filter_census_2014

April 11, 2024

1 Preprocess original census data 2014

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

[28]:
```
# Imports
import pandas as pd
from pathlib import Path
```

[29]:
```
# Paths
original_path = Path.cwd().parent / 'original_data'
original_path
```

[29]:
```
PosixPath('/home/vant/Documents/valencia/agml_workshop/inegi_censos/original_dat
a')
```

[30]:
```
# Replace 'file_path.xlsx' with the path to your Excel file
file_path = original_path/'ena14_agri02.xlsx'

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

[31]:
```
df.head()
```

[31]:
```
  Entidad  Cultivo  \
0     NaN      NaN
1     NaN      NaN
2     NaN      NaN
3    Ags.      NaN
4    Ags.  Alfalfa

  Entidad federativa y cultivo con representatividad en la muestra  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                     Aguascalientes
```

```
4                                                Alfalfa

  Superficie cultivada            Unnamed: 4 Producción
0  Superficie sembrada  Superficie cosechada        NaN
1            Hectáreas                   NaN  Toneladas
2                  NaN                   NaN        NaN
3                  NaN                   NaN        NaN
4          6461.311717           6420.884037        NaN
```

[32]:
```
df.columns
```

[32]:
```
Index(['Entidad', 'Cultivo',
       'Entidad federativa y cultivo con representatividad en la muestra',
       'Superficie cultivada', 'Unnamed: 4', 'Producción'],
      dtype='object')
```

[33]:
```
# Define new column names
column_names = ['Entidad federativa', 'Cultivo','Entidad federativa y␣
 ↪cultivo','Total superficie sembrada','Total superficie cosechada',
                'Producción total']

# Rename the columns
df.columns = column_names

# Delete rows with initial no data
df = df.drop(index=range(3)).reset_index(drop=True)
```

[34]:
```
df.head()
```

[34]:
```
  Entidad federativa      Cultivo Entidad federativa y cultivo  \
0               Ags.          NaN               Aguascalientes
1               Ags.      Alfalfa                      Alfalfa
2               Ags.       Frijol                       Frijol
3               Ags.  Maíz blanco                  Maíz blanco
4                BC.          NaN              Baja California

  Total superficie sembrada Total superficie cosechada Producción total
0                       NaN                       NaN              NaN
1                6461.311717               6420.884037              NaN
2                10156.525639              8594.411379      3797.035222
3                74292.287463              67277.37109    274683.292887
4                       NaN                       NaN              NaN
```

[35]:
```
# 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      Ags.                  Aguascalientes
4       BC.                 Baja California
9      BCS.              Baja California Sur
14    Camp.                        Campeche
17    Coah.            Coahuila de Zaragoza
25     Col.                          Colima
29    Chis.                         Chiapas
33    Chih.                        Chihuahua
39      DF.               Distrito Federal
42     Dgo.                         Durango
48     Gto.                       Guanajuato
54     Gro.                        Guerrero
58     Hgo.                         Hidalgo
62     Jal.                         Jalisco
66     Mex.                          México
69    Mich.            Michoacán de Ocampo
76     Mor.                         Morelos
80     Nay.                         Nayarit
86      NL.                      Nuevo León
91     Oax.                          Oaxaca
96     Pue.                          Puebla
100    Qro.                       Querétaro
105  Q. Roo                   Quintana Roo
109    SLP.                 San Luis Potosí
114    Sin.                         Sinaloa
118    Son.                          Sonora
122    Tab.                         Tabasco
126  Tamps.                     Tamaulipas
132   Tlax.                       Tlaxcala
134    Ver.  Veracruz de Ignacio de la Llave
140    Yuc.                         Yucatán
143    Zac.                       Zacatecas
```

[36]:
```
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()
```

[36]:
```
  Entidad federativa      Cultivo Total superficie sembrada  \
1               Ags.      Alfalfa               6461.311717
2               Ags.       Frijol              10156.525639
3               Ags.  Maíz blanco              74292.287463
5                BC.      Alfalfa              28979.202828
6                BC.      Algodón              28204.512802

  Total superficie cosechada Producción total
1                6420.884037              NaN
2                8594.411379      3797.035222
3                67277.37109    274683.292887
5               28780.852218              NaN
6               28146.903202     84765.363171
```

[37]:
```
# Extract only maize
maiz_df = df_clean[df_clean['Cultivo'].str.contains('Maíz', case=False)]
maiz_df.head(10)
```

[37]:
```
   Entidad federativa         Cultivo Total superficie sembrada  \
3                Ags.     Maíz blanco              74292.287463
12               BCS.     Maíz blanco                 5351.2698
16              Camp.     Maíz blanco             167105.616512
20              Coah.     Maíz blanco              49616.524801
28               Col.     Maíz blanco              16981.536845
32              Chis.     Maíz blanco              572650.95816
38              Chih.     Maíz blanco             151801.744651
41                DF.     Maíz blanco               4630.825465
46               Dgo.     Maíz blanco             147282.294479
47               Dgo.  Maíz forrajero              37609.693629

   Total superficie cosechada Producción total
3                67277.37109    274683.292887
12                 5147.8698       35171.0482
16              148953.002642    408859.462636
20               37750.320798     70988.073773
28               16320.717689       56074.4033
32              543991.815373   1165423.163722
38              141981.005084    488235.542461
41               4460.835624      8731.303384
46              142243.870047    507384.326924
47               36616.705641              NaN
```

[38]:
```
maiz_df.shape
```

[38]:
```
(33, 5)
```

[39]:
```
# 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()
```

[39]:

	Entidad federativa	Cultivo	Total superficie sembrada	Total superficie cosechada	Producción total
3	Aguascalientes	Maíz blanco	74292.287463	67277.37109	274683.292887
12	Baja California Sur	Maíz blanco	5351.2698	5147.8698	35171.0482
16	Campeche	Maíz blanco	167105.616512	148953.002642	408859.462636
20	Coahuila de Zaragoza	Maíz blanco	49616.524801	37750.320798	70988.073773
28	Colima	Maíz blanco	16981.536845	16320.717689	56074.4033

[40]:
```
# translate colnames to english
english_col_names = ['State',
                     'Crop',
                     'Total Cultivated area - Sown',
                     'Total Cultivated area - Harvested',
                     'Total 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)
```

[40]:

	State	Crop	Total Cultivated area - Sown	Total Cultivated area - Harvested	Total production
3	Aguascalientes	White corn	74292.287463	67277.37109	274683.292887
12	Baja California Sur	White corn	5351.2698	5147.8698	35171.0482
16	Campeche	White corn	167105.616512	148953.002642	408859.462636
20	Coahuila de Zaragoza	White corn	49616.524801	37750.320798	70988.073773
28	Colima	White corn	16981.536845	16320.717689	56074.4033

```
   Total Cultivated area - Harvested Total production
3                        67277.37109    274683.292887
12                         5147.8698       35171.0482
16                      148953.002642    408859.462636
20                       37750.320798     70988.073773
28                       16320.717689       56074.4033
```

[41]:
```python
# Define metadata
metadata = {
    "source":"INEGI Encuesta Nacional Agropecuaria 2014",
    "Production": "tonnes",
    "Areas": "hectares"
}

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

# Display the modified DataFrame
maiz_df.attrs
```

[41]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2014',
  'Production': 'tonnes',
  'Areas': 'hectares'}}
```

[42]:
```python
# Saving data
# Save DataFrame to CSV
maiz_df.to_csv('maize_data_2014.csv')

# Save metadata to a separate file (e.g., JSON)
import json
with open('maize_metadata_2014.json', 'w') as file:
    json.dump(metadata, file)
```

[43]:
```python
#Check saved data
# Load DataFrame from CSV
maiz_df2 = pd.read_csv('maize_data_2014.csv', index_col=0)

# Load metadata from JSON
with open('maize_metadata_2014.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()
```

[43]:
```
{'metadata': {'source': 'INEGI Encuesta Nacional Agropecuaria 2014',
  'Production': 'tonnes',
  'Areas': 'hectares'}}
```

[ ]:

[ ]: