What makes a movie click?

Factors affecting success and failure of movies

Shikha Chamoli, MSCS , IIT Chicago

Outline

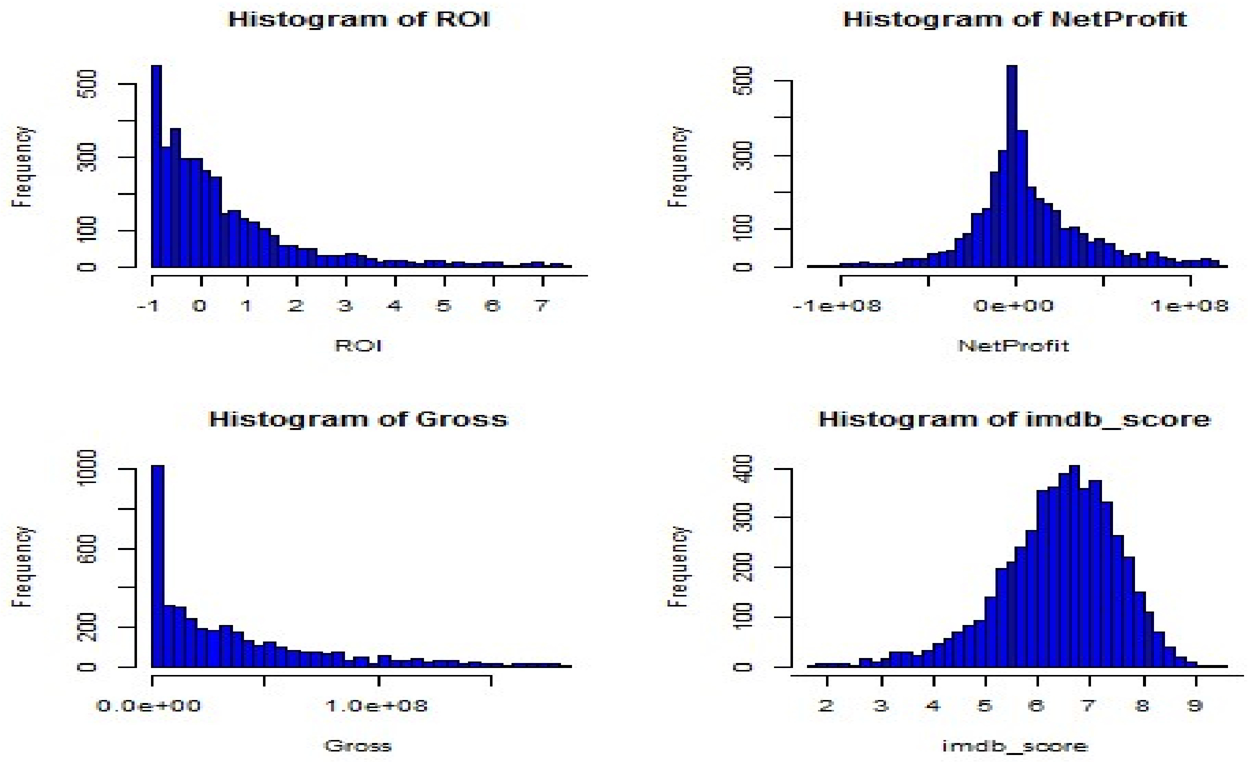
- Quantifying the success of a movie
- Identifying the underlying factors
  - Demographic factors
  - Style factors
  - Genre
  - Plot
- Factor Model:
  - Establishing a response variable
  - Objective : To explain the variance of the response variable
  - Identifying the shortcomings of the model
- Conclusions:
  - What Strategies to follow
  - What to avoid

Quantifying Success

- Financial Measures:
  - Gross Revenue (gross) : A naïve measure
  - Net Profit = gross – budget
    - Focusses on profitability of the investment
  - Return on Investment(ROI) $=\frac{gross-budget}{budget}$
    - Focusses on percentage returns
    - Makes success comparable across time

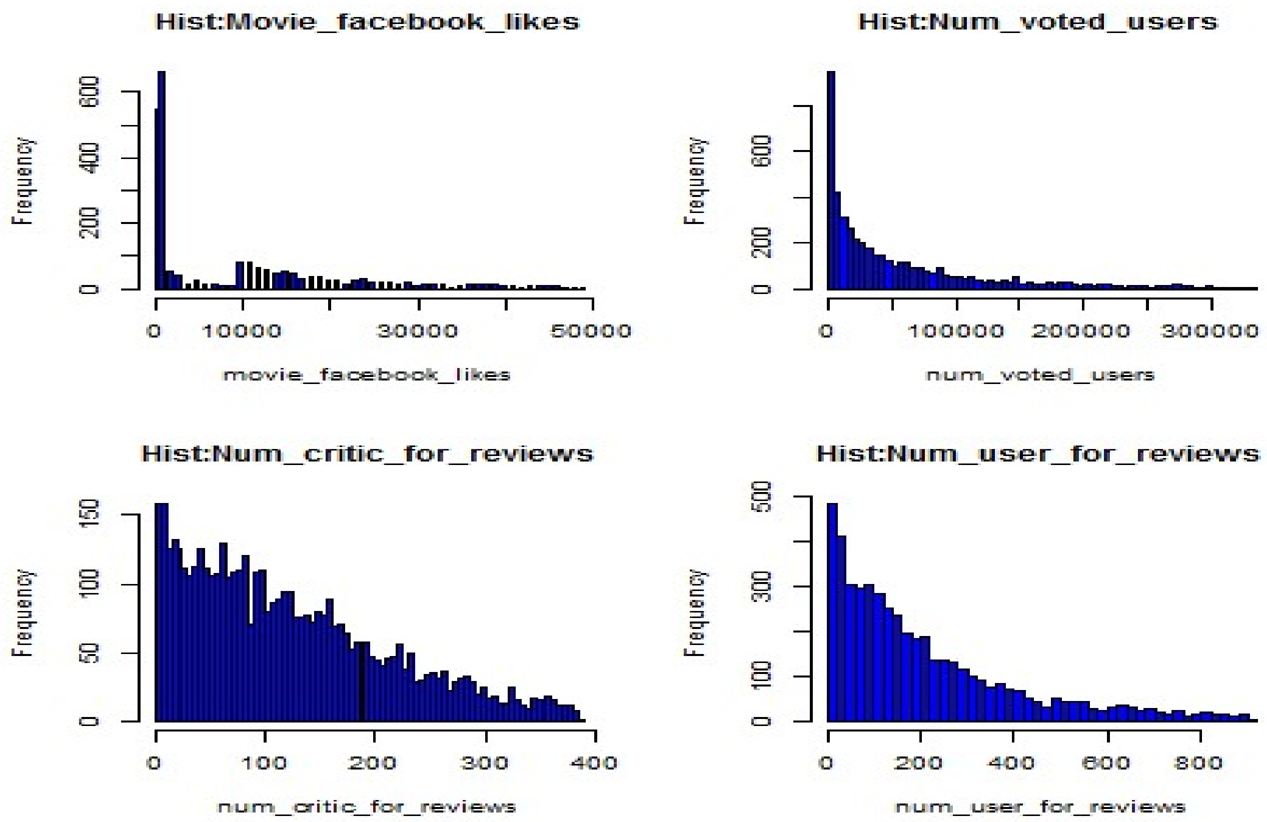
- Other Measures/Indicators:
  - Imdb score
  - Number of Users Voted
  - Number of critique reviews
  - Number of user reviews

Quantifying Success



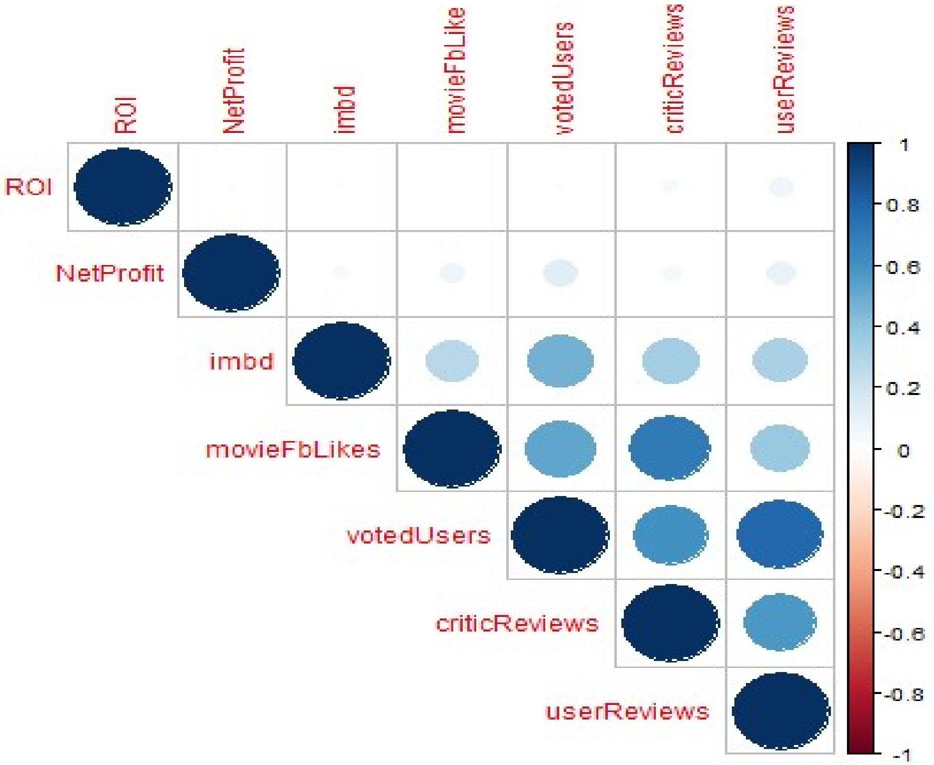
- ROI & Gross: Looks like right skewed with a thin tail.
- Net Profit : Looks normally distributed around 0
- Imbd Score: Looks normally distributed around 6.5

Quantifying Success



- Movie FB likes : No Information
- NumVotedUsers & numUserReviews:
  - Has similar distribution to ROI
- NumCriticReviews:
  - Looks somewhat correlated to ROI

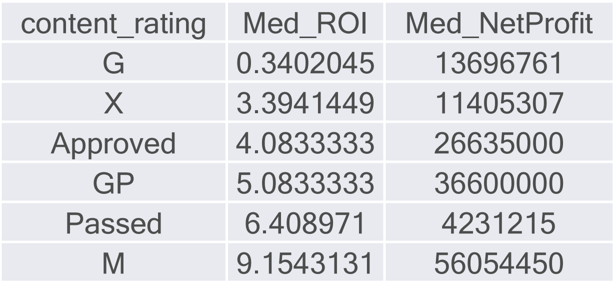
Quantifying Success



- Correlation Plot for various variables:
  - Establishes that ROI and NetProfit are indeed positively correlated with all the other indicators.
  - Confirms our observations of similar distributions in previous slides.

Identifying Factors: Demographics

- Language:
  - Among 5043 data points, 3707 are of English
  - Other 37 languages have very small number of movies.
  - Statistical Conclusions:
    - Most conclusions will be dominated by English movies
    - Wouldn’t really apply to other language movies.
- Country:
  - Among 5043 data points, 3074 are made in USA:
  - Other 116 countries have relatively small number of movies.
  - Statistical Conclusions:
    - Most conclusions will be dominated by American movies
    - Wouldn’t really apply to other language movies.
- Content Rating:

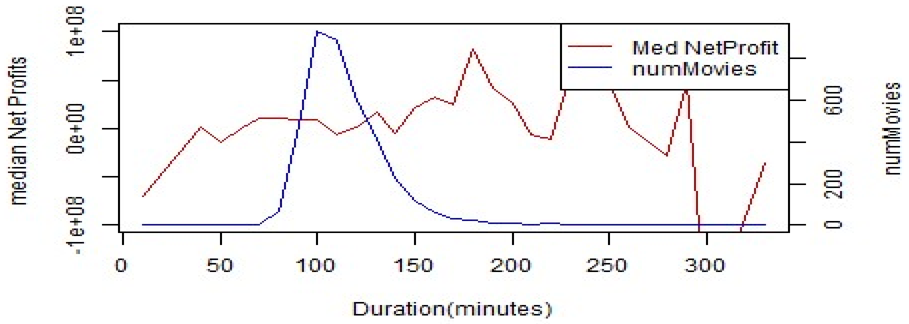


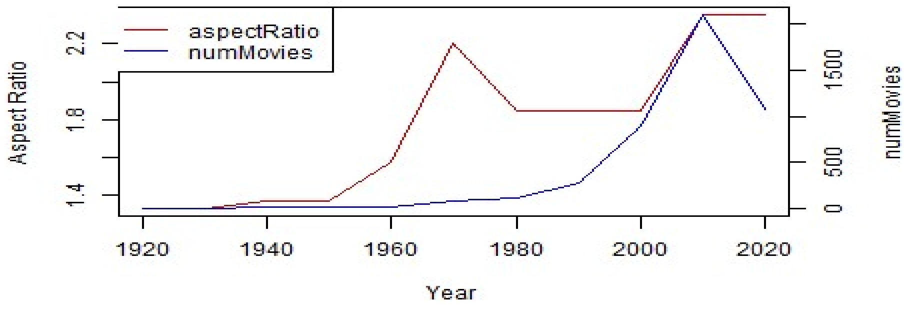
content_rating	Med_ROI	Med_NetProfit
G	0.3402045	13696761
X	3.3941449	11405307
Approved	4.0833333	26635000
GP	5.0833333	36600000
Passed	6.408971	4231215
M	9.1543131	56054450

  - GP and M are both (now) PG-13 standard which is of moderate impact
  - Data says, that these are the movies that do best business

Identifying Factors: Style

- Color: Almost all data is for color movies



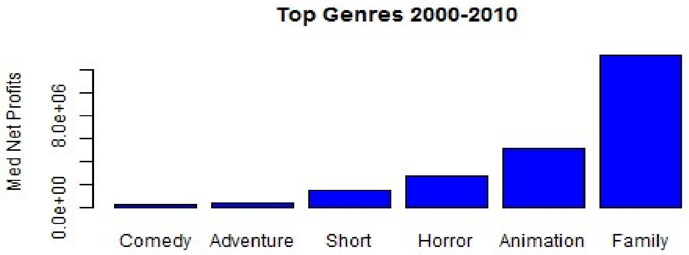


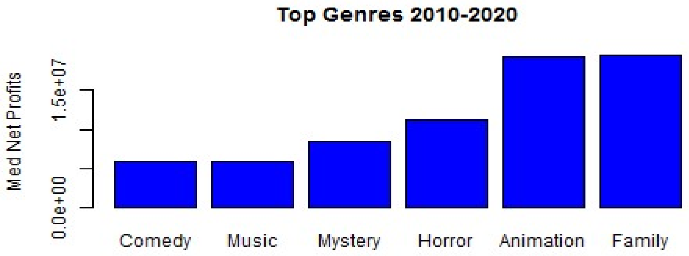
- Duration:
  - Bulk of the movies are of length 90 to 150 mins
  - Longer (100-200) are slightly better (Net Profits wise)
- Aspect Ratio: Mostly irrelevant,
  - has gradually changed over years.

Identifying Factors: Genre

Top Genres By Net Profit			
Genre	Med_ROI	Med_NetProfit	numMovies
Adventure	0.062	2853933	795
Music	0.318	3946250	161
Fantasy	0.131	4976634	517
Horror	0.490	6753840	397
Animation	0.200	11571351	199
Family	0.314	14008741	453

Top Genres By ROI			
Genre	Med_ROI	Med_NetProfit	numMovies
Comedy	0.195	2690677	1511
Animation	0.200	11571351	199
Family	0.314	14008741	453
Music	0.318	3946250	161
Horror	0.490	6753840	397
Short	4.533	909267	2

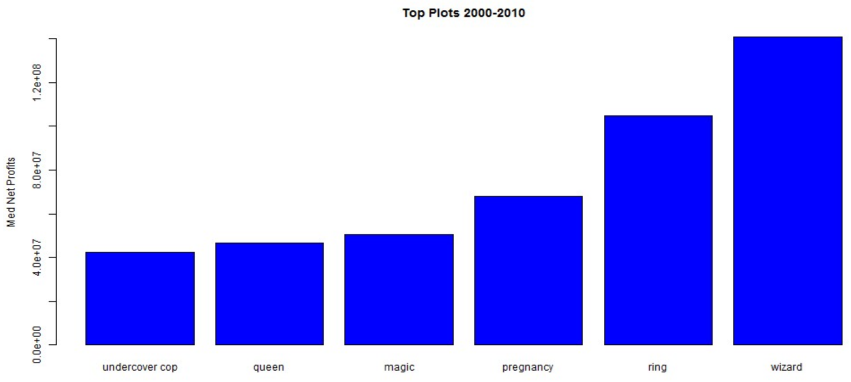


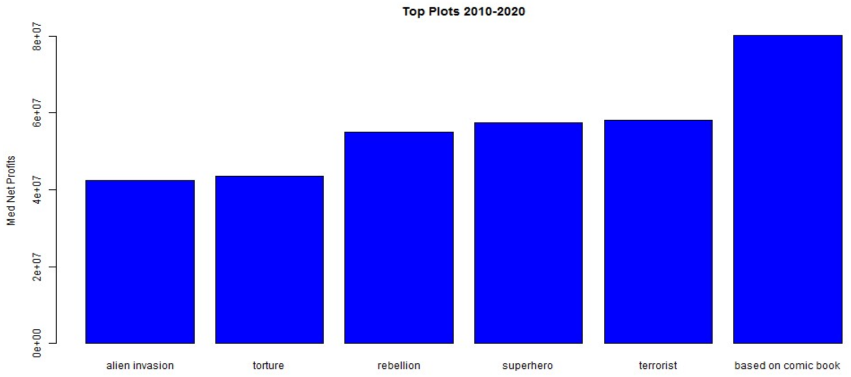


- Key Observations:
  - For high budget, Animation and Family are best (for business)
  - For constrained budget, Horror and Short Movies are best
  - Animations: Business wise this Genre has picked a lot after 2010, and as of now it is as big as Family

Identifying Factors: Plot

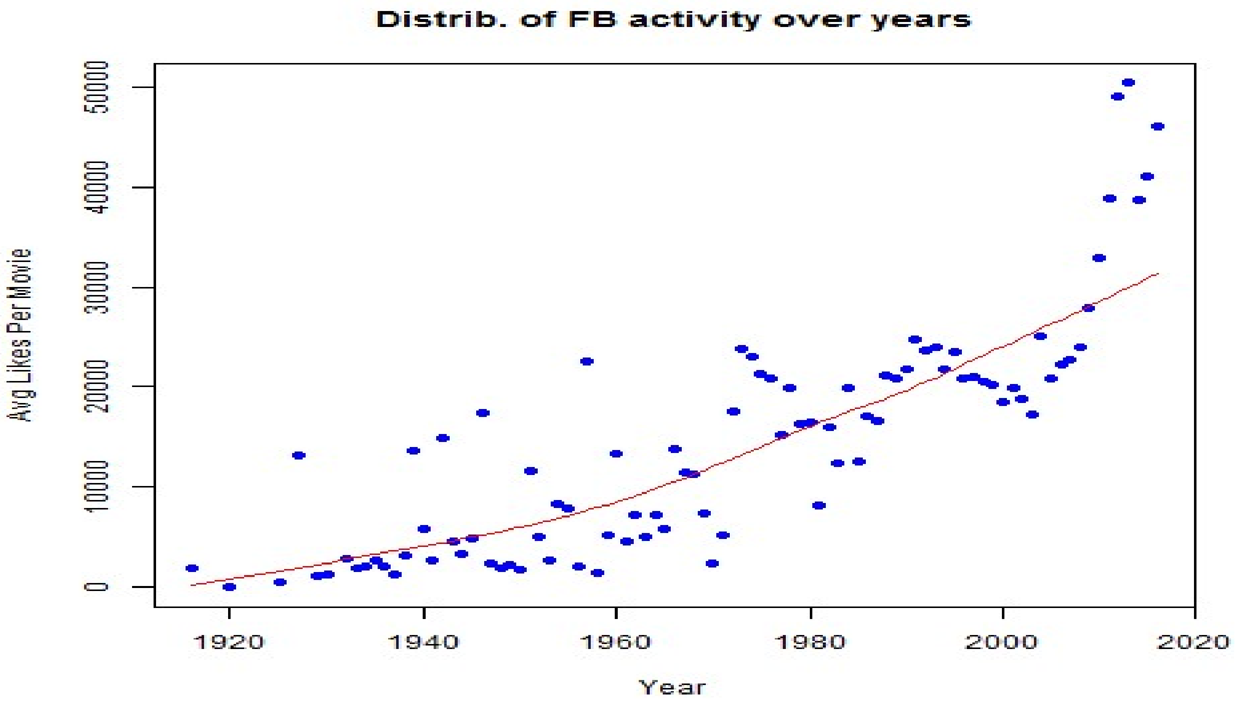
Top Plots by Number of movies			
Plot	Med_ROI	Med_NetProfit	numMovies
new york city	0.09726693	308707	75
police	0.114865	1239558	93
death	0.02067494	682273	105
murder	-0.1171151	-2057591	132
friend	0.3522218	1727544	137
love	-0.1035104	-744000	155


Top Plots 2000-2010


Top Plots 2010-2020

- Key Observations:
  - 2000-10: Most important plots words are ring and wizard
    - Its probably dominated by LOTR, and Harry Potter (both mega hits)
  - Post 2010 :
    - Most popular plot is from comic books and superheros
    - Terrorist plot has gained importance: Reflects fear in society
  - Popular among movie makers BUT Bad:
    - Love, murder , death : People have had enough of these plots

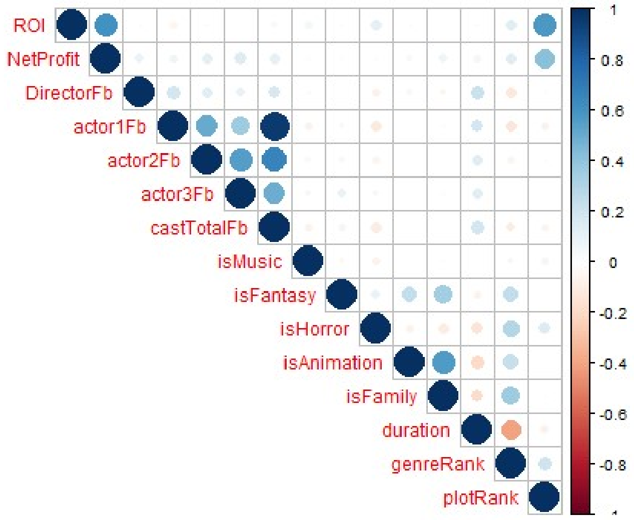
Identifying Factors: Facebook



- Key Observations:
  - Facebook didn’t exist before 2005, but people now can still vote for a 1960 movie
  - Activity in new movies are far more than those in old movies (non-linearity is clearly visible)
  - Further analysis in Model

Factor Model:

- Genre and Plot both look important but to quantify its importance, I have created 2 factor
  - GenreRank: WghtAvg of ROI of Movies in a given genre
  - PlotRank: WghtAvg of ROI of Movies with a given plot
- Factor Correlations:



  - Use these to eleminate
  - Correlated factors
  - Lack of inter factor
  - Correlations is good

Factor Model: Continued

- Two Models :
  - Model 1 : for predicting ROI
  - Model 2 : for predicting NetProfit
- X Variables:
  - FB likes of director, actor , cast
  - Indicator variables for popular genres
  - PlotRank
  - GenreRank
- Model 1 (Post dropping weak factors):
  - ROI ~ DirectorFb+isMusic+isFantasy+isHorror+plotRank
- Model 2 (Post dropping weak factors):
  - NetProfit ~ DirectorFb +castTotalFb+isHorror+isFamily+genreRank+plotRank

Model 1		R2=0.33	F-stat 384
Estimate	Std. Error	t value	Pr(>|t|)
1.59E-01	3.82E-02	4.162	3.22E-05
2.05E-04	6.28E-05	3.268	0.00109
3.08E-01	1.44E-01	2.149	0.03169
-1.75E-01	8.42E-02	-2.076	0.03796
3.14E-01	9.59E-02	3.274	0.00107
2.82E+00	6.63E-02	42.467	< 2e-16

Model 2:		R2 = 0.33	F-stat 171
Estimate	Std. Error	t value	Pr(>|t|)
-6776650.7	961551	-7.048	2.14E-12
8521.5	1313.5	6.487	9.83E-11
406.8	50.9	7.991	1.75E-15
-4450913.3	2089053.7	-2.131	0.0332
6363528.9	2003254.7	3.177	0.0015
31433483.2	7941473.8	3.958	7.69E-05
39408330.5	1379601.8	28.565	< 2e-16

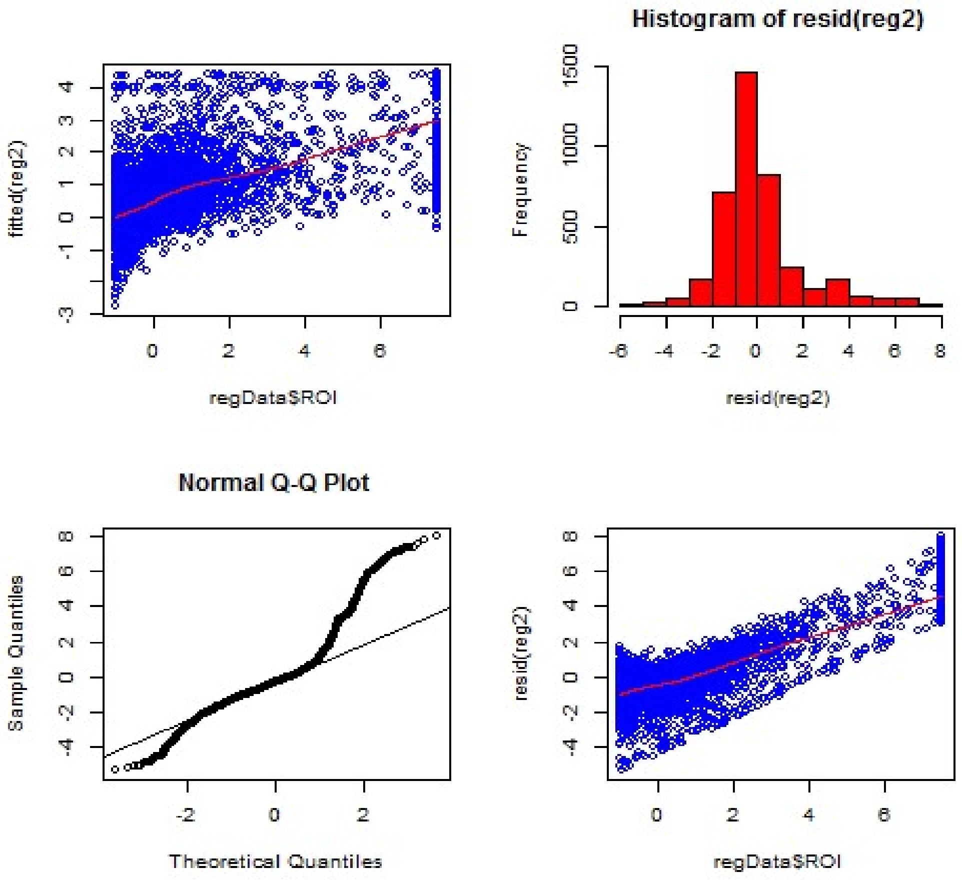
Conclusion: Strategy

- US remains the biggest market :
  - So make a movie for this market
- English remains the most dominant language.
  - No point considering foreign language
- High Budget :
  - For High Budget Project Stick to Genres like Family and Animations
  - These Genres have low ROI , but they do a lot more volume, and Net Profit is higher
- Low Budget :
  - Stick to genres like Horror and Short films
  - These have high ROI but need less investment and are smaller scale projects
- Once Decided on budget, and demography:
  - Use both the Regression Models to Optimally choose the plot , director and story line.
  - Use both the models to gain an insight on how much investment and ROI potential should be expected from any given existing project.

Thank You

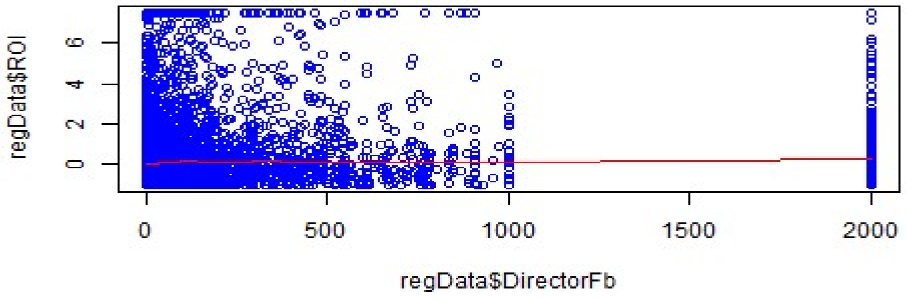
Please also read a supplementary report that I am submitting with this presentation

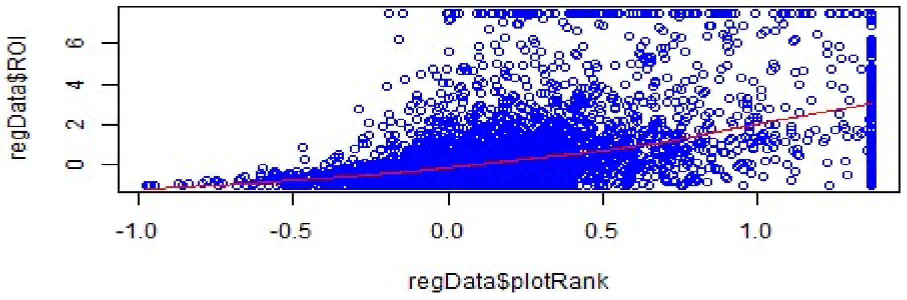
Appendix: Reg1 : Residuals



- Plot1 : Y vs Fitted : Shows model has strength
- Plot2 : Hist of Resid : resembles normality
- Plot3 : QQ plot of resid: Close to normal
- Plot4 : Should ideally be horizontal: Data not nice.

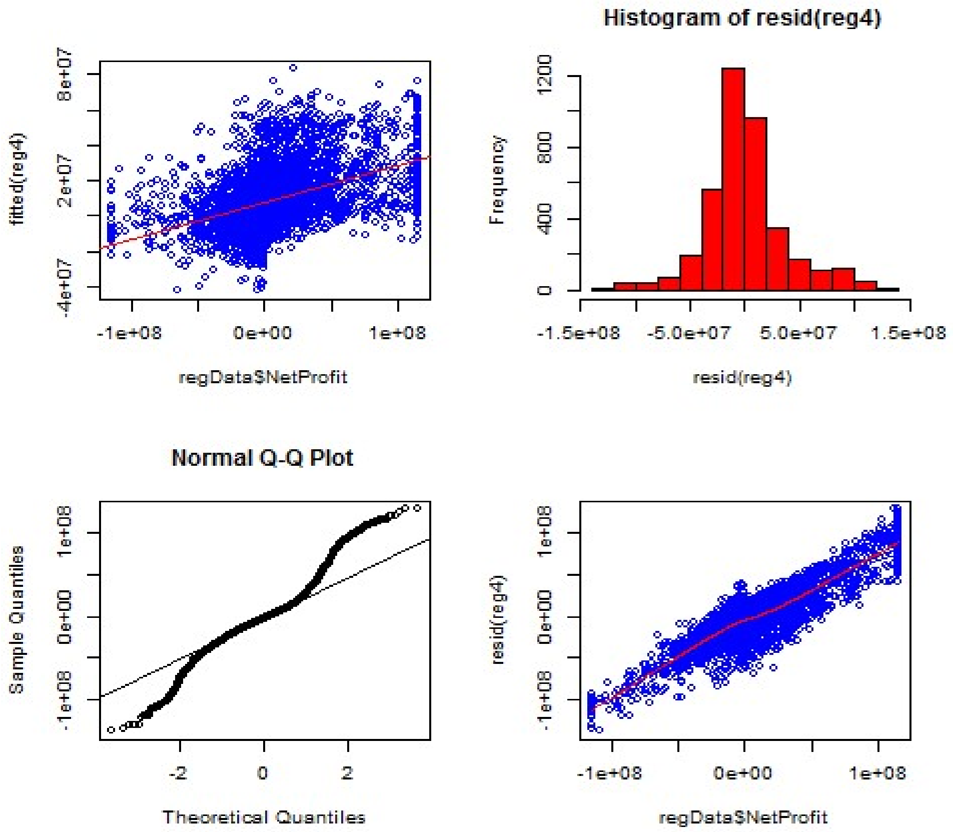
Appendix: Reg1 : XY-plots





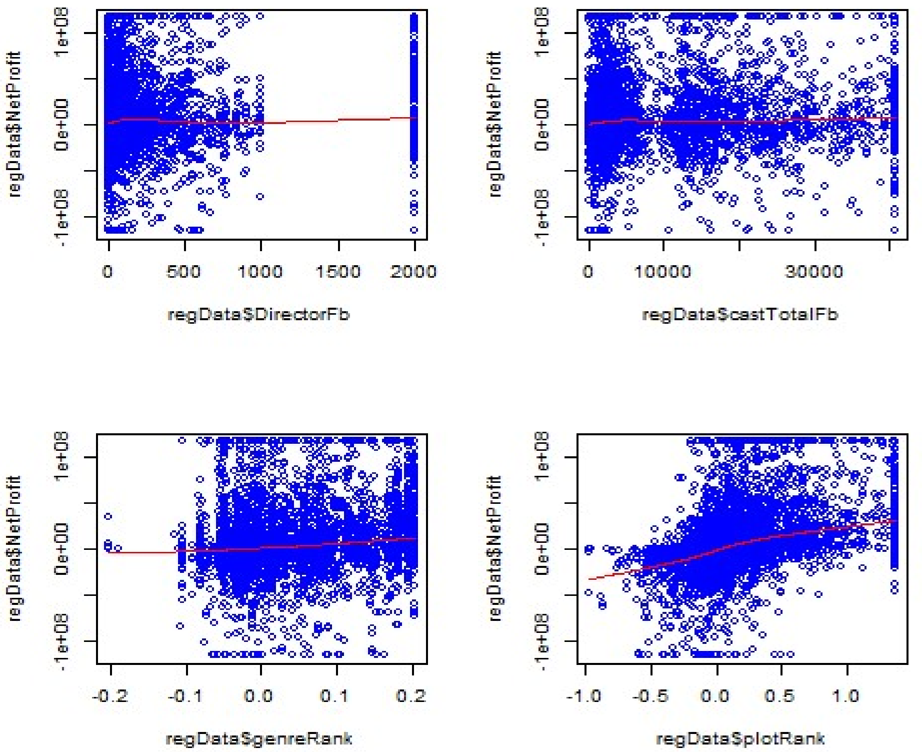
- Plot1 : ROI vs DirectorFB :
  - Relation is weak (possibly the reason behind low R-sq)
  - DirectorFb even if bounded has serious outlier problems
- Plot2 :ROI vs PlotRank:
  - Mild Non linearity, but there is clear relationship.

Appendix: Reg2 : Residuals



- Plot1 : Y vs Fitted : Shows model has strength
- Plot2 : Hist. of Resid : resembles normality
- Plot3 : QQ plot of resid: Close to normal
- Plot4 : Should ideally be horizontal: Data not nice.

Appendix: Reg2 : XY-plots



- Plot1 ,Plot2: NetProfit vs DirectorFB, castTotalFb likes:
  - Relation is weak (possibly the reason behind low R-sq)
  - Outliers remain problem
- Plot2 :NetProfit vs GenreRank, PlotRank:
  - GenreRank: Mild relationship. Weak factor.
  - PlotRank : Some Non linearity, but clear relationship.

Appendix: Reg: Key Remarks

- The most important factors (based on t-stat) is the plotRank.
  - One needs to have a good story line and plot.
- ROI Model (by t-stat):
  - Horror Genre: Most important Genre for high ROI,
  - Somewhat important in the NetProfit model too.
- Net Profit Model (by t-stat):
  - Family Genre : Most Important Genre for high NetProfit
  - Not important to ROI model
- The above conclusions are in sync with the genre and plot analysis done earlier(Slide9,10)
- Quality of data :
  - Over all quality of data is poor,
  - Deploying even more advance Datamining skills may not be of much help.
  - Data suffers from a lot of outliers
- High Fstat of both models (ROI and NetProfit):
  - Shows that the regression models are valid
- Near normal residual QQ plots in both regressions:
  - Further confirms that model assumptions are satisfied and model is valid.