10/31/2022

SAS Project

TMDB Movie Dataset

Antaliya, Ruta
CIS 5250: Visual Analytics

Contents

Introduction

Movies affect many of us powerfully because the combined impact of images, music, dialogue, lighting, sound, and computer graphics can elicit deep feelings and help us reflect on our lives. They will help us to understand our own lives, the lives of those around us and even how our society and culture operate. I chose “tmdb_movie_data” topic because of its valuable features and benefits in application development. Also, it provides information about movie name, director, and production companies between year 1960 to 2015. TMDB from (Anjana, 2020) has been adopted by start-ups and some leading companies such as Netflix, Amazon, Hot star, and Hulu to scale as well as handle their large operations. In the current scenario, as the applications are becoming increasingly complex, architecting the entire application from start to end is becoming nearly impossible. The Movie DB (TMDb) is a free and community edited database. The TMDb API Track this API is a Resource for any developers that want to integrate movie, TV show and cast data along with posters or movie fan art.

The primary objective of this dataset is to find out most Profitable movie, famous director of movie, most popular production company, movie with highest budget and longest movie of the all-time. This analysis help to quick view of the movie DB(TMDB). This dataset contains information about 10,000 movies collected from The Movie Database (TMDb), including user title, budget, revenue, cast, director, tagline, genres, release date and runtime. It contains 10866 rows and 21 columns. But after analyzing the dataset found some columns is contains null value.

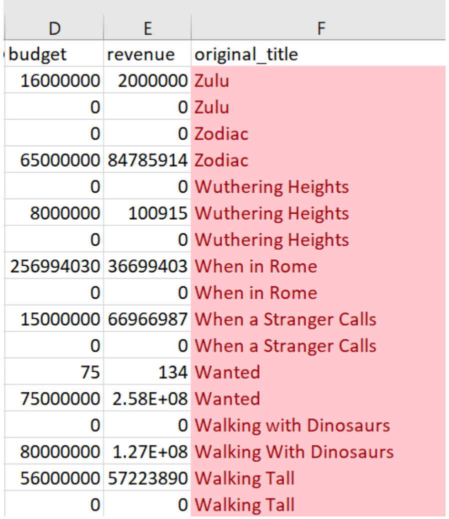
Dataset URL: https://www.kaggle.com/datasets/juzershakir/tmdb-movies-dataset

<u>Data Description</u>

Field Name	**Description**	**Example Values**
Popularity	A numeric quantity specifying the movie popularity Data Type: Float	9.432768, 13.112507, 2.586787
Budget	In which the movie was made Data Type: Number	133000000, 20000000, 38000000
Revenue	A revenue generated by movie Data Type: Number	296221663, 1327817822, 1081041287
Original_title	The title of movie Data Type: Char	Jurassic World, Minions, Iron Man 3
Director	Director of movie Data Type: Char	David Yate, Steven Spielberg, Clint Eastwood
Runtime	Length of the movie Data Type: Number	114, 160, 91
Genres	The genre of the movie, Action, Comedy, Thriller etc. Data Type: Char	Adventure Fantasy, Action, Drama, Comedy, etc.
Production_companies	Production Company of movie Data Type: Char	Universal Pictures, Warner Bros, and Paramount Pictures
Release_Date	Date was release movie Data Type: Date	12/10/2009, 11/27/2013, 12/15/1974
Release_year	Movie release year Data Type: Number	1995, 2000, 2005
Vote_count	Average vote of movie Data Type: Number	9767, 6185, 710
Movie_Ratings	Rating of the movie Data Type: Float	2.3, 5.8, 8.4

Data Cleaning

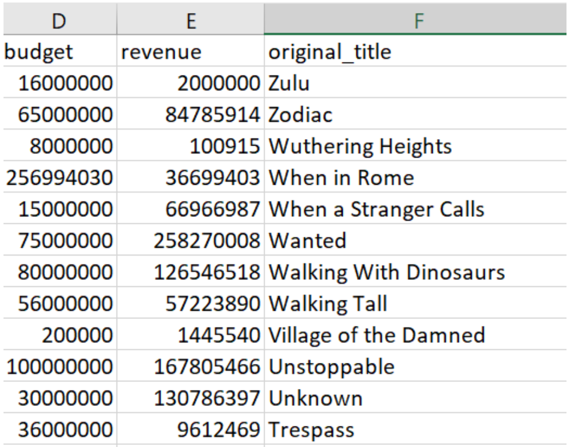
1. In the analyzing, Tmdb_movie_data dataset I saw some duplicate rows in title field. So, I did conditional formatting style to check how many duplicate rows present in title column. When the duplicate value rules applied in title field it created red cell color value so it easy to remove row with duplicate value.



D	E	F
budget	revenue	original_title
16000000	2000000	Zulu
0	0	Zulu
0	0	Zodiac
65000000	84785914	Zodiac
0	0	Wuthering Heights
8000000	100915	Wuthering Heights
0	0	Wuthering Heights
256994030	36699403	When in Rome
0	0	When in Rome
15000000	66966987	When a Stranger Calls
0	0	When a Stranger Calls
75	134	Wanted
75000000	2.58E+08	Wanted
0	0	Walking with Dinosaurs
80000000	1.27E+08	Walking With Dinosaurs
56000000	57223890	Walking Tall
0	0	Walking Tall

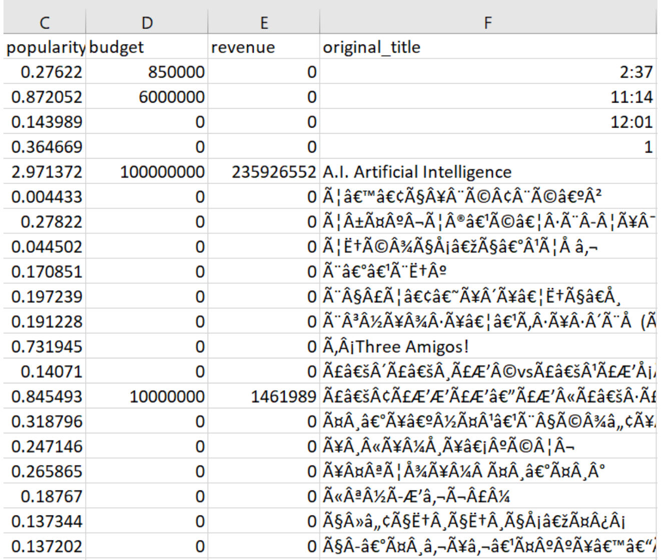
Modified data

Deleted rows based on zero budget and revenue value. After removing the duplicate value in column, the following table created. Applied Sorting filter to check.



D	E	F
budget	revenue	original_title
16000000	2000000	Zulu
65000000	84785914	Zodiac
8000000	100915	Wuthering Heights
256994030	36699403	When in Rome
15000000	66966987	When a Stranger Calls
75000000	258270008	Wanted
80000000	126546518	Walking With Dinosaurs
56000000	57223890	Walking Tall
200000	1445540	Village of the Damned
100000000	167805466	Unstoppable
30000000	130786397	Unknown
36000000	9612469	Trespass

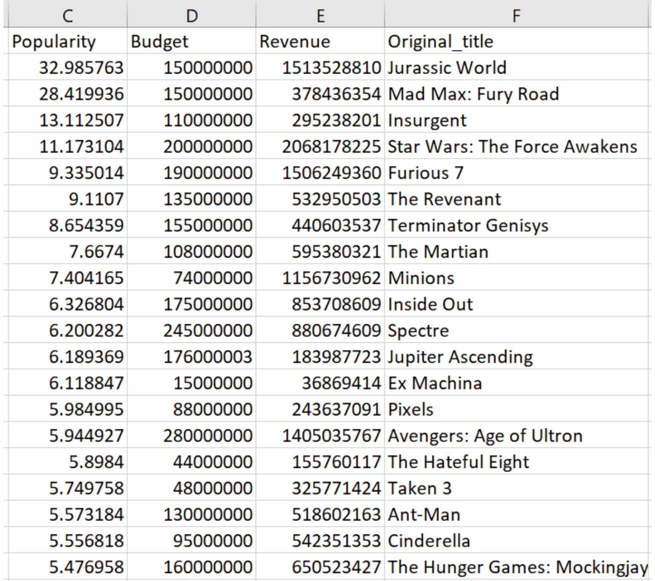
2. Applying Sorting filter found some rows with number value and some rows with special character in movie title field. So removed these rows from original title field. I can easily which movie is longest in analysis part.



C	D	E	F
popularity	budget	revenue	original_title
0.27622	850000	0	2:37
0.872052	6000000	0	11:14
0.143989	0	0	12:01
0.364669	0	0	1
2.971372	100000000	235926552	A.I. Artificial Intelligence
0.004433	0	0	Ã¦â€™â€¢Ã§Â¥Â¨Ã©Â¢Â¨Ã©â€ºÂ²
0.27822	0	0	Ã¦Â±Ã¤ÂºÂ¬Ã¦Â®â€¹Ã©â€¦Â·Ã¨ÂÂ¦Ã¥Â
0.044502	0	0	Ã¦Ë†Ã©Â¾Ã§Å¡â€žÃ§â€°Â¹Ã¦Å â‚¬
0.170851	0	0	Ã¨â€°â€¹Ã¨Ë†Âº
0.197239	0	0	Ã¨Â§Â£Ã¦â€¢â€˜Ã¥Â´Ã¥â€¦Ë†Ã§â€Å¸
0.191228	0	0	Ã¨Â³Â½Ã¥Â¾Â·Ã¥â€¦â€¹Ã‚Â·Ã¥Â·Â´Ã¨Å (Ã
0.731945	0	0	Ã‚Â¡Three Amigos!
0.14071	0	0	Ã£â€šÂ´Ã£â€šÂ¸Ã£Æ’Â©vsÃ£â€šÂ¹Ã£Æ’Å¡
0.845493	10000000	1461989	Ã£â€šÂ¢Ã£Æ’Æ’Ã£Æ’â€”Ã£Æ’Â«Ã£â€šÂ·Ã
0.318796	0	0	Ã¤Â¸â€°Ã¥â€ºÂ½Ã¤Â¹â€¹Ã¨Â§Ã©Â¾â„¢Ã¥
0.247146	0	0	Ã¥Â¸Â«Ã¥Â¼Å¸Ã¥â€¡ÂºÃ©Â¦Â¬
0.265865	0	0	Ã¥Â¤ÂªÃ¦Å¾Ã¥Â¼Â Ã¤Â¸â€°Ã¤Â¸Â°
0.18767	0	0	Ã«ÂªÂ½ÃÆ’â‚¬Ã¬Â£Â¼
0.137344	0	0	Ã§Â»â„¢Ã§Ë†Â¸Ã§Ë†Â¸Ã§Å¡â€žÃ¤Â¿Â¡
0.137202	0	0	Ã§Ââ€°Ã¤Â¸â‚¬Ã¥â‚¬â€¹Ã¤ÂºÂºÃ¥â€™â€œ

Modified Data

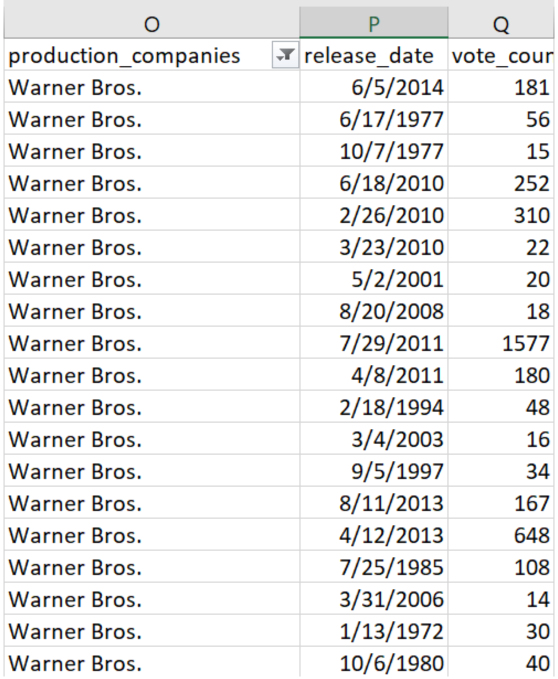
After deleting rows which contains special character and number value in original title the following table created.



C	D	E	F
Popularity	Budget	Revenue	Original_title
32.985763	150000000	1513528810	Jurassic World
28.419936	150000000	378436354	Mad Max: Fury Road
13.112507	110000000	295238201	Insurgent
11.173104	200000000	2068178225	Star Wars: The Force Awakens
9.335014	190000000	1506249360	Furious 7
9.1107	135000000	532950503	The Revenant
8.654359	155000000	440603537	Terminator Genisys
7.6674	108000000	595380321	The Martian
7.404165	74000000	1156730962	Minions
6.326804	175000000	853708609	Inside Out
6.200282	245000000	880674609	Spectre
6.189369	176000003	183987723	Jupiter Ascending
6.118847	15000000	36869414	Ex Machina
5.984995	88000000	243637091	Pixels
5.944927	280000000	1405035767	Avengers: Age of Ultron
5.8984	44000000	155760117	The Hateful Eight
5.749758	48000000	325771424	Taken 3
5.573184	130000000	518602163	Ant-Man
5.556818	95000000	542351353	Cinderella
5.476958	160000000	650523427	The Hunger Games: Mockingjay

3. Applying sorting filter found zero value in budget and revenue column. So, rows removed with zero value from these two fields for better visualization purpose.

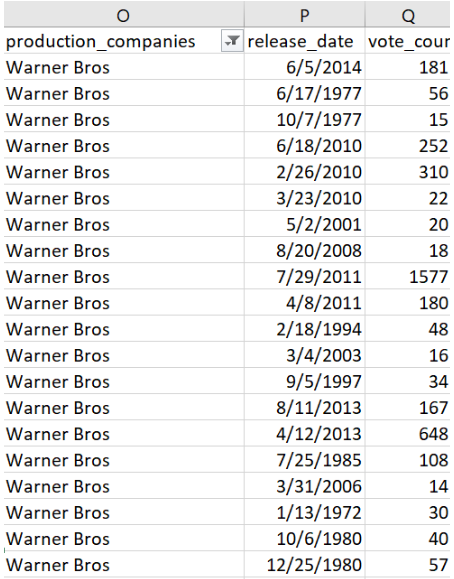
4. Sorting Production company field found some special character "." was present in this column. So applied replace filter to remove special character from production companies field.



O	P	Q
production_companies	release_date	vote_cour
Warner Bros.	6/5/2014	181
Warner Bros.	6/17/1977	56
Warner Bros.	10/7/1977	15
Warner Bros.	6/18/2010	252
Warner Bros.	2/26/2010	310
Warner Bros.	3/23/2010	22
Warner Bros.	5/2/2001	20
Warner Bros.	8/20/2008	18
Warner Bros.	7/29/2011	1577
Warner Bros.	4/8/2011	180
Warner Bros.	2/18/1994	48
Warner Bros.	3/4/2003	16
Warner Bros.	9/5/1997	34
Warner Bros.	8/11/2013	167
Warner Bros.	4/12/2013	648
Warner Bros.	7/25/1985	108
Warner Bros.	3/31/2006	14
Warner Bros.	1/13/1972	30
Warner Bros.	10/6/1980	40

Modified Data

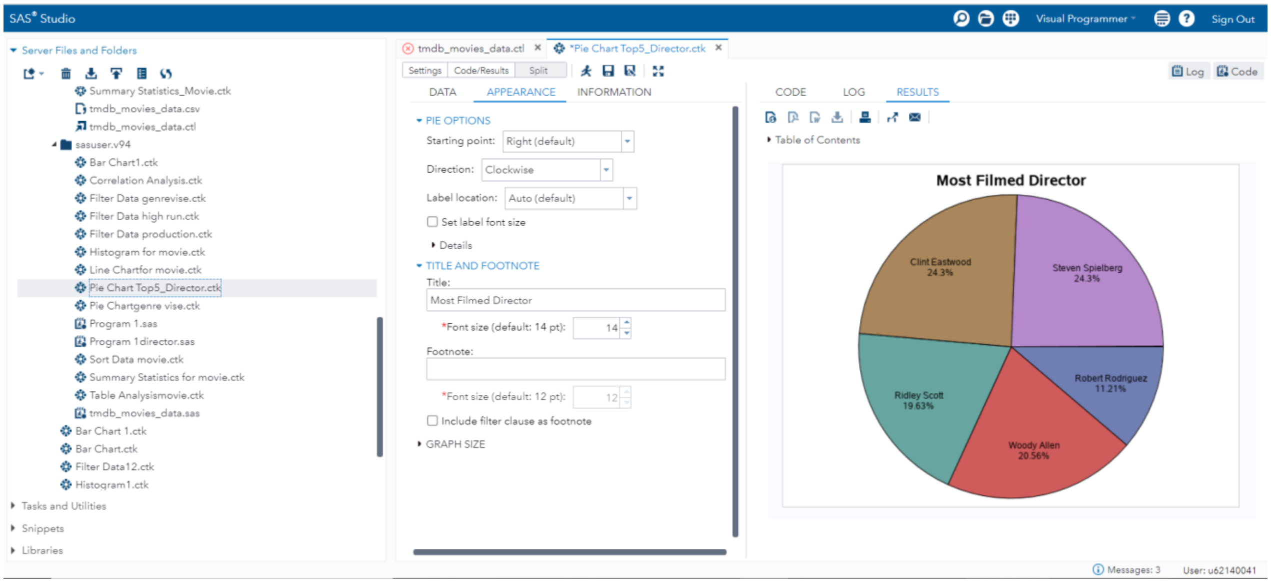
By replacing Special character with blank the following dataset created. When it will use in analyzing purpose it creates better visualization.



O	P	Q
production_companies	release_date	vote_cour
Warner Bros	6/5/2014	181
Warner Bros	6/17/1977	56
Warner Bros	10/7/1977	15
Warner Bros	6/18/2010	252
Warner Bros	2/26/2010	310
Warner Bros	3/23/2010	22
Warner Bros	5/2/2001	20
Warner Bros	8/20/2008	18
Warner Bros	7/29/2011	1577
Warner Bros	4/8/2011	180
Warner Bros	2/18/1994	48
Warner Bros	3/4/2003	16
Warner Bros	9/5/1997	34
Warner Bros	8/11/2013	167
Warner Bros	4/12/2013	648
Warner Bros	7/25/1985	108
Warner Bros	3/31/2006	14
Warner Bros	1/13/1972	30
Warner Bros	10/6/1980	40
Warner Bros	12/25/1980	57

Analysis & Data Visualization

1. Which director was most filmed?



PIE OPTIONS

Starting point: Right (default)

Direction: Clockwise

Label location: Auto (default)

☐ Set label font size

Details

TITLE AND FOOTNOTE

Title: Most Filmed Director

*Font size (default: 14 pt): 14

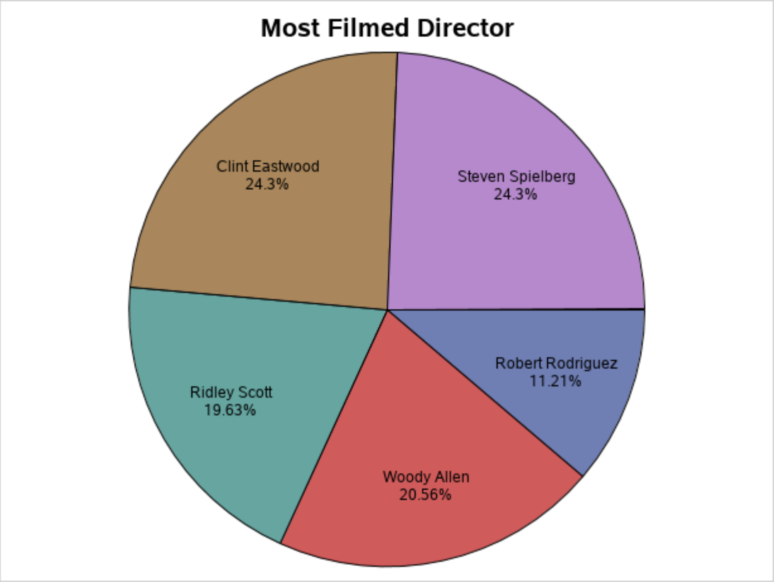
Footnote:

*Font size (default: 12 pt): 12

☐ Include filter clause as footnote

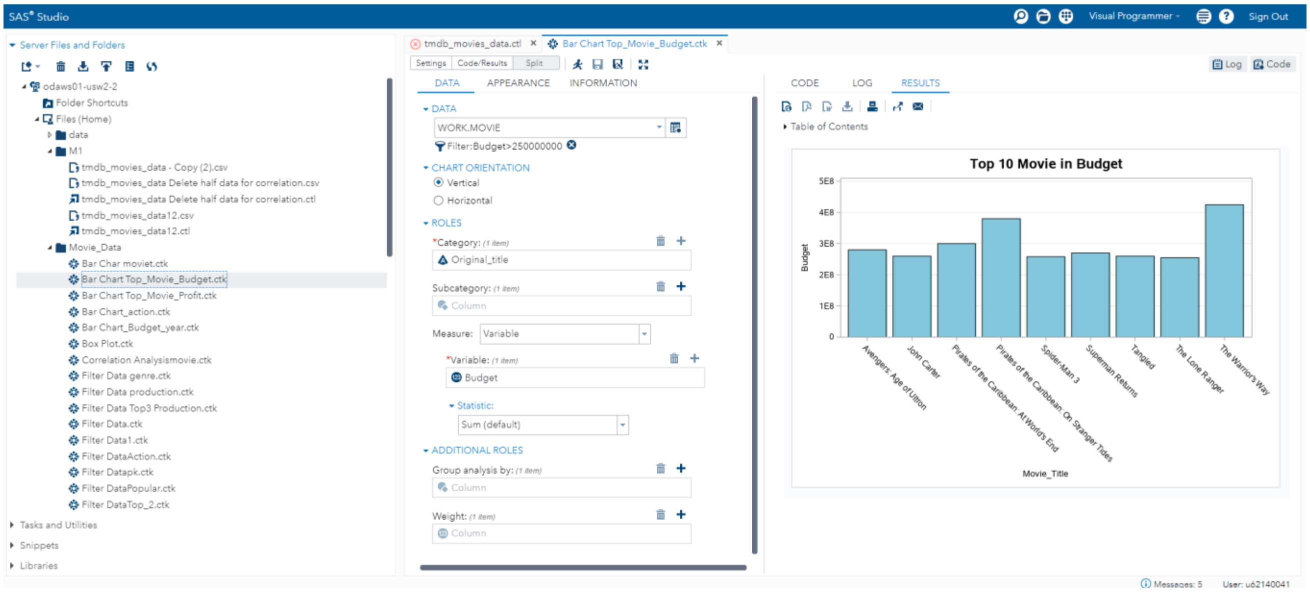
GRAPH SIZE

Messages: 3 User: u62140041



The above pie charts depict top5 director who direct maximum movie between year 1960 – 2015. It clearly shows clint Eastwood and Steven director are directed by most of the movie in these years. Approximately 50% of movie was directed by these two directors. Woody Allen is the third highest directed movie director, which directed more than 20% of movies among 5 decades. Robert Rodriguez directed 11.21% of movie which is lesser than another movie directors. 19.63% of movie was directed by Ridley Scott which is approximate similar to Woody Allen director.

2. Which Movie has highest Budget?



DATA

WORK.MOVIE

Filter:Budget>250000000

CHART ORIENTATION

● Vertical

○ Horizontal

ROLES

*Category: (1 item)
Original_title

Subcategory: (1 item)
Column

Measure: Variable

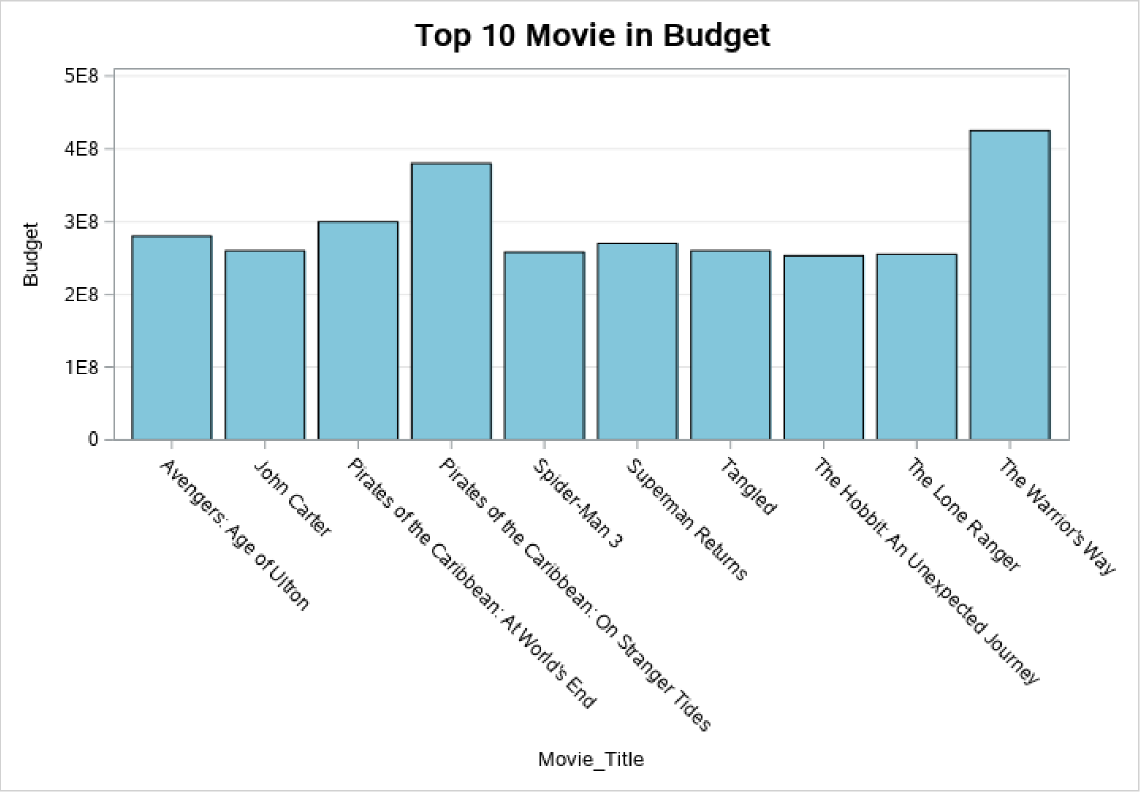
*Variable: (1 item)
Budget

Statistic:
Sum (default)

ADDITIONAL ROLES

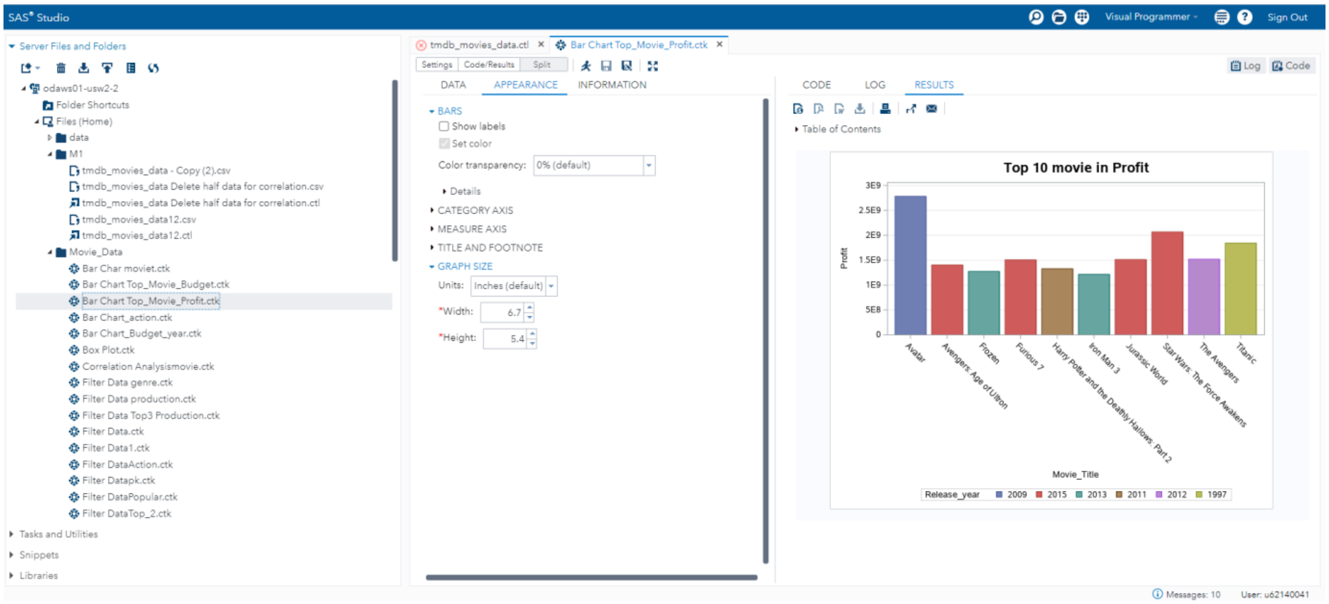
Group analysis by: (1 item)
Column

Weight: (1 item)
Column

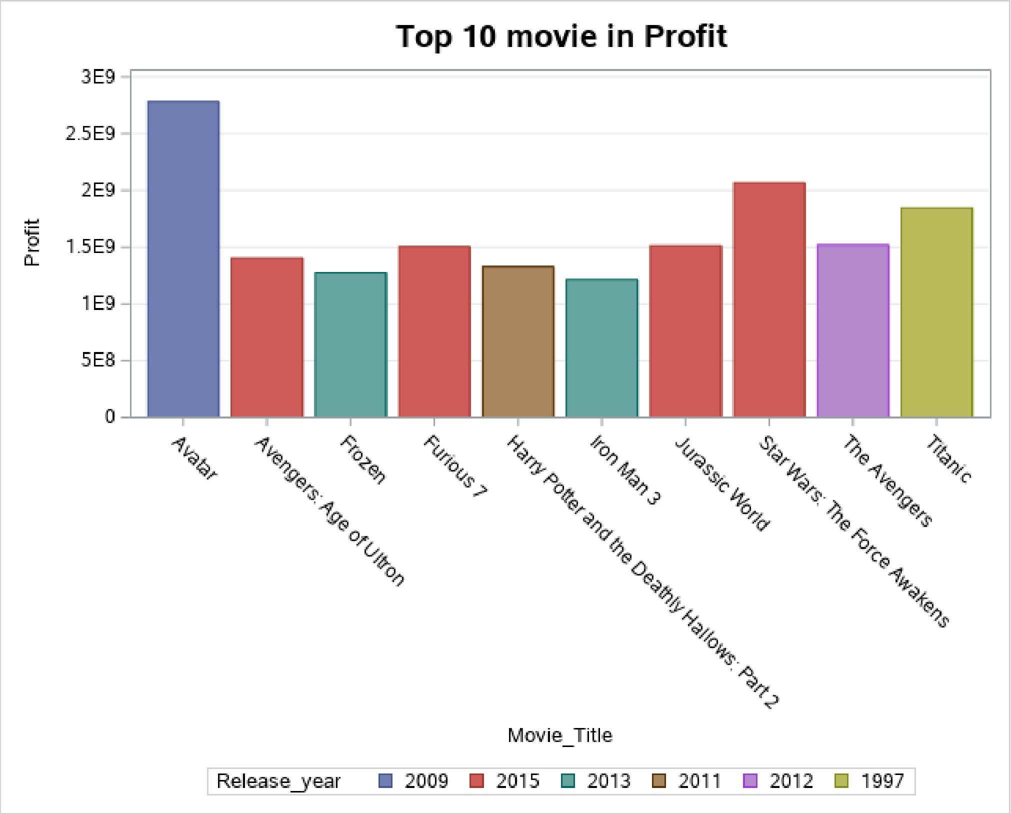


The bar chart shows the top10 movies which has highest budget between 1960 and 2015. The data is shown as billion. The warrior’s way movie had the highest budget, which was around 0.42billion. The second and third highest budget movie were Pirates of the Caribbean: On Stranger Tides and Pirates of the Caribbean: At World's End, which made from more than 0.3billion budget. Avengers: Age of Ultron movie was made from 0.28billion budget. Tangled and John Carter had made from a similar budget which was 0.26billion. Over 0.25billion budget Spider-Man 3, The Lone Ranger, and The Hobbit: An Unexpected Journey movie were made.

3. Which top 10 Movie has highest Profit?

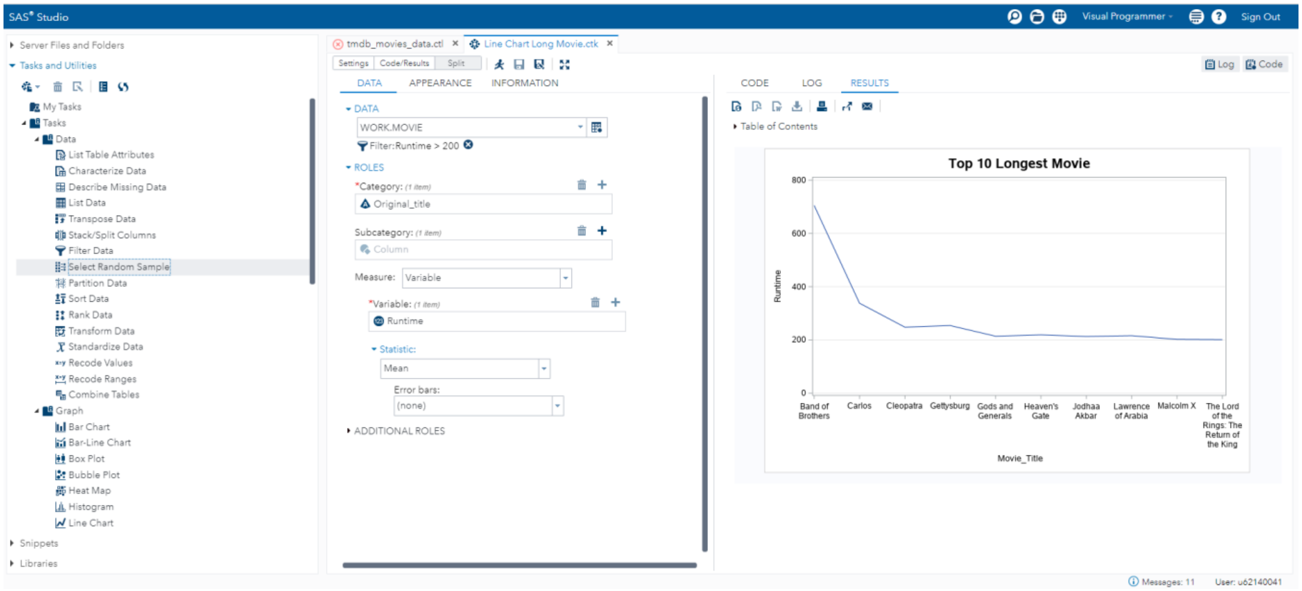


☐ Show labels
☑ Set color
Color transparency: 0% (default)
Units: Inches (default)
*Width: 6.7
*Height: 5.4



The above bar graph illustrates top 10 highest profitable movie with different release year. The above data show in Billion. Graph clearly shows Avatar movie was released in 2009 and it was most profitable movie between 1960 and 2015, it generated 2.7 billion profits. In 2015, Star Wars: The Force Awakens movie was released which profit was around 2.0 billion. Other movie like Jurassic World, Furious 7 and Avengers: Age of Ultron which released in same year and generated more than 1.4 billion profits. Iron Man 3 and Frozen was released in year 2013 which made approximately 1.2 billion profits. In the year 2012, The Avengers was released its generated 1.5 billion profit. Harry Potter and the Deathly Hallows: Part 2 was released in year 2011 which was so popular, it was become highest profit producible movie. Titanic was released in 1997 which generated 1.8billion profit until 2015.

4. **Which movie has longest runtime**?



DATA

WORK.MOVIE

Filter:Runtime > 200

ROLES

*Category: (1 item)
Original_title

Subcategory: (1 item)
Column

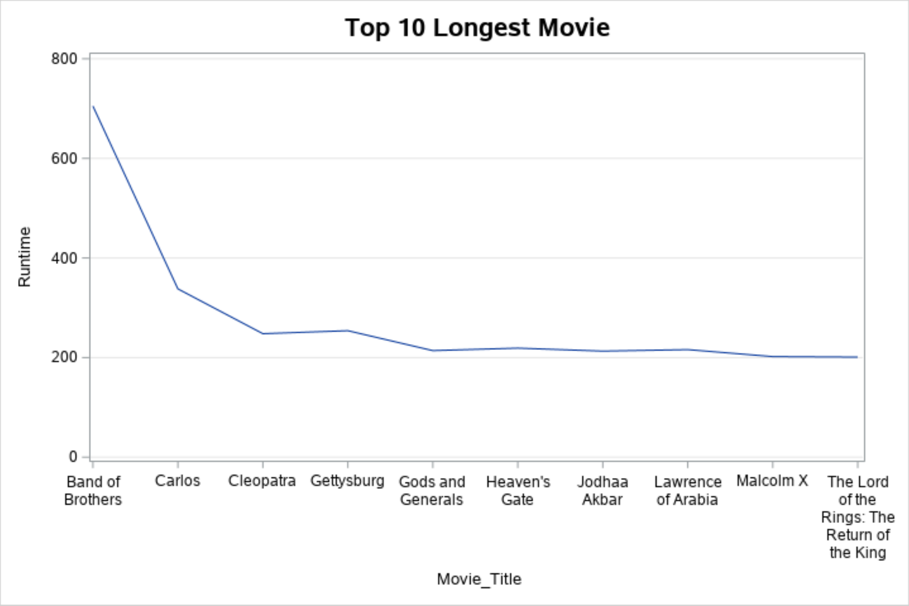
Measure: Variable

*Variable: (1 item)
Runtime

Statistic:
Mean

Error bars:
(none)

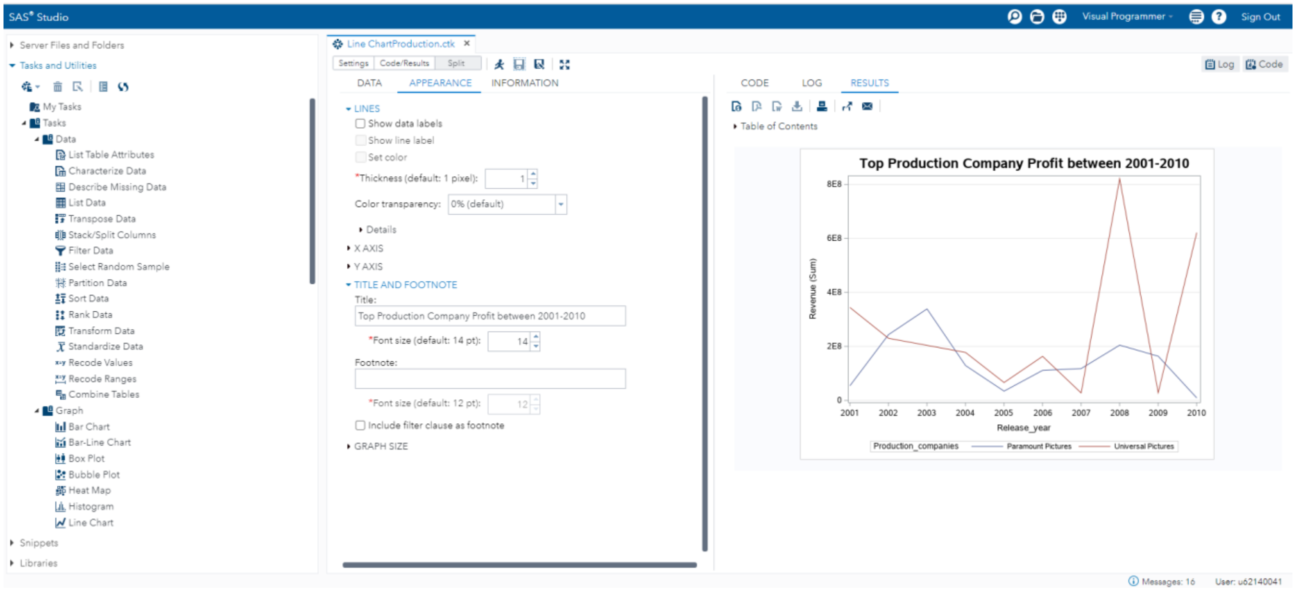
ADDITIONAL ROLES



The above line chart shows longest movie between year 1960 to 2015. The above data show as minutes. The band of brothers movie, which is based on documentary, its longest movie in 55 years.

The second longest movie is Carlos, which is 338mins movie. Gettysburg and Cleopatra are more than 245min movie. The chart shows Heaven's Gate movie is slightly longer than Gods and Generals movie. The movie Jodhaa Akbar is around 213min movie. Malcolm X movie is a bit shorter compared to Lawrence of Arabia. The Lord of the Rings: The Return of the King is approximately 200min movie.

5. Which production company has higher profit between 2001 to 2010?



Line ChartProduction.ctk

DATA | APPEARANCE | INFORMATION

LINES

□ Show data labels

□ Show line label

□ Set color

*Thickness (default: 1 pixel): 1

Color transparency: 0% (default)

Details

X AXIS

Y AXIS

TITLE AND FOOTNOTE

Title: Top Production Company Profit between 2001-2010

*Font size (default: 14 pt): 14

Footnote:

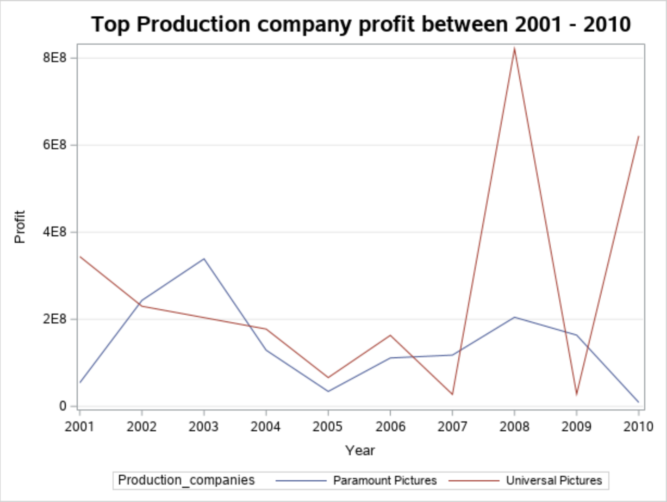
*Font size (default: 12 pt): 12

□ Include filter clause as footnote

GRAPH SIZE

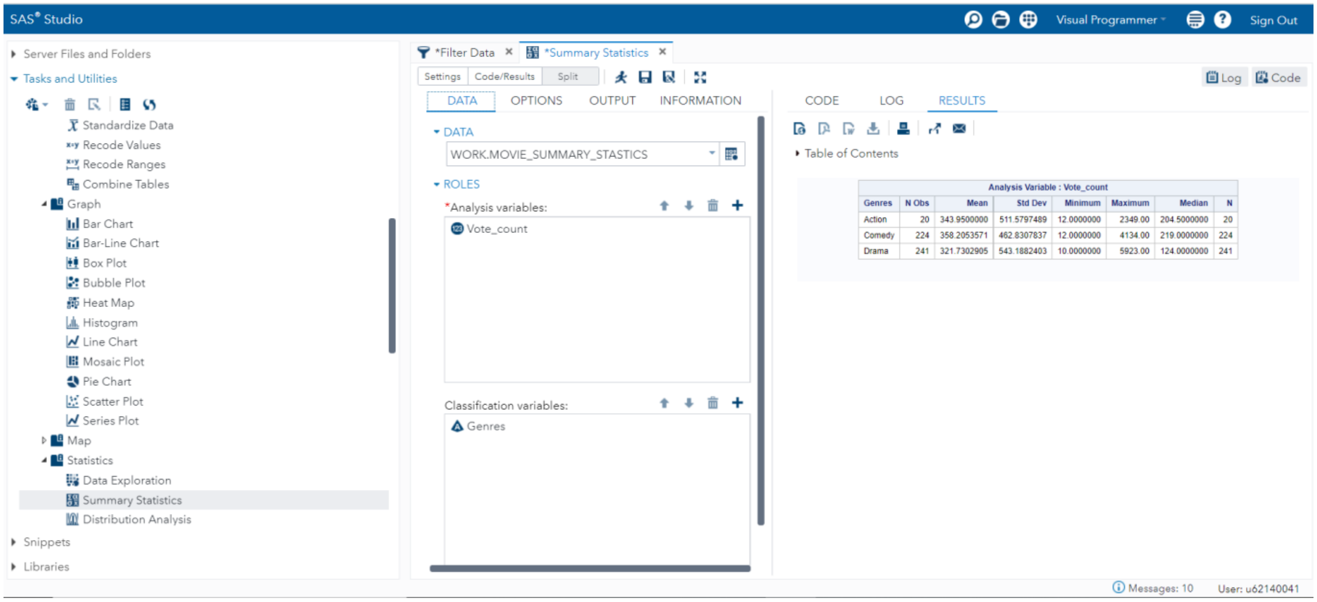
Messages: 16

User: u62140041



The above line chart represents top production companies year wise total profit in one decade. Data shows as billion. Overall, Universal pictures company has highest profit in year 2008 and Paramount pictures company has highest profit in year 2003. In the beginning of 2001 year both production company total profit was approximately 0.05 and 0.38 billion. In a year 2002 and 2003, Paramount picture company profit was increased, and Universal pictures profit was slightly decreased. Between 2004 to 2006 both production company profit was lesser than 0.2 billion. After 2007, Universal pictures profit was increased and became high in year 2008 and it suddenly decreased in 2009. Paramount Pictures production company profit was gradually shrunk after 2008 and became low at the end of year 2010, which was around 9.0million. In the year 2010, Universal pictures profit was increased about 0.6 billion.

Statical Summary



DATA

WORK.MOVIE_SUMMARY_STASTICS

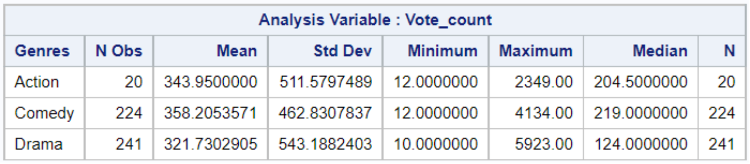
ROLES

*Analysis variables: Vote_count

Classification variables: Genres

Messages: 10

User: u62140041

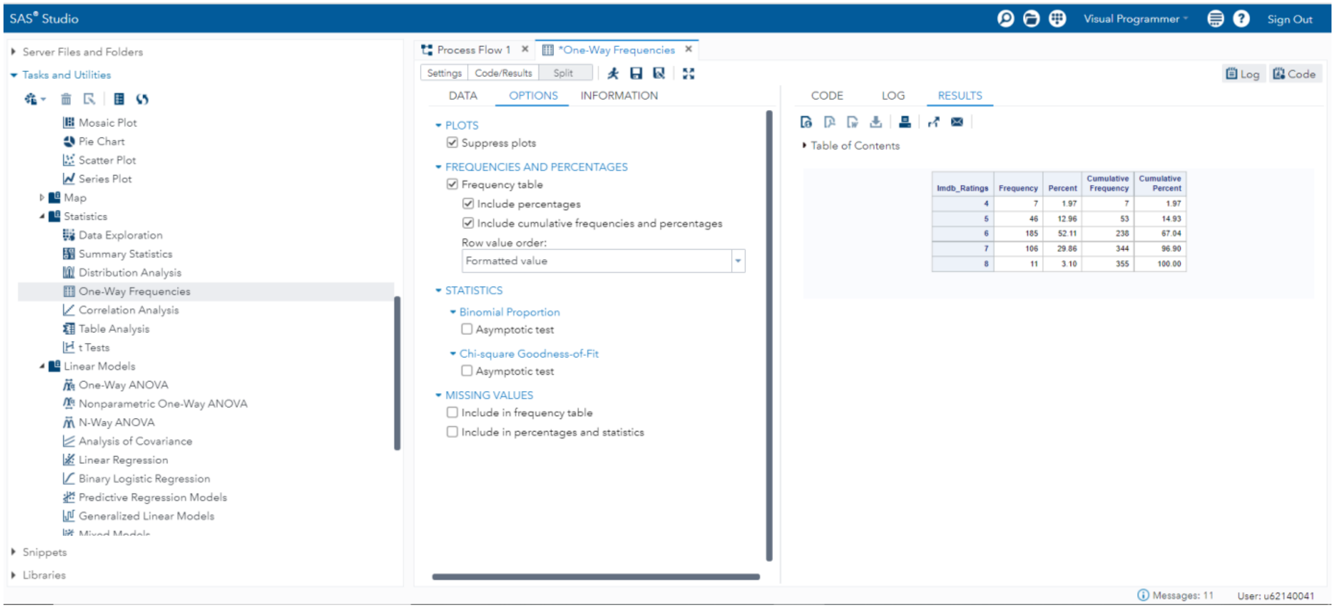


Analysis Variable : Vote_count							
Genres	N Obs	Mean	Std Dev	Minimum	Maximum	Median	N
Action	20	343.9500000	511.5797489	12.0000000	2349.00	204.5000000	20
Comedy	224	358.2053571	462.8307837	12.0000000	4134.00	219.0000000	224
Drama	241	321.7302905	543.1882403	10.0000000	5923.00	124.0000000	241

The above summary statistics summarize and provide information about most popular genres have highest vote count for a total of 485 observations. The average of vote count provides in mean column. It’s shows comedy genre is so popular because it has highest Mean which observed 358.20. Also, Drama genre has 321.73 Mean which lowest but it has highest standard deviation which is 543.18 among most popular genres observation. The std. dev describes amount of variation. The Drama genre has lowest minimum vote and highest maximum vote value. The middle number in sequence number of vote count shows in median column. The median of Action movie is 204.50 which is average among other 2.

Statical Tests

- **One-Way Frequency**



PLOTS

☑ Suppress plots

FREQUENCIES AND PERCENTAGES

☑ Frequency table

☑ Include percentages

☑ Include cumulative frequencies and percentages

Row value order: Formatted value

STATISTICS

Binomial Proportion

☐ Asymptotic test

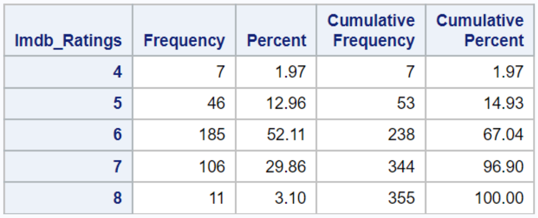
Chi-square Goodness-of-Fit

☐ Asymptotic test

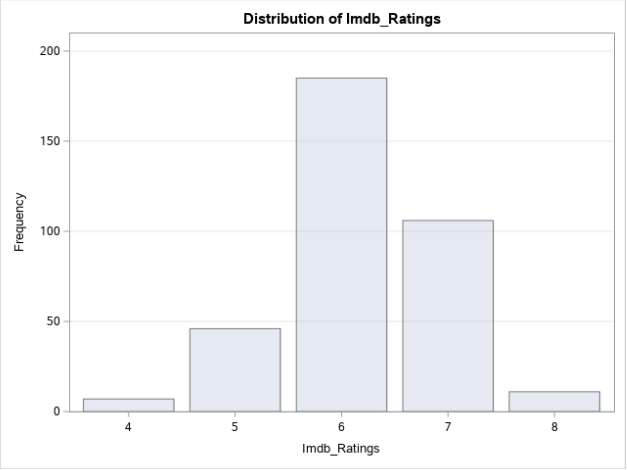
MISSING VALUES

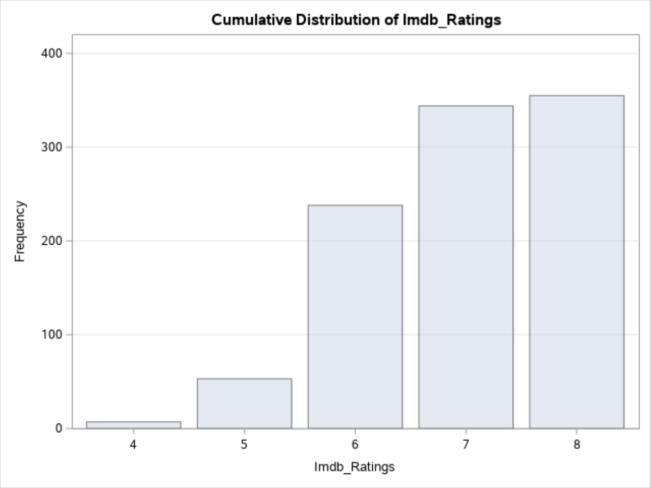
☐ Include in frequency table

☐ Include in percentages and statistics



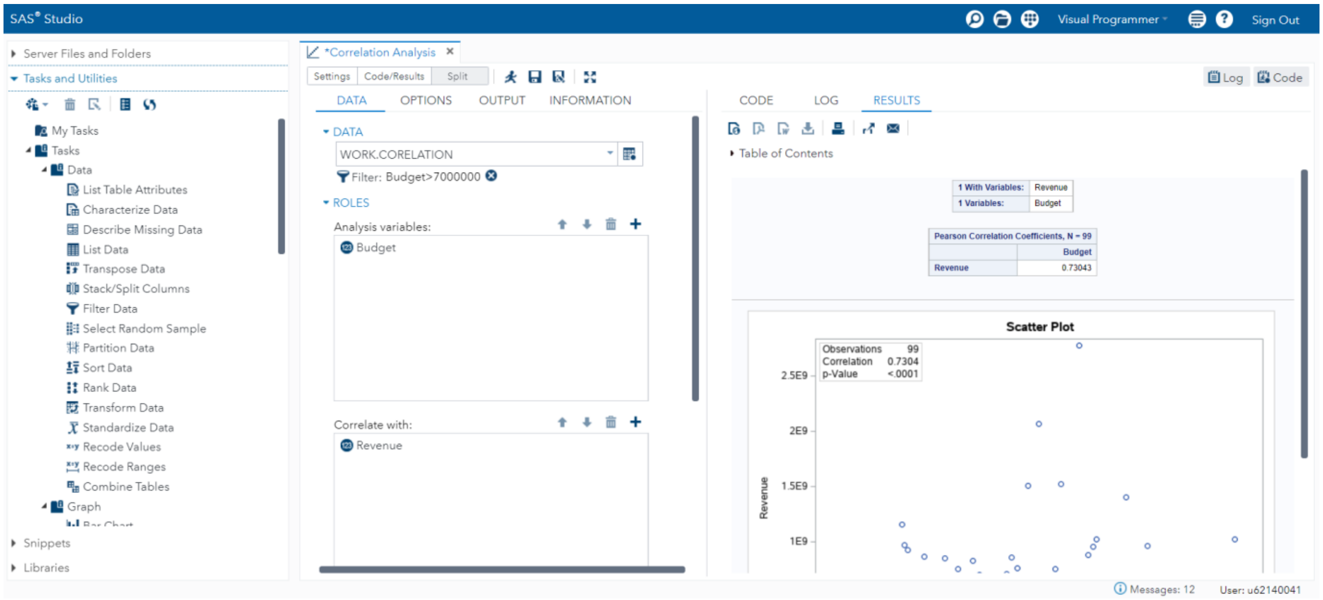
Imdb_Ratings	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	7	1.97	7	1.97
5	46	12.96	53	14.93
6	185	52.11	238	67.04
7	106	29.86	344	96.90
8	11	3.10	355	100.00





The above One-way frequency table represents the frequency, percent, cumulative frequency, and cumulative percent of the variable Imdb Ratings. There are five different movie ratings listed in the table. By looking at the table most of movies got 6 ratings with frequency of 185 which makes up 52.11% of the distribution. On the other hand, the least of movie got 4 rating with frequency of 7 which is 1.97% of distribution. Also, 4 Imdb rating has low cumulative distribution. As mentioned in the table cumulative distribution of 8 rating is very high which is 100% for 355 cumulative frequencies.

- **Correlation**



DATA

WORK.CORELATION

Filter: Budget>7000000

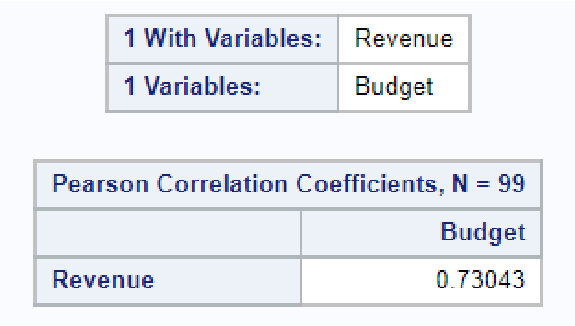
ROLES

Analysis variables:

Budget

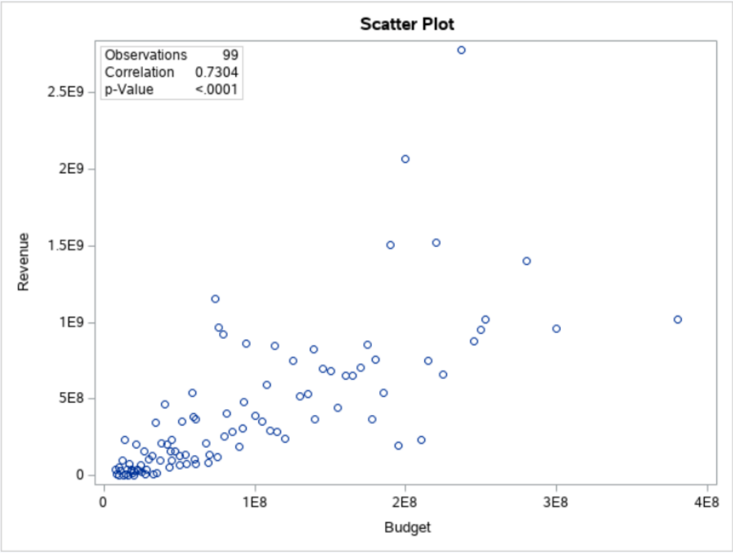
Correlate with:

Revenue



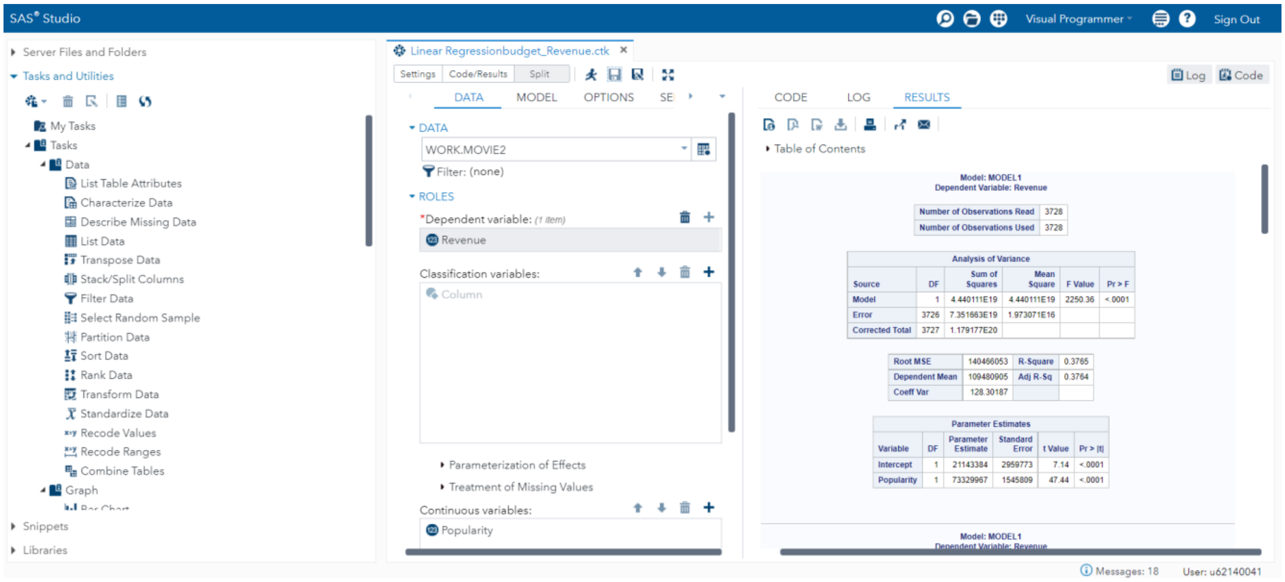
1 With Variables:	Revenue
1 Variables:	Budget

Pearson Correlation Coefficients, N = 99	
	Budget
Revenue	0.73043



The above scatter plot indicates correlation between Budget and revenue. The analysis shows analysis variable Budget is independent variable and correlate variable Revenue is dependent variable. Pearson's coefficient of correlation is defined as a linear correlation coefficient that falls in the value range of -1 to +1. Value of -1 signifies strong negative correlation while +1 indicates strong positive correlation. Observing the scatter plot, one can determine there are 99 observations, and the correlation value is 0.73% which indicate positive relation between them. So, here is a good possibility that movies with higher investments result in better revenues.

- **Linear Regression**



DATA: WORK.MOVIE2

Filter: (none)

ROLES

*Dependent variable: (1 item) Revenue

Classification variables: Column

Continuous variables: Popularity

Model: MODEL1
Dependent Variable: Revenue

Number of Observations Read	3728
Number of Observations Used	3728

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.440111E19	4.440111E19	2250.36	<.0001
Error	3726	7.351663E19	1.973071E16		
Corrected Total	3727	1.179177E20			

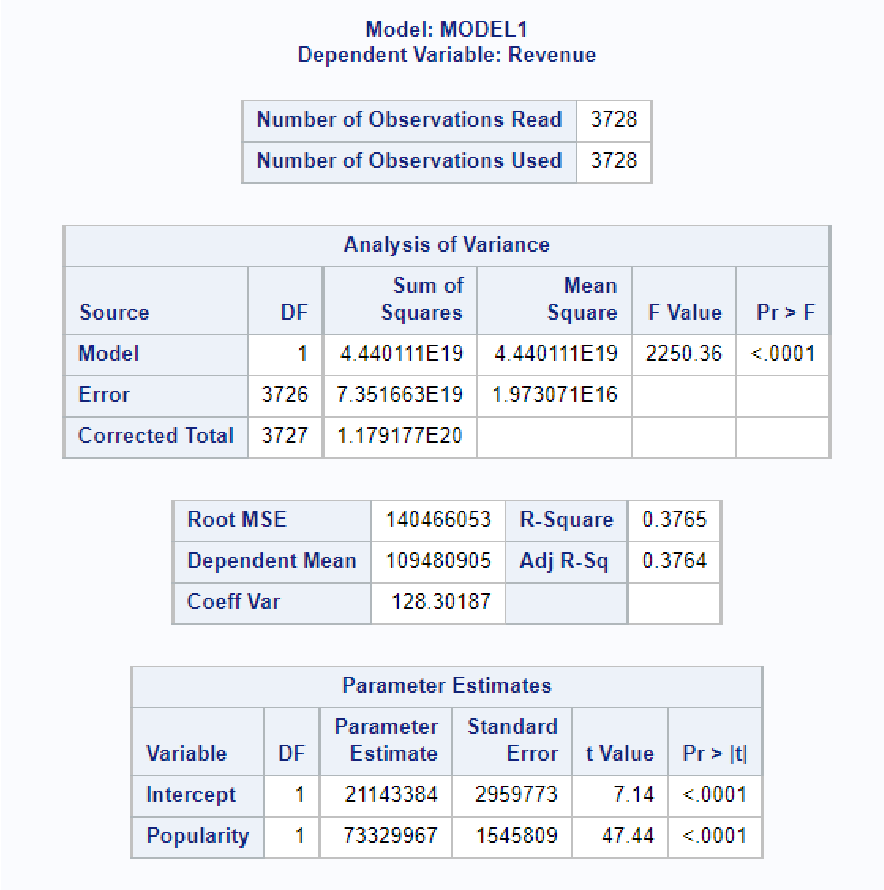
Root MSE	140466053	R-Square	0.3765
Dependent Mean	109480905	Adj R-Sq	0.3764
Coeff Var	128.30187		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21143384	2959773	7.14	<.0001
Popularity	1	73329967	1545809	47.44	<.0001

Model: MODEL1
Dependent Variable: Revenue

Messages: 18   User: u62140041



Model: MODEL1
Dependent Variable: Revenue

Number of Observations Read	3728
Number of Observations Used	3728

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.440111E19	4.440111E19	2250.36	<.0001
Error	3726	7.351663E19	1.973071E16		
Corrected Total	3727	1.179177E20			

Root MSE	140466053	R-Square	0.3765
Dependent Mean	109480905	Adj R-Sq	0.3764
Coeff Var	128.30187		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21143384	2959773	7.14	<.0001
Popularity	1	73329967	1545809	47.44	<.0001

The above linear model illustrates the relation between Popularity and Revenue variable co-efficiency significant or not. The popularity is independent variable and revenue is dependent variable. R-squared is a primary measure of how well a regression model fits the data. This statistic represents the percentage of variation in one variable those other variables explain. The linear regression formula is y = ax + b where a is slope and b is intercept.

The table show R^2 value is “0.37 (37%)” which represent there is some correlation between popularity and revenue. There are 3727 observations and there are no missing observations find. Here, assumption value p = 0.05 it is greater than 0.0001 which shows it is statically significant. This mean that any variation in the Revenue can be explained Popularity variables.

References

1. Tan, E.S. (2018, July 03). Palgrave Communication. A psychology of the film. https://doi.org/10.1057/s41599-018-0111-y
2. Anjana Vegaraju (2020, November 13) Exploratory Data Analysis on TMDB Dataset. https://medium.com/@anjana.vegaraju/exploratory-data-analysis-on-tmdb-dataset-b2c99aadf10e