LAPORAN TUGAS BESAR
WI2002 LITERASI DATA DAN INTELEGENSI ARTIFISIAL
SEMESTER II 2024-2025
Exploratory Data Analysis (EDA) dan Analisis Regresi pada Dataset Aplikasi Google Play Store dan Video Games Populer 1980-2023

Dibimbing oleh:

Dr. techn. Ir. Saiful Akbar, S.T, M.T.
Yuda Sukmana, S.Pd., M.T.

Disusun oleh:

Kevin Wirya Valerian	13524019
Stevanus Agustaf Wongso	13524020
Bryan Pratama Putra Hendra	13524067
Athilla Zaidan Zidna Fann	13524068
Philipp Hamara	13524101

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA - KOMPUTASI
INSTITUT TEKNOLOGI BANDUNG

MEI 2025

DAFTAR ISI

1. Pertanyaan Penelitian

1.1 Data Aplikasi pada Google Play Store

1. Bagaimana konten rating dari suatu aplikasi pada Google Play Store memengaruhi jumlah pengunduhan aplikasi?
2. Apakah benar bahwa aplikasi bertipe game lebih sering melakukan pembaruan dibandingkan tipe lainnya?
3. Bagaimana harga suatu aplikasi memengaruhi jumlah unduh dari aplikasi tersebut?
4. 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?

1.2 Data Video Game Terpopuler 1980 - 2023

1. Bagaimana tanggal rilis suatu *video game* dapat memengaruhi jumlah pemain dari game tersebut?
2. Bagaimana popularitas suatu genre *video game* berubah dari tahun ke tahun?
3. 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?
4. Apa yang membuat suatu *video game* memiliki angka *wishlist* yang tinggi? Apa saja variabel-variabel yang mungkin berperan dalam mempengaruhi angka tersebut?

2. Data dan Atribut Data

2.1 Data Aplikasi pada Google Play Store

A. Deskripsi Umum

Data ini berisi hasil *scraping data* mengenai aplikasi pada Google Play Store untuk menganalisis pasar aplikasi Android. Data ini mencakup nama aplikasi, kategori, rating, jumlah ulasan, ukuran, total pengunduhan, tipe berbayar atau gratis, harga, genre, tanggal pembaruan terakhir, versi saat ini, dan versi di Android.

Link sumber data : https://www.kaggle.com/datasets/bhavikjikadara/google-play-store-applications

Format data : CSV

Ukuran file data : 1,33 MB

B. Dimensi Data

Jumlah baris dan kolom : 10.841 baris x 13 kolom

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 *missing value*
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 *dollar*
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”

C. Sampel data

- Sampel data awal

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up
Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	26-Mar-17	1	2.3 and up
Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	26-Apr-18	1.1	4.0.3 and up
Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	14-Jun-18	6.1.61.1	4.2 and up
Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	20-Sep-17	2.9.2	3.0 and up
Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativity	3-Jul-18	2.8	4.0.3 and up
Text on Photo - Fonteee	ART_AND_DESIGN	4.4	13880	28M	1,000,000+	Free	0	Everyone	Art & Design	27-Oct-17	1.0.4	4.1 and up
Name Art Photo Editor - Focus n Filters	ART_AND_DESIGN	4.4	8788	12M	1,000,000+	Free	0	Everyone	Art & Design	31-Jul-18	1.0.15	4.0 and up
Tattoo Name On My Photo Editor	ART_AND_DESIGN	4.2	44829	20M	10,000,000+	Free	0	Teen	Art & Design	2-Apr-18	3.8	4.1 and up
Mandala Coloring Book	ART_AND_DESIGN	4.6	4326	21M	100,000+	Free	0	Everyone	Art & Design	26-Jun-18	1.0.4	4.4 and up
3D Color Pixel by Number - Sandbox Art Coloring	ART_AND_DESIGN	4.4	1518	37M	100,000+	Free	0	Everyone	Art & Design	3-Aug-18	1.2.3	2.3 and up
Learn To Draw Kawaii Characters	ART_AND_DESIGN	3.2	55	2.7M	5,000+	Free	0	Everyone	Art & Design	6-Jun-18		4.2 and up
Photo Designer - Write your name with shapes	ART_AND_DESIGN	4.7	3632	5.5M	500,000+	Free	0	Everyone	Art & Design	31-Jul-18	3.1	4.1 and up
350 Diy Room Decor Ideas	ART_AND_DESIGN	4.5	27	17M	10,000+	Free	0	Everyone	Art & Design	7-Nov-17	1	2.3 and up
FlipaClip - Cartoon animation	ART_AND_DESIGN	4.3	194216	39M	5,000,000+	Free	0	Everyone	Art & Design	3-Aug-18	2.2.5	4.0.3 and up
ibis Paint X	ART_AND_DESIGN	4.6	224399	31M	10,000,000+	Free	0	Everyone	Art & Design	30-Jul-18	5.5.4	4.1 and up
Logo Maker - Small Business	ART_AND_DESIGN	4	450	14M	100,000+	Free	0	Everyone	Art & Design	20-Apr-18	4	4.1 and up
Boys Photo Editor - Six Pack & Men's Suit	ART_AND_DESIGN	4.1	654	12M	100,000+	Free	0	Everyone	Art & Design	20-Mar-18	1.1	4.0.3 and up
Superheroes Wallpapers | 4K Backgrounds	ART_AND_DESIGN	4.7	7699	4.2M	500,000+	Free	0	Everyone 10+	Art & Design	12-Jul-18	2.2.6.2	4.0.3 and up
Mcqueen Coloring pages	ART_AND_DESIGN		61	7.0M	100,000+	Free	0	Everyone	Art & Design;Action & Adventure	7-Mar-18	1.0.0	4.1 and up
HD Mickey Minnie Wallpapers	ART_AND_DESIGN	4.7	118	23M	50,000+	Free	0	Everyone	Art & Design	7-Jul-18	1.1.3	4.1 and up

- Sampel data dengan jumlah unduhan terbanyak dan penilaian tertinggi

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
Subway Surfers	GAME	4.5	27722264	76M	1,000,000,000+	Free	0	Everyone 10+	Arcade	12-Jul-18	1.90.0	4.1 and up
Instagram	SOCIAL	4.5	66577313	Varies with d	1,000,000,000+	Free	0	Teen	Social	31-Jul-18	Varies with device	Varies with device
Google Photos	PHOTOGRAPHY	4.5	10859051	Varies with d	1,000,000,000+	Free	0	Everyone	Photography	6-Aug-18	Varies with device	Varies with device
WhatsApp Messenger	COMMUNICATION	4.4	69119316	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	3-Aug-18	Varies with device	Varies with device
Google	TOOLS	4.4	8033493	Varies with d	1,000,000,000+	Free	0	Everyone	Tools	3-Aug-18	Varies with device	Varies with device
Google Drive	PRODUCTIVITY	4.4	2731211	Varies with d	1,000,000,000+	Free	0	Everyone	Productivity	6-Aug-18	Varies with device	Varies with device
Google Chrome: Fast & Secure	COMMUNICATION	4.3	9642995	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	1-Aug-18	Varies with device	Varies with device
Gmail	COMMUNICATION	4.3	4604483	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	2-Aug-18	Varies with device	Varies with device
Google Play Games	ENTERTAINMENT	4.3	7165362	Varies with d	1,000,000,000+	Free	0	Teen	Entertainment	16-Jul-18	Varies with device	Varies with device
Maps - Navigate & Explore	TRAVEL_AND_LOCAL	4.3	9235373	Varies with d	1,000,000,000+	Free	0	Everyone	Travel & Local	31-Jul-18	Varies with device	Varies with device
YouTube	VIDEO_PLAYERS	4.3	25623548	Varies with d	1,000,000,000+	Free	0	Teen	Video Players & Editors	2-Aug-18	Varies with device	Varies with device
Google Chrome: Fast & Secure	COMMUNICATION	4.3	9642112	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	1-Aug-18	Varies with device	Varies with device
Maps - Navigate & Explore	TRAVEL_AND_LOCAL	4.3	9231613	Varies with d	1,000,000,000+	Free	0	Everyone	Travel & Local	31-Jul-18	Varies with device	Varies with device
Google Play Games	FAMILY	4.3	7168735	Varies with d	1,000,000,000+	Free	0	Teen	Entertainment	16-Jul-18	Varies with device	Varies with device
Google+	SOCIAL	4.2	4831125	Varies with d	1,000,000,000+	Free	0	Teen	Social	26-Jul-18	Varies with device	Varies with device
Google Street View	TRAVEL_AND_LOCAL	4.2	2129689	Varies with d	1,000,000,000+	Free	0	Everyone	Travel & Local	6-Aug-18	Varies with device	Varies with device
Skype - free IM & video calls	COMMUNICATION	4.1	10484169	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	3-Aug-18	Varies with device	Varies with device
Facebook	SOCIAL	4.1	78128208	Varies with d	1,000,000,000+	Free	0	Teen	Social	3-Aug-18	Varies with device	Varies with device
Messenger – Text and Video Chat for Free	COMMUNICATION	4	56642847	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	1-Aug-18	Varies with device	Varies with device
Hangouts	COMMUNICATION	4	3419249	Varies with d	1,000,000,000+	Free	0	Everyone	Communication	21-Jul-18	Varies with device	Varies with device
Google News	NEWS_AND_MAGAZINES	3.9	878065	13M	1,000,000,000+	Free	0	Teen	News & Magazines	1-Aug-18	5.2.0	4.4 and up
Google Play Movies & TV	VIDEO_PLAYERS	3.7	906384	Varies with d	1,000,000,000+	Free	0	Teen	Video Players & Editors	6-Aug-18	Varies with device	Varies with device
Android TV Remote Service	TOOLS		1	3.7M	1,000,000+	Free	0	Everyone	Tools	17-Jan-18	2.1.02.4477839	7.0 and up
SD card backup	TOOLS		142	Varies with d	1,000,000+	Free	0	Everyone	Tools	27-Mar-17	Varies with device	Varies with device
Tickets + PDA 2018 Exam	AUTO_AND_VEHICLES	4.9	197136	38M	1,000,000+	Free	0	Everyone	Auto & Vehicles	15-Jul-18	8.31	4.1 and up

2.2 Data Video Game Terpopuler 1980 - 2023

A. Deskripsi Umum

Kumpulan data ini berisi daftar permainan video mulai tahun 1980 hingga 2023, juga menyediakan hal-hal seperti tanggal rilis, peringkat ulasan pengguna, dan aspek-aspek lainnya. Data diambil dari Backloggd, sebuah *website* yang menyediakan akses bagi pengguna untuk dapat mengulas game yang pernah dimainkan.

Link sumber data : https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023

Format data : CSV

Ukuran file data : 2,87 MB

B. Dimensi Data

Jumlah baris dan kolom : 1.099 baris x 13 kolom

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

C. Sampel data

- Sampel data awal

Title	Release Date	Team	Ratir	Times	Numt	Genres	Summary	Reviews	Plays	Playing	Backlogs	Wishlist
Elden Ring	Feb 25, 2022	['Bandai Namco Entertainmen	4.5	3.9K	3.9K	['Adventure', 'RPG']	Elden Ring is a fantasy, action and open wor	["The first playthrough of elden ring is one of the best eperiences gamin	17K	3.8K	4.6K	4.8K
Hades	Dec 10, 2019	['Supergiant Games']	4.3	2.9K	2.9K	['Adventure', 'Brawler', 'Indie', 'RPG']	A rogue-lite hack and slash dungeon crawle	['convinced this is a roguelike for people who do not like the genre. The	21K	3.2K	6.3K	3.6K
The Legend of Zelda: Breath of t	Mar 03, 2017	['Nintendo', 'Nintendo EPD Pro	4.4	4.3K	4.3K	['Adventure', 'RPG']	The Legend of Zelda: Breath of the Wild is the	['This game is the game (that is not CS:GO) that I have played the most c	30K	2.5K	5K	2.6K
Undertale	Sep 15, 2015	['tobyfox', '8-4']	4.2	3.5K	3.5K	['Adventure', 'Indie', 'RPG', 'Turn Bas	A small child falls into the Underground, wh	['soundtrack is tied for #1 with nier automata. a super charming story a	28K	679	4.9K	1.8K
Hollow Knight	Feb 24, 2017	['Team Cherry']	4.4	3K	3K	['Adventure', 'Indie', 'Platform']	A 2D metroidvania with an emphasis on clos	['this games worldbuilding is incredible, with its amazing soundtrack a	21K	2.4K	8.3K	2.3K
Minecraft	Nov 18, 2011	['Mojang Studios']	4.3	2.3K	2.3K	['Adventure', 'Simulator']	Minecraft focuses on allowing the player to e	["Minecraft is what you make of it. Unfortunately theres no reason to do	33K	1.8K	1.1K	230
Omori	Dec 25, 2020	['OMOCAT', 'PLAYISM']	4.2	1.6K	1.6K	['Adventure', 'Indie', 'RPG', 'Turn Bas	A turn-based surreal horror RPG in which a c	["The best game I've played in my life", "omori is a game held up by it's f	7.2K	1.1K	4.5K	3.8K
Metroid Dread	Oct 07, 2021	['Nintendo', 'MercurySteam']	4.3	2.1K	2.1K	['Adventure', 'Platform']	Join intergalactic bounty hunter Samus Aran	['Have only been a Metroid fan for couple of years but I think this was w	9.2K	759	3.4K	3.3K
Among Us	Jun 15, 2018	['InnerSloth']	3.0	867	867	['Indie', 'Strategy']	Join your crew-mates in a multiplayer game	["it's a solid party game. i'm bad at lying though and it makes me feel b	25K	470	776	126
NieR: Automata	Feb 23, 2017	['PlatinumGames', 'Square Eni	4.3	2.9K	2.9K	['Brawler', 'RPG']	NieR: Automata tells the story of androids 2E	['Holy shit', 'im carrying the weight of the wooooooooooooooooorld',	18K	1.1K	6.2K	3.6K
Persona 5 Royal	Oct 31, 2019	['Atlus USA', 'Atlus']	4.4	2.7K	2.7K	['Adventure', 'RPG', 'Turn Based Strat	An enhanced version of Persona 5 with som	['Verdadeiro goty 2017, zelda é o caralho. Vai tomar no cu, Ryuji', 'This	12K	2.3K	5.1K	3K
Stray	Jul 19, 2022	['BlueTwelve Studio', 'Annapur	3.7	1.5K	1.5K	['Adventure', 'Indie']	Lost, alone, and separated from family, a str	['Press B to Meow', 'like, nya?', 'You are a cat, thats all.', 'There's no wa	7.7K	801	2.5K	3.4K
God of War	Apr 20, 2018	['Sony Interactive Entertainmer	4.2	2.9K	2.9K	['Adventure', 'Brawler', 'RPG']	God of War is the sequel to God of War III as	['freya te vejo como figura materna', 'i ruv !', 'One of the greatest games	21K	1.1K	4.8K	2.6K
Portal 2	Apr 18, 2011	['Valve', 'Electronic Arts']	4.4	2.9K	2.9K	['Adventure', 'Platform', 'Puzzle', 'Shc	Sequel to the acclaimed Portal (2007), Portal	['This is my fav game of all time, everything about it is amazing', 'soy mu	29K	471	3.9K	1.2K
Bloodborne	Mar 24, 2015	['FromSoftware', 'Sony Comput	4.5	3.4K	3.4K	['Adventure', 'RPG']	An action RPG in which the player embodies	["I'm not trying to brag, but Bloodborne wasn't much of a challenge for i	17K	1.1K	5.6K	3.3K
Celeste	Jan 25, 2018	['Extremely OK Games', 'Maddy	4.2	2.8K	2.8K	['Adventure', 'Indie', 'Platform']	Help Madeline survive her inner demons on t	['video games.', 'My absolute favorite platformer', 'im bad at platformin	20K	1.2K	5.9K	2K
Yakuza 0	Mar 12, 2015	['Ryū Ga Gotoku Studios', 'Sega	4.4	2.7K	2.7K	['Adventure', 'Brawler', 'RPG', 'Simula	The glitz, glamour, and unbridled decadenc	["THIS IS PEAK IT IS ONE OF MY FAVORITE GAMES ITS SO GOOD NOT ONL	15K	1.8K	6.4K	2K
Red Dead Redemption 2	Oct 26, 2018	['Take-Two Interactive', 'Rocks	4.4	2.9K	2.9K	['Adventure', 'RPG', 'Shooter']	Red Dead Redemption 2 is the epic tale of ou	['overrated as fuck', 'I wanna be a cowwwboyyyy, baby...', 'shoutout roc	19K	1.7K	5.5K	2.9K
Portal	Oct 10, 2007	['Valve', 'Electronic Arts']	4.1	2K	2K	['Platform', 'Puzzle', 'Shooter']	Waking up in a seemingly empty laboratory,	['did somebody say CLASSSSICCCC', "Portal's desolate labs with mind-b	28K	244	2.7K	1.1K
Super Mario Odyssey	Oct 27, 2017	['Nintendo']	4.2	2.9K	2.9K	['Adventure', 'Platform']	Explore incredible places far from the Mushr	["A really great game, just didn't stick out as much as it's predecessor 6	25K	710	2.9K	2K
Pokémon Legends: Arceus	Jan 28, 2022	['Nintendo', 'Game Freak']	3.7	1.6K	1.6K	['Adventure', 'RPG', 'Turn Based Strat	The Pokémon Legends: Arceus game honors	['The more mature story (by recent Pokémon standards) , lore ramificat	9.1K	1.6K	2.5K	2.1K
Hi-Fi Rush	Jan 25, 2023	['Tango Gameworks', 'Bethesd	4.3	926	926	['Adventure', 'Brawler', 'Music', 'Platf	As wannabe rockstar Chai, you'll fight back a	['It was a great game all round ending felt a little rushed', 'Música', 'Ban	3K	866	1.5K	2K
Metal Gear Rising: Revengeance	Feb 19, 2013	['Konami', 'PlatinumGames']	4.1	2.1K	2.1K	['Adventure', 'Brawler', 'Shooter', 'Str	Developed by Kojima Productions and Platir	['This game is so jank but so entertaining', 'RULES OF NATURE', 'Metal G	14K	492	4.2K	2K
Grand Theft Auto V	Sep 17, 2013	['Rockstar North', 'Rockstar Ga	3.8	2.1K	2.1K	['Adventure', 'Shooter']	Grand Theft Auto V is a vast open world gam	['People be rating this a 3.... BITCH BE SERIOUS!', 'incredible spectacle a	30K	829	3.2K	664
God of War Ragnarök	Nov 09, 2022	['Sony Interactive Entertainmer	4.4	1.7K	1.7K	['Adventure', 'Brawler']	God of War: Ragnarök is the ninth installmer	['Amazang !', 'Deep storytelling, intense gameplay, beautiful graphics.',	5.3K	801	2K	3.3K
Xenoblade Chronicles 3	Jul 29, 2022	['Monolith Soft', 'Nintendo']	4.4	1.4K	1.4K	['Adventure', 'RPG']	A vast world awaits in Xenoblade Chronicles	["Not my fav XB game, I didn't grow attached to the cast in general and c	3.9K	795	2.1K	2.2K
Kirby and the Forgotten Land	Mar 25, 2022	['HAL Laboratory', 'Nintendo']	4.2	1.6K	1.6K	['Adventure', 'Platform']	Join Kirby in an unforgettable journey throug	['Kirby and the Forgotten Land is one of the harder games for me to rate.	5.9K	955	2.5K	3.1K
Disco Elysium: The Final Cut	May 01, 2020	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'Indie', 'RPG']	Disco Elysium: The Final Cut is a groundbrea	["a captivating journey from start to finish. this uses its characters and	6K	1.2K	5K	2.7K
Marvel's Spider-Man	Sep 07, 2018	['Insomniac Games', 'Sony Inte	4.1	2.5K	2.5K	['Adventure', 'Brawler']	Starring the world's most iconic Super Hero,	['Platinum', 'This is a spiderman game', 'abartıldığı kadar iyi değil ama	21K	577	2.9K	2.2K
Dark Souls III	Mar 24, 2016	['Bandai Namco Entertainmen	4.2	2.4K	2.4K	['Adventure', 'RPG']	Dark Souls continues to push the boundarie	['The only soulsborne game I dont ever want to replay.', "Finally beat thi	19K	851	4.3K	2.2K
NieR Replicant ver.1.224744871	Apr 22, 2021	['Toylogic', 'Square Enix']	4.2	1.5K	1.5K	['Adventure', 'RPG']	NieR Replicant ver.1.22474487139... is an up	['Not as good as automata but still slaps', 'Tem dos piores game design	6.7K	880	4.1K	3.7K
Super Mario 64	Jun 23, 1996	['Nintendo', 'Nintendo EAD']	4.1	2.6K	2.6K	['Adventure', 'Platform']	The first three dimensional entry in the Mario	['We all know the importance of this game and what it brought to video	21K	463	2.5K	775

- Sampel data dengan penilaian tertinggi

Title	Release Date	Team	Rating	Times L	Number	Genres	Summary	Reviews	Plays	Playing	Backlogs	Wishlist
Pokémon Añil	Mar 23, 2023	['Eric Lost']		2	2	['RPG']	Pokémon Añil is the fourt	['pokemon HWHAT', 'Finally, a	1	0	1	7
Elden Ring: Shadow of the Erdtree	releases on TBD	['FromSoftware', 'Bandai Namco Enter	4.8	18	18	['Adventure', 'RPG']	An expansion to Elden Ri	['I really loved that they integr	1	0	39	146
Disco Elysium: The Final Cut	May 01, 2020	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'Indie', 'RPG']	Disco Elysium: The Final	["a captivating journey from s	6K	1.2K	5K	2.7K
Outer Wilds	May 28, 2019	['Mobius Digital', 'Annapurna Interacti	4.6	1.8K	1.8K	['Adventure', 'Indie', 'Puzzle', 'Sim	Outer Wilds is a critically-	['Replayed with my girlfriend,	7.7K	661	4.8K	3.1K
Disco Elysium	Oct 15, 2019	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'RPG', 'Turn Based S	A CRPG in which, waking	["I really enjoyed this one. The	4K	478	2.8K	1.9K
Umineko: When They Cry Chiru	Sep 15, 2009	['07th Expansion']	4.6	324	324	['Adventure', 'Visual Novel']	Umineko no Naku Koro n	['cried like a little bitch ngl', "(	1.7K	108	582	493
Bloodborne: The Old Hunters	Nov 24, 2015	['FromSoftware', 'Sony Computer Ente	4.6	266	266	['Adventure', 'RPG']	The Old Hunters is the fir	['HAHA FINALMENTE MATEI O	4.4K	68	930	616
Disco Elysium: The Final Cut	May 01, 2020	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'Indie', 'RPG']	Disco Elysium: The Final	["a captivating journey from s	6K	1.2K	5K	2.7K
Outer Wilds	May 28, 2019	['Mobius Digital', 'Annapurna Interacti	4.6	1.8K	1.8K	['Adventure', 'Indie', 'Puzzle', 'Sim	Outer Wilds is a critically-	['Replayed with my girlfriend,	7.7K	661	4.8K	3.1K
Disco Elysium	Oct 15, 2019	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'RPG', 'Turn Based S	A CRPG in which, waking	["I really enjoyed this one. The	4K	478	2.8K	1.9K
Umineko: When They Cry Chiru	Sep 15, 2009	['07th Expansion']	4.6	324	324	['Adventure', 'Visual Novel']	Umineko no Naku Koro n	['cried like a little bitch ngl', "(	1.7K	108	582	493
Hitman World of Assassination	Jan 26, 2023	['Inlusio Interactive', 'IO Interactive']	4.6	38	38	['Adventure', 'Shooter', 'Tactical']	Become Agent 47 in the u	['Aunque ya había jugado a lo	167	47	54	54
Disco Elysium: The Final Cut	May 01, 2020	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'Indie', 'RPG']	Disco Elysium: The Final	["a captivating journey from s	6K	1.2K	5K	2.7K
Outer Wilds	May 28, 2019	['Mobius Digital', 'Annapurna Interacti	4.6	1.8K	1.8K	['Adventure', 'Indie', 'Puzzle', 'Sim	Outer Wilds is a critically-	['Replayed with my girlfriend,	7.7K	661	4.8K	3.1K
Disco Elysium	Oct 15, 2019	['ZA/UM']	4.6	1.1K	1.1K	['Adventure', 'RPG', 'Turn Based S	A CRPG in which, waking	["I really enjoyed this one. The	4K	478	2.8K	1.9K
Final Fantasy XIV: Endwalker	Dec 07, 2021	['Square Enix']	4.6	426	426	['Adventure', 'RPG']	Final Fantasy XIV: Endwa	["Jugado hasta la historia bas	2.5K	454	683	391
Metal Gear Solid 3: Subsistence	Dec 22, 2005	['Konami Computer Entertainment Jap	4.6	641	641	['Adventure', 'Shooter', 'Tactical']	Metal Gear Solid 3: Subsi	['MGS 2 and 3 are passion pro	3.7K	50	949	639
Bloodborne: The Old Hunters	Nov 24, 2015	['FromSoftware', 'Sony Computer Ente	4.6	266	266	['Adventure', 'RPG']	The Old Hunters is the fir	['HAHA FINALMENTE MATEI O	4.4K	68	930	616
Final Fantasy XIV: Shadowbringers	Jul 02, 2019	['Square Enix']	4.6	401	401	['RPG']	SHADOWBRINGERS is th	['"The rains have ceased, and	3K	192	517	267
The Great Ace Attorney 2: Resolve	Aug 03, 2017	['Capcom']	4.6	386	386	['Adventure', 'Point-and-Click', 'P	The Great Ace Attorney 2:	["Greatest of all time. Zenith c	1.1K	75	771	468
Sekiro: Shadows Die Twice - GOTY Edition	Oct 28, 2020	['Activision', 'FromSoftware']	4.6	173	173	['Adventure', 'Brawler', 'RPG']	Carve your own clever pa	['Simplesmente uma perfeiçã	1.4K	153	461	373
Half-Life: Alyx	Mar 23, 2020	['Valve']	4.6	515	515	['Adventure', 'Puzzle', 'Shooter']	Half-Life: Alyx is Valve's V	["It's like I was actually there.	1.9K	210	1.2K	1.4K
Dwarf Fortress	Dec 06, 2022	['Bay 12 Games', 'Kitfox Games']	4.6	48	48	['Indie', 'RPG', 'Simulator', 'Strate	In this complex construc	['Both a fascinating, awe-insp	225	80	181	195
Metal Gear Solid 3: Snake Eater HD Edition	Nov 08, 2011	['Konami Digital Entertainment']	4.6	293	293	['Adventure', 'Shooter']	Get Metal Gear Solid 3: Sı	['Após um pequeno grande sı	2.5K	34	554	244
Tokyo Necro	Jan 29, 2016	['Nitroplus', 'JAST USA']	4.6	21	21	['Adventure', 'Visual Novel']	Nemo ante mortem beatu	['the indifferent cruelty of the	14	22	50	69
Bloodborne: Game of the Year Edition	Nov 27, 2015	['Sony Computer Entertainment', 'Fron	4.6	238	238	['RPG']	With new story details, le	['O melhor jogo da história', 'A	1.5K	57	323	255

3. Visualisasi

3.1 Data Aplikasi pada Google Play Store

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

A. 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 .xslx.
   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*.

B. 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, ",", ""), "+", ""))`

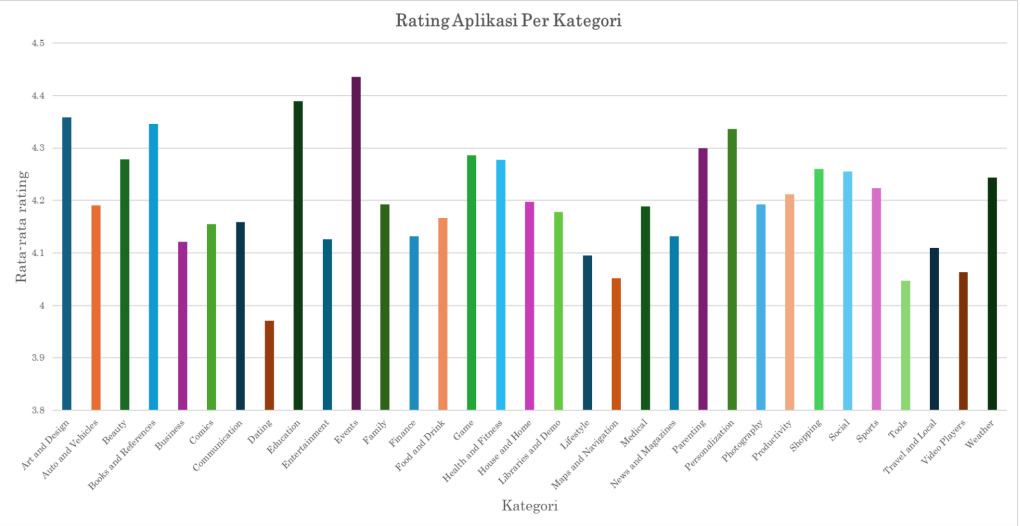
C. Pada data mentah, format data Price berupa value string karena terdapat $ di depannya, untuk standardisasi value Price, maka diperbuat kolom Price_Number dengan menggunakan formula `=VALUE(SUBSTITUTE(I2,"$",""))`

3.1.1 Visualisasi 1: Rata-rata Rating Berdasarkan Kategori

Proses Pembuatan:

1. Pilih kolom Kategori dan kolom Rating
2. Membuat PivotTable dengan Kategori sebagai rows, dan valuesnya sebagai Average of Rating. Hasilnya, PivotTable akan menunjukkan rata-rata rating untuk setiap kategori.
3. Buat visualisasi dengan grafik batang dengan sumbu-x adalah Kategori dan sumbu-y adalah rata-rata Rating
4. Memberikan label yang sesuai untuk visualisasi yang telah dibuat

Hasil Visualisasi



Hasil Insight

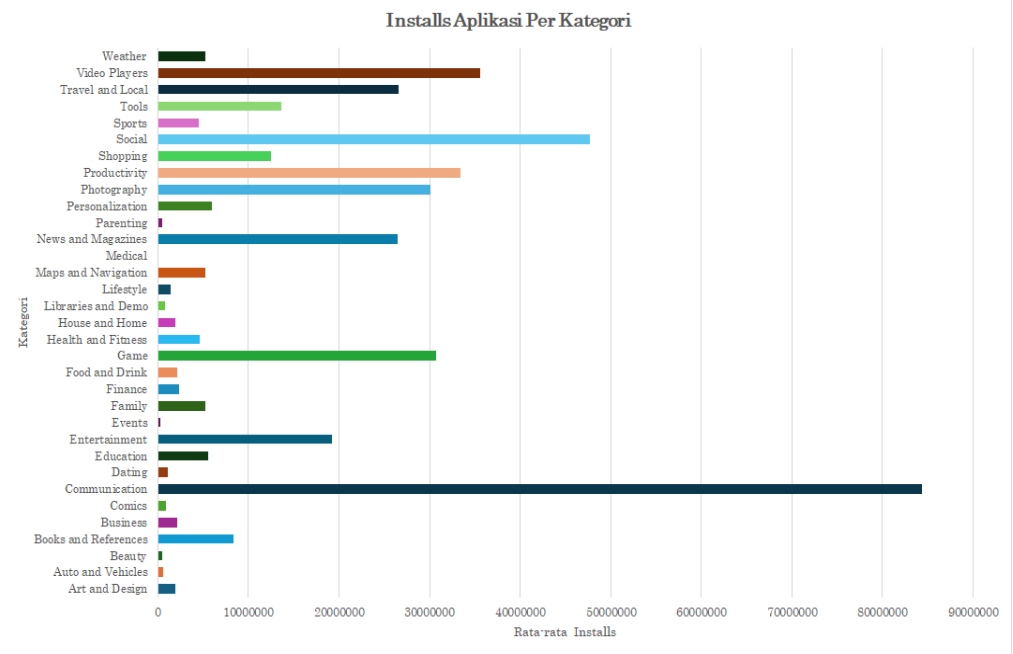
Berdasarkan grafik "Rating Aplikasi Per Kategori", dapat disimpulkan bahwa kategori aplikasi dengan rata-rata rating tertinggi adalah Events, Education, dan Art and Design, menunjukkan bahwa pengguna sangat puas terhadap aplikasi dalam kategori tersebut. Sebaliknya, kategori Dating, Maps and Navigation, 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

Proses Pembuatan:

1. Pilih kolom Kategori dan kolom Rating
2. Membuat PivotTable dengan Kategori sebagai rows, dan valuesnya sebagai Average of Installs_Clean. Hasilnya, PivotTable akan menunjukkan rata-rata Installs untuk setiap kategori.
3. Buat visualisasi dengan grafik batang dengan sumbu-x adalah rata-rata Installs dan sumbu-y adalah Kategori
4. Memberikan label yang sesuai untuk visualisasi yang telah dibuat

Hasil Visualisasi



Hasil Insight

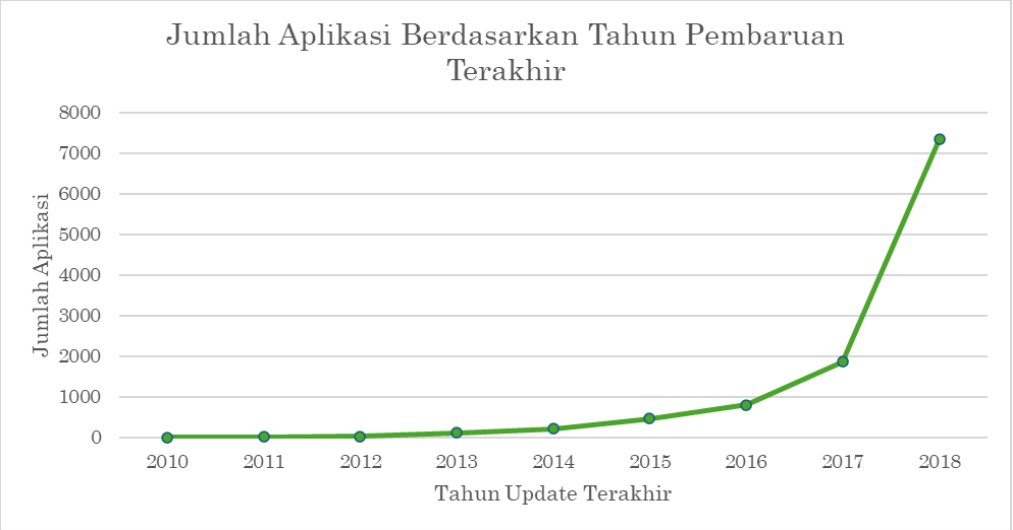
Berdasarkan grafik "Installs Aplikasi Per Kategori", terlihat bahwa kategori *Communication*, *Game*, dan *Social* mendominasi jumlah rata-rata instalasi, menandakan tingginya kebutuhan pengguna terhadap aplikasi yang bersifat interaktif dan hiburan. Sementara itu, kategori seperti *Parenting*, *Events*, 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

Proses Pembuatan:

1. Pilih kolom LastUpdated
2. Membuat PivotTable dengan Years dari kolom LastUpdated sebagai rows, dan valuesnya sebagai jumlah dari aplikasi. Hasilnya, PivotTable akan menunjukkan banyak aplikasi untuk setiap tahun pembaruan terakhir.
3. Buat visualisasi dengan grafik line dengan sumbu-x adalah tahun LastUpdated dan sumbu-y adalah Jumlah Aplikasi
4. Memberikan label yang sesuai untuk visualisasi yang telah dibuat

Hasil Visualisasi



Hasil Insight

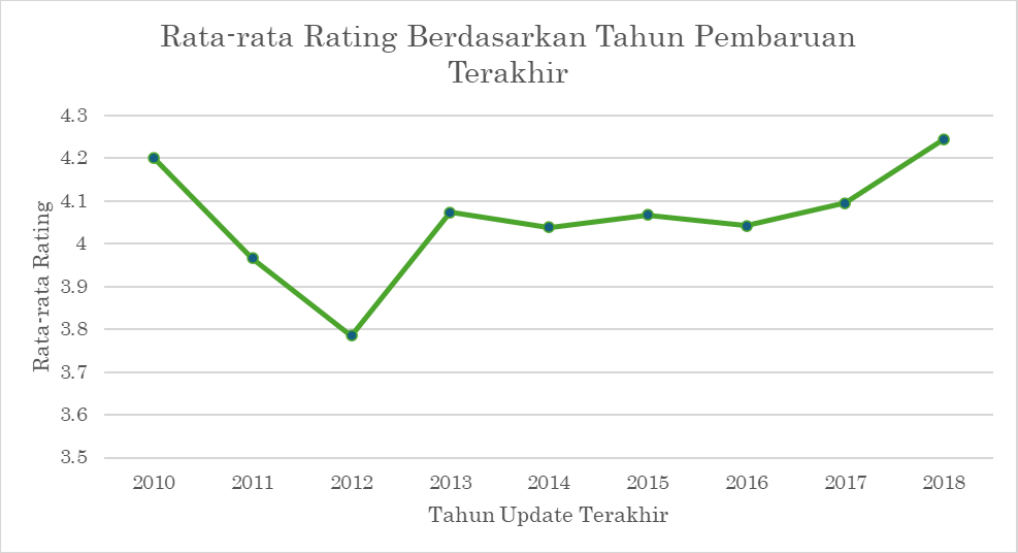
Lonjakan tajam jumlah aplikasi yang memiliki pembaruan terakhir di tahun 2018 mencerminkan kondisi aktual saat data diambil, yakni sekitar tahun tersebut, sehingga mayoritas aplikasi yang masih aktif atau relevan pada saat itu menunjukkan status *last updated* di 2018. Kecenderungan ini mengindikasikan 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

Proses Pembuatan:

1. Pilih kolom LastUpdated dan kolom Rating
2. Membuat PivotTable dengan Years dari kolom LastUpdated sebagai rows, dan valuesnya sebagai rata-rata rating. Hasilnya, PivotTable akan menunjukkan rata-rata rating aplikasi untuk setiap tahun pembaruan terakhir.
3. Buat visualisasi dengan grafik line dengan sumbu-x adalah tahun pembaruan terakhir dan sumbu-y adalah rata-rata Installs
4. Memberikan label yang sesuai untuk visualisasi yang telah dibuat

Hasil Visualisasi



Hasil Insight

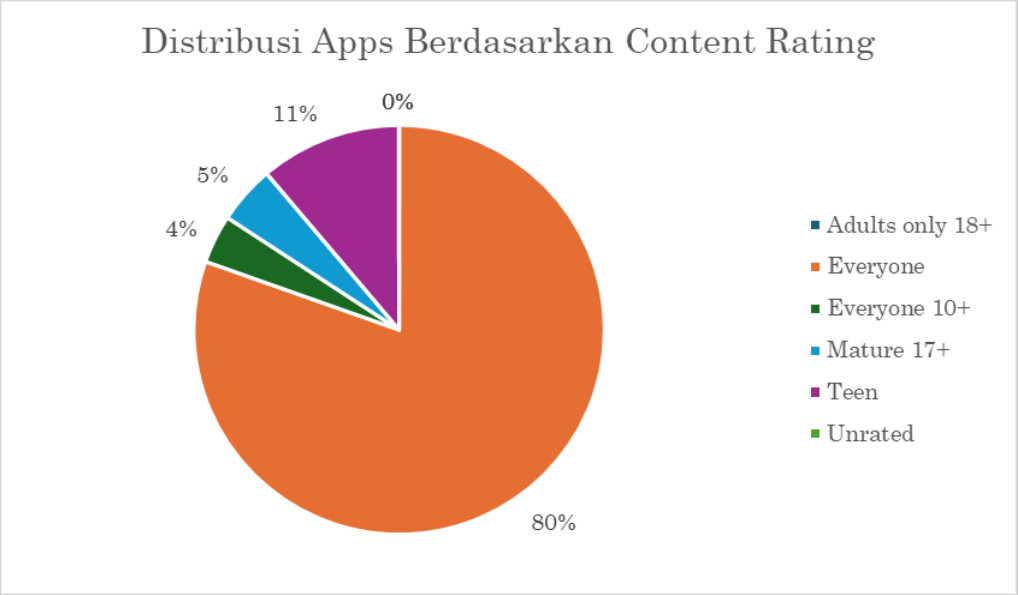
Grafik rata-rata rating berdasarkan tahun pembaruan terakhir menunjukkan adanya tren peningkatan kualitas aplikasi dari waktu ke waktu, dengan titik terendah terjadi pada tahun 2012 dan lonjakan signifikan setelahnya hingga mencapai puncak pada 2018. Pola ini mengindikasikan bahwa aplikasi yang lebih 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

Proses Pembuatan:

1. Pilih kolom Content Rating
2. Membuat PivotTable dengan Content Rating sebagai rows dan valuesnya sebagai count of apps. Hasilnya, PivotTable akan menunjukkan jumlah apps untuk setiap jenis Content Rating
3. Buat visualisasi pie chart yang mencerminkan distribusi apps berdasarkan content rating
4. Memberikan label-label yang sesuai untuk visualisasi tersebut

Hasil Visualisasi



Hasil Insight

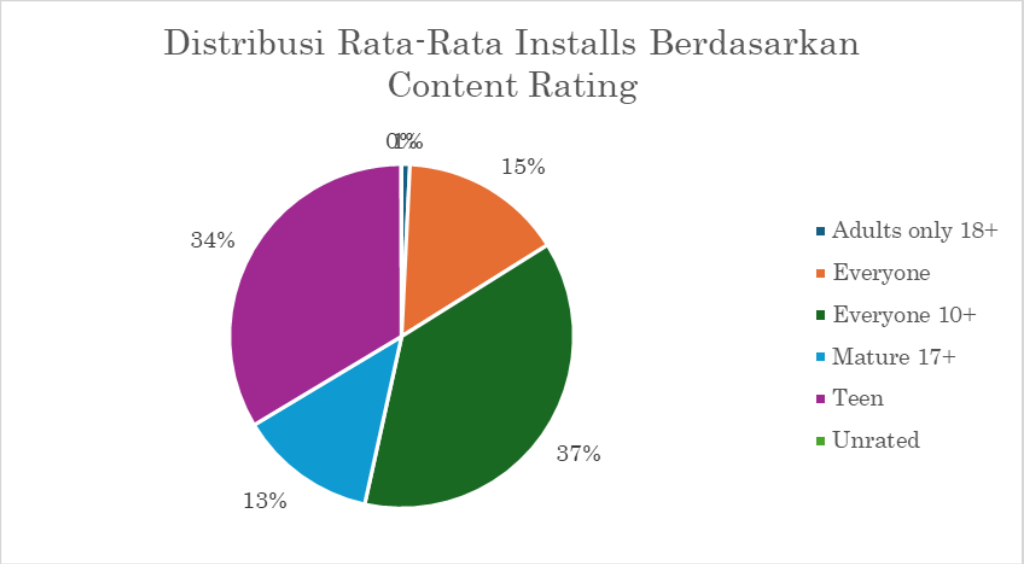
Berdasarkan pie chart distribusi aplikasi menurut content rating, terlihat bahwa mayoritas besar aplikasi di Google Play Store ditujukan untuk kategori **"Everyone"**, yaitu sebesar **80%** dari total aplikasi. Sementara itu, kategori lain seperti **"Teen"** (11%), **"Mature 17+"** (5%), dan **"Everyone 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

Proses Pembuatan:

1. Pilih kolom Content Rating dan kolom Installs_Clean
2. Membuat PivotTable dengan Content Rating sebagai rows dan valuesnya sebagai rata-rata dari Installs_Clean . Hasilnya, PivotTable akan menunjukkan rata-rata installs untuk setiap jenis Content Rating
3. Buat visualisasi pie chart yang mencerminkan distribusi rata-rata installs berdasarkan content rating
4. Memberikan label-label yang sesuai untuk visualisasi tersebut

Hasil Visualisasi



Hasil Insight

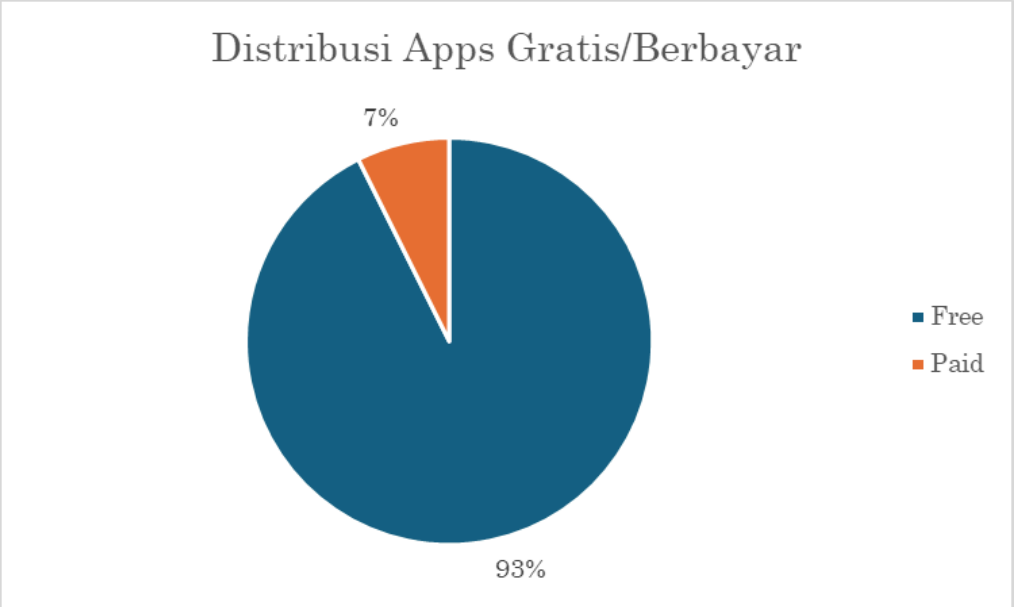
Meskipun jumlah aplikasi dengan rating "**Everyone**" mendominasi (seperti terlihat di pie chart sebelumnya), justru kategori ***"Everyone 10+"*** dan ***"Teen"*** memiliki rata-rata instalasi tertinggi per aplikasi, menandakan bahwa aplikasi-aplikasi yang ditujukan untuk remaja dan usia 10 tahun ke 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

Proses Pembuatan:

1. Pilih kolom Type
2. Membuat PivotTable dengan Type sebagai rows dan valuesnya sebagai count of apps. Hasilnya, PivotTable akan menunjukkan jumlah apps untuk jenis gratis/berbayar
3. Buat visualisasi pie chart yang mencerminkan distribusi apps berdasarkan jenis gratis/berbayar
4. Memberikan label-label yang sesuai untuk visualisasi tersebut

Hasil Visualisasi



Hasil Insight

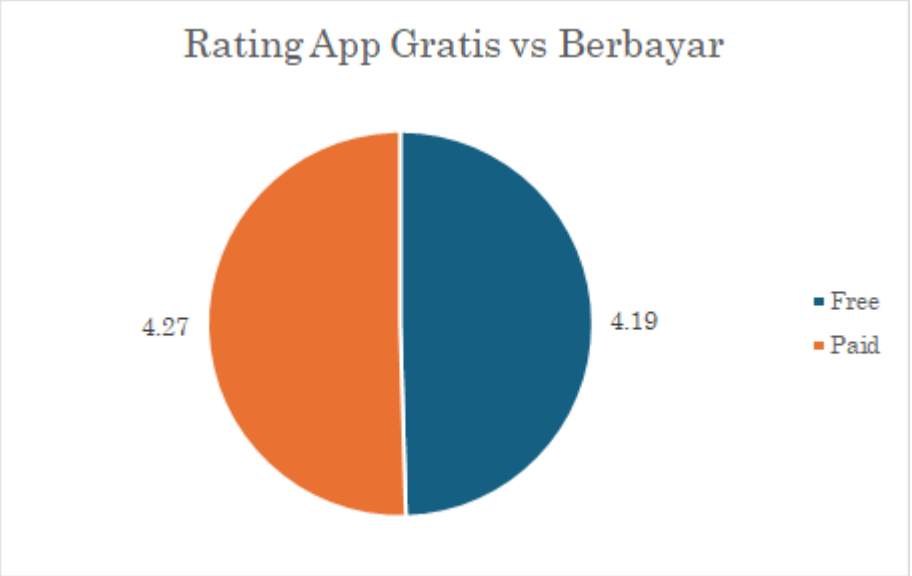
Grafik ini menunjukkan bahwa mayoritas aplikasi di Google Play Store adalah aplikasi gratis, yakni sebesar 93%, sementara hanya 7% aplikasi yang berbayar. Hal ini menandakan bahwa model distribusi aplikasi di platform ini sangat mengandalkan pendekatan freemium atau monetisasi berbasis iklan dan pembelian 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

Proses Pembuatan:

1. Pilih kolom Type dan kolom Rating
2. Membuat PivotTable dengan Type sebagai rows dan valuesnya sebagai average of rating. Hasilnya, PivotTable akan menunjukkan rata-rata ratings untuk jenis gratis/berbayar
3. Buat visualisasi pie chart yang mencerminkan distribusi rata-rata ratings berdasarkan jenis gratis/berbayar
4. Memberikan label-label yang sesuai untuk visualisasi tersebut

Hasil Visualisasi



Hasil Insight

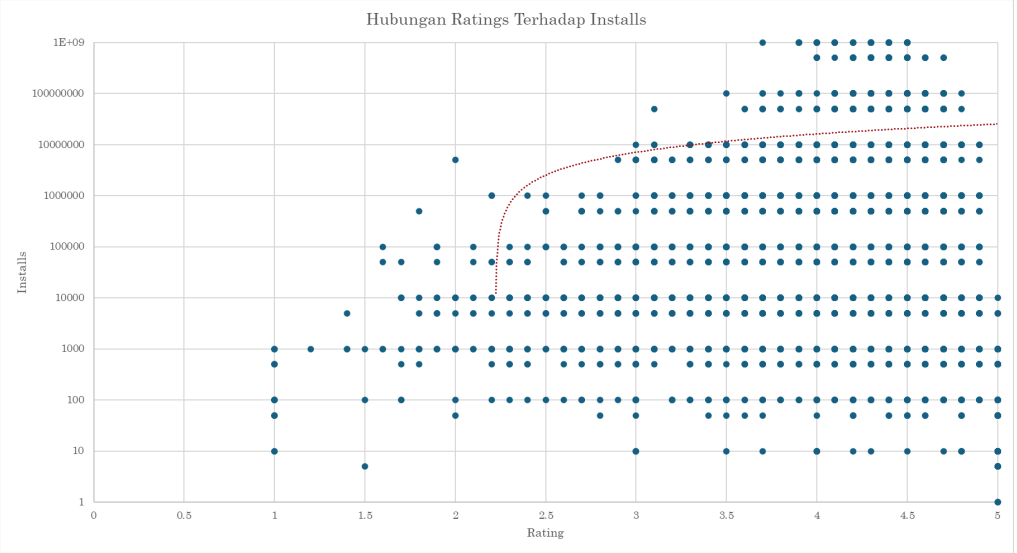
Meskipun aplikasi gratis mendominasi jumlah di Google Play Store, grafik ini menunjukkan bahwa **aplikasi berbayar memiliki rata-rata rating yang sedikit lebih tinggi** dibandingkan aplikasi gratis — yaitu **4.27** untuk berbayar versus **4.19** untuk gratis. Hal ini bisa menunjukkan bahwa pengguna cenderung 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

Proses Pembuatan:

1. Pilih kolom Rating dan kolom Installs_Clean
2. Buat Scatter Plot dengan Rating dan Installs_Clean
3. Tampilkan trendline untuk menunjukkan korelasi rating dengan jumlah installs
4. Berikan label-label yang sesuai untuk visualisasi

Hasil Visualisasi



Hasil Insight

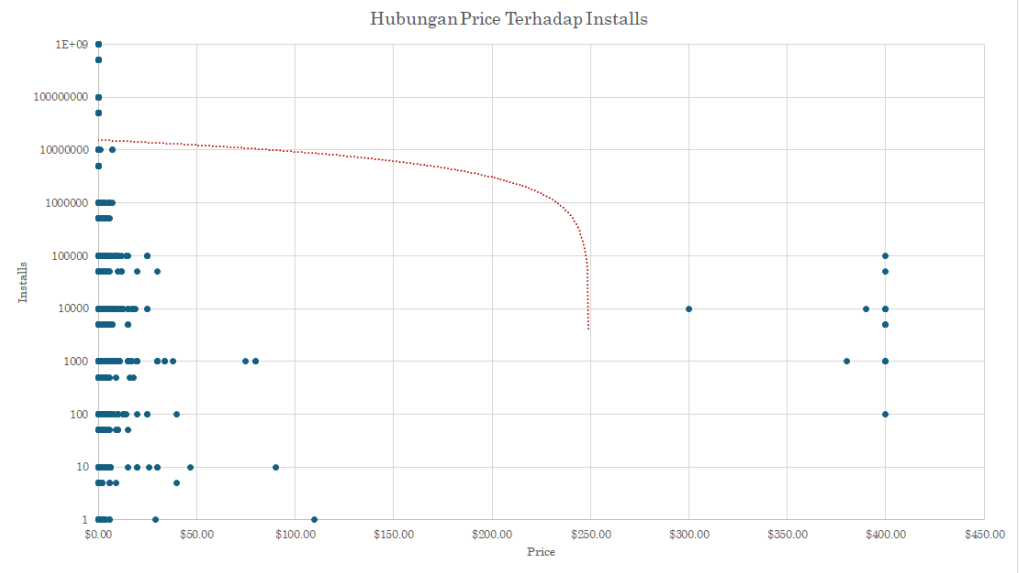
Dari data yang disajikan, dapat disimpulkan bahwa terdapat korelasi positif antara tinggi rendahnya ratings dengan jumlah installs. Semakin tinggi rating yang diberikan pengguna, semakin besar pula kemungkinan aplikasi tersebut untuk diunduh. Hal ini menunjukkan bahwa reputasi dan kepuasan pengguna, 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

Proses Pembuatan:

1. Pilih kolom Installs_Clean dan kolom Price
2. Buat Scatter Plot dengan Installs_Clean dan Price
3. Tampilkan trendline untuk menunjukkan korelasi Installs_Clean dengan Price
4. Berikan label-label yang sesuai untuk visualisasi

Hasil Visualisasi



Hasil Insight

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.

3.2 Data Video Game Terpopuler 1980 - 2023

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

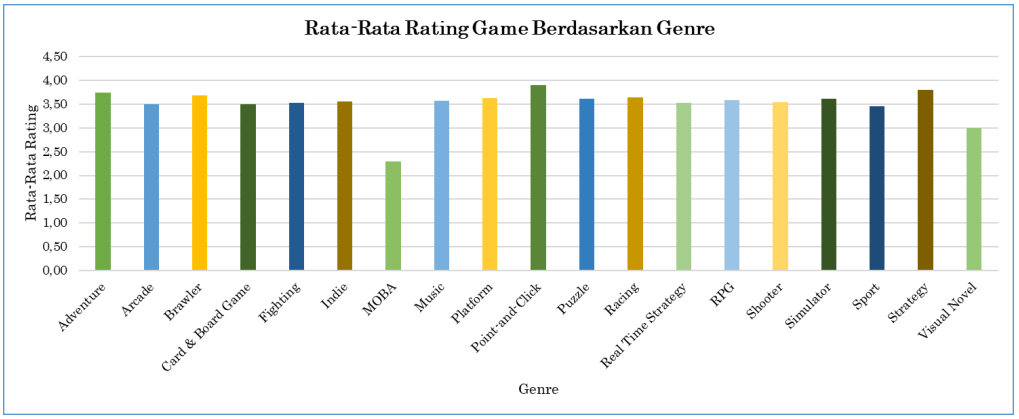
1. Pembuatan dilakukan pada Ms. Excel dengan database terdiri atas 5 sheets, yaitu
   a. SheetUtama, berisikan data mentah yang diambil dari sumber data langsung dalam format file .csv yang sudah saya konversikan menjadi .xslx.
   b. PerbandinganKategori, bersihkan data dan visualisasi untuk perbandingan 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. Angka Plays merujuk pada jumlah seluruh pemain yang pernah memainkan video game sebelumnya.
3. Angka Playing merujuk pada jumlah pemain aktif, yaitu pemain yang masih memainkan video game hingga sekarang.
4. Angka Backlogs merujuk pada jumlah pemain pasif, yaitu pemain yang memiliki akses terhadap video game, tetapi belum memainkannya.
5. Pada data mentah, kolom Genre terdiri atas beberapa genre. Oleh karena itu, kami mengambil genre utama saja dari video game tersebut dengan mengambilkan 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'].
6. 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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Rating.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!E:E) untuk menghitung rata-rata rating dengan XX merupakan sel Genre yang ingin dicari rata-rata genre ratingnya. (di SheetUtama, O adalah kolom Genre dan E adalah kolom Rating).
4. Buat visualisasi dengan tipe *bar chart* dengan sumbu-x adalah Genre dan sumbu-y adalah rata-rata rating.

Hasil Visualisasi



Hasil Insight

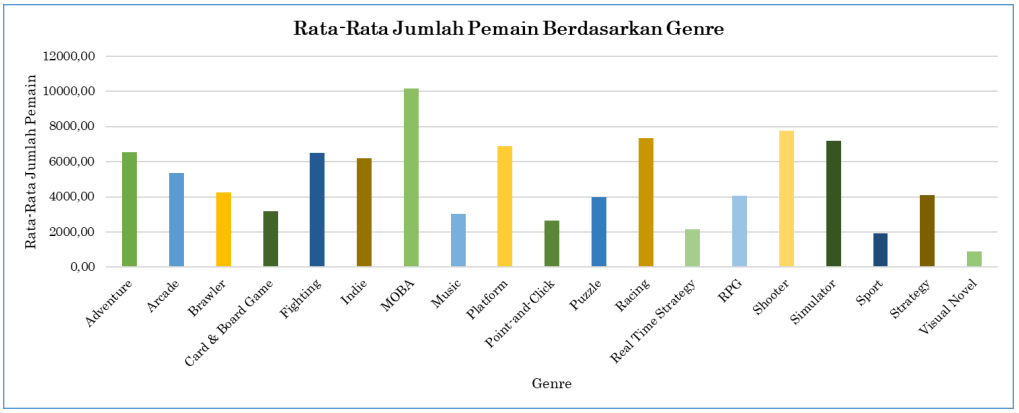
Berdasarkan diagram, video game dengan rating tertinggi memiliki genre Point-and-Click disusul oleh video game dengan Strategy dan Adventure, pada posisi 2 dan 3, berturut-turut. Genre game dengan rating paling rendah adalah MOBA. Secara umum, rata-rata rating game berada di sekitar angka 3,50.

3.2.2 Visualisasi 2: Rata-Rata Jumlah Pemain Berdasarkan Genre

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Plays.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!K:K) untuk menghitung rata-rata rating dengan XX merupakan sel Genre yang ingin dicari rata-rata jumlah pemainnya. (di SheetUtama, O adalah kolom Genre dan K adalah kolom Plays).
4. Buat visualisasi dengan tipe *bar chart* dengan sumbu-x adalah Genre dan sumbu-y adalah Rata-Rata Jumlah Pemain.

Hasil Visualisasi



Hasil Insight

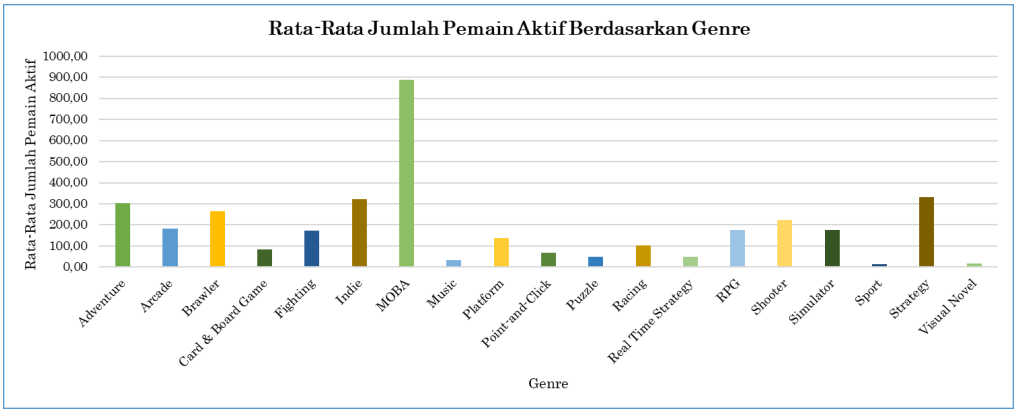
Berdasarkan grafik, MOBA adalah genre video game dengan rata-rata jumlah pemain paling tinggi disusul oleh game dengan genre Shooter dan Racing. Tiga genre video game dengan rata-rata jumlah pemain terendah adalah Real-Time Strategy, Sport, dan Visual Novel.

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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Playing.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!L:L) untuk menghitung rata-rata rating dengan XX merupakan sel Genre yang ingin dicari rata-rata jumlah pemainnya. (di SheetUtama, O adalah kolom Genre dan L adalah kolom Playing).
4. Buat visualisasi dengan tipe *bar chart* dengan sumbu-x adalah Genre dan sumbu-y adalah Rata-Rata Jumlah Pemain Aktif.

Hasil Visualisasi



Hasil Insight

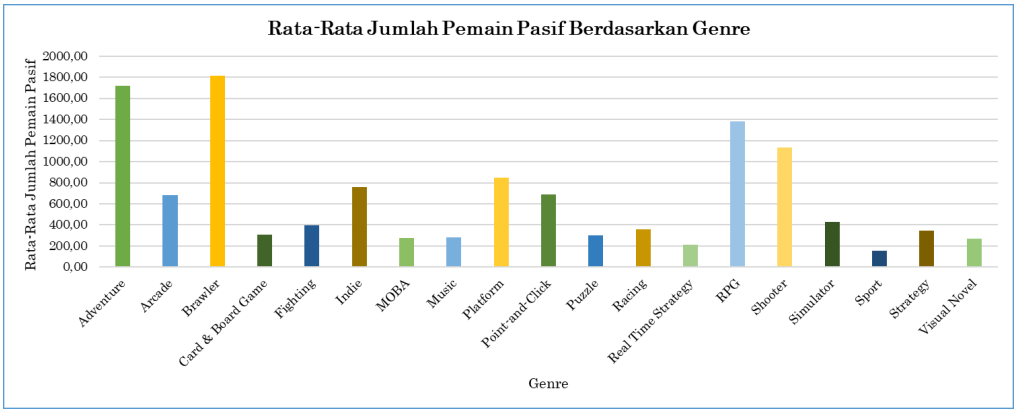
Berdasarkan diagram, MOBA adalah genre video game dengan rata-rata jumlah pemain aktif paling tinggi di angka hampir 900. Hal ini berbeda jauh dengan rata-rata jumlah pemain aktif pada video genre game lainnya. Pada posisi kedua, genre video game Strategy memiliki rata-rata jumlah pemain aktif hanya 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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Backlogs.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!M:M) untuk menghitung rata-rata rating dengan XX merupakan sel Genre yang ingin dicari rata-rata jumlah pemainnya. (di SheetUtama, O adalah kolom Genre dan M adalah kolom Backlogs).
4. Buat visualisasi dengan tipe *bar chart* dengan sumbu-x adalah Genre dan sumbu-y adalah Rata-Rata Jumlah Pemain Pasif.

Hasil Visualisasi



Hasil Insight

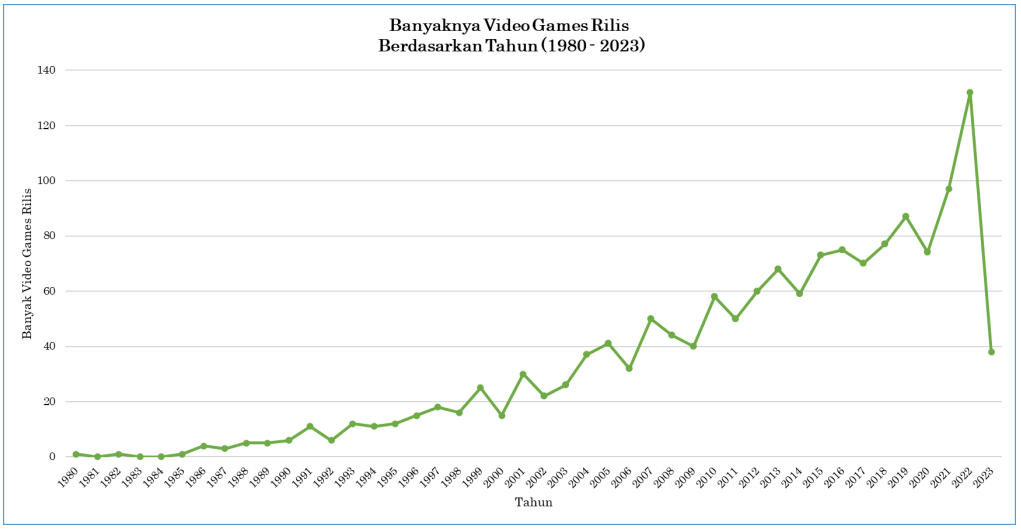
Berdasarkan diagram, Brawler adalah genre video game dengan rata-rata jumlah pemain pasif tertinggi di angka 1800, diikuti oleh Adventure dan RPG, pada posisi ke-2 dan ke-3 secara berturut-turut. Video game dengan rata-rata jumlah pemain pasif terendah adalah game dengan genre Sport. Secara umum, terdapat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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Tahun (1980 - 2023) dan Banyak Game Rilis.
2. Gunakan rumus =YEAR(Sheet1!XX) untuk mengambil tahun dari sel tertentu dengan XX merupakan sel dengan format tanggal DD/MM/YY.
3. Gunakan rumus =COUNTIF(SheetUtama!P:P; XX) untuk menghitung banyaknya kemunculan tahun tertentu di kolom P pada SheetUtama; XX merupakan sel yang menampilkan tahun tertentu.
4. Buat visualisasi dengan *line chart* dengan sumbu-x adalah Tahun dan sumbu-y merupakan Banyak Video Game Rilis.

Hasil Visualisasi



Hasil Insight

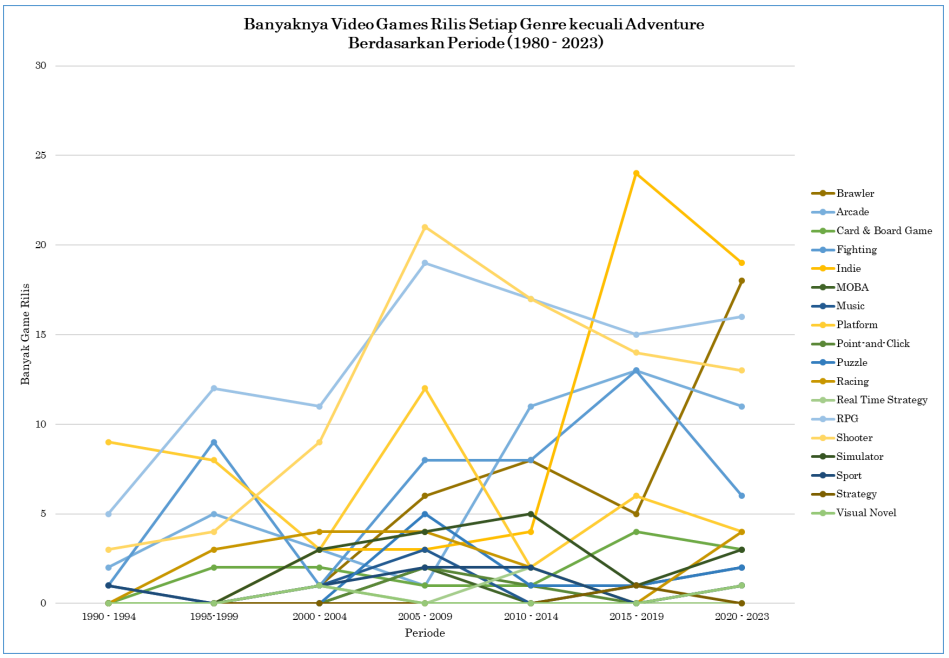
Berdasarkan grafik, peningkatan video game yang dirilis tertinggi pada tahun 2020 menuju tahun 2021. Selain itu, dapat dilihat bahwa banyak video game yang dirilis dari tahun ke tahun meningkat secara fluktuatif. Ada tahun-tahun saat video game mengalami penurunan, tetapi penurunan minor ini tidak sebanding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

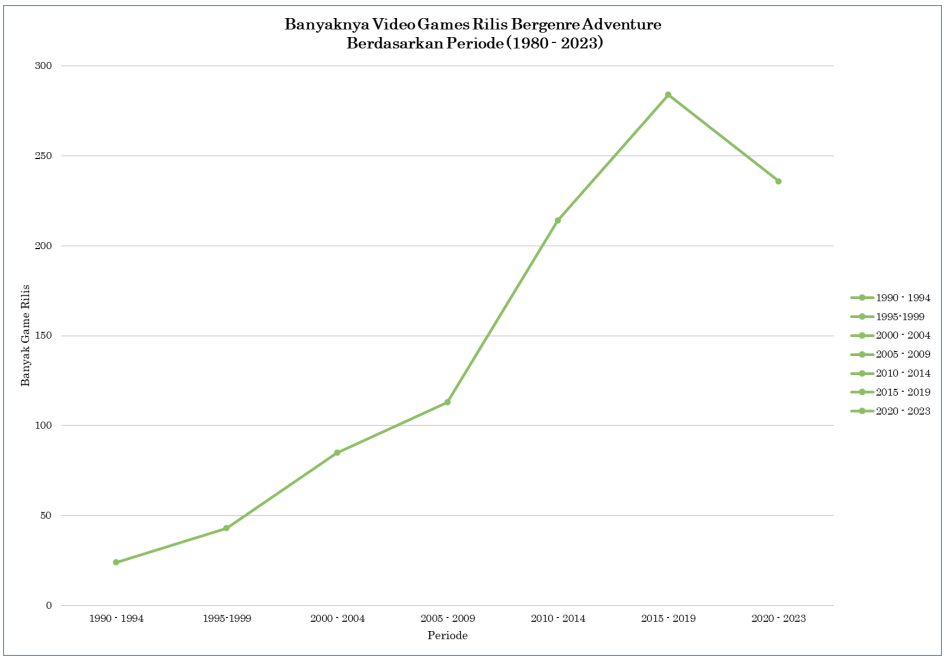
3.2.6 Visualisasi 6: Banyaknya Video Games Rilis Setiap Genre Berdasarkan Periode (1980 - 2023)

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari beberapa kolom, Genre, lalu periode dari 1990 - 1994, 1995 - 1999, .., hingga 2020 - 2023.
2. Gunakan rumus =COUNTIFS(SheetUtama!O:O; $XX; SheetUtama!P:P; ">=YYY1"; SheetUtama!P:P; "<=YYY2") untuk mengambil banyaknya rilis video game dengan genre yang berada pada sel XX dari tahun YYY1 hingga YYY2. Ingat kembali bahwa O adalah kolom genre pada SheetUtama dan P adalah kolom tahun.
3. Buat visualisasi dengan grafik garis dengan sumbu-x adalah Tahun dan sumbu-y merupakan Banyak Video Game Rilis.
4. Perbedaan warna *line chart* di sini digunakan untuk memudahkan visualisasi pembaca terkait data yang disajikan.
5. Selain itu, khusus untuk genre Adventure, kami melakukan pendekatan berbeda dikarenakan game bergenre Adventure yang dirilis terlampau banyak sehingga perlu dilakukan visualisasi pada grafik berbeda.

Hasil Visualisasi





Hasil Insight

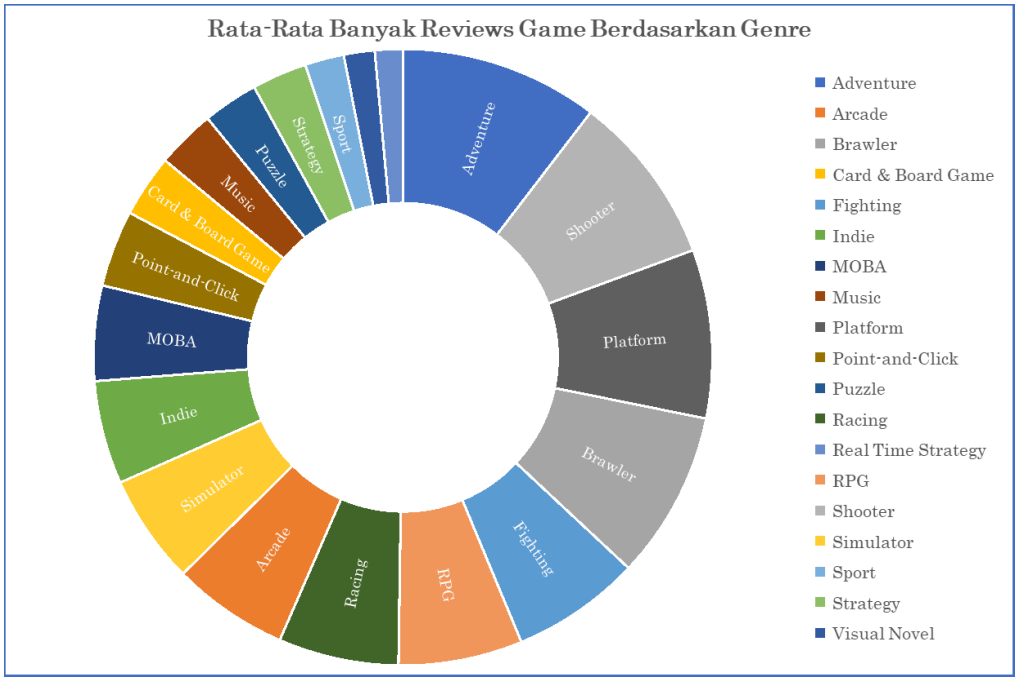
Berdasarkan grafik, video game bergenre Adventure mendominasi rilis dari periode satu ke periode lainnya. Selanjutnya, beberapa hal yang menarik perhatian adalah video game bergenre Indie yang mengalami peningkatan angka perilisan drastis dari periode 5 tahun, 2010 - 2014 ke 2015 - 2019. Banyak video 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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Number of Reviews.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!G:G) untuk menghitung rata-rata banyak review dengan XX merupakan sel Genre yang ingin dicari rata-rata banyak reviewnya. (di SheetUtama, O adalah kolom Genre dan G adalah kolom Reviews).
4. Buat visualisasi dengan tipe *Sunburst*.

Hasil Visualisasi



Hasil Insight

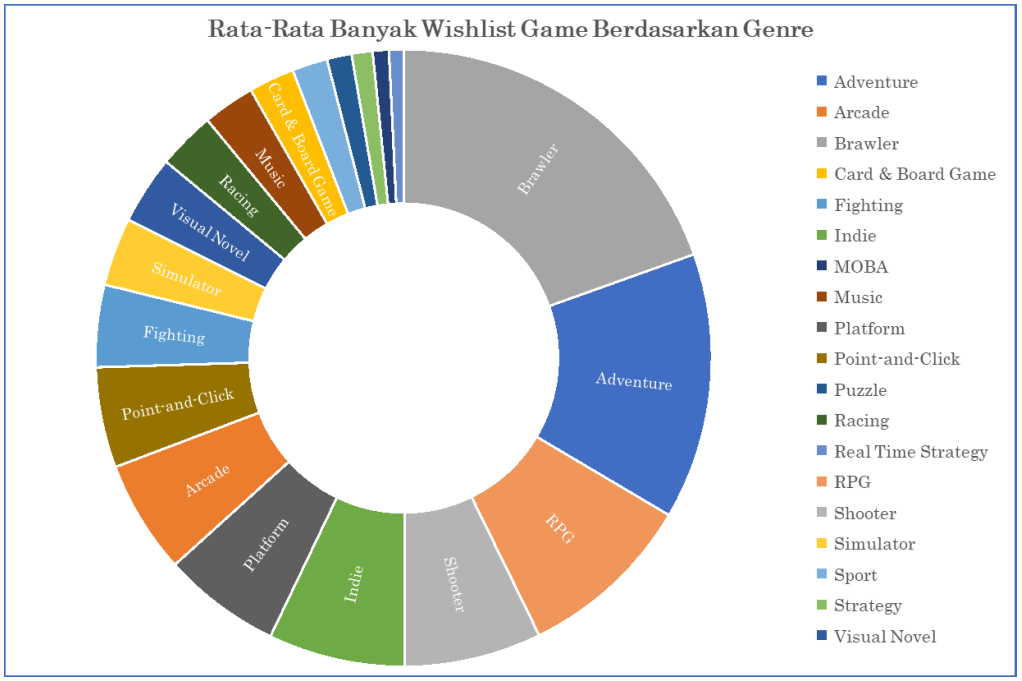
Berdasarkan diagram, video game bergenre Adventure memiliki rata-rata reviews paling tinggi di antara genre lainnya. Hal ini kemudian disusul oleh video game bergenre Shooter dan Platform, pada peringkat 2 dan 3, secara berturut-turut. Video game bergenre Visual Novel dan Real Time Strategy memiliki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

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

Proses Pembuatan:

1. Buat suatu tabel yang terdiri dari 2 kolom, Genre dan Reviews.
2. List semua genre yang ada.
3. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;XX;SheetUtama!N:N) untuk menghitung rata-rata banyak review dengan XX merupakan sel Genre yang ingin dicari rata-rata banyak wishlistnya. (di SheetUtama, O adalah kolom Genre dan N adalah kolom Wishlist).
4. Buat visualisasi dengan tipe *Sunburst.*

Hasil Visualisasi



Hasil Insight

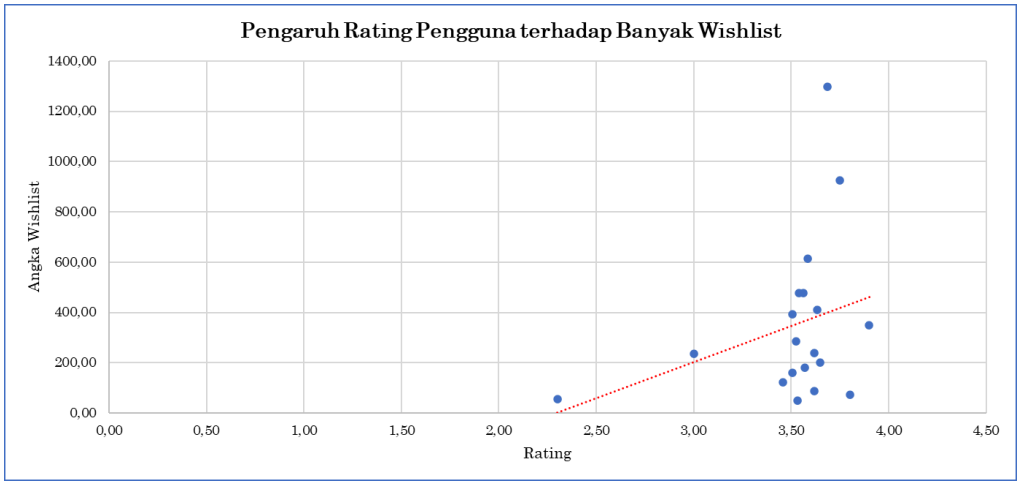
Berdasarkan diagram, banyak pemain video game yang memasukkan video game bergenre Brawler ke dalam wishlist mereka, disusul oleh video game bergenre Adventure, RPG, dan Shooter pada tiga peringkat setelahnya. Beberapa video game yang jarang dimasukkan ke dalam wishlist adalah genre Sport, Visual Novel,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

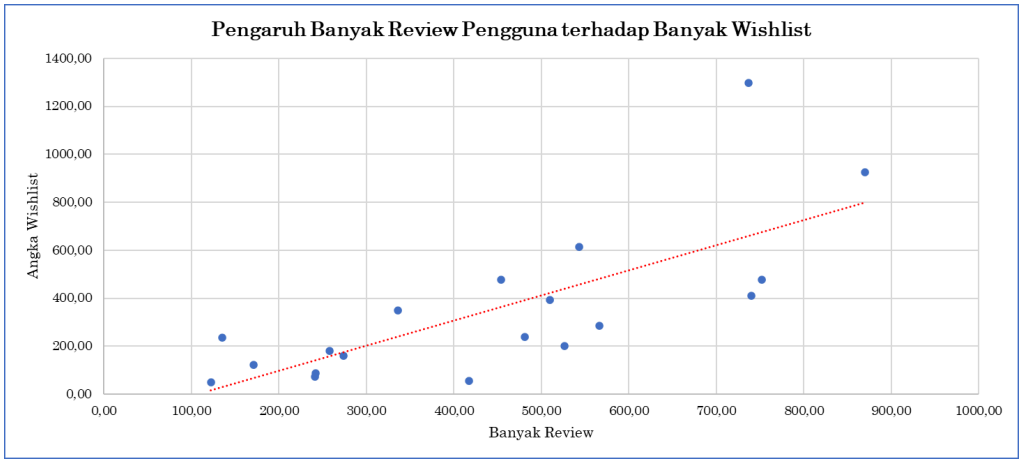
3.2.9 Visualisasi 9: Variabel Bebas yang Memengaruhi Angka Wishlist

Proses Pembuatan:

1. Amati variabel bebas apa saja yang memengaruhi wishlist pengguna. Kemudian, buat tabel berdasarkan variabel-variabel bebas tersebut.
2. Buat suatu tabel yang terdiri dari 4 kolom: Genre, Rating, Number of Reviews, dan Wishlist.
3. List semua genre yang ada.
4. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;B3;SheetUtama!E:E) untuk menghitung rata-rata rating dengan XX merupakan sel Genre yang ingin dicari rata-rata ratingnya. (di SheetUtama, O adalah kolom Genre dan E adalah kolom Rating).
5. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;B3;SheetUtama!G:G) untuk menghitung rata-rata banyak review dengan XX merupakan sel Genre yang ingin dicari rata-rata banyak reviewnya. (di SheetUtama, O adalah kolom Genre dan G adalah kolom Reviews).
6. Gunakan rumus =AVERAGEIF(SheetUtama!O:O;B3;SheetUtama!N:N) untuk menghitung rata-rata wishlistnya dengan XX merupakan sel Genre yang ingin dicari rata-rata wishlistnya. (di SheetUtama, O adalah kolom Genre dan E adalah kolom Wishlist).
7. Buat dua buah *scatter plot diagram* yang menunjukkan korelasi antara variabel Rating dan Wishlist, kemudian Number of Reviews dan Wishlist.
8. Pada diagram pertama, sumbu-x adalah Rating dan sumbu-y adalah Angka Wishlist. Di sisi lain, pada diagram kedua, sumbu-x adalah Banyak Review dan sumbu-y adalah Angka Wishlist.

Hasil Visualisasi





Hasil Insight

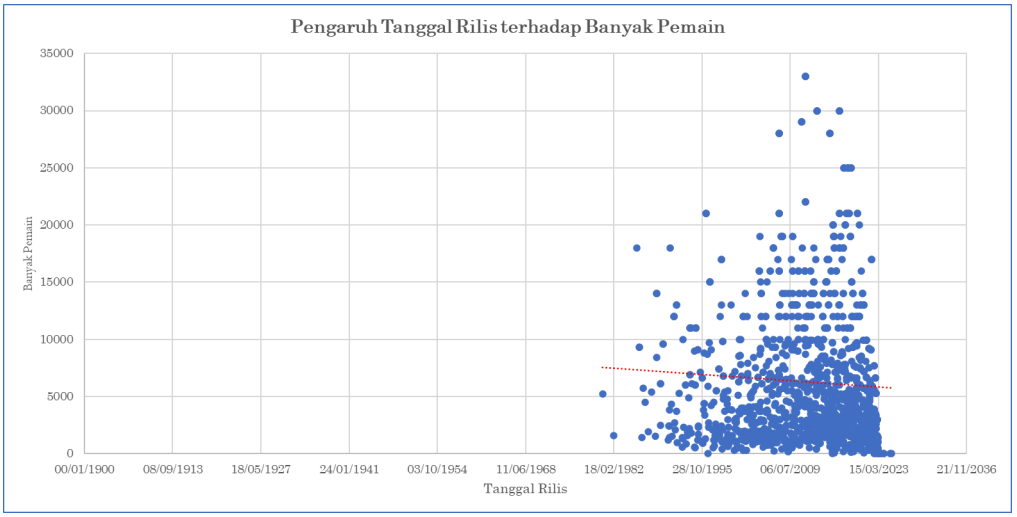
Berdasarkan kedua diagram, dapat disimpulkan bahwa rating pengguna dan banyak review memengaruhi banyak wishlist. Hal ini ditunjukkan dengan adanya korelasi positif antara kedua pasangan variabel bebas dan terikat yang ada.

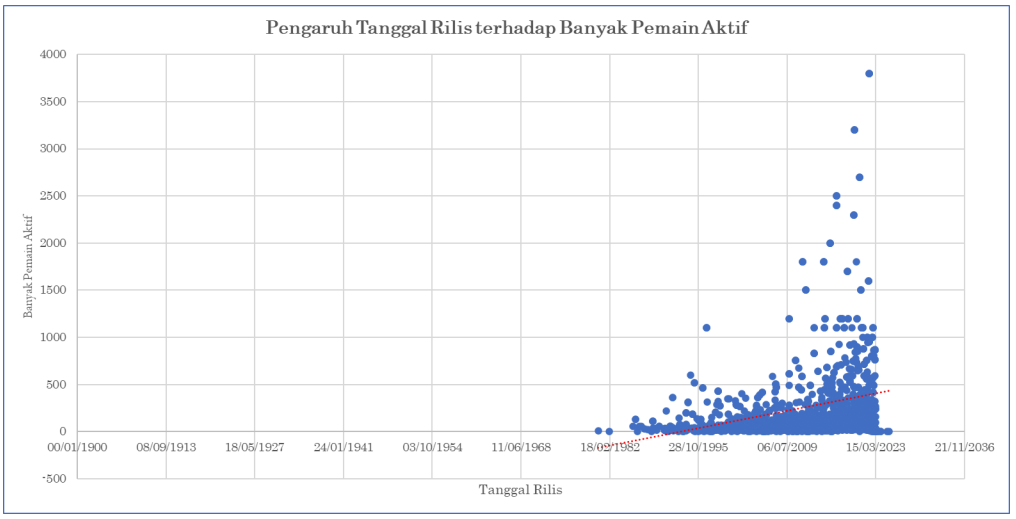
3.2.10 Visualisasi 10: Pengaruh Tanggal Rilis Terhadap Banyak Pemain

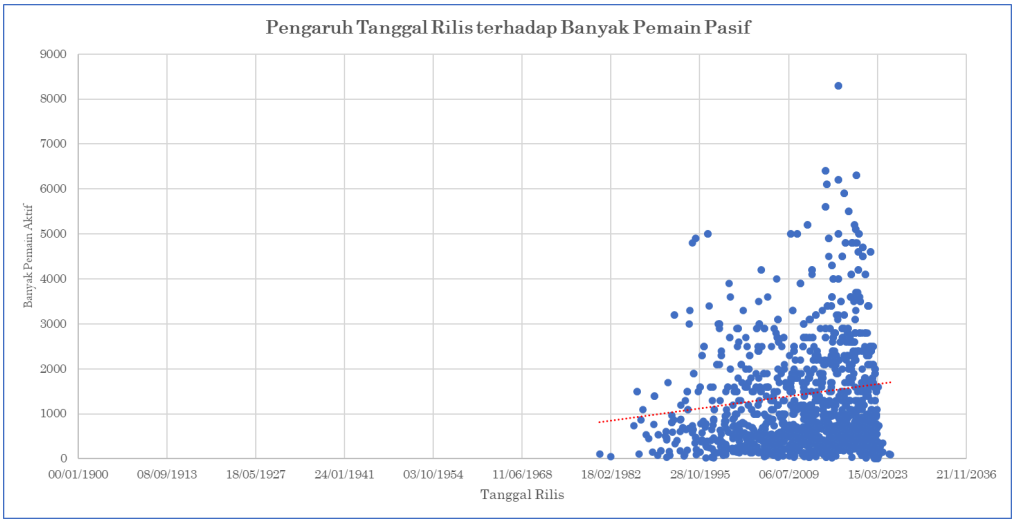
Proses Pembuatan:

1. Gunakan data yang berada pada SheetUtama untuk memproses data secara keseluruhan.
2. Buat 3 *scatter plot diagram*: 1) Tanggal Rilis terhadap Banyak Pemain (Plays), 2) Tanggal Rilis terhadap Banyak Pemain Aktif (Playing), 3) Tanggal Rilis terhadap Banyak Pemain Pasif (Backlogs).

Hasil Visualisasi







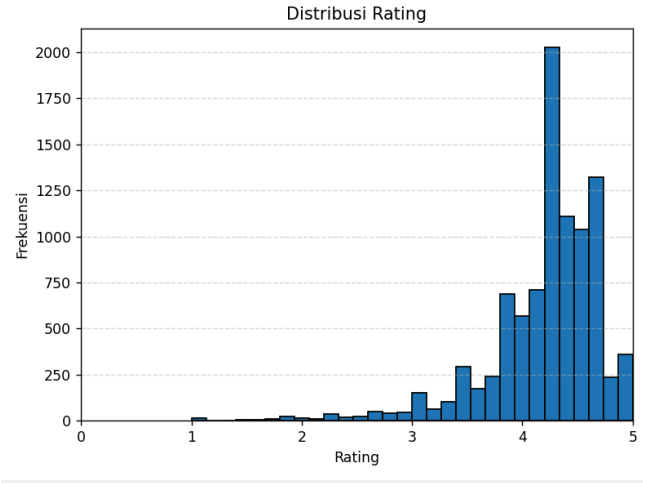
Hasil Insight

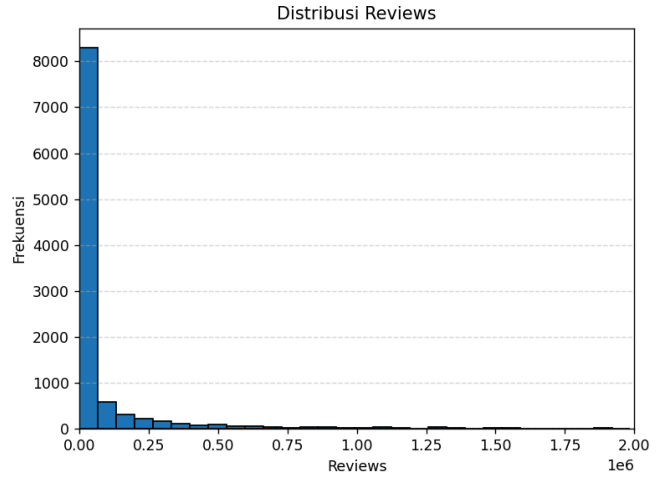
Berdasarkan ketiga diagram, diketahui bahwa tanggal rilis berkorelasi negatif dengan banyak pemain. Akan tetapi, hal ini berbeda dengan banyak pemain aktif dan pasif. Pada kasus tersebut, tanggal rilis berkorelasi positif.

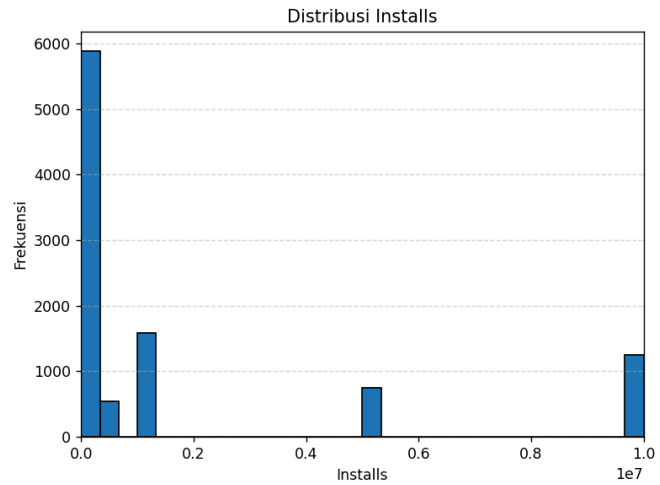
4. Statistik Deskriptif

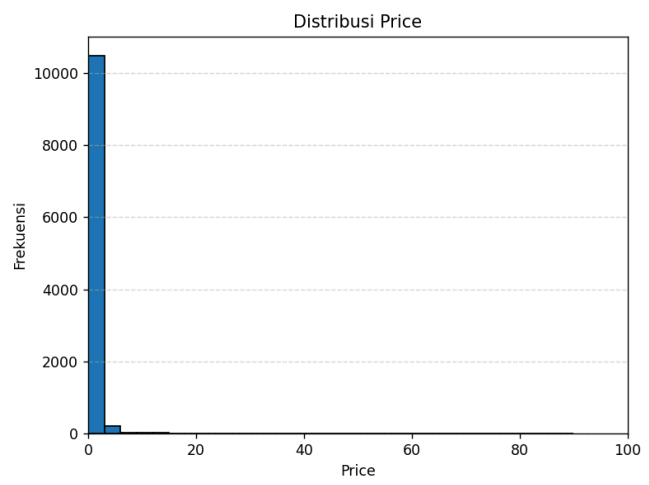
4.1 Data Aplikasi pada Google Play Store

Kategori	Mean	Std Dev	10%	25%	Median	75%	90%	Min	Max
Rating	4.19	0.54	3.6	4.0	4.3	4.5	4.7	1.0	1.900000e+01
Reviews	444152.90	2927760.60	3.0	38.0	2094.0	54775.5	464993.1	0.0	7.815831e+07
Installs	15464338.88	85029361.40	100.0	1000.0	100000.0	5000000.0	10000000.0	0.0	1.000000e+09
Price	1.03	15.95	0.0	0.0	0.0	0.0	0.0	0.0	4.000000e+02









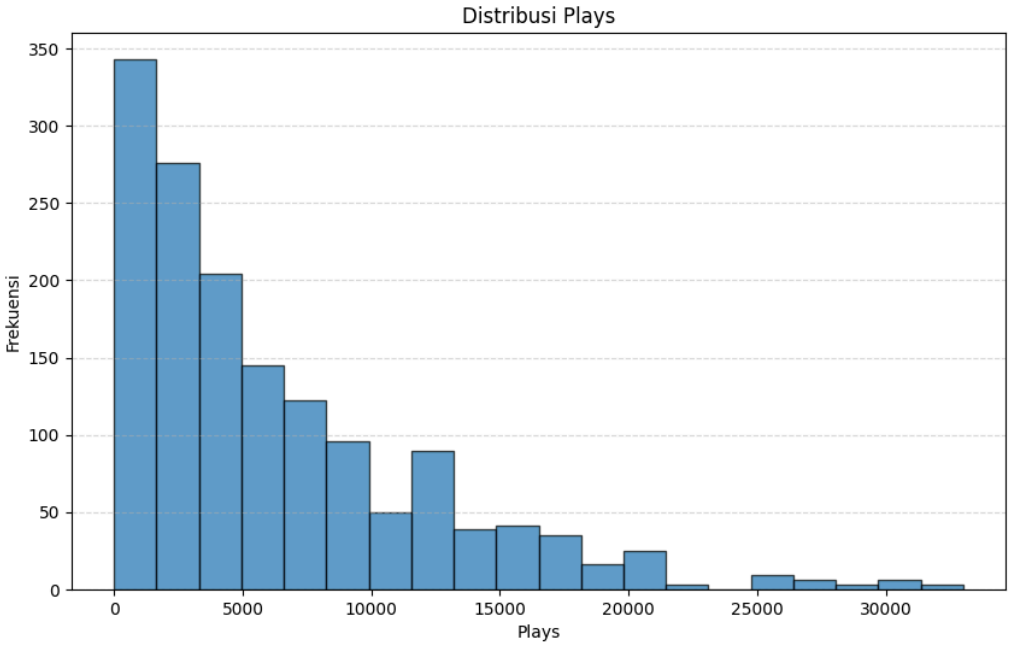
Cara Mendapatkan Data:

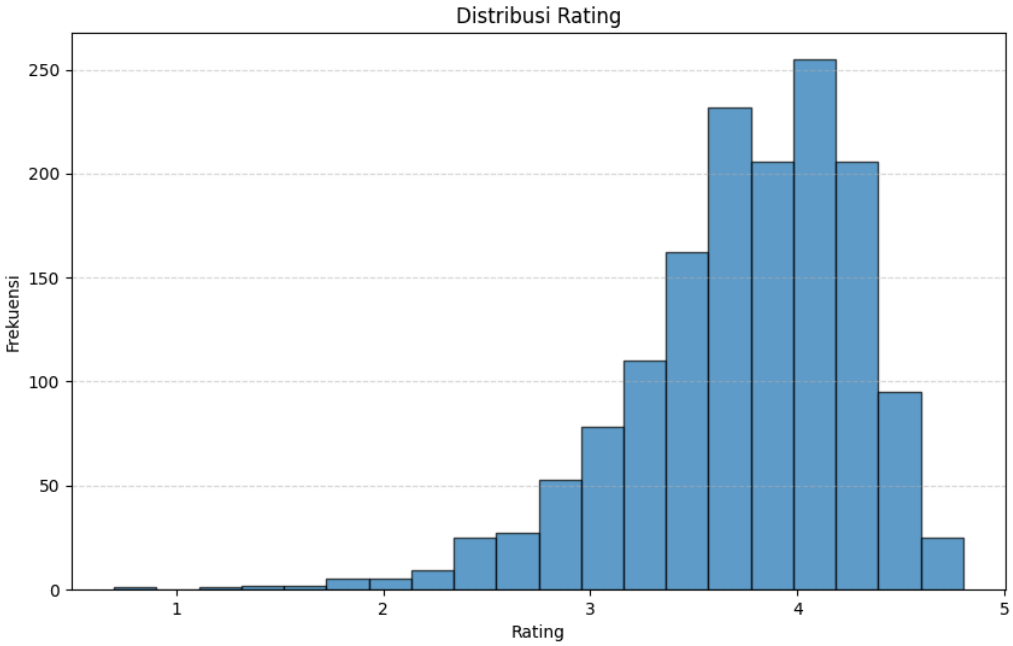
```
import os
```

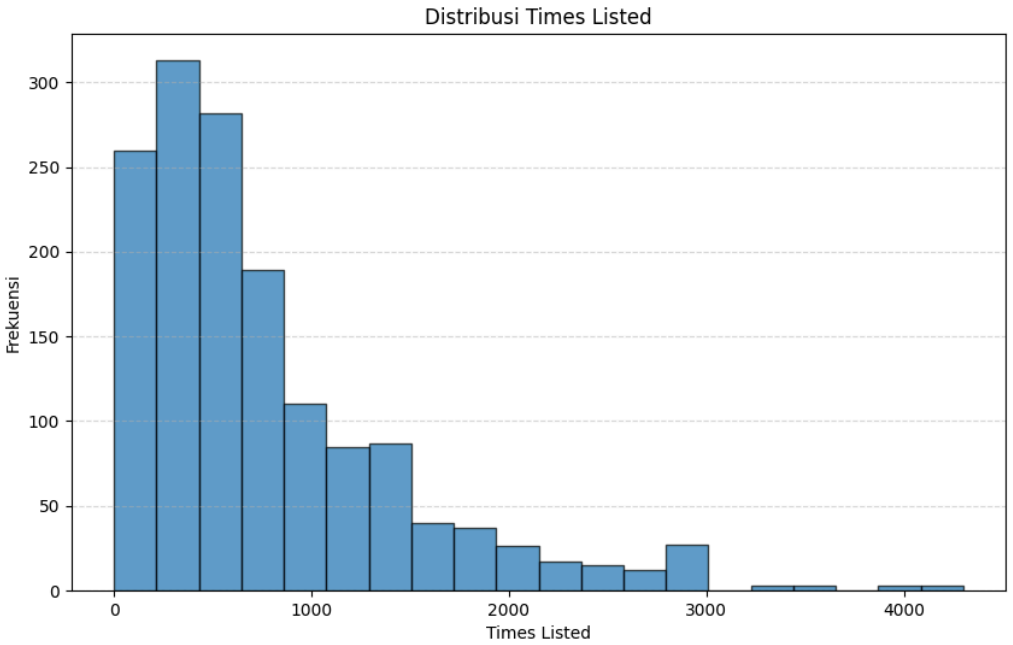
```
import pandas as pd
import matplotlib.pyplot as plt

# Setup directory
os.chdir(os.path.dirname(os.path.abspath(__file__)))

# Load data
df_google = pd.read_csv("data/googleplaystore.csv")


baris1 = "Rating"
baris2 = "Reviews"
baris3 = "Installs"
baris4 = "Price"
kolom_analisis = [baris1, baris2, baris3, baris4]


df_google[baris1] = pd.to_numeric(df_google[baris1], errors='coerce')
df_google[baris2] = pd.to_numeric(df_google[baris2], errors='coerce')

df_google[baris3] = df_google[baris3].astype(str).str.replace('[+,]', '', regex=True)
df_google[baris3] = pd.to_numeric(df_google[baris3], errors='coerce')

df_google[baris4] = df_google[baris4].astype(str).str.replace('$', '', regex=False)
df_google[baris4] = pd.to_numeric(df_google[baris4], errors='coerce')


all_stats = pd.DataFrame()

for kolom in kolom_analisis:
    data = df_google[kolom].dropna()

    statistik = {
        "Mean": data.mean(),
        "Standard Deviation": data.std(),
        "Percentile 10%": data.quantile(0.10),
        "Percentile 25%": data.quantile(0.25),
        "Median (50%)": data.median(),
        "Percentile 75%": data.quantile(0.75),
        "Percentile 90%": data.quantile(0.90),
        "Minimum": data.min(),
        "Maximum": data.max()
    }
    all_stats = pd.concat([all_stats, pd.DataFrame([statistik], index=[kolom])])

all_stats = all_stats.round(2)

# tabel
print("\n Statistik Deskriptif:")
print(all_stats)

# histogram
for kolom in kolom_analisis:
    data = df_google[kolom].dropna()


    if kolom == "Rating":
        data = data[(data >= 0) & (data <= 5)]
    elif kolom == "Price":
```

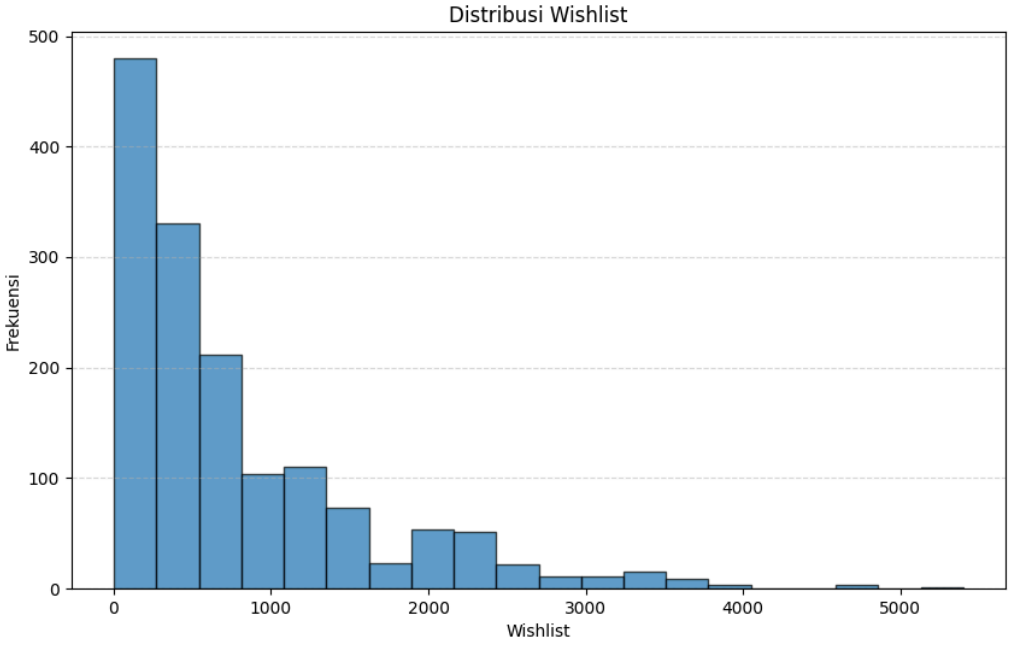
```
    data = data[(data >= 0) & (data <= 100)]
elif kolom == "Installs":
    data = data[(data >= 0) & (data <= 1e7)]
elif kolom == "Reviews":
    data = data[(data >= 0) & (data <= 2e6)]


plt.figure()
plt.hist(data, bins=30, edgecolor='black')
plt.title(f"Distribusi {kolom}")
plt.xlabel(kolom)
plt.ylabel("Frekuensi")

if kolom == "Rating":
    plt.xlim(0, 5)
elif kolom == "Price":
    plt.xlim(0, 100)
elif kolom == "Installs":
    plt.xlim(0, 1e7)
elif kolom == "Reviews":
    plt.xlim(0, 2e6)

plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

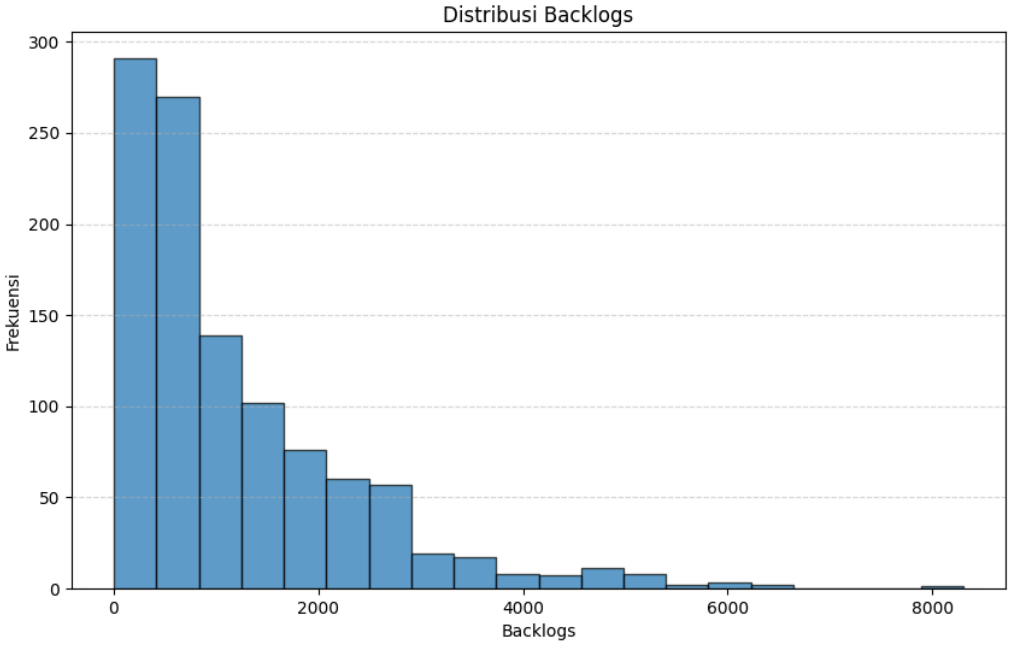
4.2 Data Video Game Terpopuler 1980 - 2023

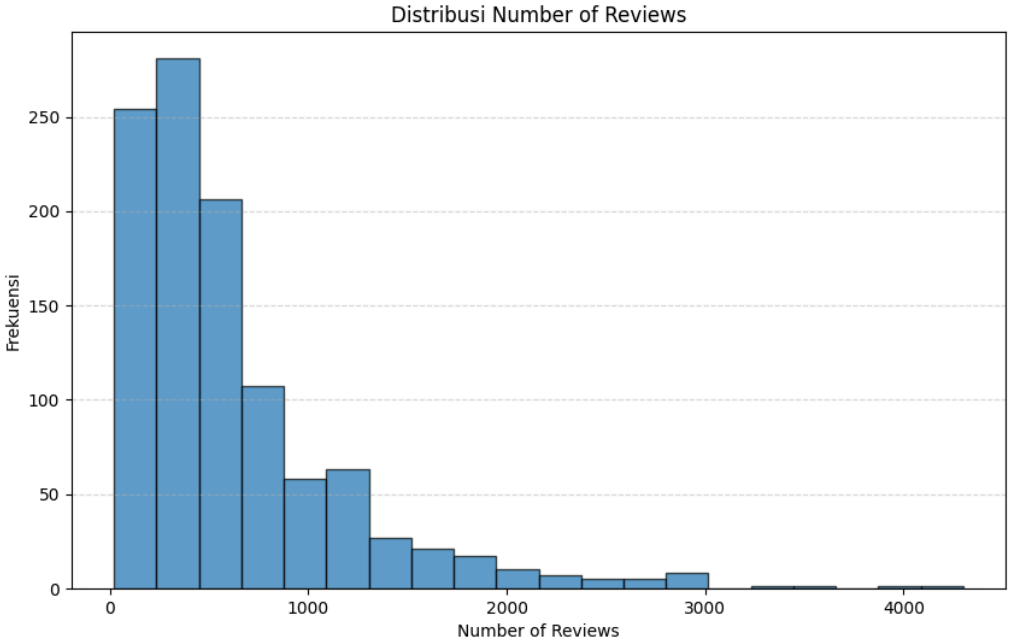


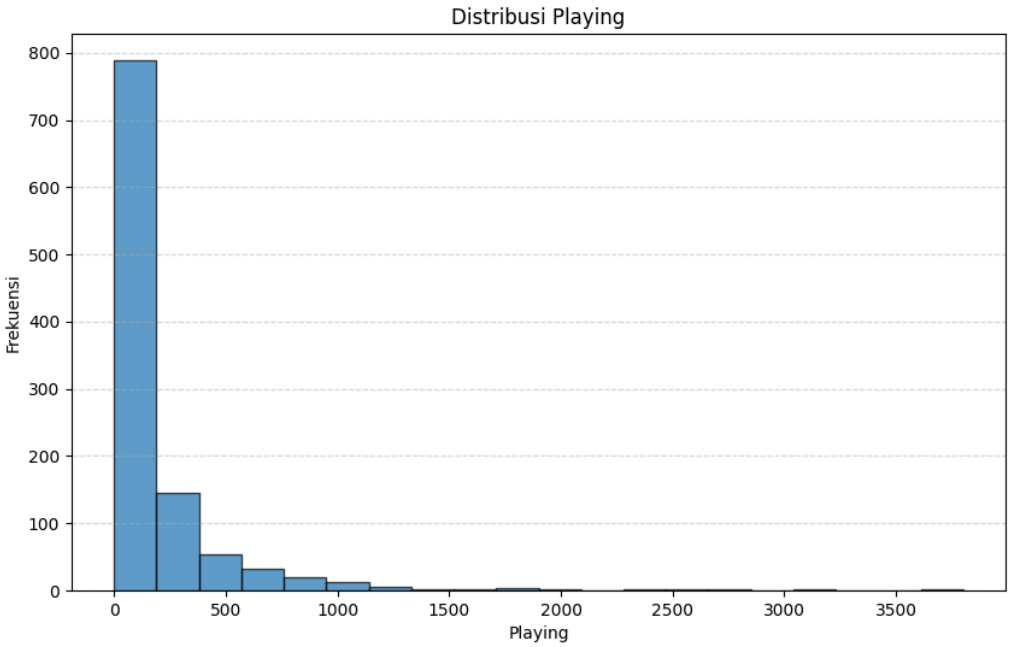



Distribusi Times Listed
Frekuensi
Times Listed
0
50
100
150
200
250
300
0
1000
2000
3000
4000


Distribusi Wishlist
Frekuensi
Wishlist
0
100
200
300
400
500
0
1000
2000
3000
4000
5000


Distribusi Backlogs
Frekuensi
Backlogs
0
50
100
150
200
250
300
0
2000
4000
6000
8000


Distribusi Number of Reviews
Frekuensi
Number of Reviews
0
50
100
150
200
250
0
1000
2000
3000
4000



Cara Mendapatkan Data Distribusi:

```
import pandas as pd
import matplotlib.pyplot as plt
import os
import sys

# Fungsi untuk konversi format angka seperti "3.9K" menjadi 3900
def konversi_k_format(val):
    if pd.isna(val):
        return None
    val = str(val).strip().upper()
    if val.endswith('K'):
        try:
            return float(val[:-1]) * 1000
        except ValueError:
            return None
    try:
        return float(val)
    except ValueError:
        return None

# Cek ketersediaan file CSV
file_path = "games.csv"
if not os.path.exists(file_path):
    print(f"File {file_path} tidak ditemukan!")
    sys.exit(1)

# Baca data dari CSV
df_games = pd.read_csv(file_path)

# Daftar kolom yang dianalisis
kolom_analisis = [“Rating”, “Times Listed”, “Number of Reviews”, “Plays”, “Playing”, “Backlogs”, “Wishlist”]
```

```
# Konversi dan analisis statistik sederhana
for kolom in kolom_analisis:
    if kolom in df_games.columns:
        if kolom == "Rating":
            df_games[kolom] = pd.to_numeric(df_games[kolom], errors='coerce')
            df_games = df_games[(df_games[kolom] >= 0) & (df_games[kolom] <= 5)]
        else:
            df_games[kolom] = df_games[kolom].apply(konversi_k_format)
            df_games = df_games[df_games[kolom] >= 0]

# Pembuatan histogram untuk setiap kolom
for kolom in kolom_analisis:
    if kolom in df_games.columns:
        data = df_games[kolom].dropna()
        if len(data) > 0:
            plt.figure(figsize=(10, 6))
            plt.hist(data, bins=20, edgecolor='black', alpha=0.7)
            plt.title(f"Distribusi {kolom}")
            plt.xlabel(kolom)
            plt.ylabel("Frekuensi")
            plt.grid(axis='y', linestyle='--', alpha=0.5)
            file_name = f"histogram_{kolom.replace(' ', '_')}.png"
            plt.savefig(file_name)
            plt.close()
```

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

Data didapatkan dengan mensorting tiap kategori data, lalu :

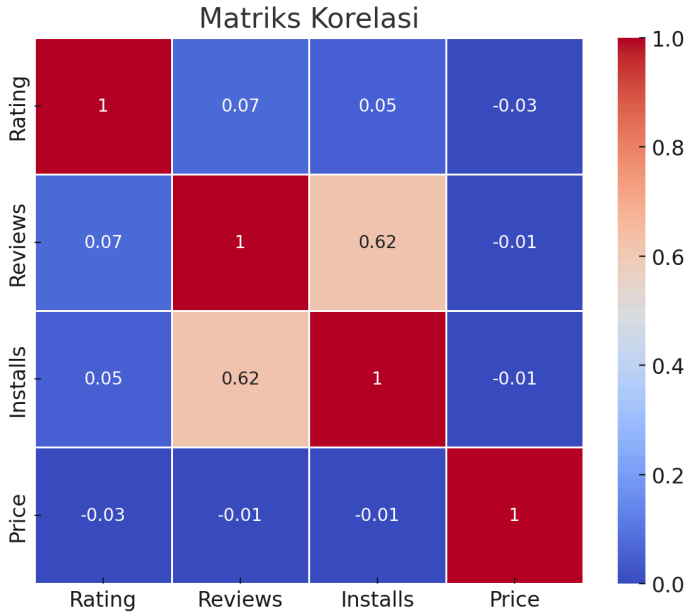
1. COUNT didapatkan dengan fungsi =COUNT(xx:xx) pada excel
2. AVERAGE didapatkan dengan fungsi =AVERAGE(xx:xx) pada excel
3. Standard Deviation didapatkan dengan fungsi =STDEV.P(xx:xx) pada excel
4. Percentile. 10% didapatkan dengan fungsi =PERCENTILE.INC(xx:xx, 0,1) pada excel
5. Percentile. 25% didapatkan dengan fungsi =PERCENTILE.INC(xx:xx, 0,25) pada excel
6. Percentile. 50% didapatkan dengan fungsi =PERCENTILE.INC(xx:xx, 0,5) pada excel
7. Percentile. 75% didapatkan dengan fungsi =PERCENTILE.INC(xx:xx, 0,75) pada excel
8. Percentile. 90% didapatkan dengan fungsi =PERCENTILE.INC(xx:xx, 0,9) pada excel
9. MAX didapatkan dengan fungsi =MAX(xx:xx) pada excel
10. MIN didapatkan dengan fungsi =MIN(xx:xx) pada excel

Note : (xx:xx) berupa range baris dan kolom, misal (E16:E1514).
Percentile INC berarti data ke 0,x percent ikut diproses.

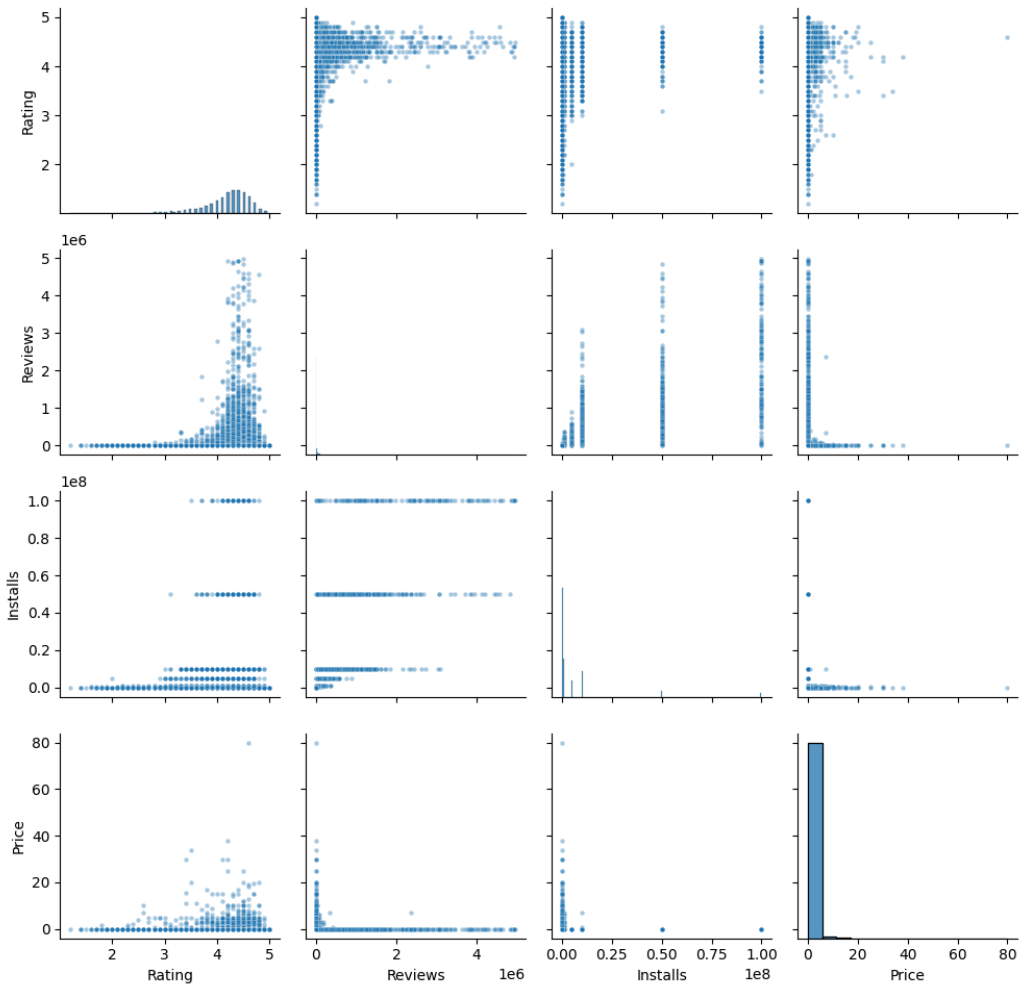
5. Korelasi

5.1 Data Aplikasi pada Google Play Store

A. Matriks Korelasi



B. Visualisasi Korelasi



1. Rating dan Price
   - Pola: Tidak terlihat pola linear yang jelas.
   - Penjelasan: Baik aplikasi gratis maupun berbayar bisa memiliki rating tinggi atau rendah.
   - Insight: Harga aplikasi tidak berkorelasi signifikan dengan rating pengguna. Artinya, mahal atau murah bukan penentu langsung kualitas yang dirasakan user.
2. Rating dan Installs
   - Pola: Sebagian besar aplikasi dengan rating tinggi justru punya instalasi yang tinggi juga, tapi dengan sebaran luas.
   - Penjelasan: Walaupun ada tren bahwa aplikasi populer sering memiliki rating baik, hubungannya cukup lemah dan tidak konsisten.
   - Insight: Banyak aplikasi populer dengan rating tinggi, tapi tidak semua aplikasi rating tinggi otomatis populer.
3. Rating dan Reviews
   - Pola: Mirip seperti dengan installs — ada kerumunan data di rating tinggi + jumlah review tinggi, tapi tetap menyebar.
   - Penjelasan: Aplikasi yang sering dipakai cenderung lebih sering direview. Tapi ada juga aplikasi dengan rating bagus tapi review sedikit.
   - Insight: Jumlah review tidak secara langsung menentukan rating. Bisa karena niche audience, atau campaign internal.
4. Price dan Reviews
   - Pola: Mayoritas data terkumpul di harga $0 dengan berbagai jumlah review.
   - Penjelasan: Aplikasi gratis jelas lebih banyak digunakan dan direview. Aplikasi berbayar cenderung punya review sedikit.
   - Insight: Aplikasi gratis jauh lebih banyak digunakan dan direview. Price tinggi cenderung menurunkan eksposur.
5. Price dan Installs
   - Pola: Sangat terpusat di harga $0 dan jumlah instalasi tinggi. Untuk aplikasi berbayar, instalasi sangat sedikit.
   - Penjelasan: Ini wajar — gratis = lebih banyak diunduh.
   - Insight: Semakin mahal aplikasi, semakin kecil kecenderungan untuk diunduh. Korelasi sangat lemah bahkan mendekati negatif.
6. Reviews dan Installs
   - Pola: Terlihat korelasi positif yang cukup kuat.
   - Penjelasan: Aplikasi dengan banyak instalasi cenderung menghasilkan banyak review.
   - Insight: Ini satu-satunya hubungan dengan korelasi yang cukup signifikan. Jumlah install memicu lebih banyak review.

Kode Python untuk mendapatkan Matriks Korelasi:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


df = pd.read_csv("data/googleplaystore_cleaned.csv")

df["Rating"] = pd.to_numeric(df["Rating"], errors="coerce")
df["Reviews"] = pd.to_numeric(df["Reviews"], errors="coerce")
df["Installs"] = pd.to_numeric(df["Installs"], errors="coerce")
df["Price"] = pd.to_numeric(df["Price"], errors="coerce")


kolom_korelasi = ["Rating", "Reviews", "Installs", "Price"]
df_subset = df[kolom_korelasi].dropna()

corr_matrix = df_subset.corr().round(2)

plt.figure(figsize=(6, 5))
sns.heatmap(
```

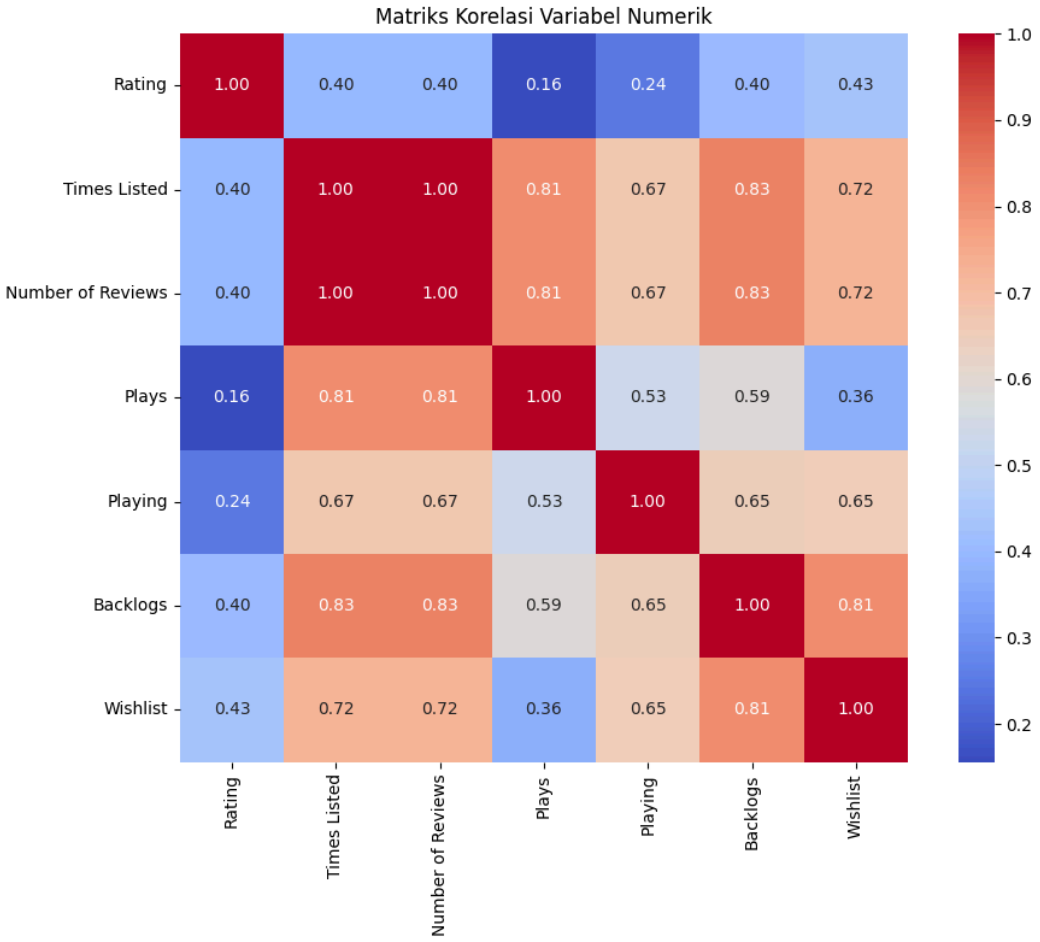
```
    corr_matrix,
    annot=True,
    cmap="coolwarm",
    vmin=0,
    vmax=1,
    square=True,
    linewidths=0.5
)
plt.title("Matriks Korelasi")
plt.tight_layout()
plt.savefig("matriks_korelasi_googleplaystore.png")
plt.show()
```

Kode Python untuk mendapatkan Visualisasi Korelasi:

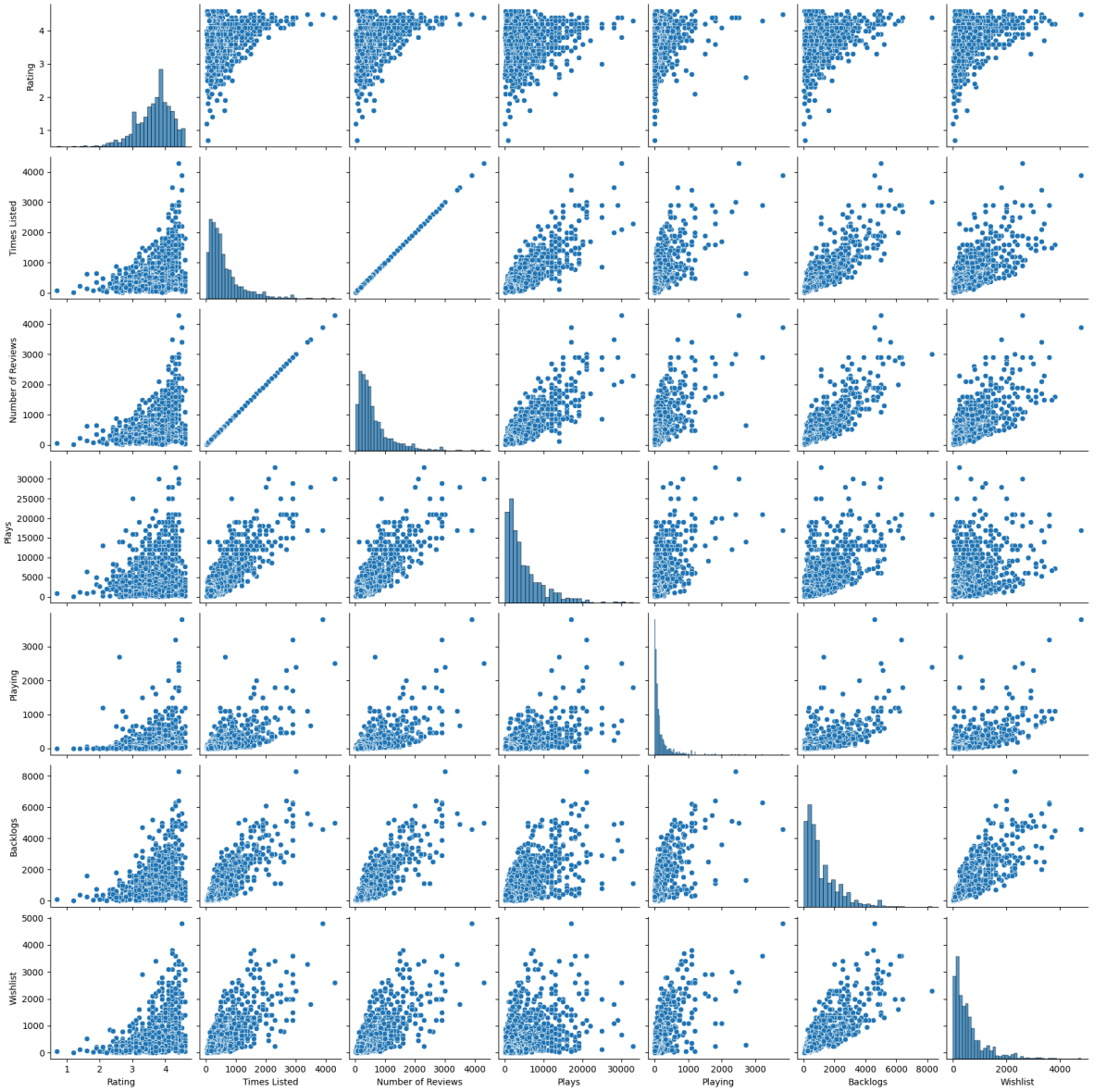
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv("data/googleplaystore_cleaned.csv")

df["Rating"] = pd.to_numeric(df["Rating"], errors="coerce")
df["Reviews"] = pd.to_numeric(df["Reviews"], errors="coerce")
df["Installs"] = pd.to_numeric(df["Installs"], errors="coerce")
df["Price"] = pd.to_numeric(df["Price"], errors="coerce")

kolom_numerik = ["Rating", "Reviews", "Installs", "Price"]
df_korelasi = df[kolom_numerik].dropna()

df_plot = df_korelasi[
    (df_korelasi["Rating"] <= 5) &
    (df_korelasi["Reviews"] <= 5e6) &
    (df_korelasi["Installs"] <= 1e8) &
    (df_korelasi["Price"] <= 100)
]


g = sns.pairplot(df_plot, corner=False, plot_kws={'alpha': 0.4, 's': 10})
g.fig.suptitle("Scatterplot antar Atribut Kuantitatif (Google Play Store)", y=1.02)

for i, var in enumerate(df_plot.columns):
    for ax in g.axes[i]:
        if ax is not None:
            ax.set_ylabel(df_plot.columns[i])
    for j, ax in enumerate(g.axes[-1]):
        if ax is not None:
            ax.set_xlabel(df_plot.columns[j])

plt.tight_layout()
plt.savefig("scatterplot_matrix_google_play_store.png")
plt.show()
```

5.2 Data Video Game Terpopuler 1980 - 2023

A. Matriks Korelasi



B. Visualisasi Korelasi



Analisis Korelasi antar Atribut

1. Rating dan Times Listed (r = 0.40)
   Terdapat korelasi lemah hingga sedang antara *Rating* dan *Times Listed*. Aplikasi yang sering muncul atau sering di-*list* cenderung memiliki rating yang cukup tinggi, meskipun tidak selalu konsisten. Artinya, ada kecenderungan bahwa aplikasi yang lebih dikenal juga berpeluang mendapatkan penilaian yang lebih baik dari pengguna.
   Insight: Makin dikenal sebuah aplikasi, makin besar peluangnya untuk mendapatkan rating yang bagus, meski bukan faktor penentu utama.
2. Rating dan Plays (r = 0.16)
   Hubungan antara *Rating* dan *Plays* sangat lemah. Berdasarkan visualisasi scatter, terlihat bahwa rating tidak banyak berubah meskipun jumlah permainan (*play*) meningkat. Data tersebar cukup luas tanpa pola linear yang jelas.
   Insight: Rating tidak tergantung pada seberapa sering aplikasi dimainkan. Aplikasi populer belum tentu punya rating tinggi, dan sebaliknya.
3. Rating dan Wishlist (r = 0.43)
   Korelasi antara *Rating* dan *Wishlist* termasuk lemah ke sedang. Terlihat bahwa semakin banyak pengguna yang menambahkan aplikasi ke wishlist, umumnya rating juga cukup tinggi.

Insight: Rating bisa menjadi indikator awal minat pengguna, tapi tidak selalu mencerminkan keputusan mereka untuk menyimpan atau tidak menyimpan aplikasi.

4. Times Listed dan Plays (r = 0.81)
   Ini adalah pasangan dengan korelasi sangat kuat. Aplikasi yang lebih sering muncul (baik di toko aplikasi atau rekomendasi) juga lebih sering dimainkan. Penyebaran titik-titik pada scatter menunjukkan tren linear yang cukup konsisten.
   Insight: Keterlihatan (visibilitas) sangat berpengaruh terhadap jumlah pemakaian. Semakin terlihat sebuah aplikasi, semakin besar kemungkinan orang akan memainkannya.
5. Times Listed dan Wishlist (r = 0.72)
   Korelasi antara *Times Listed* dan *Wishlist* juga cukup kuat. Aplikasi yang sering tampil di depan pengguna memiliki peluang lebih besar untuk dimasukkan ke dalam wishlist.
   Insight: Semakin sering sebuah aplikasi dilihat, semakin besar kemungkinan orang menyimpannya untuk dipertimbangkan di kemudian hari.
6. Plays dan Wishlist (r = 0.36)
   Terdapat korelasi cukup lemah antara *Plays* dan *Wishlist*. Meskipun ada tren bahwa aplikasi yang sering dimainkan juga cenderung disukai, banyak penyimpangan dari pola ini. Beberapa aplikasi yang sering dimainkan justru tidak terlalu banyak masuk wishlist, dan sebaliknya.
   Insight: Wishlist lebih mencerminkan minat awal daripada hasil penggunaan. Artinya, seseorang bisa tertarik pada aplikasi dan menyimpannya, tetapi belum tentu akan memainkannya, atau bisa juga memainkan aplikasi tanpa pernah menyimpannya dulu.
7. Playing dan Number of Reviews (r = 0.67)
   Terdapat Korelasi cukup kuat antara *Playing* dan *Number of Reviews*. Game yang sedang dimainkan cenderung memiliki jumlah review yang tinggi.
   Insight: Pengguna yang sedang aktif bermain lebih terlibat secara emosional, sehingga lebih terdorong untuk memberikan ulasan. Ini memperlihatkan bahwa aktivitas bermain dapat menjadi pendorong keterlibatan dalam bentuk feedback atau review, yang pada akhirnya memperkaya persepsi publik terhadap game tersebut.
8. Playing dan Backlogs (r = 0.65)
   Terdapat korelasi yang sedang-kuat antara *Playing* dan *Backlogs*. Game yang sedang dimainkan seringkali berasal dari backlog.
   Insight: Ini menunjukkan adanya pola konsumsi berurutan, di mana pemain menyimpan game terlebih dahulu, lalu memainkannya setelah waktu memungkinkan. Hal ini mencerminkan kebiasaan perencanaan konsumsi konten oleh pengguna.
9. Playing dan Wishlist (r = 0.65)
   Terdapat korelasi yang sedang-kuat antara *Playing* dan *Wishlist.* Game yang banyak masuk wishlist memiliki kemungkinan besar untuk dimainkan.
   Insight: Wishlist berperan sebagai indikator minat awal yang cukup akurat terhadap potensi permainan di masa depan. Game dengan angka wishlist tinggi umumnya akan memperoleh jumlah pemain aktif yang signifikan, terutama setelah rilis atau saat ada diskon.
10. Backlogs dan Number of Reviews (r = 0.83)
    Terdapat korelasi yang sangat kuat antara *Backlogs* dan *Number of Reviews*. Game dengan banyak review cenderung banyak masuk backlog.
    Insight: Banyaknya review menjadi indikator popularitas atau kualitas yang diyakini oleh calon pemain. Review menjadi referensi penting dalam proses pengambilan keputusan untuk menyimpan game dalam backlog sebagai game yang "layak dimainkan nanti".
11. Backlogs dan Wishlist (r = 0.81)
    Terdapat korelasi sangat kuat antara *Backlogs* dan *Wishlist*. Keduanya mencerminkan ketertarikan awal terhadap game.
    Insight: Wishlist merupakan fase minat awal, sedangkan backlog adalah bentuk lanjutan dari niat tersebut. Pengguna cenderung menambahkan game yang mereka simpan di wishlist ke backlog sebagai bagian dari daftar yang akan dimainkan ketika waktunya tiba.
12. Number of Reviews dan Wishlist (r = 0.72)

Terdapat korelasi kuat antara *Number of Reviews* dan *Wishlist*. Game yang banyak direview juga banyak masuk wishlist.
Insight: Review memiliki fungsi sebagai validasi sosial yang kuat. Game yang banyak dibicarakan cenderung lebih menarik perhatian pengguna baru, sehingga lebih sering ditambahkan ke wishlist meskipun belum dimainkan.

13. Plays dan Number of Reviews (r = 0.81)
    Terdapat korelasi sangat kuat antara *Number of Reviews* dan *Plays*. Game yang banyak dimainkan juga banyak direview.
    Insight: Jumlah pemain yang tinggi menciptakan banyak pengalaman pengguna yang mendorong aktivitas review. Hal ini memperkuat eksistensi game di ruang publik dan meningkatkan visibilitasnya bagi calon pemain baru.
14. Plays dan Backlogs (r = 0.59)
    Terdapat korelasi sedang antara *Plays* dan *Backlogs*. Game dalam backlog seringkali akhirnya dimainkan.
    Insight: Backlog berfungsi bukan hanya sebagai tempat penyimpanan pasif, tetapi juga sebagai sumber utama game yang akan dimainkan. Ini menunjukkan transisi nyata dari niat untuk bermain ke aksi yang dilakukan pengguna.

Kode Python untuk mendapatkan Matriks Korelasi dan Visualisasi Korelasi

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
import sys

# Fungsi untuk konversi format seperti "3.9K" menjadi 3900
def konversi_k_format(val):
  if pd.isna(val):
    return None
  val = str(val).strip().upper()
  if val.endswith('K'):
    try:
      return float(val[:-1]) * 1000
    except ValueError:
      return None
  try:
    return float(val)
  except ValueError:
    return None

file_path = "games_cleaned.csv"
df_games = pd.read_csv(file_path)

# Daftar kolom yang dianalisis
kolom_analisis = ["Rating", "Times Listed", "Plays", "Wishlist", “Number of Reviews”, “Playing”,
“Backlogs”]

# Konversi dan pembersihan data
for kolom in kolom_analisis:
  if kolom in df_games.columns:
    if kolom == "Rating":
      df_games[kolom] = pd.to_numeric(df_games[kolom], errors='coerce')
      df_games = df_games[(df_games[kolom] >= 0) & (df_games[kolom] <= 5)]
    else:
      df_games[kolom] = df_games[kolom].apply(konversi_k_format)
      df_games = df_games[df_games[kolom] >= 0]
```

```
# Matriks Korelasi
df_corr = df_games[kolom_analisis].dropna().corr()
plt.figure(figsize=(8, 6))
sns.heatmap(df_corr, annot=True, cmap="coolwarm", fmt=".2f", square=True)
plt.title("Matriks Korelasi")
plt.tight_layout()
plt.savefig("korelasi_matrix.png")
plt.close()

# Scatter Plot Antar Semua Kolom
sns.pairplot(df_games[kolom_analisis].dropna())
plt.savefig("scatter_pairplot.png")
plt.close()
```

6. Data Cleansing

6.1 Data Aplikasi pada Google Play Store

Sebelum masuk ke data cleansing per kolom, kita dapat memuat dan menampilkan informasi data awal sebelum data cleansing, serta melakukan langkah-langkah pre-cleansing umum untuk menangani baris anomali dan data-data duplikat.

```
import pandas as pd
import numpy as np

# Muat dataset
df = pd.read_csv('googleplaystore.csv')

print("--- INFORMASI DATA SEBELUM CLEANSING ---")
print("Informasi Umum Dataset:")
df.info()
print("\nJumlah Missing Values per Kolom:")
print(df.isnull().sum())
print("\nPersentase Missing Values per Kolom:")
print((df.isnull().sum() / len(df)) * 100)
print(f"\nJumlah Baris Duplikat Awal: {df.duplicated().sum()}")
print("-" * 50)

# --- Langkah Pre-cleansing Umum (Penanganan Baris Anomali dan Duplikat Umum) ---
# Baris ini dikenal sebagai baris yang salah parsing, 'Category' adalah '1.9'
initial_rows = len(df)
df = df[df['Category'] != '1.9']
print(f"\n[UMUM] Baris dihapus karena anomali parsing (Category = '1.9'): {initial_rows - len(df)}")

# Hapus duplikat berdasarkan kolom 'App' untuk memastikan setiap aplikasi unik
rows_before_deduplication = len(df)
df.drop_duplicates(subset=['App'], inplace=True)
print(f"[UMUM] Baris dihapus karena duplikasi (berdasarkan App): {rows_before_deduplication - len(df)}")
```

Hasil :

```
--- INFORMASI DATA SEBELUM CLEANSING ---
Informasi Umum Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      10841 non-null  int64
 1   App             10841 non-null  object
 2   Category        10841 non-null  object
 3   Rating          9367 non-null   float64
 4   Reviews         10841 non-null  object
 5   Size            10841 non-null  object
 6   Installs        10841 non-null  object
 7   Type            10840 non-null  object
 8   Price           10841 non-null  object
 9   Content Rating  10840 non-null  object
 10  Genres          10841 non-null  object
 11  Last Updated    10841 non-null  object
 12  Current Ver     10833 non-null  object
 13  Android Ver     10838 non-null  object
 14  Unnamed: 14     0 non-null      float64
 15  Installs_Clean  10841 non-null  object
dtypes: float64(2), int64(1), object(13)
memory usage: 1.3+ MB
```

```
Jumlah Missing Values per Kolom:
Unnamed: 0           0
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content Rating       1
Genres               0
Last Updated         0
Current Ver          8
Android Ver          3
Unnamed: 14      10841
Installs_Clean       0
dtype: int64

Persentase Missing Values per Kolom:
Unnamed: 0         0.000000
App                0.000000
Category           0.000000
Rating            13.596532
Reviews            0.000000
Size               0.000000
Installs           0.000000
Type               0.009224
Price              0.000000
Content Rating     0.009224
Genres             0.000000
Last Updated       0.000000
Current Ver        0.073794
Android Ver        0.027673
Unnamed: 14      100.000000
Installs_Clean     0.000000
dtype: float64

Jumlah Baris Duplikat Awal: 0
--------------------------------------------------

[UMUM] Baris dihapus karena anomali parsing (Category = '1.9'): 1
[UMUM] Baris dihapus karena duplikasi (berdasarkan App): 1181
```

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.
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Rating
# Tindakan: Hapus baris dengan missing values, pastikan dalam rentang [1.0, 5.0]
print("\n--- CLEANSING 'Rating' ---")
rows_before_rating_clean = len(df)
df.dropna(subset=['Rating'], inplace=True) # Menghapus baris yang memiliki missing values pada 'Rating'
```

```
df['Rating'] = pd.to_numeric(df['Rating'], errors='coerce') # Konversi ke numerik, ubah error menjadi NaN
df = df[(df['Rating'] >= 1.0) & (df['Rating'] <= 5.0)] # Filter untuk rentang yang valid
print(f"  Baris dihapus (missing/out-of-range): {rows_before_rating_clean - len(df)}")
```

- **Hasil**
  1463 baris dihapus (termasuk yang missing dan out-of-range). Sekarang, semua nilai Rating adalah float dan berada dalam rentang [1.0, 5.0].

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.
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Reviews
# Tindakan: Konversi ke numerik, hapus jika gagal. Terapkan batas bawah.
df['Reviews'] = df['Reviews'].astype(str).str.replace('+', '', regex=False) # Hapus '+'
df['Reviews'] = df['Reviews'].str.replace(',', '', regex=False) # Hapus ','
df['Reviews'] = pd.to_numeric(df['Reviews'], errors='coerce') # Konversi ke numerik, ubah error menjadi NaN
rows_before_reviews_clean = len(df)
df.dropna(subset=['Reviews'], inplace=True) # Hapus jika gagal konversi (NaN dari coerce)
print(f"\n--- CLEANSING 'Reviews' ---")
print(f"  Baris dihapus (gagal konversi): {rows_before_reviews_clean - len(df)}")

# Penambahan Batas Bawah untuk Reviews
min_reviews_threshold = 10
rows_before_reviews_threshold = len(df)
df = df[df['Reviews'] >= min_reviews_threshold]
print(f"  Baris dihapus (Reviews < {min_reviews_threshold}): {rows_before_reviews_threshold - len(df)}")
df['Reviews'] = df['Reviews'].astype(int) # Ubah ke integer secara permanen
```

- **Hasil**
  0 baris dihapus karena gagal konversi tipe data dasar. Sebanyak 588 baris dihapus karena memiliki kurang dari 10 ulasan. Tipe data kolom Reviews sekarang adalah int64, tanpa missing values, dan semua nilai adalah 10 atau lebih tinggi.

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.om ini adalah missing values, ditambah beberapa nilai anomali yang perlu dihapus.
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Installs
# Tindakan: Hapus karakter '+', ',', konversi ke numerik, hapus jika gagal. Terapkan batas bawah.
df['Installs'] = df['Installs'].astype(str).str.replace('+', '', regex=False)
df['Installs'] = df['Installs'].str.replace(',', '', regex=False)
df['Installs'] = pd.to_numeric(df['Installs'], errors='coerce')
rows_before_installs_clean = len(df)
df.dropna(subset=['Installs'], inplace=True) # Hapus jika gagal konversi
print(f"\n--- CLEANSING 'Installs' ---")
print(f"  Baris dihapus (gagal konversi): {rows_before_installs_clean - len(df)}")

# Penambahan Batas Bawah untuk Installs
min_installs_threshold = 1000
rows_before_installs_threshold = len(df)
df = df[df['Installs'] >= min_installs_threshold]
print(f"  Baris dihapus (Installs < {min_installs_threshold}): {rows_before_installs_threshold - len(df)}")
df['Installs'] = df['Installs'].astype(int) # Ubah ke integer secara permanen
```

- **Hasil**
  0 baris dihapus karena gagal konversi tipe data dasar. Sebanyak 203 baris dihapus karena memiliki kurang dari 1000 instalasi. Tipe data kolom Installs sekarang adalah int64, tanpa *missing values*, dan semua nilai adalah 1000 atau lebih tinggi.

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".
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Size
# Tindakan: Konversi 'M'/'k' ke Megabytes, 'Varies with device' ke NaN, lalu imputasi NaN dengan median.
def clean_size(size):
    if isinstance(size, str):
        size = size.replace(',', '') # Hapus koma jika ada
        if 'M' in size:
            return float(size.replace('M', ''))
        elif 'k' in size:
            return float(size.replace('k', '')) / 1024 # Konversi Kilobytes ke Megabytes
        elif 'Varies with device' in size:
            return np.nan # Menjadi NaN
    return np.nan # Untuk nilai yang tidak dikenal atau bukan string

df['Size'] = df['Size'].apply(clean_size)
df['Size'].fillna(df['Size'].median(), inplace=True) # Imputasi NaN pada Size dengan median
print(f"\n--- CLEANSING 'Size' ---")
print("  NaN pada kolom 'Size' diisi dengan median.")
```

- **Hasil**
  Kolom Size sekarang bertipe float64 dan tidak memiliki missing values. Size yang awalnya berupa “Varies with device” diubah menjadi median. Semua nilai ukuran kini dalam satuan Megabytes.

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.

- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Price
# Tindakan: Hapus karakter '$', konversi ke numerik, hapus jika gagal.
df['Price'] = df['Price'].astype(str).str.replace('$', '', regex=False)
df['Price'] = pd.to_numeric(df['Price'], errors='coerce')
rows_before_price_clean = len(df)
df.dropna(subset=['Price'], inplace=True) # Hapus jika gagal konversi
print(f"\n--- CLEANSING 'Price' ---")
print(f"  Baris dihapus (gagal konversi): {rows_before_price_clean - len(df)}")
```

- **Hasil**
  0 baris dihapus karena gagal konversi. Tipe data kolom Price sekarang adalah float64, tanpa *missing values*.

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%.

- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Type, Content Rating, Current Ver, Android Ver
# Tindakan: Imputasi missing values dengan modus.
columns_to_impute_mode = ['Type', 'Content Rating', 'Current Ver', 'Android Ver']
for col in columns_to_impute_mode:
    if df[col].isnull().any():
        mode_val = df[col].mode()[0]
        df[col].fillna(mode_val, inplace=True)
        print(f"\n--- CLEANSING '{col}' ---")
        print(f"  Missing values pada '{col}' diisi dengan modus: {mode_val}")
```

- **Hasil**
  Semua kolom ini sekarang tanpa *missing values*.

6.2 Data Video Game Terpopuler 1980 - 2023

Sebelum masuk ke data cleansing per kolom, kita dapat memuat dan menampilkan informasi data awal sebelum data cleansing, serta melakukan langkah-langkah pre-cleansing umum untuk menangani baris anomali dan data-data duplikat.

```
import pandas as pd
import numpy as np

# Muat dataset
df_games = pd.read_csv('games.csv')

print("--- INFORMASI DATA GAMES.CSV SEBELUM CLEANSING ---")
print("Informasi Umum Dataset Games.csv:")
df_games.info()
print("\nJumlah Missing Values per Kolom Games.csv:")
print(df_games.isnull().sum())
print("\nPersentase Missing Values per Kolom Games.csv:")
print((df_games.isnull().sum() / len(df_games)) * 100)
print(f"\nJumlah Baris Duplikat Awal Games.csv: {df_games.duplicated().sum()}")
print("\nStatistik Deskriptif Kolom Numerik Awal Games.csv:")
print(df_games.describe())
print("\nContoh 5 Baris Pertama Games.csv:")
print(df_games.head())
print("-" * 50)

# Menghapus duplikat berdasarkan kolom 'Title'
rows_before_deduplication = len(df_games)
df_games.drop_duplicates(subset=['Title'], inplace=True)
print(f"\n--- LANGKAH UMUM: Menghapus Duplikat ---")
print(f"  Baris dihapus karena duplikasi (berdasarkan Title): {rows_before_deduplication - len(df_games)}")
```

Hasil :

```
--- INFORMASI DATA GAMES.CSV SEBELUM CLEANSING ---
Informasi Umum Dataset Games.csv:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1512 entries, 0 to 1511
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0       1512 non-null   int64
 1   Title            1512 non-null   object
 2   Release Date     1512 non-null   object
 3   Team             1511 non-null   object
 4   Rating           1499 non-null   float64
 5   Times Listed     1512 non-null   object
 6   Number of Reviews  1512 non-null   object
 7   Genres           1512 non-null   object
 8   Summary          1511 non-null   object
 9   Reviews          1512 non-null   object
 10  Plays            1512 non-null   object
 11  Playing          1512 non-null   object
 12  Backlogs         1512 non-null   object
 13  Wishlist         1512 non-null   object
dtypes: float64(1), int64(1), object(12)
memory usage: 165.5+ KB

Jumlah Missing Values per Kolom Games.csv:
Unnamed: 0          0
Title               0
Release Date        0
```

```
Team                1
Rating             13
Times Listed        0
Number of Reviews   0
Genres              0
Summary             1
Reviews             0
Plays               0
Playing             0
Backlogs            0
Wishlist            0
dtype: int64

Persentase Missing Values per Kolom Games.csv:
Unnamed: 0          0.000000
Title               0.000000
Release Date        0.000000
Team                0.066138
Rating              0.859788
Times Listed        0.000000
Number of Reviews   0.000000
Genres              0.000000
Summary             0.066138
Reviews             0.000000
Plays               0.000000
Playing             0.000000
Backlogs            0.000000
Wishlist            0.000000
dtype: float64

Jumlah Baris Duplikat Awal Games.csv: 0

Statistik Deskriptif Kolom Numerik Awal Games.csv:
       Unnamed: 0       Rating
count  1512.000000  1499.000000
mean    755.500000     3.719346
std     436.621117     0.532608
min       0.000000     0.700000
25%     377.750000     3.400000
50%     755.500000     3.800000
75%    1133.250000     4.100000
max    1511.000000     4.800000

Contoh 5 Baris Pertama Games.csv:
  Unnamed: 0                                  Title  Release Date  ... Playing  Backlogs Wishlist
0          0                           Elden Ring  Feb 25, 2022  ...    3.8K      4.6K     4.8K
1          1                              Hades  Dec 10, 2019  ...    3.2K      6.3K     3.6K
2          2  The Legend of Zelda: Breath of the Wild  Mar 03, 2017  ...    2.5K        5K     2.6K
3          3                            Undertale  Sep 15, 2015  ...     679      4.9K     1.8K
4          4                         Hollow Knight  Feb 24, 2017  ...    2.4K      8.3K     2.3K

[5 rows x 14 columns]
```

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.
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Rating
# Tindakan: Hapus baris dengan missing values, pastikan dalam rentang valid.
rows_before_rating_clean = len(df_games)
df_games['Rating'] = pd.to_numeric(df_games['Rating'], errors='coerce') # Konversi ke numerik, ubah error menjadi NaN
df_games.dropna(subset=['Rating'], inplace=True) # Menghapus baris yang memiliki missing values (termasuk hasil coerce)
# Filter untuk rentang yang valid (jika ada nilai di luar range, hapus)
df_games = df_games[(df_games['Rating'] >= 0.0) & (df_games['Rating'] <= 5.0)]
print(f"\n--- CLEANSING 'Rating' ---")
print(f"  Baris dihapus (missing/out-of-range): {rows_before_rating_clean - len(df_games)}")
```

- **Hasil**
  13 baris dihapus (karena missing values). Semua nilai Rating adalah float dan berada dalam rentang [0.0, 5.0]. Tipe data kolom Rating kini adalah float64.

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.
- **Tindakan dan proses cleansing (Python)**

```
# Fungsi bantu untuk konversi string 'K'/'M' ke numerik
def convert_k_m_to_numeric(value):
    if isinstance(value, str):
        value = value.replace(',', '') # Hapus koma jika ada
        if 'K' in value:
            return float(value.replace('K', '')) * 1000
        elif 'M' in value:
            return float(value.replace('M', '')) * 1000000
    return pd.to_numeric(value, errors='coerce') # Konversi langsung atau jadi NaN jika tidak valid

columns_to_convert_and_filter = {
    'Times Listed': 10, # Batas bawah 10 kali terdaftar
    'Number of Reviews': 10, # Batas bawah 10 kali terdaftar
    'Plays': 100, # Batas bawah 100 plays
    'Playing': None, # Tidak ada batas bawah spesifik
    'Backlogs': None, # Tidak ada batas bawah spesifik
    'Wishlist': None # Tidak ada batas bawah spesifik
}

for col, threshold in columns_to_convert_and_filter.items():
    rows_before_col_clean = len(df_games)
```

```
    df_games[col] = df_games[col].apply(convert_k_m_to_numeric)
    df_games.dropna(subset=[col], inplace=True) # Hapus jika gagal konversi (NaN dari coerce)
    print(f"\n--- CLEANSING '{col}' ---")
    print(f"  Baris dihapus (gagal konversi/missing): {rows_before_col_clean - len(df_games)}")

    if threshold is not None:
        rows_before_threshold_filter = len(df_games)
        df_games = df_games[df_games[col] >= threshold]
        print(f"  Baris dihapus ({col} < {threshold}): {rows_before_threshold_filter - len(df_games)}")

    df_games[col] = df_games[col].astype(int) # Ubah ke integer secara permanen
```

- **Hasil**
  Times Listed: 0 baris dihapus karena gagal konversi, dan 2 baris dihapus karena nilai kurang dari 10. Kolom menjadi int64 dan semua nilai >= 10.
  Number of Reviews: 0 baris dihapus karena gagal konversi, dan 1 baris dihapus karena nilai kurang dari 10. Kolom menjadi int64 dan semua nilai >= 10.
  Plays: 0 baris dihapus karena gagal konversi, dan 12 baris dihapus karena nilai kurang dari 100. Kolom menjadi int64 dan semua nilai >= 100.
  Playing: 0 baris dihapus. Kolom menjadi int64.
  Backlogs: 0 baris dihapus. Kolom menjadi int64.
  Wishlist: 0 baris dihapus. Kolom menjadi int64. Semua kolom ini kini bertipe int64, tanpa missing values, dan nilai 'K'/'M' telah dikonversi dengan benar.

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).
- **Tindakan dan proses cleansing (Python)**

```
# Kolom: Team, Summary
# Tindakan: Imputasi missing values dengan modus.
columns_to_impute_mode = ['Team', 'Summary']
for col in columns_to_impute_mode:
    if df_games[col].isnull().any():
        mode_val = df_games[col].mode()[0]
        df_games[col].fillna(mode_val, inplace=True)
        print(f"\n--- CLEANSING '{col}' ---")
        print(f"  Missing values pada '{col}' diisi dengan modus: {mode_val}")
```

- **Hasil**
  Kolom Team dan Summary sekarang tanpa missing values.

7. Transformasi Data

7.1 Data Aplikasi pada Google Play Store

Berikut adalah transformasi yang dilakukan untuk beberapa atribut pada "Data Aplikasi pada Google Play Store".

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	☑	'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;".";","))
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	☑	'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. =VALUE(SUBSTITUTE(SUBSTITUTE(installs, ",", ""), "+", ""))
Type	☐	Free / Paid	Free / Paid	Tidak perlu dilakukan transformasi karena data sudah benar secara satuan dan range.
Price	☑	'$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 =VALUE(SUBSTITUTE(price,"$",""))
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.

7.2 Data Video Game Terpopuler 1980 - 2023

Berikut adalah transformasi yang dilakukan untuk beberapa atribut pada "Data Video Game Terpopuler".

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	☑	[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	☑	'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	☑	'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(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;",";"."))))))
Number of Reviews	☑	'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	☑	['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	☑	'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(plays;",";""))))))
Playing	☑	'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	☑	'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	☑	'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 Analytic Sederhana

8.1 Data Aplikasi pada Google Play Store

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.

Berikut adalah kode Python yang digunakan untuk membangun model tersebut:

```
import pandas as pd
from sklearn.linear_model import LinearRegression

df = pd.read_csv('Playstore.csv')

df = df.dropna(subset=['Rating', 'Installs'])
df['Installs'] = df['Installs'].astype(str).str.replace('[+,]', '', regex=True).astype(int)

X = df[['Rating']]
y = df['Installs']

model = LinearRegression()
model.fit(X, y)

a = model.coef_[0]
b = model.intercept_

print(f'Installs = {a:.2f} * Rating + {b:.2f}')
```

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

$$Installs(r) = ar + b$$

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:

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

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 :

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

Berikut adalah kode Python yang digunakan untuk membangun model tersebut:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

df = pd.read_csv("Playstore.csv")

df = df.dropna(subset=['Price', 'Installs'])

df['Price'] = df['Price'].astype(str).str.replace('$', '', regex=False)
df = df[df['Price'].str.replace('.', '', 1).str.isnumeric()]
df['Price'] = df['Price'].astype(float)

df = df[df['Price'] < 250]

X = df[['Price']]
y = df['Installs']

model = LinearRegression()
model.fit(X, y)

a = model.coef_[0]
b = model.intercept_
print(f"Installs(p) = {a:.2f} * p + {b:.2f}")
```

Hasil pemodelan regresi menghasilkan persamaan sebagai berikut:

$$Installs(p) = 0.00 * p + 19217171.16$$

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.

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

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

Berikut adalah kode Python yang digunakan untuk membangun model tersebut:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

df = pd.read_csv("GameList_CLEAN.csv")

X = df[['Rating']]
y = df['Wishlist']

model = LinearRegression()
model.fit(X, y)

a = model.coef_[0]
b = model.intercept_
print(f"Persamaan regresi:\nWishlist = {a:.2f} * Rating + {b:.2f}")
```

Model regresi linear yang dihasilkan adalah sebagai berikut:

$$Wishlist(r) = -735.14 * r + 3520.27$$

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

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

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

9. Kesimpulan

Analisis menunjukkan bahwa rating aplikasi hanya memiliki pengaruh lemah terhadap jumlah pengunduhan. Meskipun terdapat korelasi positif, scatter plot dan regresi linear membuktikan bahwa aplikasi dengan rating tinggi belum tentu banyak diunduh. Ini menandakan bahwa pengguna lebih mempertimbangkan faktor lain seperti kategori, visibilitas, dan harga, ketimbang semata-mata nilai rating. Berdasarkan distribusi data tahun pembaruan terakhir, aplikasi bertipe game memang lebih sering diperbarui dibanding tipe lain. Ini bisa terjadi karena game cenderung membutuhkan pembaruan berkala untuk memperbaiki bug, menambah fitur baru, atau menyesuaikan dengan versi Android terbaru, menunjukkan bahwa developer game lebih aktif dalam menjaga engagement pengguna. Harga aplikasi memiliki pengaruh besar dan negatif terhadap jumlah unduhan. Aplikasi gratis mendominasi pasar dengan jumlah unduhan jauh lebih tinggi dibanding aplikasi berbayar. Ketika harga naik, angka unduhan turun secara drastis, menandakan bahwa pengguna sangat sensitif terhadap harga, dan model freemium menjadi pendekatan yang paling berhasil di Google Play Store. Aplikasi bertipe komunikasi memang memiliki jumlah unduhan tertinggi, sebagaimana ditunjukkan dalam analisis rata-rata instalasi per kategori. Hal ini didorong oleh kebutuhan universal pengguna terhadap komunikasi, serta sifat aplikasinya yang esensial dan digunakan secara berulang. Variabel lain yang memengaruhi adalah gratis/tidaknya aplikasi, jumlah review, dan pembaruan berkala.

Tanggal rilis memengaruhi perilaku pemain secara unik: game yang dirilis lebih lama cenderung memiliki lebih banyak total pemain, karena akumulasi waktu. Namun untuk jumlah pemain aktif saat ini, game yang lebih baru memiliki angka yang lebih tinggi. Ini menunjukkan adanya siklus hidup game, di mana popularitas dan engagement berubah seiring waktu sejak perilisan. Popularitas genre game berfluktuasi antar periode, dengan genre Adventure secara konsisten mendominasi jumlah rilis dari 1980–2023. Sementara itu, genre Indie menunjukkan pertumbuhan drastis dalam dekade terakhir, mencerminkan pergeseran tren industri dan meningkatnya akses developer kecil ke pasar global. Ini menunjukkan bahwa preferensi pengguna dinamis dan genre tertentu bisa naik turun tergantung konteks waktu. Genre dengan jumlah pemain tertinggi adalah MOBA, diikuti oleh Shooter dan Racing. Namun, genre MOBA justru memiliki rating rendah, mengindikasikan bahwa faktor seperti kompetisi, komunitas, dan replayability lebih memengaruhi unduhan/pemakaian daripada kualitas persepsi. Variabel lain yang relevan meliputi waktu rilis, hype komunitas, dan exposure media. Angka wishlist dipengaruhi oleh rating, jumlah review, dan visibilitas game, namun analisis menunjukkan bahwa hubungan antara rating dan wishlist tidak kuat atau bahkan negatif. Game yang banyak di-wishlist belum tentu dimainkan, tetapi cenderung masuk backlog. Wishlist mencerminkan minat awal, dan korelasinya yang kuat dengan backlog menunjukkan bahwa pengguna sering menunda memainkan game yang mereka minati.

LAMPIRAN

Pembagian Kerja

Bagian	NIM Pekerja
Pertanyaan Penelitian	13524019 13524067
Data dan Atribut Data	13524020 13524068 13524101
Visualisasi	13524019 13524101
Statistik Deskriptif	13524020 13524067 13524068
Korelasi	13524067 13524068
Data Cleansing	13524101
Transformasi Data	13524019
Data Analytics Sederhana	13524020
Kesimpulan	13524020

Link Repositori Github

https://github.com/Nusss0/Tubes-WI2002-Exploratory_Data_Analysis_and_Regression_Analysis