Final Report of Project 2

County-Level Earthquake Risk Screener: a SQL-first earthquake risk ranking of US counties

Student Name: Reza Farzad

Faculty Mentor: Prof. Roya Ghiaseddin

ACMS 86700-02 Directed Readings

FALL Semester 2025

1- Introduction:

Earthquake risk varies widely across US regions because different areas face different levels of **hazard** (how often and how strongly the ground shakes) and **exposure** (how many buildings, people, or infrastructure could be affected). Hazard represents the natural threat, while exposure captures the scale of potential threat and damage.

This project develops a SQL-only, clear, and reproducible risk screener ranking all US counties by combining (1) a historical catalog of nearby earthquakes (as a proxy for seismic hazard), and (2) county population (as a proxy for exposure).

The goal is to build a screening tool to identify high-risk counties that can lead to additional detailed studies. The model produces: (1) a hazard score per county, (2) a exposure score per county, (3) a combined risk index based on normalized hazard and exposure scores, (4) national and state-level ranked lists, and (5) visualizations to support interpretation. The hope is that we can expand the current work later to build a full loss model.

The entire pipeline, including ETL (Extract, Transform, Load), spatial linking, hazard and exposure formulation and computation, standardization, and ranking, is implemented in **pure DuckDB SQL** with Python used only for visualization. In practice, pandas package of Python is commonly used for datasets (up to 2-3 GB) while SQL is often used for larger datasets. However, in this project, although the data size is small, SQL is chosen for the sake of practice.

2- Goals of the Project:

The project's goals are: (1) build and reform datasets including earthquake events, and county information consisting of population and spatial population centroid, (2) define a clear and interpretable risk index combining hazard and exposure, and (3) generate ranked outputs, including top 30 counties nationwide as well as top 10 states.

3- Dataset and Scope:

This research uses 5672 **earthquake events** from USGS with magnitude ≥ 5 around US, as shown in Figure 1. This data includes features including event id, time, magnitude, latitude, and longitude of events. This study also uses **county data** (downloaded from US Census Bureau) including FIPS codes, county name, state name, population, and population centroid coordinates, used for proximity linking. The geographic scope includes 50 US states, and the time window ranges from Nov. 18, 1975, to Nov. 18, 2025.

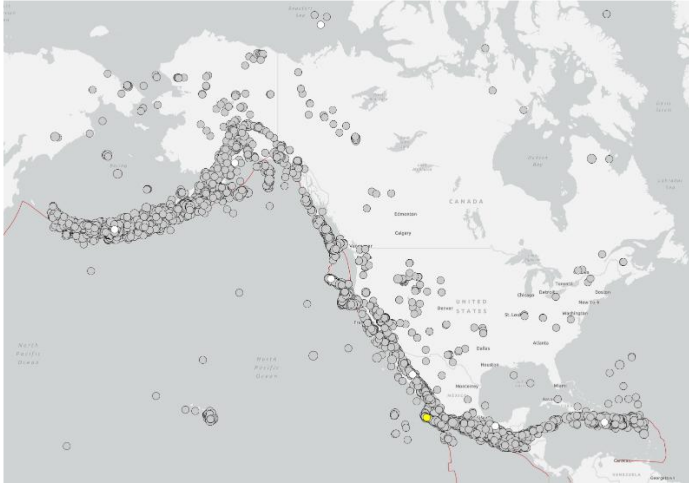


Figure 1. Location of earthquake dataset on the US map (generated by USGS earthquake catalog)

The considered **Hazard proxy** for county j is

$$H_j = \sum_{i=1}^{n_{eq}} \frac{(m_i)^{1.5}}{(1 + d_{ij})}$$

where, n_{eq} represents the number of earthquakes in the defined proximity policy around county j, m_i is the magnitude of earthquake i, and d_{ij} represents the (Haversine) distance between earthquake i and county j. The proximity policy used in this study is as follows: for earthquake with magnitude less than 7.0, the maximum considered distance is 200 km, while for others with greater magnitude the maximum considered distance is 250 km (implemented and again explained later in step 4.2). In other words, it is assumed that these earthquakes have no damaging effect beyond the maximum distance. The Haversine distance is also the distance of two points on top of a sphere. In this study, the distance between earthquakes and each county, assuming the earth is a sphere, is computed using the Haversine distance which is fairly accurate for distances less than 1000 km. The formula to compute the Haversine distance is follows:

$$d = 2R\ arcsin(\sqrt{a})$$

in which R is the radius of the earth, and a is defines as

$$a = sin^2\left(\frac{\Delta\varphi}{2}\right) + cos\varphi_1 cos\varphi_2 sin^2\left(\frac{\Delta\lambda}{2}\right)$$

where φ and λ are the latitude and longitude (in radian, not degree), respectively, for which the subscripts 1 represents earthquake and subscript 2 represents the population centroid of the county.

The considered **Exposure proxy** for county j is

$$E_j = \log{(1 + p_j)}$$

where p_j is the population of county j. It should be noted that 1 is used to avoid numerical errors, and log scale is used to avoid the counties with extremely high population dominate the risk analysis.

Finally, the **risk index (unitless)** is defined as follows using the standardized versions (z-scores) of hazard and exposure:

$$R_j = z_{H_j} \times z_{E_j}$$

4- Methodology (ETL → Linking → Hazard/Exposure & Risk):

This section includes the methods implemented in the code:

4.1 Step 1 – ETL & Cleaning

Two raw CSV files were loaded into staging tables “stg_county_raw” and “stg_eq_raw”. County cleaning involved forming a 5-digit FIPS code, ensuring valid longitude and latitude bounds, and filtering out missing or negative population values. Earthquake cleaning included converting timestamps, ensuring valid latitude and longitude, and removing events with negative depth. Final counts indicated 3221 clean counties as well as 5646 clean earthquake events. Also, there are only 72 earthquake events with magnitude larger than 7.0.

4.2 Step 2 - Linking Earthquakes to Counties

Each event was linked to counties using a Haversine distance computation implemented in SQL. The magnitude-dependent radius rule is:

- If mag < 7.0, then we include the counties within 200 km.
- If mag ≥ 7.0, then we include counties within 250 km.

Based on this rule, the total number of valid event-county links is 11773. The quality checks also indicate that 24.2% of earthquakes linked to at least one county, and 36.4% of counties linked to at least one event. The distance distribution shown in Figure 2, indicates the increasing counts up to the distance of about 200 km with the majority of distances being between 50 and 200 km. Only the few earthquakes over magnitude of 7 were considered for distances larger than 200 km, i.e., up to 250 km.

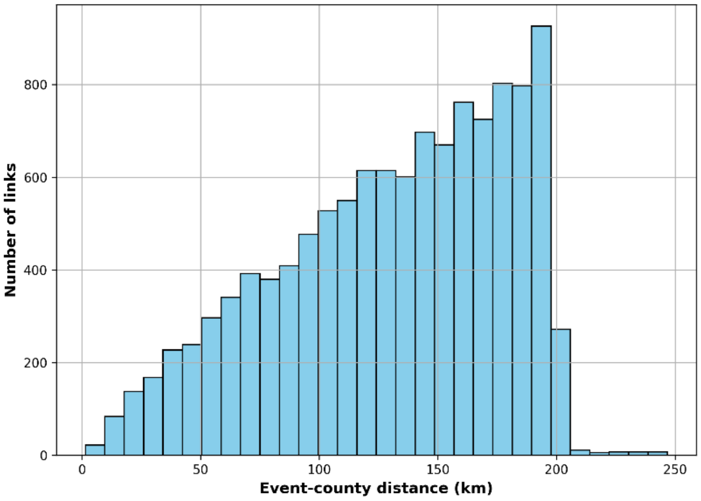
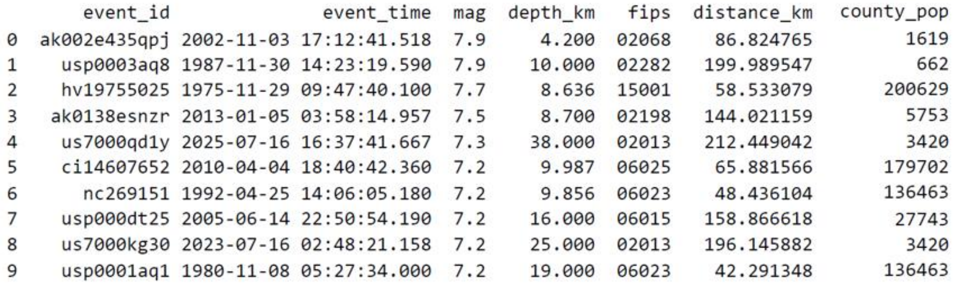


Figure 2. Histogram of distances for linked event-county pairs

A nearest-county table for the 10 largest events was also generated as shown below. This table demonstrates the top 10 largest earthquakes (affecting the counties) sorted in terms of earthquake magnitude as well as their distance with the nearest county along with the counties' population, verifying the correctness of spatial linkage for a few cases.



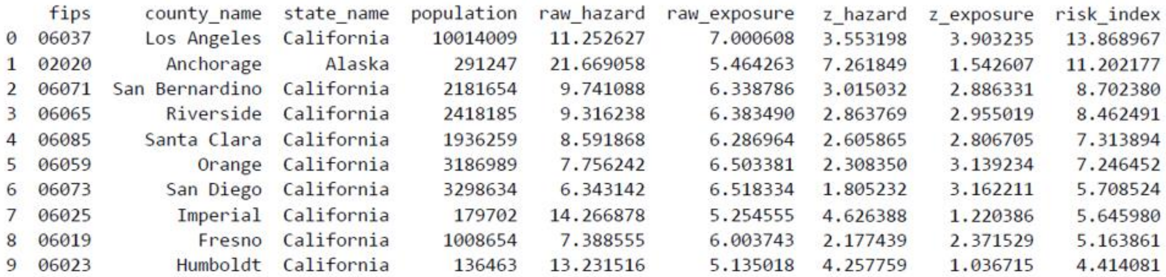
```
      event_id              event_time  mag  depth_km   fips  distance_km  county_pop
0  ak002e435qpj 2002-11-03 17:12:41.518  7.9     4.200  02068    86.824765        1619
1    usp0003aq8 1987-11-30 14:23:19.590  7.9    10.000  02282   199.989547         662
2    hv19755025 1975-11-29 09:47:40.100  7.7     8.636  15001    58.533079      200629
3   ak0138esnzr 2013-01-05 03:58:14.957  7.5     8.700  02198   144.021159        5753
4    us7000qd1y 2025-07-16 16:37:41.667  7.3    38.000  02013   212.449042        3420
5    ci14607652 2010-04-04 18:40:42.360  7.2     9.987  06025    65.881566      179702
6      nc269151 1992-04-25 14:06:05.180  7.2     9.856  06023    48.436104      136463
7    usp000dt25 2005-06-14 22:50:54.190  7.2    16.000  06015   158.866618       27743
8    us7000kg30 2023-07-16 02:48:21.158  7.2    25.000  02013   196.145882        3420
9    usp0001aq1 1980-11-08 05:27:34.000  7.2    19.000  06023    42.291348      136463
```

4.3 Step 3 – Hazard, Exposure, and Risk Index

For each county, hazard and exposure proxies (H_j and E_j), as well as their z-scores z_{H_j} and z_{E_j} were computed from which the risk index R_j was obtained. The results obtained in this step are shown in the next section.

5- Results:

The top 10 top-risk counties across nation are obtained based on the risk index value. As shown in the table below, Los Angeles is the number 1 county with the highest risk index value. This is primarily because of the high population of Los Angeles County, and then because of its exposure to hazard. The second county is Anchorage in Alaska which has extremely high hazard despite lower exposure. In general, California dominates due to frequent seismicity as well as large population.



```
    fips     county_name  state_name  population  raw_hazard  raw_exposure  z_hazard  z_exposure  risk_index
0  06037     Los Angeles  California    10014009   11.252627      7.000608  3.553198    3.903235   13.868967
1  02020       Anchorage      Alaska      291247   21.669058      5.464263  7.261849    1.542607   11.202177
2  06071  San Bernardino  California     2181654    9.741088      6.338786  3.015032    2.886331    8.702380
3  06065       Riverside  California     2418185    9.316238      6.383490  2.863769    2.955019    8.462491
4  06085     Santa Clara  California     1936259    8.591868      6.286964  2.605865    2.806705    7.313894
5  06059          Orange  California     3186989    7.756242      6.503381  2.308350    3.139234    7.246452
6  06073       San Diego  California     3298634    6.343142      6.518334  1.805232    3.162211    5.708524
7  06025        Imperial  California      179702   14.266878      5.254555  4.626388    1.220386    5.645980
8  06019          Fresno  California     1008654    7.388555      6.003743  2.177439    2.371529    5.163861
9  06023        Humboldt  California      136463   13.231516      5.135018  4.257759    1.036715    4.414081
```

A further investigation demonstrates the top 30 counties in the bar chart shown in Figure 3. According to this chart, after a sharp decrease in the quantity of risk index, the differences in quantities become smaller and smaller in the next points.

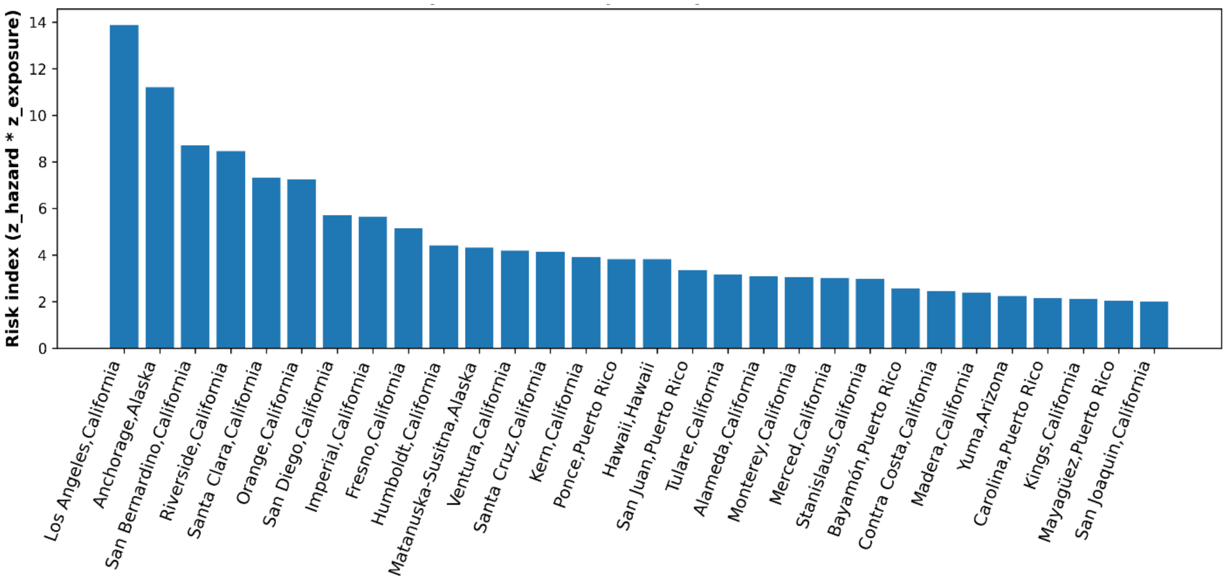


Figure 3. Top 30 counties by earthquake risk index

Scatterplot of hazard vs exposure, shown in Figure 4, indicates that the most US counties cluster at low hazard, and high hazard counties appear mostly at the top center of the plot. It should be noted that the risk index is shown in color for each county point, therefore, the highest values of risk index are located at top right section of the figure, where both exposure and hazard are highest.

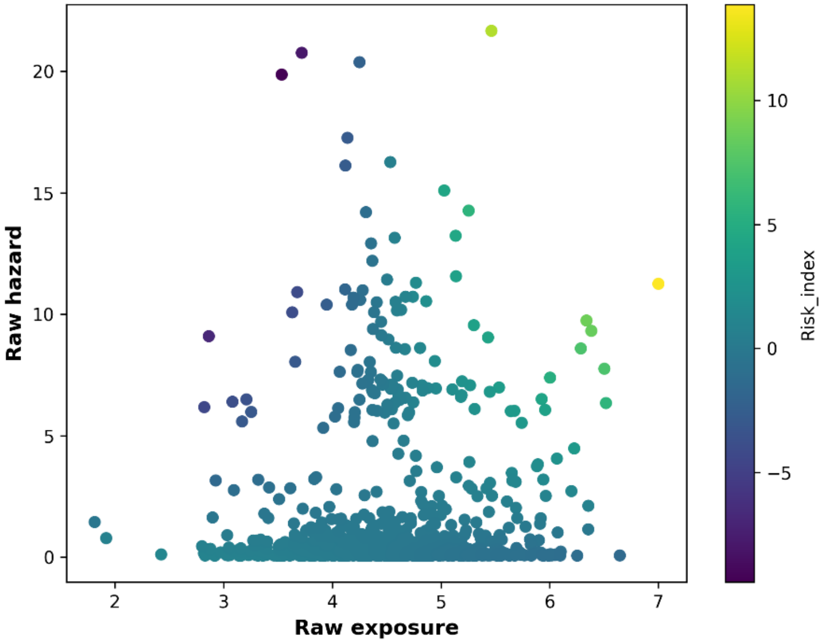
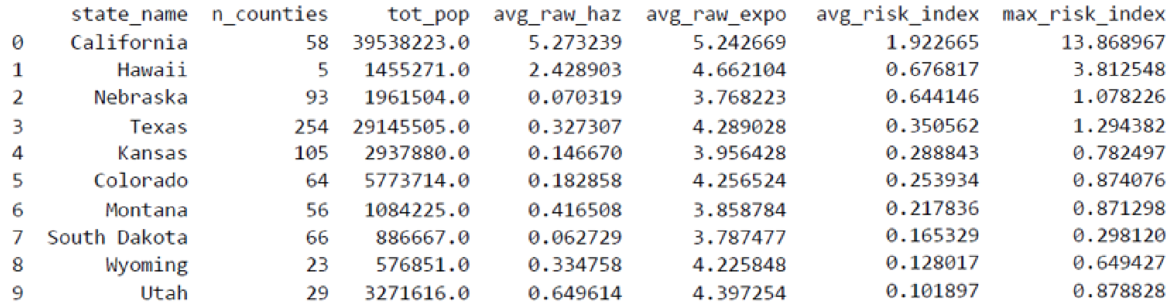


Figure 4. County hazard vs exposure (colored by risk index)

Finally, going beyond county-level investigations, a state-level exploration has been done to rank the top 10 states with the highest average risk index value across its counties. The results shown in the table below indicate that after California, Hawaii and Nebraska are ranked second and third states with the highest average risk index across counties. California and Hawaii show a significantly higher average hazard value, while Nebraska has South Dakota have the lowest hazard value among the top 10 states. The rest of the states shown in this table have moderate hazard with a high exposure quantity.



```
     state_name  n_counties     tot_pop  avg_raw_haz  avg_raw_expo  avg_risk_index  max_risk_index
0    California          58  39538223.0     5.273239      5.242669        1.922665       13.868967
1        Hawaii           5   1455271.0     2.428903      4.662104        0.676817        3.812548
2      Nebraska          93   1961504.0     0.070319      3.768223        0.644146        1.078226
3         Texas         254  29145505.0     0.327307      4.289028        0.350562        1.294382
4        Kansas         105   2937880.0     0.146670      3.956428        0.288843        0.782497
5      Colorado          64   5773714.0     0.182858      4.256524        0.253934        0.874076
6       Montana          56   1084225.0     0.416508      3.858784        0.217836        0.871298
7  South Dakota          66    886667.0     0.062729      3.787477        0.165329        0.298120
8       Wyoming          23    576851.0     0.334758      4.225848        0.128017        0.649427
9          Utah          29   3271616.0     0.649614      4.397254        0.101897        0.878828
```

6- Discussion:

This SQL-based screener correctly identifies US counties with the highest risk using a clear methodology. The risk index is unitless, interpretable, and reflects interaction between Hazard (natural threat) and Exposure (population concentration). Negative risk values occur when hazard and exposure mismatch. In other words, high hazard but low exposure (like rural area) or low hazard (near zero earthquake) but high exposure. While this is mathematically expected, we consider negative values of risk index as opportunity to ignore the corresponding counties in the current study.

The SQL workflow is fully reproducible. No geometry libraries are required. The implemented hazard model is clear and intuitive, and it can scale easily to large datasets because of the simplicity of implementation. Finally, the current project produces explainable outputs as watch lists or rankings.

The current project uses county centroids instead of full polygons and the exposure is simplified to population only. Hazard model does not include depth of earthquake, fault mechanism, or site amplification. Furthermore, there is no temporal weighting of recent vs older earthquakes.

FEMA NRI indicators can be used for validation. Separate hazard trends can be formed by decade. Housing and infrastructure can be integrated as exposure.

7- Conclusion:

This project successfully delivers a transparent SQL-first earthquake risk screener that (1) cleans and integrates county and earthquake datasets, (2) performs spatial linkage with magnitude-dependent conditions, (3) computes hazard, exposure, and risk scores, (4) produces national rankings of counties and states, and (5) generates clear and interpretable visualizations. This tool can serve as an initial screening mechanism for agencies, planners, and researchers to identify counties requiring deeper seismic risk assessment.

8- References:

- USGS Earthquake Catalog
- US Census County Populations