Import Library

In [35]:
```
import numpy as np
import pandas as pd
```

Import Datsbase

In [36]:
```
df = pd.read_csv('weight-height.csv')
```

In [37]:
```
df.head()
```

Out[37]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

In [38]:
```
df.shape
```

Out[38]:
```
(10000, 3)
```

Describe Height

In [39]:
```
df['Height'].describe()
```

Out[39]:
```
count    10000.000000
mean        66.367560
std          3.847528
min         54.263133
25%         63.505620
50%         66.318070
75%         69.174262
max         78.998742
Name: Height, dtype: float64
```

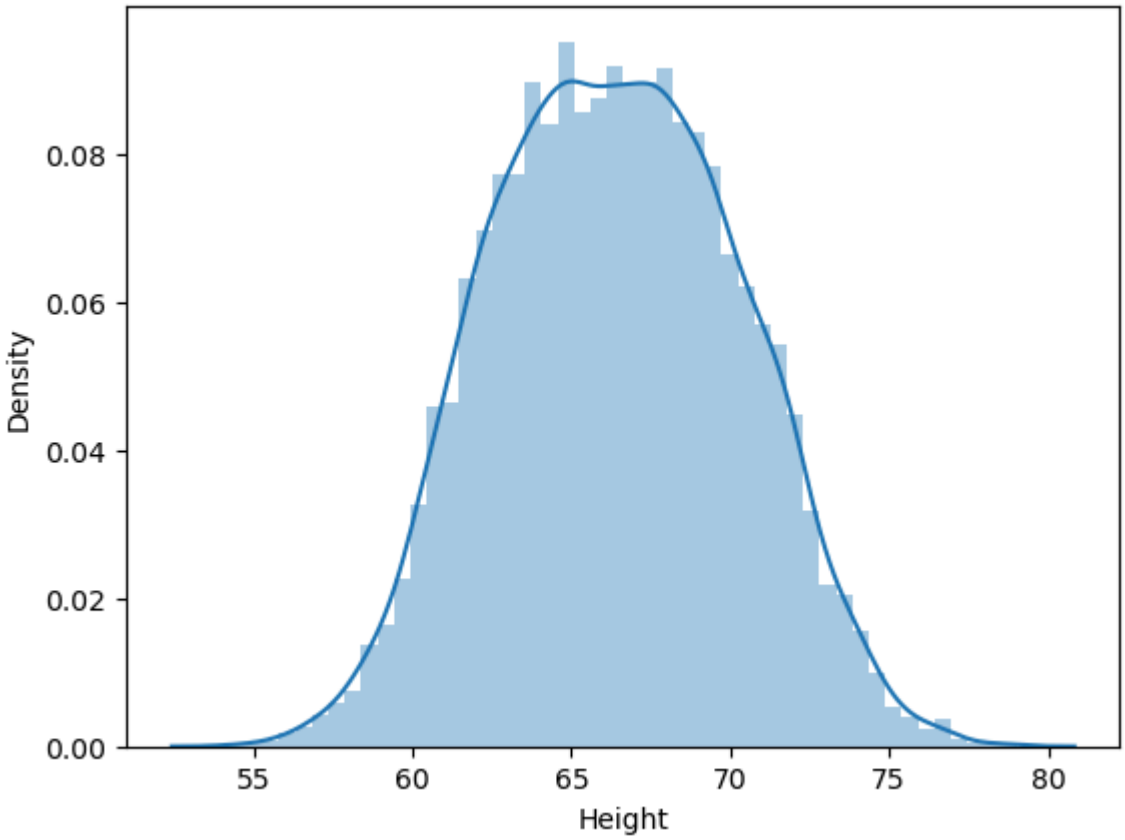
Import Seaborn and Plot

In [40]:
```
import seaborn as sns
```

In [41]:
```
sns.distplot(df['Height'])
```

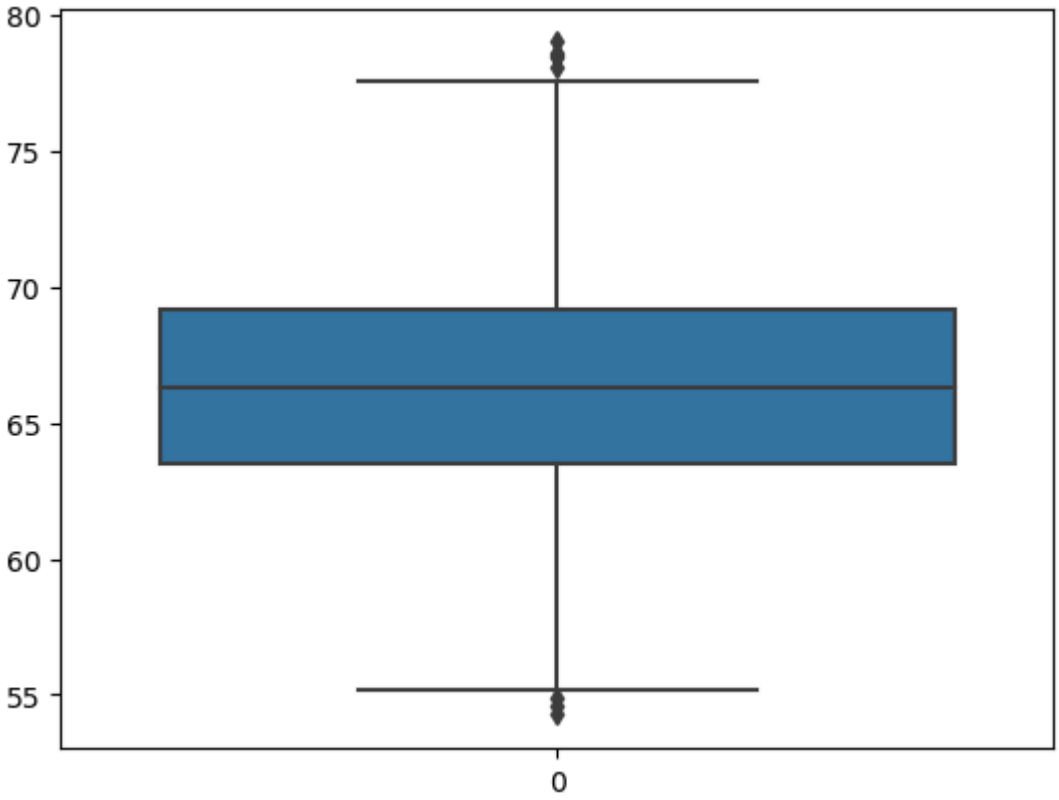
```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_11628\3945773010.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['Height'])
```

Out[41]:
```
<Axes: xlabel='Height', ylabel='Density'>
```



In [42]:
```
sns.boxplot(df['Height'])
```

Out[42]: <Axes: >



Define Upper Limit

In [43]:
```
upper_limit = df['Height'].quantile(0.99)
upper_limit
```

Out[43]: 74.7857900583366

Define Lower Limit

In [44]:
```
lower_limit = df['Height'].quantile(0.01)
lower_limit
```

Out[44]: 58.13441158671655

Set Range: If criteria is not full fill

In [45]:
```
df[(df['Height'] >= 74.78) | (df['Height'] <= 58.13)]
```

Out[45]:

	Gender	Height	Weight
23	Male	75.205974	228.761781
190	Male	76.709835	235.035419
197	Male	75.944460	231.924749
202	Male	75.140821	224.124271
215	Male	74.795375	232.635403
...	...	...	...
9761	Female	56.975279	90.341784
9825	Female	55.979198	85.417534
9895	Female	57.740192	93.652957
9904	Female	57.028857	101.202551
9978	Female	57.375759	114.192209

201 rows × 3 columns

Define new dataframe with limit

Triming throught percentile method

In [49]:
```
new_df = df[(df['Height'] <= 74.78) & (df['Height'] >= 58.13)]
```

In [50]:
```
new_df['Height'].describe()
```

Out[50]:
```
count    9799.000000
mean       66.363507
std         3.644267
min        58.134496
25%        63.577147
50%        66.317899
75%        69.119859
max        74.767447
Name: Height, dtype: float64
```

In [51]:
```
df['Height'].describe()
```

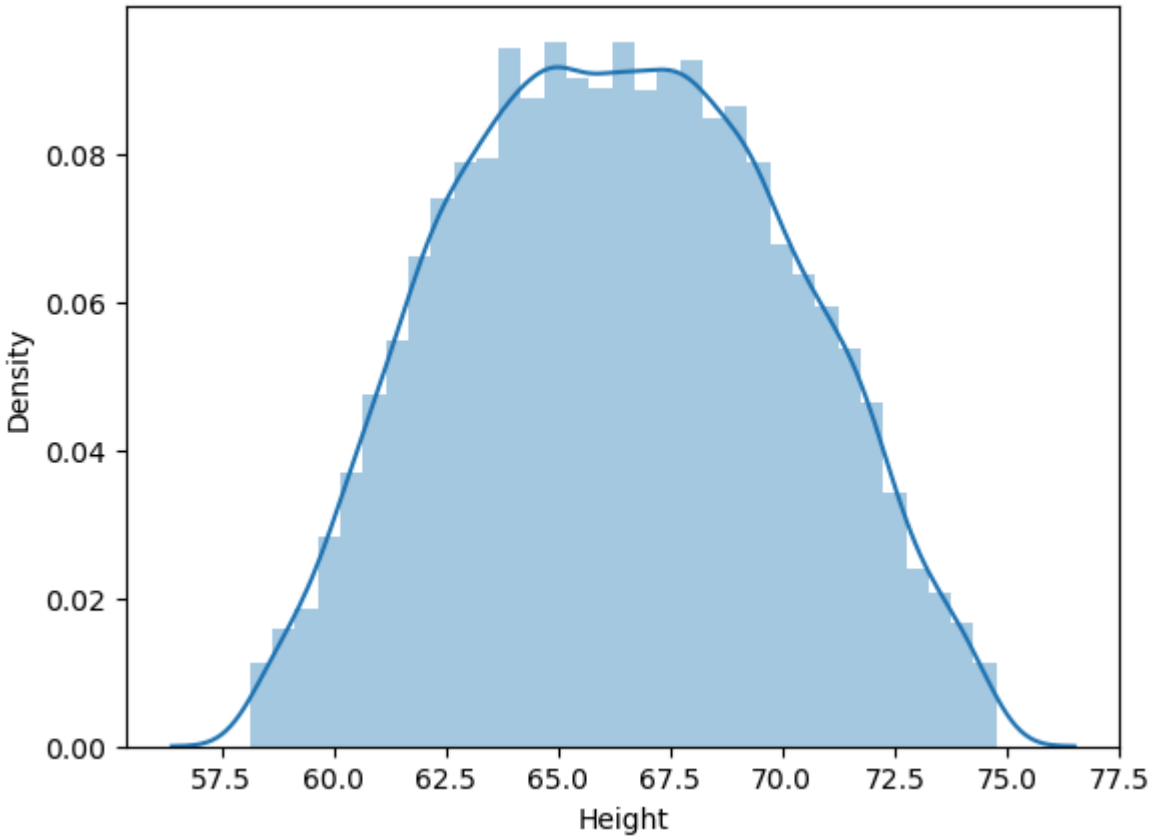
Out[51]:
```
count    10000.000000
mean        66.367560
std          3.847528
min         54.263133
25%         63.505620
50%         66.318070
75%         69.174262
max         78.998742
Name: Height, dtype: float64
```

Plot Draw after Trimming

In [53]:
```
sns.distplot(new_df['Height'])
```

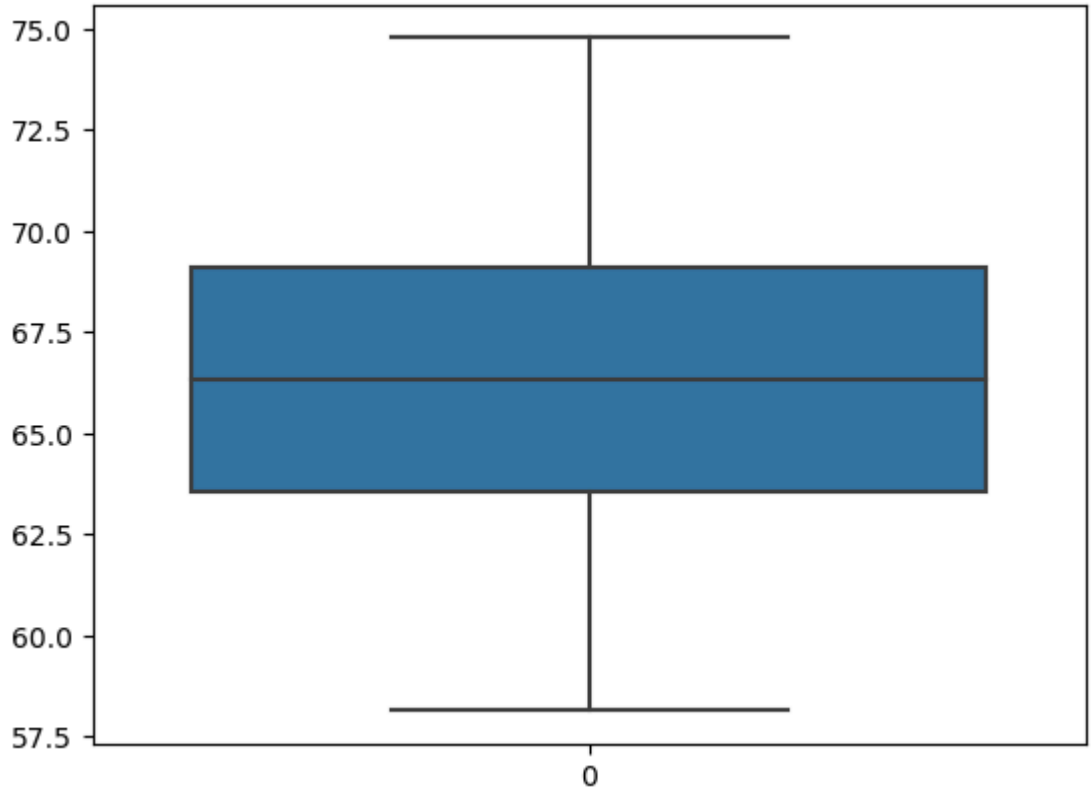
```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_11628\1622920233.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(new_df['Height'])
```

Out[53]:
```
<Axes: xlabel='Height', ylabel='Density'>
```



In [54]:
```
sns.boxplot(new_df['Height'])
```

Out[54]: <Axes: >



Capping with winsorization

In [65]:
```
df['Height'] = np.where(df['Height'] >= upper_limit,
        upper_limit,
        np.where(df['Height'] <= lower_limit,
        lower_limit,
                 df['Height']))
```

In [67]:
```
df.shape
```

Out[67]: (10000, 3)

In [68]:
```
df.head()
```

Out[68]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

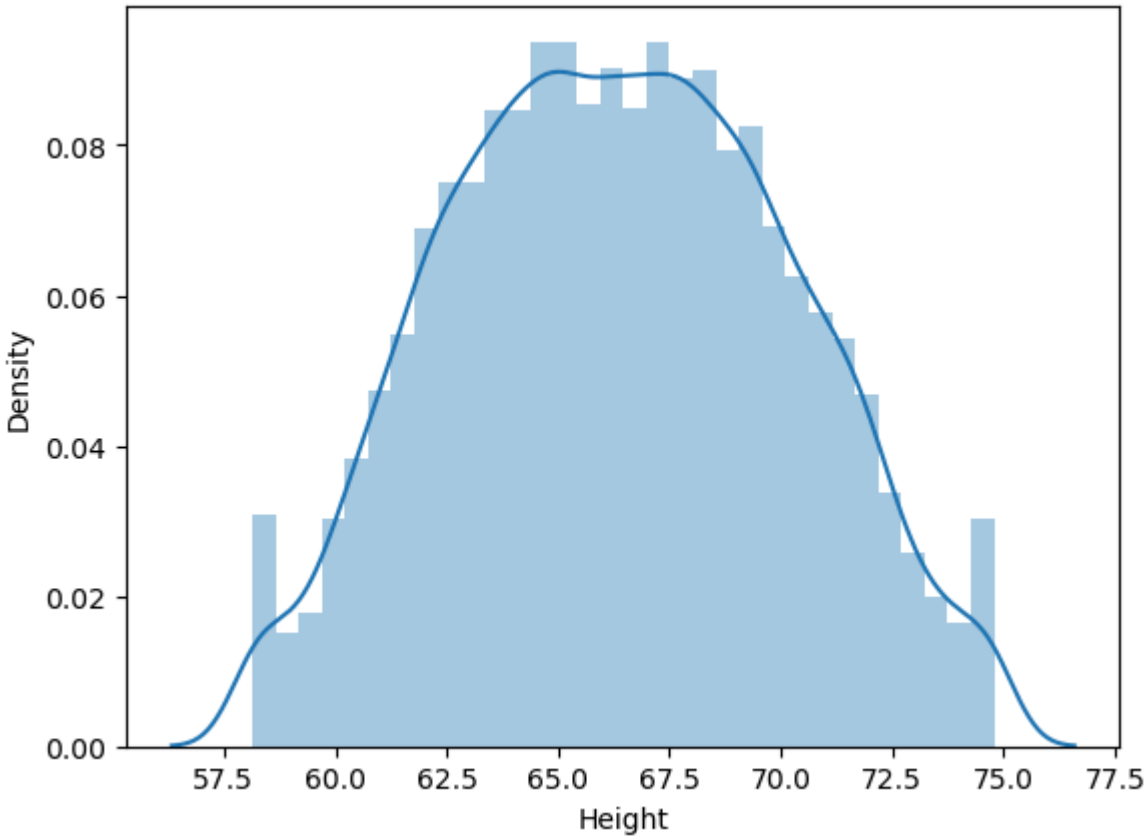
In [69]:
```
df['Height'].describe()
```

Out[69]:
```
count    10000.000000
mean        66.366281
std          3.795717
min         58.134412
25%         63.505620
50%         66.318070
75%         69.174262
max         74.785790
Name: Height, dtype: float64
```

In [70]:
```
sns.distplot(df['Height'])
```

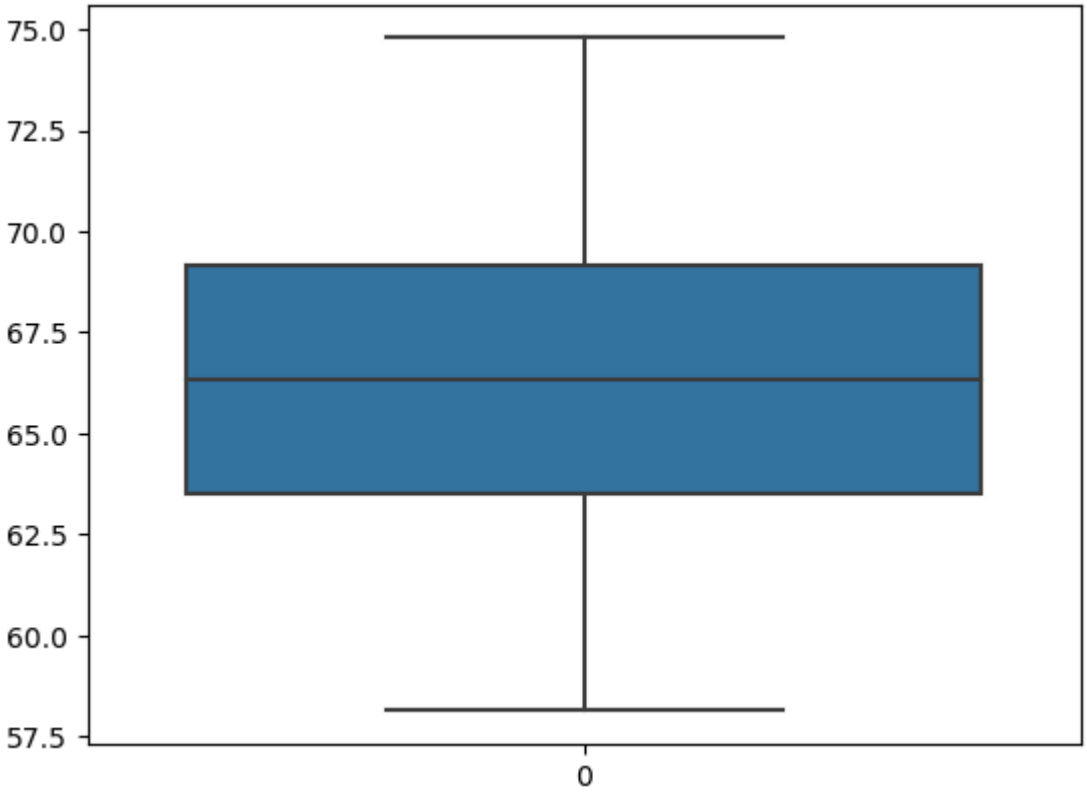
```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_11628\3945773010.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['Height'])
```

Out[70]:
```
<Axes: xlabel='Height', ylabel='Density'>
```



In [71]:
```
sns.boxplot(df['Height'])
```

Out[71]: <Axes: >



In [ ]: