Data Analytics with cognos

ASSESSMENT OF MARGINAL WORKERS IN TAMILNADU

PHASE 3

Name : LEKHASRI S
Reg no : 510421205021

Guided by: B. Moohambikai

Aim:

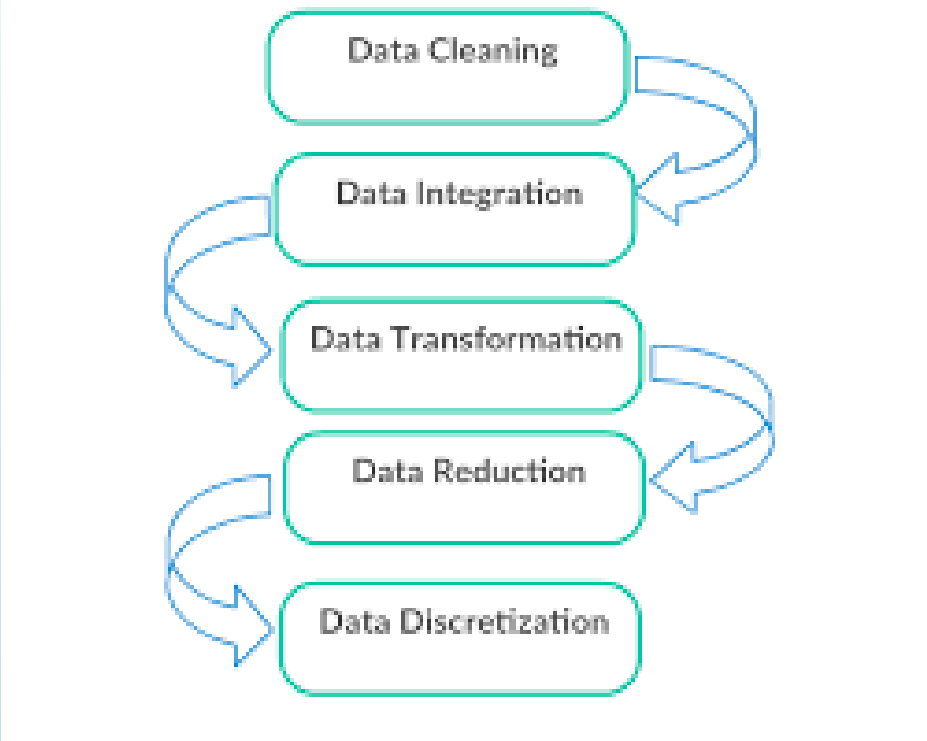
To Load and preprocessing the dataset. Start the data analysis by loading and preprocessing the dataset. Load the dataset using Python and data manipulation libraries (e.g., pandas).

Dataset:

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

Data preprocessing

Data preprocessing is the concept of changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.



Manipulation of data

Data manipulation refers to the process of adjusting data to make it organised and easier to read. Data manipulation language, or DML, is a programming language that adjusts data by inserting, deleting and modifying data in a database such as to cleanse or map the data

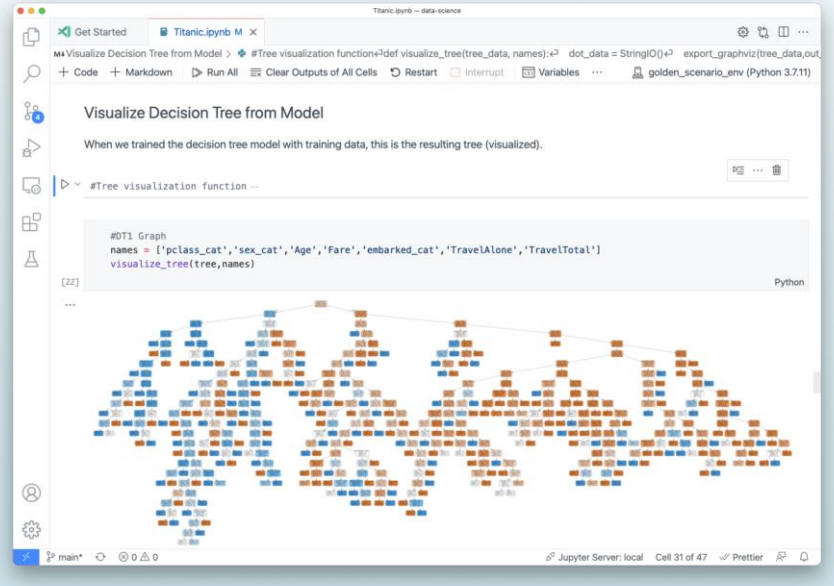
Program

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
url = https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil.csv
df = pd.read_csv(url)
# Print the first few rows of the dataframeprint(df.head())
# Get the number of unique categories
print("Unique categories:", df["Category"].nunique())

# Visualize the distribution of marginal workers across categories
plt.figure(figsize=(10, 6))sns.countplot(data=df, x="Category")
plt.title("Distribution of Marginal Workers Across Categories")
plt.xlabel("Category")
plt.ylabel("Number of Marginal Workers")
plt.show()

# Visualize the distribution of marginal workers across sex
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x="Sex")
plt.title("Distribution of Marginal Workers Across Sex")
plt.xlabel("Sex")
plt.ylabel("Number of Marginal Workers")plt.show()

# Visualize the distribution of marginal workers across age groupsplt.figure(figsize=(10, 6))
sns.countplot(data=df, x="Age")
plt.title("Distribution of Marginal Workers Across Age Groups")
plt.xlabel("Age")
plt.ylabel("Number of Marginal Workers")
plt.show()
```

Sample output

The output of the above program will be a pandas dataframe with the following columns:
Age , Industrial Category , Sex ,Scheduled Caste
The dataframe will contain the data from the linked dataset, which is about marginal workers classified by age, industrial category, sex, and scheduled caste in Tamil Nadu in 2011.
Here is an example of what the output might look like:

	Age	Industrial Category	Sex	Scheduled Caste
0	15	Agriculture & Allied	Male	Scheduled Caste
1	20	Manufacturing	Female	Not Scheduled Caste
2	25	Services	Male	Not Scheduled Caste
3	30	Construction	Female	Scheduled Caste
4	35	Trade & Commerce	Male	Not Scheduled Caste



Thank you!!!