OSTBAYERISCHE TECHNISCHE HOCHSCHULE
REGENSBURG

Informatik und Mathematik

Solving Business Cases Using Applied Data Analytics - ZADA

Data Analysis on Movies from IMDB

Lecturer: Prof. Dr. Jonathan Whitaker
Student: Nghia Le Minh
Student ID: 3397327

07/2023

Table of content

Abstract

The movie industry has been developing non-stop during the last decade or so, bringing in billions of dollars every year. This project will consider the trend of development for the movie industry, considering the genres of the movie, its popularity and the gross it brings in for the producer. The context of this data analysis is that the client is a big movie production company, looking to create a popular movie that can make a name for themselves. Therefore, gross is the biggest criteria that should be considered, and budget is only secondary. Through this analysis, I hope to answer the question, what type of movie is the most suitable to be made now.

The type of analysis that is carried out in this project are:

- General state of the movie industry in the last three decades
- The distribution of movie genres
- Development of gross and profit of top movie genres
- Top actors and directors analysis
- Sub-genres analysis
- Keyword analysis in the description of movies

Datasets

There are two datasets that are used in this analysis.

- "IMDb Movie Dataset: All Movies by Genre": this dataset consisted of multiple tables. There are 16 csv files in total, each containing movies from their respective genres, from 1894 to 2029. Basic information about movies are included in the dataset. However, the budgets of the movies are not included.
    + Values that will be used in this dataset:
        - Movie Name
        - Year
        - Description
        - Director
        - Star
        - Genre
- "Movie Industry": this dataset contains only one file, containing movies from all genres and their respective information. Despite the fact that the previous dataset also contains gross of the movies, the budget and gross from this dataset will be considered to ensure consistency in the analysis.
    + Values that will be used in this dataset:
        - Movie Name
        - Gross
        - Budget

For the loading data process, macro programming was used to modify the csv files in the first dataset and to automate the process without copy-and-pasting the code. The fields of actors and directors in the files contain multi-line values, which cannot be read correctly in the SAS program. Therefore, the file has to be read byte-by-byte first to identify the newline value and change it to the character "|". Furthermore, after combining rows from different files into one table, there are a lot of duplicates from different files, which requires cleaning. The released year is also limited to 1990-2020 to ensure that only recent data is considered.

Formatted datasets

There are two main tables after the loading process, the “full” table and “full_gross” table.

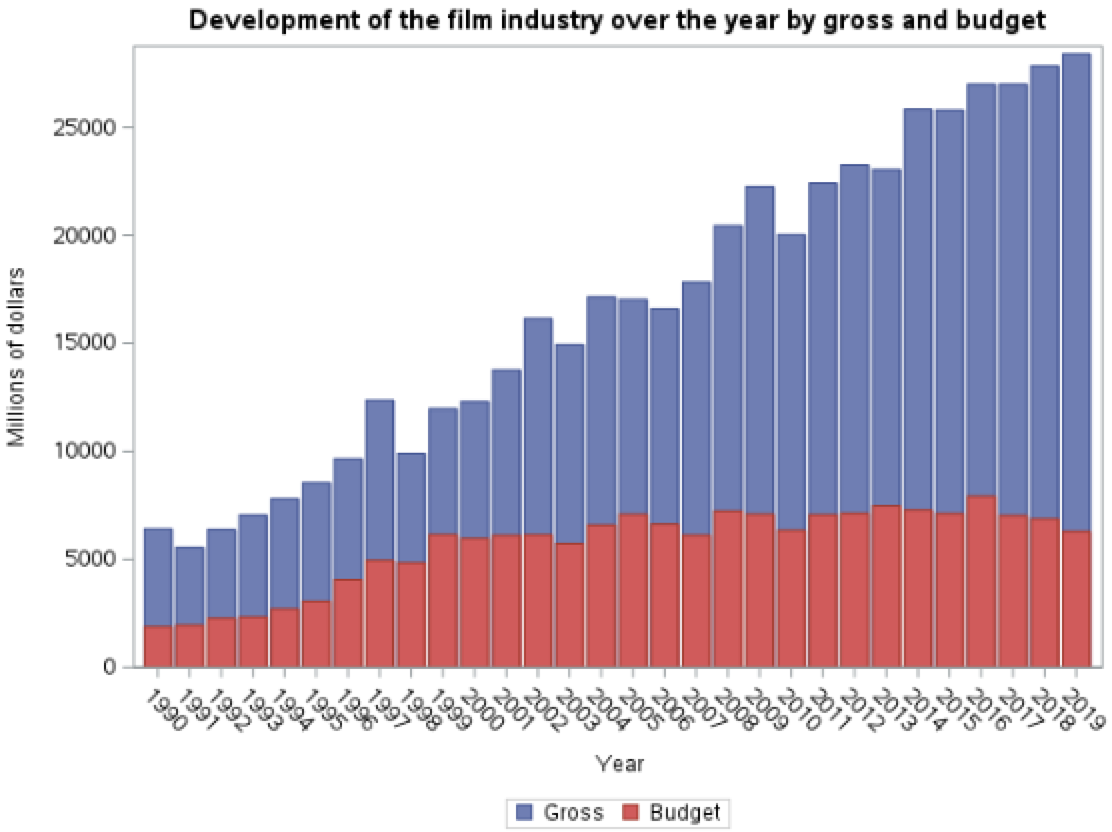
- “full” table:
    - This table contains movies from the first dataset, with duplicates removed. Furthermore, movies were also further filtered to only movies from 1990 to 2020, in order to remove movies that are too old to represent trends now and movies that have not been shown yet. This results in a table with over 100,000 rows, with actors, directors and genres separated into different columns.
    - The purpose of this table is to give a more general look into the movie industry.
    - The table will be used for section 2, 3 and 4.
- “full_gross”:
    - This table contains movies that are contained in both datasets, which has full information about its gross and budget taken from the second dataset. Fuzzy matching was first attempted to create this dataset. However, due to the size of the “full” table, it ends up taking too much time for the procedure. A simple comparison was used and results in a table with over 5000 rows. Additional columns were also created to contain the Logarithm value of gross and budget. This is due to the fact that the gross and budget value is highly skewed to the right.
    - The purpose of this table is to give a more in-depth look into the business of movie making.
    - The table will be used in section 1, 5, 6, 7, 8, 9, 10.

Data Analysis

- Notice: The unit for gross and profit used in the analysis is millions of dollars.

1) Current situation of the movie industry:

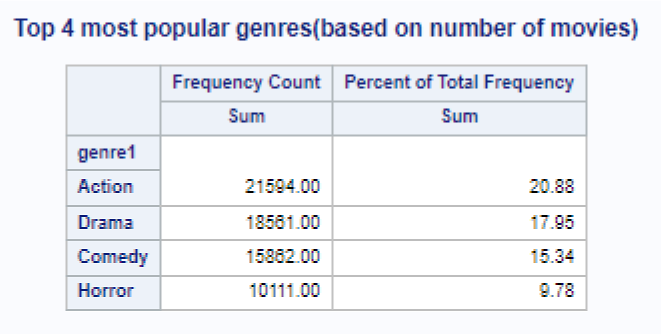
- Method: For this analysis, the gross and budget of movies for every year from 1990 to 2020 is summed up and displayed as bar charts year by year. The column for gross and budget starts at the y-value of 0.
- Result:



- Comment: The budget required to make movies in general seems to have risen slightly since 1990, and fluctuates from 2000 to 2020. In contrast, the gross of movies rise considerably year by year, which makes movie making an attractive industry to participate in now. The continuously rising gross also helps prove that movies are a popular form of media in the present day. This helps to justify the reason for this analysis.

2) Top 4 main genres by number of movie:

- Method: Using the “full” table, the number of movies in each genre are counted and then ranked based on the result. The top 4 genres are then extracted and displayed.
- Result:



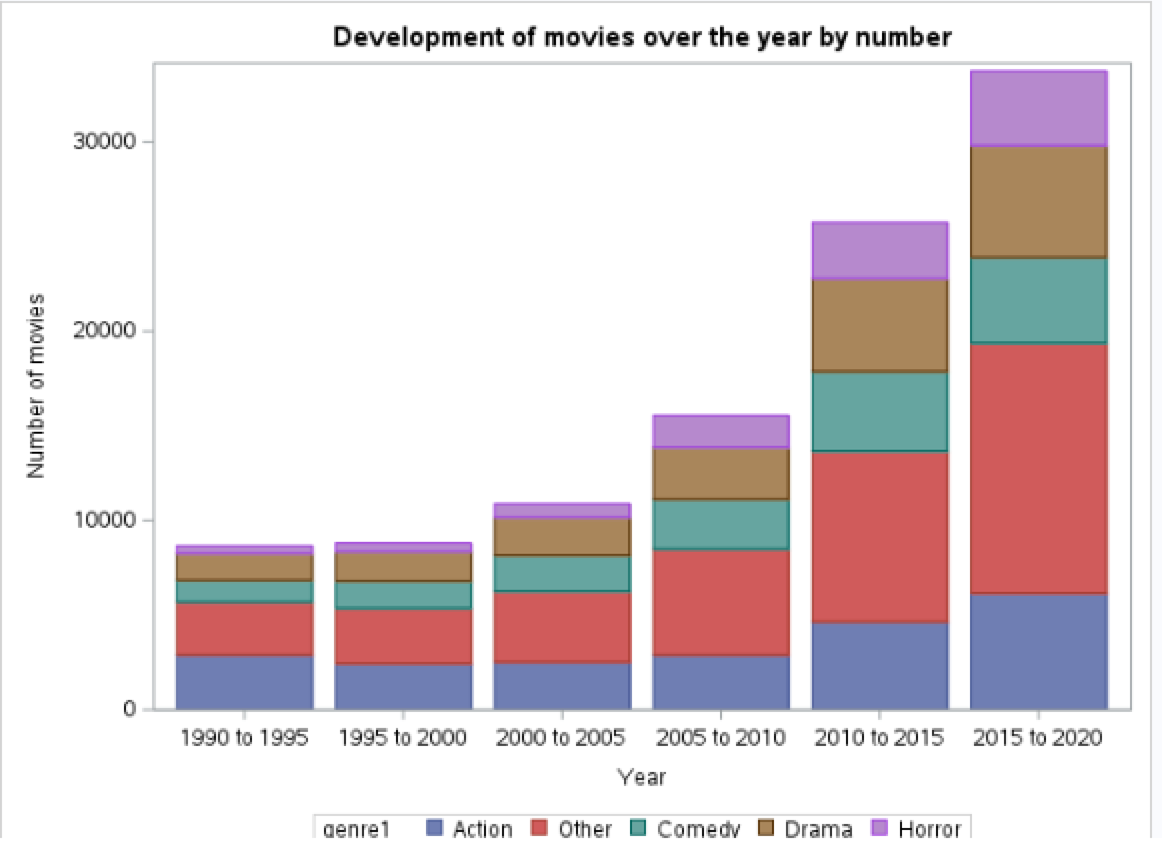
Top 4 most popular genres(based on number of movies)

	Frequency Count	Percent of Total Frequency
	Sum	Sum
genre1		
Action	21594.00	20.88
Drama	18561.00	17.95
Comedy	15862.00	15.34
Horror	10111.00	9.78

  + The top 4 genres are:
    - Action
    - Drama
    - Comedy
    - Horror
- Comment: In this case, the unit to measure the popularity is the number of movies that have been made in that genre. Only the top 4 genres were taken to further analyze because there are a lot of genres in the “full” table, reaching up to more than 20, which makes it impossible to analyze all of them. Therefore, in accordance with the client’s goal, only the popular genres are selected for analysis. The “full” table was selected as it presents a more thorough view of all of the movies.

3) Bar charts of number of movies over the year:

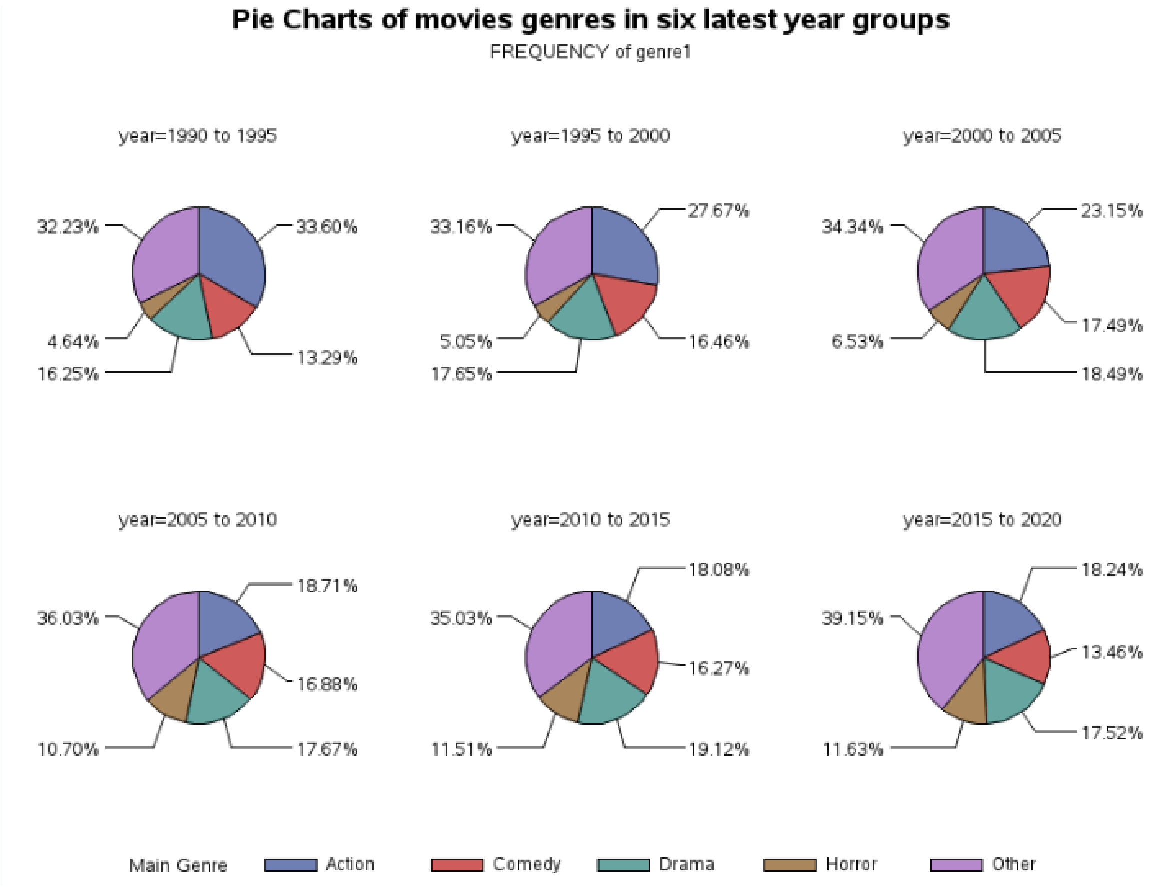
- Method: While the popular genres are kept unchanged, the other genres are grouped together into one category, "Other". All of the movies are then further categorized into groups, representing the time that it was released. Every time group represents a time period of five years, from 1990 to 2020. A bar chart is then constructed from the time group.
- Result:



- Comment: The rise of the number of movies every five years is slow at first, but from 2005 onward, the rise has been more considerable. Furthermore, every genre seems to have had a rise in the number of movies made from 1990 to 2020. The "Horror" genre has had the biggest rise from this period of time. The changes in distribution of each movie genre can be viewed in detail in the next part.

4) Distribution of movies by genres in six latest year groups:

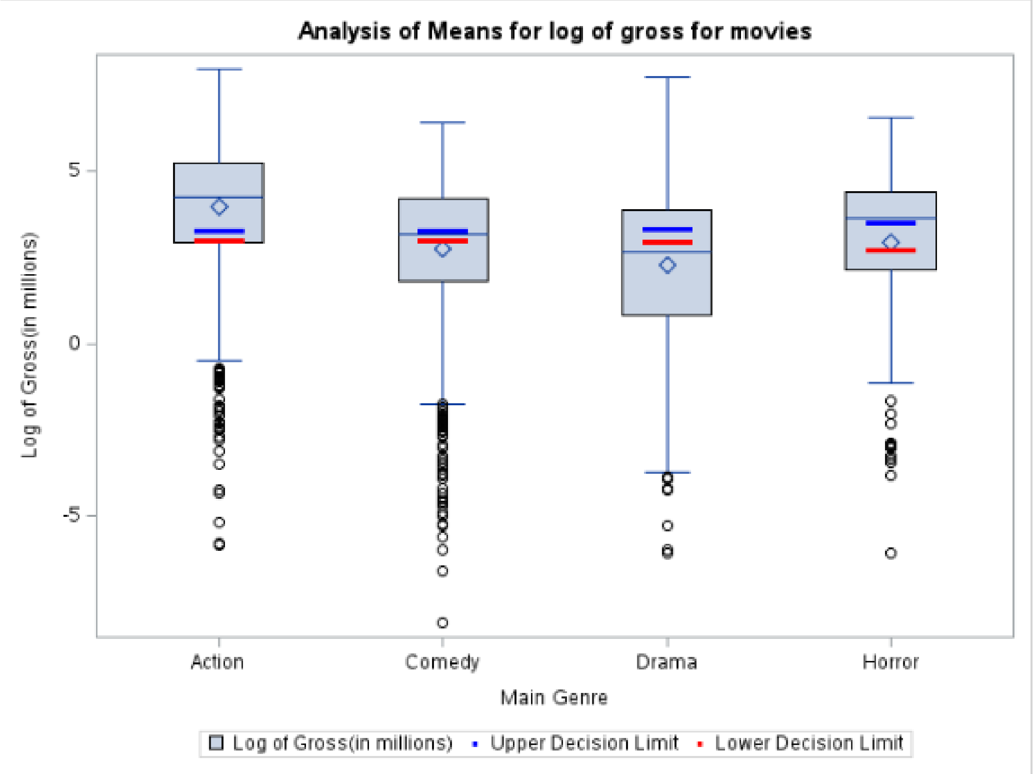
- Method: From the data from the table “full”, pie charts corresponding to each time period are created to present the distribution of movies by number.
- Result:



- Comment: The “Action” genre from 1990 to 1995 is the most prominent genre in number, accounting for approximately one-third of movies made in that year. However, as time goes on, the genre loses its dominating position, with only 18% of all movies made being Action movies by 2015-2020. “Comedy” and “Drama” movies have not changed much in the distribution in the 30 years. “Horror” has had quite a considerable rise in this 30 year period of time. Overall, the distribution of movie genres has grown to be more diversified as the other genres take up more and more percentage of movies made. This can be considered to be the result of movies becoming more and more popular as a media, which results in its attraction toward people with different tastes. Furthermore, with the advance of technology, the entry barrier into the film industry becomes lower, resulting in people being able to make movies on their own accord, without much investment.

5) Analysis of means for gross of movies by genres:

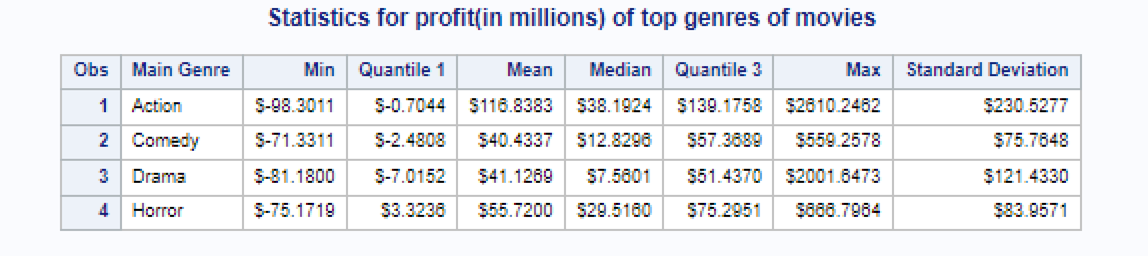
- Method: The data is first sorted by genre. Then, analysis of means is performed on all of the movie genres, comparing the means of gross for each of them with the overall mean at the significance level of 0.05. From this analysis, it can be determined which genre has gross significantly different from the overall mean. The log of gross is used as the gross value of all movies are highly skewed to the right, causing the graph to be flatten and hard to view. As the "anom" procedure does not support selected display of the genres, the upper decision limit, lower decision limit and other values are saved into a table and then merged with the "full" table after being sorted. The gross is then displayed as a boxplot, with the decision limits displayed as a line on each of the genres as a series of scatter plots.
- Result:



- Comment: The "Action" genre has significantly more gross than overall means, followed by "Horror". Meanwhile, "Comedy" and "Drama" perform approximately the same or even lower than the overall mean. Moreover, as the unit of y axis is log of gross, the difference between the gross of "Action" and the overall means is even more significant.

6) Statistics regarding profit made from movies from top genres:

- Method: Important statistics regarding the profit generated by each top genre of movie are created.
- Result:



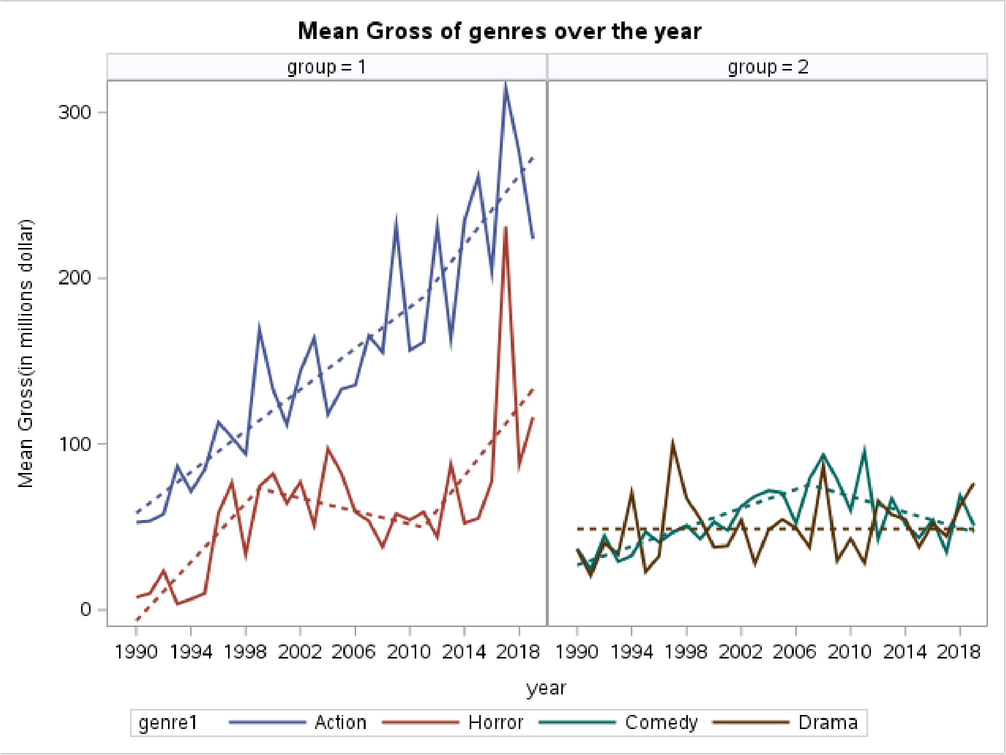
Statistics for profit(in millions) of top genres of movies

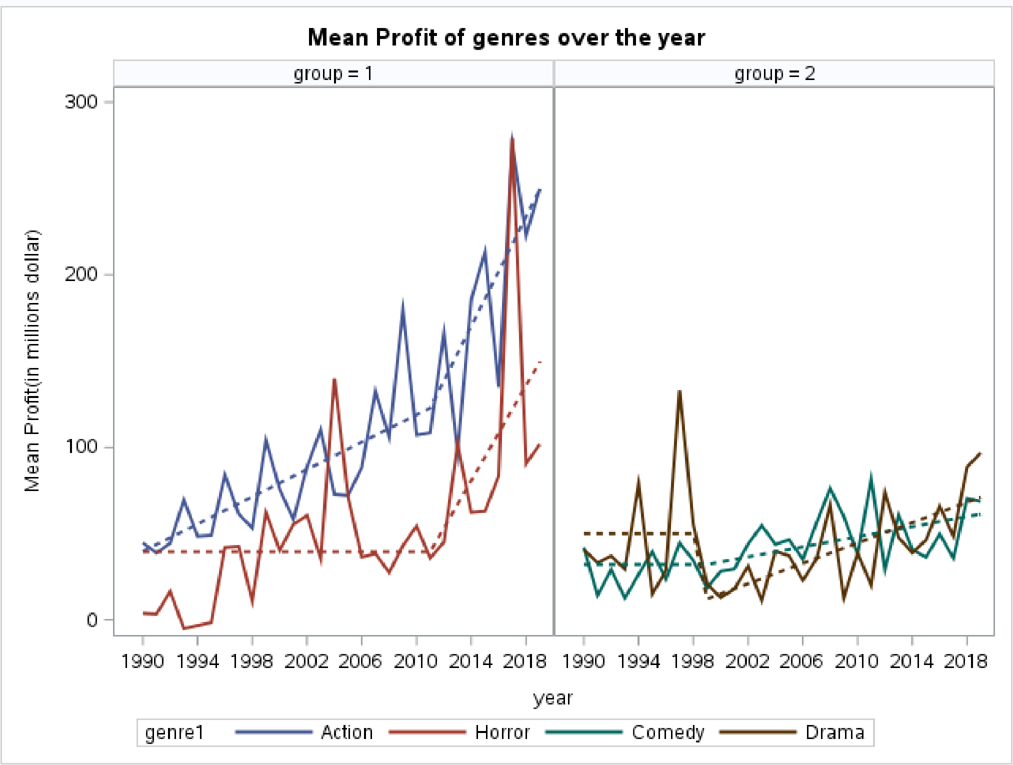
Obs	Main Genre	Min	Quantile 1	Mean	Median	Quantile 3	Max	Standard Deviation
1	Action	$-98.3011	$-0.7044	$116.8383	$38.1924	$139.1758	$2610.2462	$230.5277
2	Comedy	$-71.3311	$-2.4808	$40.4337	$12.8296	$57.3689	$559.2578	$75.7648
3	Drama	$-81.1800	$-7.0152	$41.1269	$7.5601	$51.4370	$2001.6473	$121.4330
4	Horror	$-75.1719	$3.3236	$55.7200	$29.5160	$75.2951	$666.7964	$83.9571

- Comment: From the table, it can be inferred that in terms of profit, “Horror” is the safest genre as the interquartile range of profit is entirely positive and has the third smallest standard deviation of the four top genres. However, for making a good profit, the “Action” genre has more than double the mean of profit from the “Horror” genre. Moreover, the interquartile range covers higher profit and the first quartile is only slightly smaller than that of “Horror”. “Drama” has approximately the same distribution as “Action”, however, its interquartile range is significantly more leaned towards the negative. “Comedy” does not have much potential in making a profit when compared to the other three.

7) Development of gross and profit for each genres over the year:

- Method: The mean gross and profit for each year of each genre was calculated. Then, the *adaptivereg* procedure was used to generate predicted values for each year based on the means, which is later drawn onto the graph as a trendline. The genres are split into 2 groups, one with considerable gross and one without, based on previous analysis.
- Result:

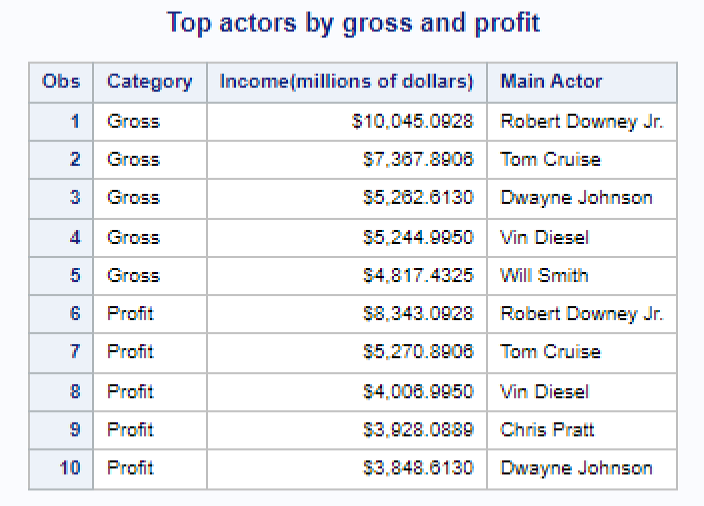




- Comment: For the first group, “Action” has been rising continuously since 1990, and the pace at which it is advancing continues to go up. This can be attributed to public interest that the genre has managed to develop since the 1990s, given its overall percentage of over 30% of all movies made in 1990-1995. “Horror” has had a significant leap in gross and profit, given its low count of movies at the start of this time period. However, despite its rising trend, the genre will have difficulty attaining the same number that “Action” has due to lacking the massive legacy that “Action” has. For “Comedy” and “Drama”, they have unstable development as the trend of both seems to be fluctuating at a very low value when compared to that of the first group. In conclusion, if the purpose of making the film is to gain exposure, choosing “Action” as the main genre is advisable, and will be considered as such moving forward.

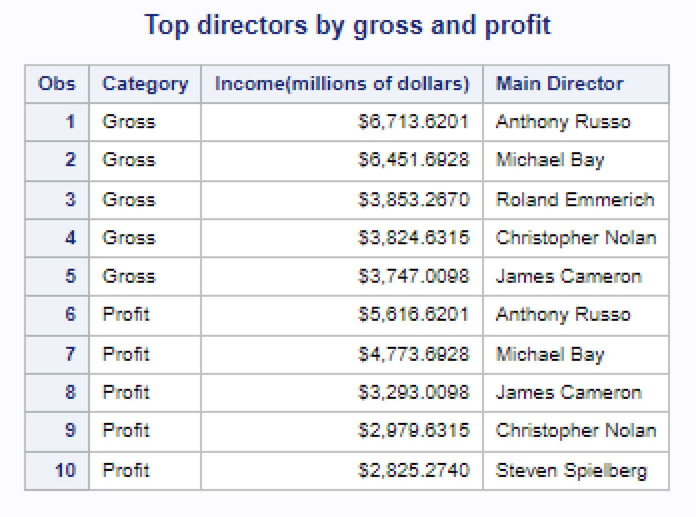
8) Top actors/actresses and directors in the Action genre

- Method: For all of the actors/actresses and directors in the “Action” genre, the sum of gross and profit of movies starring them or directed by them were calculated and then ranked. The reason why sum was used is to measure the overall popularity of the person. Furthermore, a list of movies that have both the top actors/actresses and directors was also generated as a point of reference for consideration for collaboration in the future. A correlation matrix is generated to test whether having a famous actor/actress or director helps to improve the gross of movies.
- Result:



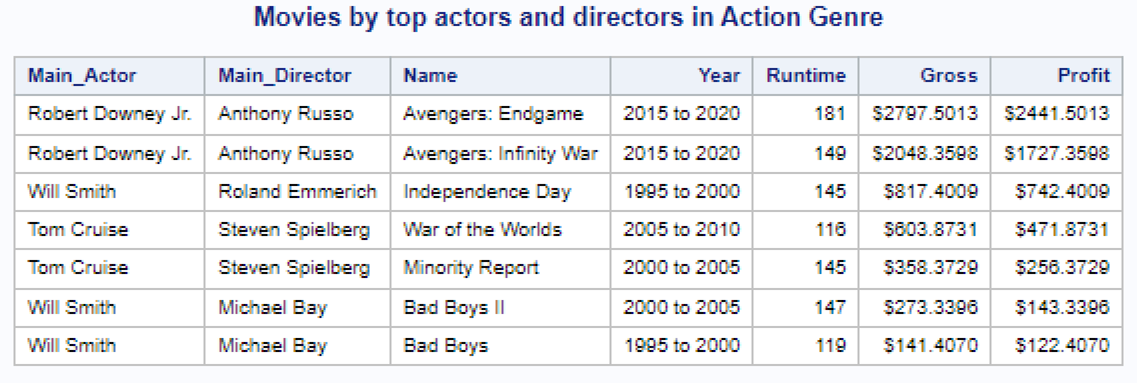
Top actors by gross and profit

Obs	Category	Income(millions of dollars)	Main Actor
1	Gross	$10,045.0928	Robert Downey Jr.
2	Gross	$7,367.8906	Tom Cruise
3	Gross	$5,262.6130	Dwayne Johnson
4	Gross	$5,244.9950	Vin Diesel
5	Gross	$4,817.4325	Will Smith
6	Profit	$8,343.0928	Robert Downey Jr.
7	Profit	$5,270.8906	Tom Cruise
8	Profit	$4,006.9950	Vin Diesel
9	Profit	$3,928.0889	Chris Pratt
10	Profit	$3,848.6130	Dwayne Johnson



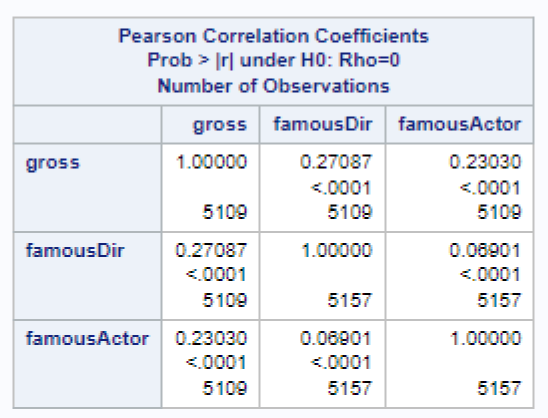
Top directors by gross and profit

Obs	Category	Income(millions of dollars)	Main Director
1	Gross	$6,713.6201	Anthony Russo
2	Gross	$6,451.6928	Michael Bay
3	Gross	$3,853.2670	Roland Emmerich
4	Gross	$3,824.6315	Christopher Nolan
5	Gross	$3,747.0098	James Cameron
6	Profit	$5,616.6201	Anthony Russo
7	Profit	$4,773.6928	Michael Bay
8	Profit	$3,293.0098	James Cameron
9	Profit	$2,979.6315	Christopher Nolan
10	Profit	$2,825.2740	Steven Spielberg



Movies by top actors and directors in Action Genre

Main_Actor	Main_Director	Name	Year	Runtime	Gross	Profit
Robert Downey Jr.	Anthony Russo	Avengers: Endgame	2015 to 2020	181	$2797.5013	$2441.5013
Robert Downey Jr.	Anthony Russo	Avengers: Infinity War	2015 to 2020	149	$2048.3598	$1727.3598
Will Smith	Roland Emmerich	Independence Day	1995 to 2000	145	$817.4009	$742.4009
Tom Cruise	Steven Spielberg	War of the Worlds	2005 to 2010	116	$603.8731	$471.8731
Tom Cruise	Steven Spielberg	Minority Report	2000 to 2005	145	$358.3729	$256.3729
Will Smith	Michael Bay	Bad Boys II	2000 to 2005	147	$273.3396	$143.3396
Will Smith	Michael Bay	Bad Boys	1995 to 2000	119	$141.4070	$122.4070



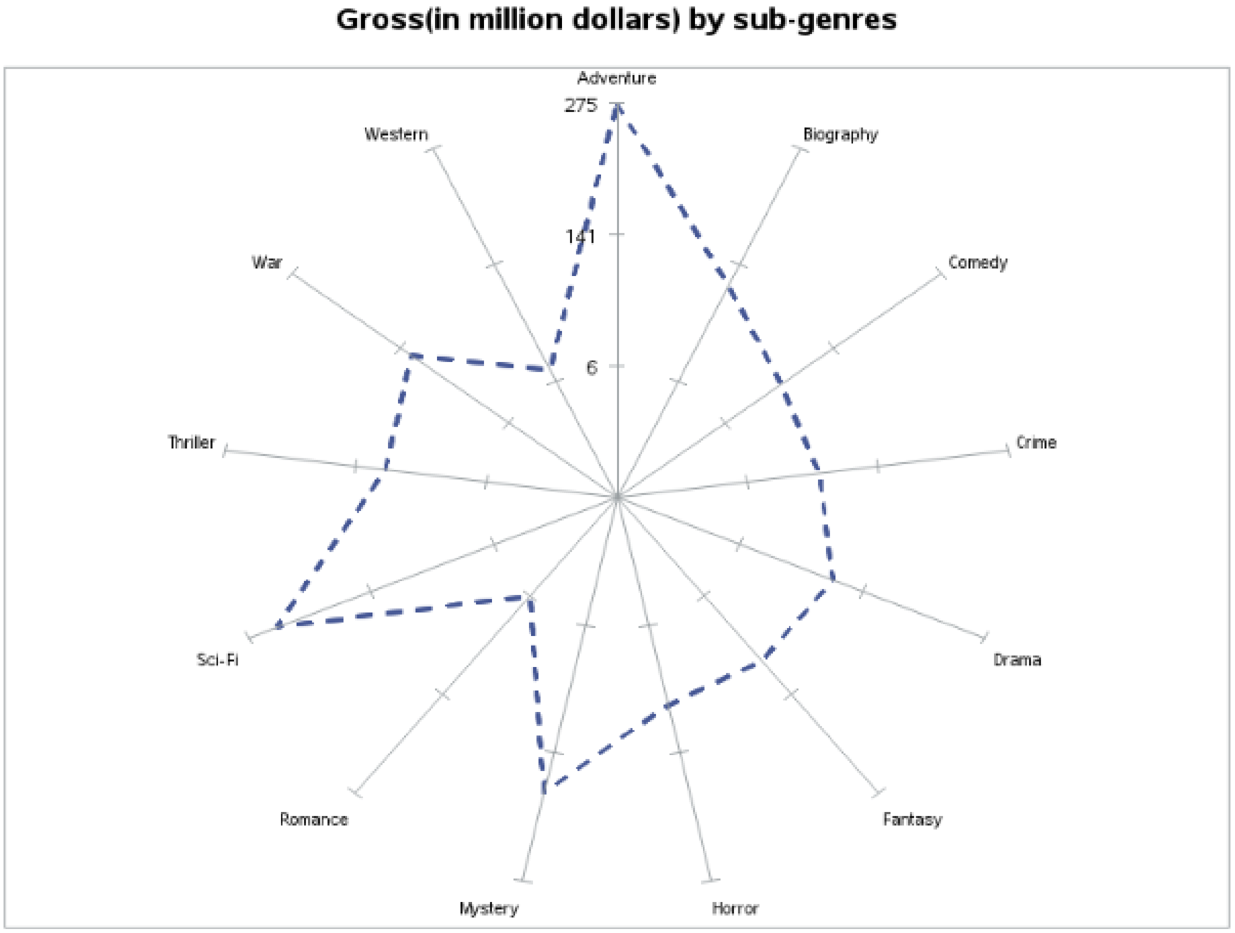
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	gross	famousDir	famousActor
gross	1.00000 5109	0.27087 <.0001 5109	0.23030 <.0001 5109
famousDir	0.27087 <.0001 5109	1.00000 5157	0.06901 <.0001 5157
famousActor	0.23030 <.0001 5109	0.06901 <.0001 5157	1.00000 5157

- Comment: There is a high correlation between gross and profit as 4 of the top actors/actresses and directors are in both categories. Furthermore, when considering the movies with participation from top personnel from both sides, the gross and profit range is from slightly over the average to overwhelmingly above it. However, as both Avengers titles are built off of an enormous franchise, its success is quite misleading, as much time and effort are spent to build the franchise from the ground up. The other movies have a more reasonable range of gross and profit from the 2 top titles, which still overperforms when compared to other titles of the same genre. Furthermore, having a famous actor/actress or director in a movie has a positive effect on the gross.

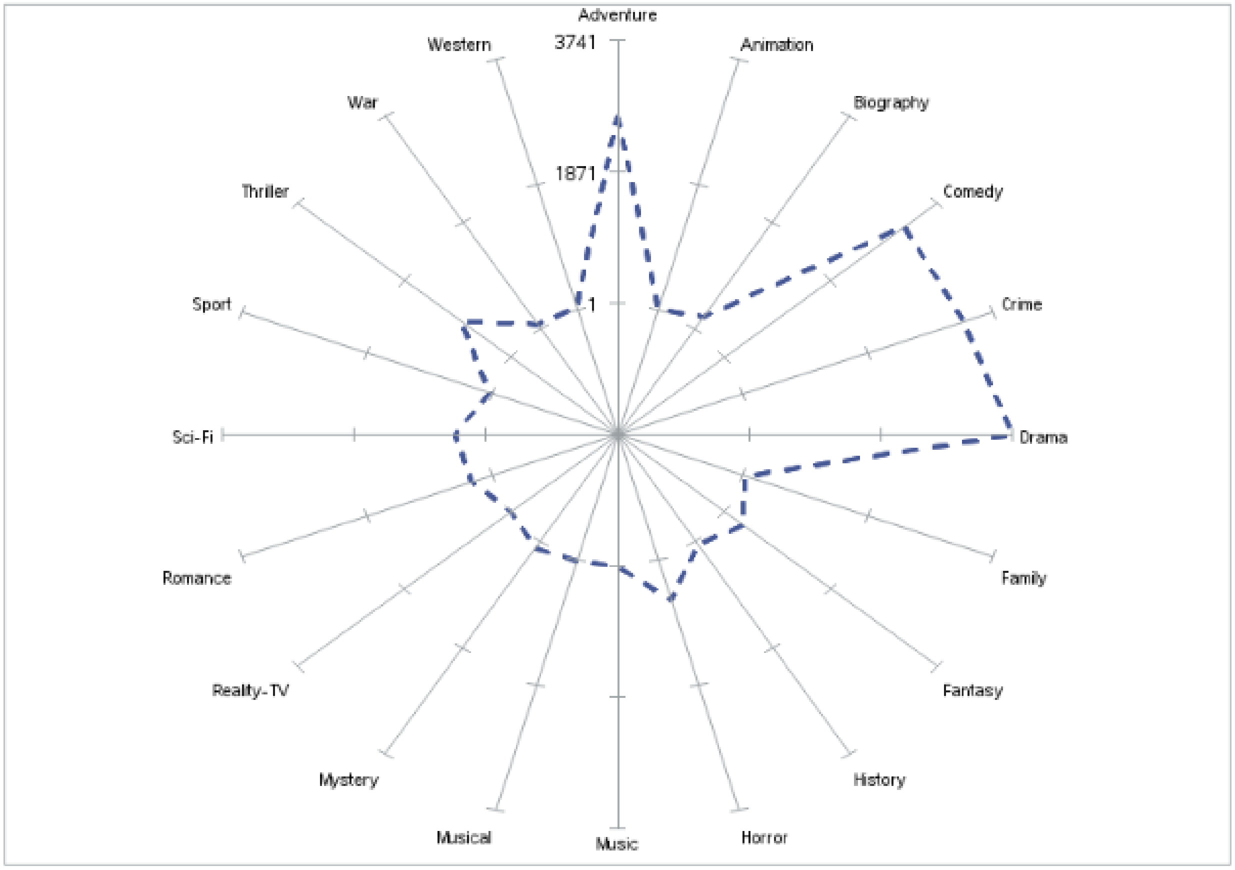
9) Analysis of sub-genres of Action:

- Method: In order to make a promising movie, the sub-genres also need to be looked into. The mean of the gross an action movie can bring is calculated and categorized into groups based on its sub-genres. Then, the calculated value is displayed into a radar chart. The number of movies in each sub-genres of action movies is also displayed as a radar chart in order to see the full picture.
- Result:



Gross(in million dollars) by sub-genres

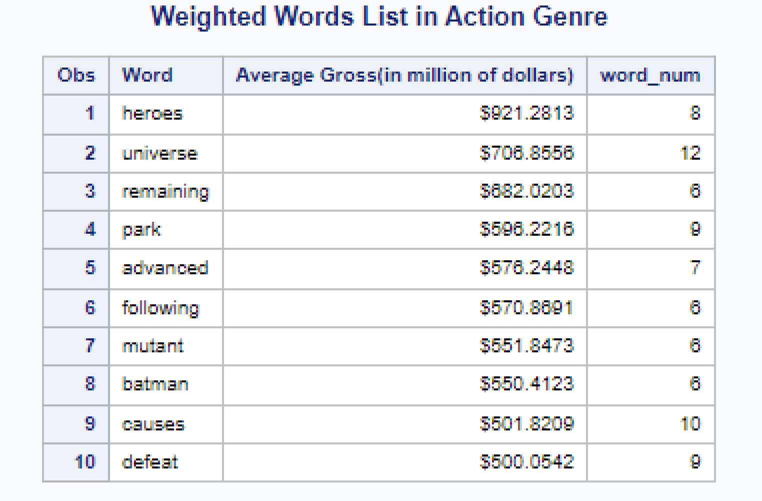
Number of Movies by sub-genres



- Comment: The *Adventure* sub-genre accounts for a significant portion of action movies, and it also seems to perform considerably well, which can signify that competition is quite fierce. For the *Comedy, Crime* and *Drama* sub-genre, despite the overwhelmingly large number of movies made, its average gross is quite underwhelming, signifying a high-risk-low-reward situation. Two of the best categories of sub-genre to work on are *Mystery* and *Sci-Fi,* given their low movie counts, low competition and high average gross that a movie can bring in.

10) List of keywords ranked based on gross of the movies:

- Method: For every word in the *Description* portion of a movie, they are extracted and recorded into a table with the word and the corresponding gross of the movies that they are used in. There is a threshold set to reduce the effect outliers have on the effect each word may have on the gross. The words are then counted and calculated the average gross that they can bring. In order to avoid any franchise or movie-specific words, there is also another threshold for the number of times the word appears. The top 10 most impactful words are chosen and displayed into a table.
- Result:



Weighted Words List in Action Genre

Obs	Word	Average Gross(in million of dollars)	word_num
1	heroes	$921.2813	8
2	universe	$706.8556	12
3	remaining	$682.0203	6
4	park	$596.2216	9
5	advanced	$576.2448	7
6	following	$570.8691	6
7	mutant	$551.8473	6
8	batman	$550.4123	6
9	causes	$501.8209	10
10	defeat	$500.0542	9

- Comment: From the words in the table, there are interesting subjects that can be used in the new movie such as “heroes”, “universe”, “mutant”,... These can be used as a guide in making the new film.

Conclusion

In conclusion, the film industry has been developing since 1990 and has become quite a profitable field. As in the context of this analysis, in order to create a movie that appeals to the general public and gain exposure for a movie studio, *Action* was advised to be chosen as the main genre of the film. This is thanks to its positive performance on both gross and profit and also its popularity, represented by the number of movies made. *Action* overperformed its runner-up, *Horror*, by a large margin. When considering personnel to work on the movies, having famous actors/actresses or directors can positively influence the gross of a movie. A list of famous actors/actresses and directors was also provided. Furthermore, when considering sub-genres of movie, *Adventure* seems to compliment the *Action* best, given its overwhelmingly high gross value. However, there is also a lot of competition for this combination. Other promising sub-genres are *Mystery* and *Sci-Fi*, thanks to its high gross and low competition, signified by low number of movies made. Lastly, keywords, weighted by the gross of movies they are used in as a part of the description, were also provided to suggest some themes that could be of use to the studio.

Reflection

From this project, I have learnt how to create a narrative through my analyses in order to create a coherent report. Furthermore, I have been able to utilize most of the commands and skills that I have learnt in this course to clarify my findings and make my point. However, I found that most of my analyses are quite surface-level and have not been able to dig deeper into the statistics side. Therefore, if you have any feedback regarding how I can improve my analysis, I will be glad to learn more from it.

Citation

[1] Importing multi-line value in CSV file:

- Google Search Query: multiline csv import sas
- Link: https://communities.sas.com/t5/SAS-Programming/import-multiple-line-value-of-csv-file/m-p/272790
- From: Ballardw

[2] Performing analysis of means:

- Google Search Query: proc anom sas
- Link: https://support.sas.com/documentation/onlinedoc/qc/142/anom.pdf

[3] Creating custom reference line for each column of boxplot:

- Google Search Query: proc sgplot vbox reference line each column
- Link: https://communities.sas.com/t5/Graphics-Programming/proc-sgplot-vbox-reference-lines/td-p/414921
- From: WCW