filter_census_2017

April 11, 2024

1 Preprocess original census data 2017

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

[76]:
```
# Imports
import pandas as pd
from pathlib import Path
```

[77]:
```
# Paths
original_path = Path.cwd().parent / 'original_data'
original_path
```

[77]:
```
PosixPath('/home/vant/Documents/valencia/agml_workshop/inegi_censos/original_dat
a')
```

[78]:
```
# Replace 'file_path.xlsx' with the path to your Excel file
file_path = original_path/'ena17_ent_agri03.xlsx'

# Read the Excel file into a Pandas DataFrame
df = pd.read_excel(file_path,skiprows=5)
```

[79]:
```
df.head()
```

[79]:
```
  Entidad Cultivo Entidad federativa y cultivo Superficie cultivada  \
0     NaN     NaN                          NaN                  NaN
1     NaN     NaN                          NaN  Superficie sembrada
2     NaN     NaN                          NaN                  NaN
3     NaN     NaN                          NaN           Hectáreas
4     NaN     NaN                          NaN           A = D + G

             Unnamed: 4 Producción Disponibilidad del agua  \
0                   NaN        NaN                Temporal
1  Superficie cosechada        NaN    Superficie sembrada
2                   NaN        NaN                    NaN
3                   NaN  Toneladas              Hectáreas
```

```
4             B = E + H  C = F + I                         D

             Unnamed: 7  Unnamed: 8           Unnamed: 9  \
0                   NaN         NaN                Riego
1  Superficie cosechada  Producción  Superficie sembrada
2                   NaN         NaN                  NaN
3                   NaN   Toneladas           Hectáreas
4                     E           F                   G

            Unnamed: 10 Unnamed: 11
0                   NaN         NaN
1  Superficie cosechada  Producción
2                   NaN         NaN
3                   NaN   Toneladas
4                     H           I
```

[80]:
```
df.columns
```

[80]:
```
Index(['Entidad', 'Cultivo', 'Entidad federativa y cultivo',
       'Superficie cultivada', 'Unnamed: 4', 'Producción',
       'Disponibilidad del agua', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9',
       'Unnamed: 10', 'Unnamed: 11'],
      dtype='object')
```

[81]:
```
# Define new column names
column_names = ['Entidad federativa', 'Cultivo','Entidad federativa y␣
 ↪cultivo','Total superficie sembrada','Total superficie cosechada',
                'Producción total','MH-temporal superficie␣
 ↪sembrada','MH-temporal superficie cosechada','MH-temporal producción',
                'MH-riego superficie sembrada','MH-riego superficie␣
 ↪cosechada','MH-riego producción']

# Rename the columns
df.columns = column_names

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

[82]:
```
# Extract key:value for state code
# Filter rows when Cultivo es NaN
df.dropna(subset=['Entidad federativa', 'Entidad federativa y cultivo'],␣
 ↪inplace=True)
estado_codigo_nombre = df.loc[df['Cultivo'].isna(), ['Entidad federativa',␣
 ↪'Entidad federativa y cultivo']]
estado_codigo_nombre = estado_codigo_nombre[~estado_codigo_nombre['Entidad␣
 ↪federativa y cultivo'].isin(["Perennes", "Anuales"])]
```

```
estado_codigo_nombre.columns = ["codigo","nombre"]
# Mostrar el nuevo DataFrame
print(estado_codigo_nombre)
```

```
     codigo                          nombre
0    01 Ags                     Aguascalientes
6     02 BC                   Baja California
9    03 BCS               Baja California Sur
16   05 Coa             Coahuila de Zaragoza
21   06 Col                           Colima
28   07 Chs                          Chiapas
33   08 Chi                        Chihuahua
39   09 CMX                  Ciudad de México
42   10 Dgo                          Durango
46   11 Gto                       Guanajuato
54   12 Gro                         Guerrero
59   13 Hgo                          Hidalgo
62   14 Jal                          Jalisco
70   15 Mex                           México
73   16 Mic               Michoacán de Ocampo
80   17 Mor                          Morelos
87   18 Nay                          Nayarit
96   19 Nln                       Nuevo León
102  20 Oax                           Oaxaca
110  21 Pue                           Puebla
115  22 Qro                        Querétaro
118  23 Qtr                     Quintana Roo
121  24 SLP                   San Luis Potosí
128  25 Sin                          Sinaloa
133  26 Son                           Sonora
138  27 Tab                          Tabasco
142  28 Tam                       Tamaulipas
149  29 Tla                         Tlaxcala
152  30 Ver  Veracruz de Ignacio de la Llave
163  31 Yuc                          Yucatán
166  32 Zac                        Zacatecas
```

[83]:
```
df.dropna(subset=['Cultivo'], inplace=True)
# Drop the column at index 3
df_clean = df.drop("Entidad federativa y cultivo", axis=1)
df_clean.head()
```

[83]:
```
   Entidad federativa      Cultivo Total superficie sembrada  \
2              01 Ags       Frijol                3358.74626
3              01 Ags  Maíz blanco               73393.422331
5              01 Ags      Guayaba                3887.692187
8               02 BC      Cebolla                  2873.0976
```

```
11             03 BCS     Jitomate                1195.71095

    Total superficie cosechada Producción total  \
2                  3201.45101      2290.601174
3                 67484.245731    201502.371934
5                  3520.011687     33612.830537
8                    2850.1976      95135.28156
11                  1172.96095     82809.648064

    MH-temporal superficie sembrada MH-temporal superficie cosechada  \
2                       2409.01026                       2251.71501
3                     61353.990581                     55444.813981
5                                0                                0
8                                0                                0
11                               0                                0

    MH-temporal producción MH-riego superficie sembrada  \
2              1120.357124                      949.736
3             82382.615314                  12039.43175
5                        0                  3887.692187
8                        0                    2873.0976
11                       0                   1195.71095

    MH-riego superficie cosechada MH-riego producción
2                         949.736          1170.24405
3                     12039.43175         119119.75662
5                     3520.011687         33612.830537
8                       2850.1976          95135.28156
11                     1172.96095         82809.648064
```

[84]:
```
# Extract only maize
maiz_df = df_clean[df_clean['Cultivo'].str.contains('Maíz', case=False)]
maiz_df.head(10)
```

[84]:
```
    Entidad federativa        Cultivo Total superficie sembrada  \
3              01 Ags    Maíz blanco               73393.422331
12             03 BCS    Maíz blanco                 4644.99695
18             05 Coa    Maíz blanco               68820.574663
24             06 Col    Maíz blanco               16200.289808
30             07 Chs    Maíz blanco             519026.048064
36             08 Chi  Maíz amarillo             250020.277205
41             09 CMX    Maíz blanco                5238.283318
45             10 Dgo    Maíz blanco             139503.011051
49             11 Gto    Maíz blanco             467291.710938
56             12 Gro    Maíz blanco             423698.821003

    Total superficie cosechada Producción total  \
```

3	67484.245731	201502.371934
12	4582.99695	33856.84485
18	37634.2971	NaN
24	16034.177048	54795.656418
30	512396.603064	1622396.828905
36	247424.592595	2339662.510603
41	4448.172839	11030.413804
45	137311.602671	503828.301416
49	438044.267889	2321067.101112
56	384094.492327	887247.840106

	MH-temporal superficie sembrada	MH-temporal superficie cosechada	\
3	61353.990581	55444.813981	
12	182.8334	173.8334	
18	52657.822338	24258.235475	
24	13304.88757	13138.77481	
30	494760.214889	488161.129639	
36	50425.419025	48125.734415	
41	5236.283318	4446.172839	
45	62508.806981	62302.706981	
49	248689.862194	221590.464553	
56	391589.90499	352491.815905	

	MH-temporal producción	MH-riego superficie sembrada	\
3	82382.615314	12039.43175	
12	93.83345	4462.16355	
18	NaN	16162.752325	
24	43156.571618	2895.402238	
30	1488474.65536	24265.833175	
36	264885.336738	199594.85818	
41	10996.413804	2	
45	93598.693845	76994.20407	
49	387168.200662	218601.848743	
56	798942.489012	32108.916013	

	MH-riego superficie cosechada	MH-riego producción
3	12039.43175	119119.75662
12	4409.16355	33763.0114
18	13376.061625	NaN
24	2895.402238	11639.0848
30	24235.473425	133922.173544
36	199298.85818	2074777.173865
41	2	34
45	75008.89569	410229.607571
49	216453.803336	1933898.900449
56	31602.676422	88305.351094

[85]:
```
maiz_df.shape
```

[85]:
```
(27, 11)
```

[86]:
```
# replace Entidad federativa codes for state names
maiz_df.loc[:,'Entidad federativa'] = maiz_df['Entidad federativa'].
  ↪map(estado_codigo_nombre.set_index('codigo')['nombre'])
maiz_df.head()
```

[86]:
```
      Entidad federativa      Cultivo Total superficie sembrada  \
3          Aguascalientes  Maíz blanco              73393.422331
12   Baja California Sur  Maíz blanco                4644.99695
18  Coahuila de Zaragoza  Maíz blanco              68820.574663
24                Colima  Maíz blanco              16200.289808
30               Chiapas  Maíz blanco             519026.048064

   Total superficie cosechada Producción total  \
3                67484.245731    201502.371934
12                 4582.99695      33856.84485
18                 37634.2971              NaN
24                16034.177048     54795.656418
30               512396.603064   1622396.828905

   MH-temporal superficie sembrada MH-temporal superficie cosechada  \
3                     61353.990581                      55444.813981
12                        182.8334                         173.8334
18                     52657.822338                     24258.235475
24                      13304.88757                      13138.77481
30                    494760.214889                    488161.129639

   MH-temporal producción MH-riego superficie sembrada  \
3             82382.615314                  12039.43175
12                93.83345                   4462.16355
18                     NaN                 16162.752325
24            43156.571618                  2895.402238
30           1488474.65536                 24265.833175

   MH-riego superficie cosechada MH-riego producción
3                    12039.43175        119119.75662
12                    4409.16355          33763.0114
18                  13376.061625                 NaN
24                   2895.402238          11639.0848
30                  24235.473425       133922.173544
```

[87]:
```
# translate colnames to english
english_col_names = ['State',
                     'Crop',
```

```
                    'Total Cultivated area - Sown',
                    'Total Cultivated area - Harvested',
                    'Total production',
                    'Water Modality - Temporary - Cultivated area - Sown',
                    'Water Modality - Temporary - Cultivated area - Harvested',
                    'Water Modality - Temporary - Production',
                    'Water Modality - Irrigation - Cultivated area - Sown',
                    'Water Modality - Irrigation - Cultivated area -␣
 ↪Harvested',
                    'Water Modality - Irrigation - Production']

maiz_df.columns = english_col_names

# translate to English crop names
# Define translations
translations = {
    'Maíz forrajero': 'Forage corn',
    'Maíz amarillo': 'Yellow corn',
    'Maíz blanco': 'White corn'
}

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

maiz_df.head(5)
```

[87]:

	State	Crop	Total Cultivated area - Sown	Total Cultivated area - Harvested	Total production	Water Modality - Temporary - Cultivated area - Sown	Water Modality - Temporary - Cultivated area - Harvested	Water Modality - Temporary - Production	Water Modality - Irrigation - Cultivated area - Sown	Water Modality - Irrigation - Cultivated area - Harvested	Water Modality - Irrigation - Production
3	Aguascalientes	White corn	73393.422331	67484.245731	201502.371934	61353.990581					
12	Baja California Sur	White corn	4644.99695	4582.99695	33856.84485	182.8334					
18	Coahuila de Zaragoza	White corn	68820.574663	37634.2971	NaN	52657.822338					
24	Colima	White corn	16200.289808	16034.177048	54795.656418	13304.88757					
30	Chiapas	White corn	519026.048064	512396.603064	1622396.828905	494760.214889					

```
   Water Modality - Temporary - Cultivated area - Harvested  \
3                                          55444.813981
12                                             173.8334
18                                         24258.235475
24                                          13138.77481
30                                        488161.129639

   Water Modality - Temporary - Production  \
3                             82382.615314
12                                93.83345
18                                     NaN
24                            43156.571618
30                           1488474.65536

   Water Modality - Irrigation - Cultivated area - Sown  \
3                                          12039.43175
12                                          4462.16355
18                                         16162.752325
24                                          2895.402238
30                                         24265.833175

   Water Modality - Irrigation - Cultivated area - Harvested  \
3                                          12039.43175
12                                          4409.16355
18                                         13376.061625
24                                          2895.402238
30                                         24235.473425

   Water Modality - Irrigation - Production
3                              119119.75662
12                               33763.0114
18                                      NaN
24                               11639.0848
30                            133922.173544
```

[88]:
```
# Define metadata
metadata = {
    "source":"INEGI Encuesta Nacional Agropecuaria 2017",
    "Production": "tonnes",
    "Areas": "hectares",
    "Note": "Last update was on 8th of January,2019 since there was en error␣
 ↪found and fixed"
}

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

```
# Display the modified DataFrame
maiz_df.attrs
```

[88]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2017',
  'Production': 'tonnes',
  'Areas': 'hectares',
  'Note': 'Last update was on 8th of January,2019 since there was en error found
and fixed'}}
```

[89]:
```
# Saving data
# Save DataFrame to CSV
maiz_df.to_csv('maize_data_2017.csv')

# Save metadata to a separate file (e.g., JSON)
import json
with open('maize_metadata_2017.json', 'w') as file:
    json.dump(metadata, file)
```

[90]:
```
#Check saved data
# Load DataFrame from CSV
maiz_df2 = pd.read_csv('maize_data_2017.csv', index_col=0)

# Load metadata from JSON
with open('maize_metadata_2017.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()
```

[90]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2017',
  'Production': 'tonnes',
  'Areas': 'hectares',
  'Note': 'Last update was on 8th of January,2019 since there was en error found
and fixed'}}
```

[ ]: