Code: ABE 65100N **|** **Name:** Environmental Informatics **|** **CRN:** 15495
Name of Student: Jibin Joseph

Assignment 07
Graphical Analysis with Python

Describe the source and format of the input data (not more than half a page)

Solution:

The source (provenance) of the data is from USGS Earthquake Hazards Program. The input data is obtained in CSV (which is a "comma separated values" ASCII text file). The csv file contains 22 fields and uses a comma to separate values. Each line of the file is a data record. The first line contains the header of fields which is useful to call the data in pandas.

The data was obtained for the past 30 days and the downloaded file (all_month.csv) contained 16062 entries. This is obtained by info() function in pandas. The data ranges from 2020-03-18T22:06:02.610Z to 2020-04-17T21:40:57.740Z.

Describe the types of analysis conducted by your script (not more than half a page)

Solution:

The python script "program-07.py" performs different types of graphical analysis on the earthquake data. This is a good initial exercise to understand the data. The following graphical analysis are performed on the data.

- Plotting the histogram of field "mag" with a binwidth of 1 and for range from 0 to 10.
- Kernel Density Plot of field "mag"
- Scatter Plot of Latitude (on y-axis) and Longitude (on x-axis)
- Normalized CDF of Earthquake Depths
- Scatter Plot of Depth vs Magnitude
- QQ Plot of Earthquake Magnitude

Import the graphical analysis figures you developed in the previous section, and write a caption that describes the figure and addresses the hints provided for each figure. Captions should be consecutively numbered.

Solution:

Reasons for genfromtxt() will not work properly with this data file:

1. Numpy.genfromtxt will not work smoothly because NumPy arrays have one dtype for the entire array, while pandas DataFrames have one dtype per column.

This may be overcome (not with this data) by defining 22 datatypes as shown below.

```
eq_data2=np.genfromtxt("all_month.csv",
                       dtype=['str','float','float','float','float','str','int','int',
                              'float','float','str','str','str','str','str','float',
                              'float','float','int','str','str','str'],
                       names=True,delimiter=',')
```

2. The downloaded data is in csv format i.e comma separated. But, the above method, will not work as "place" (14th) field contains a comma. This will cause an error stating that "got 23 columns instead of 22". This lead me to use pd.read_table function.

1. Histogram of Earthquake Magnitude

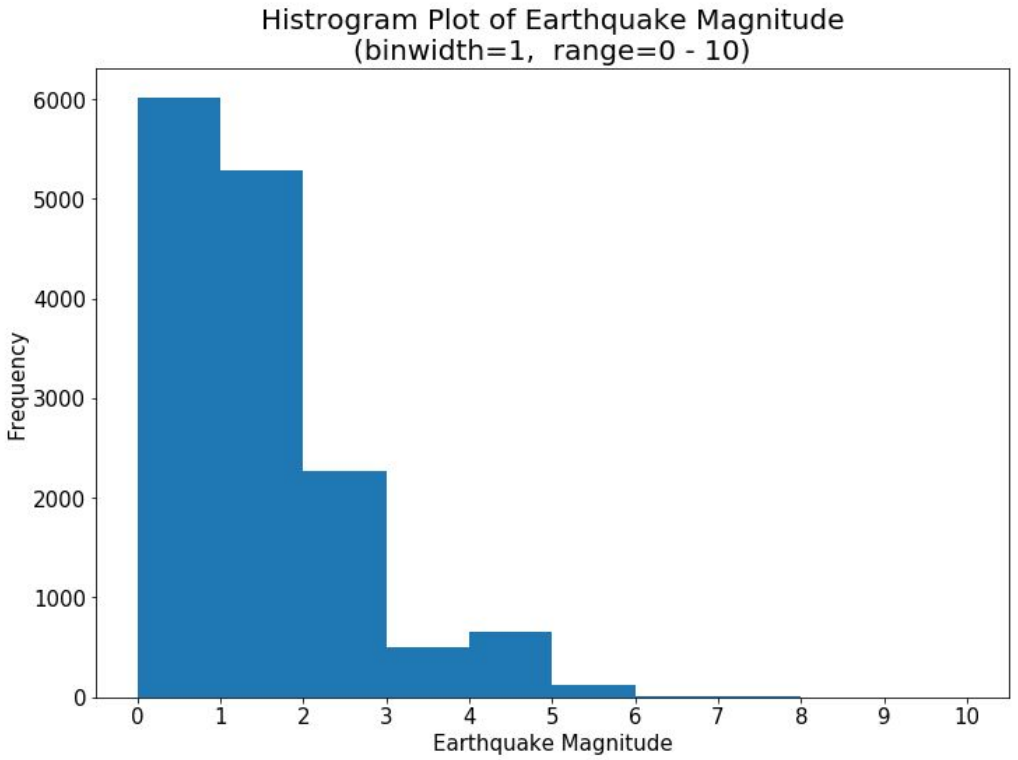

Figure 1: Histogram Plot of Earthquake with a binwidth of 1 magnitude and range from 0 to 10 magnitude.

When you plot the histogram, binwidth really controls the shape of histogram. If we choose a binwidth of 10, it will result in uniform shape.

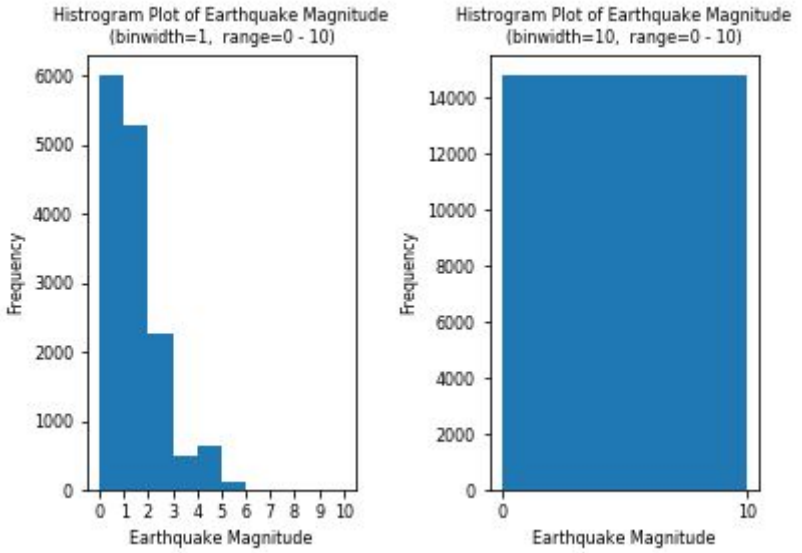

Figure 2: Effect of Binwidth

Further, due to selection of range from 0 to 10, the negatives magnitudes are filtered off. The negative magnitude earthquake is a very small earthquake which is not felt by humans. Also, a range from 0 to 20 would make the histogram looks more positively skewed.

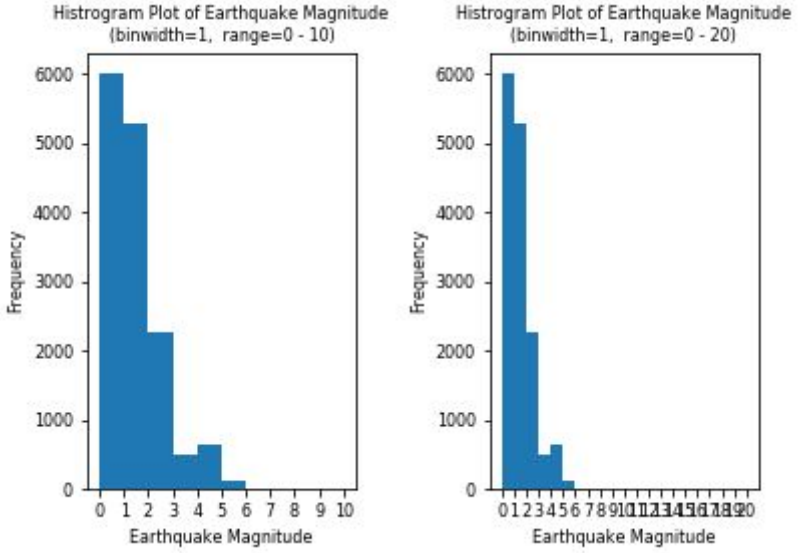

Figure 3: Effect of Range

From the binwidth of 1, it suggests a positively skewed (right-skewed) indicating tthat mean, median and mode all are different (for a normal distribution, all three are same)

2. Kernel Density Plot of Earthquake Magnitude

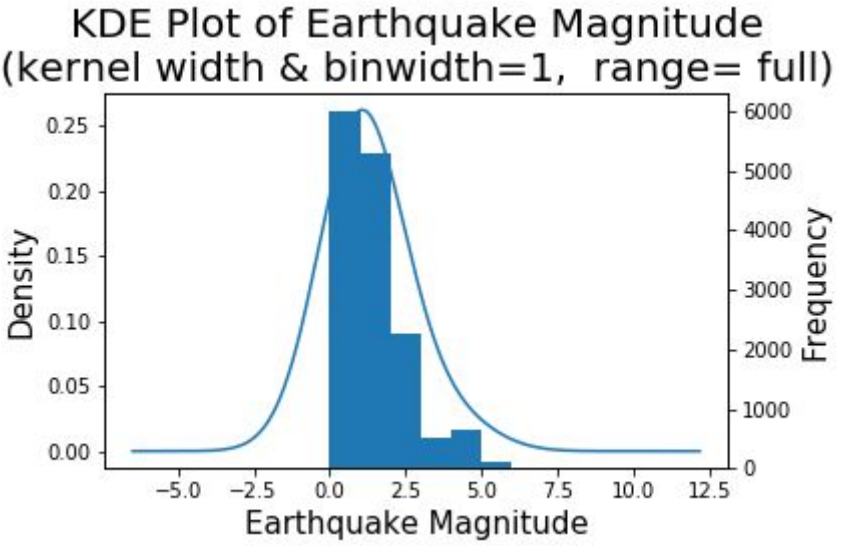

Figure 4: Kernel Density Plot of Earthquake Magnitude

plot.kde() function from pandas library uses **Gaussian Kernel** to generate the KDE plot. For the kernel width 1, a smooth curve is created.

If the width is not specified, it will use scott method to generate the plot and closely resemble the histogram showing one high peak and another peak.

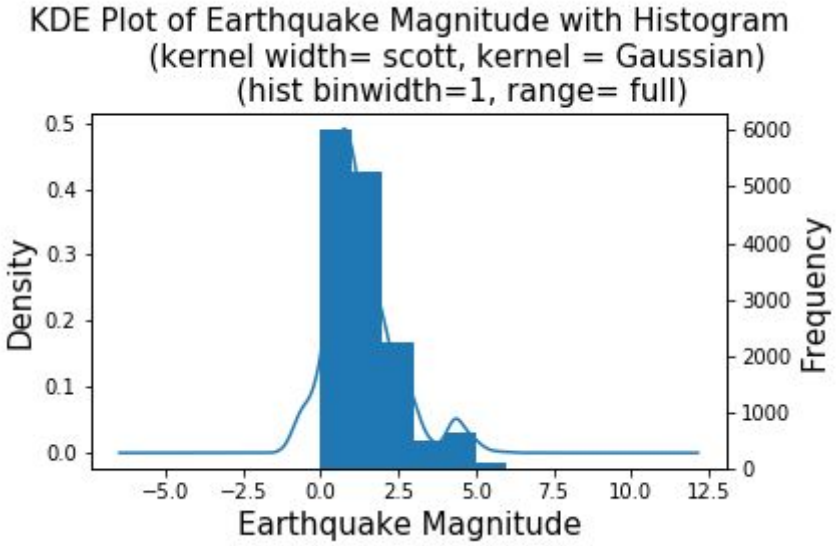

Figure 5: Kernel Density Plot of Earthquake Magnitude with bins calculated using 'scott' method

3. Scatter Plot of Latitude and Longitudes

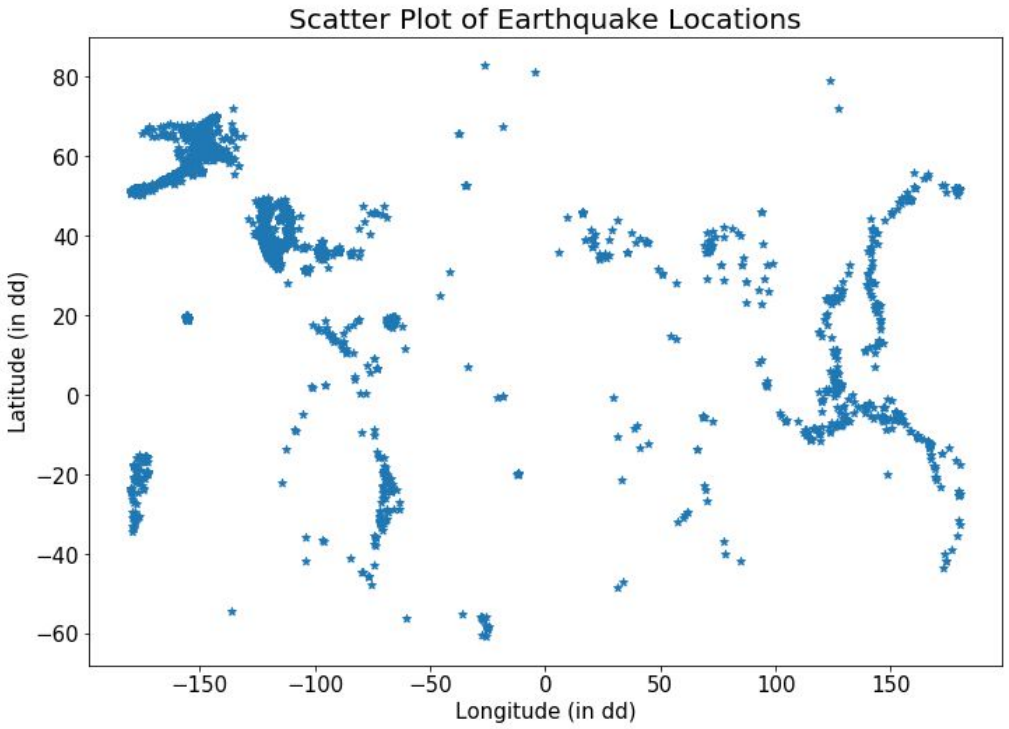

Figure 6: Scatter Plot showing the locations of earthquakes

The scatter plot shows the latitude on y axis (varies from -90 (S) to +90(N)) and longitude on x axis (varies from -180(W) to +180(E)). The data points span across the whole earth with Alaska and North America (west side) having more data points.

4. Normalized CDF of Earthquake Depths

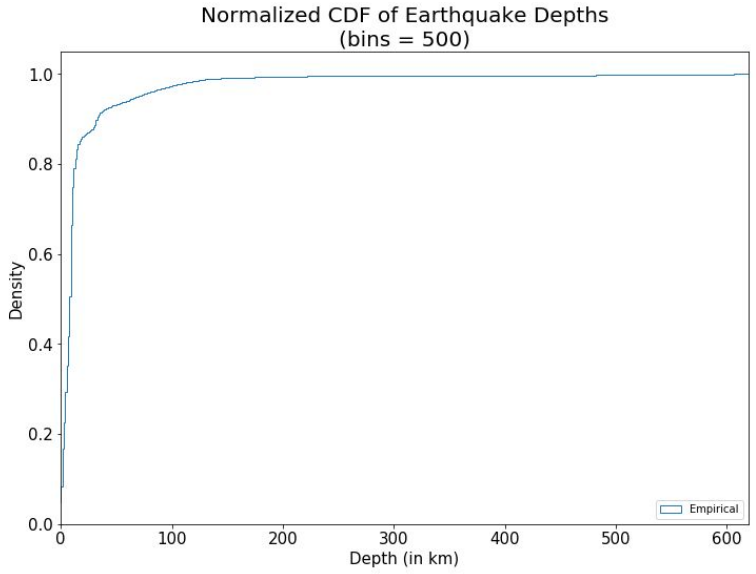

Figure 7: The plot shows the normalized cumulative distribution of Earthquake Depths

From the CDF Plot, it is evident that most of the earthquakes (that is recorded) occurred within a depth of 200 km. It may be also stated that depths greater than 300 km may be considered as outliers (Ref: https://www.andata.at/en/software-blog-reader/why-we-love-the-cdf-and-do-not-like-histograms-that-much.html)

5. Scatter Plot of Depth vs Magnitude

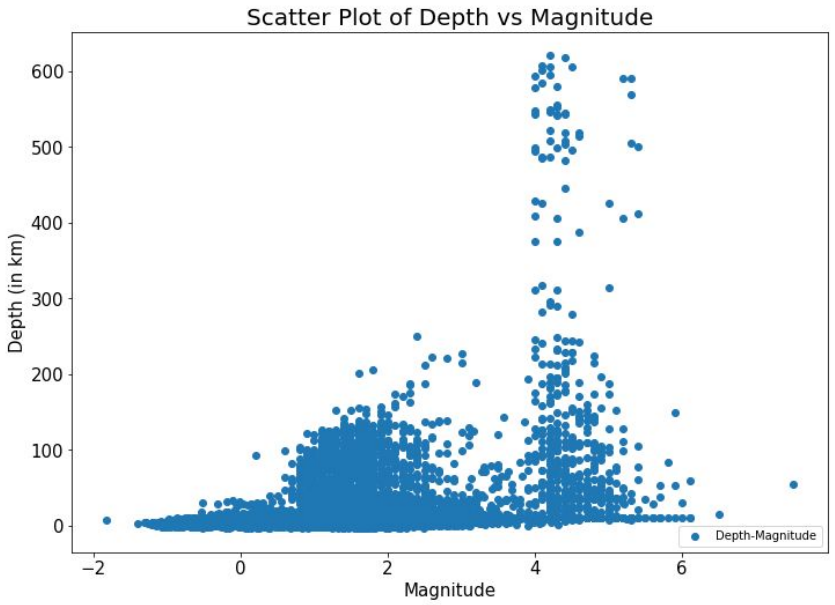

Figure 8: The scatter plot between depth (in km) and magnitude of earthquakes

From the figure, it can noticed that earthquakes of magnitudes less than 4 occurred at a shallow depth of ~200 km. But, higher magnitude earthquakes, originated at depth between 300 and 600 km. Earthquakes with magnitude less than 3 seems to be concentrated in depth of ~100 km.

6. QQ Plot of Earthquake Magnitude

For QQ plot, magnitudes greater than 0 was subsetted and different distributions were fitted. From the six fitted distributions, gumbel right skewed distribution is closely aligned with data. It was expected that pareto distribution would be a closer fit, but this could not be attained with downloaded data. Further, it seems that the higher values do not follow the distribution.

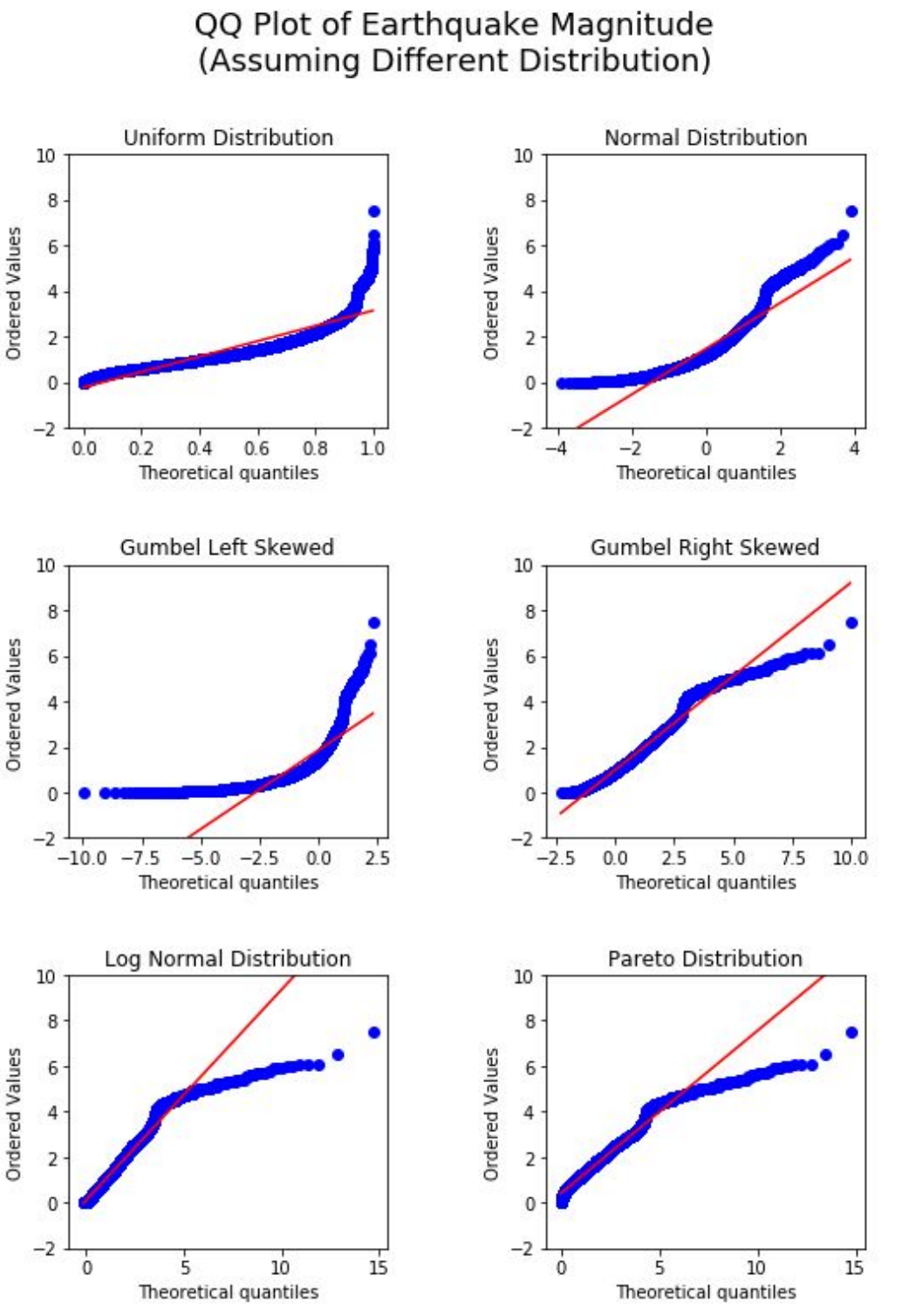

Figure 9: QQ Plot of Earthquake Magnitude