

# BUSINESS INTELLIGENCE AND DATA MINING IN SAS

## BIBO

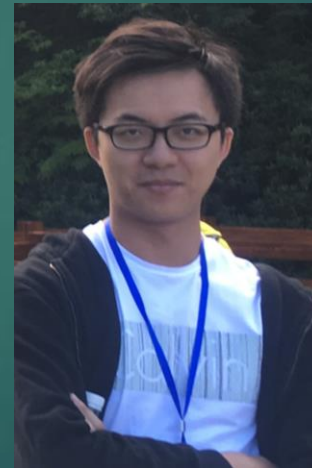
### Business Intelligence for Box Office



Kushal Ganesh Babu



Yunlu Huang



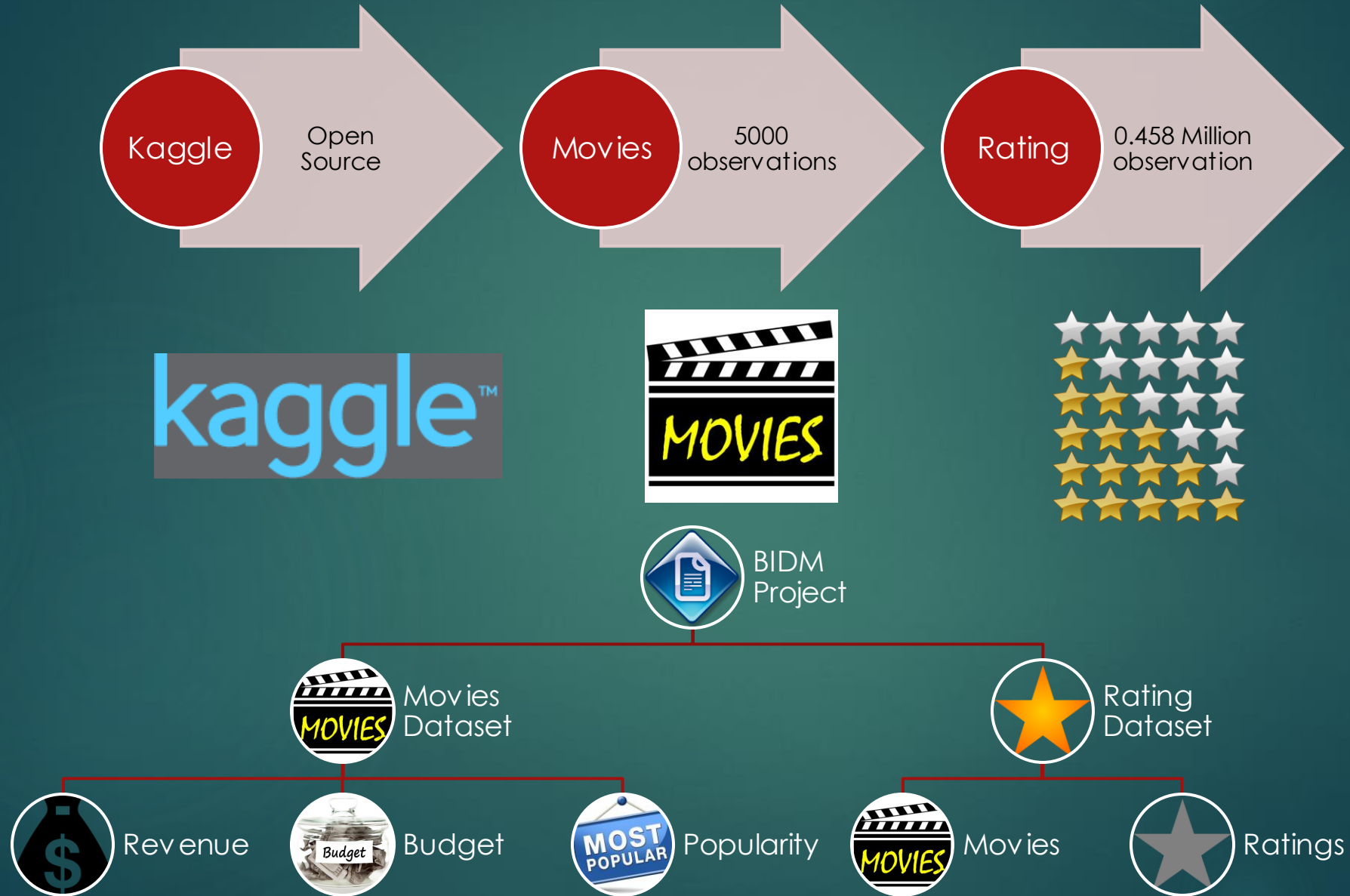
Heng Li



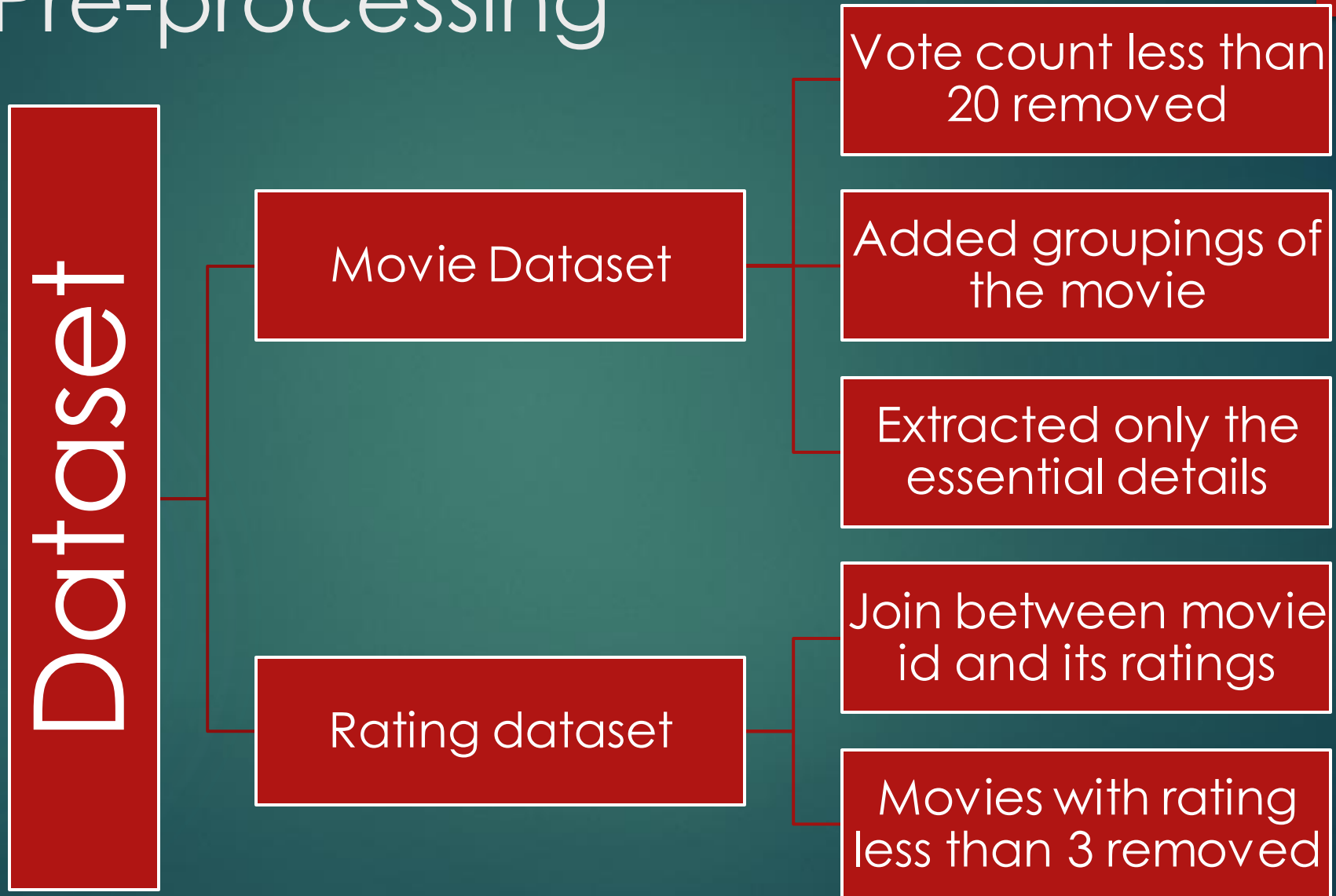
Snigdha Purohit

# About the Data

2



# Data Pre-processing



# Post-processed movies dataset samples

4

	A	B	C	D	E	F	G	H	I	J	K
1	budget	id	original_language	popularity	release_date	revenue	runtime	title	vote_average	vote_count	Vote_class
2	1000000	18206	en	2.175696	4/28/2006	10018	78	Alone With Her	6.2	20	HIGH
3	7000000	30128	en	7.155519	9/25/2009	1429299	106	I Hope They Serve Beer in Hell	5.7	20	HIGH
4	2600000	13064	en	8.001156	9/17/1993	2850263	91	Airborne	6.1	20	HIGH
5	20000000	13948	en	2.605374	10/3/2008	7000000	83	An American Carol	4.1	20	LOW
6	8000000	26306	en	8.717235	1/24/1997	589304	106	Prefontaine	6.7	21	HIGH
7	1000000	39833	en	1.444258	11/15/1956	9000000	89	Love Me Tender	5.5	21	HIGH
8	245000	3060	en	0.785744	11/5/1925	22000000	151	The Big Parade	7	21	HIGH
9	15000000	24206	en	2.609057	6/26/1996	76382	93	Silent Trigger	5	22	LOW
10	134005	58882	en	2.260315	7/15/2011	1332231	104	Snow Flower and the Secret Fan	5.3	22	HIGH
11	25000000	24264	en	4.020194	8/17/1984	5778353	117	Sheena	5	22	LOW
12	3768785	25209	en	1.726175	5/17/1950	8000000	107	Annie Get Your Gun	7.3	22	HIGH
13	4000000	40494	en	1.379354	5/30/1956	14400000	105	Trapeze	6.4	22	HIGH
14	800000	26899	en	5.292536	1/17/2003	62852	94	The Mudge Boy	7.3	23	HIGH

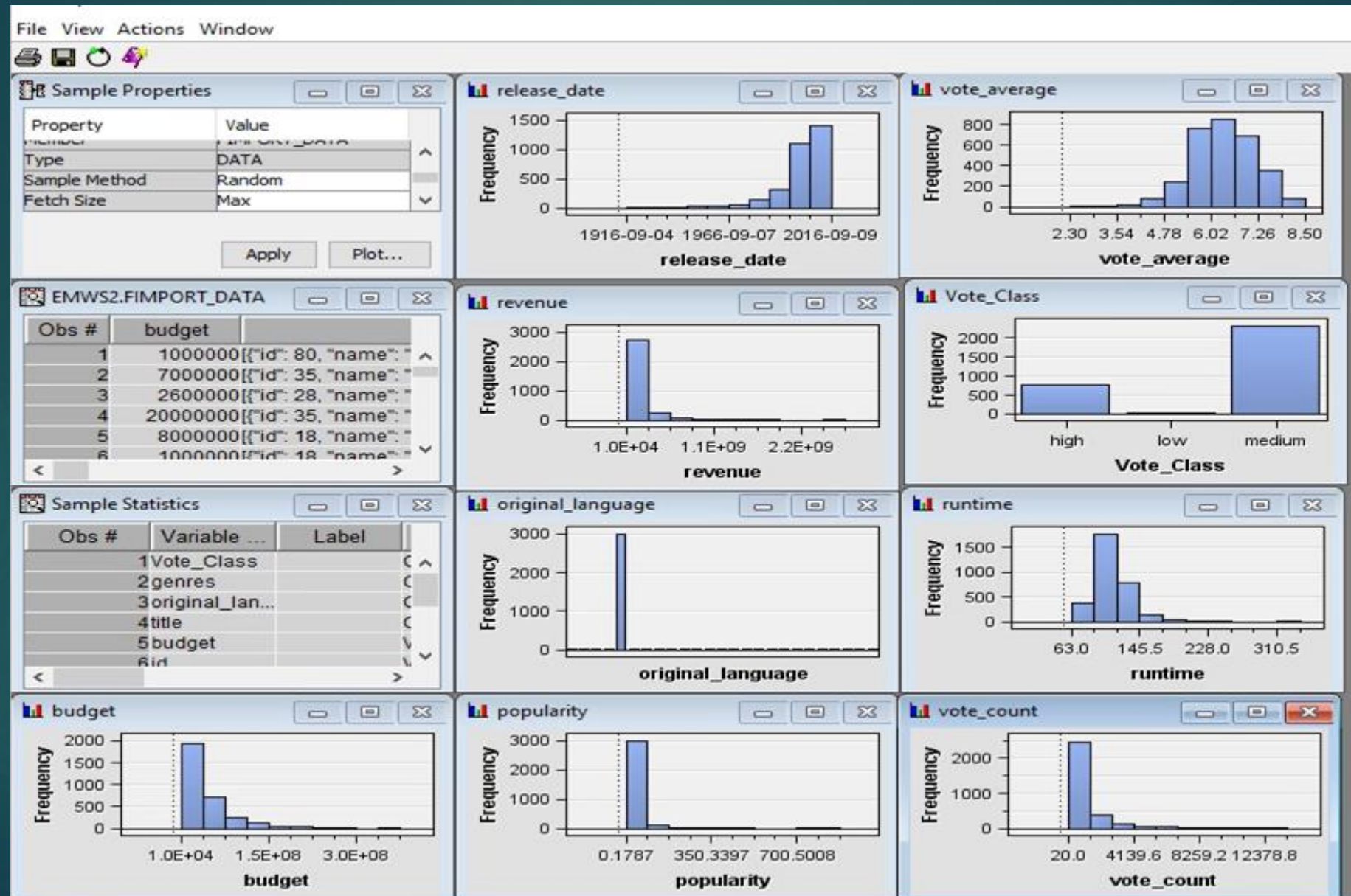
# Post-processed ratings dataset samples

userId	movieId	rating	title
1	2959	4	License to Wed
1	1246	5	Rocky Balboa
2	58	3	Pirates of the Caribbean: Dead Man's Chest
2	605	4	The Matrix Revolutions
2	25	3	Jarhead
2	628	4	Interview with the Vampire
2	377	4	A Nightmare on Elm Street
2	141	3	Donnie Darko
2	5	3	Four Rooms
2	762	3	Monty Python and the Holy Grail
3	4474	3	My Super Ex-Girlfriend
3	480	3	Monsoon Wedding
4	1911	4	The 13th Warrior
4	415	4	Batman & Robin
4	1422	4	The Departed
4	2023	5	Hidalgo
4	1597	3	Meet the Parents
4	1645	4	A Time to Kill
4	1909	4	Don Juan DeMarco
4	2355	4	Reign Over Me
4	1732	3	The Prisoner of Zenda
4	223	4	Rebecca
5	296	4	Terminator 3: Rise of the Machines
6	11	3	Star Wars



# Exploratory plots and tables

6



# Business Problem-1

7





popularity

Revenue

User-  
rating



Budget



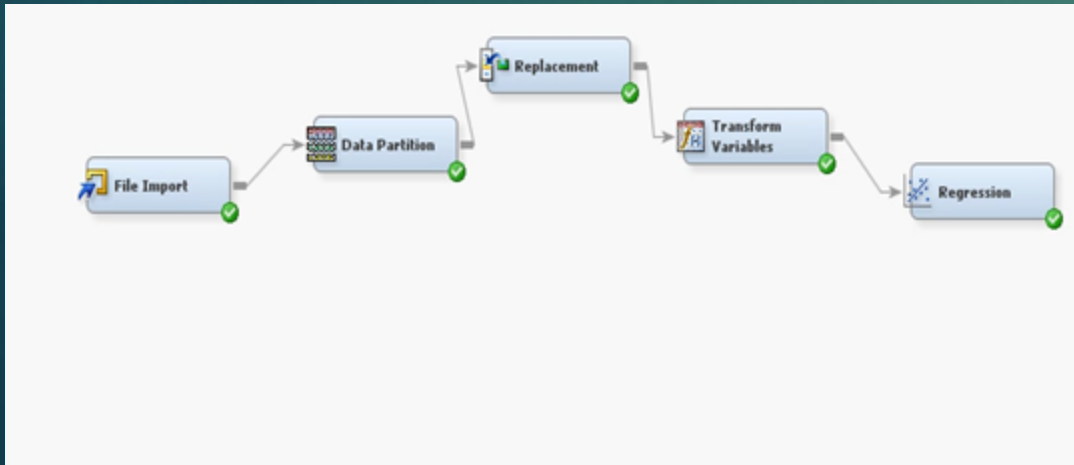


# Business Problem 1

9

**Problem Statement:** Predicting revenue of a movie from its attributes such as budget, runtime, popularity, Original language and vote average

**Algorithm:** Linear Regression



# Results from Analysis

10

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	16.0255	0.0814	196.83	<.0001
REP_budget	1	1.62E-8	7.82E-10	20.70	<.0001
REP_original_language 0	1	-0.3062	0.0769	-3.98	<.0001
REP_popularity	1	0.0250	0.00123	20.28	<.0001

**Observations:** Popularity and budget => Positive correlation with revenue  
Original language(not English) => Negative correlation with revenue

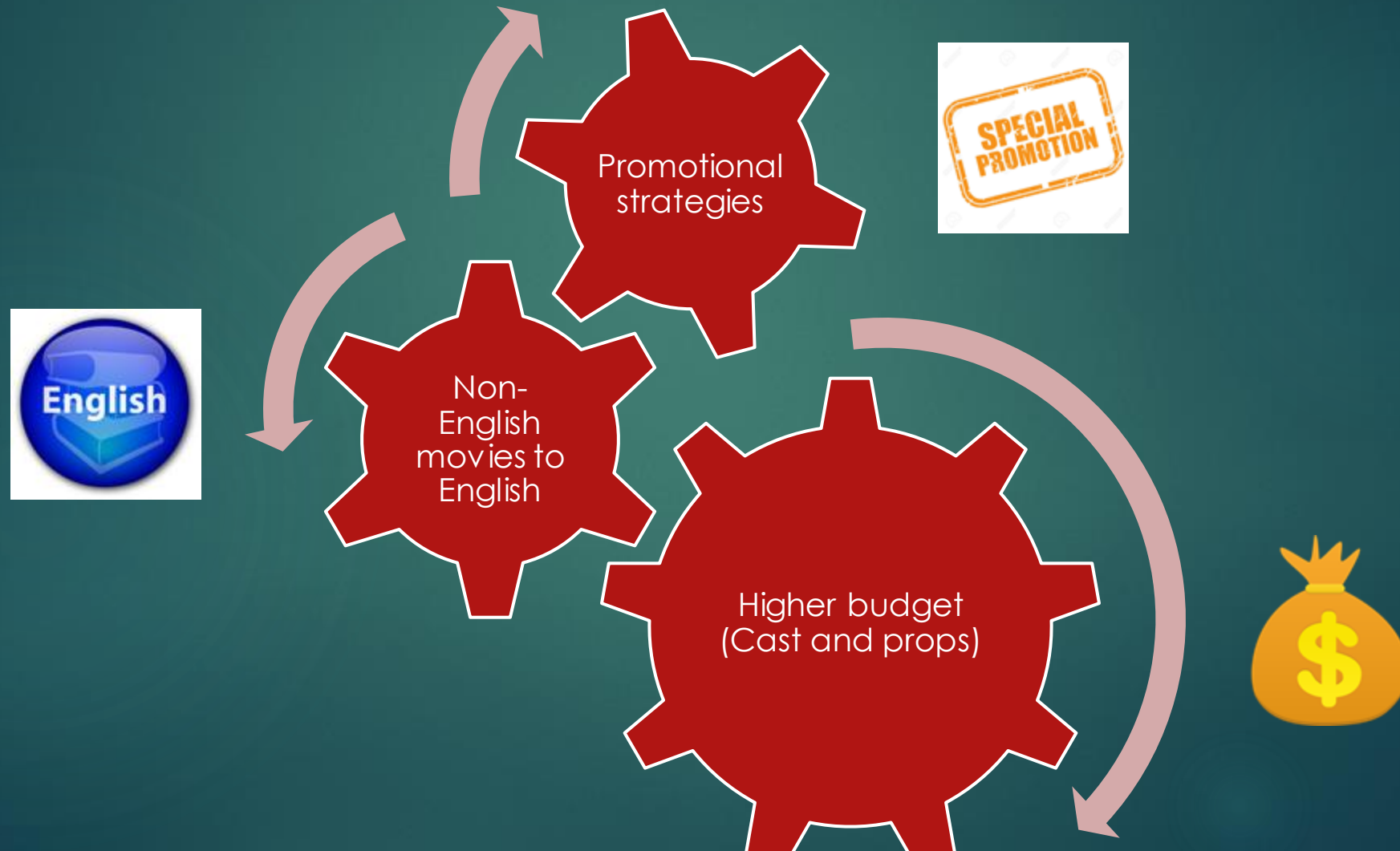
**Inference :** 1) Movies with higher budget => Better cast and props => likely to generate higher revenue

2) Movies with higher popularity => Likely to have higher revenue

3) Non-English movies => Likely to generate lesser revenue

# Business Strategies problem-1

11

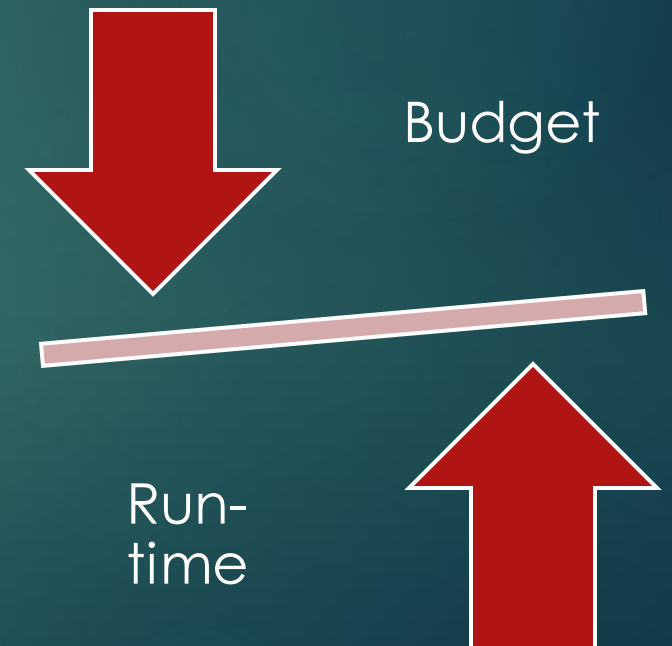


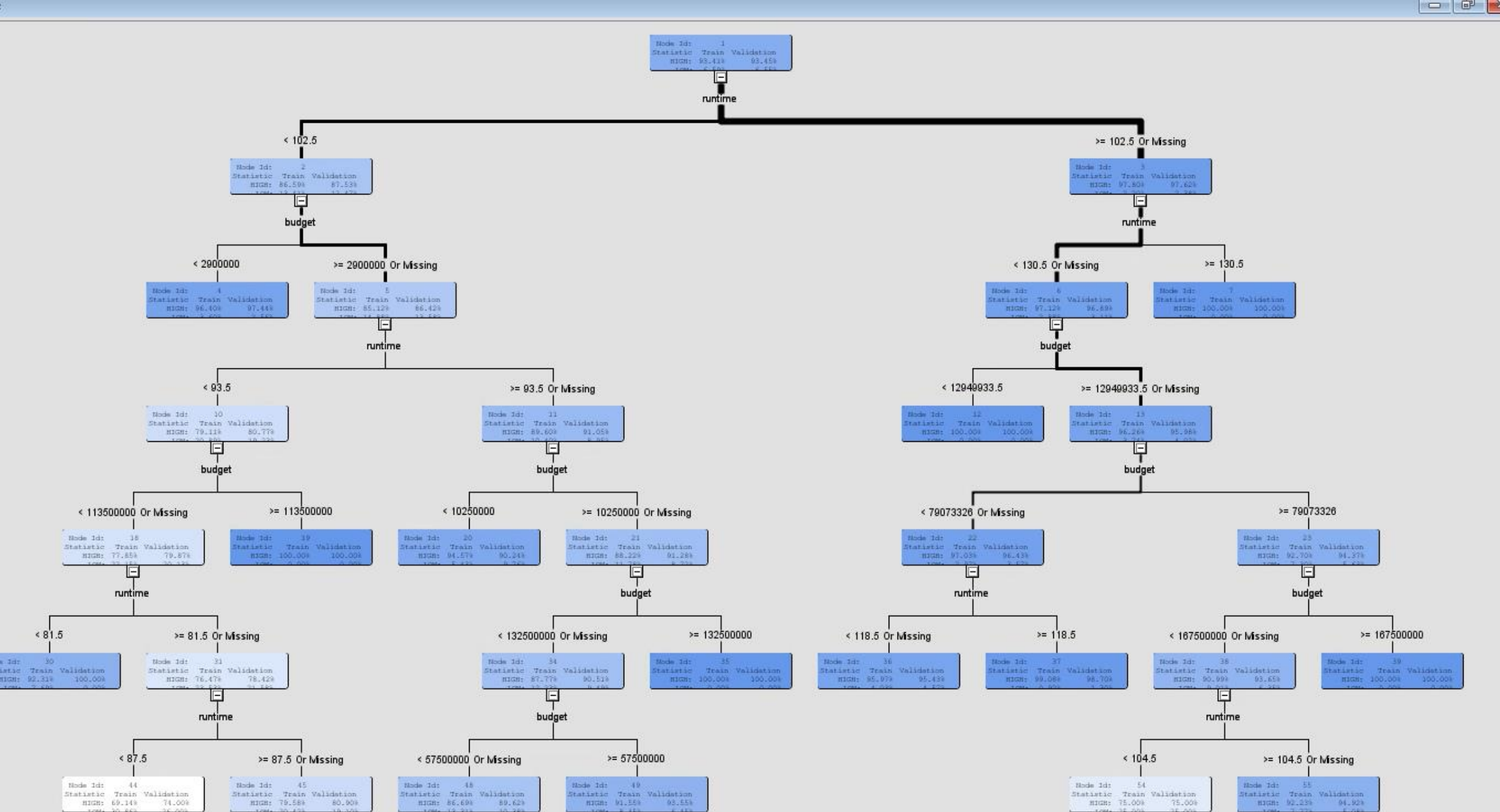
# Business Problem 2

12

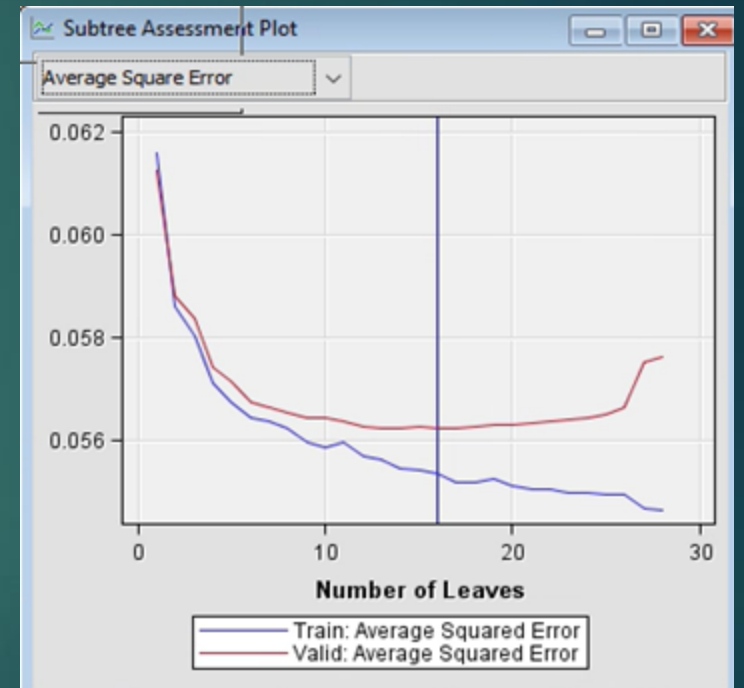
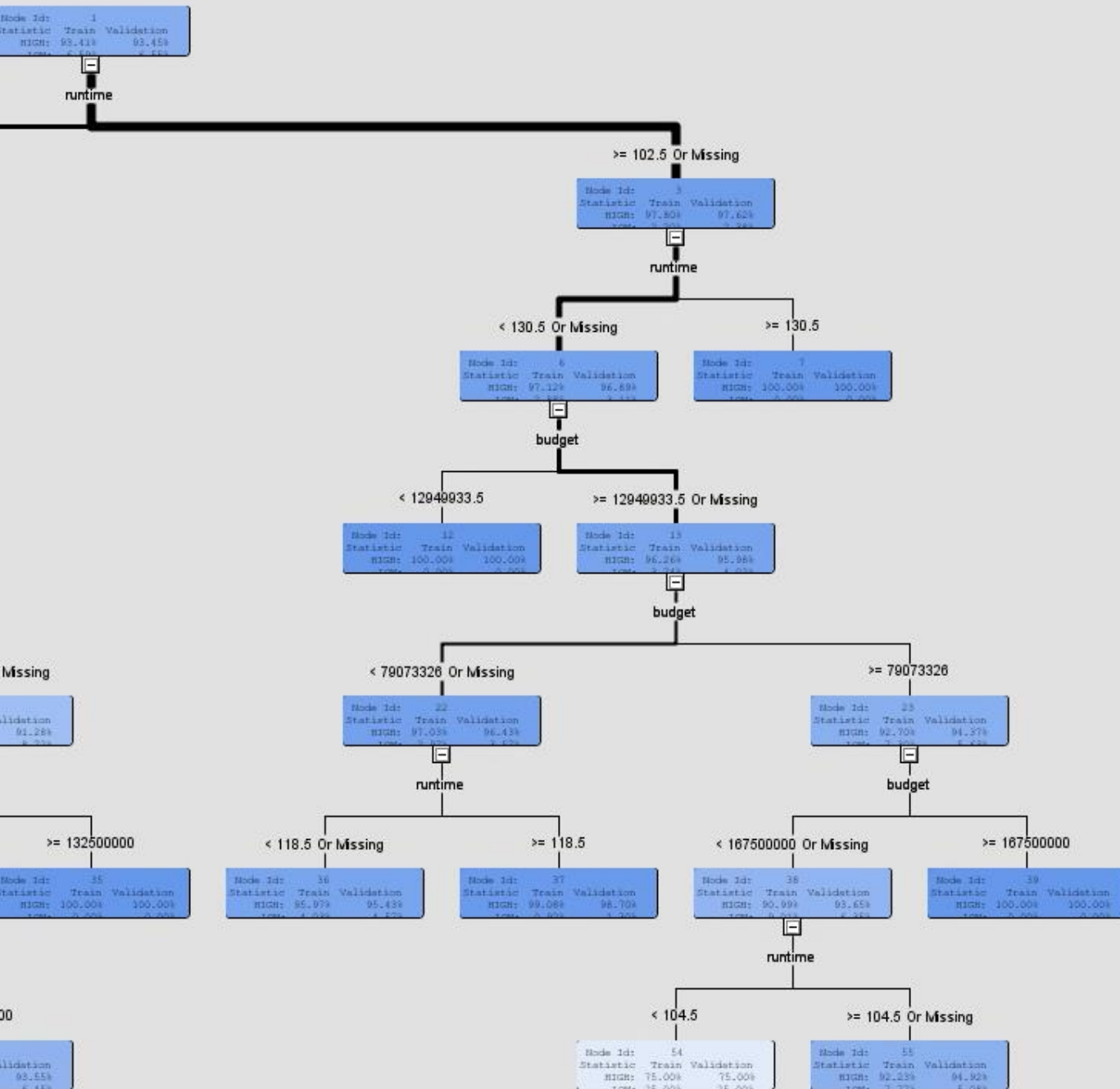
**Problem Statement:** Predicting the vote\_average/rating (HIGH or LOW) for a movie based on budget and runtime

**Algorithm:** Decision Tree









# High voting average movies have :

15

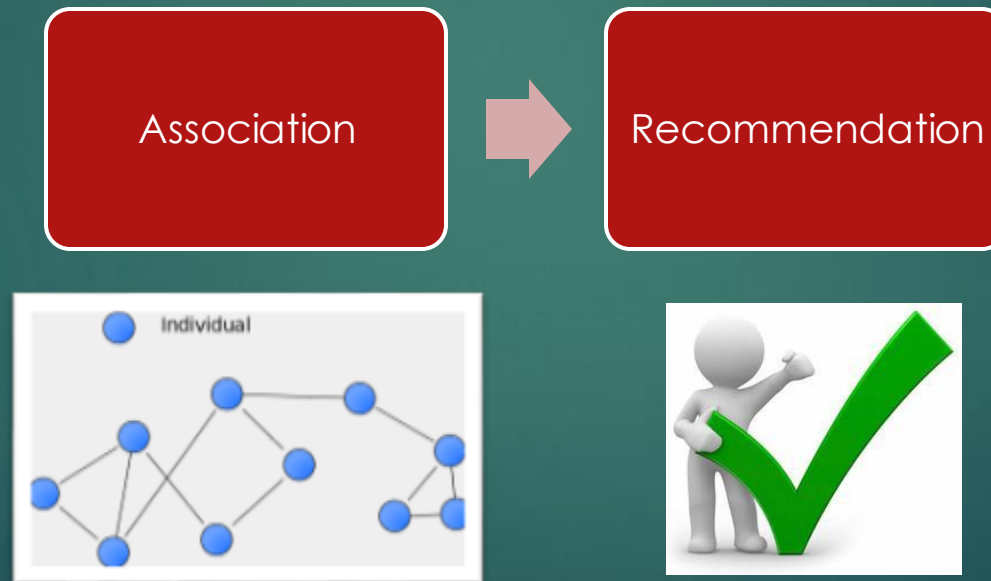
- ▶ Run-time between 102-130 minutes
- ▶ Budget between \$12.94 million and \$79.07 million

People tend to prefer movies which are not too short or too long!

Budget determines the cast and props of the movies. Better cast and props imply higher budget! => Higher voting average!

# Business Problem 3&4

16

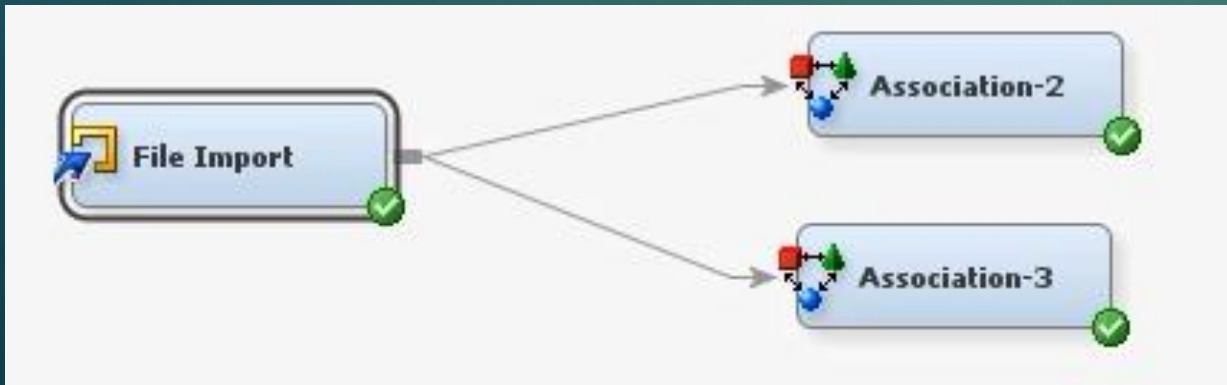
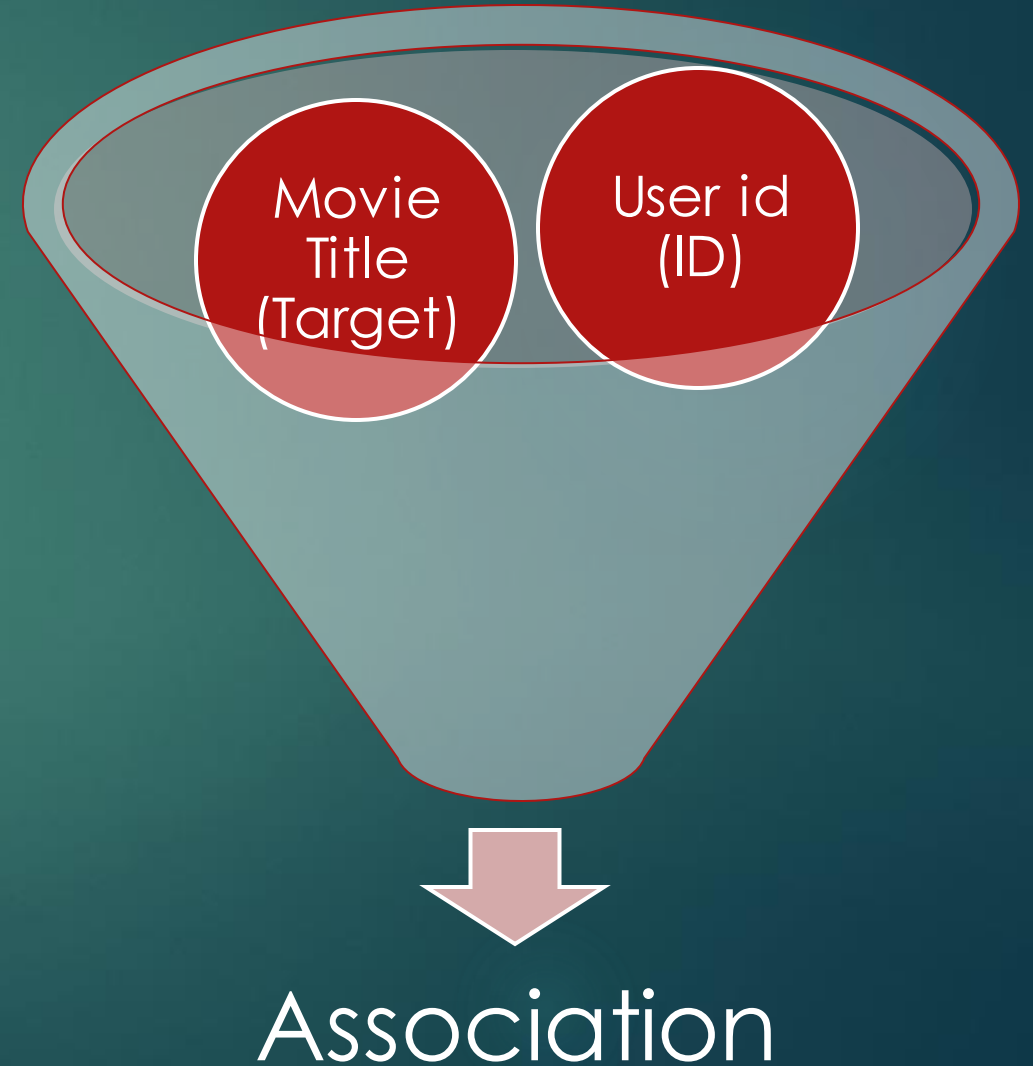


# Business Problem 3

17

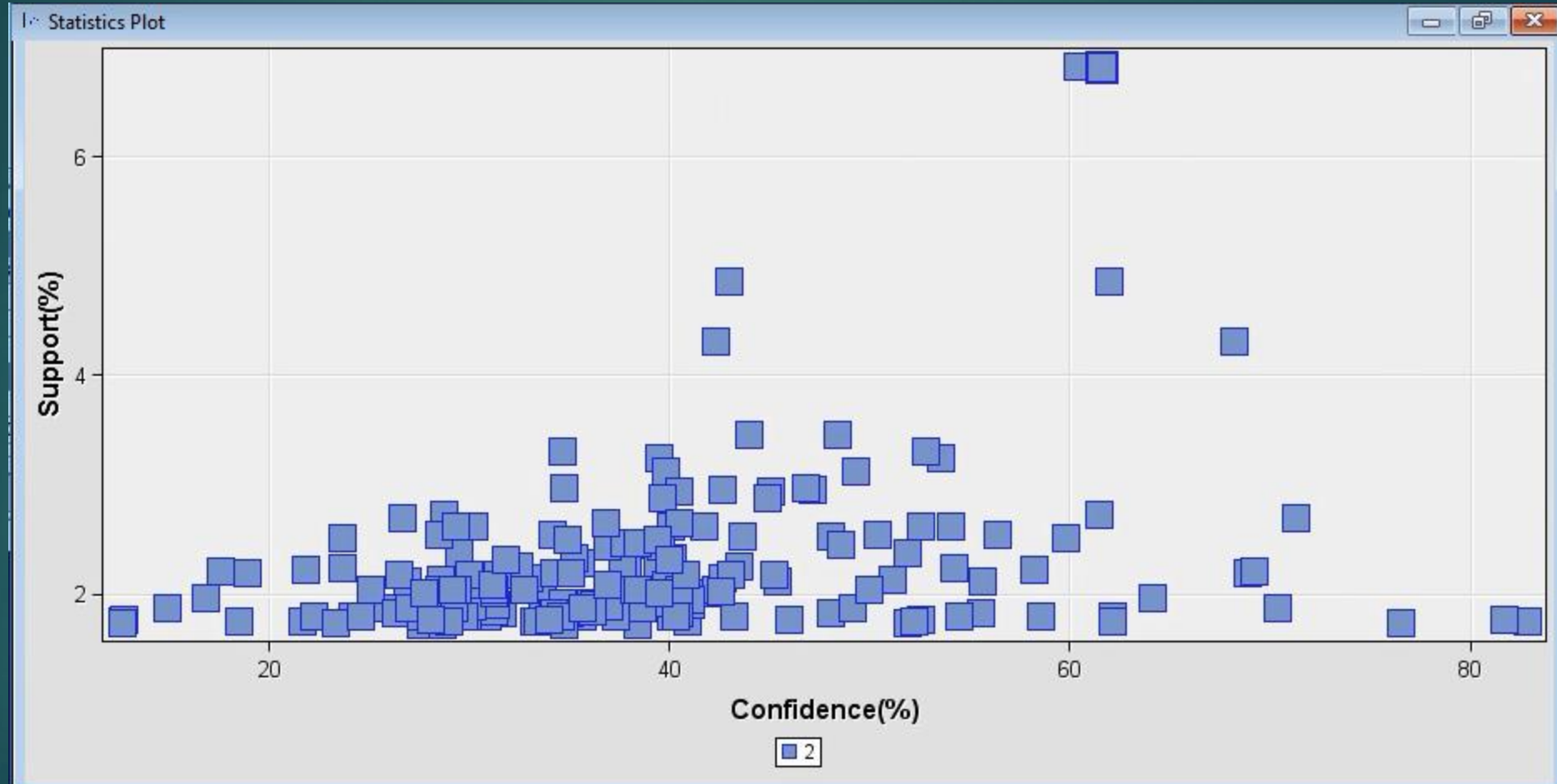
**Problem Statement:** Recommending users based on the movies they have previously rated and ratings by similar users.

**Algorithm:** Association Rule Mining



# Association Rule - 2 items

18





# Log of the results

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
2	4.75	55.52	1.82	11.68	171.00	Modern Times ==> Captain Corelli's Mandolin
2	3.28	38.34	1.82	11.68	171.00	Captain Corelli's Mandolin ==> Modern Times
2	5.05	43.22	1.80	8.56	169.00	X-Men Origins: Wolverine ==> Dogtown and Z-Boys
2	4.17	35.65	1.80	8.56	169.00	Dogtown and Z-Boys ==> X-Men Origins: Wolverine
2	3.35	27.92	1.76	8.34	165.00	Stand by Me ==> Alien
2	6.30	52.55	1.76	8.34	165.00	Alien ==> Stand by Me
2	7.45	62.13	1.80	8.34	169.00	The Wedding Planner ==> The Terminal
2	2.90	24.18	1.80	8.34	169.00	The Terminal ==> The Wedding Planner
2	5.79	48.07	2.53	8.31	237.00	Stranger Than Fiction ==> Ghost Rider
2	5.25	43.65	2.53	8.31	237.00	Ghost Rider ==> Stranger Than Fiction
2	5.79	48.04	1.83	8.30	172.00	Sunshine ==> Ghost Rider
2	3.82	31.68	1.83	8.30	172.00	Ghost Rider ==> Sunshine
2	4.81	38.49	2.01	8.01	189.00	Rosemary's Baby ==> Lolita
2	5.23	41.91	2.01	8.01	189.00	Lolita ==> Rosemary's Baby
2	5.25	40.90	1.75	7.78	164.00	Back to the Future Part III ==> 20,000 Leagues Under the Sea
2	4.27	33.27	1.75	7.78	164.00	20,000 Leagues Under the Sea ==> Back to the Future Part III
2	6.26	48.52	2.45	7.76	230.00	We Own the Night ==> Rambo III
2	5.05	39.18	2.45	7.76	230.00	Rambo III ==> We Own the Night
2	2.81	21.69	1.75	7.71	164.00	K-19: The Widowmaker ==> Brothers
2	8.06	62.12	1.75	7.71	164.00	Brothers ==> K-19: The Widowmaker
2	5.21	39.71	2.08	7.62	195.00	Rosemary's Baby ==> Interview with the Vampire
2	5.23	39.88	2.08	7.62	195.00	Interview with the Vampire ==> Rosemary's Baby
2	3.30	24.18	1.80	7.32	169.00	The Terminal ==> The Last Samurai
2	7.45	54.52	1.80	7.32	169.00	The Last Samurai ==> The Terminal
2	5.21	37.92	1.82	7.28	171.00	Lolita ==> Interview with the Vampire

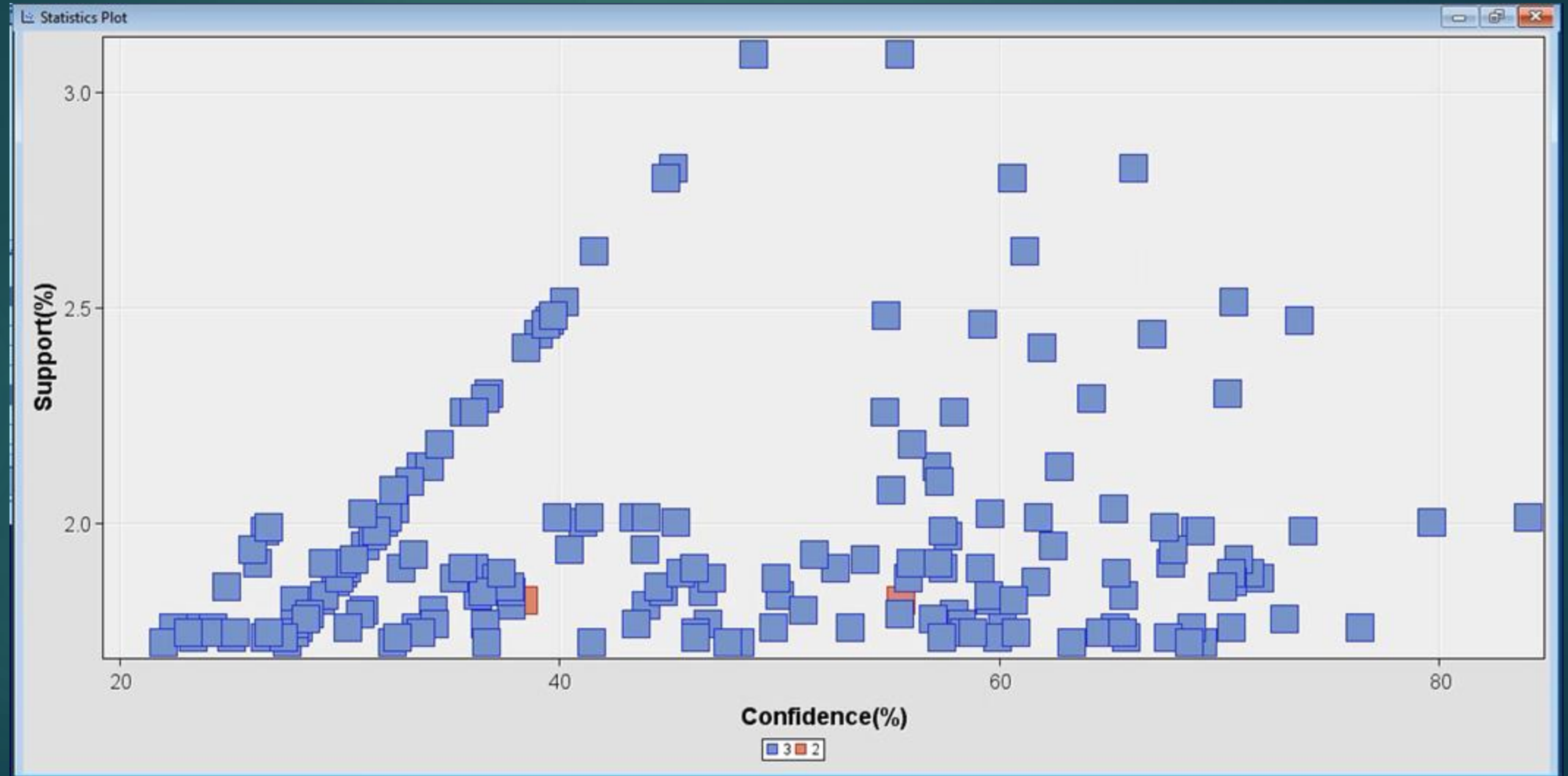
# Movies with highest association (2-way)

20



# Association Rule - 3 items

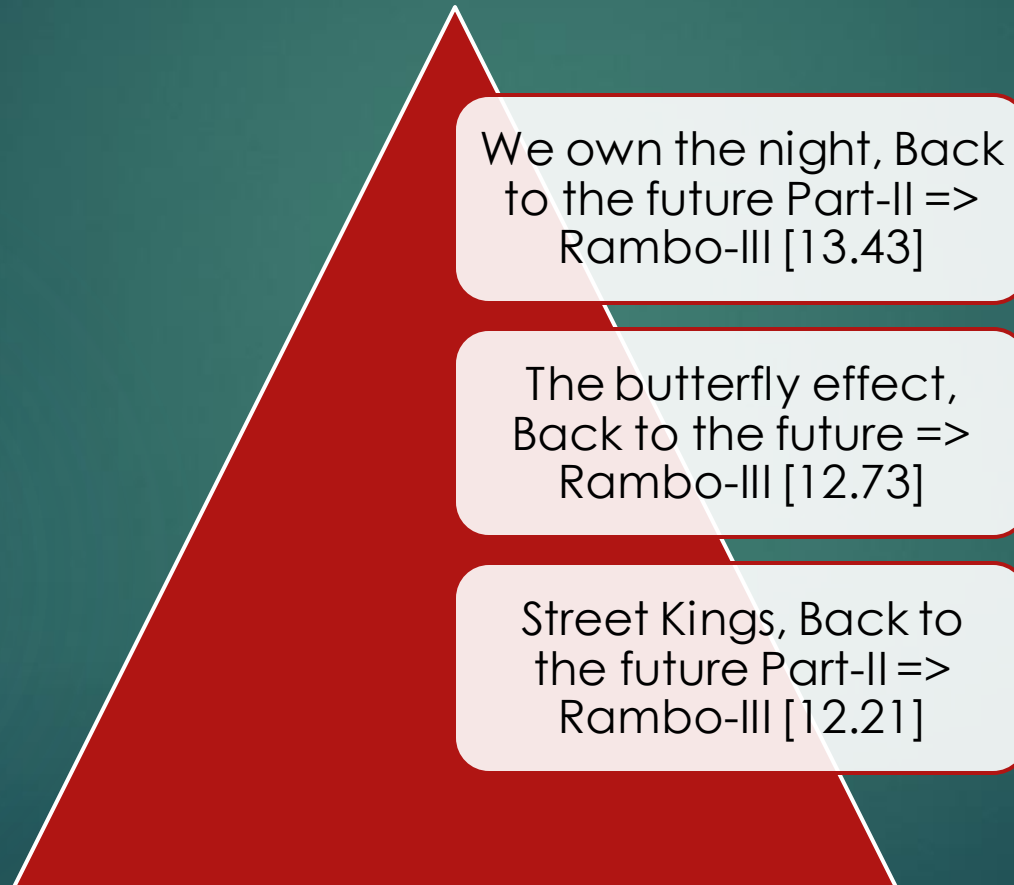
21



Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
3	6.26	84.00	2.01	13.43	189.00	We Own the Night & Back to the Future Part II ==> Rambo III
3	2.40	32.20	2.01	13.43	189.00	Rambo III ==> We Own the Night & Back to the Future Part II
3	6.26	79.66	2.00	12.73	188.00	The Butterfly Effect & Back to the Future Part II ==> Rambo III
3	2.51	32.03	2.00	12.73	188.00	Rambo III ==> The Butterfly Effect & Back to the Future Part II
3	6.26	76.39	1.76	12.21	165.00	Street Kings & Back to the Future Part II ==> Rambo III
3	2.30	28.11	1.76	12.21	165.00	Rambo III ==> Street Kings & Back to the Future Part II
3	6.26	73.81	1.98	11.80	186.00	The Thirteenth Floor & Back to the Future Part II ==> Rambo III
3	2.69	31.69	1.98	11.80	186.00	Rambo III ==> The Thirteenth Floor & Back to the Future Part II
3	3.36	39.52	2.47	11.77	232.00	Rambo III ==> Die Hard 2 & Back to the Future Part II
3	6.26	73.65	2.47	11.77	232.00	Die Hard 2 & Back to the Future Part II ==> Rambo III
2	4.75	55.52	1.82	11.68	171.00	Modern Times ==> Captain Corelli's Mandolin
2	3.28	38.34	1.82	11.68	171.00	Captain Corelli's Mandolin ==> Modern Times
3	6.26	72.93	1.78	11.66	167.00	The Golden Compass & Back to the Future Part II ==> Rambo III
3	2.44	28.45	1.78	11.66	167.00	Rambo III ==> The Golden Compass & Back to the Future Part II
3	6.26	71.84	1.88	11.48	176.00	We Own the Night & Rain Man ==> Rambo III
3	2.61	29.98	1.88	11.48	176.00	Rambo III ==> We Own the Night & Rain Man
3	6.26	71.37	1.89	11.41	177.00	We Own the Night & A Nightmare on Elm Street ==> Rambo III
3	2.64	30.15	1.89	11.41	177.00	Rambo III ==> We Own the Night & A Nightmare on Elm Street
3	6.26	70.87	1.92	11.33	180.00	Rocky Balboa & Back to the Future Part II ==> Rambo III
3	2.71	30.66	1.92	11.33	180.00	Rambo III ==> Rocky Balboa & Back to the Future Part II
3	3.56	40.20	2.51	11.30	236.00	Rambo III ==> Fever Pitch & Back to the Future Part II
3	6.26	70.66	2.51	11.30	236.00	Fever Pitch & Back to the Future Part II ==> Rambo III
3	2.64	29.81	1.86	11.28	175.00	Rambo III ==> Notes on a Scandal & Back to the Future Part II
3	6.26	70.56	1.86	11.28	175.00	Notes on a Scandal & Back to the Future Part II ==> Rambo III
3	6.26	70.52	1.89	11.27	177.00	The Prisoner of Zenda & Back to the Future Part II ==> Rambo III

# Movies with highest association (3-way)

23





# Business problem 4:

24

Finding a set of similar movies based on parameters such as revenue, budget, run-time, popularity etc.

Method: Clustering



Columns: <input type="checkbox"/> Label <input type="checkbox"/> Mining						
Name	Role /	Level	Report	Order	Drop	
id	ID	Nominal	No		No	
budget	Input	Interval	No		No	
runtime	Input	Interval	No		No	
vote_average	Input	Interval	No		No	
revenue	Input	Interval	No		No	
vote_count	Input	Interval	No		No	
year	Input	Interval	No		No	
popularity	Input	Interval	No		No	
genres	Rejected	Nominal	No		No	
Vote_Class	Rejected	Nominal	No		No	
day	Rejected	Interval	No		No	
title	Rejected	Nominal	No		No	
original_language	Rejected	Nominal	No		No	
month	Rejected	Interval	No		No	

# Segment-1

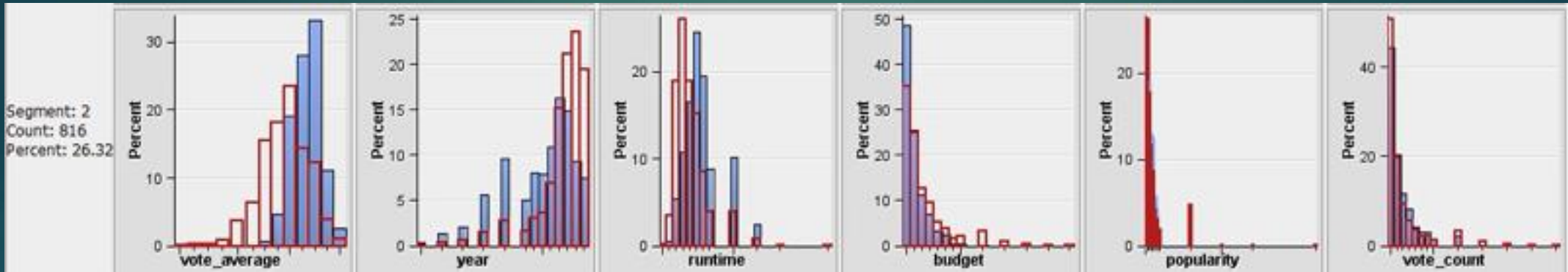
25



- Segment 1: higher revenue, higher budget, higher vote count and higher popularity ( popular movies)

# Segment-2

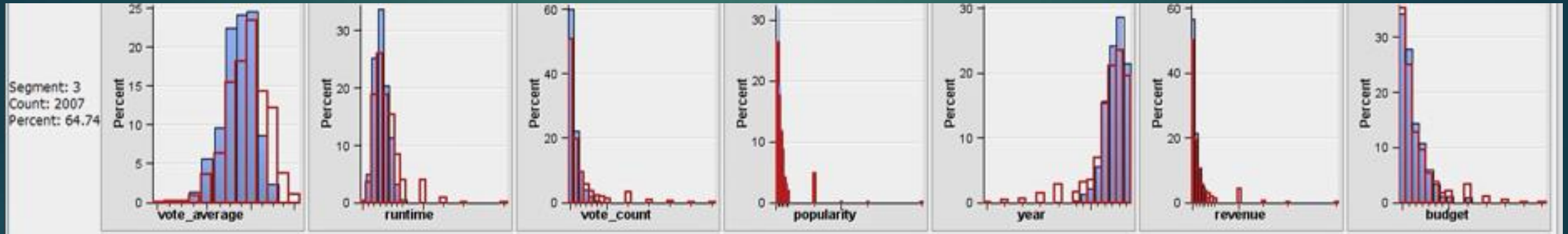
26



- ▶ segment 2: lower budget, higher vote average, lower popularity (critically acclaimed but not well promoted movie)

# Segment-3

27



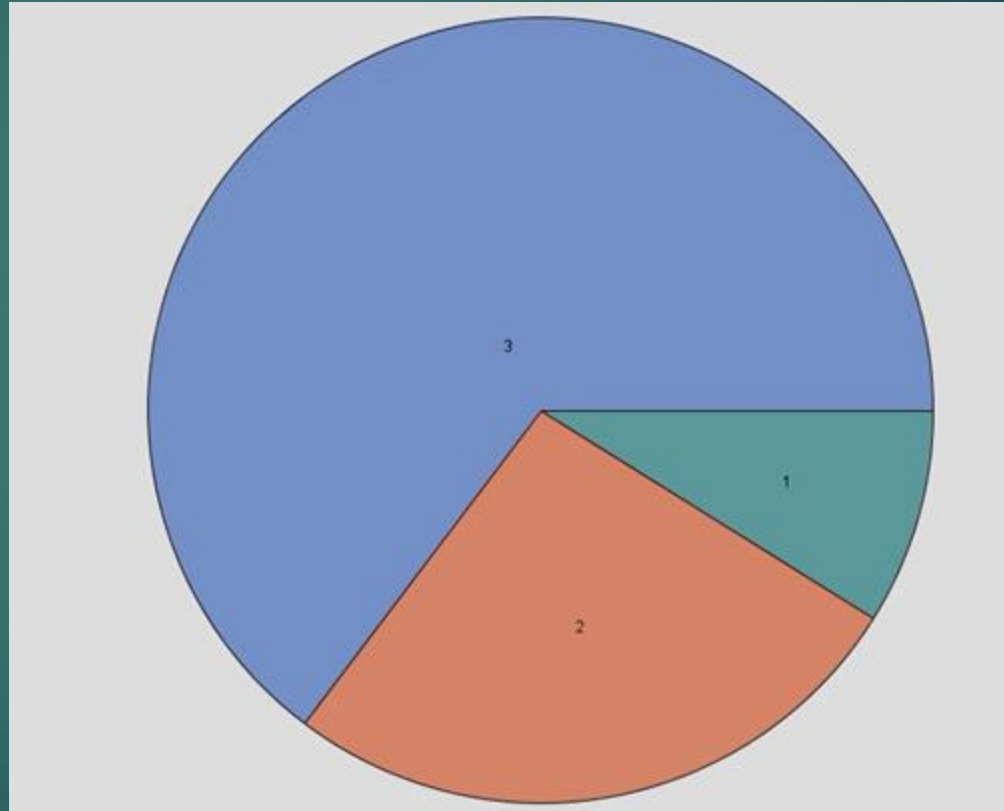
- ▶ segment 3: lower popularity, lower budget, lower vote average, comparable or lower revenue

# clustering results - pie charts

segment 1: 8.94%  
segment 2: 26.32%  
segment 3: 64.74%

Paradox Effect!

Only 8.94% movies are popular.  
Most of us watch popular movies  
So we think all the movies popular  
BECAUSE we watch Popular movies





# Results and Business strategies

29

- Movie recommendations based on user viewing habits

title	Segment Descript...
Waterworld	Cluster1
Die Another Day	Cluster1
A Christmas Carol	Cluster1
The Golden Compass	Cluster1
Penguins of Madagascar	Cluster1
The Mummy: Tomb of the Dragon E...	Cluster1
Robin Hood	Cluster1
Superman Returns	Cluster1
Monsters vs Aliens	Cluster1
The Polar Express	Cluster1
The Chronicles of Narnia: The Voya...	Cluster1
Home	Cluster1
Kung Fu Panda 3	Cluster1
The Chronicles of Narnia: Prince Ca...	Cluster1
Alice Through the Looking Glass	Cluster1
The Good Dinosaur	Cluster1
Pearl Harbor	Cluster1
Madagascar 3: Europe's Most Want...	Cluster1
Madagascar: Escape 2 Africa	Cluster1
Night at the Museum: Secret of the T...	Cluster1
Kung Fu Panda 2	Cluster1
The Last Samurai	Cluster1
Megamind	Cluster1
Exodus: Gods and Kings	Cluster1
Rise of the Guardians	Cluster1
National Treasure	Cluster1
Mission: Impossible II	Cluster1
Shrek Forever After	Cluster1

title	Segment Descript...
2001: A Space Odyssey	Cluster2
No Country for Old Men	Cluster2
A Beautiful Mind	Cluster2
American History X	Cluster2
Pan's Labyrinth	Cluster2
Into the Wild	Cluster2
The Notebook	Cluster2
Prisoners	Cluster2
Gran Torino	Cluster2
The Devil Wears Prada	Cluster2
The Conjuring	Cluster2
GoodFellas	Cluster2
Aliens	Cluster2
The Usual Suspects	Cluster2
Amélie	Cluster2
The Theory of Everything	Cluster2
American Beauty	Cluster2
Braveheart	Cluster2
The Godfather: Part II	Cluster2
District 9	Cluster2
The Social Network	Cluster2
Nightcrawler	Cluster2
Argo	Cluster2
Donnie Darko	Cluster2
Blade Runner	Cluster2
The Bourne Identity	Cluster2
Edward Scissorhands	Cluster2
Eternal Sunshine of the Spotless Mi	Cluster2

title	Segment Descript..
Alone With Her	Cluster3
I Hope They Serve Beer in Hell	Cluster3
Airborne	Cluster3
An American Carol	Cluster3
Prefontaine	Cluster3
Silent Trigger	Cluster3
Snow Flower and the Secret Fan	Cluster3
Sheena	Cluster3
The Mudge Boy	Cluster3
No End in Sight	Cluster3
Murderball	Cluster3
Lovely & Amazing	Cluster3
From Justin to Kelly	Cluster3
The Inhabited Island	Cluster3
10th & Wolf	Cluster3
Gun Shy	Cluster3
Two Girls and a Guy	Cluster3
Home Run	Cluster3
Cats Don't Dance	Cluster3
The Lost City	Cluster3
Connie and Carla	Cluster3
How Stella Got Her Groove Back	Cluster3
The Monkey King 2	Cluster3
Surfer, Dude	Cluster3
The Broken Hearts Club: A Romanti...	Cluster3
Mrs. Winterbourne	Cluster3
Spaced Invaders	Cluster3
Prophecy	Cluster3

THANK YOU  
QUESTIONS?