Data Analysis Project 1

Step 1: ASK

1) What topic are you exploring?
    - Top 250 movies according to IMDB (Internet Movie Database)
    - Company represents clients that are interested in success in movie industry.
    - The type of data appropriate for my analysis is title of movies, budget, year released, budget, total box office, genre, and rating.
    - I will be obtaining my data from a dataset from Kaggle. This dataset collected was from Chidambara where he got the data from IMBD website.
2) Business Task: I will perform a data analysis to answer questions involving this dataset about the top 250 movies according to IMBD.
    - What is the most common genre amongst this dataset?
    - What does the distribution of films look like by rating?
    - How does the budget of the movie compare to the total box office collection across the world?
    - Over time how much top movies by rating compare to the future?
3) What are your tools that will be used?
    - I will be cleaning up and manipulating the data using Microsoft Excel and through Tableau I will be supporting visualizations and key findings.

Step 2: Prepare

1) Where is the data located?
    - The data is from Kaggle directly from the IMBD website.

2) Are there issues with bias or credibility?

   - The only issue with bias here is that these movies are rated only from the IMBD website and doesn't account for other movie sites such as Rotten Tomatoes or other movie review websites. I will only be studying data from the IMBD dataset.

3) How do you verify the data's integrity?

   - This dataset is from Kaggle a reliable source for grabbing datasets.

4) Are there any problems with the data?

   - The only problem with the data is formatting issues and useless data columns that we will not be using because it won't help us answer our questions.

5) How is the data organized?

   - The data is organized by ranking of the top 250 movies according to IMBD database. The ranking is sorted by rating of a scale 1 to 10 with 10 being the top rated.

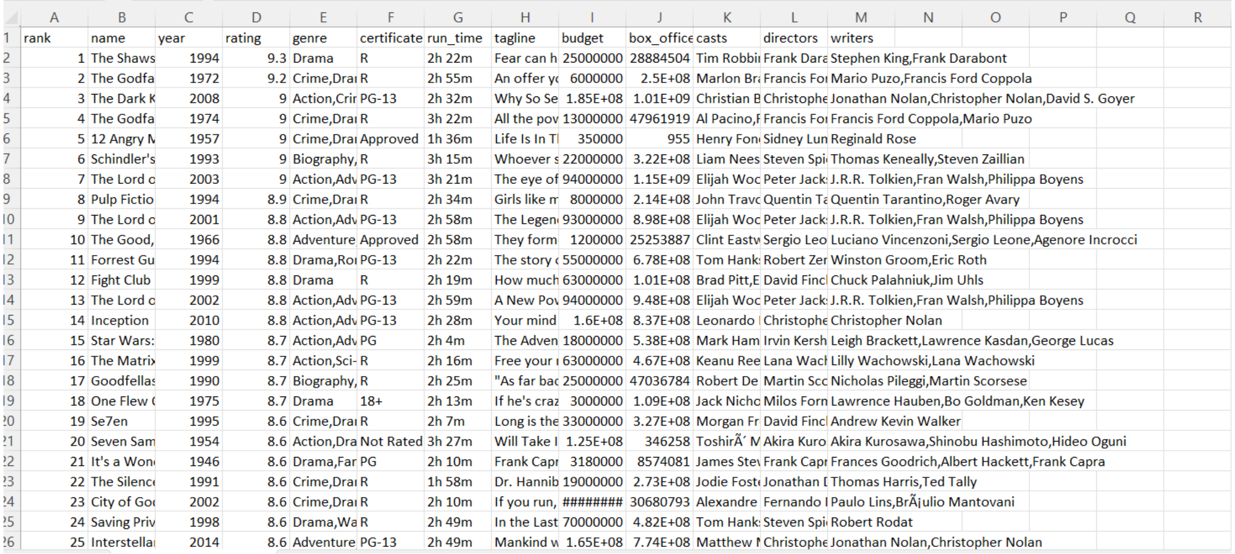
Step 3: Process

1) What tools are you choosing and why?

   - The tools that I will be using for this analysis are Microsoft Excel and Tableau to display my data analysis. I chose Microsoft Excel because this is not that big of a dataset and chose Tableau for its efficiency and graphics.
   - The steps I took to clean the data is first checking for duplicates in the table using the delete duplicate's function. Secondly, I check for formatting issues and columns that are useless to my analysis. With those columns I simply removed them from my table. I changed the budget column from currency to numerical numbers so that it would easier creating graphical representations later when I analyze the data. The next step I parsed out the genre column because it had multiple variables. I don't need to look at

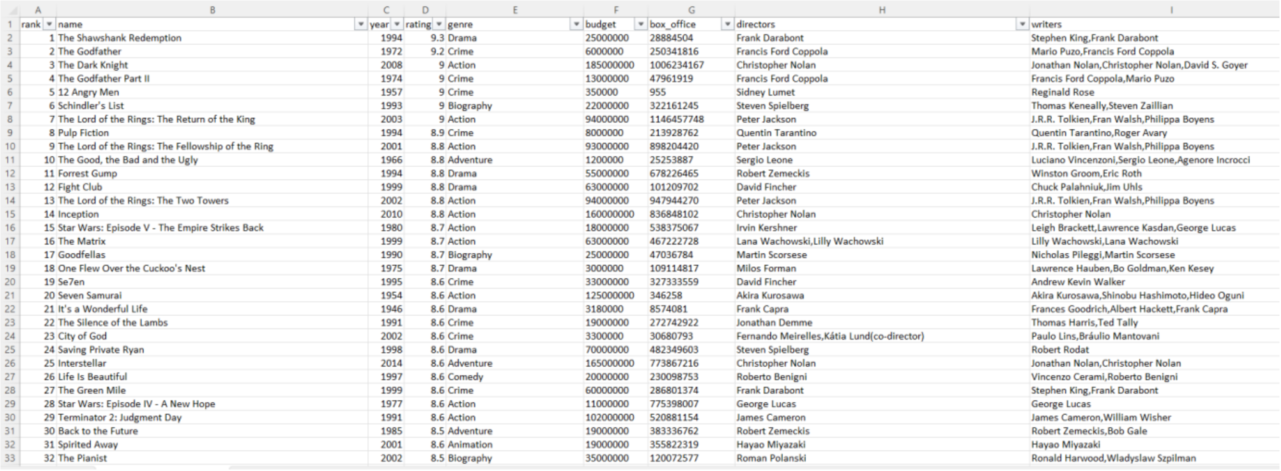
more than one genre, so I parsed out the secondary genre that the movie was categorized in.

Before Data Cleaning:



rank	name	year	rating	genre	certificate	run_time	tagline	budget	box_office	casts	directors	writers
1	The Shaws	1994	9.3	Drama	R	2h 22m	Fear can h	25000000	28884504	Tim Robbi	Frank Dara	Stephen King,Frank Darabont
2	The Godfa	1972	9.2	Crime,Dra	R	2h 55m	An offer yo	6000000	2.5E+08	Marlon Bra	Francis For	Mario Puzo,Francis Ford Coppola
3	The Dark K	2008	9	Action,Crir	PG-13	2h 32m	Why So Se	1.85E+08	1.01E+09	Christian B	Christophe	Jonathan Nolan,Christopher Nolan,David S. Goyer
4	The Godfa	1974	9	Crime,Dra	R	3h 22m	All the pov	13000000	47961919	Al Pacino,F	Francis For	Francis Ford Coppola,Mario Puzo
5	12 Angry M	1957	9	Crime,Dra	Approved	1h 36m	Life Is In Tl	350000	955	Henry Fon	Sidney Lun	Reginald Rose
6	Schindler's	1993	9	Biography,	R	3h 15m	Whoever s	22000000	3.22E+08	Liam Nees	Steven Spi	Thomas Keneally,Steven Zaillian
7	The Lord o	2003	9	Action,Adv	PG-13	3h 21m	The eye of	94000000	1.15E+09	Elijah Woo	Peter Jack	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
8	Pulp Fictio	1994	8.9	Crime,Dra	R	2h 34m	Girls like m	8000000	2.14E+08	John Travo	Quentin Ta	Quentin Tarantino,Roger Avary
9	The Lord o	2001	8.8	Action,Adv	PG-13	2h 58m	The Legen	93000000	8.98E+08	Elijah Woo	Peter Jack	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
10	The Good,	1966	8.8	Adventure	Approved	2h 58m	They form	1200000	25253887	Clint Eastw	Sergio Leo	Luciano Vincenzoni,Sergio Leone,Agenore Incrocci
11	Forrest Gu	1994	8.8	Drama,Ro	PG-13	2h 22m	The story o	55000000	6.78E+08	Tom Hank	Robert Zer	Winston Groom,Eric Roth
12	Fight Club	1999	8.8	Drama	R	2h 19m	How much	63000000	1.01E+08	Brad Pitt,E	David Fincl	Chuck Palahniuk,Jim Uhls
13	The Lord o	2002	8.8	Action,Adv	PG-13	2h 59m	A New Pov	94000000	9.48E+08	Elijah Woo	Peter Jack	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
14	Inception	2010	8.8	Action,Adv	PG-13	2h 28m	Your mind	1.6E+08	8.37E+08	Leonardo	Christophe	Christopher Nolan
15	Star Wars:	1980	8.7	Action,Adv	PG	2h 4m	The Adven	18000000	5.38E+08	Mark Ham	Irvin Kersh	Leigh Brackett,Lawrence Kasdan,George Lucas
16	The Matrix	1999	8.7	Action,Sci-	R	2h 16m	Free your	63000000	4.67E+08	Keanu Ree	Lana Wacl	Lilly Wachowski,Lana Wachowski
17	Goodfellas	1990	8.7	Biography,	R	2h 25m	"As far bac	25000000	47036784	Robert De	Martin Scc	Nicholas Pileggi,Martin Scorsese
18	One Flew (	1975	8.7	Drama	18+	2h 13m	If he's craz	3000000	1.09E+08	Jack Nicho	Milos Forn	Lawrence Hauben,Bo Goldman,Ken Kesey
19	Se7en	1995	8.6	Crime,Dra	R	2h 7m	Long is the	33000000	3.27E+08	Morgan Fr	David Fincl	Andrew Kevin Walker
20	Seven Sam	1954	8.6	Action,Dra	Not Rated	3h 27m	Will Take I	1.25E+08	346258	ToshirÃ´ M	Akira Kuro	Akira Kurosawa,Shinobu Hashimoto,Hideo Oguni
21	It's a Won	1946	8.6	Drama,Far	PG	2h 10m	Frank Capr	3180000	8574081	James Stev	Frank Capr	Frances Goodrich,Albert Hackett,Frank Capra
22	The Silenc	1991	8.6	Crime,Dra	R	1h 58m	Dr. Hannib	19000000	2.73E+08	Jodie Fost	Jonathan D	Thomas Harris,Ted Tally
23	City of Go	2002	8.6	Crime,Dra	R	2h 10m	If you run,	########	30680793	Alexandre	Fernando I	Paulo Lins,BrÃ¡ulio Mantovani
24	Saving Priv	1998	8.6	Drama,Wa	R	2h 49m	In the Last	70000000	4.82E+08	Tom Hank	Steven Spi	Robert Rodat
25	Interstella	2014	8.6	Adventure	PG-13	2h 49m	Mankind w	1.65E+08	7.74E+08	Matthew M	Christophe	Jonathan Nolan,Christopher Nolan

After Data Cleaning:



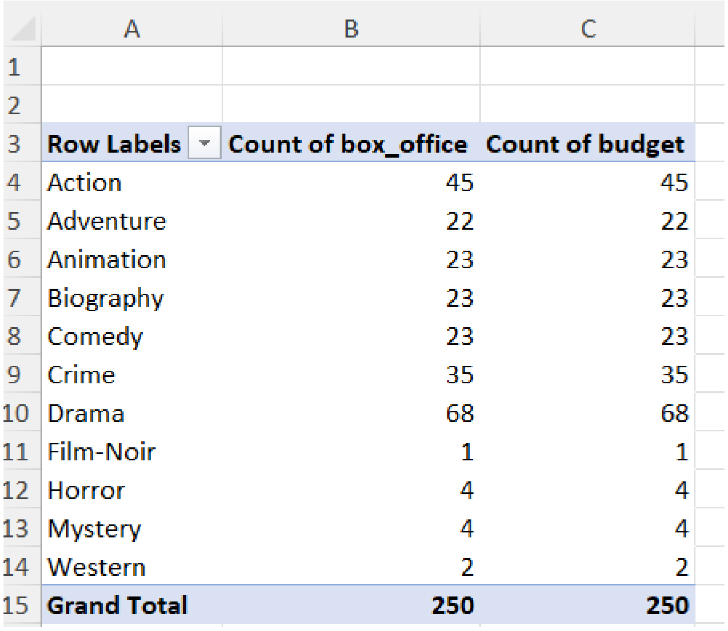
rank	name	year	rating	genre	budget	box_office	directors	writers
1	The Shawshank Redemption	1994	9.3	Drama	25000000	28884504	Frank Darabont	Stephen King,Frank Darabont
2	The Godfather	1972	9.2	Crime	6000000	250341816	Francis Ford Coppola	Mario Puzo,Francis Ford Coppola
3	The Dark Knight	2008	9	Action	185000000	1006234167	Christopher Nolan	Jonathan Nolan,Christopher Nolan,David S. Goyer
4	The Godfather Part II	1974	9	Crime	13000000	47961919	Francis Ford Coppola	Francis Ford Coppola,Mario Puzo
5	12 Angry Men	1957	9	Crime	350000	955	Sidney Lumet	Reginald Rose
6	Schindler's List	1993	9	Biography	22000000	322161245	Steven Spielberg	Thomas Keneally,Steven Zaillian
7	The Lord of the Rings: The Return of the King	2003	9	Action	94000000	1146457748	Peter Jackson	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
8	Pulp Fiction	1994	8.9	Crime	8000000	213928762	Quentin Tarantino	Quentin Tarantino,Roger Avary
9	The Lord of the Rings: The Fellowship of the Ring	2001	8.8	Action	93000000	898204420	Peter Jackson	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
10	The Good, the Bad and the Ugly	1966	8.8	Adventure	1200000	25253887	Sergio Leone	Luciano Vincenzoni,Sergio Leone,Agenore Incrocci
11	Forrest Gump	1994	8.8	Drama	55000000	678226465	Robert Zemeckis	Winston Groom,Eric Roth
12	Fight Club	1999	8.8	Drama	63000000	101209702	David Fincher	Chuck Palahniuk,Jim Uhls
13	The Lord of the Rings: The Two Towers	2002	8.8	Action	94000000	947944270	Peter Jackson	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
14	Inception	2010	8.8	Action	160000000	836848102	Christopher Nolan	Christopher Nolan
15	Star Wars: Episode V - The Empire Strikes Back	1980	8.7	Action	18000000	538375067	Irvin Kershner	Leigh Brackett,Lawrence Kasdan,George Lucas
16	The Matrix	1999	8.7	Action	63000000	467222728	Lana Wachowski,Lilly Wachowski	Lilly Wachowski,Lana Wachowski
17	Goodfellas	1990	8.7	Biography	25000000	47036784	Martin Scorsese	Nicholas Pileggi,Martin Scorsese
18	One Flew Over the Cuckoo's Nest	1975	8.7	Drama	3000000	109114817	Milos Forman	Lawrence Hauben,Bo Goldman,Ken Kesey
19	Se7en	1995	8.6	Crime	33000000	327333559	David Fincher	Andrew Kevin Walker
20	Seven Samurai	1954	8.6	Action	125000000	346258	Akira Kurosawa	Akira Kurosawa,Shinobu Hashimoto,Hideo Oguni
21	It's a Wonderful Life	1946	8.6	Drama	3180000	8574081	Frank Capra	Frances Goodrich,Albert Hackett,Frank Capra
22	The Silence of the Lambs	1991	8.6	Crime	19000000	272742922	Jonathan Demme	Thomas Harris,Ted Tally
23	City of God	2002	8.6	Crime	3300000	30680793	Fernando Meirelles,Kátia Lund(co-director)	Paulo Lins,Bráulio Mantovani
24	Saving Private Ryan	1998	8.6	Drama	70000000	482349603	Steven Spielberg	Robert Rodat
25	Interstellar	2014	8.6	Adventure	165000000	773867216	Christopher Nolan	Jonathan Nolan,Christopher Nolan
26	Life Is Beautiful	1997	8.6	Comedy	20000000	230098753	Roberto Benigni	Vincenzo Cerami,Roberto Benigni
27	The Green Mile	1999	8.6	Crime	60000000	286801374	Frank Darabont	Stephen King,Frank Darabont
28	Star Wars: Episode IV - A New Hope	1977	8.6	Action	11000000	775398007	George Lucas	George Lucas
29	Terminator 2: Judgment Day	1991	8.6	Action	102000000	520881154	James Cameron	James Cameron,William Wisher
30	Back to the Future	1985	8.5	Adventure	19000000	383336762	Robert Zemeckis	Robert Zemeckis,Bob Gale
31	Spirited Away	2001	8.6	Animation	19000000	355822319	Hayao Miyazaki	Hayao Miyazaki
32	The Pianist	2002	8.5	Biography	35000000	120072577	Roman Polanski	Ronald Harwood,Wladyslaw Szpilman

Step 4: Analyze

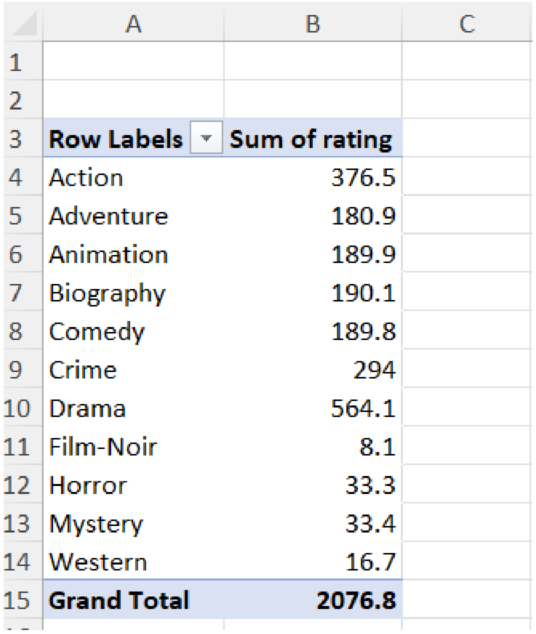
1) How should you organize your data to perform analysis on it?

   - The data was organized by proper sorting and filtering the data.

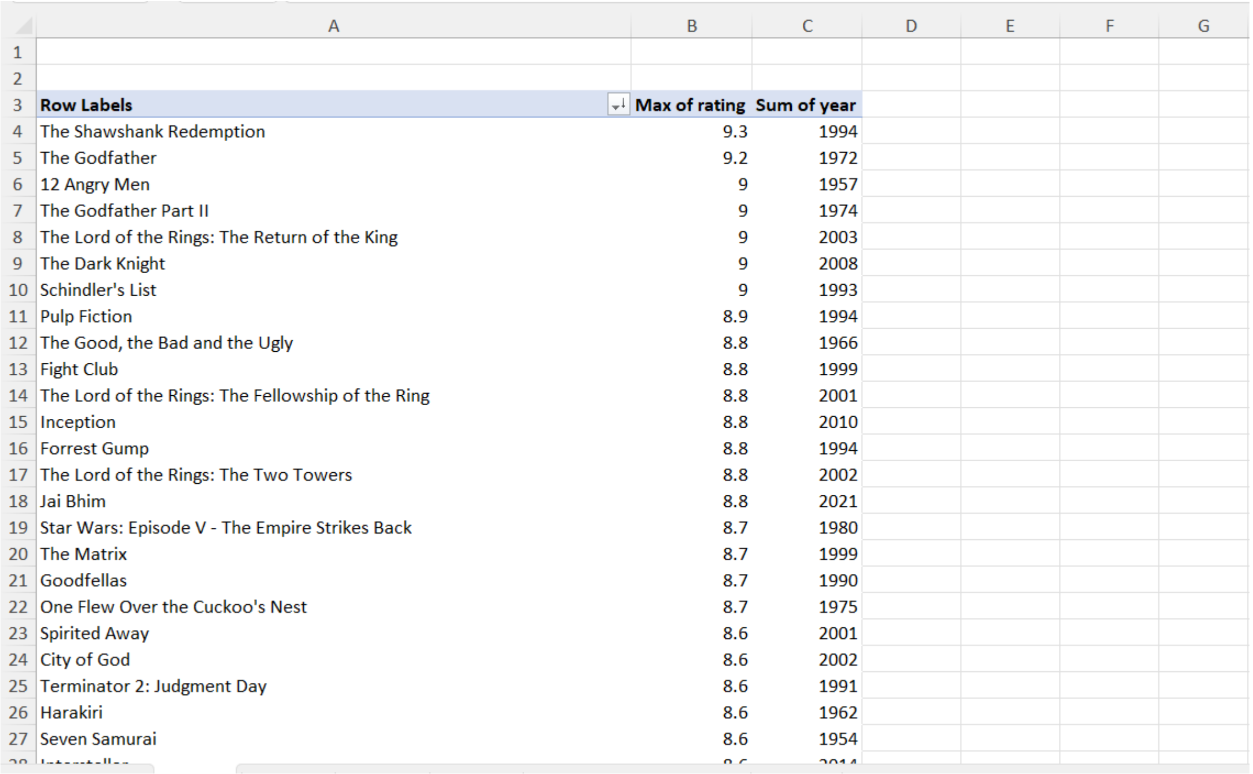
2) Data was analyzed using a pivot table with the title of the movies and the budget as rows and ratings and the year the movie made as values. After analysis of the pivot table, the top-rated movie is The Shawshank Redemption made in 1994 with a budget of 25000000. The top 10 movies only had 2 of them coming from before the 2000's. Another pivot table compares the genre and ratings show that the top 3 categories of movies were Drama, Action, and Biography with Drama having the most top rated movies. The most box-office movies were Avengers: Endgame, Avengers: Infinity War, and Spider-Man: No Way Home. The highest budget movie was Princess Mononoke with a budget of $240,00,00,000. From the data, the more recent movies from the 2000's spent more money than the movies made in the before the 2000's. I also created a new column called total revenue that calculated how much the movie had with the budget subtracted by how much the movie's box office was.



	A	B	C
1			
2			
3	**Row Labels**	**Count of box_office**	**Count of budget**
4	Action	45	45
5	Adventure	22	22
6	Animation	23	23
7	Biography	23	23
8	Comedy	23	23
9	Crime	35	35
10	Drama	68	68
11	Film-Noir	1	1
12	Horror	4	4
13	Mystery	4	4
14	Western	2	2
15	**Grand Total**	**250**	**250**



	A	B	C
1			
2			
3	**Row Labels**	**Sum of rating**	
4	Action	376.5	
5	Adventure	180.9	
6	Animation	189.9	
7	Biography	190.1	
8	Comedy	189.8	
9	Crime	294	
10	Drama	564.1	
11	Film-Noir	8.1	
12	Horror	33.3	
13	Mystery	33.4	
14	Western	16.7	
15	**Grand Total**	**2076.8**	



	A	B	C	D	E	F	G
1							
2							
3	**Row Labels**	**Max of rating**	**Sum of year**				
4	The Shawshank Redemption	9.3	1994				
5	The Godfather	9.2	1972				
6	12 Angry Men	9	1957				
7	The Godfather Part II	9	1974				
8	The Lord of the Rings: The Return of the King	9	2003				
9	The Dark Knight	9	2008				
10	Schindler's List	9	1993				
11	Pulp Fiction	8.9	1994				
12	The Good, the Bad and the Ugly	8.8	1966				
13	Fight Club	8.8	1999				
14	The Lord of the Rings: The Fellowship of the Ring	8.8	2001				
15	Inception	8.8	2010				
16	Forrest Gump	8.8	1994				
17	The Lord of the Rings: The Two Towers	8.8	2002				
18	Jai Bhim	8.8	2021				
19	Star Wars: Episode V - The Empire Strikes Back	8.7	1980				
20	The Matrix	8.7	1999				
21	Goodfellas	8.7	1990				
22	One Flew Over the Cuckoo's Nest	8.7	1975				
23	Spirited Away	8.6	2001				
24	City of God	8.6	2002				
25	Terminator 2: Judgment Day	8.6	1991				
26	Harakiri	8.6	1962				
27	Seven Samurai	8.6	1954				
28	Interstellar	8.6	2014				

Step 5: Share

1) Were you able to answer the business question?

   - Yes, I was able to answer the business question after analyzing the data.
   - The most common genre amongst the movies was Drama. The comparison with the budget and total box office numbers is that the top movies didn't have to spend as much money to make bigger box office numbers. The distribution amongst the movies is that the older movies before the 2000's were rated higher than movies after the 2000's. Overtime the future movies doesn't look like they are going to surpass the older movies ratings.

2) What story does your data tell?

   - The story that the data tells me is that older movies before the 2000's are harder to surpass for future upcoming movies. It doesn't matter how much a movie has to spend on making the movie because even the top spending movie did not have the biggest box office numbers. Also, for upcoming movies the specific genre that looks to make the most success is Drama movies.

3) Who is your audience?
    - My audience is future producers and directors that are looking to make a successful movie.

Step 6: Act

1) How could your team and business apply your insights?
    - They could apply my insights by looking at the numbers and visuals to understand what type of movies make the most success.
    - I was able to display my findings through a Tableau Dashboard titled Top 250 Movies (IMDB) Dashboard. This dashboard is able to show ranking of movies, most popular genre, and a timeline of rankings showing where top rated movies are headed in the future.
    - Link to Tableau Dashboard: https://public.tableau.com/views/Top250MoviesIMDB_17084544490060/Dashboard1?:language=en-US&:sid=&:display_count=n&:origin=viz_share_link