Exploring the correlation between video game violence and game popularity

1 Introduction, Aims and Objectives

1.1 Introduction

In recent years the game industry has skyrocketed and the market for video games has been on a constant rise, expecting to expand as years go by. Hence, it is only natural that games and their players be analysed to see the type of games and its specialities that makes certain games popular among people. Violence is something that has become a kind of taboo in our society. It is frowned upon and condoned for obvious reasons, to keep the peace among people. However, I have always felt that violence might be an inherent part of human nature or moreso since violence has been disallowed in society, human craving for it will increase. Since humans are unable to experince the feeling of violence in reality, perhaps they might turn to games seeking for the experience of violence or for the thrill of being able to do something that is prohibited in reality. This has led me to explore the topic, does game violence have a relation to game popularity, which is what I will be studying in this report.

1.2 Aims and objectives

1.2.1 Background

Video game violence has been increasingly discussed and analysed as more violent and gory video games have appeared. There has been a substantial amount of research done on the affect of violence in video games to the habits of these game's players in reality. However, there has not been as much research done on the popularity of games in correlation to the level of violence in games. Thus, I have chosen to research more on this and try to analyse some data sets and conclude if violence has an impact on viedo game sales.

1.2.2 Aims and objectives

In this project, I would like to explore:

- How the datasets were chosen
- How the data has been cleaned and organised
- Limitations and ethical concerns of using the data
- How the data has been studied and analysed to look for relationships between the level of violence in games and its popularity in terms of sales. I will look mainly at game sales,game genre and their esrb rating during analysis

- Draw a conclusion based on the data analysis on how video game violence affects the game's sales

2 Data

2.1 Choice of dataset

I have chosen to use 2 datasets. The first dataset is a Video Games Sales dataset from Kaggle.
[Gregory Smith]. ([October 2016]). [Video Game Sales], [Version 2]. Retrieved [10/12/22] from https://www.kaggle.com/datasets/gregorut/videogamesales
This dataset contains a list of video games and has 16596 games recorded. It was generated using a scrape of vgchartz.com. It contains numerous fields of data with columns such as the rank,title,genres and sales of the games which are crucial information for me to determine how sales of the games are affected. The data is also complete without any missing informtion and is formatted in a csv file which i can access easily
The Fields of data are:

- Rank - Ranking of overall sales
- Name - The title of the game
- Platform - Platform that the games was released
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

The second dataset is a Video Games Rating By 'ESRB' from Kaggle.
[Mohammed Alhamad]. ([January 2021]). [Video Games Rating By 'ESRB'], [Version 3]. Retrieved [10/12/22] from [https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb?select=Video_games_esrb_rating.csv].
This dataset contains 1895 games with the Entertainment Software Rating Board(ESRB) rating along with many content descriptors relating to game violence that might be present in the game. ESRB rating is an American self-regulatory organization that assigns games an appropriate audience age based on the content descriptors such as the appearance of blood, violence, nudity, coarse language etc. This dataset is essesntial for me to be able to analyse the level of violence in each game since it contains the esrb rating as well as any presence of violent related features such as blood, animated blood, cartoon violence, violence, etc, and compare it to the game’s sales levels. The dataset is also complete without missing information and was updated 2 years ago which is still recent hence the data is decently reliable. It is also presented in a well formatted csv file that i can use readily hence i chose to use this dataset.
The ESRB ratings are as follows:

- RP-Rating Pending
- EC-Early Childhood
- E-Everyone
- E 10+ -Everyone 10+
- T-Teen
- M-Mature
- A-Adult

There are two datasets that I had considered but decided not to use eventually. The first is Video Game Sales and Ratings by Kendall Gillies (https://www.kaggle.com/datasets/kendallgillies/video-game-sales-and-ratings). I did not use this dataset as her data was taken from another user whos license could not be found hence i did not know of the copyright claim issues. Furthermore, there was incomplete information in certain columns like game rating and the dataset also did not include the game rank which is extremely useful for my research.
The second dataset i decided not to use is this Video games dataset from the University of Portsmouth(https://researchportal.port.ac.uk/en/datasets/video-games-dataset). The dataset is from 2013 which is a substantial period of time and I concluded the data would be outdated and it would also not be able to match very well with the ESRB rating dataset I was going to use. In addition, certain columns are unexplained, for example block 4,block2, YearReleasedSq, these columns have no description hence i was unsure what was being recorded. Even in columns such as Usedprice and InUsedPrice I was unsure what was the unit of measurement for the axis.Thus i chose not to use this dataset.

2.2 Ethical considerations

I have 2 datasets that I will be basing my research on, the Video Games Sales dataset and the Video Games Rating By 'ESRB' dataset.
As for the Video Games Sales dataset, the dataset was created by Gregory Smith and the creator has given access to his GitHub repository that has provided the script as to how the data was scrapped from the internet. The creator has issued the GNU General Public License v3.0 , stating that public users are allowed to use the data for commercial use, distribution, modification, patent use and private use. Permission to this dataset is conditioned on making available complete source code of licensed works and modifications. Copyright and license notices must be preserved.
The following dataset is The Video Games Rating By 'ESRB' dataset. This dataset was created by Mohammed Alhamad. The creator has also made this a public notebook with the CC0 1.0 Universal (CC0 1.0) Public Domain License, giving full access for anyone to use the dataset freely.

As for other ethical considerations regarding this project, I do not feel that any research summaries or analysis will lead to any dangerous or harmful implications or assumptions.This research only analyses is video game violence increases sales but does not instigate or suggest that there should be more violence in video games.

2.3 Limitations and constraints of data

A limitation of the Video Games Sales dataset is that dataset is 6 years old and has not been updated in more recent years hence is not completely reliable in performing a research for todays age. Furthermore, there is a difference in time periods between the 2 datasets as one was 6 years ago while the Video Games Rating By 'ESRB' dataset is only 2 years ago hence the games in each dataset might not match up as well as hoped.
Another limitation is that I am unable to know which games exactly contain violence and to what extent. The Video Games Sales dataset only provides me with the genre of the game but I will not be able to tell if there is violence in genres such as adventure or Action hence my analysis might not be able to fully encapsulate all the games that contain violence.

3 Data Modification and Analysis

I will first import the required libraries:

In [ ]:
```
import pandas as pd
import re, math
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns
```

The first thing to do is to read the 2 csv files and create the dataframes

In [ ]:
```
#reading the first csv data file containing the esrb ratings
sheet1 = pd.read_csv('C:/Users/../Pwd Mid Term/Video_games_esrb_rating.csv')
sheet1.head()

#reading the second csv data file containing the game genre and game sales data
sheet2 = pd.read_csv('C:/Users/../Pwd Mid Term/vgsales.csv')
sheet2.head()

#creating dataframe for both csv files
df1 = sheet1
df2 = sheet2
```

Comparing of data between two datasets

What I have done is to compare the two datasets and find similar titles of games so that I can use the data from both csv files for the same game. Since the titles from the 2 datasets are not exactly the same, in that they might have a different spacing or punctuation, I used a method known as cosine similarity to compare the two titles. The titles are made into lowercase words, non-alphanumeric characters are removed and the text is then made into a vector. The vector representation of the title can then be used in the cosine similarity formula to derive a value of similarity between 0 and 1. A value of 1 would indicate that the titles are exactly alike, while a value of 0 would indicate that the titles have no similarity. I used this to compare the two game titles from the dataframes and after trying out a few different values, I decided that a cosine similarity value of 0.67 would be the best to provide me with the titles of similar games in both data files. Although this method is not full proof and in my new dataframe I will have still have one or two mismatched games that are not actually the same game but names have a higher similarity value than 0.67. However, I still went with this method to create a new dataframe as majority of the game titles have been matched correctly.

Once i have done that i proceed to create a new dataframe for these games that have been title matched and create a new csv file

In [12]:

```
#compiles a regular expression object after reading the words and removing special
titleText = re.compile(r'\w+')

#calculating the cosineSimilarity value using its formula
def cosineSimilarity(titleVector1, titleVector2):
    intersection = set(titleVector1.keys()) & set(titleVector2.keys())
    numerator = np.sum([np.dot(titleVector1[x],titleVector2[x]) for x in intersect
    vectorSum1 = np.sum([titleVector1[x]**2 for x in titleVector1.keys()])
    vectorSum2 = np.sum([titleVector2[x]**2 for x in titleVector2.keys()])
    denominator = np.dot(math.sqrt(vectorSum1), math.sqrt(vectorSum2))

    if not denominator:
        return 0.0
    else:
        return float(numerator) / denominator

#returns all non-overlapping matches from the titleText function
def wordToVector(textss):
    return Counter(titleText.findall(textss))

#takes in the 2 game titles and makes them into lowercase words and calls wordToVe
def similarityValue(title1, title2):
    title1 = wordToVector(title1.strip().lower())
    title2 = wordToVector(title2.strip().lower())

    return cosineSimilarity(title1, title2)

#creating a dictionary for the data to be used in the dataframe
data = {"Title1":[],"Title2":[],"Similarity Score":[],"Rank":[],"Game Genre":[],"Y
        "EU Sales":[],"JP Sales":[],"Other Sales":[],"Global Sales":[],"ESRB ratin

#looping throguh all the df2 titles and checking if the similarity between each of
for x in range(1894):
    for y in range(1211):
        similarityScore = similarityValue(df1["Title"][x],df2["Name"][y])
        if similarityScore > 0.67:
            data["Title1"].append(df1["Title"][x])
            data["Title2"].append(df2["Name"][y])
            data["Similarity Score"].append(similarityScore)
            data["Rank"].append(df2["Rank"][y])
            data["Game Genre"].append(df2["Genre"][y])
            data["Year"].append(df2["Year"][y])
            data["NA Sales"].append(df2["NA_Sales"][y])
            data["EU Sales"].append(df2["EU_Sales"][y])
            data["JP Sales"].append(df2["JP_Sales"][y])
            data["Other Sales"].append(df2["Other_Sales"][y])
            data["Global Sales"].append(df2["Global_Sales"][y])
            data["ESRB rating"].append(df1["esrb_rating"][x])

#adjusting the display for the output data
pd.options.display.max_columns = None
pd.options.display.max_rows = None

#creating new dataframe for the new data
df3 = pd.DataFrame.from_dict(data)

# sorting by game ranking
```

```
df3.sort_values("Rank", inplace=True)

# dropping duplicate values of title2 and only keeping the top ranked value of that
df3.drop_duplicates(subset="Title2",
                    keep='first', inplace=True)

#creating a new csv file containing new dataframe for later analysis
df3.to_csv('C:/Users/../Downloads/NewGameData.csv', index=False, header=True)
df3.head(20)
```

Out[12]:

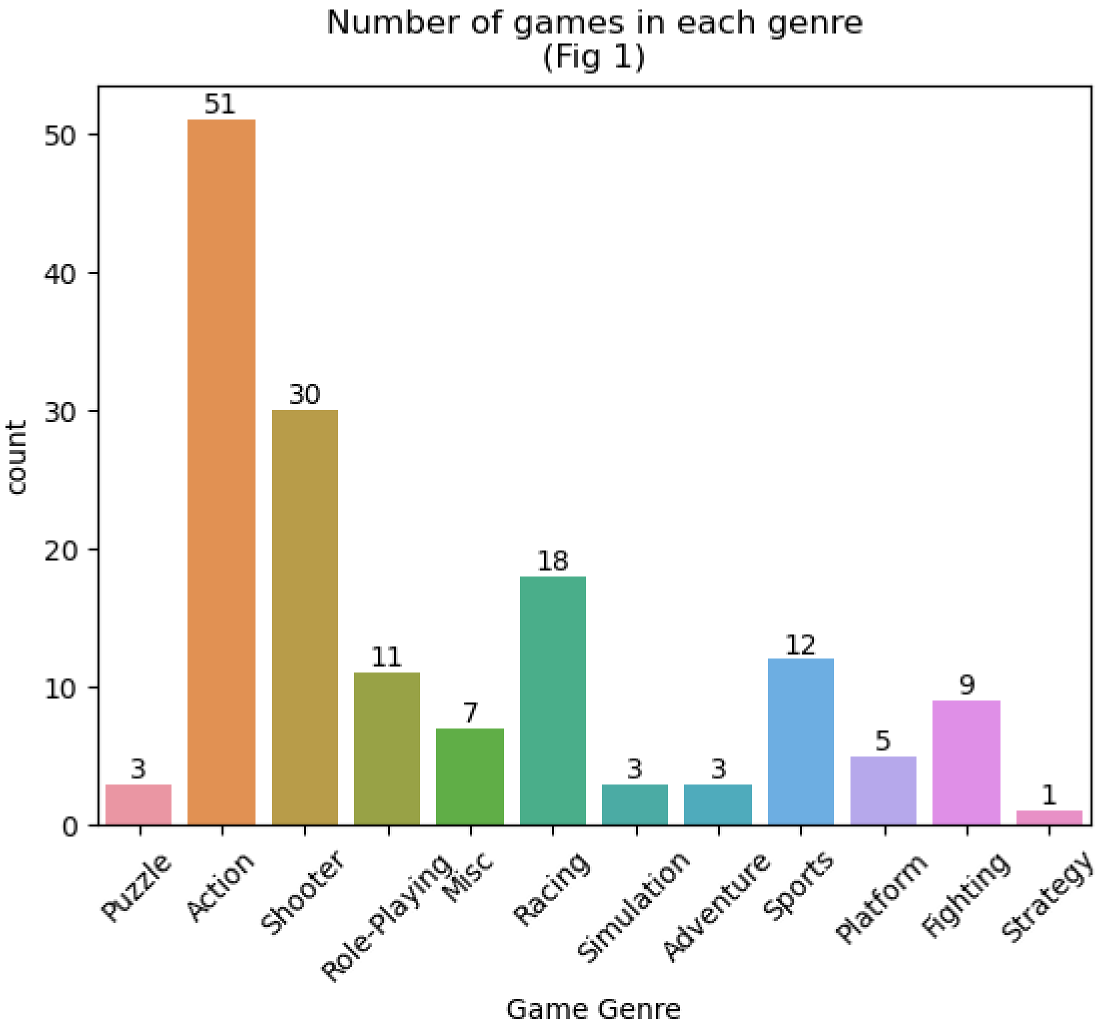
	Title1	Title2	Similarity Score	Rank	Game Genre	Year	NA Sales	EU Sales	JP Sales	Other Sales	Glot Sal
228	Tetris Effect	Tetris	0.707107	6	Puzzle	1989.0	23.20	2.26	4.22	0.58	30.
158	Grand Theft Auto V	Grand Theft Auto V	1.000000	17	Action	2013.0	7.01	9.27	0.97	4.14	21.
159	Grand Theft Auto V	Grand Theft Auto: San Andreas	0.670820	18	Action	2004.0	9.43	0.40	0.41	10.57	20.
161	Grand Theft Auto V	Grand Theft Auto: Vice City	0.670820	25	Action	2002.0	8.41	5.49	0.47	1.78	16.
127	Call of Duty®: Modern Warfare®	Call of Duty: Modern Warfare 3	0.912871	30	Shooter	2011.0	9.03	4.28	0.13	1.32	14.
27	Call of Duty®: Black Ops Cold War	Call of Duty: Black Ops	0.845154	32	Shooter	2010.0	9.67	3.73	0.11	1.13	14.
28	Call of Duty®: Black Ops Cold War	Call of Duty: Black Ops 3	0.771517	34	Shooter	2015.0	5.77	5.81	0.35	2.31	14.
29	Call of Duty®: Black Ops Cold War	Call of Duty: Black Ops II	0.771517	35	Shooter	2012.0	4.99	5.88	0.65	2.52	14.
128	Call of Duty®: Modern Warfare®	Call of Duty: Modern Warfare 2	0.912871	37	Shooter	2009.0	8.52	3.63	0.08	1.29	13.
162	Grand Theft Auto V	Grand Theft Auto III	0.750000	39	Action	2001.0	6.99	4.51	0.30	1.30	13.
164	Grand Theft Auto V	Grand Theft Auto IV	0.750000	52	Action	2008.0	6.76	3.10	0.14	1.03	11.
131	Call of Duty®: Modern Warfare®	Call of Duty: Ghosts	0.670820	62	Shooter	2013.0	6.72	2.63	0.04	0.82	10.
92	FINAL FANTASY VII Remake	Final Fantasy VII	0.866025	67	Role-Playing	1997.0	3.01	2.47	3.28	0.96	9.
133	Call of Duty®: Modern Warfare®	Call of Duty 4: Modern Warfare	0.912871	71	Shooter	2007.0	5.91	2.38	0.13	0.90	9.
299	Minecraft	Minecraft	1.000000	73	Misc	2013.0	5.58	2.83	0.02	0.77	9.
379	The Elder Scrolls V: Skyrim VR	The Elder Scrolls V: Skyrim	0.912871	76	Role-Playing	2011.0	5.03	2.86	0.10	0.85	8.
344	Star Wars™ Battlefront™	Star Wars Battlefront	0.750000	93	Shooter	2015.0	2.93	3.29	0.22	1.23	7.

	Title1	Title2	Similarity Score	Rank	Game Genre	Year	NA Sales	EU Sales	JP Sales	Other Sales	Glob Sal
	II	(2015)									
53	Call of Duty: Warzone	Call of Duty: Advanced Warfare	0.670820	94	Shooter	2014.0	2.80	3.30	0.14	1.37	7.
402	Just Dance 2018	Just Dance	0.816497	103	Misc	2009.0	3.51	3.03	0.00	0.73	7.
95	Need for Speed HEAT	Need for Speed Underground	0.750000	105	Racing	2003.0	3.27	2.83	0.08	1.02	7.

Here I have drawn a bar graph of the number of games in each genre. From observation, Action games have the highest number but shooter games are second highest. This shows that shooter games, which involve violence, are ranked decently high and have a high popularity

In [48]:

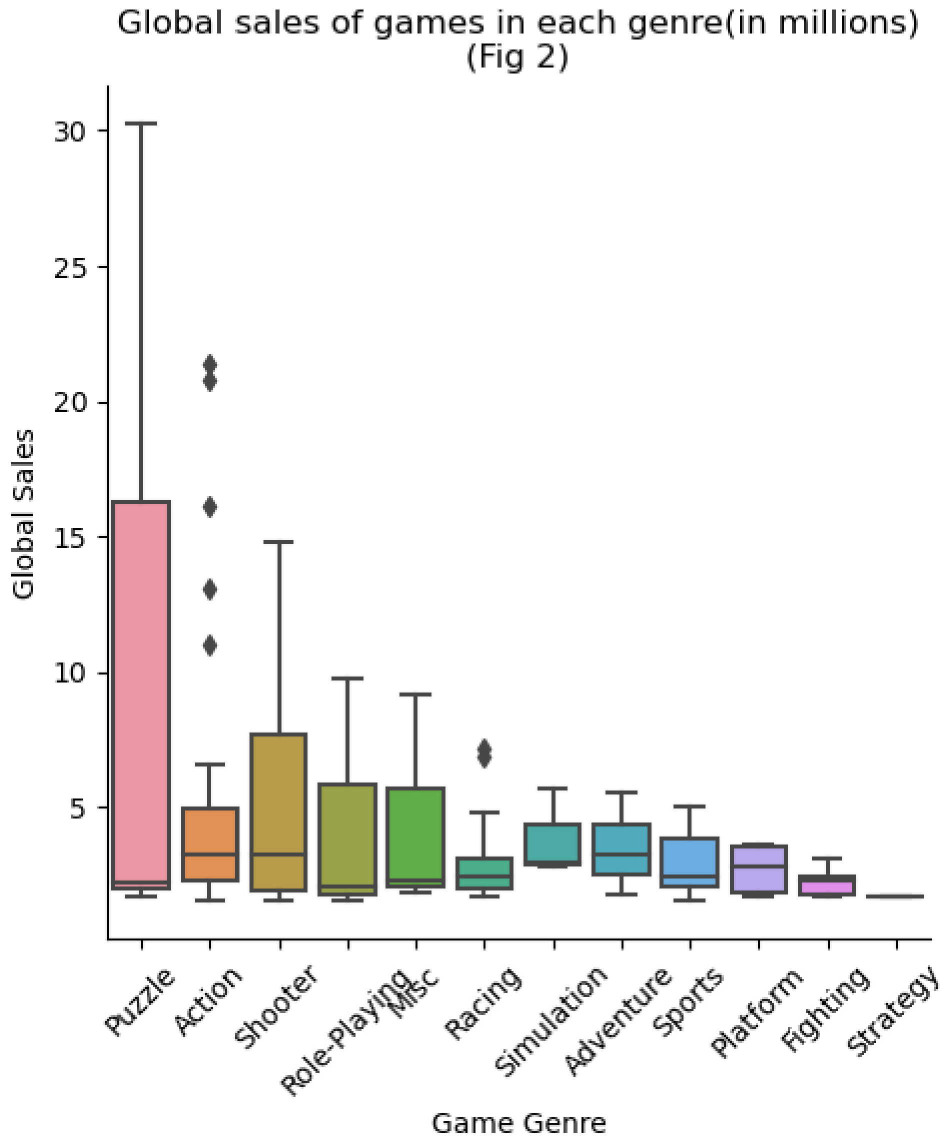
```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

df = pd.read_csv('C:/Users/../Pwd Mid Term/NewGameData.csv')
#drawing the bar graph
graph = sns.countplot(x= "Game Genre", data=df)
graph.bar_label(graph.containers[0])
plt.title("Number of games in each genre\n(Fig 1)")
plt.xticks(rotation = 45)
plt.show()
```



I have drawn a box plot to show the global sales of each genre. The main box in each genre shows the inter-quartile range (IQR) which is where 75% of the sales values lie. The bar in the middle of the box represents the median. The lines that extend from the box are where the highest and lowest sale values lie. Any other data that is above that line is an outlier. Yet again, from this plot, it can be seen that the global sales of shooter games is second highest behind puzzle games which implies that violent video games are still rather high in terms of sales.

In [49]:

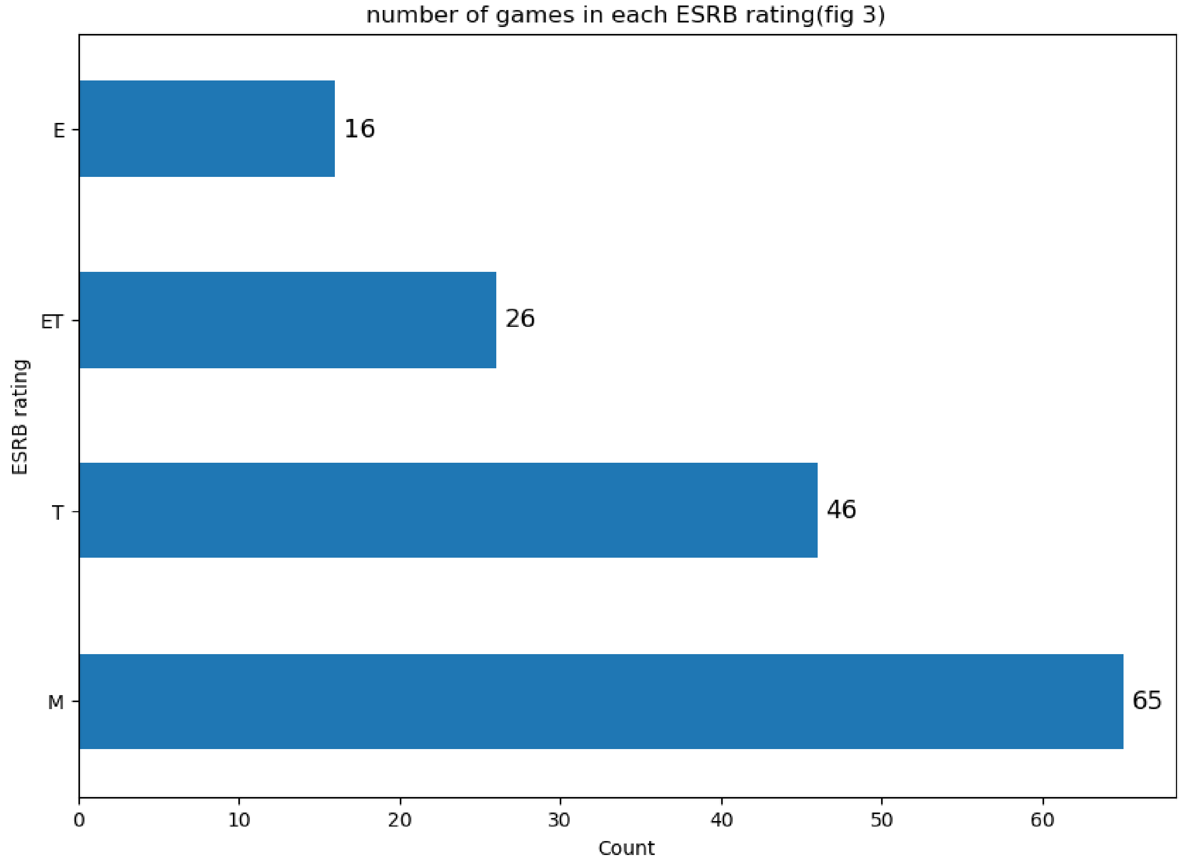
```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

pd.read_csv('C:/Users/../Pwd Mid Term/NewGameData.csv')
#drawing the box plot
sns.catplot(x="Game Genre", y="Global Sales", kind="box", data=df)
plt.title("Global sales of games in each genre(in millions)\n(Fig 2)") #title of ti
plt.xticks(rotation=45) # rotating the x axis labels
plt.show()
```



Here I have plotted a bar graph counting the number of games in each ESRB rating. As observed, the highest number of games lies in the M(Mature) category. This category has the highest number of games that contain violence or other inappropriate features. Violence will definitely be the highest in this category of games.

In [30]:

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

pd.read_csv('C:/Users/../Pwd Mid Term/NewGameData.csv')
#plotting a bar graph
ax = df['ESRB rating'].value_counts().plot(kind='barh', figsize=(10,7), title="numl
ax.set_ylabel("ESRB rating") #label for the y axis
ax.set_xlabel("Count")  #label for the x axis
for bars in ax.containers:
    ax.bar_label(bars,padding=4, fontsize=13)
plt.show()
```



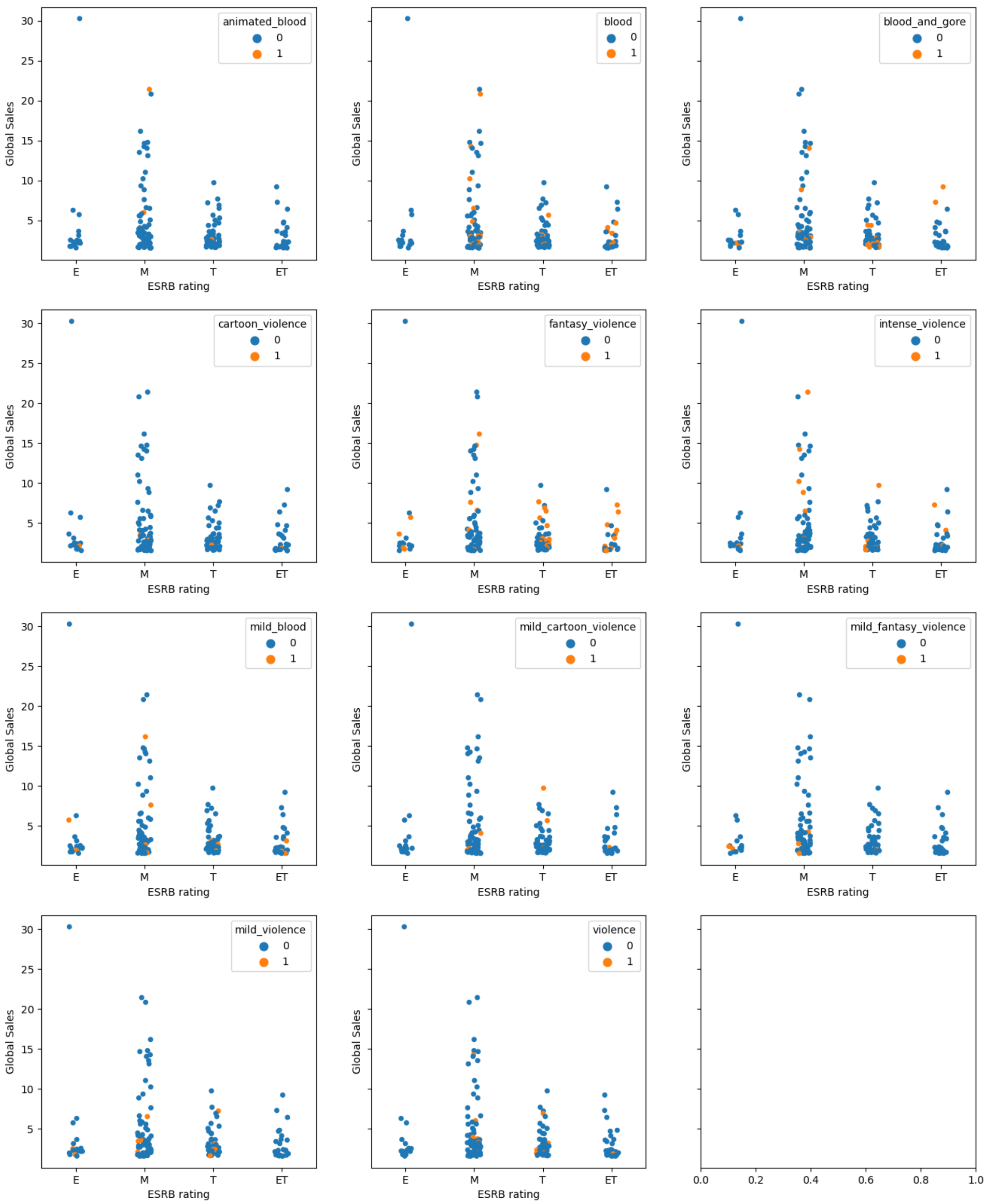
In this following part, I have first created a dictionary to collate the content descriptors that relate to my research topic of game violence since not all the content descriptors were related to that. I then looped over the titles in the new csv file I created ealier and searched for those titles in the "Video Games Rating By 'ESRB' " dataframe so that I can find the content descriptors for those games. I then printed out the number of games in each content descriptor and then proceeded to create a grid for the scatterplots. I have drawn a scatter plot for the global sales against the ESRB rating for each content descriptor.

In [23]:
```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

df1 = pd.read_csv('C:/Users/../Pwd Mid Term/NewGameData.csv')
df2 = pd.read_csv('C:/Users/../Pwd Mid Term/Video_games_esrb_rating.csv')

#creating a dictionary for the data to be used in the dataframe
data = {"Title":[],"ESRB rating":[],"Global Sales":[],"animated_blood":[],"blood":
        "fantasy_violence":[],"intense_violence":[],"mild_blood":[],"mild_cartoon_
        "mild_violence":[],"violence":[]}

#looping over the values in df1 and checking if the title is in df2 and writing the
for x in range(153):
    for y in range(1895):
        if df1["Title1"][x] == df2["Title"][y]:
            data["Title"].append(df1["Title1"][x])
            data["ESRB rating"].append(df1["ESRB rating"][x])
            data["Global Sales"].append(df1["Global Sales"][x])
            data["animated_blood"].append(df2["animated_blood"][x])
            data["blood"].append(df2["blood"][x])
            data["blood_and_gore"].append(df2["blood_and_gore"][x])
            data["cartoon_violence"].append(df2["cartoon_violence"][x])
            data["fantasy_violence"].append(df2["fantasy_violence"][x])
```

```
            data["intense_violence"].append(df2["intense_violence"][x])
            data["mild_blood"].append(df2["mild_blood"][x])
            data["mild_cartoon_violence"].append(df2["mild_cartoon_violence"][x])
            data["mild_fantasy_violence"].append(df2["mild_fantasy_violence"][x])
            data["mild_violence"].append(df2["mild_violence"][x])
            data["violence"].append(df2["violence"][x])

df3 = pd.DataFrame.from_dict(data)
print(df3['animated_blood'].value_counts(),df3['blood'].value_counts(),df3['blood_
      df3['cartoon_violence'].value_counts(),df3['fantasy_violence'].value_counts(
      df3['mild_blood'].value_counts(),df3['mild_cartoon_violence'].value_counts()
      df3['mild_violence'].value_counts(),df3['violence'].value_counts())
#creating a grid for the many scatterplots
fig, axes = plt.subplots(4, 3, figsize=(16, 20), sharey=True)
fig.suptitle('Comparing the global sales in each esrb rating for every violence typ

# scatter plot for violence content descriptor:animated_blood
sns.stripplot(ax=axes[0,0], x="ESRB rating",y="Global Sales", hue="animated_blood"
# scatter plot for violence content descriptor:blood
sns.stripplot(ax=axes[0,1], x="ESRB rating",y="Global Sales", hue="blood", data=df
# scatter plot for violence content descriptor:blood_and_gore
sns.stripplot(ax=axes[0,2], x="ESRB rating",y="Global Sales", hue="blood_and_gore"
# scatter plot for violence content descriptor:cartoon_violence
sns.stripplot(ax=axes[1,0], x="ESRB rating",y="Global Sales", hue="cartoon_violenc
# scatter plot for violence content descriptor:fantasy_violence
sns.stripplot(ax=axes[1,1], x="ESRB rating",y="Global Sales", hue="fantasy_violenc
# scatter plot for violence content descriptor:intense_violence
sns.stripplot(ax=axes[1,2], x="ESRB rating",y="Global Sales", hue="intense_violenc
# scatter plot for violence content descriptor:mild_blood
sns.stripplot(ax=axes[2,0], x="ESRB rating",y="Global Sales", hue="mild_blood", da
# scatter plot for violence content descriptor:mild_cartoon_violence
sns.stripplot(ax=axes[2,1], x="ESRB rating",y="Global Sales", hue="mild_cartoon_vi
# scatter plot for violence content descriptor:mild_fantasy_violence
sns.stripplot(ax=axes[2,2], x="ESRB rating",y="Global Sales", hue="mild_fantasy_vi
# scatter plot for violence content descriptor:mild_violence
sns.stripplot(ax=axes[3,0], x="ESRB rating",y="Global Sales", hue="mild_violence",
# scatter plot for violence content descriptor:violence
sns.stripplot(ax=axes[3,1], x="ESRB rating",y="Global Sales", hue="violence", data

plt.show()
```

```
0    150
1      3
Name: animated_blood, dtype: int64 0    127
1     26
Name: blood, dtype: int64 0    133
1     20
Name: blood_and_gore, dtype: int64 0    148
1      5
Name: cartoon_violence, dtype: int64 0    118
1     35
Name: fantasy_violence, dtype: int64 0    136
1     17
Name: intense_violence, dtype: int64 0    141
1     12
Name: mild_blood, dtype: int64 0    146
1      7
Name: mild_cartoon_violence, dtype: int64 0    145
1      8
Name: mild_fantasy_violence, dtype: int64 0    141
1     12
Name: mild_violence, dtype: int64 0    141
1     12
Name: violence, dtype: int64
```

Comparing the global sales in each esrb rating for every violence type(Fig 4)



From firgure 4, Cartoon violence, animated blood and blood are the three highest content descriptors that have appeared in games. However, even though they have appeared in a higher number of games, they are not mainly in games with high sales and are rather in average or lower selling games. In all the scatterplots, it can be observed that the category M(Mature) has the highest sales, the games in this category by nature have the highest amount of violence. A surprising conclusion from these scatterplots is also that there is an appearance of violence even in categories like E(Everyone) and ET(Everyone 10+) especially in the fantasy violence content descriptor, although the sales are not very high.

4 Summary

4.1 Data analysis conclusion

Overall, comparing the analysis from differents charts, figure 1 showed that action games have the highest number of sales but shooter games are second highest. Unfortunately there is no way of knowing if those action games have violence but there is a good chance that a number of them will contain violence. Using that analysis, game violence has actually allowed for decently high game sales. From figure 2, the anaysis of the box plot has shown that the global sales of shooter games is second highest behind puzzle games. Puzzle games has definitely dominated sales in this plot but shooter games has also performed well with its highest sales still being significantly above other genres. In figure 3, it can be seen that the largest number of games comes from the M(Mature) category and from the ESRB official webpage,it is stated that for the M category the content may contain intense violence and blood and gore hence the level of violence is definitely higher in this category than the rest. From figure 3, since it was concluded that the M category of games has the largest number of games, looking at the M category in figure 4, from the sctatterplots it has clearly the highest sales even in all content descriptor categories. Surprisingly, the three highest content descriptors that appeared in games most did not contribute to the highest selling games. But overall, it is clear that the category with the most violence has contributed the greatest to sales.

4.2 Conclusion

In summary, though I cannot say for certain that my analysis has provided a very clear conclusion, I think there has been sufficient evidence to show that the presence of violence in video games does lead to higher sales compared to the majority of other games. Even though my research has many missing elements such as not being able to know the presence of violence in every single game, I feel that on a general note, from the data analysis, violent video games have performed better than most games in terms of sales.

5 References

5.1 References

- https://www.kaggle.com/datasets/gregorut/videogamesales
- https://github.com/GregorUT/vgchartzScrape/blob/master/LICENSE
- https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb?select=Video_games_esrb_rating.csv
- https://www.esrb.org/
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

- https://stackoverflow.com/questions/5268929/compare-two-csv-files-and-search-for-similar-items