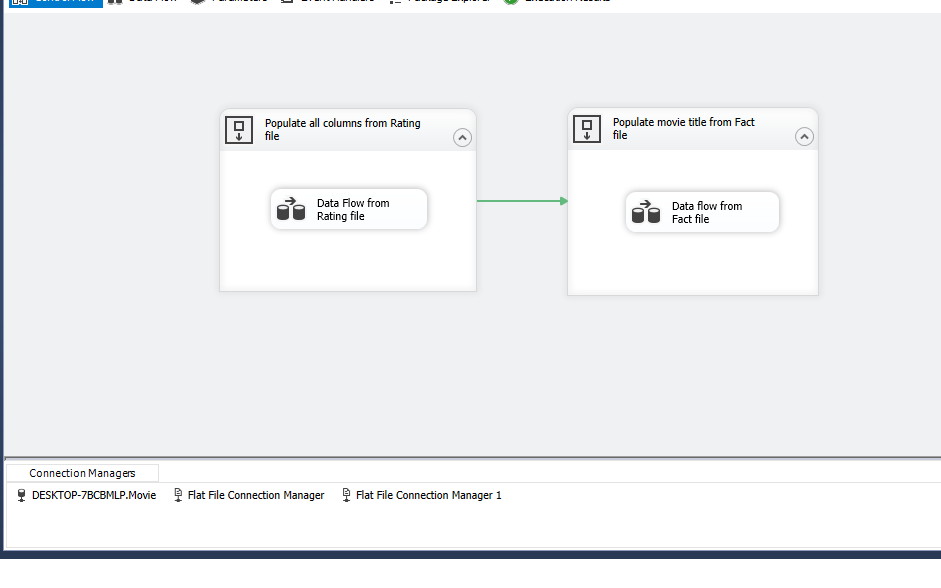
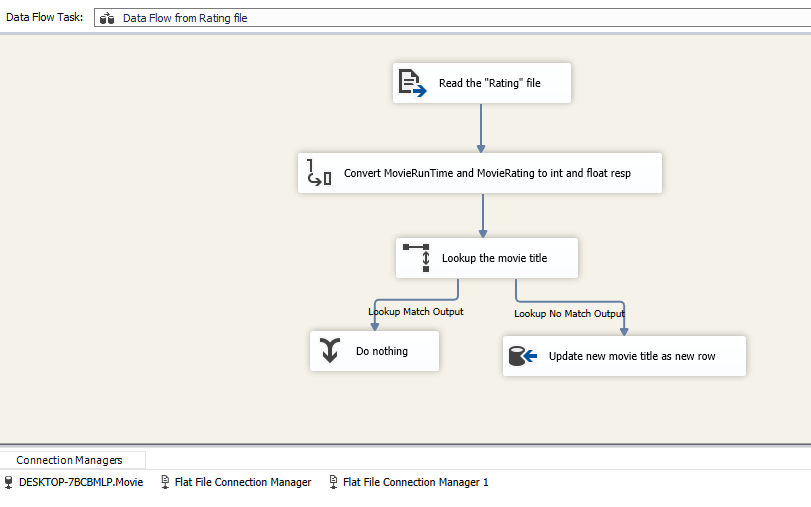
Data Flow of movie dimension and fact table:

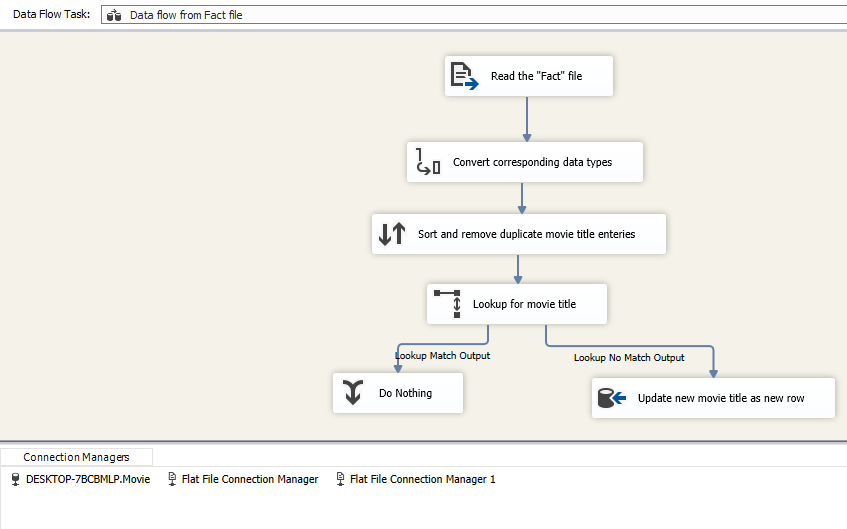
First, we scrapped data from the given links into files name Ratings.txt and fact.txt. After that we created two sequence containers to populate the dimMovie table in the Movie data base. As we want to collate the data from two files into one dimension table. Following is the screen shot for the same.



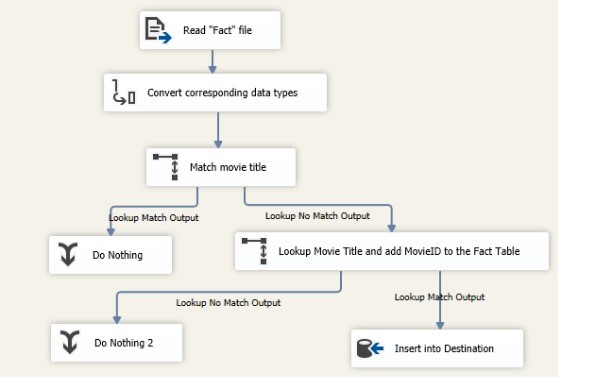
Following is the data flow for the first sequence container. In this data flow we first read the rating.txt file. Then we did the necessary data conversion operations (we converted MovieRunTime and MovieRating into integer and float). Then we lookup for the movie title, if there is a match we do nothing and if there is no match then output is added in the dimMovie table.



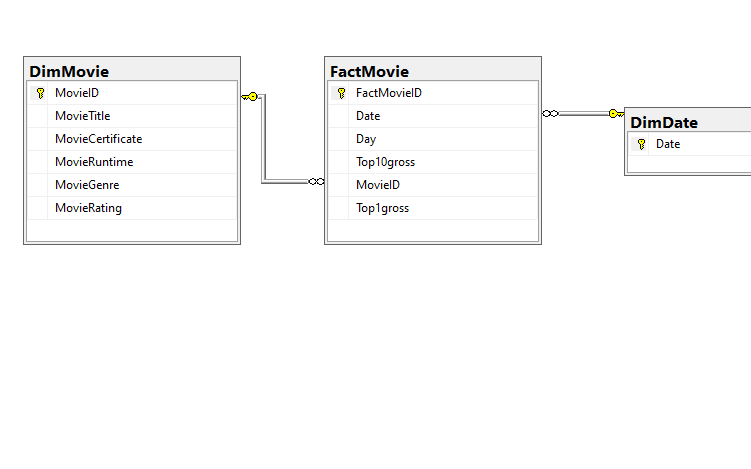
Following is the data flow for second sequence container. We first read the fact.txt file. Then we converted to corresponding data types. After that we sorted and removed duplicate movie tiltes. We do nothing if title matches and if not then we added to the dimMovie table.



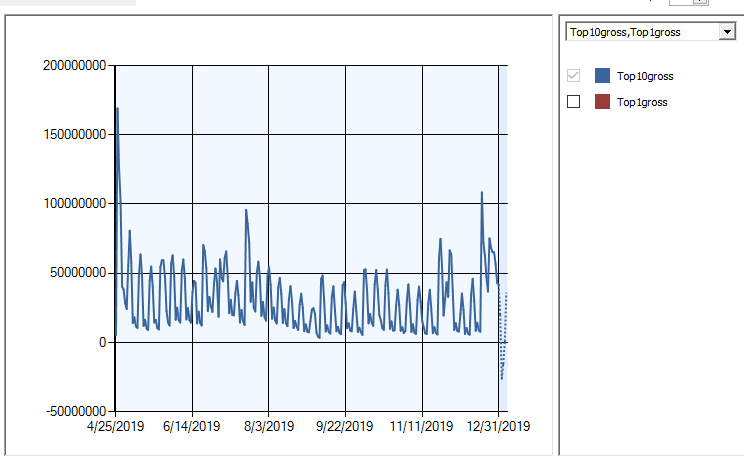
Following is the data flow for fact movie table. In this we read fact.txt file, We did data conversion for corresponding data types (Day, Date, Top10gross and Top1gross. We lookup for movie tiltes, If they match we do nothing, if they don’t match we go to another lookup. Now we then lookup for Movie Title and update the MovieID to fact table.

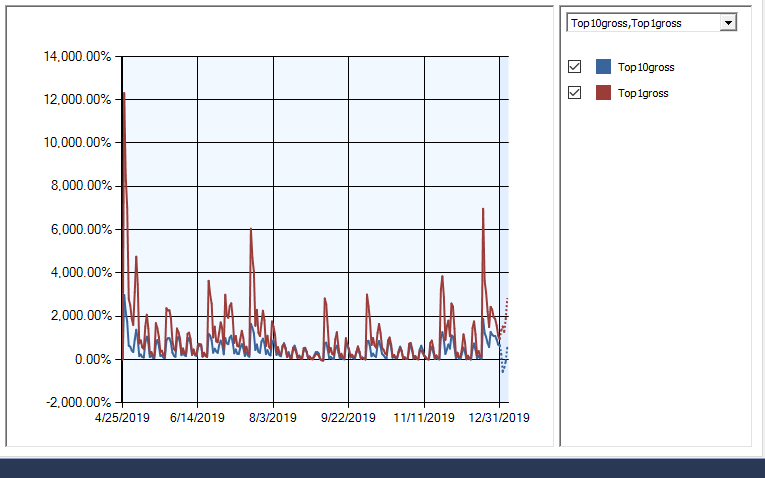


Following is the database diagram for Movie database.



The following figure shows the variation of Top10gross over a period of approx. eight months.





Questions:

*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?*

Time Series chart shows the variation of Top10gross over a duration of April to December.

With new release of movie the revenue peaks towards mid of the release week.

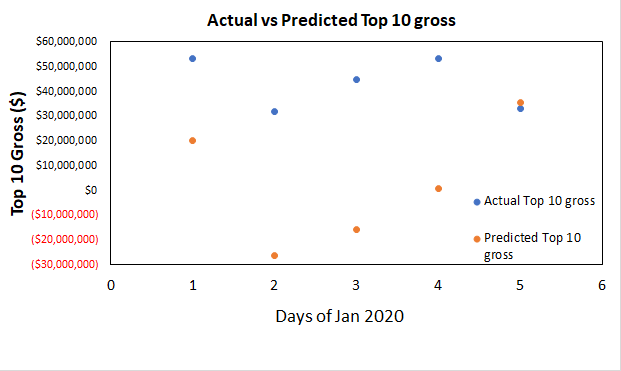
We see periodic spikes in top10gross that can be attributed to new movie releases causing higher gross on first few initial days and then we can see a dip in the later part of the month. Periodic cycle observed is aprrox. one week. It is interesting to note that the Top10gross peaks towards the end of the year which can be attributed to the holiday season.

*Question:*

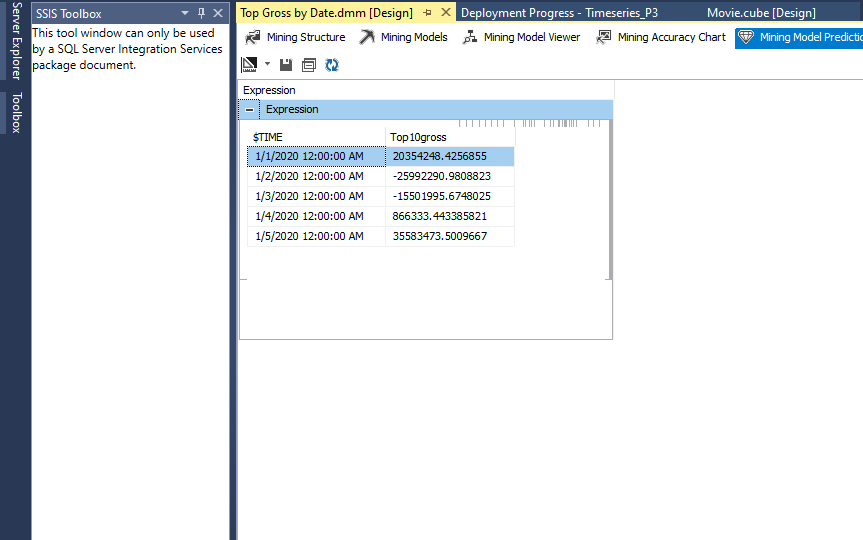
*How good is this prediction (look at the website for actual values)?*

It can be observed that prediction for the fifth day is accurate whereas perdictions for rest of the days do not match.

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Days** | **Actual Top 10 gross** | **Predicted Top 10 gross** |
| 1/1/2020 0:00 | 1 | $53,187,474 | $20,354,248.43 |
| 1/2/2020 0:00 | 2 | $31,762,542 | -$25,992,290.98 |
| 1/3/2020 0:00 | 3 | $44,715,346 | -$15,501,995.67 |
| 1/4/2020 0:00 | 4 | $53,199,800 | $866,333.44 |
| 1/5/2020 0:00 | 5 | $33,304,942 | $35,583,473.50 |



Following is the screenshot of predictions.



Question:

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

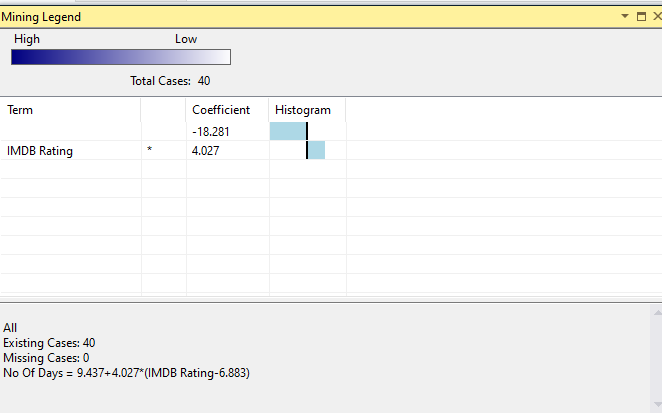
Regression equation is as follows:

Y=a+bx

No of days= 9.437+4.027(IMDB Rating-6.883)

Hypothesis: Higher the movie rating, more number of days it will remain as the top grossing movie.

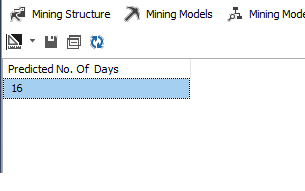
The model supports the hypothesis. It can be observed that Movie Ratings and No of days are strongly dependent on each other. A higher movie rating leads to more no of days it will remain top.



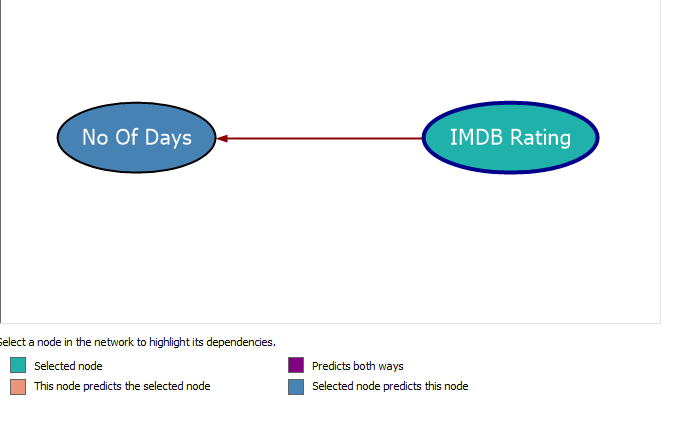
Question:

Predict how long a movie with rating 8.5 will remain as the number one release. **Take a screenshot of the result.**

From the screenshot rating of 8.5 will remain 16 days as number one release.



The image below shows that IMDB Rating predicts No of Days i.e it supports our hypothesis



Similarly the following figure shows No of Days predict IMDB ratings which implies movies staying for longer duration most likely would have higher IMDB rating.

