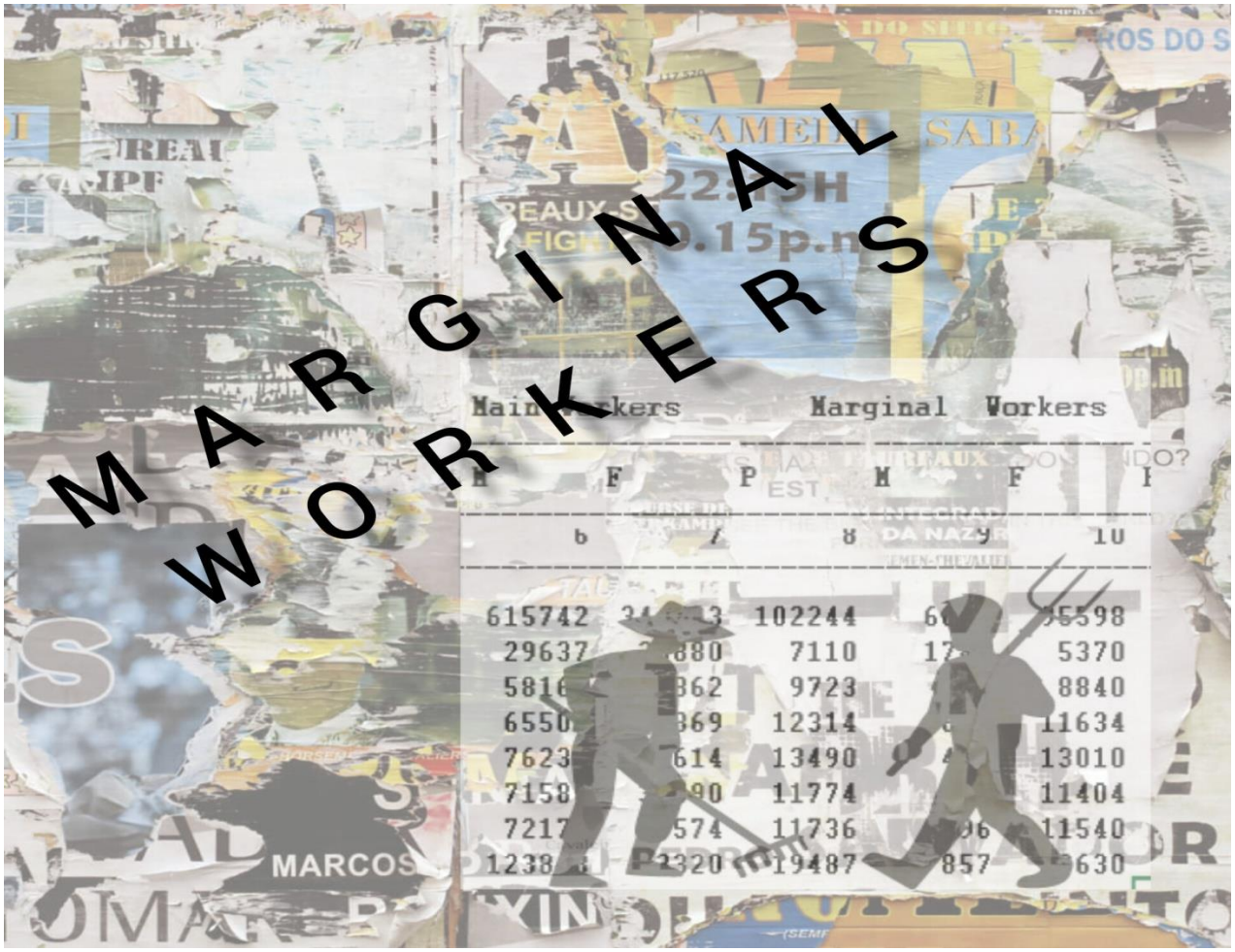
ASSESSMENT OF MARGINAL WORKERS IN TAMILNADU

PHASE3 PROJECT



INTRODUCTION:

- In this part will begin building project by loading and preprocessing the dataset.
- The data analysis by loading and preprocessing the dataset .
- Dataset can be loaded using Python and data manipulation library like pandas is used.

PREPROCESSING:

- Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.

IMPORT LIBRARIES:

- Begin by importing the necessary libraries, which typically include pandas for data manipulation and potentially other libraries for visualization and analysis.

LOAD THE DATASET:

- Load your dataset using pandas. The most common format is a CSV file, but pandas can handle various other formats, such as Excel, SQL databases, or even web APIs.

DATASET:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	Table Cod	State Code	District Co	Area Name	Total/Rur	Age group	Worked fo	Worked fo	Worked fo	Worked fo	Worked fo	Worked fo	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial	Industrial
2	B08065C	'33	'000	State - TAl	Total	Total	1200828	588003	618825	220386	93368	122018	64235	34632	29603	907752	404844	502908	29410	15269	13142	2853	1862	991	21486	7843	13643	62368	37520	24848	1903	1467
3	B08065C	'33	'000	State - TAl	Total	'5-14	27791	14125	13666	2447	1247	1200	1710	825	885	6398	3130	3268	190	107	83	9	9	0	182	67	115	818	383	435	3	3
4	B08065C	'33	'000	State - TAl	Total	'15-34	514340	259580	254760	32423	43892	48531	24863	12711	12152	345420	152368	192452	9430	5443	3987	1174	839	335	9583	3115	6468	40249	23897	16352	852	632
5	B08065C	'33	'000	State - TAl	Total	'35-59	542581	251957	290624	99202	40891	58311	29892	15927	13765	450052	192771	257281	15744	8230	7514	1436	860	576	9461	3526	5935	18976	11705	7271	926	698
6	B08065C	'33	'000	State - TAl	Total	'60+	115103	52833	62270	27165	13465	13700	7930	5151	2779	165325	55730	49595	4028	2470	1558	234	154	80	2249	1127	1122	2252	1490	762	125	71
7	B08065C	'33	'000	State - TAl	Total	Age not sta	1013	528	485	149	73	76	40	18	22	557	245	312	18	18	0	0	0	0	11	8	3	73	45	28	3	3
8	B08065C	'33	'000	State - TAl	Rural	Total	966645	458738	508307	174443	73653	100790	59637	32089	27448	824698	364131	460567	19758	11033	8725	1738	1181	557	15349	5526	9823	36310	21494	14816	959	715
9	B08065C	'33	'000	State - TAl	Rural	'5-14	17239	8713	8526	1977	985	992	1443	684	759	6005	2922	3083	144	80	64	6	6	0	142	47	95	551	250	301	0	0
10	B08065C	'33	'000	State - TAl	Rural	'15-34	406847	198575	208272	71674	31917	40057	22933	11766	11167	316885	138622	178263	6887	3905	2778	732	543	189	7023	2264	4759	24055	14506	10349	398	318
11	B08065C	'33	'000	State - TAl	Rural	'35-59	444800	195573	249227	77922	29888	48114	27799	14887	12912	406147	172178	233969	10307	5468	4839	882	567	315	6487	2344	4143	9547	5880	3767	434	371
12	B08065C	'33	'000	State - TAl	Rural	'60+	97011	52498	44513	22446	10902	11544	7425	4835	2590	95151	50192	44959	2618	1564	1044	108	75	33	1689	863	826	1211	760	451	37	27
13	B08065C	'33	'000	State - TAl	Rural	Age not sta	748	375	373	124	51	73	37	17	20	510	217	293	12	12	0	0	0	0	8	8	0	46	28	18	0	0
14	B08065C	'33	'000	State - TAl	Urban	Total	234183	129265	104918	46943	25705	21238	4598	2443	2155	83754	40713	43041	9652	5235	4417	1125	671	454	6157	2317	3820	26158	16116	9942	1850	751
15	B08065C	'33	'000	State - TAl	Urban	'5-14	10552	5412	5140	470	262	208	267	141	126	393	208	185	46	27	19	3	3	0	40	20	20	267	133	134	3	3
16	B08065C	'33	'000	State - TAl	Urban	'15-34	107493	60985	46508	20443	11975	8474	1930	945	985	28535	14346	14189	2743	1534	1209	442	296	146	2560	851	1709	15394	9391	6003	464	374
17	B08065C	'33	'000	State - TAl	Urban	'35-59	97781	52384	45397	20280	10883	10397	1893	1040	853	43805	20593	23212	5437	2762	2675	554	293	261	2974	1182	1792	9329	5845	3484	492	327
18	B08065C	'33	'000	State - TAl	Urban	'60+	18032	10335	7757	4719	2563	2156	505	316	189	10174	5538	4636	1420	906	514	126	79	47	580	264	296	1041	730	311	88	44
19	B08065C	'33	'000	State - TAl	Urban	Age not sta	265	149	116	25	22	3	3	1	2	47	28	19	6	6	0	0	0	0	3	0	3	27	17	10	3	3
20	B08065C	'33	'602	District - T	Total	Total	74448	39295	35153	15866	8004	7862	3066	1663	1403	42579	20345	22234	1519	1025	494	63	47	16	1529	755	774	8114	5484	2630	245	193
21	B08065C	'33	'602	District - T	Total	'5-14	2521	1284	1237	147	82	65	122	56	66	330	154	176	12	12	0	0	0	0	17	3	14	126	73	53	0	0
22	B08065C	'33	'602	District - T	Total	'15-34	33568	18049	15519	6529	3654	2875	1225	632	593	15591	7257	8334	570	397	183	27	21	6	649	337	312	5487	3630	1857	136	93
23	B08065C	'33	'602	District - T	Total	'35-59	32568	16771	15797	7718	3529	4189	1414	792	622	22192	10446	11746	788	532	256	36	26	10	699	320	379	2310	1637	673	103	70
24	B08065C	'33	'602	District - T	Total	'60+	5716	3147	2569	1465	739	726	305	183	122	4441	2476	1965	149	94	55	0	0	0	164	95	69	173	136	37	6	6
25	B08065C	'33	'602	District - T	Total	Age not sta	75	44	31	7	0	7	0	0	0	25	12	13	0	0	0	0	0	0	0	0	0	18	8	10	0	0
26	B08065C	'33	'602	District - T	Rural	Total	55577	28082	27495	12131	5653	6478	2804	1511	1293	39766	18880	20886	1262	852	410	38	26	12	1080	528	552	4748	3114	1634	160	98
27	B08065C	'33	'602	District - T	Rural	'5-14	1424	743	681	114	61	53	99	43	56	319	149	170	12	12	0	0	0	0	12	0	12	104	62	42	0	0
28	B08065C	'33	'602	District - T	Rural	'15-34	23985	12377	11608	4713	2443	2270	1109	566	543	14527	6700	7827	484	325	159	18	12	6	468	242	226	3398	2210	1188	66	52
29	B08065C	'33	'602	District - T	Rural	'35-59	25421	12417	13004	6104	2574	3530	1320	740	580	20723	9675	11048	650	444	206	20	14	6	456	202	254	1148	774	374	74	46
30	B08065C	'33	'602	District - T	Rural	'60+	4718	2516	2202	1193	575	618	276	162	114	4175	2327	1848	116	70	46	0	0	0	144	84	60	82	60	22	0	0
31	B08065C	'33	'602	District - T	Rural	Age not sta	49	29	20	7	0	7	0	0	0	22	9	13	0	0	0	0	0	0	0	0	0	16	8	8	0	0
32	B08065C	'33	'602	District - T	Urban	Total	18871	11213	7658	3735	2351	1384	262	152	110	2813	1485	1328	257	173	84	25	21	4	449	227	222	3366	2370	996	95	71
33	B08065C	'33	'602	District - T	Urban	'5-14	1097	541	556	33	21	12	23	13	10	11	5	6	0	0	0	0	0	0	5	3	2	22	11	11	0	0
34	B08065C	'33	'602	District - T	Urban	'15-34	9583	5672	3911	1816	1211	605	116	66	50	1064	557	507	66	61	25	9	9	0	181	95	86	2089	1420	669	50	41
35	B08065C	'33	'602	District - T	Urban	'35-59	7947	4354	2793	1614	955	659	94	52	42	1469	771	698	138	88	50	16	12	4	243	118	125	1162	863	299	29	24
36	B08065C	'33	'602	District - T	Urban	'60+	998	631	367	272	164	108	29	21	8	266	149	117	33	24	9	0	0	0	20	11	9	91	76	15	6	6
37	B08065C	'33	'602	District - T	Urban	Age not sta	26	15	11	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0
38	B08065C	'33	'603	District - C	Total	Total	33748	19313	14435	6351	3954	2397	740	327	413	819	513	306	141	114	27	22	17	5	470	233	237	3529	2565	964	138	112
39	B08065C	'33	'603	District - C	Total	'5-14	2749	1483	1266	140	74	66	91	43	48	25	19	6	0	0	0	0	0	0	3	3	0	40	15	25	0	0
40	B08065C	'33	'603	District - C	Total	'15-34	17431	9836	7595	3102	1981	1121	379	153	226	380	236	154	48	36	12	6	6	0	163	99	64	2028	1454	574	49	44

PROGRAM:

1) Display the first few rows of the dataset:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
print(df.head())
```

ouput:

```
   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
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]
```

2) Get basic information about the dataset:

```
import pandas as pd

csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'

df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')

print(df.info())
```

output:

```
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

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

memory usage: 320.3+ KB

None

3) Summary statistics for numerical columns:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
print(df.describe())
```

output:

```
Worked for 3 months or more but less than 6 months -  Persons  \
count                                   5.940000e+02
mean                                    1.617277e+04
std                                     7.607172e+04
min                                     0.000000e+00
25%                                     2.872500e+02
50%                                     2.225500e+03
75%                                     9.628500e+03
max                                     1.200828e+06

Worked for 3 months or more but less than 6 months - Males  \
count                                     594.000000
mean                                     7932.700337
std                                     36864.822704
min                                         0.000000
```

25%	147.250000
50%	1147.000000
75%	4770.500000
max	589003.000000

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

count	594.000000
mean	8240.067340
std	39259.545337
min	0.000000
25%	144.000000
50%	1076.000000
75%	4887.500000
max	611825.000000

Worked for less than 3 months - Persons \

count	594.000000
mean	2981.629630
std	13909.621137
min	0.000000
25%	27.000000
50%	430.000000
75%	1775.250000
max	221386.000000

Worked for less than 3 months - Males \

count	594.000000
mean	1338.289562
std	6127.047670
min	0.000000
25%	14.250000
50%	198.500000
75%	774.250000
max	99368.000000

Worked for less than 3 months - Females \

count	594.000000
mean	1643.340067
std	7808.832522
min	0.000000
25%	13.000000
50%	213.000000

```
75%                          946.500000
max                       122018.000000

 Industrial Category - A - Cultivators - Persons  \
count                          594.000000
mean                           865.117845
std                           4274.458077
min                              0.000000
25%                              9.000000
50%                             69.500000
75%                            466.000000
max                          64235.000000


Industrial Category - A - Cultivators - Males  \

count                          594.000000
mean                           466.424242
std                           2298.072295
min                              0.000000
25%                              5.000000
50%                             35.500000
75%                            244.250000
max                          34632.000000

 Industrial Category - A - Cultivators - Females  \
count                          594.000000
mean                           398.693603
std                           1978.682322
min                              0.000000
25%                              4.000000
50%                             32.000000
75%                            204.750000
max                          29603.000000
[8 rows x 63 columns]
```

4) Count unique values in each column:

```
import pandas as pd

csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'

df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
```

```
print(df.nunique())
```

Output:

```
Table Code                                  1
State Code                                  1
District Code                              33
Area Name                                  33
Total/ Rural/ Urban                         3
                                   ...
Industrial Category - R to U - HHI - Males        120
Industrial Category - R to U - HHI - Females      187
Industrial Category - R to U - Non HHI - Persons  397
Industrial Category - R to U - Non HHI - Males    314
Industrial Category - R to U - Non HHI - Females  342
Length: 69, dtype: int64
```

5) Count missing values in each column:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
print(df.isnull().sum())
```

output:

```
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) Delete duplicate values:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
df = df.drop_duplicates()
print('duplicated values are deleted')
```

output:

```
duplicated values are deleted
```

7) change column datatype:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
df['Table Code'] = df['Table Code'].astype('string')
print(‘datatype changed’)
```

output:

```
datatype changed
```

8)data preprocess:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
df.to_csv('preprocessed_data.csv', index=False)
print('data can be preprocessed')
```

output:

```
data can be preprocessed
```

9) Drop duplicate row:

```
import pandas as pd
csv_file_path = '/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv'
df = pd.read_csv('/content/DDW_B06SC_3300_State_TAMIL_NADU-2011.csv')
df = df.dropna()
print('dropped')
```

OUTPUT:

```
drooped
```