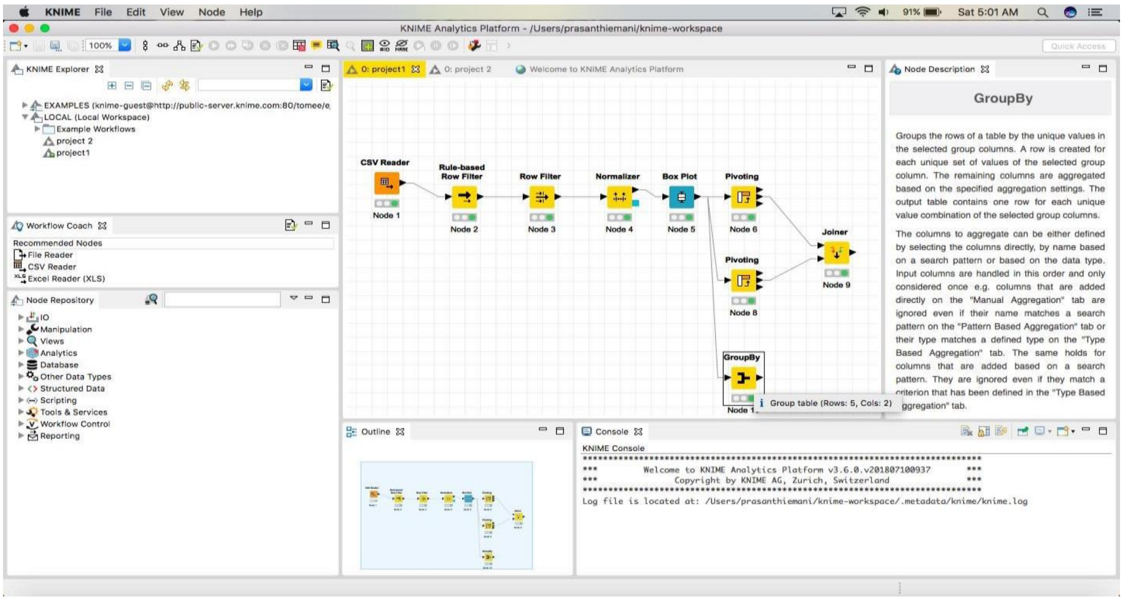
EXPERIMENT-15

EXTRACT TRANSFORM LOAD (ETL) AND OLAP OPERATION USING KNIME TOOL

OUTPUT:



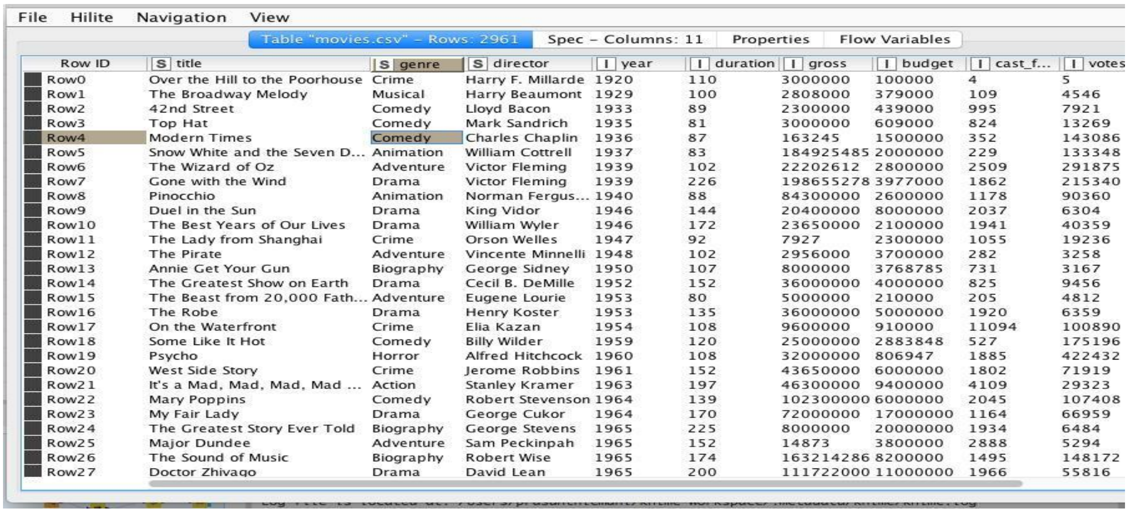
GroupBy

Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns.

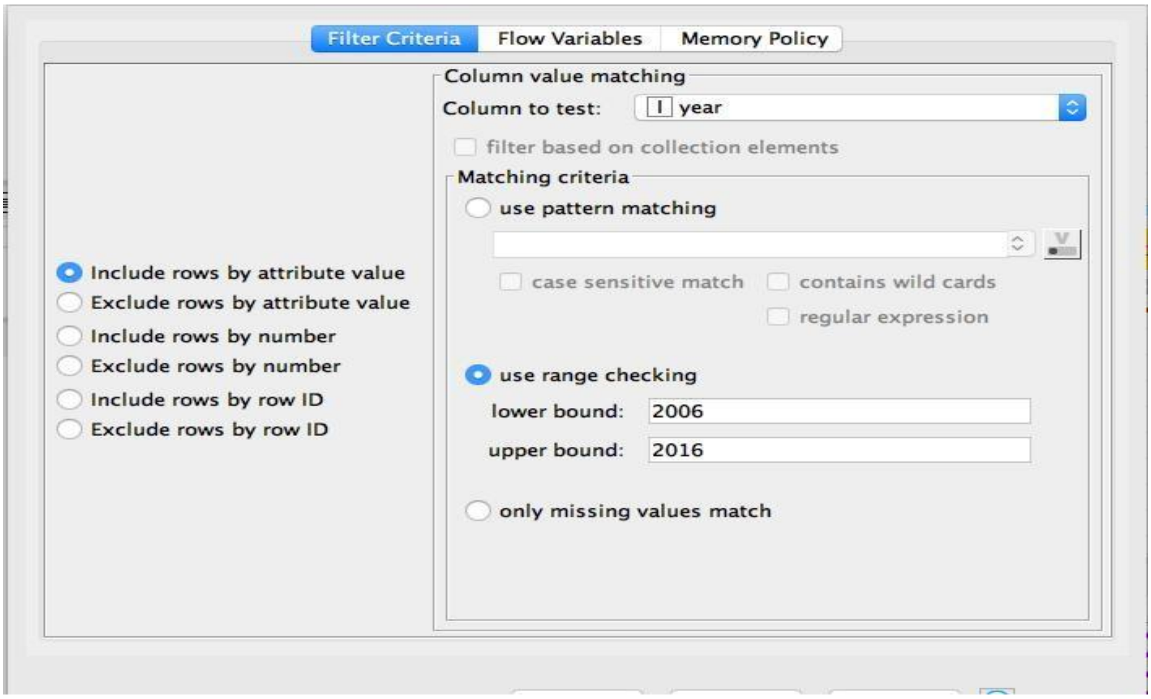
The columns to aggregate can be either defined by selecting the columns directly, by name based on a search pattern or based on the data type. Input columns are handled in this order and only considered once e.g. columns that are added directly on the "Manual Aggregation" tab are ignored even if their name matches a search pattern on the "Pattern Based Aggregation" tab or their type matches a defined type on the "Type Based Aggregation" tab. The same holds for columns that are added based on a search pattern. They are ignored even if they match a criterion that has been defined in the "Type Based Aggregation" tab.

KNIME Console

```
*****************************************************************************
***        Welcome to KNIME Analytics Platform v3.6.0.v201807100937       ***
***              Copyright by KNIME AG, Zurich, Switzerland               ***
*****************************************************************************
Log file is located at: /Users/prasanthiemani/knime-workspace/.metadata/knime/knime.log
```



Row ID	S title	S genre	S director	I year	I duration	I gross	I budget	I cast_f...	I votes
Row0	Over the Hill to the Poorhouse	Crime	Harry F. Millarde	1920	110	3000000	100000	4	5
Row1	The Broadway Melody	Musical	Harry Beaumont	1929	100	2808000	379000	109	4546
Row2	42nd Street	Comedy	Lloyd Bacon	1933	89	2300000	439000	995	7921
Row3	Top Hat	Comedy	Mark Sandrich	1935	81	3000000	609000	824	13269
Row4	Modern Times	Comedy	Charles Chaplin	1936	87	163245	1500000	352	143086
Row5	Snow White and the Seven D...	Animation	William Cottrell	1937	83	184925485	2000000	229	133348
Row6	The Wizard of Oz	Adventure	Victor Fleming	1939	102	22202612	2800000	2509	291875
Row7	Gone with the Wind	Drama	Victor Fleming	1939	226	198655278	3977000	1862	215340
Row8	Pinocchio	Animation	Norman Fergus...	1940	88	84300000	2600000	1178	90360
Row9	Duel in the Sun	Drama	King Vidor	1946	144	20400000	8000000	2037	6304
Row10	The Best Years of Our Lives	Drama	William Wyler	1946	172	23650000	2100000	1941	40359
Row11	The Lady from Shanghai	Crime	Orson Welles	1947	92	7927	2300000	1055	19236
Row12	The Pirate	Adventure	Vincente Minnelli	1948	102	2956000	3700000	282	3258
Row13	Annie Get Your Gun	Biography	George Sidney	1950	107	8000000	3768785	731	3167
Row14	The Greatest Show on Earth	Drama	Cecil B. DeMille	1952	152	36000000	4000000	825	9456
Row15	The Beast from 20,000 Fath...	Adventure	Eugene Lourie	1953	80	5000000	210000	205	4812
Row16	The Robe	Drama	Henry Koster	1953	135	36000000	5000000	1920	6359
Row17	On the Waterfront	Crime	Elia Kazan	1954	108	9600000	910000	11094	100890
Row18	Some Like It Hot	Comedy	Billy Wilder	1959	120	25000000	2883848	527	175196
Row19	Psycho	Horror	Alfred Hitchcock	1960	108	32000000	806947	1885	422432
Row20	West Side Story	Crime	Jerome Robbins	1961	152	43650000	6000000	1802	71919
Row21	It's a Mad, Mad, Mad, Mad ...	Action	Stanley Kramer	1963	197	46300000	9400000	4109	29323
Row22	Mary Poppins	Comedy	Robert Stevenson	1964	139	102300000	6000000	2045	107408
Row23	My Fair Lady	Drama	George Cukor	1964	170	72000000	17000000	1164	66959
Row24	The Greatest Story Ever Told	Biography	George Stevens	1965	225	8000000	20000000	1934	6484
Row25	Major Dundee	Adventure	Sam Peckinpah	1965	152	14873	3800000	2888	5294
Row26	The Sound of Music	Biography	Robert Wise	1965	174	163214286	8200000	1495	148172
Row27	Doctor Zhivago	Drama	David Lean	1965	200	111722000	11000000	1966	55816



● Include rows by attribute value
○ Exclude rows by attribute value
○ Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Column value matching

Column to test: I year

□ filter based on collection elements

Matching criteria

○ use pattern matching

□ case sensitive match □ contains wild cards
□ regular expression

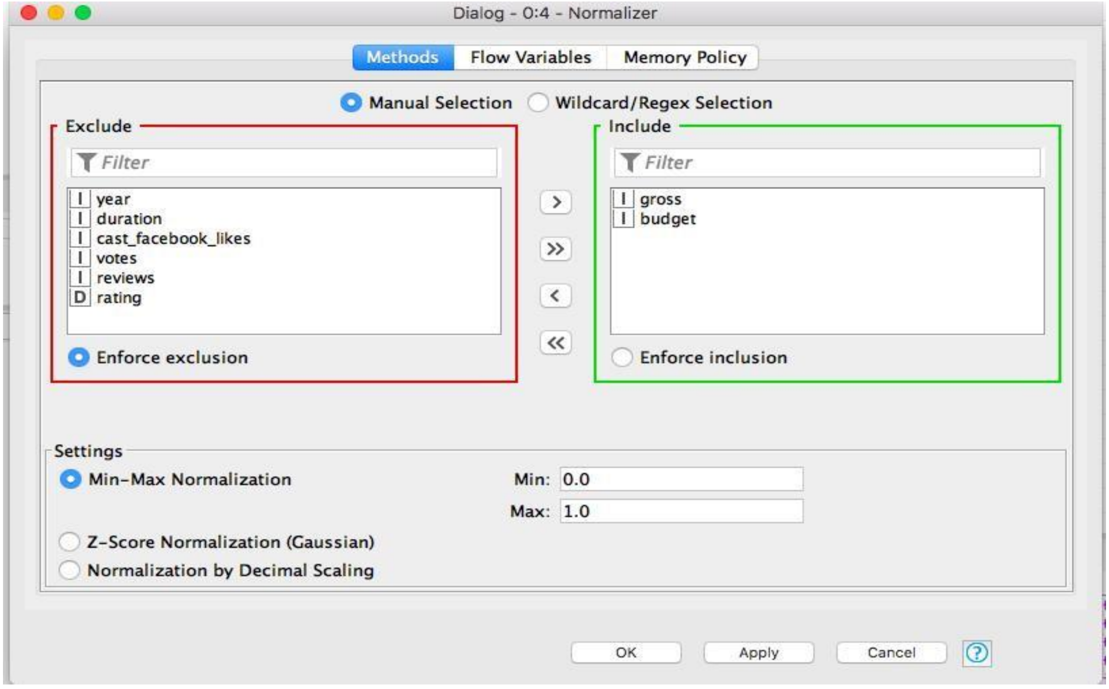
● use range checking

lower bound: 2006

upper bound: 2016

○ only missing values match

NORMALIZER:



● Manual Selection ○ Wildcard/Regex Selection

Exclude

I year
I duration
I cast_facebook_likes
I votes
I reviews
D rating

● Enforce exclusion

Include

I gross
I budget

○ Enforce inclusion

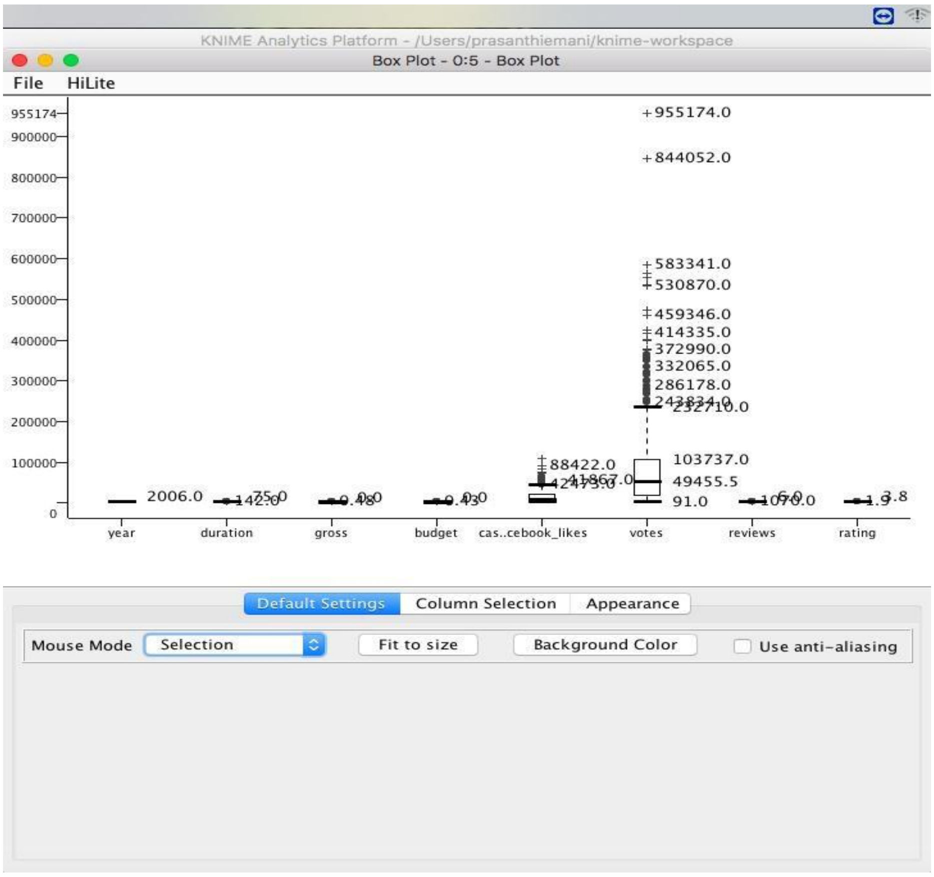
Settings

● Min-Max Normalization Min: 0.0 Max: 1.0

○ Z-Score Normalization (Gaussian)

○ Normalization by Decimal Scaling

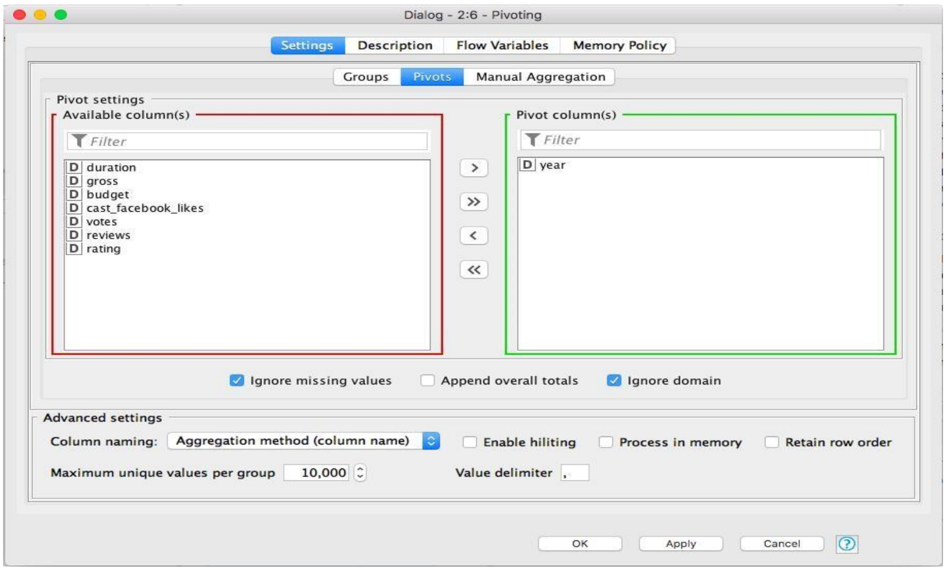
BOXPLOT:



Mouse Mode: Selection

□ Use anti-aliasing

PIVOTING:



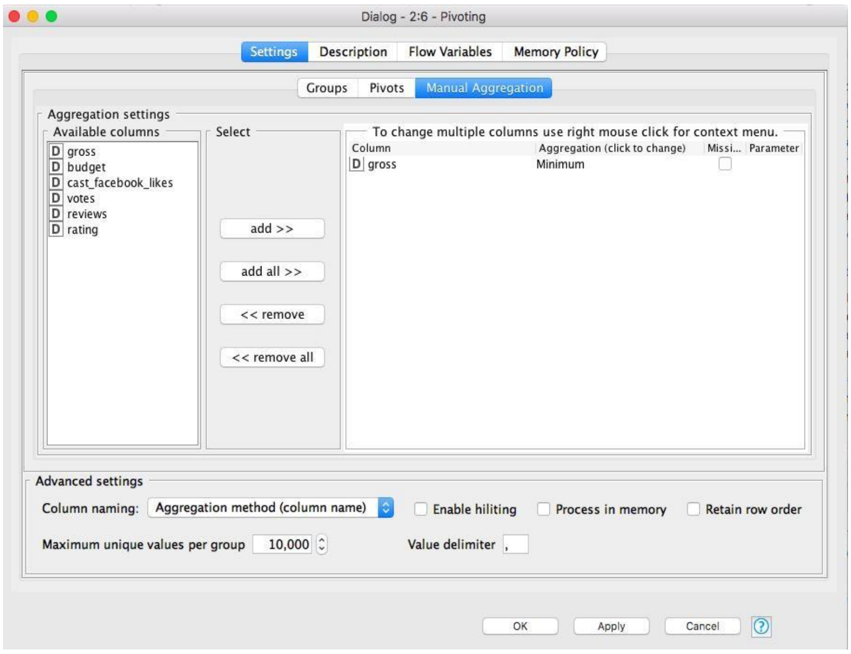
Available column(s):
D duration
D gross
D budget
D cast_facebook_likes
D votes
D reviews
D rating

Pivot column(s):
D year

☑ Ignore missing values □ Append overall totals ☑ Ignore domain

Column naming: Aggregation method (column name) □ Enable hiliting □ Process in memory □ Retain row order

Maximum unique values per group 10,000 Value delimiter ,



Dialog - 2:6 - Pivoting

Settings | Description | Flow Variables | Memory Policy

Groups | Pivots | Manual Aggregation

Aggregation settings

Available columns:
- D gross
- D budget
- D cast_facebook_likes
- D votes
- D reviews
- D rating

Select: add >> | add all >> | << remove | << remove all

To change multiple columns use right mouse click for context menu.

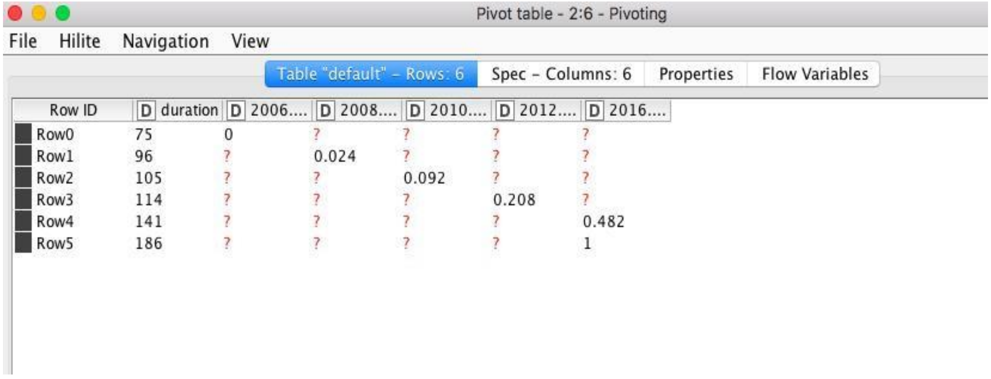
Column	Aggregation (click to change)	Missi...	Parameter
D gross	Minimum	□	

Advanced settings

Column naming: Aggregation method (column name) □ Enable hiliting □ Process in memory □ Retain row order

Maximum unique values per group 10,000 Value delimiter ,

OK | Apply | Cancel



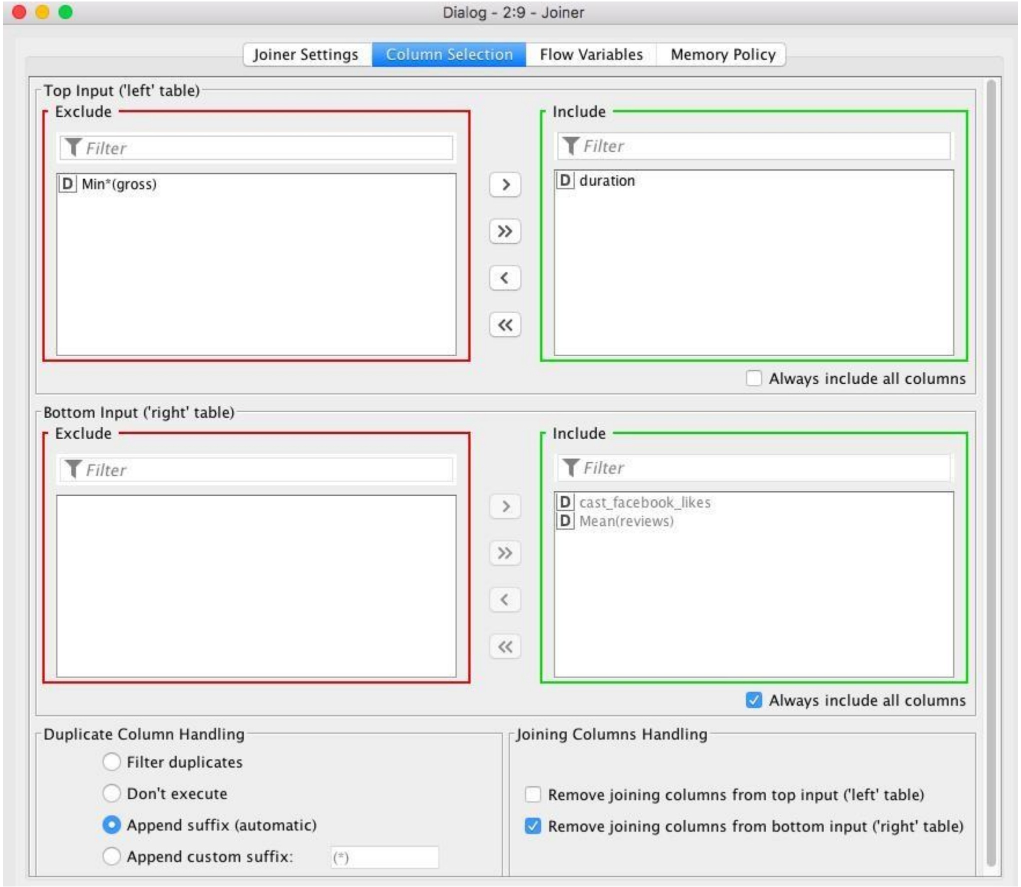
Pivot table - 2:6 - Pivoting

File Hilite Navigation View

Table "default" - Rows: 6 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	D duration	D 2006....	D 2008....	D 2010....	D 2012....	D 2016....
Row0	75	0	?	?	?	?
Row1	96	?	0.024	?	?	?
Row2	105	?	?	0.092	?	?
Row3	114	?	?	?	0.208	?
Row4	141	?	?	?	?	0.482
Row5	186	?	?	?	?	1

JOINER:



Top Input ('left' table)

Exclude: Filter; D Min*(gross)

Include: Filter; D duration

☐ Always include all columns

Bottom Input ('right' table)

Exclude: Filter

Include: Filter; D cast_facebook_likes; D Mean(reviews)

☑ Always include all columns

Duplicate Column Handling

○ Filter duplicates

○ Don't execute

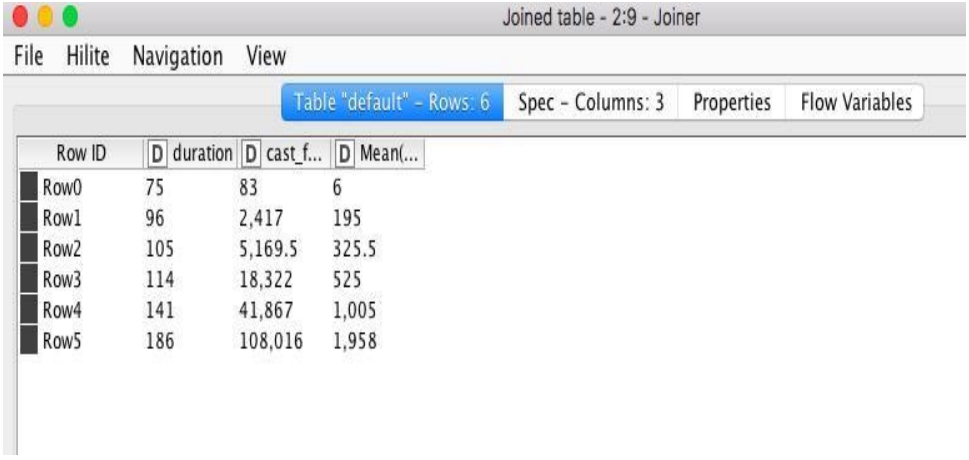
● Append suffix (automatic)

○ Append custom suffix: (*)

Joining Columns Handling

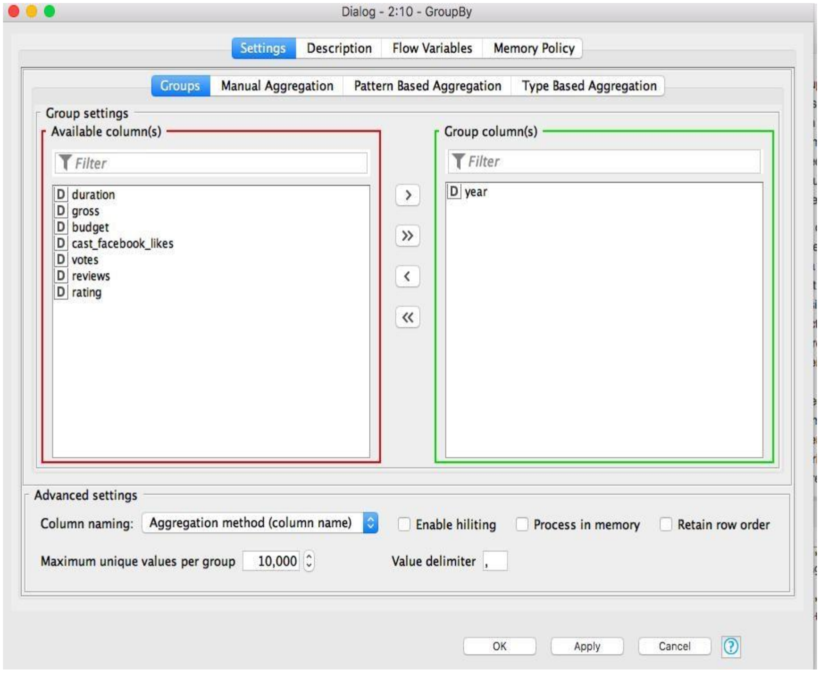
☐ Remove joining columns from top input ('left' table)

☑ Remove joining columns from bottom input ('right' table)



Row ID	D duration	D cast_f...	D Mean(...
Row0	75	83	6
Row1	96	2,417	195
Row2	105	5,169.5	325.5
Row3	114	18,322	525
Row4	141	41,867	1,005
Row5	186	108,016	1,958

GROUP BY:



Group settings

Available column(s)

- D duration
- D gross
- D budget
- D cast_facebook_likes
- D votes
- D reviews
- D rating

Group column(s)

- D year

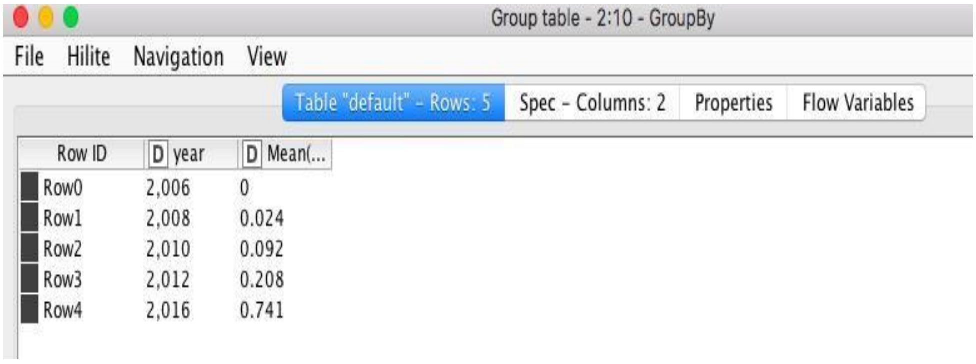
Advanced settings

Column naming: Aggregation method (column name)

□ Enable hiliting □ Process in memory □ Retain row order

Maximum unique values per group 10,000

Value delimiter ,



Row ID	D year	D Mean(...
Row0	2,006	0
Row1	2,008	0.024
Row2	2,010	0.092
Row3	2,012	0.208
Row4	2,016	0.741