Phase 5

PROJECT DOCUMENTATION & SUBMISSION

Date	**31-10-2023**
Team ID	**4498**
Project Name	**Assessment of Marginal Workers in Tamil Nadu- A Socioeconomic Analysis**

Problem Statement:

Objective: The aim of this data analytics project is to comprehensively assess and compare the living conditions and income levels of male and female marginal workers in the state of Tamil Nadu, India.

Problem identified:

The project involves analyzing the demographic characteristics of marginal workers in Tamil Nadu based on their age, industrial category, and sex. The objective is to perform a socioeconomic analysis and create visualizations to represent the distribution of marginal workers across different categories. This project includes defining objectives, designing the analysis approach, selecting appropriate visualization types, and performing the analysis using Python and data visualization libraries.

Introduction:

This analysis delves into the socioeconomic dynamics of marginal workers in Tamil Nadu, India. We will examine key demographics, including age, industrial category, and gender distribution, using insightful visualizations to highlight the challenges faced by this vulnerable group. By shedding light on their profiles, this study aims to inform targeted policies and initiatives that can uplift the marginalized workforce in Tamil Nadu.

Data: We possess a dataset comprising inputs about marginal workers(e.g., age, location, number of workers, education etc.) along with their corresponding incomes. This data will be used to train and evaluate and analyse the data.

Link: https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil

LITERATURE SURVEY

1. "SOCIOECONOMIC PROBLEMS OF MARGINAL SALT PRODUCER- A STUDY WITH REFERENCE TO TAMILNADU" [2022]

The research findings indicate that certain factors, such as gender, age, and marital status, show little correlation with production problems. However, education, family size, and work experience are found to significantly impact these issues. Moreover, the study identifies critical challenges faced by marginal salt producers, including restricted market access, low demand, productivity issues, and compromised product quality. These challenges collectively hinder the sustainability of their operations. The significance of this research lies in its comprehensive examination of the multifaceted challenges faced by marginal salt producers. By considering the influence of socioeconomic factors on production challenges, the paper provides valuable insights for policymakers, researchers, and stakeholders aiming to address the concerns of this segment.

2. "BIG DATA ANALYTICS FRAMEWORK FOR DIGITAL GOVERNMENT" [2020]

The paper introduces a Big Data Analytics Framework for Digital Government, aimed at guiding governments in their digital transformation efforts. The framework includes seven components: Infrastructure, Human Resource Development, Data Governance, Data Catalog, Data Exchange, Laws and Regulations, and Smart and Open Government. These components work together to formalize the transformation process, standardize infrastructure, develop a skilled workforce, manage data effectively, facilitate data exchange, ensure legal compliance, and achieve data-driven decision-making and transparency. The framework is designed to be adaptable to various government contexts and emphasizes the significance of data in addressing contemporary challenges.

3. “RURAL EMPLOYMENT AND DISPARITIES IN TAMILNADU: A QUANTITATIVE ANALYSIS “[2020]

The paper delves into the intricate dynamics of the rural labor market in the context of Tamil Nadu, India, spanning a period of forty years from 1981 to 2011. A prominent observation is the declining contribution of the primary sector, particularly agriculture, to the Gross State Domestic Product (GSDP) in Tamil Nadu. This decline, from a substantial 25% in 1981 to a mere 11% in 2016-17, indicates a notable shift in economic structure.In summary, this paper emerges as a pivotal contribution to the understanding of rural labor markets in Tamil Nadu. The paper's nuanced examination of shifting patterns, coupled with its insightful recommendations, reflects its potential to inform policy decisions aimed at addressing the evolving dynamics of rural employment and disparities within the state.

4. “CLUSTERING-BASED ON PREDICTIVE ANALYSIS TO IMPROVE SCIENTIFIC DATA DISCOVERY” [2020]

The paper discusses a clustering-based predictive analytics approach to enhance the data discovery process within data-intensive projects at the US Department of Energy. Oak Ridge National Laboratory. The primary aim is to improve the accuracy of recommending relevant scientific data to users based on their preferences, search patterns, and past data usage history. The proposed algorithm constructs user profiles using explicit and implicit interactions with the data search system. These interactions are then used to create a training dataset for predicting and recommending data that align with user interests. The paper outlines the use of clustering techniques, content-based filtering, and collaborative filtering to achieve this. The study demonstrates the application of this approach to the Atmospheric Radiation Measurement (ARM) project as an example. The goal is to simplify data recommendations, increase visibility of popular data products, and guide users towards data that meets their needs.

5. "COPING WITH MISSING DATA IN AN UNOBTRUSIVE MONITORING SYSTEM FOR OFFICE WORKERS" [2022]

The presented text discusses the implementation of unobtrusive monitoring systems for tracking daily activities and physiological signs, particularly among older adults. This is driven by societal needs due to the aging population and technological advancements like wearable devices and IoT. The SmartWork project aims to create an AI system that monitors workers' health and well-being, especially in office environments. The system collects health data through wearable devices and work tools, processes the information, and provides adaptive work support. The text delves into technical details about using Laplacian matrix completion for imputing missing data in multidimensional time-series collected by wearable devices. The experimental study compares this approach with other methods for different scenarios, highlighting its effectiveness in completing missing data. The study acknowledges the limitations of the current exploration but emphasizes the potential for using this technique in unobtrusive monitoring systems to support workers' health and productivity.

DESIGN THINKING

Design Thinking Approach

Empathize:

To develop we need to empathize a comprehensive socio-economic analysis of marginal workers in Tamil Nadu, India, with a focus on understanding their demographic characteristics based on age, industrial category, and gender. Create informative visualizations to facilitate data-driven policy decisions and interventions aimed at improving the well-being of this vulnerable segment of the population.

Actions:

- Data from surveys conducted among marginal workers or their households to gather detailed information.

- Information on existing government programs and policies related to marginalized workers, as these may impact the data.

- Segment marginal workers based on age, gender, and industrial category to create profiles.

Define:

Based on our understanding of the problem and the users' needs, we will define clear objectives and success criteria for our project.

Objectives:

- Conduct Exploratory Data Analysis(EDA) to understand the distribution and characteristics of the data, identify outliers, and detect any missing values or inconsistencies.

- Develop predictive models to estimate income levels or employment status based on demographic and educational attributes, providing insights into the factors influencing these outcomes.

Ideate:

Brainstorm potential solutions and approaches to address the problem. This phase involves thinking creatively and considering various algorithms and techniques for socio-economic analysis.

Actions:

- Explore different machine learning algorithms such as decision trees, random forests.

-Experiment with feature engineering techniques to enhance model performance.

-Consider incorporating external data sources (e.g., educational attainment, skill levels, impact income and job opportunities) to improve analysis.

Prototype:

Create a prototype of the machine learning model and the user interface for socio-economic analysis.

Actions:

- Develop a IBM cognos analysis for data pre-processing, model training, and evaluation.

- Create a interface tools like Django to allow users to input details of marginal workers.

- Test the prototype with a subset of the dataset to ensure it meets performance objectives.

Test :

Evaluate the model's performance using appropriate metrics and gather feedback from users.

Actions:

- Split the dataset into training and testing sets.

- Train the model on the training set and evaluate it on the testing set.

- We use metrics such as Age distribution, Gender-based, Industrial category, Educational based to maintain the performance of the model.

- Collect user feedback on the web interface for usability and accuracy.

Implement:

When the prototype achieves its defined objectives and favourable feedback, proceed with the complete implementation.

Actions:

- Perform in-depth analysis and draw conclusions from the dataset.

- Train the final machine learning model using the entire dataset.

- Develop a production-ready web application that incorporates the model.

Iterate:

Continues enhancement is necessary. Collect user feedback and iterate on both the model and interface to improve accuracy and user-friendliness in the context of assessing the socio-economic status of marginal workers in Tamil Nadu.

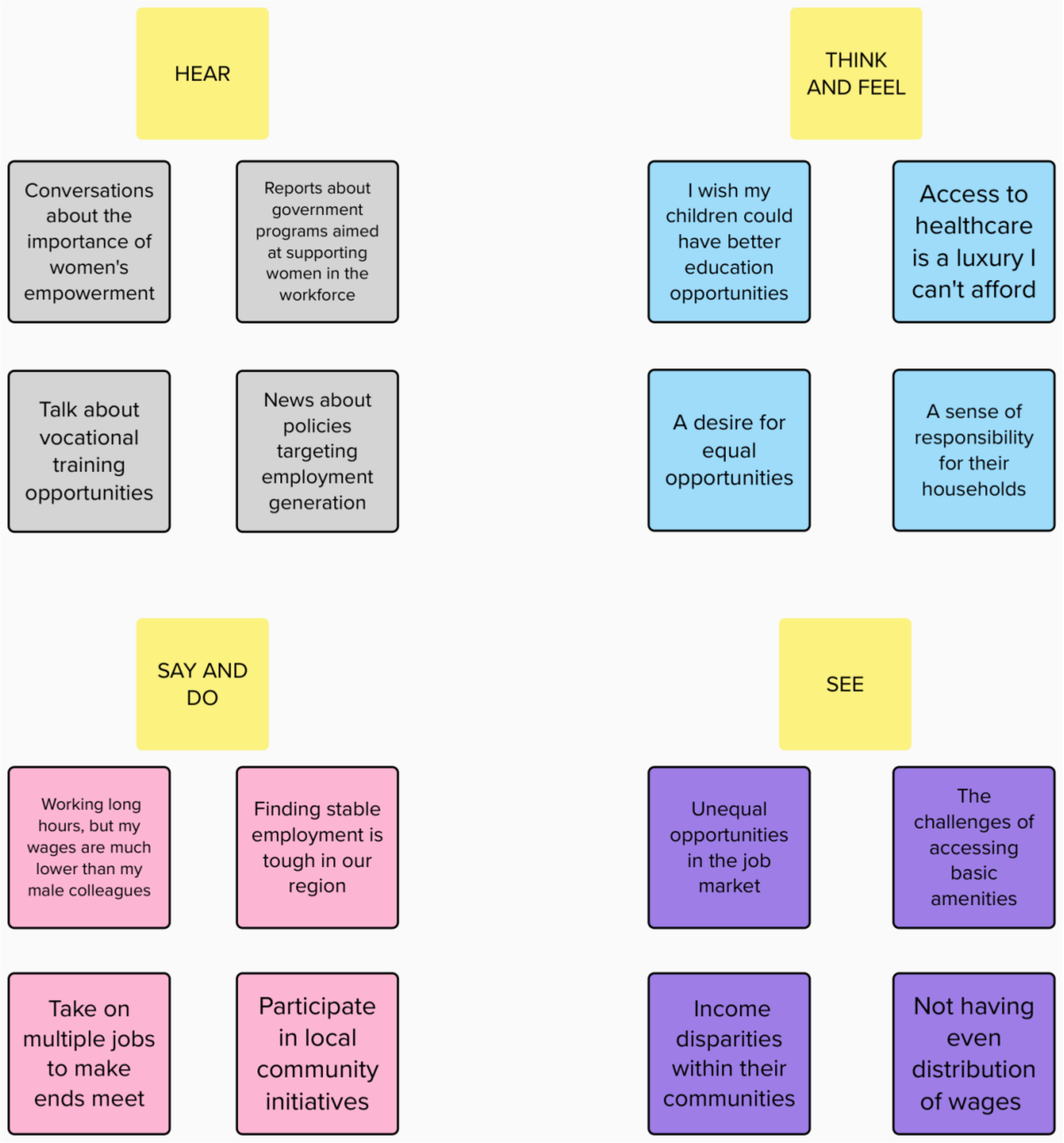
Actions:

- Continuously monitor the model's performance and periodically retrain it using the most recent data.

- Actively incorporate user feedback to enhance the web interface's usability and effectiveness.

- Continuously assess the long-term impact of your interventions on the socio-economic status of marginal workers. This involves tracking changes over time and adjusting strategies accordingly.

TECHNOLOGY ARCHITECTURE



HEAR

Conversations about the importance of women's empowerment

Reports about government programs aimed at supporting women in the workforce

Talk about vocational training opportunities

News about policies targeting employment generation

THINK AND FEEL

I wish my children could have better education opportunities

Access to healthcare is a luxury I can't afford

A desire for equal opportunities

A sense of responsibility for their households

SAY AND DO

Working long hours, but my wages are much lower than my male colleagues

Finding stable employment is tough in our region

Take on multiple jobs to make ends meet

Participate in local community initiatives

SEE

Unequal opportunities in the job market

The challenges of accessing basic amenities

Income disparities within their communities

Not having even distribution of wages

1. Data Collection and Storage:

- **Data Sources**: Utilize external sources, such as Kaggle, for obtaining datasets containing all the marginal workers data based on the working category.

-Data Storage: Data storage is centralized to store and manage the socio-economic data.

2. Data Preprocessing:

- **Data Cleaning**: Clean the dataset by handling missing values, removing duplicates, and addressing data quality issues.

- **Feature Engineering**: Create relevant features from the dataset, such as extracting domain information and content analysis.

- **Data Transformation**: Perform encoding of categorical features, scaling of numerical features, and normalization as necessary.

3. Model Development and Training:

- **IBM Cognos Analysis**: Utilize IBM Cognos Analysis feature for model development and analysis to:

  - Begin integrating relevant data to ensure data quality and consistency.

  - Create data models that define how different data elements are related.

  - Create customized reports and interactive dashboards.

  - Use different visualization option.

4. Model Deployment:

- **IBM Cognos Deployment Space:** Create a deployment space within IBM Cognos for deploying the selected machine learning model.

- **Web Service:** These are used for data integration, data preparation and data transformation that enables us to access, cleanse and structure data from various sources making it ready for analysis with cognos.

5. Security and Access Control:

- **Authorization**: It involves defining roles and responsibilities for users.

- **Access Control:** Configure permissions and access policies to control who can access the dashboards.

6. Real-time Prediction:

-**Data Streams:** It involves processing data streams, where data flow continuously in and needs to be analysed and acted upon immediately.

- **Low Latency:** It needs to have low latency to make decisions quickly and delay should be minimal.

7. Monitoring and Maintenance:

-**Performance Monitoring:** Regularly monitor the performance of the systems and applications.

- **Logging and Alerts:** Set up error tracking and log monitoring to identify and address issues proactively.

- **Quality assurance:** Implement quality assurance processes to enhance the system’s performance and reliability.

8. Scalability and Redundancy:

- **Load Balancing**: Implement load balancing to distribute users requests evenly across multiple Cognos servers.

- **Database Redundancy:** Implement database redundancy by using clustering or replication for the database that stores Cognos content and metadata.

9. Data Feedback:

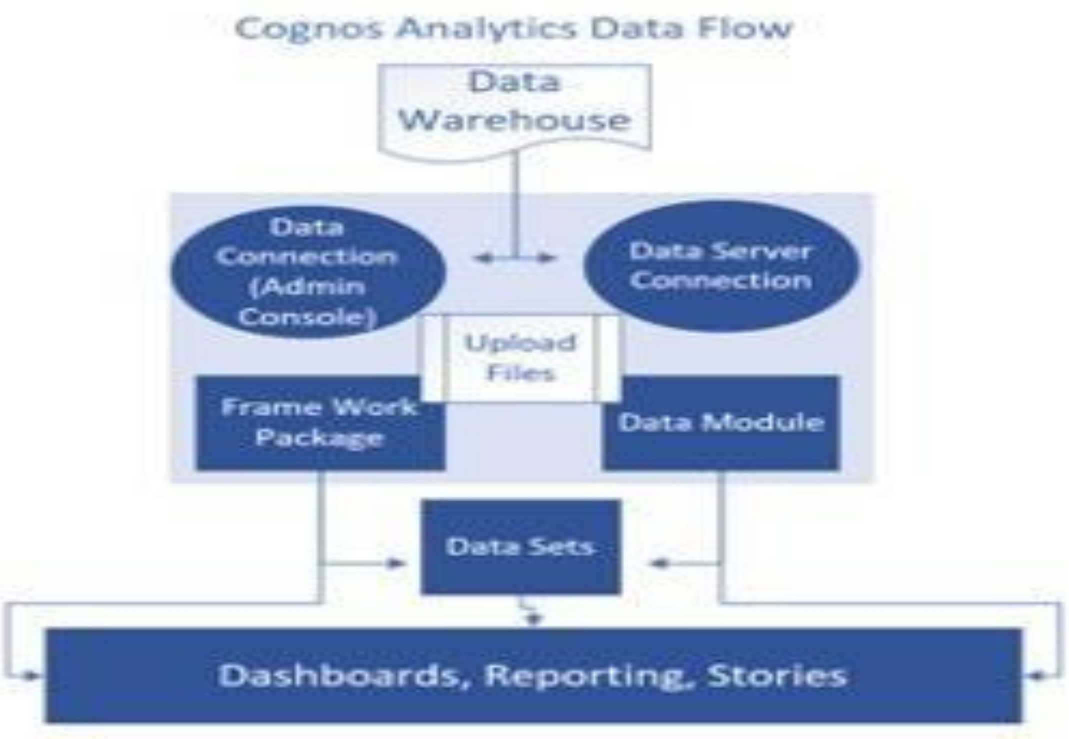
- **User Surveys:** Create surveys or questionnaires for users or the target audience of your data analysis model. Ask questions about the usefulness, clarity, and effectiveness of the model's outputs can be asked.

- **Documentation and Training:** Provide clear documentation and training resources for users to understand how to effectively use the data analysis model. Make these resources accessible and user-friendly.

10. Reporting and Visualization:

- Generate reports and visualizations to provide insights into the model's performance and effectiveness.

THE DIAGRAM REPRESENTS THE TECHNOLOGY ARCHITECTURE :


Cognos Analytics Data Flow
Data Warehouse
Data Connection (Admin Console)
Data Server Connection
Upload Files
Frame Work Package
Data Module
Data Sets
Dashboards, Reporting, Stories

This technology architecture outlines a comprehensive solution for deploying. It focuses on scalability, security, real-time prediction, and ongoing maintenance to ensure the system's effectiveness.

MODULES DESCRIPTION

1. Data Collection and Storage Module:

- **Objective:** This module focuses on gathering and storing datasets for further process of analysis.

- **Key Tasks:**

  - Designing data collection methods and surveys.

- Establishing efficient data storage and organization.

- Maintaining and updating the data repository for ongoing analysis.

2. Data Preprocessing Module:

- Objective: Prepare the dataset for model training by cleaning, transforming, and engineering features.

- Key Tasks:

- Identifying and rectifying missing or inconsistent data values.

- Converting data into suitable formats and scales for analysis.

- Data transformation, including encoding categorical features and scaling numerical features.

3. Model Development and Training Module:

- Objective: To create accurate analytical models to gain deeps insights into the socio-economic conditions of marginal workers in Tamil Nadu.

- Key Tasks:

- Data preprocessing and cleaning.

- Feature selection and engineering.

- Model evaluation and validation.

4. Model Deployment Module:

- Objective: Utilize IBM Cognos Analysis feature for model development and analysis.

- Key Tasks:

- Begin integrating relevant data to ensure data quality and consistency.

- Create data models that define how different data elements are related.

- Create customized reports and interactive dashboards.

- Use different visualization option.

5. Security and Access Control Module:

- Objective: To ensure the confidentiality, integrity and controlled access to sensitive socioeconomic data in the analysis of marginal workers in Tamil Nadu.

- Key Tasks:

- Implementing strong user authentication.
- Ensuring data encryption and secure transmission.

6. Real-time Prediction Module:

- Objective: To provide accurate prediction on the workers in each industrial category.

- Key Tasks:

- Real-time data integration and processing.
- Monitoring and updating models as new data becomes available.

ALGORITHM AND TECHNOLOGY USED

1. Data Collection and Preprocessing:

- Technology: Python (for data handling)
- Description: Collect the dataset from Kaggle, Preprocess the data by handling missing values, encoding categorical features, scaling numerical features, and any other necessary data transformations.

2. Model Development:

- Technology: IBM COGNOS ANALYTICS.
- Algorithm: Linear Regression
- Description: IBM Cognos Analysis feature to automate the model development process. Cognos performs the following tasks:

- Upload dataset.
- Create dashboard.
- Perform visualization
- Generates Professional report.

3. Model Evaluation and Selection:

- Technology: IBM Cognos Analytics.

- Algorithm: Statistical Analysis

- Description: Data analysis are wide range of algorithms some are Descriptive statistics such as (Mean, Median, Mode, Variance, Standard Deviation) and Regression Analysis (Linear regression, Logistic regression). Evaluate these models using performance metrics like accuracy and precision to select the best-performing model.

4. Model Deployment:

- Technology: IBM Cognos Analytics

- Algorithm: Deployment is content of data analytics may differ in each algorithm.

- Description: Data Warehouse Deployment store and manage large volumes of data for analytics. Data analytics deployment in this context involves designing data warehouses, creating schemas, and optimizing data storage for efficient querying and reporting.

5. Security and Access Control:

- Technology: IBM Cognos Analytics

- Description: Cognos supports data-level security, allowing administrators to restrict access to specific data within reports or data sources. This is crucial when certain users should only see a subset of the data.

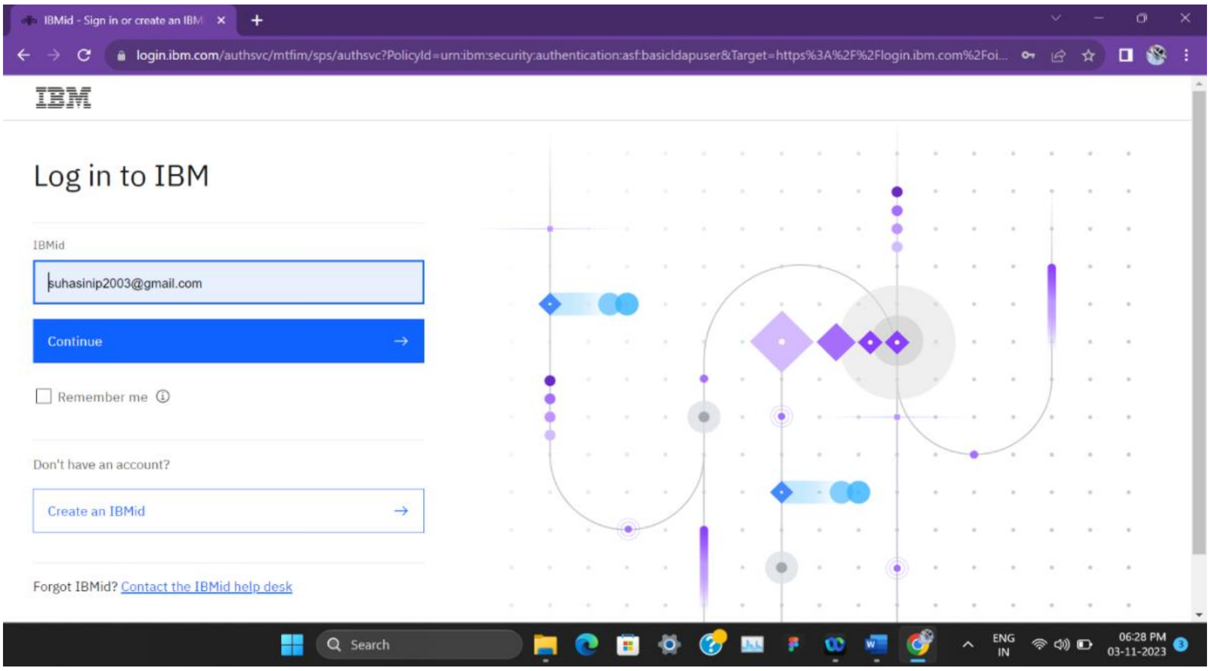
6. Real-time Prediction:

- Technology: Real-time prediction in the context of data analytics

- Description: Ability to generate predictions or make decisions instantly as new data becomes available. This process involves applying a trained model to incoming data to provide immediate insights or actions.

PROJECT DEVELOPMENT STEPS AND SCREENSHOT

Step 1: Account creation and create a new dashboard in IBM Cognos Analytics



IBMid - Sign in or create an IBM

login.ibm.com/authsvc/mtfim/sps/authsvc?PolicyId=urn:ibm:security:authentication:asf:basicldapuser&Target=https%3A%2F%2Flogin.ibm.com%2Foi...

Log in to IBM

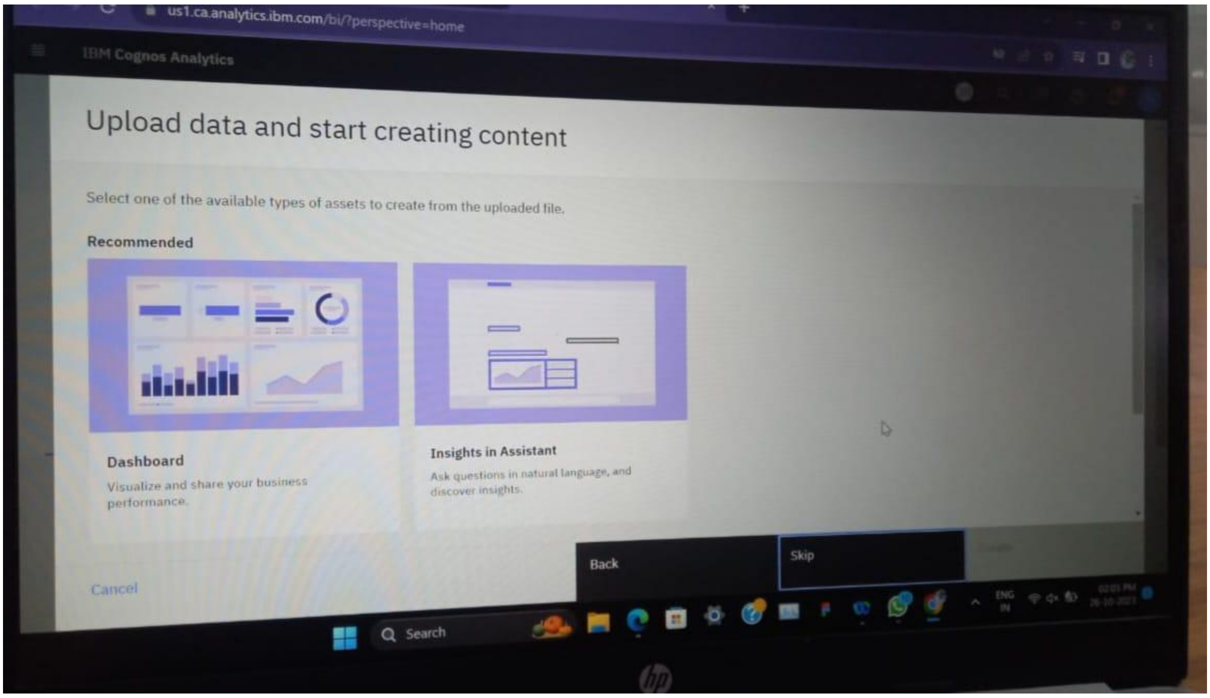
IBMid: suhasinip2003@gmail.com

□ Remember me

Don't have an account?

Forgot IBMid? Contact the IBMid help desk

Step 2 : Upload the dataset to the dashboard



Upload data and start creating content

Select one of the available types of assets to create from the uploaded file.

Recommended

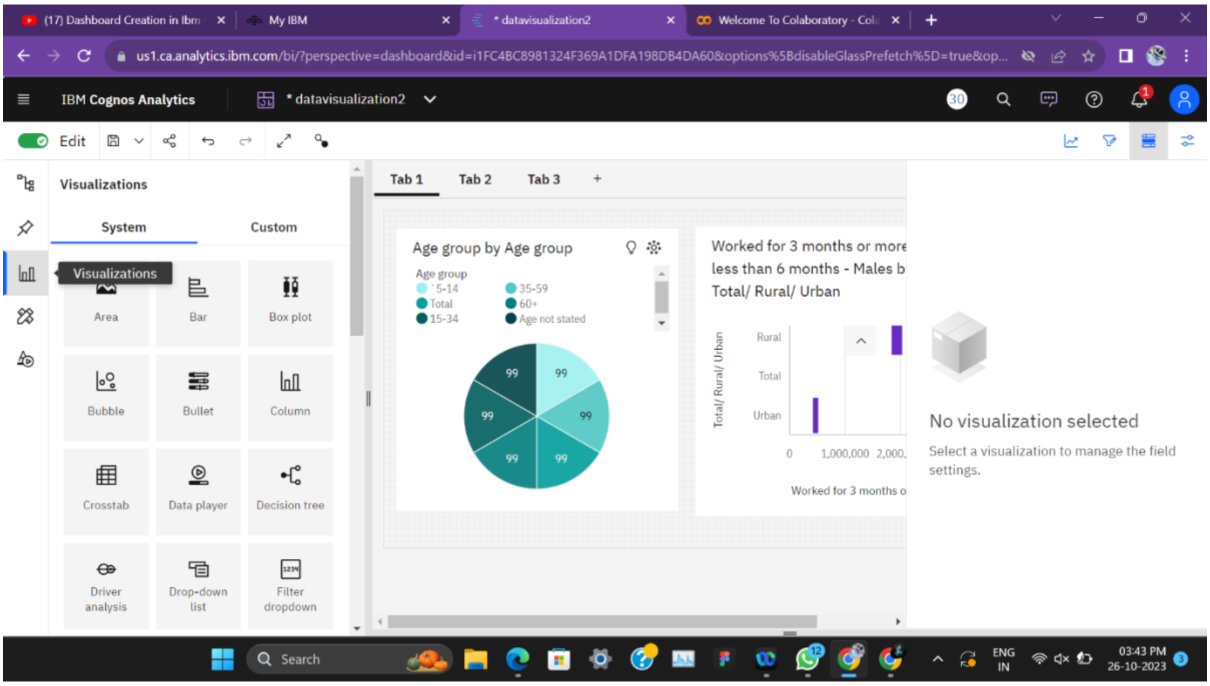
Dashboard

Visualize and share your business performance.

Insights in Assistant

Ask questions in natural language, and discover insights.

Step 3 : Choose the Visualization needed and perform the analysis



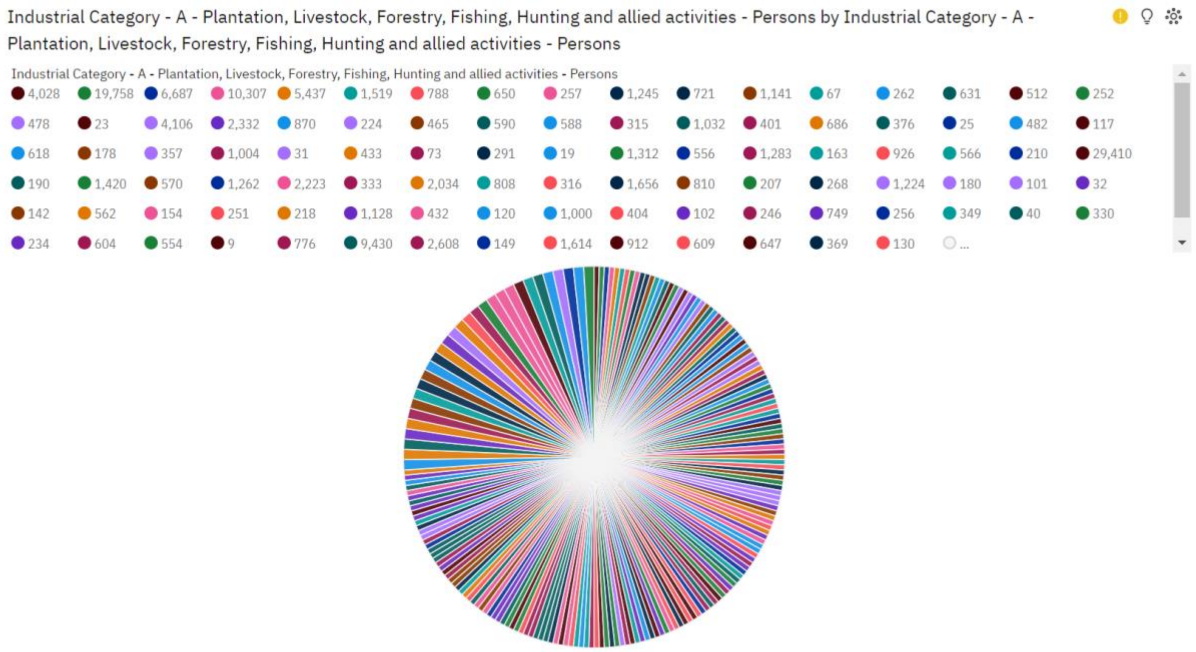
Age group by Age group

Worked for 3 months or more less than 6 months - Males b Total/ Rural/ Urban

No visualization selected

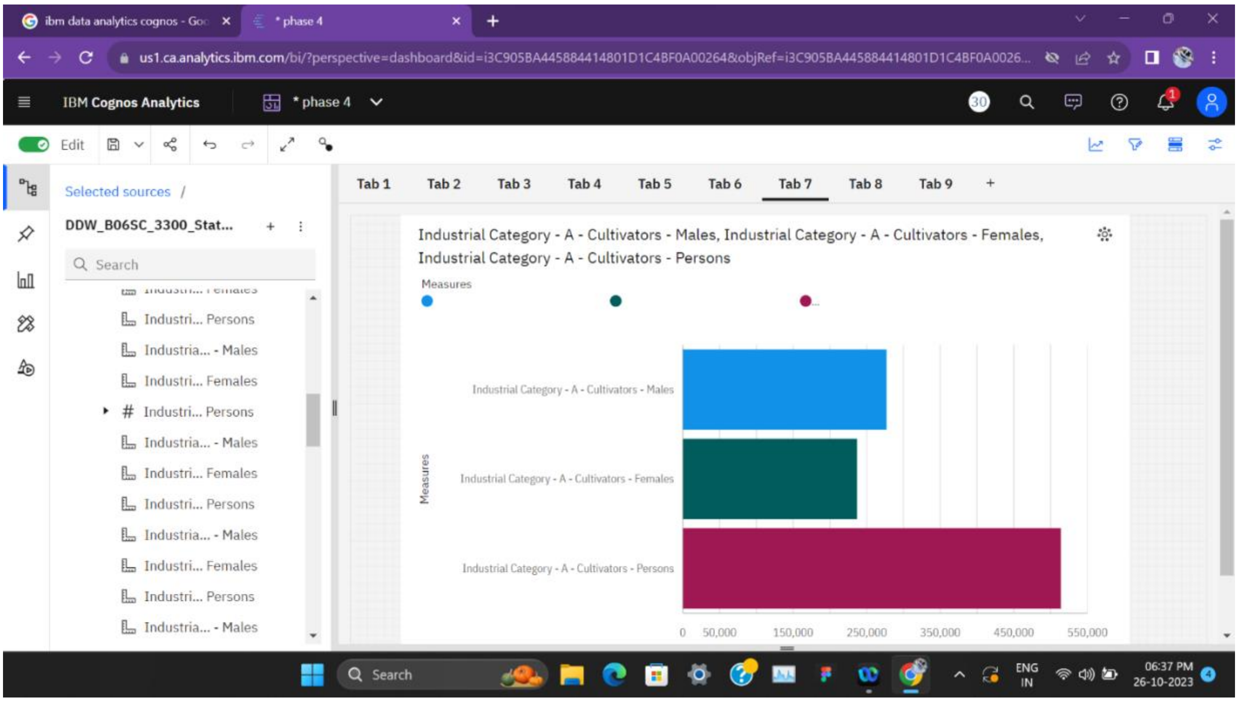
Select a visualization to manage the field settings.

Step 4: Data visualization is done the Industrial Category A .



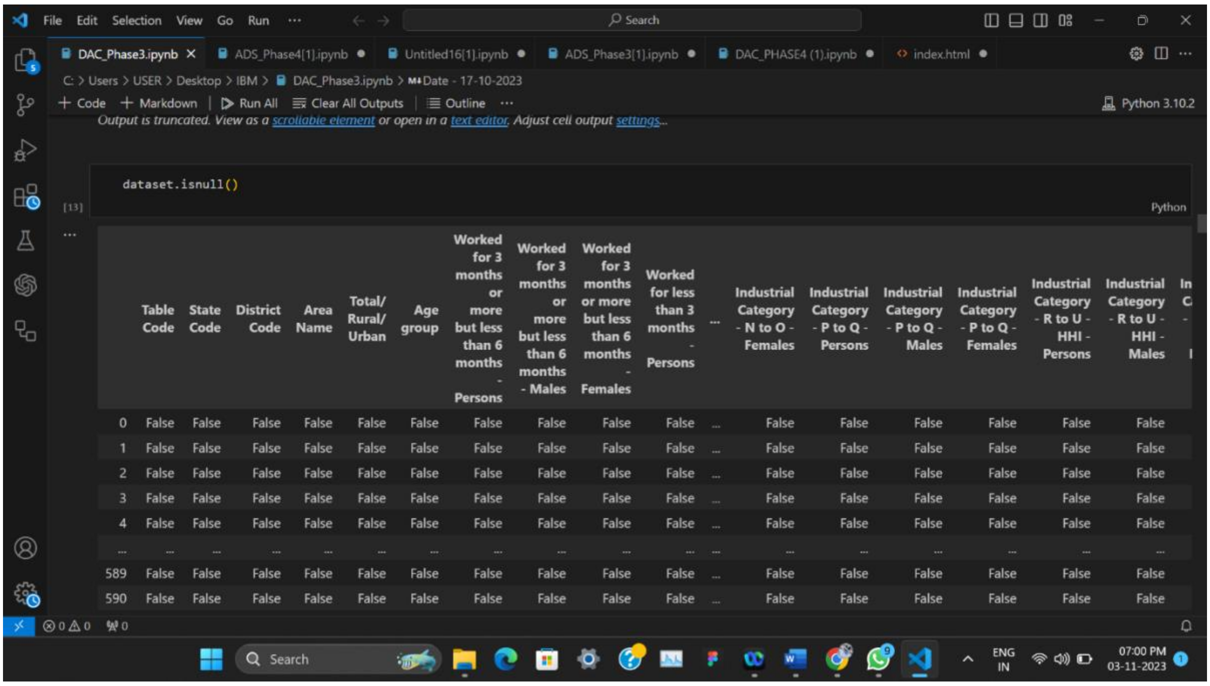
Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Persons by Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Persons

Step 5: Data Visualization on Industrial Category A between Male and Female.



Industrial Category - A - Cultivators - Males, Industrial Category - A - Cultivators - Females, Industrial Category - A - Cultivators - Persons

Step 6: Perform Data analysis using Jupyter Notebook.


File Edit Selection View Go Run ···
Search
DAC_Phase3.ipynb
ADS_Phase4[1].ipynb
Untitled16[1].ipynb
ADS_Phase3[1].ipynb
DAC_PHASE4 (1).ipynb
index.html
C: > Users > USER > Desktop > IBM > DAC_Phase3.ipynb > M↓Date - 17-10-2023
+ Code + Markdown | ▷ Run All ☰ Clear All Outputs | ☰ Outline ···
Python 3.10.2
Python
07:00 PM
03-11-2023

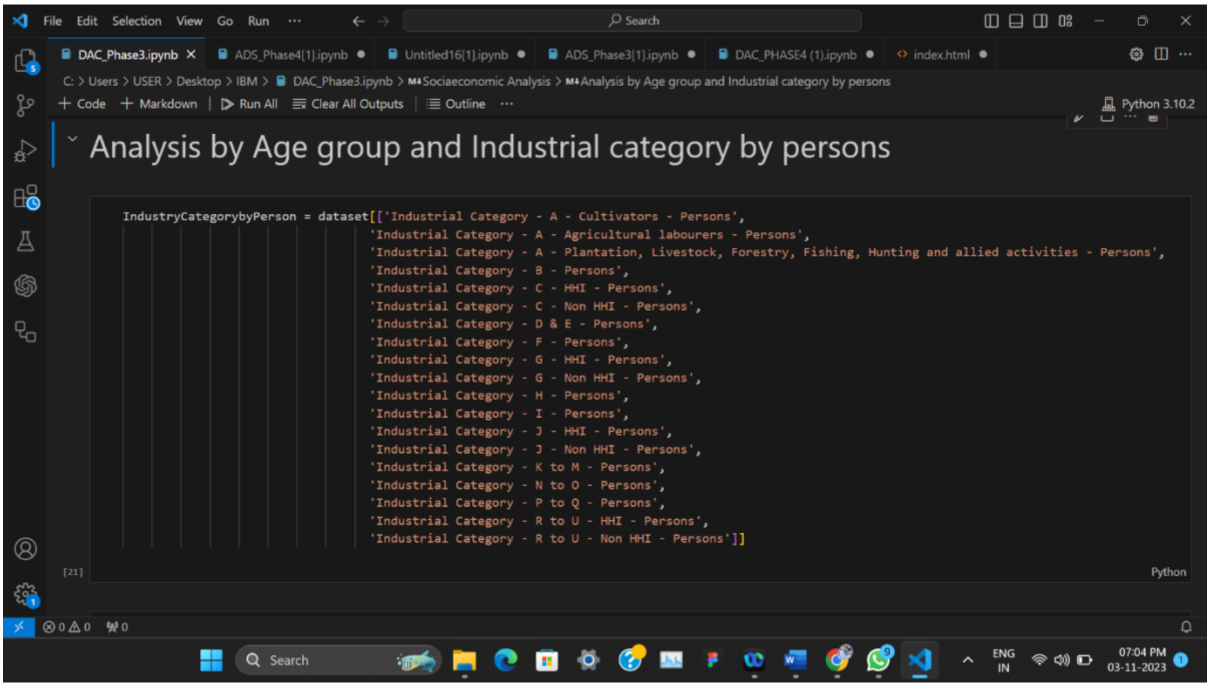
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
dataset.isnull()
```

[13]

	Table Code	State Code	District Code	Area Name	Total/ Rural/ Urban	Age group	Worked for 3 months or more but less than 6 months - Persons	Worked for 3 months or more but less than 6 months - Males	Worked for 3 months or more but less than 6 months - Females	Worked for less than 3 months - Persons	...	Industrial Category - N to O - Females	Industrial Category - P to Q - Persons	Industrial Category - P to Q - Males	Industrial Category - P to Q - Females	Industrial Category - R to U - HHI - Persons	Industrial Category - R to U - HHI - Males
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
589	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
590	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False

Step 7: Analysis by Age group and Industrial category persons.

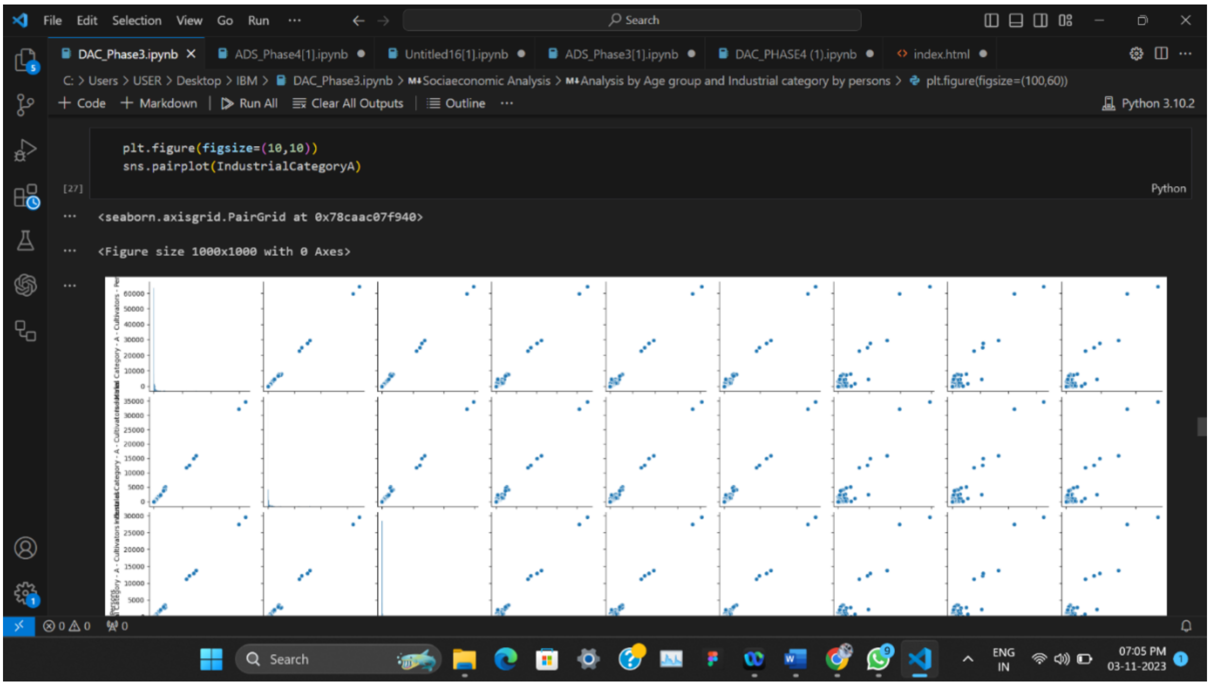

File Edit Selection View Go Run ···
Search
DAC_Phase3.ipynb
ADS_Phase4[1].ipynb
Untitled16[1].ipynb
ADS_Phase3[1].ipynb
DAC_PHASE4 (1).ipynb
index.html
C: > Users > USER > Desktop > IBM > DAC_Phase3.ipynb > M↓Sociaeconomic Analysis > M↓Analysis by Age group and Industrial category by persons
+ Code + Markdown | ▷ Run All ☰ Clear All Outputs | ☰ Outline ···
Python 3.10.2
Python
07:04 PM
03-11-2023

Analysis by Age group and Industrial category by persons

```
IndustryCategorybyPerson = dataset[['Industrial Category - A - Cultivators - Persons',
                                    'Industrial Category - A - Agricultural labourers - Persons',
                                    'Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Persons',
                                    'Industrial Category - B - Persons',
                                    'Industrial Category - C - HHI - Persons',
                                    'Industrial Category - C - Non HHI - Persons',
                                    'Industrial Category - D & E - Persons',
                                    'Industrial Category - F - Persons',
                                    'Industrial Category - G - HHI - Persons',
                                    'Industrial Category - G - Non HHI - Persons',
                                    'Industrial Category - H - Persons',
                                    'Industrial Category - I - Persons',
                                    'Industrial Category - J - HHI - Persons',
                                    'Industrial Category - J - Non HHI - Persons',
                                    'Industrial Category - K to M - Persons',
                                    'Industrial Category - N to O - Persons',
                                    'Industrial Category - P to Q - Persons',
                                    'Industrial Category - R to U - HHI - Persons',
                                    'Industrial Category - R to U - Non HHI - Persons']]
```

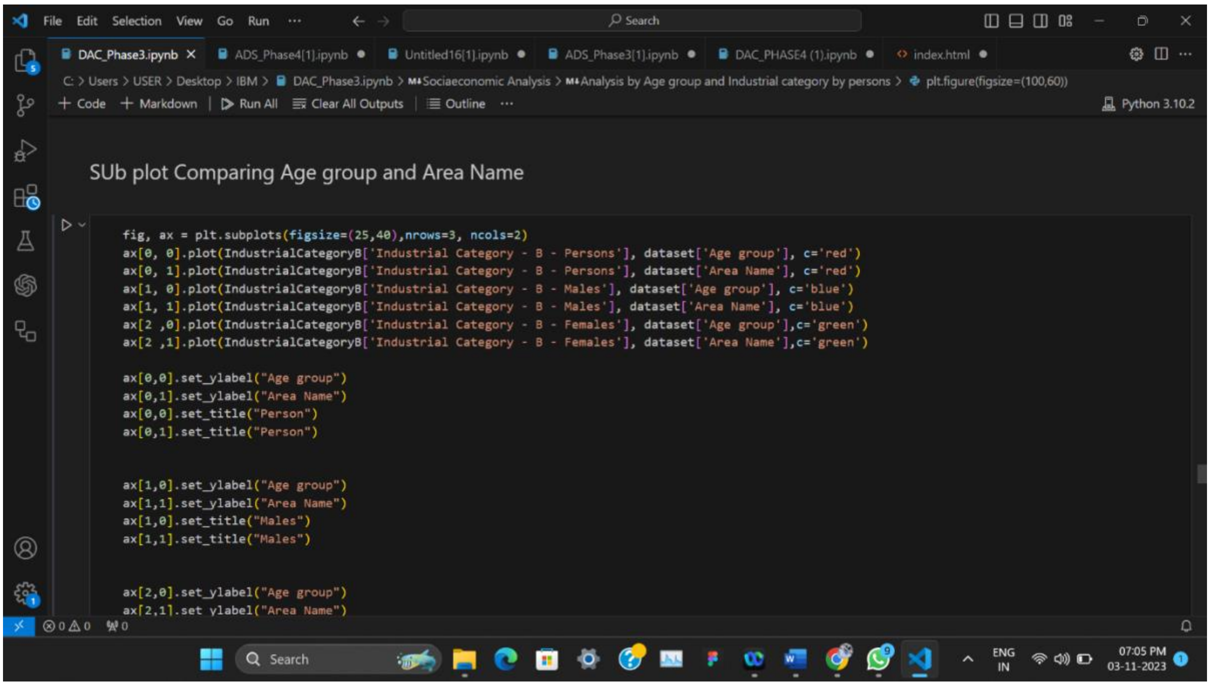
[21]

Step 8: Ploting the industrial category A.


File Edit Selection View Go Run
Search
DAC_Phase3.ipynb
ADS_Phase4[1].ipynb
Untitled16[1].ipynb
ADS_Phase3[1].ipynb
DAC_PHASE4 (1).ipynb
index.html
C: > Users > USER > Desktop > IBM > DAC_Phase3.ipynb > Sociaeconomic Analysis > Analysis by Age group and Industrial category by persons > plt.figure(figsize=(100,60))
+ Code + Markdown Run All Clear All Outputs Outline
Python 3.10.2
[27]
Python
Search
ENG
IN
07:05 PM
03-11-2023

```
plt.figure(figsize=(10,10))
sns.pairplot(IndustrialCategoryA)
```

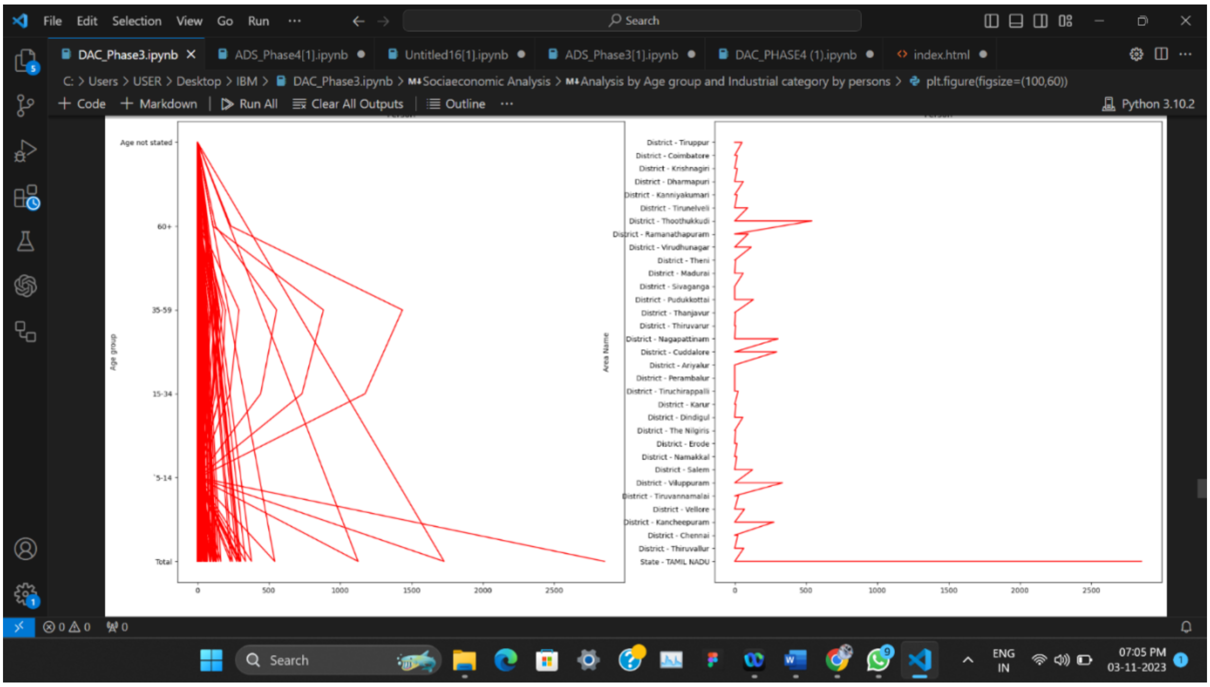
```
<seaborn.axisgrid.PairGrid at 0x78caac07f940>

<Figure size 1000x1000 with 0 Axes>
```

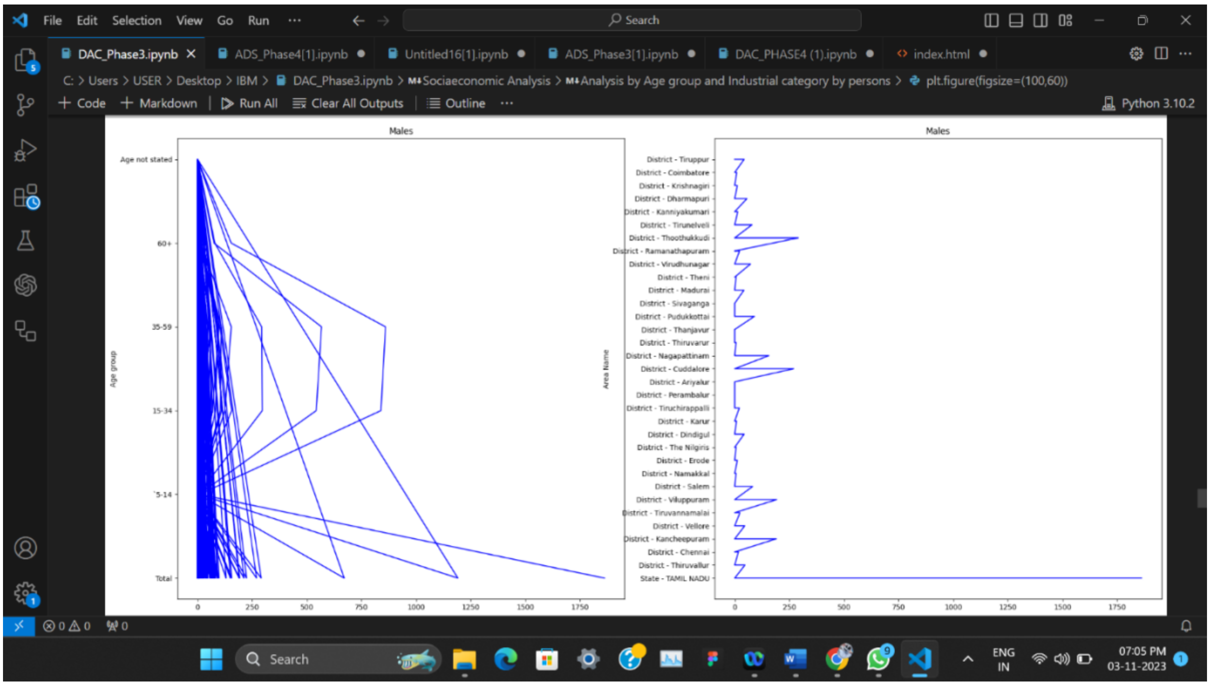


SUb plot Comparing Age group and Area Name

```
fig, ax = plt.subplots(figsize=(25,40),nrows=3, ncols=2)
ax[0, 0].plot(IndustrialCategoryB['Industrial Category - B - Persons'], dataset['Age group'], c='red')
ax[0, 1].plot(IndustrialCategoryB['Industrial Category - B - Persons'], dataset['Area Name'], c='red')
ax[1, 0].plot(IndustrialCategoryB['Industrial Category - B - Males'], dataset['Age group'], c='blue')
ax[1, 1].plot(IndustrialCategoryB['Industrial Category - B - Males'], dataset['Area Name'], c='blue')
ax[2 ,0].plot(IndustrialCategoryB['Industrial Category - B - Females'], dataset['Age group'],c='green')
ax[2 ,1].plot(IndustrialCategoryB['Industrial Category - B - Females'], dataset['Area Name'],c='green')

ax[0,0].set_ylabel("Age group")
ax[0,1].set_ylabel("Area Name")
ax[0,0].set_title("Person")
ax[0,1].set_title("Person")


ax[1,0].set_ylabel("Age group")
ax[1,1].set_ylabel("Area Name")
ax[1,0].set_title("Males")
ax[1,1].set_title("Males")


ax[2,0].set_ylabel("Age group")
ax[2,1].set_ylabel("Area Name")
```

Step 9: Analysis on Male and Female based on all Industrial Category.





CONCLUSION

The assessment of marginal workers in Tamil Nadu is a crucial undertaking to address their socio-economic challenges. Through data analysis and a user-friendly web interface, this project aims to provide valuable insights and support evidence-based decision-making. Continuous improvement, user feedback, and model refinement are essential for the success and relevance of this initiative. Ultimately, this endeavor strives to improve the well-being and inclusion of marginal workers, fostering positive socio-economic changes in Tamil Nadu.