

Final case Analysis

Shruti Daund

Applied Business statistics 2

Objectives of the case study are as follows

- 1) Data cleaning and filtering, Descriptive data analysis and visualization
- 2) Regression model building for both continuous and binary response variables
- 3) To test ability to make predictions on unseen data.

The final case analysis is based on the movie industry dataset with 23 variables. Along with these 23 variables the success or failure of a movie depends on some other factors also like Prime time, Critics, Controversy, advertisements, the popularity of actors, etc. Controversial movies are either stalled or when released make a huge fortune at box office. The prime time given to movies also affects the overall success of a movie. The advertisements before the release hold great importance for success. Also the time period in which the movie was released matters a lot for example movies released during war time or curfews or national emergency don't make good at box office. Initial days of filmmaking that is 1890's these factors mentioned above were not very relevant for the success of films made, because humans were experiencing a new technology and many were not even aware of its existence.

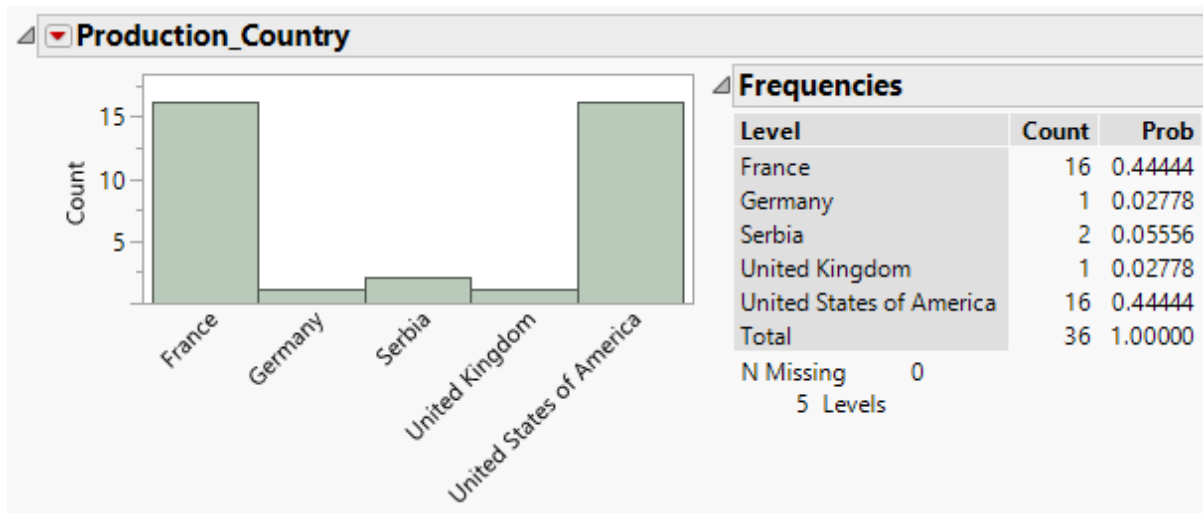
The first movie ever made in human history was just 2.11 seconds long in 1888. The dataset includes data about movie industry more than a century starting from 1900 to 2019. Starting from very few movies in early 1900's to hundreds and thousands of movies releasing every month all over the world, the movie industry has truly developed with the inception of colour movies in 1918. As of year 2018 the global box office was worth \$41.7 billion. According to data provided by statista for the box office revenue from 1980's to 2018 for North America the average ticket price in 2018 was 9.14 that is 2 dollar increase from 2008. Warner bros, Walt Disney, Marvel studios have always been the cherry on the cake among all other production houses due to the movies they have given to the world.

In the last five decades the major change movie industry has gone through is advertisements. Due to availability of a huge source of platform for advertising

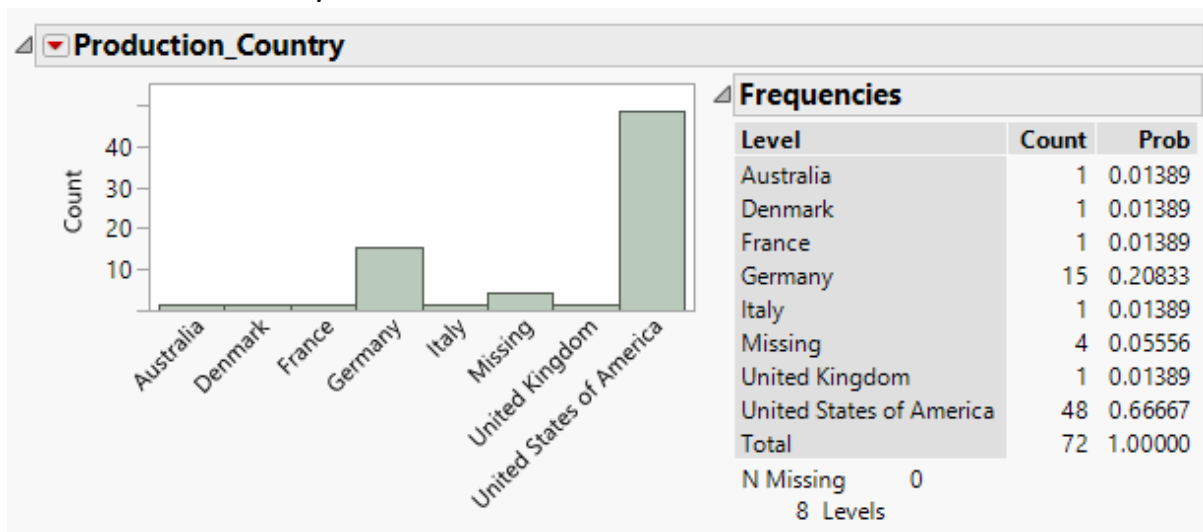
films before they release the income generated has also increased. Sometimes even though the movies are a flop the songs released making a good fortune. In early 1970 more importance was given to the story and plot, the success of a movie was dependent on this factor. But the scenario has changed in last 25 years where advertisements, cast, high budget sets matters the most.

Data cleaning and descriptive analysis

The original dataset consists of 34,996 values with 23 variables. Spoken language which had less than 25 movies released was removed and there were some missing values in the dataset were also removed. The clean dataset now consists of 29,204 values.

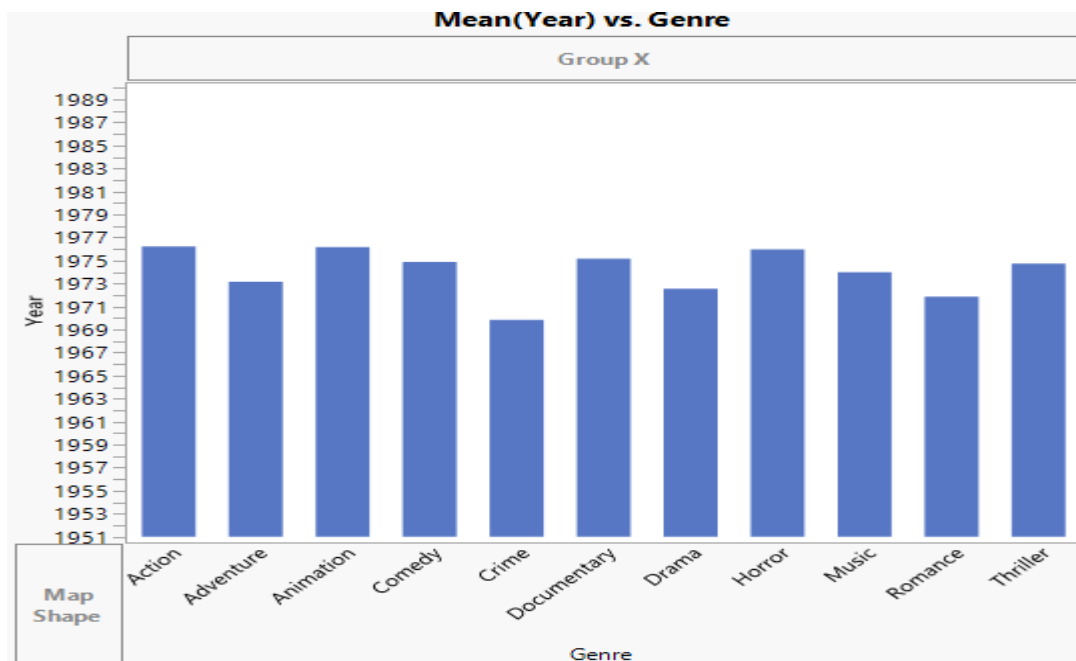


In the era of silent films from 1900 to 1913 the United States and France were on top in releasing movies, United Kingdom, Siberia, Germany also contributed in the movie industry.

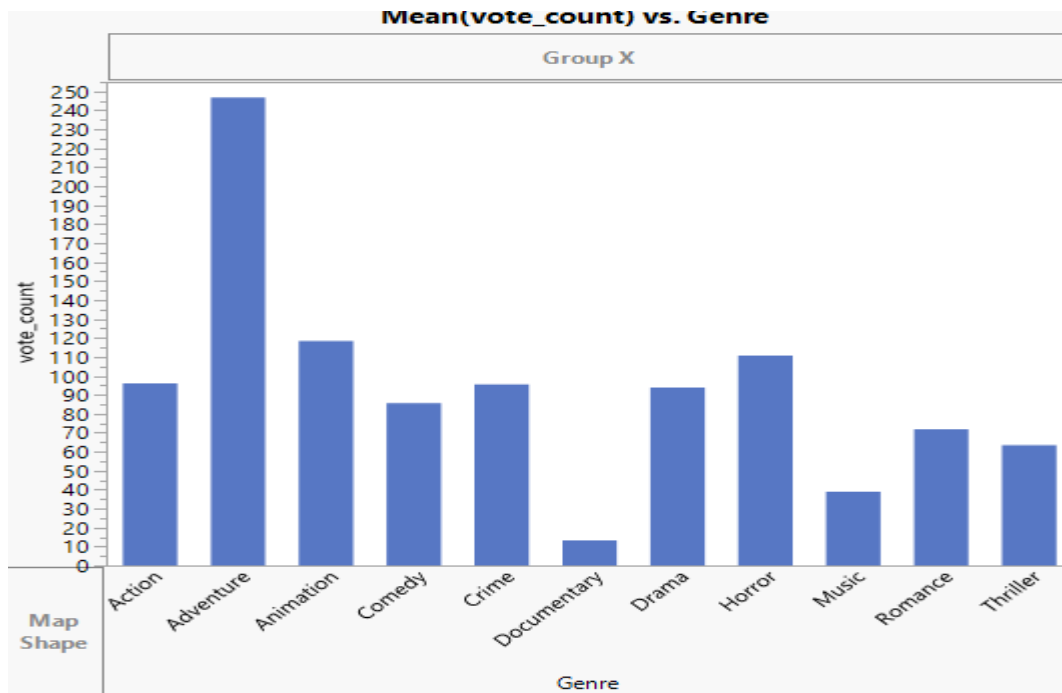


During the World War 1 period The United States which decided to keep itself neutral from the war ended up making more number of films than any other country, followed by Germany. The next 20 years after World War 1 saw a boom in the movie industry. Along with United States and European countries the Asians also entered into filmmaking. Countries like India, Japan, China were on the forefront.

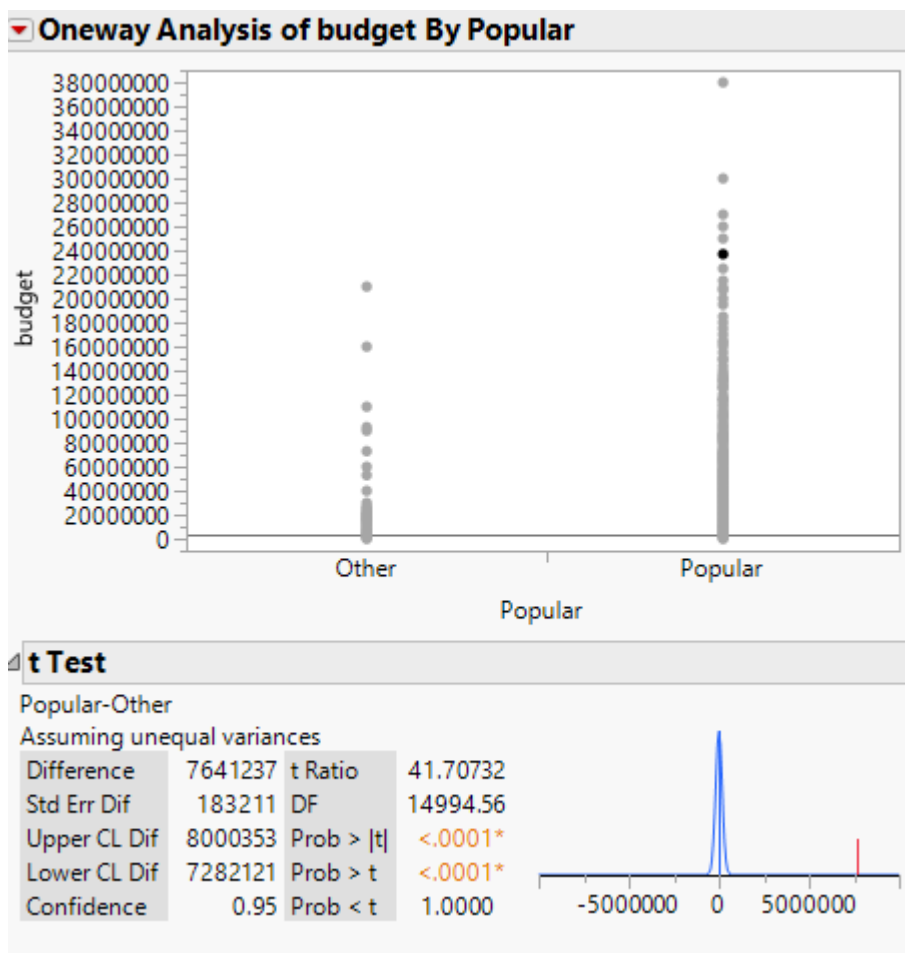
The next 10 years from 1940 to 1950 during the World War 2 and Post World War2 period the movie industry was not affected except the year 1944 when the War ended saw a decline in the number of movies released in that year. Again during these 10 years The United States was on the top with 854 movies released in that decade.



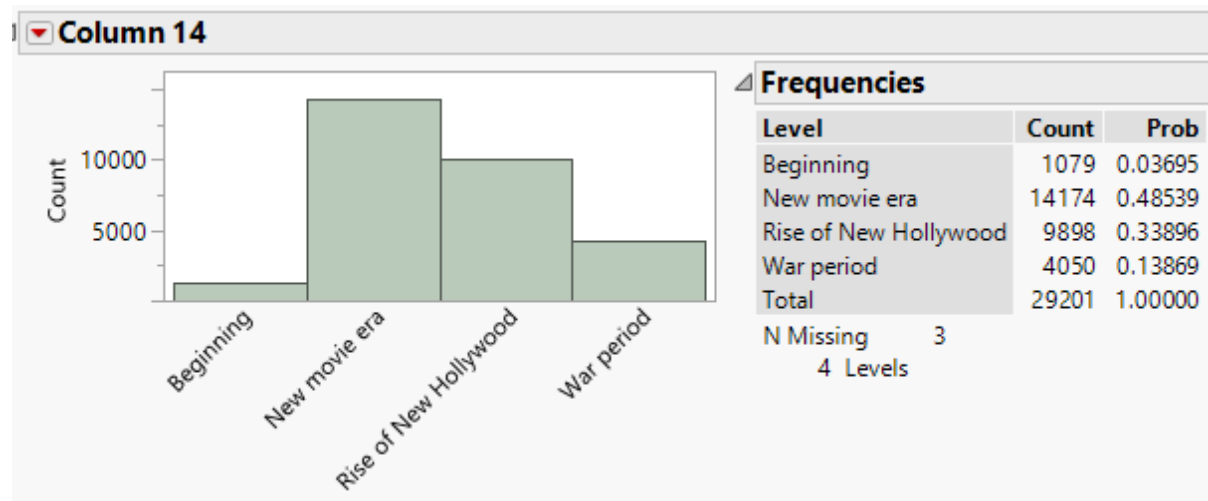
In the Cold War period from 1950 to 1989 with creative directors like Alfred Hitchcock, Sergio Leone and others the movie industry saw a rapid growth in filmmaking as with growing technology and its availability to a large population. More movies were released of the genre Action, animation, horror.



The Adventure movies were applauded more by the people in this period followed by animation, action and horror.



The above test shows that higher the budget higher are the chances of the movie becoming popular.



Movie data for more than a century is present so I have categorized it into 5 time periods:

- ❖ 1900 – 1940 : Beginning
- ❖ 1941 to 1970 : war period
- ❖ 1971 to 1999 : rise of new Hollywood
- ❖ 2000 to 2019 : new movie era

Contingency Table

		Genre										
		Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Horror	Music	Romance	Thriller
Count												
Total %												
Col %												
Row %												
Column 14	Beginning	66	46	58	363	96	19	316	35	25	43	12
		0.23	0.16	0.20	1.24	0.33	0.07	1.08	0.12	0.09	0.15	0.04
		1.85	4.03	5.23	5.02	6.51	0.94	4.14	1.49	2.66	5.91	1.19
		6.12	4.26	5.38	33.64	8.90	1.76	29.29	3.24	2.32	3.99	1.11
	New movie era	1621	433	541	3397	470	1489	3681	1076	583	339	544
		5.55	1.48	1.85	11.63	1.61	5.10	12.61	3.68	2.00	1.16	1.86
		45.37	37.92	48.78	46.96	31.86	73.90	48.28	45.75	61.96	46.63	53.97
		11.44	3.05	3.82	23.97	3.32	10.51	25.97	7.59	4.11	2.39	3.84
	Rise of New Hollywood	1515	420	378	2533	451	418	2424	958	238	228	335
		5.19	1.44	1.29	8.67	1.54	1.43	8.30	3.28	0.82	0.78	1.15
		42.40	36.78	34.08	35.02	30.58	20.74	31.79	40.73	25.29	31.36	33.23
		15.31	4.24	3.82	25.59	4.56	4.22	24.49	9.68	2.40	2.30	3.38
	War period	371	243	132	941	458	89	1204	283	95	117	117
		1.27	0.83	0.45	3.22	1.57	0.30	4.12	0.97	0.33	0.40	0.40
		10.38	21.28	11.90	13.01	31.05	4.42	15.79	12.03	10.10	16.09	11.61
		9.16	6.00	3.26	23.23	11.31	2.20	29.73	6.99	2.35	2.89	2.89
	Total	3573	1142	1109	7234	1475	2015	7625	2352	941	727	1008
		12.24	3.91	3.80	24.77	5.05	6.90	26.11	8.05	3.22	2.49	3.45

As we see from above table the beginning of the movie era comedy movies were released more in number followed by Drama. The following era of World War 2 saw more number of drama movies been released. The new Hollywood

released more comedy movies. Also the new movie era released more drama followed by comedy. From the above observations we can conclude that comedy movies are released more in number.

Regression

You are expected to create two kinds of regression models: MLR and Logistic regression.

- For MLR the response variables are **Revenue** and **Likeability**. Create two separate models for these responses and compare models in terms of their *predictor variables* and their *ability to predict*. Describe in your own words important details of the above two models

Response revenue					
Summary of Fit					
RSquare			0.731506		
RSquare Adj			0.731368		
Root Mean Square Error			28436986		
Mean of Response			10049964		
Observations (or Sum Wgts)			29201		
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	15	6.43e+19	4.287e+18	5300.926	
Error	29185	2.3601e+19	8.087e+14		Prob > F
C. Total	29200	8.7901e+19			<.0001*
Parameter Estimates					
Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
vote_count	1	1	1.5365e+19	19000.20	<.0001*
Column 14	3	3	3.1675e+16	13.0563	<.0001*
budget	1	1	8.2988e+18	10262.42	<.0001*
Genre	10	10	4.9551e+16	6.1276	<.0001*
Effect Details					

The Rsquare is 0.731506 with variables vote_count, column 14, budget, Genre which gave the new column for prediction formula for revenue.

Response Likeability					
Summary of Fit					
RSquare			0.359148		
RSquare Adj			0.358621		
Root Mean Square Error			3.327482		
Mean of Response			3.1963		
Observations (or Sum Wgts)			29204		
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	24	181057.91	7544.08	681.3571	
Error	29179	323073.92	11.07		Prob > F
C. Total	29203	504131.83			<.0001*
Parameter Estimates					
Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
vote_average	1	1	8478.860	765.7834	<.0001*
Genre	10	10	9473.042	85.5575	<.0001*
vote_count	1	1	30500.613	2754.717	<.0001*
Month	11	11	2952.976	24.2458	<.0001*
budget	1	1	20381.391	1840.782	<.0001*
Effect Details					

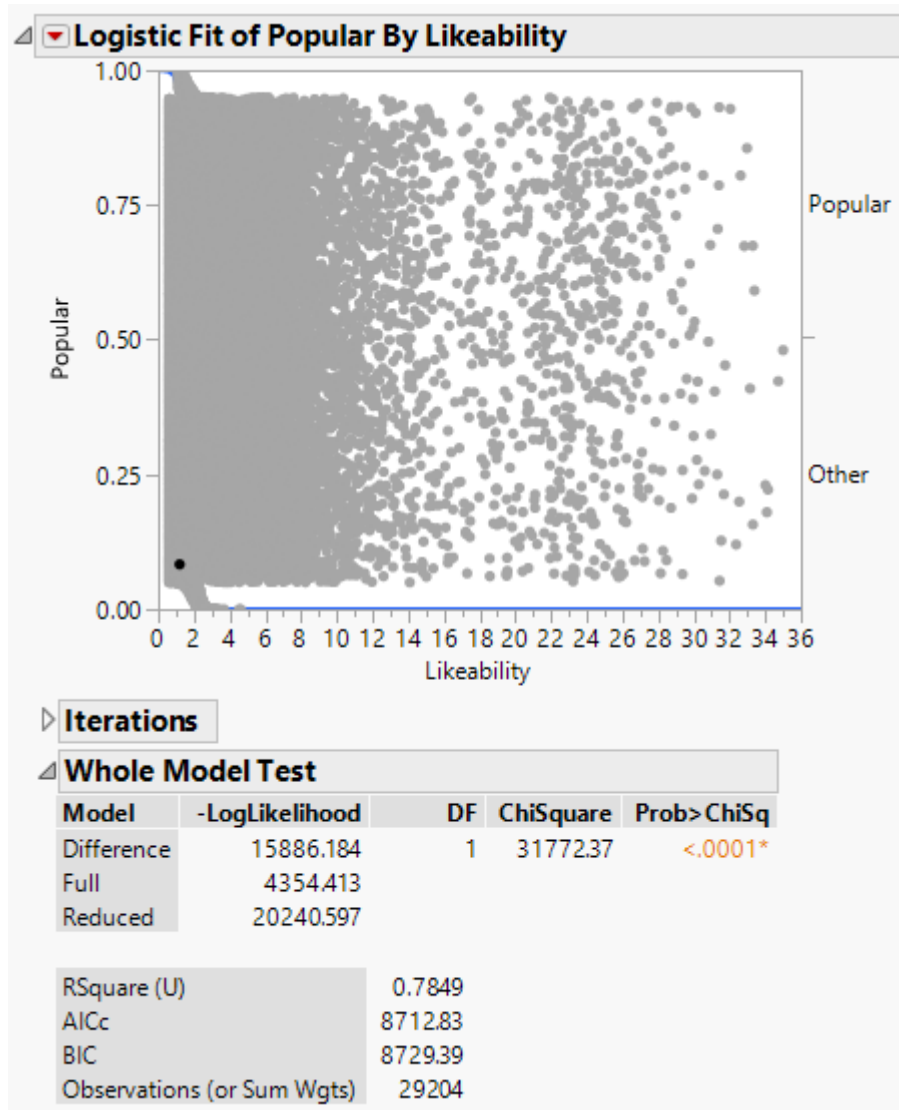
The Rsquare for likeability is 0.359148 with variables vote_average, genre, vote_count, month and budget. The F-ratio is 681.3571. A new column for prediction of likeability was created.

The 100 values from the excel file were also added before making the predictions.

Predictions

Spoken_Lai	Production	Genre	adult	budget	original_lar	vote_aver	vote_count	release_da	Year	Month	Day	runtime	tagline	video	ID	pred formula	likeability	pred formula
Deutsch	United Stat	Metropolit	Music	FALSE	0	en	8	2	#####	2013	3	2	280	TRUE	1	2.514556		18302479
???	Japan	Square Eni	Action	FALSE	0	ja	6.8	450	7/14/2005	2005	7	14	101 Is it for the	FALSE	2	21.8935		3.87E+08
English	United Stat	Imagine En	Thriller	FALSE	1.25E+08	en	6.6	4973	5/17/2006	2006	5	17	149 Seek the tr	FALSE	3	2.921375		775146.8
English	United Stat	Double Hel	Horror	FALSE	0	en	5.8	93	10/31/1981	1988	10	31	80 When you i	FALSE	4	3.888535		776471.4
Español	United Stat	MLG Produ	Adventure	FALSE	0	en	6.4	119	#####	2006	8	8	73 To save hui	FALSE	5	1.767313		-2335046
English	United Stat	Incendiary	Horror	FALSE	0	en	3.8	2	#####	2009	1	1	90 You can't e	FALSE	6	5.288612		-4181428
English	United Stat	Adam & Ev	Adventure	TRUE	150000	en	10	1	#####	2009	9	8	265 In a lawless	FALSE	7	2.443761		-2072340
????????	Israel	United Cha	Missing	FALSE	0	en	5.4	8	1/22/2015	2015	1	22	92	FALSE	8			
English	United Stat	Missing	Drama	FALSE	0	en	6.1	91	1/23/2006	2006	1	23	96	FALSE	9	2.581089		2838778
English	United Stat	Universal H	Action	FALSE	3800000	en	5	37	#####	2002	4	9	99	FALSE	10	2.765911		-1723509
English	France	ARTE Franc	Romance	FALSE	0	en	6.3	57	1/20/2001	2001	1	20	119 Every Wed	FALSE	11	2.886681		1413524
English	United Stat	David Fost	Mystery	FALSE	0	en	6.5	31	2/15/1985	1985	2	15	103 A time bet	FALSE	12			
English	United Stat	NBC Studio	Drama	FALSE	0	en	6.6	7	1/16/2000	2000	1	16	89 Live Speller	FALSE	13	4.11953		-2015160
Missing	Missing	Missing	Missing	FALSE	0	en	9	1	#####	1990	9	2	100	FALSE	14			
???	Japan	Production	Animation	FALSE	3200000	ja	7.8	108	2/23/2007	2007	2	23	105	FALSE	15	3.51843		264448.6
English	United Stat	Missing	Document	FALSE	0	en	10	1	#####	1997	1	1	60	FALSE	16	3.466237		32224691
English	United Stat	RKO Radio	Thriller	FALSE	2000000	en	7.9	639	8/15/1946	1946	8	15	103 Notorious	FALSE	17	2.712372		-5217318
Missing	Missing	Missing	Missing	FALSE	0	en	6.5	2	11/13/2001	2007	11	13	80	FALSE	18			
English	United Kin	Missing	Comedy	FALSE	0	en	6.5	10	#####	1974	1	1	90 Serviced wi	FALSE	19	3.360671		10860090
English	Australia	Walt Disne	Animation	FALSE	0	en	6.1	263	2/29/2000	2000	2	29	79	FALSE	20	1.353742		-1699860
English	United Stat	MoIAM Eni	Comedy	FALSE	0	en	2.5	2	7/15/2006	2006	7	15	100	FALSE	21	3.419134		-2575709
Français	France	Zadig Prodi	Document	FALSE	0	en	7	2	9/19/2006	2006	9	19	52	FALSE	22	2.601591		11228418
Italiano	Italy	Vera Films	Drama	FALSE	0	it	8.3	223	#####	1970	2	9	115 When you'l	FALSE	23	1.053296		-2015160
Dansk	Denmark	Egmont Fili	Comedy	FALSE	0	da	3	1	11/24/1981	1989	11	24	80	FALSE	24	2.313573		157376.3
English	United Stat	Vernon-Ser	Horror	FALSE	60000	en	4.3	8	9/22/1965	1965	9	22	79	FALSE	25	3.477871		-3864666
???	Japan	Total Medi	Action	FALSE	0	ja	6.3	13	5/15/2010	2010	5	15	73 A bloody A	FALSE	26	3.38492		-2895241
English	United Stat	Missing	Comedy	FALSE	700000	en	6	8	#####	2004	2	6	90	FALSE	27	2.735023		877266
English	France	Medusa Pr	Action	FALSE	0	en	5.3	35	7/22/1983	1983	7	22	96 In The Year	FALSE	28	4.534176		18763288
English	United Stat	Jacmac Fil	Thriller	FALSE	8500000	en	6.6	209	#####	1991	3	8	97 They're a n	FALSE	29	3.012751		-2727598
English	United Stat	Full Body	P Drama	FALSE	0	en	6.4	13	#####	1995	10	1	93 Feelings an	FALSE	30	1.052829		-3979468
Missing	Missing	Missing	Music	FALSE	0	en	4	2	#####	2010	1	1	90 The "Foot"	FALSE	31	0.691274		-1356678
English	Canada	Pope Prodi	Drama	FALSE	0	en	5	1	#####	2009	3	8	89	FALSE	32	4.888323		33130279
Deutsch	Australia	Wim Wend	Thriller	FALSE	23000000	de	6.7	73	#####	1991	9	12	280 ... the ulti	FALSE	33	2.337299		-5042181
English	United Stat	Buffalo Spe	Comedy	FALSE	0	en	4.3	6	7/24/2009	2009	7	24	107 It takes a vi	FALSE	34	2.590959		424190.9
English	United Stat	Four-Leaf P	Comedy	FALSE	0	en	5.5	16	#####	1964	3	6	96 Elvis is bac	FALSE	35	20.17921		3.63E+08
English	United Stat	Golden Me	Adventure	FALSE	1.65E+08	en	6.6	3119	#####	2004	11	10	100 This holida	FALSE	36	3.927759		-2378831
English	United Stat	Talking Mo	Comedy	FALSE	0	en	8	1	11/21/2011	2014	11	21	63	FALSE	37	3.42746		-348695
???	Hong Kong	Shaw Broth	Action	FALSE	0	zh	7.1	7	7/27/1972	1972	7	27	123	FALSE	38	3.776942		-1788511
Missing	Missing	Missing	Missing	FALSE	0	en	9	1	#####	1962	1	1		FALSE	39			
No Language	Germany	Hochschule	Drama	FALSE	0	de	7.9	46	5/25/1989	1989	5	25	7	FALSE	40	2.0972		-1461196
Česky	Czech Repu	Missing	Animation	FALSE	0	cs	6.5	5	#####	1954	1	1	12	FALSE	41	2.227018		-1612291
Eesti	Estonia	Taska Film	Comedy	FALSE	0	et	5.1	4	#####	2006	10	6	90	FALSE	42	4.463689		-2619493
Missing	Missing	Missing	Missing	TRUE	0	en	10	1	2/18/2001	2001	2	18	200	FALSE	43			
Italiano	Italy	Transglobe	Comedy	FALSE	0	en	4.3	3	#####	1972	1	1	83	FALSE	44	2.757794		182703
??????	India	Bollywood	Drama	FALSE	0	hi	5.8	65	5/21/2010	2010	5	21	123	FALSE	45	2.38995		-1746468
Deutsch	Australia	A-Mark Ent	Horror	FALSE	0	en	5.2	53	#####	2009	10	1	86 Lead us noi	FALSE	46	3.755382		-4258725
???	Japan	J Storm	Drama	FALSE	0	ja	6.8	4	2/14/2015	2015	2	14	103	FALSE	47	2.375618		486532.1
English	United Kin	Feelgood F	Drama	FALSE	0	en	6.7	104	1/20/2015	2015	1	20	90	FALSE	48	2.067866		-2709723
Italiano	Italy	Missing	Missing	FALSE	0	it	5.3	31	11/23/2001	2007	11	23		FALSE	49			
English	United Stat	Sony Pictur	Drama	FALSE	0	en	6.8	34	#####	2010	5	9	87	FALSE	50	3.089155		-2015160
Magyar	Missing	Missing	Animation	FALSE	0	hu	7	1	2/18/1988	1988	2	18		FALSE	51	-0.12705		-1524722
English	United Stat	Missing	Horror	FALSE	0	en	1.2	6	1/16/2006	2006	1	16	81	FALSE	52	3.229199		-1944142
English	United Stat	Par-Par Pro	Crime	FALSE	0	en	5.5	11	#####	1982	2	5	107 A Controve	FALSE	53	3.133913		-6220579
???	Japan	Nikkatsu Cr	Drama	FALSE	0	ja	6	3	12/25/2001	2004	12	25	94	FALSE	54	1.959871		-3979468
Nederlands	Netherlands	EMI Films	Music	FALSE	0	en	4.3	2	4/28/2006	2006	4	28	139	FALSE	55	1.868868		22173158
English	United Stat	Paramount	Drama	FALSE	15000000	en	5.2	16	#####	1993	2	12	99 Don't get n	FALSE	56	2.303384		-3935684
English	Missing	Missing	Missing	FALSE	0	en	6	3	#####	2010	11	11	121	TRUE	57			
English	India	Mid Day M	Crime	FALSE	700000	hi	7.6	33	8/13/2004	2004	8	13	143 The shocki	FALSE	58	2.740843		-6220579
English	Australia	Missing	Comedy	FALSE	0	en	6.2	3	#####	2004	11	5	90	TRUE	59	2.553951		-523832
English	United Stat	Missing	Document	FALSE	0	en	7.3	3	#####	1981	1	1	76	FALSE	60	1.892332		2121280
Srpski	Serbia	Avala Film	Drama	FALSE	0	sr	7.1	15	9/22/1967	1967	9	22	79	FALSE	61	2.030275		-614062
English	United Stat	Interstate	Horror	FALSE	0	en	5.2	33	#####	1987	11	6	90 Look what'	FALSE	62	4.669802		22430296
English	United Stat	David Fost	Comedy	FALSE	10000000	en	5.8	250	#####	1988	7	6	110 The advent	FALSE	63	2.889438		-304911
??????	India	Trimurti Fil	Crime	FALSE	0	hi	5.8	8	#####	1989	7	7	173 A Volcanic	FALSE	64	2.239946		-4141028
English	United Stat	Metro-Gol	Romance	FALSE	0	en	6	3	1/31/1936	1936	1	31	113 Jeanette M	FALSE	65	2.631257		-1776745
English	United Stat	Big Shoe Pr	Comedy	FALSE	0	en	6.3	4	6/15/2003	2003	6	15	77	FALSE	66	2.959327		73916.36
English	United Kin	Filmways P	Drama	FALSE	0	en	6.1	8	12/21/1961	1969	12	21	117	FALSE	67	2.712379		-4023252
Missing	Missing	Missing	Missing	FALSE	0	it	8	1	#####	2014	1	11	8	FALSE	68			
English	United Stat	Reel Life F	Document	FALSE	25000	en	10	2	#####	2007	4	10	1 The true st	FALSE	69	1.377528		-528345
????/???	South Kore	Jininsa Film	Action	FALSE	0	ko	6.5	9	#####	2004	10	8	112	FALSE	70	3.367846		-3168391
English	Missing	Missing	TV Movie	FALSE	0	en	6.9	24	11/20/2001	2004	11	20	88	FALSE	71			
English	United Stat	Touchstone	Drama	FALSE	16400000	en	8.3	5641	#####	1989	6	2	128 He was the	FALSE	72	1.915477		-3979468
Deutsch	Germany	Missing	Drama	FALSE	0	de	4.3	2	#####	2003	10	1	92	FALSE	73	1.743339		1.93E+08
English	United Kin	Revolution	Adventure	FALSE	1E+08	en	7.1	1335	12/25/2001	2003	12	25	113 All children	FALSE	74	8.564309		1.01E+08
English	United Stat	Mandalay	Action	FALSE	50000000	en	5.8	780	9/30/2005	2005	9	30	110 Treasure hi	FALSE	75	3.969114		203001.2
English	United Kin	BBC Films	Drama	FALSE	0	en	6.6	101	9/25/2009	2009	9	25	108	FALSE	76	2.649811		-3454056
Français	France	Le Pacte	Comedy	FALSE	0	fr	5.4	14	#####	2009	9	2	105	FALSE	77	3.11686		132932.6
English	United Stat	Cineville	Drama	FALSE	0	en	6.3	18	#####	1992	7	10	101 When Shac	FALSE	78	3.086082		-1198490
???	Japan	Daiei Moti	Action	FALSE	0	ja	6.7	11	#####	1962	12	1	104 The Film th	FALSE	79	1.959133		-1569589
English	United Stat	Columbia P	Comedy	FALSE	0	en	4.6	6	#####	1964	1	1	101 Who will b	FALSE	80	3.021832		249053.6
???	Hong Kong	Shaw Broth	Action	FALSE	0	zh	6.5	12	#####	1968	4	3	89	FALSE	81	3.424767		-1788511
English	United Stat	Metro-Gol	Drama	FALSE	0	en	6	1	12/26/1951	1950	12	26	76 SOUTH SEA	FALSE	82	1.809898		-4023252
???	Hong Kong	Jing's Prodi	Comedy	FALSE	0	en	4	1	#####	2007	3	8		FALSE	83	2.546815		-911905
P???????	United Stat	CineTel Fil	Action	FALSE	0	en	5.2	40	#####	2010	2	9	88	FALSE	84	2.948279		7575639
English	France	Carousel Pi																

- b) For Logistic regression the binary response variable is **Popular**. Describe the model and comment if the predictor variables are similar as the above MLR models. Also analyze the accuracy and confusion matrix for the final model.



Fit Details				
Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	9.24708501	0.1387056	4444.5	<.0001*
Likeability	-5.3264188	0.082377	4180.8	<.0001*
For log odds of Other/Popular				
Covariance of Estimates				
Cov				
	Intercept	Likeability		
Intercept	0.0192	-0.011		
Likeability	-0.011	0.0068		

The Rsquare is 0.7849 for Popular.