Game Sales Analysis

Eunike Hedriani Pardede

Table of Contents

Objective

Analyze the game sales by release date, its series, publishers and developers

Research Questions

1. Which game is the oldest and the newest games?
2. Which publisher published most of the games?
3. Which developer developed most of the games?
4. Which series is the most sales?
5. Which series have the most games?

Data Walkthrough

- Dataset : https://docs.google.com/spreadsheets/d/10poofg-l8DMdtUgGy8mOpra2lHmA9EQC7drmF9AyYHA/edit#gid=1485085913
- 177 rows , 7 columns
- Features:
  1. Name
  2. Sales (in millions)
  3. Series
  4. Release (date)
  5. Genre
  6. Developer
  7. Publisher

Data Processing

Import Data

1. Import libraries
2. Import dataset
3. Inspect the data

Clean Data

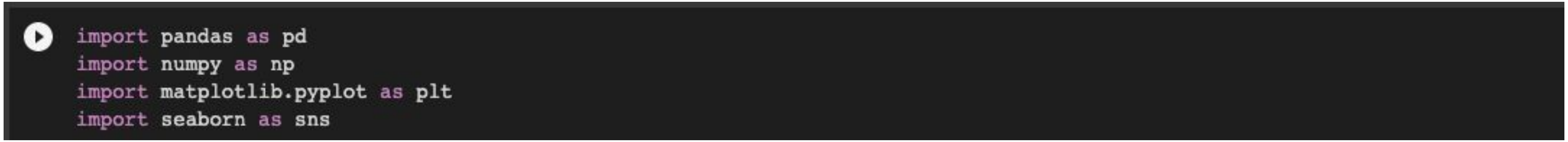
1. Change the datatype
2. Check null values
3. Clean ‘Publisher’ column

Explore & Visualize Data

1. Which game is the oldest and the newest games?
2. Which publisher published most of the games?
3. Which developer developed most of the games?
4. Which series is the most sales?
5. Which series have the most games?

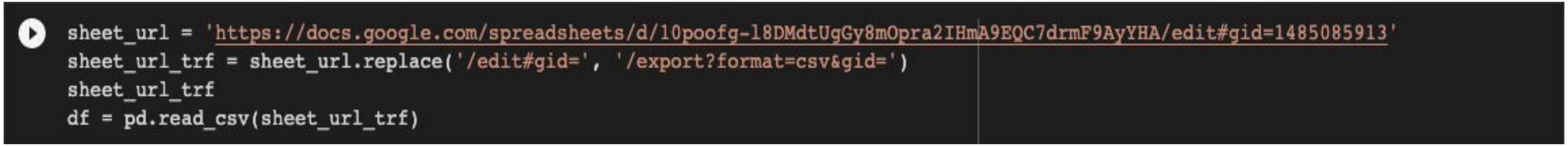
Import Data

1.Import libraries



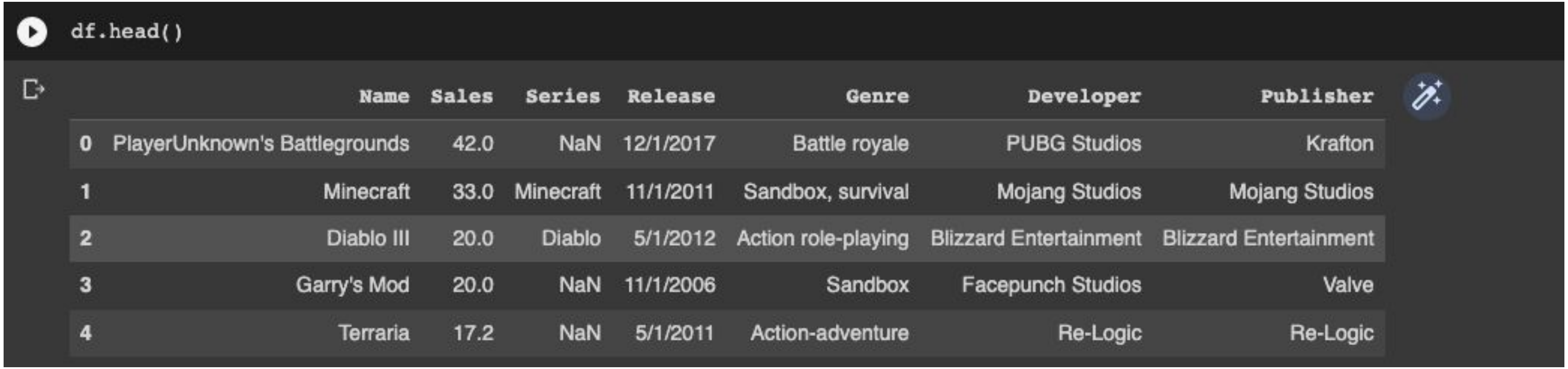
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Import dataset



```
sheet_url = 'https://docs.google.com/spreadsheets/d/10poofg-l8DMdtUgGy8mOpra2IHmA9EQC7drmF9AyYHA/edit#gid=1485085913'
sheet_url_trf = sheet_url.replace('/edit#gid=', '/export?format=csv&gid=')
sheet_url_trf
df = pd.read_csv(sheet_url_trf)
```

3. Inspect the data

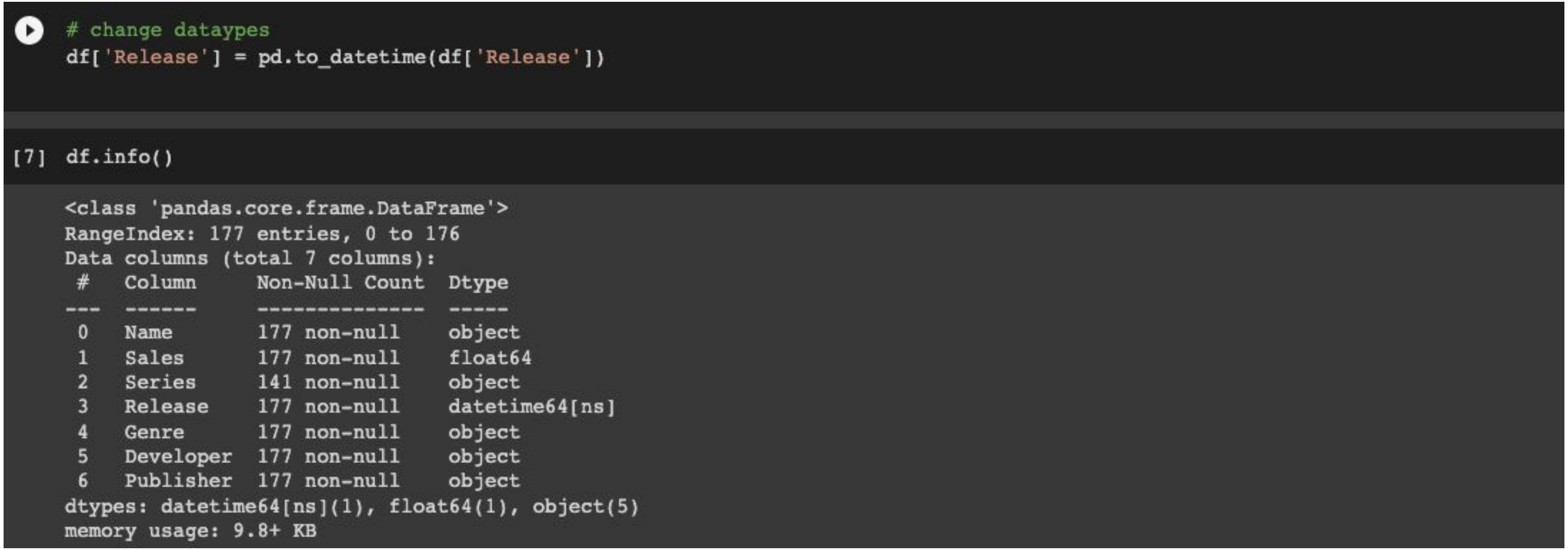


```
df.head()
```

```
                            Name  Sales     Series    Release                Genre               Developer               Publisher
0  PlayerUnknown's Battlegrounds   42.0        NaN  12/1/2017         Battle royale            PUBG Studios                  Krafton
1                      Minecraft   33.0  Minecraft  11/1/2011     Sandbox, survival          Mojang Studios           Mojang Studios
2                      Diablo III  20.0     Diablo   5/1/2012  Action role-playing  Blizzard Entertainment  Blizzard Entertainment
3                     Garry's Mod  20.0        NaN  11/1/2006               Sandbox       Facepunch Studios                    Valve
4                        Terraria  17.2        NaN   5/1/2011      Action-adventure                Re-Logic                 Re-Logic
```

Clean Data (1/4)

1. Change the datatype of ‘Release’ column from ‘object’ to ‘datetime’



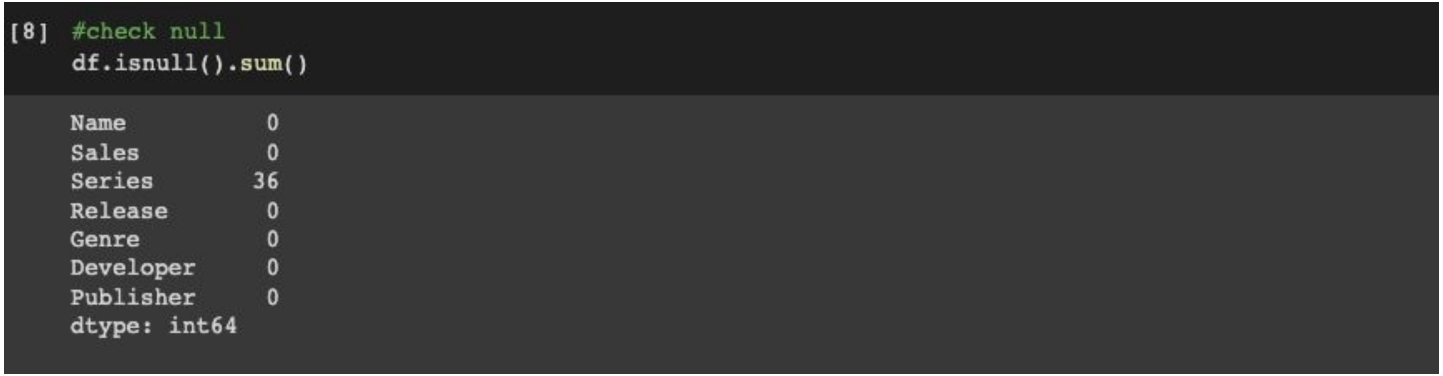
```
# change dataypes
df['Release'] = pd.to_datetime(df['Release'])
```

```
[7] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 177 entries, 0 to 176
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Name       177 non-null    object
 1   Sales      177 non-null    float64
 2   Series     141 non-null    object
 3   Release    177 non-null    datetime64[ns]
 4   Genre      177 non-null    object
 5   Developer  177 non-null    object
 6   Publisher  177 non-null    object
dtypes: datetime64[ns](1), float64(1), object(5)
memory usage: 9.8+ KB
```

Clean Data (2/4)

2. Check the null values



```
[8] #check null
    df.isnull().sum()
```

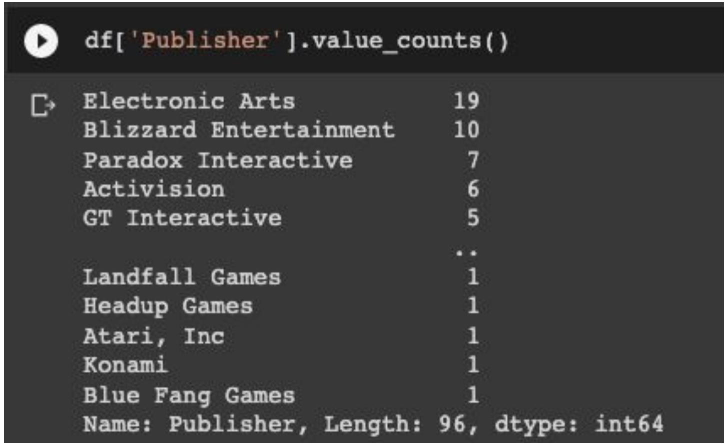
```
Name          0
Sales         0
Series       36
Release       0
Genre         0
Developer     0
Publisher     0
dtype: int64
```

The number of null values on ‘Series’ column are **36** ,which is around **20%** from total rows (177).
So, I kept those null values, instead of dropping them for avoiding the wrong conclusion.

Clean Data (3/4)

3. Clean ‘Publisher’ Column

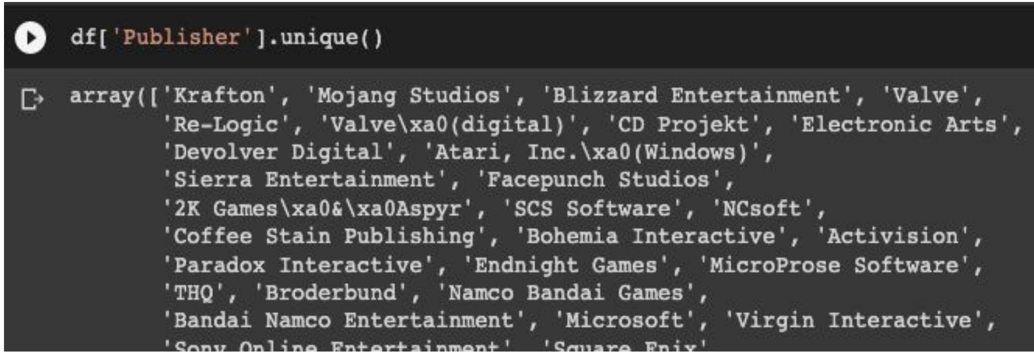
a. This was initial count of games based on ‘Publisher’ column. There were **96 publishers** in total. But, some of them had similar name (duplicated name with wrong spelling)



```
df['Publisher'].value_counts()

Electronic Arts          19
Blizzard Entertainment   10
Paradox Interactive       7
Activision                6
GT Interactive            5
                         ..
Landfall Games            1
Headup Games              1
Atari, Inc                1
Konami                    1
Blue Fang Games           1
Name: Publisher, Length: 96, dtype: int64
```

b. Check the unique values from ‘Publisher’ column. It was found that there were so many duplicate values with similar name, such as:
'Brøderbund' ↣ **'Broderbund'**
'ConcernedApe[f]' ↣ **'ConcernedApe'**, etc
Therefore, it needs to be cleaned.

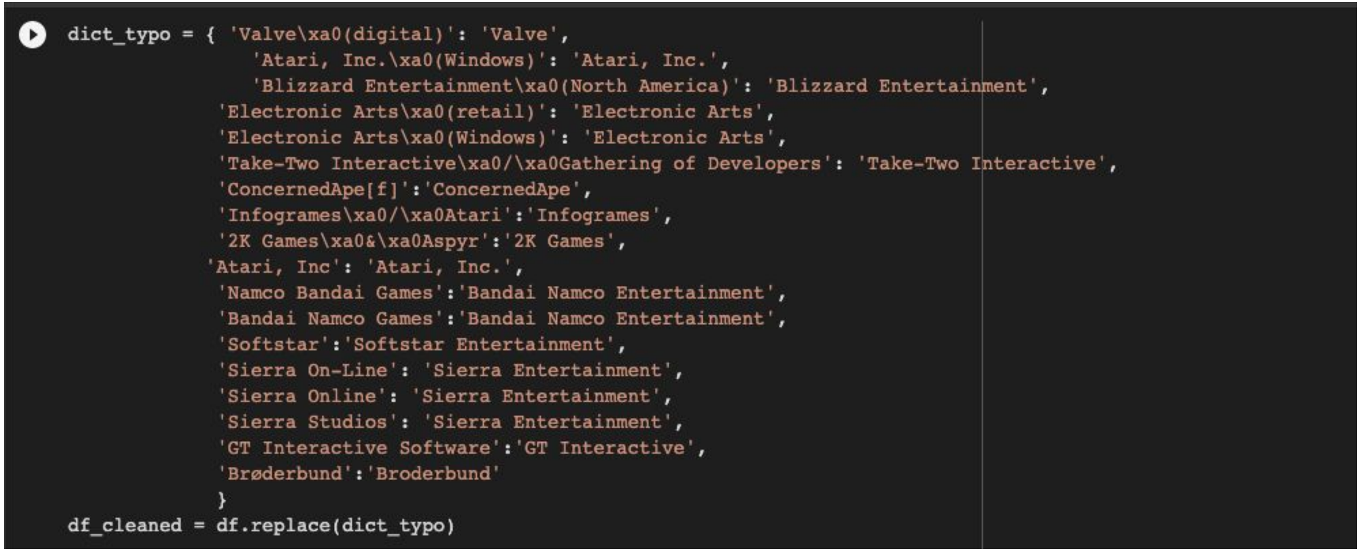


```
df['Publisher'].unique()

array(['Krafton', 'Mojang Studios', 'Blizzard Entertainment', 'Valve',
       'Re-Logic', 'Valve\xa0(digital)', 'CD Projekt', 'Electronic Arts',
       'Devolver Digital', 'Atari, Inc.\xa0(Windows)',
       'Sierra Entertainment', 'Facepunch Studios',
       '2K Games\xa0&\xa0Aspyr', 'SCS Software', 'NCsoft',
       'Coffee Stain Publishing', 'Bohemia Interactive', 'Activision',
       'Paradox Interactive', 'Endnight Games', 'MicroProse Software',
       'THQ', 'Broderbund', 'Namco Bandai Games',
       'Bandai Namco Entertainment', 'Microsoft', 'Virgin Interactive',
       'Sony Online Entertainment', 'Square Enix',
```

Clean Data (4/4)

c. Replace some publisher name that have similar names



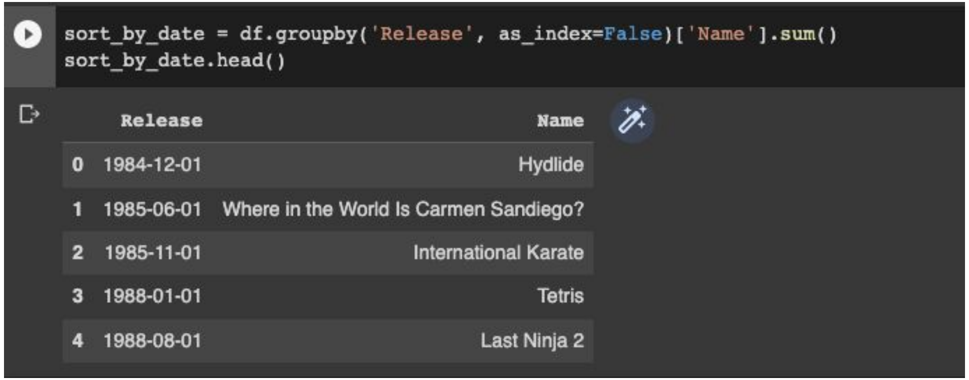
```
dict_typo = { 'Valve\xa0(digital)': 'Valve',
                'Atari, Inc.\xa0(Windows)': 'Atari, Inc.',
                'Blizzard Entertainment\xa0(North America)': 'Blizzard Entertainment',
             'Electronic Arts\xa0(retail)': 'Electronic Arts',
             'Electronic Arts\xa0(Windows)': 'Electronic Arts',
             'Take-Two Interactive\xa0/\xa0Gathering of Developers': 'Take-Two Interactive',
             'ConcernedApe[f]':'ConcernedApe',
             'Infogrames\xa0/\xa0Atari':'Infogrames',
             '2K Games\xa0&\xa0Aspyr':'2K Games',
            'Atari, Inc': 'Atari, Inc.',
             'Namco Bandai Games':'Bandai Namco Entertainment',
             'Bandai Namco Games':'Bandai Namco Entertainment',
             'Softstar':'Softstar Entertainment',
             'Sierra On-Line': 'Sierra Entertainment',
             'Sierra Online': 'Sierra Entertainment',
             'Sierra Studios': 'Sierra Entertainment',
             'GT Interactive Software':'GT Interactive',
             'Brøderbund':'Broderbund'
             }
df_cleaned = df.replace(dict_typo)
```

1. Which game is the oldest and the newest games?

1. Use groupby ‘Release’ date to find the oldest and newest games

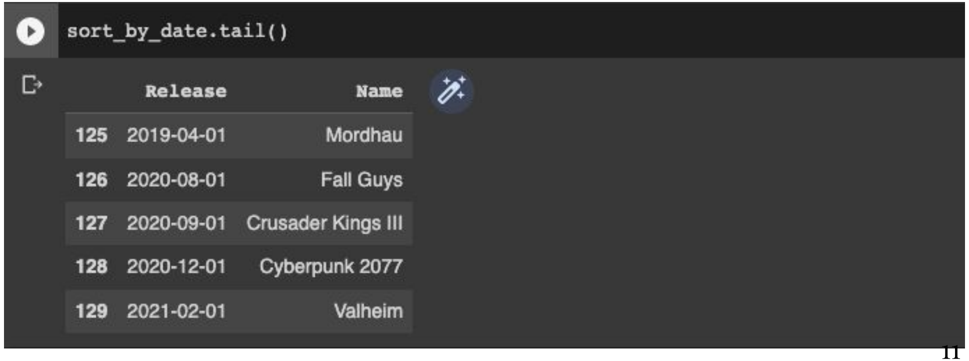
2. Use head to find the oldest (index: 0)
   Answer: Hydlide 1984-12-01

3. Use tail to find the newest (index: 129)
   Answer: Valheim 2021-02-01



```
sort_by_date = df.groupby('Release', as_index=False)['Name'].sum()
sort_by_date.head()
```

```
     Release                                      Name
0    1984-12-01                                Hydlide
1    1985-06-01   Where in the World Is Carmen Sandiego?
2    1985-11-01                    International Karate
3    1988-01-01                                  Tetris
4    1988-08-01                             Last Ninja 2
```

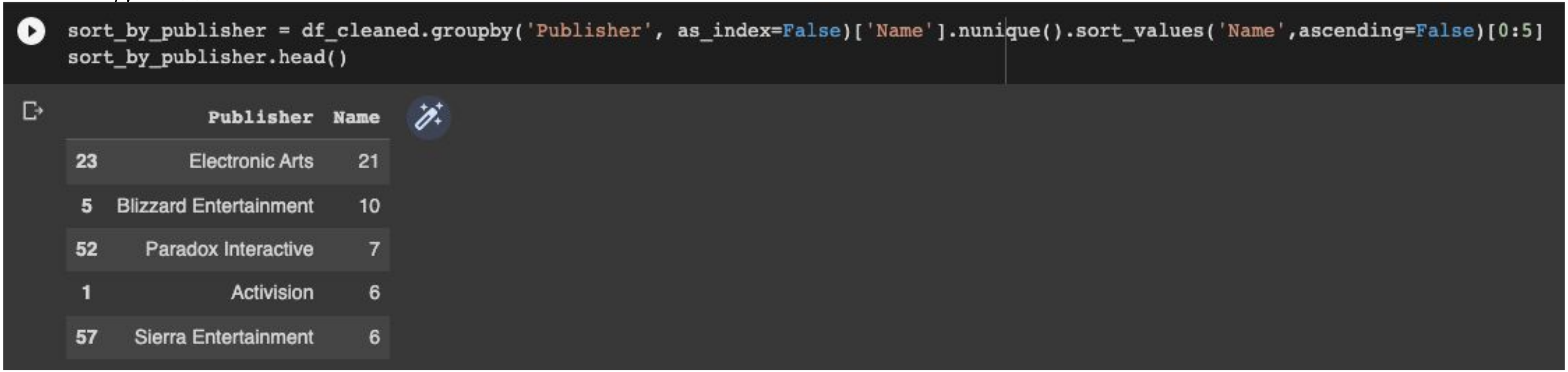


```
sort_by_date.tail()
```

```
      Release                Name
125   2019-04-01          Mordhau
126   2020-08-01        Fall Guys
127   2020-09-01   Crusader Kings III
128   2020-12-01   Cyberpunk 2077
129   2021-02-01          Valheim
```

2. Which publisher published most of the games? (1/2)

1. Perform groupby ‘Publisher’ to count the number of games.

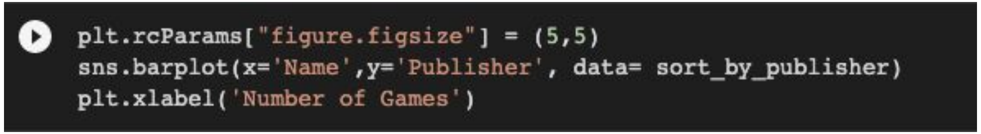


```
sort_by_publisher = df_cleaned.groupby('Publisher', as_index=False)['Name'].nunique().sort_values('Name',ascending=False)[0:5]
sort_by_publisher.head()
```

	Publisher	Name
23	Electronic Arts	21
5	Blizzard Entertainment	10
52	Paradox Interactive	7
1	Activision	6
57	Sierra Entertainment	6

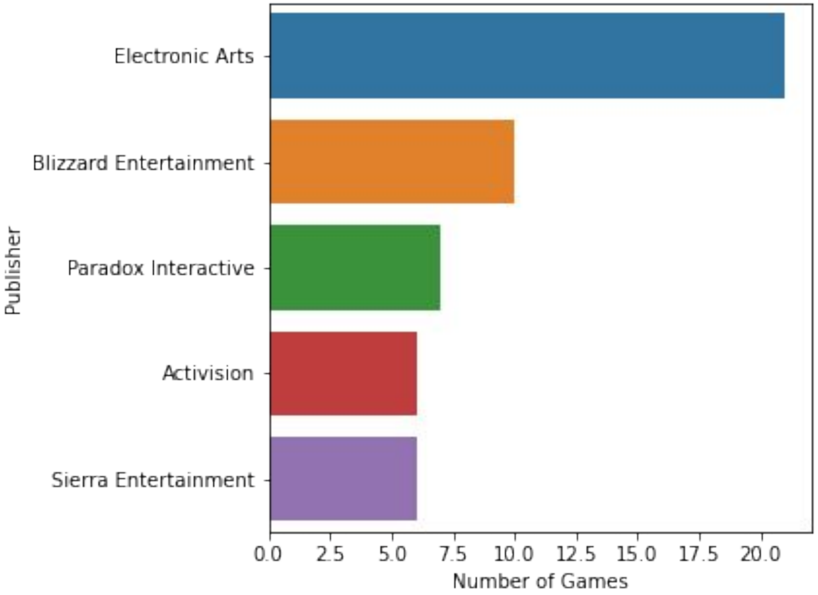
2. Which publisher published most of the games? (2/2)

2. Using barplot from Seaborn to show the graph



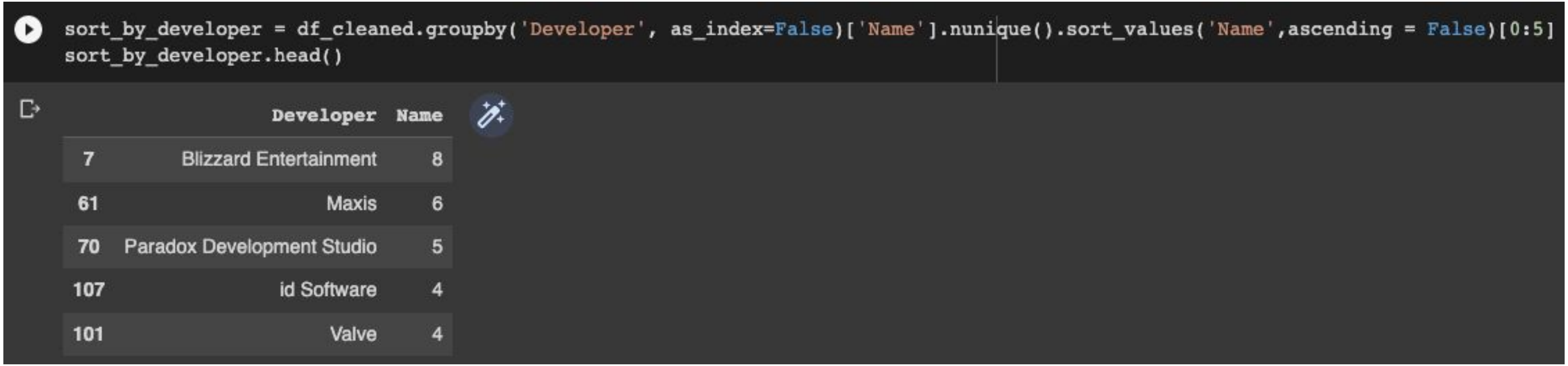
```
plt.rcParams["figure.figsize"] = (5,5)
sns.barplot(x='Name',y='Publisher', data= sort_by_publisher)
plt.xlabel('Number of Games')
```

Electronic Arts is the top 1 that published most games with **21 games.**
Followed by Blizzard Entertainment (10 games) and Paradox Interactive (7 games).



3. Which developer developed most of the games?(1/2)

1. Perform groupby ‘Developer’ to count the number of games of each developer.

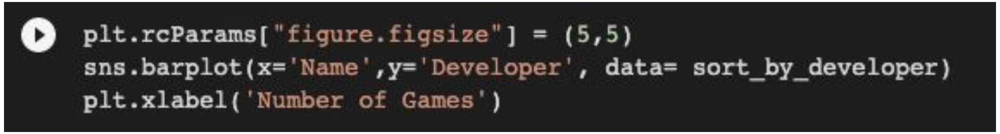


```python
sort_by_developer = df_cleaned.groupby('Developer', as_index=False)['Name'].nunique().sort_values('Name',ascending = False)[0:5]
sort_by_developer.head()
```

	Developer	Name
7	Blizzard Entertainment	8
61	Maxis	6
70	Paradox Development Studio	5
107	id Software	4
101	Valve	4

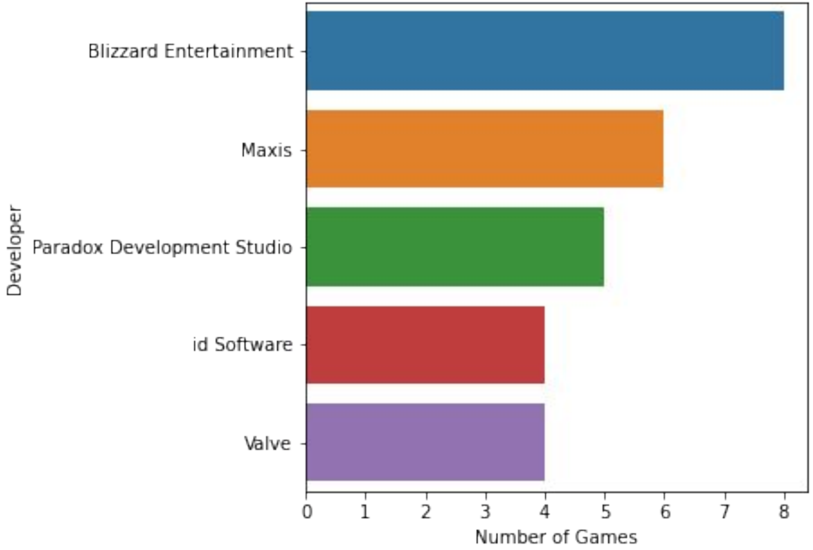
3. Which developer developed most of the games?(2/2)

2. Using barplot from Seaborn to visualize the data



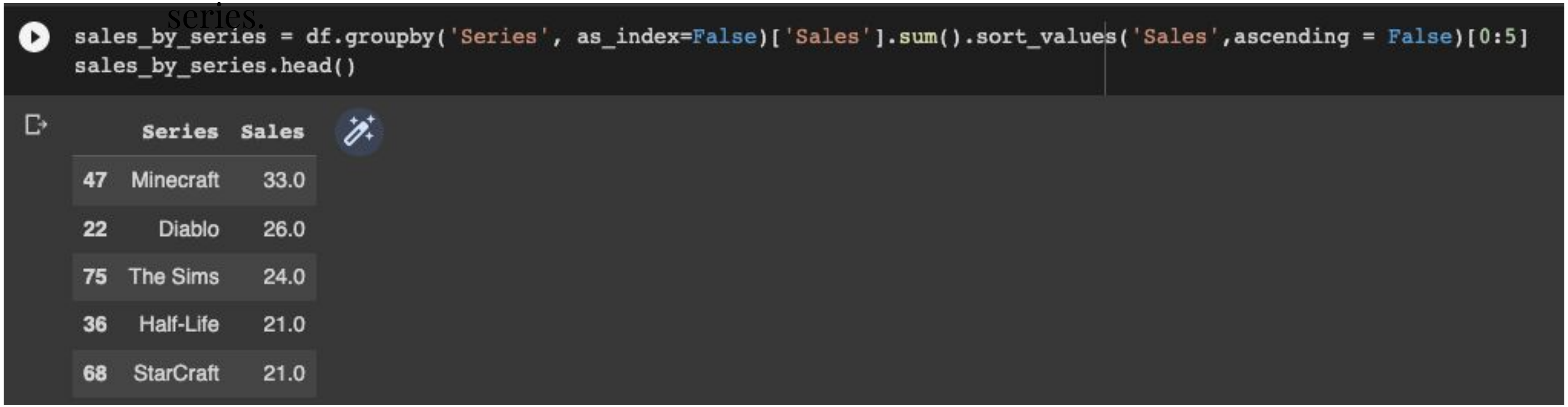
```
plt.rcParams["figure.figsize"] = (5,5)
sns.barplot(x='Name',y='Developer', data= sort_by_developer)
plt.xlabel('Number of Games')
```

Blizzard Entertainment is the top 1 that developed most games with **8 games.** Followed by Maxis (6 games) and Paradox Development Studio (5 games).



4. Which series is the most sales?(1/2)

1. Perform groupby ‘Series’ to summarize the total sales of each series.

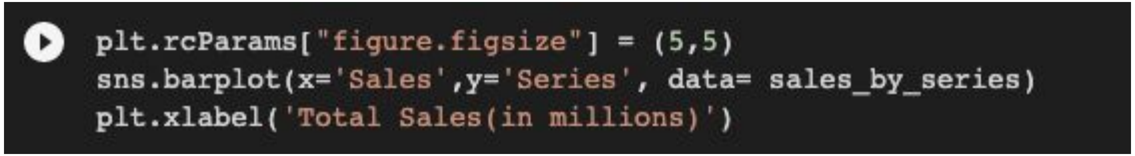


```
sales_by_series = df.groupby('Series', as_index=False)['Sales'].sum().sort_values('Sales',ascending = False)[0:5]
sales_by_series.head()
```

```
       Series  Sales
47  Minecraft   33.0
22     Diablo   26.0
75   The Sims   24.0
36  Half-Life   21.0
68  StarCraft   21.0
```

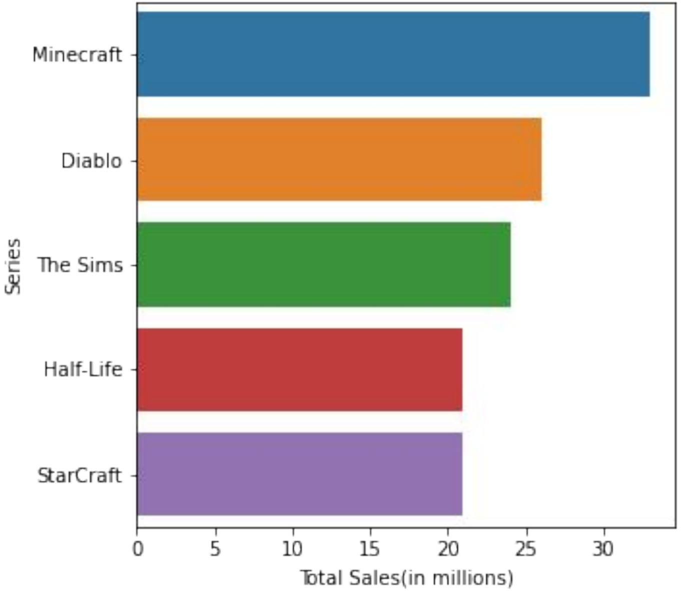
4. Which series is the most sales?(2/2)

2. Using barplot from Seaborn to visualize the data



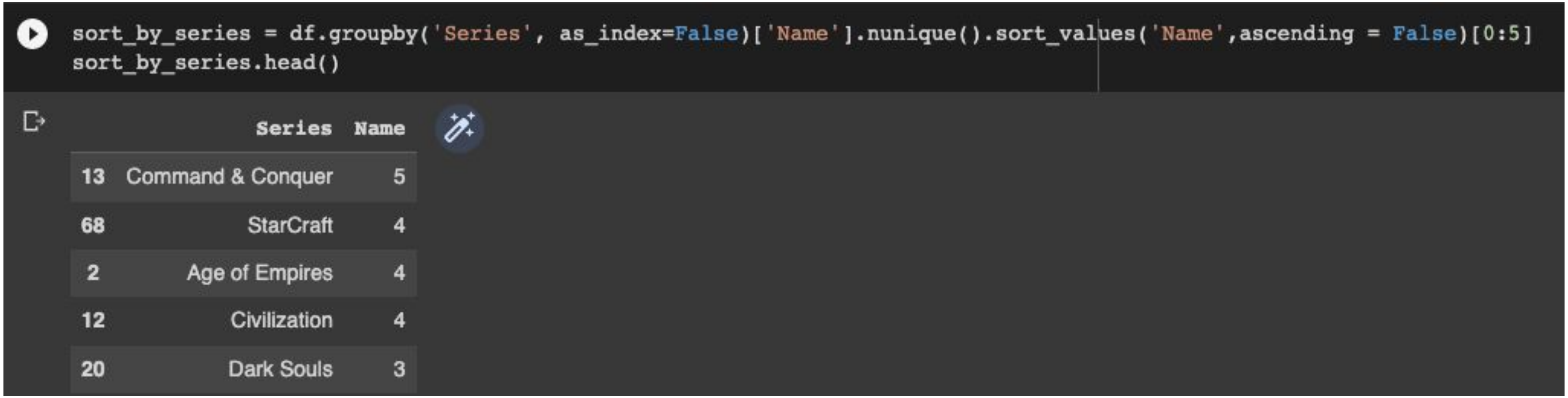
```
plt.rcParams["figure.figsize"] = (5,5)
sns.barplot(x='Sales',y='Series', data= sales_by_series)
plt.xlabel('Total Sales(in millions)')
```

Minecraft is the most popular series with sales of **33 millions .**
Followed by Diablo (26 millions) and The Sims (24 millions).



5. Which series have the most games?(1/2)

1. Groupby 'Series' to count the number of games of each series.

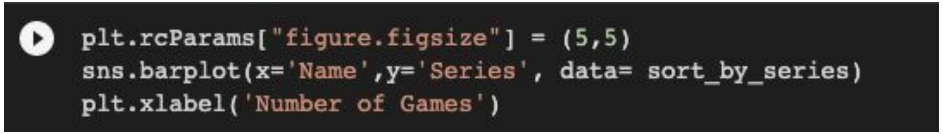


```
sort_by_series = df.groupby('Series', as_index=False)['Name'].nunique().sort_values('Name',ascending = False)[0:5]
sort_by_series.head()
```

	Series	Name
13	Command & Conquer	5
68	StarCraft	4
2	Age of Empires	4
12	Civilization	4
20	Dark Souls	3

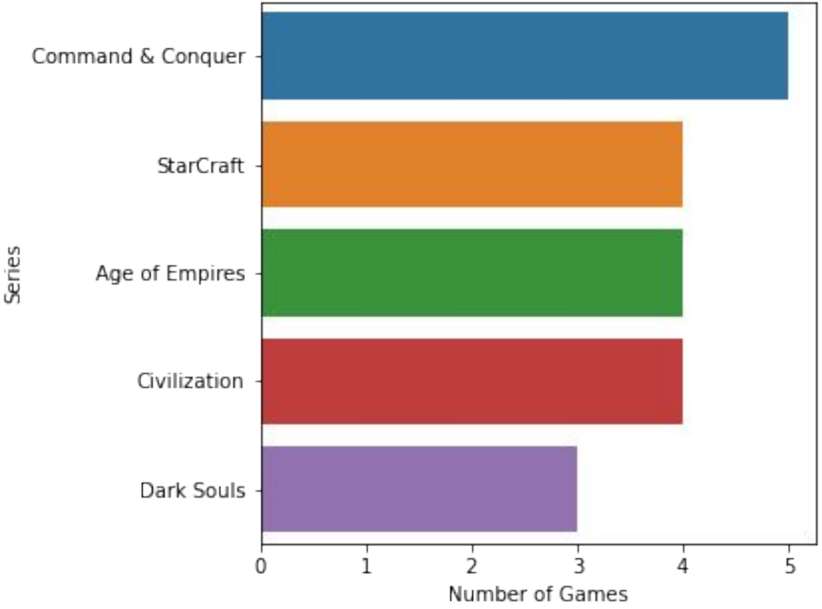
5. Which series have the most games?(2/2)

2. Using barplot from Seaborn to visualize the data



```
plt.rcParams["figure.figsize"] = (5,5)
sns.barplot(x='Name',y='Series', data= sort_by_series)
plt.xlabel('Number of Games')
```

Command & Conquer has the most games series with 5 games.
Followed by StarCraft, Age of Empires and Civilization with 4 games.



Conclusion

1. Hydlide is the oldest games ever, which was released on Dec 1984. Meanwhile, Valheim is the newest game, was released on Feb 2021.
2. Electronic Arts has published the most of games compared other publishers.
3. Blizzard Entertainment is the most productive developer.
4. In terms of the sales, Minecraft is the most popular series.
5. Command & Conquer has the most games series.