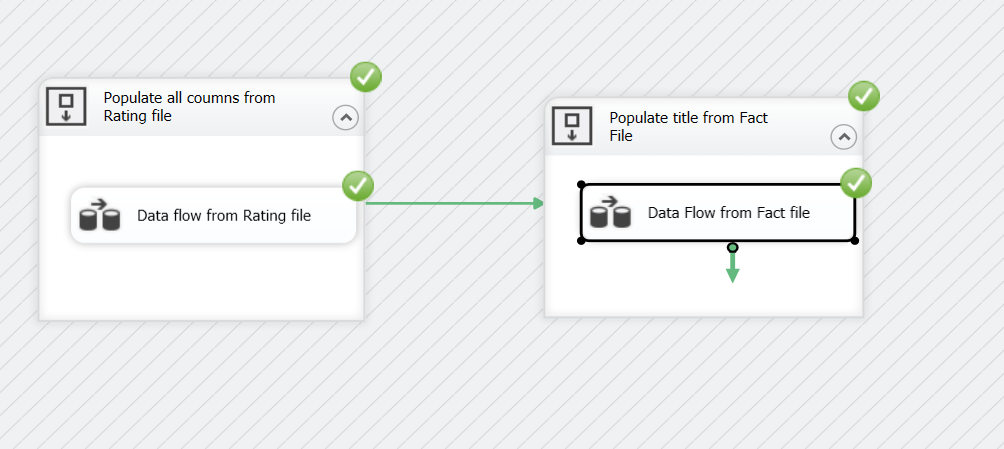
**Project3**

|  |  |  |
| --- | --- | --- |
| **Name** | **StudentID** | **Email** |
| Gyati Mittal | ji5288 | gmittal4@horizon.csueastbay.edu |
| Swati Hardasani | qd6852 | shardasani@horizon.csueastbay.edu |

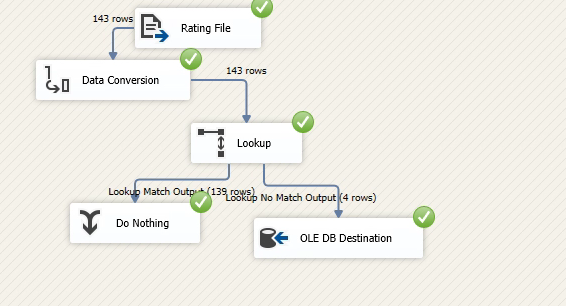
**Populate the Data:**

In SQL Server Data Tools (SSDT), create a multidimensional SSIS project to populate the DimMovie table. Drag two **Sequence Containers** in the Control Flow window. Then drag a Data Flow task inside each of these containers. Connect the first container to the second by dragging the green arrow as shown below.

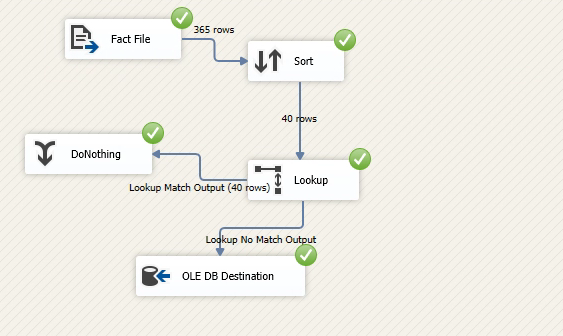


The first data flow task will populate the DimMovie dimension from Rating.txt file. Once the data is added to the dimension, the data flow task in the second container will automatically execute and populate the remaining movie titles from Fact.txt.

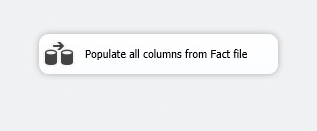
**Inside First Sequence Container:**

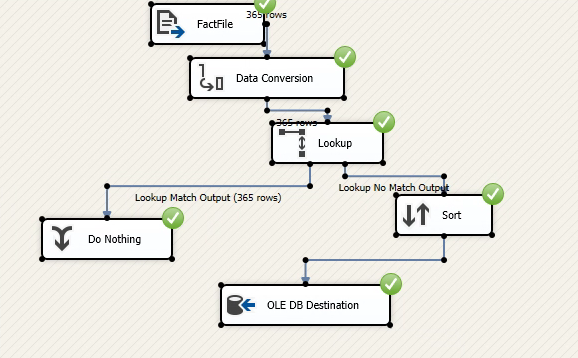


**Inside Second Sequence Container:**



**Populate data from Fact file to FactMovie table:**





Briefly describe the data flow of the Movie dimension and the fact table:

Movie Table:

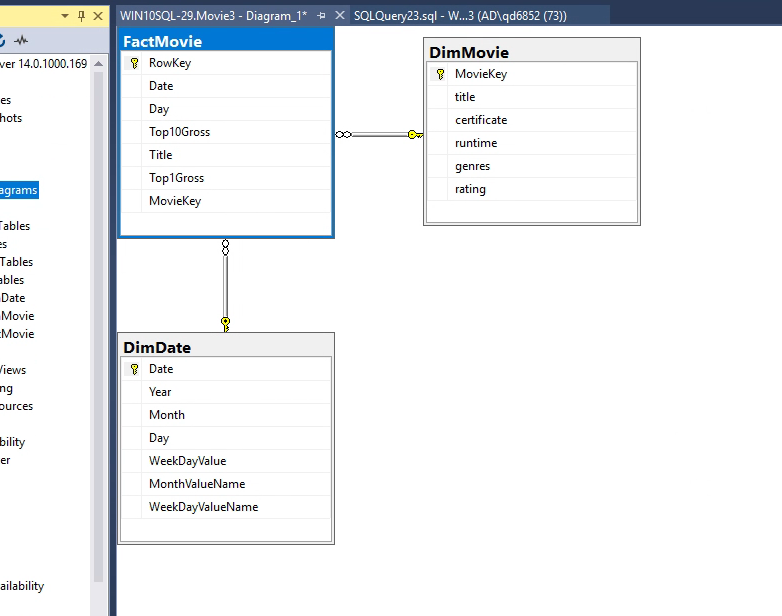
In the first sequence container populate all the data from the rating file to the DimMovie dimension. Here we convert the data types for runtime and time. Then we map the flat file with the movie table through lookup, sort it through with title and generate the output.

In the second sequence container, we populate the remaining titles and generate the output to the DimMovie dimension. First, we browse the flat file of Fact and then ignore the duplicates values through the sort, use the lookup for mapping the values of flat file with the DimMovie dimension and produce the output.

Fact Table:

We populate the all the data from fact file to the FactMovie table. We browse the fact flat file of and convert the data types for Date, day, Top10Gross and Top1Gross and then use lookup for mapping the flat file with the FactMovie table. We sort it through title and generate the output.

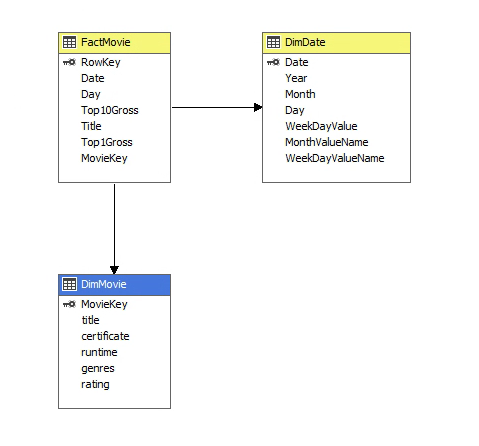
**Database Diagram:**



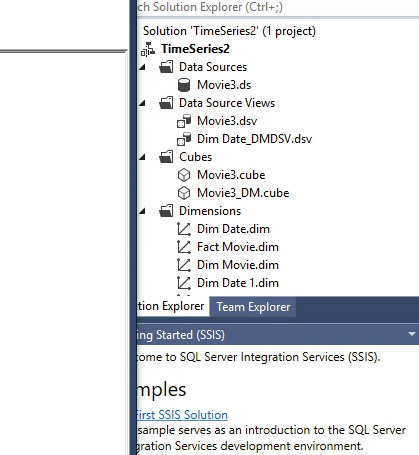
**Data Mining:**

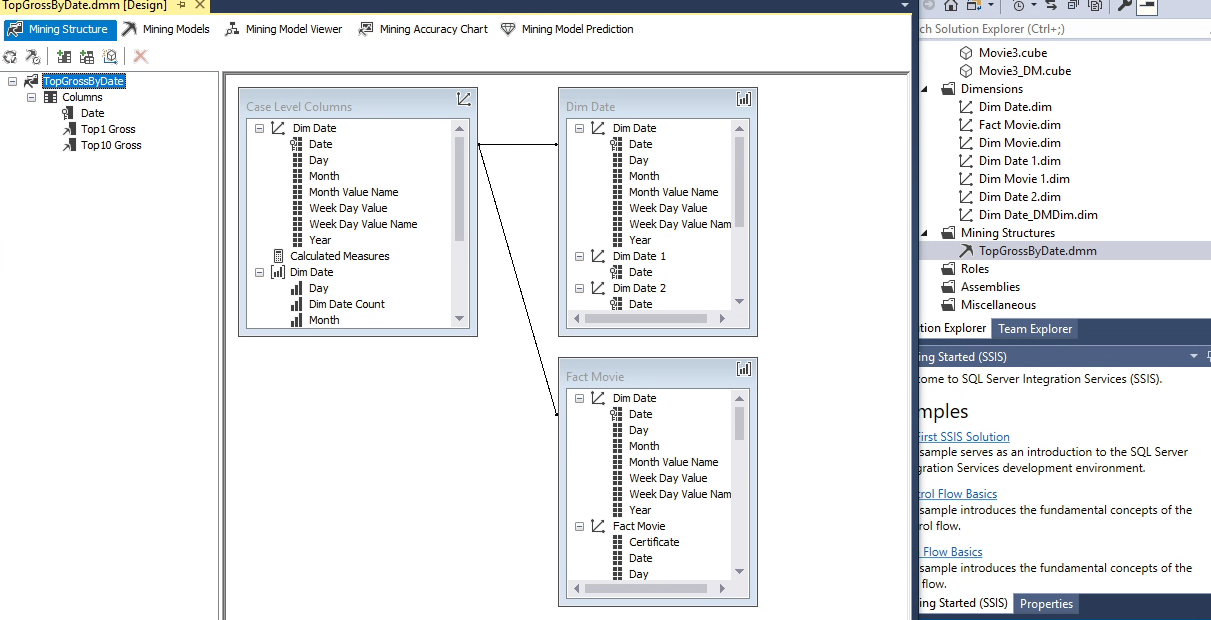
**Time Series Analysis:**

**Diagram of Cube:**

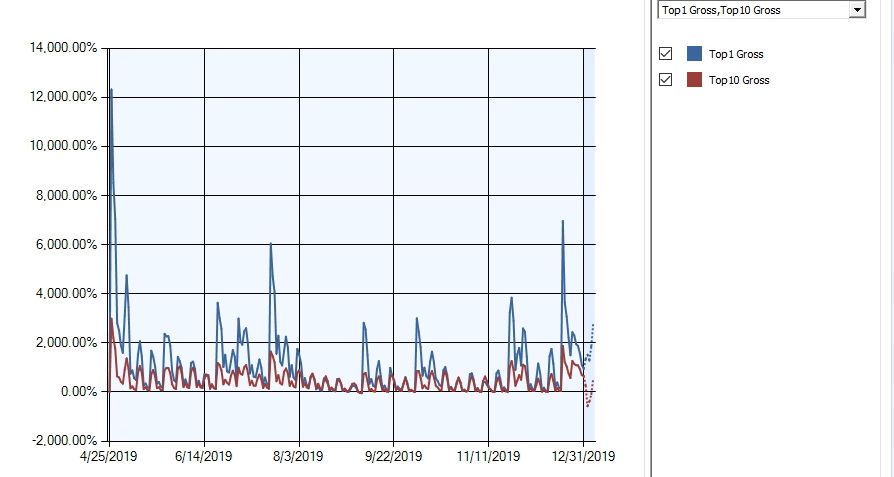


Connect to the Analysis Services Server in SSMS. Create an Analysis Services Multidimensional and Data Mining project with the name **Time Series**. Create an OLAP cube from the movie database. Create a new mining structure by right-clicking the **Mining Structure** folder in the Solution Explorer. Use the existing cube to do the mining structure. Select Time Series mining model. Select the Date dimension as the Source Cube Dimension. Select Date as the Case Key. Select Top10Gross and Top1Gross as the Case Level Columns. Top10Gross and Top1Gross should also be used as the input and predictable attributes. You can select Date dimension to slice the source cube. Name the Data mining and the model as **Top Gross by date**. Check the box to create cube using mining model. Right-click the ‘Top Gross by date.dmm’ data mining model in the solution explorer and click Process to deploy the model. Click Run to process mining structure after successful deployment. Open the **Mining Model Viewer** tab.

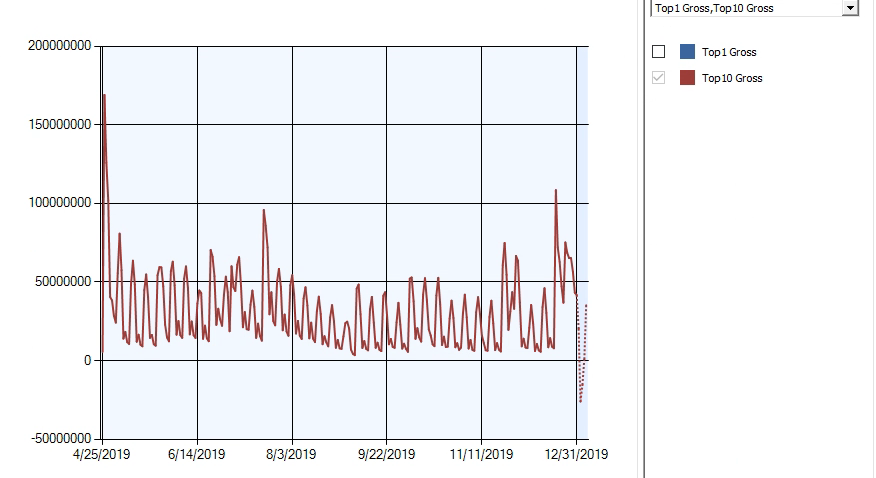




**Screenshot of the time series chart.**



**Uncheck the box of Top1Gross**



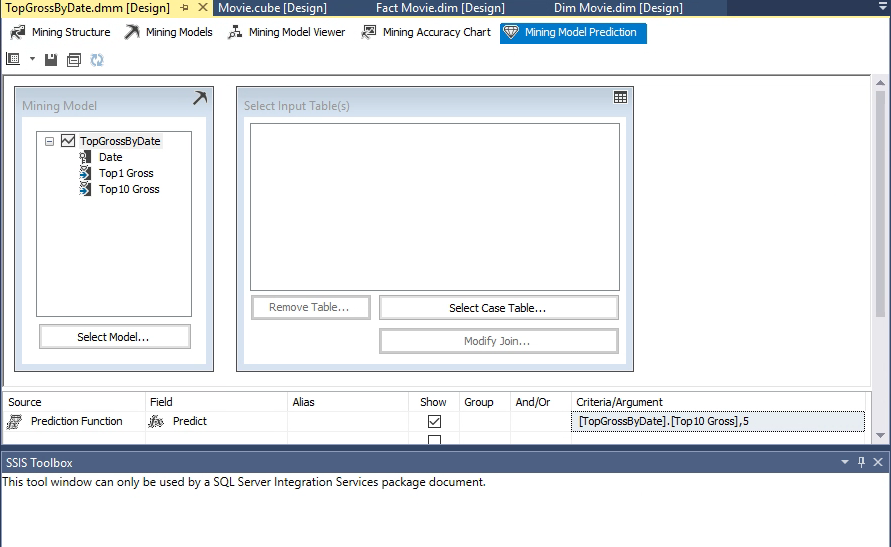
What do you infer from the time series chart? Why there are periodic spikes in Top10Gross revenue? Give suitable reason for periodic cycles. What is your guess about the number of days in each cycle?

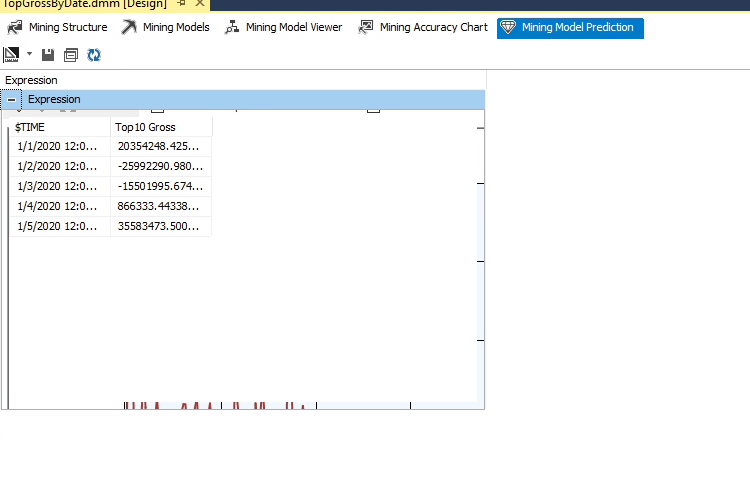
As per the above first time series chart when Top 1 Gross and Top 10 Gross are selected, the revenue spikes periodically. The reason is that people go to cinema usually on weekends and holidays. The top movie is making more revenue than the top 10 movies, which can be explained by the tendency of people going to the movie which is most popular that weekend.

Similarly, in the second chart when Top 10 Gross is checked, we can see periodic spikes on weekends. For example, 22 September 2019 was a Sunday, hence the spike.

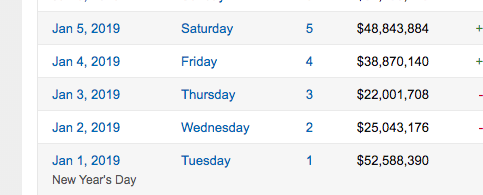
My guess is that each cycle is 7 days.

Click on the drop-down arrow below the Mining Structure tab and expand the expression to see the predicted revenues for next 5 days.





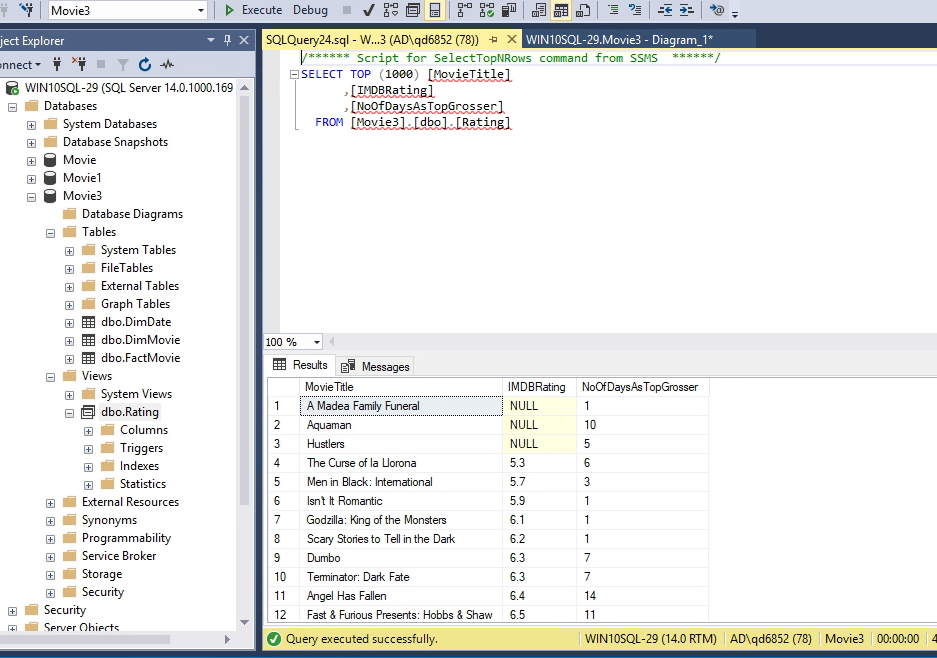
How good is this prediction (look at the website for actual values)?

**

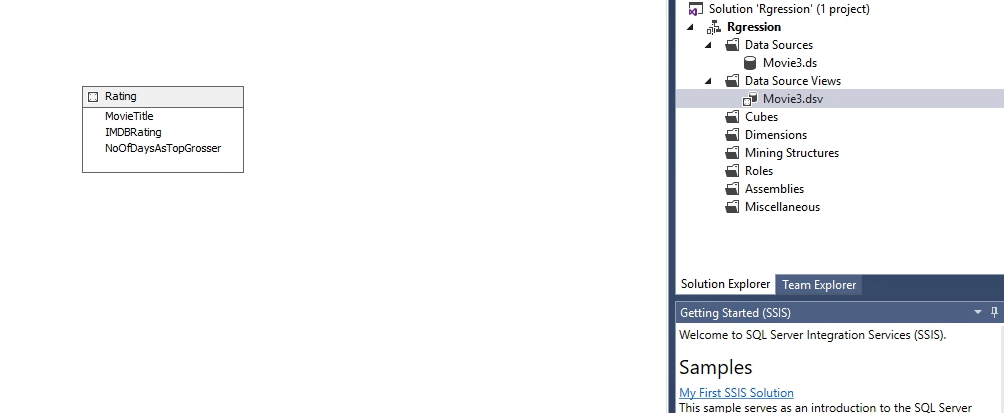
The predicted value does not seem close to the actual values from the website.

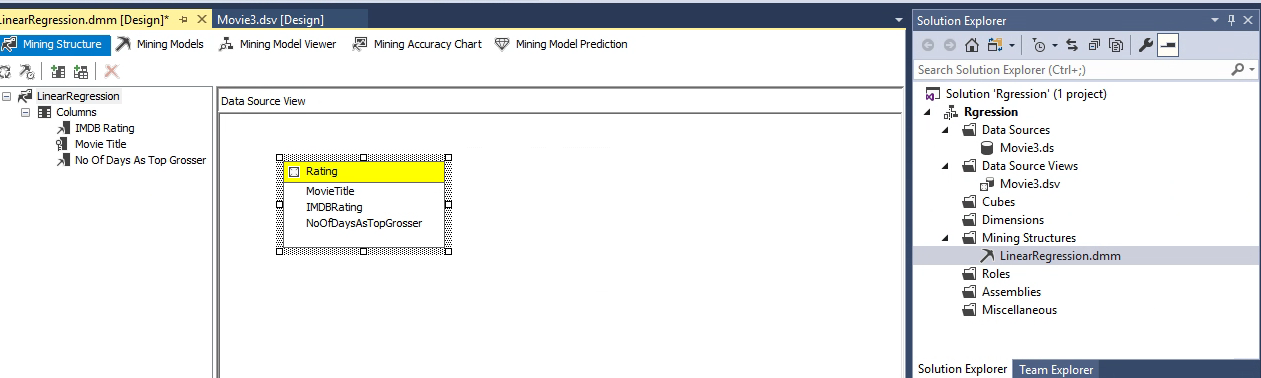
**Linear Regression Analysis**:

**Create View:**

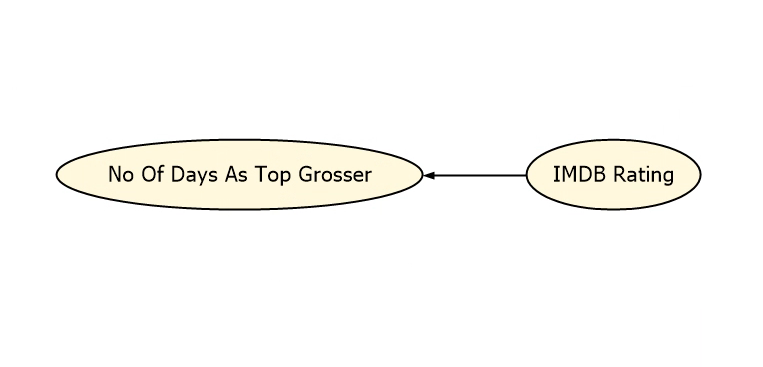
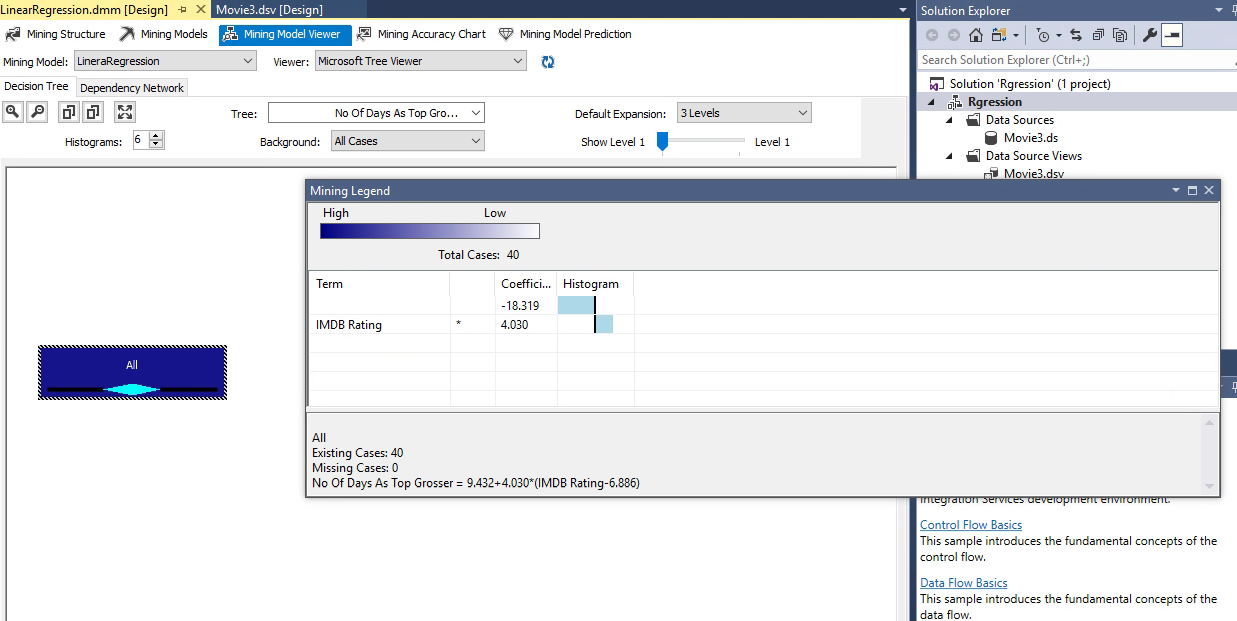


Create another SSAS Multidimensional and Data Mining project with the name **Regression**. Add the Rating view as the source. Create a linear regression data mining model with NoOfDays as the predictable attribute. Both IMDbRating and NoOfDays will be inputs and Movie Title will be the key. Use zero percentage data for testing. Deploy and process the model.





**Screenshot of Mining Model Viewer tab:**



What is the regression equation? Do the model estimates support the hypothesis (do not consider the significance level as it is not provided)?

**Regression Equation:**

No Of Days As Top Grosser = 9.432+4.030\*(IMDB Rating-6.886)

Yes, this model supports the hypothesis that higher the IMDB rating, more will be the number od days of that movie as top grosser. Number of days as top grosser increases linearly with the IMDB rating of that movie.

Click on the Mining Model Prediction tab and select Rating as the Case table. Right click to select Singleton query. Predict how long a movie with rating 8.5 will remain as the number one release.

