Predicting Opening Week Box Office performance of Hollywood Movies

Team: BIDM Project A1
Members :
Deepshikha Yadav (61110463)
Shanawaz Janmohamed (61119007)
Vibha Naryan (61110350)
Udayan Dasgupta (61110544)
Santosh P N (61110163)

Agenda

1. Objective
2. Snapshot of Data
3. Model Building
4. Key Findings
5. Further Improvements
6. Appendix

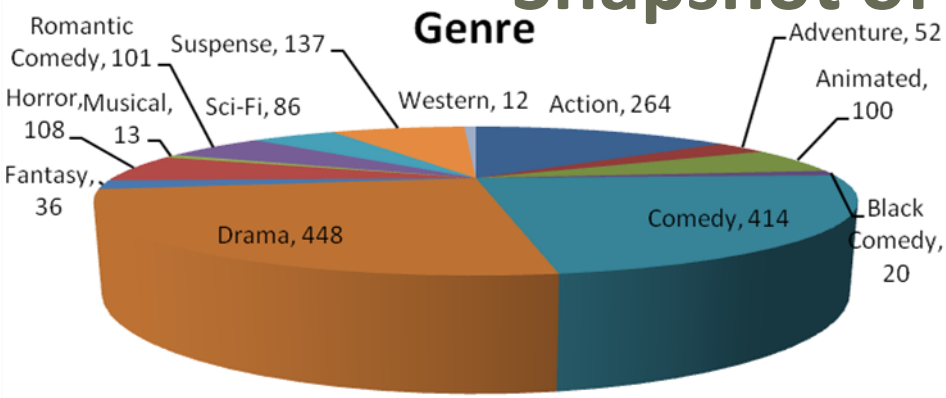
Logo

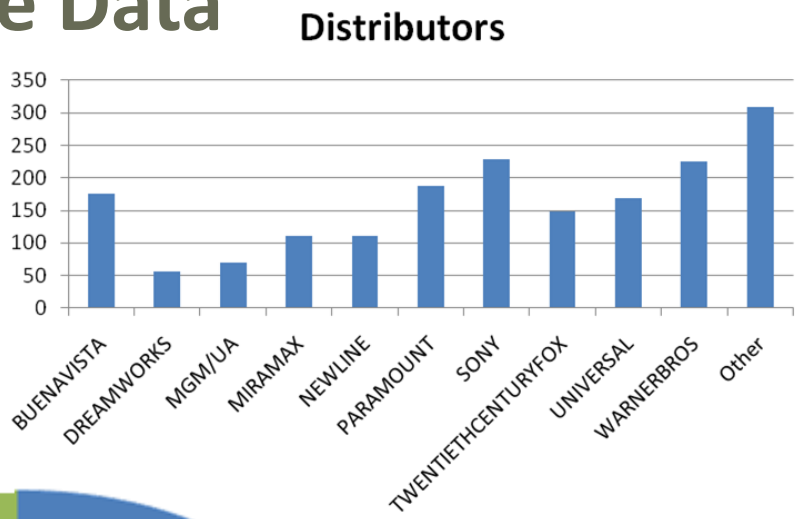
Project Objective

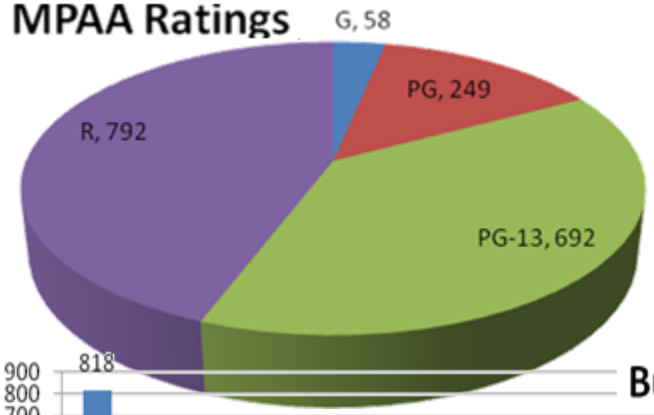
- To predict Opening Box Office Revenue of a movie

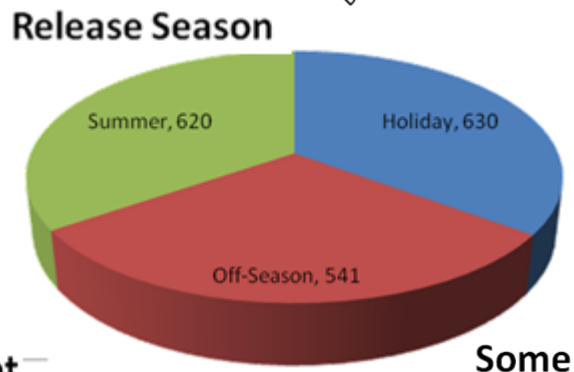
- To Predict whether a movie will break even in first weekend

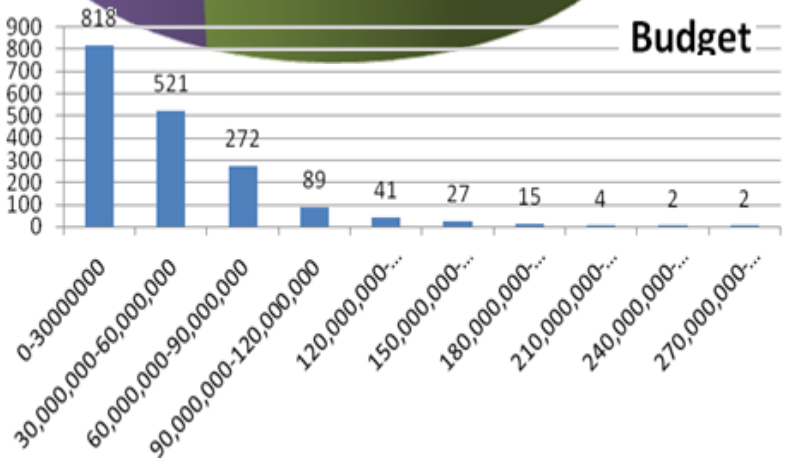
Snapshot of the Data

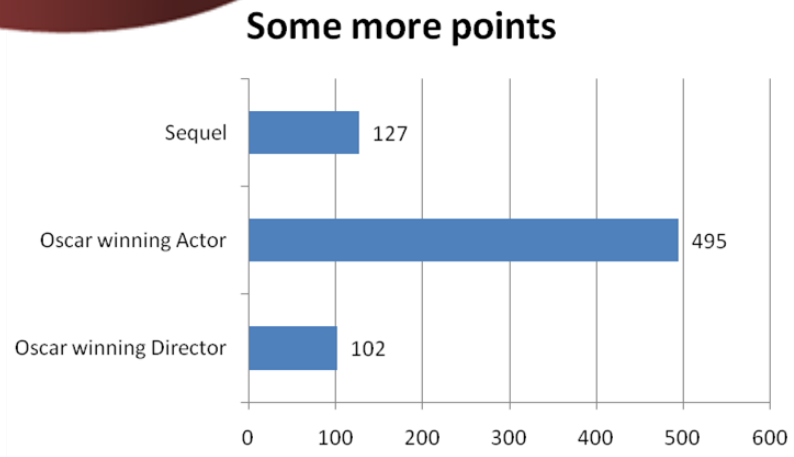












Logo

Model Building

1. Regression Tree
2. Multi-Linear Regression
3. Naïve Bayes
4. Classification Tree
5. Logistic Regression
6. Ensemble Model

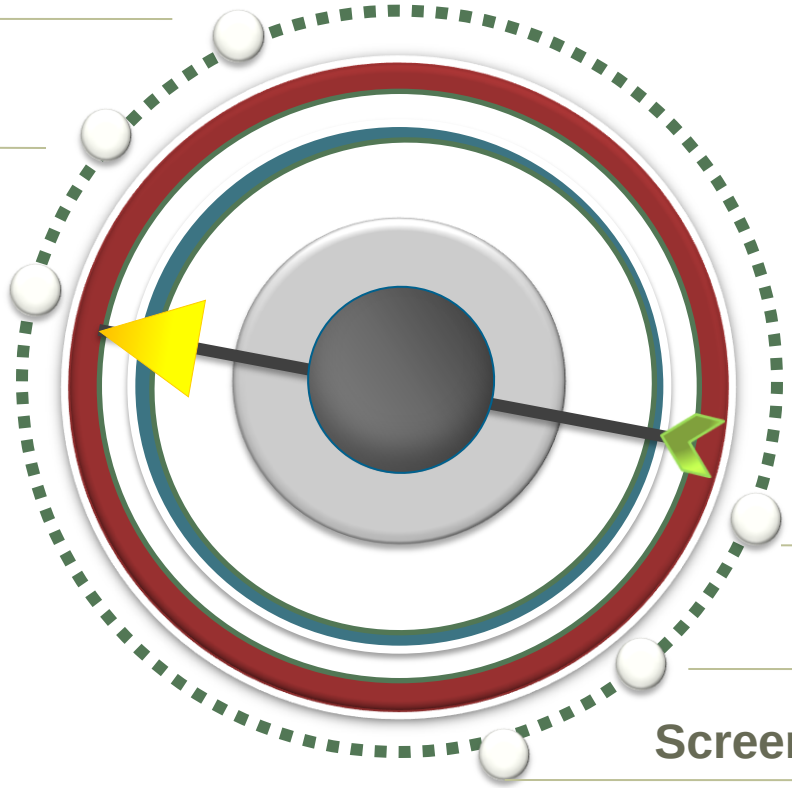
Key Findings

Key Revenue Drivers

Adventure, Drama
Horror , Sequel

Summer

Distributed by
- ❑Paramount
- ❑Warner bros
- ❑Dream –work
- ❑Miramax
- ❑20th century fox



Type of Content

Release timing

Screen and Budget

Further Improvements

Current Model Findings

- We were able to predict breakeven with accuracy of 70%
- We were not able to predict box office revenue

Simonoff's Model

Log domestic = -0.164+1.09 Log 1st weekend

Additional Parameters

- Marketing Budget distribution across various channels
- Duration(Running time of movie)
- Lead actors and actress salary

Logo

Thank You

Appendix

Logo

Data Source: Hollywood Movies

Original data

- www.imdb.com
- http://www.imdb.com/title/tt0116191/
- http://www.imdb.com/title/tt0116191/parentalguide#certification
- http://www.1728.com/page8.htm

Additional data

- www.the-numbers.com
- www.leesmovieinfo.com
- www.boxofficemojo.com

- ❑ 1800 data points for training and validation
- ❑ 800 data points as test data
- ❑ Data from year 1996-2010

Column Names

- Breakeven
- ROI
- Box office opening week revenues
- Budget
- MPAA Ratings
- Oscar Actor/Director/Producer
- KIM (Sex/ Violence /Profanity)
- Genre
- Distributor
- opening week/month/year

Bollywood data set

- http://www.bollywoodtrade.com/box-office/movies-domestic.htm
- http://en.wikipedia.org/wiki/Bollywood_films_of_2010

Logo

Data Exploration

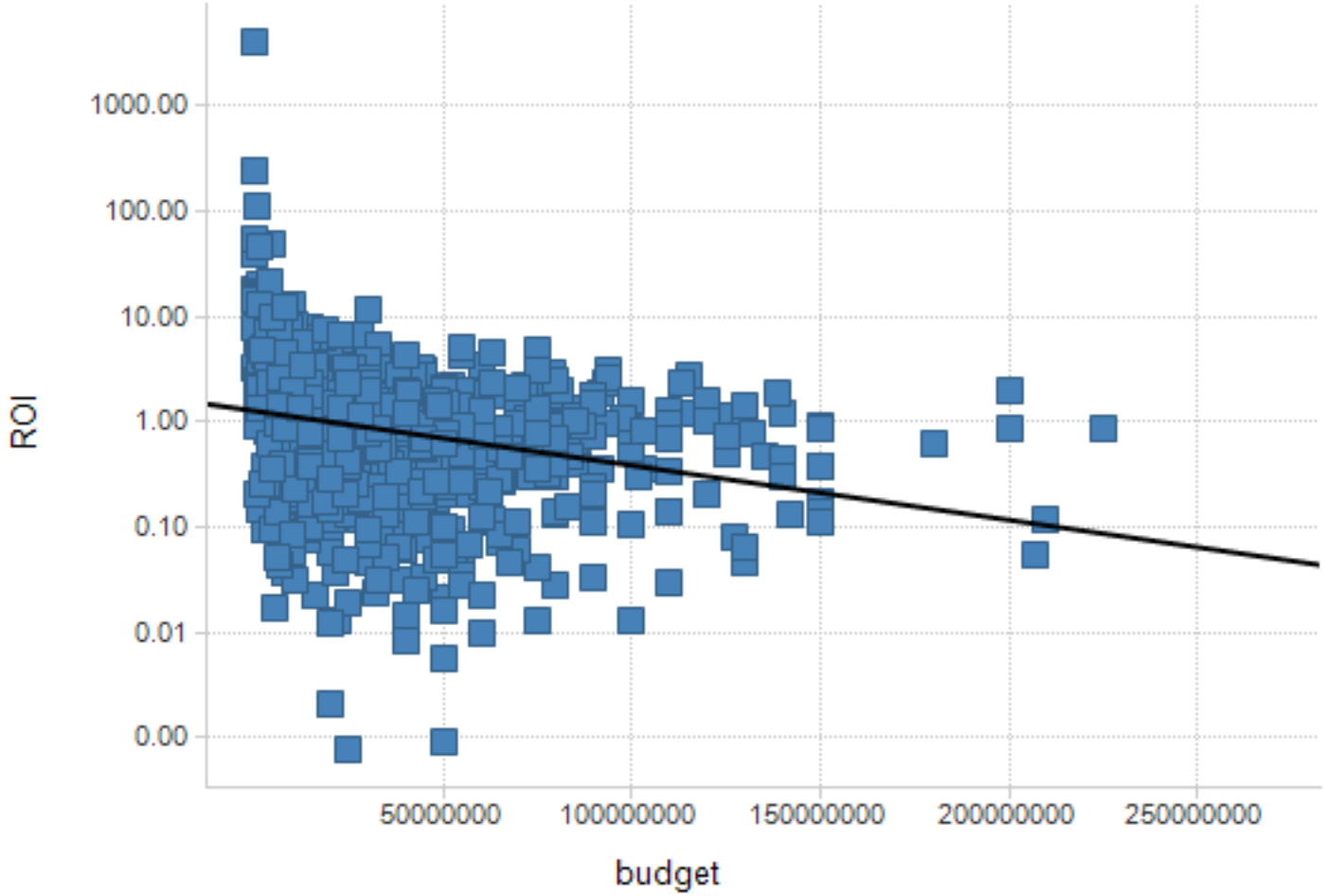
Data Relationships (2) (Linear Regression)

Y (numerical)	X (numerical)	p-value	FStat	RSq	R
ROI	budget	1.33E-001	2.25	0.00	-0.0

Data table:
Data Relationships (2)

Marking:

Data Relationships (2) (Details)



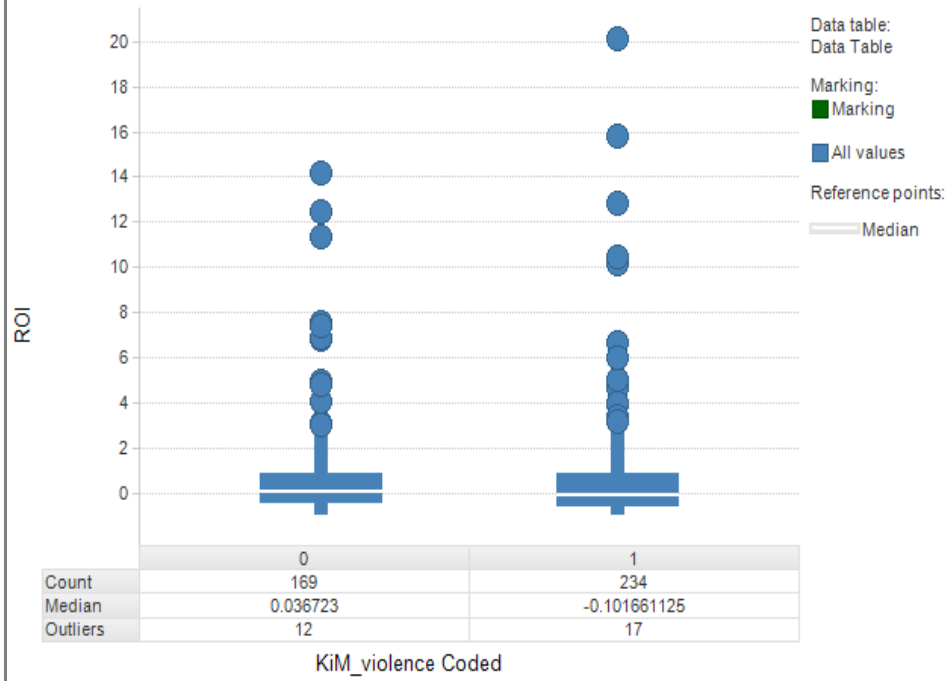
Data table:
Data Table

Marking:
Marking

Marker by
(Row Number)

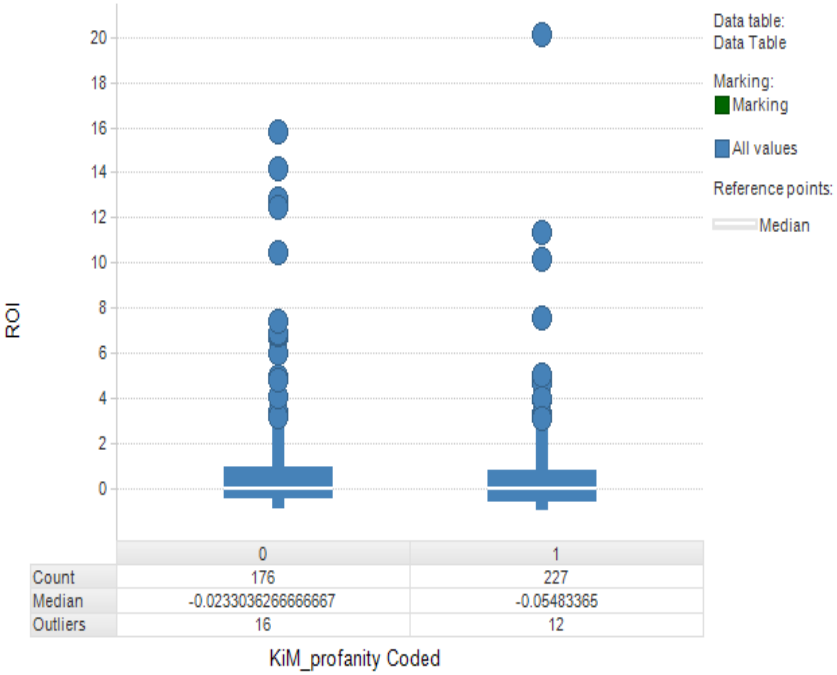
All values

Box Plot

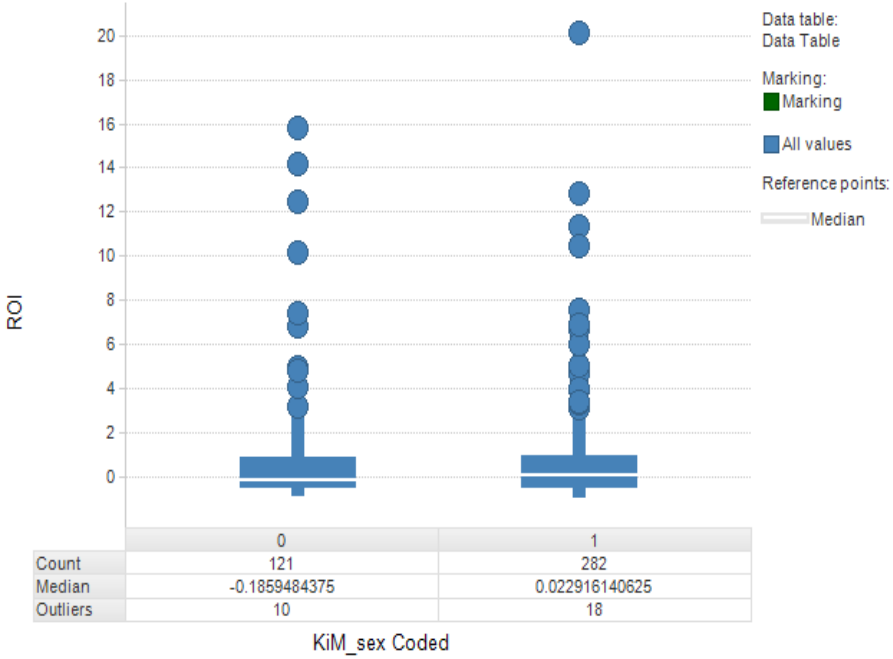


	0	1
Count	169	234
Median	0.036723	-0.101661125
Outliers	12	17

Box Plot



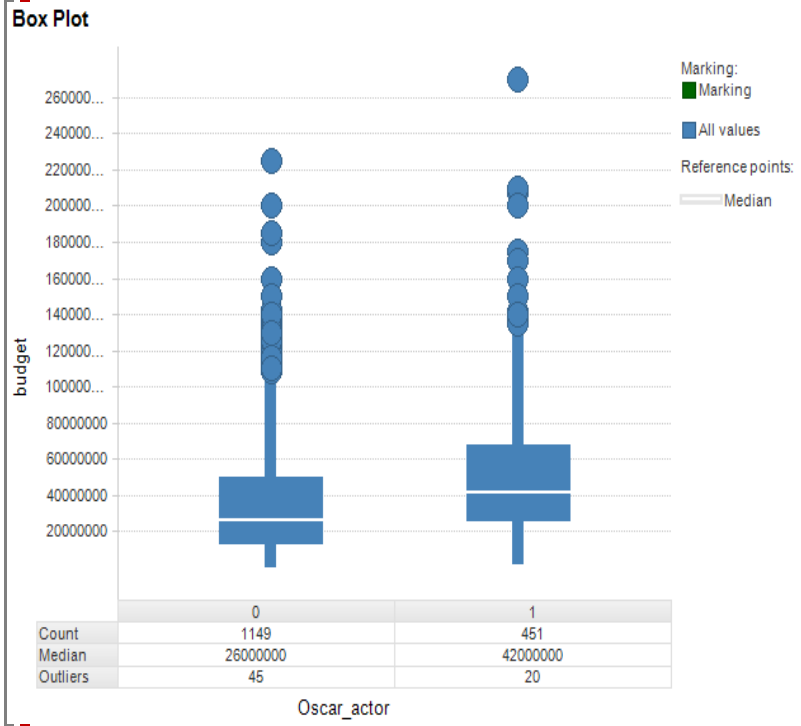
	0	1
Count	176	227
Median	-0.0233036266666667	-0.05483365
Outliers	16	12

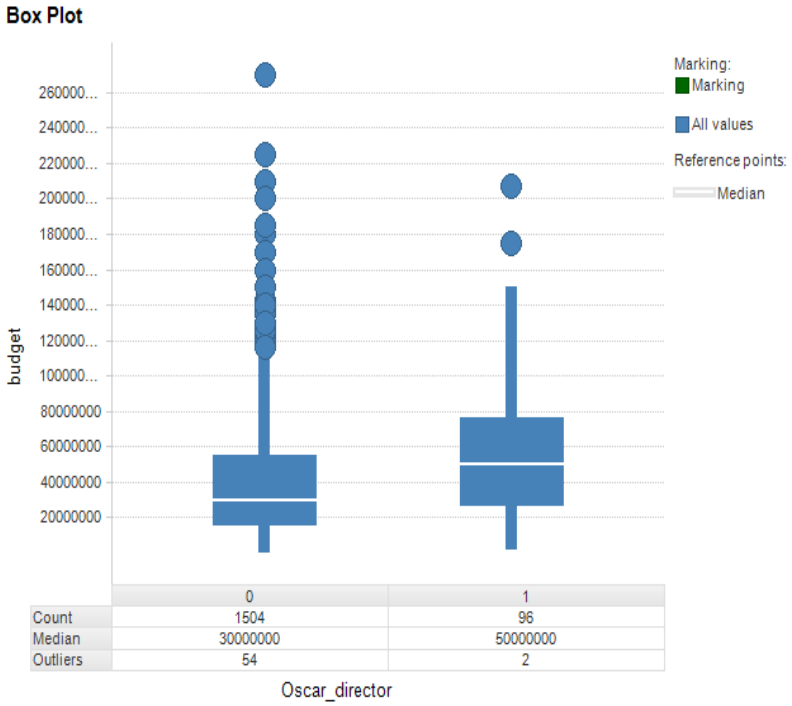


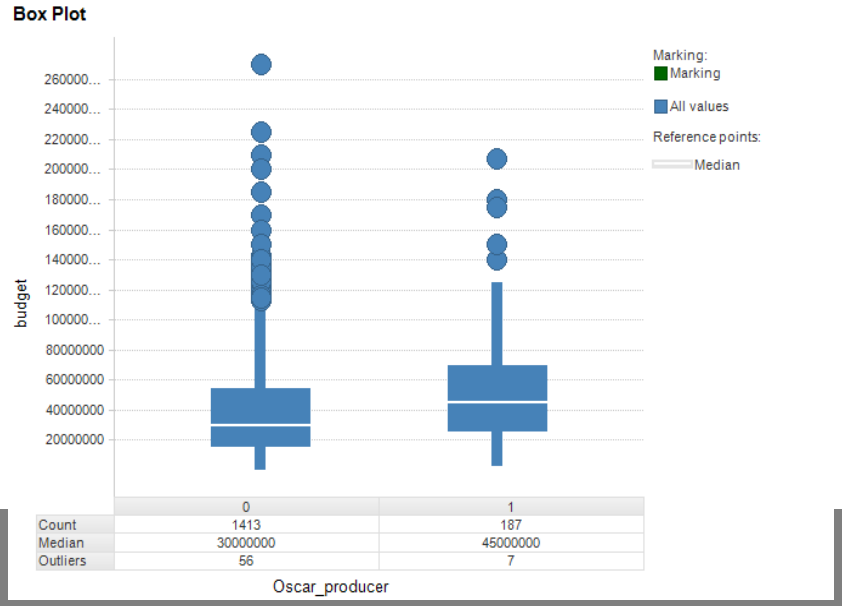
	0	1
Count	121	282
Median	-0.1859484375	0.022916140625
Outliers	10	18

ROI for a movie is negatively correlated to the presence of violence and profanity. However, the data shows that ROI has a positive trend in relation to presence of sexual content.

Logo



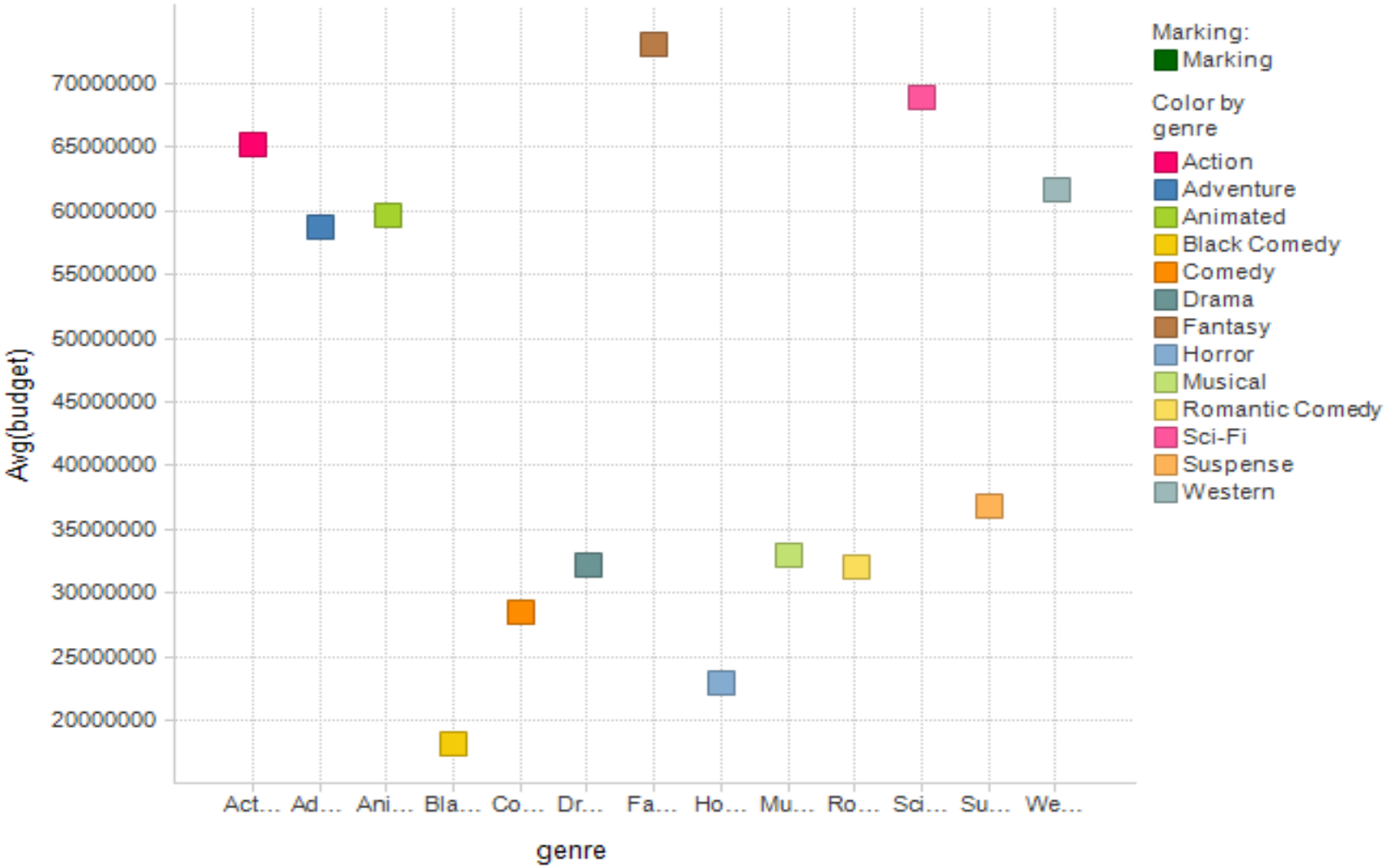




Budget for a movie shows an increasing trend in relation to the presence of Oscar actor , director or producer.

Logo

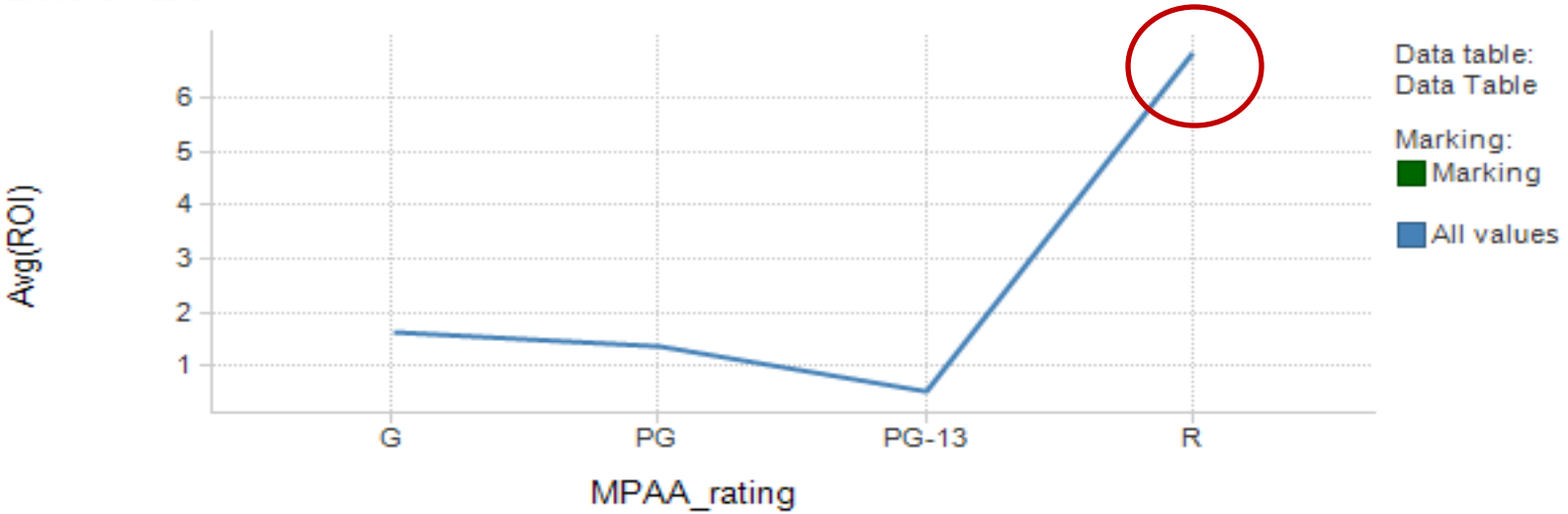
Scatter Plot



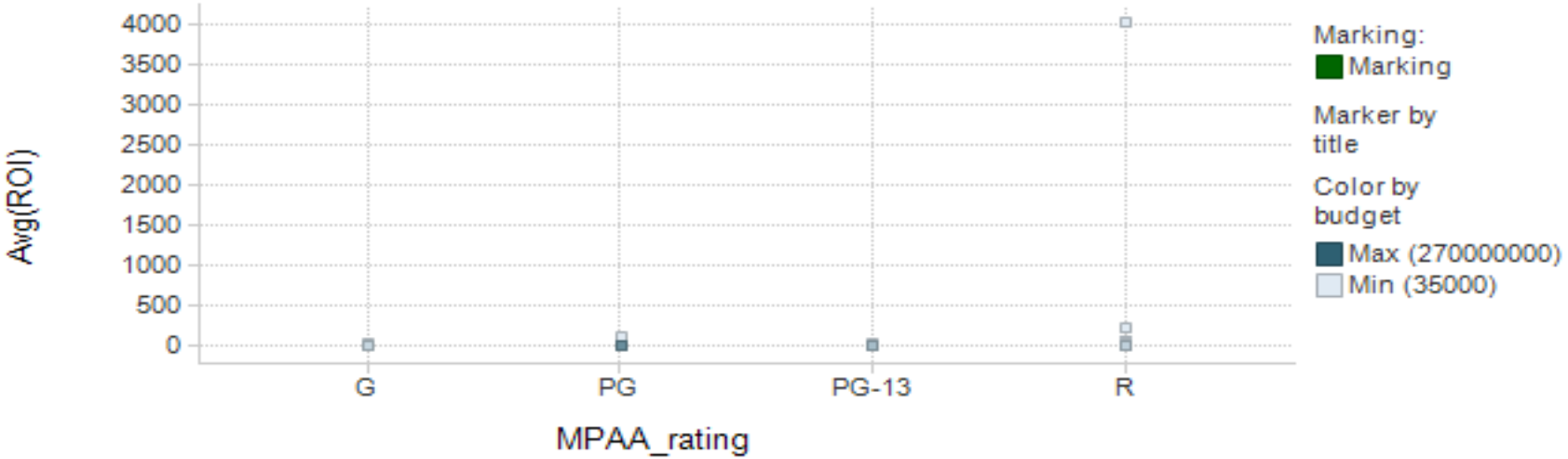
Average budget is typically high for movies belonging to Action , fantasy, Animation , Sci-Fi genre.

Logo

Line Chart



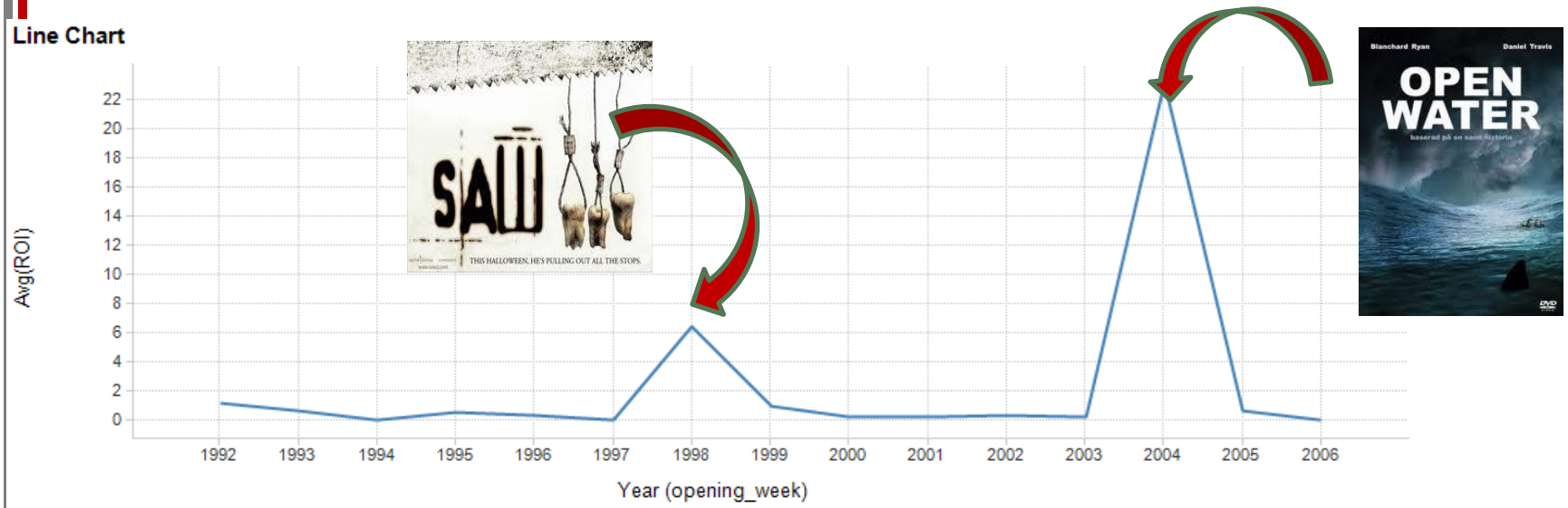
Scatter Plot



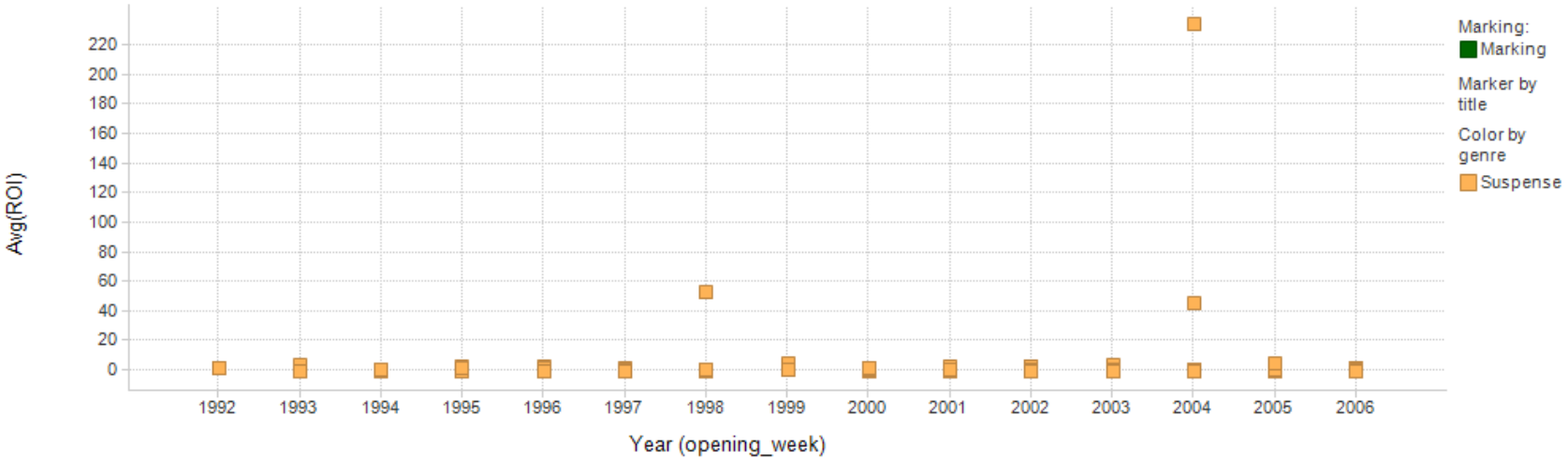
Avg ROI for ‘R’ rated movies tends to be high as compared to PG and PG-13 movies

Logo

Line Chart

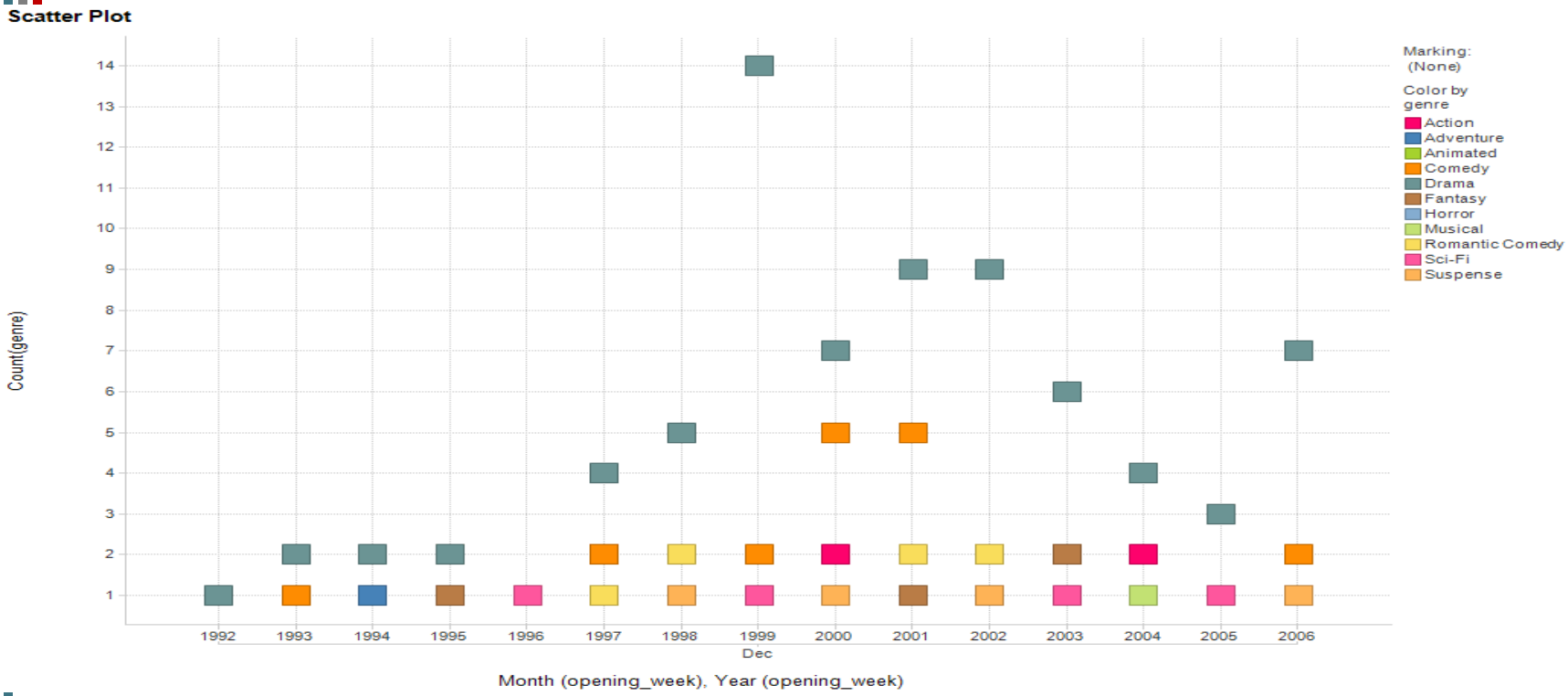


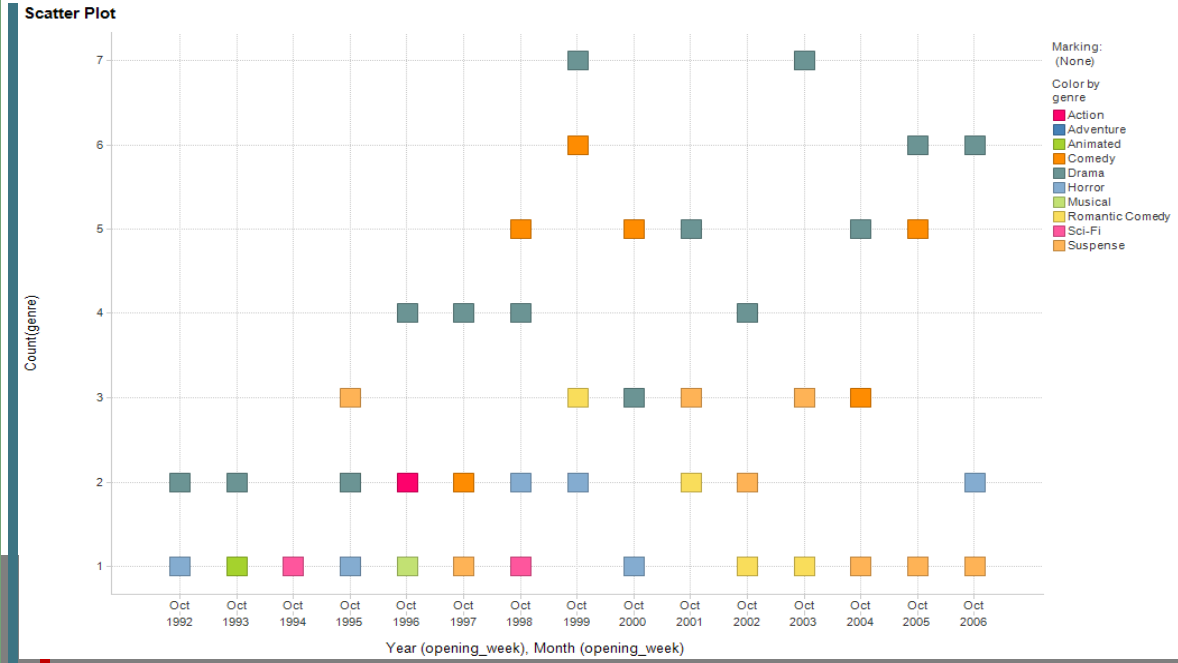
Scatter Plot



The suspense genre data shows spikes in the data for Avg ROI for movies like Open Water and SAW which are medium budget movies but have earned a huge revenue.

Logo





In the month of December, most of the movies belong to drama genre.

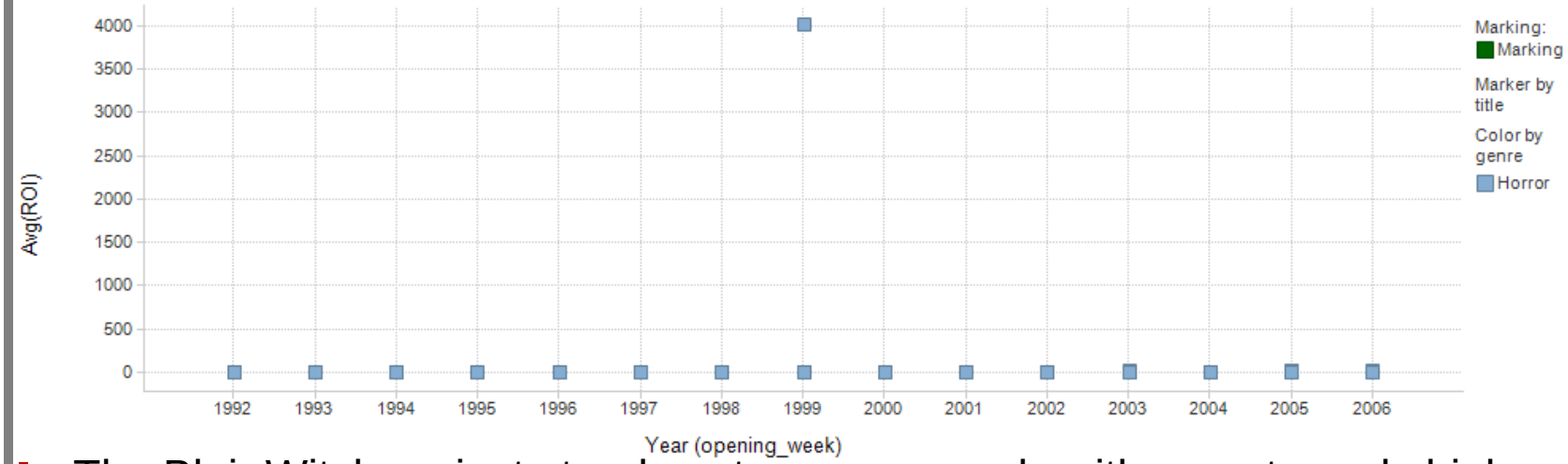
However, in the month of October from the year 2001 to 2005 , the data shows a pattern of genres preferred namely – Drama, comedy, suspense.

Logo

Line Chart

The Blair Witch Project

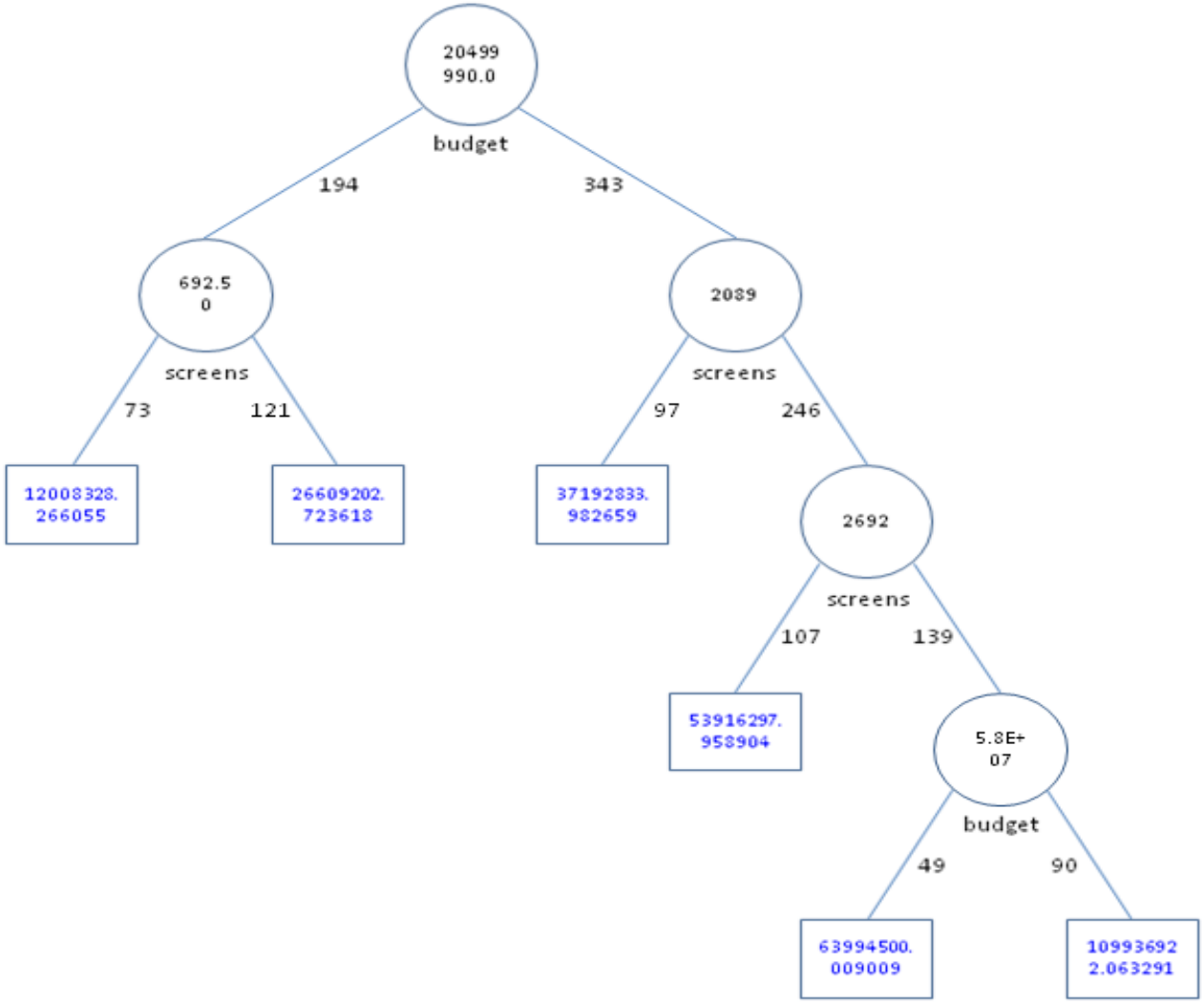
Scatter Plot



The Blair Witch project stands out as an anomaly with an extremely high ROI as compared to other movies of the same genre. This anomaly is due to extremely low budget of the movie (35000$)

Logo

Model : Regression Tree(Continuous Response)



Validation Data scoring - Summary
sing Prune Tree)

Total sum of squared errors	RMS Error	Average Error
1.53695E+18	53498702.35	1362057.907

Logo

Regression Model(Continuous Response)

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3.91168594	13.14394665	0.76613033	26401.87891
genre_Horror	100.4289246	21.06678963	0.00000246	333723.9688
screens	-0.01027392	0.0048622	0.0350982	74345.84375
Opening_Wknmer	15.29835224	9.2486763	0.09873728	46816.76953
budget	0.00000016	0.00000014	0.027727968	5629.838379
KiM_sex Coded	13.48590279	11.87761402	0.025675398	16797.4082
KiM_violence Coded	22.88528633	10.14943409	0.02457906	59309.75781
KiM_profanity Coded	-17.4772415	11.15459824	0.011779463	43049.92969

Residual df	888
Multiple R-squared	0.035889134
Std. Dev. estimate	132.4241791
Residual SS	15572110

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
407462.0569	27.54586684	-5.93001635

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
269392.9576	27.43162946	-6.22902859

Logo

Naïve Bayes

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	157	108
0	123	149

Error Report			
Class	# Cases	# Errors	% Error
1	265	108	40.75
0	272	123	45.22
Overall	537	231	43.02

Test Data scoring - Summary Report

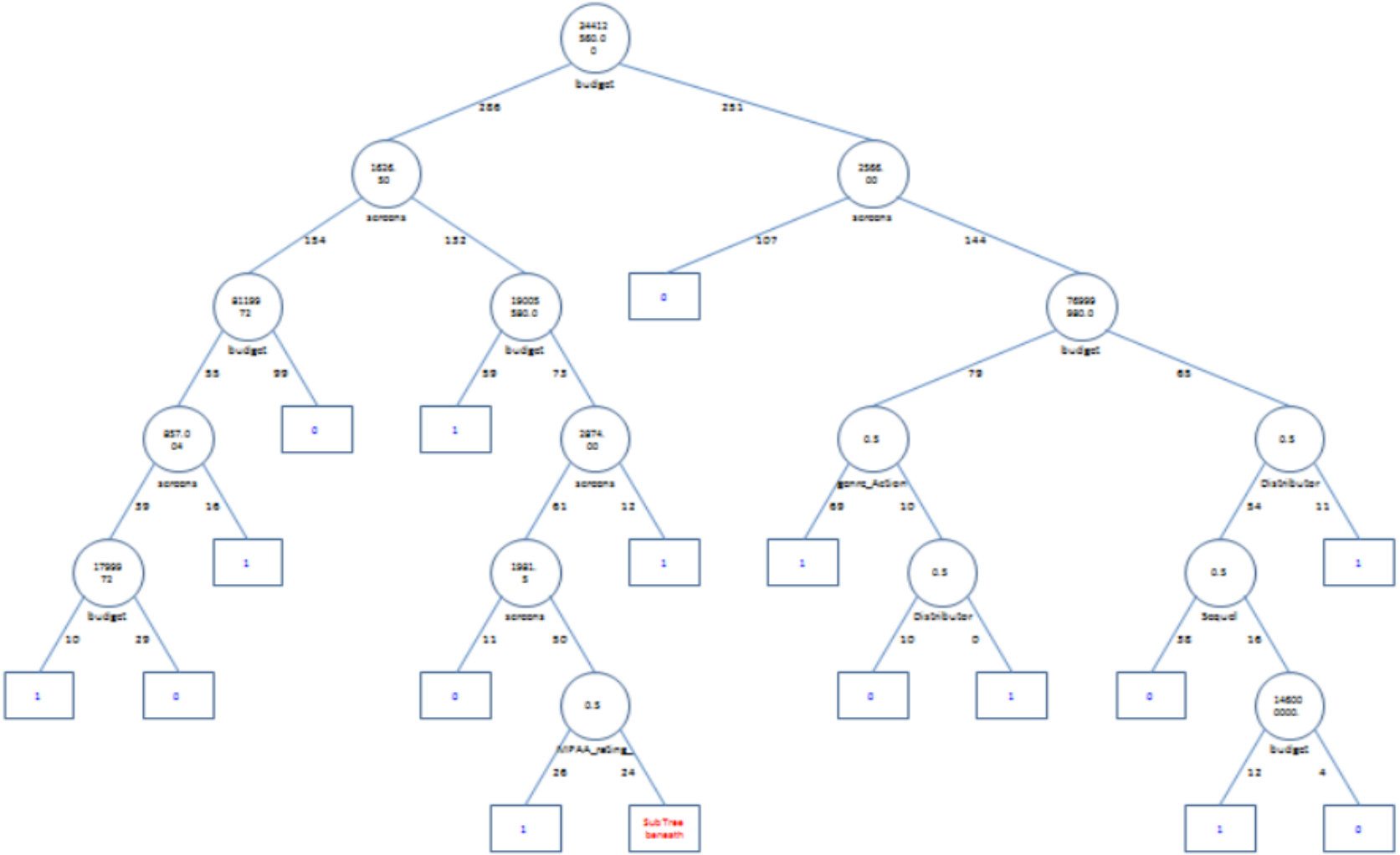
Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	109	74
0	78	97

Error Report			
Class	# Cases	# Errors	% Error
1	183	74	40.44
0	175	78	44.57
Overall	358	152	42.46

Logo

Model :Classification Tree



Logo

Model : Logistic Regression(Categorical Response)

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	168	97
0	120	152

Error Report			
Class	# Cases	# Errors	% Error
1	265	97	36.60
0	272	120	44.12
Overall	537	217	40.41

Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	126	57
0	74	101

Error Report			
Class	# Cases	# Errors	% Error
1	183	57	31.15
0	175	74	42.29
Overall	358	131	36.59

Logo

Ensemble Model with Optimization

	Logistic Regression	Class Tree	Naïve Bayes	Regression Tree	Multi-Linear Reg	imized Mc	Cutoff	Error
	0.2	0.5	0.1	0.1	0.1	1		
Actual Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Average	0.5	32.23%
0	0	0	0	0	1	0.100	0	0
1	1	1	1	1	0	0.900	1	0
1	0	0	0	0	1	0.100	0	1
1	0	0	1	0	1	0.200	0	1
0	1	1	1	1	0	0.900	1	1

	Logistic Regression	Class Tree	Naïve Bayes	Regression Tree	Multi-Linear Reg	Optimized Model
Validation Data	40.4%	37.4%	43.0%	43.2%	53.3%	32.2%
Test Data	36.8%	37.7%	42.3%	42.9%	53.9%	29.0%

```
CLASSIFICATION DATA
-------------------
Total Cells
  Numerics
    Adjustables
      Continuous
      Free
      Integers/Binari
    Constants
    Formulas
  Strings
  Constraints
Nonlinears
Coefficients
```

0	0	0	0	1	0	0.100	0	
1	1	0	0	1	0	0.300	0	Error
0	1	0	1	1	0	0.400	0	
0	0	1	0	1	1	0.700	1	Error
1	1	1	1	1	0	0.900	1	
1	0	0	0	0	0	0.000	0	Error
1	0	0	0	0	1	0.100	0	Error
0	0	0	0	1	1	0.200	0	
1	0	0	0	0	0	0.000	0	Error
0	0	0	0	0	0	0.000	0	
1	1	1	1	1	0	0.900	1	
1	1	1	1	1	1	1.000	1	
1	0	0	0	1	0	0.100	0	Error

Logo