Final Project Submission

Please fill out:

- Student name: Stephen Gomes
- Student pace: self paced
- Scheduled project review date/time: 2022.10.26 15:00
- Instructor name: Joe Comeaux
- Blog post URL: https://github.com/steve-gomes/steve-gomes.github.io/blob/main/_posts/2022-10-15-your-new-blog-post.md

Current Movie Industry Analysis

Author: Stephen Gomes

Summary

The task assigned is to advise Microsoft on their launch of a new movie studio. They are in need of insight as to the type of movies to produce in order to be most successful at the box office. There are many levers we can pull when deciding on making a movie - total budget, genre of movie, staff (directors, writers, actors, run length, release date, etc.

After exploration I focussed on budget, genre, run length and release scheduling as the main drivers.

The recommendations I make based on data analysis are:

- Focus your studio on Action & Adventure movies as these genres are the most profitable
- Release films primarily in Winter & Summer seasons, as box office sales show moviegoing is highest in these seasons
- Budget up to 300M per film, as ROI and Profit increases with budget until 300M at which point it drops off
- Keep movie runtime in the sweet spot of 90-120 minutes length as higher run times do not result in higher sales

Business Problem

Microsoft is launching a new business studio and in need of insight as to the types of movies to launch to maximize their successs at the box office. In answering the question of what

maximizes success, I focused on revenue & ROI as the measures of success, and did not use any movie ratings. The thesis I had was that while having well rated movies is nice, it is not something you can directly control for like genre, staffing, budget, or release date. Similarly, the ultimate measure of success in business is sales, not product reviews.

Data Inputs

For this analysis I used the following datasets & sources:

IMDB - Reference data for movie genre

Rotten Tomatoes - Reference data for movie release date, length, writers, diretors, plus box office sales

The Numbers - Movie budget data

Minneapolis Fed - Inflation data

Basics - imports & file loading

In [1]:
```
# Your code here - remember to use markdown cells for comments as well!

# setup all our imports
import json
import pandas as pd
import numpy as np
import matplotlib
import sqlite3
import requests
import matplotlib.pyplot as plt

%matplotlib inline

# clear some warnings on edits to copies of dataslice
pd.options.mode.chained_assignment = None  # default='warn'
```

In [2]:
```
# read in all our input files
info = pd.read_csv('zippedData/rt.movie_info.tsv.gz', sep="\t") # USED
budgets = pd.read_csv('zippedData/tn.movie_budgets.csv.gz') # USED
# open DB file for query, get list of tables, USED
conn = sqlite3.connect('zippedData/im.db')
tbls = pd.read_sql("""SELECT name FROM sqlite_master WHERE type = 'table';"""

# SUPPLIED INPUTS NOT USED BELOW
# reviews file had a utf8 encoding error with default pandas read encoding
reviews = pd.read_csv('zippedData/rt.reviews.tsv.gz', sep="\t", encoding = "I
bom = pd.read_csv('zippedData/bom.movie_gross.csv.gz') # not used
tmdb = pd.read_csv('zippedData/tmdb.movies.csv.gz') # not used
```

In [3]:
```
tbls
```

Out[3]:

	name
0	movie_basics
1	directors
2	known_for
3	movie_akas
4	movie_ratings
5	persons
6	principals
7	writers

In [4]:
```
# select info from movie_basics table to enrich budgets with genre
movie_basics = pd.read_sql("""SELECT primary_title AS movie,start_year AS yea
conn.close()
```

Introducing a new dataset - CPI

We are comparing & aggregating dollar figures across time for budgets and box office sales. When comparing dollar figures it is important to use "real" not "nominal" numbers, and therefore to inflation adjust across years to a fixed year. In this case we chose current year, 2022, as the baseline. So we use inflation data from the Minneapolis Fed, who have a convenient consumer price index time series going back to the early 1900s.

In [5]:
```
# we need to adjust historical dollars to 2022 level, let's get CPI data from
# webscrape directly with pandas & requests
# grab CPI of Minn. Fed site
html_page = requests.get('https://www.minneapolisfed.org/about-us/monetary-po
webdf_list = pd.read_html(html_page.text) # pull the table from the html
webdf = webdf_list[0]
webdf
```

Out[5]:

	Year	Annual Average CPI(-U)	Annual Percent Change (rate of inflation)
0	1913	9.9	NaN
1	1914	10.0	1.3%
2	1915	10.1	0.9%
3	1916	10.9	7.7%
4	1917	12.8	17.8%
...	...	...	...
105	2018	251.1	2.4%
106	2019	255.7	1.8%
107	2020	258.8	1.2%
108	2021	271.0	4.7%
109	2022*	294.4	8.6%

110 rows × 3 columns

Data cleaning & manipulation

CPI data

Using the CPI timeseries, we create a multiplier that adjusts each year to the 2022 dollar figure. We use this later on all tables containing dollar figures.

In [6]:

```
# cleanup CPI data
# rename CPI column, remove inflation rate, create annual multiplier column w
cpi = webdf.rename({'Year' : 'year' , 'Annual Average CPI(-U)': 'CPI'}, axis=
cpi = cpi.iloc[: , :-1]
cpi['CPI_mult'] = cpi.iloc[-1]['CPI':].div(cpi.loc[:,'CPI':]).astype(np.float
# clean up current year * label, cast to int
cpi.loc[cpi['year'] == '2022*', 'year'] = '2022'
cpi['year'] = cpi['year'].astype(np.int64)
cpi
```

Out[6]:

	year	CPI	CPI_mult
0	1913	9.9	29.737374
1	1914	10.0	29.440000
2	1915	10.1	29.148515
3	1916	10.9	27.009174
4	1917	12.8	23.000000
...	...	...	...
105	2018	251.1	1.172441
106	2019	255.7	1.151349
107	2020	258.8	1.137558
108	2021	271.0	1.086347
109	2022	294.4	1.000000

110 rows × 3 columns

Budget table

Using the provided movie budget data from The Numbers we do the following

- Cleanup data types for date and dollar figures
- Join on movie_basics table from IMDB, which contains title->genre mapping
- Remove rows we cannot find a genre for
- Join on CPI data from MN Fed, and adjust the dollar figure columns in new 2022 tagged columns
- Compute ROI
- Bucket budgets into 100M buckets

In [7]:
```
budgets
```

Out[7]:

	id	release_date	movie	production_budget	domestic_gross	worldwide_gross
0	1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
1	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
2	3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
3	4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
4	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747
...	...	...	...	...	...	...
5777	78	Dec 31, 2018	Red 11	$7,000	$0	$0
5778	79	Apr 2, 1999	Following	$6,000	$48,482	$240,495
5779	80	Jul 13, 2005	Return to the Land of Wonders	$5,000	$1,338	$1,338
5780	81	Sep 29, 2015	A Plague So Pleasant	$1,400	$0	$0
5781	82	Aug 5, 2005	My Date With Drew	$1,100	$181,041	$181,041

5782 rows × 6 columns

In [8]:
```
movie_basics
```

Out[8]:

	movie	year	genres
0	Sunghursh	2013	Action,Crime,Drama
1	One Day Before the Rainy Season	2019	Biography,Drama
2	The Other Side of the Wind	2018	Drama
3	Sabse Bada Sukh	2018	Comedy,Drama
4	The Wandering Soap Opera	2017	Comedy,Drama,Fantasy
...	...	...	...
146139	Kuambil Lagi Hatiku	2019	Drama
146140	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	Documentary
146141	Dankyavar Danka	2013	Comedy
146142	6 Gunn	2017	None
146143	Chico Albuquerque - Revelações	2013	Documentary

146144 rows × 3 columns

In [9]:

```
# cleanup budgets data
clean_budgets = budgets
clean_budgets['release_date'] = pd.to_datetime(clean_budgets['release_date'])
clean_budgets['year'] = clean_budgets['release_date'].dt.year
clean_budgets = clean_budgets.merge(movie_basics,on=['year','movie'],how='lef
clean_budgets = clean_budgets.loc[~clean_budgets['genres'].isnull()]
# join CPI info & inflate $ to 2022 figures
clean_budgets = clean_budgets.merge(cpi, on='year', how='left')
clean_budgets["worldwide_gross"] = clean_budgets["worldwide_gross"].replace("
clean_budgets["production_budget"] = clean_budgets["production_budget"].repla
clean_budgets['gross2022'] = clean_budgets['worldwide_gross'] * clean_budgets
clean_budgets['budget2022'] = clean_budgets['production_budget'] * clean_budg
clean_budgets['net2022'] = clean_budgets['gross2022'] - clean_budgets['budget
clean_budgets["roi"] = 100 * clean_budgets["net2022"] / clean_budgets["budget

# bin movie budgets into 50M buckets
bins = [0, 0.5e8, 1e8, 1.5e8, 2e8, 2.5e8, 3e8, 3.5e8, 4e8, 4.5e8, 5e8, 5.5e8,
blabel = [0.5e8, 1e8, 1.5e8, 2e8, 2.5e8, 3e8, 3.5e8, 4e8, 4.5e8, 5e8, 5.5e8,


clean_budgets['budget2022bin'] = pd.cut(clean_budgets['budget2022'], bins,lab

clean_budgets
```

Out[9]:

	id	release_date	movie	production_budget	domestic_gross	worldwide_gross	yea
0	2	2011-05-20	Pirates of the Caribbean: On Stranger Tides	410600000.0	$241,063,875	1.045664e+09	201
1	3	2019-06-07	Dark Phoenix	350000000.0	$42,762,350	1.497624e+08	201
2	4	2015-05-01	Avengers: Age of Ultron	330600000.0	$459,005,868	1.403014e+09	201
3	7	2018-04-27	Avengers: Infinity War	300000000.0	$678,815,482	2.048134e+09	201
4	9	2017-11-17	Justice League	300000000.0	$229,024,295	6.559452e+08	201
...	...	...	...	...	...	...	.
1536	45	2017-01-27	Emily	27000.0	$3,547	3.547000e+03	201
1537	49	2015-09-01	Exeter	25000.0	$0	4.897920e+05	201
1538	52	2015-12-01	Dutch Kills	25000.0	$0	0.000000e+00	201
1539	59	2011-11-25	The Ridges	17300.0	$0	0.000000e+00	201
1540	62	2014-12-31	Stories of Our Lives	15000.0	$0	0.000000e+00	201

1541 rows × 15 columns

Info table

Using the provided movie info data from Rotten Tomatoes we do the following:

- Remove blank box_office sales and non-dollar box office numbers
- Cleanup data types for dates, $ numbers and runtime
- Create year, month & season columns
- Join on CPI data from MN Fed, and adjust the dollar figure columns in new 2022 tagged columns

In [10]:
```
info
```

Out[10]:

	id	synopsis	rating	genre	director	writer	t
		This gritty,					

0	1	fast-paced, and innovative police...	R	Action and Adventure|Classics|Drama	William Friedkin	Ernest Tidyman	
1	3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	
2	5	Illeana Douglas delivers a superb performance ...	R	Drama|Musical and Performing Arts	Allison Anders	Allison Anders	
3	6	Michael Douglas runs afoul of a treacherous su...	R	Drama|Mystery and Suspense	Barry Levinson	Paul Attanasio|Michael Crichton	
4	7	NaN	NR	Drama|Romance	Rodney Bennett	Giles Cooper	
...	...	...	...	...	...	...	
1555	1996	Forget terrorists or hijackers -- there's a ha...	R	Action and Adventure|Horror|Mystery and Suspense	NaN	NaN	
1556	1997	The popular Saturday Night Live sketch was exp...	PG	Comedy|Science Fiction and Fantasy	Steve Barron	Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner	
1557	1998	Based on a novel by Richard Powell, when the l...	G	Classics|Comedy|Drama|Musical and Performing Arts	Gordon Douglas	NaN	
1558	1999	The Sandlot is a coming-of-age story about a g...	PG	Comedy|Drama|Kids and Family|Sports and Fitness	David Mickey Evans	David Mickey Evans|Robert Gunter	
1559	2000	Suspended from the force, Paris cop Hubert is ...	R	Action and Adventure|Art House and Internation...	NaN	Luc Besson	

1560 rows × 12 columns

In [11]:
```
# cleanup info table to just non-null box office sales
# has genre, director, genre, runtime
clean_info = info.loc[(~info['box_office'].isnull()) & (info['currency']=='$'
clean_info['theater_date'] = pd.to_datetime(clean_info['theater_date'])
clean_info['year'] = clean_info['theater_date'].dt.year
clean_info['month'] = clean_info['theater_date'].dt.month

# join CPI info & inflate $ to 2022 figures
clean_info = clean_info.merge(cpi, on='year', how='left')
clean_info["box_office"] = clean_info["box_office"].replace("[$,]", "", regex
clean_info['box2022'] = clean_info['box_office'] * clean_info['CPI_mult']
clean_info['runtime'] = clean_info['runtime'].replace("[minutes]", "", regex=

# map months to seasons
m2s = {1.0: "Winter", 2.0: "Winter", 3.0: "Spring", 4.0: "Spring", 5.0: "Spri
clean_info['season'] = clean_info['month'].map(m2s)
clean_info
```

Out[11]:

	id	synopsis	rating	genre	director	writer	theate
0	3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	2012
1	10	Some cast and crew from NBC's highly acclaimed...	PG-13	Comedy	Jake Kasdan	Mike White	200
2	13	Stewart Kane, an Irishman living in the Austra...	R	Drama	Ray Lawrence	Raymond Carver|Beatrix Christian	2006
3	14	"Love Ranch" is a bittersweet love story that ...	R	Drama	Taylor Hackford	Mark Jacobson	2010
4	22	Two-time Academy Award Winner Kevin Spacey giv...	R	Comedy|Drama|Mystery and Suspense	George Hickenlooper	Norman Snider	201
...	...	...	...	...	...	...	

335	1980	A band of renegades on the run in outer space ...	PG-13	Action and Adventure|Science Fiction and Fantasy	Joss Whedon	Joss Whedon	2005
336	1981	Money, Fame and the Knowledge of English. In I...	NR	Comedy|Drama	Gauri Shinde	Gauri Shinde	2012
337	1985	A woman who joins the undead against her will ...	R	Horror|Mystery and Suspense	Sebastian Gutierrez	Sebastian Gutierrez	2007
338	1986	Aki Kaurismaki's The Man Without a Past opens ...	PG	Art House and International|Comedy|Drama	NaN	NaN	2002
339	1996	Forget terrorists or hijackers -- there's a ha...	R	Action and Adventure|Horror|Mystery and Suspense	NaN	NaN	2006

340 rows × 18 columns

Genre cleanup

Before analyzing by genre, we need to expand out the genre column to its consituent genres.

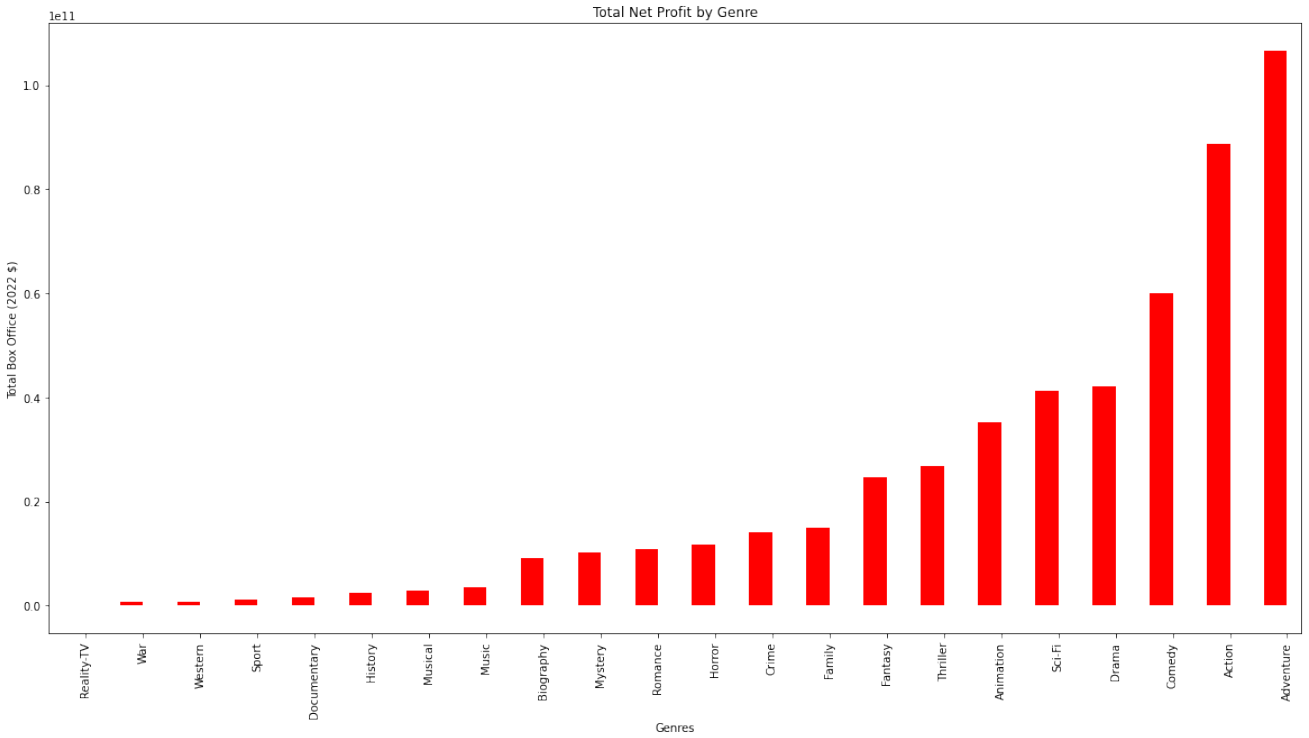
The movies are tagged with all the applicable genres that pertain to the film, but we want to analyze each genre.

In [12]:
```
# explode out multi-genre films to each of their component genres
g2n = clean_budgets[['genres','net2022']]
g2n['genres']=g2n['genres'].apply(lambda x : x.split(','))
g2n = g2n.join(pd.concat([g2n.pop('genres').explode()],axis=1))
```

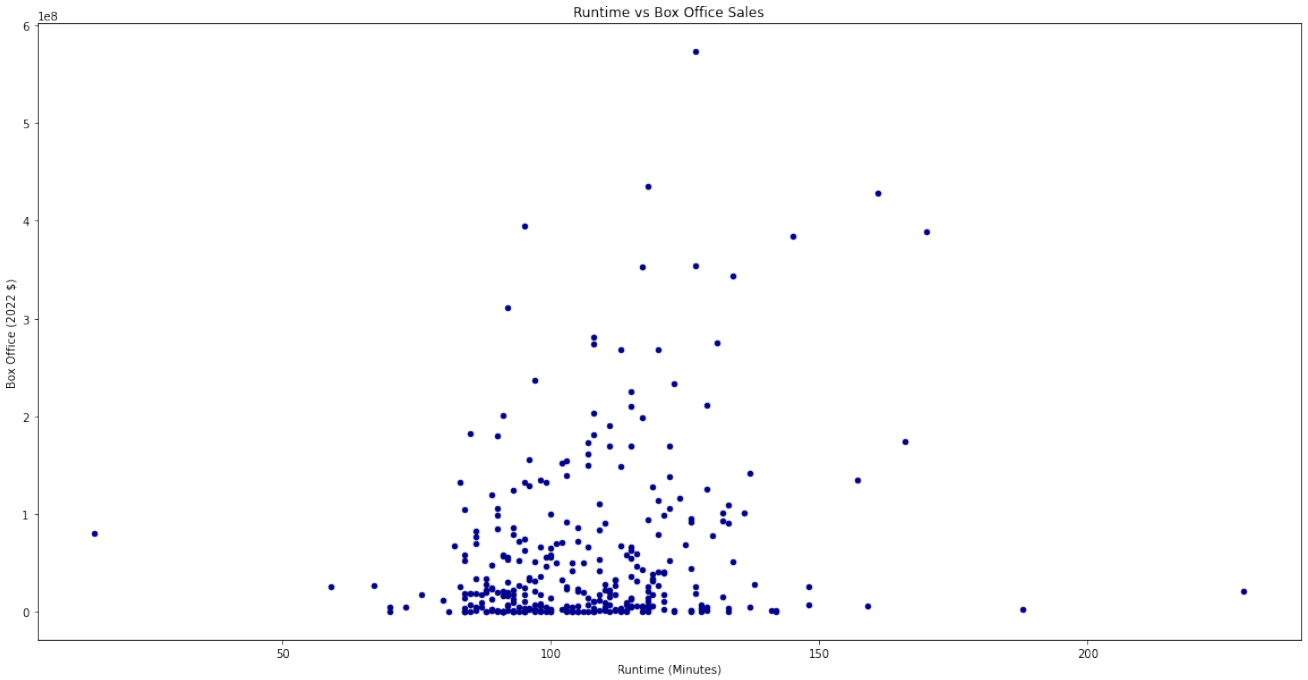
Genre vs Profit visualizations

- Immediately this bar chart makes a lot of sense and has clear takeaways

In [13]:
```
# visualize total box office sales by genre
tot_net = g2n[['genres','net2022']].groupby('genres').sum()
tot_net = tot_net.sort_values(by=['net2022'])
plt.rcParams['figure.figsize'] = (20, 10)
fig = plt.figure() # Create matplotlib figure
ax0 = fig.add_subplot(111) # Create matplotlib axes
width = 0.4
tot_net.net2022.plot(kind='bar', color='red', ax=ax0, width=width, position=1
ax0.set_title('Total Net Profit by Genre')
ax0.set_xlabel('Genres')
ax0.set_ylabel('Total Box Office (2022 $)')
plt.show()
```



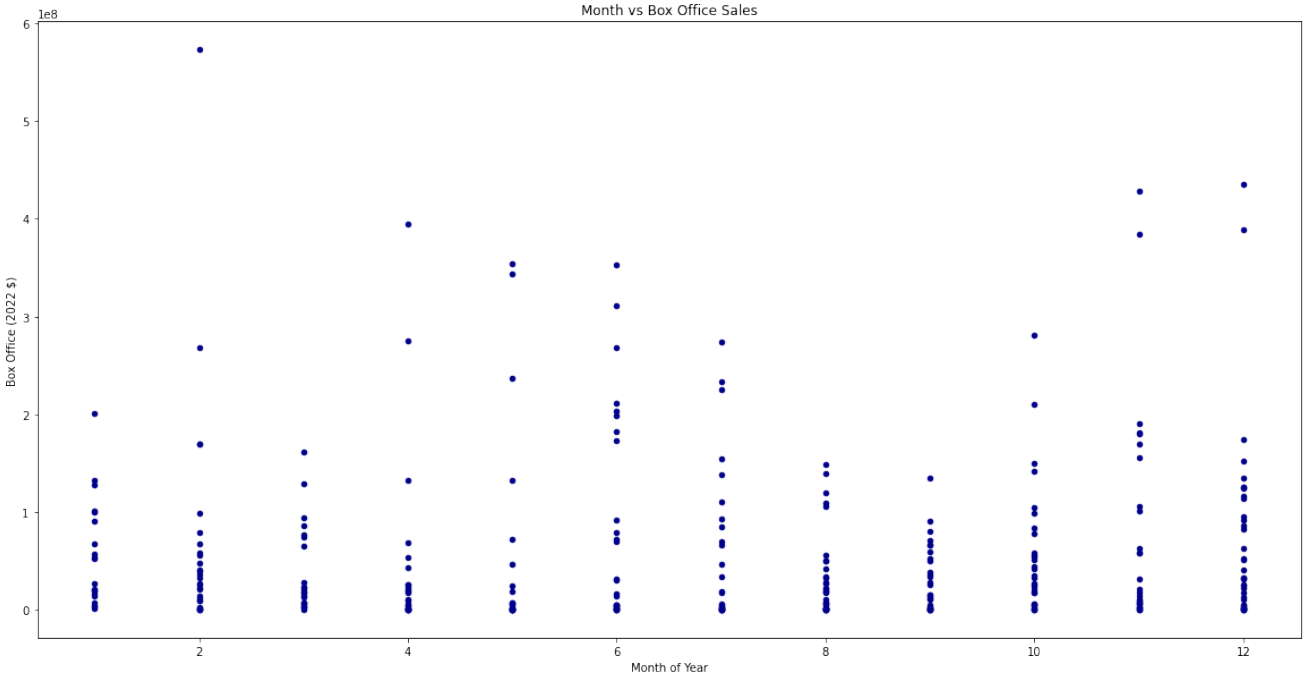
In [14]:
```
ax1 = clean_info.plot.scatter(x='runtime',y='box2022',c='DarkBlue')
ax1.set_title('Runtime vs Box Office Sales')
ax1.set_xlabel('Runtime (Minutes)')
ax1.set_ylabel('Box Office (2022 $)')
plt.show()
```



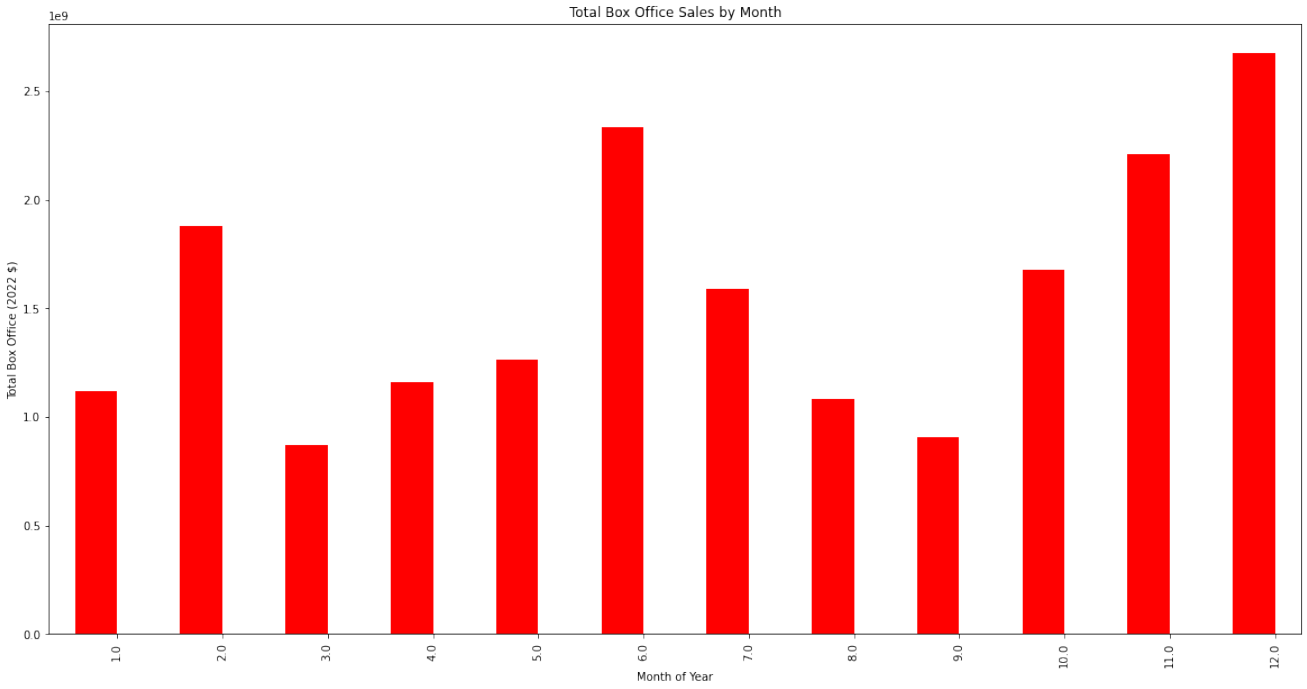
Release date visualizations

- First we try a scatter of Month vs Box Office Sales.. hard to read!
- Next we try a bar chart for Month vs Box Office Sales, interesting but noisy
- Finally we settle on mapping the months into seasons and visualizing Season vs Box Office Sales - nice

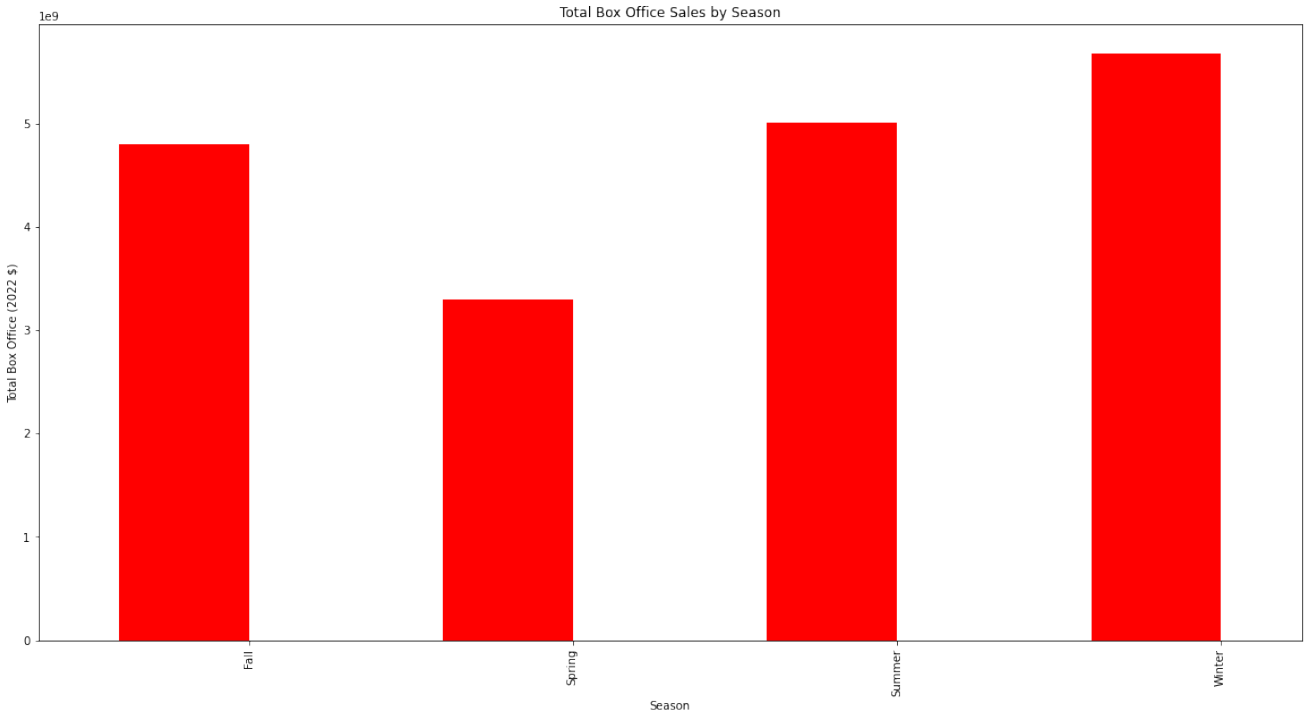
In [15]:
```
# Try Month vs Sales scatter
ax2 = clean_info.plot.scatter(x='month',y='box2022',c='DarkBlue')
ax2.set_title('Month vs Box Office Sales')
ax2.set_xlabel('Month of Year')
ax2.set_ylabel('Box Office (2022 $)')
plt.show()
```



In [16]:
```
# Try Total Box office by Month of Year
# visualize total box office sales by genre
tot_mo = clean_info[['month','box2022']].groupby('month').sum()
plt.rcParams['figure.figsize'] = (20, 10)
fig = plt.figure() # Create matplotlib figure
ax3 = fig.add_subplot() # Create matplotlib axes
tot_mo.box2022.plot(kind='bar', color='red', ax=ax3, width=width, position=1)
ax3.set_title('Total Box Office Sales by Month')
ax3.set_xlabel('Month of Year')
ax3.set_ylabel('Total Box Office (2022 $)')
plt.show()
```



In [17]:
```
# Try Total Box office by Season of Year
# visualize total box office sales by genre
tot_season = clean_info[['season','box2022']].groupby('season').sum()
plt.rcParams['figure.figsize'] = (20, 10)
fig = plt.figure() # Create matplotlib figure
ax4 = fig.add_subplot() # Create matplotlib axes
tot_season.box2022.plot(kind='bar', color='red', ax=ax4, width=width, positio
ax4.set_title('Total Box Office Sales by Season')
ax4.set_xlabel('Season')
ax4.set_ylabel('Total Box Office (2022 $)')
plt.show()
```

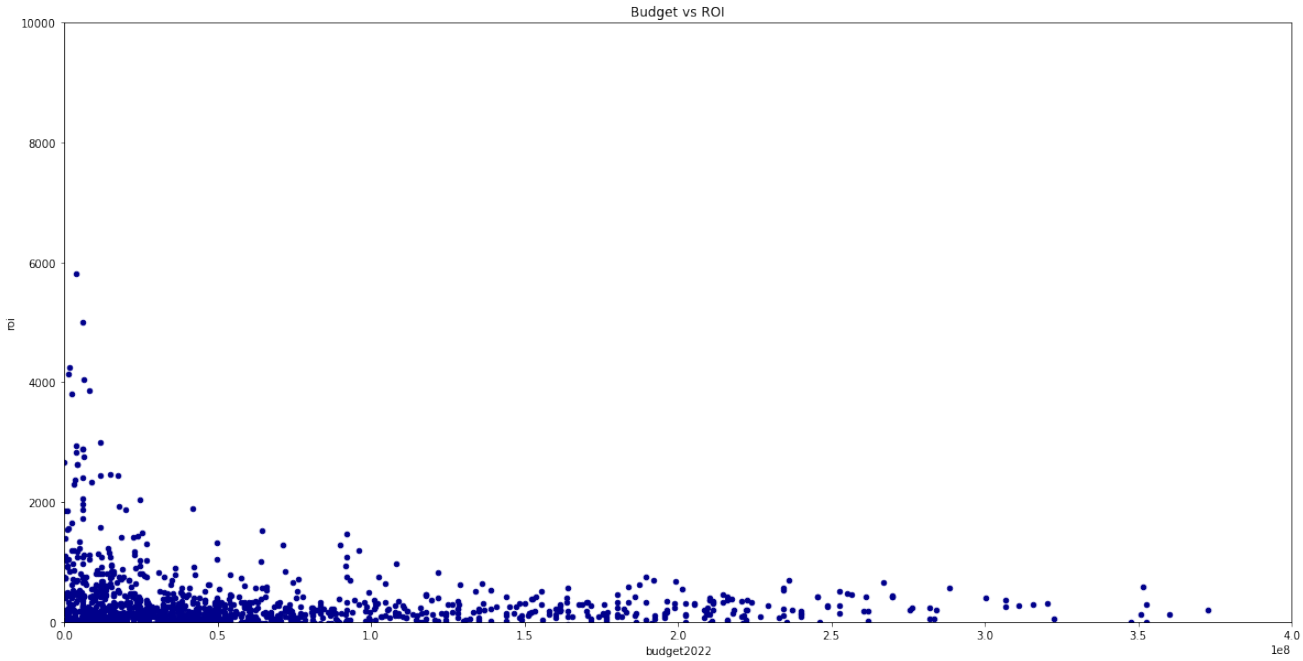


Budget vs ROI visualization

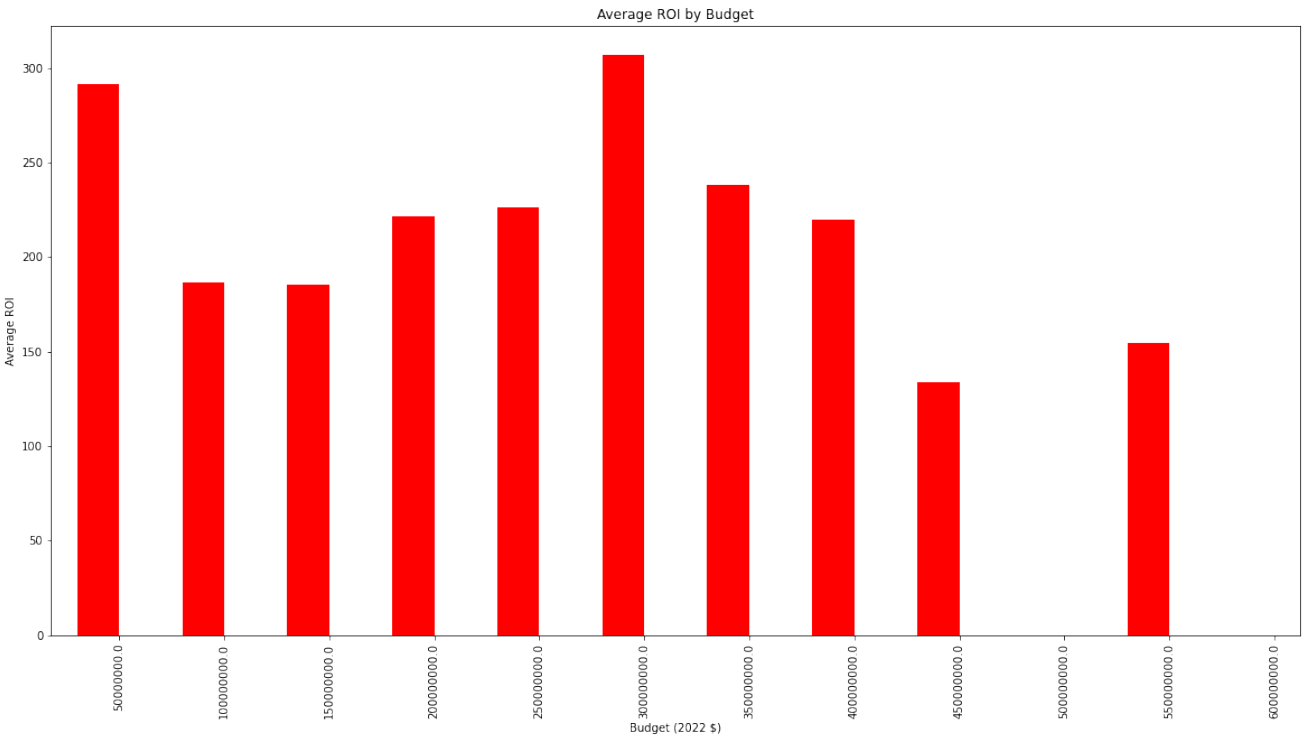
- First we try a scatter plot, and find it a bit hard to read
- Next we try bucketing the budgets and doing a bar chart - much better!

In [18]:
```
#budget2022 vs #net2022

# clean_budgets
ax5 = clean_budgets.plot.scatter(x='budget2022',y='roi',c='DarkBlue')
plt.xlim(0, 4e8)
plt.ylim(0, 10000)
ax5.set_title('Budget vs ROI')
ax2.set_xlabel('Budget (2022 $)')
ax2.set_ylabel('ROI')
plt.show()
```



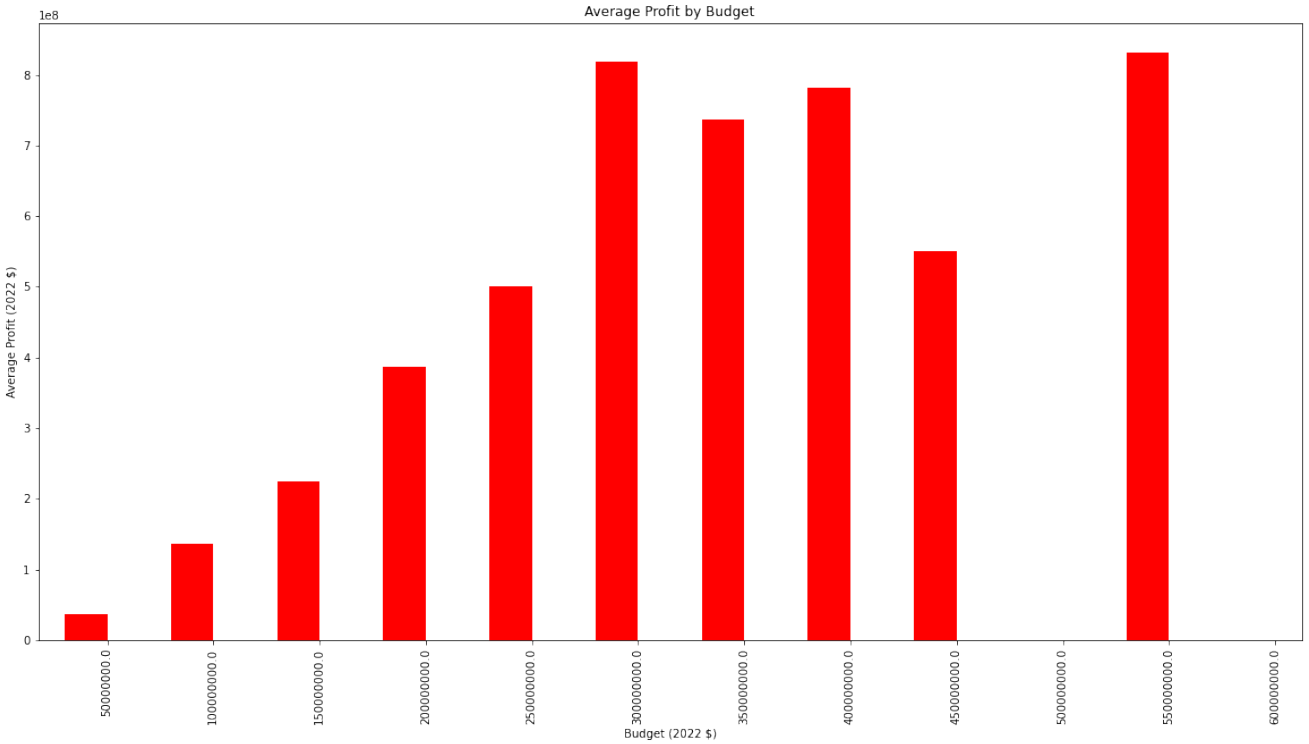
In [19]:
```
# bucketed ROI by budget graph
avg_roi = clean_budgets[['budget2022bin','roi']].groupby('budget2022bin').mea
plt.rcParams['figure.figsize'] = (20, 10)
fig = plt.figure() # Create matplotlib figure
ax6 = fig.add_subplot() # Create matplotlib axes
avg_roi.roi.plot(kind='bar', color='red', ax=ax6, width=width, position=1)
ax6.set_title('Average ROI by Budget')
ax6.set_xlabel('Budget (2022 $)')
ax6.set_ylabel('Average ROI')
plt.show()
```



Budget vs Profit visualization

- The bucketed budget vs ROI bar chart makes sense, so lets use the same to look at Budget vs Profit
- This makes the relationship of mony in to mony out even more clear

In [20]:
```
# bucketed ROI by budget graph
avg_net = clean_budgets[['budget2022bin','net2022']].groupby('budget2022bin')
plt.rcParams['figure.figsize'] = (20, 10)
fig = plt.figure() # Create matplotlib figure
ax7 = fig.add_subplot() # Create matplotlib axes
avg_net.net2022.plot(kind='bar', color='red', ax=ax7, width=width, position=1
ax7.set_title('Average Profit by Budget')
ax7.set_xlabel('Budget (2022 $)')
ax7.set_ylabel('Average Profit (2022 $)')
plt.show()
```



Recommendations & Next Steps

Recommendation

- Focus your studio on Action & Adventure movies as these genres are the most profitable
- Release films primarily in Winter & Summer seasons, as box office sales show moviegoing is highest in these seasons
- Budget up to 300M per film, as ROI and Profit increases with budget until 300M at which point it drops off
- Keep movie runtime in the sweet spot of 90-120 minutes length as higher run times do not result in higher sales

Next Steps

Multi-factor drill downs such as:

- Optimal budget for Action & Adventure genres individually
- Drill into day-of-week and week-of-year time slices
- Look at how time of year & budget interact
- Etc

In [21]:
```
from IPython.display import HTML
from IPython.display import IFrame

IFrame(src="https://www.youtube.com/embed/b9434BoGkNQ", width="560", height="
```

Out[21]:

In [ ]: