 ISSN: 3110-8652

Exploratory Data Analysis (EDA) Pada *Dataset* Penjualan *Video Game Global* (1980-2020)

https://doi.org/10.28932/jste.v1i1.13190

Received: 29 Agustus 2025 | *Revised*: 27 September 2025 | *Accepted*: 29 September 2025

Adhitya Putra Pratama Tisnadja[✉#1], Setia Budi[*2]

[#]*Program Studi Teknik Informatika, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha*
Jl. Prof. drg. Surya Sumantri No.65, Bandung, Jawa Barat 40164, Indonesia
[1]2172024@maranatha.ac.id

[*]*Program Studi Magister Ilmu Komputer, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha*
Jl. Prof. drg. Surya Sumantri No.65, Bandung, Jawa Barat 40164, Indonesia
[2]setia.budi@it.maranatha.edu

[✉]Corresponding author: 2172024@maranatha.ac.id

How to cite this article:
A. P. P. Tisnadja, S. Budi, "Exploratory Data Analysis (EDA) Pada Dataset Penjualan Video Game Global (1980-2020)," *Journal of Smart Technology and Engineering*, vol. 1, no. 1, pp. 71-81, 2025, https://doi.org/10.28932/jste.v1i1.13190

Abstrak — **Perkembangan pesat industri video game menegaskan pentingnya analisis data penjualan untuk memahami preferensi pasar secara global. Penelitian ini menerapkan metode Exploratory Data Analysis (EDA) guna menganalisis dataset penjualan video game global yang diperoleh dari Kaggle. Tujuan dari penelitian ini adalah untuk mengidentifikasi faktor-faktor yang memengaruhi keberhasilan penjualan video game, seperti platform, genre, penerbit (publisher), serta distribusi regional. Metode EDA yang digunakan meliputi tahapan eksplorasi data, pembersihan data, visualisasi, serta analisis statistik dengan memanfaatkan Python dan pustaka pendukung seperti Pandas, Matplotlib, dan Seaborn. Hasil analisis menunjukkan bahwa wilayah Amerika Utara memberikan kontribusi terbesar terhadap penjualan video game secara global, dengan PlayStation 2 sebagai platform yang paling dominan. Adapun genre yang paling populer meliputi Action, Sports, dan Shooter. Selain itu, Nintendo dan Electronic Arts merupakan penerbit dengan tingkat penjualan tertinggi. Penelitian ini diharapkan dapat memberikan wawasan dan referensi yang bermanfaat bagi pengembang maupun penerbit video game dalam menyusun strategi pemasaran yang efektif, adaptif, dan berbasis data.**

Kata Kunci— **analisis data; EDA; game; penjualan video game; visualisasi data.**

Exploratory Data Analysis (EDA) on the Global Video Game Sales Dataset (1980-2020)

Abstract — ***The rapid development of the video game industry highlights the importance of analyzing sales data to understand global market preferences. This research applies the Exploratory Data Analysis (EDA) method to analyze a global video game sales dataset sourced from Kaggle. The objective is to identify factors influencing game sales success, such as platform, genre, publisher, and regional distribution. The EDA method involves data exploration, cleaning, visualization, and statistical analysis using Python and libraries such as Pandas, Matplotlib, and Seaborn. The analysis reveals that North America contributes the most to global video game sales, with the PlayStation 2 being the dominant platform. The most popular genres are Action, Sports, and Shooter. Furthermore, publishers like Nintendo and Electronic Arts are among the top contributors in terms of sales. This research provides valuable insights for game developers and publishers in designing effective and data-driven marketing strategies.***

Keywords— ***data analysis, data visualization, EDA, game, video game sales.***

I. PENDAHULUAN

Industri video game telah berkembang dalam beberapa tahun terakhir. Game menjadi salah satu hiburan terbesar di dunia dengan nilai pasar yang terus meningkat. Perkembangan teknologi game mendorong inovasi dalam industri ini. Selain itu, tren seperti esport semakin memperkaya pengalaman bermain dan menarik lebih banyak pemain. Setiap tahunnya, banyak game dirilis di berbagai platform mulai dari PC, konsol, hingga mobile. Namun, tidak semua game yang dirilis berhasil mencatat angka penjualan yang tinggi. Beberapa game mampu menarik perhatian pasar global. Oleh karena itu, faktor-faktor yang mempengaruhi tingkat penjualan video game seperti genre, platform, strategi penyebaran pemasaran wajib dipahami oleh pelaku industri game agar dapat merancang strategi pemasaran yang lebih efektif.

Salah satu metode yang dapat digunakan untuk menganalisis data penjualan video game adalah Exploratory Data Analysis (EDA). EDA merupakan sebuah metode analisis data yang dikembangkan untuk membantu masalah-masalah dalam analisis data secara efisien [1]. Dengan metode ini, berbagai informasi penting seperti tren penjualan, platform, serta faktor-faktor lain yang berkontribusi terhadap kesuksesan sebuah game di pasar global.

Hasil dari EDA dapat memberikan wawasan untuk mengambil keputusan bisnis untuk industri video game. Penerapan metode EDA pada data penjualan video game global dapat membantu dalam menentukan jenis game yang diminati, platform yang potensial, dan strategi pemasaran yang lebih efektif. Selain itu, informasi dari EDA juga dapat digunakan untuk mengidentifikasi pasar yang penting, seperti potensi penyebaran game ke wilayah tertentu seperti Amerika Utara, Eropa, Jepang, dan wilayah lainnya di dunia.

Secara keseluruhan, penerapan EDA dalam analisis data penjualan video game dapat membantu pelaku industri dalam merancang strategi yang lebih berbasis data dan mengurangi risiko kegagalan di pasar global. Dengan pemahaman yang lebih mendalam mengenai tren industri, pengembang dan penerbit dapat mengoptimalkan untuk menciptakan produk yang lebih menarik bagi konsumen. Di zaman digital ini, keputusan yang didukung oleh data menjadi kunci utama dalam meraih kesuksesan di industri video game.

II. KAJIAN TEORI

A. *Exploratory Data Analysis*

Exploratory Data Analysis adalah sebuah metode analisis data yang dikembangkan untuk membantu masalah-masalah dalam analisis data secara efisien [1]. EDA dengan cepat mendeskripsikan jumlah baris ataupun kolom dalam kumpulan data, data yang hilang, tipe data, dan pratinjau data. Dalam menggambarkan distribusi data, EDA menggunakan *bar charts, histogram, box plot.* Selain itu, EDA juga menghitung dan menggambarkan hubungan (korelasi) antar variabel menggunakan *heat map.* EDA berperan penting dalam menganalisis kumpulan data guna merangkum karakteristik statistiknya yang berfokus pada 4 aspek utama, yaitu ukuran pemusatan (*mean, modus*, dan *median*), ukuran penyebaran (standar deviasi dan varians), bentuk distribusi, dan keberadaan outlier. Untuk mendukung analisis tersebut, terdapat beberapa teknik analisis data dan visualisasi yang digunakan secara luas [2].

1) *Data Exploration: Data Exploration adalah tahap pertama dalam analisis data. Pada tahap ini, dapat memahami isi serta karakteristik dari kumpulan data. Tahap ini memberikan informasi mengenai ukuran data, memungkinkan untuk mengidentifikasi nilai yang hilang, serta menemukan kemungkinan hubungan antar data. Visualisasi data dilakukan dengan menggunakan data dalam bentuk tabel untuk memahami karakteristiknya.*

2) *Data Cleaning: Data Cleaning adalah proses perbaikan masalah yang sistematis atau kesalahan dalam data yang tidak terstruktur. Terdapat banyak alasan mengapa data dapat memiliki nilai yang salah, seperti kesalahan ketik, data yang rusak, duplikasi. Data Cleaning menjadi tahapan yang biasanya dilakukan terlebih dahulu sebelum tahapan persiapan data yang lainnya [3].*

3) Hasil Analisis: *Pada metode ini, data yang kompleks divisualisasikan melalui penggunaan bagan, grafik, dan tabel. Sebagian manusia dapat memproses informasi dengan lebih baik melalui bagan dan grafik. Melalui tahap ini, memudahkan untuk menyampaikan konsep. Visualisasi data juga membantu mengidentifikasi area yang perlu diperbaiki dan memperjelas faktor-faktor*

B. *Analisis Korelasi*

Analisis Korelasi adalah analisis statistik yang berusaha untuk mencari hubungan atau pengaruh antara dua buah variabel atau lebih [9]. Proses ini untuk mempelajari kekuatan hubungan tersebut berdasarkan data statistik yang tersedia [10]. Variabel dibagi ke dalam dua bagian, di antaranya variabel bebas (*Independent Variable*) yaitu variabel yang keberadaannya tidak dipengaruhi oleh variabel lain dan variabel terikat (*Dependent Variable*) yaitu variabel yang keberadaannya dipengaruhi oleh variabel yang lain [9].

Dalam penelitian ini, perhitungan korelasi antar kolom menggunakan fungsi *.corr*() dari pustaka *pandas* [12]. Koefisien korelasi *Pearson* memiliki rentang nilai antara -1 dan +1. Jika korelasi linier antara x dan y positif akan menghasilkan $r > 0$. Sedangkan jika korelasi linier antara x dan y negatif menghasilkan $r < 0$. Jika nilai $r = 0$ menunjukkan tidak adanya hubungan [11]. Untuk menguji hubungan antara dua variabel numerik, digunakan fungsi *pearsonr()* dari pustaka *scipy.stats.* Fungsi ini menghitung koefisien korelasi Pearson (r) dan *p-value,* yang digunakan untuk menentukan apakah hubungan yang diamati bersifat signifikan secara statistik atau tidak [13].

C. Python

Python semakin banyak digunakan dalam aplikasi ilmiah yang secara tradisional didominasi oleh R, MATLAB, Stata, SAS, dan lingkungan penelitian *open-source* lainnya [7]. *Python* adalah bahasa interpretatif yang dianggap mudah dipelajari dan berfokus pada keterbacaan kode. *Python* menangani pembuatan aplikasi yang mengandung kata *kunci big data, data mining, deep learning, data science,* hingga *machine learning* [4].

Dalam beberapa tahun terakhir, dukungan *library* Python semakin baik telah menjadikannya alternatif yang kuat. Dikombinasikan dengan kelebihan Python dalam pemrograman umum, ini merupakan pilihan yang sangat baik sebagai bahasa tunggal untuk membangun aplikasi yang berpusat pada data [5]. *Python* didukung oleh *library* yang di dalamnya menyediakan fungsi analisis data dan fungsi *machine learning*, data *preprocessing tools*, serta visualisasi data. Hal ini membuat Python menjadi bahasa pemrograman yang populer pada bidang *data science* dan analisis data [6].

D. Pandas

Pandas adalah paket dari Python yang menyediakan struktur data yang cepat, fleksibel, dan ekspresif. Pandas dirancang untuk mempermudah dan membuat intuitif pekerjaan dengan data "relasional" atau "berlabel". Pandas bertujuan menjadi komponen dasar tingkat tinggi yang mendukung analisis data di dunia nyata menggunakan Python. Selain itu, pandas memiliki tujuan untuk menjadi alat analisis data *open-source* yang paling kuat dan fleksibel dalam bahasa pemrograman. Pandas dibangun di atas NumPy dan dimaksudkan untuk berintegrasi dengan baik dalam lingkungan komputasi ilmiah dengan banyak *library* yang lainnya.

Pandas sesuai dengan berbagai jenis data di antaranya data tabular seperti tabel SQL atau *spreadsheet* Excel, data *ordered* dan *unordered*, data matriks, dan bentuk lain dari kumpulan data observasi ataupun statistik. Terdapat struktur data utama dalam pandas di antaranya *Series* (1 dimensi) dan *DataFrame* (2 dimensi) yang dirancang untuk menangani sebagian besar kasus penggunaan umum di berbagai bidang seperti keuangan, statistik, ilmu sosial, dan berbagai bidang teknik.

Terdapat beberapa kelebihan pandas di antaranya penanganan data yang hilang, kemudahan dalam mengonversi data yang tidak beraturan, penyelarasan data secara otomatis, kemampuan mengubah ukuran secara fleksibel. Untuk fitur seperti *group by* dibuat menjadi fleksibel untuk melakukan operasi *split-apply-combine* pada kumpulan data [7].

E. Matplotlib

Matplotlib adalah *library* tingkat rendah yang menyediakan berbagai macam plot 2 dimensi dan 3 dimensi yang dapat disesuaikan, *termasuk scatter plot, line plot, histogram,* dan lainnya. Matplotlib dibangun di atas *array* NumPy dan sangat kompatibel dengan *library* Python lainnya, seperti Pandas, NumPy, dan scikit-learn. Matplotlib telah banyak digunakan dalam penelitian sebagai alat visualisasi data.

Beberapa cara matplotlib membantu dalam penelitian di antaranya visualisasi data yang kompleks, Matplotlib menyediakan beragam opsi visualisasi termasuk *line plot, scatter plot, histogram, heatmap,* dan lainnya. Hal ini memungkinkan peneliti membuat visualisasi yang menyampaikan data dan tren yang kompleks dengan mudah. Selanjutnya, kemampuan untuk interaksi secara langsung, Matplotlib memungkinkan visualisasi interaktif yang dapat dimanipulasi secara *real-time,* sehingga peneliti dapat mengeksplorasi data dan mengidentifikasi pola. Selanjutnya, konsistensi hasil, Matplotlib menyediakan cara untuk membuat visualisasi berkualitas publikasi yang dapat dengan mudah diperbanyak dalam makalah penelitian, presentasi, dan materi lainnya [8].

F. Seaborn

Seaborn merupakan *library* visualisasi data yang banyak dimanfaatkan oleh peneliti dari berbagai bidang untuk menyajikan data dalam bentuk visual guna memperoleh wawasan yang lebih mendalam. Berikut adalah beberapa cara Seaborn membantu dalam penelitian di antaranya, Analisis data eksploratori, Seaborn menyediakan antarmuka tingkat tinggi untuk membuat grafik statistik yang menarik dan informatif. *Library* ini memiliki banyak fungsi untuk membuat berbagai jenis *plot* seperti *scatter plot, bar plot, heatmap,* dan lainnya. Plot ini memungkinkan peneliti untuk mengekplorasi data mereka serta mengidentifikasi pola dan hubungan. Selanjutnya, Seaborn bersifat informatif dan menarik secara visual yang menjadikannya alat yang efektif untuk menyampaikan hasil penelitian yang lebih luas. Langkah selanjutnya adalah pemodelan statistik, Seaborn menyediakan berbagai fungsi untuk menyajikan hasil dari model statistik seperti model regresi dan analisis faktor. Yang terakhir, Seaborn memungkinkan peneliti untuk menyesuaikan plot sesuai kebutuhan. Sebagai contoh, peneliti dapat mengatur warna, jenis huruf, dan gaya plot agar sesuai dengan preferensi atau persyaratan publikasi [8].

III. ANALISIS DAN RANCANGAN SISTEM

A. Analisis Kolom Penjualan (NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales)

Analisis ini bertujuan untuk mengidentifikasi tren penjualan *video game* berdasarkan wilayah yaitu Amerika Utara (*NA_Sales*), Eropa (*EU_Sales*), Jepang (*JP_Sales*), wilayah lainnya (*Other_Sales*), dan penjualan global (*Global_Sales*) dari tahun ke tahun. Melalui visualisasi penjualan per tahun di masing-masing wilayah dapat diperoleh wawasan mengenai tahun-tahun dengan performa penjualan terbaik, serta menggambarkan preferensi pasar yang berbeda di setiap wilayah.

Selain itu, dilakukan juga analisis untuk rata-rata penjualan di setiap wilayah yang bertujuan untuk mengidentifikasi besarnya rata-rata penjualan di setiap wilayah. Dengan memahami sebaran rata-rata ini, perusahaan dapat mengidentifikasi wilayah dengan potensi pasar terbesar.

B. Analisis Game dan Genre di Setiap Wilayah

Penentuan *genre* dan alur *game* yang tepat sangat berpengaruh pada daya tarik game di pasar global. Genre yang populer di satu wilayah mungkin berbeda dengan wilayah lainnya, sehingga pemahaman ini menjadi kunci. Dengan memahami faktor ini, pengembang dapat merancang *game* yang lebih sesuai dengan selera pasar masing-masing wilayah.

Analisis ini bertujuan untuk mengidentifikasi *game* dan *Genre* yang paling diminati di masing-masing wilayah berdasarkan penjualan tertinggi per tahun. Informasi ini dapat digunakan untuk menentukan strategi pemasaran sesuai dengan tren pasar di setiap wilayah. Selain itu, analisis ini juga dapat membantu perusahaan dalam mengidentifikasi *game* yang memiliki potensi untuk diperluas ke pasar lain.

C. Platform dengan Penjualan Tertinggi

Platform memainkan peran penting dalam kesuksesan penjualan *video game*. Setiap platform seperti *PC, PS,* atau *platform* lainnya memiliki karakteristik yang berbeda. *Platform* yang populer menawarkan fitur yang mendukung untuk pengalaman bermain video game, sehingga menarik banyak pemain.

Analisis ini bertujuan untuk mengidentifikasi *Platform* yang paling sering digunakan oleh *game* dengan penjualan tertinggi secara keseluruhan. Dengan analisis ini, dapat diketahui *Platform* yang dominan dalam mendukung game populer di pasar global. Informasi mengenai Platform ini bermanfaat bagi pengembang game untuk menentukan *Platform* prioritas saat merilis game baru. Pada analisis ini juga dilakukan analisis lebih lanjut diantaranya analisis distribusi penjualan *platform.*

1) *Distribusi Penjualan Platform:* Fokus utama analisis ini membandingkan *Platform* yang dominan di setiap wilayah. Analisis ini bertujuan untuk mengetahui *Platform* yang banyak digunakan di wilayah tertentu. Dengan mengetahui Platform yang mendominasi di Amerika Utara, Eropa, Jepang, dan wilayah lainnya memudahkan pemasaran dan pengembangan *game.*

D. Analisis Penjualan Berdasarkan Publisher

Publisher game menjadi peran yang krusial dalam keberhasilan pemasaran *game* itu sendiri. Selain bertanggung jawab atas penyebaran, *publisher* berperan dalam promosi, penentuan pasar target, dan strategi penjualan. *Publisher* berpengalaman seharusnya memahami tren pasar dan memiliki jaringan luas untuk kesuksesan game di berbagai wilayah.

Fokus utama analisis ini adalah mengamati *Publisher* yang berhasil memasarkan *game* dengan penjualan tinggi. Analisis ini dapat memberikan wawasan mengenai *Publisher* yang konsisten dalam menghasilkan *game* di berbagai wilayah maupun secara *global*. Analisis ini penting bagi perusahaan yang berencana mengembangkan *game,* karena dapat memberikan referensi mengenai *Publisher* yang mampu menghadirkan *game* berkualitas tinggi sesuai dengan tren pasar.

IV. IMPLEMENTASI

A. Persiapan Dataset

Pada tahap ini, dilakukan proses persiapan data dengan mengimpor beberapa *library* yang diperlukan *dataset* utama yang akan dianalisis. Pustaka yang digunakan di antaranya Pandas, Matplotlib, Seaborn. Setelah itu, dataset diimpor ke dalam Python menggunakan Pandas dan disimpan pada *dataframe* yang diberi nama variabel df. Setelah data berhasil dimuat, dilakukan eksplorasi awal untuk memahami struktur, isi, serta kualitas data secara umum sebelum masuk ke tahap analisis yang lebih mendalam.

1) *Tampilan Dataset:* Pada tahapan awal, dilakukan pemeriksaan awal terhadap isi dataset menggunakan perintah *df.head().* Fungsi ini digunakan untuk menampilkan lima baris awal dari *dataframe* yang diberi nama df. Tujuan dari tahapan ini untuk memberikan gambaran awal mengenai struktur dan isi data.

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37

Gambar 1. Tampilan Data Teratas

2) *Informasi Struktur Dataset:* Perintah *df.info()* digunakan untuk menampilkan informasi mengenai struktur dataset yang telah diimpor ke dalam *dataframe* df. Fungsi ini memberikan detail penting seperti total baris, jumlah kolom, jumlah nilai *non-null* di setiap kolom, tipe data masing-masing kolom. *Dataset* ini memiliki total 16.292 baris dan 11 kolom.

Semua kolom dalam *dataset* ini memiliki jumlah nilai *non-null* yang lengkap, yang berarti tidak ada nilai yang hilang dalam *dataset.* Tipe data dalam *dataset* terdiri dari 3 jenis di antaranya *int64, float64,* dan *object.*

```
<class 'pandas.core.frame.DataFrame'>
Index: 16291 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16291 non-null  int64
 1   Name          16291 non-null  object
 2   Platform      16291 non-null  object
 3   Year          16291 non-null  float64
 4   Genre         16291 non-null  object
 5   Publisher     16291 non-null  object
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.5+ MB
```

Gambar 2. Informasi Dataset

3) *Identifikasi Missing Value:* Pada tahap ini, dilakukan identifikasi terhadap nilai-nilai kosong (*missing values*) dalam dataset menggunakan perintah *df.isnull().sum().* Informasi ini penting untuk mengetahui seberapa banyak data yang hilang pada masing-masing kolom.

```
Rank             0
Name             0
Platform         0
Year           271
Genre            0
Publisher       58
NA_Sales         0
EU_Sales         0
JP_Sales         0
Other_Sales      0
Global_Sales     0
dtype: int64
```

Gambar 3. Hasil Identifikasi Missing Value

4) *Identifikasi Missing Value:* Pada tahap ini, dilakukan identifikasi terhadap nilai-nilai kosong (*missing values*) dalam dataset menggunakan perintah *df.isnull().sum().* Informasi ini penting untuk mengetahui seberapa banyak data yang hilang pada masing-masing kolom. Berdasarkan hasil yang ditampilkan, diketahui bahwa sebagian besar kolom tidak memiliki *missing value.* Namun demikian, terdapat 2 kolom yang memiliki data yang hilang di antaranya kolom *Year* memiliki *missing value* sebanyak 271 dan kolom *Publisher* memiliki *missing value* sebanyak 58.

```
Rank           0
Name           0
Platform       0
Year           0
Genre          0
Publisher      0
NA_Sales       0
EU_Sales       0
JP_Sales       0
Other_Sales    0
Global_Sales   0
dtype: int64
```

Gambar 4. Pembersihan Data Missing Value

5) *Pembersihan Data:* Proses ini dilakukan pembersihan data dari nilai-nilai yang hilang (*missing values*) yang dapat berdampak negatif terhadap hasil analisis. Berdasarkan hasil perintah *df.isnull().sum(),* ditemukan bahwa terdapat nilai kosong pada beberapa kolom yaitu kolom Year sebanyak 271 baris dan kolom Publisher sebanyak 58 baris. Untuk menghindari potensi kesalahan dalam analisis, semua baris yang memiliki nilai kosong dihapus dari dataset.

 ISSN: 3110-8652

```
df.duplicated().sum()
```

```
0
```

Gambar 5. Jumlah Duplikasi Data Pada Dataset

6) *Deteksi Duplikasi Data*: Pada tahap ini, bertujuan untuk mendeteksi adanya data yang terduplikasi dalam *dataset*. Perintah *df.duplicated().sum()* digunakan untuk menghitung jumlah baris yang merupakan duplikat dari baris lainnya dalam *dataframe* df. Didapatkan hasil 0, yang berarti tidak terdapat baris yang terduplikasi dalam *dataset*. Dapat disimpulkan bahwa data yang digunakan telah bebas dari duplikasi.

```
df.duplicated().sum()
```

```
0
```

Gambar 6. Jumlah Duplikasi Data Pada Dataset

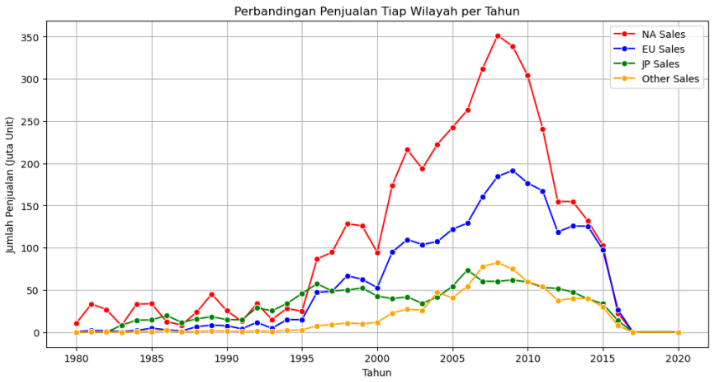
7) *Statistik Deskriptif Dataset:* Perintah *df.describe()* digunakan untuk menampilkan statistik deskriptif dari kolom-kolom numerik dalam *dataframe* df. Fungsi ini secara otomatis menghitung nilai statistik seperti jumlah data (*count*), rata-rata (*mean*), standar deviasi (*std*), nilai minimum dan maksimum, nilai kuartal.

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16598.000000	16598.000000	16598.000000	16598.000000	16598.000000
mean	0.264667	0.146652	0.077782	0.048063	0.537441
std	0.816683	0.505351	0.309291	0.188588	1.555028
min	0.000000	0.000000	0.000000	0.000000	0.010000
25%	0.000000	0.000000	0.000000	0.000000	0.060000
50%	0.080000	0.020000	0.000000	0.010000	0.170000
75%	0.240000	0.110000	0.040000	0.040000	0.470000
max	41.490000	29.020000	10.220000	10.570000	82.740000

Gambar 7. Statistik Deskriptif Dataset

B. Analisis Penjualan per Wilayah

Pada analisis ini, kolom *NA_Sales*, *EU_Sales*, *JP_Sales*, dan *Other_Sales* menjadi variabel target yang dianalisis untuk mengetahui kontribusi masing-masing wilayah terhadap total penjualan *video game*. Keempat kolom ini mencerminkan jumlah unit yang terjual di wilayah Amerika Utara, Eropa, Jepang, dan wilayah lainnya, serta dapat memberikan gambaran bagaimana distribusi pada empat wilayah tersebut. Analisis ini bertujuan untuk melihat distribusi penjualan per wilayah dari tahun ke tahun. Dengan mengelompokkan data berdasarkan tahun rilis game, dapat diamati bagaimana tren penjualan berkembang di setiap wilayah.

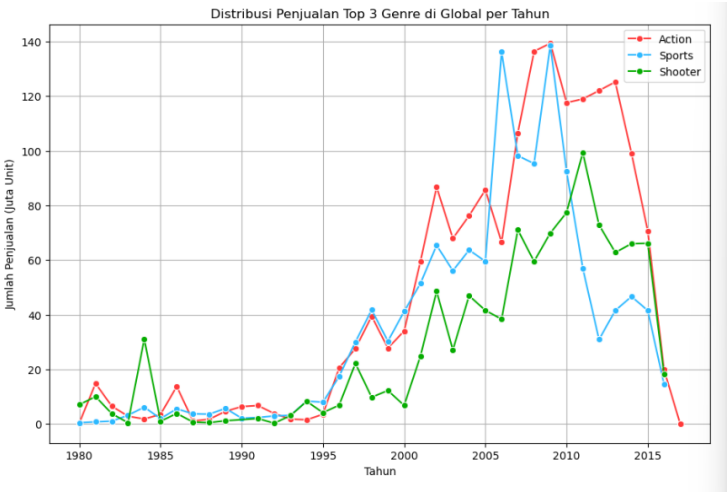


Gambar 8. Tren Penjualan Tiap Wilayah per Tahun

C. Analisis Distribusi Penjualan Genre

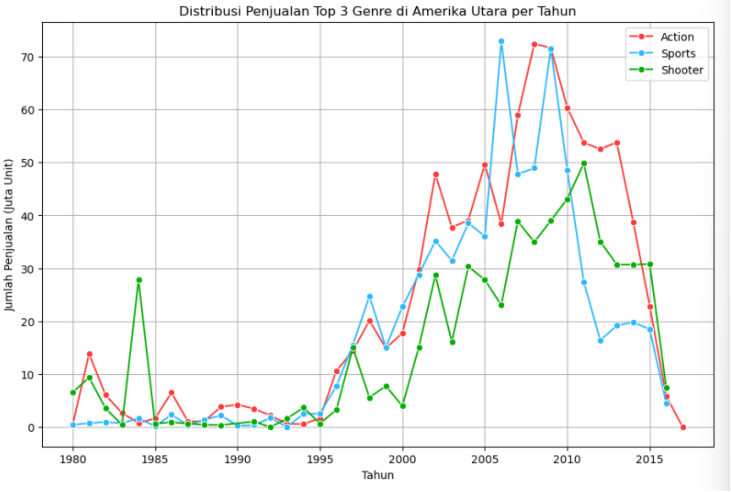
Analisis ini dilakukan dengan mengelompokkan data berdasarkan kolom *Genre* dan menghitung total penjualan dari masing-masing *genre*. Data penjualan yang digunakan mencakup Global_Sales, serta penjualan per wilayah yaitu *NA_Sales*, *EU_Sales*, *JP_Sales*, dan *Other_Sales*. Hasil analisis ini divisualisasikan dalam bentuk *line plot*, karena grafik ini mampu menunjukkan pola distribusi penjualan dari masing-masing *genre* terutama 3 *genre* dengan nilai total penjualan terbesar.

 ISSN: 3110-8652



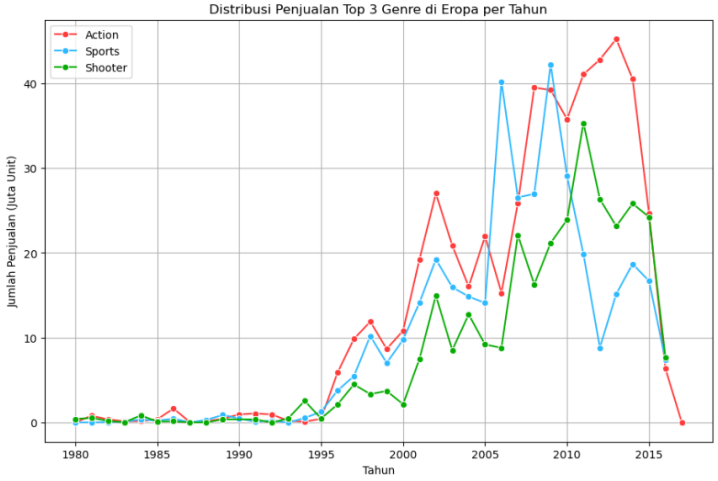
Gambar 9. Distribusi Penjualan Top 3 Genre di Global

Pada visualisasi yang dihasilkan, puncak tahun penjualan untuk *genre action* pada tahun 2009, *genre sports* pada tahun 2009, dan *genre shooter* pada tahun 2011. Selanjutnya, analisis distribusi terhadap *genre game* di masing-masing wilayah. Yang pertama akan dianalisis yaitu wilayah Amerika Utara. Analisis ini dilakukan menggunakan visualisasi *line plot.* Difokuskan pada 3 *genre* teratas di Amerika Utara diantaranya *action*, *sports*, dan *shooter*.



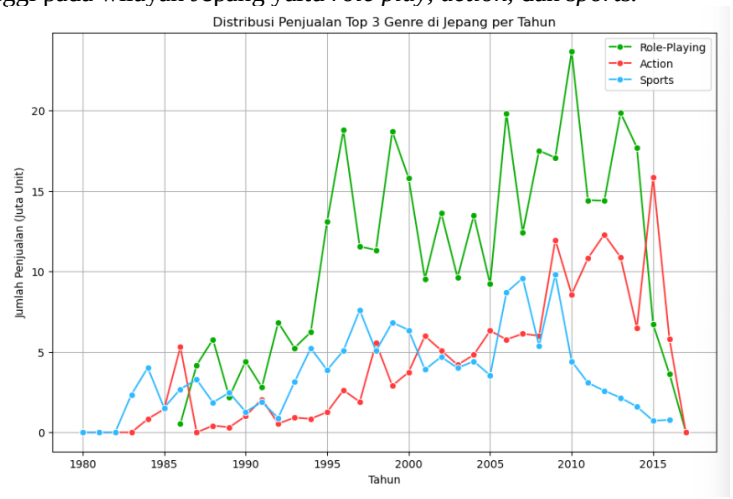
Gambar 10. Distribusi Penjualan Top 3 Genre di Amerika Utara

Pada visualisasi yang dihasilkan, puncak tahun penjualan untuk *genre action* pada tahun 2008, *genre sports* pada tahun 2006, dan *genre shooter* pada tahun 2011. Selanjutnya, yang akan dianalisis distribusi penjualan *genre* pada wilayah Eropa. Difokuskan pada 3 *genre* tertinggi pada wilayah Eropa yaitu *action*, *sports*, dan *shooter*. Setelah itu, akan ditampilkan judul *game* yang rilis pada tahun puncak penjualan di wilayah Eropa.



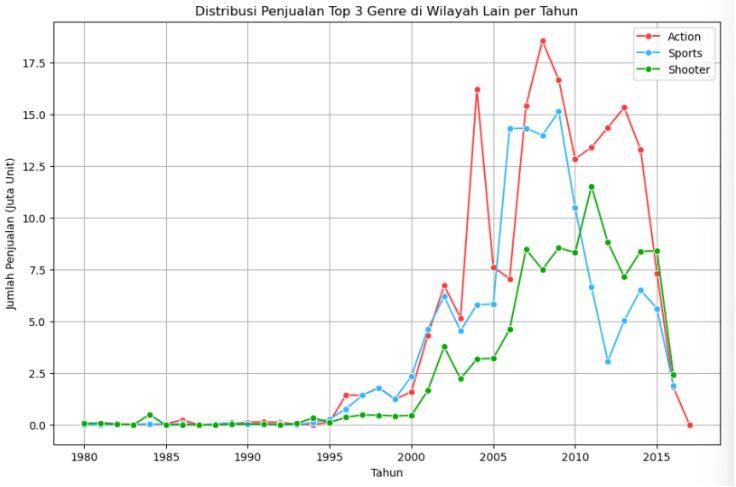
Gambar 11. Distribusi Penjualan Top 3 Genre di Eropa

Pada visualisasi yang dihasilkan, puncak tahun penjualan untuk *genre action* pada tahun 2013, *genre sports* pada tahun 2009, dan *genre shooter* pada tahun 2011. Selanjutnya, yang akan dianalisis distribusi penjualan *genre* pada wilayah Jepang. Difokuskan pada 3 *genre* tertinggi pada wilayah Jepang yaitu *role play*, *action*, dan *sports.*



Gambar 12. Distribusi Penjualan Top 3 Genre di Jepang

Pada visualisasi yang dihasilkan, puncak tahun penjualan untuk *genre role-playing* pada tahun 2010, *genre action* pada tahun 2015, dan *genre sports* pada tahun 2009. Pada bagian ini, terdapat analisis distribusi penjualan *genre* yang terakhir yaitu wilayah lainnya (*Other_Sales*), selain Amerika Utara, Eropa, dan Jepang. Pada analisis ini difokuskan pada 3 *genre* teratas pada kolom ini.



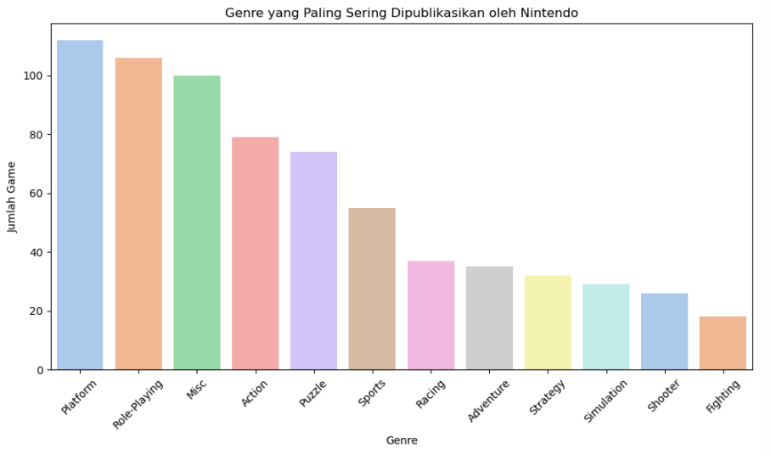
Gambar 13. Distribusi Penjualan Top 3 Genre di Wilayah Lain

Pada visualisasi yang dihasilkan, puncak tahun penjualan untuk *genre action* pada tahun 2008, *genre sports* pada tahun 2009, dan *genre shooter* pada tahun 2011. Dapat disimpulkan bahwa sebagian besar semua wilayah menyukai *game* dengan *genre action*, *sports*, dan *shooter.* Namun, terdapat perbedaan pada wilayah Jepang, di mana *genre* dengan total penjualan tertingginya adalah *genre role-playing.*

D. Analisis Perbandingan Genre Game Berdasarkan Publisher

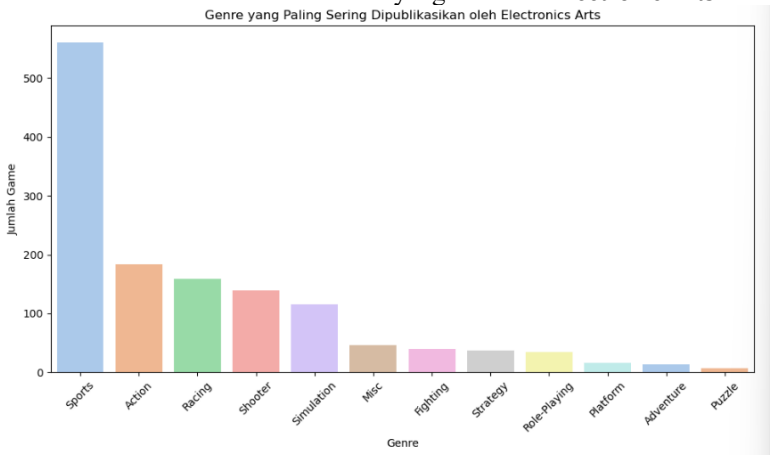
Pada bagian ini, dilakukan analisis untuk mengetahui *genre* apa saja yang paling sering dipublikasikan oleh beberapa *publisher*. Pada analisis ini difokuskan pada 3 *publisher* dengan total penjualan tertinggi antara lain *Nintendo*, *Electronic Arts*, dan *Activision*. Analisis ini bertujuan untuk memahami fokus pengembangan dan distribusi *game* dari masing-masing *publisher* berdasarkan *genre* dan dapat menjadi acuan bagi *developer* atau *publisher* lain untuk mengikuti keberhasilan *publisher* ini dalam meraih pasar global maupun lokal.

 ISSN: 3110-8652



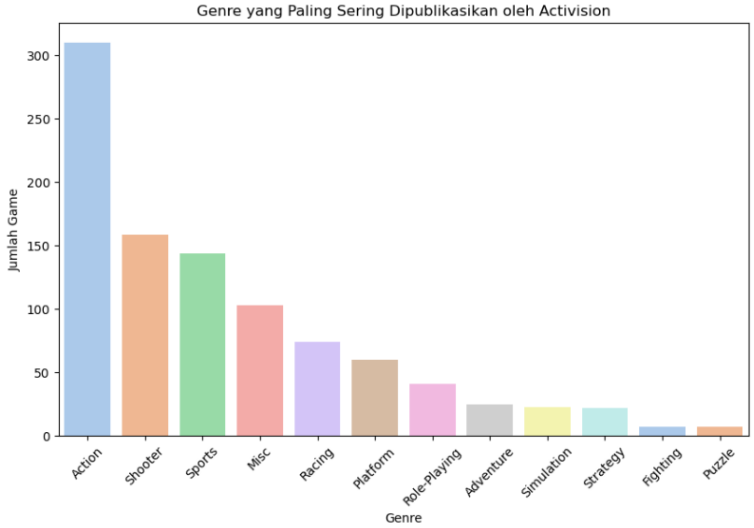
Gambar 14. Genre Game Yang Dipublikasikan Oleh Nintendo

Hasil dari visualisasi menunjukkan, *genre platform* dengan jumlah *game* terbanyak, di atas 100 game. Visualisasi ini menunjukkan bahwa Nintendo memiliki fokus utama pada *genre Platform* dan *Role-Playing*, yang sesuai dengan *game* populer mereka seperti *Super Mario*. Selanjutnya, mengidentifikasi *genre* yang dipublikasikan oleh *Electronic Arts*. Data yang digunakan dalam analisis ini difilter berdasarkan kolom Publisher yang bernilai ”*Electronic Arts*”.



Gambar 15. Genre Game Yang Dipublikasikan Oleh Electronic Arts

Electronic Arts berfokus pada game dengan *genre sports* dengan jumlah game mencapai lebih dari 500 *game*. *Game electronic arts* yang cukup populer di antaranya FIFA dan Madden NFL. Selanjutnya, mengidentifikasi *genre* yang paling sering dipublikasikan oleh *publisher Activision.* Data yang digunakan untuk analisis ini difilter melalui kolom *Publisher* yang bernilai “*Activision*”.

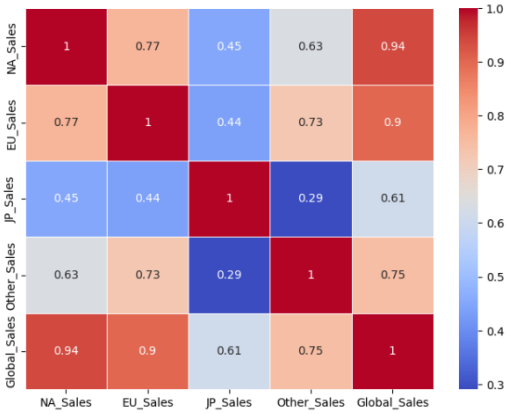


Gambar 16. Genre Game Yang Dipublikasikan Oleh Activision

Hasil dari visualisasi menunjukkan bahwa *genre game* yang paling sering dipublikasikan oleh Activision adalah *genre Action*, mendominasi dengan lebih dari 300 judul. Hal ini menunjukkan bahwa Activision lebih mengutamakan *genre action* dan *shooter*. *Game* yang cukup terkenal di antaranya *LEGO Indiana Jones, Call Of Duty,* dan *Spider-Man*.

E. Matriks Korelasi

Analisis ini bertujuan untuk mengetahui hubungan atau korelasi antara beberapa variabel numerik yang terdapat dalam *dataset* seperti *Global_Sales, NA_Sales, EU_Sales, JP_Sales,* dan *Other_Sales.* Dari analisis ini, dapat diketahui wilayah mana yang paling berkontribusi terhadap total penjualan *Global_Sales.* Bentuk visualisasi yang dipakai untuk analisis ini adalah *heatmap,* karena mampu merepresentasikan hubungan antar kolom numerik. Pada grafik ini menunjukkan nilai korelasi antar 2 variabel, dengan warna yang menunjukkan kekuatan antar 2 variabel tersebut.



Gambar 17. Matriks Korelasi

Dapat dilihat bahwa seluruh variabel penjualan wilayah (*NA_Sales, EU_Sales, JP_Sales,* dan *Other_Sales*) memiliki hubungan positif terhadap *Global_Sales.* Hal ini terlihat dari matriks korelasi, di mana nilai korelasi tertinggi ditunjukkan oleh *NA_Sales* sebesar 0.94, yang mengindikasikan adanya hubungan yang sangat kuat antara penjualan di Amerika Utara dengan total penjualan global. *EU_Sales* juga menunjukkan hubungan yang sangat kuat dengan nilai korelasi sebesar 0.90, sedangkan *Other_Sales* memiliki nilai korelasi sebesar 0.75 yang termasuk dalam kategori kuat. Sementara itu, *JP_Sales* memiliki korelasi paling rendah sebesar 0.61, namun masih menunjukkan adanya hubungan yang cukup kuat dengan penjualan global.

	Region	r	p-value
0	NA_Sales	0.941047	0.0
1	EU_Sales	0.902836	0.0
2	JP_Sales	0.611816	0.0
3	Other_Sales	0.748331	0.0

Gambar 18. Perhitungan p-value

Sementara itu, berdasarkan hasil perhitungan nilai p menggunakan fungsi *pearsonr*(), diperoleh *p-value* sebesar 0.0 untuk seluruh variabel penjualan wilayah terhadap *Global_Sales*. Nilai ini menunjukkan bahwa hubungan korelasi yang ditemukan bersifat signifikan secara statistik.

V. KESIMPULAN

Penelitian ini menerapkan metode *Exploratory Data Analysis* (EDA) pada *dataset* penjualan *video game* global untuk memahami faktor-faktor yang memengaruhi kesuksesan penjualan. Hasil analisis menunjukkan bahwa wilayah Amerika Utara dan Eropa merupakan pasar utama dengan kontribusi terbesar terhadap total penjualan global, berdasarkan nilai korelasi yang sangat kuat terhadap *Global_Sales* serta p-value yang signifikan secara statistik.

Dalam hal *genre*, *Action*, *Sports*, dan *Shooter* menjadi *genre* paling populer secara global, meskipun terdapat perbedaan preferensi di beberapa wilayah seperti Jepang yang lebih menyukai genre *Role-Playing.* setiap *publisher* cenderung memiliki fokus tertentu terhadap *genre game* yang mereka terbitkan. Beberapa *publisher* secara konsisten menerbitkan *game* dari satu atau dua *genre* utama yang menjadi ciri khas atau kekuatan mereka di pasar. Misalnya, Nintendo lebih banyak memfokuskan diri pada *genre platform* dan *role-playing.* Sementara itu, *publisher* lain seperti *Electronic Arts* cenderung mendominasi *genre sports, dan* Activision banyak menerbitkan *game* dengan *genre action*

Proses EDA yang dilakukan meliputi eksplorasi awal, pembersihan data, visualisasi, hingga analisis korelasi berhasil memberikan wawasan yang komprehensif terhadap tren penjualan video game global. Temuan ini dapat digunakan sebagai dasar bagi pengembang dan publisher dalam menyusun strategi pemasaran, pengembangan produk, dan ekspansi pasar yang lebih tepat sasaran dan berbasis data.

DAFTAR PUSTAKA

[1] Behrens, J., & Yu, C.-H. (2003). Exploratory Data Analysis. Dalam J. Schinka, & W. Velicer, Handbook of Psychlogy (hal. 42). Hoboken: Wiley.
[2] Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 4727-4728.
[3] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, 2020.
[4] Enterprise, J. (2019). Python Untuk Programmer Pemula. Jakarta: PT Elex Media Komputindo.
[5] W. McKinney, Pyton for Data Analysis, United States: 0'Reilly Media, 2022.
[6] Pane, S. F., & Saputra, Y. A. (2020). Big Data: Classification Behavior Menggunakan Python. Bandung: Kreatif Industri Nusantara.
[7] McKinney, W. (2012). Pandas: powerful Python data analysis. 1.
[8] Addepalli, L., & Ali, W. (2023). Assessing the Perfomance of Python Data Visualization Libraries: A review. International Journal of Computer Engineering in research Trends, 32.
[9] Muhson, A. (2006). TEKNIK ANALISIS KUANTITATIF. 2-3.
[10] *A., J. (2018). Correlation analysis. IRIS.*
[11] Franzese, M., & Luliano, A. (2018). Correlation Analysis. In S. Ranganathan, *Encyclopedia of Bioinformatics and Computational Biology* (p. 709). Elsevier.
[12] pandas. (n.d.). pandas.DataFrame.corr. Retrieved from pandas: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html
[13] SciPy Developers, "scipy.stats.mstats.pearsonr SciPy v1.13.0 Manual", SciPy.org. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.pearsonr.html.