Written Assessment Questions – Data Scientist

Data Scientist Shared Certificate

Scenario: The Department of State and USAID Joint Strategic Plan includes goal 4, objective 4.3: *"Protect our personnel, information, and physical infrastructure from 21st century threats"*. The Secretary is asking for assessments about natural disasters and the potential impact on our U.S. Embassy and Consulate personnel and facilities. You have been asked to provide a rapid preliminary analysis of the risk of earthquakes to our overseas missions.

You have been provided a data set from NOAA's National Centers for Environmental Information containing "significant earthquakes" and associated features from 1970 to present.

Additional information about the dataset can be found at: https://data.noaa.gov/metaview/page?xml=NOAA/NESDIS/NGDC/MGG/Hazards/iso/xml/G012153.xml&view=getDataView

A list of current U.S. Embassies and Consulates can be found at: https://www.usembassy.gov/

Question 1: Study Design – Please discuss in detail your approach to this analysis. Study Design may Include how you defined the question, data, methodology, and analytic design or your methodology to data exploration and explanation, the assumptions taken, how you would ensure the integrity and accuracy of any results, and any considerations for communication to a varied audience.

Question 2: Data Analysis – What are key findings found in your analysis? Which embassies are most at risk? Are there groups of embassies that share similar features or profiles with respect to earthquake risk? Were there outliers or areas with data quality issues? Please describe how you conducted your data analysis, including which tools you used, why you chose them, and data preprocessing steps you took, and justify the results of your analysis.

Question 3: Communication – Please describe in detail how you would communicate your results to a diverse, non-technical audience. Provide an executive summary of your analysis for senior management.

Question 4: Hypothetically, if a member of your team completed this work using software or methodology in which you are not proficient, describe how you would validate their work.

Question 1: Study Design – Please discuss in detail your approach to this analysis. Study Design may Include how you defined the question, data, methodology, and analytic design or your methodology to data exploration and explanation, the assumptions taken, how you would ensure the integrity and accuracy of any results, and any considerations for communication to a varied audience.

1.Study Design: Study purpose

Leadership requested a "rapid preliminary assessment," and the Department's strategy for achieving Strategic Objective 4.3 calls for "the ability to share information in real-time to assist in mitigating risks." As such, this study prioritized speed and produced findings within a few hours. By generating a portrait of earthquake risk across our embassies so quickly, we can effectively respond to leadership's urgent request for insight, obtain their input, and pursue deeper inquiry if required.

Research Questions

1. What embassies face the greatest risk of earthquakes, based on the frequency and severity of earthquakes recorded between 1970 and the present?
2. What patterns are seen within the data?

Rationale

Leadership has requested a rapid analysis of the risk of earthquakes to our overseas missions. We should balance speed and rigor to effectively respond. This set of research questions can be answered quickly and responsibly using available data and lays the foundation for a crucial next step: the discussion and development of post-specific mitigation strategies that account for aging infrastructure and the history of earthquakes in the area. We know from the Department's strategic plan that leadership is interested in "post-specific strategies" for mitigating risks to infrastructure and personnel.

Study Design: Methodology

Data

The main sources of data used in our exploratory analysis are:

- **Significant Earthquakes from 1970 to 2022.** This dataset was provided along with the prompt, and appears to have been sourced from NOAA's National Centers for Environmental Information database. The dataset includes information about significant earthquakes including geolocation, magnitude, time and year of event, various data points on the earthquake impact including recorded deaths, injuries and damage caused by the earthquake, as well as other data related to the earthquake event. This is the primary source of risk-related data that we utilized in our exploratory notebook.
- **Embassy Data.** This data was sourced from an open-source repo that tracks locations and contact information for embassies and consulates around the world (https://database-of-embassies.github.io). It should be noted that while this source is

comprehensive, it does appear to be outdated and does not contain some newer U.S-operated embassies. However, we have yet to find a more reliable and readily accessible list of embassies and consulates. For future iterations, we would explore alternative datasets, as well as scraping the US embassy website for location information. Additionally, we could leverage internal US State Department data which may have specific information such as granular

- **Peak Ground Acceleration Data (PGA).** This data was pulled from the api provided by the European Facilities for Earthquake Hazard and Risk (EFEHR). This data contains the PGA score for the locations of the embassies. The PGA score represents the predicted maximum ground acceleration in an area within 50 years, with at least 10% probability. It effectively gives us an idea of how severe a seismic event might be within 50 years for the embassy locations. It should be noted that EFEHR's api appears to have limitations w.r.t retrieving scores for all of our embassy locations. For future iterations, we would utilize a more reliable source for seismic hazard score

Study Design: Considerations for Communication

The decision-making context and audience should drive our communication strategy. Our most senior decision-makers need a concise executive summary that answers their topline question, establishes the credibility of the analysis, and suggests well-reasoned options for next steps. Mid-level staffers may need to dig deeper and should be empowered to build on my analysis and pursue new lines of inquiry. They may need an extended presentation on the study, a dashboard, or my raw code and files depending on the extent to which they want to understand and explore the data.

For a presentation, I would default to this structure for communicating key information, relying on plain language and data visualization to convey insights more effectively:

- Mission problem
- Study overview
- Findings
- Limitations
- Recommendations
- Appendix
    a. Study design
    b. Methodology
    c. Data tables

The goal is to shorten time-to-insight and time-to-action. Depending on the audience's professional experience and the decision at hand, they may tolerate more or less jargon and may be eager to dive into highly technical questions–or avoid them altogether for a discussion about the bigger picture. Ideally, I would have a chance to learn about my audience's needs and proficiency before communicating to them. I often engage stakeholders prior to analysis and probe them about their background, what decision they're trying to make, what thresholds matter or trigger the decision, and other details that illuminate what they need to know and how they want to act. Aligning data storytelling to the larger decision-making context helps ensure a relevant, productive conversation with stakeholders.

Question 2: Data Analysis – What are key findings found in your analysis? Which embassies are most at risk? Are there groups of embassies that share similar features or profiles with respect to earthquake risk? Were there outliers or areas with data quality issues? Please describe how you conducted your data analysis, including which tools you used, why you chose them, and data preprocessing steps you took, and justify the results of your analysis.

Exploratory Data Analysis:

Data Preprocessing, Feature Engineering and Algorithm Development

In the exploratory notebook, we chose to utilize an analytical framework by designing a risk score based on aggregating features of each earthquake up to the embassy level. To do this we first needed to determine which earthquakes in the data correspond to which embassies. In particular, we want to know which embassies are located "close enough" to an earthquake to be within the area impacted by that earthquake. Some initial research indicates that determining this radius for an earthquake is complex and requires more information including factors such as soil composition. However for quick exploratory purposes we used a heuristic sourced online that calculates the radius based on the earthquake's magnitude.

$$earthquake\ radius\ (meters) = e^{(\frac{earthquake\ magnitude}{1.01} - 0.13)} * 1000$$

Once we have calculated the effect radius for each earthquake in the data, we determine if for each embassy which earthquakes are close to it by comparing the earthquake radius to the distance from the epicenter to the embassy. If the epicenter-to-embassy distance is shorter than the earthquake radius (see earthquake 1 in the image below), then we will include this earthquake in the feature aggregation for that embassy. If the distance to the embassy is longer than the earthquake radius (see earthquake 2 below), then we will ignore this earthquake during feature aggregation.

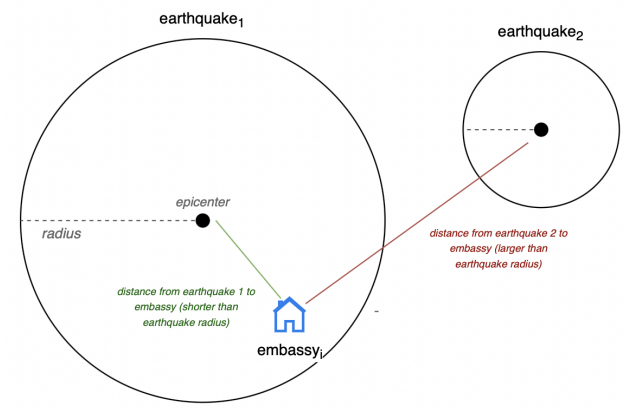


fig. diagram representing how each earthquake is determined to be close enough to an embassy location

Once we have determined which earthquakes are 'close enough' to which embassies, we create a lookup table, called Severity LUT, that effectively maps which earthquakes are close enough to which embassy. The table includes the embassies on the row and the earthquakes on the column, and the cells contain a flag of 0 or 1, where 1 indicates the embassy and earthquake at that cell's coordinates are close, while a 0 indicates that they are not close enough.

	earthquake 1	earthquake 2	earthquake 3	…	earthquake n
embassy a	0	1	0	…	0
embassy b	1	0	0	…	1
…	…	…	…	…	…
embassy z	0	0	1	…	1

fig. Example lookup table for identifying which earthquakes occured nearby which embassy locations

We also create an "adjustment factor" which is meant to scale the value of an earthquake by how far away it is from an embassy. The rationale is that even though we have the heuristic above for determining which earthquake is 'close enough' to an embassy, we don't know how much the embassy would be affected by the full force of the earthquake. The impact of an earthquake (generally) decreases as we are further away from it.

Again, the calculation for this is complex and requires more data and research to accurately determine, so we will create a simple heuristic that scales the value such that it rapidly decreases the further away from the epicenter.

$$severity\ adjustment\ =\ max(\ 0,\ 1\ -\sqrt{\frac{distance\ to\ embassy}{earthquake\ radius}}\ )$$

This produces a factor that equals 1 when the embassy location is exactly at the epicenter, and equals 0 when the distance to the embassy is greater or equal to the earthquake radius.

The adjustment allows us to crudely scale the impact of an earthquake on the area where the embassy is located. For example let's assume the follow for a given earthquake and embassy:

- *earthquake magnitude = 5*
- *distance from earthquake epicenter to embassy location = 75,000 meters*
- *total deaths caused by earthquake = 50,000*

Using the heuristic above to calculate the earthquake radius, we get roughly 124026 meters.
Plugging this into the severity adjustment equation we get *max(0, 1-sqrt(75000/124026))* = 0.222 .
This can then be used to adjust the total deaths, 50000 * 0.222 = 11,100.

Now that we have the lookup table and the severity adjustment, we can create a set of features for each

embassy that aggregate the values for all earthquakes in the effect radius, 'close enough' to that embassy. The features created are:

Total Adjusted Total Deaths - Sum of the *Total Deaths* for each nearby earthquake, scaled using the severity adjustment
Total Adjusted Total Missing - Sum of the *Total Missing* for each nearby earthquake, scaled using the severity adjustment
Total Adjusted Total Injuries - Sum of the *Total Injuries* for each nearby earthquake, scaled using the severity adjustment
Total Adjusted Total Damage ($Mil) - Sum of the *Total Damage ($Mil)* for each nearby earthquake, scaled using the severity adjustment
Total Adjusted Total Houses Damaged - Sum of the *Total Houses Damaged* for each nearby earthquake, scaled using the severity adjustment
Total Adjusted Total Houses Destroyed - Sum of the T*otal Houses Destroyed* for each nearby earthquake, scaled using the severity adjustment
Average MMI - The average *MMI (Int)* for each nearby earthquake
Num Earthquake - The count of nearby earthquakes to the embassy
Num Tsunami - The count of nearby earthquakes to the embassy that are associated with a tsunami
Num Volcano - The count of nearby earthquakes to the embassy that are associated with a volcano
PGA - Represents the predicted Peak Ground Acceleration with 10% exceedance within 50 years for the area of the embassy. Effectively, it provides the expected max strength of a seismic event within the location with at least 10% probability of occurring.

Note that the 'Total Adjusted' metrics above follow the following eq:

$$total\_adj\_metric(EM_i) = \sum_{j=0}^{n} metric(EQ_j) * severity\_lut(i,j) * severity\_adjustment(EM_i, EQ_j)$$

where,
EM_i = embassy i
EQ_j = earthquake j
$metric(EQ_j)$ = the metric value for earthquake j. The metric is whichever we are calculating it for a given feature, e.g. for Total Adjusted Total Deaths, the metric is the Total Deaths column provided in the original data

Once each feature is aggregated to the embassy level we need to determine how to combine these into a single risk score. For future analysis we propose that not all of the features carry equal weight when determining risk, and so we would collaborate with Subject matter Experts to determine what the relative weighting should be. However for quick exploratory analysis we will assume a naive average weighting - that is, that each feature calculated above will be equally weighted in calculating the final risk score.
Also, In order to avoid skewing the scores because of a difference in the magnitude of the values, we will standardize each feature across all embassies, so that the values for each standardized feature is between 0 and 1.

$$Risk(EM_i) = \sum_{j=0} \alpha_j * std\_feature_j(EM_i)$$

where,
EM_i = embassy i
$std\_feature_j(EM_i)$ = value of the standardized feature j, for embassy i. The features are those listed above in bold.
α_j = weighting for feature j

Note that as mentioned, we are taking the naive assumption that the weights for all features are equally, therefor α = 1 for all features.

This means that the final risk score for each embassy is just the sum of each of the 7 features we calculate above.

Unsupervised Machine Learning: Embassy Earthquake Data Clustering

This study analyzed the similarities of the embassies, w.r.t. the earthquake data, by performing clustering of the features calculated in the above section. We chose to not include other embassy-related features (such as raw location, embassy vs. consulate, etc), as our focus was to understand how the embassies group based on the features that we believed were relevant to earthquake-related risk. However we do explore afterwards how clustered embassies group geographically.

We chose to use agglomerative clustering as it tends to work well for relatively small datasets and feature sets. We considered other clustering methods but ultimately dismissed them for various reasons. K-means -based methods perform poorly when the natural clusters in the data are of varying size and density. Density-based methods such as DBSCAN and OPTICS are robust to varying cluster density and can automatically determine the number of clusters, however they will classify some points as outliers. Since we are mainly interested in partitioning all of the embassies into “similar” groups, we decided to use a hierarchical method such as agglomerative clustering, which would assign a cluster to every embassy.

As a preprocessing step to clustering, all features were standardized to remove the mean and scale to unit variance. This step effectively reduces the chance that clustering would be biased towards features with significantly high magnitudes compared to other features.

Four was chosen as the number of clusters after visual inspection of the data. In order to make visualizing the data easier, we performed dimensionality reduction using PCA, and then plotted the first 2 and first 3 principal component scores.

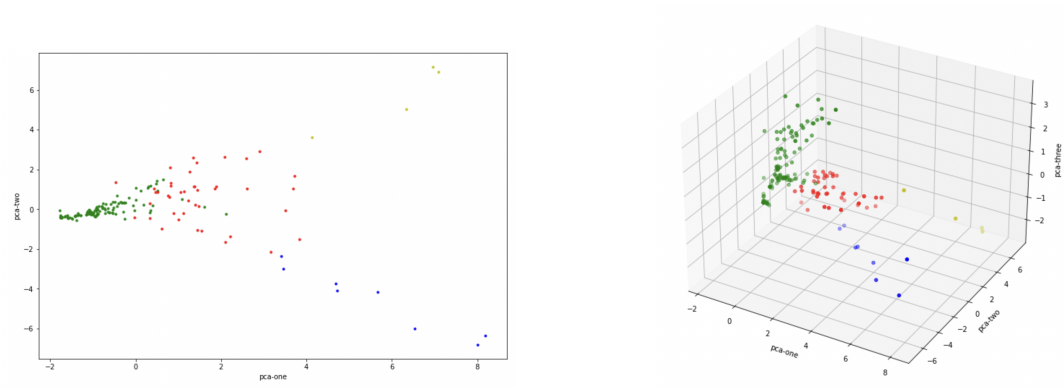


fig. 2d and 3d scatterplots displaying the PCA scores for the feature data. The colors represent the clusters identified using Agglomerative Clustering.

While PCA was used to visualize the data and aid in selecting the number of clusters, it was not used as a preprocessing step in the actual clustering of the data. This was based mainly on the fact that we currently do not have many features in our dataset (11), so our data doesn't currently suffer from very high dimensionality. For future analysis though, we would propose creating and incorporating many more features, in which case dimensionality reduction would be beneficial when clustering.
It should be noted that some of our features are known to be highly correlated, in which case PCA could be a useful approach for decorrelating the data prior to clustering. This is again something we propose exploring in a full study of this task.

The results of clustering showed strong alignment with geographic location of embassies as well as with known fault lines

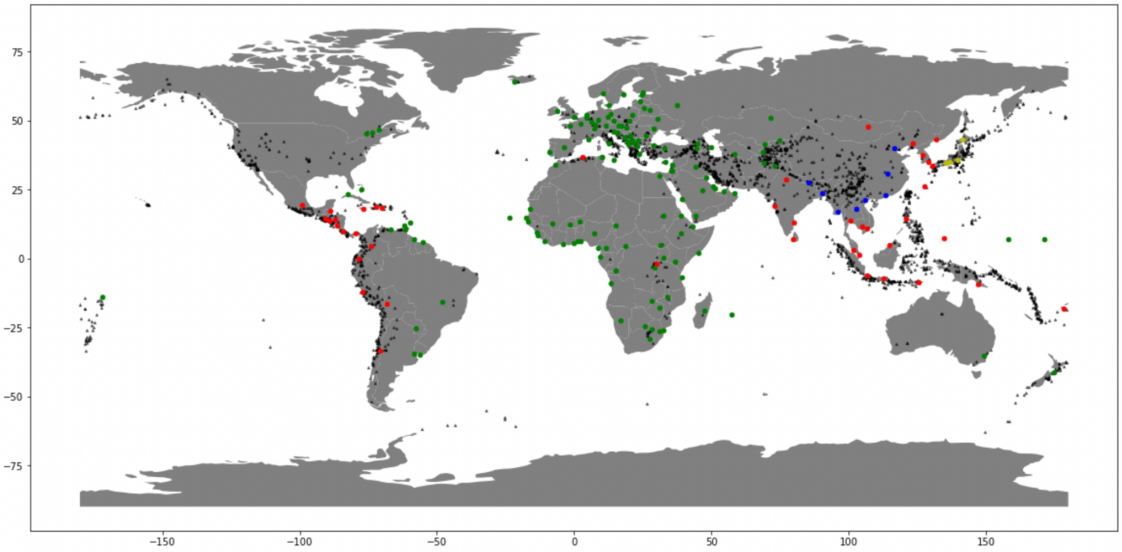


fig. World map displaying the embassy locations (colored circles) along with the earthquake locations from our dataset (black dots). The colors represent the clusters identified.

As seen in the map above, the clusters tend to group into geographic locations:

- Green shows primarily Europe, Africa, Western Asia, some of Oceania, and North America and the eastern part of South America.
- Red shows Southern and Eastern Asian (except China and Japan), Central America, and the western part of South America.
- Blue is focused primarily in and around China.
- Yellow is focused primarily in and around Japan.

When we look at the avg feature values for each cluster we see clear patterns:

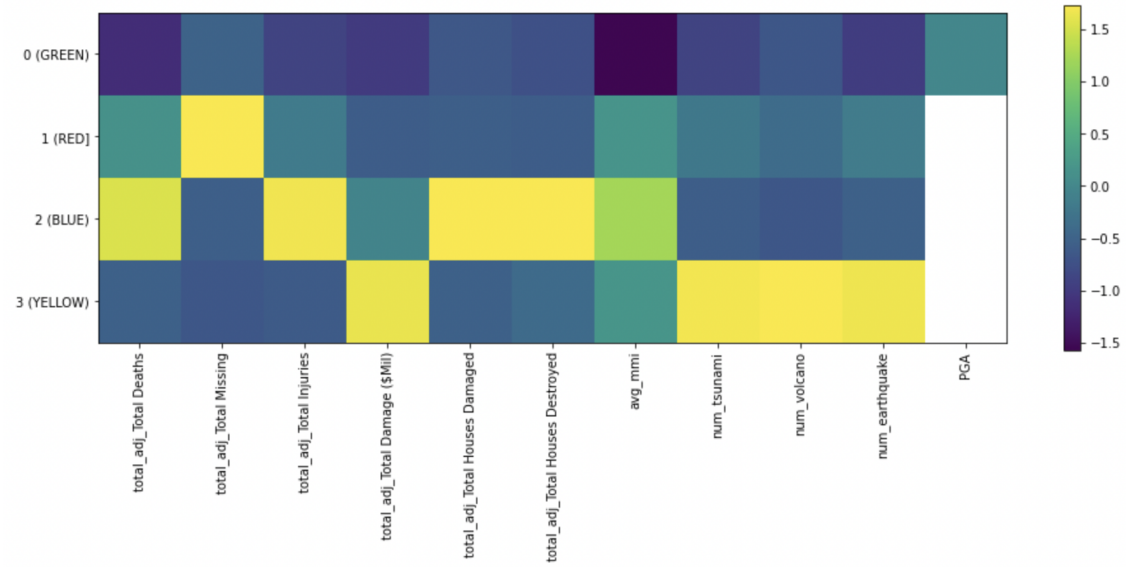


fig. intensity map showing the relative mean feature values for each cluster. Note that a color of white represents indicates null. Note that due to limitations of our source for PGA, we see 3 of the 4 clusters do not have pga scores.

The average values for the green cluster are relatively low across all features. We see low recorded impact (total death, damage, destruction) as well as low seismic activity. This aligns well with the geographic locations for the green-clustered embassies. Note that while we see in the map above a fair amount of seismic activity around the embassy regions in the mediterranean and middle east, the number of earthquakes that were calculated to be 'nearby' those embassies - based on our heuristic for earthquake radius - is relatively low. Of course, there are strong limitations with our heuristic and we would explore more sophisticated and accurate radius calculations for a more in-depth study.

The average values for the red cluster show slightly higher values across the board, compared to the previous cluster, indicating more nearby seismic activity and higher recorded impact. These embassies also tend to be in areas near ocean coastlines.

The average values for the blue cluster are focused around China, and show some of the highest impact for total deaths, injuries and also impact on housing, as well as high average earthquake MMI. However

the number of nearby earthquakes is relatively low, indicating these embassies are in areas with lower seismic activity but with high potential for damage, due to high magnitude earthquakes. Note that these locations may also be in areas with high population density, which would explain the high values for the total death, injuries and housing damage. For a more in-depth study, we would propose incorporating new socio-economic and demographic data that would help us better explore the similarities in our clusters.

The average values for the yellow cluster are focused around Japan, and show an interesting combination of low recorded impact from death, injury and housing damage, but very high damage in terms of overall cost, as well as very high seismic activity along with volcano and tsunami activity.

Manual Evaluation of Ranking Results

total_adj_Total Deaths	total_adj_Total Missing	total_adj_Total Injuries	total_adj_Total Damage ($Mil)	total_adj_Total Houses Damaged	total_adj_Total Houses Destroyed	avg_mmi	num_tsunami	num_volcano	num_earthquake	PGA	country	city	cluster	SCORE
237181.691426	0.000000	619070.123453	117025.853969	3.469689e+06	9.144981e+05	9.142857	7.0	0.0	13.0	NaN	People's Republic of China	Beijing	2	23.019281
140915.896456	0.000000	243976.292954	119015.564964	6.615292e+06	1.703355e+06	8.727273	11.0	0.0	19.0	NaN	People's Republic of China	Wuhan	2	22.348888
44563.824170	0.000000	31244.639038	231537.926139	2.117990e+05	2.094105e+05	7.500000	61.0	2.0	91.0	NaN	Japan	Nagoya	3	21.325659
39716.170834	0.000000	23514.816461	219146.137493	2.295491e+05	1.518125e+05	7.388889	63.0	2.0	96.0	NaN	Japan	Akasaka	3	20.988660
137435.417714	0.000000	107968.201756	85605.778563	5.944660e+06	1.643572e+06	9.200000	8.0	0.0	20.0	NaN	Vietnam	Hanoi	2	18.194553
47745.782731	0.000000	38757.662717	247604.352926	2.040594e+05	2.596769e+05	7.636364	57.0	1.0	83.0	NaN	Japan	Kita-ku	3	17.678802
113415.877255	0.000000	89906.949734	100278.071831	4.813707e+06	1.240869e+06	7.846154	16.0	0.0	25.0	NaN	People's Republic of China	Guangzhou	2	15.391115
139237.852177	0.000000	71376.312038	67616.252448	3.922589e+06	1.147460e+06	9.285714	9.0	0.0	19.0	NaN	Laos	Vientiane	2	12.944484
121587.715174	10.896498	75995.885898	55510.106379	3.749663e+06	1.181066e+06	7.600000	10.0	0.0	34.0	NaN	Bangladesh	Dhaka	2	12.862428
257664.677492	215.461004	251769.355195	7966.701103	2.212114e+05	1.195757e+05	8.250000	7.0	0.0	15.0	NaN	Haiti	Tabarre	1	11.906108

fig. Top 10 embassies based on calculated risk score. Note that the features used in the risk calculation are first standardized prior to calculation. They are displayed above in their pre-standardized values.

Reviewing the top scoring embassies, we see they are primarily in east and southeast asia. In particular we see the majority of the top 10 embassies are just in China and Japan (6 out of 10, combined). It seems there are two main factors for these embassies having the highest calculated risk: having a high impact from earthquakes w.r.t death injuries, damage, destruction, cost, and also having very high seismic activity especially associated with tsunamis.
This aligns with what we found in the cluster analysis, where we two groupings; one with very high impact from death, injury and destruction but somewhat above average seismic activity, while the other showed lower impact from death and injury, but significantly high impact from cost and also a high amount of seismic activity along with tsunamis. We see that those clusters (2 and 3) dominate the top 10 list here, with only a single embassy from a different cluster.
Interestingly the embassy in Haiti - ranked 10th and placed in the 1st cluster - shows one of the highest impact from death and injury, which would explain its position in the risk rankings despite having low impact from cost and relatively low seismic activity compared to the other embassies in the list.

Cluster	0	1	2	3
Embassy count	142	42	8	4

fig. Table displaying the number of embassies assigned to each cluster

Key Assumptions & Limitations

Our assumptions and limitations for this analysis include the following:

Data Missingness

In consideration of the time for this project, we chose in our exploratory analysis to rely primarily on the earthquake data provided in order to define risk for embassies. However the data contains limited information about each earthquake and recorded statistics about its impact related to death, injury, cost and structural damage.
Furthermore, we show in the notebook that much of the data for these metrics is missing, anywhere from 40% to over 80%. This makes it significantly more difficult to generate reliable risk features from this data, as much of it is missing.

We also found that there were no zeroes present for many of the features related to impact-related columns (e.g., death, injury, damage and cost). The concern is that we may not be able to distinguish between a null representing missing data versus a null representing a value of zero. As per NOAA's description of their data representing earthquakes with *"Moderate damage (approximately $1 million or more), 10 or more deaths, Magnitude 7.5 or greater, Modified Mercalli Intensity X or greater, or the earthquake generated a tsunami"*, we might expect to see possible zero values for death or injury. We were unable to find any clarifying documentation for this even after reviewing the data dictionary available on NOAA's website. Being unable to distinguish between missing and zero-valued data could significantly impact the results of our current ranking, as well as any future attempts to address missing data via imputation methods, since some of the null values may in fact be zeros and their values would be artificially inflated via imputation.
Note that in our exploratory analysis we chose to treat null values in the provided earthquake data as zeros when creating our risk features. The exception to this is with PGA, which is a feature specific to each embassy and was left null if we were unable to retrieve it from our source for PGA values.

Data Sources

As mentioned above, we limited ourselves primarily to the provided earthquake data for computing risk features. In future iterations, however, we would propose incorporating more data sources, including:

- **Socio-economic and demographic data** for the regions of each embassy location.

- **Probabilistic seismic hazard scores**. For the exploratory notebook we included PGA scores, however the source we pulled from was unreliable and we were not able to retrieve scores for each embassy location. We also chose to use PGA as it was ubiquitous in most of the literature related to predicting and measuring seismic hazard, and it appeared to be more readily available online compared to other scores. In future iterations we would research more reliable sources for seismic hazard scores, as well as explore alternative scores to PGA.
- **Infrastructure data** for the regions of each embassy location.
- **IDENT RISK 2022 Database,** which is a suite of qualitative and analytical measures related to humanitarian crises and disasters.

Methodology

As mentioned above, our approach to missing data in the exploratory notebook was to replace nulls with zeros. However, for future iterations we would utilize more effective imputation strategies such as stochastic boosting or expectation-maximization approaches, which should be more reliable than our current zero-imputation step.

We also did not perform any other type of adjustment on the earthquake data prior to using it to create the embassy risk features. This includes the option of age-adjusting our features to account for differences due to when the earthquake occurred. One example of this would be to scale the impact-related features (death, injury, cost, damage). We know that the recorded costs in our data are listed as dollar amounts for the given year, and have not been adjusted to reflect current day value. This means that costs for much older earthquakes may appear low, but would be much higher when adjusted for the value today. This ultimately could skew our risk towards earthquakes that occurred more recently, whose cost amounts are closer to today's dollar value. This is also the case with population and infrastructure, as older earthquakes would reflect an impact on populations that are likely smaller than they are today. For infrastructure, it's possible that modern buildings would be able to withstand seismic activity better, and so it is not necessarily the case that adjusting for modern day would result in higher numbers. For future analysis, we would address each of the mentioned examples by incorporating the date related to demographics, infrastructure and dollar inflation, which we would use to adjust the values for each earthquake.

With regards to how we created risk features in our exploratory notebook, we made two major simplifying assumptions in order to attribute earthquake values to each embassy.
The first was how we determine whether an earthquake is near enough to the location of an embassy for it to be relevant w.r.t. the risk to that embassy. We decided to consider an earthquake near enough if the embassy is within the earthquake's radius of effect. Accurately determining this area of effect is complex and requires more data and analysis, and is often not symmetric depending on factors such as depth and soil type. However for the purposes of exploratory analysis we chose to assume a symmetric area of effect based solely on the earthquake's recorded magnitude, using an equation for radius we sourced online. Another key limitation with our equation for earthquake radius is that we were unable to find more information on calculating simplified earthquake radius, and the equation we sourced is not validated.

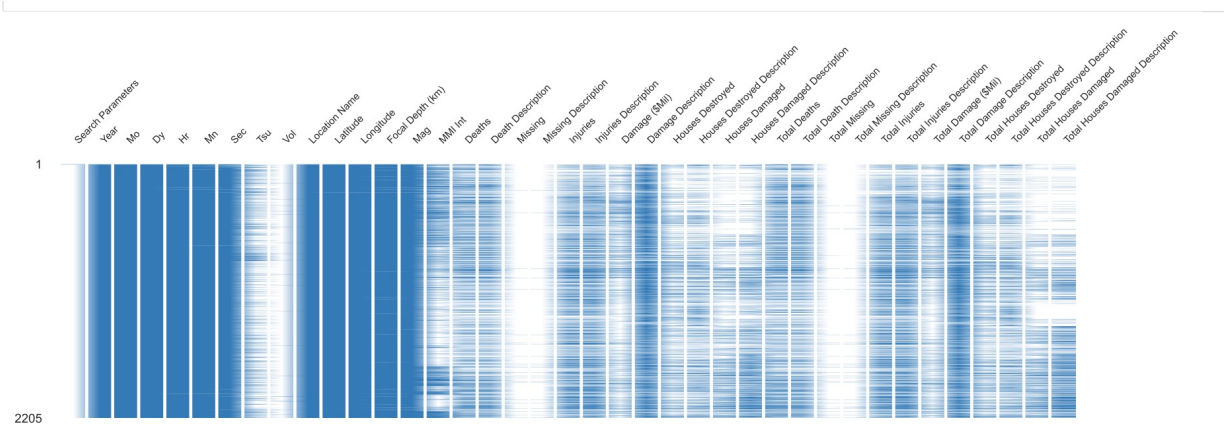
The second assumption used for creating our risk features in determining for earthquakes considered nearby and embassy just how much of its impact should be attributed to the embassy's location. This is another complex problem that requires more data and research to accurately determine. For future iteration we would try to source earthquake values for death, injury, damage and cost broken out by the areas around the earthquake, so we could more accurately determine the counts for the area where the embassy is located. However for the exploratory notebook we chose to assume that the impact of the earthquake drops off rapidly as we move away from the epicenter towards the length of its radius (calculated in the previous step). This assumption that the severity of an earthquake decreases as you move away from its epicenter is *generally* true, however there is no common simplifying equation for calculating this. Instead we chose to create our own scaling factor that we felt approximated our understanding of this effect based on limited research.

Another major simplifying assumption made in the exploratory analysis is to weigh each risk feature equally when calculating the combined risk score. The significance of each feature towards defining risk for our embassies requires a combination of expert opinion as well as stakeholder feedback. Working with SMEs (subject matter experts) we would want to understand which features are most relevant to estimating both the likelihood of a significant seismic event occurring for each embassy, as well as the severity of that event on the safety and operations for each embassy. Incorporating stakeholder feedback would be invaluable in understanding which aspects of risk are most important to them and to our embassies.

Data Integrity

I would ensure the integrity and accuracy of any results by comparing the work to ground truth data. Additionally, to ensure data integrity, I would utilize statistical analysis such as Fisher's exact test for categorical data and Wilcoxon rank-sum test for continuous variables (number of deaths, etc.).

Data Quality - Missingness



The decision-making context and audience should drive our communication strategy. Key limitations include the amount of missing values within our dataset.

Variable	Count	Percentage
Damage Description	203	18.6
Total Damage Description	244	22.4
Total Injuries Description	455	41.8
Injuries Description	465	42.7
Total Death Description	508	46.6
Total Deaths	516	47.4
Death Description	522	47.9
Total Injuries	522	47.9
Deaths	531	48.8
Injuries	531	48.8
MMI Int	595	54.6
Total Houses Destroyed Description	699	64.2
Houses Destroyed Description	710	65.2
Houses Damaged Description	724	66.5
Total Houses Damaged Description	748	68.7
Total Damage ($Mil)	827	75.9
Damage ($Mil)	831	76.3
Total Houses Destroyed	833	76.5
Houses Destroyed	841	77.2
Houses Damaged	859	78.9

Total Houses Damaged	870	79.9
Total Missing	1072	98.4
Total Missing Description	1072	98.4
Missing Description	1073	98.5
Missing	1073	98.5

Table of Missing in Relevant Earthquake Dataset

Relevant Earthquake Data
Number of variables: 44
Number of observations: 1089
Missing cells: 20256 (42.3%)

Tools and Technologies Used

The analysis was performed in a jupyter notebook using the programming language Python. We chose to structure the analysis in a jupyter notebook since this allows us to provide an organized and linear record of the work performed, while also enabling for comments as well as graphics to be embedded in the resulting notebook. Jupyter also allows us to export the notebook as a self-contained html file, which can be shared without the need for python or jupyter to be installed in order to view. Additionally, python is an open-source framework.

Regarding python packages, we utilized the following that are not already part of the python core:

- pandas : Pandas is an excellent package for manipulating and viewing tabular data, which is what we primarily used for this project
- pandas_profiling : This is an add-on that creates comprehensive and easily-consumable profile reports on a dataset. It provides many of the calculations one would perform when evaluating a dataset, including missingness, correlation, descriptive statistics and duplication identification. This was used heavily in helping us understand the nature and limitations of our data.
- geopandas : This package enables us to easily plot our data onto geographic maps. We used this in the cluster analysis to view how embassy clusters were positioned around the world, and view the locations of the earthquakes in our data.
- haversine : This was used as a convenience for calculating the haversine distance between embassies and earthquakes, which was used in the step to determine which earthquakes were nearby which embassy locations
- sklearn: Scikit is a powerful package with many statistical and machine learning methods built-in. We used it in various parts of the analysis, including performing PCA for visualizing our clusters in reduced dimension, standardizing our data as a preprocessing step prior to clustering, and for performing the actual clustering using the AgglomerativeClustering method.

One other tool we utilized was the API provided by the European Facilities for Earthquake Hazard and Risk (EFEHR), which was used to retrieve the PGA scores for the embassies. Unfortunately due to issues

and limitations with their server, we were unable to retrieve scores for all embassies. For future iterations we will consider a more reliable source for PGA and other hazard scores.

Results & Conclusion

Based on the analysis, we have identified the embassies with the highest risk (see the Manual Evaluation of Ranking Results section). We have also observed that the embassies with the highest risk are primarily clustered around Japan, China and the immediate surrounding area. The most significant contributions to risk for these embassies includes for very high impact w.r.t potential death, injury and housing damage for the embassies in and near China, as well as the high potential cost and high seismic activity associated with tsunamis for the embassies in Japan.

Discussion & Future Directions

For the full study, I would propose taking a similar approach to what was done in the exploratory notebook, and utilize a rules-based approach for determining risk. However there are key areas that I would improve upon, namely:

- Incorporate new relevant data sources:
    - Tsunami Data: The earthquake dataset already has a column that allows us to connect an earthquake to a related tsunami. We could then merge the existing dataset with NOAA's tsunami data in order to add new features that might be valuable w.r.t impact from tsunamis caused by earthquakes
    - Political Stability Index:
    - GSHAP: The Global Seismic Hazard Assessment Program produces global probabilistic scores of significant ground movement, effectively helping us understand which areas are at higher risk of an earthquake occurring within a certain amount of time. This likelihood of future seismic events would be a valuable metric to incorporate into total risk score.
    - Population Density
    - Infrastructure age and resiliency data
    - Country indicator level data
    - Geospatial data and image data

Opportunities to improve upon current work include development of additional metrics, for edge cases and further refinement of the heuristics with the guidance of subject matter experts (SMEs) and a variety of stakeholders.

- Best practices - Interventions are deemed effective based on critical review of multiple research and evaluation studies
- Best experiences - Interventions show promise of being effective based on prior experience
- Best processes - Original intervention that you create based on good planning processes, involvement of the target population, and theories

Question 2

Question 2: Data Analysis – What are key findings found in your analysis? Which embassies are most at risk? Are there groups of embassies that share similar features or profiles with respect to earthquake risk? Were there outliers or areas with data quality issues? Please describe how you conducted your data analysis, including which tools you used, why you chose them, and data preprocessing steps you took, and justify the results of your analysis.

(Please see study plan above for full details of the findings from the analysis performed, including data preprocessing steps such as utilizing zero imputation for metrics outside of Magnitude in the final ranking dataset.)

Key findings from the Exploratory Data Analysis

What embassies face the greatest risk of earthquakes, based on the frequency and severity of earthquakes recorded between 1970 and the present?

- Refer to table above
- Embassies did seem to share similar profiles (4 clusters)

What patterns were found in the data?

- Four distinct clusters were found in the embassy data

Outliers in the data

- We found that the embassy in Haiti was in the top10 ranked embassies w.r.t. calculated risk, despite the other 9 in the list being located in Asia, and primarily in East and Southeast Asia. Compared to the other embassies locations in the top10 list, Haiti had relatively low seismic activity and low recorded impact related to cost and damage. However, the Haiti embassy did have one of the highest recorded impact related to death and injury, which seem to be the main drivers for the embassy's placement in the top10 list

Question 3: Communication – Please describe in detail how you would communicate your results to a diverse, non-technical audience. Provide an executive summary of your analysis.

Executive summary

Where do we see the greatest risk of earthquakes to missions overseas?

The Secretary requested a "rapid preliminary assessment" of the risk of earthquakes to missions overseas. This study combines a variety of measures to calculate an earthquake risk score for each embassy and rank them.

Key findings

1. Our ranking algorithm suggests that the top embassies at risk are in East Asia.
2. This work provides a preliminary framework for ranking the top embassies based on risk to earthquakes.

Study overview

- This study represents an overall risk score and ranks embassies utilizing a rules-based system.
- The key limitation of this study relates to data quality utilized for the analysis. This database also has known issues which are described in the details provided on NOAA's description page for the data.

Recommendations

1. **Strengthen the analysis to better target mitigation.** This study prioritized speed to produce a list of high-risk embassies within a few hours. We recommend strengthening the analysis to understand where mitigation may have the greatest ROI on mission outcomes. Ideally, our assessment of risk would account for the activities at each post and their relationship to top diplomatic priorities; the number and types of personnel employed there; and the capacity of each relevant region to recover from an earthquake, as measured by economic resilience in the face of past earthquakes and the population's access to reliable infrastructure, such as utilities. Such an analysis would yield more precise answers on where we face the greatest risk to strategic progress (e.g., where an earthquake could knock out power for weeks to an embassy engaged in information technology work central to one of Biden's signature diplomatic initiatives.)

2. **Consider alerting high-risk embassies and convening a discussion between staff responsible for managing enterprise risk.** The development of post-specific mitigation strategies requires collaboration with post-level staff and enterprise risk management teams. Consider convening key staff to review the findings and discuss mitigation options.

3. **Explore solutions for mitigating the risk of earthquakes to infrastructure and personnel.** A preliminary search uncovered a variety of reports from GAO on how the Federal government can better manage the risks of earthquakes. We could engage GAO or the targets of their earthquake-related audits (for example, U.S. Geological Survey's Earthquake Hazards Program) for recommendations on how to best mitigate the risk of earthquakes. Additionally, we should assess the available evidence on which mitigation strategies work, as demonstrated by research (ideally, rigorous studies that employ quasi- or experimental methods), and analyze how we might apply them to the high-risk posts identified by this analysis.

Question 4: Hypothetically, if a member of your team completed this work using software or methodology in which you are not proficient, describe how you would validate their work.

If a member of my team completed this work using a software or methodology in which I am not proficient, I would consider the following approaches for validating their work:

Depending on time constraints and the particular software used, I may consider quickly learning how to operate the software in order to run, test and evaluate their work. This option would be highly dependent

on the software used. If the work was performed primarily in a coding language that I'm not proficient in then I may be inclined to take this approach, as most of the learning I would need to do would be on the language syntax and the particulars of any packages used in the code.

If I were unable to take the above approach then I would consider trying to reproduce my colleague's work based on their documentation in another programming language or software that I am proficient in. This is a well-established method of validating work via reproduction that I have experience performing. One drawback to this would be the inefficiencies caused by duplication of efforts and extra the time taken to effectively recreate the original work. However, in cases where time constraints are less of a concern and validating the implementation of the original work is important this would be a reasonable approach.

One aspect of the validation process that is independent of the software used would be the reference review. I would make sure to review the references and research cited in the work, and use those to confirm the assumptions and methodologies used in the analysis.

References

1. American Society of Civil Engineers. (2021). *Failure to Act: Economic Impacts of Status Quo Investment Across Infrastructure Systems*. https://infrastructurereportcard.org/wp-content/uploads/2021/02/FTA_Econ_Impacts_Status_Quo-1.pdf

2. Bocquier, P., Beguy, D., Zulu, E. M., Muindi, K., Konseiga, A., & Yé, Y. (2011). Do migrant children face greater health hazards in slum settlements? Evidence from Nairobi Kenya. *Journal of Urban Health, 88*(2), 266–281.

3. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

5. Columbia Edu: Lamont-Doherty Earth Observatory. *Are Seismic Hazard Predictions Effective?* [Poster]. Are Seismic Hazard Predictions Effective? https://www.ldeo.columbia.edu/sites/default/files/uploaded/file/Educational%20Material/2010%20Intern%20Posters/Hannon_poster.pdf

6. EM-DAT Public. (2002). *EM-DAT Guidelines: Data Entry, Field Description/Definition*. EM-DAT: The International Disaster Database. Retrieved July 2, 2022, from https://public.emdat.be/about

7. Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (1996). *Neural network design* (Vol. 20). Boston: Pws Pub.

8. Internal migration to Nairobi's slums: Linking migrant streams to sexual risk behavior. *Health & place, 17*(1), 86–93.

9. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18–22.

10. M. Pagani, J. Garcia-Pelaez, R. Gee, K. Johnson, V. Poggi, R. Styron, G. Weatherill, M. Simionato, D. Viganò, L. Danciu, D. Monelli (2018). Global Earthquake Model (GEM) Seismic Hazard Map (version 2018.1 - December 2018), DOI: 10.13117/GEM-GLOBAL-SEISMIC-HAZARD-MAP-2018.1

11. National Centers for Environmental Information (NCEI). (2022). *Dataset Overview | National Centers for Environmental Information (NCEI)*. National Centers for Environmental Information. Retrieved July 2, 2022, from https://data.noaa.gov/metaview/page?xml=NOAA/NESDIS/NGDC/MGG/Hazards/iso/xml/G012153.xml&view=getDataView

12. Support-vector networks. *Machine Learning, 20*(3), 273–297. Greif, M. J., & Dodoo, F. N. A. (2011).

13. U.S. Department of State. (2022). *US Embassy*. US Embassy. Retrieved July 2, 2022, from https://www.usembassy.gov

city	latit	long	type	cluster	SCORE	INFORM RISK	RISK CLASS	HAZARD & EXPOSURE	Earthquake
Beijing	39.953	116.459	embassy	2	23.01928067	4.1	Medium	6.7	7.2
Wuhan	30.59527778	114.27	consulate general	2	22.34888844	4.1	Medium	6.7	7.2
Tokyo	35.66861	139.74328	embassy	3	20.9886596	2.2	Low	5.4	10
Hanoi	21.0215396	105.8189851	embassy	2	18.19455264	3.7	Medium	5.7	4.1
Guangzhou	23.119444444	113.315	consulate general	2	15.39111495	4.1	Medium	6.7	7.2
Dhaka	23.7967	90.4221	embassy	2	12.86242829	5.8	High	7.4	9.2
Tabarre	18.563861	-72.24909	embassy	1	11.9061079	6.3	High	5.8	9.7
Shenyang	41.783333333	123.4263889	consulate general	1	10.12515849	4.1	Medium	6.7	7.2
Islamabad	33.725	73.117	embassy	0	9.713914893	6	High	7	9.3
Algiers	36.7558	3.0392	embass	1	9.50637	4	Medium	5.1	8.8

			y		2826				
Surabaya	-7.2836546	112.6479533	consulate general	1	9.271955304	4.6	Medium	7.3	8.9