Laporan Tugas Kecil 2 IF3170

Exploratory Data Analysis

Disusun oleh:

Kelompok 49

Muhammad Hanan	13521041
Vieri Fajar Firdaus	13521099

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika - Institut Teknologi Bandung

Jl. Ganesha 10, Bandung 40132

2023

Jawaban

1. Statistik dasar (mean, min, max, dll)

   Untuk mencari statistik dasar dari file data_train.csv, kita akan memisahkan setiap kolom yang ada sesuai dengan jenisnya. Jenis ini terdiri dari nominal, ordinal dan kuantitatif. Setiap jenis ini memiliki statistik data yang berbeda beda. Berikut merupakan statik dasar setiap golongan dan penjelasannya.

   a. Nominal

      Untuk jenis nominal, kolom yang tergolong dalam jenis ini adalah kolom blue, dual_sim, four_g, tree_g, touch_screen, wifi. Kemudian statistik dasar yang akan dicari yaitu modus, frekuensi, dan presentase. Hasilnya adalah seperti berikut.

	Modus	Frekuensi 0	Frekuensi 1	Persentase 0	Persentase 1
blue	0	709	691	0.506429	0.493571
dual_sim	1	696	704	0.497143	0.502857
four_g	1	658	742	0.470000	0.530000
three_g	1	335	1065	0.239286	0.760714
touch_screen	0	715	685	0.510714	0.489286
wifi	0	707	693	0.505000	0.495000

Gambar 2.1 Statistik dasar Nominal

   b. Ordinal

      Untuk jenis ordinal, kolom yang tergolong dalam jenis ini adalah kolom price_range. Kemudian statistik dasar yang akan dicari yaitu modus, median, Q1, Q2, dan IQR. Hasilnya adalah seperti berikut.

	Price Range
Modus	0.0
Median	1.0
Q1	0.0
Q3	2.0
IQR	2.0

Gambar 2.2. Statistik dasar Ordinal

   c. Kuantitatif

Untuk jenis kuantitatif, kolom yang tergolong dalam jenis ini adalah kolom clock_speed, fc, int_memory, m_dep, mobile_wt, n_cores, pc, px_height, px_width, ram, sc_h, sc_w, dan talk_time. Kemudian statistik dasar yang akan dicari yaitu mean, median, modus, standar deviasi, variasi, range, Q1, Q2, Q3, IQR, skewness, dan kurtosis. Hasilnya adalah seperti berikut.

	Mean	Median	Modus	Standar Deviasi	Variasi	Range	Kuartil 1	Kuartil 2	Kuartil 3	IQR	Skewness	Kurtosis
battery_power	1237.145714	1219.0	772, 1068, 1330, 1872, 1949	430.051785	1.849445e+05	1497.0	864.75	1219.0	1602.00	737.25	0.041901	-1.168068
clock_speed	1.521714	1.5	0.5	0.814723	6.637740e-01	2.5	0.70	1.5	2.20	1.50	0.166399	-1.329523
fc	4.275000	3.0	0	4.324170	1.869845e+01	19.0	1.00	3.0	7.00	6.00	1.020324	0.293404
int_memory	31.962143	32.0	27	18.162970	3.298935e+02	62.0	16.00	32.0	48.00	32.00	0.063166	-1.227200
m_dep	0.507857	0.5	0.1	0.288539	8.325488e-02	0.9	0.20	0.5	0.80	0.60	0.059116	-1.266823
mobile_wt	139.375714	139.0	182	35.400803	1.253217e+03	120.0	108.00	139.0	169.00	61.00	0.020013	-1.210202
n_cores	4.481429	4.0	4	2.279836	5.197653e+00	7.0	2.00	4.0	7.00	5.00	0.019913	-1.232209
pc	9.917143	10.0	10	6.080023	3.696668e+01	20.0	5.00	10.0	15.00	10.00	0.028708	-1.163876
px_height	643.177857	561.0	88, 347, 526	444.628980	1.976949e+05	1960.0	273.75	561.0	950.25	676.50	0.659456	-0.316229
px_width	1251.717143	1247.0	1247	428.982850	1.840263e+05	1498.0	876.50	1247.0	1627.50	751.00	0.004023	-1.176025
ram	2106.731429	2102.0	1229, 3142	1078.347277	1.162833e+06	3742.0	1201.00	2102.0	3035.75	1834.75	0.029403	-1.186141
sc_h	12.285714	12.0	17	4.204198	1.767528e+01	14.0	9.00	12.0	16.00	7.00	-0.103474	-1.183273
sc_w	5.665000	5.0	1	4.372234	1.911643e+01	18.0	2.00	5.0	9.00	7.00	0.670751	-0.334641
talk_time	11.042143	11.0	15	5.399052	2.914976e+01	18.0	6.00	11.0	16.00	10.00	-0.009319	-1.192018

Gambar 2.3 Statistik dasar Kuantitatif.

2. *Duplicate value*

Nilai duplikat mengacu pada keberadaan data yang serupa atau identik dalam suatu kumpulan data. Dalam pengelolaan data atau basis data, nilai duplikat dapat muncul dalam satu atau lebih entitas (baris atau catatan) dalam suatu koleksi data.

Untuk mencari nilai duplikat ini kami mengimplementasikannya dengan membuat fungsi bernama check_duplicate_values yang meneriman parameter data berupa DataFrame. Untuk hasil pengeksekusian fungsinya adalah seperti berikut.

```python
def check_duplicate_values(data):
    duplicate_rows = data[data.duplicated()]

    if duplicate_rows.shape[0] == 0:
        print("Tidak terdapat duplicate value")
    else:
        print("Terdapat duplicate value")
        print(duplicate_rows)

check_duplicate_values(df)
```

```
Tidak terdapat duplicate value
```

Gambar 2.2. Fungsi untuk mengecek nilai duplikat dan hasil eksekusinya.

Dari hasil eksekusi fungsi check_duplicate_values tersebut, dilihat bahwa tidak terdapat nilai yang duplikat, artinya semua data yang ada itu identik.

3. *Missing value*

Ketidaktersediaan nilai, yang juga dikenal sebagai *missing value*, mengacu pada situasi di mana data yang seharusnya ada dalam suatu kumpulan data tidak ada atau tidak tercatat. Keadaan ini sering kali disebabkan oleh berbagai alasan dan dapat memiliki konsekuensi penting dalam analisis data dan proses pengambilan keputusan.

Untuk mencari *missing value* ini kami mengimplementasikannya dengan membuat fungsi check_missing_values yang akan menerima sebuah data berupa DataFrame. Untuk hasil pengeksekusian fungsinya adalah seperti berikut

```python
def check_missing_values(data):
    missing_values = data.isnull().sum()

    if missing_values.sum() == 0:
        print("Tidak terdapat missing values")
    else:
        print("Terdapat missing values: ")
        for column, count in missing_values.items():
            if count > 0:
                print(f"{column}: {count} missing values")

check_missing_values(df1)
```

```
Tidak terdapat missing values
```

Gambar 2.3. Fungsi untuk mengecek *missing value* dan hasil eksekusinya.

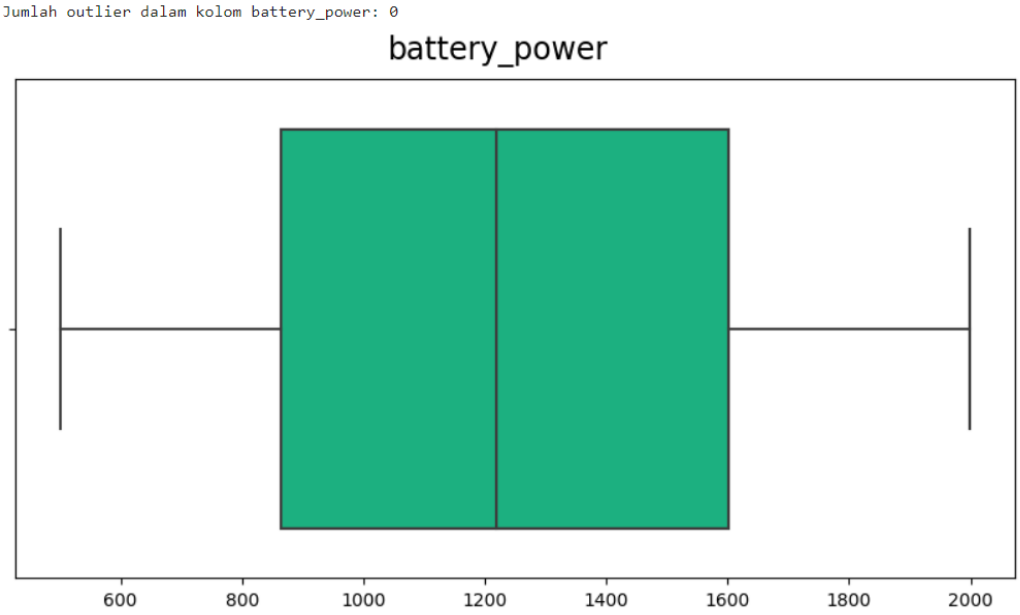
Dari hasil eksekusi fungsi check_missing_values tersebut, dilihat bahwa tidak terdapat nilai yang hilang, artinya semua data itu ada.

4. *Outlier*

Dalam dataset yang tersedia, terdapat 14 atribut numerik, dan kita akan memfokuskan perhatian pada ke-14 kolom tersebut dalam upaya untuk mengidentifikasi outlier, sebab mereka dianggap lebih berhubungan dengan kasus outlier. Visualisasi berupa grafik box plot

digunakan untuk memudahkan identifikasi data outlier karena memberikan wawasan tentang distribusi data, nilai median, kuartil, dan jangkauan interkuartil. Dalam konteks ini, berikut adalah hasil identifikasi outlier pada setiap kolom tersebut.

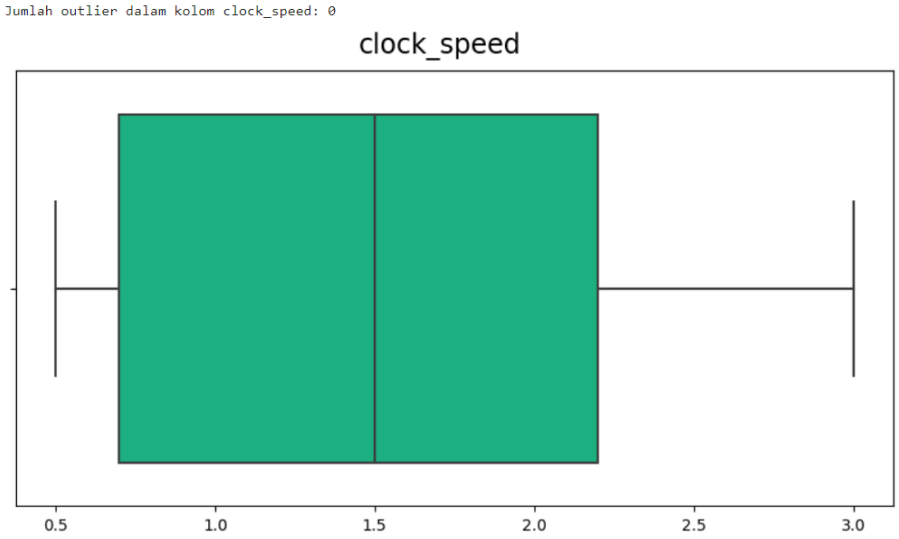
a. Kolom battery_power



Gambar 4.1. Boxplot kolom battery_power

Tidak terdapat nilai ekstrim pada kolom "battery_power," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

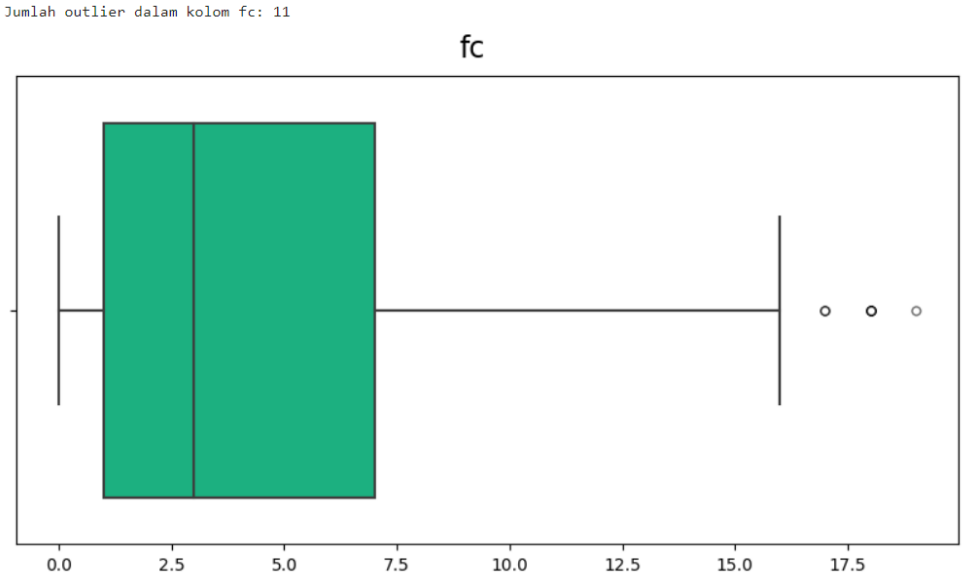
b. Kolom clock_speed



Gambar 4.2. Boxplot kolom clock_speed

Tidak terdapat nilai ekstrim pada kolom "clock_speed," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

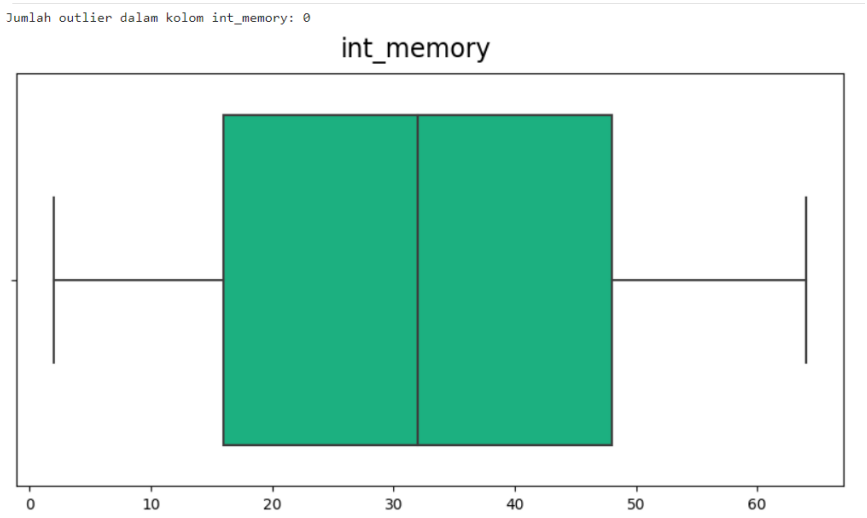
c. Kolom fc



Gambar 4.3. Boxplot kolom fc

Terdapat nilai ekstrim pada kolom "fc," hal ini terlihat dari grafik box plot yang menunjukkan adanya outlier di atas batas atas. Terdapat 11 data yang berada pada outlier.

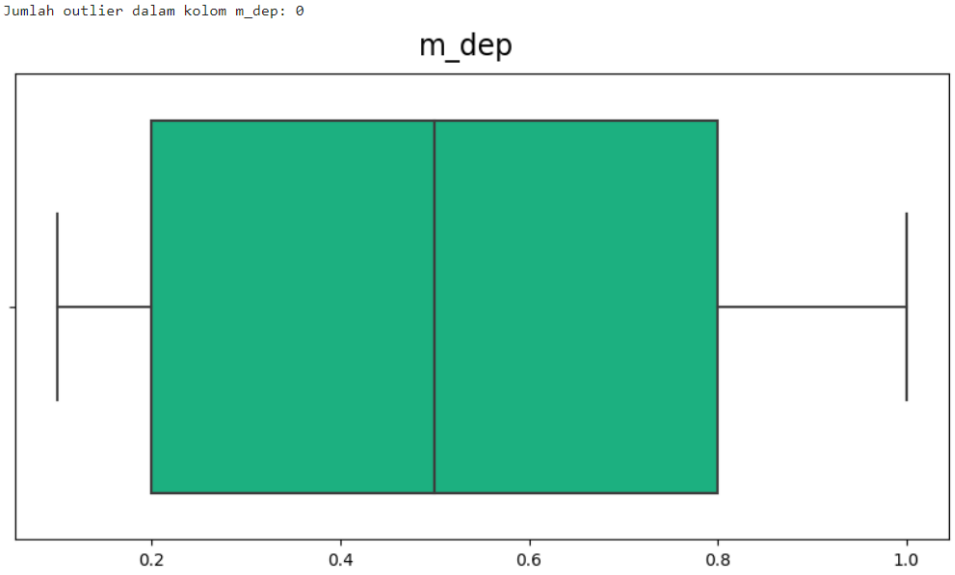
d. Kolom int_memory



Gambar 4.4. Boxplot kolom int_memory

Tidak terdapat nilai ekstrim pada kolom "int_memory," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

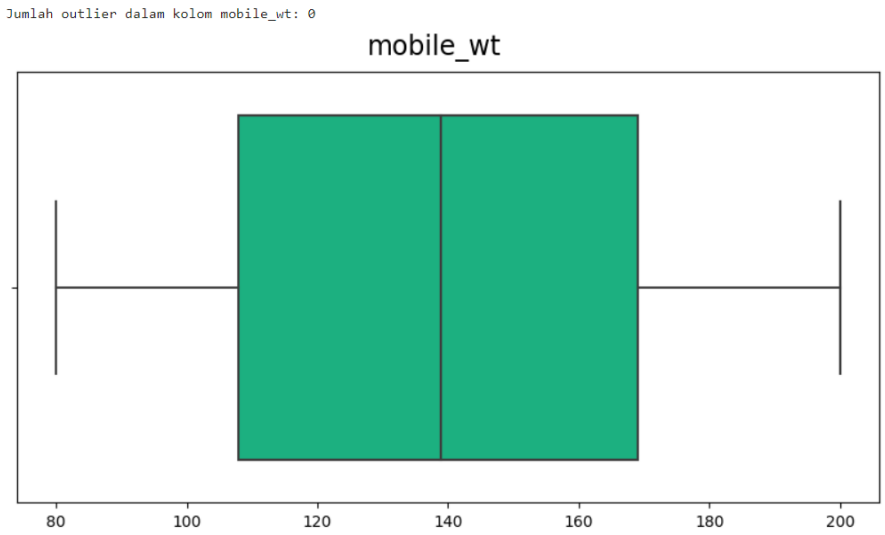
e. Kolom m_dep



Gambar 4.5. Boxplot kolom m_dep

Tidak terdapat nilai ekstrim pada kolom "m_dep," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

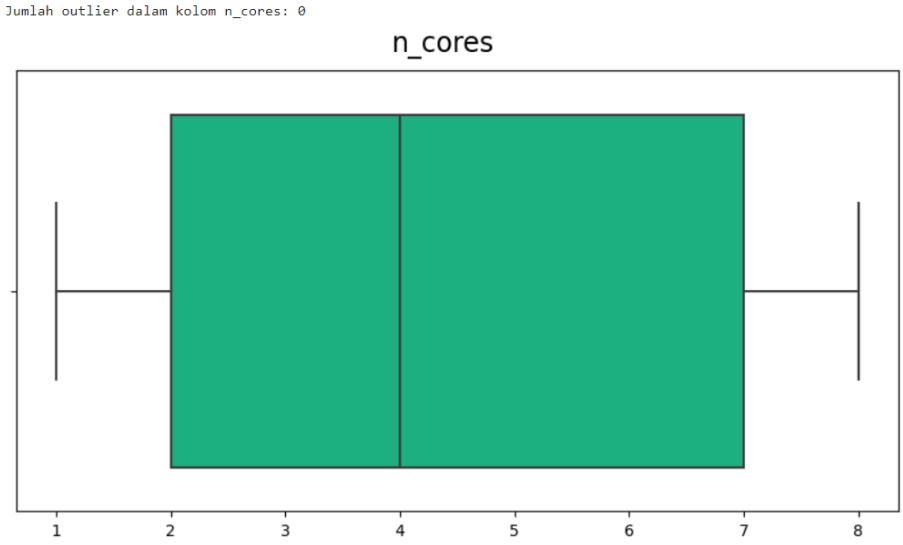
f. Kolom mobile_wt



Gambar 4.6. Boxplot kolom mobile_wt

Tidak terdapat nilai ekstrim pada kolom "mobile_wt" hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

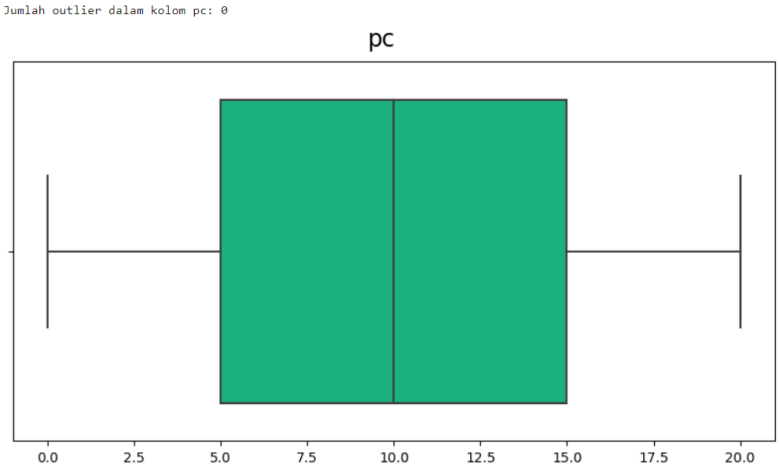
g. Kolom n_cores



Gambar 4.7. Boxplot kolom n_cores

Tidak terdapat nilai ekstrim pada kolom "n_cores," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

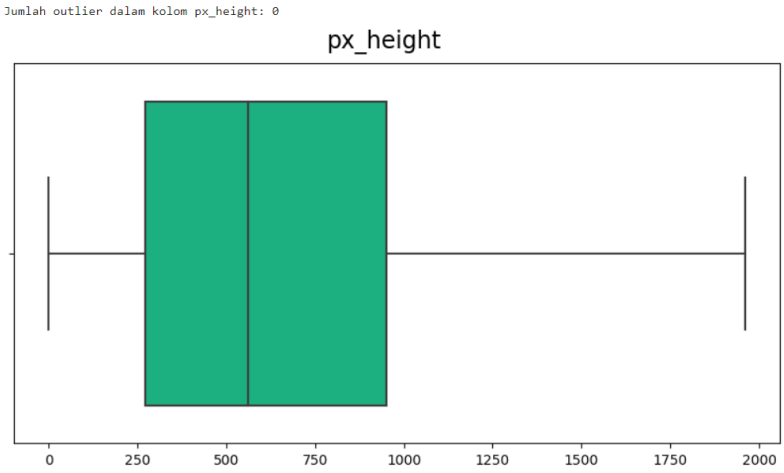
h. Kolom pc



Gambar 4.8. Boxplot kolom pc

Tidak terdapat nilai ekstrim pada kolom "pc," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

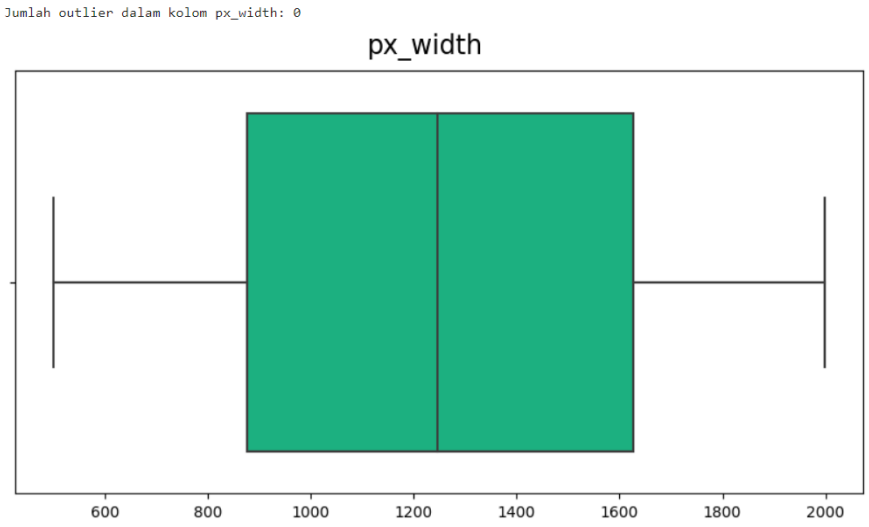
i. Kolom px_height



Gambar 4.9. Boxplot kolom px_height

Tidak terdapat nilai ekstrim pada kolom "px_height," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

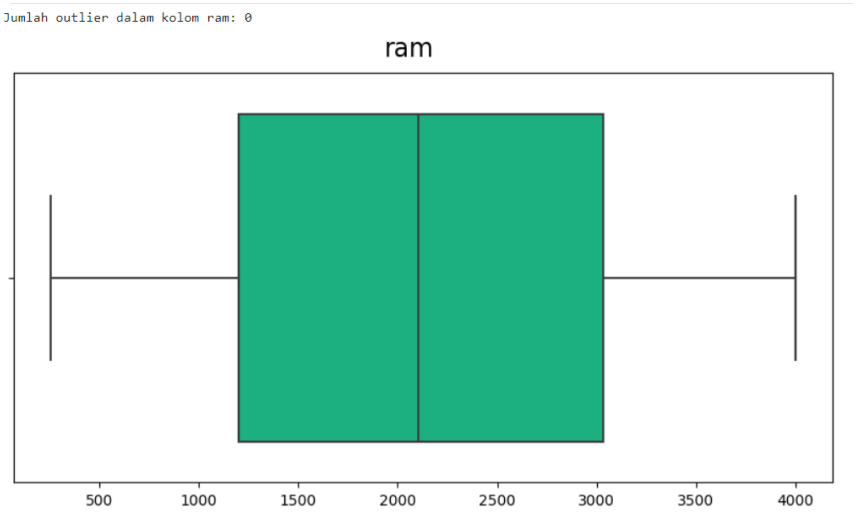
j. Kolom px_width



Gambar 4.10. Boxplot kolom px_width

Tidak terdapat nilai ekstrim pada kolom "px_width" hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

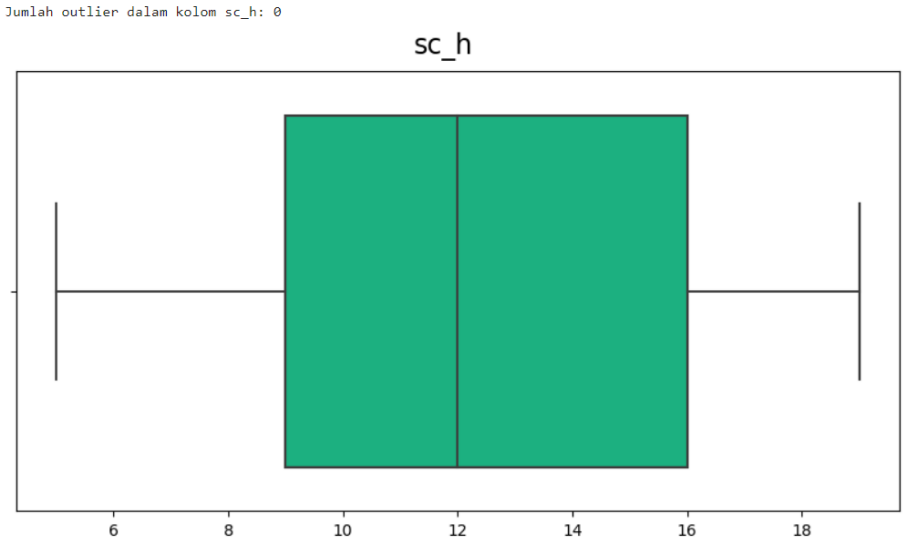
k. kolom ram



Gambar 4.11. Boxplot kolom ram

Tidak terdapat nilai ekstrim pada kolom "ram" hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

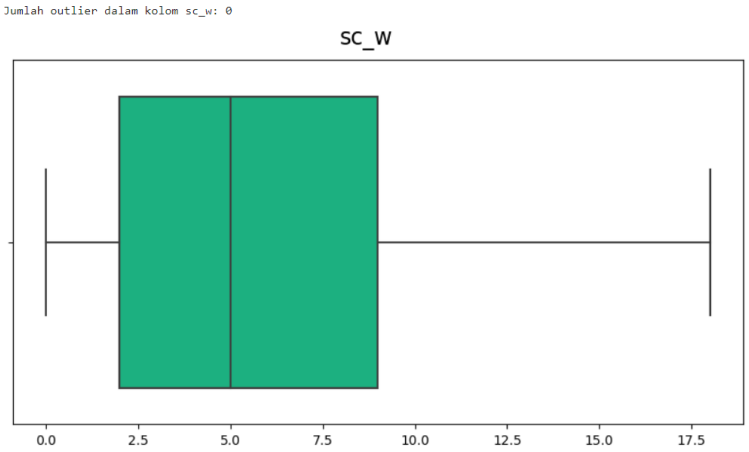
l. Kolom cs_h



Gambar 4.12. Boxplot kolom sc_h

Tidak terdapat nilai ekstrim pada kolom "sc_h" hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

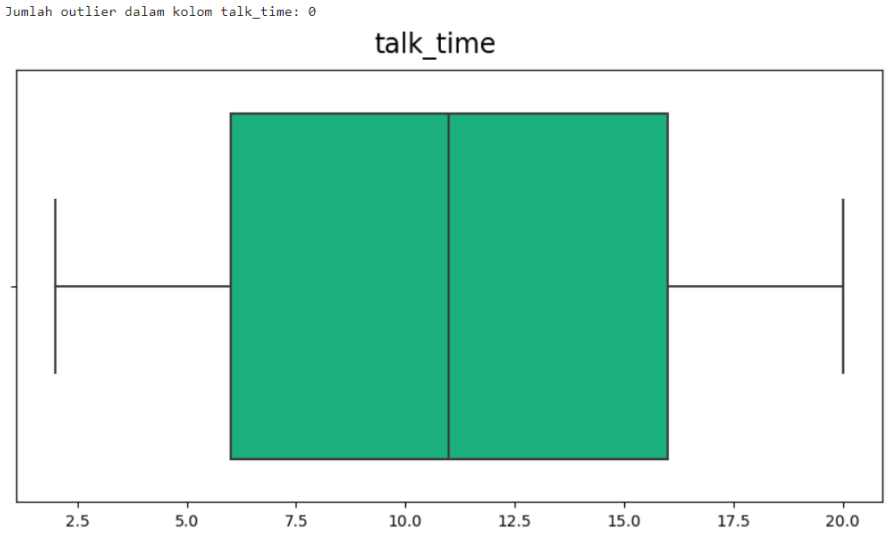
m. Kolom sc_w



Gambar 4.13. Boxplot kolom sc_w

Tidak terdapat nilai ekstrim pada kolom "sc_w," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

n. Kolom talk_time



Gambar 4.14. Boxplot kolom talk_time

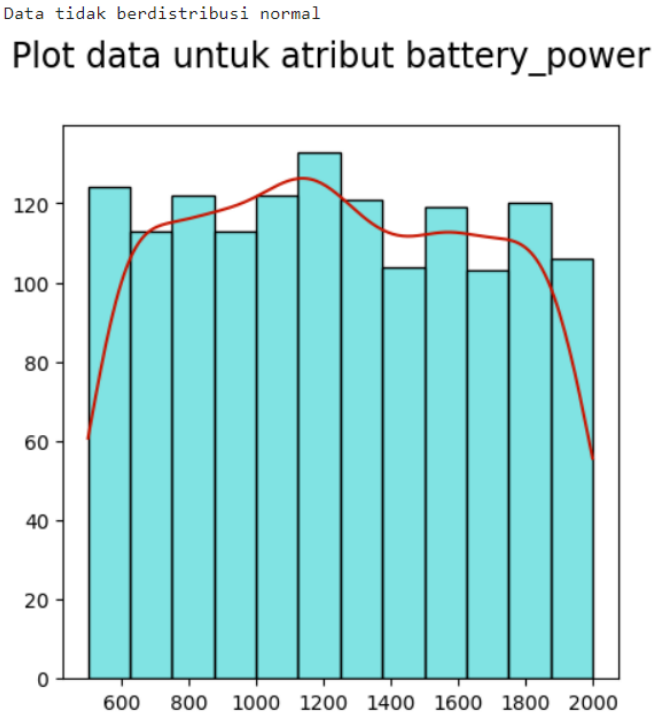
Tidak terdapat nilai ekstrim pada kolom "talk_time," hal ini terlihat dari grafik box plot yang tidak menunjukkan adanya outlier baik di bawah batas bawah maupun di atas batas atas.

5. Distribusi data dan histogram/barchart

Dalam dataset yang tersedia, terdapat 14 atribut numerik, dan untuk data numerik akan dibuat histogram, selain itu akan dilakukan analisis kurtosis untuk mengetahui persebaran data apakah tersebar secara normal atau tidak. Untuk data set non numerik akan dilakukan ditampilkan secara barchart

a. Untuk kolom numerik: distribusi data (plot dan analisis kurtosis)

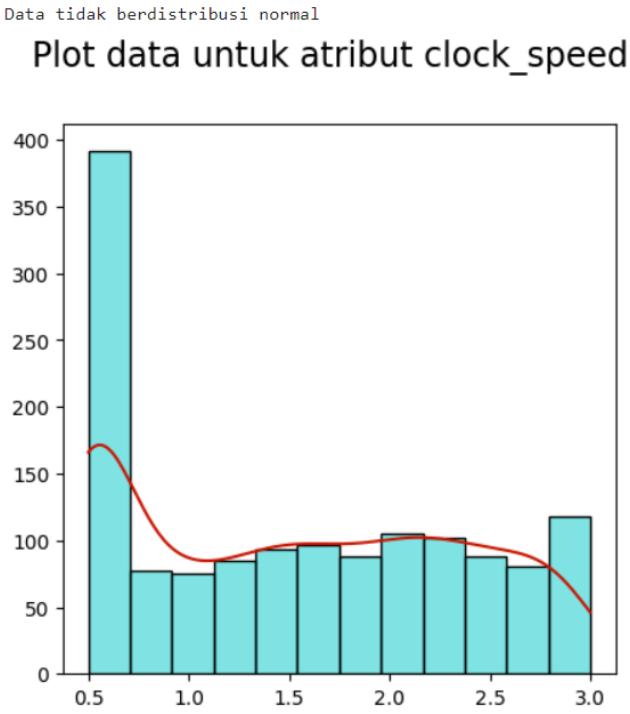
- battery_power



Gambar 5.1. Plot data kolom battery_power

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

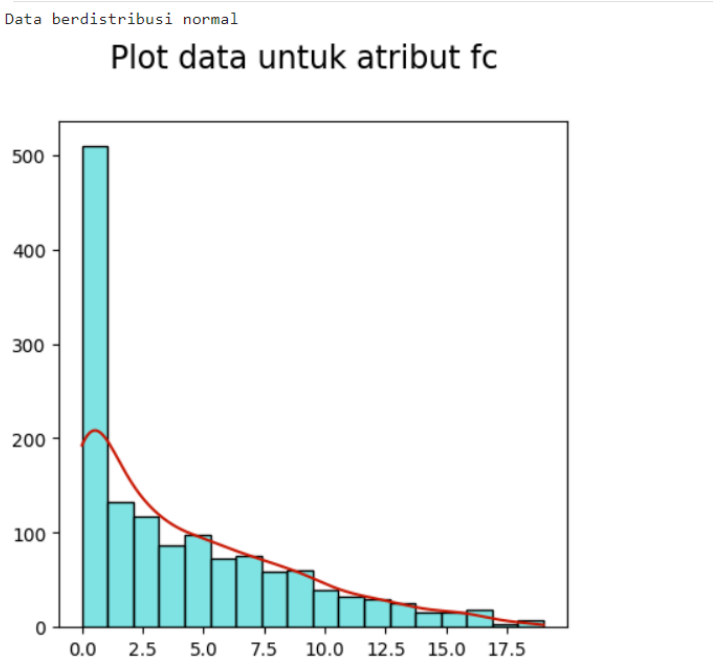
- clock_speed



Gambar 5.2. Plot data kolom clock_speed

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

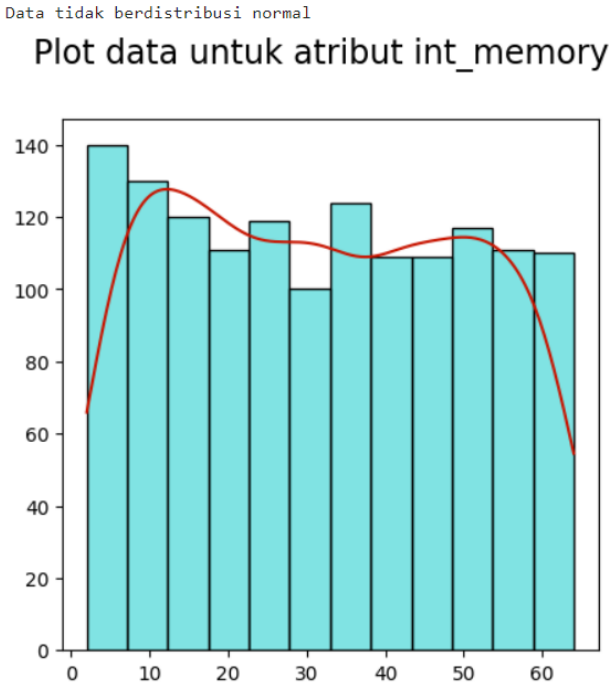
- fc



Gambar 5.3. Plot data kolom fc

Karena nilai mutlak dari kurtosis atribut ini kurang dari 1, maka data pada kolom ini terdistribusi normal

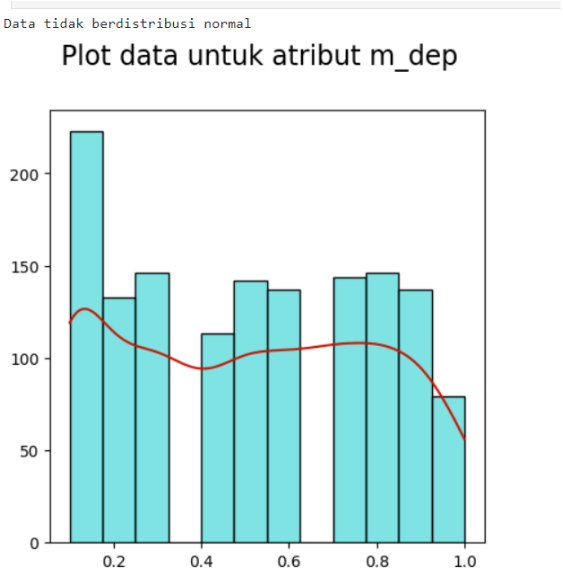
- int_memory



Gambar 5.4. Plot data kolom int_memory

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

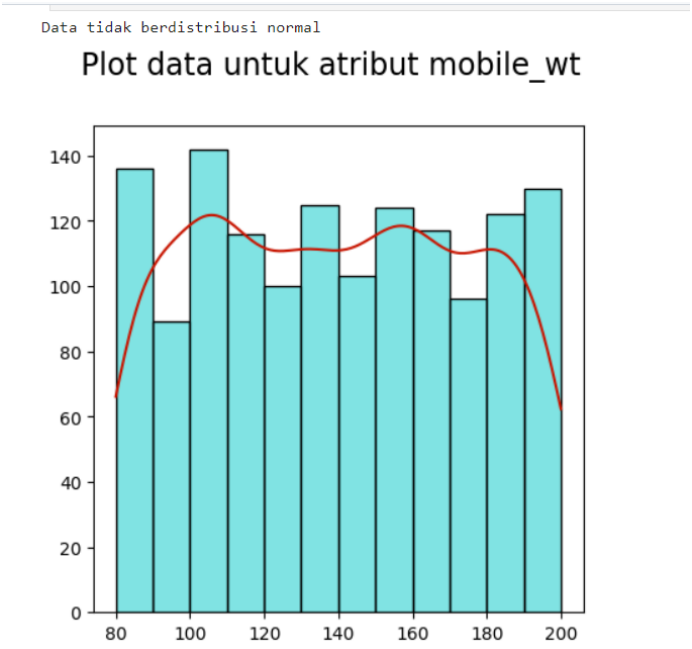
- m_dep



Gambar 5.5. Plot data kolom m_dep

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

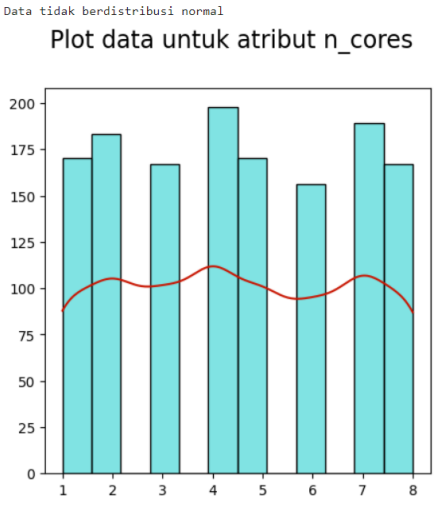
- mobile_wt



Gambar 5.6. Plot data kolom mobile_wt

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

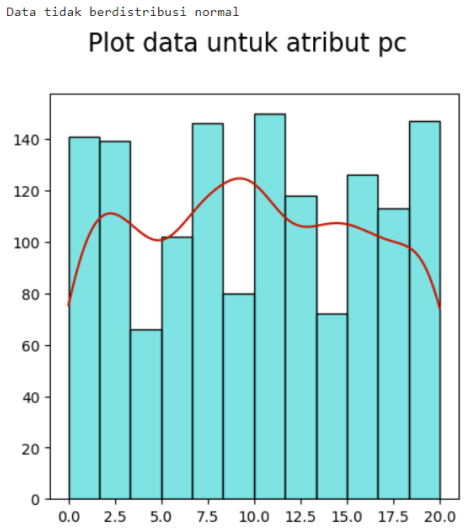
- n_cores



Gambar 5.7. Plot data kolom n_cores

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

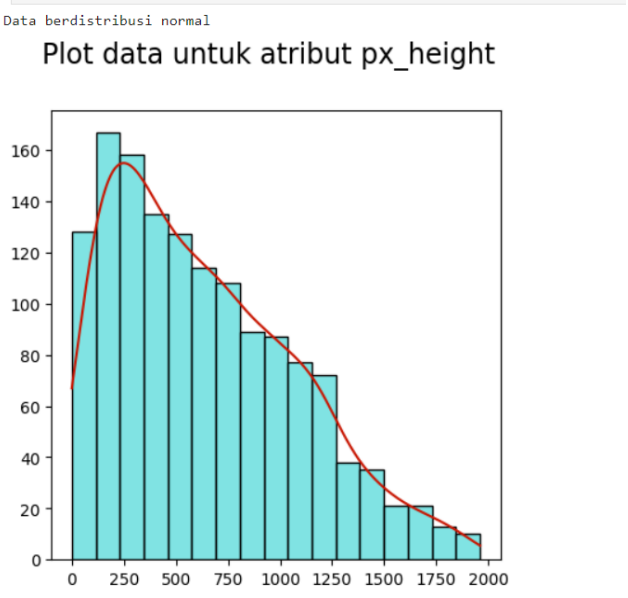
- pc



Gambar 5.8. Plot data kolom pc

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

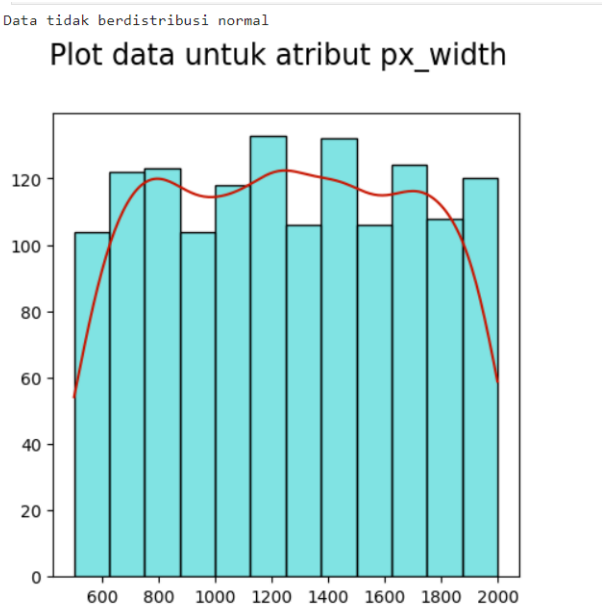
- px_height



Gambar 5.9. Plot data kolom px_height

Karena nilai mutlak dari kurtosis atribut ini kurang dari 1, maka data pada kolom ini terdistribusi normal

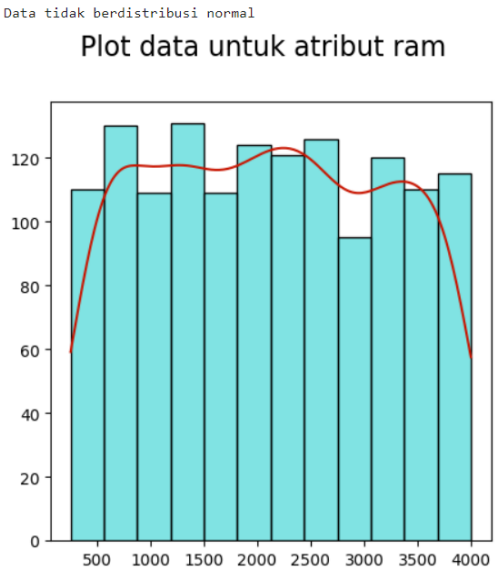
- px_width



Gambar 5.10. Plot data kolom px_width

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal.

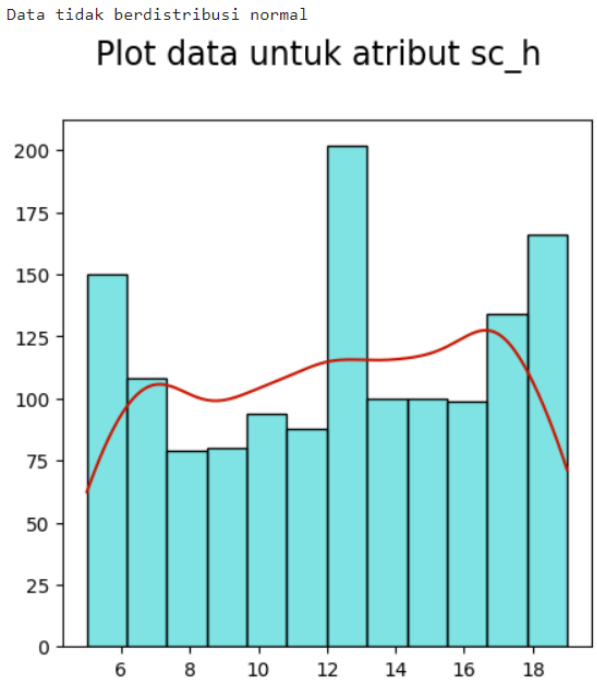
- ram



Gambar 5.11. Plot data kolom ram

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal.

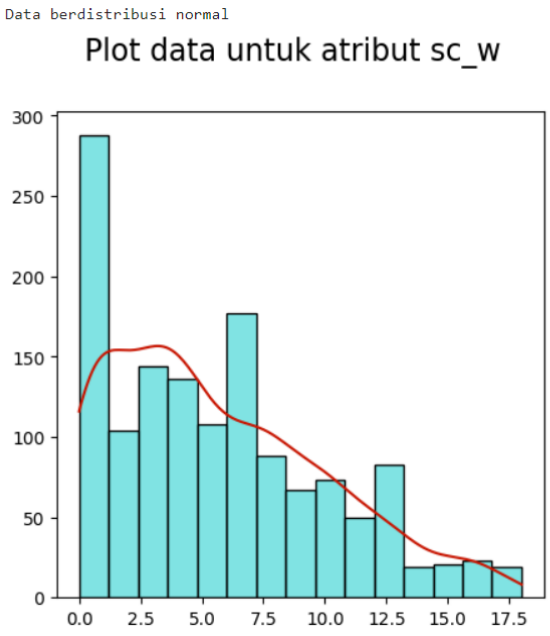
- sc_h



Gambar 5.12. Plot data kolom sc_h

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

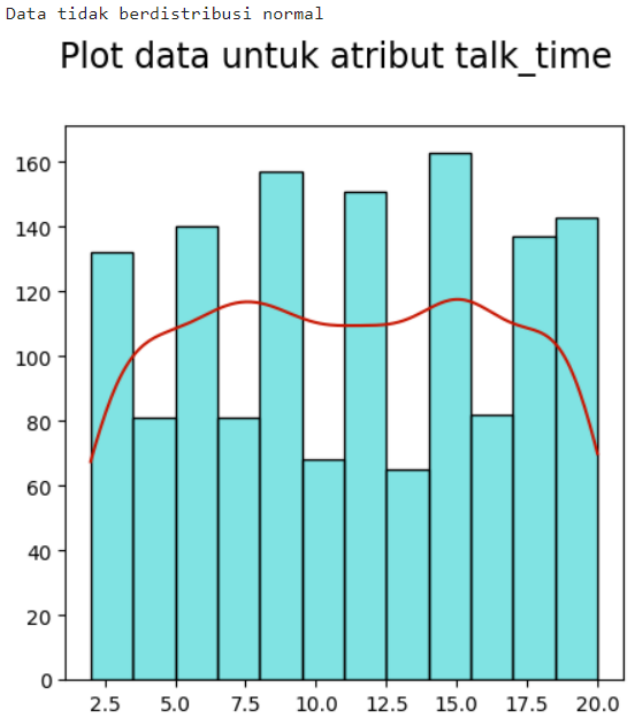
- sc_w



Gambar 5.13. Plot data kolom sc_w

Karena nilai mutlak dari kurtosis atribut ini kurang dari 1, maka data pada kolom ini terdistribusi normal

- talk_time

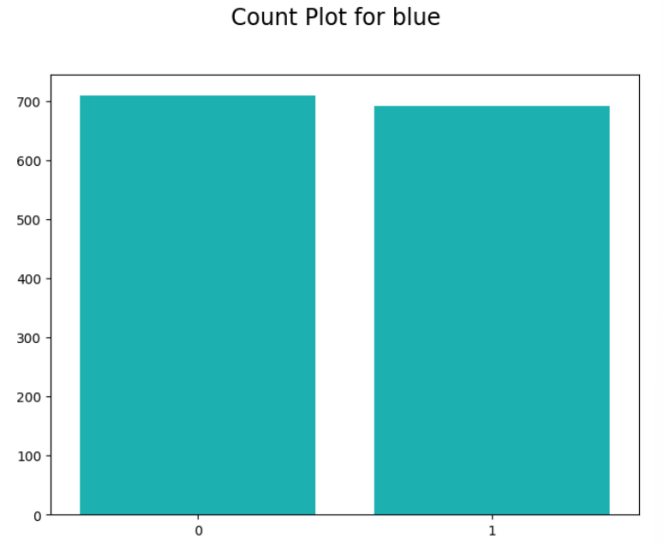


Gambar 5.14. Plot data kolom talk_time

Karena nilai mutlak dari kurtosis atribut ini lebih dari 1, maka data pada kolom ini tidak terdistribusi normal

b. Untuk kolom non numerik: barchart

- blue

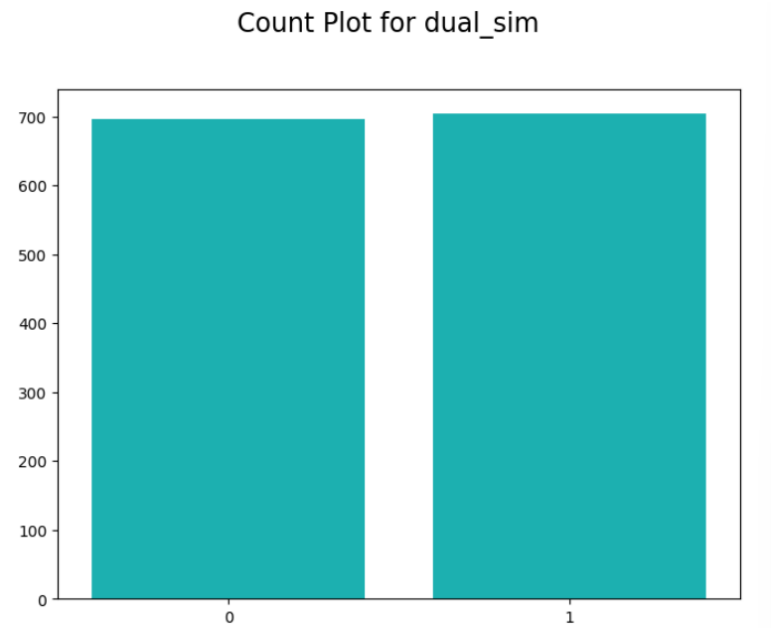


Gambar 5.15. Barchart kolom blue

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
709	691

- dual_sim

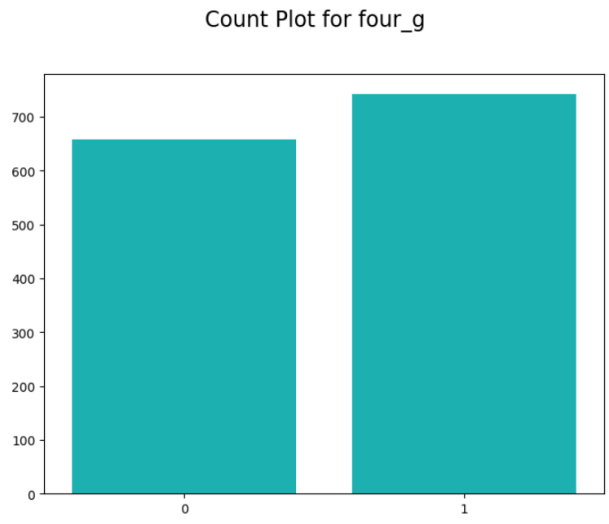


Gambar 5.16. Barchart kolom dual_sim

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
704	696

- four_g

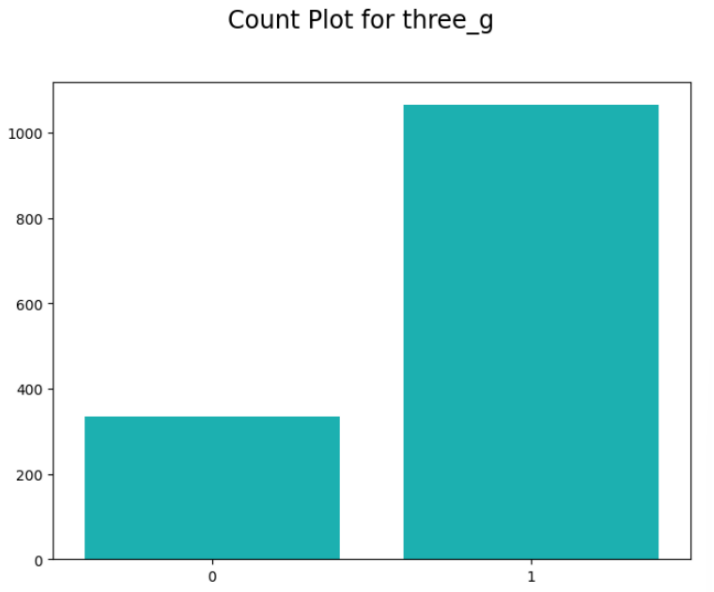


Gambar 5.17. Barchart kolom four_g

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
742	658

- three_g

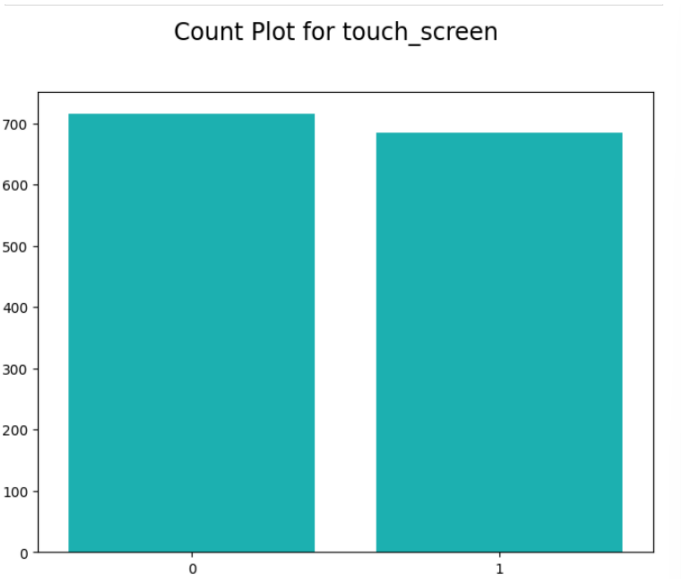


Gambar 5.18. Barchart kolom thee_g

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
1065	334

- touch_screen

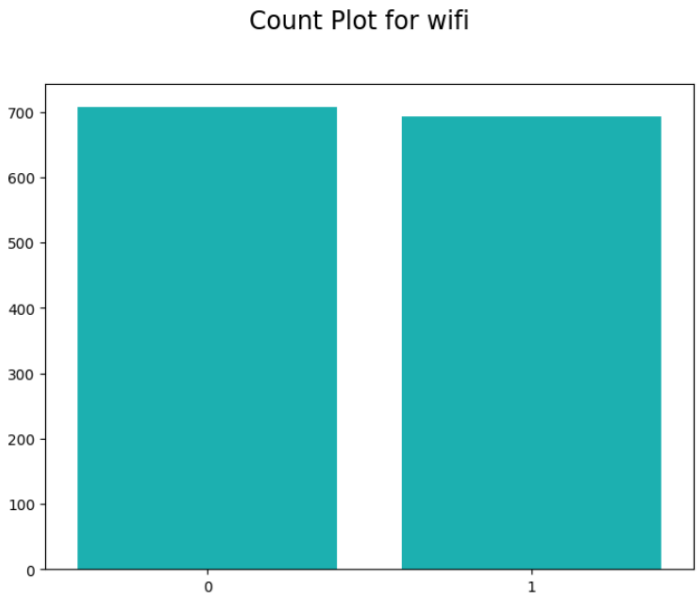


Gambar 5.19. Barchart kolom touch_screen

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
715	685

- wifi

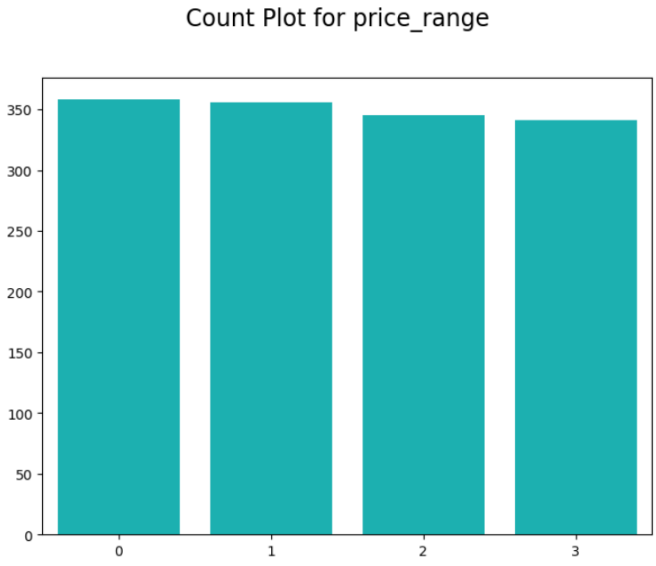


Gambar 5.20. Barchart kolom wifi

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1
707	693

- price_range



Gambar 5.21. Barchart kolom price_range

Frekuensi data :

Frekuensi nilai 0	Frekuensi nilai 1	Frekuensi nilai 2	Frekuensi nilai 3
358	356	345	341

6. Korelasi dengan kolom target

Korelasi adalah suatu konsep dalam statistika yang mengukur sejauh mana dua variabel berkaitan satu sama lain. Dalam permasalah kali ini akan dilakukan pencarian korelasi terhadap atribut price_range yang berperan sebagai target

Untuk kolom numerik, korelasi dicari melalui fungsi corr dari library pandas, fungsi corr merupakan fungsi untuk menentukan korelasi, apabila nilai mutlak dari korelasi lebih dari 0.5 maka data berkorelasi, apabila nilai mutlak dari korelasi kurang dari 0.5 maka atribut tidak berkorelasi

Untuk kolom non-numerik, korelasi dicari melalui ANOVA (analysis of variance) karena ANOVA digunakan untuk membandingkan rata-rata populasi bukan ragam populasi. Sehingga untuk jenis data nominal dan ordinal cocok apabila menggunakan ANOVA

a. Untuk kolom numerik: distribusi data (plot dan analisis kurtosis)

- battery_power

```
correlation_numerik("battery_power")
```

```
Korelasi antara battery_power dan price range adalah 0.18480092449553084
Kolom battery_power tidak berpengaruh terhadap price range
```

- clock_speed

```
correlation_numerik("clock_speed")
```

```
Korelasi antara clock_speed dan price range adalah 0.014031254818008083
Kolom clock_speed tidak berpengaruh terhadap price range
```

- fc

```
correlation_numerik("fc")
```

```
Korelasi antara fc dan price range adalah -0.003842010298191734
Kolom fc tidak berpengaruh terhadap price range
```

- int_memory

```
correlation_numerik("int_memory")
```

```
Korelasi antara int_memory dan price range adalah 0.026175706877841595
Kolom int_memory tidak berpengaruh terhadap price range
```

- m_dep

```
correlation_numerik("m_dep")
```

```
Korelasi antara m_dep dan price range adalah 0.0012049180209846337
Kolom m_dep tidak berpengaruh terhadap price range
```

- mobile_wt

```
correlation_numerik("mobile_wt")
```

```
Korelasi antara mobile_wt dan price range adalah -0.07476875048323661
Kolom mobile_wt tidak berpengaruh terhadap price range
```

- n_cores

```
correlation_numerik("n_cores")
```

```
Korelasi antara n_cores dan price range adalah -0.0005823306285452805
Kolom n_cores tidak berpengaruh terhadap price range
```

- pc

```
correlation_numerik("pc")
```

```
Korelasi antara pc dan price range adalah -0.005214430491652989
Kolom pc tidak berpengaruh terhadap price range
```

- px_height

```
correlation_numerik("px_height")
```

```
Korelasi antara px_height dan price range adalah 0.15883273548307963
Kolom px_height tidak berpengaruh terhadap price range
```

- px_width

```
correlation_numerik("px_width")
```

```
Korelasi antara px_width dan price range adalah 0.1787126901102656
Kolom px_width tidak berpengaruh terhadap price range
```

- ram

```
correlation_numerik("ram")
```

```
Korelasi antara ram dan price range adalah 0.9183192307843839
Kolom ram berpengaruh terhadap price range
```

- sc_h

```
correlation_numerik("sc_h")
```

```
Korelasi antara sc_h dan price range adalah 0.012148883173074988
Kolom sc_h tidak berpengaruh terhadap price range
```

- sc_w

```
correlation_numerik("sc_w")
```

```
Korelasi antara sc_w dan price range adalah 0.019911698810365006
Kolom sc_w tidak berpengaruh terhadap price range
```

- talk_time

```
correlation_numerik("talk_time")
```

```
Korelasi antara talk_time dan price range adalah 0.011112731754754877
Kolom talk_time tidak berpengaruh terhadap price range
```

b. Untuk kolom non numerik: barchart

- blue

```
correlation_non_numerik("blue")
```

```
Kolom blue berpengaruh terhadap price range
Kolom blue berkorelasi positif terhadap price range
```

- dual_sim

```
correlation_non_numerik("dual_sim")
```

```
Kolom dual_sim berpengaruh terhadap price range
Kolom dual_sim berkorelasi positif terhadap price range
```

- four_g

```
correlation_numerik("four_g")
```

```
Korelasi antara four_g dan price range adalah 0.0005508484718002661
Kolom four_g tidak berpengaruh terhadap price range
```

- three_g

```
correlation_non_numerik("three_g")
```

```
Kolom three_g berpengaruh terhadap price range
Kolom three_g berkorelasi positif terhadap price range
```

- touch_Screen

```
correlation_non_numerik("touch_screen")
```

```
Kolom touch_screen berpengaruh terhadap price range
Kolom touch_screen berkorelasi positif terhadap price range
```

- wifi

```
correlation_non_numerik("wifi")
```

```
Kolom wifi berpengaruh terhadap price range
Kolom wifi berkorelasi positif terhadap price range
```

Referensi

- https://www.qualtrics.com/experience-management/research/anova/#:~:text=by%3A%20Aaron%20Carpenter-,What%20is%20ANOVA%3F,more%20unrelated%20samples%20or%20groups.
- https://pandas.pydata.org/
- https://seaborn.pydata.org/
- https://scipy.org/