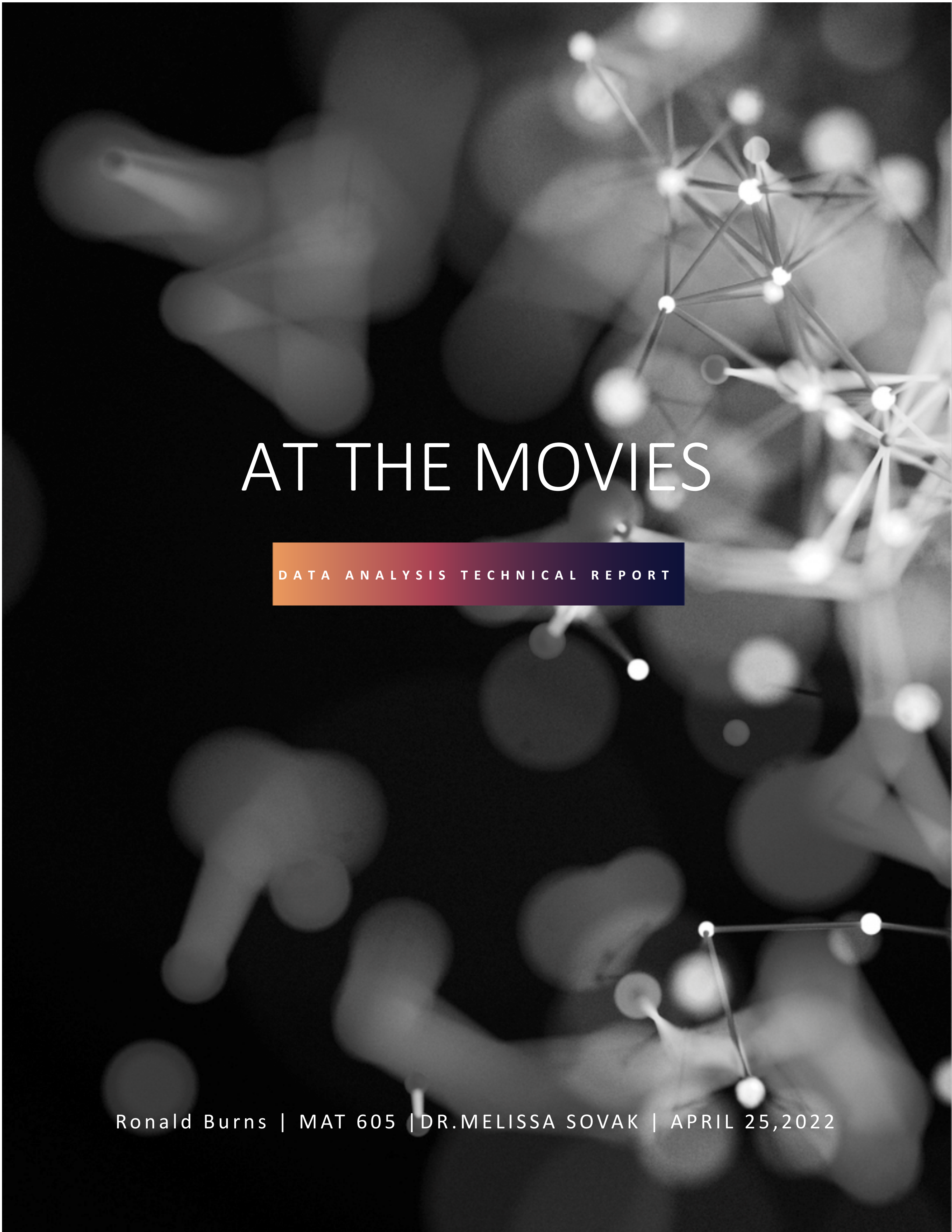
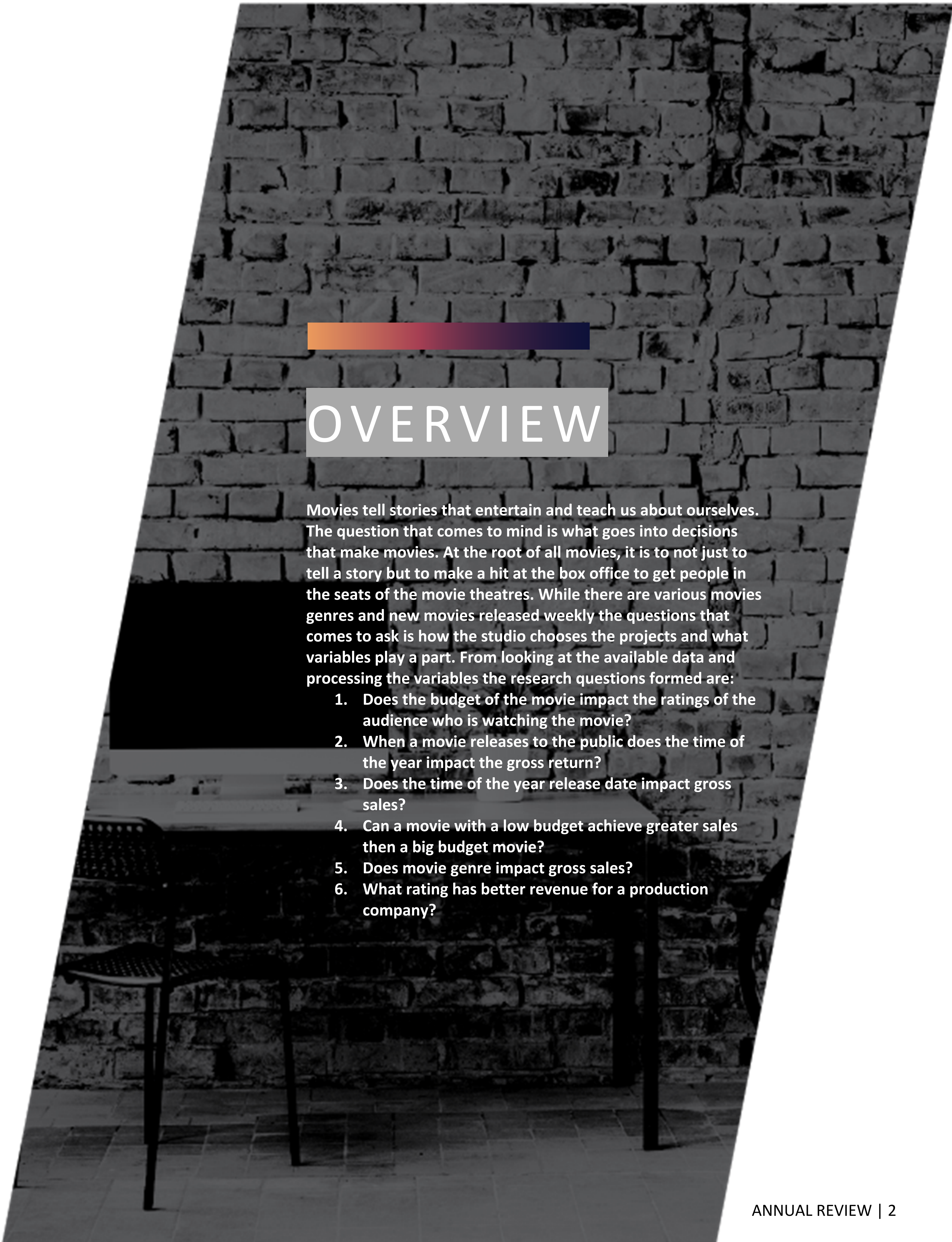


AT THE MOVIES

DATA ANALYSIS TECHNICAL REPORT

Ronald Burns | MAT 605 |DR.MELISSA SOVAK | APRIL 25,2022



OVERVIEW

Movies tell stories that entertain and teach us about ourselves. The question that comes to mind is what goes into decisions that make movies. At the root of all movies, it is to not just to tell a story but to make a hit at the box office to get people in the seats of the movie theatres. While there are various movies genres and new movies released weekly the questions that comes to ask is how the studio chooses the projects and what variables play a part. From looking at the available data and processing the variables the research questions formed are:

1. Does the budget of the movie impact the ratings of the audience who is watching the movie?
2. When a movie releases to the public does the time of the year impact the gross return?
3. Does the time of the year release date impact gross sales?
4. Can a movie with a low budget achieve greater sales then a big budget movie?
5. Does movie genre impact gross sales?
6. What rating has better revenue for a production company?

INFORMATION

WHERE THE DATA COME FROM

The data was obtained from the website known as Data World. The collector of the data is named James Gaskin who compiled a list of movies from 1989 to 2001. The movies consist of various genres and ratings with information that will help answer the questions that are being asked.
https://data.world/jamesgaskin/movies/workspace/file?filename=view

Other data was pursued from IMDB or other data sources but did not have the complete information necessary to address the questions.

VARIABLES

MPAA Rating
Data Type: Character
Meaning:
G: Kid Friendly
PG: Parental Guidance
PG-13: 13 years of age and older
R: Over 17/18

Budget
Data Type: Long
Meaning: Amount it cost to create the film

Release Date
Data Type: Date
Meaning: The date that the movie was released into movie theatres

Genre
Data Type: Character
Meaning: The type of movie it is. Romance, Comedy, Horror, etc.

Gross
Data Type: Long
Meaning: Amount of money the movie made in the box office

HOW THE DATA WAS CLEANED

The data was prepared and cleaned through SAS

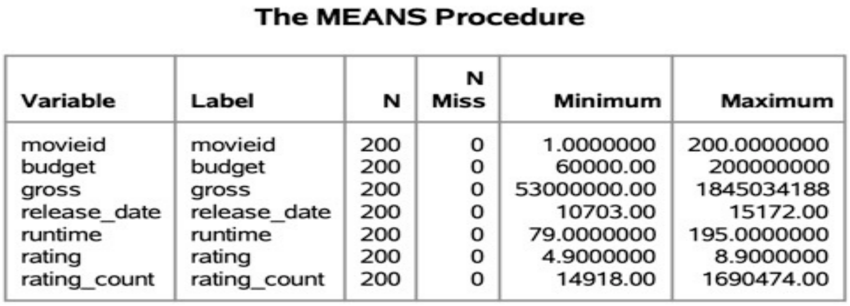
The first coding used was:

```
proc freq data = work.import;
tables title mpaa_rating budget gross release_date genre runtime rating rating_count/ nocum nopercent;
run;
```

This allowed to make sure there were not duplicate name titles and how many movies were under each genre.

To verify there were no missing data the following code was used.

```
proc means data = work.import n nmiss min max;
run;
```



The MEANS Procedure

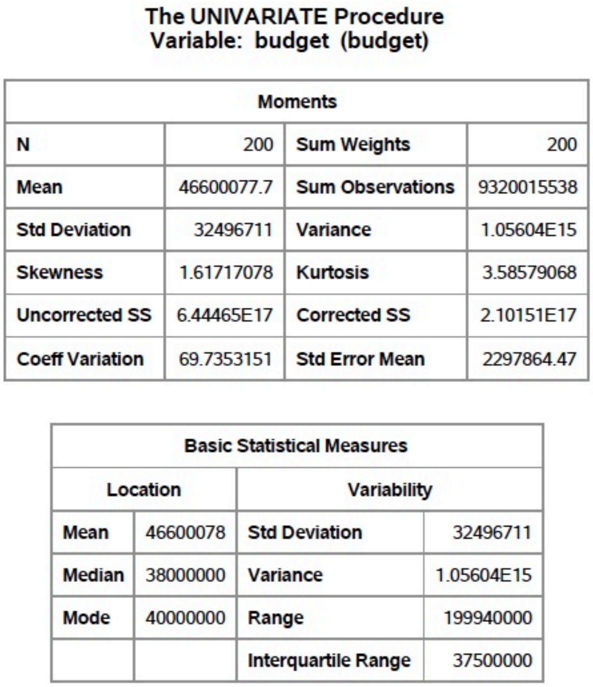
Variable	Label	N	N Miss	Minimum	Maximum
movieid	movieid	200	0	1.0000000	200.0000000
budget	budget	200	0	60000.00	200000000
gross	gross	200	0	53000000.00	1845034188
release_date	release_date	200	0	10703.00	15172.00
runtime	runtime	200	0	79.0000000	195.0000000
rating	rating	200	0	4.9000000	8.9000000
rating_count	rating_count	200	0	14918.00	1690474.00

As the table shows there were not missing values to cause an issue with the analysis.

Since the data shows to have no missing values, the next process was to test if the information in each variable was equally spread.

The code used used:

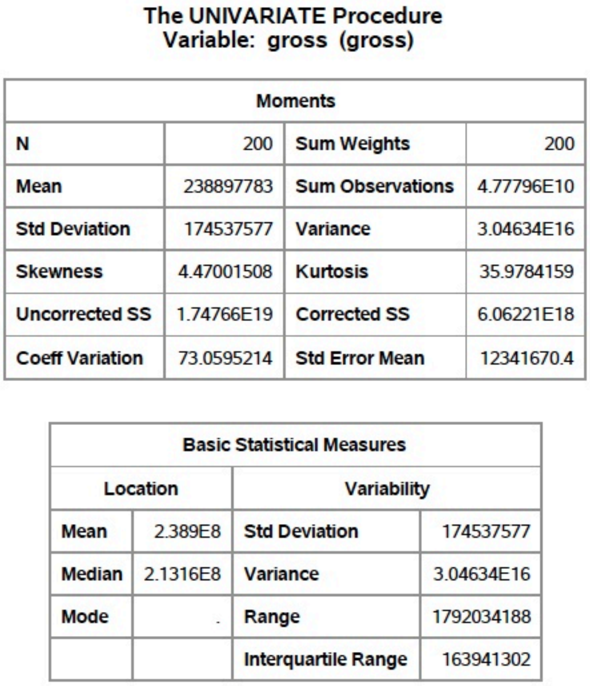
```
proc univariate data=work.import;
run;
```



The UNIVARIATE Procedure
Variable: budget (budget)

Moments			
N	200	Sum Weights	200
Mean	46600077.7	Sum Observations	9320015538
Std Deviation	32496711	Variance	1.05604E15
Skewness	1.61717078	Kurtosis	3.58579068
Uncorrected SS	6.44465E17	Corrected SS	2.10151E17
Coeff Variation	69.7353151	Std Error Mean	2297864.47

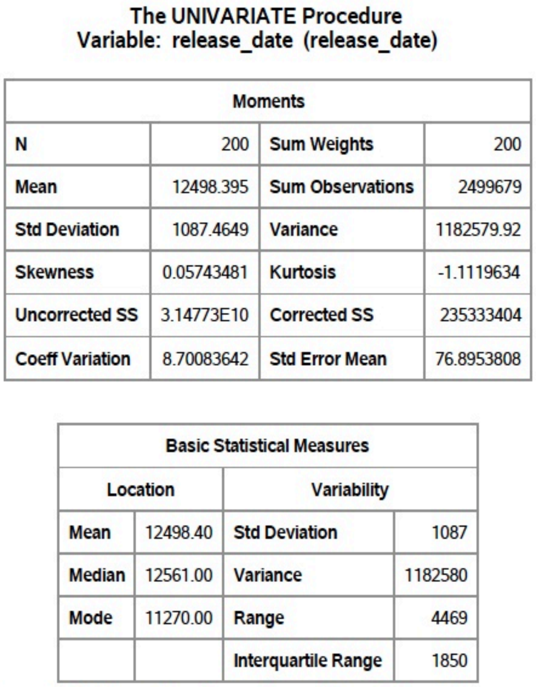
Basic Statistical Measures			
Location		Variability	
Mean	46600078	Std Deviation	32496711
Median	38000000	Variance	1.05604E15
Mode	40000000	Range	199940000
		Interquartile Range	37500000



The UNIVARIATE Procedure
Variable: gross (gross)

Moments			
N	200	Sum Weights	200
Mean	238897783	Sum Observations	4.77796E10
Std Deviation	174537577	Variance	3.04634E16
Skewness	4.47001508	Kurtosis	35.9784159
Uncorrected SS	1.74766E19	Corrected SS	6.06221E18
Coeff Variation	73.0595214	Std Error Mean	12341670.4

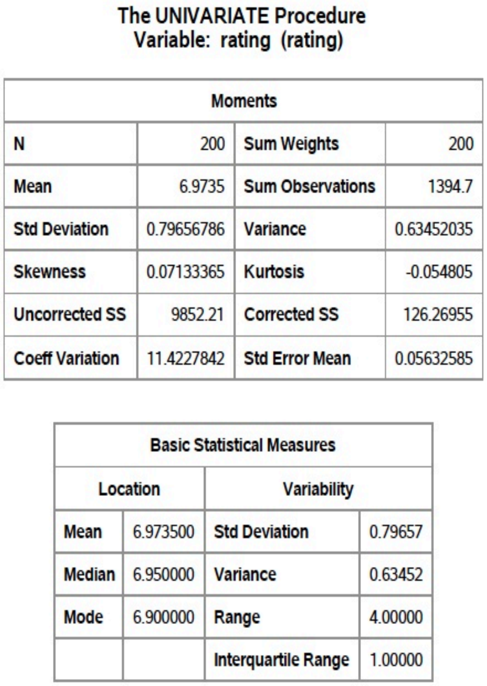
Basic Statistical Measures			
Location		Variability	
Mean	2.389E8	Std Deviation	174537577
Median	2.1316E8	Variance	3.04634E16
Mode	.	Range	1792034188
		Interquartile Range	163941302



The UNIVARIATE Procedure
Variable: release_date (release_date)

Moments			
N	200	Sum Weights	200
Mean	12498.395	Sum Observations	2499679
Std Deviation	1087.4649	Variance	1182579.92
Skewness	0.05743481	Kurtosis	-1.1119634
Uncorrected SS	3.14773E10	Corrected SS	235333404
Coeff Variation	8.70083642	Std Error Mean	76.8953808

Basic Statistical Measures			
Location		Variability	
Mean	12498.40	Std Deviation	1087
Median	12561.00	Variance	1182580
Mode	11270.00	Range	4469
		Interquartile Range	1850



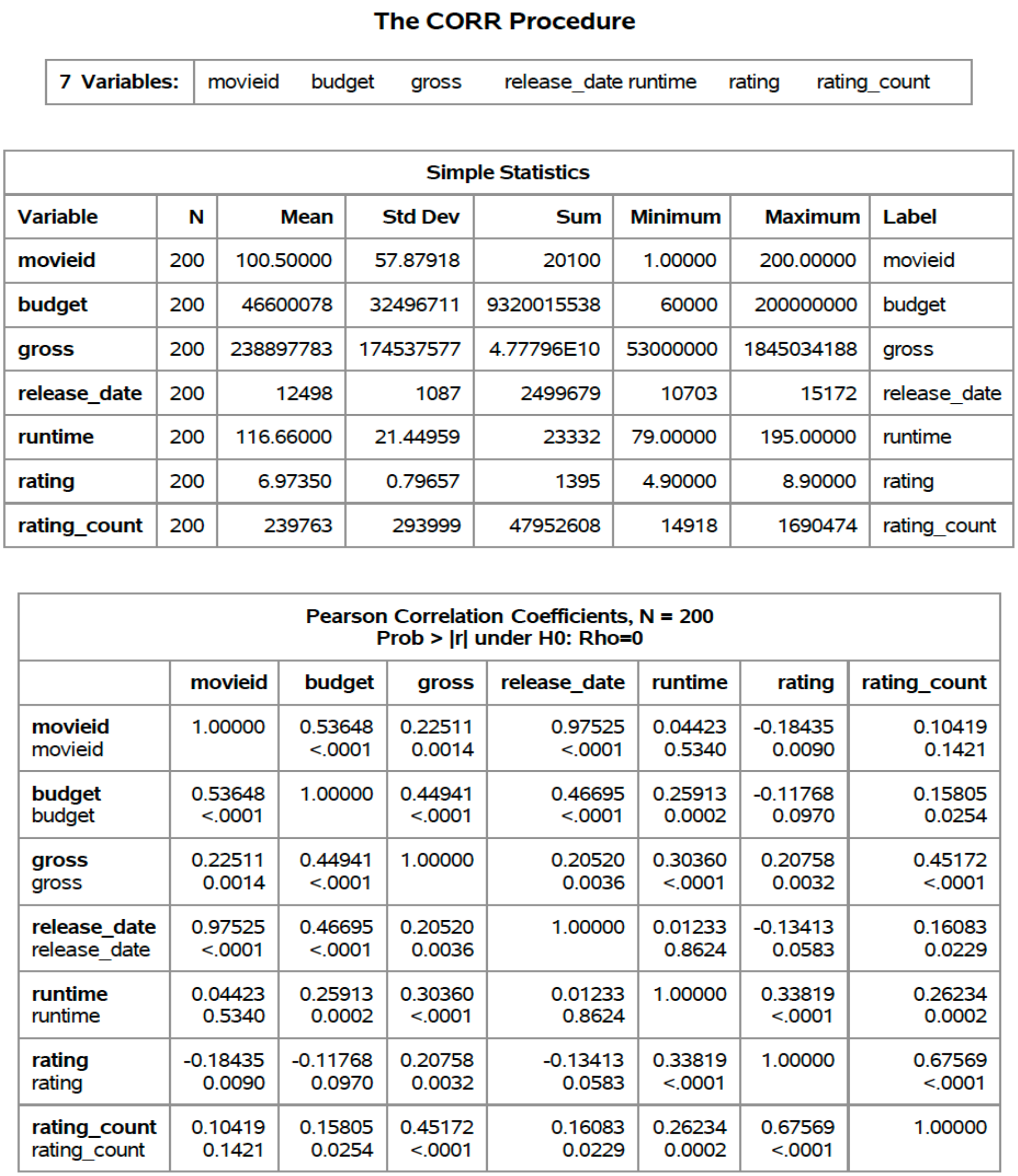
The UNIVARIATE Procedure
Variable: rating (rating)

Moments			
N	200	Sum Weights	200
Mean	6.9735	Sum Observations	1394.7
Std Deviation	0.79656786	Variance	0.63452035
Skewness	0.07133365	Kurtosis	-0.054805
Uncorrected SS	9852.21	Corrected SS	126.26955
Coeff Variation	11.4227842	Std Error Mean	0.05632585

Basic Statistical Measures			
Location		Variability	
Mean	6.973500	Std Deviation	0.79657
Median	6.950000	Variance	0.63452
Mode	6.900000	Range	4.00000
		Interquartile Range	1.00000

Based on the information returned gross provided information that there was no mode so there were not consistent values that occurred more often, which is a good tell that budget of a movie does not always repeat. As expected, the budget showed there is a consistency in the amount or close to the same value of other projects.

The next part of the analysis was to see if there is any correlation between the variables and if there was a need to investigate closer.
The code used was:
proc corr data=work.import;
run;



The CORR Procedure

7 Variables:	movieid	budget	gross	release_date	runtime	rating	rating_count

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
movieid	200	100.50000	57.87918	20100	1.00000	200.00000	movieid
budget	200	46600078	32496711	9320015538	60000	200000000	budget
gross	200	238897783	174537577	4.77796E10	53000000	1845034188	gross
release_date	200	12498	1087	2499679	10703	15172	release_date
runtime	200	116.66000	21.44959	23332	79.00000	195.00000	runtime
rating	200	6.97350	0.79657	1395	4.90000	8.90000	rating
rating_count	200	239763	293999	47952608	14918	1690474	rating_count

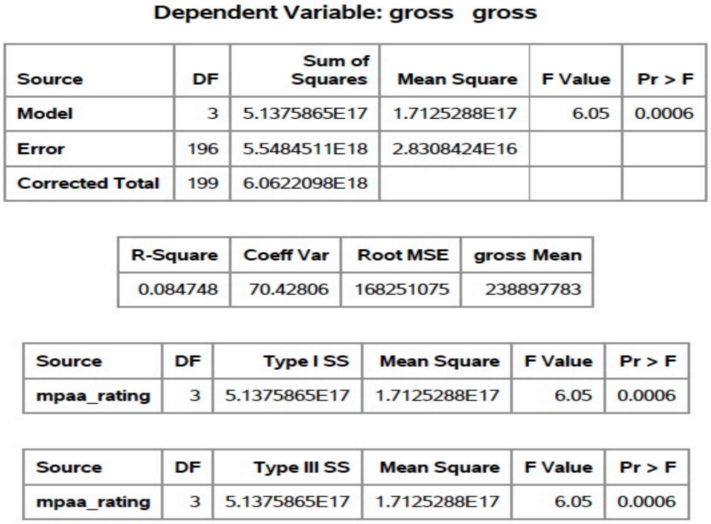
Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	movieid	budget	gross	release_date	runtime	rating	rating_count
movieid movieid	1.00000	0.53648 <.0001	0.22511 0.0014	0.97525 <.0001	0.04423 0.5340	-0.18435 0.0090	0.10419 0.1421
budget budget	0.53648 <.0001	1.00000	0.44941 <.0001	0.46695 <.0001	0.25913 0.0002	-0.11768 0.0970	0.15805 0.0254
gross gross	0.22511 0.0014	0.44941 <.0001	1.00000	0.20520 0.0036	0.30360 <.0001	0.20758 0.0032	0.45172 <.0001
release_date release_date	0.97525 <.0001	0.46695 <.0001	0.20520 0.0036	1.00000	0.01233 0.8624	-0.13413 0.0583	0.16083 0.0229
runtime runtime	0.04423 0.5340	0.25913 0.0002	0.30360 <.0001	0.01233 0.8624	1.00000	0.33819 <.0001	0.26234 0.0002
rating rating	-0.18435 0.0090	-0.11768 0.0970	0.20758 0.0032	-0.13413 0.0583	0.33819 <.0001	1.00000	0.67569 <.0001
rating_count rating_count	0.10419 0.1421	0.15805 0.0254	0.45172 <.0001	0.16083 0.0229	0.26234 0.0002	0.67569 <.0001	1.00000

Since the analysis was providing details associated with budget and gross one-way ANOVA was used to analyze the distribution of the Budget and the Gross based on the MPAA rating.

The code used was:

```
proc glm data=WORK.IMPORT;
        class mpaa_rating;
        model gross=mpaa_rating;
        means mpaa_rating / hovtest=levene welch plots=none;
        lsmeans mpaa_rating / adjust=tukey pdiff alpha=.05;
        run;
quit;
```



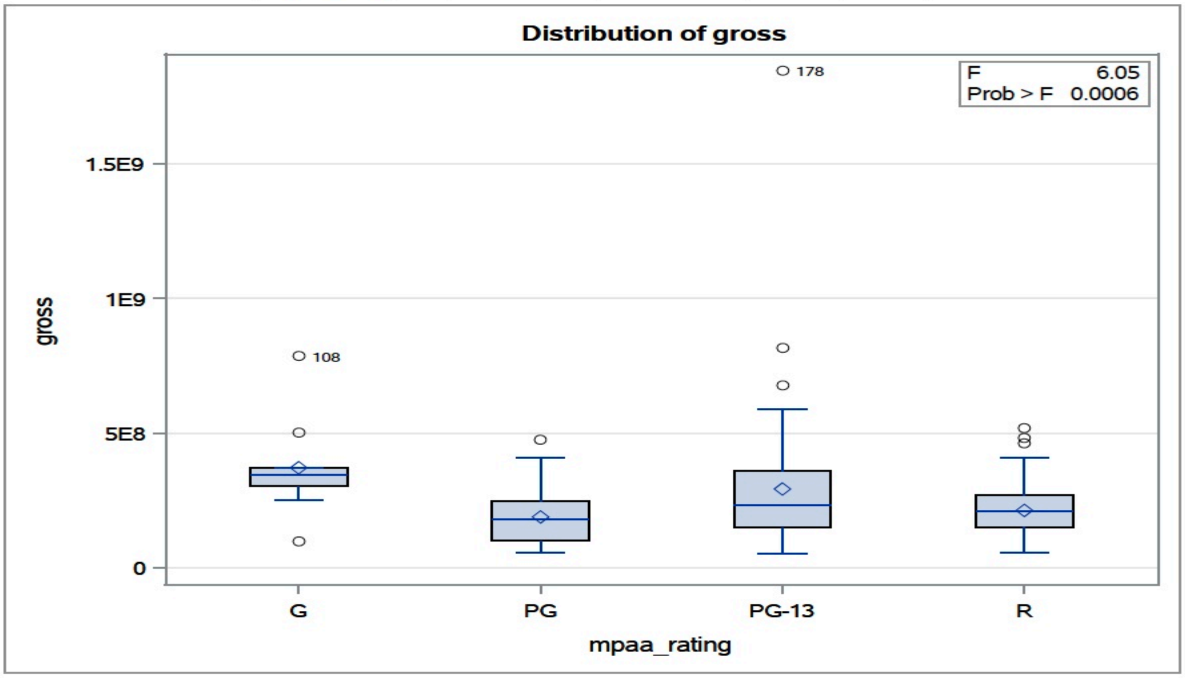
Dependent Variable: gross gross

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.1375865E17	1.7125288E17	6.05	0.0006
Error	196	5.5484511E18	2.8308424E16		
Corrected Total	199	6.0622098E18			

R-Square	Coeff Var	Root MSE	gross Mean
0.084748	70.42806	168251075	238897783

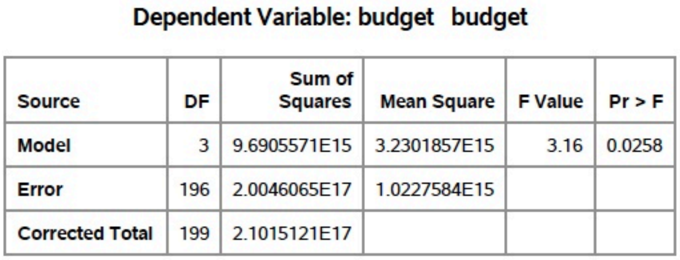
Source	DF	Type I SS	Mean Square	F Value	Pr > F
mpaa_rating	3	5.1375865E17	1.7125288E17	6.05	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mpaa_rating	3	5.1375865E17	1.7125288E17	6.05	0.0006



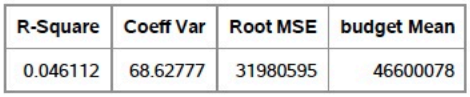
The code used was:

```
proc glm data=WORK.IMPORT;
        class mpaa_rating;
        model budget=mpaa_rating;
        means mpaa_rating / hovtest=levene welch plots=none;
        lsmeans mpaa_rating / adjust=tukey pdiff alpha=.05;
        run;
quit;
```

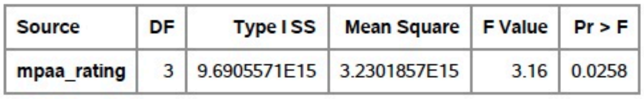


Dependent Variable: budget budget

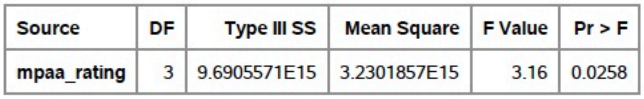
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.6905571E15	3.2301857E15	3.16	0.0258
Error	196	2.0046065E17	1.0227584E15		
Corrected Total	199	2.1015121E17			



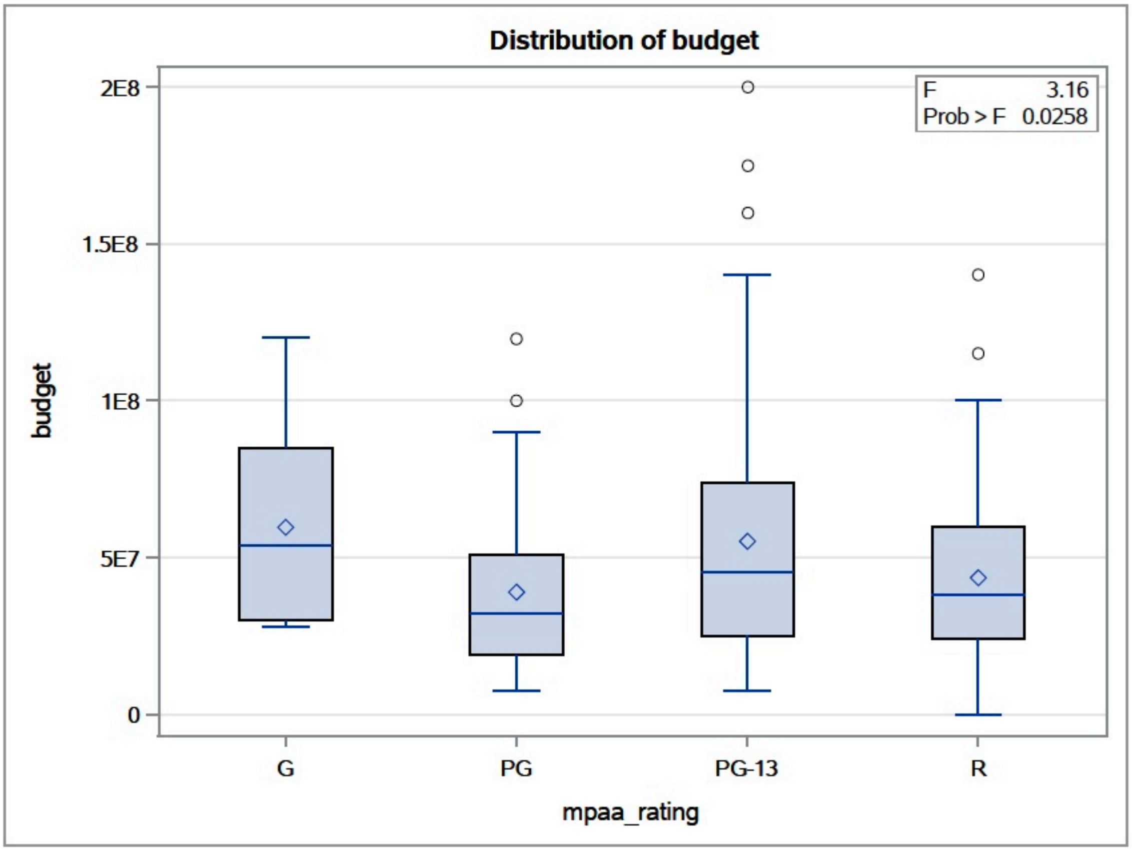
R-Square	Coeff Var	Root MSE	budget Mean
0.046112	68.62777	31980595	46600078



Source	DF	Type I SS	Mean Square	F Value	Pr > F
mpaa_rating	3	9.6905571E15	3.2301857E15	3.16	0.0258



Source	DF	Type III SS	Mean Square	F Value	Pr > F
mpaa_rating	3	9.6905571E15	3.2301857E15	3.16	0.0258

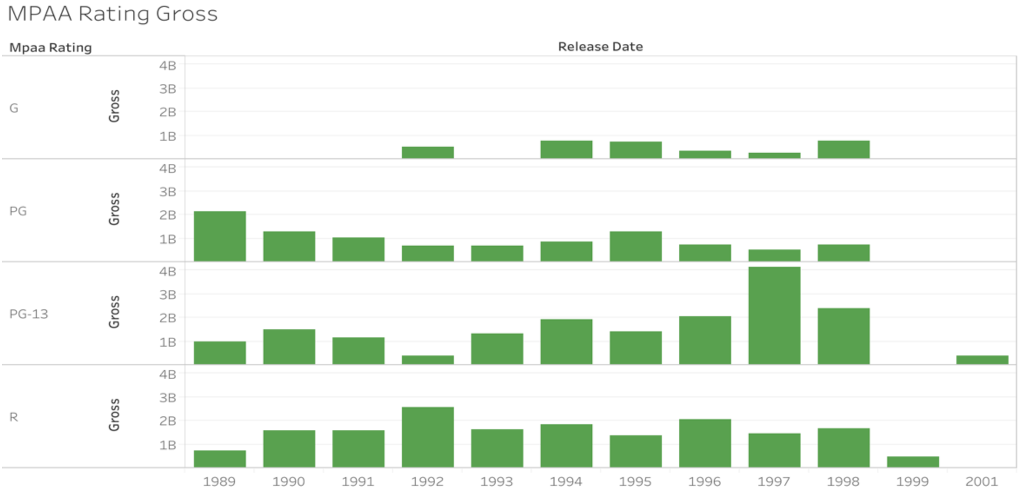


Based on the analysis provided in SAS there is a significant gain in PG-13 movies in gross and the budget which was starting this was not expected outcome over the 10 years. Though the popularity of PG-13 and R has grown there were some movies with MPAA rating that show outside the expected norms.

THE VISUALS

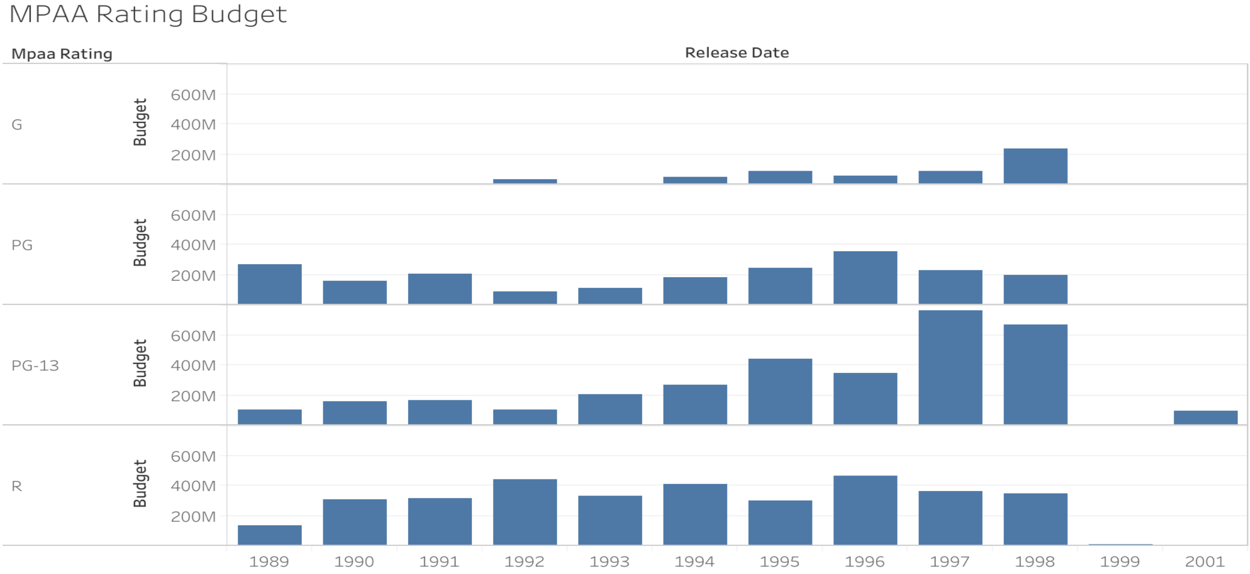
MPAA RATING GROSS VISUAL

This visual created in Tableau shows how much gross profits each movie rating in each year. This visual was chosen because it clearly shows changes that were occurring in the box office. As noted in the visual below PG13 and was on the rise in the box office showing a shift to more sensitive subjects.



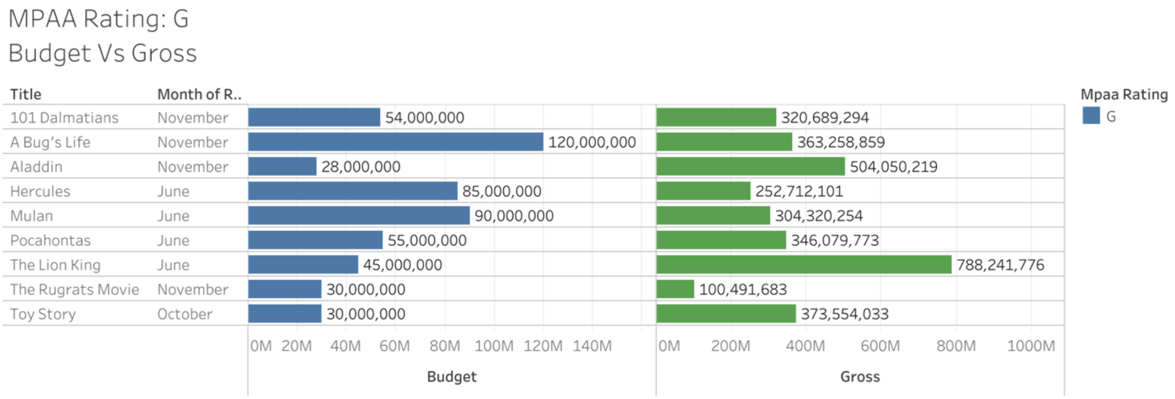
MPAA RATING BUDGET VISUAL

This visual was also created in Tableau and the budget each movie rating had in each year. This visual was chosen because it clearly shows PG-13 and R content was favored by the studios



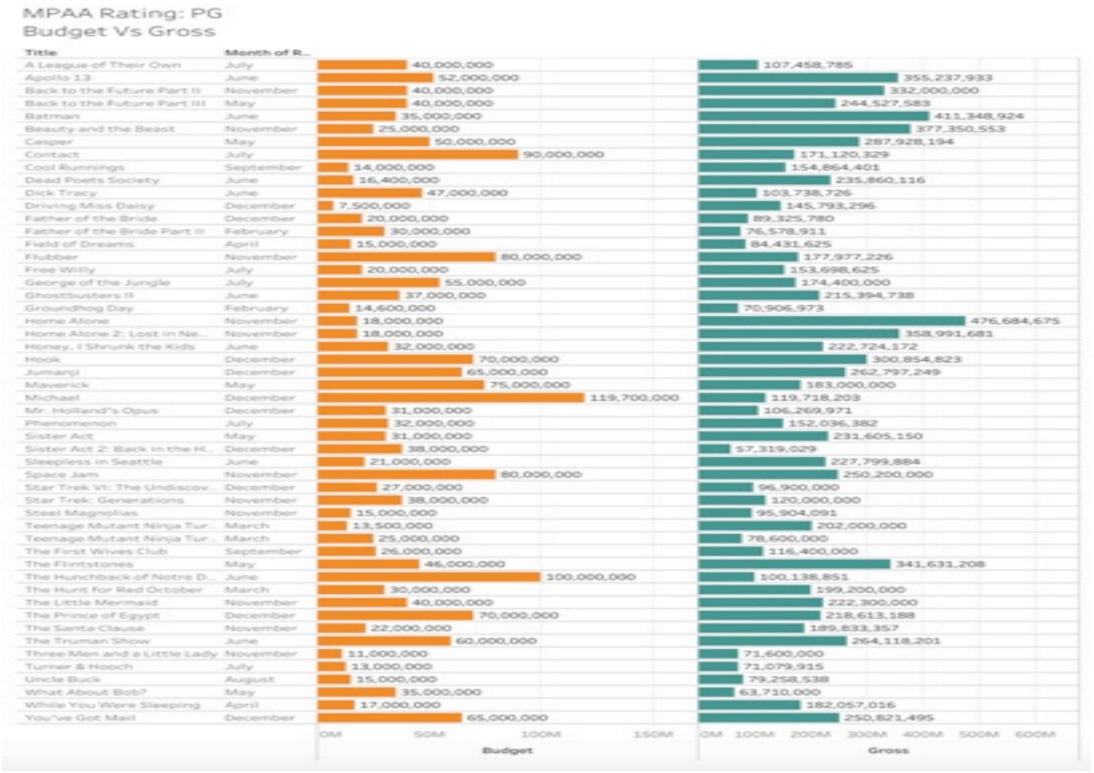
MPAA RATING "G" BUDGET VS GROSS VISUAL

Tableau was also used to show a breakdown in the budget versus the gross revenue made by a few of the movies under the G MPAA rating. As can be seen below the movies listed all had successful box office sales but as can be seen Lion king which was low budgeted made the largest revenue which answers one of the research questions of Budget VS gross.



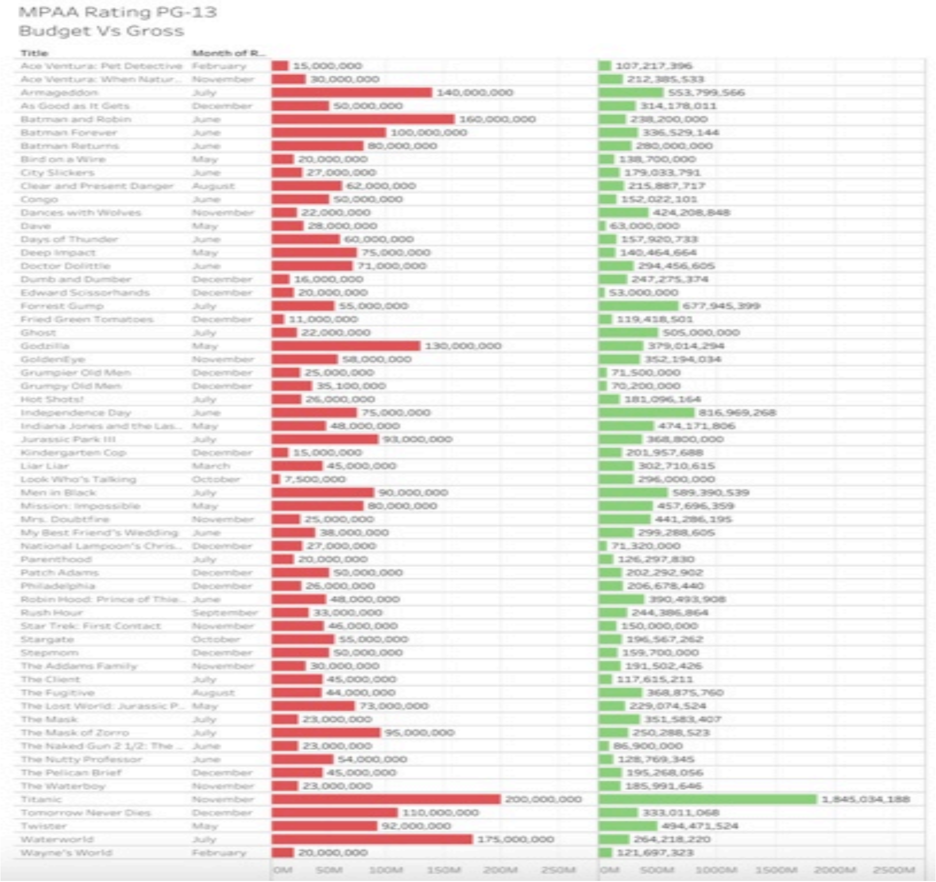
MPAA RATING "PG" BUDGET VS GROSS VISUAL

Tableau was also used to show a breakdown in the budget versus the gross revenue made by the movies under the PG MPAA rating. Another large grouping shows gross sales were more than the budget but movies near holidays have better sales. The largest revenue under PG was Home alone which only costed the production company 18 million and generated 476 million.



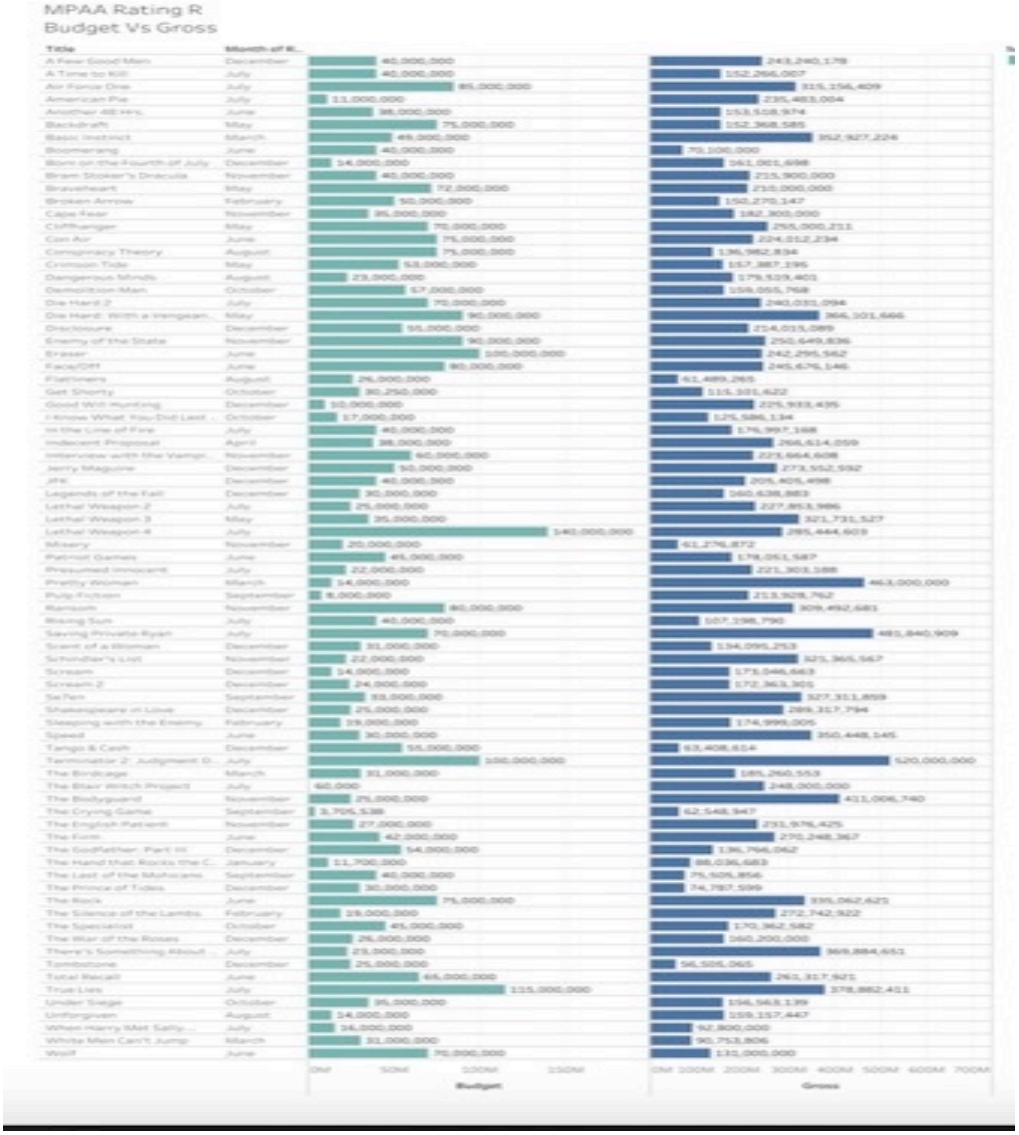
MPAA RATING "PG13" BUDGET VS GROSS VISUAL

Since Tableau was able to process visualizations to answer the questions it was used for the next two visuals as well.   The current trend seen in the previous slide show that budget does not equal a blockbuster at the box office. As can be seen in PG 13 the budget is very close to the Gross revenue and almost breaking even for the movie. The biggest blockbuster in this visual was Titanic.



MPAA RATING "R" BUDGET VS GROSS VISUAL

Visual was created by Tableau since was able to create these visuals. As noticed during this time frame many movies with the Genre R were created by production companies. Many of the movies were a mixture of action, horror, and adult comedy. The biggest blockbuster was Terminator 2.



THE METHODS & CONCLUSION

The reason for the methods I used was to identify various information to understand why movie makers focus on certain genres and how much budget is being used versus the amount of revenue that is coming in. SAS was used to clean up the data and make sure there were no variables out of place which then can be corrected if necessary. Tableau was used to help make distinctions between variables. Though SAS can create visuals to help Tableau was able to create more vivid visuals to identify trends in the data. Overall, both methods allowed the data to processed and provide answers to the research questions.

As a reminder the questions that were created to find the answers were:

1. **Does the budget of the movie impact the ratings of the audience who is watching the movie?**
   Answer: No, we saw in the data movies like lion king, Titanic and Terminator did not have large budgets, but they made large revenue for the box office and their studios,

2. **When a movie releases to the public does the time of the year impact the gross return?**
   Answer: Yes, Around November and December showed great returns for the studios. Home alone showed this specifically.

3. **Does the time of the year release date impact gross sales?**
   Answer: This was also answered movies in summertime or not near any holiday weekends did not show the revenue that holiday time movies did.

4. **Can a movie with a low budget achieve greater sales then a big budget movie?**
   Answer: Yes, many low budget movies made more than high budget movies. Lion King showed this to be true alone.

5. **Does movie genre impact gross sales?**
   Answer: Yes, PG-13 and R had more box office sales over the other categories suggesting more adults were watching movies and interested.

6. **What rating has better revenue for a production company?**
   Answer: PG-13 is the best revenue stream followed by the R rated movies. As shown over the 10 years PG13 movies were on the rise and R movies did not drop as G and PG did.