Final Project Submission

Please fill out:

- Student name: Ryan Keats
- Student pace: self paced / part time / full time
- Scheduled project review date/time: September 11, 11:59pm.
- Instructor name: Hardik Idnani

In [1]:
```
# Firstly, I have imported pandas and numpy to help me work on my d
import pandas as pd
import numpy as np

# I have imported matplotlib to be able to peform my graphs.
import matplotlib.pyplot as plt

# I have adjusted the data sets with this code to clean the numbers
import pandas as pd
pd.set_option('display.float_format', lambda x: '${:,.2f}'.format(x
```

In [2]:
```
# This is the first data set I have used and defined to start my pr
movie_info = pd.read_csv("zippedData/bom.movie_gross.csv.gz")
movie_info.head(20)
```

Out[2]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	$415,000,000.00	652000000	2010
1	Alice in Wonderland (2010)	BV	$334,200,000.00	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	$296,000,000.00	664300000	2010
3	Inception	WB	$292,600,000.00	535700000	2010
4	Shrek Forever After	P/DW	$238,700,000.00	513900000	2010
5	The Twilight Saga: Eclipse	Sum.	$300,500,000.00	398000000	2010
6	Iron Man 2	Par.	$312,400,000.00	311500000	2010
7	Tangled	BV	$200,800,000.00	391000000	2010
8	Despicable Me	Uni.	$251,500,000.00	291600000	2010
9	How to Train Your Dragon	P/DW	$217,600,000.00	277300000	2010
10	Clash of the Titans (2010)	WB	$163,200,000.00	330000000	2010
11	The Chronicles of Narnia: The Voyage of the Da...	Fox	$104,400,000.00	311300000	2010
12	The King's Speech	Wein.	$135,500,000.00	275400000	2010
13	Tron Legacy	BV	$172,100,000.00	228000000	2010
14	The Karate Kid	Sony	$176,600,000.00	182500000	2010
15	Prince of Persia: The Sands of Time	BV	$90,800,000.00	245600000	2010
16	Black Swan	FoxS	$107,000,000.00	222400000	2010
17	Megamind	P/DW	$148,400,000.00	173500000	2010
18	Robin Hood	Uni.	$105,300,000.00	216400000	2010
19	The Last Airbender	Par.	$131,800,000.00	187900000	2010

In [3]:
```
# I have used .iloc to display another way I searched the data fram
movie_info.iloc[0:10,:]
```

Out[3]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	$415,000,000.00	652000000	2010
1	Alice in Wonderland (2010)	BV	$334,200,000.00	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	$296,000,000.00	664300000	2010
3	Inception	WB	$292,600,000.00	535700000	2010
4	Shrek Forever After	P/DW	$238,700,000.00	513900000	2010
5	The Twilight Saga: Eclipse	Sum.	$300,500,000.00	398000000	2010
6	Iron Man 2	Par.	$312,400,000.00	311500000	2010
7	Tangled	BV	$200,800,000.00	391000000	2010
8	Despicable Me	Uni.	$251,500,000.00	291600000	2010
9	How to Train Your Dragon	P/DW	$217,600,000.00	277300000	2010

In [4]:
```
# I have used .columns to outline the columns of the data set.
movie_info.columns
```

Out[4]:
```
Index(['title', 'studio', 'domestic_gross', 'foreign_gross', 'year
'], dtype='object')
```

In [5]:
```
# I have chosen .shape to show the size of the first data set I am
movie_info.shape
```

Out[5]:
```
(3387, 5)
```

In [6]:
```
# I have sorted the coloumn domestic_gross here, displaying from hi
movie_info = movie_info.sort_values('domestic_gross', ascending = F
movie_info.head(20)
```

Out[6]:

	title	studio	domestic_gross	foreign_gross	year
1872	Star Wars: The Force Awakens	BV	$936,700,000.00	1,131.6	2015
3080	Black Panther	BV	$700,100,000.00	646900000	2018
3079	Avengers: Infinity War	BV	$678,800,000.00	1,369.5	2018
1873	Jurassic World	Uni.	$652,300,000.00	1,019.4	2015
727	Marvel's The Avengers	BV	$623,400,000.00	895500000	2012
2758	Star Wars: The Last Jedi	BV	$620,200,000.00	712400000	2017
3082	Incredibles 2	BV	$608,600,000.00	634200000	2018
2323	Rogue One: A Star Wars Story	BV	$532,200,000.00	523900000	2016
2759	Beauty and the Beast (2017)	BV	$504,000,000.00	759500000	2017
2324	Finding Dory	BV	$486,300,000.00	542300000	2016
1875	Avengers: Age of Ultron	BV	$459,000,000.00	946400000	2015
729	The Dark Knight Rises	WB	$448,100,000.00	636800000	2012
1131	The Hunger Games: Catching Fire	LGF	$424,700,000.00	440300000	2013
3081	Jurassic World: Fallen Kingdom	Uni.	$417,700,000.00	891800000	2018
0	Toy Story 3	BV	$415,000,000.00	652000000	2010
2767	Wonder Woman	WB	$412,600,000.00	409300000	2017
1128	Iron Man 3	BV	$409,000,000.00	805800000	2013
2322	Captain America: Civil War	BV	$408,100,000.00	745200000	2016
735	The Hunger Games	LGF	$408,000,000.00	286400000	2012
2762	Jumanji: Welcome to the Jungle	Sony	$404,500,000.00	557600000	2017

In [7]:
```
# I have added the .value_counts() of the studio's production infor
movie_info['studio'].value_counts().head(10)
```

Out[7]:
```
IFC      166
Uni.     147
WB       140
Fox      136
Magn.    136
SPC      123
Sony     110
BV       106
LGF      103
Par.     101
Name: studio, dtype: int64
```

In [8]:
```
# I have also added .describe() to show general information of the
movie_info['studio'].describe()
```

Out[8]:
```
count     3382
unique     257
top        IFC
freq       166
Name: studio, dtype: object
```

In [9]:
```
# Make column names easier to use
movie_info.columns = movie_info.columns.str.lower().str.replace(' '

# Drop unnecessary columns
movie_info.drop(columns = ['studio', 'foreign_gross'], inplace=True
```

In [10]:
```
# I wanted to show the annual domestic_gross average, by using .mea
movie_info.groupby('year').mean()
```

Out[10]:

	domestic_gross
year	
2010	$31,445,592.57
2011	$25,350,524.43
2012	$27,675,842.23
2013	$31,282,115.64
2014	$26,439,228.90
2015	$24,613,375.04
2016	$25,989,960.96
2017	$34,166,456.87
2018	$36,010,421.75

In [11]:
```
# In addition to showing the average annual domestic_gross, I wante
movie_info.groupby('year').sum()
```

Out[11]:

	domestic_gross
year	
2010	$10,156,926,399.00
2011	$10,064,158,200.00
2012	$10,876,605,997.00
2013	$10,792,329,897.00
2014	$10,337,738,499.00
2015	$11,051,405,394.00
2016	$11,253,653,097.00
2017	$10,933,266,198.00
2018	$11,091,209,899.00

In [12]:
```
# I have adjusted the data sets with this code to clean the numbers
import pandas as pd
pd.set_option('display.float_format', lambda x: '{:,.2f}'.format(x)
```

In [13]:
```
# I wanted to get a description of the domestic_gross and the Inner
movie_info['domestic_gross'].describe()
```

Out[13]:
```
count          3,359.00
mean      28,745,845.07
std       66,982,498.24
min              100.00
25%          120,000.00
50%        1,400,000.00
75%       27,900,000.00
max      936,700,000.00
Name: domestic_gross, dtype: float64
```

In [14]:
```
# I have grouped the movies by year, to illustrate their yearly pro
movie_info.groupby('year').count().describe()
```

Out[14]:

	title	domestic_gross
count	9.00	9.00
mean	376.33	373.22
std	51.44	51.23
min	308.00	308.00
25%	328.00	323.00
50%	395.00	391.00
75%	400.00	397.00
max	450.00	449.00

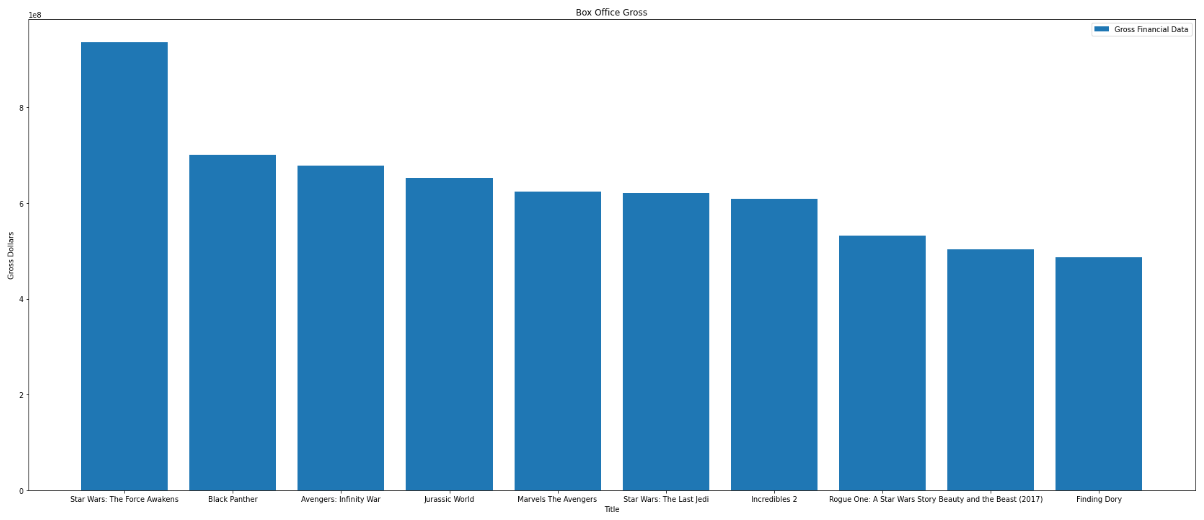
In [15]:
```
# I have added the value_counts() before .mean() to get the average
# Which is also displayed with a different code in the above cell.
movie_info['year'].value_counts().mean()
```

Out[15]:
```
376.3333333333333
```

In [16]:
```
# I have shown how many movies are made each with .value_counts() a
movie_info['year'].value_counts().sort_values()
```

Out[16]:
```
2018    308
2017    321
2010    328
2013    350
2014    395
2011    399
2012    400
2016    436
2015    450
Name: year, dtype: int64
```

In [17]:
```
# I have chosen to use bar graphs as I feel they are the most simpl

y = [ 936700000.0, 700100000.0, 678800000.0, 652300000.0, 623400000
    608600000.0, 532200000.0, 504000000.0, 486300000.0 ]
x = range(10)
labels = ['Star Wars: The Force Awakens', 'Black Panther', 'Avenger
          'Marvels The Avengers', 'Star Wars: The Last Jedi', 'Incr
        'Beauty and the Beast (2017)', 'Finding Dory' ]

# Create the plot
fig, ax = plt.subplots(figsize=(29,12))


ax.bar(x, y, tick_label = labels)

ax.set_title('Box Office Gross')
ax.set_ylabel('Gross Dollars')
ax.set_xlabel('Title');

ax.legend(['Gross Financial Data'], loc=1);

plt.style.use('ggplot')
```



In [18]:
```
# I wanted to add a command/comments cell gap between each of my da
# Also, for clarity and uniformity to display that I am moving onto
```

In [19]:
```
#I have imported pandas and numpy to help me work on my data sets.
import pandas as pd
import numpy as np

# I have imported matplotlib to be able to peform my graphs.
import matplotlib.pyplot as plt

# I have adjusted the data sets with this code to clean the numbers
import pandas as pd
pd.set_option('display.float_format', lambda x: '{:,.2f}'.format(x)
```

In [20]:
```
#This is the second data set I have used to get my findings from.
movie_info2=pd.read_csv("zippedData/imdb.title.basics.csv.gz")
movie_info2.head()
```

Out[20]:

	tconst	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.00	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.00	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.00	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.00	Comedy,Drama,Fantasy

In [21]:
```
# I have used .columns to outline the columns of the data set.
movie_info2.columns
```

Out[21]:
```
Index(['tconst', 'primary_title', 'original_title', 'start_year',
       'runtime_minutes', 'genres'],
      dtype='object')
```

In [22]:
```
# I have chosen .shape to show the size of the first data set I am
movie_info2.shape
```

Out[22]:
```
(146144, 6)
```

In [23]:
```
# Make column names easier to use
movie_info2.columns = movie_info2.columns.str.lower().str.replace('

# Drop unnecessary columns
movie_info2.drop(columns = ['tconst', 'original_title'], inplace=Tr
```

In [24]:
```
#I wanted to use the .value_count() feature to display the top genr
movie_info2['genres'].value_counts().head(10)
```

Out[24]:
```
Documentary                 32185
Drama                       21486
Comedy                       9177
Horror                       4372
Comedy,Drama                 3519
Thriller                     3046
Action                       2219
Biography,Documentary        2115
Drama,Romance                2079
Comedy,Drama,Romance         1558
Name: genres, dtype: int64
```

In [25]:
```
#I have used .describe() on the runtime to outline the (IQR).
movie_info2['runtime_minutes'].describe()
```

Out[25]:
```
count   114,405.00
mean         86.19
std         166.36
min           1.00
25%          70.00
50%          87.00
75%          99.00
max      51,420.00
Name: runtime_minutes, dtype: float64
```

In [26]:
```
# I wanted to add a command/comments cell gap between each of my da
# Also, for clarity and uniformity to display that I am moving onto
```

In [27]:
```
#I have imported pandas and numpy to help me work on my data sets.
import pandas as pd
import numpy as np

# I have imported matplotlib to be able to peform my graphs.
import matplotlib.pyplot as plt
```

In [28]:
```
#This is the third data set I have used to get my findings from.
movie_info3=pd.read_csv("zippedData/imdb.title.ratings.csv.gz")
movie_info3.head()
```

Out[28]:

	tconst	averagerating	numvotes
0	tt10356526	8.30	31
1	tt10384606	8.90	559
2	tt1042974	6.40	20
3	tt1043726	4.20	50352
4	tt1060240	6.50	21

In [29]:
```
# I have used .columns to outline the columns of the data set.
movie_info3.columns
```

Out[29]:
```
Index(['tconst', 'averagerating', 'numvotes'], dtype='object')
```

In [30]:
```
# I have chosen .shape to show the size of the third data set I am
movie_info3.shape
```

Out[30]:
```
(73856, 3)
```

In [31]:
```
# Make column names easier to use
movie_info3.columns = movie_info3.columns.str.lower().str.replace('

# Drop unnecessary columns
movie_info3.drop(columns = ['tconst'], inplace=True )
```

In [32]:
```
# I have sorted the coloumn numvotes here, displaying from high to
movie_info3 = movie_info3.sort_values('numvotes', ascending = False
movie_info3.head(10)
```

Out[32]:

	averagerating	numvotes
63498	8.80	1841066
8738	8.40	1387769
24920	8.60	1299334
38058	8.40	1211405
48221	8.10	1183655
39356	8.20	1035358
3140	8.10	1005960
25777	8.10	948394
60518	8.00	820847
63506	7.20	795227

In [33]:
```
# I have displayed a different way to produce the above information
movie_info3 = movie_info3.sort_values(by=['numvotes','averagerating
movie_info3.tail(10)
```

Out[33]:

	averagerating	numvotes
63506	7.20	795227
60518	8.00	820847
25777	8.10	948394
3140	8.10	1005960
39356	8.20	1035358
48221	8.10	1183655
38058	8.40	1211405
24920	8.60	1299334
8738	8.40	1387769
63498	8.80	1841066

In [34]:
```
# I have displayed the IQR with this code.
movie_info3 = movie_info3.sort_values('averagerating', ascending =
movie_info3.describe()
```

Out[34]:

	averagerating	numvotes
count	73,856.00	73,856.00
mean	6.33	3,523.66
std	1.47	30,294.02
min	1.00	5.00
25%	5.50	14.00
50%	6.50	49.00
75%	7.40	282.00
max	10.00	1,841,066.00