Nama : Al Fitra Nur Ramadhani

NIM : 202210370311264

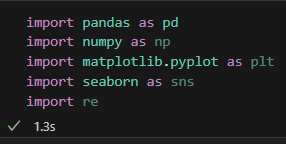
Mata Kuliah : Data, Information & Knowledge

**Laporan PreProcessing Metadata Aplikasi Google Play Store**

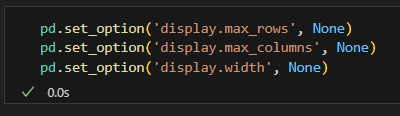
1. **Deskripsi**

Laporan ini bertujuan untuk menjelaskan tahapan PreProcessing dataset metadata aplikasi dari Google Play Store yang berisi informasi seperti nama aplikasi, kategori, rating, jumlah ulasan, ukuran aplikasi, jumlah instalasi, tipe (gratis/berbayar), harga, rating konten, genre, tanggal pembaruan terakhir, versi saat ini, dan versi Android yang didukung. Dataset ini diambil dari Kaggle yang awalnya ada 10.842 rows , kemudian setelah melewati beberapa tahapan preprocessing menjadi sekitar 9660 rows.

1. **Link Kaggle =** [**Google Play Store Apps**](https://www.kaggle.com/datasets/lava18/google-play-store-apps)
2. **Implementasi Step by Step**
3. **Menampilkan metadata Information Details**

****

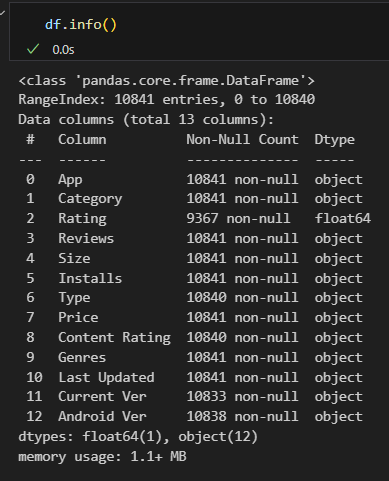
* Import beberapa library yang dibutuhkan, pandas (Pengolahan data), numpy (Perhitungan Numerik), matplotlib (visualisasi Grafik), seaborn (Visualisasi Statistik), re (Ekspresi Reguler)

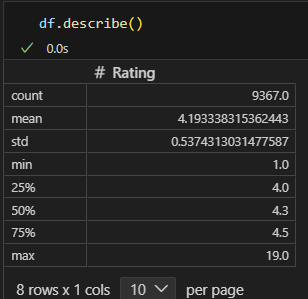


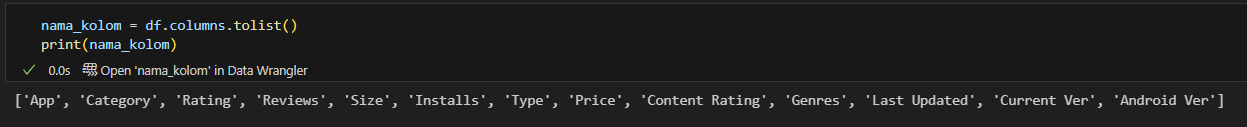
**-** Atur untuk menampilkan semua baris, kolom dan tampilan penuh



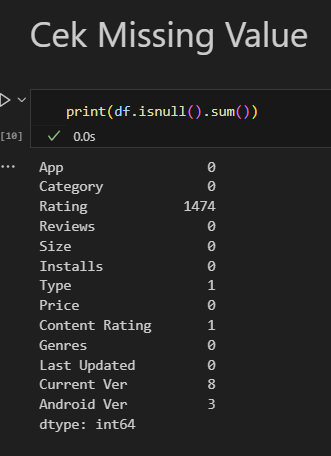
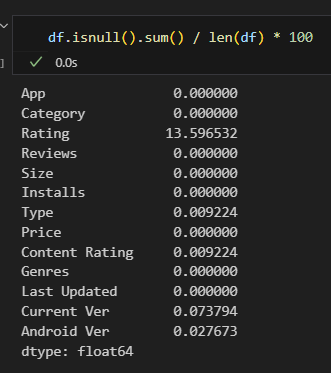
**-** Import data csv nya untuk dibaca di kernel python

****

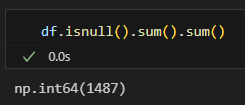
* Menampilkan Informasi Dataframe
* Jumlah data ada 10841 entries dan 13 kolom
* Beberapa tipe data tidak sesuai dengan hakikatnya seperti Rating, Reviews, Size, Installs, Price, dan Last Updated
* Beberapa data ditemukan missing value di kolom Rating, Type, Content Rating, Cureent Ver, dan Android Ver
* Statistik Deskriptif, menyangkut mean , count, standar deviasi, dll untuk melihat sebaran data

****

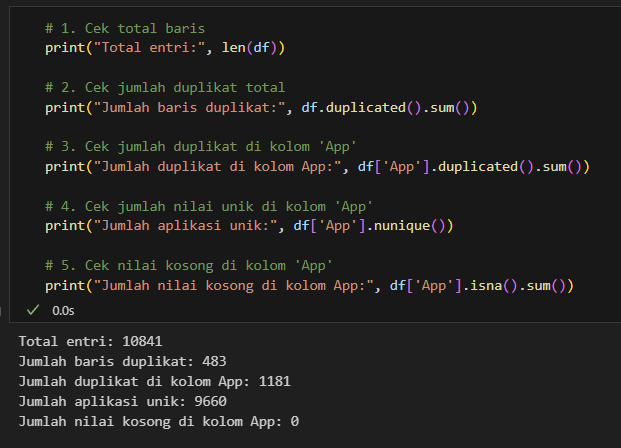
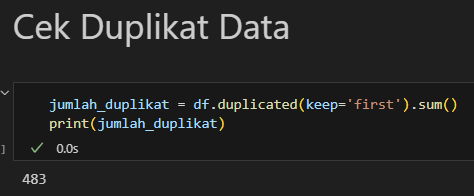
* Semua nama kolom yang ada, agar mudah untuk di analisis
* ===================================================================

****

**Cek Missing Value**

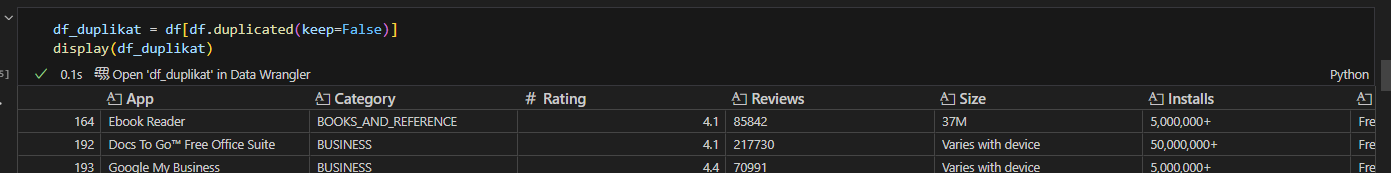
* Informasi detail mengenai kolom mana saja yang missing, ada di kolom Rating, Type, Content rating, Current Ver dan Android Ver
* Total Missing Value ada 1487 cell
* ****Dan ditampilkan persenan missing value terhadap keseluruhan data , paling tinggi ada di kolom Reviews yaitu 13,6%

**==============================================================**

****

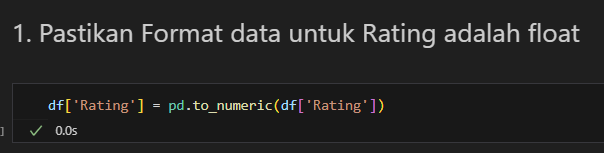
**Cek Data Duplikat**

* Setelah di cek duplikat data ternyata ada 1181 duplikat dengan mengacu pada kolom App

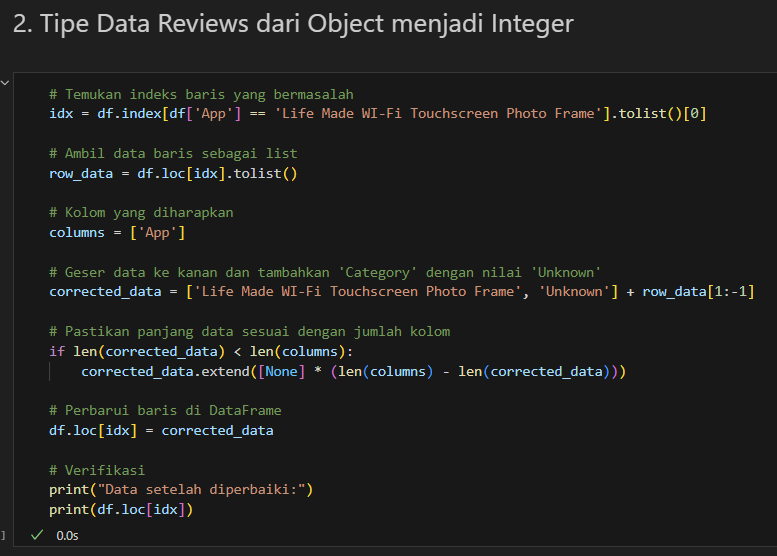
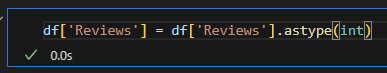
****

1. **Data Preprocessing**

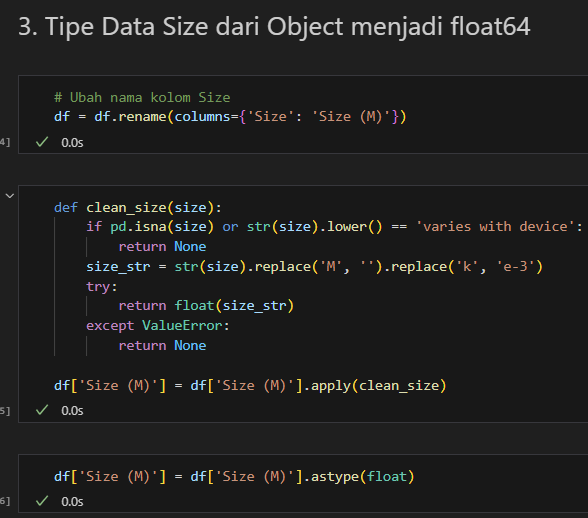
* **Ubah Data Type**

****

* **-** Tipe data numeric untuk Rating

****

* Ternyata setelah di cek di Kolom Reviews ada di kolom App (Life Made WI-Fi Touchscreen Photo Frame) datanya kegeser ke kanan sebanyak 1
* Setelah di proses kemudian ubah tipe datanya menjadi integer



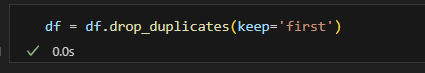
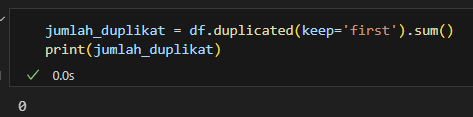
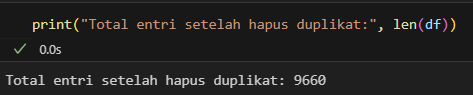
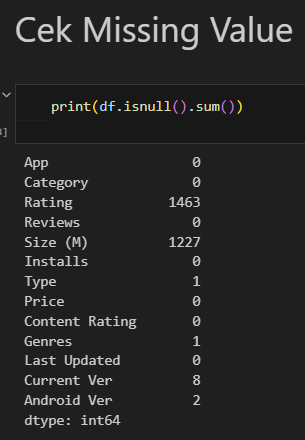
* Rename kolom untuk Size menjadi Size (M) untuk memberikan penjelasan lebih baik, karena akan di ubah tipe data nya menjadi float
* Cleaning format data yang Dimana yang mengandung huruf M maka akan dihapus



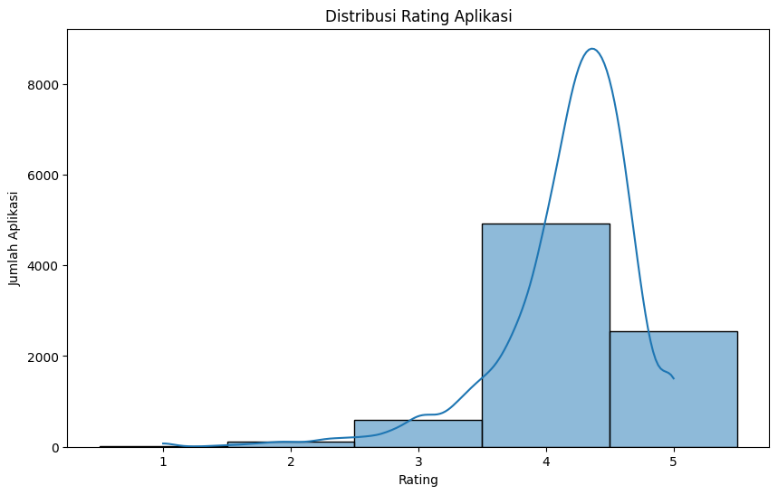
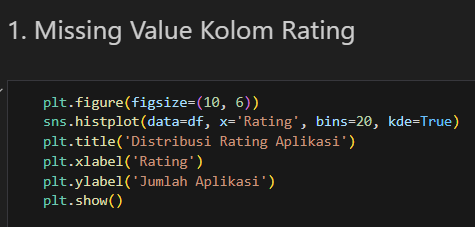
* Hapus karakter symbol yang ada di kolom Installs dan ubah tipe data menjadi Integer

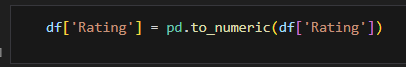
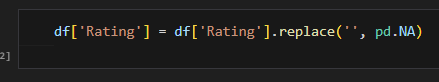
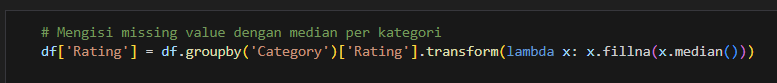
* Ubah tipe data Price menjadi float dan menghapus symbol symbol agar lebih konsisten untuk numeric

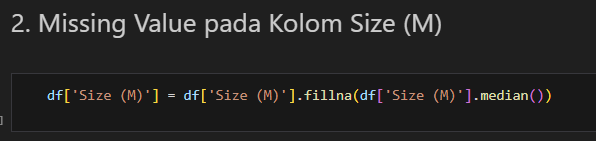


* Ubah tipe data Last Updated menjadi Date Time
* **Cek & Drop Duplikat**
* ****
* Drop duplikat di kolom App , dan setiap baris yang duplikat dan gunakan first untuk menjaga data pertama tidak ke drop
* Periksa lagi data nya apakah sudah clear
* Total entri setelah duplikat ada 9660 rows
* **Cek Missing Value Again**
* Missing Value = Rating, Size (M), Type, Genres, Current Ver, Android Ver

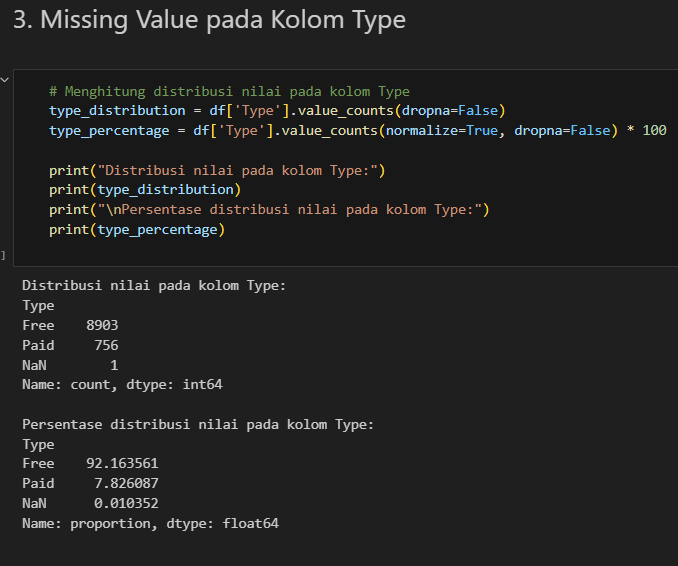
**# Kolom Rating**

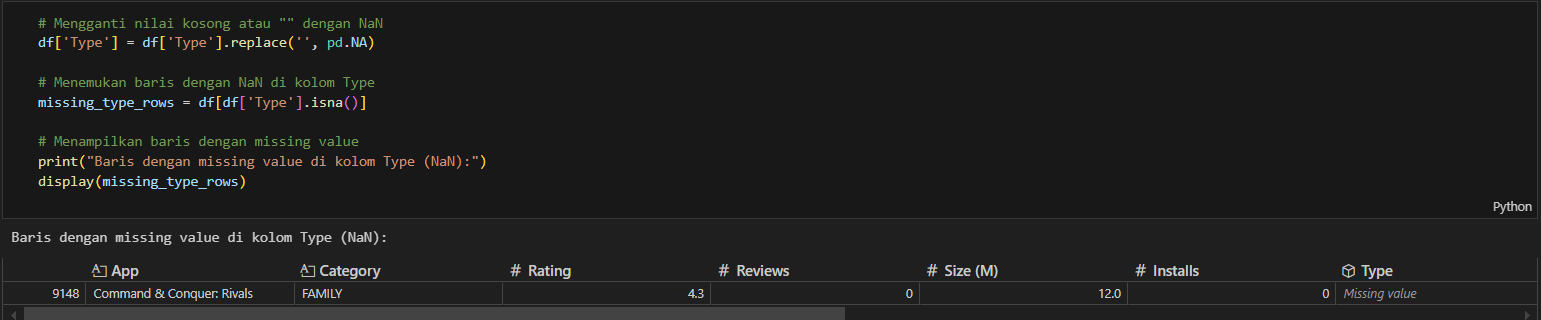


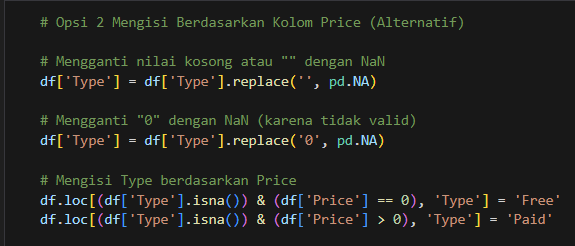
* Setelah di cek distribusi data lewat histogram ternyata ada data outlier , data lebih dari 5, yang seharusnya rentang value untuk rating kisaran 1-5
* 
* Mengisi missing value di Kolom Rating dengan median karena lebih robust untuk distribusi data dan tidak terpengaruh oleh outlier nya

**# Kolom Size (M)**

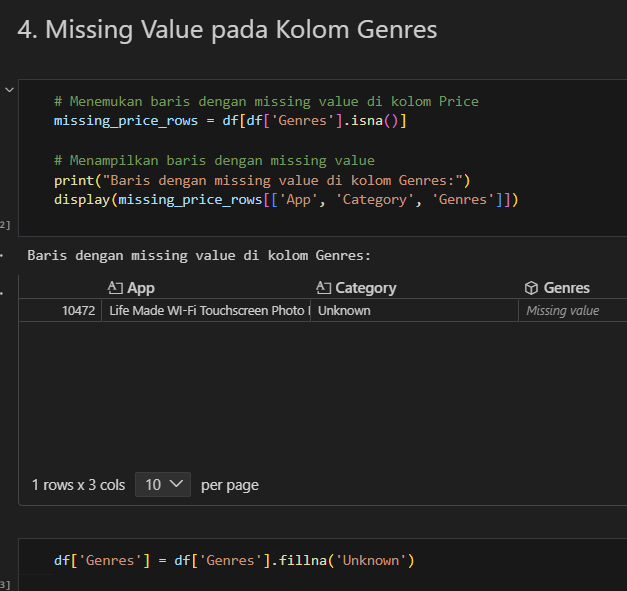
* Isi missing value dengan median karena lebih robust terhadap distribusi data dan data di kolom Size (M) bersifat ordinal

**# Kolom Type**

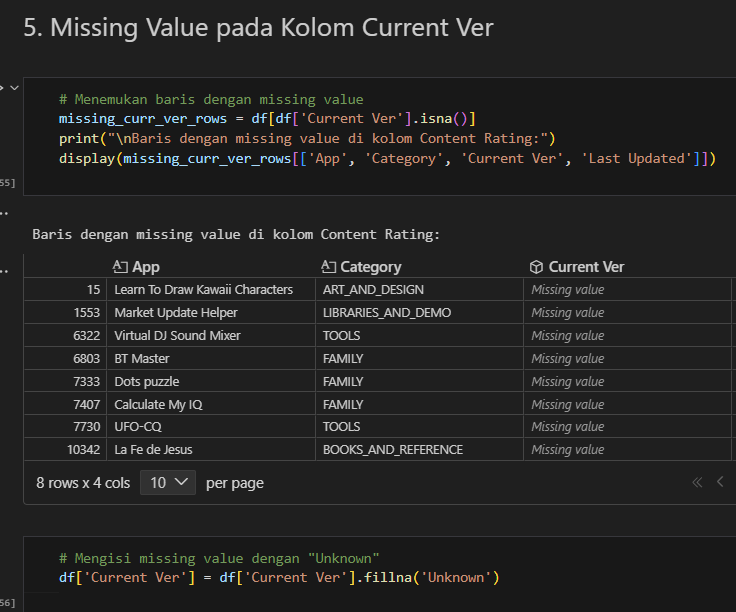
****



* Mengisi missing value dengan melihat kolom price , karena price nya 0 maka diisi dengan free, jika > 0 maka diisi dengan paid

**# Kolom Genres**

* Karena Data tidak tahu source nya maka diisi Unknown agar cari aman

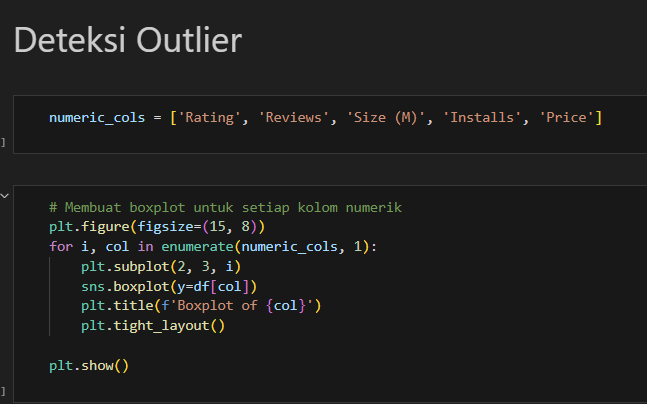
# Kolom Current Ver

* Mengisi missing value dengan Unknown karena tidak tahu source nya seperti apa

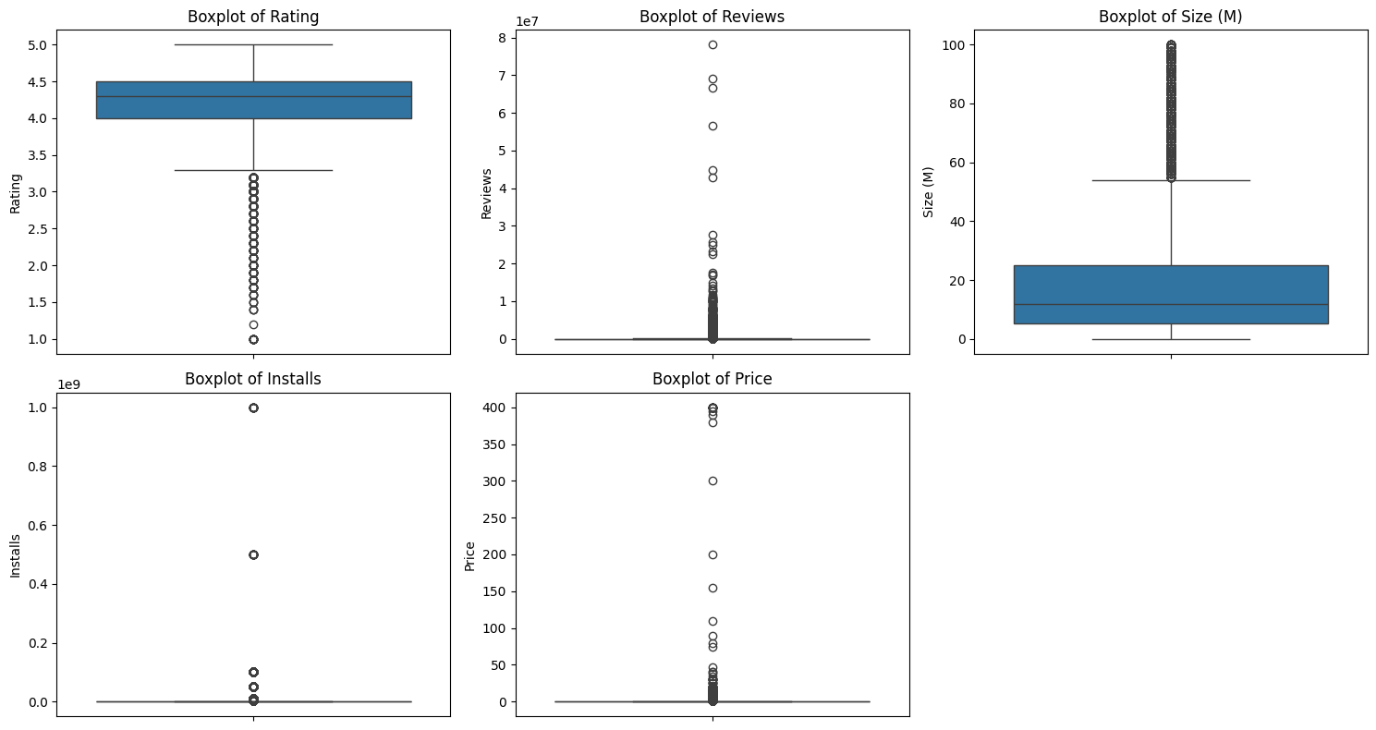
**# Kolom Android Ver**

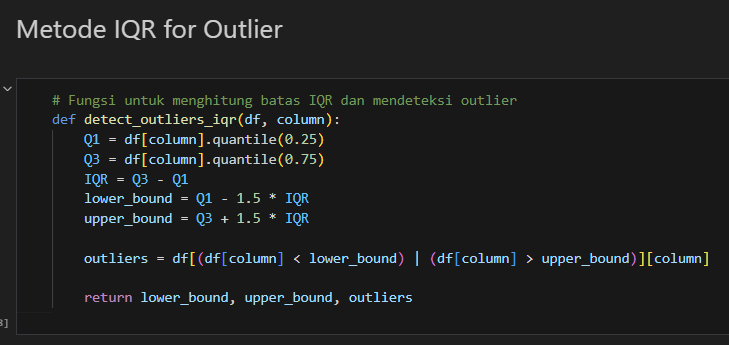
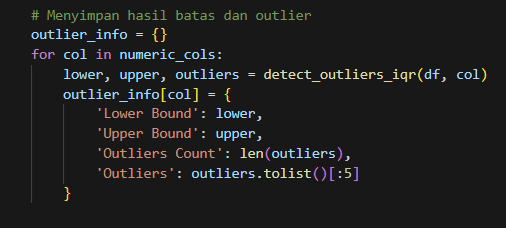
* Isi missing value dengan Unknown agar cari aman karena source nya tidak diketahui, karena data nya kategorikal maka ini opsi terbaik menurut saya

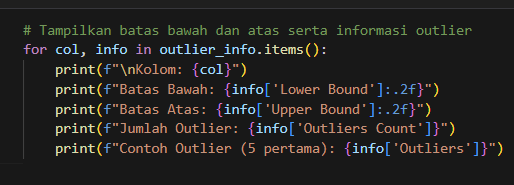
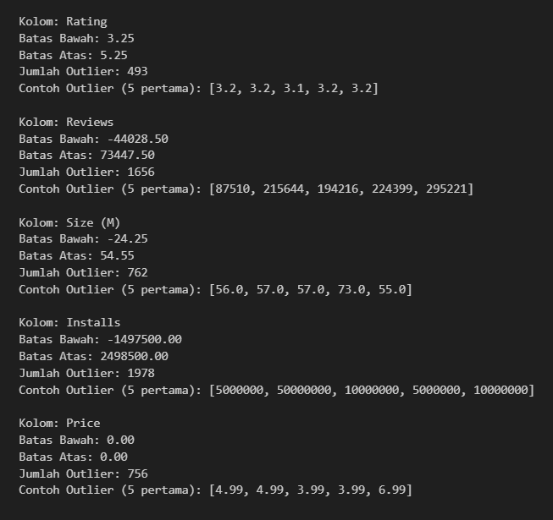
1. **Deteksi Outlier**

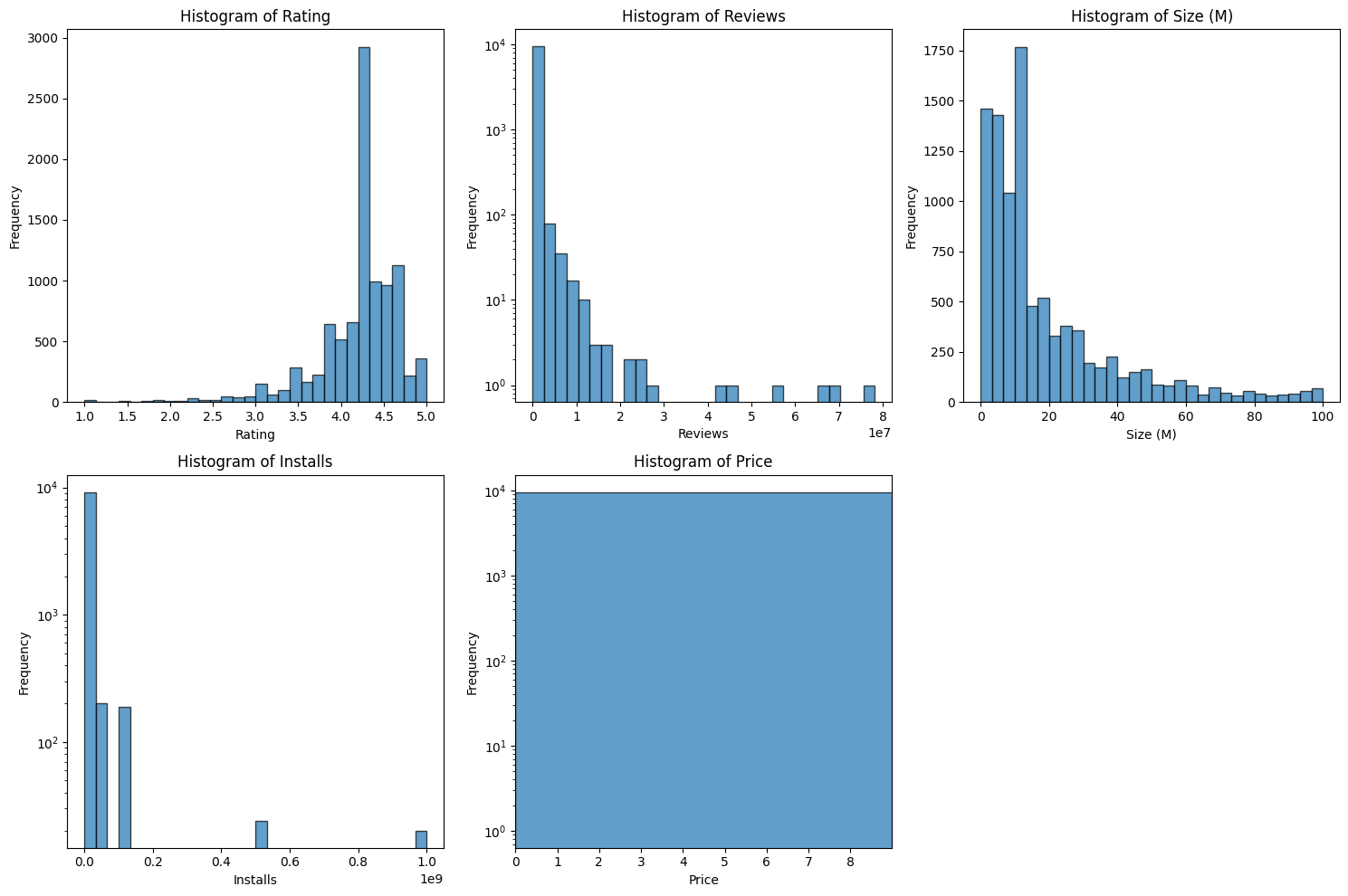
****

* Deteksi Outlier dengan boxplot untuk kolom numerik = Rating, Reviews, Size (M), Installs, price
* Untuk Distribusi data nya masih aman jika dilihat dari konteks kolom nya



**# Metode IQR for Outlier**

****

**# Histogram untuk melihat distribusi data**