filter_census_2019

April 11, 2024

1 Preprocess original census data 2019

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

[37]:
```
# Imports
import pandas as pd
from pathlib import Path
```

[38]:
```
# Paths
original_path = Path.cwd().parent / 'original_data'
original_path
```

[38]:
```
PosixPath('/home/vant/Documents/valencia/agml_workshop/inegi_censos/original_dat
a')
```

[39]:
```
# Replace 'file_path.xlsx' with the path to your Excel file
file_path = original_path/'ena19_ent_agri02.xlsx'

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

[40]:
```
df.head()
```

[40]:
```
  Entidad federativa Cultivo seleccionado Superficie cultivada  \
0                NaN                   NaN                 Total
1                NaN                   NaN                   NaN
2                NaN                   NaN  Superficie sembrada
3                NaN                   NaN                   NaN
4                NaN                   NaN           Hectáreas

             Unnamed: 3           Unnamed: 4            Unnamed: 5  \
0                   NaN    Modalidad hídrica                   NaN
1                   NaN          De temporal                   NaN
2  Superficie cosechada  Superficie sembrada  Superficie cosechada
3                   NaN                   NaN                   NaN
```

```
4                   NaN                   NaN                   NaN

            Unnamed: 6            Unnamed: 7 Producción         Unnamed: 9  \
0                   NaN                   NaN        NaN                NaN
1             De riego                   NaN      Total  Modalidad hídrica
2  Superficie sembrada  Superficie cosechada        NaN      Bajo temporal
3                   NaN                   NaN        NaN                NaN
4                   NaN                   NaN  Toneladas                NaN

  Unnamed: 10
0         NaN
1         NaN
2  Bajo riego
3         NaN
4         NaN
```

[41]:
```
df.columns
```

[41]:
```
Index(['Entidad federativa', 'Cultivo seleccionado', 'Superficie cultivada',
       'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7',
       'Producción', 'Unnamed: 9', 'Unnamed: 10'],
      dtype='object')
```

[42]:
```
# Define new column names
column_names = ['Entidad federativa', 'Cultivo','Total superficie␣
↪sembrada','Total superficie cosechada',
                'MH-temporal superficie sembrada','MH-temporal superficie␣
↪cosechada','MH-riego superficie sembrada','MH-riego superficie cosechada',
                'Producción total', 'MH-temporal producción','MH-riego␣
↪producción']

# Rename the columns
df.columns = column_names

# Delete rows with initial no data
df = df.drop(index=range(6)).reset_index(drop=True)
```

[43]:
```
df.dropna(subset=['Cultivo'], inplace=True)
df.head()
```

[43]:
```
    Entidad federativa      Cultivo Total superficie sembrada  \
0      Baja California      Anuales                       NaN
1      Baja California      Cebolla                3722.67205
2      Baja California  Maíz blanco                  635.8649
3      Baja California  Trigo grano              82894.463903
4  Baja California Sur      Anuales                       NaN
```

```
  Total superficie cosechada MH-temporal superficie sembrada  \
0                        NaN                              NaN
1                  3721.64905                                0
2                    635.8649                                0
3                80554.267803                          3846.1392
4                        NaN                              NaN

  MH-temporal superficie cosechada MH-riego superficie sembrada  \
0                              NaN                           NaN
1                                0                    3722.67205
2                                0                      635.8649
3                           3474.2976                 79048.324703
4                              NaN                           NaN

  MH-riego superficie cosechada Producción total MH-temporal producción  \
0                           NaN              NaN                      NaN
1                   3721.64905     89666.876496                        0
2                     635.8649       6286.43536                        0
3                  77079.970203    427174.268979             12188.23215
4                           NaN              NaN                      NaN

  MH-riego producción
0                 NaN
1        89666.876496
2          6286.43536
3       414986.036829
4                 NaN
```

[44]:
```
# Extract only maize
maiz_df = df[df['Cultivo'].str.contains('Maíz', case=False)]
maiz_df.head(10)
```

[44]:
```
     Entidad federativa        Cultivo Total superficie sembrada  \
2       Baja California    Maíz blanco                  635.8649
5   Baja California Sur    Maíz blanco                 5596.0026
16              Chiapas    Maíz blanco            412950.845178
21            Chihuahua  Maíz amarillo            287635.764182
22            Chihuahua    Maíz blanco              43774.04167
27              Durango    Maíz blanco             75580.912031
31             Guerrero    Maíz blanco             461825.98459
35              Hidalgo    Maíz blanco            242242.611318
40              Jalisco    Maíz blanco            442494.440692
44     Estado de México    Maíz blanco            321974.216186

   Total superficie cosechada MH-temporal superficie sembrada  \
2                    635.8649                                0
5                     5511.54                         102.6676
```

```
16              395995.111668                    396831.302078
21              281470.829138                     51017.316192
22                38248.48574                      33206.05267
27               63446.434402                     34176.125614
31              394182.231747                    421360.558126
35              183654.094897                    177064.869218
40              437268.560967                    349235.980705
44              305118.049947                     260516.15259

   MH-temporal superficie cosechada MH-riego superficie sembrada  \
2                                 0                     635.8649
5                          102.6676                     5493.335
16                    379875.568568                   16119.5431
21                     48237.309148                  236618.44799
22                      27989.88314                     10567.989
27                      22161.969365                  41404.786417
31                     355092.107873                  40465.426464
35                     118655.229597                    65177.7421
40                      344452.31133                  93258.459987
44                     244362.539205                  61458.063596

   MH-riego superficie cosechada Producción total MH-temporal producción  \
2                         635.8649       6286.43536                      0
5                         5408.8724     37869.150169                44.0004
16                      16119.5431   1084385.734431         1024625.797131
21                   233233.51999   2481299.064627          232423.322577
22                     10258.6026     101107.46144            33516.17514
27                   41284.465037    298564.161223           27602.381701
31                   39090.123874    731011.750509          621558.761184
35                     64998.8653    601215.282388          169105.300928
40                   92816.249638   3410048.905136         2580426.755376
44                   60755.510742    621496.910713          459468.140854

   MH-riego producción
2           6286.43536
5          37825.149769
16           59759.9373
21       2248875.74205
22           67591.2863
27       270961.779522
31       109452.989325
35        432109.98146
40        829622.14976
44       162028.769858
```

[45]:
```
maiz_df.shape
```

[45]: (19, 11)

[46]:
```
# translate colnames to english
english_col_names = ['State', 'Crop','Total Cultivated area - Sown','Total␣
 ↪Cultivated area - Harvested',
                'Water Modality - Temporary - Cultivated area - Sown',
                'Water Modality - Temporary - Cultivated area - Harvested',
                'Water Modality - Irrigation - Cultivated area - Sown',
                'Water Modality - Irrigation - Cultivated area - Harvested',
                'Total production',
                'Water Modality - Temporary - Production',
                '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)
```

[46]:

	State	Crop	Total Cultivated area - Sown	Total Cultivated area - Harvested	Water Modality - Temporary - Cultivated area - Sown
2	Baja California	White corn	635.8649	635.8649	0
5	Baja California Sur	White corn	5596.0026	5511.54	102.6676
16	Chiapas	White corn	412950.845178	395995.111668	396831.302078
21	Chihuahua	Yellow corn	287635.764182	281470.829138	51017.316192
22	Chihuahua	White corn	43774.04167	38248.48574	

```
22                                        33206.05267

   Water Modality - Temporary - Cultivated area - Harvested  \
2                                                      0
5                                                102.6676
16                                           379875.568568
21                                            48237.309148
22                                             27989.88314

   Water Modality - Irrigation - Cultivated area - Sown  \
2                                                635.8649
5                                                5493.335
16                                             16119.5431
21                                            236618.44799
22                                               10567.989

   Water Modality - Irrigation - Cultivated area - Harvested Total production  \
2                                                635.8649              6286.43536
5                                                5408.8724            37869.150169
16                                             16119.5431          1084385.734431
21                                            233233.51999          2481299.064627
22                                              10258.6026            101107.46144

   Water Modality - Temporary - Production  \
2                                        0
5                                  44.0004
16                           1024625.797131
21                            232423.322577
22                              33516.17514

   Water Modality - Irrigation - Production
2                                6286.43536
5                               37825.149769
16                                59759.9373
21                              2248875.74205
22                                 67591.2863
```

[47]:
```
# Define metadata
metadata = {
    "source":"INEGI Encuesta Nacional Agropecuaria 2019",
    "Production": "tonnes",
    "Areas": "hectares",
    "Note": "Data for states of Aguascalientes, Coahuila, and Quintana Roo is␣
 ↪not published because the collected information from the selected crops is␣
 ↪insufficient to obtain estimated data."
}
```

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

# Display the modified DataFrame
maiz_df.attrs
```

[47]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2019',
  'Production': 'tonnes',
  'Areas': 'hectares',
  'Note': 'Data for states of Aguascalientes, Coahuila, and Quintana Roo is not
published because the collected information from the selected crops is
insufficient to obtain estimated data.'}}
```

[48]:
```
# Saving data
# Save DataFrame to CSV
maiz_df.to_csv('maize_data_2019.csv')

# Save metadata to a separate file (e.g., JSON)
import json
with open('maize_metadata_2019.json', 'w') as file:
    json.dump(metadata, file)
```

[49]:
```
#Check saved data
# Load DataFrame from CSV
maiz_df2 = pd.read_csv('maize_data_2019.csv', index_col=0)

# Load metadata from JSON
with open('maize_metadata_2019.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()
```

[49]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2019',
  'Production': 'tonnes',
  'Areas': 'hectares',
  'Note': 'Data for states of Aguascalientes, Coahuila, and Quintana Roo is not
published because the collected information from the selected crops is
insufficient to obtain estimated data.'}}
```

[ ]:

[ ]: