## Data Analysis

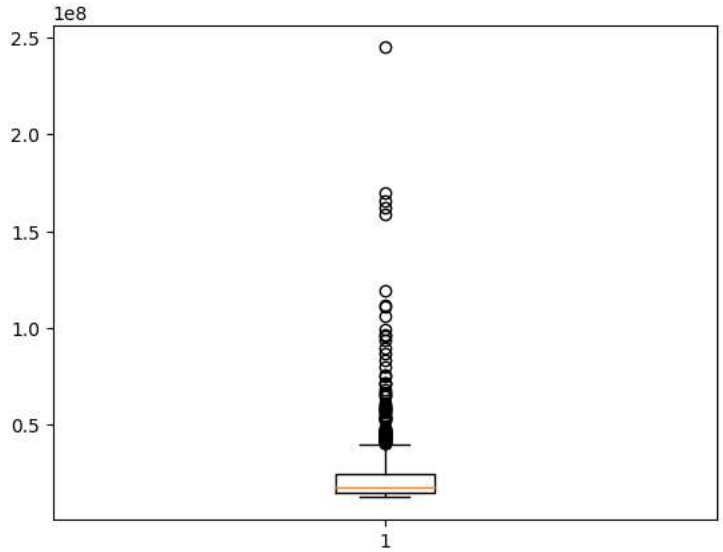
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px

df = pd.read_csv('/content/Global YouTube Statistics.csv', encoding = 'unicode_escape', on_bad_lines='skip')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995 entries, 0 to 994
Data columns (total 28 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   rank                                    995 non-null    int64
 1   Youtuber                                995 non-null    object
 2   subscribers                             995 non-null    int64
 3   video views                             995 non-null    float64
 4   category                                949 non-null    object
 5   Title                                   995 non-null    object
 6   uploads                                 995 non-null    int64
 7   Country                                 873 non-null    object
 8   Abbreviation                            873 non-null    object
 9   channel_type                            965 non-null    object
 10  video_views_rank                        994 non-null    float64
 11  country_rank                            879 non-null    float64
 12  channel_type_rank                       962 non-null    float64
 13  video_views_for_the_last_30_days        939 non-null    float64
 14  lowest_monthly_earnings                 995 non-null    float64
 15  highest_monthly_earnings                995 non-null    float64
 16  lowest_yearly_earnings                  995 non-null    float64
 17  highest_yearly_earnings                 995 non-null    float64
 18  subscribers_for_last_30_days            658 non-null    float64
 19  created_year                            990 non-null    float64
 20  created_month                           990 non-null    object
 21  created_date                            990 non-null    float64
 22  Gross tertiary education enrollment (%) 872 non-null    float64
 23  Population                              872 non-null    float64
 24  Unemployment rate                       872 non-null    float64
 25  Urban_population                        872 non-null    float64
 26  Latitude                                872 non-null    float64
 27  Longitude                               872 non-null    float64
dtypes: float64(18), int64(3), object(7)
memory usage: 217.8+ KB
```

## Central tendency and dispersion

Let us consider the attribute `subscribers` and analyse the measures of central tendency and spread.

```
attr = 'subscribers'
plt.boxplot(df[attr])
plt.show()
```

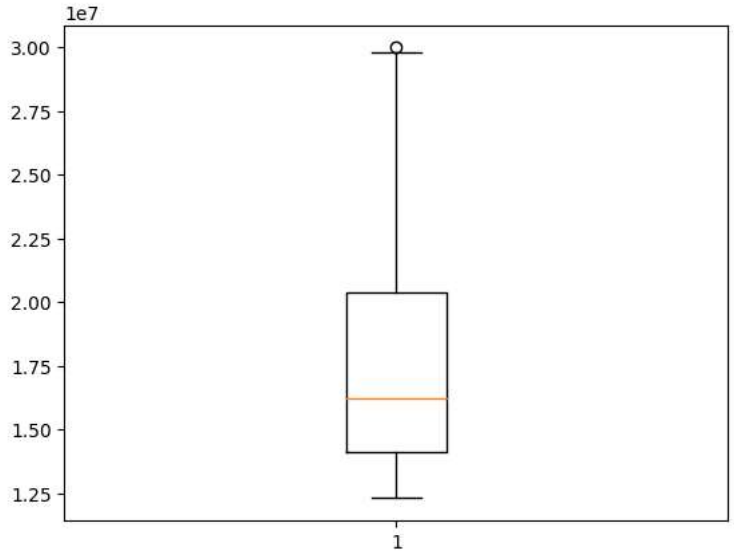


Set the threshold as 3e7 and remove all values greater than the threshold.

```
threshold=3e7
df.drop(df.loc[df[attr] > threshold].index, inplace=True)
df.shape
```

```
(825, 28)
```

```
plt.boxplot(df[attr])
plt.show()
```




**a) Central Tendency** - From the boxplot it is clear that `subscribers` attribute values are skewed to the right. For skewed data, Median is a better measure of central tendency than Mean since mean is sensitive to extreme values in the dataset.

```
df['subscribers'].median()
```

```
16200000.0
```

**b) Dispersion** - We will use Quartiles to analyse the spread of the attribute values.

```
print("Q1: ", np.quantile(df['subscribers'], 0.25))
print("Q2 (median): ", np.quantile(df['subscribers'], 0.5))
print("Q3: ", np.quantile(df['subscribers'], 0.75))
```

```
Q1:  14100000.0
Q2 (median):  16200000.0
Q3:  20400000.0
```

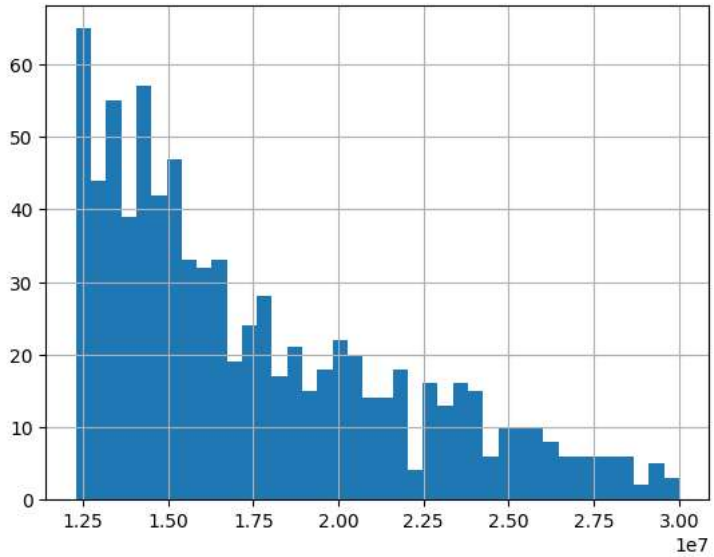
This means that 50% of the values are between 1.41e7 and 2.04e7.

## Data Visualization

Now let's plot a histogram to examine the distribution of values of the `subscribers` attribute.

This is a right skewed distribution since the `subscribers` values are for the top YouTube channels. If there was data for all Youtube channels, the distribution would have looked more like a Normal Distribution.:

```
df[attr].hist(bins=40)
plt.show()
```


60
50
40
30
20
10
0
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
1e7