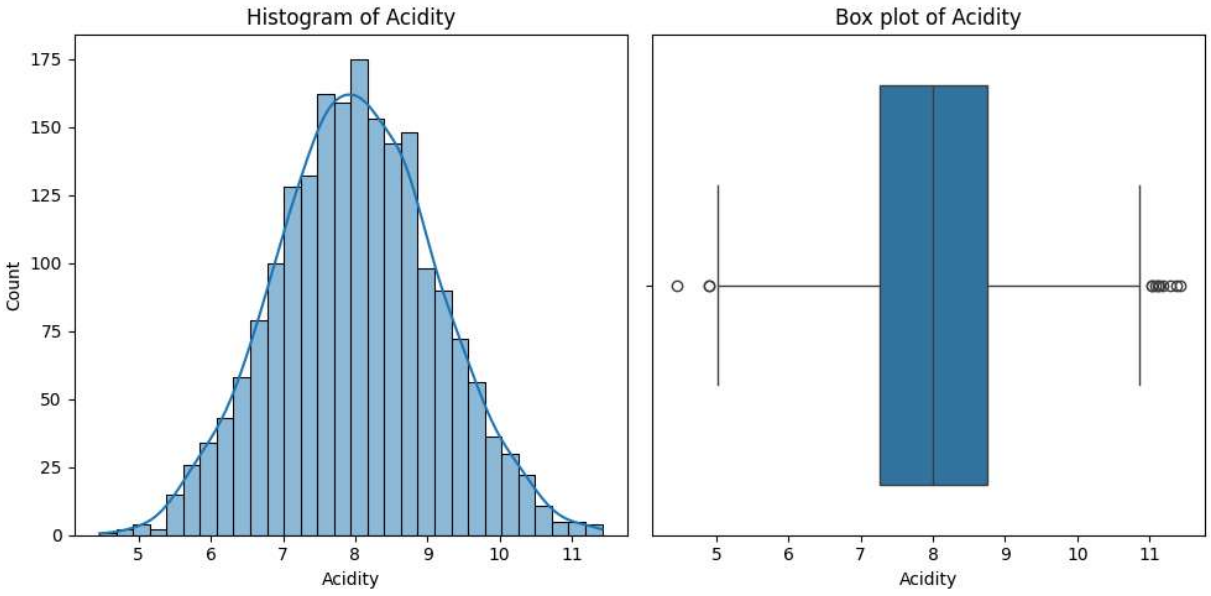
Nomor 3

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import shapiro
```

In [ ]:
```
df = pd.read_csv('banana.csv')
df = df.drop(df.columns[0], axis=1)

# penghapusan data anomali
df = df.drop(df[df['Price'] == 0].index)
df = df.drop(df[df['Price'] == -1].index)
df = df.drop(df[df['Firmness'] == 2.0].index)
df = df.drop(df[df['Country_of_Origin'] == 'undefined'].index)
df_num = df.select_dtypes(include=['number'])
df_str = df.select_dtypes(include=['object'])
```

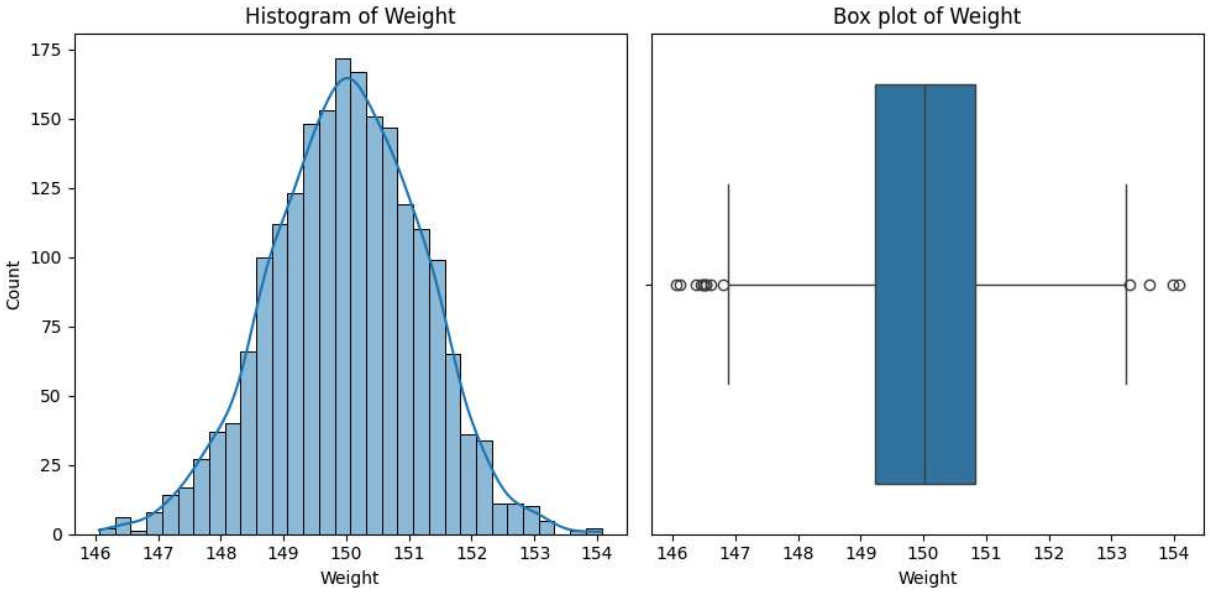
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Acidity'], kde=True)
plt.title(f'Histogram of Acidity')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Acidity'], orient='h')
plt.title(f'Box plot of Acidity')

plt.tight_layout()
plt.show()
```



Histogram `Acidity` memiliki persebaran data yang merata di sekitar titik puncaknya. Dapat dilihat juga bahwa sebagian besar data berkumpul di sekitar nilai tengah. Maka dapat disimpulkan kolom `Acidity` terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 berada di tengah-tengah kotak IQR. Hal ini berarti bahwa persebaran data terdistribusi secara simetris disekitar median. Selain itu, terdapat lebih banyak outlier pada range atas dibandingkan dengan range bawah.

In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Weight'], kde=True)
plt.title(f'Histogram of Weight')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Weight'], orient='h')
plt.title(f'Box plot of Weight')

plt.tight_layout()
plt.show()
```

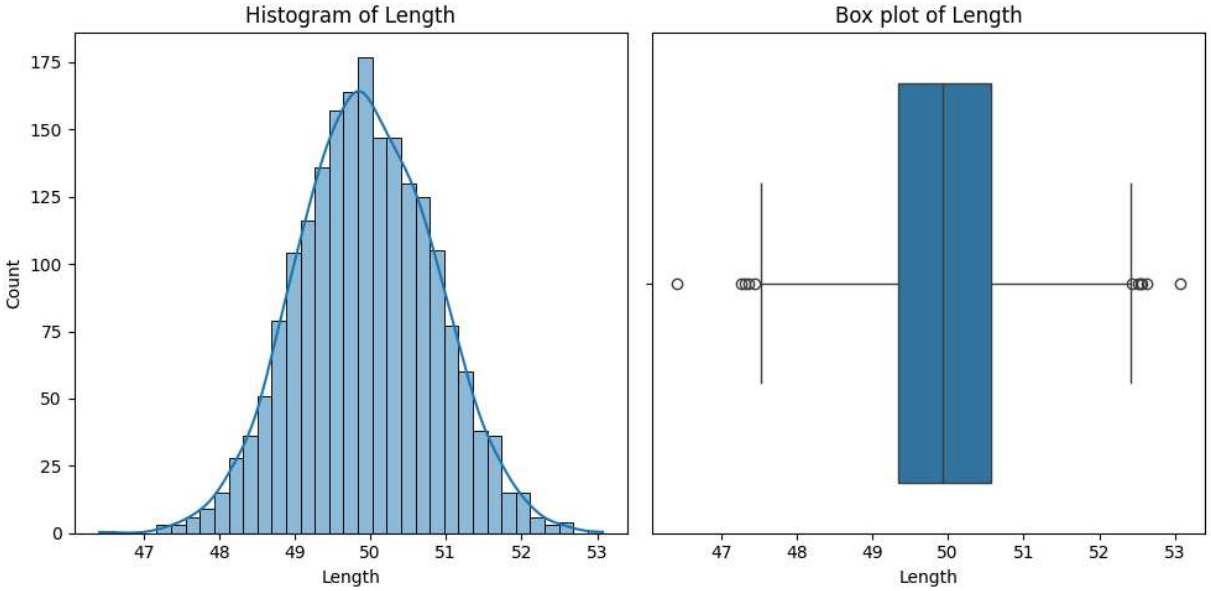


Histogram `Weight` memiliki persebaran data yang merata di sekitar titik puncaknya. Dapat dilihat juga bahwa sebagian besar data berkumpul di sekitar nilai tengah. Maka dapat disimpulkan kolom `Weight` terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 berada di tengah-tengah kotak IQR. Hal ini berarti bahwa persebaran data terdistribusi secara simetris disekitar median. Selain itu, terdapat lebih banyak outlier pada range bawah dibandingkan dengan range atas.

In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Length'], kde=True)
```

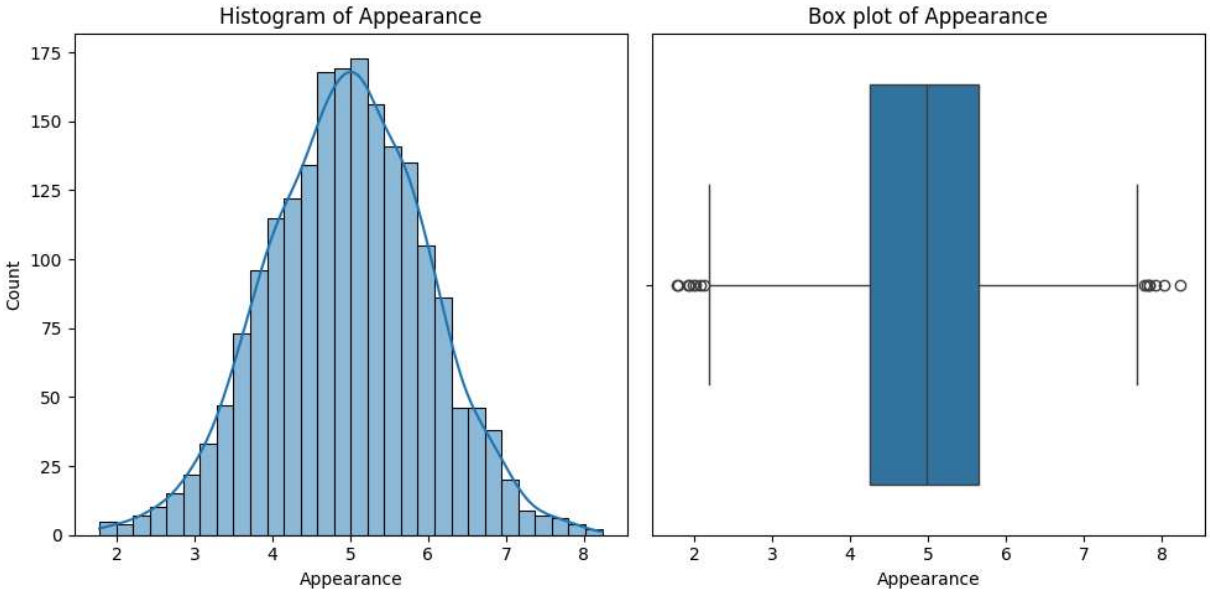
```
plt.title(f'Histogram of Length')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Length'], orient='h')
plt.title(f'Box plot of Length')

plt.tight_layout()
plt.show()
```



Walaupun persebaran data disekitar titik puncak histogram ini terlihat tidak terlalu merata, tetapi bentuk garisnya masih terlihat normal. Dapat dilihat juga bahwa sebagian besar data berkumpul di sekitar nilai tengah. Maka dapat disimpulkan kolom `Length` masih terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 hampir berada di tengah-tengah kotak IQR. Hal ini berarti bahwa persebaran data hampir terdistribusi secara simetris disekitar median.

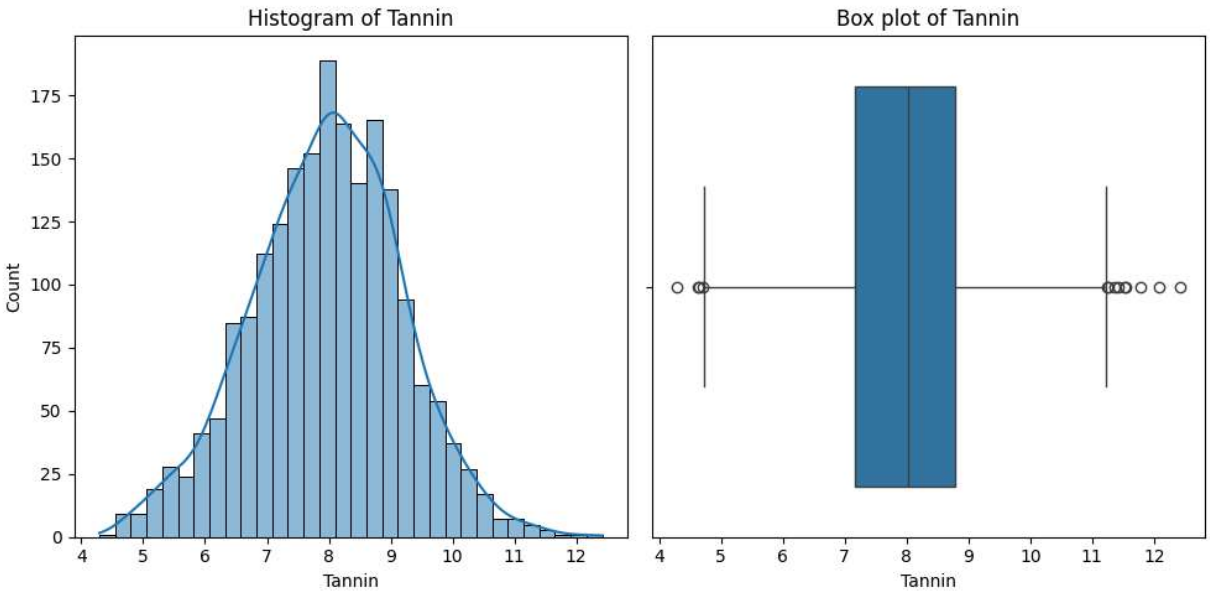
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Appearance'], kde=True)
plt.title(f'Histogram of Appearance')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Appearance'], orient='h')
plt.title(f'Box plot of Appearance')

plt.tight_layout()
plt.show()
```



Walaupun persebaran data disekitar titik puncak histogram ini terlihat tidak terlalu merata, tetapi bentuk garisnya masih terlihat normal. Dapat dilihat juga bahwa sebagian besar data berkumpul di sekitar nilai tengah. Maka dapat disimpulkan kolom `Appearance` masih terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 hampir berada di tengah-tengah kotak IQR. Hal ini berarti bahwa persebaran data hampir terdistribusi secara simetris disekitar median. Selain itu, terdapat lebih banyak outlier pada range bawah dibandingkan dengan range atas.

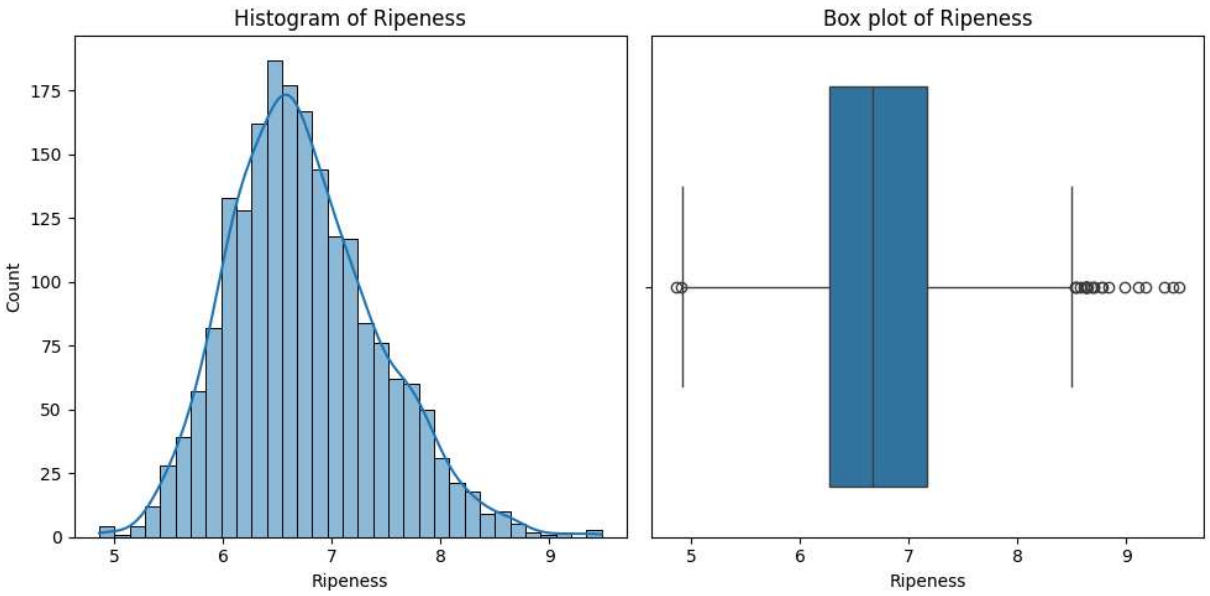
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Tannin'], kde=True)
plt.title(f'Histogram of Tannin')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Tannin'], orient='h')
plt.title(f'Box plot of Tannin')

plt.tight_layout()
plt.show()
```



Pada histogram `Tannin`, dapat dilihat bahwa pesebaran data disekitar titik puncak terlihat tidak merata. Karena itu kolom `Tannin` tidak terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 terlihat melenceng dari tengah. Juga terlihat outlier yang sangat jauh pada range atas.

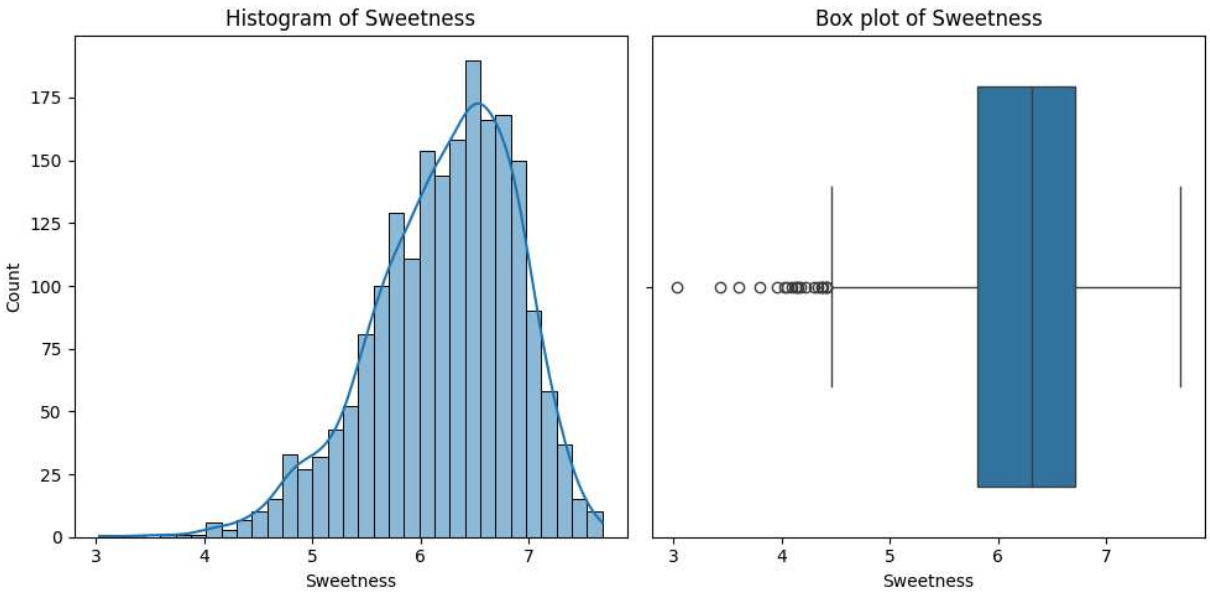
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Ripeness'], kde=True)
plt.title(f'Histogram of Ripeness')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Ripeness'], orient='h')
plt.title(f'Box plot of Ripeness')

plt.tight_layout()
plt.show()
```



Terlihat jelas bahwa distribusi pada histogram lebih condong ke arah kiri. Maka dapat disimpulkan bahwa kolom `Ripeness` memiliki distribusi *Positively Skewed*.

Jenis distribusi ini memiliki karaktreistik garis Q2 yang lebih berada di kiri pada kotak IQR serta outlier yang lebih banyak tersebar di range atas.

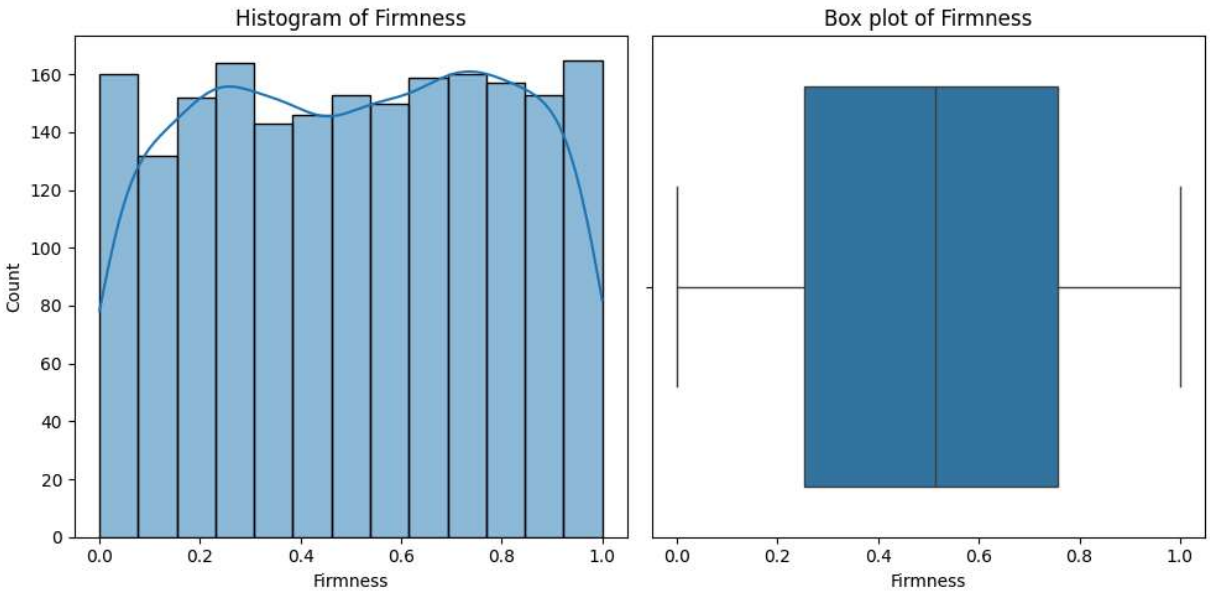
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Sweetness'], kde=True)
plt.title(f'Histogram of Sweetness')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Sweetness'], orient='h')
plt.title(f'Box plot of Sweetness')

plt.tight_layout()
plt.show()
```



Terlihat jelas bahwa distribusi pada histogram lebih condong ke arah kanan. Maka dapat disimpulkan bahwa kolom `Sweetness` memiliki distribusi *Negatively Skewed*.

Jenis distribusi ini memiliki karaktreistik garis Q2 yang lebih berada di kanan pada kotak IQR serta outlier yang lebih banyak tersebar di range bawah.

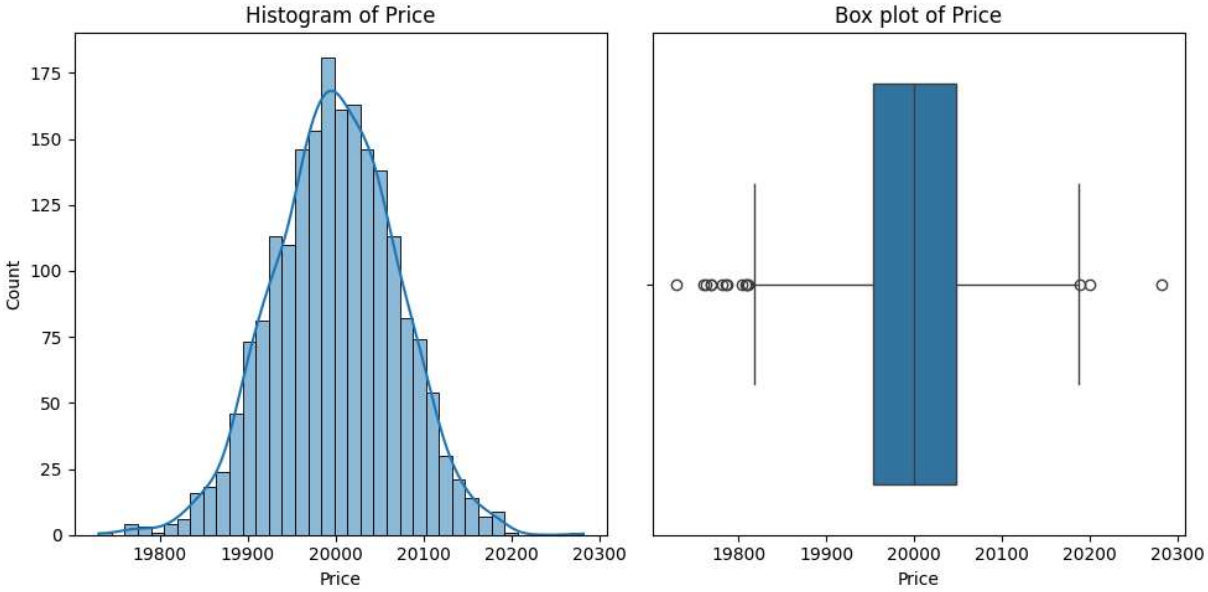
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Firmness'], kde=True)
plt.title(f'Histogram of Firmness')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Firmness'], orient='h')
plt.title(f'Box plot of Firmness')

plt.tight_layout()
plt.show()
```



Pada histogram ini terlihat bahwa jumlah kemunculan setiap nilai `Firmness` hampir sama. Maka dapat dibilang jenis distribusi dari kolom ini adalah *Uniform Distribution*.

Karakteristik dari *Uniform Distribution* adalah garis Q2 yang berada di tengah-tengah kotak IQR serta tidak adanya outlier dalam persebaran data.

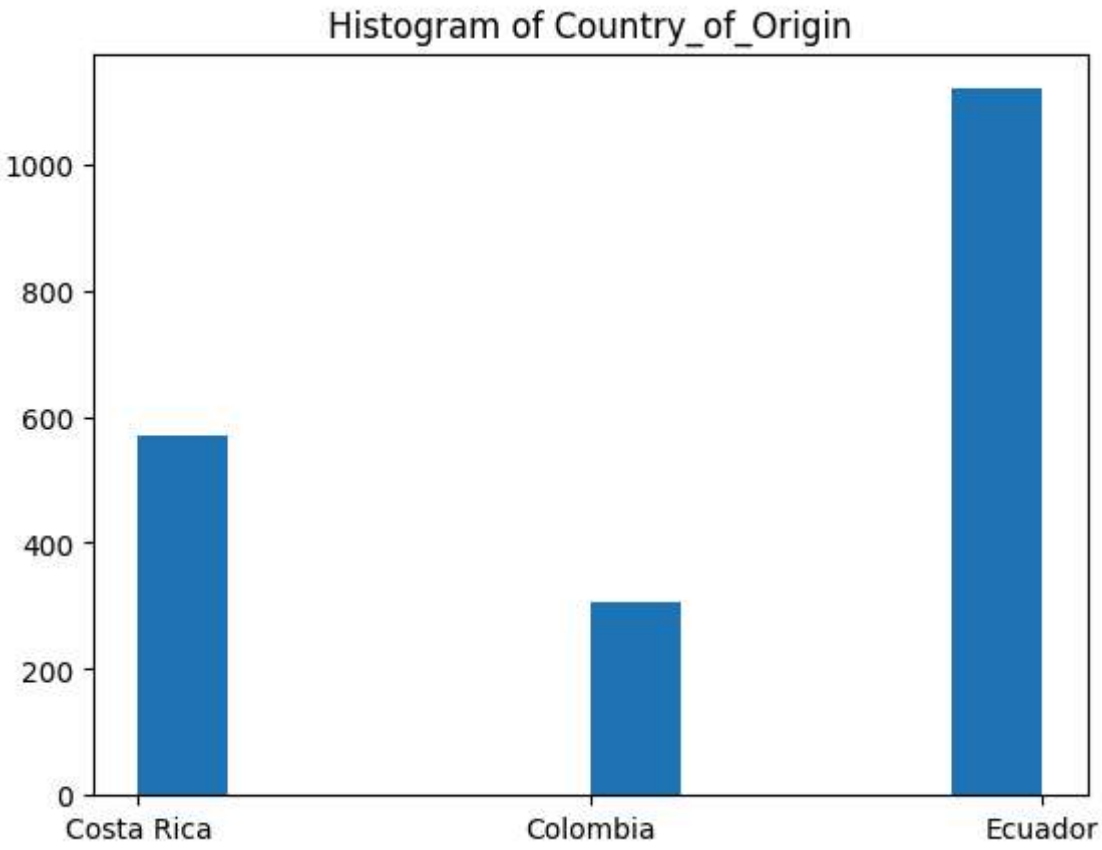
In [ ]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.histplot(df_num['Price'], kde=True)
plt.title(f'Histogram of Price')

plt.subplot(1, 2, 2)
sns.boxplot(x=df_num['Price'], orient='h')
plt.title(f'Box plot of Price')

plt.tight_layout()
plt.show()
```



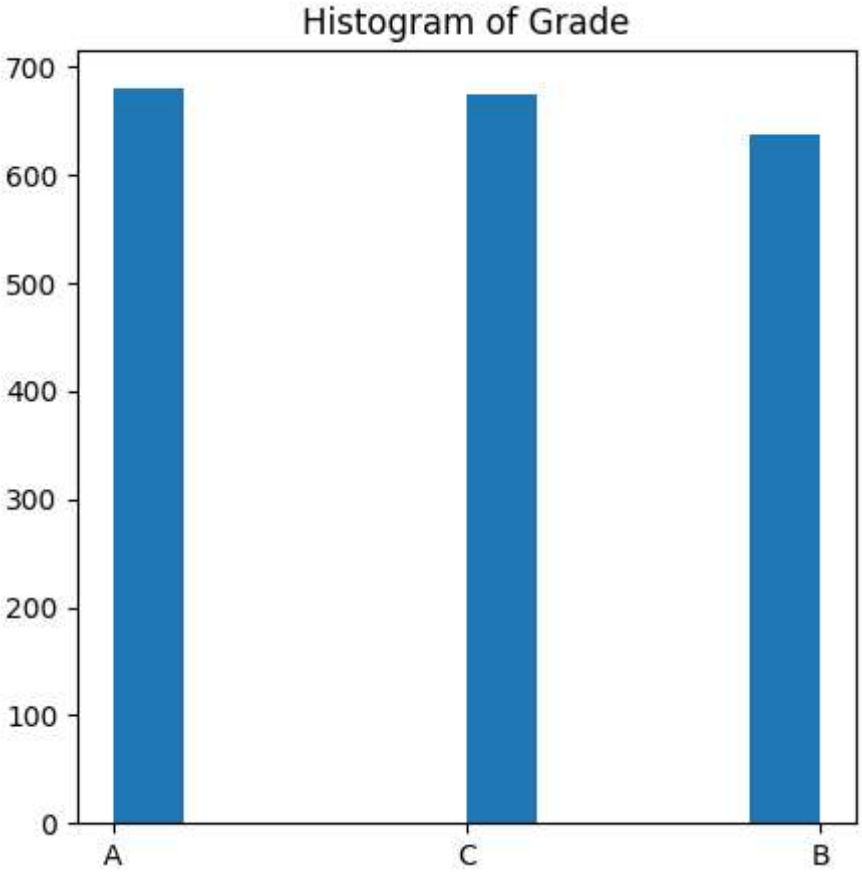
Histogram `Price` memiliki persebaran data yang merata di sekitar titik puncaknya. Dapat dilihat juga bahwa sebagian besar data berkumpul di sekitar nilai tengah. Maka dapat disimpulkan kolom `Price` terdistribusi normal.

Sedangkan pada boxplot, dapat dilihat garis Q2 berada di tengah-tengah kotak IQR. Hal ini berarti bahwa persebaran data terdistribusi secara simetris disekitar median. Selain itu, terdapat lebih banyak outlier pada range bawah dibandingkan dengan range atas.

In [ ]:
```
plt.hist(df_str['Country_of_Origin'], bins=10)
plt.title(f'Histogram of Country_of_Origin')

plt.show()
```



In [ ]:
```
plt.figure(figsize=(5, 5))

plt.hist(df_str['Grade'], bins=10)
plt.title(f'Histogram of Grade')

plt.show()
```


Histogram of Grade
700
600
500
400
300
200
100
0
A
C
B