Metadata

Title: AK_EDA

Author: Ami Kano

Date: March 12, 2023

Comments:

This is a quick exploratory data analysis of the two tables provided by the project sponsor.

Variables, null values, amount of unique values, and distribution of some variables were explored.

Set-up

In [1]:

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

In [3]:

```
# 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 [4]:

```
database.list_collection_names()
```

Out[4]:

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

In [8]:

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

Exploratory Data Analysis

Quick Look at Data

In [9]:

```
record_1867
```

Out[9]:

	_id	SourceSteward	SourceLocCity	SourceLocState	SourceTitle
0	63e9361e59c84387372abcb2	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
1	63e9361e59c84387372abcaf	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
2	63e9361e59c84387372abcc1	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
3	63e9361e59c84387372abcb5	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
4	63e9361e59c84387372abcbc	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
...	...	...	...	...	..
12358	63e9362759c84387372aecca	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
12359	63e9362759c84387372aeccb	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
12360	63e9362759c84387372aece6	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
12361	63e9362759c84387372aecf1	Library of Virginia	Richmond	Virginia	County Persona Property Taxes
12362	63e9362759c84387372aecf3	Library of Virginia	Richmond	Virginia	County Persona Property Taxes

12363 rows × 50 columns

In [10]:

```
record_1782
```

Out[10]:

	_id	SourceSteward	SourceLocCity	SourceLocState	SourceTitle
0	63e9351559c84387372ab951	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
1	63e9351559c84387372ab97c	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
2	63e9351559c84387372ab958	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
3	63e9351559c84387372ab95d	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
4	63e9351559c84387372ab965	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
...	...	...	...	...	...
857	63e9351559c84387372abc5a	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
858	63e9351559c84387372abc66	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
859	63e9351559c84387372abc7e	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
860	63e9351559c84387372abc9a	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes
861	63e9351559c84387372abca5	Fluvanna County Historical Society	Palmyra	Virginia	County Personal Property Taxes

862 rows × 24 columns

Column Values

In [11]:

```
record_1867.columns
```

```
Out[11]: Index(['_id', 'SourceSteward', 'SourceLocCity', 'SourceLocState',
                'SourceTitle', 'SourceType', 'SourceDateYearCreated', 'SourceCreator',
                'SourceLocCreatedCounty', 'SourceAuthorName', 'EventTitle',
                'EventLocJurisdictionCounty', 'EventDateYear', 'PersonSurname',
                'PersonGivenNames', 'PersonNameSuffix', 'PersonEventRole',
                'PersonTaxCountHorsesMules', 'PersonTaxValueHorsesMules',
                'PersonTaxCountCattle', 'PersonTaxValueCattle', 'PersonTaxCountSheep',
                'PersonTaxValueSheep', 'PersonTaxCountHogs', 'PersonTaxValueHogs',
                'PersonTaxCountCarriageWagon', 'PersonTaxValueCarriageWagon',
                'PersonTaxValueFurnishings', 'PersonTaxValueJewelry',
                'PersonTaxValueAggregatePersonlProperty', 'PersonTaxStateAll',
                'PersonTaxLeviedLand', 'PersonTaxTotalCountyValue', 'EventImageLink',
                'PersonsTaxedCountWMalesover21', 'PersonTaxCountWMalesover16',
                'PersonNameAlternate', 'PersonTaxCountWatches', 'PersonTaxValueWatche
         s',
                'PersonTaxCountClocks', 'PersonTaxValueClocks',
                'PersonTaxCountMusicalInstruments', 'PersonTaxValueMusicalInstruments',
                'PersonTaxCommissionerRemarks', 'PersonsTaxedCountNMalesover21',
                'PersonTaxCountNMalesover16', 'PersonRoleLocSurnameEmployer',
                'PersonRoleGivenNamesEmployer', 'PersonRoleLocResidence',
                'PersonTaxValueMoniesSchC1'],
               dtype='object')
```

```
In [12]: record_1782.columns
```

```
Out[12]: Index(['_id', 'SourceSteward', 'SourceLocCity', 'SourceLocState',
                'SourceTitle', 'SourceType', 'SourceDateYearCreated', 'SourceCreator',
                'SourceLocCreatedCounty', 'SourceAuthorName', 'EventTitle',
                'EventLocJurisdictionCounty', 'EventDateYear', 'PersonSurname',
                'PersonGivenNames', 'PersonRaceNotation', 'PersonCountTaxableTithes',
                'PersonCountTaxableEnslavedPersons', 'PersonTaxCountHorsesMules',
                'PersonTaxCountCattle', 'PersonEventRole', 'EventArchiveLink',
                'PersonNameSuffix', 'PersonTaxCommissionerRemarks'],
               dtype='object')
```

Check for Null Values

```
In [13]: record_1867.isnull().sum()
```

Out[13]:

```
_id                                           0
SourceSteward                                 0
SourceLocCity                                 0
SourceLocState                                0
SourceTitle                                   0
SourceType                                    0
SourceDateYearCreated                         0
SourceCreator                                 0
SourceLocCreatedCounty                        0
SourceAuthorName                              0
EventTitle                                    0
EventLocJurisdictionCounty                    0
EventDateYear                                 0
PersonSurname                                 2
PersonGivenNames                             36
PersonNameSuffix                           9715
PersonEventRole                               0
PersonTaxCountHorsesMules                  8733
PersonTaxValueHorsesMules                  8728
PersonTaxCountCattle                       8245
PersonTaxValueCattle                       8248
PersonTaxCountSheep                       10996
PersonTaxValueSheep                       10996
PersonTaxCountHogs                         8809
PersonTaxValueHogs                         8794
PersonTaxCountCarriageWagon               11040
PersonTaxValueCarriageWagon               11052
PersonTaxValueFurnishings                  8159
PersonTaxValueJewelry                     11833
PersonTaxValueAggregatePersonlProperty     7372
PersonTaxStateAll                           321
PersonTaxLeviedLand                       10537
PersonTaxTotalCountyValue                  8694
EventImageLink                                0
PersonsTaxedCountWMalesover21              7279
PersonTaxCountWMalesover16                10741
PersonNameAlternate                       11266
PersonTaxCountWatches                     11632
PersonTaxValueWatches                     11194
PersonTaxCountClocks                      11198
PersonTaxValueClocks                      10650
PersonTaxCountMusicalInstruments          12121
PersonTaxValueMusicalInstruments          11919
PersonTaxCommissionerRemarks              10211
PersonsTaxedCountNMalesover21              6147
PersonTaxCountNMalesover16                10255
PersonRoleLocSurnameEmployer               5519
PersonRoleGivenNamesEmployer               5757
PersonRoleLocResidence                     9878
PersonTaxValueMoniesSchC1                 12360
dtype: int64
```

In [14]:

```
record_1782.isnull().sum()
```

Out[14]:

```
_id                                    0
SourceSteward                          0
SourceLocCity                          0
SourceLocState                         0
SourceTitle                            0
SourceType                             0
SourceDateYearCreated                  0
SourceCreator                          0
SourceLocCreatedCounty                 0
SourceAuthorName                       0
EventTitle                             0
EventLocJurisdictionCounty             0
EventDateYear                          0
PersonSurname                        636
PersonGivenNames                       0
PersonRaceNotation                    39
PersonCountTaxableTithes             663
PersonCountTaxableEnslavedPersons    757
PersonTaxCountHorsesMules            657
PersonTaxCountCattle                 665
PersonEventRole                        0
EventArchiveLink                       0
PersonNameSuffix                     859
PersonTaxCommissionerRemarks         809
dtype: int64
```

Unique Values

record_1867

In [27]:

```
for column in ["SourceSteward",
                "SourceLocCity",
                "SourceLocState",
                "SourceTitle",
                "SourceType",
                "SourceDateYearCreated",
                "SourceCreator",
                "SourceLocCreatedCounty",
                "SourceAuthorName",
                "EventTitle",
                "EventLocJurisdictionCounty",
                "EventDateYear"]:

    print(column+":", record_1867[column].unique())
```

```
SourceSteward: ['Library of Virginia']
SourceLocCity: ['Richmond']
SourceLocState: ['Virginia']
SourceTitle: ['County Personal Property Taxes']
SourceType: ['Government Record']
SourceDateYearCreated: [1867]
SourceCreator: ['Cumberland County' 'Buckingham County' 'Fluvanna County'
 'Louisa County District 1' 'Louisa County District 2' 'Orange County']
SourceLocCreatedCounty: ['Cumberland' 'Buckingham' 'Fluvanna' 'Louisa' 'Orang
e']
SourceAuthorName: ['R B Trent' '1 J E Morgan' '2 Wm K Saunders' 'O B Thomas'
 'John A Perkins' 'Robert F Moss' 'G W Wright' 'GW Wright']
EventTitle: ['Personal Property Tax Recorded']
EventLocJurisdictionCounty: ['Cumberland' 'Buckingham' 'Fluvanna' 'Louisa' 'Lo
uisa ' 'Orange']
EventDateYear: [1867 1868 1869]
```

In [28]:

```
for column in ['PersonSurname',
               'PersonGivenNames', 'PersonNameSuffix', 'PersonEventRole',
               'PersonTaxCountHorsesMules', 'PersonTaxValueHorsesMules',
               'PersonTaxCountCattle', 'PersonTaxValueCattle', 'PersonTaxCountS
               'PersonTaxValueSheep', 'PersonTaxCountHogs', 'PersonTaxValueHogs
               'PersonTaxCountCarriageWagon', 'PersonTaxValueCarriageWagon',
               'PersonTaxValueFurnishings', 'PersonTaxValueJewelry',
               'PersonTaxValueAggregatePersonlProperty', 'PersonTaxStateAll',
               'PersonTaxLeviedLand', 'PersonTaxTotalCountyValue', 'EventImageI
               'PersonsTaxedCountWMalesover21', 'PersonTaxCountWMalesover16',
               'PersonNameAlternate', 'PersonTaxCountWatches', 'PersonTaxValueW
               'PersonTaxCountClocks', 'PersonTaxValueClocks',
               'PersonTaxCountMusicalInstruments', 'PersonTaxValueMusicalInstru
               'PersonTaxCommissionerRemarks', 'PersonsTaxedCountNMalesover21',
               'PersonTaxCountNMalesover16', 'PersonRoleLocSurnameEmployer',
               'PersonRoleGivenNamesEmployer', 'PersonRoleLocResidence',
               'PersonTaxValueMoniesSchC1']:

    if len(record_1867[column].unique()) >= 5:
        display_value = record_1867[column].unique()[0:5]
    else:
        display_value = record_1867[column].unique()

    print(column+":", display_value)
```

```
PersonSurname: ['Alderson' 'Allen' 'Amonett' 'Anderson' 'Amos']
PersonGivenNames: ['Thomas' 'Joseph L' 'Benj A' 'Jno T' 'Charles']
PersonNameSuffix: ['Est' nan 'MD' 'and sons' 'Miss']
PersonEventRole: ['taxpayer' 'resident and taxpayer']
PersonTaxCountHorsesMules: [ 4.  3.  5. nan  1.]
PersonTaxValueHorsesMules: [225. 135. 300.  nan 100.]
PersonTaxCountCattle: [ 9.  3. 13. nan  6.]
PersonTaxValueCattle: [170.  75. 250.  nan 150.]
PersonTaxCountSheep: [14. nan 12. 30.  3.]
PersonTaxValueSheep: [20. nan 25. 60.  5.]
PersonTaxCountHogs: [ 6.  8. 16. nan  5.]
PersonTaxValueHogs: [30. 34. 48. nan 40.]
PersonTaxCountCarriageWagon: [ 1. nan  2.  3.  8.]
PersonTaxValueCarriageWagon: [ 50.  nan  10. 300.  20.]
PersonTaxValueFurnishings: [ 50. 100.  25.  nan 150.]
PersonTaxValueJewelry: [20.  6. 18. nan 40.]
PersonTaxValueAggregatePersonlProperty: [565. 310. 981.  25. 100.]
PersonTaxStateAll: [1.7  1.65 3.55 0.68 0.9 ]
PersonTaxLeviedLand: [42. 41. 89. 17. 23.]
PersonTaxTotalCountyValue: [0.42 1.41 2.89 1.17 1.23]
EventImageLink: ['https://onesharedstory.org/HBCP/files/original/e9b003282b83e
abf0582b6b711df6bf0.pdf'
 'https://onesharedstory.org/HBCP/files/original/51eed9f02c794805091f40beaec80
32b.pdf'
 'https://onesharedstory.org/HBCP/files/original/acb92f5a7e347a3ee664a08e1f602
f99.pdf'
 'https://onesharedstory.org/HBCP/files/original/a1f47370685597126ee701750f23c
e4e.pdf'
 'https://onesharedstory.org/HBCP/files/original/9f525de1ff4466a60258041701244
7d5.pdf']
PersonsTaxedCountWMalesover21: [nan  1.  3.  2.  4.]
PersonTaxCountWMalesover16: [nan  1.  2.  3.  5.]
PersonNameAlternate: [nan 'Benjamin A' 'John T' 'John L' 'Richard W']
PersonTaxCountWatches: [nan  1.  2. 15. 25.]
PersonTaxValueWatches: [ nan  25.  10. 100.  80.]
PersonTaxCountClocks: [nan  2.  1. 10.  3.]
PersonTaxValueClocks: [nan  5. 12. 10.  2.]
PersonTaxCountMusicalInstruments: [nan  1.  2.  5.]
PersonTaxValueMusicalInstruments: [ nan 200. 150. 300.  50.]
PersonTaxCommissionerRemarks: [nan 'W' 'Bazaar Farm' 'Fork Farm' 'for 1867']
PersonsTaxedCountNMalesover21: [nan  1.  2.  4.  5.]
PersonTaxCountNMalesover16: [nan  1.  2.  4.  3.]
PersonRoleLocSurnameEmployer: [nan 'Hubard' 'McRae' 'Jones' 'Flanigan']
PersonRoleGivenNamesEmployer: [nan 'E W' 'E S' 'H A' 'Matt']
PersonRoleLocResidence: [nan 'Cartersville' 'Wood Lawn' 'at J S Durham' 'Hoope
rs Rock']
PersonTaxValueMoniesSchC1: [  nan 5000. 3000.]
```

record_1782

In [23]:

```
for column in ["SourceSteward",
               "SourceLocCity",
               "SourceLocState",
               "SourceTitle",
               "SourceType",
               "SourceDateYearCreated",
               "SourceCreator",
               "SourceLocCreatedCounty",
               "SourceAuthorName",
               "EventLocJurisdictionCounty",
               "EventDateYear"]:
   print(column+":", record_1782[column].unique())
```

```
SourceSteward: ['Fluvanna County Historical Society']
SourceLocCity: ['Palmyra']
SourceLocState: ['Virginia']
SourceTitle: ['County Personal Property Taxes']
SourceType: ['Government Record']
SourceDateYearCreated: [1782]
SourceCreator: ['Fluvanna County']
SourceLocCreatedCounty: ['Fluvanna']
SourceAuthorName: ['Jos Haden' 'Thomas Napier' 'John Ware' 'Benj Anderson'
 'Samuel Richardson' 'Samuel Richarson' 'Roger Thompson']
EventLocJurisdictionCounty: ['Fluvanna County']
EventDateYear: [1782]
```

In [26]:

```
for column in ["PersonSurname",
               "PersonGivenNames",
               "PersonRaceNotation",
               "PersonCountTaxableTithes",
               "PersonCountTaxableEnslavedPersons",
               "PersonTaxCountHorsesMules",
               "PersonTaxCountCattle",
               "PersonEventRole",
               "EventTitle",
               "EventArchiveLink",
               "PersonNameSuffix",
               "PersonTaxCommissionerRemarks"]:

    if len(record_1782[column].unique()) >= 5:
        display_value = record_1782[column].unique()[0:5]
    else:
        display_value = record_1782[column].unique()

    print(column+":", display_value)
```

```
PersonSurname: ['Stone' nan 'Rice' 'Jiles' 'Farrow']
PersonGivenNames: ['Caleb' 'Hannah' 'Sue' 'Dick' 'Sam']
PersonRaceNotation: ['W' 'NN' 'N' nan 'w']
PersonCountTaxableTithes: [ 1. nan  3.  2.]
PersonCountTaxableEnslavedPersons: [ 9. nan  1.  4. 13.]
PersonTaxCountHorsesMules: [ 5. nan  2.  1.  4.]
PersonTaxCountCattle: [41. nan  7.  4.  3.]
PersonEventRole: ['Tax Payer' 'Person Taxed As Property']
EventTitle: ['Caleb Stone Personal Property Tax Recorded'
 'William Bernard Personal Property Tax Recorded'
 'John Ashlin Personal Property Tax Recorded'
 'Joseph Haden Personal Property Tax Recorded'
 'Samuel Richardson Personal Property Tax Recorded']
EventArchiveLink: ['http://piedmontvahistory.org/archives14/items/show/1809'
 'http://piedmontvahistory.org/archives14/items/show/1810'
 'http://piedmontvahistory.org/archives14/items/show/1814'
 'http://piedmontvahistory.org/archives14/items/show/1815'
 'http://piedmontvahistory.org/archives14/items/show/1816']
PersonNameSuffix: [nan 'Jr' 'estate of ' 'Col']
PersonTaxCommissionerRemarks: [nan ' ' 'estate of ' 'At the Fork' 'At the Ferr
y']
```

Distribution of Values

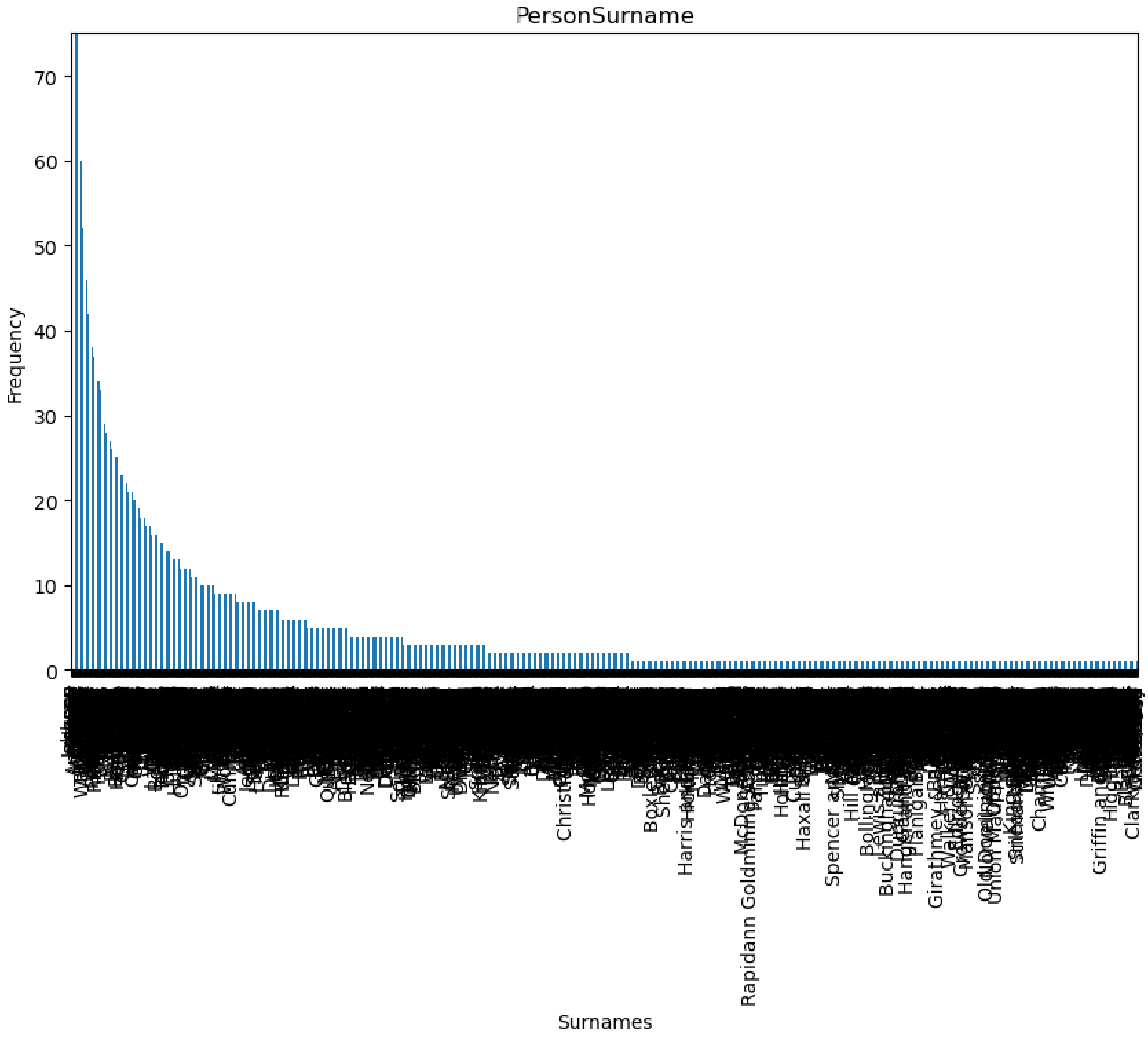
Name-related Variables

In [31]:

```
ax = record_1867['PersonSurname'].value_counts().plot(kind='bar',
                                                      figsize=(10,6),
                                                      ylim=(0,75),
                                                      title="PersonSurname")
ax.set_xlabel("Surnames")
ax.set_ylabel("Frequency")
```

Out[31]:

```
Text(0, 0.5, 'Frequency')
```

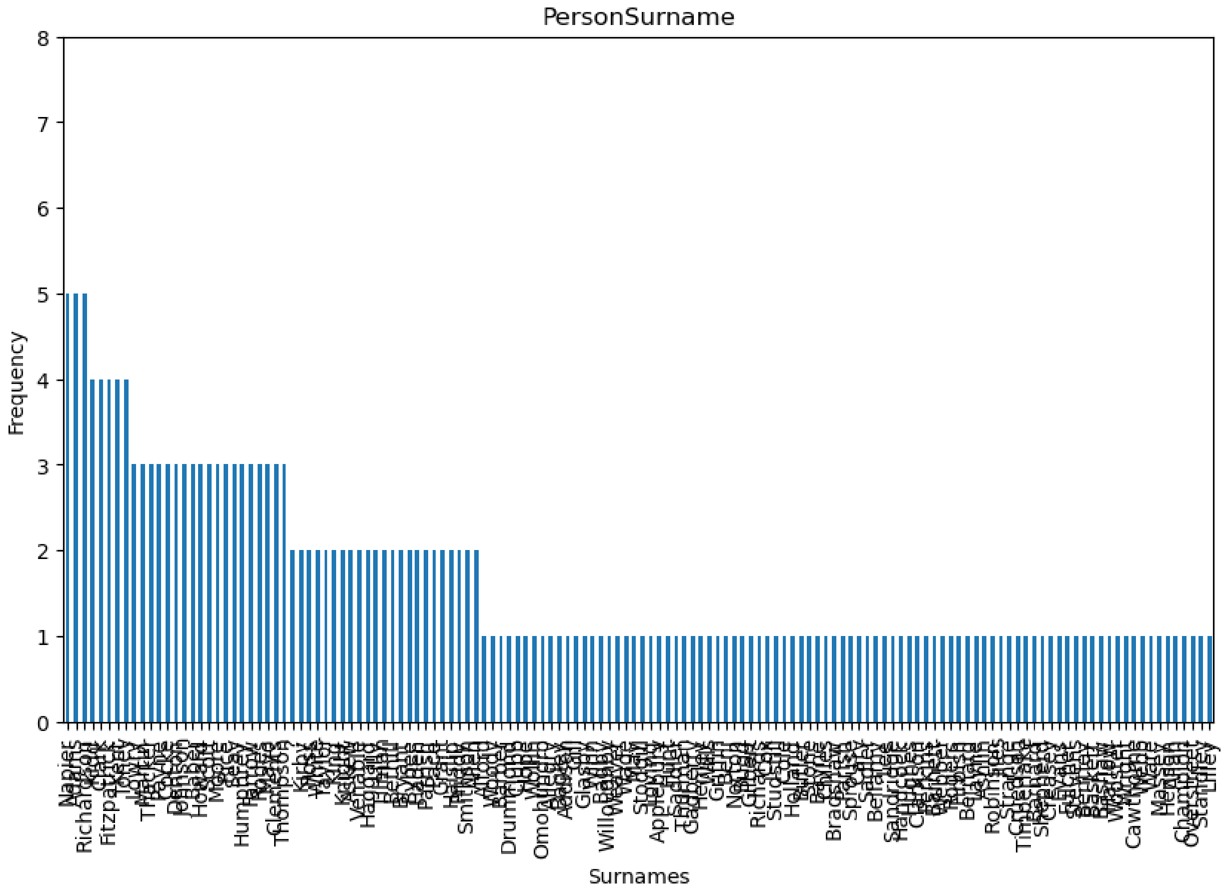


In [34]:

```
ax = record_1782['PersonSurname'].value_counts().plot(kind='bar',
                                                      figsize=(10,6),
                                                      ylim=(0,8),
                                                      title="PersonSurname")
ax.set_xlabel("Surnames")
ax.set_ylabel("Frequency")
```

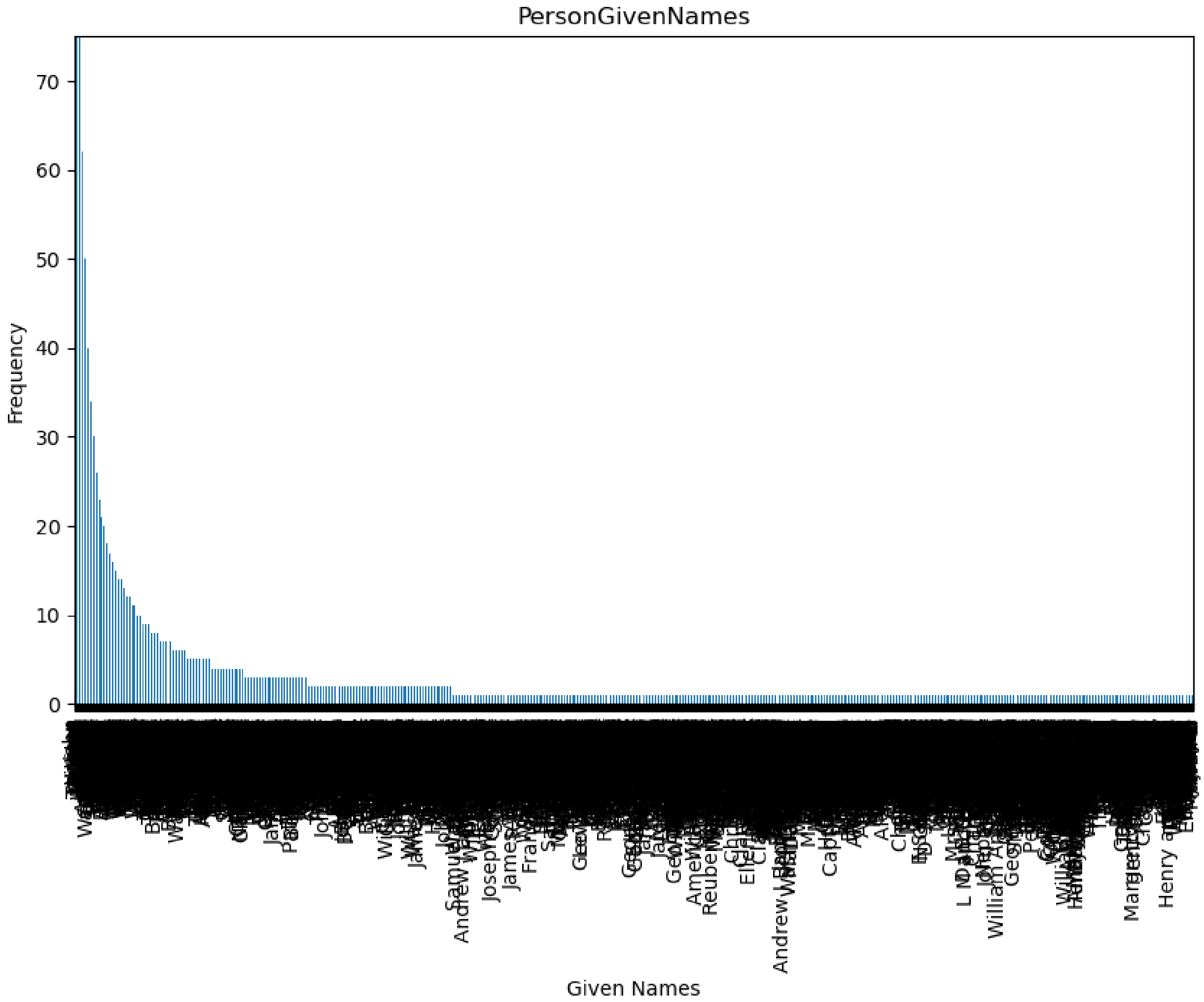
Out[34]:

```
Text(0, 0.5, 'Frequency')
```

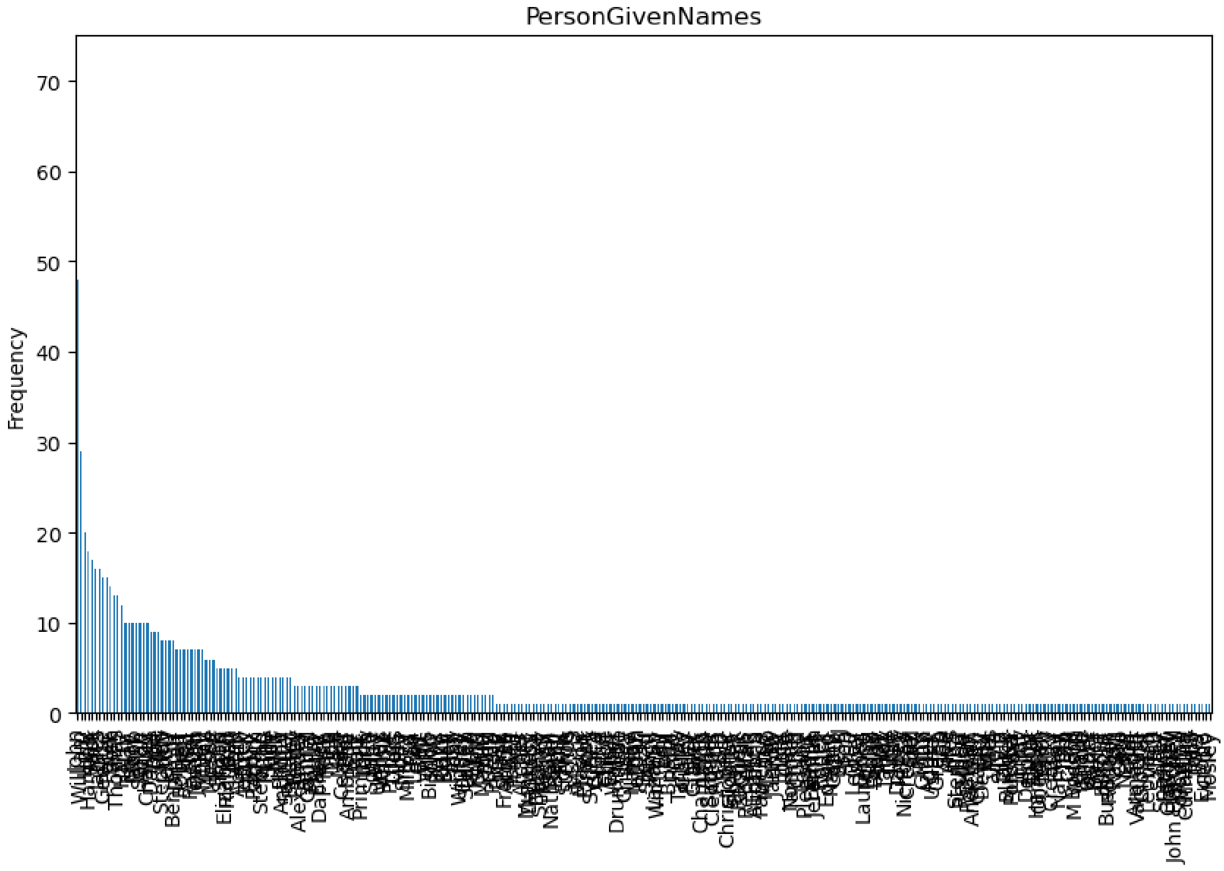


```
In [38]: ax = record_1867["PersonGivenNames"].value_counts().plot(kind='bar',
                                                          figsize=(10,6),
                                                          ylim=(0,75),
                                                          title="PersonGivenNames")
         ax.set_xlabel("Given Names")
         ax.set_ylabel("Frequency")
```

```
Out[38]: Text(0, 0.5, 'Frequency')
```



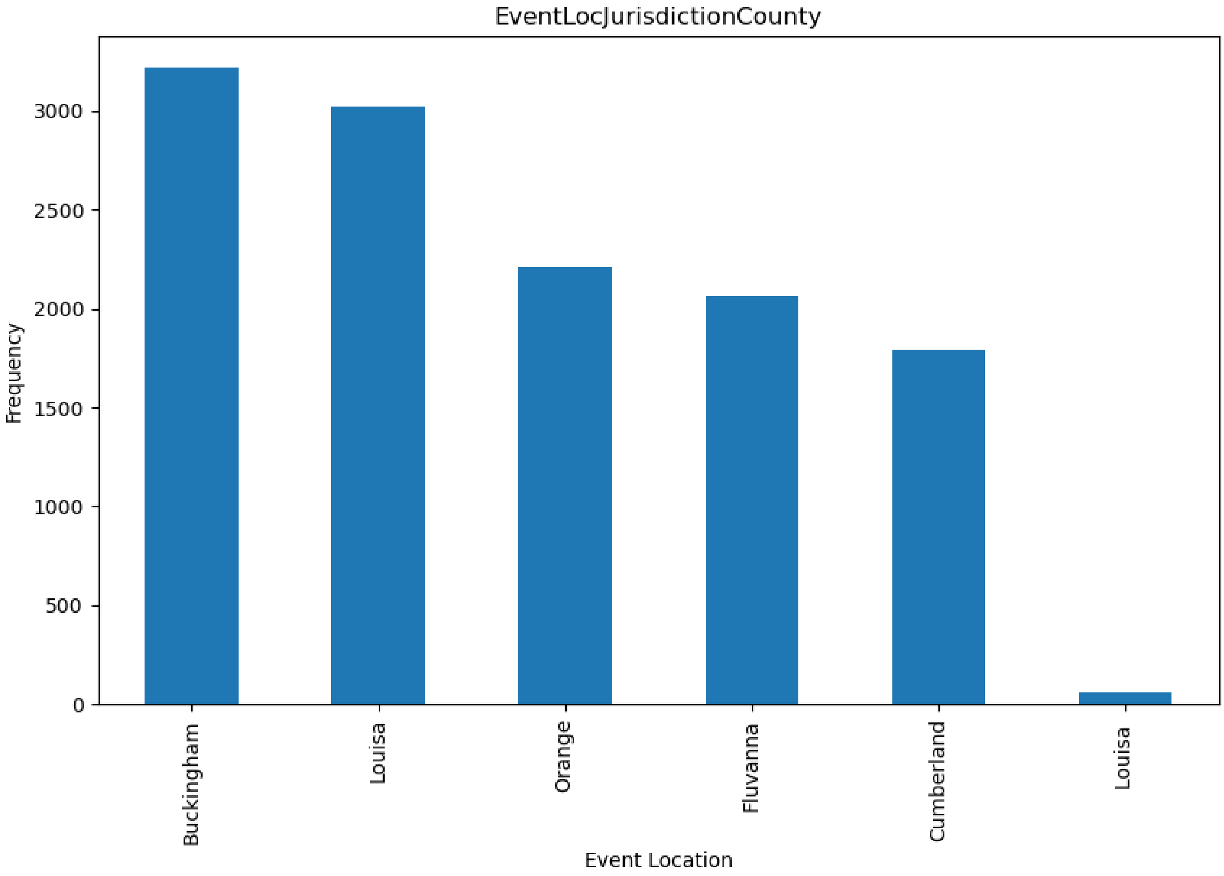
```
In [37]: ax = record_1782["PersonGivenNames"].value_counts().plot(kind='bar',
                                                                   figsize=(10,6),
                                                                   ylim=(0,75),
                                                                   title="PersonGivenNames")
         ax.set_xlabel("Given Names")
         ax.set_ylabel("Frequency")

Out[37]: Text(0, 0.5, 'Frequency')
```



Location-related Variables

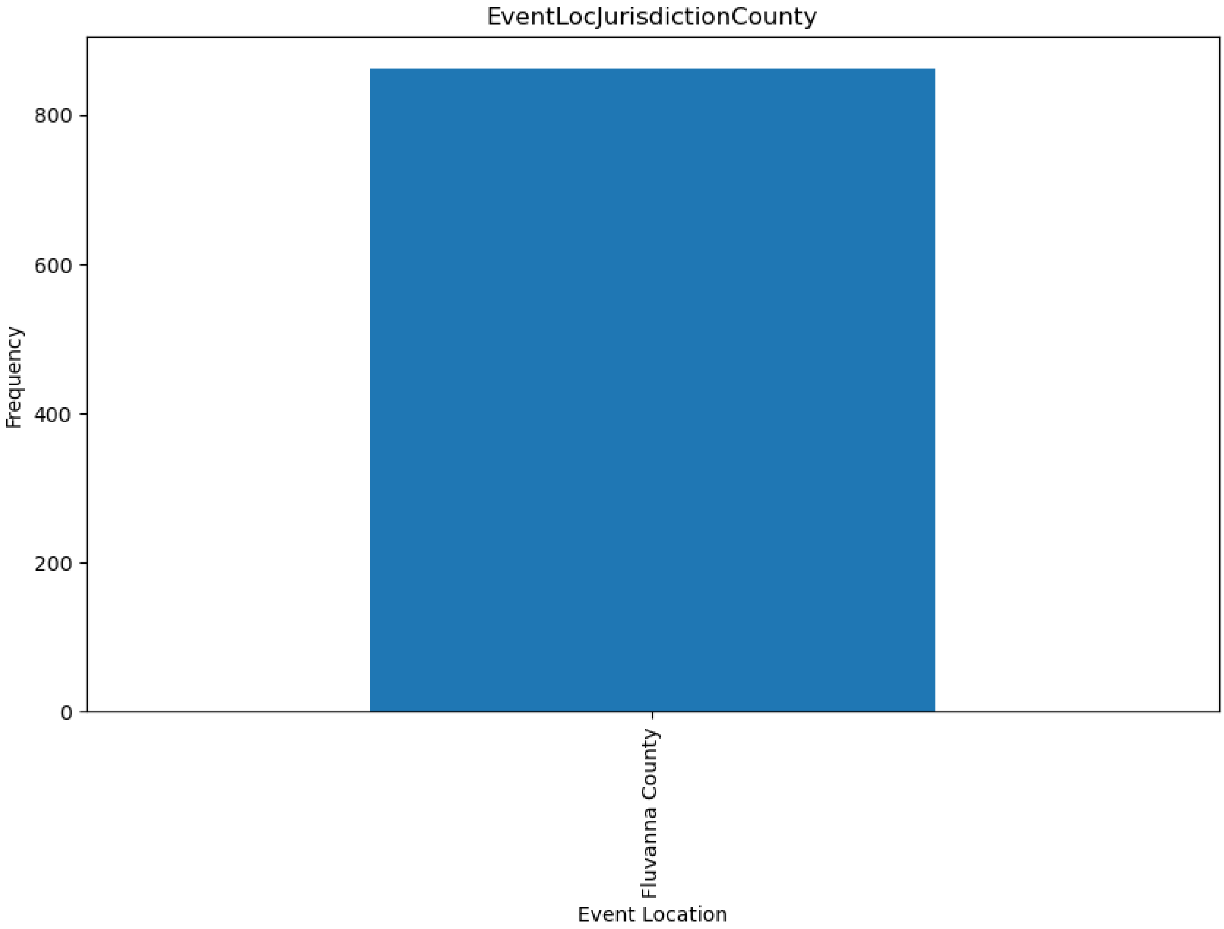
In [40]:
```
ax = record_1867['EventLocJurisdictionCounty'].value_counts().plot(kind='bar',
                                                      figsize=(10,6),
                                                      # ylim=(0,75),
                                                      title='EventLocJurisdictionCounty')
ax.set_xlabel("Event Location")
ax.set_ylabel("Frequency")
```

Out[40]:
```
Text(0, 0.5, 'Frequency')
```



```
In [41]: ax = record_1782['EventLocJurisdictionCounty'].value_counts().plot(kind='bar',
                                                                         figsize=(10,6),
                                                                         # ylim=(0,75),
                                                                         title='EventLocJurisdictionCounty')
         ax.set_xlabel("Event Location")
         ax.set_ylabel("Frequency")
```

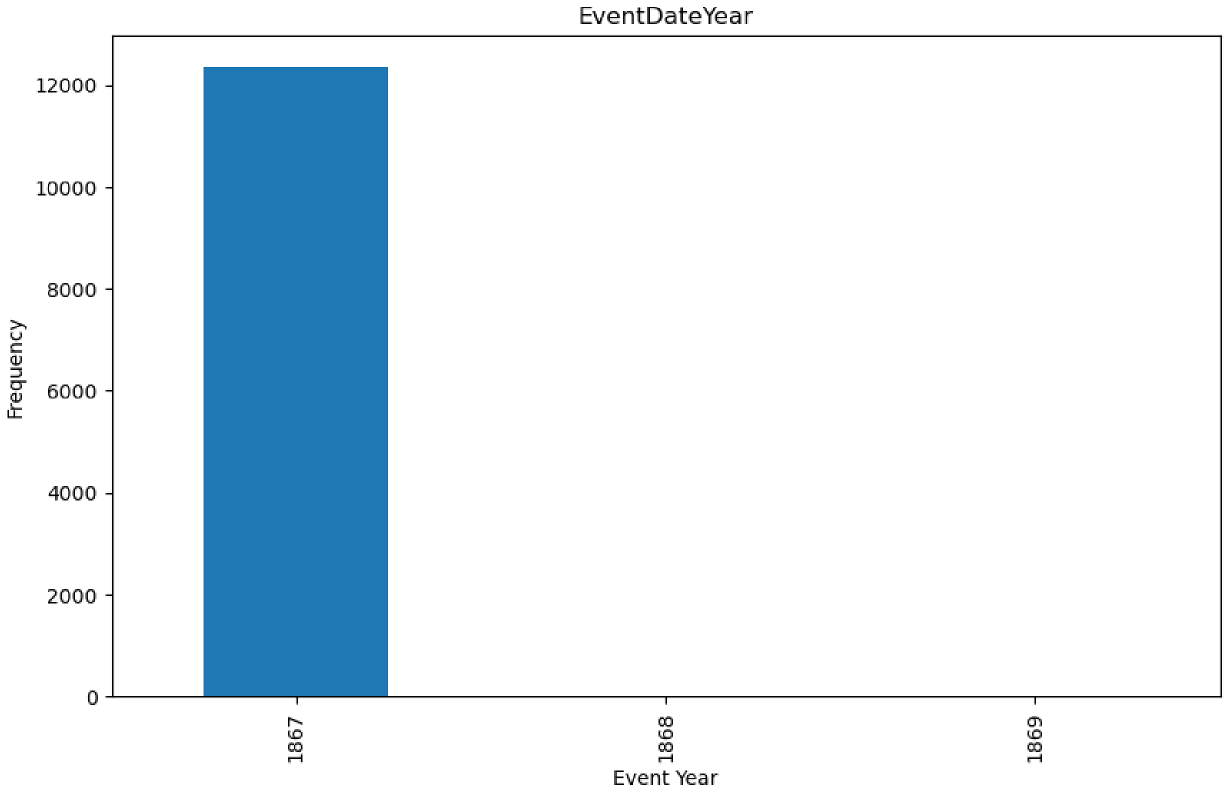
```
Out[41]: Text(0, 0.5, 'Frequency')
```



Time-related Variable

```
In [42]: ax = record_1867["EventDateYear"].value_counts().plot(kind='bar',
                                                    figsize=(10,6),
                                                    # ylim=(0,75),
                                                    title="EventDateYear")
         ax.set_xlabel("Event Year")
         ax.set_ylabel("Frequency")
```

```
Out[42]: Text(0, 0.5, 'Frequency')
```

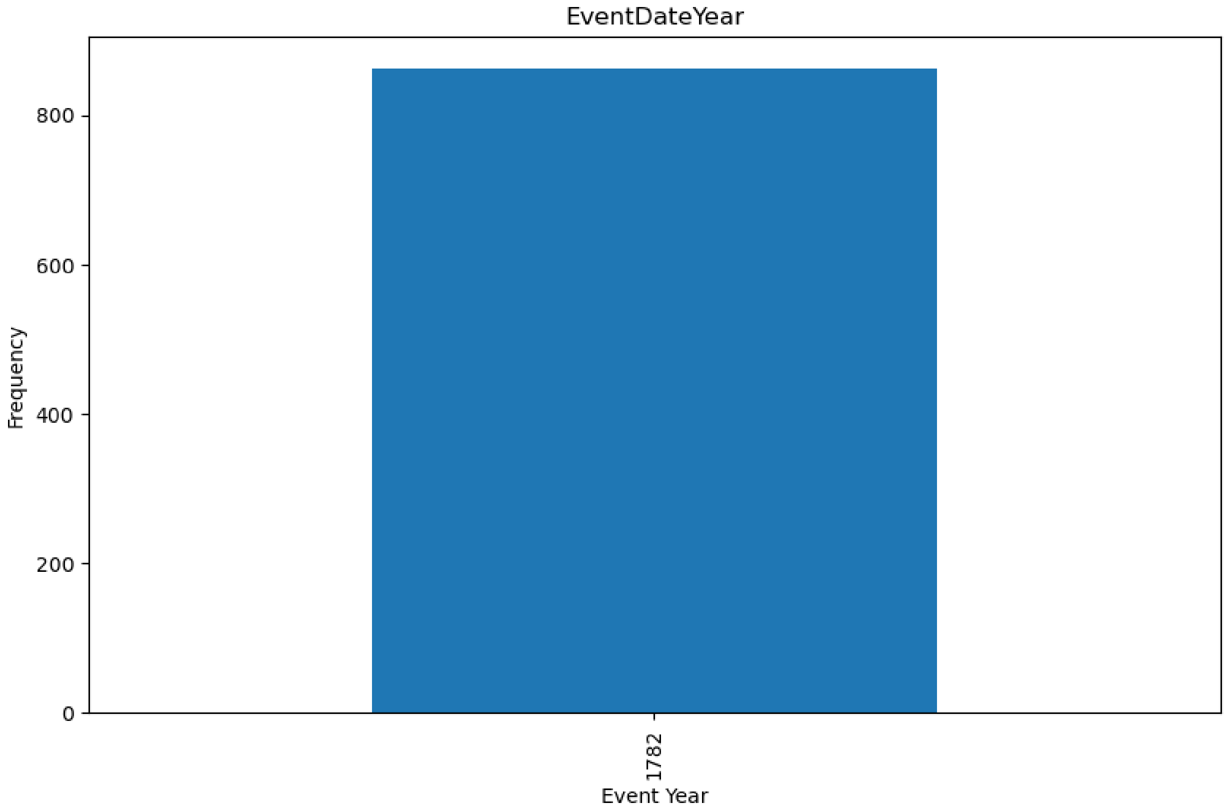


In [43]:

```
ax = record_1782["EventDateYear"].value_counts().plot(kind='bar',
                                    figsize=(10,6),
                                    # ylim=(0,75),
                                    title="EventDateYear")
ax.set_xlabel("Event Year")
ax.set_ylabel("Frequency")
```

Out[43]:

```
Text(0, 0.5, 'Frequency')
```


EventDateYear
Frequency
800
600
400
200
0
1782
Event Year