Data Analytics With Cognos

Phase-5 Project Documentation and Submission

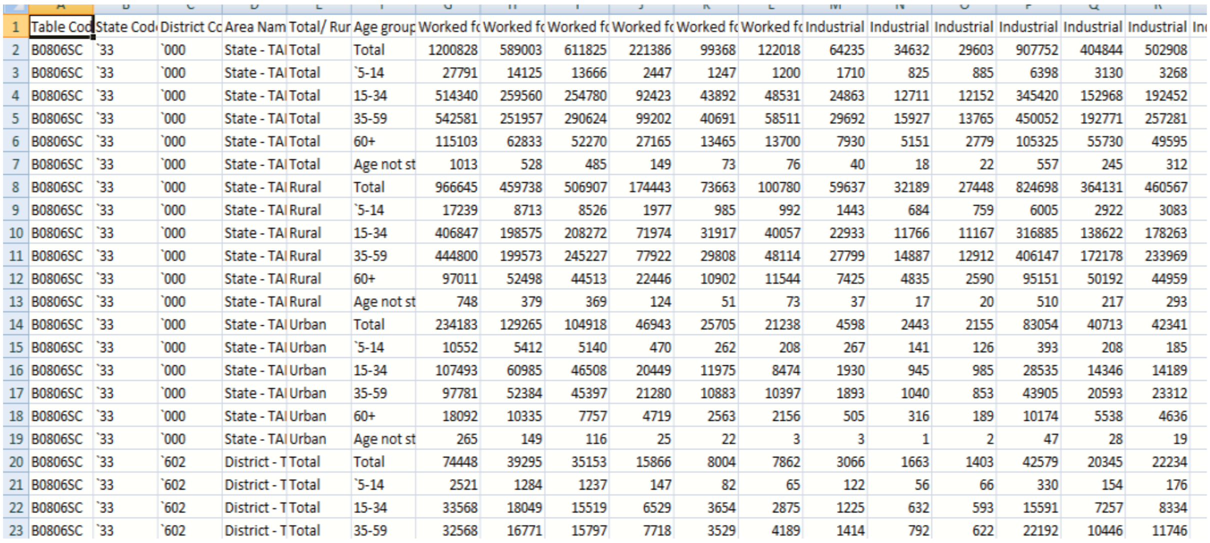
TN Marginal Workers Assessment

Project's Objectives*:*

- *In this project's we determine the analysis* approach, *visualization types and code implementation using python,these analysis provides the insights of data analysis and demographic characteristics of marginal workers in TamilNadu.*
- ***Analysis approach:*** *The analysis approach involves the data collection,data cleaning,exploratory of data analysis,data modeling and this approach has the techniques such as statistical analysis,data visualizations to insights the data extraction for decision making and predictions of given dataset.*
- ***Visualization types****:In data analytics,various visualizations are used to present and interpret data effectively.The types are Line charts,Bar chats,Scatter plots,Histograms,and Pie charts.*
- ***Code implementation****:In this we use Python provide libraries such as pandas,matplotlib and Numpy for implementing the given dataset based on the requirements.*
- ***Data analysis****:In the data analysis process the following steps are involved as follows as data collection,data cleaning,exploratory data analysis(EDA) and so on.*
- ***Demographic analysis****: It involves the specific characteristics of a population such as age,gender,education and industry evaluation by using the given dataset.*

GIVEN DATASET:

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



1	Table Cod	State Cod	District Cc	Area Nam	Total/ Rur	Age group	Worked fc	Worked fc	Worked fc	Worked fc	Worked fc	Worked fc	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	In
2	B0806SC	`33	`000	State - TAI	Total	Total	1200828	589003	611825	221386	99368	122018	64235	34632	29603	907752	404844	502908	
3	B0806SC	`33	`000	State - TAI	Total	`5-14	27791	14125	13666	2447	1247	1200	1710	825	885	6398	3130	3268	
4	B0806SC	`33	`000	State - TAI	Total	15-34	514340	259560	254780	92423	43892	48531	24863	12711	12152	345420	152968	192452	
5	B0806SC	`33	`000	State - TAI	Total	35-59	542581	251957	290624	99202	40691	58511	29692	15927	13765	450052	192771	257281	
6	B0806SC	`33	`000	State - TAI	Total	60+	115103	62833	52270	27165	13465	13700	7930	5151	2779	105325	55730	49595	
7	B0806SC	`33	`000	State - TAI	Total	Age not st	1013	528	485	149	73	76	40	18	22	557	245	312	
8	B0806SC	`33	`000	State - TAI	Rural	Total	966645	459738	506907	174443	73663	100780	59637	32189	27448	824698	364131	460567	
9	B0806SC	`33	`000	State - TAI	Rural	`5-14	17239	8713	8526	1977	985	992	1443	684	759	6005	2922	3083	
10	B0806SC	`33	`000	State - TAI	Rural	15-34	406847	198575	208272	71974	31917	40057	22933	11766	11167	316885	138622	178263	
11	B0806SC	`33	`000	State - TAI	Rural	35-59	444800	199573	245227	77922	29808	48114	27799	14887	12912	406147	172178	233969	
12	B0806SC	`33	`000	State - TAI	Rural	60+	97011	52498	44513	22446	10902	11544	7425	4835	2590	95151	50192	44959	
13	B0806SC	`33	`000	State - TAI	Rural	Age not st	748	379	369	124	51	73	37	17	20	510	217	293	
14	B0806SC	`33	`000	State - TAI	Urban	Total	234183	129265	104918	46943	25705	21238	4598	2443	2155	83054	40713	42341	
15	B0806SC	`33	`000	State - TAI	Urban	`5-14	10552	5412	5140	470	262	208	267	141	126	393	208	185	
16	B0806SC	`33	`000	State - TAI	Urban	15-34	107493	60985	46508	20449	11975	8474	1930	945	985	28535	14346	14189	
17	B0806SC	`33	`000	State - TAI	Urban	35-59	97781	52384	45397	21280	10883	10397	1893	1040	853	43905	20593	23312	
18	B0806SC	`33	`000	State - TAI	Urban	60+	18092	10335	7757	4719	2563	2156	505	316	189	10174	5538	4636	
19	B0806SC	`33	`000	State - TAI	Urban	Age not st	265	149	116	25	22	3	3	1	2	47	28	19	
20	B0806SC	`33	`602	District - T	Total	Total	74448	39295	35153	15866	8004	7862	3066	1663	1403	42579	20345	22234	
21	B0806SC	`33	`602	District - T	Total	`5-14	2521	1284	1237	147	82	65	122	56	66	330	154	176	
22	B0806SC	`33	`602	District - T	Total	15-34	33568	18049	15519	6529	3654	2875	1225	632	593	15591	7257	8334	
23	B0806SC	`33	`602	District - T	Total	35-59	32568	16771	15797	7718	3529	4189	1414	792	622	22192	10446	11746	

ANALYSIS APPROACH:

- *The approaches for this analysis are:*
  - *Data collection*
  - *Data preprocessing*
  - *Choosing a cluster algorithm*
  - *Validation*
  - *Labelling*

Data collection:

- *Gather relevant data from various sources,ensuring that the insights align with the intended outcomes.*
- *The given dataset information are taken from the TN Marginal Workers In TamilNadu*

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

Data preprocessing:

- *The clean and preprocess the data and this involves checking missing values .*

Choosing a cluster algorithm and validation:

- *Some clustering algorithms are K-means,Hierarchical clustering,DBSCAN.*
- *The clustering algorithm involves structure of your data and the desired outcomes.*
- *It’s assess the quality of the cluster for this clustering with mixed data types (e.g:numerical age and categories),we need to use appropriate validation matrics.*

Labelling:

- *Assign labels to the clusters,representing different age groups and industrial categories from the given dataset.*
- *It assigns meaningful tags or categories to data points for supervised machine learning to predict the accurate predictions.*
- *It’s serves the development and evaluation of predictive models*

Program for analysis approach:

```
  # Import necessary libraries and gather data from various sources

Step 1:Define objectives

Step 2: Data collection

import pandas as pd

# Example: Reading data from a CSV file

data = pd.read_csv('data.csv')

# Step 3: Data Cleaning and Preprocessing

# Remove inconsistencies, handle missing values, and preprocess the data

# Example: Dropping missing values

data = data.dropna()

# Step 4: Exploratory Data Analysis (EDA)

# Explore data characteristics and relationships using visualizations and summary statistics
# Example: Generating a basic visualization
import matplotlib.pyplot as plt
plt.scatter(data['feature1'], data['feature2'])
plt.title('Scatter plot of feature1 vs. feature2')
plt.show()
# Step 5: Statistical Analysis and Modeling
from sklearn.linear_model import LinearRegression
X = data[['feature1']]
y = data['target']
model = LinearRegression()
model.fit(X, y)
```

DEMOGRAPHIC ANALYSIS:

- *Demographic analysis involves examining and interpreting data related to the charactersistics of a population.*
- *This process aims to insights various demographic factors such as age,gender,income ,industry categories and so on.*
- *It includes the data collection,cleaning and preprocessing by exploratory data analysis(EDA) to understand the distribution of different demographic variables.*

To perform the demographic ananlysis using the given dataset are as the following steps,they are:

- *Explore the dataset*
- *Load the data*
- *Filter required or relevant variables*
- *Create visualizations*
- *Conduct comparative analysis*

Load the data: *Load the dataset of the given dataset by using the python libraries such as pandas.*
Explore the data: *Explore the content and structure of given dataset and check the missing values in the given dataset.*
Filter relevant variables: *Identify the required demographic variables such as age,gender and industrial categories based on the dataset .*
Create visualizations: *Represent the given dataset in the format of histograms,bar charts,pie charts and so on. By the demographic analysis we identify the age distribution,industrial categories and gender distribution of marginal workers .*

VISUALIZATION:

- *Visualization in data analytics is the representation of complex data sets.*
- *It's presents the information in amore understandable and accessible format.*

Visualization types:

- *In data analytics,various visualization types are commonly used to represent data in understandable form.*
- *It includes line charts,bar charts,pie charts,scatter plots,histograms,heatmaps,box plots,area charts,choropleth maps and so on.*
- *This representations gives the data in the given dataset in a more understandable and insightful manner.*
  Scatter plot: *The scatter plot shows the relationship between the two variables in the given dataset.*
  Bar charts: *It is useful for the representation of comparing categories of data in the given dataset.*
  Line charts: *It is the ideal representation for showing trends over the time.*
  Histograms: *It is for the visualization of distributing of the numerical data.*

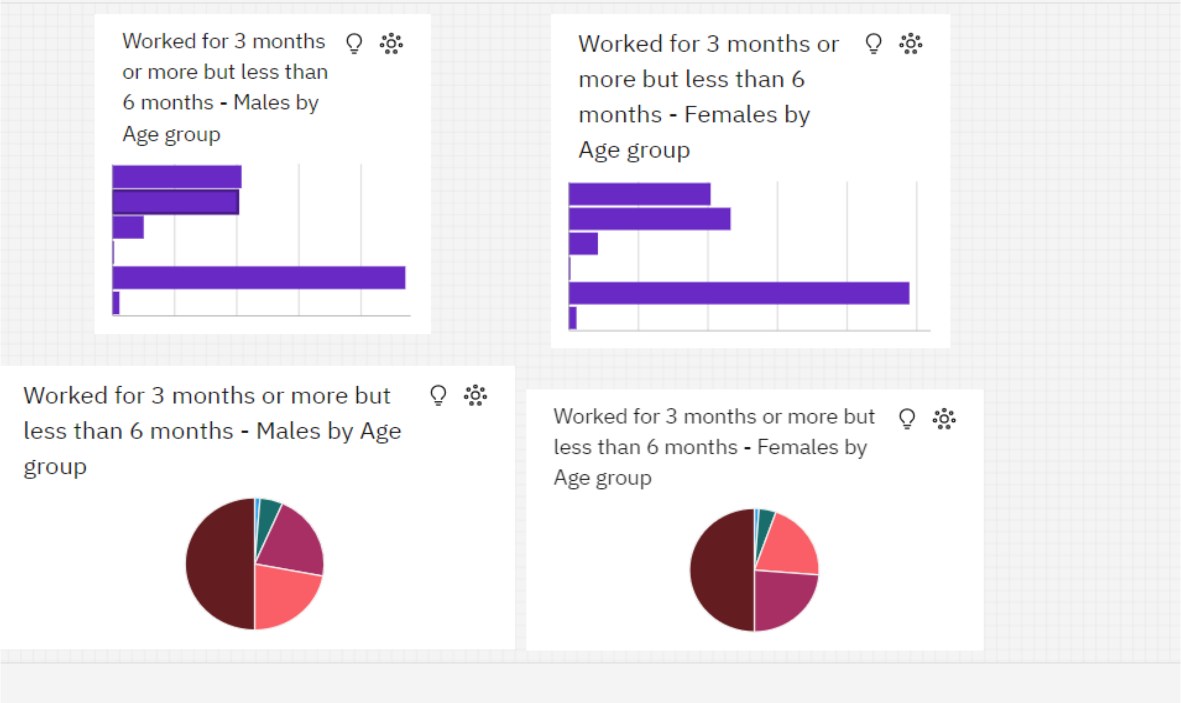
__Heatmaps:__ It is suitable for displaying the metrics where the values are represented as colors.
__Bubble charts:__ combine data points with different sizes,It is useful for the relationships between three variables.

Include the example outputs of data analysis and visualizations:
Let us consider the TN Marginal Workers in TamilNadu dataset to data analysis process and visualizations.The following steps are followed for the analysis and to visualize the data that provided in the given dataset.

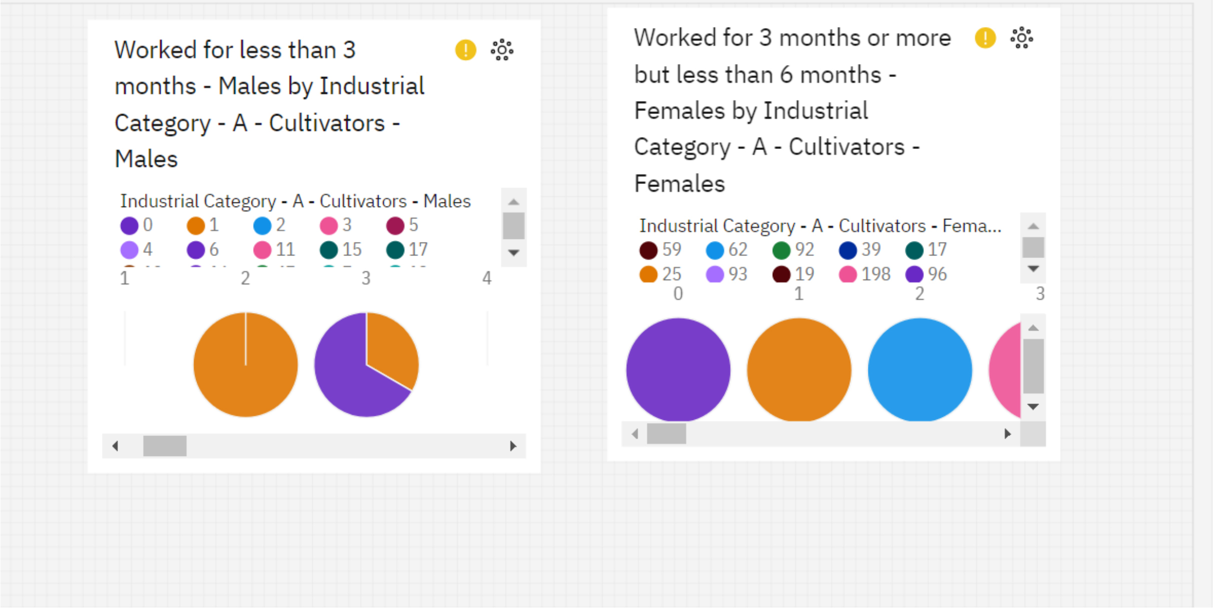
```
import matplotlib.pyplot as plt
# Load the dataset
data = pd.read_csv(“ https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil ’’)
# Filter data for marginal workers
marginal_workers = data[data[‘age'] == 'age group]
# Calculate the distribution based on age, industrial category, and sex
age_distribution = marginal_workers['age'].value_counts()
category_distribution =
marginal_workers['industrial_category'].value_counts()
sex_distribution = marginal_workers['sex'].value_counts()
# Create visualizations
plt.figure(figsize=(10, 6))
plt.subplot(1, 3, 1)
age_distribution.plot(kind='pie', title='Age Distribution of Marginal
Workers')
plt.xlabel('Age')
plt.ylabel('Count')
plt.subplot(1, 3, 2)
category_distribution.plot(kind='plot', title='Industrial Category
Distribution of Marginal Workers')
plt.xlabel('Industrial Category')
plt.ylabel('Count')
plt.subplot(1, 3, 3)
```

```
sex_distribution.plot(kind='pie', title='Sex Distribution of Marginal
Workers')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```

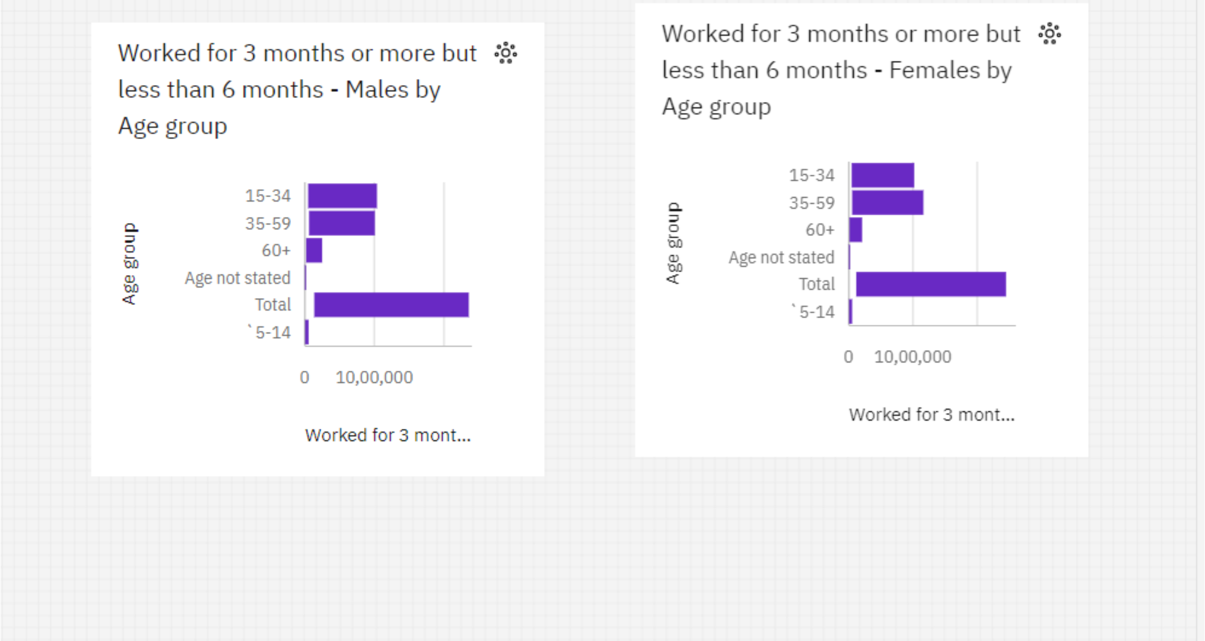
sample output for gender distribution:

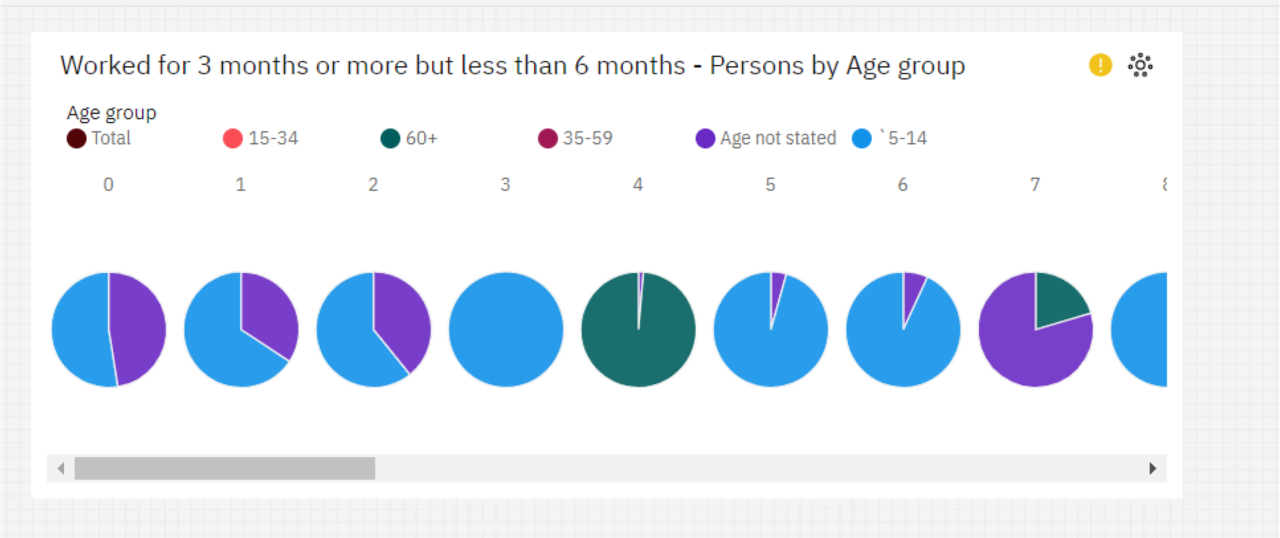

Worked for 3 months or more but less than 6 months - Males by Age group
Worked for 3 months or more but less than 6 months - Females by Age group
Worked for 3 months or more but less than 6 months - Males by Age group
Worked for 3 months or more but less than 6 months - Females by Age group

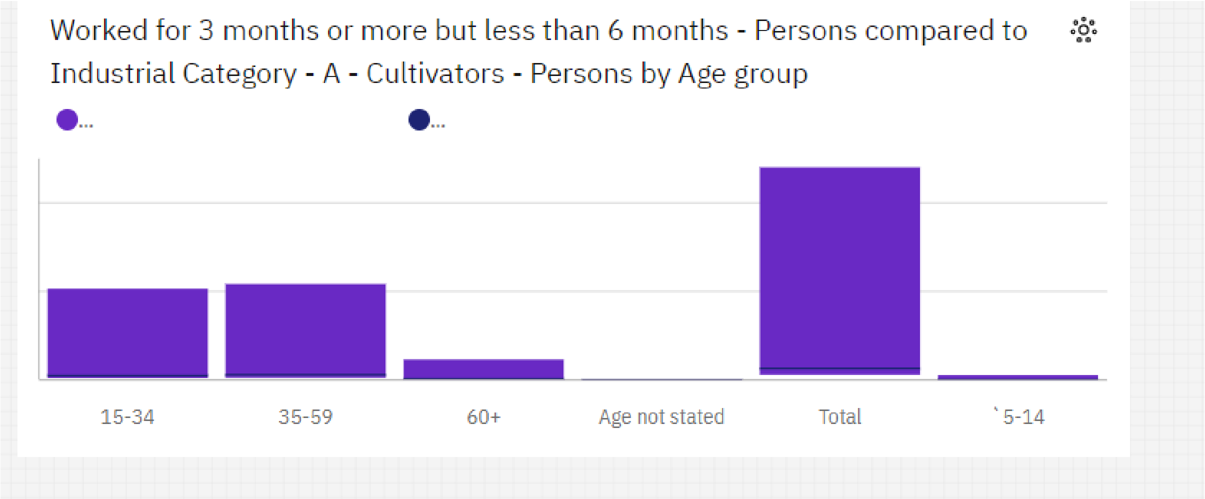
sample output for industrial categories distribution:


Worked for less than 3 months - Males by Industrial Category - A - Cultivators - Males
Industrial Category - A - Cultivators - Males
● 0 ● 1 ● 2 ● 3 ● 5
● 4 ● 6 ● 11 ● 15 ● 17
1 2 3 4
Worked for 3 months or more but less than 6 months - Females by Industrial Category - A - Cultivators - Females
Industrial Category - A - Cultivators - Fema...
● 59 ● 62 ● 92 ● 39 ● 17
● 25 ● 93 ● 19 ● 198 ● 96
0 1 2 3

Sample output for age group distribution:




Worked for 3 months or more but less than 6 months - Persons by Age group
Age group
● Total
● 15-34
● 60+
● 35-59
● Age not stated
● `5-14
0
1
2
3
4
5
6
7


Worked for 3 months or more but less than 6 months - Persons compared to Industrial Category - A - Cultivators - Persons by Age group
●...
●...
15-34
35-59
60+
Age not stated
Total
`5-14

CONCLUSION:

- *This aims to analyze the demographic characteristics of marginal workers in Tamil Nadu. By defining clear objectives, outlining the analysis approach, and selecting appropriate visualization, we will work towards a comprehensive understanding of this important demographic group.*
- *Clustering analysis helps identify natural patterns and groupings within the data.*
- *It can reveal how individuals or entities with certain age groups tend to belong to specific industrial categories.*
- *Clustering analysis is a valuable tool for identifying relationships between age groups and industrial categories.*
- *In this document the process of demographic analysis and creating the data visualizations using various libraries such as matplotlib,seaborn are done by using the given dataset of TN Marginal TamilNadu dataset.we observe the marginal workers multiple age groups,examination of industrial categories and gender distribution,these are all express the vital process of marginal workers.These are the concepts explained and visualized in this phase of TN Marginal Workers TamilNadu in India.*