

**LAPORAN TUGAS BESAR
PENGENALAN KOMPUTASI
ANALISIS DATA PENJUALAN VIDEO GAME**



**Kelompok 09 – KU1102 Pengenalan Komputasi Kelas 18
Dosen: Dr. Fazat Nur Azizah, ST, M.Sc.**

19622014
19622144
16522034
19622224

Jonathan Wiguna
Melati Anggraini
Kean Malik Aji Santoso
Kayla Dyara

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
NOVEMBER 2022**

DAFTAR ISI

DAFTAR ISI	2
BAB 1 PENDAHULUAN	3
1.1 Latar Belakang	3
1.2 Rumusan Masalah	3
1.3 Tujuan	3
BAB II ANALISIS DATA	4
2.1 Deskripsi Data dan File	4
2.2 Karakteristik Data	6
2.3 Statistik Data	7
2.4 Visualisasi Data	10
2.5 Korelasi Data	17
2.6 Data Cleansing	17
BAB III KESIMPULAN	20
3.1 Kesimpulan	20
DAFTAR PUSTAKA	21
LAMPIRAN	22

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Video gim telah menjadi hiburan favorit bagi sebagian besar dari kita. Perkembangan zaman membuat perkembangan video gim makin pesat. Bisnis video gim ini juga sangat menjanjikan karena menghasilkan miliaran dolar setiap tahunnya. Pada Juli 2018, video game menghasilkan US\$134,9 miliar per tahun dalam penjualan global. Riset dari Analisis Ampere menunjukkan tiga poin: sektor ini telah tumbuh secara konsisten setidaknya sejak 2015 dan berkembang 26% dari 2019 hingga 2021, mencapai rekor \$191 miliar; pasar game dan layanan global diperkirakan menyusut 1,2% setiap tahun menjadi \$188 miliar pada tahun 2022; industri ini tidak tahan resesi.

Tentunya, penjualan video gim ini menyumbang data yang cukup besar sehingga diperlukan analisis data untuk mengetahui perkembangan penjualan industri video gim. Dari perkembangan tersebut, kita dapat mengetahui tren pasar dan perkembangan video gim sehingga nantinya monetisasi video gim diharapkan bisa menyumbang profit yang besar. Oleh karena itu, dalam bahasan laporan kali ini, kami akan menganalisis dan melakukan pembersihan data penjualan video gim yang didapatkan dari Kaggle yang merupakan rumah bagi banyak kumpulan data dan kompetisi semacam itu. Kumpulan data tersebut berisi informasi mengenai penjualan video gim di berbagai wilayah, seperti Amerika Utara, Eropa, Jepang dan secara global, sekaligus memberikan informasi mendetail mengenai gim yang terdiri atas urutan (*Rank*), nama gim (*Name*), platform (*Platform*), tahun rilis (*Year*), genre (*Genre*), dan perilis (*Publisher*).

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, rumusan masalahnya sebagai berikut.

- 1) Jika data “kotor”, bagaimana cara melakukan “pembersihan”-nya?
- 2) Bagaimana deskripsi data dan *file*-nya?
- 3) Bagaimana karakteristik data-datanya?
- 4) Bagaimana statistik data-datanya?
- 5) Bagaimana visualisasi data-datanya?
- 6) Bagaimana korelasi-korelasi antardata, khususnya data numerik?

1.3 Tujuan

Berdasarkan rumusan masalah yang telah disebutkan, tujuannya sebagai berikut.

- 1) mengetahui dan melakukan pembersihan terhadap data jika data “kotor”,
- 2) mengetahui deskripsi data dan *file*-nya beserta cara mengetahuinya,
- 3) mengetahui karakteristik data-datanya beserta caranya,
- 4) mengetahui statistik data-datanya beserta cara memperoleh data statistiknya,
- 5) mengetahui visualisasi data-datanya beserta cara memvisualisasikan data-datanya, dan
- 6) mengetahui korelasi antardata numerik.

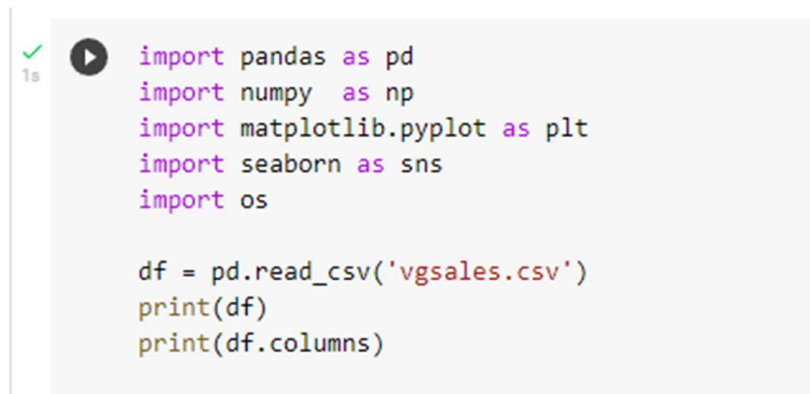
BAB II ANALISIS DATA

2.1 Deskripsi Data dan File

A. Tujuan Analisis data

Sebelum data dianalisis, data diimpor terlebih dahulu dengan modul pandas. Kelompok kami mengambil data tentang penjualan video game. Data tentang penjualan video game tersebut diambil dari <https://www.kaggle.com/datasets/arslanali4343/sales-of-video-games> dengan format *data comma-separated values* (CSV). Tujuan dari analisis data ini adalah untuk mengetahui :

1. Sepuluh gim terlaris berdasarkan penjualan global
2. Banyaknya *platform* per gim (20 teratas)
3. Platform dengan produksi gim terbanyak
4. 10 publisher game terbanyak
5. Jumlah perilis gim per tahun
6. Penjualan rata-rata tiap tahun
7. Banyaknya game per genre
8. Penjualan NA,EU,JP,10 game terlaris
9. Scatter plot antara penjualan NA,EU,JP, terhadap penjualan global



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

df = pd.read_csv('vgsales.csv')
print(df)
print(df.columns)
```

Gambar 1: Cara Baca Data, Impor Data, Impor Modul, dan Menampilkan Nama Atribut yang Ada

B. Informasi Data

Tahap selanjutnya adalah mencari informasi-informasi terkait dengan data yang sudah diambil.



```
# Format data
print(f'Format data tersebut adalah csv.')
```

```
# Asal data
source_data= 'https://www.kaggle.com/datasets/arslanali4343/sales-of-video-games?resource=download'
print(f'Data diambil dari : {source_data}')
```

```
# Dimensi data
baris, kolom = df.shape
print(f'Dimensi : {kolom} kolom & {baris} baris")
```

```
# Ukuran file
size = os.path.getsize('vgsales.csv')
print(f"Ukuran file : {size} bytes")
# df.info()
```

```
#Kolom dan kategori jenis data
print("Kategori kolom : ")
kolom1 = list(df.columns)
for index in range(len(kolom1)):
    print(str(index+1) + ", " + str(kolom1[index]) + " (" + str(df.dtypes[str(kolom1[index]])) + ")")
print()
```

Format data tersebut adalah csv.
 Data diambil dari : <https://www.kaggle.com/datasets/arslanali4343/sales-of-video-games?resource=download>
 Dimensi : 11 kolom & 16598 baris
 Ukuran file : 1372188 bytes
 Kategori kolom :
 1. Rank (category)
 2. Name (category)
 3. Platform (category)
 4. Year (category)
 5. Genre (category)
 6. Publisher (category)
 7. NA_Sales (float64)
 8. EU_Sales (float64)
 9. JP_Sales (float64)
 10. Other_Sales (float64)
 11. Global_Sales (float64)

Gambar 2, 3: Cara dan Hasil dalam Menjelaskan Informasi Data

Kesimpulan yang dapat diambil adalah data dengan jenis “int64” dan “float64” merupakan angka, tetapi “int64” hanya merupakan bilangan bulat dan “float64” angka yang tidak bula. Data-data kategorikal terdiri atas kategorikal nominal, kategorikal ordinal, dan kategorikal binari.

C. Jenis Data Pada Setiap Atribut

Tabel 1: Tabel Penjelasan Atribut

No	Nama Atribut	Deskripsi	Jenis Data
1	Rank	Urutan ranking rata-rata penjualan gim	Categorical Ordinal
2	Name	Nama gim	Categorical Nominal
3	Platform	Platform agar gim bisa dimainkan (PC,PS4,PSP,dll)	Categorical Nominal
4	Year	Tahun gim dirilis	Categorical Ordinal
5	Genre	Genre gim	Categorical Nominal
6	Publisher	Perusahaan yang merilis gim	Categorical Nominal
7	NA_Sales	Penjualan di Amerika Selatan	Quantitative Continues
8	EU_Sales	Penjualan di Eropa (dalam juta)	Quantitative Continues
9	JP_Sales	Penjualan di Jepang (dalam juta)	Quantitative Continues
10	Other_Sales	Penjualan di negara-negara selain negara yang telah disebutkan (dalam juta)	Quantitative Continues

11	Global_Sales	Total penjualan dunia (dalam juta)	Quantitative Continues
----	--------------	------------------------------------	------------------------

2.2 Karakteristik Data

Untuk menganalisis karakteristik dari dataset, kami membagi karakteristik data menjadi 2 bagian, yaitu data kuantitatif dan data kategorikal. Berikut *source code* yang digunakan untuk menentukan karakteristik tiap data :

1. Data Kuantitatif

```
# Karakteristik Data
df.columns
# Berdasarkan hasil dari program "df.columns", dapat diketahui bahwa data yang kuantitatif adalah
kuantitatif = ['Rank', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales']
for i in kuantitatif :
    print(i)
    print(f"min : {df[i].min()}")
    print(f"max : {df[i].max()}")
    print("Banyak data kosong = ", df[i].isnull().sum())
    print()
```

Gambar 4: Cara Menentukan Karakteristik Data Kuantitatif

Berdasarkan *source code*, informasi-informasi yang diberikan adalah nilai maksimal dari masing-masing atribut data, nilai minimum dari masing-masing atribut data, dan banyaknya data kosong. Hasilnya dapat dilihat pada gambar di bawah ini.

```
NA_Sales
min : 0.0
max : 41.49
Banyak data kosong = 0

EU_Sales
min : 0.0
max : 29.02
Banyak data kosong = 0

JP_Sales
min : 0.0
max : 10.22
Banyak data kosong = 0

Other_Sales
min : 0.0
max : 10.57
Banyak data kosong = 0

Global_Sales
min : 0.01
max : 82.74
Banyak data kosong = 0
```

Gambar 5: Hasil Run Source Code Penentuan Karakteristik Data Kuantitatif

2. Data Kategorikal

```
kategorikal = ['Name', 'Platform', 'Year', 'Genre', 'Publisher']
for i in kategorikal :
    print("=", i, "=")
    print(df[i].unique())
    print("Banyak data kosong = ", df[i].isnull().sum())
    print("")
```

Gambar 6: Cara Menentukan Karakteristik Data Kategorikal

Berdasarkan *source code*, informasi-informasi yang diberikan adalah elemen-elemen atribut data dan banyaknya data kosong. Hasil *run* dari *source code* dapat dilihat pada gambar di bawah ini.

```
* Rank *
[1, 2, 3, 4, 5, ..., 16596, 16597, 16598, 16599, 16600]
Length: 16598
Categories (16598, int64): [1, 2, 3, 4, ..., 16597, 16598, 16599, 16600]
Banyak data kosong = 0

* Name *
['Wii Sports', 'Super Mario Bros.', 'Mario Kart Wii', 'Wii Sports Resort', 'Pokemon Red/Pokemon Blue', ..., 'Chou Ezaru wa Akai Hana: Koi wa 1
Length: 11493
Categories (11493, object): ['98 Koshien', '.hack//G.U. Vol.1//Rebirth',
                             '.hack//G.U. Vol.2//Reminisce', '.hack//G.U. Vol.2//Reminisce (jp sales)', ...,
                             'thinkSMART: Chess for Kids', 'uDraw Studio', 'uDraw Studio: Instant Artist',
                             'wwe Smackdown vs. Raw 2006']
Banyak data kosong = 0

* Platform *
['Wii', 'NES', 'GB', 'DS', 'X360', ..., 'NG', 'TG16', '3DO', 'GG', 'PCFX']
Length: 31
Categories (31, object): ['2600', '3DO', '3DS', 'DC', ..., 'WiiU', 'X360', 'XB', 'XOne']
Banyak data kosong = 0

* Year *
[2006.0, 1985.0, 2008.0, 2009.0, 1996.0, ..., 1987.0, 1980.0, 1983.0, 2020.0, 2017.0]
Length: 40
Categories (39, float64): [1980.0, 1981.0, 1982.0, 1983.0, ..., 2015.0, 2016.0, 2017.0, 2020.0]
Banyak data kosong = 271

* Genre *
['Sports', 'Platform', 'Racing', 'Role-Playing', 'Puzzle', ..., 'Simulation', 'Action', 'Fighting', 'Adventure', 'Strategy']
Length: 12
Categories (12, object): ['Action', 'Adventure', 'Fighting', 'Misc', ..., 'Shooter', 'Simulation',
                          'Sports', 'Strategy']
Banyak data kosong = 0

* Publisher *
['Nintendo', 'Microsoft Game Studios', 'Take-Two Interactive', 'Sony Computer Entertainment', 'Activision', ..., 'Inti Creates', 'Takuyo', 'In
Length: 579
Categories (578, object): ['10TACLE Studios', '1C Company', '20th Century Fox Video Games', '2D Boy',
                          ..., 'imageepoch Inc.', 'inXile Entertainment', 'mixi, Inc',
                          'responDESIGN']
Banyak data kosong = 58
```

Gambar 7, 8: Hasil *Run Source Code* Penentuan Karakteristik Data Kuantitatif

2.3 Statistik Data

A. Sampel Data

Sampel data yang ditunjukkan adalah mencari lima data pada baris pertama, data terbesar dan terkecil berdasarkan atribut `Global_Sales`, data dari sebuah beberapa atribut, dan data dengan syarat tertentu. Berikut ini beberapa *source code* dan hasil *run source code* dari masing-masing sampel data.

1. Data Lima Baris Pertama

```
# sampel data beberapa data pada 5 baris pertama
df.head()
```

	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 9: Kode Program dan Hasil Data Lima Baris Pertama

2. Data Terbesar dan Terkecil Berdasarkan Data pada Atribut `Global_Sales`

```
# Data dengan nilai Global_Sales terbesar
imax = df["Global_Sales"].idxmax()
df[imax:imax+1]
```

	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

```
# Data dengan nilai Global_Sales terkecil
imin = df["Global_Sales"].idxmin()
df[imin:imin+1]
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
15980	15983	Turok	PC	2008.0	Action	Touchstone	0.0	0.01	0.0	0.0	0.01

Gambar 10: Kode Program dan Hasil Data Terbesar dan Terkecil Berdasarkan Data pada Atribut Global_Sales

3. Data dari Atribut Nama dan Platform

```
# Data dari baris ke-5 sampai dengan ke-10 pada kolom Name dan Platform
df.loc[4:10, 'Name' : 'Platform']
```

	Name	Platform
4	Pokemon Red/Pokemon Blue	GB
5	Tetris	GB
6	New Super Mario Bros.	DS
7	Wii Play	Wii
8	New Super Mario Bros. Wii	Wii
9	Duck Hunt	NES
10	Nintendogs	DS

Gambar 11: Kode Program dan Hasil Data dari Atribut Nama dan Platform

4. Data dengan Atribut Other_Sales di atas Lima

Data dengan kolom Other_Sales di atas 5

df.loc[df['Other_Sales'] >= 5]

	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
17	18	Grand Theft Auto: San Andreas	PS2	2004.0	Action	Take-Two Interactive	9.43	0.40	0.41	10.57	20.81
47	48	Gran Turismo 4	PS2	2004.0	Racing	Sony Computer Entertainment	3.01	0.01	1.10	7.53	11.66

Gambar 12: Kode Program dan Hasil Data dengan Atribut Other_Sales di atas Lima

B. Mengurutkan Data

Data yang dapat diurutkan pada dataset ini ialah atribut data kuantitatif, contohnya atribut Global_Sales. Data diurutkan mulai dari terkecil hingga terbesar atau sebaliknya. Berikut ini adalah *source code* dan hasil dari beberapa data-data yang diurutkan berdasarkan syarat-syarat tertentu.

1. Pengurutan Data Berdasarkan Global_Sales secara Menaik

Mengurutkan Global_Sales dengan menggunakan command untuk mengurutkan data secara menaik
df.sort_values(["Global_Sales"], ascending=[1])

Rank		Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01
16188	16191	Toro to Morimori	PS3	2009.0	Misc	Sony Computer Entertainment	0.00	0.00	0.01	0.00	0.01
16187	16190	Jewel Quest II	PC	2007.0	Puzzle	Avanquest	0.00	0.01	0.00	0.00	0.01
16186	16189	BattleForge	PC	2009.0	Strategy	Electronic Arts	0.00	0.01	0.00	0.00	0.01
16185	16188	Tantei Jinguuji Saburo: Hai to Diamond	PSP	2009.0	Adventure	Arc System Works	0.00	0.00	0.01	0.00	0.01
...
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74

16598 rows x 11 columns

Gambar 13: Kode Program dan Hasil Pengurutan Data Berdasarkan Global_Sales secara Menaik

2. Pengurutan Data Berdasarkan Global_Sales secara Menurun

Mengurutkan Global_Sales dengan menggunakan command untuk mengurutkan data secara menurun
df.sort_values(["Global_Sales"], ascending=[0])

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
...
16186	16189	BattleForge	PC	2009.0	Strategy	Electronic Arts	0.00	0.01	0.00	0.00	0.01
16187	16190	Jewel Quest II	PC	2007.0	Puzzle	Avanquest	0.00	0.01	0.00	0.00	0.01
16188	16191	Toro to Morimori	PS3	2009.0	Misc	Sony Computer Entertainment	0.00	0.00	0.01	0.00	0.01
16189	16192	Sonic & All-Stars Racing Transformed	PC	2013.0	Racing	Sega	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows x 11 columns

Gambar 14: Kode Program dan Hasil Pengurutan Data Berdasarkan Global_Sales secara Menurun

C. Statistik Setiap Atribut

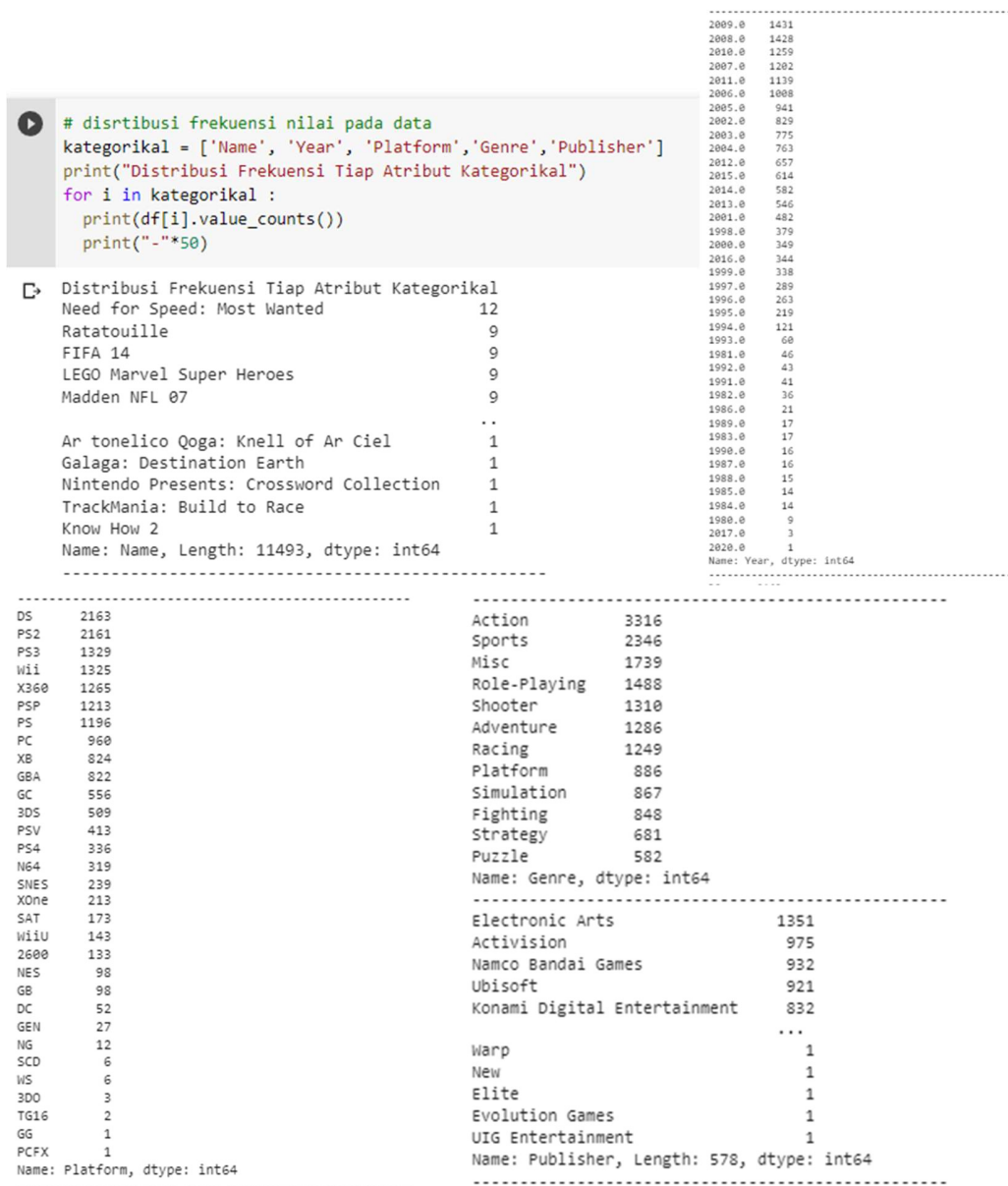
1. Rata-Rata, Standar Deviasi, Persentil, dan Eksentrum Data-Data Kuantitatif

```
# rata-rata dan standar deviasi, percentile (10%,20%,50%,75%,90%), dan ekstremum (nilai maksimum dan minimum)
print('rata-rata, standar deviasi, persentil, dan eksentrum')
print('\n')
df_baru = df[df.columns[6:]]
print(df_baru.describe(percentiles=[.1,.25,.5,.75,.9]))
print("-"*75)
```

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16598.000000	16598.000000	16598.000000	16598.000000	16598.000000
mean	0.263768	0.145951	0.077808	0.048063	0.537441
std	0.816594	0.505121	0.309293	0.188588	1.555028
min	0.000000	0.000000	0.000000	0.000000	0.010000
10%	0.000000	0.000000	0.000000	0.000000	0.020000
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
90%	0.610000	0.350000	0.180000	0.110000	1.210000
max	41.490000	29.020000	10.220000	10.570000	82.740000

Gambar 15, 16: Kode Program dan Hasil Data Rata-Rata, Standar Deviasi, Persentil, dan Eksentrum Data-Data Kuantitatif

2. Distribusi Frekuensi

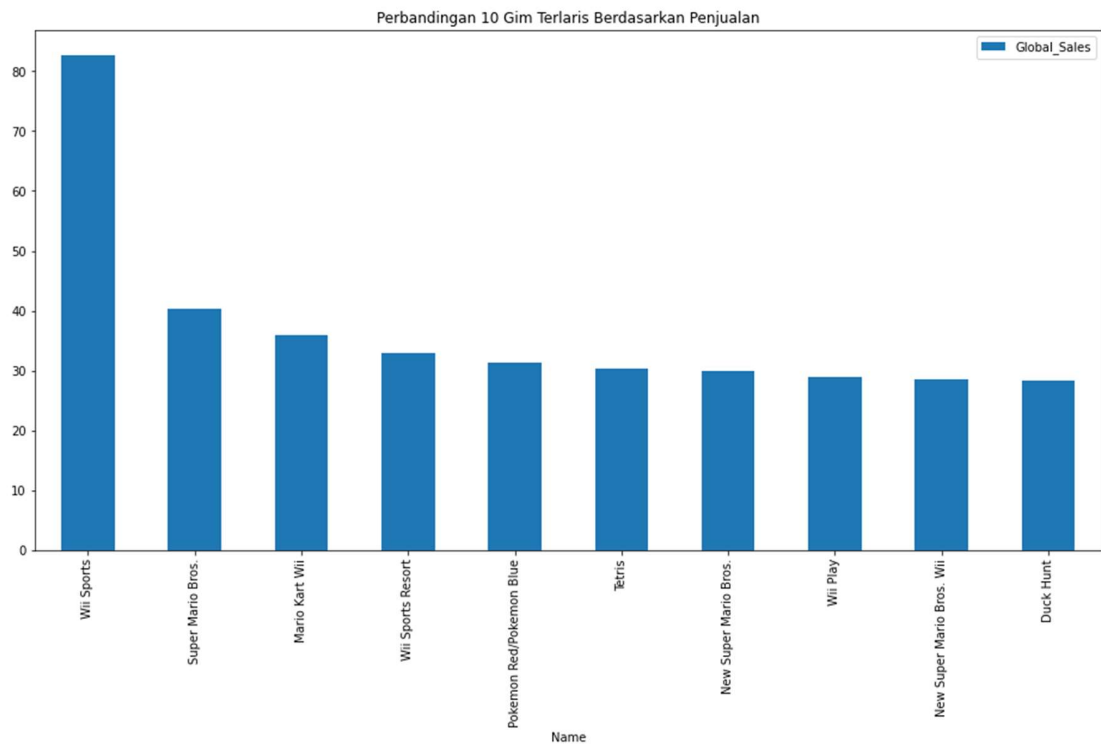


Gambar 17, 18, 19, 20: Kode Program dan Hasil Data Distribusi Frekuensi

2.4 Visualisasi Data

A. Bar Chart Sepuluh Gim Terlaris Berdasarkan Penjualan Global

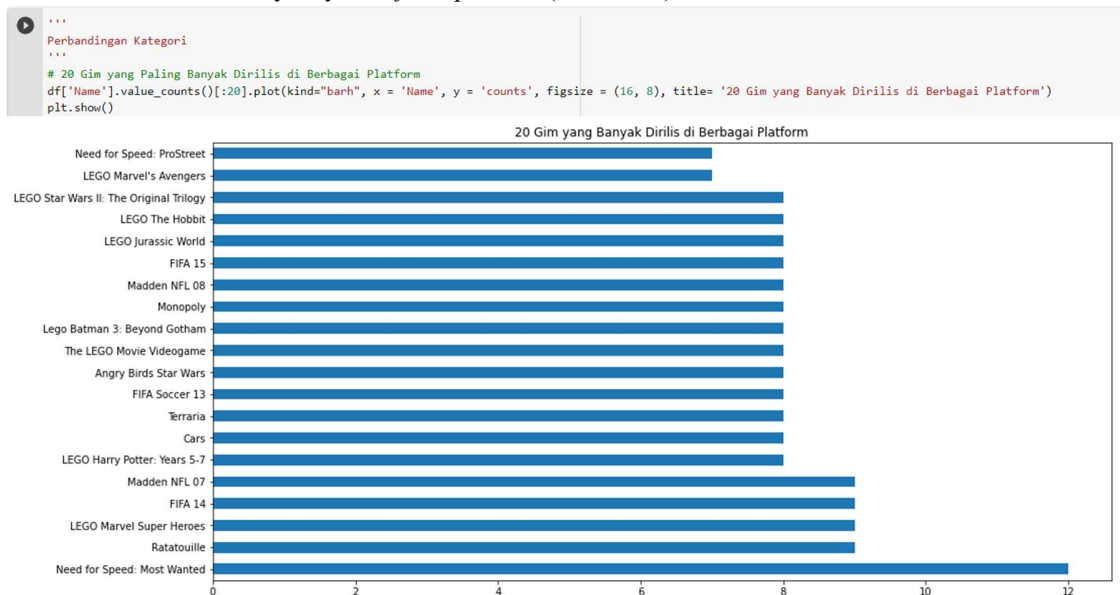




Gambar 21, 22: Visualisasi Sepuluh Gim Terlaris Berdasarkan Penjualan Global_Sales

Berdasarkan grafik, kami dapat memperoleh informasi bahwa Wii Sport merupakan gim yang memiliki nilai penjualan global paling tinggi. Selain itu, sepuluh gim teratas berdasarkan penjualan global didominasi oleh gim-gim yang platform asalnya adalah Wii. Visualisasi ini termasuk visualisasi perbandingan kategori.

B. Bar Horizontal Chart Banyaknya Platform per Gim (20 Teratas)

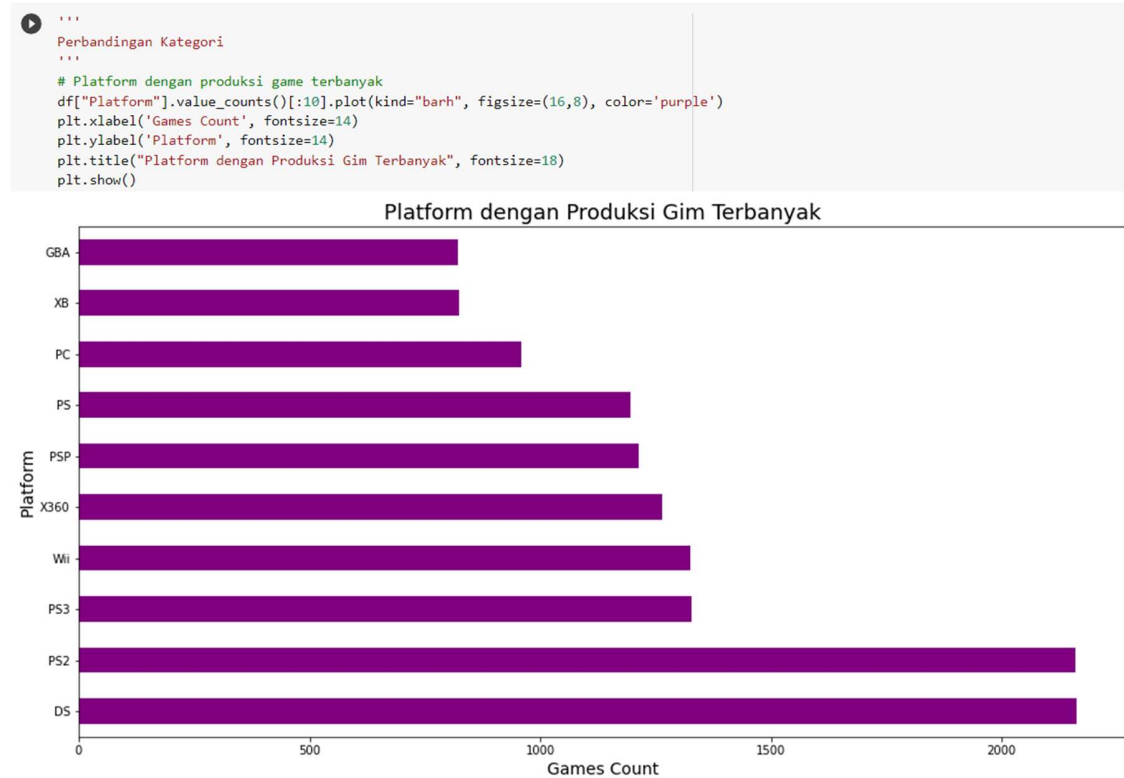


Gambar 23, 24: Visualisasi 20 Gim Terbanyak Dirilis Berdasarkan Platform

Berdasarkan grafik, kami memperoleh informasi bahwa gim-gim yang paling banyak dirilis didominasi oleh gim-gim LEGO lalu diikuti dengan FIFA. Dari 20 gim yang ditampilkan, gim dengan jumlah

perilisan di berbagai platform terbanyak adalah Need for Speed: Most Wanted. Visualisasi ini termasuk ke dalam visualisasi perbandingan kategori.

C. Bar Horizontal Chart Platform dengan Produksi Gim Terbanyak

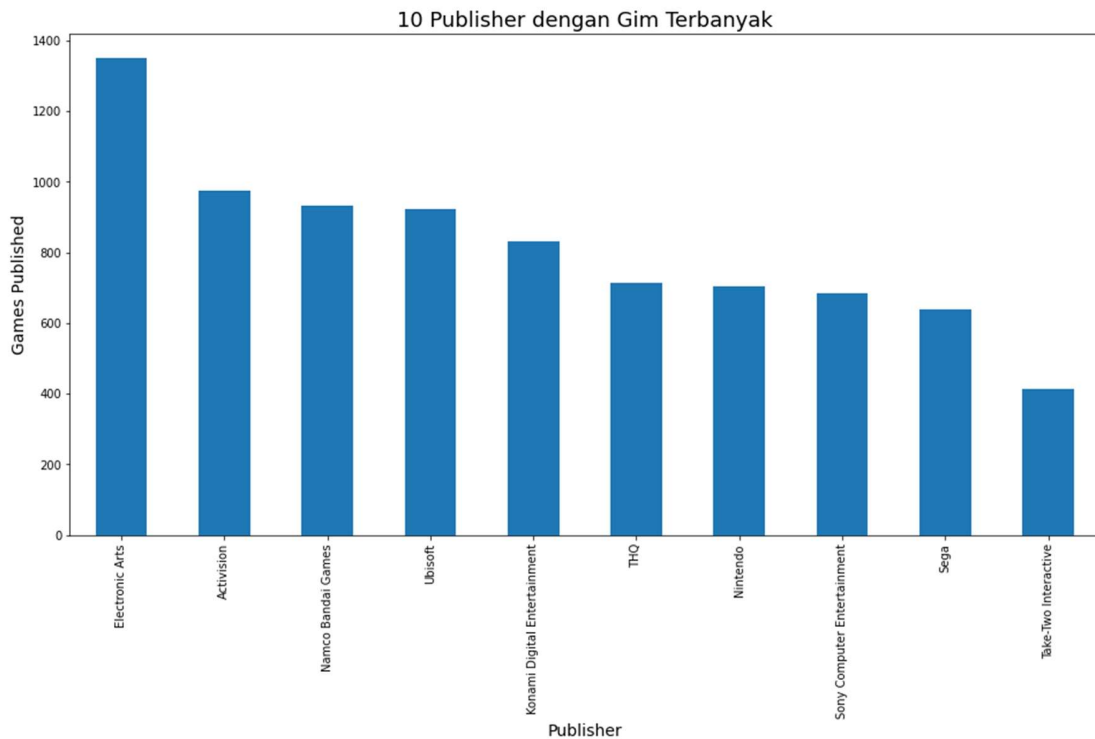


Gambar 25, 26: Visualisasi *Platform* dengan Produksi Gim Terbanyak

Berdasarkan grafik di atas, kami memperoleh informasi bahwa *platform* yang memproduksi gim terbanyak adalah DS lalu diikuti oleh PS2. Visualisasi ini termasuk ke dalam visualisasi perbandingan kategori.

D. Bar Chart Sepuluh Publisher dengan Produksi Gim Terbanyak

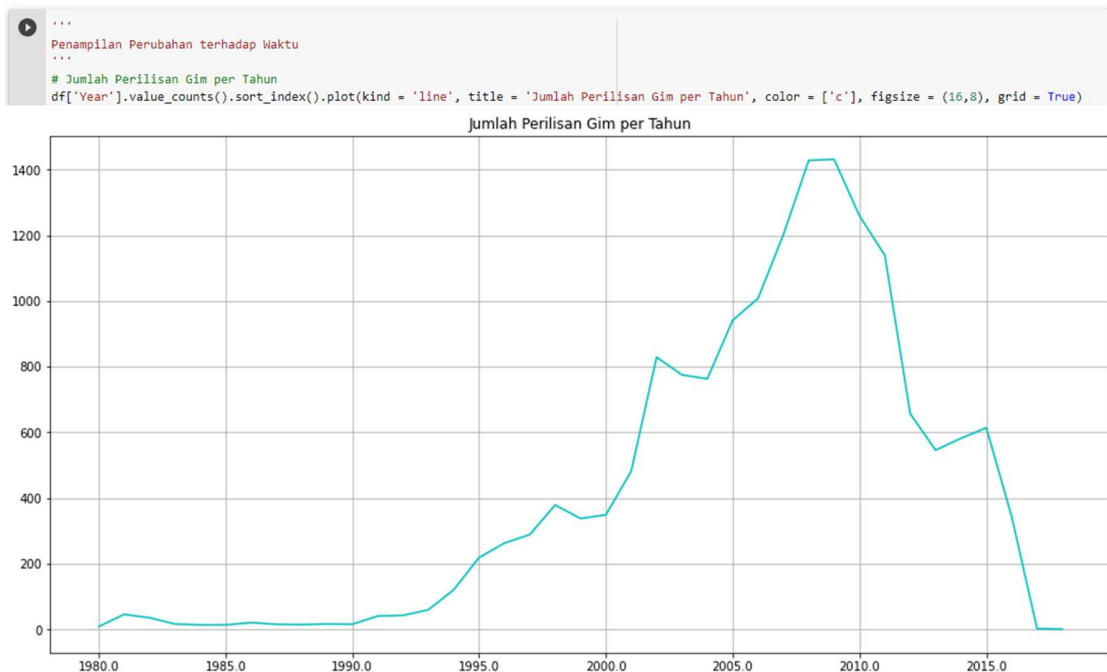




Gambar 27, 28: Visualisasi Sepuluh Publisher dengan Gim Terbanyak

Berdasarkan grafik, informasi yang kami dapatkan adalah Electronic Arts menjadi *publisher* yang merilis gim terbanyak. Visualisasi ini termasuk ke dalam visualisasi perbandingan kategori.

E. Line Chart Jumlah Perilisan Gim per tahun

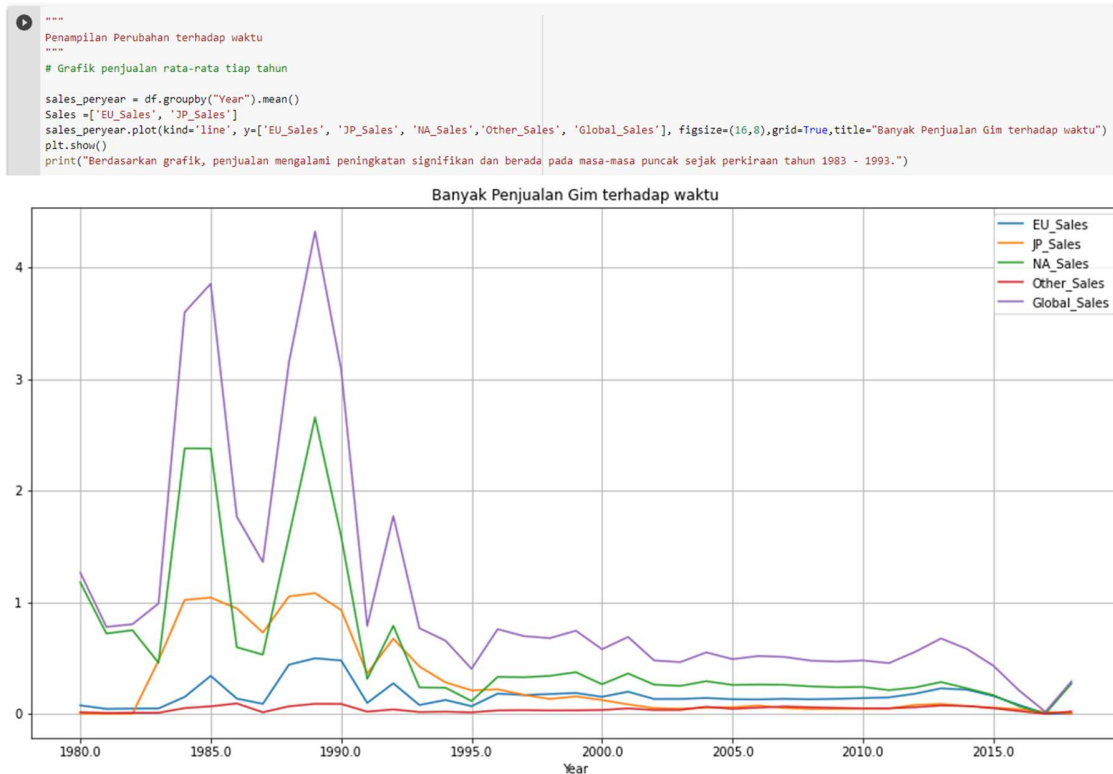


Gambar 29, 30: Visualisasi Jumlah Perilisan Gim per Tahun

Berdasarkan grafik, kami memperoleh informasi bahwa jumlah perilisan gim mengalami peningkatan yang signifikan dari 1990-an sampai tahun 2008 atau 2009. Puncaknya berada di tahun 2008

atau 2009 dengan jumlah gimnya mencapai lebih dari 1400 gim. Namun, jumlah gim langsung mengalami penurunan mulai tahun 2008 atau 2009. Visualisasi ini termasuk ke dalam visualisasi perubahan terhadap waktu.

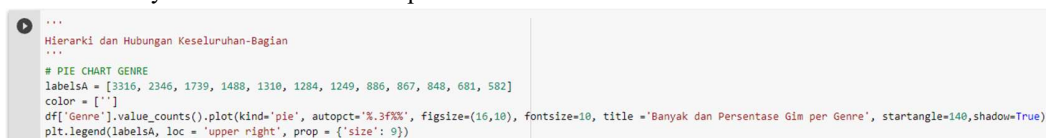
F. Line Chart Penjualan Rata-rata per Tahun

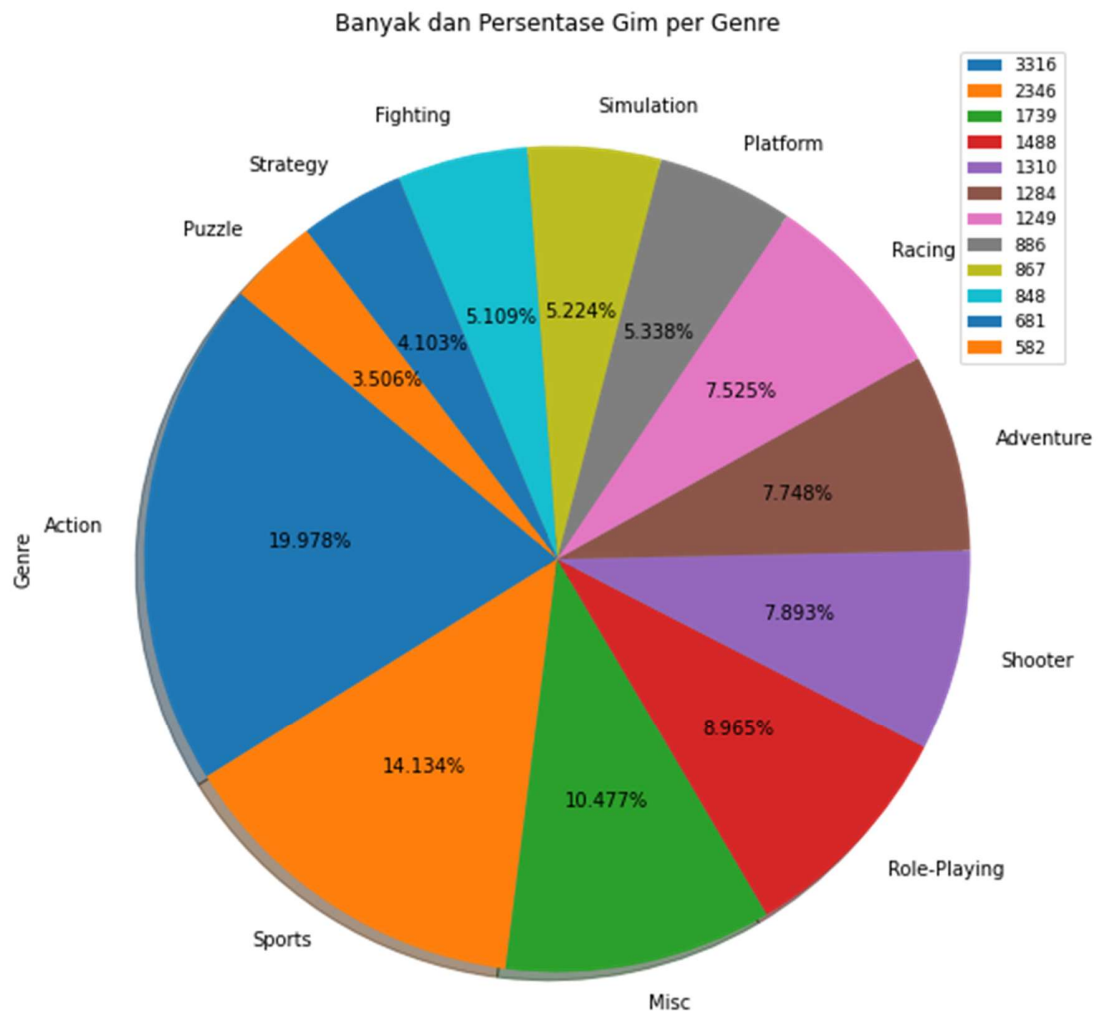


Gambar 31, 32: Visualisasi Banyak Penjualan Gim terhadap Waktu

Berdasarkan grafik, menurut interval tahun yang diperkirakan, kami memperoleh informasi bahwa penjualan global dan di EU, JP, NA, dan Other mengalami peningkatan yang signifikan dan berada di saat-saat yang tertinggi (masa puncak) pada tahun 1983 - 1993. Visualisasi ini termasuk ke dalam visualisasi perubahan terhadap waktu.

G. Pie Chart Banyak dan Persentase Gim per Genre



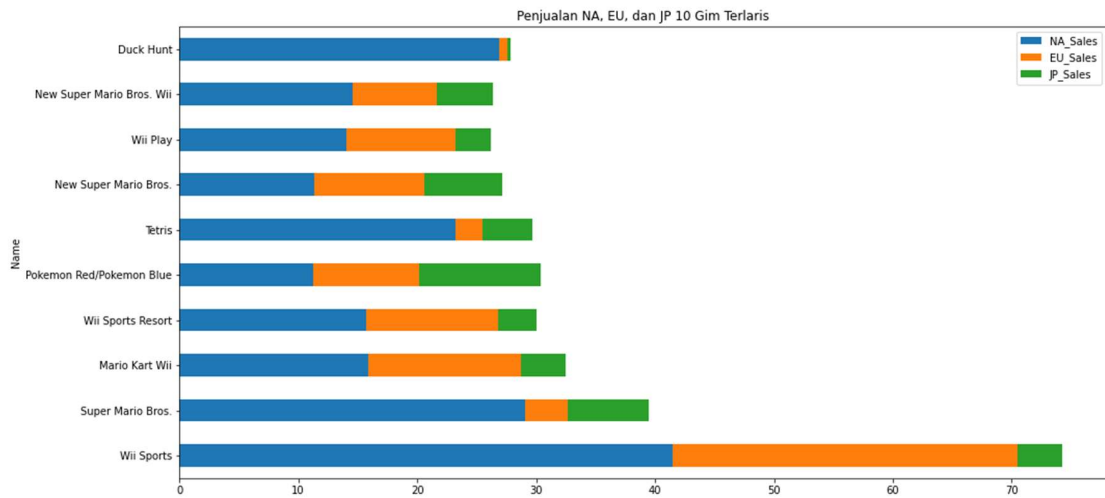


Gambar 33, 34: Visualisasi Banyak dan Persentase Gim per Genre

Berdasarkan grafik, kami memperoleh informasi bahwa ada 12 genre gim. Genre gim yang memiliki jumlah gim terbanyak berdasarkan genre adalah *action* dengan jumlah gimnya mencapai 3316 gim, sedangkan jumlah gim paling sedikit berdasarkan genre adalah *puzzle* dengan jumlah gimnya mencapai 582 gim. Visualisasi ini termasuk ke dalam visualisasi hierarki dan hubungan keseluruhan-bagian.

H. Stacked Bar Horizontal Penjualan NA, EU, dan JP Sepuluh Gim Terlaris

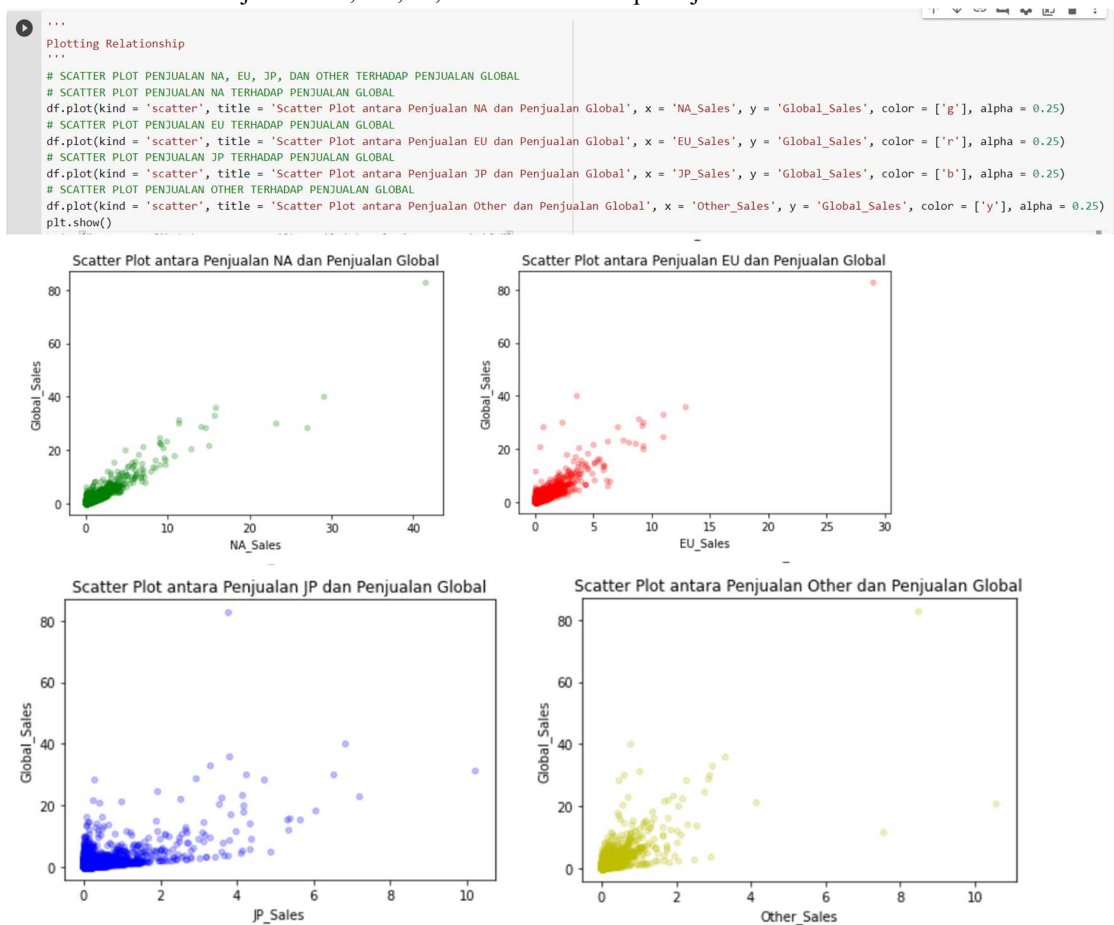
```
# STACKED BAR CHART 10 GIM TERLARIS TERHADAP PENJUALAN DI NA, EU, DAN JP
df[:10].plot(kind = 'barh', x = 'Name', y = ['NA_Sales', 'EU_Sales', 'JP_Sales'], title = "Penjualan NA, EU, dan JP 10 Gim Terlaris", stacked = 1, figsize=(16,8))
plt.show()
```

Gambar 35, 36: Visualisasi Penjualan NA, EU, dan JP Sepuluh Gim Terlaris

Berdasarkan grafik, dapat diketahui bahwa penjualan NA lebih dominan daripada penjualan EU dan JP. Khusus untuk Tetris, Pokemon Red/Pokemon Blue, dan Super Mario Bros., penjualan JP lebih besar daripada penjualan EU. Selain ketiga game tersebut, semua gim memiliki penjualan EU yang lebih besar daripada penjualan JP. Visualisasi ini termasuk ke dalam visualisasi hierarki dan hubungan keseluruhan-bagian.

I. Scatter Plot antara Penjualan NA, EU, JP, dan Other terhadap Penjualan Global

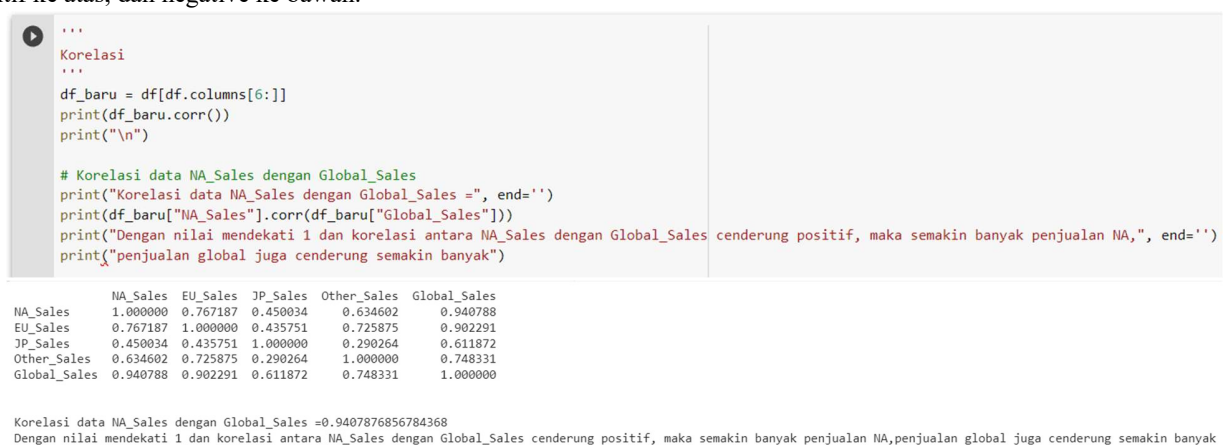


Gambar 37, 38, 39, 40, 41: Visualisasi Scatter Plot Penjualan NA, EU, dan Other terhadap Penjualan Global

Berdasarkan keempat grafik, kami memperoleh informasi bahwa semua grafik hubungan menampilkan nilai korelasi yang positif. Grafik hubungan yang memiliki nilai korelasi positif terbesar adalah grafik hubungan antara penjualan NA terhadap penjualan global karena titik-titik terdistribusi rapat terhadap suatu garis dengan gradien positif, sedangkan grafik hubungan yang memiliki nilai korelasi positif terkecil adalah grafik penjualan hubungan antara penjualan JP dan penjualan global karena titik-titiknya cenderung terdistribusi secara menyebar terhadap garis yang memiliki gradien positif. Keempat visualisasi ini termasuk ke dalam *plotting relationship*.

2.5 Korelasi Data

Data yang akan divisualisasikan adalah korelasi antar semua atribut kuantitatif yang ada di dataset. Angka menunjukkan hubungan linear antar variabel, semakin menuju satu (1) semakin berkorelasi, semakin menuju nol (0) semakin tidak berkorelasi. Tanda positif (+) dan negatif (-) menunjukkan apakah linear menuju keatas atau kebawah, positif ke atas, dan negative ke bawah.



Gambar 42, 43: Korelasi Data

Dengan nilai mendekati satu dan korelasi antara NA_Sales dengan Global_Sales cenderung positif, maka semakin banyak penjualan NA, penjualan global juga cenderung semakin banyak. Begitu juga dengan korelasi antara EU_Sales, JP_Sales, dan Other_Sales dengan Global_Sales.

2.6 Data Cleansing

Data ini merupakan data yang kotor sehingga perlu dilakukan pembersihan dan perbaikan agar keseluruhan data dapat dipakai dalam analisis data. *Data cleansing* dilakukan sebelum analisis data dimulai. Dalam *data cleansing* data ini, kami membagi prosesnya menjadi dua proses, yakni

A. Data Salah Tempat



Gambar 44: Pengecekan Data “Salah Tempat”

Pada proses ini, kami melakukan pengecekan terhadap data-data yang “salah tempat”. Kami melakukan pengecekan dengan melihat isi data dari masing-masing atribut data yang merupakan hasil dari menjalankan perintah “df[i].unique()”. Isi data tersebut dapat dilihat di Gambar 7 dan Gambar 8.

```
# MISPLACED DATA 2
# Pengecekan data berdasarkan kesalahan di kolom Platform
df.loc[(df['Platform'] == '2007') | (df['Platform'] == '2010')]
# Jika hasilnya kosong, data sudah dilakukan perbaikan.
```

index	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
11593	11595	Boku no Natsuyasumi 3: Hokkoku Hen: Chiisana Boku no Dai Sougen??PS3	2007	Adventure	Sony Computer Entertainment	0	0.0	0.08	0.0	0.08	NaN
13538	13540	B's-LOG Party??PSP	2010	Adventure	Idea Factory	0	0.0	0.04	0.0	0.04	NaN

Gambar 45: Pengecekan Baris-Baris Data “Salah Tempat”

Kemudian, kami mem-*print out* baris data yang memiliki kesalahan data dengan *source code* dan hasil *run* dari *source code*. Baris-baris data yang muncul dapat dilihat di Gambar 45. Kami menjadi tahu bahwa kedua baris data masing-masing memiliki data yang “bergeser” ke kiri, seperti atribut “Platform” baris pertama, yakni “2007”, yang seharusnya terletak di atribut “Year”.

```
# MISPLACED DATA 3
print("Berdasarkan kolom kode sebelumnya, dapat diketahui bahwa terdapat dua baris yang salah.")
print("Agar data dapat dipakai, kedua baris data perlu diperbaiki.")
# Perbaikan data index 11593
df['Name'].replace("Boku no Natsuyasumi 3: Hokkoku Hen: Chiisana Boku no Dai Sougen??PS3", "Boku no Natsuya: Hokkoku Hen: Chiisana Boku no Dai Sougen", inplace = True)
df['Platform'].replace("2007", "PS3", inplace = True)
df['Year'].replace("Adventure", "2007", inplace = True)
df['Genre'].replace("Sony Computer Entertainment", "Adventure", inplace = True)
df['Publisher'].replace("0", "Sony Computer Entertainment", inplace = True)
# Hasil perbaikan data index 11593
print(df.iloc[11593])

# MISPLACED DATA 4
# Perbaikan data index 13538
df['Name'].replace("B's-LOG Party??PSP", "B's-LOG Party", inplace = True)
df['Platform'].replace("2010", "PSP", inplace = True)
df['Year'].replace("Adventure", "2010", inplace = True)
df['Genre'].replace("Idea Factory", "Adventure", inplace = True)
df['Publisher'].replace("0", "Idea Factory", inplace = True)
# Hasil perbaikan data index 13538
print(df.iloc[13538])
```

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
11804	Bomberman Max 2: Red Advance	GBA	2002	Puzzle	Vivendi Games	0.05	0.02	0.0	0.0	0.07
13783	Katekyoo Hitman Reborn! Battle Arena 2 - Spiri...	PSP	2009	Fighting	Marvelous Interactive	0.0	0.0	0.04	0.0	0.04

Gambar 46, 47: Perbaikan Baris-Baris Data “Salah Tempat”

Selanjutnya, kami melakukan perbaikan dari kedua baris data yang memiliki kesalahan data. Kami melakukan perbaikan berdasarkan Gambar 45. Di atribut “Name” dari masing-masing baris data, dapat terlihat sebuah *string* yang terletak di nama *platform* yang ditulis setelah “??”, yakni “PS3” di baris pertama dan “PSP” di baris kedua. “PS3” dan “PSP” merupakan data kategorikal yang berada di atribut “Platform” dan buktinya dapat diketahui dengan “df[‘Platform’].unique()” yang dapat dilihat di Gambar 7 dan Gambar 8. Kami melakukan perbaikan data dengan menggunakan *method replace* per atribut data sehingga hasilnya

dapat dilihat di Gambar 46 untuk baris data pertama yang salah dan Gambar 47 untuk baris data kedua yang salah.

B. Perbaikan Jenis Data



```
# DATA TYPE FIXING
# Mengubah beberapa jenis data yang awalnya objek menjadi kategorikal karena telah dilakukan perbaikan dan pembersihan data.
df['Name'] = df['Name'].astype('category')
df['Platform'] = df['Platform'].astype('category')
df['Year'] = df['Year'].astype('category')
df['Genre'] = df['Genre'].astype('category')
df['Publisher'] = df['Publisher'].astype('category')
```

Gambar 48: Kode Program Perbaikan Jenis Data

Pada bagian ini, kami melakukan perbaikan terhadap tipe-tipe atribut data kategorikal. Sebelum melakukan perbaikan, kami melihat bahwa atribut data-data kategorikal memiliki tipe data *object* yang kami definisikan sebagai tipe data yang memiliki elemen-elemen data campuran, seperti campuran antara string dengan bilangan. Hal tersebut tentu tidaklah benar karena setiap atribut data merepresentasikan satu jenis data saja, yakni kategorikal atau *category*. Berdasarkan alasan tersebut, kami memutuskan untuk melakukan perbaikan dengan mengubah tipe-tipe atribut data kategorikal yang sebelumnya *object* menjadi tipe data *category* dengan kode program yang terlihat pada Gambar 7 dan Gambar 8.

Sebenarnya, *dataset* yang kami pilih untuk dianalisis memiliki kekosongan data, yakni atribut “Year” sebanyak 271 data dan atribut “Publisher” sebanyak 58 data, setelah dilakukan perbaikan terhadap *misplaced data* dan tipe-tipe data. Jika kedua jenis kekosongan itu digabung, akan diperoleh banyaknya kekosongan data adalah 307 baris data. Perlu diketahui bahwa angka total kekosongan data diperoleh itu tidak sesuai dengan $271 + 58$ karena adanya irisan antara kekosongan di atribut “Year” dengan kekosongan di atribut “Publisher”. Berdasarkan hal tersebut, kami dapat menghitung galat dengan rumus berikut.

$$\phi = [\Sigma(n) \div \Sigma(N)] \times 100\%$$

$$\phi = [307 \div 16598] \times 100\%$$

$$\phi = 1.85\%$$

dengan variabel “n” sebagai banyaknya baris data yang kosong, variabel “N” sebagai banyaknya baris data, dan ϕ sebagai galat. Kami dapat memperoleh galat sebesar 1.85%. Berdasarkan hal tersebut, dapat diasumsikan bahwa analisis data yang kami lakukan jika tanpa dilakukan perbaikan terhadap data-data kosong memiliki keroran sebesar 1.85% atau kebenaran analisis kami sebesar 98.15% terhadap analisis data yang dilakukan jika dilakukan perbaikan terhadap data-data kosong. Karena tidak terlalu besarnya galat ($< 5\%$), kami memutuskan untuk membiarkan kekosongan data karena tidak memengaruhi secara signifikan analisis data yang kami lakukan.

BAB III KESIMPULAN

3.1 Kesimpulan

Pada pengerjaan tugas analisis data ini, dilakukan analisis terhadap data penjualan video gim. Sebelum dilakukan analisis, kami melakukan *data cleansing* yang dibagi menjadi dua proses, yakni perbaikan data “salah tempat” dan perbaikan jenis data. Hasil *data cleansing* dapat dilihat dari Gambar 44 - Gambar 48. Analisis data yang pertama kami lakukan adalah deskripsi data dan file yang terdiri atas tujuan analisis, informasi data, dan jenis data setiap atribut. Hasil analisis pertama dapat dilihat di Tabel 1 serta Gambar 1, Gambar 2, dan Gambar 3. Analisis data kedua adalah karakteristik data yang bertujuan untuk menjelaskan data kategorik dan data kuantitatif dari data kami. Hasil analisis kedua dapat dilihat dari Gambar 4 - Gambar 8. Analisis data ketiga adalah statistik data yang terdiri atas sampel data, pengurutan data, dan statistik data serta bertujuan untuk menjelaskan deskripsi statistik dari setiap atribut data. Hasil analisis ketiga dapat dilihat dari Gambar 9 - Gambar 20. Analisis data selanjutnya adalah visualisasi data yang dapat dibagi menjadi perbandingan kategori, perubahan terhadap waktu, hierarki dan hubungan keseluruhan-bagian, dan *plotting relationship*. Hasil analisis ini dapat dilihat dari Gambar 21 - Gambar 41. Terakhir, kami menganalisis korelasi antar data. Hasil analisis ini dapat dilihat di Gambar 42 dan Gambar 43.

DAFTAR PUSTAKA

- Kaggle. (2022, July 22). *Sales Of Video Games*. Www.kaggle.com.
<https://www.kaggle.com/datasets/arslanali4343/sales-of-video-games?resource=download>
- pandas. (2019). *pandas: powerful Python data analysis toolkit — pandas 0.25.3 documentation*. Pydata.org.
<https://pandas.pydata.org/pandas-docs/stable/index.html>

LAMPIRAN

LAMPIRAN 1 TABEL KONTRIBUSI

No.	Nama	NIM	Bagian Tubes
1.	Jonathan Wiguna	19622014	<ul style="list-style-type: none">- Deskripsi dan Karakteristik Data- Membuat PPT- Edit Video Presentasi- Merekomendasikan Dataset
2.	Melati Anggraini	19622144	<ul style="list-style-type: none">- Visualisasi Data- Membuat Laporan Bab I dan Bab II yang Berkaitan dengan Visualisasi Data- Membuat PPT- Menyunting Program (Kesalahan-Kesalahan Pemrograman)
3.	Kean Malik Aji Santoso	16522034	<ul style="list-style-type: none">- Visualisasi Data- Data Cleansing- Menyunting Laporan (Kebahasaan, Estetika, dan Kesalahan-Kesalahan Data)- Membuat Laporan (Bab I dan Bab II Mengenai Data Cleansing dan Beberapa Visualisasi Data)
4.	Kayla Dyara	19622224	<ul style="list-style-type: none">- Statistik dan Korelasi- Membuat PPT- Membuat Laporan Bab II (Statistik dan Sampel Data)

LAMPIRAN 2 TAUTAN SOURCE CODE

<https://colab.research.google.com/drive/1U8XJWyuqsY6ZjKPX7YzCGrMoeuqwgmwR?usp=sharing>

LAMPIRAN 3 TAUTAN VIDEO PRESENTASI

<https://youtu.be/OOt14w5BDqc>