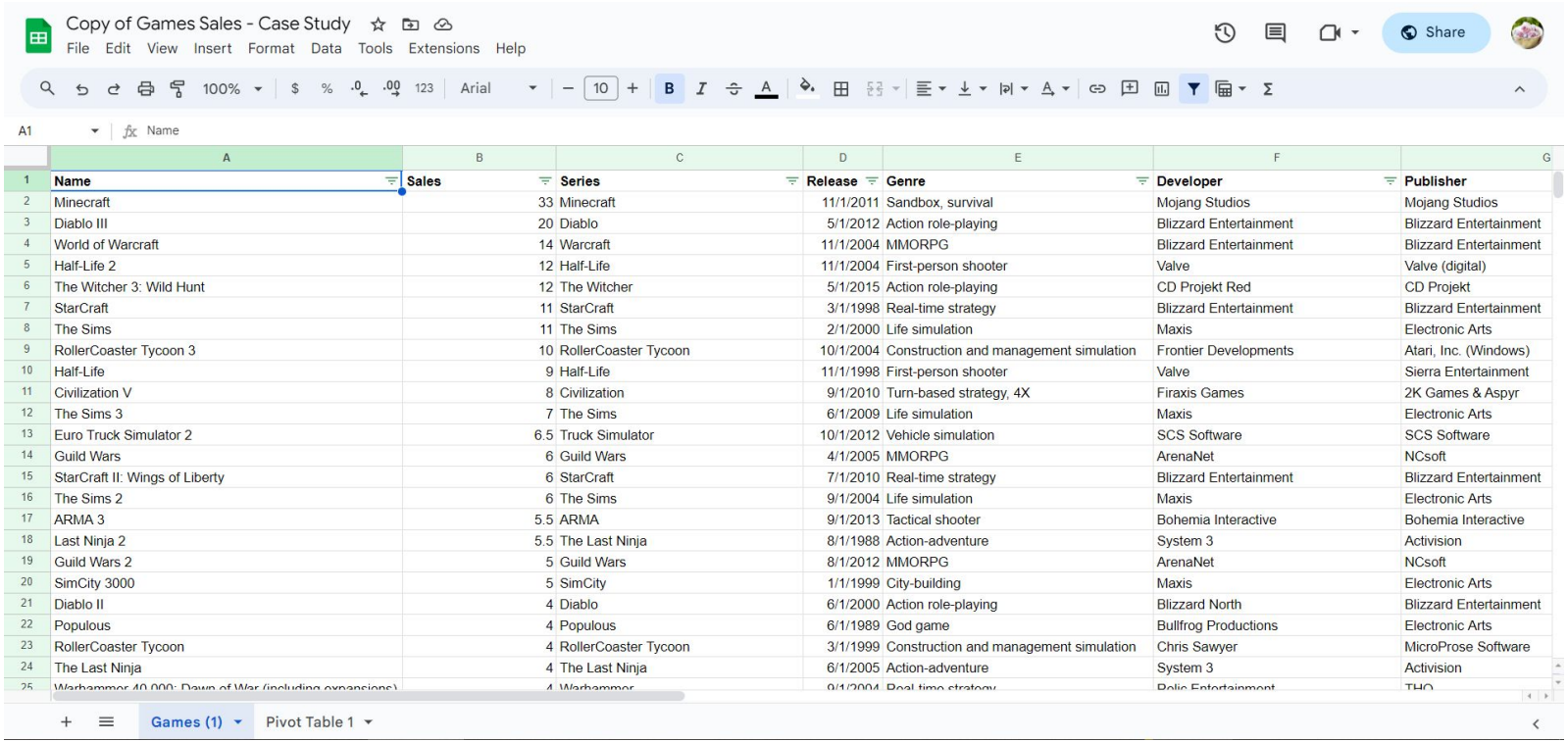
# Games Sales

Nur Fatih Alam

Case Study Mini Course: Data Analytics RevoU Batch Oct 2, 2023

Tools: Google Colaboratory (Python)

# Dataset from Google Sheets




| | Name | Sales | Series | Release | Genre | Developer | Publisher |
|---|---|---|---|---|---|---|---|
| 2 | Minecraft | 33 | Minecraft | 11/1/2011 | Sandbox, survival | Mojang Studios | Mojang Studios |
| 3 | Diablo III | 20 | Diablo | 5/1/2012 | Action role-playing | Blizzard Entertainment | Blizzard Entertainment |
| 4 | World of Warcraft | 14 | Warcraft | 11/1/2004 | MMORPG | Blizzard Entertainment | Blizzard Entertainment |
| 5 | Half-Life 2 | 12 | Half-Life | 11/1/2004 | First-person shooter | Valve | Valve (digital) |
| 6 | The Witcher 3: Wild Hunt | 12 | The Witcher | 5/1/2015 | Action role-playing | CD Projekt Red | CD Projekt |
| 7 | StarCraft | 11 | StarCraft | 3/1/1998 | Real-time strategy | Blizzard Entertainment | Blizzard Entertainment |
| 8 | The Sims | 11 | The Sims | 2/1/2000 | Life simulation | Maxis | Electronic Arts |
| 9 | RollerCoaster Tycoon 3 | 10 | RollerCoaster Tycoon | 10/1/2004 | Construction and management simulation | Frontier Developments | Atari, Inc. (Windows) |
| 10 | Half-Life | 9 | Half-Life | 11/1/1998 | First-person shooter | Valve | Sierra Entertainment |
| 11 | Civilization V | 8 | Civilization | 9/1/2010 | Turn-based strategy, 4X | Firaxis Games | 2K Games & Aspyr |
| 12 | The Sims 3 | 7 | The Sims | 6/1/2009 | Life simulation | Maxis | Electronic Arts |
| 13 | Euro Truck Simulator 2 | 6.5 | Truck Simulator | 10/1/2012 | Vehicle simulation | SCS Software | SCS Software |
| 14 | Guild Wars | 6 | Guild Wars | 4/1/2005 | MMORPG | ArenaNet | NCsoft |
| 15 | StarCraft II: Wings of Liberty | 6 | StarCraft | 7/1/2010 | Real-time strategy | Blizzard Entertainment | Blizzard Entertainment |
| 16 | The Sims 2 | 6 | The Sims | 9/1/2004 | Life simulation | Maxis | Electronic Arts |
| 17 | ARMA 3 | 5.5 | ARMA | 9/1/2013 | Tactical shooter | Bohemia Interactive | Bohemia Interactive |
| 18 | Last Ninja 2 | 5.5 | The Last Ninja | 8/1/1988 | Action-adventure | System 3 | Activision |
| 19 | Guild Wars 2 | 5 | Guild Wars | 8/1/2012 | MMORPG | ArenaNet | NCsoft |
| 20 | SimCity 3000 | 5 | SimCity | 1/1/1999 | City-building | Maxis | Electronic Arts |
| 21 | Diablo II | 4 | Diablo | 6/1/2000 | Action role-playing | Blizzard North | Blizzard Entertainment |
| 22 | Populous | 4 | Populous | 6/1/1989 | God game | Bullfrog Productions | Electronic Arts |
| 23 | RollerCoaster Tycoon | 4 | RollerCoaster Tycoon | 3/1/1999 | Construction and management simulation | Chris Sawyer | MicroProse Software |
| 24 | The Last Ninja | 4 | The Last Ninja | 6/1/2005 | Action-adventure | System 3 | Activision |
| 25 | Warhammer 40,000: Dawn of War (including expansions) | 4 | Warhammer | 9/1/2004 | Real-time strategy | Relic Entertainment | THQ |



https://docs.google.com/spreadsheets/d/18nCexUyyqZ2g74BalhoLut8qYd2UuTzC7kLKCQ8v1d0/edit#gid=1485085913

# Business Questions

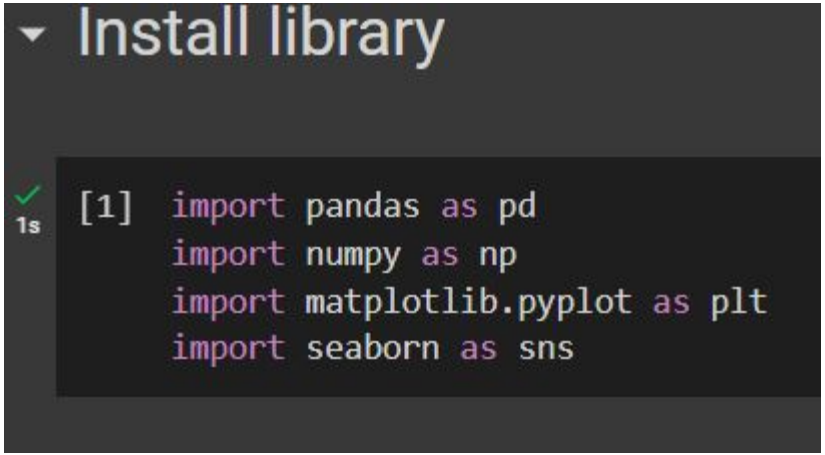
**Context**

Gaming industry is an interesting field to explore, it would be fun knowing who is the most popular publishers and developers and which games are the most popular.

**Questions**

- Which game is the oldest and the newest games in that dataset?
- Which publisher published most of the games?
- Which developer developed most of the games?
- Which series has the most sales?
- Which series have the most games?

# Install library




Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Seaborn is a Python data visualization library based on matplotlib.

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Import data

## Import data

```
[2]  # read data from google sheet
     sheet_url = 'https://docs.google.com/spreadsheets/d/18nCexUyyqZ2g74BalhoLut8qYd2UuTzC7kLKCQ8v1d0/edit#gid=1485085913'
     sheet_url_trf = sheet_url.replace('/edit#gid=', '/export?format=csv&gid=')
     print(sheet_url_trf)
     df = pd.read_csv(sheet_url_trf)
```

```
https://docs.google.com/spreadsheets/d/18nCexUyyqZ2g74BalhoLut8qYd2UuTzC7kLKCQ8v1d0/export?format=csv&gid=1485085913
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 139 entries, 0 to 138
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Name       139 non-null    object
 1   Sales      139 non-null    float64
 2   Series     139 non-null    object
 3   Release    139 non-null    object
 4   Genre      139 non-null    object
 5   Developer  139 non-null    object
 6   Publisher  139 non-null    object
dtypes: float64(1), object(6)
memory usage: 7.7+ KB
```

```
[4]  df.columns
```

```
Index(['Name', 'Sales', 'Series', 'Release', 'Genre', 'Developer',
       'Publisher'],
      dtype='object')
```

Use syntax to be able to read data from Google Sheets and convert it into CSV form and make sure the data is correct

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Cleaning

```
[5] # change data type 'Release' to datetime
    df['Release'] = pd.to_datetime(df['Release'])
    df = df.drop_duplicates()
    df.info()
    df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 139 entries, 0 to 138
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Name       139 non-null    object
 1   Sales      139 non-null    float64
 2   Series     139 non-null    object
 3   Release    139 non-null    datetime64[ns]
 4   Genre      139 non-null    object
 5   Developer  139 non-null    object
 6   Publisher  139 non-null    object
dtypes: datetime64[ns](1), float64(1), object(5)
memory usage: 8.7+ KB
```

| | Name | Sales | Series | Release | Genre | Developer | Publisher |
|---|---|---|---|---|---|---|---|
| 0 | Minecraft | 33.0 | Minecraft | 2011-11-01 | Sandbox, survival | Mojang Studios | Mojang Studios |
| 1 | Diablo III | 20.0 | Diablo | 2012-05-01 | Action role-playing | Blizzard Entertainment | Blizzard Entertainment |
| 2 | World of Warcraft | 14.0 | Warcraft | 2004-11-01 | MMORPG | Blizzard Entertainment | Blizzard Entertainment |
| 3 | Half-Life 2 | 12.0 | Half-Life | 2004-11-01 | First-person shooter | Valve | Valve (digital) |
| 4 | The Witcher 3: Wild Hunt | 12.0 | The Witcher | 2015-05-01 | Action role-playing | CD Projekt Red | CD Projekt |

Make sure to data clean first. Remove duplicate data and change data type to correctly such as date time.

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Visualization

Which game is the oldest and the newest games in that dataset?

```
[6]  df[['Name','Release']].sort_values('Release').head()
```

| | Name | Release |
|---|---|---|
| 107 | Hydlide | 1984-12-01 |
| 24 | Where in the World Is Carmen Sandiego? | 1985-06-01 |
| 72 | International Karate | 1985-11-01 |
| 128 | Tetris | 1988-01-01 |
| 16 | Last Ninja 2 | 1988-08-01 |

The oldest game is Hydlide

```
[7]  df[['Name','Release']].sort_values('Release', ascending=False).head()
```

| | Name | Release |
|---|---|---|
| 90 | Crusader Kings III | 2020-09-01 |
| 96 | Divinity: Original Sin II | 2017-09-01 |
| 76 | Nier: Automata | 2017-03-01 |
| 104 | Hearts of Iron IV | 2016-06-01 |
| 42 | 7 Days to Die | 2016-06-01 |

The newest game is Crusader Kings III

By using sort_values() can get answer. The oldest game is Hydlide and the newest game is Valheim

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Visualization

Which publisher published most of the games?

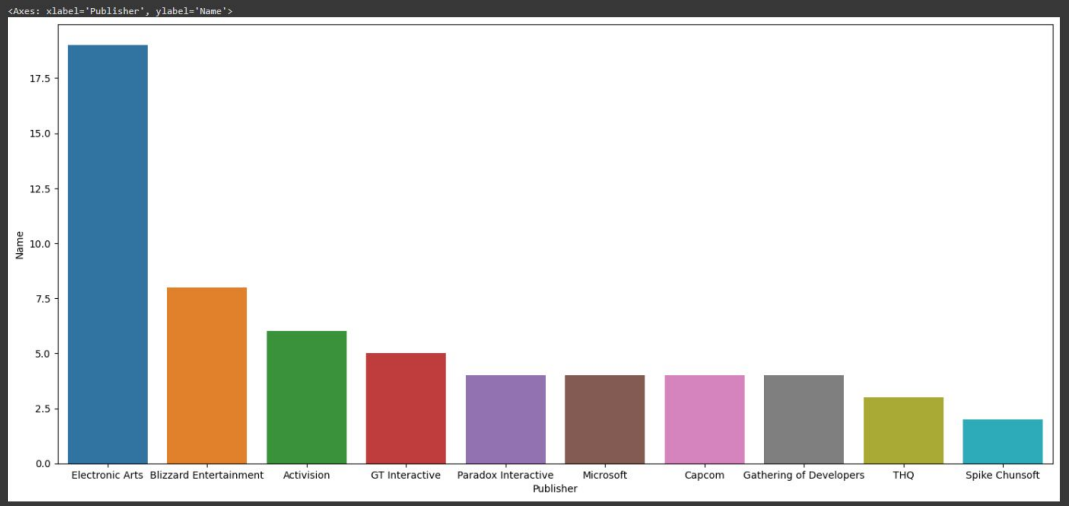
```
[8] agg_publisher = df.groupby('Publisher', as_index=False)['Name'].nunique()
    agg_publisher.sort_values('Name', ascending=False)
```

| | Publisher | Name |
|---|---|---|
| 20 | Electronic Arts | 19 |
| 7 | Blizzard Entertainment | 8 |
| 2 | Activision | 6 |
| 25 | GT Interactive | 5 |
| 43 | Paradox Interactive | 4 |
| ... | ... | ... |
| 29 | Impressions Game | 1 |
| 30 | Infogrames | 1 |
| 31 | Infogrames / Atari | 1 |
| 33 | Konami | 1 |
| 68 | id Software | 1 |

69 rows × 2 columns

The answer is Electronic Arts with 19 published games.

```
# data visualization
plt.rcParams["figure.figsize"] = (18,8)
sns.barplot(x='Publisher',y='Name', data = agg_publisher.sort_values('Name', ascending=False).head(10))
```




Can aggregate/group by publisher and distinct count the Name of games. Answer is Electronic Arts with 19 published games.

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Visualization

Which developer developed most of the games?

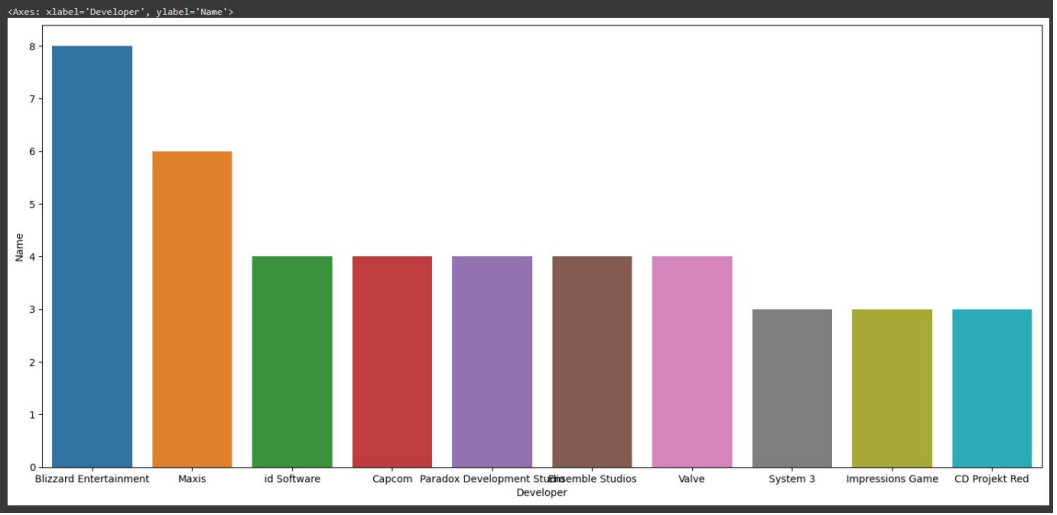
```
[10] agg_dev = df.groupby('Developer', as_index=False)['Name'].nunique()
     agg_dev.sort_values('Name', ascending=False)
```

| | Developer | Name |
|---|---|---|
| 4 | Blizzard Entertainment | 8 |
| 41 | Maxis | 6 |
| 80 | id Software | 4 |
| 10 | Capcom | 4 |
| 49 | Paradox Development Studio | 4 |
| ... | ... | ... |
| 47 | Namco Bandai Games | 1 |
| 48 | Origin Systems | 1 |
| 50 | PlatinumGames | 1 |
| 12 | Chris Sawyer | 1 |
| 0 | 3D Realms | 1 |

81 rows × 2 columns

The answer is Blizzard Entertainment with 8 developed games

```
[11] # data visualization
     sns.barplot(x='Developer', y='Name', data = agg_dev.sort_values('Name', ascending=False).head(10))
```




Can aggregate/group by developer and distinct count the Name of games. Answer is Blizzard Entertainment with 8 developed games.

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Visualization

Which series has the most sales?

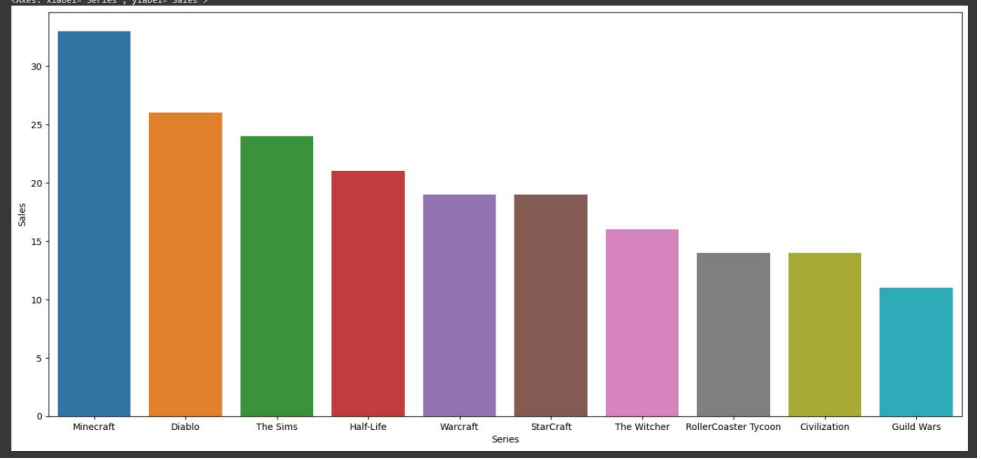
```
[12] agg_seriess = df.groupby('Series', as_index=False).agg({'Sales':'sum','Name':'nunique'})
     agg_seriess.sort_values('Sales', ascending=False)
```

| | Series | Sales | Name |
|---|---|---|---|
| 47 | Minecraft | 33.0 | 1 |
| 22 | Diablo | 26.0 | 3 |
| 75 | The Sims | 24.0 | 3 |
| 36 | Half-Life | 21.0 | 2 |
| 85 | Warcraft | 19.0 | 3 |
| ... | ... | ... | ... |
| 60 | RoboCop | 1.0 | 1 |
| 46 | Microsoft Flight Simulator | 1.0 | 1 |
| 58 | Railroad Tycoon | 1.0 | 1 |
| 56 | Psychonauts | 1.0 | 1 |
| 90 | Zork | 1.0 | 1 |

91 rows × 3 columns

The answer is Minecraft with 33 sales

```
[13] # data visualization
     sns.barplot(x='Series', y='Sales', data = agg_seriess.sort_values('Sales', ascending=False).head(10))
```




The answer is Minecraft with 33 sales

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb

# Data Visualization

Which series have the most games?

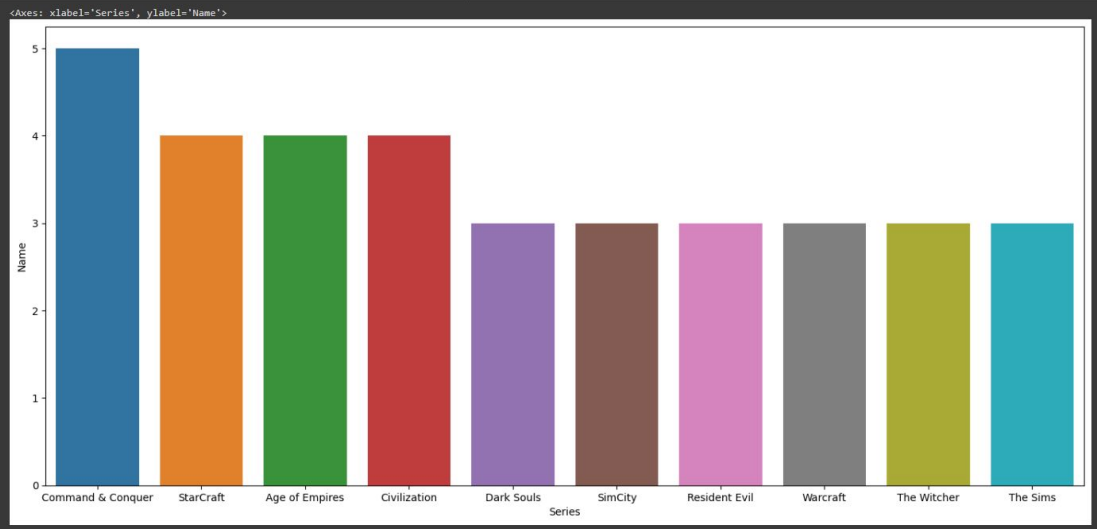
```
[ ] agg_series = df.groupby('Series', as_index=False)['Name'].nunique()
    agg_series.sort_values('Name', ascending=False)
```

| | Series | Name |
|---|---|---|
| 13 | Command & Conquer | 5 |
| 68 | StarCraft | 4 |
| 2 | Age of Empires | 4 |
| 12 | Civilization | 4 |
| 20 | Dark Souls | 3 |
| ... | ... | ... |
| 41 | International Karate | 1 |
| 40 | Hydlide | 1 |
| 38 | Hearts of Iron | 1 |
| 37 | Harry Potter | 1 |
| 90 | Zork | 1 |

91 rows × 2 columns

The answer is Command & Conquer with 5 games

```
[15] sns.barplot(x='Series', y='Name', data = agg_series.sort_values('Name', ascending=False).head(10))
```




The answer is Command & Conquer with 5 games

https://github.com/alamfatih/DataAnalytics-Portfolio-Fatih/blob/main/CaseStudy(Games_Sales).ipynb


# Thank you

Feel free to give me criticism and suggestions