ALY6030 Assignment3 Public Housing Inspection Star Schema

Yinan Zhou

March 23, 2025

1. How many facts are there in this dataset?

There are **three main facts**:

Identified Facts	**Type of Facts**
COST_OF_INSPECTION_IN_DOLLARS	*Additive* fact – we can sum this over multiple inspections
INSPECTION_SCORE	*Semi-additive* – meaningful when averaged, but not necessarily summed.
INSPECTION_DATE	*Non-additive* – useful for grouping, filtering, or ordering, but not additive.

2. How many dimensions are there in this dataset?

There are **six main dimensions**:

- INSPECTION_ID: acts as a primary key.
- PUBLIC_HOUSING_AGENCY_NAME
- INSPECTED_DEVELOPMENT_NAME
- INSPECTED_DEVELOPMENT_ADDRESS
- INSPECTED_DEVELOPMENT_CITY
- INSPECTED_DEVELOPMENT_STATE

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?

To support senior management's need to view both detailed inspection-level data and monthly summaries of inspection costs, I would use **two types of fact tables**: a **transactional fact table** and a **periodic snapshot fact table**.

The **transactional fact table** captures individual inspection events. Each row in this table represents a single inspection, including attributes such as the inspection date, cost, score, and references to related dimension tables (e.g., public housing agency, development, location, and date). This structure allows for detailed, drill-down analysis of inspection-level data, enabling management to examine specific inspections by agency, region, or timeframe.

The **periodic snapshot fact table** summarizes inspection metrics at a fixed interval—such as monthly—aggregating total inspection costs, average inspection scores, and the number of inspections per public housing agency. This enables management to monitor trends over time, track performance and spending, and compare agencies or regions month over month.

Using both types of fact tables provides flexibility: the transactional fact table supports granular analysis, while the periodic snapshot fact table enables high-level trend reporting and executive summaries. Together, they offer a complete view of inspection performance and cost behavior across different timeframes.

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.

Since public housing agency names and addresses can change over time—often due to mergers or reorganizations—and senior management is interested in tracking these changes historically, the most appropriate approach is to implement a **Slowly Changing Dimension (SCD) Type 2**.

Justification:
SCD Type 2 is ideal when it's necessary to preserve the history of changes in dimension data. In this case, if a public housing agency changes its name or location, we want to maintain a historical record of how it was identified at the time of each inspection.
SCD Type 2 supports:

- Tracking the full history of name and address changes
- Retaining multiple versions of the same agency in the dimension table
- Linking each fact record (e.g., inspection) to the correct historical version of the agency

Under this model, the "dim_public_housing_agency" table would include additional columns such as:

- effective_date
- expiration_date or current_flag
- A surrogate key as the primary key (instead of using a natural agency ID)

Each time an agency's name or address changes, a new row is inserted with the updated values, while the previous version is retained to support historical analysis.

Why Not Types 0, 1, or 3?

- **Type 0** (no changes) would ignore updates entirely—this fails to meet management's need to track agency mergers or address/name changes.
- **Type 1** (overwrite changes) would update the agency name/address in place, **losing historical context**.
- **Type 3** (store previous value in a separate column) allows limited history tracking (usually only one prior version), which is **not sufficient** for frequent, ongoing mergers or renaming.

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"):
   - *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.

Objective

Senior management requested a summary identifying **Public Housing Agencies (PHAs)** that experienced a **cost increase in their most recent inspections**. The goal was to:

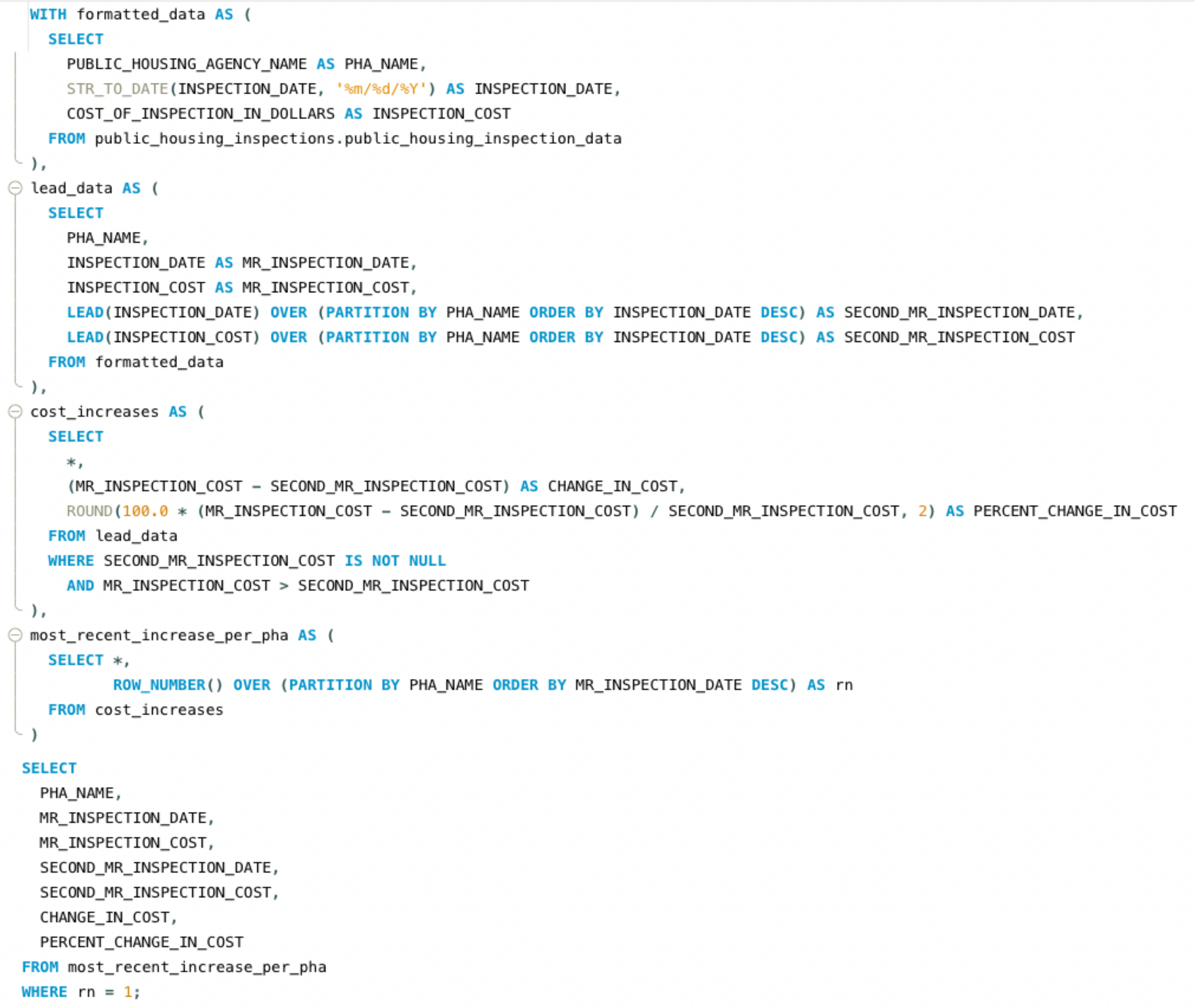
- Compare each PHA's two most recent inspections,
- Identify those with a cost increase,
- Report the cost difference and percentage increase,
- Ensure that each PHA appears **only once** in the final results.

Methodology

Inspection data is stored in a table named public_housing_inspection_data under the public_housing_inspections schema. The inspection dates are stored as text strings in the format MM/DD/YYYY. To allow accurate date comparisons and ordering, I first used "STR_TO_DATE" to convert the text dates to MySQL DATE type.

Using the "LEAD" window function, I retrieved each PHA's second most recent inspection immediately following their most recent one, based on descending inspection dates. I then calculated both the absolute change in cost and the percentage increase.

To avoid clutter and ensure clarity, I applied a "ROW_NUMBER" window function to rank cost increases and selected only the most recent one for each PHA.



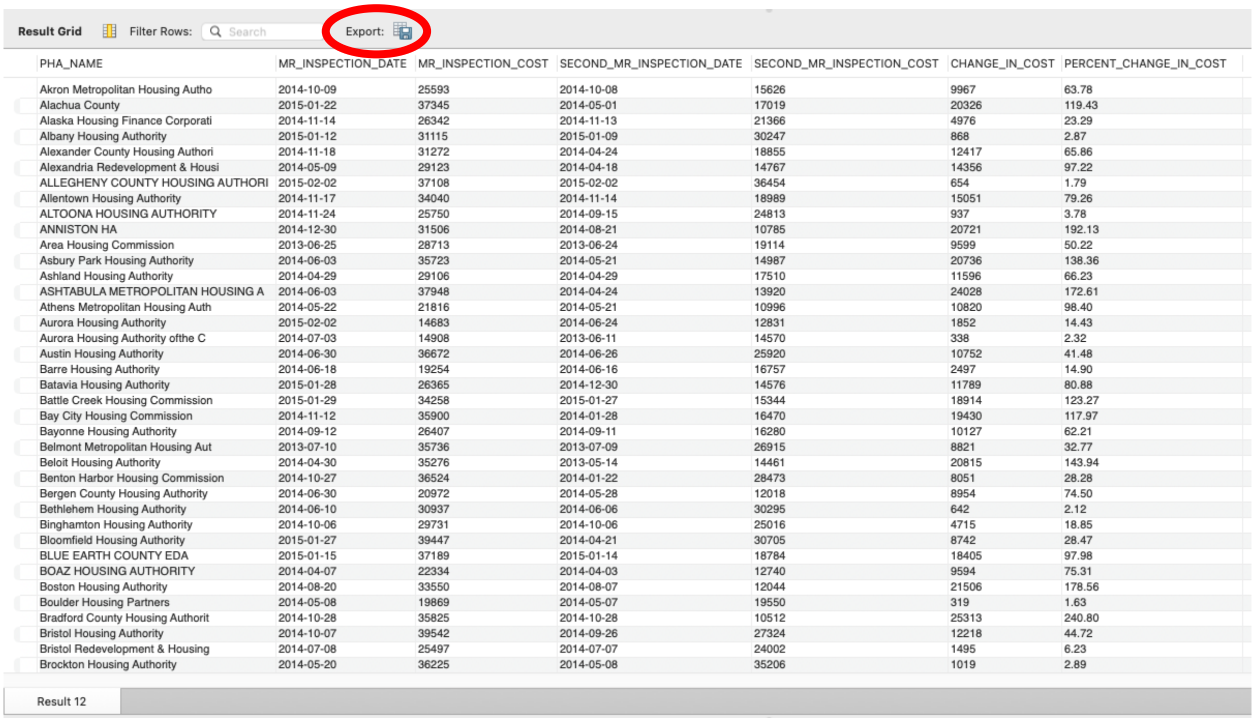
```
WITH formatted_data AS (
  SELECT
    PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
    STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') AS INSPECTION_DATE,
    COST_OF_INSPECTION_IN_DOLLARS AS INSPECTION_COST
  FROM public_housing_inspections.public_housing_inspection_data
),
lead_data AS (
  SELECT
    PHA_NAME,
    INSPECTION_DATE AS MR_INSPECTION_DATE,
    INSPECTION_COST AS MR_INSPECTION_COST,
    LEAD(INSPECTION_DATE) OVER (PARTITION BY PHA_NAME ORDER BY INSPECTION_DATE DESC) AS SECOND_MR_INSPECTION_DATE,
    LEAD(INSPECTION_COST) OVER (PARTITION BY PHA_NAME ORDER BY INSPECTION_DATE DESC) AS SECOND_MR_INSPECTION_COST
  FROM formatted_data
),
cost_increases AS (
  SELECT
    *,
    (MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) AS CHANGE_IN_COST,
    ROUND(100.0 * (MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) / SECOND_MR_INSPECTION_COST, 2) AS PERCENT_CHANGE_IN_COST
  FROM lead_data
  WHERE SECOND_MR_INSPECTION_COST IS NOT NULL
    AND MR_INSPECTION_COST > SECOND_MR_INSPECTION_COST
),
most_recent_increase_per_pha AS (
  SELECT *,
         ROW_NUMBER() OVER (PARTITION BY PHA_NAME ORDER BY MR_INSPECTION_DATE DESC) AS rn
  FROM cost_increases
)

SELECT
  PHA_NAME,
  MR_INSPECTION_DATE,
  MR_INSPECTION_COST,
  SECOND_MR_INSPECTION_DATE,
  SECOND_MR_INSPECTION_COST,
  CHANGE_IN_COST,
  PERCENT_CHANGE_IN_COST
FROM most_recent_increase_per_pha
WHERE rn = 1;
```

Results

The final output includes the following columns: PHA_NAME, MR_INSPECTION_DATE, MR_INSPECTION_COST, SECOND_MR_INSPECTION_DATE, SECOND_MR_INSPECTION_COST, CHANGE_IN_COST, and PERCENT_CHANGE_IN_COST.

Only PHAs with **at least two inspections**, and a **cost increase** are included in the result. Each PHA appears **exactly once**. This dataset can then be exported from MySQL Workbench as a CSV file for management review.



PHA_NAME	MR_INSPECTION_DATE	MR_INSPECTION_COST	SECOND_MR_INSPECTION_DATE	SECOND_MR_INSPECTION_COST	CHANGE_IN_COST	PERCENT_CHANGE_IN_COST
Akron Metropolitan Housing Autho	2014-10-09	25593	2014-10-08	15626	9967	63.78
Alachua County	2015-01-22	37345	2014-05-01	17019	20326	119.43
Alaska Housing Finance Corporati	2014-11-14	26342	2014-11-13	21366	4976	23.29
Albany Housing Authority	2015-01-12	31115	2015-01-09	30247	868	2.87
Alexander County Housing Authori	2014-11-18	31272	2014-04-24	18855	12417	65.86
Alexandria Redevelopment & Housi	2014-05-09	29123	2014-04-18	14767	14356	97.22
ALLEGHENY COUNTY HOUSING AUTHORI	2015-02-02	37108	2015-02-02	36454	654	1.79
Allentown Housing Authority	2014-11-17	34040	2014-11-14	18989	15051	79.26
ALTOONA HOUSING AUTHORITY	2014-11-24	25750	2014-09-15	24813	937	3.78
ANNISTON HA	2014-12-30	31506	2014-08-21	10785	20721	192.13
Area Housing Commission	2013-06-25	28713	2013-06-24	19114	9599	50.22
Asbury Park Housing Authority	2014-06-03	35723	2014-05-21	14987	20736	138.36
Ashland Housing Authority	2014-04-29	29106	2014-04-29	17510	11596	66.23
ASHTABULA METROPOLITAN HOUSING A	2014-06-03	37948	2014-04-24	13920	24028	172.61
Athens Metropolitan Housing Auth	2014-05-22	21816	2014-05-21	10996	10820	98.40
Aurora Housing Authority	2015-02-02	14683	2014-06-24	12831	1852	14.43
Aurora Housing Authority ofthe C	2014-07-03	14908	2013-06-11	14570	338	2.32
Austin Housing Authority	2014-06-30	36672	2014-06-26	25920	10752	41.48
Barre Housing Authority	2014-06-18	19254	2014-06-16	16757	2497	14.90
Batavia Housing Authority	2015-01-28	26365	2014-12-30	14576	11789	80.88
Battle Creek Housing Commission	2015-01-29	34258	2015-01-27	15344	18914	123.27
Bay City Housing Commission	2014-11-12	35900	2014-01-28	16470	19430	117.97
Bayonne Housing Authority	2014-09-12	26407	2014-09-11	16280	10127	62.21
Belmont Metropolitan Housing Aut	2013-07-10	35736	2013-07-09	26915	8821	32.77
Beloit Housing Authority	2014-04-30	35276	2013-05-14	14461	20815	143.94
Benton Harbor Housing Commission	2014-10-27	36524	2014-01-22	28473	8051	28.28
Bergen County Housing Authority	2014-06-30	20972	2014-05-28	12018	8954	74.50
Bethlehem Housing Authority	2014-06-10	30937	2014-06-06	30295	642	2.12
Binghamton Housing Authority	2014-10-06	29731	2014-10-06	25016	4715	18.85
Bloomfield Housing Authority	2015-01-27	39447	2014-04-21	30705	8742	28.47
BLUE EARTH COUNTY EDA	2015-01-15	37189	2015-01-14	18784	18405	97.98
BOAZ HOUSING AUTHORITY	2014-04-07	22334	2014-04-03	12740	9594	75.31
Boston Housing Authority	2014-08-20	33550	2014-08-07	12044	21506	178.56
Boulder Housing Partners	2014-05-08	19869	2014-05-07	19550	319	1.63
Bradford County Housing Authorit	2014-10-28	35825	2014-10-28	10512	25313	240.80
Bristol Housing Authority	2014-10-07	39542	2014-09-26	27324	12218	44.72
Bristol Redevelopment & Housing	2014-07-08	25497	2014-07-07	24002	1495	6.23
Brockton Housing Authority	2014-05-20	36225	2014-05-08	35206	1019	2.89