

Entity Resolution, ER Diagram, Normalization, Hampel X84

Instructor: Aditya Parameswaran

Covers Lecture(s): 13 – 16

I Entity Resolution

In this section, we will work with an album sales dataset. Here is a sneak peek:

name	artist	track_count	name	day	count
Red (Deluxe Edition)	Taylor Swift	22	After LIKE	1	2
thank u, next	Ariana Grande	12	After Like	2	-1
Bloom	Red Velvet	11	After like (cd)	4	7
After LIKE	IVE	2	After LIKE - ive	6	7
			Red - Deluxe	5	-1
albums			sales		

One problem with this dataset is that the album names are not standardized. For example, “After Like”, “After LIKE”, “After Like (CD)” all refer to the same album, but their string representations are not exactly the same.

1. Write a query to join the *albums* table with the *sales* using strict equality condition. What is going to go wrong with this approach?

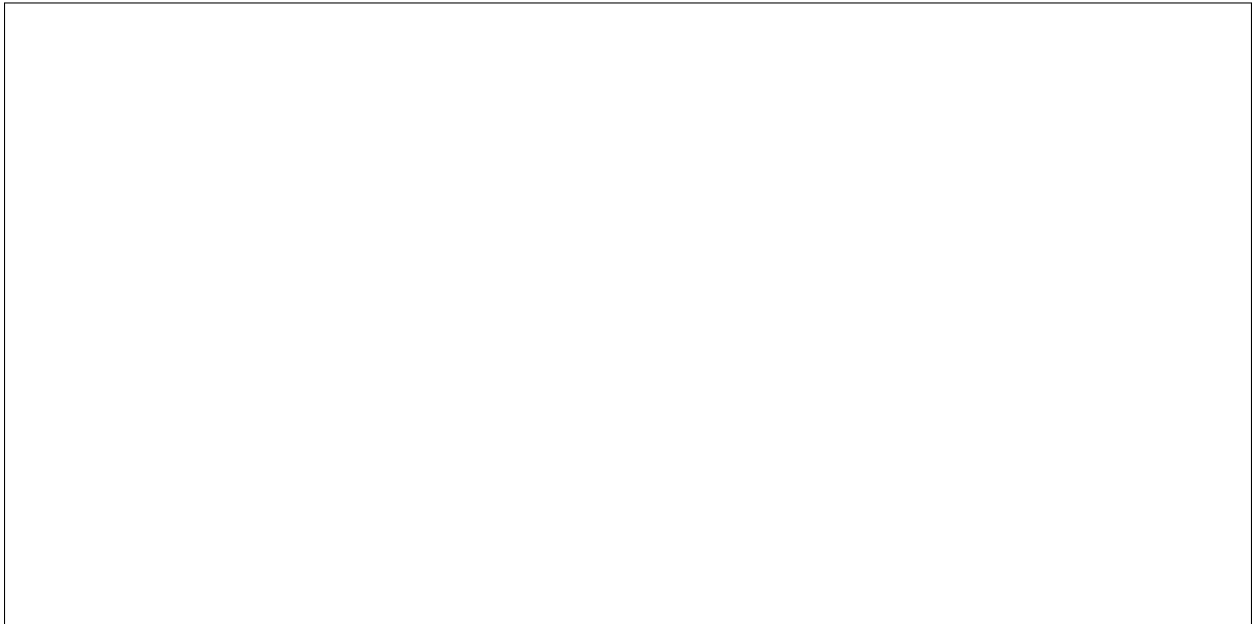
2. One solution is to use the Levenshtein distance. The `levenshtein()` distance function can be helpful to find strings that are similar but not identical. This algorithm computes the minimal number of insertions, deletions, and mutations. Let us apply this to our dataset. Write a query to compute the Levenshtein distance between every pair of names in the two tables, sorted by this distance.

3. Create a materialized view `joined_sales` that joins the tables on the condition that two entries match if their Levenshtein distance is less than 10.

II Entity Relationship Diagrams

4. We want to store sports teams and their players in our database. Draw an ER diagram corresponding to data given below:

- Every **team** has a unique team_name and a stadium where they play.
- Each **coach** has a unique name.
- Each **player** has a unique player_id, a name and an average score.
- A **player** *plays for* a **team**. This relationship also records the position of the player. A player can play at most one team, but may also be a free agent and not play on any team. A team needs at least 1 player.
- A **player** can *captain* at most one **team**. A team has exactly one captain.
- A **coach** can *coach* at most one **team**. A team may have many coaches, and needs at least 1 coach.



5. Translate the ER diagram into a relational schema.



III Functional Dependency and Normalization

A functional dependency (FD) is a form of constraint between 2 sets of attributes in a relation.

- Two attributes satisfy $X \rightarrow Y$ if and only if for every pair of tuples (t, u) , the following holds: if $t.X = u.X$, then $t.Y = u.Y$.

- Three attributes satisfy $XY \rightarrow Z$ if and only if for every pair of tuples (t, u) , the following holds: if $t.X = u.X$ and $t.Y = u.Y$, then $t.Z = u.Z$.

For this section, consider the student_course_info table.

student_id	student_name	course_id	course_name	professor
101	Alice	C001	Math	Dr. Smith
102	Bob	C001	Math	Dr. Smith
103	Charlie	C002	English	Dr. Brown
104	Dave	C002	English	Dr. Brown
105	Eve	C003	History	Dr. Clark

Table 1: student_course_info

6. Write down three functional dependencies that **may hold** based on the schema and the sample data.

7. Perform normalization on this table. Your result should have 3 tables: **students**, **courses**, **enrollments**.

8. Suppose we know that $X \rightarrow Y$ and $Y \rightarrow Z$. Does $X \rightarrow Z$ hold? Prove your answer using contradiction.

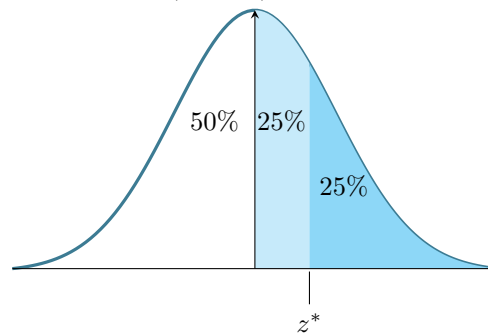
IV [Optional] Hampel X84

Hampel X84 is a robust outlier detection technique because it has an ideal breakdown point of 50% (the proof of the breakdown point is beyond the scope of this course).

9. In this exercise, we will prove the following statement: selecting outliers 1.4826 MAD away from the median is equivalent to selecting outliers 1 standard deviation away from the mean, if the data follows a normal distribution.

- (a) Write down the definition of Median Absolute Deviation (MAD). What is the intuitive meaning of MAD?

- (b) Let us try to derive the magic number 1.4826. First, let us find the MAD of a standard normal distribution. Let Z be a random variable following a standard normal distribution with $\mu = 0$ and $\sigma = 1$, find z^* such that $P(Z < z^*) = 0.75$. Is z^* the MAD? Hint: Use `scipy.stats.norm`.



- (c) Continuing the derivation of 1.4826, what is the mathematical relation between the MAD and the standard deviation in a normal distribution?

10. Now let us apply the Hampel X84 method on the following dataset to detect outliers: $\{2, 4, 6, 8, 9, 11, 30\}$.

- (a) Compute the median.

(b) Compute the MAD.

(c) Find the outliers using the Hampel X84 method.