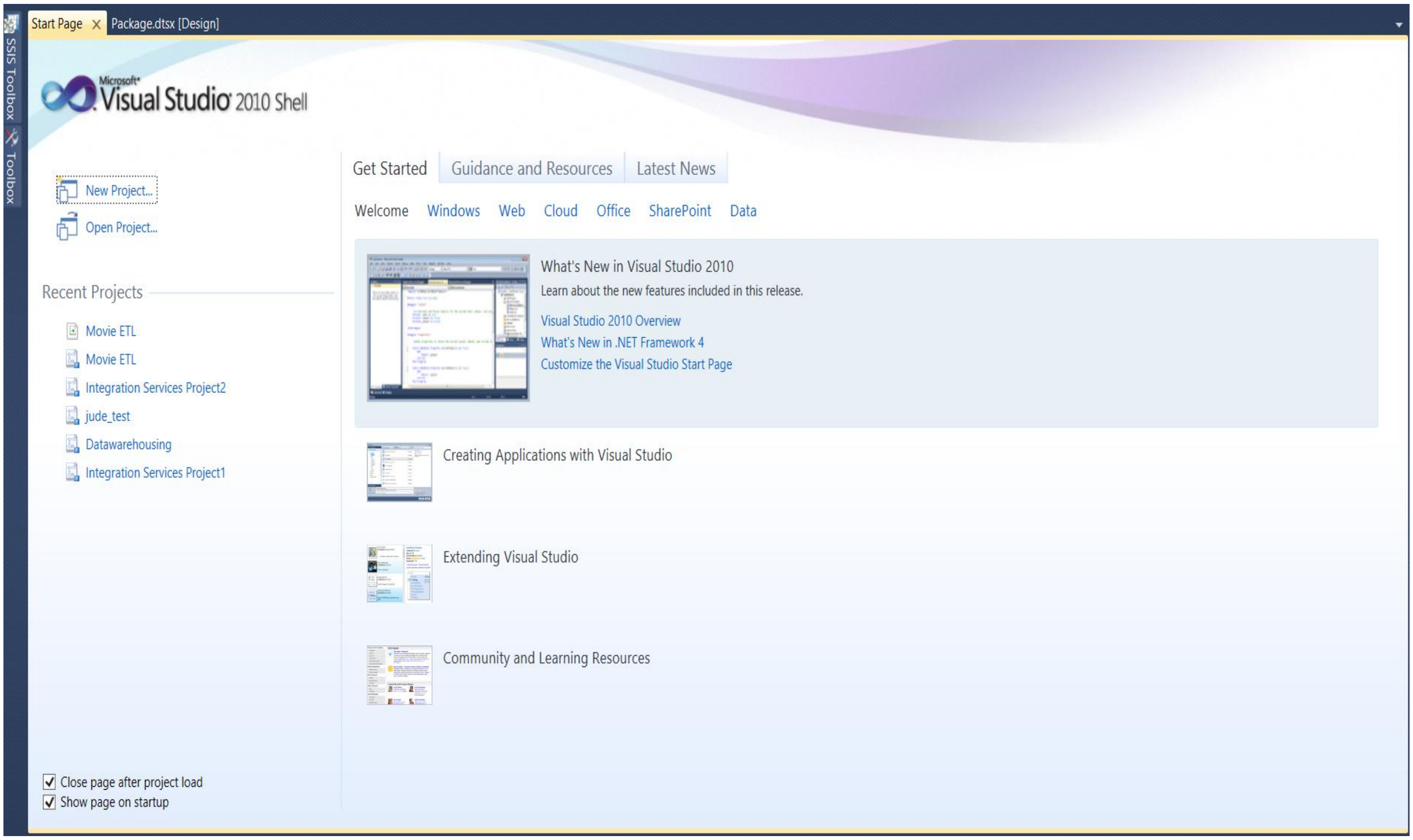
“Movie Management System”

Data Mining of the Application

Mohamed Niyaz

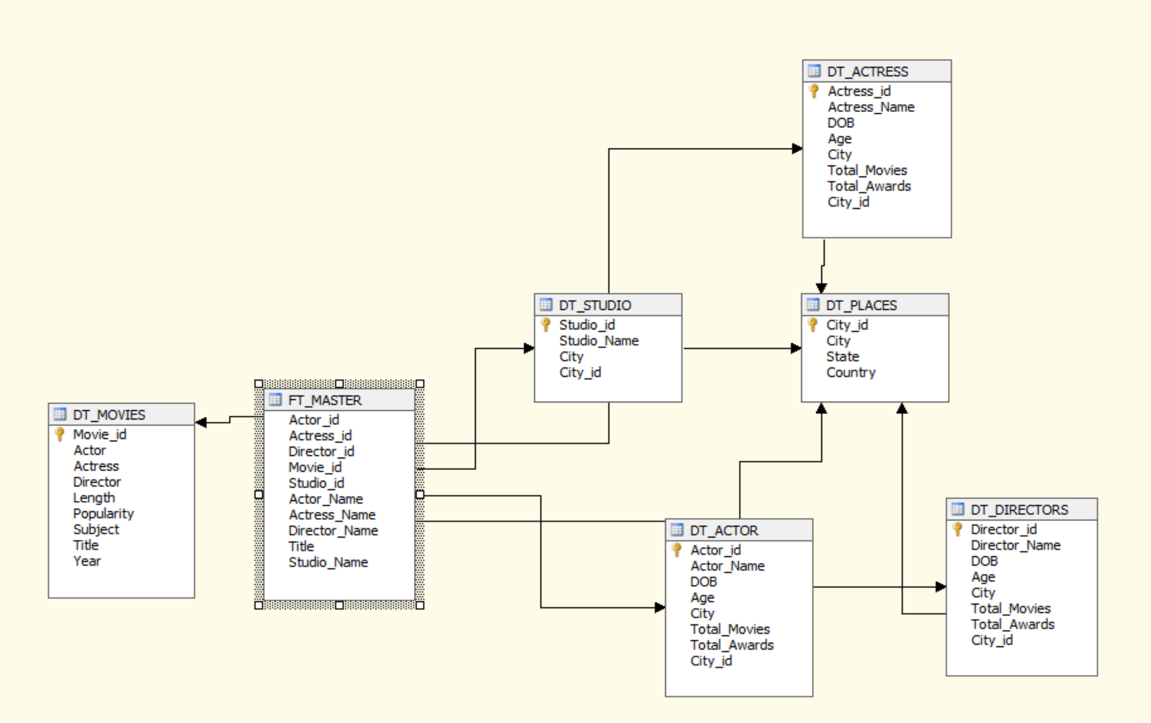

Start Page
Package.dtsx [Design]
SSIS Toolbox
Toolbox
Microsoft Visual Studio 2010 Shell
New Project...
Open Project...
Recent Projects
Movie ETL
Movie ETL
Integration Services Project2
jude_test
Datawarehousing
Integration Services Project1
Get Started
Guidance and Resources
Latest News
Welcome
Windows
Web
Cloud
Office
SharePoint
Data
What's New in Visual Studio 2010
Learn about the new features included in this release.
Visual Studio 2010 Overview
What's New in .NET Framework 4
Customize the Visual Studio Start Page
Creating Applications with Visual Studio
Extending Visual Studio
Community and Learning Resources

☑ Close page after project load
☑ Show page on startup

1. Mining Introduction

Data mining consists of a set of statistical techniques for analyzing observational data. The techniques are employed in the analysis step of the "Knowledge Discovery in Databases" process, or KDD. These tools discover patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

2. Mining Design



Our Data Mining Process is carried through **Microsoft BI** tool and below screen shot explains the high level overview of the **Schema**

The below Screenshot and process depicts different parts of Data Mining Process.

Tools Used: Microsoft SSAS

3. A Description of our mining scenarios and implementation

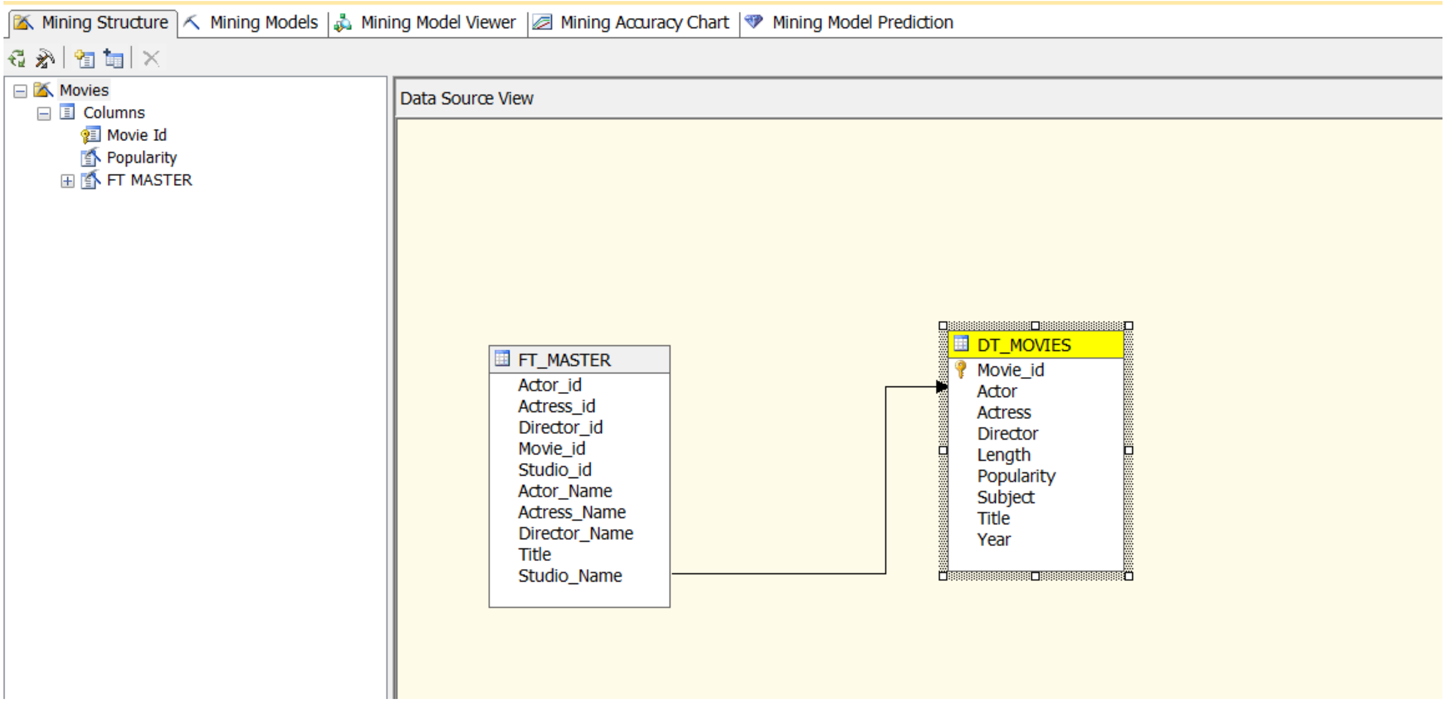
We designed and implemented two mining scenarios for this phase:

Scenario 1:
The movie popularity system used by the IMDB uses a simple 0-100 scale. These values are given definite values for example 5, 10 where 0 being the minimum and 100 being the maximum. Therefore, in order to better understand this popularity system we want to cluster the movies by popularities. This will allow us to find breaking points between clusters and use these to apply labels to certain number ranges. In other words we hope to use clustering to identify clusters of movie ratings and assign values to each cluster.

4. Implementation

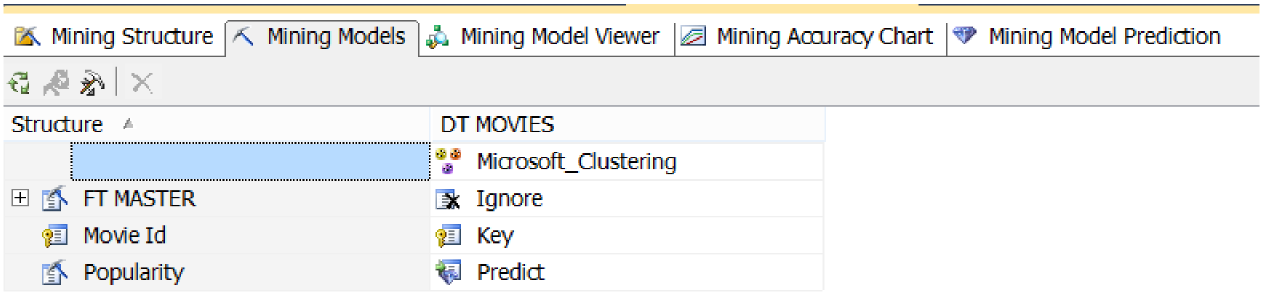
We implemented mining scenario using Sql server analysis service.
Steps Involved:
1. Created a new SSAS project.
2. Gave connection to Sql Server 2012 and added an existing Movies database to the data source.
3. Created Data source view from the dimension and fact tables present in the database.
4. Created a cube structure using the fact table.
5. Created Mining structure, we used "Microsoft clustering algorithm". Added "DT_MOVIES" table as "Case" table and "FT_MASTER" table as the "nested" one.
6. Gave "rating" as prediction attribute and "Movie_id" as input attribute and "Actor_Id"as the key.

Mining Structure for the clustering Algorithm



Overall relation of a Case table and a Nested Table

Mining model for the clustering algorithm

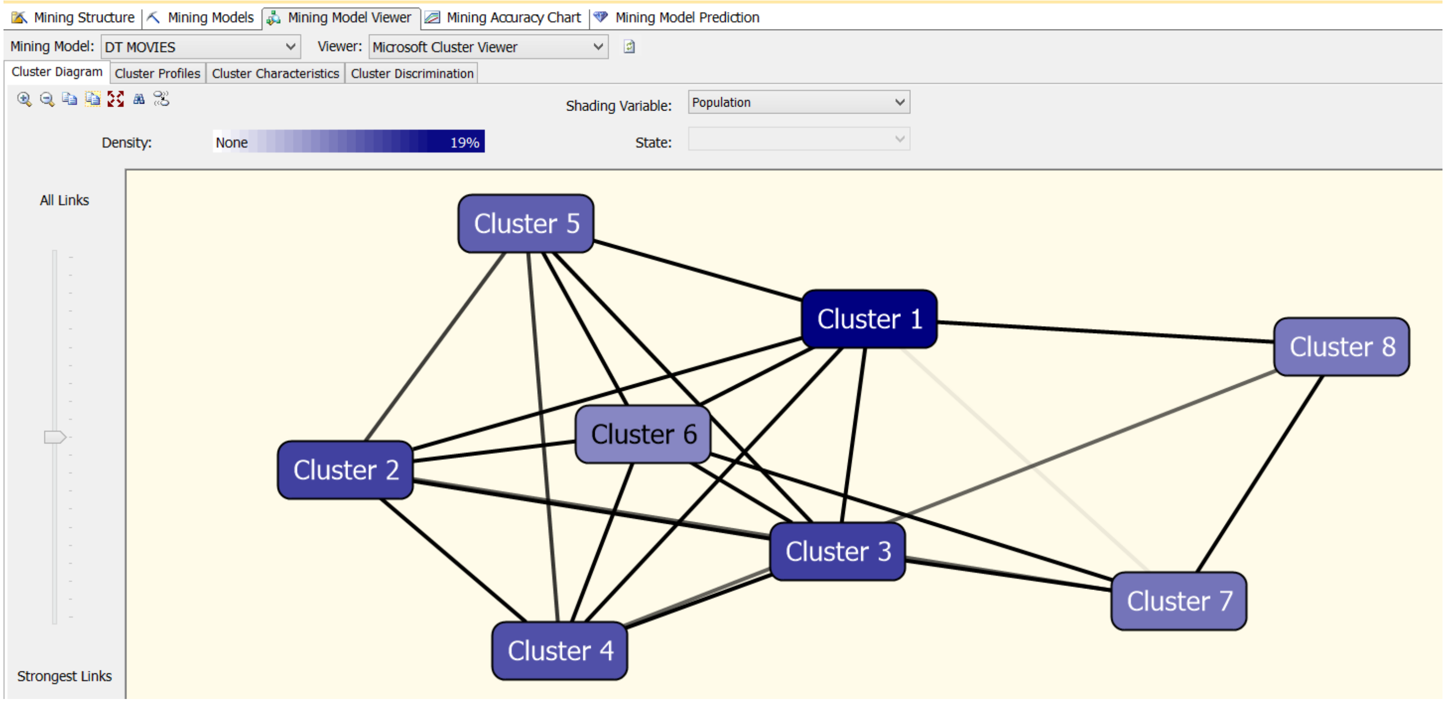


Structure	DT MOVIES
	Microsoft_Clustering
FT MASTER	Ignore
Movie Id	Key
Popularity	Predict

The Mining Model elaborates the relation between the nested table "FT_MASTER" and case table "DT_MOVIES". The attribute Popularity serves to be the Predict Column and Movie_id attribute being the Primary Key. The Data's are clustered using both the attributes and the calculation happens with Predict.

Screenshots of working Cluster Algorithms

Given below is the result of the clustering algorithm we have used on our movie database for the above scenario. The links between the clusters shows how strongly they are correlated based on the distance. Darker the color, stronger is the link. The Cluster with Strong Color pattern indicates the more insight of data's grouped with in the desired range. The Cluster 1 is the darkest of all which gives an overview depicting more rates are given in this desired range.



Mining Model: DT MOVIES

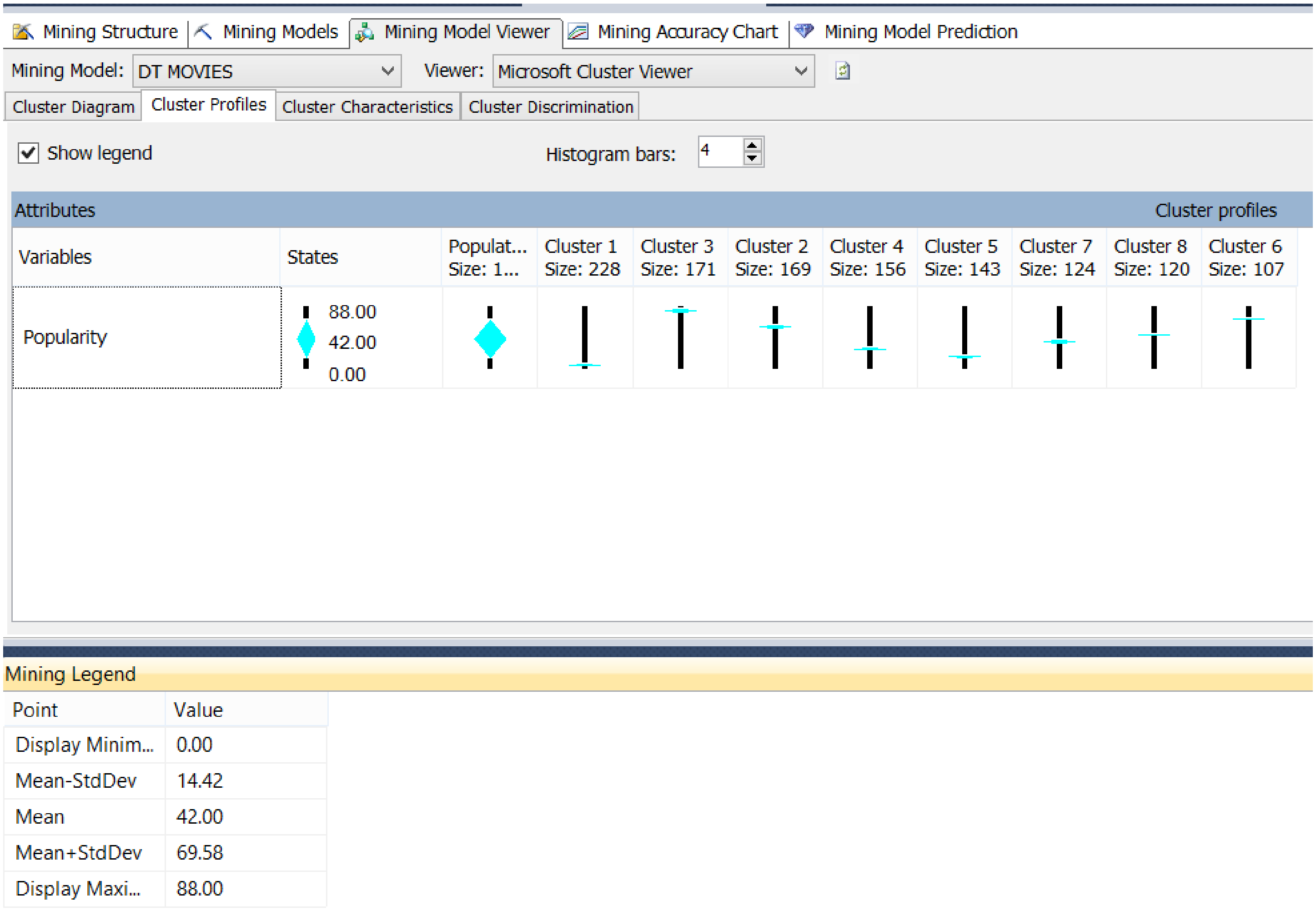
Viewer: Microsoft Cluster Viewer

Shading Variable: Population

Density: None 19%

State:

Cluster Profile of the Algorithm is used



Mining Model: DT MOVIES

Viewer: Microsoft Cluster Viewer

☑ Show legend

Histogram bars: 4

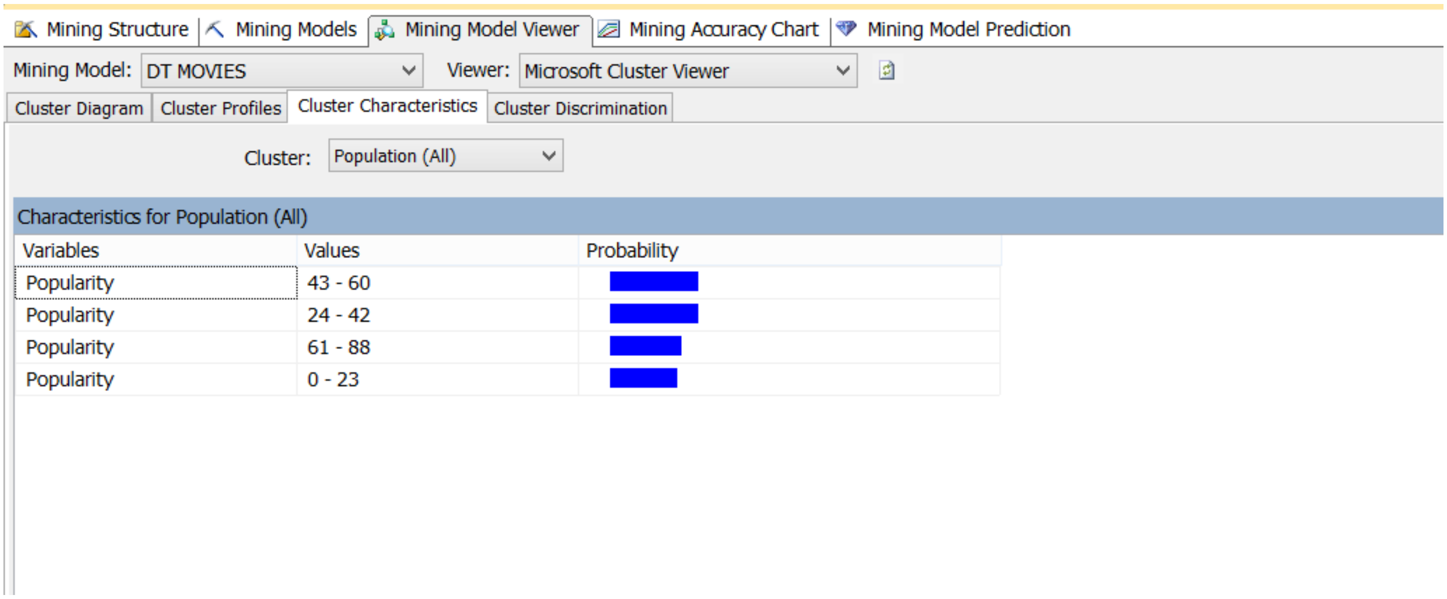
Variables	States	Populat... Size: 1...	Cluster 1 Size: 228	Cluster 3 Size: 171	Cluster 2 Size: 169	Cluster 4 Size: 156	Cluster 5 Size: 143	Cluster 7 Size: 124	Cluster 8 Size: 120	Cluster 6 Size: 107
Popularity	88.00 42.00 0.00									

Mining Legend

Point	Value
Display Minim...	0.00
Mean-StdDev	14.42
Mean	42.00
Mean+StdDev	69.58
Display Maxi...	88.00

Here we see the averages and ranges of the individual clusters. The average popularity for the movies is 42 and if you will notice the most clusters have a varying quality to them for example Cluster 2 is very low, Cluster 3 is very low and Cluster stands to be entire variant to other two being very high. The Histograms formed classifies the number of data's in each buckets the each cluster.

Cluster Characteristics



Mining Model: DT MOVIES

Viewer: Microsoft Cluster Viewer

Cluster: Population (All)

Characteristics for Population (All)

Variables	Values	Probability
Popularity	43 - 60	■
Popularity	24 - 42	■
Popularity	61 - 88	■
Popularity	0 - 23	■

The Popularity ranges have been well split among the rated values, the values and chart are well determined in the decreasing manner. If you see for example more number of users have rated movies between 43 -60 and least was rated between 0 to 23. While the first two ranges all have approximately the same probability the low range has a much lower probability. This is good since you do not want to see a lot of movies with such a low rating

5. Interesting Mining Results

The cluster are created with proper weightages, the buckets are split in a fashion that the popularities are spread across the buckets properly. For Example the clustered one is sized accordingly with average and has the maxium size 1218. The Most interesting of all was grouping of clusters, there were 6 clusters formed with right set of data distribution.

6. Challenges

- Data Mining operations can be performed only the cleanse data, so getting the cleanse data was one among the challenges.
- Data mining operations have to be performed on fact and dimension tables, forming multiple relations and hierarchy was really challenging.
- Forming the clusters was the biggest challenge among all these, we have to make sure the entity relation is perfectly right and data are fully cleansed.
- Finding the best relation between the nested table and the case table was really hard.