In [6]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [7]:
```
df=pd.read_csv("imdb.csv")
df
```

Out[7]:

	Rank	Title	Genre	Description	Director	Actors	Year
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016
...	...	...	...	...	...	...	...
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016

1000 rows × 12 columns

In [10]:
```
df.head(10)
```

Out[10]:

	Rank	Title	Genre	Description	Director	Actors	Year	(M
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	
5	6	The Great Wall	Action,Adventure,Fantasy	European mercenaries searching for black powde...	Yimou Zhang	Matt Damon, Tian Jing, Willem Dafoe, Andy Lau	2016	
6	7	La La Land	Comedy,Drama,Music	A jazz pianist falls for an aspiring actress i...	Damien Chazelle	Ryan Gosling, Emma Stone, Rosemarie DeWitt, J....	2016	
7	8	Mindhorn	Comedy	A has-been actor best known for playing the ti...	Sean Foley	Essie Davis, Andrea Riseborough, Julian Barrat...	2016	
8	9	The Lost City of Z	Action,Adventure,Biography	A true-life drama, centering on British explor...	James Gray	Charlie Hunnam, Robert Pattinson, Sienna Mille...	2016	
9	10	Passengers	Adventure,Drama,Romance	A spacecraft traveling to a distant colony pla...	Morten Tyldum	Jennifer Lawrence, Chris Pratt, Michael Sheen,...	2016	

In [11]:
```
df.tail(10)
```

Out[11]:

	Rank	Title	Genre	Description	Director	Actors	Year	Ru (Mi
990	991	Underworld: Rise of the Lycans	Action,Adventure,Fantasy	An origins story centered on the centuries-old...	Patrick Tatopoulos	Rhona Mitra, Michael Sheen, Bill Nighy, Steven...	2009	
991	992	Taare Zameen Par	Drama,Family,Music	An eight-year-old boy is thought to be a lazy ...	Aamir Khan	Darsheel Safary, Aamir Khan, Tanay Chheda, Sac...	2007	
992	993	Take Me Home Tonight	Comedy,Drama,Romance	Four years after graduation, an awkward high s...	Michael Dowse	Topher Grace, Anna Faris, Dan Fogler, Teresa P...	2011	
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	
994	995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012	
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014	

	Rank	Title	Genre	Description	Director	Actors	Year	Ru (Mi
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	

In [12]:
```
df.shape   #row,column
```

Out[12]:
```
(1000, 12)
```

In [13]:
```
# identify each column and analyse the no of observation there.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                1000 non-null   int64  
 1   Title               1000 non-null   object 
 2   Genre               1000 non-null   object 
 3   Description         1000 non-null   object 
 4   Director            1000 non-null   object 
 5   Actors              1000 non-null   object 
 6   Year                1000 non-null   int64  
 7   Runtime (Minutes)   1000 non-null   int64  
 8   Rating              1000 non-null   float64
 9   Votes               1000 non-null   int64  
 10  Revenue (Millions)  872 non-null    float64
 11  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
```

In [14]:
```
#analyse the missing values
df.isnull().sum()
```

Out[14]:
```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore              64
dtype: int64
```

In [15]:
```
#remove missing values
films=df.dropna(axis=0)
films.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 838 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                838 non-null    int64  
 1   Title               838 non-null    object 
 2   Genre               838 non-null    object 
 3   Description         838 non-null    object 
 4   Director            838 non-null    object 
 5   Actors              838 non-null    object 
 6   Year                838 non-null    int64  
 7   Runtime (Minutes)   838 non-null    int64  
 8   Rating              838 non-null    float64
 9   Votes               838 non-null    int64  
 10  Revenue (Millions)  838 non-null    float64
 11  Metascore           838 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 85.1+ KB
```

In [16]:
```
# understanding basic statistics(min ranking,revenue,avg,std etc)
films.describe()
```

Out[16]:

	Rank	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
count	838.000000	838.00000	838.000000	838.000000	8.380000e+02	838.000000	838.000000
mean	485.247017	2012.50716	114.638425	6.814320	1.932303e+05	84.564558	59.575179
std	286.572065	3.17236	18.470922	0.877754	1.930990e+05	104.520227	16.952416
min	1.000000	2006.00000	66.000000	1.900000	1.780000e+02	0.000000	11.000000
25%	238.250000	2010.00000	101.000000	6.300000	6.127650e+04	13.967500	47.000000
50%	475.500000	2013.00000	112.000000	6.900000	1.368795e+05	48.150000	60.000000
75%	729.750000	2015.00000	124.000000	7.500000	2.710830e+05	116.800000	72.000000
max	1000.000000	2016.00000	187.000000	9.000000	1.791916e+06	936.630000	100.000000

In [17]:
```
#exploratory data analysis(EDA)
#display the movie which runs greater than 170 mints.
films[films["Runtime (Minutes)"]>170]["Title"]
```

Out[17]:
```
82     The Wolf of Wall Street
88           The Hateful Eight
267                Cloud Atlas
311            La vie d'Adèle
Name: Title, dtype: object
```

In [18]:
```
films[films["Runtime (Minutes)"]>170][["Title","Genre","Rating"]]
```

Out[18]:

	Title	Genre	Rating
82	The Wolf of Wall Street	Biography,Comedy,Crime	8.2
88	The Hateful Eight	Crime,Drama,Mystery	7.8
267	Cloud Atlas	Drama,Sci-Fi	7.5
311	La vie d'Adèle	Drama,Romance	7.8

In [20]:
```
#avg revenue generated wrt year also showing the values in smaller to larger
films.groupby("Year")["Revenue (Millions)"].mean().sort_values(ascending=True)
```

Out[20]:
```
Year
2016     55.566111
2015     80.725596
2014     85.433656
2006     87.255610
2013     88.084643
2007     93.074091
2011     93.703333
2008     97.525417
2010    105.081579
2012    110.103065
2009    115.742000
Name: Revenue (Millions), dtype: float64
```

In [21]:
```
#display the movie which runs greater than 170 mints and rating greater than 9.0
films[(films["Runtime (Minutes)"]>170) & (films["Rating"]>7.0)]
```

Out[21]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)
82	83	The Wolf of Wall Street	Biography,Comedy,Crime	Based on the true story of Jordan Belfort, fro...	Martin Scorsese	Leonardo DiCaprio, Jonah Hill, Margot Robbie,M...	2013	180
88	89	The Hateful Eight	Crime,Drama,Mystery	In the dead of a Wyoming winter, a bounty hunt...	Quentin Tarantino	Samuel L. Jackson, Kurt Russell, Jennifer Jaso...	2015	187
267	268	Cloud Atlas	Drama,Sci-Fi	An exploration of how the actions of individua...	Tom Tykwer	Tom Hanks, Halle Berry, Hugh Grant, Hugo Weaving	2012	172
311	312	La vie d'Adèle	Drama,Romance	Adèle's life is changed when she meets Emma, a...	Abdellatif Kechiche	Léa Seydoux, Adèle Exarchopoulos, Salim Kechio...	2013	180

In [22]:
```
# show the top 10 votes with runtime and also the rank ,title,votes,runtime,year,direc
```

```
films.nlargest(10,["Votes","Runtime (Minutes)"])[["Title","Rank","Year","Runtime (Minu
```

Out[22]:

	Title	Rank	Year	Runtime (Minutes)	Rating	Votes	Director	Genre
54	The Dark Knight	55	2008	152	9.0	1791916	Christopher Nolan	Action,Crime,Drama
80	Inception	81	2010	148	8.8	1583625	Christopher Nolan	Action,Adventure,Sci-Fi
124	The Dark Knight Rises	125	2012	164	8.5	1222645	Christopher Nolan	Action,Thriller
36	Interstellar	37	2014	169	8.6	1047747	Christopher Nolan	Adventure,Drama,Sci-Fi
76	The Avengers	77	2012	143	8.1	1045588	Joss Whedon	Action,Sci-Fi
144	Django Unchained	145	2012	165	8.4	1039115	Quentin Tarantino	Drama,Western
77	Inglourious Basterds	78	2009	153	8.3	959065	Quentin Tarantino	Adventure,Drama,War
99	The Departed	100	2006	151	8.5	937414	Martin Scorsese	Crime,Drama,Thriller
87	Avatar	88	2009	162	7.8	935408	James Cameron	Action,Adventure,Fantasy
64	The Prestige	65	2006	130	8.5	913152	Christopher Nolan	Drama,Mystery,Sci-Fi

In [23]:

```
# total no of movies launched by each year
dataset=films["Year"].value_counts().reset_index()
movies=dataset.rename(columns={"index":"Year","Year":"Count"})
movies
```

Out[23]:

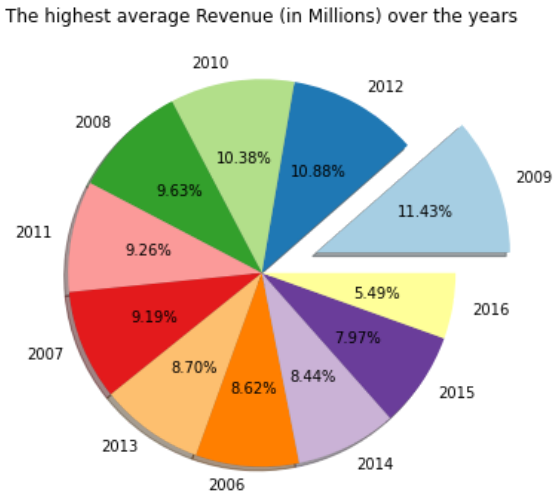
	Year	Count
0	2016	198
1	2015	109
2	2014	93
3	2013	84
4	2012	62
5	2011	57
6	2010	57
7	2008	48
8	2009	45
9	2007	44
10	2006	41

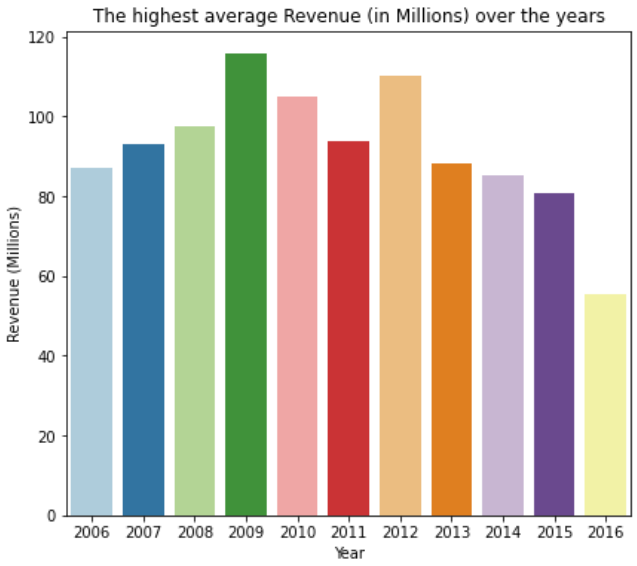
In [24]:
```
# represents movie by each year that has generated highest revenue in million.
films.groupby("Year")[["Title","Runtime (Minutes)"]].max()
```

Out[24]:

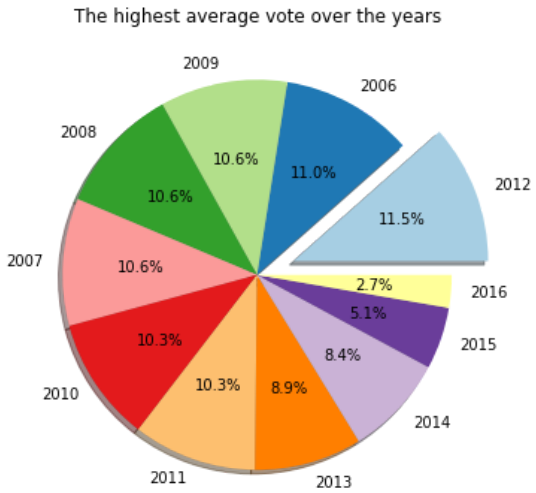
Year	Title	Runtime (Minutes)
2006	X-Men: The Last Stand	154
2007	Zodiac	169
2008	Zack and Miri Make a Porno	166
2009	Zombieland	170
2010	Unstoppable	148
2011	Your Highness	158
2012	Zero Dark Thirty	172
2013	World War Z	180
2014	X-Men: Days of Future Past	169
2015	Youth	187
2016	Zootopia	163

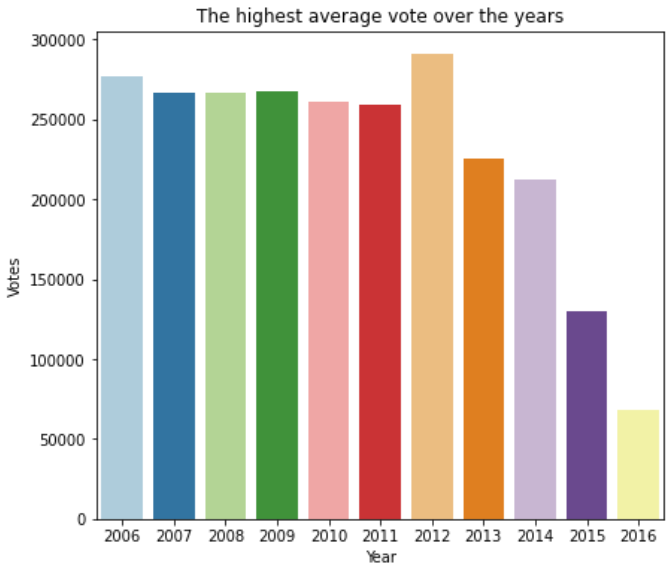
In [25]:
```
#Apart from above exploratory analysis, let's use some graphical approaches to analyse

highest_revenue=films.groupby("Year")["Revenue (Millions)"].mean().sort_values(ascendi
# Pie chart
plt.figure(figsize=(15,6))
plt.subplot(1,2,1)
colors=sns.color_palette("Paired")
space=(0.3,0,0,0,0,0,0,0,0,0,0)
plt.pie(highest_revenue[ "Revenue (Millions)"],
labels=highest_revenue["Year"],
autopct="%0.2f%%",
explode=space,
shadow=True,
colors=colors)
plt.title("The highest average Revenue (in Millions) over the years")
# Barplot
plt.subplot(1,2,2)
sns.barplot(x="Year",y="Revenue (Millions)",data=highest_revenue, palette=colors)
plt.title("The highest average Revenue (in Millions) over the years")
plt.show()
```





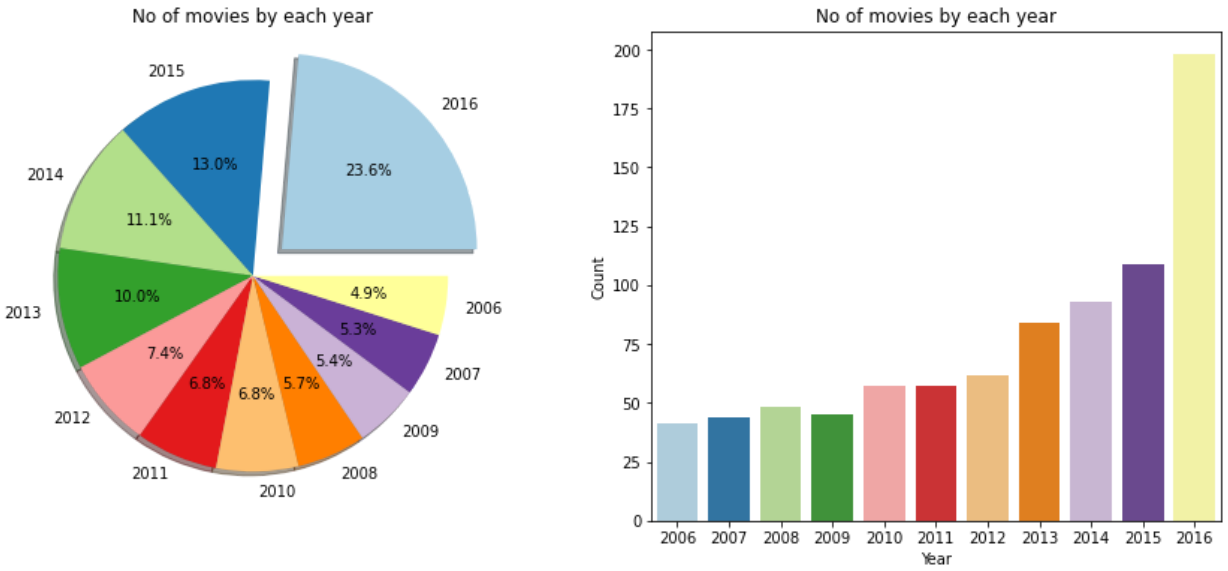
In [26]:
```
#In the same way, highest voting per year can also analysed with the help of following

highest_votes=films.groupby("Year")["Votes"].mean().sort_values(ascending=False).reset
# Pie chart
plt.figure(figsize=(15,6))
plt.subplot(1,2,1)
colors=sns.color_palette("Paired")
space=(0.2,0,0,0,0,0,0,0,0,0)
plt.pie(highest_votes[ "Votes"],
labels=highest_votes["Year"],
autopct="%1.1f%%",
explode=space,
shadow=True,
colors=colors)
plt.title("The highest average vote over the years")
# Barplot
plt.subplot(1,2,2)
sns.barplot(x="Year",y="Votes",data=highest_votes, palette=colors)
plt.title("The highest average vote over the years")
plt.show()
```



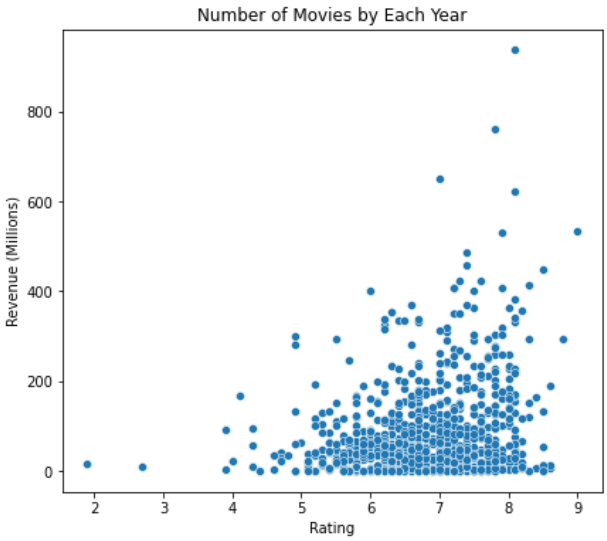


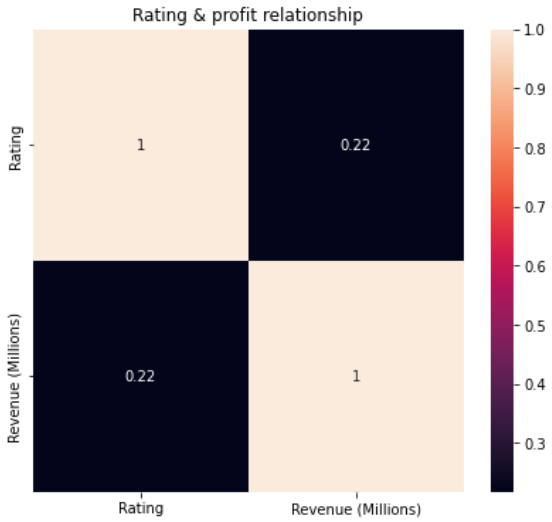
In [29]:
```
# Let's view number of movies by each year with the help of following command followed
```

```
plt.figure(figsize=(15,6))
plt.subplot(1,2,1)
colors=sns.color_palette("Paired")
space=(0.2,0,0,0,0,0,0,0,0,0,0)
plt.pie(movies["Count"],
labels=movies["Year"],
autopct="%1.1f%%",
explode=space,
shadow=True,
colors=colors)
plt.title("No of movies by each year")
# Barplot
plt.subplot(1,2,2)
sns.barplot(x="Year",y="Count",data=movies, palette=colors)
plt.title("No of movies by each year")
plt.show()
```



In [54]:
```
# identify the relationship between rating of movies and revenue generated by movies.

# Relationship between Revenue and Rating
plt.figure(figsize=(15,6))

plt.subplot(1,2,1)

sns.scatterplot(x="Rating",y="Revenue (Millions)", data=films)
plt.title("Number of Movies by Each Year")

plt.subplot(1,2,2)

df=films[["Rating", "Revenue (Millions)"]]
sns.heatmap(df.corr(),annot=True)
plt.title("Rating & profit relationship")
plt.show()
```

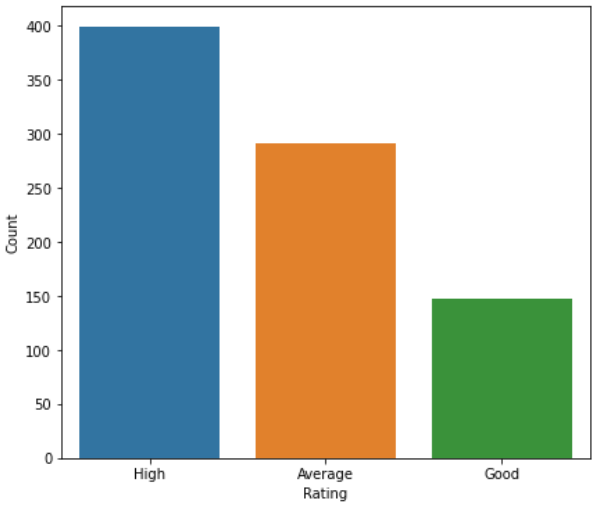


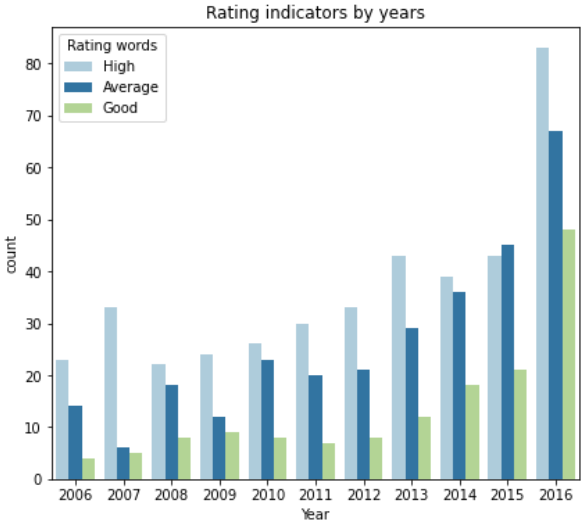


In [56]:
```
# Let's create a column of rating to specify categories of high, good and average base

def rating(movie):
    if movie.Rating>=7.0:
        return "High"
    elif movie.Rating<=6.0:
        return "Good"
    else:
        return "Average"
films["Rating words"]=films.apply(rating,axis="columns")
```

```
C:\Users\SREEJA BISWAS\AppData\Local\Temp\ipykernel_2768\913324727.py:10: SettingWith
CopyWarning:
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/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  films["Rating words"]=films.apply(rating,axis="columns")
```

In [55]:
```
# The following figure is used to create sub-dataset, followed by command to visualise

data_set2=films[["Rating words"]].value_counts().reset_index()
rating=data_set2.rename(columns={"Rating words":"Rating",0:"Count"})

plt.figure(figsize=(15,6))

plt.subplot(1,2,1)

sns.barplot(x="Rating",
y="Count",
data=rating)

plt.subplot(1,2,2)

sns.countplot(x="Year",hue="Rating words", data=films,palette=colors)
plt.title("Rating indicators by years")
plt.show()
```


400
350
300
250
200
150
100
50
0
Count
High
Average
Good
Rating


Rating indicators by years
Rating words
High
Average
Good
80
70
60
50
40
30
20
10
0
count
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year

In [ ]: