**EDA of Earthquake Dataset**

**Introduction**

Earthquakes are natural disasters that can cause significant damage to infrastructure and pose a threat to human lives. The study of earthquakes and their patterns is crucial for understanding their behavior, predicting future occurrences, and implementing effective mitigation strategies. With the availability of large earthquake datasets, exploratory data analytics using Python can play a vital role in uncovering valuable insights and patterns within the seismic data.

**Problem Statement**

The objective of this project is to perform an exploratory data analytics on the earthquake dataset records between the year 1969 to 2018 and with magnitude 5.5 and greater. By leveraging Python's data analysis library pandas and its visualization libraries, we aim to gain a comprehensive understanding of the earthquake patterns. Therefore, we have these tasks:

1. What's the total number of earthquakes recorded?
2. What is the relationship between the earthquake magnitudes and depths, and how does it impact our understanding of the earthquake events?
3. What are the 10 largest earthquakes from 1969 - 2018?
4. Find how many times each category of the earthquake occurred?
5. Fetch the records for all aftershocks within a week of the initial earthquake for the top greatest earthquake?
6. Find the count of the distinct cause of the earthquakes over the years?
7. Fetch the records where the earthquakes were caused by **'explosion'** only with respect to the place, magnitude and year occurred?
8. Which years have the highest number of earthquake recorded?
9. Find the most recent earthquake caused by nuclear explosion?
10. Find the places with the highest record of earthquake occurrence?

**Dataset**

The earthquake dataset used for this analysis is sourced from Socratica on Github. [https://github.com/socratica/sql]. It consists of a single CSV file named "earthquake.csv" containing the following columns:

* Earthquake\_id: The unique ID of each earthquake event
* Occurred\_on: The date and time of the earthquake event
* Latitude: The latitude coordinate of the earthquake location
* Longitude: The longitude coordinate of the earthquake location
* Depth: The depth of the earthquake event in kilometers
* Magnitude: The magnitude of the earthquake event
* Calculation\_method: Used to measure the earthquake size.
* Network\_id: Earthquake unique identifier
* Place: The region where the earthquake occurred
* Cause: The cause of the earthquake event

**Tools Used**

The analysis is performed using the following tools and libraries:

* Python: Programming language used for data manipulation and analysis
* Jupyter Notebook: Interactive environment for running Python code and documenting the analysis
* pandas: Library for data manipulation and analysis
* NumPy: Library for numerical computations
* matplotlib: Library for data visualization
* seaborn: Library for statistical data visualization

**Analysis Steps**

1. **Data Loading and Inspection**
   * Importing the necessary python libraries
   * Load the earthquake dataset into a pandas DataFrame.
   * Display the first few rows of the dataset to get an overview.
   * Check the dimensions of the dataset (number of rows and columns).
   * Identified the data types of each column and checked for missing values.
2. **Data Cleaning**
   * Converted the data type of the ‘occurred\_on’ column from object type to datetime using pd.to\_datetime function.
   * Extracted date, year, month\_name, and time columns from the ‘occurred\_on’ column.
   * Created a new column **'quake\_category'** that categorised the earthquakes into **great**, **major**, **strong**, and **moderate** earthquakes using python’s IF-ELSE function statement.
   * Checked for missing values, duplicates and outliers in the dataset.
3. **Data Exploration**
   * Examined the total number of earthquake recorded within the time.
   * Analyzed the number earthquakes occurrence by year by plotting a horizontal bar graph of the number of earthquakes over year.
   * Investigated the relationship between earthquake magnitudes and depths using a scatter plot.
   * Fetched record from the dataset for the 10 largest earthquakes.
   * Examined the count of each category of earthquake using the Seaborn count plot.
   * Analyzed for the distinct cause of the earthquakes.
   * Analyzed the distribution of the earthquakes by regions of the world.

**Results and Findings**

The analysis of the earthquake dataset revealed the following insights:

* A total of **23119** earthquakes of magnitude 5.5 and greater were recorded between the years 1969-2018.
* There is a weak positive correlation between earthquake magnitudes and depths. The corr\_coef of **0.03** indicates that the relationship between earthquake magnitudes and depths is quite scattered and does not exhibit a strong linear pattern.
* The **9.1** magnitude near the coast of Honchu, Japan in 2011 is the joint largest with the **9.1** off the coast of Indonesia, which caused the devastating Boxing Day tsunami in 2004.
* The high magnitude earthquakes were discovered to spur numerous aftershocks.
* A total of **40 (0.17%)** **great** earthquakes, **665 (2.88%)** **major** earthquakes and **6357 (27.50%)** **strong** earthquakes were recorded between 1969 – 2018. Most of the earthquake categories were of moderate type with a percentage of 69.45%.
* Our findings also showed that **99.23%** of the earthquakes were caused by natural earthquake activity while **0.75%** was caused nuclear explosion.
* Year 2011 has the highest number of earthquake occurrence with a record of **713** earthquakes. Generally, the number of earthquakes was found to have increased over the years, with some fluctuations. However, this could be due to advancements in seismic monitoring technology and increased reporting rather than an actual increase in earthquake frequency.
* Geographically, we found that earthquakes are not evenly distributed around the world. Certain Oceania regions, such as Vanuatu, Fiji region, Tonga, Papua New Guinea, Solomon Island and Asian regions such as Japan, Indonesia experienced a higher concentration of seismic activity. This information can be crucial for disaster preparedness and risk assessment in vulnerable areas.

**Conclusion**

This python exploratory data analysis project has provided valuable insights into the earthquake dataset of magnitude 5.5+ and spanning from 1969 to 2018. By leveraging Python's libraries such as pandas, numpy, matplotlib, seaborn and the Jupyter notebook, we were able to gain a deeper understanding of the seismic activity during this extensive time period.

This earthquake dataset project serves as a foundation for further investigations, such as predictive modeling or advanced statistical analyses. By continuing to explore and analyze earthquake datasets, we can make significant strides in mitigating the impact of earthquakes and ensuring the safety and well-being of communities worldwide. Also, further research could be conducted to explore other relationships such as geological phenomena, climate factors, or other external variables in more detail.