TUGAS BESAR PENGKOM 2

KELOMPOK 9

KELOMPOK 9

16521321
MARIA SOPHIE DEWANTO

16521331
SALOMO REINHART GREGORY MANALU

16521341
MUHAMMAD RIDZKY SYAFTIRA

16521351
FAKIH ANUGERAH PRATAMA

DESKRIPSI

Deskripsi data dan file

Google PlayStore Android App Data (2.3 Million+ App Data)

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("/content/PlayStore_Cleansed_Dataset_Final.csv")
df
```

df

	App Name	App Id	Category	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Free	Price	Currency	Size	Minimum Android	Developer Id	Developer Website	Developer Email	Released	Last Updated	Content Rating	Privacy Policy	Ad Supported	In App Purchases	Editors Choice	Scraped Time
0	Gakondo	com.ishakwe.gakondo	Adventure	0.0	0.0	10	10	15	True	0.0	USD	0.0 - 50000000.0	7.1 and up	Jean Confident Irénée NIYIZIBYOSE	https://beniyizibyose.tk/#/	jean21101999@gmail.com	2020	Feb 26, 2020	Everyone	https://beniyizibyose.tk/projects/	False	False	False	2021-06-15 20:19:35
1	Ampere Battery Info	com.webservais.batteryinfo	Tools	4.4	64.0	5000	5000	7662	True	0.0	USD	0.0 - 50000000.0	5.0 and up	Webservais	https://webservais.netlify.app/	webservais@gmail.com	2020	May 06, 2021	Everyone	https://dev4phones.wordpress.com/licencia-de-uso/	True	False	False	2021-06-15 20:19:35
2	Vibook	com.doantiepvien.crm	Productivity	0.0	0.0	50	50	58	True	0.0	USD	0.0 - 50000000.0	4.0.3 and up	Cabin Crew	Tidak Tersedia	vnacrewit@gmail.com	2019	Aug 19, 2019	Everyone	https://www.vietnamairlines.com/vn/en/terms-an...	False	False	False	2021-06-15 20:19:35
3	Smart City Trichy Public Service Vehicles 17UC...	cst.stJoseph.ug17ucs548	Communication	5.0	5.0	10	10	19	True	0.0	USD	0.0 - 50000000.0	4.0.3 and up	Climate Smart Tech2	http://www.climatesmarttech.com/	climatesmarttech2@gmail.com	2018	Oct 13, 2018	Everyone	Tidak Tersedia	True	False	False	2021-06-15 20:19:35
4	GROW.me	com.horodyski.grower	Tools	0.0	0.0	100	100	478	True	0.0	USD	0.0 - 50000000.0	4.1 and up	Rafał Milek-Horodyski	http://www.horodyski.com.pl	rmilekhorodyski@gmail.com	2020	Nov 12, 2018	Everyone	http://www.horodyski.com.pl	False	False	False	2021-06-15 20:19:35
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3996	VPGO	com.telerik.VPGO	Productivity	3.9	24.0	5000	5000	6663	True	0.0	USD	0.0 - 50000000.0	5.0 and up	Visual Planning, Stilog	https://www.visual-planning.com/	stilog.dev@stilog.com	2017	Apr 28, 2020	Everyone	https://www.visual-planning.com	False	False	False	2021-06-15 20:23:26
3996	Clamping Tools for machines	c.specialbite.spannmittel	Books & Reference	0.0	0.0	10	10	12	True	0.0	USD	0.0 - 50000000.0	5.0 and up	John Weninger	https://specialbite1001.wixsite.com/spannmittel	specialbite1001@gmail.com	2019	May 27, 2020	Everyone	https://docs.wixstatic.com/ugd/271cdc_7d403d8e...	True	False	False	2021-06-15 20:23:26
3997	CrossFit Bermondsey	com.pushpress.crossfitbermondsey	Health & Fitness	0.0	0.0	50	50	55	True	0.0	USD	0.0 - 50000000.0	5.0 and up	PushPress, Inc.	https://www.crossfitbermondsey.com/	info@crossfitbermondsey.com	2020	Nov 25, 2020	Everyone	https://cfbermondsey.pushpress.com/open/policy...	False	False	False	2021-06-15 20:23:26
3998	The Bean NYC	com.chownow.thebeannyc	Food & Drink	0.0	0.0	10	10	12	True	0.0	USD	0.0 - 50000000.0	5.0 and up	ChowNow	http://www.chownow.com	eng-android@chownow.com	2018	Jul 31, 2019	Everyone	http://www.chownow.com/privacy-policy	False	False	False	2021-06-15 20:23:26
3999	Hot Anime Girls Pictures	com.vardex.hot_anime_girls_pictures	Personalization	3.7	55.0	5000	5000	9118	True	0.0	USD	0.0 - 50000000.0	4.1 and up	vardex	https://vardexapps.blogspot.com	vardexapps@gmail.com	2021	May 20, 2021	Mature 17+	https://vardexapps.blogspot.com/p/privacy.html	True	False	False	2021-06-15 20:23:26

4000 rows × 24 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 24 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   App Name           4000 non-null   object
 1   App Id             4000 non-null   object
 2   Category           4000 non-null   object
 3   Rating             3962 non-null   float64
 4   Rating Count       4000 non-null   float64
 5   Installs           4000 non-null   int64
 6   Minimum Installs   4000 non-null   int64
 7   Maximum Installs   4000 non-null   int64
 8   Free               4000 non-null   bool
 9   Price              4000 non-null   float64
 10  Currency           4000 non-null   object
 11  Size               4000 non-null   object
 12  Minimum Android    4000 non-null   object
 13  Developer Id       4000 non-null   object
 14  Developer Website  4000 non-null   object
 15  Developer Email    4000 non-null   object
 16  Released           4000 non-null   object
 17  Last Updated       4000 non-null   object
 18  Content Rating     4000 non-null   object
 19  Privacy Policy     4000 non-null   object
 20  Ad Supported       4000 non-null   bool
 21  In App Purchases   4000 non-null   bool
 22  Editors Choice     4000 non-null   bool
 23  Scraped Time       4000 non-null   object
dtypes: bool(4), float64(3), int64(3), object(14)
memory usage: 640.8+ KB
```

Data ini diambil dari https://www.kaggle.com/gauthamp10/google-playstore-apps

KARAKTERISTIK DATA

Penjelasan Karakteristik Atribut/Kolom data

Atribut/Kolom	Jenis	Makna	Penjelasan	Bahasa Pemrograman
App Name	Kategorikal - Nominal	Nama Aplikasi	Banyak value: 3997	df["App Name"].value_counts()
App Id	Kategorikal - Nominal	Id Aplikasi	Banyak value: 4000	df["App Id"].value_counts()
Category	Kategorikal - Nominal	Kategori Aplikasi	Banyak value: 48	df["Category"].value_counts()
Rating	Kategorikal - Ordinal	Nilai Penilaian	Banyak value: 41	df["Rating"].value_counts()
Rating Count	Kuantitatif - Discrete	Banyak Penilaian	Range value: 0.0 – 1264773.0	df.min()[Rating Count] df.max()[Rating Count]
Installs	Kuantitatif - Discrete	Pengunduhan	Range value: 0 – 100000000	df.min()[Installs] df.max()[Installs]
Minimum Installs	Kuantitatif - Discrete	Pengunduhan Minimum	Range value: 0 – 100000000	df.min()[Minimum Installs] df.max()[Minimum Installs]
Maximum Installs	Kuantitatif - Discrete	Pengunduhan Maximum	Range value: 0 – 125445533	df.min()[Maximum Installs] df.max()[Maximum Installs]
Free	Kategorikal - Binary	Aplikasi Gratis	True (3915), False (85)	df["Free"].value_counts()
Price	Kategorikal - Nominal	Harga Aplikasi	Banyak value: 20	df["Price"].value_counts()
Currency	Kategorikal - Nominal	Mata Uang	Banyak value: 1 ▪ USD (4000)	df["Currency"].value_counts()
Size	Kategorikal - Nominal	Ukuran Aplikasi	Banyak value: 7	df["Size"].value_counts()
Minimum Android	Kategorikal - Nominal	Versi Android Minimum	Banyak value: 25	df["Minimum Android"].value_counts()
Developer Id	Kategorikal - Nominal	Id Developer	Banyak value: 3770	df["Developer Id"].value_counts()
Developer Website	Kategorikal - Nominal	Website Developer	Banyak value: 2541	df["Developer Website"].value_counts(
Developer Email	Kategorikal - Nominal	Email Developer	Banyak value: 3775	df["Developer Email"].value_counts()
Released	Kategorikal - Nominal	Tahun Keluar	Banyak value: 13	df["Released"].value_counts()
Last Updated	Kategorikal - Nominal	Tanggal Pembaruan Terakhir	Banyak value: 1421	df["Last Update"].value_counts()

Atribut/Kolom	Jenis	Makna	Penjelasan	Bahasa Pemrograman
Content Rating	Kategorikal - Nominal	Kategori Umur	Banyak value: 4 ▪ Everyone (3507) ▪ Teen (325) ▪ Mature 17+ (104) ▪ Everyone 10+ (64)	df[“Content Rating”].value_counts()
Privacy Policy	Kategorikal - Nominal	Kebijakan Pribadi Aplikasi	Banyak value: 3063	df[“Privacy Policy”].value_counts()
Ad Supported	Kategorikal - Binary	Dukungan Iklan	Terdiri dari 2 value: False (2076), True (1924)	df[“Ad Supported”].value_counts()
In App Purchase	Kategorikal - Binary	Pembelian Dalam Aplikasi	Terdiri dari 2 value: False (3696), True (304)	df[“In App Purchase”].value_counts()
Editors Choice	Kategorikal - Binary	Pilihan *Editor's*	Terdiri dari 2 value: False (3998), True (2)	df[“Editors Choice”].value_counts()
Scraped Time	Kategorikal - Nominal	Waktu Aplikasi Dihapus	Terdiri dari 232 value	df[“Scraped Time].value_counts()

Statistik

Statistik dan sampel data

Mengambil nilai 10 data awal

```
df[['App Name', 'Category', 'Rating', 'Installs',
    'Free', 'Released', 'Developer Id']][:10]
```

	App Name	Category	Rating	Installs	Free	Released	Developer Id
0	Gakondo	Adventure	0.0	10	True	2020	Jean Confident Irénée NIYIZIBYOSE
1	Ampere Battery Info	Tools	4.4	5000	True	2020	Webserveis
2	Vibook	Productivity	0.0	50	True	2019	Cabin Crew
3	Smart City Trichy Public Service Vehicles 17UC...	Communication	5.0	10	True	2018	Climate Smart Tech2
4	GROW.me	Tools	0.0	100	True	2020	Rafal Milek-Horodyski
5	IMOCCI	Social	0.0	50	True	2018	Imocci GmbH
6	unlimited 4G data prank free app	Libraries & Demo	4.5	1000	True	2019	android developer779
7	The Everyday Calendar	Lifestyle	2.0	500	True	2019	Mozaix LLC
8	WhatsOpen	Communication	0.0	10	True	0	Yilver Molina Hurtatiz
9	Neon 3d Iron Tech Keyboard Theme	Personalization	4.7	50000	True	2019	Free 2021 Themes for Emoji keyboard

10 data tertinggi berdasar jumlah install dari terbesar ke yang terkecil

```
df[['App Name', 'Category', 'Rating', 'Installs', 'Free',
 'Released', 'Developer Id']].sort_values(['Installs'],
ascending=[0])[:10]
```

	App Name	Category	Rating	Installs	Free	Released	Developer Id
1313	Screen Recorder, Video Recorder, V Recorder Ed...	Video Players & Editors	4.6	100000000	True	2018	VideoShow EnjoyMobi Video Editor & Video Maker...
3136	Archery Master 3D	Sports	4.2	100000000	True	2015	TerranDroid
1500	Girly Lock Screen Wallpaper with Quotes	Personalization	4.1	10000000	True	2017	Thalia Graphic Image Fusion
92	Car Games Revival: Car Racing Games for Kids	Racing	3.9	10000000	True	2016	Abstron Pvt Ltd
885	Komoot — Cycling, Hiking & Mountain Biking Maps	Health & Fitness	4.2	10000000	True	2014	komoot GmbH
3361	Recarga Celular, Pagar Contas, Boletos e Pix	Finance	4.3	10000000	True	2012	RecargaPay
3305	Factory Inc.	Simulation	4.4	10000000	True	2019	menghuangu inc
561	Web Browser & Fast Explorer	Social	4.4	10000000	True	2016	iJoysoft
631	Piano Melody	Education	4.2	10000000	True	2010	Learn To Master
1524	Kids Airport Adventure	Simulation	4.0	10000000	True	2017	YovoGames

```
df.sort_values(['Minimum Installs', 'Maximum Installs'],
ascending=[0, 1])[['App Name', 'Category', 'Rating', 'Installs',
'Minimum Installs', 'Maximum Installs', 'Released']][:10]
```

	App Name	Category	Rating	Installs	Minimum Installs	Maximum Installs	Released
1313	Screen Recorder, Video Recorder, V Recorder Ed...	Video Players & Editors	4.6	100000000	100000000	102657325	2018
3136	Archery Master 3D	Sports	4.2	100000000	100000000	125445533	2015
2594	AppLock Lite	Tools	4.3	10000000	10000000	10399530	2020
764	Bridge Constructor Playground FREE	Simulation	4.1	10000000	10000000	11050132	2013
885	Komoot — Cycling, Hiking & Mountain Biking Maps	Health & Fitness	4.2	10000000	10000000	11930352	2014
1563	Opus Player - WhatsApp Audio Search and Organize	Music & Audio	3.9	10000000	10000000	12408547	2017
1500	Girly Lock Screen Wallpaper with Quotes	Personalization	4.1	10000000	10000000	12805205	2017
561	Web Browser & Fast Explorer	Social	4.4	10000000	10000000	13629503	2016
631	Piano Melody	Education	4.2	10000000	10000000	13882293	2010
1524	Kids Airport Adventure	Simulation	4.0	10000000	10000000	16153296	2017

Mengambil nilai 10 data pertama dengan Install < 100000

```
df.loc[df['Installs'] < 100000, ['App Name', 'Last Updated', 'Installs']][:10]
```

	App Name	Last Updated	Installs
0	Gakondo	Feb 26, 2020	10
1	Ampere Battery Info	May 06, 2021	5000
2	Vibook	Aug 19, 2019	50
3	Smart City Trichy Public Service Vehicles 17UC...	Oct 13, 2018	10
4	GROW.me	Nov 12, 2018	100
5	IMOCCI	Dec 20, 2019	50
6	unlimited 4G data prank free app	Sep 27, 2019	1000
7	The Everyday Calendar	Jun 21, 2019	500
8	WhatsOpen	Dec 07, 2018	10
9	Neon 3d Iron Tech Keyboard Theme	Oct 07, 2020	50000

Menampilkan statistik penting data

```
df.describe(percentiles=[.1, .25, .5, .75, .9])[1:]
```

	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Price
mean	2.122400	1.535866e+03	1.206866e+05	1.206866e+05	1.784452e+05	0.113113
std	2.109082	2.791034e+04	2.320042e+06	2.320042e+06	2.787123e+06	1.953821
min	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
10%	0.000000	0.000000e+00	1.000000e+01	1.000000e+01	1.700000e+01	0.000000
25%	0.000000	0.000000e+00	5.000000e+01	5.000000e+01	7.600000e+01	0.000000
50%	2.500000	5.000000e+00	5.000000e+02	5.000000e+02	6.245000e+02	0.000000
75%	4.300000	3.700000e+01	5.000000e+03	5.000000e+03	6.408750e+03	0.000000
90%	4.700000	2.878000e+02	5.000000e+04	5.000000e+04	5.742670e+04	0.000000
max	5.000000	1.264773e+06	1.000000e+08	1.000000e+08	1.254455e+08	99.990000

Persebaran data 'Category' dengan groupby()

```
df.groupby('Category')['Category'].count()[:10]
```

```
df.groupby('Category')['Category'].count()[:10]
```

```
Category
Action                 34
Adventure              41
Arcade                 83
Art & Design           30
Auto & Vehicles        28
Beauty                 24
Board                  14
Books & Reference     216
Business              248
Card                   10
Name: Category, dtype: int64
```

Persebaran data 'Category' dengan value_counts()

```
df['Category'].value_counts()[:10]
```

```
Education            417
Music & Audio        289
Business             248
Tools                238
Entertainment        221
Lifestyle            218
Books & Reference    216
Productivity         159
Personalization      141
Health & Fitness     138
Name: Category, dtype: int64
```

Persebaran data 'Size' dengan value_counts()

df['Size'].value_counts()

```
df['Size'].value_counts()
```

```
0.0 - 50000000.0                3571
50000000.0 - 100000000.0         268
Varies with device               121
100000000.0 - 150000000.0         33
150000000.0 - 200000000.0          4
200000000.0 - 250000000.0          2
500000000.0 - 550000000.0          1
Name: Size, dtype: int64
```

Persebaran data 'Minimum Android' dengan value_counts()

```
df['Minimum Android'].value_counts()
```

```
4.1 and up            1036
5.0 and up             697
4.4 and up             681
4.0.3 and up           345
4.0 and up             266
4.2 and up             201
6.0 and up             148
2.3 and up             110
5.1 and up             102
4.3 and up              75
Varies with device      68
7.0 and up              59
2.3.3 and up            43
2.2 and up              39
2.1 and up              33
3.0 and up              27
1.6 and up              17
8.0 and up              13
7.1 and up              11
Tidak Tersedia          10
3.1 and up               6
2.0 and up               4
1.5 and up               4
3.2 and up               4
1.0 and up               1
Name: Minimum Android, dtype: int64
```

Persebaran data 'Released' dengan value_counts()

```
df['Released'].value_counts()
```

```
df['Released'].value_counts()

2020                928
2019                826
2018                577
2017                479
2021                329
2016                293
2015                196
Tidak Tersedia      121
2014                115
2013                 64
2012                 40
2011                 26
2010                  6
Name: Released, dtype: int64
```

Persebaran data 'Content Rating' dengan value_counts()

```
df['Content Rating'].value_counts()
```

```
df['Content Rating'].value_counts()
```

```
Everyone        3507
Teen             325
Mature 17+       104
Everyone 10+      64
Name: Content Rating, dtype: int64
```

Pengetahuan
yang didapat

sort_values()

Menggunakan sort_values(), kita dapat melakukan pengurutan terhadap data sesuai dengan kolom yang kita inginkan dengan pengurutan yang kita inginkan (dari tebesar ke terkecil atau sebaliknya).

describe()

Menggunakan describe(), kita dapat mengetahui dengan jelas data-data statistik penting atribut numerikal yang dimiliki dataset kita secara keseluruhan. Kita dapat mengetahui dengan jelas rata-rata, standar deviasi, nilai minimum dan maksimum, serta persentil dari data.

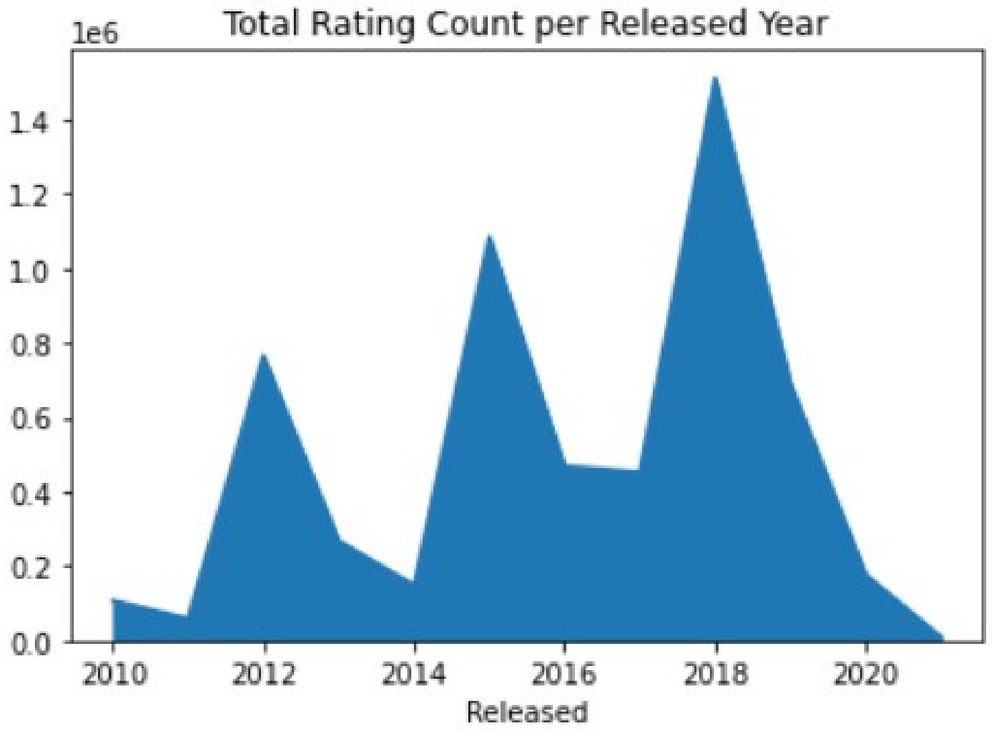
value_counts()

Kita dapat menentukan persebaran data menggunakan beberapa cara, di antaranya menggunakan value_counts() dan groupby().count() yang memiliki fungsi serta kelebihan dan kekurangan masing-masing.

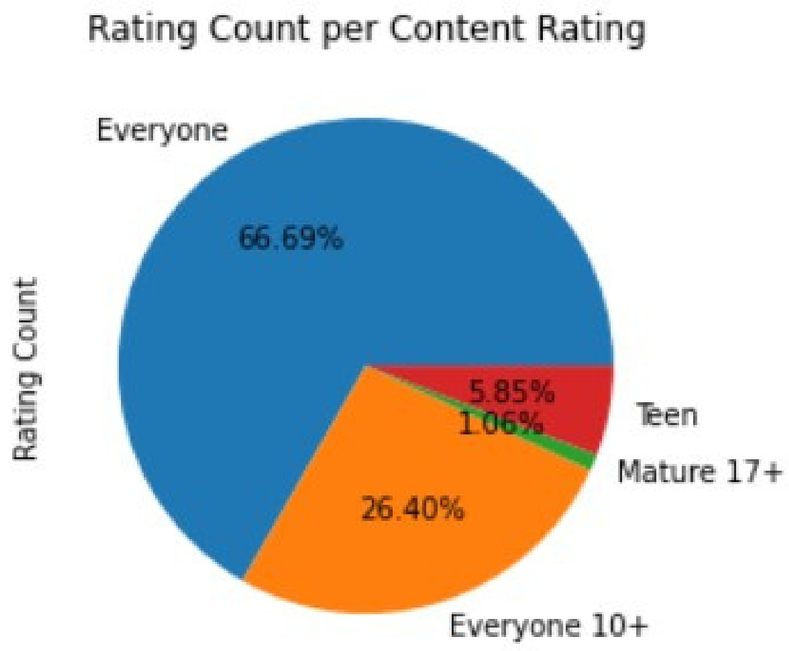
VISUALISASI

Bentuk visual data berupa grafik

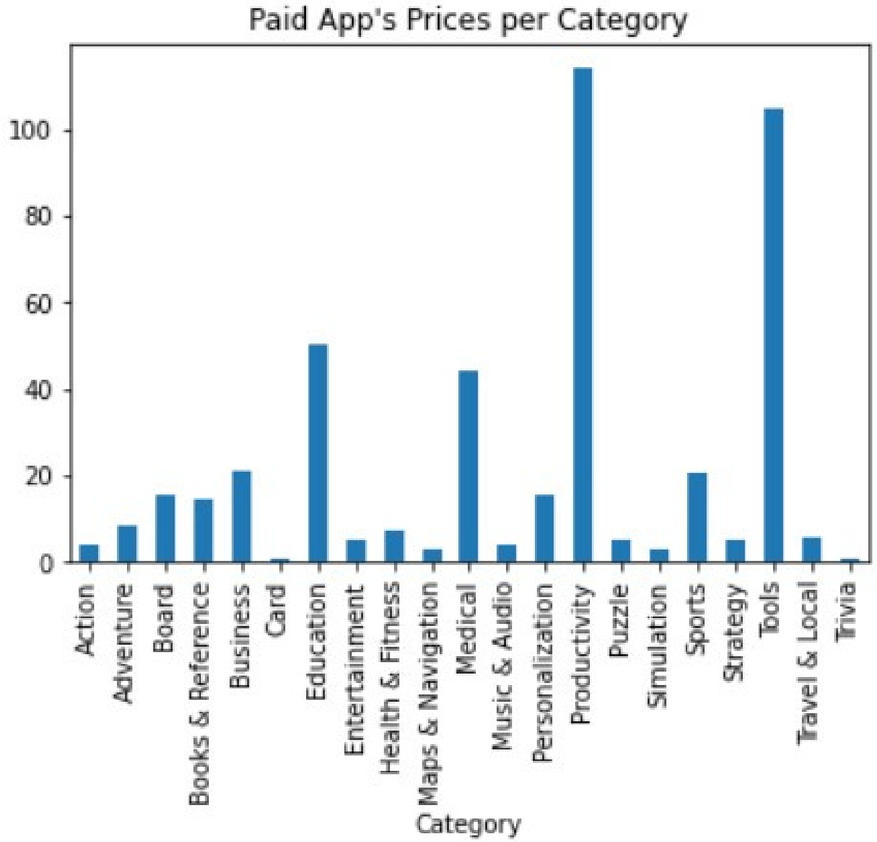
```
df1 = df[df["Released"] != "Tidak Tersedia"].groupby("Released")["Rating Count"].sum()
df1.plot(kind = "area", x = "Category", y = "Rating", title = "Total Rating Count per Released Year")
plt.show()
```



```
df2 = df.groupby("Content Rating")["Rating Count"].sum()
df2.plot(kind = "pie", autopct = "%1.2f%%" , title = "Rating Count per Content Rating")
plt.show()
```



```
df2 = df[df["Free"] == False].groupby("Category")["Price"].sum()
df2.plot(kind = "bar", x = "Price", y = "Category", title = "Paid App's Prices per Category")
plt.show()
```



KORELASI

Korelasi antara semua atribut kuantitatif

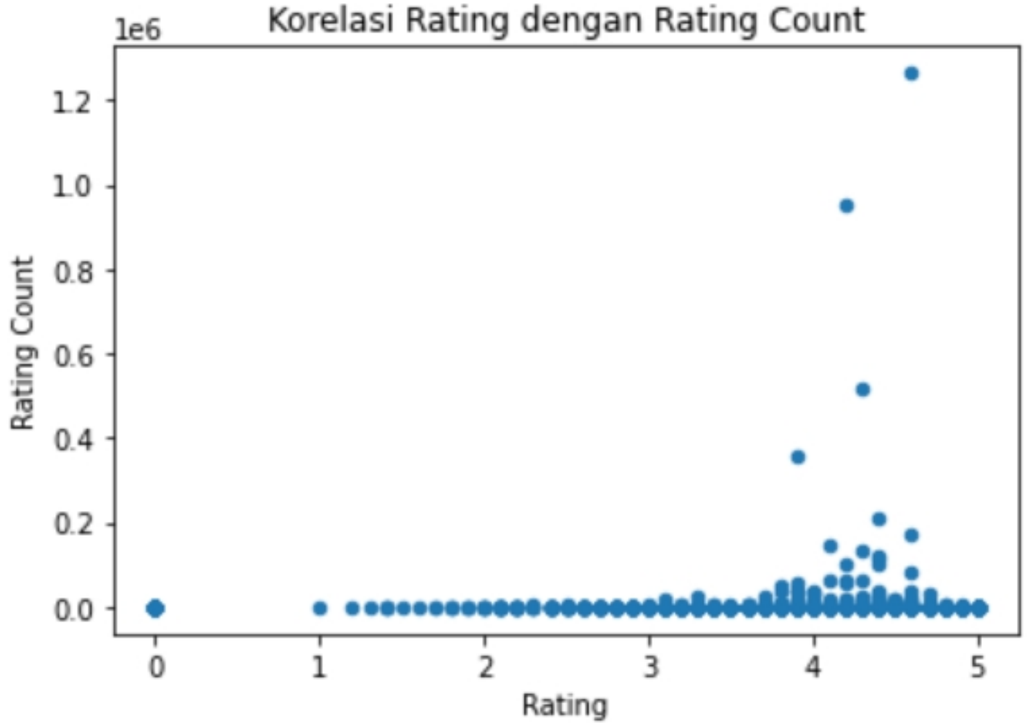
Korelasi antara 6 atribut kuantitatif

	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Price
Rating	1.000000	0.056158	0.051421	0.051421	0.061957	0.014702
Rating Count	0.056158	1.000000	0.927963	0.927963	0.912511	-0.000658
Installs	0.051421	0.927963	1.000000	1.000000	0.978603	-0.002707
Minimum Installs	0.051421	0.927963	1.000000	1.000000	0.978603	-0.002707
Maximum Installs	0.061957	0.912511	0.978603	0.978603	1.000000	-0.003238
Price	0.014702	-0.000658	-0.002707	-0.002707	-0.003238	1.000000

```
df["Rating"].corr(df["Rating Count"])
```

```
0.05615831164311739
```

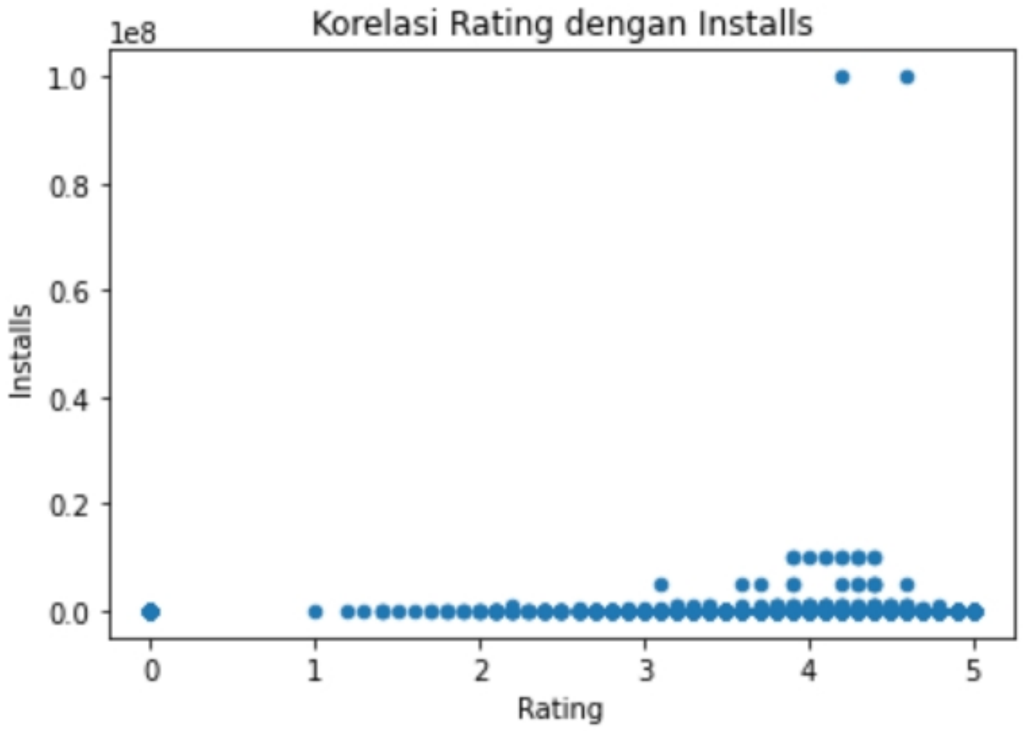
```
df.plot(kind="scatter",x="Rating",y="Rating Count",title="Korelasi Rating dengan Rating Count")
plt.show()
```



```
df["Rating"].corr(df["Installs"])
```

```
0.05142104556495695
```

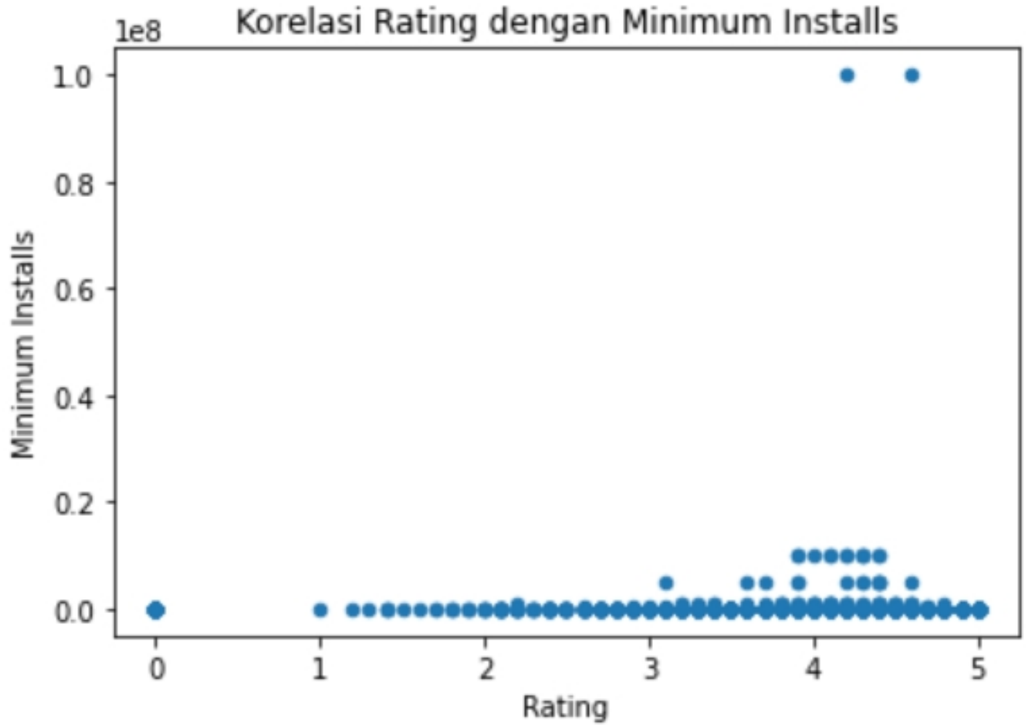
```
df.plot(kind="scatter",x="Rating",y="Installs",title="Korelasi Rating dengan Installs")
plt.show()
```



```
df["Rating"].corr(df["Minimum Installs"])
```

```
0.05142104556495695
```

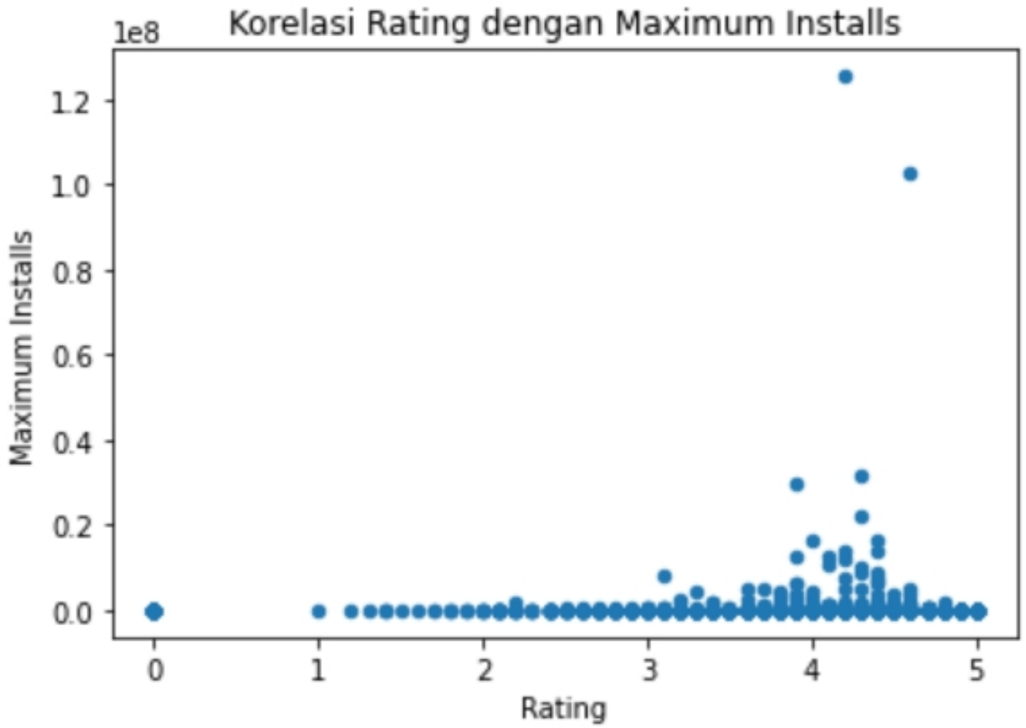
```
df.plot(kind="scatter",x="Rating",y="Minimum Installs",title="Korelasi Rating dengan Minimum Installs")
plt.show()
```



```
df["Rating"].corr(df["Maximum Installs"])
```

```
0.06195653433428596
```

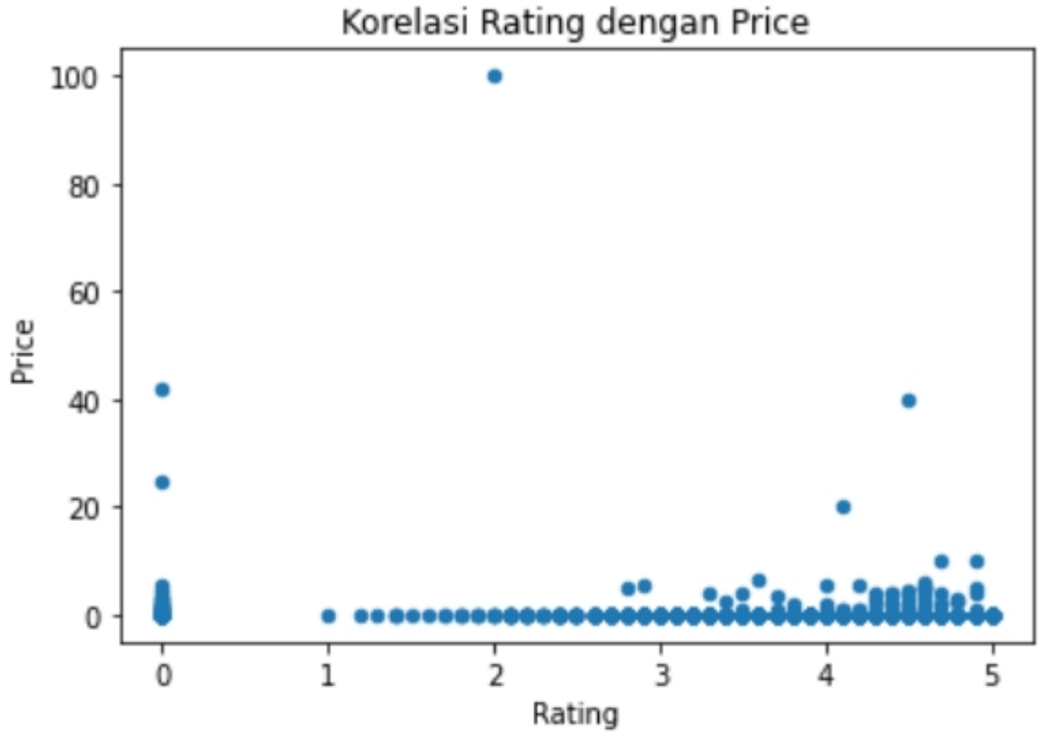
```
df.plot(kind="scatter",x="Rating",y="Maximum Installs",title="Korelasi Rating dengan Maximum Installs")
plt.show()
```



```
df["Rating"].corr(df["Price"])
```

```
0.014701998534002676
```

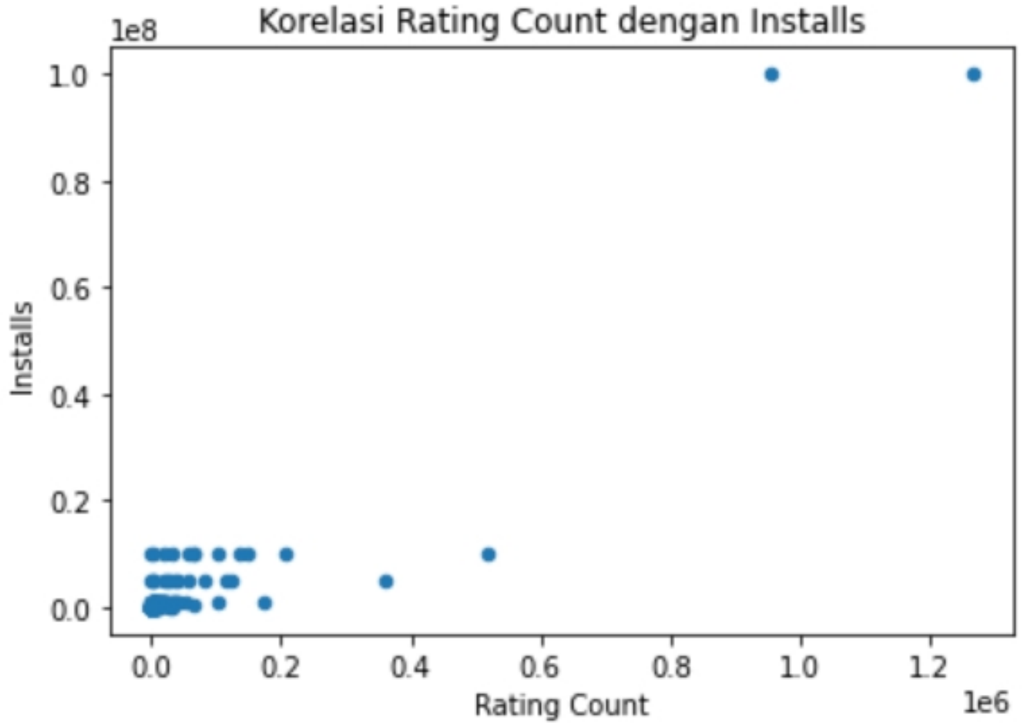
```
df.plot(kind="scatter",x="Rating",y="Price",title="Korelasi Rating dengan Price")
plt.show()
```



```
df["Rating Count"].corr(df["Installs"])
```

```
0.9279625309166728
```

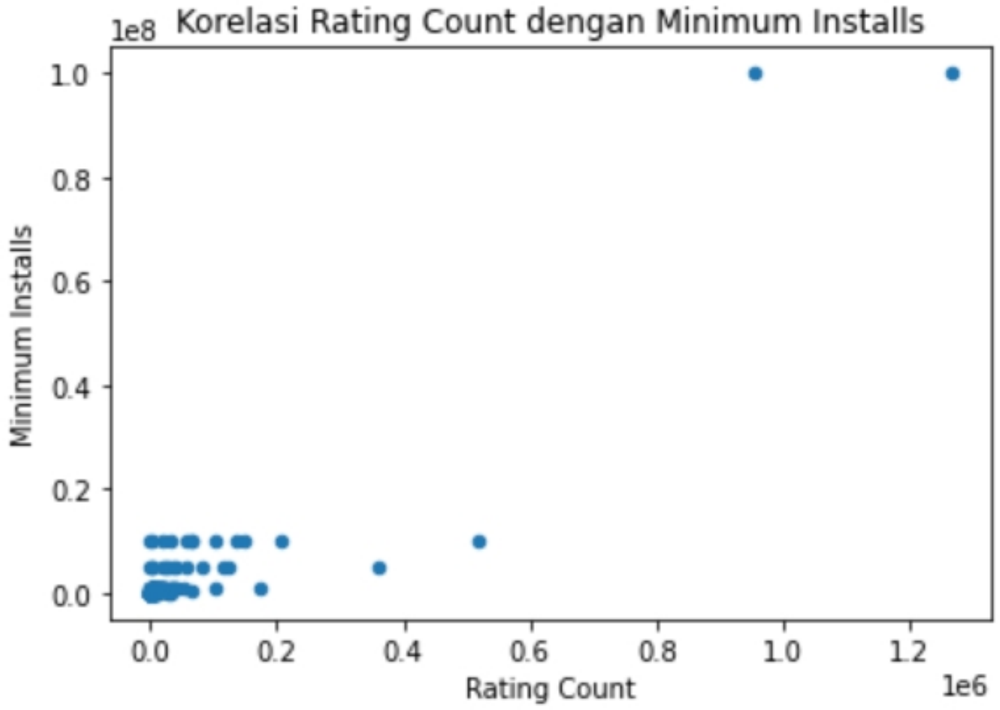
```
df.plot(kind="scatter",x="Rating Count",y="Installs",title="Korelasi Rating Count dengan Installs")
plt.show()
```



```
df["Rating Count"].corr(df["Minimum Installs"])
```

```
0.9279625309166728
```

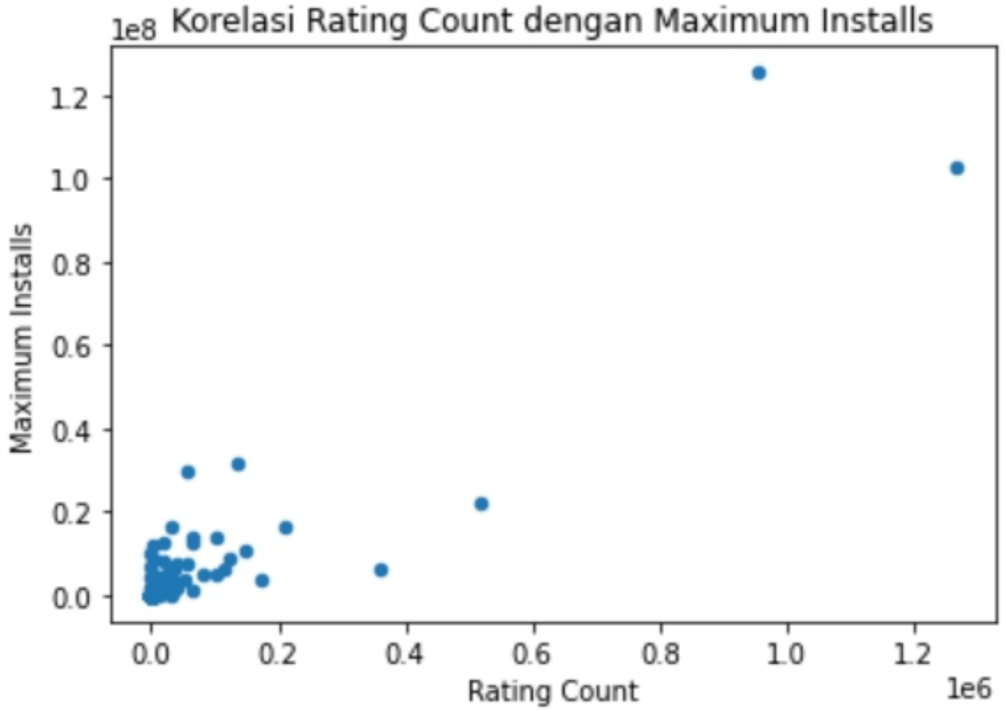
```
f.plot(kind="scatter",x="Rating Count",y="Minimum Installs",title="Korelasi Rating Count dengan Minimum Instal
lt.show()
```



```
df["Rating Count"].corr(df["Maximum Installs"])
```

```
0.9125109499475784
```

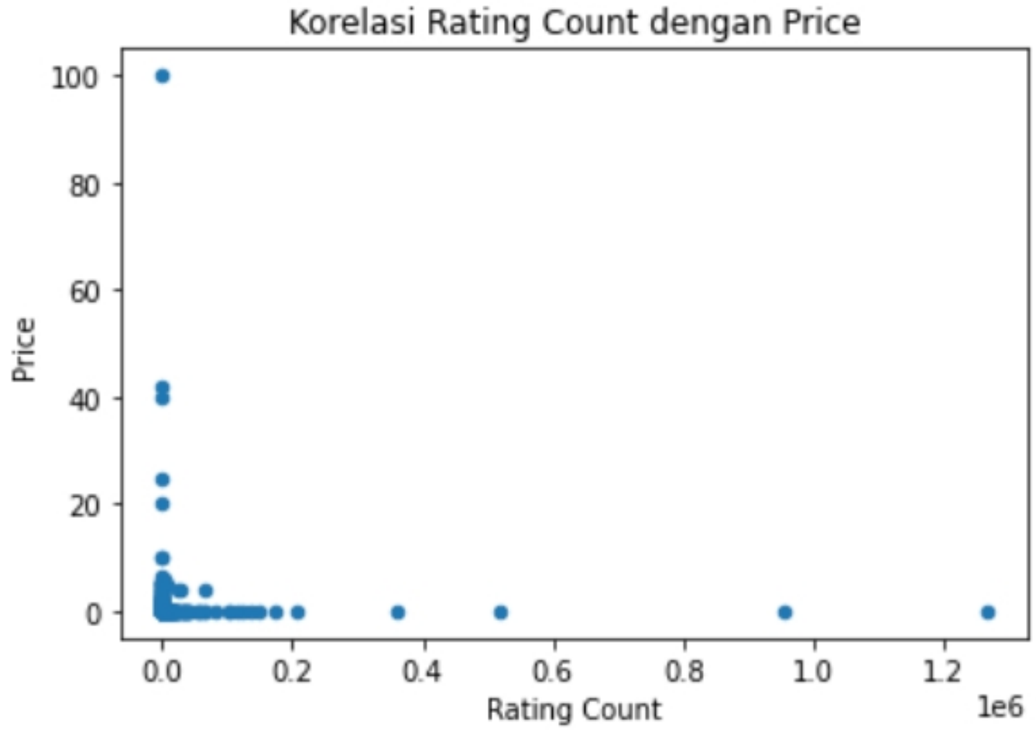
```
df.plot(kind="scatter",x="Rating Count",y="Maximum Installs",title="Korelasi Rating Count dengan Maximum Insta
plt.show()
```



```
df["Rating Count"].corr(df["Price"])
```

```
-0.0006582581022293049
```

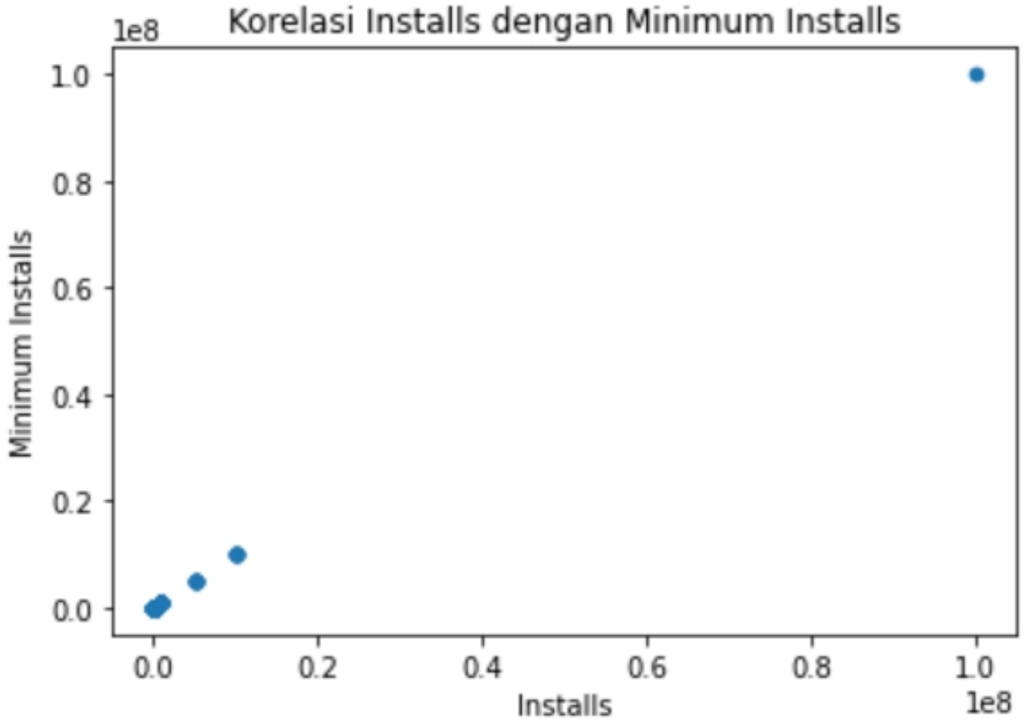
```
df.plot(kind="scatter",x="Rating Count",y="Price",title="Korelasi Rating Count dengan Price")
plt.show()
```



```
df["Installs"].corr(df["Minimum Installs"])
```

```
1.0
```

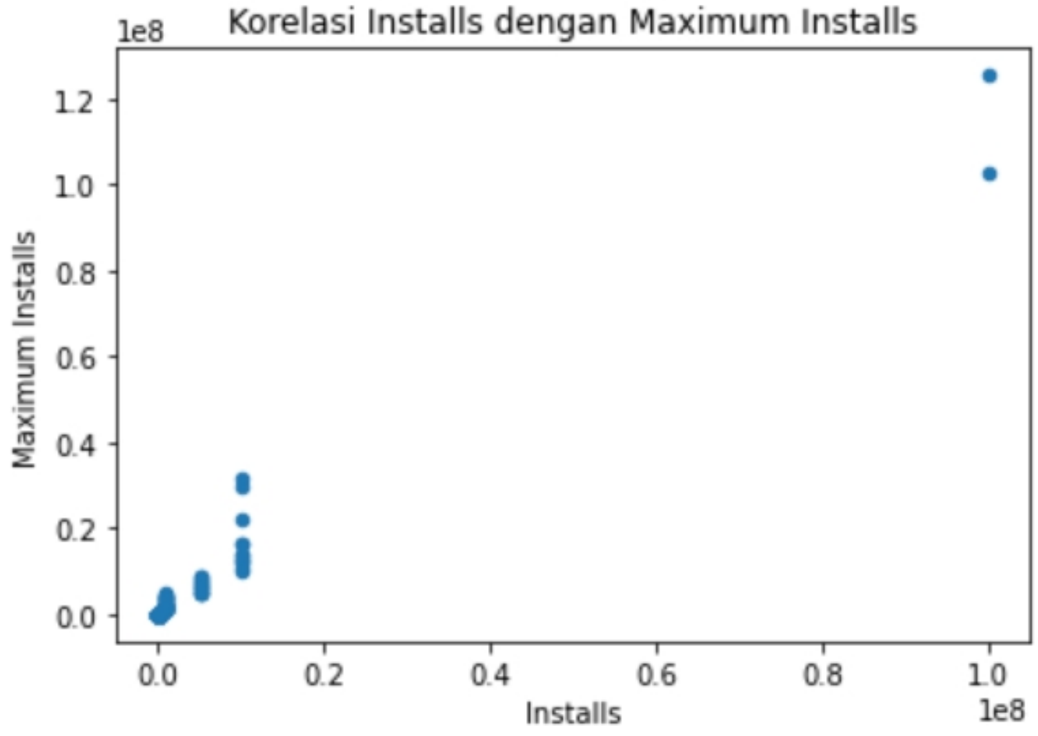
```
df.plot(kind="scatter",x="Installs",y="Minimum Installs",title="Korelasi Installs dengan Minimum Installs")
plt.show()
```



```
df["Installs"].corr(df["Maximum Installs"])
```

```
0.9786027288024505
```

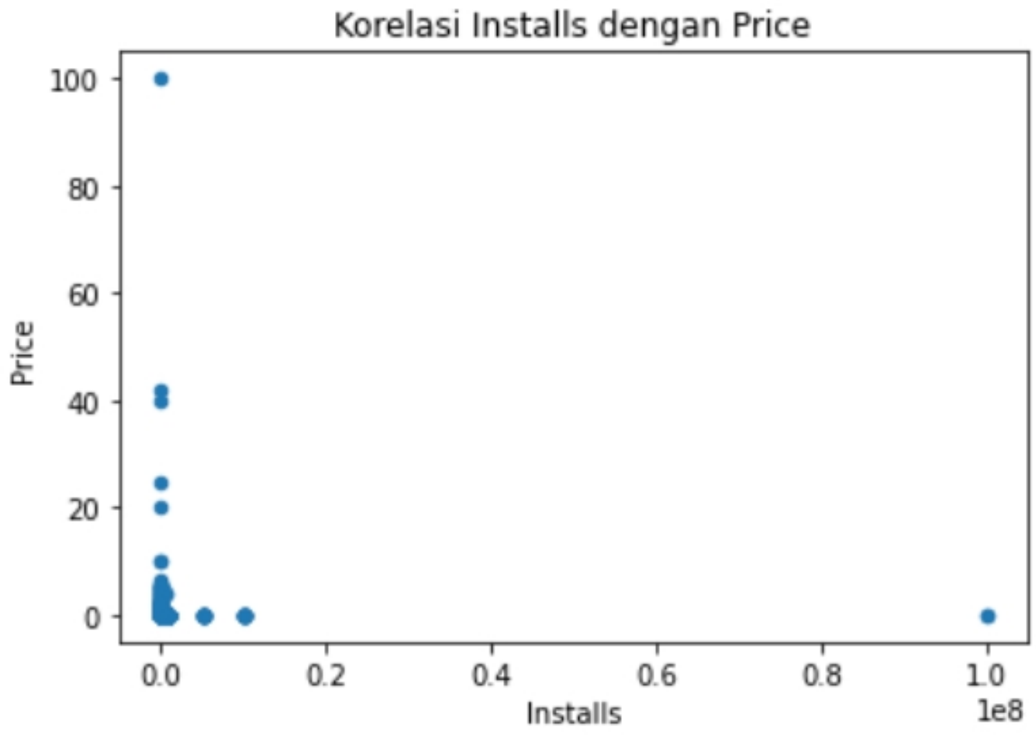
```
df.plot(kind="scatter",x="Installs",y="Maximum Installs",title="Korelasi Installs dengan Maximum Installs")
plt.show()
```



```
df["Installs"].corr(df["Price"])
```

```
-0.002706507233076278
```

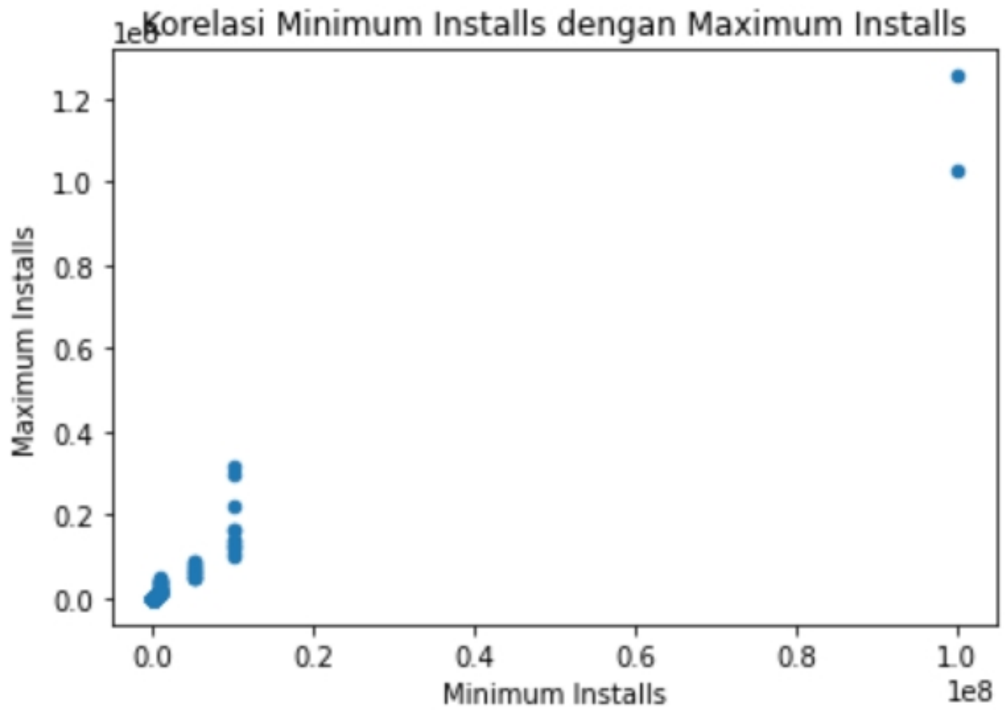
```
df.plot(kind="scatter",x="Installs",y="Price",title="Korelasi Installs dengan Price")
plt.show()
```



```
df["Minimum Installs"].corr(df["Maximum Installs"])
```

```
0.9786027288024505
```

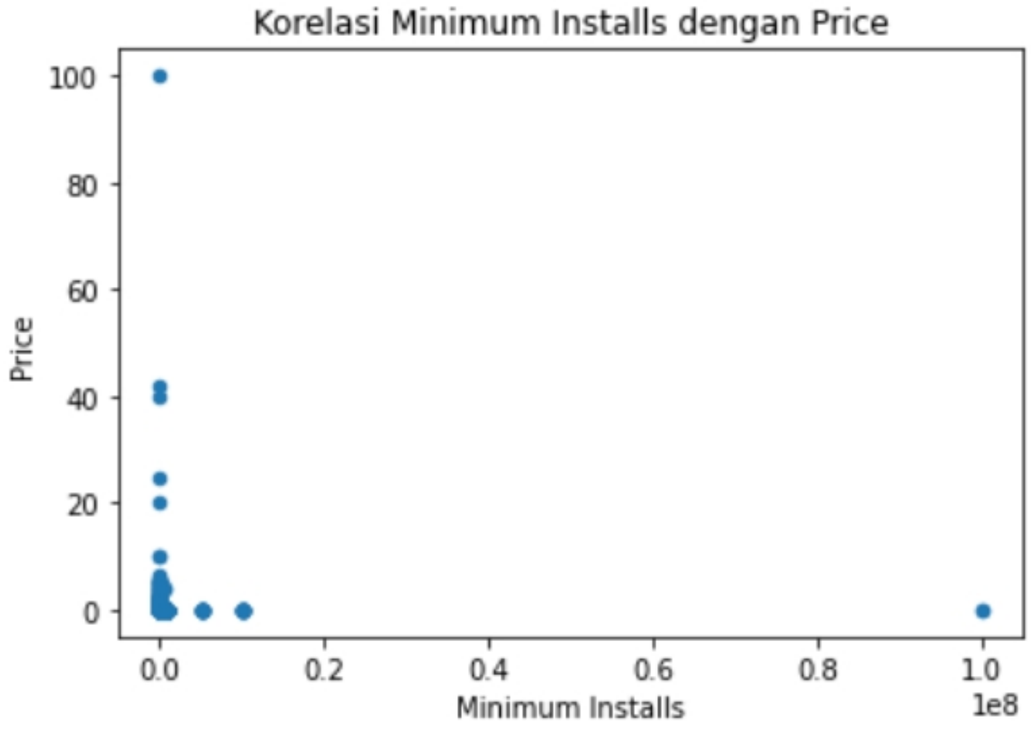
```
df.plot(kind="scatter",x="Minimum Installs",y="Maximum Installs",title="Korelasi Minimum Installs dengan Maximu
plt.show()
```



```
df["Minimum Installs"].corr(df["Price"])
```

```
-0.002706507233076278
```

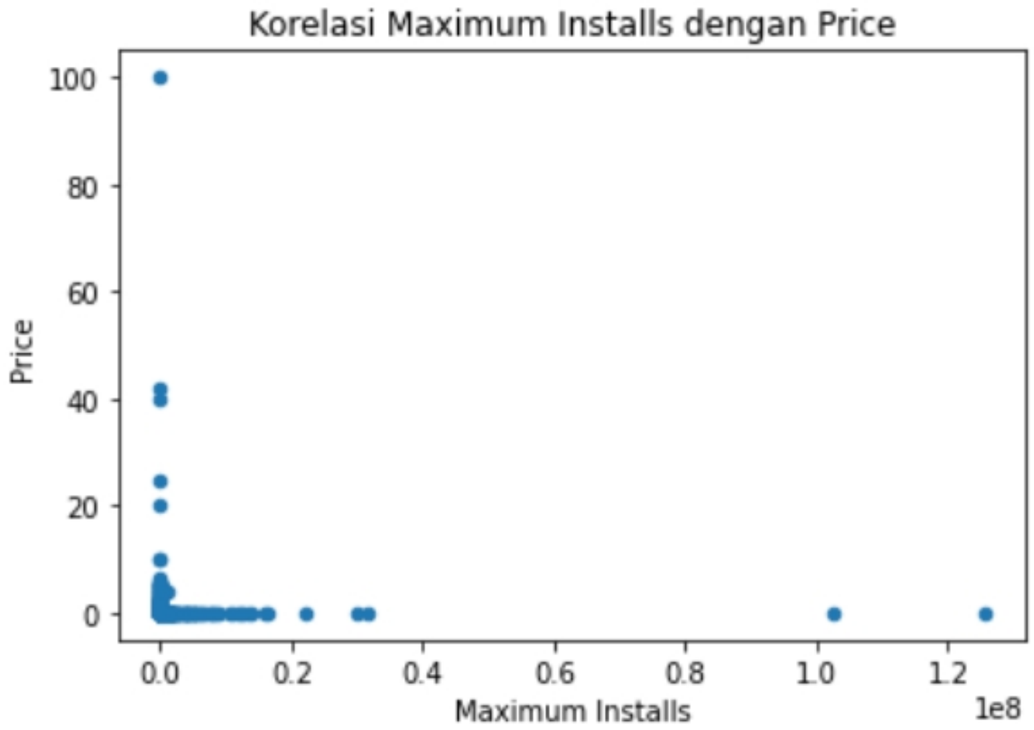
```
df.plot(kind="scatter",x="Minimum Installs",y="Price",title="Korelasi Minimum Installs dengan Price")
plt.show()
```



```
df["Maximum Installs"].corr(df["Price"])
```

```
-0.0032376605896525935
```

```
df.plot(kind="scatter",x="Maximum Installs",y="Price",title="Korelasi Maximum Installs dengan Price")
plt.show()
```



Data Cleansing

Data Cleansing and Re-Formatting

Menampilkan jumlah dan persentase data yang kotor

```
# CLEANSING DATA

import pandas as pd
df = pd.read_csv('/content/drive/MyDrive/data pengkom ITB/PlayStore Raw Dataset.csv')

# Nilai NaN
print('Banyak data yang kotor :')
for i in df.columns:
  if df[i].isnull().sum() > 0:
    print(i, '-->', df[i].isnull().sum(), 'data yang kotor')
print()

# Persen data kotor
print('Persen data yang kotor :')
for i in df.columns:
  if df[i].isnull().sum() > 0:
    print(i, '-->', str(df[i].isnull().sum()/len(df.index)*100) + '%')
```

Keluaran

```
Banyak data yang kotor :
Rating --> 38 data yang kotor
Rating Count --> 38 data yang kotor
Minimum Android --> 10 data yang kotor
Developer Website --> 1337 data yang kotor
Released --> 121 data yang kotor
Privacy Policy --> 723 data yang kotor

Persen data yang kotor :
Rating --> 0.95%
Rating Count --> 0.95%
Minimum Android --> 0.25%
Developer Website --> 33.425%
Released --> 3.025%
Privacy Policy --> 18.075%
```

Melakukan penanganan terhadap nilai NaN (numpy.nan)

```
# Mencari kolom yang memiliki nilai NaN
# column with_nan = [i for i in df.columns if df[i].isnull().sum() > 0]
   # keluaran : ['Rating', 'Rating Count', 'Minimum Android', 'Developer Website', 'Released', 'Privacy Policy']
column_to_replace_with_notavailable = ['Minimum Android', 'Developer Website', 'Released', 'Privacy Policy']
column_to_replace_with_0 = ['Rating Count']
column_to_be_left_as_is = ['Rating'] # dibiarkan NaN

# Mengganti NaN dengan nilai 0, data type numpy.int64
for i in column_to_replace_with_0:
  df[i].replace(np.nan, np.int64(0), inplace=True)

# Mengganti NaN dengan 'Tidak Tersedia'
for i in column_to_replace_with_notavailable:
  df[i].replace(np.nan, 'Tidak Tersedia', inplace=True)
```

Melakukan re-formatting pada kolom 'Released' dan 'Installs'

```
# Mem-format kembali kolom 'Released' menjadi tahun saja
for i in df.index:
  if df.loc[i, "Released"] != 'Tidak Tersedia':
    x = df.loc[i, "Released"].split(' ')
    df.loc[i, "Released"] = x[2] # format awal yaitu mmm dd, yyyy

# Mem-format kembali kolom 'Installs' dengan menghilangkan tanda koma(,) dan lebih(+)
# lalu mengubahnya dari str menjadi numpy.int64
for i in df.index:
  # Format awal : xxx,xxx+
  # Agar menjadi : xxxxxx
  x = df.loc[i, "Installs"]
  x = x.replace('+', '')
  x = x.replace(',', '')
  df.loc[i, "Installs"] = np.int64(x) # Value menjadi int
```

Melakukan re-formatting pada kolom 'Minimum Android'

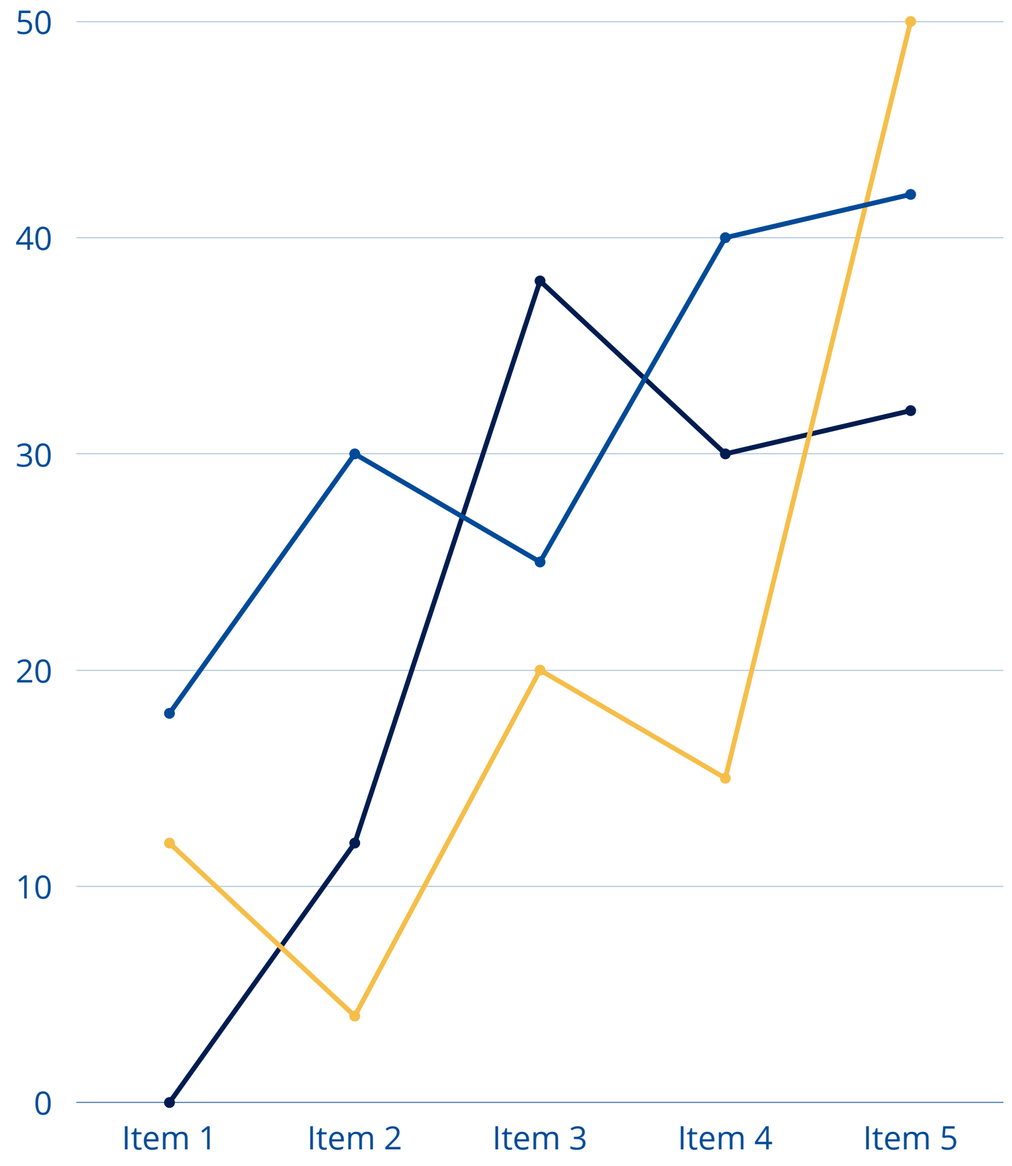
```
# Minimum Android
for i in df.index:
  if df.loc[i, 'Minimum Android'] == '4.0.3 - 7.1.1':
    df.loc[i, 'Minimum Android'] = '4.0.3 and up'
  elif df.loc[i, 'Minimum Android'] == '4.4W and up':
    df.loc[i, 'Minimum Android'] = '4.4 and up'
```

Melakukan re-formatting pada kolom 'Size'

```
# size
Batas_Size = [i*10**8/2 for i in range(20,-1, -1)]
for i in df.index:
  # Mem-format kembali variabel menjadi bilangan bulat
  df.loc[i, 'Size'] = df.loc[i, 'Size'].replace(',', '')
  if 'k' in df.loc[i, 'Size']:
    df.loc[i, 'Size'] = np.int64(np.float(df.loc[i, 'Size'].replace('k', '')))*1000
  elif 'M' in df.loc[i, 'Size']:
    df.loc[i, 'Size'] = np.int64(np.float(df.loc[i, 'Size'].replace('M', '')))*1000000

  # Mengkelaskan kembali nilai variabel berdasarkan batas yang dibuat di Batas_Size
  if type(df.loc[i, 'Size']) == np.int64:
    j = 0
    while (df.loc[i, 'Size'] <= Batas_Size[j]):
      j+=1

    df.loc[i, 'Size'] = str(Batas_Size[j]) + ' - ' + str(Batas_Size[j-1])
```

Melakukan re-formatting pada kolom 'Size'

```
# size
Batas_Size = [i*10**8/2 for i in range(20,-1, -1)]
for i in df.index:
  # Mem-format kembali variabel menjadi bilangan bulat
  df.loc[i, 'Size'] = df.loc[i, 'Size'].replace(',', '')
  if 'k' in df.loc[i, 'Size']:
    df.loc[i, 'Size'] = np.int64(np.float(df.loc[i, 'Size'].replace('k', '')))*1000
  elif 'M' in df.loc[i, 'Size']:
    df.loc[i, 'Size'] = np.int64(np.float(df.loc[i, 'Size'].replace('M', '')))*1000000

  # Mengkelaskan kembali nilai variabel berdasarkan batas yang dibuat di Batas_Size
  if type(df.loc[i, 'Size']) == np.int64:
    j = 0
    while (df.loc[i, 'Size'] <= Batas_Size[j]):
      j+=1

    df.loc[i, 'Size'] = str(Batas_Size[j]) + ' - ' + str(Batas_Size[j-1])
```



Kesimpulan

Kesimpulan yang dapat diambil dari tugas besar ini yaitu analisis sangat diperlukan dalam mengerti sebuah data. Analisis data dapat memberikan persepsi dan pemahaman secara cepat terhadap karakteristik atau elemen yang ada dalam data. Analisis data juga dapat memperlihatkan masalah-masalah dan data-data yang tidak sesuai (tidak masuk akal/kosong).

Terimakasih Banyak

Selesai