filter_census_2022

April 10, 2024

1 Preprocess original census data 2022

- Open original census data
- Extract all rows for maize
- Rename variables to english
- Save file as csv

[1]:
```
# Imports
import pandas as pd
from pathlib import Path
```

[2]:
```
# Paths
original_path = Path.cwd().parent / 'original_data'
original_path
```

[2]:
```
PosixPath('/home/vant/Documents/valencia/agml_workshop/inegi_censos/original_dat
a')
```

[3]:
```
# Replace 'file_path.xlsx' with the path to your Excel file
file_path = original_path/'ca2022_agr01.xlsx'

# Read the Excel file into a Pandas DataFrame
df = pd.read_excel(file_path,skiprows=4)
```

[4]:
```
# Now you can work with your DataFrame 'df'
# For example, you can print the first few rows:
df.head(3)
```

[4]:
```
  Entidad federativa Municipio Cultivo  \
0                NaN       NaN     NaN
1                NaN       NaN     NaN
2                NaN       NaN     NaN

  Entidad federativa, municipio y cultivo  \
0                                     NaN
1                                     NaN
2                                     NaN
```

```
  Unidades de producción agropecuaria activas  \
0                                         NaN
1                                       Total
2                                         NaN

                      Unnamed: 5 Superficie cultivada            Unnamed: 7  \
0                            NaN                  NaN                   NaN
1  Con agricultura a cielo abierto  Superficie sembrada  Superficie cosechada
2                            NaN                  NaN                   NaN

  Producción       Modalidad hídrica          Unnamed: 10  \
0        NaN                Temporal                  NaN
1        NaN  Unidades de producción  Superficie sembrada
2        NaN                     NaN                  NaN

           Unnamed: 11 Unnamed: 12             Unnamed: 13  \
0                  NaN         NaN                   Riego
1  Superficie cosechada  Producción  Unidades de producción
2                  NaN         NaN                     NaN

          Unnamed: 14           Unnamed: 15 Unnamed: 16
0                 NaN                   NaN         NaN
1  Superficie sembrada  Superficie cosechada  Producción
2                 NaN                   NaN         NaN
```

[5]:
```
df.iloc[0:5,0:6]
```

[5]:
```
  Entidad federativa Municipio Cultivo  \
0                NaN       NaN     NaN
1                NaN       NaN     NaN
2                NaN       NaN     NaN
3                NaN       NaN     NaN
4                NaN       NaN     NaN

  Entidad federativa, municipio y cultivo  \
0                                     NaN
1                                     NaN
2                                     NaN
3                                     NaN
4                                     NaN

  Unidades de producción agropecuaria activas                       Unnamed: 5
0                                         NaN                              NaN
1                                       Total  Con agricultura a cielo abierto
2                                         NaN                              NaN
3                                         NaN                              NaN
4                                           A                             B<=A
```

[6]:
```
df.columns
```

[6]:
```
Index(['Entidad federativa', 'Municipio', 'Cultivo',
       'Entidad federativa, municipio y cultivo',
       'Unidades de producción agropecuaria activas', 'Unnamed: 5',
       'Superficie cultivada', 'Unnamed: 7', 'Producción', 'Modalidad hídrica',
       'Unnamed: 10', 'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13',
       'Unnamed: 14', 'Unnamed: 15', 'Unnamed: 16'],
      dtype='object')
```

[7]:
```
# Define new column names
column_names = ['Entidad federativa', 'Municipio', 'Cultivo',
       'Entidad federativa, municipio y cultivo',
       'Unidades de producción agropecuaria activas-Total', 'Unidades de␣
 ↪producción agropecuaria activas-Con agricultura a cielo abierto',
       'Superficie cultivada-sembrada', 'Superficie cultivada-cosechada',␣
 ↪'Producción', 'MH-temporal-unidad-produccion',
       'MH-temporal-superficie-sembrada', 'MH-temporal-superficie-cosechada',␣
 ↪'MH-temporal-produccion', 'MH-riego-unidad-produccion',
       'MH-riego-superficie-sembrada', 'MH-riego-superficie-cosechada',␣
 ↪'MH-riego-produccion']

# Rename the columns
df.columns = column_names

# Delete rows with no data
df = df.drop(index=range(5)).reset_index(drop=True)
```

[8]:
```
df.head(5)
```

[8]:

	Entidad federativa	Municipio	Cultivo	Entidad federativa, municipio y cultivo	Unidades de producción agropecuaria activas-Total
0	00 NAL	NaN	NaN	Estados Unidos Mexicanos	4629134
1	00 NAL	NaN	Anuales	Anuales	NaN
2	00 NAL	NaN	Algodón	Algodón	
3	00 NAL	NaN	Amaranto	Amaranto	
4	00 NAL	NaN	Arroz	Arroz	

```
2                                                 4666
3                                                 4532
4                                                 2952

  Unidades de producción agropecuaria activas-Con agricultura a cielo abierto  \
0                                            4366995
1                                                NaN
2                                               4665
3                                               4461
4                                               2950

  Superficie cultivada-sembrada Superficie cultivada-cosechada   Producción  \
0                 22937029.2021                  18864519.5362          NaN
1                           NaN                            NaN          NaN
2                  169271.2457                   157702.2879  686092.3256
3                    4933.7853                     4659.1731    6265.1429
4                   31854.1177                    30417.4643  172702.6519

  MH-temporal-unidad-produccion MH-temporal-superficie-sembrada  \
0                       3673541                  16788300.4445
1                           NaN                            NaN
2                            75                      1075.7468
3                          4119                      4615.0523
4                          1393                     15804.7797

  MH-temporal-superficie-cosechada MH-temporal-produccion  \
0                    13081544.6158                    NaN
1                              NaN                    NaN
2                       1040.0138               2089.543
3                       4349.2943              5690.5142
4                       14862.1306             71087.1147

  MH-riego-unidad-produccion MH-riego-superficie-sembrada  \
0                     831305                 6148728.7576
1                        NaN                          NaN
2                       4592                168195.4989
3                        366                    318.733
4                       1565                   16049.338

  MH-riego-superficie-cosechada MH-riego-produccion
0                  5782974.9204                NaN
1                           NaN                NaN
2                   156662.2741        684002.7826
3                      309.8788           574.6287
4                   15555.3337        101615.5372
```

[9]:
```
# Drop the column at index 3
df_clean = df.drop("Entidad federativa, municipio y cultivo", axis=1)

# Display the modified DataFrame
df_clean.head()
```

[9]:

	Entidad federativa	Municipio	Cultivo	Unidades de producción agropecuaria activas-Total	Unidades de producción agropecuaria activas-Con agricultura a cielo abierto	Superficie cultivada-sembrada	Superficie cultivada-cosechada	Producción	MH-temporal-unidad-produccion	MH-temporal-superficie-sembrada	MH-temporal-superficie-cosechada	MH-temporal-produccion
0	00 NAL	NaN	NaN	4629134	4366995	22937029.2021	18864519.5362	NaN	3673541	16788300.4445	13081544.6158	NaN
1	00 NAL	NaN	Anuales	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	00 NAL	NaN	Algodón	4666	4665	169271.2457	157702.2879	686092.3256	75	1075.7468	1040.0138	2089.543
3	00 NAL	NaN	Amaranto	4532	4461	4933.7853	4659.1731	6265.1429	4119	4615.0523	4349.2943	5690.5142
4	00 NAL	NaN	Arroz	2952	2950	31854.1177	30417.4643	172702.6519	1393	15804.7797	14862.1306	71087.1147

```
  MH-riego-unidad-produccion MH-riego-superficie-sembrada  \
0                     831305                  6148728.7576
1                        NaN                           NaN
2                       4592                   168195.4989
3                        366                      318.733
4                       1565                    16049.338

  MH-riego-superficie-cosechada MH-riego-produccion
0                  5782974.9204                 NaN
1                           NaN                 NaN
2                   156662.2741         684002.7826
3                      309.8788            574.6287
4                   15555.3337         101615.5372
```

[ ]:

[10]:
```
# Filter records for maize (in spanish maíz or Maíz)
# obtener todos los registros con cultivo que contenga: Maíz forrajero Maíz␣
 ↪grano amarillo Maíz grano blanco
# Filter the DataFrame based on the condition
df_clean.dropna(subset=['Cultivo'], inplace=True)
df_clean.dropna(subset=['Municipio'], inplace=True)
maiz_df = df_clean[df_clean['Cultivo'].str.contains('Maíz', case=False)]
maiz_df.head(20)
```

[10]:

	Entidad federativa	Municipio	Cultivo
93	01 AGS	001 Aguascalientes	Maíz forrajero
94	01 AGS	001 Aguascalientes	Maíz grano amarillo
95	01 AGS	001 Aguascalientes	Maíz grano blanco
122	01 AGS	002 Asientos	Maíz forrajero
123	01 AGS	002 Asientos	Maíz grano amarillo
124	01 AGS	002 Asientos	Maíz grano blanco
148	01 AGS	003 Calvillo	Maíz forrajero
149	01 AGS	003 Calvillo	Maíz grano amarillo
150	01 AGS	003 Calvillo	Maíz grano blanco
175	01 AGS	004 Cosío	Maíz forrajero
176	01 AGS	004 Cosío	Maíz grano amarillo
177	01 AGS	004 Cosío	Maíz grano blanco
198	01 AGS	005 Jesús María	Maíz forrajero
199	01 AGS	005 Jesús María	Maíz grano amarillo
200	01 AGS	005 Jesús María	Maíz grano blanco
224	01 AGS	006 Pabellón de Arteaga	Maíz forrajero
225	01 AGS	006 Pabellón de Arteaga	Maíz grano amarillo
226	01 AGS	006 Pabellón de Arteaga	Maíz grano blanco
252	01 AGS	007 Rincón de Romos	Maíz forrajero
253	01 AGS	007 Rincón de Romos	Maíz grano amarillo

```
    Unidades de producción agropecuaria activas-Total  \
93                                                1591
94                                                  29
95                                                 802
122                                                690
123                                                 52
124                                               2053
148                                                262
149                                                 15
150                                                836
175                                                337
176                                                 10
177                                                421
198                                                347
199                                                 11
200                                                649
224                                                525
225                                                 10
226                                                250
252                                                893
253                                                 11

    Unidades de producción agropecuaria activas-Con agricultura a cielo abierto
\
93                                                1591
94                                                  29
95                                                 802
122                                                690
123                                                 52
124                                               2053
148                                                262
149                                                 15
150                                                836
175                                                337
176                                                 10
177                                                421
198                                                347
199                                                 11
200                                                649
224                                                525
225                                                 10
226                                                250
252                                                893
253                                                 11

    Superficie cultivada-sembrada Superficie cultivada-cosechada   Producción  \
```

93	13719.6093	13083.3126	339873.3523
94	233.835	230.8305	997.6648
95	6019.7733	5826.5034	22008.7552
122	4587.2771	4320.7626	152522.1552
123	177.805	144.952	506.9692
124	9618.8066	8185.7351	13855.6288
148	972.6302	857.1883	14532.262
149	47.44	45.44	196.4423
150	2016.9773	1797.7592	4869.0259
175	1450.8873	1418.2261	60974.8832
176	31.75	31.7374	302.2972
177	1551.1395	1402.3324	5745.0826
198	1965.3007	1856.2078	63158.2197
199	32.88	29.38	152.0192
200	2075.3256	1889.8108	6092.0952
224	4137.6417	3991.5303	188613.2108
225	40.55	38.55	288.1506
226	1081.8235	1001.0702	2304.0697
252	6277.7995	6052.7947	264204.8649
253	47.75	47.3622	410.8173

	MH-temporal-unidad-produccion	MH-temporal-superficie-sembrada	\
93	1257	9406.1929	
94	29	233.835	
95	754	5556.7985	
122	392	2152.3979	
123	47	157.805	
124	1777	8322.2844	
148	209	784.7755	
149	12	42.19	
150	672	1668.2322	
175	154	458.1881	
176	3	3.29	
177	216	733.1451	
198	207	1026.7239	
199	7	25.7	
200	450	1616.8298	
224	199	939.6671	
225	3	14.27	
226	194	826.0055	
252	394	2205.501	
253	5	14.95	

	MH-temporal-superficie-cosechada	MH-temporal-produccion	\
93	8855.2401	104486.6145	
94	230.8305	997.6648	
95	5364.0286	18359.1613	

122	1961.1395	14881.8086
123	125.952	363.4392
124	6922.3776	6065.2098
148	678.2547	4719.2554
149	40.19	150.4148
150	1449.2141	2707.3644
175	427.9796	3637.1987
176	3.29	12.0743
177	601.0542	746.6715
198	929.9297	9102.5794
199	23.7	101.097
200	1435.6928	3143.8161
224	803.9485	7391.4385
225	12.27	20.35
226	749.2502	830.3591
252	1990.4347	15929.072
253	14.5635	61.3493

	MH-riego-unidad-produccion	MH-riego-superficie-sembrada	\
93	380	4313.4164	
94	0	0	
95	55	462.9748	
122	348	2434.8792	
123	5	20	
124	312	1296.5222	
148	59	187.8547	
149	3	5.25	
150	182	348.7451	
175	194	992.6992	
176	7	28.46	
177	219	817.9944	
198	153	938.5768	
199	4	7.18	
200	212	458.4958	
224	371	3197.9746	
225	7	26.28	
226	75	255.818	
252	549	4072.2985	
253	7	32.8	

	MH-riego-superficie-cosechada	MH-riego-produccion
93	4228.0725	235386.7378
94	0	0
95	462.4748	3649.5939
122	2359.6231	137640.3466
123	19	143.53
124	1263.3575	7790.419

```
148                    178.9336          9813.0066
149                        5.25            46.0275
150                    348.5451          2161.6615
175                    990.2465         57337.6845
176                     28.4474           290.2229
177                    801.2782          4998.4111
198                    926.2781         54055.6403
199                        5.68            50.9222
200                     454.118          2948.2791
224                   3187.5818        181221.7723
225                       26.28            267.8006
226                      251.82           1473.7106
252                     4062.36         248275.7929
253                     32.7987            349.468
```

[11]:
```
maiz_df.shape
```

[11]:
```
(6433, 16)
```

[12]:
```
maiz_df.columns
```

[12]:
```
Index(['Entidad federativa', 'Municipio', 'Cultivo',
       'Unidades de producción agropecuaria activas-Total',
       'Unidades de producción agropecuaria activas-Con agricultura a cielo
abierto',
       'Superficie cultivada-sembrada', 'Superficie cultivada-cosechada',
       'Producción', 'MH-temporal-unidad-produccion',
       'MH-temporal-superficie-sembrada', 'MH-temporal-superficie-cosechada',
       'MH-temporal-produccion', 'MH-riego-unidad-produccion',
       'MH-riego-superficie-sembrada', 'MH-riego-superficie-cosechada',
       'MH-riego-produccion'],
      dtype='object')
```

[13]:
```
english_col_names = ['State', 'Municipality', 'Crop',
 'Active agricultural production units - Total',
 'Active agricultural production units - With open agriculture',
 'Cultivated area - Sown',
 'Cultivated area - Harvested',
 'Production',
 'Water Modality - Temporary - Production unit',
 'Water Modality - Temporary - Cultivated area - Sown',
 'Water Modality - Temporary - Cultivated area - Harvested',
 'Water Modality - Temporary - Production',
 'Water Modality - Irrigation - Production unit',
 'Water Modality - Irrigation - Cultivated area - Sown',
 'Water Modality - Irrigation - Cultivated area - Harvested',
 'Water Modality - Irrigation - Production']
```

[14]:
```
maiz_df.columns = english_col_names

maiz_df.head(5)
```

[14]:
```
      State        Municipality                 Crop  \
93   01 AGS  001 Aguascalientes       Maíz forrajero
94   01 AGS  001 Aguascalientes  Maíz grano amarillo
95   01 AGS  001 Aguascalientes    Maíz grano blanco
122  01 AGS        002 Asientos       Maíz forrajero
123  01 AGS        002 Asientos  Maíz grano amarillo

    Active agricultural production units - Total  \
93                                          1591
94                                            29
95                                           802
122                                          690
123                                           52

    Active agricultural production units - With open agriculture  \
93                                                  1591
94                                                    29
95                                                   802
122                                                  690
123                                                   52

    Cultivated area - Sown Cultivated area - Harvested   Production  \
93              13719.6093                  13083.3126  339873.3523
94                 233.835                    230.8305     997.6648
95              6019.7733                   5826.5034   22008.7552
122             4587.2771                   4320.7626  152522.1552
123                177.805                     144.952     506.9692

    Water Modality - Temporary - Production unit  \
93                                          1257
94                                            29
95                                           754
122                                          392
123                                           47

    Water Modality - Temporary - Cultivated area - Sown  \
93                                           9406.1929
94                                             233.835
95                                           5556.7985
122                                          2152.3979
123                                            157.805

    Water Modality - Temporary - Cultivated area - Harvested  \
```

```
93                                          8855.2401
94                                           230.8305
95                                          5364.0286
122                                         1961.1395
123                                           125.952

    Water Modality - Temporary - Production  \
93                               104486.6145
94                                  997.6648
95                                18359.1613
122                               14881.8086
123                                 363.4392

    Water Modality - Irrigation - Production unit  \
93                                           380
94                                             0
95                                            55
122                                          348
123                                            5

    Water Modality - Irrigation - Cultivated area - Sown  \
93                                          4313.4164
94                                                 0
95                                           462.9748
122                                         2434.8792
123                                                20

    Water Modality - Irrigation - Cultivated area - Harvested  \
93                                          4228.0725
94                                                 0
95                                           462.4748
122                                         2359.6231
123                                                19

    Water Modality - Irrigation - Production
93                               235386.7378
94                                         0
95                                3649.5939
122                             137640.3466
123                                  143.53
```

[15]:
```python
# translate to English crop names
# Define translations
translations = {
    'Maíz forrajero': 'Forage corn',
    'Maíz grano amarillo': 'Yellow corn',
    'Maíz grano blanco': 'White corn'
```

```
}

# Replace the values in the "Cultivo" column with their English translations
maiz_df.loc[:, "Crop"] = maiz_df["Crop"].replace(translations)

# Display the modified DataFrame
maiz_df.head()
```

[15]:

	State	Municipality	Crop	Active agricultural production units - Total	Active agricultural production units - With open agriculture	Cultivated area - Sown	Cultivated area - Harvested	Production	Water Modality - Temporary - Production unit	Water Modality - Temporary - Cultivated area - Sown
93	01 AGS	001 Aguascalientes	Forage corn	1591	1591	13719.6093	13083.3126	339873.3523	1257	9406.1929
94	01 AGS	001 Aguascalientes	Yellow corn	29	29	233.835	230.8305	997.6648	29	233.835
95	01 AGS	001 Aguascalientes	White corn	802	802	6019.7733	5826.5034	22008.7552	754	5556.7985
122	01 AGS	002 Asientos	Forage corn	690	690	4587.2771	4320.7626	152522.1552	392	
123	01 AGS	002 Asientos	Yellow corn	52	52	177.805	144.952	506.9692	47	

```
122                                          2152.3979
123                                            157.805

    Water Modality - Temporary - Cultivated area - Harvested  \
93                                           8855.2401
94                                            230.8305
95                                           5364.0286
122                                          1961.1395
123                                            125.952

    Water Modality - Temporary - Production  \
93                              104486.6145
94                                 997.6648
95                                18359.1613
122                               14881.8086
123                                 363.4392

    Water Modality - Irrigation - Production unit  \
93                                            380
94                                              0
95                                             55
122                                           348
123                                             5

    Water Modality - Irrigation - Cultivated area - Sown  \
93                                           4313.4164
94                                                  0
95                                            462.9748
122                                          2434.8792
123                                                20

    Water Modality - Irrigation - Cultivated area - Harvested  \
93                                           4228.0725
94                                                  0
95                                            462.4748
122                                          2359.6231
123                                                19

    Water Modality - Irrigation - Production
93                              235386.7378
94                                       0
95                                3649.5939
122                             137640.3466
123                                  143.53
```

[16]:
```
# Define metadata
metadata = {
```

```
    "source":"INEGI Censo Agropecuario 2022",
    "Production": "tonnes",
    "Areas": "hectares"
}

# Store metadata in attributes or dictionaries
maiz_df.attrs['metadata'] = metadata

# Display the modified DataFrame
maiz_df.attrs
```

[16]:
```
{'metadata': {'source': 'INEGI Censo Agropecuario 2022',
  'Production': 'tonnes',
  'Areas': 'hectares'}}
```

[27]:
```
# Saving data
# Save DataFrame to CSV
maiz_df.to_csv('maize_data_2022.csv')

# Save metadata to a separate file (e.g., JSON)
import json
with open('maize_metadata_2022.json', 'w') as file:
    json.dump(metadata, file)
```

[28]:
```
#Check saved data
# Load DataFrame from CSV
maiz_df2 = pd.read_csv('maize_data_2022.csv', index_col=0)

# Load metadata from JSON
with open('maize_metadata_2022.json', 'r') as file:
    metadata = json.load(file)

# Assign metadata back to the DataFrame
maiz_df2.attrs['metadata'] = metadata

#maiz_df2.attrs
maiz_df2.head()
```

[28]:

	State	Municipality	Crop	\
93	01 AGS	001 Aguascalientes	Forage corn	
94	01 AGS	001 Aguascalientes	Yellow corn	
95	01 AGS	001 Aguascalientes	White corn	
122	01 AGS	002 Asientos	Forage corn	
123	01 AGS	002 Asientos	Yellow corn	

	Active agricultural production units - Total	\
93	1591	

```
94                                                29
95                                               802
122                                              690
123                                               52

    Active agricultural production units - With open agriculture  \
93                                                  1591
94                                                    29
95                                                   802
122                                                  690
123                                                   52

     Cultivated area - Sown  Cultivated area - Harvested   Production  \
93               13719.6093                   13083.3126  339873.3523
94                 233.8350                     230.8305     997.6648
95                6019.7733                    5826.5034   22008.7552
122               4587.2771                    4320.7626  152522.1552
123                177.8050                     144.9520     506.9692

    Water Modality - Temporary - Production unit  \
93                                          1257
94                                            29
95                                           754
122                                          392
123                                           47

     Water Modality - Temporary - Cultivated area - Sown  \
93                                          9406.1929
94                                           233.8350
95                                          5556.7985
122                                         2152.3979
123                                          157.8050

     Water Modality - Temporary - Cultivated area - Harvested  \
93                                          8855.2401
94                                           230.8305
95                                          5364.0286
122                                         1961.1395
123                                          125.9520

     Water Modality - Temporary - Production  \
93                              104486.6145
94                                997.6648
95                               18359.1613
122                              14881.8086
123                                363.4392
```

```
    Water Modality - Irrigation - Production unit  \
93                                            380
94                                              0
95                                             55
122                                           348
123                                             5

     Water Modality - Irrigation - Cultivated area - Sown  \
93                                          4313.4164
94                                             0.0000
95                                           462.9748
122                                         2434.8792
123                                           20.0000

     Water Modality - Irrigation - Cultivated area - Harvested  \
93                                          4228.0725
94                                             0.0000
95                                           462.4748
122                                         2359.6231
123                                           19.0000

     Water Modality - Irrigation - Production
93                                235386.7378
94                                     0.0000
95                                  3649.5939
122                               137640.3466
123                                  143.5300
```

[ ]: