**Nearest Neighbors Approach**

For finding similar companies

Objective:   
To find similar companies by using nearest neighbors’ approach.

# Approach:

The is a two-step process:

Step 1: Selecting the right population using Taxpayer Segmentation.

1. Nature of Business

Please note - *Taxpayer Type* has been excluded from the selection criteria, because of the data issues.

Step 2: Nearest Neighbor Algorithm

Similarity of taxpayers is gauged based on the Euclidean distance amongst them, calculated based on weighted distances measured along several dimensions. The dimensions chosen for calculating the nearest neighbors distance are:

|  |  |  |
| --- | --- | --- |
| **Dimension** | **Distance** | **Weight** |
| Turnover (Normalized) | D1 = difference between company’s Normalized Turnover | W3=0.4 |
| GP Ratio (Normalized) | D2 = difference between company’s Normalized GP ratios | W4=0.4 |
| Profit (Normalized) | D3 = difference between company’s Normalized Profit | W5=0.1 |
| Geography (Pincode Category- MMI) | D4 =   * 0 if the company’s pincode category is same * 1 if the company’s pincode category is different | W6=0.1 |

*Normalization:*

Normalized value of a particular feature **X** has been calculated using the formula below:

**X\_Normalized = [X-MIN(X)] / [MAX(X)-MIN(X)]**

Where **MIN(X)** is minimum value of feature **X** in X population and **MAX(X)** is maximum value of feature **X** in X population.

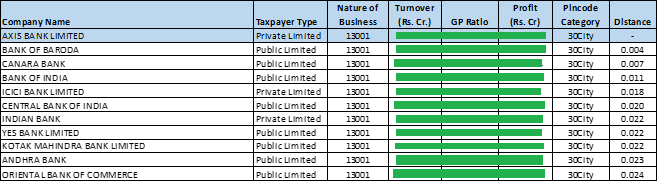
*Weighted Euclidean Distance:*

The nearest neighbors distance is calculated as –

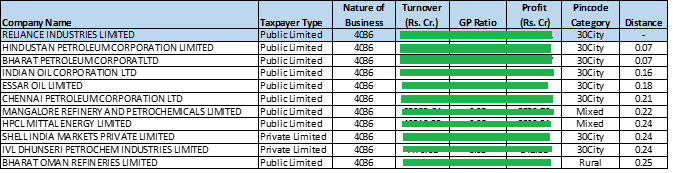
SQRT ((W1\*D1)\*\*2 + (W2\*D2)\*\*2+(W3\*D3)\*\*2+(W4\*D4)\*\*2)

# Outcome: Top 10 similar companies are populated for the concerned PAN based on the approach above. Please find two working examples below.

**Example 1:** Axis Bank (based on PAN search)

****

**Example 2:** Reliance Industries (based on PAN search)

****

# Future Scope:

Optimization can be refined at three levels:

1. Weights
2. Distance
3. Additional Parameters