Big Data Programming Assignment 4 Name: Hama Earvin Bako

```
import numpy as np import matplotlib as mpl import pandas as pd import requests

url1_csv = "https://earthquake.usgs.gov/fdsnws/event/1/query?format=csv&starttime=2016-01-01&endtime=2017-01-01&minmagnitude=4"

url2_csv = "https://earthquake.usgs.gov/fdsnws/event/1/query?format=csv&starttime=2017-01-01&endtime=2018-07-01&minmagnitude=4"

url3_csv = "https://earthquake.usgs.gov/fdsnws/event/1/query?format=csv&starttime=2018-07-01&endtime=2019-10-02&minmagnitude=4"

df1 = pd.read_csv(url1_csv, usecols = ["time", "latitude", "longitude", "place", "mag", "depth"])

df1.tail(5)

df2 = pd.read_csv(url2_csv, usecols = ["time", "latitude", "longitude", "place", "mag", "depth"])

df2.head(5)

df3 = pd.read_csv(url3_csv, usecols = ["time", "latitude", "longitude", "place", "mag", "depth"])

df3.tail(5)

frames = [df1, df2, df3]
```

```
df = pd.concat(frames, ignore_index = True)
```

```
df.sort_values("time")
```

1) Use describe to get the basic statistics of all the columns (5 points)

```
df.describe()
```

2) Get the top 10 earthquakes by magnitude (5 points)

```
df_sorted = df.sort_values( by ='mag')
```

```
df_sorted.tail(5)
```

3) Handle all Null/empty data by filling it with zeros (10 points)

```
df = df.fillna(0)
```

4) Find the top 10 places where the strongest earthquakes occurred (15 points)

```
df_sortedplace = df.sort_values(by = "mag")
```

```
df_sortedplace.tail(10).place
```

5) Find the top 10 places where the weakest earthquakes occurred (15 points

```
df_sortedplace.head(10).place
```

6) On a per-year basis, use a bar chart to plot the number of earthquakes for each of the

following magnitude groups ranges: Group 1: [4,4.5), Group 2: [4.5,5), Group 3: [5,6), Group 4: [6,7), Group 5: (7,MAX]. Pay close attention to the group ranges. (20 points) Please add labels and colors to the plot.

```
df_group1 = df[df['mag'] < 4.5]

import matplotlib.pyplot as plt plt.hist(df_group1['mag'])

df_group2 = df[df['mag'] < 5]

df_group2 = df_group2[df_group2['mag'] >= 4.5]

plt.hist(df_group2['mag'])

df_group3 = df[df['mag'] < 7]

df_group3 = df_group3[df_group3['mag'] >= 6]

plt.hist(df_group3['mag'])

df_group4 = df[df['mag'] > 7]

plt.hist(df_group4['mag'])
```

7) Find the 10 countries with the highest number of earthquakes (30 points) (Note: Yes, this is only countries, not full place)

```
df_place = df['place']

df_place = df_place[df_place.duplicated() == True]
```

df_place.head(10)

```
In [268]: df_place.describe()
```

```
Out[268]: count                            13970
          unique                            5251
          top       South of the Fiji Islands
          freq                              1042
          Name: place, dtype: object
```

df_place.max()

8) Analyze the distribution of the Earthquake magnitudes. This is, make a histogram of the Earthquake count versus magnitude. Make sure to use a Logarithmic scale. What sort of relationship do you see? (20 points) Please add labels and colors to the plot.

df_mag = df['mag']

plt.hist(df_mag)

We can see from this histogram that eathquakes of higher magnitude occur less frequently than earthquakes

# of lower magnitude.

9) Analyze the distribution of the Earthquake depths. This is, make a histogram of the Earthquake count versus depth. Make sure to use a Logarithmic scale. What sort of relationship do you see? (20 points)

```
plt.hist(df['depth'])
```

We can see from this histogram that earthquakes of greater depth occur less frequently than eathquakes of lesser depth.

10) Visualize the locations of earthquakes by making a scatterplot of their latitude and longitude. (20 points)

```
rng = np.random.RandomState(0) x = df['latitude'] y = df['longitude'] colors = rng.rand(53020) size = 100*rng.rand(100) plt.scatter(x, y, c=colors, s = size, alpha=0.3, cmap='inferno') plt.colorbar()
```

In [ ]: