1. Introduction

We chose a real-time earthquake dataset provided by the U.S.Geological Survey (USGS), a reputable U.S. government scientific agency that provides authoritative, open-access data on natural phenomena such as earthquakes, volcanoes, water resources, and land use. The USGS collects this data through a global network of seismic monitoring stations that continuously detect and record seismic activity. Each record in the dataset contains detailed attributes, such as location (latitude and longitude), time of occurrence, and more (detailed below).

The dataset is publicly available and can be accessed through the USGS Earthquake Hazards Program via https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php. The feed provides real-time event data that is updated every minute and can be filtered/chosen per time range (past hour, week, or month), focusing on earthquakes but also including quarry blasts and explosions. We decided to download a static version of “All Earthquakes” under “Per month” and use this single version for our project. This ensures our findings are based on a fixed dataset and allows us to avoid confusion, unmatched data, or any discrepancies.

We are Team A:

- Luna Gulec: 4th year BCom student specialising in BTM (Business Technology Management) with a minor in Data Science. I am interested in global datasets, primarily being able to spot different geographic trends.
- Ericson Ho: 4th year BCS student. I am interested in exploring datasets related to natural disasters
- Logan Wu: 4th year BSc Student specialising in Statistics. I am interested in exploring any data relations and data insights.
- Athena Wong: 4th year BCom student specialising in BTM (Business Technology Management) with a minor in Data Science. I am interested in exploring any datasets.

Our intended audience includes students and the general public, researchers, policymakers, government agencies, emergency response departments, urban planners, and professionals interested in understanding global earthquake activity and trends. We aim to make complex seismic data more interpretable and engaging through multiple visual analytics, allowing audiences without a technical background in geology or seismology to grasp meaningful insights about where and how frequently earthquakes occur, as well as any potential trends within the dataset.

This project demonstrates how open government data can be used for real-world data analysis and visualization. For researchers, government agencies, and policymakers, it highlights possible geospatial or temporal trends that could inform about any gaps in research or further study that could be pursued on seismic risk, preparedness, and effects. Policymakers and urban planners can also benefit from such data, analysis, and visualization by identifying regions with higher seismic frequencies and informing responses for infrastructure and emergency response planning. Therefore, our motivation is to understand global earthquake data and showcase how data visualization can make such scientific information accessible, as well as improve our own skills in data analysis and visualization in Python (specifically Pandas and Altair) and storytelling skills.

---

2. About the Data

A. Data Abstraction

- **Dataset:** USGS Earthquake Data (All Earthquakes, Past Month)
- **Source:** U.S. Geological Survey (USGS)
- **URL:** https://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_month.csv
- **License:** U.S. Public Domain
- **Date Accessed:** October 23, 2025
- **Time Period:** Past 30 days

The original dataset (all_month.csv) contained 8,371 records across 22 attributes. Data cleaning involved handling missing values in measurement-related columns (horizontalError, magError, magNst, dmin, gap, nst) and dropping incomplete rows, resulting in 6,281 complete records.

We also created three new ordinal variables to support interpretability in later analyses:

- **mag_ordinal** → categorizing earthquakes into Minor, Light, Moderate, Strong, and Major based on magnitude thresholds.
- **depth_ordinal** → grouping events as Negative, Shallow, Intermediate, or Deep based on depth in km.
- **gap_level** → classifying azimuthal gap coverage as high, moderate-high, moderate-low, or poor based on quartile cutoffs.

The resulting processed dataset (ordinal_data.csv) serves as the foundation for the following Data Abstraction and Exploratory Data Analysis sections.

Attribute Table

Attribute Name	Attribute Type	Data Semantics	Cardinality
time	Quantitative (Sequential, Temporal)	Timestamp of earthquake occurrence in ISO 8601 format	6,281
latitude	Quantitative (Sequential)	Spatial coordinate (°), range: ~–65.6 to 85.1	5,074
longitude	Quantitative (Sequential)	Spatial coordinate (°), range: ~–179.9 to 179.9	5,265
depth	Quantitative (Sequential)	Depth below Earth's surface (km), range: approx –3.2 to 643.1	3,330
mag	Quantitative (Sequential)	Earthquake magnitude on a standard scale, range: approx –2.1 to 7.6	492
magType	Categorical (Nominal)	Type of magnitude calculation	7
nst	Quantitative (Sequential)	Number of seismic stations that recorded the event	146
gap	Quantitative (Sequential)	Largest azimuthal gap between adjacent stations (°)	319
dmin	Quantitative (Sequential)	Horizontal distance from epicenter to nearest station (°)	4,440
rms	Quantitative (Sequential)	Root-mean-square travel time residual (s)	159
net	Categorical (Nominal)	Network code of contributing organization	12
id	Categorical (Nominal)	Unique identifier for each earthquake event	6,281
updated	Quantitative (Sequential, Temporal)	Timestamp of last data update (ISO 8601)	6,281
place	Categorical (Nominal)	Human-readable text description of location	3,334
type	Categorical (Nominal)	Type of seismic event (e.g., earthquake, quarry blast, explosion)	3
horizontalError	Quantitative (Sequential)	Uncertainty estimate of horizontal position (km)	1,885

depthError	Quantitative (Sequential)	Uncertainty estimate of depth (km)	2,415
magError	Quantitative (Sequential)	Uncertainty estimate of magnitude	2,165
magNst	Quantitative (Sequential)	Number of stations used to calculate magnitude	208
status	Categorical (Ordinal)	Review status: automatic (computer-generated) < reviewed (human-verified)	2
locationSource	Categorical (Nominal)	Network code that originally determined the location	12
magSource	Categorical (Nominal)	Network code that originally determined the magnitude	12
mag_ordinal	Categorical (Ordinal)	Categorized magnitude level: Minor (<4), Light (4–4.9), Moderate (5–5.9), Strong (6–6.9), Major (7+)	5
depth_ordinal	Categorical (Ordinal)	Categorized depth level: Shallow (0–70 km), Intermediate (70–300 km), Deep (>300 km), Negative (above surface)	4
gap_level	Categorical (Ordinal)	Binned azimuthal gap level representing quality of coverage (Low, Moderate, High, Very High)	4

B. Exploratory Data Analysis

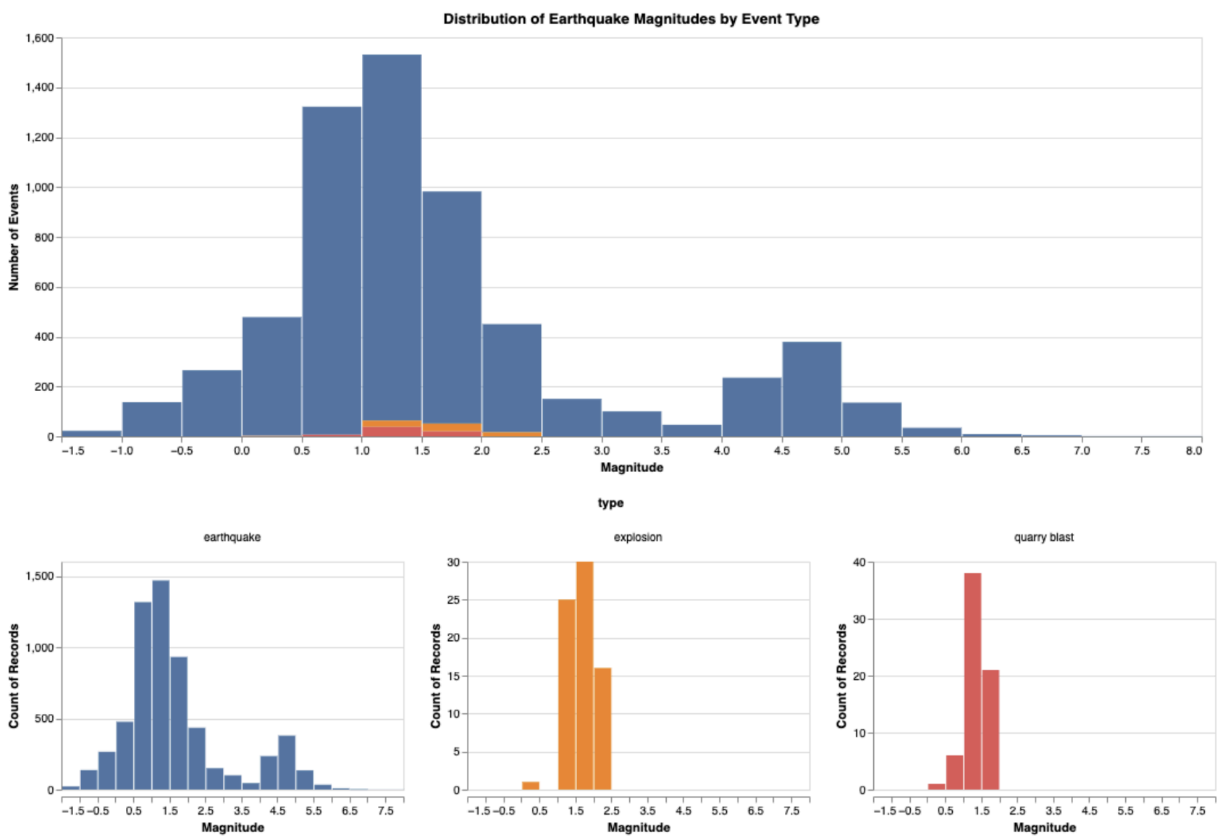
General

Team member	Numerical Attributes	Temporal Attributes	Ordinal Attributes	Nominal Attributes	Type of Analysis
Luna Gulec	depth, magnitude	-	depth_ordinal, mag_ordinal	type, net, magType	Univariate (distribution summaries, histograms, calculations), Bivariate (depth vs. magnitude scatterplot + trendline + correlation coefficient, boxplots), Multivariate (depth vs. magnitude by type and network, mag_ordinal frequency by network, magType, heatmaps)
Ericson Ho	latitude, longitude, depth, magnitude, gap	-	depth_ordinal, mag_ordinal, gap_level	net	Univariate (distribution summaries, histograms), Bivariate (Boxplot net vs mag boxplot and density diagram, latitude vs mag proportional histogram, longitude vs mag proportional histogram, mag vs depth density diagram), Multivariate (latitude vs longitude vs magnitude vs depth plot, mag vs depth vs gap plot, correlation diagram of all attributes, latitude vs longitude vs earthquake count heat map)
Logan	horizontalErro	time	Gap_level,	status	Univariate (), Bivariate (),

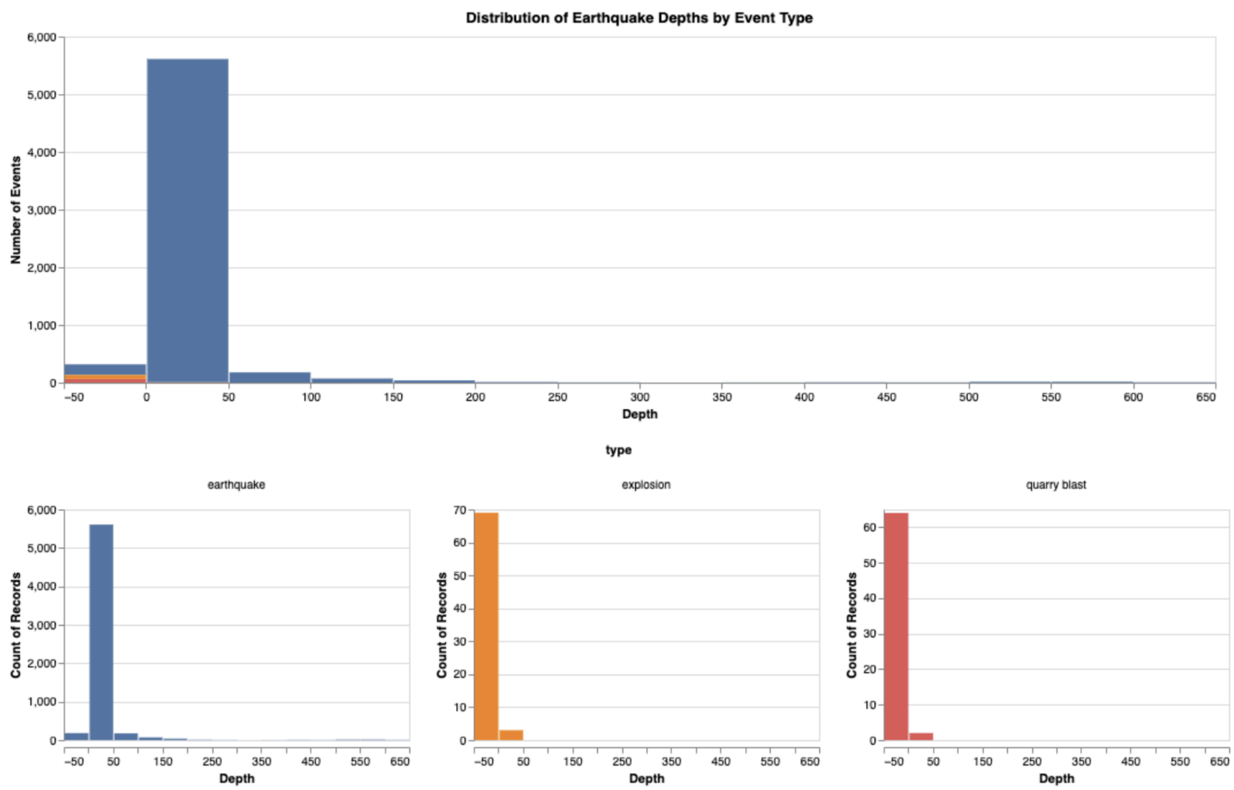
Wu	r, depthError, magError, rms		mag_ordinal, depth_ordinal		Multivariate ()
Athena Wong	nst, gap, dmin, magError, depthError, rms, mag, latitude, longitude	time	gap_level, nst_bin, dmin_bin, mag_level	status, net, region, place	Univariate (distributions of nst, gap, dmin, error metrics via histograms and summary statistics), Bivariate (nst vs magError density heatmap, gap vs nst 2D histogram, monitoring_score vs small_event_rate scatter, error metrics by status boxplots), Multivariate (small event detection rate by nst_bin x gap_level heatmap, magnitude distributions faceted by nst_bin and gap_level, regional comparisons incorporating multiple monitoring dimensions)

EDA 1: LUNA

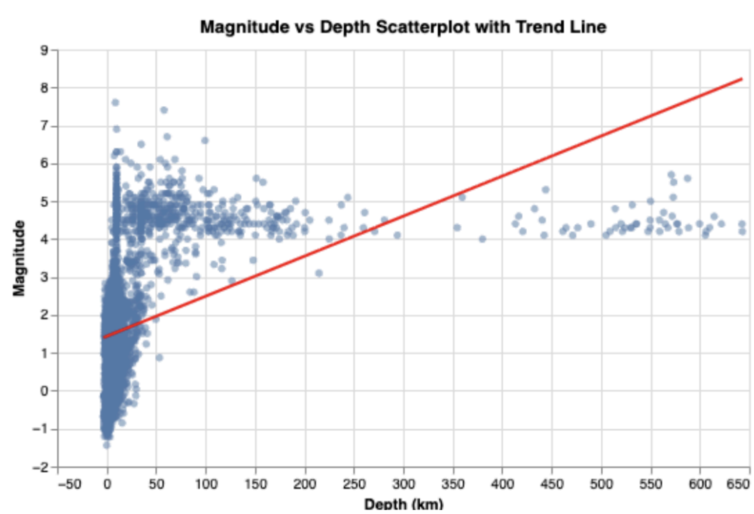
A. Earthquake Magnitude vs. Depth



- Purpose: View the distribution of earthquake magnitudes by event type
- Insight: There is one dominant peak with a magnitude of around 1.5, the magnitude of almost 1,600 events. Faceting by event type with independent scales shows that earthquakes and quarry blasts peak with a magnitude of around 1.5, whereas explosions peak slightly higher near 2.0.



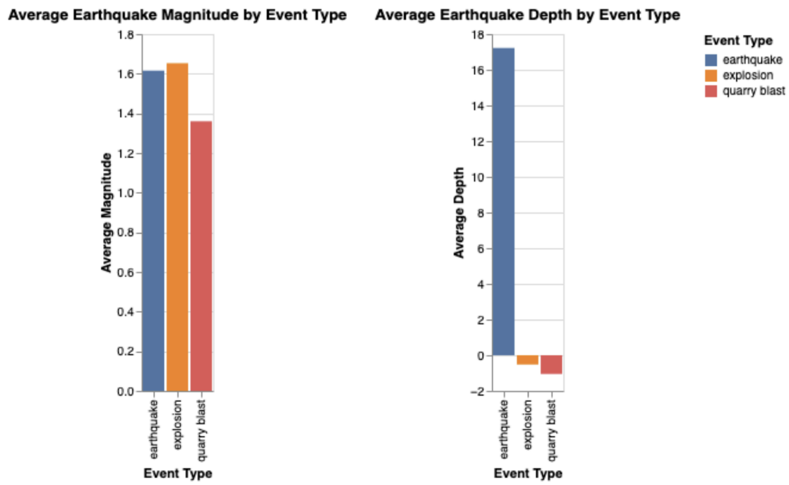
- Purpose: View the distribution of earthquake depths by event type
- Insight: The most common is 0 to 50 depth. After depth = 50, the number of events with increasing depths decreases steadily until depth = 250. There are a couple of events (almost negligible) with depths of 500 to 600 meters.




```
Correlation coefficient: 0.397
```

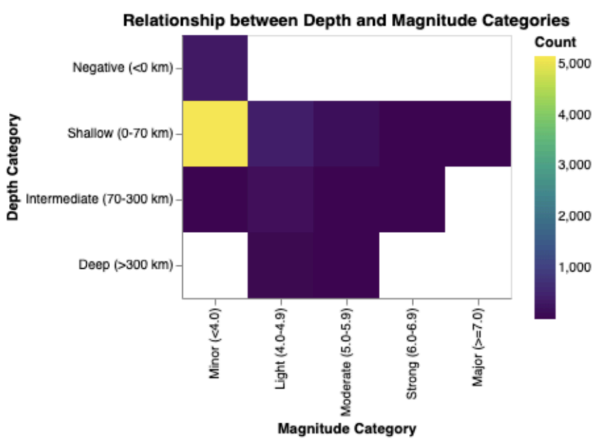
- Purpose: View the correlation between magnitude and depth
- Insight: The correlation between depth and magnitude is 0.397, which is weakly positive.

B. Average Earthquake Magnitude and Depth by Event Type



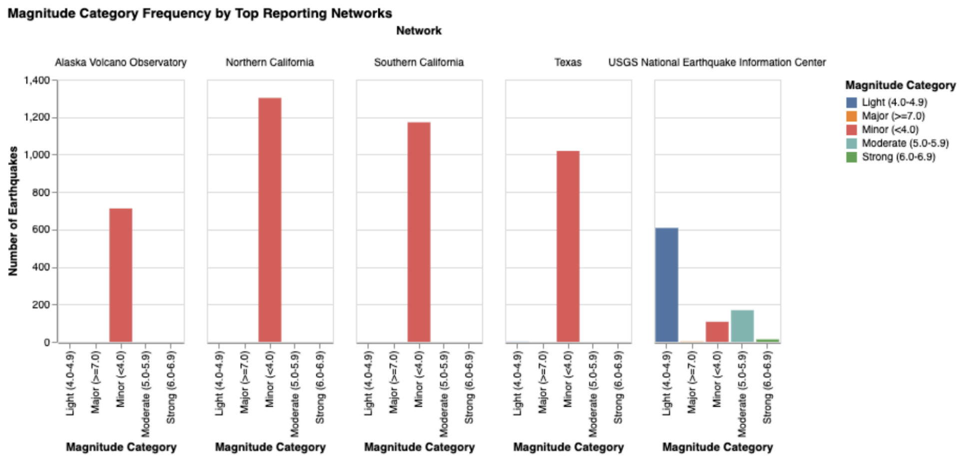
- Purpose: View how average earthquake magnitude and depth differs by event type.
- Insight: Average magnitude is pretty consistent among event types, whereas average depth shows clear differences between earthquakes (much higher) and both explosions and quarry blasts.

C. Magnitude and Depth Categories (Ordinal Analysis)



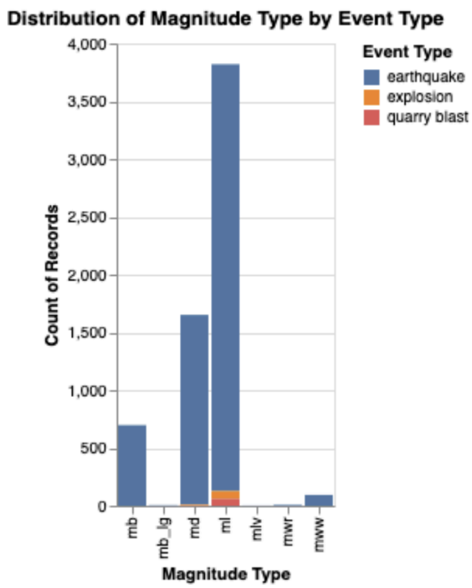
- Purpose: View the relationship between mag_ordinal and depth_ordinal
- Insight: Most events occur in shallow zones (0-70 km) and are classified as Minor (magnitude < 4.0). It is also interesting to note that all Major earthquakes (magnitude >
- = 7.0) are shallow, which contradicts the initial assumption that major earthquakes must be deep and intense

D. Frequency of mag_ordinal or depth_ordinal by top 5 networks

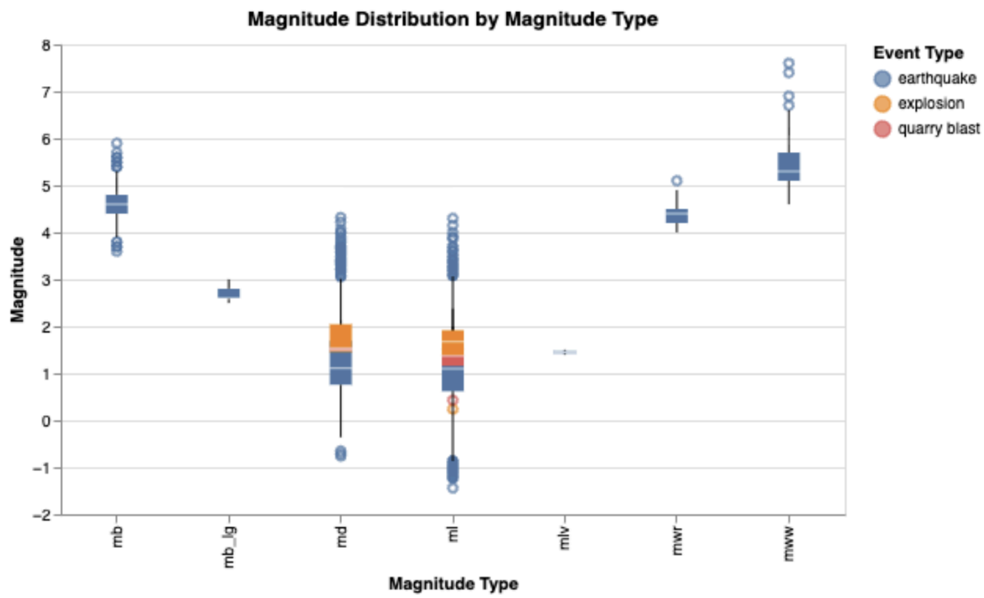


- Purpose: Compare magnitude categories across major networks to see how reporting differs regionally.
- Insight: California networks (NC, CI) have the highest volumes of Minor events, whereas USGS reports fewer but more varying events.

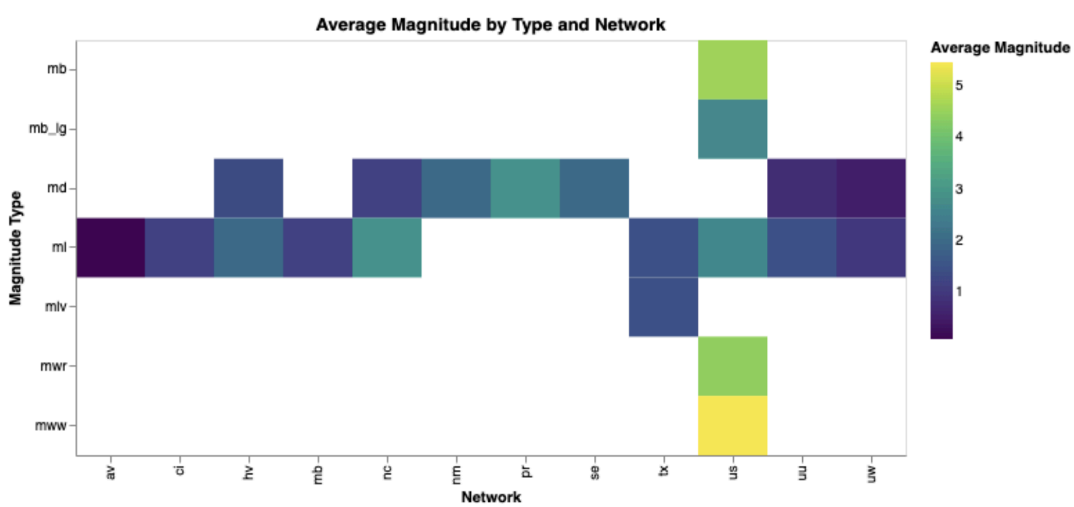
E. Exploring Magnitude Calculations (magType)



- Purpose: View how frequently each magType occurs and for what event type.
- Insight: Most events were reported using local ml and duration md scales, with ml being the most common, which aligns with their definitions as local and duration, as they are based measures for small, nearby earthquakes, and as we saw earlier, more minor earthquakes (smaller magnitude) are more common in this dataset.



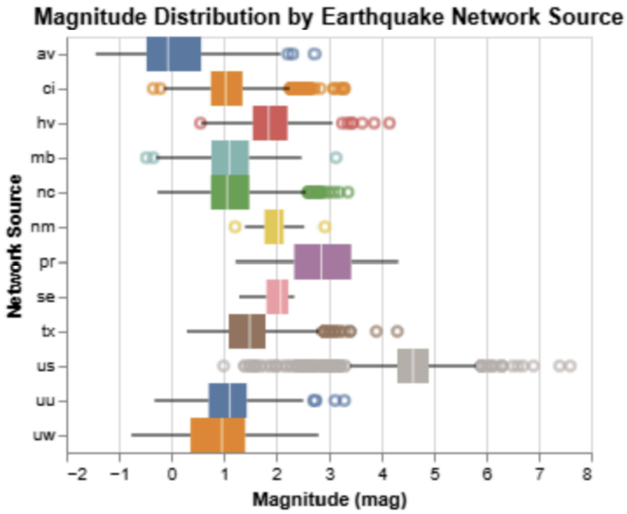
- Purpose: View magnitude distribution by magnitude measurement type, coloured by event type.
- Insight: All are used for earthquakes, but ml and md are the only types used to measure the magnitude of explosions and quarry blasts. Also, md and ml measure the most events, with the biggest range and interquartile range.



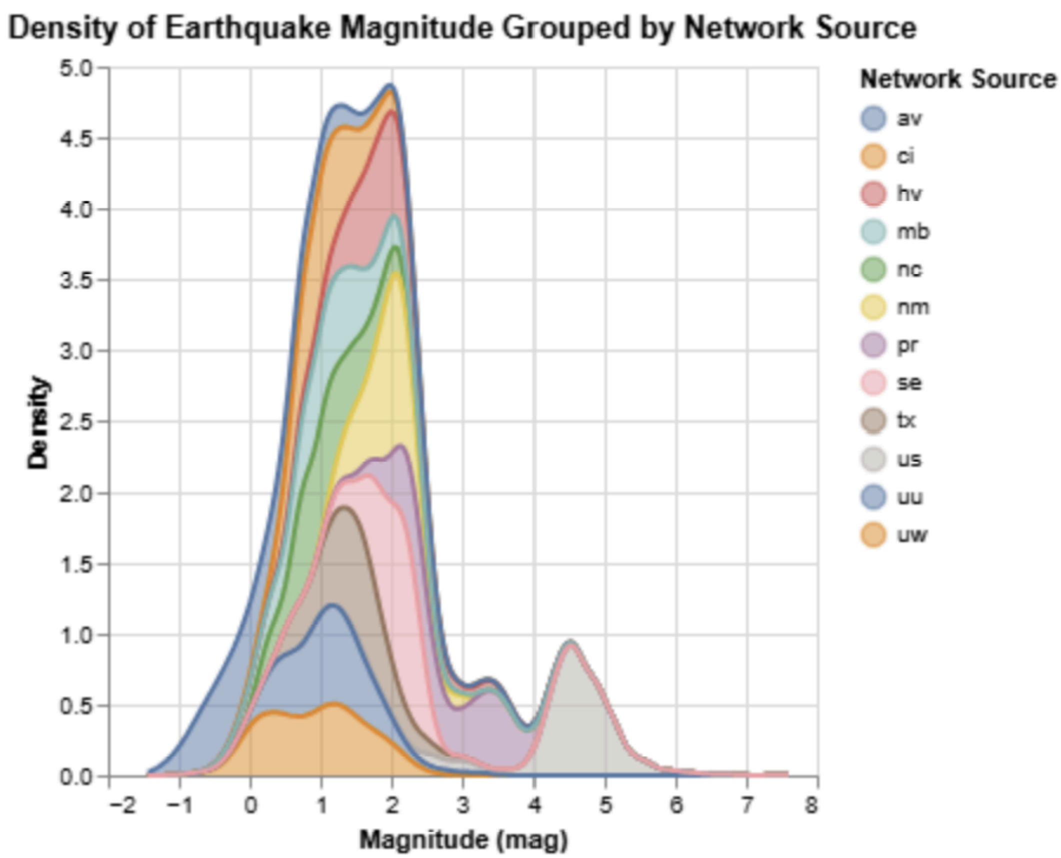
- Purpose: Explore how average magnitude varies across the different recording networks and magnitude measurement types.
- Insight: The U.S. network (us) records some of the highest average magnitudes, particularly for mww, mb, and mwr, whereas local networks such as ci and nc (Northern and Southern California) primarily record lower to moderate magnitudes, typically using local magnitude scales like ml or md. Also, magnitude types ml and md show a broader spread across multiple networks.

EDA 2: ERICSON

A. Earthquake Network



Despite the different locations, each network has pretty similar and consistent results, with the US being the exception. Our interpretation is that the US is likely recording the most earthquakes around the world and that could contribute to more outliers.

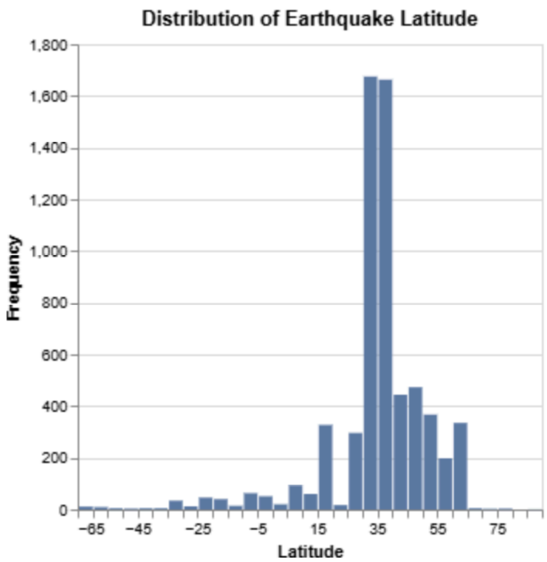


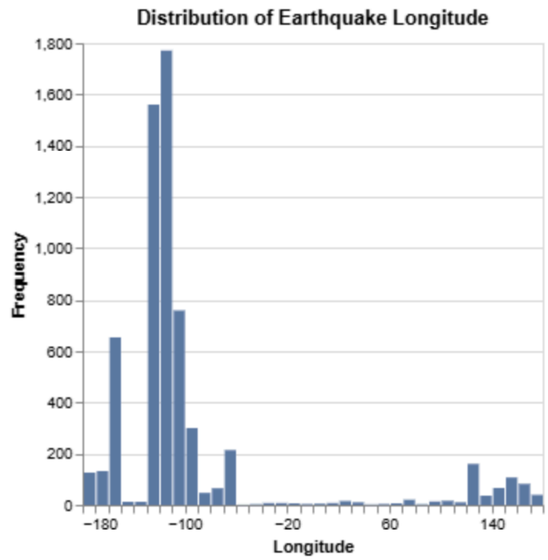
The graph clearly separates the goals of the seismic networks:

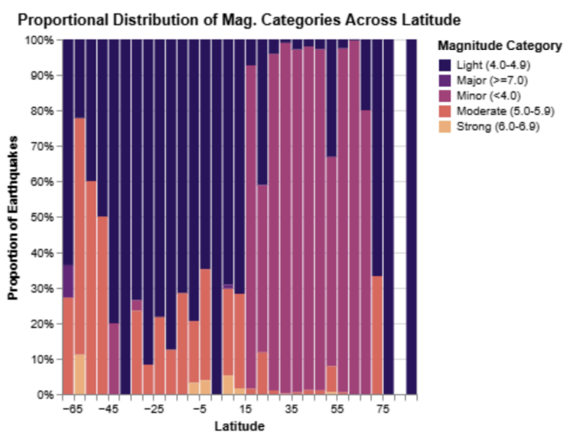
- High-Density Regional Networks: ('ci', 'hv', 'nc', etc.) prioritize detection completeness and high resolution for small-to-micro earthquakes to study local fault mechanics.
- Broad-Scope Global Network: ('us') prioritizes breadth of coverage and the reliable recording of moderate-to-large events that pose a hazard.

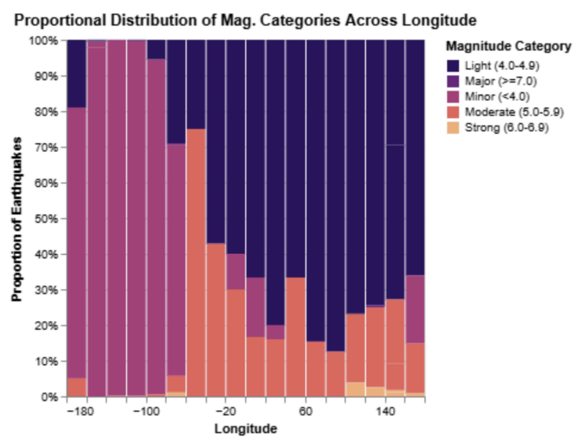
B. Geolocation vs Earthquake Attributes

The following graphs are a collection basic bar charts of the frequency of earthquakes in relation to the respective attributes of interest:



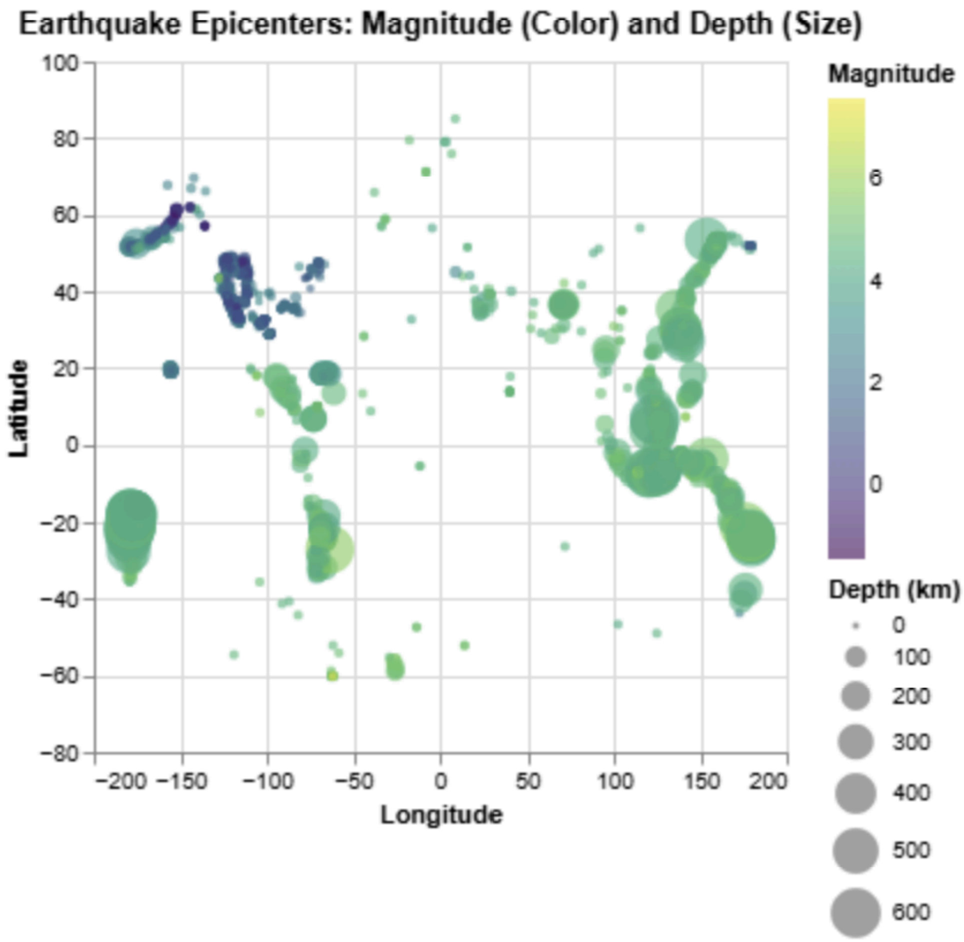




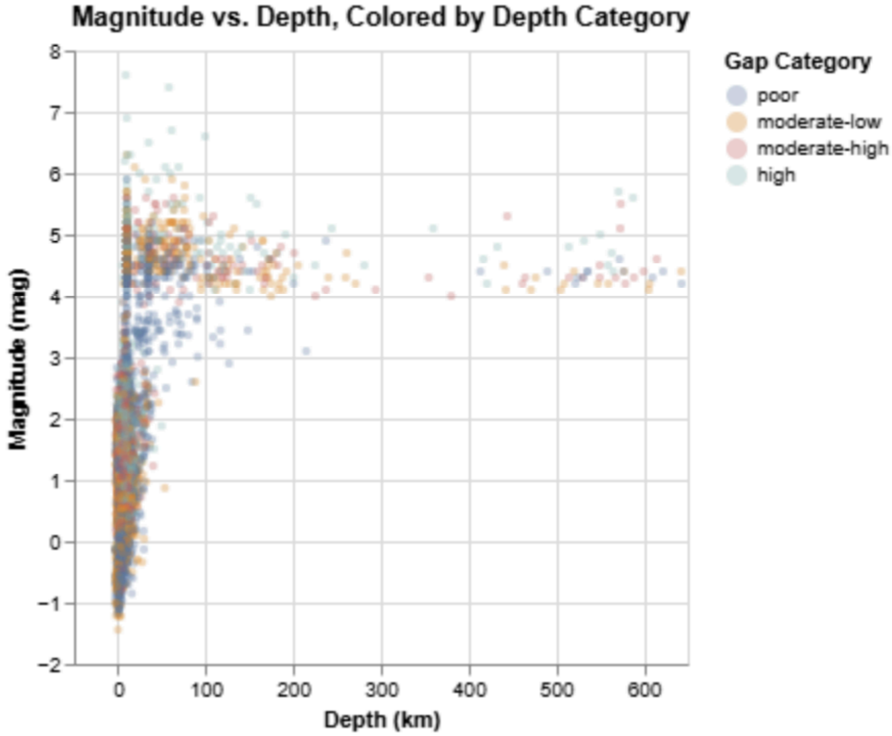


The chart(left) visually demonstrates that while small earthquakes occur everywhere, the most severe seismic hazard, represented by the proportion of Strong and Major quakes, is highly concentrated geographically, particularly at the extremes of the sampled latitude range (high northern and high southern latitudes), corresponding to major plate boundaries like the Circum-Pacific Belt.

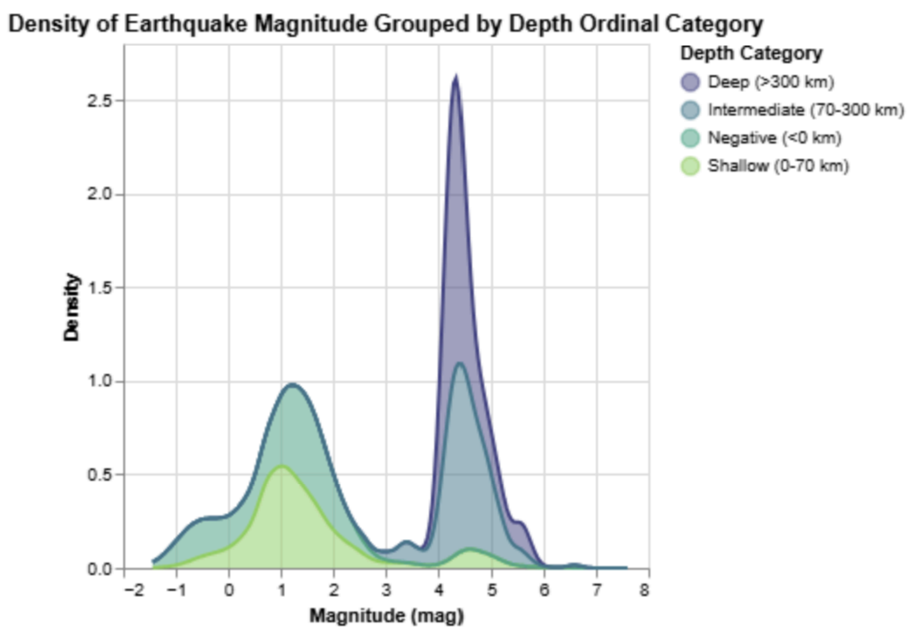
The chart(right) graphically demonstrates that the seismic hazard (the relative proportion of larger, potentially damaging quakes) is concentrated along the Pacific margins (the Ring of Fire), particularly in the Western Hemisphere longitudes, while activity elsewhere is dominated by small, non-destructive events



- Circum-Pacific Belt (Ring of Fire): The most striking feature is the dense clustering of earthquakes along the edges of the Pacific Ocean (roughly Longitude -180 to -70 and 100 to 180). This region is the famous "Ring of Fire," where the majority of the world's large, shallow, and deep earthquakes occur, primarily due to subduction zones.
- Absence in Continental Interiors: There are very few epicenters in the centers of major continental landmasses (e.g., central Africa, central North America, central South America, central Eurasia), confirming that most significant seismicity occurs at plate boundaries
- Mid-Ocean Ridges: A scattered, linear pattern of small, shallow earthquakes is visible in the middle of the Atlantic Ocean and the southern Pacific (e.g., near Longitude -10, 0, and 100). These correspond to mid-ocean ridges (divergent boundaries), which typically produce many small, shallow quakes



The overwhelming majority of the data points are clustered between 0 km and 100 km depth. The density of points drops off dramatically after 100 km. This confirms that most earthquakes, regardless of magnitude, occur in the Earth's brittle upper crust and lithosphere where rocks are cold and rigid enough to store and release elastic strain energy. This is a fundamental concept in seismology.



Shallow (0–70 km)

- Distribution: Bimodal — a large peak at *M ≈ 1.5* (many small local quakes) and a smaller peak at *M ≈ 4.5* (moderate-to-large quakes).

- Significance: Occur in the brittle crust, spanning the widest magnitude range (below 0 to ~7.5). Most destructive earthquakes occur here.

Intermediate (70–300 km)

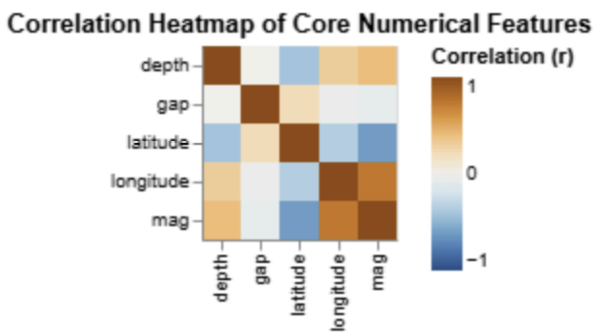
- Distribution: Similar to shallow quakes but dominated by small magnitudes (*M ≈ 1.5*) and contributing to the *M ≈ 4.5* peak.
- Significance: Happen within subducting slabs; capable of small to moderate magnitudes but fewer large quakes than shallow ones.

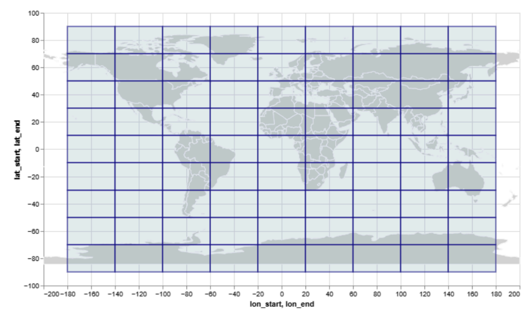
Deep (≥300 km)

- Distribution: Very narrow, strongly peaked around *M ≈ 4.5*. No micro-quakes (*M < 2*), and rarely exceed *M 6*.
- Significance: Large quakes cannot occur due to high temperature and pressure making rocks ductile.

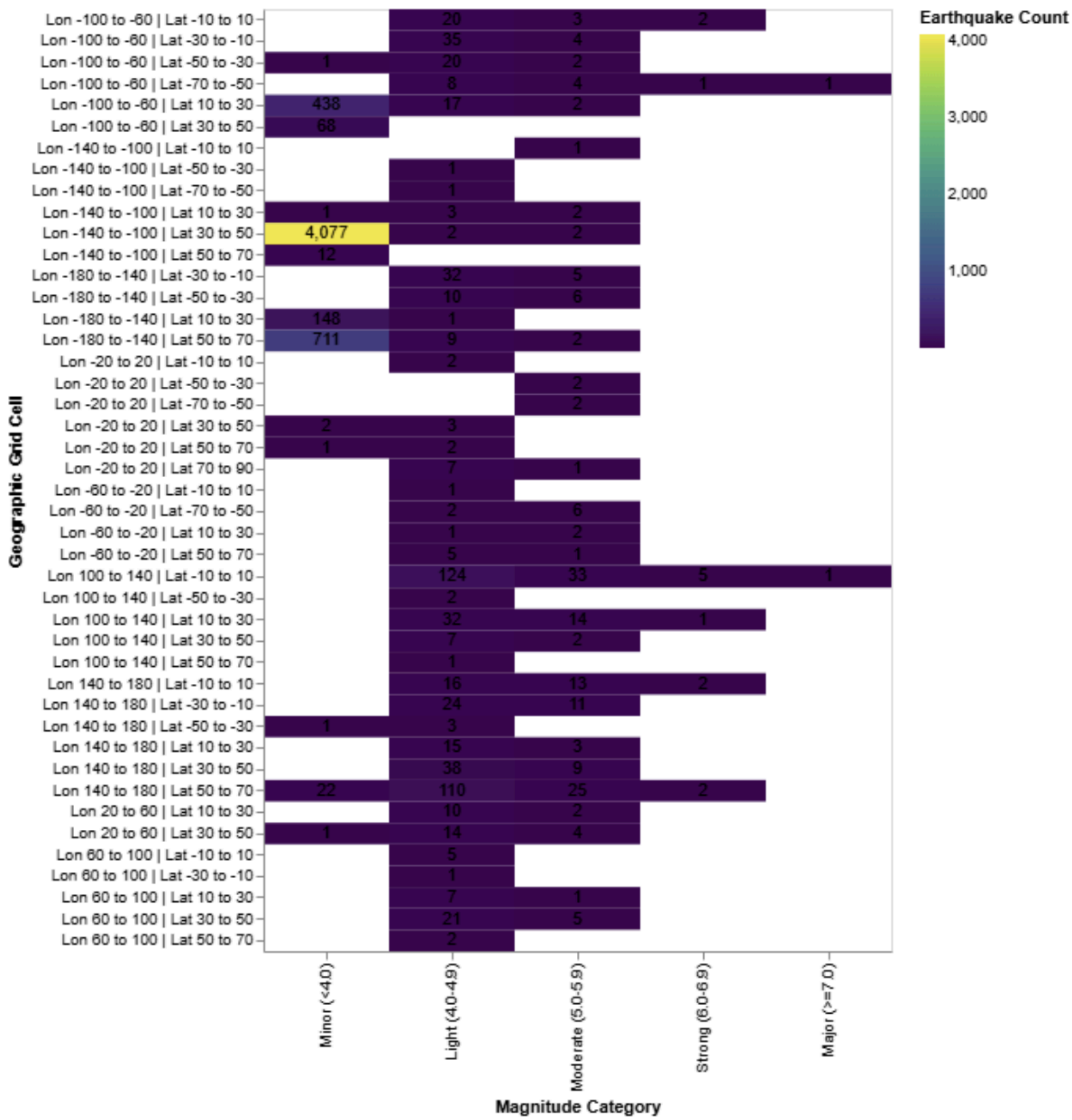
Negative (<0 km)

- Distribution: Smallest group, concentrated at *M < 1.0*.
- Significance: Represents highly sensitive detections of micro-events, often from precise measurement methods; not true “above-surface” quakes.





There is a strong correlation between longitude vs mag and latitude vs mag. Let's dive deeper to understand the distribution of earthquakes based on regions by assigning the grid to the world map.



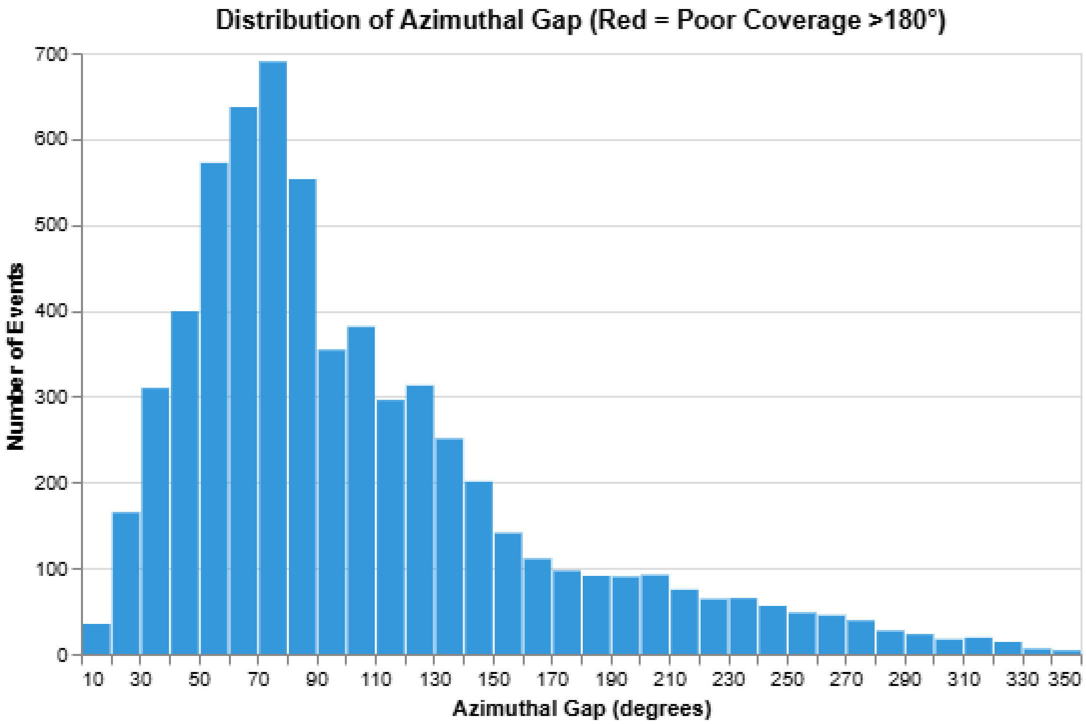
High Sensitivity vs. True Activity: The cell Lon -140 to -100, Lat 30 to 50 (4,077 minor quakes) has the highest count. This is a region (like California) with high tectonic activity and an extremely dense seismic network, allowing for the detection of many more minor quakes compared to less instrumented regions.

Relative Hazard: While the overall counts are dominated by Minor quakes, the cells with the most quakes in the Moderate (5.0-5.9) and Strong (6.0-6.9) categories (e.g., Lon 100-140) indicate the areas with the highest frequency of potentially damaging events.

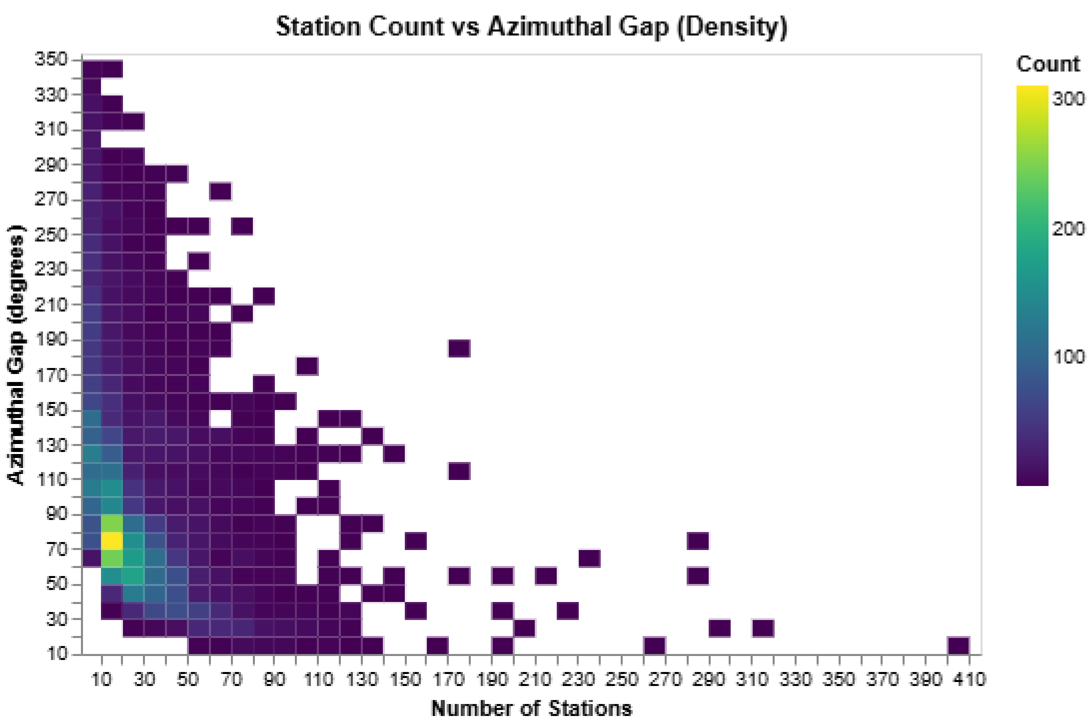
EDA 3: LOGAN

[Logan received an extension from Dr. K due to medical reasons. He will submit the rest on his own time]

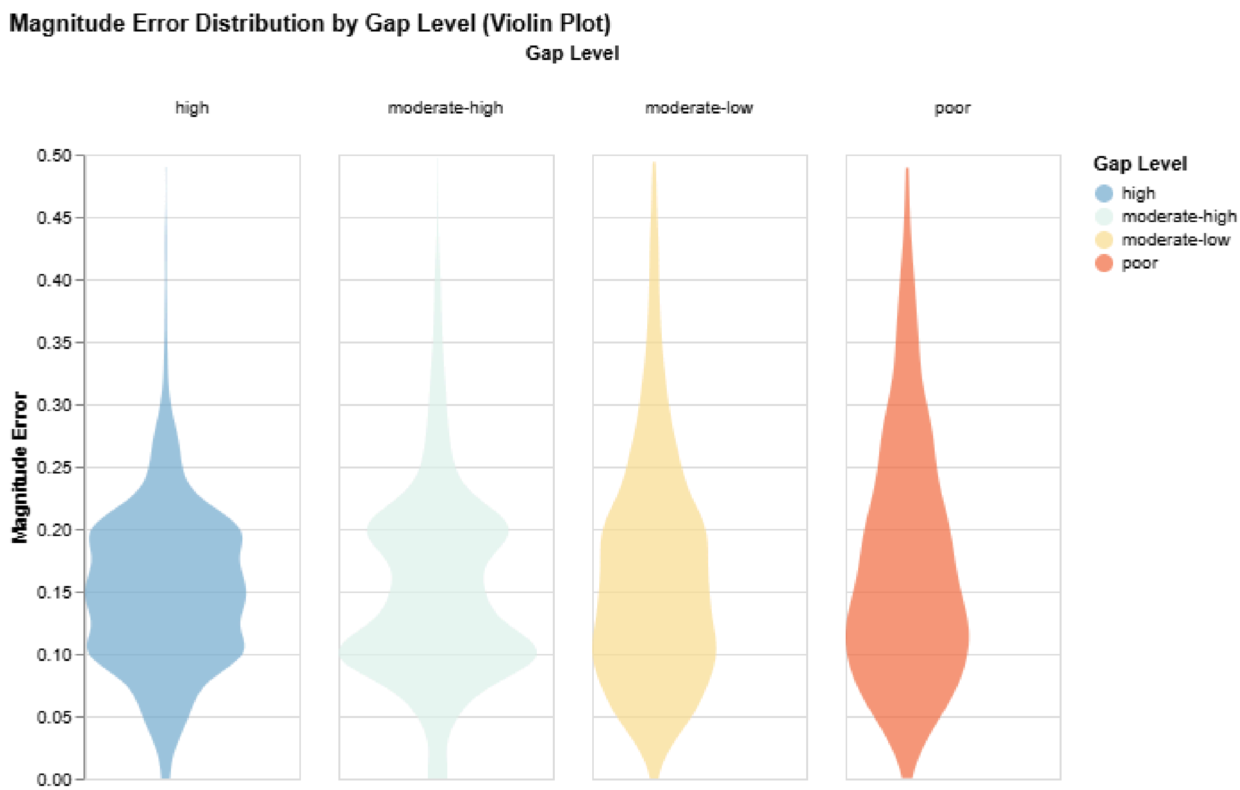
EDA 4: ATHENA



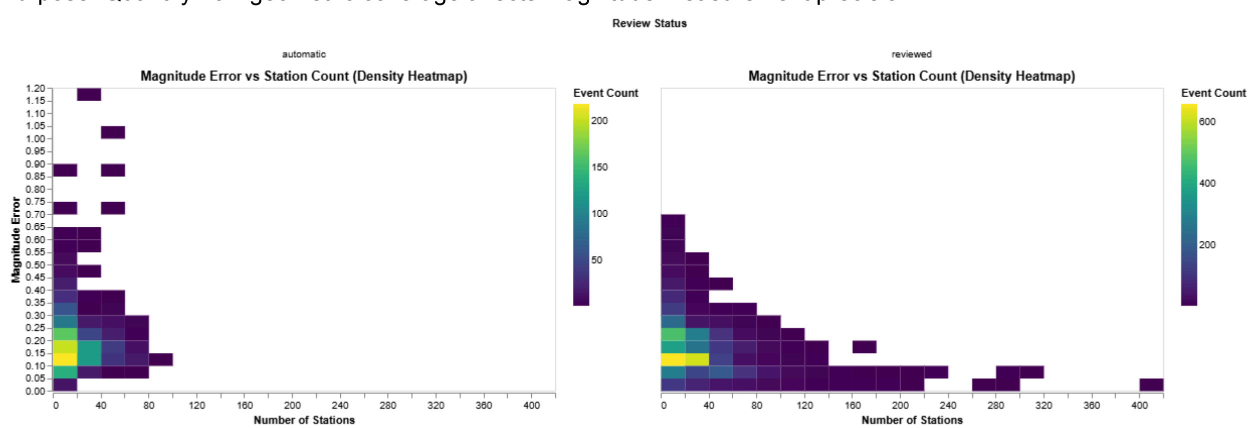
Purpose: Examine the geometric coverage quality - how well stations surround earthquake epicenters.



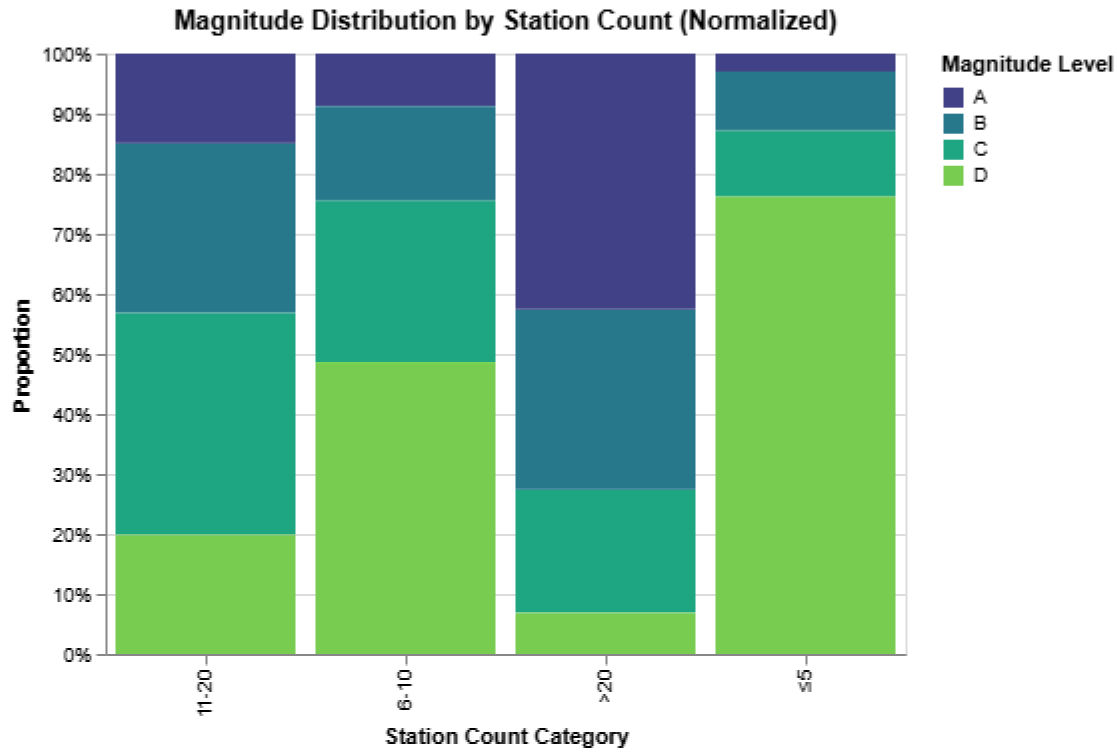
Purpose: Understand whether more stations automatically means better geometric coverage.



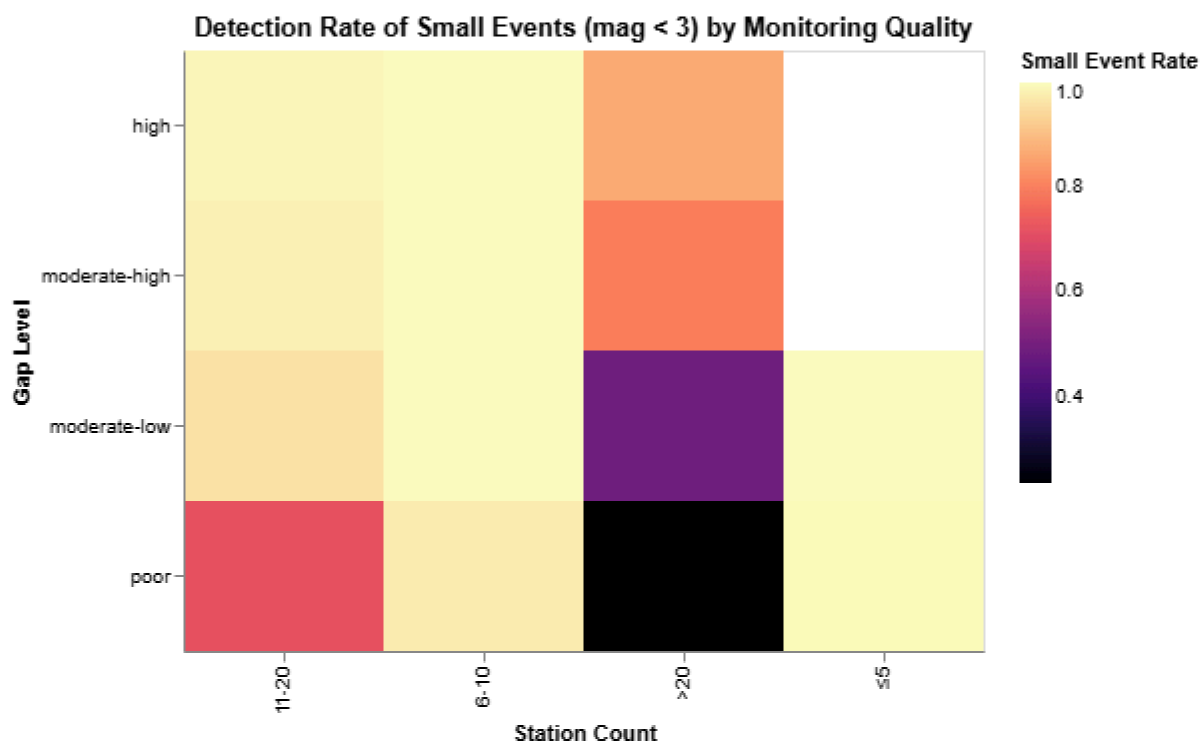
Purpose: Quantify how geometric coverage affects magnitude measurement precision.



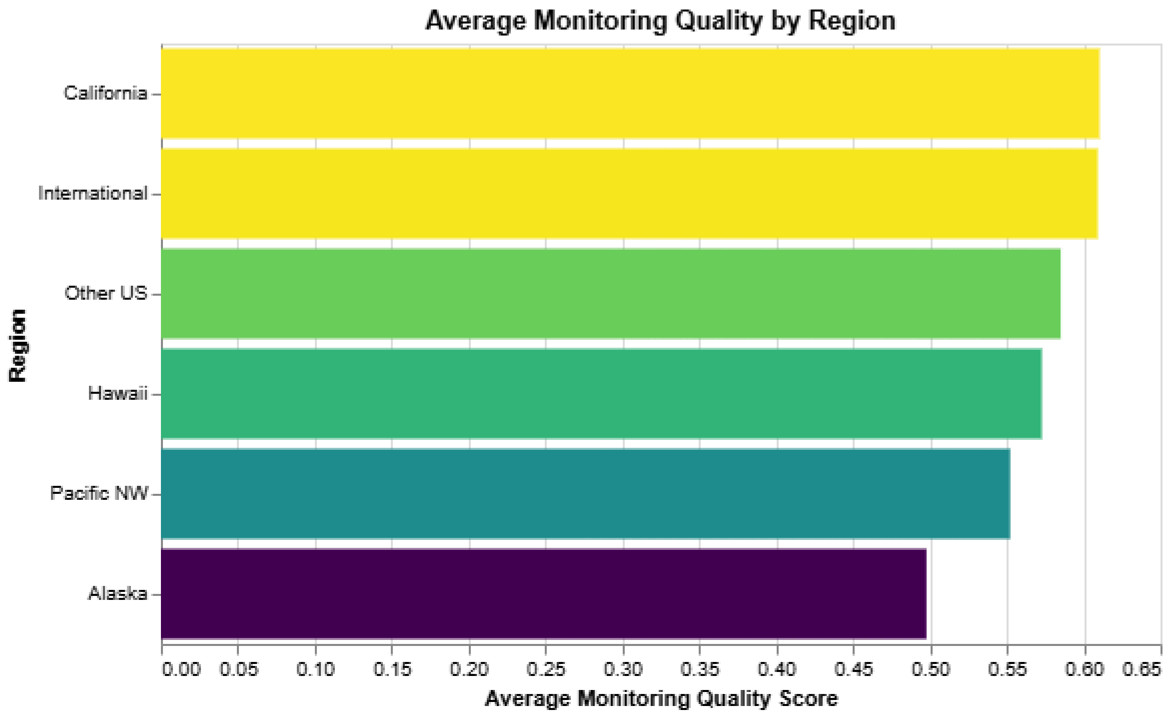
Purpose: Examine whether more stations reduce measurement uncertainty.



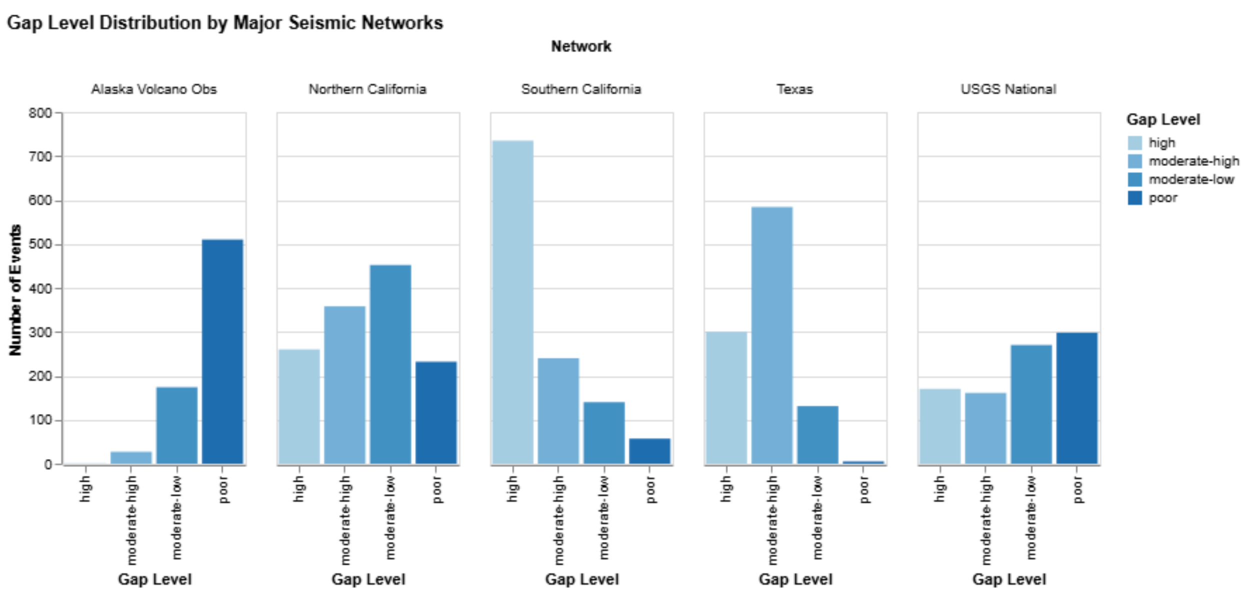
Purpose: Test whether well-monitored areas detect proportionally more small earthquakes.



Purpose: Quantify exactly how much detection bias exists across different monitoring configurations.



Purpose: Understand how different seismic networks prioritize coverage quality.



Purpose: Understand how different seismic networks prioritize coverage quality.

<u>**3. Focus of Inquiry**</u>

3.1. LUNA

Inquiry Theme: *Magnitude and Depth Patterns Across Event, Network, and Measurement Types.*

This inquiry examines how the characteristics of seismic events, particularly magnitude and depth, interact across different event types and reporting networks, while also considering how magnitude measurement systems used to observe and quantify their strength influence these interpretations. Earthquake records represent a mixture of physical signals and observational context; so, by jointly studying magnitude, depth, event type, reporting network, and magnitude measurement types, the project aims to understand the interactions and how these relationships are interpreted to provide a comprehensive understanding of how physical and measurement factors shape our interpretation of global seismic events. The goal is to uncover systematic patterns in how earthquakes are recorded and represented, identifying both physical relationships (i.e. do deeper quakes tend to be stronger?) and observational effects driven by network or methodological differences (e.g. how measurement type affects comparability).

Analytic Question 1:
How does the relationship between earthquake depth and magnitude differ across event types, and do these patterns hold consistently across major recording networks?

- <u>Low-level analytic tasks:</u>
  - Characterize distribution: Compute and visualise distributions between magnitude and depth within each event type and top reporting networks, potentially with histograms, bar chart, grouped counts.
  - Correlate and compare: Create scatterplots, trendlines, and correlation coefficients of magnitude vs. depth within each event type and network. Possibly also using correlation heatmaps.
  - Visualize relationships: Create faceted scatterplots and trendlines of magnitude vs. depth, faceted by event type and/or network.
  - Find anomalies: Summarise average magnitude and depth ranges by event type and network to identify outliers or deviations, and find unusual distributions. Potentially with box plots, heatmaps.

Analytic Question 2:
How do different magnitude calculations (magType) influence the observed relationship between magnitude and depth, and what do these differences reveal about network reporting practices?

- <u>Low-level analytic tasks:</u>
  - Characterize distribution: Explore the prevalence of each magnitude measurement type across event types and networks. Compare magnitude distribution for each magType. Potentially with histograms, bar charts, grouped counts.
  - Find relationships: Examine whether the relationship between magnitude and depth varies systematically across magTypes. For each magType, compute mag-depth (faceted) scatterplots or grouped bar charts, compare mean and median magnitudes.
  - Mapping: Identify which networks primarily report which magTypes or rely most heavily on specific magTypes, and whether those choices align with event profiles (e.g. local small events vs. large).
  - Find anomalies: Identify magType-network combinations that produce outlier magnitudes (i.e. unexpectedly high or low mag for a magType) and evaluate whether these outliers reflect measurement differences or true seismic anomalies. Potentially with scatterplots, boxplots, heatmaps.

3.2. ERICSON

Inquiry Theme: Geolocation and its relation with other earthquake attributes

This investigation aims to explore the spatial and quantitative characteristics of global earthquake activity to uncover patterns that can inform seismic hazard awareness and risk mitigation. By analyzing the distribution of earthquake magnitudes, depths, and station gaps across different geographic regions, the study seeks to identify where seismic events are most concentrated and how their physical characteristics vary globally. Through visual and statistical analyses, such as mapping epicenters, examining correlations among seismic parameters, and detecting regional anomalies, this inquiry will highlight high-activity zones and regions with particularly hazardous profiles, those combining high magnitudes with shallow depths. Lastly, the theme focuses on integrating spatial analytics and data-driven insights to better understand earthquake behavior and guide regionally targeted preparedness efforts.

Analytic Question 1:

What is the overall spatial and quantitative profile of the earthquake dataset, specifically identifying the primary geographic clusters and the general trends in magnitude, depth, and the seismic station gap?

<u>Low-level analytic tasks:</u>

- Characterize Overall Distribution (Global):
  - Plot all earthquake epicenters on a world map, colored by Magnitude Category and sized by Magnitude. This will immediately show where the bulk of the seismic activity is concentrated globally.
  - Visualize the overall distributions of magnitude, depth, and gap using histograms or density plots to understand the central tendencies, spread, and skewness of each key feature.
- Correlate and compare:
  - Quantify the relationships among mag, depth, latitude, longitude, and gap using a correlation heatmap or channels within a scatterplot
  - Specifically, examine the relationship between magnitude and depth, and magnitude and gap using scatter plots with trendlines to see if higher magnitudes correlate with deeper events or larger gaps between recording stations.
- Find Anomalies:
  - Identify Geographic Skew by analyzing the proportional distribution of magnitude categories across binned latitude and longitude to identify the general coordinates that concentrate the highest counts of large events

Analytic Question 2:

Which specific, high-activity geographic regions exhibit earthquake characteristics (high magnitude, shallow depth) that pose the greatest potential hazard, thereby informing regionally focused preventative measures?

<u>Low-level analytic tasks:</u>

- Characterize Overall Distribution (Global):
  - Profile Regional Hazard: For the top high-activity regions, create Box Plots or Violin Plots to compare: Median Magnitude and Median Depth (especially identifying regions with shallow, high-magnitude events). Use grouped counts to show the count of Strong or Major events (Magnitude >= 6.0) within each region.
- Correlate and compare:
  - Compare Network Reporting: Analyze which major reporting networks (net) are responsible for reporting events in these high-hazard regions to understand monitoring distribution.
- Find Anomalies:
  - Identify High-Hazard Regions: Clean, group, and extract the top N most frequent geographic regions from the place field that have also recorded the highest average or maximum magnitudes (identifying regional outliers).
  - Spatially Isolate Hazard Zones **(Attempt if possible)** : Focus the geographic plot onto a region of interest and use density heatmaps or Hexbin plots to precisely pinpoint the latitude/longitude micro-zones within that region that concentrate the largest, shallowest events (identifying micro-spatial anomalies in hazard intensity).

3.3. LOGAN

[Logan received an extension from Dr. K due to medical reasons. He will submit the rest on his own time]

3.4. ATHENA

Inquiry theme: *Understanding How Seismic Network Geometry and Monitoring Infrastructure Shape Earthquake Detection and Measurement Quality.*

This inquiry investigates the critical but often overlooked relationship between the spatial configuration of seismic monitoring networks and the quality, completeness, and reliability of earthquake data. When an earthquake occurs, its detection and accurate measurement depend fundamentally on the geometry of nearby seismic stations: how many stations record it (nst), how those stations are distributed around the epicenter (azimuthal gap), and how close the nearest station is (dmin). These geometric factors directly influence both what we can detect and how precisely we can measure it.

Analytic Question 1:
How does seismic network geometry (station count, azimuthal gap, nearest station distance) influence earthquake measurement uncertainty, and do these relationships vary by review status or geographic region?

- Low-level analytic tasks:
    - Correlate: Compute correlation coefficients between monitoring variables (nst, gap, dmin) and error metrics (magError, depthError, rms) to quantify relationship strength.
    - Compare: Compare error distributions across gap_level categories using boxplots and statistical tests to identify meaningful differences.
    - Characterize Distribution: Examine the distribution of monitoring quality scores across regions to understand baseline network capabilities.
    - Filter & Aggregate: Separate automatic versus reviewed events to assess whether human review compensates for poor geometric coverage.
    - Find Extremes: Identify events with exceptionally high errors despite good coverage, or low errors despite poor coverage, to understand outlier patterns.

Analytic Question 2:
Does better seismic monitoring enable detection of more small-magnitude earthquakes, and how do detection biases vary across regions and network configurations?

- Low-level analytic tasks:
    - Compute Derived Value: Calculate small-event detection rates (proportion of mag < 3) for different monitoring quality bins.
    - Correlate: Assess relationship between monitoring_score and small_event_share across regions to quantify detection bias.
    - Compare: Compare magnitude distributions between high-quality and low-quality monitoring conditions using normalized stacked bars.
    - Cluster: Identify natural groupings of regions by their monitoring characteristics and detection patterns.
    - Find Anomalies: Detect regions with unusual detection patterns that deviate from expected monitoring-detection relationships**.**

---

4. Preliminary Sketches (~250 - 500 words)

4.1. LUNA

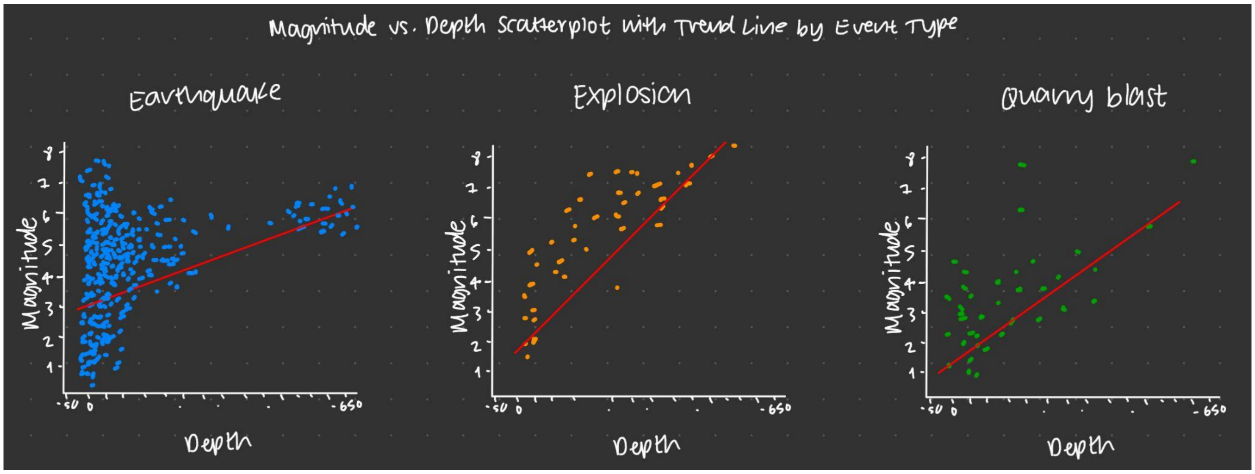
Low-fidelity sketches

Sketch 1

AQ1: *How does the relationship between earthquake depth and magnitude differ across event types, and do these patterns hold consistently across major recording networks?*

Task: Correlate, compare, find relationships.

<u>Sketch:</u> Scatterplot + Trendline, faceted by Event Type



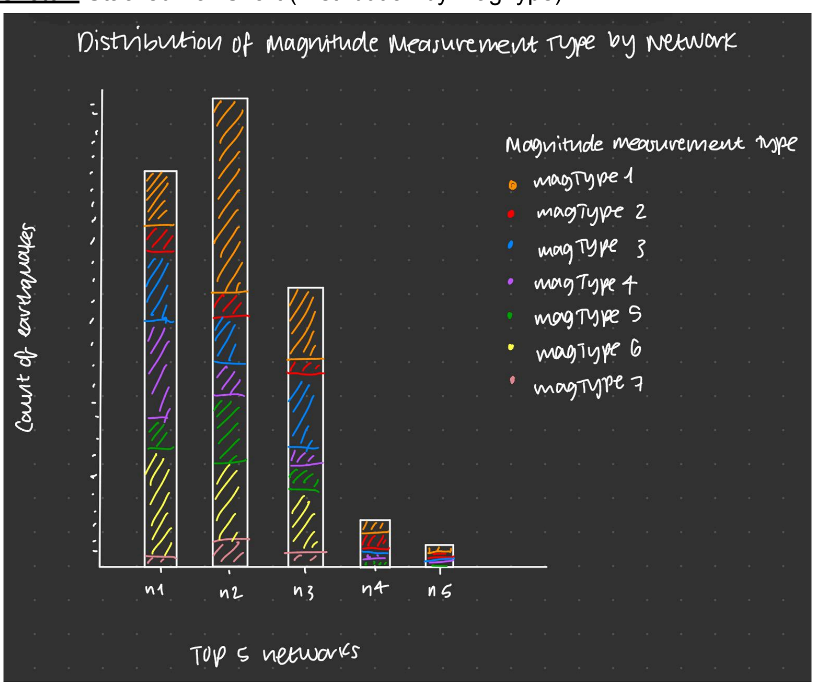
- I created a similar scatterplot + trendline for magnitude and depth and calculated correlation coefficient in general, but this dives deeper into the event types by faceting
- <u>Strength:</u>
    - Shows depth-magnitude relationship clearly across event types
- <u>Critiques:</u>
    - This is just a sample (i.e. I don't know the correct distribution) – scales may need to be independent, which can affect interpretability and comparability.
    - It could get crowded if it has too many points (especially earthquakes).

Sketch 2

<u>AQ2:</u> *How do different magnitude calculations (magType) influence the observed relationship between magnitude and depth, and what do these differences reveal about network reporting practices?*

<u>Task:</u> Characterize distribution, compare.

<u>Sketch:</u> Stacked Bar Chart (Distribution by magType)



- This stacked bar chart aims to show multiple things to understand the relationship and interactions between networks, magType, and count of earthquakes:

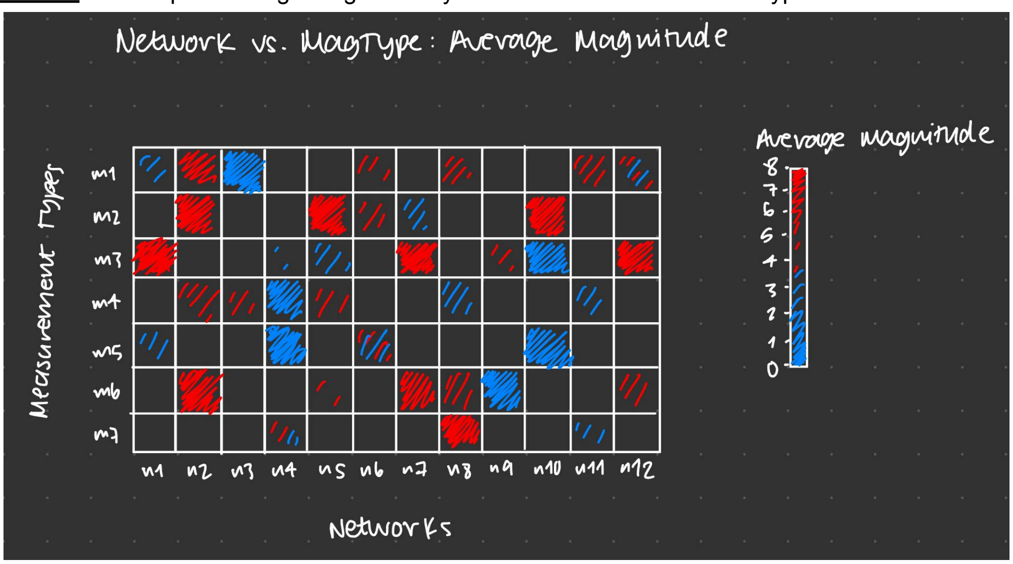
    - How many earthquakes the top 5 networks report
    - Within these 5 networks (bars), what is the distribution of magTypes, i.e. how much of this total count is calculated using each of the 7 magnitude measurement types
- Strengths:
    - Shows which magTypes (7) are reported for earthquakes and network (top 5) in a compact visualisation.
    - Clearly shows both total activity per network and relative distribution of magTypes.
    - Easier to interpret than grouped bars with 7 clusters per network (too many clusters).
- Critiques:
    - It's harder to directly compare counts of individual magTypes between networks – you have to visually estimate segment heights.
    - There are 7 magTypes and 12 networks here, so it can get busy if too many networks are included (here, I used top 5 networks like I did in EDA – this would improve clarity but could limit analysis and we may miss some patterns).

Sketch 3

AQ2: *How do different magnitude calculations (magType) influence the observed relationship between magnitude and depth, and what do these differences reveal about network reporting practices?*

Task: Find anomalies/outliers, view relationships

Sketch: Heatmap of average magnitude by networks and measurement types.

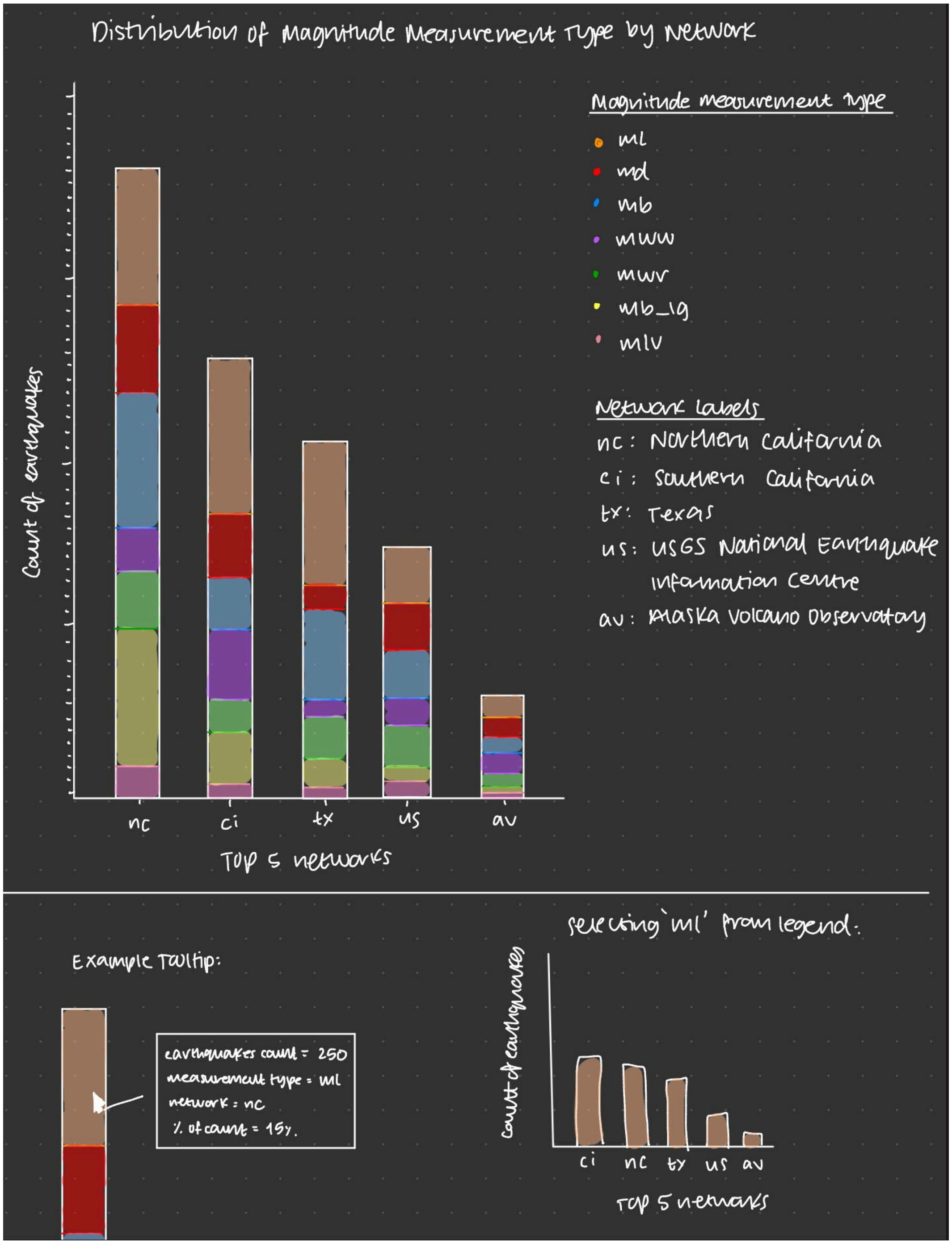


- This heatmap aims to show:
    - Which networks report the most earthquakes and what the average magnitude of those earthquakes are by colour
    - Which measurement types (magType) are used most and what the average magnitude of those are by colour
    - How does this differ between networks and measurement types, understanding patterns (i.e. do local networks report less with lower magnitudes and use measurement types used more for smaller events?) and any anomalies or outliers (i.e. bigger networks in more earthquake-prone areas that should report higher magnitudes and use more suitable measurement types, but use different types completely)
- Strengths:
    - Summarises network-magType patterns in one compact view, which highlights anomalies
- Critiques:
    - Loses individual data points.
    - Need something specific (i.e., a star or a circle to mark outliers) to explicitly show outliers in a more intuitive and interpretable way.

- ○ May be less intuitive for someone unfamiliar with heatmaps

High-fidelity sketch

AQ2: *How do different magnitude calculations (magType) influence the observed relationship between magnitude and depth, and what do these differences reveal about network reporting practices?*
Task: Characterize distribution, compare.
Improved High-fidelity Sketch: Stacked Bar Chart (Distribution by magType)



- This improved sketch shows a sample of how the stacked bar chart would like like, sorted in descending total count of earthquakes reported and with legends for colours (magnitude measurement type, magType) an the names of each of the top 5 networks.

- I would want to implement a tooltip for increased clarity and more details (bottom left example) for when you hover over a part of a bar.
- My vision for this chart is to be interactive (clickable). In the bottom right, I have an example of what would happen if you were to select ‘ml’ from the ‘Magnitude Measurement Type’ legend – the bar chart would only show bars for ‘ml’, and it would automatically sort in descending order.

4.2. ERICSON

Low-fidelity sketches

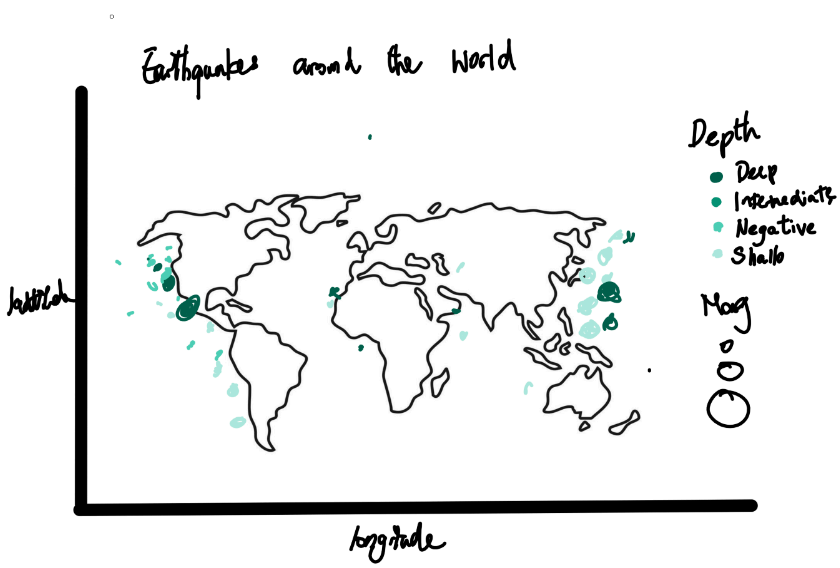
Sketch 1

AQ:

What is the overall spatial and quantitative profile of the earthquake dataset, specifically identifying the primary geographic clusters and the general trends in magnitude, depth, and the seismic station gap?

Task: Characterize Global Distribution, Compare and quantify the attributes of all earthquakes

Sketch:



Strengths:

- Strong visualization and very easy to relate each single earthquake to its relative geolocation
- Easy to interpret latitude and longitude
- Can compare each single data entry in depth and magnitude

Critiques:

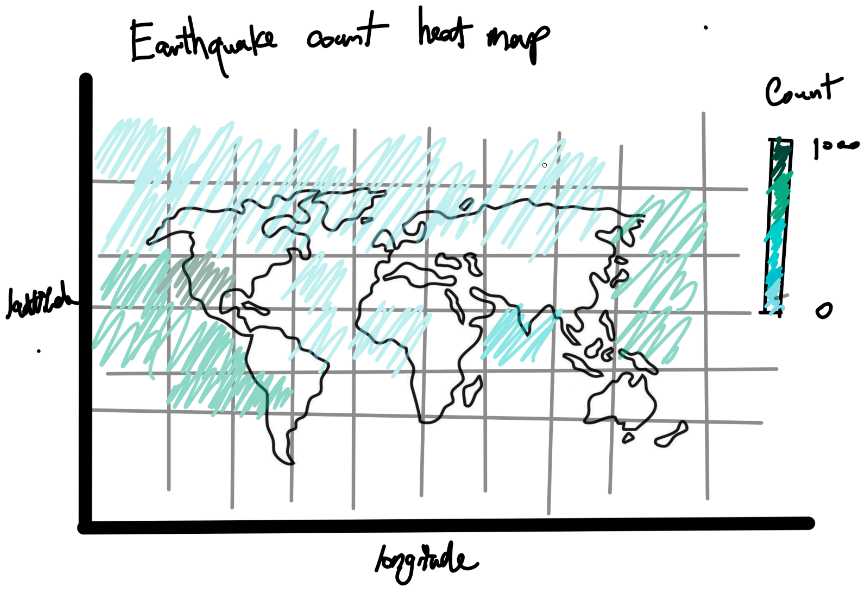
- There are 6000+ entries, areas where there are many data points, it is going to be very hard to tell the exact frequency from one place to another(i.e. 100 data points crowded in one area visually looks similar to 500 data points, it is hard for the audience to interpret the difference in trends).

Sketch 2

AQ: *Which specific, high-activity geographic regions exhibit earthquake characteristics (high magnitude, shallow depth) that pose the greatest potential hazard, thereby informing regionally focused preventative measures?*

Task: Characterize Global Distribution, Compare and quantify the count of all earthquakes, find abnormality regions with outstanding high/low counts

Sketch:



Strengths:

- Can easily quantify the difference in frequency based on region
- Can compare the difference in frequency based on region

Critiques:

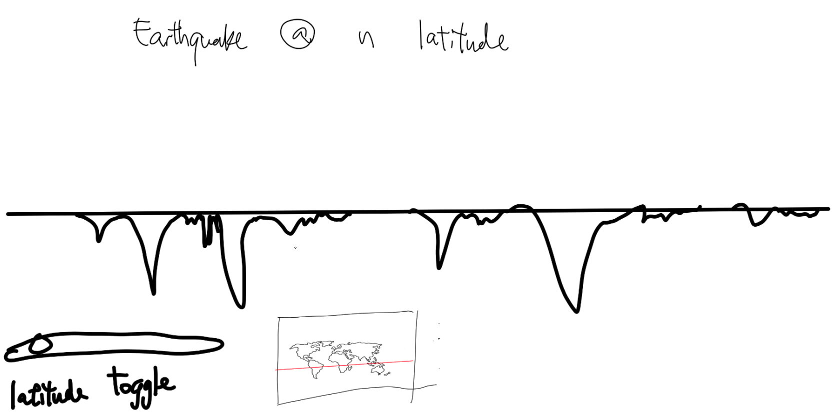
- Lost other earthquakes attributes other than frequency
- Lost individual data points information

Sketch 3

AQ: *Which specific, high-activity geographic regions exhibit earthquake characteristics (high magnitude, shallow depth) that pose the greatest potential hazard, thereby informing regionally focused preventative measures?*

Task: Compare and quantify the attributes of all earthquakes, find abnormality regions with outstanding high/low counts

Sketch:



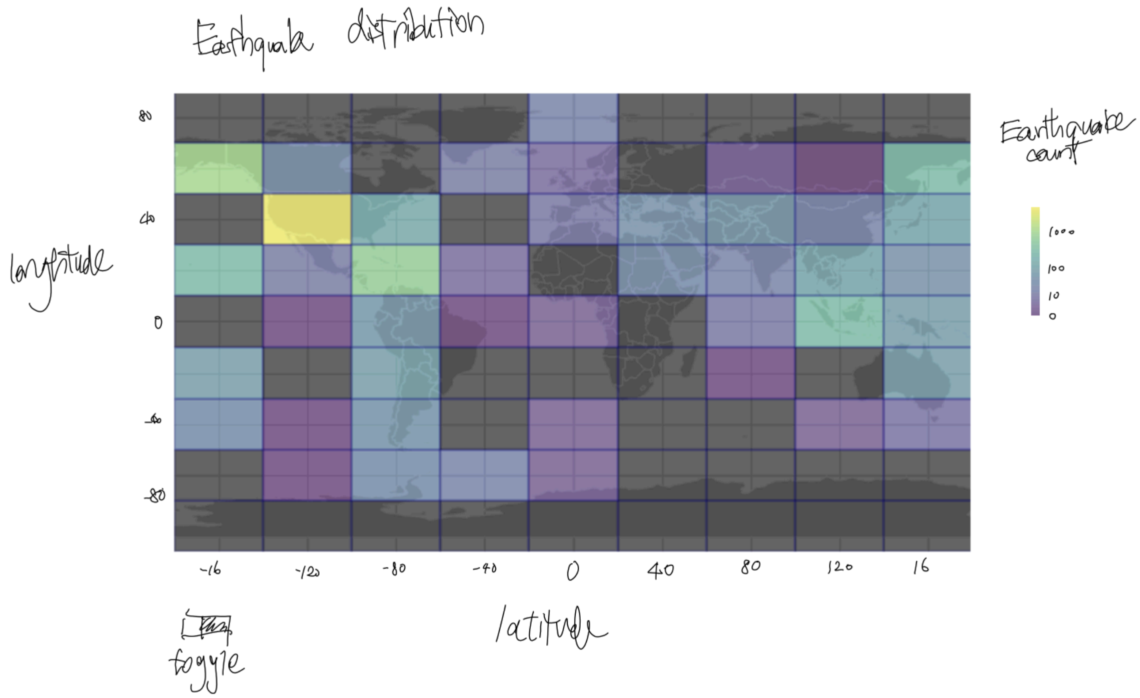
Strengths:

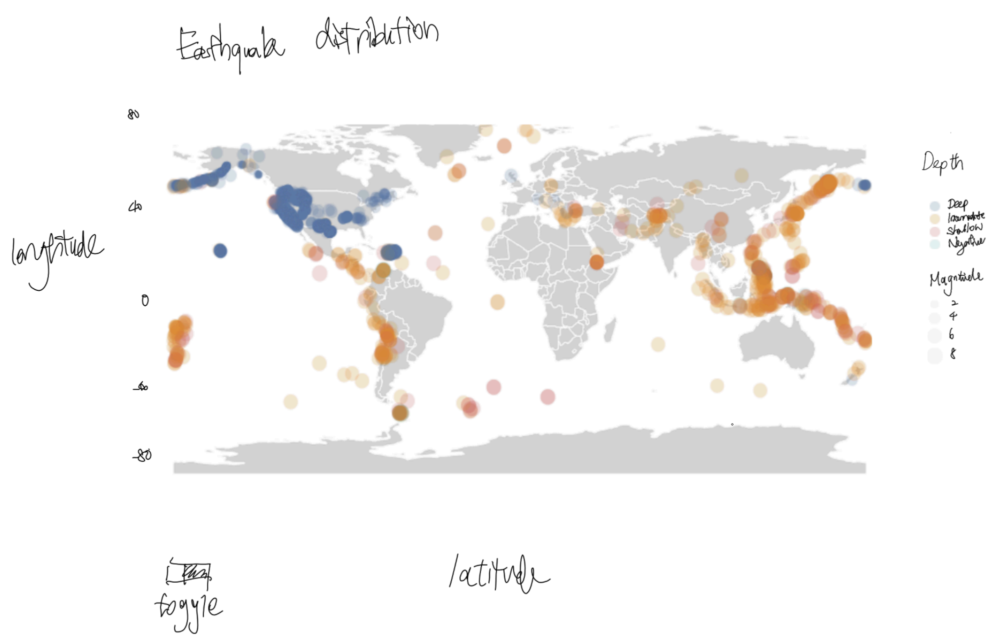
- By toggling along the latitude, the audience can capture all the details of each earthquake according to the longitude.

Critiques:

- Lost the overview of all earthquakes globally
- Hard to compare trends from a global perspective

High-fidelity sketch





In the Hi-Fi Sketch, we try to address both questions by using a toggle button so that the audience could toggle between the grid mode where you can see trends and count distribution by region better and scatterplot mode where you can see how individual earthquakes are being distributed.

In this graph we would also implement dynamic information that would update between the two modes when you hover over the data points.

4.3 LOGAN

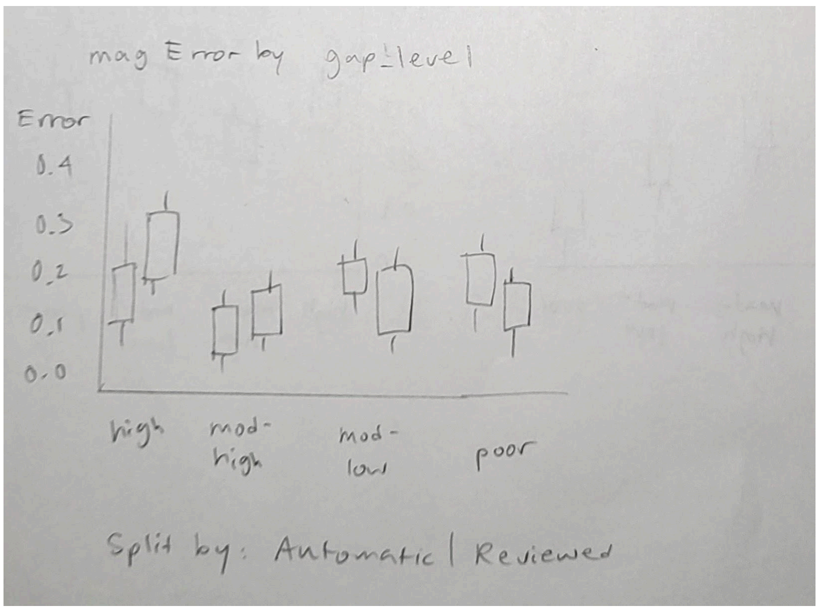
[Logan received an extension from Dr. K due to medical reasons. He will submit the rest on his own time]

4.4. ATHENA

Sketch 1: Magnitude Error by Gap Level

AQ: *How does seismic network geometry (station count, azimuthal gap, nearest station distance) influence earthquake measurement uncertainty, and do these relationships vary by review status or geographic region?*
Task: Compare and Characterize Distribution

Sketch:



Strengths:

- Shows clear progression. Error increases as gap quality worsens from high to poor.
- Side by side comparison shows reviewed events consistently have lower errors.
- Simple to implement in Altair
- No overplotting

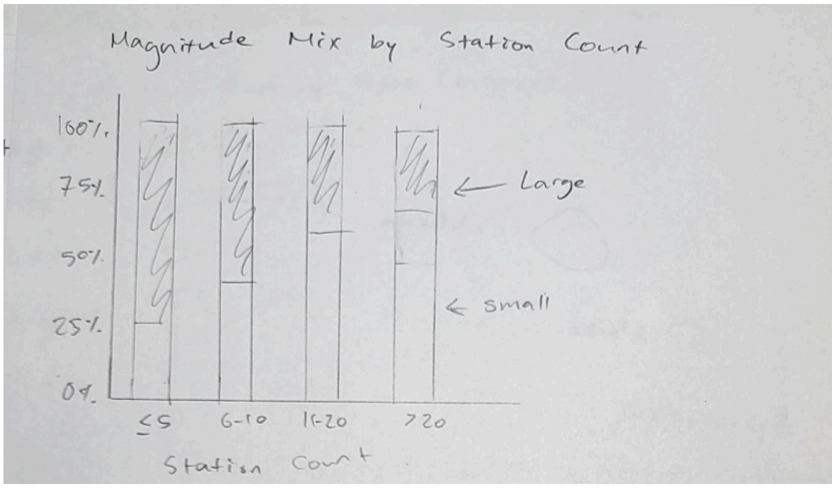
Critiques:

- Only examines gap, doesn't show stations count (nst) or distance (dmin) relationships
- Categorical grouping loses information about continuous relationships
- Focuses on one monitoring dimensions and doesn't show how nst and gap interact

Sketch 2: Magnitude Distribution by Station Count

AQ: *Does better seismic monitoring enable detection of more small-magnitude earthquakes, and how do detection biases vary across regions and network configurations?*
Task: Compare, Compute Derived Value, Characterize Distribution

Sketch:



<u>Strengths:</u>

- Detection bias is easily identified. Small event proportion increases 3× from poor to good monitoring.
- Normalized stacking makes proportions directly comparable.
- Simple and familiar chart type and is easy to understand
- Percentages directly answer “how much bias?”

<u>Critiques:</u>

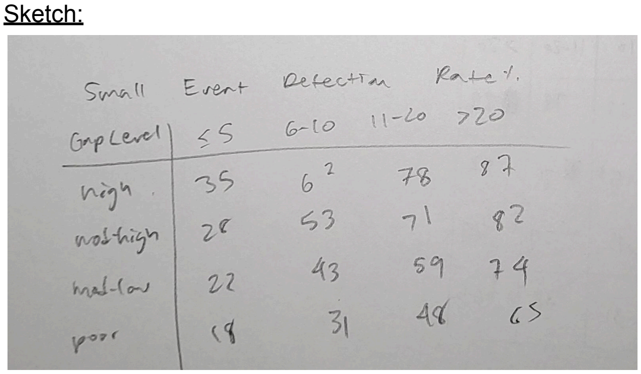
- Only shows station count dimension and ignores the azimuthal gap
- Doesn’t show regional variation
- Normalization hides that well-monitored areas also have more total events.

Sketch 3: Detection Rate Heatmap

<u>AQ:</u> *Does better seismic monitoring enable detection of more small-magnitude earthquakes, and how do detection biases vary across regions and network configurations?*

<u>Task:</u> Compare, Find Anomalies, Correlate

<u>Sketch:</u>



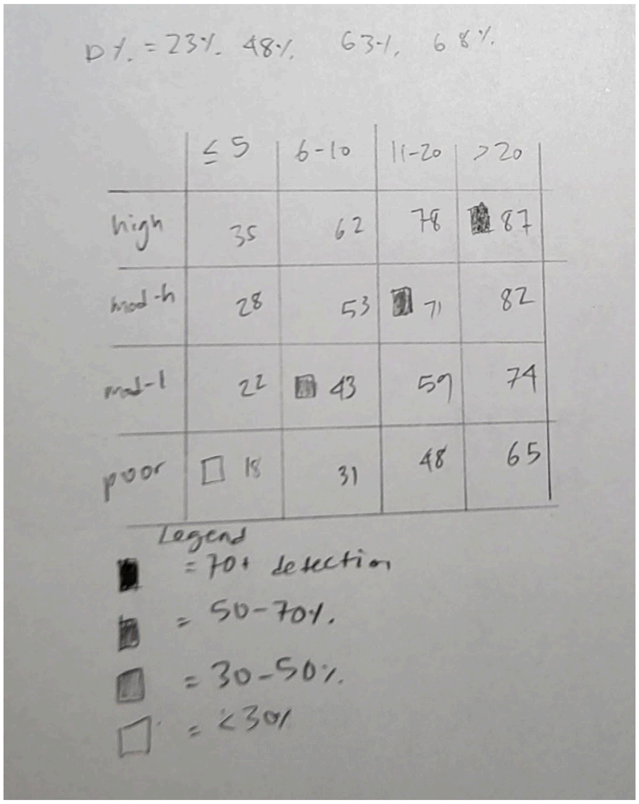
<u>Strengths:</u>

- Shows both monitoring dimensions simultaneously (nst and gap)
- Quantifies bias precisely
- Colour and numbers mean quick visual scan and precise reading
- Compact and all 16 condition combinations in one view.
- Directly answers “how do detection biases vary across configurations?”

<u>Critiques:</u>

- Requires pre-calculating detection rates (more data prep)
- Regional variation not visible (could add as separate annotation)

High-fidelity sketch



Sketch 3 provides the most comprehensive answer to AQ2 by simultaneously showing how detection bias varies across both network dimensions (station count and geometric coverage). While sketch 2 demonstrates detection bias exists, sketch 3 quantifies exactly how much bias occurs under each monitoring configuration, revealing that the best conditions (>20 stations + high coverage) detect 87% of small earthquakes while the worst (<= 5 stations + poor coverage) detect only 18%.

Sketch 3 provides the most comprehensive answer to AQ2 by simultaneously showing how detection bias varies across both network dimensions (station count and geometric coverage).

Implementation Plan:

- Use mark_rect() for heatmap with viridis color scale
- Overlay mark_text() showing detection rates (format='.2f')
- Conditional text color (white on dark, black on light)
- Add regional annotations comparing California vs Alaska
- Properties: width=400, height=300

---

5. Next Steps

Team-wide plan

Our team has outlined a detailed plan to progress from PD2 to PD3, ensuring all tasks are completed efficiently and collaboratively. With about 2.5 weeks before PD3 (Dashboard) is due on Friday, 28th November, we aim to stick to the following team-wide outline:

- Sun 16 Nov: Finalize ideas for how to implement sketches in Python Altair, select specific visualizations.
- Sun 23 Nov: Complete initial visualizations and integrate into a dashboard template.
- Wed 26 Nov: First complete draft of dashboard with all interactions functional, meet in/after lecture.

- Thu 27 Nov: Attend lab, troubleshoot remaining issues, decide on any edits, request help from TAs where necessary.
- Fri 28 Nov: Submit finalized PD3 dashboard.

Key milestones and responsibilities

	Luna Gulec	**Ericson Ho**	**Logan Wu**	**Athena Wong**
16-Nov	Decide final visualizations; prep Altair code	Decide final visualizations; prep Altair code	Decide final visualizations; prep Altair code	Decide final visualizations; prep Altair code
23-Nov	Implement viz; integrate into dashboard	Implement viz; integrate into dashboard	Implement viz; integrate into dashboard	Implement viz; integrate into dashboard
26-Nov	Refine interactions; help integrate team sections	Refine interactions; help integrate team sections	Refine interactions; help integrate team sections	Refine interactions; help integrate team sections
27-Nov	Review and troubleshoot dashboard	Review and troubleshoot dashboard	Review and troubleshoot dashboard	Review and troubleshoot dashboard
28-Nov	Final draft, submit	Final draft, submit	Final draft, submit	Final draft, submit

Final comments

This rough timeline allows for iterative review, ensures all members have clear responsibilities, and allows us to stick to a timely plan. Following PD3, we will allocate time for feedback and preparation for the final video submission leading up to the PD4 due date on Friday, 5th December.