**TN Marginal Workers Assessment**

**Phase 2: Innovation**

**Objective:**

To conduct clustering analysis to identify patterns among different industrial categories and age groups of marginal workers in Tamil Nadu.

**1. Data Preprocessing:**

* Purge any incomplete information and anomalous events from the accumulated dataset.
* Convert ordinal variables into numeric representation such as one-hot encoding and label encoding.

**2. Clustering Analysis:**

* Choose suitable clustering algorithms, including K-means, hierarchical clustering, and DBSCAN.
* Estimate the best number of clusters by means of methods like as the elbow method or silhouette analysis.
* Use the selected clustering algorithm to find similarities between industries and different age groups of margin workers in the pre-processed data.
* Use techniques like scatter plots and heatmaps, for instance, visualization of clustering results.

**3. Interpretation and Analysis:**

* Examine the clustering results to ascertain attributes that may define separate clusters.
* Compare the characteristics of marginal workers in various classifications with regard to age groups.

**Innovative design**

For the specific project of assessing marginal workers in Tamil Nadu based on different industrial categories and age groups, a \*\*combination of K-Means Clustering and Hierarchical Clustering\*\* methods would be suitable:

**1. K-Means Clustering:**

**Suitability:** The K-means is a viable method for dividing data into separate, non-overlapping subsets. This seems to be the case for your project, since you have shown an interest in specific industrial categories and age groups, which means that the number of clusters can be reasonable predefined.

**Application:** K-means can be used to cluster marginal workers on numerical features that include age, income, education level, and years of experience. It enables you to detect distinct numbers within these numerical attributes.

**2. Hierarchical Clustering:**

**Suitability:** If you want to look at hierarchical structure of data, use hierarchical clustering. Unlike others that call for pre-specification of the number of clusters, it is beneficial in identifying broader patterns and sub-patterns regarding the marginal workers.

**Application:** The hierarchical clustering can be used to determine clusters at multiple levels of granularity. For instance, a firm may start with broad clusters that categorically represent various industrial categories before further subdividing these clusters into sub-clusters that represent different age groups within each category.

Integrating both will provide overall picture of the data. For numerical attributes, K-means should be used to identify specific patterns in age groups and income levels. At the same time, use hierarchical clustering to investigate broader patterns, like different types of industry, and then continue to narrow these categories to the level of age segments.