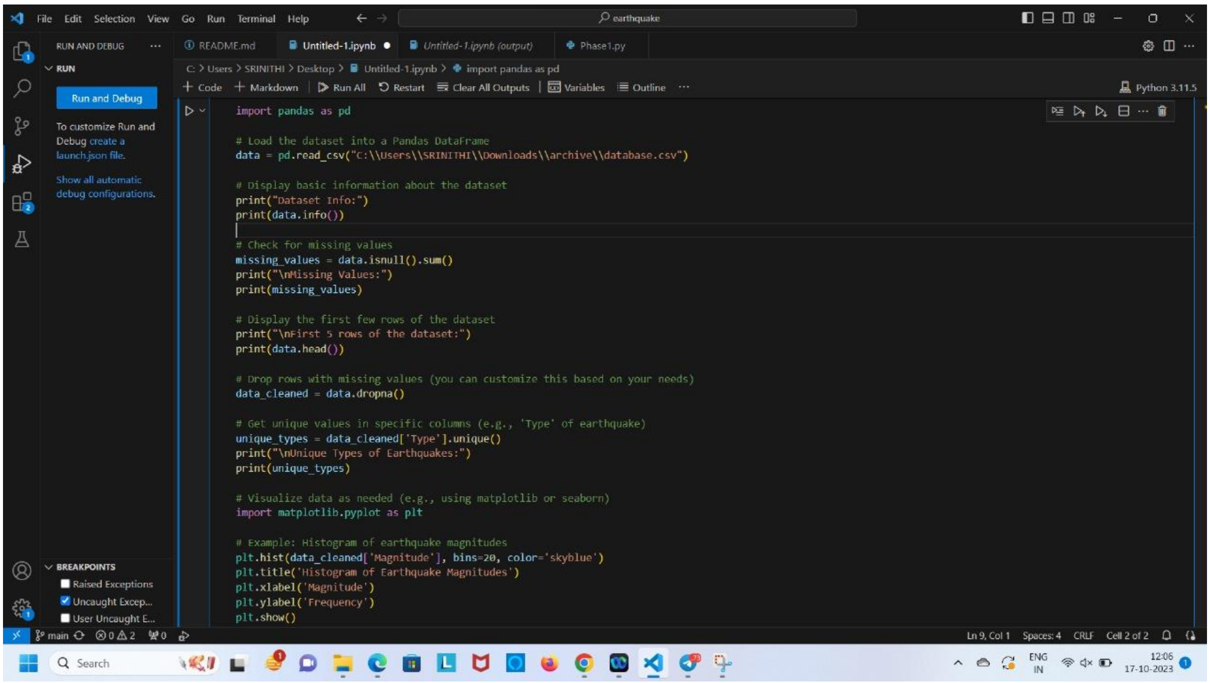
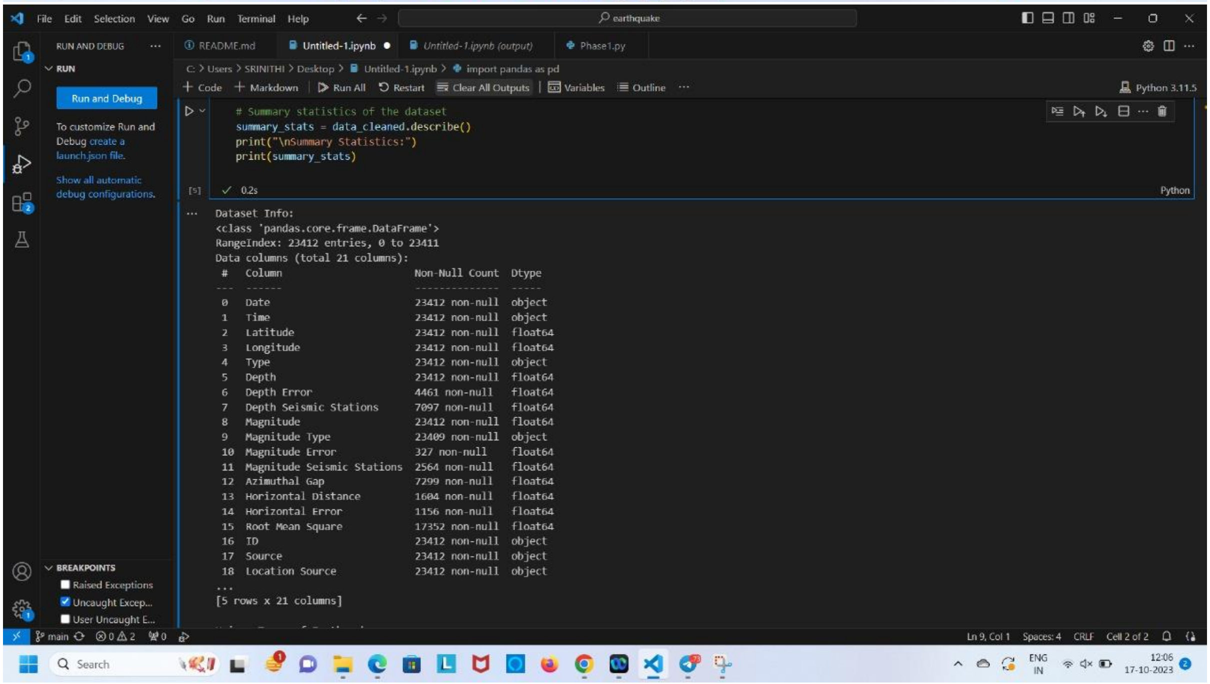
PROCESSING AND LOADING THE DATASET

- A dataset is a collection of data that may be used to train a model.
- How to load a dataset? Datasets are loaded from a dataset loading script that downloads and generates the dataset. However, you can also load a dataset from any dataset repository on the Hub without a loading script! Begin by creating a dataset repository and upload your data files. Now you can use the load_dataset () function to load the dataset.
- To load and preprocess the earthquake dataset from Kaggle, you can follow these steps:
-
- 1. Download the dataset: Visit the Kaggle website (https://www.kaggle.com/datasets/usgs/earthquake-database) and download the dataset file to your local machine.
-
- 2. Import necessary libraries: In a programming language like Python, import the required libraries such as pandas and numpy for data manipulation and analysis.
-
- 3. Load the dataset: Use the appropriate function from the pandas library (e.g., `read_csv()`) to load the dataset file into a pandas DataFrame.
-
- 4. Preprocess the dataset: Perform necessary preprocessing steps based on your analysis goals. This may include handling missing values, data cleaning, feature engineering, or transforming data types.
-
- 5. Analyze the dataset: Utilize the available analysis tools and techniques to gain insights from the dataset. This can involve statistical analysis, data visualization, or machine learning algorithms.
-
- Remember to refer to the documentation and resources provided by Kaggle for specific instructions on loading and preprocessing the earthquake dataset.

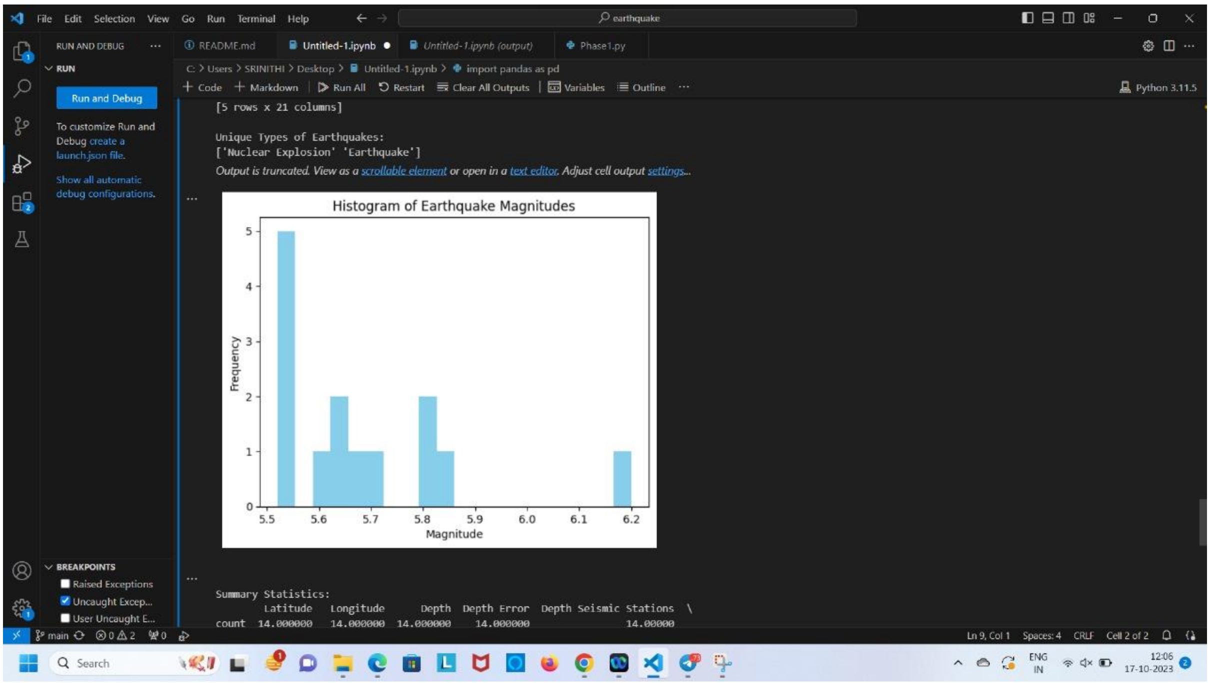


```python
import pandas as pd

# Load the dataset into a Pandas DataFrame
data = pd.read_csv("C:\\Users\\SRINITHI\\Downloads\\archive\\database.csv")

# Display basic information about the dataset
print("Dataset Info:")
print(data.info())

# Check for missing values
missing_values = data.isnull().sum()
print("\nMissing Values:")
print(missing_values)

# Display the first few rows of the dataset
print("\nFirst 5 rows of the dataset:")
print(data.head())

# Drop rows with missing values (you can customize this based on your needs)
data_cleaned = data.dropna()

# Get unique values in specific columns (e.g., 'Type' of earthquake)
unique_types = data_cleaned['Type'].unique()
print("\nUnique Types of Earthquakes:")
print(unique_types)

# Visualize data as needed (e.g., using matplotlib or seaborn)
import matplotlib.pyplot as plt

# Example: Histogram of earthquake magnitudes
plt.hist(data_cleaned['Magnitude'], bins=20, color='skyblue')
plt.title('Histogram of Earthquake Magnitudes')
plt.xlabel('Magnitude')
plt.ylabel('Frequency')
plt.show()
```



```python
# Summary statistics of the dataset
summary_stats = data_cleaned.describe()
print("\nSummary Statistics:")
print(summary_stats)
```

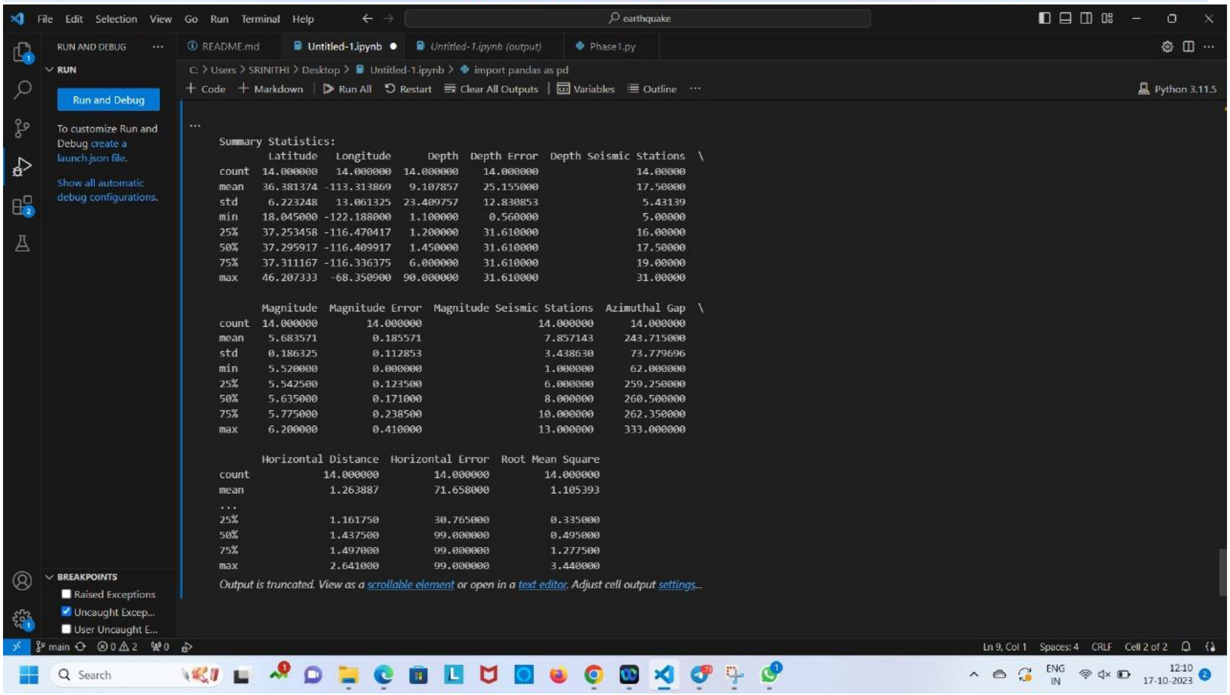
```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23412 entries, 0 to 23411
Data columns (total 21 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Date                        23412 non-null  object
 1   Time                        23412 non-null  object
 2   Latitude                    23412 non-null  float64
 3   Longitude                   23412 non-null  float64
 4   Type                        23412 non-null  object
 5   Depth                       23412 non-null  float64
 6   Depth Error                 4461 non-null   float64
 7   Depth Seismic Stations      7097 non-null   float64
 8   Magnitude                   23412 non-null  float64
 9   Magnitude Type              23409 non-null  object
 10  Magnitude Error             327 non-null    float64
 11  Magnitude Seismic Stations  2564 non-null   float64
 12  Azimuthal Gap               7299 non-null   float64
 13  Horizontal Distance         1604 non-null   float64
 14  Horizontal Error            1156 non-null   float64
 15  Root Mean Square            17352 non-null  float64
 16  ID                          23412 non-null  object
 17  Source                      23412 non-null  object
 18  Location Source             23412 non-null  object
...
[5 rows x 21 columns]
```


File Edit Selection View Go Run Terminal Help
earthquake
RUN AND DEBUG
RUN
Run and Debug
To customize Run and Debug create a launch.json file.
Show all automatic debug configurations.
BREAKPOINTS
□ Raised Exceptions
☑ Uncaught Excep...
□ User Uncaught E...
README.md
Untitled-1.ipynb
Untitled-1.ipynb (output)
Phase1.py
C: > Users > SRINITHI > Desktop > Untitled-1.ipynb > import pandas as pd
+ Code + Markdown Run All Restart Clear All Outputs Variables Outline
Python 3.11.5
Histogram of Earthquake Magnitudes
Frequency
0
1
2
3
4
5
5.5
5.6
5.7
5.8
5.9
6.0
6.1
6.2
Magnitude
main
Ln 9, Col 1 Spaces: 4 CRLF Cell 2 of 2
Search
ENG IN
12:06
17-10-2023

```
[5 rows x 21 columns]

Unique Types of Earthquakes:
['Nuclear Explosion' 'Earthquake']
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
Summary Statistics:
        Latitude   Longitude      Depth  Depth Error  Depth Seismic Stations  \
count  14.000000   14.000000  14.000000    14.000000                14.00000
```


File Edit Selection View Go Run Terminal Help
earthquake
RUN AND DEBUG
RUN
Run and Debug
To customize Run and Debug create a launch.json file.
Show all automatic debug configurations.
BREAKPOINTS
□ Raised Exceptions
☑ Uncaught Excep...
□ User Uncaught E...
README.md
Untitled-1.ipynb
Untitled-1.ipynb (output)
Phase1.py
C: > Users > SRINITHI > Desktop > Untitled-1.ipynb > import pandas as pd
+ Code + Markdown Run All Restart Clear All Outputs Variables Outline
Python 3.11.5
main
Ln 9, Col 1 Spaces: 4 CRLF Cell 2 of 2
Search
ENG IN
12:10
17-10-2023

```
Summary Statistics:
        Latitude   Longitude      Depth  Depth Error  Depth Seismic Stations  \
count  14.000000   14.000000  14.000000    14.000000                14.00000
mean   36.381374 -113.313869   9.107857    25.155000                17.50000
std     6.223248   13.061325  23.409757    12.830853                 5.43139
min    18.045000 -122.188000   1.100000     0.560000                 5.00000
25%    37.253458 -116.470417   1.200000    31.610000                16.00000
50%    37.295917 -116.409917   1.450000    31.610000                17.50000
75%    37.311167 -116.336375   6.000000    31.610000                19.00000
max    46.207333  -68.350900  90.000000    31.610000                31.00000

       Magnitude  Magnitude Error  Magnitude Seismic Stations  Azimuthal Gap  \
count  14.000000        14.000000                   14.000000      14.000000
mean    5.683571         0.185571                    7.857143     243.715000
std     0.186325         0.112853                    3.438630      73.779696
min     5.520000         0.000000                    1.000000      62.000000
25%     5.542500         0.123500                    6.000000     259.250000
50%     5.635000         0.171000                    8.000000     260.500000
75%     5.775000         0.238500                   10.000000     262.350000
max     6.200000         0.410000                   13.000000     333.000000

       Horizontal Distance  Horizontal Error  Root Mean Square
count            14.000000         14.000000         14.000000
mean              1.263887         71.658000          1.105393
...
25%               1.161750         30.765000          0.335000
50%               1.437500         99.000000          0.495000
75%               1.497000         99.000000          1.277500
max               2.641000         99.000000          3.440000
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...