In [1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import re
import math
import matplotlib.pyplot as plt
import seaborn as sns

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the

import os
for dirname, _, filenames in os.walk('D:/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the curren
```

```
D:/kaggle/input\Alzheimer Disease and Healthy Aging Data In US.csv
```

load data

In [2]:

```
file_path = 'D:/kaggle/input/Alzheimer Disease and Healthy Aging Data In US.csv'
data = pd.read_csv(file_path)
data.columns
```

```
D:\anaconda3\envs\pytorch\lib\site-packages\IPython\core\interactiveshell.py:3441: D
typeWarning: Columns (13,14) have mixed types.Specify dtype option on import or set
low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[2]:

```
Index(['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Datasource',
       'Class', 'Topic', 'Question', 'Data_Value_Unit', 'DataValueTypeID',
       'Data_Value_Type', 'Data_Value', 'Data_Value_Alt',
       'Low_Confidence_Limit', 'High_Confidence_Limit', 'Sample_Size',
       'StratificationCategory1', 'Stratification1', 'StratificationCategory2',
       'Stratification2', 'Geolocation', 'ClassID', 'TopicID', 'QuestionID',
       'LocationID', 'StratificationCategoryID1', 'StratificationID1',
       'StratificationCategoryID2', 'StratificationID2'],
      dtype='object')
```

数据摘要

标称数据包括:
LocationAbbr;LocationDesc;Datasource;Class;Topic;Question;Data_Value_Unit;DataValueType;StratificationCate
其中有意义的包括:
LocationDesc;Class;Topic;Question;Data_Value_Type;DataValueTypeID;Stratification1;Stratification2 而StratificationCategory2与Datasource为unique 数值数据包括:
YearStart;YearEnd;Data_Value;Data_Value_Alt;Low_Confidence_Limit;High_Confidence_Limit;Sample_size 其中有意义的包括: Data_Value,YearStart.,YearEnd,Data_Value_Alt与Data_Value值完全重合。数据集中缺失Sample_size。

数据清洗

考虑到确实Sample_Size,因而Data_Value中以MEAN为量纲的数据无法使用，为统一数据集中的量纲，首先对数据集进行过滤

In [3]:

```
data.columns = data.columns.str.replace('-', '_').str.lower()
data = data.loc[data['data_value_type']=="Percentage"]
data=data[['yearstart','yearend','locationdesc','class','topic','question','data_value_type','data_v
data
```

Out[3]:

	yearstart	yearend	locationdesc	class	topic	question	data_value_type
0	2020	2020	Hawaii	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage
1	2017	2017	Idaho	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage
2	2017	2017	Idaho	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage
4	2020	2020	Indiana	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage
5	2020	2020	Iowa	Overall Health	Prevalence of sufficient sleep	Percentage of older adults getting sufficient ...	Percentage
...	...	...	...	...	...	...	...
214456	2018	2018	Wyoming	Screenings and Vaccines	Colorectal cancer screening	Percentage of older adults who had either a ho...	Percentage
214458	2015	2015	Wyoming	Smoking and Alcohol Use	Current smoking	Percentage of older adults who have smoked at ...	Percentage
214459	2017	2017	Wyoming	Overall Health	Self-rated health (fair to poor health)	Percentage of older adults who self-reported t...	Percentage
214460	2016	2016	Wyoming	Overall Health	Fall with injury within last year	Percentage of older adults who have fallen and...	Percentage

	yearstart	yearend	locationdesc	class	topic	question	data_value_type
214461	2018	2018	Wyoming	Smoking and Alcohol Use	Binge drinking within past 30 days	Percentage of older adults who reported binge ...	Percentage

In [4]:

```
nominals = ['locationdesc','class','topic','question','data_value_type','datavaluetypeid','stratific
numerics = ['data_value','yearstart','yearend']
# Convert column names into snake_case.
# data = data.iloc[[2]]

# Make views and dowloads numeric.
for col in ['data_value']:
#     data[col] = data[col].str.replace(',','')
    data[col] = data[col].astype('float')

# Output formte
pd.options.display.float_format = '{:.2f}'.format
data
```

Out[4]:

	yearstart	yearend	locationdesc	class	topic	question	data_value_type	d
0	2020	2020	Hawaii	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage	
1	2017	2017	Idaho	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage	
2	2017	2017	Idaho	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage	
4	2020	2020	Indiana	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage	
5	2020	2020	Iowa	Overall Health	Prevalence of sufficient sleep	Percentage of older adults getting sufficient ...	Percentage	
...	...	...	...	...	...	...	...	
214456	2018	2018	Wyoming	Screenings and Vaccines	Colorectal cancer screening	Percentage of older adults who had either a ho...	Percentage	
214458	2015	2015	Wyoming	Smoking and Alcohol Use	Current smoking	Percentage of older adults who have smoked at ...	Percentage	
214459	2017	2017	Wyoming	Overall Health	Self-rated health (fair to poor health)	Percentage of older adults who self-reported t...	Percentage	

	yearstart	yearend	locationdesc	class	topic	question	data_value_type	d
214460	2016	2016	Wyoming	Overall Health	Fall with injury within last year	Percentage of older adults who have fallen and...	Percentage	
214461	2018	2018	Wyoming	Smoking and Alcohol Use	Binge drinking within past 30 days	Percentage of older adults who reported binge ...	Percentage	

197929 rows × 12 columns

标称属性

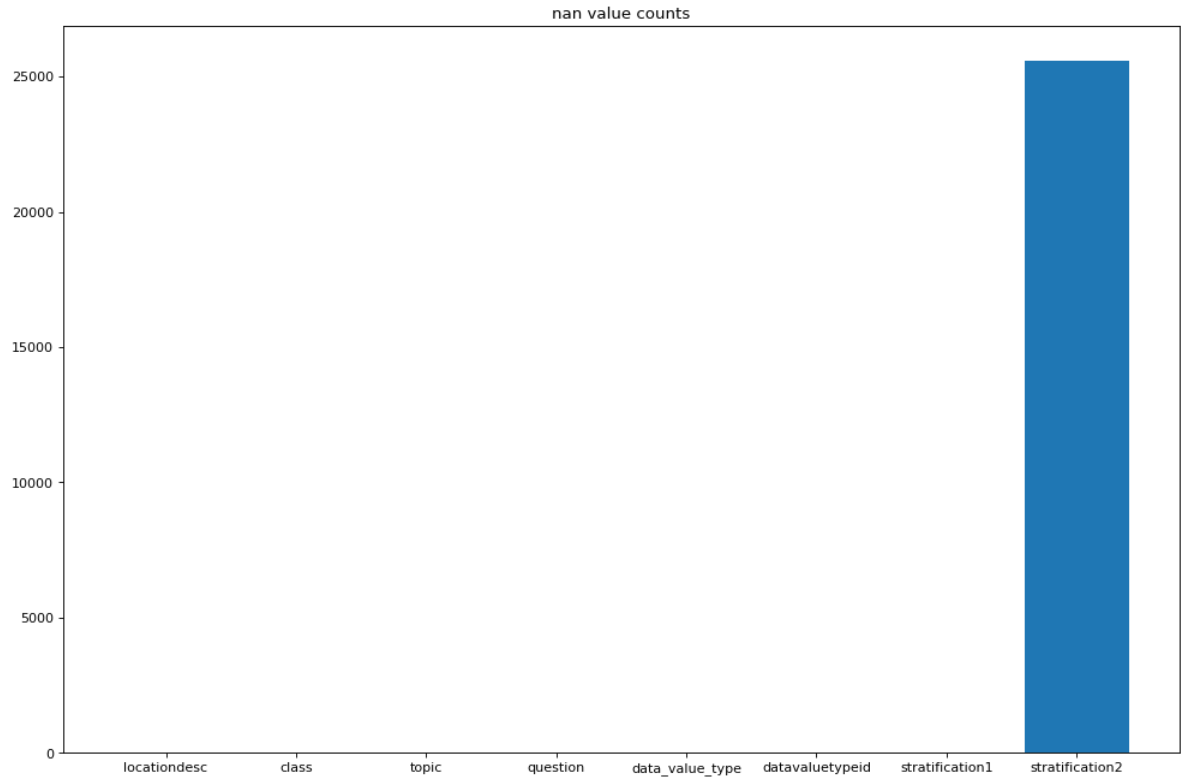
标称属性的缺失值的个数

In [5]:

```
ax = nominals
ay = []
plt.figure(figsize=(15,10), dpi=80)
for attr in nominals:
    freq = 5
    ay.append(data[attr].isna().sum())
plt.bar(ax, ay)
plt.title(f'nan value counts')
```

Out[5]:

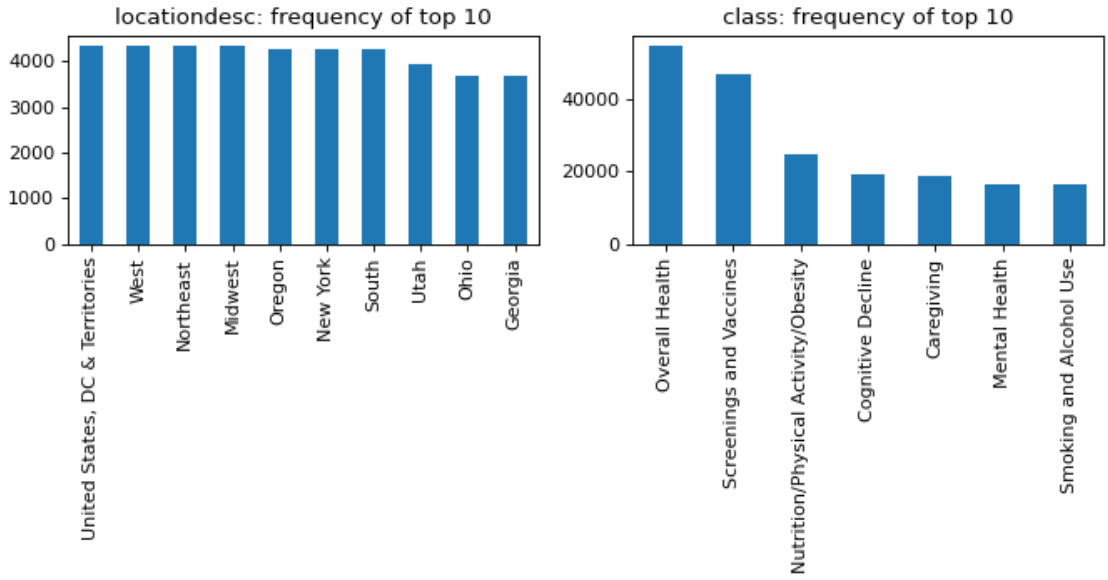
```
Text(0.5, 1.0, 'nan value counts')
```



标称属性的每个可能取值的频数 由.value_counts()取得，这里仅展示频度前五

In [6]:

```
index = 1
plt.figure(figsize=(10,60), dpi=80).subplots_adjust(hspace=6)
plt.figure(1)
col = 2
row = int(len(nominals) / col) + 1
for attr in nominals:
    plt.subplot(row, col, index)
    index += 1
    freq = 10
    data[attr].value_counts().head(freq).plot.bar()
    plt.title(f'{attr}: frequency of top {freq}')
```



数值属性

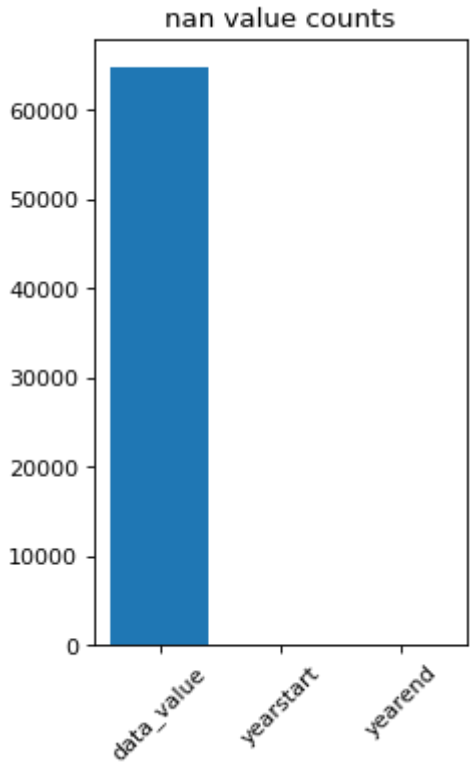
数值属性的缺失值个数

In [7]:

```
ax = range(len(numerics))
ay = []
plt.figure(figsize=(3,5), dpi=80)
for attr in numerics:
    freq = 5
    ay.append(data[attr].isna().sum())
plt.bar(ax, ay)
plt.xticks(ax, numerics, rotation=45)
plt.title(f'nan value counts')
```

Out[7]:

```
Text(0.5, 1.0, 'nan value counts')
```



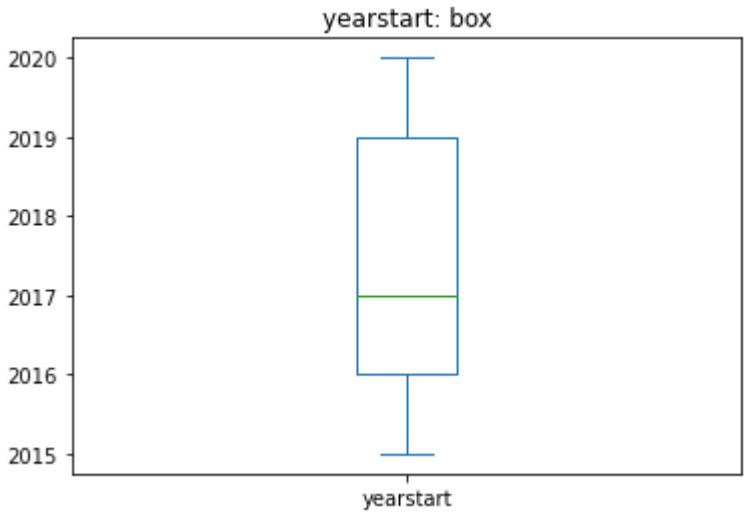
数据属性的五数、盒图

yearstart 处理时仅关注开始年份信息

In [8]:

```
attr = 'yearstart'
print(data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

```
count   197929.00
mean      2017.37
std          1.79
min       2015.00
25%       2016.00
50%       2017.00
75%       2019.00
max       2020.00
Name: yearstart, dtype: float64
```

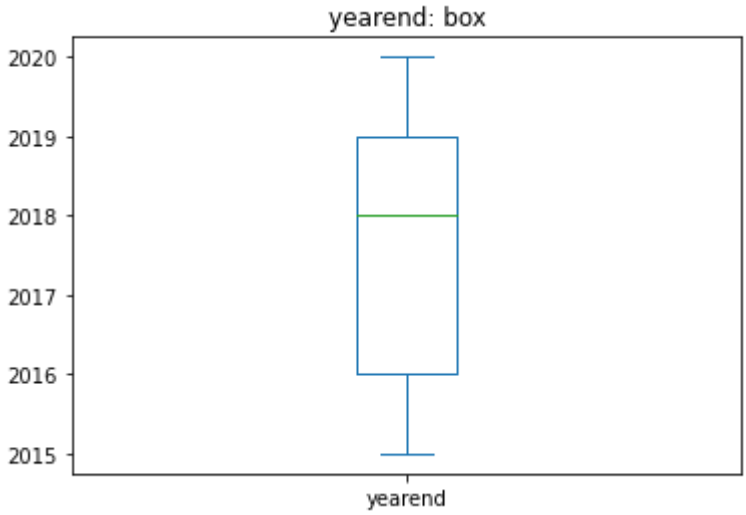


yearend
处理时仅关注终止年份信息

In [9]:

```
attr = 'yearend'
print(data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

```
count   197929.00
mean      2017.65
std          1.78
min       2015.00
25%       2016.00
50%       2018.00
75%       2019.00
max       2020.00
Name: yearend, dtype: float64
```

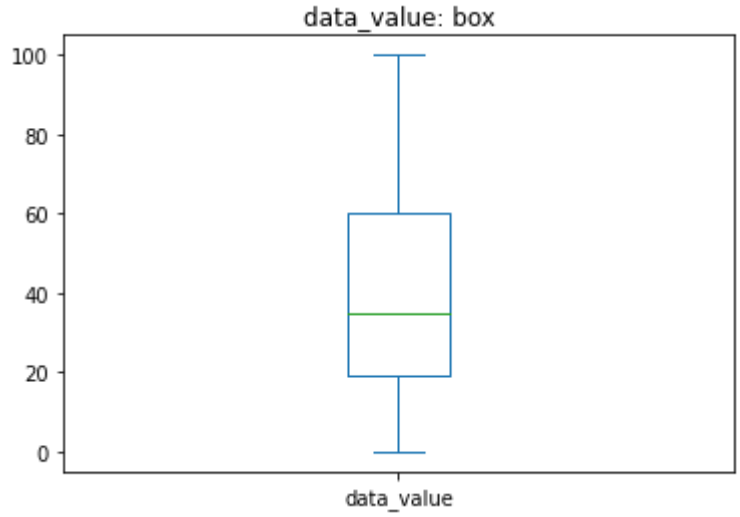


data value

In [10]:

```
attr = 'data_value'
print(data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

```
count   133262.00
mean        40.03
std         24.42
min          0.00
25%         19.30
50%         35.10
75%         60.00
max        100.00
Name: data_value, dtype: float64
```



缺失值处理

剔除 剔除value为空的数据后仅剩133262条数据 远少于原数据量197929

In [11]:

```
new_data = data[data['data_value'].notnull()]
new_data
```

Out[11]:

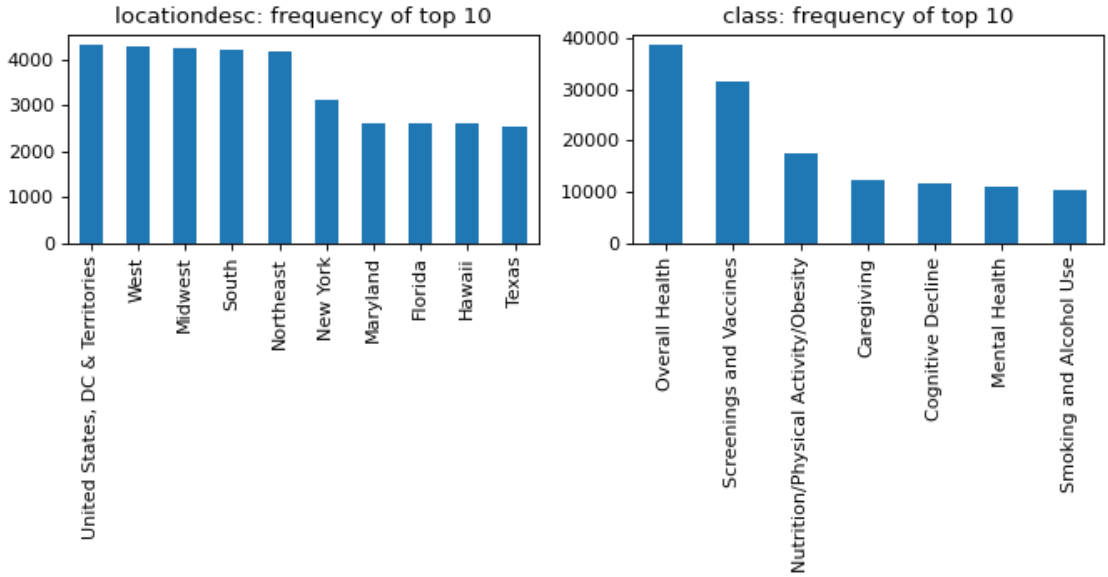
	yearstart	yearend	locationdesc	class	topic	question	data_value_type	d
0	2020	2020	Hawaii	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage	
1	2017	2017	Idaho	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage	
2	2017	2017	Idaho	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage	
4	2020	2020	Indiana	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage	
5	2020	2020	Iowa	Overall Health	Prevalence of sufficient sleep	Percentage of older adults getting sufficient ...	Percentage	
...	...	...	...	...	...	...	...	
214451	2015	2020	Wyoming	Caregiving	Provide care for someone with cognitive impair...	Percentage of older adults who provided care f...	Percentage	
214452	2016	2016	Wyoming	Overall Health	Self-rated health (good to excellent health)	Percentage of older adults who self-reported t...	Percentage	
214454	2015	2015	Wyoming	Cognitive Decline	Need assistance with day-to-day activities bec...	Percentage of older adults who reported that a...	Percentage	
214455	2015	2020	Wyoming	Cognitive Decline	Need assistance with day-to-day activities bec...	Percentage of older adults who reported that a...	Percentage	
214460	2016	2016	Wyoming	Overall Health	Fall with injury within last year	Percentage of older adults who have fallen and...	Percentage	

133262 rows × 12 columns

标称属性变化

In [14]:

```
index = 1
plt.figure(figsize=(10,60), dpi=80).subplots_adjust(hspace=6)
plt.figure(1)
col = 2
row = int(len(nominals) / col) + 1
for attr in nominals:
    plt.subplot(row, col, index)
    index += 1
    freq = 10
    new_data[attr].value_counts().head(freq).plot.bar()
    plt.title(f'{attr}: frequency of top {freq}')
```

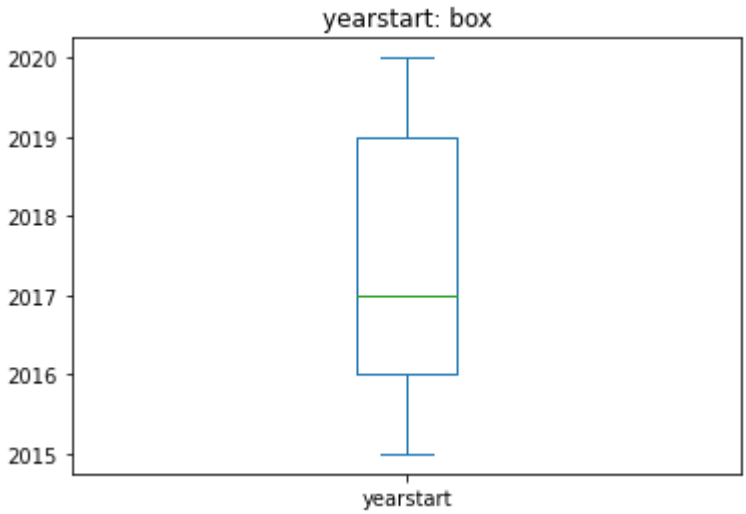


数值属性变化

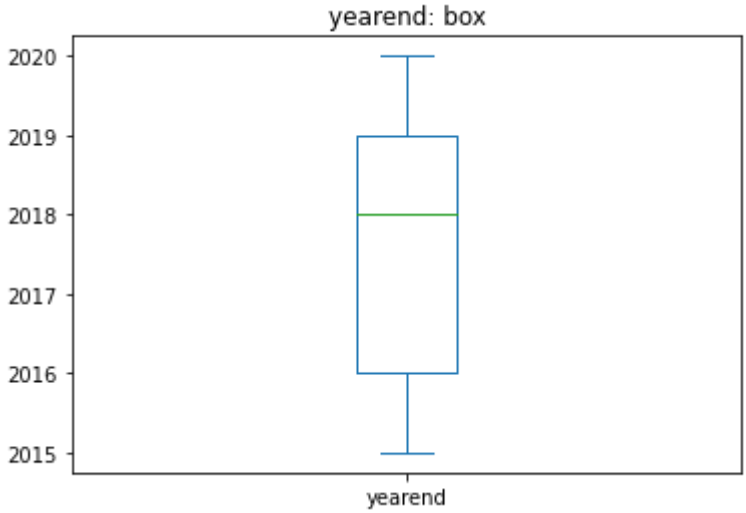
In [16]:

```
attr = 'yearstart'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'yearend'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'data_value'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

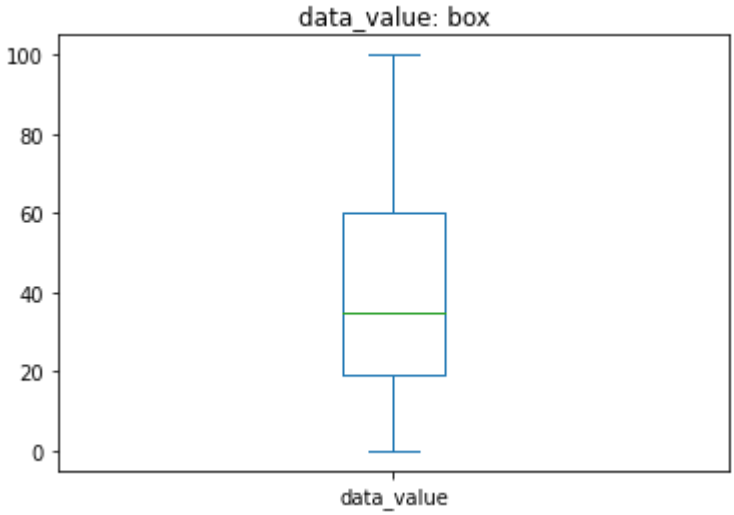
```
yearstart
count   133262.00
mean      2017.38
std          1.77
min       2015.00
25%       2016.00
50%       2017.00
75%       2019.00
max       2020.00
Name: yearstart, dtype: float64
```



```
yearend
count   133262.00
mean      2017.63
std          1.77
min       2015.00
25%       2016.00
50%       2018.00
75%       2019.00
max       2020.00
Name: yearend, dtype: float64
```




```
data_value
count   133262.00
mean        40.03
std         24.42
min          0.00
25%         19.30
50%         35.10
75%         60.00
max        100.00
Name: data_value, dtype: float64
```



最高频率值填补

In [17]:

```
attrs = nominals + numerics
new_data = data.copy(deep=True)
for attr in attrs:
    most = data[attr].value_counts().index[0]
    new_data[attr] = data[attr].fillna(most)
new_data
```

Out[17]:

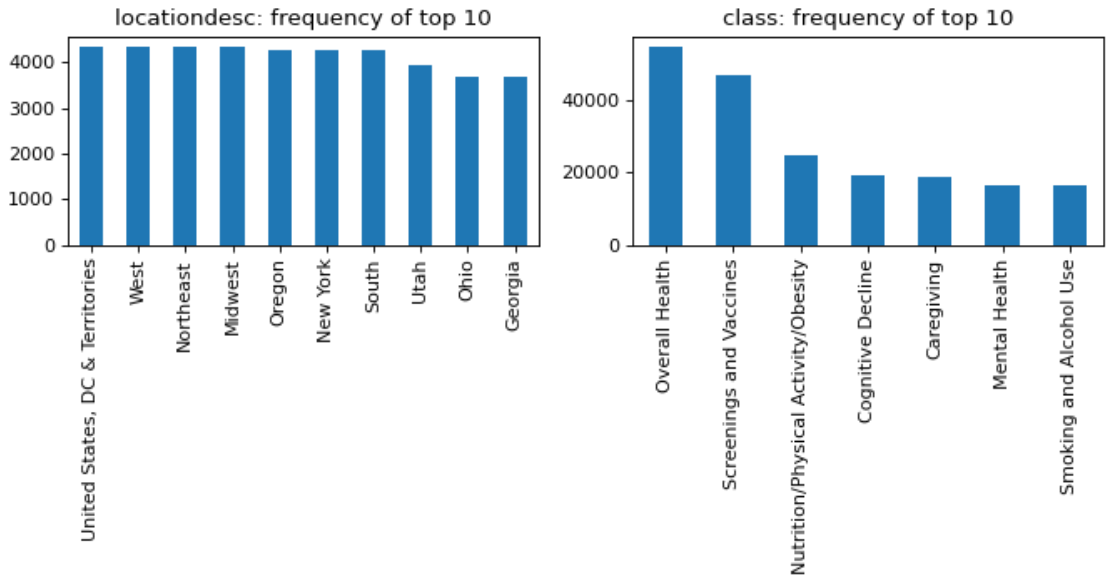
	yearstart	yearend	locationdesc	class	topic	question	data_value_type
0	2020	2020	Hawaii	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage
1	2017	2017	Idaho	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage
2	2017	2017	Idaho	Overall Health	Arthritis among older adults	Percentage of older adults ever told they have...	Percentage
4	2020	2020	Indiana	Mental Health	Lifetime diagnosis of depression	Percentage of older adults with a lifetime dia...	Percentage
5	2020	2020	Iowa	Overall Health	Prevalence of sufficient sleep	Percentage of older adults getting sufficient ...	Percentage
...	...	...	...	...	...	...	...
214456	2018	2018	Wyoming	Screenings and Vaccines	Colorectal cancer screening	Percentage of older adults who had either a ho...	Percentage
214458	2015	2015	Wyoming	Smoking and Alcohol Use	Current smoking	Percentage of older adults who have smoked at ...	Percentage
214459	2017	2017	Wyoming	Overall Health	Self-rated health (fair to poor health)	Percentage of older adults who self-reported t...	Percentage
214460	2016	2016	Wyoming	Overall Health	Fall with injury within last year	Percentage of older adults who have fallen and...	Percentage

	yearstart	yearend	locationdesc	class	topic	question	data_value_type
214461	2018	2018	Wyoming	Smoking and Alcohol Use	Binge drinking within past 30 days	Percentage of older adults who reported binge ...	Percentage

标称属性变化

In [18]:

```
index = 1
plt.figure(figsize=(10,60), dpi=80).subplots_adjust(hspace=6)
plt.figure(1)
col = 2
row = int(len(nominals) / col) + 1
for attr in nominals:
    plt.subplot(row, col, index)
    index += 1
    freq = 10
    new_data[attr].value_counts().head(freq).plot.bar()
    plt.title(f'{attr}: frequency of top {freq}')
```

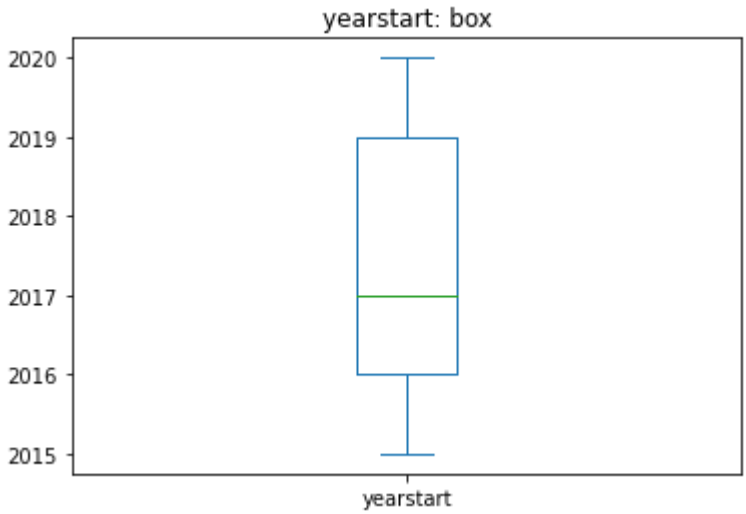


数值属性变化

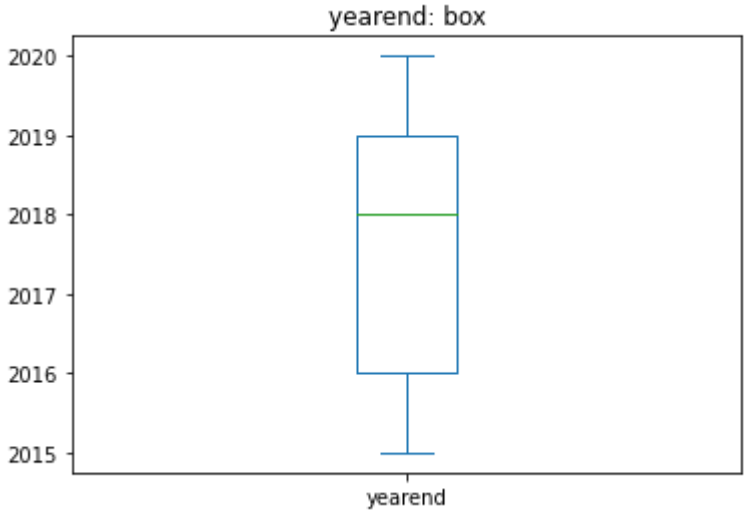
In [19]:

```
attr = 'yearstart'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'yearend'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'data_value'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

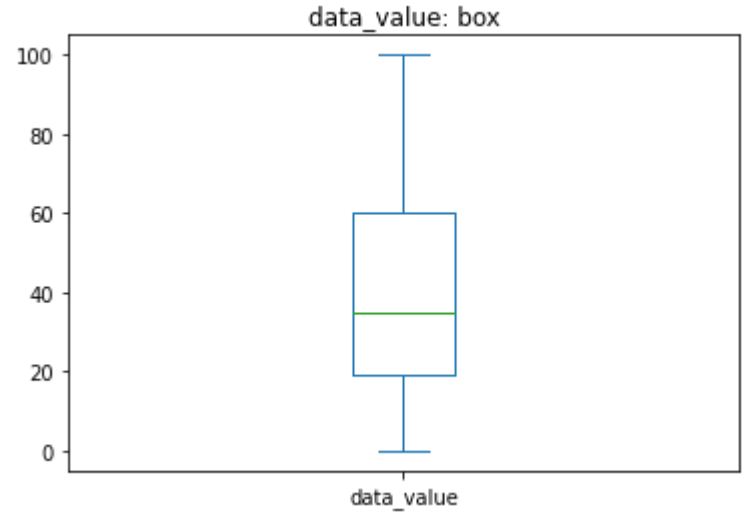
```
yearstart
count   197929.00
mean      2017.37
std          1.79
min       2015.00
25%       2016.00
50%       2017.00
75%       2019.00
max       2020.00
Name: yearstart, dtype: float64
```



```
yearend
count   197929.00
mean      2017.65
std          1.78
min       2015.00
25%       2016.00
50%       2018.00
75%       2019.00
max       2020.00
Name: yearend, dtype: float64
```




```
data_value
count   197929.00
mean        30.57
std         24.20
min          0.00
25%         11.10
50%         19.80
75%         44.70
max        100.00
Name: data_value, dtype: float64
```



相关关系填补，可见data_value与调查的起始年份相关性较小，因此无法使用改方法填补缺失值

In [20]:

```
new_data = data.copy(deep=True)
corr_matrix = new_data.corr()
corr_matrix
```

Out[20]:

	yearstart	yearend	data_value
yearstart	1.00	0.79	0.05
yearend	0.79	1.00	0.00
data_value	0.05	0.00	1.00

基于相似性 利用impyute工具 对几个数值属性进行填补

In [21]:

```
pip install impyute
```

```
Requirement already satisfied: impyute in d:\anaconda3\envs\pytorch\lib\site-package
s (0.0.8)
Requirement already satisfied: scikit-learn in d:\anaconda3\envs\pytorch\lib\site-pa
ckages (from impyute) (1.0.1)
Requirement already satisfied: scipy in d:\anaconda3\envs\pytorch\lib\site-packages
(from impyute) (1.7.1)
Requirement already satisfied: numpy in d:\anaconda3\envs\pytorch\lib\site-packages
(from impyute) (1.21.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in d:\anaconda3\envs\pytorch\lib
\site-packages (from scikit-learn->impyute) (3.0.0)
Requirement already satisfied: joblib>=0.11 in d:\anaconda3\envs\pytorch\lib\site-pa
ckages (from scikit-learn->impyute) (1.1.0)
Note: you may need to restart the kernel to use updated packages.
```

In [22]:

```
from impyute import fast_knn
features = ['yearstart', 'yearend','data_value']
new_data = data.copy(True)
new_data[features] = pd.DataFrame(fast_knn(np.array(new_data[features]), k=2), columns=features)
new_data.isnull().any()
```

```
D:\anaconda3\envs\pytorch\lib\site-packages\impyute\imputation\cs\fast_knn.py:113: R
untimeWarning: invalid value encountered in true_divide
  weights = distances/np.sum(distances)
```

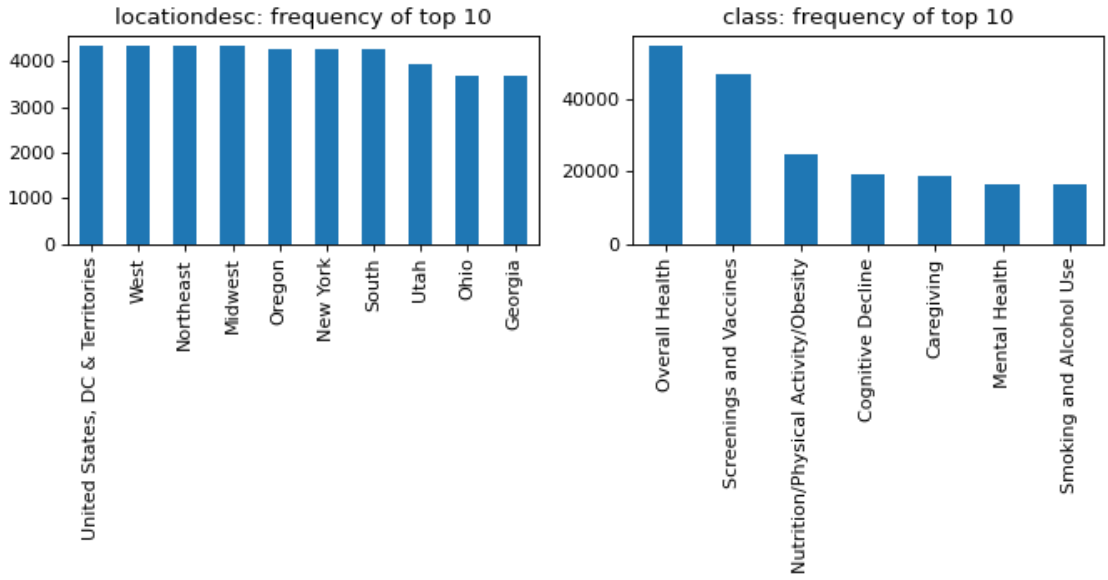
Out[22]:

```
yearstart                 True
yearend                   True
locationdesc             False
class                    False
topic                    False
question                 False
data_value_type          False
data_value                True
datavaluetypeid          False
stratification1          False
stratificationcategory2   True
stratification2           True
dtype: bool
```

标称属性变化

In [23]:

```
index = 1
plt.figure(figsize=(10,60), dpi=80).subplots_adjust(hspace=6)
plt.figure(1)
col = 2
row = int(len(nominals) / col) + 1
for attr in nominals:
    plt.subplot(row, col, index)
    index += 1
    freq = 10
    new_data[attr].value_counts().head(freq).plot.bar()
    plt.title(f'{attr}: frequency of top {freq}')
```

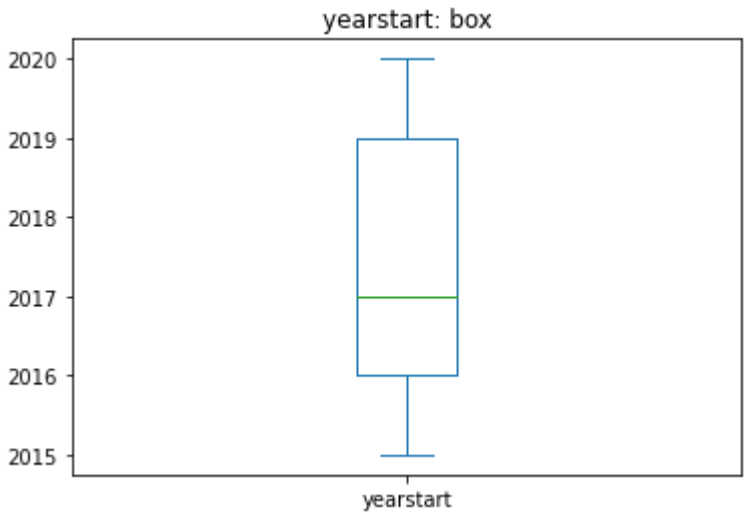


数值属性变化

In [24]:

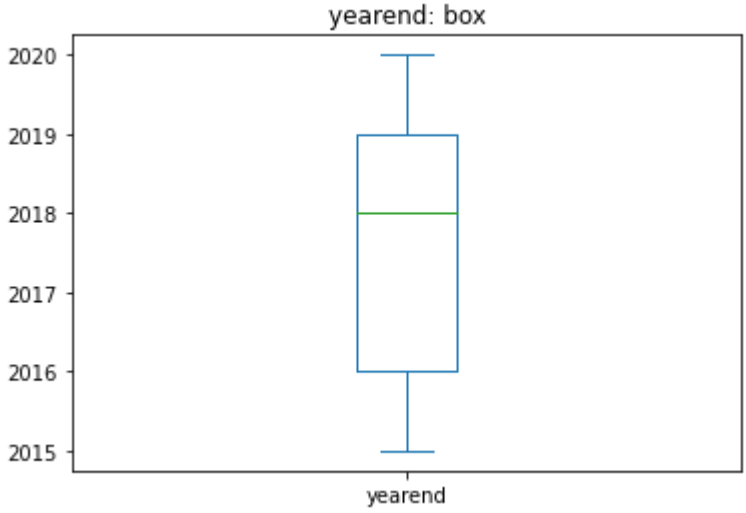
```
attr = 'yearstart'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'yearend'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
attr = 'data_value'
print(attr)
print(new_data[attr].describe())
visit = pd.DataFrame(data[attr])
visit.plot.box()
plt.title(f'{attr}: box')
plt.show()
```

```
yearstart
count   182664.00
mean      2017.37
std          1.79
min       2015.00
25%       2016.00
50%       2017.00
75%       2019.00
max       2020.00
Name: yearstart, dtype: float64
```

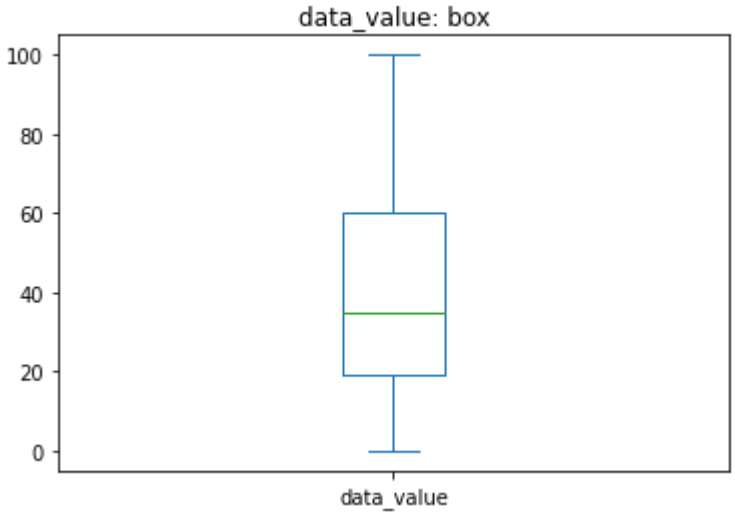



```
yearend
count   182664.00
mean      2017.65
std          1.78
min       2015.00
25%       2016.00
50%       2018.00
75%       2019.00
```

```
max       2020.00
Name: yearend, dtype: float64
```




```
data_value
count   122975.00
mean        40.01
std         24.43
min          0.00
25%         19.30
50%         35.10
75%         60.00
max        100.00
Name: data_value, dtype: float64
```



In [ ]: