## College code: 4212

Register num: 421221243001

**ASSESSMENT OF MARGINAL WORKERS IN TAMIL NADU**

**DATA ANALYTICS WITH COGNOS:GROUP2PHASE:3**

This phase involves in designing of the steps that defining in each phase of the

previous documentation this involves importing necessary functions, data processing and so on in this phase we have to begin our project by loading and preprocessing the dataset.

The IBM suggests using the jupyter notebook for loading and preprocess the dataset:

Here for this project title we need to define the loading the libraries, understand the data and visualize the missing values.

For this certain inputs are defined for this project.in this phase each of the input lines of the project is given as follows:

# phase3

## October 17, 2023

[1]:

**import pandas as pd import numpy as np import missingno as msno**

[2]:

df = pd.read\_csv('WA\_Fn-UseC\_-Telco-Customer-Churn.csv')

[3]:

df.head()

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [3]: | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | \ |
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No |  |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes |  |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Yes |  |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | No |  |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Yes |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | MultipleLines | InternetService OnlineSecurity | … | DeviceProtection | \ |
| 0 | No | phone service | DSL No | … | No |  |
| 1 |  | No | DSL Yes | … | Yes |  |
| 2 |  | No | DSL Yes | … | No |  |
| 3 | No | phone service | DSL Yes | … | Yes |  |
| 4 |  | No | Fiber optic No | … | No |  |

TechSupport StreamingTV StreamingMovies Contract PaperlessBilling \

1. No No No Month-to-month Yes
2. No No No One year No
3. No No No Month-to-month Yes
4. Yes No No One year No
5. No No No Month-to-month Yes

PaymentMethod MonthlyCharges TotalCharges Churn

1. Electronic check 29.85 29.85 No
2. Mailed check 56.95 1889.5 No
3. Mailed check 53.85 108.15 Yes
4. Bank transfer (automatic) 42.30 1840.75 No
5. Electronic check 70.70 151.65 Yes

[5 rows x 21 columns]

[4]:

df.shape

[4]: (7043, 21)

[5]:

df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 7043 entries, 0 to 7042 Data columns (total 21 columns):

# Column Non-Null Count Dtype

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 customerID | 7043 | non-null |  | object |
| 1 gender | 7043 | non-null |  | object |
| 2 SeniorCitizen | 7043 | non-null |  | int64 |
| 3 Partner | 7043 | non-null |  | object |
| 4 Dependents | 7043 | non-null |  | object |
| 5 tenure | 7043 | non-null |  | int64 |
| 6 PhoneService | 7043 | non-null |  | object |
| 7 MultipleLines | 7043 | non-null |  | object |
| 8 InternetService | 7043 | non-null |  | object |
| 9 OnlineSecurity | 7043 | non-null |  | object |
| 10 OnlineBackup | 7043 | non-null |  | object |
| 11 DeviceProtection | 7043 | non-null |  | object |
| 12 TechSupport | 7043 | non-null |  | object |
| 13 StreamingTV | 7043 | non-null |  | object |
| 14 StreamingMovies | 7043 | non-null |  | object |
| 15 Contract | 7043 | non-null |  | object |
| 16 PaperlessBilling | 7043 | non-null |  | object |
| 17 PaymentMethod | 7043 | non-null |  | object |
| 18 MonthlyCharges | 7043 | non-null |  | float64 |
| 19 TotalCharges | 7043 | non-null |  | object |
| 20 Churn | 7043 | non-null |  | object |

dtypes: float64(1), int64(2), object(18) memory usage: 1.1+ MB

[6]:

df.columns.values

1. : array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'], dtype=object)

[7]:

df.dtypes

1. : customerID object gender object

SeniorCitizen int64

Partner object

Dependents object

tenure int64

PhoneService object

MultipleLines object

InternetService object

OnlineSecurity object

OnlineBackup object DeviceProtection object TechSupport object

StreamingTV object

StreamingMovies object

Contract object PaperlessBilling object PaymentMethod object

MonthlyCharges float64

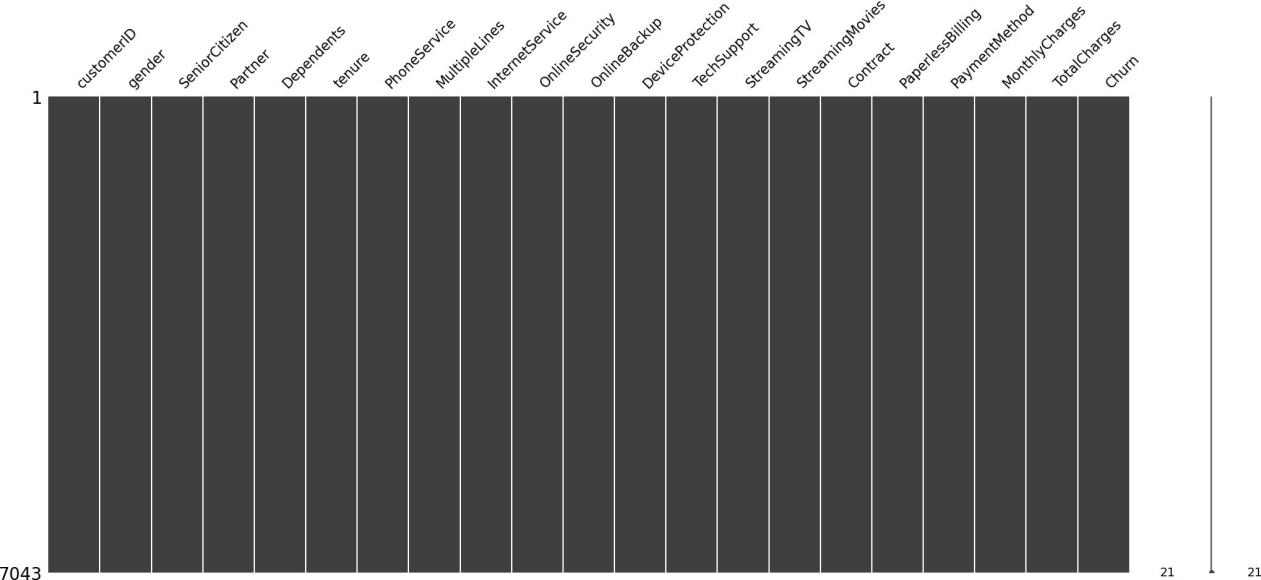
TotalCharges object

Churn object

dtype: object

[8]:

msno.matrix(df);



[9]:

df = df.drop(['customerID'], axis = 1) df.head()

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [9]: | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | \ |
| 0 | Female | 0 | Yes | No | 1 | No |  |
| 1 | Male | 0 | No | No | 34 | Yes |  |
| 2 | Male | 0 | No | No | 2 | Yes |  |
| 3 | Male | 0 | No | No | 45 | No |  |
| 4 | Female | 0 | No | No | 2 | Yes |  |

MultipleLines InternetService OnlineSecurity OnlineBackup \

1. No phone service DSL No Yes
2. No DSL Yes No
3. No DSL Yes Yes
4. No phone service DSL Yes No
5. No Fiber optic No No

DeviceProtection TechSupport StreamingTV StreamingMovies Contract \

1. No No No No Month-to-month
2. Yes No No No One year
3. No No No No Month-to-month
4. Yes Yes No No One year
5. No No No No Month-to-month

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | PaperlessBilling | PaymentMethod | MonthlyCharges | TotalCharges | \ |
| 0 | | Yes | Electronic check | 29.85 | 29.85 |  |
| 1 | | No | Mailed check | 56.95 | 1889.5 |  |
| 2 | | Yes | Mailed check | 53.85 | 108.15 |  |
| 3 | | No | Bank transfer (automatic) | 42.30 | 1840.75 |  |
| 4 | | Yes | Electronic check | 70.70 | 151.65 |  |
|  | Churn   1. No 2. No 3. Yes 4. No 5. Yes | |  | | | |
| [10]: | df['TotalCharges'] df.isnull().sum() | | = pd.to\_numeric(df.TotalCharges, errors='coerce') | | | |
| [10]: | gender | | 0 | | | |
|  | SeniorCitizen | | 0 | | | |
|  | Partner | | 0 | | | |
|  | Dependents | | 0 | | | |
|  | tenure | | 0 | | | |
|  | PhoneService | | 0 | | | |
|  | MultipleLines | | 0 | | | |
|  | InternetService | | 0 | | | |
|  | OnlineSecurity | | 0 | | | |

|  |  |
| --- | --- |
| OnlineBackup | 0 |
| DeviceProtection | 0 |
| TechSupport | 0 |
| StreamingTV | 0 |
| StreamingMovies | 0 |
| Contract | 0 |
| PaperlessBilling | 0 |
| PaymentMethod | 0 |
| MonthlyCharges | 0 |
| TotalCharges | 11 |
| Churn  dtype: int64 | 0 |

[11]:

df[np.isnan(df['TotalCharges'])]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [11]: | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | \ |
| 488 | Female | 0 | Yes | Yes | 0 | No |  |
| 753 | Male | 0 | No | Yes | 0 | Yes |  |
| 936 | Female | 0 | Yes | Yes | 0 | Yes |  |
| 1082 | Male | 0 | Yes | Yes | 0 | Yes |  |
| 1340 | Female | 0 | Yes | Yes | 0 | No |  |
| 3331 | Male | 0 | Yes | Yes | 0 | Yes |  |
| 3826 | Male | 0 | Yes | Yes | 0 | Yes |  |
| 4380 | Female | 0 | Yes | Yes | 0 | Yes |  |
| 5218 | Male | 0 | Yes | Yes | 0 | Yes |  |
| 6670 | Female | 0 | Yes | Yes | 0 | Yes |  |
| 6754 | Male | 0 | No | Yes | 0 | Yes |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | MultipleLines | InternetService |  | OnlineSecurity | \ |
| 488 | No | phone service | DSL |  | Yes |  |
| 753 |  | No | No | No | internet service |  |
| 936 |  | No | DSL |  | Yes |  |
| 1082 |  | Yes | No | No | internet service |  |
| 1340 | No | phone service | DSL |  | Yes |  |
| 3331 |  | No | No | No | internet service |  |
| 3826 |  | Yes | No | No | internet service |  |
| 4380 |  | No | No | No | internet service |  |
| 5218 |  | No | No | No | internet service |  |
| 6670 |  | Yes | DSL |  | No |  |
| 6754 |  | Yes | DSL |  | Yes |  |

488

OnlineBackup

No

DeviceProtection

Yes

TechSupport \

Yes

753 No internet service No internet service No internet service 936 Yes Yes No 1082 No internet service No internet service No internet service 1340 Yes Yes Yes

[12]:

df[df['tenure'] == 0].index

3331 No internet service No internet service No internet service 3826 No internet service No internet service No internet service 4380 No internet service No internet service No internet service 5218 No internet service No internet service No internet service 6670 Yes Yes Yes

6754 Yes No Yes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | StreamingTV |  | StreamingMovies | | Contract | PaperlessBilling | | \ |
| 488 |  | Yes |  | No | | Two year | Yes | |  |
| 753 | No | internet service | No | internet service | | Two year | No | |  |
| 936 |  | Yes |  | Yes | | Two year | No | |  |
| 1082 | No | internet service | No | internet service | | Two year | No | |  |
| 1340 |  | Yes |  | No | | Two year | No | |  |
| 3331 | No | internet service | No | internet service | | Two year | No | |  |
| 3826 | No | internet service | No | internet service | | Two year | No | |  |
| 4380 | No | internet service | No | internet service | | Two year | No | |  |
| 5218 | No | internet service | No | internet service | | One year | Yes | |  |
| 6670 |  | Yes |  | No | | Two year | No | |  |
| 6754 |  | No |  | No | | Two year | Yes | |  |
|  | PaymentMethod | | | | MonthlyCharges | TotalCharges | | Churn | |
| 488 | Bank transfer (automatic) | | | | 52.55 | NaN | | No | |
| 753 | Mailed check | | | | 20.25 | NaN | | No | |
| 936 | Mailed check | | | | 80.85 | NaN | | No | |
| 1082 | Mailed check | | | | 25.75 | NaN | | No | |
| 1340 | Credit card (automatic) | | | | 56.05 | NaN | | No | |
| 3331 | Mailed check | | | | 19.85 | NaN | | No | |
| 3826 | Mailed check | | | | 25.35 | NaN | | No | |
| 4380 | Mailed check | | | | 20.00 | NaN | | No | |
| 5218 | Mailed check | | | | 19.70 | NaN | | No | |
| 6670 | Mailed check | | | | 73.35 | NaN | | No | |
| 6754 | Bank transfer (automatic) | | | | 61.90 | NaN | | No | |

[12]: Int64Index([488, 753, 936, 1082, 1340, 3331, 3826, 4380, 5218, 6670, 6754],

dtype='int64')

[13]:

df.drop(labels=df[df['tenure'] == 0].index, axis=0, inplace=**True**) df[df['tenure'] == 0].index

1. : Int64Index([], dtype='int64') [14]:

df.fillna(df["TotalCharges"].mean())

1. : gender SeniorCitizen Partner Dependents tenure PhoneService \

0 Female 0 Yes No 1 No

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Male |  | 0 | No |  | No | 34 | Yes |
| 2 | Male |  | 0 | No |  | No | 2 | Yes |
| 3 | Male |  | 0 | No |  | No | 45 | No |
| 4 | Female |  | 0 | No |  | No | 2 | Yes |
| … | … | … | … | … | … |  | … |  |
| 7038 | Male |  | 0 | Yes |  | Yes | 24 | Yes |
| 7039 | Female |  | 0 | Yes |  | Yes | 72 | Yes |
| 7040 | Female |  | 0 | Yes |  | Yes | 11 | No |
| 7041 | Male |  | 1 | Yes |  | No | 4 | Yes |
| 7042 | Male |  | 0 | No |  | No | 66 | Yes |

MultipleLines InternetService OnlineSecurity OnlineBackup \

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | No | phone | service |  | DSL |  | No |  | Yes | |
| 1 |  |  | No |  | DSL |  | Yes |  | No | |
| 2 |  |  | No |  | DSL |  | Yes |  | Yes | |
| 3 | No | phone | service |  | DSL |  | Yes |  | No | |
| 4  …  7038 |  |  | No  …  Yes | Fiber  … | optic  DSL | … | No  Yes | … | No  No | |
| 7039 |  |  | Yes | Fiber | optic |  | No |  | Yes | |
| 7040 | No | phone | service |  | DSL |  | Yes |  | No | |
| 7041 |  |  | Yes | Fiber | optic |  | No |  | No | |
| 7042 |  |  | No | Fiber | optic |  | Yes |  | No | |
|  | DeviceProtection | | | TechSupport | StreamingTV | | StreamingMovies | | Contract | \ |
| 0 | No | | | No | No | | No | | Month-to-month |  |
| 1 | Yes | | | No | No | | No | | One year |  |
| 2 | No | | | No | No | | No | | Month-to-month |  |
| 3 | Yes | | | Yes | No | | No | | One year |  |
| 4 | No | | | No | No | | No | | Month-to-month |  |
| … | … | | | … | … | | … | | … |  |
| 7038 | Yes | | | Yes | Yes | | Yes | | One year |  |
| 7039 | Yes | | | No | Yes | | Yes | | One year |  |
| 7040 | No | | | No | No | | No | | Month-to-month |  |
| 7041 | No | | | No | No | | No | | Month-to-month |  |
| 7042 | Yes | | | Yes | Yes | | Yes | | Two year |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PaperlessBilling |  | PaymentMethod | MonthlyCharges \ |
| 0 | Yes |  | Electronic check | 29.85 |
| 1 | No |  | Mailed check | 56.95 |
| 2 | Yes |  | Mailed check | 53.85 |
| 3 | No | Bank | transfer (automatic) | 42.30 |
| 4 | Yes |  | Electronic check | 70.70 |
| … | … |  | … | … |
| 7038 | Yes |  | Mailed check | 84.80 |
| 7039 | Yes | Credit card (automatic) | | 103.20 |
| 7040 | Yes | Electronic check | | 29.60 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 7041 | Yes | | Mailed check | 74.40 |
| 7042 | Yes Bank | | transfer (automatic) | 105.65 |
|  | TotalCharges | Churn | | |
| 0 | 29.85 | No | | |
| 1 | 1889.50 | No | | |
| 2 | 108.15 | Yes | | |
| 3 | 1840.75 | No | | |
| 4  … 7038 | 151.65  … …  1990.50 | Yes  No | | |
| 7039 | 7362.90 | No | | |
| 7040 | 346.45 | No | | |
| 7041 | 306.60 | Yes | | |
| 7042 | 6844.50 | No | | |

[7032 rows x 20 columns]

[15]:

df.isnull().sum()

1. : gender 0

SeniorCitizen 0

Partner 0

Dependents 0

tenure 0

PhoneService 0

MultipleLines 0

InternetService 0

OnlineSecurity 0

OnlineBackup 0

DeviceProtection 0

TechSupport 0

StreamingTV 0

StreamingMovies 0

Contract 0

PaperlessBilling 0

PaymentMethod 0

MonthlyCharges 0

TotalCharges 0

Churn 0

dtype: int64

[16]:

df["SeniorCitizen"]= df["SeniorCitizen"].map({0: "No", 1: "Yes"}) df.head()

1. : gender SeniorCitizen Partner Dependents tenure PhoneService \

0 Female No Yes No 1 No

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Male | No | No | No | 34 | Yes |
| 2 | Male | No | No | No | 2 | Yes |
| 3 | Male | No | No | No | 45 | No |
| 4 | Female | No | No | No | 2 | Yes |

MultipleLines InternetService OnlineSecurity OnlineBackup \

1. No phone service DSL No Yes
2. No DSL Yes No
3. No DSL Yes Yes
4. No phone service DSL Yes No
5. No Fiber optic No No

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract \ |
| 0 | No | No | No | No | Month-to-month |
| 1 | Yes | No | No | No | One year |
| 2 | No | No | No | No | Month-to-month |
| 3 | Yes | Yes | No | No | One year |
| 4 | No | No | No | No | Month-to-month |

PaperlessBilling PaymentMethod MonthlyCharges TotalCharges \

1. Yes Electronic check 29.85 29.85
2. No Mailed check 56.95 1889.50
3. Yes Mailed check 53.85 108.15
4. No Bank transfer (automatic) 42.30 1840.75
5. Yes Electronic check 70.70 151.65

Churn

1. No
2. No
3. Yes
4. No
5. Yes

[17]:

df["InternetService"].describe(include=['object', 'bool'])

[17]: count 7032

unique top freq

3

Fiber optic

3096

Name: InternetService, dtype: object

[18]:

numerical\_cols = ['tenure', 'MonthlyCharges', 'TotalCharges'] df[numerical\_cols].describe()

1. : tenure MonthlyCharges TotalCharges count 7032.000000 7032.000000 7032.000000

mean 32.421786 64.798208 2283.300441

|  |  |  |  |
| --- | --- | --- | --- |
| std | 24.545260 | 30.085974 | 2266.771362 |
| min | 1.000000 | 18.250000 | 18.800000 |
| 25% | 9.000000 | 35.587500 | 401.450000 |
| 50% | 29.000000 | 70.350000 | 1397.475000 |
| 75% | 55.000000 | 89.862500 | 3794.737500 |
| max | 72.000000 | 118.750000 | 8684.800000 |

# phase-3

## October 18, 2023

[2]:

**import pandas as pd import numpy as np import missingno as msno**

[4]:

df = pd.read\_csv("C:/Users/BALAJI/Downloads/

↪DDW\_B06SC\_3300\_State\_TAMIL\_NADU-2011.csv")

[5]:

df.head()

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [5]: |  | Table Code | State Code | District | Code | Area Name Total/ Rural/ | Urban | \ |
|  | 0 | B0806SC | `33 |  | `000 State | - TAMIL NADU | Total |  |
|  | 1 | B0806SC | `33 |  | `000 State | - TAMIL NADU | Total |  |
|  | 2 | B0806SC | `33 |  | `000 State | - TAMIL NADU | Total |  |
|  | 3 | B0806SC | `33 |  | `000 State | - TAMIL NADU | Total |  |
|  | 4 | B0806SC | `33 |  | `000 State | - TAMIL NADU | Total |  |

Age group Worked for 3 months or more but less than 6 months - Persons \ 0 Total 1200828

1 `5-14 27791

2 15-34 514340

3 35-59 542581

4 60+ 115103

Worked for 3 months or more but less than 6 months - Males \ 0 589003

1 14125

2 259560

3 251957

4 62833

Worked for 3 months or more but less than 6 months - Females \ 0 611825

1 13666

2 254780

3 290624

4 52270

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Worked for | | less than 3 months - Persons | | | | | | | … | \ |
| 0 | | 221386 | | | | | | | … |  |
| 1 | | 2447 | | | | | | | … |  |
| 2 | | 92423 | | | | | | | … |  |
| 3 | | 99202 | | | | | | | … |  |
| 4 | | 27165 | | | | | | | … |  |
| 0 | Industrial | Category | - | N | to | O | - | Females  3565 | \ | |
| 1 |  |  |  |  |  |  |  | 11 |  | |
| 2 |  |  |  |  |  |  |  | 1754 |  | |
| 3 |  |  |  |  |  |  |  | 1619 |  | |
| 4 |  |  |  |  |  |  |  | 175 |  | |

Industrial Category - P to Q - Persons \ 0 11080

1 122

2 7536

3 3205

4 211

Industrial Category - P to Q - Males \ 0 4019

1 71

2 2718

3 1131

4 93

Industrial Category - P to Q - Females \ 0 7061

1 51

2 4818

3 2074

4 118

Industrial Category - R to U - HHI - Persons \ 0 16833

1 427

2 8346

3 6591

4 1457

Industrial Category - R to U - HHI - Males \ 0 4266

1 169

2 2127

3 1487

[6]:

df.shape

4 483

Industrial Category - R to U - HHI - Females \ 0 12567

1 258

2 6219

3 5104

4 974

Industrial Category - R to U - Non HHI - Persons \ 0 122088

1 19305

2 68929

3 26498

4 7065

Industrial Category - R to U - Non HHI - Males \ 0 55801

1 9774

2 32803

3 9675

4 3394

Industrial Category - R to U - Non HHI - Females 0 66287

1 9531

2 36126

3 16823

4 3671

[5 rows x 69 columns]

[6]: (594, 69)

[7]:

df.info()

<class 'pandas.core.frame.DataFrame'> 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

1. Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Persons 594 non-null int64
2. Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Males 594 non-null int64
3. Industrial Category - A - Plantation, Livestock, Forestry, Fishing, Hunting and allied activities - Females 594 non-null int64
4. 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

[8]:

df.columns.values

* 1. : array(['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',

'Worked for less than 3 months - Males', 'Worked for less than 3 months - Females',

'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',

[9]:

df.dtypes

'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'], dtype=object)

* 1. : Table Code object

State Code object

District Code object

Area Name object

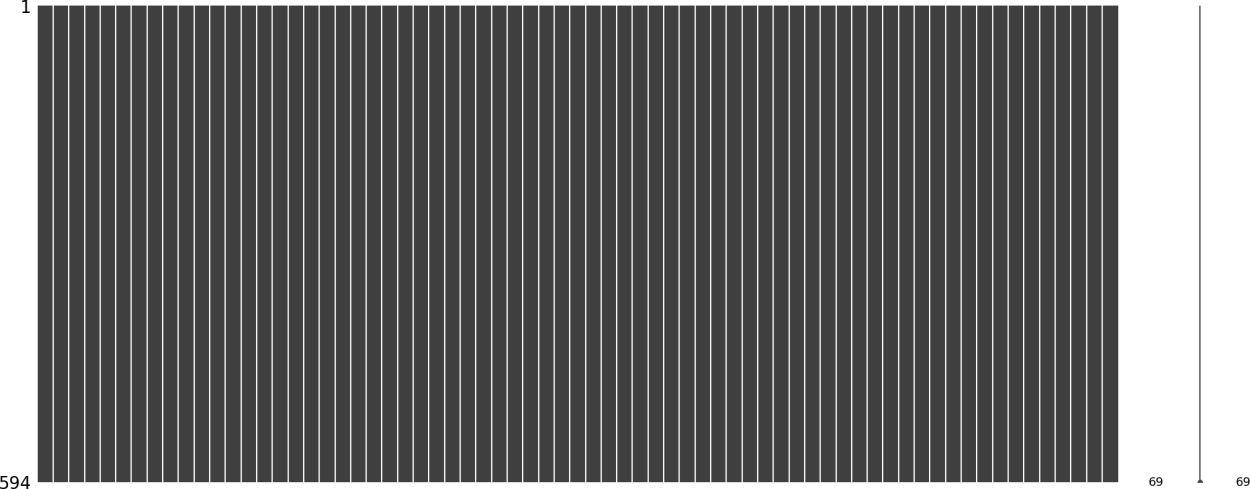
Total/ Rural/ Urban object

… Industrial Category - R to U - HHI - Males int64

Industrial Category - R to U - HHI - Females int64 Industrial Category - R to U - Non HHI - Persons int64 Industrial Category - R to U - Non HHI - Males int64 Industrial Category - R to U - Non HHI - Females int64 Length: 69, dtype: object

[10]:

msno.matrix(df);



[12]:

df = df.drop(['Table Code'], axis = 1) df.head()

[12]: State Code District Code Area Name Total/ Rural/ Urban Age group \

1. `33 `000 State - TAMIL NADU Total Total
2. `33 `000 State - TAMIL NADU Total `5-14
3. `33 `000 State - TAMIL NADU Total 15-34
4. `33 `000 State - TAMIL NADU Total 35-59
5. `33 `000 State - TAMIL NADU Total 60+

Worked for 3 months or more but less than 6 months - Persons \ 0 1200828

1 27791

2 514340

3 542581

4 115103

Worked for 3 months or more but less than 6 months - Males \ 0 589003

1 14125

2 259560

3 251957

4 62833

Worked for 3 months or more but less than 6 months - Females \ 0 611825

1 13666

2 254780

3 290624

4 52270

Worked for less than 3 months - Persons \ 0 221386

1 2447

2 92423

3 99202

4 27165

Worked for less than 3 months - Males … \ 0 99368 …

1 1247 …

2 43892 …

3 40691 …

4 13465 …

Industrial Category - N to O - Females \ 0 3565

1 11

2 1754

3 1619

4 175

Industrial Category - P to Q - Persons \ 0 11080

1 122

2 7536

3 3205

4 211

Industrial Category - P to Q - Males \ 0 4019

1 71

2 2718

3 1131

4 93

Industrial Category - P to Q - Females \ 0 7061

1 51

2 4818

3 2074

4 118

Industrial Category - R to U - HHI - Persons \ 0 16833

1 427

2 8346

3 6591

[20]:

df[df['Age group'] == 0].index

4 1457

Industrial Category - R to U - HHI - Males \ 0 4266

1 169

2 2127

3 1487

4 483

Industrial Category - R to U - HHI - Females \ 0 12567

1 258

2 6219

3 5104

4 974

Industrial Category - R to U - Non HHI - Persons \ 0 122088

1 19305

2 68929

3 26498

4 7065

Industrial Category - R to U - Non HHI - Males \ 0 55801

1 9774

2 32803

3 9675

4 3394

Industrial Category - R to U - Non HHI - Females 0 66287

1 9531

2 36126

3 16823

4 3671

[5 rows x 68 columns]

1. : Index([], dtype='int64')

[21]:

df.drop(labels=df[df['Age group'] == 0].index, axis=0, inplace=**True**) df[df['Age group'] == 0].index

1. : Index([], dtype='int64')

[26]:

df.fillna(df["Industrial Category - R to U - HHI - Persons"].mean())

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [26]: |  | State | Code | District | Code |  | Area Name | Total/ | Rural/ | Urban | \ |
|  | 0 |  | `33 |  | `000 | State - | TAMIL NADU |  |  | Total |  |
|  | 1 |  | `33 |  | `000 | State - | TAMIL NADU |  |  | Total |  |
|  | 2 |  | `33 |  | `000 | State - | TAMIL NADU |  |  | Total |  |
|  | 3 |  | `33 |  | `000 | State - | TAMIL NADU |  |  | Total |  |
|  | 4 |  | `33 |  | `000 | State - | TAMIL NADU |  |  | Total |  |
|  | .. |  | … | … |  |  | … |  | … |  |  |
|  | 589 |  | `33 |  | `633 | District | - Tiruppur |  |  | Urban |  |
|  | 590 |  | `33 |  | `633 | District | - Tiruppur |  |  | Urban |  |
|  | 591 |  | `33 |  | `633 | District | - Tiruppur |  |  | Urban |  |
|  | 592 |  | `33 |  | `633 | District | - Tiruppur |  |  | Urban |  |
|  | 593 |  | `33 |  | `633 | District | - Tiruppur |  |  | Urban |  |

|  |  |  |
| --- | --- | --- |
| 0 | Age group  Total | \ |
| 1 | `5-14 |  |
| 2 | 15-34 |  |
| 3 | 35-59 |  |
| 4 | 60+ |  |
| .. | … |  |
| 589 | `5-14 |  |
| 590 | 15-34 |  |
| 591 | 35-59 |  |
| 592 | 60+ |  |
| 593 | Age not stated |  |

Worked for 3 months or more but less than 6 months - Persons \ 0 1200828

1 27791

2 514340

3 542581

4 115103

.. …

589 272

590 3285

591 3672

592 696

593 2

Worked for 3 months or more but less than 6 months - Males \ 0 589003

1 14125

2 259560

3 251957

4 62833

|  |  |
| --- | --- |
| .. | … |
| 589 | 129 |
| 590 | 1654 |
| 591 | 1769 |
| 592 | 399 |
| 593 | 1 |

Worked for 3 months or more but less than 6 months - Females \ 0 611825

1 13666

2 254780

3 290624

4 52270

.. …

589 143

590 1631

591 1903

592 297

593 1

Worked for less than 3 months - Persons \ 0 221386

1 2447

2 92423

3 99202

4 27165

.. …

589 18

590 473

591 522

592 111

593 0

Worked for less than 3 months - Males … \ 0 99368 …

1 1247 …

2 43892 …

3 40691 …

4 13465 …

.. … …

589 6 …

590 238 …

591 247 …

592 50 …

593 0 …

Industrial Category - N to O - Females \

0 3565

1 11

2 1754

3 1619

4 175

.. …

589 0

590 20

591 33

592 0

593 0

Industrial Category - P to Q - Persons \ 0 11080

1 122

2 7536

3 3205

4 211

.. …

589 0

590 44

591 35

592 3

593 0

Industrial Category - P to Q - Males \ 0 4019

1 71

2 2718

3 1131

4 93

|  |  |
| --- | --- |
| .. … |  |
| 589 | 0 |
| 590 | 15 |
| 591 | 12 |
| 592 | 0 |
| 593 | 0 |

Industrial Category - P to Q - Females \ 0 7061

1 51

2 4818

3 2074

4 118

.. …

589 0

590 29

591 23

592 3

593 0

Industrial Category - R to U - HHI - Persons \ 0 16833

1 427

2 8346

3 6591

4 1457

.. …

589 0

590 62

591 36

592 10

593 0

Industrial Category - R to U - HHI - Males \ 0 4266

1 169

2 2127

3 1487

4 483

.. …

589 0

590 6

591 9

592 3

593 0

Industrial Category - R to U - HHI - Females \ 0 12567

1 258

2 6219

3 5104

4 974

.. …

589 0

590 56

591 27

592 7

593 0

Industrial Category - R to U - Non HHI - Persons \ 0 122088

1 19305

2 68929

|  |  |
| --- | --- |
| 3 | 26498 |
| 4 | 7065 |
| .. | … |
| 589 | 228 |
| 590 | 675 |
| 591 | 279 |
| 592 | 81 |
| 593 | 0 |

[27]:

df.isnull().sum()

Industrial Category - R to U - Non HHI - Males \ 0 55801

1 9774

2 32803

3 9675

4 3394

.. …

589 104

590 247

591 103

592 35

593 0

Industrial Category - R to U - Non HHI - Females 0 66287

1 9531

2 36126

3 16823

4 3671

.. …

|  |  |  |
| --- | --- | --- |
| 589 |  | 124 |
| 590 |  | 428 |
| 591 |  | 176 |
| 592 |  | 46 |
| 593 |  | 0 |
| [594 | rows x 68 columns] |  |

|  |  |  |
| --- | --- | --- |
| [27]: State Code |  | 0 |
| District Code |  | 0 |
| Area Name |  | 0 |
| Total/ Rural/ | Urban | 0 |
| Age group |  | 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: 68, dtype: int64

[29]:

df["Worked for less than 3 months - Persons"].describe(include=['object',␣

↪'bool'])

[29]: count 594.000000

mean 2981.629630

std 13909.621137

min 0.000000

25% 27.000000

50% 430.000000

75% 1775.250000

max 221386.000000

Name: Worked for less than 3 months - Persons, dtype: float64

[31]:

numerical\_cols = ['Industrial Category - R to U - HHI - Persons', 'Age group',␣

↪'Industrial Category - R to U - HHI - Males'] df[numerical\_cols].describe()

[31]: Industrial Category - R to U - HHI - Persons \ count 594.000000

mean 226.707071

std 1039.953069

min 0.000000

25% 0.000000

50% 27.000000

75% 126.750000

max 16833.000000

Industrial Category - R to U - HHI - Males count 594.000000

mean 57.454545

std 265.230865

min 0.000000

25% 0.000000

50% 7.500000

75% 32.000000

max 4266.000000

[ ]: