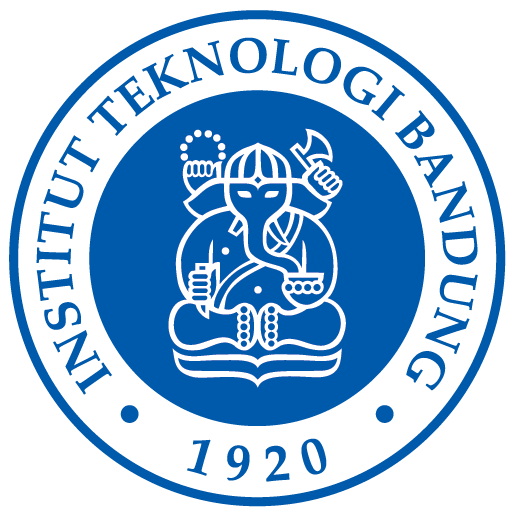
**LAPORAN TUGAS BESAR PENGENALAN KOMPUTASI**

**KELAS KU1102-15**

**ANALISIS DATA GOOGLE PLAYSTORE ANDROID APP JUNI 2021**



**KELOMPOK 9 :**

**MARIA SOPHIE DEWANTO (16521321)**

**SALOMO REINHART GREGORY MANALU (16521331)**

**MUHAMMAD RIDZKY SYAFTIRA (16521341)**

**FAKIH ANUGERAH PRATAMA (16521351)**

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA**

**INSTITUT TEKNOLOGI BANDUNG**

**2021**

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc86914095)

[PEMBAGIAN TUGAS 3](#_Toc86914096)

[LAPORAN INTI 4](#_Toc86914097)

[1. Tugas 3 : Deskripsi Data dan File 4](#_Toc86914098)

[2. Tugas 4 : Karakteristik 5](#_Toc86914099)

[3. Tugas 5 : Statistik 7](#_Toc86914100)

[4. Tugas 6 : Visualisasi 12](#_Toc86914100)

[5. Tugas 7 : Korelasi 13](#_Toc86914100)

[6. Tugas 8 : Cleansing Data 19](#_Toc86914100)

[KESIMPULAN 24](#_Toc86914101)

[LESSON LEARNED 24](#_Toc86914102)

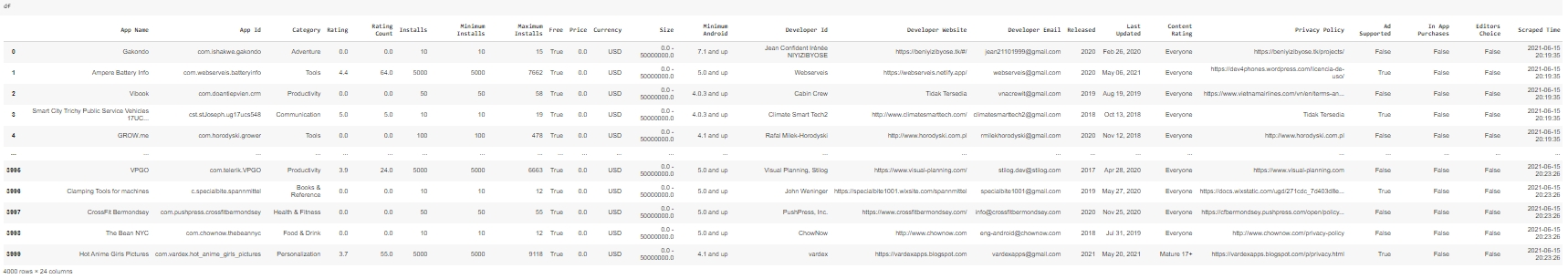
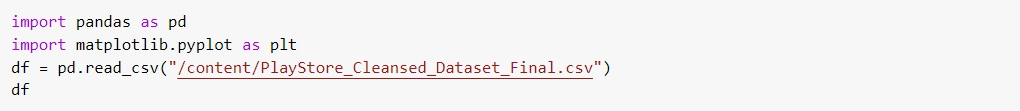
# PEMBAGIAN TUGAS

|  |  |
| --- | --- |
| Maria Sophie Dewanto | * Menentukan data yang akan dianalisis * Membuat deskripsi data (tugas 3) * Membuat korelasi data (tugas 7) * Menyusun PPT |
| Salomo Reinhart Gregory Manalu | * Menentukan data yang akan dianalisis * Membuat laporan * Menyusun PPT |
| Muhammad Ridzky Syaftira | * Menentukan data yang akan dianalisis * Membuat karakteristik data (tugas 4) * Membuat visualisasi data (tugas 6) * Menyusun PPT |
| Fakih Anugerah Pratama | * Menentukan data yang akan dianalisis * Membuat statistik data (tugas 5) * Membuat data cleansing (tugas 8) * Menyusun PPT |

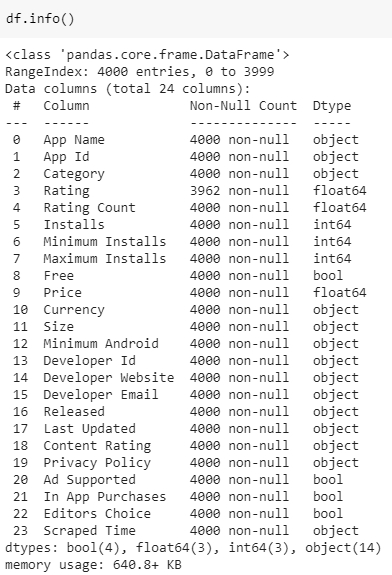
**LAPORAN INTI**

1. **Tugas 3 (Deskripsi Data dan File)**

Data yang kita analisis merupakan Google PlayStore Android App Data (2.3 Million+ App Data) yang diambil pada bulan Juni 2021.

Data ini memiliki format .csv dimana dapat diakses dengan pandas menggunakan code :

Data ini berisikan:



Penjelasan kolom:

1. App Name : nama dari aplikasi
2. App Id : Package name
3. Category : Kategori aplikasi
4. Rating : Average rating
5. Rating Count: Number of rating
6. Installs : Approximate install count
7. Minimum Installs: Approximate minimum app install count
8. Maximum installs: Approximate maximum app install count
9. Free : Whether the app is free or paid
10. Price: App price
11. Currency: App currency
12. Size: Size of application package
13. Minimum Android: Minimum android version supported
14. Developer id: Developer Id in Google Playstore
15. Developer Website: Website of the developer
16. Developer Email: id email pengembang
17. Released: Tanggal peluncuran aplikasi di Google Playstore
18. Last updated: Tanggal pembaruan aplikasi terakhir
19. Content Rating: Tingkat kematangan aplikasi
20. Privacy Policy: Kebijakan privasi dari pengembang
21. Ad Supported: Dukungan iklan di aplikasi
22. In App Purchases: Pembelian dalam Aplikasi
23. Editors Choice: Apakah dinilai sebagai Pilihan Editor.
24. Scraped Time: Tanggal-waktu dalam GMT

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

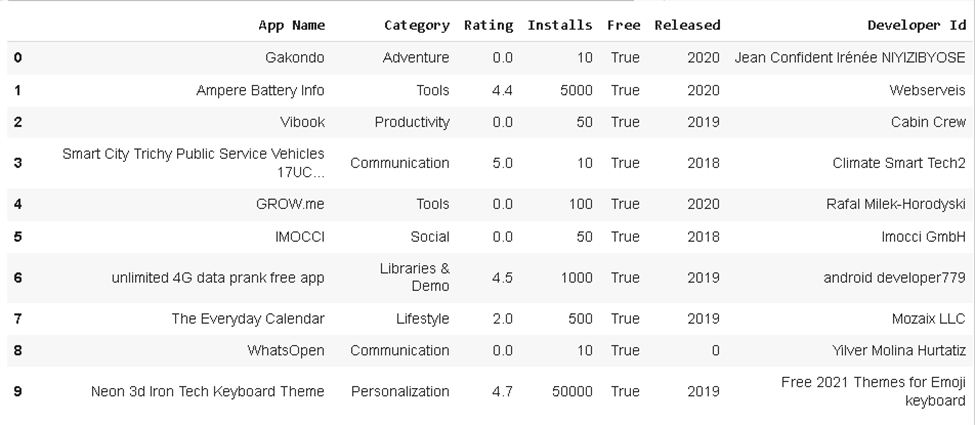
Data terdiri dari 4000 baris dan 24 kolom dengan ukuran file data 640.8+ KB.

1. **Tugas 4 (Karakteristik 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 | Kuantitatif - Discrete | Harga Aplikasi | Range value: 0.0 – 99.99 | df.min()[“Price”]  df.max()[“Price”] |
| 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() |
| 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 Privasi 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 | Penghargaan Pilihan *Editor’s* | Terdiri dari 2 value: False (3998), True (2) | df[“Editors Choice”].value\_counts() |
| Scraped Time | Kategorikal - Nominal | Tanggal-Waktu dalam GMT | Terdiri dari 232 value | df[“Scraped Time].value\_counts() |

1. **Tugas 5 (Statistik)**
   1. Mengambil nilai 10 data awal dengan kolom ['App Name', 'Category', 'Rating', 'Installs', 'Free', 'Released', 'Developer Id'] .

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



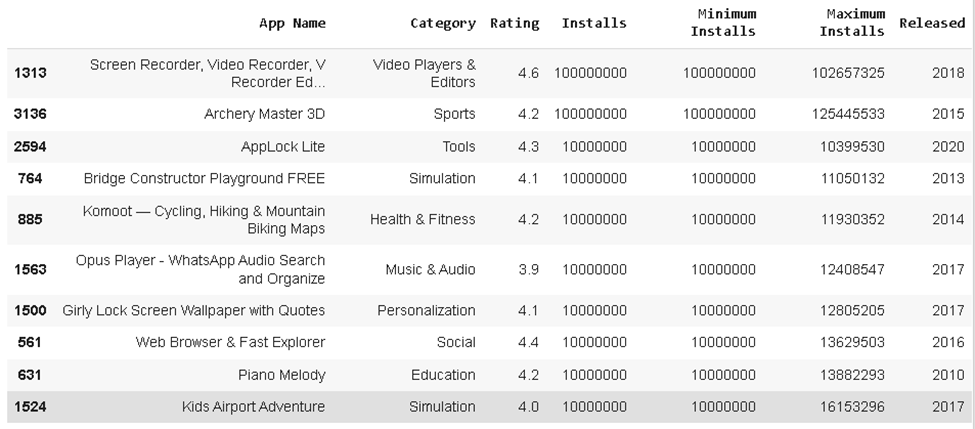
* 1. Mengambil nilai 10 data tertinggi dari data yang telah diurutkan berdasar jumlah install dari terbesar ke yang terkecil dengan kolom ['App Name', 'Category', 'Rating', 'Installs', 'Free', 'Released', 'Developer Id']

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



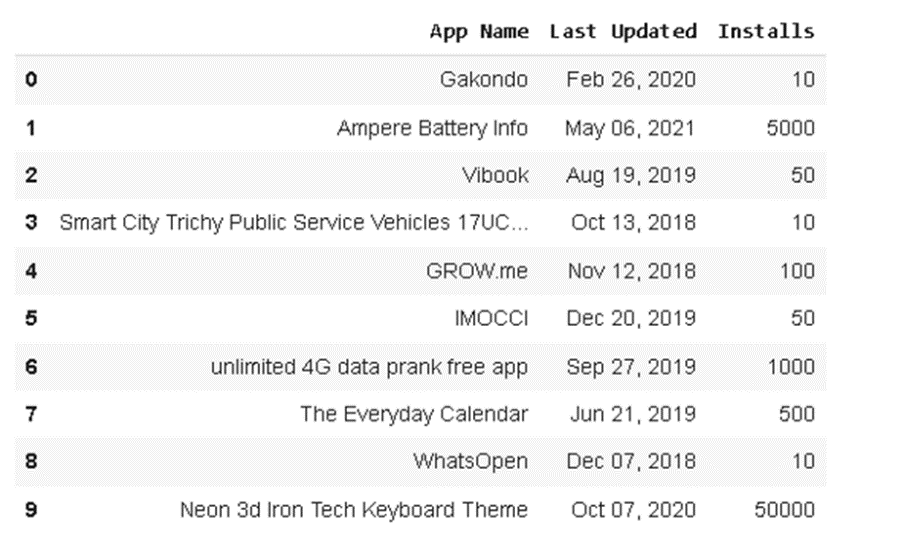
* 1. Mengambil nilai 10 data tertinggi dari data yang telah diurutkan berdasar jumlah minimum install dari terbesar ke yang terkecil dan maksimum install dari terkecil ke terbesar dengan kolom ['App Name', 'Category', 'Rating', 'Installs', 'Free', 'Released', 'Developer Id']

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



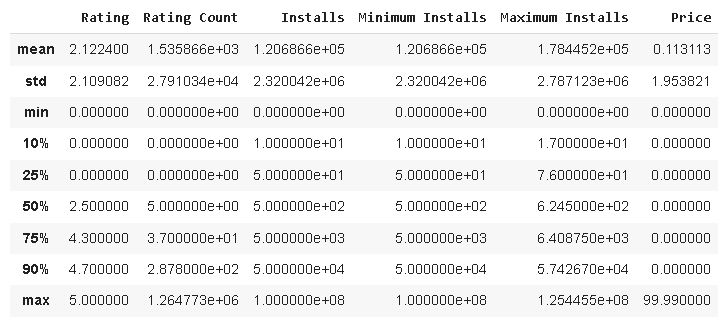
* 1. Menampilkan 10 data awal dengan jumlah install kurang dari 100000 dengan kolom ['App Name', 'Last Updated', 'Installs']

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



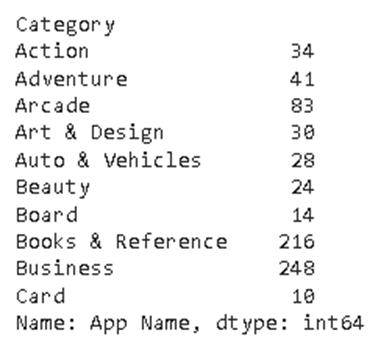
* 1. Menampilkan rata-rata dan standar deviasi, persentil, serta nilai maksimum dan minimum dari data-data numerical

df.describe(percentiles=[.1, .25, .5, .75, .9])[1:] # tidak mengikutkan count



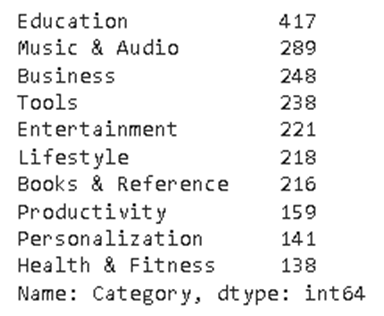
* 1. Menampilkan 10 data awal jumlah app yang telah dikelompokkan berdasarkan kategori menggunakan .groupby() dan .count()

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

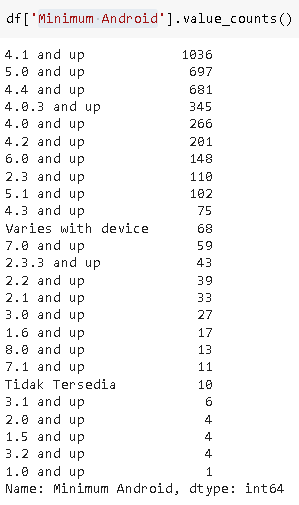


* 1. Menampilkan persebaran data 'Category' menggunakan value\_counts() df['Category'].value\_counts()[:10]

df['Category'].value\_counts()[:10]



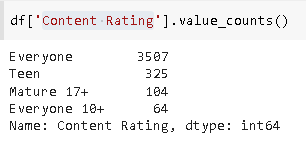
* 1. Menampilkan persebaran data 'Minimum Android' menggunakan value\_counts()



* 1. Menampilkan persebaran data 'Released' menggunakan value\_counts()



* 1. Menampilkan persebaran data 'Content Rating' menggunakan value\_counts()



* 1. Pengetahuan yang didapat
     1. 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). Pengurutan data sangatlah bermanfaat saat kita ingin mengetahui seberapa besar range data dan perbandingan besar nilai-nilai tertentu yang berdekatan, sehingga memudahkan proses analisis data secara keseluruhan. Sebagai contoh didapat bahwa ada 2 anggota dataset (app Google PlayStore) yang perkiraan jumlah install-nya lebih tinggi dari data yang lain.
     2. Sebagai contoh didapat bahwa ada 2 anggota dataset (app Google PlayStore) yang perkiraan jumlah install-nya lebih tinggi dari data yang lain.

Sebagai contoh, didapat persentil 10% sampai 90% data 'Price' bernilai 0, yang menandakan lebih dari 90% keseluruhan anggota data memiliki harga 0 atau gratis.

* + 1. 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. Mengetahui persebaran data juga sangatlah penting dalam melakukan analisis data.

1. **Tugas 6 (Visualisasi)**
   1. Grafik Perbandingan

Chart

Description automatically generated with low confidence

Insight: Berdasarkan grafik tersebut, kita dapat mengetahui kategori aplikasi berbayar yang memiliki penghasilan dari yang terkecil hingga terbesar.

Chart

Description automatically generated

Insight: Berdasarkan grafik diatas, kita dapat mengetahui rata-rata penilaian untuk setiap kategori kedewasaan.

* 1. Grafik Perubahan terhadap Waktu

Chart, line chart

Description automatically generated

Insight: Berdasarkan grafik tersebut, kita dapat mengetahui banyak pengunduhan tiap tahun keluarannya aplikasi.

Chart

Description automatically generated

Insight: Berdasarkan data diatas, kita dapat mengetahui banyaknya penilaian yang didapatkan tiap tahun pengeluaran aplikasi.

* 1. Grafik Hierarki

Graphical user interface, text, application

Description automatically generated

Insight: Berdasarkan diagram diatas, kita dapat mengetahui persentase banyaknya penilaian setiap kategori kedewasaan.

Graphical user interface, text, application, email

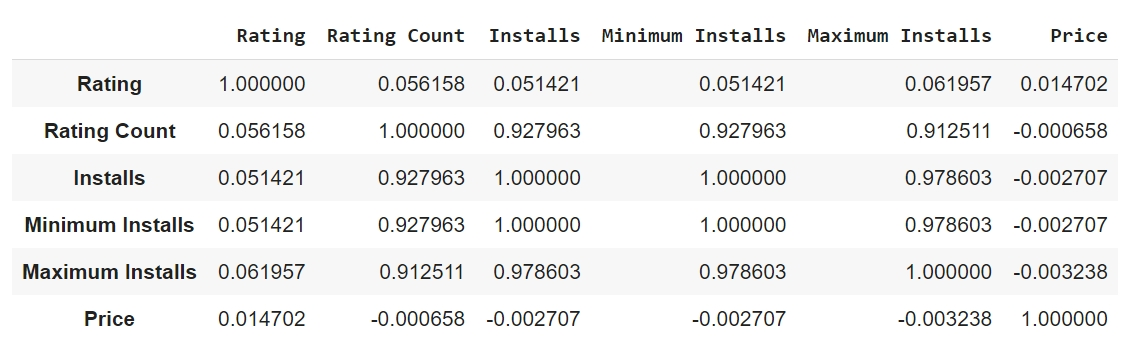
Description automatically generated

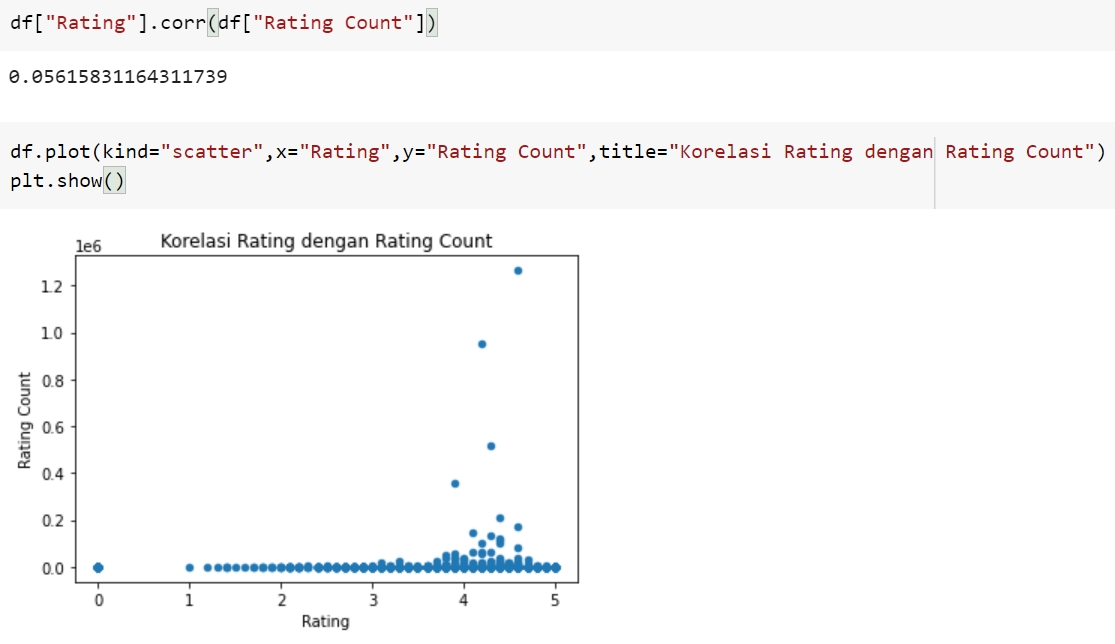
Insight: Berdasarkan diagram diatas, kita dapat mengetahui perbandingan total harga antara aplikasi dengan konten berbayar dan aplikasi tanpa konten berbayar.

Plotting relationship: dapat dilihat pada bagian korelasi

1. **Tugas 7 (Korelasi)**

Data ini memiliki 6 atribut kuantitatif yaitu pada bagian rating, rating count, installs, minimum installs, maximum installs, dan price. Berikut adalah korelasi antara setiap atribut kuantitatif tersebut :

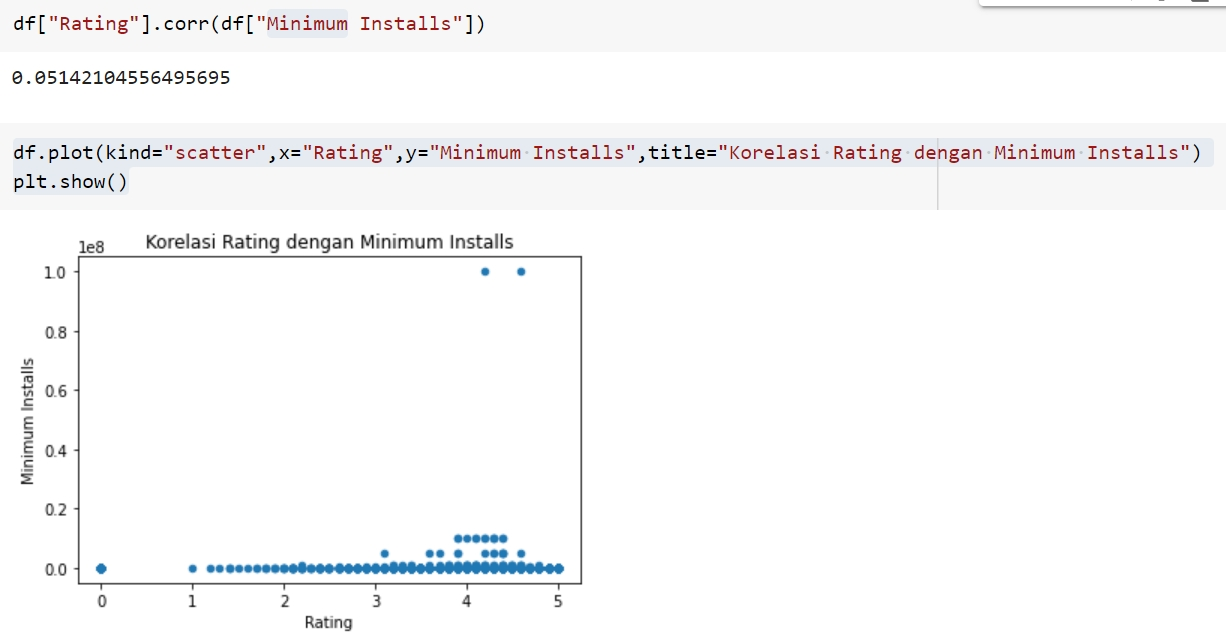


* 1. Koefisien korelasi ( r ) antara Rating dengan Rating accountr mendekati 0 sehingga tidak ada hubungan antara kedua variable
  2. Koefisien korelasi ( r ) antara Rating dengan Installs

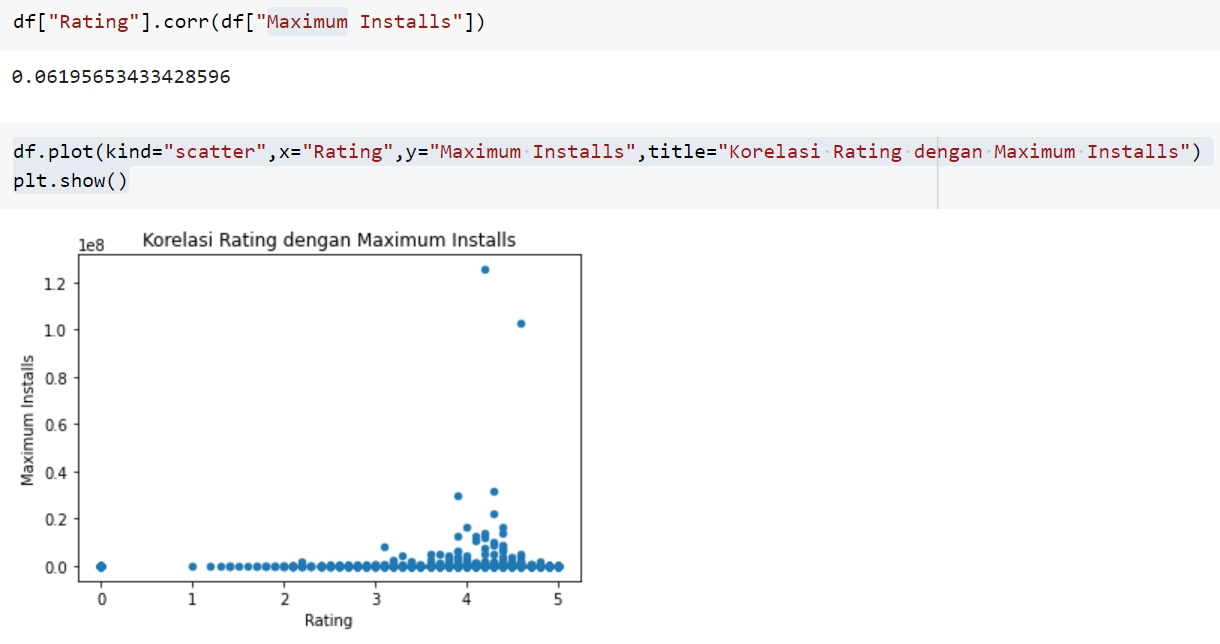


r mendekati 0 sehingga tidak ada hubungan antara kedua variabel

* 1. Koefisien korelasi ( r ) antara Rating dengan Minimum Installs

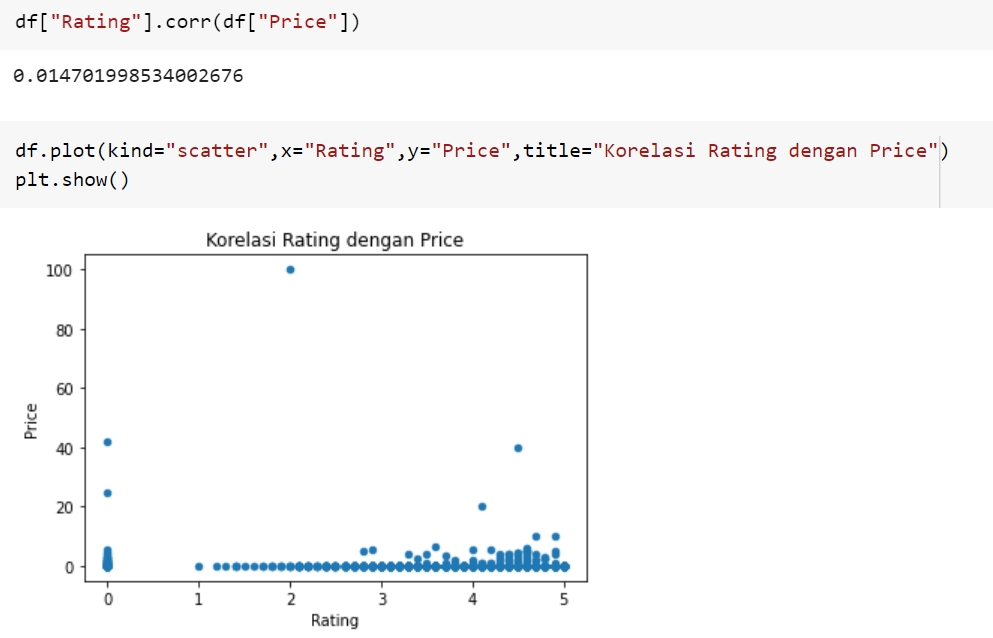
r mendekati 0 sehingga tidak ada hubungan antara kedua variabel

* 1. Koefisien korelasi ( r ) antara Rating dengan Maximum Installs



r mendekati 0 sehingga tidak ada hubungan antara kedua variabel

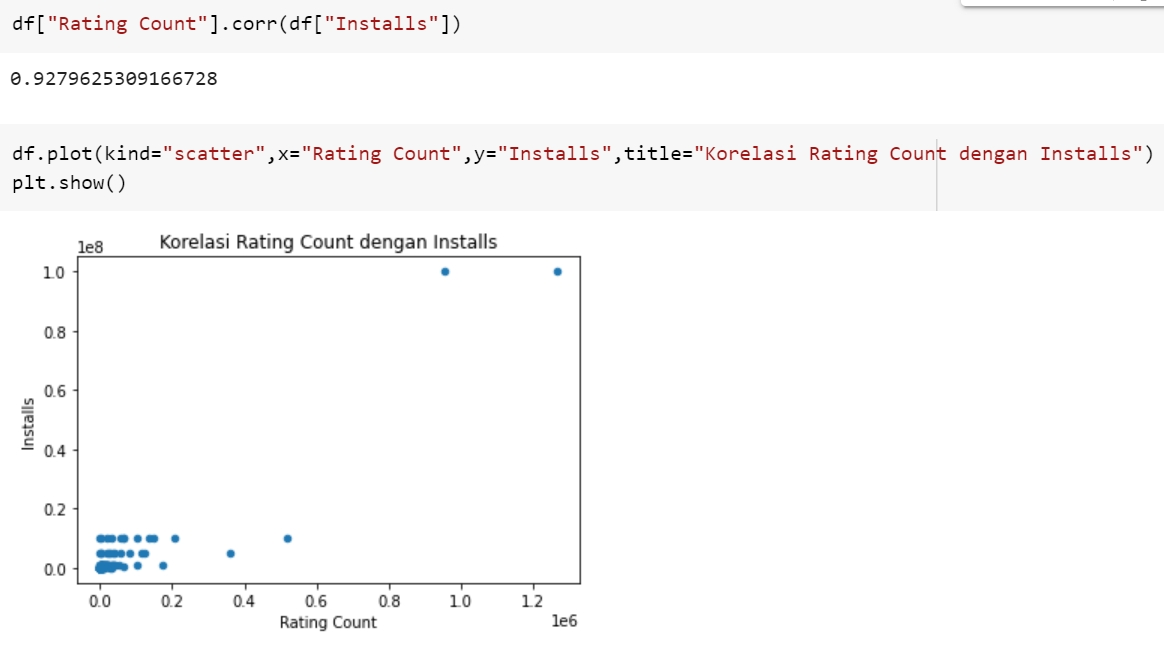
* 1. Koefisien korelasi ( r ) antara Rating dengan Price



No correlation

r mendekati 0 sehingga tidak ada hubungan antara kedua variabel

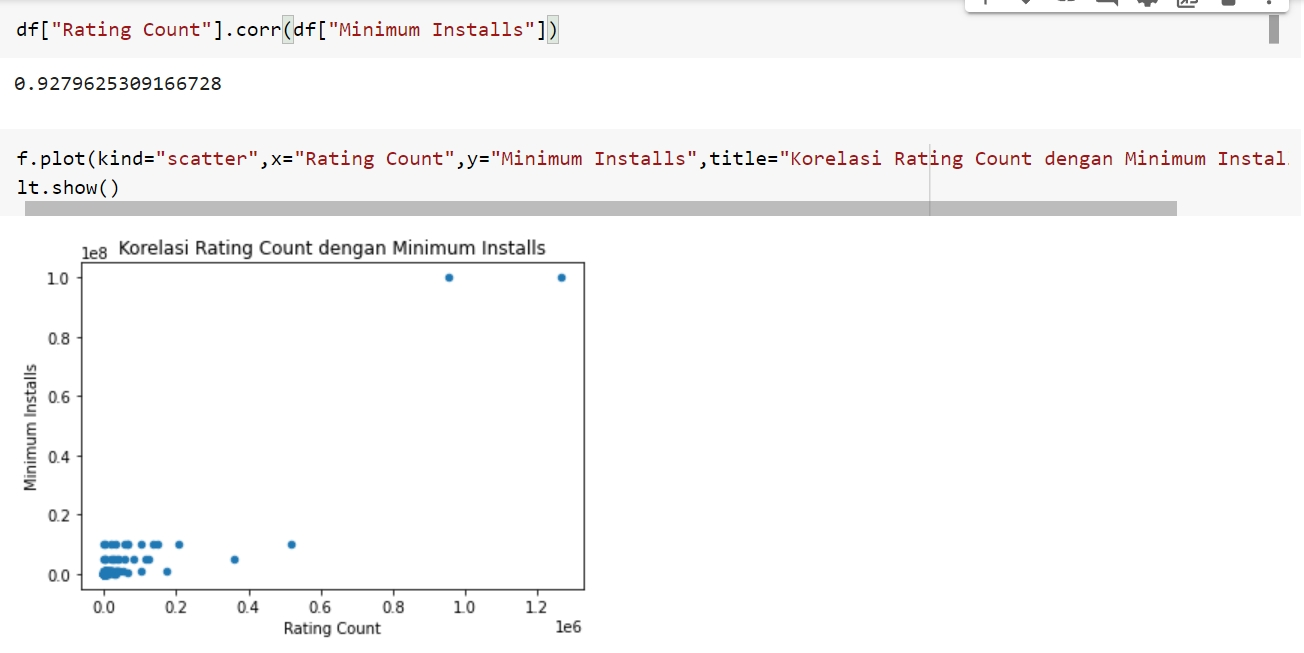
* 1. Koefisien korelasi ( r ) antara Rating count dengan Installs



High positive correlation

r positif sehingga apabila satu variabel membesar variabel lain juga akan membesar

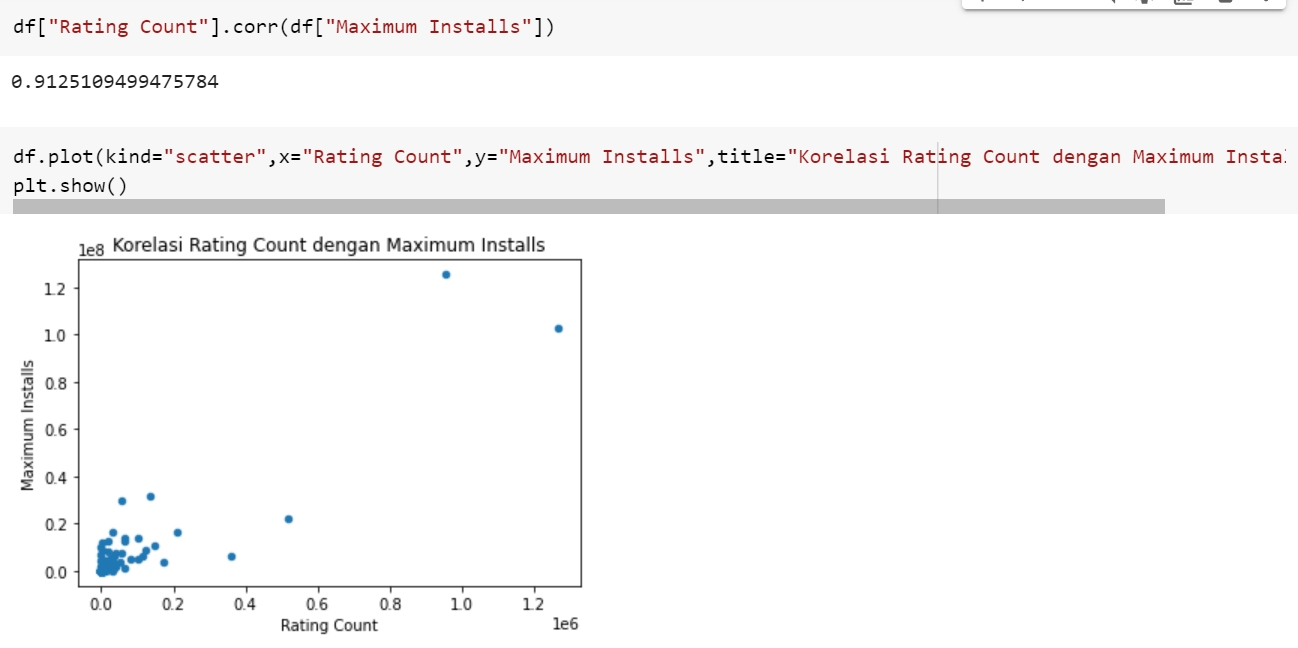
* 1. Koefisien korelasi ( r ) antara Rating count dengan Minimum Installs



High positive correlation

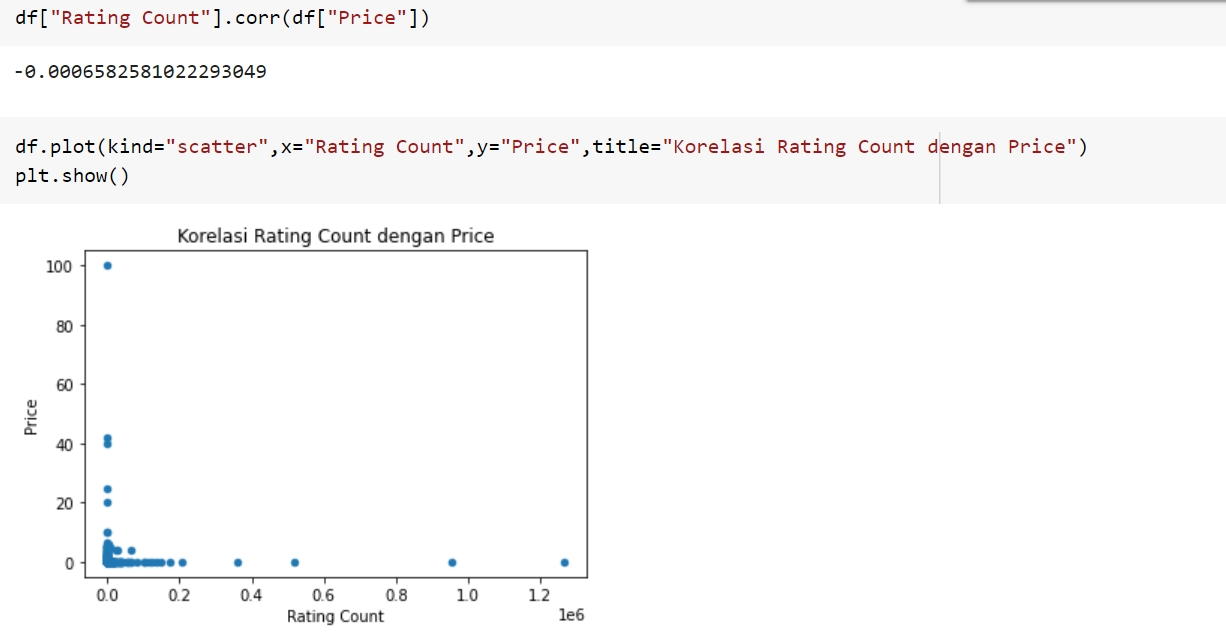
r positif sehingga apabila satu variabel membesar variabel lain juga akan membesar

* 1. Koefisien korelasi ( r ) antara Rating count dengan Maximum Installs



High positive correlation

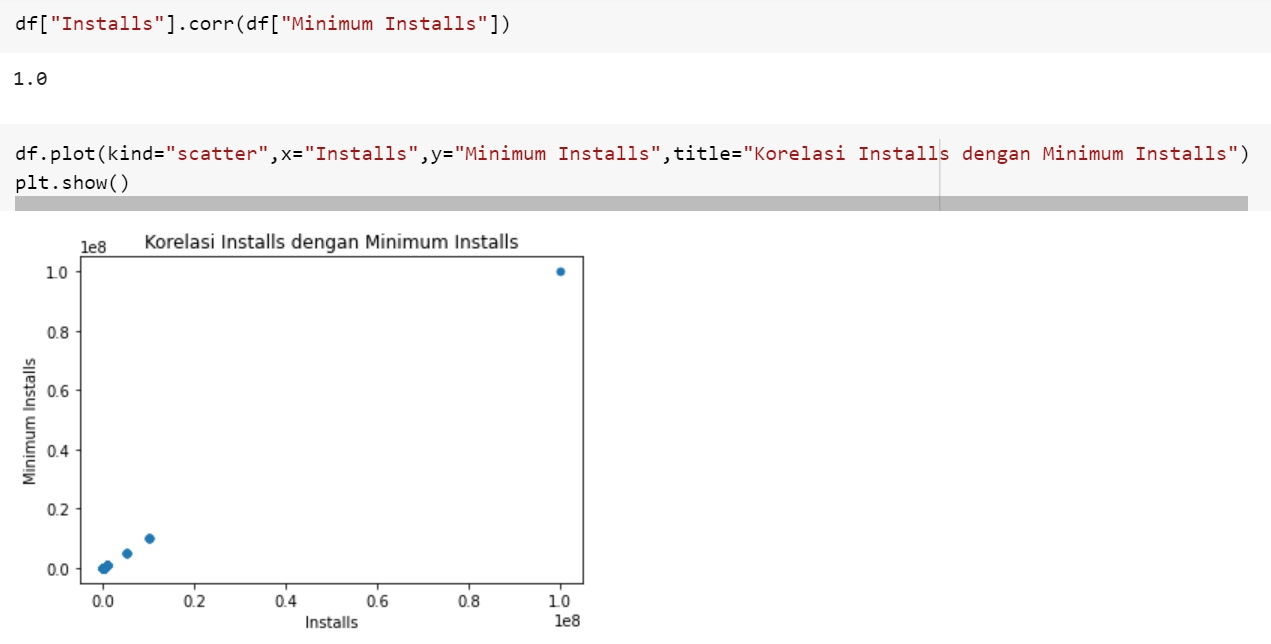
* 1. Koefisien korelasi ( r ) antara Rating count dengan Price



No correlation

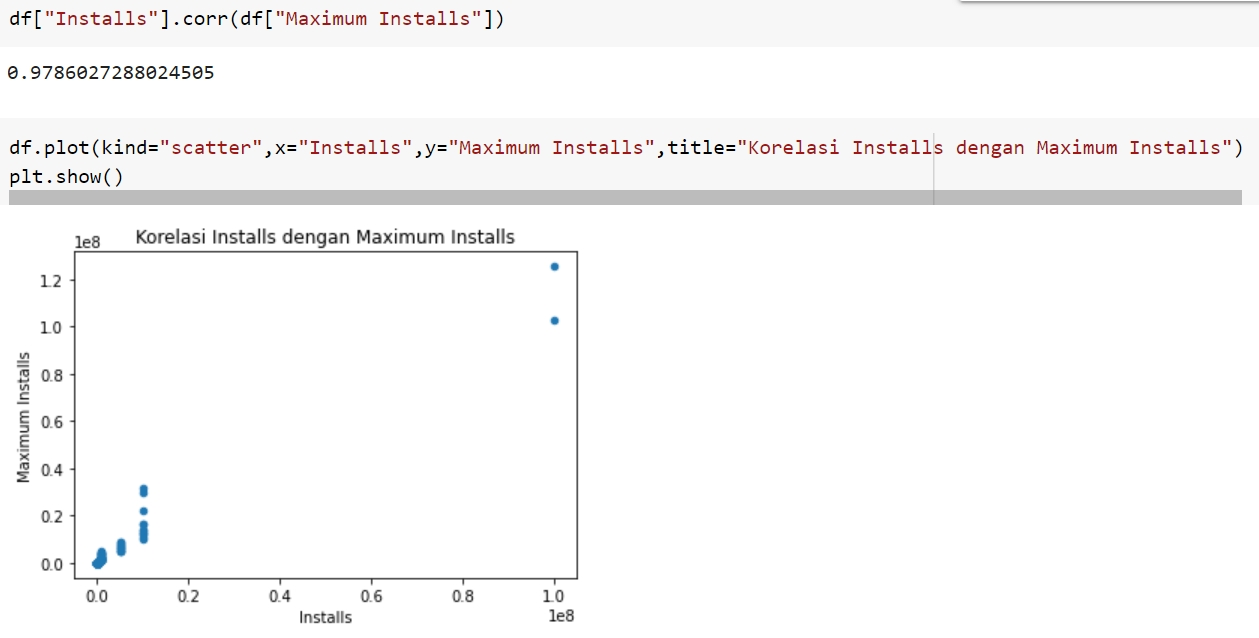
r mendekati 0 sehingga tidak ada hubungan antara kedua variabel

* 1. Koefisien korelasi ( r ) antara Installs dengan Minimum Installs

****

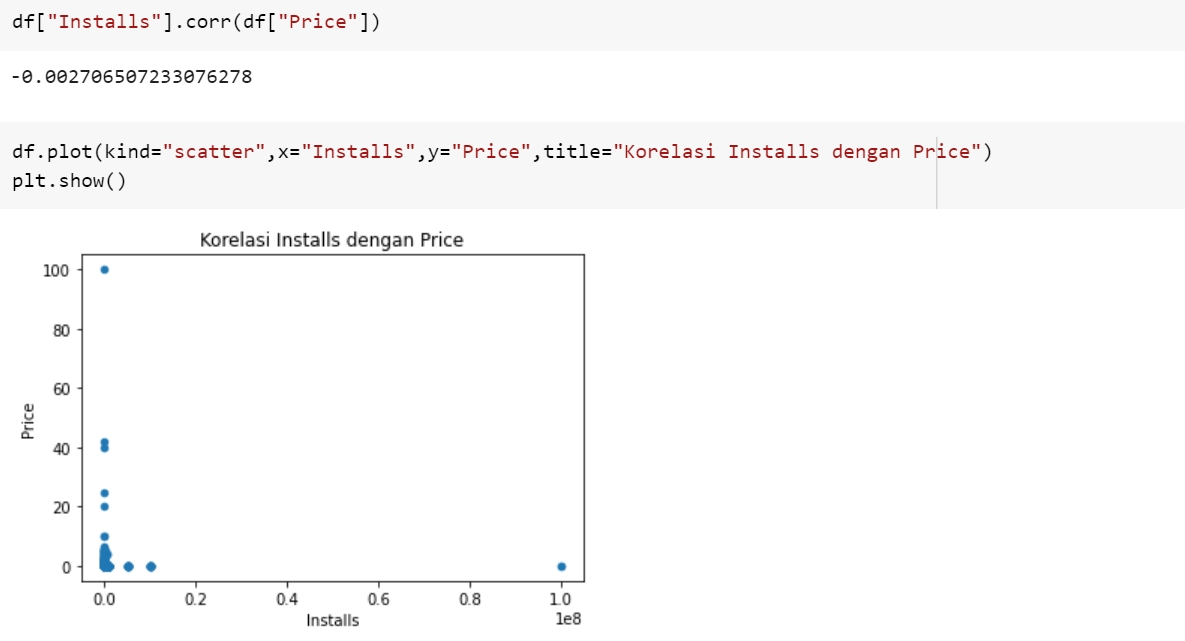
Perfect positive correlation

* 1. Koefisien korelasi ( r ) antara Installs dengan Maximum Installs

****

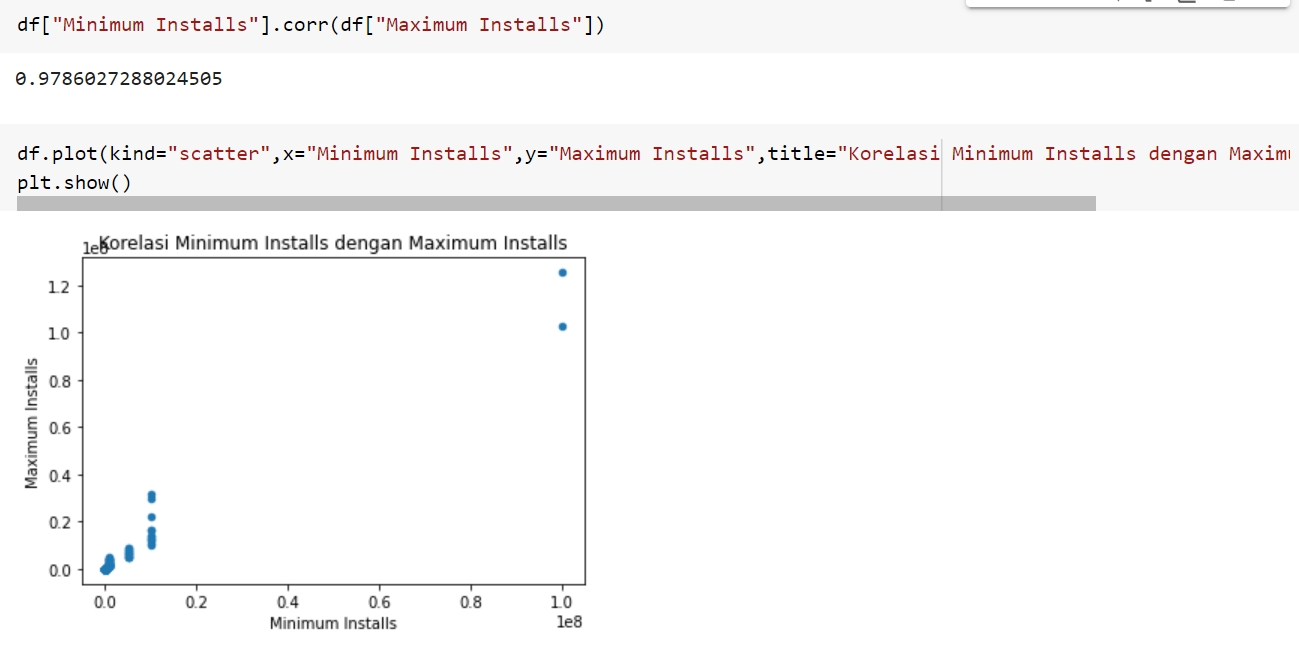
High positive correlation

* 1. Koefisien korelasi ( r ) antara Installs dengan Price

****

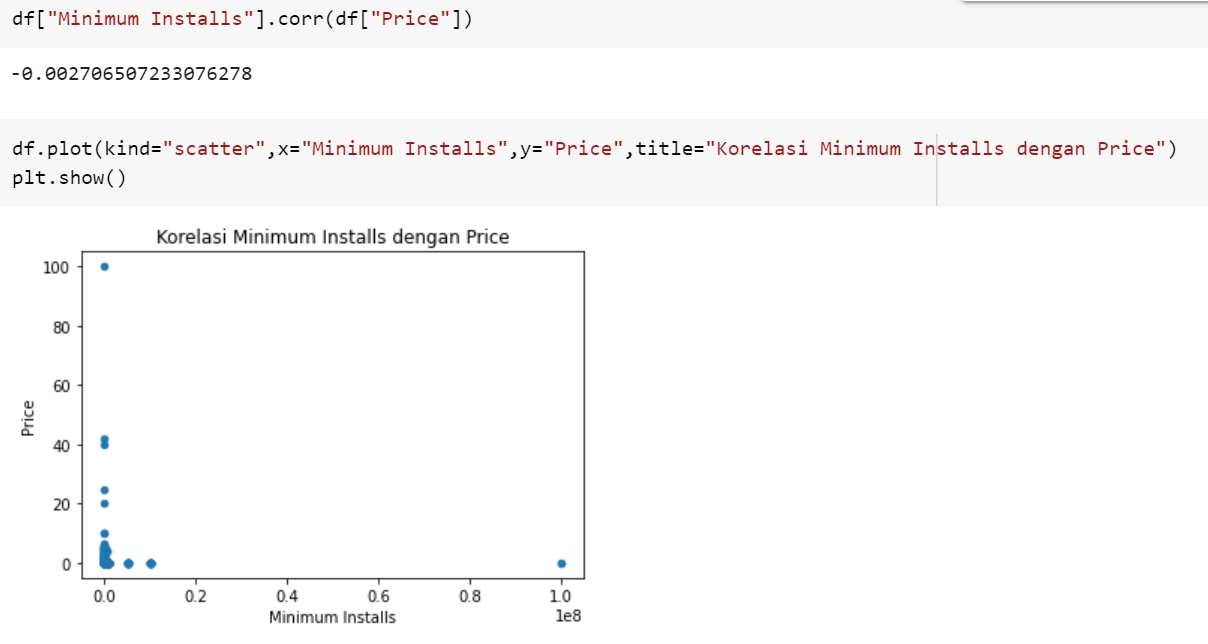
No correlation

* 1. Koefisien korelasi ( r ) antara Minimum Installs dengan Maximum Installs

****

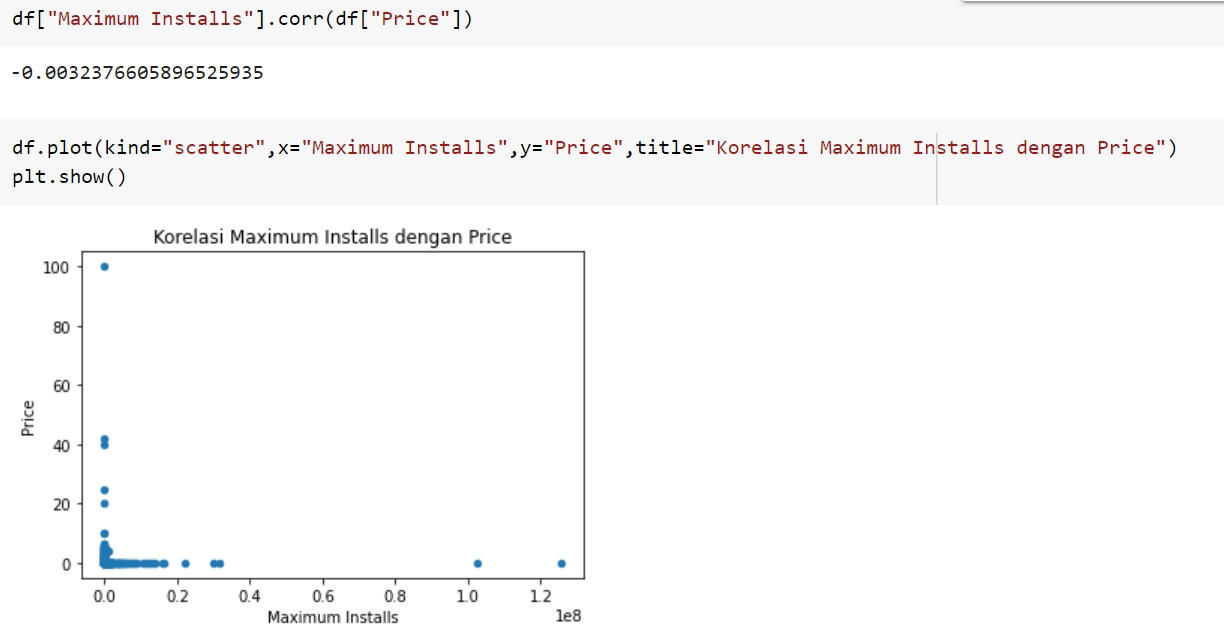
High positive correlation

* 1. Koefisien korelasi ( r ) antara Minimum Installs dengan Price

****

No correlation

* 1. Koefisien korelasi ( r ) antara Maximum Installs dengan Price

****

No correlation

1. **Tugas 8 (Data Cleansing)**
   1. Nilai NaN

Pertama-tama dicari terlebih dahulu kolom mana saja yang memiliki nilai NaN, dapat dicari dengan :

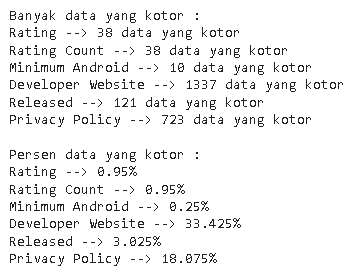
column\_with\_nan = [i for i in df.columns if df[i].isnull().sum() > 0]

# keluaran : ['Rating', 'Rating Count', 'Minimum Android', 'Developer Website', 'Released', 'Privacy Policy']

Dapat dilihat ada 6 kolom yang memiliki nilai NaN. Lalu, kita juga dapat melakukan pengecekan terhadap banyaknya nilai NaN yang dimiliki masing-masing kolom sebagai tingkat kekotorannya.  Selain itu, kita juga bisa menuliskan berapa persen data yang kotor tiap kolomnya.



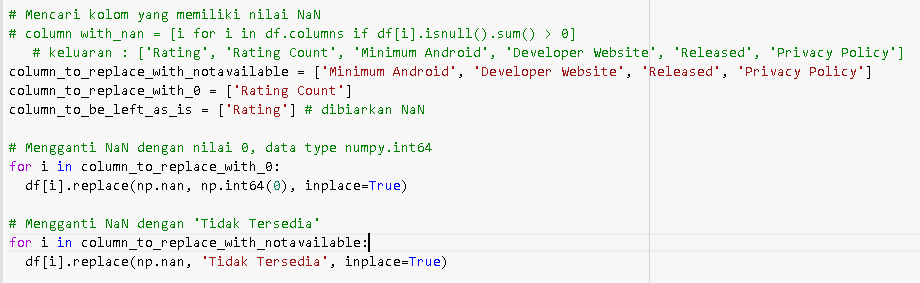
Keluaran :



Setelah itu, kolom dibagi menurut penanganannya. Kolom ['Minimum Android', 'Developer Website', 'Released', 'Privacy Policy'] akan diubah nilai NaN-nya menjadi nilai string “Tidak Tersedia” karena kolom tersebut berisi data non-numerikal yang memang tidak tersedia saat dilakukan pengambilan.

Kolom ['Rating Count'] akan diubah menjadi nilai 0 karena saat dilakukan pengambilan data, kolom ini hanya sebagai jumlah atau banyaknya data kolom ['Rating'] yang tidak tersedia.

Kolom ['Rating'] sendiri akan dibiarkan apa adanya karena memang tidak memiliki data.



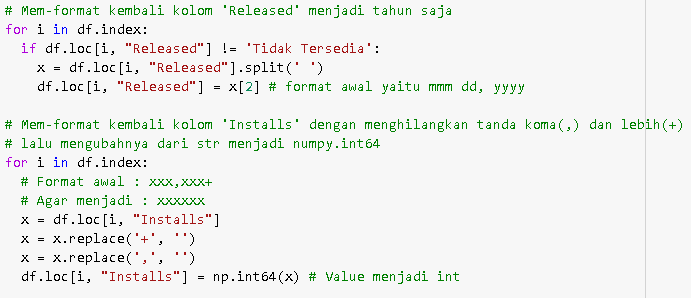
* 1. Nilai Kolom "Released"

Kolom ini memiliki nilai dengan format [mmm dd, yyyy]. Sebagai contoh, df.loc[1, 'Released'] memiliki keluaran ‘May 21, 2020’. Agar kolom ini tidak memiliki kategori yang terlalu luas, kelompok kami memutuskan untuk melakukan penyederhanaan kategori terhadap kolom ini sehingga menjadi sebatas tahun saja. Sebagai contoh, keluaran yang diharapkan dari df.loc[1, 'Released'] adalah ‘2020’.

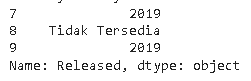
* 1. Nilai kolom "Installs"

Sebagai kolom yang menyatakan jumlah Install aplikasi, kolom ini seharusnya merupakan atribut numerikal yang dapat dilakukan operasi numerik. Namun kenyataannya dari dataset yang didapat, kolom ini memiliki tipe data string yang bukan merupakan ciri atribut numerikal. Selain itu hampir semua anggota kolom ini juga memiliki tanda koma(,) dan lebih(+).

Diharapkan, data kolom ini dapat diubah dari str menjadi data dengan tipe numpy.int64 atau bilangan bulat.

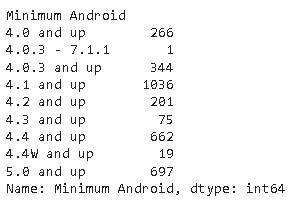


Keluaran df.loc[7:9, 'Released'] setelah dilakukan data cleansing :



* 1. Nilai kolom 'Minimum Android'

df.groupby('Minimum Android')['Minimum Android'].count()[11:20]akan menghasilkan keluaran :

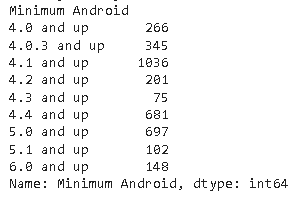


Dapat dilihat ada 2 variabel atribut yang memiliki kemiripan sehingga dapat digabung menjadi satu agar banyak variasi juga semakin berkurang sehingga dapat mempermudah visualisasi. Variabel yang dimaksud adalah 4.0.3 - 7.1.1 dan 4.4W and up yang dapat diubah menjadi 4.0.3 and up dan 4.4 and up berturut-turut.

Metode berikut dapat digunakan untuk melakukan tugas tersebut :

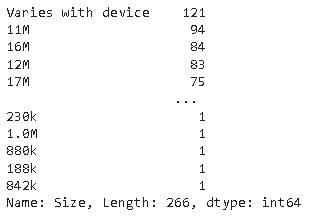


Didapatkan :



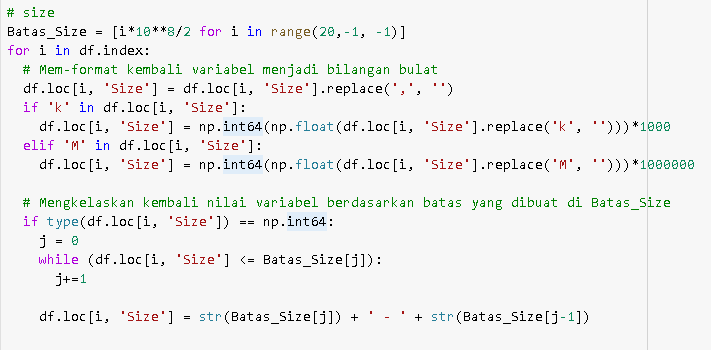
* 1. Nilai kolom 'Size'

df['Size'].value\_counts()akan menghasilkan keluaran :



dengan 266 jenis nilai berbeda.

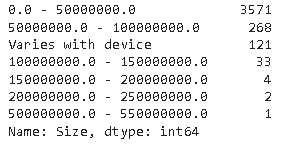
Agar atribut ini dapat lebih dimanfaatkan, kelompok kami memutuskan untuk membuat format nilai baru yaitu dengan mengkelaskan ukuran file sesuai dengan rentangnya, yaitu Batas\_Size = [i\*10\*\*8/2 for i in range(20,-1, -1)] dengan keluaran [1000000000.0, 950000000.0, 900000000.0, …, 100000000.0, 50000000.0, 0.0].



Untuk setiap nilai dalam atribut 'Size' yang tidak bernilai ‘Varies with device’, nilai yang awalnya bertipe str dengan format xxxk (menandakan xxx ribu bita) atau xxxM (menandakan xxx juta bita) akan diubah menjadi nilai bertipe np.int64.

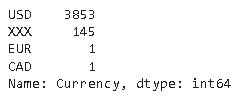
Setelah didapatkan nilai dalam bentuk bilangan bulat, variabel akan dikelaskan menurut rentang yang sudah ditentukan dalam Batas\_Size.

df['Size'].value\_counts()akan menghasilkan keluaran :



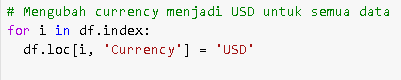
* 1. Nilai 'Currency'

df['Currency'].value\_counts()akan menghasilkan keluaran :



Setelah di-cek kembali menggunakan  df[(df['Currency'] != 'USD') & (df['Free'] != True)] keluaran menghasilkan dataframe kosong.

Hal ini menandakan bahwa semua item dalam dataset dengan variabel 'Currency' bukan 'USD' pastilah gratis. Sehingga, kami memutuskan untuk menyamakan semua data dalam kolom 'Currency' menjadi bernilai 'USD'.



Hal ini kami lakukan untuk memastikan data pada kolom 'Price' selalu konsisten karena memiliki mata uang yang sama. Data 'Price' yang konsisten akan memudahkan proses visualisasi data.

# 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).

# LESSON LEARNED

Dari tugas besar ini, kami telah mempelajari cara untuk menganalisis dan memvisualisasikan data dengan jumlah yang banyak ataupun sedikit. Data yang kami analisis adalah data Google Playstore Android App Juni 2021. Kami melakukan analisis dan visualisasi data dengan melakukan beberapa hal. Hal tersebut adalah mendeskipsikan data, membuat karakteristik data, membuat statistik data, korelasi, visualisasi dan juga data cleansing. Dari hal-hal yang kami lakukan ini, kami dapat melihat hubungan (korelasi) antar elemen dalam data dan juga dapat melihat visualisasi dari data tersebut, serta data yang tidak sesuai. Kami juga menyadari bahwa suatu masalah dapat lebih cepat diselesaikan dengan bekerja sama, pada tugas kali ini kami belajar cara untuk bekerja sama dengan baik dalam kurun waktu tertentu sehingga hal ini mengasah kemampuan kami untuk bekerja dalam tim dan time management kami.