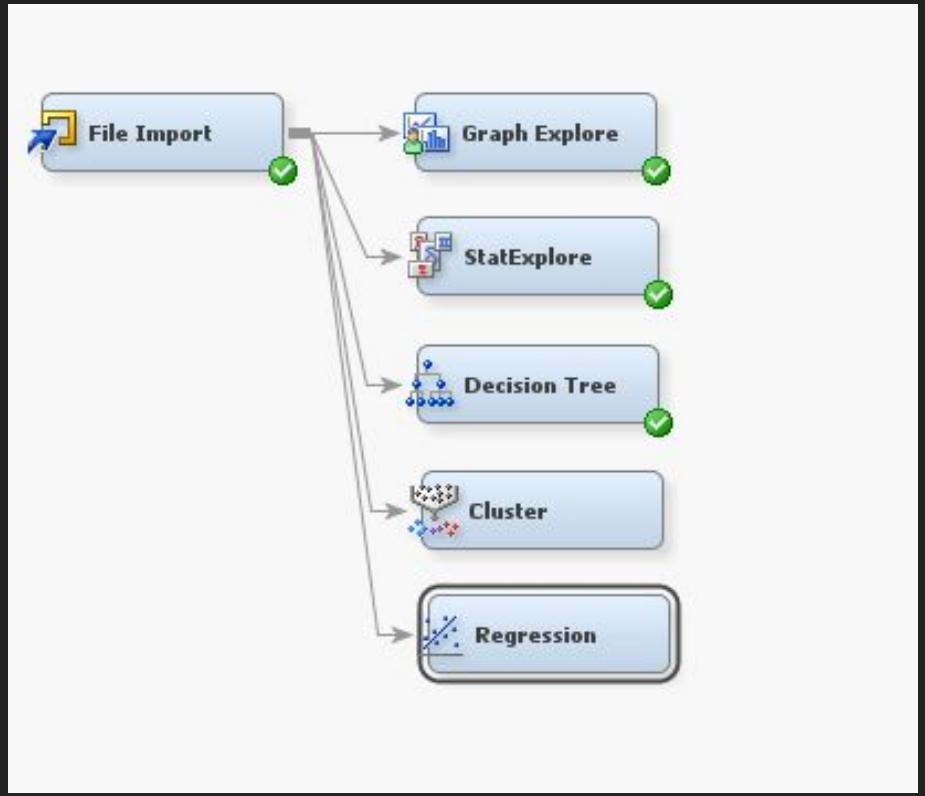
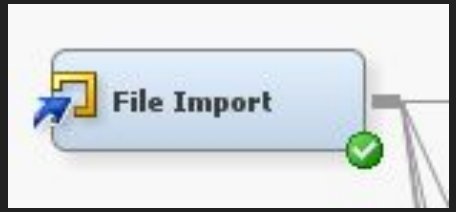
Using SAS Enterprise Miner to Analyze Netflix Data

By: Ashley Krause

- Secondary Source: Kaggle
  - Column 1
  - Title
  - Year (1905 - 2022)
  - Kind
  - Genre
  - Rating
  - Vote
  - Country
  - Language
  - Cast
  - Director
  - Composer
  - Writer
  - Runtime


File Import
Graph Explore
StatExplore
Decision Tree
Cluster
Regression



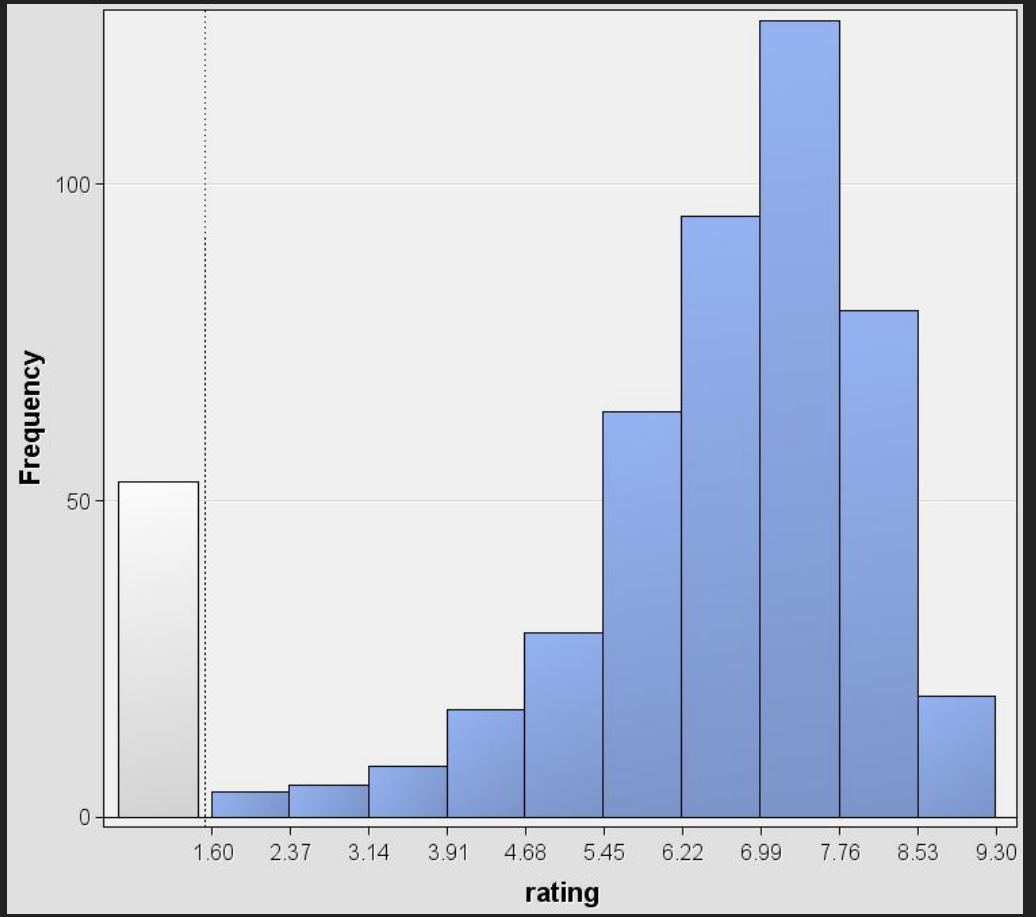
Variables - FIMPORT

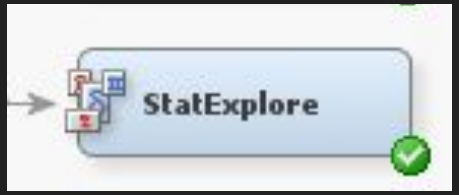
(none) not Equal to ...

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
cast	Text	Nominal	No		No	.	.
Column1	ID	Interval	No		No	.	.
composer	Text	Nominal	No		No	.	.
country	Text	Nominal	No		No	.	.
director	Input	Nominal	No		No	.	.
genre	Text	Nominal	No		No	.	.
kind	Input	Nominal	No		No	.	.
language	Text	Nominal	No		No	.	.
rating	Input	Interval	No		No	.	.
runtime	Input	Interval	No		No	.	.
title	Text	Nominal	No		No	.	.
vote	Input	Interval	No		No	.	.
writer	Text	Nominal	No		No	.	.
year	Input	Interval	No		No	.	.


Graph Explore


Frequency
100
50
0
1.60
2.37
3.14
3.91
4.68
5.45
6.22
6.99
7.76
8.53
9.30
rating



Class Variable Summary Statistics
(maximum 500 observations printed)

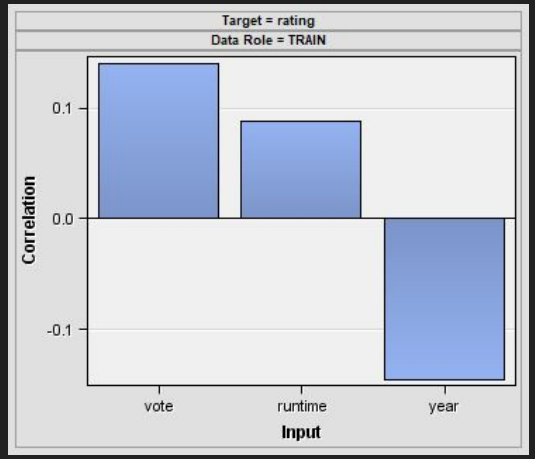
Data Role=TRAIN

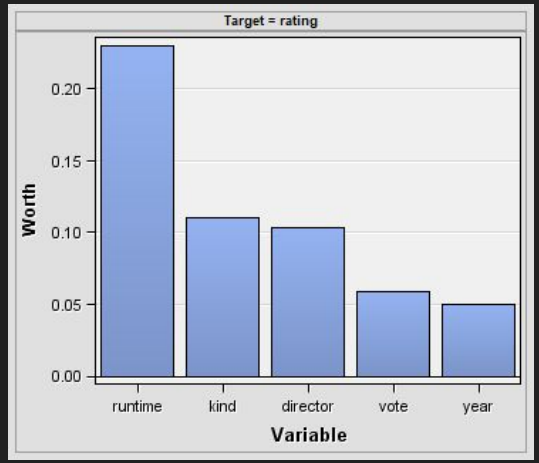
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	director	INPUT	513	138		20.26	['Akira Kurosawa']	0.59
TRAIN	kind	INPUT	8	0	movie	56.80	video movie	14.54

Interval Variable Summary Statistics
(maximum 500 observations printed)

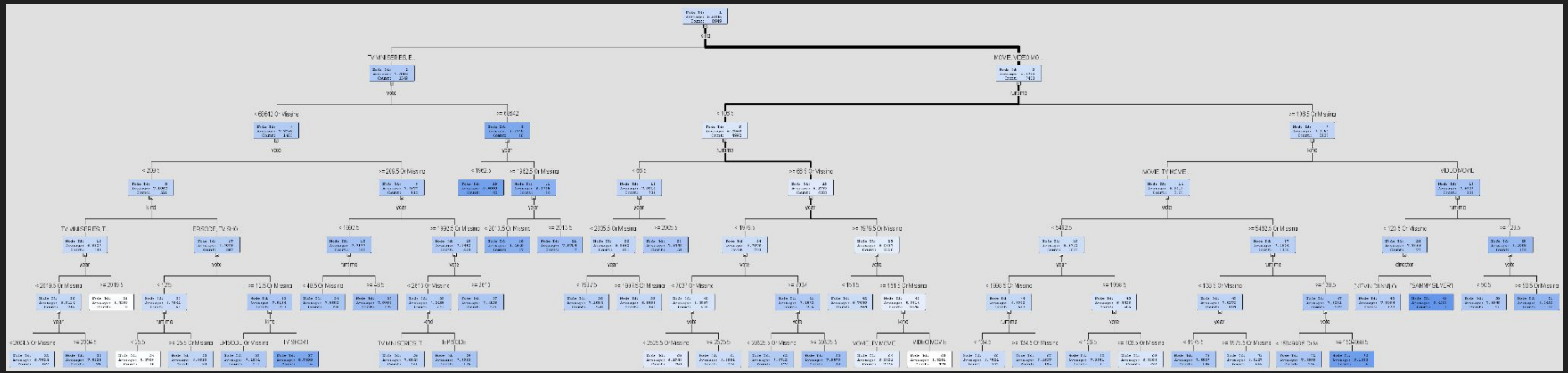
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
rating	INPUT	6.680635	1.285113	8949	807	1	6.9	9.6	-0.81107	0.757476
runtime	INPUT	98.11594	63.68129	8763	993	1	94	1620	8.410935	141.9943
vote	INPUT	21218.21	98048.73	8949	807	5	1535	2462087	12.20194	201.5009
year	INPUT	1994.74	16.24509	9756	0	1905	1999	2023	-1.46315	3.023835






Decision Tree



Cluster Summary for 1 Cluster

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	1.096628	0.2742	1.0181

Total variation explained = 1.096628 Proportion = 0.2742

Cluster 1 will be split because it has the largest second eigenvalue, 1.018102, which is greater than the MAXEIGEN=1 value.

Clustering algorithm converged.

Cluster Summary for 2 Clusters

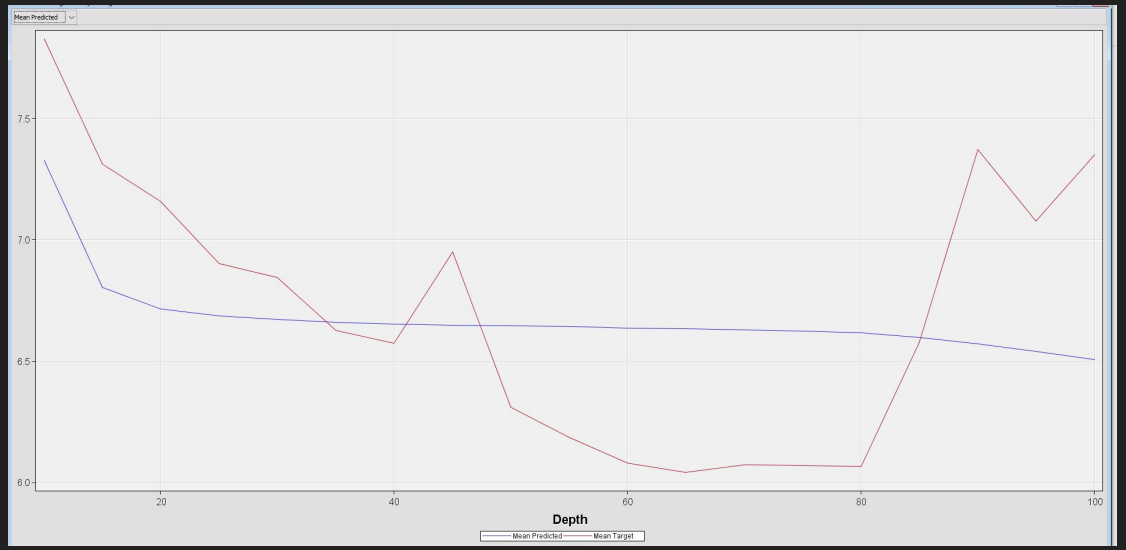
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	2	2	1.061685	0.5308	0.9383
2	2	2	1.031453	0.5157	0.9685

Total variation explained = 2.093138 Proportion = 0.5233

2 Clusters		R-squared with Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	runtime	0.5308	0.0010	0.4696
	vote	0.5308	0.0007	0.4695
Cluster 2	VAR1	0.5157	0.0025	0.4855
	year	0.5157	0.0000	0.4843

No cluster meets the criterion for splitting.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	1.096628	0.2742	0.2742	1.018102	0.0445	
2	2.093138	0.5233	0.5157	0.968547	0.5157	0.4855



utput

The DMREG Procedure

Model Information

Training Data Set	WORK.EM_DMREG.VIEW
DMDB Catalog	WORK.REG_DMDB
Target Variable	rating
Target Measurement Level	Interval
Error	Normal
Link Function	Identity
Number of Model Parameters	2
Number of Observations	8453

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	108.089400	108.089400	65.89	<.0001
Error	8451	13863	1.640360		
Corrected Total	8452	13971			

Model Fit Statistics

R-Square	0.0077	Adj R-Sq	0.0076
AIC	4185.5230	BIC	4187.5239
SBC	4199.6075	C(p)	2.0000

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.4701	0.0260	249.29	<.0001
runtime	1	0.00178	0.000219	8.12	<.0001

---
* Score Output

---
* Report Output