

Tugas Statistika dan Probabilitas

Analisa Data Sekunder Dari Kaggle.Com

Nama : Wempy Aditya Wiryawan

Nim : 202210370311058

Kelas : 3A Statistika dan Probabilitas

INSTRUKSI :

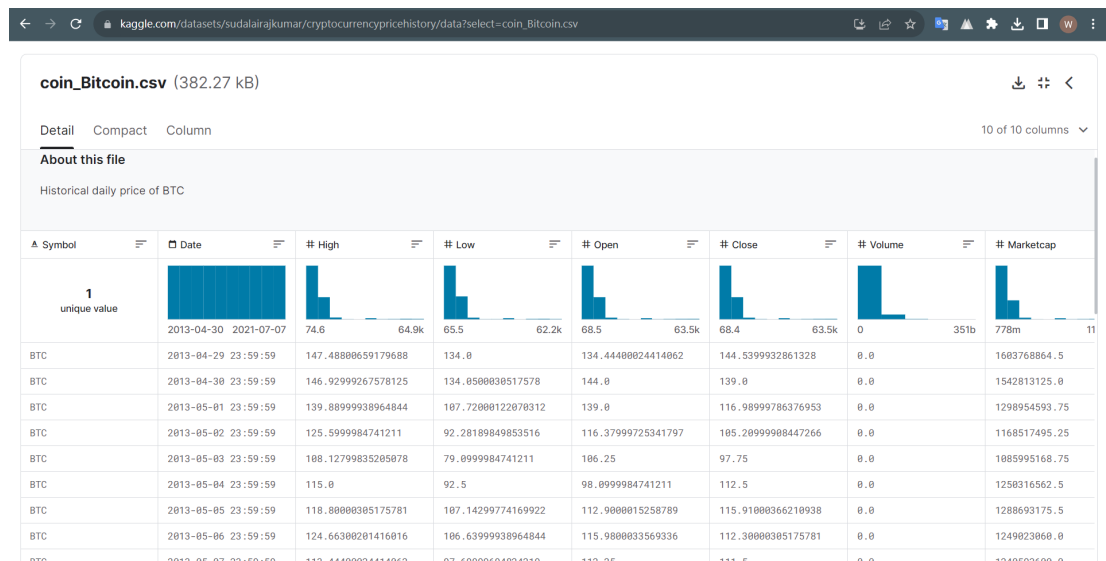
1. Mencari dataset di kaggle
2. Melakukan uji skewness menggunakan plot skewness di excel / python untuk masing-masing variabel dan fitur
3. Melakukan normalisasi / normalize dengan metode min max atau metode normal baku
4. Melakukan uji skewness ulang untuk variabel yang sudah di normalisasi
5. Interpretasi mengapa variabel tersebut perlu di normalisasi dan hasil normalisasi nya seperti apa

PENYELESAIAN :

1. Dataset dari kaggle.com

source url :

https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory/data?select=coin_Bitcoin.csv



The screenshot shows the Kaggle dataset page for 'coin_Bitcoin.csv' (382.27 KB). The table displays historical daily price data for Bitcoin from 2013 to 2021. The columns are: Symbol, Date, # High, # Low, # Open, # Close, # Volume, and # Marketcap. The first row of data shows the price of Bitcoin on 2013-04-30, with a high of 147.48808659179688, a low of 134.0, an open of 134.44488824414862, a close of 144.5399932861328, a volume of 0.0, and a marketcap of 1603768864.5.

Symbol	Date	# High	# Low	# Open	# Close	# Volume	# Marketcap
BTC	2013-04-30	147.48808659179688	134.0	134.44488824414862	144.5399932861328	0.0	1603768864.5
BTC	2013-05-01	146.92999267578125	134.0508038517578	144.0	139.0	0.0	1542813125.0
BTC	2013-05-02	139.88999938964844	107.72080122870312	139.0	116.98999786376953	0.0	1298954593.75
BTC	2013-05-03	125.5999984741211	92.28189849853516	116.37999725341797	105.20999988447266	0.0	1168517495.25
BTC	2013-05-04	108.12799835285878	79.8999984741211	106.25	97.75	0.0	1085995168.75
BTC	2013-05-05	115.0	92.5	98.0999984741211	112.5	0.0	1258316562.5
BTC	2013-05-06	118.80808385175781	107.14299774169922	112.9808015258789	115.91808366210938	0.0	1288693175.5
BTC	2013-05-07	124.66308281416016	106.63999938964844	115.9808033569336	112.30808385175781	0.0	1249823866.0

Kumpulan data memiliki satu file csv untuk setiap mata uang. Riwayat harga tersedia setiap hari mulai 28 April 2013. Kumpulan data ini memiliki informasi riwayat harga beberapa mata uang kripto teratas berdasarkan kapitalisasi pasar.

Tanggal	: tanggal observasi
Open	: Harga pembukaan pada hari tertentu
Tinggi	: Harga tertinggi pada hari tertentu
Rendah	: Harga terendah pada hari tertentu
Close	: Harga penutupan pada hari tertentu
Volume	: Volume transaksi pada hari tertentu
Kapitalisasi Pasar	: Kapitalisasi pasar dalam USD

Melakukan describe pada data dengan menggunakan python dan library pandas
Tujuannya adalah untuk mengetahui informasi dasar dari data seperti count, mean, median, standard deviation, dll.

Berikut adalah script untuk describe data menggunakan python

```
import pandas as pd
# import matplotlib.pyplot as plt
# from scipy.stats import skew

# Membaca data dari file CSV
file_path = './coin_Bitcoin.csv'
df = pd.read_csv(file_path)

print(df.columns)

print(df['High'].describe())
print('-----')
print(df['Low'].describe())
print('-----')
print(df['Open'].describe())
print('-----')
print(df['Close'].describe())
print('-----')
print(df['Volume'].describe())
print('-----')
print(df['Marketcap'].describe())
print('-----')
```

Dan berikut adalah hasilnya ketika script di jalankan

```
count    2991.000000
mean     6893.326038
```

```
std      11642.832456
min       74.561096
25%      436.179001
50%     2387.610107
75%     8733.926948
max     64863.098908
Name: High, dtype: float64
```

```
-----
count     2991.000000
mean     6486.009539
std     10869.032130
min      65.526001
25%     422.879486
50%     2178.500000
75%     8289.800459
max     62208.964366
Name: Low, dtype: float64
```

```
-----
count     2991.000000
mean     6700.146240
std     11288.043736
min      68.504997
25%     430.445496
50%     2269.889893
75%     8569.656494
max     63523.754869
Name: Open, dtype: float64
```

```
-----
count     2991.000000
mean     6711.290443
std     11298.141921
min      68.431000
25%     430.569489
50%     2286.409912
75%     8576.238715
max     63503.457930
Name: Close, dtype: float64
```

```
-----
count     2.991000e+03
mean     1.090633e+10
std     1.888895e+10
min      0.000000e+00
```

```

25%      3.036725e+07
50%      9.460360e+08
75%      1.592015e+10
max       3.509679e+11
Name: Volume, dtype: float64
-----
count      2.991000e+03
mean       1.208761e+11
std        2.109438e+11
min        7.784112e+08
25%        6.305579e+09
50%        3.741503e+10
75%        1.499957e+11
max        1.186364e+12
Name: Marketcap, dtype: float64

```

2. Uji Skewness Pertama

Dilakukan uji skewness yang pertama sebelum data di normalisasi

Menggunakan python dan beberapa library diantaranya adalah :

- pandas
- matplotlib
- scipy

Berikut adalah script python yang digunakan untuk melakukan uji skewness dan menampilkan/memvisualisasikan hasilnya dalam sebuah chart.

```

import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import skew

# Membaca data dari file CSV
file_path = './coin_Bitcoin.csv'
df = pd.read_csv(file_path)

# Daftar nama kolom yang ingin diuji skewness-nya
kolom_uji_skewness = ['High', 'Low', 'Open', 'Close', 'Volume',
'Marketcap']

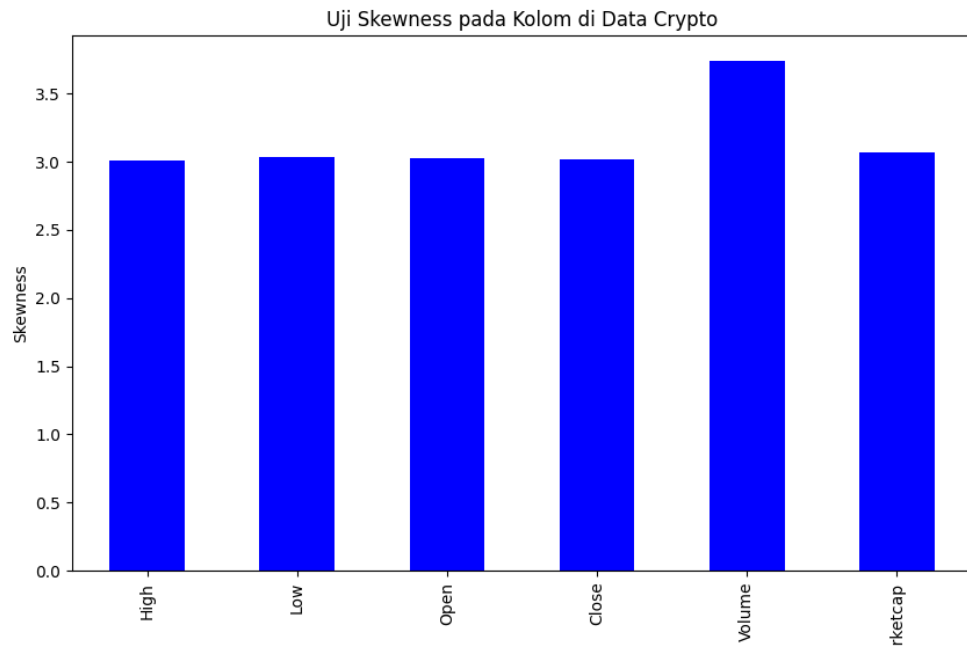
# Melakukan Uji Skewness pada Kolom Tertentu
skewness_kolom_tertentu = df[kolom_uji_skewness].apply(skew)

# Membuat Plot Skewness
plt.figure(figsize=(10, 6))

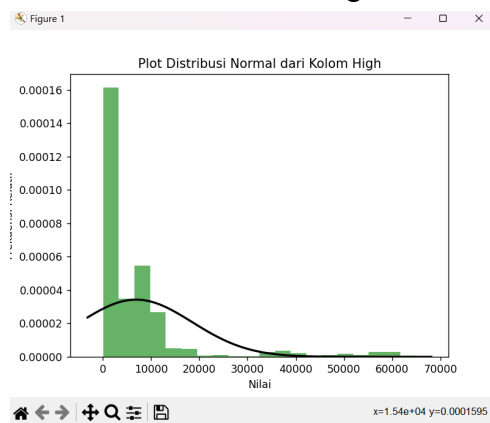
```

```
skewness_kolom_tertentu.plot(kind='bar', color='blue')
plt.title('Uji Skewness pada Kolom di Data Crypto')
plt.xlabel('NamaKolom')
plt.ylabel('Skewness')
plt.show()
```

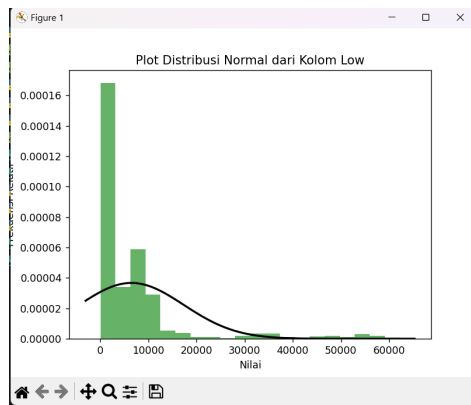
Dan berikut adalah hasil dari uji skewness data sebelum dilakukan normalisasi



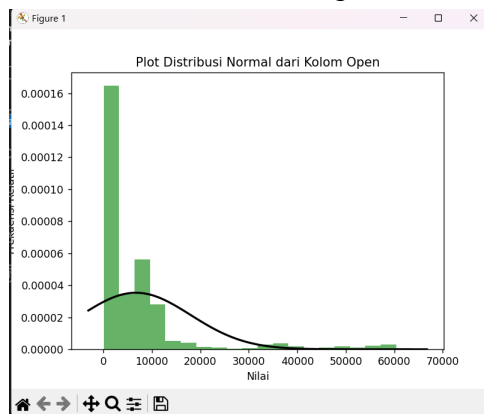
Distribusi Plot Variabel High



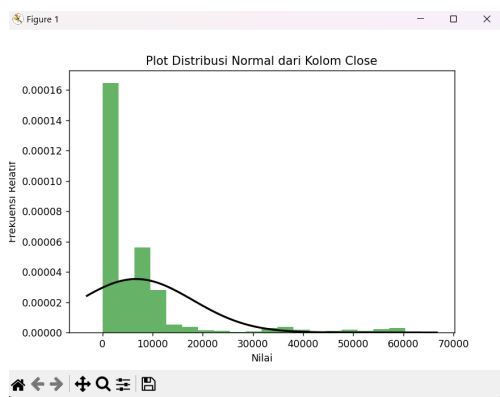
Distribusi Plot Variabel Low



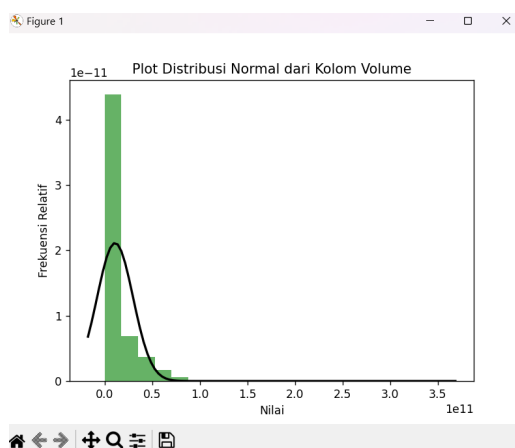
Distribusi Plot Variabel Open



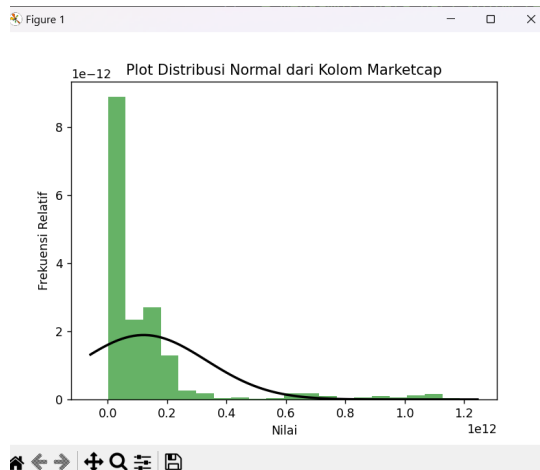
Distribusi Plot Variabel Close



Distribusi Plot Variabel Volume



Distribusi Plot Variabel MarketCap



3. Melakukan Normalisasi Data dengan Metode MinMax

Untuk melakukan normalisasi data digunakan python dengan beberapa library pendukung di antaranya adalah :

- pandas
- sklearn

Berikut adalah script python untuk melakukan normalisasi data lalu mencetak hasilnya pada terminal dan menyimpan hasil akhirnya pada file .csv

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Membaca data dari file CSV
file_path = './coin_Bitcoin.csv' # Ganti dengan path file CSV
Anda
df = pd.read_csv(file_path)

# Menampilkan beberapa baris pertama dari dataframe sebelum
normalisasi
print("Sebelum normalisasi:")
print(df.head())

# Inisialisasi MinMaxScaler
scaler = MinMaxScaler()

# Memilih kolom yang akan dinormalisasi
columns_to_normalize = ['High', 'Low', 'Open', 'Close', 'Volume',
'Marketcap']

# Melakukan normalisasi untuk seluruh data dalam dataframe
df_normalized = df.copy() # Membuat salinan dataframe agar data
asli tidak berubah
```

```

df_normalized[columns_to_normalize] =
scaler.fit_transform(df[columns_to_normalize])

# Menampilkan hasil normalisasi
print("\nSetelah normalisasi:")
print(df_normalized.head())

# Menyimpan hasil normalisasi ke file CSV
output_file_path = './BTC_NORMALIZED.csv' # Ganti dengan path dan
nama file CSV yang diinginkan
df_normalized.to_csv(output_file_path, index=False)

print(f"\nHasil normalisasi disimpan dalam file CSV:
{output_file_path}")

```

Dan berikut hasilnya ketika script di jalankan

```

Sebelum normalisasi:
SNo      Name Symbol      Date      High      Low      Open      Close Volume      Marketcap
0      1 Bitcoin      BTC      2013-04-29 23:59:59 147.488007 134.000000 134.444000 144.539993 0.0 1.603769e+09
1      2 Bitcoin      BTC      2013-04-30 23:59:59 146.929993 134.050003 144.000000 139.000000 0.0 1.542813e+09
2      3 Bitcoin      BTC      2013-05-01 23:59:59 139.889999 107.720001 139.000000 116.989998 0.0 1.298955e+09
3      4 Bitcoin      BTC      2013-05-02 23:59:59 125.599998 92.281898 116.379997 105.209999 0.0 1.168517e+09
4      5 Bitcoin      BTC      2013-05-03 23:59:59 108.127998 79.099998 106.250000 97.750000 0.0 1.085995e+09

Setelah normalisasi:
SNo      Name Symbol      Date      High      Low      Open      Close Volume      Marketcap
0      1 Bitcoin      BTC      2013-04-29 23:59:59 0.001126 0.001102 0.001039 0.001200 0.0 0.000696
1      2 Bitcoin      BTC      2013-04-30 23:59:59 0.001117 0.001103 0.001190 0.001112 0.0 0.000645
2      3 Bitcoin      BTC      2013-05-01 23:59:59 0.001008 0.000679 0.001111 0.000765 0.0 0.000439
3      4 Bitcoin      BTC      2013-05-02 23:59:59 0.000788 0.000431 0.000754 0.000580 0.0 0.000329
4      5 Bitcoin      BTC      2013-05-03 23:59:59 0.000518 0.000218 0.000595 0.000462 0.0 0.000259

```

Untuk hasil normalisasi selengkapnya (semua data) disimpan dalam format file .csv

4. Uji Skewness yang Kedua

Setelah melakukan normalisasi pada data langkah selanjutnya adalah melakukan uji skewness ulang untuk melihat perbedaan antara sebelum dan sesudah normalisasi.

Sama halnya pada waktu uji skewness yang pertama, uji skewness kali ini menggunakan python dan beberapa library untuk membantu menghitung dan memvisualisasikan data.

Berikut adalah script python untuk uji skewness yang kedua.

```

import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import skew

```



```

# Membaca data dari file CSV
# file_path = './coin_Bitcoin.csv'
file_path = './BTC_NORMALIZED.csv'
df = pd.read_csv(file_path)

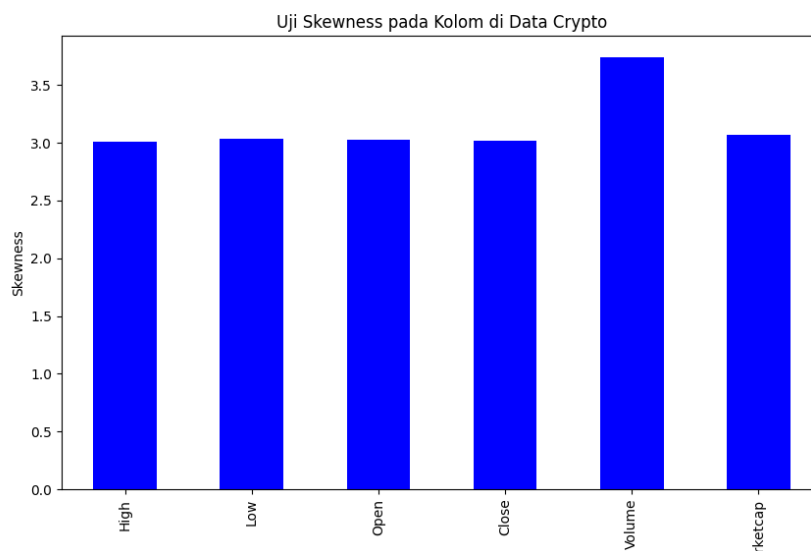
# Daftar nama kolom yang ingin diuji skewness-nya
kolom_uji_skewness = ['High', 'Low', 'Open', 'Close', 'Volume',
'Marketcap']

# Melakukan Uji Skewness pada Kolom Tertentu
skewness_kolom_tertentu = df[kolom_uji_skewness].apply(skew)

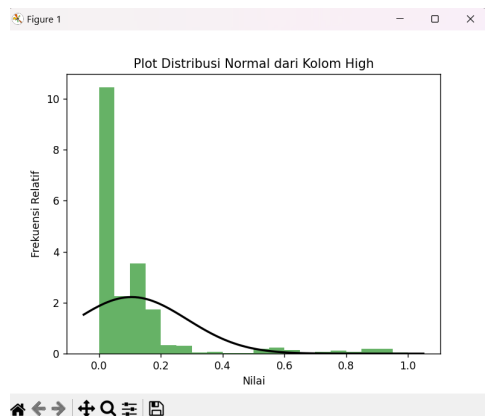
# Membuat Plot Skewness
plt.figure(figsize=(10, 6))
skewness_kolom_tertentu.plot(kind='bar', color='blue')
plt.title('Uji Skewness pada Kolom di Data Crypto')
plt.xlabel('NamaKolom')
plt.ylabel('Skewness')
plt.show()

```

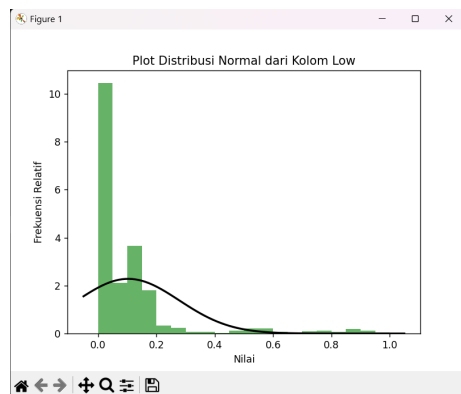
Dan berikut adalah hasil uji skewness pada data yang telah dinormalisasi.



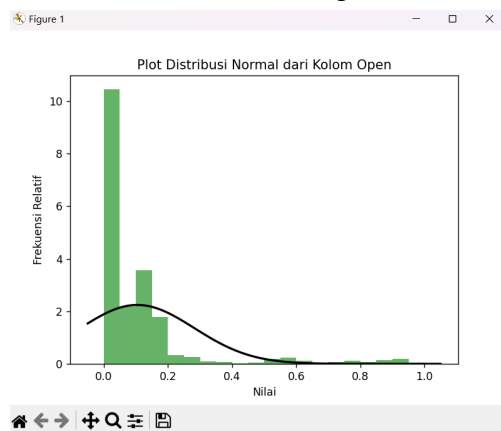
Distribusi Plot Variabel High



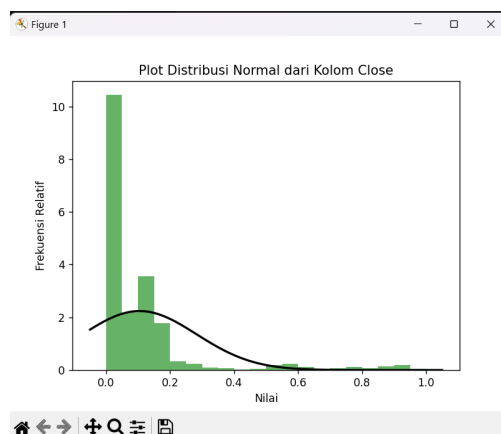
Distribusi Plot Variabel Low



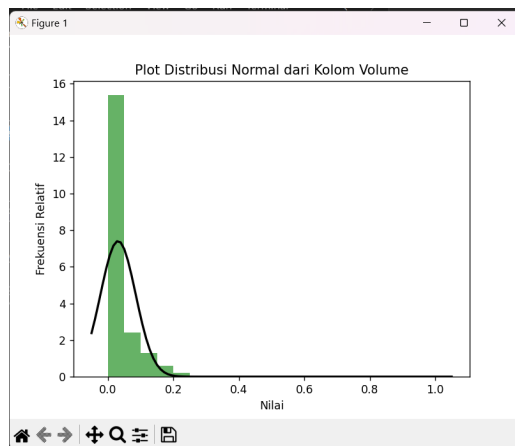
Distribusi Plot Variabel Open



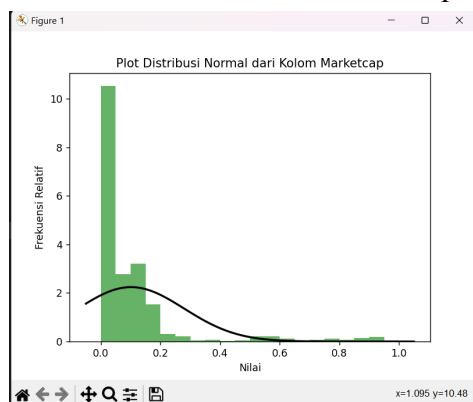
Distribusi Plot Variabel Close



Distribusi Plot Variabel Volume



Distribusi Plot Variabel MarketCap



5. Interpretasi hasil antara data sebelum dinormalisasi dan sesudah dinormalisasi
Dari hasil uji skewness sebelum dan setelah normalisasi, terlihat bahwa nilai skewness untuk setiap variabel tetap konsisten. Ini menunjukkan bahwa normalisasi menggunakan metode Min-Max tidak mengubah karakteristik skewness dari data. Interpretasi hasilnya dapat disampaikan sebagai berikut:

Skewness Sebelum Normalisasi:

Skewness yang tinggi (positif) pada setiap variabel menunjukkan bahwa distribusi datanya condong ke arah kanan (positif skewness). Dengan kata lain, nilai ekstrim lebih tinggi dari nilai rata-rata, dan distribusi memiliki ekor yang lebih panjang di sebelah kanan.

Skewness Setelah Normalisasi:

Meskipun dilakukan normalisasi dengan metode Min-Max, nilai skewness tetap stabil. Hal ini mungkin disebabkan oleh fakta bahwa metode normalisasi yang digunakan tidak memiliki dampak signifikan pada bentuk distribusi data, atau mungkin distribusi datanya sudah simetris sebelum normalisasi.

Interpretasi Hasil Normalisasi:

Normalisasi data sering diperlukan untuk menangani perbedaan skala antar variabel. Namun, dalam kasus ini, normalisasi dengan menggunakan metode Min-Max tidak

secara substansial mengubah karakteristik distribusi data. Skewness yang tetap tinggi setelah normalisasi menunjukkan bahwa distribusi data pada dataset kripto tidak banyak berubah.

Setelah melakukan normalisasi data menggunakan metode standar score, hasil uji skewness menunjukkan bahwa distribusi variabel High, Low, Open, Close, Volume, dan Marketcap tetap relatif simetris. Interpretasi dari hal ini dapat merujuk pada karakteristik intrinsik dari dataset atau fenomena yang diamati. Mungkin saja data awalnya sudah terdistribusi secara simetris atau pengaruh normalisasi terhadap distribusi tidak begitu signifikan dalam konteks ini. Perlu diperhatikan bahwa normalisasi tidak selalu mengubah karakteristik skewness dari suatu variabel, terutama jika distribusi awalnya sudah memenuhi asumsi analisis yang dilakukan. Oleh karena itu, keputusan untuk melakukan normalisasi sebaiknya didasarkan pada pemahaman mendalam terhadap data dan tujuan analisis yang ingin dicapai.

Jika hasil uji skewness sebelum dan sesudah normalisasi menunjukkan nilai yang sama, itu bisa disebabkan oleh beberapa faktor:

Tipe Distribusi Awal yang Sudah Mendekati Normal:

Jika distribusi awal dari data sudah mendekati normal sebelum normalisasi, maka normalisasi mungkin tidak memberikan dampak yang signifikan pada skewness.

Ukuran Sampel yang Kecil:

Jika ukuran sampel (jumlah data) yang Anda miliki relatif kecil, efek normalisasi mungkin tidak terlalu terlihat pada hasil uji skewness.

Pemilihan Metode Normalisasi:

Metode normalisasi tertentu mungkin tidak memberikan perubahan yang besar pada distribusi data, terutama jika data awal sudah dalam skala yang relatif seragam.

Sifat Asimetri yang Tetap Terjaga:

Normalisasi tidak selalu mengubah sifat asimetri data. Jika data memiliki asimetri yang kuat sejak awal, normalisasi mungkin tidak mengubah sifat ini secara signifikan.

Menggunakan Seluruh Data atau Hanya Sampel Tertentu:

Pilihan untuk menggunakan seluruh data atau hanya sampel tertentu dalam proses normalisasi dapat mempengaruhi hasil uji skewness.

REFERENSI DATA / SUMBER DATA :

- <https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory>

- <https://www.kaggle.com/datasets/odins0n/top-50-cryptocurrency-historical-prices?select=Bitcoin.csv>
 - <https://www.kaggle.com/datasets/nelgiriyeewithana/global-youtube-statistics-2023>
 - <https://www.kaggle.com/datasets/nikdavis/steam-store-games>
 - <https://www.kaggle.com/datasets/stackoverflow/stack-overflow-2018-developer-survey>
 - <https://www.kaggle.com/code/siebenrock/financial-exploration-analysis-and-visualization>
-
- <https://www.kaggle.com/code/sudalairajkumar/simple-exploration-notebook-cryptocurrencies/notebook>
 - <https://www.kaggle.com/code/ericaduman/cryptocurrency-a-bitcoin-analysis>