In [ ]:
```
#**** PROJECT INTRODUCTION ****#
#The data that we are currently investigating and exploring is VG_sales.csv.
#We wanted to get a better idea of what this dataset contained.
#After exploring it we saw that it mainly gave the total sales of different types
#regions, and then combined all the other ones into "other".
#The biggest questions we have for this data is what genres sold
#the most games and to see if there are any big differences between NA, EU, and J
#We think that we can gather good data
```

In [ ]:
```
#**** CHANGES ****#
#So there hasnt been many changes right now, the biggest question that comes to m
#is how can we link the sales of different
#genres to whether or not certain genres or games in general effect peoples healt
```

In [ ]:
```
#**** DATA CLEANING ****#
#Luckily for the VG_sales.csv we didnt have to clean much at all,
#all the columns were describe well and the data was easy
#to understand. It was more of how we wanted to explore the data.
```

In [2]:
```
#**** EDA Structure****#

#The data that we have been messing around with is tabular
#and this allowed us to easily import it directly to either
#google sheets or excel. This allowed us to easily figure out
#what was going on in the data as well as know what each column
#represented.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

tweets = pd.read_csv("vg_sales.csv", na_filter=False)
tweets.head()
```

Out[2]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	

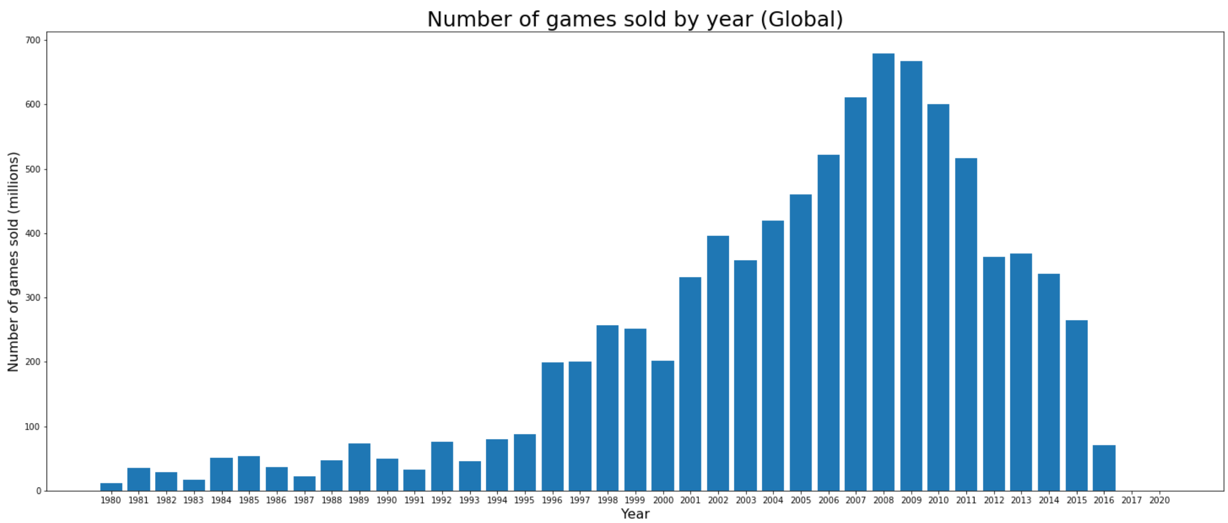
In [ ]:
```
#**** EDA Granularity ****#
#The granularity of the data was also very straight forward,
#as you can see each column is labled very straight forward
#The columns are ranks, the name of the game, year,
#publisher, and then how well it did in each of those regions. The final
#column adds all of the percentages up.
#The data just seems to be the % of sales per region and then it ranks all games
#on the global sales %.
```

In [ ]:
```
#**** EDA Scope ****#
#The scope of this data doesnt directly show whether
#or not video games effect people in a negative or positive way, but from
#this we can see what kind of games people around the world,
#as well as certain regions, enjoy most. This allows us to
#to ask more questions and try to see if their are
#relationships between certain genres and certain health conditions occuring
#in  people. Since the range of the games sold is very wide,
#we believe that we can apply our findings here to other hypotheses
#down the road.
```

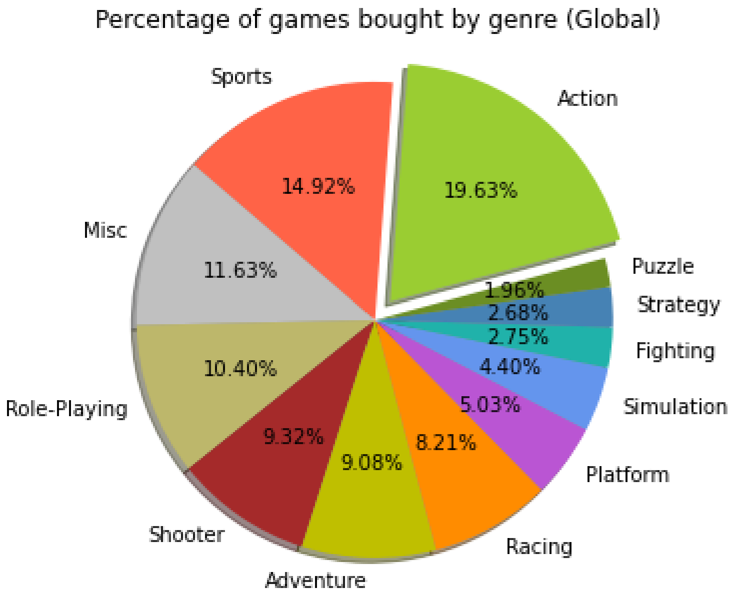
In [13]:
```
#**** EDA Temporality And Faithfulness ****#
#The time frame for the data ranges from 1980-2020.
#Some rows contain N/A but we cleaned it to allow us to work better.
#The data is also faithful, we looked up the amount of games sold and
#compared it to some of them and saw that they are extremely close.

#**** EDA conclusion ****#
#We found a lot of cool things while looking at the data, and noticed some very c
#An example of this would be how most games sold were around 2008 and 2009,
#which we assumed to be around the time that Xbox Live was extremely big.
#the biggest challenge that comes to  mind is trying to figure out a way if
#we can link this to our main idea or scrap it.
```

In [14]:
```
sales = tweets.groupby('Year')['Global_Sales'].sum()
sales = pd.DataFrame(sales)
sales = sales[sales.index != 'N/A']
labels = sales.index
plt.bar(labels, sales['Global_Sales'])
plt.title('Number of games sold by year (Global)', fontsize=25)
plt.xlabel('Year', fontsize=16)
plt.ylabel('Number of games sold (millions)', fontsize=16)
fig = plt.gcf()
fig.set_size_inches(25,10)
plt.show()
```



In [12]:
```
#We have more visualizations under the dataset folder on github in the sales.ipyr
sales = tweets.groupby('Genre')['Global_Sales'].sum().sort_values(ascending=False
labels = tweets['Genre'].value_counts().index
explode = (0.1, 0, 0, 0, 0, 0, 0 ,0 ,0 ,0 ,0 ,0)
colors = ['yellowgreen', 'tomato', 'silver', 'darkkhaki', 'brown', 'y', 'darkorar
pie = plt.pie(sales, labels=labels, autopct='%1.2f%%', explode=explode, shadow=Tr
plt.title('Percentage of games bought by genre (Global)')
plt.axis('equal')
fig = plt.gcf()
fig.set_size_inches(5,5)
plt.show()
```



In [ ]:
```
#**** ML ****#
#We currently do not have a complete Machine Learning analysis so far,
#but we do have some ideas that we want to implement
#because we have a lot of features within the data.
#The current Machine learning analysis that we are working on
#is to find a relationship between if people get to the same levels of anger,
#for example road rage, as they do when they
#play violent video game genres, such as action and/or shooters.
#The current baseline that we have in mind for this analysis
#would be that 50% of the people that get "car angry" have recently
#played some sort of violent video game genre.
#We expect our Machine Learning model to be more accurate than 50% on
#this specific type of topic. We think this because after
#analyzing the data from the vg_aggression file, we can see that there is a
#trend that correlates violent games to road rage.
#The testing code is also on the github under machine learning 418 folder.
```

In [ ]:
```
#**** REFLECTION ****#
#So far the hardest part of the project, as stated above, is trying to find
#a correlation between certain genres of games
#and different types of impacts that they can cause.
#Basically we want to figure out a way to see if different genres
#of games cause different types of behaviours of people or
#different mental states and/or things from occuring.
#Right now our initial insight on this question is yes,
#since there are a lot of studies and data out there that have
#studied how different types of games effect peoples mental states.
#So we believe that we can manage to do it and link these
#two things. At this point there are a few concrete things that we can show,
#mainly different tpyes of genres being the most
#popular all around the world. As well as certain years
#and different publishers being the most popular. Going forward,
#with the data that we have, our biggest obstacle is
#figuring out how we can correlate these things that we have learned
#and how it effects peoples health. We do believe that
#we are on track, but we for sure have to spend more time exploring
#the other data set that is mainly showing us different
#effects, good or bad, that different games and genres can impose
#onto mental health or phsyical health. But we do believe
#that the current project and hypothese that we currently have
#are worth preceding and learning about.
```

In [ ]:
```
#**** NEXT STEPS ****#
#Our next step is to go through the aggression.csv file and
#explore it thoroughly so we have a good understanding of it.
#And then clean up the data and get exactly what we need from it,
#and essentially see if we need to go look for more data
#to explore and perform EDA/cleaning on it.
#We're also going to figure the Machine Learning analysis that
#we want to use exactly and apply it accordingly to the
#hypothesis that we have in mind for it.
#And lastly we have to go through the video game aggression csv file
#and apply data cleaning techniques to it so it is easier to understand.
```