Metadata

Title: AK_clean

Author: Ami Kano

Date: March 12, 2023

Comments:

This is an attempt at cleaning the given data.

Set-up

In [1]:

```
from pymongo import MongoClient
import numpy as np
import pandas as pd
```

In [2]:

```
# URI is specific to Ami's login credentials
uri = "mongodb+srv://DS6013_Students_Ami:DS6013_Students_AK@countyrecords.4cdfg

# connect to database
client = MongoClient(uri)
database = client['TaxRecords']
```

In [3]:

```
database.list_collection_names()
```

Out[3]:

```
['Tax_Record_1867', 'Tax_Record_1782']
```

In [4]:

```
record_1867 = pd.DataFrame(list(database['Tax_Record_1867'].find()))
record_1782 = pd.DataFrame(list(database['Tax_Record_1782'].find()))
```

Cleaning/Preparing Data

In [5]:

```
# drop seemingly irrelevant or redundant columns

record_1867 = record_1867.drop(['_id', 'PersonTaxCountHorsesMules', 'PersonTaxV
       'PersonTaxCountCattle', 'PersonTaxValueCattle', 'PersonTaxCountSheep',
       'PersonTaxValueSheep', 'PersonTaxCountHogs', 'PersonTaxValueHogs',
       'PersonTaxCountCarriageWagon', 'PersonTaxValueCarriageWagon',
       'PersonTaxValueFurnishings', 'PersonTaxValueJewelry',
       'PersonTaxValueAggregatePersonlProperty', 'PersonTaxStateAll',
       'PersonTaxLeviedLand', 'PersonTaxTotalCountyValue', 'EventImageLink',
       'PersonsTaxedCountWMalesover21', 'PersonTaxCountWMalesover16',
       'PersonTaxCountWatches', 'PersonTaxValueWatches',
       'PersonTaxCountClocks', 'PersonTaxValueClocks',
       'PersonTaxCountMusicalInstruments', 'PersonTaxValueMusicalInstruments',
       'PersonTaxCommissionerRemarks', 'PersonsTaxedCountNMalesover21',
       'PersonTaxCountNMalesover16', 'PersonTaxValueMoniesSchC1'], axis=1)
```

In [6]:

```
record_1782 = record_1782.drop(['_id', 'PersonCountTaxableTithes',
       'PersonCountTaxableEnslavedPersons', 'PersonTaxCountHorsesMules',
       'PersonTaxCountCattle', 'EventArchiveLink',
       'PersonTaxCommissionerRemarks'], axis=1)
```

In [7]:

```
# lowercase text

for text_col in list(record_1867.select_dtypes(include=['object']).columns):
    record_1867[text_col] = record_1867[text_col].str.lower()

for text_col in list(record_1782.select_dtypes(include=['object']).columns):
    record_1782[text_col] = record_1782[text_col].str.lower()
```

In [8]:

```
# replace NaN with empty string

record_1867 = record_1867.fillna('')
record_1782 = record_1782.fillna('')
```

In [9]:

```
# make `EventTitle` the index

record_1867['EventTitle'] = "1867 "+record_1867['EventTitle'].astype(str)
record_1867 = record_1867.set_index('EventTitle')

record_1782 = record_1782.set_index('EventTitle')
```

In [10]:

```
record_1867.head()
```

Out[10]:

	SourceSteward	SourceLocCity	SourceLocState	SourceTitle	SourceType	SourceDat
EventTitle						
1867 personal property tax recorded	library of virginia	richmond	virginia	county personal property taxes	government record	
1867 personal property tax recorded	library of virginia	richmond	virginia	county personal property taxes	government record	
1867 personal property tax recorded	library of virginia	richmond	virginia	county personal property taxes	government record	
1867 personal property tax recorded	library of virginia	richmond	virginia	county personal property taxes	government record	
1867 personal property tax recorded	library of virginia	richmond	virginia	county personal property taxes	government record	

In [11]:

```
record_1782.head()
```

Out[11]:

	SourceSteward	SourceLocCity	SourceLocState	SourceTitle	SourceType	SourceDat
EventTitle						
caleb stone personal property tax recorded	fluvanna county historical society	palmyra	virginia	county personal property taxes	government record	
william bernard personal property tax recorded	fluvanna county historical society	palmyra	virginia	county personal property taxes	government record	
caleb stone personal property tax recorded	fluvanna county historical society	palmyra	virginia	county personal property taxes	government record	
john ashlin personal property tax recorded	fluvanna county historical society	palmyra	virginia	county personal property taxes	government record	
john ashlin personal property tax recorded	fluvanna county historical society	palmyra	virginia	county personal property taxes	government record	