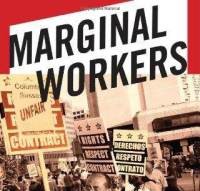
**Assessment of Marginal Workers in TamilNadu**

**TEAM MEMBER :** RAMKUMAR.R

**PROJECT TITLE:** **TN Marginal Workers Assessment**



**OBJECTIVE :**

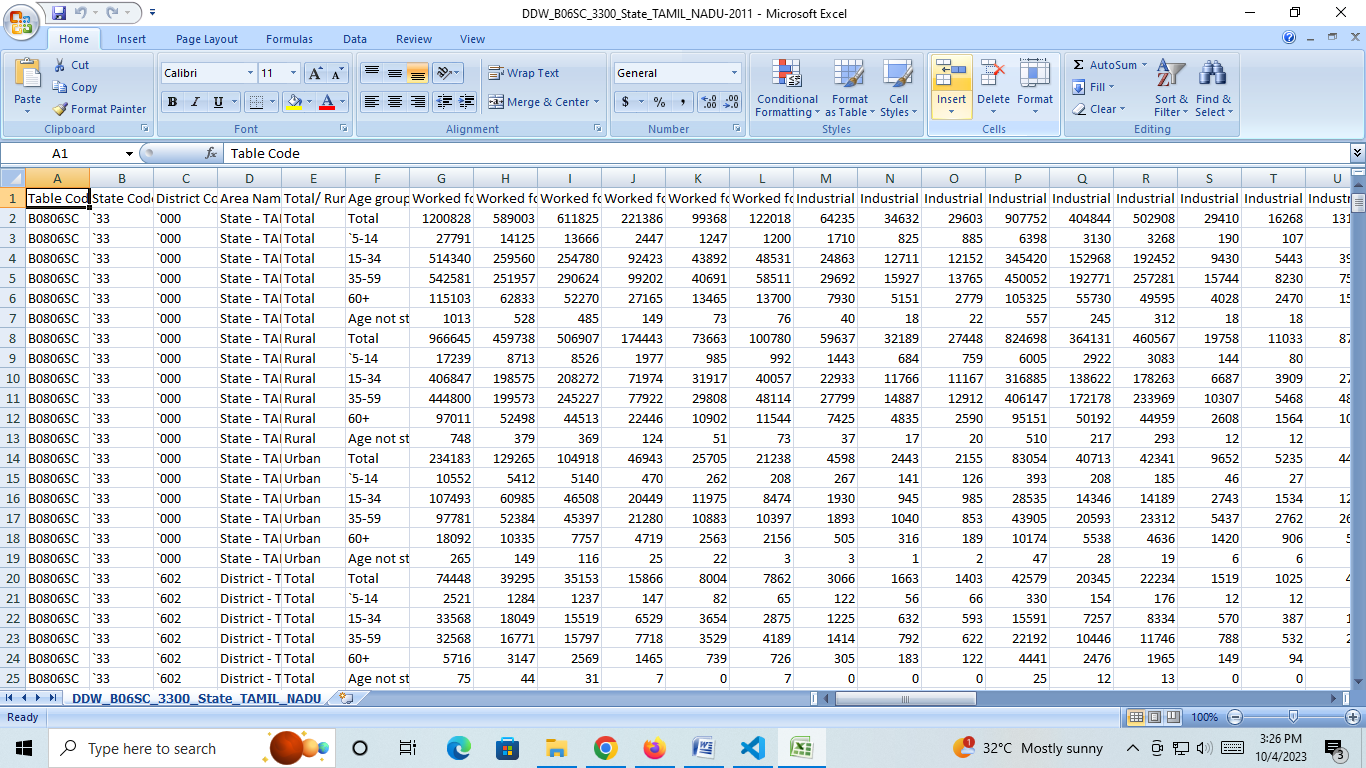
The Objective of this project is conducting clustering analysis to identify patterns among different industrial categories and age groups.

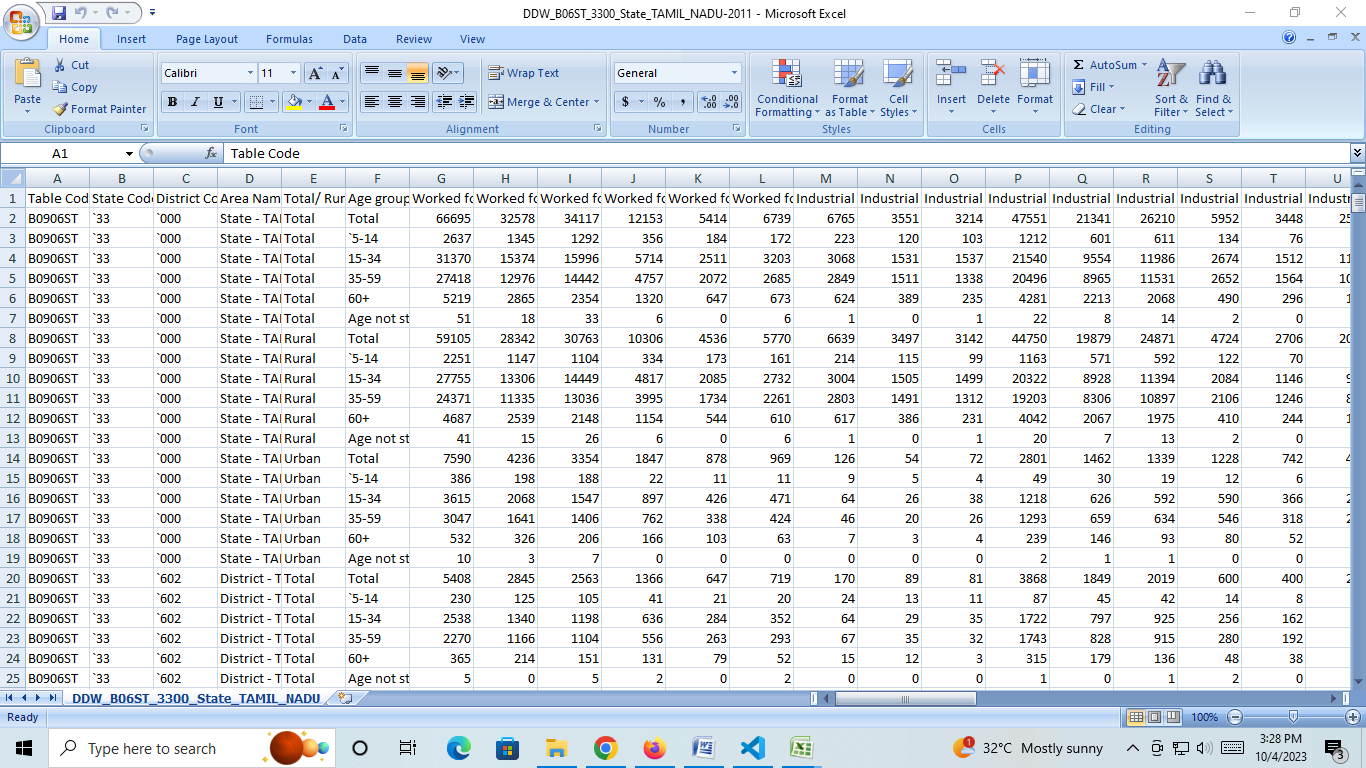
**Content For Project Phase 2 :**

Consider conducting clustering analysis to identify patterns among different industrial categories and age groups.

**Data Source :**

**Data Source :** [**https://tn.data.gov.in/catalog/marginal-workers-classified-age-industrial-category-and-sex-census-2011-india-and-states**](https://tn.data.gov.in/catalog/marginal-workers-classified-age-industrial-category-and-sex-census-2011-india-and-states)

****

****

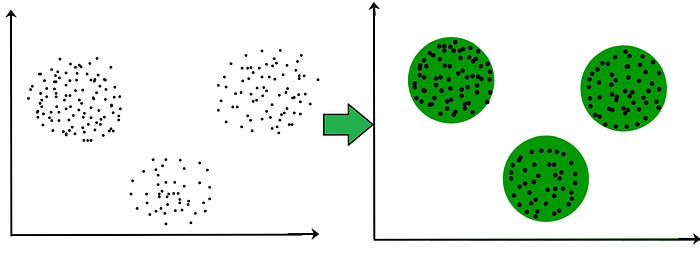
**Project Steps:**

**Phase 2: Innovation**

* Consider conducting clustering analysis to identify patterns among different industrial categories and age groups.
* The most common use of cluster analysis is classification. Subjects are separated into groups so that each subject is more similar to other subjects in its group than to subjects outside the group.
* We will initially focus on clustering procedures that result in the assignment of each subject to one, and only one, class. Subjects within a class are usually assumed to be indistinguishable from one another. Thus, we assume that the underlying structure of the data involves an unordered set of discrete classes. In some cases we may also view these classes as hierarchical in nature, with some classes divided into subclasses. Clustering procedures can be viewed as “pre-classificatory” in the sense that the researcher has not used prior judgment to partition the subjects (rows of the data matrix). However, it is assumed that some of the objectives are heterogeneous; that is, that “clusters” exist.

**INTRODUCTION :**

* Cluster analysis, like reduced space analysis (factor analysis), is concerned with data matrices in which the variables have not been partitioned beforehand into criterion versus predictor subsets.
* The objective of cluster analysis is to find similar groups of subjects, where “similarity” between each pair of subjects means some global measure over the whole set of characteristics. In this article we discuss various methods of clustering and the key role that distance plays as measures of the proximity of pairs of points.



**Algorithms for cluster analysis**

* Cluster analysis uses machine learning algorithms to group things, in our case, customers, into similar groups or clusters.
* There are two main algorithms used in clustering, K-Means analysis and K-Medoids analysis.

**K-Means cluster analysis**

The K-means algorithm divides a single cluster into K different clusters. It does this by finding organically similar data points and assigning each one to a cluster with similar characteristics. K-means clustering works by constantly trying to find a centroid (a data point that represents the mean or center of the cluster). The end clusters will each have a centroid and data points that are closer to the centroid compared to the other centroids.

There are pros and cons to using K-means cluster analysis:

Pros

* Simple, popular method
* Guarantees convergence
* Offers a good estimate of centroids’ initial positions

Cons

* You must specify the number of clusters
* Depends on random initial values, so it may be inconsistent in different runs
* Data may need to be scaled before clustering

K-means cluster analysis is widely used across several verticals, from determining urban traffic patterns for Uber drivers to segmenting customers based on interests, purchase history, or buying behaviors.

**K-Medoids cluster analysis**

The K-medoids algorithm is similar to the K-means algorithm, but instead of using centroids, it establishes medoids (the least dissimilar data points). The medoid is an actual point in the data set, where the centroid in K-means is an average (mean). K-medoids cluster analysis is more accurate because it is not sensitive to outliers like K-means can be.

There are pros and cons to using K-medoids cluster analysis:

Pros

* Easy to understand and implement
* Fast
* Less sensitive to outliers

Cons

* May have differing results in runs because the first k-medoids are chosen randomly
* Not suitable for non-spherical groups because it focuses on the proximity of the data points rather than the connectivity

The K-medoids algorithm is used in facial recognition software because it uses an actual data point and is more robust to outliers. If K-means were used, it would be less effective because there is no real data point to start with, but rather a centroid with mixed features from several photos.

**Program:**

**Assessment of Marginal Workers in TamilNadu**

import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.preprocessing import StandardScaler  
from sklearn.metrics import silhouette\_score  
from sklearn.cluster import KMeans  
from matplotlib import pyplot

# Load dataset  
df = pd.read\_csv("/Users/nailamolooicloud.com/Downloads/Mall\_Customers.csv")  
del df['CustomerID']  
print(df.head)

**OUTPUT :**

**Gender Age Annual Income (k$) Spending Score (1–100)  
Male 19 15 39  
Male 21 15 81  
Female 20 16 6  
Female 23 16 77  
Female 31 17 40**

**Submission:**

* Share the GitHub repository link containing the project's code and dataset.
* Provide instructions on how to replicate the analysis, load the dataset, perform calculations, and create visualizations using Python.
* Include a summary of the key findings from the demographic analysis and visualizations.