and enhancements. This will make it easy and effective to recognize and the findings.

We have included the metadata information so that any future user can understand and expand this project. This project also makes sure that any future user, who is interested in the work have proper documentation. This will make the future user can utilize our work effectively and easily.

3.3 Describe how your presentation/visualization supports the goal of the data science investigation and highlight any value that was gained, Min 3-4 sentences

The hypothesis was acknowledged through the visualizations and modeling done on the merged dataset. The results of our analysis are explained through the poster. Through the visualization and presentation, we were able to identify and validate the correlation between the landslide and other factors like precipitation, earthquake, etc. For the future work, additional factors like soil data, vegetation cover, etc. could be added to our merged dataset to improve the analysis and provide more accurate results.

4 Describe your overall data management plan for the results for questions 1,2, and 3 using the 9 categories from assignment 2, Min 1-2 sentences for each category

- **Logical collections:** All the datasets are gathered from NASA's website and other Government organizations. The logical collections used in our project are: Precipitation data, Power Plant data, Earthquake data, Topography data, etc. All these logical collections are cleaned and preprocessed and then they are merged together for the data analysis.
- **Physical data handling:** All the datasets were gathered and stored as .csv files. They are uploaded to Google Drive as a sharing medium among the team members. It is also uploaded to the "Dataset" folder in the Github repository for data archival.
- **Interoperability support:** All the gathered datasets are in the .csv or .h5 format. So, the datasets are easily machine readable through the use of Microsoft Excel, Google Sheets, etc. It is also easily accessible on different platforms. These datasets are therefore easily usable for data analysis in Python or R.
- **Security support:** All the datasets, python code scripts used for data merging, analysis, modeling, etc. are uploaded to the Google drive as well as on the class GitHub repository. All the files are safe on Google Drive from unauthorized access as only the team members have access to it. The files on the Github repository are accessible by anyone who has the repository link, but only team members can make any changes to the files in the repository. Contributors in the GitHub repository can be added by the team members only.
- **Knowledge and information discovery:** The project data and files can be accessed from the GitHub repository. Anyone who wants to explore or work with the project's datasets, analysis, etc. can clone the repository and build upon it. The correlation between landslide data and other factors can be built upon for future study as a part of Information Discovery.
- **Data ownership:** All the datasets were collected from NASA's and U.S. Government websites and so the owner of the datasets are NASA and U.S. Govt.. Besides that, all the project code, .ipynb files, and processed merged dataset are owned by the RPI. These all project data are accessible by the Professor and RPI through GitHub and LMS.
- **Data distribution and publication:** The correlation between the landslide data along with other contributing factors can be of great help to the Governmental bodies in prevention of the disasters. Anyone who wants to research more on this, can build more upon the analysis done easily by cloning the project. The analysis done can be used for the journals, research papers, etc.
- **Metadata collection, management:** All the datasets were gathered along with the Metadata and data descriptions. The Metadata standards and conventions required no changes and were easily usable, since they were downloaded from NASA's and U.S. Government websites. These standards were easily usable in the project's analysis with the use of python libraries. The metadata is also kept on the GitHub repository.
- **Persistence:** The project data files are available on the GitHub repository. It has no expiry date and is available until any team member decides to delete it.

5 Create a poster (poster templates are available on LMS). Please submit your poster on LMS using the same naming scheme mentioned above. Mandatory peer-evaluation form must be submitted within 12 hours of the final project submission on LMS in order to receive the presentation and class participation grade.

We submitted our poster in LMS and uploaded it in the GitHub repository.

References

[CKSC00] R Chambers, P Kokic, P Smith, and M Cruddas. Winsorization for identifying and treating outliers in business surveys. In *Proceedings of the Second International Conference on Establishment Surveys*, pages 717–726. American Statistical Association Alexandria, Virginia, 2000.

[cut21] Pandas.cut() method in python, Jul 2021.

[gps] gpsvisualizer. elevationdata. https://www.gpsvisualizer.com/elevation.

[HIF] HIFLD. powerplantdata. https://hifld-geoplatform.opendata.arcgis.com/datasets/power-plants/.

[isn19] Python: Pandas series.isnull(), Feb 2019.

[Mun] Thilanka Munasinghe. Presentations shared in class.

[NAS] NASA. precipitationdata. https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGM_06/summary.

[USG] USGS. Earthquakedata. https://earthquake.usgs.gov/earthquakes/search/.