**IBM NAANMUTHALVAN**

**PHASE 4 PROJECT**

**TEAM MEMBERS:**

**A.AHAMED IBRAHIM(820421205002)(samadhasanalthauf@gmail.com)**

**J.JAYACHANDAR(820421205033)(bugal890@gmail.com)**

**K.HARI KRISHNA(820421205027)(harikrishna22032004@gmail.com)**

**S.BALAMURUGAN(820421205013)(balamurugan610001@gmail.com)**

**DOMAIN NAME :**

**DATA ANALYTICS WITH IBM COGNOS**

**PROJECT NAME:**

**ASSESSMENT OF TN MARGINAL WORKERS –A SOCIO-ECONOMIC ANALYSIS.**

**1. Introduction**

**1.1 Project Overview**

* Perform the demographic analysis and create visualizations.
* Calculate the distribution of marginal workers based on age, industrial category, and sex using data aggregation and manipulation.
* Create visualizations using data visualization libraries (e.g., Matplotlib)

**1.2 Data Source**

**DatasetLink:**[**https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil**](https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil)

**2. Data Exploration**

**2.1 Data Loading**

Data loading is the process of bringing external data into your program or environment so that you can work with it in your analysis, processing, or modeling tasks**.** We use file I/O operations to open and read the data like pandas.

**2.2 Data Cleaning**

Data cleaning, also known as data preprocessing, is a critical step in the data analysis process. It involves identifying and addressing issues or inconsistencies in your dataset to ensure that the data is accurate, complete, and in the right format for analysis.

Data cleaning tasks using Python and the Pandas library, which is a popular choice for data manipulation and cleaning

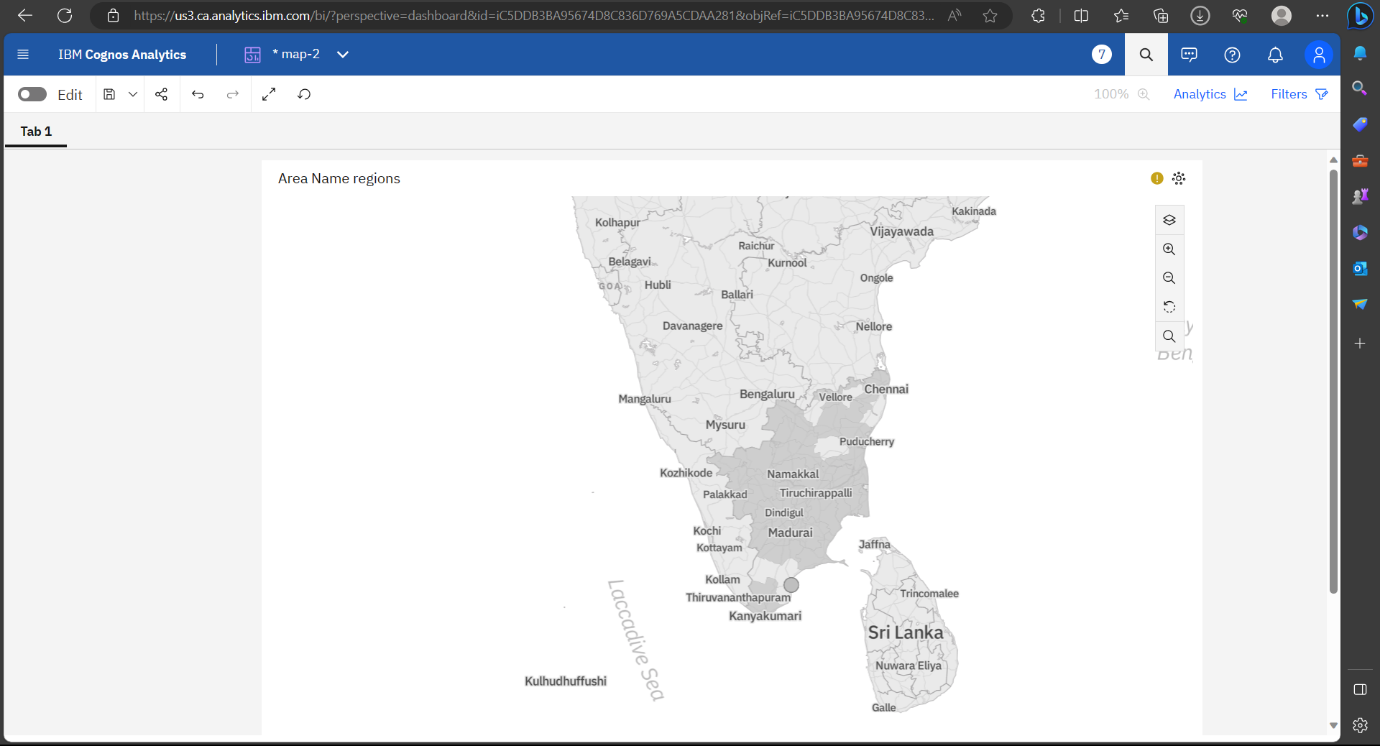
**3.Data Visualization**

**3.1 Visualizations**

Data visualization is a critical component of data analytics. It involves representing data graphically to help analysts and decision-makers understand the data's patterns, relationships, and insights more easily.

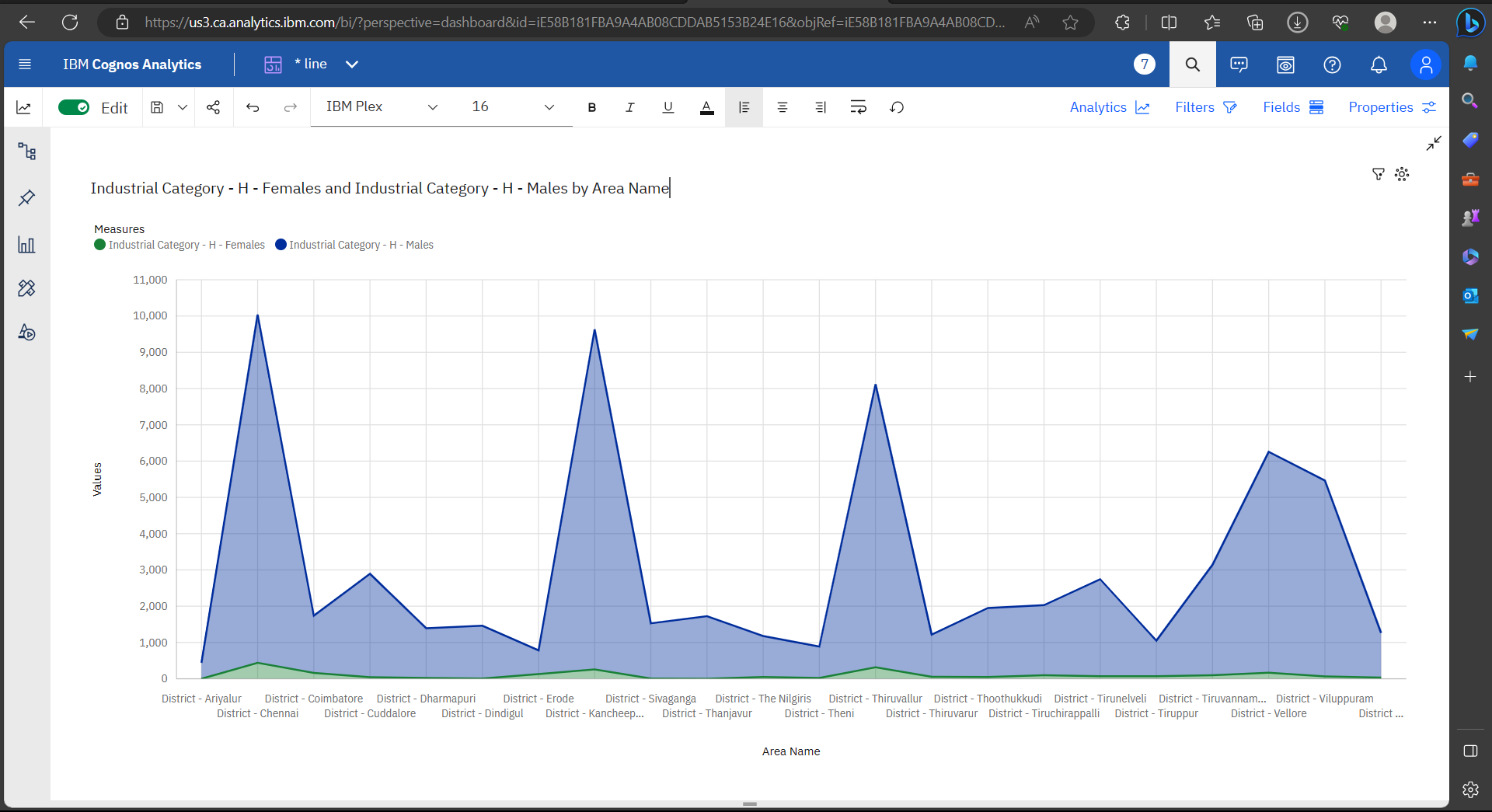
With the help of IBM Cognos, we use map for distribution of worker in different regions, use line charts, Box plot, Point graph, Point graph and Driver analysis and Include code and visualizations for age, industrial category, and sex using python.

**Distribution of workers by region:**



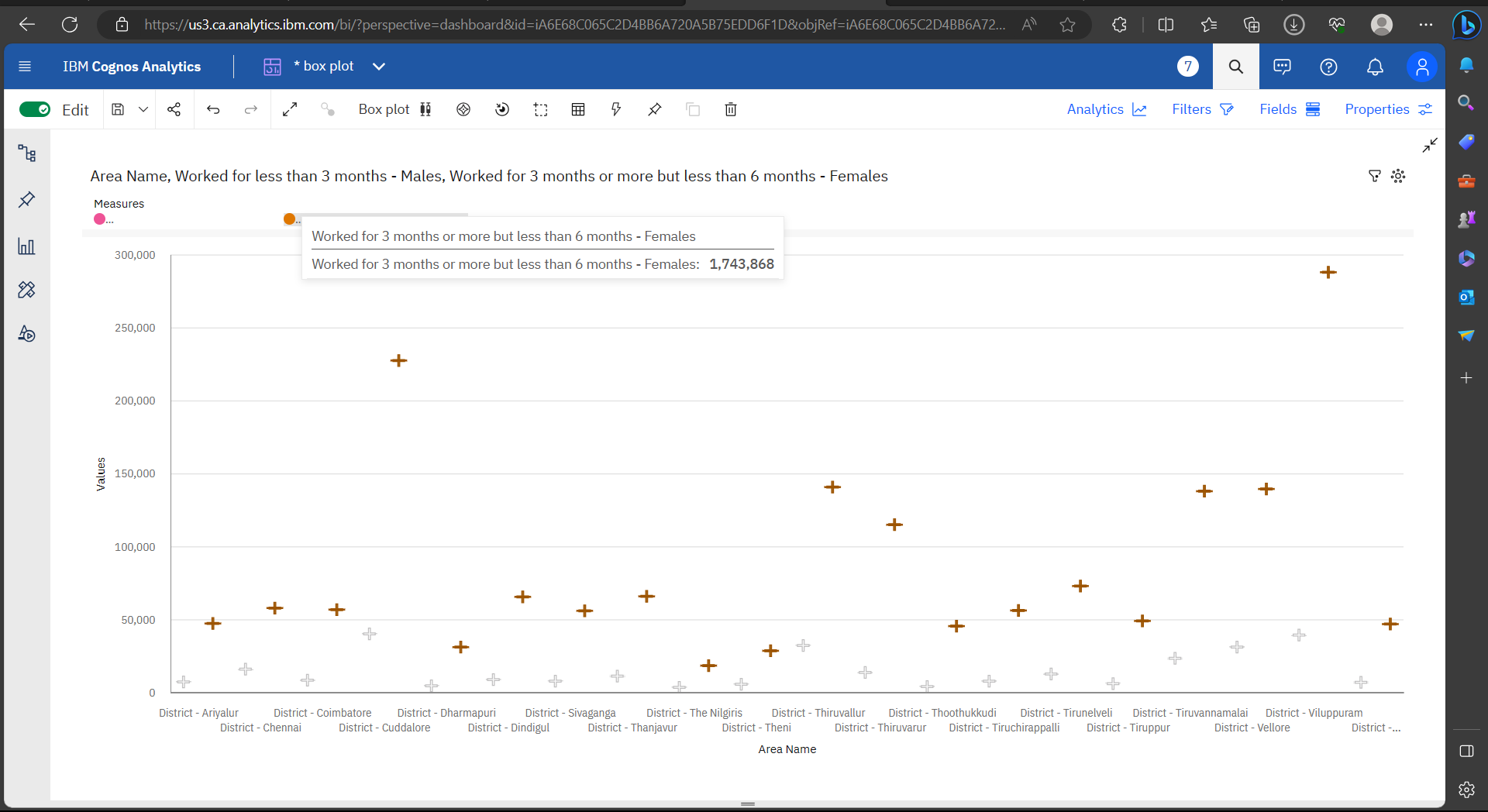
**Line graph:**

The demographic analysis visualization.

Industrial Category - H - Females and Industrial Category - H - Males by Area Name

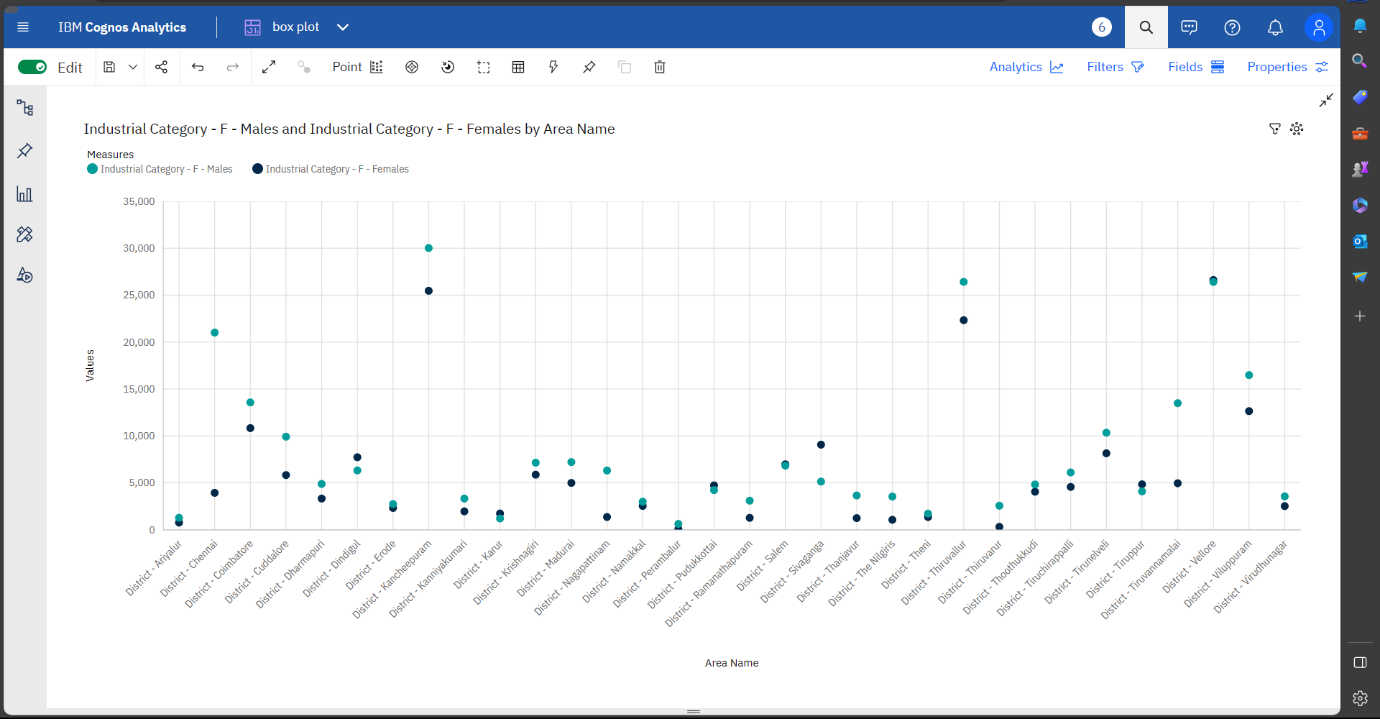
**Box plot**

Area Name, Worked for less than 3 months - Males, Worked for 3 months or more but less than 6 months - Females



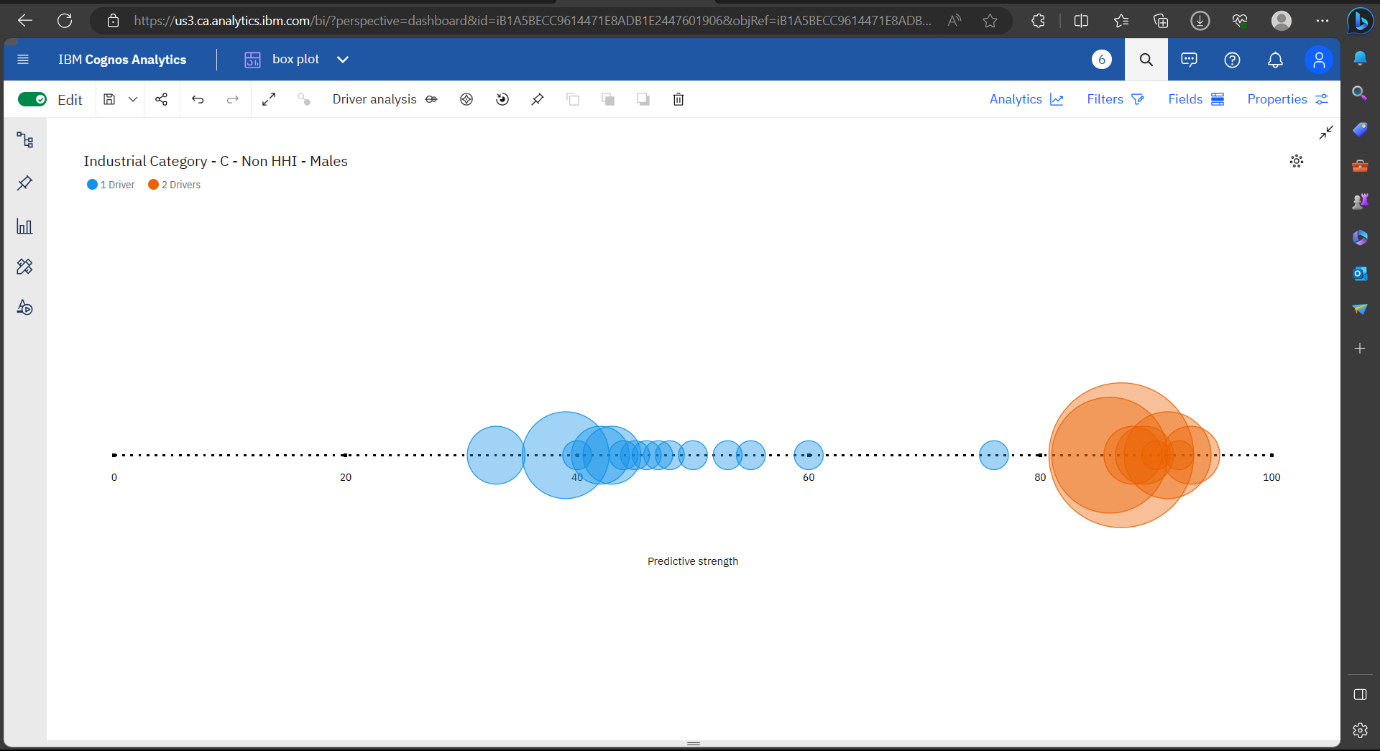
**Point graph:**

Point graph for Industrial Category - F - Males and Industrial Category - F - Females by Area Name



**Driver analysis**

Driver analysis for Industrial Category - C - Non HHI – Males

`

**DATA VISUALIZATION USING PYTHON**

import pandas as pd

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

from sklearn.cluster import Birch

csv\_file\_path = 'Project.csv'

data = pd.read\_csv(csv\_file\_path)

print(df.isnull().sum())

**output:**

Table Code 0

State Code 0

District Code 0

Area Name 0

Total/ Rural/ Urban 0

..

Industrial Category - R to U - HHI - Males 0

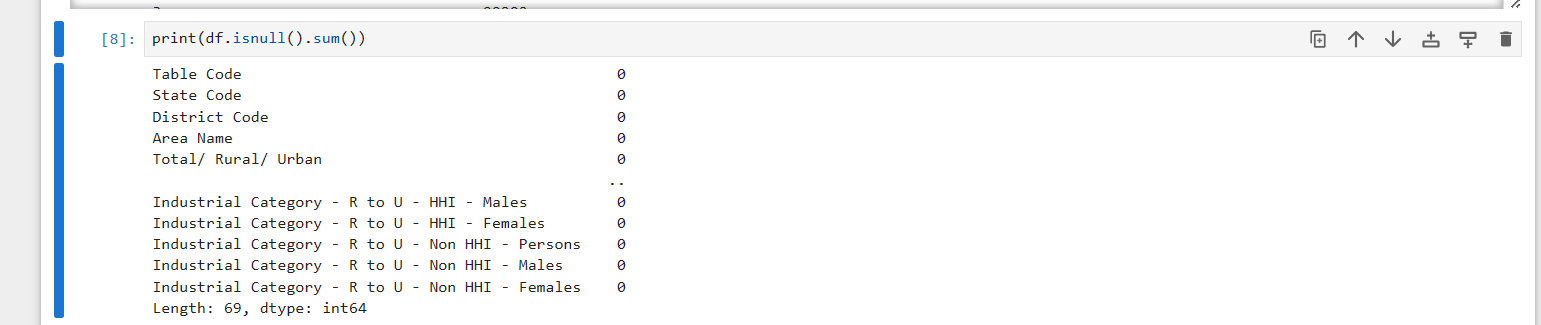
Industrial Category - R to U - HHI - Females 0

Industrial Category - R to U - Non HHI - Persons 0

Industrial Category - R to U - Non HHI - Males 0

Industrial Category - R to U - Non HHI - Females 0

Length: 69, dtype: int64



data=df.dropna()

data = data[(data['Age group'] != 'Total') & (data['Age group'] != 'Age not stated')]

data['Age group'] = data['Age group'].apply(lambda x: (int(x.split('-')[0]) + int(x.split('-')[1])) // 2 if '-' in x else int(x))

X = data[['Age group']].values

Y=data[['Industrial Category - A - Cultivators - Females']].values

Z=data[['Industrial Category - A - Cultivators - Males']].values

birch = Birch(threshold=5.5, n\_clusters=2)

birch.fit(X)

cluster\_labels = birch.predict(X)

data['cluster'] = cluster\_labels

fig = plt.figure()

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(X, Y, Z, c= data['cluster'], cmap='RdBu')

ax.set\_xlabel('Age group')

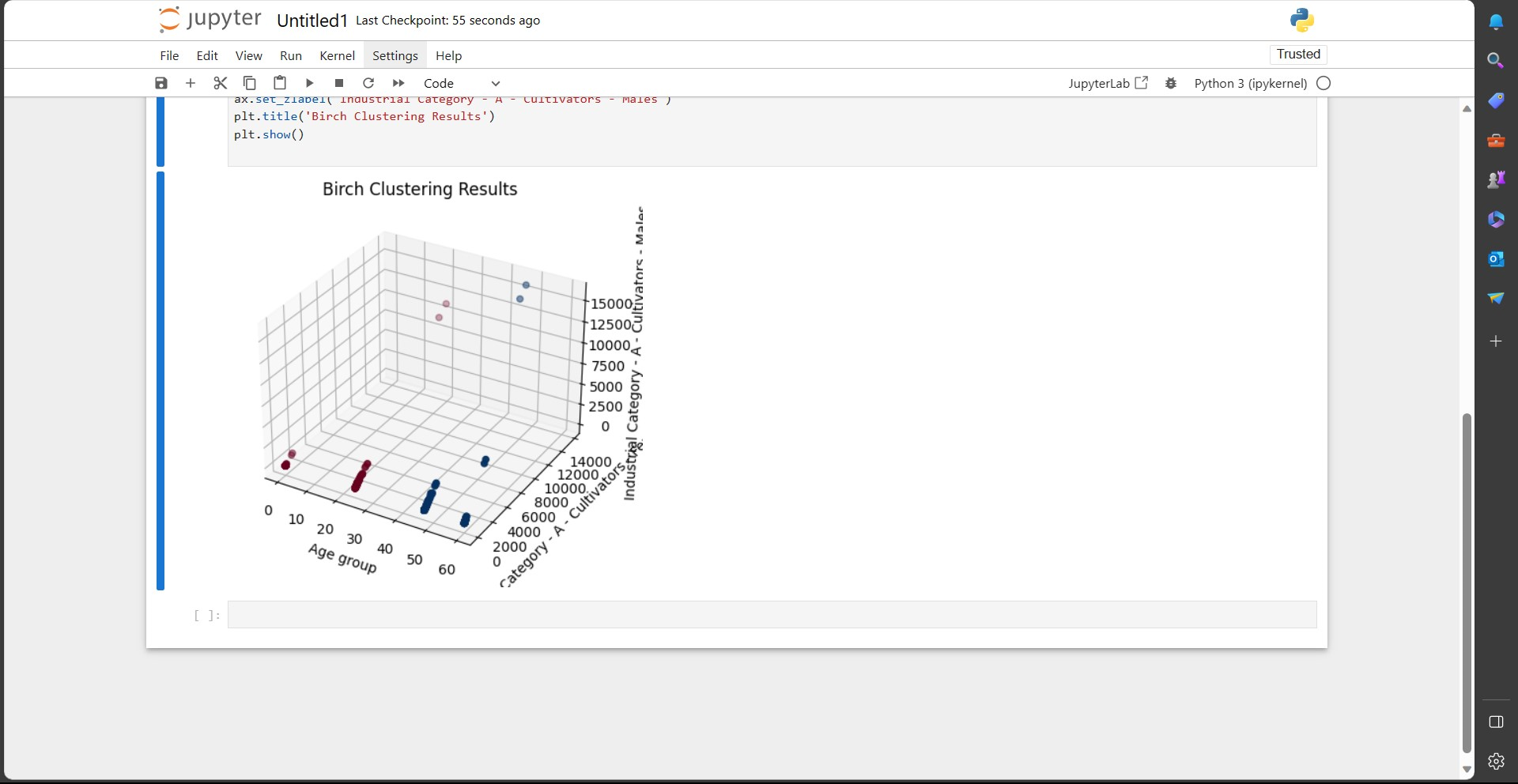
ax.set\_ylabel('Industrial Category - A - Cultivators - Females')

ax.set\_zlabel('Industrial Category - A - Cultivators - Males')

plt.title('Birch Clustering Results')

plt.show()

**Out put:**



**4.Conclusion:**

In this project, we set out to analyze and visualize demographic data related to marginal workers in Tamil Nadu, India. The primary objectives were to understand the distribution of marginal workers based on age, industrial category, and sex, and to derive insights from the data.