Alzheimer Disease and Healthy Aging Data in US的数据分析与处理

1. 摘要分析和可视化

摘要分析

首先编写分析数据摘要的函数，如下。其中`calc_5num_abstract()`函数用来计算数值属性的五数概括，`analyze_abstract_github()`用来进行数据摘要的提取，之后`print_abstract()`函数用来在控制台打印分析结果，`save_abstract()`用于将分析结果保存为.csv格式。

In [1]:
```
import pandas as pd
import numpy as np
import os


def calc_5num_abstract(arr: np.ndarray):
    n_max = np.nanmax(arr)
    n_min = np.nanmin(arr)
    q1 = np.percentile(arr, 25)
    median = np.nanmedian(arr)
    q3 = np.percentile(arr, 75)
    return n_min, q1, median, q3, n_max


def analyze_abstract_github(df: pd.DataFrame) -> dict:
    print("analyzing data")
    analyze_results = {}
    for col in df.columns:
        if df[col].dtype == object:
            count = df[col].value_counts()
            df_meta = {col: count.index, "count": count.values}
            analyze_results[col] = pd.DataFrame(df_meta)
        else:
            values = df[col].values
            mask = np.isnan(values)
            if len(values[np.logical_not(mask)]) == 0:
                print(f"The column {col} has no data.")
                continue
            n_min, q1, median, q3, n_max = calc_5num_abstract(values[np.logical_
            null_calc = df[col].isnull().sum()
            res = pd.DataFrame(
                {"min": [n_min], "Q1": [q1], "median": [median], "Q3": [q3], "ma
            analyze_results[col] = res
    return analyze_results


def print_abstract(result):
    for k, v in result.items():
        print(k, v.to_string(index=False, max_rows=20), sep="\n")
        print()
```

```
def save_abstract(path, result):
    os.makedirs(path, exist_ok=True)
    for k, v in result.items():
        v.to_csv(os.path.join(path, f"{k}.csv"), index=False)
```

下面的代码用来实现摘要分析。

In [2]:
```
print("Analyzing Alzheimer Disease and Healthy Aging Data In US.csv")
df_alzheimer_data = pd.read_csv("data/alzheimer/Alzheimer Disease and Healthy Ag
res = analyze_abstract_github(df_alzheimer_data)
result_dir = "result/alzheimer/"
print_abstract(res)
save_abstract(result_dir, res)
```

```
Analyzing Alzheimer Disease and Healthy Aging Data In US.csv
C:\Users\LEGION\AppData\Local\Temp\ipykernel_22900\265366865.py:2: DtypeWarnin
g: Columns (13,14) have mixed types. Specify dtype option on import or set low_
memory=False.
  df_alzheimer_data = pd.read_csv("data/alzheimer/Alzheimer Disease and Healthy
Aging Data In US.csv")
```

```
analyzing data
The column Sample_Size has no data.
YearStart
 min     Q1  median     Q3  max  blanks
2015 2016.0  2017.0 2019.0 2020       0

YearEnd
 min     Q1  median     Q3  max  blanks
2015 2016.0  2018.0 2019.0 2020       0

LocationAbbr
LocationAbbr  count
          US   4644
        WEST   4638
         NRE   4614
         MDW   4611
          OR   4565
          NY   4557
         SOU   4542
          UT   4222
          OH   3955
          GA   3951
         ...    ...
          NC   3349
          WA   3348
          MT   3348
          DE   3346
          NH   3284
          VT   3278
          MA   3174
          PR   2797
          GU   2703
          VI    503

LocationDesc
                   LocationDesc  count
United States, DC & Territories   4644
                           West   4638
                      Northeast   4614
                        Midwest   4611
                         Oregon   4565
                       New York   4557
                          South   4542
                           Utah   4222
                           Ohio   3955
                        Georgia   3951
                            ...    ...
                 North Carolina   3349
                     Washington   3348
                        Montana   3348
                       Delaware   3346
                  New Hampshire   3284
                        Vermont   3278
                  Massachusetts   3174
                    Puerto Rico   2797
                            Guam   2703
                 Virgin Islands    503

Datasource
Datasource  count
```

BRFSS 214462

Class

Class	count
Overall Health	71694
Screenings and Vaccines	46867
Nutrition/Physical Activity/Obesity	24851
Cognitive Decline	19180
Caregiving	18671
Mental Health	16600
Smoking and Alcohol Use	16599

Topic

Topic	count
Obesity	8300
Influenza vaccine within past year	8300
Physically unhealthy days (mean number of days)	8300
Frequent mental distress	8300
Current smoking	8300
Lifetime diagnosis of depression	8300
No leisure-time physical activity within past month	8300
Self-rated health (fair to poor health)	8299
Self-rated health (good to excellent health)	8299
Binge drinking within past 30 days	8299
...	...
Severe joint pain among older adults with arthritis	4064
Provide care for a friend or family member in past month	3848
Expect to provide care for someone in the next two years	3797
Provide care for someone with cognitive impairment within the past month	3682
Duration of caregiving among older adults	3681
Intensity of caregiving among older adults	3663
Up-to-date with recommended vaccines and screenings - Women	3280
Up-to-date with recommended vaccines and screenings - Men	3271
Mammogram within past 2 years	3271
Pap test within past 3 years	3242

Question

Question	count
Percentage of older adults who are currently obese, with a body mass index (BMI) of 30 or more	8300
Percentage of older adults who reported influenza vaccine within the past year	8300
Physically unhealthy days (mean number of days in past month)	8300
Percentage of older adults who are experiencing frequent mental distress	8300
Percentage of older adults who have smoked at least 100 cigarettes in their entire life and still smoke every day or some days	8300
Percentage of older adults with a lifetime diagnosis of depression	8300
Percentage of older adults who have not had any leisure time physical activity in the past month	8300
Percentage of older adults who self-reported that their health is "fair" or "poor"	8299
Percentage of older adults who self-reported that their health is "good", "very good", or "excellent"	8299
Percentage of older adults who reported binge drinking within the past 30 days	8299

```
...    ...
                                                    Severe joint pain due to ar
thritis among older adults with doctor-diagnosed arthritis   4064
                                          Percentage of older adults who provide
d care for a friend or family member within the past month   3848
Percentage of older adults currently not providing care who expect to provide c
are for someone with health problems in the next two years   3797
               Percentage of older adults who provided care for someone with de
mentia or other cognitive impairment within the past month   3682
                                          Percentage of older adults who provide
d care to a friend or family member for six months or more   3681
                                             Average of 20 or more ho
urs of care per week provided to a friend or family member   3663
                                         Percentage of older adult women w
ho are up to date with select clinical preventive services   3280
                                           Percentage of older adult men w
ho are up to date with select clinical preventive services   3271
                                                  Percentage of older adult w
omen who have received a mammogram within the past 2 years   3271
                                          Percentage of older adult women with a
n intact cervix who had a Pap test within the past 3 years   3242

Data_Value_Unit
Data_Value_Unit  count
              % 197929
         Number  16533

DataValueTypeID
DataValueTypeID  count
          PRCTG 197929
           MEAN  16533

Data_Value_Type
Data_Value_Type  count
     Percentage 197929
           Mean  16533

Data_Value
 min   Q1  median   Q3   max  blanks
 0.0 15.3    32.5 56.8 100.0   69833

Data_Value_Alt
 min   Q1  median   Q3   max  blanks
 0.0 15.3    32.5 56.8 100.0   69833

Low_Confidence_Limit
Low_Confidence_Limit  count
                 5.4    350
                 5.1    318
                 4.8    314
                 5.3    313
                   5    305
                 4.7    298
                 4.9    298
                 4.6    296
                 5.6    296
                 5.7    292
                 ...    ...
                99.6      1
                98.1      1
```

```
                98.1      1
                99.5      1
                98.8      1
                96.8      1
                98.4      1
                 0.9      1
                97.8      1
                99.4      1

High_Confidence_Limit
High_Confidence_Limit  count
                  6.5    216
                  5.8    193
                  6.8    192
                  6.7    189
                  7.5    186
                  5.5    186
                  6.6    185
                  6.9    177
                    6    177
                  6.4    177
                  ...    ...
                  100      2
                  2.3      2
                  1.4      2
                  1.9      1
                  2.5      1
                  1.7      1
                  2.8      1
                  2.8      1
                  1.5      1
                  1.6      1

StratificationCategory1
StratificationCategory1  count
              Age Group 214462

Stratification1
  Stratification1  count
          Overall  71919
      50-64 years  71528
65 years or older  71015

StratificationCategory2
StratificationCategory2  count
         Race/Ethnicity 134959
                 Gender  51834

Stratification2
         Stratification2  count
     White, non-Hispanic  27633
                Hispanic  27525
     Black, non-Hispanic  26968
Native Am/Alaskan Native  26571
  Asian/Pacific Islander  26262
                  Female  26091
                    Male  25743

Geolocation
                     Geolocation  count
```

```
POINT (-120.1550313 44.56744942)   4565
POINT (-75.54397043 42.82700103)   4557
POINT (-111.5871306 39.36070017)   4222
POINT (-82.40426006 40.06021014)   3955
POINT (-83.62758035 32.83968109)   3951
POINT (-76.60926011 39.29058096)   3919
POINT (-157.8577494 21.30485044)   3907
POINT (-85.77449091 35.68094058)   3879
POINT (-84.71439027 44.66131954)   3796
POINT (-78.45789046 37.54268067)   3758
                                ...    ...
POINT (-79.15925046 35.46622098)   3349
POINT (-109.4244206 47.06652897)   3348
POINT (-120.4700108 47.52227863)   3348
POINT (-75.57774117 39.00883067)   3346
POINT (-71.50036092 43.65595011)   3284
POINT (-72.51764079 43.62538124)   3278
POINT (-72.08269067 42.27687047)   3174
    POINT (-66.590149 18.220833)   2797
    POINT (144.793731 13.444304)   2703
    POINT (-64.896335 18.335765)    503

ClassID
ClassID  count
    C01  71694
    C03  46867
    C02  24851
    C06  19180
    C07  18671
    C05  16600
    C04  16599

TopicID
TopicID  count
  TNC04   8300
  TSC08   8300
  TOC01   8300
  TMC01   8300
  TAC01   8300
  TMC03   8300
  TNC03   8300
  TOC07   8299
  TOC08   8299
  TAC03   8299
    ...    ...
  TOC12   4064
  TGC01   3848
  TGC02   3797
  TGC05   3682
  TGC03   3681
  TGC04   3663
  TSC11   3280
  TSC10   3271
  TSC01   3271
  TSC03   3242

QuestionID
QuestionID  count
       Q13   8300
       Q18   8300
```

```
       Q08   8300
       Q03   8300
       Q17   8300
       Q27   8300
       Q16   8300
       Q32   8299
       Q33   8299
       Q21   8299
       ...    ...
       Q44   4064
       Q36   3848
       Q37   3797
       Q40   3682
       Q38   3681
       Q39   3663
       Q11   3280
       Q10   3271
       Q12   3271
       Q20   3242

LocationID
 min   Q1  median   Q3  max  blanks
   1 18.0    33.0 49.0 9004       0

StratificationCategoryID1
StratificationCategoryID1  count
                      AGE 214462

StratificationID1
StratificationID1  count
      AGE_OVERALL  71919
             5064  71528
           65PLUS  71015

StratificationCategoryID2
StratificationCategoryID2  count
                     RACE 134959
                   GENDER  51834
                  OVERALL  27669

StratificationID2
StratificationID2  count
          OVERALL  27669
              WHT  27633
              HIS  27525
              BLK  26968
              NAA  26571
              ASN  26262
           FEMALE  26091
             MALE  25743
```

运行结束后，可以在`./result/alzheimer`目录中找到摘要分析结果。对于标称数据，表格包含两列：可能取值及其频数；对于数值数据，表格中包含6列，分别是`min`（最小值）、`Q1`（第一四分位数）、`median`（中位数）、`Q3`（第三四分位数）、`max`（最大值）、`blanks`（缺失数据个数。

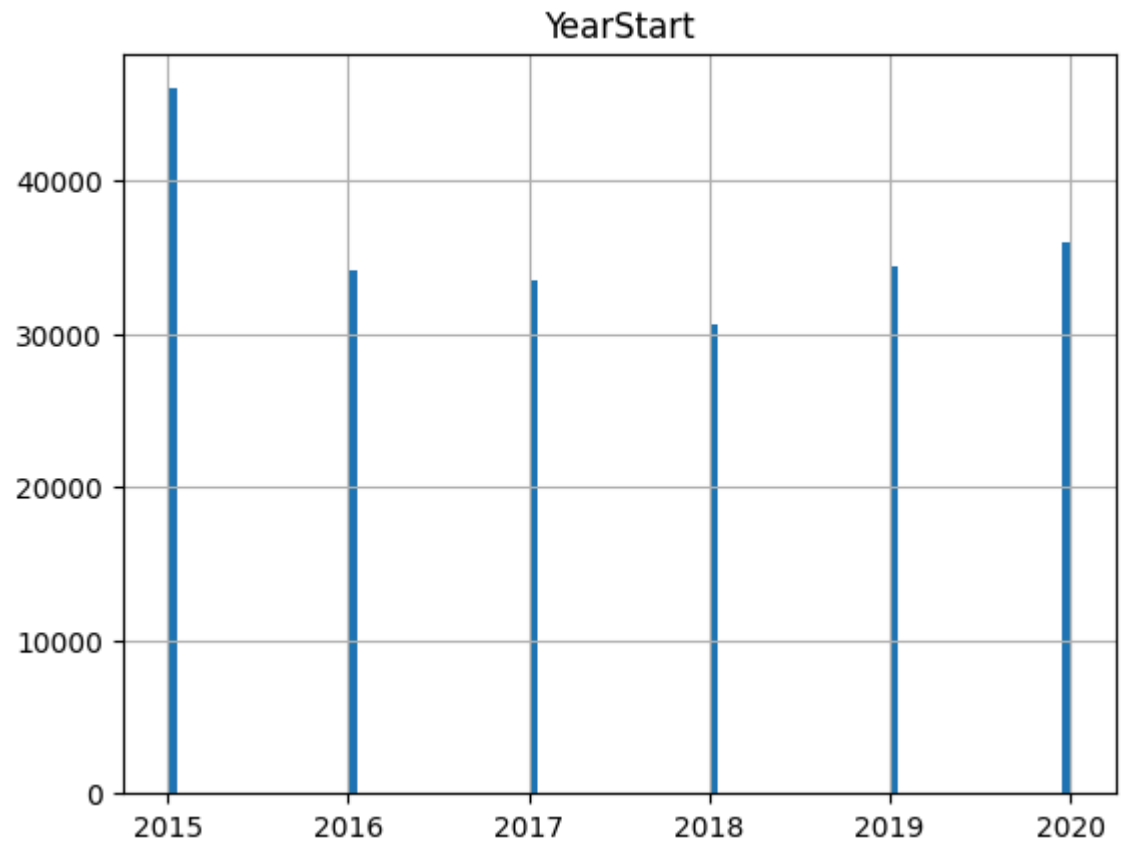
数据可视化

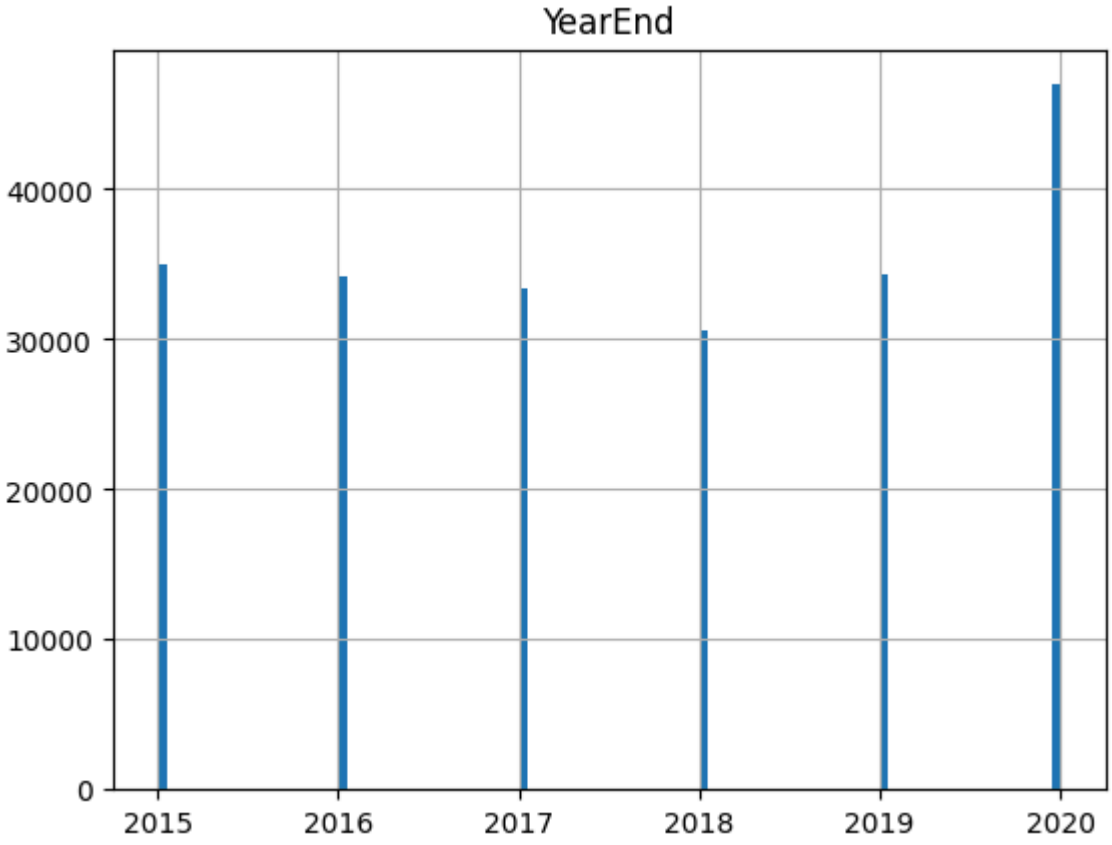
下面对Alzheimer Disease and Healthy Aging Data In US.csv的数据进行可视化，包含直方图、盒图。下面分别对两种图编写可视化函数。

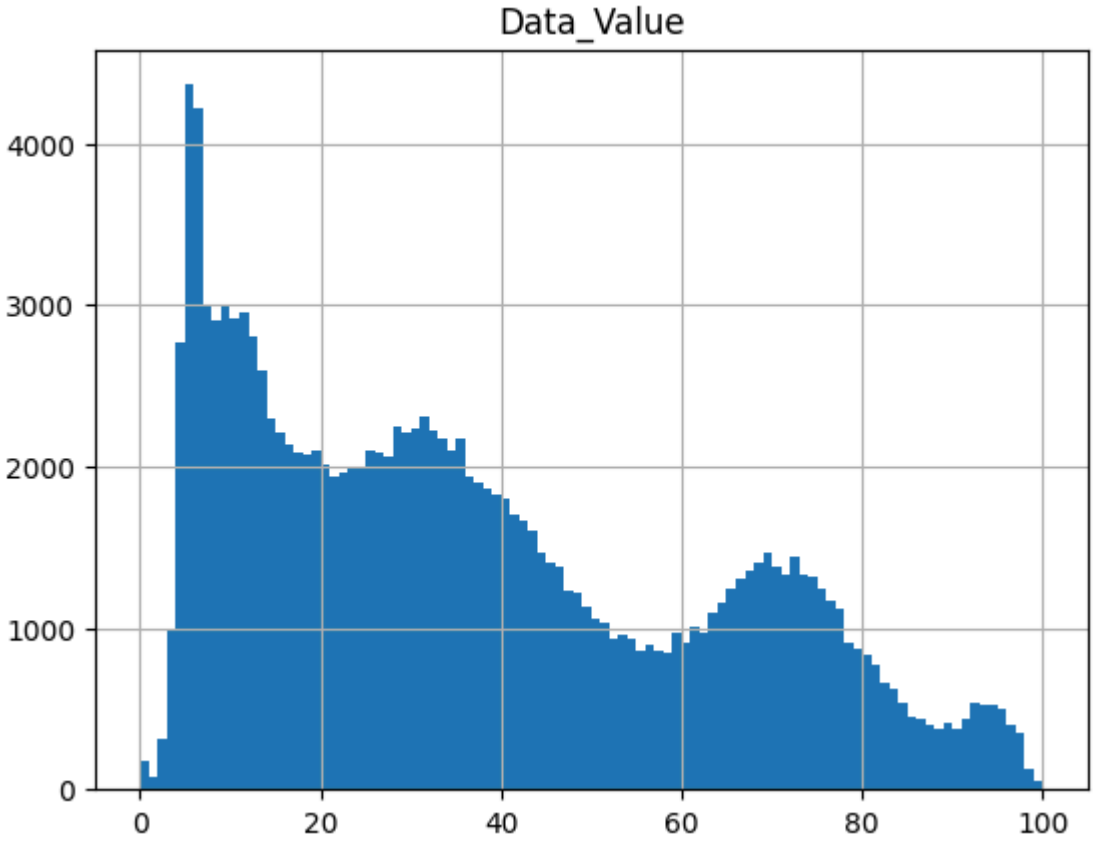
In [3]:
```
def draw_histogram(df: pd.DataFrame):
    for col in df.columns:
        if df[col].dtype != object and len(df[~df[col].isnull()]) > 0:
            df.hist(column=col, bins=100)


def draw_boxgram(df: pd.DataFrame, columns=None):
    if columns is None:
        df.boxplot()
    else:
        df.boxplot(column=columns)
```

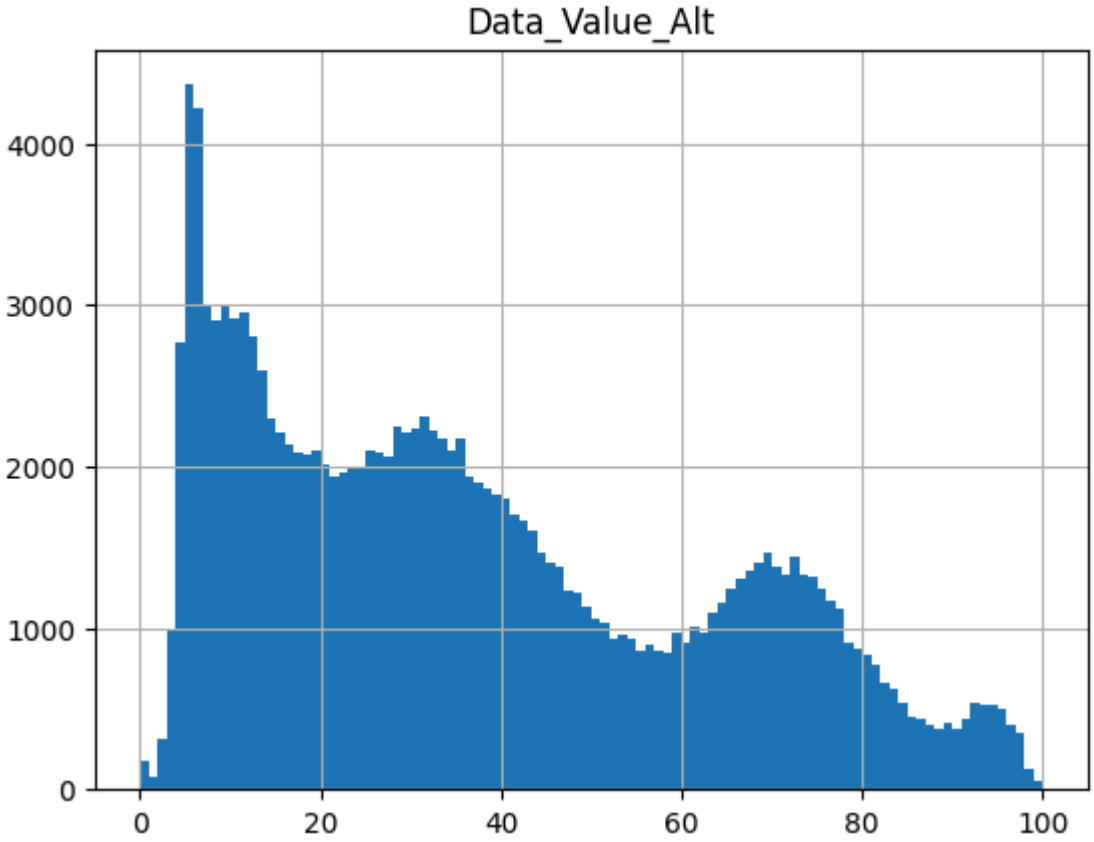
直方图

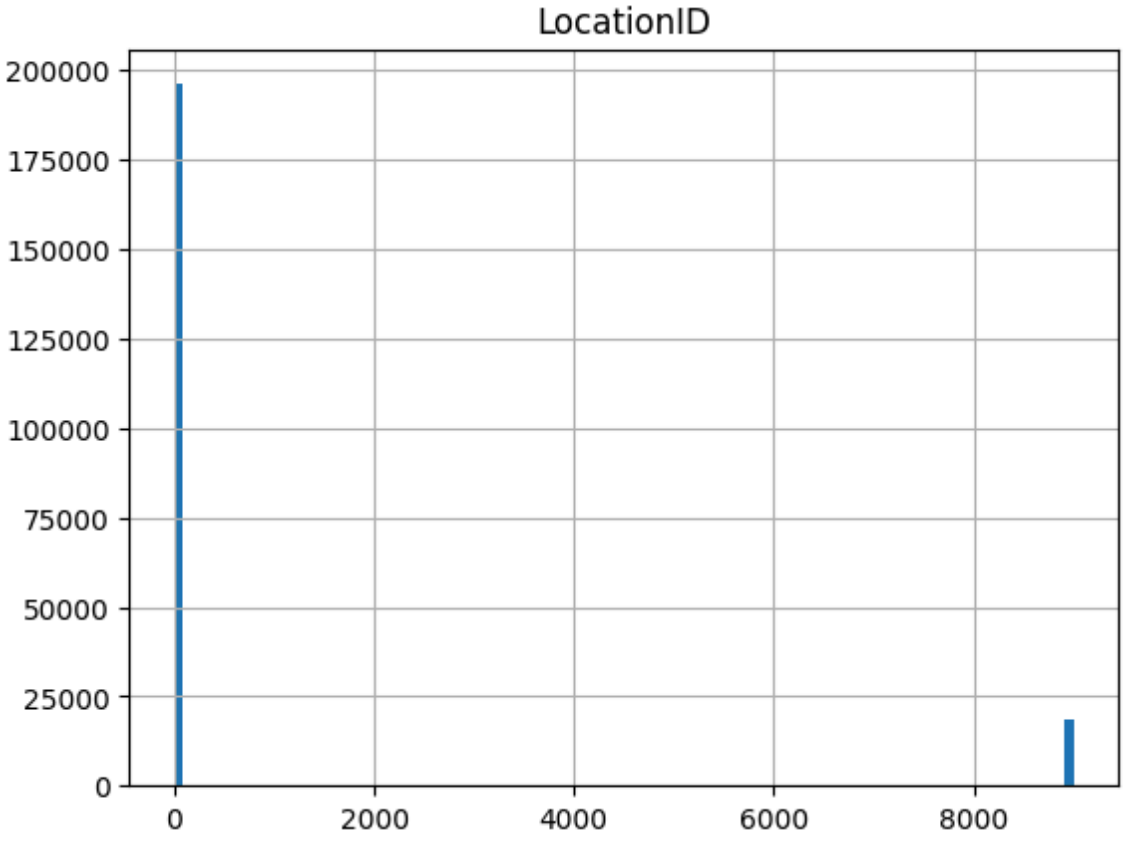
In [4]:
```
draw_histogram(df_alzheimer_data)
```




YearEnd
40000
30000
20000
10000
0
2015
2016
2017
2018
2019
2020

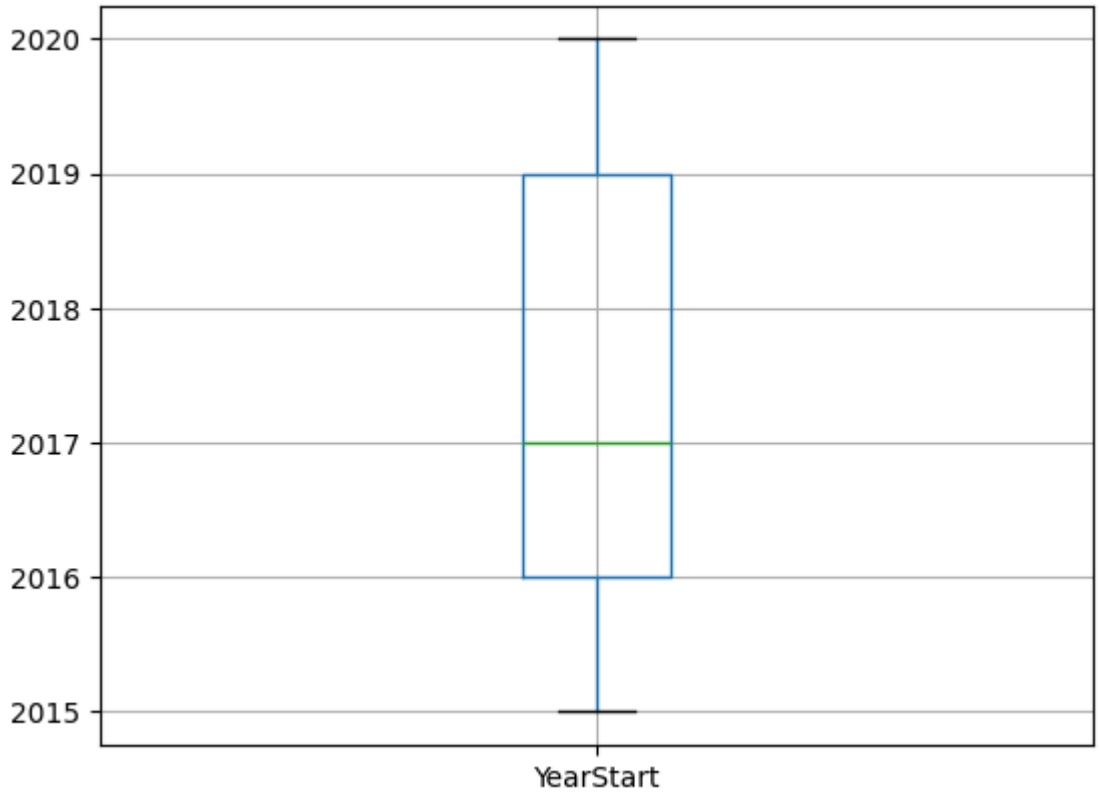

Data_Value
4000
3000
2000
1000
0
0
20
40
60
80
100



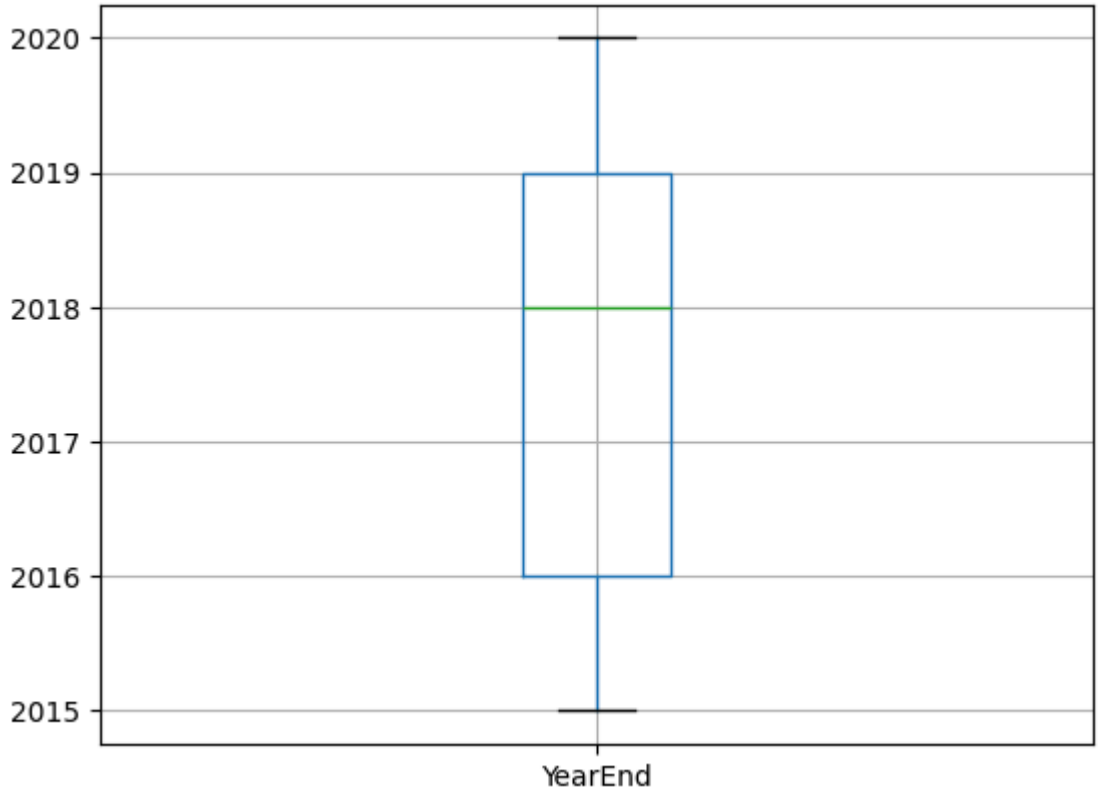


盒图

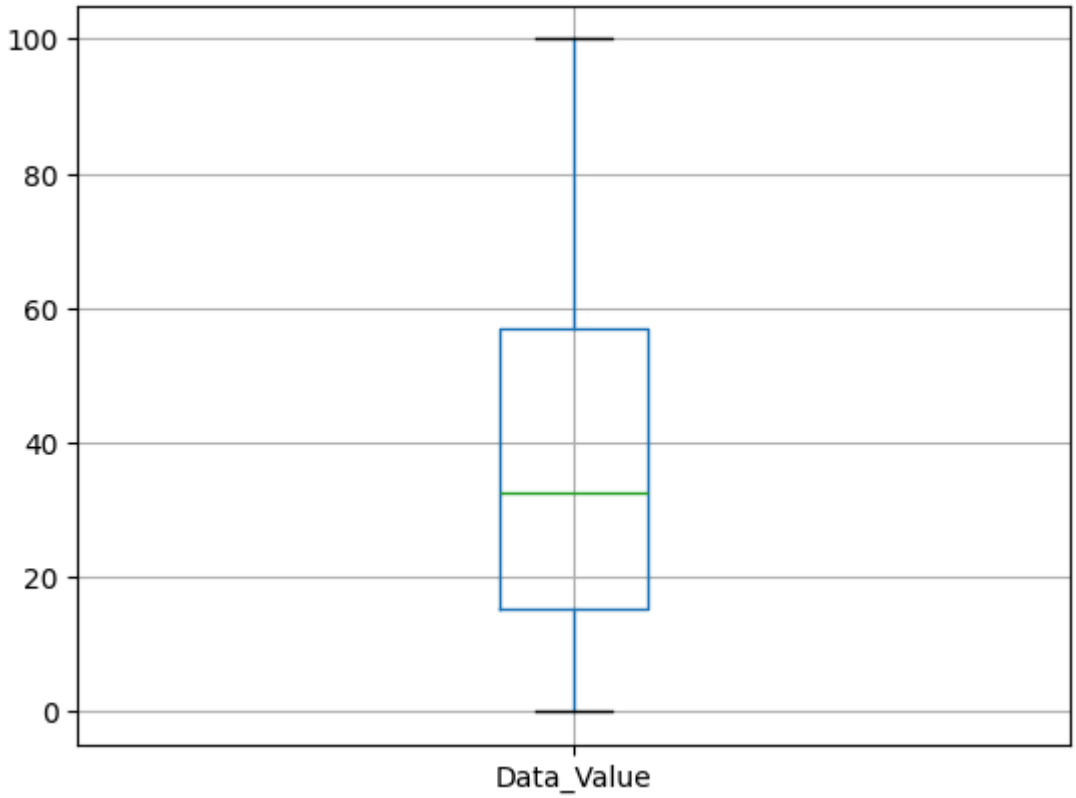
In [5]:
```
draw_boxgram(df_alzheimer_data, columns="YearStart")
```



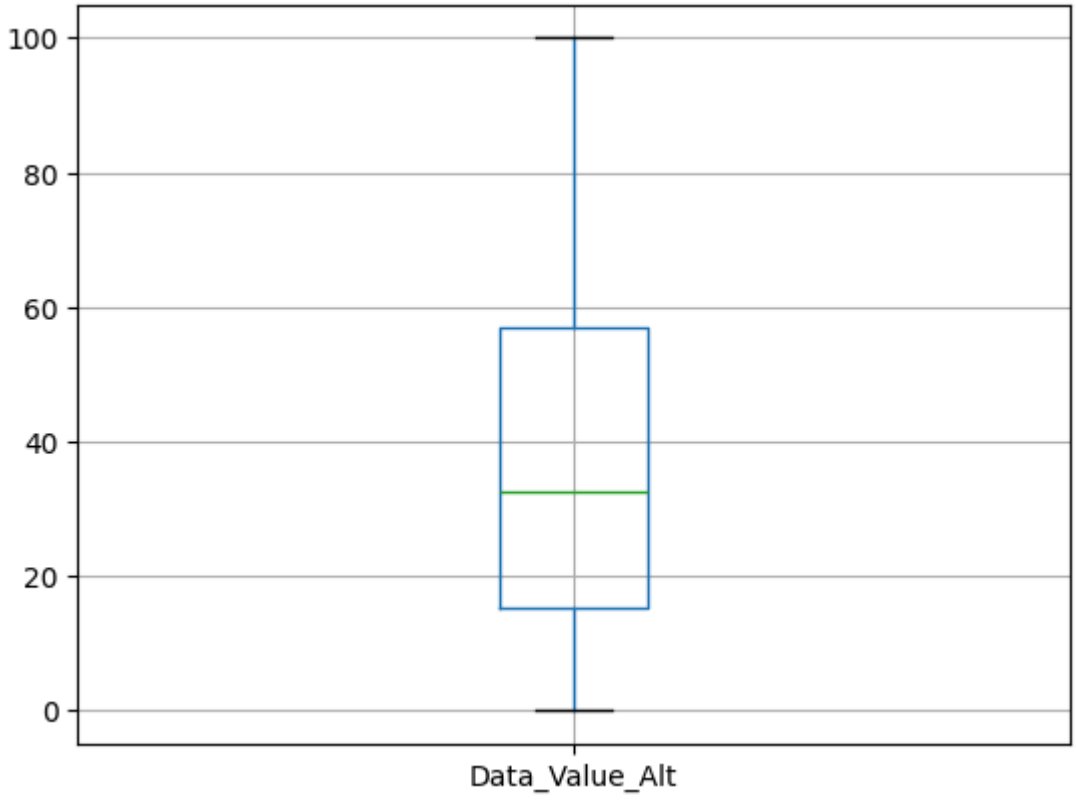
In [6]:
```
draw_boxgram(df_alzheimer_data, columns="YearEnd")
```



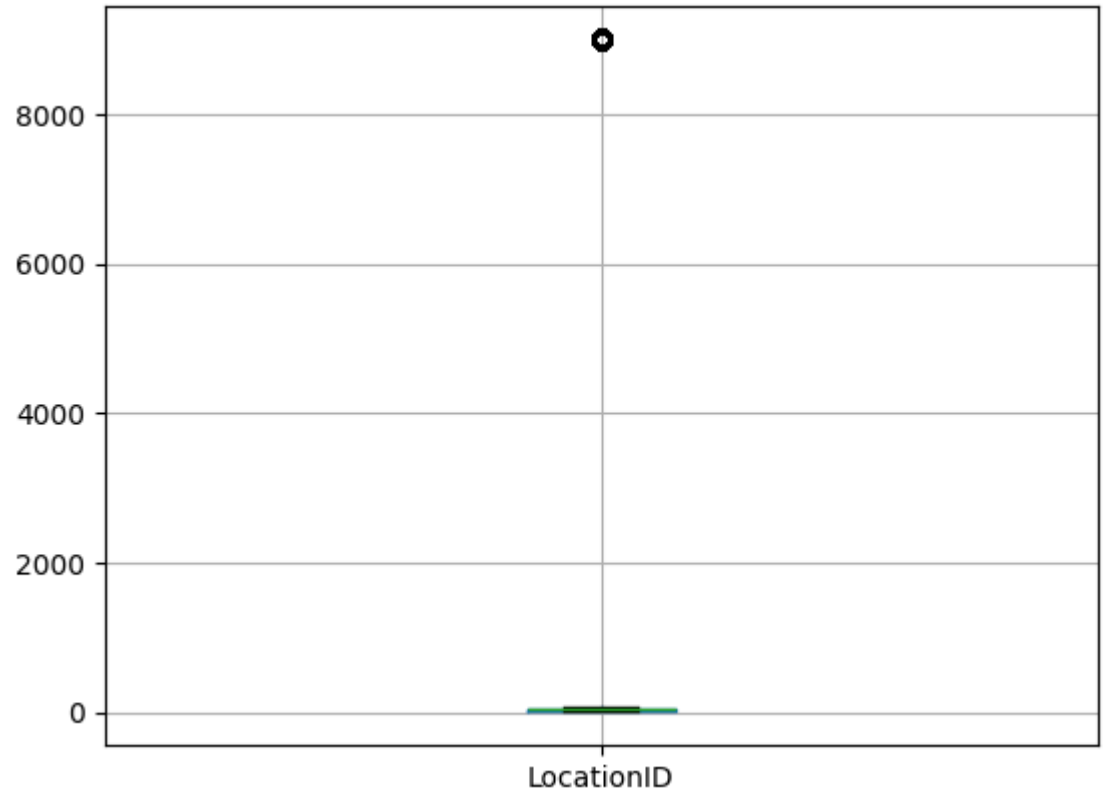
In [7]:
```
draw_boxgram(df_alzheimer_data, columns="Data_Value")
```



In [8]:
```
draw_boxgram(df_alzheimer_data, columns="Data_Value_Alt")
```



In [9]:
```
draw_boxgram(df_alzheimer_data, columns="LocationID")
```



2. 数据缺失处理

先寻找缺失的列

In [10]:
```
for col in df_alzheimer_data.columns:
    if df_alzheimer_data[col].isnull().sum() > 0 and df_alzheimer_data[col].dtyp
        print(col)
```

```
Data_Value
Data_Value_Alt
Sample_Size
```

首先是缺失数据剔除。由于Sample_Size列均为null，因此剔除该列：

In [11]:
```
print(df_alzheimer_data.shape)
```

```
(214462, 29)
```

In [12]:
```
df = df_alzheimer_data.copy()
del df["Sample_Size"]
print(df.shape)
```

```
(214462, 28)
```

删除了一列数据。

第二个是用最高频率值填补缺失值

In [13]:
```
null_lines = df_alzheimer_data["Data_Value"].isnull()
print(df_alzheimer_data[null_lines]["Data_Value"])
```

```
32        NaN
33        NaN
42        NaN
47        NaN
48        NaN
          ..
214456    NaN
214457    NaN
214458    NaN
214459    NaN
214461    NaN
Name: Data_Value, Length: 69833, dtype: float64
```

In [14]:
```
freq_max = df_alzheimer_data["Data_Value"].value_counts().index[0]
print(freq_max)
```

```
6.0
```

In [15]:
```
df = df_alzheimer_data["Data_Value"].fillna(freq_max)
print(df[null_lines])
```

```
32        6.0
33        6.0
42        6.0
47        6.0
48        6.0
         ...
214456    6.0
214457    6.0
214458    6.0
214459    6.0
214461    6.0
Name: Data_Value, Length: 69833, dtype: float64
```

可以看到缺失值均被填充为了2.0（最大频率值）。

下面使用平均数来填充缺失值。

In [16]:
```
mean = df_alzheimer_data["Data_Value"].mean()
print(mean)
```

```
37.34195562439067
```

In [17]:
```
df = df_alzheimer_data["Data_Value"].fillna(mean)
print(df[null_lines])
```

```
32        37.341956
33        37.341956
42        37.341956
47        37.341956
48        37.341956
            ...
214456    37.341956
214457    37.341956
214458    37.341956
214459    37.341956
214461    37.341956
Name: Data_Value, Length: 69833, dtype: float64
```

空值均被填充为了平均数

下面使用前后值来填充

In [18]:
```
df = df_alzheimer_data["Data_Value"].fillna(method="pad")
print(df[null_lines])
```

```
32        68.5
33        68.5
42         7.6
47        41.5
48        41.5
          ...
214456    18.7
214457    18.7
214458    18.7
214459    18.7
214461    10.6
Name: Data_Value, Length: 69833, dtype: float64
```