**ANALYSIS OF EPICENTRE AND DEPTH OF EARTHQUAKES WITH FREQUENCY AND THEIR CORRESPONDING MAGNITUDES**

**GROUP MEMBERS :**

NAME : Sanjay Chari NAME : Aditya Shankaran

SRN : PES1201700278 SRN : PES1201700710

Department of CSE Department of CSE

PES University PES University

Bengaluru, India Bengaluru, India

Email Id : [sanjaychari2xy2@gmail.com](mailto:sanjaychari2xy2@gmail.com) Email Id :

adityashankaran10@gmail.com

**ABSTRACT :**

The following project will contain analysis of the latitude and longitude of earthquakes and their relationship with the corresponding magnitude. Through the given dataset, it is possible to find the expected frequency of an earthquake in a given latitude and longitude range and also find the expected magnitude through correlation analysis. In addition, the following project will attempt to find a correlation, if any, between the depth of an earthquake and its corresponding magnitude. Also involved in the project is prediction of magnitude, latitude and longitude of future earthquakes using regression.

**INTRODUCTION :**

The following project contains an analysis of data collected of major earthquakes from the year 1965. An earthquake is the shaking of the surface of the Earth, resulting from the sudden release of energy in the Earth's lithosphere that creates seismic waves. Earthquakes can range in size from those that are so weak that they cannot be felt to those violent enough to toss people around and destroy whole cities.

Before beginning the description of the project in detail, it would be necessary to introduce a few technical terms related to the analysis being carried out:

Latitude : Latitude refers to the angular distance of a place north or south of the earth's equator, usually expressed in degrees and minutes.

Longitude : Longitude refers to the angular distance of a place east or west of the Greenwich Meridien, usually expressed in degrees and minutes.

Magnitude : The magnitude of an earthquake refers to its intensity and is measured using the Richter Scale.

Depth : the depth offocus or focaldepth refers to the depth at which an earthquake occurs.

Depth error : Depth error refers to the error in the measurement of the depth of an earthquake.

Through the following project, we analyse the frequency of earthquakes with respect to latitude and longitude of their epicentre and also attempt to study the correlation, if any, between the latitude and longitude of an earthquake and its magnitude. In addition, we make an attempt to find the correlation, if any, between the depth of an earthquake and its corresponding magnitude.

All descriptive analysis related to the project has been performed on the Anaconda platform. The descriptive analysis of the attributes of the dataset consists of computing mean, median, variance, standard deviation, interquartile range for each of the attributes; along with suitable graphs for each of the attributes.

Also included in the project is a program that makes the user input a range of latitudes and longitudes and displays the probability , the expected magnitude and the expected depth of an earthquake happening in the area. The program has been coded taking into consideration the existing values of attributes in the dataset.

**DATASET :**

The dataset selected for the project contains data of all major earthquakes (magnitude 5.5 and above) from the year 1965. The given dataset was obtained from https://www.kaggle.com/usgs/earthquake-database, a website which contains many influential datasets that have scope for analysis. The given dataset is a large dataset consisting of 23412 rows and 21 columns. Each row of the dataset refers to an earthquake that occurred on a specific date with its epicentre at a given location, along with other data about it. Out of the 21 columns, analysis will mainly be focussed on 5 of them, namely;

Date

Latitude

Longitude

Magnitude

Depth

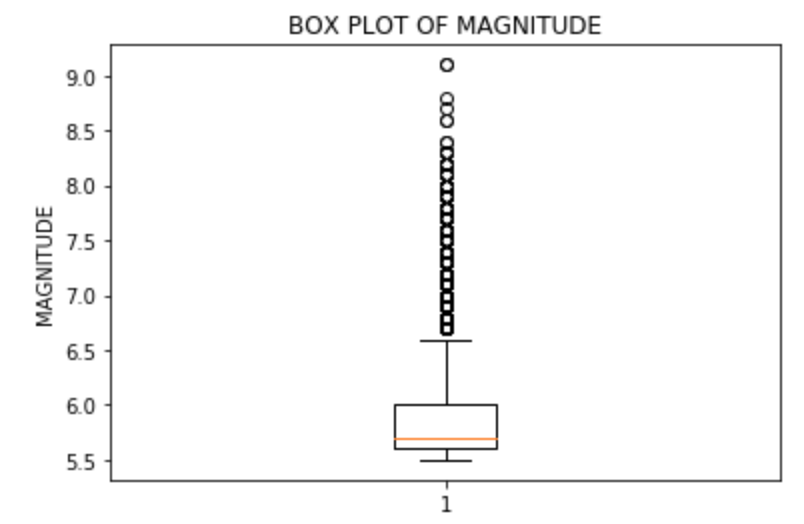
**PRE PROCESSING(Data Cleaning) :**

Data Cleaning is the process of detecting and removing faulty data records from a dataset. The term faulty in this case could refer to inconsistent, inaccurate, corrupt, irrelevant or dirty data. It could also refer to a typographical error.

Data Cleaning is generally achieved by removing the outliers from a given dataset. An outlier is any data value that falls outside the interval defined with lower limit as 25th percentile – 1.5 times the interquartile range and higher limit as 75th percentile + 1.5 times the interquartile range. Interquartile range is defined as the difference between the value of the 75th percentile and the 25th percentile of any dataset.

Outliers can easily be identified using a box plot. For example, a box plot of the

magnitude of earthquakes from the above dataset is given below.



In the above graph, the lowest horizontal edge of the box represents the 25th percentile or Q1. The red line represents the 50th percentile or Q2. The highest horizontal edge of the box represents the 75th percentile or Q3. IQR is Q3-Q1.

The horizontal lines below and above the box represent the minimum and maximum extremes respectively. The outliers in the data are represented by the circular dots.

Outliers can be removed using Anaconda platform. A code for removing outliers from the above dataset is given below.

import numpy as np

import pandas as pd

#reading from csv file

df=pd.read\_csv("/Users/Sanjay/Documents/SECOND YEAR STUFF/IDS/PROJECT/database.csv")

#removing outliers

q1,q2,q3=np.percentile(df.Magnitude,[25,50,75])

max=q3+1.5\*(q3-q1)

min=q1-1.5\*(q3-q1)

for i, j in df.Magnitude.iteritems():

if(j>max or j<min):

df.drop(df.index[i])

Generally, outliers should be removed from the dataset. However, in the above case, outliers are an integral part of analysis of the dataset as the outliers in the above dataset refer to high magnitude earthquakes that are by no means inconsequential or inaccurate data.

Thus, there is no need or purpose to perform data cleaning on the above dataset as the outliers are vital to the data analysis in this scenario.

**DESCRIPTIVE ANALYSIS :**

Descriptive analysis of a given dataset consists of finding measures of central tendency(mean, median, mode); measures of spread(range, interquartile range, variance, standard deviation); plotting suitable graphs to make inferences about the given data.

As Anaconda is capable of processing large data in a few seconds, sampling of the dataset is not necessary. Moreover, sampling of the above dataset is highly likely to lead to random errors as the dataset has positive skewness with respect to the magnitude and depth attributes which could lead to incorrect results. This skewness has not been eliminated by data cleaning as the higher values are essential to the analysis.

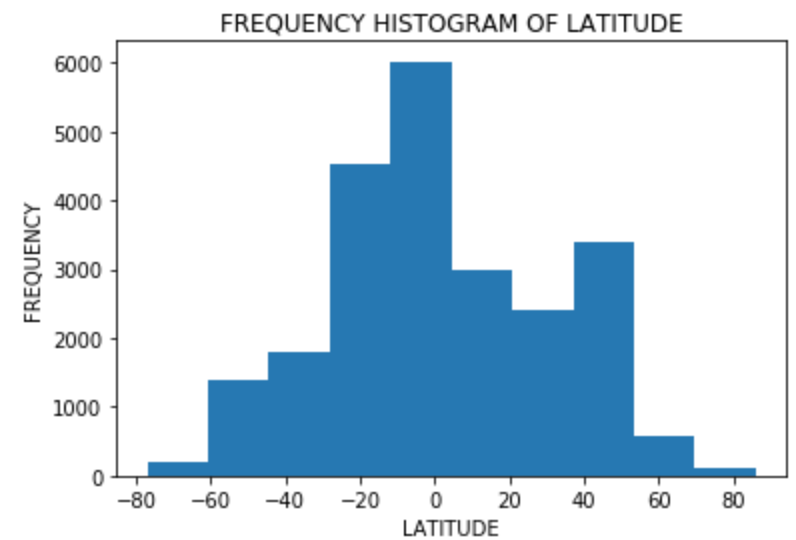
All descriptive analysis of the above dataset was performed using Anaconda software. The modules used in doing so are numpy, matplotlib, pandas, seaborn, scipy. Functions to find mean, median, standard deviation are present in the numpy module as numpy.mean(), numpy.median(), numpy.std(). All graph plotting functions are present in matplotlib.plot.

A table of summary of measures of central tendency and spread for attributes involved in the analysis is shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Column | Unit | Mean | Median | Range | IQR | Variance | S.D. |
| Latitude | Degrees | **1.68** | **-3.57** | **163.08** | **44.84** | **906.61** | **30.11** |
| Longitude | Degrees | **39.64** | **103.98** | **360** | **221.37** | **15752.76** | **125.51** |
| Magnitude | - | **5.88** | **5.7** | **3.6** | **0.4** | **0.18** | **0.42** |
| Depth | km | **70.77** | **33.0** | **701.1** | **39.48** | **15040.57** | **122.64** |

Descriptive analysis through graphs is a good way to make meaningful inferences on a data attribute. The descriptive analysis through graphs for important attributes is shown below.

**Latitude :**

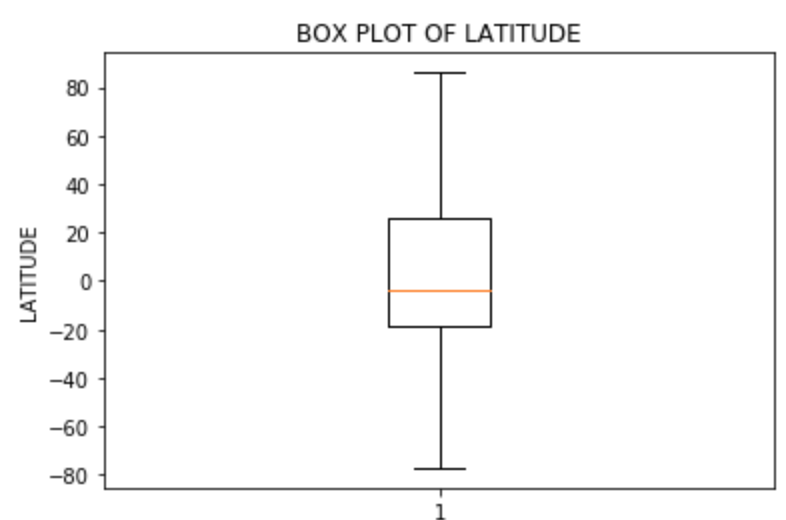


Through the above histogram, the following inferences were made :

Most of the earthquakes take place in the range of latitudes -5 to +5.

These latitudes are very close to the equator. Thus, it can be inferred that areas close to the equator have higher probability of receiving an earthquake than other areas.

The distribution is approximately normal in nature, except for a few inconsistencies. As can be seen in the graph, the number of earthquakes reduces as we approach the poles. Thus, we can infer that the probability of an earthquake happening in a given latitude reduces as we move towards the poles.

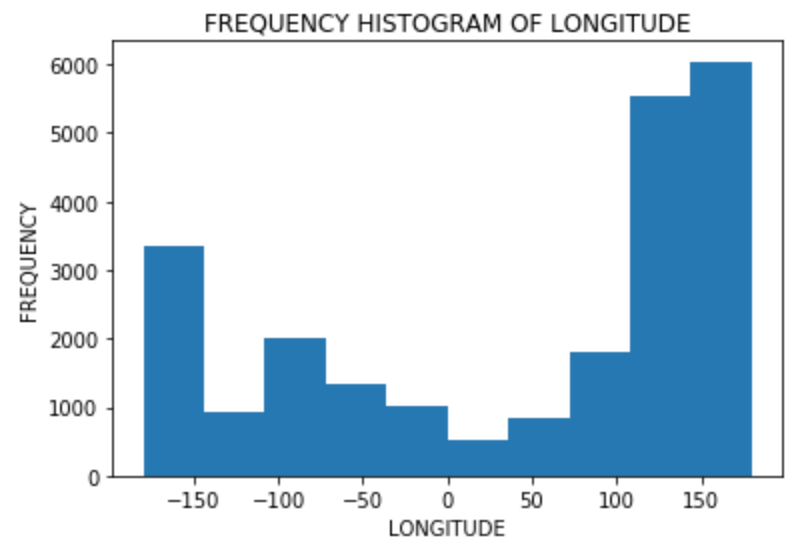


Through the above box plot, the following inferences were made :

As can be seen in the plot, Q2 is closer to Q1 than it is to Q3. Thus, we can conclude that most of the earthquakes are concentrated between Q2 and Q1 and the latitudes between Q2 and Q3 are sparse in number of earthquakes.

The median of the data is close to 0, as can be seen in the plot. Thus, we can infer that the central value of the earthquake latitudes is located around the equator.

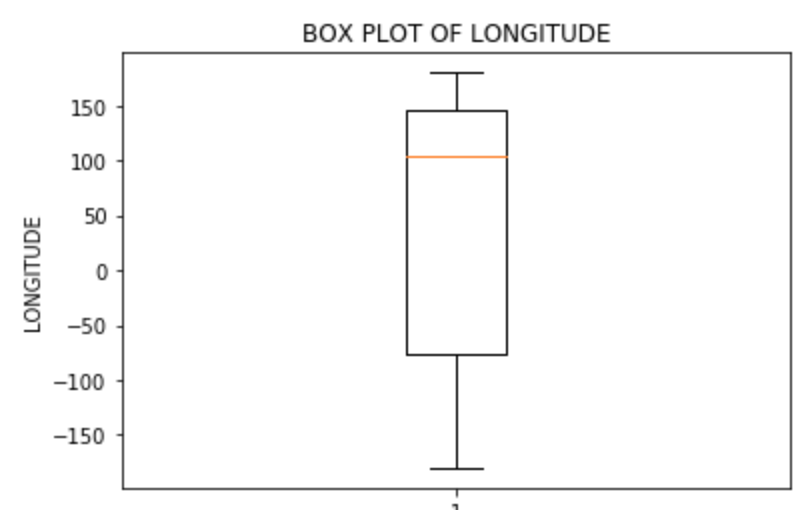
**Longitude :**



Through the above histogram, we can make the following inferences :

The mode class of the data seems to be 140-180. Thus, it can be inferred that this range of longitudes has the highest probability of receiving earthquakes among all longitudes.

The longitude with lowest number of earthquakes can be seen to be 0. 0 in the above case refers to the Greenwich meridian. Thus, it can be inferred that areas around the Greenwich meridian have lowest probability of receiving earthquakes.

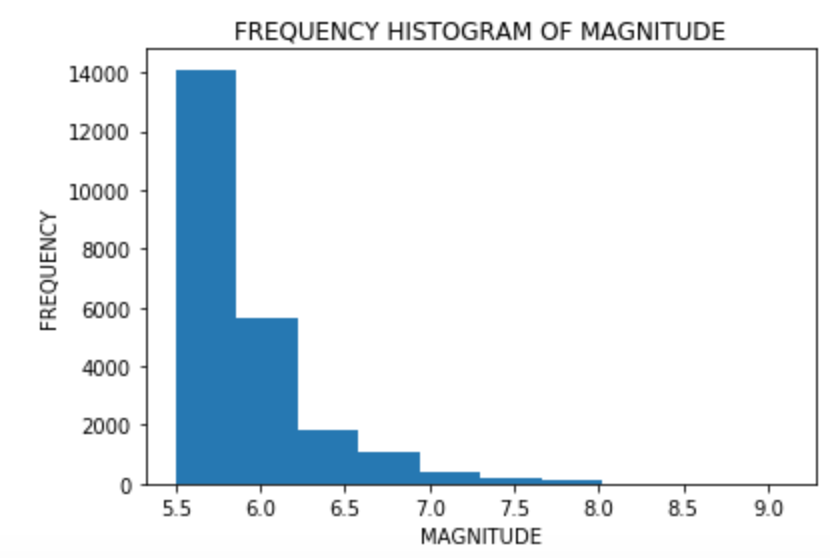
The above graph is moderately left skewed in nature. Thus, it can be inferred that higher longitudes receive more earthquakes than lower longitudes.

Through the above box plot, the following inferences were made :

Q2 is located very close to Q3. Thus, we can infer that most of the earthquakes took place in higher longitudes.

The higher extreme is closer to Q3 than the lower extreme is to Q1. Thus, we can infer that data is more concentrated closer to the higher extreme than the lower extreme.

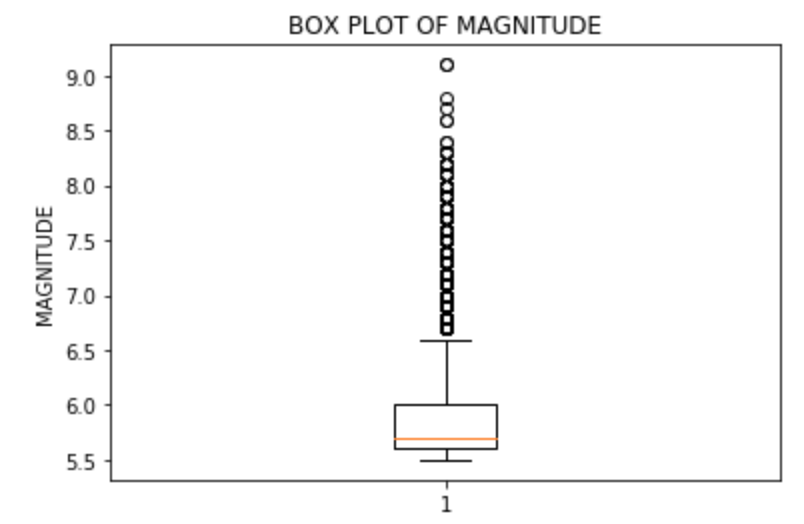
**Magnitude :**



Through the above histogram, the following inferences were made :

The mode class of the data is 5.5-5.75. Thus, it can be inferred that most of the earthquakes that occur are likely to be of low magnitude.

It can be inferred that high magnitude earthquakes are very rare as the graph is heavily right skewed.

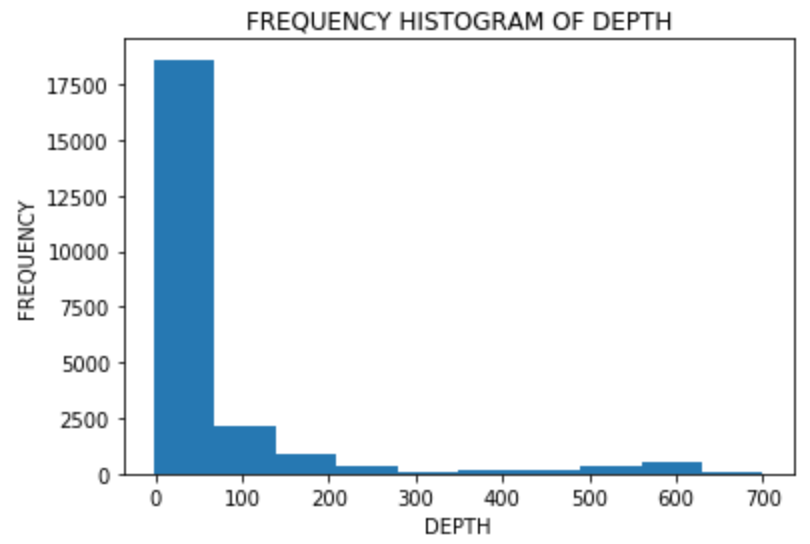


From the above box plot, the following inferences can be made :

The outliers represent the high magnitude earthquakes. As can be seen, there are countable number of outliers in the plot. Thus, it can be inferred that the number of high magnitude earthquakes are countable.

Q2 is very close to Q1 and the lower extreme is very close to Q1. Thus, it can be inferred that 50% of the earthquakes are between 5.5 and 5.75.

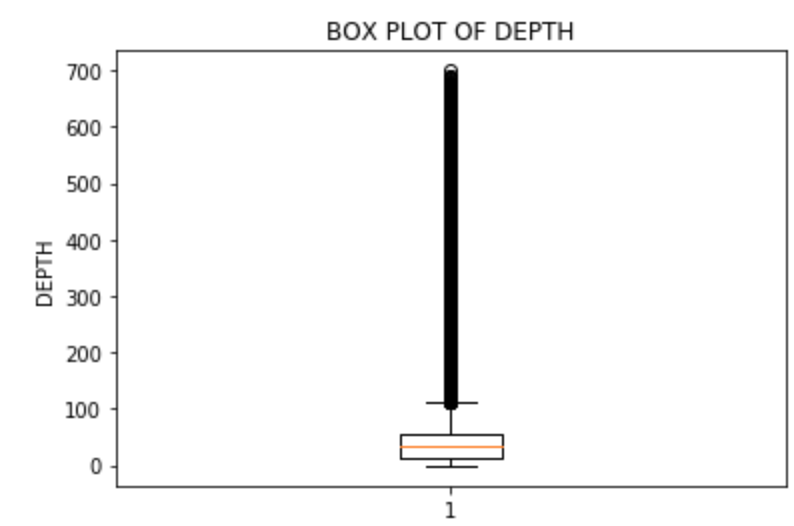
**Depth :**



From the above histogram, the following inferences can be made :

The mode class is 0-75. This implies that the highest number of earthquakes have their epicentre in this depth range.

The graph is heavily right skewed. This implies that greater depth earthquakes are very rare.

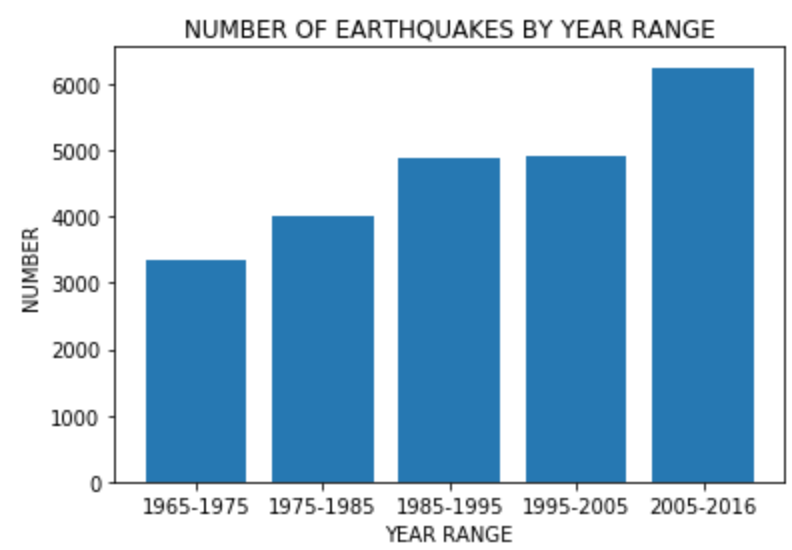


From the above box plot, the following inferences can be made :

The lower extreme,Q1,Q2,Q3 are all very close to each other. This implies that 75% of the earthquakes have depth of epicentre between 0 km and 60 km.

The number of outliers are very high in number. This implies that the number of high depth epicentre earthquakes are fairly substantial in number.

**Year :**

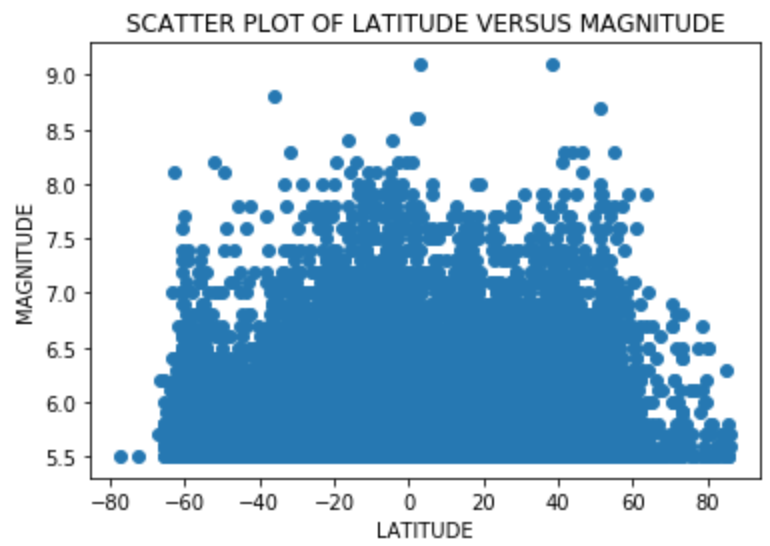


From the above bar graph, the following inferences were made.

The number of earthquakes in the years 1985-1995 and the years 1995-2005 appear to be almost the same.

The number of earthquakes appears to be monotonically increasing. Thus, it can be inferred that, as the years pass by, tectonic activity seems to be getting more and more violent.

**COMPARATIVE GRAPHS :**

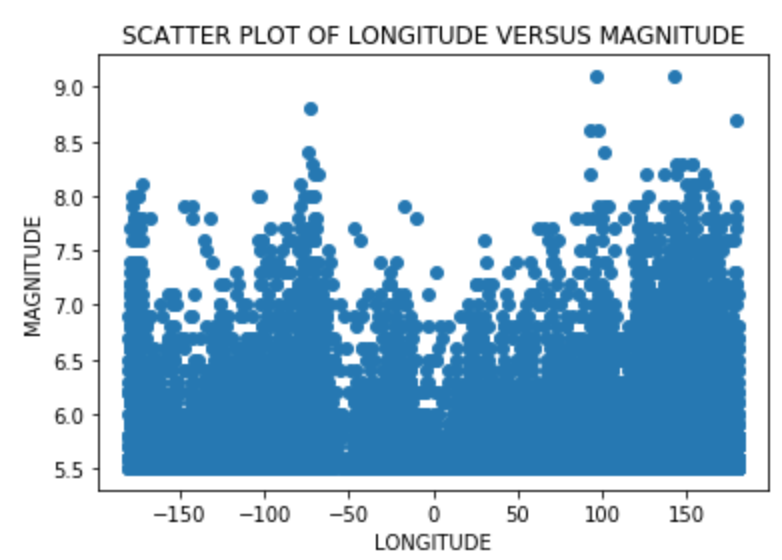


From the above scatter plot, the following inferences were made.

The earthquakes in the magnitude range 7.5 to 8.0 mainly seem to be clustered in the latitudes -20 to 0. This range also happens to be the approximate mode class for the frequency of earthquakes with respect to latitude. Thus, it can be inferred that the latitudes -20 to 0 are the most prone to both relatively high magnitude earthquakes and intermediate magnitude earthquakes. This could also point to a possible correlation between magnitude and frequency of earthquakes in a particular latitude.

The lower magnitude earthquakes are scattered evenly along all latitudes except for the poles(the left and right extremes). Thus, it can be inferred that low magnitude earthquakes have almost same likelihood of occurring at any latitude.

The left extreme has only 2 scatter points while the right extreme has a countable number of scatter points. Thus, it can be inferred that the Antarctic region is the safest when it comes to earthquakes and the Arctic region is a close second.



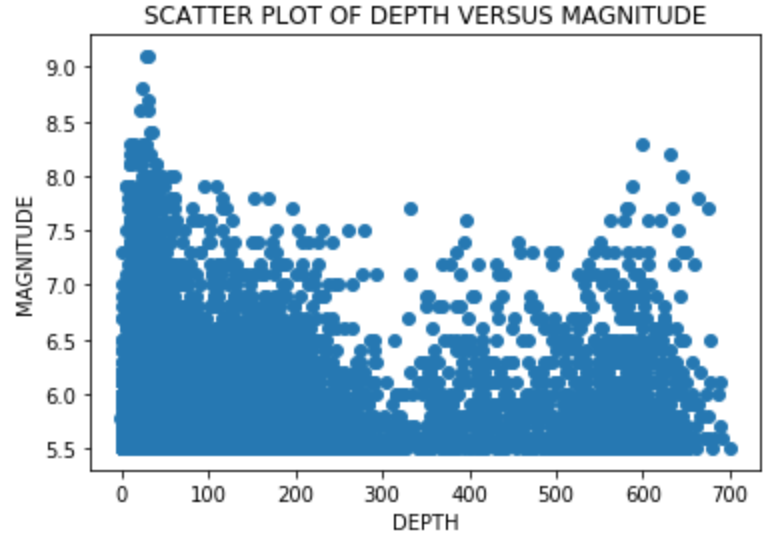
From the above scatter plot, the following inferences were made.

The earthquakes in the magnitude range 7.5-8.25 are clustered in the longitudes 120 to 150. This is close to the mode class of the frequency of earthquakes with respect to longitude. Thus, it can be inferred that the longitudes 120 to 150 are the most prone to both relatively high magnitude earthquakes and intermediate magnitude earthquakes. This could also point to a possible correlation between magnitude and frequency of earthquakes in a particular longitude.

The lower magnitude earthquakes appear to be scattered evenly along all longitudes. Thus, it can be inferred that low magnitude earthquakes have almost same likelihood of occurring at any longitude.

The range of longitudes -50 to 50 have no instance of an earthquake above magnitude 8.0 and only a countable number of instances of an earthquake above magnitude 7.0. Thus, it can be inferred that the probability of a high magnitude earthquake in the range of longitudes -50 to 50 is least when compared to other longitudes.

There seem to be five distinct local maxima in the graph where points are concentrated densely and also extend up to high magnitude. These are the longitudes -180, +180, -67, +100, +133(-180 and +180 refer to the left and right extremes of the graph). Thus, it can be inferred that the hotspot for earthquakes in general (high and low magnitude) is the set of longitudes {-180,-67,+100,+133,+180}.



From the above scatter plot, the following inferences were made.

High magnitude earthquakes seem to be clustered in the depth range 0 to 20. Thus, we can infer that all high magnitude earthquakes are most likely to have shallow epicentres.

Low magnitude earthquakes(below 5.75) appear to be evenly scattered across all depths. Thus, it can be inferred that there is no specific correlation between earthquakes with their magnitude below 5.75 and the depth of their epicentre.

The depth 335 appears to be local minima. Also, the depth range 300 to 350 has very few high magnitude earthquakes. Thus, it can be inferred that earthquakes with epicentres in the range 300-350 have very low probability of being high magnitude earthquakes.

The number of earthquakes, high or intermediate magnitude, appear to start increasing again in the range of depth 550 to 650. Thus, it can be inferred that earthquakes with their epicentres in the depth range 550 to 650 have a fairly high probability of being intermediate or high magnitude earthquakes.

**CUMULATIVE INFERENCES :**

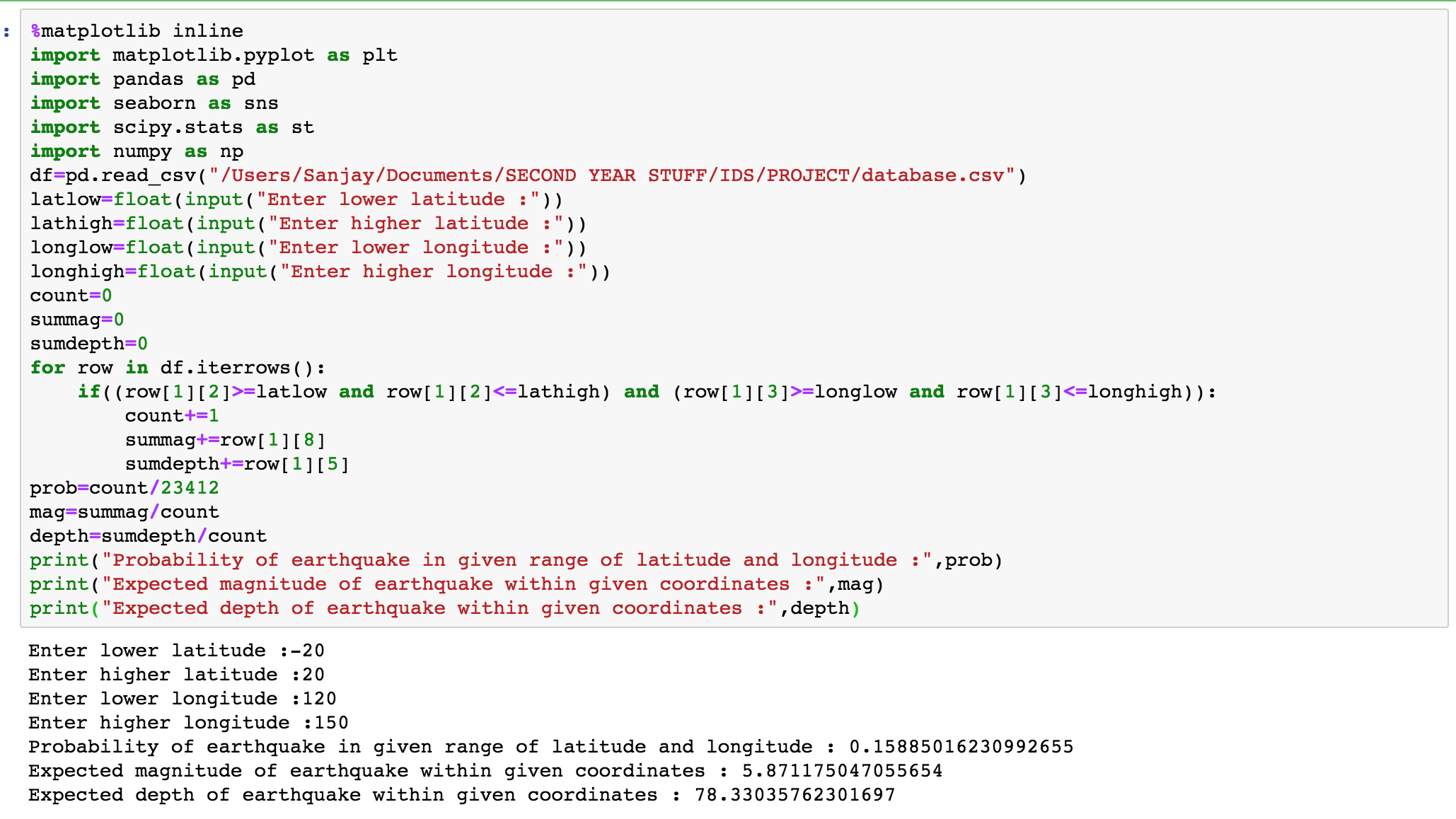
The latitude range -10 to 0 and the longitude range 140 to 150 fall in the mode class of frequency and magnitude with respect to both latitude and longitude. Thus, it can be inferred that any areas in the range of latitude and longitude coordinates { (x,y) | x∈(-10,0) ∧ y∈(140,150) } are extremely prone to earthquakes, both high magnitude and low magnitude.

The latitudes -90 to -70 and +70 to +90 and the longitude range 0 to 10 fall in the least class of frequency and magnitude with respect to both latitude and longitude. Thus, it can be inferred that areas close to the north pole or south pole and in the longitude range 0 to 10 have the least probability of an earthquake.

A very small number of earthquakes have the depth of their epicentres more than 50 km. This is the reason why most earthquakes, including the high magnitude ones, have their epicentre within 10 km below sea level. However, out of the few earthquakes that have their epicentre below 500 km below the ground, a significant percentage of them have high magnitude. Thus, it can be inferred that, given that an earthquake has a high depth epicentre (more than 500 km below the ground), the probability of it being high magnitude is relatively high.

**PREDICTIVE PROGRAM :**

Also included in the project is a program that takes range of latitudes and longitudes as input from the user and prints the probability, expected magnitude, expected depth of epicentre of an earthquake happening in the given range. A screenshot of a sample test case is displayed below. The .ipynb file of the code has been sent to the concerned faculty’s email id.



**INFERENTIAL ANALYSIS :**

**HYPOTHESIS TESTING :**

Hypothesis Testing consists of making a null hypothesis and an alternate hypothesis and checking for the validity of either of them.

In the project, the following hypotheses were made for all the columns in the dataset.

Null Hypothesis(H0) : μ=μ0

Alternate Hypothesis(Ha) : μ is not equal to μ0 ,

where μ0 represents the value of population mean for a given column of the dataset found using the np.mean() function.

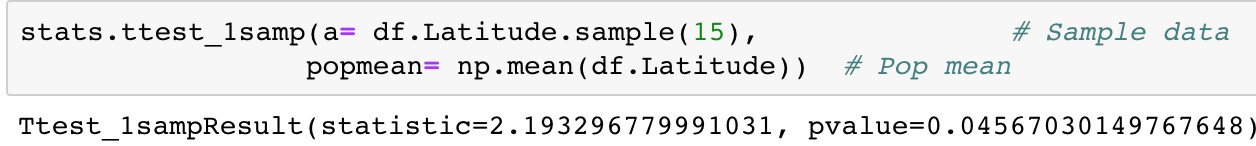
For the testing, the ttest\_lsamp() function found in the scipy.stats module has been used. The function returns two values, a t value and a p value.

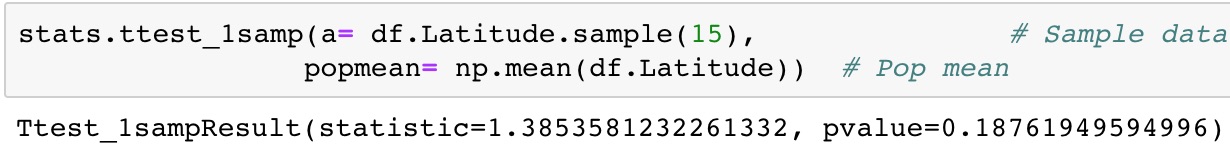
A confidence level of 95% has been used for the above test.

Thus, if the p value if less than 0.05, H0 will be rejected and Ha will be accepted to be true.

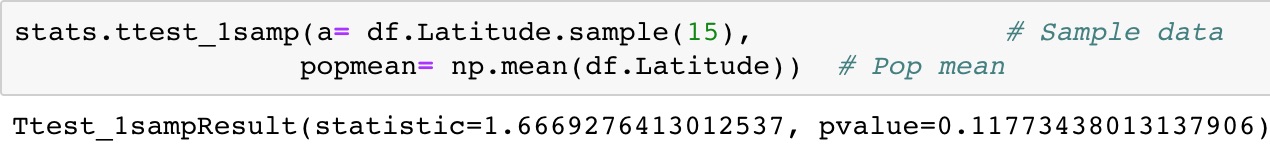
**Latitude :**

Listed below are various trials for hypothesis testing with respect to latitude.

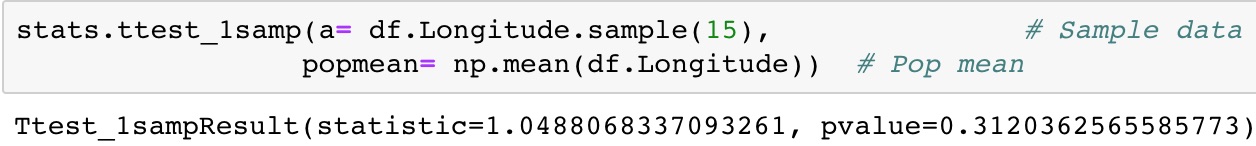
In the above code snippet, a p value of 0.0457 is obtained. Thus, the null hypothesis H0 can be rejected, as p<0.05, which is the significance level chosen. Thus, it can be concluded that the mean of the above sample would not be equal to the population mean μ0.



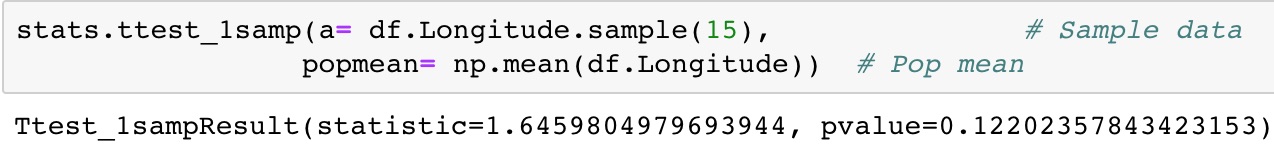
In the above code snippet, a p value of 0.1876 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.

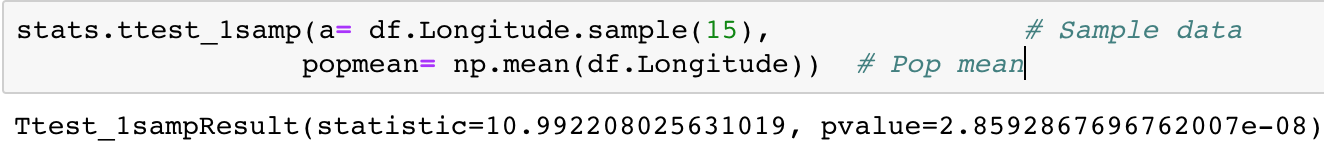
In the above code snippet, a p value of 0.1177 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.

**Longitude:**

Listed below are various trials for hypothesis testing with respect to longitude.  


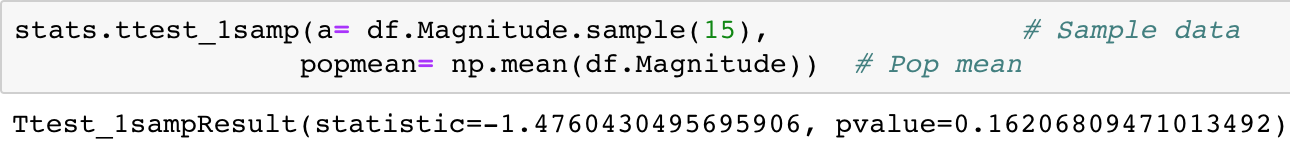
In the above code snippet, a p value of 0.3120 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.

In the above code snippet, a p value of 0.1220 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.

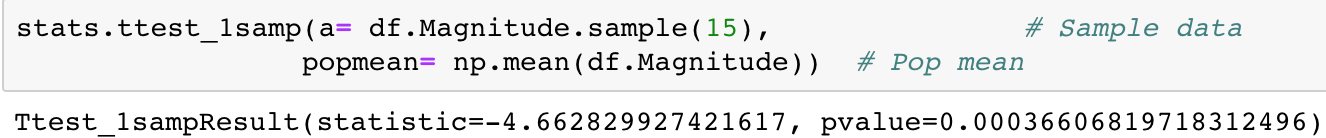


In the above code snippet, a p value of 2.85e-08 is obtained. Thus, the null hypothesis H0 can be rejected, as p<0.05, which is the significance level chosen. Thus, it can be concluded that the mean of the above sample would not be equal to the population mean μ0.

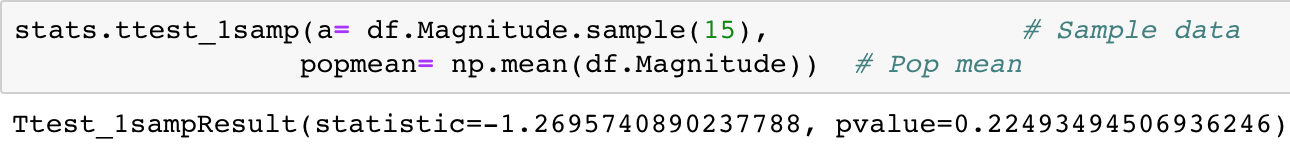
**Magnitude :**



In the above code snippet, a p value of 0.1621 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.



In the above code snippet, a p value of 0.0004 is obtained. Thus, the null hypothesis H0 can be rejected, as p<0.05, which is the significance level chosen. Thus, it can be concluded that the mean of the above sample would not be equal to the population mean μ0.

In the above code snippet, a p value of 0.2249 is obtained. As this p value is greater than the given significance level(0.05), it can be concluded that both the null hypothesis H0 and the alternate hypothesis Ha are plausible. The sample mean may or may not be equal to the population mean.

**CONFIDENCE INTERVAL :**

Confidence interval is a range of values within which the population parameter will most likely be found based on the value of a sample statistic.

Listed below is a code snippet along with a plot indicating confidence intervals for latitude.

As can be seen in the above image, the proportion of confidence intervals covering population mean is 0.94. This implies that the accuracy of confidence interval constructed in the above case is 94%, that is, 94% of confidence intervals constructed in the above case have the population mean within them.

**SIMPLE LINEAR REGRESSION :**

Simple Linear Regression is a useful approach to make approximately accurate predictions about future data values. It makes use of the least square fit method to construct a line plot that most accurately represents the trends in the data. In our dataset, simple linear regression has been used to make predictions about the expected frequency, latitude, longitude and magnitude of earthquakes in 2017.

**Frequency :**

The below linear regression plot was constructed for the frequency of earthquakes using Anaconda.

A close up of a map

Description automatically generated

As can be seen, the number of earthquakes seems to be increasing as the years pass by. The exact equation of the least square fit line extracted from the above plot is :



Using the above equation, predictions can be made about the number of earthquakes that would occur in the years to follow. This has been done in a later part of this article.

**Most Likely Latitude of Earthquakes :**

The below linear regression plot was constructed for the most likely latitude of earthquakes using Anaconda.

A close up of a map

Description automatically generated

As can be seen, the most likely latitude of earthquakes seems to be moving south as the years pass by. The exact equation of the least square fit line extracted from the above plot is :



Using the above equation, predictions can be made about the most likely latitude of earthquakes that would occur in the years to follow. This has been done in a later part of this article.

**Most Likely Longitude of Earthquakes :**

The below linear regression plot was constructed for the most likely latitude of earthquakes using Anaconda.

A screenshot of a cell phone

Description automatically generated

As can be seen, the most likely longitude of earthquakes seems to be moving eastward as the years pass by. The exact equation of the least square fit line extracted from the above plot is :

A close up of a mans face

Description automatically generated

Using the above equation, predictions can be made about the most likely longitude of earthquakes that would occur in the years to follow. This has been done in a later part of this article.

**Most Likely Magnitude of Earthquakes :**

The below linear regression plot was constructed for the most likely magnitude of earthquakes using

Anaconda.

.A screenshot of a cell phone

Description automatically generated

As can be seen, the most likely magnitude of earthquakes seems to be reducing as the years pass by. The exact equation of the least square fit line extracted from the above plot is :

Using the above equation, predictions can be made about the most likely magnitude of earthquakes that would occur in the years to follow. This has been done in a later part of this article.

**PREDICTIONS MADE USING LINEAR REGRESSION :**

Using the equations found in the above segment, predictions for the statistics of the year 2017 can be made with a fairly high level of accuracy.

The following values were predicted for the year 2017 :

A screenshot of a cell phone

Description automatically generated

Thus, using regression, it is possible to make fairly accurate predictions with respect to the dataset under study.

**RESULTS AND DISCUSSION :**

Earthquakes are a natural force of destruction and, through this article, an attempt was made to predict their behaviour so as to be aware of their whereabouts in the future. Data science techniques such as simple linear regression and graphs have helped us in doing so. A short list of all the inferences made from the beginning of this article are listed below.

The latitudes close to the equator and the longitudes close to the poles are extremely prone to earthquakes, both high magnitude and low magnitude.

The latitude ranges -90 to -70 and +70 to +90 and the longitude range 0 to 10 are the areas least prone to earthquakes.

If an earthquake has a high depth epicentre, it is most likely to be of high magnitude.

The total number of earthquakes occurring every year seems to be increasing over the years. This could be a foreboding omen that man’s interference with nature has finally begun to take its toll.

The epicentre of earthquakes has begun shifting south-eastward. This could indicate that the Earth’s tectonic plates are still changing and could indicate a change in the geography of the Earth close to Australia.

The magnitude of earthquakes seems to be going down over the years. This could be a consequence of the increase in frequency of earthquakes, as high magnitude earthquakes are very rare.

On the whole, as the number of earthquakes seem to be going up by the year, it is time for us to be alert and data science can help us in preventing a dystopian future.