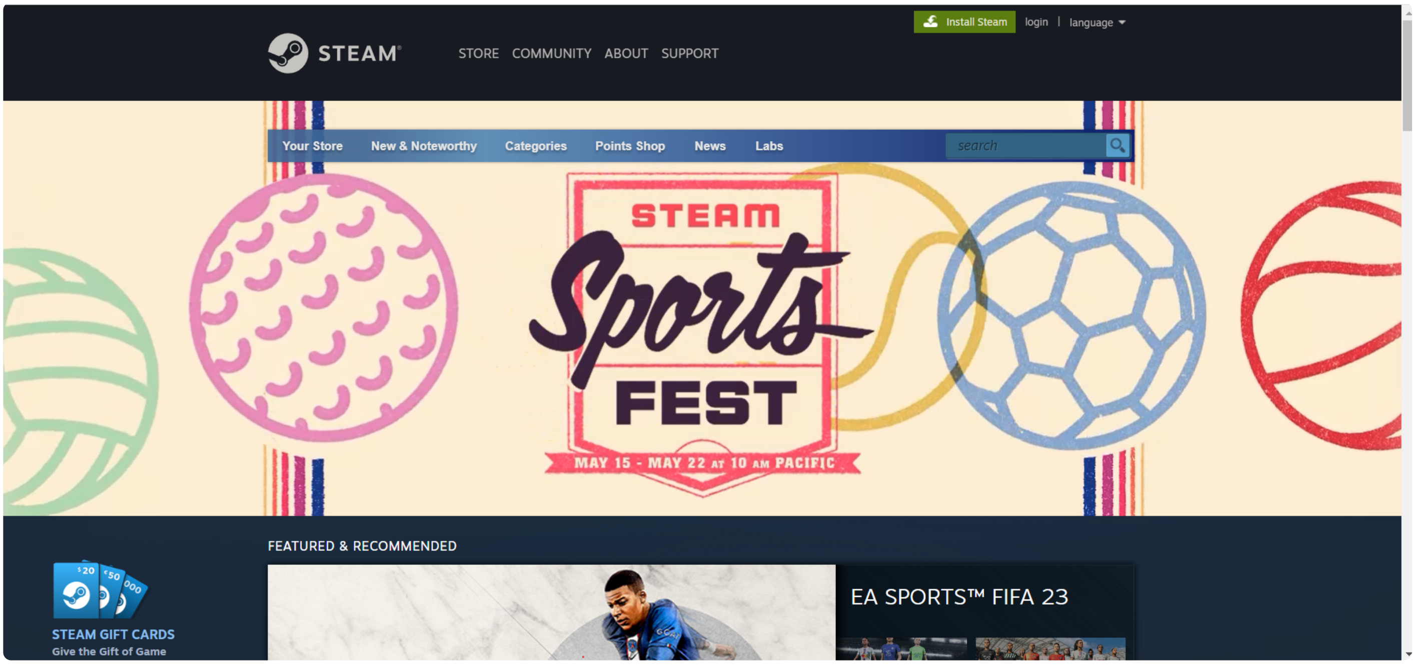
Exploratory Data Analysis and Visualization of Steam Games Dataset



Steam is a digital platform created by Valve Corporation to serve as a distributor of PC games. The Steam client allows users to install PC games online directly to their cloud drives after purchase.

What is Exploratory Data Analysis?

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Here's an outline of steps we'll follow:

1. Select a real-world dataset.
2. Download dataset using opendataset.
3. Perform data preparation & cleaning.
4. Perform exploratory analysis & visualization.
5. Ask & answer questions about the data.
6. Summary and Conclusion.

Download selected dataset from kaggle using opendatasets

We are installing opendatasets library to download steam dataset

Numpy and pandas are also installed using pip to analise and filter data

```
!pip install numpy pandas-profiling jovian --upgrade --quiet
```

```
import pandas as pd
import numpy as np
import jovian
```

```
!pip install opendatasets --upgrade --quiet
```

```
import opendatasets as od
```

Steam games dataset url is named as steam_games_url

```
steam_games_url = 'https://www.kaggle.com/datasets/mikekzan/steam-games-dlcs?select=ste
```

```
od.download(steam_games_url)
```

```
Skipping, found downloaded files in "./steam-games-dlcs" (use force=True to force
download)
```

```
steam_games_csv = "steam-games-dlcs/steam.csv"
```

We are now reading that csv file as steam_games_df

```
steam_games_df = pd.read_csv(steam_games_csv)
```

Steam games data frame is as shown below

```
steam_games_df
```

	appid	type	name	required_age	dlc	fullgame	supported_languages	developers	publish
0	10	game	Counter-Strike	0	NaN	NaN	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Va
1	20	game	Team Fortress Classic	0	NaN	NaN	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Va
2	30	game	Day of Defeat	0	NaN	NaN	['English', 'French', 'German', 'Italian', 'Sp...	['Valve']	['Va
3	40	game	Deathmatch Classic	0	NaN	NaN	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Va
4	50	game	Half-Life: Opposing Force	0	NaN	NaN	['English', 'French', 'German', 'Korean']	['Gearbox Software']	['Va
...	...	...	...	...	...	...	...	...	
102499	2028023	dlc	Total War Saga: FALL OF THE SAMURAI – Blood Pack	18	NaN	{'appid': '201271', 'name': 'A Total War Saga:...	['Czech', 'English', 'French', 'German', 'Ital...	['The Creative Assembly']	['SE

	appid	type	name	required_age	dlc	fullgame	supported_languages	developers	publish
102500	2028055	dlc	Tom Clancy's Ghost Recon Future Soldier - Seas...	0	NaN	{'appid': '212630', 'name': "Tom Clancy's Ghos...	['Danish', 'Dutch', 'English', 'French', 'Germ...	['Ubisoft Paris', 'Red Storm Entertainment']	['Ubis
102501	2028056	dlc	Worms Revolution Season Pass	0	NaN	{'appid': '200170', 'name': 'Worms Revolution'}	['English', 'French', 'German', 'Italian', 'Po...	['Team17 Digital Ltd.']	['Tean Digital I
102502	2028062	dlc	Call of Duty®: Black Ops II Season Pass	0	NaN	{'appid': '202970', 'name': 'Call of Duty®: Bl...	['English', 'French', 'German', 'Italian', 'Sp...	['Treyarch']	['Activisi
102503	2028850	dlc	Bioshock Infinite: Columbia's Finest	17	NaN	{'appid': '8870', 'name': 'BioShock Infinite'}	['English', 'French', 'German', 'Italian', 'Ja...	['Irrational Games', 'Virtual Programming (Lin...	['

102504 rows × 26 columns

dataframe `.columns` function provides information about the columns

```
steam_games_df.columns
```

```
Index(['appid', 'type', 'name', 'required_age', 'dlc', 'fullgame',
       'supported_languages', 'developers', 'publishers', 'packages',
       'platforms', 'categories', 'genres', 'achievements', 'release_date',
       'supported_audio', 'coming_soon', 'price', 'review_score',
       'total_positive', 'total_negative', 'rating', 'owners',
       'average_forever', 'median_forever', 'tags'],
      dtype='object')
```

`.info` function provides data about the column,, non-null count and data type

```
steam_games_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102504 entries, 0 to 102503
Data columns (total 26 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   appid                102504 non-null  int64 
 1   type                 102504 non-null  object
 2   name                 102504 non-null  object
 3   required_age         102504 non-null  int64 
 4   dlc                  9696 non-null    object
 5   fullgame             34607 non-null   object
 6   supported_languages  102352 non-null  object
 7   developers           102463 non-null  object
 8   publishers           102464 non-null  object
 9   packages             81153 non-null   object
 10  platforms            102504 non-null  object
 11  categories           102398 non-null  object
 12  genres               102311 non-null  object
 13  achievements         102504 non-null  float64
 14  release_date         95676 non-null   object
 15  supported_audio      49698 non-null   object
 16  coming_soon          102504 non-null  bool
 17  price                90670 non-null   float64
 18  review_score         102504 non-null  float64
 19  total_positive       102504 non-null  float64
 20  total_negative       102504 non-null  float64
 21  rating               102504 non-null  float64
 22  owners               102504 non-null  object
 23  average_forever      102504 non-null  int64
 24  median_forever       102504 non-null  int64
 25  tags                 61048 non-null   object
dtypes: bool(1), float64(6), int64(4), object(15)
memory usage: 19.6+ MB
```

`.describe` function provides information about mathematical information for numeric columns

```
steam_games_df.describe()
```

	appid	required_age	achievements	price	review_score	total_positive	total_negati
count	1.025040e+05	102504.000000	102504.000000	90670.000000	102504.000000	1.025040e+05	102504.00000
mean	1.082187e+06	1.423896	13.790594	6.847773	2.303715	4.754627e+02	73.1602
std	5.137251e+05	312.349047	155.164937	10.865827	3.199680	1.257478e+04	2906.3706
min	1.000000e+01	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.0000
25%	6.507375e+05	0.000000	0.000000	1.590000	0.000000	0.000000e+00	0.0000
50%	1.092725e+06	0.000000	0.000000	3.990000	0.000000	2.000000e+00	0.0000
75%	1.523995e+06	0.000000	8.000000	8.990000	6.000000	1.700000e+01	5.0000
max	2.028850e+06	99999.000000	9821.000000	999.000000	9.000000	2.949363e+06	733480.0000

After going through the dataframe i have selected these columns for visualization

```
selected_cols = ['type', 'name', 'required_age', 'supported_languages', 'developers',
                 'publishers', 'platforms', 'categories', 'genres', 'achievements', 're
                 'supported_audio', 'review_score', 'total_positive', 'total_negative',
```

For some numeric columns size of the dtype has been decreased inorder to decrease the size of the dataframe

```
selected_dtypes = {
    'requires_age' : 'int16',
    'achievements' : 'float16',
    'review_score' : 'float16',
    'total_positive' : 'float32',
    'total_negative' : 'float32',
    'rating' : 'float16'
}
```

After making those change we are reading our required data

```
steam_games_df1 = pd.read_csv(steam_games_csv,
                              usecols = selected_cols,
                              dtype = selected_dtypes,
                              parse_dates = ['release_date'])
```

```
steam_games_df1
```

	type	name	required_age	supported_languages	developers	publishers	platforms	categori
0	game	Counter-Strike	0	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Valve']	['windows', 'mac', 'linux']	['Multi-playe 'PvP', 'Onlir PvI 'Shared/
1	game	Team Fortress Classic	0	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Valve']	['windows', 'mac', 'linux']	['Multi-playe 'PvP', 'Onlir PvI 'Shared/
2	game	Day of Defeat	0	['English', 'French', 'German', 'Italian', 'Sp...	['Valve']	['Valve']	['windows', 'mac', 'linux']	['Multi-playe 'Valve An Che: enablec
3	game	Deathmatch Classic	0	['English', 'French', 'German', 'Italian', 'Ko...	['Valve']	['Valve']	['windows', 'mac', 'linux']	['Multi-playe 'PvP', 'Onlir PvI 'Shared/
4	game	Half-Life: Opposing Force	0	['English', 'French', 'German', 'Korean']	['Gearbox Software']	['Valve']	['windows', 'mac', 'linux']	['Single-playe 'Multi-playe 'Valve Anti-
...	...	...	...	...	...	...	...	
102499	dlc	Total War Saga: FALL OF THE SAMURAI – Blood Pack	18	['Czech', 'English', 'French', 'German', 'Ital...	['The Creative Assembly']	['SEGA']	['windows']	['Single-playe 'Multi-playe 'Co-op', 'Do
102500	dlc	Tom Clancy's Ghost Recon Future Soldier - Seas...	0	['Danish', 'Dutch', 'English', 'French', 'Germ...	['Ubisoft Paris', 'Red Storm Entertainment']	['Ubisoft']	['windows']	['Single-playe 'Multi-playe 'Co-op', 'Do
102501	dlc	Worms Revolution Season Pass	0	['English', 'French', 'German', 'Italian', 'Po...	['Team17 Digital Ltd.']	['Team17 Digital Ltd']	['windows']	['Single-playe 'Multi-playe 'Co-op', 'Sh

	type	name	required_age	supported_languages	developers	publishers	platforms	categori
102502	dlc	Call of Duty®: Black Ops II Season Pass	0	['English', 'French', 'German', 'Italian', 'Sp...	['Treyarch']	['Activision']	['windows']	['Single-playe 'Multi-playe 'Co-op', 'Do
102503	dlc	Bioshock Infinite: Columbia's Finest	17	['English', 'French', 'German', 'Italian', 'Ja...	['Irrational Games', 'Virtual Programming (Lin...	['2K']	['windows', 'linux']	['Single-playe 'Downloadab Content', 'Ste

102504 rows × 17 columns

Handling missing & duplicate data

Missing data in Pandas is indicated using np.nan. We can find the number of missing values in each column of a dataframe using the following expression:

.isna function provides information about the count of nan values, based on that we have to filter the data

```
steam_games_df1.isna().sum()
```

```
type                      0
name                      0
required_age              0
supported_languages     152
developers               41
publishers               40
platforms                 0
categories              106
genres                  193
achievements              0
release_date           6828
supported_audio       52806
coming_soon               0
review_score              0
total_positive            0
total_negative            0
rating                    0
dtype: int64
```

For some nan values we have to remove the row which are having nan values for better understanding

```
steam_games_df1.drop(steam_games_df1[steam_games_df1.required_age > 100].index, inplace
```

```
sample_df = steam_games_df1.dropna(subset = ['supported_audio', 'release_date', 'genres
```

Replace the missing values in the columns `developers and publishers` using the most common value in each column.

```
most_common_developers = sample_df.developers.mode()[0]
most_common_publishers = sample_df.publishers.mode()[0]
```

```
sample_df.developers.fillna(most_common_developers, inplace = True)
sample_df.publishers.fillna(most_common_publishers, inplace = True)
```

```
/opt/conda/lib/python3.9/site-packages/pandas/core/generic.py:6392:
SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
sample_df.isna().sum()
```

```
type                  0
name                  0
required_age          0
supported_languages   0
developers            0
publishers            0
platforms             0
categories            0
genres                0
achievements          0
release_date          0
supported_audio       0
coming_soon           0
review_score          0
total_positive        0
total_negative        0
rating                0
dtype: int64
```

Creating seperate columns for year and month from release_date column

```
sample_df['Year'] = sample_df.release_date.dt.year
sample_df['Month'] = sample_df.release_date.dt.month
```

```
/tmp/ipykernel_46/2826042935.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2826042935.py:2: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

As of now we are in 2023 so data which is present above 2023 is not correct, so we are removing data greater than 2023

```
sample_df.drop(sample_df[sample_df.Year > 2022].index, inplace=True)
```

```
/opt/conda/lib/python3.9/site-packages/pandas/core/frame.py:4906:
SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Removing [] from the columns which are containing them

```
sample_df['supported_languages'] =sample_df.supported_languages.str.strip('[ ]')
sample_df['developers'] =sample_df.developers.str.strip('[ ]')
sample_df['publishers'] =sample_df.publishers.str.strip('[ ]')
sample_df['platforms'] =sample_df.platforms.str.strip('[ ]')
sample_df['categories'] =sample_df.categories.str.strip('[ ]')
sample_df['supported_audio'] =sample_df.supported_audio.str.strip('[ ]')
sample_df['genres'] =sample_df.genres.str.strip('[ ]')
```

```
/tmp/ipykernel_46/2907351487.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:2: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:3: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:4: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:5: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:6: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_46/2907351487.py:7: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
sample_df['games'] = 'games'
```

```
/tmp/ipykernel_46/2554607113.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
sample_df
```

	type	name	required_age	supported_languages	developers	publishers	platforms	cat
0	game	Counter-Strike	0	'English', 'French', 'German', 'Italian', 'Kor...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Multi- 'PvP', 'Shar
6	game	Half-Life	0	'English', 'French', 'German', 'Italian', 'Kor...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- 'Multi- 'PvP', '
9	game	Half-Life 2	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achieven 'St
16	game	Half-Life 2: Episode One	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achieven 'C
17	game	Portal	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achieven 'C
...	...	...	...	...	...	...	...	

	type	name	required_age	supported_languages	developers	publishers	platforms	cat
102474	game	SailSim	0	'English'	'Demetris Rouslan Zavorotnitsienko'	'Demetris Rouslan Zavorotnitsienko'	'windows'	'Single
102493	game	Penguins Can Fly	0	'English'	'Joey Cook'	'Joey Cook'	'windows'	'Single- 'Multi- 'Co-op
102494	game	DarkSelf: Other Mind	18	'English', 'Portuguese'	'TiagoChefe Studio'	'TiagoChefe Studio'	'windows'	'Single- 'Ca ava
102500	dlc	Tom Clancy's Ghost Recon Future Soldier - Seas...	0	'Danish', 'Dutch', 'English', 'French', 'Germa...	'Ubisoft Paris', 'Red Storm Entertainment'	'Ubisoft'	'windows'	'Single- 'Multi- 'Co-op',
102501	dlc	Worms Revolution Season Pass	0	'English', 'French', 'German', 'Italian', 'Pol...	'Team17 Digital Ltd.'	'Team17 Digital Ltd'	'windows'	'Single- 'Multi- 'Co-op'

45681 rows × 20 columns

```
steam_df = sample_df.copy()
```

Performing exploratory analysis & visualization

To begin, let's install and import the libraries. We'll use the matplotlib.pyplot module for basic plots like line & bar charts. It is often imported with the alias plt. We'll use the seaborn module for more advanced plots. It is commonly imported with the alias sns

```
!pip install matplotlib seaborn plotly --upgrade --quiet
```

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import plotly.express as px
```

What is the highest rating that a game got?

```
steam_df.rating.max().round()
```

```
98.0
```

98 of 100 is the highest rating that a game has got

Game which has got the highest positive rate?

```
highest_positive_rate = steam_df.sort_values('total_positive', ascending = False).head(
highest_positive_rate[['name', 'total_positive']]
```

	name	total_positive
25	Counter-Strike: Global Offensive	2949363.0
4604	Grand Theft Auto V	1030822.0
9323	Tom Clancy's Rainbow Six® Siege	777553.0
141	Garry's Mod	679681.0
3722	Rust	553246.0

`Counter-Strike: Global Offensive` has got highest number of likes

Which game has got highest rating?

```
highest_rating = steam_df.sort_values('rating', ascending = False).head(10)
highest_rating[['name', 'rating']]
```

	name	rating
23	Portal 2	97.6250
54655	Hades	97.3750
17	Portal	96.8750
60227	ULTRAKILL	96.8125
19727	Half-Life: Alyx	96.7500
51079	The Henry Stickmin Collection	96.6250
21	Left 4 Dead 2	96.5625
5360	The Witcher® 3: Wild Hunt	96.5625
3663	The Binding of Isaac: Rebirth	96.5000
30576	Phasmophobia	96.2500

`Portal 2` is the highest rated game in steam

Which game has got the least rating?

```
least_rating = steam_df.sort_values('rating', ascending = True).head(10)
least_rating[['name', 'rating']]
```

	name	rating
10504	Jurassic Island: The Dinosaur Zoo	17.078125
38433	CODE VEIN: Frozen Empress	17.343750
96285	Tricolour Lovestory TrueEnd	17.859375
4670	Game Tycoon 1.5	18.218750
55242	XIII	18.640625
38434	CODE VEIN: Lord of Thunder	18.796875
17800	Sid Meier's Civilization® VI: Vikings Scenario...	19.328125
40607	NEW LIFE	20.234375
64390	Blood Bowl 2 - DEATH ZONE	20.953125
21350	Tom Clancy's Ghost Recon® Wildlands - Narco Road	21.203125

`Jurassic Island: The Dinosaur Zoo` is the least rated game in steam games

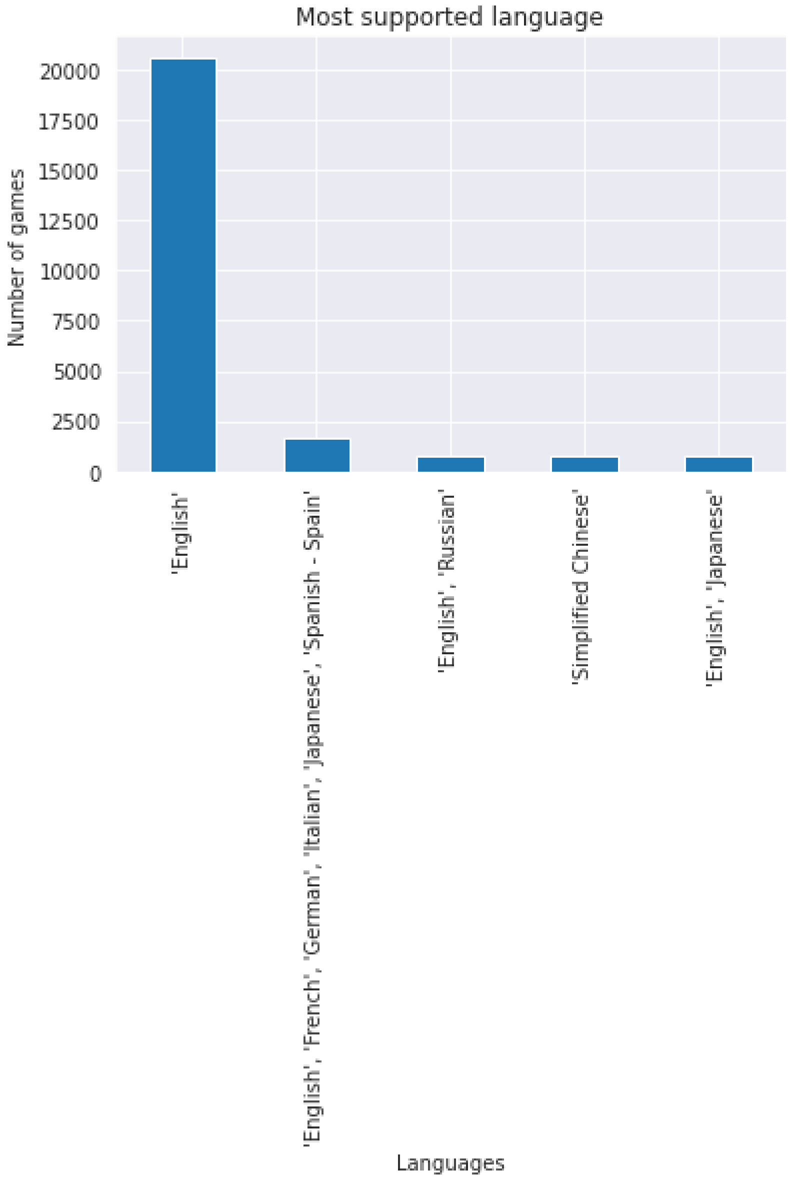
```
steam_df
```

	type	name	required_age	supported_languages	developers	publishers	platforms	cate
0	game	Counter-Strike	0	'English', 'French', 'German', 'Italian', 'Kor...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Multi- 'PvP', 'Shai
6	game	Half-Life	0	'English', 'French', 'German', 'Italian', 'Kor...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- 'Multi- 'PvP', '
9	game	Half-Life 2	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achiever 'St
16	game	Half-Life 2: Episode One	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achiever 'C
17	game	Portal	0	'Danish', 'Dutch', 'English', 'Finnish', 'Fren...	'Valve'	'Valve'	'windows', 'mac', 'linux'	'Single- Achiever 'C
...	...	...	...	...	...	...	...	
102474	game	SailSim	0	'English'	'Demetris Rouslan Zavorotnitsienko'	'Demetris Rouslan Zavorotnitsienko'	'windows'	'Single-
102493	game	Penguins Can Fly	0	'English'	'Joey Cook'	'Joey Cook'	'windows'	'Single- 'Multi- 'Co-op
102494	game	DarkSelf: Other Mind	18	'English', 'Portuguese'	'TiagoChefe Studio'	'TiagoChefe Studio'	'windows'	'Single- 'Ca ava
102500	dlc	Tom Clancy's Ghost Recon Future Soldier - Seas...	0	'Danish', 'Dutch', 'English', 'French', 'Germa...	'Ubisoft Paris', 'Red Storm Entertainment'	'Ubisoft'	'windows'	'Single- 'Multi- 'Co-op',
102501	dlc	Worms Revolution Season Pass	0	'English', 'French', 'German', 'Italian', 'Pol...	'Team17 Digital Ltd.'	'Team17 Digital Ltd'	'windows'	'Single- 'Multi- 'Co-op'

45681 rows × 20 columns

What is the most supported language?

```
most_supportred_languages = steam_df.supported_languages.value_counts().head()
most_supportred_languages.plot(kind = 'bar')
plt.title('Most supported language')
plt.xlabel('Languages')
plt.ylabel('Number of games');
```



English is the most supported_language in steam games

In which year does more number of games have been launched?

```
yearly_release = steam_df.groupby('Year')[['name']].count().sort_values('name', ascendi
yearly_release
```

```
      name
Year
2021  9180
2020  6719
2019  5849
2018  5462
2022  4715
2017  4686
2016  3433
2015  2403
2014  1660
```

	name
Year	
2013	831

In the year `2021` more number of games are launched by steam

Which genre has got more number of positive likes?

```
genre_likes = steam_df.groupby('genres')[['total_positive']].count().sort_values('total
genre_likes
```

	total_positive
genres	
'Action'	3932
'Action', 'Casual', 'Indie'	2567
'Action', 'Indie'	2015
'Casual', 'Simulation'	1945
'Action', 'Adventure', 'Indie'	1733
'Adventure', 'Indie'	1447
'Casual', 'Indie'	1310
'Action', 'Adventure'	1182
'RPG'	999
'Simulation'	959

`Action` genre has got more number of positive likes.

Here we have removed 0 value from numeric columns for better visualizatio of treemaps or sunburst maps

```
non_zero_rating = steam_df.copy()
```

```
non_zero_rating.drop(non_zero_rating[non_zero_rating.rating == 0].index, inplace = True
non_zero_rating.drop(non_zero_rating[non_zero_rating.total_positive == 0].index, inplac
non_zero_rating.drop(non_zero_rating[non_zero_rating.achievements == 0].index, inplace 
non_zero_rating.drop(non_zero_rating[non_zero_rating.total_negative == 0].index, inplac
non_zero_rating.drop(non_zero_rating[non_zero_rating.total_negative == 0].index, inplac
```

Which genere type games are more in number?

```
genre_df= steam_df.groupby('genres')[['name']].count().sort_values('name', ascending=Fa
genre_df
```

	name
genres	
'Action'	3932
'Action', 'Casual', 'Indie'	2567

	name
genres	
'Action', 'Indie'	2015
'Casual', 'Simulation'	1945
'Action', 'Adventure', 'Indie'	1733
'Adventure', 'Indie'	1447
'Casual', 'Indie'	1310
'Action', 'Adventure'	1182
'RPG'	999
'Simulation'	959

Action genre type games are more in number

Which type of games has got more positive likes?

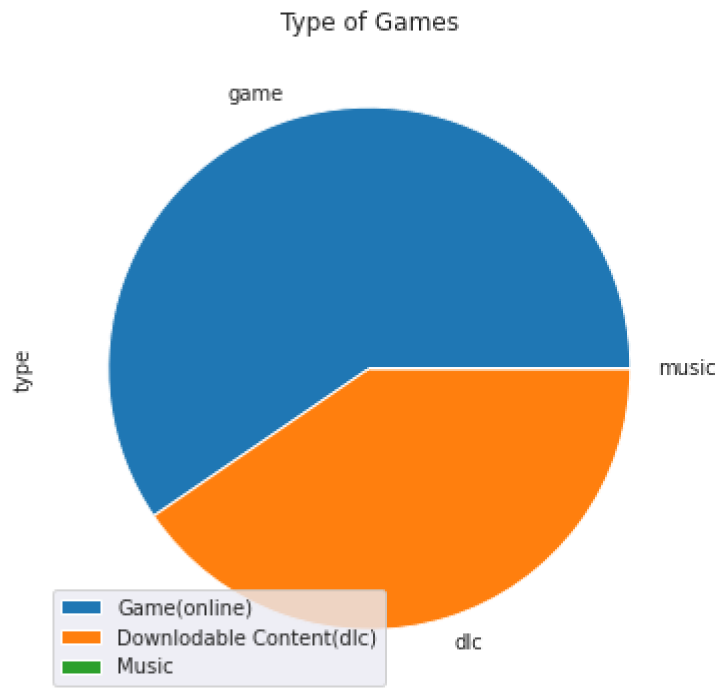
```
steam_df.groupby('type')[['total_positive']].count().sort_values('total_positive', asce
```

	total_positive
type	
game	27201
dlc	18479
music	1

Ask & answer questions about the data

1. Which type of games are more in number wether online or downlodable contents (dlc)?

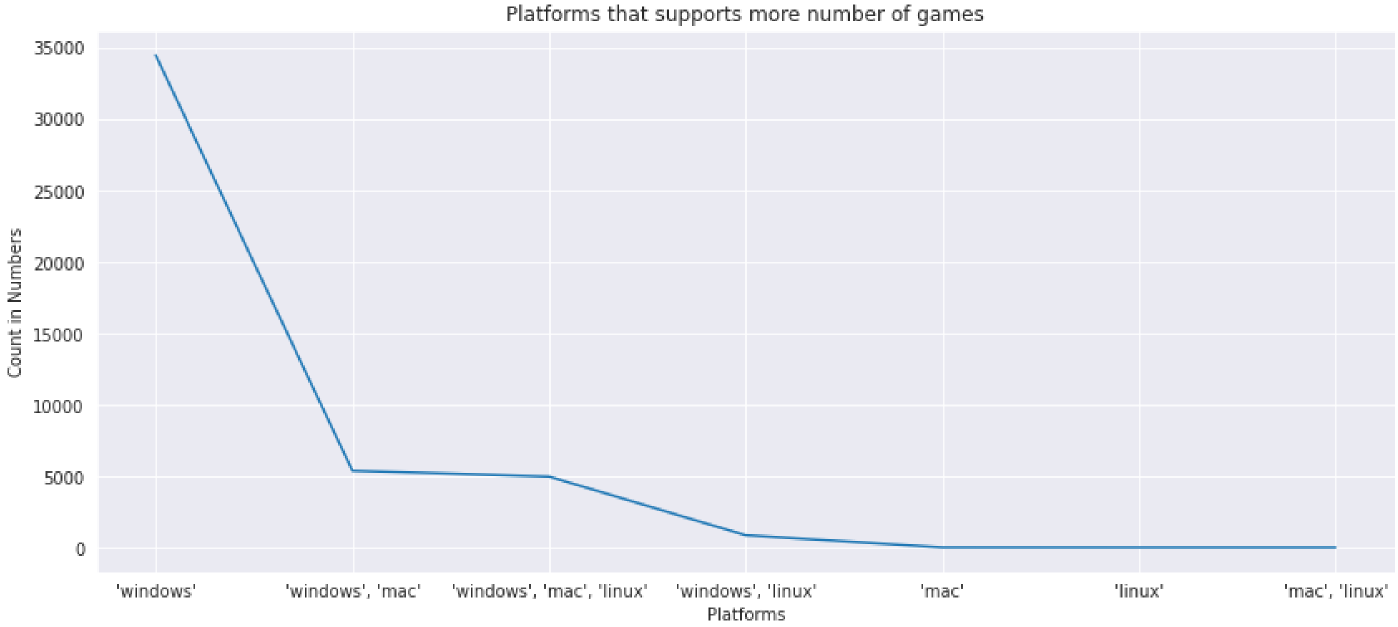
```
plt.figure(figsize=(10, 6))
steam_df.type.value_counts().head(10).plot(kind = 'pie')
plt.title('Type of Games')
plt.legend(['Game(online)', 'Downlodable Content(dlc)', 'Music']);
```



In steam games there are more number of online games than downloadable contents

2. Which operating system does most of the games supports?

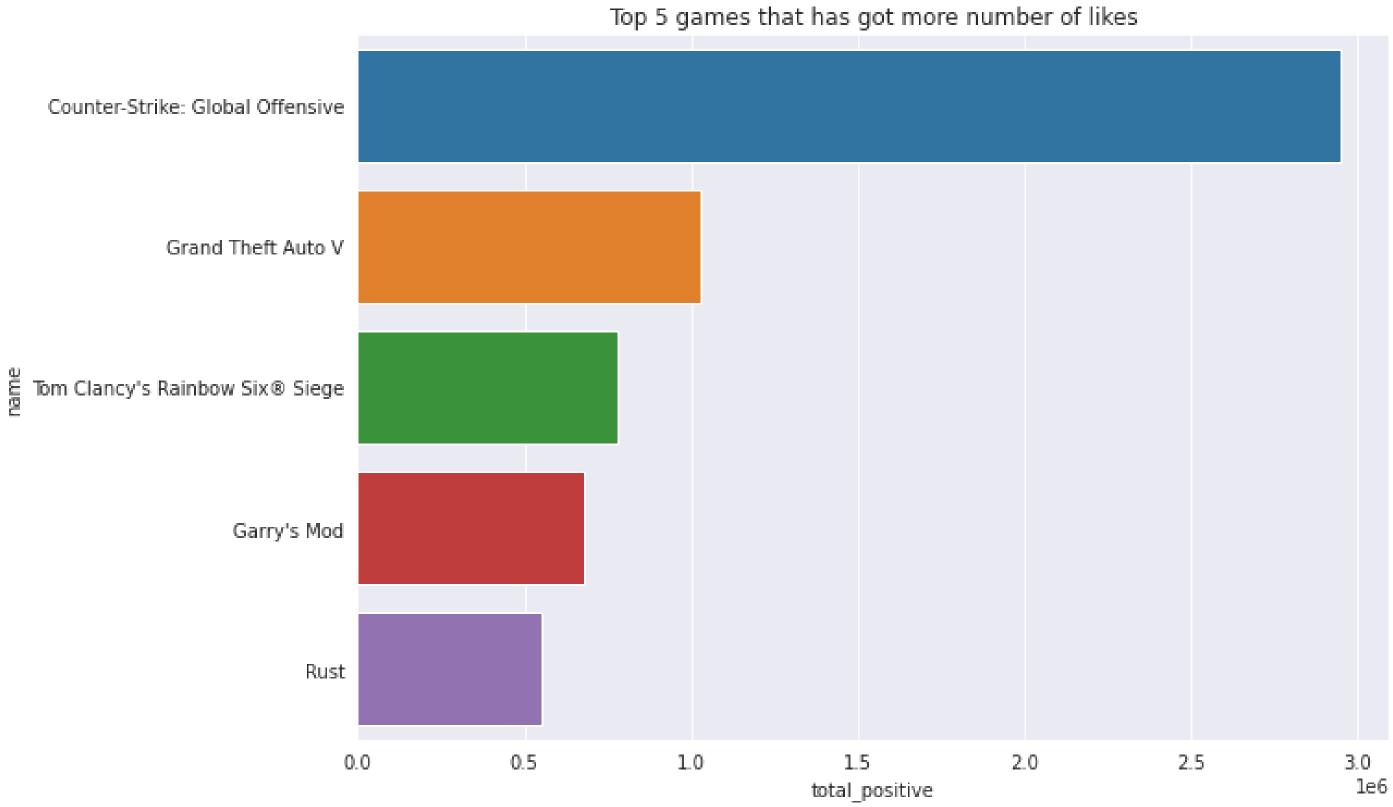
```
plt.figure(figsize=(14, 6))
steam_df.platforms.value_counts().plot(kind = 'line')
plt.xlabel('Platforms')
plt.ylabel('Count in Numbers')
plt.title('Platforms that supports more number of games');
```



Windows is the most supported operating system for many games

3. Which game has got the highest likes? which genre does it belongs to?

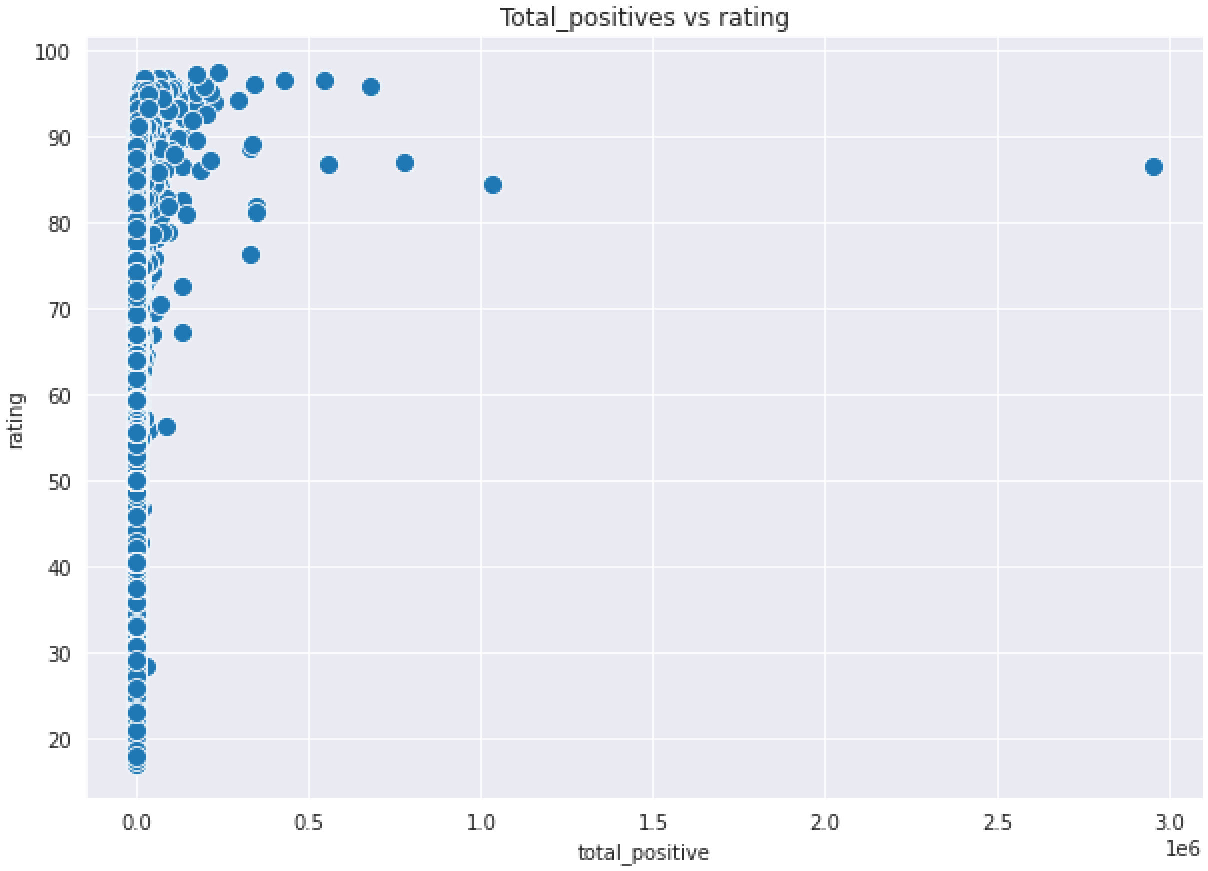
```
plt.figure(figsize = (10, 7))
sns.barplot(x = 'total_positive', y = 'name', data= highest_positive_rate)
plt.title('Top 5 games that has got more number of likes');
```



Counter strike : Global offensive has got more number of likes from the players, so that steam games can launch more versions of it to make profits.

4. Does high positive means good rating?

```
plt.figure(figsize = (10, 7))
sns.set_style("darkgrid")
sns.scatterplot(x = steam_df.total_positive, y = steam_df.rating, s = 100)
plt.title('Total_positives vs rating');
```

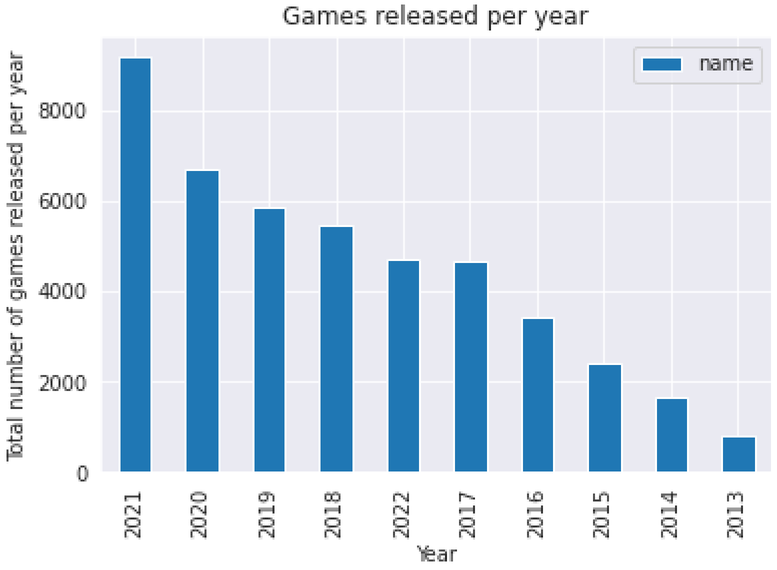


No, high rated games are not getting more likes because, Steam reviews are 50% reliable because there are unknown organizations that are paid to write false reviews. These false reviewers can be identified by the steam accounts that contain 1 to 10 paid games on the account in addition to a large majority of free to play games.

5. In which year does more number of games have been released?

```
plt.figure(figsize = (12, 7))
yearly_release.plot(kind = 'bar')
plt.ylabel('Total number of games released per year')
plt.title('Games released per year');
```

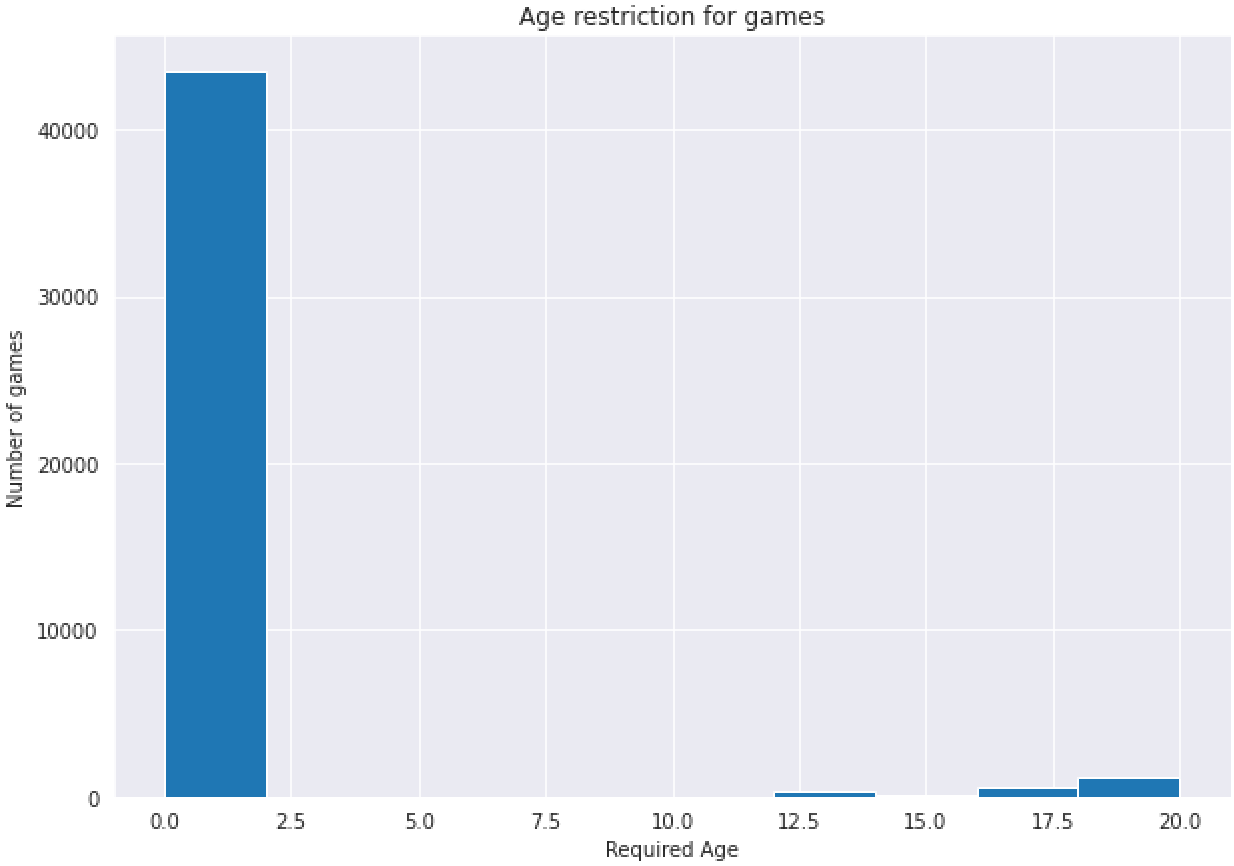
```
<Figure size 864x504 with 0 Axes>
```



From the graph we can conclude that number of games publishing per year are gradually increasing year by year but in 2022 publishings are suddenly droped down by 30%

6. Can anyone play games in steam games or is there any age restriction?

```
plt.figure(figsize = (10, 7))
plt.hist(steam_df.required_age)
plt.xlabel('Required Age')
plt.ylabel('Number of games')
plt.title('Age restriction for games');
```



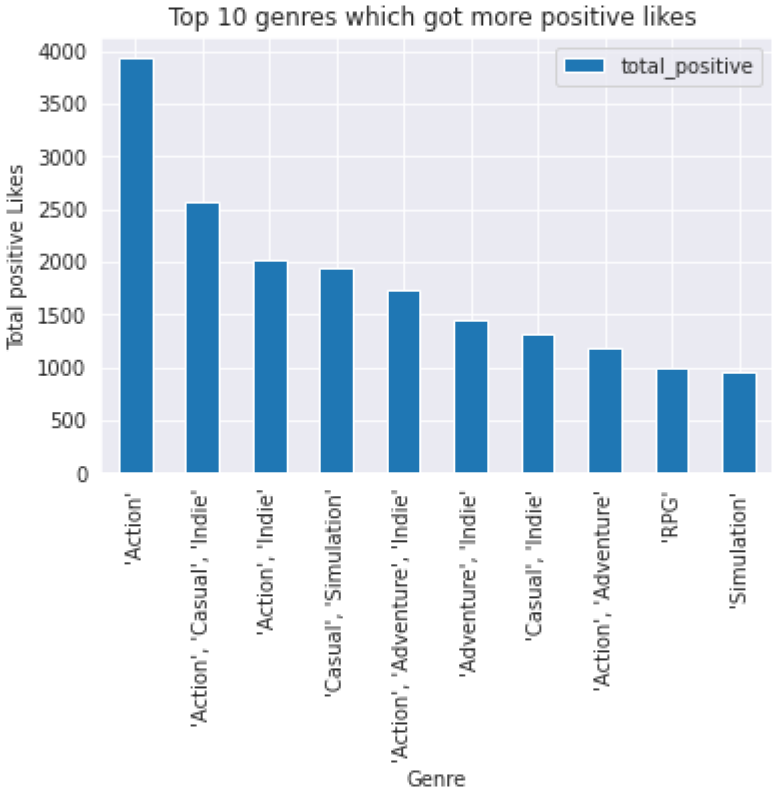
Yes, there is age restriction for few games but many of the games in steam games are having no age limits

7. Which type of genres are launching in more number and which type of them are getting more likes are they same?

```
plt.figure(figsize = (10, 7))
genre_likes.plot(kind = 'bar')
plt.xlabel('Genre')
plt.ylabel('Total positive Likes')
plt.title('Top 10 genres which got more positive likes ')
```

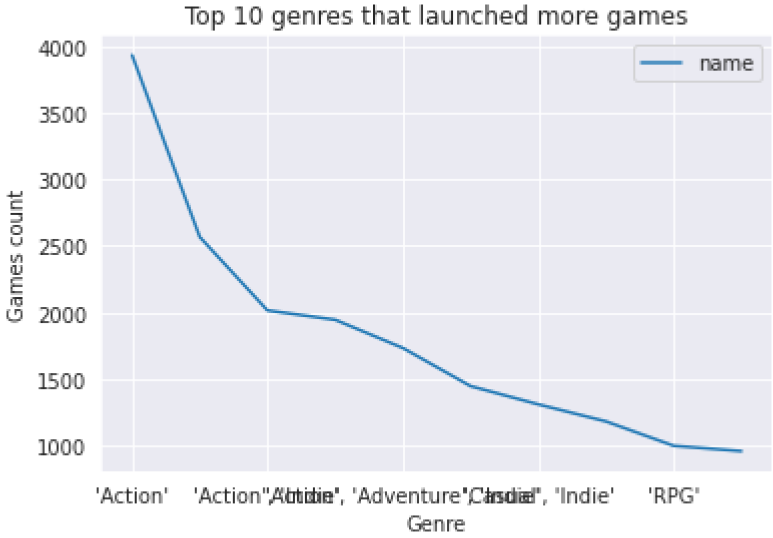
```
Text(0.5, 1.0, 'Top 10 genres which got more positive likes ')
```

```
<Figure size 720x504 with 0 Axes>
```



```
plt.figure(figsize=(14, 6))
genre_df.plot(kind ='line')
plt.xlabel('Genre')
plt.ylabel('Games count')
plt.title('Top 10 genres that launched more games');
```

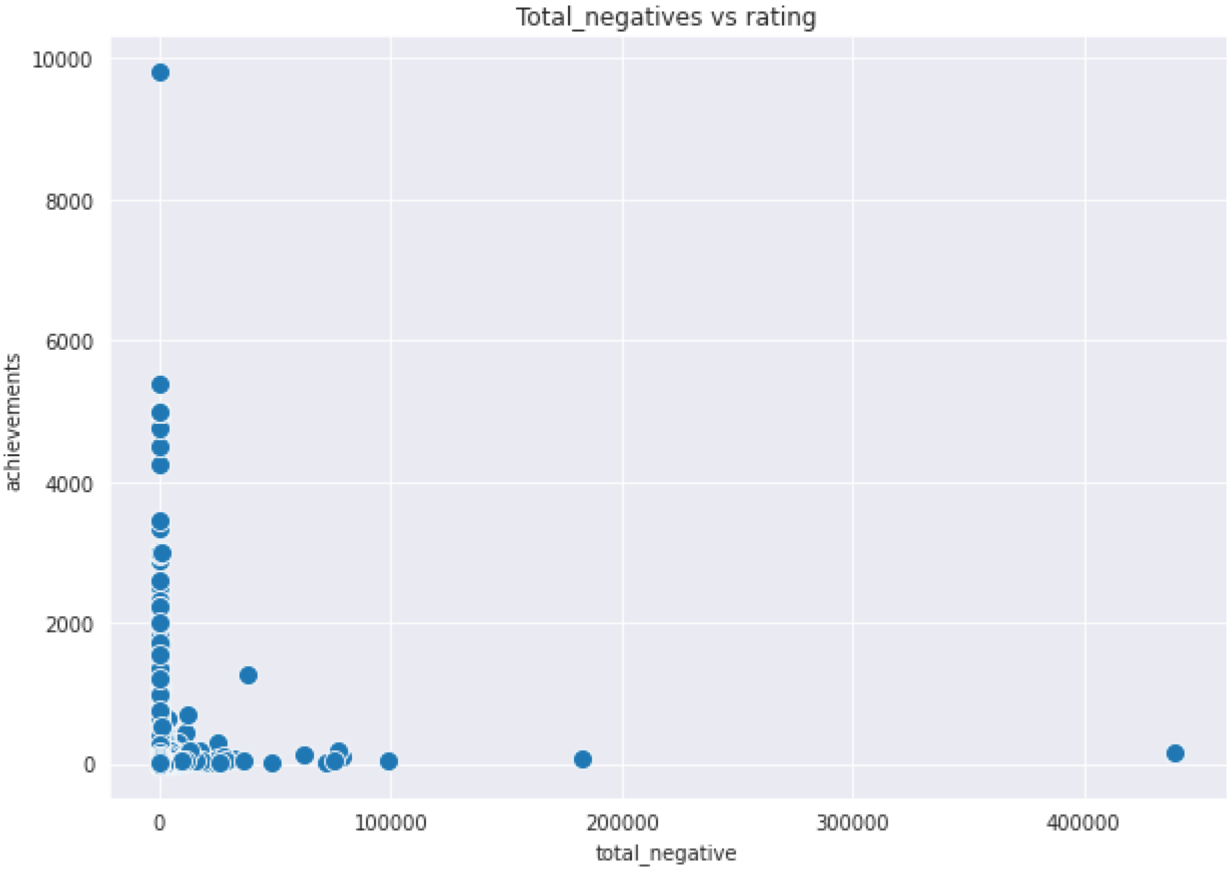
```
<Figure size 1008x432 with 0 Axes>
```



Steam games are launching more number of action games and players are also more intrested in action games

8. Does less negative means more achievements?

```
plt.figure(figsize = (10, 7))
sns.set_style("darkgrid")
sns.scatterplot(x = non_zero_rating.total_negative, y = non_zero_rating.achievements, s
plt.title('Total_negatives vs rating');
```

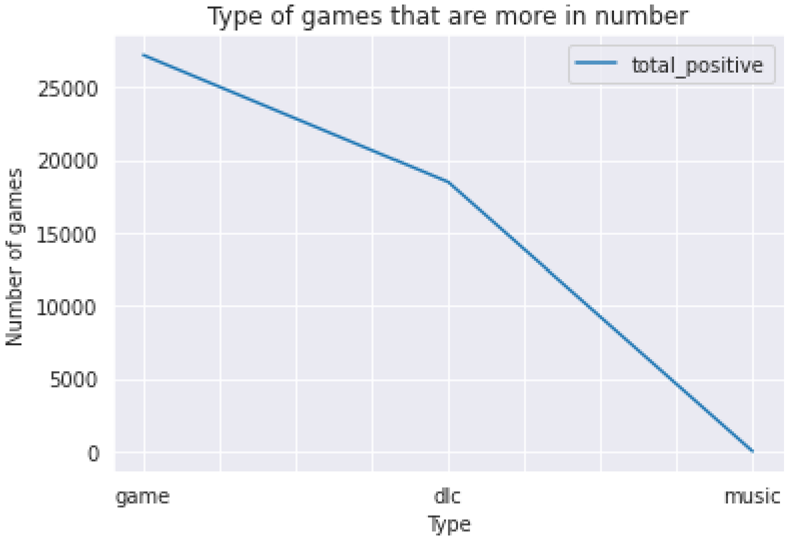


Yes, games with less negative likes are having more achievements

9. Does steam games is launching the type of games what players are liking most?

```
plt.figure(figsize=(10, 7))
steam_df.groupby('type')[['total_positive']].count().sort_values('total_positive', asce
plt.xlabel('Type')
plt.ylabel('Number of games')
plt.title('Type of games that are more in number');
```

```
<Figure size 720x504 with 0 Axes>
```



Yes, steam game is launching more number of online games which has got more positive likes from players

**10. Which type of genres are published more and which type of games has got more positive likes and ratings? **

```
fig = px.sunburst(non_zero_rating,
                  path=['games', 'genres', 'publishers'],
                  values='total_positive',
                  color='rating',
                  color_continuous_scale='RdBu',
                  title = 'Games that has got more positive likes and ratings based on t
fig.show()
```

There are more nuber of `action` games and they are published by `valve` , they have also got more positive likes

Summary and Conclusion

Here's what we have covered in this notebook:

1. Select a real-world dataset.
2. Download dataset using opendataset.
3. Perform data preparation & cleaning.
4. Perform exploratory analysis & visualization.
5. Ask & answer questions about the data.

From my observations

- In steam there are more number of `online games` , people are also liking more online games when compared with downlodable content.
- `English` is the most supported language for many games.
- Steam has age restrictions for players but for many games there is no age minimum age, but for few games maximum age restriction is 21.
- `Windows` is the most supported os for many games, mostly people are also using windows pc's if their browsing or downloads may decrease the they have to check whether people are using windows or mac os.
- Number of games publishing per year are gradually increasing year by year but in 2022 publishings are suddenly droped down by 30%.
- `Counter-Strike: Global Offensive` has got more positive likes from players which belongs to `action` genre and players also like more action games.
- High rated games does not get more number of positive likes, so high rating is not directly proporsnal to positive rating.
- There are more nuber of action games and they are published by valve, they have also got more positive likes, so that steam is launching what people like the most.

Future Scope

- Further we can analyse which games has got least likes and to which genre they belongs to.
- Some columns have not be analysed like publishers, packages, categories, supported audio, comming soon ...
- So that in future we can analyse them and can conclude the result from those observations.

References

- Kaggle
- Opendatasets
- Pandas Tutorial
- Numpy
- W3 School
- Visualization