Top 1000 Highest Grossing Movies

Thant, Phue P

CIS-5250 – Visual Analytics

SAS Project

CONTENTS

About the Dataset 2

Data Description: 3

Data Cleaning 4

Data Visualization 8

Statistical Analysis 1

Summary Statistics 1

One-Way Frequencies 3

Correlation Analysis 1

Linear Regression 3

References 5

TOP 1000 HIGHEST GROSSING MOVIES

ABOUT THE DATASET

Movies industry has become a large revenue generating industry. And the top movies are generating billion dollars sales domestically and internationally, as well as all over the world. The release of new movies and their enormous sales generated from their success have sparked my interest in what attributes have been associated with highly successful and earned movies releases both in the US and around the world. There are many movie datasets available online and I have decided to choose this "Top 1000 highest grossing movies" which is especially for Hollywood movies. Hollywood movies are the top movie industry which has the highest earning around the world. As we are also based in Los Angeles, which is best known for the city of Hollywood, this topic has attracted me to do the analysis for this project.

The objective of this study is to find out which factors plays a role in the success of a movie in terms of sales domestically, internationally, and world sales for Top Hollywood Movies. The below analysis questions are for the movie makers which they should consider before making the decision to release a movie in order to reach the objective. I also wanted to take on this project from the consultant point of view or a quick glance of this industry.

This dataset contains information about the top 1000 highest grossing Hollywood films. It is up to date as of 10th January 2022. It contains over 900 rows and 11 columns.

Source of the dataset (URL) – Top 1000 Highest Grossing Movies | Kaggle

Data Description:

No.	Field Name	Description	Example Values
1	Movie Title	Name of the Movie Data Type: String	Star Wars: Episode VII - The Force Awakens
2	Movie Year	Year of the Movie Data Type: Integer	2015, 2016, 2008, 2022
3	Movie Release Month	Movie Release Month Data Type: String	January, February, March
4	Movie Release Date	Movie Release Date Data Type: Integer	1,24,16,4,1,19,10
5	Movie Release Year	Movie Release Year Data Type: Integer	2015, 2019, 2009, 2018
6	Domestic Sales	Domestic Sales amount in $ Data Type: Float	$936,662,225.00 $858,373,000.0
7	International Sales	International Sales amount in $ Data Type: Float	$1,132,859,475.00 $1,939,128,328.00 $2,086,738,578.00
8	World Sales	World Sales amount in $ Data Type: Float	$2,069,521,700.00 $2,797,501,328.00 $2,847,246,203.00
9	Movie Runtime	Duration of the movie Data Type: Number	2:18 3:1 2:42
10	Movie Runtime (Total Minutes)	Duration of the movie in total Minutes Data Type: Number	138 181 162
11	Distributor	Distributor of the Movie Data Type: String	Walt Disney Studios Motion Pictures Walt Disney Studios Motion Pictures Twentieth Century Fox
12	Genre	Genre of movie (1 movie may have many genres) Data Type: String	['Action', 'Adventure', 'Sci-Fi'] ['Action', 'Adventure', 'Drama', 'Sci-Fi']
13	License	Type of Movie license Data Type: String	PG-13, G, R
14	Movie Info	Summary of the Story Data Type: String	Lion prince Simba and his father are targeted by his bitter uncle, who wants to ascend the throne himself.

DATA CLEANING

1. Data split for Movie title and Year. This data cleaning technique demonstrates the technique of split column. This Data concept takes a string from one column and pushes it to another. I utilized a filter to select the field I wanted to change. Next I inserted a column after the Title field. Then I selected Data, Split Colum Wizard and separated the text using a delimiter (. After I cleaned the data by removing the parentheses.

Original Data

	Title
0	Star Wars: Episode VII - The Force Awakens (2015)
1	Avengers: Endgame (2019)
2	Avatar (2009)
3	Black Panther (2018)
4	Avengers: Infinity War (2018)
5	Spider-Man: No Way Home (2021)
6	Titanic (1997)
7	Jurassic World (2015)
8	The Avengers (2012)

Modified Data

Movie Title	Movie Year
Star Wars: Episode VII - The I	2015
Avengers: Endgame	2019
Avatar	2009
Black Panther	2018
Avengers: Infinity War	2018
Spider-Man: No Way Home	2018
Titanic	1997
Jurassic World	2015
The Avengers	2012

2. This data cleaning technique involves removing NA values and replacing them with values from IMDB. Then I parsed Release Date column into Month, Date and Year in using the split column technique. This involved two techniques filing NA values with logical replacements and separating a Multi Attribute Value into separate fields.

Original Data

B	E
Title	Release Date
Star Wars: Episode VII - The Force Awakens (2015)	16-Dec-15
Avengers: Endgame (2019)	24-Apr-19
Avatar (2009)	16-Dec-09
Black Panther (2018)	NA
Avengers: Infinity War (2018)	NA
Spider-Man: No Way Home (2021)	NA
Titanic (1997)	19-Dec-97
Jurassic World (2015)	10-Jun-15
The Avengers (2012)	25-Apr-12
Star Wars: Episode VIII - The Last Jedi (2017)	13-Dec-17
Incredibles 2 (2018)	NA

Modified Data

Movie Title	Movie Year	Release Month	Release Date	Release Year
Star Wars: Episode VII - The I	2015	January	1	2015
Avengers: Endgame	2019	April	24	2019
Avatar	2009	December	16	2009
Black Panther	2018	January	1	2018
Avengers: Infinity War	2018	January	1	2018
Spider-Man: No Way Home	2018	January	1	2018
Titanic	1997	December	19	1997
Jurassic World	2015	June	10	2015
The Avengers	2012	April	25	2012

3. For the next data cleaning technique, I changed data types. The sales value had a generic number, but to understand the data better I altered the column to represent US Currency in dollars. I altered the data type by selecting the appropriate type I wanted to change to in the Home ribbon. Change the data type for the Sales from General to Currency.

Original Data

Domestic Sales (in $)	International Sale	World Sales (in $)
936662225	1132859475	2069521700
858373000	1939128328	2797501328
760507625	2086738578	2847246203
700426566	647171407	1347597973
678815482	1369544272	2048359754
675813257	868642706	1544455963
659363944	1542283320	2201647264
652385625	1018130819	1670516444
623357910	895457605	1518815515
620181382	712517448	1332698830
608581744	634507500	1243089244

Modified Data

Domestic Sales (in $)	International Sales (in $)	World Sales (in $)
$936,662,225.00	$1,132,859,475.00	$2,069,521,700.00
$858,373,000.00	$1,939,128,328.00	$2,797,501,328.00
$760,507,625.00	$2,086,738,578.00	$2,847,246,203.00
$700,426,566.00	$647,171,407.00	$1,347,597,973.00
$678,815,482.00	$1,369,544,272.00	$2,048,359,754.00
$675,813,257.00	$868,642,706.00	$1,544,455,963.00
$659,363,944.00	$1,542,283,320.00	$2,201,647,264.00
$652,385,625.00	$1,018,130,819.00	$1,670,516,444.00
$623,357,910.00	$895,457,605.00	$1,518,815,515.00
$620,181,382.00	$712,517,448.00	$1,332,698,830.00

4. I changed data type into Time and added the Total Minutes column for movie runtime. For the movie run times to get the actual total minutes, I altered the “Movie Runtime” column to Time format and added total minutes column.

Original Data

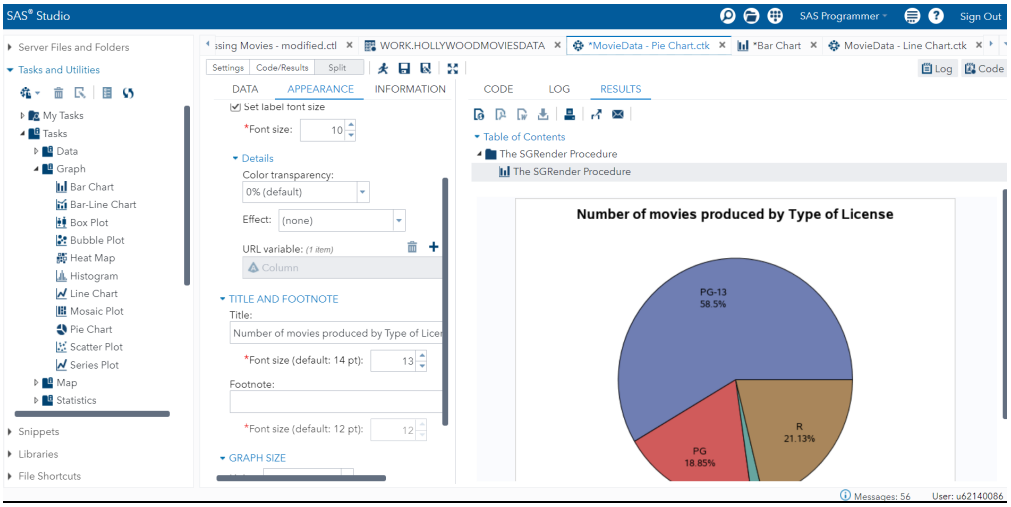
H	I	J	K
World Sales (in $)	Genre	Movie Runtime	License
2069521700	['Action', 'Adv	2 hr 18 min	PG-13
2797501328	['Action', 'Adv	3 hr 1 min	PG-13
2847246203	['Action', 'Adv	2 hr 42 min	PG-13
1347597973	['Action', 'Adv	2 hr 14 min	NA
2048359754	['Action', 'Adv	2 hr 29 min	NA
1544455963	['Action', 'Adv	2 hr 28 min	NA
2201647264	['Drama', 'Ron	3 hr 14 min	PG-13
1670516444	['Action', 'Adv	2 hr 4 min	PG-13
1518815515	['Action', 'Adv	2 hr 23 min	PG-13
1332698830	['Action', 'Adv	2 hr 32 min	PG-13
1243089244	['Action', 'Adv	1 hr 58 min	NA
1662899439	['Adventure', '	1 hr 58 min	PG
1005973645	['Action', 'Crin	2 hr 32 min	PG-13

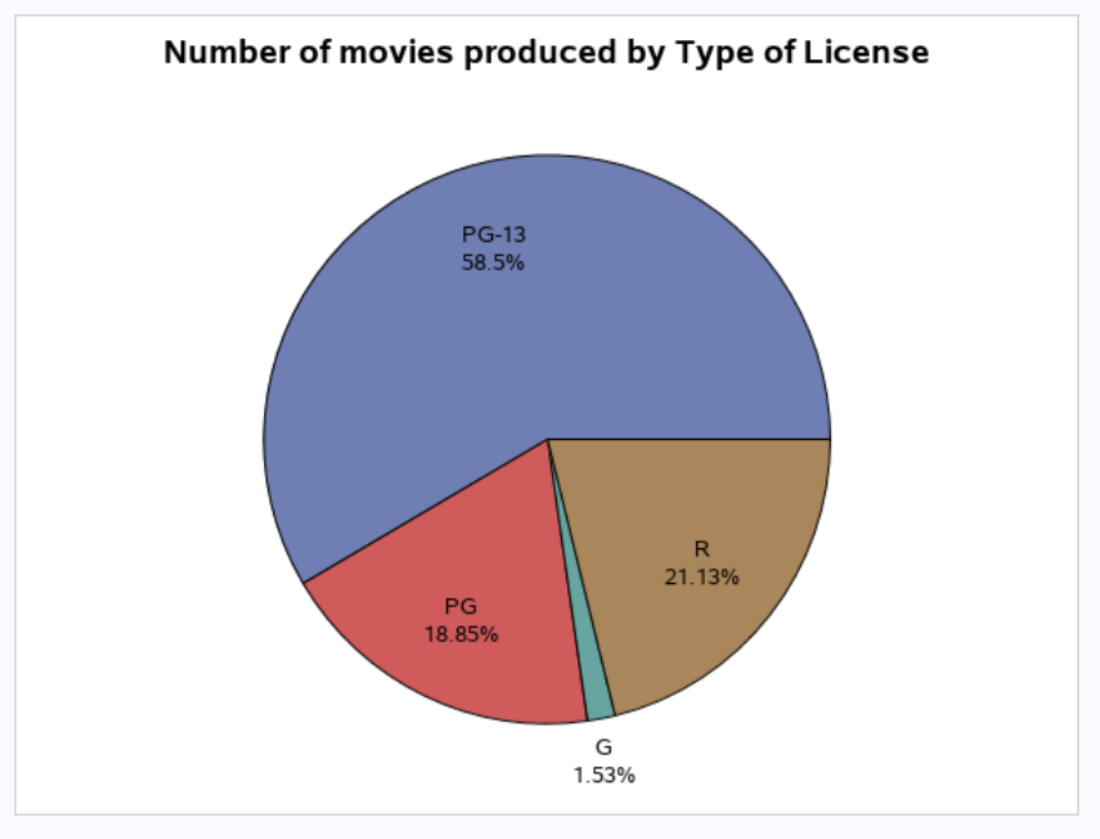
Modified Data

Movie Runtime	Movie Runtime (Total Minutes)
2:18	138
3:1	181
2:42	162
2:14	134
2:29	149
2:28	148
3:14	194
2:4	124
2:23	143
2:32	152
1:58	118
1:58	118

DATA VISUALIZATION

What are the number of movies produced by license?



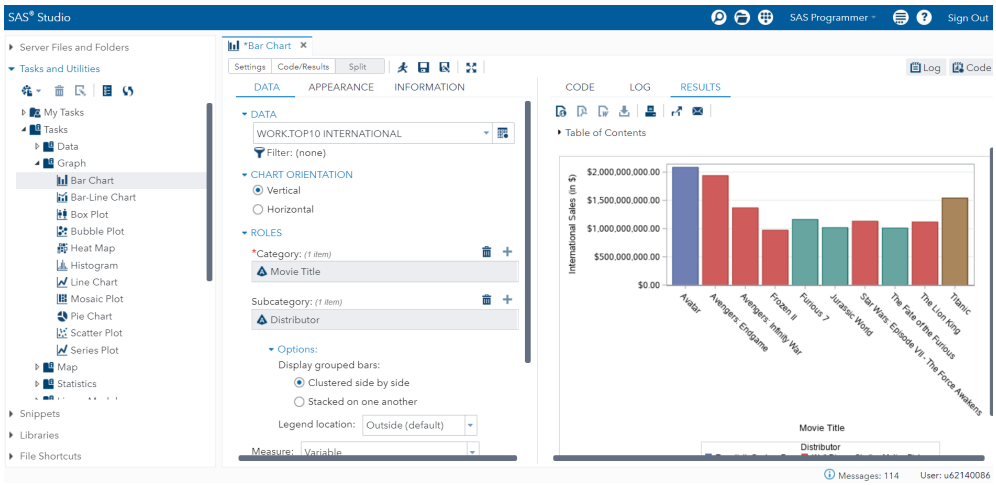


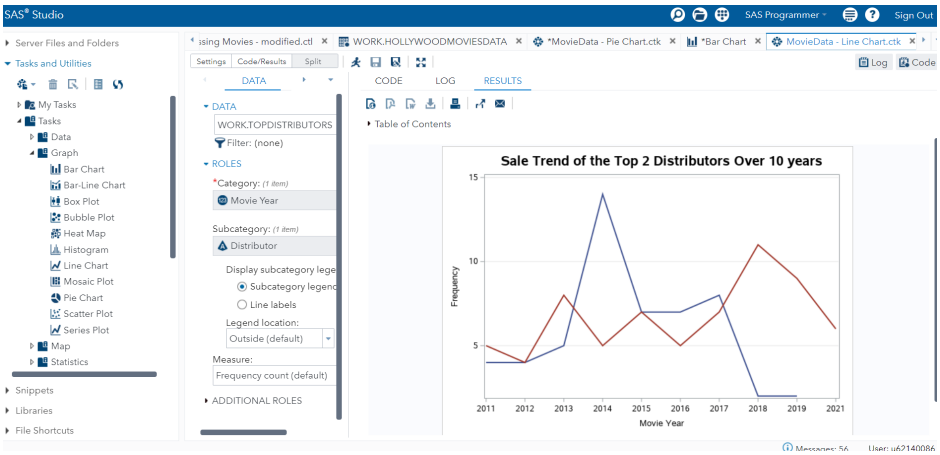
License Types

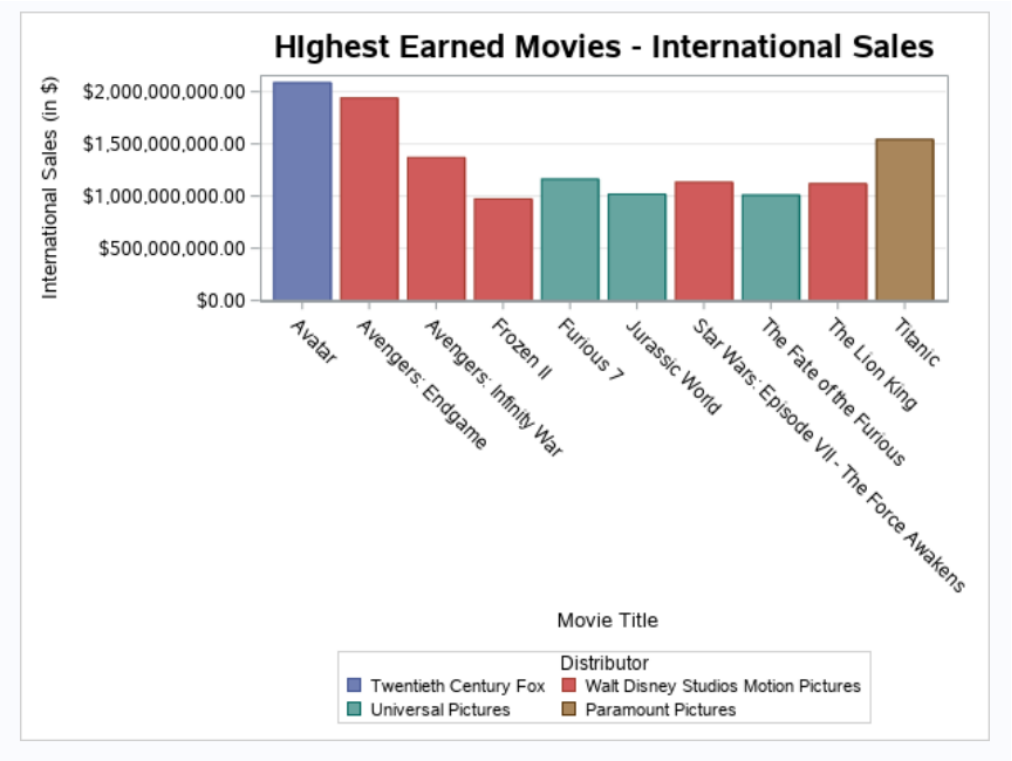
- License PG – Parental Guidance is Suggested.
- License PG-13 – Parents cautioned-to determine whether their children under age 13 are appropriate to watch.
- License R- Children under age 17 are restricted to watch.
- License G- All ages can watch.

We can clearly see that in this Pie chart license type PG-13 have the most percentage which is 58.5%, it means that most of the movies have this license type. Then we have license-R with 21.13%, on second number then license-PG with 18.85% at third number and lastly, we have G license with 1.53%, which have the least number of movies.

Which are the highest earned distributors and what is their sale trends?

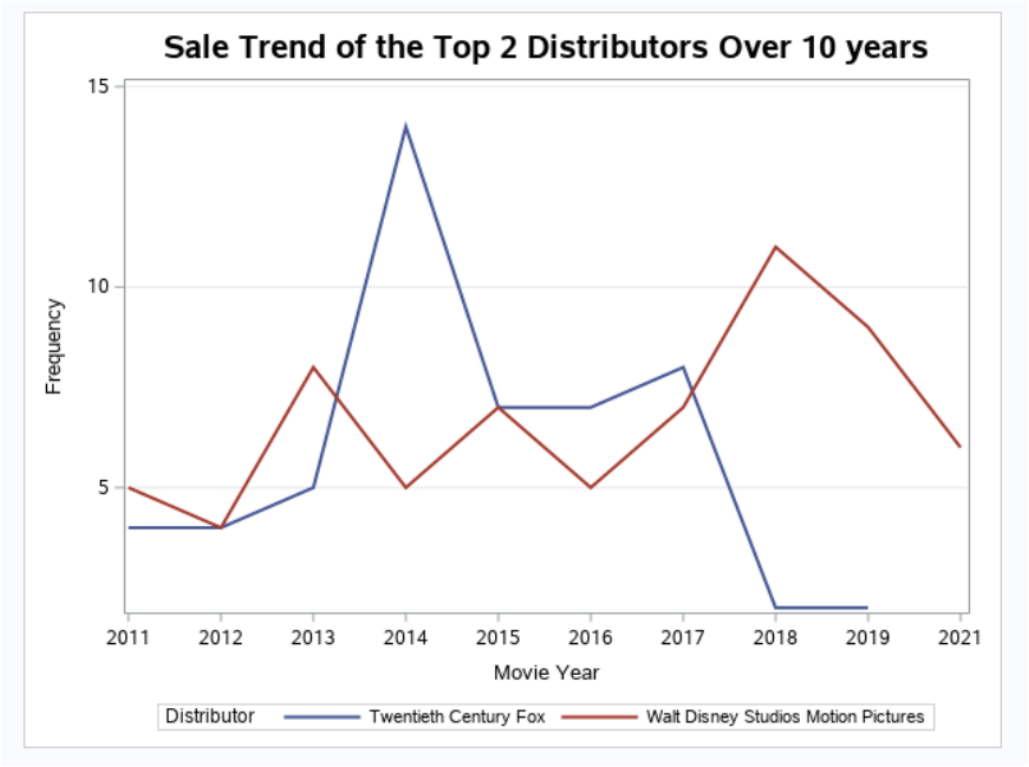






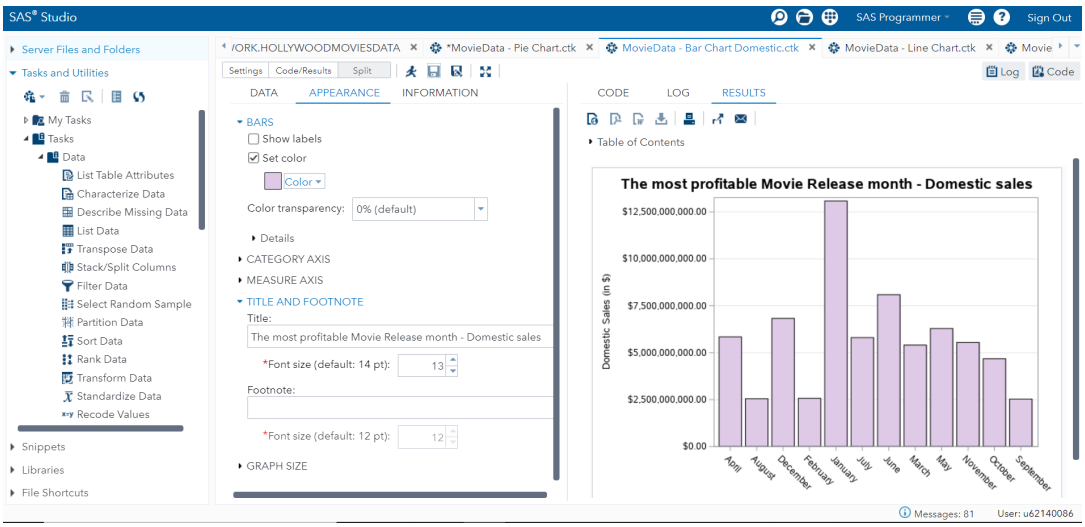
In the bar chart the color of bar represents the movie distributor that which movie is produced by which distributor. Blue is for Twentieth Century Fox, red is for Walt Disney studios Motion Pictures, green is for Universal pictures and brown is for Paramount pictures. We can observe that Avatar has highest sales among all the movies, and it is produced by Twentieth Century Fox. Walt Disney studios Motion Pictures has released the highest number of movies and collectively they

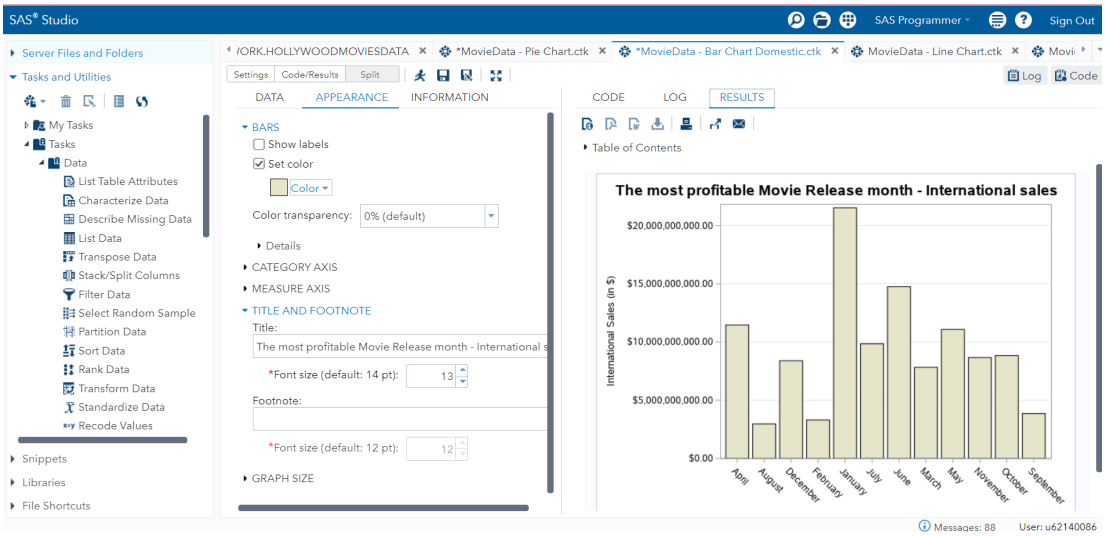
made highest sales and revenue. Then we have Universal pictures and lastly Paramount pictures with least sales.

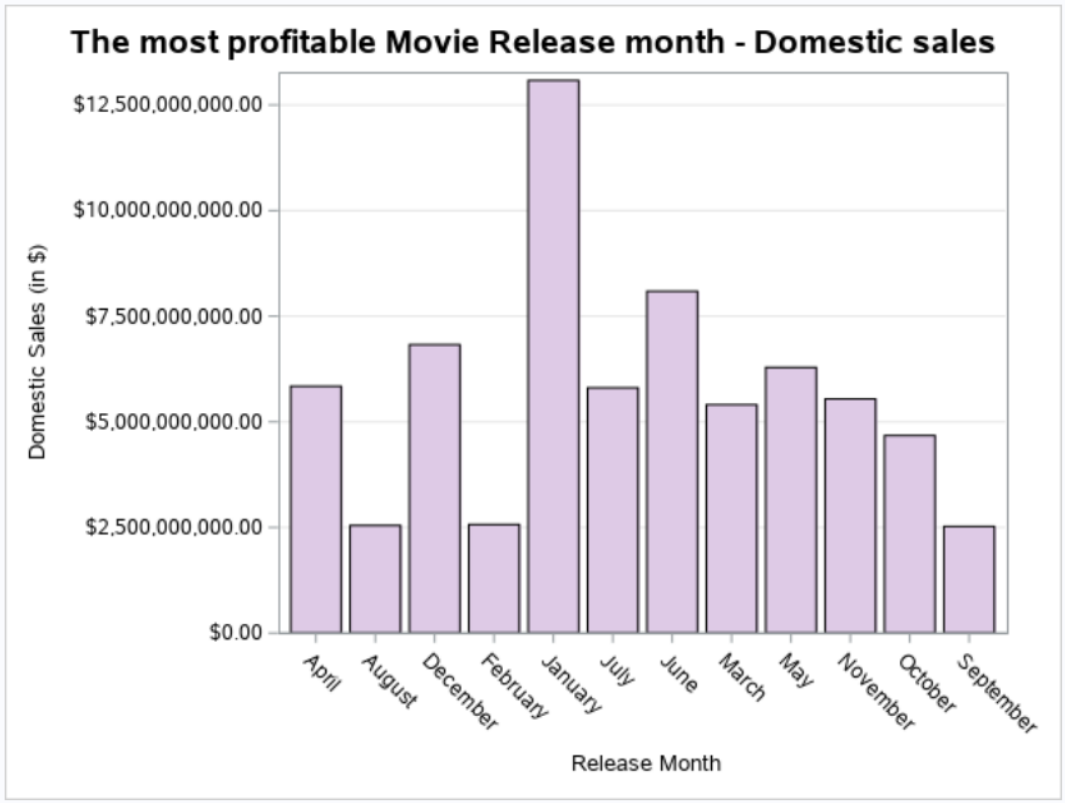


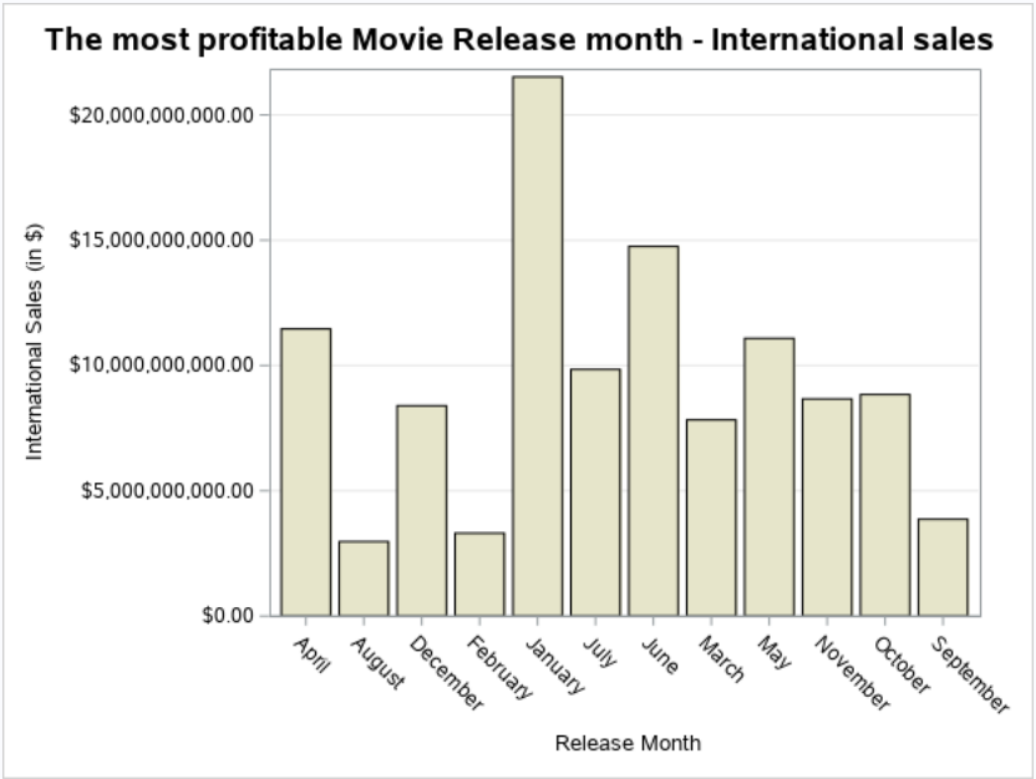
In this line graph, we can observe sale trend of Twentieth century fox and Walt Disney motion pictures over 10 years, the graph is between movie year and frequency. So, we can see that highest frequency of sale trend for TCF is in year 2014 and WDMP's in 2018. The overall performance of both fluctuates over the years but in 2012 both have same frequency. The least frequency of TCF is in 2018 and 2019 which is like almost 0. The least frequency of WDMP is in year 2012.

What are most profitable Movie Release month – Domestic and International Sales?





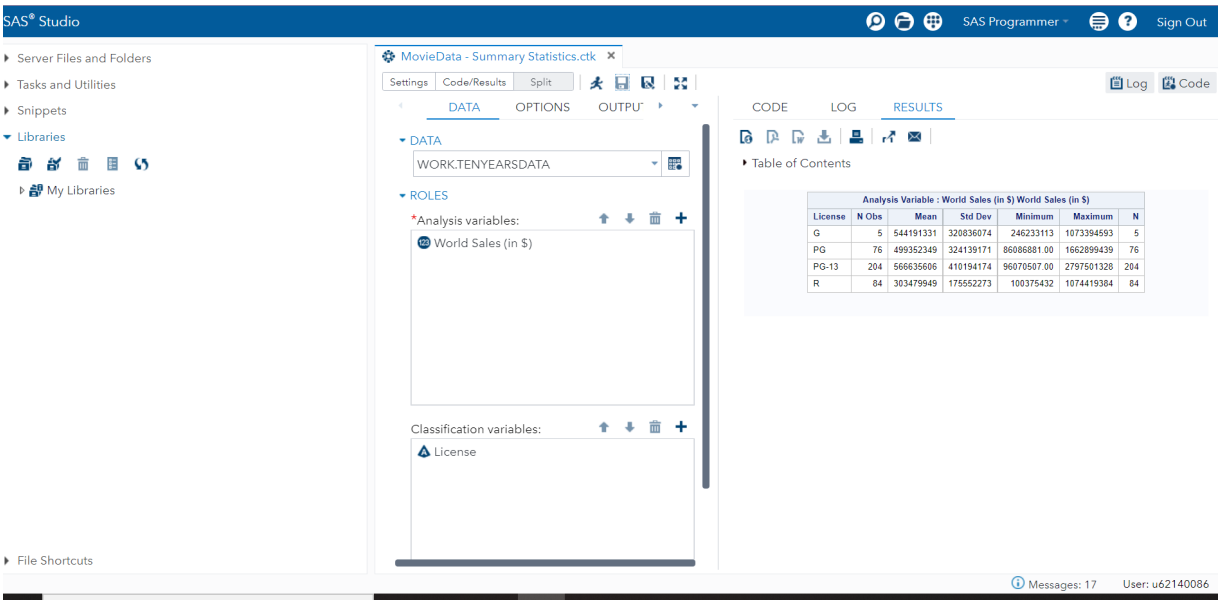

The most profitable Movie Release month - Domestic sales
$12,500,000,000.00
$10,000,000,000.00
$7,500,000,000.00
$5,000,000,000.00
$2,500,000,000.00
$0.00
Domestic Sales (in $)
April
August
December
February
January
July
June
March
May
November
October
September
Release Month


The most profitable Movie Release month - International sales
$20,000,000,000.00
$15,000,000,000.00
$10,000,000,000.00
$5,000,000,000.00
$0.00
International Sales (in $)
April
August
December
February
January
July
June
March
May
November
October
September
Release Month

In two bar chart we can observe that both graphs are between domestic and international sales and the release month of movies. Here we can see that in which month movies have more sales and in which month sales are less. So, in these two bar charts we can see that the most profitable month for movies is January in both domestic and international sales graphs. The overall profit in domestic sales graph for January is $12,500,000,000 and overall profit in international sales graph for January is $20,000,000,000. The months in which profit is least are February, August, and September in domestic sales graph, and the amount is $2,500,000,000 for all these months. In international sales graph the least profit is in August which is less than $5,000,000,000 which is almost $2,500,000,000.

STATISTICAL ANALYSIS

SUMMARY STATISTICS



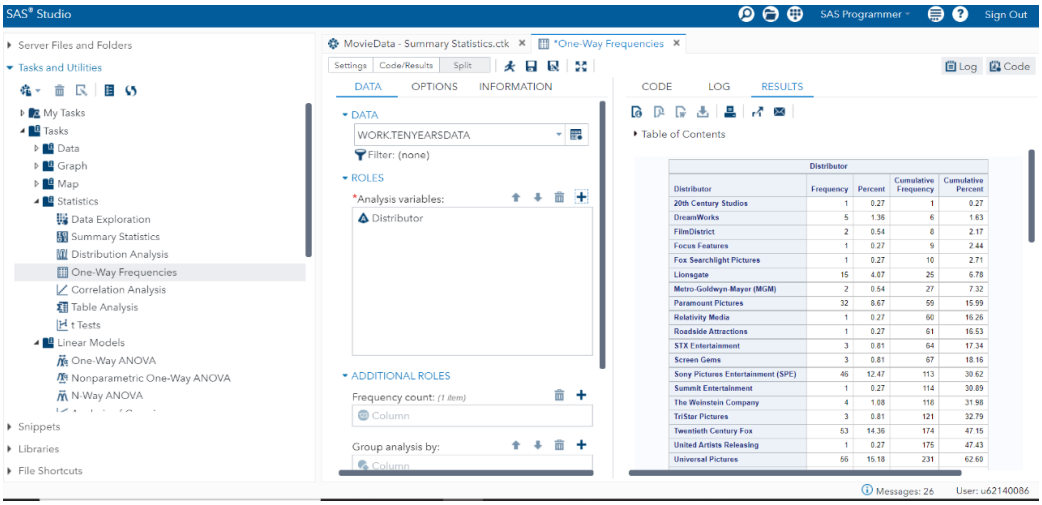
Analysis Variable : World Sales (in $) World Sales (in $)						
License	N Obs	Mean	Std Dev	Minimum	Maximum	N
G	5	544191331	320836074	246233113	1073394593	5
PG	76	499352349	324139171	86086881.00	1662899439	76
PG-13	204	566635606	410194174	96070507.00	2797501328	204
R	84	303479949	175552273	100375432	1074419384	84

Here we have calculated mean by sum up all movies' sales with the corresponding license type. Then we calculate Standard deviation by using its formula: xi is sales of each movie, μ is the mean of license and N is total number of observations

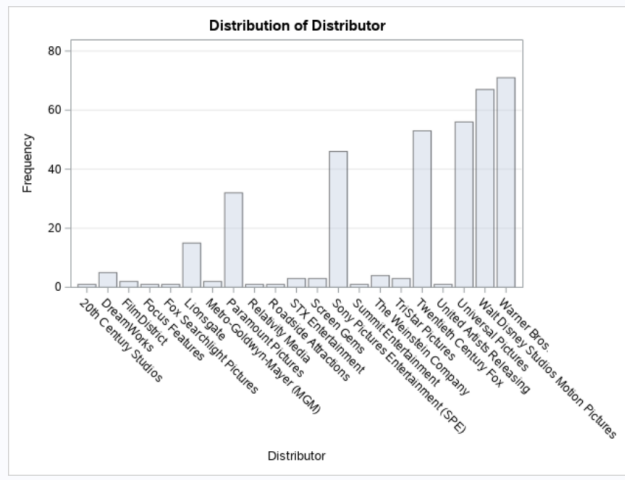
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

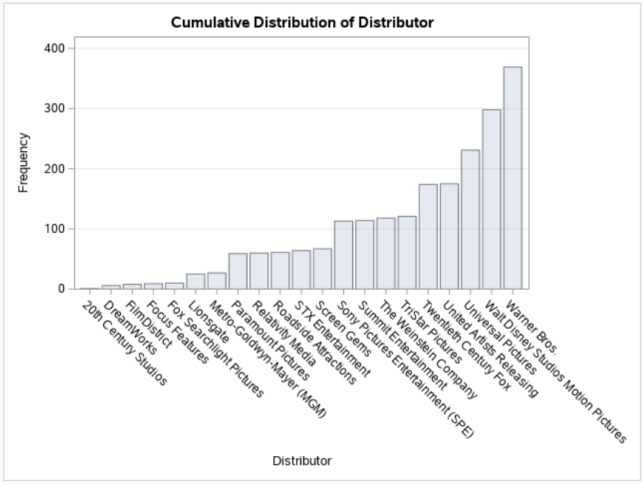
Also, minimum sales and maximum sales over the years. PG-13 license has highest mean which is 566635606 and it also has highest standard deviation 410194174. PG-13 has highest minimum sales and maximum sales. R-license has lowest mean 303479949, standard deviation 175552273, maximum 1074419384 and minimum sales 100375432 which means License R is the lowest type of license in Sales.

ONE-WAY FREQUENCIES



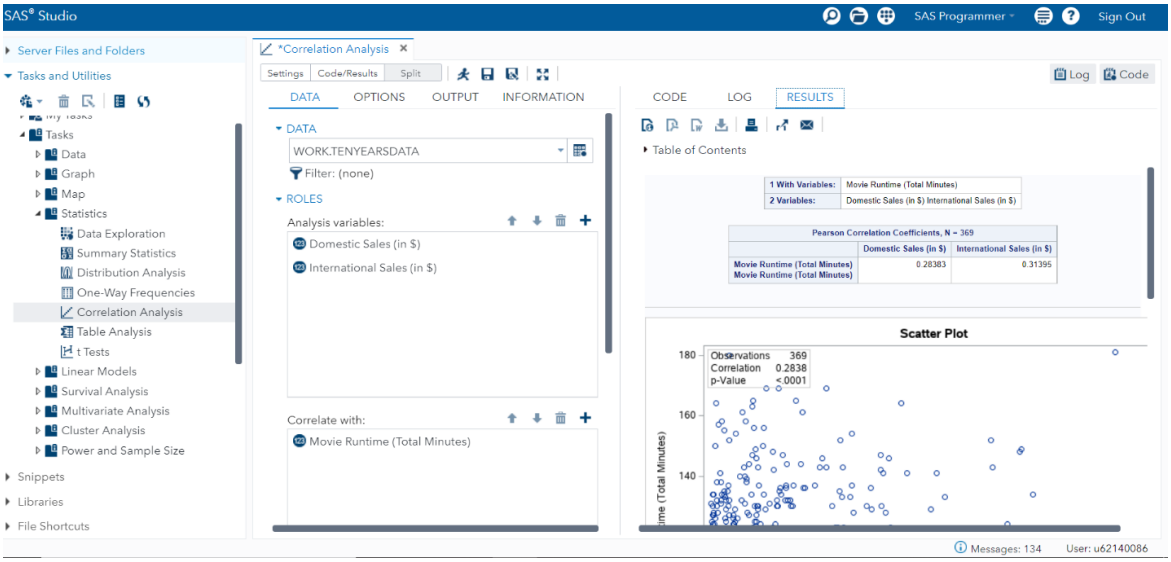
Distributor				
Distributor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
20th Century Studios	1	0.27	1	0.27
DreamWorks	5	1.36	6	1.63
FilmDistrict	2	0.54	8	2.17
Focus Features	1	0.27	9	2.44
Fox Searchlight Pictures	1	0.27	10	2.71
Lionsgate	15	4.07	25	6.78
Metro-Goldwyn-Mayer (MGM)	2	0.54	27	7.32
Paramount Pictures	32	8.67	59	15.99
Relativity Media	1	0.27	60	16.26
Roadside Attractions	1	0.27	61	16.53
STX Entertainment	3	0.81	64	17.34
Screen Gems	3	0.81	67	18.16
Sony Pictures Entertainment (SPE)	46	12.47	113	30.62
Summit Entertainment	1	0.27	114	30.89
The Weinstein Company	4	1.08	118	31.98
TriStar Pictures	3	0.81	121	32.79
Twentieth Century Fox	53	14.36	174	47.15
United Artists Releasing	1	0.27	175	47.43
Universal Pictures	56	15.18	231	62.60
Walt Disney Studios Motion Pictures	67	18.16	298	80.76
Warner Bros.	71	19.24	369	100.00





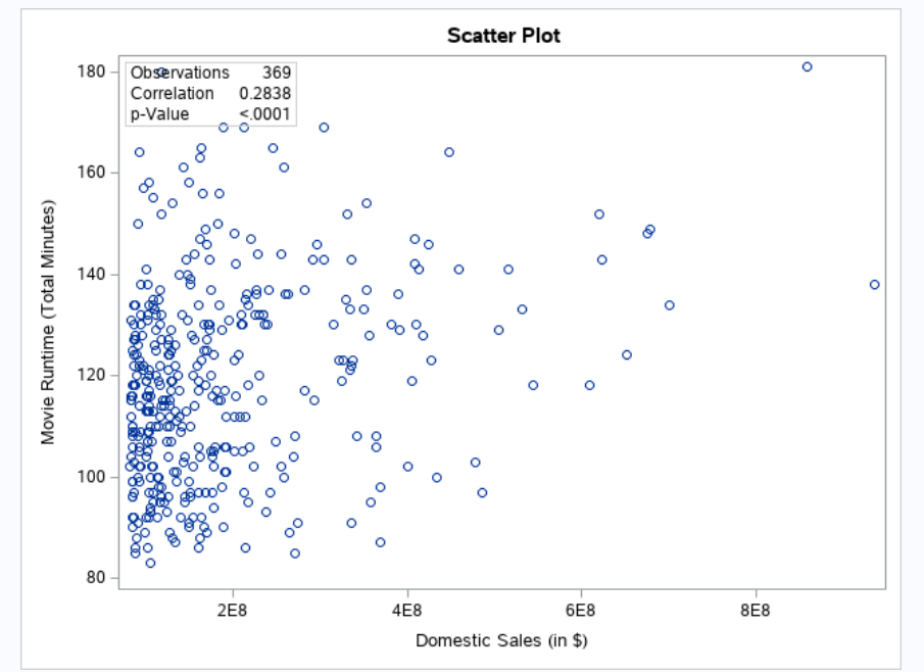
This one-way frequency table shows frequency, frequency percentage, cumulative frequency, and its percentage of different movie distributors. There are 21 distributors which are listed in table. If we observe we will know that Walt Disney Studios Motion Pictures has highest earning profits among all the distributors, with the highest frequency of 67 and percentage is 18.6%. The cumulative frequency is 298 and percentage is 80.76%. On the other hand, the distributor who has the least frequency are 20th Century Studios, focus features, Fox Searchlight Pictures, Relativity Media, Roadside Attractions, Summit Entertainment, United Artists Releasing with frequency of 1 and percentage 0.27%. 20t Century Studios has least cumulative frequency which is 1 and percentage is 0.27%.

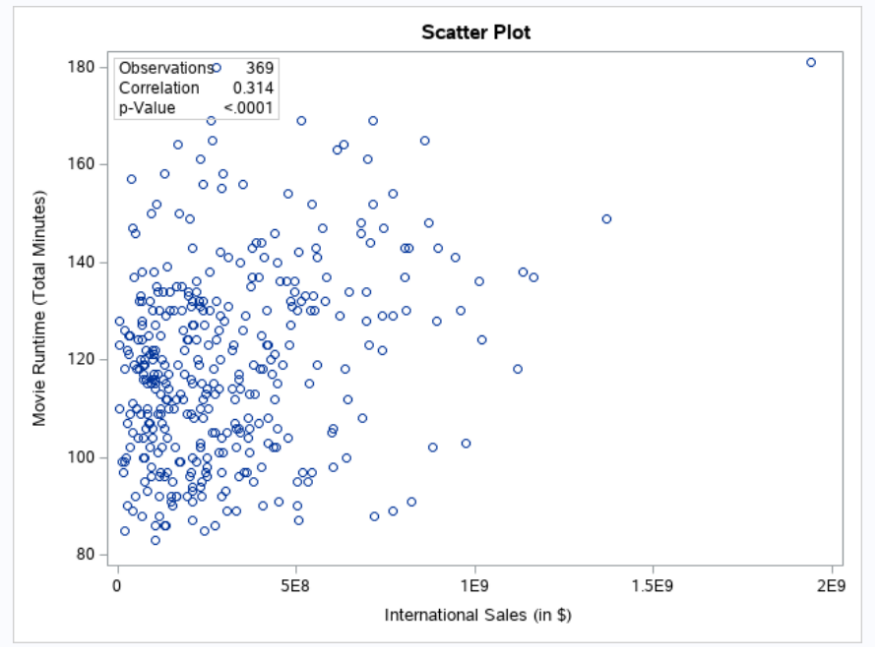
CORRELATION Analysis



1 With Variables:	Movie Runtime (Total Minutes)
2 Variables:	Domestic Sales (in $) International Sales (in $)

Pearson Correlation Coefficients, N = 369		
	Domestic Sales (in $)	International Sales (in $)
Movie Runtime (Total Minutes) Movie Runtime (Total Minutes)	0.28383	0.31395

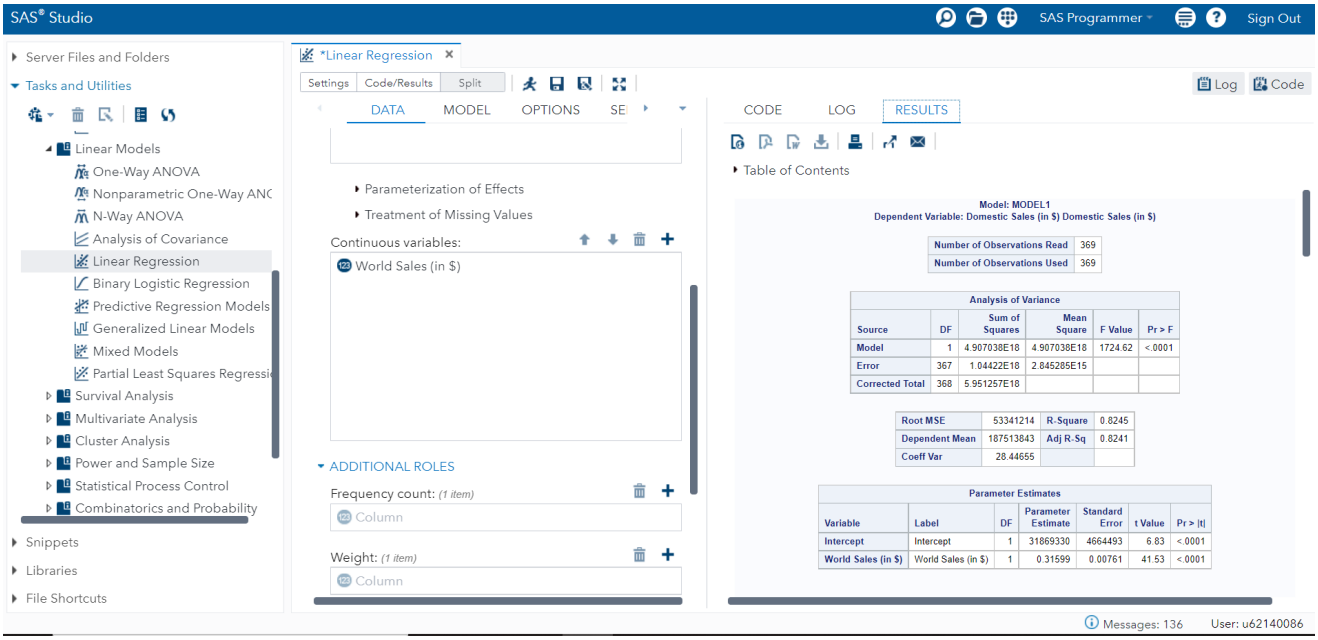




The table shows that the correlation coefficient of Movie Runtime (total minutes) and Domestic sales (in $) is 0.28383. And correlation coefficient of Movie Runtime (total minutes) and international sales (in $) is 0.31395. Correlation coefficient “r” indicates the degree of linear

relationship between the two variables (Movie runtime and Domestic sales) and (Movie runtime and International Sales). Here, there is a poor positive correlation (between Movie runtime and domestic sales) and fair positive correlation (between Movie runtime and international sales).

LINEAR REGRESSION



Model: MODEL1
Dependent Variable: Domestic Sales (in $) Domestic Sales (in $)

Number of Observations Read	369
Number of Observations Used	369

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.907038E18	4.907038E18	1724.62	<.0001
Error	367	1.04422E18	2.845285E15		
Corrected Total	368	5.951257E18			

Root MSE	53341214	R-Square	0.8245
Dependent Mean	187513843	Adj R-Sq	0.8241
Coeff Var	28.44655		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	31869330	4664493	6.83	<.0001
World Sales (in $)	World Sales (in $)	1	0.31599	0.00761	41.53	<.0001

The table indicates whether the correlation coefficient (R) is significant or not. As the p-value (Sig.) is less than 0.0001, R is statistically significant. Here, the R-square value is 0.8245, which indicates that 82.45% variation in World sales can be explained by Domestic sales. The rest of the variation (17.55%) is due to other factors. And the value of “a” is 0.31599 and the value of “b” is 31869330, which means that if the domestic sales increases by 1, world sales will increase 0.31599. The table also shows the significance (p-value), which is greater than 0.0001.

REFERENCES

Top 1000 Highest Grossing Movies. (2022, January 15). Kaggle.

https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies

Song, C. (2018, November 29). *Film & Cinema Industry Analysis – Industrial Engineering Era*.

http://iera.name/film-cinema-industry-analysis/

Investigate TMDb Movie Dataset. (2018, September 15). Kaggle.

https://www.kaggle.com/code/deepak525/investigate-tmdb-movie-dataset