Assignment 3: Retail Sales Star Schema

Background:

Public housing in the United States is administered by local or regional Public Housing Agencies (PHAs), under the supervision of the Federal Department of Housing & Urban Development (HUD). These agencies conduct periodic inspections of the developments in their jurisdiction and assign an inspection score for that development at a certain cost to the US Taxpayers.

You are a Developer hired by HUD to assist with a dimensional model for the inspection data, and to provide a key analysis for Senior Management as your deliverables.

You have access to a flat file of inspections including the PHA name, development name being inspected (along with address), as well as the date of inspection, inspection score which is expressed as a ***ratio or percentage*** from 0 to 100, and the cost of performing that inspection in dollars.

Your boss in IT wants you to answer the following questions based on your inspection of the dataset:

1.
    a. How many facts are there in this dataset?
    b. Which facts do you identify?
    c. For the facts that you identify, what type of facts are they?
2.
    a. How many dimensions are there in this dataset?
    b. Which dimensions do you identify?
3. 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?
4. 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, how should we handle this slowly changing dimension? Select from types 0, 1, 2, or 3 from the Kimball reading. Justify your answer.
5. 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"):
    a. **PHA_NAME**
    b. **MR_INSPECTION_DATE**
    c. **MR_INSPECTION_COST**
    d. **SECOND_MR_INSPECTION_DATE**
    e. **SECOND_MR_INSPECTION_COST**
    f. **CHANGE_IN_COST**
    g. **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 (in dollars), 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.

Submit three files:

1. Your answers to the questions above in a .pdf file
2. .csv file with the final output from question 5
3. .sql file from question 5

Example:

- YoniDvorkis_Assignment3.pdf
- YoniDvorkis_Assignment3.csv
- YoniDvorkis_Assignment3.sql