TAMILNADU MARGINAL WORKERS ASSESSMENT

Data Analytics with cognos – Phase 1

DOCUMENTATION

Team Members:

1.KARTHIKEYAN N (211521243081)

2.HARI KRISHNAMOORTHY M (211521243061)

3.ANJEEVARAGAVAN R (211521243014)

4.BALAJI V (211521243029)

5.KARTHIKEYAN P (211521243082)

Phase 1: Problem Definition and Design Thinking

PROJECT DEFINITON:

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.

Dataset Link:

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

IN DESIGN THINKING, WHAT WE ARE GOING TO SEE IN THIS DOCUMENTATION ARE:

1.PROJECT OBJECTIVES

2.ANALYSIS APPROACH

3.VISUALIZATION SELECTION

Project objectives:

Define objectives such as analyzing marginal worker demographics, understanding age and gender distribution, and exploring industrial categories.

Analysis Approach:

Plan the steps to extract, clean, and analyze the dataset to derive insights.

Visualization Selection:

Determine suitable visualization types (e.g., bar charts, pie charts, heatmaps) to represent demographic distributions effectively.

Design Thinking:

Project Objectives:

Firstly, we are going to explore the analyzing Marginal Workers Demographics and detail observation of Dataset.

Marginal Workers Demographics:

According to the dataset, It consists of code for state, district and area name that shows the details as below,

Table Code:

B0806SC

State Code:

`33 – Tamil Nadu

District Code:

000 – Tamil Nadu

602 – Thiruvallur

603 – Chennai

604 – Kancheepuram

605 – Vellore

606 – Tiruvannamalai

607 – Villupuram

608 – Salem

609 – Namakkal

610 – Erode

611 – The Nilgiris

612 – Dindigul

613 – Karur

614 – Tiruchirappalli

615 – Perambalur

616 – Ariyalur

617 – Cuddalore

618 – Nagapattinam

619 – Tiruvarur

620 – Thanjavur

621 – Pudukkottai

622 – Sivaganga

623 – Madurai

624 – Theni

625 – Virudhunagar

626 – Ramanathapuram

627 – Thoothukkudi

628 – Tirunelveli

629 – Kanniyakumari

630 – Dharmapuri

631 – Krishnagiri

632 – Coimbatore

633 – Tiruppur

The above mentioned are the code with their respective locations.

Total/Rural/Urban:

This column in the data set shows the type of location that inside the respective districts, here Total represents the sum of the Rural and Urban areas given in the State or Districts.

TOTAL = RURAL + URBAN

Total – represents the whole state.

Rural – represents the villages in state or districts.

Urban – represents the town or city in state or districts.

Age Groups:

In this Column, the ages are divided into certain groups as below,

5 – 14

15 – 34

35 – 59

60+

Age not stated

Genders:

In this data set, there are two types of genders are categorized, they are

Male

Female

Persons = Male + Female

Industrial Categories:

In Data set, Columns are categorized in the event of Industries that shows the type of industries and different phenomenon of works, the below following keys shows the type of industries that are categorized respective to the persons,

A: Agriculture, Forestry and Fishing;

B: Mining and Quarrying;

C: Manufacturing;

D: Electricity, Gas, steam and Air conditioning Supply;

E: Water Supply; (Sewerage, Waste Management and remediation activities);

F: Construction;

G: Wholesale and Retail Trade (Repair of motor vehicles and motor cycles);

H: Transportation and Storage;

I: Accommodation and food service activities;

J: Information and Communication;

K: Financial and Insurance activities;

L: Real Estate activities;

M: Professional, Scientific and Technical activities;

N: Administrative and support service activities;

O: Public Administration and Defence, Compulsory Social Security;

P: Education;

Q: Human Health and Social Work activities;

R: Arts, Entertainment and recreation;

S: Other Service Activities;

T: Activities of Households as Employers: Undifferentiated Goods and Services;

U: Activities of Extra-Territorial Organisations and Bodies.

Work Duration:

In Data set, some of the columns are categorized based on the duration of workers respects with the genders and ages, the duration types that shown in the dataset are:

>3 Months or More but less than 6 Months

>Worked for less than 3 Months

Those columns are describe the time period of person worked.

Gender Distribution:

In this data set, columns are categorized in industrial or works respects with the genders and ages, the below shown are gender distribution presents in the dataset,

Industrial Category - A - Cultivators – Persons;

Industrial Category - A - Cultivators – Males;

Industrial Category - A - Cultivators – Females;

Industrial Category - A - Agricultural labourers – Persons;

Industrial Category - A - Agricultural labourers – Males;

Industrial Category - A - Agricultural labourers – Females;

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

Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities – Males;

Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities – Females;

Industrial Category - B – Persons;

Industrial Category - B – Males;

Industrial Category - B – Females;

Industrial Category - C - HHI – Persons;

Industrial Category - C - HHI – Males;

Industrial Category - C - HHI – Females;

Industrial Category - C - Non HHI – Persons;

Industrial Category - C - Non HHI – Males;

Industrial Category - C - Non HHI – Females;

Industrial Category - D & E – Persons;

Industrial Category - D & E – Males;

Industrial Category - D & E – Females;

Industrial Category - F – Persons;

Industrial Category - F – Males;

Industrial Category - F – Females;

Industrial Category - G - HHI – Persons;

Industrial Category - G - HHI – Males;

Industrial Category - G - HHI – Females;

Industrial Category - G - Non HHI – Persons;

Industrial Category - G - Non HHI – Males;

Industrial Category - G - Non HHI – Females;

Industrial Category - H – Persons;

Industrial Category - H – Males;

Industrial Category - H – Females;

Industrial Category - I – Persons;

Industrial Category - I – Males;

Industrial Category - I – Females;

Industrial Category - J - HHI – Persons;

Industrial Category - J - HHI – Males;

Industrial Category - J - HHI – Females;

Industrial Category - J - Non HHI – Persons;

Industrial Category - J - Non HHI – Males;

Industrial Category - J - Non HHI – Females;

Industrial Category - K to M – Persons;

Industrial Category - K to M – Males;

Industrial Category - K to M – Females;

Industrial Category - N to O – Persons;

Industrial Category - N to O – Males;

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;

Industrial Category - R to U - HHI – Females;

Industrial Category - R to U - Non HHI – Persons;

Industrial Category - R to U - Non HHI – Males;

Industrial Category - R to U - Non HHI – Females;

ANALYSIS APPROACH:

The steps for conducting a demographic analysis of marginal workers in Tamil Nadu using Python and data visualization libraries:

1. **Define Objectives**:

Objective 1: Demographic Analysis

- Define the criteria for identifying marginal workers.

- Analyze the demographic characteristics, including age, industrial category, and sex.

Objective 2: Socioeconomic Analysis

- Explore socioeconomic factors affecting marginal workers.

- Include variables related to education, income, and access to social services.

2. **Data Extraction and Cleaning**:

- Obtain datasets containing information on marginal workers in Tamil Nadu.

- Clean the data, handling missing values, outliers, and ensuring data consistency.

- **Handle Missing Values:**
    - Identify and address missing values using appropriate methods like imputation or removal.
- **Data Transformation:**
    - Standardize formats, units, or any inconsistencies in the dataset.
- **Remove Duplicates:**
    - Check for and remove any duplicate entries in the dataset.
- **Handle Outliers:**
    - Identify and address outliers that could skew the analysis.

3. **Data Exploration**:

- Use Pandas for initial exploration and summary statistics.

- Visualize the distribution of demographic variables using histograms and other relevant plots.

4. **Demographic Analysis**:

- Segment the data based on age groups, industrial categories, and sex.

- Calculate proportions and percentages to understand the composition of each segment.

5. **Socioeconomic Analysis**:

- Include variables related to education levels, income, and access to social services.

- Use descriptive statistics and visualizations to analyze socioeconomic factors.

6. **Data Visualization**:

Matplotlib and Seaborn:

- Use Matplotlib and Seaborn for creating visualizations.

- Plot bar charts for demographic distributions and box plots for socioeconomic factors.

Pandas Plotting:

- Leverage Pandas plotting capabilities for quick visualizations.

7. **Interactive Visualizations**:

- Consider using libraries like Plotly for interactive visualizations.

- Interactive charts can enhance the exploration of data and insights.

8. **Geospatial Analysis** (Optional):

- If applicable, consider geospatial visualizations using libraries like GeoPandas for mapping the distribution of marginal workers across regions in Tamil Nadu.

9. **Statistical Analysis**:

- Use statistical tests to identify significant differences in demographic and socioeconomic variables.

- Perform correlation analysis to explore relationships.

10. **Documentation and Reporting**:

- Document the data analysis process, including cleaning steps, analysis techniques, and visualization choices.

- Prepare a detailed report with key findings, insights, and actionable recommendations.

11. **Code Organization**:

- Organize Python code into functions or classes for modularity and reproducibility.

- Add comments and documentation within the code for clarity.

12. **Ethical Considerations**:

- Ensure compliance with ethical standards and data privacy regulations.

- Anonymize data if necessary to protect individual privacy.

13. **Stakeholder Engagement**:

- Engage with stakeholders to validate findings and gather additional insights.

- Incorporate feedback into the analysis.

14. **Iterative Analysis**:

- Be open to iterations based on initial findings and stakeholder feedback.

15. **Knowledge Sharing**:

- Share knowledge and insights gained during the analysis with relevant stakeholders and the broader community if applicable.

By following these steps, you can systematically conduct a demographic and socioeconomic analysis of marginal workers in Tamil Nadu, creating insightful visualizations using Python and data visualization libraries.

VISUALIZATION SELECTION:

The specific demographic data available (districts in the state, genders, ages, work duration, and industrial categories), here are visualization selections:

1. **District-wise Distribution**:

- Visualization Type: Bar Chart or Choropleth Map

- Description: Illustrate the distribution of marginal workers across different districts in Tamil Nadu.

2. **Age Distribution Across Districts**:

- Visualization Type: Stacked Bar Chart or Grouped Bar Chart

- Description: Display the age distribution of marginal workers in each district.

3. **Gender Distribution Across Districts**:

- Visualization Type: Stacked Bar Chart or Grouped Bar Chart

- Description: Show the gender distribution of marginal workers in each district.

4. **Work Duration Analysis**:

- Visualization Type: Box Plot or Violin Plot

- Description: Explore the distribution of work duration among marginal workers in different districts.

5. **Industrial Category Composition Across Districts**:

- Visualization Type: Stacked Bar Chart or Grouped Bar Chart

- Description: Illustrate the distribution of marginal workers across different industrial categories within each district.

6. **Age Distribution by Gender**:

- Visualization Type: Grouped Bar Chart

- Description: Compare the age distribution of male and female marginal workers.

7. **Work Duration by Industrial Category**:

- Visualization Type: Grouped Bar Chart or Box Plot

- Description: Compare the work duration across different industrial categories.

8. **District-wise Age and Gender Composition**:

- Visualization Type: Multi-level Bar Chart or Stacked Bar Chart

- Description: Visualize the age and gender composition of marginal workers in each district.

9. **Comparative Analysis Across Districts**:

- Visualization Type: Radar Chart or Spider Chart

- Description: Compare demographic characteristics (age, gender, work duration) across multiple districts.

10. **Interactive Dashboard** (Optional):

- Visualization Type: Interactive dashboard using tools like Dash or Streamlit.

- Description: Combine multiple visualizations for an interactive exploration of demographic data across districts.

11. **Industrial Category Distribution by Age Group**:

- Visualization Type: Stacked Bar Chart or Heatmap

- Description: Explore how different age groups are distributed across various industrial categories.

12. **Age Distribution Overlaid with Industrial Categories**:

- Visualization Type: Stacked Area Chart or Layered Bar Chart

- Description: Illustrate the combined distribution of age groups and industrial categories.