DATA WAREHOUSING AND SQL  
  
  
ALY6030, SPRING 2020  
MODULE 5 PROJECT ASSIGNMENT

WEEK 5: RETAIL SALES STAR SCHEMA

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

SUBMITTED TO: PROF. DAYA RUDRAMOORTHI

DATE: 06/21/2020

**Introduction**

The assignment aims at providing practical experience in developing SQL statements and understanding concepts. We have performed inspection of the dataset to provide key analysis on the public housing in Unites States, for demonstration to the senior management.

**Analysis**

**Question 1 (15 points)**

How many facts are there in this dataset?

As we know, a fact table is used in implementation of data warehouse design. A fact table is preferred in the center of a snowflake or star schema. Moreover, a fact table shows facts of a particular dataset. Also, a fact table record captures a measurement. The ‘Public Housing Inspections’ contains two facts.

Which facts do you identify?

After analyzing the CSV file, it was observed that the table has two facts, namely, COST\_OF\_INSPECTION\_IN\_DOLLARS and INSPECTION\_SCORE, as these records have metric measurements.

For the facts that you identify, what type of facts are they?

We know that, a fact table stores different types of measures such as additive, non-additive and semi-additive. The COST\_OF\_INSPECTION\_IN\_DOLLARS is an additive measure as we can calculate the summation of different costs with respect to the INSPECTION\_ID and INSPECTION\_DATE. However, the INSPECTION\_SCORE is a non-additive fact as this feature cannot be added amongst all the dimensions.

**Question 2 (15 points)**

How many dimensions are there in this dataset?

As we know that, dimension tables store the dimensions and attributes that describe the objects in the fact table. The dataset consists of seven dimensions.

Which dimensions do you identify?

The identified seven dimensions are:

INSPECTION\_ID, PUBLIC\_HOUSING\_AGENCY\_NAME, INSPECTED\_DEVELOPMENT\_NAME, INSPECTED\_DEVELOPMENT\_ADDRESS, INSPECTED\_DEVELOPMENT\_CITY, INSPECTED\_DEVELOPMENT\_STATE and INSPECTION\_DATE

**Question 3 (20 points)**

Senior management is interested in viewing the facts identified above, at both the inspection level, as well as a periodic summary of inspection costs for each month. Based on this context, if you were to store these data in a set of fact tables, which type (or types) of fact tables would you use and why?

In order to maintain a periodic summary of records, using a Transaction Fact Table is preferred, where every event has a date column showing when a particular event occurred. Also, a transaction table has an identifier column which identifies each record. This will help in displaying date tracked records to the senior management.

**Question 4 (20 points)**

Senior Management is also concerned with changes in the names and addresses of the public housing agency names since they tend to get merged with other agencies on a frequent basis.

Based on this context, how would handle this slowly changing dimension? Select from types 0,1,2, or 3 from the Kimball reading. Justify your answer.

The Slowly Changing Dimensions change slowly over a period of time. It is important to have a track of the slowly changing dimension to maintain a historical record of data. As we know, In the, Type 0 approach, some of the data can remain the same as it was recently inserted into the database and this approach might overwrite the remaining data. Moreover, in the Type 1 approach, the old value would be overwritten by the new value and there is no historical record of the old value kept in the database. Although, the Type 2 approach will have maintain all the history of dimension changes that are kept in the database, but this approach could be very expensive database operation and therefore not recommended. In addition, the Type 3 will ensure addition of a new column for inserting the new records. This type keeps only the old and new values within the database table. Therefore, this approach would be effective if used to maintain the historical data for the public housing dataset.

**Question 5 (30 points)**

Finally, Senior Management is interested in a subset of this data, for only those PHAs that saw an increase in the $$ cost of performing an inspection in their jurisdiction. Since none of them are SQL programmers, they’ve asked your help in performing this analysis by providing a file as your final deliverable with the following columns:

Note that MR stands for “most recent”:

**PHA\_NAME,**

**MR\_INSPECTION\_DATE,**

**MR\_INSPECTION\_COST,**

**SECOND\_MR\_INSPECTION\_DATE,**

**SECOND\_MR\_INSPECTION\_COST,**

**CHANGE\_IN\_COST**

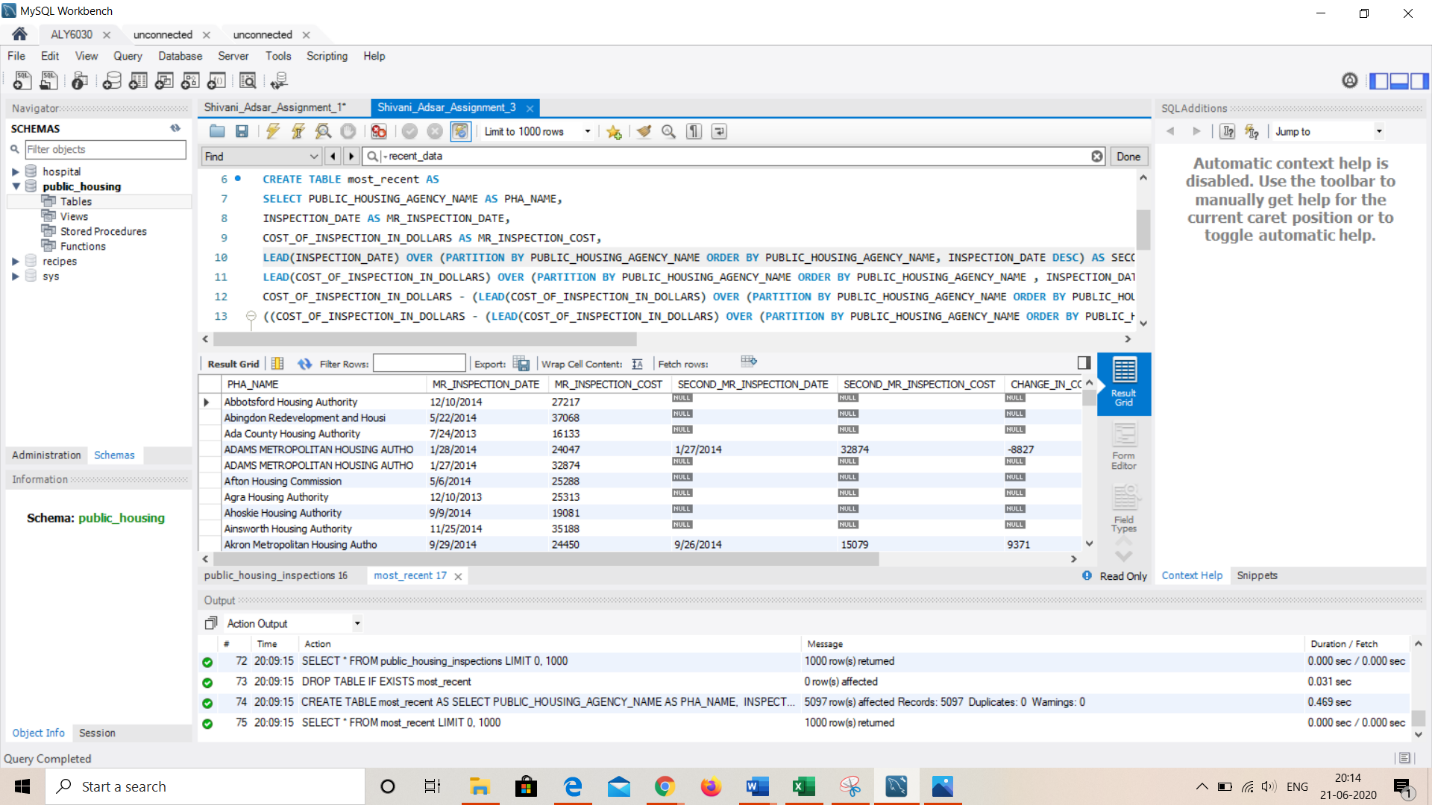
**PERCENT\_CHANGE\_IN\_COST**

Management has asked that you perform this function using lead or lag functions in SQL.

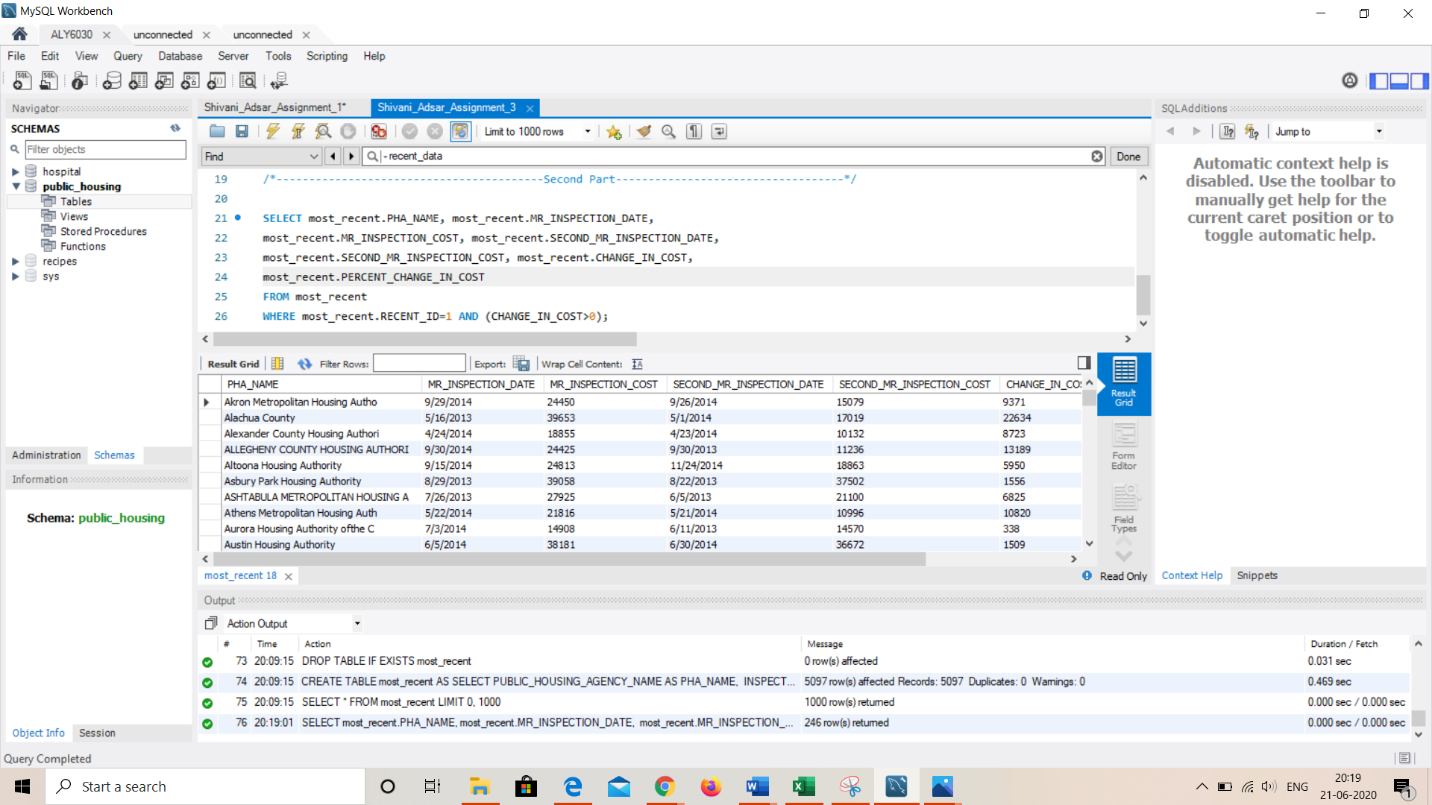
However, they’re concerned that the files when imported into MySQL Workbench may not properly refer to dates using the correct format. If that is the case, they’ve asked you to investigate how best to convert dates from TEXT to Date format so that the lead/lag functions work as expected.

They’ve also asked that you filter your dataset to only those PHAs that saw an increase in $$ cost, and that you only list the PHA once with no duplicates to avoid noisy data.

Naturally, this would also require you to filter out PHAs that only performed one inspection, so they’ve asked you to remove those as well.



*Fig.1: Use of Lead Function in SQL*



*Fig.2: Most Recent Change in Cost Output*

**References**

[1] Verbeeck, K. (2018, May 30). SQL Server Window Functions LEAD and LAG. Retrieved from https://www.mssqltips.com/sqlservertutorial/9127/sql-server-window-functions-lead-and-lag/#:~:text=LAG and LEAD,by changing the sort order.

[2] SQL Server LAG() Function By Practical Examples. (2020, April 11). Retrieved from https://www.sqlservertutorial.net/sql-server-window-functions/sql-server-lag-function/