Nomor 4

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) # ini ngapus id karna id dari pandas udah ada j
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'])
```

Salah satu ciri grafik histogram yang terdistribusi normal adalah memiliki bentuk yang sama dengan **Kurva Gauss**. Jika dilihat hasil visualisasi pada nomor sebelumnya, yang bentuknya mirip dengan Kurva Gauss adalah kolom `Acidity` , `Weight` , `Length` , `Appearance` , dan `Price` . Karena itu, dapat disimpulkan kelima kolom tersebut terdistribusi normal.

Sedangkan kolom `Tanin` , `Ripeness` , `Sweetness` , dan `Firmness` tidak terdistribusi normal karena bentuknya yang berbeda dari **Kurva Gauss**.

Dapat dilihat pada visualisasi kolom `Tanin` , terdapat puncak yang tidak berada ditengah serta beberapa puncak yang tidak tersebar merata di kiri dan kanannya.

Kemudian untuk kolom `Ripeness` dan `Sweetness` , kedua kolom itu memiliki jenis *Skewed Distribution*. Maksudnya adalah distribusi yang condong ke salah satu sisi. Seperti yang dapat dilihat, `Ripeness` lebih condong ke arah kiri atau *Positively Skewed* dan `Sweetness` lebih condnong ke arah kanan atau *Negatively Skewed*.

Terakhir untuk kolom `Firmness` , jenis distribusinya termasuk ke dalam distribusi *Uniform* karena setiap nilai memiliki kemungkinan yang hampir sama untuk muncul. Hal ini dapat dilihat dari histogram yang menunjukkan bahwa jumlah data hampir sama, dengan sedikit variasi yang terlihat.

Analisis ini kami validasi kembali dengan menggunakan `Shapiro-Wilk Test` dengan kode dibawah.

In [ ]:
```
for column in df_num.columns:
  stat, p = shapiro(df_num[column])
  alpha = 0.05
  # print("nilai p nya",p)
  if p > alpha:
```

```
    print(f'{column} is normally distributed according to Shapiro-Wilk Test')
  else:
    print(f'{column} is NOT normally distributed according to Shapiro-Wilk Test')
```

```
Acidity is normally distributed according to Shapiro-Wilk Test
Weight is normally distributed according to Shapiro-Wilk Test
Length is normally distributed according to Shapiro-Wilk Test
Appearance is normally distributed according to Shapiro-Wilk Test
Tannin is NOT normally distributed according to Shapiro-Wilk Test
Ripeness is NOT normally distributed according to Shapiro-Wilk Test
Sweetness is NOT normally distributed according to Shapiro-Wilk Test
Firmness is NOT normally distributed according to Shapiro-Wilk Test
Price is normally distributed according to Shapiro-Wilk Test
```