In [ ]:
```
import re
import sys
import os
sys.path.append(os.path.abspath(os.path.join("../scripts")))
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display
import plots
# loading data set from csv file
sw_df = pd.read_csv("..\\DATA\\Swahili-Data-sets-ava\\Data\\text\\metadata_TZ.csv")
display(sw_df)
```

	filename	transcription	filepath	sample_rate	duration
0	SWH-05-20101106_16k-emission_swahili_05h30_-_0...	rais wa tanzania jakaya mrisho kikwete	SWH-05-20101106/SWH-05-20101106_16k-emission_s...	16000	3.140000
1	SWH-05-20101106_16k-emission_swahili_05h30_-_0...	yanayo andaliwa nami pendo pondo idhaa ya kisw...	SWH-05-20101106/SWH-05-20101106_16k-emission_s...	16000	3.100000
2	SWH-05-20101106_16k-emission_swahili_05h30_-_0...	inayokutangazia moja kwa moja kutoka jijini da...	SWH-05-20101106/SWH-05-20101106_16k-emission_s...	16000	3.650000
3	SWH-05-20101106_16k-emission_swahili_05h30_-_0...	juma hili bara la afrika limeshuhudia raia wa ...	SWH-05-20101106/SWH-05-20101106_16k-emission_s...	16000	3.900000
4	SWH-05-20101106_16k-emission_swahili_05h30_-_0...	wakipiga kura ya maoni ilikufanya mabadiliko ya	SWH-05-20101106/SWH-05-20101106_16k-emission_s...	16000	2.940000
...	...	...	...	...	...
10175	SWH-15-20110310_16k-emission_swahili_15h00_-_1...	na somo lile lililopokelewa kule kenya	SWH-15-20110310/SWH-15-20110310_16k-emission_s...	16000	2.500062
10176	SWH-15-20110310_16k-emission_swahili_15h00_-_1...	ambapo mtu aliyeshindwa kwenye uchaguzi	SWH-15-20110310/SWH-15-20110310_16k-emission_s...	16000	2.910000
10177	SWH-15-20110310_16k-emission_swahili_15h00_-_1...	ni kauli yake mchambuzi wa masuala ya siasa	SWH-15-20110310/SWH-15-20110310_16k-emission_s...	16000	2.950000
10178	SWH-15-20110310_16k-emission_swahili_15h00_-_1...	mwanasheria anayemtetea rais wa zamani wa liberia	SWH-15-20110310/SWH-15-20110310_16k-emission_s...	16000	2.590000
10179	SWH-15-20110310_16k-emission_swahili_15h00_-_1...	na kesi yake ya kubadilishana almasi na silaha...	SWH-15-20110310/SWH-15-20110310_16k-emission_s...	16000	5.010000

10180 rows × 5 columns

In [ ]:
```
# viewing how many values we have in the sample rate column of our dataset
sw_df['sample_rate'].value_counts()
```

Out[ ]:
```
16000    10180
Name: sample_rate, dtype: int64
```

In [ ]:
```
# viewing data information
sw_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10180 entries, 0 to 10179
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   filename       10180 non-null  object 
 1   transcription  10180 non-null  object 
 2   filepath       10180 non-null  object 
 3   sample_rate    10180 non-null  int64  
 4   duration       10180 non-null  float64
dtypes: float64(1), int64(1), object(3)
memory usage: 397.8+ KB
```

In [ ]:
```
# selecting two columns for detecting any outliers in the audio and text fies related
dur_sw_df = sw_df[['duration', 'filename']]
```

In [ ]:
```
dur_sw_df.describe()
```

Out[ ]:

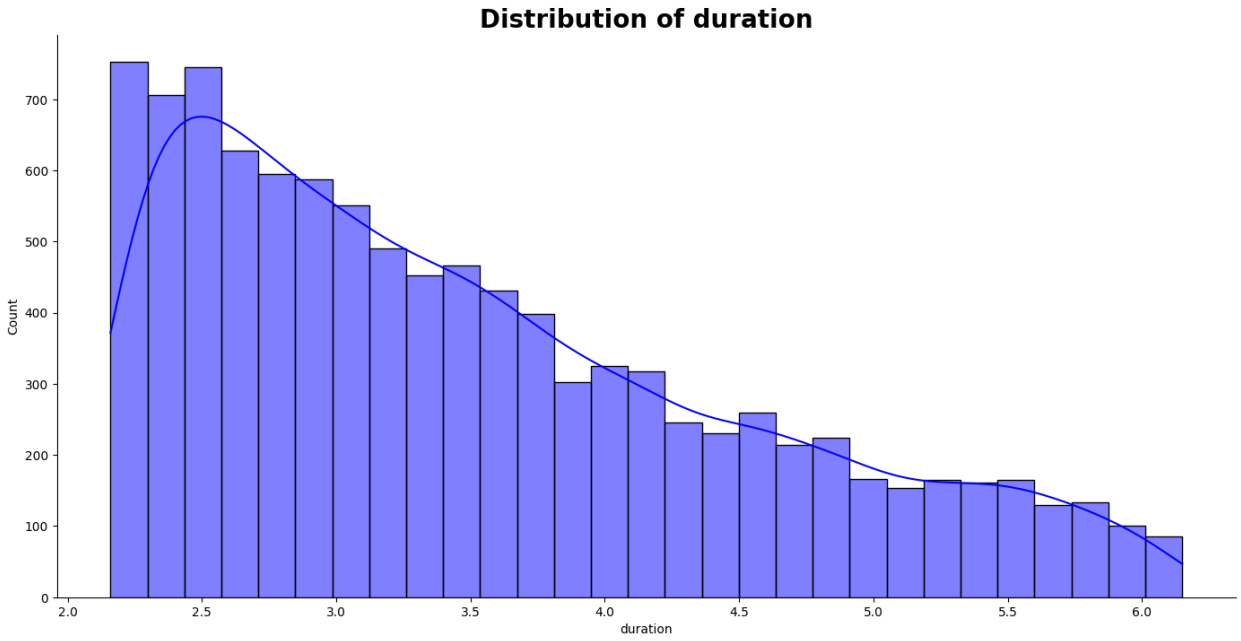
	duration
count	10180.000000
mean	3.504845
std	1.024975
min	2.159750
25%	2.650000
50%	3.279906
75%	4.179953
max	6.150000

The maximum duration is 6.1 seconds

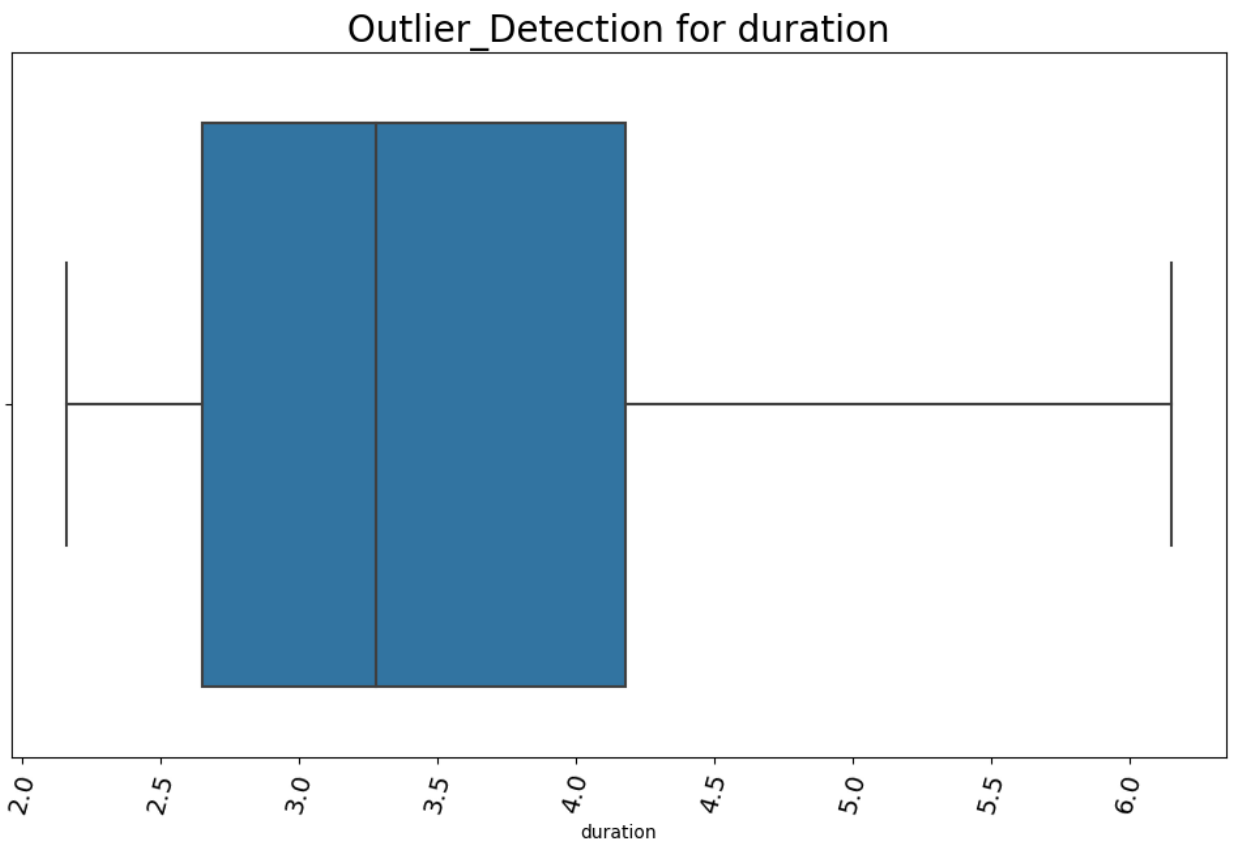
The Minimum duration is 2.1 seconds

In [ ]:
```
# the distribution of the duration shown in a histplot
plots.plot_hist(dur_sw_df, 'duration', 'blue')
```

```
<Figure size 900x700 with 0 Axes>
```



In [ ]:
```
plots.plot_box(dur_sw_df, 'duration', 'Outlier_Detection for duration')
```



The Above plot shows us that there are no outliers.

Pre Processing using regular expression...

In [ ]:
```
#  Removing any empty spaces in the data set specifically transcription column

for transcription in sw_df:
   sw_df['transcription'] = sw_df['transcription'].str.replace(r'<UNK>.*',' ',regex=True)

sw_df.to_csv('metadata_analyzed_1.csv')
```