TN MARGINAL WORKERS ASSESSMENT

PHASE 3

INTRODUCTION:

The goal of this phase is to preprocessing the dataset which involves fixing missing values,encoding categorical data,scaling numerical features,dealing with outliers and to ensure the data is accurate and ready for meaningful analysis.

STEPS INVOLVED IN PREPROCESSING:

1.IMPORT THE NECESSARY LIBRARIES AND LOAD THE DATASET:

import pandas as pd

import scipy

import numpy as np

from sklearn.preprocessing import MinMaxScaler

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read\_csv("C:\\Users\poovi\Downloads\Tnmarginalworkers.csv")

Print(df.head())

Table Code ... Industrial Category - R to U - Non HHI - Females

0 B0806SC ... 66287

1 B0806SC ... 9531

2 B0806SC ... 36126

3 B0806SC ... 16823

4 B0806SC ... 3671

[5 rows x 69 columns]

Print(df.info())

RangeIndex: 594 entries, 0 to 593

Data columns (total 69 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Table Code 594 non-null object

1 State Code 594 non-null object

2 District Code 594 non-null object

3 Area Name 594 non-null object

4 Total/ Rural/ Urban 594 non-null object

5 Age group 594 non-null object

6 Worked for 3 months or more but less than 6 months - Persons 594 non-null int64

7 Worked for 3 months or more but less than 6 months - Males 594 non-null int64

8 Worked for 3 months or more but less than 6 months - Females 594 non-null int64

9 Worked for less than 3 months - Persons 594 non-null int64

10 Worked for less than 3 months - Males 594 non-null int64

11 Worked for less than 3 months - Females 594 non-null int64

12 Industrial Category - A - Cultivators - Persons 594 non-null int64

13 Industrial Category - A - Cultivators - Males 594 non-null int64

14 Industrial Category - A - Cultivators - Females 594 non-null int64

15 Industrial Category - A - Agricultural labourers - Persons 594 non-null int64

16 Industrial Category - A - Agricultural labourers - Males 594 non-null int64

17 Industrial Category - A - Agricultural labourers - Females 594 non-null int64

18 Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Persons 594 non-null int64

19 Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Males 594 non-null int64

20 Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Females 594 non-null int64

21 Industrial Category - B - Persons 594 non-null int64

22 Industrial Category - B - Males 594 non-null int64

23 Industrial Category - B - Females 594 non-null int64

24 Industrial Category - C - HHI - Persons 594 non-null int64

25 Industrial Category - C - HHI - Males 594 non-null int64

26 Industrial Category - C - HHI - Females 594 non-null int64

27 Industrial Category - C - Non HHI - Persons 594 non-null int64

28 Industrial Category - C - Non HHI - Males 594 non-null int64

29 Industrial Category - C - Non HHI - Females 594 non-null int64

30 Industrial Category - D & E - Persons 594 non-null int64

31 Industrial Category - D & E - Males 594 non-null int64

32 Industrial Category - D & E - Females 594 non-null int64

33 Industrial Category - F - Persons 594 non-null int64

34 Industrial Category - F - Males 594 non-null int64

35 Industrial Category - F - Females 594 non-null int64

36 Industrial Category - G - HHI - Persons 594 non-null int64

37 Industrial Category - G - HHI - Males 594 non-null int64

38 Industrial Category - G - HHI - Females 594 non-null int64

39 Industrial Category - G - Non HHI - Persons 594 non-null int64

40 Industrial Category - G - Non HHI - Males 594 non-null int64

41 Industrial Category - G - Non HHI - Females 594 non-null int64

42 Industrial Category - H - Persons 594 non-null int64

43 Industrial Category - H - Males 594 non-null int64

44 Industrial Category - H - Females 594 non-null int64

45 Industrial Category - I - Persons 594 non-null int64

46 Industrial Category - I - Males 594 non-null int64

47 Industrial Category - I - Females 594 non-null int64

48 Industrial Category - J - HHI - Persons 594 non-null int64

49 Industrial Category - J - HHI - Males 594 non-null int64

50 Industrial Category - J - HHI - Females 594 non-null int64

51 Industrial Category - J - Non HHI - Persons 594 non-null int64

52 Industrial Category - J - Non HHI - Males 594 non-null int64

53 Industrial Category - J - Non HHI - Females 594 non-null int64

54 Industrial Category - K to M - Persons 594 non-null int64

55 Industrial Category - K to M - Males 594 non-null int64

56 Industrial Category - K to M - Females 594 non-null int64

57 Industrial Category - N to O - Persons 594 non-null int64

58 Industrial Category - N to O - Males 594 non-null int64

59 Industrial Category - N to O - Females 594 non-null int64

60 Industrial Category - P to Q - Persons 594 non-null int64

61 Industrial Category - P to Q - Males 594 non-null int64

62 Industrial Category - P to Q - Females 594 non-null int64

63 Industrial Category - R to U - HHI - Persons 594 non-null int64

64 Industrial Category - R to U - HHI - Males 594 non-null int64

65 Industrial Category - R to U - HHI - Females 594 non-null int64

66 Industrial Category - R to U - Non HHI - Persons 594 non-null int64

67 Industrial Category - R to U - Non HHI - Males 594 non-null int64

68 Industrial Category - R to U - Non HHI - Females 594 non-null int64

dtypes: int64(63), object(6)

memory usage: 320.3+ KB

2.STATISTICAL ANALYSIS:

print(df.describe())

Worked for 3 months or more but less than 6 months - Persons ... Industrial Category - R to U - Non HHI - Females

count 5.940000e+02 ... 594.000000

mean 1.617277e+04 ... 892.754209

std 7.607172e+04 ... 3988.125301

min 0.000000e+00 ... 0.000000

25% 2.872500e+02 ... 30.500000

50% 2.225500e+03 ... 135.000000

75% 9.628500e+03 ... 500.000000

max 1.200828e+06 ... 66287.000000

[8 rows x 63 columns]

3.CHECK THE OUTLIERS:

fig, axs = plt.subplots(9,1,dpi=95, figsize=(7,17))

i = 0

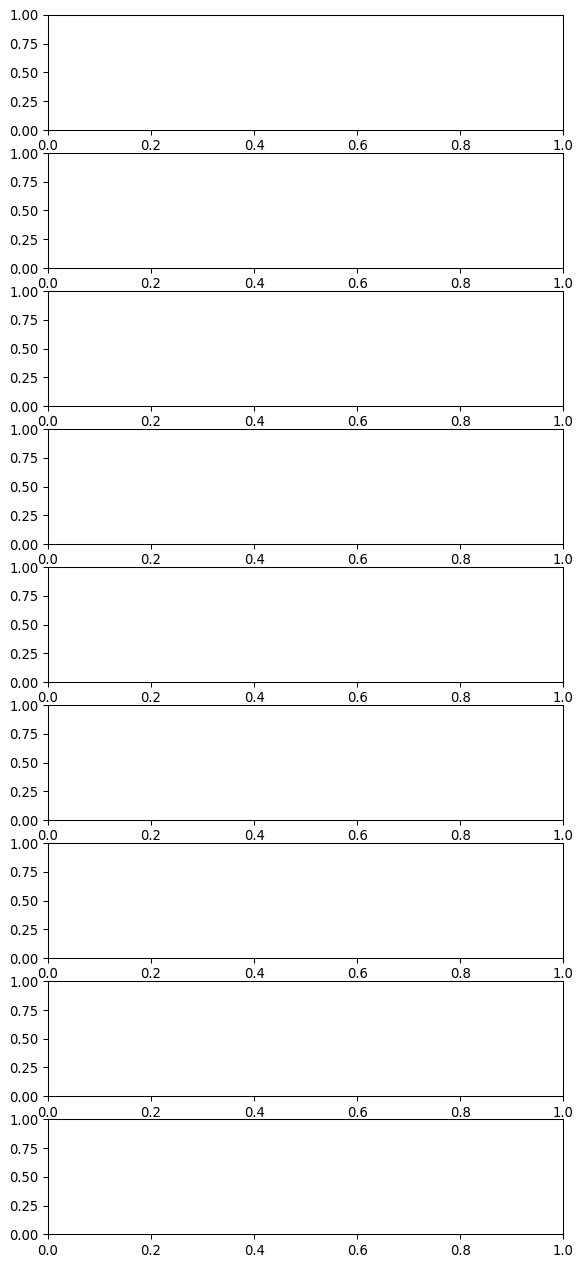
for col in df.columns:

axs[i].boxplot(df[col], vert=False)

axs[i].set\_ylabel(col)

i+=1

plt.show()



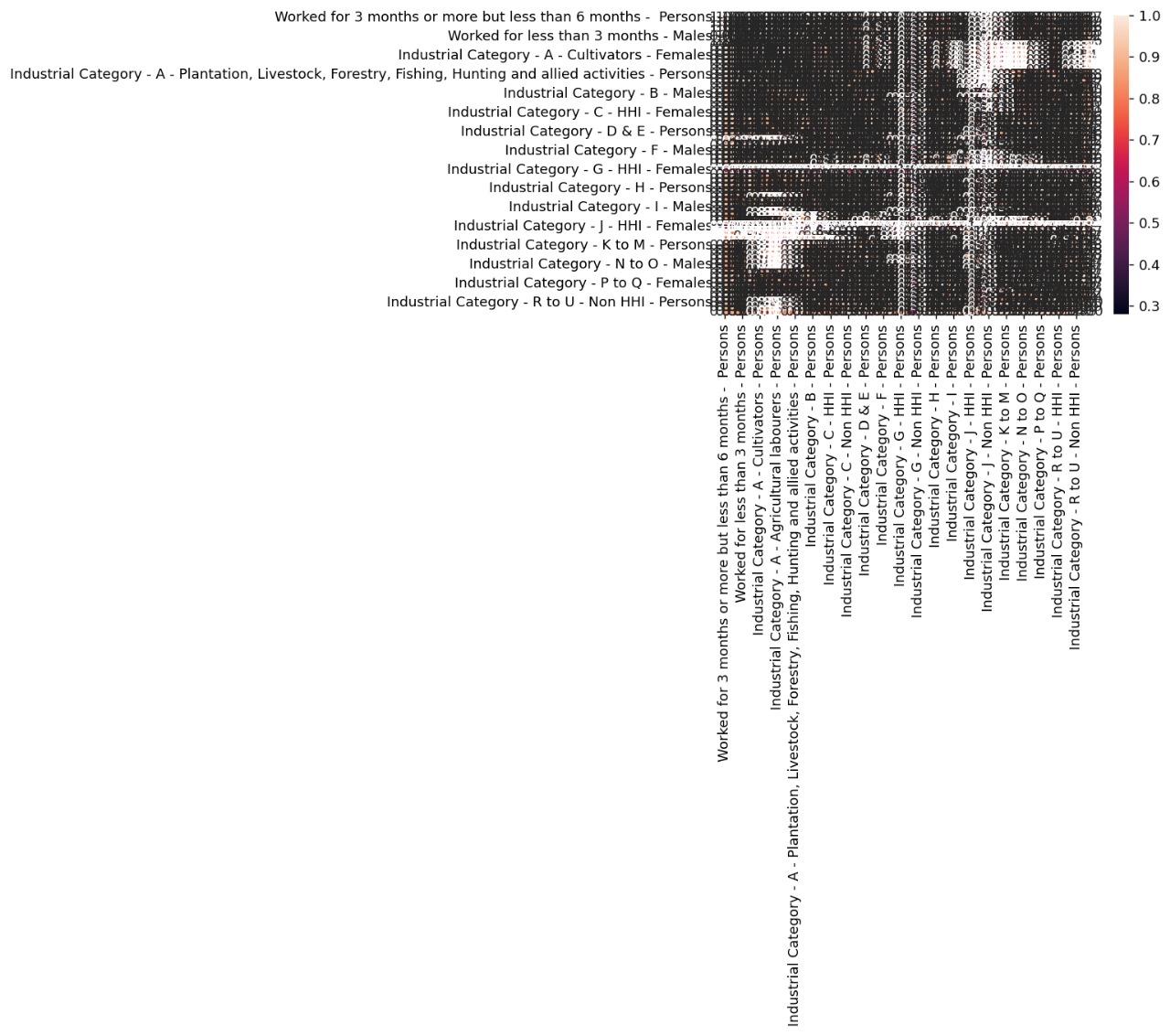
5.CORRELATION:

corr = df.corr()

plt.figure(dpi=130)

sns.heatmap(df.corr(), annot=True, fmt= '.2f')

plt.show()



5.MISSING VALUE:

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

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

6.VISUALIZATIONS

HISTOGRAM:

marginal\_workers\_data = df['Industrial Category - A - Cultivators - Persons']

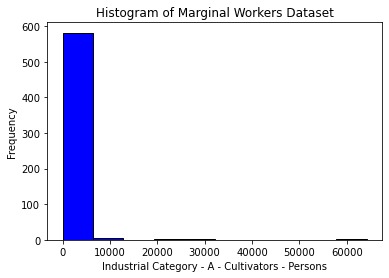
plt.hist(marginal\_workers\_data, bins=10, color='blue', edgecolor='black')

plt.xlabel('Industrial Category - A - Cultivators - Persons')

plt.ylabel('Frequency')

plt.title('Histogram of Marginal Workers Dataset')

plt.show()



PIECHART:

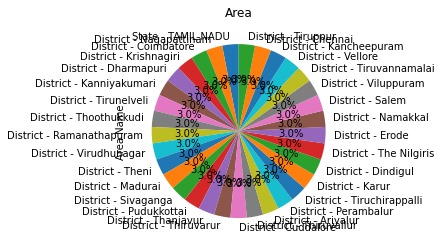
selected\_column = df['Area Name']

value\_counts = selected\_column.value\_counts()

value\_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90)

plt.title('Area')

plt.show()



CONCLUSION:

In this process we load the dataset and done the statistical analysis by using describe() function.Then we find the outliers,correlation,missing values and represent the dataset in visual manner using plots.