

CISC 3810 Database Systems

HW 7: *k-Nearest Neighbors for Similarity*

Our company is doing well in the greater Los Angeles area, and wants to open some new locations.

Our most successful location is in the Los Angeles zip code 93591.

We have a list of 319 zipcodes in the area that we can consider for new locations.

Your job is to recommend the 10 best candidate zip codes for us to consider, based on their *demographic* similarity to our most successful location (zip 93591).

To find similarity, we will define a *distance function* that measures demographic similarity, and apply that function to find the demographic distance between our most successful zip code and every other zip code.

For this case, we will treat three properties like physical dimensions, and use them to compute the *Euclidian Distance* between zip codes.

The three properties that we will use are the following¹ :

(Median Age, Male Percentage of the Population, Average Household Size)²

The *Euclidian distance*, $d(\text{point } \mathbf{p}, \text{point } \mathbf{q})$, is defined as follows:

$$\begin{aligned} d(\mathbf{p}, \mathbf{q}) &= d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}. \end{aligned}$$

This is just the good old-fashioned geometric distance formula, a.k.a. the Pythagorean Theorem.

You can now think of each zip code as being a point in 3D space with coordinates (x, y, z), where

x = median age,
y = male percentage of the population, and
z = average household size.

Once we compute the distance between each candidate zip code and our target zip code (93591), all we need to do is order the zip codes by their demographic "distance" from 93591, and take the 10 smallest distances.

These are the 10 "nearest neighbors" in 3D feature space - where our features are *demographic properties*.

Your Assignment:

Download the .csv data file.
Create a table with the proper data types for the census data.
Load the data from the .csv file into your table.
Compute the 3-Dimensional demographic distance
 between each zip code and zip code 93591.
Sort the results and report the Top 10 closest zip codes by demographic distance.
 These will be your recommended candidate locations.

Hints:

Watch your integer division vs. float division - you may have to cast a few things to float.
Watch for division by zero - you may need to impose some conditions WHERE columns != 0.

Begin by selecting only the values for zip code 93591 into a table (the 'target' table).

Make sure you have the male (or female) population percentage in your 93591 table
 AND also in the main "census" table - which contains all the rest of the zip codes.

SELECT the *distance computation* from your "target" table *CROSS JOIN*ed with
 the "census" table. This will give you the distance between the zip code 93591
 and every other zip code.

From there, you've done the hard part - now just order by distance, and take the top 10.

*A file with the correct answers has been provided, and you can use it as a guide to compare
 your answers against.*

Requirements:

Submit: one .sql file with the queries you used to get your answers,
 and also one separate .txt file with your final output.

**ALL HOMEWORK SUBMISSIONS WHICH ARE COPIED AND PASTED INTO THE
 BODY OF THE EMAIL WILL RECEIVE ZERO CREDIT.**

-
1. For simplicity, we will not adjust the scale of any of the dimensions. In real life,
 we would need to *scale* the dimensions to be of similar size. I.e., adjust them all to be
 nearly equal in magnitude, so that some dimensions do not heavily outweigh the others,
 due to large differences in magnitude.
 2. Note that if we have percent males, we also have percent females - they are not independent.
 You can use either one.