```
import pandas as pd
import numpy as np

#Load Dataset
games=pd.read_csv('games.csv')
vgsales=pd.read_csv('vgsales.csv')

#Overview
print("games.csv ----")
print(games.shape)
print(games.head())

print("vgsales.csv ----")
print(vgsales.shape)
print(vgsales.head())
```

```
games.csv ----
(1512, 14)
   Unnamed: 0                                Title  Release Date  \
0           0                           Elden Ring  Feb 25, 2022
1           1                                Hades  Dec 10, 2019
2           2  The Legend of Zelda: Breath of the Wild  Mar 03, 2017
3           3                            Undertale  Sep 15, 2015
4           4                        Hollow Knight  Feb 24, 2017

                                          Team  Rating Times Listed  \
0    ['Bandai Namco Entertainment', 'FromSoftware']     4.5         3.9K
1                          ['Supergiant Games']     4.3         2.9K
2  ['Nintendo', 'Nintendo EPD Production Group No...     4.4         4.3K
3                          ['tobyfox', '8-4']     4.2         3.5K
4                             ['Team Cherry']     4.4           3K

  Number of Reviews                                        Genres  \
0             3.9K                            ['Adventure', 'RPG']
1             2.9K          ['Adventure', 'Brawler', 'Indie', 'RPG']
2             4.3K                            ['Adventure', 'RPG']
3             3.5K  ['Adventure', 'Indie', 'RPG', 'Turn Based Stra...
4               3K               ['Adventure', 'Indie', 'Platform']

                                          Summary  \
0  Elden Ring is a fantasy, action and open world...
1  A rogue-lite hack and slash dungeon crawler in...
2  The Legend of Zelda: Breath of the Wild is the...
3  A small child falls into the Underground, wher...
4  A 2D metroidvania with an emphasis on close co...

                                          Reviews Plays Playing Backlogs  \
0  ["The first playthrough of elden ring is one o...   17K    3.8K     4.6K
1  ['convinced this is a roguelike for people who...   21K    3.2K     6.3K
2  ['This game is the game (that is not CS:GO) th...   30K    2.5K       5K
3  ['soundtrack is tied for #1 with nier automata...   28K     679     4.9K
4  ["this games worldbuilding is incredible, with...   21K    2.4K     8.3K

  Wishlist
0     4.8K
1     3.6K
2     2.6K
3     1.8K
4     2.3K
vgsales.csv ----
(16598, 11)
   Rank                     Name Platform    Year         Genre Publisher  \
0     1               Wii Sports      Wii  2006.0        Sports  Nintendo
1     2         Super Mario Bros.      NES  1985.0      Platform  Nintendo
2     3            Mario Kart Wii      Wii  2008.0        Racing  Nintendo
3     4         Wii Sports Resort      Wii  2009.0        Sports  Nintendo
4     5  Pokemon Red/Pokemon Blue       GB  1996.0  Role-Playing  Nintendo

   NA_Sales  EU_Sales  JP_Sales  Other_Sales  Global_Sales
0     41.49     29.02      3.77         8.46         82.74
1     29.08      3.58      6.81         0.77         40.24
2     15.85     12.88      3.79         3.31         35.82
3     15.75     11.01      3.28         2.96         33.00
4     11.27      8.89     10.22         1.00         31.37
```

```
# Basic info
print("\nGames Info:")
games.info()
print("\nMissing in Games:")
print(games.isnull().sum())
```

```
print("\nVGSales Info:")
vgsales.info()
print("\nMissing in VGSales:")
print(vgsales.isnull().sum())

# Check duplicates
print("\nGames duplicates:", games.duplicated().sum())
print("VGSales duplicates:", vgsales.duplicated().sum())
```

```
 12  Backlogs           1512 non-null   object
 13  Wishlist           1512 non-null   object
dtypes: float64(1), int64(1), object(12)
memory usage: 165.5+ KB

Missing in Games:
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

VGSales Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16598 non-null  int64
 1   Name          16598 non-null  object
 2   Platform      16598 non-null  object
 3   Year          16327 non-null  float64
 4   Genre         16598 non-null  object
 5   Publisher     16540 non-null  object
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB

Missing in VGSales:
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

Games duplicates: 0
VGSales duplicates: 0
```

```
# Replace empty strings, whitespace, and "N/A" with NaN
vgsales["Publisher"] = vgsales["Publisher"].replace(r'^\s*$', np.nan, regex=True)
vgsales["Publisher"] = vgsales["Publisher"].replace("N/A", np.nan)
```

```
print(vgsales["Publisher"].isnull().sum())
```

```
58
```

```
# Keep missing Year as NaN, but convert valid ones to int
vgsales["Year"] = pd.to_numeric(vgsales["Year"], errors="coerce").astype("Int64")
```

```
# --- FIX FOR MULTILINE TEXT & SPECIAL CHARACTERS ---

# Replace line breaks and extra spaces in text columns
games["Summary"] = games["Summary"].astype(str).str.replace(r"[\r\n]+", " ", regex=True)
games["Title"] = games["Title"].astype(str).str.replace(r"[\r\n]+", " ", regex=True)
games["Team"] = games["Team"].astype(str).str.replace(r"[\r\n]+", " ", regex=True)

# Remove weird non-UTF8 characters (like emojis)
games["Summary"] = games["Summary"].str.encode("ascii", "ignore").str.decode("ascii")
games["Title"] = games["Title"].str.encode("ascii", "ignore").str.decode("ascii")
games["Team"] = games["Team"].str.encode("ascii", "ignore").str.decode("ascii")
```

```
# Rename the first column (currently no name) to "S.No."
games.rename(columns={games.columns[0]: "S.No."}, inplace=True)
```

```
# Handle numeric columns
num_cols = ["Rating", "Times Listed", "Number of Reviews",
            "Plays", "Playing", "Backlogs", "Wishlist"]

for col in num_cols:
    games[col] = pd.to_numeric(games[col], errors="coerce")
    games[col] = games[col].fillna(games[col].median())

# Clean Genres → take only the first if multiple
games["Genres"] = games["Genres"].astype(str).str.split(",").str[0].str.strip().str.title()

# Normalize text columns
text_cols = ["Title", "Team", "Summary"]
for col in text_cols:
    games[col] = games[col].astype(str).str.strip()
```

```
# Remove brackets and quotes from Team and Genres columns
games["Team"] = games["Team"].astype(str).str.replace(r"[\[\]']", "", regex=True)
games["Genres"] = games["Genres"].astype(str).str.replace(r"[\[\]']", "", regex=True)
games["Reviews"] = games["Reviews"].astype(str).str.replace(r'["\[\]\']', '', regex=True)

# Also remove extra spaces after commas (for neatness)
games["Team"] = games["Team"].str.replace(", ", ",").str.replace(",", ", ")
games["Genres"] = games["Genres"].str.replace(", ", ",").str.replace(",", ", ")
games["Reviews"] = games["Reviews"].str.replace(r"[\r\n]+", " ", regex=True).str.strip()

# Display first 5 rows to confirm
games[["Team", "Genres","Reviews"]].head()
```

	Team	Genres	Reviews
0	Bandai Namco Entertainment, FromSoftware	Adventure	The first playthrough of elden ring is one of ...
1	Supergiant Games	Adventure	convinced this is a roguelike for people who d...
2	Nintendo, Nintendo EPD Production Group No. 3	Adventure	This game is the game (that is not CS:GO) that...
3	tobyfox, 8-4	Adventure	soundtrack is tied for #1 with nier automata. ...
4	Team Cherry	Adventure	this games worldbuilding is incredible, with i...

```
#Release Date Format
games['Release Date'] = pd.to_datetime(games['Release Date'], errors='coerce').dt.strftime('%Y-%m-%d')
```

```
print("First 10 rows of games.csv")
display(games.head(10))

print("First 10 rows of vgsales.csv")
display(vgsales.head(10))
```

First 10 rows of games.csv

	S.No.	Title	Release Date	Team	Rating	Times Listed	Number of Reviews	Genres	Summary	Reviews	Plays	Playing	Backlogs
0	0	Elden Ring	2022-02-25	Bandai Namco Entertainment, FromSoftware	4.5	424.5	424.5	Adventure	Elden Ring is a fantasy, action and open world...	The first playthrough of elden ring is one of ...	442.0	100.0	461.0
1	1	Hades	2019-12-10	Supergiant Games	4.3	424.5	424.5	Adventure	A rogue-lite hack and slash dungeon crawler in...	convinced this is a roguelike for people who d...	442.0	100.0	461.0
2	2	The Legend of Zelda: Breath of the Wild	2017-03-03	Nintendo, Nintendo EPD Production Group No. 3	4.4	424.5	424.5	Adventure	The Legend of Zelda: Breath of the Wild is the...	This game is the game (that is not CS:GO) that...	442.0	100.0	461.0
3	3	Undertale	2015-09-15	tobyfox, 8-4	4.2	424.5	424.5	Adventure	A small child falls into the Underground, wher...	soundtrack is tied for #1 with nier automata. ...	442.0	679.0	461.0
4	4	Hollow Knight	2017-02-24	Team Cherry	4.4	424.5	424.5	Adventure	A 2D metroidvania with an emphasis on close co...	this games worldbuilding is incredible, with i...	442.0	100.0	461.0
5	5	Minecraft	2011-11-18	Mojang Studios	4.3	424.5	424.5	Adventure	Minecraft focuses on allowing the player to ex...	Minecraft is what you make of it. Unfortunatel...	442.0	100.0	461.0
6	6	Omori	2020-12-25	OMOCAT, PLAYISM	4.2	424.5	424.5	Adventure	A turn-based surreal horror RPG in which a chi...	The best game Ive played in my life, omori is ...	442.0	100.0	461.0
7	7	Metroid Dread	2021-10-07	Nintendo, MercurySteam	4.3	424.5	424.5	Adventure	Join intergalactic bounty hunter Samus Aran in...	Have only been a Metroid fan for couple of yea...	442.0	759.0	461.0
8	8	Among Us	2018-06-15	InnerSloth	3.0	867.0	867.0	Indie	Join your crew-mates in a multiplayer game of ...	its a solid party game. im bad at lying though...	442.0	470.0	776.0
9	9	NieR: Automata	2017-02-23	PlatinumGames, Square Enix	4.3	424.5	424.5	Brawler	NieR: Automata tells the story of androids 2B,...	Holy shit, im carrying the weight of the woooo...	442.0	100.0	461.0

First 10 rows of vgsales.csv

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
5	6	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
6	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.50	2.90	30.01
7	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.20	2.93	2.85	29.02
8	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.70	2.26	28.62
9	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31

```
# Download Cleaned Data
games.to_csv("cleaned_games.csv", index=False, encoding="utf-8-sig")
vgsales.to_csv("cleaned_vgsales.csv", index=False, encoding="utf-8-sig")

from google.colab import files
files.download("cleaned_games.csv")
files.download("cleaned_vgsales.csv")
```

```
cleaned_games = pd.read_csv("cleaned_games.csv")
cleaned_vgsales = pd.read_csv("cleaned_vgsales.csv")

# Combine side-by-side based on index
combined = pd.concat([cleaned_games, cleaned_vgsales], axis=1)

# Check
combined.head()
```

	Title	Release Date	Team	Rating	Times Listed	Number of Reviews	Genres	Plays	Playing	Backlogs	...	Name	Platform	Year
0	Elden Ring	25-02-22	Bandai Namco Entertainment, FromSoftware	4.5	424.5	424.5	Adventure	442.0	100.0	461.0	...	Wii Sports	Wii	2006.0
1	Hades	10-12-19	Supergiant Games	4.3	424.5	424.5	Adventure	442.0	100.0	461.0	...	Super Mario Bros.	NES	1985.0
2	The Legend of Zelda: Breath of the Wild	03-03-17	Nintendo, Nintendo EPD Production Group No. 3	4.4	424.5	424.5	Adventure	442.0	100.0	461.0	...	Mario Kart Wii	Wii	2008.0
3	Undertale	15-09-15	tobyfox, 8-4	4.2	424.5	424.5	Adventure	442.0	679.0	461.0	...	Wii Sports Resort	Wii	2009.0
4	Hollow Knight	24-02-17	Team Cherry	4.4	424.5	424.5	Adventure	442.0	100.0	461.0	...	Pokemon Red/Pokemon Blue	GB	1996.0

5 rows × 21 columns

```
combined["Game_Title"] = combined["Title"].fillna(combined["Name"])
```

```
combined["Game_Title"] = combined["Game_Title"].astype(str).str.strip()
combined["Game_Title"] = combined["Game_Title"].str.replace(r"[^A-Za-z0-9\s]", "", regex=True)
combined["Game_Title"] = combined["Game_Title"].str.replace(r"\s+", " ", regex=True)
combined["Game_Title"] = combined["Game_Title"].str.title()
```

```
combined.drop(columns=["Title", "Name"], inplace=True)
```

```
combined.columns
```

```
Index(['Release Date', 'Team', 'Rating', 'Times Listed', 'Number of Reviews',
       'Genres', 'Plays', 'Playing', 'Backlogs', 'Wishlist', 'Platform',
       'Year', 'Genre', 'Publisher', 'NA_Sales', 'EU_Sales', 'JP_Sales',
       'Other_Sales', 'Global_Sales', 'Game_Title'],
      dtype='object')
```

```
#Release Date Format
combined['Release Date'] = pd.to_datetime(combined['Release Date'], errors='coerce').dt.strftime('%Y-%m-%d')
```

```
/tmp/ipython-input-668084456.py:2: UserWarning: Could not infer format, so each element will be parsed individually, falling back t
  combined['Release Date'] = pd.to_datetime(combined['Release Date'], errors='coerce').dt.strftime('%Y-%m-%d')
```

```
combined.head()
```

	Release Date	Team	Rating	Times Listed	Number of Reviews	Genres	Plays	Playing	Backlogs	Wishlist	Platform	Year	Genre	Publisher
0	2022-02-25	Bandai Namco Entertainment, FromSoftware	4.5	424.5	424.5	Adventure	442.0	100.0	461.0	334.0	Wii	2006.0	Sports	Nintendo
1	2019-10-12	Supergiant Games	4.3	424.5	424.5	Adventure	442.0	100.0	461.0	334.0	NES	1985.0	Platform	Nintendo
2	2017-03-03	Nintendo, Nintendo EPD Production Group No. 3	4.4	424.5	424.5	Adventure	442.0	100.0	461.0	334.0	Wii	2008.0	Racing	Nintendo
3	2015-09-15	tobyfox, 8-4	4.2	424.5	424.5	Adventure	442.0	679.0	461.0	334.0	Wii	2009.0	Sports	Nintendo
4	2017-02-24	Team Cherry	4.4	424.5	424.5	Adventure	442.0	100.0	461.0	334.0	GB	1996.0	Role-Playing	Nintendo

```
# Save combined dataset as a new CSV file
combined.to_csv("merged_games_data.csv", index=False, encoding="utf-8-sig")

# Download it
from google.colab import files
files.download("merged_games_data.csv")
```