



Tugas Besar (STEI) Literasi Data dan Intelelegensi Artifisial

WI2002

Presented by Garam dan Madu

Dataset yang diteliti



Data Aplikasi Pada Google Play

Link dataset:

<https://www.kaggle.com/datasets/bhavikjikadara/google-play-store-applications>

Data Video Game Terpopuler 1980 - 2023

Link dataset:

<https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023>

Pertanyaan Penelitian



Data Aplikasi pada Google Play

- Bagaimana konten rating dari suatu aplikasi pada Google Play Store memengaruhi jumlah pengunduhan aplikasi?
- Apakah benar bahwa aplikasi bertipe game lebih sering melakukan pembaruan dibandingkan tipe lainnya?
- Bagaimana harga suatu aplikasi memengaruhi jumlah unduh dari aplikasi tersebut?
- Apakah benar bahwa aplikasi bertipe komunikasi memiliki mayoritas jumlah unduh yang lebih tinggi? Apa saja variabel-variabel yang mungkin berperan dalam mempengaruhi jumlah unduh aplikasi tersebut?

Data Video Game Terpopuler 1980 - 2023

- Bagaimana tanggal rilis suatu video game dapat memengaruhi jumlah pemain dari game tersebut?
- Bagaimana popularitas suatu genre video game berubah dari tahun ke tahun?
- Jenis genre video game apa yang memiliki jumlah unduh paling tinggi dari 1980-2023? Apa saja variabel-variabel yang mungkin berperan dalam mempengaruhi jumlah unduh aplikasi tersebut
- Apa yang membuat suatu video game memiliki angka wishlist yang tinggi? Apa saja variabel-variabel yang mungkin berperan dalam mempengaruhi angka tersebut?

Deskripsi dan Karakteristik Data

Data Aplikasi pada Google Play:

Atribut	Arti	Tipe Data	Karakteristik
App	Nama Aplikasi	Kategorikal (Nominal)	Teks bebas, Unik (berbeda antara satu sama lain)
Category	Kategori Aplikasi	Kategorikal (Nominal)	Terdiri dari banyak kategori
Rating	Rating pengguna	Kuantitatif	Range 0.0-5.0, banyak <i>missing value</i>
Reviews	Jumlah ulasan pengguna	Kuantitatif	Nilai numerik
Size	Ukuran aplikasi	Kuantitatif	Nilai Numerik
Installs	Banyaknya unduhan	Kategorikal (Ordinal)	Kategori berdasarkan range nilai
Type	Gratis atau berbayar	Kategorikal (Nominal, Biner)	Hanya 2 kategori
Price	Harga aplikasi	Kuantitatif	Harga 0 = gratis, dikuantifikasi dalam mata uang <i>dollar</i>
Content Rating	Target/Minimal Usia	Kategorikal (Nominal)	Dibagi berdasarkan range umur
Genres	Genre dari aplikasi	Kategorikal (Nominal)	Bisa terdiri dari 1 atau lebih genre yang dipisahkan dengan ;
Last Updated	Tanggal pembaruan terakhir	Time-Series	Berisi tanggal
Current Ver	Versi aplikasi sekarang	Kategorikal (nominal)	Ada versi yang "Varies with devices"
Android Ver	Versi aplikasi di android sekarang	Kategorikal (ordinal)	Ada versi yang "Varies with devices"

Deskripsi dan Karakteristik Data

Data Video Game Terpopuler 1980 - 2023:

Atribut	Arti	Tipe Data	Karakteristik
Title	Judul <i>video game</i>	Kategorikal (Nominal)	Teks bebas, unik (berbeda antara satu sama lain)
Release Date	Tanggal rilis <i>video game</i>	Time-Series	Berisi Tanggal
Team	Nama tim <i>developer</i>	Kategorikal	Teks
Rating	Rating rata-rata	Kuantitatif	Range 0.0-5.0
Times Listed	Jumlah <i>users</i> yang melakukan <i>list</i> pada <i>video game</i>	Kuantitatif	Nilai Numerik
Number of Reviews	Jumlah <i>users</i> yang mengulas <i>video game</i>	Kuantitatif	Nilai numerik
Genres	Genre dari game	Kategorikal	Terdiri dari banyak kategori
Summary	Summary dari tim <i>developer</i>	Kategorikal (Nominal)	Teks tentang isi game, mayoritas unik
Reviews	Ulasan dari <i>user</i>	Kategorikal (Nominal)	Kumpulan ulasan pengguna
Plays	Jumlah <i>users</i> yang pernah memainkan game	Kuantitatif	Nilai numerik
Playing	Jumlah <i>users</i> yang sedang memainkan game	Kuantitatif	Nilai numerik
Backlogs	Jumlah <i>users</i> yang memiliki game, tapi belum memainkan	Kuantitatif	Nilai numerik
Wishlist	Jumlah <i>users</i> yang berharap memainkan game	Kuantitatif	Nilai numerik

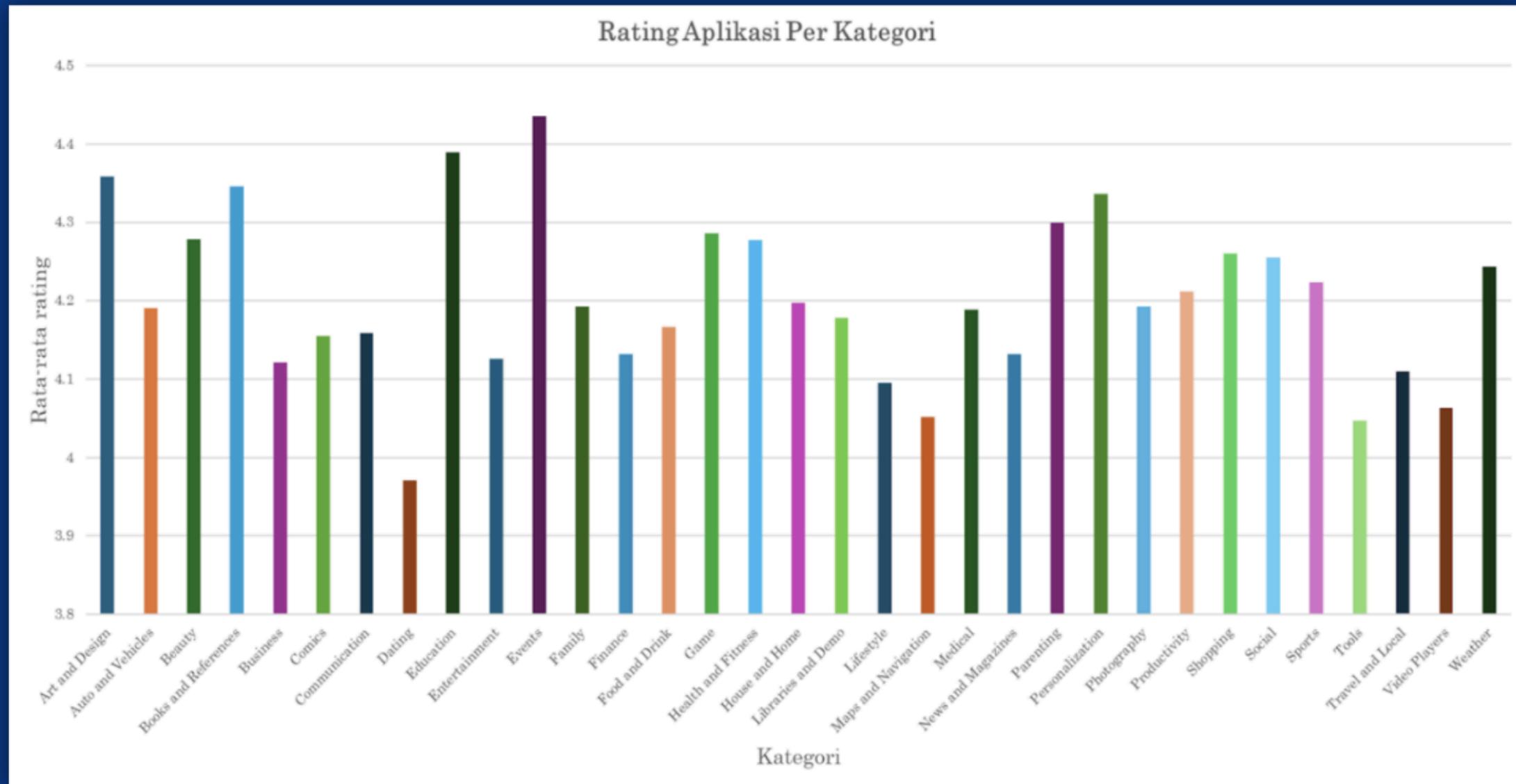
Visualisasi Data Aplikasi pada Google Play

Garam & Madu

Pada metode visualisasi yang kami gunakan di bagian ini, ada beberapa hal yang perlu diperhatikan:

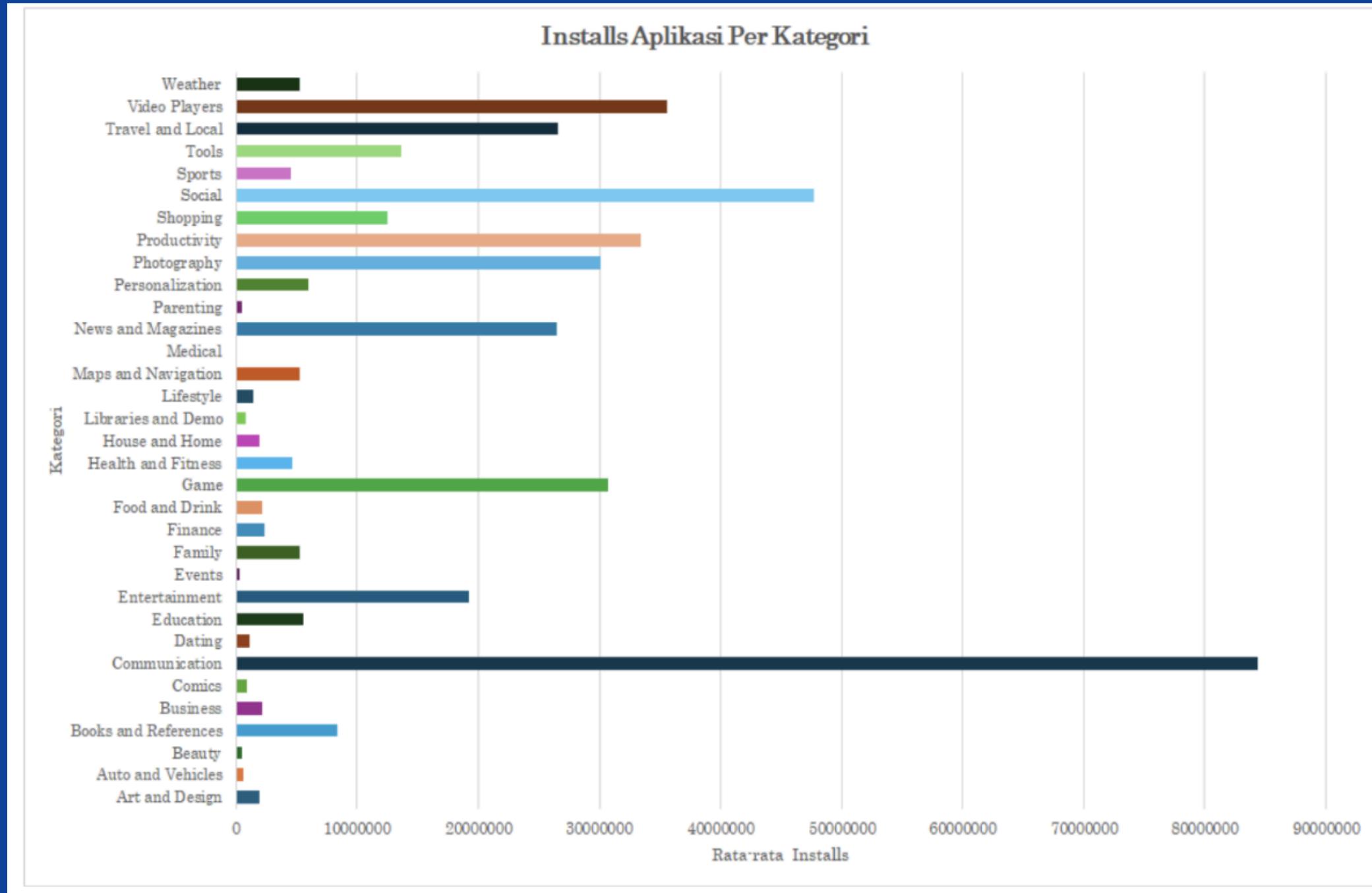
1. Pembuatan dilakukan pada Ms. Excel dengan database terdiri atas 5 sheets, yaitu
 - a. SheetUtama, berisikan data mentah yang diambil dari sumber data langsung dalam format file .csv yang telah dikonversikan menjadi format file .xlsx.
 - b. PerbandinganKategori, bersihkan data dan visualisasi untuk perbandingan kategori.
 - c. PerubahanTerhadapWaktu, bersihkan data dan visualisasi untuk penampilan perubahan terhadap waktu.
 - d. PenampilanHierarkis, berisikan dan visualisasi untuk penampilan hierarki dan hubungan keseluruhan bagian.
 - e. PlottingRelationships, berisikan dan visualisasi untuk plotting relationships
2. Pada data mentah, format data Installs memiliki + di akhir setiap value, untuk standardisasi data Install, maka diperbuat kolom Installs_Clean yang bervalue integer menggunakan formula
 $=VALUE(SUBSTITUTE(SUBSTITUTE(G2, ",", ""), "+", ""))$
3. Pada data mentah, format data Price berupa value string karena terdapat \$ di depannya, untuk standardisasi value Price, maka diperbuat kolom Price_Number dengan menggunakan formula
 $=VALUE(SUBSTITUTE(I2,"$",""))$

3.1.1 Visualisasi 1: Rata-rata Rating Berdasarkan Kategori



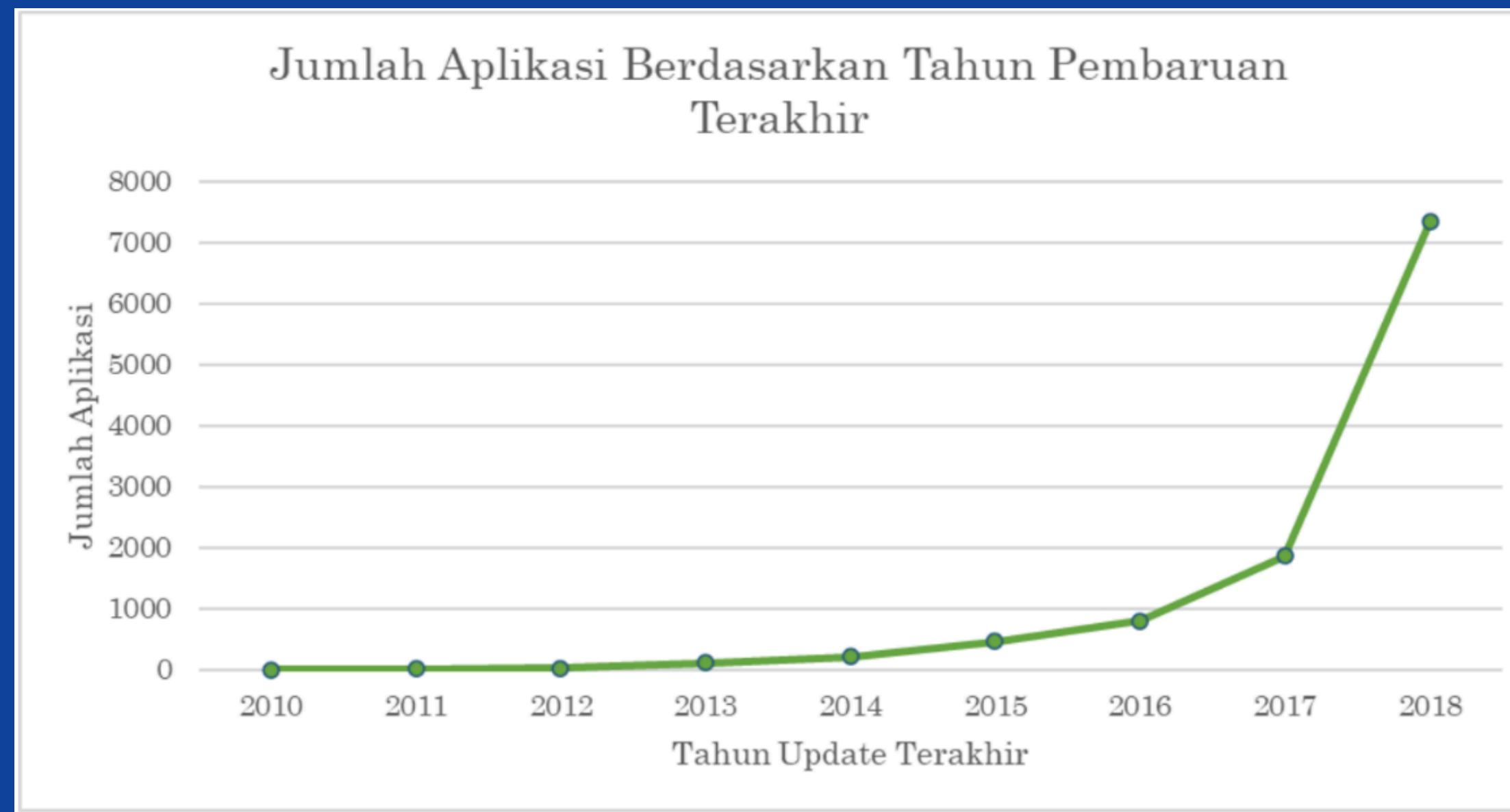
Berdasarkan grafik "Rating Aplikasi Per Kategori", dapat disimpulkan bahwa **kategori aplikasi dengan rata-rata rating tertinggi adalah Events, Education, dan Art and Design**, menunjukkan bahwa pengguna sangat puas terhadap aplikasi dalam kategori tersebut. Sebaliknya, kategori Dating, Maps and Navigation, serta Video Players memiliki rating terendah, yang mengindikasikan perlunya peningkatan dari segi kualitas, fitur, atau pengalaman pengguna. Secara umum, sebagian besar kategori memiliki rating antara 4.1 hingga 4.3, dan rata-rata rating secara keseluruhan bernilai sekitar 4.19 mencerminkan standar kualitas yang cukup konsisten di Google Play Store. Hal ini menunjukkan bahwa aplikasi pendidikan, kreativitas, dan acara cenderung memberikan nilai lebih bagi pengguna, sementara pengembang di kategori dengan rating rendah perlu lebih memperhatikan kebutuhan dan ekspektasi penggunanya.

3.1.2 Visualisasi 2: Rata-rata Jumlah Installs Berdasarkan Kategori



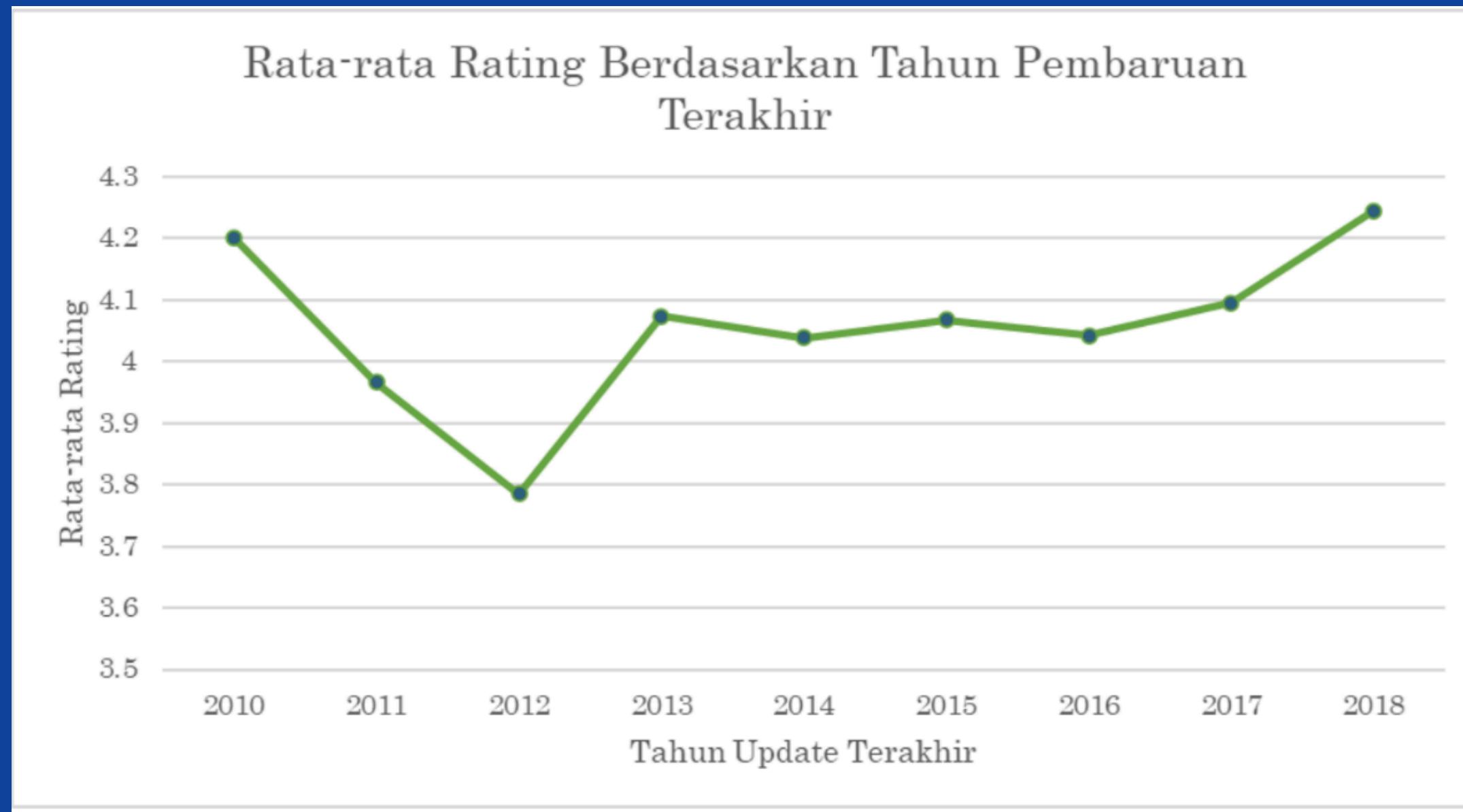
Berdasarkan grafik "Installs Aplikasi Per Kategori", terlihat bahwa **kategori Communication, Game, dan Social mendominasi jumlah rata-rata instalasi**, menandakan tingginya kebutuhan pengguna terhadap aplikasi yang bersifat interaktif dan hiburan. Sementara itu, kategori seperti Parenting, Events, dan Libraries and Demo memiliki rata-rata instalasi yang sangat rendah, yang bisa mencerminkan segmen pengguna yang terbatas atau kurangnya eksposur aplikasi dalam kategori tersebut. Ketimpangan yang cukup besar antara kategori populer dan kurang populer ini menunjukkan adanya peluang bagi pengembang untuk mengeksplorasi ceruk pasar yang belum tergarap maksimal. Selain itu, tingginya instalasi tidak selalu mencerminkan kualitas, sehingga penting untuk mempertimbangkan metrik lain seperti rating untuk mendapatkan gambaran menyeluruh mengenai kepuasan pengguna.

3.1.3 Visualisasi 3: Jumlah Aplikasi Berdasarkan Tahun Pembaruan Terakhir



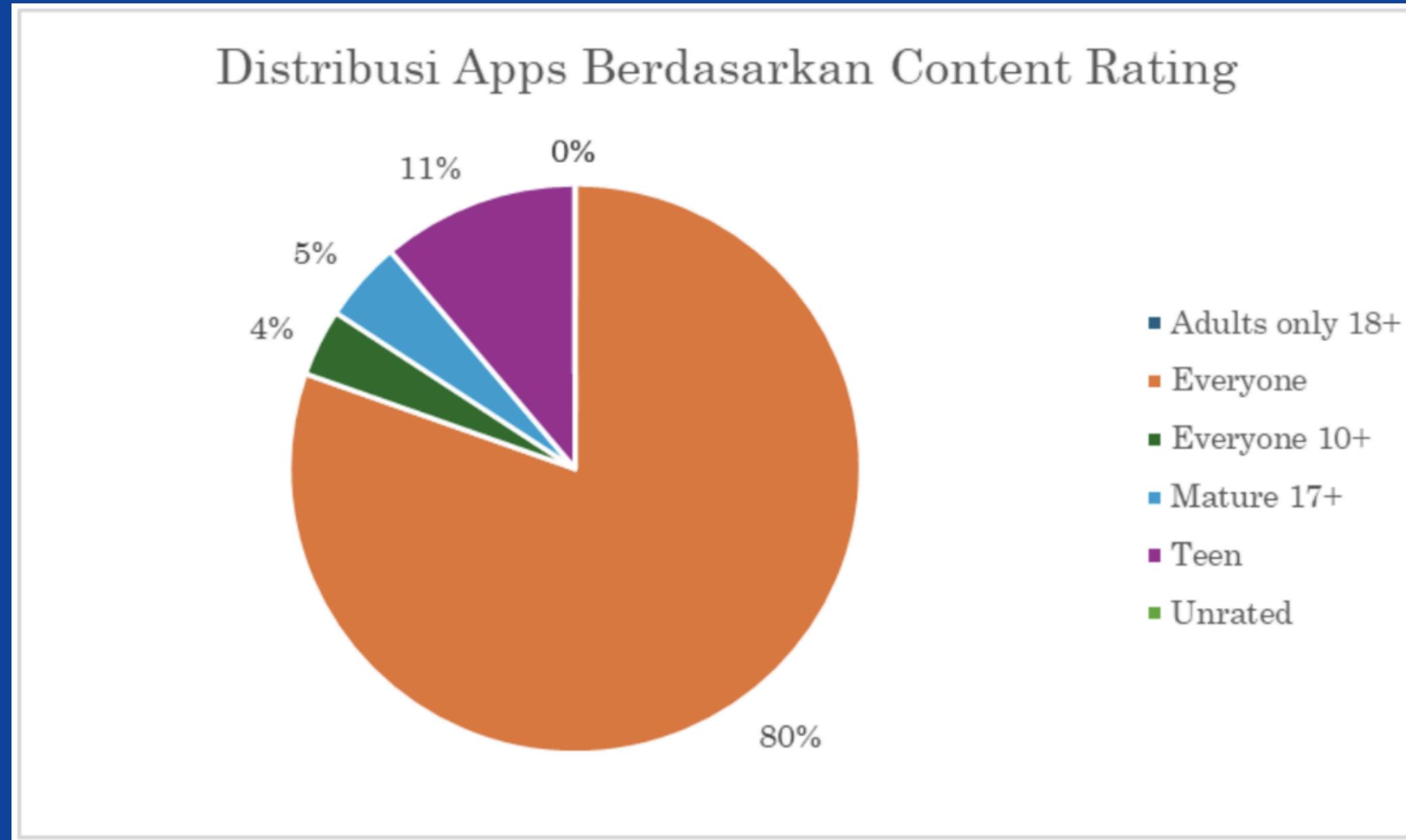
Lonjakan tajam jumlah aplikasi yang memiliki pembaruan terakhir di tahun 2018 mencerminkan kondisi aktual saat data diambil, yakni sekitar tahun tersebut, sehingga **majoritas aplikasi yang masih aktif atau relevan pada saat itu menunjukkan status last updated di 2018**. Kecenderungan ini mengindikasikan bahwa banyak pengembang melakukan pembaruan rutin untuk menjaga kompatibilitas aplikasi mereka dengan versi Android terbaru dan mengikuti standar keamanan serta performa terkini yang berlaku saat itu.

3.1.4 Visualisasi 4: Rata-rata Rating Berdasarkan Tahun Pembaruan Terakhir



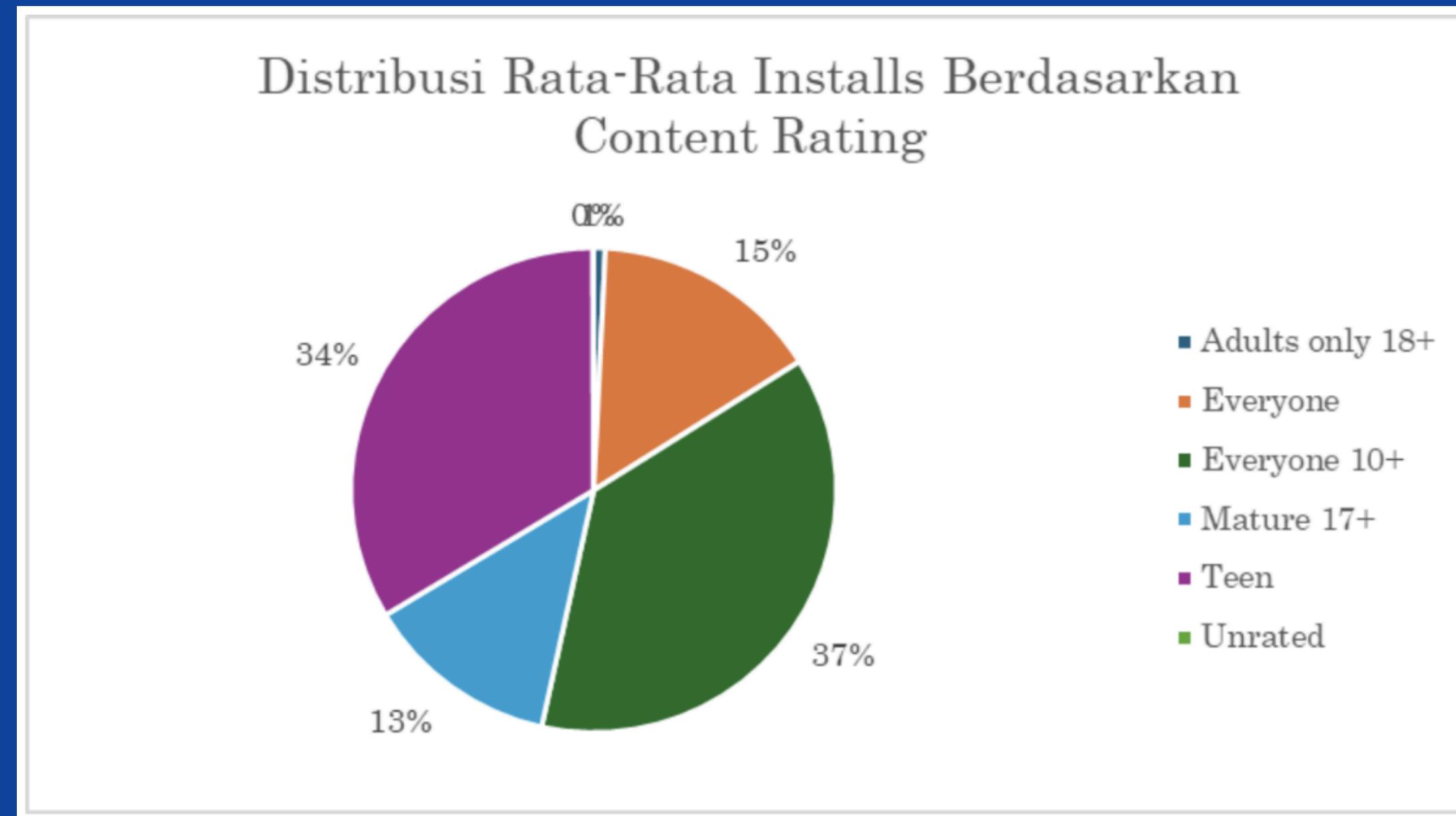
Grafik rata-rata rating berdasarkan tahun pembaruan terakhir menunjukkan adanya tren peningkatan kualitas aplikasi dari waktu ke waktu, dengan titik terendah terjadi pada tahun 2012 dan lonjakan signifikan setelahnya hingga mencapai puncak pada 2018. Pola ini mengindikasikan bahwa aplikasi yang lebih sering diperbarui atau masih aktif dikelola cenderung memiliki rating yang lebih tinggi dibanding aplikasi lama yang tidak diperbarui, menandakan bahwa pembaruan rutin kemungkinan besar meningkatkan kepuasan pengguna. Hal ini memperkuat asumsi bahwa developer yang terus melakukan perbaikan dan penyesuaian terhadap kebutuhan pengguna cenderung memperoleh kepercayaan dan respons positif yang lebih besar di platform.

3.1.5 Visualisasi 5: Distribusi Apps Berdasarkan Content Rating



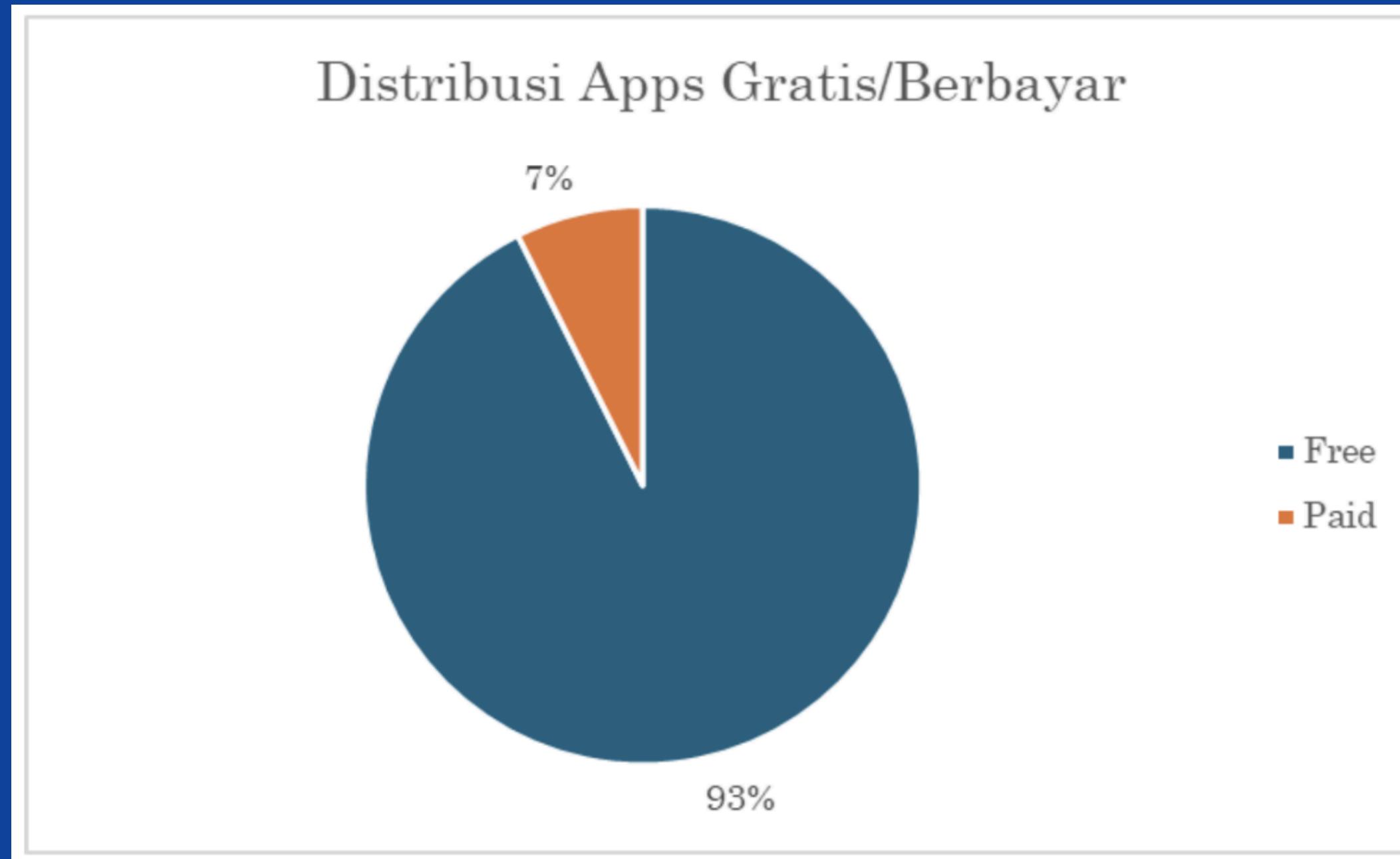
Berdasarkan pie chart distribusi aplikasi menurut content rating, terlihat bahwa **majoritas besar aplikasi di Google Play Store ditujukan untuk kategori "Everyone"**, yaitu sebesar 80% dari total aplikasi. Sementara itu, kategori lain seperti "Teen" (11%), "Mature 17+" (5%), dan "Everyone 10+" (4%) memiliki porsi jauh lebih kecil. Kategori "Adults only 18+" dan "Unrated" hampir tidak signifikan, masing-masing menyumbang 0% dari total. Hal ini mengindikasikan bahwa ekosistem aplikasi di platform ini sangat berfokus pada konten yang dapat diakses oleh pengguna dari berbagai usia, khususnya anak-anak dan keluarga, menunjukkan arah pengembangan aplikasi yang inklusif dan ramah untuk semua kalangan.

3.1.6 Visualisasi 6: Distribusi Rata-rata Installs Berdasarkan Content Rating



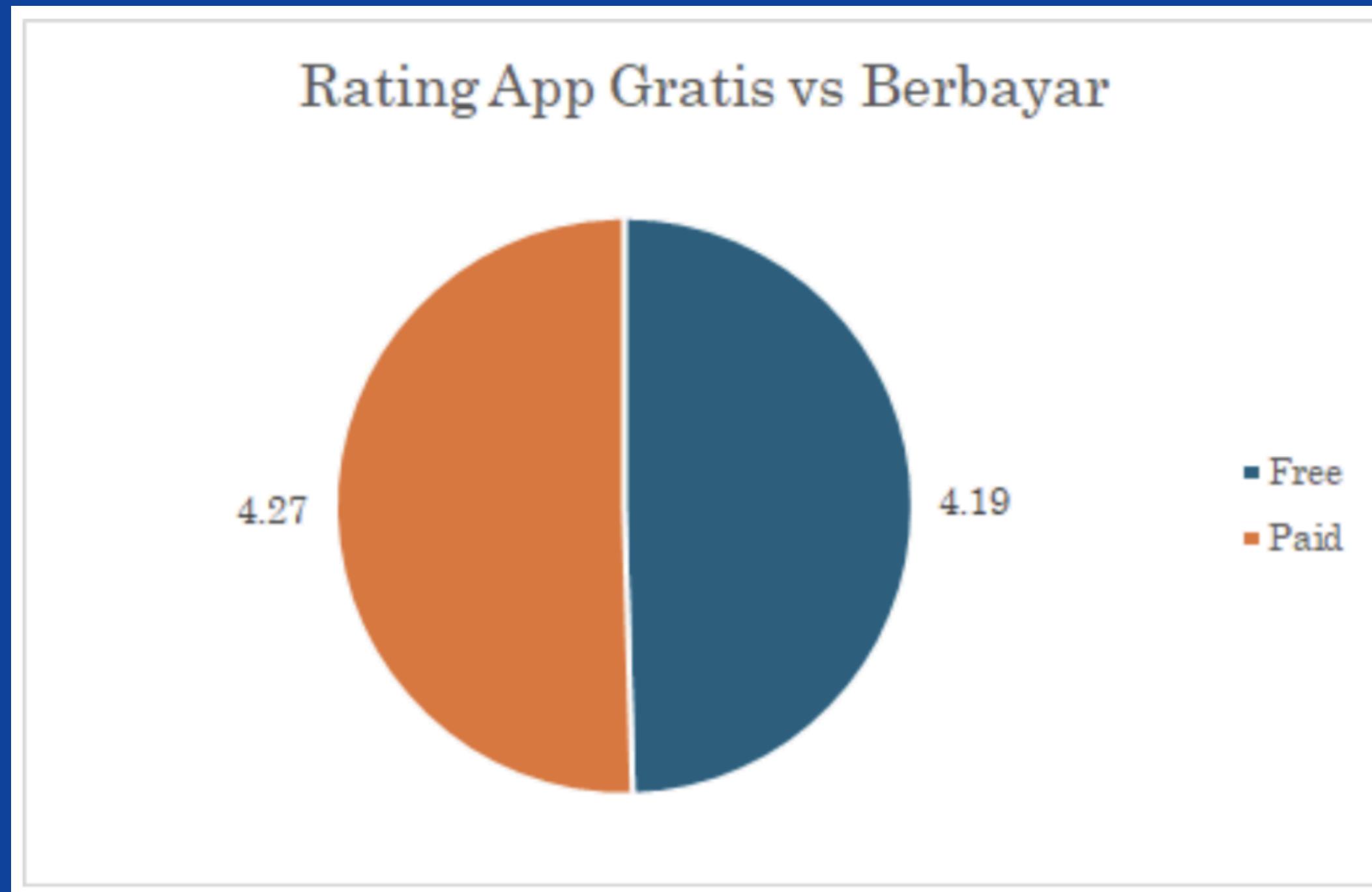
Meskipun jumlah aplikasi dengan rating "Everyone" mendominasi (seperti terlihat di pie chart sebelumnya), justru **kategori "Everyone 10+" dan "Teen"** memiliki **rata-rata instalasi tertinggi per aplikasi**, menandakan bahwa aplikasi-aplikasi yang ditujukan untuk remaja dan usia 10 tahun ke atas lebih menarik atau lebih sering diunduh oleh pengguna, dibandingkan aplikasi yang ditujukan untuk semua umur secara umum.

3.1.7 Visualisasi 7: Distribusi Apps Gratis/Berbayar



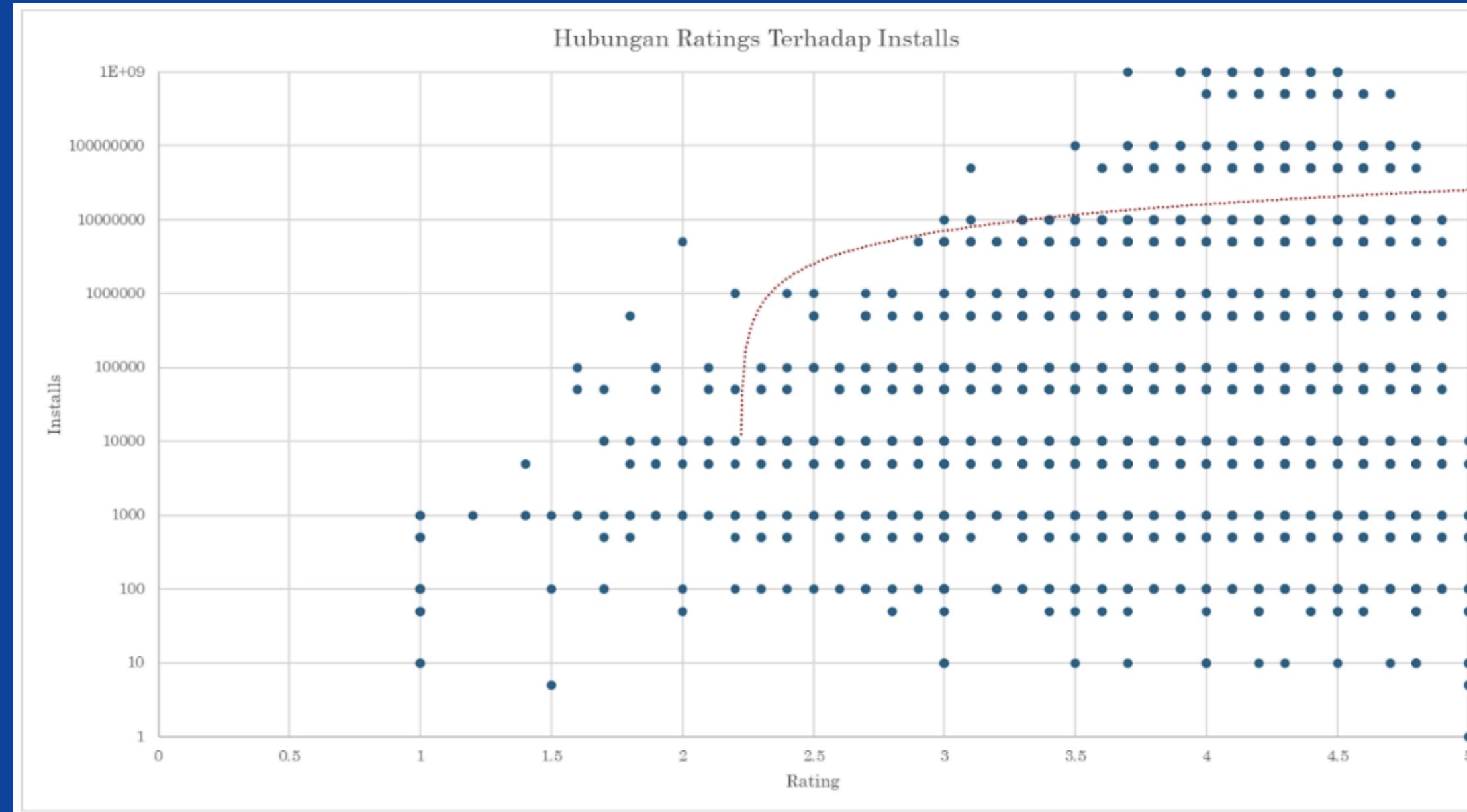
Grafik ini menunjukkan bahwa mayoritas aplikasi di Google Play Store adalah aplikasi gratis, yakni sebesar 93%, sementara hanya 7% aplikasi yang berbayar. Hal ini menandakan bahwa model distribusi aplikasi di platform ini sangat mengandalkan pendekatan freemium atau monetisasi berbasis iklan dan pembelian dalam aplikasi (in-app purchase), dibandingkan dengan model pembayaran langsung di muka. Dominasi aplikasi gratis juga mencerminkan preferensi pasar pengguna Android yang cenderung lebih memilih aplikasi tanpa biaya awal, sehingga pengembang lebih terdorong untuk menarik pengguna lewat akses gratis sebagai strategi pertumbuhan awal.

3.1.8 Visualisasi 8: Rating App Gratis vs Berbayar



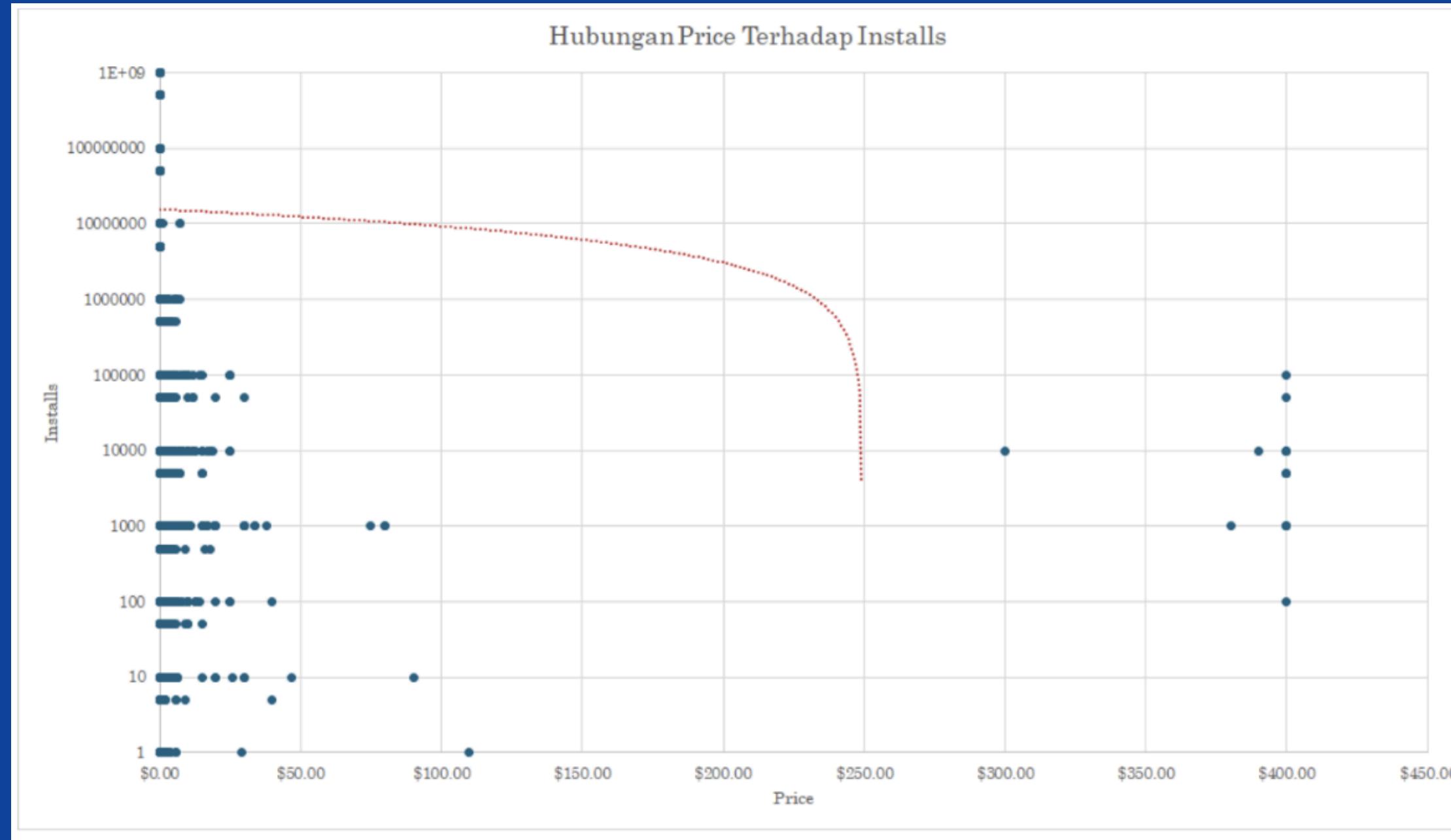
Meskipun aplikasi gratis mendominasi jumlah di Google Play Store, grafik ini menunjukkan bahwa **aplikasi berbayar memiliki rata-rata rating yang sedikit lebih tinggi dibandingkan aplikasi gratis – yaitu 4.27 untuk berbayar versus 4.19 untuk gratis**. Hal ini bisa menunjukkan bahwa pengguna cenderung memberikan penilaian lebih tinggi terhadap aplikasi berbayar, mungkin karena ekspektasi yang lebih tinggi sejalan dengan kualitas, fitur premium, atau pengalaman pengguna yang lebih baik yang mereka dapatkan. Sementara itu, aplikasi gratis yang tersedia dalam jumlah besar bisa jadi memiliki kualitas yang bervariasi, sehingga berdampak pada rating rata-ratanya.

3.1.9 Visualisasi 9: Hubungan Ratings Terhadap Installs



Dari data yang disajikan, dapat disimpulkan bahwa terdapat **korelasi positif antara tinggi rendahnya ratings dengan jumlah installs**. Semakin tinggi rating yang diberikan pengguna, semakin besar pula kemungkinan aplikasi tersebut untuk diunduh. Hal ini menunjukkan bahwa reputasi dan kepuasan pengguna, yang tercermin dari ratings, menjadi faktor krusial dalam menarik minat pengguna baru untuk menginstal aplikasi. Oleh karena itu, pengembang perlu memprioritaskan kualitas dan pengalaman pengguna untuk mencapai ratings yang tinggi, yang pada akhirnya akan mendorong peningkatan jumlah installs.

3.1.10 Visualisasi 10: Hubungan Price Terhadap Installs



Data menunjukkan bahwa **harga memiliki dampak ekstrem terhadap jumlah installs**. Aplikasi gratis (Rp0) mendominasi dengan angka mencapai 1 miliar installs. Namun, jumlah ini anjlok drastis menjadi sekitar 10 juta saat harga aplikasi mencapai Rp50, dan terus merosot hingga hanya puluhan installs di kisaran harga Rp400. Pola ini mengonfirmasi tingginya sensitivitas pasar aplikasi mobile terhadap harga, dengan batas toleransi yang jelas berada di bawah Rp50. Temuan ini sekaligus membuktikan superioritas model bisnis freemium dalam menarik pengguna massal dibandingkan dengan model berbayar konvensional.

Visualisasi

Data Video Game Terpopuler 1980 - 2023

Garam & Madu

Pada metode visualisasi yang kami gunakan di bagian ini, ada beberapa hal yang perlu diperhatikan:

1. Pembuatan dilakukan pada Ms. Excel dengan database terdiri atas 5 sheets, yaitu
 - a. SheetUtama, berisikan data mentah yang diambil dari sumber data langsung dalam format file .csv yang sudah saya konversikan menjadi .xlsx.
 - b. PerbandinganKategori, bersihkan data dan visualisasi untuk perbandingan kategori.
 - c. PerubahanterhadapWaktu, bersihkan data dan visualisasi untuk penampilan perubahan terhadap waktu.
 - d. PenampilanHierarkis, berisikan dan visualisasi untuk penampilan hierarki dan hubungan keseluruhan bagian.
 - e. PlottingRelationships, berisikan dan visualisasi untuk plotting relationships.
1. Angka Plays merujuk pada jumlah seluruh pemain yang pernah memainkan video game sebelumnya.
2. Angka Playing merujuk pada jumlah pemain aktif, yaitu pemain yang masih memainkan video game hingga sekarang.
3. Angka Backlogs merujuk pada jumlah pemain pasif, yaitu pemain yang memiliki akses terhadap video game, tetapi belum memainkannya.
4. Pada data mentah, kolom Genre terdiri atas beberapa genre. Oleh karena itu, kami mengambil genre utama saja dari video game tersebut dengan mengambil genre yang disebut pertama kali dengan formula berikut

```
=MID(A1, FIND("'", A1)+1, FIND("'", A1, FIND("'", A1)+1) - FIND("'", A1) - 1)
```

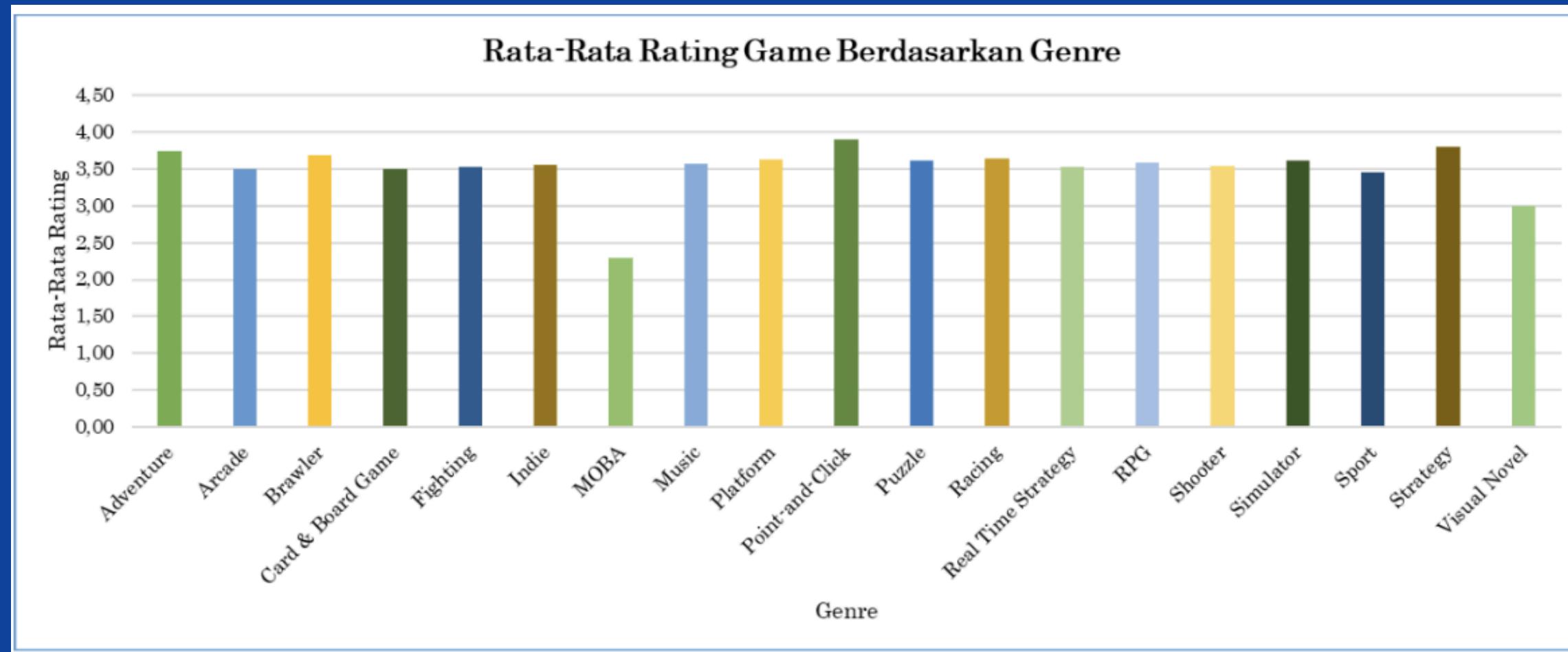
sebab contoh formatnya adalah ['Adventure', 'RPG', 'Turn Based Strategy'].

1. Pada data mentah, kolom Times Listed, Number of Reviews, Plays, Playing, Backlogs, dan Wishlist memiliki data yang berakhiran K, M, dan B yang menunjukkan ribuan, jutaan, dan miliaran. Oleh karena itu, saya mengolah format tersebut menjadi format number dengan formula berikut.

```
=IF(ISNUMBER(K2); K2;  
    IF(RIGHT(K2;1)="K"; VALUE(SUBSTITUTE(LEFT(K2;LEN(K2)-1); "."); ",")) * 1000;  
    IF(RIGHT(K2;1)="M"; VALUE(SUBSTITUTE(LEFT(K2;LEN(K2)-1); "."); ",")) * 1000000;  
    IF(RIGHT(K2;1)="B"; VALUE(SUBSTITUTE(LEFT(K2;LEN(K2)-1); "."); ",")) * 1000000000;  
    VALUE(SUBSTITUTE(K2; ","; "."))))))
```

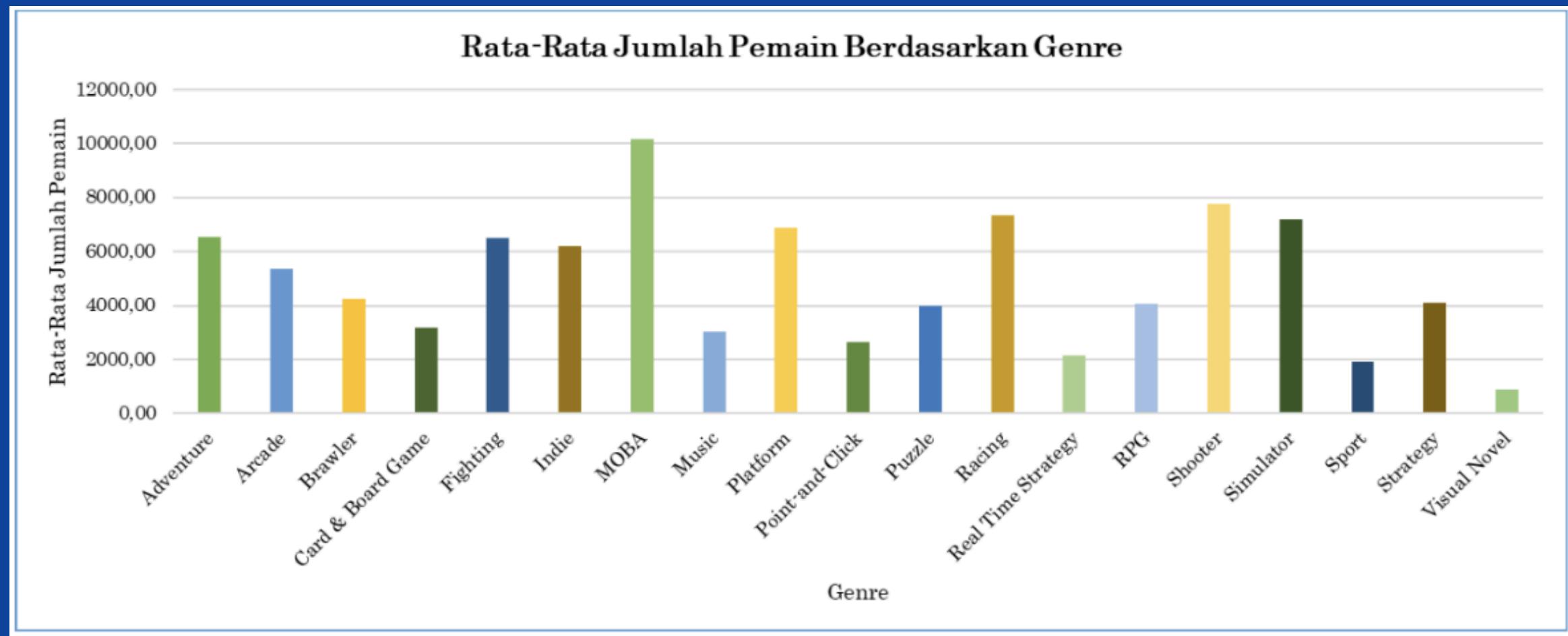
Perlu diperhatikan bahwa pada Ms. Excel yang kami gunakan, default format decimal menggunakan "," bukan "." sehingga perlu dilakukan substitusi.

3.2.1 Visualisasi 1: Rata-Rata Rating Game Berdasarkan Genre



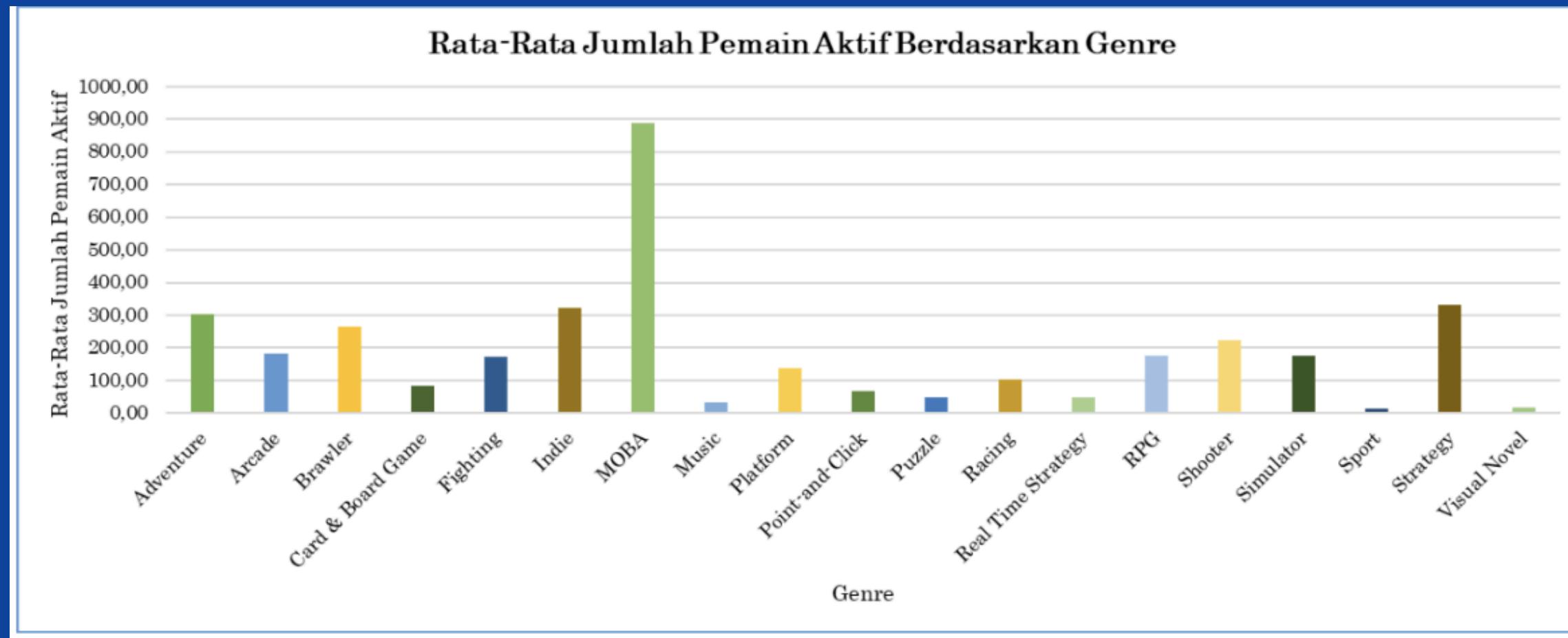
Berdasarkan diagram, **video game dengan rating tertinggi memiliki genre Point-and-Click** disusul oleh video game dengan Strategy dan Adventure, pada posisi 2 dan 3, berturut-turut. Genre game dengan rating paling rendah adalah MOBA. Secara umum, rata-rata rating game berada di sekitar angka 3,50.

3.2.2 Visualisasi 2: Rata-Rata Jumlah Pemain Berdasarkan Genre



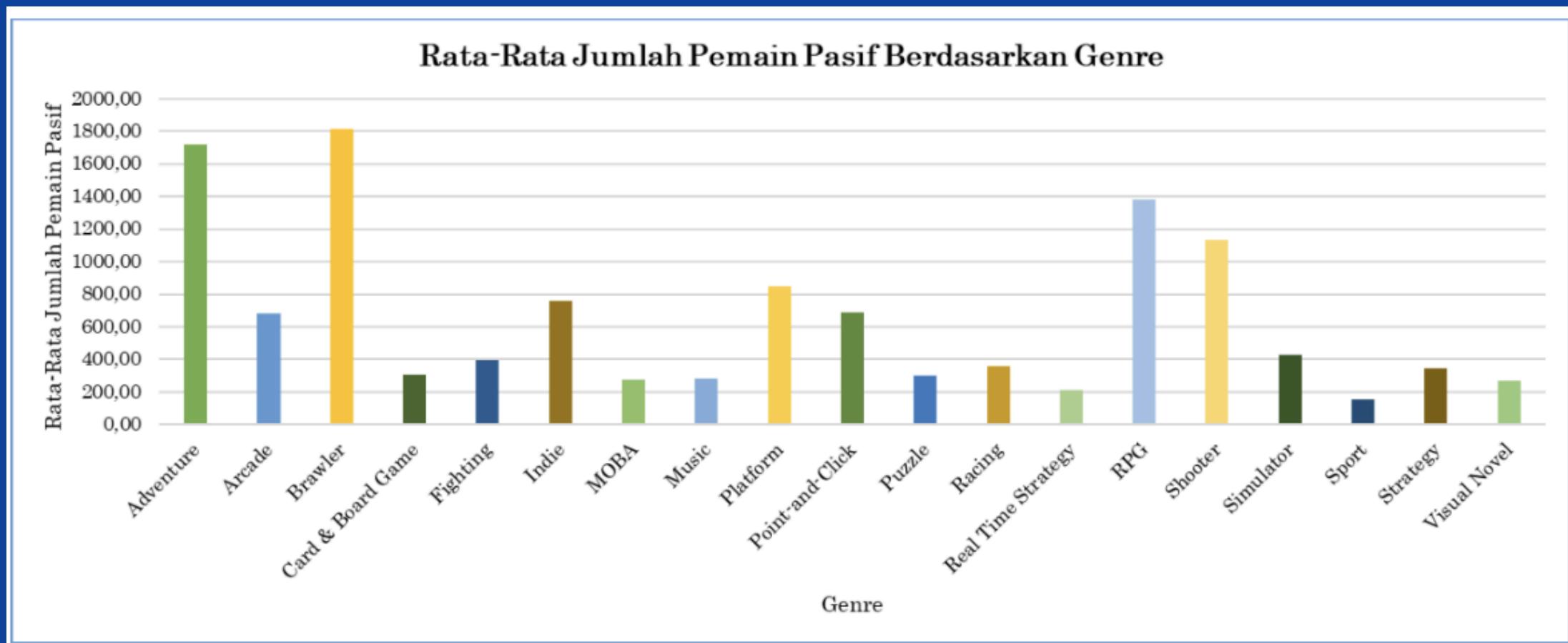
Berdasarkan grafik, **MOBA adalah genre video game dengan rata-rata jumlah pemain paling tinggi** disusul oleh game dengan genre Shooter dan Racing. Tiga genre video game dengan rata-rata jumlah pemain terendah adalah Real-Time Strategy, Sport, dan Visual Novel.

3.2.3 Visualisasi 3: Rata-Rata Jumlah Pemain Aktif Berdasarkan Genre



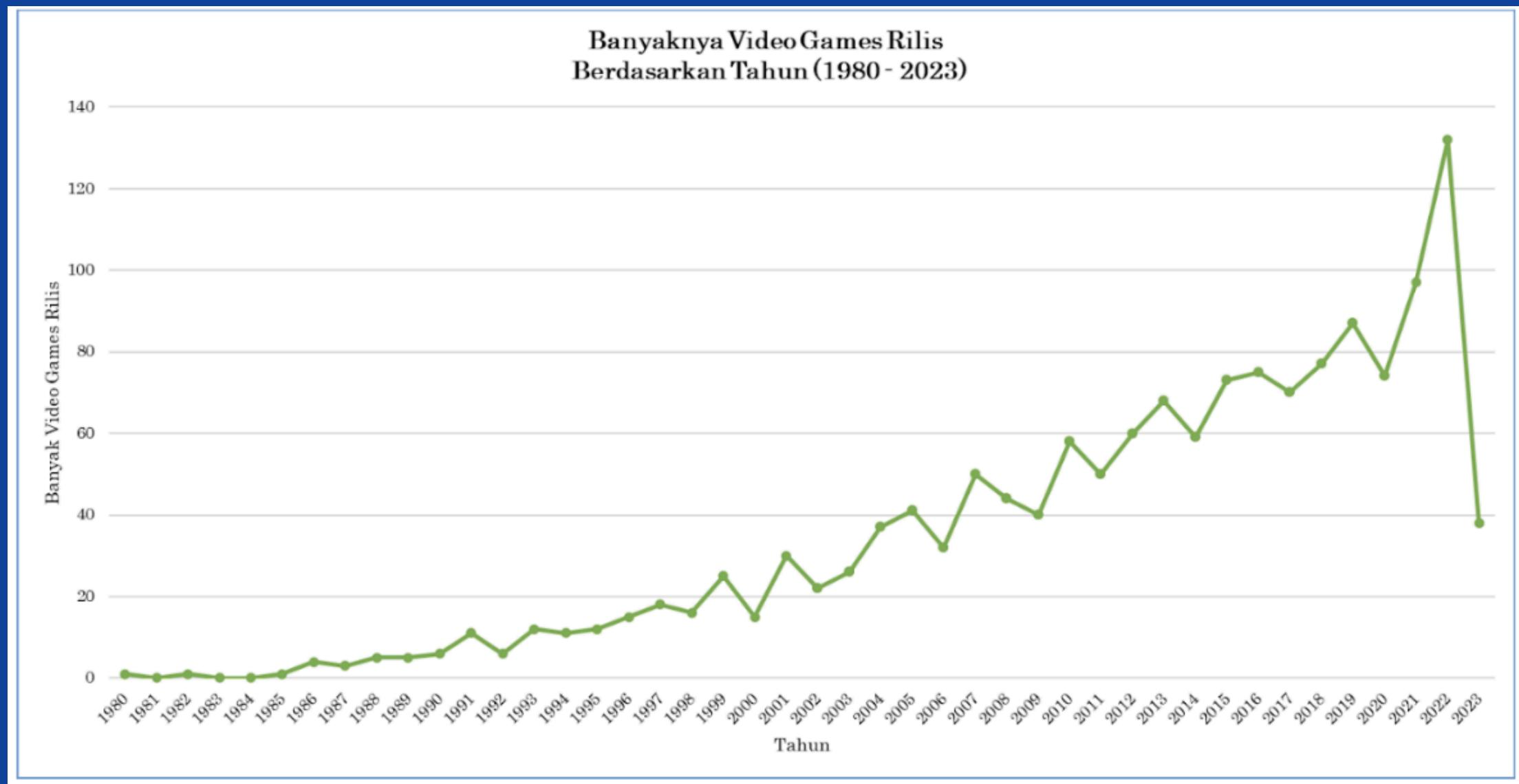
Berdasarkan diagram, **MOBA adalah genre video game dengan rata-rata jumlah pemain aktif paling tinggi di angka hampir 900**. Hal ini berbeda jauh dengan rata-rata jumlah pemain aktif pada video genre game lainnya. Pada posisi kedua, genre video game Strategy memiliki rata-rata jumlah pemain aktif hanya di sekitar angka 300. Rata-rata jumlah pemain aktif terendah terdapat pada video game bergenre Music, Sport, dan Visual Novel.

3.2.4 Visualisasi 4: Rata-Rata Jumlah Pemain Pasif Berdasarkan Genre



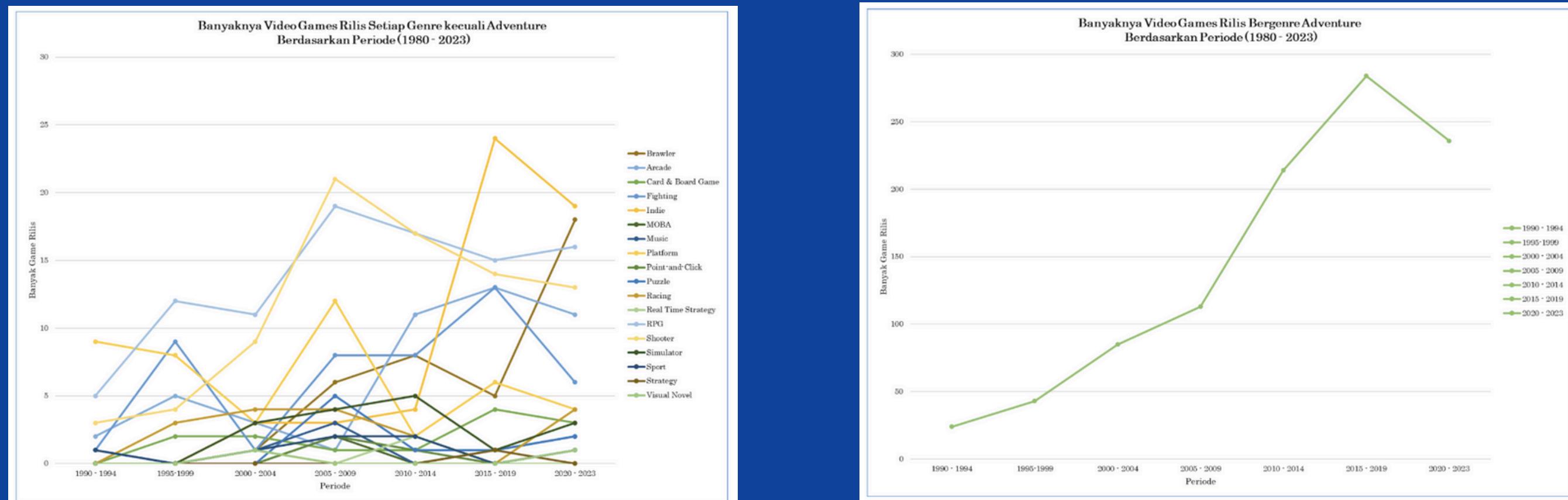
Berdasarkan diagram, **Brawler** adalah **genre video game dengan rata-rata jumlah pemain pasif tertinggi di angka 1800**, diikuti oleh Adventure dan RPG, pada posisi ke-2 dan ke-3 secara berturut-turut. Video game dengan rata-rata jumlah pemain pasif terendah adalah game dengan genre Sport. Secara umum, terdapat perbedaan yang mencolok pada angka rata-rata video game bergenre Brawler, Adventure, RPG, dan Shooter dengan genre lainnya.

3.2.5 Visualisasi 5: Banyaknya Video Games Rilis Berdasarkan Tahun (1980 - 2023)



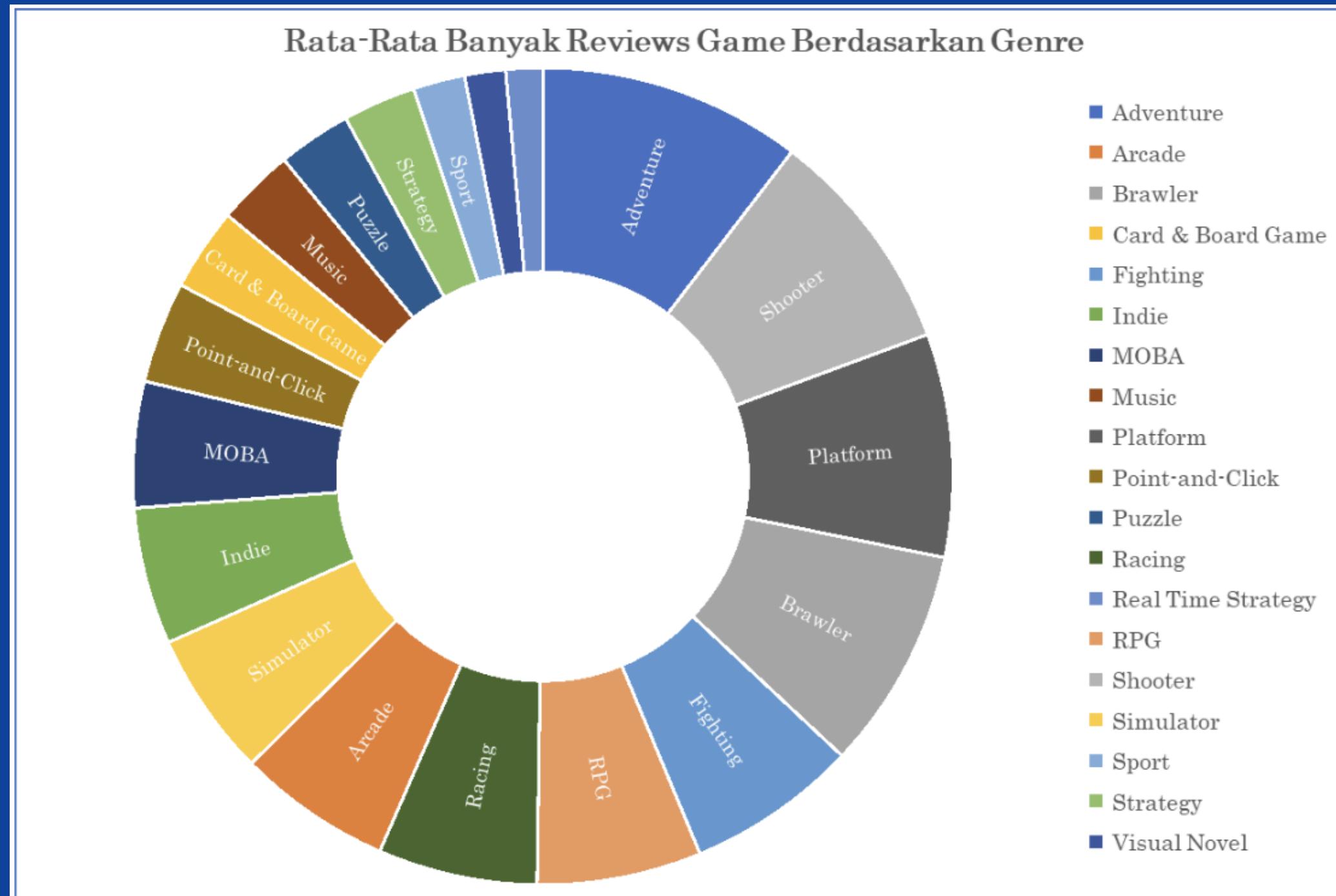
Berdasarkan grafik, **peningkatan video game yang dirilis tertinggi pada tahun 2020 menuju tahun 2021**. Selain itu, dapat dilihat bahwa banyak video game yang dirilis dari tahun ke tahun meningkat secara fluktuatif. Ada tahun-tahun saat video game mengalami penurunan, tetapi penurunan minor ini tidak sebanding dengan peningkatan video game yang dirilis dari waktu ke waktu.

3.2.6 Visualisasi 6: Banyaknya Video Games Rilis Setiap Genre Berdasarkan Periode (1980 - 2023)



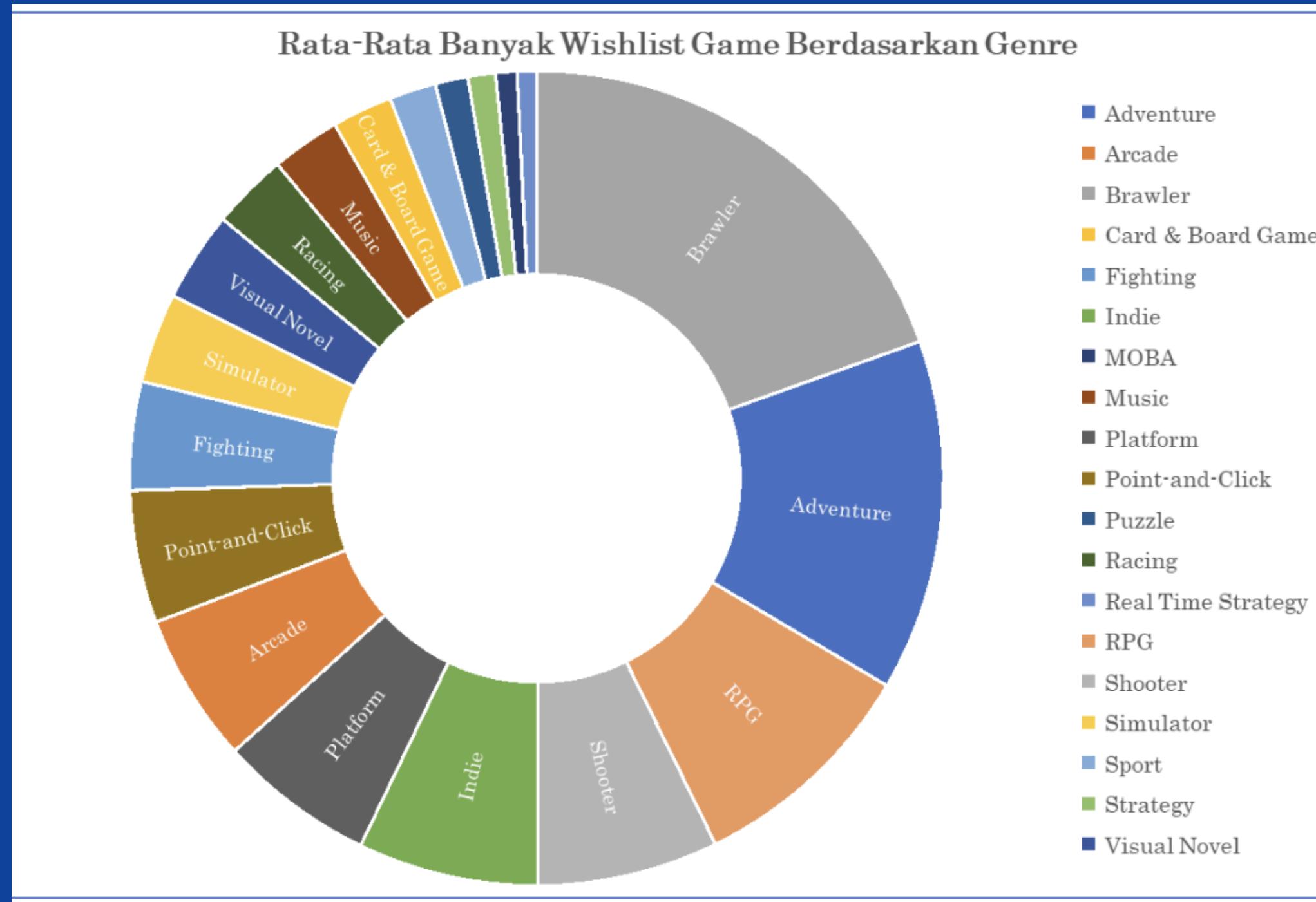
Berdasarkan grafik, **video game bergenre Adventure mendominasi rilis dari periode satu ke periode lainnya**. Selanjutnya, beberapa hal yang menarik perhatian adalah video game bergenre Indie yang mengalami peningkatan angka perilisan drastis dari periode 5 tahun, 2010 - 2014 ke 2015 - 2019. Banyak video game yang rilis pada setiap genre memiliki tren fluktuatif yang tidak bisa diprediksi, kecuali pada genre Adventure yang monoton naik dengan mempertimbangkan bahwa pada periode terakhir baru berlangsung selama 4 tahun.

3.2.7 Visualisasi 7: Rata-Rata Banyak Reviews Berdasarkan Genre



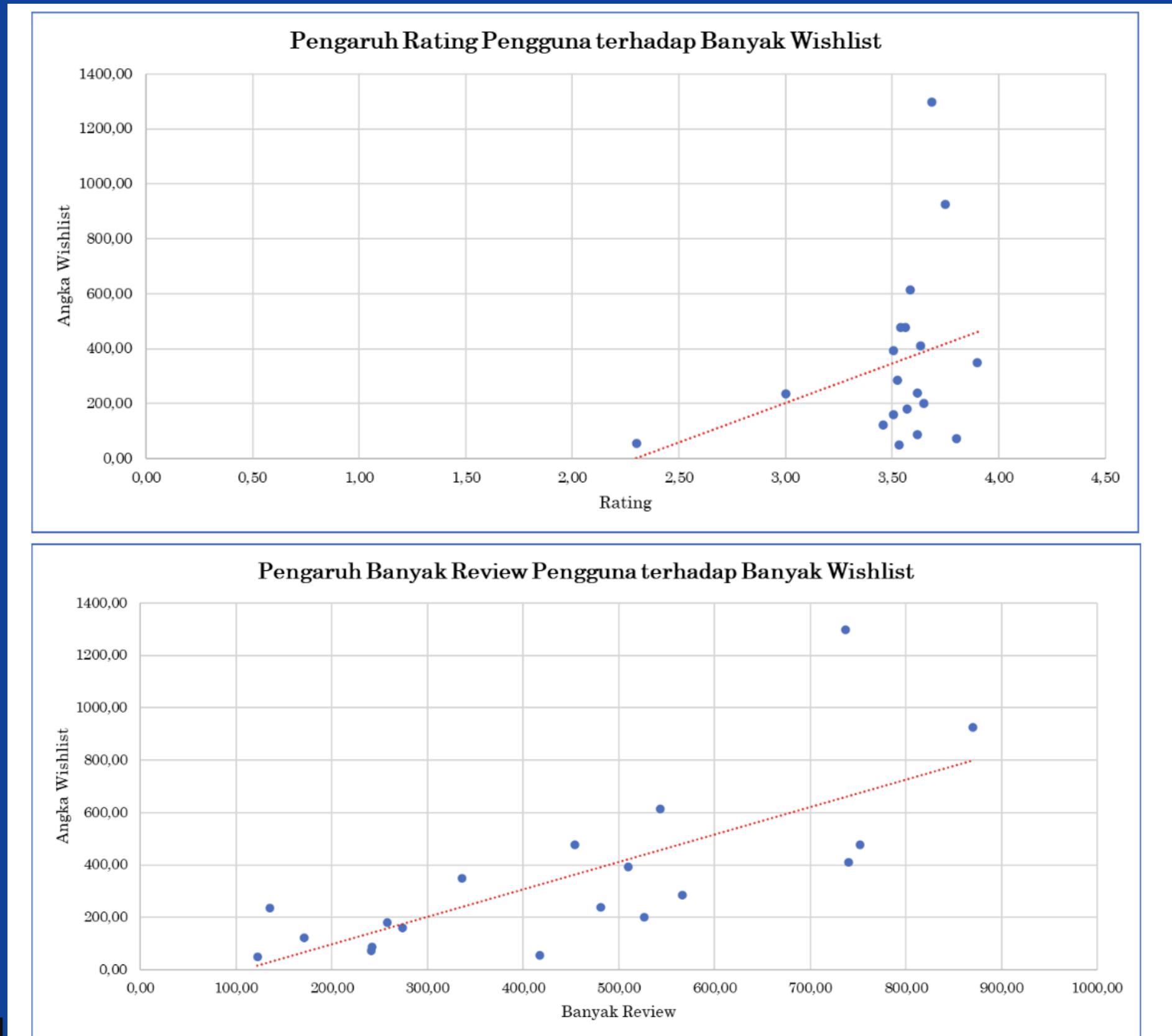
Berdasarkan diagram, **video game bergenre Adventure memiliki rata-rata reviews paling tinggi** di antara genre lainnya. Hal ini kemudian disusul oleh video game bergenre Shooter dan Platform, pada peringkat 2 dan 3, secara berturut-turut. Video game bergenre Visual Novel dan Real Time Strategy memiliki rata-rata banyak review paling rendah di antara genre lainnya.

3.2.8 Visualisasi 8: Rata-Rata Banyak Wishlist Game Berdasarkan Genre



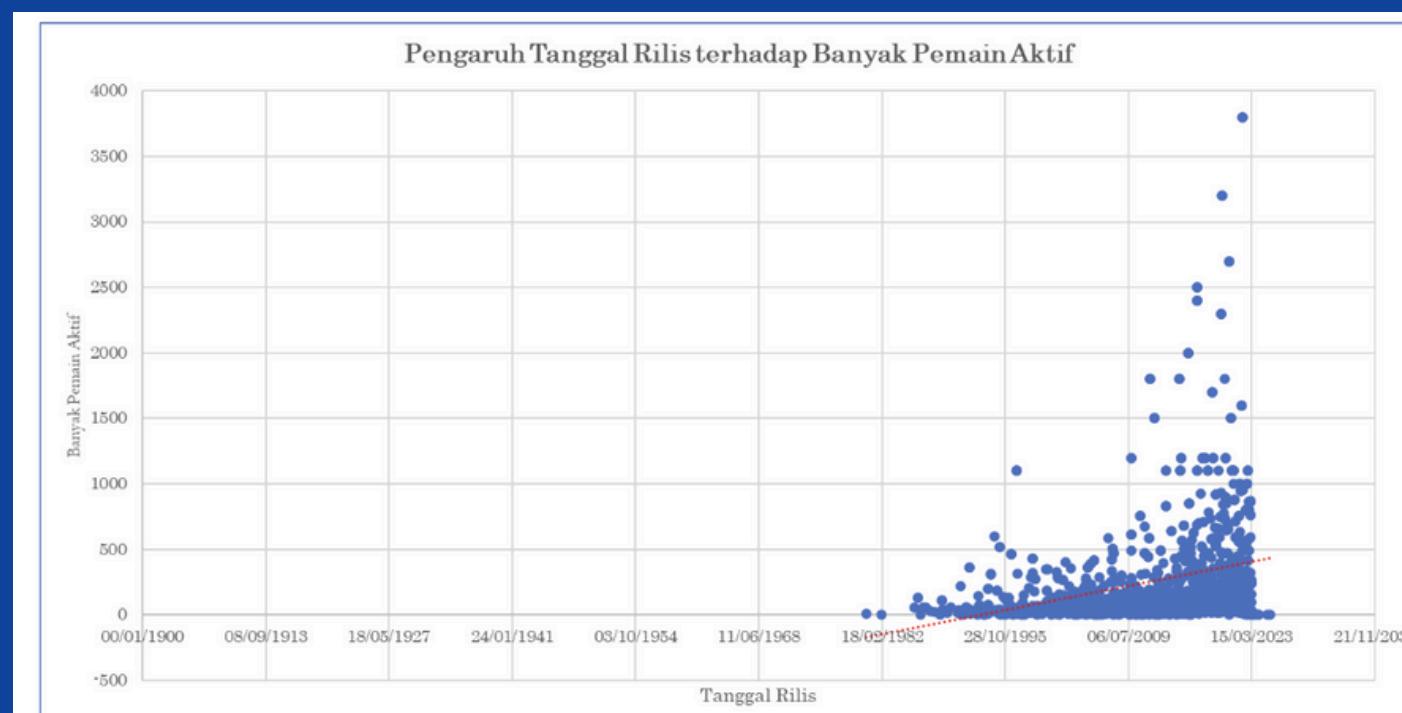
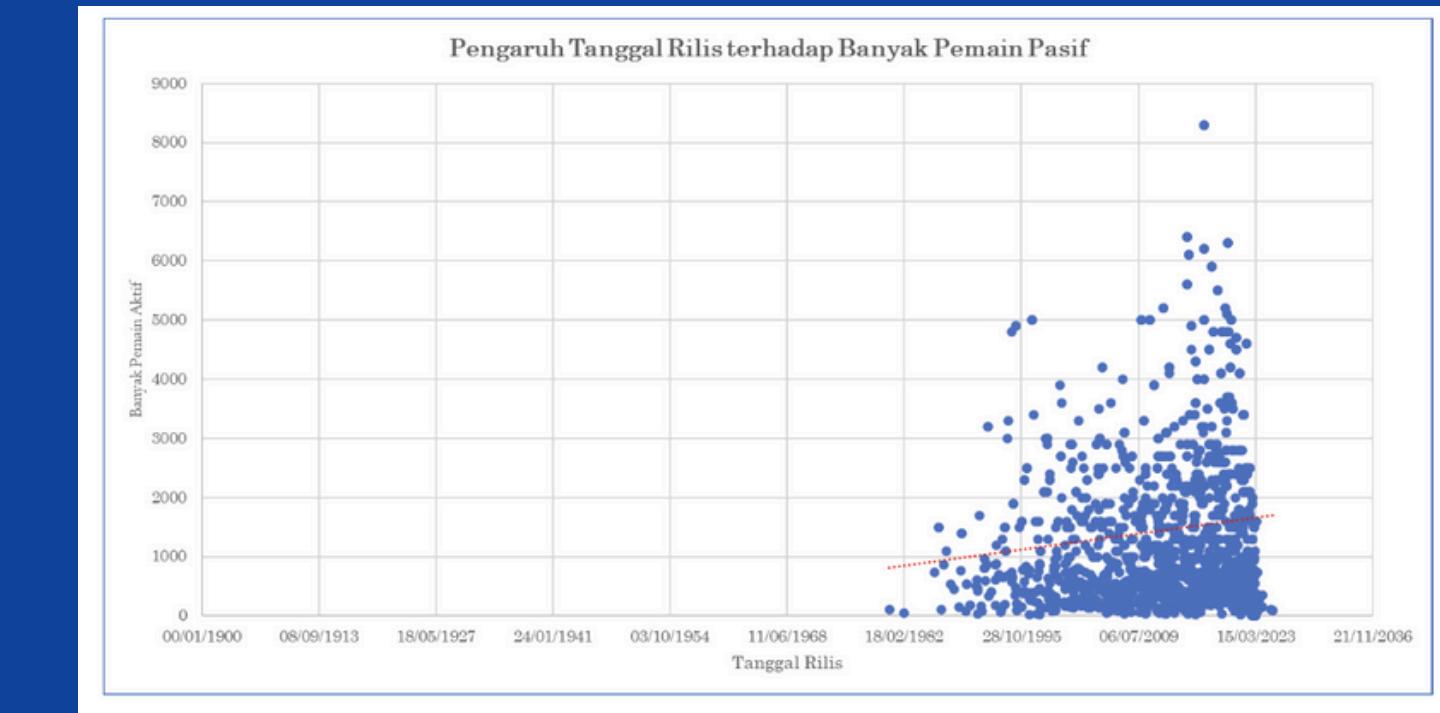
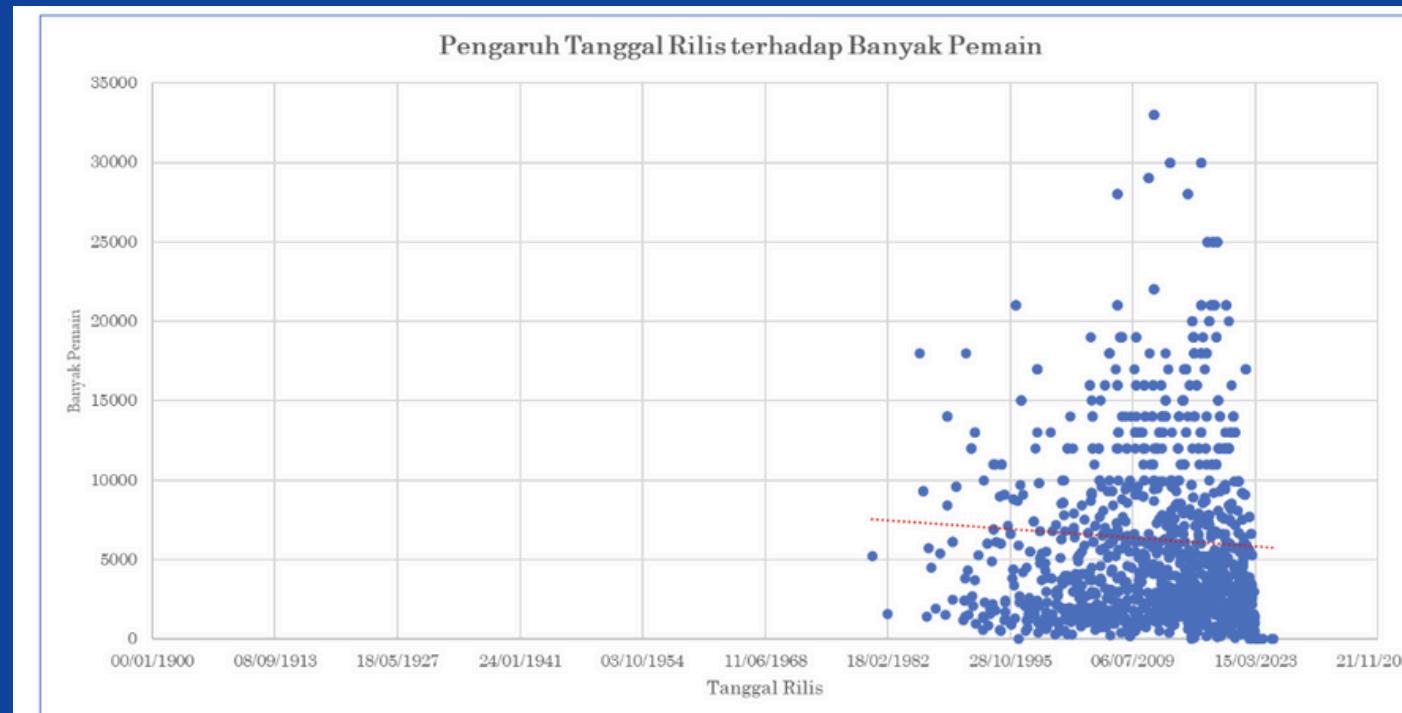
Berdasarkan diagram, **banyak pemain video game yang memasukkan video game bergenre Brawler ke dalam wishlist mereka**, disusul oleh video game bergenre Adventure, RPG, dan Shooter pada tiga peringkat setelahnya. Beberapa video game yang jarang dimasukkan ke dalam wishlist adalah genre Sport, Visual Novel, dan Strategy.

3.2.9 Visualisasi 9: Variabel Bebas yang Memengaruhi Angka Wishlist



Berdasarkan kedua diagram, dapat disimpulkan bahwa **rating pengguna dan banyak review memengaruhi banyak wishlist**. Hal ini ditunjukkan dengan adanya korelasi positif antara kedua pasangan variabel bebas dan terikat yang ada.

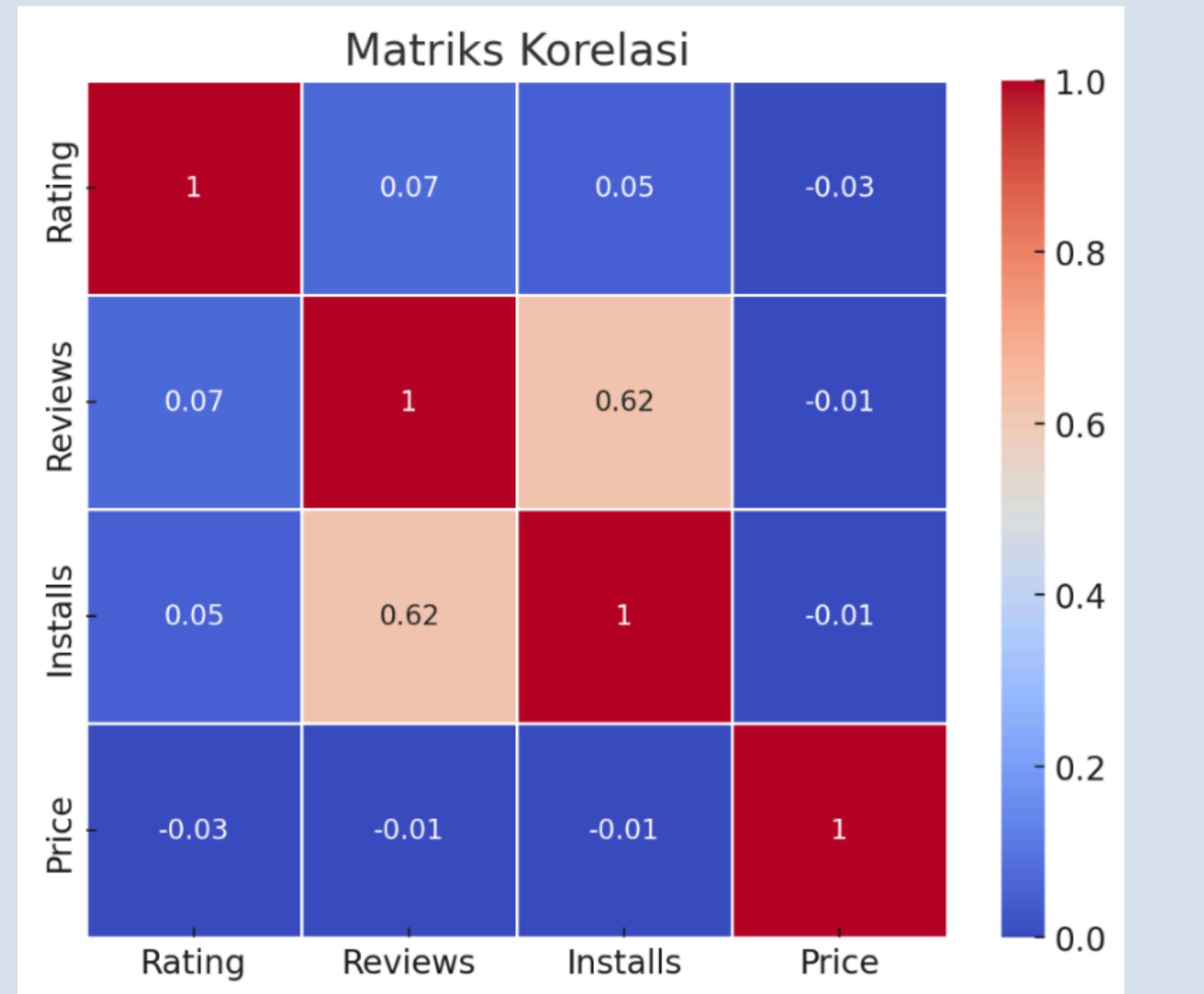
3.2.10 Visualisasi 10: Pengaruh Tanggal Rilis Terhadap Banyak Pemain



Berdasarkan ketiga diagram, diketahui bahwa **tanggal rilis berkorelasi negatif dengan banyak pemain**. Akan tetapi, hal ini berbeda dengan banyak pemain aktif dan pasif. Pada kasus tersebut, tanggal rilis berkorelasi positif.

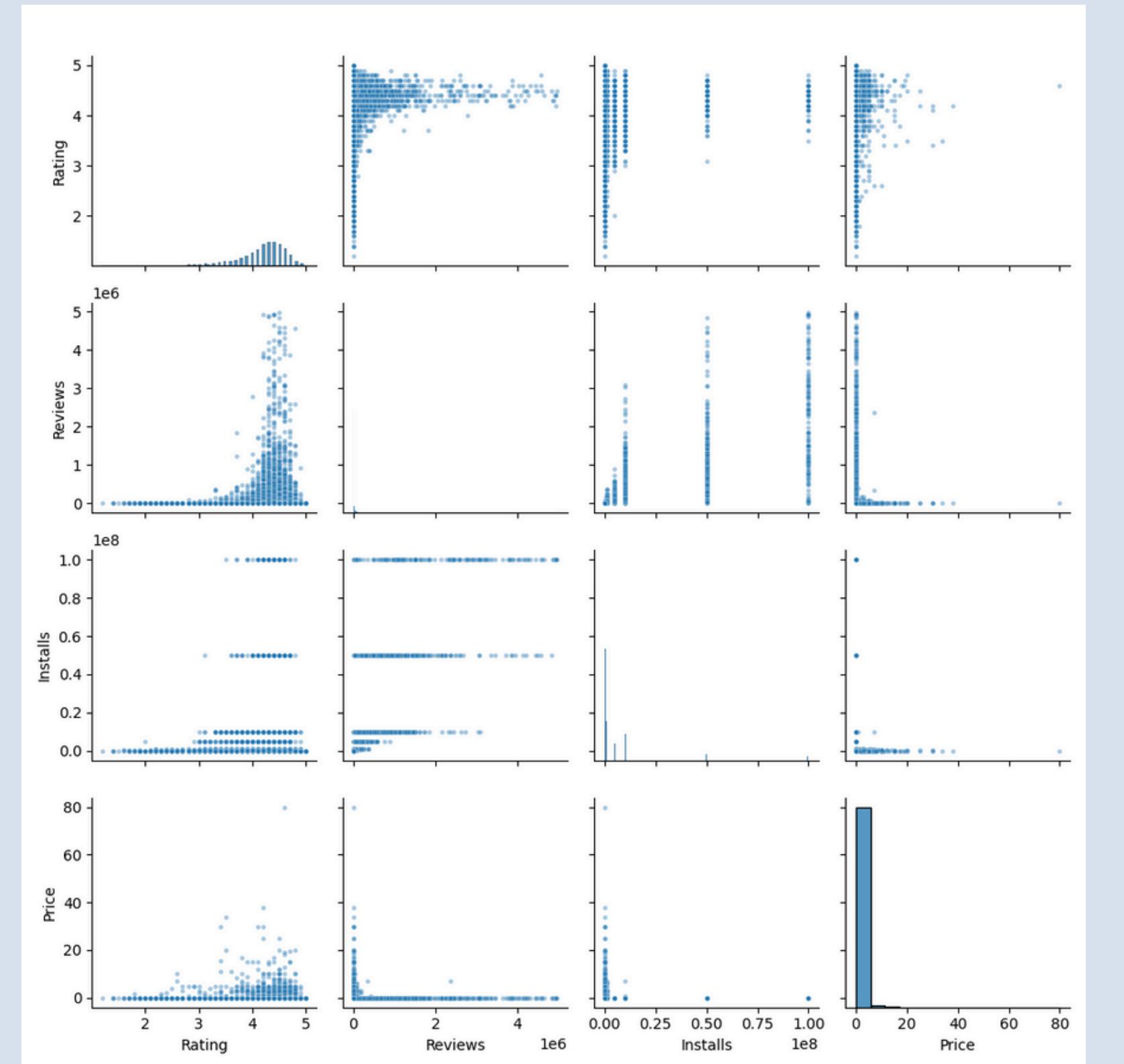
Korelasi

Google Play Store

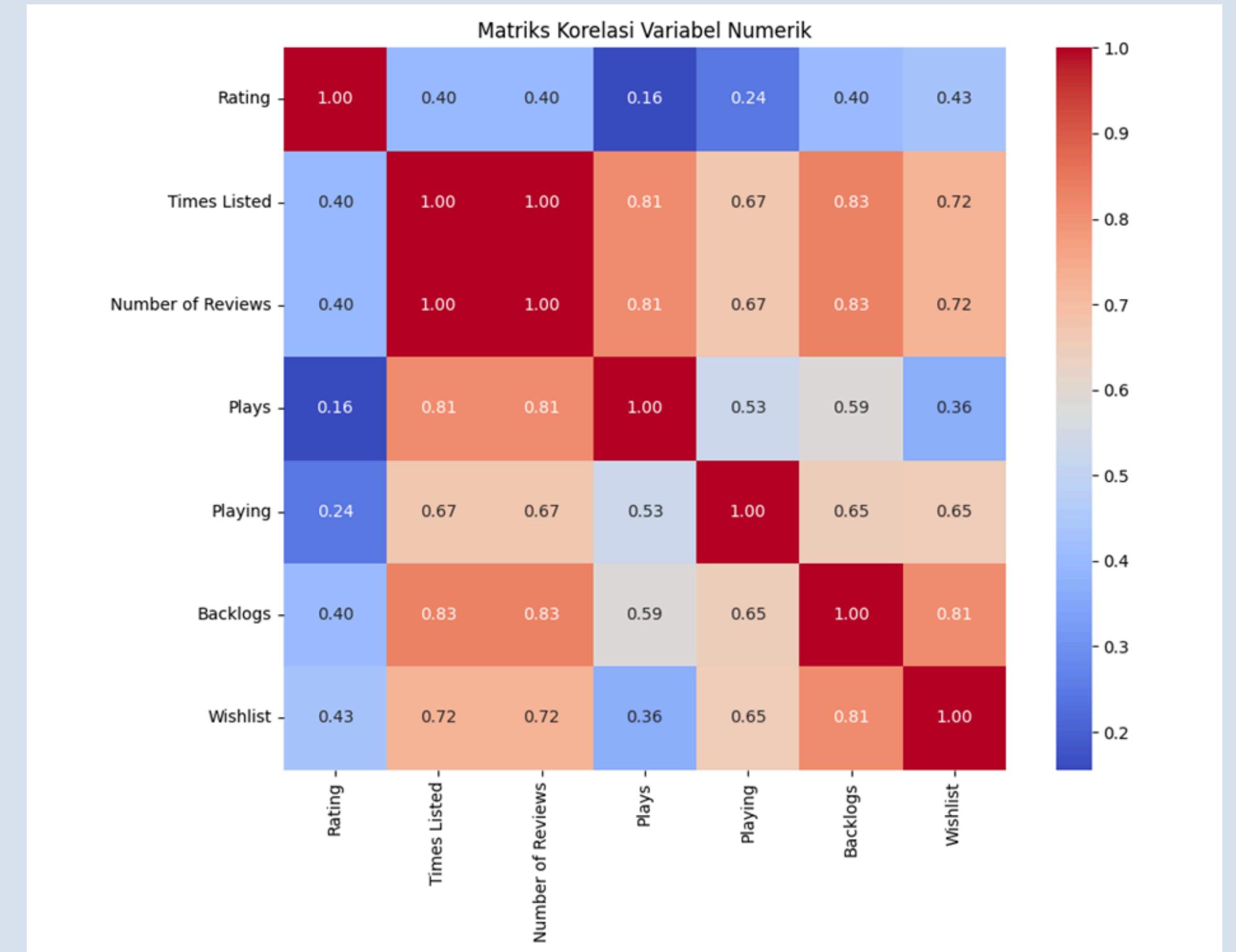


Korelasi

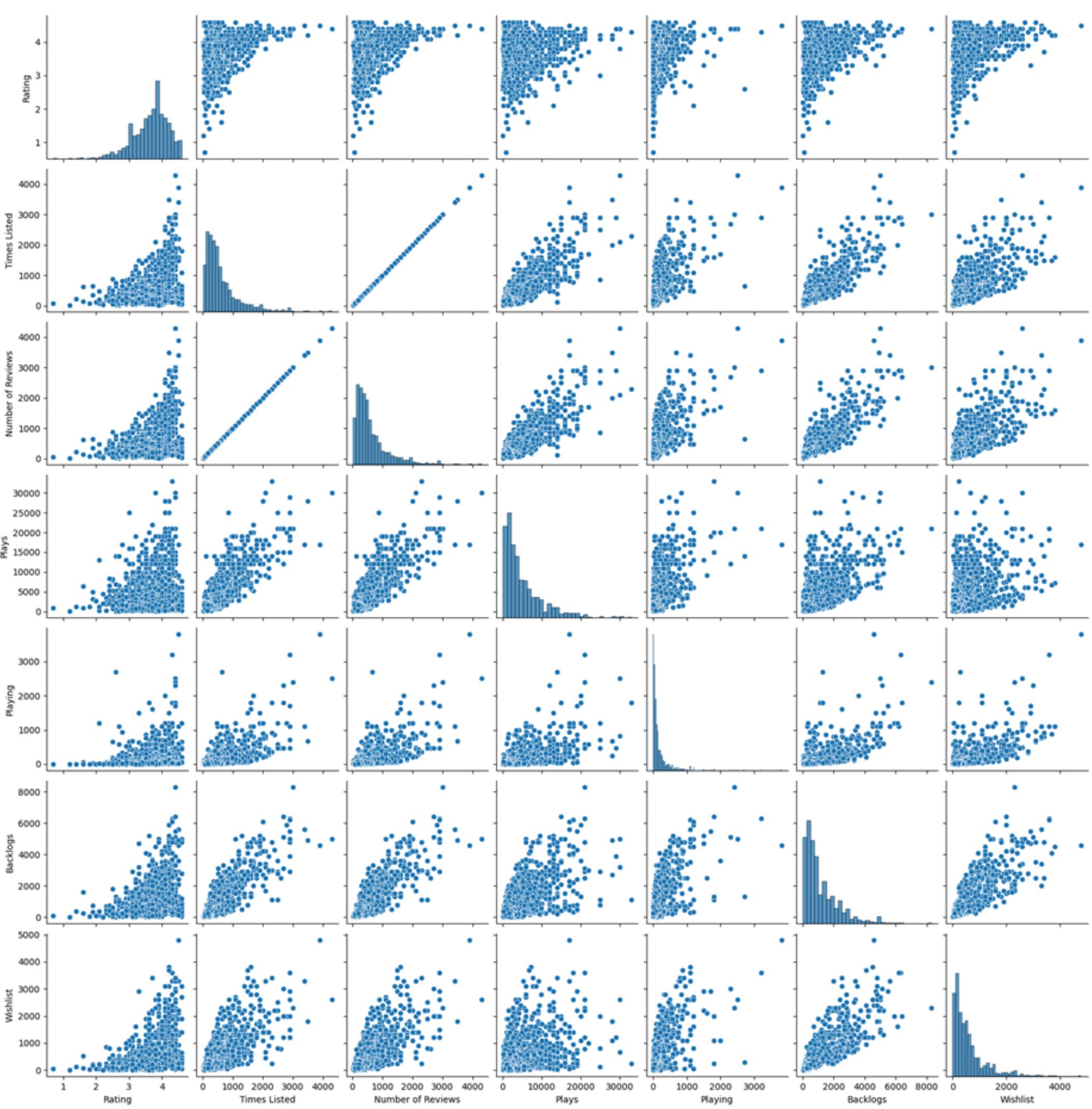
Google Play Store



Korelasi Video Games



Korelasi Video Games



Statistik Deskriptif

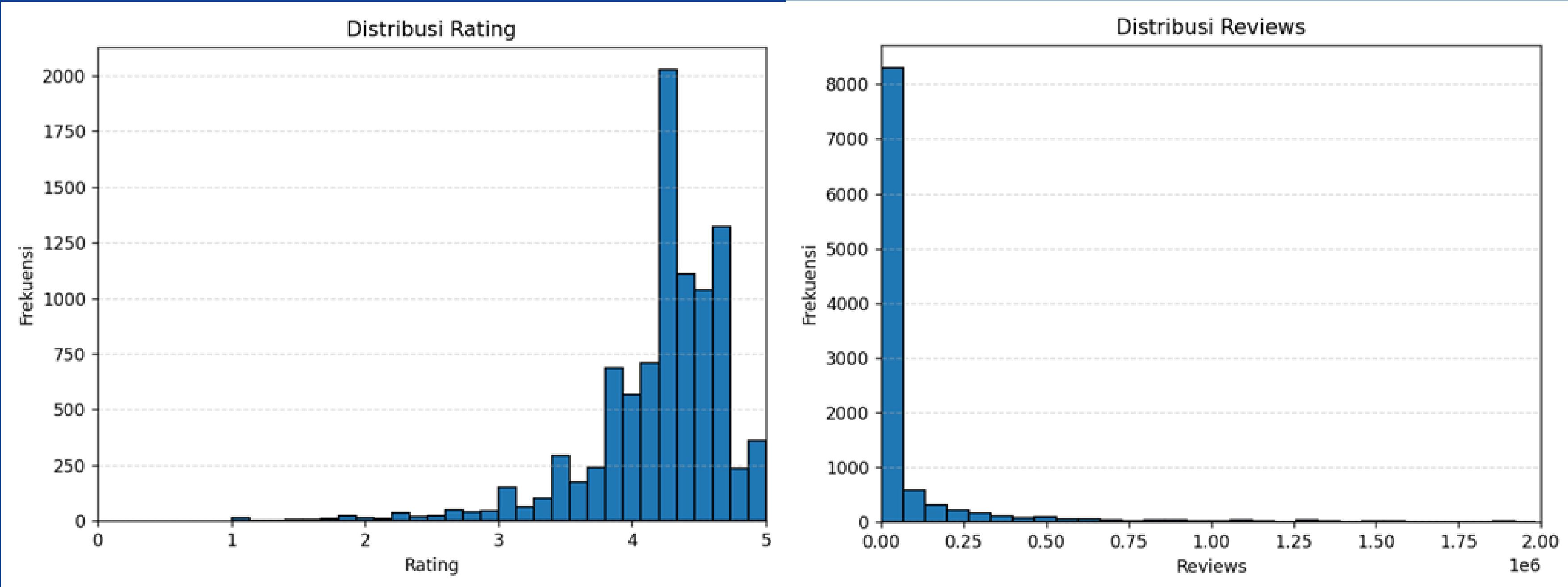
Google Play Store

Kategori	Mean	Std Dev	10%	25%	Median	75%	90%	Min	Max
Rating	4.19	0.54	3.6	4	4.3	4.5	4.7	1	19
Reviews	444152.9	2927760.6	3	38	2094	54775.5	464993.1	0	78158310
Installs	15464338.88	85029361.4	100	1000	100000	5000000	10000000	0	1000000000
Price	1.03	15.95	0	0	0	0	0	0	400



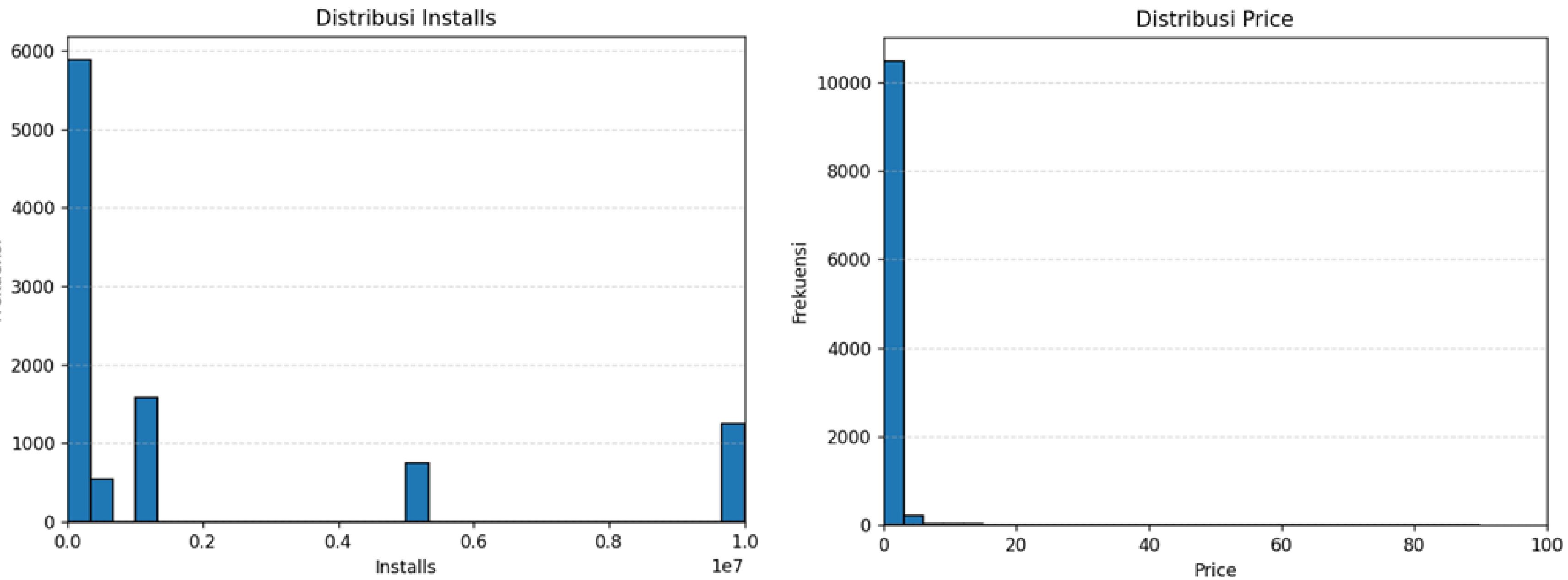
Statistik Deskriptif

Google Play Store



Statistik Deskriptif

Google Play Store



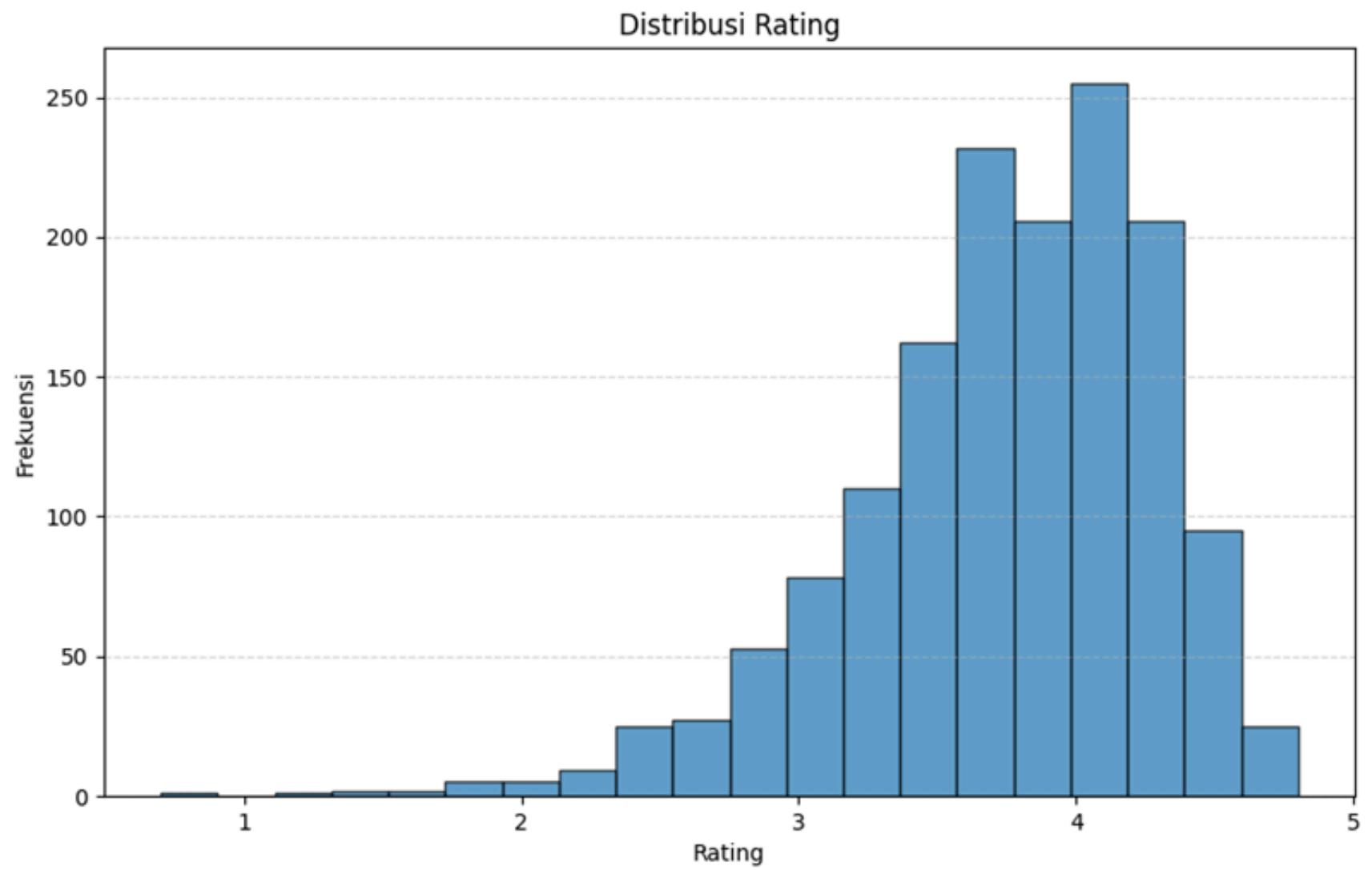
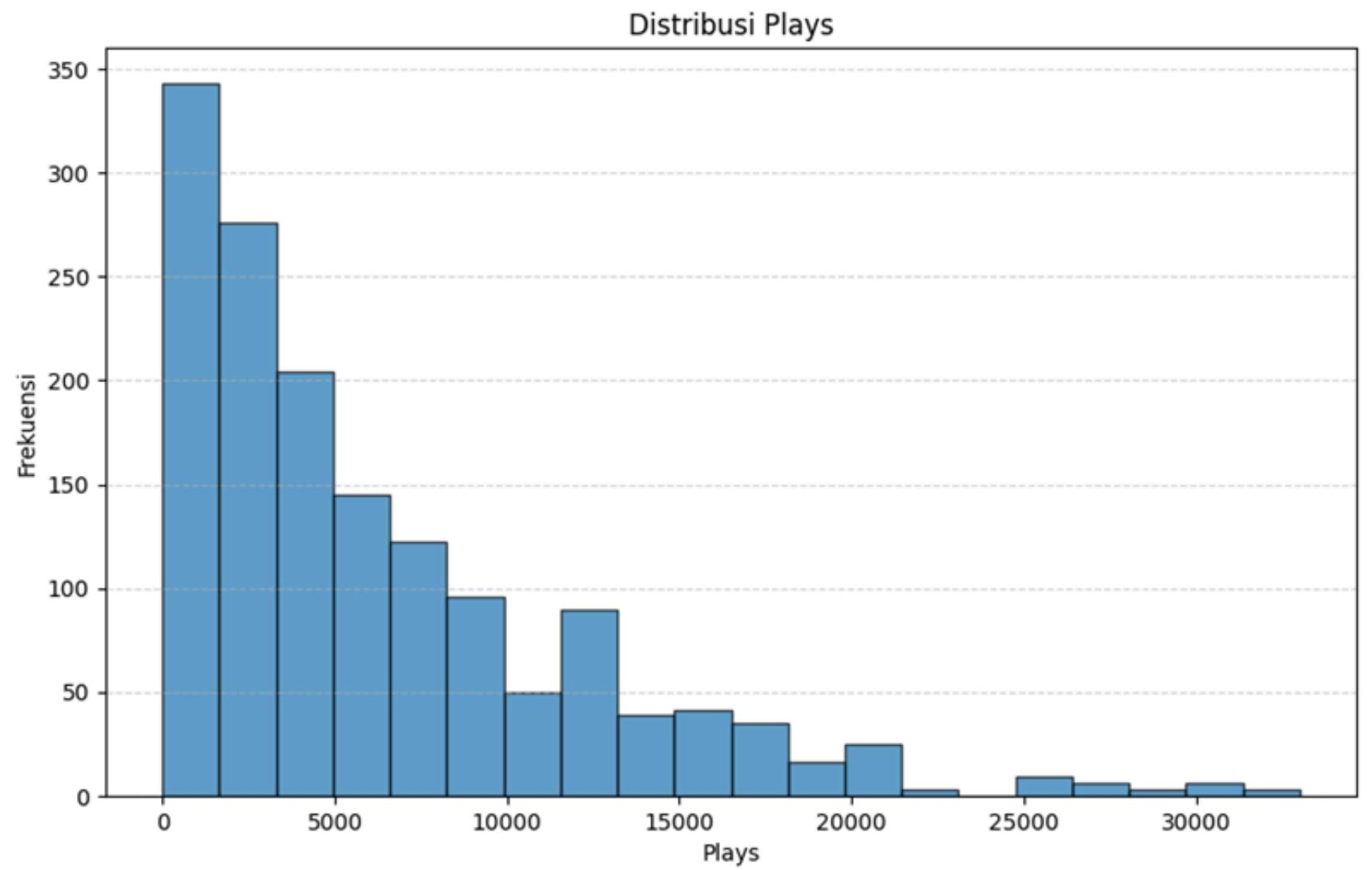
Statistik Deskriptif

Video Games

KATEGORI	RATING	Time Listed	Plays	Wishlist
COUNT	1498	1498	1498	1498
AVERAGE	3.71	751.20	6253.58	780.54
S.Deviation	0.53	660.05	5894.98	800.99
P.10%	3	147.8	875.5	97.2
P.25%	3.4	281	1800	212
P.50%	3.8	545	4200	496
P.75%	4.1	985.5	9100	1100
P.90%	4.3	1700	14000	2000
MAX	4.8	4300	33000	5400
MIN	0.7	0	0	2

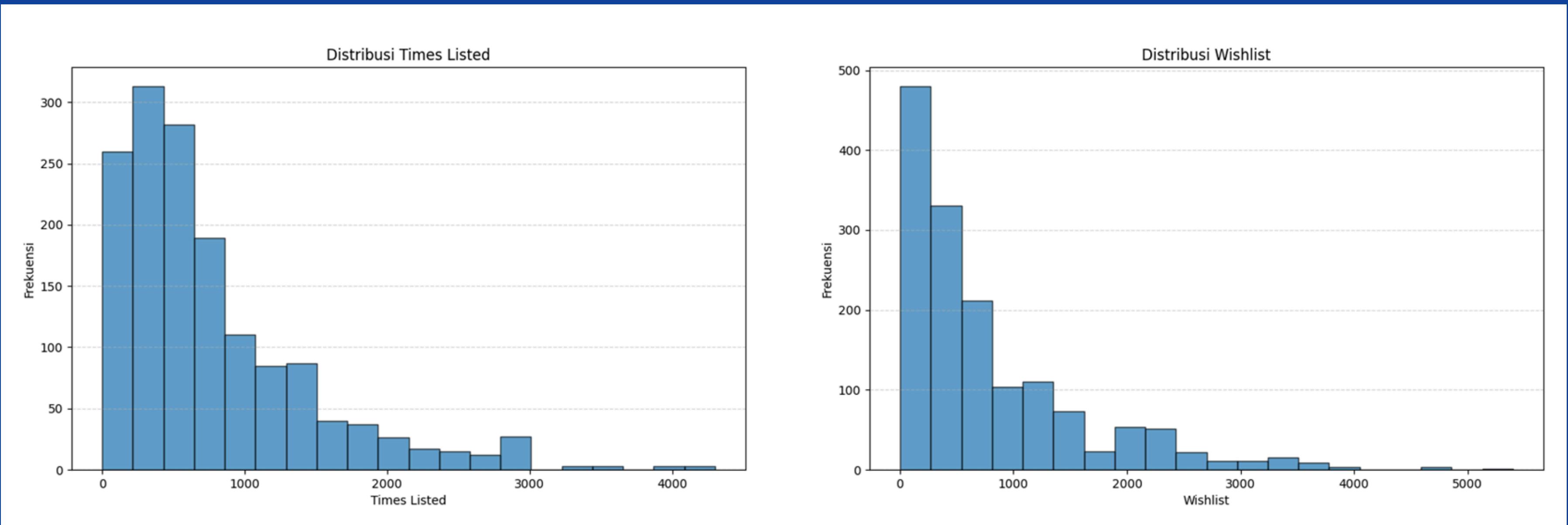
Statistik Deskriptif

Video Games



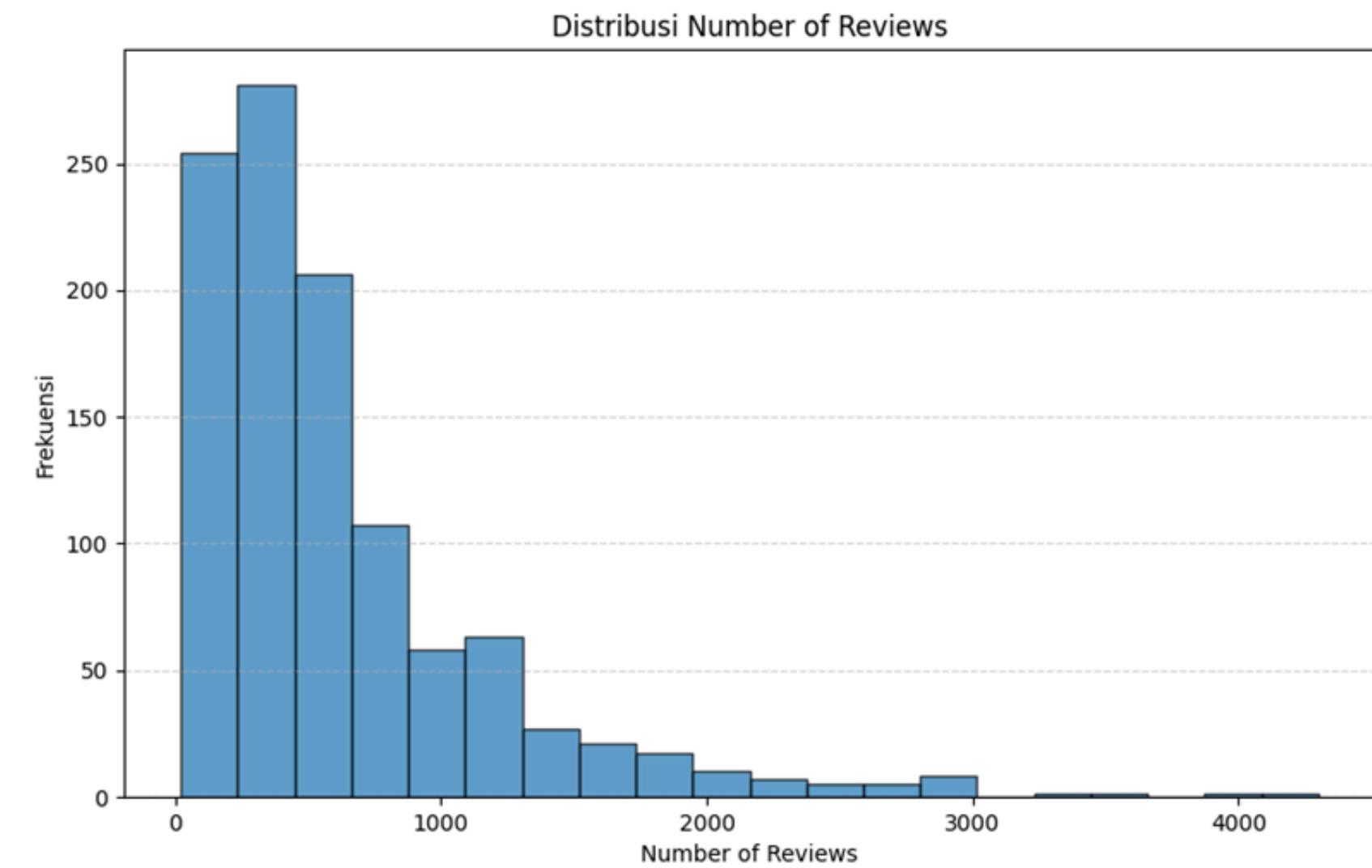
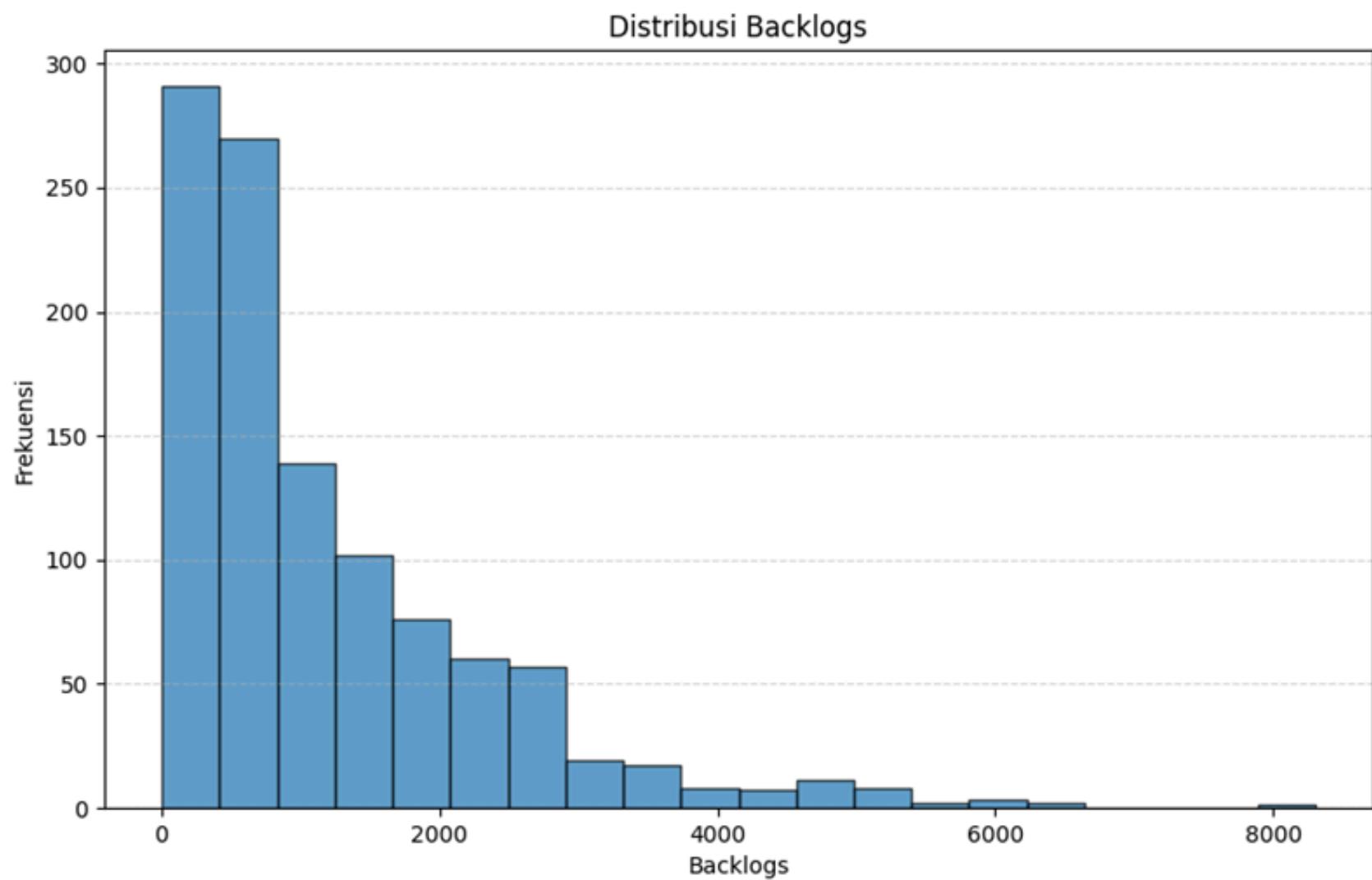
Statistik Deskriptif

Video Games



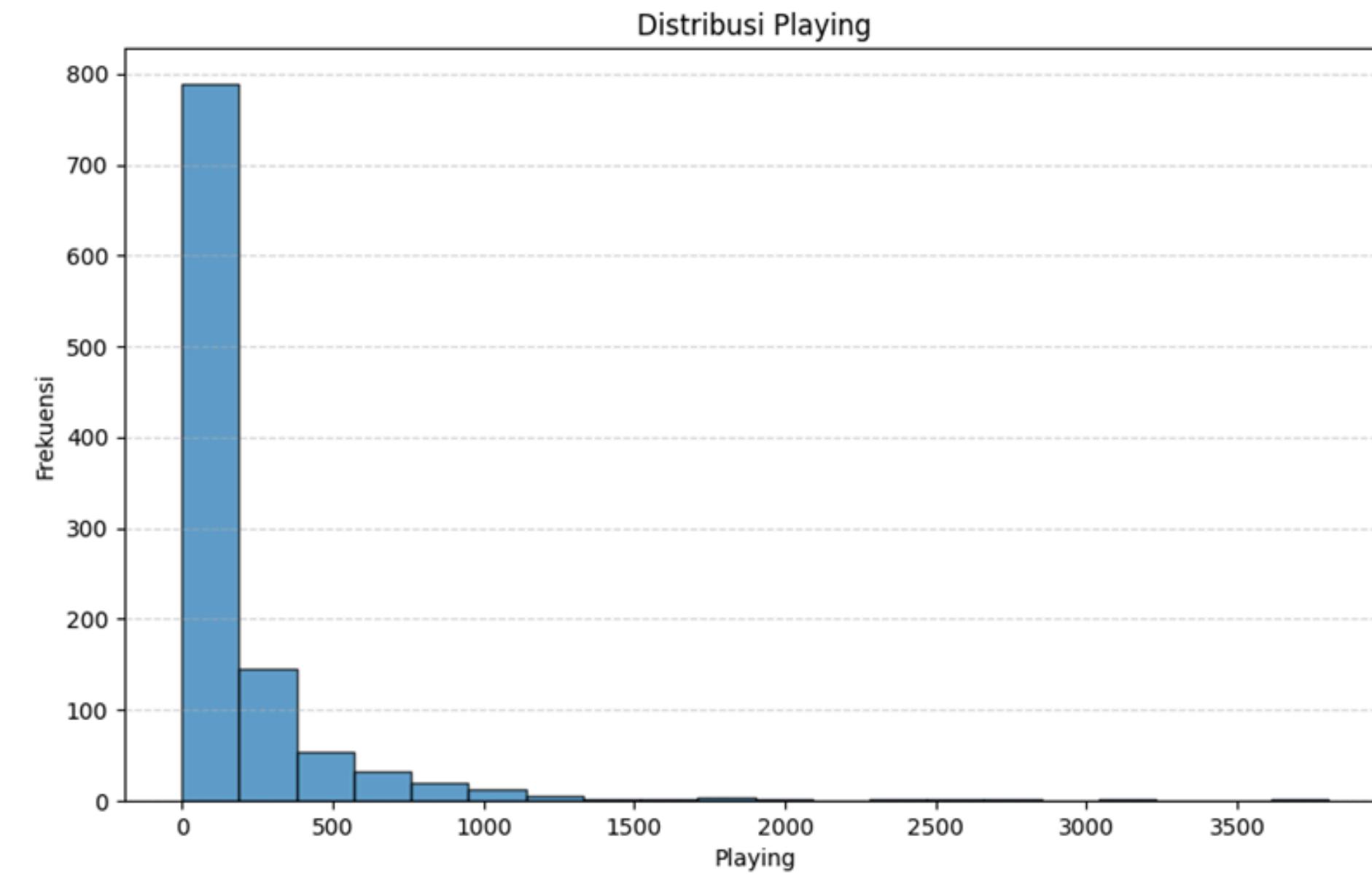
Statistik Deskriptif

Video Games



Statistik Deskriptif

Video Games



Data Cleansing



6.1 Data Aplikasi pada Google Play

6.1.1 Variabel dependen 1: Rating

- Deskripsi Kekotoran
- Karakteristik: Kolom Rating seharusnya berisi nilai numerik floating-point antara 1.0 hingga 5.0.
- Missing Values: Sebanyak 1474 baris dari data awal pada kolom ini adalah NaN.
- Nilai Anomali: Ditemukan satu nilai 19.0 yang berada di luar rentang valid 1.0-5.0.
- Estimasi Persentase Kotor: Sekitar 13,6% dari data awal pada kolom ini adalah missing values, ditambah beberapa nilai anomali yang perlu dihapus.

6.1.2 Variabel dependen 2: Reviews

- Deskripsi Kekotoran
- Karakteristik: Kolom Reviews seharusnya berisi jumlah ulasan sebagai bilangan bulat (integer). Namun, data ini masih dalam format string (object) dan ada potensi nilai non-numerik yang fundamental.
- Nilai Sangat Rendah: Aplikasi dengan jumlah ulasan sangat rendah (misalnya kurang dari 10) mungkin tidak memberikan informasi yang representatif atau stabil untuk analisis.
- Estimasi Persentase Kotor: Sebagian besar nilai perlu konversi tipe data. Sebagian kecil mungkin terlalu rendah untuk relevansi analisis.

6.1.3 Variabel dependen 3: Installs

- Deskripsi Kekotoran
- Karakteristik: Kolom Installs seharusnya berisi jumlah instalasi sebagai bilangan bulat (integer). Namun, data ini masih dalam format string (object) dan mengandung karakter khusus (+, ,).
- Nilai Sangat Rendah: Aplikasi dengan jumlah instalasi sangat rendah (misalnya kurang dari 1000) mungkin tidak relevan untuk analisis tren pasar yang lebih luas.
- Estimasi Persentase Kotor: Semua nilai perlu dibersihkan karakternya dan diubah tipenya. Sebagian kecil mungkin terlalu rendah untuk relevansi analisis. ini adalah missing values, ditambah beberapa nilai anomali yang perlu dihapus.

6.1.4 Variabel dependen 4: Size

- Deskripsi Kekotoran
- Karakteristik: Kolom Size seharusnya berisi ukuran aplikasi sebagai nilai numerik (misalnya dalam Megabytes atau Kilobytes). Namun, data ini masih dalam format string (object) dengan berbagai satuan ('M', 'k') dan nilai non-numerik ("Varies with device").
- Estimasi Persentase Kotor: Semua nilai dalam kolom ini perlu diubah formatnya. Sekitar 15-20% data Size kemungkinan berisi "Varies with device".

Data Cleansing



6.1 Data Aplikasi pada Google Play

6.1.5 Variabel dependen 5: Price

- Deskripsi Kekotoran
- Karakteristik: Kolom Price seharusnya berisi harga aplikasi sebagai nilai numerik floating-point. Namun, data ini masih dalam format string (object) dan mengandung karakter \$.
- Estimasi Persentase Kotor: Semua nilai perlu dibersihkan karakternya dan diubah tipenya.

6.1.6 Variabel Kategorikal dengan Missing Values: Type, Content Rating, Current Ver, Android Ver

- Deskripsi Kekotoran
- Karakteristik: Kolom-kolom ini (Type, Content Rating, Current Ver, Android Ver) adalah kategorikal atau berisi versi string.
- Missing Values: Masing-masing memiliki sedikit missing values.
- Estimasi Persentase Kotor: Bervariasi, antara 0.009% hingga 0.07%.

Snippet Kode Python

```
1 import pandas as pd
2 import numpy as np
3
4 # Muat dataset
5 df_games = pd.read_csv('games.csv')
6
7 print(" --- INFORMASI DATA GAMES.CSV SEBELUM CLEANSING ---")
8 print("Informasi Umum Dataset Games.csv:")
9 df_games.info()
10 print("\nJumlah Missing Values per Kolom Games.csv:")
11 print(df_games.isnull().sum())
12 print("\nPersentase Missing Values per Kolom Games.csv:")
13 print((df_games.isnull().sum() / len(df_games)) * 100)
14 print(f"\nJumlah Baris Duplikat Awal Games.csv: {df_games.duplicated().sum()}")
15 print("\nStatistik Deskriptif Kolom Numerik Awal Games.csv:")
16 print(df_games.describe())
17 print("\nContoh 5 Baris Pertama Games.csv:")
18 print(df_games.head())
19 print("-" * 50)
20
21 # Menghapus duplikat berdasarkan kolom 'Title'
22 rows_before_deduplication = len(df_games)
23 df_games.drop_duplicates(subset=['Title'], inplace=True)
24 print(f"\n--- LANGKAH UMUM: Menghapus Duplikat ---")
25 print(f" Baris dihapus karena duplikasi (berdasarkan Tit
26
27 # Kolom: Rating
28 # Tindakan: Hapus baris dengan missing values, pastikan
29 rows_before_rating_clean = len(df_games)
30 df_games['Rating'] = pd.to_numeric(df_games['Rating'],
31 df_games.dropna(subset=['Rating'], inplace=True) # Meng
32 # Filter untuk rentang yang valid (jika ada nilai di lu
33 df_games = df_games[(df_games['Rating'] >= 0.0) & (df_g
34 print(f"\n--- CLEANSING 'Rating' ---")
35 print(f" Baris dihapus (missing/out-of-range): {rows_b
36
37 # Fungsi bantu untuk konversi string 'K'/M' ke numerik
38 def convert_k_m_to_numeric(value):
39     if isinstance(value, str):
40         value = value.replace(',', '') # Hapus koma jika
41         if 'K' in value:
42             return float(value.replace('K', '')) * 1000
43         elif 'M' in value:
44             return float(value.replace('M', '')) * 1000000
45     return pd.to_numeric(value, errors='coerce') # Konver
46
47 columns_to_convert_and_filter = {
48
49     'Times Listed': 10, # Batas bawah 10 kali terdaftar
50     'Number of Reviews': 10, # Batas bawah 10 kali terdaftar
51     'Plays': 100, # Batas bawah 100 plays
52     'Playing': None, # Tidak ada batas bawah spesifik
53     'Backlogs': None, # Tidak ada batas bawah spesifik
54     'Wishlist': None # Tidak ada batas bawah spesifik
55 }
56
57 for col, threshold in columns_to_convert_and_filter.items():
58     rows_before_col_clean = len(df_games)
59     df_games[col] = df_games[col].apply(convert_k_m_to_numeric)
60     df_games.dropna(subset=[col], inplace=True) # Hapus jika gagal konversi (NaN dari coerce)
61     print(f"\n--- CLEANSING '{col}' ---")
62     print(f" Baris dihapus (gagal konversi/missing): {rows_b
63
64     if threshold is not None:
65         rows_before_threshold_filter = len(df_games)
66         df_games = df_games[df_games[col] >= threshold]
67         print(f" Baris dihapus ({col} < {threshold}): {rows_b
68
69         df_games[col] = df_games[col].astype(int) # Ubah ke integer secara permanen
70
71 # Kolom: Team, Summary
72 # Tindakan: Imputasi missing values dengan modus.
73 columns_to_impute_mode = ['Team', 'Summary']
74 for col in columns_to_impute_mode:
75     if df_games[col].isnull().any():
76         mode_val = df_games[col].mode()[0]
77         df_games[col].fillna(mode_val, inplace=True)
78         print(f"\n--- CLEANSING '{col}' ---")
79         print(f" Missing values pada '{col}' diisi dengan modus: {mode_val}")
80
81
82
83 # --- HASIL KESELURUHAN SETELAH CLEANSING ---
84 print("\n--- INFORMASI DATA GAMES.CSV SETELAH SEMUA CLEANSING ---")
85 print("Informasi Umum Dataset Games.csv:")
86 df_games.info()
87
88 print("\nJumlah Missing Values per Kolom Games.csv Setelah Cleansing:")
89 print(df_games.isnull().sum())
90 print("\nPersentase Missing Values per Kolom Games.csv Setelah Cleansing:")
91 print((df_games.isnull().sum() / len(df_games)) * 100)
92 print("\nStatistik Deskriptif Kolom (termasuk non-numerik):")
93 print(df_games.describe(include='all'))
94 print("\n5 Baris Pertama Data Games.csv Setelah Cleansing:")
95 print(df_games.head())
96
97 df_games.to_csv('games_cleaned.csv', index=False)
```

Data Cleansing



6.2 Data Video Game Terpopuler 1980 - 2023

6.2.1 Variabel dependen 1: Rating

- Deskripsi Kekotoran
- Karakteristik: Kolom Rating berisi rating numerik game (misalnya 4.5, 3.8). Rentang yang valid adalah 0.0 hingga 5.0.
- Missing Values: Sebanyak 13 baris memiliki NaN (data kosong).
- Nilai Anomali: Tidak ada nilai anomali yang jelas di luar rentang (min 0.7, max 4.8).
- Estimasi Persentase Kotor: Sekitar 0.86% dari data adalah missing values.

6.2.2 Variabel: Times Listed, Number of Reviews, Plays, Playing, Backlogs, Wishlist

- Deskripsi Kekotoran
- Karakteristik: Kolom-kolom ini seharusnya berisi hitungan numerik. Namun, data ini masih dalam format string (object) dan mengandung akhiran 'K' (untuk ribuan) atau 'M' (untuk jutaan) yang perlu dikonversi.
- Nilai Sangat Rendah (khusus untuk Times Listed, Number of Reviews, dan Plays): Game yang terdaftar, terreview atau dimainkan sangat sedikit mungkin tidak memberikan informasi yang representatif atau stabil untuk analisis.
- Estimasi Persentase Kotor: Semua nilai perlu dibersihkan karakternya dan diubah tipenya. Sebagian kecil mungkin terlalu rendah untuk relevansi analisis.

6.2.3 Variabel: Team dan Summary

- Deskripsi Kekotoran
- Karakteristik: Kolom Team (tim pengembang) dan Summary (ringkasan game) adalah kolom teks/kategorikal.
- Missing Values: Masing-masing memiliki 1 missing value.
- Estimasi Persentase Kotor: Sangat kecil (sekitar 0.066% masing-masing).

Snippet Kode Python

```
● ● ●
1 import pandas as pd
2 import numpy as np
3
4 # Muat dataset
5 df_games = pd.read_csv('games.csv')
6
7 print("---- INFORMASI DATA GAMES.CSV SEBELUM CLEANSING ---")
8 print("Informasi Umum Dataset Games.csv:")
9 df_games.info()
10 print("\nJumlah Missing Values per Kolom Games.csv:")
11 print(df_games.isnull().sum())
12 print("\nPersentase Missing Values per Kolom Games.csv:")
13 print((df_games.isnull().sum() / len(df_games)) * 100)
14 print(f"\nJumlah Baris Duplikat Awal Games.csv: {df_games.duplicated().sum()}")
15 print("\nStatistik Deskriptif Kolom Numerik Awal Games.csv:")
16 print(df_games.describe())
17 print("\nContoh 5 Baris Pertama Games.csv:")
18 print(df_games.head())
19 print("-" * 50)
20
21 # Menghapus duplikat berdasarkan kolom 'Title'
22 rows_before_deduplication = len(df_games)
23 df_games.drop_duplicates(subset=['Title'], inplace=True)
24 print(f"\n--- LANGKAH UMUM: Menghapus Duplikat ---")
25 print(f" Baris dihapus karena duplikasi (berdasarkan Title)")
26
27 # Kolom: Rating
28 # Tindakan: Hapus baris dengan missing values, pastikan data masih valid
29 rows_before_rating_clean = len(df_games)
30 df_games['Rating'] = pd.to_numeric(df_games['Rating'], errors='coerce')
31 df_games.dropna(subset=['Rating'], inplace=True) # Menghapus baris dengan rating yang tidak valid
32 # Filter untuk rentang yang valid (jika ada nilai di luar range)
33 df_games = df_games[(df_games['Rating'] >= 0.0) & (df_games['Rating'] <= 5.0)]
34 print(f"\n--- CLEANSING 'Rating' ---")
35 print(f" Baris dihapus (missing/out-of-range): {rows_before_rating_clean - len(df_games)}")
36
37 # Fungsi bantu untuk konversi string 'K'/'M' ke numerik
38 def convert_k_m_to_numeric(value):
39     if isinstance(value, str):
40         value = value.replace(',', '') # Hapus koma jika ada
41         if 'K' in value:
42             return float(value.replace('K', '')) * 1000
43         elif 'M' in value:
44             return float(value.replace('M', '')) * 1000000
45     return pd.to_numeric(value, errors='coerce') # Konversi ke numerik
46
47 columns_to_convert_and_filter = {
48     'Number of Reviews': 10, # Batas bawah 10 kali terdaftar
49     'Plays': 100, # Batas bawah 100 plays
50     'Playing': None, # Tidak ada batas bawah spesifik
51     'Backlogs': None, # Tidak ada batas bawah spesifik
52     'Wishlist': None # Tidak ada batas bawah spesifik
53 }
54
55 for col, threshold in columns_to_convert_and_filter.items():
56     rows_before_col_clean = len(df_games)
57     df_games[col] = df_games[col].apply(convert_k_m_to_numeric)
58     df_games.dropna(subset=[col], inplace=True) # Hapus jika gagal konversi (NaN dari coerce)
59     print(f"\n--- CLEANSING '{col}' ---")
60     print(f" Baris dihapus (gagal konversi/missing): {rows_before_col_clean - len(df_games)}")
61
62     if threshold is not None:
63         rows_before_threshold_filter = len(df_games)
64         df_games = df_games[df_games[col] >= threshold]
65         print(f" Baris dihapus ({col} < {threshold}): {rows_before_threshold_filter - len(df_games)}")
66
67     df_games[col] = df_games[col].astype(int) # Ubah ke integer secara permanen
68
69 # Kolom: Team, Summary
70 # Tindakan: Imputasi missing values dengan modus.
71 columns_to_impute_mode = ['Team', 'Summary']
72 for col in columns_to_impute_mode:
73     if df_games[col].isnull().any():
74         mode_val = df_games[col].mode()[0]
75         df_games[col].fillna(mode_val, inplace=True)
76         print(f"\n--- CLEANSING '{col}' ---")
77         print(f" Missing values pada '{col}' diisi dengan modus: {mode_val}")
78
79
80 # --- HASIL KESELURUHAN SETELAH CLEANSING ---
81 print("\n\n--- INFORMASI DATA GAMES.CSV SETELAH SEMUA CLEANSING ---")
82 print("Informasi Umum Dataset Games.csv:")
83 df_games.info()
84 print("\nJumlah Missing Values per Kolom Games.csv Setelah Cleansing:")
85 print(df_games.isnull().sum())
86 print("\nPersentase Missing Values per Kolom Games.csv Setelah Cleansing:")
87 print((df_games.isnull().sum() / len(df_games)) * 100)
88 print("\nStatistik Deskriptif Kolom (termasuk non-numerik):")
89 print(df_games.describe(include='all'))
90 print("\n5 Baris Pertama Data Games.csv Setelah Cleansing:")
91 print(df_games.head())
92 df_games.to_csv('games_cleaned.csv', index=False)
```

```
● ● ●
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
```

Transformasi Data



7.1 Data Aplikasi pada Google Play

Nama Atribut	Transformasi?	Satuan/Range Awal	Satuan/Range Akhir	Penjelasan
#		Numerikal	Numerikal	Tidak perlu dilakukan transformasi karena tidak termasuk variabel yang ingin dianalisis.
App		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena sudah benar dan sesuai.
Category		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena data sudah benar dan sesuai.
Rating	V	'5.4'	'5,4'	Perlu dilakukan transformasi karena pada Ms. Excel yang berada dalam region Indonesia atau dengan format desimal dalam bentuk koma (,), data harus diubah agar dapat dianalisis. Untuk melakukan transformasi data, kami menggunakan rumus pada Ms. Excel, yaitu <code>=VALUE(SUBSTITUTE(rating;".",";"))</code>
Reviews		Numerikal	Numerikal	Tidak perlu dilakukan transformasi data karena data sudah eksak dan sesuai.
Size		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi data karena data ukuran aplikasi tidak digunakan untuk analisis sebagai suatu variabel.
Installs	V	'5,000,000+'	'5000000'	Perlu dilakukan transformasi karena data dalam satuan awal masih belum terukur dan bisa saja naik atau turun sewaktu-waktu jika digunakan nilai eksak. Oleh karena itu, satuan akhir diputuskan berdasarkan batas minimal jumlah unduhan yang ada. Untuk mengetahuinya, digunakan formula Ms. Excel sebagai berikut. <code>=VALUE(SUBSTITUTE(SUBSTITUTE(installs, ",",""), "+",""))</code>
Type		Free / Paid	Free / Paid	Tidak perlu dilakukan transformasi karena data sudah benar secara satuan dan range.

Transformasi Data



7.1 Data Aplikasi pada Google Play

Nama Atribut	Transformasi?	Satuan/Range Awal	Satuan/Range Akhir	Penjelasan
Price	V	'\$0.99'	'0.99'	Perlu dilakukan transformasi karena pada format mata uang dollar AS, data perlu disederhanakan menjadi 0.99 agar bisa dilakukan analisis lebih sederhana. Proses penyederhanaan ini menggunakan formula <code>=VALUE(SUBSTITUTE(price,"\$","",))</code>
Content Rating		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi data karena content rating yang ada sudah terdefinisi jelas.
Genres		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi data karena genre bukan menjadi fokus utama pertanyaan penelitian.
Last Updated		Time-series	Time-series	Tidak perlu dilakukan transformasi data karena perhitungan time-series menggunakan PivotTable langsung mengubah data khusus time-series.
Current Ver		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena version sudah dalam format yang tepat.
Android Ver		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena version sudah dalam format yang tepat.

Transformasi Data



7.2 Data Video Game Terpopuler 1980 - 2023

Nama Atribut	Transformasi?	Satuan/Range Awal	Satuan/Range Akhir	Penjelasan
#		Numerikal	Numerikal	Tidak perlu dilakukan transformasi karena tidak termasuk variabel yang ingin dianalisis.
Title		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena sudah benar dan sesuai.
Release Date	V	[Bulan] [Tanggal], [Tahun]	MM/DD/YYYY	Perlu dilakukan transformasi karena jika digunakan data Time-Series dengan format awal, Ms. Excel tidak dapat membacanya dengan baik. Dengan menggunakan tools "Transform Data" dari Ms. Excel, data dapat terbaca dan disesuaikan dengan format yang diharapkan.
Team		Kategorikal	Kategorikal	Tidak perlu dilakukan karena asal usul tim pembuat video game tidak menjadi salah satu variabel yang dianalisis.
Rating	V	'5.4'	'5,4'	Perlu dilakukan transformasi karena pada Ms. Excel yang berada dalam region Indonesia atau dengan format desimal dalam bentuk koma (,), data harus diubah agar dapat dianalisis. Untuk melakukan transformasi data, kami menggunakan rumus pada Ms. Excel, yaitu =VALUE(SUBSTITUTE(rating;".",","))
Times Listed	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(times_listed); times_listed; IF(RIGHT(times_listed;1)="K";VALUE(SUBSTITUTE(LEFT(times_listed;LEN(times_listed)-1);".","")) * 1000; IF(RIGHT(times_listed;1)="M";VALUE(SUBSTITUTE(LEFT(times_listed;LEN(times_listed)-1);".","")) * 1000000; IF(RIGHT(times_listed;1)="B"; VALUE(SUBSTITUTE(LEFT(times_listed;LEN(times_listed)-1);".",""))*1000000000;VALUE(SUBSTITUTE(times_listed;".","")))))

Transformasi Data



7.2 Data Video Game Terpopuler 1980 - 2023

Nama Atribut	Transformasi?	Satuan/Range Awal	Satuan/Range Akhir	Penjelasan
Number of Reviews	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, seluruh satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(reviews); reviews; IF(RIGHT(reviews;1)="K";VALUE(SUBSTITUTE(LEFT(reviews;LEN(reviews)-1);".",".")) * 1000; IF(RIGHT(reviews;1)="M";VALUE(SUBSTITUTE(LEFT(reviews;LEN(reviews)-1);".",".")) * 1000000; IF(RIGHT(reviews;1)="B";VALUE(SUBSTITUTE(LEFT(reviews;LEN(reviews)-1);".",".")); 1000000000*VALUE(SUBSTITUTE(reviews;".","."))))*)
Genres	V	['Adventure', 'Indie', 'RPG', 'Turn Based Strategy']	Adventure	Perlu dilakukan transformasi karena untuk mendeteksi suatu video game yang memiliki beberapa genre, analisis data akan menjadi sulit karena genre suatu game tidak tentu banyaknya. Oleh sebab itu, perlu dilakukan perubahan dengan rumus Ms. Excel berikut. =MID(genres, FIND("""", genres)+1, FIND("""", genres, FIND("""", genres)+1) - FIND("""", genres) - 1)
Summary		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena sangat sulit untuk melakukan analisis data secara kualitatif sebab membutuhkan tools yang lebih lanjut, seperti AI.
Reviews		Kategorikal (Nominal)	Kategorikal (Nominal)	Tidak perlu dilakukan transformasi karena sangat sulit untuk melakukan analisis data secara kualitatif sebab membutuhkan tools yang lebih lanjut, seperti AI.
Plays	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, seluruh satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(plays); plays; IF(RIGHT(plays;1)="K";VALUE(SUBSTITUTE(LEFT(plays;LEN(plays)-1);".",".")) * 1000; IF(RIGHT(plays;1)="M";VALUE(SUBSTITUTE(LEFT(plays;LEN(plays)-1);".",".")) * 1000000; IF(RIGHT(plays;1)="B";VALUE(SUBSTITUTE(LEFT(plays;LEN(plays)-1);".",".")); 1000000000*VALUE(SUBSTITUTE(LEFT(plays;LEN(plays)-1);".","."))))*)

Transformasi Data



7.2 Data Video Game Terpopuler 1980 - 2023

Nama Atribut	Transformasi?	Satuan/Range Awal	Satuan/Range Akhir	Penjelasan
Playing	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, seluruh satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(playing); playing; IF(RIGHT(playing;1)="K";VALUE(SUBSTITUTE(LEFT(playing;LEN(playing)-1); ".",".")) * 1000; IF(RIGHT(playing;1)="M";VALUE(SUBSTITUTE(LEFT(playing;LEN(playing)-1); ".",".")) * 1000000; IF(RIGHT(playing;1)="B";VALUE(SUBSTITUTE(LEFT(playing;LEN(playing)-1); ".",".")) * 1000000000; VALUE(SUBSTITUTE(playing; ".","."))))
Backlogs	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, seluruh satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(backlogs); backlogs; IF(RIGHT(backlogs;1)="K";VALUE(SUBSTITUTE(LEFT(backlogs;LEN(backlogs)-1); ".",".")) * 1000; IF(RIGHT(backlogs;1)="M";VALUE(SUBSTITUTE(LEFT(backlogs;LEN(backlogs)-1); ".",".")) * 1000000; IF(RIGHT(backlogs;1)="B";VALUE(SUBSTITUTE(LEFT(backlogs;LEN(backlogs)-1); ".",".")) * 1000000000; VALUE(SUBSTITUTE(backlogs; ".","."))))
Wishlist	V	'1.9K' '1.9M' '1.9B'	'1900' '1900000' '1900000000'	Perlu dilakukan transformasi karena sistem akan sulit mengenali satuan dalam bentuk 'K', 'M', dan 'B' yang mana bukan merupakan satuan formal. Oleh karena itu, seluruh satuan ini perlu ditransformasikan menjadi satuan dengan format numerik. =IF(ISNUMBER(wishlist); wishlist; IF(RIGHT(wishlist;1)="K";VALUE(SUBSTITUTE(LEFT(wishlist;LEN(wishlist)-1); ".",".")) * 1000; IF(RIGHT(wishlist;1)="M";VALUE(SUBSTITUTE(LEFT(wishlist;LEN(wishlist)-1); ".",".")) * 1000000; IF(RIGHT(wishlist;1)="B";VALUE(SUBSTITUTE(LEFT(wishlist;LEN(wishlist)-1); ".",".")) * 1000000000; VALUE(SUBSTITUTE(wishlist; ".","."))))

8. Data Analytics Sederhana

8.1 Data Aplikasi pada Google Play

8.1.1 Model Regresi Linear antara Rating dan Jumlah Install

Dalam bagian ini, kita membangun sebuah model regresi linear sederhana yang bertujuan untuk memprediksi jumlah installs sebuah aplikasi berdasarkan rating aplikasi tersebut. Dataset yang digunakan merupakan data aplikasi dari Google Play Store, yang sebelumnya telah dibersihkan dan difilter agar hanya menyisakan data dengan kolom Rating dan Installs yang valid.

Langkah-langkah umum yang dilakukan untuk menyusun model ini adalah sebagai berikut:

1. Filter dataset agar hanya berisi aplikasi yang memiliki nilai Rating dan Installs yang valid (tidak kosong).
2. Konversi nilai Installs ke bentuk numerik, dengan menghapus simbol + dan , agar dapat diproses sebagai angka.
3. Gunakan Rating sebagai variabel independen (sumbu-x), dan Installs sebagai variabel dependen (sumbu-y).
4. Buat model regresi linear menggunakan LinearRegression dari pustaka sklearn.linear_model.
5. Ambil koefisien regresi (kemiringan/slope dan intersep) sebagai dasar penyusunan formula model.

Berdasarkan model regresi yang telah dibentuk, diperoleh formula prediktif sebagai berikut:

$$\text{Installs}(r)=ar+b$$

dengan

- r : nilai rating aplikasi
- a : koefisien kemiringan regresi = 9128238.10
- b : intersep (konstanta) = -20759976.51

Model regresi ini mengindikasikan bahwa setiap peningkatan sebesar 1 poin dalam nilai rating aplikasi berkorelasi dengan peningkatan jumlah unduhan sekitar 9,1 juta kali. Namun, nilai intersep yang negatif mengimplikasikan bahwa jika rating bernilai 0, prediksi jumlah unduhannya justru bernilai negatif, yang jelas tidak logis.

Sebagai ilustrasi, jika dimasukkan nilai rating sebesar 4.5 ke dalam model, diperoleh:

$$\text{Installs}(4.5)=9128238.10 \times 4.5 - 20759976.51 \approx 20.542.875 \text{ kali}$$

8. Data Analytics Sederhana

8.1 Data Aplikasi pada Google Play

8.1.2 Model Regresi Linear antara Harga dan Jumlah Install

Pada bagian ini, dibangun sebuah model regresi linear sederhana untuk memprediksi jumlah install berdasarkan harga aplikasi yang tercantum pada Google Play Store. Analisis dilakukan menggunakan dataset yang telah melalui proses pembersihan sebelumnya.

Langkah-langkah penyusunan model :

- Dataset difilter untuk menghapus baris dengan nilai kosong pada kolom Price dan Installs.
- Kolom Price dibersihkan dari simbol "\$" dan dikonversi menjadi numerik (float).
- Model regresi linear dibentuk menggunakan Price sebagai variabel independen dan Installs sebagai variabel dependen.
- Penerapan dilakukan dengan libraryLinearRegression dari sklearn.linear_model.

Hasil pemodelan regresi menghasilkan persamaan sebagai berikut:

$$\text{Installs}(p) = 0.00 * p + 19217171.16$$

Model ini menunjukkan bahwa harga aplikasi tidak memiliki pengaruh signifikan terhadap jumlah install berdasarkan data yang tersedia. Koefisien slope yang bernilai nol menandakan bahwa model mempelajari hubungan yang sangat lemah, atau bahkan tidak ada, antara harga dan jumlah unduhan. Hal ini kemungkinan besar disebabkan karena dominasi aplikasi gratis dalam dataset, sehingga variasi harga yang berbayar tidak cukup untuk menghasilkan pola yang bermakna.

Data Analytics Sederhana

8.2 Data Video Game Terpopuler 1980 - 2023

Dalam bagian ini, dilakukan pemodelan regresi linear sederhana untuk memprediksi jumlah wishlist pada suatu game berdasarkan nilai rating yang dimiliki. Model ini disusun dengan menggunakan data video game dari dataset yang telah dibersihkan dan diproses agar hanya memuat nilai numerik yang valid.

Langkah-langkah Penyusunan Model :

- Data difilter untuk menghapus baris dengan nilai kosong pada kolom Rating dan Wishlist.
- Kolom Rating digunakan sebagai variabel independen (fitur), dan Wishlist sebagai variabel dependen (target).
- Model regresi linear dibangun menggunakan library `sklearn.linear_model.LinearRegression`.

Model regresi linear yang dihasilkan adalah sebagai berikut:

$$\text{Wishlist}(r) = -735.14 * r + 3520.27$$

Sebagai ilustrasi, jika suatu game memiliki rating sebesar 4.5, maka jumlah wishlist dapat diprediksi dengan:

$$\text{Wishlist}(4.5) = -735.14 * 4.5 + 3520.27 \approx 212.12$$

Artinya, model memperkirakan bahwa game dengan rating 4.5 akan memiliki sekitar **212** wishlist.

Kesimpulan

Google Play Store



Analisis menunjukkan bahwa rating aplikasi hanya memiliki pengaruh lemah terhadap jumlah unduhan. Meskipun terdapat korelasi positif, aplikasi dengan rating tinggi belum tentu populer, karena pengguna lebih mempertimbangkan kategori aplikasi, visibilitas di Play Store, dan harga.

Sementara itu, aplikasi bertipe game memang lebih sering diperbarui dibandingkan tipe lainnya. Hal ini karena game membutuhkan pembaruan rutin untuk memperbaiki bug, menambah fitur, atau adaptasi ke versi Android terbaru – mencerminkan tingkat engagement tinggi dari developer game.

Dari sisi harga, ditemukan bahwa harga memiliki pengaruh negatif yang kuat terhadap jumlah unduhan. Aplikasi gratis mendominasi, sedangkan aplikasi berbayar cenderung jarang diunduh. Hal ini menandakan bahwa pengguna sangat sensitif terhadap harga, dan pendekatan freemium terbukti paling sukses.

Kategori juga berperan besar: aplikasi bertipe komunikasi memiliki jumlah unduhan tertinggi. Ini karena sifat aplikasinya esensial dan digunakan berulang oleh hampir semua pengguna. Variabel lain yang turut berpengaruh terhadap jumlah unduhan adalah: apakah aplikasi gratis, jumlah review, dan frekuensi update.

Kesimpulan

Data Video Game Terpopuler



Tanggal rilis memengaruhi pola pemain secara menarik: game lama memiliki lebih banyak total pemain (karena waktu akumulasi), sedangkan game baru cenderung punya pemain aktif lebih tinggi. Ini menunjukkan adanya siklus hidup game, dari booming hingga turun.

Popularitas genre berubah secara dinamis tiap tahun. Genre Adventure mendominasi dulu, namun dalam dekade terakhir genre Indie naik pesat. Ini mencerminkan perubahan tren industri dan munculnya lebih banyak developer kecil di pasar global.

Untuk jumlah pemain tertinggi, genre MOBA unggul, diikuti Shooter dan Racing. Namun menariknya, MOBA justru memiliki rating rendah. Artinya, aspek seperti kompetisi, komunitas, dan replayability lebih penting dibanding rating semata dalam memengaruhi engagement.

Terakhir, angka wishlist ternyata tidak selalu sejalan dengan rating. Wishlist lebih mencerminkan minat awal, tapi tidak selalu berlanjut ke aktivitas bermain. Korelasi yang tinggi antara wishlist dan backlog menunjukkan bahwa banyak pengguna hanya “berencana main”, bukan langsung main.



Terima Kasih
atas Atensinya

Tugas Besar LiDIA